# DATA ANALYTIC AND PRICE MODELLING USING ENSEMBLE METHODS FOR HIGH RISE RESIDENTIAL IN KUALA LUMPUR

MUHAMMAD IZZUDDIN BIN AHAMAD SFHAFI

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR THE MASTER OF DATA SCIENCE

2018

# Contents

# Table of Figures

# Abstract

House, most of the time is the most expensive purchase in a person's lifetime. Yet, amongst of the biggest challenges by individual purchaser or seller is to determine the  fair price of the real estate property. If buyer pay above the market price, they will end up with a bigger mortgage that can cost them a high monthly instalment, on the seller side, selling below market value can cost them of lower return of their investment..

The objective of this research project is to conduct data analytic study and build machine learning model that posses profound accuracy in determining the housing price. Around 39000 house's transaction from January 2008 to March 2018 for 953 high rise residential in Kuala Lumpur are extracted. Python programming language was used in this research from data acquisition, data cleaning, data exploration, data visualization and finally data modelling.

Our end goals is to investigates factors that influence the housing price and build several ensemble machine learning models, compare their accuracy and propose the best model that can predict the future housing price based on the  model performance.

# 1. Introduction

## 1.1 Research Objective

Over the past several decades, machine learning model has been utilized extensively to predict real estate property value all over the world. The recent high growth of machine learning utilization especially in the artificial neural network (ANN) was credited to the increase of high computational power and  exponential growth of data. Yet, since from the last decade, there was not many substantial local research that utilize machine learning technique in the context of Malaysia's housing market that can be found online.

For the last 10 years, the change of the market and economic cause the inconsistent price of housing. In 2008 to 2014, the Malaysia House Price Index rise from around 1% to be around 12% in 2014 and drop

to 4% in 2015 before hovering around 6% in the recent year (Malaysia House Price Index, n.d.) In 2016, the volume of residential transaction drop from 14% while total transacted value drop to 11% compare to the year before. This cause predicament amongst home buyer and seller on how to accurately assess the price of a house and what type of features that determine the desirability of a certain house compare to others on the market

The diversity of housing characteristics pose challenges to determine accurate price for the a particular property. The price of a property is substantially influenced by it own unique set of structural features(Completion Date, level of storey, total number of block, Tenure, gymnasium, swimming pool, level of security), location features(distance to amenities such as shopping mall and highways) , neighbourhood features (crime rate, population density and income level) and macroeconomic features (Overnight Policy rate, Inflation rate, unemployment rate)

To develop a model that can accurately predict a price of a house is difficult due to numerous of reasons. First, it is not easy to acquire or access the data set of the mentioned attributes above to train the model. Second, the value of house may not be consistent due to specific features across different area or geographic regions. For example, people in the rural tend to prefer a terrace house with a land area compare to people living in the city, which tend to prefer high rise property due to the price and surrounding facilities. Due to these reasons, the machine learning  model tend to be developed focusing on a specific regions. Under this project, Location features will be included to account the spatial effect.

Despite of all these obstacles, previous research has shown that it is possible to assess the actual value of a house. The objective of this research is to investigate that features that contribute in determining the value of a house and develop a model that can predict the future value of a house with a profound accuracy

# 2. Literature Review

## 2.1 Previous Research Review

Hedonic regression model has been widely used in since 1970's to develop prediction model for housing value. Hedonic pricing model allow the housing price to be assessed internally (structural) and externally(location, neighbourhood and macroeconomic) by breaking down the heterogenous attributes of a house into individual components.

Kain and Quigley (Quigley, 1970) develop hedonic model that include the structural features and location features of the house and study the relationship between the housing price and the house characteristic. While Kenskin (Kenskin, 2008) improve the model by including socio-economic characteristics such as neighbourhood quality satisfaction, distance to amenities and facilities.

Malpezzi (Malpezzi, 2003) stated that hedonic method is one of the way that a house price can be evaluated. In which, by decomposing the heterogenous features of a house into measurable prices and quantities. Therefore using, hedonic method and regression analysis can help to investigate how each attribute help to determine the final overall price. Maelpezzi suggested hedonic approach also consider the age of the properties in which time series attribute also being taken into the consideration.

Typically, hedonic model does account for spatial factors as well, Basu and Thibodeau (Basu, 1998) explain that spatial dependence exist for a few reason. For example, houses in a same region tend to share similar characteristic such as similar design structure. Another reason, house in the same region have the similar distance to the surrounding amenities and other central business district. All of these spatial factor contribute to correlation between the features stated and the final price of the house.

However, technique mention above applied the multiple regression analysis were criticized in other researches. Mainly due to the potential problem relating to the model estimation and assumption. This is amongst the factors of the rise of ensemble method. The idea of ensemble models is to combine multiple independent model and averaging their result to improve the accuracy. Ensemble models can be categorized as three. Bagging, used to reduce variance of the prediction model. Boosting, is aimed

on transforming a collection of weak learner into a strong predictor. Finally, stacking, aimed to decrease prediction biased in addition to variance (Yakov Frayman, 2007)

Fan, Ong and Koh (Gang-Zhi Fan, 2006) stated that multiple regression framework pose problem such as market disequilibrium, selecting independent features and identifying supply and demand. Fan, Ong and Koh claimed that  one of the bagging method, tree based model approach is a better alternative to hedonic multiple regression for predicting housing price. This claim was further supported by Ceh, Kilibarda, Lisec and Bajat (Marjan Ceh ˇ, 2018) which showed that tree based approach model (Random Forest) performed significantly higher compare to hedonic multiple regression model.

For boosting algorithm, XGBoost is the recent boosting ensemble method that raise to fame due it highly scalable end to end tree boosting system and novel sparsity aware calculation (Tianqi Chen, 2016). This features lead to fast execution speed and better performance compare to other model. We will apply XGBoost in our study despite there is not much industrial application that can be found.

For Stacking, Frayman, Rolfe and Webb claimed that stacked ensemble performed better compared to the conventional machine learning model (Yakov Frayman, 2007)  Frayman, Rolfe and Webb achieved better performing model by combining linear regression, multilayer perceptron, logistic regression and KNN as the based model and used Multilayer Perceptron as the final generalization model. For the selection of the based model for the stack ensemble, we will used the base model similar to Frayman, Rolfe and Webb's.

# 3. Methodology

## 3.1 Area of study

Kuala Lumpur is the national capital of Malaysia. The boundary of the city encompasses of 243km2. It is located between the confluence of Klang and Gombak rivers, and situated in the centre of Selangor state. As 2016, the city cater the population of 1.76 million people in the area of just 94 kilometer. This translate to a very high population density of 17,310 people per square mile of area. By 2017, almost 464,000 thousand units are available in Kuala Lumpur area. This research will solely focus on the high rise residential units available in Kuala Lumpur

## 3.2 Data Acquisition (Web Scrapping)

### 3.2.1 Master Data's Scrapping

The Web Scrapping Process was automated using python programming package Selenium and Beautiful Soup. From www.brickz.my/transactions/residential/kuala-lumpur/non-landed/?range=2008+Jan-,information such as Name, Tenure, Median Price PSf and Number of Transactions are derived from January 2008 to March 2018. 1037 names of high rise residential was obtained from this first stage of web scrapping activity.



| Project/Township | Tenure/Type | Median Price Psf ⓘ | Median Price | Filed Transactions˅ | | |
|---|---|---|---|---|---|---|
| **OUG PARKLANE**<br>OLD KLANG ROAD, KUALA LUMPUR | FREEHOLD<br>SERVICE RESIDENCE | 442 | 420,000 | 437 Transactions | ♥ | ✉ |
| **THE SCOTT GARDEN**<br>OLD KLANG ROAD, KUALA LUMPUR | FREEHOLD<br>SERVICE RESIDENCE | 604 | 470,000 | 338 Transactions | ♥ | ✉ |
| **VERVE SUITES**<br>MONT KIARA, KUALA LUMPUR | FREEHOLD<br>SERVICE RESIDENCE | 1,060 | 698,000 | 287 Transactions | ♥ | ✉ |
| **SRI PENARA**<br>CHERAS, KUALA LUMPUR | LEASEHOLD<br>APARTMENT | 237 | 153,000 | 283 Transactions | ♥ | ✉ |
| **REGALIA RESIDENCE**<br>KL CITY CENTRE, KUALA LUMPUR | FREEHOLD<br>SERVICE RESIDENCE | 846 | 490,000 | 240 Transactions | ♥ | ✉ |

*Figure 1: All the names of high rise residential in KL (Jan 2008-March 2018) from brickz.com*

Table 1: Result of Web Scraping process on KL High Rise Residential

|  | Name | Type | Street | Tenure | No of Transaction | Median Price Psf |
|---|---|---|---|---|---|---|
| 1. | OUG Parklene | Service Residence | Old Klang Road | Freehold | 435 | 442 |
| 2. | The Scott Garden | Service Residence | Old Klang Road | Freehold | 604 | 604 |
| 3. | The Verve Suites | Service Residence | Mont Kiara | Freehold | 287 | 1060 |
| 4. | Sri Penara | Flat | Cheras | Leasehold | 232 | 279 |
| 5 | Regalia Service Apartment | Service Residence | KlCC | Freehold | 238 | 238 |

### 3.2.2 Transactions Web Scrapping

Each name from this 1037 property's name will be used to further extract individual property information such as SPA Date (Date of the Transaction), Address, Rooms, Lot Size and Price through the same method of web scrapping.



| 338 Transactions from Jun 2010 to Dec 2017 | | | | | | | 1 2 3 … 17 » |
|---|---|---|---|---|---|---|---|
| SPA Date^ | Address | Building Type | Tenure | Rooms | Lot Size | Price Psf | Price |
| 18/12/2017 | C-20-13, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 2 | 775 ft² | 813 | 630,000 |
| 04/12/2017 | A-12-12, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 764 ft² | 785 | 600,000 |
| 09/10/2017 | B-12-26, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 775 ft² | 516 | 400,000 |
| 09/10/2017 | B-16-6, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 764 ft² | 550 | 420,000 |
| 06/10/2017 | A-10-13A, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 775 ft² | 516 | 400,000 |
| 06/10/2017 | B-13A-11, JALAN KLANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 764 ft² | 628 | 480,000 |
| 05/10/2017 | A-18-35, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 1,389 ft² | 475 | 660,000 |
| 20/09/2017 | B-23A-18, JALAN KELANG LAMA | SERVICE RESIDENCE | FREEHOLD | 0 | 764 ft² | 825 | 630,000 |

Figure 2: Sample of Transaction record for a single residential name from brickz.com

*Table 2: Sample's result of the web scrapping activities of each individual properties from brickz.com*

|  | Name | SPA Date | Address | Rooms | Lot Size | Price |
|---|---|---|---|---|---|---|
| 1. | Oug Parklene | 27/02/2018 | D2-18-16, Jalan 1/152 | 3 | $950\ ft^2$ | 450,000 |
| 2. | Oug Parklene | 13/02/2018 | A-13-7, Jalan 1/152 | 3 | $950\ ft^2$ | 474, 800 |
| 3. | Oug Parklene | 17/01/2018 | D1-27-1, Jalan 1/152 | 3 | $950\ ft^2$ | 409, 000 |
| 4. | Scott Garden | 18/12/2017 | C-20-13, Jalan Kelang Lama | 2 | $775\ ft^2$ | 638,000 |
| 5. | Scott Garden | 04/10/2017 | A-12-12. Jalan Kelang Lame | 0 | $764\ ft^2$ | 600,000 |

### 3.2.3 Structural Information's Scrapping

Information such as year of completion, developer name, maintenance fee, total number of block, total number of storey and total unit available for a particular high rise property were scrapped from different property review websites (order is based on the google page ranking) such as;

- www.propsocial.com

- www.propwall.com

- www.durianproperty.com

- www.landbar.com

The web scrapping is done in the sequence of the google page ranking. Meaning that the web scrapping will start by scrapping the highest ranking property review's website, www.propsocial.com. Any information that is not available from this website will be then scrapped in the www.propwall.com and so on.
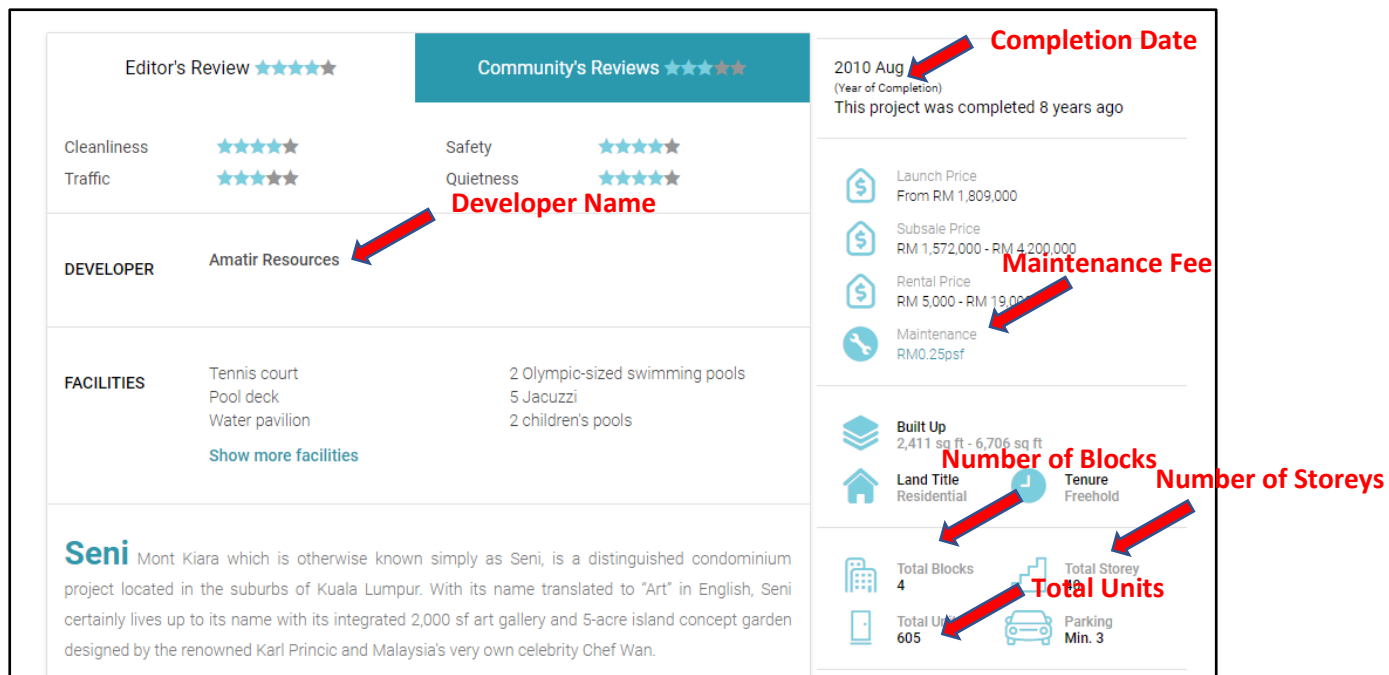
*Figure 3: Example of extracted information from www.propsocial.my*

*Table 3: Sample of Structural Information scrapped from the Property Review's Website*

| | Name | Completion Date | Developer | Maintenance Fee | Number of Block | Total Unit | Total Storeys |
|---|---|---|---|---|---|---|---|
| 1. | Oug Parklene | 2015 | Akisama | 0.18 | 11 | 4429 | 33 |
| 2. | Scott Garden | 2011 | Aston Villa | 0.15 | 3 | 1080 | 3 |
| 3. | Sri Penara | 2000 | PPC Glomac | 50/month | 4 | 2040 | 17 |
| 4. | Regalia | 2012 | PPC Glomac | 0.33 | 3 | 1033 | 17 |
| 5. | Verves Suites | 2013 | Bukit Kiara Properties | RM 0.33psf | 4 | 933 | 37 |

The accumulated transaction for all the individual high rise residential was 39,200 rows. For any further missing data that is not available through all the website mentioned above were scrapped from numerous forums and social medias. Although housing information acquired from the forum or social media is deemed lower quality and authenticity, we still implement these lower tier of information into our final dataset with the assumption that it is better than nothing.
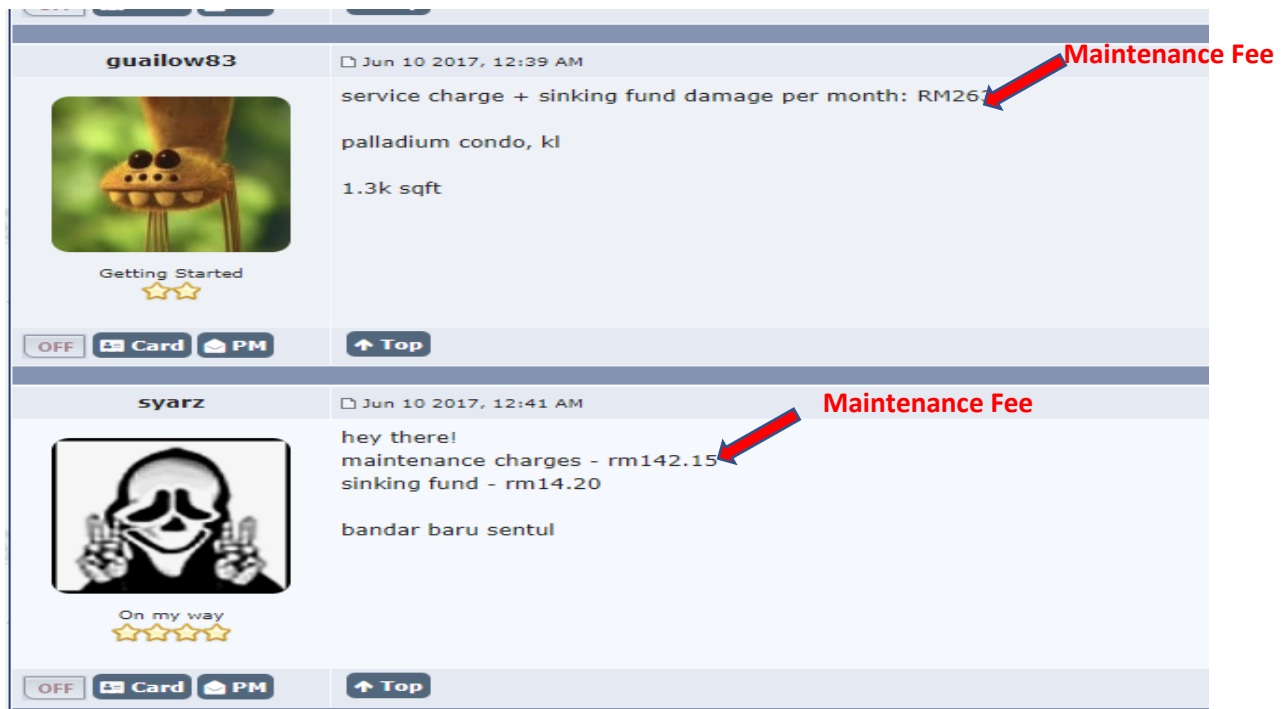
*Figure 4: Example of the maintenance fee acquired from lowyat.com forum website*

Any further leftover missing value will be fill up using imputation method discussed in the cleaning section. Due to the uncertainty of the contents and complexity of the website's structure. The information from these websites will be extracted manually instead of using automation.

### 3.2.4 Property's Coordinate

The spatial coordinate for each property was acquired from Google Map API service using Python looping process on each property's name from master data .

*Table 4: Coordinate Sample of Google Map API*

|   | Name | Latitude | Longitude |
|---|------|----------|-----------|
| 1 | OUG Parklene | 3.069821 | 101.6565 |
| 2 | The Scott Garden | 3.095009 | 101.6747 |
| 3 | Sri Penara | 3.164146 | 101.6926 |
| 4 | Verve Suite | 3.167698 | 101.646 |

### 3.2.5 Overnight Policy Rate

For macroeconomic factor, Overnight Policy Rate (OPR) is selected to study the macroeconomic factor toward housing price. OPR is the interest rate which financial institution lend to another institution overnight. The chain of OPR trigger the chain of event that effect the base lending rate (BLR). The BLR determine the interest rate of lending instrument such as property mortgage and personal loan.

| OPR Decision and Statement | | | |
|---|---|---|---|
| Date | Change in OPR (%) | New OPR Level (%) | Monetary Policy Statement |
| 10 May 2018 | 0 | 3.25 | View Statement |
| 07 Mar 2018 | 0 | 3.25 | View Statement |
| 25 Jan 2018 | +0.25 | 3.25 | View Statement |
| 09 Nov 2017 | 0 | 3.00 | View Statement |
| 07 Sep 2017 | 0 | 3.00 | View Statement |
| 13 Jul 2017 | 0 | 3.00 | View Statement |
| 12 May 2017 | 0 | 3.00 | View Statement |
| 02 Mar 2017 | 0 | 3.00 | View Statement |
| 19 Jan 2017 | 0 | 3.00 | View Statement |
| 23 Nov 2016 | 0 | 3.00 | View Statement |

*Figure 5: Sample of OPR rate from BNM's website*

The reason only single macroeconomic factor was selected due to the reasons of others macroeconomic indicator such as GDP and inflation rate was lagging indicator. Which mean that these indexes were formulated after the event has already occurred. The OPR acquired will be from Jan 2008 – March 2018

All the acquired data are then integrated and can be summarized as below

### Table 5: Type of Variables

| | Variables | Description | Format | Type |
|---|---|---|---|---|
| 1. | Name | Name of the property | Category | Neighbourhood |
| 2. | Type | Type of property- flat, apartment or condominium | Category | Structural |
| 3. | Address | Address of the property | Category | Neighbourhood |
| 4. | Street | Name if the neighbourhood | Category | Neighbourhood |
| 5. | Tenure | Freehold or leasehold | Category | Structural |
| 6. | Size | Size of the property | Integer | Structural |
| 7. | Room | Total number of room | Integer | Structural |
| 8. | Completion | Year of completion | Integer | Structural/Time |
| 9. | Developer | Name of Developer/ Developer reputation | Category | Structural |
| 10. | Maintenance fee | Maintenance charge per square foot area | Float | Structural |
| 11. | Block | Total number of block | Integer | Structural |
| 12. | Total Storey | Total number of Storey of particular building | Integer | Structural |
| 13 | Level Number | Level number of storey for a particular transacted property | Integer | Structural |
| 14. | Total Unit | Total unit available for a particular high rise property | Integer | Structural |
| 15. | Sale Date | Transaction sale date | Date | Time |
| 16. | Price | Sale price of the property | Integer | Target |
| 17. | Latitude | Latitude coordinate | Float | Neighbourhood |
| 18. | Longitude | Longitude coordinate | Float | Neighbourhood |

| | | | | |
|---|---|---|---|---|
| 19. | OPR | Overnight Policy Rate | Float | Macroeconomic |
| 20 | Number of Transaction | Total Number of Transaction for each residential | Integer | Structural/Time |

All the dataset was imported and manipulated using pandas and numpy package

## 3.3 Data Cleaning

Data cleaning is a process of identifying and removing inaccurate or corrupted data from the data frame. The data cleaning process is vital to prevent false conclusion derived from the final model. Several type of data cleaning method will be discussed in the next sub section.

### 3.3.1 Duplicate Data

There were several duplicate rows based on our initial assessment. All of these duplicate rows were found to be originated from the brickz website. We used duplicated() on the column Name and Type to assess the duplicated rows.

*Table 6: Sample of duplicate value*

| Row | Name | Complete | Type | Street | No of Transaction |
|---|---|---|---|---|---|
| 10. | Seni Mont Kiara | 2010 | Condominium | Mont Kiara | 191 |
| 11. | Seni Mont Kiara | 2010 | Condominium | Jinjang | 1 |
| 12. | Seni Mont Kiara | 2010 | Condominium | Jinjang | 1 |
| 23. | The Leafz@Dedaun Sungai Besi | 2015 | Condominium | Sungai Besi | 154 |
| 24. | The Leafz@Dedaun Sungai Besi | 2015 | Condominium | Salak South | 1 |
| 25 | The Leafz@Dedaun Sungai Besi | 2015 | Condominium | Salak South | 1 |

Based on the sample above, we found that the same subset of Name and Type resulted in different street and No of transaction. As it is almost impossible for only one transaction to occur in the span of 10 years of transaction, we decide to omit all the duplicated row based on subset of Name and Type that has only Number of Transaction of 1.

### 3.3.2 Incomplete and Missing  Data

Any missing numerical value ('Maintenance', 'Block', 'Total Unit', 'Level Number', 'Complete') will be filled up using median value imputation method of the intended missing feature. The imputation method use statiscal inference that substitute missing value based on other similar available information. The first tier imputation will group the property based on "Street" feature, follow  by "Type" feature. The price will be filled using median "Price" transformation.  Any further missing value will be filled up under the second tier imputation by grouping the property based on "Type" followed by filling up using median value imputation of the intended feature. The reason behind such selection is that same type of  house in a same neighbourhood tend to share similar characteristics.

### 3.3.3 Outliers

Outliers is the data  that possessed extreme value or highly deviated from normal distribution. We used Descriptive Statistic Table to assess any outlier in our data set.

Table 7: Descriptive Statistic Table

| | Mean | Median | S.D | Min | Max |
|---|---|---|---|---|---|
| Price (RM) | 627,668 | 450,000 | 701,960 | 5,000 | 34,300,000 |
| Lot Size (ft2) | 1167 | 1021 | 710 | 11 | 36,484 |
| Year of Completion | 2003 | 2006 | 9.43 | 1950 | 2017 |
| Maintenance Fee (RM) | 0.21 | 0.19 | 0.12 | 0.01 | 2.00 |
| Number of Blocks | 3.21 | 3.00 | 3.21 | 1.00 | 41.00 |
| Total Storey | 20.05 | 18.00 | 10.28 | 2.00 | 60.00 |
| Level of Storey | 11.15 | 10.00 | 8.37 | 0 | 51.00 |
| No of Transaction | 89 | 72 | 71 | 1 | 435 |
| Number of Room | 2.75 | 3 | 1.13 | 0 | 118 |

| | | | | | |
|---|---|---|---|---|---|
| Year of Sale | 2014.62 | 2015 | 1.89 | 2008 | 2018 |
| Month of Sale | 6.86 | 7 | 3.41 | 1 | 12 |
| OPR rate | 3.08 | 3.0 | 0.19 | 2.0 | 3.5 |

From the Descriptive Statistic Table, Lot Size and Number of Room possessed extreme value on both min and max end. It is impossible for a house in Malaysia to has the size of 11 $sqft^2$ or to has 118 number of rooms. We now try to find out any property that is below 100 $sqft^2$, above 20,000 or has rooms more than 10 using pandas.

*Table 8: Sample of Result for the Outlier Values*

| Row | Name | Lot Size | Type | Room | Price |
|---|---|---|---|---|---|
| **522** | The Scott Garden | 775 | Service Residence | 24 | 400,000 |
| **6920** | Amadesa | 1238 | Condominium | 118 | 460,000 |
| **16934** | Palazzio Sunway | 10258 | Condominium | 4 | 5,510,000 |
| **16962** | Palazzio Sunway | 10183 | Condominium | 6 | 4,920,750 |
| **18892** | Pertama Residency | 10409 | Service Residence | 1 | 5,500,000 |
| **18981** | Pelangi Indah | 11011 | Condominium | 1 | 180,000 |

We found out the extreme value that did not make sense. For example, it is impossible for a house of size 1238 sqft2 to has 118 number of rooms. Another example, Average selling price for extremely high size as shown by Pelangi Indah (11,000 sqft2 sold at just Rm180,000). Instead of removing this outlier value, we will replace the outliers through group by name and average imputation method.

## 3.3 Data Filtering ,Substitution and Sorting

This section explained how the data filtering, substitution and sorting are made.

- Remove string based value from any numerical features and convert them into numerical type (int or float). This is applied to Year of Completion, Total Storeys, Number of Blocks, Number of Units, Price and OPR.

- Substitute any ordinal category (category with ranking) data into integer. For 'Type' feature, Flat = 1 Apartment = 2, Condominium/ Service Residence = 3. For 'Tenure', Leasehold = 0 and Freehold =1. The increasing number represent higher quality of feature

- Substitute any nominal value (category without ranking) into integer. The 'Street' feature which in total of 41 unique value are converted into integer using label encoder. The conversion of numerical value is done for visualization interest and the 'Street' feature will be drop before the final modelling.

- Under 'Maintenance' variable. Small portion of the value consist of object type "/month". Any strings text was stripped off and convert into numerical. The '/month" value was converted into fee per square foot area by diving the value by the size of the property

- The 'Developer' features were ranked based on The Edge Malaysia Property Excellence Award Top 30 Property Developer. Due to difficulty to rank the developer, all the developer that made through the Edge list will be given rank 1, while other developer that that did not make the list will be given value of 0

- 'Median Price Psf' will be omitted before the final modelling because it is highly correlated with target variable, 'Price'. 'Median Price Psf' will be used only for visualization and EDA purpose.

- 'Longitude' and 'Latitude' based on kmean Euclidean distance. Total 10 clusters will be chosen as the parameter under the kmean clustering. Both of 'Zone' and 'Longitude' & 'Latitude' will be run separately in the final model to see the differences in the model performance.

- Several features are found out to be inconsistence. For example, the "No of Transaction" which represent total number of transactions volume made for a particular property from 2008 to 2018 acquired from Brickz.com is not consistence with total transaction volume with another similar property transaction portal ([www.propertyadvisor.com](www.propertyadvisor.com)). This variable will be also excluded from the final modelling.

- The dataset will be sorted through the ascending date order

## 3.4 Feature Engineering

Feature engineering is the process of transforming the raw data into features that better represent the underlying predictive model, resulting in improved model accuracy on unseen data. Feature engineering is used to aid visualization and data exploration as well. We will apply feature engineering to generate new set of features

- For the time series, a new features 'SPAMonth' and 'SPAYear' will be generated from the 'Date of Sale' variable as most machine learning model is not able to handle date type data directly. The 'SPAMonth' will be used to observe any seasonality effect while the 'SPAYear' will be used to assess the trend effect. Date column will be dropped before the final modelling process.

- 'Street', 'Address', 'Longitude' and 'Latitude' represent the spatial coordinate effect and highly correlated to each other. For the final model, 'Street' and 'Address' will be omitted. Another new features called 'Zone' will be generated from 'Longitude' and 'Latitude' by kmean clustering. We will apply both Zone and Coordinate columns separately in the final model and observe the performance differences.

- From 'Maintenance' feature, a new feature features 'MaintenanceRng' will be generated which represent the maintenance fee range value. The range value will consist of range 0-0.1,0-0.15,0.15-0.20,0.20-0.25,0.25-0.30, 0.30-0.35, 0.35-0.40, 0.40-0.45, 0.45-0.50, 0.50-2.0. The

- The Level Number features is generated from the Address feature. For example, A1-10-12, Vista Komanwel A. The A1 represent the block, 10 represent the level number and 12

designated unit for the house. We use pandas manipulation technique to extract the level number from the address.

## 3.4 Exploratory Data Analysis

The purpose of exploratory data analysis (EDA) is to find meaningful pattern or insights, often with the assistance of visualization methods. EDA is also used to investigate the correlation between dependent and independent variables, or amongst independent variables themselves.

Our main interest is this study is to find factors that influence our dependent variable (Price) by investigating the relationship between the independent variables (Complete, Type, Developer, Maintenance, Block, Total Storeys, Level of Stroey, Total Unit, Tenure, Latitude, Longitude, Zone, Rooms, SPAYear,SPAMonth and OPR)

Our dataset consist of 3 type of high rise residential: flat-199 units, apartment-159 units and condominium-581 units. Figure 3.4a will showed total count of each type of property along with the total transaction for each type of high rise residential throughout the study period



*Figure 6: Total Building Count & Total Transaction Count*

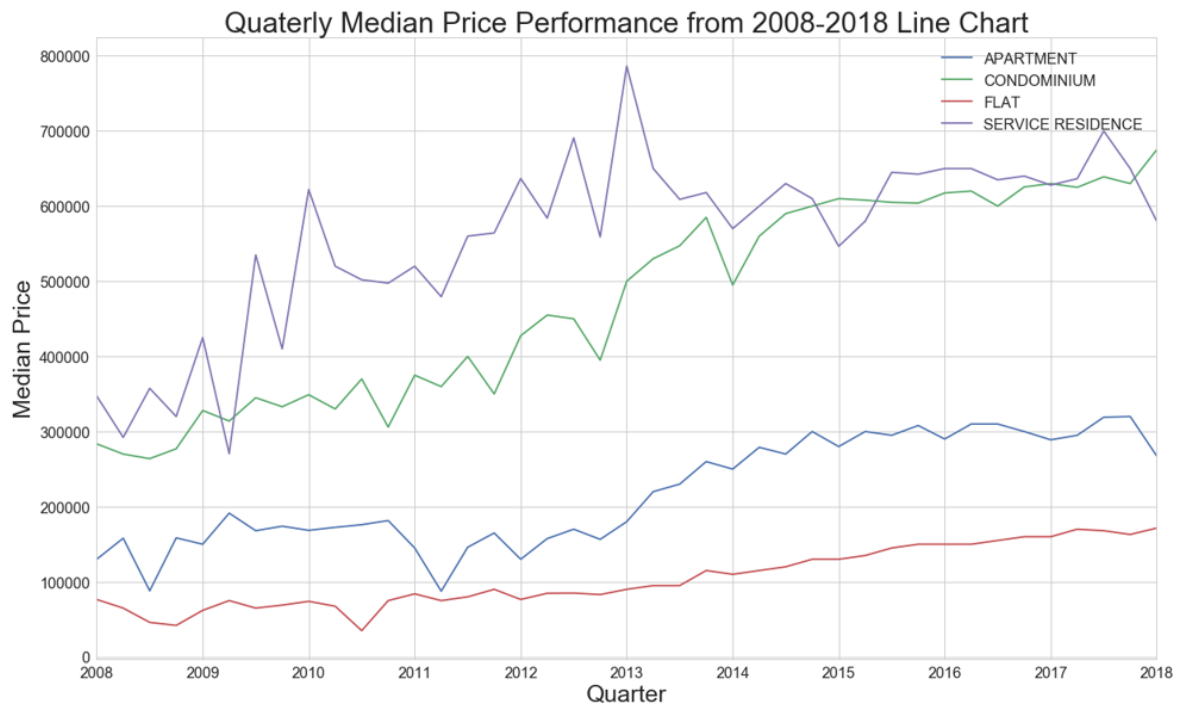## 3.4.2 Price Trend Analysis



*Figure 7: Quarterly Median Price Performance 2008-2018*
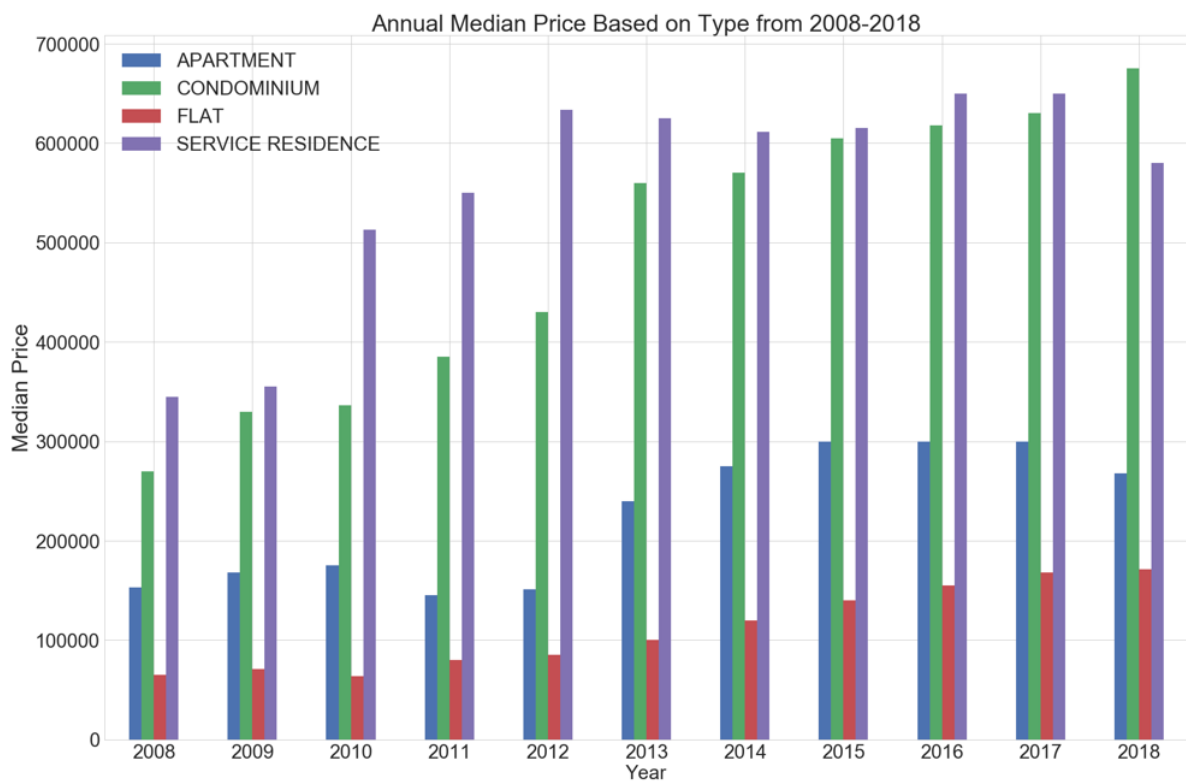


*Figure 8: Annual Median Price Trend*

The price for condominium and service residence showed tremendous growth of almost 100% from 2008 to 2013 and then start to become stagnant from 2016 to 2018. Price for flat and apartment start to pick up traction after year 2012. Despite the late explosion of price growth for flat, it can be observed that flat still show consistent annual growth up to recent year compare to other two type of property. Apartment showed significant decline of median price on 2018 but it to early to conclude that the price of apartment is declining since the period of this study for 2018 only cover the first quarter of 2018.

### 3.4.3 Geographic Analysis – Boundary Analysis

We used gmaps package to assess the geographical distribution (Latitude & Longitude) in this research which consist of 951 location. The total area of coverage will be approximately around 243km2. It can be observed that luxury residential such as condominiums tend to populate hot spot area such as Mont Kiara and KLCC. While cheapest high rise residential such as flat tend to populate the area nearing to the boundary of Kuala Lumpur.
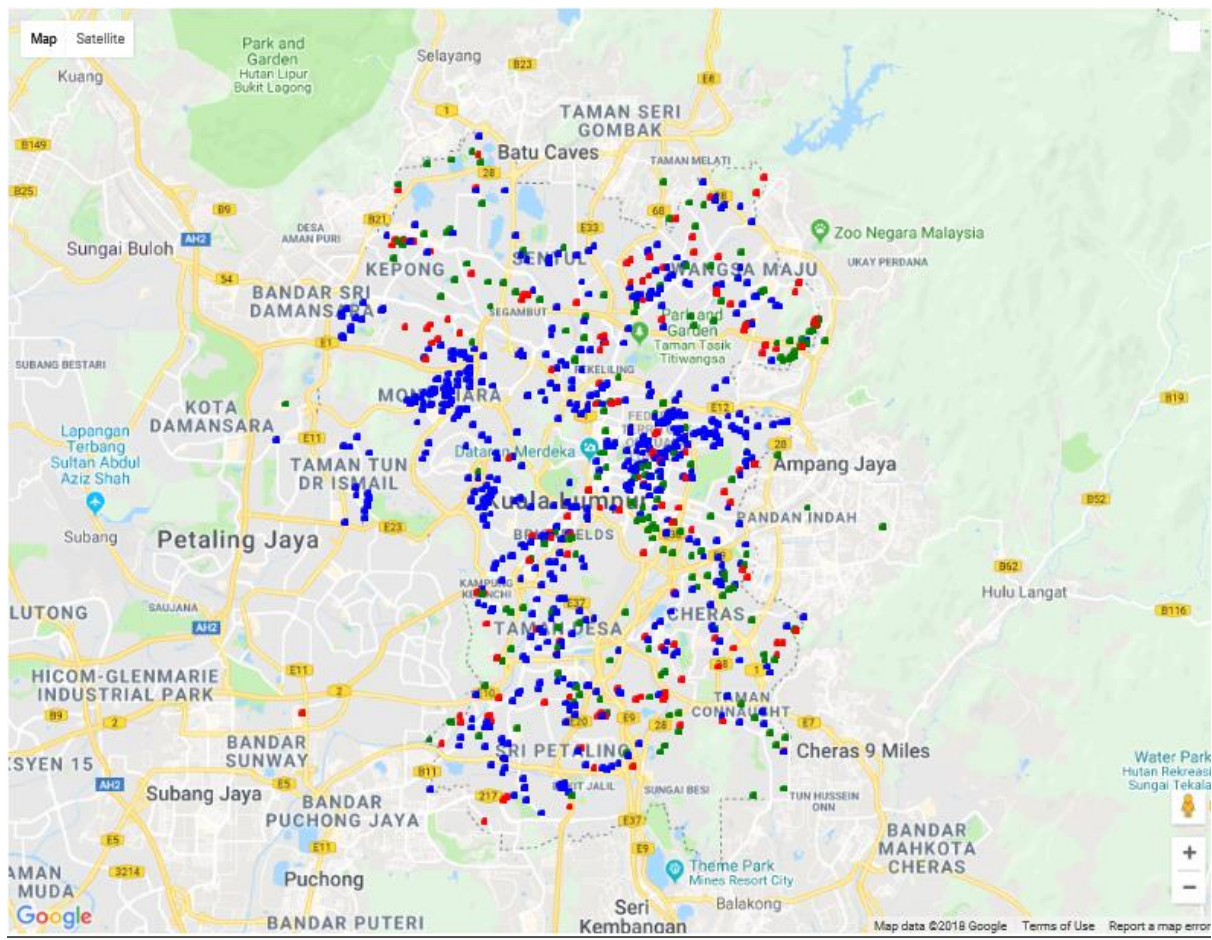


*Figure 9: Property Type Distribution with colour coded*

### 3.4.4 Heatmap Analysis

Heatmap is used to assess the intensity of price per square feet using the same google map package
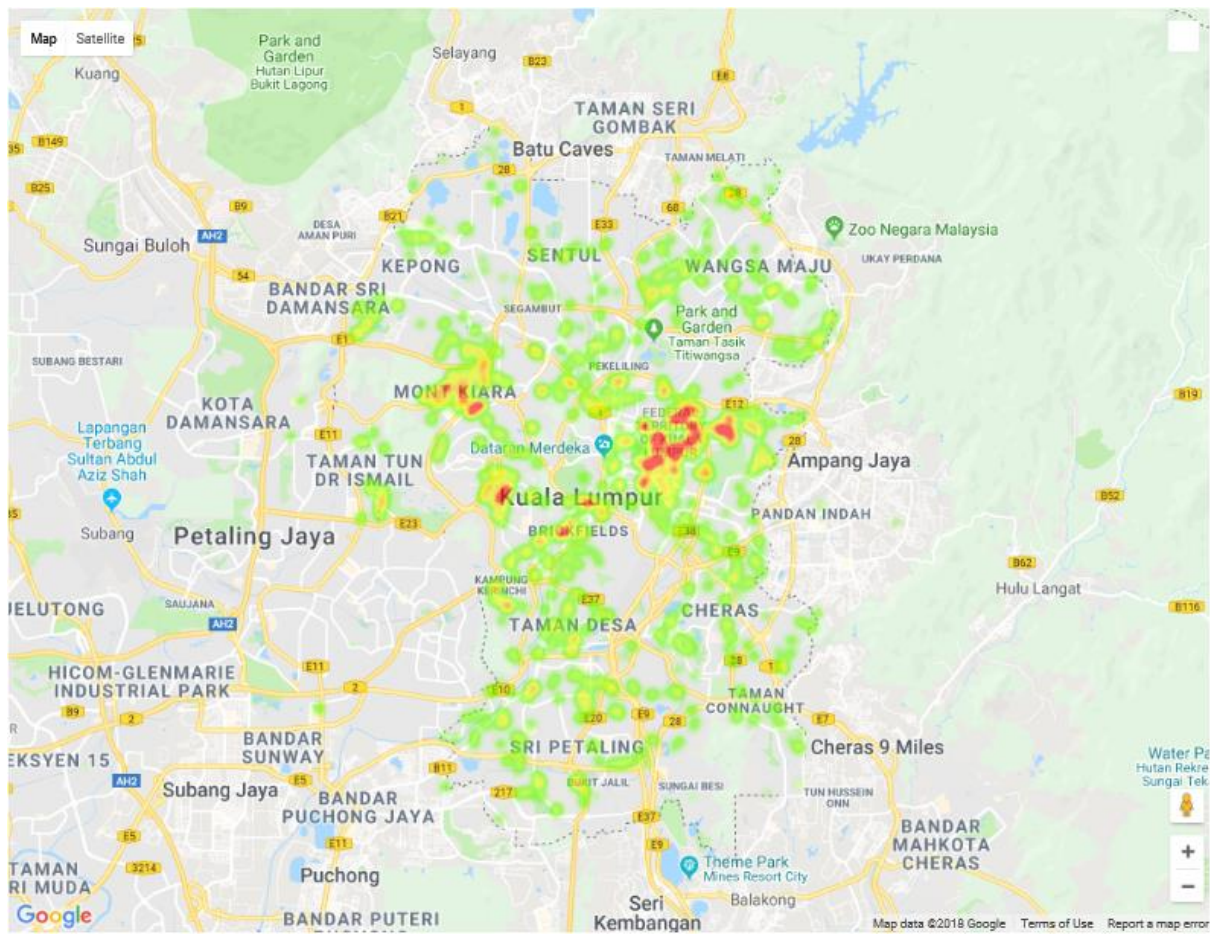


*Figure 10: Heatmap for Price Per Square of Area*

From the heatmap, it can be observed that highest price per square foot concentrated in KLCC, Mont Kiara and Bangsar area. Boxplot is used to further assess the distribution of the price across the 41 street/neighbourhood in Kuala Lumpur.
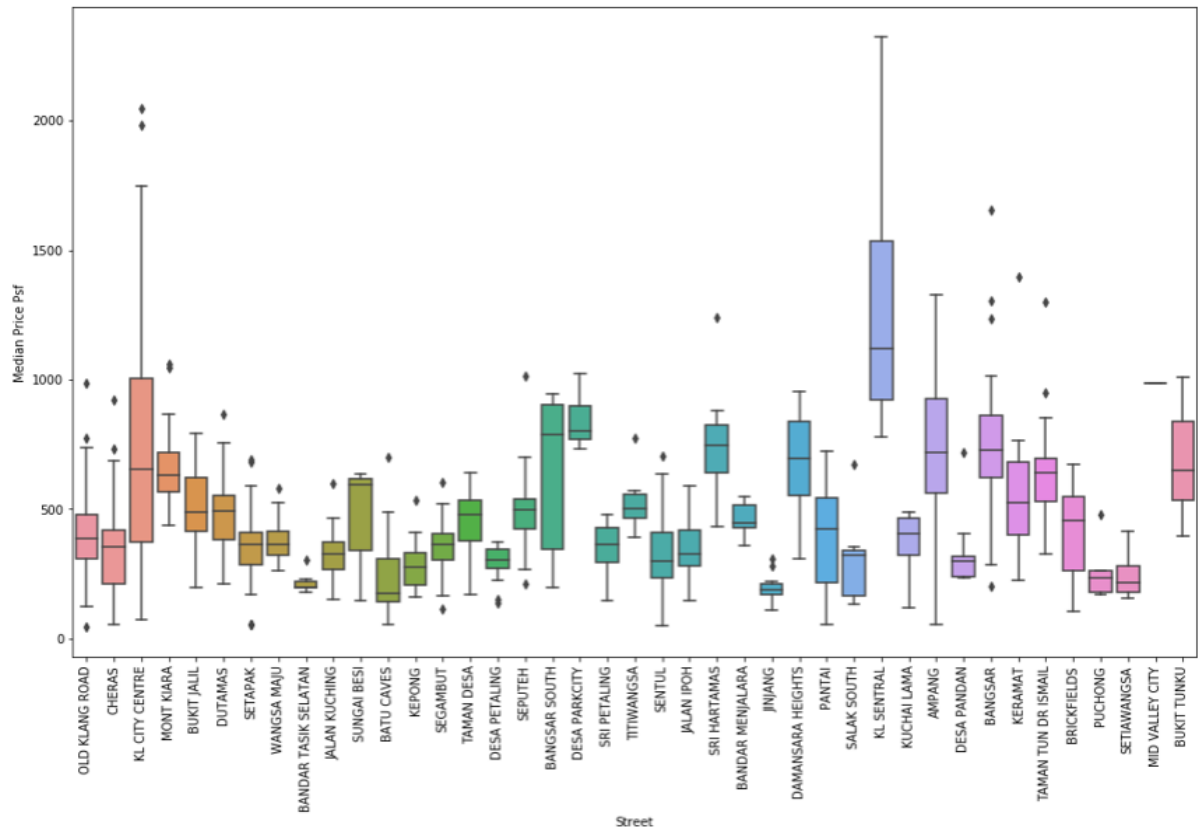
*Figure 11: Boxplot of median price based on the Neighbourhood*

It is confirmed that neighbourhood such as Mont Kiara, KLCC, Bangsar and KL Sentral have higher median price per square area compare to the surrounding neighbourhood. All of these area also located in the centre of Kuala Lumpur, which the economic activity is the highest.

### 3.4.5 Zone Clustering

Coordinate (Latitude and Longitude) and Street posed complexity to categorize due to the high number. An alternative spatial variable was generated by clustering the coordinate into 10 zones using k-mean clustering. Separate model will also be generated using this zone clustering variable over the spatial coordinate variable (longitude and latitude)
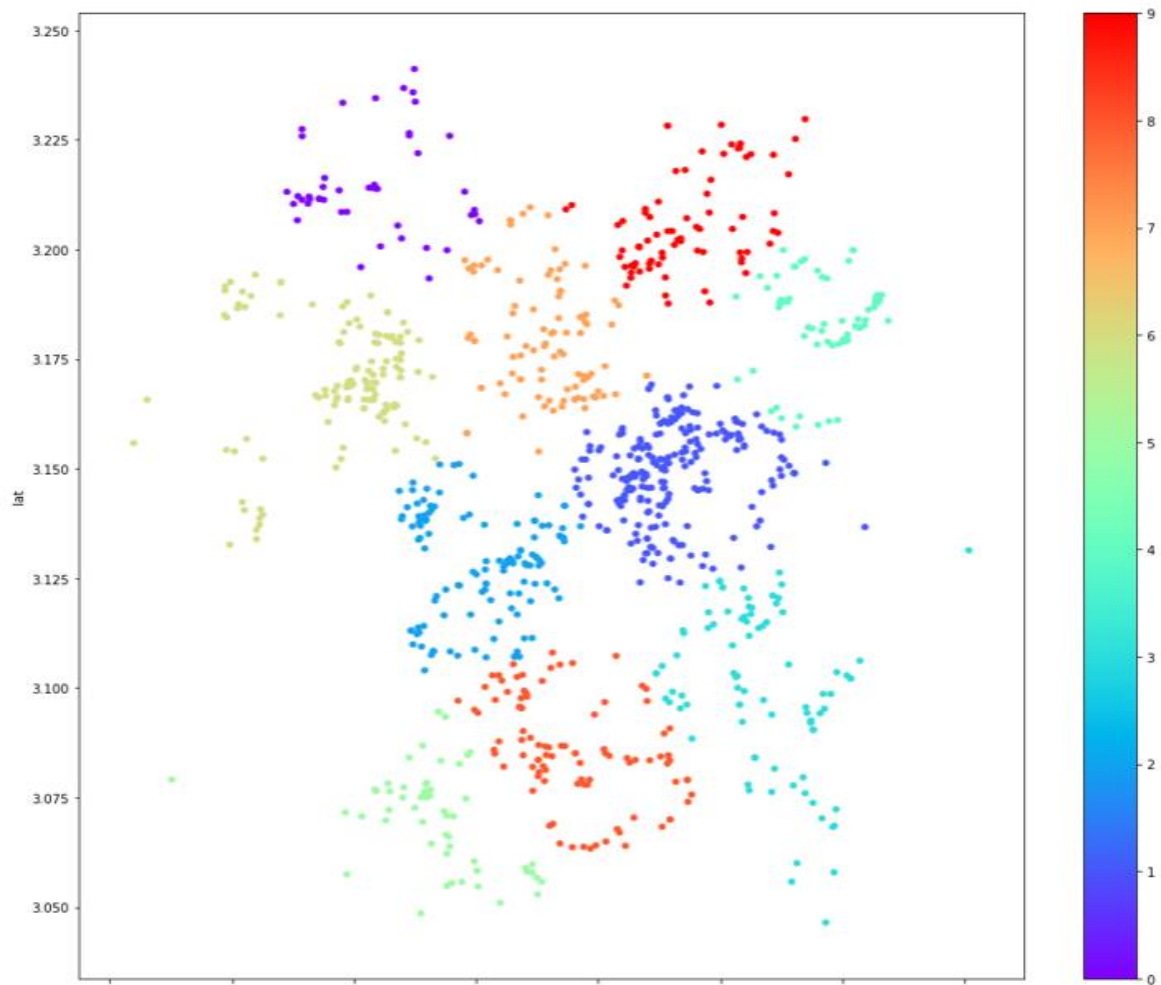


*Figure 12: Scatter Plot based on 10 cluster of Kmean*

## 3.4.6 Structural Analysis
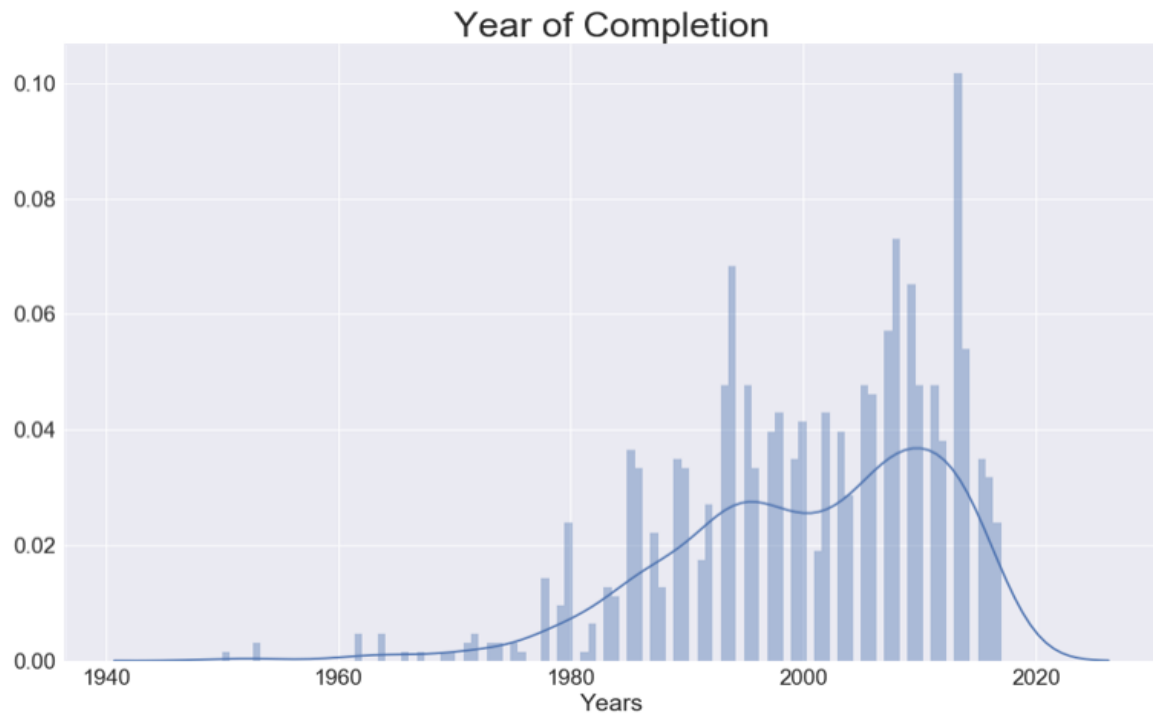
### 3.4.6.1 Year of Completion



*Figure 13: Histogram of Year of Build*

Based on our dataset, most of the residentials was built after year 1990, with the most transaction made up from the residentials that are recently built.
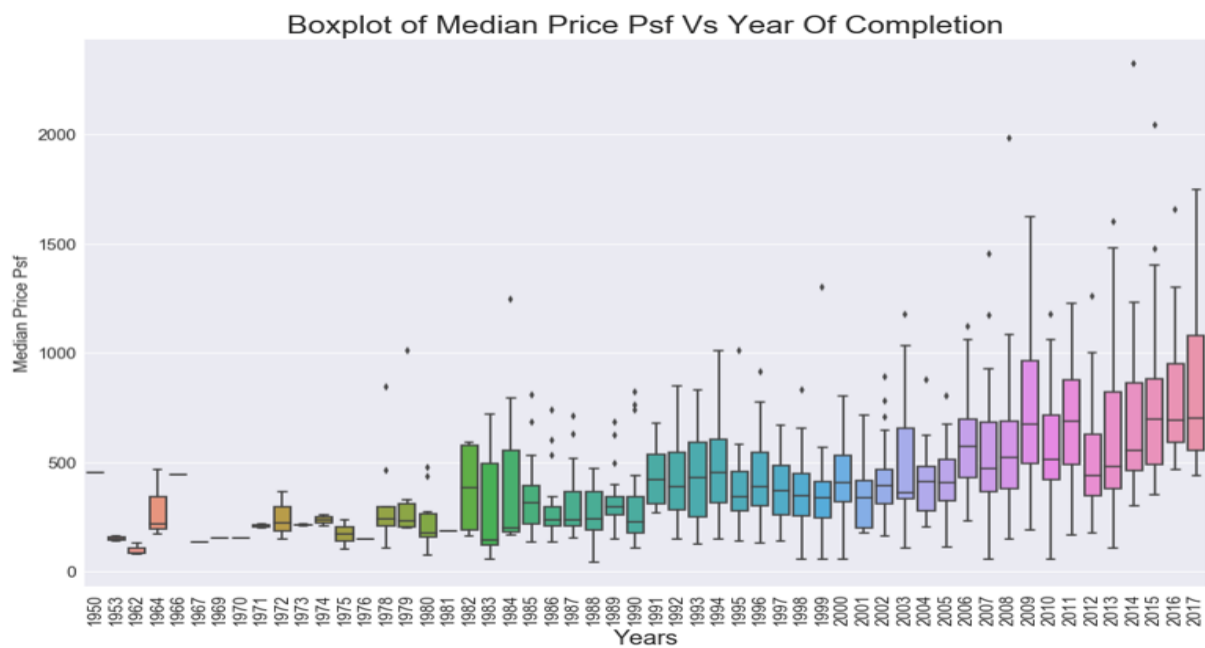


*Figure 14: Boxplot of Median Price Psf Vs Year of Completion*

Based on the boxplot, the recent completed residential tend to command higher price per square foot. From both of Year of Completion histogram and Boxplot, we hypothesize that the house buyer tend to buy recently completed residential and more tolerance toward the higher price due to more facilities and up to date design offered by newly built residential.

### 3.4.7 Macroeconomic Analysis (OPR)

Around 2007-2009, subprime mortgage bubble took place in USA. Malaysia was not spared from this recession. Central Bank of Malaysia decided to lower the Overnight Policy Rate to the lowest rate of 2.0% in 2009 as the countermeasure of this economic recession. This step is taken to encourage financial institution to lend money to the end consumer at the cheaper rate which to spur back the economic activities. Using OPR rate for 2008 to 2018, we use line chart to measure if the housing price is correlated with OPR rate. Our initial assumption is that house price should drop if the OPR rate is high and vise versa.
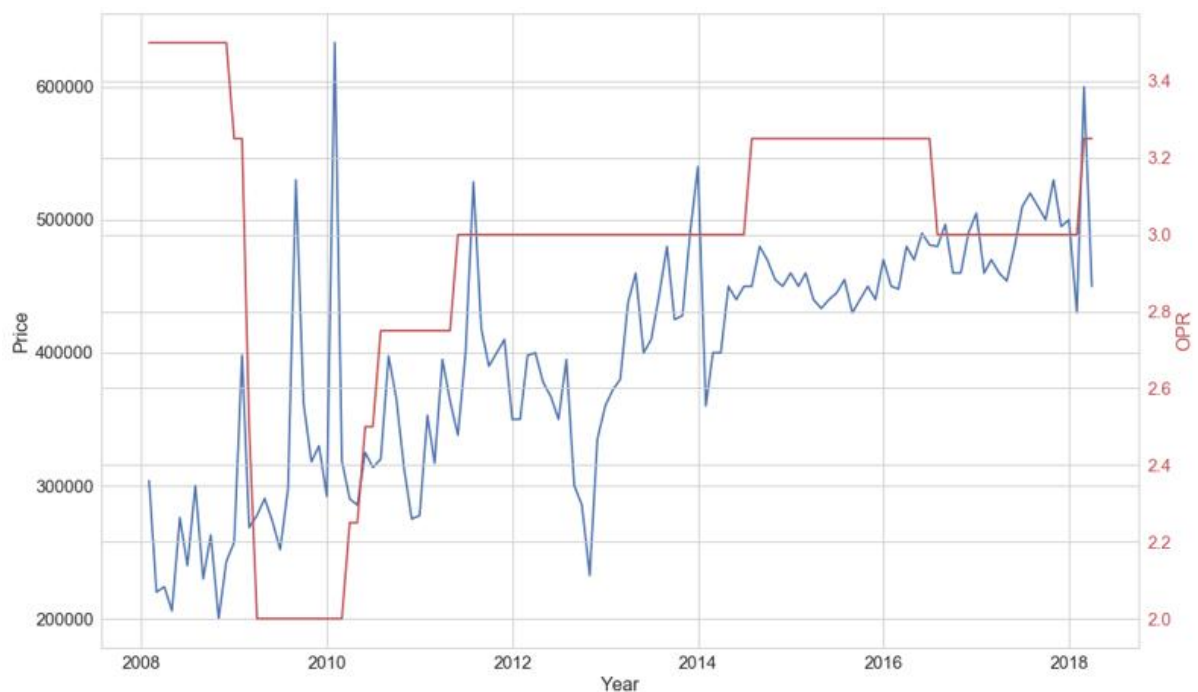


*Figure 15:Monthly Median Price Based on OPR*

Based on the monthly median price, we could not see any direct correlation between monthly median housing price with OPR rate. Nevertheless, we will further study the feature of importance for OPR on the housing price under the Modelling chapter.

### 3.4.8 Correlation Analysis Between Independent Variables and The Target

Continuous variables consist of Price, Lot Size, Year Build, Maintenance Fee, Number of Blocks, Total Storey, Level of Storey Transacted, Number of Room, Number of Transaction, Month of Transaction, Year of Transaction and OPR rate. The table below show the aggregated statistic summary for each continuous variable. Scatter plot is used to see the correlation between target variable (Price) and all the independent continuous variables.
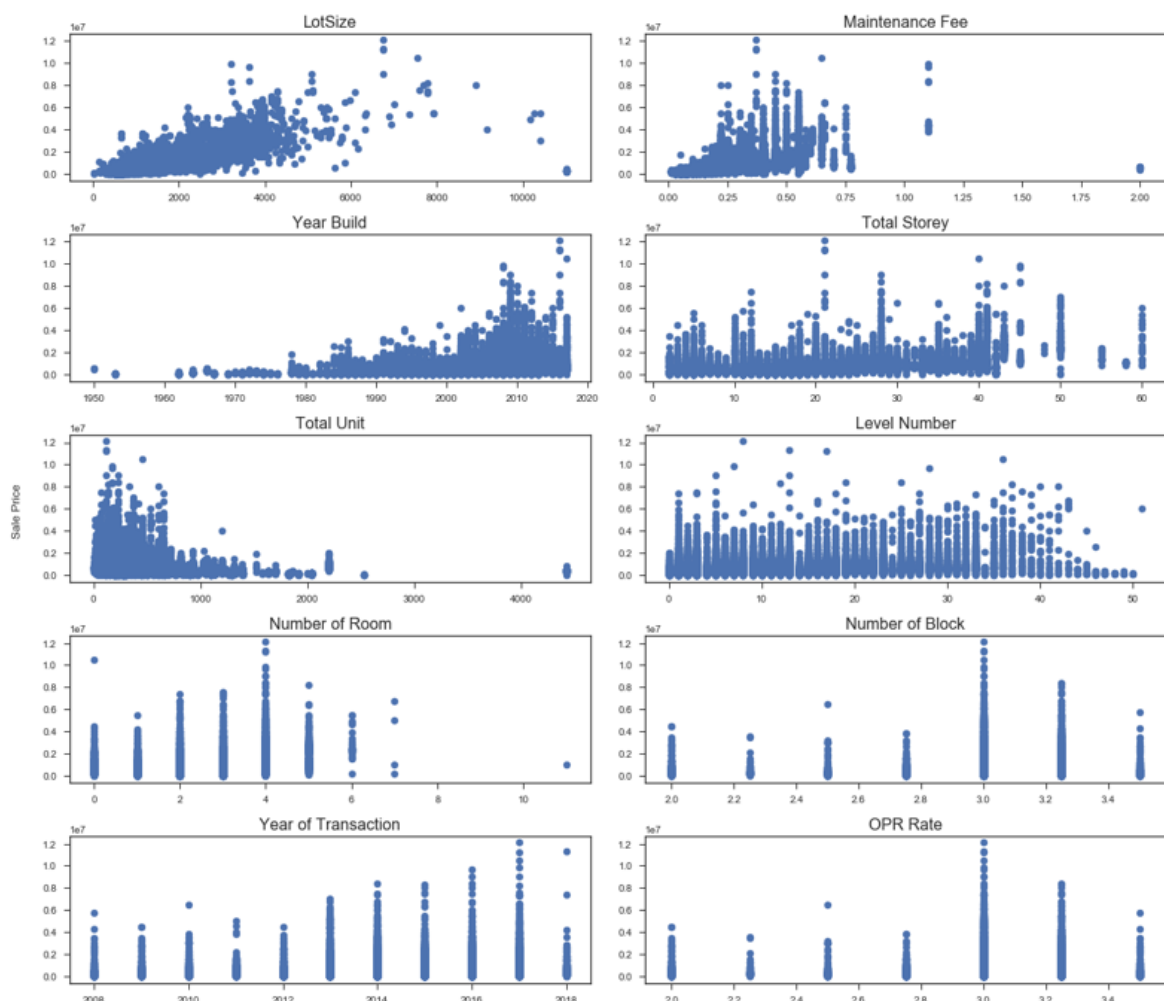


*Figure 16: Correlation between Price and other continuous variable*

It can be observed that there are several variables were showing positive correlation with Price. These variables that showed strong positive correlation are "Size", "Maintenance Fee", "Year Build" and "Year of Transaction". 'Total Unit' and 'Number of Block' indicated negative correlation which imply that the house price tend to reduce as the total unit of house in a particular residential building increase.

Pearson correlation is also used to see the correlation between the structural variables. In the context of housing attributes, several features showed strong correlation between each other. For example, "Total Unit" showed strong positive correlation with "Number of Block". This is the due to higher number of block able to accommodate higher number of total unit of house for a particular residential. This factor is also true for the correlation between "Level Number" and "Total Storey". Overall, the independent variable showed little multicollinearity which fulfil our linear regression assumptions.
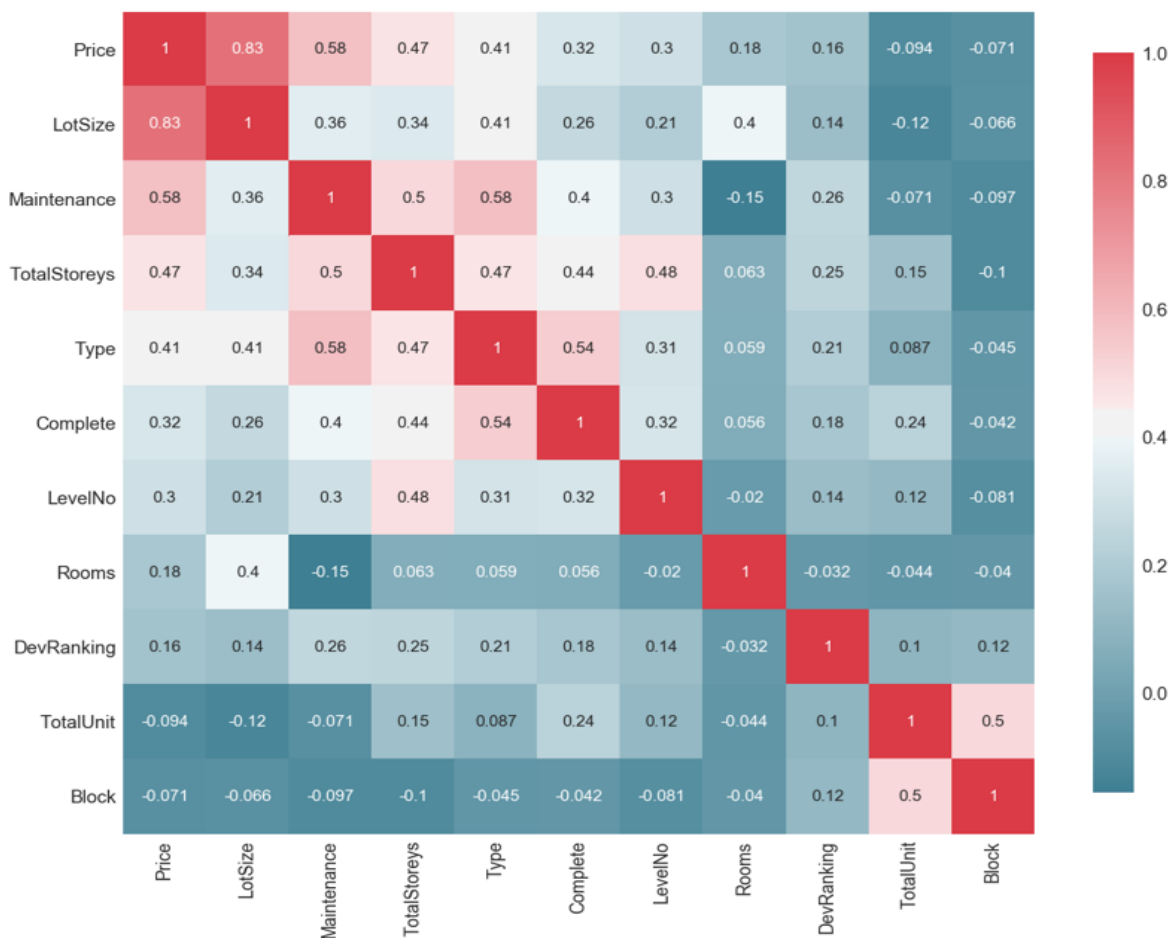


*Figure 17: Heatmap of Pearson correlation between structural variables*
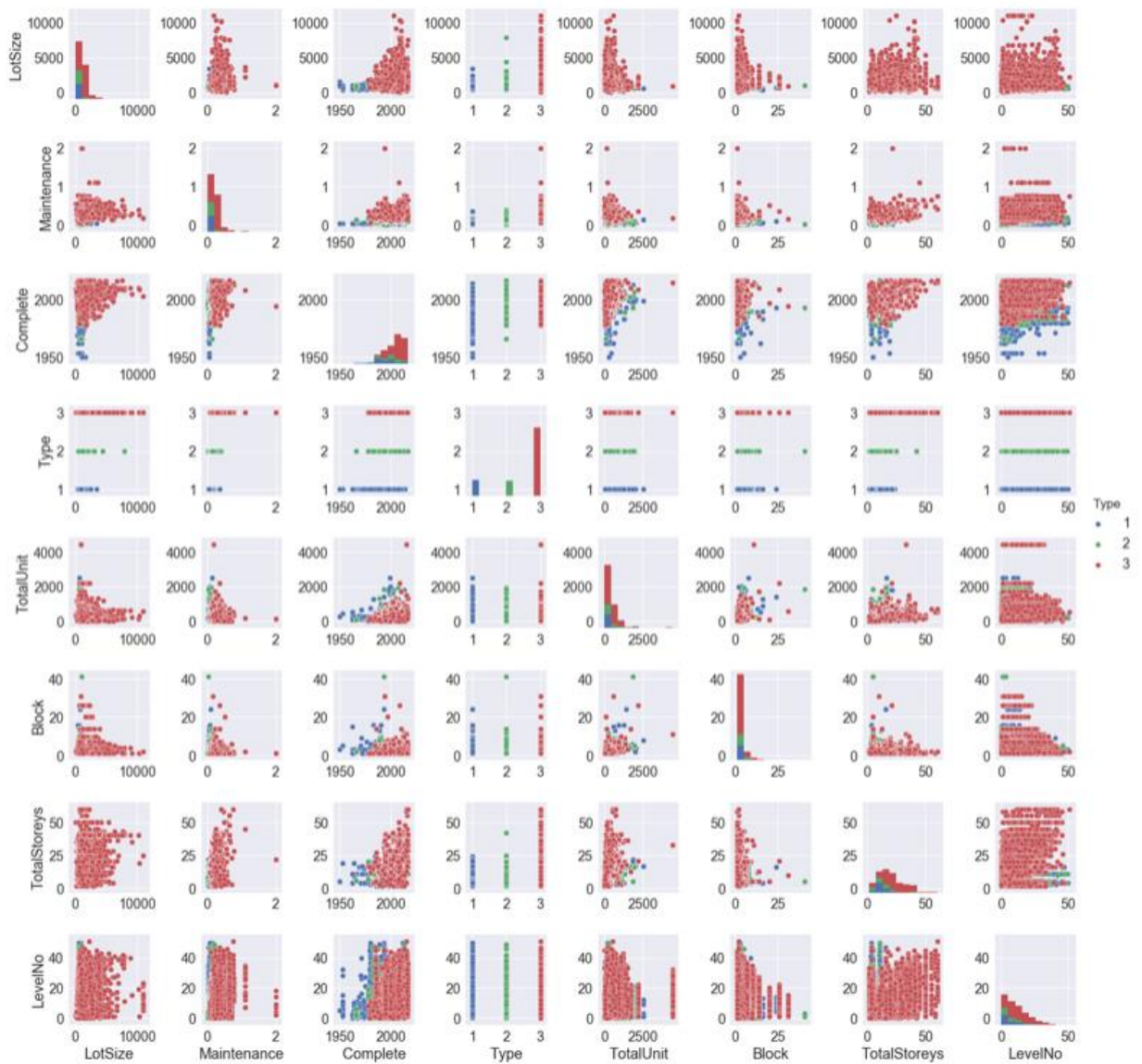
*Figure 18:Pair plot of independent structural variables*

Pair plot of figure 12 is a used to visualize and further assess the correlation between independent structural variables. Based on our observation on the pair plot, there is no strong correlation displayed between independent variables. We can safely say that there is minimum multicollinearity issue that affect our final models.

### 3.4.9 Price Distribution Analysis & Data Transformation
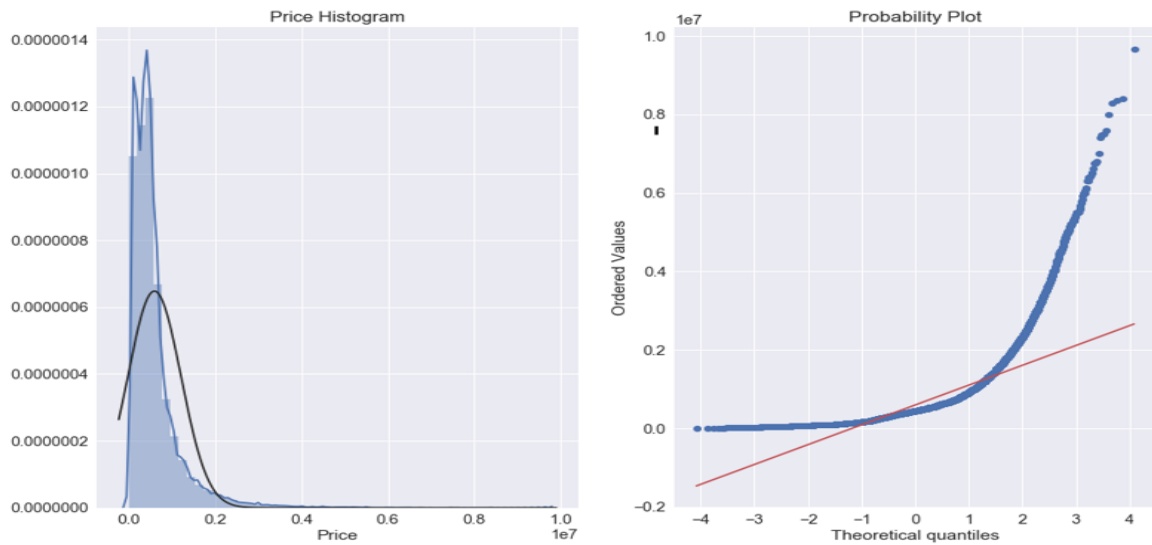


*Figure 19: Price Distribution & q-q plot*

The Price Histogram is not normally distributed and highly skewed to the right tail. Skewness is undesirable for the final modelling as most of the machine learning algorithms make regards the shape (distribution) of data namely normal distribution. The Q-Q plot further indicate how much is the deviation of our Price distribution from the normal distribution.

We use log transformation to transform the Price into normal distribution. Also, using log transformation help to minimize the impact of outliers



*Figure 20: Price Distribution & q-q plot after log  transformation*

# 4. Data Modelling

Ensemble learning is a technique that combine multiple learners to increase the prediction power of the model. Under this section, we will apply 3 ensemble method which are bagging (Random Forest), Boosting (Extreme Gradient Boost) and stack generalization (several combination of different algorithm). We use the pre-built ensemble models from sklearn library instead of building one from the scratch.

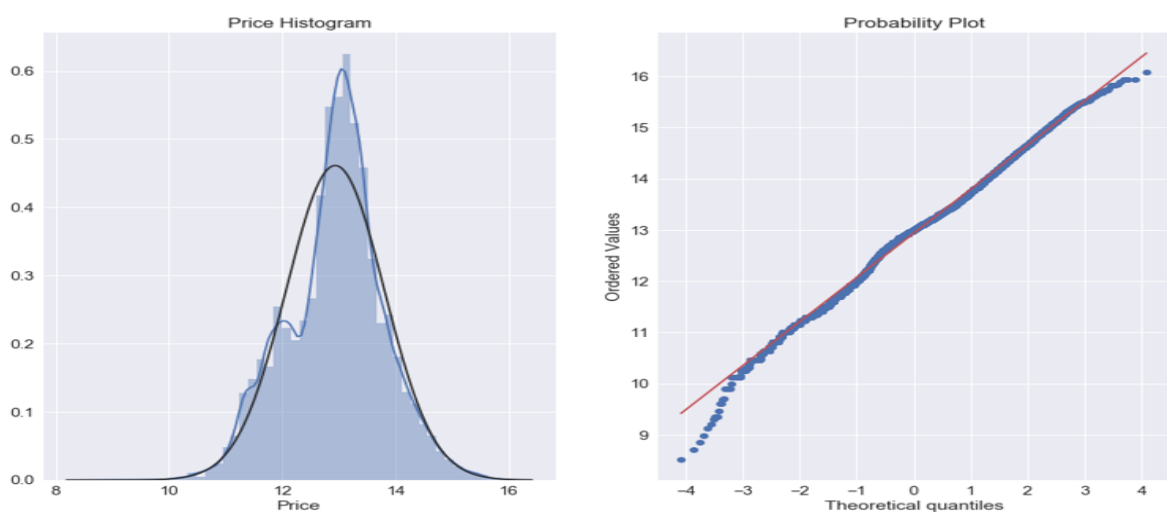The final dataset after all the pre-processing activities discussed under the previous chapter will be 30949 rows and 17 columns. We will sort the data set into based on transaction date whereby the oldest transaction will be on the top. 80% of this data will be training set and the remaining 20% of the most recent transaction will become our test set. The target variable, Price for the test set, will be removed. By the end of the modelling activities, we will compare the result derived from the prediction model of the test set with the actual value.

## 4.1 Bagging (Random Forest)

Bagging which stand for bootstrap aggregation, involves creating samples from the dataset with replacement. Random Forest is one of the machine learning model under Bagging. It give prediction based on majority voting (for classification case) or averaging (for regression case). Each tree (learner) in random forest is build independently from each other. For example, N is the size of the sample for the dataset, and M is the total number of tree. A subset of m<M Is randomly chosen to grow the each tree on a bootstrap sample. Unpruned tree is grown from each of the bootstrap sample. The best split among random sample of m variable is selected based on the information gain formula. After large number of trees are generated, the final prediction are averaged over the different trees.. Random forest require minimum preparation of the dataset and minimize overfitting. Another advantage of using Random forest under sklearn package is the ability to generated separate holdout set than can be used as the validation set. This hold out set which also called Out of Bag (OOB) can be used to test if the model is overfitting.

The hyperparameters for this random forest:

*Table 9: Parameter for Random Forest*

| Parameters | Description | Value |
|---|---|---|
| n_estimator | Number of tree estimator | 300 |
| min_sample_leaf=5 | Minimum number of sample in the leaf node | 5 |
| max_features | Number of features to consider when looking for the best split | 0.50 |
| oob_true | Out of the bag feature, Random forest only use 67% of the bootstrapped data for training, the remaining 33% which also called OOB can be used for validation set | True |

## 4.2 Boosting (XGBoost)

Boosting algorithm trained based on the data iteratively and update the instances weight according to the different between predicted and actual values. The update of the weight depend of the algorithm that being used. Misclassified learner get weight increase while correct classification get weight reduction. This lead to the classifier give more attention to the misclassified one.

We use XGBoost algorithm due to the reputation that it received. In the recent year, XGBoost has been used extensively especially in Kaggle competition. This due to the several advantages that this algorithm possessed. First, it is extremely fast and second, it performance is top notch especially when it comes to tubular dataset.

| Parameters | Description | Value |
|---|---|---|
| colsample_bytree | Subsample of columns when constructing each tree | 0.04 |
| gamma | Minimum loss reduction required to make further partition on leaf node of tree | 0.045 |
| Learning_rate | Step size shrinkage used in update to prevent overfitting | 0.07 |
| max_depth | Maximum depth of tree | 20 |
| min_child_weight | Minimum sum of instance weight needed in a child | 1.5 |
| n_estimators | Number of learner | 300 |
| reg_alpha | L1 regularization on weight | 0.65 |
| reg_lambda | L2 regularization on weight | 0.45 |
| subsample | Subsample ratio in the traini9ng instance | 0.95 |

## 4.3 Stacking

Unlike previous 2 methods which use multiple models of the same kind to improve accuracy, Stacking Generalization Ensemble is an ensemble method that use different kind of models. What is unique about stack ensemble is it's ability to train new model by combining previously trained model on the same dataset. The initial level of classifiers were trained separately and their results are combined using another classifier. The result of the stack ensemble can be simply achieved by averaging the initial classifiers prediction or by using weighted sum.

Compare to the previous two ensemble models. The hyperparameter for each classifier inside the stack ensemble need to be tuned separately. We use linear regression, ridge regularization, random forest, gradient boosting light and neural network(MLP) as the 0 level model (initial model) and linear regression as the second level model. The reasons behind selecting such model is because they are

different pattern of generalization model and have different biases thus can lead to efficient

ensemble. As for the level 1 learner, linear regression is selected due to the simplicity of this model

*Table 11: Models and Parameter for Stack Ensemble*

| Model | Description | Value |
|---|---|---|
| Linear Regression | Use Basic Linear Regression method | n_job = 1 |
| Ridge | Linear least square with l2 regularization | alpha = 4.84 |
| Random Forest Regressor | Random forest regressor model | n_estimators = 40, max_depth = 3, |
| Gradient Boost Regressor | Additive model in a forward stage wise fashion | n_estimator =40, max_depth=2 |
| Multi Layer Perceptron | Multilayer Perceptron Regressor that optimize stochastic gradient descent | Hidden_layer_size=(90,90), alpha=2 |
| 1 Level Model Linear regression | Use Basic Linear Regression method | N_job = 1 |

# 5. Results

The aim of this chapter is to assess the performance for each model. Next, we will investigate the variables which have the strong feature of importance using the result derived from the models.

## 5.1 Accuracy

We use the coefficient of determination, $R^2$ and Root Mean Square Error (RMSE) to determine the accuracy of prediction for all our ensemble model. $R^2$ is the proportion of the variance in the dependent variable that is predictable from the independent variables. It measure of how closed the data to the fitted regression line. The highest core for $R^2$ will be 1 and the worst is 0. While RMSE is used to measure the differences between value (sample or population values) predicted by a model and the actual value.

The dataset for the accuracy assessment will be divided into train and test set. The accuracy comparison between train set and test set is used to determine if the model is overfitting (higher score of train set over test set indicate overfitting)

*Table 12: Accuracy Score*

| Model | RMSE train | RMSE test data | $R^2$ Train | $R^2$ Test | OOB |
|-------|-----------|----------------|-------------|------------|-----|
| Random Forest | 0.174 | 0.235 | 0.96 | 0.92 | 0.93 |
| XGBoost | 0.159 | 0.23 | 0.98 | 0.92 | NA |
| Stack Ensemble | 0.313 | 0.319 | 0.87 | 0.86 | NA |

Stack Ensemble performed the worst compare to the Random Forest and XGBoost with around 0.86 $R^2$ score. Both Random Forest and XGBoost showed similar $R^2$ score at 0.92, yet XGBoost showed less RMSE of both train set and test set compare to Random Forest.. Despite the higher score of RMSE train set to RMSE test set, we concluded that our model is not overfitted due the high OOB score of the random forest.

## 5.2 Feature of Importance

Since the stacking ensemble is a combination between different algorithm, it is nearly impossible for us to assess the features of importance. So we ignore the feature of importance for stacking ensemble and focus on the other two, Random Forest and XGBoost. After all, this two ensemble accuracy performance is much higher compare to the stack ensemble.

*Table 13: Comparison of features of important between ensemble model*

| Variables | Random Forest | XGBoost | Stacking |
|---|---|---|---|
| Type | 0.308 | 0.013 | NA |
| Size | 0.280 | 0.154 | NA |
| Maintenance Fee | 0.220 | 0.073 | NA |
| Total Storey | 0.052 | 0.076 | NA |
| Year Build | 0.025 | 0.073 | NA |
| Latitude | 0.020 | 0.146 | NA |
| Longitude | 0.016 | 0.111 | NA |
| Total Unit | 0.011 | 0.071 | NA |
| Tenure | 0.011 | 0.011 | NA |
| Rooms | 0.010 | 0.023 | NA |
| Level Number | 0.006 | 0.126 | NA |
| Number of Block | 0.005 | 0.036 | NA |
| OPR | 0.005 | 0.026 | NA |
| Developer Reputation | 0.001 | 0.006 | NA |
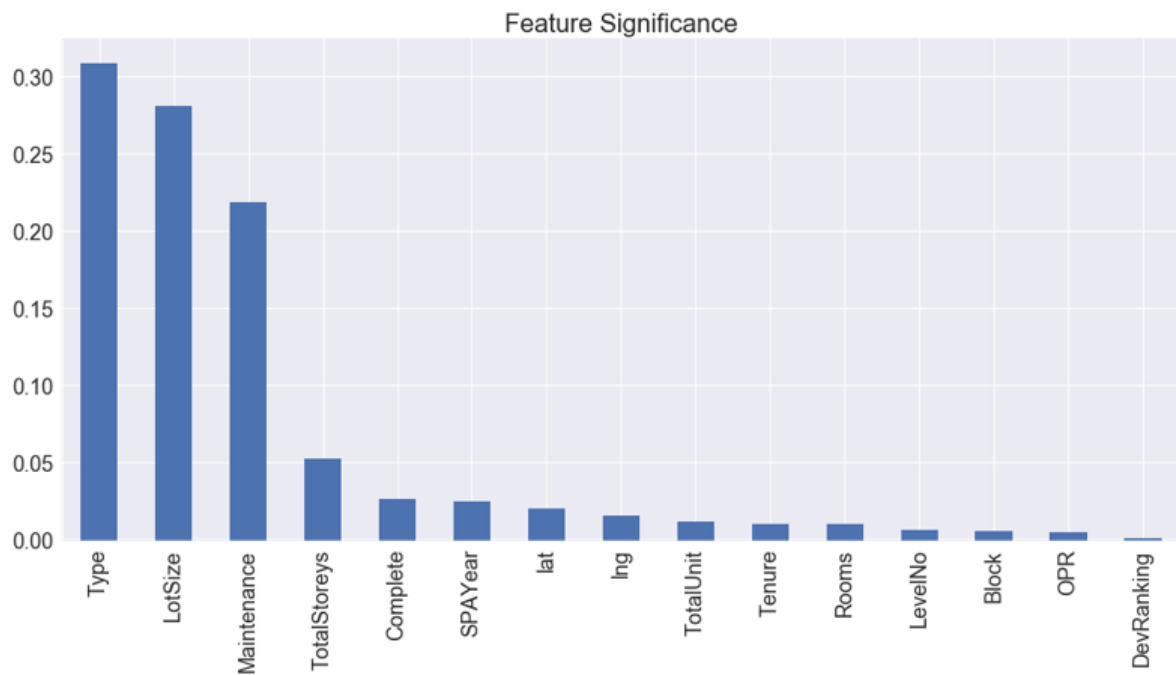
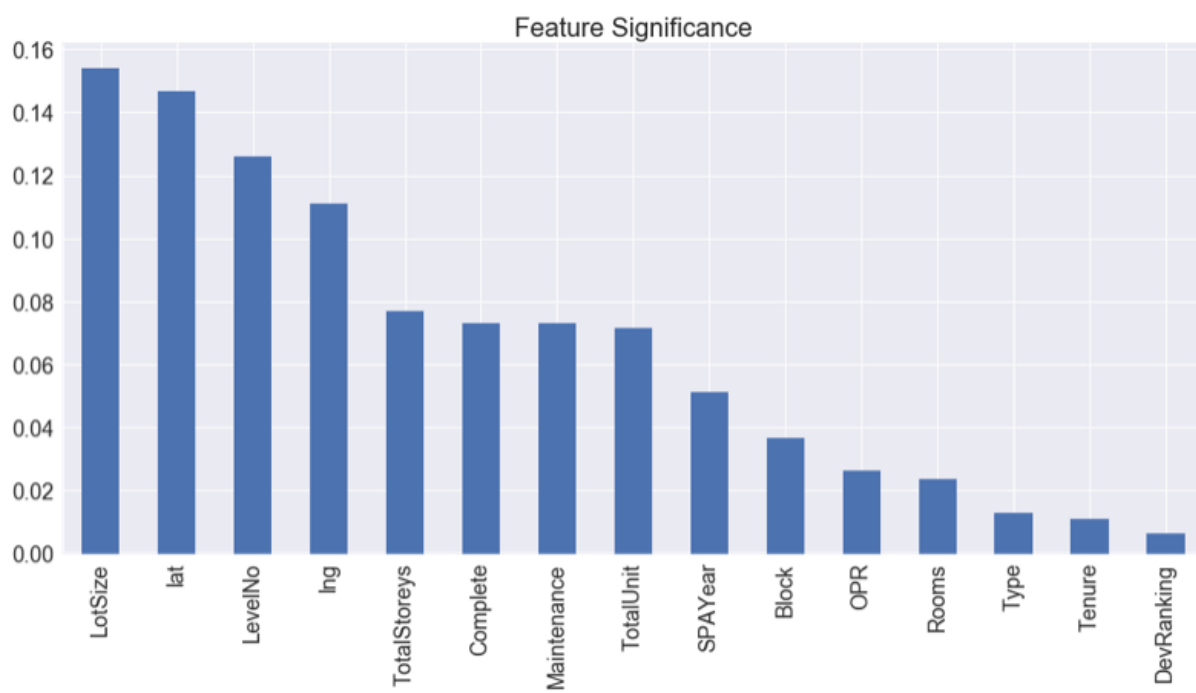*Figure 21: Features Importance for Random Forest*



*Figure 22: Feature Importance for XGBoost*

Surprisingly, Both Random Forest and XGBoost showed different features of importance. Random forest select Type, Size and Maintenance Fee is the top 3 most important feature while XGBoost choose Size, latitude and Level Number as the most importance feature. Yet, both model agree that Size is the most important feature in determining the price of the house. For the final prediction model, we will chose XGBoost because not just it possess higher accuracy score, but the feature of importance shown by XGboost tend be more uniformly distributed amongst its variable.

We used test set to predict the price and compare the predicted price with the actual price. The figure 20 showed the sample of predicted house price value with the actual price value.

| | Name | Address | LotSize | Date | ActualPrice | PredictedPrice |
|---|---|---|---|---|---|---|
| 1 | 11 MONT KIARA | D-30-2, JALAN KIARA 1 | 3315 | 20/10/2016 | 2,500,000 | 2,807,221 |
| 2 | WANGSA MAJU SEKSYEN 1 | 102, OFF JALAN 3/27A, OFF JALAN 1/27A | 531 | 20/10/2016 | 238,000 | 167,483 |
| 3 | DESA PANDAN FLAT BLOCK B | B13-4-6, DESA PANDAN | 721 | 20/10/2016 | 130,000 | 171,576 |
| 4 | SERI MAS | 5-9-1, JALAN 4/89A | 958 | 20/10/2016 | 365,000 | 327,317 |
| 5 | BUKIT ANGGERIK | 35-3, JALAN 2/154 | 603 | 20/10/2016 | 56,000 | 88,729 |
| 6 | PRIMA DAMANSARA | 8D-1-1 (D31), JALAN CHEMPENAI | 1356 | 20/10/2016 | 1,300,000 | 1,241,326 |
| 7 | VISTA MUTIARA | B-18-19, BATU 61/2, JALAN KEPONG | 1001 | 20/10/2016 | 485,000 | 405,991 |
| 8 | SUTERA BUKIT TUNKU | B-3-2, JALAN TUN ISMAIL | 5522 | 20/10/2016 | 3,700,000 | 2,854,916 |
| 9 | LAMAN SURIA | 00-05, OFF JALAN BUKIT KIARA BAYU | 506 | 20/10/2016 | 67,100 | 388,375 |
| 10 | PUTRA SURIA RESIDENSI | B-15-6, JALAN JELAWAT 2 | 748 | 20/10/2016 | 380,000 | 354,116 |
| 11 | ROYAL DOMAIN SRI PUTRAMAS 2 | B2-26-01, JALAN PUTRAMAS | 1173 | 20/10/2016 | 560,000 | 550,423 |
| 12 | ROYAL DOMAIN SRI PUTRAMAS 2 | A1-15-07, JALAN PUTRAMAS | 1130 | 20/10/2016 | 580,000 | 557,044 |
| 13 | SUTERA BUKIT TUNKU | B-3-1, JALAN TUN ISMAIL | 5522 | 20/10/2016 | 3,700,000 | 2,854,916 |
| 14 | HAMPSHIRE RESIDENCES | M1B-4-210, PERSIARAN HAMPSHIRE OFF JALAN AMPANG | 2852 | 20/10/2016 | 2,187,000 | 2,138,242 |
| 15 | KETUMBAR HEIGHTS | A-10-5, JALAN 6/95B | 755 | 21/10/2016 | 315,000 | 306,683 |
| 16 | BAYU TASIK 2 | A-4-9, JALAN PERMAISURI 5 | 926 | 21/10/2016 | 420,000 | 389,686 |
| 17 | WINNER COURT B | 16-4-10, JALAN 1/125A | 829 | 21/10/2016 | 320,000 | 284,758 |
| 18 | THE OVAL | 36-1, LORONG KUDA | 7793 | 21/10/2016 | 7,500,000 | 7,773,202 |
| 19 | MANDA'RINA COURT | B-3A-1, JALAN 10/142 | 775 | 21/10/2016 | 360,000 | 311,992 |
| 20 | SERI MALAYSIA | 09-27, JALAN 3/141 | 657 | 21/10/2016 | 150,000 | 111,764 |
| 21 | SEGAR APARTMENT | D-04-173, JALAN 1A | 667 | 21/10/2016 | 160,000 | 181,202 |
| 22 | MERCU SUMMER SUITES | B-23-05, 8 JALAN CENDANA | 493 | 21/10/2016 | 480,000 | 488,550 |
| 23 | BUSTAN SHAMELIN | 13-2-9, JALAN 2/91A | 1023 | 21/10/2016 | 330,000 | 274,463 |
| 24 | PKNS KAMPUNG BARU | 20R, RAJA MUDA MUSA | 843 | 21/10/2016 | 225,000 | 190,488 |

*Figure 23: Housing Price Prediction Sample using XGBoost*

# 6. Conclusion

We have successfully doing data analytic in finding the importance feature that affect the housing price. We also managed to develop prediction model that capable of predicting the housing price with profound accuracy. We concluded that we met our objective

# Bibliography

Basu, A. a. (1998). Anlysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*.

Gang-Zhi Fan, S. E. (2006). Determinants of House Price: A Decision Tree Approach. *Urban Studies, Vol 43, No 12*, 2310-2315.

Kenskin, B. (2008). Hdonic Analysis of Price in the Instanbul Housing Market. *International Journal of Strategic Property Management*.

*Malaysia House Price Index*. (n.d.). Retrieved from https://tradingeconomics.com/malaysia/housing-index.

Malpezzi. (2003). Hedonic pricing models: A selective and Applied Review. *Housing Economic & Public Policy*.

Marjan Ceh ̌, M. K. (2018). Estimating the Performance of Random Forest versus Multiple Regression for Prediciting Prices of the Apartment. *International Journal of Geo-Information*.

Quigley, J. F. (1970). Measuring theValue of Housing Quality. *Jornal of American Statiscal Association*.

Tianqi Chen, C. G. (2016). XGBoost: A Scalable Tree Boosting System.

Yakov Frayman, B. F. (2007). Solving Regression Problems Using Competitive Ensemble Models.