**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 Homw 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)

**Data Pre-processing**

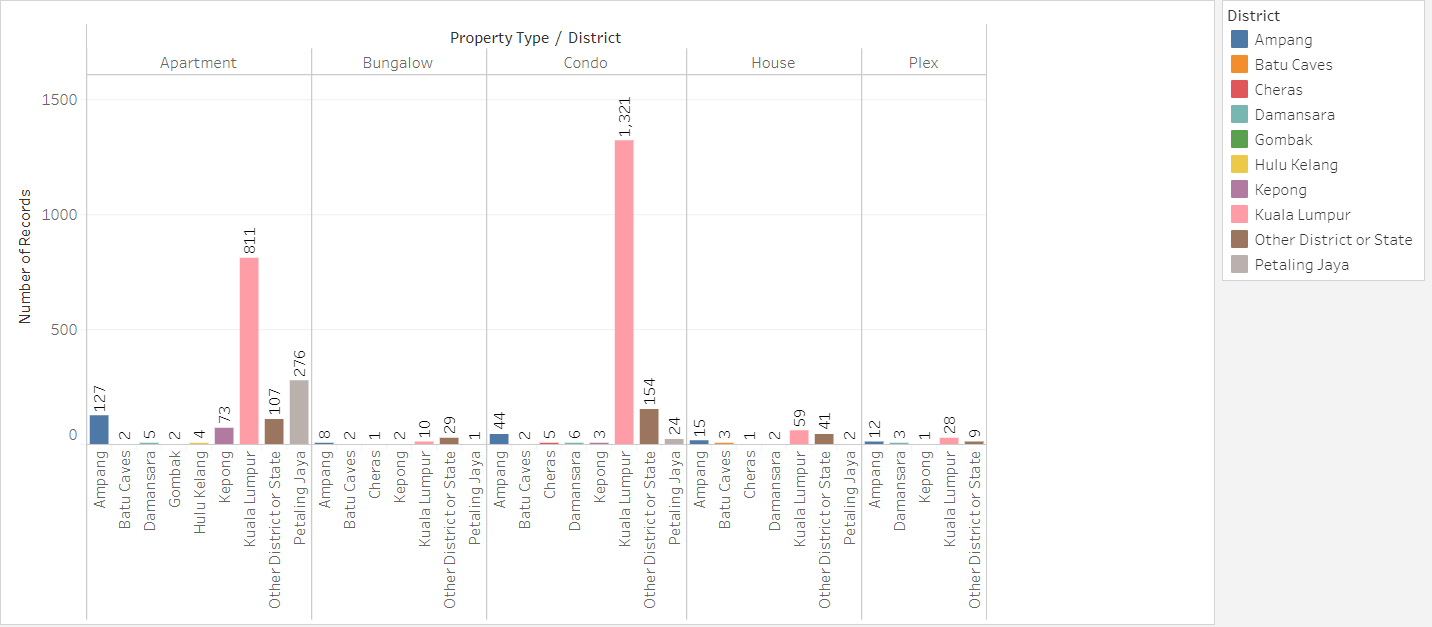
Next, the data will undergo some pre-processing or clean up. Below are some of the pre-processing data that has been done:

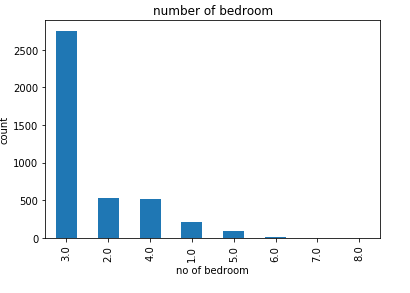
* Remove duplicate data
* Remove null data
* Remove the irrelevant column (example: URL, image,address,property\_details, etc)
* Added new column for district
* Replace and identify the district available in the given address
* Remove outliers and weird observation (example sqft 99999999)
* Filter and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

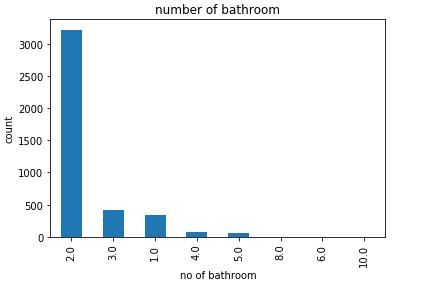
The complete or clean dataset has contained 4106 rows with 7 column or attributes.

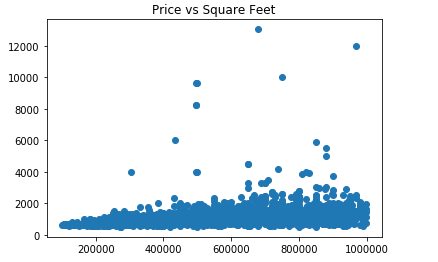
**Data Exploration or Visualization**

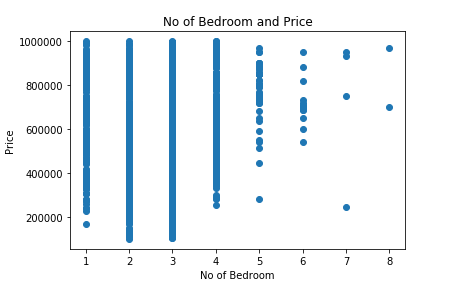
Next, we are going to see some visualization for the data. This visualization of data will involve the scatter plots and bar chart. This will involve between price and others variable. Below is the visualization of the data:

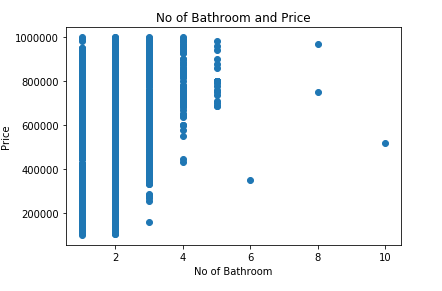


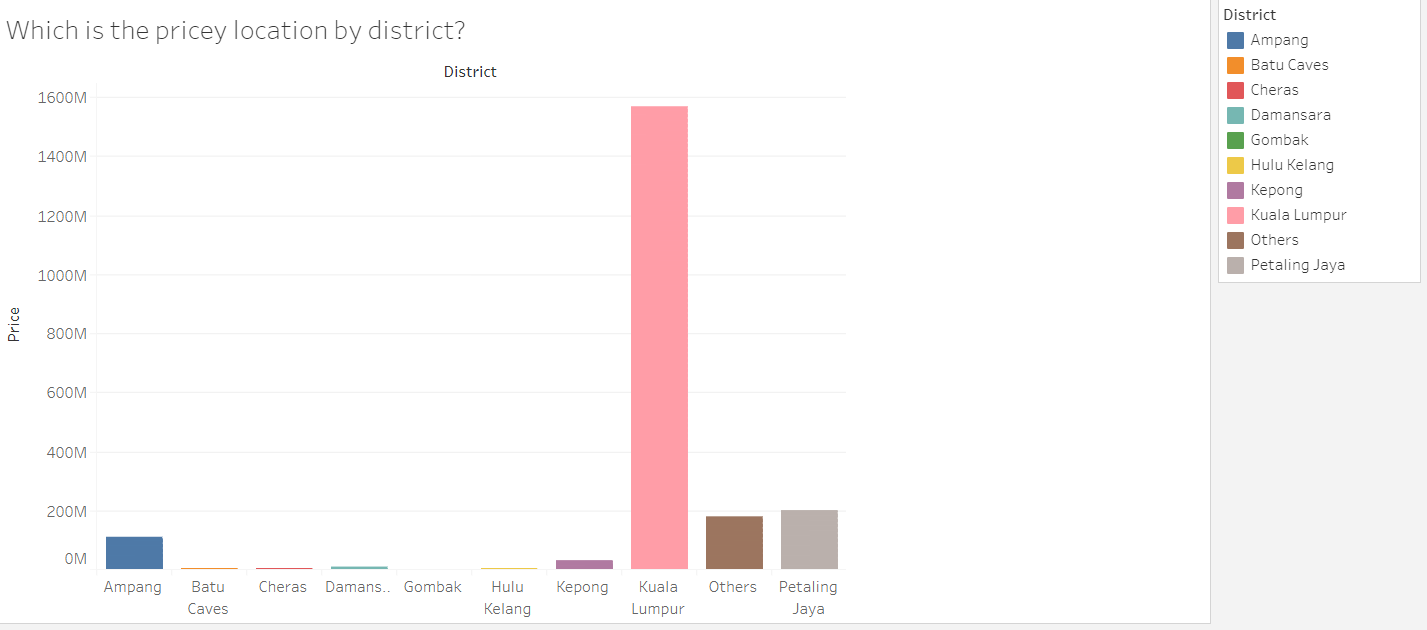


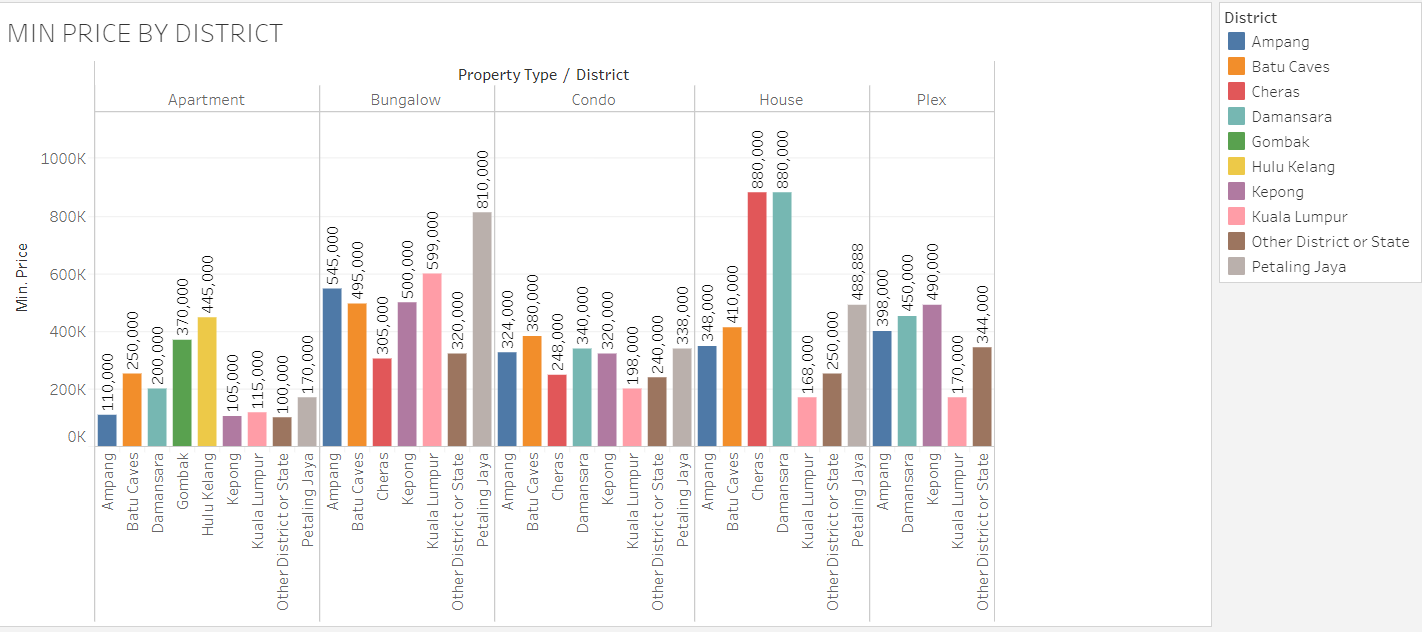




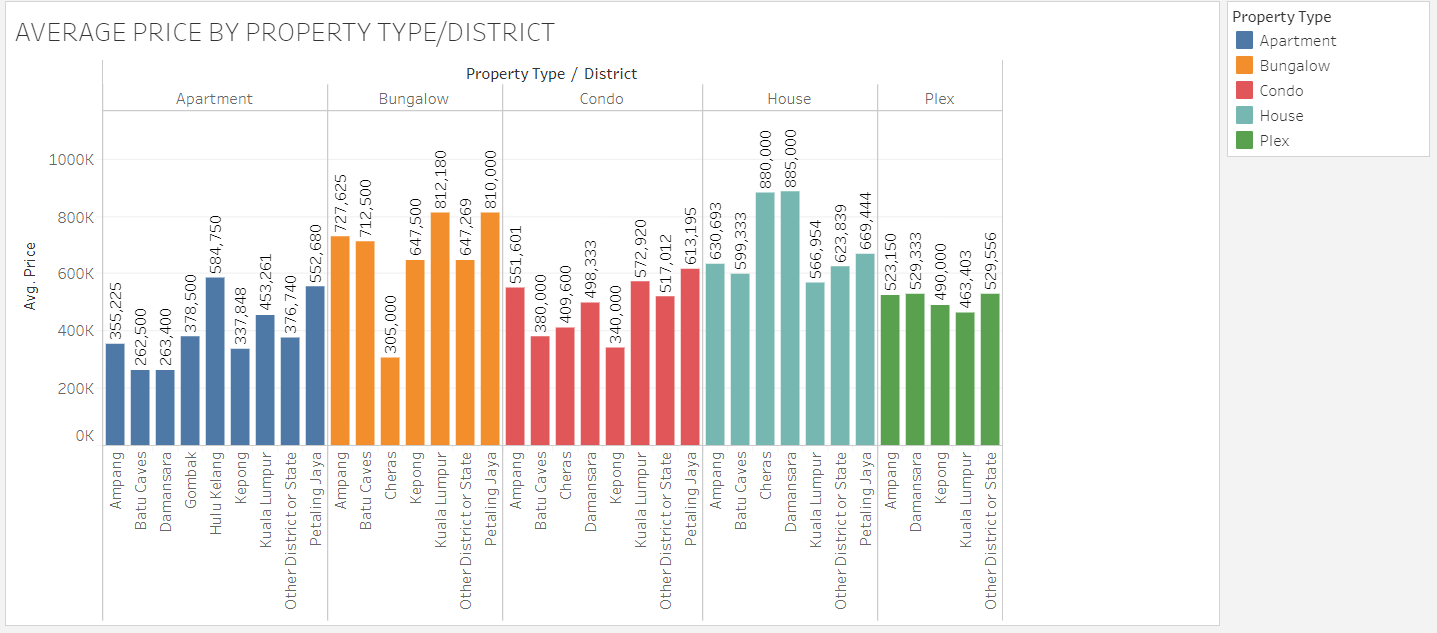










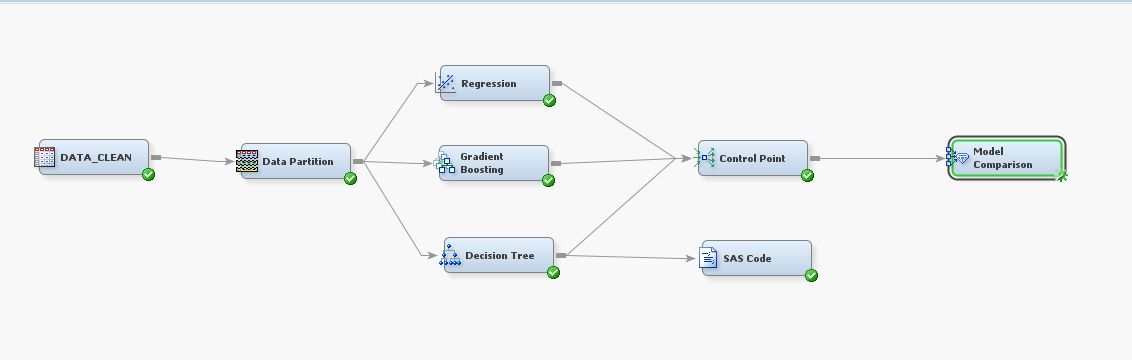


**Model Comparison**

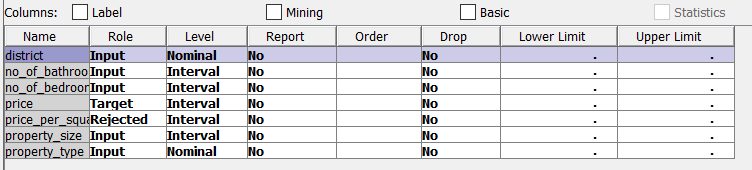
In this project, below are the dependent and independent variable for the data.

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

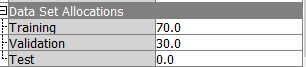
For this project, there are 3 models were use which is Linear Regression, Gradient Boosting and Decision Tree. The idea was to identify the best model which give the least average squared error, as it an indicator of the difference in the predicted value vs true value of the house price. The Lower Average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



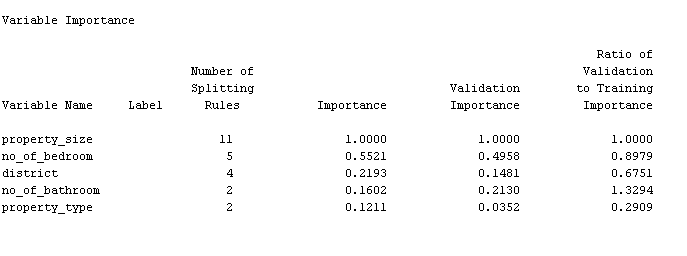
The data was partition into 70% for Training and 30% for validation.



**Result**

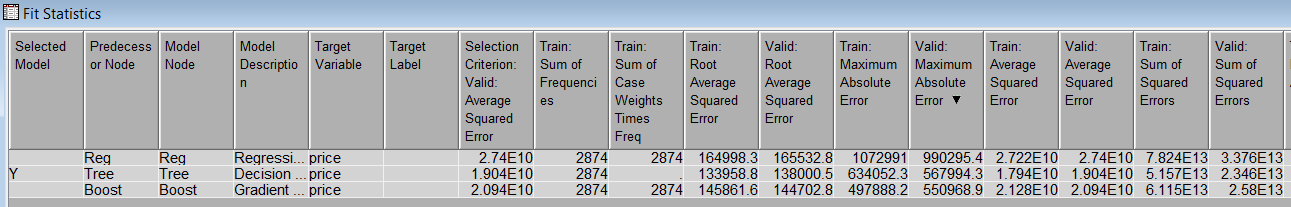
1. Identifying the important variable

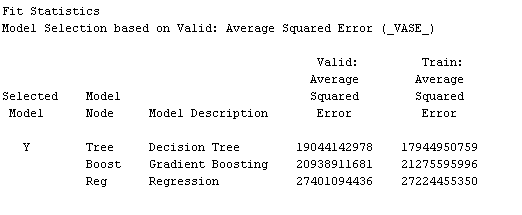
The decision tree model has identified the variables with the largest importance for the information gain to the target variable. As we can see, property\_size and other below variable are important variable that related or on estimation to the price or target variable.



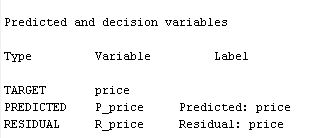
1. Model comparison Results

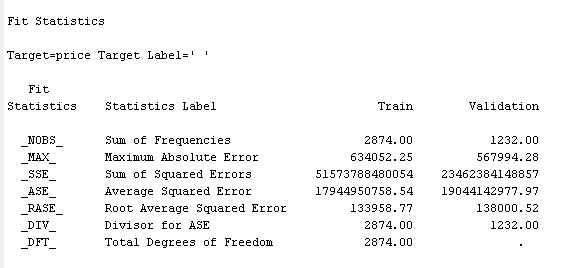
Based on the model comparison results, we see that Decision Tree model is the model with the least average squared error. We are using Linear Regression model, Gradient Boosting Model and Decision Tree model. The result of the all three model as in below.



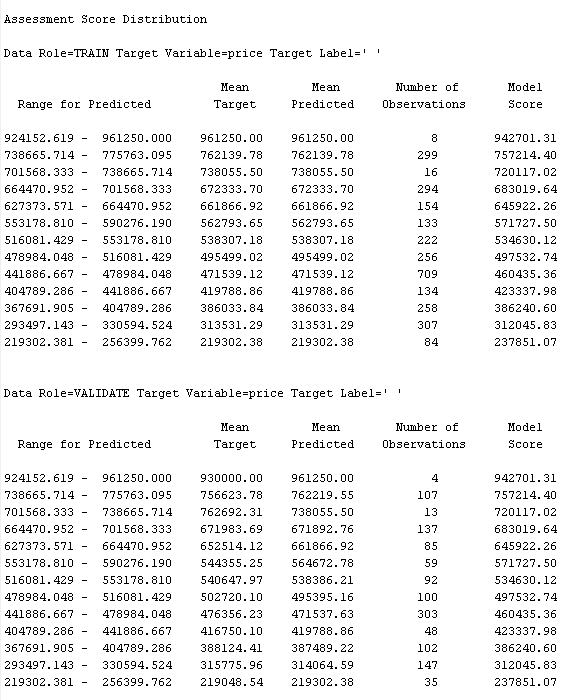


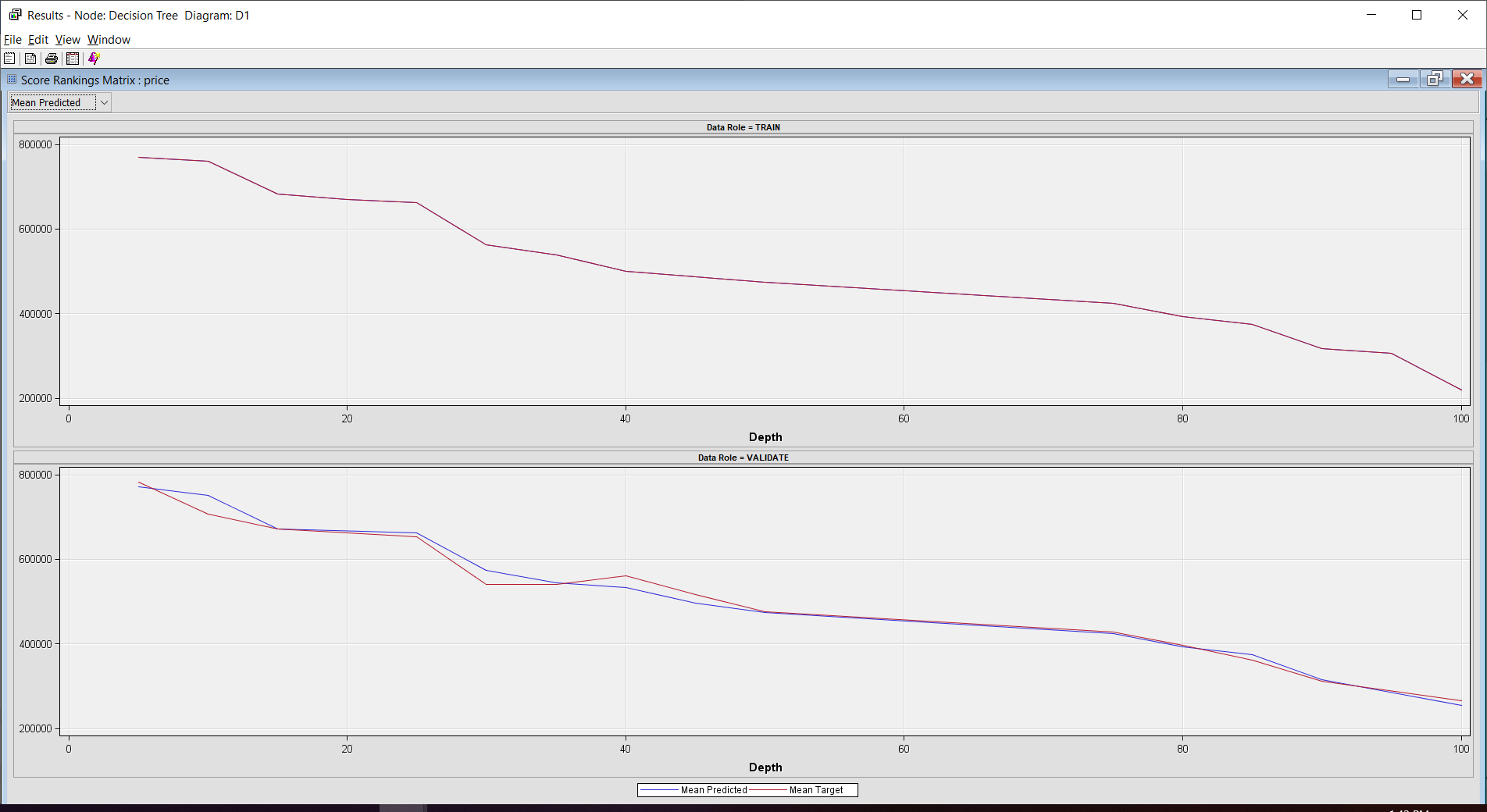
1. Decision Tree Model Result





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:

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.
3. The code available here: <https://github.com/FatinNabilah1/HomeTrovit.git>