**HOUSE PRICE PREDICTION**

WQD 7005 – DATA MINING PROJECT

Name: Fatin Nabilah Abdul Raman

Matric No.: WQD190011

**Introduction**

For prospective buyers, developers, investors, appraisers, tax assessors, and other players in the real estate market, such as mortgage lenders and insurers, it is critical to have an accurate prediction of the house price [1]. Traditional estimation of house prices is based on a buy and sell price comparison without an accepted standard and a certification process. The development of a house price prediction model therefore helps to fill a significant information gap and increase the real estate market's performance [2].

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses in Boston’s suburbs. Once we get a good fit, we will use this model to predict the monetary value of a house located at the Kuala Lumpur area. A model like this would be very valuable for those buyers and a real estate agent who could make use of the information provided in a daily basis. This project we set to both discover some more insightful findings about Kuala Lumpur house property and to use machine learning to try and predict property values across the city. For this project, what are we trying to answer?

* What are the important features and factors that impact house prices?
* Can we build a model focusing on these important features and estimate accurately the cost of a house?

To answer the above question, the data will be tested and analyse for more details. Therefore, the machine learning model that will be use are Linear Regression, Gradient Boosting and Decision Tree. This model will be run or tested using the SAS Enterprise Miner with some training and validation dataset.

**Objective**

The objective for this project are:

* Interpret data into useful information
* To mine the listing price information from Home Trovit data in Kuala Lumpur
* To estimate the selling price of house based on a set of predictor variables

**Dataset**

The data was crawl from the Home Trovit website(<https://homes.trovit.my/>) and we only focus on the Kuala Lumpur state. 7903 row of data has been retrieved or collected with 12 column or features.

The features can be summarised as below:

* Title: This is the title or property name.
* Location: This the address of the property
* property\_details: This is the details of the property.
* url: This URL of the property.
* Image:This is Image of the property.
* source\_info:This is the source where this property is listed.
* published\_days: This is the numbers of days this property been published.
* Price: This is the Price of the property.
* no\_of\_bedroom: This is the Number of Bedroom available for the property.
* no\_of\_bathroom: This is the Number of Bathroom available for the property.
* property\_size: This is the property size in square feet for the property.
* property\_type: This is the property type for the property. (eg : Plex, House, Condo)

Due to the limitation of the HomeTrovit page which they can only display up to 100 pages with 25 property per page, the data was extracted from the recent published and been filter as per the property type available. There are 5 type of property has been extracted which is Apartment, Condo, Bungalow, House and Plex. Home Trovit website have this feature where we can select the property type that we want to extract.

**Data Pre-processing**

Next, the data will undergo some pre-processing or clean up. Below are some of the pre-processing data that has been done:

* Remove duplicate data
* Remove null data
* Remove the irrelevant column (example: URL, image,address,property\_details, etc)
* Added new column for district
* Replace and identify the district available in the given address
* Remove outliers and weird observation (example sqft 99999999)
* Replace and remove the symbols for price
* Filter and select the data that the price below than 1000000 or 1 million

The complete or clean dataset has contained 4106 rows with 7 column or attributes. The District was been classified to 12 groups as below:

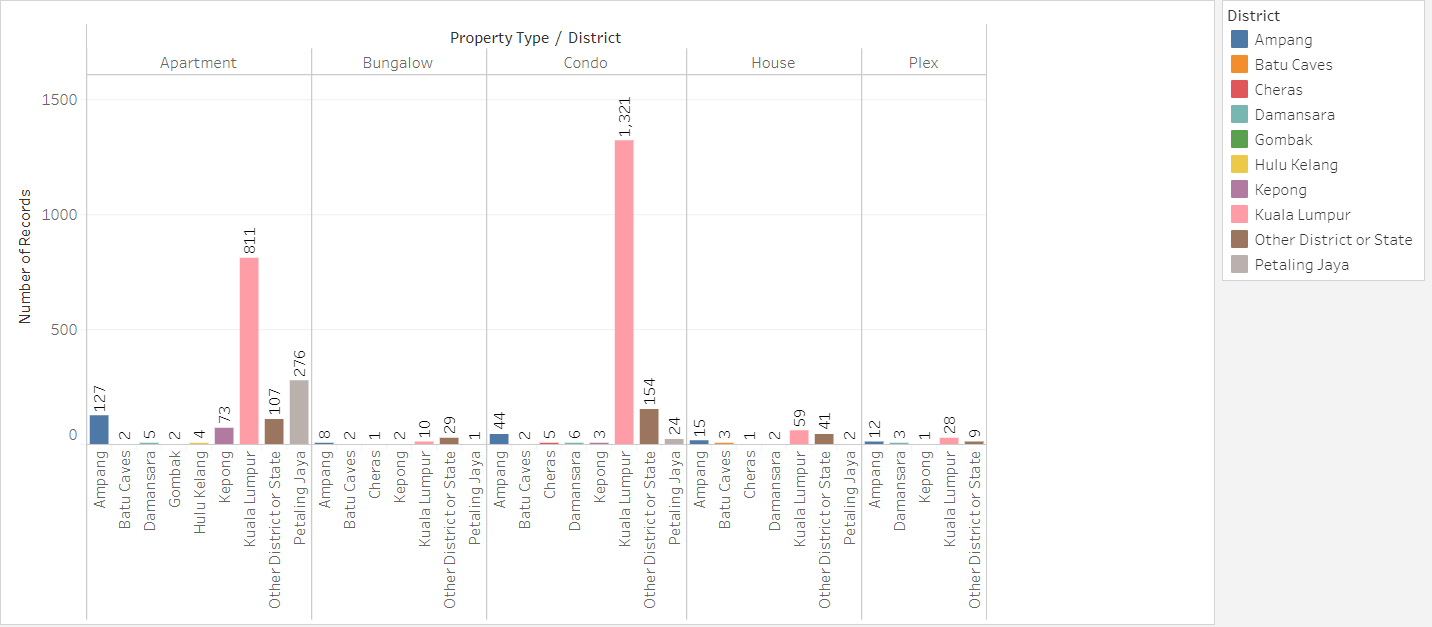
* [Ampang](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/ampang.html)
* [Batu Caves](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/batu_caves.html)
* [Cheras](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/cheras.html)
* [Damansara](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/damansara.html)
* [Gombak](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/gombak.html)
* [Hulu Kelang](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/hulu_kelang.html)
* [Kepong](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/kepong.html)
* [Petaling](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/petaling.html)
* [Petaling Jaya](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/petaling_jaya.html)
* [Sentul](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/sentul.html)
* [Setapak](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/setapak.html)
* [Sungai Besi](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/sungai_besi.html)
* [Kuala Lumpur](https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/kuala_lumpur.html)

The list or districts was retrieved or obtain from the online website (<https://geographic.org/streetview/malaysia/kuala_lumpur/kuala_lumpur/index.html>). However only 9-10 district was present in the dataset and there are some other state was also mention in the data which later been identify as the Other District/State.

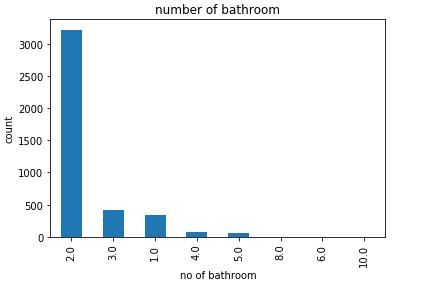
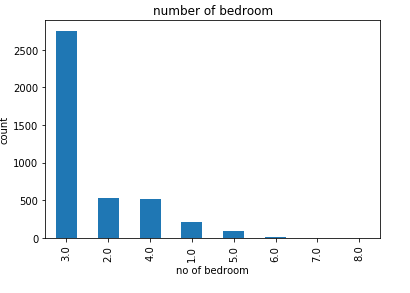
**Data Exploration or Visualization**

Next, we are going to see some visualization for the data. This visualization of data will involve the scatter plots and bar chart. This will involve between price and others variable. This is where we are going to see or to present insights so that it enables the decision makers to see analytics presented visually, so that they can grasp the pattern. Below is the visualization of the data:

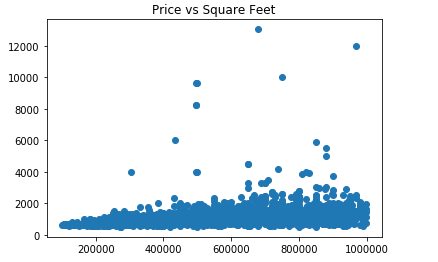
For the first image, it is show that the total of records for the Property Types per District which was showed that Property Type: Condo with District: Kuala Lumpur has the highest record (1321) in the dataset along with the Property Type: Apartment with District: Kuala Lumpur for the second highest record (811). This District Kuala Lumpur is the other district such as Bukit Bintang, KL Sentral and Lake Gardens, Bukit Nanas etc which is the most expensive land.



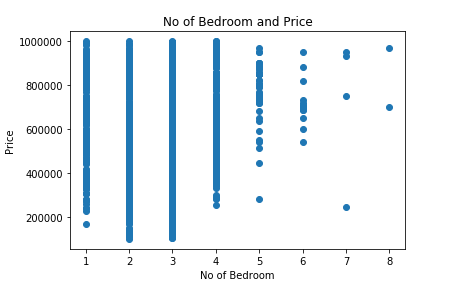
The next image, it is showing the bar graph of the frequency of number of bedroom and the frequency of number of bathrooms that available in dataset. It is shows that the most common house or property has 3 number of bedroom and 2 bathroom.

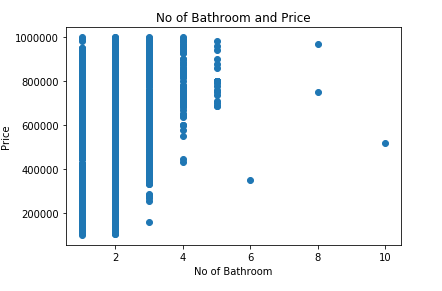


Next, is the scatter plot graph for Price Vs Property Size in Square Feet. It is shows that less 2000 square feet of living area has the range of price RM400000 to RM1000000. And some also shows that the larger living area or property size, the higher the price. It also shows that some of house or property offering less 2000 square feet living area with less than RM400000 price.

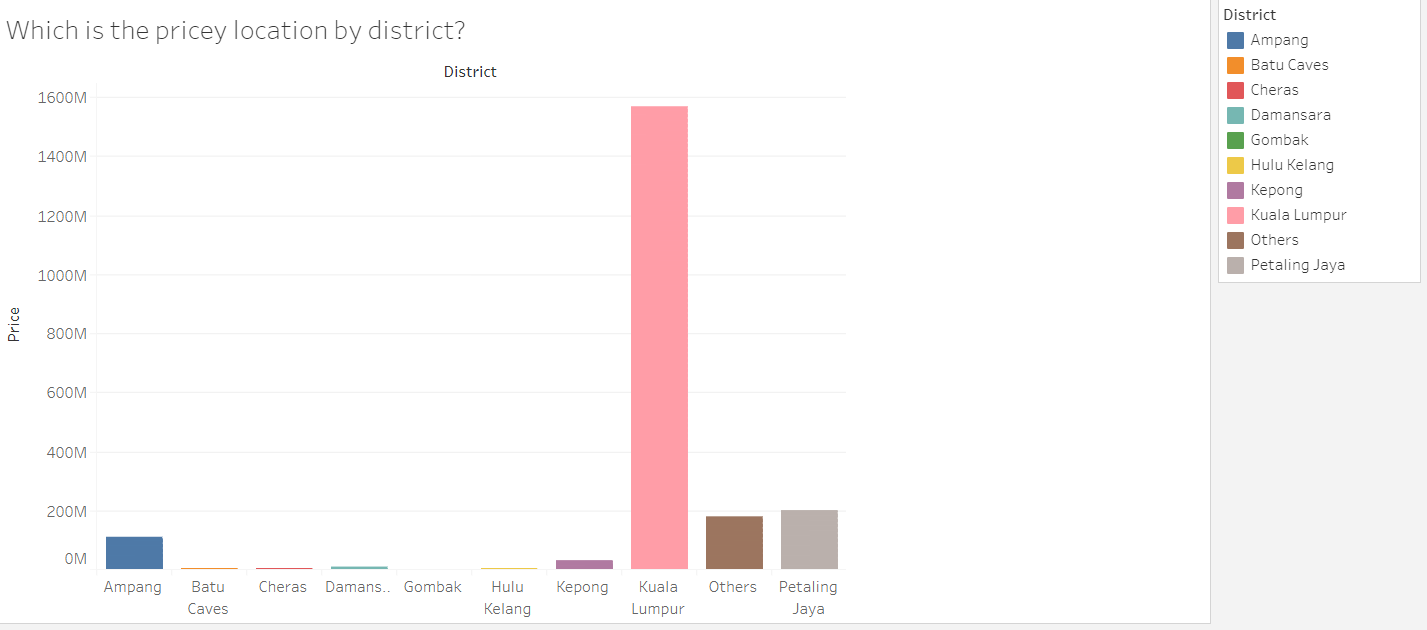


The next image, the scatter plot graph of the Number of Bedroom VS Price and Number of Bathroom VS Price. For this, we can identify that there 2 and 3 bedroom has the range of price RM200000 to RM1000000. And we also can see 1-2 have a room of 5 and 7 with price range below RM300000. Besides that, we can identify that there 2 number of bathrooms the range of price RM200000 to RM1000000. And we also can see 1-2 have a bathroom of 6 and 10 with price range below RM600000.

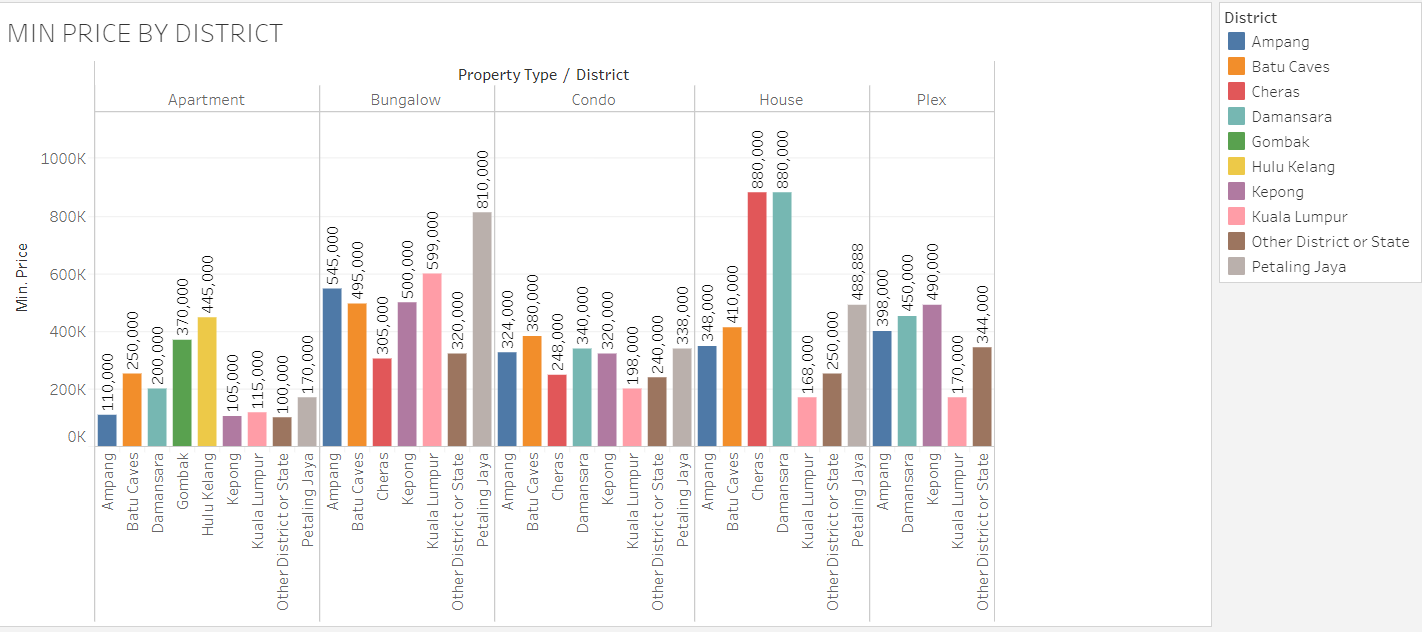




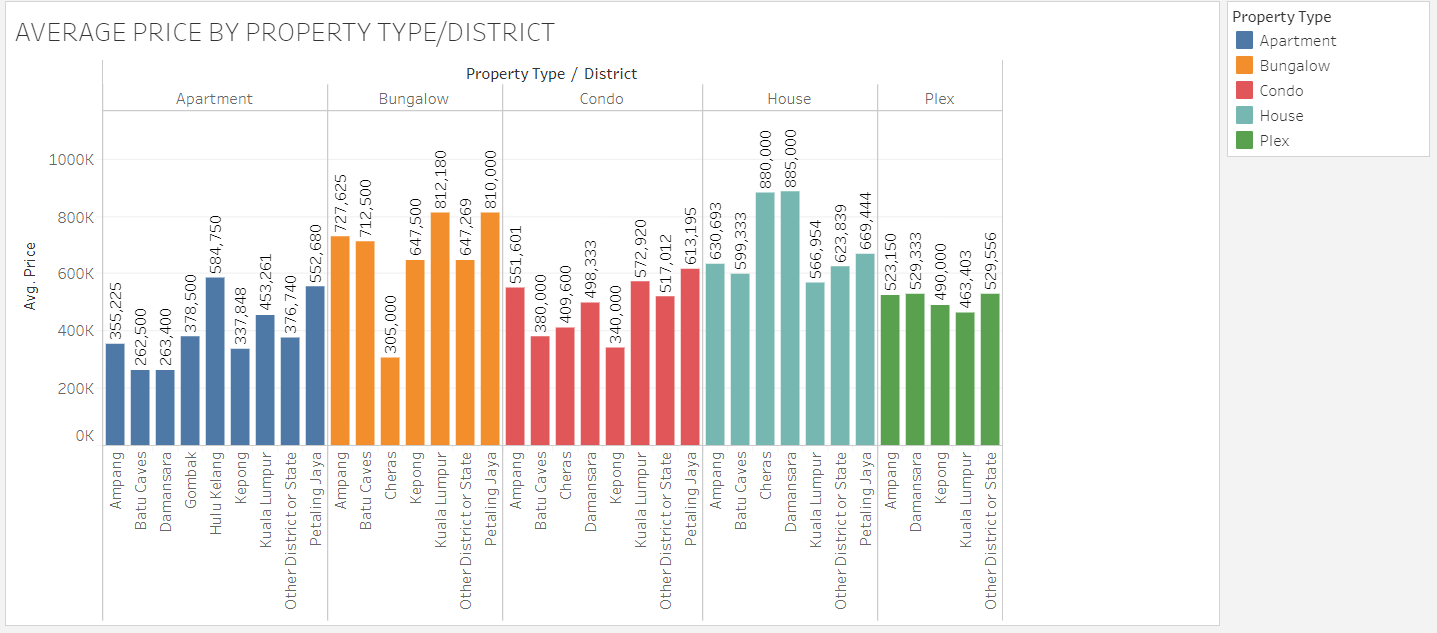
For the next image, it is a graph to identify which is the expensive location or district. It shows that Kuala Lumpur has the most expensive house or property. As per mention before this, Kuala Lumpur district is consist of Bukit Bintang, Kuala Lumpur City Centre and etc which has the most expensive land in Kuala Lumpur. Besides that, we can also see that some of the other district has the most lowest price such as Cheras, Batu Caves, Damansara, Gombak and Hulu Kelang.



The next image, is the graph to visualize the minimum, maximum and the average price of the house or property per District and Property Type in the dataset. So we can see that the Min Price graph is the Property Type: Apartment Price : RM100000 and District : Others and Property Type: Apartment Price : RM105000 and District : Kepong. For the Maximum Price Graph is Property Type: Condo Price : RM999000 and District : Kuala Lumpur.







**Model Comparison**

In this project, below are the dependent and independent variable for the data.

|  |  |
| --- | --- |
| Dependent Variable | Independent Variable |
| Price | No\_of\_bedroom |
| No\_of\_bathroom |
| District |
| Property\_size |
| Property\_type |

The clean data with the csv format file will be converted to SAS dataset or sas7bdat file format using the SAS Studio. Below are the code for the conversion:

*/\* Generated Code (MYLIB) \*/*

*/\* Source File: HomeTrovit\_clean.csv \*/*

*/\* Source Path: /folders/myfolders/sasuser.v94 \*/*

*%web\_drop\_table(MYLIB.DATA\_CLEAN);*

*FILENAME REFFILE '/folders/myfolders/sasuser.v94/HomeTrovit\_clean.csv';*

*PROC IMPORT DATAFILE=REFFILE*

*DBMS=CSV*

*OUT=MYLIB.IMPORT;*

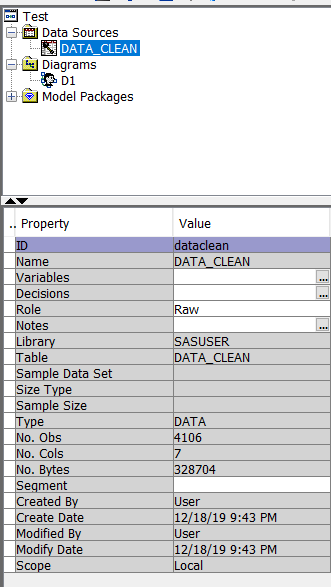
*GETNAMES=YES;*

*RUN;*

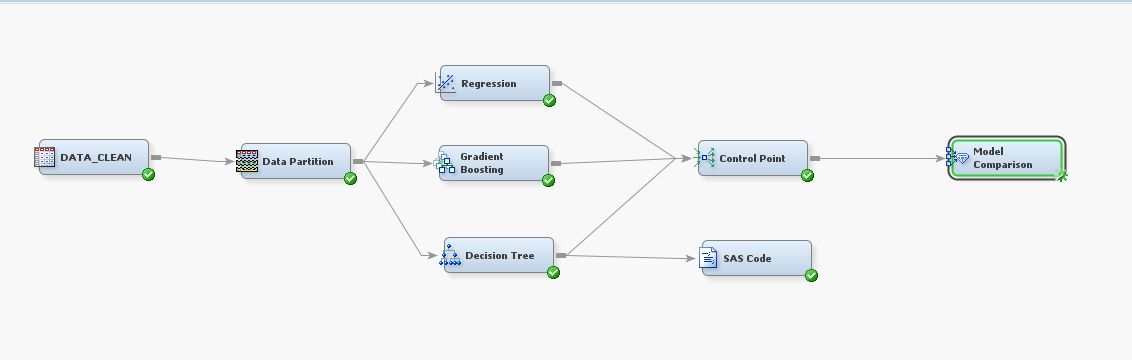
*PROC CONTENTS DATA=MYLIB.DATA\_CLEAN; RUN;*

*%web\_open\_table(MYLIB.DATA\_CLEAN);*

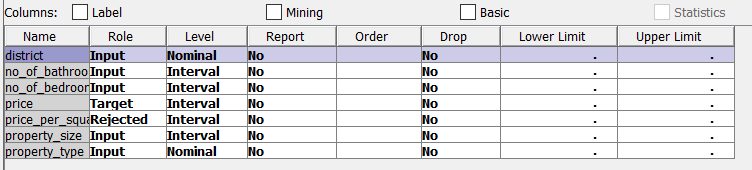
The clean data with sas7bdat will be loaded and imported as a Data Sources in the SAS Enterprise Miner with Role = Raw.



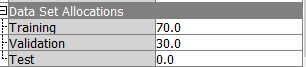
For this project, there are 3 models were use which is Linear Regression, Gradient Boosting and Decision Tree. The idea was to identify the best model which give the least average squared error, as it an indicator of the difference in the predicted value vs true value of the house price. The Lower Average squared error show indicates a better model.



Above shown a SAS Enterprise Miner diagram. We use model comparison node to identify the model with least Average Squared error. Thus, we set the price as a target variable and other variable as below.



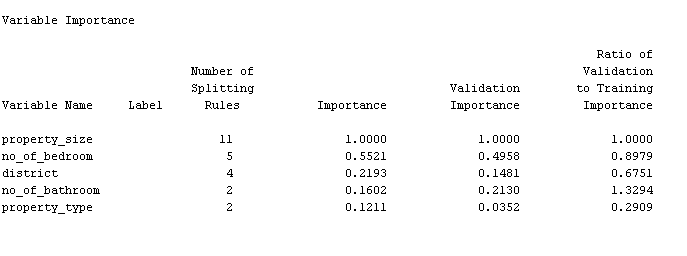
The data was partition into 70% for Training and 30% for validation.



**Result**

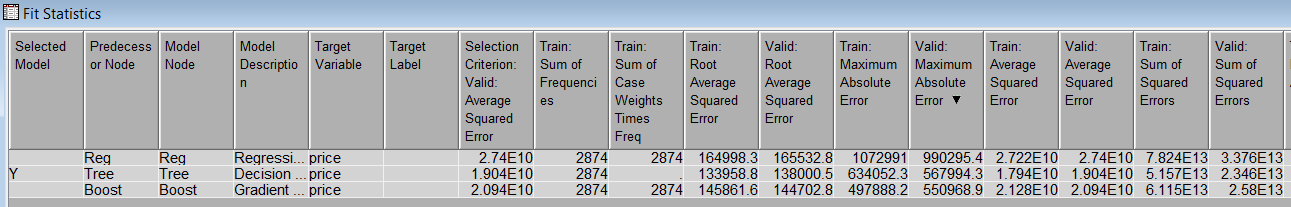
1. Identifying the important variable

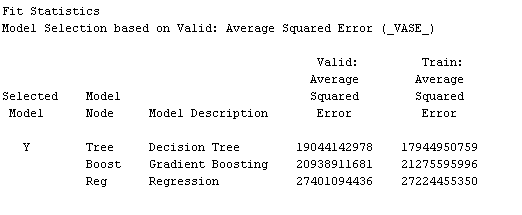
The decision tree model has identified the variables with the largest importance for the information gain to the target variable. As we can see, property\_size and other below variable are important variable that related or on estimation to the price or target variable.



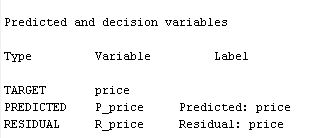
1. Model comparison Results

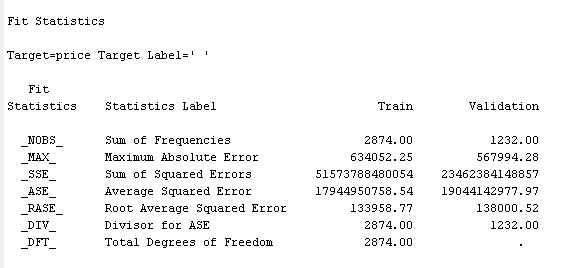
Based on the model comparison results, we see that Decision Tree model is the model with the least average squared error. We are using Linear Regression model, Gradient Boosting Model and Decision Tree model. The result of the all three model as in below.



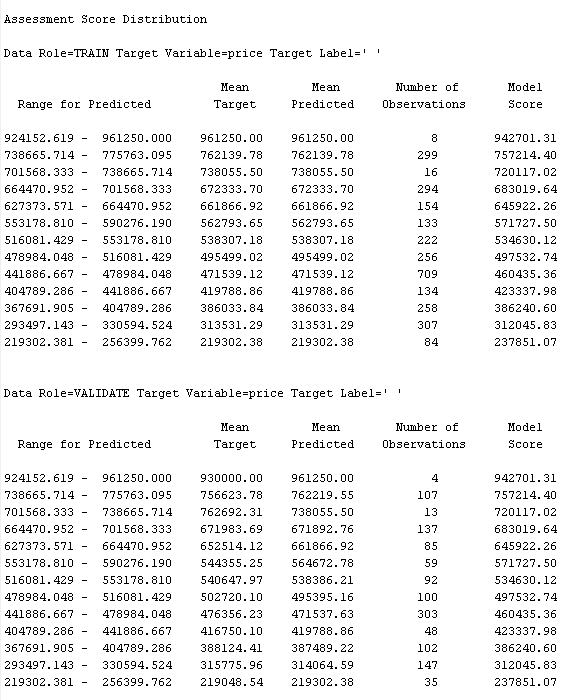


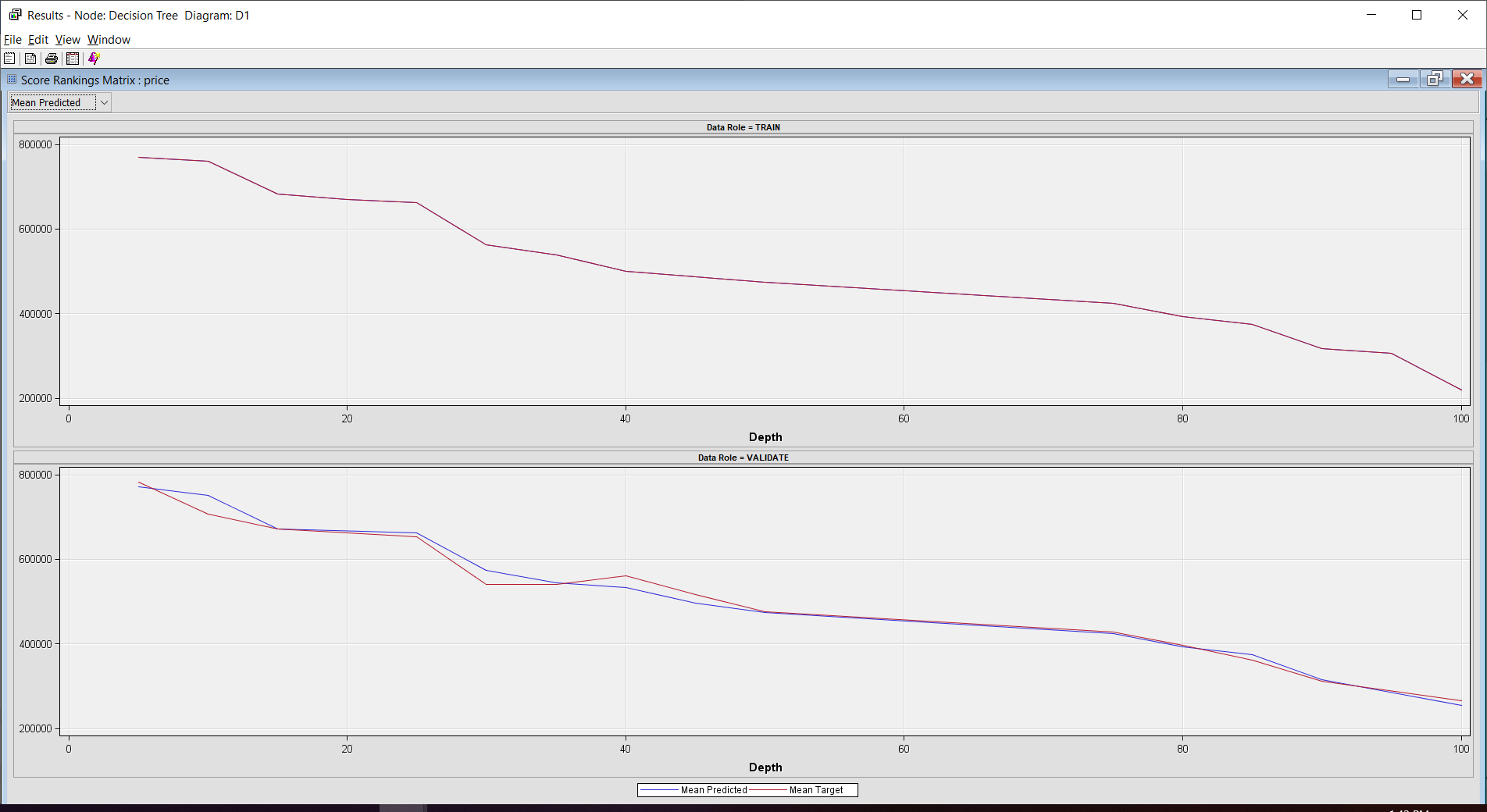
1. Decision Tree Model Result





It is show that the Root average squared error for the model 138000. This means that on the average the difference in predicted value from the actual value is 138000.





**Conclusion**

It is as expected that the property\_size variable would have a direct relation to Price and it is found that property\_type is the least important variable. However, the data is lacking with the other property details such as availability of garage or parking, year builts, floor number – for those apartment or condo or terrace house and etc so that we can identify other variable that might affect the valuation of the house price.

Reference:

To conduct this project the following tools have been used :

● Python 3.7

● Pandas (Library)

● BeautifulSoap (Library)

● Numpy (Library)

● SAS Studio

● SAS Enterprise Miner

The code and data are available here: <https://github.com/FatinNabilah1/HomeTrovit.git>

1. Frew, J., and G. D. Jud, 2003. Estimating The Value of Apartment Buildings, The J. Real Estate Res., 25: 77 - 86.
2. Calhoun, C. A., 2003. Property Valuation Models and House Price Indexes for The Provinces of Thailand: 1992 – 2000. Housing Finance International, 17: 31 – 41.