CONDOMINIUM PRICES IN SINGAPORE

By: HOANG UYEN LE

1. Introduction

1.1 Business Problem

With National Development Minister Lawrence Wong's comments on the increasing supply of new private residential units in Singapore in the upcoming years (Wong, 2017), developers will be facing stronger competition in the industry. As shown in Figure 01, the number of planned private residential units is increasing from 2018 to 2019.

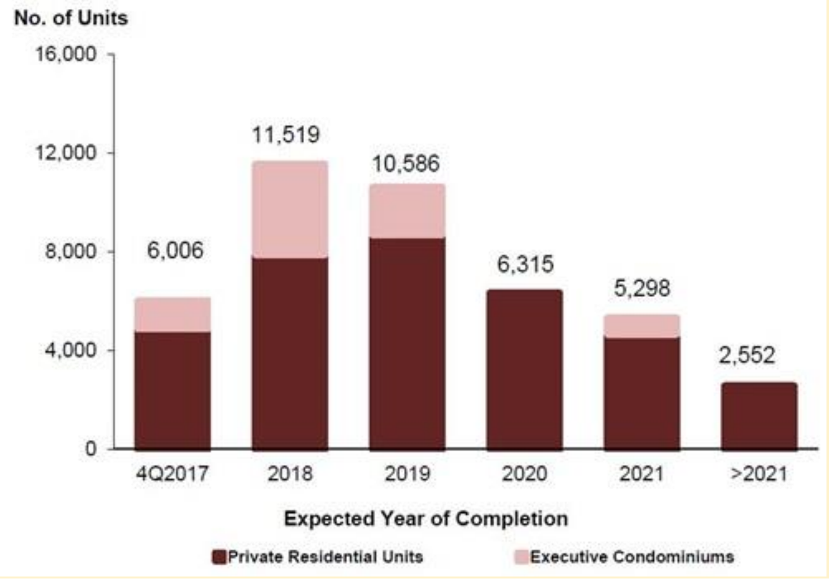


Figure 1: Number of planned Private Residential Units and Executive Condominiums in Singapore from 2017 to 2021

Furthermore, there is a rising demand of private residential units in Singapore as shown by an increasing number of private residential units sold by developers in Figure 2.

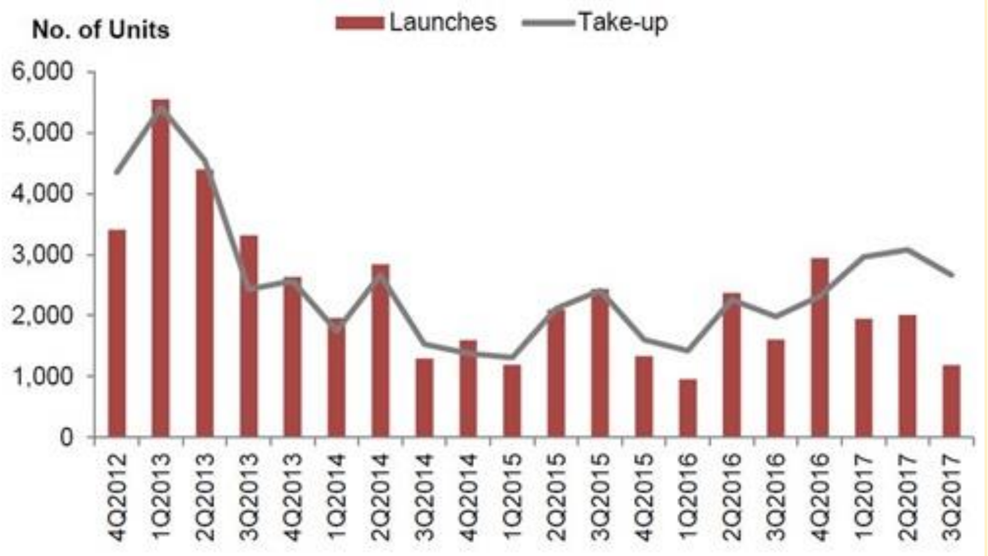


Figure 2: Number of Private Residential Unit Launches and Take-up in Singapore from 2012 to 2017

It is imperative for developers to be able to compete for these buyers by staying attractive in the market. The outcome of this report aims to provide models for developers to price their condominiums according to its variables. The two models selected are Classification Tree and Multiple Regression Model. Developers can price their new condominiums based on results of the model application

1.2 Managerial Decisions

This report aims to provide insights and outcomes by analysing using past data of private residential units, specifically condominiums. The two models selected are Classification Tree and Multiple Regression Model. This excludes Executive condominiums which are developed by the government to provide more affordable private housing to residents in Singapore. Therefore, the pricing of Executive Condominiums differ from that of other Condominiums of interest. Developers need to make pricing decisions for both resale and new flats from observing the willingness to pay of buyers based on historical data. The factors that affect this willingness to buy include the area of the flats, whether it's resale or new sale, leasehold of the flat, completion year, et cetera.

2. Data

2.1 Dataset

- Obtained from Singapore Real Estate Information System (REALIS)
https://spring.ura.gov.sg/lad/ore/login/mobile.cfm
- Including 8525 rows and 12 columns.
- Containing residential unit transactions that occurs from June 1st 2018 to May 31th 2019.
- Description of variables:
  + Area (sqm)
  + Transacted Price ($)
  + Unit Price ($ psm)
  + Unit Price ($ psf)
  + Sale Date
  + Tenure
  + Completion Date
  + Type of Sale
  + Purchaser Address Indicator
  + Postal Code
  + Planning Region
  + Planning Area

2.2 Data Cleaning stage

2.2.1 Adding Class variable based on transacted price

Range of transacted price was applied for classifiers involving class determination. The 25th percentile and 75th percentile of transacted price are used to segment transacted price into Low, Mid and High classes respectively. For condominiums with transacted price at most S$996,000, these condominiums are classified as the Low class. For condominiums with transacted price at least S$1,780,000, these condominiums are classified as the High class. For condominiums with transacted price between S$996,000 and S$1,780,00, these condominiums are classified as the Mid class.

2.2.2 Precluded Data

Due to the limitations imposed by R software in which only a maximum of 32 categories can be processed, 5 categories are removed from the planning area. These planning areas are Downtown Core, Mandai, Museum, Orchard and Outram. They are chosen due to the minimal condominium sale data in these areas.

2.2.3 NA and Outliers values.

The original dataset included 8525 rows. After checking some statistics index and skewness as well as the distribution of each variables, I removed 89 rows, and had a data with 8436 observation to work on it.

2.2.4 Variables to be modified

There are various existing categorical variables that needs to be modified before the model can be constructed, as they could not be measured on a quantitative scale. While handing the data set, Tenure was considered as complex character string. Getting rid of the unused information in each cells, this column turned to be category variables with 7 options.

3. Data visualization

When benchmarking the price time series and the transaction records (Figure 3 – 4), I can see that the price data seems to be a leading indicator (by 1-2 months) of the number of sale transaction: months of high prices will be followed by low volume of transaction and vice versa. For example, the months of January - February 2019 saw downward trend of housing price, and then in March I noticed an increase in the number of transactions. Then, in April when the prices started to surge, I can see the number of transaction dropped to two-thirds of that of February level. I can think of this as a market self-correcting mechanism to stabilize the housing prices.

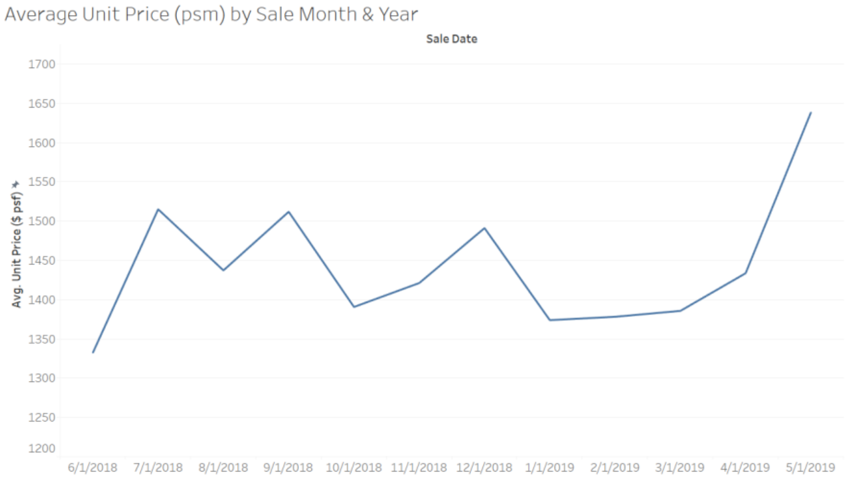


Figure 3; Unit price ($psm) time series

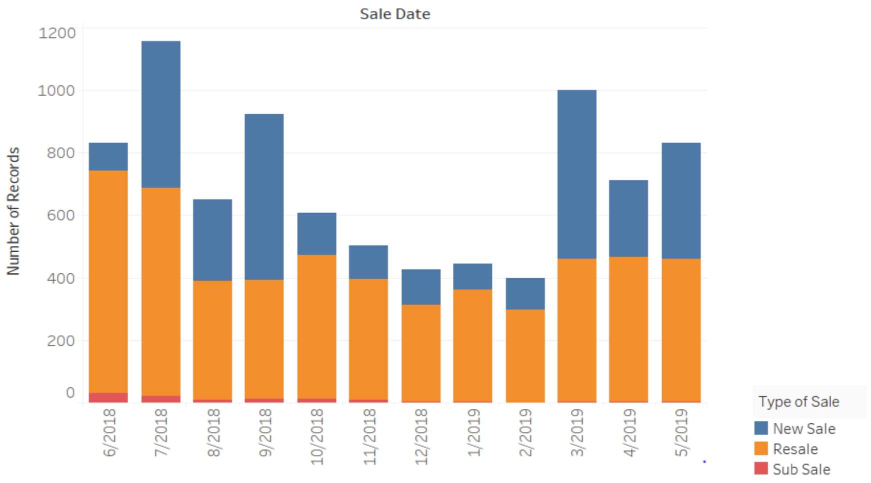


Figure 4: Transaction records by month

The original dataset has 2 factors are location-related, which are Planning Region and Planning Area. There are 5 planning regions and 37 planning areas (5 of which were excluded in this analysis, see note below in Data cleaning section). I can see that there are variations in the average unit price for different regions: the North and Central regions, known to be mature residential areas with good supply of nearby neighborhood amenities, have slightly higher average price, while the North East, North and West regions offer lesser housing choices and are offered lower price due to higher inaccessibility to the central area, where the majority of the office locations are. Toa Payoh area specifically enjoys a significantly high unit price, primarily due to it being central and also one of the oldest, most established residential hubs in Singapore. (Figure 5, 6)

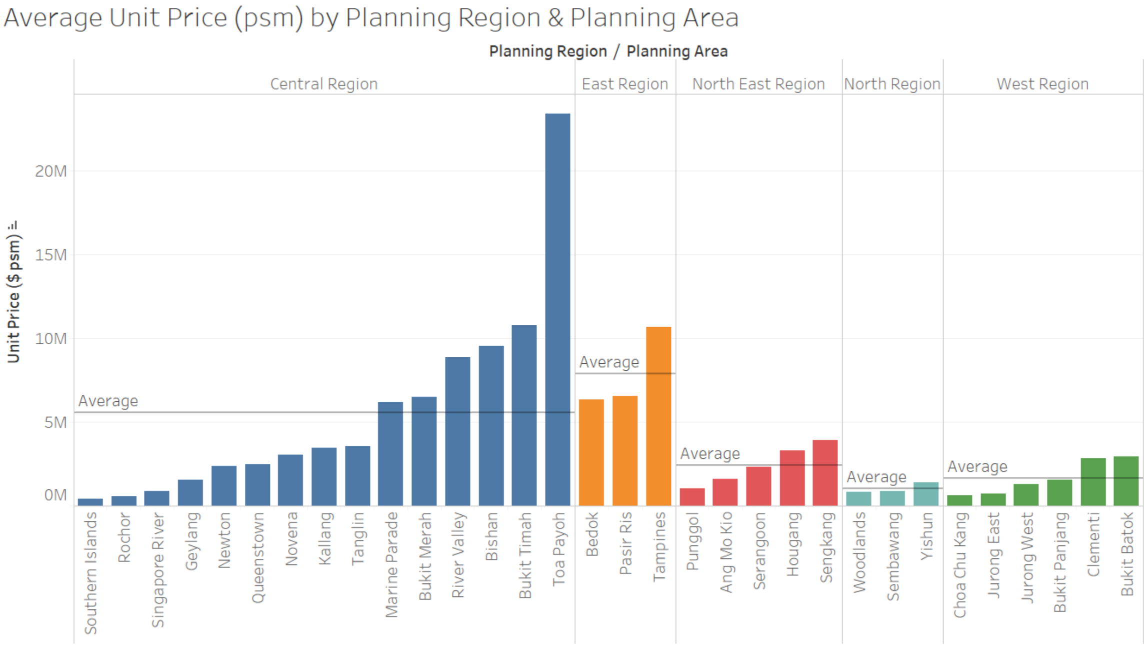


Figure 5: Average unit price (psm) by Planning Area and region

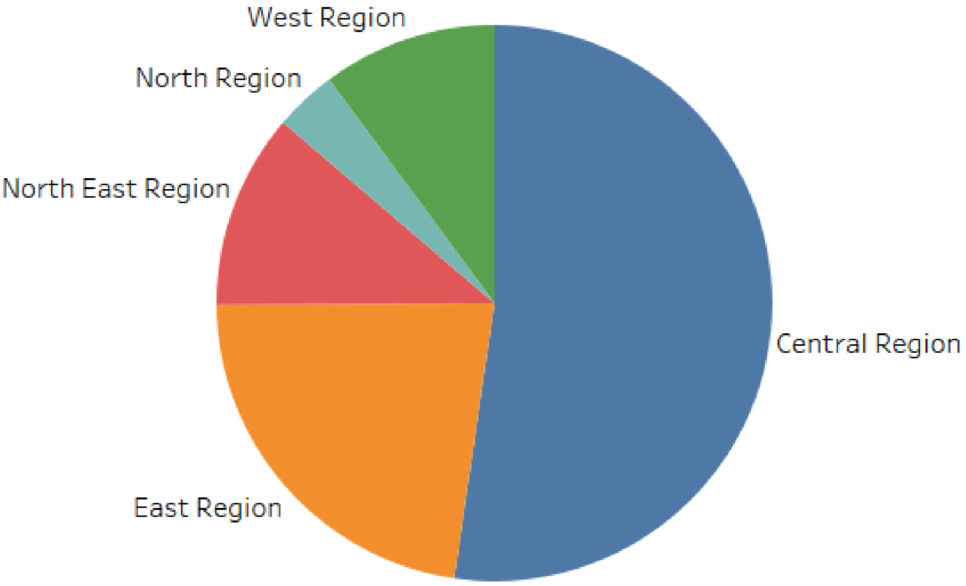


Figure 6: Distribution of housing transaction during 1-year period.

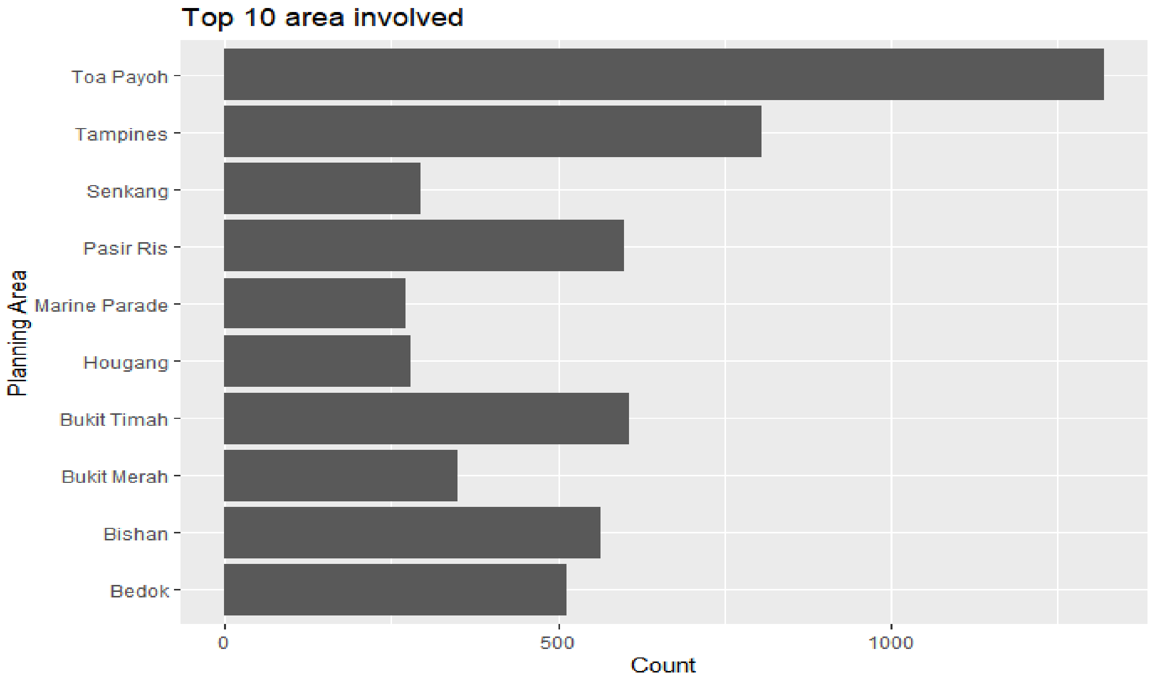


Figure 7

4. Multiple Regression

Regression analysis allows understanding of how the world operates and allows making of predictions. To obtain improved fit to the data, several explanatory variables could be used in the regression equation, and being known as multiple regression. The regression equation is still estimated by the least squares method.

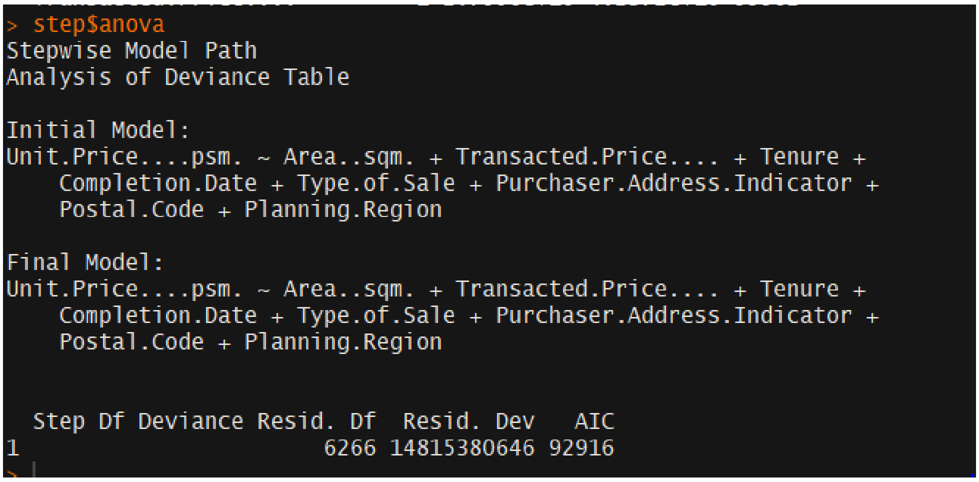
For building multiple regression model, I use different methods, specifically, by hand, by using stepIC() backwards, by sing LASSO method and PCR (Principle Components Regression) to built the models; then fitting them one by one into the test dataset. After that, comparing the predict accuracy of these model together to chose which on is the best models to predict dependent variable. This time, I chose Unit Price ($psm) to be predictors variables.

With Manual multiple linear regression

Equation:

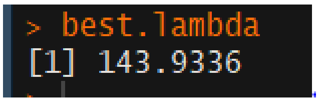
Unit Price ($psm) = (1.776e+04) – 99.69 * Area + (6.066e-03) * Transacted Price
– 1037 * Resale + 120 * Sub-sale + 271.3 * Private
– 233.4 * East region – 1909 * Northeast region
– 2803 * North region – 1857 * West region

With Step backward method: Based on Stepbackward result, the model contains 3 most significant variable named: Area, Transacted Price, Tenure, Completion Date, Type of Sale, Purchaser Address Indicator, Postal Code and Planning Region



```
> step$anova
Stepwise Model Path 
Analysis of Deviance Table

Initial Model:
Unit.Price....psm. ~ Area..sqm. + Transacted.Price.... + Tenure + 
    Completion.Date + Type.of.Sale + Purchaser.Address.Indicator + 
    Postal.Code + Planning.Region

Final Model:
Unit.Price....psm. ~ Area..sqm. + Transacted.Price.... + Tenure + 
    Completion.Date + Type.of.Sale + Purchaser.Address.Indicator + 
    Postal.Code + Planning.Region


  Step Df Deviance Resid. Df  Resid. Dev   AIC
1                       6266 14815380646 92916
```

With LASSO: I choose the best lambda i.e. first **lambda.min** where minimum error observed is 143.9336



```
> best.lambda
[1] 143.9336
```

With PCR: Use the validationplot() function either set to RMSEP, MSEP, and/or R2 to determine how many principle components should be used in the final model. In the case, it looks like 4 components are enough to explain more than 90% of the variability in the data.

Prediction accuracy: After using different methods to create 4 models and assessing model performance by calculating measures of prediction accuracy, I will choose models which has smallest RSS, MAE, RMSE and largest R2. Table below summarizes these measures of 4 models, and can be witnessed clearly from it, Lasso is the best model for predicting.

	R squared	**RSS**	**MAE**	**RMSE**
Manual model	0.8645	6796543263	1134.968	1795.171
Step backward	0.895	5267613977	1580.406	1075.927
LASSO	0.89773	5535801247	1044.6287	942.741
PCR	0.88568	6458439504	1100.6439	1263.475

5. Classification Tree

Classification Trees are one of the commonly used decision trees for predictive modelling. It predicts a categorical target variable and is often binary. Every single branch node represents a single variable and a split point of the variable, and the selected variable is able to generate the highest node purity. The leaf nodes contain the training data from which the most commonly occurring class is used to predict the class of the testing data. It is simple to understand and can be used to be presented to managers who are not familiar with other complex predictive model.

The Classification Tree that is generated will attempt to predict, for each condominium, which of a set of classes that condominium belongs to. Confusion matrix will be used to derive the in-sample and out-of-sample accuracy.

The variables actually used in the classification tree are:

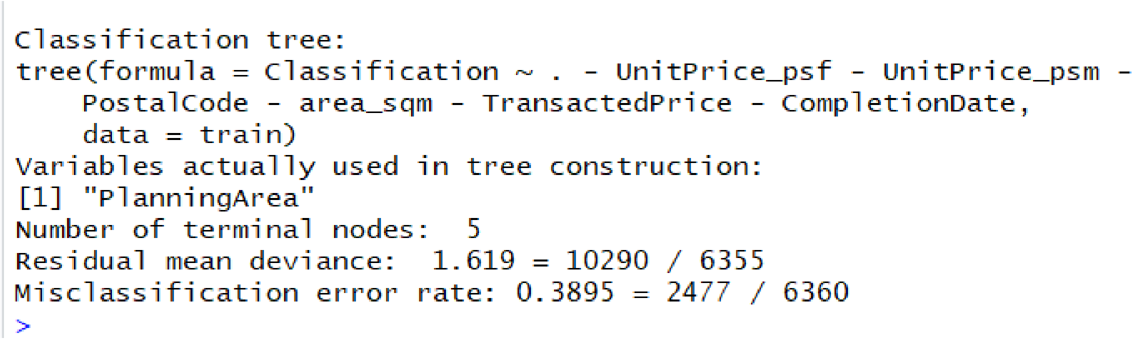
- Sale month & year
- Tenure
- Type of Sale (new sale, resale, or sub-sale)
- Type of Purchaser (based on purchaser address)
- Planning Region
- Planning Area* (*additional data processing, see note below)*

Using the tree package, I face a few limitations:

- Categorical variables cannot have more than 32 distinct values. This thus render values such as postal code, or area of the unit irrelevant and thus removed from the analysis.

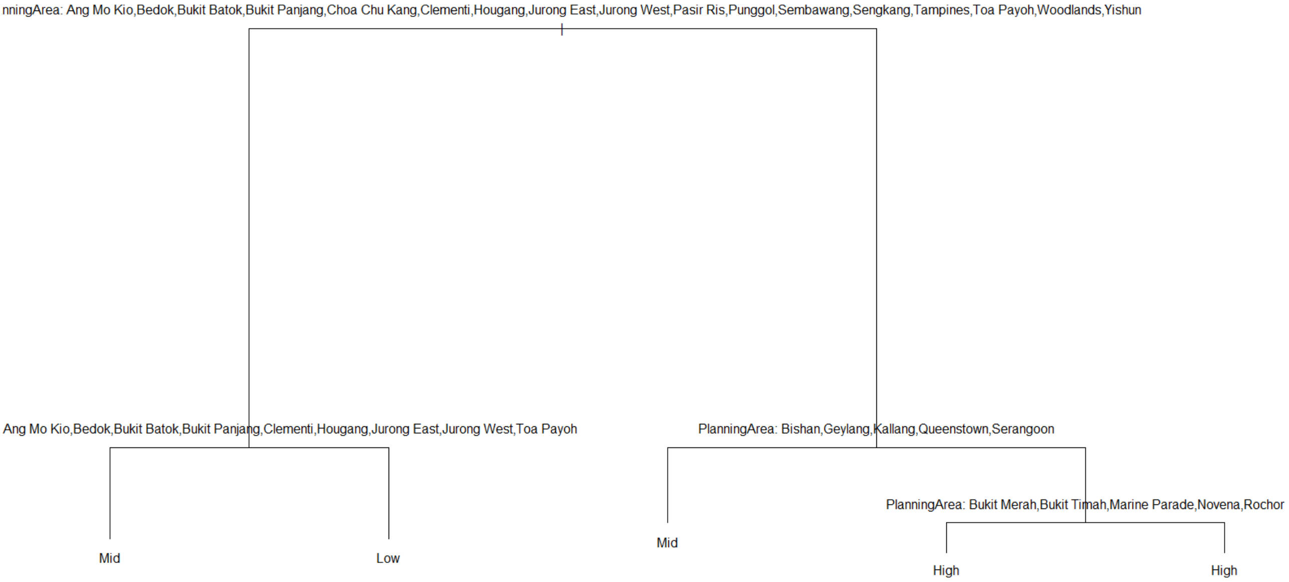
- Also, for Planning Area, there were originally 37 unique values; thus I have decided to remove the 5 areas with lowest number of transactions, namely: Outram (1), Museum (4), Orchard (8), Downtown Core (12), Mandai (20)

The classification tree method is first used on the train data set, which yields the result:

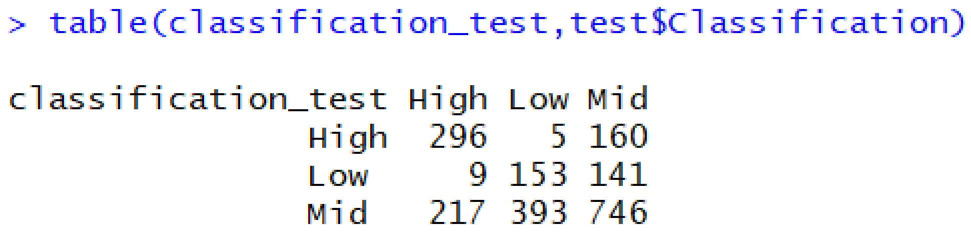


```
Classification tree:
tree(formula = Classification ~ . - UnitPrice_psf - UnitPrice_psm -
    PostalCode - area_sqm - TransactedPrice - CompletionDate,
    data = train)
Variables actually used in tree construction:
[1] "PlanningArea"
Number of terminal nodes:  5
Residual mean deviance:  1.619 = 10290 / 6355
Misclassification error rate: 0.3895 = 2477 / 6360
>
```

In the graphical form, the decision tree looks relatively simple:

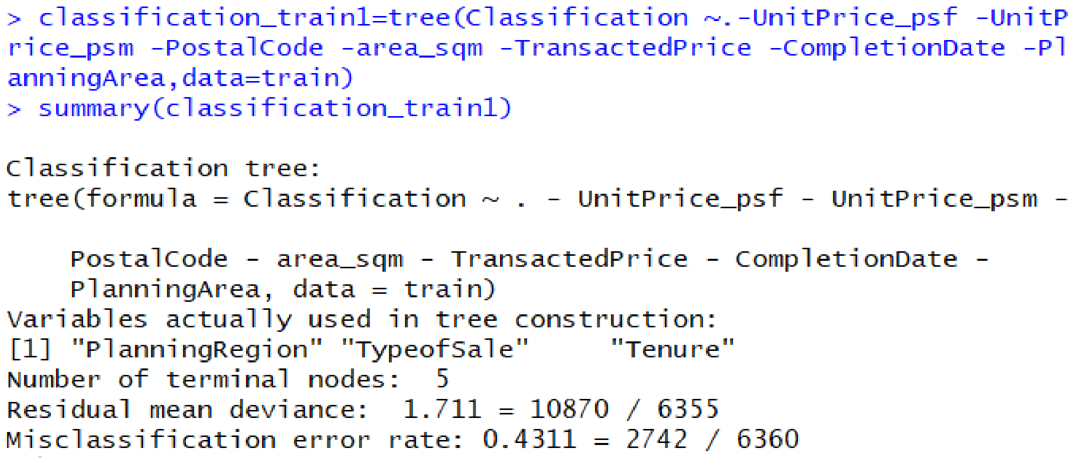

nningArea: Ang Mo Kio,Bedok,Bukit Batok,Bukit Panjang,Choa Chu Kang,Clementi,Hougang,Jurong East,Jurong West,Pasir Ris,Punggol,Sembawang,Sengkang,Tampines,Toa Payoh,Woodlands,Yishun
Ang Mo Kio,Bedok,Bukit Batok,Bukit Panjang,Clementi,Hougang,Jurong East,Jurong West,Toa Payoh
PlanningArea: Bishan,Geylang,Kallang,Queenstown,Serangoon
PlanningArea: Bukit Merah,Bukit Timah,Marine Parade,Novena,Rochor
Mid
Low
Mid
High
High

When applying the model to the train set and compare the prediction made by the model and the actual classification, I can use to calculate the correct predictions to be 56.36%, which means slightly higher misclassification error than that in the train set.



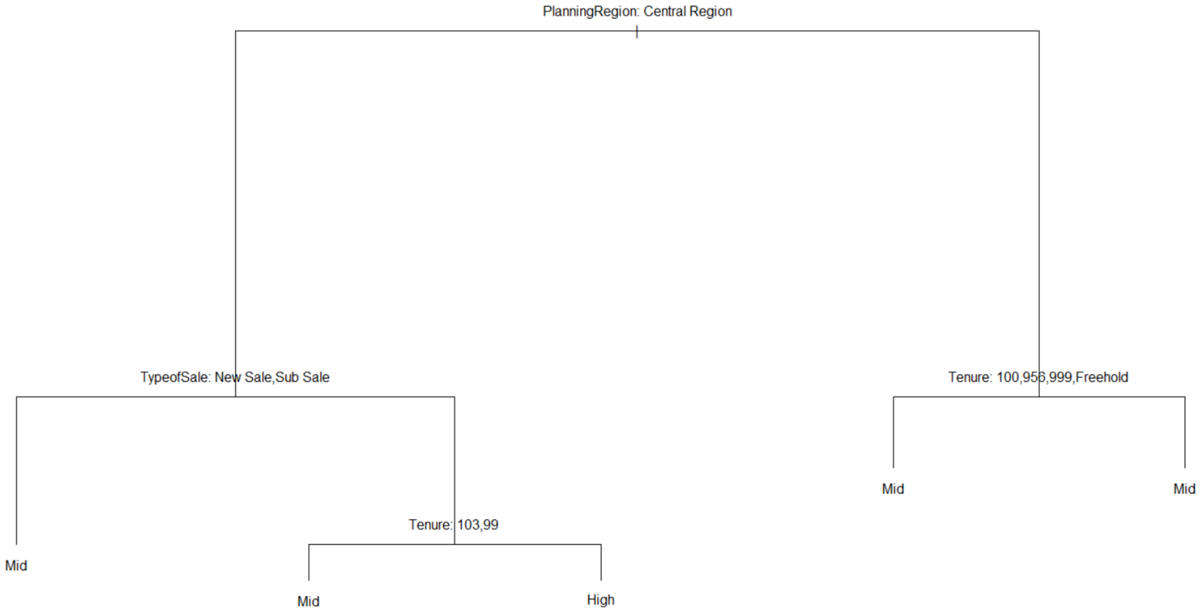
```
> table(classification_test,test$Classification)

classification_test High Low Mid
               High  296   5 160
               Low     9 153 141
               Mid   217 393 746
```

As I notice, in the first classification tree, there was only 1 variable used in the final decision tree (Planning Area). To understand more about what other factors that could significantly affect the unit price, a second model is run, this time excluding the Planning Area variable from the analysis. In this way, I can see other important factors, though less than Planning Area, that affect the housing price. The results are as shown:

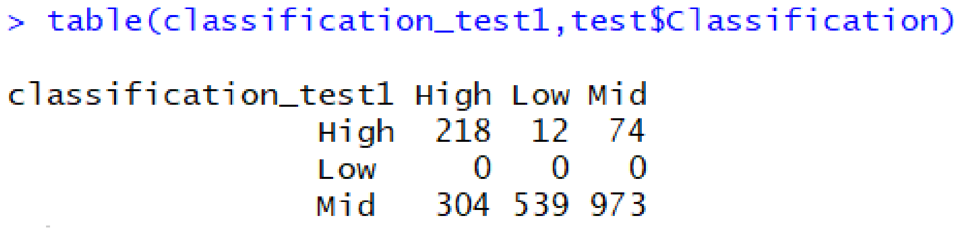


```
> classification_train1=tree(Classification ~.-UnitPrice_psf -UnitP
rice_psm -PostalCode -area_sqm -TransactedPrice -CompletionDate -Pl
anningArea,data=train)
> summary(classification_train1)

Classification tree:
tree(formula = Classification ~ . - UnitPrice_psf - UnitPrice_psm - 

    PostalCode - area_sqm - TransactedPrice - CompletionDate - 
    PlanningArea, data = train)
Variables actually used in tree construction:
[1] "PlanningRegion" "TypeofSale"     "Tenure"        
Number of terminal nodes:  5 
Residual mean deviance:  1.711 = 10870 / 6355 
Misclassification error rate: 0.4311 = 2742 / 6360 
```

Another location variable, Planning Region is still in the final decision tree. This time, I can see that Type of Sale and Tenure also have a role in determining the price of housing in various location. Another point to note is that this model fails to classify transactions that were classified as Low in the dataset.


PlanningRegion: Central Region
TypeofSale: New Sale,Sub Sale
Tenure: 100,956,999,Freehold
Mid
Mid
Tenure: 103,99
Mid
Mid
High

In terms of its accuracy, this model offers an error rate of 43.11% on the train set and 56.2% correction prediction rate on the test set, which is surprisingly only slightly lower accuracy rate than that of the first model.



```
> table(classification_test1,test$Classification)

classification_test1 High Low Mid
                High  218  12  74
                Low     0   0   0
                Mid   304 539 973
```

6. Conclusion

Results by all 4 models are highly robust. The accuracy levels and Mean Square Error values exhibit very consistent results with low standard deviation. This high level of alignment to the data is not done at the risk of overfitting, as the out-of-sample accuracy level and Mean Squared Errors are also highly consistent. The robustness also shows through the consistency of the variables that are selected in the 5 partitions for each model, as well as the common selected variables that I can see across all 4 models. For example, I can conclude that Planning areas and Areas of the unit are very likely variables that determine the transacted prices, as they appear in all 5 models' results. Even more, specifically for decision trees, there is a high correlation even in the sequence of the variables selected in different trees of different partitions, indicating the consistency in the ranking of relative importance of the variables in determining the class (for Classification tree)

A comparison may not be most appropriate in this case as one analysis is a regression and another one is a classification. Thus, an area of further investigation will be to deploy regression tree, a form of decision tree that will be more aligned in terms of comparison to the multiple regression method. Also, I can explore other external variables that are not currently in the dataset that determine the housing price, such as distance to public transport amenities (bus interchange, subway), or the reputation of the developer of the estate.

There are a few noteworthy observations from my results that allow us to conclude that this study serves as a good preliminary and exploratory step for subsequent studies in uncovering different factors that affect housing prices.