Recommendation system for Airbnb in New York City

STAT 456 Final Project

Code name: Jordan

December 03, 2019

Contents

1 Abstract

In this project, method to build a recommendation system using principal components analysis and clustering analysis will be provided. Recommendation system is used by multiple IT companies to make personalized recommendation for their customers and help them to make a decision for their products. Lodging company, such as, Airbnb and Trivago are overlooking power of the recommendation system, therefore, the recommendation system for this project will be built based on Airbnb data in New York city. Recommendation system will be powered by clustering analysis, including k-means clustering and model-based clustering will be performed on two different plots. One plot will be generated with coordinate data of the accomodation and other being generated with principal components of different datasets. While making plots based on principal components, analysis will be done to obtain deeper understanding of the dataset.

2 Introduction

2.1 Background

Recommendation system is one of the most popular and powerful engines that is being used for the internet products. Netflix, Youtube, Amazon and other IT companies use recommendation system to make personalized recommendation for their customers and help them to make a decision for their products. While multiple companies use this machine learning based system, when it comes to booking accomadation for traveling recommendation system is hard to be found. Lodging company, such as, Airbnb and Trivago does not provide any recommendation based on customers' booking history. Everytime customers plan on a new trip, customers need to rely on filter and compare their options. While I was traveling to New York City earlier this year, I have spent quite a long time to choose the perfect accomodation for me. To stop this from happening again when I visit New York city next time, I decided to make a recommendation system using methods that I have learned from this course.

Data to make recommendation system for New york city was obtained through insideairbnb.com, this website provides data of airbnb accomodation from all over the world at different time of the year. From this website, I obtained dataset of New york city from July, 2019. To get more useful data to book accomodation for vacation trip, I chose data from July. Total of 15 variables were chosen from this datset with approximately 50,000 observations. The list of variables and their descriptions are described in Table 1.

Table 1. List of Variables in the Dataset

Variable names	Categorical/Quantitative	Details
ID	Categorical	Housing ID
Name	Categorical	Name of the housing
Host Name	Categorical	Host name
Neighborhood Group	Categorical	Neighboorhood such as Brooklyn
Neighborhood	Categorical	Specific town such as Williamsburg
Latitude	Quantitative	Latitude of the place
Longitude	Quantitative	Longitude of the place
Room type	Categorical	Type of housing such as private room or shared room
Price	Quantitative	Price of the housing in dollars
Minimum nights	Quantitative	Amount of nights required to book this housing
Number of reviews	Quantitative	Number of reviews for the housing
Reviews per month	Quantitative	Number of reviews per month
Calculated host listings	Quantitative	Number of housing this host has
Availability	Quantitative	Number of days this housing is available in year
Ratings	Quantitative	Ratings on the scale of 100

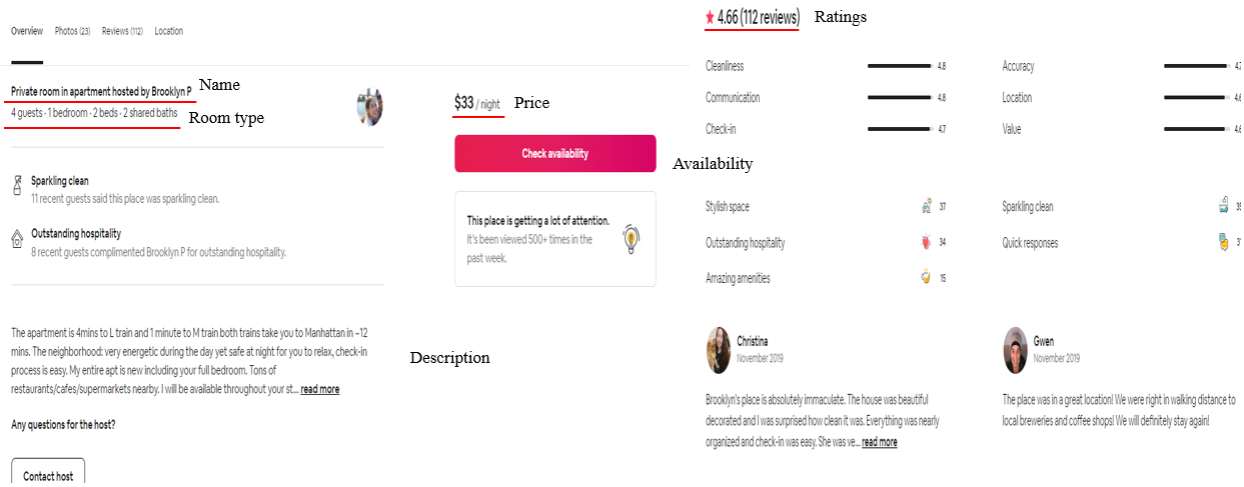

Figure 1: Airbnb example

Figure 2: NYC Map

2.2 Data cleaning & Sampling

Total of 15 variables were chosen based on how well it would represent the accomodation. When people are booking place to stay for their trip there are certain aspects that they would consider. First, the location of the accomodation, since we are considering people that are visiting the city to travel, customers would consider somewhere close to subway or landmarks of the city (Latitude, Longitude, Neighborhood). Second, the price of the accomadation, of course customers would consider cheapest one among the similar options they have (Price). Third, the ratings and review of the accomodation, especially for Airbnb, since the accomodation is not owned by company but normal people in the city, customers would consider if the place they are booking is reliable and if the place is good enough to stay (Ratings, Reviews). Fourth, availabilty of the accomoadation, of course the place should be available the date when customers are visiting the city (Availability, Minimum nights, Host listings). Fifth, what type of place they are staying, customers would consider the information of the place they are staying and decide if they like it or not (Room type, Name).

Since the dataset is very dirty and not have been cleaned at all by Airbnb I cleansed the data before I performed the analysis. First, I deleted all the data points that contains missing values, which will also help to reduce the dataset at the same time. Secondly, I deleted all the outliers, such as, ratings that are higher than 100, price that is higher than $1000 (since this is recommendation system for traveling) or that is $0, housing that has 0 availability, and data that contains different type of data. By doing this preprocessing the dataset was reduced to approximately 25,000 observations.

Then, to get more visible and computationally efficient data, I perfromed stratified random sampling based on the neighborhood group. Same ratio of neighborhood group (Bronx, Brooklyn, Manhattan, Queens, Staten Island) were sampled from raw data. The size of the sampled data is 1000.

3 Goals

- Explore the datasets for deeper understandings

  After obtaining clean and compact dataset, I will explore quantitavie variables through scatterplot and bivariate boxplots. Additionally, unique plot like starplot can be used to exlore the dataset. For, categorical variables I will create a word cloud to get the gist of what aspects are hosts selling to the customers. Also, by using coordinate date map of New york city can be expressed using data points.

- Reduce dimension and describe the dataset with principal components analysis

  Then, after exploring datasets and getting a deeper understanding of the data, we can divide into different cases using different variables. With these different cases we can use principal components analysis to express our data into 2 dimensional plots. Scree plots can be used to choose how many principal components should be used and biplots can be used to understand better about the plots using principal components.

- Build clustering models to group similar accomodations

  After expressing the dataset with principal components and plotting in 2 dimensions, I will build clustering models using two differnet methods, k-means and model based clustering. I will apply different variables for different cases and get multiple clustering models.

- Make a recommendation based on booking history

  After building mutiple clustering model, I will apply the place that I stayed earlier this year in New york city which I really liked to the models, and get recommendation based on the different clustering models.

4 Result

4.1 Data exploration

The simplest way to explore data is to plot the data that we have. However, in our dataset we have two types of variables. Categorical variables and quantitative variables. Categorical variable will be explored later after exploring quantitative variables. The quantitative variables that we have are 'Minimum nights', 'Number of reviews', 'Reviews per month', 'Calculated host listings', 'Availability', 'Ratings, and coordinates of the location. Coordinate of the location will be dealt later on by mapping them.

To make the plot look better I will divide these quantitative variables into two groups. According to what variables are the most important aspects on booking the accomodation. I have interviewed several friends, what aspects do they consider the most among qunantitative variables that I have when it comes to booking accomodation for their trips. The answer from all my friends were consistent, with them being 'Price', 'Ratings' and 'Reviews'. Therefore, I divided the quantitative variables into two groups, then plotted their scatterplot with bivariate boxplots. Find Table 2 to see these groups.

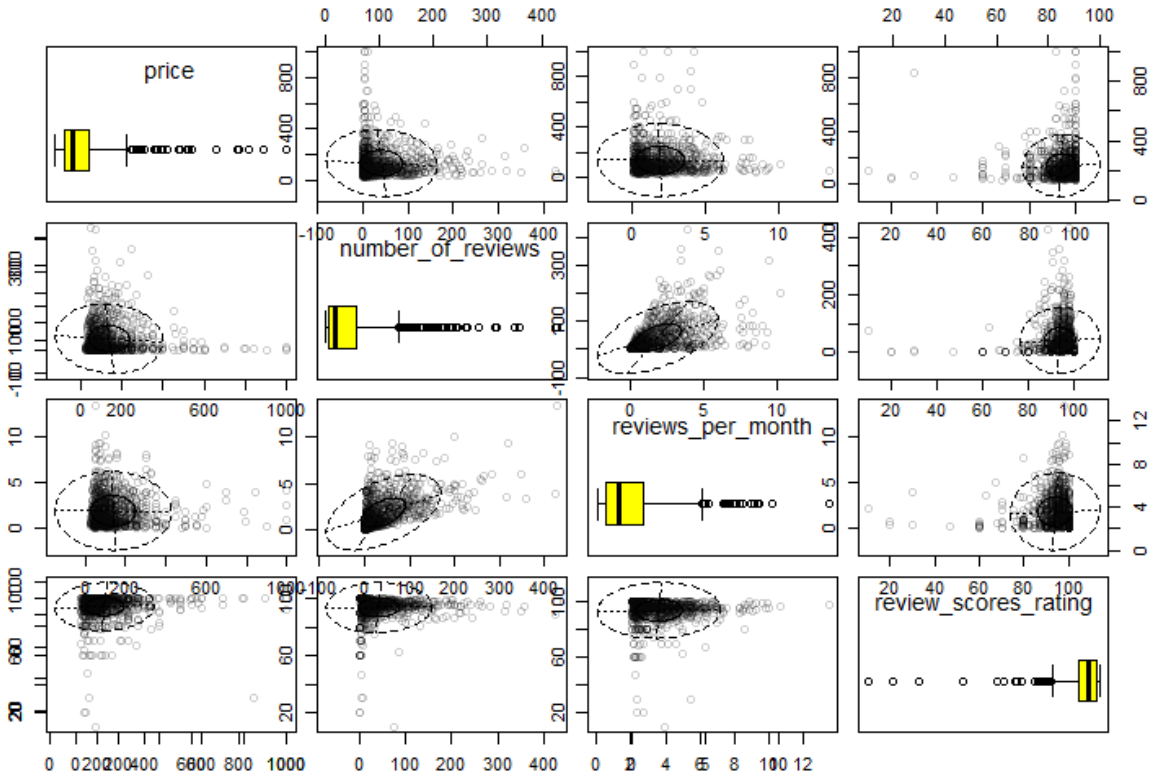

Figure 3: Scatterplot 1

Looking at this plot in Figure 3, first we can identify that there are clear minimum values and maximum value for the variables. For example, maximum value for the rating is 100, and minimum value for the price is $0. Also, since the accomodations are tended to have similar settings to compete with each other, data points are fairly condensed. Other than that we can also identify that 'Number of reviews' and 'Reviews per month' are linearly correlated, 'Ratings' variable not having significant correlation with other variables. From this we can conclude that

Airbnb data is unique dataset, sicnce customers can have different experience even on the same accomodation. We can also see that there are some outliers in all bivariate boxplots, however, I will keep these outliers, since this kind of Airbnb data is unique in every single one of them, therefore, they might carry important information.

Table 2. Groups of quantitative variables

Group 1	Group 2
Price	Minimum nights
Number of reviews	Calculated listings
Reviews per month	Availabilty
Ratings	

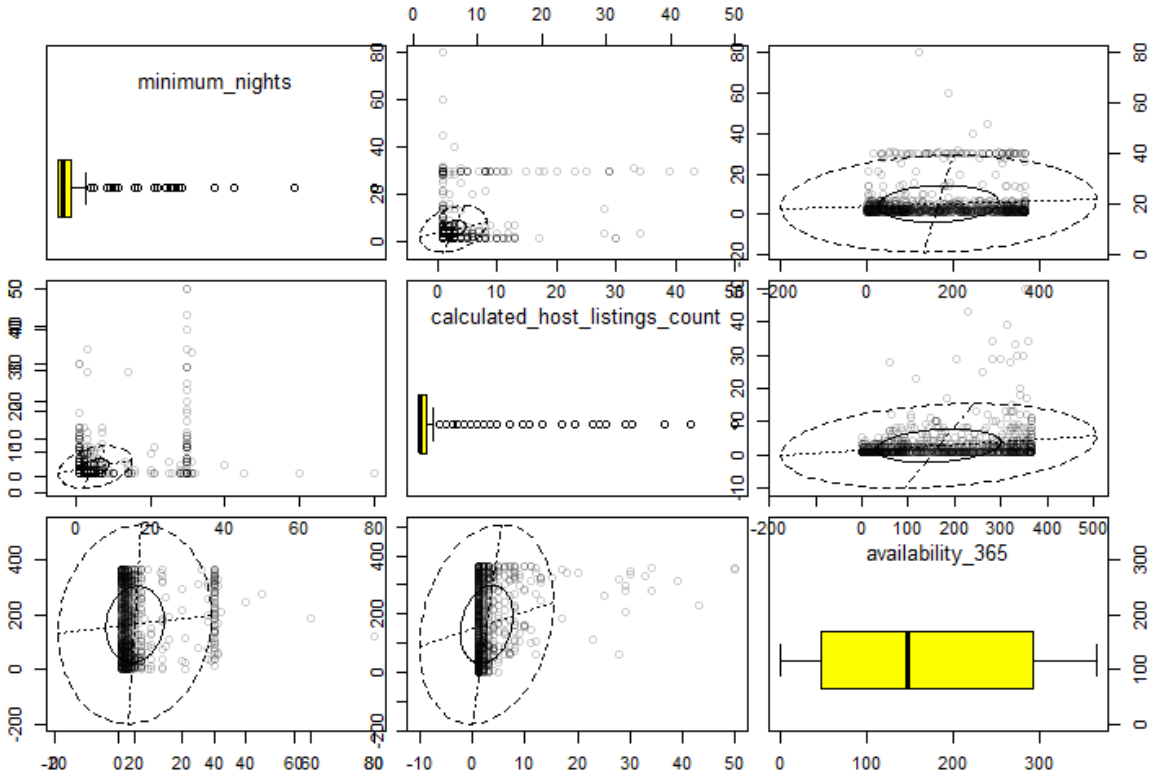

Figure 4: Scatterplot 2

Figure 4 is the scatterplot for Group 2. By looking at these variables in group 2 we can tell that these variables are related to the availability of the accomodations. However from the scatterplot, we can not identify any relationship between the variables since most of the accomodations' 'Minimum night' and 'Calculated listings' values are low. But, these variables contains important information about the availability of the accomodations, therefore, it should not be overlooked.

To find hidden information in the dataset and get different point of view to look at the dataset, we can use some of the unique plots. Star plot is one of the unique plots that could be considered. By using star plot we can identify some unique data points.

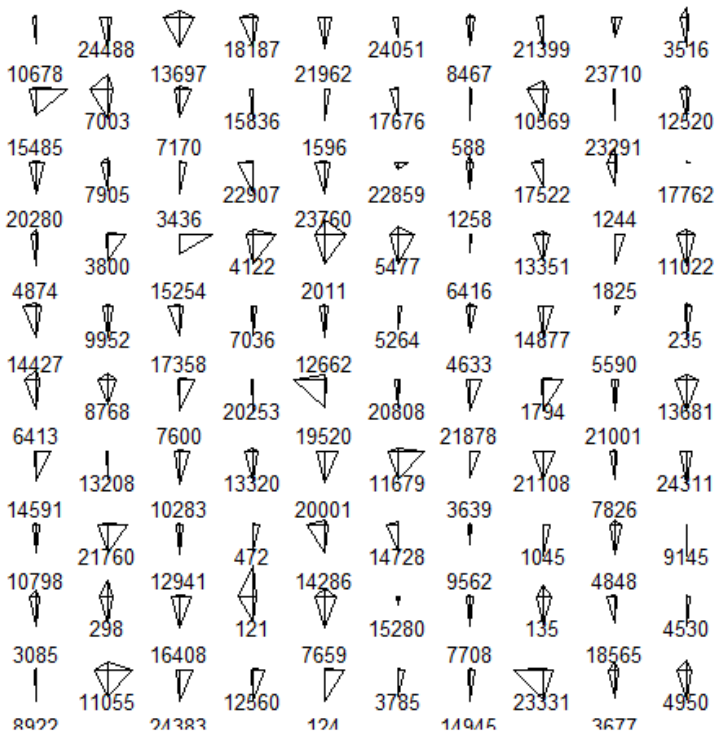

Figure 5: Starplot

Figure 5 is the starplot based on four variables from Group 1 in Table 2. First 100 data points are selected to make the plot visible. For this star plot, +x indicates 'Price' variable, -x indicates 'Number of reviews' variable, +y indicates 'Reviews per month variable' and -y indicates 'Ratings' variable. From this plot we can identify some unique data points, for example, '121'(2nd last tow, 4th column) seems like the most valuable accomodation with being large on -y, +y and -x and short on +x. This means that those data will be cheap, have high ratings and have many reviews.

For categorical variable like, name of the accomodation, I will create word cloud using raw data, since word clouds tend to look better with sufficient amount of data.

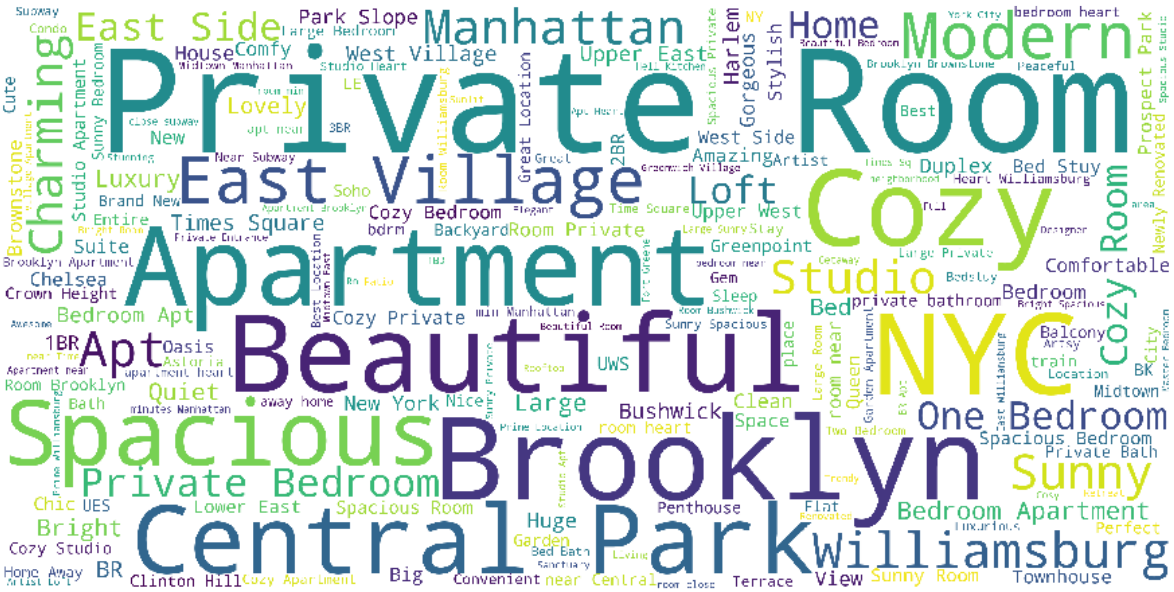

Figure 6: Word Cloud based on name

From Figure 6, the word cloud that is based on the name of the accomodation, we can identify big words like, 'Private Room', 'Brooklyn', 'Apartment', 'Beautiful', 'Central Park', 'Cozy', 'East Village', and 'Spacious'. Name of the accomodation is one of the most important aspects and first thing customers would know about the place. According to this word cloud, host are selling their place by adding the location and room type of the place.

To check if these two categorical variables are actually important in our dataset, we can check the mean difference of the ratings between differnet levels of these variables. From our dataset, we can conclude that there are significant difference between differnet room type and different neighborhood group.

Table 3. Average ratings for different room type

Room type	Average ratings
Entire home/apt	94.77
Private room	93.14
Shared room	90

Table 4. Average ratings for different neighborhood

Neighborhood	Average ratings
Bronx	92.03
Brooklyn	94.41
Manhattan	93.79
Queens	92.84
Staten Island	95.5

Table 3 and Table 4 indicate that using the room type variable and the location variable is important and should not be overlooked. However, these two variables are categorical variable so it is not able to use it for principal components analysis or clustering. Therefore, I will change theses two categorical variables to quantitative variable by using one hot encoding(dummy variables).

Figure 7 is a map of New york city created with location coordinates of the accomodation in our dataset.

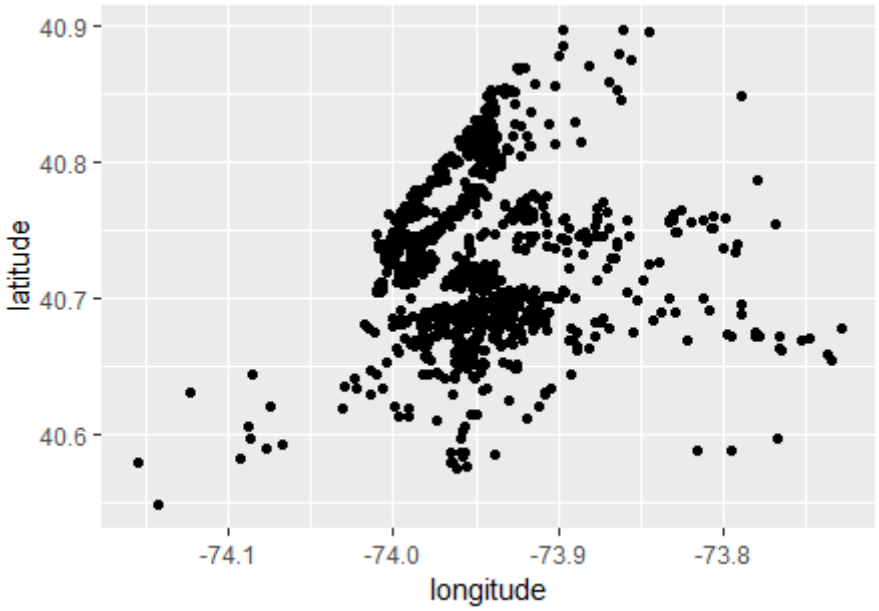

Figure 7: NYC map based on coordinate data

From this map that is expressed with coordinate data points, we can easily identify central park which is empty place in Manhattan area, and East river which runs through Brooklyn and Manhattan. We can also easily identify Staten island which is bottom left part of the map.

4.2 Principal Components Analysis

From data exploration we learned that, location(neighborhood groups) variable and room type variable are important aspects that need to be considered. Also, 'Price', 'Ratings', 'Reviews' variables are most important variables that need to be considered too. Therefore, I will divide the dataset into three groups to perform principal components analysis. Since 'Number of reviews' and 'Reviews per month' variable is positively correlated, I will just use 'Reviews per month' variable and denote it as best variables. Find Table 5 to see these dataset groups.

Table 5. Different datasets used for analysis

Dataset 1	Dataset 2	Dataset 3
Price	Price	Price
Ratings	Ratings	Ratings
Review per month	Reviews per month	Reviews per month
Number of Reviews	Location Dummy	Room type Dummy
Calculated listings		
Availability		
Minimum nights		
Location Dummy		
Room type Dummy		

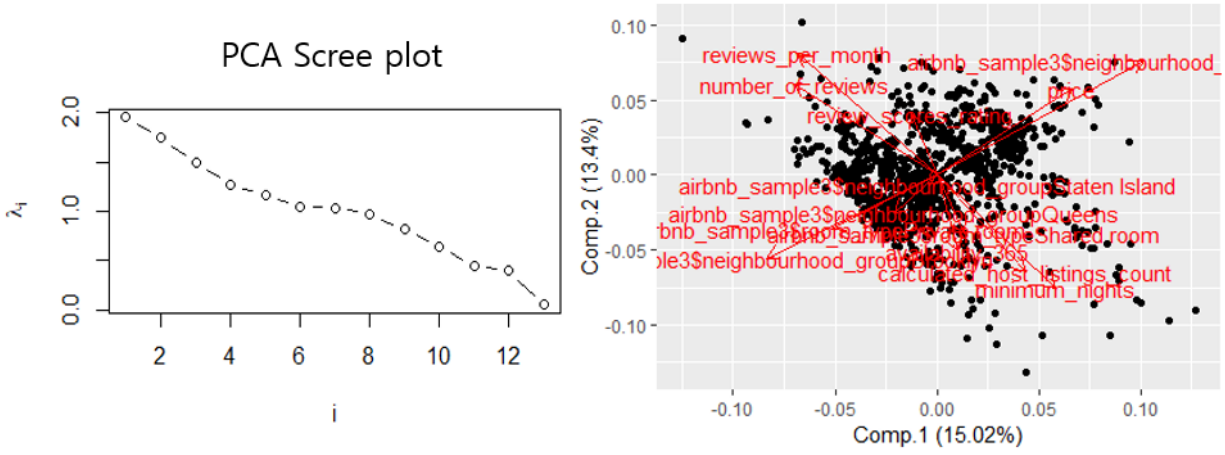

Figure 8: PCA using all variables

Figure 8 includes the scree plot for PCA and PCA biplot of PC1 and PC2 using first dataset. From scree plot we can see that since there are a lot of variables used PC1 and PC2 only represents few variance of the whole dataset. From PCA biplot, we can see how the plot is plotted based on these variables. From this we can see that 'Manhattan' and 'Price' variable is in the same direction, which indicates that accomodation in Manhattan is highly dependent on the price of the accomodation. We can also see that 'Queens', 'Brooklyn', 'Staten Island' and 'Private room' is in the same direction, which indicates that accomadation in these neighborhood has similar attributes with a lot of them having room type of 'private room. Other than that 'Availability' varaibles and 'Shared room' is in the same direction which indicates there are some relation between these two variables.

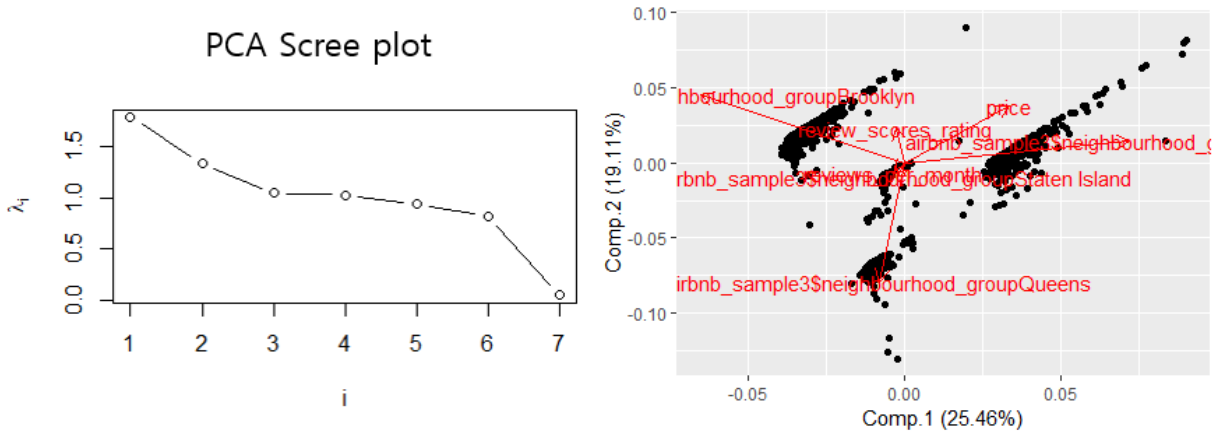

Figure 9: PCA using Location and best variables

Figure 9 includes the scree plot for PCA and PCA biplot of PC1 and PC2 using second dataset. From scree plot we can see that variance that PC1 and PC2 represents increased compared to first dataset since there are less variables. From PCA biplot, we can see how the plot is plotted based on these variables. First thing we can see is that it is divided into three disticnt group each highly affected by different locations. We can also see that 'Manhattan' and 'Price' variable is in the same direction like the biplot from the first dataset. Other than that we can see that 'Ratings' and 'Reviews per month' variables have their own direction which means that 'Ratings' and 'Reviews per month' are equally important for the different groups.

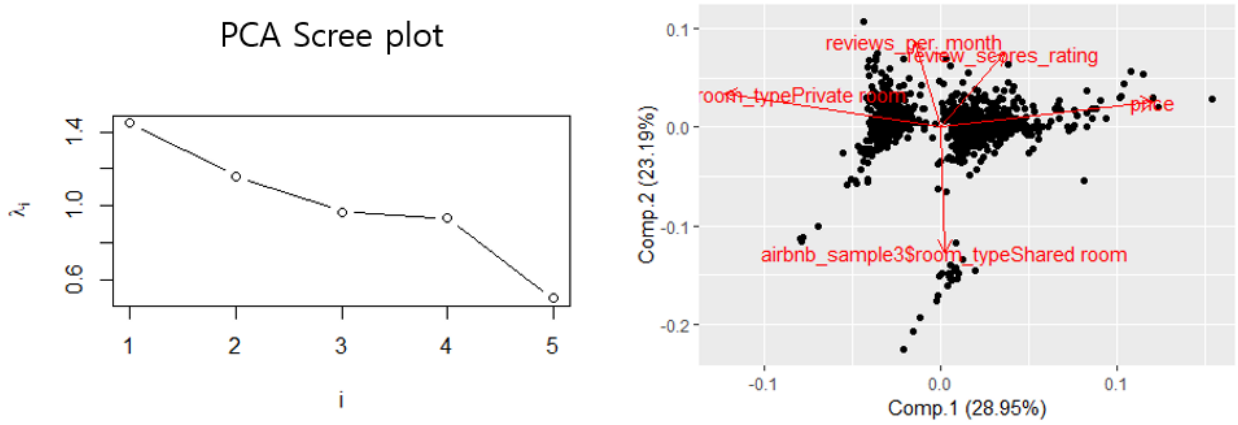

Figure 10: PCA using Room type and best variables

Figure 10 includes the scree plot for PCA and PCA biplot of PC1 and PC2 using third dataset. From scree plot we can see that variance that PC1 and PC2 represents increased being total of 53%, higher compared to first and second dataset since there are less variables. From PCA biplot, we can see how the plot is plotted based on these variables. First thing we can see is that it is divided into three disticnt group each highly affected by different room types. We can also see that 'Entire home/apt' and 'Price' variable is in the same direction which indicates booking entire home and apt can be correlated. Other than that we can see that 'Ratings' and 'Reviews per month' variables have their own direction which means that 'Ratings' and 'Reviews per month' are equally important for the different groups just like in second dataset.

4.3 Clustering

Now since we have PCA plot and also the map of new york city, I will perform two different cluster methods, which are k-means clustering and model-based clustering on two different plots, PCA plot and the map. By plotting on two different plot it will provide different point of view to look at the each clustering method.

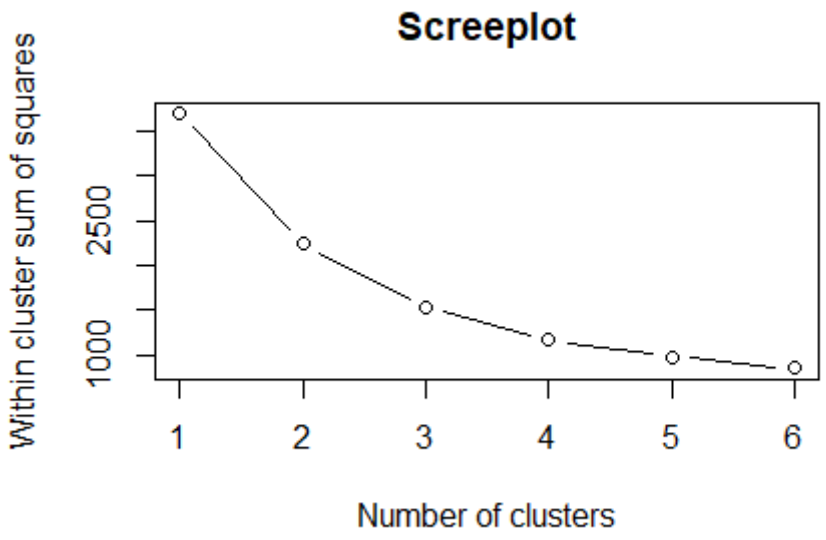

Figure 11: Scree plot for k-means using all variables

Figure 11 is the scree plot for the k-means clustering on first datset. From this scree plot we can find the elbow point on 4, therefore, we can conclude that optimal setting for hyperparameter for first dataset k is 4.

Figure 12: K-means clustering using all variables

Figure 12 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots k-means clustering(k=4) is performed with each cluster having different colors of data points on first dataset. From PCA plot, it is well clustered among condensed data points

and clusters are easy to identify. 'Manhattan' and 'Price' variables are clustered together as cluster 1 with red color. 'Availability' and 'Shared room' are clustered togehter as cluster 2 with green color. 'Queens', 'Brooklyn', 'Staten Island' and 'Private room' are clustered together as cluster 3 with blue clolor. 'Reviews' and 'Ratings' are clustered together as cluster 4 with purple color. All as estimated from PCA biplots. From the map, we can see that Manhattan area is mostly red color with some purple color. Brooklyn, Staten Island and Queens area is mixed with blue color and purple color. From this we can conclude that Review and ratings are equally important for any neighborhood group.

Figure 13: Model-based clustering using all variables

Figure 13 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots model-based clustering(Model name: VEV) is performed with each 9 cluster having different colors of datapoints on first dataset. From PCA plot, each cluster from k-means clusters are divided into mutiple cluster, all adding up to 9 clusters. It is hard to interpret what each of these cluster indicates from PCA plot, but possibly each cluster meaning each direction of the variables. From the map, purple datapoints are condensed at Brooklyn, which can indicate that cluter 8 is 'Brooklyn' variable. we can also see that dark green, cluster 3 is condensed at Manhattan which can indicate that cluster 3 is 'Manhattan'. From this we can guess that cluster 9 might indicate 'Price' variable, since it had same direction with 'Manhattan'. We can also guess cluster 1 is 'Availability' since it is distributed in every neighborhood and having same strong direction to bottom right. Other than that since green cluster, cluster 4 is distributed all over the neighborhood, we can guess that this cluster might indicate 'Price' or 'Ratings'.

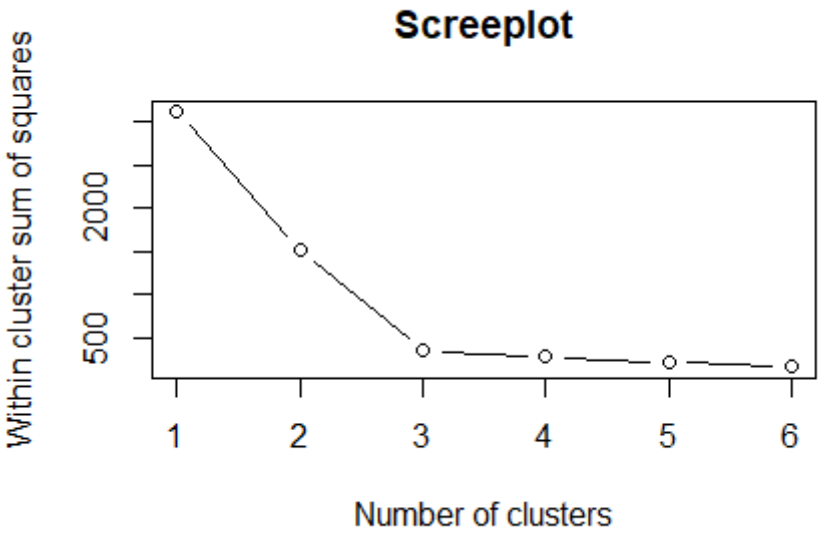

Figure 14: Scree plot for k-means using Location and best variables

Figure 14 is the scree plot for the k-means clustering on second datset. From this scree plot we can find the elbow point on 3, therefore, we can conclude that optimal setting for hyperparameter k for second dataset is 3.

Figure 15: K-means clustering using Location and best variables

Figure 15 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots k-means clustering(k=3) is performed with each cluster having different colors of data points on second dataset. From PCA plot, it is well clustered among three distinct neighborhood group which is easy to identify. We can say that Manhattan area is clustered as cluster 3 with blue color. Brookyln area is clustered as cluster 1 wtih red color. Queens and Staten Island clustered as cluster 2 with green color. From the map we can see that Bronx area has mostly red data points, therefore, Bronx having similar attribute as Brooklyn.

Figure 16 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots model-based clustering(Model name: VVV) is performed with each 9 cluster having different colors of datapoints on second dataset. From PCA plot, each cluster from

Figure 16: Model-based clustering using Location and best variables

k-means clusters are divided into mutiple cluster, all adding up to 9 clusters. We can tell that Manhattan cluster from k-means is now divided into three groups and Brooklyn cluster from k-means is now divided into two groups. Queens cluster from k-means is also divided into three group, while Staten Island retained as one group. From the map, we can get a different conclusion from k-means clustering. We can see that Bronx is rather related to Staten Island than Brooklyn, which is plausible since these two neighborhoods are most far away from the downtown New York. We can also guess that Manhattan, Brooklyn, Queens are divided into multiple groups according to other variables like 'Price', 'Ratings' and 'Reviews per month'.

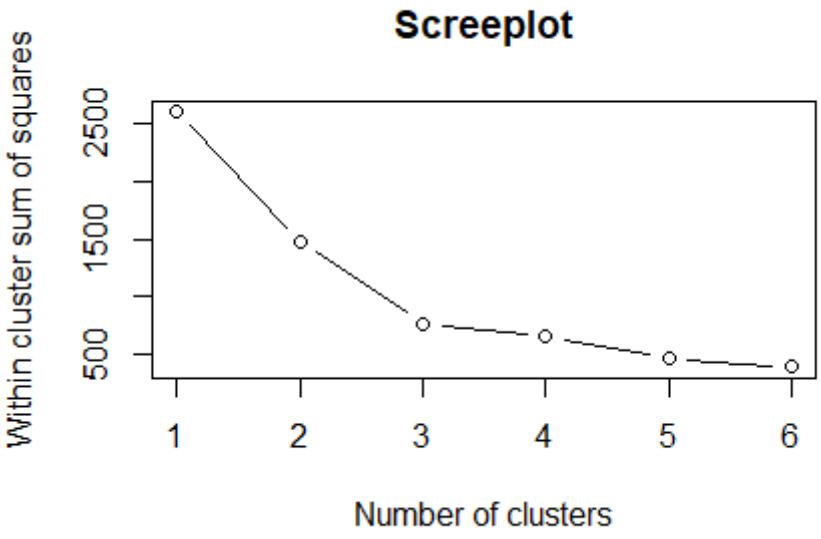

Figure 17: Scree plot for k-means using Room type and best variables

Figure 17 is the scree plot for the k-means clustering on third datset. From this scree plot we can find the elbow point on 3, therefore, we can conclude that optimal setting for hyperparameter k for third dataset is 3.

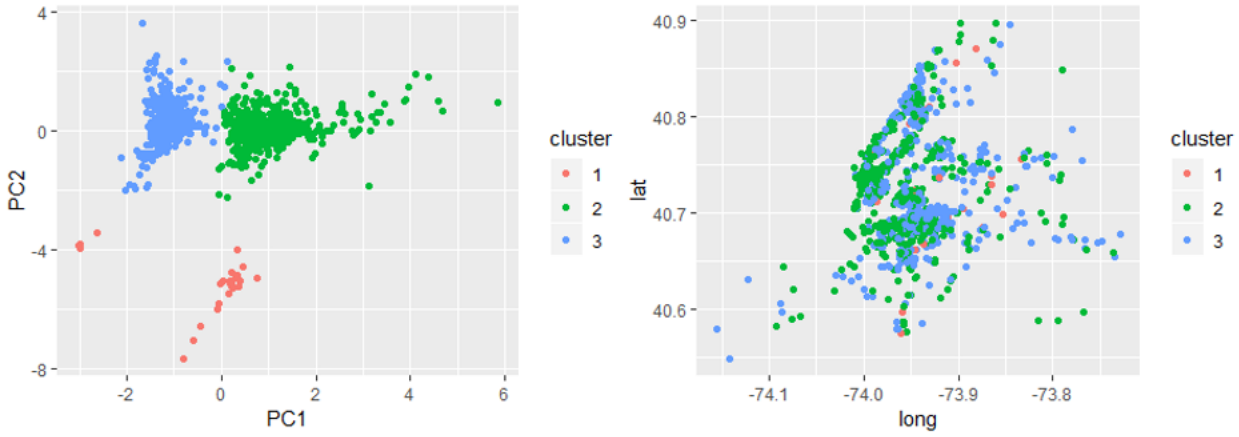

Figure 18: K-means clustering using Room type and best variables

Figure 18 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots k-means clustering(k=3) is performed with each cluster having different colors of data points on third dataset. From PCA plot, it is well clustered among three distinct room type group which is easy to identify. We can say that 'Entire home/apt' is clustered as cluster 2 with green color. 'Private room' is clustered as cluster 3 with blue color. 'Shared room' is clustered as cluster 1 with red color. From the map we can see that Shared room are mostly distributed in Brooklyn, Queens and Bronx. Manhattan having more green color, which is 'Entire home/apt' and Rest of the neighborhood having more blue color, which is 'Private room'.

Figure 19: Model-based clustering using Room type and best variables

Figure 19 includes two plots, on the left is PCA plot and on the right the map of New York city. On these plots model-based clustering(Model name: VEV) is performed with each 7 cluster having different colors of datapoints on third dataset. From PCA plot, each cluster from k-means clusters are divided into mutiple cluster, all adding up to 7 clusters. We can tell that 'Entire home/apt' cluster divided into two groups and 'Private room' cluster from k-menas is now divided into three group. While green data points, cluster 3 existing both 'Entire home/apt' and 'Private room' group. We can also identify that 'Shared room' retained as one group. From this we can conclude that 'Shared room' is the least preferred room type. From the map, we can see that brown data points, cluster 2 being more condensed in Manhattan and sky blue data points, cluter 5 being distributed all the neighborhood. From this we can guess that cluster 2, being 'Entire home/apt' variable and cluster 5, being 'Price' variable using prior knowledge that we learned. We can also guess green data points, cluster 3 being 'Ratings' and 'Reviews per month' variable.

4.4 Recommendation

To get recommendation from these clustering models, I will apply data points I reserved in the dataset, which is actual place that I have stayed earlier this year.

Name	Host name	Location	Room type	Price	Reviews per month	Ratings
Jades place	Jade	Williamsburg, Brooklyn	Private room	65	5.01	99

I absolutely loved staying at this place, since it was cheap and had reliable host with high ratings and also located in one of the main place at Brooklyn. So I want to get recommendation using my clustering model that is similar to this place. I will use my model-based clustering model(VVV) that was built with second dataset. Since I do not really care if it is Entire home or just private room, but I would want to stay at same location, Brooklyn.

By using this model the recommendation that I get is any accomodation that is in cluster 3 from model-based clustering model(VVV) that was built with second dataset. From Figure 16 it is dark green data point. Among this data point I would prefer accomodation that has reviews more than 50, ratings higher than 95, price lower than $100. Additionally this time I want to stay at entire home/apt.

Then I get the recommendation of one specific accomodation.

Name	Host name	Location	Room type	Price	Number of reviews	Ratings
2000sq 3 story townhouse	Lindsay	Boerum Hill, Brooklyn	Entire home	99	140	96

Perfect! This will be my next place that I will stay when I visit New York next time.

5 Conclusion

5.1 Summary of Findings

From this project there were some new knowledge that was obtained. First, from the data exploration we learned that location(neighborhood groups) variable and room type variable are important aspects that need to be considered. Also, 'Price', 'Ratings', 'Reviews' variables are most important variables that need to be considered too. Secondly, we learned that there are some relationship betweeen different variables through principal components analysis. For example, Manhattan area was higly related with room type of 'Entire home/apt' and 'Price'. Other than that, we learned that 'Ratings' and 'Reviews per month' are equally important for the different groups. Lastly, we could learn that Bronx is rather related to Staten Island than Brooklyn and 'Shared room' is the least preferred room type from clustering analysis. Overall, from using differnet clustering methods we were able to get good recommendation based on booking history and individual preference.

5.2 Error Analysis

There could be some possible errors that might have been overlooked thorughout this project. First, the dataset that I used to perform analysis is only based on 1000 observation. This dataset also includes multiple outliers, that might have affected the result of the analysis. Secondly, the PCA plots that was plotted for differnet dataset did not represent much of the variance of the whole dataset. This might have affected the dataset since it is only based on small ratio of the dataset. Lastly, my interpretation of the clustering analysis might possibly have some errors.

5.3 Further Studies

For further studies, we can make recommendation system for different cities, possibly using different kinds of clustering methods besides k-means and model-based clustering models and more variables. It will be interesting to see other clustering methods including deep clustering using neural network give better recomendation result. Since Airbnb does not have recommendation system at the moment, it would be fascinating if they build a recommendation system for there application similar to what I built.