**To what extent is there a correlation between electricity price demand and various weather factors within a state?**

From going to bed with the heater on to waking up and pouring yourself a bowl of cereal, we subconsciously make decisions on how much electricity we use based on our current mood. However, we do not know to what extent this is a fact, furthermore there are a lot of factors that affect our mood whether it be the time of day, what we plan to do or as we want to investigate today, the weather outside.

Categorizing the weather outside as good or bad is very broad and can be considered an over classification. Our report plans to go into deeper detail on each weather factor and slowly use a regression analysis to decide the level and type of impact each explanatory variable has.

***Aim***

Our aim is to see what kind of relationship weather factors have on the price demand and use our developed relations to try and predict the direction of a price change. Optimistically, we will be able to see a vivid relationship and use this information to forecast the price demand of electricity. Three key stakeholders that we plan to appeal to are everyday consumers, electricity companies and state governments. An everyday consumer might simply take a very approach to this information, use it for menial things like interest, global awareness, and budgeting. On the other hand, electricity companies may use this information and precisely predict their change in supply and make business decisions. In conjunction, state governments can study this information and choose to subsidize electricity companies as a means of increasing overall social well-being.

***Datasets***

The two data sets of interest include the Price Demand of Electricity data 2021 and the Weather Brisbane 2021. The reason why we have only chosen Brisbane and not the other cities is because Brisbane has the most price surges on average showing a higher volatility rate. This is rather unexpected as Melbourne is more known to have a more drastic weather change.

Both datasets are given in CSV files. The included information on the Electricity data on the state of the price demand and whether it was an instance of a price surge in half an hour increments every day for a year (2021 to 2022) for every region in Australia. Our Weather data gave us information on the temperature, rainfall, sunshine, wind, humidity, MSL pressure and cloud on each day in a year.

The weather factors what we have chosen to use:

* Rainfall
* Mean Temperature
* Range of Temperature
* Sunshine hours
* Humidity and temperature at 9am
* Humidity and temperature at 3pm

Weather factors that we have chosen to exclude are maximum wind speed and wind direction as they are not a good indication of what happened that day. A strong gust of wind or if the wind is south does not give us any information on the weather that day.

Price demand factors we wish to predict:

* Mean demand
* Range of demand
* Range of demand at 9am (between 8am to 10am)
* Range of demand at 3pm (between 2pm to 4 pm)

Since data is grouped by day for the Weather in Brisbane, we will aggregate the values for the price demand data to be grouped by day. Although Brisbane was chosen from the overall high number of price surges, it has been excluded from a factor we would like to predict because we are grouping our variables by day and most days don't have any price surges. We have also decided to specifically look at the data at 9am and 3pm as these are significant times in our daily routine.

***Pre-processing and wrangling***

To be able to use our data, we must first clean and then aggregate our data into the desired values. For our weather data we dropped the columns we did not need and changed the format of the day to be in datetime. We also renamed our columns for better readability and for our price demand data, we removed all the other regions and split the settlement date into two parts datetime and time of day. This is so that we can take the time of day and pick our specific times across all days.

To figure out our derived data, we calculated the range of the temperature and the mean temperature for each day. Maximum and minimum temperature are not really a good indication of what happened that day. Humidity and temperature at 9am and 3pm have been chosen as data at specific times are very interesting. These variables were chosen over cloud amount, wind direction, MSL pressure and wind speed as they don’t give us a good indication of what the weather was like and do not vary a lot. For our price demand data, the data was grouped by day calculating the average and range of demand and the range of demand at 9am and 3pm (+- 1 hour). This is because we wanted to see the level of variability during this time frame. Regex was performed as the time was given in the form of hour:minutes. All our aggregated values were saved as a new csv file each to 2 decimal places for readability.

Finally the 2 aggregated tables were then merged using an inner join and used to plot different graphs. They were joined on the ‘date’ column of each dataset as they were both converted into datetime format. An inner join was performed as it will match up the dates correctly and remove dates that were not in both tables (deleting missing values).

***Analysis Methods***

***Methods***

Our aim was to develop a relationship between various explanatory factors to develop an understanding of the changes in price demand. Since all our data is given numerically, clustering or classification techniques would be an oversimplification and also inappropriate. As our data is continuous (links with time) we have decided to link our variables visually by using scatter plots and heat maps of correlation. The strength and direction of our scatter plots were assessed via a regression analysis.

***Train test split***

Instead of using a conventional train test split, which introduces the possibility of unlucky splits, we used a k-fold testing linear regression model. This model splits the data into 5 folds and then trains and tests the model on the various folds and averages the results. This leads to our model not being overfitted to a particular split.

***Preliminary Analysis***

An incentive for our research question sprouted from the visual similarities from simply graphing the change in mean temperature over time and the change in mean demand over time (figures 1.1 and 1.2). Simultaneously looking at these graphs we can see that they have a congruence in seasonality. At the start of the year the temperature and demand are at their first highest peak which steadily decreased to the lowest trough at approximately June for temperature and September for demand. As these months differ it may show the degree that weather can be used to explain demand. Both graphs then steadily increase towards the end of the year, reaching their second higher peak again at the start of the year. This is very significant because there is a very high fluctuation in the mean demand at its second peak in comparison to the mean temperature, this implies that there could be other unseen factors that had a greater impact during this time.

Looking at our first factor in figures 2.1 and 2.2 we can see that there is no relationship between rainfall and mean or range of demand. The majority of days have 0 rainfall which means that a lot of our demand values are based on days that have 0 rain; having 0 rain is not an indication of how the weather was on that day. Even on days that have a lot of rain the mean and range of demand tend to be average. Although rainfall does not seem to be an explanatory variable in Brisbane, it does not mean this is the same for every city; if our research was extended to other cities ie Darwin the results may differ.

Moving on to figures 3.3, 3.3.1 and 3.4 where mean temperature was used as the explanatory variable, we can see that it shares a quartic relationship with the mean demand, no relationship with range of demand. Having no relationship with the range of demand is logical as the average temperature of the day should have no indication of how much the demand changes. Interestingly a quartic relationship is shared between mean temperature and mean demand which means that at extreme lows and highs of the mean temperature the demand is at its highest and at the median temperature the demand is at its lowest (turning point). An argument for this is that whether the temperature is hot or cold we end up using more electricity in order to counter this extreme weather. We tend to use heaters and aircons more when the weather isn’t comfortable; when the weather is comfortable (approximately 20 to 25 degrees) we don’t use them at all.

Studying figures 4.1 and 4.2 we can examine the depth of using the variability of temperature as an explanatory variable for the changes in demand. In figure 4.1 there is a negative correlation however in figure 4.2 there is a positive correlation; this shows that the higher the variability in temperature the lower the mean demand but the higher the variability in demand. One sensible implication we can make is that on days the temperature is more volatile people tend to quickly change the level of electricity causing a joint high volatility in the change in demand. The opposite trend is seen for mean demand, and this may be as the range in temperature increases, it we know it generally means that the weather became cold (because our data is relating to Brisbane) and a low mean demand for electricity may show that the consumer may have a preference for colder weather. This exact same pattern is seen using sunshine hours as the explanatory variable as seen in figures 5.1 and 5.2. Although we can make an explanation as to why sunshine affects electricity use, it is more since sunshine itself could be the explanatory variable for the range in temperature.

Chart

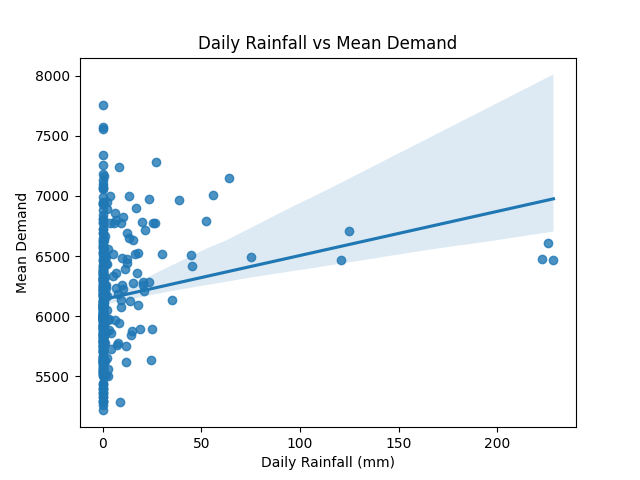
Description automatically generatedChart, line chart

Description automatically generated

*Figure 1.1: Mean Temperature plotted over time (left)*

*Figure 1.2: Mean Demand plotted over time (right)*

Chart, scatter chart

Description automatically generated

*Figure 2.1: Rainfall feature plotted against mean demand (left)*

*Figure 2.2: Rainfall feature plotted against range of demand (right)*

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 3.1: Mean Temperature feature plotted against mean demand(left)*

*Figure 3.1.1: Linearized Mean Temperature feature plotted against mean demand (right)*

Chart, scatter chart

Description automatically generated

*Figure 3.2: Mean Daily Temperature plotted against range of demand*

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

*Figure 4.1: Range of Daily Temperature feature plotted against mean demand(left)*

*Figure 4.2: Range of Daily Temperature feature plotted against mean demand (right)*

*Figure 5.1: Hours of Sunshine feature plotted against mean demand(left)*

*Figure 5.2: Hours of Sunshine feature plotted against mean demand (right)*

***Modelling***

In figures 6.1 there is a heatmap of the Pearson correlation matrix between all explanatory factors is plotted and in 6.2 there is a heatmap of the Pearson correlation matrix between the explanatory factors and the target variables to assist in feature selection to train our model. In table 6.2 we can see all the explanatory variables that we have retained for both models have a correlation of at least 20% with the target variable. For mean demand we are training against quadrupled mean temperature (linearized mean temperature), range temperature and sunlight hours. For the range of demand, we are training against the range of temperature and sunlight hours.

After doing our preliminary analysis, our model was trained against 4 explanatory variables which were the daily (linearized) mean temperature, range temperature, sunshine hours, and rainfall to predict the mean and range of demand. The model was trained using k-folds methods explained earlier on. To validate our model, we used k-folds validation methods to derive the RMSE (root mean squared error) and the r2 score of the predictions of the model. We also calculated the residuals and plotted them against the predicted values and against each individual feature (see appendix) to perform a residual analysis.

***Discussion***

In figure 6.2 our highest correlation for mean demand was quadrupled mean temperature with a Person correlation of 0.69. For range of demand the highest correlation was 0.52 and 0.57 explained by range of temperature and sunshine hours that day. Rainfall in response variables had the lowest Pearson correlation both below 0.2 and due to this weak relationship is probably not an explanatory variable.

Our table 8 suggests 49.5% of the change in daily mean demand is explained by the change in [] and 33% of the change in the range of demand is explained by []. Our RMSE value of 0.125 and 0.1363 shows that our 5-fold cross validation is quite accurate to the prediction. Overall, there is a correlation between weather factors and the change in price demand however the relationship is only moderate. Our residual graphs display a constant variance with no clear patterns proving a linear regression was appropriate for our model.

From figures 7.1 and 7.2 we can see that at 9am temperature is a much better variable for the of change in demand range at this hour, on the other hand at 3pm we have the exact antithesis where humidity is a better explanation for the change in demand range at this hour. In the morning, everyone begins their day and thus their electricity usage is based on the weather outside as they plan their morning. In the afternoon everyone is relaxed and at home and thus are not fussed about the temperature outside and maybe perhaps the humidity inside. To conclude in the morning approximately 60% of the change in electricity demand is explained by the temperature at 9am and in the afternoon 50% of the change in electricity demand is explained by the change in humidity.

***Chart, treemap chart

Description automatically generated***

*Figure 6.1: Heat map of explanatory variables and their correlation with each other*

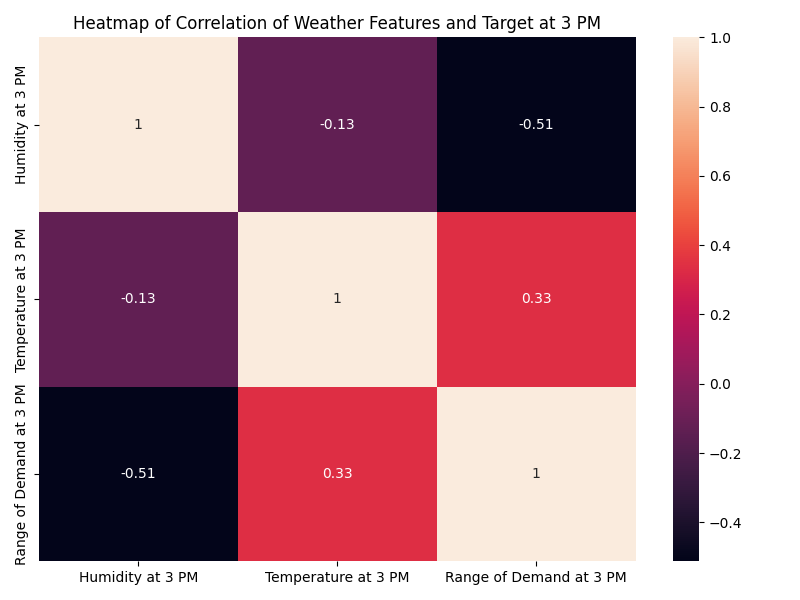
Chart, funnel chart

Description automatically generated

*Figure 6.2: Heat map of explanatory variables and their correlation with response variables*

*Table 8: 5 fold cross validation r2 and rmse for each response variable*

|  |  |  |
| --- | --- | --- |
| Target | 5-fold CV r2 | 5-fold CV rmse (scaled) |
| demand\_mean | 0.4951 | 0.125 |
| demand\_range | 0.3318 | 0.1363 |

 Chart, treemap chart

Description automatically generated

*Figure 7.1: Weather features at 9am heatmap(left)*

*Figure 7.2: Weather features at 3pm heatmap (right)*

***Evaluation***

Our report may not have reached our optimistic goal that a higher proportion of the change in demand is explained by weather factors however in the real-world situations a person’s daily usage of electricity is not only limited by weather factors. From figure 7.1, 7.2, 1.1 we can see that at different times of day and time of year has a different impact to the demand and therefore maybe unseen secondary factors heavily effect our usage of electricity. A contemporary example is the current pandemic going on where on days of look down people use more electricity opposed to when there isn’t one.

Accessing the accuracy of our report we can see that from figure 6.1 shows that sunlight hours and range of temperature are collinear, which might make our model unreliable. We also used data from January to March twice from 2 different years which may be what caused denser clusters of certain datapoints and skewed our linear regression. To better our experiment next time, adding more weather factors, not having missing variables and using the same number of days in each month will improve the accuracy and precision of our report.

To further extend our studies we can look at other cities across Australia and possibly develop a correlation between cities, their geographical location and their history of electricity demand.

***Appendix***

Chart, scatter chart

Description automatically generated

*Figure 9: Linear regression analysis graphs and residual plots for response variables.*