Assignment 1 Report

Student name: Jielin Feng, Zedong Wen
SID: 480019484, 500334863

Problem

In recent years, house purchasing had become a significant event that most people need to deal with. It is worth noting that the housing price may be affected by various factors (Wen & Goodman, 2003). Studying the housing price plays an important role for not only buyers but also sellers. Since housing prices can bring huge influence to the real estate market, which is a primary economic market in Melbourne, house purchasing strongly influences the economic development of the country (Saad, Alqatan & Arslan, 2021). The purpose of this project is to analyze the real historical transaction data set (Melbourne Housing data set) (Kaggle, 2018) to understand in-depth which factors have a profound impact on Melbourne housing prices, and on this basis to find an efficient model to predict housing prices based on specific attributes. Meanwhile, it is expected to forecast the future housing price in various regions within 10 years.

Based on the data from reliable sources, the problems need to be solved are: 1. How to find and simplify the key attributes related to the changes in housing prices; 2. How to use these data to effectively forecast housing prices (based on different types of methods); 3. How to improve the accuracy of prediction models; 4. How to forecast the trend of housing prices in different regions in the next 10 years based on the available data.

After obtaining an effective prediction model, the model can enable buyers and sellers to make better decisions. In addition, it can formulate corresponding policies for the state to regulate and control housing prices and market management in different regions.

Data：

The data selected was retrieved from "https://www.kaggle.com/dansbecker/melbourne-housing-snapshot", it was created by Tony Pino (Kaggle, 2016), then processed further by DanB (Kaggle, 2018). The information contains in the dataset was collected from publicly available results posted every week from Domain.com.au ("Domain.com.au | Real Estate & Properties For Sale & Rent", 2021). The dataset was downloaded in" CSV" format, and the size is approximately 2.1 MB. The actual dataset contains 13580 instances described by 21 attributes list from left to right: **Suburb**, **Address**, **Rooms**, **Type**, **Price**, **Method**, **SellerG**, **Date**, **Distance**, **Postcode**, **Bedroom2**, **Bathroom**, **Car**, **Landsize**, **BuildingArea**, **YearBuilt**, **CouncilArea**, **Lattitude**, **Longitude**, **Regionname** and **Propertycount**. Further information and explanation of each attribute are provided in supporting notes (Kaggle, 2016).

The data is then uploaded into R for further pre-processing and analysis. After importing the data, data cleaning was applied. As can be shown in Appendix 1, 11,887 values are missing. Among the missing values, 62 are from **Car**, 6,450 are from **BuildingArea**, and 5375 are from **YearBulit**. Since the missing values of **BuildingArea** and **YearBulit** are significantly large, which occupy 47.50% and 39.58% of the column respectively. To avoid bad accuracy caused by the existence of too many missing values in further study, it was decided to directly remove these two attributes. While the missing values of **Car** are 62, which only occupy 0.46% respectively, the missing values for **Car** are relatively small, and it is inappropriate to use interpolation to fill in the data as the value of car attribute is between 1 and 5. Therefore, 62 rows with

the missing value of **Car** attribute were cleaned. The final cleared data set contains 13518 rows and 19 attributes. Some examples of these missing values are shown in Appendix 1.

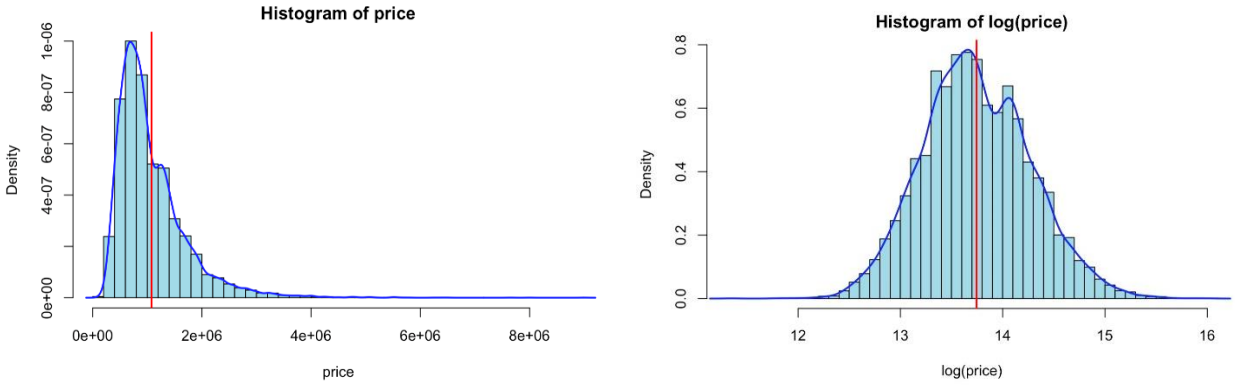

Figure 1-(a) Histogram of price and (b) Histogram of log(price)

Fig. 1 (a) and (b) show histograms of price and log (price). From Figure 1(a), the distribution of price is right skewed. It varies widely and has a long tail. The mean for housing price in Melbourne is $1074796. When housing price exceeds a certain level, the higher the housing price, the fewer people are willing to pay for it. Normalisation is applied by log the price as Figure 1(b) shows. The log (price) is considered to have a normal distribution. Therefore, log (price) will be used as the target variable in the stage of model building and evaluation.

To facilitate correlation analysis, the data set is divided into categorical data and numerical data as can be shown in the data summary table in Appendix 3. In order to better discover correlation between each attribute, categorical data is transformed into numerical data. The correlation coefficient obtained by Pearson correlation analysis of the processed data set is shown in Appendix 2. It can be seen that **Price** is significantly positively correlated with **Rooms**, **Bathroom**, and **Bedroom2**; slightly positively correlated with **Car** and **Postcode**; and slightly negatively correlated with **Distance**. It is evidence that price is correlated with some attributes, deeper analysis of correlation will be further explored in project 2.

Proposal

For stage 2 of the project, we are aiming to propose an optimal housing price prediction model based on various types of variables and visualise the results. The specific steps can be explained as follows:

1. Find out the factors affecting housing prices;
2. Simplify the characteristics of the factors that affect housing prices (feature selection);
3. Use different methods to establish several housing price forecast models, for instance regression is considered to be used as one of the methods.
4. Evaluate the effectiveness and accuracy of the model;
5. Forecast the future trend of housing prices in different regions based on spatial distribution, time series forecasting is considered to be used.

To reach the goal, the data set will be divided into smaller sets (training set and test set) for establishing and verifying the model. Several algorithms methods are considered, for instance, regression, neural network and XGBoost, which will be further developed in stage 2 of the project. In order to propose optimal models, produce accurate Statistics information and display quality visualisation, Python is considered to be used as the major tool in stage 2 of the project.

Reference

Domain.com.au | Real Estate & Properties For Sale & Rent. (2021). Retrieved 13 September 2021, from https://www.domain.com.au/

Kaggle 2016, *Melbourne Housing Market*, Retrieved 14 September 2021, from https://www.kaggle.com/anthonypino/melbourne-housing-market

Kaggle 2018, *Melbourne Housing Snapshot*, Retrieved 14 September 2021, from https://www.kaggle.com/dansbecker/melbourne-housing-snapshot

Saad, A. B., Alqatan, A., & Arslan, M. (2021). House Price Shock and Business Cycle: The French Case. *Scientific Annals of Economics and Business*, *68*(1), 115–127. https://doi.org/10.47743/saeb-2021-0007

Wen, H., & Goodman, A. C. (2013). Relationship between urban land price and housing price: Evidence from 21 provincial capitals in China. *Habitat International*, *40*, 9–17. https://doi.org/10.1016/j.habitatint.2013.01.004

Appendix 1-Missing data patterns

The charts in the following figure show patterns and counts of missing data in data pre-processing stage.

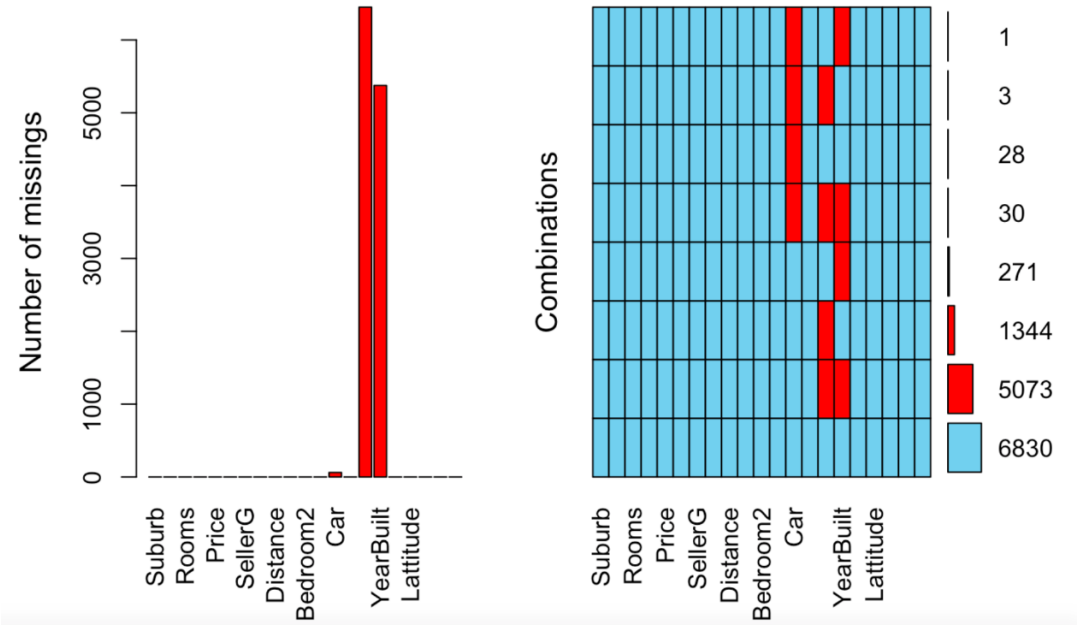

Figure 1 - Missing data patterns in data pre-processing

Appendix 2- Numeric variable correlation

The chart in the following figure shows correlations between numeric variable after pre-processing.

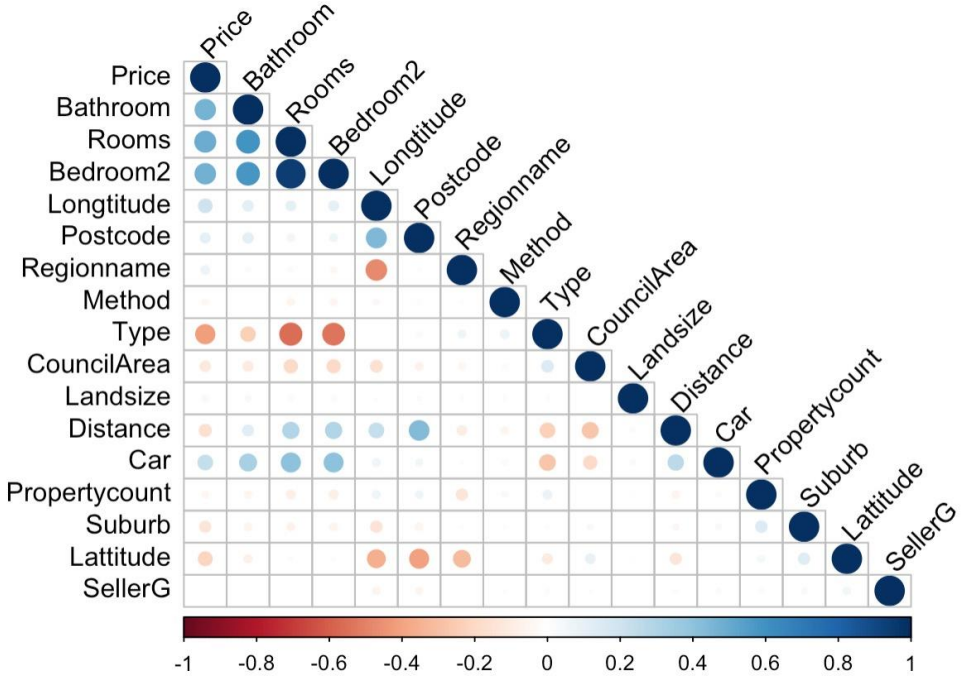

Figure 2- Numeric variable correlation

Appendix 3- R code

As can be shown in code.html.