New York City Airbnb Analysis

Zoom recording:
https://usc.zoom.us/rec/play/ORG8ffj2p8pH_H6oKM0A5iE5r0KFMcQl48NBJekmUO224lp7wIx6iYGtEY4KtOMJQnU_uXtnJwWMApq2.o0AKPBZmDHlSj2Ik?startTime=1669107661000

ISE 535 Final Project

Group 10: Shao-Chi Wu / Goureesh Yarlagadda / Szu-Yu Chen / Xingyao Tang

EDA – Project Overview

- Business objectives
  - The marketing manager from the Airbnb hopes to do an analysis ***on different factors compared to the price*** throughout the year for individual listing.
- Data Source
  - A dataset of airbnb listing and availability of the year within different district of the New York City
  - Data consisting of 16 attributes and 48,896 observations where each observation represents a single reservation
  - The data is obtained from the website:https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data

EDA – Initial Data Review

Interpretation for Large Data Missing

We observed that the last_review, last_review_in_day, reviews_per_month, have 10 thousands missing values :

- We conclude that the large data missing makes sense here. Because when a listing has no review, the variables related to review will have no info (N/A).
- Other than the missing data in the review columns, the dataset appears to have no obvious error
- We added the interaction term “review_performance” with a few details in mind:
  - This interaction terms help solving the problem of having large missing data.
  - We think that it’s more accurate if we consider number of the review and the last_review date together, it give us an idea of ***how good*** a review is.
  - Formula: **rescale[1/last_review_in_days, to c(0.5,1.5)] X number of reviews**

EDA - Variable Analysis

Summary of Attributes After Dropping outliers and irrelevant variables

Numeric Variables:

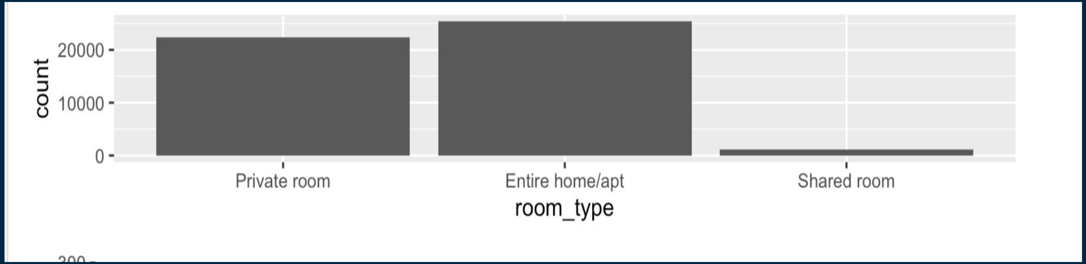
Variable	Description	Logical Group
latitude	Latitude	Location
longitude	Longitude	
number_of_reviews	No. of reviews for the listing	Review Info
last_review	The date of the last review	
reviews_per_month	Average no. of the reviews per month	
last_review_in_days	How old a the last review's listing is in days.	Review Performance
review_performance	Interaction term that normalizes the performance of a review	
minimum_nights	Latitude	Booking
availability_365	Longitude	

*The text in red are artificial variables added afterward.

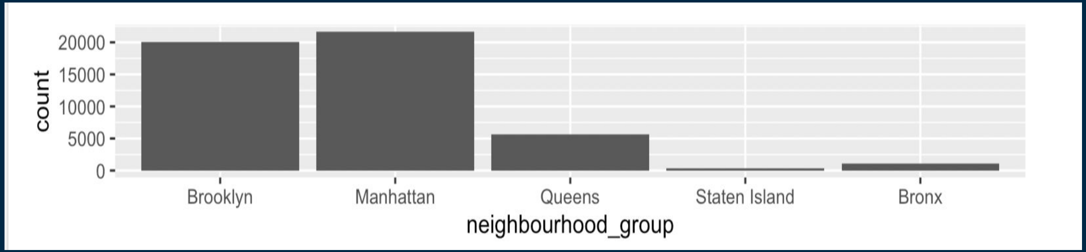
Factor Variables:

Variable	Description	Logical Group
neighbourhood_group	Location of the listing	Neighbour
neighbourhood	District of the listing	
calculated_host_listings_count	Total No. of listing for a host	Host
availability_365	Number of days a listing is available during a year.	
room_type	The type of space	

Univariate Analysis

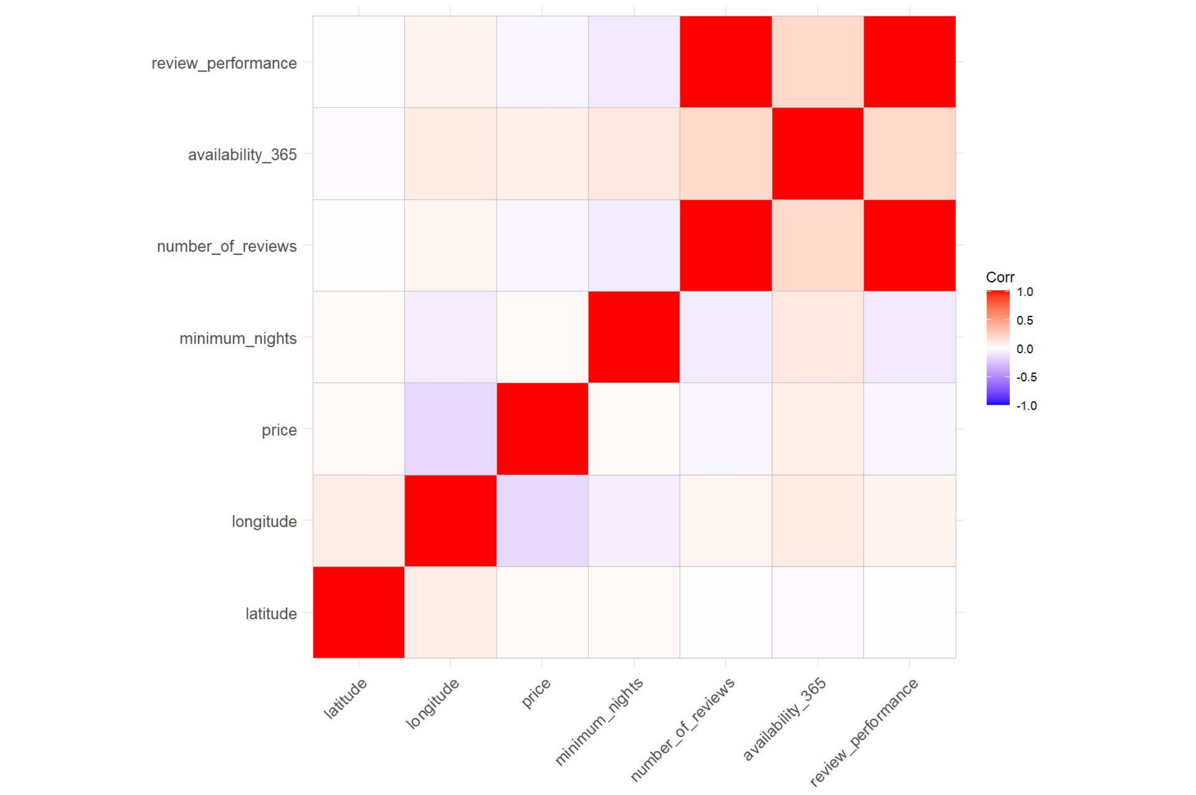


Considering the total number of rooms , “Shared room” is the least preferred type to be converted to an Airbnb



Close to 80% of the Airbnb listings are located in Brooklyn and Manhattan

Bivariate Analysis



Price and Longitude might have a negative relationship.

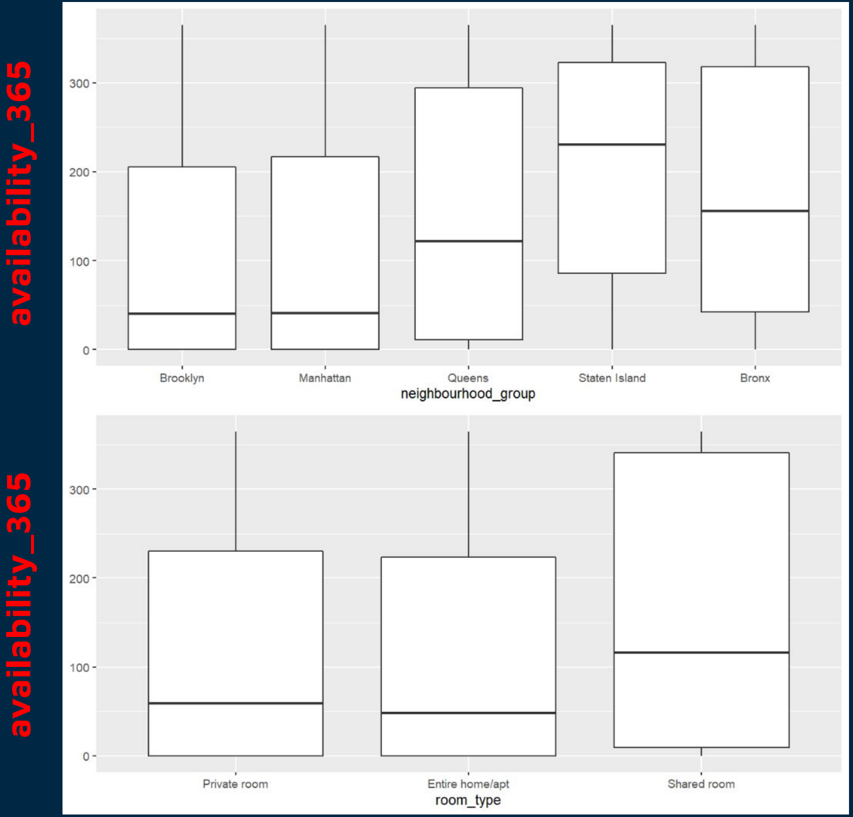
Price and availability_365 might have a positive relationship.

T test

	P- value	Confidence Interval	Estimated sample correlation
Price vs Longitude	< 2.2 e-16	[214.3177 , 218.2477]	-0.1552956
Price vs availability_365	< 2.2 e-16	[25.12793 , 29.81915]	0.07830582

Based on t tests and correlation tests:
There is true correlation between price and longitude, and if price increased by $1, the longitude will decrease 0.155
Also, we can state that price and availability_365 have true correlation, and if increased by $1, the availability days in year might increase 0.078

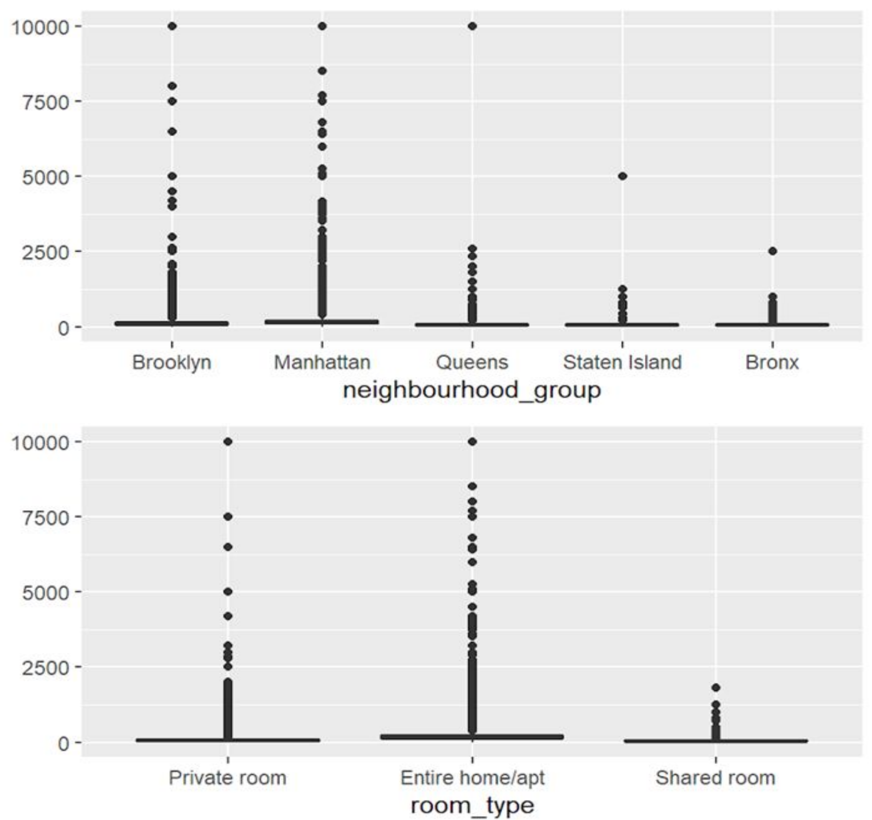
Bivariate Analysis



The upper graph shows that the available days for different neighborhood groups are different, this might be due to the fact that some areas are popular for visiting.

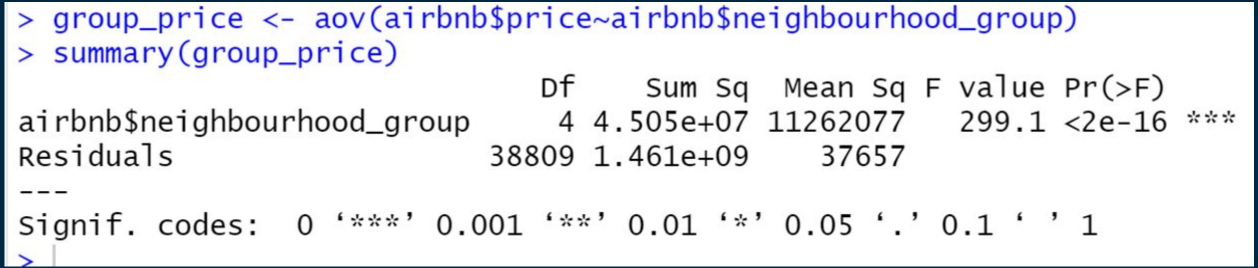
For the lower graph, we found that there are more shared rooms than two other types of rooms. There are more share rooms available in a year because it is cheaper than others and might be easier to be booked. So hosts are more likely to release more shared rooms than other two types.

Bivariate Analysis



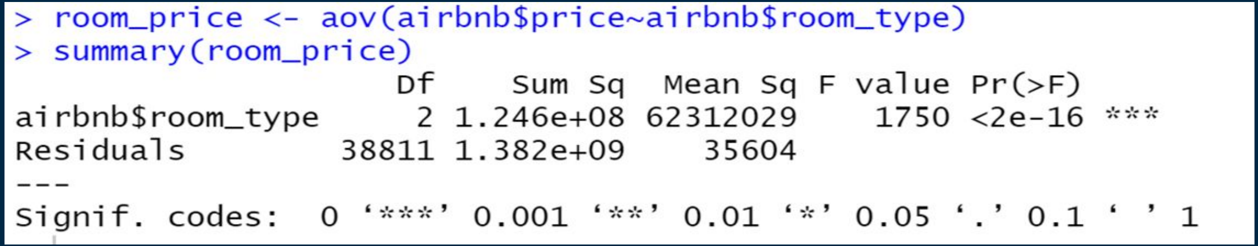
Based on the graph, we guess that variables of "neighbourhood_group" and "room_type" might influence the price.

Statistical Data Analysis – ANOVA



```
> group_price <- aov(airbnb$price~airbnb$neighbourhood_group)
> summary(group_price)
                              Df    Sum Sq  Mean Sq F value Pr(>F)    
airbnb$neighbourhood_group     4 4.505e+07 11262077   299.1 <2e-16 ***
Residuals                  38809 1.461e+09    37657                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The average price in different neighborhoods groups might not be the same.



```
> room_price <- aov(airbnb$price~airbnb$room_type)
> summary(room_price)
                    Df    Sum Sq  Mean Sq F value Pr(>F)    
airbnb$room_type     2 1.246e+08 62312029    1750 <2e-16 ***
Residuals        38811 1.382e+09    35604                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Different type of rooms might have different price

We might say no matter that airbnb is located in which neighborhoods groups and what types of rooms, the price each night seems no difference

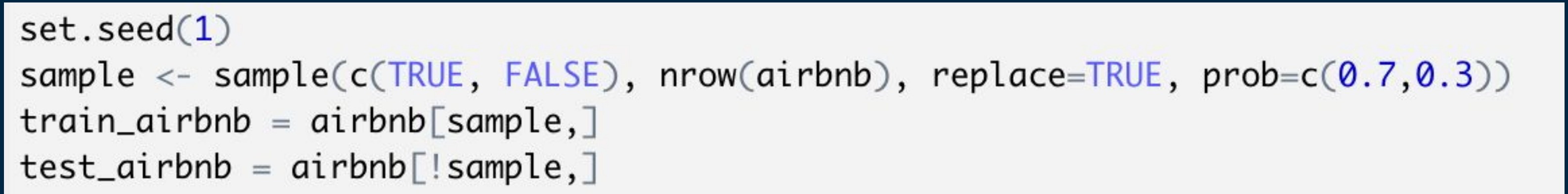
Decision tree model

Regression tree

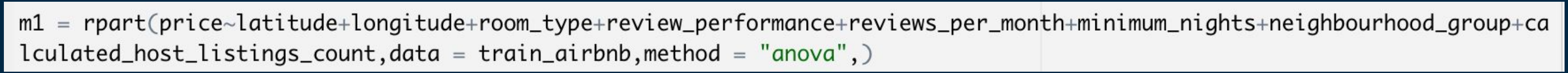
01

Regression Tree Model

Even though the decision tree model is not suitable for this dataset, but we still try using this model to do analysis.
We decide to use these factors(latitude+longitude+room_type+review_performance+reviews_per_month+minimum_nights+neighbourhood_group+calculated_host_listings_count) to train this model.
And split the data into train and test dataset.



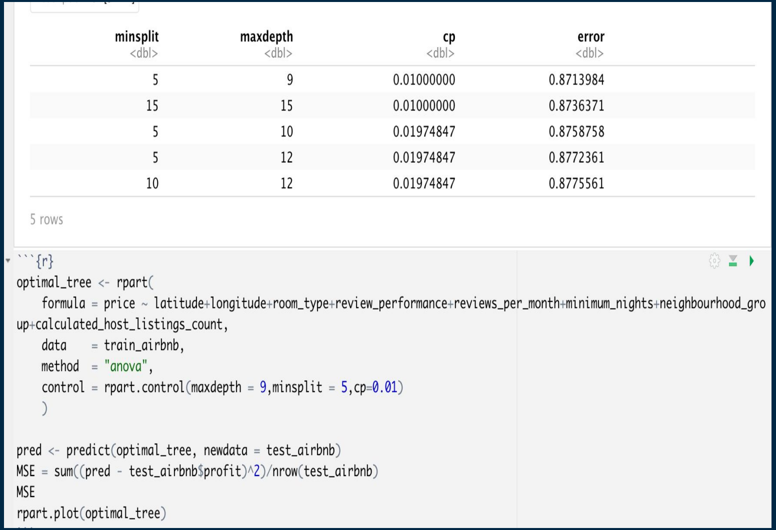
```
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(airbnb), replace=TRUE, prob=c(0.7,0.3))
train_airbnb = airbnb[sample,]
test_airbnb = airbnb[!sample,]
```



```
m1 = rpart(price~latitude+longitude+room_type+review_performance+reviews_per_month+minimum_nights+neighbourhood_group+ca
lculated_host_listings_count,data = train_airbnb,method = "anova",)
```

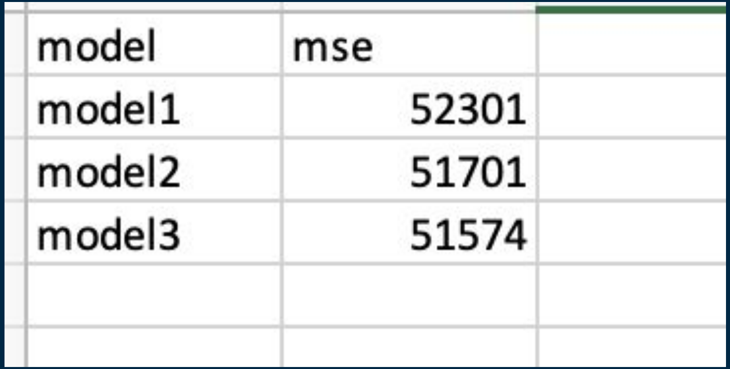
Regression tree

We prune the tree and find the optimal tree



```
 minsplit   maxdepth           cp       error
    <dbl>      <dbl>        <dbl>       <dbl>
        5          9   0.01000000   0.8713984
       15         15   0.01000000   0.8736371
        5         10   0.01974847   0.8758758
        5         12   0.01974847   0.8772361
       10         12   0.01974847   0.8775561
5 rows
```

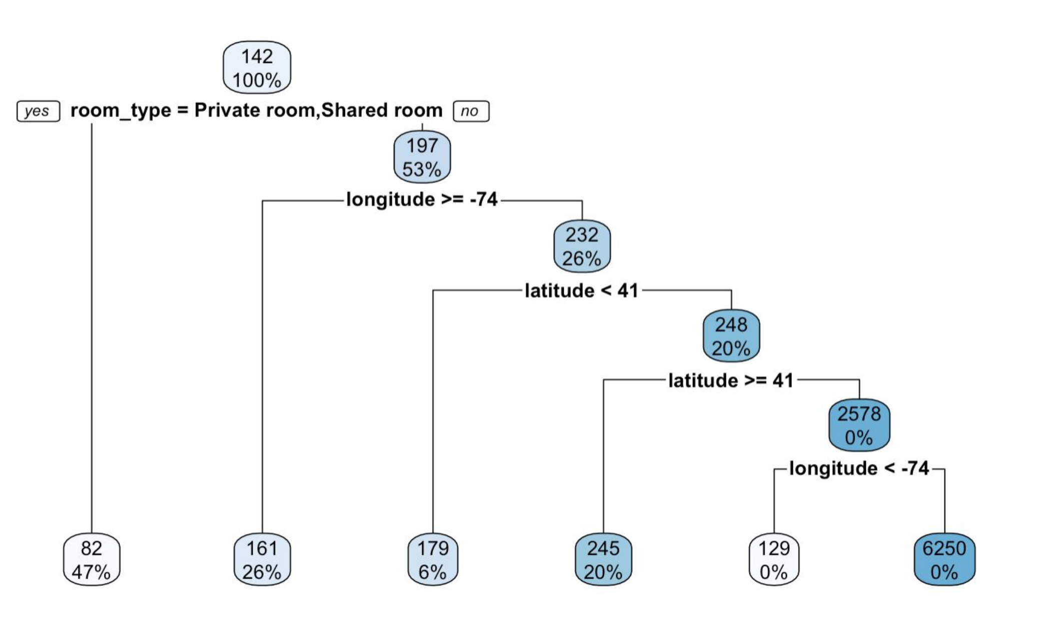
~~~
```{r}
optimal_tree <- rpart(
    formula = price ~ latitude+longitude+room_type+review_performance+reviews_per_month+minimum_nights+neighbourhood_gro
up+calculated_host_listings_count,
    data     = train_airbnb,
    method   = "anova",
    control  = rpart.control(maxdepth = 9,minsplit = 5,cp=0.01)
    )

pred <- predict(optimal_tree, newdata = test_airbnb)
MSE = sum((pred - test_airbnb$profit)^2)/nrow(test_airbnb)
MSE
rpart.plot(optimal_tree)
~~~



model	mse
model1	52301
model2	51701
model3	51574

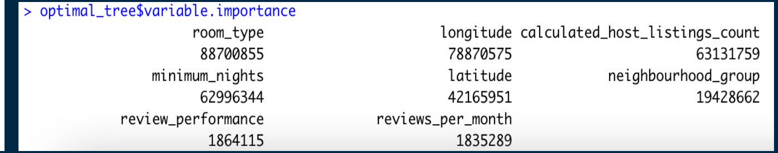
We have three models and mse decrease from 52301 to 51576

Result About the Optimal Tree



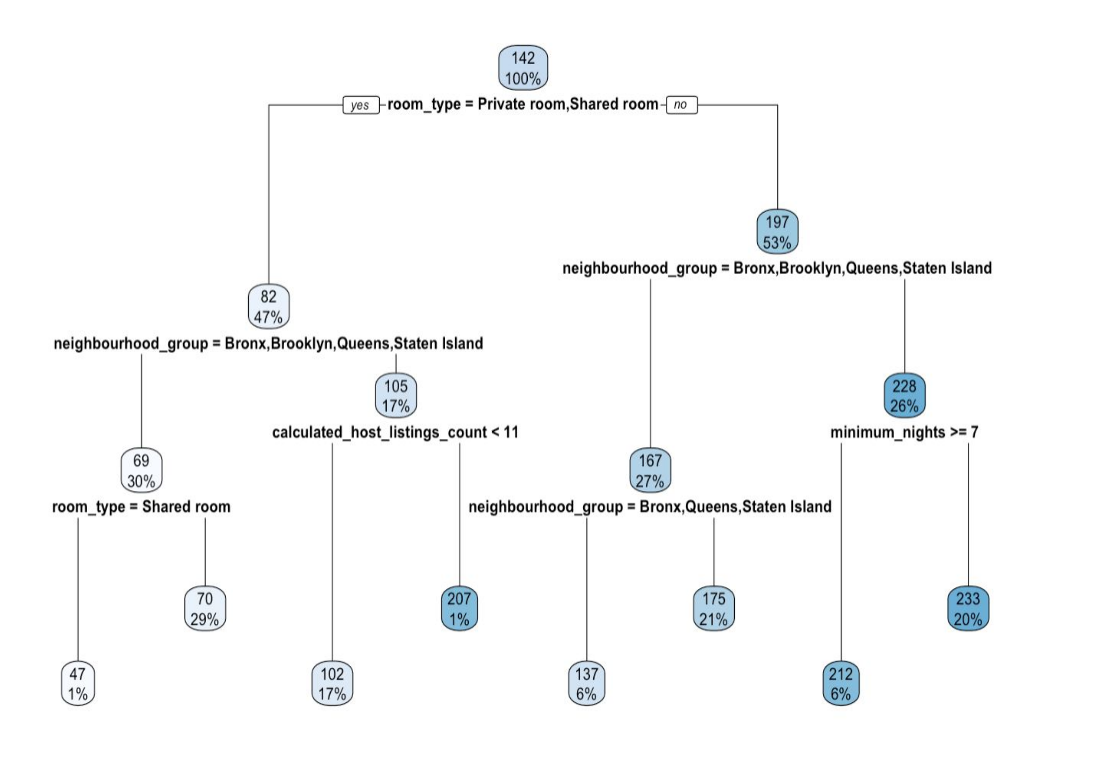
Room_type is the most important factor. If the room_type is Private room or Shared room, the average price is likely 82. Longitude is also a significant factor, if longitude >= -74 then 26% average rental price will be 161.

Checking the variable importance, we find room_type, longitude, host_listings_count play a important role in the tree model



```
> optimal_tree$variable.importance
                   room_type                    longitude calculated_host_listings_count
                    88700855                     78870575                     63131759
              minimum_nights                     latitude          neighbourhood_group
                    62996344                     42165951                     19428662
          review_performance            reviews_per_month
                     1864115                      1835289
```

Result about my tree


142
100%
yes
room_type = Private room,Shared room
no
82
47%
neighbourhood_group = Bronx,Brooklyn,Queens,Staten Island
69
30%
room_type = Shared room
47
1%
70
29%
105
17%
calculated_host_listings_count < 11
102
17%
207
1%
197
53%
neighbourhood_group = Bronx,Brooklyn,Queens,Staten Island
167
27%
neighbourhood_group = Bronx,Queens,Staten Island
137
6%
175
21%
228
26%
minimum_nights >= 7
212
6%
233
20%

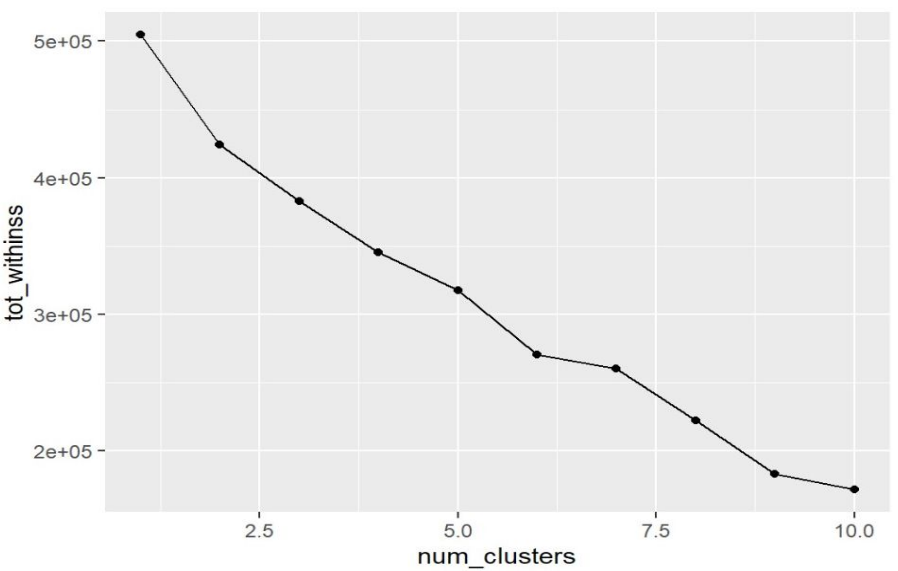
In this tree we used room_type+review_performance+minimum_nights+neighbourhood_group+calculated_host_listings_count.

We can see the room_type is the most important variable in this tree. The second important variable is neighbourhood_group, the least is calculated_host_listings_count. If the room is not apartment and not located in Manhattan and the calculated_host_listings_count is less than 11, the 17% of the room's average price is 102.

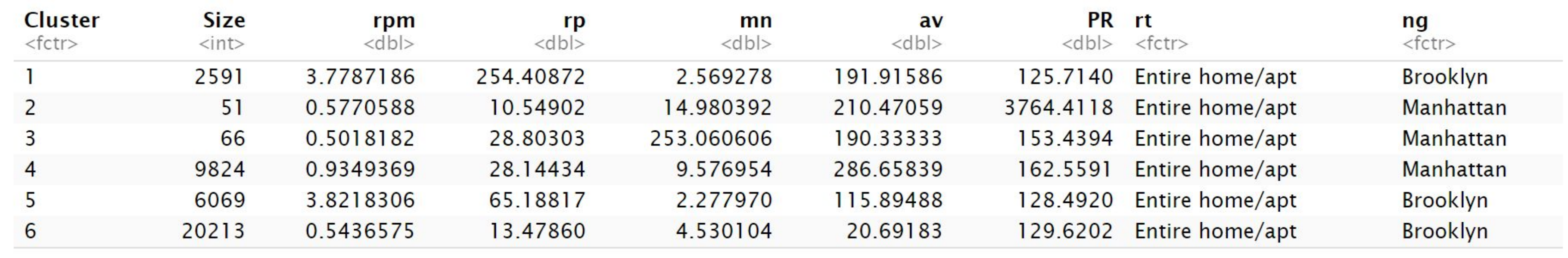
Cluster Analysis

02

Optimal Clusters

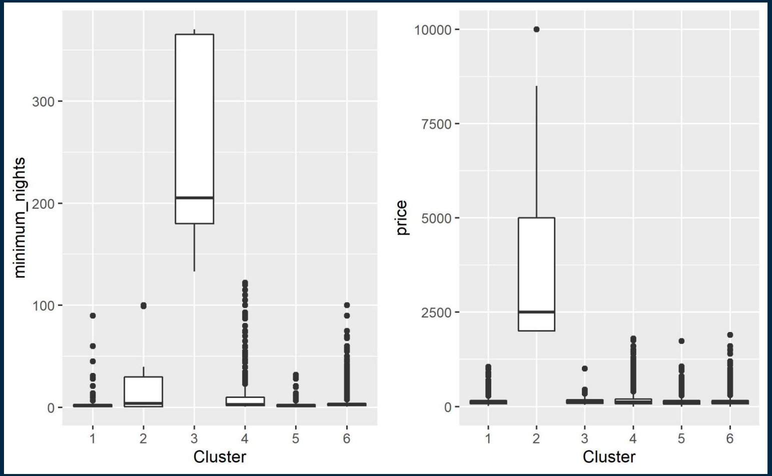


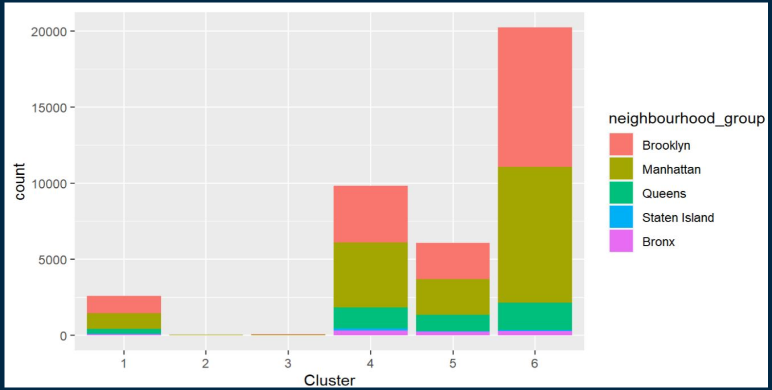
- Integrated dummy variables from room type and neighbourhood_group
- Number of clusters: 6 [depicted from the scree plot on the left]
- Hopkins Statistic **0.0184**

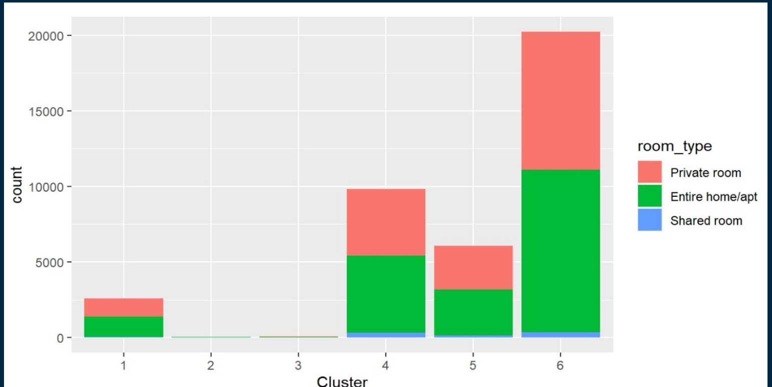


Cluster	Size	rpm	rp	mn	av	PR	rt	ng
<fctr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<fctr>	<fctr>
1	2591	3.7787186	254.40872	2.569278	191.91586	125.7140	Entire home/apt	Brooklyn
2	51	0.5770588	10.54902	14.980392	210.47059	3764.4118	Entire home/apt	Manhattan
3	66	0.5018182	28.80303	253.060606	190.33333	153.4394	Entire home/apt	Manhattan
4	9824	0.9349369	28.14434	9.576954	286.65839	162.5591	Entire home/apt	Manhattan
5	6069	3.8218306	65.18817	2.277970	115.89488	128.4920	Entire home/apt	Brooklyn
6	20213	0.5436575	13.47860	4.530104	20.69183	129.6202	Entire home/apt	Brooklyn

Further Analysis







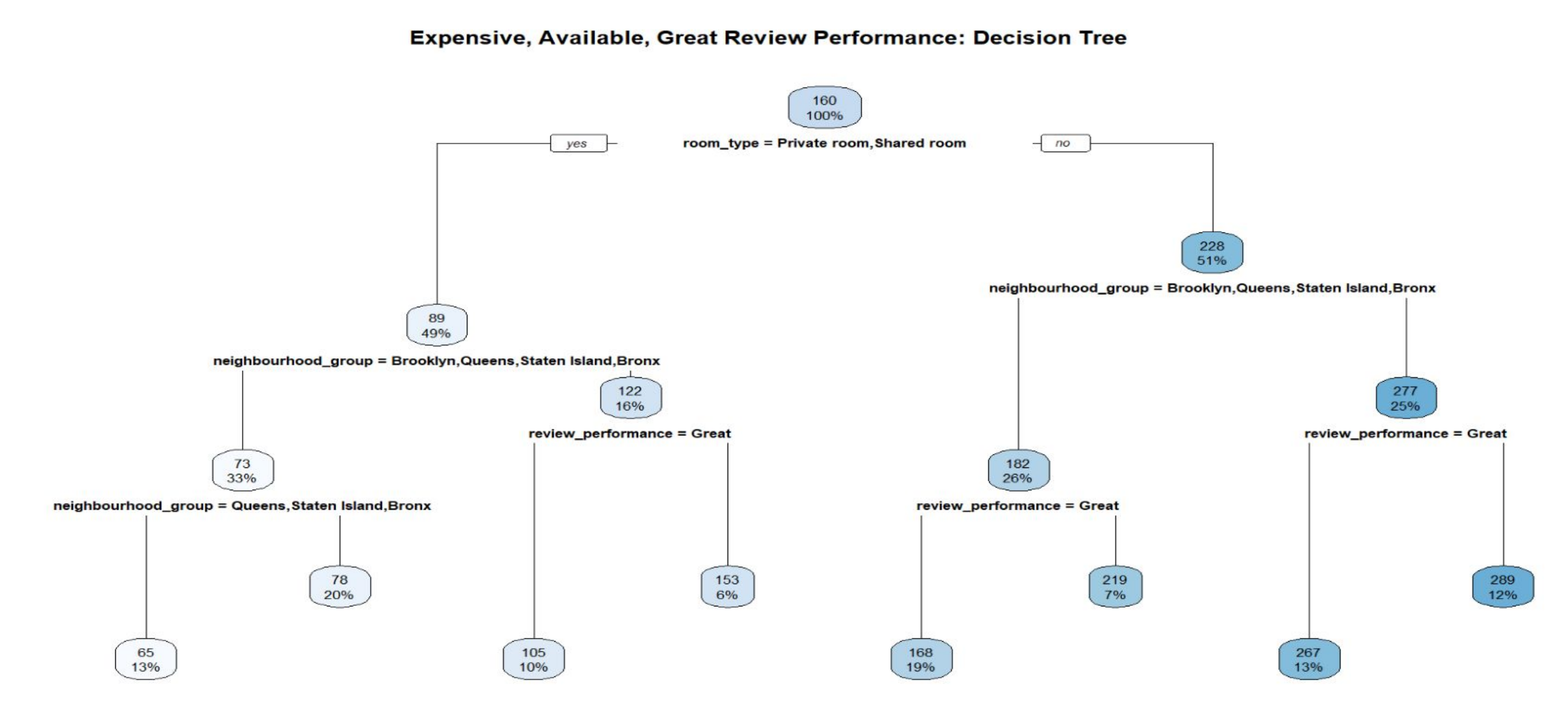
Cluster 2

Airbnbs tend to be expensive and slightly occupied for longer relatively

Cluster 3

Airbnb's tend to be occupied for longer stays

Decision Tree[Cluster 2]


Expensive, Available, Great Review Performance: Decision Tree
160
100%
yes
room_type = Private room,Shared room
no
228
51%
neighbourhood_group = Brooklyn,Queens,Staten Island,Bronx
89
49%
neighbourhood_group = Brooklyn,Queens,Staten Island,Bronx
122
16%
review_performance = Great
277
25%
review_performance = Great
73
33%
neighbourhood_group = Queens,Staten Island,Bronx
182
26%
review_performance = Great
78
20%
153
6%
219
7%
289
12%
65
13%
105
10%
168
19%
267
13%

Linear Regression Model

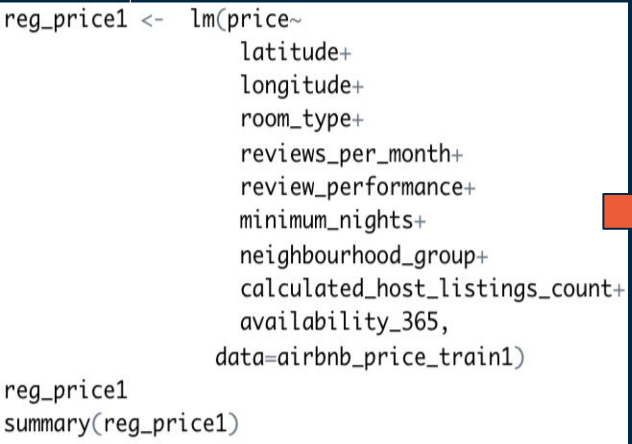
03

Linear Regression Model

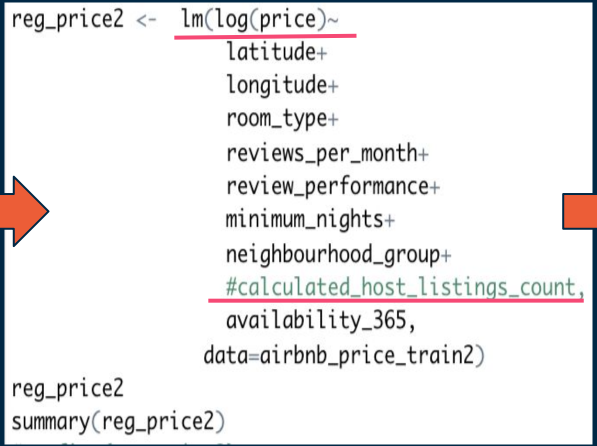
Three Regression Model Comparison

- The dataset is splitted for: 70% training and 30% testing
- We removed the price that’s equal to 0, since it is meaningless.

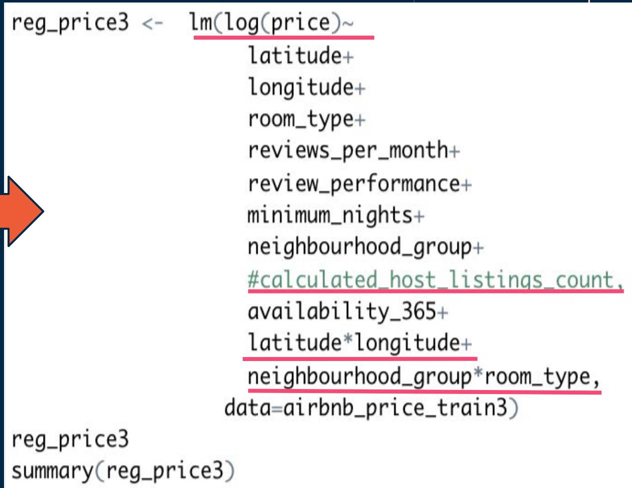
The progression of optimizing regression model:



```
reg_price1 <-  lm(price~
                    latitude+
                    longitude+
                    room_type+
                    reviews_per_month+
                    review_performance+
                    minimum_nights+
                    neighbourhood_group+
                    calculated_host_listings_count+
                    availability_365,
                  data=airbnb_price_train1)
reg_price1
summary(reg_price1)
```



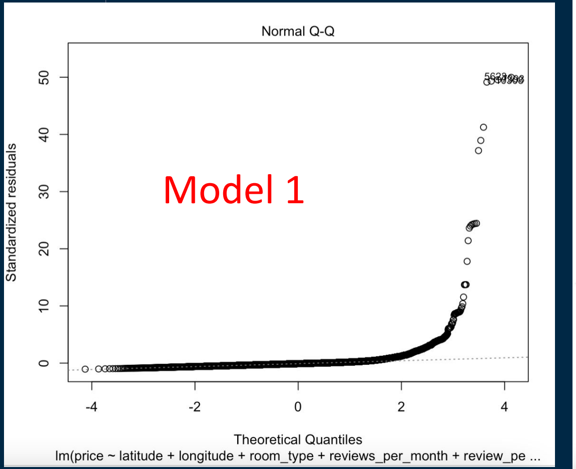
```
reg_price2 <-  lm(log(price)~
                    latitude+
                    longitude+
                    room_type+
                    reviews_per_month+
                    review_performance+
                    minimum_nights+
                    neighbourhood_group+
                    #calculated_host_listings_count,
                    availability_365,
                  data=airbnb_price_train2)
reg_price2
summary(reg_price2)
```

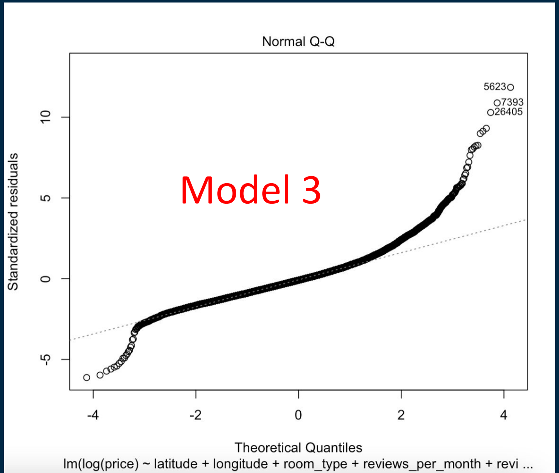


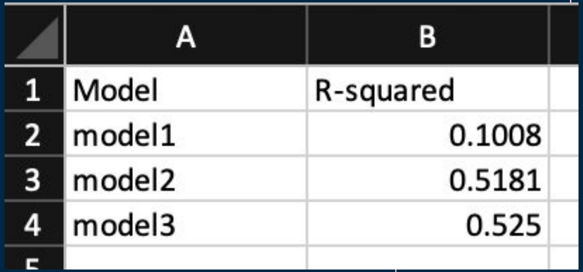
```
reg_price3 <-  lm(log(price)~
                    latitude+
                    longitude+
                    room_type+
                    reviews_per_month+
                    review_performance+
                    minimum_nights+
                    neighbourhood_group+
                    #calculated_host_listings_count,
                    availability_365+
                    latitude*longitude+
                    neighbourhood_group*room_type,
                  data=airbnb_price_train3)
reg_price3
summary(reg_price3)
```

Linear Regression Model

Regression model Summary





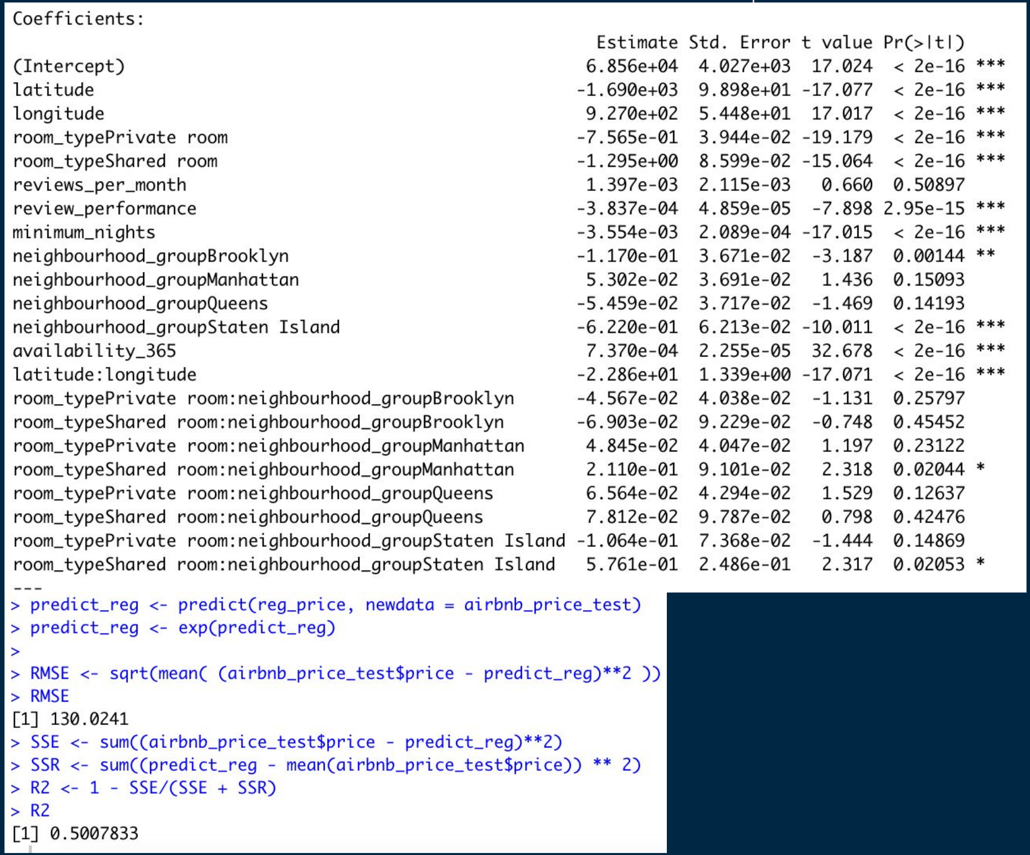


	A	B
1	Model	R-squared
2	model1	0.1008
3	model2	0.5181
4	model3	0.525

- Based on the result, we can see that Model 3 has improved a lot from Model 1.
- The QQ plot shows that the model 1's data hardly increase as the quantiles increase; whereas, the model 3 for the most part, follows an increasing trend except for the extreme values at the beginning and at the end of the line.

Linear Regression Model

Regression model Summary



```
Coefficients:
                                                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                          6.856e+04  4.027e+03  17.024  < 2e-16 ***
latitude                                            -1.690e+03  9.898e+01 -17.077  < 2e-16 ***
longitude                                            9.270e+02  5.448e+01  17.017  < 2e-16 ***
room_typePrivate room                               -7.565e-01  3.944e-02 -19.179  < 2e-16 ***
room_typeShared room                                -1.295e+00  8.599e-02 -15.064  < 2e-16 ***
reviews_per_month                                    1.397e-03  2.115e-03   0.660  0.50897    
review_performance                                  -3.837e-04  4.859e-05  -7.898 2.95e-15 ***
minimum_nights                                      -3.554e-03  2.089e-04 -17.015  < 2e-16 ***
neighbourhood_groupBrooklyn                         -1.170e-01  3.671e-02  -3.187  0.00144 ** 
neighbourhood_groupManhattan                         5.302e-02  3.691e-02   1.436  0.15093    
neighbourhood_groupQueens                           -5.459e-02  3.717e-02  -1.469  0.14193    
neighbourhood_groupStaten Island                    -6.220e-01  6.213e-02 -10.011  < 2e-16 ***
availability_365                                     7.370e-04  2.255e-05  32.678  < 2e-16 ***
latitude:longitude                                  -2.286e+01  1.339e+00 -17.071  < 2e-16 ***
room_typePrivate room:neighbourhood_groupBrooklyn   -4.567e-02  4.038e-02  -1.131  0.25797    
room_typeShared room:neighbourhood_groupBrooklyn    -6.903e-02  9.229e-02  -0.748  0.45452    
room_typePrivate room:neighbourhood_groupManhattan   4.845e-02  4.047e-02   1.197  0.23122    
room_typeShared room:neighbourhood_groupManhattan    2.110e-01  9.101e-02   2.318  0.02044 *  
room_typePrivate room:neighbourhood_groupQueens      6.564e-02  4.294e-02   1.529  0.12637    
room_typeShared room:neighbourhood_groupQueens       7.812e-02  9.787e-02   0.798  0.42476    
room_typePrivate room:neighbourhood_groupStaten Island -1.064e-01  7.368e-02  -1.444  0.14869    
room_typeShared room:neighbourhood_groupStaten Island  5.761e-01  2.486e-01   2.317  0.02053 *  
---
> predict_reg <- predict(reg_price, newdata = airbnb_price_test)
> predict_reg <- exp(predict_reg)
> 
> RMSE <- sqrt(mean( (airbnb_price_test$price - predict_reg)**2 ))
> RMSE
[1] 130.0241
> SSE <- sum((airbnb_price_test$price - predict_reg)**2)
> SSR <- sum((predict_reg - mean(airbnb_price_test$price)) ** 2)
> R2 <- 1 - SSE/(SSE + SSR)
> R2
[1] 0.5007833
```

- Looking at the coefficient chart, we can observe that the variable such as longitude, latitude and Shared_room in room_type are overall more important to the model.
- For our prediction, we got the result of RMSE(Root Mean Square Error): 130.02 and R2: 0.5, which are both acceptable

Conclusion

- Regression tree analysis shows that room_type, longitude, host_listings_count play important roles in the New York City airbnb price.
- Cluster analysis shows that clear relationships between price to location and room type and minimum number of nights to location and room type.
- In linear regression model, the dataset shows little to none relationship between price and the rest of the parameters, but has reasonable correlation to the data once we take the log of the price.

The price tends to decrease when the room type is not “Entire home/apt”, and when a listing located towards North or East of New York.

Thank You