EDA for Hypothetical Housing Market

James To

15/10/2020

Task 1: Data Analysis and Exploratory Data Analysis (EDA)

Explain what is the iterative cycle of EDA including all the steps involved.

Data Analytics is the exploration of data through visualizations, summarizing statistics, and modeling that lead to conclusions and insights into problems, prospects, and discoveries. The 5 steps of data analysis are as follows:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

One form of data analysis is EDA, which is primarily used for providing descriptive analysis by searching for trends, correlations, or other relationships between the variables to generate ideas or hypotheses. In essence, EDA is developing a greater understanding of data through a series of steps. These are as follows:

1. Formulate your question based on the data
2. Search for answers to the questions in the given data set by:
    a. Cleaning the data
    b. Transforming the data
    c. Visualize the data and summarize the statistics
3. Re-evaluate the question and look to generate new questions

EDA can be explained as an iterative cycle where each cycle starts with a new question, answering that said question, and then re-evaluating and/or creating new questions. This cycle can be repeated until all questions pertaining to the data are exhausted. The ultimate goal of each cycle is to gain a better understanding of the data. Furthermore, this becomes the foundation for modeling, external research, and the search for external data sets once the EDA is satisfied.

Task 2: EDA for Cydney property data

Describe the objecteve of EDA in your case with reasons.

The Cydney property market data was provided to *Old South Wales Consulting (OSWC)* to be analysed for our client, *Best Homes*. The goal of the project is to find what factors affect the price of properties other than just the size. Hence, by doing a preliminary analysis using the EDA method, we can garner a better understanding of the data.

To perform the EDA, the data must be examined such that the EDA methodology is suitable. This elementary examination would include checking the number of variables, identifying the variables as continuous or discrete, and gaining a broad idea of the data's potential for visualization and predictive modelling.

The data set contains twelve variables - ten numerical and two character - with 19,990 data points. These variables are *price*, *who posted it*, *if it's under construction*, *RERA approval status*, *number of rooms*, *type of property*, *size of the property*, *ready to move*, *resale or not*, *longitude*, *latitude*, *number of supermarkets nearby*, and *number of parks nearby*. These variables are indeed sufficient for EDA due to the diverse variety of and type of variables. However, they must be cleaned and checked during the EDA as well.

Task 3: The Exploratory Data Analysis

Follow the steps of exploratory data analysis described in Task 1 to conduct the whole process of EDA.

This report will go through the questions asks and answered in each iterative cycle of the EDA. This exploration concluded in three iterative cycles where *OSWC* were happy to report the findings.

Iterative Cycle 1

Based on the given data set and the context of the project, two starting questions were asked to analyze the data.

1. What are the highlights and underlying problems with the data?
2. What are the correlations between each variable?

To begin the EDA, the data must be first imported and cleaned. This cleaning process also acts as a gateway to our first question.

Prior to the EDA, the the csv file containing the raw data in imported into RStudio which was achieved through the following line of code:

```
data <- read.csv("trainset.csv") #import data
```

Data manipulation was also employed to correct variable types into factors and numerical. This was to ensure the data would be easily used for a variety of functions (i.e. plotting and summaries).

Iterative Cycle 1: Question 1

Clean the Data To get a better understanding of the data and what needs to be cleaned, a summary was assessed (**Appendix 3.1.1**).

The summary showcased a median price of $610K and a mean price of $1Mil for properties in the data set. In addition, the median size is 1,170 square feet and mean size is 1,351 square feet. However, there were a few issues that needed to be addressed, such as the variable types and outliers, in order to clean the data and improve the data quality.

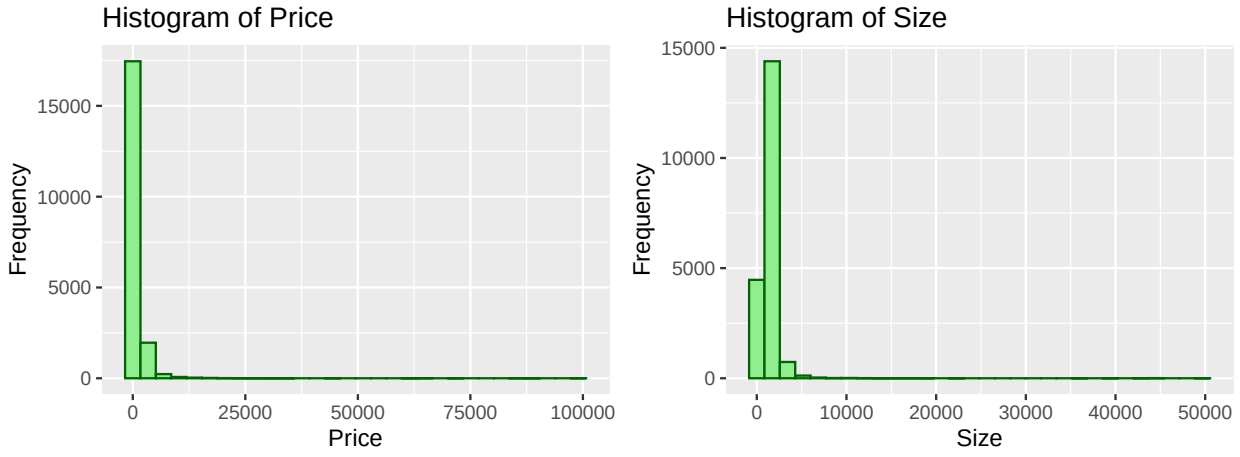
Missing Data The summary showcased that there were a variety of missing values in almost all of the variables. Upon closer inspection, there were 27 rows of data that are affected by the missing values. To address this issue, these rows were removed from the data. Data imputation using the *mice* package was considered, however, the missing rows made up less than 0.1% of the data, which made removal the most efficient approach.

Duplicates There were 73 identical, duplicated rows of data in the data set and they were removed. Removal of duplicates in a combination of longitude and latitude was considered as these houses would be double-listings. However, without the knowledge of the characteristics of the property, they were not removed in consideration for apartment complexes.

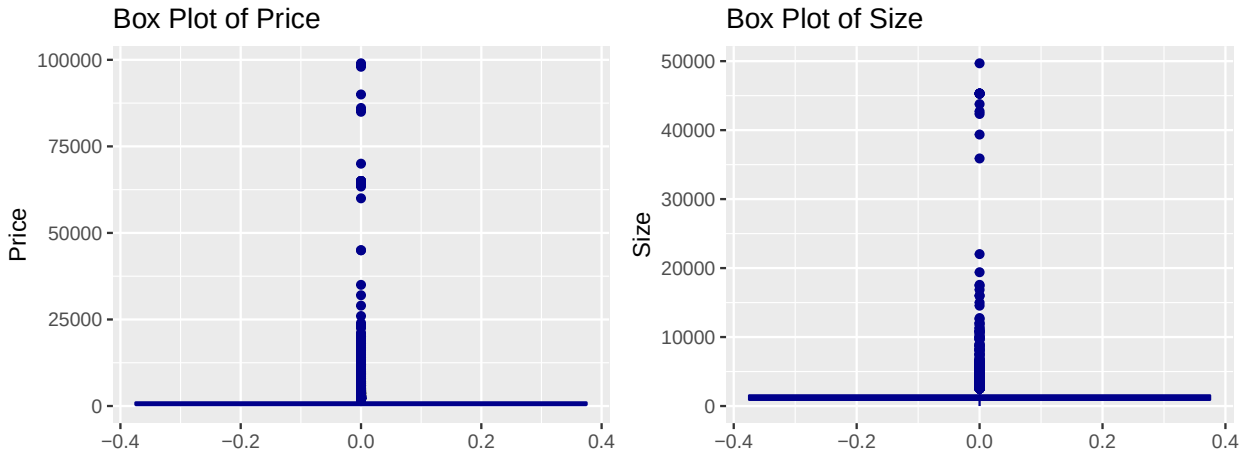
Incorrect Values Based on the summary, a number of incorrect values could be spotted as shown by the the maximum values. The maximum values of number of supermarkets nearby and number of parks nearby were 999,999. Using logical domain knowledge, it was concluded that these were incorrect values, thus deleting the rows containing them.

There were two other variables with incorrect values - *Type of Property (ToP)* and *Under Construction (UnCon)*. In the ToP variable, there was the incorrect capitalization of the "BHK" value where some of the values were spelt "bHK", and these values were corrected to be "BHK". In the categorical variable for *UnCon*, there are supposed to be only two values - 'zero' for not under construction and 'one' for being under construction - but it was found that values 'two' and 'three' were present. The rows containing these values were removed from the data set.

Outliers The range of values for both variables *Price* and *Size* were very extensive as displayed by the histograms below.



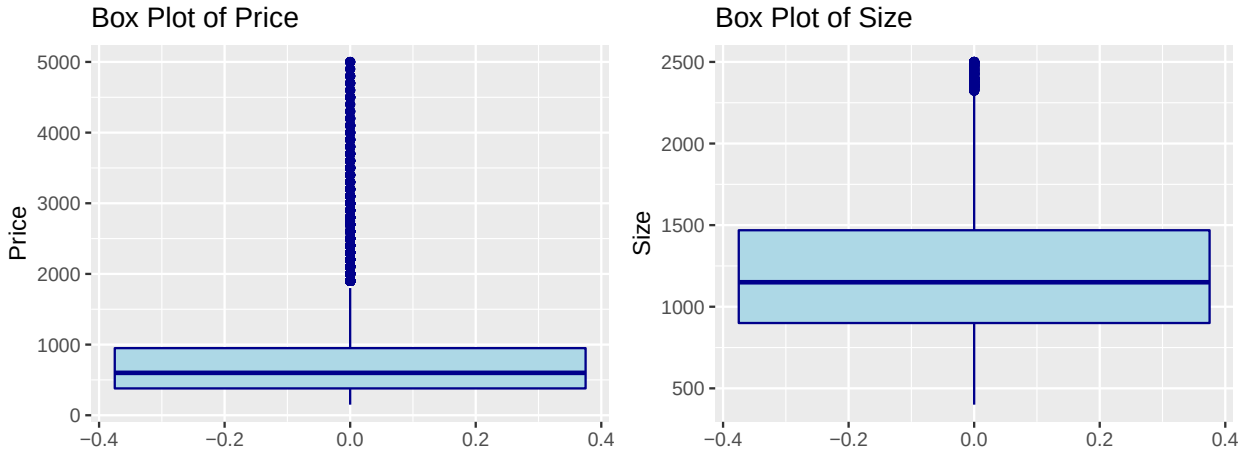
The data was heavily skewed with a large amount of values on the lower tail and a sudden drop in values of price greater than $2.5M and size greater than 10,000 feet square. This was further accentuated by the box plots below.



The points on the box plot indicate the outliers, which are most prominent above the upper bound. These upper outliers are calculated using the interquartile range - between the lower and upper quartile - plus the upper bound ($Q3+1.5IQR$). The initial thought was to remove these outliers from the data set as a majority of them were nonsensical. However, removing such a large number of data points could be detrimental to the end goal of the project. Hence, boundaries for both *Price* and *Size* were made based on domain knowledge.

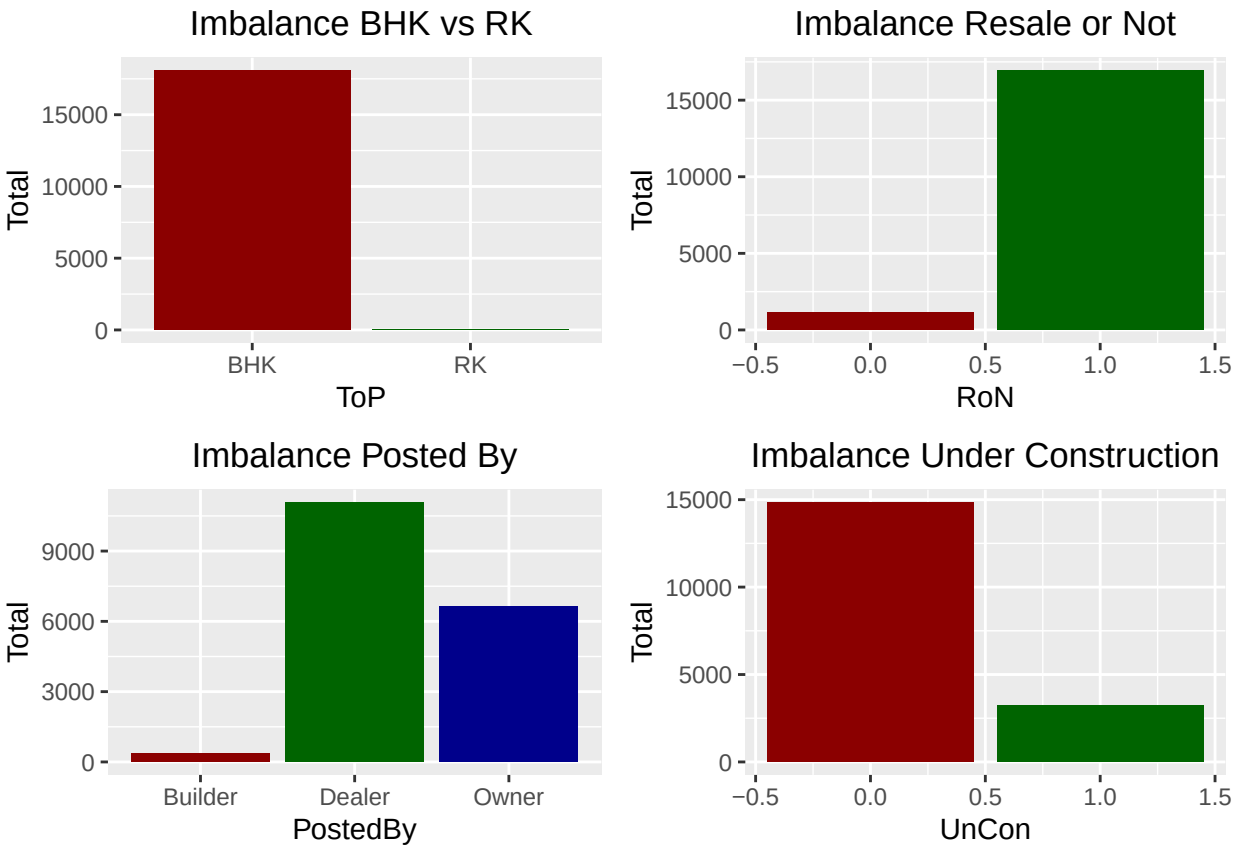
Given that the city of Sydney is a reasonable comparison to Cydney, boundaries were chosen based upon Sydney's housing market. Given the cheapest house prices in Sydney are approximately \$200K, and standard living size is greater than 400 feet squared (37 metre squared), the lower boundary was set to prices under \$150K and house sizes under 400 feet squared. For the upper boundary, although prices in Sydney can reach upwards of \$100Mil, the lack of data points with a price above \$2.5Mil and upper whisker of the box plot being set at \$1.9Mil, makes \$5Mil a generous compromise for the upper limit. As for the upper boundary for size, it was set to square feet (232 metre squared), which is larger than two soccer fields. Anything above of below these boundaries were removed.

The box plots below are the new data sets.



This process lead to a total of 18,071 data points left, which is 90% of the original data. A summary fo the cleaned data can be found in **Appendix 3.1.2**.

Imbalance of Variables There was a resounding imbalance of data for four categorical variables - Type of Property, Resale or Not, Under Construction, and Posted By. This is showcased by the four plots below.



Observations:

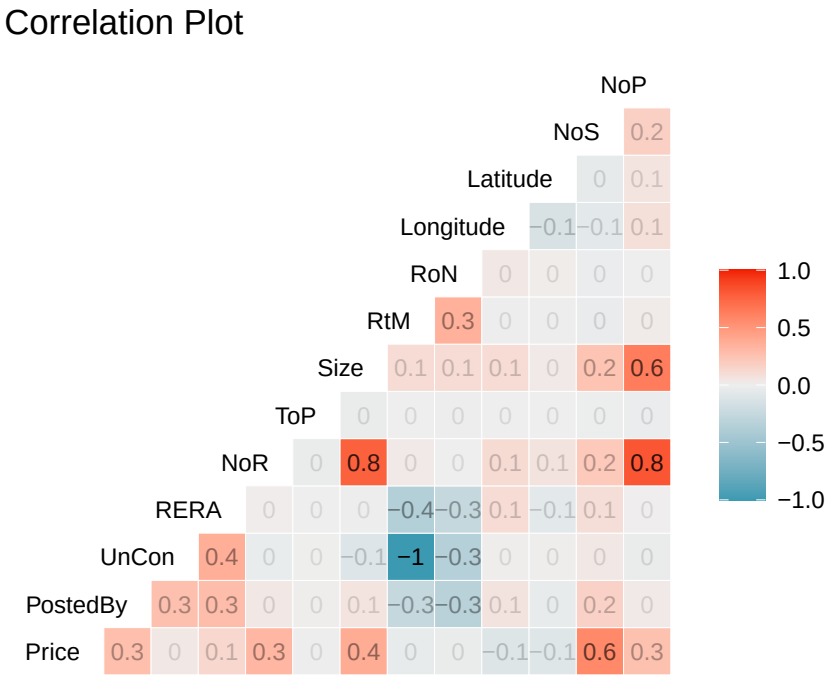
- *Type of Property* essentially only contained properties with bedroom, hall and kitchen (BHK). There were only four properties with rooms and kitchen only.
- Majority of houses are not resale
- Builders have the least listings (only 367 listings)
- Majority of the homes are not under construction (3,244 were under construction)

This puts into question the usefulness of the variables to help determine factors that affect house prices. This could be particularly troublesome for modelling, as there would be heavy biases to certain variables. In future, the data set would need to be more balances for a better analysis or simply removed.

Iterative Cycle 1: Question 2

Looking at the correlation

Using a correlation matrix, we are able to identify the relationships between variables.



The correlation plot helped identify the various correlations between the variables. The correlation between *Price* and *Size* was found to be 0.4, which is lower than anticipated, but is relatively high to other variables. However, without the outliers removed from *Price*, the correlation between *Price* and *Size* was found to be 0.8. This aligns more with expectations because typically the larger houses are more expensive.

There were a few other notable correlations between other variables. Unsurprisingly, the correlation between *NoR* and *Size* was found to be very high at 0.8 and, as expected, the negative correlation between properties that are ready to move and under construction. Furthermore, a notably high correlation of 0.8 between Price and number of supermarkets nearby is present in the data. In addition, the correlation between the Number of Parks and the Number of Rooms is unexpectedly high at 0.8, which may be a coincidence.

The correlation between price and number of supermarkets was particularly interesting, and was further explored in Iterative Cycle 2.

Iterative Cycle 2

For the previous cycle, it was found that there were numerous issues with the data quality, and after the cleaning the correlations between variables were explored. Rather than explore the relationship between price and size, it was decided that taking a closer look at the relationship between other variables should be explored. The question was asked as follows:

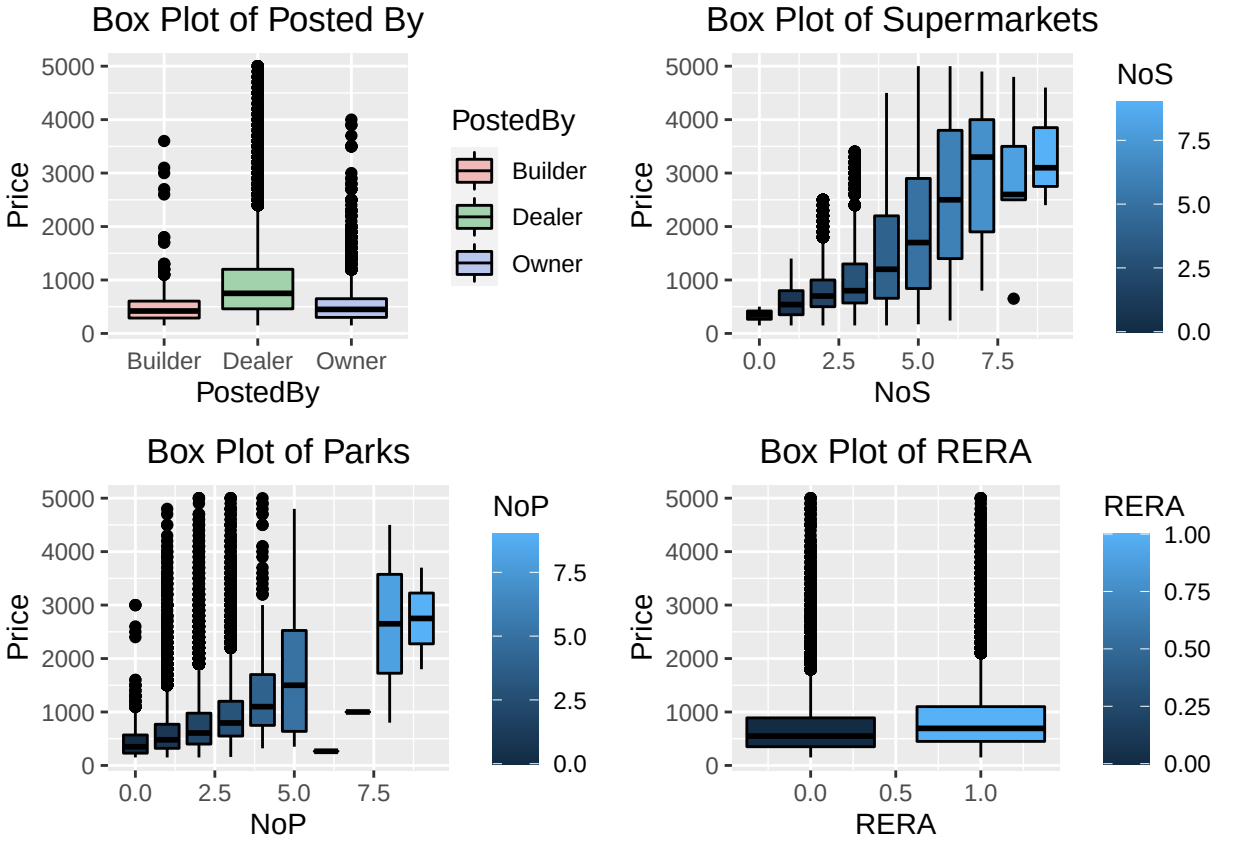
- What is the relationship between price and other variables?

This question aims to grasp a better understanding between some other variables and the price.

Iterative Cycle 2: Question 1

Relationships between Price and Other Variables

Four variables were selected to be analysed further and these were “Posted By”, “Number of Supermarkets”, “Number of Parks”, and "RERA. To explore these variables relationship with price, the following box plots were graphed:



When the “PostedBy” box plot was explored, there was clearly a higher price listed by dealers than both builders and owners. Dealers posted much more expensive properties than both owners and builders, with a mean of $966K in comparison to $528K and $522K. An explanation for this could be due to intermediary fees, purposeful increase in prices, or the properties sold by the dealer are more higher value. However, it was noted that there were a larger number of outliers in all three categories.

The large correlation of 0.8 between price and number of supermarkets nearby discovered in the last iterative cycle was further accentuated by the box plot, which displays the higher price as the number of supermarkets increase. This is surprising considering that the majority of people only shop at one supermarket, however, this could also be a possible indicator of living in a central location, which is generally priced higher. This formed a basis for a new question explored in Iterative Cycle 3.

The number of parks nearby also affects seems to have a relationship with price. As the number of parks nearby increases, so does the price of the property, however, there is hardly any data for houses with Number of Parks more than six. This showcases the unreliability of this variable, as it’s unrealistic to assume that a house has more than three parks nearby. To improve the data quality, the properties with more than six parks could be grouped. The number of parks also put into question the locations of these properties, which is further explored in Iterative Cycle 3.

Based on the RERA box plot, if the property is RERA approved, then the property is worth slightly more. The mean price for properties that are RERA approved was found to be $924K and those not approved

were $738K. It is difficult to assume anything about the variable as the description provided is very minimal. Regardless, as with any certification, it is logical for a certified property to be worth more.

Iterative Cycle 3

As mentioned in the previous iterative cycle, a question in regards to geolocation was materialized when realizing the number of supermarkets nearby heavily correlated with the price. This was further extrapolated to question if these locations were within central areas. Furthermore, what is there an increase in value when living in a central district. The hypothesis for this investigation was that:

"Geolocation does affect the price of the property."

The questions asked are as follows.

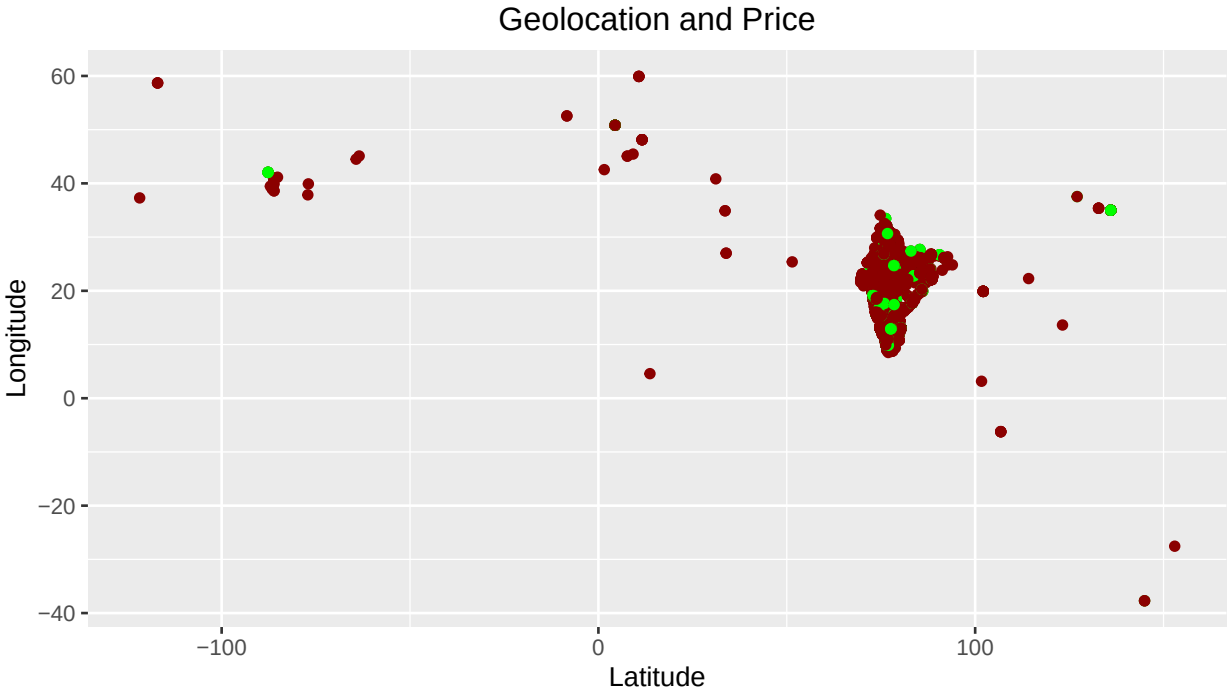
1. What is the relationship between Price and Geolocation?
2. What is the relationship between Supermarkets, Parks and Geolocation?

Before answering these questions, it must be acknowledged that the latitude-longitude co-ordinates provided cannot be compared to the real world. When put on a world map (Appendeix 3.3.1) the locations are spread around such that there are houses in both the North and South Pole. For the purposes of the EDA, these locations are mapped out as if they represent a single city.

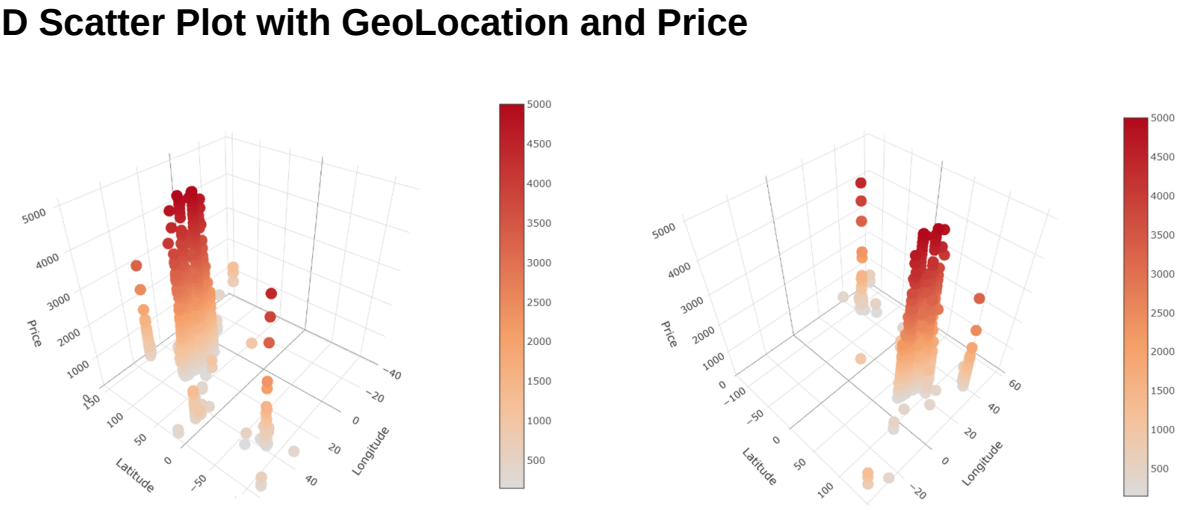
Iterative Cycle 2: Question 2

Relationship between Price and Location

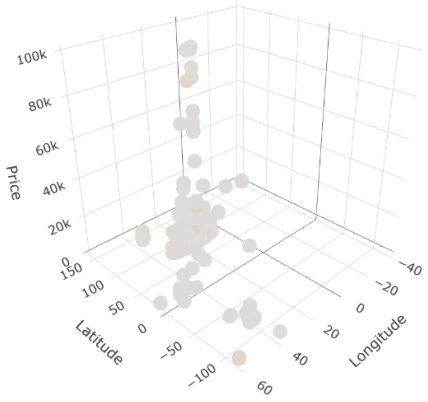
The scatterplot below depicts the locations of each property, the x-axis is the latitude and the y-axis is the longitude. The dark red points are properties under $1Mil and the green dots are the properties above $1Mil.



The cluster surrounding co-ordinates (75, 20) has the majority of the data in it - 17,921 data points. In this cluster there are 3,729 properties priced above $1Mil, this is 20% of the properties in the cluster. Outside the cluster there are 150 properties, 31 of which are above $1Mil, indicating that the location of the house does not dictate the house price as both have a 20% chance of being above $1Mil. The 3D plot depicts this below.



This 3D scatter plot has the x-axis and y-axis as latitude and longitude respectively, and a z-axis to depict the price of the property. Furthermore, the darker the plot point, the more expensive the property. As stated previously, there are numerous properties that are over $1Mil both inside the cluster and outside. However, this was the catalyst for a reconsideration of the outliers, which was mentioned in Iterative Cycle 1: Question 1. The below plot demonstrates that without outliers removed, there is clearly more expensive homes within the clusters.

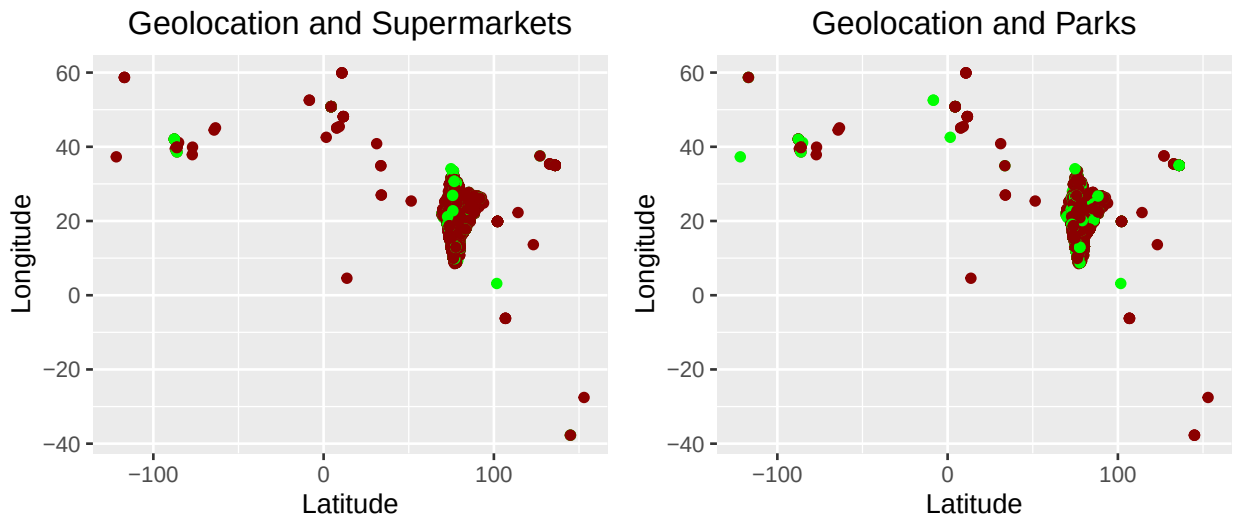


Within the cluster, there is a resounding majority of properties that are over $20Mil. Though these are considered outliers, given the expensive house price in Sydney's CBD these would need to be considered for a more accurate model to predict price with location. This was evidence that the location of the property did in fact affect the price.

Iterative Cycle 3: Question 2

Geolocation, Supermarktets, and Parks

Similarly to the previous question, plotting the co-ordinates and highlighting the Number of Supermarkets and Number of Parks results in the following scatter plots.



The left figure highlights the locations with more than two supermarkets nearby. A majority of these locations are with the central cluster, which is to be expected as this cluster could be considered the main city. However, one concern with the data is that the number of supermarkets is not homogeneous across the same location. This emphasizes the flaws in the data.

The right figure highlights the locations with more than two parks nearby. These locations are surprisingly present in bother the central cluster and outside of it. Those outside the central cluster are expected as these locations could be deemed more rural - this is on the basis the data is comparable to Sydney. The initial impression of the vast majority of these locations within the central cluster is that the data is incorrect. However, given the large number of parks that are in major cities such as Sydney and New York, this follows.

Task 4: EDA Conclusion and Model Recommendation

Conclusion of EDA

A summary of each iterative cycle below:

Iterative Cycle 1 The first cycle constituted cleaning the data and finding correlations, which was in respect to the pertaining questions. The data had a large number of issues with it. these included outliers, duplicates, and missing values. A total of 10% of the original raw data was removed during this process. Furthermore, it was found there were imbalances in the variables, which would need to be dealt with - either by resampling or new data - for modelling. A correlation plot was made to find the correlations between the variables, and a few notable ones stuck out. The highest positive correlations were between number of number of rooms and size, and number of parks and number of rooms, both with a correlation of 0.8. The largest negative correlation of -1 was between under construction and ready to move. In future, for modelling, one of each in a pari would need to be removed to avoid multicolinearity.

Iterative Cycle 2 To explore the relationship with price further, box plots for four variables - PostedBy, NoS, NoP, and RERA - were explored, and each variable had unique takeaways. For the "PostedBy" variable, it was shown that dealers posted much more expensive properties than both owners and builders, with a mean of \$966K in comparison to \$528K and \$522K. For the NoS variable, it was found that the more supermarkets nearby, the higher the price. This was speculated to be due to location, which would have more supermarkets in a central district. Similarly, the number of parks also had a positive relationship with the price of the property. Both of these were further investigated in the next iterative cycle. Finally, if the property was RERA approved, then the price was higher.

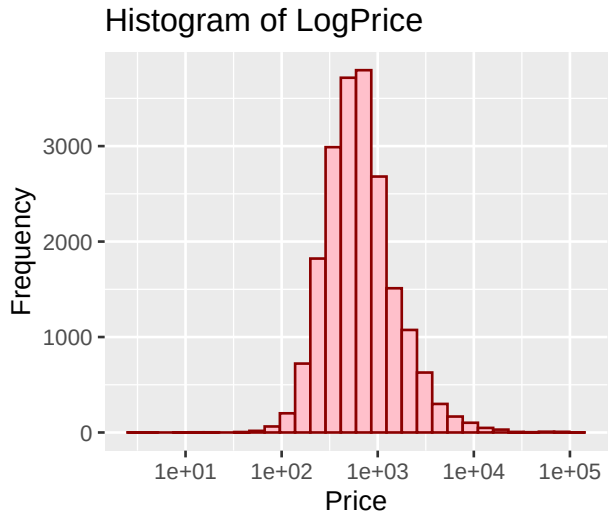
Iterative Cycle 3 As mentioned in the previous iterative cycle, the location of the properties were examined along with price, numbr of supermarkets, and parks. A disclaimer to this analysis was also provided due to the questionable co-ordinates. Given that Cydney is a city, the longitude and latitude were assumed to be scaled down. A central cluster was identified in the data, providing some insight as to the characteristics of Cydney. On initial analysis, it was found that property prices were not significantly different inside the central cluster or outside. The likelihood for a property of \$1Mil was 20% in both the cluster and outside. However, the price outliers addressed in Iterative Cycle 1 were re-evaluated, and it was discovered that extremely expensive properited were all within the central cluster. Furthermore, the number of supermarkets and parks were very much within the central cluster, giving evidence that the evidence that this cluster is a central district is true.

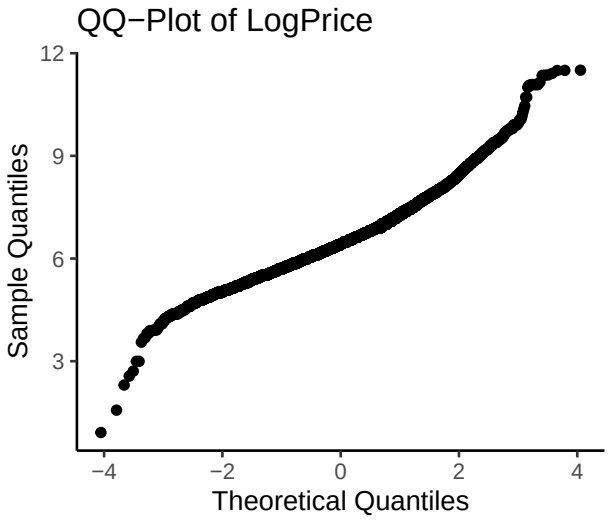
The EDA helped attain a greater understanding of the housing market of Cydney. By examining a variety of variables, we were able to identify factors that may significantly affect the price. Furthermore, we identified variables that may not be as useful for the forecasting price. Regardless, this analysis successfully provided the knowledge to choose certain models.

Model Recommendation

Given the skewed nature of the original price variable that will be the response variable, it is obvious that linear modelling techniques are not appropriate for the modelling. Hence, non-linear techniques would be far better suited for this.

Log-Transformed Linear Regression Given that the log price has a likeness to a normal distribution, a Log-transformed Linear Regression could be used.





However, the tails of the distribution are not normal distributed (as shown by the qqplot), hence this model might be inaccurate for more expensive or cheaper homes.

Generalise Linear Model (GLM) Given that price is continuous and non-linear, GLMs are arguably the most suited for the modelling. To add more to this, there is a large mix of variable of both categorical and numerical, and GLMs are able to handle these diverse variable types and cater for the different distributions. Within this process, feature selection would need to be performed as to remove the covariate variables. Furthermore, some of the imbalanced variable would need to be manipulated or removed due to the lack of usefulness due to how largely one-sided the values are - example being the non-existent "RK" values in the *ToP* variable. In summary, the best model for forecasting price would be the GLM.

Conclusion

In summary, the data for the housing market of Cydney was analysed using the EDA methodology, going through three cycles. This provided *OSWC* a greater understanding of all data and ideas for future modelling. Two types of models were suggested to begin modelling; Log-Transformed and GLM. Once modelling is completed, *OSWC* will be able to provide the findings and results to our client *Best Homes*.

Appendix

Appendix 3.1.1

```
##      Price           PostedBy             UnCon             RERA
##  Min.   :    2.5   Length:19909       Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  380.0   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :  610.0   Mode  :character   Median :0.0000   Median :0.0000
##  Mean   : 1069.9                      Mean   :0.1792   Mean   :0.3146
##  3rd Qu.: 1000.0                      3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :99000.0                      Max.   :3.0000   Max.   :1.0000
##  NA's   :7                            NA's   :4        NA's   :3
##       NoR             ToP                Size            RtM
##  Min.   : 1.000   Length:19909       Min.   :    3   Min.   :0.0000
##  1st Qu.: 2.000   Class :character   1st Qu.:  900   1st Qu.:1.0000
##  Median : 2.000   Mode  :character   Median : 1170   Median :1.0000
##  Mean   : 2.392                      Mean   : 1351   Mean   :0.8211
##  3rd Qu.: 3.000                      3rd Qu.: 1545   3rd Qu.:1.0000
##  Max.   :20.000                      Max.   :49690   Max.   :1.0000
##                                      NA's   :6       NA's   :1
##       RoN           Longitude         Latitude           NoS
##  Min.   :0.0000   Min.   :-37.71   Min.   :-121.76   Min.   :     0.0
##  1st Qu.:1.0000   1st Qu.: 18.45   1st Qu.:  73.80   1st Qu.:     1.0
##  Median :1.0000   Median : 20.91   Median :  77.32   Median :     2.0
##  Mean   :0.9332   Mean   : 21.29   Mean   :  76.89   Mean   :    52.2
##  3rd Qu.:1.0000   3rd Qu.: 26.90   3rd Qu.:  77.79   3rd Qu.:     3.0
##  Max.   :1.0000   Max.   : 59.91   Max.   : 152.96   Max.   :999999.0
##                                    NA's   :4         NA's   :4
##       NoP
##  Min.   :     0.0
##  1st Qu.:     1.0
##  Median :     2.0
##  Mean   :    52.1
##  3rd Qu.:     2.0
##  Max.   :999999.0
##  NA's   :2
```

Appendix 3.1.2

```
##      Price           PostedBy             UnCon             RERA
##  Min.   : 150.0   Length:18071       Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 380.0   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 600.0   Mode  :character   Median :0.0000   Median :0.0000
##  Mean   : 796.4                      Mean   :0.1795   Mean   :0.3143
##  3rd Qu.: 950.0                      3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :5000.0                      Max.   :1.0000   Max.   :1.0000
##       NoR             ToP                Size             RtM
##  Min.   : 1.000   Length:18071       Min.   : 400.0   Min.   :0.0000
##  1st Qu.: 2.000   Class :character   1st Qu.: 900.2   1st Qu.:1.0000
##  Median : 2.000   Mode  :character   Median :1150.1   Median :1.0000
##  Mean   : 2.328                      Mean   :1205.4   Mean   :0.8205
##  3rd Qu.: 3.000                      3rd Qu.:1468.8   3rd Qu.:1.0000
##  Max.   :10.000                      Max.   :2500.0   Max.   :1.0000
```

```
##       RoN           Longitude         Latitude           NoS
##  Min.   :0.0000   Min.   :-37.71   Min.   :-121.76   Min.   :0.000
##  1st Qu.:1.0000   1st Qu.: 18.11   1st Qu.:  73.85   1st Qu.:1.000
##  Median :1.0000   Median : 20.74   Median :  77.34   Median :2.000
##  Mean   :0.9365   Mean   : 21.24   Mean   :  77.03   Mean   :1.753
##  3rd Qu.:1.0000   3rd Qu.: 26.90   3rd Qu.:  78.11   3rd Qu.:2.000
##  Max.   :1.0000   Max.   : 59.91   Max.   : 152.96   Max.   :9.000
##       NoP
##  Min.   :0.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :1.834
##  3rd Qu.:2.000
##  Max.   :9.000
```

Appendix 3.3.1