Weather Indicators and Their Influence on Real Estate Prices in Melbourne, Australia

**Introduction**

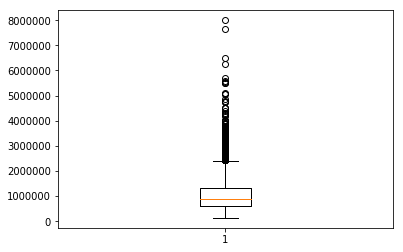
The purpose of this report is to analyze the real estate trends in Melbourne, Australia and whether the weather the day of sale of the real estate demonstrates an impact on the final price. The data is two datasets, one of real estate sold over a period, and another of weather attributes for everyday for a period. They are combined based on their dates.

When deciding on this project, to accomplish something that would answer some questions I already had. I also wanted this project to focus on areas that I am interested in. So, I was interested in exploring the housing market in Australia because I am interested in visiting Australia soon, and I have an interest in real estate. As well, I was curious to see if there are possible trends with final purchase price of real estate and the weather of that day.

I wanted to explore if there is a psychological factor in a buyer’s choice. Moreover, with the rise of climate change, weather becomes more volatile. Already Semenenko and Yoo have identified that volatile weather affects real estate prices in their research called *Climate Change and Real Estate Prices.* Specifically, they found that lower volatility correlates with higher prices. However, they did not research which weather indicators caused their results. So, putting some interests together with the knowledge of Australia’s current climate crisis and the lack of research, this project was created.

This project explores whether simple weather indicators relate to a buyer’s purchase price. Specifically, are there weather indicators that effect the buyer’s final purchase price of real estate? If so, by how much? As it turns out, weather does have an influence over the final sale price of real estate, however minute. The indicators sunshine and max temp had the most influence of all the indicators. In this report, there is a demonstration of the data including their source, information on attributes, and charts reporting summary statistics. There is a description of the method used to gathering and manipulating the data, extracting key information, and preparing the data. An explanation of the analysis found with the algorithms and including any visuals. Finally, the results measuring the accuracy of a model with weather indicators and a model without weather indicators.

**Data**

The data for this project was taken from Kaggle. The weather dataset is from the Bureau of Meteorology of the Commonwealth of Australia. They compiled this data from taking measurements from different weather stations across the greater Melbourne area. The housing dataset is provided by an author who retrieved the data from publicly available data from Domain.com.au. The raw datasets of the weather include 145,461 rows and 23 columns and the housing market includes 34,858 rows and 21 columns. Of the weather dataset, the column variables are dates, locations, temperature, rainfall, sunshine, and different wind descriptors. These variables are populated with dates, location by name of city, and numbers in the form of floats. Of course, this dataset will need to be cleaned to single out the specific location to be Melbourne, Australia. Of the dataset, the columns variables are dates, locations, different housing variables, and prices. These variables are populated with dates, location by address and suburb, something that will need to be transformed into similar format as the weather data, and price in AUS dollars. As well, the housing variables include characters to describe the type of property sold, size of the property and number of rooms are integers. Also included are the sellers and method, however they were not needed. The target variable price, from the housing dataset, is skewed right with the summary statistics listed below.

|  |
| --- |
| count 5634.00 |
| mean 1076078.24 |
| std 672594.31 |
| min 131000.00 |
| 25% 623875.00 |
| 50% 888000.00 |
| 75% 1340750.00 |
| max 8000000.00 |

Figure 1: Statistics of Sale Price of Houses

Figure 2: Boxplot of Sale Price of Real Estate

Moreover, the weather indicators, the variables of interest, mostly demonstrated normal spread except for Rainfall. In the analysis is the explanation for standardizing and filtering these variables, but for a clearer picture the variables standardized is pictured in boxplot form in Figure 3 below.

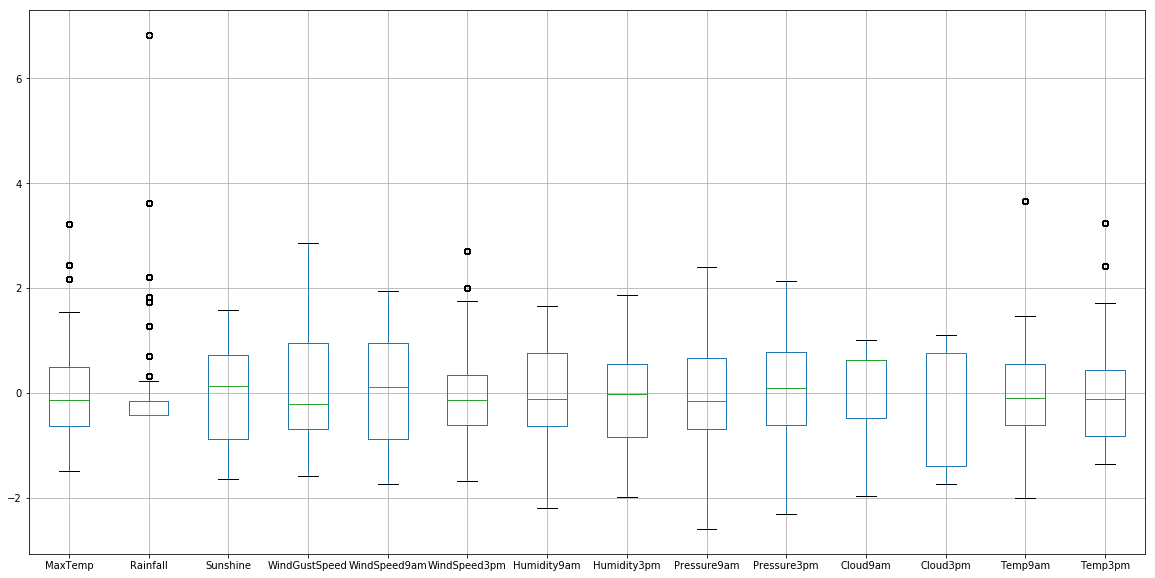


Figure 3: Boxplots of Weather Indicators Standardized

**Method**

The methodology used was simple. First, the two datasets were downloaded from their respective Kaggle repositories. They were then imported as DataFrame objects into the notebook. When first analyzing the data, there were a lot of null values littered in each dataset. Some of these values were compounded in some columns that were removed. As well, some columns deemed unnecessary were removed, such as ‘Seller’ from the housing dataset and ‘RainTomrrow’ from the weather dataset. Next, the date columns for each dataset were transformed from text to DateTime using the method pd.to\_datetime. This allowed for the datasets to then merge, using the merge(), method easily on the key ‘Date’. As one dataset, there were still many null values. Reluctantly, quite a few more rows were removed as well as the column ‘Location’.

Once the dataset did not include any more null values, the correlation of the price of each real estate against the weather indicators were measured. To visualize this, the package Seaborn was used. The analysis section will discuss the numbers for this.

Prices were then analyzed by the location categories ‘Regionname’ and ‘CouncilArea’. These groupings are definitely an important distinction when discussing price of real estate. However, there were problems with attempting to use LabelEncoder() and HotOneEncoder() which led to their removal in the model training and analysis. As well, the prices were analyzed by date, type of real estate, and if it rained that day. There was an attempt as well to categorize the date column to their different seasons and use that to train the model. However, with similar issues as the location categories, this attempt and attempts on the other categorical data were removed. These problems are something future research should include as they could potentially prove to be meaningful in the analysis.

Moving forward, the weather indicator columns were standardized. The goal of this transformation was to allow for measuring the regression to be more accurate. The package sklearn.preprocessing was used to import StandardScaler. In completing this, as seen in Figure 3 above, all columns were then centered around 0. This created a way to observe these columns easily but resulted in no difference in the measure of the regression.

Then, the data was prepared for training. There were two subsets created, X and y, used for train\_test\_split(). This resulted in four NumPy arrays: X\_train, X\_test, y\_train, y\_test. After this, XGBoost was used to train the model. This algorithm was chosen because it can easily demonstrate which features of the data were more important to the accuracy than others. Finally, the results and the feature accuracy were printed.

**Analysis**

After cleaning the data, the analysis began with measuring the correlations of the columns against the column ‘Price’. In this, it was revealed that there was not any significant correlation between the price and any weather indicators. The scale used in the visual below is auto set and it ranges from [0.12,-0.08]. This range demonstrates a very weak correlation.



Figure 4: Correlation Between Real Estate Indicators and Weather Indicators

As previously mentioned, after analyzing the correlation, the data was split into different categorical groups. Doing this demonstrated that the Price of real estate differed significantly by Region and by Council. Some places averaged $400,000 while others $1,600,000. This makes sense as real estate prices are dependent on location. As well, splitting the data by whether day of sale had rain also showed a difference in mean by $60,000. This split probably was the strongest in favor of the research question, however as previously mentioned it was left out of the results. Splitting the data by type of real estate also demonstrated significant differences per category average price. The difference from the lowest: $589694.67 and the highest: $1,305935.84 probably attributes to the largest lost in accuracy in the model used. Finally, in categorizing the data by date, there was a slight increase in average prices when Australia was in its summer (winter months for northern hemisphere). Unfortunately, there were difficulties separating the data into the seasons and applying this categorization to the model. In all, the analysis of these categorical variables demonstrates that more research is necessary, including the use of LabelEncoder and OneHotEncoder to create a dataset that can train from these categorizations.

Next, preparing the data to be trained, train\_test\_split() was used, and the categorical variables previously mentioned were removed. Then the package XGBoost was used to create a model xgb.XGBRegressor(objective=’reg:linear’, seed=99, verbosity=0). Since the prices are continuous linearly, the objective was set to ‘reg:linear’. Also, the seed was used to generate consistent results every time the model was used. It was then fit to the data using xg\_reg.fit(X\_train, y\_train) where X\_train and y\_train are the attributes and targets used for training. Then a prediction array was created using xg\_reg.predict(X\_test) where X\_test is the attributes used for testing, and this was used to measure the Mean Squared Error of the model.

Once the model was fit and the predictions were made, the packages cross\_val\_score and RepeatedKFold were used to measure the accuracy of the model. These packages are general packages to measure the accuracy of a model. As well, the package mean\_squared\_error was used to measure the mean\_squared\_error of the model. MSE is also an important metric because it gives insight into how the model performed. In the model, the accuracy was found to be 76.9% accurate with a standard deviation of 3.8%. As well, the MSE came out to as 107,452,694,717.53242. When the features of importance were plotted, the most important weather feature was Sunshine at an F score of 123. This paled in comparison to the overall important feature of Land size with an F score of 895. These results are shown below in Figure 5.

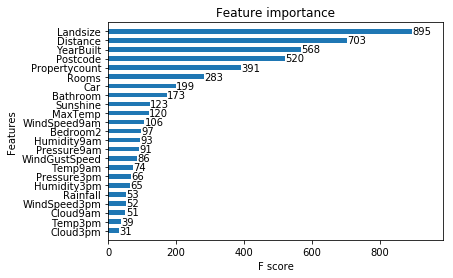


Figure 5: Features of Importance of Model with Weather Indicators

Finally, interested in how a model could perform without the weather indicators, once more a model was fit and measured as before. This time the weather features were left out. The accuracy of the model was found to be 77.1% with a standard deviation of 4.2%. The MSE of the model was 110,105,607,694.02255. Below in Figure 6 are the features of importance of this model.

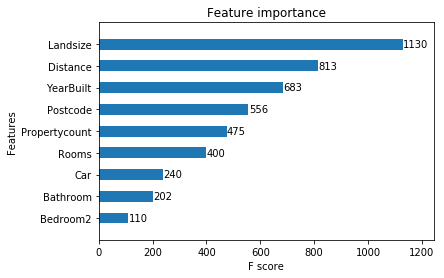


Figure 6: Features of Importance of Model Without Weather Features

**Results**

From the two model results, inferences can be made. In comparison of the accuracy of the models, the one without the weather features is slightly better. This makes sense as more features can negatively influence a linear regressive model. However, in comparison of MSE, the model with weather features performed better, where MSE is smaller by 2,652,912,976.49. This could be explained by some weather features performing better than the housing feature ‘Bedrooms’. As previously mentioned, the strongest weather feature did not have a large influence over the resulting model, but it appears that in comparing MSE’s there is an influence in the final sale price of real estate.

**Conclusion**

When deciding to sell real estate there are many factors a seller needs to consider. The timing of the sale, the buyer, and other considerations affect the bottom line. However, it appears that the weather of the day of sale can affect the bottom line as well. From this report, it is demonstrated that weather does not overly influence the sale price of real estate but can have a little influence. Sunshine and max temp had the most influence over the final price, indicating that indeed, warmer and sunnier days do show higher sale price of real estate.

Looking to the future, research needs to be done on the categorical variables described in the analysis. As they were explained, they probably can help the model’s accuracy and further describe the influence of weather over a price. As well, more research should be done on the weather of real estate showings, negotiations, etc. This way, a greater understanding of the weather can lead to a greater understanding of the psychology of a buyer and the final sale price.

As well, this project had to remove the use of GEOJson from the final deliverable. There was a time constraint problem. This is something that also should be researched further and included in future analyses. The use will allow for easy-to-understand visuals that any person could play around with.

**Appendix**

References

Pino, Tony. Melbourne Housing Market: Melbourne housing clearance data from Jan

2016. Kaggle.com, 2019. [Melbourne Housing Market | Kaggle](https://www.kaggle.com/anthonypino/melbourne-housing-market?select=Melbourne_housing_FULL.csv)

Semenenko & Yoo. Climate Change and Real Estate Prices. Canadian Center of Science

and Education, 2019.

Young, Joe. Rain in Australia: Predict next-day in Australia. Kaggle.com, 2020.

[Rain in Australia | Kaggle](https://www.kaggle.com/jsphyg/weather-dataset-rattle-package)