## **Melbourne Housing Analysis Report**

## Problem Description:

## 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 to not be severely affected due to the recession, it did fall prey to the universal housing crisis. Post 2008, the housing prices in Australia were increasing at a higher rate relative to income growth. Given the rise in house pricing at an alarming rate, some economists have theorized that the Australian property market, especially in cities like Melbourne, has been significantly overpriced and should be due for a price correction. The dataset of Housing Market Data in Melbourne allows us to understand the patterns of house prices during 2016-18 by taking various factors into consideration. We try to answer the following questions from the perspective of a real estate investment firm :

## · How have the prices changed in recent years?

## · Which sectors are the more expensive vs reasonably priced?

## · Which property is the best suited as an investment opportunity?

## Importance of the problem:

## The city of Melbourne is currently experiencing a housing bubble which economists expect to burst soon. Using the dataset and modelling, we are trying to discover trends within the data that could help us understand which areas are the most cost effective. We chose this data problem as while it does deal exclusively with Melbourne, housing bubbles are ubiquitous. Using this dataset, we may be able to understand the patterns behind house pricing. Analysis and modelling of the data would not be beneficial or limited to Australian house markets but may help one with house pricing within other countries too.

# **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 if the market is cooling; 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 to extract non-linear seasonality of the time series, if present. We then partition the data by mid-2017 to form our train (n=10,134) and test (n=7,563) sets. The idea is to check model R2 for all our models on the same test (hold-out) to ensure comparability. 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). We then go on to see what features are best drivers for the Price.

Given both best performing models are tree-based, we plot a decision tree (Figure 9) to see 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 in deriving property pricing. For RandomForest and LightGBM, we explore multiple model explainability strategies and align on using SHAP importance given it is model agnostic and provides directionality in insights.

The first few important variables 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 it’s expected price. Intuitively, the larger the Landsize, the higher the house price. Similarly, properties in the Southern metropolitan region are associated with positive shap value, consecutively signifying an increase in price. This is expected given EDA of Price vs regionname showed southern metropolitan prices being relatively higher. Features such as #Rooms, #Bathrooms also show unidirectional impact on price with price increasing with higher #Rooms. While property count comes high up in importance, SHAP doesn’t show a clean 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 Regionname came up as one of the most important parameters with clear directionality, we decided to split up our stable time series plot by region to see for any patterns. We see the market for the Southern Metropolitan region has been booming whereas the housing market for all other regions was cooling off. This explains the stable overall line chart as market dynamics are opposite in movement in different regions. Surprisingly, we do not see a lot of seasonality in our time series as the month flag variables do not come up in the SHAP summary plot. We do see the week (week of the year) variable suggesting there is some increasing trend in prices during the course of an year irrespective of region, house-type or real-estate agents.

Our recommendation based on the overall analysis would be to invest in a house in the Southern Metropolitan region, if possible under Boroondara or Bayside city council with decent to high #Rooms, bathrooms etc. ideally at the start of year. For best returns, the preferred real estate agent while selling the house is Marshall.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Melbourne Housing Analysis Report**

## 

## Problem Description:

Since 2000, house prices in Australia have appreciated by 150%. Given the rise in house pricing at an alarming rate, some economists have theorized that the Australian property market, especially in cities like Melbourne is significantly overpriced and is due for a price correction. However, the prices have only increased since 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 can try to answer the following questions:

## How much have house prices increased in recent years?

* What are the best areas for investment in Melbourne?

## Importance of the problem:

## The city of Melbourne is currently experiencing a boom in housing prices which against economist’s beliefs is only going to rise in the coming years. Using the dataset and modelling, we will try to discover trends within the data that could help us understand which areas are the most cost effective. We chose this data problem as while it does deal exclusively with Melbourne, real estate investment is ubiquitous. Using this dataset, we may be able to understand the patterns behind house pricing. Analysis and modelling of the data would not only be beneficial or limited to Australian house markets but may help one with house pricing within other countries too.

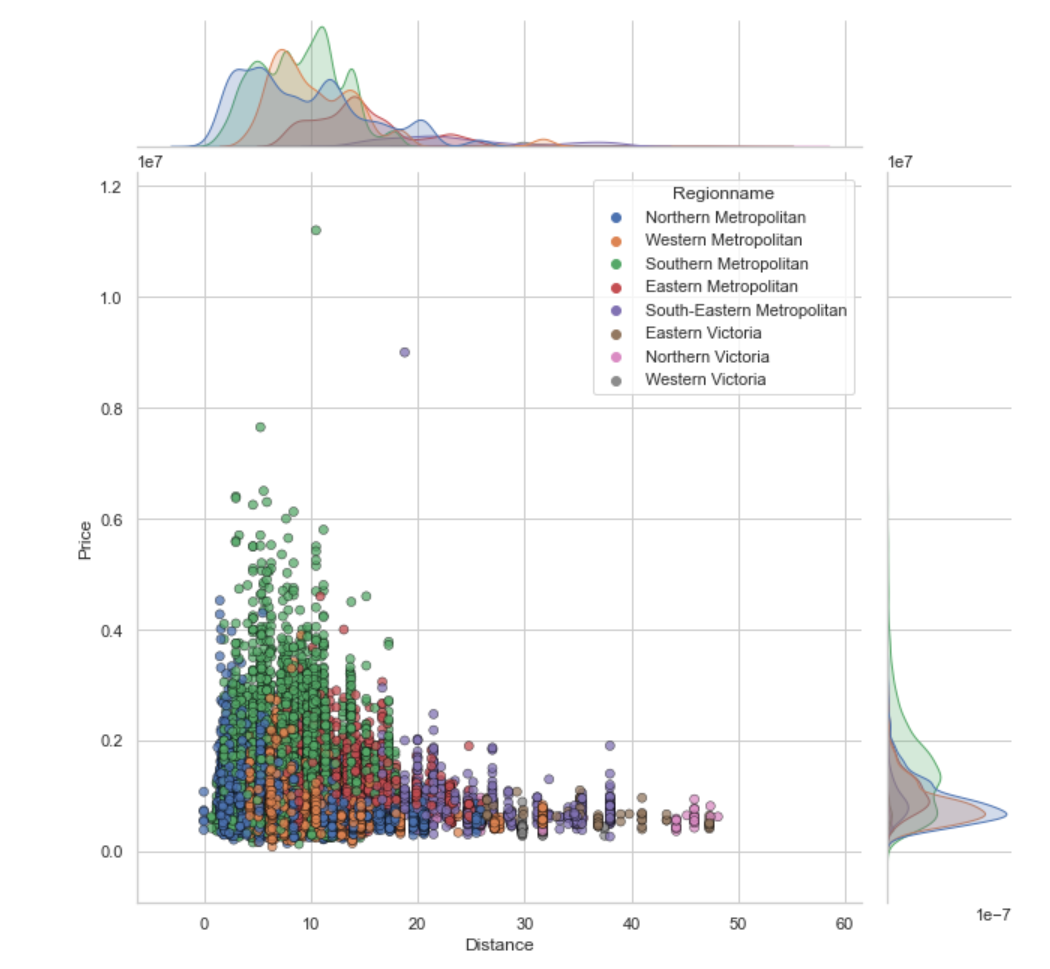
# **Exploratory analysis**

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, 14 variables had missing values. After computing the table of missing values, we decided to not 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 as 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 as well. 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 of price based on a time series from 2016 to 2018.

From the graph, we do not see a clear pattern to demonstrate if the market is cooling; the graph shows a stable market where prices have been moving within the same level for years.

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



We see the joint distribution plot of price vs distance, 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 scatter plot of rooms 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.

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. We also feature engineer temporal components such as week, month and year from the Date variable. We use year and week as continuous to capture time series trend and use month as categorical to extract non-linear seasonality of the time series, if present. We partition the data by mid-2017 to form our train and test sets. The idea is to check model R2 for all our models on the same test (hold-out) to ensure comparability. 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). We then go on to see what features are best drivers for the Price.

1. questions - needed answer: how's the price moving recently put the first graph of the timeline vs price ---- we see that there is no obvious pattern since the price is relatively stable(even though it fluctuate for a period) 2. we wonder whether or no the graph reflect the true situation or is there some information hidden behind the graph --- and then we did EDA for the variables to understand them better(include remove outliers, missing values, and we found some variables have stronger linear relationship with price and so on)

3 modeling ----- we found RF and boosting give us highest accuracy so we can trust these models and then use these model to find the importance of variables (some process of how to separate the variables include separate region into southern , etc.) 4 we found region-- southern is the most importance qualifier for price ---- so that we deep dive the timeline graph (here goes for the second timeline graph) and see southern region prices rising whereas market cooling off for remaining regions ( and it explains why the first graph looks like stable because southern goes up and others go down --- they balance) 5 conclusion ---- the market for other region is cooling off except southern and if people want to invest in Melbourne, he should invest in the southern area (within this area , we can talk about the first 6 importance variables such as the closer the distance the better or somethings like this)

**Appendix**

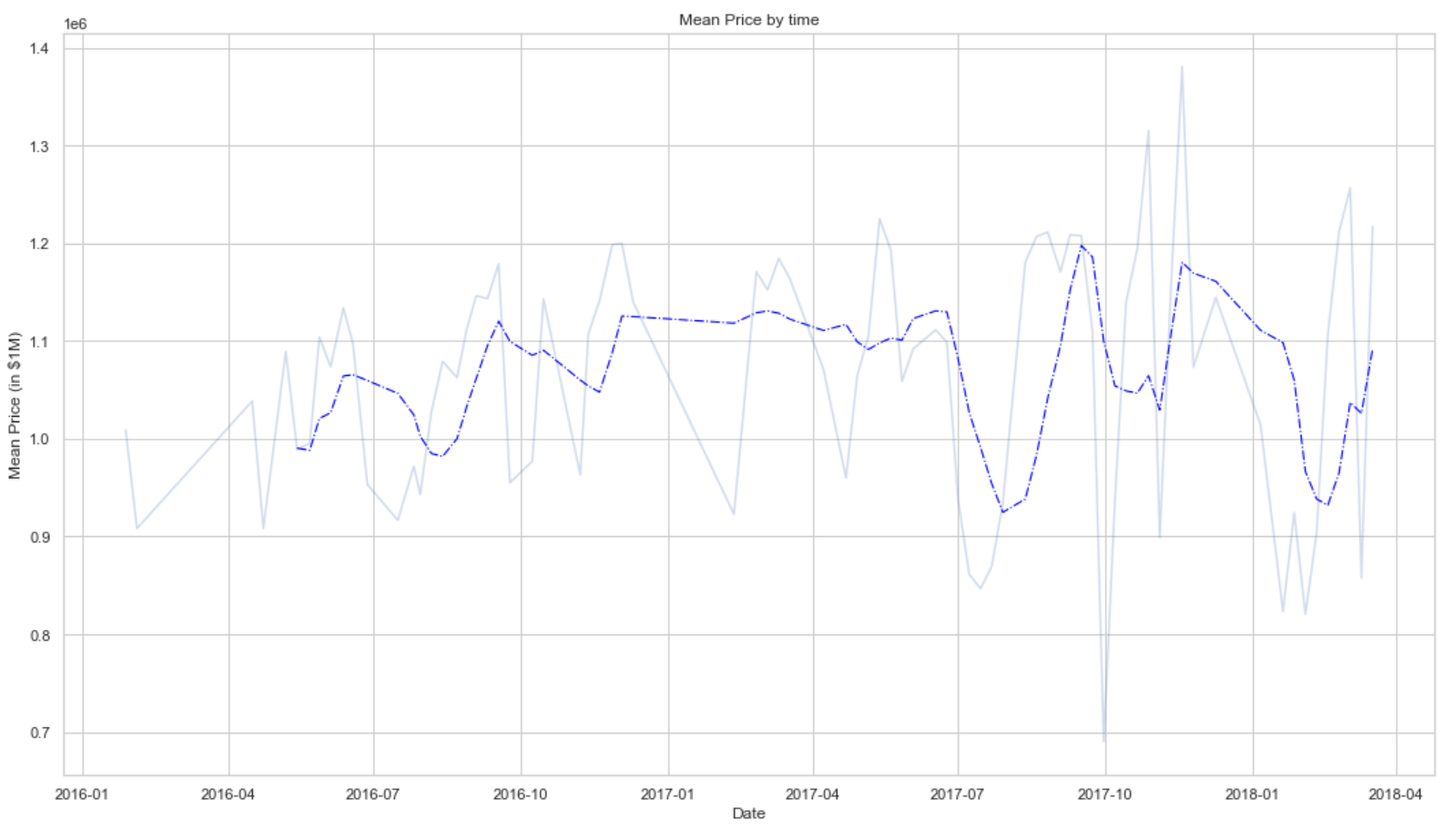


Figure 1

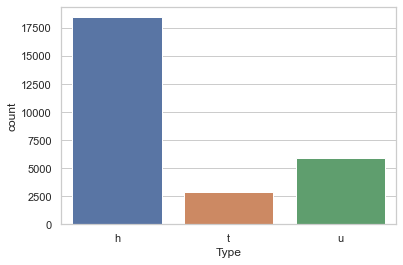


Figure 2

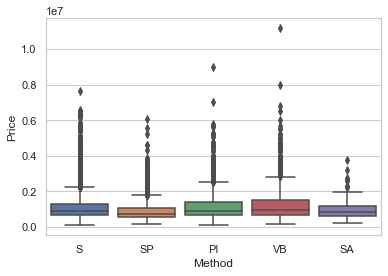


Figure 3

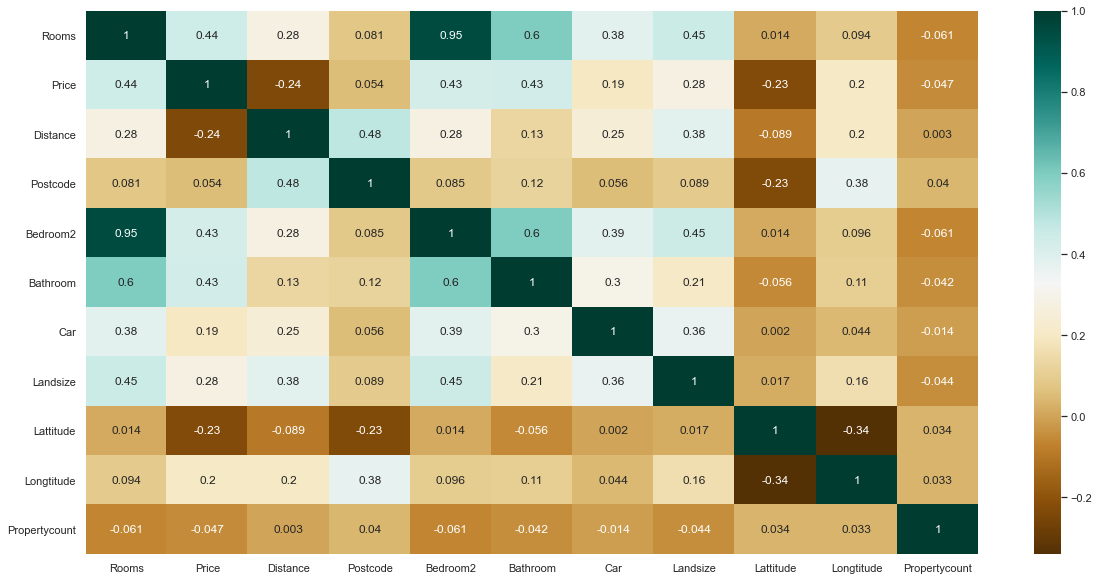


Figure 4

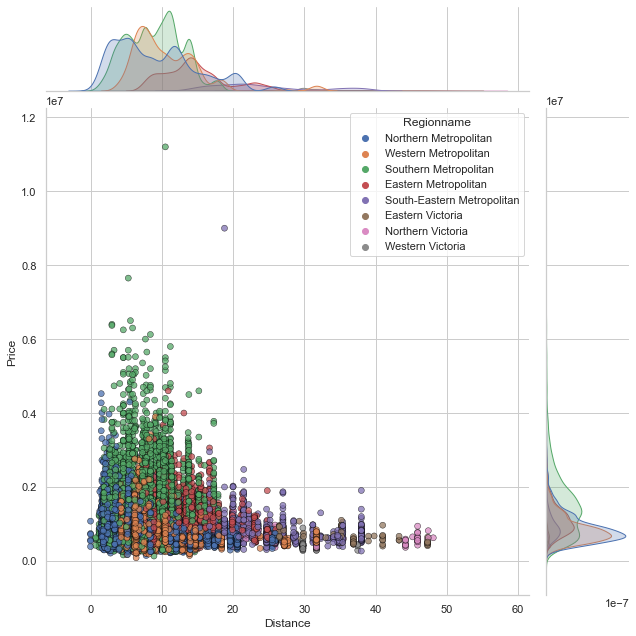


Figure 5

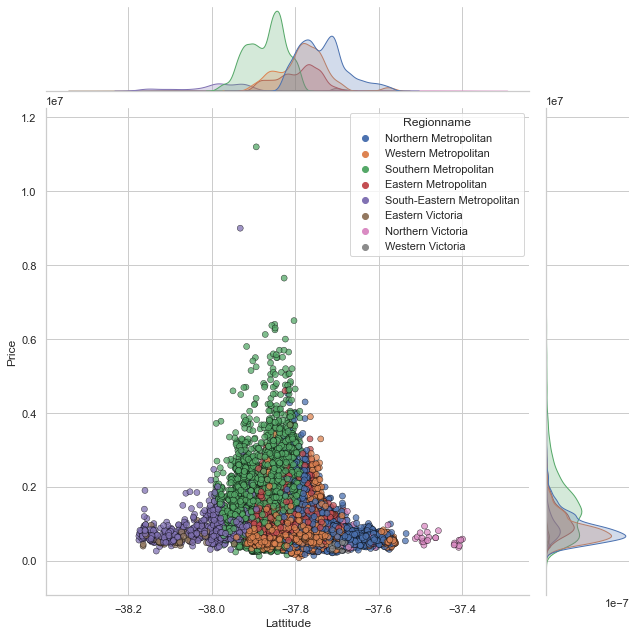


Figure 6

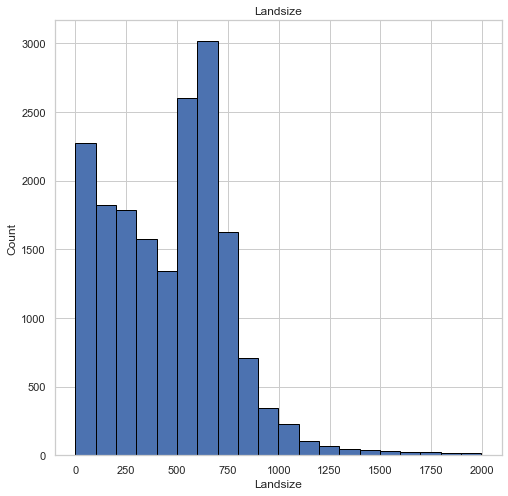


Figure 7

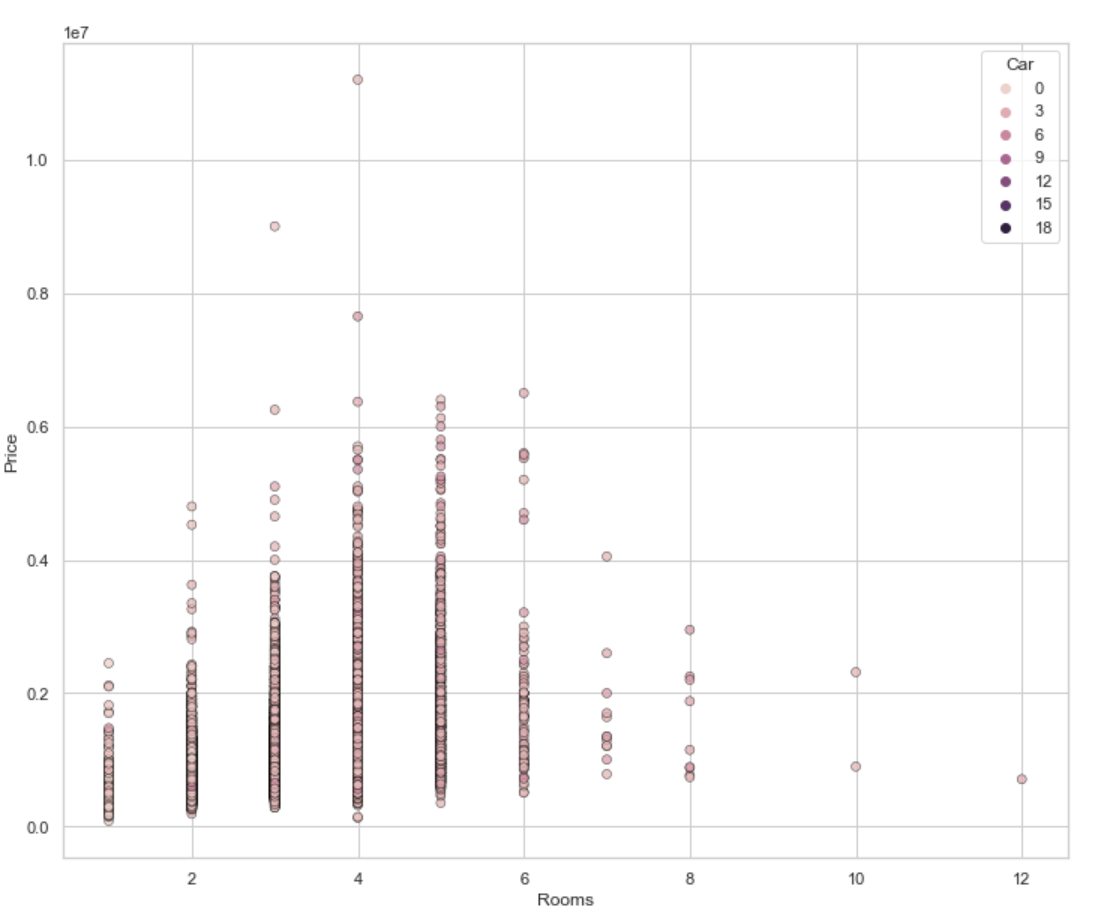


Figure 8

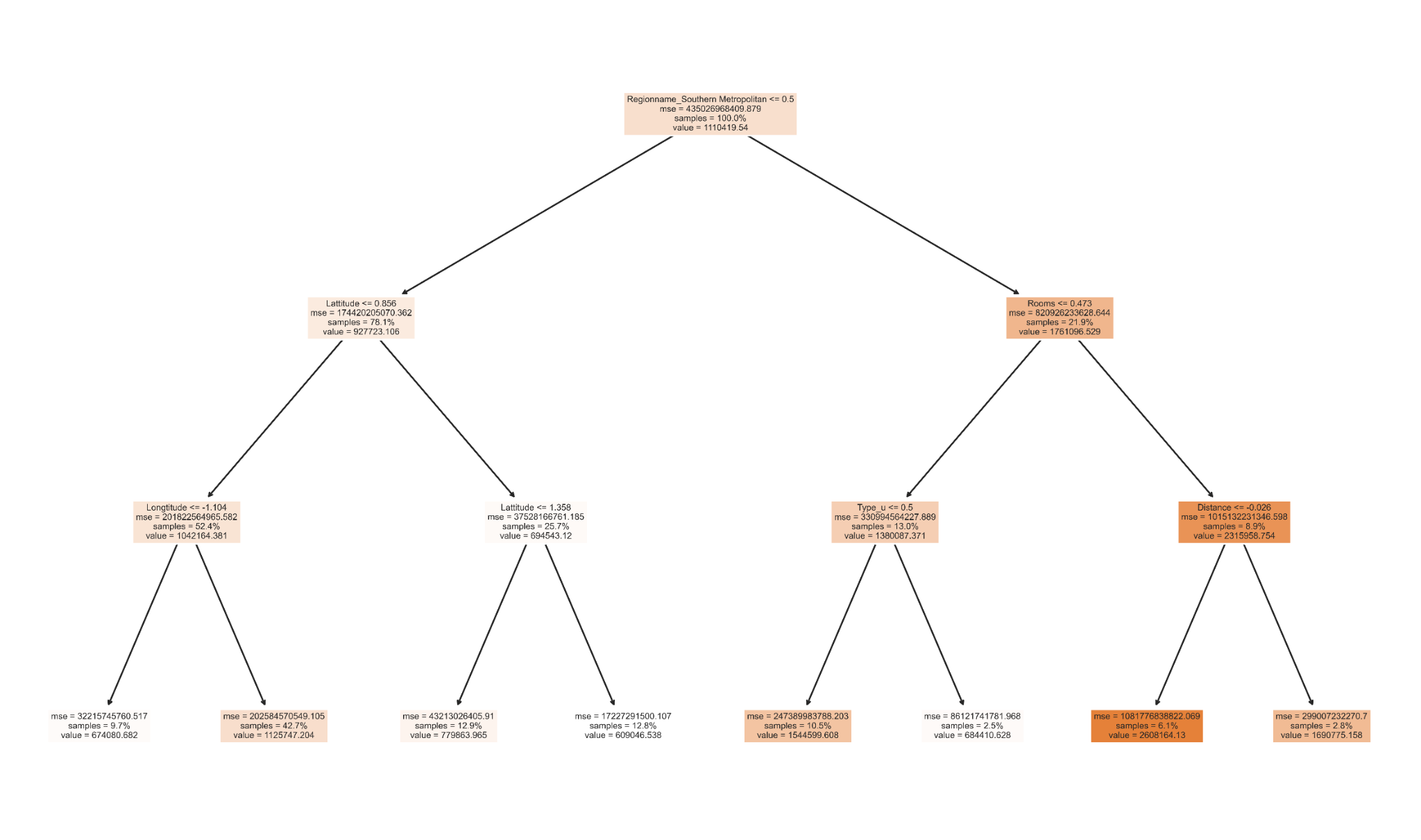


Figure 9