Inland Norway University

Rena

Randi Anglevik

Anette Landråk Nes

Master Thesis

How important is weather as a predictor for the demand for ski lift passes in the alpine skiing industry?

Master's in Economics and Management
Profile of Digital Management and Business Analytics

2021

Abstract

This thesis aims to examine and interpret the importance of weather variables as predictors for demand in the Norwegian alpine skiing industry. A specific skiing facility has provided a unique data set, containing their daily sales data from the winter seasons of 2014/2015 to 2019/2020. The sales data is used in combination with simulated weather forecast data to develop linear regression forecast models. The predictive performance of the models is compared statistically to analyse the importance of weather variables for predictive accuracy. The main findings show that the importance of temperature, snow depth and precipitation for predictive purposes is low. Seasonal variables, such as day of the week and public holidays, appears to be of greater importance as predictors of demand. The authors find no statistically significant improvement in the predictive ability of models with weather variables compared to models without.

Acknowledgements

We would like to express our sincere gratitude to our advisor, Dr. Daumantas Bloznelis, for his invaluable help and guidance. He has been available around the clock, and never failed to answer any of our questions. The process of writing this thesis has been challenging, but also highly educational. We have stepped far outside our comfort zones with regards to machine learning techniques being used, which has required us to dive into theories and topics that were unbeknown to us before we started working with this thesis. The learning curve has been steep, leaving us with far more knowledge in the field of prediction than we anticipated. Appreciation is also directed at Professor Erik Haugom and the alpine skiing facility for giving us access to data and for swift correspondence concerning further inquiries.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Demand in the alpine skiing industry exhibits strong variations, with big seasonal fluctuations (Malasevska, 2017). Alpine skiing facilities usually face increased demand during weekends and demand peaks at holidays, while activity typically is low in the early and late season. In addition, weather can be an important predictor of demand. Bad weather naturally makes skiing less enjoyable, leading customers to postpone the visit or seek other activities. Good weather, on the other hand, can attract more customers to the slopes than otherwise. Ideal weather conditions for visitors are generally characterized by a fine balance of little precipitation, but still with sufficient snow depth, and sun with clear skies, yet not with temperatures that will either melt the snow or be too cold for people to stay out in. A thorough understanding of the variables that influence demand for alpine skiing lift passes is crucial for facilities operating in the industry.

This thesis aims to get a better understanding of weather as a predictor for demand in the alpine skiing industry. The focus will be on the importance of weather variables for predictive accuracy. Producing accurate forecasts is hard, requiring statistical models with estimated parameters (Diebold, 2017, p. 14).

Our research will contribute to the iPaaSki project, in which the main objective is to create value in the alpine skiing industry by developing and implementing new and innovative pricing schemes (iPaaSki, n.d.). The aim of this thesis is not concerned with dynamic pricing per se, but with developing models that can be used as an operational planning tool by the facilities or by academics. The contribution is thus both practical and theoretical, providing models of practical use along with a deeper understanding of the role of weather for predictive accuracy in the alpine skiing industry.

1.1 The Norwegian downhill skiing market and weather climate

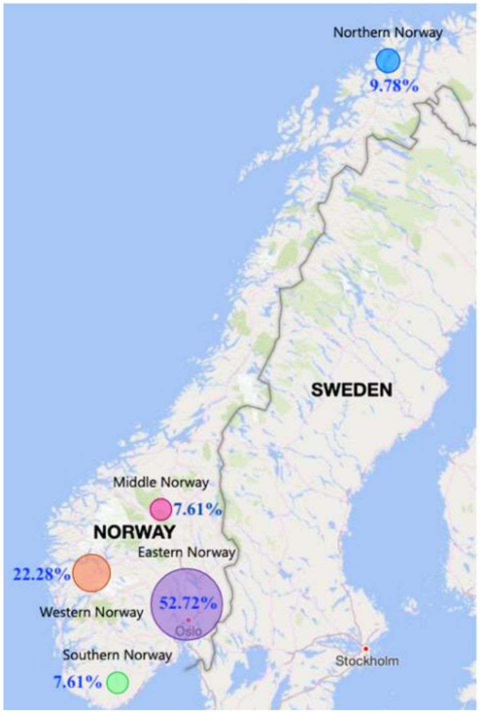


Figure 1: Distribution of skiing facilities by region

Norway can offer over two hundred alpine skiing facilities, in which the majority are small and medium-sized with merely one or two ski lifts (Norske alpinanlegg og fjelldestinasjoner, n.d.). The spatial distribution of all the facilities in Norway, gathered from (Ski Info, 2021), are displayed in the map in Figure 1. Over half of the facilities are located in Eastern Norway, where all of the facilities in the iPaaSki project are also located.

According to statistics, there have been registered some decreased popularity in skiing activities in the last decade (Dalen & Gram, 2020; Tuv, 2019). While the reason could be related to a decline in natural snow in the last seasons, it could also be due to changes in demographic elements or perhaps increased competition with international facilities. The adult customer group is seen as one of the most important customer groups, and an ageing population could therefore be of relevant significance for future developments in the market (Vanat, 2020). There have furthermore been found that skiing is most popular amongst those with higher education (Dalen & Gram, 2020). This can further be linked to the cabin market, as people investing in secondary property usually have a higher income, possibly due to higher education. Along with a growing cabin market, one can also expect an increase in the demand for alpine skiing activities, seeing that the cabin owners often buy them to gain access to alpine facilities.

From an industry report of the alpine industry in Norway of the winter season of 2018/19, the season ended with a 3% decrease from the previous season, before experiencing an increase in the early season of 2019/20 (Alpinanleggenes Landsforening, n.d.). In other words, it seems like some fluctuations from season to season in the industry are normal. There have however been some challenges in terms of less natural snow, varying weather, and heatwaves in certain locations, yet with upswings in demand during the holidays, especially around Easter (Alpinanleggenes Landsforening, n.d.).

In recent reports, the average temperature in Norway points to an increasing tendency, and the Inland region was pointed out to be one of the warmest areas in the country with a deviation of 7-8°C above normal in the winter season of 2019/20 (Grinde et al., 2020). There have also been signs of a significant increase in precipitation in general, but the Inland area on the other hand, experienced a decrease, having the lowest amount of precipitation nationwide during the same season. Increased temperature and less precipitation could lead to challenges with less natural snow. Seeing that the majority of the alpine skiing facilities are small and medium-sized, a lack of natural snow could have a significant impact on the industry. Even if the facilities have access to the needed resources, snow production is a complex process needing distinct circumstances to be fulfilled (Kulturdepartementet, 2011). The consequences of low snow depth for the resourceful facilities could therefore be just as bad as any other facility as long as the needed circumstances are crippled. If there is a need for snow production in the first place, chances are the climate does not grant the desired conditions for making snow of high quality.

1.2 Research question

The importance of seasonal variables, such as holidays and day of the week, on demand for alpine skiing, is well-documented in the literature. Additionally, the characteristics of consumers and the individual skiing facilities influence demand as well. These characteristics can range from real income, level of skiing proficiency, number of slopes, and accessibility. The importance of weather variables is, however, more debated. There are numerous studies concerning the effect of weather on demand for alpine skiing, but they are spread in terms of geography, time units used, aggregation levels, and measurements of the skier demand, making it hard to compare and draw concrete conclusions (Falk & Vieru, 2017). Whereas some studies find weather to be statistically significant for demand, others claim that the effects are small, and outperformed by other predictors (Malasevska et al., 2017).

This thesis aims to provide a better understanding of the importance of weather as a predictor of demand. To achieve this goal, the thesis will examine the following research question:

How important is weather as a predictor for the demand for ski lift passes in the alpine skiing industry?

The research question will be addressed by analysing historical sales data from an alpine skiing facility, in combination with weather forecast data. Machine learning will be utilized to develop predictive models, in which some models include both seasonal and weather variables, while others only contain seasonal variables. These are compared using the Diebold-Mariano test to determine if there is any statistically significant difference in the predictive accuracy of models with and without weather variables to analyse the importance of weather variables as a predictor of demand.

We hypothesize that adding weather forecast data to a predictive model will enhance its predictive performance. Customers plan their behaviour, especially recreational activities such as alpine skiing, an activity that most people do not have in their immediate environment. Seeing as bad weather typically decreases demand, weather forecast could lead to more accurate predictions by controlling for the effect of weather. On the other hand, the weather could prove to only be an important predictor when the weather is extreme, meaning that weather within the normal range is of little importance for predictive accuracy. It could also be possible that the importance of weather is so small compared to other factors such as seasonal variables that they do not yield more accurate predictions.

1.3 The alpine skiing facility

One of the alpine skiing facilities in the iPaaSki project provided us with raw data and is located in the Inland region. The facility will remain anonymous and will hereby be mostly referred to as the facility. This facility gave access to data with a time horizon of almost 6 full seasons. Some of the sources used to gather information about the facility will reveal the location and name and will therefore not be disclosed. The main sources are however their website and information they have provided directly.

The facility distinguishes between a high- and a low season, with prices adjusted accordingly, being lower in the low season compared with the high season. This is a type of pricing differentiation intended to draw more visitors during the low season, not only because lower prices will increase demand, but also because some ski-lifts are closed due to lack of snow and reduced usage. The visitors, therefore, receive less value in the low season than in the high season, which should be reflected in the price. The high season is set between December 26th and the last day of Easter, which varies from late March till late April, while the low season includes the remaining periods, being before December 26th and after Easter. Seeing that the season usually starts in mid-November and ends soon after Easter, the low season is very short compared to the high season. In the early season, the facility is only open at weekends, and on days with bad weather, they sometimes decide to keep closed. With fluctuations in weather, visitors can check the conditions through a web camera on the facility's website, with added information concerning temperature, wind, sun, precipitation, etc.

The alpine facility is of larger size and has stated that they are well equipped with resources to cover snow production. They also cooperate with another skiing facility nearby, offering more challenging slopes, and ski lift passes bought on either of these facilities can thus be used at both locations. By cooperating, they can provide a broader service which could affect the demand positively. The nearby facility has suffered from severe losses in revenue, resulting in having to close one of the ski-lifts a couple of years back. It was here pointed to a lack of natural snow and snow production machines of their own, which indicate that even though the two facilities share ski lift passes, they do not share resources.

Along with the facilities, other complementary services are offered in the nearby area. This includes many different winter activities, such as skiing school, snowmobile, snow rafting, ice fishing, dog sledging, and winter expedition. In addition, there is a climbing park, a shooting simulator, and one could also attend festivals, go on mountain trips, bobsleigh, and much more, but these latter activities are mainly available in the summer when the skiing facilities are closed. The alpine facility itself has not made any efforts of attracting visitors other than through ordinary marketing, but the complementary activities could affect demand, nonetheless.

1.4 Delimitation

There are numerous alpine skiing facilities, both domestic and international, so there had to be set a limit to what research objectives to include in this study. A natural focus was alpine skiing facilities in Norway. The assumption is, however, that demand for alpine ski lift passes is influenced by the same types of variables across countries, but there could very well be national differences, especially with regards to the weather. Further delimitation was made based on availability, as we ended up using the raw data of only one facility in the IPaaSki project. Limiting the number of facilities to one, allowed us to generalize our results across time instead of across facilities.

Because the demand for alpine skiing fluctuates throughout the season, we narrowed our focus further by type of pass and type of customer group. The day passes were used exclusively, leaving out other lift passes such as seasonal passes for a longer period or just a few hours. We found that day passes were most popular and therefore believed they would do a better job of capturing the fluctuations in demand than other types of passes. Within the day pass category, there was also an abundant number of different customer groups, in which many overlapped one another or was not consistent over the different seasons. The customer group of adults seemed to be the most consistent in the dataset and historically an important customer group, and the research was therefore further limited to this customer group.

2 Theoretical framework

This thesis aims to provide a better understanding of predictors of demand for ski lift passes in Norway. It is driven by the theoretical foundation that already exists in the literature and will therefore be given an introduction to the theoretical framework the thesis is based on. To begin with, a brief chapter on the earlier findings of demand for alpine skiing activities will be provided to help situate the reader in the context of alpine skiing demand. An introduction to demand and supply theory will then follow before the theoretical framework is concluded with some theory of machine learning.

2.1 Earlier findings of demand for alpine skiing activities

When it comes to the effect weather has on demand in the alpine skiing industry, there are some differing conclusions. Several additional factors are mentioned to be of importance when forecasting the demand for alpine skiing, and there lies a challenge in understanding tourist's perceptions and reactions to anticipate potential shifts in the demand (Gössling et al., 2012).

Results from research covered by Shih et al. (2009) suggest that weather variables such as temperature, snow depth, and wind chill have a statistically significant impact on sales of downhill ski lift passes. The authors did, however, find that day of the week and holidays have the greatest impact on the demand in the United Stated. Weather furthermore tends to have less of an impact on the demand for skiing when observed over a longer period, such as over a whole winter season (Falk & Vieru, 2017). It will therefore be relevant to consider the data frequencies being used (Gómez Martín, 2005). A study from Romania uses the same method as Shih et al. (2009), being multiple linear regression, only with yearly data frequency rather than that of daily. It concluded with temperature and tourism having a negative relationship, although variables such as day of the week and holiday were not accounted for in this research (Surugiu et al., 2010). Falk (2013) found that winter tourism demand is indeed significantly related to various weather conditions, however with an emphasis on the relationships being of minor significance. This could indicate that other variables might be of greater importance than the weather, also in the research of Surugiu et al. (2010), although not being reported due to them missing from the analysis.

Other than weather variables, relative prices and real income are significant determinants of the number of skier visitors in the long run. The change of relative prices has the largest impact on winter tourism demand, followed by real income and lastly snow depth (Falk, 2015). It is also argued that the demand for alpine skiing is dependent on factors such as the physical characteristics of the skiing facility, individual’s skiing ability, cost of skiing, leisure time, skiing budget as well as weather conditions (Malasevska et al., 2017). Related to cost and budgets, Holmgren and McCracken (2014) found that in Utah, when all skiers had access to several skiing facilities with similar snow density and weather, the majority chose from the facilities closest to the airport. Availability and transportation costs was likely a significant factor in this case. However, Falk (2013) found that the effect of travel costs was bigger for foreign tourists than for domestic.

Many factors that contribute to the increase of demand are not possible to control, but by being attentive to them, numerous measures can be taken to make advantage of it. Holmgren & McCracken (2014) encourage facilities to aid in the increase of demand by differentiating through expanding, making improvements such as faster chairlifts, snow parks, snowmaking machines and increased lodging opportunities.

2.2 Demand and supply

2.2.1 The basics of supply and demand

In microeconomic theory, a market is comprised of consumers and producers. The producers produce and offer a commodity or a service, and the consumers consume the commodity or service. The demand and supply in a market can be illustrated by the demand and supply curves, as shown in **Error! Reference source not found.** below, in which quantity marks the x-axis and price marks the y-axis. The market is said to be in equilibrium when the demand for a commodity or service equals the supply of that commodity or service (Pindyck & Rubinfeld, 2018, p. 25). This can be found where the two curves intersect, and the corresponding quantity and price is called the equilibrium, or market-clearing, quantity, and price.

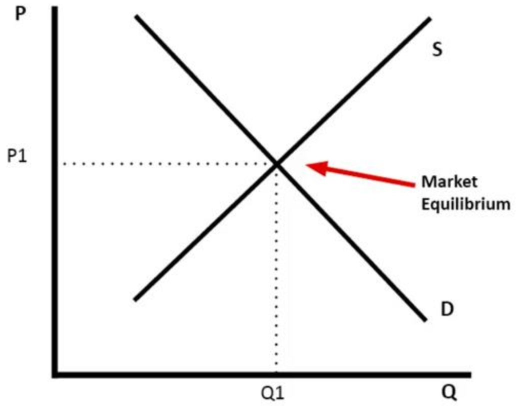


Figure 2: Market equilibrium (Pindyck & Rubinfeld, 2018)

The demand curve represents the relationship between the quantity of a good that consumers are willing to buy and the price of the good (Pindyck & Rubinfeld, 2018, p. 23). The demand curve has a negative slope, indicating a negative relationship between price and quantity, when the price drops, the consumed quantity increases. The supply curve, on the other hand, has a positive slope, meaning that the relationship between price and quantity is positive. As the market price increases, the producers are willing to sell more units.

Price is an important mediating variable in any market, and it represents the contradictory desires of consumers and producers. Consumers want a low price to consume big quantities of a good, while producers want the price to be high to produce big quantities. Both sides actively use whatever power they inhabit to influence the price in the desired direction. Customers can, for instance, shift to substitute goods if they are unsatisfied with the price/quality ratio of a commodity, while many producers engage in price wars and offer discounts to attract customers and gain market shares.

The relationship between price and consumption quantity is well-established in the literature, but various other variables influence supply and demand, both at the market level and at the local level. At the market level, big macroeconomic factors, such as economic growth and unemployment rates, heavily influences both supply and demand (Holden, 2016, p. 88). The supply side is also influenced by regulations set in place by governments, costs, and technology, to name a few. The demand side can be influenced by income levels, price of substitutions, and weather. The list of influencing factors is of course much longer, and it usually varies to some extent between markets. Supply and demand are also influenced by local factors, such as local regulations and availability, and factors that are specific for individual suppliers or consumers. Demand for alpine ski lift passes at a specific facility would in light of this, most likely, increase if the facility increased the number of slopes or the number of complimentary activities.

Markets are not necessarily always in equilibrium, and there are many potential reasons why. External shocks to a specific industry or an entire economy can, for instance, shift the demand or the supply curve. The direction and the size of the shift depend on whether the shock is positive or negative. Over time the market mechanism, also called the invisible hand, will move the market towards equilibrium, at least in a completely free and unregulated market (Smith, 2008). This is a slow process and explains why big shocks on the economy, such as the Covid-19 crisis or the financial crisis of 2008, have long-lasting effects on GDP, unemployment, and currency. An understanding of these principles is key to understanding how markets function and how they may recover from external shocks.

2.2.2 Individual versus market demand

There is a critical distinction to be made between the individual demand curve and the market demand curve. The individual demand curve is the demand function of one consumer and will vary from person to person. At the individual level, consumers can be modelled as if being ultimately interested in maximizing their utility. This utility can stem from numerous sources, ranging from leisure activities such as alpine skiing, reading, or going to a concert, to the consumption of more physical products, such as eating food or buying new clothes. The utility individuals gain from different products or activities is highly variable, depending on their individual preferences (Pindyck & Rubinfeld, 2018, p. 79). Some may favour skiing over reading a book, while others would much rather spend resources on going to a concert. Regardless of their disposition towards different sources of utility, the goal is always to maximize the utility. Consumers are, however, restricted by budget and time constraints, which means that they have to prioritize the consumption of goods or activities they believe will bring the most utility (Falk & Vieru, 2017). The market demand curve, on the other hand, is the aggregated demand of all individuals, thus representing the total quality demanded by all consumers.

We also need to draw the line between the market demand function and a price response function (PRF). While the first represents the entire market's response to changes in price, the latter describes how demand changes for a single producer as the single producer charges different prices (Haugom, 2015, p. 54). The market demand function represents changes in demand at the market level, while the PRF represents changes in demand at the individual producer's level. This is an important distinction because two companies competing in the same market can face different price-response functions. The difference in PRFs can stem from several factors, including how effective any marketing campaigns are, how the customer perceives the quality the different companies deliver, and location (Malasevska, 2017).

2.3 Machine learning

To be able to make accurate predictions, several factors of significance need to be considered. Not to mention the complexity of the data itself, we are met with great amounts of information that is not necessarily structured. With the help of machine learning, it could enable the uncovering as well as interpretation of valuable underlying patterns that otherwise would be difficult to unveil with our bare minds. Edwards (2018) puts it in short, explaining that machine learning is a tool for turning information into knowledge. In addition to assisting in revealing the relevant results, it can also uncover the underlying patterns, providing a deeper understanding of the problem by working its way through a learning process to enhance its performance.

Depending on the desired outcome, machine learning can be applied through different forms of learning. While supervised machine learning uses both established inputs and outputs to predict something new, unsupervised learning has no output data, leaving the algorithm with no guidelines (James et al., 2013, p. 26). With unsupervised learning, the problem is typically less defined than in supervised learning, which can expose relevant patterns that would otherwise have remained undetected. In some cases, a combination of these learning methods could address the issue better, with a small part of labelled data being merged with a large unlabelled dataset to enforce semi-supervised learning (Edwards, 2018). This could be useful when a certain bias is desired, but still leaving the possibility for new discoveries open. A more complex type of machine learning uses rewards and punishments through reinforced learning to generate desired behaviours. Although most problems fall into the supervised and unsupervised learning categories, semi-supervised as well as reinforced learning have been able to produce some remarkable results. Within the domain of supervised learning, several classical statistical learning methods operate, such as linear and logistic regression, GAM, boosting, and support vector machines. With unsupervised learning, on the other hand, having no output data to supervise the analysis, other statistical learning methods are needed, such as clustering for instance (James et al., 2013, p. 27).

2.3.1 The no free lunch theorem

There are many different algorithms to choose from when approaching a problem, and seldom one superior algorithm. The theorem of *no free lunch* is highly acclaimed in machine learning, which states that no one single algorithm is universally the best-performing algorithm for all problems (James et al., 2013, p. 29). The idea behind this theorem is that all machine learning algorithms are based on a priori assumptions, and the performance of a machine learning algorithm is highly dependent on how well these assumptions align with reality (Mavuduru, 2020). Choosing an algorithm also means choosing a set of assumptions regarding the problem situation. If the assumptions are well aligned, the model performance will be good, but if they are misaligned, the model will not perform well. A model can thus perform well on a problem in which the assumptions hold up, but there is no guarantee the model will perform well under other circumstances, as the a priori assumptions may not work. The price paid for lunch is thus the limiting assumptions accompanying an algorithm, which simplifies reality and fail in certain situations. The choice of the better model is dependent on the research problem and the size and structure of the data at hand, and the best performing algorithm is often revealed through plain old trial and error (Seif, 2021).

2.3.2 The bias-variance trade-off

The goal of prediction models is to gain an estimation that provides the best possible forecast of the unseen test set, in which the training set is only used to discover the patterns that help establish a method for this purpose. For the error in the test to be as low as possible, a statistical learning method is needed to achieve both low variance and low bias at the same time (James et al., 2013, p. 34). Low variance does however come at the expense of high bias, and vice versa. The goal is therefore to find a good balance between the two.

Bias refers to errors that will follow when working with real-life problems. Seeing that not all information can be accounted for in complicated issues, simple models typically cause a misrepresentation in terms of bias, as a consequence of simplifying the relationships. Variance can be explained as the variability of values predicted by a model across different possible training sets (James et al., 2013). When the model's complexity is high, it can lead to high variance by having an over-focus on every part of the training set. Logistic and linear regression are typical examples of simpler models that tend to have more bias, while more complex models such as neural networks tend to overfit, thus resulting in high variance. Too much variance or bias can cause the predictions to fit the data set too well or too poorly, which is referred to as overfitting and underfitting. Underfitted models suffer from high bias, while overfitted models usually lead to high variance. The best models for a given problem are therefore to be placed somewhere in the middle of the two extremes of bias and variance (Mavuduru, 2020). This balance is what we refer to as the bias-variance trade-off.

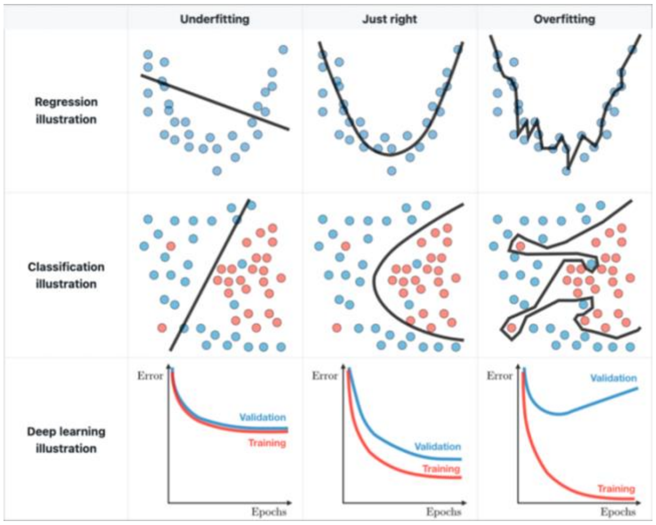


Figure 3: Underfitting and overfitting (Amidi & Amidi, 2018)

Models affected by high bias do not fit the training data well, leading to a particular high error on test data. When applying the new knowledge gathered from the dataset, the training data in the machine learning process needs to be as generalized as possible to avoid unusual data points being overly accounted for while also making sure significant patterns are not being ignored (Edwards, 2018). Whereas high variance over-focuses on the data points by including outliers and data not relevant to the pattern due to failure of generalizing the data, high bias can miss important underlying patterns by generalizing the data too much. By having a certain degree of both bias and variance, they can collectively make a model that follows the trends better and thus gains validation that is more realistic when applied to new data. By including more data and/or regularization, it can help stabilize high variance, while possible ways to combat high bias include increasing the model's complexity, adding more features, or training the model longer (Amidi & Amidi, 2018). An increase of the model complexity does reduce the bias at the expense of increasing the variance and vice versa. However, with the bias-variance trade-off in mind, the disadvantages of each occurrence are made as low as possible to produce the best model complexity and thus the lowest total error.

2.3.3 Time series

Time series forecasting is an example of machine learning in which the data is used to track events or measures that are to be observed and aggregated over time (Lai, 2020). To forecast future values of the time series, the dynamic relationships in the past or present data should be representative of the future. However, seeing that structural change patterns can be of either gradual or abrupt character, this is not always the case (Diebold, 2017). Trend, seasonal and cyclic patterns are mentioned to be different types of time series patterns (Hyndman and Athanasopoulus, 2018). When there is discovered either a long-term increase or decrease in the data, linear or not, it is referred to as a trend. If the pattern from the data appears to be affected by seasonal factors, like the day of the week or time of the year, it indicates a seasonal pattern. When no indication of any fixed frequencies exists, but the time series data still display a pattern, it goes under the term cyclic pattern (Hyndman & Athanasopoulus, 2018). One type of time series pattern does not necessarily exclude another, but when it comes to choosing a forecasting method, it is important to be aware of which pattern one is working with to find a method that is capable of apprehending the underlying patterns and thereafter more likely generate a reliable result.

When evaluating forecast accuracy, it is common to separate the data into two parts, having one larger part (often 70%-80%) for training data and a smaller part (often 20%-30%) for testing data. The size of the two parts does however depend on the length of the sample and the desired forecasting scope. The model is made based on the training data, in which the goal is to estimate parameters of a forecasting method before using it on the test data to assess how well it performs on new, though similar data (Hyndman & Athanasopoulus, 2018). In some situations, the available data may be limited, rendering the data size too small to make a reliable forecasting model at the given point in time. Time series cross-validation is a way to use current data to predict future data, one step at a time. Figure 4 visualize this principle, in which the blue observations are the training set and the red form the test set. However, when the training set is small, the earliest observations are not considered test sets due to unreliability. After forecasting for the later data points, the accuracy is checked (when the training set is large enough), before the same forecasted data points are added into the next training dataset. This can be seen as cross-validation on a rolling basis, in which the forecasted data is being used to forecast further data points, thus rolling forward in time (Hyndman & Athanasopoulus, 2018).

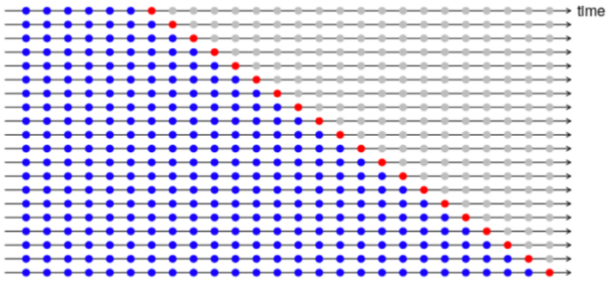


Figure 4: Time series cross-validation (Hyndman & Athanasopoulus, 2018)

When monitoring current data to be used in time series, it is more probable to detect relevant patterns effectively when the period in the time series is longer, thus being more difficult in short time series (Kirshners & Borisov, 2012). When the time series is shorter, it can result in arbitrary factors or outliers becoming seemingly significant and further make the result lack credibility. To secure credible results from time series, it is essential to ensure the integrity of the data being studied (Kirshners & Borisov, 2012). For instance, temperature from day to day usually does not differ very much, meaning the temperature on January 2nd can be highly correlated to that on January 1st while comparing temperature of a date in June will not likely be as correlated with a date in January.

2.3.4 Fourier series

Fourier series can be used when modelling seasonality, especially for seasons that inhibit a long seasonal period (Hyndman & Athanasopoulus, 2018). Najera (2021) explains that *"the Fourier Series is simply a long, intimidating function that breaks down* ***any*** *periodic function into a simple series of sine and cosine waves".* This series of sine and cosine waves converts a signal from the time domain to the frequency domain and this converted signal can be used to model seasonality. Other methods, such as ARIMA, are better suited for data with shorter seasonal periods, for example hours in a day, or days in a month (Hyndman, n.d.). The seasonality of our data stretches over a considerably longer period, with daily observations over a period of six months. With time series data, such as the one we are using in our research, there will be a considerable within-year seasonal cycle, and Fourier terms are well-suited for modelling these.

When using a Fourier series, one must determine how many Fourier terms to use. An increase in the number of terms leads to a better fit to the data, as it allows for greater flexibility within a season. It naturally follows that too high a number of Fourier terms can lead to overfitting the data, while too few terms can lead to underfitting. The inherent trade-off between variance and bias also applies when setting the number of Fourier terms, which needs to be considered when used for modelling. Allowing a great number of Fourier terms will lead to low bias at the expense

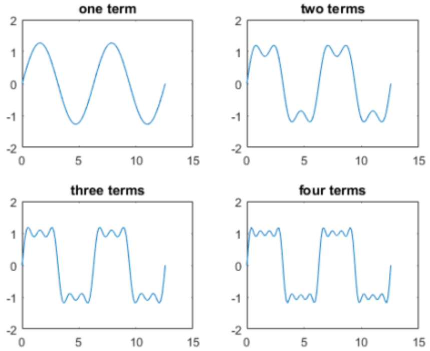


Figure 5: Fourier terms (Bower, n.d)

of high variance, and vice versa. As always, finding a good balance between the two is key, since both overfitting and underfitting causes problems.

2.3.5 Regression analysis

Regression is as mentioned a type of supervised learning with a numeric output that is useful when predicting numerous independent variables, such as temperature for a given day, probability of an event, and much more (Edwards, 2018). There are many different ways to use regression, depending on the goal and/or the data, and some approaches relevant to this research will be further addressed.

2.3.5.1 Linear regression

One of the most basic approaches is simple linear regression concerns a single predictor variable X being used as a base to predict a quantitative response on Y (James et al., 2013, p. 61). This approach assumes that X and Y can be expressed by a relatively linear relationship. The linear relationship can be written mathematically, as shown below, in which b_0 and b_1 are the coefficients that represent the intercept and slope terms, respectively.

$$Y \approx b_0 + b_1X$$

The coefficients and the p-value that the regression analysis provides, will help in interpreting whether the relationships are statistically significant as well as the nature of the relationship(s). While the coefficients describe the mathematical relationship between the dependent- and the independent variable, the p-value for these coefficients reveals if these relationships are indeed statistically significant at a given significance level (Frost, 2017).

Before testing relationships using linear regression, there will be a null hypothesis (H_0) claiming that the independent variables do not correlate with the dependent variable. To determine whether this is true or not, the p-value of each independent variable is used to test H_0. The p-value has to be less than or equal to the significance level to claim a relationship between independent variables and the dependent variable and thus reject H_0 with a high degree of certainty. The significance level can vary depending on how much evidence one requires before rejecting H_0. The lower the significance level, the more evidence is required from the data. A significance level of 0.05 is typically used, which means that there is a 5% risk of rejecting H_0, thus concluding that a correlation exists when it does not (Frost, 2017). This does on the other hand mean that we can with 95% certainty claim that there is a correlation, which is fairly good odds. If the p-value is greater than the significance level, there is not enough evidence in the sample to conclude that there is a correlation. However, this does not mean H_0 is true, but merely that it cannot be rejected. H_0 can never be proven, only disproven. When dealing with several independent variables, a common practice is to remove variables that are not statistically significant to keep them from reducing the model's precision (Frost, 2017). Hyndman (2011) does however discredit this, claiming that statistical tests are not made to select variables but to test hypotheses. In forecasting, it is possible for an insignificant coefficient associated with a variable to be useful, as it is also possible for a significant variable to be better omitted.

Linear regression models use the ordinary least squares (OLS) approach to calculate the coefficient estimates from the data sample (Oleszak, 2019). The goal is to estimate the parameters in a way that minimizes the sum of squared residuals. Linear regression models are a relatively inflexible approach, as it only generates linear functions. They usually have higher bias than variance, which makes them prone to underfitting the data (James et al., 2013, p. 35). The main source of error from linear regression models is therefore not its sensitivity to small variations in the training data but stem from the prior assumptions in the model being misaligned with reality (Mavuduru, 2020).

2.3.5.2 Ridge regression

Regularization is an extension of the linear model framework, and a technique to combat overfitting a model. Specifically, linear regression operates by selecting coefficients for every independent variable that seeks to minimize a loss function, and since large coefficients can cause overfitting, regularization is used to modify the loss function by penalizing the large coefficients (James et al., 2013, p. 215). Ridge regression is a type of regularization, often referred to as L2 regularization, and it uses the hyperparameter lambda (λ) as a way to tune the penalty (Machine Learning with R, n.d.). The value of λ is chosen by using cross-validation, aiming to minimize the sum of square errors on the validation sets. A λ of 0 indicates that the penalty term has no effect, and that ridge regression will produce the same results as OLS. As λ increases, the penalty term becomes more effective, shrinking the coefficients closer to zero. The shrinkage penalty of the larger coefficients is λ times the sum of squares of the coefficients (Machine Learning with R, n.d.).

Originally, ridge regression was developed to combat data when independent variables are collinear, thus making ridge regression a tool to combat multicollinearity in linear regressions. For predictive purposes, multicollinearity is not a problem, but the Ridge estimator presents a shrinkage estimator which can make it useful in forecasting after all (Elliott & Timmermann, 2016, p. 72). Ridge regression is therefore optimized for predictions, as the shrinkage of coefficient estimates towards zero combats overfitting and makes the model work better on new data compared to unregularized models (Gupta, 2017).

Even though the OLS method finds the coefficients that seemingly fit the data best, it does not consider if any variable is more or less important than others, thus being unbiased (Qshick, 2019). Ridge regression's advantage over least squares is rooted in the bias-variance trade-off. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance along with increased bias (James et al., 2013, p. 217). From what we know about the bias-variance trade-off, having no bias does not produce the lowest total error, and ridge regression provide some added bias on the important variables to modify the model for the better.

2.3.6 Loss function

The goal with forecasting is naturally to make as accurate predictions as possible. Predictions will, however, never be completely identical to the actual outcome, and will therefore always have some level of error associated with them. Depending on the situation, certain errors can be far more costly than others, and the loss function tells us how costly or painful certain errors are by adding penalties accordingly (Diebold, 2017). An error in the field of medicine, for example, can literally make the difference between life and death. Elliott & Timmermann, (2016, p. 13) defines the loss function (L) as a description of how costly it will be to implement an imperfect forecast (f) based on the outcome (Y), possibly with other observed data as well (Z). Because one wants to avoid making errors that result in higher costs, one might end up favouring a less accurate model as long as it has an emphasis on avoiding more costly errors.

The two main types of loss functions for regression analysis are quadratic and absolute loss. While quadratic loss measures the average of squared errors, absolute loss measures the average of absolute errors (Parmar, 2018). The quadratic loss thus penalizes larger errors more than absolute loss and is thus more sensitive to outliers.

In the case of predicting the number of visitors in alpine skiing facilities, overpredicting visitors could result in investing more than they otherwise would in the production of snow, extra staff in cafeterias or shops, and perhaps even extension or renovation of property and slopes when looking at the long run. Underpredicting, on the other hand, can lead to poorer customer experiences if customers are forces to wait in line at lifts and cafeterias because of understaffing. Norwegian alpine skiing facilities are far from utilizing their full capacity, but underprediction can still be painful given that resources such as staffing are needed for daily operations (Malasevska et al., 2017). If the facility is to use dynamic pricing schemes, prediction error could have great impacts on profits, as the price is set according to predicted demand.

3 Method

We have chosen to apply a quantitative approach to our research problem. Quantitative approaches are concerned with phenomena that can be measured and quantified and are frequently used when the goal is to map the prevalence of phenomena or to examine the relationship between different objects or factors (Johannessen et al., 2020, p. 23). A clear drawback of quantitative approaches is its failure to capture information that is unquantifiable, but still important for the phenomena in question, but it has the advantage of collecting data from a large number of units and generalizing the results from sample to population (Oppen et al., 2020, p. 31). Given that we are interested in weather as a predictor for demand, we need to establish the relationship between demand and factors that can influence it. A quantitative approach was thus a natural choice.

Data from two different sources, the facility and the Norwegian Meteorological Institute, will be used to develop regression models. Some include both seasonal and weather variables, while others only contain seasonal variables. By comparing them we can gain a better understanding of the importance of weather variables for the predictive performance of a model. Our initial hypothesis is that weather data will improve model performance by contributing to a better understanding of the underlying factors that influence demand for alpine skiing lift passes. The models will be developed by using the validation-set approach, which means that we are making use of supervised machine learning algorithms for model development.

Research conducted by Makridakis & Hibon (2000) concluded with the most accurate forecasts not necessarily being produced by more statistically complex methods, but often rather by simpler ones, such as linear regression. The results do however differ from the length of the forecast horizon, but the rule-based forecasting (RBF) method, which includes linear regression, was time and again exemplified as a well-performing method compared to those of higher complexity in various scenarios. Similar studies, researching the importance of weather, has also utilized regression models, making the same approach a natural choice for our research.

3.1 Data Sample

When applying a quantitative approach to a research problem it is customary to make some evaluation of the sample and the sample selection process. There is a distinction to be made between a population and a sample. A population is the entire group of people or objects that a specific research question applies to. It is, however, hardly even possible to collect data from the entire population, so samples are used to make inferences about the population (Oppen et al., 2020, p. 68). A sample is a subset of the population from which we collect data and is used to make estimations of the data generating process in the population. The sample has to be representative to generalize the findings to the rest of the population (Johannessen et al., 2020, p. 58)

We have made a clear delimitation in our thesis, choosing to only use data from one alpine skiing facility from the Inland region in Norway. If we considered all alpine skiing facilities in the Inland region, or all facilities in Norway for that matter, as our population, we would have problems with representativeness, given that our sample selection process was one of convenience and not one of probability. This would in turn make it harder to generalize the results. However, using time series data brings about some subtleties regarding the distinction between population and sample. Given that our models are to predict demand at a specific location, the repeated observations from this location are the population, not other alpine skiing facilities. Our sample is thus the historical observations we have available from the population. We know that there is some data generating process that generates the observations at the specific facility, but this process is unknown to us. To foresee what the process will generate next we need to learn more about the data generating process.

Given that we are interested in one skiing facility, and that we observe that their daily sales data over a longer period, there are no sample selection problems. The generalization is over time, not across skiing facilities, so we are less concerned with the representativeness of the skiing facility for the purposes of generalization to other facilities. We do, however, hypothesize that our findings may apply to similar skiing facilities in the Inland region of Norway, but any generalization to other facilities or regions needs to consider the representativeness of the facility we collected data from. Generalizing the results to other facilities that have different characteristics may lead to poor results. It is not automatically given that a small facility has the same data-generating process as our facility, or that the demand at a facility located in other parts of Norway is influenced by the same factor as a facility in the Inland region. Demand may, for instance, be more influenced by weather conditions in the other parts of the country, where the weather fluctuates more than in the Inland region, but customers may also be less sensitive to price if there is a greater distance between different facilities, limiting the available options severely. These are all considerations to be made if the results are to be generalized to other facilities than the facility we study, but we believe it could be an appropriate foundation.

3.2 Data collection

The dataset is comprised of data from two different sources. The first source is the alpine skiing facility, which gave us access to historical sales data, and the second source is the Norwegian Meteorological Institute (MET), which offers historical weather data for free from their webpage www.seklima.met.no. Data from these sources were combined into one dataset, containing the foundation of variables used to develop models.

Both the data from the facility and the data from MET can be described as time series data, as it is a sequence of numerical data points in successive order. Using time series data does bring about some subtleties regarding the methods applied. We are, for instance, less concerned with the representativeness of the data, but we still need to make some assessment of the data we use, and the data sources themselves, to ensure valid and reliable results. A discussion of the data from the two different sources, and a discussion of their validity, reliability, and privacy concerns follows below.

3.2.1 Data from the alpine skiing facilities

From the alpine skiing facility, we received historical sales data. The data is retrieved directly from their internal systems, and primarily contains information on historical sales of alpine ski lift passes. The raw data was comprised of daily sales from November 2014 to March 2020, thus covering almost 6 full seasons. The raw data contained 15 variables. A full list of these is shown in Table 1, along with a short description of the various variables.

Table 1: Variable descriptions of the sales data

Variable	Description
Date	Date of sale.
Register	Which cash register the passes were sold at. No filtering available, meaning that we received aggregated sales across all cash registers.
Pool	No filtering available, all observations marked with *All Pools*.
Ticket type	Type of pass sold. Contained 54 different types of passes. The most common ones were day pass, season pass, and single pass.
Customer group	Customer group. 13 different types. Differentiate between adults, senior citizens, children, and youth. Own group for companies.
Schedule	The time of day the pass is valid. Day, evening, X-hours.
Sales (1)	The number of sold passes.
Sales (2)	Revenue from the passes.
Annulled (1)	The number of passes annulled.
Annulled (2)	Amount of revenue annulled.
Refunded (1)	The number of passes paid back. Contained no variation. All observations marked with 0.
Refunded (2)	Costs of passes paid back. Contained no variation. All observations marked with 0.
Total (1)	The total number of passes sold in each observation. Sales (1) minus annulled (1).
Total (2)	The total daily revenue for each observation. Sales (2) minus annulled (2).
Total (3)	Percentage of revenue to the total daily revenue. Total (2) divided by every total (2) observation with the shared date.

The data contained one row for each type of pass sold to each different customer group on each day. All day passes sold to adults on one specific day are shown on one row, while all day passes sold to children on the same day are shown on the row below. The total number of observations/lines for each day thus varied, depending on how many types of passes they sold, and to how many customer groups they sold the passes to. On the 3rd of April 2018, for instance, there were a total of 28 observations, while 10th of April 2018 there were only 8 observations. The same day of the week, one week apart, and a considerable difference in the number of observations. Some level of difference is to be expected, especially between different days of the week and between certain weeks. More customers are to be expected during the weekend, as more people have time off work to pursue recreational activities. We also expect higher demand for ski lift passes on public holidays, such as Christmas and Easter. When more people travel to the facilities to ski, the demand for ski lift passes increases, and we would expect higher variation in types of passes sold and groups of customers, given that there are more visitors at the facility.

The data contains historical sales of alpine skiing lift passes, but it does not necessarily reflect actual visits to the alpine skiing facilities. Some types of passes are flexible and allow for visits to the facility on days other than the one when the pass was bought. Season passes, for instance, are usually bough early in the season, but grants access to the facility throughout the season, while punch cards grant access to the ski lifts a certain number of times without specifying the day of consumption. Thus, we cannot determine when the customers actually visited the facility when buying these types of flexible passes. For day passes the data reflects actual visits in a better way, given that most people use the pass the same day they buy it. The raw data contains information on annulled or refunded passes, showing both the number of passes sold, the number of passes annulled, and finally, the total number of actual passes sold. There is, however, a possibility that some customers bought the ski lift passes but were unable to use them and unable to get them refunded, but this would only represent a small source of error, as most customers would get their pass refunded immediately if they were unable to use it.

There are several sources for price and demand data, ranging from market data to surveys, experiments and expert judgements (Haugom, 2015, p. 68). The different sources of data have their strengths and weaknesses, and these need to be addressed properly, as they can influence the results. Market data, such as the sales data received from the facility, is often used in these types of demand models, as the data is readily available and cheap to obtain (Haugom, 2015, p. 68). It also has the advantage of reflecting actual buying behaviour. If the data had been gathered by surveys instead, the data would not reflect actual buying behaviour, which would represent a problem with the data. Saying that someone is going to buy a certain number of goods at a given price does not mean that that person is going to do so. The data from the facility do reflect the historical buying behaviour of customers, but that does not necessarily guarantee that they reflect future buying behaviour. Market data is a good source of data when the market is stable, but a change in the market could, theoretically, render historical data useless (Haugom, 2015, p. 68). The market for alpine skiing has not undergone any drastic change over the years we have data for, even though there are always continuous changes to any market with regards to supply and demand. Covid-19 can, for instance, have great impacts on the industry, both in the short and long run, which could render our models less relevant. In addition to only reflecting historical buying behaviour, market data can also be of limited use if there only have been small price changes (Haugom, 2015, p. 68). Small price variations may give limited information on how consumers behave with regards to changes in price. This may be especially true for recreational activities, as consumers are free to maximize the utility in a more liberal way than they are with necessary goods like fuel and food. If the data only exhibits small variations in price, the data may be a poor basis to assess future behaviour to big price changes.

Gaining access to sales data directly from the facility grants data with both good validity and good reliability. The total number of passes sold is a good measurement of demand, and the data is generated automatically by daily operations. The fact that all sold, annulled, and total passes are given shows that the system handles annulled sales properly, either because of misregistration or refunded passes. The only reliability issue with the data is that the facility sometimes registers sales the following day if there has been low activity at the facility. Some sales are therefore registered at the wrong date. This is, however, only the case with a few days with low activity, and the number of passes affected is therefore too small to cause any major problems in the analysis. There are no concerns with the privacy of the data used, as the data is aggregated and can by no means be traced back to individual customers.

3.2.2 Weather data

The literature is full of references to the importance of weather variables on outdoor recreational activities, with some contradictory results. To test if including weather variables in a model leads to better predictions, we needed to obtain weather variables, as these were not given in the data from the facility. A distinction is to be made between actual weather data and weather forecast data. The first is measured in real-time and reflects the weather conditions observable at a given time, while the latter reflects a forecast of weather conditions at a specified time in advance. There are discrepancies between predicted and observed weather, and this discrepancy grows with the length of the forecast.

Both forecast and actual weather data can be useful to determine the demand, but only forecast data can be used in a predictive model. The reason is twofold. Firstly, the idea is that if a facility wants to determine how many customers to expect a certain number of days ahead, they must make use of weather forecasts to determine the impact of weather. When forecasting the number of customers visiting during the upcoming weekend, there is no actual weather data available to the facilities, only forecast data. Since the facility must use forecast data, so does any predictive model. Secondly, the weather forecast is central to the customers' decision to go skiing. The customers plan, and value their leisure time. According to microeconomic theory, they will make choices that maximize their utility, and weather conditions can increase or decrease the perceived utility of spending the day in the slopes. Bad weather can make it less enjoyable to go skiing, which reduces the perceived utility for the consumer (Shih et al., 2009). Reduced utility brought about by poor weather conditions can thus lead to potential customers choosing other recreational activities than alpine skiing, which will reduce demand for lift passes. Some customers will of course spontaneously decide to go skiing, so actual weather is of some importance. It cannot, however, be used in a prediction model, as actual weather information is not available in advance.

Forecast data proved impossible to obtain within the framework of this thesis. Instead, we turned our focus to MET's free service, www.seklima.met.no, where historical weather data can be easily extracted. Data were available at different aggregation levels, but daily data was the lowest time aggregated level in common for temperature, precipitation, and snow depth. The daily temperature is measured as the arithmetic mean of hourly temperatures, precipitation is the total daily amount, and snow depth is measured at a given time each day. Data can be obtained for selected regions and selected periods, meaning that we were able to collect data on the specific area of the Inland region where the facility is located, and simultaneously filter on the desired period. The historical weather data was further used to simulate weather forecast data.

The data collected directly from the SeKlima service have good reliability and validity, as the data is collected automatically in real-time from a trustworthy source. The only drawback with using the SeKlima service is the limited variables available. The service is being rolled out this spring, so there were a limited number of available variables. The list of potential weather elements on the site is long, but most of them did not contain any data. We assume that the service will become better over time and that they continuously work on providing more data.

The simulated forecast data is of bigger concern and represents a source of error in the models. The problem is that the forecast data is simulated and not historical forecasts. Any error in the simulation will thus lead to biased and incorrect predictions. There are always errors when simulating data, and this is especially true given that MET's forecasts vary in accuracy depending on weather type and season.

3.3 Data preparation and wrangling

Raw data is seldom ready to be analysed and modelled without any form of preparation. A central part of developing a model is therefore to ensure that the data is of good quality and in a format that is applicable for modelling (Hair et al., 2018). A model based on poor data will always give poor results. This is well known amongst those working with machine learning and modelling, and has led to the phrase *garbage in, garbage out* - if the data put into the algorithm is garbage, then the algorithm will give garbage in return (Rose & Fischer, 2011). There are several steps involved in a data preparation process, including cleaning, structuring, and enriching the data. Hair et al (2018) also stress the need to examine and explore the relationship among variables before applying any algorithms.

3.3.1 Creating new variables

Both the data from the facility and the data from MET were structured according to individual calendar dates but contained a limited number of variables. There are numerous variables discussed as possible predictors for demand in the literature, including seasonal variables, customer-related variables, and facility-related variables. The scope of this thesis did not allow for the collection of data on all possible predictors, as it would require the collection of sensitive data. Seasonal variables were, however, natural to include, as they account for a fair amount of the fluctuation in demand throughout the season, and they were computable based on the data already at hand. Other variables were also created, and some original variables were transformed. A full list of these can be found in Table 2, align with a short description. Some of the variables require a more detailed description.

Table 2: New variables added to the dataset

Variable	Description
Price	Added price information within each season.
The logarithm of total passes	To prevent any forecast of negative demand, the dependent variable of total passes was transformed into its logarithm.
Relative date	Date variable arranging each date in relation to January 1st within each season. Reflects the linear trend throughout each season.
Fourier term	Fourier terms were added to model seasonality.
High season	Dummy variable in which the low season is 0 and the high season is 1.
Closed	Dummy variable indicating whether the facility was open or closed to control for days with no sold passes.
Weekdays	Categorized in numbers from Sunday as 1 to Saturday as 7.
Christmas vacation	Dummy variable accounting for the days in the Christmas vacation.
Winter vacation	Dummy variable accounting for the days in the Winter vacation.
Easter vacation	Dummy variable accounting for the days in the Easter vacation.
Christmas day	Dummy variable accounting for Christmas day.
2nd Christmas day	Dummy variable accounting for 2nd Christmas day.
New Year's Day	Dummy variable accounting for New Year's Day.
Palm Sunday	Dummy variable accounting for Palm Sunday.
Maundy Thursday	Dummy variable accounting for Maundy Thursday.
Good Friday	Dummy variable accounting for Good Friday.
1st day of Easter	Dummy variable accounting for 1st day of Easter.
2nd day of Easter	Dummy variable accounting for 2nd day of Easter.
May 1st	Dummy variable accounting for May 1st.
Cold	Observations are marked as cold if the temperature is -10 °C. There were 76 cases of cold days.
Ice cold	Observations are marked as cold if the temperature is -15 °C or below. There were 12 cases of ice-cold days.
Rainfall	To distinguish between snowfall and rainfall, precipitation above 2.5mm with temperatures above 2°C should be registered as rain.

3.3.1.1 Price

For this study, having a variable with the price for the ski lift passes is central when forecasting demand, as the price is one of the variables customers could be greatly affected by (Falk, 2015; Holmgren & McCracken, 2014). There was, however, no variable in the data from the facility that contained price information directly, and it was therefore obtained through the facility's website. Seeing that the prices found on the website only were available within the current season, the prices of the previous seasons were calculated by using the consumer price index (CPI), which later were verified by the facility itself. There is some level of error related to the price variable. The facility sometimes offers free passes to accompanying employees, while others have gotten passes for 20% or 50% off. Discounted prices can lead to higher demand, as the service becomes cheaper. We were unable to control for the fact that some customers receive discounted prices because the data from the facility did not contain price information on individual sales, which could be a source of error.

3.3.1.2 Holiday and vacation

Holidays and vacations mark days when most people have time off work and school to pursue recreational activities. Dummy variables for each holiday and vacations were added to the data to control for the impact of these days on demand. Holidays only include public holidays, thus leaving out Easter Eve, Christmas Eve and New Year's Eve, as they are not public holidays, but rather anniversaries. Vacations, on the other hand, include these individual holidays along with and other crammed days and weekend related to certain holidays. The Winter vacation is two full weeks instead of one, as the week of the vacation depends on what part of the country one lives in.

There are correlations between holidays and vacations, as many public holidays are part of a vacation. Christmas day and the 2nd day of Christmas are for example both public holidays and part of the Christmas vacation. This may lead to some misleading coefficient estimates for the two variables in the models, as the models can struggle to distinguish the effect of one from the effect of the other. The high correlation between the two is, however, not a problem for predictive purposes.

3.3.2 Simulating weather forecast data

The weather data was used to create ARIMA models for temperature, precipitation, and snow depth, predicting one day ahead. The models were cross validated using rolling windows and evaluated on mean absolute error (MAE). ARIMA stands for autoregressive integrated moving average and is used to describe the autocorrelations in the data (Hyndman & Athanasopoulus, 2018).

Table 3: ARIMA models MAE

Weather Variable	MAE
Temperature	1.88
Precipitation	2.58
Snow depth	2.58

The performance of the ARIMA forecasting models could be improved. Quarterly verification reports published by MET show that their predictions are more accurate than the ARIMA models (Homleid, n.d.). This is not surprising, as weather forecasting is a complicated field of science, and MET make use of a wide array of variables in their predictions. The ARIMA models solely base their predictions on data from previous days, which is a big oversimplification. Considering this, the performance is not bad.

More accurate predictions than the ARIMA predictions were obtained by simulating forecast data. The actual weather data from MET was polluted with the same level of error as MET's forecasts, obtained from their verification reports. This way, the data reflects the accuracy of MET's forecasts, and the simulated forecast data could be used as regressors in the demand forecast models. Simulating forecast data means that the weather variables used in the model are measured at some other time than the dependent and seasonal variables. This allows for predictions, not just merely in sample adaptation. The accuracy of the forecast is dependent on a lot of factors and vary across weather types and seasons (Homleid, n.d.). We have used the mean of the standard deviation of error and MAE throughout the year. They are 1.7 and 1.5 for temperature, and 2.5 and 1.5 for precipitation, respectively. Performance measurements were not available for snow depth, so historical data was used instead. The variations in snow depth are, however, small, so using historical data instead of the simulated forecast is not a big problem. Using the mean of the performance metrics given in the quarterly reports does not account for the differences in accuracy caused by weather types and seasonal variations, which represent another source of error for the weather variables.

3.3.3 Missing observations and discrepancies in the data

There were several instances of entire observations missing from the sales data. Missing data can reduce the statistical power of a study, and can also lead to biased estimates, which ultimately lead to an invalid conclusion (Kang, 2013).

The main sources of the missing observations were caused by either the facility being closed on the day in question or by delays and errors in the registration system of the facility. The facility is closed during periods with low activity and is additionally obliged to close down ski lifts if the wind is too strong (from 18 meters pr second). However, this usually only affects the ski lifts with the highest altitude, meaning the facility still can accommodate visitors on these days. Delays in the registration system of the facility sometimes occur on days when the activity is sparse, and the number of passes sold is reported on the following day instead. However, seeing that the facility tries to filter out the days with few visitors by keeping closed, the number of passes with postponed registration is limited, making this particular error small for our purpose.

Based on the information available, it is not possible to distinguish the two sources of missing observations precisely from another, but seeing that both sources indicate very low demand, introducing any missing observation into the dataset with 0 sold passes could help the models to better capture the periods with lower demand and thus gain more accurate results. Since all missing observations were of days with zero or a low number of sold passes, the missing observations are not random, which induce bias in the forecasting model (Hyndman & Athanasopoulus, 2018). Seeing that delays in registration only happens on a few occasions, all missing observations were therefore registered as the facility being closed. This was further used in the creation of a new dummy variable to indicate whether the facility was open or closed on any given day throughout the season.

There were also a few instances in the data we received from the facility containing observations with a negative number of passes sold. This is the result of the facility refunding more passes than they sold that day. To avoid days with negative demand in the analysis we replaced all days with a negative number of sold passes as having 0 sold passes, although not being closed. Additionally, there were also some discrepancies found in the data between total passes and total revenue. Dividing total revenue by the total number of passes sold did not add up to the price set by the facility. This can be explained by the facility providing various discounts and free passes to employees.

3.4 Data visualization

Data visualization is a great way to gain a transparent view of a situation. It was therefore natural to display the sales data through dashboard visualizations using Power BI to help uncover trends and insights that are hidden in the data. This could concern effects on demand by day of the week or holidays for instance. It would also be interesting to see how the demand has changed over the seasons at this particular facility. This overview will hopefully guide us to the discovery of some interesting information that will account for better interpretation of the data before conducting the machine learning techniques. Beforehand, we naturally had some assumptions of trends concerning the activity across weekdays and holidays which will be addressed further under the relevant sections.

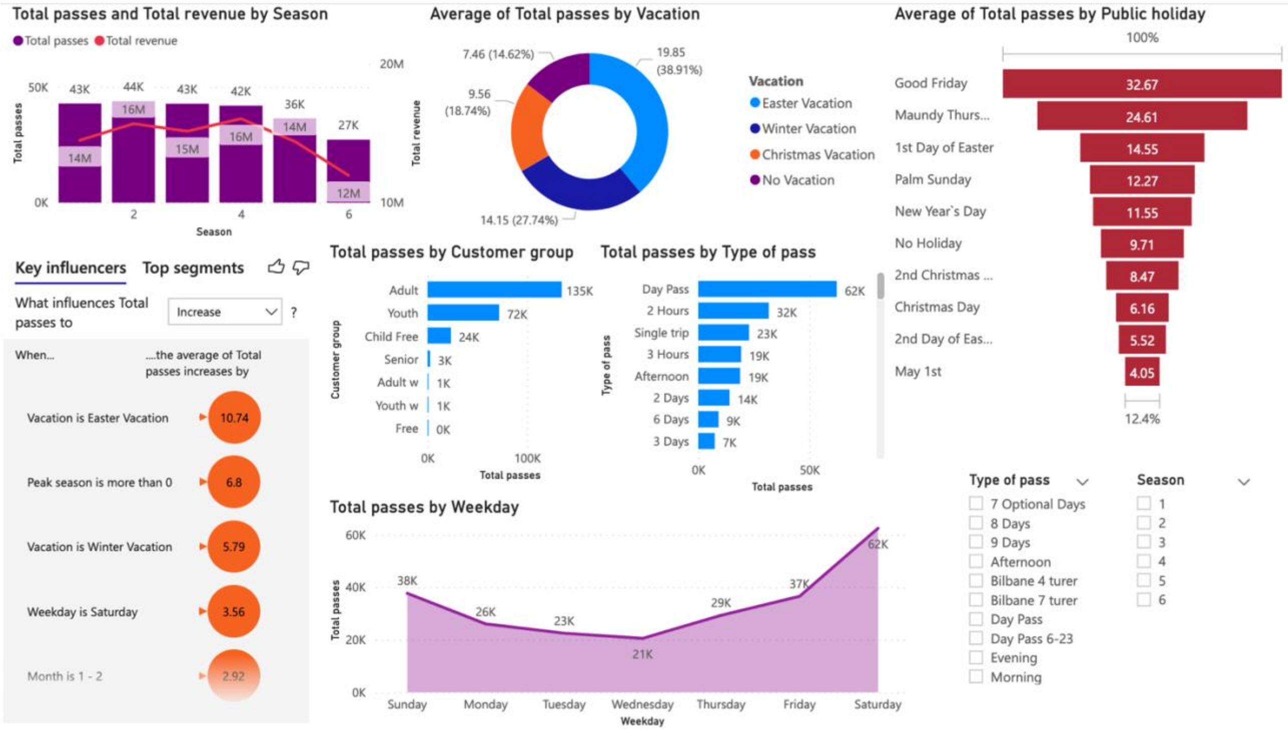


Figure 6: Dashboard of the sales data

3.4.1 Types of passes

The sales data contained about 50 different types of passes, in which many were overlapping or only including data from a very limited period. It was, therefore, important to investigate if there were any consistent types of passes across all the seasons and if any distinct passes should be removed for the data to become more reliable in the analyses. The dashboard in Figure 6 shows that of all the passes sold, the Day Pass was by far the most sold, followed by the passes of 2 Hours and Single Trip. In fact, all the five most sold pass types are valid up to one day, indicating that most customers buy the pass the same day they consume it. All of the mentioned passes proved to be sold actively through each season. Seeing that the day passes account for most of the tickets sold on a daily basis, it will likely be able to represent the overall pattern without noise from irregular types of passes.

3.4.2 Number of passes sold

The number of daily passes sold varies from 0 to 406. To assess what a reasonable error rate is when it comes to predicting the demand, the distribution of the number of passes sold each day are of interest. Figure 7 display the distribution in categories from 0-25, 26-50, 51-75, 76-100 and 100+ sales per day. Most days have a total number of sales between 0 and 25 passes, consisting of about 640 days, from a total of 985 days. The number of days with less than 25 passes sold, therefore, make up about 65% of the total number of days, whilst the four remaining consists of about 50-150 days each, adding to the remaining 45% of the days. Knowing that the mode of passes sold each day lies in the category 0 and 25, indicate that the range of acceptable errors are fairly low when it comes to evaluating the models on mean absolute error.

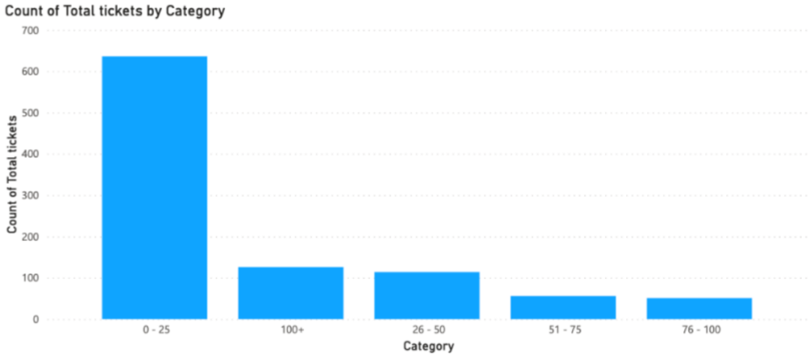


Figure 7: Count of total daily passes sold by category

3.4.3 Customer groups

The total number of customer groups is much lower than that of types of passes, with 13 categories. Several categories do, however, ultimately address the same customer group, but have been separated due to different spelling. There has not been provided any information regarding the differences of these overlapping categories, but most of them have very few observations, with only one providing the more representative selection to its belonging customer group. The bar chart of Total passes by customer group shows a clear prevalence in the Adult customer group, accounting for about 40% of all the passes, followed by Youth. The Adult customer group appear to be a natural segmentation for our analysis, filtering out both overlapping and infrequent categories.

3.4.4 Weekdays

One of the assumptions we had beforehand was a higher demand during the weekends than on the weekdays. By looking at the dashboard in Figure 6, this assumption was reinforced. In the area chart Total passes by weekday, a clear trend of more visitors during the weekend compared to the weekdays is detected, peaking on Saturday. There appears to be an escalation starting at the lowest point, being Wednesday, gradually building up till Saturday, before it de-escalates from Saturday up till Wednesday again. This could be affected by the facility keeping closed during the mid-week in the early season.

3.4.5 Public holidays and vacation

Another interesting point to look at is whether public holidays and vacations affect the number of visitors, which we assume it would. The average sold passes in different vacations and public holidays shown in the visualizations from Figure 6 are categorized into numbers for simplicity in Table 4 and Table 5.

Table 4: Average number of passes sold daily during different vacations

Vacation	Regular days	Christmas vacation	Winter vacation	Easter vacation
Average passes	7.46	9.56	14.15	19.85

Table 5: Average number of passes sold on Public holidays

Holiday	Regular days	Christmas day	2nd Christmas day	New Year's Day	Palm Sunday	Maundy Thursday	Good Friday	1st day of Easter	2nd day of Easter	May 1st
Average passes	9.71	6.16	8.47	11.55	12.27	24,61	32,67	14,55	5.52	4.05

According to the average number of passes sold during the different vacations shown in Table 4, there are sold about seven passes on average on a regular day, which amounts to barely 15% of the total average of daily sale. All the days marked as vacation represent higher average sales than that of regular days, but the highest average daily sales are clearly on vacation number three, being the Easter vacation (39%). Easter vacation is one of the few vacations that include several consecutive public holidays. The fact that most people get time off due to the public holidays it includes, is a probable explanation as to why this vacation dominates in terms of increased sales. For many people, school vacations often consist of several regular workdays, meaning that not everybody has the luxury of getting time off during all of these vacations, especially adults with full-time jobs. The specific public holidays that the vacation often revolve around, however, could produce different results.

When looking into the average number of passes sold on the different public holidays shown in Table 5, Good Friday has the highest percentage of daily passes sold, followed by Maundy Thursday and 1st day of Easter. All of these days are public holidays related to Easter which is not surprising. The public holiday with the least influence on daily sale is May 1st. This date, however, only occur in two of the seasons, seeing that the season usually ends before this date. These visualizations are based on the entire dataset, without any filtering, which can affect the overall results. The variety of seasonal passes, for instance, only provides us with information about when they were bought, not when they were used, which can disturb the results and create some discrepancies.

3.4.6 Demand across seasons

By viewing the bar chart of total passes and total revenue by season in Figure 6, it does not show any significant changes in activity across the seasons, other than a reduction of about 14% in passes sold from season 4 to 5. Season 6 stands out with much lower activity compared to the other seasons, but there are missing a couple of months' worth of data here, including the Easter holiday which typically has a significant impact on the activity. It is therefore not possible to make an accurate conclusion of this particular season based on this visualization. The line in the bar chart represents total revenue by season, but we do not have full faith in its reliability. The numbers in the chart are, however, summed together by seasons, meaning the discrepancies on a day-to-day basis might not matter. With this taken into account, we notice that there has been some differing in relative changes in revenue compared to that of number of passes sold. Based on historical prices, there has been a gradual increase from season to season, so we would expect higher revenue even if the number of visitors stays the same.

The number of passes sold does not necessarily reflect the number of visits, seeing that several passes are seasonal and thus only registered when bought and not when used. This could indicate that the revenue was decreased due to possibly being more sale of the cheaper passes as the price had risen from last season, and visitors did perhaps downgrade their usual choice of pass to save some money. There is also added a slicer of seasons in the bottom right in Figure 6, to make it possible to check the statistics for each visualization regarding the specific seasons. By viewing each season individually, the trends were very consistent, and no considerable differences in the overview were found compared to the overall data.

3.4.7 Weather elements

The actual weather data we collected from the SeKlima service was central to the weather forecast data we produced. The distribution of the historical data is shown below, with temperature being depicted at the top, precipitation in the middle and snow depth at the bottom.

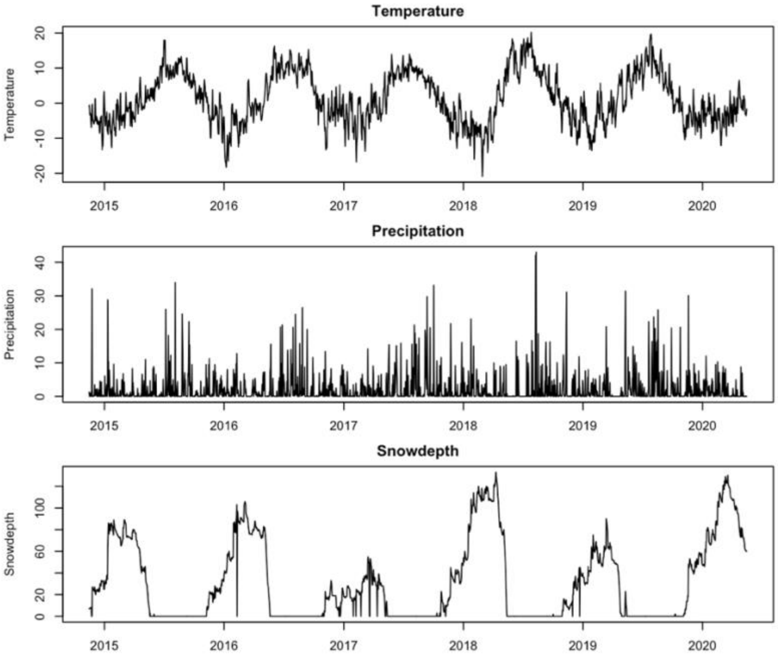


Figure 8: Weather elements distribution – Temperature, precipitation and snow depth

It is clear that all three weather variables have seasonal variations. This dependency is easiest to observe for temperature and snow depth, which move in opposite directions as the seasons pass. The temperatures are low during the winter, and high during the summer, while the snow depth is high during the winter and zero during the summer. The first observation of our weather variables is November 15th, 2014, meaning that the numbers on the x-axis represent the number of consecutive days after the first observation. The temperature normally varies between +15 °C and -15 °C, depending on the season. The distribution of snow depth shows that the third and fifth season (season 2016/2017 and 2018/2019) had less snow than the three other seasons. Snow depth is measured at the nearest weather measurement location to the skiing facility, which means that the snow at the facility's slopes may deviate due to the facilities ability to produce snow using snow cannons. Precipitation during the observed period also show sign of seasonality, but these do not always coincide with the pattern observed in temperature and snow depth.

3.4.8 Filtering the data

The inspection of the dataset showed numerous pass types and many customer groups. Including all of these in our analysis may yield poor results, as many of the pass types and customer groups are either overlapping or non-consistent over seasons. The pass types and customer groups which are non-consistent cannot be used for predictive purposes. We thus decided to further narrow the data used for model development to one pass type and one customer group. The obvious choice is the group that constitutes the biggest percentage of the total, which is daily passes and adult customers. Day passes have the added advantage of reflecting daily variations in demand, which seasonal passes or passes for multiple days fail to do. Our research is concerned with how weather affects the demand for alpine ski lift passes, and the weather is measured daily. Having a pass type that has the same aggregation level as the weather data is necessary to determine how daily variations in weather affects demand.

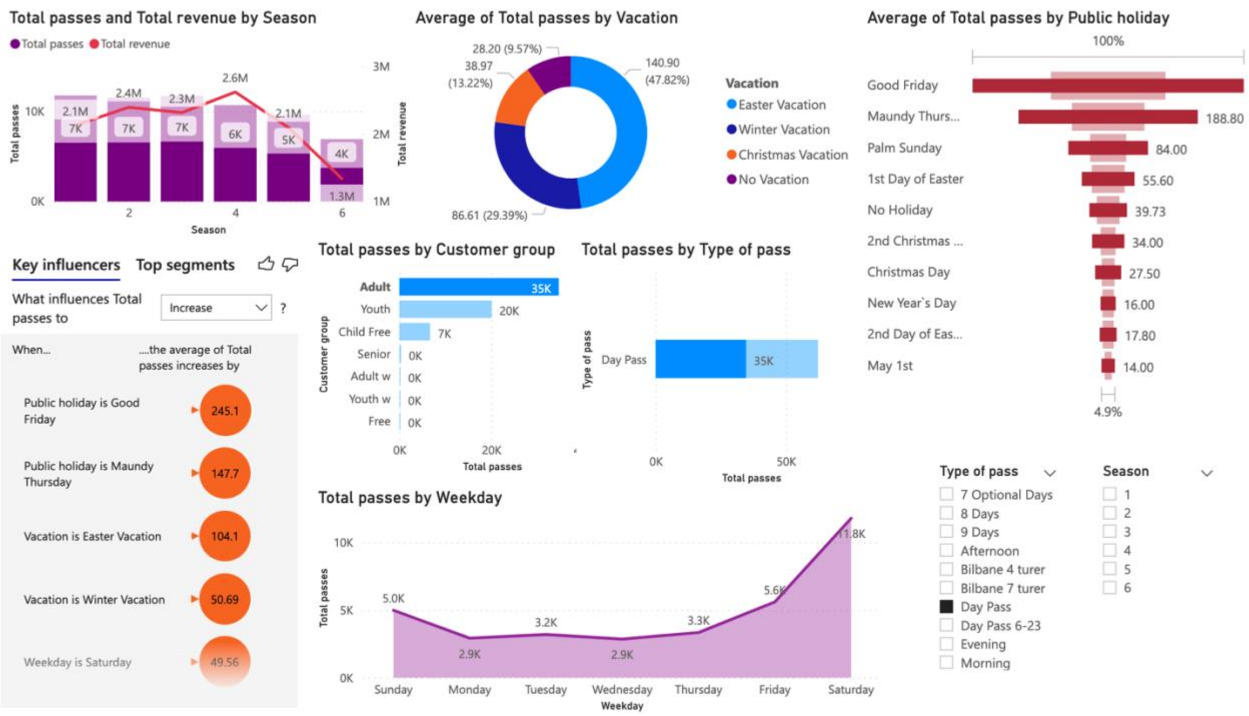


Figure 9: Dashboard of the sales data filtered on adult day passes

By narrowing the data down to adult day passes, the number of observations ends on 828 days in total. Seeing that the dataset includes observations over 923 distinct days in total, it means that this selection accounts for about 90% of all the included days. After filtering on the day pass with an emphasis on the customer group adult, it presents the same patterns as it did when including all of the data, as seen in Figure 9. This includes similar trends of activity across seasons, increased sales on different holidays and public holidays as well as the popularity of the different weekdays. This implies that if narrowing the dataset down to adult day passes, still provides a representative selection. In Figure 9, the key influencer indicates that when weekday is Saturday, there is an average of almost 50 more passes sold compared to all other weekdays, and this constitutes about 17% of all the average daily passes sold. The influence of holidays has also significantly shifted to an average of almost 73 more passes sold daily during Winter vacation and Easter vacation. Seeing that Winter vacation does not consist of any public holidays, this is somewhat surprising, as most adults probably do not have time off work and other obligations during this period. The adult day pass has been consistent through all of the seasons, and it is, therefore, possible that some misrepresentations in the dashboard from Figure 6, such as outliers and/or small sample sizes within the different pass types and customer groups has been filtered out. By narrowing down to the more frequent pass types, it could provide us with a better picture of the actual day-to-day activity with less noise and discrepancies. We do however emphasize that this is an overall discussion of apparent information concerning the data from only one alpine facility, and we draw no tangible conclusions from these dashboards alone.

3.4.9 Updated dataset

Most of the variables from the initial sales data were found to be abundant for the purpose of this research. Before moving forward with the analysis, the dataset was therefore updated by adding the new variables and merging the remainder of the sales data with the weather data, as shown in Table 6.

Table 6: Updated dataset

Variable	Description
Date	Date of sale.
Total passes	The number of passes sold.
Price	Price information. Differentiates in low- and high season.
The logarithm of total passes	To prevent any forecast of negative demand, the dependent variable of total passes was transformed into its logarithm.
Relative date	Date variable arranging each date in relation to January 1st within each season. Reflects the linear trend throughout each season.
Fourier term	Fourier terms were added to model seasonality.
High season	Dummy variable with the low season marked as 0 and high season as 1.
Closed	Dummy variable indicating whether the facility was open or closed to control for days with no sold passes.
Weekdays	Categorized in numbers from Sunday as 1 to Saturday as 7.
Christmas vacation	Dummy variable accounting for the days in the Christmas vacation.
Winter vacation	Dummy variable accounting for the days in the Winter vacation.
Easter vacation	Dummy variable accounting for the days in the Easter vacation.
Christmas day	Dummy variable accounting for Christmas day.
2nd Christmas day	Dummy variable accounting for 2nd Christmas day.
New Year's Day	Dummy variable accounting for New Year's Day.
Palm Sunday	Dummy variable accounting for Palm Sunday.

Maundy Thursday	Dummy variable accounting for Maundy Thursday.
Good Friday	Dummy variable accounting for Good Friday.
1st day of Easter	Dummy variable accounting for 1st day of Easter.
2nd day of Easter	Dummy variable accounting for 2nd day of Easter.
May 1st	Dummy variable accounting for May 1st.
Cold	Observations are marked as cold if the temperature is -10 °C.
Ice-cold	Observations are marked as cold if the temperature is -15 °C.
Rainfall	Observations marked with rainfall.
Temperature	Measured daily temperature in Celsius degrees.
Precipitation	Measured in millilitres.
Snow depth	Measured snow depth in centimetres.

4 Model presentation and analysis

In this section, the different demand forecast models and their results will be presented. The models will be presented and evaluated based on different criteria, such as adjusted R^2 and mean absolute error (MAE). There was developed several models, which varied in terms of explanatory power and forecast accuracy. The models presented are a thus a selection of the models that was further developed. The presentation consists of simplified regression results, but the full regressions outputs can be found in Appendix A to Appendix F. Also see Appendix G for the models' R-script.

The models presented are all multiple linear regression models. Linear regression models use the OLS approach to calculate the coefficient estimates that best fit the data, but it does not necessarily calculate the coefficient estimates that produces the lowest error (Qshick, 2019). The regression models were therefore also trained using ridge regression as an estimator, which adds some bias and shrinks the regression coefficients towards zero (James et al., 2013, p. 215). Six models are presented, and seasonal variables form the basis of all of them. Weather is added as a predictor in two of the models, enabling us to make comparisons between the models with only seasonal variables and the equivalent models with both seasonal and weather variables. Regression models are a commonly used forecasting technique, and historical sales data is well-suited for regression analysis (Chambers et al., 1971). Furthermore, linear regression models score high on interpretability, which is a great advantage for research problems resulting in models being used by people who are not trained in statistics (James et al., 2013, p. 25).

4.1 Creating the models

There were several steps in common for all the models developed. The first key step was to divide the dataset into two parts: a training set and a validation set. This is known as the validation set approach when the training set is used to fit the model, and the fitted model is used to predict the responses for the validation set (James et al., 2013, p. 176). The validation set is then used to estimate the test error rate, an indicator of how good the model performs. For many machine learning applications, in which the assumption of independence is upheld, the split is usually done randomly. However, given that our data is time series data, the observations are not independent. The split was therefore made non-random. The data was split according to seasons, using the first five seasons as a training set and the last season as a test set. In other words, the first five seasons were used to train the models, and the sixth and final season was used to test the model.

All models presented below use the *logarithm of passes sold*, also referred to as *demand*, as the dependent variable. Models with *passes sold* as the dependent variable were also trained, but led to many cases of negative predictions, especially for weekdays in the low season. The *logarithm of total passes* was therefore used to prevent negative point predictions, which leads to some subtleties regarding the interpretation of the coefficient estimates. The independent variables effect on *demand* cannot be interpreted directly. If the estimated coefficient of a variable is β it does not mean that a marginal increase in that variable will lead to an increase in *demand* equal to β. Rather, it means that a marginal change in the variable is associated with a $100\beta\%$ change in *demand* (Stock & Watson, 2019). The exception is the variable *price*, which is an endogenous variable. *Demand* affect *price* and *price* affects *demand*, meaning that we cannot interpret the effect of *price* in the same way as the other variables.

The coefficient estimates inform us whether the relationship between the independent and dependent variables is positive or negative. For the dummy variables, the coefficient estimates are informative for comparisons, as greater coefficient estimates translate to a greater effect on *demand*. If the *Christmas vacation* coefficient estimate is smaller than the one of *Easter vacation*, it implies that *Easter vacation* has a greater impact on *demand* than *Christmas vacation*.

The regression output indicates what variables are statistically significant at different significance levels. Determining the appropriate significance level is a bigger problem in forecasting than in other research settings. Statistical significance is, however, less important in forecasting models (Armstrong, 2007). In fact, some leading researchers in forecasting argue that p-values and statistical significance have been offered too much attention and that more emphasis should be given to the predictive ability of a model (Kostenko & Hyndman, n.d.). Including variables that are not statistically significant in a model poses no problem per se, as statistical tests were designed to test hypotheses, not to select variables (Hyndman, 2011). Two highly correlated variables can give good predictions but may get insignificant coefficient estimates because it is hard to distinguish their contribution to the model. Our presentation will therefore be less concerned with statistical significance per se, and more concerned with point estimates and their economic significance.

4.2 Evaluating model performance

The models need to be evaluated on some criteria. For in-sample fit, adjusted R^2 is a good evaluation metric. The evaluation of adjusted R^2 is entirely problem-specific and varies between different fields of science, but high numbers are generally desirable. Low adjusted R^2 does not automatically imply that a model is poor (Moksony, 1999). Low adjusted R^2 usually implies both a high mean squared error (MSE) and MAE for any fixed level of variance of Y, but some phenomena have high levels of irreducible error, which prohibits high adjusted R^2.

The goal of forecasting is to produce accurate predictions, in which evaluating a model based on how well it predicts out of sample is the preferred approach. Comparing the out-of-sample forecast errors – the one-step-ahead forecast errors – of different models gives information on the predictive ability of those models. Note, however, that there is a general tendency for out of sample forecast accuracies to be disappointing compared to within-sample fit (Chatfield, 2005).

The models presented in this thesis are evaluated on MAE, one of the most common accuracy measures for scale-dependent errors (Hyndman & Athanasopoulus, 2018). The ridge regression models have also been tuned to minimize MAE. In forecasting, an error can be described as the difference between an observed value and its forecast. An error does not necessarily translate to a mistake, but partly to the unpredictable part of the observation. There will, of course, be some level of mistakes present in the predictions, but there will also be some level of irreducible error (James et al., 2013, p. 18).

Deciding what type of loss function to use is best left to those it affects, which in this case is the alpine facility. They know best if overpredicting or underpredicting causes them the most pain, as they have in-depth knowledge of their finance and operations. The facility did not inform us of what error caused them the most pain, resulting in us using MAE as a loss function. MAE, also called absolute loss, is a symmetric loss function, meaning that loss is increasing at each side of the origin and that the loss increase at a constant rate with the size of the error (Diebold, 2017, p. 37). Under absolute loss, an error will be equally painful in both directions, meaning that under- or overpredicting by 1 visitor is just as bad. With absolute loss, the optimal point prediction is the conditional median of y (Diebold, 2017, p. 38).

4.3 Model presentation

4.3.1 Model 1 – Linear regression with seasonal variables – OLS

Table 7: Model 1 – OLS regression with seasonal variables

Model 1		
Adjusted R^2	0.639	
MAE	22.67	
Coefficients:	Estimate	
Intercept	22.54	**[1]
Relative date	0.004	***
Monday (2)	-1.303	***
Tuesday (3)	-1.051	***
Wednesday (4)	-1.123	***
Thursday (5)	-1.021	***
Friday (6)	-0.216	
Saturday (7)	1.176	***
Christmas day (1)	0.837	.
2nd Christmas day (2)	-0.076	
New Year's Day (3)	-0.850	.
Palm Sunday (4)	1.200	*
Maundy Thursday (5)	0.764	
Good Friday (6)	0.364	
1st day of Easter (7)	-1.111	*
2nd day of Easter (8)	-0.876	.
May 1st (9)	0.113	
Christmas vacation (1)	1.077	***
Winter vacation (2)	1.587	***
Easter vacation (3)	2.293	***
High season	4.303	***
Season	0.634	*
Closed	-0.686	***
Price	-0.066	**

The first model is a multiple linear regression model, using OLS to estimate the coefficients. Model 1 estimates *demand* as a function of *relative date*, *day of the week*, all *public holiday* and *vacation dummies*, *high season*, *season*, *closed* and *price*. The results of the regression analysis are shown in Table 7.

As is evident from the regression output, the independent variables are key contributors to *demand* for alpine ski lift passes. The explanatory power of the model, measured by adjusted R^2, is 0.639. This is a decent number and means that 64% of the variation in *demand* can be explained by the independent variables. The explanatory power could, of course, be higher, but some phenomena are characterized by a low signal to noise ratio, and for these phenomena, one cannot expect the adjusted R^2 to be close to 1. The signal-to-noise ratio is what proportion of the data is determined by the process of interest versus nuisance variation (Vandekerckhove et al., 2015).

Adjusted R^2 is a measure of in-sample fit and does not address how well the model performs on new data. This can be evaluated by MAE, which in this model is 22.67. This number is reported in the number of passes sold, meaning that the parameter has been transformed back to its original

[1] The asterisks indicate the significance codes for each variable: 0 = `***´, 0.001 = `**´, 0.01 = `*´, 0.05 = `.´, 0.1 = `´

form through exponentiation. An MAE of 22.67 implies that, on average, the forecast's distance from the true value will be 22.67. The MAE measure is calculated from the test set, meaning that the accuracy can deviate less or more for future observations, especially if the pandemic will have any significant impacts on the industry.

For comparison, consider the MAE of a naïve forecast. A naïve forecast is an estimating technique that uses actual values for previous periods to forecast a future period. Using the long-term historical average of passes sold as a prediction for the entire sixth season is thus a naïve forecast. The historical average of passes sold in the training set is 36.2. Using this as a prediction for all observations in the test set gives an MAE of 35.3. Model 1 has an MAE of 22.67, which is considerably lower than the MAE of the naïve forecast. This shows that the model has established some important causal relationships between the dependent and independent variables, which has led to better prediction accuracy compared to a naïve forecast.

Relative date has a positive coefficient, which implies that there is a positive linear trend regarding total passes over each season. *Relative date* does not model seasonality, only a linear trend. *Day of the week* is undoubtedly important, both statistically and economically. When *Sunday* serves as a reference point, *Monday* to *Friday* have lower *demand*, while *Saturday* has higher *demand*. This translates to increased *demand* during the weekend, which is of economic significance for the facility when they plan their operations.

All vacations have a positive impact on *demand*. This is not surprising, as people get time off work. *Easter vacation* has the greatest impact of the three, with a coefficient estimate that is twice as big as *Christmas*. This is supported by looking at the dataset and calculating the average number of passes sold during the different vacations. The average of all observations that are *not* part of any vacation is 22 passes sold. The same number for the *Christmas*-, *Winter*-, and *Easter vacations* are 39, 87 and 140 passes, respectively. The increase in passes sold is more than twice as big for *Easter vacation* as for *Christmas vacation*, but the average numbers are calculated from the entire data sample, not just the test sample. The calculation of average sales during the different vacations supports the model's claim that *Easter vacation* has the greatest positive effect on *demand*, and that the effect is of great economic significance.

Holidays have a mixed impact on *demand*, with some having positive coefficients, while others negative. Negative coefficient estimates for some holidays are not surprising. *Christmas day* and

2nd Christmas day are perhaps days most people spend at family dinners instead of on the slopes, and *New Year's Day* is perhaps spent on the couch. More surprising is the negative coefficient estimates of the *1st* and *2nd day of Easter*, as *Easter vacation* usually brings about greater activity for the alpine facilities. The *Easter vacation* dummy shows a great positive impact on *demand*. Negative coefficient estimates for the *1st* and *2nd day of Easter* is, therefore, an odd finding. As the coefficient size for *Easter (*2.293) is greater than that of both the *1st day of Easter (*-1.111) and the *2nd day of Easter (*-0.876), the net effects of these holidays are still positive.

The variables *price* and *closed* are both statistically significant, with a negative impact on *demand*, while the *high season* has a positive impact on *demand*. There are high levels of correlation (0.78) between *price* and *high season,* as the facility price differentiates between high- and low season. Multicollinearity is, however, not as important when forecasting (Elliott & Timmermann, 2016). *Closed* refers to days when the facility is closed and naturally shows that total passes sold decrease on these days. *Season* has a positive coefficient, indicating that *demand* increases each passing season.

To better understand how the coefficient estimates are to be interpreted, consider the following example. What happens when *Palm Sunday* dawns on us? Assume we are to compare a *Palm Sunday* to any other *Sunday* in the *high season*. Both days are set in the same high season, so the price is the same, meaning that neither the variables of *season*, *price* nor *high season* change. The model uses *Sunday* as a reference point, so we do not need to control for *day of the week* either. For simplicity, we ignore the changes in *relative date* as well, as we only are interested in the effect of *Palm Sunday* in isolation. There are two variables affected by *Palm Sunday,* the first being the *Palm Sunday* dummy, and the latter being the *Easter vacation* dummy. The coefficient estimates are 1.20 (*Palm Sunday)* and 2.29 (*Easter vacation).* Any changes in these variables will be from 0 to 1 as they are both dummy variables, and the change will lead to a 100%*1.20 + 100%*2.29 = 349% change in Y. If the day in question had not been *Palm Sunday*, but just a regular *Sunday* and the number of passes sold had been 35, the same number would have been 35*349% = 122 passes if it had been a *Palm Sunday*, ceteris paribus.

4.3.2 Model 2 – Linear regression with seasonal variables – Ridge

Table 8: Model 2 – Ridge regression with seasonal variables

Model 2	
MAE	21.15
Coefficients:	Estimate
Intercept	0.897
Relative date	0.004
Monday (2)	-1.136
Tuesday (3)	-0.884
Wednesday (4)	-0.965
Thursday (5)	-0.872
Friday (6)	-1.01
Saturday (7)	1.247
Christmas day (1)	0.819
2nd Christmas day (2)	-0.036
New Year's Day (3)	-0.761
Palm Sunday (4)	1.263
Maundy Thursday (5)	0.895
Good Friday (6)	0.544
1st day of Easter (7)	-0,778
2nd day of Easter (8)	-0.713
May 1st (9)	0.108
Christmas vacation (1)	1.009
Winter vacation (2)	1.522
Easter vacation (3)	2.093
High season	0.766
Season	-0.144
Closed	-0.725
Price	0.004

The second model has the same independent variables as in Model 1 but uses ridge regularization to estimate the coefficients. The only difference between the two models is the estimation method, so their point estimates can be compared directly. Ridge regression does not produce any t- or p-values, so no judgement about statistical significance can be drawn for Model 2.

MAE is 21.15, which is slightly lower than the MAE of Model 1. This could mean that estimating with ridge leads to more accurate predictions, or it could simply be a result of luck. It is apparent from the coefficient estimates that they have been shrunk towards zero compared to the corresponding estimates in Model 1.

Coefficient estimates for all *days of the week*, *closed*, and all *holiday* and *vacation* dummies are similar, both in size and in direction of the relationship. This indicates that these coefficients only have undergone low levels of shrinkage, and their effect on *demand* can be interpreted the same way as in Model 1. The bigger differences are found for *high season*, *price* and *season*. *High seaso*n has been shrunk manyfold, from 4.3 in Model 1 to 0.77 in Model 2, while *price* and *season* have changed coefficient signs. The high correlation between the two may explain why their coefficient estimates are so variable between the models, as they can have a hard time distinguishing the effect of *season* from that of *price.*

Highly variable point estimates with possibly changing signs are typical for highly correlated variables as ridge regression is doing its job of fighting the effects of multicollinearity. The point estimates of such correlated variables are often more sensible when estimated by ridge than by

OLS. In this particular case, the point estimates appear more sensible in the first model, as ridge regression has given *price* a positive coefficient estimate, which contradicts the negative relationship between *price* and *demand* which is thoroughly documented in the literature. The coefficient size is, however, small, indicating that a price increase only leads to a minimal, nearly non-existent, increase in *demand*.

Following up on the example from Model 1 with *Palm Sunday*, its effect on *demand* has decreased. The coefficient estimates of *Palm Sunday* and *Easter vacation* are 1.26 and 2.09, respectively, leading to a combined change in Y of 100%*1.26 + 100%*2.09 = 335%. This translates to an increase in the number of sold passes from 35 to 117, which is smaller than the increase of Model 1. The difference in predicted passes sold is only 5, which is quite small. This is because the coefficient sizes for the relevant variables only have been shrunk a little. For variables with bigger changes in coefficient estimates, such as *high season*, the effect on *demand* would be considerably smaller in Model 2 than in Model 1. The reason is that OLS estimation does not consider which independent variables are more important, leading to unbiased coefficients that produce the lowest Residual Sum of Squares (Qshick, 2019). Ridge regression, on the other hand, accepts that some variables are more important, and thus treats each predictor differently. Therefore, some variables are more penalized than others, resulting in different coefficient sizes than with OLS. The small size reduction of *Palm Sunday* and *Easter vacation* indicates that these variables are important predictors for *demand*.

4.3.3 Model 3 – Linear regression with Fourier terms for seasonality - OLS

Model 3 is another linear regression model, but the linear trend over one season represented by *relative date* has been replaced with *Fourier* terms to model seasonality. It has been replaced instead of kept because of the exact multicollinearity between *relative date* and *Fourier* that we would otherwise get. Weather variables are not included in this model.

The regression output is quite long, as the *Fourier* series is comprised of 16 pairs of *Fourier* terms, totalling 32 *Fourier* variables. The output presented excludes the first 15 pairs of *Fourier* variables, because the output would be too long if included. In this model, 16 pairs of *Fourier* terms are optimal for MAE. Reducing or increasing the number of pairs have a positive impact on the in-sample fit, but simultaneously increase the out-of-sample MAE.

Table 9: Model 3 – OLS regression with Fourier terms for seasonality

Model 3		
Adjusted R^2	0.697	
MAE	21.86	
Coefficients:	Estimate	
Intercept	978000000	
Fourier16sin	220	
Fourier16cos	-1950	
Monday (2)	-1.310	***
Tuesday (3)	-1.090	***
Wednesday (4)	-1.160	***
Thursday (5)	-1.040	***
Friday (6)	-0.221	.
Saturday (7)	1.200	***
Christmas day (1)	0.216	
2nd Christmas day (2)	-0.068	
New Year's Day (3)	-1.260	**
Palm Sunday (4)	1.380	**
Maundy Thursday (5)	0.858	.
Good Friday (6)	0.464	
1st day of Easter (7)	-1.540	**
2nd day of Easter (8)	-1.270	*
May 1st (9)	0.438	
Christmas vacation (1)	0.586	*
Winter vacation (2)	0.115	
Easter vacation (3)	2.550	***
High season	2.350	.
Season	0.434	.
Closed	-0.666	***
Price	-0.046	*

Replacing *relative date* with *Fourier* has led to better in-sample fit, as adjusted R^2 has increased from 63.9 to 69.7. Given that the model is comprised of 16 pairs of *Fourier* terms, this is perhaps not surprising. *Relative date* has replaced 32 *Fourier* variables, making the model more flexible. MAE is 21.86, and has decreased compared to Model 1, but increased compared to Model 2. This difference is, once again, small.

The coefficient sizes have changed compared to those of Model 1. Table 9 excludes the first 15 pairs of *Fourier* terms, but the coefficient estimates for *Fourier* terms are big compared to the remaining variables. Most variables have coefficient estimates between -2 and 2, but the *Fourier* variables have estimates of a much grander scale, with *Fourier16cos* having one of -1950. Modelling with *Fourier* has a great impact on the coefficient estimates of the model. There are no changes to the direction of the relationship between dependent and independent variables. We see the same effects linked to *day of the week* as in the previous models, with lower *demand* on weekdays compared to the weekend.

In a model with *Fourier* terms, the effect of *Palm Sunday* is even greater than that of both previous models. With coefficient estimates of 1.38 and 2.550, the combined effect of *Palm Sunday* on *demand* is 393%, resulting in 137 sold passes. This is a higher prediction than both previous models

4.3.4 Model 4 – Linear regression with Fourier terms for seasonality – Ridge

Model 4 has the same independent variables as Model 3 but uses ridge regularization instead of OLS to estimate the coefficients.

The model’s MAE is 20.26, which is lower than all the previously presented models. This could be due to better predictions, or simply because of luck. The difference is, however, small.

Table 10: Model 4 – Ridge regression with Fourier terms for seasonality

Model 4	
MAE	20.26
Coefficients:	Estimate
Intercept	2.356
Fourier16sin	-0.106
Fourier16cos	0.112
Monday (2)	-1.140
Tuesday (3)	-0.923
Wednesday (4)	-0.990
Thursday (5)	-0.892
Friday (6)	-0.121
Saturday (7)	1.254
Christmas day (1)	0.340
2nd Christmas day (2)	-0.085
New Year’s Day (3)	-0.822
Palm Sunday (4)	1.256
Maundy Thursday (5)	0.944
Good Friday (6)	0.621
1st day of Easter (7)	-0.974
2nd day of Easter (8)	-0.865
May 1st (9)	0.816
Christmas vacation (1)	0.559
Winter vacation (2)	0.320
Easter vacation (3)	2.134
High season	0.371
Season	-0.088
Closed	-0.678
Price	0.000

The coefficient size of the *Fourier* terms has been greatly reduced in absolute terms, compared to those of Model 3. *Fourier16cos* have been reduced in size from 1950 to 0.112, by applying ridge regularization. Ridge regression has produced coefficient estimates in the same range as those found in Model 1 and Model 2, thus greatly reducing the big coefficient estimates introduced by modelling seasonality with *Fourier*. The remaining coefficient estimates are similar to the estimates of Model 3, with small differences in individual point estimates. There are no differences in direction of the relationships either, except for the variables *price* and *season*. The coefficient estimates can therefore be interpreted in the same way, with increased *demand* during weekends and vacations.

It is interesting to compare the different coefficient estimates to each other. Looking at the different *vacation* dummies, *Easter vacation* is the one with the biggest coefficient estimate, and therefore with the greatest impact on *demand*. Its coefficient estimate is almost four times as big as that of *Christmas vacation* and six and a half times as big as *Winter vacation*. This suggests that *Easter vacation* has a four times bigger effect on *demand* than *Christmas vacation*, and six and a half times bigger than *Winter vacation,* all other things being equal. Note that this does not control for the effects of any individual holiday within each vacation. Some holidays, such *as the 2nd Christmas Day* and *1st day of Easter* have negative coefficients, even though the vacation they belong to has a positive effect on *demand*.

4.3.5 Model 5 – Linear regression with seasonal variables and weather variables - OLS

Table 11: Model 5 – OLS regression with seasonal variables and weather variables

Model 5		
Adjusted R^2	0.699	
MAE	21.72	
Coefficients:	Estimate	
Intercept	1490000000	
Fourier16sin	604	
Fourier16cos	-2040	
Monday (2)	-1.300	***
Tuesday (3)	-1.090	***
Wednesday (4)	-1.150	***
Thursday (5)	-1.060	***
Friday (6)	-0.229	.
Saturday (7)	1.190	***
Christmas day (1)	0.232	
2nd Christmas day (2)	-0.023	
New Year's Day (3)	-1.290	**
Palm Sunday (4)	1.350	**
Maundy Thursday (5)	0.874	.
Good Friday (6)	0.481	
1st day of Easter (7)	-1.520	**
2nd day of Easter (8)	-1.290	**
May 1st (9)	0.416	
Christmas vacation (1)	0.587	*
Winter vacation (2)	0.186	
Easter vacation (3)	2.580	***
High season	2.730	***
Season	0.517	.
Closed	-0.683	***
Price	-0.054	*
Temperature	0.023	**
Precipitation	-0.009	
Snow depth	0.001	

The next model builds on the models with *Fourier* terms for seasonality but adds weather variables as well. The reason for developing the Fourier models further instead of the *relative date* models is not based solely on predictive performance. There were only small differences between their performance, with *Fourier* terms performing slightly better. This could be due to superiority or simply due to luck. We know that both variables imply a linear trend over each season, as there is exact collinearity between the two, but *Fourier* offers rich patterns through seasonal variation as well. The reason for adding weather variables to the *Fourier* models is the seasonal variation it offers, which *relative date* fails to provide.

It is natural to compare Model 5 with Model 3, as the only difference between the two is the weather variables. The adjusted R^2 of Model 5 is almost equal to that of Model 3, with a 0.002 increase. MAE, on the other hand, has decreased from 21.86 in Model 3 to 21.72 in Model 5. The minimal changes in adjusted R^2 and MAE are remarkably low considering that three extra variables have been added, especially considering that these are documented in the literature as important predictors for *demand*.

Out of the added weather variables, only *temperature* has a statistically significant effect on *demand*. The relationship is positive, indicating that increased temperatures lead to increased *demand*. The coefficient of *temperature* is 0.023, meaning that an increase of *temperature* by 1 °C will, on average, lead to a 2.3% increase in the number of sold passes. Both *precipitation* and *snow depth* are without any statistical significance. *Snow depth* has a positive effect, meaning that increased levels of snow lead to increased *demand*. *Precipitation*, on the other hand, has a negative effect, meaning *demand* decrease as *precipitation* increase. Statistical insignificance by itself presents no big problems for predictive purposes.

4.3.6 Model 6 – Linear regression with seasonal variables and weather variables – Ridge

Model 6 consists of the same variables as Model 5, but the coefficient sizes are estimated with ridge regularization instead of OLS. The change in estimation approach has thus led to a reduction in MAE from 21.72 to 20.53. An MAE of 20.53 is the second lowest of all presented models. It is, however, not lower than that of Model 4, which has the same variables as Model 6, except for the weather variables. When applying ridge regularization, it thus appears that adding weather variables makes for poorer predictions and that *demand* is best predicted by seasonal variables alone. This contradicts the findings with OLS estimation, in which the weather variables lead to a slight increase in predictive accuracy. However, the differences in MAE are small, and could very well be the result of luck.

There are some interesting changes in the coefficient estimates compared to those of Model 5. Applying ridge regression has, once again, had a great impact on the size of the *Fourier* variables. These have been shrunk manyfold compared to those of Model 5. The remaining variables have changed less, but many have smaller coefficient sizes.

Table 12: Model 6 – Ridge regression with seasonal variables and weather variables

Model 6	
MAE	20.53
Coefficients:	Estimate
Intercept	2.189
Fourier16sin	-0.095
Fourier16cos	0.105
Monday (2)	-1.138
Tuesday (3)	-0.918
Wednesday (4)	-0.984
Thursday (5)	-0.898
Friday (6)	-0.126
Saturday (7)	1.252
Christmas day (1)	0.373
2nd Christmas day (2)	-0.070
New Year's Day (3)	-0.843
Palm Sunday (4)	1.225
Maundy Thursday (5)	0.975
Good Friday (6)	0.628
1st day of Easter (7)	-0.940
2nd day of Easter (8)	-0.857
May 1st (9)	0.796
Christmas vacation (1)	0.588
Winter vacation (2)	0.342
Easter vacation (3)	2.136
High season	0.405
Season	-0.090
Closed	-0.696
Price	0.001
Temperature	0.018
Precipitation	-0.013
Snow depth	0.003

The most interesting coefficient estimates of Model 6 are those of the weather variables. While most of the other variables have shrunken coefficient estimates, both *precipitation* and *snow depth* have bigger coefficient sizes as a result of ridge regularization. Their relationships maintain the same direction, but both variables' coefficient estimates have increased in size. *Precipitation* has increased in terms of absolute numbers, from -0.009 to -0.013, while *snow depth* has increased from 0.001 to 0.003. The increase is small in terms of numbers, but as the coefficient sizes were small to begin with, the change is big in relative terms. This suggests that ridge regression attributes greater importance to *snow depth* and *precipitation* than OLS regression does. *Temperature*, on the other hand, has a decreased coefficient estimate. An increase in expected *temperature* of 1 °C will now lead to a 1.8% increase in *demand*, compared to a 2.3% increase from Model 5.

4.4 Model comparison

Table 13 shows how well the different models perform in terms of MAE. There are only small differences between the models, indicating that they perform almost equally well on new data. Some of our unreported models, on the other hand, performed considerably worse, with adjusted R^2 between 0.3 and 0.4, having MAE over 38. Of the reported models, the best performing model has an MAE of 20.26, while the worst performing model has an MAE of 22.67.

Table 13: MAE comparison

	MAE
Model 1	22.67
Model 2	21.15
Model 3	21.86
Model 4	20.26
Model 5	21.72
Model 6	20.53

Model 1, Model 3, and Model 5 are linear regression models with OLS estimator, while Model 2, Model 4, and Model 6 are linear regression models with ridge regularization. The ridge regression models perform better than their OLS counterparts, for all three combinations of independent variables. The results also indicate that the models including *relative date* instead of *Fourier* performed worse, having higher MAE for both estimation methods compared to the models including *Fourier* terms. The effect of adding weather remains unclear, as the models with weather variables included gave a somewhat better forecast when using OLS, but worse when using ridge regularization, compared to the models without weather variables.

The apparent approach when selecting the best forecasting model is to choose the model with the smallest error measurement (Zaiontz, n.d.), which in this case would be Model 4. It is, however, unclear if this model performs better because of it being superior, or if it is due to luck. We must estimate the likelihood of the outcome being a result of chance or superiority. This can be achieved by applying the Diebold-Mariano test, which tests the null hypothesis (H_0), claiming there to be no difference in expected predictive loss from two forecasts (Diebold & Mariano, 1995). If the hypothesis is rejected, we can conclude that the two models do not have an equal expected predictive loss in the population. Failing to reject H_0, on the other hand, does not mean that the two models have an equal expected predictive loss, but that there is simply not enough evidence in the data to claim otherwise (Diebold & Mariano, 1995).

Valid inference regarding the predictive performance of two models requires the models of interest to be selected before their performance is observed (Hansen, 2010). Simply choosing to compare the best and worst-performing model will generally lead to invalid inference regarding statistical significance. The inference from such a procedure will be too liberal. Thus, if we reject H_0, this may be due to either the test having a larger than nominal size or the effect being real. However, if we fail to reject H_0 even in the presence of a positive size distortion, we have stronger support of H_0, than we would under correct size. The two models should therefore be chosen before their performance is observed, not because of it. To determine if adding weather variables to a model leads to better predictions, one model with weather variables and one without will be compared using the Diebold-Mariano test with a significance level of .05. If the p-value is lower than the significance level, H_0 is rejected, and we conclude that the two models have an unequal expected predictive loss in the population.

Both Model 4 and Model 6 use ridge regularization, performing better than their OLS counterparts. The only difference between Model 4 and Model 6 is the added weather variables in Model 6. The difference in MAE between the two is small, in slight favour of Model 4. They are not chosen because of their observed performance, but rather because they are counterparts, allowing us to examine the importance of weather variables. They will therefore be compared using the Diebold-Mariano test to estimate the likelihood of the outcome being a result of chance or superiority.

The p-value obtained from the Diebold-Mariano test is 0.627. Seeing that the p-value is greater than the significance level, it indicates that there is not enough evidence in the data to reject H_0, claiming an equal expected predictive loss. There could very well be a difference in the performance of the two models, but there is simply not enough evidence in the data to make such a claim. The difference in the predictive performance of the two models is statistically insignificant. The principle of parsimony states that simpler models should be preferred over complex ones, all other things being equal (Vandekerckhove et al., 2015). Under the principle of parsimony, Model 4 is preferable over Model 6, given that the added complexity of Model 6 in terms of additional variables does not produce more accurate predictions.

4.4.1 Model analysis

Figure 10 display the in-sample fit on Model 4. The black line presents the actual number of sold passes in the training set, while the blue line illustrates the forecasted number of sold passes. They coincide a lot, suggesting that the in-sample fit is quite good. The red line shows the forecast errors and portrays the difference between actual values and forecasted values. The forecast errors are evenly distributed around zero, suggesting that the forecasts miss the actual values in both directions. The largest errors are found at the peaks and valleys of the red line.

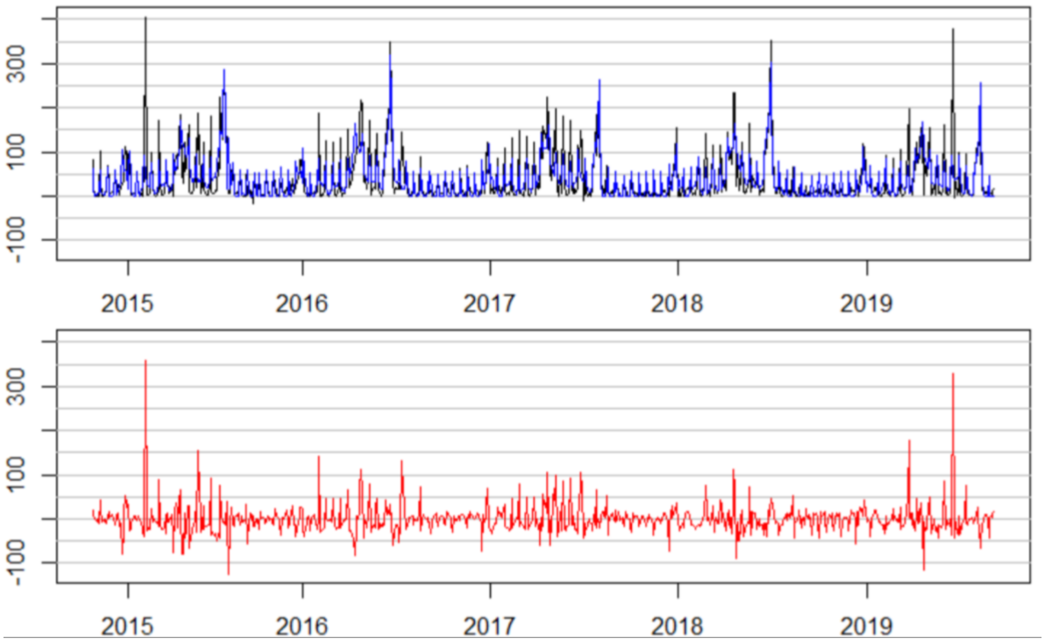


Figure 10: In-sample fit of Model 4

More interesting than the in-sample fit is the out-of-sample forecast errors, showing how well the model predicts new data. Figure 11 can be interpreted in the same way as Figure 10 but the predictions and forecast errors are only shown for the test set, being season six. The blue line follows the black line closely, but there are still forecast errors to be found.

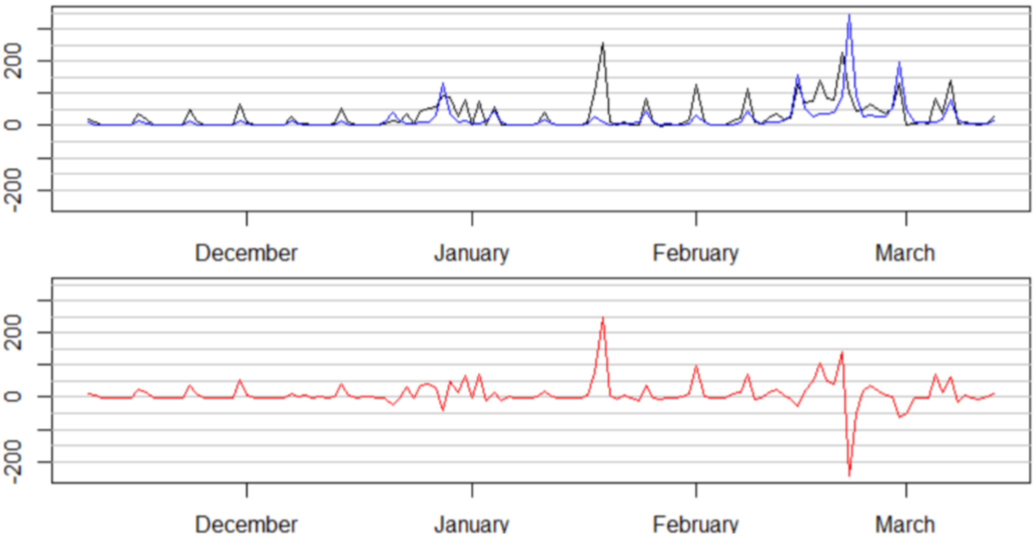


Figure 11: Out-of-sample fit of Model 4

To see if there were any patterns to the larger forecast errors, the observations with prediction errors above 45 in absolute numbers were inspected further. This is an arbitrary number but is chosen because it is roughly twice as large as the MAE of the presented models. We thus categorize errors twice the size MAE as being large. Out of the 126 observations in the test set, only 16 had forecasts errors greater than 45. Three of these stood out, with forecast errors of 246, -299 and 126. This suggests that the model predicts well on most days, but that it also makes a few considerably big errors.

The 16 observations with the largest errors were unevenly distributed throughout the days of the week, which suggests that the model struggles to predict *demand* accurately on certain days of the week. Seven of the observations fell on *Saturday*, and three on *Sunday*, telling us that there are made greater errors during the weekends compared to the weekdays. This is not surprising, as weekends typically lead to big increases in *demand*, and bigger errors are to be expected when *demand* is peaking. This is natural and not necessarily a flaw of the model. Predicting 5 when 10 is correct gives the same absolute error as predicting 250 when 255 is correct, even though the latter has a substantially lower relative error than the first.

Table 14: Large errors distributed by day of the week

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	2	0	2	1	7	3

The data is only comprised of observation up to March 13th, 2020. The span of observations in the test sample makes it hard to make any judgements on how well the model predicts the number of passes sold for the different months of the year. March is cut short in the test sample, and there are no test data for April and May. Furthermore, November is also short on observations, as the season starts mid-November. The 16 most erroneous predictions are therefore distributed unevenly among months, with peaks in January, February, and March. That does not necessarily mean that the model is flawed in any direction. It could be the result of an uneven number of observations for the different months, although the peak in February may be linked to *Winter vacation*.

Table 15: Large errors distributed by month

November	December	January	February	March	April	May
1	1	3	8	3	0	0

In the test sample, there are 13 days categorized under *Christmas vacation* and 14 under *Winter vacation*. The prediction errors are large on one of the *Christmas vacation* days and seven of the 14 *Winter vacation* days. The large errors during the *Winter vacation* were evenly distributed throughout the two weeks, suggesting no pattern concerning which part of the country has the vacation. The model thus appears to make greater errors on some of the days that are out of the ordinary, as half of the large errors occur on days that are part of a vacation. The increased error on these days is expected, as they bring about peaks in *demand*. Note that the test sample excludes the 2020 *Easter vacation*. *Easter vacation* is the most important holiday for the industry, with peaks in *demand* as people enjoy time off work. The coefficient estimates for *Easter vacation* were the greatest of all *vacation* dummies, indicating a significant economic effect on the number of visitors. The models´ performance on *Easter vacation* is unknown. None of the observations with large errors occurs on public holidays.

The test set does not contain a full *sixth season*, seeing that observations after March 13th were not included in the sales data. This could influence the predictive performance of the models. If *Easter vacation* typically leads to high peaks in *demand*, and the model is good at capturing the effect of *Easter vacation*, not having the vacation in the test set may lead to poorer accuracy than if it was included. Given the evidence that the model performs badly on some vacations, the predictive performance of the model may have been worse if the test sample contained a full season. The argument only holds true if the 2019/2020 season was to be considered a normal season, which it cannot. The lock-down enforced by the government in March led to closed doors for the remainder of the season, making it a highly irregular end of the season. Not having data after March 13th is perhaps good, as the forced shut-down would have led to abnormal *demand* for the remainder of the season, which the models would have no way to foresee based on historical data.

Weather variables were not a part of Model 4 but inspecting the weather on the observations with the largest forecast error showed some interesting patterns. There were high levels of *precipitation* (above 5mm) in the simulated weather forecast on five of the 16 days, suggesting that *precipitation* could be a potential source of larger out-of-sample forecast errors. *Snow depth* showed no pattern, and none of the days was categorized as *cold (temperature* below -10 °C) or *ice-cold (temperature* below -15°C), suggesting that cold temperatures are not a big source of forecast error. It could be coincidental that so many of the observations with large forecast errors had high levels of *precipitation,* as the models including weather variables did not produce any more statistically significant predictions.

The importance of weather as a predictor could be affected by the size of the test sample, but it is hard to determine the direction of impact. If the weather during the period missing from the test set is of great importance for *demand,* it could make weather less important in the models compared to in reality. Contrary, if the weather during the period missing is of little importance, it may not influence the importance of weather as a predictor at all. The statistically insignificant weather variables may partly be the result of a short test sample but could also reflect reality well.

The analysis suggests that Model 4 makes the biggest out-of-sample forecast errors on *Saturday, Sunday, Winter vacation* and for high levels of *precipitation*. Comparing the forecasted *demand* to the actual number of passes sold show that the model makes errors in both directions, both overpredicting and underpredicting the number of sold passes. The errors show a clear pattern with sold passes, meaning that it makes larger errors on days with many visitors compared to days with few visitors. Even though 65% of the observations in the data set has between 0 and 25 sold passes, only one of the 16 days with large errors fall in this category. Low-activity days are thus underrepresented in this case, supporting the claim that days with high *demand* brings about the greatest errors.

Three days stood out with exceptionally large forecast errors. Two of them were on a *Saturday* and a *Sunday* during the *Winter vacation,* which may explain why the model had a hard time predicting the number of passes sold. Two of the conditions associated with large errors were present for these observations, being both part of a weekend and part of a vacation. Furthermore, the two days in question are found in the weekend separating the two different weeks of the *Winter vacation*, presenting an overlap for all parts of the country. The third observation was an ordinary *Sunday* in January, in no association with neither a vacation nor any form of extreme weather. The model predicted just 11 sold passes, while the actual number was 258, making the forecast error 247. The remaining *Sundays* in *January* in the test set had three, seven, and 10 passes sold, so a prediction of 11 does not seem too extreme. The large forecast error for this day is more likely the result of a day with abnormal *demand* caused by a variable our model does not control for, such as activities at the facility or a big marketing campaign.

5 Discussion

5.1 Weather variables

The Diebold-Mariano test showed no statistically significant prediction accuracy improvement for the ridge regression model with weather variables compared to the model without weather variables. This does not automatically mean that weather variables are poor predictors of demand. There are several possible explanations as to why H_0 cannot be rejected.

The first reason is that there simply is not enough evidence in the data to reject H_0. The weather could, very well, be an important predictor of demand, but our data may not be sufficient to prove it. This is a likely explanation, given that the models are based on simulated forecast data, not actual forecasts developed by the Norwegian Meteorological Institute (MET). The simulation may give rise to many problems, which can conceal the true effect of weather as a predictor. Any error in simulating forecast data can lead to biased data and unreliable results. The biggest source of error in the forecast simulation is that it is based on the mean of accuracy across season and weather type, thus neglecting to account for the within-season fluctuations in accuracy. Our forecast data does not reflect that some weather types are easier to accurately predict, nor that accuracy varies between different parts of the country and months of the year. We used the mean of accuracy across the season, which may have produced simulations with high levels of error.

Another possible reason is that there may be issues with the types of weather variables included in the models. We used temperature, precipitation, and snow depth as predictors, but there could be other important variables, with some suggestions in the literature for important predictors being wind and cloudiness (Falk, 2013; Shih et al., 2009). Our selection of weather variables was limited but seeing that the variables are well-documented in the literature as significant predictors, we felt confident about these variables none the less. If other factors such as cloudiness and wind chill are important for demand, our models could possibly have performed better if they were included. However, considering the performance of the weather variables included, we do not believe that adding more weather variables would have greatly influenced our results.

Additionally, the models with weather variables use them as continuous numeric variables and try to establish a linear relationship between the variables and demand. It can be hypothesized that weather has a non-linear effect on demand. If weather within the normal range has little effect on demand, but extreme weather has a big effect, the relationship is perhaps not best described as a linear one. It can be easy to understand that skiing is less attractive if the temperatures are blazing cold, or if there are high levels of precipitation. Allowing for non-linear relationships may thus improve the importance of weather as a predictor. This has, however, been explored without much success. Models using nonlinear functions of the weather variables were also developed for this particular case, but they did not yield any superior forecasts and have thus not been reported. This is contrary to the findings of Malasevska et al (2017), who used data from facilities in the same region as our facility and found a non-linear relationship between wind chill temperature and the number of visitors. If the temperature were below -9.5°C, a temperature increase had a positive effect on the demand, while if the temperature were above -9.5°C, a higher temperature led to lower demand. Given their results, the cold and ice-cold dummies should have a negative impact on demand. We did indeed find a negative effect of cold, although not for ice-cold, but neither of the variables were statistically significant. Introducing them to a model led to lower accuracy, suggesting that they did not improve the predictive performance of the model. The importance of weather for predictive purposes can vary across locations, especially if weather only impacts demand at extreme conditions. Locations with high variations in temperature, snow depth and precipitation may thus find predictions based on weather forecasts to be more accurate compared to locations with low variations.

Yet another possible reason is that weather is of little importance as a predictor. This is in support of some earlier findings, in which weather variables are found to be statistically significant, although having a small total effect on demand (Falk, 2015; Shih et al., 2009). If snow depth indeed does have little impact on the demand at the facility, it could possibly be verified by them having sufficient resources to produce their own snow. Generally, the industry is dependent on sufficient levels of snow to keep the slopes open, but larger facilities have equipment available to produce their own snow when the conditions allow it. It can therefore be hypothesized that snow depth is more important for smaller facilities that have not invested in resources for snow production. The production of snow is, however, an intricate matter, requiring specific conditions to gain the desired snow quality. Even with the right conditions in place, the common person would not be able to recognize it, possibly inducing a psychological aspect with visitors not necessarily being aware of the possibilities of snow production or associate it with poor snow quality.

If the larger facilities were to experience lower demand along with decreased natural snow depth, other factors contributing to the melting of snow could be at fault. Snow melts not only by higher temperatures but also by rain. The temperature does not need to increase much before the precipitation comes down as rain rather than snow. When the snow melts due to a significant downpour, lower demand is expected under this circumstance alone, but it could possibly lead to only the variable of snow depth being captured as a significant contributor due to collinearity. Precipitation, such as rain, does have a more unreliable pattern as the downfall can vary significantly in a matter of hours, while snow depth typically will change more gradually. Precipitation could therefore induce a larger standard error which then again makes it less likely to become statistically significant when having collinearity to a more stable variable such as snow depth.

Weather can also cause spatial substitution, meaning the visitors substitute the activity either by choosing a different ski facility or substitute to another activity entirely (Malasevska and Haugom, 2019). An increase in temperature could very likely have a positive effect on other recreational activities, such as fishing, hiking or going to amusement parks. This could have direct effects on demand, but since they are not represented in the data, it could cause complementary variables such as snow depth and temperature to falsely present themselves as the founding determinant for the changes in the dependent variable. According to Tuv (2019), there is, however, no evidence of an increase in other activities connected to the decreased skiing activity.

5.2 Seasonal variables and price

Our visualizations and analysis of the data point to some significant patterns concerning seasonal variables, such as weekdays and holidays in particular. The facility discloses that when they choose to close the entire facility for whole days, it is mainly due to seasonal variables. This provides an argument for seasonal variables being of higher importance when it comes to predicting demand, compared to the weather, at least when using our particular models and data.

Of the three vacations, Easter vacation is by far being presented as the most important predictor both from empirical research and in our models. In comparison to Easter vacation, both the Christmas- and the Winter vacation is of less economic significance, with Christmas vacation being a touch more important than Winter vacation. Christmas vacation does not include many subsequent public holidays, and with the few that does, usually follows other family traditions.

There are, however, many who choose to take time off in the entire period between Christmas Eve and New Year's Eve. During Winter vacation, it is mostly school pupils that have the privilege of a vacation as there are no public holidays attached to it. On Easter vacation, on the other hand, most people have time off for five consecutive holidays, which makes it more desirable to travel. The skiing industry chooses to end the high season just past Easter vacation, depending on how late into spring it falls. This indicates how important this particular holiday is for the demand, as the facility are willing to extend the season as far as to the beginning of May, well into spring. Despite temperatures increasing in the spring, the revenue gathered from the Easter vacation has such a massive influence on the total result, that other possible challenges are somewhat overlooked in terms of normal practices. This includes investing more resources in snow production than otherwise for instance. When looking at the effect of the different public holidays, there are some peculiarities. This can perhaps be explained by the presence of the highly correlated vacation dummies, causing the model to fail in distinguishing the effect of vacation from the public holidays, resulting in some odd coefficient estimates.

Besides the public holidays and the different vacations, there has been presented a clear pattern in terms of day of the week, which was as expected. Already from the first visualizations, this was indicated, but the models verified it further with Saturday being the biggest positive contributor of demand, followed by Sunday and Friday. The decision to keep the facility closed during mid-week in the early season makes sense when looking at these results, as there is a lower demand on the weekdays compared with the weekends. There is also less demand on Friday compared to that of Sunday, but this is of minimal effect in contrast to the other weekdays. On any regular weekends, the visitors usually use Friday to travel, meaning a late arrival and a desire to settle in after travelling likely causes the demand to be a bit lower than on Saturday and Sunday. The importance of the seasonal variables day of the week as well as vacation furthermore coincide with the findings of Shih et al. (2009).

The price of the adult day passes is only differentiated between high- and low season, not leaving too much of a foundation for analysis. There has furthermore been minimal variation in price from season to season, leading to a small impact on demand. This is one of the main problems with using historical data to predict the future, as more variation in price could have generated very different results, thus providing a different perspective on the effect it has on demand. If there were to be implemented price changes in the lines of dynamic pricing or so, the data would be of poor basis to assess the future behaviour to these kinds of changes. Future studies of price-response functions could model the effect of a wider range of price levels.

5.3 The Covid-19 pandemic

When the Covid-19 pandemic arrived, basically every industry was affected, being it good or bad. The alpine industry depended on governmental support packages, as they had to limit their services significantly due to the restrictions set by the authorities. Seeing that the horizon of the data we have at hand ends in March 2020, right at the beginning of the pandemic, it impairs our ability to make assessments concerning its effect on the future demand based on the available raw data. The pandemic could however impose a great upcoming impact on the industry. Despite the domestic demand likely increasing from the closed borders and thus lessen the competition from facilities abroad, it would not nearly be possible to capitalize it. In the short run, the consequences will be in the lines of closed lifts and restrictions in the number of visitors allowed, but it may also result in lasting repercussion, such as facilities having to shut down ski lifts to reduce costs, or even forcing smaller facilities to shut down entirely.

The pandemic can also lead to long-lasting changes in business structures and work flexibility. Combined with increased activity in the cabin market, this may lead to people spending more time at their cabins, not only during the weekends and vacations. This could furthermore result in a change of demand during the different days of the week, and also lead to higher sales of seasonal passes at the expense of day passes, thus impacting the structure of the demand. The aftermath of the pandemic is yet to be known, but significant changes in supply and demand could possibly render our forecast model(s) less accurate.

6 Summary, limitations, and further research

6.1 Summary

This thesis aims to provide a better understanding of the importance of weather as a predictor of demand. Time series data from one alpine skiing facility in the Inland region of Norway is used to develop forecast regression models, predicting demand one day forward. By conducting a statistical comparison of models with and without weather variables, we find that temperature, precipitation, and snow depth does not prove to be of much importance for predictive accuracy. This is rather interesting, seeing that it contradicts the weighted emphasis on weather in many earlier studies as well as our initial hypothesis. In fact, it seems as if seasonal variables, such as day of the week and holidays, provides more or equally accurate predictions alone, even though the difference in predictive accuracy is statistically insignificant with the data at hand. While the weather still might be of greater importance under different circumstances, such as more extreme weather conditions, there appear to be other variables that could possibly be of greater importance for predictive accuracy. Seeing that seasonal factors are more stable and predictable than weather, this could actually be an uplifting result, as the complicated issue of weather may not need to consume as many resources as one might have initially thought.

6.2 Implications

Earlier research on demand in the alpine skiing industry has examined weather as a predictor for demand. These studies show varying results, but they are, however, spread in terms of geography, aggregation level, weather variables analysed, and methods applied. Our research adds to the pool of knowledge by examining how temperature, precipitation and snow depth contribute to the demand for adult day passes at an alpine skiing facility in the Inland region of Norway.

Precipitation has a negative effect on demand, while the effects of temperature and snow depth are positive. This is no revolutionary finding, but somewhat more surprising is the fact that we find no statistically significant improvement in the predictive accuracy of a forecast model by adding weather variables. This suggests that, although certain weather variables may be important for demand forecasting at other aggregation levels, their contribution to forecasting at a single facility is limited. As most of the earlier research is focused on the demand outside the Nordic countries (with the first formal study looking into Norwegian skiing facilities being in 2017), this research can offer a theoretical perspective of how weather affects the demand in an area that is not as greatly represented in the literature. Seeing that there are many contradicting findings on the subject, the results of this thesis furthermore support some of these views, meaning that it could be possible to relate more to other studies despite them being based on different geographies and aggregations.

There is also a practical value in the regression models developed, as they can be used by alpine skiing facilities for varied purposes. A less complex model, only containing seasonal variables, could be more applicable for the facility to use on their own, as it does not require the linkage to weather forecast data. The models could be used for both short-term and long-term decision-making. In the short run, on days or periods with low predicted demand, the facility could introduce measures to increase demand, such as campaigns or family activities. Additionally, the models can be used to regulate staffing and complementary products and services. On days with high demand extra staffing should be put in place to prevent long queues, and sufficient levels of food and beverages should be ordered for the cafés to ensure that they can serve their customers. In the long run, the regression models can be used for planning and management activities, and they can also be used as a basis for implementing dynamic pricing.

6.3 Limitations

There are some limitations in our approach to the research question. The limitations have been somewhat addressed throughout this paper, but in summarization, they are mainly linked to the weather forecast data. Simulating forecast data through the use of accuracy measures obtained from MET instead of forecast data represent a source of error in the models. The result could possibly be a misrepresentation of weather as a predictor for demand. Further research using forecast data obtained from a reliable source could improve the understanding of weather's influence on demand in a forecasting setting.

6.4 Further research

Seeing that our approach included simulated weather forecast data, it could be of interest to conduct a similar study based on actual forecast data collected from MET. Using historical forecast data could resolve the problems encountered in this thesis as a result of simulated forecast data and provide a better understanding of the importance of weather as a predictor for demand in the alpine skiing industry.

Furthermore, the scope and size of the Covid-19's impact on the industry are yet to be known. A further investigation into the short- and long-term effects of the pandemic could be interesting, especially since any major changes to the supply and/or demand side of the market could influence the performance and the relevance of the regression models developed in this thesis.

Lastly, it could be interesting to add additional customer groups and types of passes to a forecast analysis, to get a better understanding of the demand for ski lift passes. It could be possible that both the seasonal and weather variables have a greater impact on different pass types, such as the two- or three-hour passes. Our models are delimited to adult day passes, but there could be differences in demand between customer groups and different passes that our analysis did not examine. If the facility is to use our models, they could add the remaining daily passes and customer groups to get a model that possibly better captures the total demand for their product and services. Or even better, if they are able to detect when the passes of longer duration are being used, the entire demand can be captured in full.

References

Alpinanleggenes Landsforening. (n.d.). *Bransjerapport Alpinbransjen i Norge 2018/19*. https://indd.adobe.com/view/cc2b0cf0-73a2-4ccc-bd99-b7bd3b9ffe8f

Amidi, A., & Amidi, S. (2018, September 9). *Machine Learning Tips and Tricks Cheatsheet*. Stanford University. https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, *23*(2), 321–327. https://doi.org/10.1016/j.ijforecast.2007.03.004

Bower, T. (n.d.). *Fourier Series*. Retrieved March 1, 2021, from http://faculty.salina.k-state.edu/tim/mVision/freq-domain/fourier.html

Chambers, J. C., Mullick, S. K., & Smith, D. D. (1971, July 1). How to Choose the Right Forecasting Technique. *Harvard Business Review*. https://hbr.org/1971/07/how-to-choose-the-right-forecasting-technique

Chatfield, C. (2005). Time-series forecasting. *Significance*, *2*(3), 131–133. https://doi.org/10.1111/j.1740-9713.2005.00117.x

Dalen, H. B., & Gram, K. H. (2020, October 28). *Skigåing har blitt mindre populært*. Statistics Norway. https://www.ssb.no/kultur-og-fritid/artikler-og-publikasjoner/skigaing-har-blitt-mindre-populaert

Diebold, F. X. (2017). *Forecasting*. Department of Economics, University of Philadelphia. https://www.sas.upenn.edu/~fdiebold/Teaching221/Forecasting.pdf

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, *13*(3), 253–263.

Edwards, G. (2018, November 18). *Machine Learning | An Introduction*. Towards Data Science. https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0

Elliott, G., & Timmermann, A. (2016). *Economic forecasting*. Princeton University Press.

Falk, M. (2013). Impact of Long-Term Weather on Domestic and Foreign Winter Tourism Demand: Long-Term Weather and Winter Tourism. *International Journal of Tourism Research*, *15*(1), 1–17. https://doi.org/10.1002/jtr.865

Falk, M. (2015). The Demand for Winter Sports: Empirical Evidence for the Largest French Ski-Lift Operator. *Tourism Economics*, *21*(3), 561–580. https://doi.org/10.5367/te.2013.0366

Falk, M., & Vieru, M. (2017). Demand for downhill skiing in subarctic climates. *Scandinavian Journal of Hospitality and Tourism*, *17*(4), 388–405. https://doi.org/10.1080/15022250.2016.1238780

Frost, J. (2017, April 12). How to Interpret P-values and Coefficients in Regression Analysis. *Statistics By Jim*. http://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/

Gómez Martín, M. a B. (2005). Weather, climate and tourism a geographical perspective. *Annals of Tourism Research*, *32*(3), 571–591. https://doi.org/10.1016/j.annals.2004.08.004

Gössling, S., Scott, D., Hall, C. M., Ceron, J.-P., & Dubois, G. (2012). Consumer behaviour and demand response of tourists to climate change. *Annals of Tourism Research*, *39*(1), 36–58. https://doi.org/10.1016/j.annals.2011.11.002

Grinde, L., Mamen, J., Tunheim, K., & Tveito, O. E. (2020). Været i Norge. Klimatologisk månedsoversikt. Februar 2020 og vintersesongen 2019/20. *Meterologisk Institutt*, *2*, 38.

Gupta, P. (2017, November 16). *Regularization in Machine Learning*. Medium. https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a

Hair, J. F., Black, W., Anderson, R., & Babin, B. (2018). *Multivariate Data Analysis* (8th ed.). Cengage Learning EMEA.

Hansen, P. R. (2010). *A Winner's Curse for Econometric Models: On the Joint Distribution of In-Sample Fit and Out-of-Sample Fit and its Implications for Model Selection*. 39.

Haugom, E. (2015). *INN3027/1 Pricing and Revenue*. 106.

Holden, S. (2016). *Makroøkonomi*. Cappelen Damm akademisk.

Holmgren, M. A., & McCracken, V. A. (2014). What Affects Demand for "The Greatest Snow On Earth?" *Journal of Hospitality Marketing & Management*, *23*(1), 1–20. https://doi.org/10.1080/19368623.2012.746212

Homleid, M. (n.d.). *Verification of Operational Weather Prediction Models December 2019 to February 2020*. 83.

Hyndman, R. J. (n.d.). *Fourier terms for modelling seasonality*. Retrieved February 24, 2021, from https://pkg.robjhyndman.com/forecast/reference/fourier.html

Hyndman, R. J. (2011, March 14). *Statistical tests for variable selection*. https://robjhyndman.com/hyndsight/tests2/

Hyndman, R. J., & Athanasopoulus, G. (2018). *Forecasting: Principles and Practice (2nd ed)* [E-reader version]. OTexts. https://Otexts.com/fpp2/

iPaaSki. (n.d.). Objective and Work Packages. *IPaaSki*. Retrieved March 21, 2021, from https://www.ipaaski.com/about/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Johannessen, A., Christoffersen, L., & Tufte, P. A. (2020). *Forskningsmetode for økonomisk-administrative fag* (4. utgave.). Abstrakt forlag. https://www.nb.no/search?q=oaiid:"oai:nb.bibsys.no:999920086378402202"&mediatype=bøker

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kirshners, A., & Borisov, A. (2012). A Comparative Analysis of Short Time Series Processing Methods. *Information Technology and Management Science*, *15*(1). https://doi.org/10.2478/v10313-012-0009-4

Kostenko, A. V., & Hyndman, R. J. (n.d.). *Forecasting without significance tests?* 5.

Kulturdepartementet. (2011). *Veileder Alpinanlegg*.

Lai, K. (2020, February 27). *Time Series Analysis and Weather Forecast in Python*. Medium. https://medium.com/@llmkhoa511/time-series-analysis-and-weather-forecast-in-python-e80b664c7f71

Machine Learning with R. (n.d.). *Ridge and Lasso Regression Models*. Machine Learning with R. http://wavedatalab.github.io/machinelearningwithr/post4.html

Makridakis, S., & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. https://doi.org/10.1016/S0169-2070(00)00057-1

Malasevska, I. (2017). *Innovative pricing approaches in the alpine skiing industry* [Doctoral thesis]. Inland Norway University of Applied Sciences.

Malasevska, I., & Haugom, E. (2019). Alpine skiing demand patterns. *Scandinavian Journal of Hospitality and Tourism*, *19*(4–5), 390–403. https://doi.org/10.1080/15022250.2018.1539924

Malasevska, I., Haugom, E., & Lien, G. (2017). Modelling and forecasting alpine skier visits. *Tourism Economics : The Business and Finance of Tourism and Recreation*, *23*(3), 669–679. https://doi.org/10.5367/te.2015.0524

Mavuduru, A. (2020, November 12). *What "no free lunch" really means in machine learning*. Medium. https://towardsdatascience.com/what-no-free-lunch-really-means-in-machine-learning-85493215625d

Moksony, F. (1999). Small is beautiful. The use and interpretation of R2 in social research. *Szociológiai Szemle*. https://www.academia.edu/3880005/Small_is_beautiful_The_use_and_interpretation_of_R2_in_social_research

Norske alpinanlegg og fjelldestinasjoner. (n.d.). *Norske alpinanlegg og fjelldestinasjoner*. Alpin Og Fjell. Retrieved March 25, 2021, from https://alpinogfjell.no/om-oss

Oleszak, M. (2019, November 12). *Regularization: Ridge, Lasso and Elastic Net*. DataCamp Community. https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

Oppen, M., Mørk, B. E., & Haus, E. (2020). *Kvantitative og kvalitative metoder i merkantile fag: En introduksjon* (1. utgave.). Cappelen Damm akademisk.

Parmar, R. (2018, September 2). *Common Loss functions in machine learning*. Medium. https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23

Pindyck, R. S., & Rubinfeld, D. L. (2018). *Microeconomics* (9th ed.; Global ed.). Pearson Education.

Qshick. (2019, January 3). *Ridge Regression for Better Usage*. https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Rose, L. T., & Fischer, K. W. (2011). Garbage In, Garbage Out: Having Useful Data Is Everything. *Measurement: Interdisciplinary Research & Perspective*, *9*(4), 222–226. https://doi.org/10.1080/15366367.2011.632338

Seif, G. (2021, January 25). *Selecting the best Machine Learning algorithm for your regression problem*. Medium. https://towardsdatascience.com/selecting-the-best-machine-learning-algorithm-for-your-regression-problem-20c330bad4ef

Shih, C., Nicholls, S., & Holecek, D. F. (2009). Impact of Weather on Downhill Ski Lift Ticket Sales. *Journal of Travel Research*, *47*(3), 359–372. https://doi.org/10.1177/0047287508321207

Ski Info. (2021). *Skisteder Norge: Oversikt over skianlegg med beliggenhet og høyde over havet*. Skiinfo. /norge/statistikk.html

Smith, A. (2008). *An inquiry into the nature and causes of the wealth of nations* (A selected ed.). Univertsity Press.

Stock, J. H., & Watson, M. W. (2019). *Introduction to Econometrics* (4th ed.). Pearson Education Limited. https://www.adlibris.com/no/bok/introduction-to-econometrics-update-global-edition-9781292071312

Surugiu, C., Dincă, A.-I., & Micu, D. (2010). *Tourism Destinations Vulnerable to Climate Changes: An Econometric Approach on Predeal Resort*. *62*(1), 111–120.

Tuv, N. (2019, April 16). *Vi går mindre på ski enn før*. Statistics Norway. https://www.ssb.no/kultur-og-fritid/artikler-og-publikasjoner/vi-gar-mindre-pa-ski-enn-for

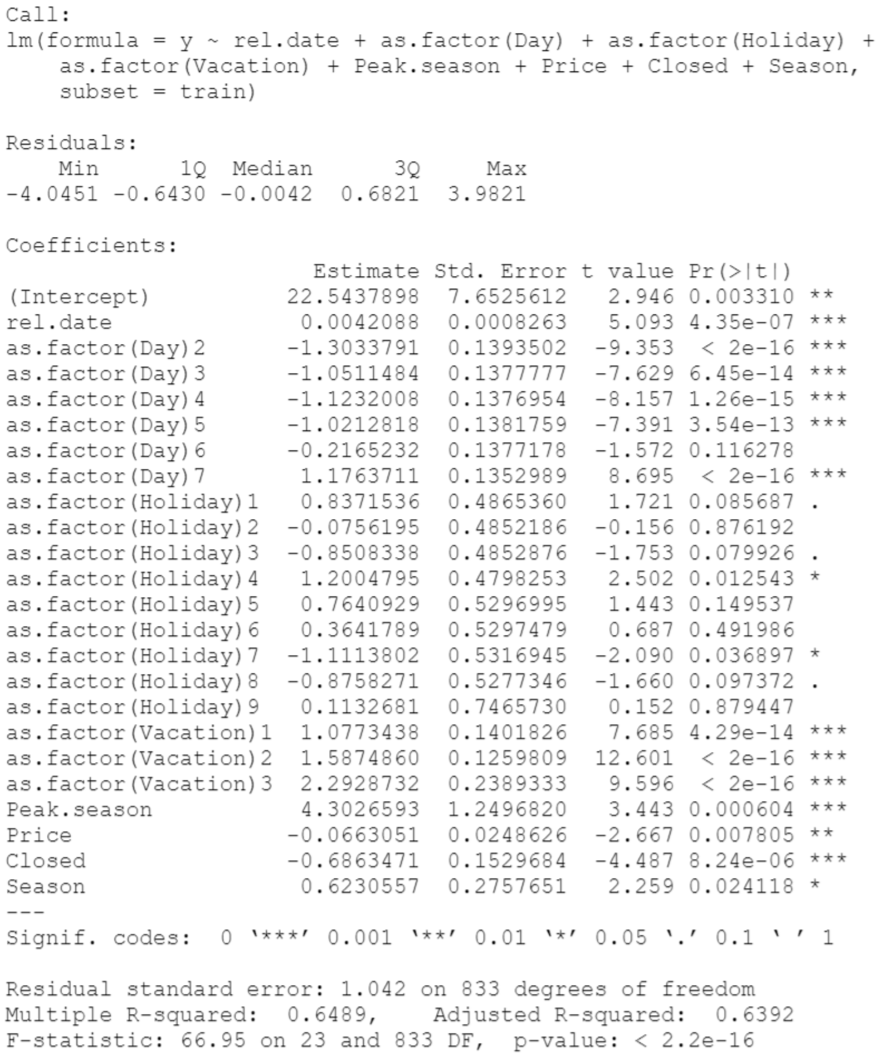
Vanat, L. (2020). *Ski resorts*. https://vanat.ch/international-report-on-snow-mountain-tourism-copy-1

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2015). *Model Comparison and the Principle of Parsimony*. The Oxford Handbook of Computational and Mathematical Psychology. https://doi.org/10.1093/oxfordhb/9780199957996.013.14

Zaiontz, C. (n.d.). *Diebold-Mariano Test*. Retrieved March 23, 2021, from https://www.real-statistics.com/time-series-analysis/forecasting-accuracy/diebold-mariano-test/

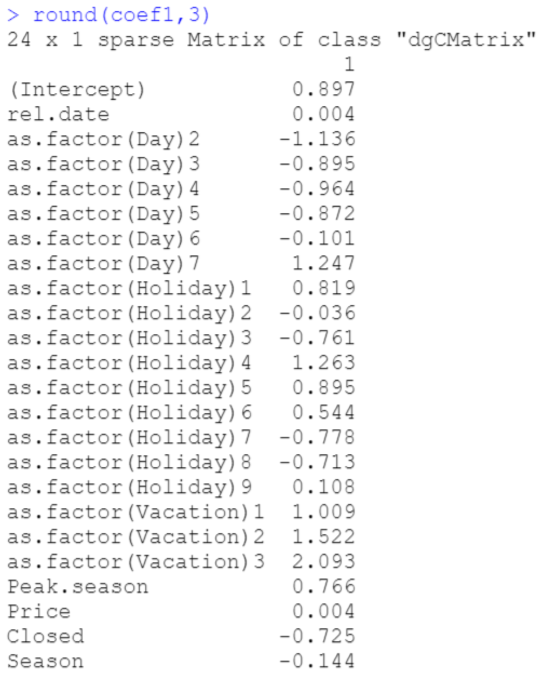
Appendices

Appendix A – Model 1 regression output



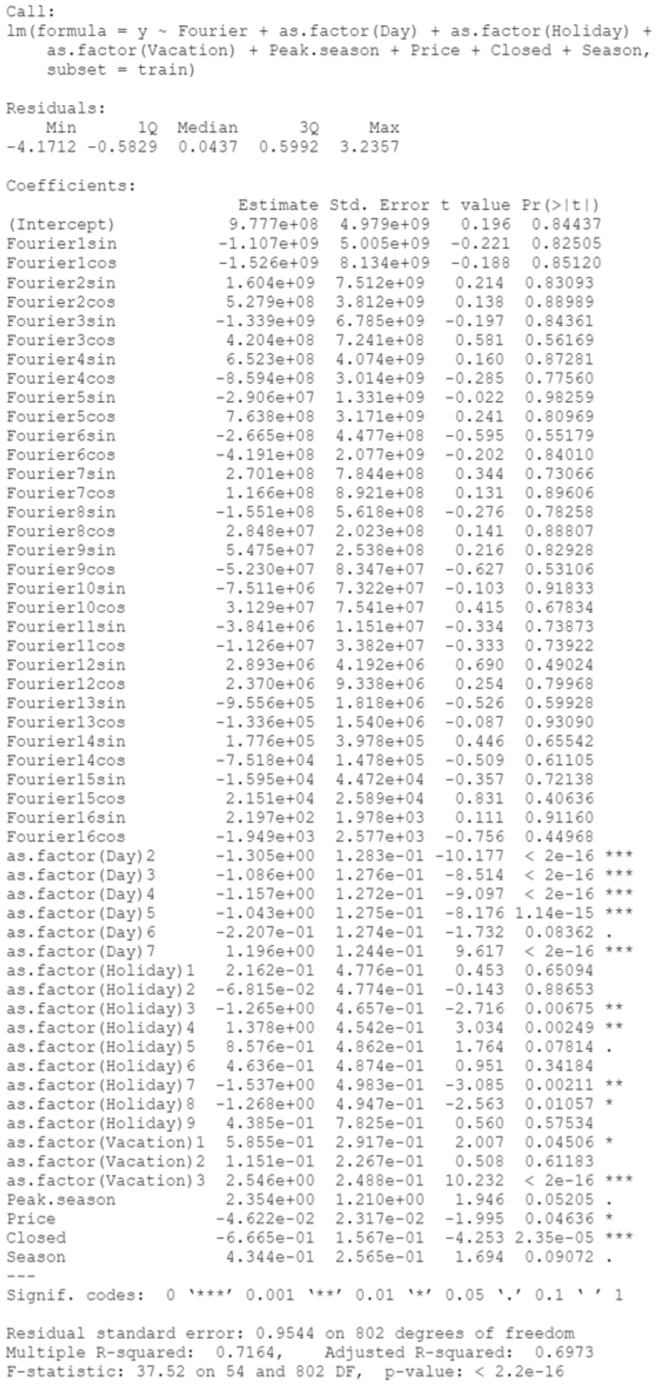
```
Call:
lm(formula = y ~ rel.date + as.factor(Day) + as.factor(Holiday) +
    as.factor(Vacation) + Peak.season + Price + Closed + Season,
    subset = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0451 -0.6430 -0.0042  0.6821  3.9821 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)          22.5437898  7.6525612   2.946 0.003310 ** 
rel.date              0.0042088  0.0008263   5.093 4.35e-07 ***
as.factor(Day)2      -1.3033791  0.1393502  -9.353  < 2e-16 ***
as.factor(Day)3      -1.0511484  0.1377777  -7.629 6.45e-14 ***
as.factor(Day)4      -1.1232008  0.1376954  -8.157 1.26e-15 ***
as.factor(Day)5      -1.0212818  0.1381759  -7.391 3.54e-13 ***
as.factor(Day)6      -0.2165232  0.1377178  -1.572 0.116278    
as.factor(Day)7       1.1763711  0.1352989   8.695  < 2e-16 ***
as.factor(Holiday)1   0.8371536  0.4865360   1.721 0.085687 .  
as.factor(Holiday)2  -0.0756195  0.4852186  -0.156 0.876192    
as.factor(Holiday)3  -0.8508338  0.4852876  -1.753 0.079926 .  
as.factor(Holiday)4   1.2004795  0.4798253   2.502 0.012543 *  
as.factor(Holiday)5   0.7640929  0.5296995   1.443 0.149537    
as.factor(Holiday)6   0.3641789  0.5297479   0.687 0.491986    
as.factor(Holiday)7  -1.1113802  0.5316945  -2.090 0.036897 *  
as.factor(Holiday)8  -0.8758271  0.5277346  -1.660 0.097372 .  
as.factor(Holiday)9   0.1132681  0.7465730   0.152 0.879447    
as.factor(Vacation)1  1.0773438  0.1401826   7.685 4.29e-14 ***
as.factor(Vacation)2  1.5874860  0.1259809  12.601  < 2e-16 ***
as.factor(Vacation)3  2.2928732  0.2389333   9.596  < 2e-16 ***
Peak.season           4.3026593  1.2496820   3.443 0.000604 ***
Price                -0.0663051  0.0248626  -2.667 0.007805 ** 
Closed               -0.6863471  0.1529684  -4.487 8.24e-06 ***
Season                0.6230557  0.2757651   2.259 0.024118 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.042 on 833 degrees of freedom
Multiple R-squared:  0.6489,    Adjusted R-squared:  0.6392 
F-statistic: 66.95 on 23 and 833 DF,  p-value: < 2.2e-16
```

Appendix B – Model 2 regression output



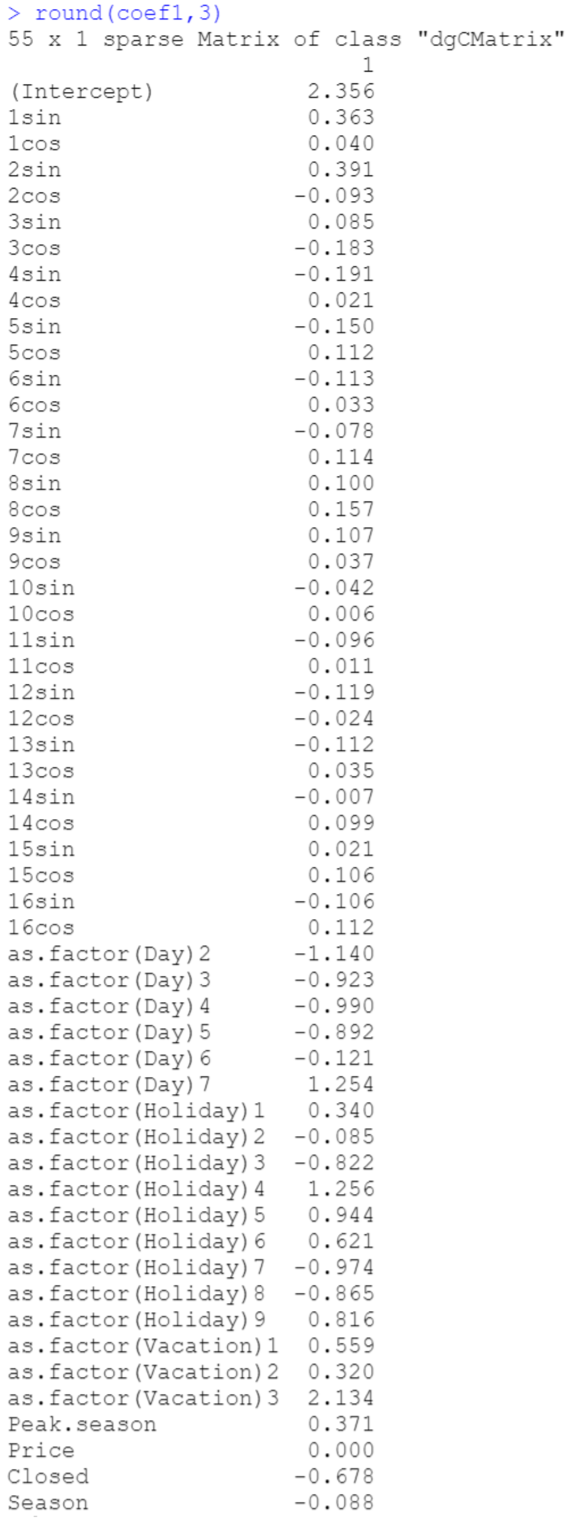
```
> round(coef1,3)
24 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)              0.897
rel.date                 0.004
as.factor(Day)2         -1.136
as.factor(Day)3         -0.895
as.factor(Day)4         -0.964
as.factor(Day)5         -0.872
as.factor(Day)6         -0.101
as.factor(Day)7          1.247
as.factor(Holiday)1      0.819
as.factor(Holiday)2     -0.036
as.factor(Holiday)3     -0.761
as.factor(Holiday)4      1.263
as.factor(Holiday)5      0.895
as.factor(Holiday)6      0.544
as.factor(Holiday)7     -0.778
as.factor(Holiday)8     -0.713
as.factor(Holiday)9      0.108
as.factor(Vacation)1     1.009
as.factor(Vacation)2     1.522
as.factor(Vacation)3     2.093
Peak.season              0.766
Price                    0.004
Closed                  -0.725
Season                  -0.144
```

Appendix C – Model 3 regression output



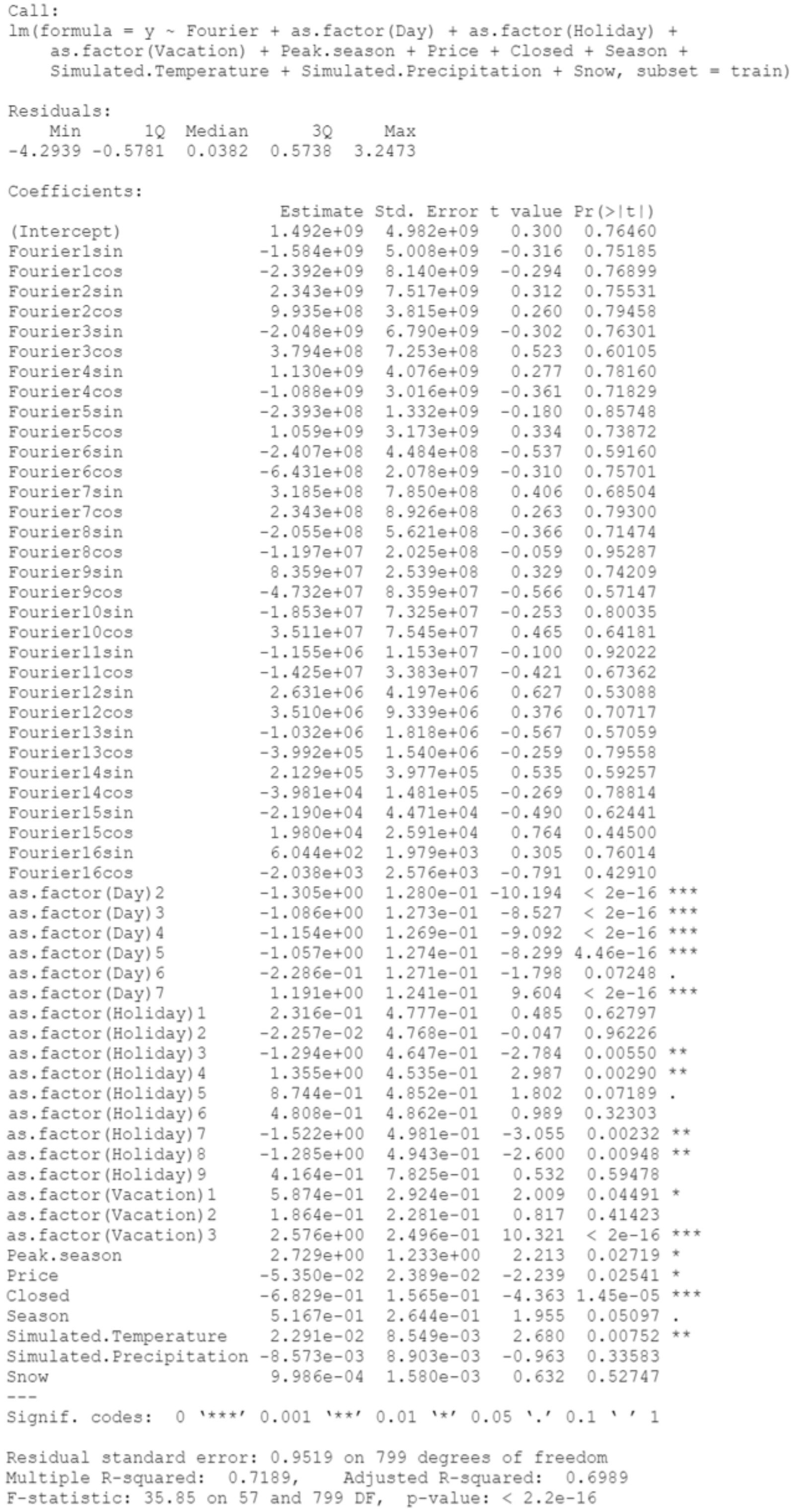
```
Call:
lm(formula = y ~ Fourier + as.factor(Day) + as.factor(Holiday) + 
    as.factor(Vacation) + Peak.season + Price + Closed + Season, 
    subset = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1712 -0.5829  0.0437  0.5992  3.2357 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           9.777e+08  4.979e+09   0.196  0.84437    
Fourier1sin          -1.107e+09  5.005e+09  -0.221  0.82505    
Fourier1cos          -1.526e+09  8.134e+09  -0.188  0.85120    
Fourier2sin           1.604e+09  7.512e+09   0.214  0.83093    
Fourier2cos           5.279e+08  3.812e+09   0.138  0.88989    
Fourier3sin          -1.339e+09  6.785e+09  -0.197  0.84361    
Fourier3cos           4.204e+08  7.241e+08   0.581  0.56169    
Fourier4sin           6.523e+08  4.074e+09   0.160  0.87281    
Fourier4cos          -8.594e+08  3.014e+09  -0.285  0.77560    
Fourier5sin          -2.906e+07  1.331e+09  -0.022  0.98259    
Fourier5cos           7.638e+08  3.171e+09   0.241  0.80969    
Fourier6sin          -2.665e+08  4.477e+08  -0.595  0.55179    
Fourier6cos          -4.191e+08  2.077e+09  -0.202  0.84010    
Fourier7sin           2.701e+08  7.844e+08   0.344  0.73066    
Fourier7cos           1.166e+08  8.921e+08   0.131  0.89606    
Fourier8sin          -1.551e+08  5.618e+08  -0.276  0.78258    
Fourier8cos           2.848e+07  2.023e+08   0.141  0.88807    
Fourier9sin           5.475e+07  2.538e+08   0.216  0.82928    
Fourier9cos          -5.230e+07  8.347e+07  -0.627  0.53106    
Fourier10sin         -7.511e+06  7.322e+07  -0.103  0.91833    
Fourier10cos          3.129e+07  7.541e+07   0.415  0.67834    
Fourier11sin         -3.841e+06  1.151e+07  -0.334  0.73873    
Fourier11cos         -1.126e+07  3.382e+07  -0.333  0.73922    
Fourier12sin          2.893e+06  4.192e+06   0.690  0.49024    
Fourier12cos          2.370e+06  9.338e+06   0.254  0.79968    
Fourier13sin         -9.556e+05  1.818e+06  -0.526  0.59928    
Fourier13cos         -1.336e+05  1.540e+06  -0.087  0.93090    
Fourier14sin          1.776e+05  3.978e+05   0.446  0.65542    
Fourier14cos         -7.518e+04  1.478e+05  -0.509  0.61105    
Fourier15sin         -1.595e+04  4.472e+04  -0.357  0.72138    
Fourier15cos          2.151e+04  2.589e+04   0.831  0.40636    
Fourier16sin          2.197e+02  1.978e+03   0.111  0.91160    
Fourier16cos         -1.949e+03  2.577e+03  -0.756  0.44968    
as.factor(Day)2      -1.305e+00  1.283e-01 -10.177  < 2e-16 ***
as.factor(Day)3      -1.086e+00  1.276e-01  -8.514  < 2e-16 ***
as.factor(Day)4      -1.157e+00  1.272e-01  -9.097  < 2e-16 ***
as.factor(Day)5      -1.043e+00  1.275e-01  -8.176 1.14e-15 ***
as.factor(Day)6      -2.207e-01  1.274e-01  -1.732  0.08362 .  
as.factor(Day)7       1.196e+00  1.244e-01   9.617  < 2e-16 ***
as.factor(Holiday)1   2.162e-01  4.776e-01   0.453  0.65094    
as.factor(Holiday)2  -6.815e-02  4.774e-01  -0.143  0.88653    
as.factor(Holiday)3  -1.265e+00  4.657e-01  -2.716  0.00675 ** 
as.factor(Holiday)4   1.378e+00  4.542e-01   3.034  0.00249 ** 
as.factor(Holiday)5   8.576e-01  4.862e-01   1.764  0.07814 .  
as.factor(Holiday)6   4.636e-01  4.874e-01   0.951  0.34184    
as.factor(Holiday)7  -1.537e+00  4.983e-01  -3.085  0.00211 ** 
as.factor(Holiday)8  -1.268e+00  4.947e-01  -2.563  0.01057 *  
as.factor(Holiday)9   4.385e-01  7.825e-01   0.560  0.57534    
as.factor(Vacation)1  5.855e-01  2.917e-01   2.007  0.04506 *  
as.factor(Vacation)2  1.151e-01  2.267e-01   0.508  0.61183    
as.factor(Vacation)3  2.546e+00  2.488e-01  10.232  < 2e-16 ***
Peak.season           2.354e+00  1.210e+00   1.946  0.05205 .  
Price                -4.622e-02  2.317e-02  -1.995  0.04636 *  
Closed               -6.665e-01  1.567e-01  -4.253 2.35e-05 ***
Season                4.344e-01  2.565e-01   1.694  0.09072 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9544 on 802 degrees of freedom
Multiple R-squared:  0.7164,	Adjusted R-squared:  0.6973 
F-statistic: 37.52 on 54 and 802 DF,  p-value: < 2.2e-16
```

Appendix D – Model 4 regression output



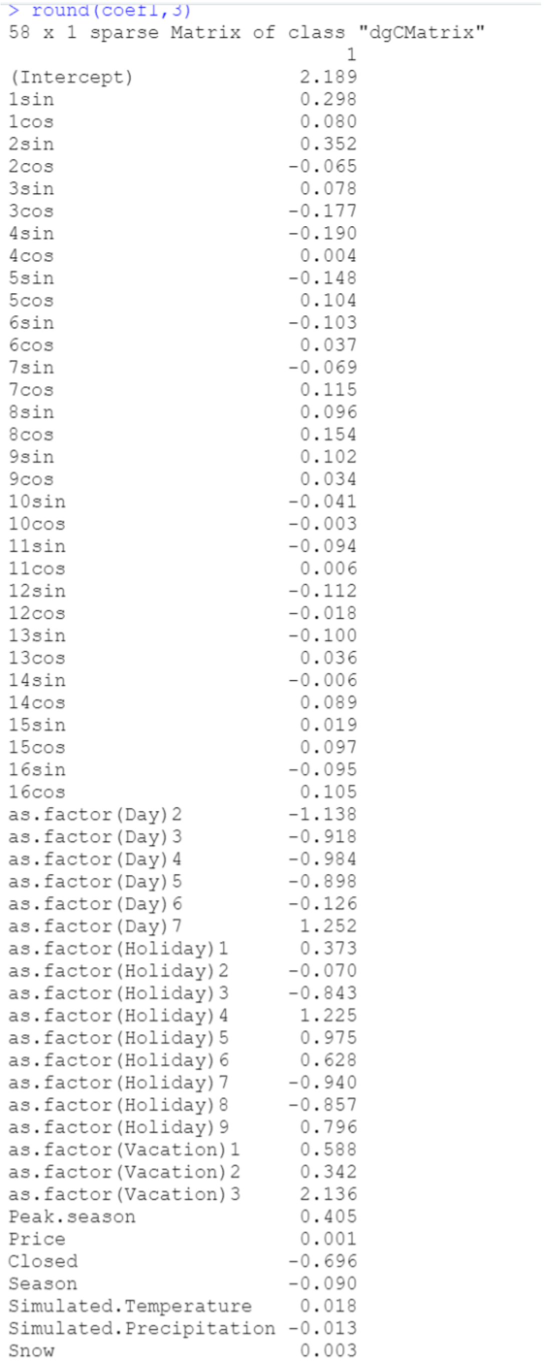
```
> round(coef1,3)
55 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)             2.356
1sin                    0.363
1cos                    0.040
2sin                    0.391
2cos                   -0.093
3sin                    0.085
3cos                   -0.183
4sin                   -0.191
4cos                    0.021
5sin                   -0.150
5cos                    0.112
6sin                   -0.113
6cos                    0.033
7sin                   -0.078
7cos                    0.114
8sin                    0.100
8cos                    0.157
9sin                    0.107
9cos                    0.037
10sin                  -0.042
10cos                   0.006
11sin                  -0.096
11cos                   0.011
12sin                  -0.119
12cos                  -0.024
13sin                  -0.112
13cos                   0.035
14sin                  -0.007
14cos                   0.099
15sin                   0.021
15cos                   0.106
16sin                  -0.106
16cos                   0.112
as.factor(Day)2        -1.140
as.factor(Day)3        -0.923
as.factor(Day)4        -0.990
as.factor(Day)5        -0.892
as.factor(Day)6        -0.121
as.factor(Day)7         1.254
as.factor(Holiday)1     0.340
as.factor(Holiday)2    -0.085
as.factor(Holiday)3    -0.822
as.factor(Holiday)4     1.256
as.factor(Holiday)5     0.944
as.factor(Holiday)6     0.621
as.factor(Holiday)7    -0.974
as.factor(Holiday)8    -0.865
as.factor(Holiday)9     0.816
as.factor(Vacation)1    0.559
as.factor(Vacation)2    0.320
as.factor(Vacation)3    2.134
Peak.season             0.371
Price                   0.000
Closed                 -0.678
Season                 -0.088
```

Appendix E – Model 5 regression output



```
Call:
lm(formula = y ~ Fourier + as.factor(Day) + as.factor(Holiday) +
    as.factor(Vacation) + Peak.season + Price + Closed + Season +
    Simulated.Temperature + Simulated.Precipitation + Snow, subset = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2939 -0.5781  0.0382  0.5738  3.2473

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.492e+09  4.982e+09   0.300  0.76460
Fourier1sin            -1.584e+09  5.008e+09  -0.316  0.75185
Fourier1cos            -2.392e+09  8.140e+09  -0.294  0.76899
Fourier2sin             2.343e+09  7.517e+09   0.312  0.75531
Fourier2cos             9.935e+08  3.815e+09   0.260  0.79458
Fourier3sin            -2.048e+09  6.790e+09  -0.302  0.76301
Fourier3cos             3.794e+08  7.253e+08   0.523  0.60105
Fourier4sin             1.130e+09  4.076e+09   0.277  0.78160
Fourier4cos            -1.088e+09  3.016e+09  -0.361  0.71829
Fourier5sin            -2.393e+08  1.332e+09  -0.180  0.85748
Fourier5cos             1.059e+09  3.173e+09   0.334  0.73872
Fourier6sin            -2.407e+08  4.484e+08  -0.537  0.59160
Fourier6cos            -6.431e+08  2.078e+09  -0.310  0.75701
Fourier7sin             3.185e+08  7.850e+08   0.406  0.68504
Fourier7cos             2.343e+08  8.926e+08   0.263  0.79300
Fourier8sin            -2.055e+08  5.621e+08  -0.366  0.71474
Fourier8cos            -1.197e+07  2.025e+08  -0.059  0.95287
Fourier9sin             8.359e+07  2.539e+08   0.329  0.74209
Fourier9cos            -4.732e+07  8.359e+07  -0.566  0.57147
Fourier10sin           -1.853e+07  7.325e+07  -0.253  0.80035
Fourier10cos            3.511e+07  7.545e+07   0.465  0.64181
Fourier11sin           -1.155e+06  1.153e+07  -0.100  0.92022
Fourier11cos           -1.425e+07  3.383e+07  -0.421  0.67362
Fourier12sin            2.631e+06  4.197e+06   0.627  0.53088
Fourier12cos            3.510e+06  9.339e+06   0.376  0.70717
Fourier13sin           -1.032e+06  1.818e+06  -0.567  0.57059
Fourier13cos           -3.992e+05  1.540e+06  -0.259  0.79558
Fourier14sin            2.129e+05  3.977e+05   0.535  0.59257
Fourier14cos           -3.981e+04  1.481e+05  -0.269  0.78814
Fourier15sin           -2.190e+04  4.471e+04  -0.490  0.62441
Fourier15cos            1.980e+04  2.591e+04   0.764  0.44500
Fourier16sin            6.044e+02  1.979e+03   0.305  0.76014
Fourier16cos           -2.038e+03  2.576e+03  -0.791  0.42910
as.factor(Day)2        -1.305e+00  1.280e-01 -10.194  < 2e-16 ***
as.factor(Day)3        -1.086e+00  1.273e-01  -8.527  < 2e-16 ***
as.factor(Day)4        -1.154e+00  1.269e-01  -9.092  < 2e-16 ***
as.factor(Day)5        -1.057e+00  1.274e-01  -8.299 4.46e-16 ***
as.factor(Day)6        -2.286e-01  1.271e-01  -1.798  0.07248 .
as.factor(Day)7         1.191e+00  1.241e-01   9.604  < 2e-16 ***
as.factor(Holiday)1     2.316e-01  4.777e-01   0.485  0.62797
as.factor(Holiday)2    -2.257e-02  4.768e-01  -0.047  0.96226
as.factor(Holiday)3    -1.294e+00  4.647e-01  -2.784  0.00550 **
as.factor(Holiday)4     1.355e+00  4.535e-01   2.987  0.00290 **
as.factor(Holiday)5     8.744e-01  4.852e-01   1.802  0.07189 .
as.factor(Holiday)6     4.808e-01  4.862e-01   0.989  0.32303
as.factor(Holiday)7    -1.522e+00  4.981e-01  -3.055  0.00232 **
as.factor(Holiday)8    -1.285e+00  4.943e-01  -2.600  0.00948 **
as.factor(Holiday)9     4.164e-01  7.825e-01   0.532  0.59478
as.factor(Vacation)1    5.874e-01  2.924e-01   2.009  0.04491 *
as.factor(Vacation)2    1.864e-01  2.281e-01   0.817  0.41423
as.factor(Vacation)3    2.576e+00  2.496e-01  10.321  < 2e-16 ***
Peak.season             2.729e+00  1.233e+00   2.213  0.02719 *
Price                  -5.350e-02  2.389e-02  -2.239  0.02541 *
Closed                 -6.829e-01  1.565e-01  -4.363 1.45e-05 ***
Season                  5.167e-01  2.644e-01   1.955  0.05097 .
Simulated.Temperature   2.291e-02  8.549e-03   2.680  0.00752 **
Simulated.Precipitation -8.573e-03  8.903e-03  -0.963  0.33583
Snow                    9.986e-04  1.580e-03   0.632  0.52747
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9519 on 799 degrees of freedom
Multiple R-squared:  0.7189,    Adjusted R-squared:  0.6989
F-statistic: 35.85 on 57 and 799 DF,  p-value: < 2.2e-16
```

Appendix F – Model 6 regression output



```
> round(coef1,3)
58 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)              2.189
1sin                     0.298
1cos                     0.080
2sin                     0.352
2cos                    -0.065
3sin                     0.078
3cos                    -0.177
4sin                    -0.190
4cos                     0.004
5sin                    -0.148
5cos                     0.104
6sin                    -0.103
6cos                     0.037
7sin                    -0.069
7cos                     0.115
8sin                     0.096
8cos                     0.154
9sin                     0.102
9cos                     0.034
10sin                   -0.041
10cos                   -0.003
11sin                   -0.094
11cos                    0.006
12sin                   -0.112
12cos                   -0.018
13sin                   -0.100
13cos                    0.036
14sin                   -0.006
14cos                    0.089
15sin                    0.019
15cos                    0.097
16sin                   -0.095
16cos                    0.105
as.factor(Day)2         -1.138
as.factor(Day)3         -0.918
as.factor(Day)4         -0.984
as.factor(Day)5         -0.898
as.factor(Day)6         -0.126
as.factor(Day)7          1.252
as.factor(Holiday)1      0.373
as.factor(Holiday)2     -0.070
as.factor(Holiday)3     -0.843
as.factor(Holiday)4      1.225
as.factor(Holiday)5      0.975
as.factor(Holiday)6      0.628
as.factor(Holiday)7     -0.940
as.factor(Holiday)8     -0.857
as.factor(Holiday)9      0.796
as.factor(Vacation)1     0.588
as.factor(Vacation)2     0.342
as.factor(Vacation)3     2.136
Peak.season              0.405
Price                    0.001
Closed                  -0.696
Season                  -0.090
Simulated.Temperature    0.018
Simulated.Precipitation -0.013
Snow                     0.003
```

Appendix G – R script

```r
#---------- Functions

# Check whether a package is installed
is_installed=function(pckg){ nzchar(system.file(package=pckg)) }

fcst_eval=function(actual,predicted){
  e=predicted-actual; y=actual
  MAE=mean(abs(e))
  MAEy=mean(abs(y))
  MSE=mean(e^2)
  MSEy=mean(y^2)
  return(list(MAE=MAE,MAEy=MAEy,MSE=MSE,MSEy=MSEy))
}

#---------- Load packages
for(package in c("forecast","glmnet","mgcv","car","relaimpo")){
  print(paste0("Loading package: ",package))
  if(!is_installed(package)) install.packages(package)
  library(package=package,character.only=TRUE)
  print(paste0("Loading package: ",package))
}

#---------- Laste data
X=read.csv(file=file.choose(), header=T)       #import the data file
n=nrow(X)

#---------- Total.passes to its logarithm
attach(X)
y=Total.passes
y[y<1]=1 # treat all days with 1 or fewer visitors as having 1 visitor
logy=log(y)
X=cbind(X,logy); rm(y,logy)

#---------- Weather forecast data simulation
set.seed(1)
X$noise = rnorm(n=length(Temperature),mean=mean(Temperature),sd=1.7)
X$Simulated.Temperature = X$Temperature + X$noise

set.seed(1)
X$noise2 = rnorm(n=length(Temperature),mean=mean(Precipitation),sd=2.5)
X$Simulated.Precipitation = X$Precipitation + X$noise2

#---------- Relative date within a season
date=as.Date(Dag,format="%m/%d/%Y")
rel.date=NA # relative date
for(s in unique(Season)){
  rel.date[Season==s] = difftime( time1=date[Season==s],
time2=as.Date(paste0("01/01/",s+2014),format="%m/%d/%Y"), units="days" )
}
X=cbind(X,date,rel.date); rm(date, rel.date)

#---------- Temperature categorized
cold  =rep(0,n); cold  [(Temperature<=-10) & (Temperature>-15)]=1
icecold=rep(0,n); icecold[Temperature<=-15]=1
X=cbind(X,cold,icecold); rm(cold,icecold)
```

```
detach(X); attach(X)
sum(cold)
sum(icecold)

plot(Temperature,type="l")
abline(h=seq(from=-30,to=30,by=5),col="grey"); abline(h=-
10,lty="dashed",lwd=2,col="lightblue"); abline(h=-
15,lty="dashed",lwd=2,col="blue")
abline(v=which(cold  ==1),col="lightblue")
abline(v=which(icecold==1),col="blue")

#---------- Rain
if(FALSE){
  rain=as.numeric((Temperature>2) & (Precipitation>2.5)) # +2 degrees and 2.5
mm precipitation
  sum(rain)
}

#---------- Fourier
# Create Fourier terms (instead of daily dummies) to account for the within-
Year seasonal cycle
# how many pairs of Fourier terms: more --> high variance, low bias; fewer --
> low variance, high bias
K=80
K=40
K=20
K=16
K=10
K=5
Fourier=matrix(NA,nrow=n,ncol=2*K);
colnames(Fourier)[(1:K)*2  ]=paste0(1:K,"cos")
colnames(Fourier)[(1:K)*2-1]=paste0(1:K,"sin")
for(k in 1:K){
  sin1=sin(2*pi*k*rel.date/365.25)
  cos1=cos(2*pi*k*rel.date/365.25)
  #plot(sin1,type="l",col="red"); lines(cos1,col="blue")
  Fourier[,k*2-1]=sin1
  Fourier[,k*2  ]=cos1
}

#---------- Linear regression: in sample
train=which(Season< 6)
test =which(Season==6)

y=Total.passes   # model Total.passes directly
y=logy            # model logarithm of y

m1=lm(y~rel.date+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.s
eason+Price+Closed+Season,subset=train); summary(m1)  #Model 1
m1=lm(y~Fourier+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.se
ason+Price+Closed+Season,subset=train); summary(m1)   #Model 3
m1=lm(y~Fourier+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.se
ason+Price+Closed+Season+Simulated.Temperature+Simulated.Precipitation+Snow,s
ubset=train); summary(m1) #Model5
```

```
#Plot in-sample fit
ylim1=range(m1$resid,m1$fitted,m1$resid+m1$fitted)
#plot(m1$resid+m1$fitted,ylim=ylim1,type="l"); lines(m1$resid,col="red");
lines(m1$fitted,col="blue")
lwd1=1
par(mfrow=c(2,1),mar=c(2,2,0.5,0.5))
at=c(34,200,379,557,736,911)
labels=as.character(c(2015:2020))
plot(m1$resid+m1$fitted,ylim=ylim1,type="l", xaxt="n");
lines(m1$fitted,col="blue"); abline(h=seq(from=-200,to=500,by=50),col="grey")
axis(side=1, at=at,labels=labels)
#  abline(v=which(Ferie==1),col="lightgreen")
#  abline(v=which(Helligdag==1),col="lightgreen")
lines(m1$resid+m1$fitted,lwd=lwd1); lines(m1$fitted,lwd=lwd1,col="blue")
plot(m1$resid,ylim=ylim1,type="l", xaxt="n",col="red"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
axis(side=1, at=at,labels=labels)
# abline(v=which(Ferie==1),col="lightgreen")
# abline(v=which(Helligdag==1),col="lightgreen")
lines(m1$resid,lwd=lwd1,col="red")
par(mfrow=c(2,1))

#---------- Time-series out of sample / cross-validetion with expanding
windows
train=which(Season< 6)
test =which(Season==6)

y=Total.passes # model Total.passes directly
y=logy          # model logarithm of Total.passes

e=y_hat=lower=upper=rep(NA,n) # prediction errors (forecast errors), point
predictions, lower end of 90% prediction interval, upper end of 90%
prediction interval
for(t in which(Season==6)){

#m1=lm(y~rel.date+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.
season+Price+Closed+Season,data=X,subset=c(1:(t-1)))  #model 1

m1=lm(y~Fourier+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.se
ason+Price+Closed+Season, data=X, subset=c(1:(t-1)))  #model 3

#m1=lm(y~Fourier+as.factor(Day)+as.factor(Holiday)+as.factor(Vacation)+High.s
eason+Price+Closed+Season+Simulated.Temperature+Simulated.Precipitation+Snow,
data=X, subset=c(1:(t-1))) #model 5
  f1=predict(m1,newdata=X,interval="prediction",level=0.9)[t,]
  y_hat[t]=f1["fit"]
  lower[t]=f1["lwr"]
  upper[t]=f1["upr"]
}

fcst_eval(actual=y[test]            ,predicted=    y_hat[test] )
fcst_eval(actual=Total.passes[test],predicted=exp(y_hat[test]))

err=Total.passes[test]-exp(f1)
round(err,2)

#Plot out-of-sample fit
```

```
e=y-y_hat
ylim1=range(y,y_hat,e,na.rm=TRUE)
par(mfrow=c(2,1),mar=c(2,2,0.5,0.5))
lwd1=1
plot(y,ylim=ylim1,type="l"); lines(y_hat,col="blue"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
#abline(v=which(Vacation==1),col="lightgreen")
#abline(v=which(Holyday==1),col="lightgreen")
lines(y,lwd=lwd1);
lines(y_hat,lwd=lwd1,col="blue")
plot(e,ylim=ylim1,type="l",col="red"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
#abline(v=which(Vacation==1),col="lightgreen")
#abline(v=which(Holyday==1),col="lightgreen")
lines(e,lwd=lwd1,col="red")
par(mfrow=c(2,1))

coef(m1)

#---------- Ridge regresjon: out of sample
train=which(Season< 6)
test =which(Season==6)

y=Total.passes # model Total.passes directly
y=logy          # model logarithm of Total.passes

alpha1=0 # ridge
# Cross-validate optimal regularization intensity using an automatically-
generated grid of values for regularization intensity, with MAE as
performance metric

#Model 2
Xmat=cbind(rel.date,model.matrix(~as.factor(Day))[,-
1],model.matrix(~as.factor(Holiday))[,-
1],model.matrix(~as.factor(Vacation))[,-1],High.season,Price,Closed,Season)
#Model 4
Xmat=cbind(Fourier,model.matrix(~as.factor(Day))[,-1],
model.matrix(~as.factor(Holiday))[,-1],model.matrix(~as.factor(Vacation))[,-
1],High.season,Price,Closed,Season)
#Model 6
Xmat=cbind(Fourier,model.matrix(~as.factor(Day))[,-1],
model.matrix(~as.factor(Holiday))[,-1],model.matrix(~as.factor(Vacation))[,-
1],High.season,Price,Closed,Season,Simulated.Temperature,Simulated.Precipitat
ion,Snow)

m1=glmnet(x=Xmat[train,],y=y[train],alpha=alpha1)

plot(m1)
cvfit=cv.glmnet(x=Xmat[train,],y=y[train],alpha=alpha1,type.measure="mae",nfo
lds=66)
cvfit
plot(cvfit)
# Plots MAE against regularization intensities
mincvloss=min(cvfit$cvm)
# Minimal (w.r.t. regularization intensity) LOOCV loss
mincvloss
y_hat=predict(m1,s=cvfit$lambda.min,newx=Xmat[test,])
```

```
#Evaluate model performance
fcst_eval(actual=y[test]            ,predicted=    y_hat )
fcst_eval(actual=Total.passes[test],predicted=exp(y_hat))

coef1=coef(cvfit, s="lambda.min")
round(coef1,3)

err=Total.passes[test]-exp(y_hat)
round(err,2)

#Plot out-of-sample fit
e=Total.passes[test]-exp(y_hat)
ylim1=range(Total.passes[test],exp(y_hat),e,na.rm=TRUE)
par(mfrow=c(2,1),mar=c(2,2,0.5,0.5))
lwd1=1
at=c(23,54,85,114)
labels=as.character(c("December", "January", "February", "March"))
plot(Total.passes[test],ylim=ylim1,type="l", xaxt="n");
lines(exp(y_hat),col="blue"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
axis(side=1, at=at,labels=labels)
#abline(v=which(Vacation==1),col="lightgreen")
#abline(v=which(Holiday==1),col="lightgreen")
lines(Total.passes[test],lwd=lwd1);
lines(exp(y_hat),lwd=lwd1,col="blue")
plot(e,ylim=ylim1,type="l",col="red", xaxt="n"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
axis(side=1, at=at,labels=labels)
#abline(v=which(Vacation==1),col="lightgreen")
#abline(v=which(Holyday==1),col="lightgreen")
lines(e,lwd=lwd1,col="red")
par(mfrow=c(2,1))

#DIEBOLD MARIANO TEST
model4_err=err
model6_err=err

test_loss_model4  =abs(model4_err) # test loss of model 4
test_loss_model6  =abs(model6_err) # test loss of model 6
dm.test(e1=test_loss_model4, e2=test_loss_model6,
alternative="two.sided",h=1,power=1)

#================ Näive forecast
avr=mean(Total.passes[train])
avr

fcst_eval(actual=Total.passes[test],predicted=avr)

#=============== Extra calculations for model presentation
noholiday =which(Vacation==0)
christmas =which(Vacation==1)
winter    =which(Vacation==2)
easter    =which(Vacation==3)

avr=mean(Total.passes[noholiday])
avr
```

```
avr=mean(Total.passes[christmas])
avr
avr=mean(Total.passes[winter])
avr
avr=mean(Total.passes[easter])
avr

#============================== Weather forecast
W=read.csv(file=file.choose(), header=T)       #import the data file
n=nrow(W)

par(mfrow=c(3,1),mar=c(0.1,2,0.1,0.5))
plot(W$Temperatur,type="l")
plot(W$Nedboer   ,type="l")
plot(W$Snodybde  ,type="l")
par(mfrow=c(1,1),mar=c(2,2,0.5,0.5))

#library(forecast)
x=W$Temperatur
x=W$Nedboer
x=W$Snodybde

# Choose one of the two following lines
m=auto.arima(x)         # demonstration of ARIMA (using automated model
selection)
m=ets(x)                # demonstration of exponential smoothing (using
automated model selection)
print(summary(m))
forecast(m,h=10)
plot(forecast(m,h=10))

# Setting up time series cross validation (rolling windows)
w=round(n*0.7) # window size
m=n-w          # how many windows within the sample
f=rep(NA,n)    # "out of sample" forecasts

# EITHER...
# Forecast with exponential smoothing (using automated model selection)
for(i in 1:m){ # Runs for 5 seconds or so
  model=ets(x[i:(i+w-1)])
  f[w+i]=as.numeric(forecast(model,h=1)$mean) # change h=1 to h=7 for 7 days
ahead;
  # some more changes would be needed to make ends meet
}

# ...OR
# Forecast with ARIMA (using automated model selection)
print(Sys.time()); for(i in 1:m){ # Runs for 10 minutes or so !!!
  if(i%%10==0) print(paste0(Sys.time()," i = ",i," of m = ",m))
  model=auto.arima(x[i:(i+w-1)])
  f[w+i]=as.numeric(forecast(model,h=1)$mean) # change h=1 to h=7 for 7 days
ahead;
  # some more changes would be needed to make ends meet
}; print(Sys.time())

# Obtain errors, evaluate forecast accuracy / estimated expected loss
test=c((w+1):(n-1)) # index of out of sample data
```

```
e=x-f               # "out of sample" forecast errors
ylim1=range(x,f,e,na.rm=TRUE)
par(mfrow=c(2,1),mar=c(2,2,0.5,0.5))
plot(x[test],ylim=ylim1,type="l"); lines(f[test],col="blue");
abline(h=seq(from=-200,to=500,by=50),col="grey")
lines(x[test]);
lines(f[test],col="blue")
plot(e[test],ylim=ylim1,type="l",col="red"); abline(h=seq(from=-
200,to=500,by=50),col="grey")
lines(e[test],col="red")
par(mfrow=c(2,1))
acf(e[test])  #         autocorrelation ( ACF) plot: is there any signal
remaining in the errors?
pacf(e[test]) # partial autocorrelation (PACF) plot: is there any signal
remaining in the errors?

fcst_eval(actual=x[test],predicted=f[test])

summary(X)
#===============================
```