

## Introduction

This report aims to inspect the data and perform necessary data cleaning at the beginning, followed by finding the best predictor of Sale Price by using different model selection formats. Using several statistical tests to indicate the relationship between different variables. Then use the training set to evaluate different data mining methods, like regression, classification, association rule and clustering to analyze our data and draw some useful conclusions.

## Methodology and Results

There are 5891 observations and 30 variables in apartment sale. There is no missing value and we have divided data into 2 main categories. Data names total sum of school and facilities. There are 14 discrete variables and 16 continuous variables. As we have 30 columns, there are unlikely to append the exact same record for different observations. We have found 316 duplicate columns and removed all of them.

### Outlier

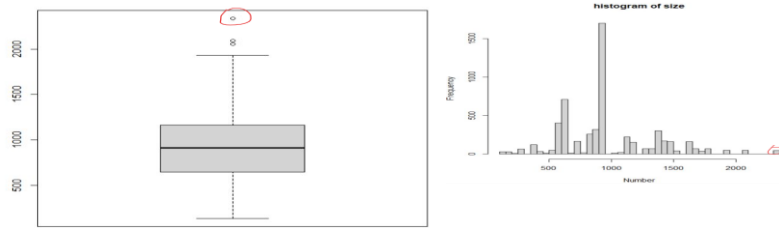
In order to prevent some extreme value from affecting our data analysis record, we have defined outlier as the variable out of range of  $\text{mean} \pm 3 \text{ standard deviation}$ . Based on the data summary, we have found that the column for size, floor and space are likely to append outlier in it.

### Outlier (Parking Space on ground)



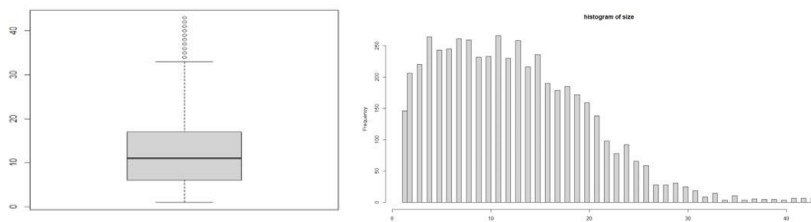
For the column for parking space on ground, we have found that there are 298 counts of observation which have a high similarity compared with other observations, like having the same apartment type, year build, subway station, N\_parking Space, which means they are from the same building. As there are lots of extremely high values in different columns in order to prevent misleading of the final results. We would like to remove them all.

## Outlier (Size)



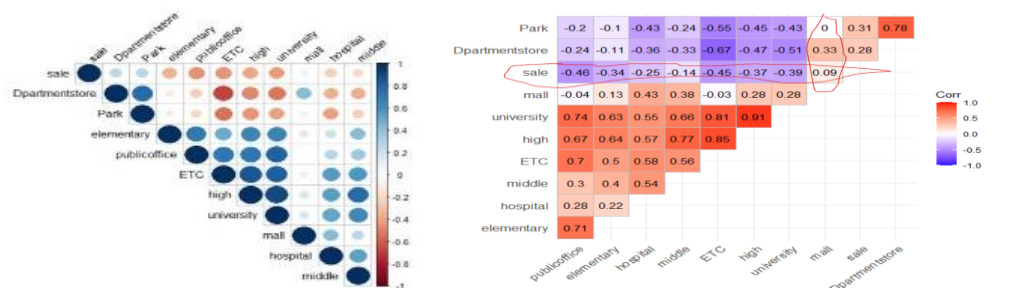
We have found 131 observations which are outliers. Because the number of observations is small, and it is normal to append those values. So, we are not going to remove these outliers.

## Outlier (Size)



We have found 60 observations are defined as outlier. Because the number of observations is small, and it is normal to append the floor above 40. So, we are not going to remove these outliers.

## Variable Selection



As we have found that most of the variable has neither negative or low correlation with the SalePrice. We chose to use N\_FacilitiesNearByTotal. and N\_SchoolNearByTotal to summarize the all the above variables.

## Model Selection for Sale Price

In order to estimate the relationship between the apartment sale price and other variables, a regression model has been selected to predict and forecast the Sale Price of different independent variables. As the dataset are sufficiently large, we separate the data into training and testing model (ratio: 7.5:2.5), using training data to select

the best multiple linear regression model for multivariate data and use testing data to test the fitness of model.

Model Diagnostics			
Model	Forward	Backward (Lowest MSE and AIC)	Stepwise

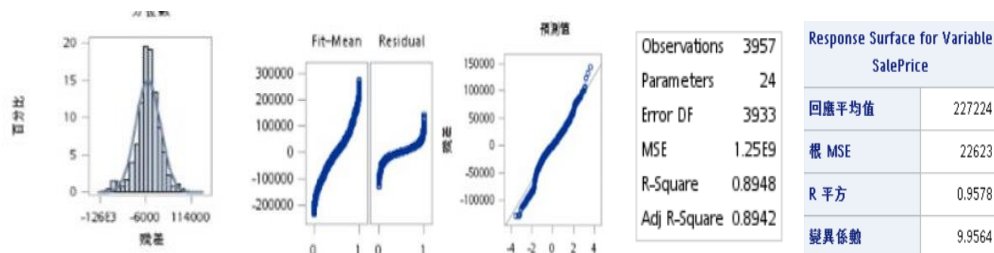
Number of estimators	26	23	24
R <sup>2</sup>	0.8948	0.8948	0.8948
Adjusted R <sup>2</sup>	0.8941	0.89418	0.8942
RMSE	35293	35285	35286
AIC	82897.88	82893.08	82894.28
F	1286.08	1454.44	1393.8
Pr>F	<0.001	<0.001	<0.001

By comparing the models in the model diagnostics table, Minimized RMSE model is the most suitable model for predicting the value of quality. It shows the best performance on Adjusted R<sup>2</sup>, the smallest difference between the actual value predicted by a model (Smallest RMSE) and P-value (<0.0001) is relatively low. Hence, we selected this model which shown below

SalePrice=

$\beta_0 + \beta_1 \text{YrBuild} + \beta_2 \text{YrSold} + \beta_3 \text{MonthSold} + \beta_4 \text{Size} + \beta_5 \text{Floor} + \beta_6 \text{N\_parking\_Ground} + \beta_7 \text{N\_parking\_basement} + \beta_8 \text{TimeToBusStop} + \beta_9 \text{corridor} + \beta_{10} \text{AptType} + \beta_{11} \text{Bangoge} + \beta_{12} \text{Chisungmarket} + \beta_{13} \text{Daegu} + \beta_{14} \text{Kyungbukunihospital} + \beta_{15} \text{Myungduk} + \beta_{16} \text{nosubway nearby} + \beta_{17} \text{subwayin05mins} + \beta_{18} \text{subwayin510mins} + \beta_{19} \text{SubwayIn1015mins} + \beta_{20} \text{N\_APT} + \beta_{21} \text{N\_manager} + \beta_{22} \text{N\_facNearBy\_Total} + \beta_{23} \text{N\_schoolNearBy\_Total}$

Since backward selection model has the lowest MSE and AIC, which means if we include more variable to estimate the SalePrice, model will become more complicated, and the accuracy of the model will become lower. Since number of elevators have 0.2082483 correlation and are not selected in the backward selection method, it seems to have a very weak relationship with SalePrice.



Since the residual for backward selection model seems to be identical distributed independently, which means there are no bias in estimating the value for Sale Price, thus the model has a great ability to estimate the value for Sale Price and does not need to include any non-linear terms.

## Regression model

Parameter estimator				
Variable Name	Variable	Parameter estimation	T value	Pr>t
-	Intercept	-27627731	-53.66	<0.0001
YearBuilt	X1	1265.01608	6.77	<0.0001
YrSold	X2	12482	55.79	<0.0001
MonthSold	X3	1503.66774	8.78	<0.0001
Size	X4	137.38868	66.63	<0.0001
Floor	X5	1215.49620	15.52	<0.0001

N_parking_Ground	X6	-287.25900	-29.01	<0.0001
N_parking_basement	X7	-680.72887	-21.58	<0.0001
TimeToBusStop	X8	-39763	-25.14	<0.0001
Corridor	X9	-458729	-27.04	<0.0001
AptType	X10	-202510	-19.73	<0.0001
Bangoge	X11	-359103	-24.20	<0.0001
Chisungmarket	X12	90864	9.79	<0.0001
Daegu	X13	-289490	-22.44	<0.0001
Kyungbukunihospital	X14	-28395	-9.27	<0.0001
Myungduk	X15	-54317	-10.84	<0.0001
Nosubwaynearby	X16	235619	21.17	<0.0001
Subwayin05mins	X17	367896	25.8	<0.0001
Subwayin510mins	X18	466772	24.13	<0.0001
Subwayin1015mins	X19	397583	27.51	<0.0001
N_Apt	X20	33641	22.41	<0.0001
N_manager	X21	34073	22.84	<0.0001
N_facNearBy_Total	X22	29854	23.31	<0.0001
N_SchoolNearBy_Total	X23	-5468.37259	-9.9	<0.0001

Since we have found that our model has a very good performance in estimating the value of saleprice and lots of variables have a very high T-value, which means the parameter estimation is effective for studying the relationship between the SalePrice and other variables. For example, since size has a very high t-value (66.63), which means that when size increase by 1, the SalePrice will likely increase by 137.3, so size and salePrice have a positive relationship.

## Data Analysis

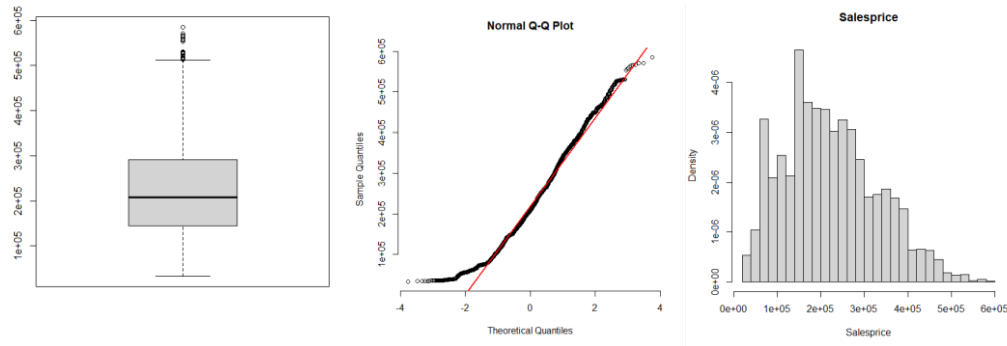
```
Two Sample t-test

data: top$SalePrice and bottom$SalePrice
t = 107.06, df = 5914, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 169273 175588
sample estimates:
mean of x mean of y
 307319.1 134888.6

Two Sample t-test

data: top$SalePrice and bottom$SalePrice
t = 107.06, df = 5914, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 168280.4 176580.6
sample estimates:
mean of x mean of y
 307319.1 134888.6
```

We compare the sales price with the top 50% of data and the bottom 50% of data. The result shows that both p-value is 2.2e-16, which is smaller than 0.05. Therefore, we can reject the null hypothesis at 95% and 99% confidence interval. There is a difference between the means of both samples.



```
> sd(top$SalePrice)
[1] 73701.54
> sd(bottom$SalePrice)
[1] 47324.19
> |
```

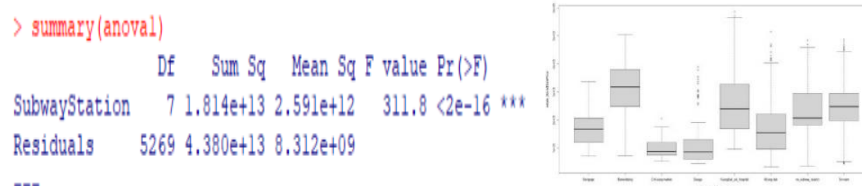
These pictures show the box plot of the Sales Price and the standard deviation of the top 50% sales price and the bottom 50% sales price. The standard deviation of top sales price is greater than bottom sales price. It is more dispersed of a third quantile compared with the first quantile. In QQ-plot of the sale price, the right slanting line occurred in the first quantile which means that there is a few or no extreme value in the first quantile. The histogram of the sales prices also proves that the third quantile has no extreme value.

### Comparing the SalePrice with different Hallway Type



From the regression model, we can find that using the HallwayType to guess the SalePrice, the  $R^2$  is 0.4993 and the F value is super big, which means it has a very close relationship between the HallwayType and the SalePrice. And from the box plot, we can find that terraced and mixed is higher than the value for corridor. Thus, we can conclude that corridor and mixed will usually have a higher of SalePrice

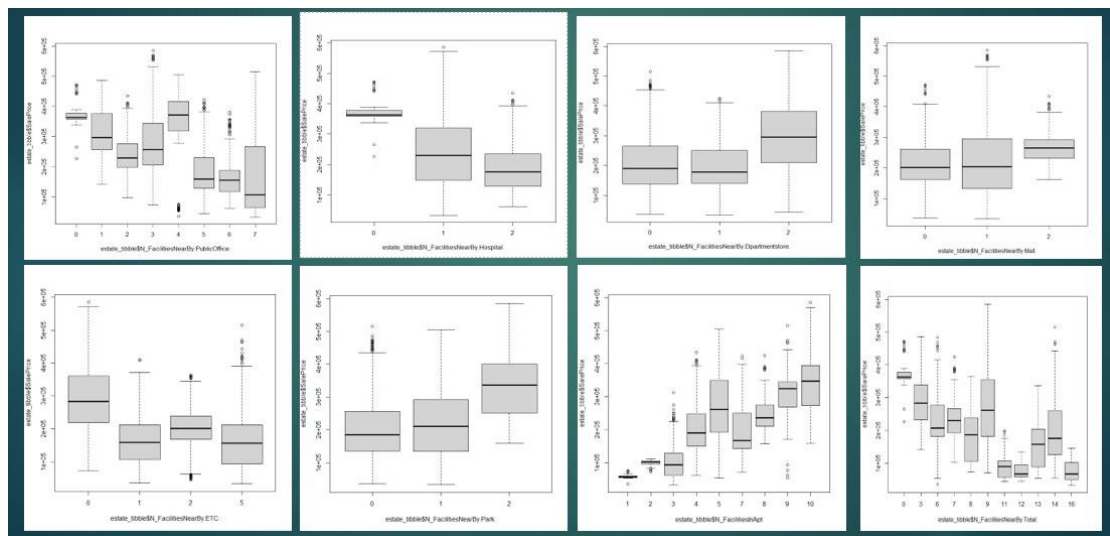
### Comparing the SalePrice with different Subway Station



參數估計值					
變數	DF	參數估計值	標準誤差	t 值	Pr >  t
Intercept	8	246182	3578.81430	68.79	<.0001
Bangoge	8	-65165	5087.00507	-12.81	<.0001
Banwoldang	8	63193	4923.58201	12.83	<.0001
Chilsungmarket	8	-147745	9224.37443	-16.02	<.0001
Daegu	8	-114485	10517	-10.89	<.0001
Kyungbukunihospital	8	36520	4459.27587	8.19	<.0001
Myungduk	8	-78511	4282.33667	-18.33	<.0001
nosubwaynearby	8	-13681	5860.24384	-2.33	0.0196

As the P-value less than 0.01, which means there are different mean of SalePrice among different Subway station, comparing the result with our regression model, we can find that Banwoldang, Kyungbukunihospital will usually have a higher sale price, but Daegu, Chilsungmarket will usually have a lower sale price.

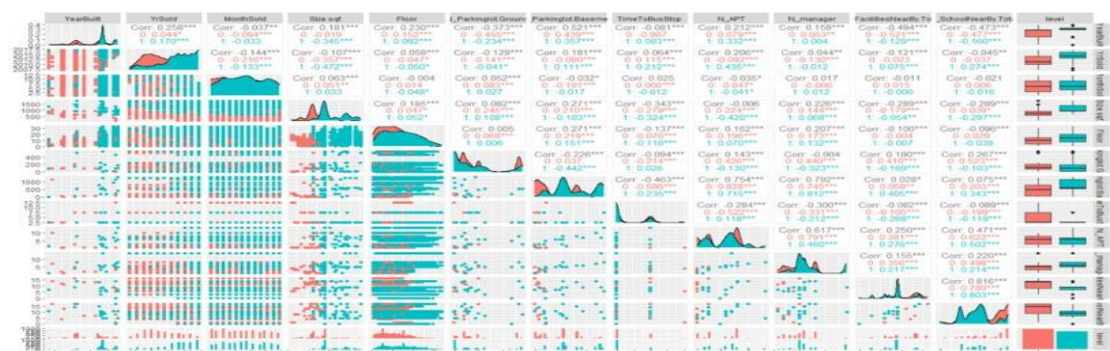
Comparing the result with different Facility



The figures above are the number of public offices, hospitals, department stores, malls, ETC, parks, in house facility which is near the house. Besides in house facility, most of them have a weak relationship between them with SalePrice. But we can also find that the increased number in house facility, the higher sale price it is.

Summary

We separate dataset into two parts. Group 1 is for top 50% of sale price and the remaining part are grouped for group2.



By considering the distribution for two different group, correlation and the result in regression, we can conclude that YearBuilt, YearSold and size has a very clear and positive relationship between SalePrice, which means when the value is better, the



sale price will be higher. For the HallwayType, terraced are likely to have a high sale price. The total number of schools has a negative relationship for SalePrice. For floor, time to bus stop, number of elevator/APT/Manager, they have a very weak relationship with SalePrice

## Classification

In order to study the relationship between SalePrice and other variables, we have classified saleprice into 3-part, quartile 0-30% are considered as low, quartile 30%-70% are considered as median and the remaining part are considered as high. We would like to conduct some classification method, using variables to predict which class observations belong to and choose the best model. So, we would split the data into training and testing model (Ratio 7:3) to test the fitness of model and prevent overfitting.

## SVM

```
> svmFit
Support Vector Machines with Linear Kernel

3167 samples
16 predictor
3 classes: 'High', 'Low', 'Median'

No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 3167, 3167, 3167, 3167, 3167, 3167, ...
Resampling results:

Accuracy Kappa
0.9010975 0.8497868

Tuning parameter 'C' was held constant at a value of 1
```

	High	Low	Median
High	567	0	64
Low	0	581	53
Median	47	36	762

```
> kappa(confusion_table)
Estimate Std.Err 2.5% 97.5% P-value
kappa    0.856 0.009729 0.8369 0.8751 0
```

SVM is a classification method which tries to use some plane to separate data into different classes in a high dimensional place. There are 2 indicators which are accuracy and Cohens' kappa. Accuracy means the correct rate for estimating the right class for our observations and Cohens' kappa is the Metrix that compares the observations accuracy and expected accuracy. As you can see, both are decent, no matter the training and testing set, which means there is a true relationship between the variable and the class, and this method is efficient to predict the true class.

## Knn method

```
k-Nearest Neighbors

3167 samples
16 predictor
3 classes: 'High', 'Low', 'Median'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 2850, 2851, 2850, 2850, 2851, 2851, ...
Resampling results across tuning parameters:

k Accuracy Kappa
5 0.8950737 0.8406947
7 0.8944418 0.8396911
9 0.8939140 0.8380225
11 0.8888630 0.8310831
13 0.8878148 0.8294107
15 0.8862325 0.8269578
17 0.8828410 0.8216662
19 0.8786479 0.8150766
21 0.8761206 0.8111098
23 0.8753875 0.8099487
```

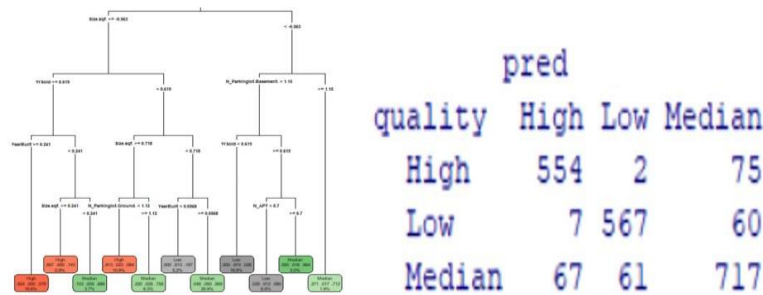
	High	Low	Median
High	551	1	79
Low	2	583	49
Median	53	57	735

```
Estimate Std.Err 2.5% 97.5% P-value
kappa    0.8267 0.01056 0.806 0.8474 0
```

Knn method is based on computing the distances between the tested and the training data, using a distance function to identify their nearest neighbors. The value of k is the measure of the distance and we have found that it has the best performance when k=5. Same as SVM, accuracy, Cohens' kappa and it does not have an obvious bias in

estimating the class in the confusion matrix. Therefore, Knn is effective to predict the true class.

## Decision tree



```
> kappa(confusion_table)
      Estimate Std.Err   2.5%   97.5% P-value
kappa   0.8046 0.01112 0.7828 0.8263      0
```

Decision tree is a regression decision tree builds regression and classification models in the form of a tree structure. Because full model trees are too complicated, we will only show a part of our tree model. As you can see, size, yearsold, n\_parking space on ground appears on the top of the tree, which means they have the lowest MSE to determine the nodes of tree. So, tree models identify these 3 variables are the most efficient variable to determine the sale price class, thus they have a very high importance. But tree models are not a good model for estimating the right class of tree due to overfitting. So, from our testing data, we can see that the performance index, like kappa, accuracy and the confusion matrix have a slightly worse performance compared with Knn and SVM method.

## Random forest method

```
> rf
Call:
randomForest(formula = quality ~ ., data = trainSet)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 4
OOB estimate of error rate: 8.58%

      pred
quality High Low Median
High    608  0   23
Low      0 607  27
Median  10 19 816

> kappa(confusion_table)
      Estimate Std.Err   2.5%   97.5% P-value
kappa   0.9446 0.006207 0.9324 0.9568      0
```

Random decision forest constructs a multitude of decision trees at training data and select the best performance for decision tree. As we can find that the performance index is quite good, thus, it is also a good method to determine the class

## Model diagnostic

Model	SVM	Knn	Decision tree	Random forest
Kappa	0.856	0.8257	0.8046	0.9446
Error	200	244	272	78
Precision for	0.9234528	0.9092409	0.8821656	0.9838188



High				
Precision for Median	0.8668942	0.8516802	0.8415493	0.9422633
Precision for Low	0.9416532	0.9095164	0.9	0.9696486
Recall for High	0.8985737	0.8732171	0.8779715	0.9635499
Recall for Median	0.9164038	0.8698225	0.09588125	0.9656805
Recall for Low	0.9017751	0.9195584	0.8943218	0.9574132

As we can find that for Random Forest method, it has the highest kappa, precision (Ratio for finding the actual class) and recall (Ratio for predicting the class correct), which means random forest has no bias in estimating the class for model, so we will choose random forest method for our classification model.

### Importance table

	Mean decreasing Gini
Size.sqf	307.4064
YearSold	246.6740
HallwayType	97.8132
YearBuilt	84.7765
N_Parkinglot_Basement	80.7527
Floor	49.8059
TimeToSubway	45.6806
N_Parkinglot.Ground.	42.5325
N_APT	35.3805
N_manager	33.9366
SubwayStation	33.5723
N_SchoolNearBy_Total	32.2521
N_FacilitiesNearBy.Total.	25.4254
TimeToBusStop	6.4268
AptManagerType	2.4271

This table indicates the importance of different variables for selecting the right class through random forest method. The higher the Mean decreasing Gini is, the more accuracy for determining the type of class. From this table, we have found that size/YearSold/HallwayType/YearBuild and N\_parking space on basement have the highest importance for determining the class. Thus, if we want to increase the SalePrice, we should prioritize these factors.

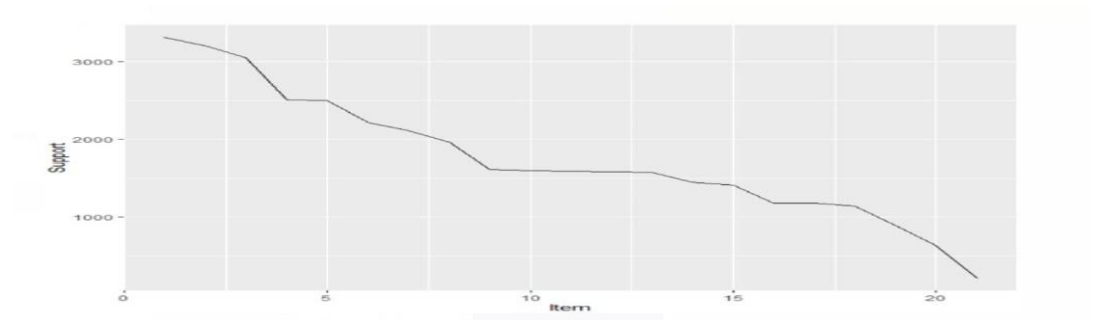
## Association

### Data selection

As mentioned above, several useless factors would be eliminated. The remaining factors are salesprice, hallway type, yearbuilt, yearsold, total facilities nearby (N\_FacilitiesNearBy.Total.), total school nearby (N\_SchoolNearBy\_Total), size(sqf). Since the association method requires categorial data. The numerical data (saleprice, total facilities nearby, total school nearby, size(sqf) )would be converted into categorical data by separating into three groups, low ( $<30\%$ ), median ( $30\% \leq x < 70\%$ ),

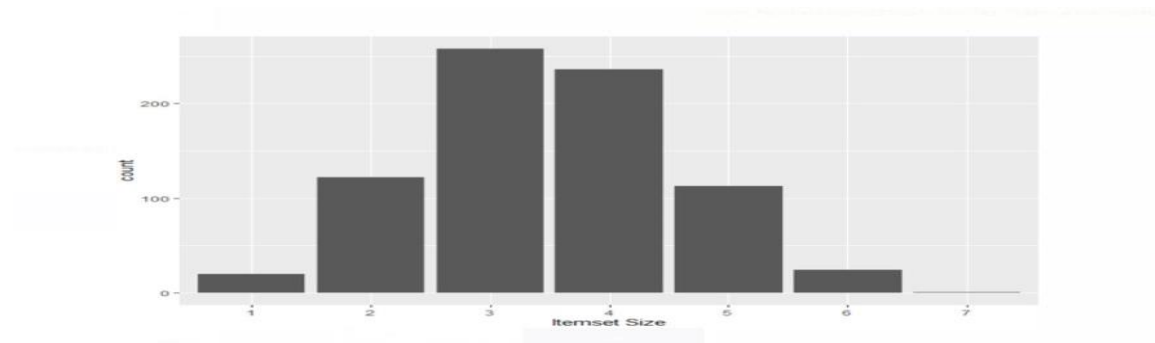
high ( $\geq 70\%$ ). Yearbuilt would be counted in 10 years as one group (e.g. 1987-1996) while yearsold would be counted in 5 years as one group (e.g. 2006-2011).

### Inspect Transaction (Item frequency)



In the association analysis, support = 0.05 would be set. With the low standard of support, items are included in a wide range.

### Frequent Itemsets (geom\_bar)



In the association analysis, support = 0.05 and confidence = 0.9 would be set. The frequent itemsets graph reveals the itemset size of our dataset is 1 to 7. The itemset size is mainly located in 3 or 4, which both exceed 200 counts.

### Scatterplot of Association rule visualization



By the scatter plot of association rule, there are total 684 rules, which conform the standard of support  $> 0.05$  and confidence  $> 0.9$  in this research.

### Association result (confidence = 1)

LHS	RHS	support	confidence	coverage	lift	count	phi	gini
All	All	All	All	All	All	All	All	All
[2] {TotalN_school=High}	{TotalN_Facilities=High}	0.224	1.000	0.224	4.472	1,180.000	1.000	0.347
[3] {TotalN_Facilities=High}	{TotalN_school=High}	0.224	1.000	0.224	4.472	1,180.000	1.000	0.347
[25] {YearBuilt10years=1997-2006,HallwayType=mixed}	{TotalN_Facilities=Median}	0.142	1.000	0.142	5.903	751.000	0.902	0.229
[31] {YearBuilt10years=1987-1996,TotalN_school=High}	{TotalN_Facilities=High}	0.185	1.000	0.185	4.472	977.000	0.888	0.274
[32] {YearBuilt10years=1987-1996,TotalN_Facilities=High}	{TotalN_school=High}	0.185	1.000	0.185	4.472	977.000	0.888	0.274
[160] {YearBuilt10years=1997-2006,HallwayType=mixed,TotalN_school=Median}	{TotalN_Facilities=Median}	0.121	1.000	0.121	5.903	636.000	0.820	0.189

For the confidence = 1, which means when left-hand side appears, the right-hand side appears also, 3 results could be observed. When the total number of schools exceed 70%, the total number of facilities would exceed 70%. When the year built fell into 1987-1996, the total number of schools would exceed 70%. When year built fell into 1987-1996, the total number of facilities would exceed 70%.

### Association result (RHS = low salesprice )

LHS	RHS	support	confidence	coverage	lift	count	phi	gini
All	{SalePricequality=Low}	All	All	All	All	All	All	All
[41] {YearBuilt10years=1987-1996,Size.sqf.quality=Low}	{SalePricequality=Low}	0.116	0.900	0.129	2.997	614.000	0.504	0.107
[85] {YrSold5years=(2006,2011],Size.sqf.quality=Low}	{SalePricequality=Low}	0.115	0.990	0.116	3.297	608.000	0.546	0.125
[1] {HallwayType=corridor}	{SalePricequality=Low}	0.112	0.932	0.120	3.104	592.000	0.510	0.109
[21] {Size.sqf.quality=Low,HallwayType=corridor}	{SalePricequality=Low}	0.104	1.000	0.104	3.329	547.000	0.519	0.113
[319] {YrSold5years=(2006,2011],Size.sqf.quality=Low,HallwayType=mixed}	{SalePricequality=Low}	0.074	0.987	0.074	3.287	388.000	0.425	0.076
[144] {YrSold5years=(2011,2017],Size.sqf.quality=Low,HallwayType=corridor}	{SalePricequality=Low}	0.064	1.000	0.064	3.329	336.000	0.398	0.067
[208] {YearBuilt10years=1987-1996,YrSold5years=(2006,2011],Size.sqf.quality=Low}	{SalePricequality=Low}	0.063	1.000	0.063	3.329	333.000	0.396	0.066
[132] {YearBuilt10years=1987-1996,Size.sqf.quality=Low,HallwayType=corridor}	{SalePricequality=Low}	0.058	1.000	0.058	3.329	304.000	0.377	0.060

During this research, we tried to explore the factors that are associated with sales price. From the above association analysis, low size(sqf), year built in 1987-1996, hallway type is corridor, total number of schools  $\geq 70\%$ , total number of facilities  $\geq 70\%$  are the 5 factors which are associated with low salesprice. Therefore, whenever one of the above 5 reasons appears, the probability of low sales price would increase.

### Association result (RHS = High salesprice )

LHS	RHS	support	confidence	coverage	lift	count	phi	gini
All	{SalePricequality=High}	All	All	All	All	All	All	All
[254] {Size.sqf.quality=High,HallwayType=terraced,TotalN_school=Median}	{SalePricequality=High}	0.065	0.917	0.070	3.065	341.000	0.372	0.058
[494] {Size.sqf.quality=High,HallwayType=terraced,TotalN_Facilities=Low,TotalN_school=Median}	{SalePricequality=High}	0.065	0.917	0.070	3.065	341.000	0.372	0.058
[252] {YearBuilt10years=2007-2015,Size.sqf.quality=High,TotalN_school=Median}	{SalePricequality=High}	0.064	0.941	0.068	3.148	337.000	0.378	0.060
[491] {YearBuilt10years=2007-2015,Size.sqf.quality=High,TotalN_Facilities=Low,TotalN_school=Median}	{SalePricequality=High}	0.064	0.941	0.068	3.148	337.000	0.378	0.060
[488] {YearBuilt10years=2007-2015,Size.sqf.quality=High,HallwayType=terraced,TotalN_school=Median}	{SalePricequality=High}	0.064	0.941	0.068	3.148	337.000	0.378	0.060
[639] {YearBuilt10years=2007-2015,Size.sqf.quality=High,HallwayType=terraced,TotalN_Facilities=Low,TotalN_school=Median}	{SalePricequality=High}	0.064	0.941	0.068	3.148	337.000	0.378	0.060

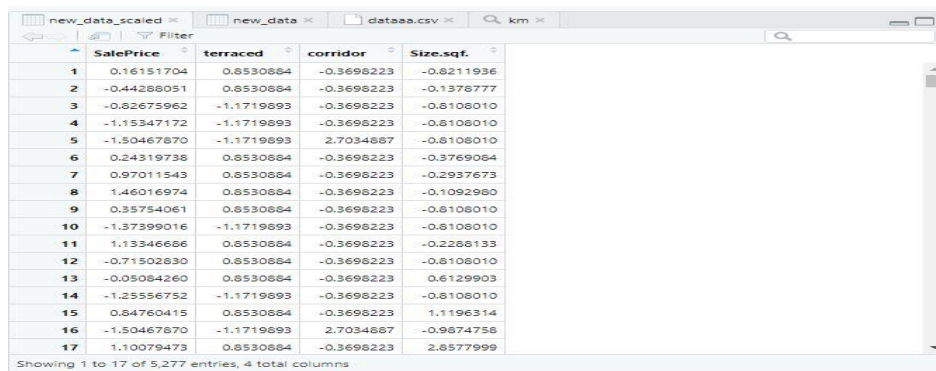
Besides the association with low salesprice, factors that associated with high salesprice are also conducted. From the above association analysis, high size(sqf), year built in 2007-2015, hallway type is terraced, total number of schools nearby is median are the 4 factors which associated with high sales price. As a result, whenever one of the above 4 reasons appears, the probability of high sales price would increase.

## Clustering

### Data preparation

For the data preparation, we filter out the most important factors for clustering through the importance table mentioned above. Hence, we choose the factors of Saleprice, HallwayType and size(sqf). For normalization, we turn the HallywayType data into two columns which are terraced and corridor. Thus, we perform the scale

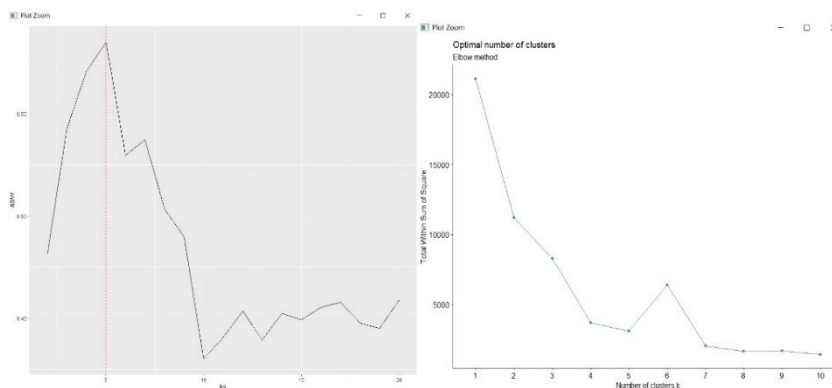
function.



	SalePrice	terraced	corridor	Size.sqf.
1	0.16151704	0.8530884	-0.3698223	-0.8211936
2	-0.44288051	0.8530884	-0.3698223	-0.1376777
3	-0.62675962	-1.1719893	+0.3698223	-0.6106010
4	-1.15347172	-1.1719893	-0.3698223	-0.6106010
5	-1.50467670	-1.1719893	2.7034687	-0.6106010
6	0.24319738	0.8530884	-0.3698223	-0.3769064
7	0.97011543	0.8530884	+0.3698223	-0.2937673
8	1.46016974	0.8530884	-0.3698223	-0.1092980
9	0.35754061	0.8530884	-0.3698223	-0.6106010
10	-1.37399016	-1.1719893	-0.3698223	-0.6106010
11	1.13346686	0.8530884	+0.3698223	-0.2286133
12	-0.71502830	0.8530884	-0.3698223	-0.6106010
13	-0.05084260	0.8530884	-0.3698223	0.6129903
14	-1.25556752	-1.1719893	-0.3698223	-0.6106010
15	0.84760415	0.8530884	+0.3698223	1.1196314
16	-1.50467670	-1.1719893	2.7034687	-0.9874758
17	1.10079473	0.8530884	-0.3698223	2.6577999

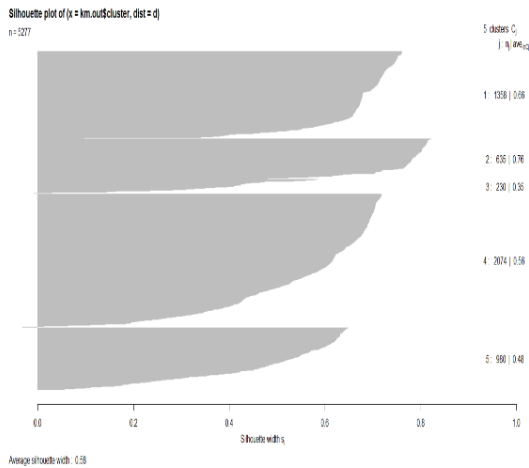
## ASW plot & Elbow method

In order to define how many clusters that we need to divide in, we perform asw plot to select the best k. We can find the best value of k by maximizing ASW and we set the KS range to be 2 to 20. Therefore, in this plot we can see that the best k is equal to 5 which indicates the red line. For the Elbow method, we can see the at the value of five, it is a significant turning point which looks like an elbow. Hence, we also chose 5 as the best k.



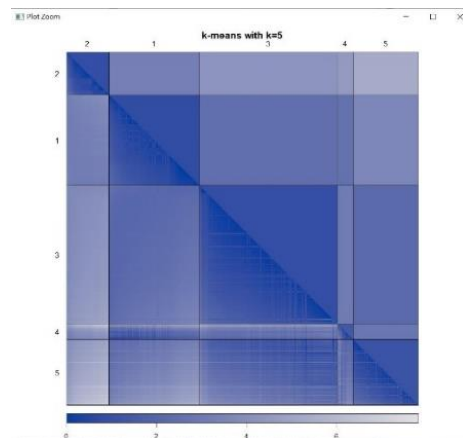
## Silhouette plot

We use silhouette plot to evaluate the quality of cluster. We can see the average silhouette width is equal to 0.58 and the mean of the clusters are 0.66,0.76,0.36,0.56 and 0.48 respectively. These numbers are close to 1 but not -1 which indicates the cohesion and separation of all clusters are decent with only a little bit outlier smaller than 0 on the left-hand side which are acceptable.



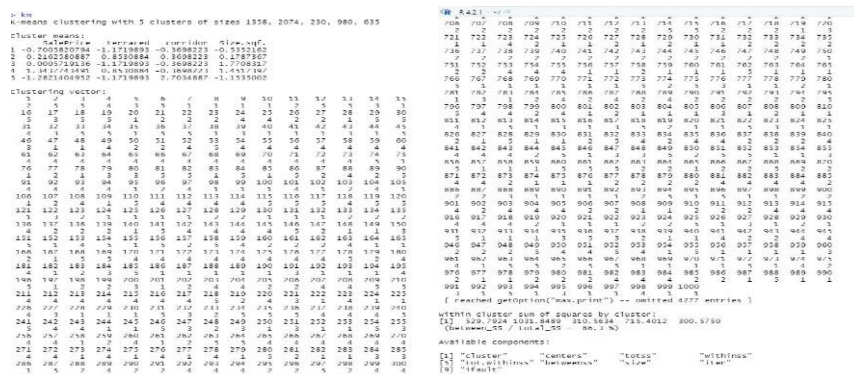
## Dissimilarity plot

In the dissimilarity plot for Euclidean Distance, it shows the average dissimilarity in the lower part of the diagonal. We can see a very dark shadow that appeared near the lower part of the diagonal which indicate obvious clusters are formed by k=5 as they are not similar.



## K-mean with cluster statistic

After setting k=5, the total sum of squares by cluster is 86.3% which means it maintains a very high clustering accuracy. And the cluster mean of each variable are shown in the graph. In the cluster statistic, we can notice the different information like vector of cluster minimum distances of a point in the cluster to a point of another cluster is 3.07331100, 0.01633607, 0.04084017 and 0.01633607. Maximum cluster diameter is 4.967707 and so on.



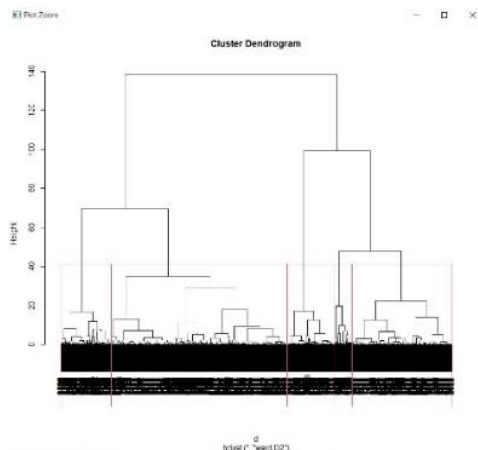
```

> cluster.stats(c, kmccluster)
$
[1] 3277
$cluster.number
[1] 5
$cluster.size
[1] 1358 2074 230 980 835
$min_cluster_size
[1] 230
$noisen
[1] 0
$diameter
[1] 2.4204188 1.107227 1.136988 1.147283 4.015819
$average.distance
[1] 0.7096642 0.8511484 1.4113510 1.0513541 0.8005638
$median.distance
[1] 0.7480188 0.7807438 1.3374410 0.9729375 0.7019770
$separation
[1] 0.04084017 0.03267213 0.04084017 0.03267213 0.07831100
$average.toother
[1] 2.860789 2.727000 3.044647 3.094674 4.187148
$separation.matrix
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.00000000 2.0250777 0.04084017 2.12785905 1.073311
[2,] 2.02507771 0.00000000 2.02518437 0.03267213 3.680514
[3,] 0.04084017 2.02518437 0.00000000 2.02507771 1.073979
[4,] 2.12785905 0.03267213 2.02507771 0.00000000 1.680704
[5,] 1.07331100 3.68051358 1.07397959 1.68070357 0.000000
$ave.between.matrix
      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.000000 2.418409 2.529435 3.585715 3.313434
[2,] 2.418409 0.000000 3.038877 2.111128 4.195520
[3,] 2.529435 3.038877 0.000000 2.795458 4.158434
[4,] 3.585715 2.111128 2.795458 0.000000 5.293680
[5,] 3.313434 4.195520 4.158434 5.293680 0.000000
$average.between
[1] 2.068791
$average.within
[1] 0.888365
[1] 10142262
$within
[1] 2778444
$max.diameter
[1] 4.015819
$min.separation
[1] 0.03267213
$within.cluster.ss
[1] 2888.181
$twu.avg.allwidth
      1      2      3      4      5
0.6627000 0.5390375 0.3495278 0.4750213 0.7354736
$avg.stlwidth
[1] 0.3849852
$g2
NULL
$g3
NULL
$spread.unjamme
[1] 0.7320261
$dummy
[1] 0.008155818
$dummy2
[1] 1.511405
$entropy
[1] 1.450358
$br.ratio
[1] 0.2894679
$ch
[1] 8932.034
$wfdgap
[1] 0.7015030 0.5450788 1.4063397 0.4800395 1.0835556
$wfdstgap
[1] 1.44854
$index
[1] 0.2858886

```

## Dendrogram

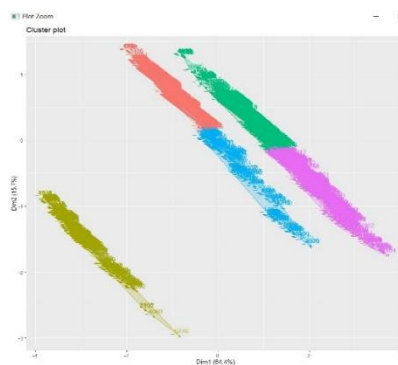
The dendrogram is formed through the complete method and it shows how many clustering the dataset can formed. The red line divides the dendrogram into 5 clusters.



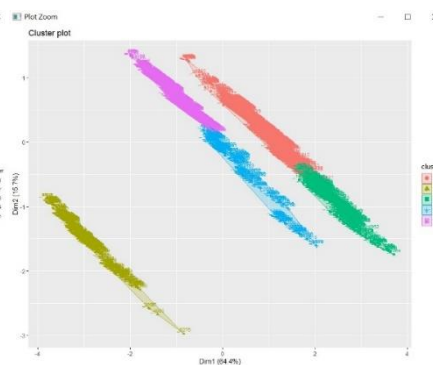
## K-mean cluster plot vs hierarchical cluster plot

After printing out the cluster plots by both k-mean and hierarchical method, we can see that the clustering effect are quite good with cohesion and separation. However, for the hierarchical cluster plot, the red area and green area results in an overlapping. Therefore, using k-mean clustering will be more reliable and preferred in this case.

### K-mean



### Hierarchical



## Anova table

For the Anova table, Saleprice has the highest significant codes which is three stars, it



means that saleprice is one of the most important factors for choosing the apartments. For the p-value, it is smaller than 0.05 so we can reject the null hypothesis, we have sufficient evidence to say that there is a significant difference between the means.

```
> summary(anova1)
              Df      Sum Sq   Mean Sq F value    Pr(>F)
cluster        1 3.374e+13 3.374e+13    6313 <2e-16 ***
Residuals    5275 2.820e+13 5.345e+09
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

## GG pairs plot & correlation

The gg pairs plot indicates how each cluster can be characterized. For instance, size has the highest correlation with Saleprice in cluster 2 but the lowest correlation with saleprice in cluster 4. For the box plot, we can notice the quartiles distribution in the 5 clusters. For the Saleprice, data has the highest sale price with the largest interquartile range and maximum and minimum in fifth cluster. For the third cluster, it is antithetical to the fifth cluster. For correlation table, we can see that cluster 1,3 consist of both terraced and corridor, cluster2 and 4 belong to terraced and cluster5 belongs to corridor.



## Conclusion

### 1. Positive impacts

By regression, correlation and distribution plot, we have found that Year build, Year sold and size, have a positive impact on the sale price, which means the value of these variables are bigger, the sale price is bigger. From our classification result, we have found that size, year build, year old and hallway type has the biggest impact on the sale price. Therefore, these variables are the priority to consider if we want to increase the sale price.

### 2. Others relationship

From association rules, we have found that when the size is larger, year build is latest. Secondly, we discovered that hallway type appears to be terraced. Moreover, if the number of schools near apartment equals to median, the sale price usually becomes higher. We have successfully put data into 10 clusters and have a difference comparing with different clusters.