By: Abhinav Sharma, Jesper Li, Mervyn Giritharan & Prashanti Kodi

Melbourne Housing Analysis Report

Problem Description:

Since 2000, house prices in Australia have appreciated by 150%. Given the alarming rate at which houses have been priced up, some economists have theorized that the Australian property market, especially in cities like Melbourne, may have been overpriced and may be due for a price correction. However, prices have only increased since then and are expected to rise even more in the future. With such high prices, it has become a topic of discussion on what the ideal places are to invest in real-estate within Melbourne. The dataset of Housing Market Data 2018 in Melbourne allows us to understand the patterns of house prices by taking various factors into consideration. We sought to understand the drivers of property prices and answer the following questions from the perspective of real estate investment businesses:

- How much have house prices risen in recent years?
- What are the best areas to invest in Melbourne?

Importance of the problem:

The global financial crisis of 2007 – 2008 was touted to be the most serious financial crisis since the Great Depression, prior to the COVID-19 recession. While Australia was one of the few countries not severely affected by the recession, it did fall victim to the universal housing crisis. The city of Melbourne is said to be experiencing fluctuations in house prices, making it necessary to analyze the market to pinpoint safe and lucrative investment avenues. Using the dataset and modelling, we will try to discover trends in the data which could help us understand which properties are best for investments. We chose this data problem as while it does deal exclusively with Melbourne, but real estate investment is ubiquitous. Using this dataset, we may be able to understand the patterns behind house pricing. The analysis and modelling of the data will not only be beneficial or limited to Australian housing markets, but may also help with house pricing in other countries.

Exploratory analysis:

The dataset was found on kaggle and consists of 21 variables for 34857 observations. Variables include Price, Suburb, Address, Rooms, Type, Method, SellerG, Date, Distance, Postcode, Bedroom, Bathroom, Car, Landsize, BuildingArea, YearBuilt, CouncilArea, Latitude, Longitude, Regionname, and Propertycount. Out of these variables, there are 14 variables that have missing values. After computing the table of missing values, we decided not to use the variables BuildingArea and YearBuilt because 60.6 percent and 55.4 percent of their total values were missing, respectively. Also, the variables Address, Suburb and Postcode will not be used in the analysis because they are unique for each observation and do not provide any information of statistical significance, and we have the same information coming from the latitude and longitude variables. For the rest variables, we divided them into two groups: categorical variables and numerical variables. Categorical variables include Type, Method, SellerG, Date, and Regionname, and number variables include Price, Rooms, Distance, Bedroom, Bathroom, Car, Landsize, CouncilArea, Latitude, Longitude, and Propertycount. We drop the data that the variable price is NA before we start our exploratory data analysis.

Before doing exploratory data analysis, we create a graph (Figure 1) of price based on a time series from 2016 to 2018. From the graph, we do not see a clear pattern to demonstrate how the market is moving; the graph shows a stable market where prices have been moving within the same level for years. To better understand the hidden information behind the graph, we explore the dataset. To begin the EDA, we look into the categorical variables.

1. SellerG: There are 310 different sellers. However, only around 10 percent of the sellers sold more than 100 properties, and almost 30 percent of sellers only sold one property. Therefore, most of the data tends to be unique and will not be used in analysis.
2. Type: There are three types of properties in the dataset: houses, townhouses, and units. The histogram (Figure 2) shows that the number of houses is much more than the sum of the number of townhouses and the units. A dummy variable is created to represent whether the property is a house.
3. Regionname:This variable is correlated with Distance, latitude, and longitude. We will talk about these variables together.
4. Method: There are five methods of selling the properties in the dataset. The boxplot (Figure 3) shows that methods of sale such as direct sale, property passed in and vendor bid have slightly higher price than properties sold prior, and sold after auction categories.

The correlation matrix (Figure 4) of all the numerical variables indicates the significant correlation of housing price with respect to number of rooms, distance from CBD, number of bedrooms and bathrooms, land size and latitude. Therefore, we can now focus on the bivariate relationship for variables most correlated with price.

1. We see the joint distribution plot of price vs distance (Figure 5), segregated by region name shows a non-linear relationship and implies a negative correlation between price and distance from the central business district. It also suggests that the areas with the

highest demand are the metropolitan areas, especially the Southern metropolitan area. The joint distribution plot of price vs latitude (Figure 6) provides the same insight; houses between -38 to -37 latitude, the metropolitan areas, tend to be more expensive than other areas.

2. The histogram of Landsize is incredibly skewed because of the presence of a few houses with extremely large land size. For example, the size of the largest house is analogous to the size of Vatican City. These data are definitely incorrect entries. We remove outliers by following a slightly lenient definition of outlier: outliers are data points with price below the first quartile minus 3 times of interquartile range or above the third quartile plus 3 times of the interquartile range. After removing the outliers, the skewness is greatly reduced, as observed from an updated histogram (Figure 7) of land size with most land sizes between 0 to 800 square feet.
3. The scatter plot of rooms (Figure 8) is a bit surprising. It shows that houses with four to five rooms are relatively more expensive. We would have expected bigger houses to be more expensive, but it may be that bigger houses are far from city centers, thus bringing prices down. It also implies location has a stronger relationship with price.

The removal of outliers greatly reduces the null values in our dataset. For the remaining null values, we impute the missing values of Latitude, longitude with the mean of the full dataset, and the missing values of Bedroom, Bathroom, and car spots with the mode of the full dataset.

Solution and insights:

Post EDA, we feature engineer temporal components such as week, month and year from the Date variable. We use year and week as continuous to capture the trend of time series and use month as categorical variable to extract potential nonlinear seasonality in the time series. We then partition the data by mid-2017 to form our training (n=10,134) and test (n=7,563) sets. The idea is to check model R2 of all models on the same test (hold-out) to ensure comparability. Given that we can’t predict the past using the future, we refrained from using cross-validation in modeling and relied on hold-out error metrics to validate models. Using the pipeline functionality, we try trees, KNN, RandomForest, Boosting by XGB and LightGBM with different hyperparameters and observe that test R2 is best from RandomForest (0.788) and lightGBM (0.812). Then, we go on to see which features are the best drivers of the Price.

Since both best performing models are tree-based, we plot a decision tree (Figure 9) and see that the first few nodes are created by partitioning based on Regionname, Latitude, #Rooms, Type and Distance from center. So, it is evident that location has the most important impact when deriving property pricing. For RandomForest and LightGBM, we explore multiple model interpretability strategies and align on using SHAP importance, which is model agnostic and provides directionality in insights.

The first few common important variables from RF feature importance (Figure 10), LGB feature importance (Figure 11) as well as SHAP summary plot (Figure 12) are Distance from CBD, Latitude, Longitude, Landsize and Southern Metropolitan Region. We observe that the farther away a property is from the CBD, the lower its expected price. Intuitively, the larger the Land Size, the higher the price. Similarly, properties in the Southern metropolitan region are associated with a positive shap value, which consecutively indicates an increase in price. This is expected because EDA of Price vs regionname showed relatively higher prices in the southern metropolitan area as well as could be due to proximity to the sea (Figure 13)

Features such as #Rooms, #Bathrooms also show a unidirectional impact on price, with price increasing as the number of rooms increase. While Propertycount comes high up in the relative importance, SHAP doesn’t show a clear association between feature value and impact. Few additional one hot encoded variables such as Bayside City Council, Boroondara City Council, Seller Marshall, Type being housing also affect price prediction implying these variables are helpful in deriving price prediction.

Given that Regionname is one of the most important parameters and has a clear directional nature, we decided to split up our stable time series plot by region to see any patterns. We see that the market for the Southern Metropolitan region has been booming whereas the housing market in all other regions was cooling off (Figure 14). This explains the stability in the overall line chart, as the market dynamics in different regions are moving in opposite directions. Surprisingly, we do not see much seasonality in our time series as the month flag variables do not appear in the SHAP summary plot. We do see the week (week of the year) variable suggesting some trend of price increasing over the course of the year, regardless of region, house-type or real-estate agents.

Based on the overall analysis, our recommendation would be to invest in a house in the Southern Metropolitan region, if possible, under Boroondara or Bayside city council, preferably early in the year with a decent to high number of rooms and bathrooms. For the best returns on investment, the preferred real estate agent when selling the house is Marshall.

Appendix

Code reference: https://github.com/abhinav-sharma-6167/DS-Python-Project/blob/main/Python%20Project%20-%20Melbourne%20Housing%20Analysis.ipynb

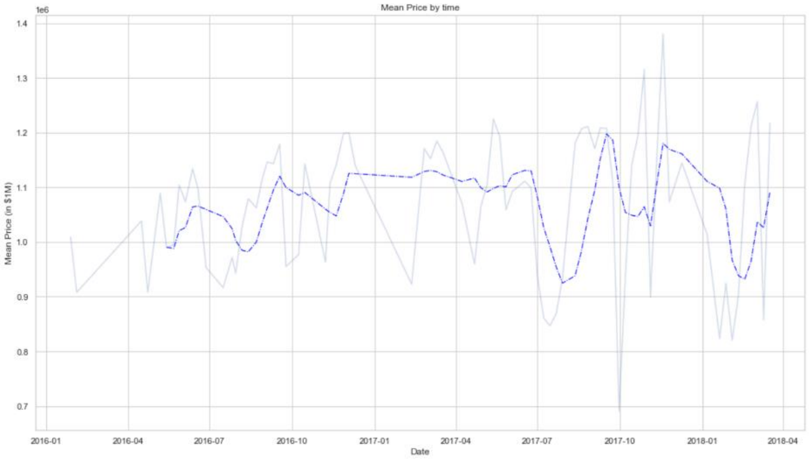

Figure 1

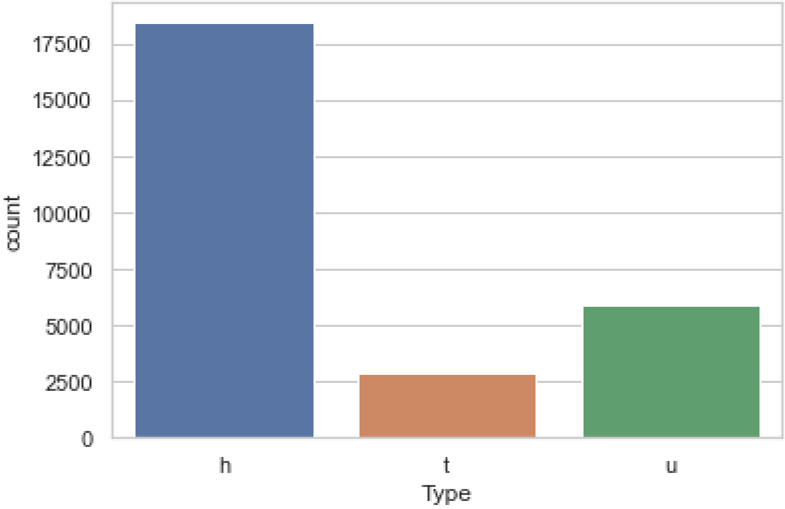

Figure 2

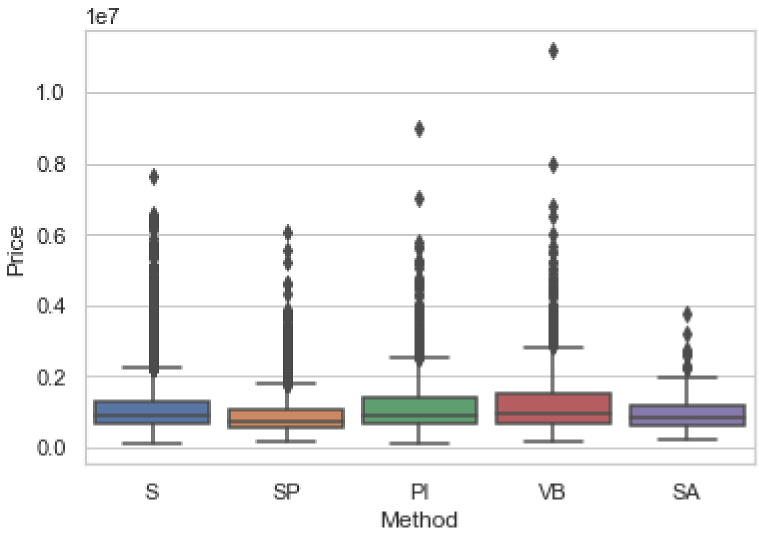

Figure 3

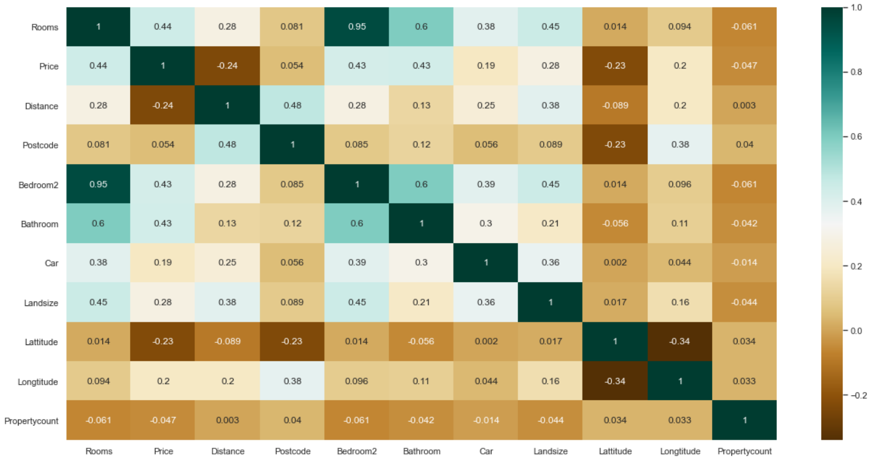

Figure 4

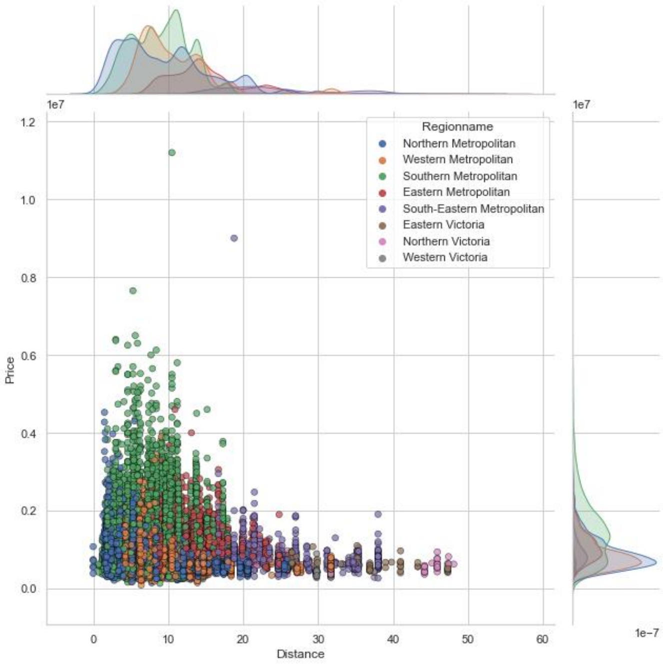

Figure 5

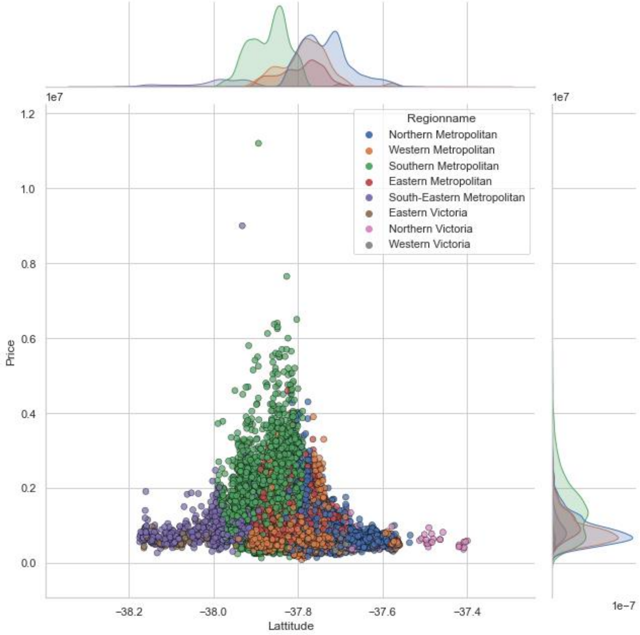

Figure 6

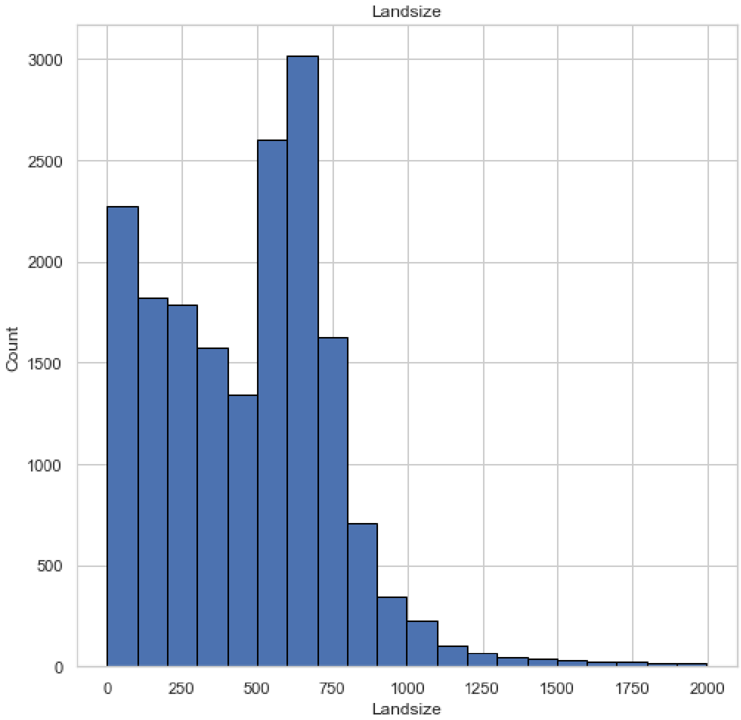

Figure 7

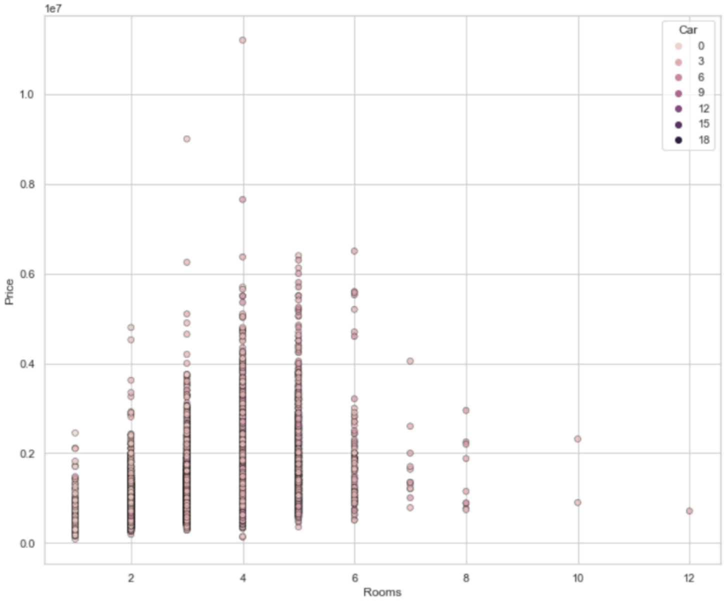

Figure 8

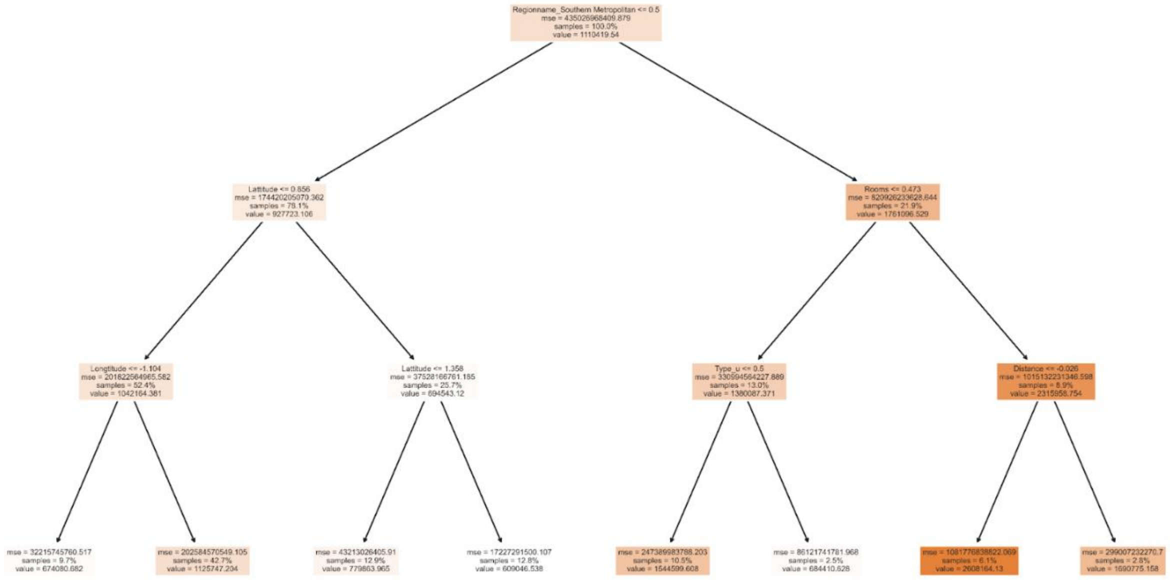

Figure 9

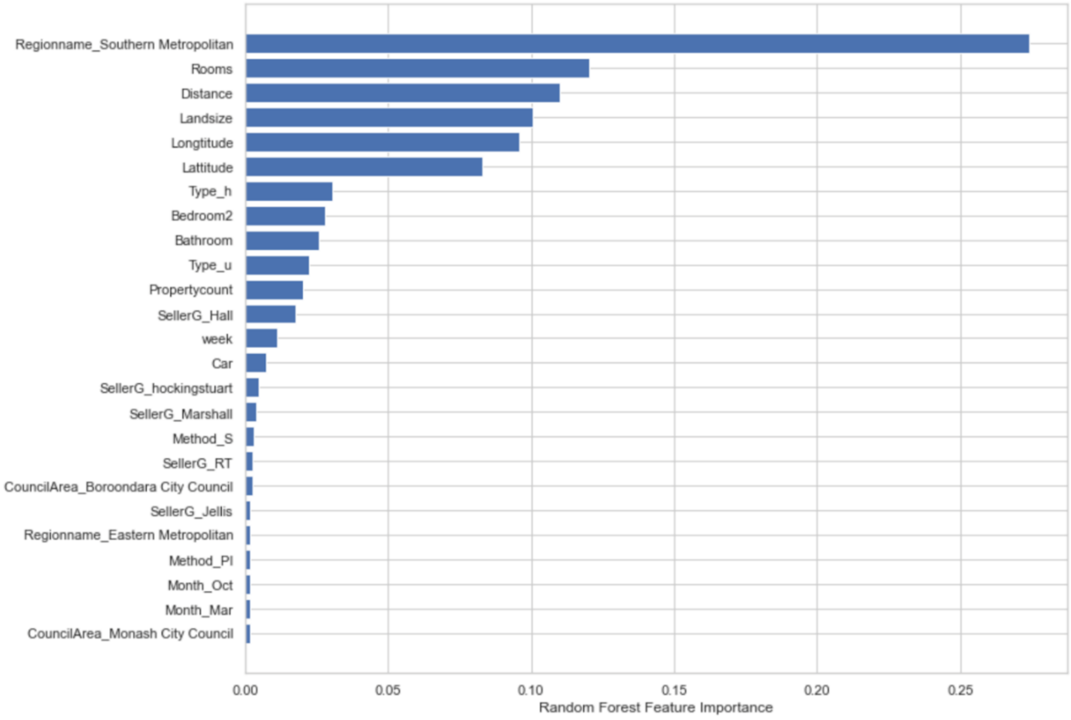

Figure 10

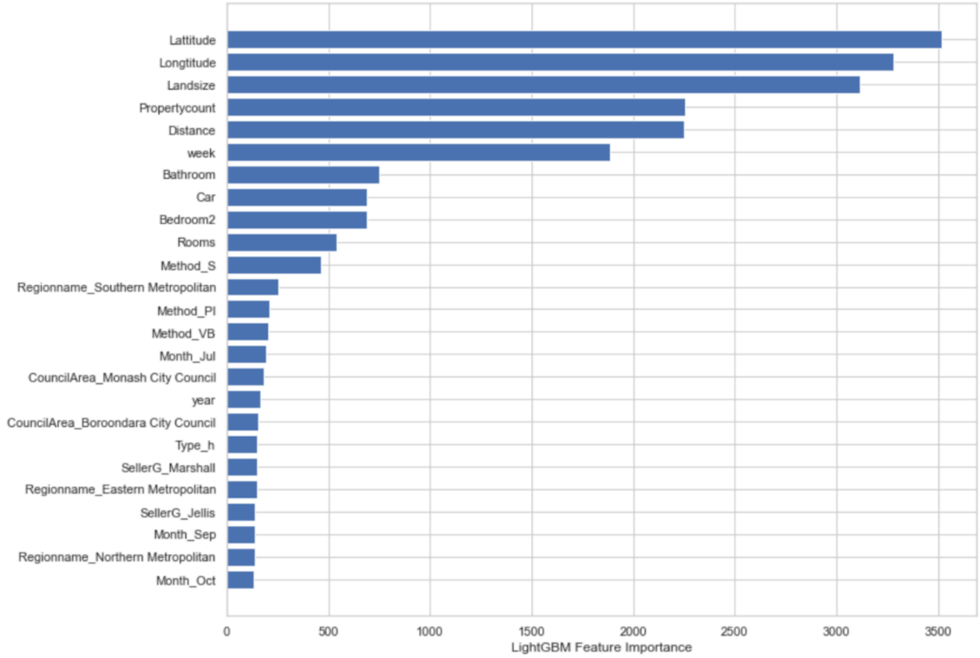

Figure 11

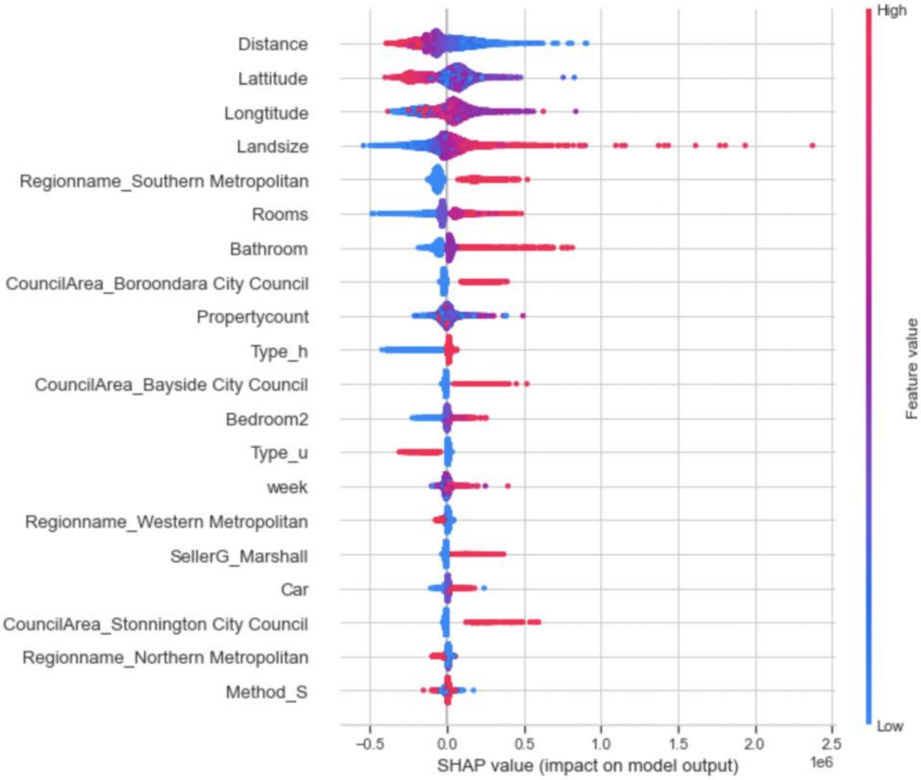

Figure 12

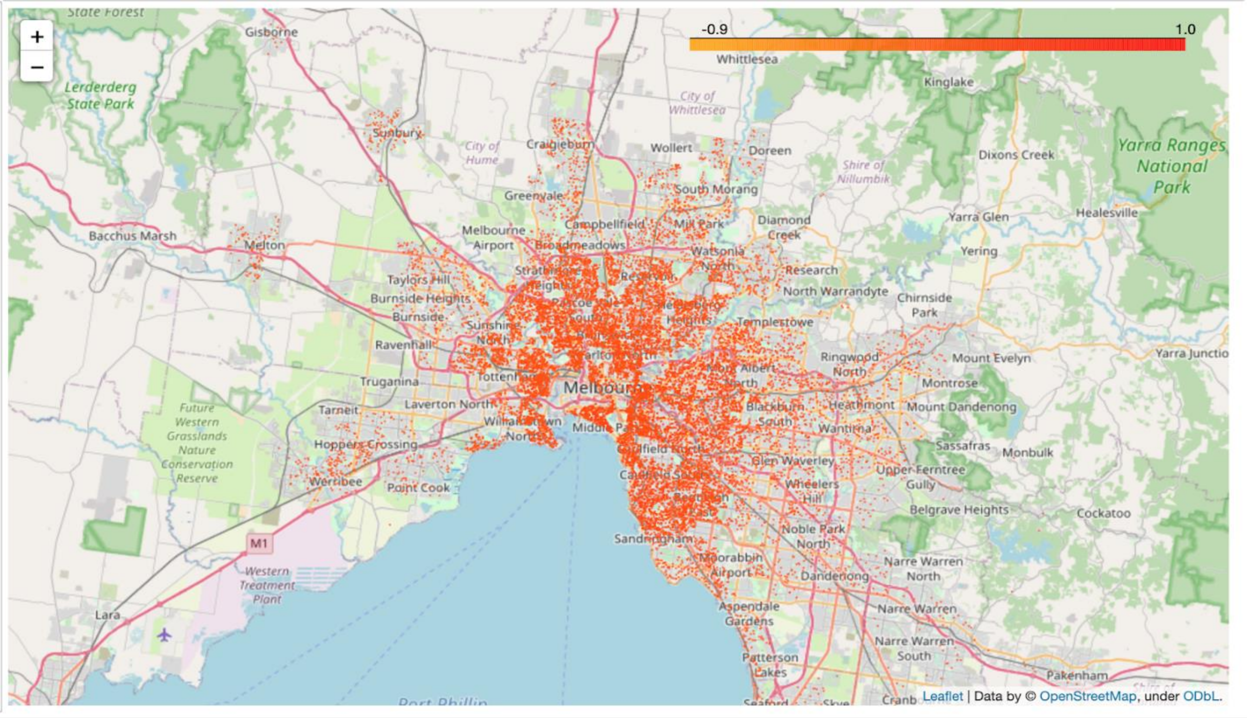

Figure 13

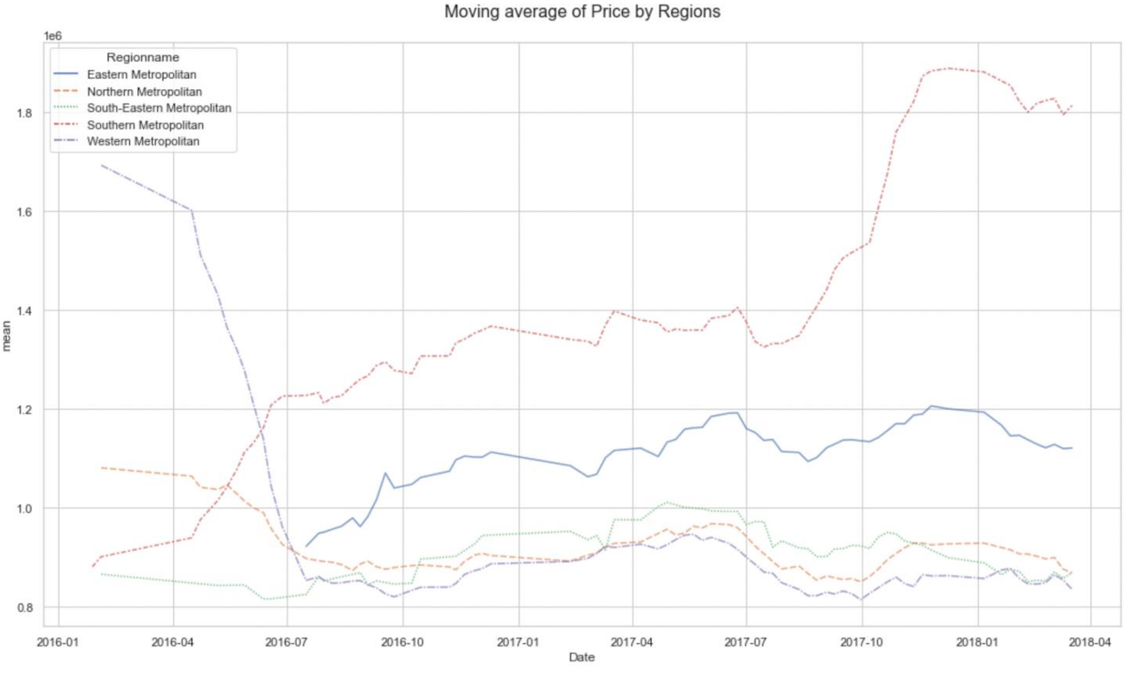

Figure 14