Assignment 2: Predicting energy demand

GitHub link: https://github.com/Alphaomegainfinity/energy_consumption_vs_weather/tree/main

Word count: 2487

Group 5 members:
Heather Konzman
Sureshini Mendis
Hien Bui
Halley Ngoc Pham

Introduction:

Energy consumption is one of the most important variables affecting any country's economy. Energy usage varies depending on a number of circumstances, including weather. The goal of this paper is to create a model that anticipates maximum daily energy use and pricing based on historical weather data.

Checking dataset:

The weather dataset contains key weather indicators for the city of Melbourne for each day between November 2022 and April 2023 and extracted from the Australian Energy Market Operator (AEMO).

The other dataset of 'price_demand' contains energy price and demand metrics for the Australian state of Victoria between November 2022 and April 2023.

Firstly, our team imported several libraries to set up the necessary dependencies for data analysis and machine learning tasks.

After importing, we implemented the function “data_checking.” This function is well presented to check for missing values and incorrect value types, returned as a report dataframe.

Data cleaning:

Before importing the datasets, we used Excel to change the format of the “Time of maximum wind gust” column to: HH:MM:SS, adding seconds and creating uniformity to allow for easier processing in Python.

After importing and checking the weather dataset, we removed all entirely empty columns, since they include missing values. We then used the Pandas “dropna()” method with the “thresh" parameter to drop all rows with more than 4 missing values. We chose this instead of imputing this data, because this suggests a problem with collecting data on that day, the existing row values may be inconsistent, and there was only 1 row affected.

Viewing the original weather dataset, we observed many blank string values in the “9am wind direction” column. To properly work with this data, we replaced the blank strings with NaN

values. The regular expression r's*$' matches only whitespace characters in strings. Then, we imputed the NaN values to the most frequent categorical value of the column. We did not drop these rows entirely because it would mean losing 9 rows worth of other weather variable data that was present, and gave us a chance to try out another method.

We then wanted to change the wind directions columns to a numeric type, with the idea of using compass directions. Assistance with the method to do so was found on stack overflow,[1] with verification of the correct degrees from wikipedia.[2]

We also observed the inconsistency of using “calm,” in the “9am wind speed” column; this was replaced with numerical value “0”. The data type of the date column was changed from object type to datetime64 type. The data types of “Time of maximum wind gust” and “9am wind speed” columns were changed from object to numeric. These steps allow for further numerical comparison and analysis with the rest of the feature columns. The lambda function was utilised to apply the necessary step to the column of time values, with zfill adding zeroes where necessary so the time data uses a consistent length of numerals. (Assistance with this method came from Stack Overflow.[3])

For the price & demand dataset, we calculated the sum of the Demand for each day, changed the Date_Time column label to Date only, and changed the data type to datetime for ease of further analysis. We then merged the two datasets with an inner join so that for Dates they shared, the weather features and the total Demand would be included in one dataframe.

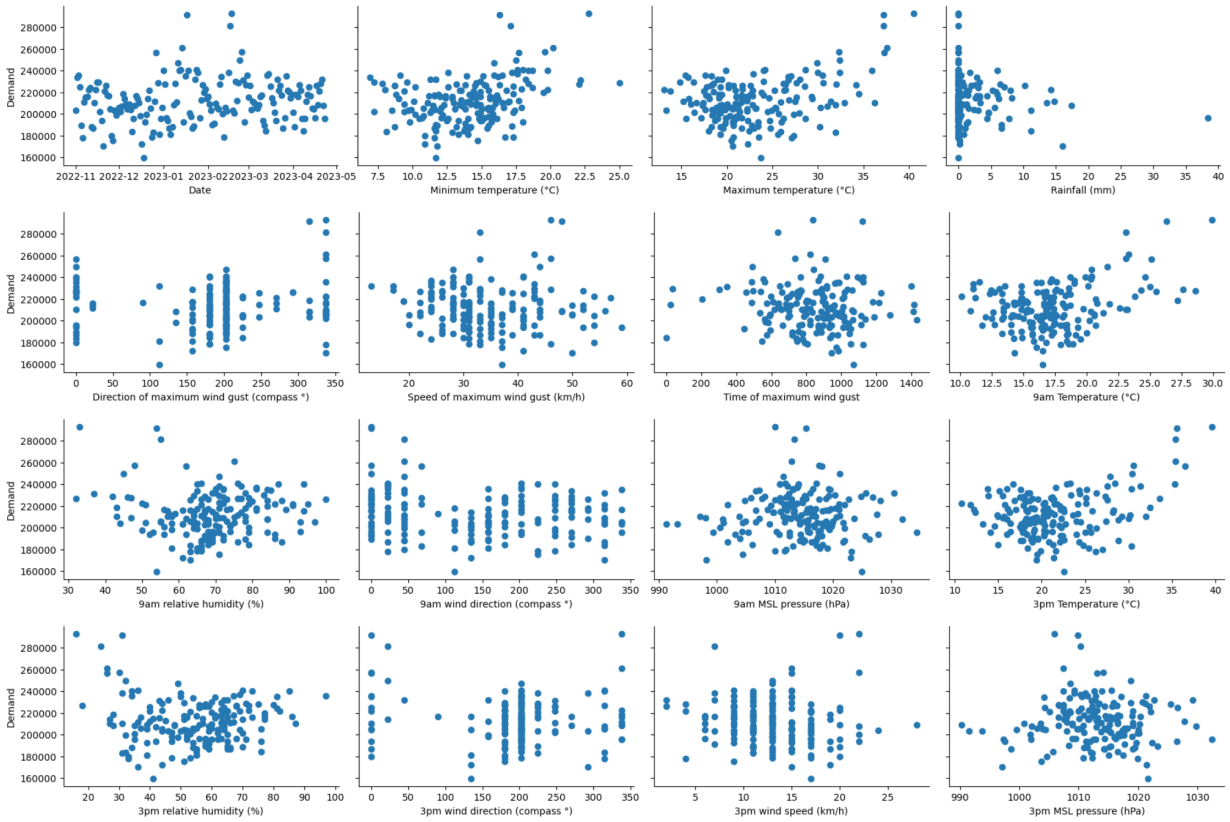
[1] https://stackoverflow.com/questions/67683766/convert-cardinal-wind-directions-to-degrees
[2] https://en.wikipedia.org/wiki/Points_of_the_compass
[3] https://stackoverflow.com/questions/74631092/pandas-zfill-multiple-items-in-single-cell

Model Building and Testing:

We chose to use a regression model as we're trying to predict Demand, which is a continuous numerical output. When choosing features for our model, we excluded the price variable, because it's dependent on demand.[4]

We started with correlation analysis to see how closely other variable changes correlate with change in Demand. We used scatter plots for this, utilising the Seaborn library which provides clear visualising with built-in labelling.[5]

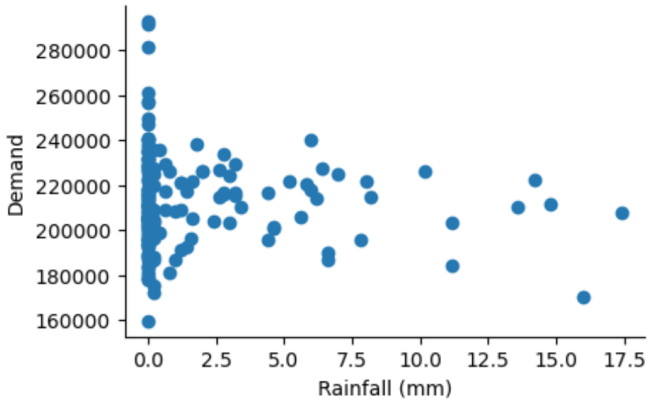
Correlation scatter plots of all variables vs Demand:



While there are outliers in many plots that could potentially be explained by contributing factors to the real-world weather data, the rainfall plot had a particularly noticeable, single high outlier. As this outlier would skew the results when using any of the rainfall data, we imputed this value to the mean of the column, instead.

[4] https://www.amber.com.au/blog/what-drives-the-wholesale-price-up-or-down
[5] https://seaborn.pydata.org/generated/seaborn.PairGrid.html#seaborn.PairGrid

Rainfall vs Demand after imputing outlier:

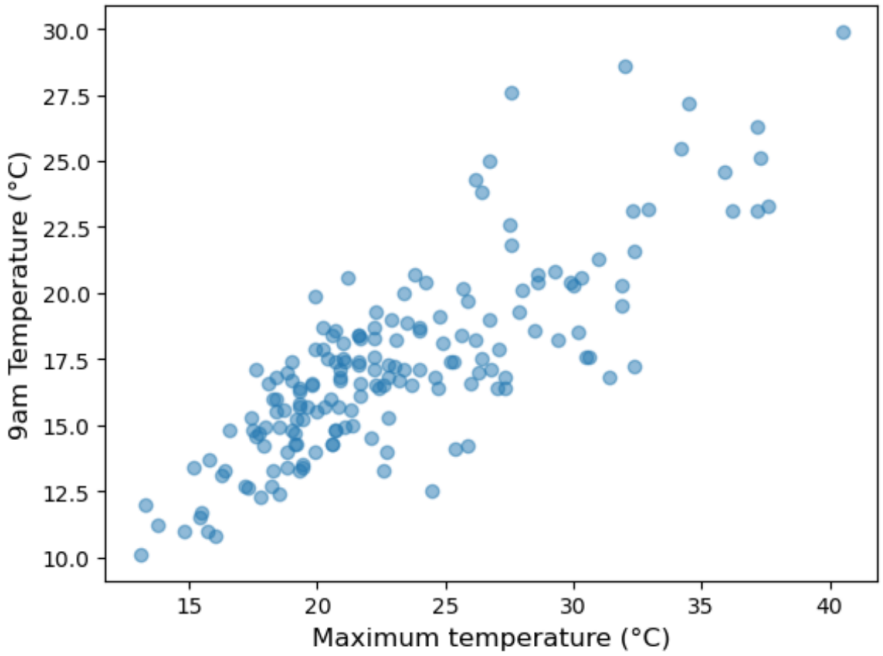


We used Pearson correlation to analyse relationships because we're using linear regression, the only type of regression we've covered in the course. A table of Pearson correlation metrics of Demand with all other features helped give further context, along with the scatterplots.

```
Pearson's correlation for all variables vs Demand:
9am MSL pressure (hPa)                          0.012934
3pm MSL pressure (hPa)                         -0.018447
9am wind speed (km/h)                           0.020679
9am relative humidity (%)                      -0.047605
Speed of maximum wind gust (km/h)              -0.059182
Time of maximum wind gust                      -0.062124
Rainfall (mm)                                  -0.064197
3pm relative humidity (%)                      -0.065690
Direction of maximum wind gust (compass °)      0.083736
3pm wind speed (km/h)                          -0.089362
3pm wind direction (compass °)                 -0.091559
9am wind direction (compass °)                 -0.195098
Minimum temperature (°C)                        0.344513
3pm Temperature (°C)                            0.355698
9am Temperature (°C)                            0.360563
Maximum temperature (°C)                        0.369480
Demand                                          1.000000
Name: Demand, dtype: float64
```

Ideally we wanted to use features that are most closely correlated with Demand. The Pearson metric table shows that this is "Maximum temperature." When choosing multiple features for multiple regression, though, we had to make sure they weren't too closely correlated with each other. Logically, we wouldn't want to use multiple temperature variables together. To check and illustrate this, we show the correlation value and plot "Maximum temperature" to the next most correlated feature, "9am temperature":

Examining correlation of variables for mutual regression: 9am temp vs Max temp:

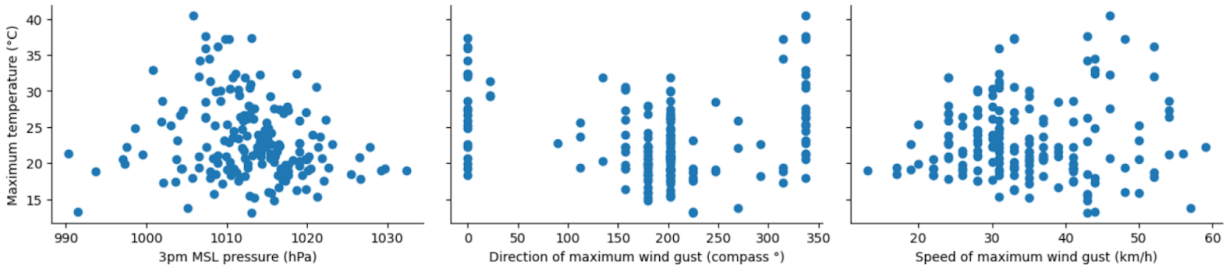



```
Pearson correlation of Maximum temperature (°C) and 9am Temperature (°C):
: 0.8059054261472441
```

We used another table of Pearson correlation to compare the rest of the features to “Maximum temperature,” as well as scatter plots, to choose these features. We were looking for features that had less than a .3 Pearson correlation (low or very low) with Maximum temperature.

```
Pearson's correlation for all variables vs Maximum temperature:
9am MSL pressure (hPa)                        0.003719
Time of maximum wind gust                    -0.015394
3pm wind speed (km/h)                         0.039839
3pm wind direction (compass °)               -0.045573
9am wind speed (km/h)                        -0.046376
Speed of maximum wind gust (km/h)             0.049710
Direction of maximum wind gust (compass °)   -0.104574
3pm MSL pressure (hPa)                       -0.157008
9am relative humidity (%)                    -0.301264
Rainfall (mm)                                -0.337381
Demand                                        0.369480
Minimum temperature (°C)                      0.516981
9am wind direction (compass °)               -0.558837
3pm relative humidity (%)                    -0.648110
9am Temperature (°C)                          0.805905
3pm Temperature (°C)                          0.966331
Maximum temperature (°C)                      1.000000
Name: Maximum temperature (°C), dtype: float64
```

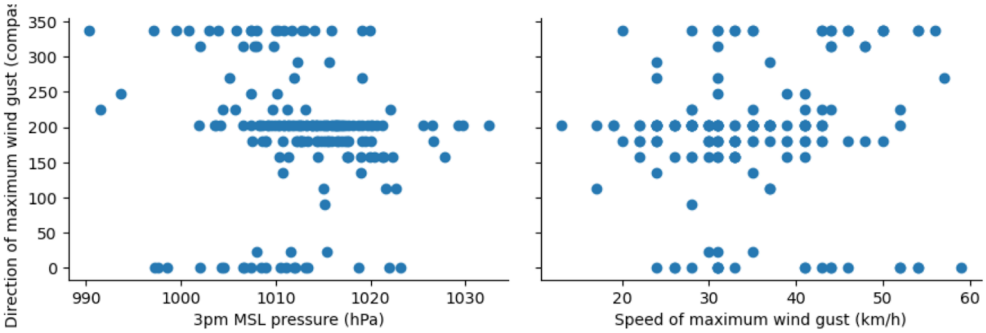
Correlation scatter plots of potential other features vs Maximum temperature:



We also needed to check the correlation of the rest of the features with each other. As '3pm MSL pressure (hPa)' and 'Speed of maximum wind gust (km/h)' were too closely correlated with each other, we did not use these together in any models.

Pearson numbers and correlation scatter plots of more variables to each other:

```
Pearson correlation of '3pm MSL pressure (hPa)' and 'Direction of maximum wind gust (compass °)':
-0.08070515404146197
Pearson correlation of '3pm MSL pressure (hPa)' and 'Speed of maximum wind gust (km/h)':
-0.5151689304551603
Pearson correlation of 'Direction of maximum wind gust (compass °)' and 'Speed of maximum wind gust (km/h)':
0.06866547642845494
```



Some of the chosen features, on their own, show a low or very low correlation with Demand, but we still needed to test them to see if they contribute to the model effectiveness. With so many highly correlated features (for example, four temperature feature columns, three wind speed feature columns), we were limited in the features that could be used together.

We tested five models with different feature sets, including four multiple regression and one simple linear regression model. We used the standard scaler on each data set. Using a testing function we built, we split the data into the train and test set, training on 80% of the data. The function then fits the linear model to the training data, predicts the y values based on the test data, and displays the evaluation metrics. Displayed are the predicted and actual values of the test data, the RMSE, MAE, and R2 of both the training and test data. Mean Absolute Error is a metric that is used in machine learning to measure the performance of regression models. It

calculates the average absolute difference between actual and predicted values. MAE was used due to its simplicity, reliability and accuracy.[6]

Model 1: features: 'Maximum temperature (°C)', '3pm MSL pressure (hPa)', 'Direction of maximum wind gust (compass °)'

```
Model: LinearRegression
Train RMSE: 18177.57
Train MAE: 14663.53
Train R2: 0.11

Test RMSE: 23115.77
Test MAE: 18637.59
Test R2: 0.22

Train score: 0.1091468031349686
Test Score: 0.22137205783333103
```

Model 2: features: 'Maximum temperature (°C)', 'Speed of maximum wind gust (km/h)', 'Direction of maximum wind gust (compass °)'

```
Model: LinearRegression
Train RMSE: 18157.92
Train MAE: 14613.29
Train R2: 0.11

Test RMSE: 22978.50
Test MAE: 18450.50
Test R2: 0.23

Train score: 0.11107156939055529
Test Score: 0.2305921796435052
```

Model 3: features: 'Maximum temperature (°C)', '3pm MSL pressure (hPa)'

```
Model: LinearRegression
Train RMSE: 18223.25
Train MAE: 14571.71
Train R2: 0.10

Test RMSE: 23540.17
Test MAE: 18864.05
Test R2: 0.19

Train score: 0.10466413776659289
Test Score: 0.1925188443618051
```

6 https://deepchecks.com/glossary/mean-absolute-error/#:~:text=Mean%20Absolute%20Error%20(MAE)%20is,effectiveness%20of%20a%20regression%20model

Model 4: features: 'Maximum temperature (°C)', 'Direction of maximum wind gust (compass °)'

```
Model: LinearRegression
Train RMSE: 18179.42
Train MAE: 14670.86
Train R2: 0.11

Test RMSE: 23164.13
Test MAE: 18628.02
Test R2: 0.22

Train score: 0.1089655111186495
Test Score: 0.2181109143630895
```

Model 5: feature: 'Maximum temperature (°C)'

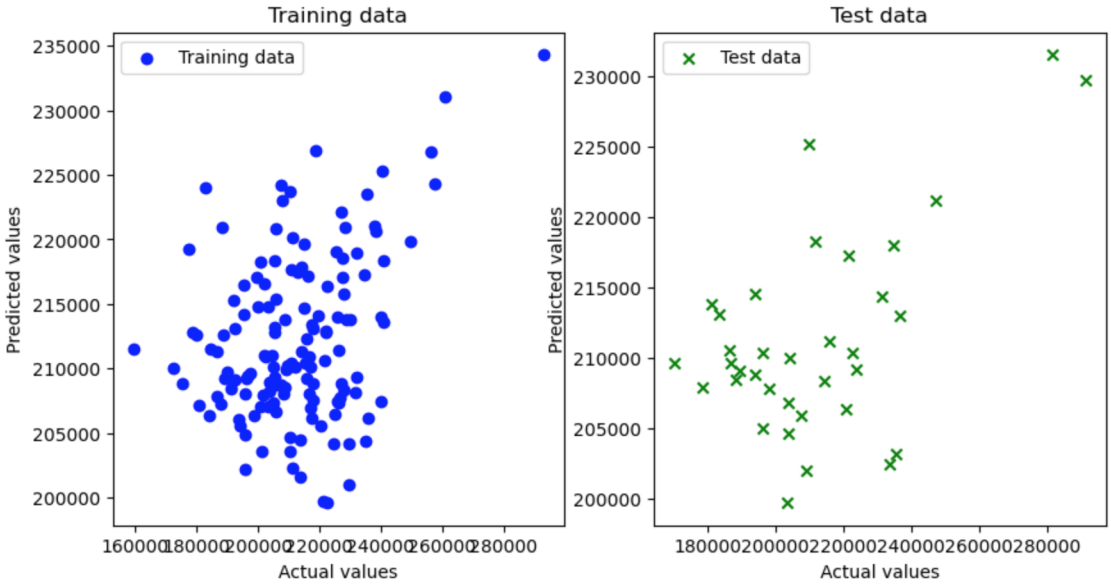
```
Model: LinearRegression
Train RMSE: 18224.27
Train MAE: 14575.44
Train R2: 0.10

Test RMSE: 23569.82
Test MAE: 18855.16
Test R2: 0.19

Train score: 0.10456329385471985
Test Score: 0.19048391831215272
```

Based on the R2 values, Model 2 seemed to be the most effective model, with 23% of the change in Demand explained by the change in the features.

Model 2 Predicted vs Actual values:



To analyse Model 2 more thoroughly, we also tested it with the MinMax scaler and didn't find any noticeable difference in the evaluation numbers:

```
Model: LinearRegression
Train RMSE: 18157.92
Train MAE: 14613.29
Train R2: 0.11

Test RMSE: 22978.50
Test MAE: 18450.50
Test R2: 0.23

Train score: 0.11107156939055529
Test Score: 0.2305921796435052
```

Further analysis of Model 2:

We then examined the full set of 34 test data values, versus the predicted values.The biggest gaps in the predictions seem to come when the actual values are very low or very high in the dataset.

	Model 2 Predicted values	Model 2 Actual values		Model 2 Predicted values	Model 2 Actual values
0	213131.031519	183406.37	**18**	188189.75	208453.329365
1	209167.209725	224000.36	**19**	207480.29	205895.537413
2	218257.834028	211616.90	**20**	233630.93	202435.944603
3	207922.716991	178602.78	**21**	203646.69	206877.122644
4	203220.000909	235672.34	**22**	189468.07	209106.031735
5	206388.416628	220897.61	**23**	231482.20	214393.706102
6	225233.558821	209996.47	**24**	181070.73	213795.377030
7	229771.188763	291349.26	**25**	186610.74	209690.551791
8	231529.387747	281608.67	**26**	214581.88	208372.427819
9	221183.531416	247335.97	**27**	236634.11	213033.503527
10	211187.398673	215967.78	**28**	196362.57	205031.017790
11	217283.705043	221523.82	**29**	198101.44	207847.078429
12	199735.421560	203251.28	**30**	193852.36	214557.078288
13	202001.598192	209028.17	**31**	222631.25	210387.340088
14	218031.569898	234738.93	**32**	204370.31	210012.643654
15	210394.796335	196419.80	**33**	186415.69	210544.504364
16	208844.679633	193799.87	**34**	203832.63	204622.761397
17	209664.475467	170252.78			

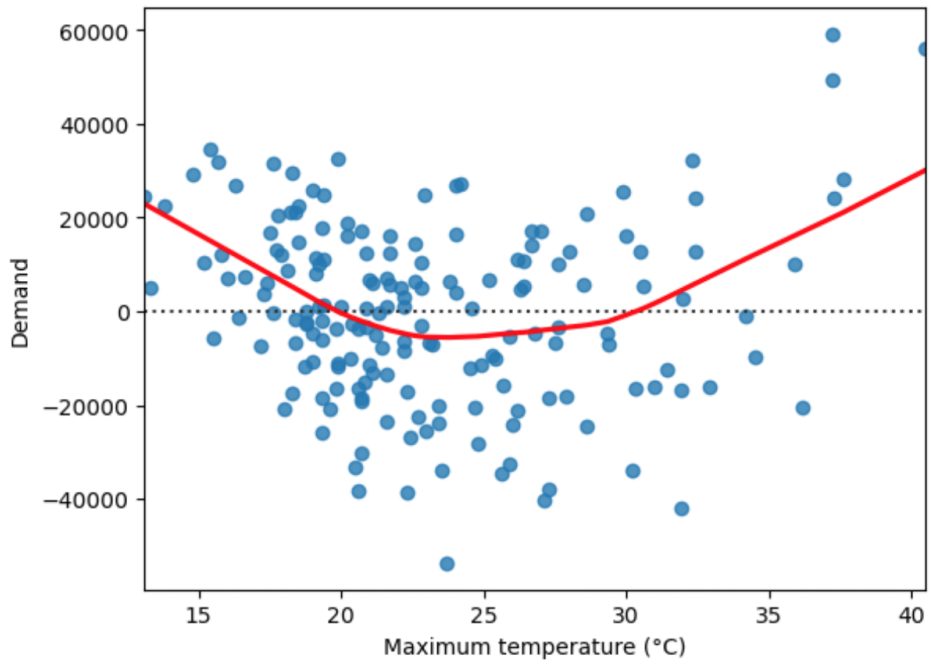
We also used the kfold cross validation to see if our testing was valid. Based on testing 10 splits, the average R2 was much lower than the initial test, showing that our split was “lucky” and the model is much less effective than initially shown. This R2 value would mean that only 3% of the change in Demand could be explained by the change in features, using this model.

```
R2 scores: [-0.0984061912563674, 0.052431706628107255, 0.2285442530078846, -0.05160510530684648, -0.4303914361235577, -0.0524839
5791495253, 0.15181797520827878, 0.19732736034328757, 0.10194350493203042, 0.2453180704643858]
Average r2 score:0.03444961799822503
```

Due to this, we also tried the kfold cross validation on Model 1, which had the second highest r2 score in initial testing. With the kfold tests this model also had a very low average r2 score, but still lower than Model 2.

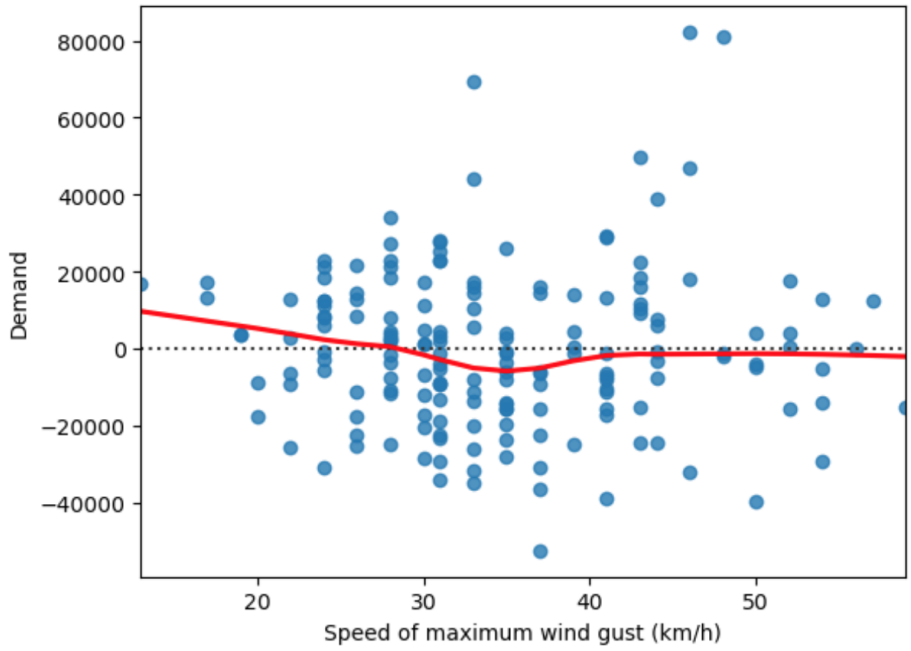
```
R2 scores: [-0.14518382372918937, 0.007152663566971307, 0.2593265294525122, -0.02335231053011544, -0.4599207803929979, -0.137476
82231158898, 0.19173691105408852, 0.20066751780926295, 0.15801897478029814, 0.23254883045128705]
Average r2 score:0.028351769015052854
```

Residual analysis of the features used in our testing is shown in scatter plot charts with a fit line, utilising the seaborn library.[7] Looking at each feature individually, the Maximum temperature residuals are not linear, and there is a clear curve showing they are not independent, showing that the relationship between the data is not linear and a linear model is not the best type of model to use. The predictions at the lowest and highest ends of the temperature spectrum will be less accurate.

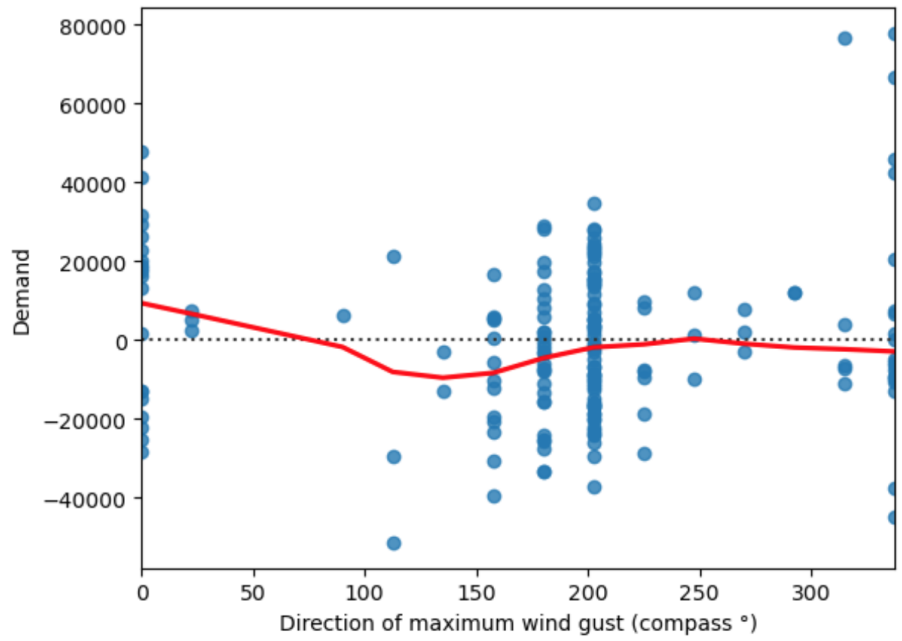


[7] https://seaborn.pydata.org/generated/seaborn.residplot.html

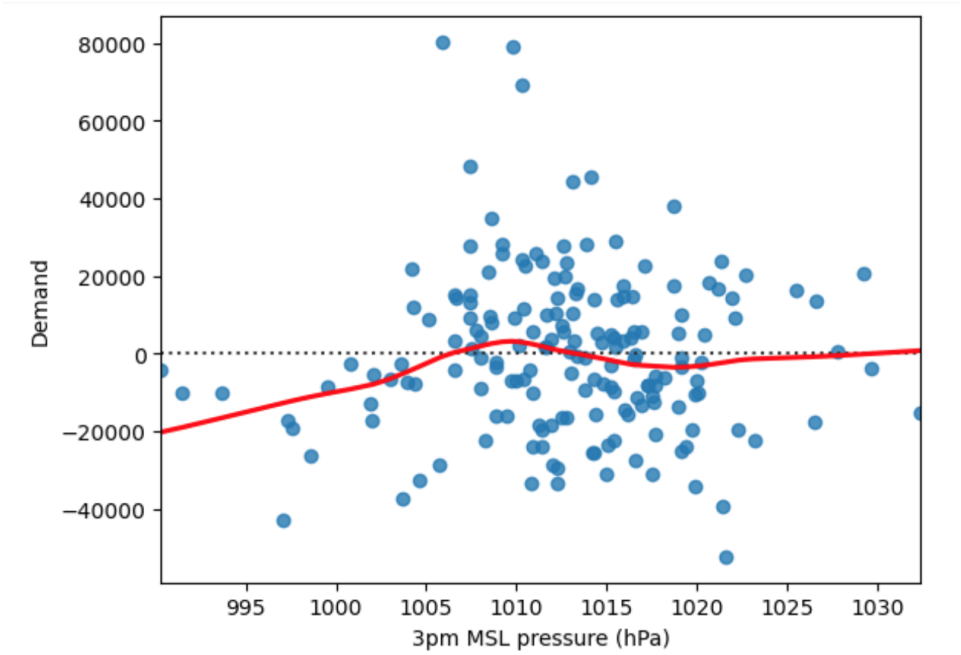
The “Speed of maximum wind gust” residuals look better, but with some lack of constant variance at the lowest and highest ends of the spectrum.



The “Direction of maximum wind gust” residuals lack constant variance, with wind directions around 200 degrees (SSW), 337.5 (NNW) and 0 degrees(North) showing much more variance. There may be a “hidden” variable at play, such as temperature changes that are associated with those wind directions in particular.



The 3pm MSL pressure residuals lack some constant variance around the middle of the spectrum but mostly seem to meet the assumptions.



Model 2 conclusions:

Our best model, Model 2, is shown not to be very effective after using kfold cross validation. Residual analysis shows that the relationships may not be linear and we would need to learn a non-linear regression model to analyse this further.

The most highly correlated linear value, Maximum temperature, seems to have a non-linear relationship with Demand. Reflecting on how energy is used in our own homes, it makes sense that if the Maximum temperature is low, such as in winter, more energy would be used in heating. In temperate weather, less energy would be used, and in hotter weather (when this feature value is higher) more energy would be used, again, for cooling. This would mean there is an inverted bell curve shape to the data of Max temp vs Demand, and a linear model shouldn't be used.

Referencing Model 2 features, the "Speed of max wind gust" and "Direction of max wind gust" would be associated with "wind chill" and how cold it feels. According to the CSIRO, Australian homes are particularly "leaky" and draughts can add up to 20% to energy bills.[8] Thus in our model, we would think these features would be associated with higher Demand as more energy is used for heating. The Pearson correlation of these features with Demand is low, though they do contribute to the model's effectiveness. The residual analysis of the "Direction of max wind gust" also suggests that the relationship is not linear. Increased domain knowledge of weather

[8] https://www.csiro.au/en/news/All/Articles/2020/May/testing-the-leakiness-of-australian-homes

feature relationships (e.g. are northerly winds lowering temperatures more than southerly winds?), as well as learning about different types of regression models, may allow us to build a model that better utilises these features and is more effective.

We would need to learn a non-linear regression model to analyse this further. Sklearn includes many other types of regression models.[9] We see that in a Random Forest Regressor model, the "3pm temperature" data would be the most important to the Demand data.
Using our test_model function to test several other regression models, the Gradient Boosting Regressor and Extra Trees Regressor models showed better performance, with lower RMSE values and a higher R2 value in both training and testing sets.

Research on current trends:

The climate in Melbourne in 2022 presented some significant trends in both rainfall and temperature. The level of rainfall was above average across Melbourne, with some months experiencing exceptionally wet conditions. Also, there was a spate of warm nights which seem to align with recognizable climate change trends and global weather patterns.

Rainfall
The range of rainfall varied from 116% to 141% over the year. October and November were notably the wettest months of the year. There were widespread thunderstorms across Eastern Australia in November.[10] Some of this is attributed to the combination of La Nina and the negative Indian Dipole which refers to warmer Ocean temperatures surrounding Indonesia, and cooler waters around Africa. This phenomenon increases the prevalence of low pressure systems over the south east of Australia as well as the level of moisture in the air.[11] [12]

Temperature
The temperatures in Melbourne over the period being analysed had incredible variability from highs of 40.5°C and lows of 15.8°C. While the days were relatively cool, the nights were quite warm at an average of 15.2°C more than a degree above the long term average of 14.0°C. This was consecutively the 27th year with noticeably warmer nights.[8]
The biggest reason for such dramatic weather events seems to be due to climate change. This is why we see increased amounts of rain but also heatwaves and bushfires.[13] According to the

[9] https://scikit-learn.org/stable/modules/classes.html#

[10] Australia Weather: Weekly weather from the Bureau of Meteorology: Sunday November, 2022 - The Global Herald

[11] https://www.abc.net.au/news/2023-03-01/australia-summer-weather-wrap-lower-temperatures/102033972

[12] https://thelatch.com.au/weather-summer-australia-2022-2023/

[13] https://www.abc.net.au/news/2022-08-06/what-is-negative-indian-ocean-dipole-and-does-it-mean-more-rain/101305822

World Meteorological Organization, 2022 was on track to be one of the warmest years on record. The Australian Bureau of Meteorology also reveals that the heating due to climate change has increased Australian temperatures by more than 1.47°C since 1910.[10]

The dramatic variability of temperatures limits the effectiveness of a model which relies on data collected over a short period of time (less than 1 year), and for which maximum temperature is the strongest predictor. More data over a longer period of time would give us more context, and more material to train an effective model. Climate change contributing to the unpredictability of current weather vs. historical trends, also means that simpler models such as the types we have learned about, will be less appropriate.

Limitations:

Even though the analysis provided valuable insights into the connection between daily weather conditions and energy demand, there are some limitations which must be taken into consideration.

1. The team are not subject matter experts in weather features and the effects of weather patterns which may impact energy consumption, and despite our research efforts, more domain knowledge would be helpful.
2. The analysis relies on data collected over a short time frame which may not properly capture seasonal trends or patterns which could impact energy consumption and pricing over time
3. Even though the weather data is specific to Melbourne, the price and demand data covers the whole of Victoria. Since energy consumption and pricing can vary across different regions, the model's applicability to other locations could be limited.
4. External factors such as economic, and social factors which could also influence energy use, are not included in the data provided.