COMP20008 - Assignment 2 Final Report

Contents

Aim…….3

Datasets……..3

Pre-processing and Wrangling……4

Analysis Methods……...5

Methods……..5

Training-test Split……..5

Preliminary Analysis…….5

Modelling…….20

Discussion……20

Evaluation……22

References……23

Aim

This project aims to examine the relationship between several weather factors and energy demand in Australia, particularly Melbourne city. The result may be beneficial for energy retailers in Australia as it provides valuable information. In addition, it will also be helpful to people working in the agriculture industry and to traders within the commodity market. Businesses can use this information to have a better understanding of consumer behaviours in regard to seasonal energy consumption. Acquiring this knowledge helps to expand the understanding of weather forecasts with the total demand, which ultimately could improve liveability in Melbourne. Therefore, we formulated the research question as follows:

> *How do changes in different weather factors impact the total demand in Victoria?*

Datasets

The weather datasets were sourced from the Bureau Meteorology website consists of .csv files of weathers from four major cities in Australia, such as Melbourne, Sydney, Brisbane, and Adelaide measured from February 2021 until March 2022. The weather data of the cities is sourced from the Bureau Meteorology website, which measures the minimum temperature (°C), maximum temperature (°C), rainfall (mm), evaporation (mm), sunshine (hours), the direction of a maximum wind gust, speed of maximum wind gust (km/h), time of maximum wind gust, 9 am temperature (°C), 9 am relative humidity(%), 9 am cloud amount (oktas), 9 am wind direction, 9 am wind speed (km/h), 9 am MSL pressure (hPa),3 pm temperature (°C), 3 pm relative humidity (%), 3 pm cloud amount (oktas), 3 pm wind direction, 3pm wind speed (km/h) and 3 pm MSL pressure (hPa). One of the quality issues we noticed in the dataset was on the evaporation (mm) column as it majorly consists of missing values so the analysis will not be accurate on that particular attribute. Hence, we decided to exclude it.

The energy demand data was in .csv as well and it is sourced from Victoria and New South Wales, Queensland and South Australia with four key variables, region, settlement data, total demand and price surge.It is collected from February 18th 2021, until March 18th 2022.

The energy demand and price surges on 9 am and 3 pm are the target variables for our investigation.

Pre-processing and Wrangling

Before cleaning the data, we need to restructure it to ensure inconsistency in each column. For instance, in the Brisbane weather dataset, the word “Calm” appears in the integer column. Hence, we parsed each column that does not have an appropriate format to an appropriate replacement. This makes manipulating data more accessible and makes the prediction more accurate for our analysis.

Imputation is needed since the weather dataset contains missing values. We choose to use KNN imputation rather than mode, mean or median because weather on each consecutive day is more likely to be similar or close enough to each other, making prediction more accurate. As the data comes in different formats we needed to normalise the weather factors columns to be able to find the relationship between them and our target variables (total demand and price surge).

To merge the datasets, we used the date, region, and time as the keys and we utilised Pandas data frame and regex expressions in doing so. First, we split the settlement date column in the price demand dataset into two-column dates and times. We also formalised the date format to YYYY-MM-DD and added a region column to each weather dataset. Since our target variable (the total demand) differs dramatically depending on the season, we found that splitting the

weather data into four seasons is a more reasonable approach. In doing so, we could enhance our analysis which ultimately helps our intended audience in making decisions.

Analysis Methods

Methods

Our investigation goal is to find the relationship between continuous variables which are the total demand and different weather patterns so we took that into consideration while choosing the analytical methods. Using feature selection to investigate which weather features are more related to total demand is critical for our research question. Therefore, we used Pearson correlation. Furthermore, we used linear regression models as a supervised method since it allows us to inspect which weather patterns are related to total demand. We choose this technique instead of other methods such as clustering and classification since it helps our research question more. These methods will assign records to classes and clusters which is not suitable for our target variable (total demand) since it is a continuous variable.

Training-test Split

Splitting is an essential step as it enables us to evaluate the fit of our model and improve our analysis. We split the datasets into two parts: training and test sets. The training set was used for the preliminary analysis, while the test set was used to depict the generalisations of the model. The method used is k-fold cross-validation, as it assesses the model without a bias and reduces the chances of getting an unlucky split which could hinder the generalisation of the model.

Preliminary Analysis

We create a scatter plot to understand the relationship of weather features as determinants of total demand preliminary. We also deducted the outliers as can be seen in the scatter plot. We calculated the respective Pearson correlation for the data with and without outliers and while there were changes in the coefficients it was not significant so we decided to include the outliers. We plotted the Pearson correlation coefficient on a heat map so that we could get an insight into the changes in the 9 am and 3 pm values.

From figure 2 and 4, we can see that there is a linear relationship between some of the weather factors (temperature, MSL pressure, cloud amount, relative humidity, rainfall, wind speed, sunshine) and the total energy demand. The linear relationship between the variables demonstrates that Pearson correlation is the appropriate method to further investigate their relationship. Looking at the heatmaps of each season, we can see that some weather patterns have a strong relationship while others have a weak one. This suggests that some weather patterns have a different relationship with total demand. For example, maximum temperature has a high negative correlation in the winter and a positive one in the winter. Also, total demand between 9 am and 3 pm in almost all seasons has little fluctuations. Some of the weather features had a high correlation, but we excluded them since it does not satisfy the assumptions of Pearson correlation in our analysis. This is evident in the scatter plot of the feature.

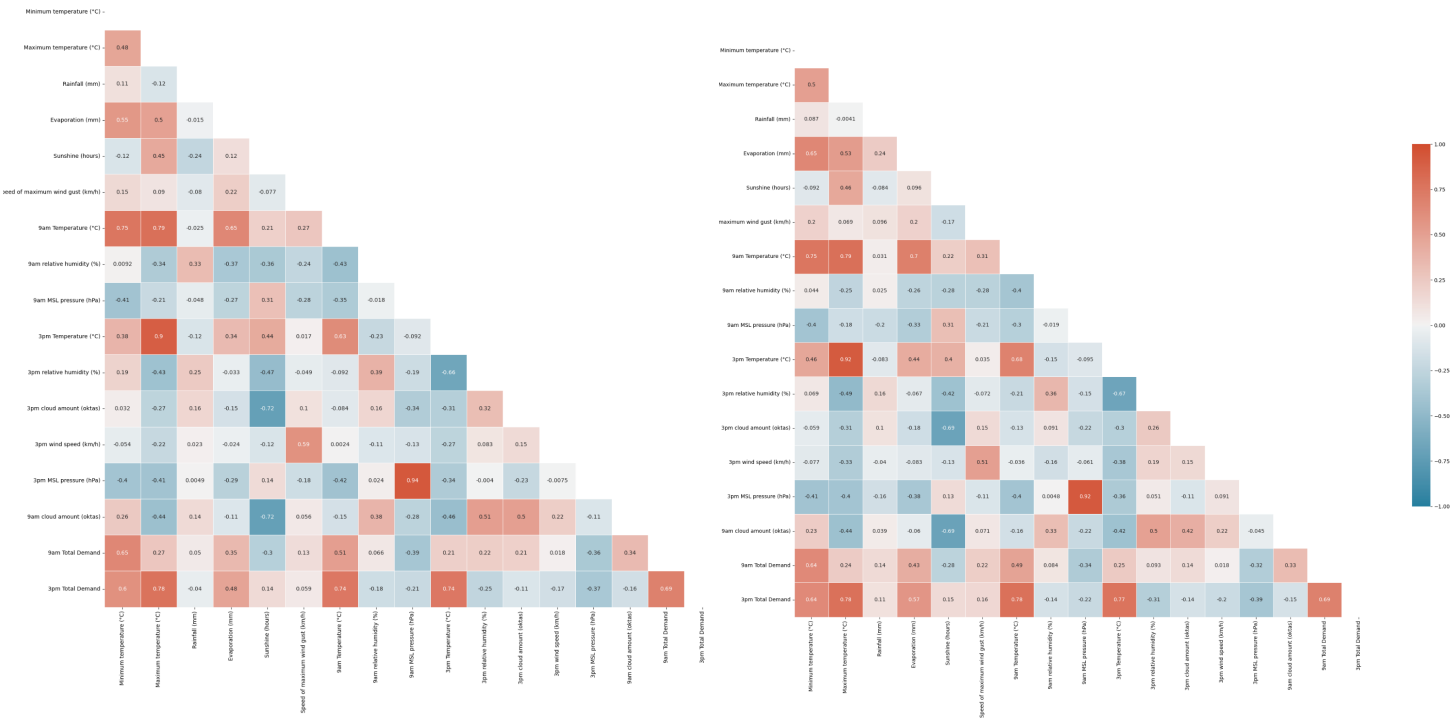


Figure 1: Pearson correlation coefficients of all variables during Summer

With outliers (left). Outliers removed (right)

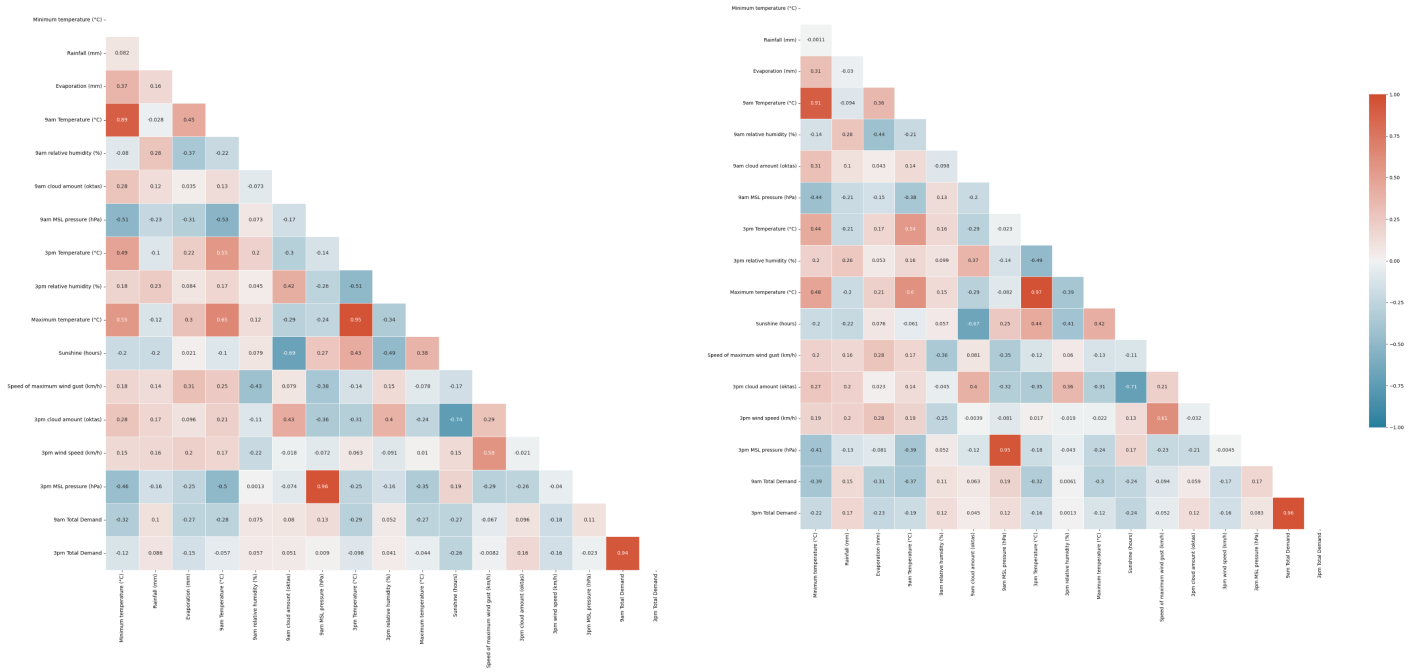


Figure 2: Pearson correlation coefficients of all variables during Autumn

With outliers (left). Outliers removed (right)

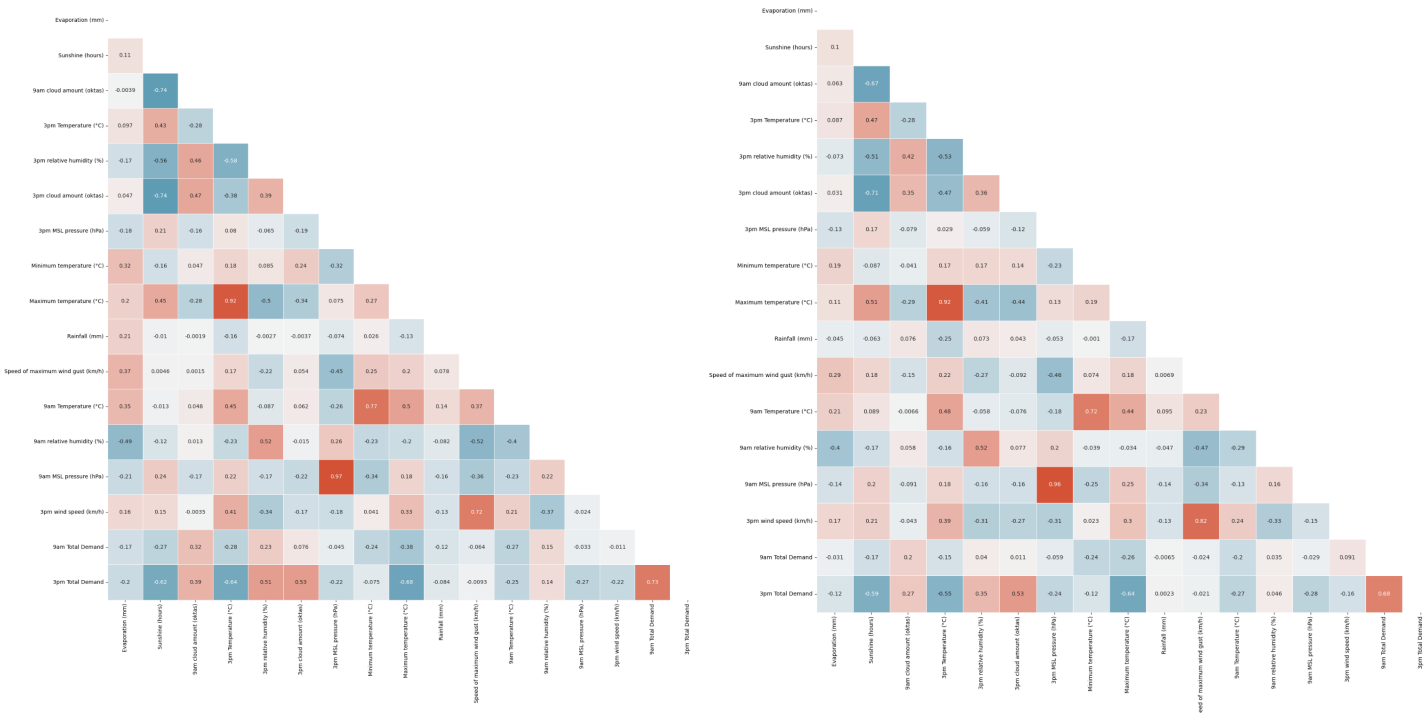


Figure 3: Pearson correlation coefficients of all variables during Winter

With outliers (left). Outliers removed (right)

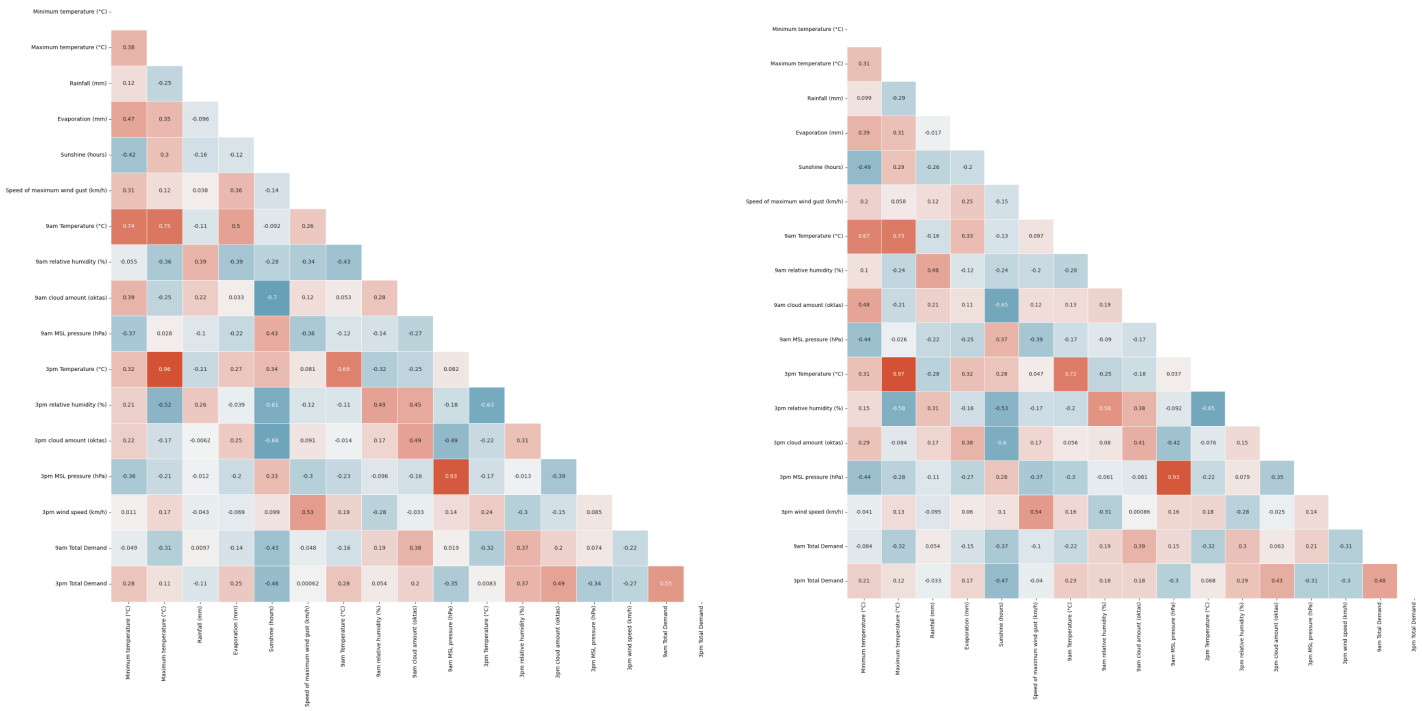


Figure 4: Pearson correlation coefficients of all variables during Spring

With outliers (left). Outliers removed (right)

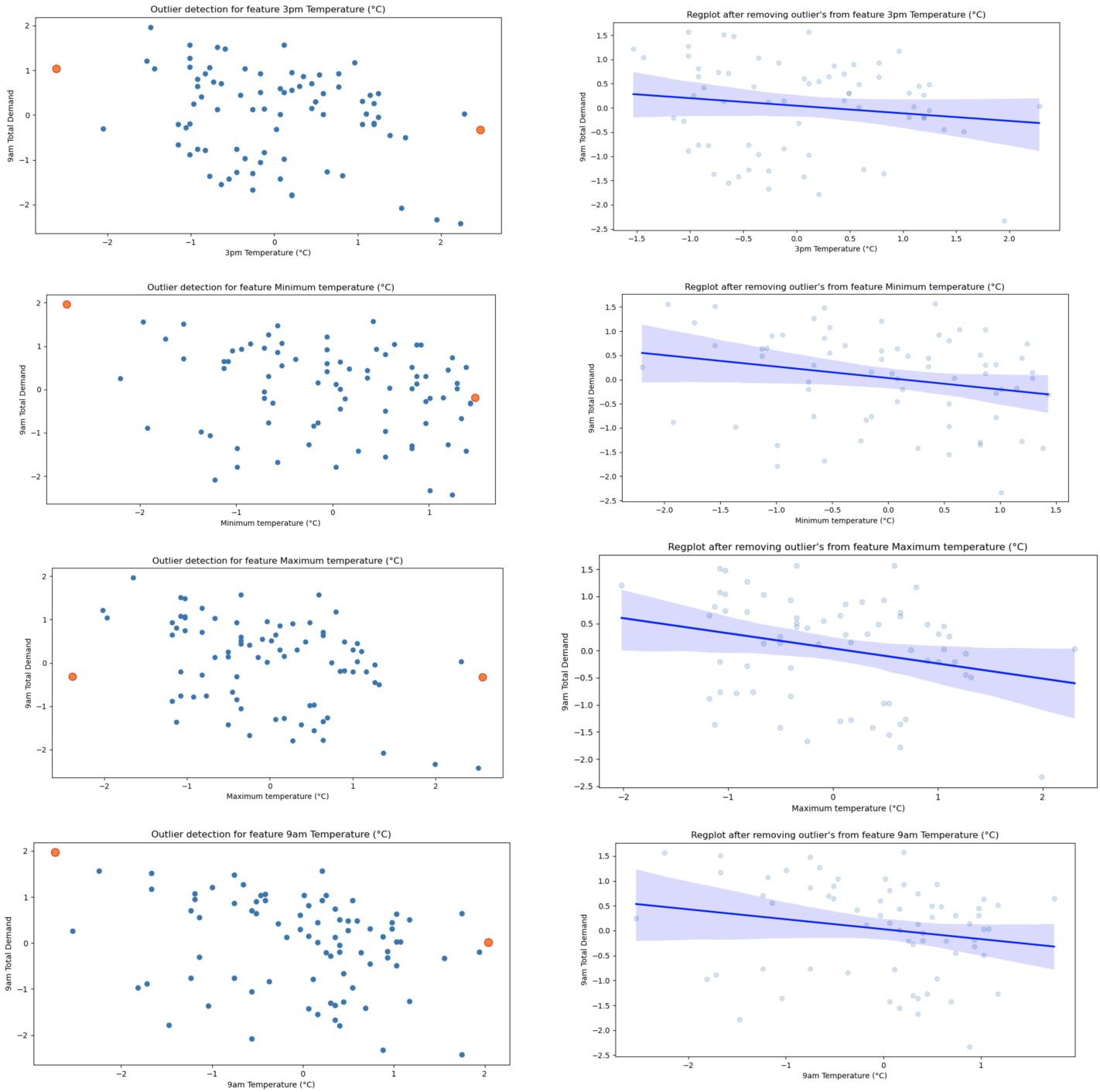


Figure 5: 9am Total Demand against Temperature features during Winter

With outliers (left). Outliers removed (right)

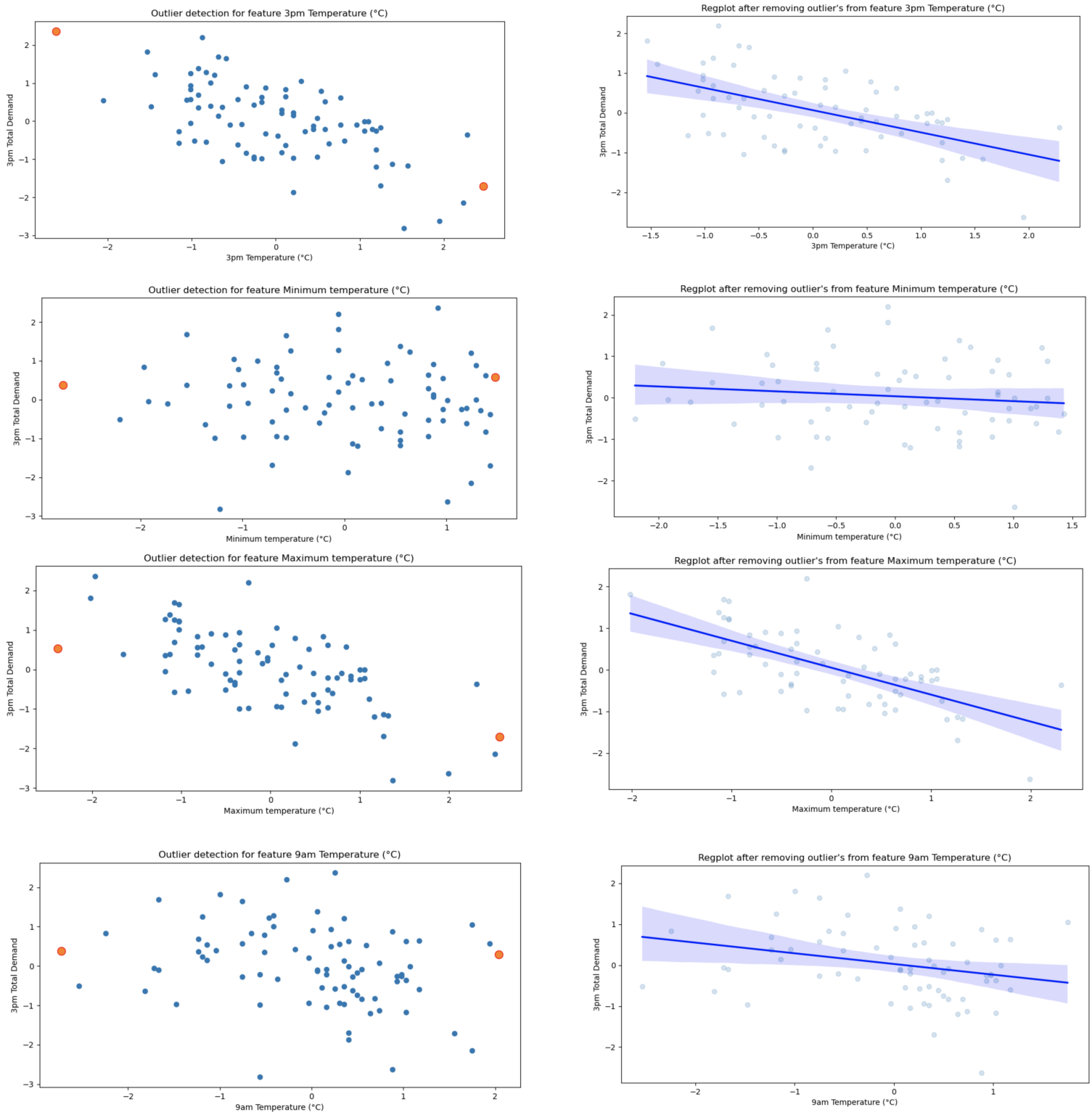


Figure 6: 3pm Total Demand against Temperature features during Winter

With outliers (left). Outliers removed (right)

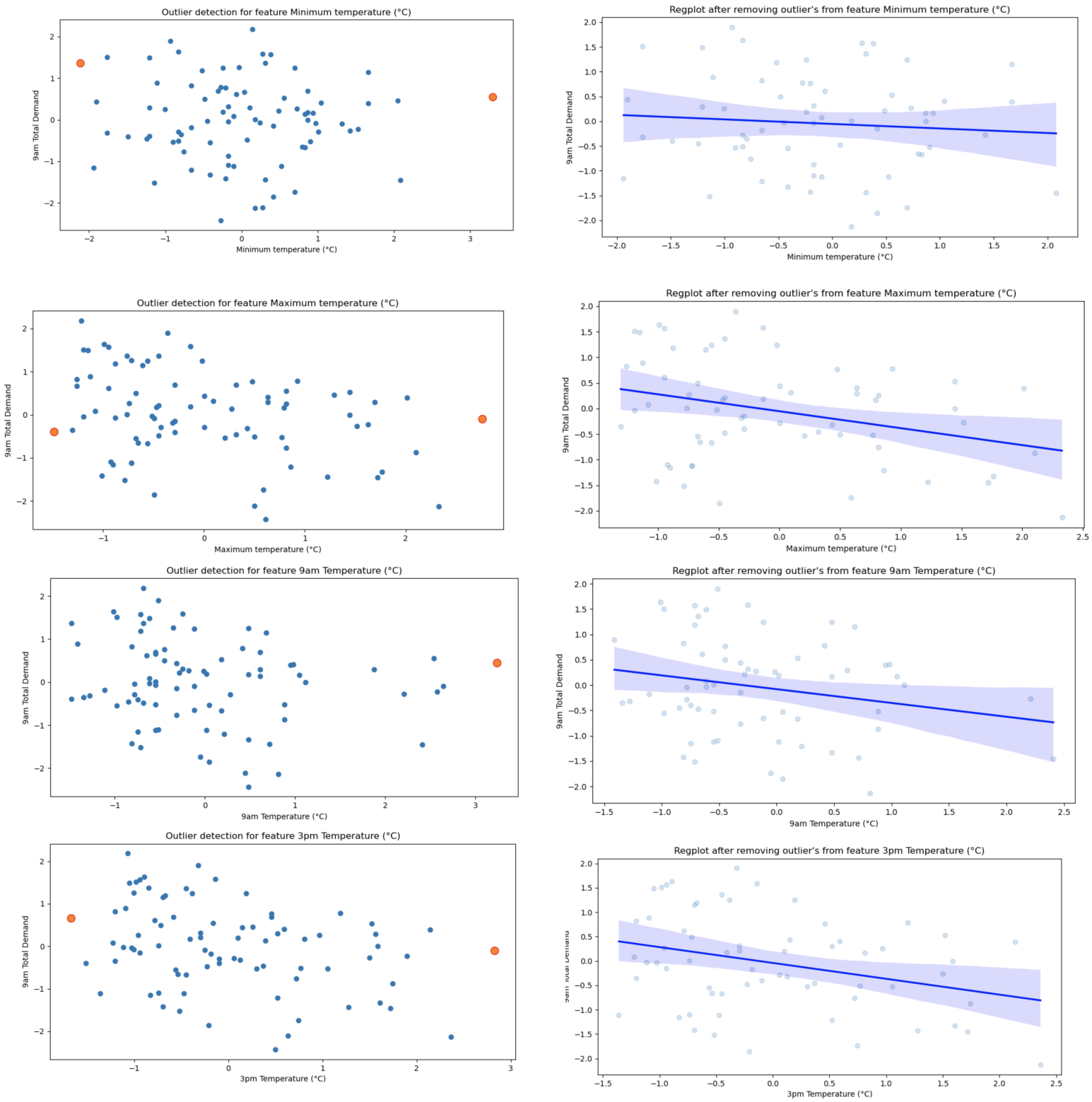


Figure 7: 9am Total Demand against Temperature features during Spring

With outliers (left). Outliers removed (right)

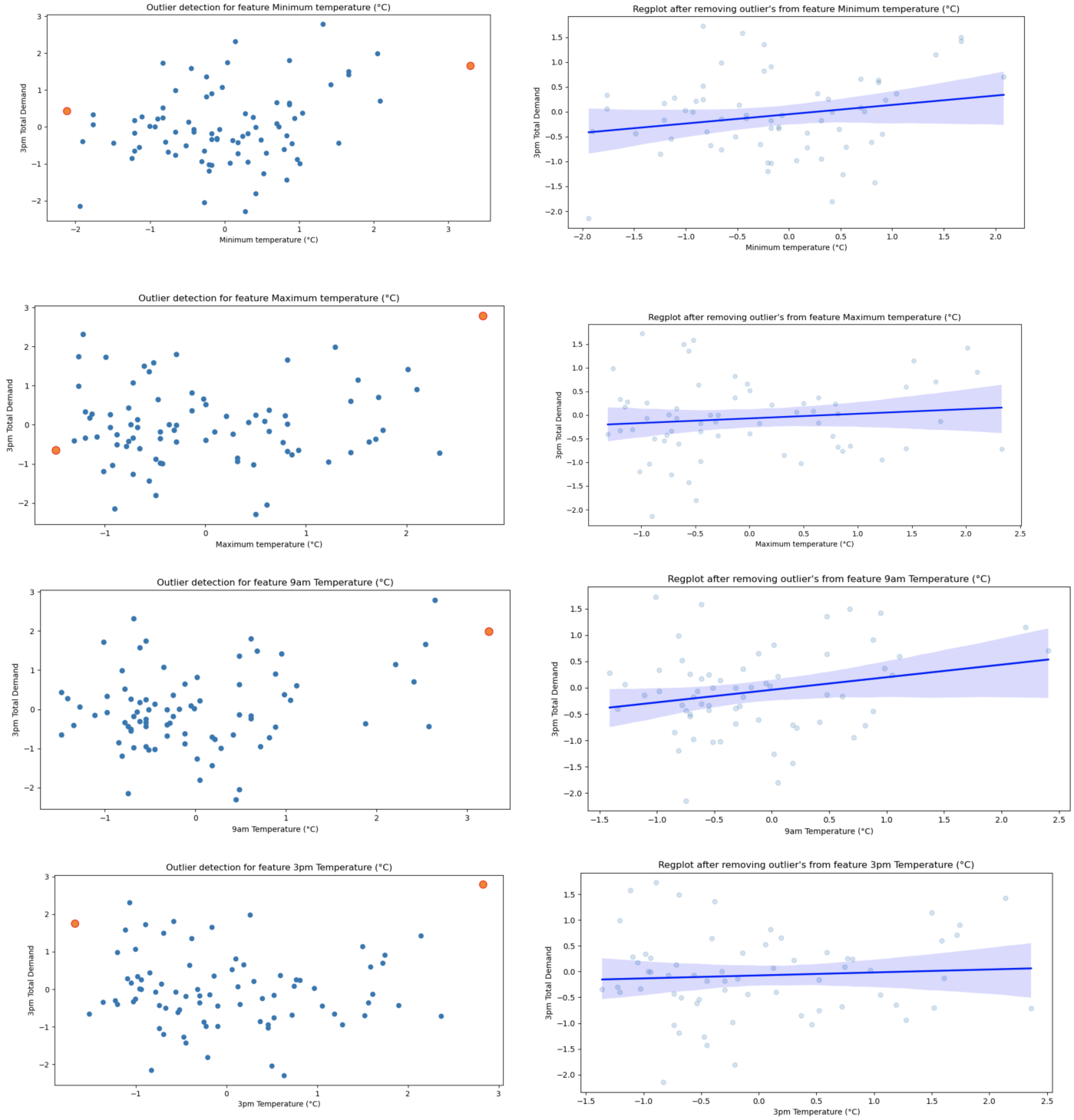


Figure 8: 3pm Total Demand against Temperature features during Spring

With outliers (left). Outliers removed (right)

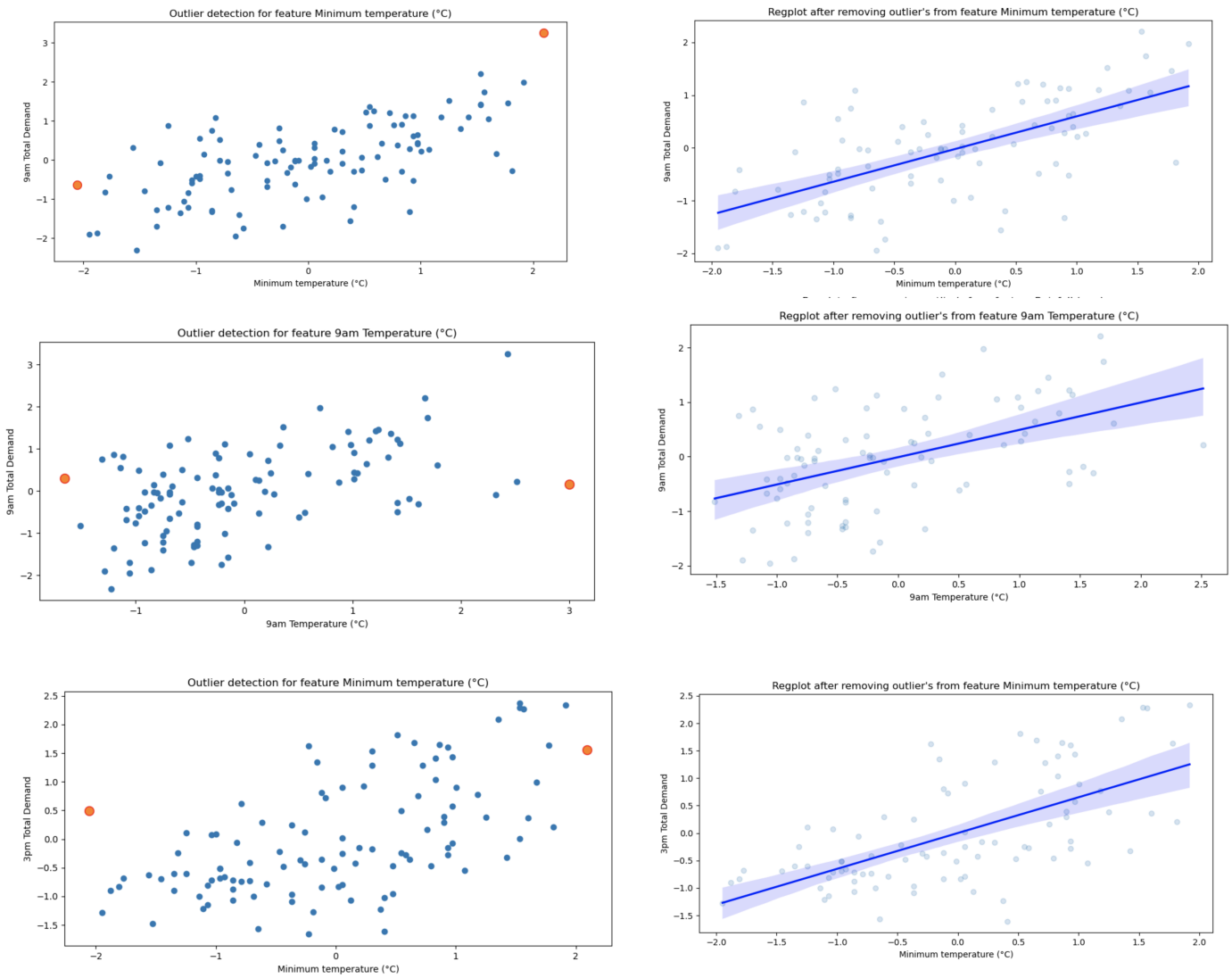


Figure 9: 9am Total Demand against Temperature features during Summer

With outliers (left). Outliers removed (right)

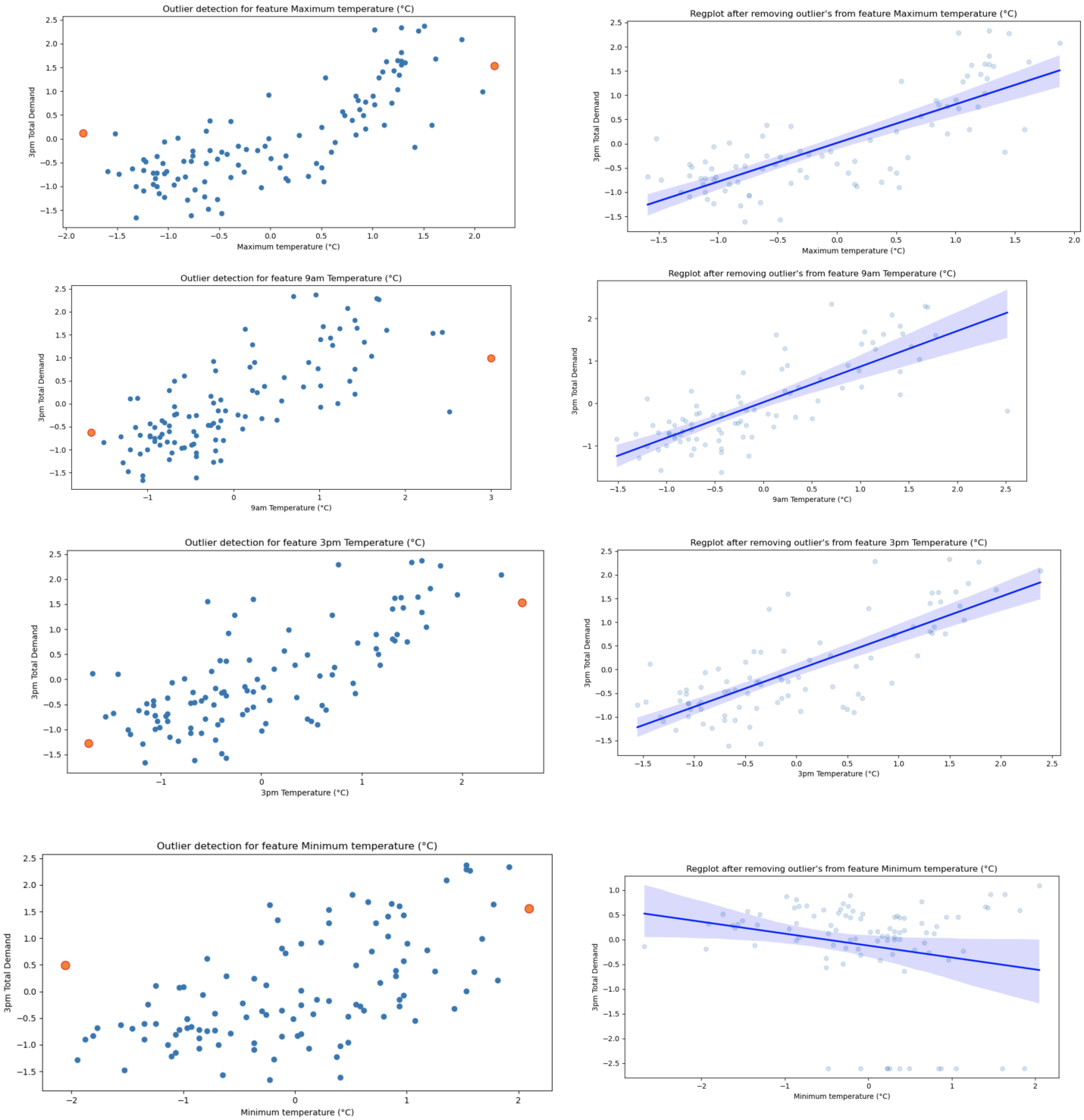
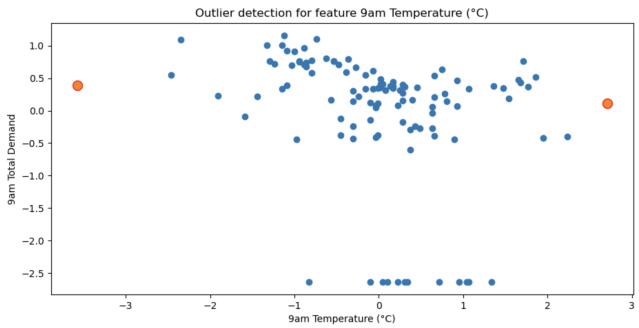
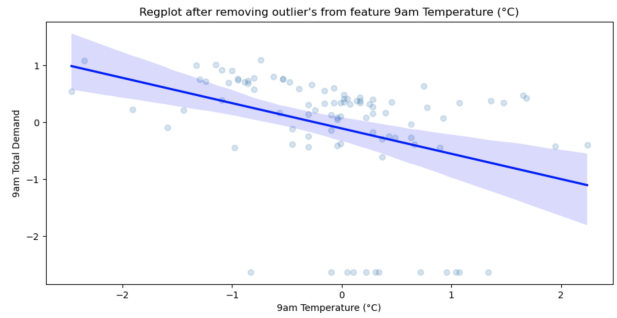
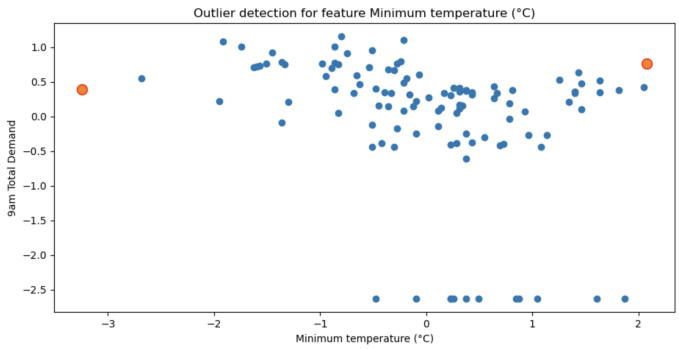


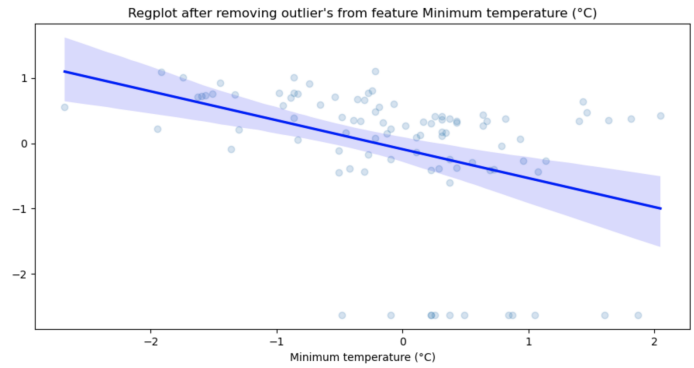
Figure 10: 3pm Total Demand against Temperature features during Summer

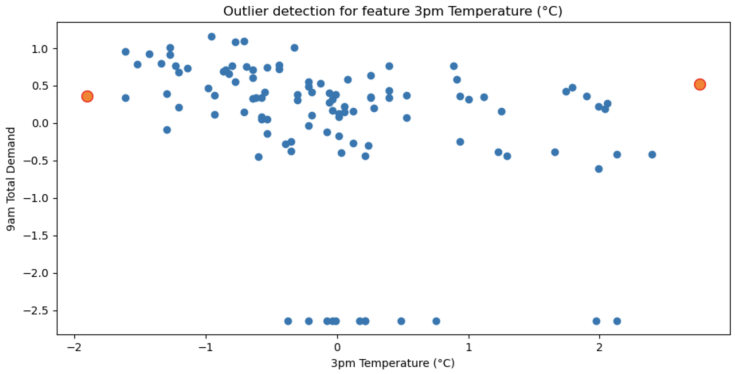
With outliers (left). Outliers removed (right)

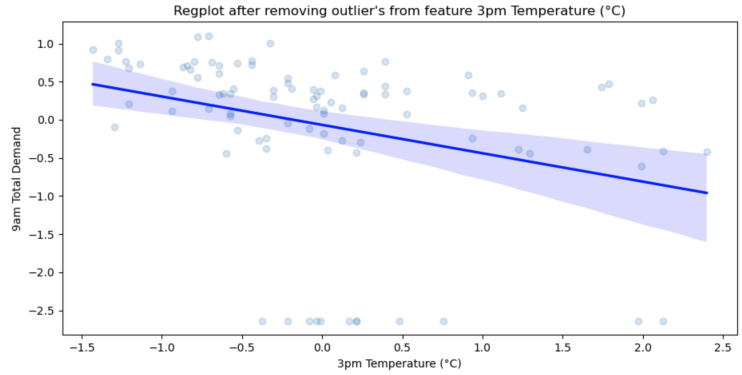

Outlier detection for feature 9am Temperature (°C)
9am Total Demand
9am Temperature (°C)

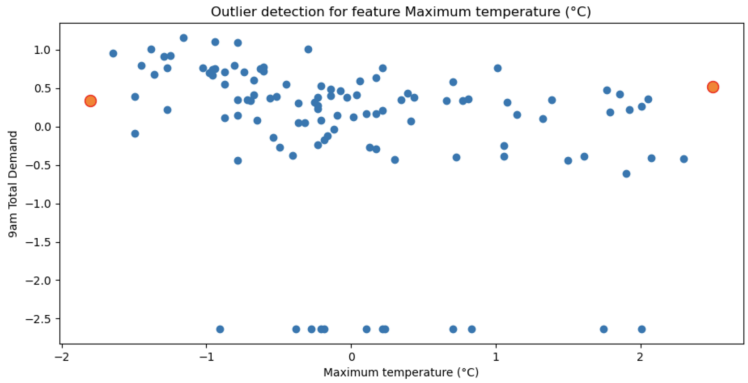

Regplot after removing outlier's from feature 9am Temperature (°C)
9am Total Demand
9am Temperature (°C)


Outlier detection for feature Minimum temperature (°C)
9am Total Demand
Minimum temperature (°C)


Regplot after removing outlier's from feature Minimum temperature (°C)
Minimum temperature (°C)


Outlier detection for feature 3pm Temperature (°C)
9am Total Demand
3pm Temperature (°C)


Regplot after removing outlier's from feature 3pm Temperature (°C)
9am Total Demand
3pm Temperature (°C)


Outlier detection for feature Maximum temperature (°C)
9am Total Demand
Maximum temperature (°C)

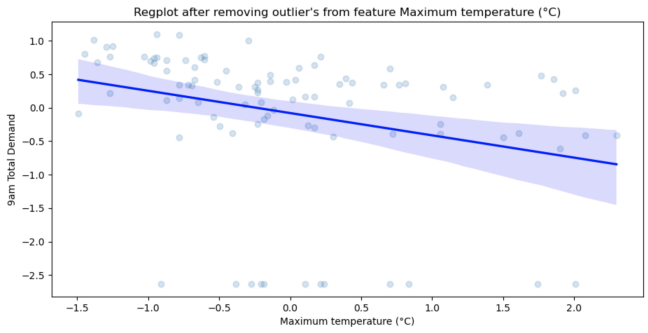

Regplot after removing outlier's from feature Maximum temperature (°C)
9am Total Demand
Maximum temperature (°C)

Figure 11: 3pm Total Demand against Temperature features during Autumn

With outliers (left). Outliers removed (right)

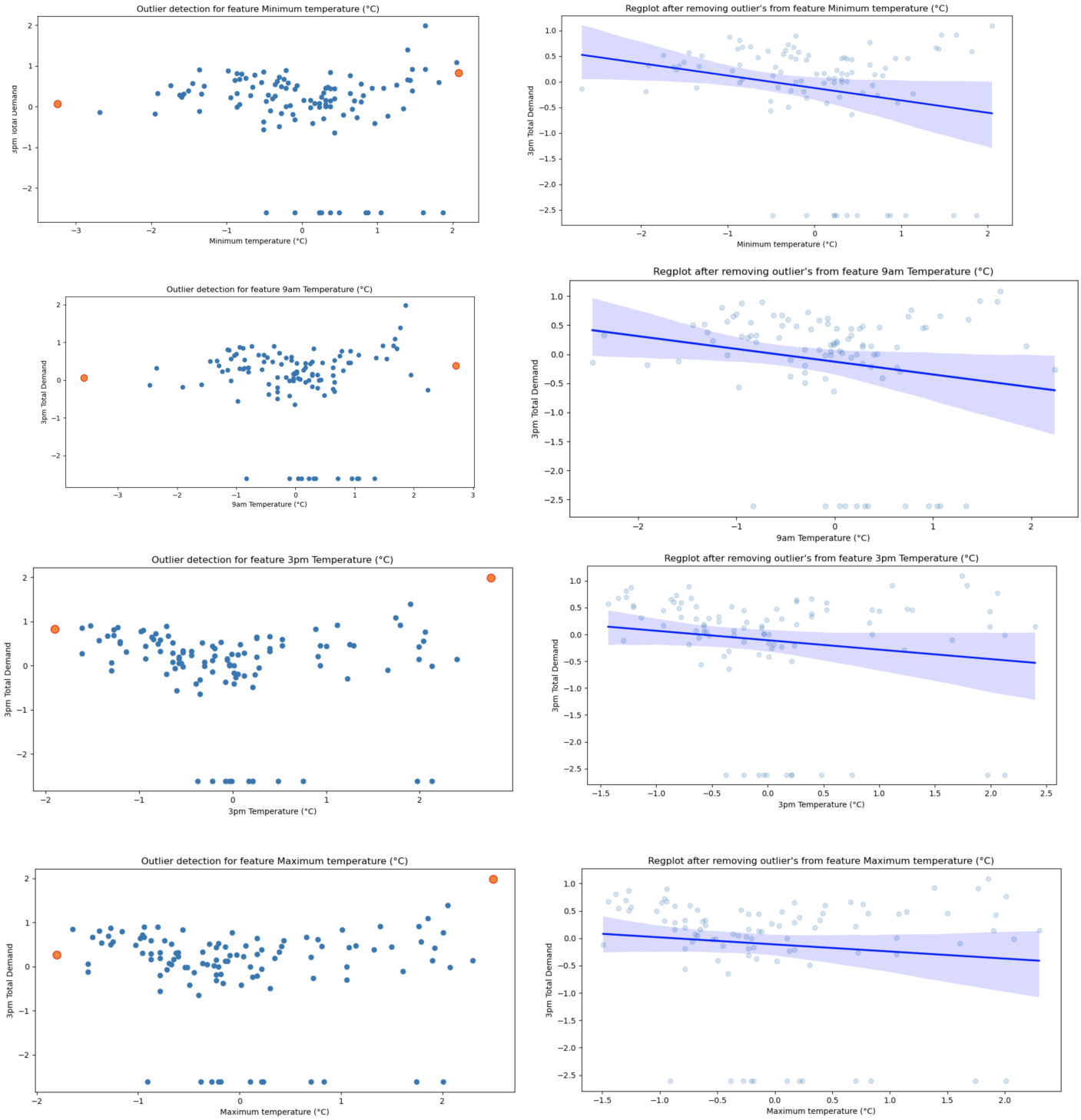


Figure 12: 3pm Total Demand against Temperature features during Spring

With outliers (left). Outliers removed (right)

Modelling

We used regression models to understand the meaningfulness of the relationship between weather features and total demand and account. We used linear regression to fit the model to account for the collinearity between the weather features variables as seen in the heat map above. To improve the reliability of the model we used 10-fold cross validation and averaged the calculations. Furthermore, we calculated the respective R_2 and MSE for the training and test set (see appendix Table1). The regression models assume homoscedasticity which is apparent in the graphs as the residuals are randomly distributed (see appendix). Hence, the assumption of linear regression was satisfied. We fit the model in multiple regression models but we will discuss only the variable with the highest correlation coefficient.

Discussion

In this section, we interpret the models and discuss the significance of the results. From the heatmap, we can see that the 3 pm total demand has a strong relationship (negative correlation) with maximum temperature, sunshine and 3 pm temperature during the winter season. Its correlation coefficients are (-0.64, -0.59 and -0.55) respectively. The three variables also observe a high correlation within each other, indicating that the maximum temperature is when the weather is sunny, and it is at 3 pm. This could suggest that when the weather is warmer during the winter season, people tend to utilise less energy as they do not need it as much, ultimately lowering the total demand for the energy. The 9 am total demand had a small relationship with all the explanatory variables. An worth noting observation is that the total

demand for both the 3 pm and 9 am have almost no relationship with the rainfall, indicating that rainfall does not have a relationship with rain. Furthermore, the 9 am and 3 pm have a high positive correlation, indicating that there is slight fluctuation between the total demand during these hours. We can observe the opposite result in the summer heatmap as the total demand at 9 am and 3 pm has a positive high correlation with the temperature. This also suggests the opposite implication as people tend to utilise energy more during the summer by turning on air conditioning which increases the total demand of energy. Spring season results seem to have the same relationship as the summer but in a modulated manner, similarly to autumn and the winter season. A worth noting observation is that 9 am and 3 pm total demand in the autumn is almost identical, reinforcing the idea of slight fluctuation between these two time slots.

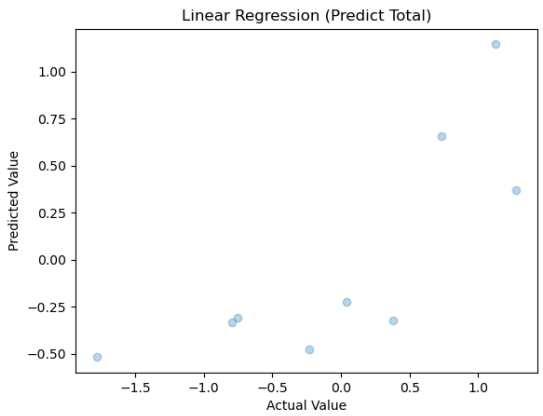
These correlations are reinforced in the fitting of our model, as we observed a meaningful relationship between weather factors and total demand in three seasons: summer, winter. After the 10-fold CV, the mean R2 is positive, meaning that the model is a predictor of total demand. The MSE also is low, meaning that our model is not fitted poorly. Hence, it is reasonable to state that weather patterns (particularly temperature) have a meaningful relationship with total demand. However, the spring and autumn season had the opposite result which means that further investigation is needed. This could be because during these seasons weather tends to fluctuate more. This conclusion is advantageous as it helps retail energy businesses to understand customers consuming behaviour in each season and make useful decisions accordingly.

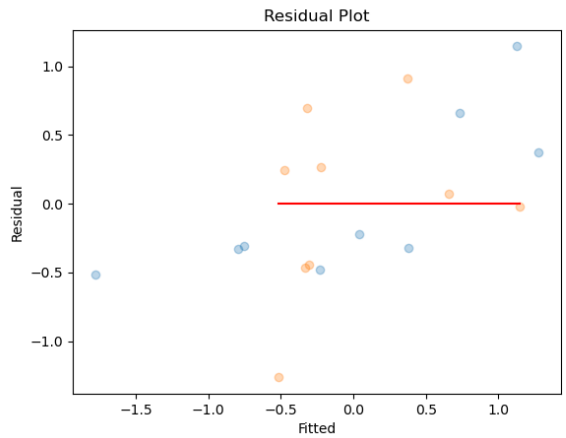
Evaluation

In this part, we look at the limits of our findings and make recommendations for further study. Firstly, the results obtained consider one weather factor against the total demand with little to no regard to the interaction between these factors against the total demand. Therefore, there is a need to investigate the relationship between our explanatory variables more to get a better insight into how that relates to the total demand. Future research may examine that using multiple regression instead of linear. Secondly, since our explanatory variables were mostly given at 9:00 am and 3:00 pm, we analysed that with maximum/minimum against the total demand at 9 am and 3 pm. However, it is vital to explore the relationship in further context to include other times so that it is understood in a well-defined manner. Finally, the price surge and total demand are linked somehow since higher prices result in higher total demand, so examining this relationship and weather patterns is essential. Therefore, in future studies, these factors should be included to provide more accurate results and enhance the understanding of the weather patterns as determinants of total demand.

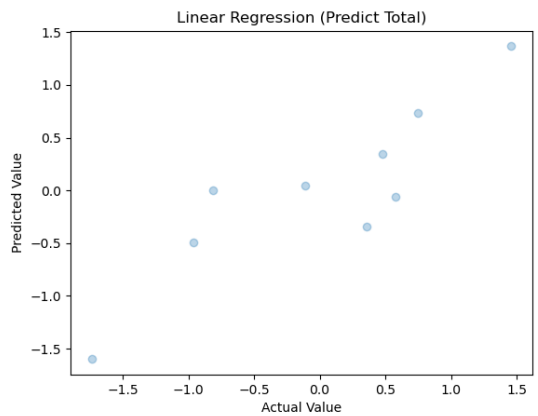
APPENDIX

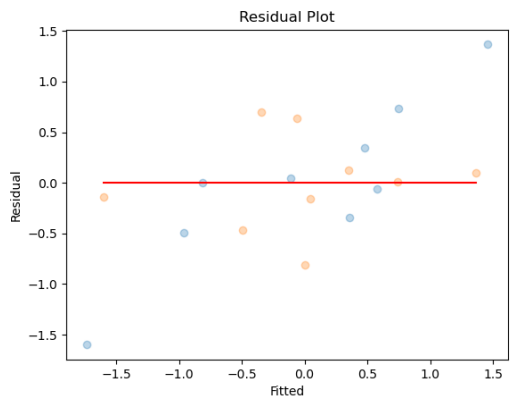
Residual analysis plots 9am (winter)



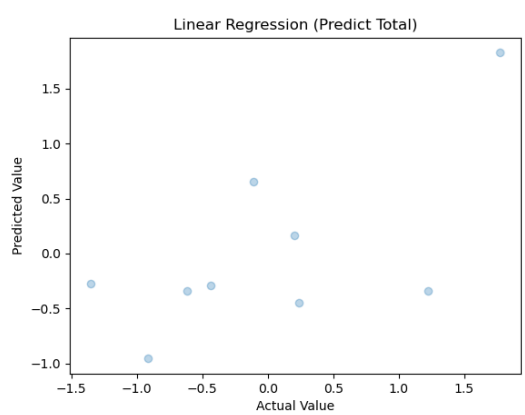


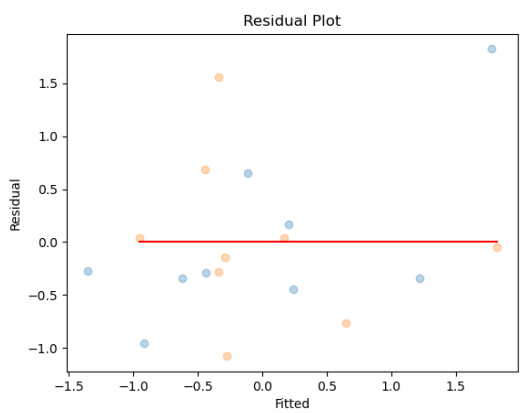
Residual analysis plots 3am (winter)



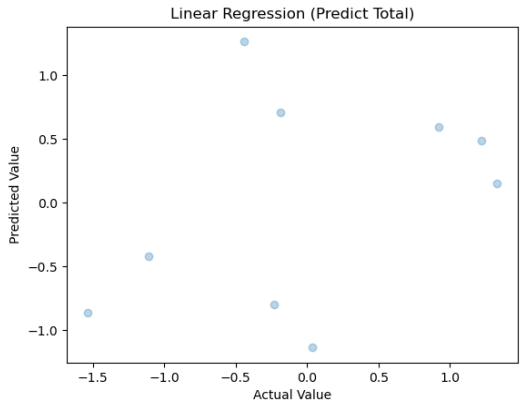


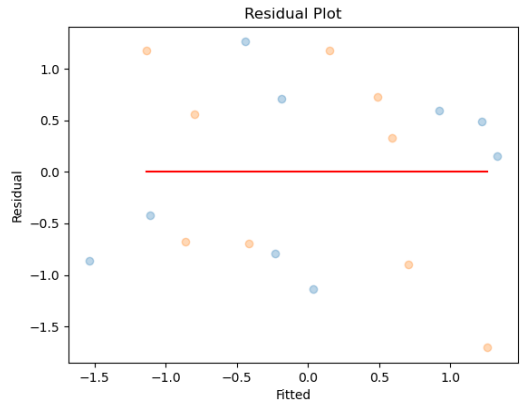
Residual analysis plots 9am (spring)



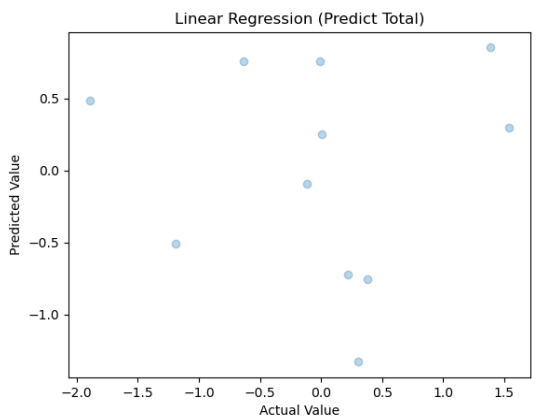


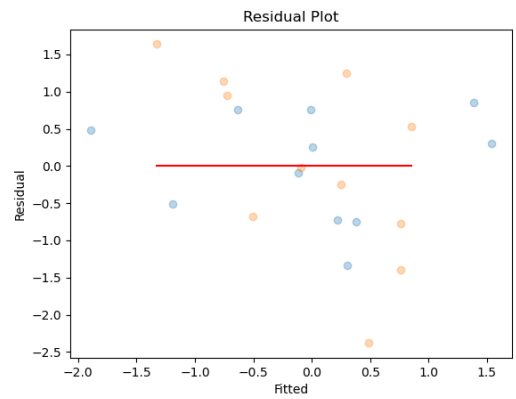
Residual analysis plots 3pm (spring)



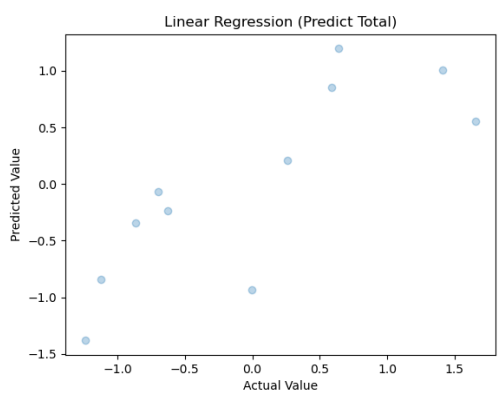


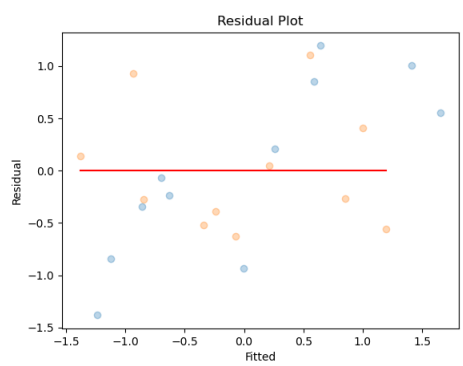
Residual analysis plots 9am (summer)



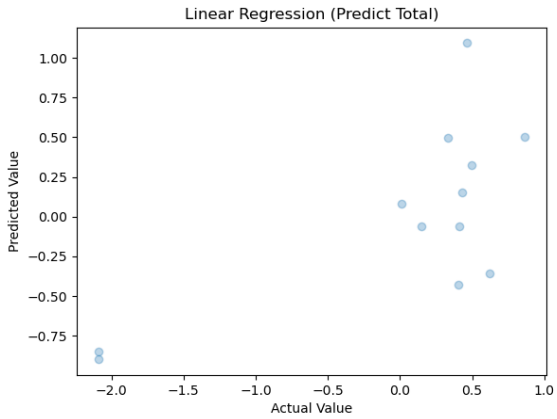


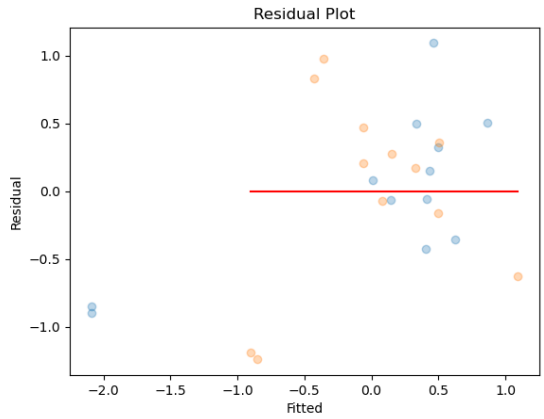
Residual analysis plots 3pm (summer)



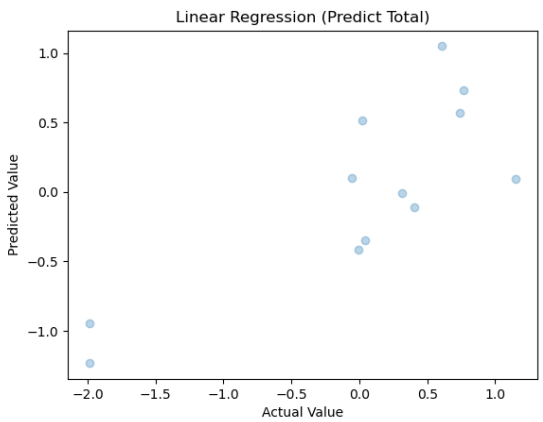


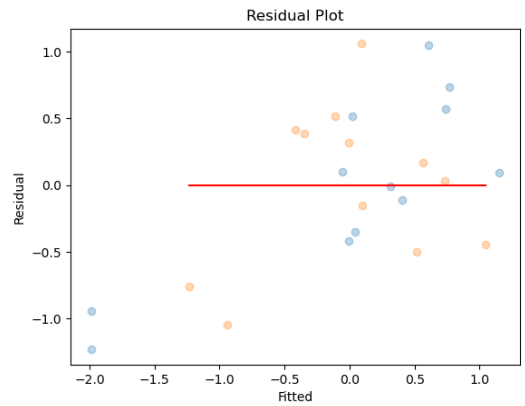
Residual analysis plots 3pm (autumn)





Residual analysis plots 9am (autumn)





3pm total demand vs weather pattern

	Autumn	Winter	Spring	Summer
R2 test set	0.4958306439168726	0.7698121083720781	0.04731965513431047	0.6442157913501607
MSE test set	0.46215524307620015	0.2046114592248194	0.9309508045638313	0.3234401896816721

References

Adelaide (west terrace / ngayirdapira), South Australia daily weather observations. Adelaide (West Terrace / Ngayirdapira), SA - March 2022 - Daily WeatherObservations. (2021). Retrieved from http://www.bom.gov.au/climate/dwo/202203/html/IDCJDW5081.202203.shtml

Aggregated price and Demand Data. AEMO. (2021,February 1). Retrieved from https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/aggregated-data

Brisbane, Queensland Daily weather observations. Brisbane, Qld - March 2022 - Daily Weather Observations. (2021). Retrieved from http://www.bom.gov.au/climate/dwo/202203/html/IDCJDW4019.202203.shtml

Gunjal, S. (2021). *Tutorial: K Fold Cross Validation*. Kaggle.com. Retrieved 19 May 2022, from https://www.kaggle.com/code/satishgunjal/tutorial-k-fold-cross-validation/notebook.

Gunjal, S. (2021). *Advanced Reg Techniques (Linear Models) Top 6%*. Kaggle.com. Retrieved 19 May 2022, from https://www.kaggle.com/code/satishgunjal/advanced-reg-techniques-linear-models-top-6/notebook.

Long-term temperature record Australian climate observations reference network – surface air temperature (ACORN-SAT). Long-term temperature record: Australian Climate Observations Reference Network – Surface Air Temperature. Retrieved May 21, 2022, from http://www.bom.gov.au/climate/data/acorn-sa

Melbourne, Victoria daily weather observations. Melbourne, Vic - March 2022 - Daily Weather Observations. (2021). Retrieved from http://www.bom.gov.au/climate/dwo/202203/html/IDCJDW3050.202203.shtmlt/

Missing Data Imputation Using sklearn. Minkyung's blog. (2020). Retrieved 19 May 2022, from https://mkang32.github.io/python/2020/11/21/Missing-data-imputation-using-sklearn.html.

Pandey, P. (2020). *A Guide to Handling Missing Values in Python*. Kaggle.com. Retrieved 19 May 2022, from https://www.kaggle.com/code/parulpandey/a-guide-to-handling-missing-values-in-python/notebook.

Sydney, New South Wales daily weather observations. Sydney, NSW - March 2022 - Daily Weather Observations. (2021). Retrieved from http://www.bom.gov.au/climate/dwo/202203/html/IDCJDW2124.202203.shtml