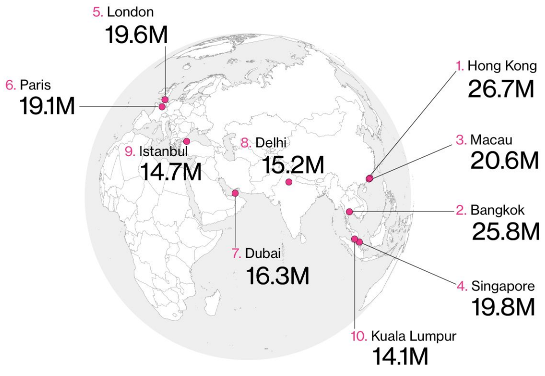
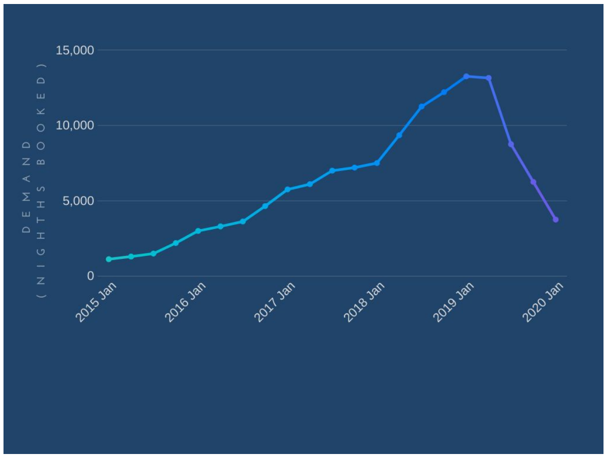
Introduction

With more than 800 million guest arrivals and 5.6 millions of active listings globally, Airbnb provides unique travelling experiences by offering millions of spaces to stay powered by local hosts worldwide(Airbnb, 2020). Take Hong Kong as an example, there are 6,760 unique active listings on Airbnb Hong Kong and around 85% of the listings mainly located in 6 out of 18 districts in Hong Kong, while all of them are located either in Tsim Sha Tsui region or Hong Kong Island region (Airbitcs, 2020). According to statistics from Bloomberg (2019), Hong Kong remains the first place for international arrivals in 2019 with the amount of around 26.7 millions, despite months of social movement (see picture 1). Besides, the demand for nights booked in Airbnb before social movement and COVID-19 kept soaring steadily (see picture 2). It indicates the positive prospect of Airbnb accommodation in Hong Kong, along with rising demand and supply.



Sources: Euromonitor International, Bloomberg



Picture 1: Bloomberg

Picture 2: Airbitcs

However, varying characteristics of spaces drive unique prices of the accommodations, which may weaken its edges. There may be a mismatch between price and actual situation of the accommodation. For travellers, if they choose a staycation that is more expensive than it should be, they fail to make the best choices for their trips. Simultaneously, for hosts, when they set the price too high, their rooms may be less attractive to travellers. On the other hand, when they set the price too low, they cannot get appropriate profits. Therefore, this report aims at helping predict the price of every distinct accommodation by applying in total 4 methods including linear regression, clustering, regression tree and random forest, followed by concluding the limitations of those models and introducing possible future works for improvement.

Overall goals

The goal of this project is to generate predictions and results to help both travellers and the airbnb hosts. For travellers, the objective is to help them to make wiser choices on accommodation with appropriate and reasonable prices, according to their preferences. This project also aims at facilitating the hosts to set a more reasonable price according to the situation, neighbourhood and facilities of their airbnb to enhance competitiveness of their airbnb and gain reasonable rental return.

Data preparation

The data used in this project is the historical data of airbnb in Hong Kong up until 25/10/2020, collected from the website *Inside Airbnb*. There are 74 variables in the original dataset. Among all the variables provided by the website, we selected 13 useful variables, which are the *price, district, room_type, bathroom_type, accommodates, bathrooms, bedrooms, beds, reviews, review_scores, host_response_rate, host_acceptance_rate* and *host_total_listings*:

price refers to the unique price of the accommodation and is stored in integer type

district refers to the location of the accommodation and is stored in character type

room_type refers to the type of accommodation and is stored in character type

bathroom_type refers to the type of bathroom and is stored in character type

accommodates refers to the amount of guests that can be accommodated and is stored in integer type

bathrooms refers to the number of bathroom in the airbnb and is stored in number type

bedrooms refers to the number of bedrooms in the airbnb and is stored in integer type

beds refers to the number of beds in the airbnb and is stored in integer type

reviews refers to the number of reviews that the airbnb receives and is stored in integer type

review_scores refers to the average scores of rating the airbnb receives and is stored in integer type

host_response_rate refers to the rate of the host responding to guests and is stored in number type

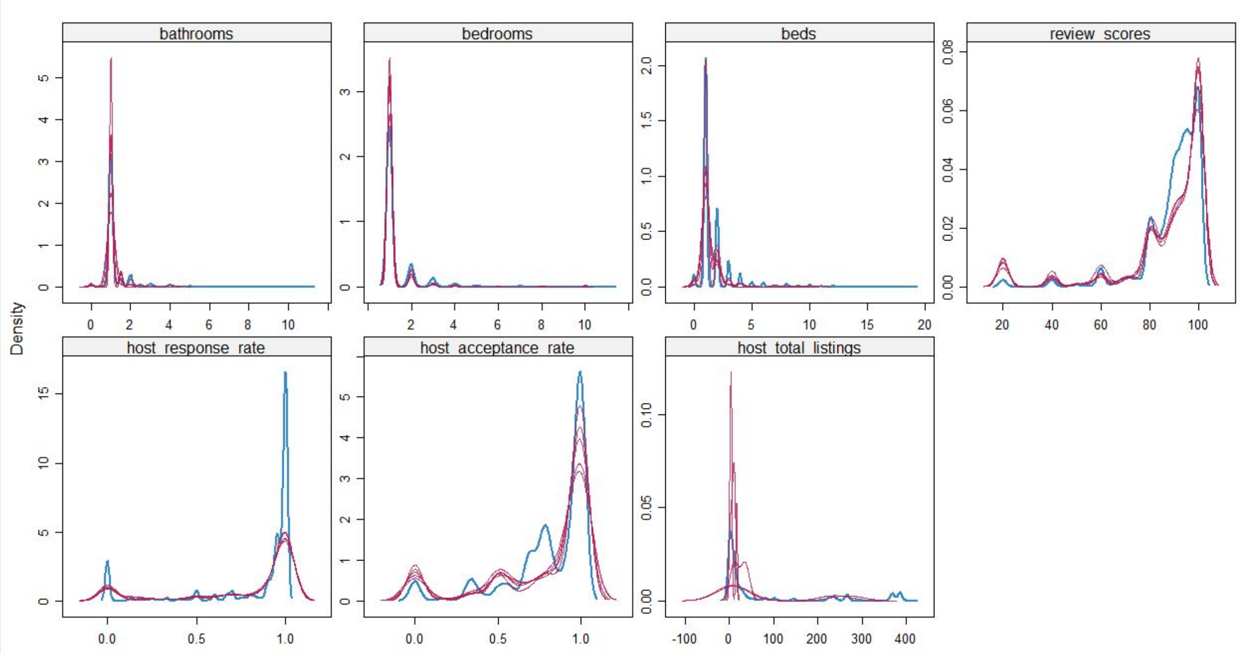
host_acceptance_rate refers to the rate of the host accepting reservation requests and booking inquiries and is stored in number type

host_total_listings refers to the number of airbnb the host has and is stored in integer type.

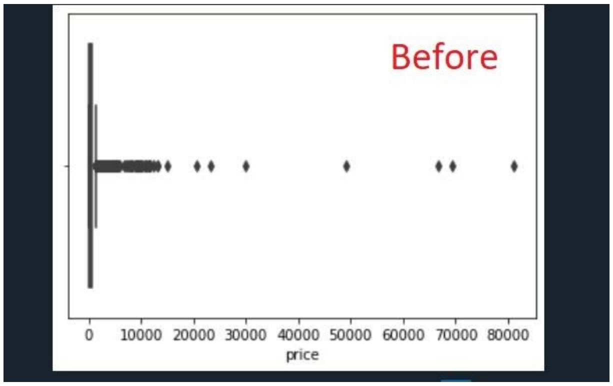
The original dataset is messy and disordered. In order to create a relatively better dataset for model building, we cleaned data in various ways.

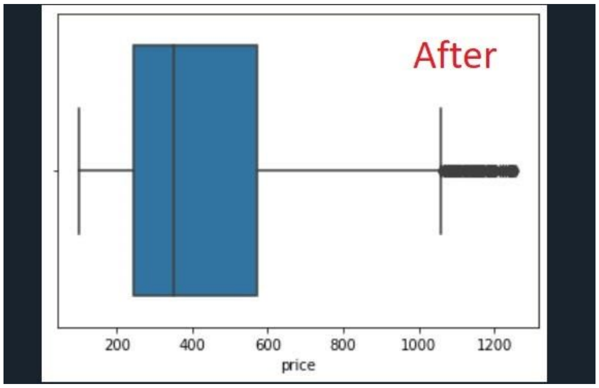
Firstly, we deleted unrelated letters and symbols for every cell and dropped rows with unrelated data. Besides, 13 variables were separated into two types, numeric and non-numeric.

Then, we imputed missing data using MICE Package in R (Multivariate Imputation Chained Equations). The package predicts missing data according to a similar distribution of existing data, which would be more accurate than using mean or median to predict missing data. The distribution of the imputed data (pink) is matching the distribution of the existing data (blue).



After that, we removed outliers for every variable for the accuracy and consistency of the models. For the dependent variable *price*, unreasonable price data points lower than $100 were dropped. Also, we dropped price data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR, by using the inter quartile range method.





For all other independent variables, we used the z-score method. We dropped data points with standard scores over 3, which were the 0.3% outliers assuming normal distribution.

The dataset was ready to use after the above data cleaning.

Model application

Due to limitations of every method and absence of some other variables, this report tries to apply in total 4 methods to increase accuracy and get more profound findings on price prediction.

Model 1: Multiple Linear Regression

Through applying multiple linear regression, this model targets at spotting the significant variables that affect price determination. Noticing most of the selected variables are continuous data, the method is run for finding the best fit line for predicting the output. At the end, by observing the R squared, fitness of the model will be evaluated and possible ways for improvement will be introduced.

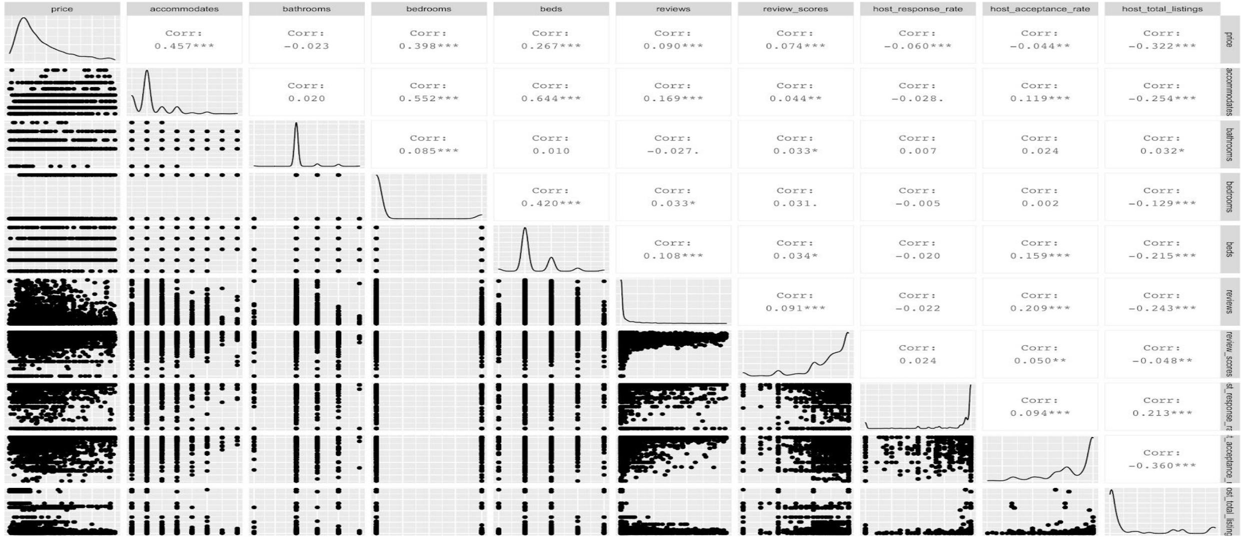
This model takes *price* as a dependent variable, while taking all other variables as independent variables, which include *district, room type, bathroom type, accommodates, bathrooms, bedrooms, beds, reviews, review scores, host response rate, host acceptance rate* and *host total listings*. Among them, there are three variables which do not contain continuous values, which are *district, room type, bathroom type.* Using the "as.factor" function in R, these variables can be modeled into categorical variables in order to shed some light into every part of the variable. To be more specific, categorizing them can evaluate the effect of each district, each room type and each bathroom type on determining price.

At first, the variable *price* is randomly assigned into the training set (qTrain) and test set (qTest) with SplitRatio equal to 0.7. The model will first regress on the training set, followed by regressing on the test set so as to get the out-of-sample R square.

```
> airbnb=read.csv("airbnb.csv")
> library(caTools)
> set.seed(123)
> split=sample.split(airbnb$price,SplitRatio = 0.7)
> qTrain=subset(airbnb,split==TRUE)
> qTest=subset(airbnb,split==FALSE)
> nrow(qTrain)
[1] 3805
> nrow(qTest)
[1] 1590
> qTrain$district=as.factor(qTrain$district)
> qTrain$room_type=as.factor(qTrain$room_type)
> qTrain$bathroom_type=as.factor(qTrain$bathroom_type)
```

We can observe that there are respectively 3805 and 1590 observations in the training set and test set. For the purpose of avoiding unbiased results caused by multicollinearity, correlations among all continuous variables are plotted.

```
> library(GGally)
> ggpairs(qTrain1,axisLabels = "none")
```



It is observed that there is no perfect correlation among variables with continuous values.

The next step is to apply the built in function "lm" to regress dependent variable *Price* on all other independent variables.

```
> predictprice=lm(price~.,data = qTrain)
```

In the variable *district*, *Central & Western* is set as a baseline. In the variable *room type*, *Entire home/apartment* is set as a baseline. In the variable *bathroom type*, *Private bathroom* is set as a baseline. After regressing on the training set, a formula with estimated coefficients is generated:

Price = 374.82 - 67.05 × *Eastern* + 2.46 × *Islands* - 93.84 × *Kowloon City* + 215.21 × *Kwai Tsing* - 78.76 × *Kwun Tong* - 250.26 × *North* - 58.24 × *Sai Kung* - 45.76 × *Sha Tin* - 180.13 × *Sham Shui Po* + 113.83 × *Southern* - 222.62 × *Tai Po* + 70.98 × *Tsuen Wan* - 83.95 × *Tsuen Mun* - 83.68 × *Wan Chai* - 100.18 × *Wong Tai Sin* - 132.73 × *Yau Tsim Mong* - 186.37 × *Yuen Long* - 60.20 × *Hotel room* - 97.39 × *Private room* - 82.70 × *Shared room* - 39.68 × *Shared bathroom* + 58.63 × *accommodates* + 1.49 × *bathrooms* + 152.12 × *bedrooms* - 3.00 × *beds* - 0.18 × *reviews* + 0.79 × *review scores* - 11.78 × *host response rate* - 99.08 × *host acceptance rate* - 0.50 × *host total listings*

Significance

signf.codes	variables
0“***”	Eastern, Kowloon City, Kwai Tsing, North, Sham Shui Po, Wan Chai, Yau Tsim Mong, Yuen Long, Private room, Shared room, Shared bathroom, accommodates, bedrooms, host acceptance rate, host total listings
0.001“**”	Tai Po, Hotel room, review scores

R squared = 0.4109; Adjusted R squared: 0.4062; Residual standard error: 196 on 3774 degrees of freedom

In general, variables *district, room type, bathroom type, accommodations, bedrooms, host acceptance rate* and *review scores* play a relatively more important role in determining price. Hosts are able to determine more suitable prices by valuing more those variables, while travellers can manage to make better staycation choices based on locations, different room types, amount of bedrooms and the capacity of the room, etc.

Predict the output on test set (qTest) for obtaining the out-of-sample data:

```
> predictTest=predict(predictprice, newdata = qTest)
> SSE=sum((predictTest-qTest$price)^2)
> SST=sum((qTest$price-mean(qTrain$price))^2)
> 1-SSE/SST
[1] 0.410776
```

We get a similar R squared from the test set compared to the training set, which means that the linear regression model can get similar accuracy for prediction and forecasting. However, it is noticed that the R squared is relatively low with the ratio of 0.4109 in the sample, which indicates the model needs more variables for improvement. It should be noted that the intercept is also with 0(***) significance, which demonstrates that other variables besides those included in the model would be important for building the model since the intercept reveals the average effect of other variables on the dependent variable *Price*. Future works such as text analysis on comments and adding more numeric variables may be beneficial to improving the multiple linear regression model.

Model 2 : Clustering

There are 18 districts in Hong Kong, for example the Central and Western District, Southern District, Yau Tsim Mong. Based on the fact that infrastructures, area, location and transportation network in every district are different, it is possible that airbnb in different districts have different average price or facilities.

By using the method of clustering, all airbnb records in the database can be divided into some clusters. This facilitates us to analyse the data in the groups with similar intra-cluster characteristics and their respective characteristics.

The K-means method clustering is used in this project as it requires less computational effort and is much simpler than the hierarchical clustering. K is set to be 10 and we perform the clustering according to the neighbourhood (the district) of the airbnb. To facilitate the use of clustering, some amendments are done on the excel file specifically for this model, which is adding 18 columns for the neighbourhoods and the binary value of each airbnb according to their location (0 indicates the airbnb is not in that neighbourhood; 1 indicates the airbnb is in that neighbourhood).

Normalisation of data is also carried out. Variables like bathrooms, bedroom are being normalised by minusing the mean and dividing the standard deviation. For example:

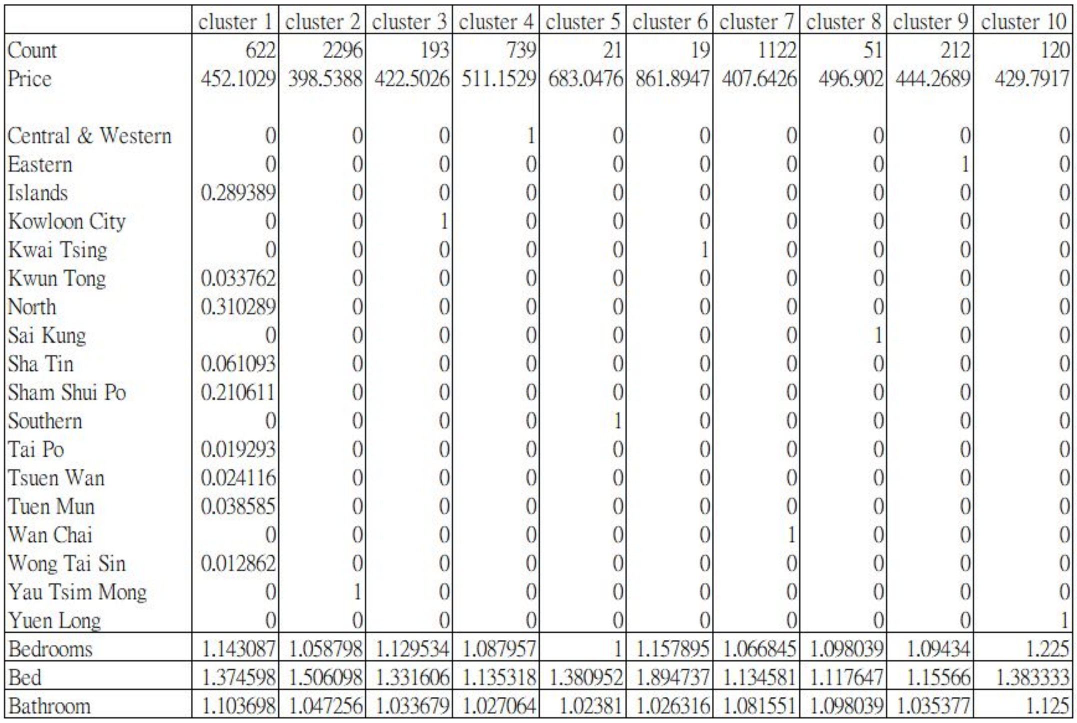
```
>s=sd(airbnb_normal$bathrooms)
>x=mean(airbnb_normal$bathrooms)
>airbnb_normal$beds=(airbnb_normal$bathrooms-x)/s
```

After normalisation, clustering is performed.

```
>set.seed(3000)
```

```
>k=10
>km.out=kmeans(airbnb_normal[14:31],k,nstart=25)
>km.clusters=km.out$cluster
```

The following is the result of clustering:



	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10
Count	622	2296	193	739	21	19	1122	51	212	120
Price	452.1029	398.5388	422.5026	511.1529	683.0476	861.8947	407.6426	496.902	444.2689	429.7917
Central & Western	0	0	0	1	0	0	0	0	0	0
Eastern	0	0	0	0	0	0	0	0	1	0
Islands	0.289389	0	0	0	0	0	0	0	0	0
Kowloon City	0	0	1	0	0	0	0	0	0	0
Kwai Tsing	0	0	0	0	0	1	0	0	0	0
Kwun Tong	0.033762	0	0	0	0	0	0	0	0	0
North	0.310289	0	0	0	0	0	0	0	0	0
Sai Kung	0	0	0	0	0	0	0	1	0	0
Sha Tin	0.061093	0	0	0	0	0	0	0	0	0
Sham Shui Po	0.210611	0	0	0	0	0	0	0	0	0
Southern	0	0	0	0	1	0	0	0	0	0
Tai Po	0.019293	0	0	0	0	0	0	0	0	0
Tsuen Wan	0.024116	0	0	0	0	0	0	0	0	0
Tuen Mun	0.038585	0	0	0	0	0	0	0	0	0
Wan Chai	0	0	0	0	0	0	1	0	0	0
Wong Tai Sin	0.012862	0	0	0	0	0	0	0	0	0
Yau Tsim Mong	0	1	0	0	0	0	0	0	0	0
Yuen Long	0	0	0	0	0	0	0	0	0	1
Bedrooms	1.143087	1.058798	1.129534	1.087957	1	1.157895	1.066845	1.098039	1.09434	1.225
Bed	1.374598	1.506098	1.331606	1.135318	1.380952	1.894737	1.134581	1.117647	1.15566	1.383333
Bathroom	1.103698	1.047256	1.033679	1.027064	1.02381	1.026316	1.081551	1.098039	1.035377	1.125

It can be observed that with k=10, which means dividing the data into 10 clusters according to their neighbourhood, 9 neighbourhoods will be grouped as 1 cluster, and the remaining neighbourhoods will be grouped in 9 separate clusters. It is expected that the clustering result will be different if a different k is used. However, it is clear that every cluster has its own characteristics: its average price, average number of facilities included, i.e. bedrooms, bed, bathrooms.

Finding:

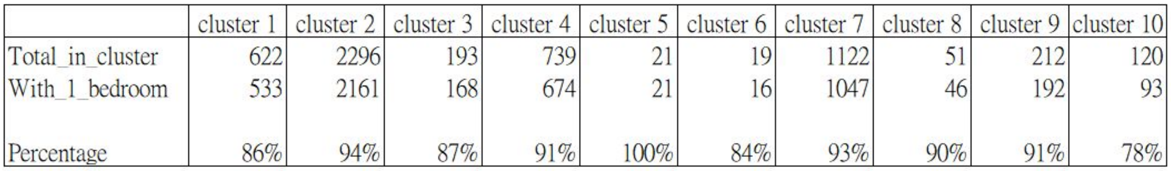
Clustering is regarded as one of the unsupervised learning methods. Dependent variable is not set in clustering and hence no prediction will be made. However, based on the grouping result of this method, we can analyse the k clusters and their characteristics one by one.

For example, the mean price of airbnb, mean number of facilities like bedroom in each clusters can be found by tapply:

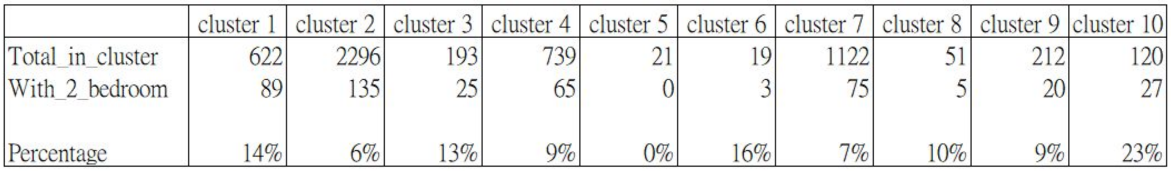
```
>tapply(airbnb$price,km.clusters,mean)
>tapply(airbnb$bedrooms,km.clusters,mean)
```

Percentage of airbnb with specific requirement among all airbnb in each clusters can also be found by:

```
>rn=which(airbnb$bedrooms == 2)
>airbnb[rn,]
>test=km.clusters[rn]
>table(test)
```



	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10
Total_in_cluster	622	2296	193	739	21	19	1122	51	212	120
With_1_bedroom	533	2161	168	674	21	16	1047	46	192	93
Percentage	86%	94%	87%	91%	100%	84%	93%	90%	91%	78%



	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7	cluster 8	cluster 9	cluster 10
Total_in_cluster	622	2296	193	739	21	19	1122	51	212	120
With_2_bedroom	89	135	25	65	0	3	75	5	20	27
Percentage	14%	6%	13%	9%	0%	16%	7%	10%	9%	23%

Analysis:

The findings in clustering can benefit travellers and hosts.

For travellers with specific requirements on the airbnb, they can take the result as reference and focus on finding an airbnb in certain districts based on their preferences. For example, a traveller with a larger family requires a 2-bedroom airbnb. Based on the result above, it is suggested that he can consider finding airbnb in cluster 10 (Yuen Long). As the percentage of 2-bedroom airbnb is the highest in all clusters (23%), they are more likely to find a 2-bedroom airbnb in Yuen Long, compared with that in cluster 5 (Southern), where all available airbnb is 1-bedroom.

For hosts, it is crucial for them to set the rent of their airbnb reasonably, therefore, setting the rent similar to the mean price of airbnb which is similar to their airbnb is suggested. As clustering divides all airbnb into different clusters with high similarity of intra-group characteristics, all airbnb in the same cluster is regarded as “similar”. Therefore, hosts can consider setting the rent similar to the average price of the cluster which their airbnb is included in. For example, hosts with an airbnb in Sai Kung (cluster 8) can set the rent at around $497.

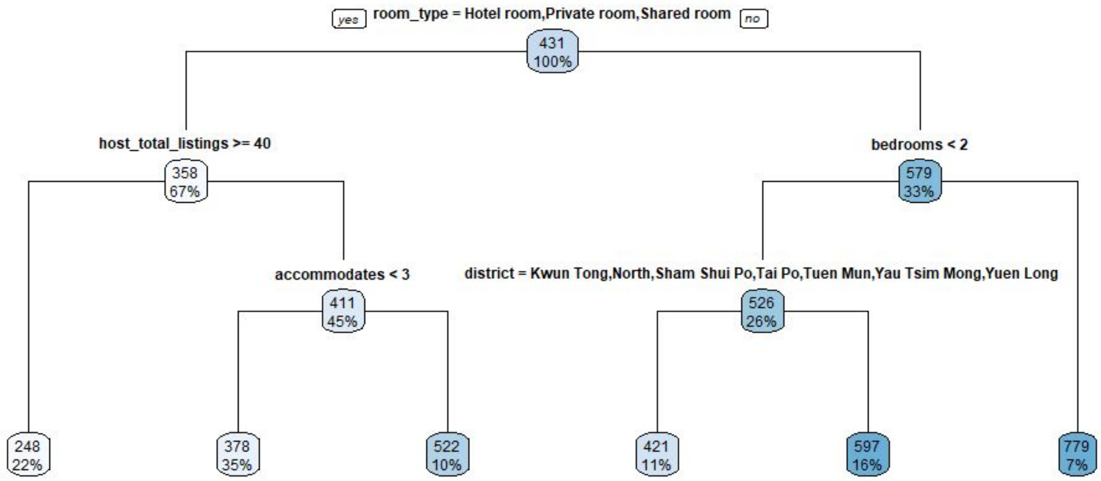
For investors or hosts that would like to invest in the property market and turn the house into an airbnb business, it is essential for them to consider the possible rental return. As airbnb in different districts may have a very different rent, it is best for investors or hosts to buy a house in districts with high average rent of airbnb. The clustering result and the average price of airbnb in each cluster may give them some insights. For example, if an investor is considering investing in either the Southern District or Wan Chai, the result found will suggest them to invest in the Southern District over Wan Chai as the average price (rental return to hosts) in the former is much higher than the latter.

Model 3: Regression Tree

Decision tree method is a useful machine learning method in prediction, which arranges observations into predicted regions. We used two regression trees here to build models that explain the price of airbnb in an interpretive way. We kept *price* as our dependent variable and other 12 independent variables same as the linear regression. We used a simple regression tree method with a minimum bucket constraint as our first model and a k-fold validation method as our second model. Our target is to find important variables that affect the price and to find the model with the best explanatory power.

First model:

```
>library(rpart)
>library(rpart.plot)
>Tree1=rpart(price~., data=Train, method='anova', minbucket=250)
>rpart.plot(Tree1, type=4)
```


yes
room_type = Hotel room,Private room,Shared room
no
431
100%
host_total_listings >= 40
358
67%
bedrooms < 2
579
33%
accommodates < 3
411
45%
district = Kwun Tong,North,Sham Shui Po,Tai Po,Tuen Mun,Yau Tsim Mong,Yuen Long
526
26%
248
22%
378
35%
522
10%
421
11%
597
16%
779
7%

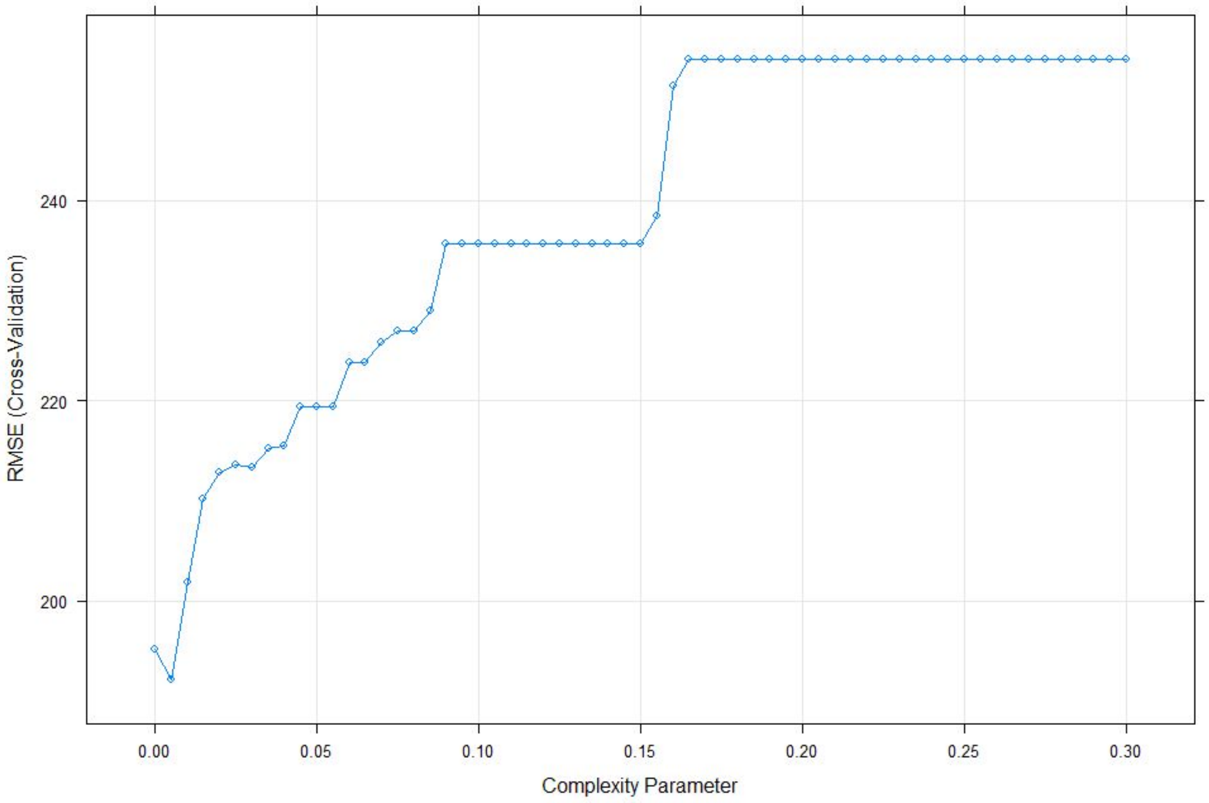
For the first model, considered about 5400 observations in total, we used minbucket=250 to restrict the number of points in each subset in order to have a balance between accuracy and overfitting issue. Hotel room, private room and shared room types tend to have lower prices (~$358) than the entire apartment type (~$579). Besides, homes that are able to accommodate more people would have higher prices. Accommodations located in Kwun Tong, North, Sham Shui Po, Tai Po, Tuen Mun, Yau Tsim Mong and Yuen Long have averagely lower prices than other districts. The most important factor in determining accomodation price is room type. Host total listings and amount of bedrooms are important to determine the price. Surprisingly, variables district and accommodates in this model are relatively less important than the other three variables above.

```
>library(forecast)
>Prediction1 = predict(Tree1, newdata=qTest, type='matrix')
>accuracy(Prediction1, qTest$price)
```

After running the models with the test set, we got a RMSE (root mean squared error) =200.4416. RMSE is a measure of accuracy here, which measures the variation between observed data and predicted data.

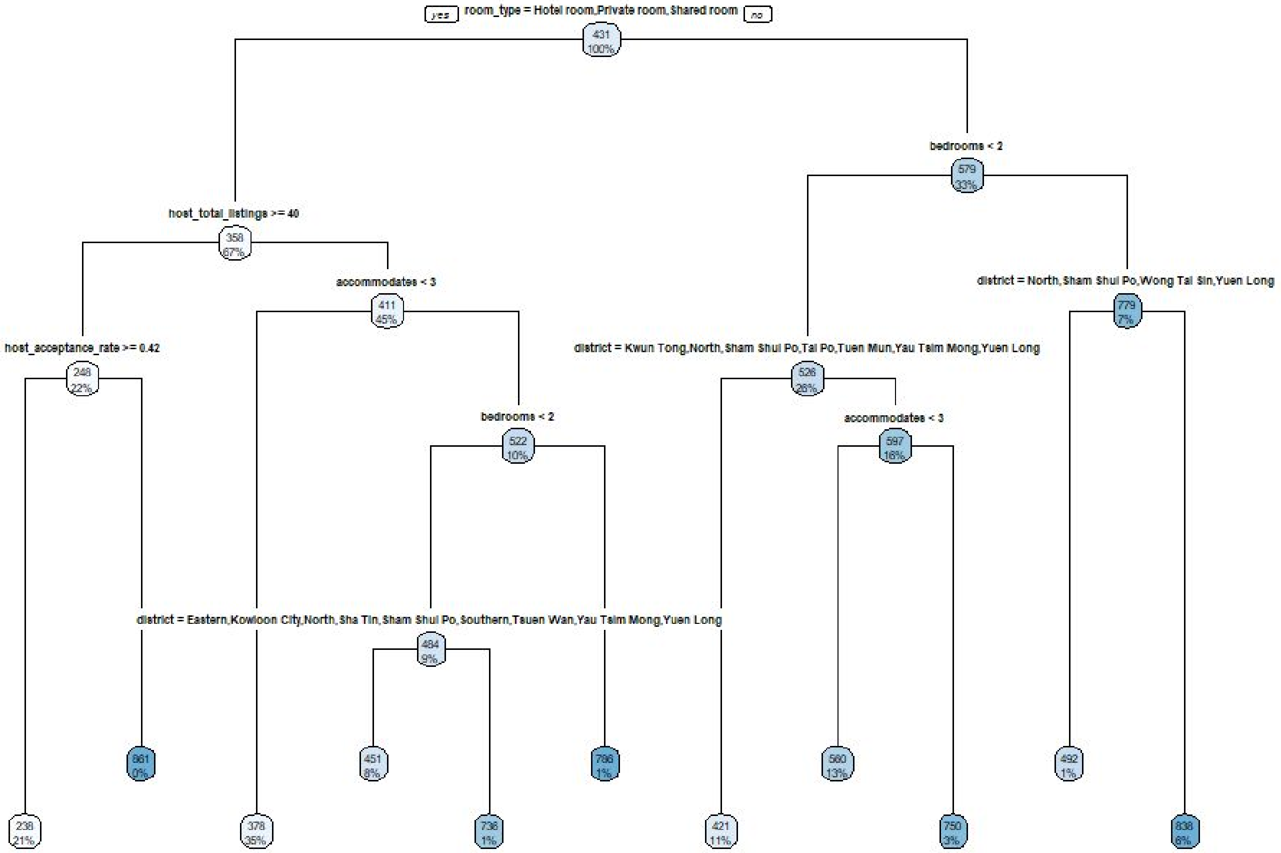
Second model:

```
>library(caret)
>library(e1071)
>set.seed(123)
>numFolds = trainControl(method = "cv", number = 10)
>cpGrid = expand.grid(cp = seq(0,0.3,0.005))
>result = train(price~., qTrain, method = "rpart", trControl = numFolds, tuneGrid = cpGrid)
>plot(result)
```

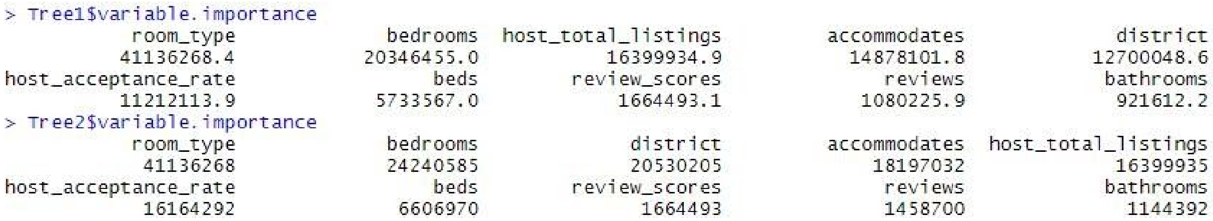


Using the cross-validation method with 10 folds, we found that cp=0.01 was the best fit to the model with lowest prediction error. The parameter measures trade-off between complexity and accuracy. Although the RMSE could be even smaller if we picked cp < 0.01, there may be an overfitting issue. Therefore we chose cp=0.01 here.

```
>Tree2=rpart(price ~., data = qTrain,method="anova", cp = 0.01)
>rpart.plot(Tree2,type=1)
```


yes
room_type = Hotel room,Private room,Shared room
no
431
100%
host_total_listings >= 40
358
67%
bedrooms < 2
579
33%
accommodates < 3
411
45%
district = North,Sham Shui Po,Wong Tai Sin,Yuen Long
779
7%
host_acceptance_rate >= 0.42
248
22%
district = Kwun Tong,North,Sham Shui Po,Tai Po,Tuen Mun,Yau Tsim Mong,Yuen Long
526
26%
bedrooms < 2
522
10%
accommodates < 3
597
16%
district = Eastern,Kowloon City,North,Sha Tin,Sham Shui Po,Southern,Tsuen Wan,Yau Tsim Mong,Yuen Long
484
9%
861
0%
451
8%
786
1%
560
13%
492
1%
238
21%
378
35%
736
1%
421
11%
750
3%
838
6%

The second model is more complicated than the first model. It has similar characteristics with the first model but it has more leaves in the tree. If we look into the importance of each variable, variable district is more important than host total listings now, compared to the first model.



```
> Tree1$variable.importance
           room_type             bedrooms  host_total_listings         accommodates             district
          41136268.4           20346455.0           16399934.9           14878101.8           12700048.6
host_acceptance_rate                 beds        review_scores              reviews            bathrooms
          11212113.9            5733567.0            1664493.1            1080225.9             921612.2
> Tree2$variable.importance
           room_type             bedrooms             district         accommodates  host_total_listings
            41136268             24240585             20530205             18197032             16399935
host_acceptance_rate                 beds        review_scores              reviews            bathrooms
            16164292              6606970              1664493              1458700              1144392
```

```
Prediction2 = predict(Tree2,newdata=qTest, type='matrix')
accuracy(Prediction2,qTest$price)
```

After running the models with the test set, we got a RMSE (root mean squared error) =192.3258. The second model has a lower prediction error than the first model and the accuracy is higher.

In conclusion ,we should use the second model as our final choice as it has a lower RMSE. We also found that the most important variables in descending order are *room_type*, *bedrooms*, *district*, *accommodates* and *host_total_listings*. For *room_type*, Hotel room, private room and shared room types tend to have lower prices. For *district*, Kwun Tong, North, Sham Shui Po, Tai Po, Tuen Mun, Yau Tsim Mong, Wong Tai Sin and Yuen Long have averagely lower prices.

Model 4: Random Forest

The model starts off by factorizing some variables such as *accommodates, bathrooms, bedrooms* and *beds*. The factorization is to prevent the modelling engine from presuming a trend correlated with the scalar amount of those variables. Comparison was made between before and after the factorization using the same model mechanism. The regression model result was similar yet very slight improvement was found after factorizing two of the variables: *bathrooms* and *bedrooms*.

```
$ % Var Explained
TestRF (Original Data) : 59.45
TestRF1 (Data with 4 factorized variables) : 59.3
TestRF2 (Data with 2 factorized variables) : 59.99
# ntree = 500, mtry =10
```

There are two important variables to be input in the model, *ntree* and *mtry*. The former one represents the total number of decision trees to be set up in the model consisting of the voting where as the later one decides the number of variables concerned in each of the trees, the former one affects the size of the forest whereas the later one affects the size of each tree (depth and length).

There are numerous ways to find the optimal solution that fit the model best. In this study, the most straightforward way was used. For *ntree*, the same regression was made with the change only on the number of trees to find the one with highest % Var Explained.

```
>airbnb = read.csv(“airbnb.csv”)
>set.seed(88)
>RF500 = randomForest (price~., data = airbnb , ntree = 500)
>RF1000 = randomForest (price~., data = airbnb , ntree = 1000)
>RF2000 = randomForest (price~., data = airbnb , ntree = 2000)
#default mtry value in this regression is total number of variables /3
```

The result showed that the RF1000 has the highest % var Explained, hence *ntree* is denoted to be optimal at *ntree*=1000.

Furtheron, the *mtry* is another variable to be decided at its best value. R's own mechanism was used in this case to lookup the optimal value of *mtry* with minimum Out of Bag (OOB) error. The function *tuneRF* was used to find out the value, whereas using *ntree*=1000.

```
> print(mtry)
  mtry OOBError
3    3 25808.77
4    4 25495.62
6    6 25408.35
```

The result shown that the best value of *mtry* is 6.

The model is then put to regression by the above fine tuning result.

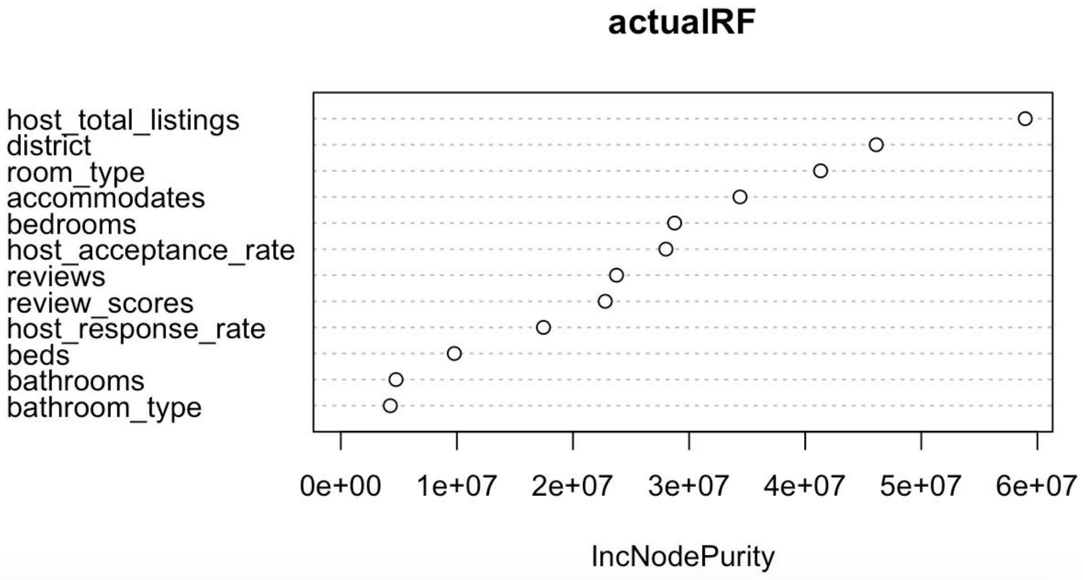
```
> actualRF = randomForest(price ~. , data = airbnb, ntree = 1000, mtry = 6)
> print(actualRF)

Call:
 randomForest(formula = price ~ ., data = airbnb, ntree = 1000,      mtry = 6)
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 6

          Mean of squared residuals: 25366.13
                    % Var explained: 60.12
```

It can be shown that the % Var explained in the model is at 60.12, which means around 60% of the figures can be predicted by the model of random forest.

Looking into the data, there are four variables that are comparatively more significant than others. By *varImpPlot()*, the *IncNodePurity* can be shown, as the rating of each variable in the contribution of building the accuracy of the model.



From above, it can be told that the *host_total_listing* is the most significant variable whereas the *district* is the second and so on. It coherently inherited the same result from Model 1 and further proved that these two variables are the most important factor in determining the price of the listing.

Summary

In general, the four models separately provide different insights to various means and implications for the stakeholders to consider.

It is agreed by different models that the *district*, *host_total_listing*, *room_type* and *accommodates* are the significant factors that determine the price. Which given that one of the important observations across all the models, that the variance explained is relatively low compared to a robust model (generally reaching around 50% of the variance explained). The intercept is having a high significance which shows that there are some omitted variables which might also have determining effects towards the dependent variable, the pricing model.

Overall, the four models individually provide four different outcomes which can be generalized to some criterias as reference to the stakeholders, the models can also be providing some reference figures for potential cases that are not yet listed on airbnb yet or any cases that wished to reprice their properties.

The implication to the travellers can be interpreted as some consideration criteria when they are looking for a suitable accommodation, and the variables listed can become one of the concerns during the planning process. Whereas for the house owners, this can be taken as a reference of how the price should be set to satisfy the need and demand of the market, or if the owner preferred to perform a promotion, what is the suitable value of price that should be set to attract the travellers. The models have provided some insights for the owners to generate a suitable range of price.

For investors and potential house acquisitors, they can take into consideration of how would be the potential rental return for buying different properties with different onset, while the owners can further plan regarding hardwares to satisfy the need of the market, they can maximise the potential capital-return rate on their investment.

Limitation and improvements

1. This research is limited to the third-party data source which the completeness and data integrity is not secured. The dataset may just be a sample of total data in Hong Kong. The variables themselves have incomplete data points and some of them may not be listed in the dataset. For example, each district has different amounts of data points. Some districts with only a few data points would make them not significant in model and have very high variance in practice. Some districts with large amounts of data points would definitely have smaller variance and easier to have higher significance in the model. Yau Tsim Wong has 2296 data points while Wong Tai Sin has only 8 data points. When doing a price prediction on a specific district, there will be high variation if the amount of data points is too small. The explanatory power of the model will be seriously affected in practice.

   We could improve this by scraping all data ourselves on Airbnb websites using web scraping techniques, in order to create a more unbiased sample.

2. More variables are also desired in this research as the regression result shows that there might be some possible missing variables that can contribute to the robustness, reliability and accuracy of the regression model. For example, the overall beauty level of web page design for each listing, the amount of accommodation photo uploaded for each listing and the amount of adjectives used in describing accommodation by hosts for each listing. These are the variables that would affect the users decision when selecting accommodation.

   We could improve this by doing text analytics and web scraping. Finding the amount of adjectives could be useful to measure the attractive level in terms of text. Finding the number of pictures could be useful to measure the attractive level in terms of graphics.

3. 4 models having 4 interpretations may be complicated for users to compare results in practice. For the multiple linear regression model, it provides information regarding the

assigned value of each variable, but it can be further broken down by the clustering method. Clustering can provide a refined subset for the regression model to generate more precise and concise figures, whereas the random forest model and regression tree model can provide a reference as an alternative model for evaluating the suitable recommended price for the property.

We could combine the 4 models into 1 model by using ensemble methods. It produces a more accurate result than using 4 models separately. However, this topic is not easy for us now and we would like to explore more in the future.

Dataset source:

http://data.insideairbnb.com/china/hk/hong-kong/2020-10-25/data/listings.csv.gz

Reference

Airbitcs. (2020, April 21). A snapshot of the Hong Kong Airbnb market amidst the pandemic. Retrieved from https://airbtics.com/analyzing-hong-kong-airbnb/

Airbnb. (2020). About Us. Retrieved from https://news.airbnb.com/about-us/

McDonald, A. & Look, C. (2019, December 3). These Are the World's Most Popular City Destinations in 2019. *Bloomberg Travel*. Retrieved from https://www.bloomberg.com/news/articles/2019-12-03/these-are-the-world-s-most-popular-city-destinations-in-2019