Data analysis and Data visualization

for New York City Airbnb open data 2019 in R

Pengfei Ma

Boston University

CS 544

Prof. Heather

Oct.12th 2020

Abstract

As a travel lover, traveling on vacation is the most exciting thing. From eastern coast to California, from Texas to Yellowstone. Road trip and flights are two major ways to arrive the city. However, no matter which way to travel, housing need to be considered in both two travelling. Nowadays, a lot of housing websites came out, a very attractive housing application named Airbnb is very popular. Analyzing the open data from Airbnb would be really useful for the future travelling and could know a city much better.

This paper will collect all the steps of data analyzing of New York City Airbnb 2019 open data by R. Explaining the methods and collecting all the results. The steps of analyzing in this project is importing data, analyzing data, reflecting the results and conclusion. This project will use regular statistic data such as mean, standard deviation, distributions and strata in statistic to do the calculation. In addition, this project will use sum(), srswr(), inclusionprobabilities(), UPsystematic(), strata() functions in R code as the main body. “plotly” and “sampling” are the packages used in this project as well. Eventually, plotting method: barplot, plot, pie, histogram in R will be used to do the data visualization. The conclusion of the projects will be the reflection based on the analyzing results above. RStudio will be the only one IDE used for all calculation and plotting in this project.

The major columns used in this project is the “neighborhood_group”, “price”, and “number of reviews”. And the goal of this project is exploring three information, three relationships and sampling for the data. Three information that want to be obtained is:

1. How many airbnbs in each borough, which has the most and which has the least?
2. Which borough has the highest price and which has the lowest?
3. What are their proportions?

The three relationships that this project want to explore is:

1. The relationship between price and location.
2. The relationship between price and minimum nights
3. The relationship between price and its location.

Three information exploration will use statistic calculation, probability functions to gain the result. Three relationships will use probability density function (PDF), cumulative density function (CDF), distribution analysis, and central limit theorem (CLT) to concern these relationships. Finally, the application of sampling on Airbnb open data will use sample random sampling, systematic sampling, inclusion probability and strata to show the result.

Keywords: Statistic functions, Plotting, RStudio, PDF, CDF, sampling, CLT

Project map

Preparing the data

1. Picking the data set
2. Importing data

Data analyzing

I. Categorical variable analysis of the number of airbnbs in different boroughs in NYC
   1. Setting data
   2. Visualization of airbnbs in each borough
      - Barplot of the number of airbnbs in each borough in NYC
      - Pie chart for the proportion of airbnbs in each borough

II. Numerical variable analysis of the price of airbnbs in NYC
   1. Setting data
   2. Visualization for the price
      - Barplot of the frequency of the price
      - Barplots of the frequency of the price in each borough
      - Summaries of the price in each borough
      - Boxplots of the price in each borough in one diagram

III. Comparation of the relationship between price, minimum night, number of reviews
   1. Pairs diagram of the combination of the relationship of three variables
   2. Scatterplot of the relationship of price and minimum night
   3. Scatterplot for the relationship of price and number of reviews
   4. Boxplots of price, minimum nights, and number of reviews

IV. Distribution Analysis
   1. Plotting the data of number of reviews in airbnb open data
   2. Scatterplot for the probability of each number of reviews appeared
   3. Boxplot of the number of reviews
   4. Summary of the number of reviews and its standard deviation
   5. Probability density function for the number reviews as normal distribution
   6. Cumulative distribution function for the number reviews as normal distribution

V. Applicability of the Central Limit Theorem for Price

1. Setting data
2. Central Limit Theorem visualization for price
    - Histogram for CLT of sample size 10, 20, 30, 40 for price
3. Central Limit Theorem visualization for each borough
    - CLM for Manhattan
    - CLM for Brooklyn
    - CLM for Bronx
    - CLM for Queens
    - CLM for Staten Island

VI. Sampling for airbnb open data

1. Setting data
2. Sample random sampling
    - Table of 5000 random variables drew in airbnb
    - Barplot of 5000 random variables drew in airbnb
    - Table of proportion of random variables drew in each borough
    - Pie chart of the proportion of random variables drew in each borough
3. Systematic sampling
    - Table of systematic sampling with size = 5000 for each borough
    - Barplot of systematic sampling with size = 5000 for each borough
    - Table of the proportion of systematic sampling with size = 5000 for each borough
    - Pie chart of the proportion of systematic sampling with size = 5000 for each borough
4. Inclusion probability
    - Table of inclusion probability
    - Barplot of inclusion probability
    - Table of the proportion of inclusion probability of the number of airbnbs
    - Pie chart for the proportion of inclusion probability of the number of airbnbs
5. Stratified
    - Setting data
    - Table and description of stratified

Conclusion

References

Preparing the data

1. Picking the data set

Based on the choices of the data set, this project decided to choose Kaggle (*https://www.kaggle.com/datasets)* as the resource of data website. The New York City Airbnb Open Data by author *Dgomonov* which was updated in 2019 will be considered as the primary data set in this project. The AB_NYC_2019.csv file is the only data file used in this project.

2. Importing Data

Firstly, downloaded the *AB_NYC_2019.csv* file from *https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv* into one folder on desktop named CS544_project. Then created a new R script in the same folder. Then, setting the path of the R script by the code

```
setwd("~/Desktop/CS544_Project/")
```

Moreover, importing data and assigning data to name “airbnb” by

```
airbnb <- read.csv(file = 'AB_NYC_2019.csv')
```

Setting the packages (plotly, sampling) that will be used in this project

```
library(plotly)
library(sampling)
```

By using R code

```
View(airbnb)

> ncol(airbnb)
[1] 16
> nrow(airbnb)
[1] 48895
```

It could be clearly seen that there are 16 columns in airbnb and provide detailed information about the data. 16 columns are “id”, “name”, “host_id”, “host_name”, “neighborhood_group”, “neighborhood”, “latitude”, “longtitude”, “room_type”, “price”, “minimum_nights”, “number_of_reviews”, “last_review”, “reviews_per_hour”, “calculated_host_listing_count”, “availability_365”.

id	name	host_id	host_name	neighbourhood_group	neighbourhood	latitude	longitude	room_type	price	minimum_nights	number_of_reviews	last_review	reviews_per_month	calculated_host_listings_count	availability_365
2539	Clean & quiet apt home by the park	2787	John	Brooklyn	Kensington	40.64749	-73.97237	Private room	149	1	9	2018-10-19	0.21	6	365
2595	Skylit Midtown Castle	2845	Jennifer	Manhattan	Midtown	40.75362	-73.98377	Entire home/apt	225	1	45	2019-05-21	0.38	2	355
3647	THE VILLAGE OF HARLEM....NEW YORK !	4632	Elisabeth	Manhattan	Harlem	40.80902	-73.9419	Private room	150	3	0			1	365
3831	Cozy Entire Floor of Brownstone	4869	LisaRoxanne	Brooklyn	Clinton Hill	40.68514	-73.95976	Entire home/apt	89	1	270	2019-07-05	4.64	1	194
5022	Entire Apt: Spacious Studio/Loft by central park	7192	Laura	Manhattan	East Harlem	40.79851	-73.94399	Entire home/apt	80	10	9	2018-11-19	0.10	1	0
5099	Large Cozy 1 BR Apartment In Midtown East	7322	Chris	Manhattan	Murray Hill	40.74767	-73.975	Entire home/apt	200	3	74	2019-06-22	0.59	1	129
5121	BlissArtsSpace!	7356	Garon	Brooklyn	Bedford-Stuyvesant	40.68688	-73.95596	Private room	60	45	49	2017-10-05	0.40	1	0
5178	Large Furnished Room Near B'way	8967	Shunichi	Manhattan	Hell's Kitchen	40.76489	-73.98493	Private room	79	2	430	2019-06-24	3.47	1	220
5203	Cozy Clean Guest Room - Family Apt	7490	MaryEllen	Manhattan	Upper West Side	40.80178	-73.96723	Private room	79	2	118	2017-07-21	0.99	1	0
5238	Cute & Cozy Lower East Side 1 bdrm	7549	Ben	Manhattan	Chinatown	40.71344	-73.99037	Entire home/apt	150	1	160	2019-06-09	1.33	4	188
5295	Beautiful 1br on Upper West Side	7702	Lena	Manhattan	Upper West Side	40.80316	-73.96545	Entire home/apt	135	5	53	2019-06-22	0.43	1	6
5441	Central Manhattan/near Broadway	7989	Kate	Manhattan	Hell's Kitchen	40.76076	-73.98867	Private room	85	2	188	2019-06-23	1.50	1	39
5803	Lovely Room 1, Garden, Best Area, Legal rental	9744	Laurie	Brooklyn	South Slope	40.66829	-73.98779	Private room	89	4	167	2019-06-24	1.34	3	314
6021	Wonderful Guest Bedroom in Manhattan for SINGLES	11528	Claudio	Manhattan	Upper West Side	40.79826	-73.96113	Private room	85	2	113	2019-07-05	0.91	1	333
6090	West Village Nest - Superhost	11975	Alina	Manhattan	West Village	40.7353	-74.00525	Entire home/apt	120	90	27	2018-10-31	0.22	1	0
6848	Only 2 stops to Manhattan studio	15991	Allen & Irina	Brooklyn	Williamsburg	40.70837	-73.95352	Entire home/apt	140	2	148	2019-06-29	1.20	1	46
7097	Perfect for Your Parents + Garden	17571	Jane	Brooklyn	Fort Greene	40.69169	-73.97185	Entire home/apt	215	2	198	2019-06-28	1.72	1	321
7322	Chelsea Perfect	18946	Doti	Manhattan	Chelsea	40.74192	-73.99501	Private room	140	1	260	2019-07-01	2.12	1	12
7726	Hip Historic Brownstone Apartment with Backyard	20950	Adam And Charity	Brooklyn	Crown Heights	40.67592	-73.94694	Entire home/apt	99	3	53	2019-06-22	4.44	1	21
7750	Huge 2 BR Upper East Cental Park	17985	Sing	Manhattan	East Harlem	40.79685	-73.94872	Entire home/apt	190	7	0			2	249
7801	Sweet and Spacious Brooklyn Loft	21207	Chaya	Brooklyn	Williamsburg	40.71842	-73.95718	Entire home/apt	299	3	9	2011-12-28	0.07	1	0
8024	CBG CtyBGd HelpsHaiti rm#1:1-4	22486	Lisel	Brooklyn	Park Slope	40.68069	-73.97706	Private room	130	2	130	2019-07-01	1.09	6	347
8025	CBG Helps Haiti Room#2.5	22486	Lisel	Brooklyn	Park Slope	40.67989	-73.97798	Private room	80	1	39	2019-01-01	0.37	6	364
8110	CBG Helps Haiti Rm #2	22486	Lisel	Brooklyn	Park Slope	40.68001	-73.97865	Private room	110	2	71	2019-07-02	0.61	6	304
8490	MAISON DES SIRENES1,bohemian apartment	25183	Nathalie	Brooklyn	Bedford-Stuyvesant	40.68371	-73.94028	Entire home/apt	120	2	88	2019-06-19	0.73	2	233
8505	Sunny Bedroom Across Prospect Park	25326	Gregory	Brooklyn	Windsor Terrace	40.65599	-73.97519	Private room	60	1	19	2019-06-23	1.37	2	85
8700	Magnifique Suite au N de Manhattan - vue Cloitres	26394	Claude & Sophie	Manhattan	Inwood	40.86754	-73.92639	Private room	80	4	0			1	0
9357	Midtown Pied-a-terre	30193	Tommi	Manhattan	Hell's Kitchen	40.76715	-73.98533	Entire home/apt	150	10	58	2017-08-13	0.49	1	75
9518	SPACIOUS, LOVELY FURNISHED MANHATTAN BEDROOM	31374	Shon	Manhattan	Inwood	40.86482	-73.92106	Private room	44	3	108	2019-06-15	1.11	3	311
9657	Modern 1 BR / NYC / EAST VILLAGE	21904	Dana	Manhattan	East Village	40.7292	-73.98542	Entire home/apt	180	14	29	2019-04-19	0.24	1	67

The airbnb file provide the details of the airbnbs in 5 boroughs of New York City (Manhattan, Brooklyn, Bronx, Queens, Staten Island) with total 48895 airbnbs. The most attractive information in this data file is the price and number of reviews. Also, these two columns of data are the most related to a customer.

Data analyzing

I. Categorical variable analysis of the number of airbnbs in different boroughs in NYC

1. Setting data

Firstly, assigning the number of airbnbs in different boroughs with the name of boroughs in NYC to "names_groupsInNYC".

```
names_groupsInNYC <- table(airbnb$neighbourhood_group)
```

Secondly, gaining only the number of airbnbs in each borough and assign it to "numberOfAirbnbs".

```
numberOfAirbnbs <- as.numeric(table(airbnb$neighbourhood_group))
```

Finally, calculating the proportion of the number of airbnbs in the whole NYC, and assign to name "proportionEachBoroughs".

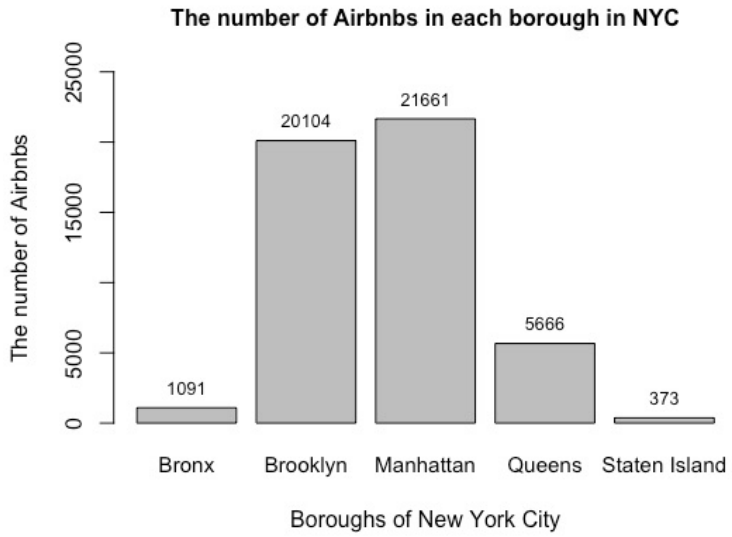
```
proportionEachBoroughs <- names_groupsInNYC / nrow(airbnb)
```

2. Visualization of airbnbs in each borough

- Barplot of the number of airbnbs in each borough in NYC

Assigning the barplot to name "group_plot_number", and added the number above the bar of the plot.

```
group_plot_number<-barplot(names_groupsInNYC, names = names(names_groupsInNYC),
                    main = "The number of Airbnbs in each borough in NYC",
                    xlab = "Boroughs of New York City",
                    ylab = "The number of Airbnbs",
                    ylim = c(0,25000), cex.names = 0.95, cex.main = 1)

text(x = group_plot_number, y = numberOfAirbnbs, label = numberOfAirbnbs,
     pos = 3, cex = 0.8)
```



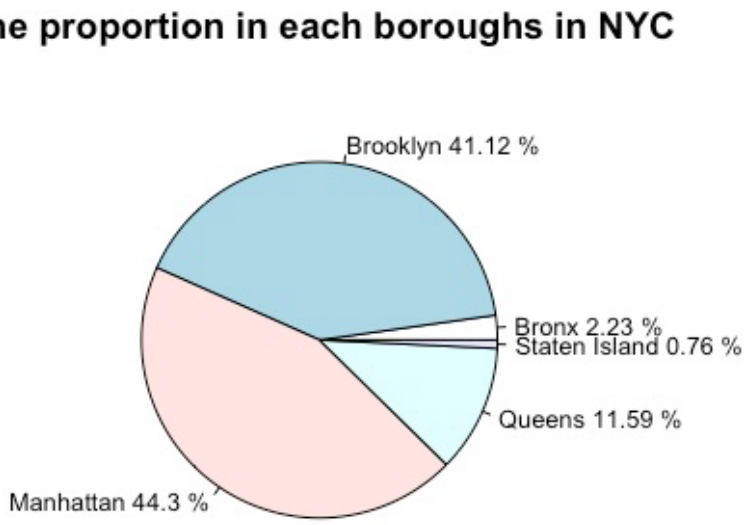
From the barplot above, it clearly shows that Manhattan has the greatest number of airbnbs in the whole NYC, which has 21661 airbnbs. Conversely, Staten Island has the least number of airbnbs, which is 373 airbnbs.

- Pie chart for the proportion of airbnbs in each borough

  At first, creating the labels of the proportion. Assigning it to name "percent_labels"

  Secondly, creating a pie chart and assign it to name "group_plot_percent"

```
percent_labels <- names(proportionEachBoroughs)
percent <- paste(round(proportionEachBoroughs,4)*100,"%")
percent_labels<-paste(percent_labels, percent)

group_plot_percent<-pie(proportionEachBoroughs, labels = percent_labels,
                        main = "The proportion in each boroughs in NYC",
                        cex = 0.8)
```



From the pie chart of the proportion of airbnbs in each borough in NYC, Manhattan occupied the most percentage of the number of airbnbs which is 44.3%. Brooklyn is NO.2 which has 41.12%. Queens has 11.59%. Bronx has 2.23%. And Staten Island has the least percentage, which is 2.23%.

II. Numerical variable analysis of the price of airbnbs in NYC

1. Setting data

Creating the subset of 5 boroughs (Manhattan, Brooklyn, Bronx, Queens, Staten Island) and assign them to names “manhattan”, “brooklyn”, “bronx”, “queens”, “staten”.

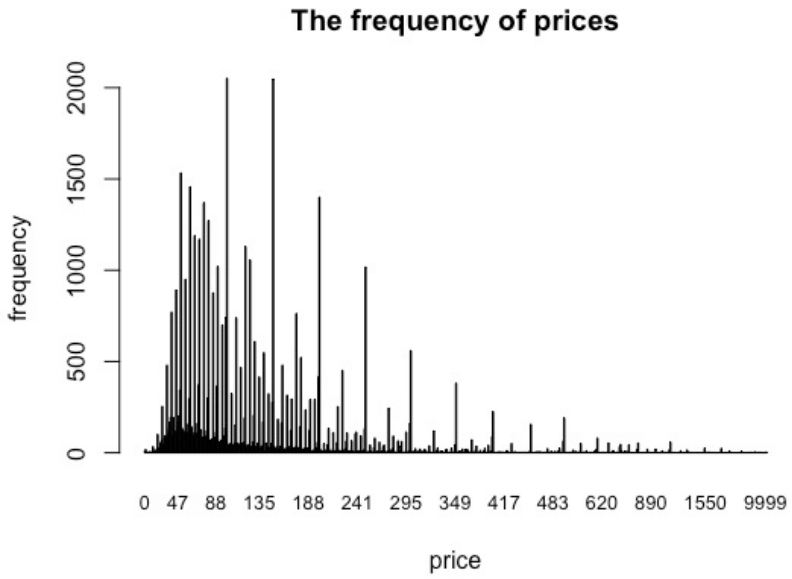
```
manhattan <- subset(airbnb, airbnb$neighbourhood_group == "Manhattan")
brooklyn <- subset(airbnb, airbnb$neighbourhood_group == "Brooklyn")
bronx <- subset(airbnb, airbnb$neighbourhood_group == "Bronx")
queens <- subset(airbnb, airbnb$neighbourhood_group == "Queens")
staten <- subset(airbnb, airbnb$neighbourhood_group == "Staten Island")
```

2. Visualization for the price

- Barplot of the frequency of the prices

By using table to get the frequency of each number in price and apply into barplot function to see the frequency of prices.

```
barplot(table(airbnb$price), main = "The frequency of prices",
      xlab = "price", ylab = "frequency", cex.names = 0.8)
```

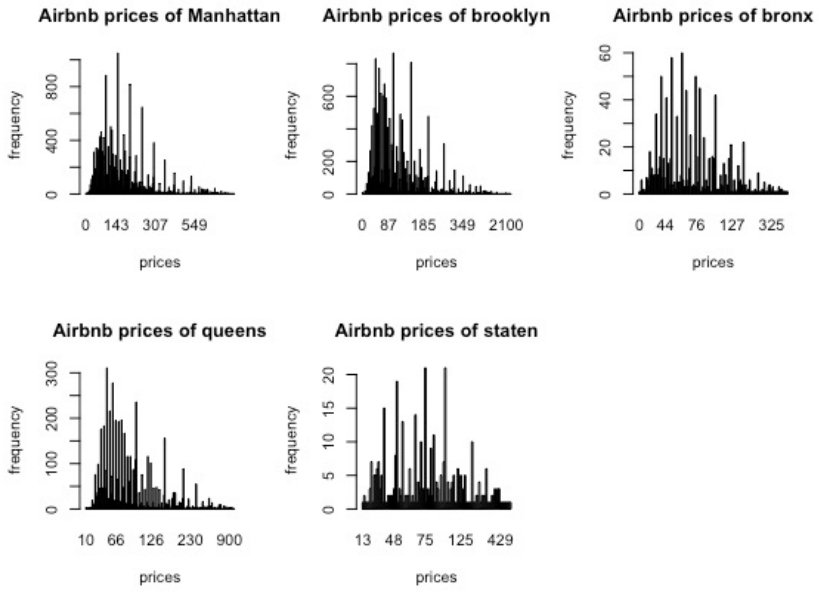


The barplot of the frequency of prices determined that most prices are concentrated on the interval 0 to 188. It could be separated to three intervals. [0, 88), [88, 188), and [188, 10000). From the diagram, the frequencies in the interval [0, 88] are higher than other intervals. This could imply that the price of most airbnbs are in [0, 88].

- Barplots of the frequency of the price in each borough

By using barplot functions to each subset. Assigning them to one diagram.

```
barplot(table(manhattan$price), main = "Airbnb prices of Manhattan", xlab = "prices", ylab = "frequency")
barplot(table(brooklyn$price), main = "Airbnb prices of brooklyn", xlab = "prices", ylab = "frequency")
barplot(table(bronx$price), main = "Airbnb prices of bronx", xlab = "prices", ylab = "frequency")
barplot(table(queens$price), main = "Airbnb prices of queens", xlab = "prices", ylab = "frequency")
barplot(table(staten$price), main = "Airbnb prices of staten", xlab = "prices", ylab = "frequency")
```



From the barplot of Manhattan, it shows that most prices located in the interval [0,143]. In the barplot of Brooklyn, most prices located in the interval [0,87]. In the barplot of Bronx, most prices are in the interval [0,76]. In the barplot of Queens, most prices are in the interval [10, 96], which 96 is the middle number of 66 and 126. In the barplot of Staten Island, most prices are in the interval [48,125]. From the combination of these five barplots, the assumption came out is that Manhattan has the highest price of airbnbs and Bronx has the lowest price. However, the information showed so far cannot provide an accurate result. Therefore, by applying summary function to these five data will provide statistic result to determine the assumption, which is more accurate and reliable.

- Summaries of the price in each borough

  Applying summary function to each price column in each subset to gain six important statistical data to prove the assumption.

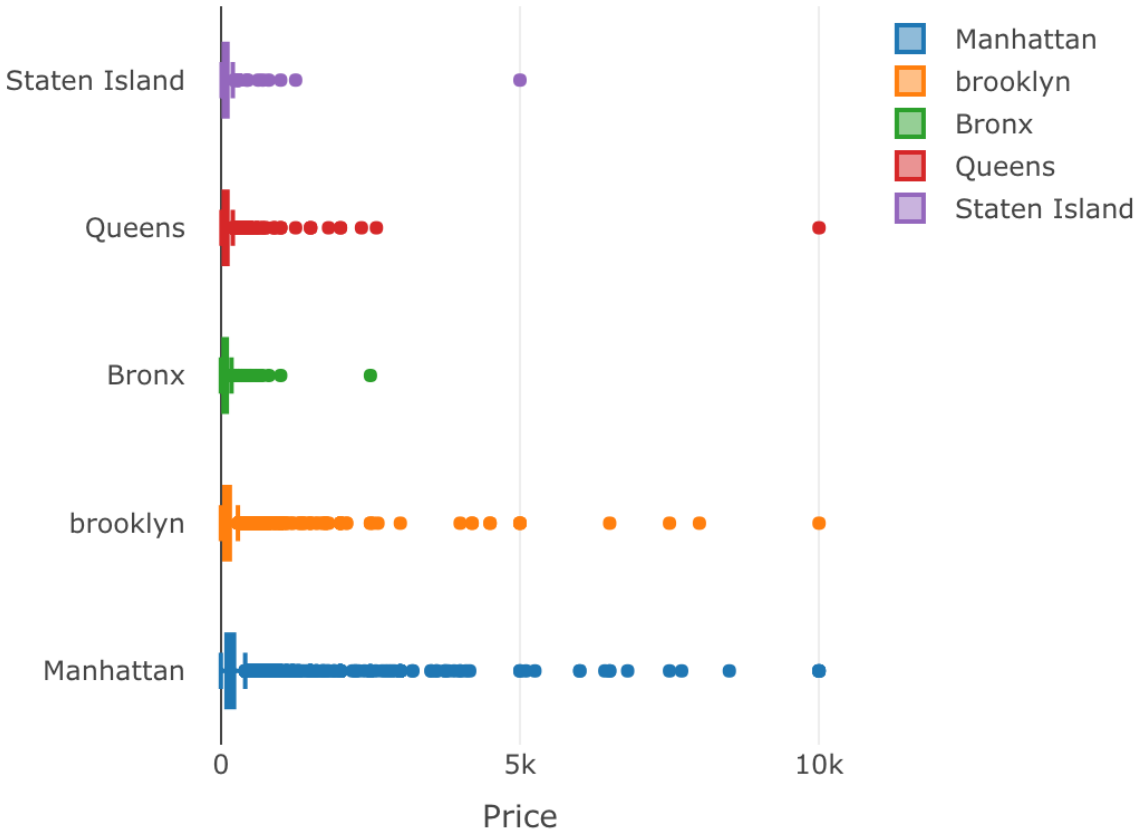
```
> summary(manhattan$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    95.0   150.0   196.9   220.0 10000.0
> summary(brooklyn$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    60.0    90.0   124.4   150.0 10000.0
> summary(bronx$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    45.0    65.0    87.5    99.0  2500.0
> summary(queens$price)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   10.00    50.00    75.00    99.52   110.00 10000.00
> summary(staten$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   13.0    50.0    75.0   114.8   110.0  5000.0
```

From five summaries above, by looking at the means of each summary. Manhattan has the highest mean, which imply that the average price of airbnb in Manhattan is the biggest. Conversely, Bronx has the lowest average price. Moreover, by looking at the median of each summary, it maintains the same results that Manhattan has the largest number and Bronx has the lowest. Based on all statistical information obtained above. The assumption was proved that Manhattan has the highest airbnb price, and Bronx has the lowest.

- Boxplots of the price in each borough in one diagram

  By using package “plotly”, do the data visualization. Combining boxplots of five columns of the data into one diagram to see their quantiles, outliers, and six statistical data.

```
NYC_price <- plot_ly(manhattan, x = ~manhattan$price, type="box", name = 'Manhattan')
NYC_price <- add_trace(NYC_price, x = ~brooklyn$price, name = 'brooklyn')
NYC_price <- add_trace(NYC_price, x = ~bronx$price, name = 'Bronx')
NYC_price <- add_trace(NYC_price, x = ~queens$price, name = 'Queens')
NYC_price <- add_trace(NYC_price, x = ~staten$price, name = 'Staten Island')
NYC_price <- layout(NYC_price, xaxis = list(title = 'Price'))
NYC_price
```

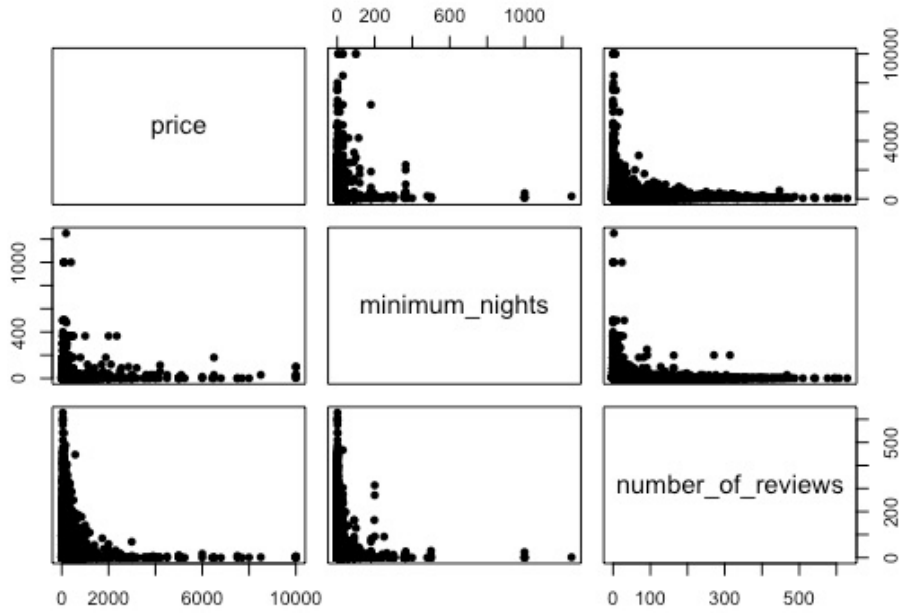


III. Comparation of the relationship between price, minimum night, number of reviews

1. Pairs diagram of the combination of the relationship of three variables

```
pairs(airbnb[,10:12], main = "Compare", pch=16)
```

Comparation of price, minimum night, and number of reviews



From the diagram above could clearly do the comparation to each two of the variables. There are three results could be determined.

Firstly, by compared of price and minimum_nights. More minimum nights need, lower price will be offered. And converse to be true that higher price gives less need of minimum nights.

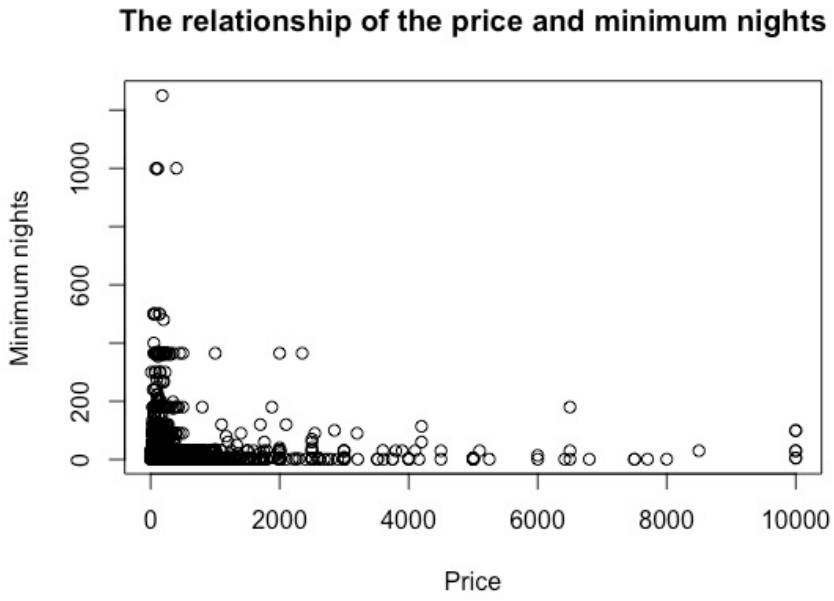
Additionally, for price and number_of_reviews. More number of reviews, then the price gets lower. Conversely, higher price gets a smaller number of reviews.

Last but not least, for minimum_nights and number_of_reviews. More number of reviews implies fewer minimum nights. Moreover, more minimum nights needed, a smaller number of reviews provided.

2. Scatterplot of the relationship of price and minimum nights

For a better vision of the relationship between price and minimum nights. There will be a scatterplot for this relationship to prove the result, and provide a better vision.

```
plot(airbnb$price, airbnb$minimum_nights,
     main = "The relationship of the price and minimum nights",
     xlab = "Price", ylab = "Minimum nights")
```

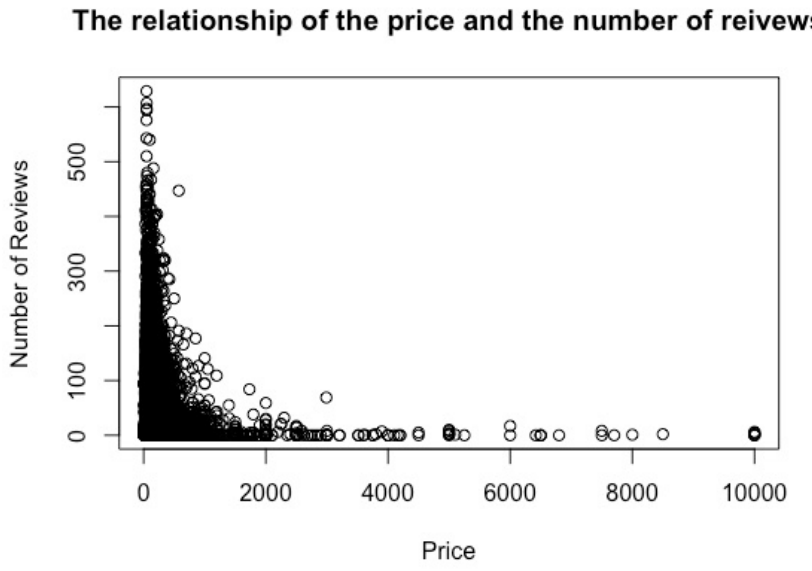


From the scatterplot showed above, the relationship between price and minimum nights is that higher price offered, fewer minimum nights needed. Which is as same as the result determined in the pairs diagram above.

3. Scatterplot for relationship of price and number of reviews

For the same reason, in order to provide a better vision of the relationship of price and number of reviews. By using plot() to provide a scatterplot to prove the result.

```
plot(airbnb$price, airbnb$number_of_reviews,
     main = "The relationship of the price and the number of reivews",
     xlab = "Price", ylab = "Number of Reviews")
```

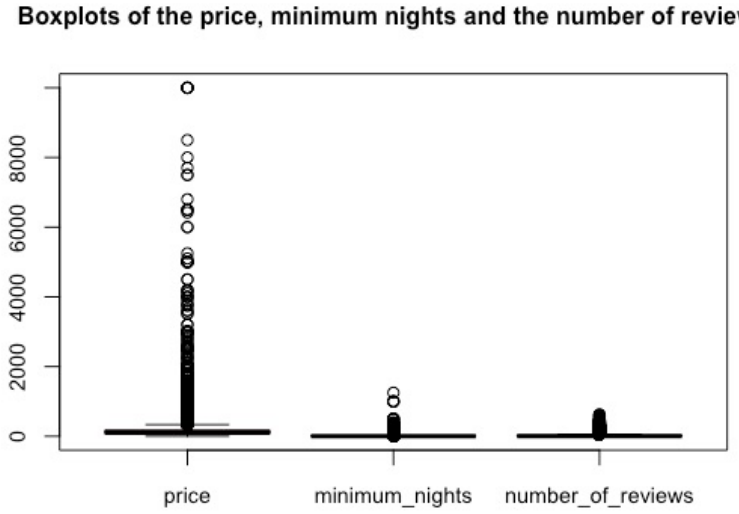


From the scatterplot, which clearly show the relationship that higher price implies a smaller number of reviews. Which proved the results.

4. Boxplots of price, minimum nights, and number of reviews

Applying the combination of boxplots of these three data to see their statistical number.

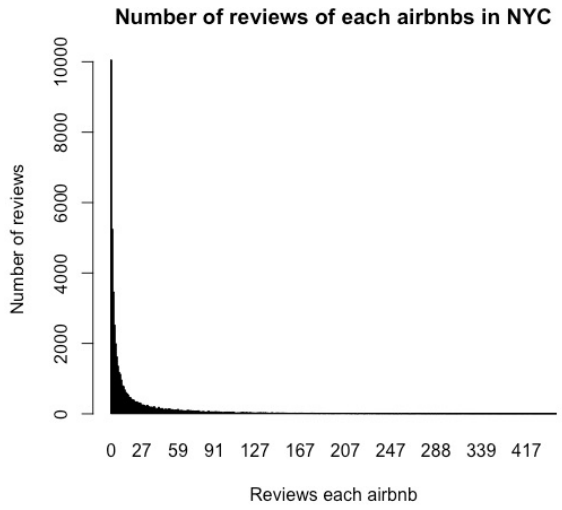
```
boxplot(airbnb[,10:12], col = c("red", "blue", "green"),
        main = "Boxplots of the price, minimum nights and the number of reviews",
        cex.main = 1, cex.axis = 0.9)
```



IV. Distribution Analysis

1. Plotting the data of number of reviews in airbnb open data

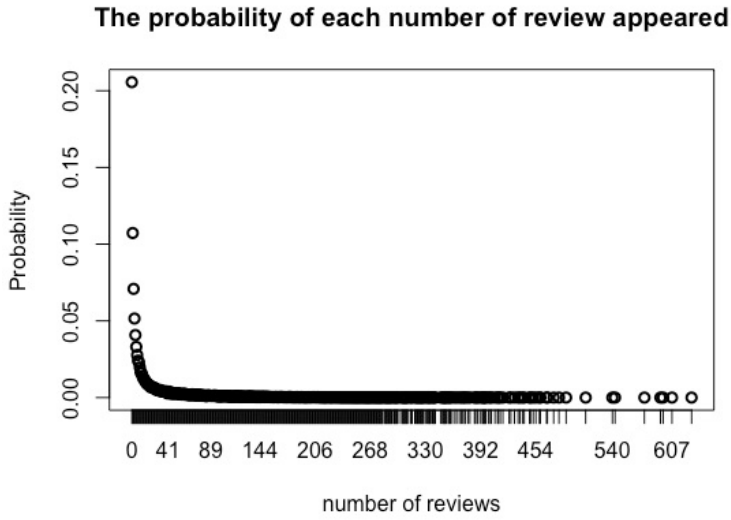
```
barplot(table(airbnb$number_of_reviews),
     main = "Number of reviews of each airbnbs in NYC",
     xlab = "Reviews each airbnb", ylab = "Number of reviews")
```



This plots clearly shows most of airbnbs has the number of reviews between 0 to 27.

2. Scatterplot for the probability of each number of reviews appeared

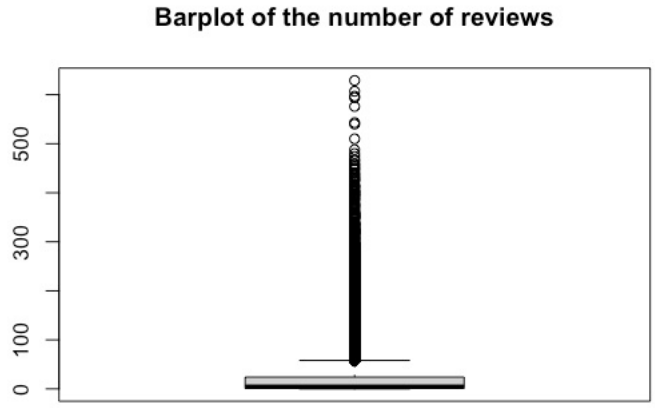
```
plot(table(airbnb$number_of_reviews)/nrow(airbnb), type = "p",
     main = "The probability of each number of review appeared",
     xlab = "number of reviews", ylab = "Probability")
```



This scatterplot shows the probabilities of the number of reviews got by each airbnb. It is less possible to get more reviews.

3. Boxplot for the number of reviews

```
boxplot(airbnb$number_of_reviews, main = "Barplot of the number of reviews")
```



Using boxplot to show the statistical information for the numerical column number_of_reviews.

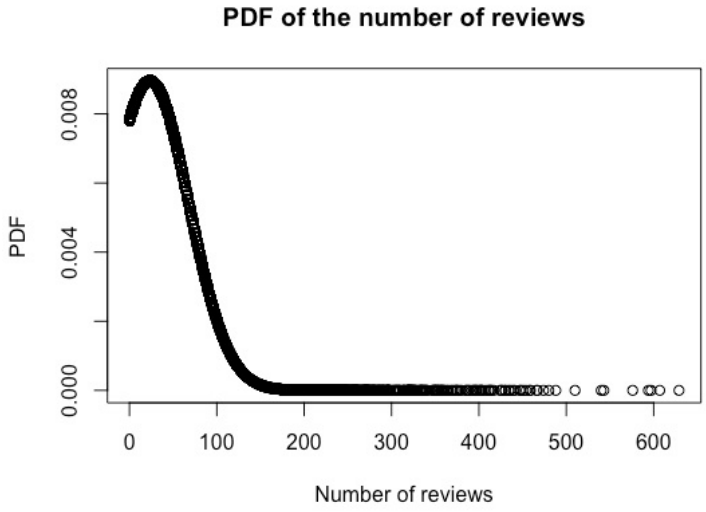
4. Summary of the number of reviews and its standard deviation

```
> summary(airbnb$number_of_reviews)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    5.00   23.27   24.00  629.00
> sd(airbnb$number_of_reviews)
[1] 44.55058
```

This is the six statistical number and standard deviation of number_of_reviews by using summary() and sd() functions.

5. Probability density function for the number reviews as normal distribution

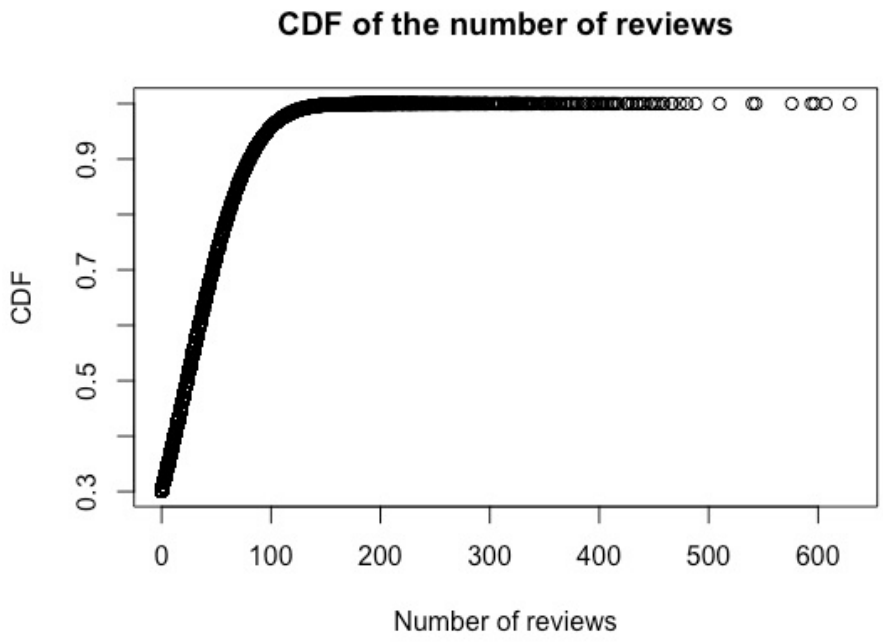
```
pdf<-dnorm(airbnb$number_of_reviews, mean = 23.27, sd = 44.55058)
plot(airbnb$number_of_reviews, pdf, type = "p",
     main = "PDF of the number of reviews",
     xlab = "Number of reviews", ylab = "PDF")
```



From the diagram of the probability density function, it could imply that more probabilities concentrated on [0,300], and the density is increasing from 0 to 600. As a result, an airbnb has a higher probability to receive fewer reviews.

6. Cumulative distribution function for the number reviews as normal distribution

```
cdf<-pnorm(airbnb$number_of_reviews, mean = 23.27, sd = 44.55058)

plot(airbnb$number_of_reviews, cdf, type = "p",
     main = "CDF of the number of reviews",
     xlab = "Number of reviews", ylab = "CDF")
```



From above diagram, it could come up with the same result of PDF, which is with the increasing number of reviews, the probability is decreasing declining significantly. With the decreasing of the slope, the probability is decreasing. Also, the density is decreasing with the increasing number of reviews.

V. Applicability of the Central Limit Theorem for Price

1. Setting data

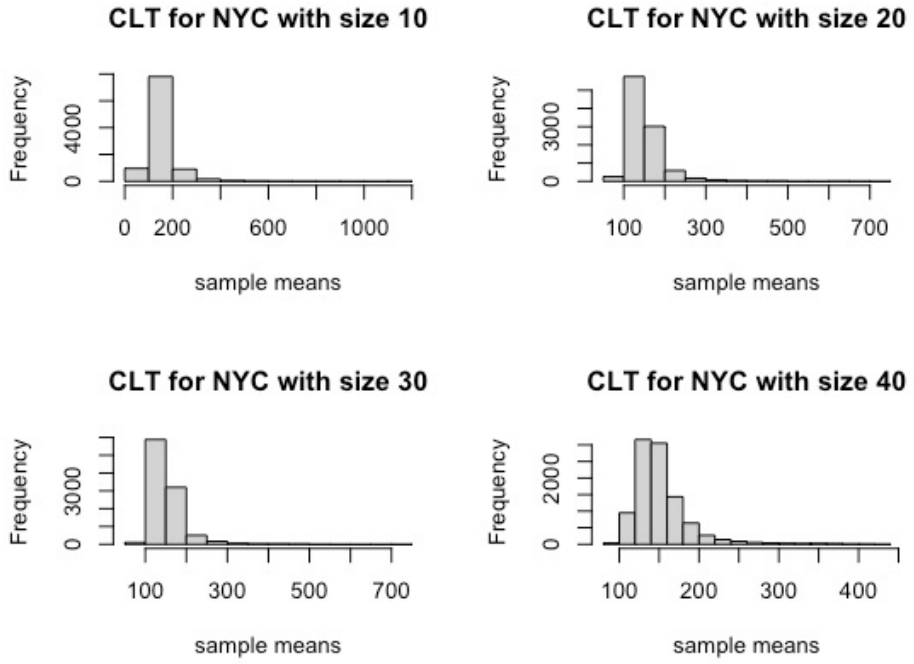
```
sample_draw <- 10000
sample_size1 <- 10
sample_size2 <- 20
sample_size3 <- 30
sample_size4 <- 40

xbar_10 <- numeric(sample_draw)
xbar_20 <- numeric(sample_draw)
xbar_30 <- numeric(sample_draw)
xbar_40 <- numeric(sample_draw)
```

Setting the data for central limit theorem. In this part, 4 different sample size 10, 20, 30, 40 will be used for each one method to do the data visualization for price of the whole NYC, and for each borough. 10000 samples will be draw randomly with replacement and equal probability.

2. Central Limit Theorem visualization for price

- Histogram for CLT of sample size 10, 20, 30, 40 for price

```
for (i in 1: sample_draw) {
  xbar_10[i] <- mean(sample(airbnb$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  xbar_20[i] <- mean(sample(airbnb$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  xbar_30[i] <- mean(sample(airbnb$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  xbar_40[i] <- mean(sample(airbnb$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(xbar_10, main = "CLT for NYC with size 10", xlab = "sample means")
hist(xbar_20, main = "CLT for NYC with size 20", xlab = "sample means")
hist(xbar_30, main = "CLT for NYC with size 30", xlab = "sample means")
hist(xbar_40, main = "CLT for NYC with size 40", xlab = "sample means")
```

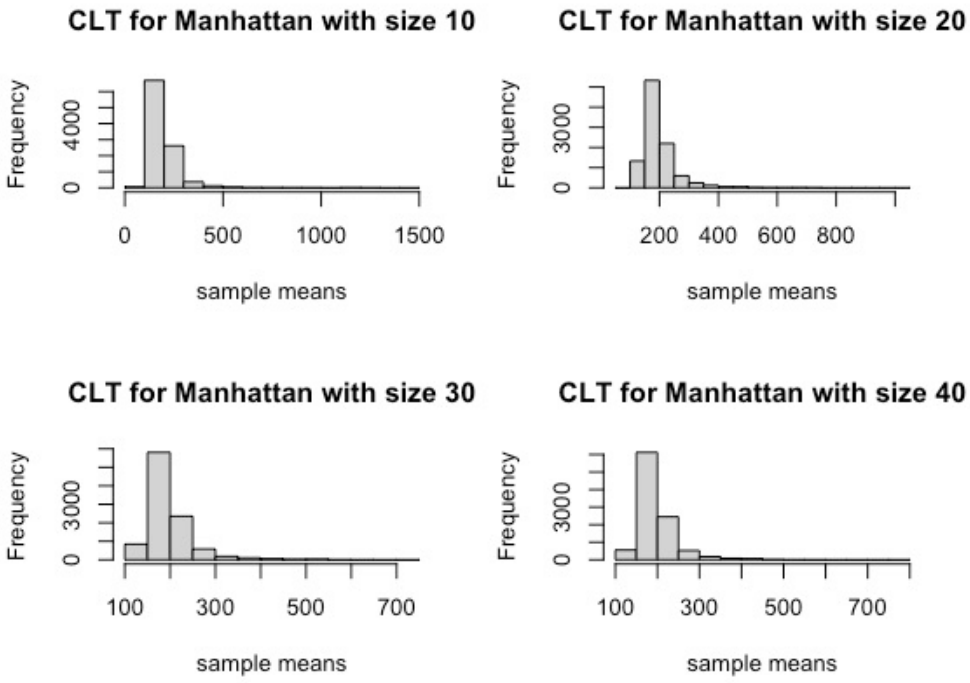


With each sample size, the means are all focus on 100 to 200. From the diagram, it could be determined that the price column is a right-skewed diagram. From the large data set airbnb, the sample means are normal distributed. And it is normal distributed in all four sample sizes. Next step, applying central limit theorem to each borough and doing data visualization for each one.

3. Central Limit Theorem visualization for each borough

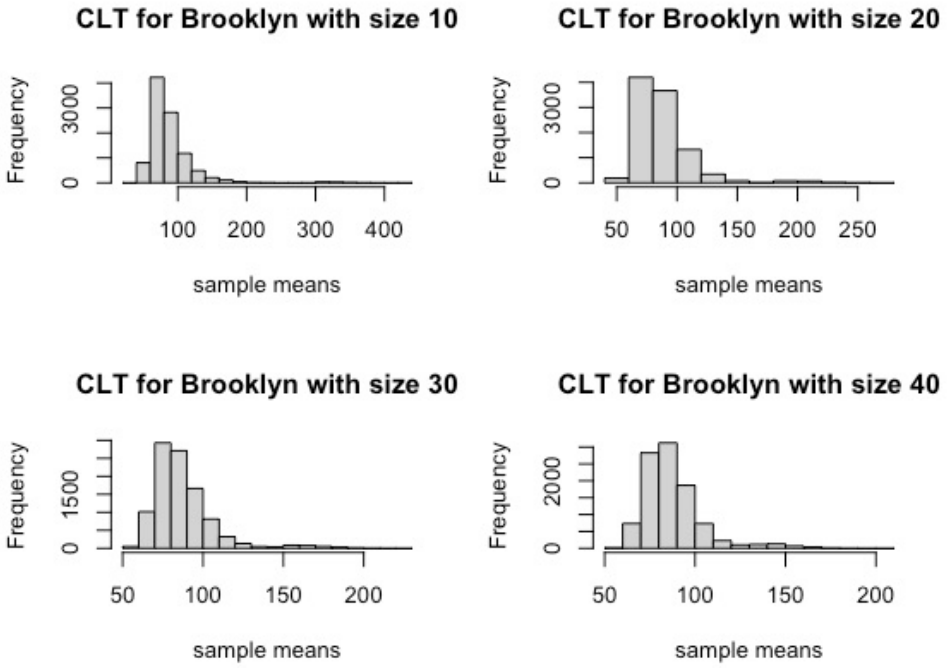
- CLT for Manhattan

```r
mbar_10 <- numeric(sample_draw)
mbar_20 <- numeric(sample_draw)
mbar_30 <- numeric(sample_draw)
mbar_40 <- numeric(sample_draw)

for (i in 1: sample_draw) {
  mbar_10[i] <- mean(sample(manhattan$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  mbar_20[i] <- mean(sample(manhattan$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  mbar_30[i] <- mean(sample(manhattan$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  mbar_40[i] <- mean(sample(manhattan$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(mbar_10, main = "CLT for Manhattan with size 10", xlab = "sample means")
hist(mbar_20, main = "CLT for Manhattan with size 20", xlab = "sample means")
hist(mbar_30, main = "CLT for Manhattan with size 30", xlab = "sample means")
hist(mbar_40, main = "CLT for Manhattan with size 40", xlab = "sample means")
```



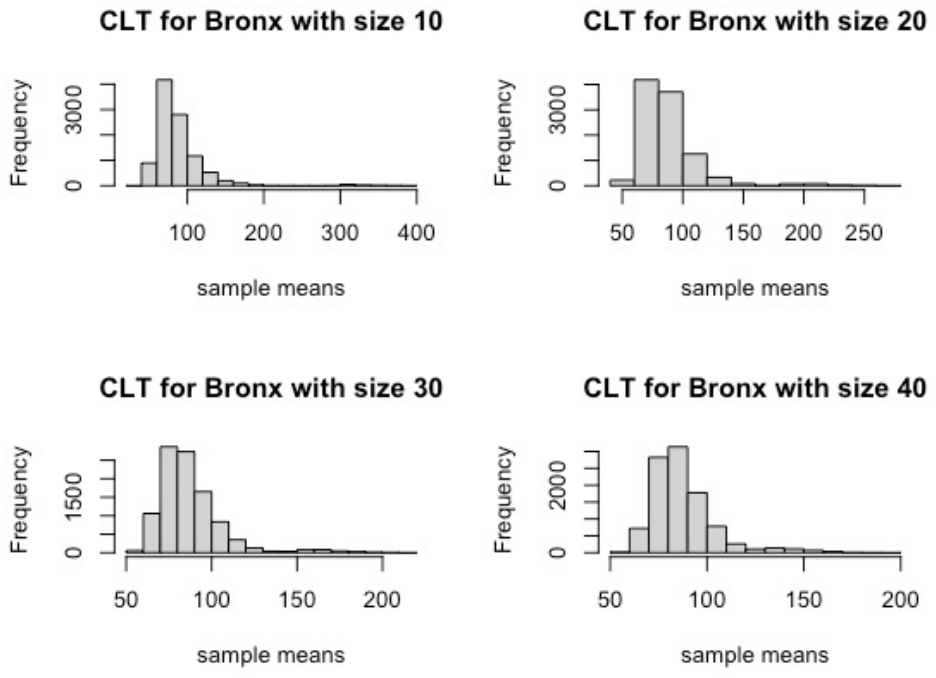
- CLT for Brooklyn

```
bbar_10 <- numeric(sample_draw)
bbar_20 <- numeric(sample_draw)
bbar_30 <- numeric(sample_draw)
bbar_40 <- numeric(sample_draw)

for (i in 1: sample_draw) {
  bbar_10[i] <- mean(sample(brooklyn$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  bbar_20[i] <- mean(sample(brooklyn$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  bbar_30[i] <- mean(sample(brooklyn$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  bbar_40[i] <- mean(sample(brooklyn$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(bbar_10, main = "CLT for Brooklyn with size 10", xlab = "sample means")
hist(bbar_20, main = "CLT for Brooklyn with size 20", xlab = "sample means")
hist(bbar_30, main = "CLT for Brooklyn with size 30", xlab = "sample means")
hist(bbar_40, main = "CLT for Brooklyn with size 40", xlab = "sample means")
```



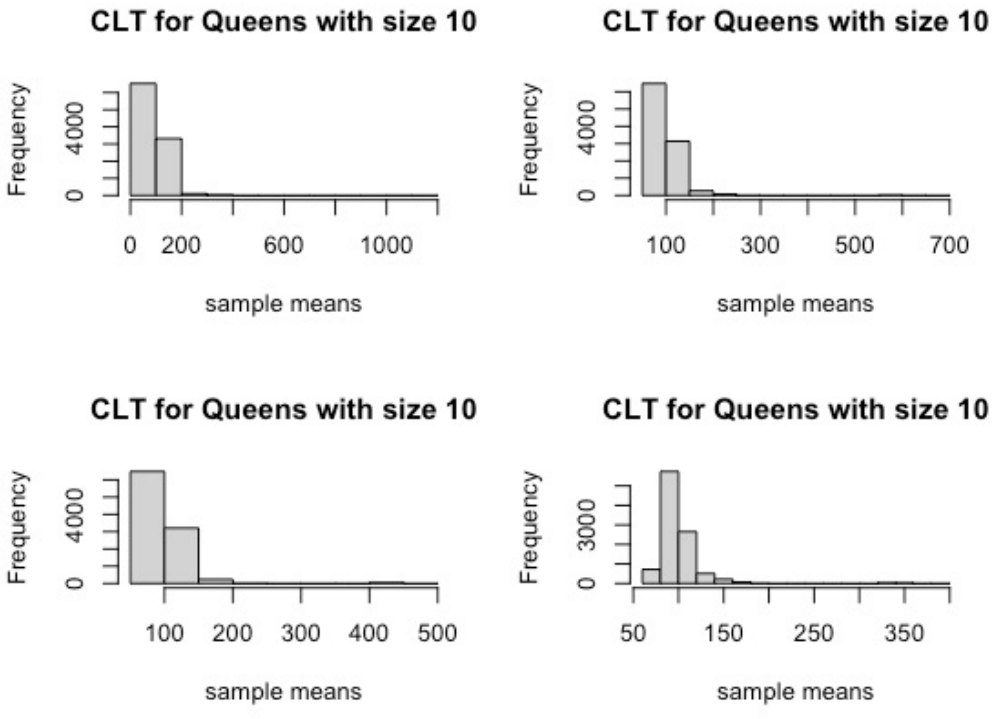
- CLT for Bronx

```
brbar_10 <- numeric(sample_draw)
brbar_20 <- numeric(sample_draw)
brbar_30 <- numeric(sample_draw)
brbar_40 <- numeric(sample_draw)

for (i in 1: sample_draw) {
  brbar_10[i] <- mean(sample(bronx$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  brbar_20[i] <- mean(sample(bronx$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  brbar_30[i] <- mean(sample(bronx$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  brbar_40[i] <- mean(sample(bronx$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(brbar_10, main = "CLT for Bronx with size 10", xlab = "sample means")
hist(brbar_20, main = "CLT for Bronx with size 20", xlab = "sample means")
hist(brbar_30, main = "CLT for Bronx with size 30", xlab = "sample means")
hist(brbar_40, main = "CLT for Bronx with size 40", xlab = "sample means")
```



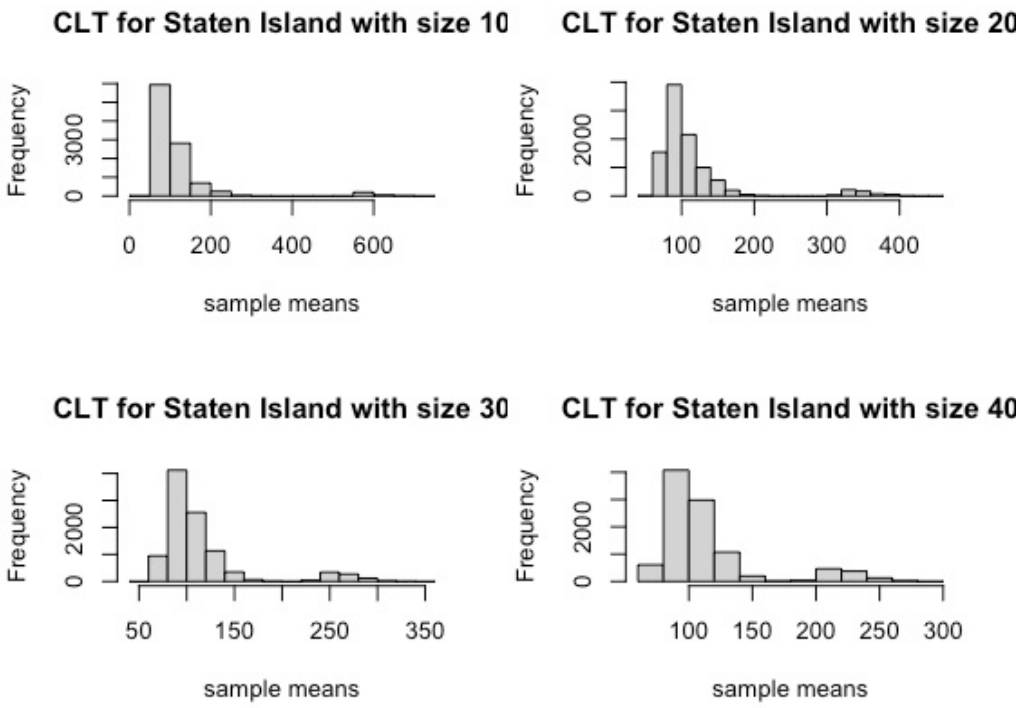
- CLT for Queens

```
qbar_10 <- numeric(sample_draw)
qbar_20 <- numeric(sample_draw)
qbar_30 <- numeric(sample_draw)
qbar_40 <- numeric(sample_draw)

for (i in 1: sample_draw) {
  qbar_10[i] <- mean(sample(queens$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  qbar_20[i] <- mean(sample(queens$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  qbar_30[i] <- mean(sample(queens$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  qbar_40[i] <- mean(sample(queens$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(qbar_10, main = "CLT for Queens with size 10", xlab = "sample means")
hist(qbar_20, main = "CLT for Queens with size 10", xlab = "sample means")
hist(qbar_30, main = "CLT for Queens with size 10", xlab = "sample means")
hist(qbar_40, main = "CLT for Queens with size 10", xlab = "sample means")
```



- CLT for Staten Island

```
sbar_10 <- numeric(sample_draw)
sbar_20 <- numeric(sample_draw)
sbar_30 <- numeric(sample_draw)
sbar_40 <- numeric(sample_draw)

for (i in 1: sample_draw) {
  sbar_10[i] <- mean(sample(staten$price, size = sample_size1, replace = FALSE))
}

for (i in 1: sample_draw) {
  sbar_20[i] <- mean(sample(staten$price, size = sample_size2, replace = FALSE))
}

for (i in 1: sample_draw) {
  sbar_30[i] <- mean(sample(staten$price, size = sample_size3, replace = FALSE))
}

for (i in 1: sample_draw) {
  sbar_40[i] <- mean(sample(staten$price, size = sample_size4, replace = FALSE))
}

par(mfrow=c(2,2))
hist(sbar_10, main = "CLT for Staten Island with size 10", xlab = "sample means")
hist(sbar_20, main = "CLT for Staten Island with size 20", xlab = "sample means")
hist(sbar_30, main = "CLT for Staten Island with size 30", xlab = "sample means")
hist(sbar_40, main = "CLT for Staten Island with size 40", xlab = "sample means")
```



VI. Sampling for airbnb open data

1. Setting data

```
size5000<-5000
NrowAirbnb <- nrow(airbnb)
numItems <- ceiling(NrowAirbnb / size5000)
```

Randomly draw 5000 samples from the whole airbnb data set by using sampling method to draw different samples.

2. Sample random sampling

- Table of 5000 random variables drew in Airbnb

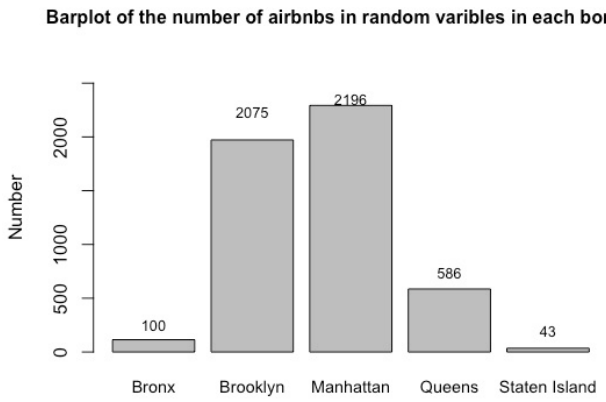
```
NYC_sampling1 <- srswr(size5000, nrow(airbnb))

rows <- (1:nrow(airbnb))[NYC_sampling1!=0]
rows <- rep(rows, NYC_sampling1[NYC_sampling1 != 0])
rs_NYC <- airbnb[rows,]
```

```
> table(rs_NYC$neighbourhood_group)

        Bronx      Brooklyn     Manhattan        Queens Staten Island
          114          1972          2293           586            35
```

By using sample random sampling, 5000 samples were drawing. By table() function, it shows that 114 samples from Bronx, 1972 samples from Brooklyn, 2293 samples from Manhattan, 586 samples from Queens, and 35 samples from Staten Island.

- Barpolot of 5000 random variables drew in airbnb

```
numberOfsrs <- as.numeric(table(rs_NYC$neighbourhood_group))

NYC_srs<-barplot(table(rs_NYC$neighbourhood_group),
        main = "Barplot of the number of airbnbs in random varibles in each borough",
        ylab = "Number", ylim = c(0,2700), cex.main = 1, cex.names = 0.9)

text(x = NYC_srs, y = numberOfsrs, label = numberOfsrs,
     pos = 3, cex = 0.8)
```



Doing the data visualization for the samples drew by sample random sampling.

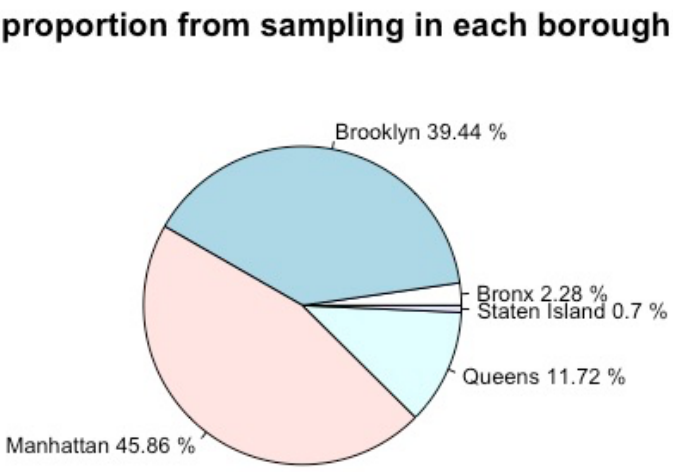
- Table of proportion of random variables drew in each borough

```
> table(rs_NYC$neighbourhood_group)/size5000

        Bronx      Brooklyn     Manhattan        Queens Staten Island
       0.0228        0.3944        0.4586        0.1172        0.0070
```

Now, calculate the proportion of sampling drew from each borough by using table(). Next, using pie chart to see a better version of the data.

- Pie chart of the proportion of random variables drew in each borough

```
table(rs_NYC$neighbourhood_group)/size5000

NYC_srs_prob <- table(rs_NYC$neighbourhood_group)/size5000

NYC_srs_prob_label <- names(NYC_srs_prob)
percent_srs <- paste(round(NYC_srs_prob,4)*100,"%")
NYC_srs_prob_label<-paste(NYC_srs_prob_label, percent_srs)


NYC_prob_srs<-pie(NYC_srs_prob, labels = NYC_srs_prob_label,
                                    main = "The proportion from sampling in each borough",
                                    cex = 0.8)
```



From the pie chart shows Manhattan occupied 45.86% of 5000 samples, which is the most. Brooklyn has the second the greatest number of samples and occupied 39.44%. Queens is the next occupied 11.72%. Bronx stands 2.28%. Staten Island has the least number of samples out of 5000, and it stands for 0.7%.

3. Systematic sampling

- Table of systematic sampling with size = 5000 for each borough

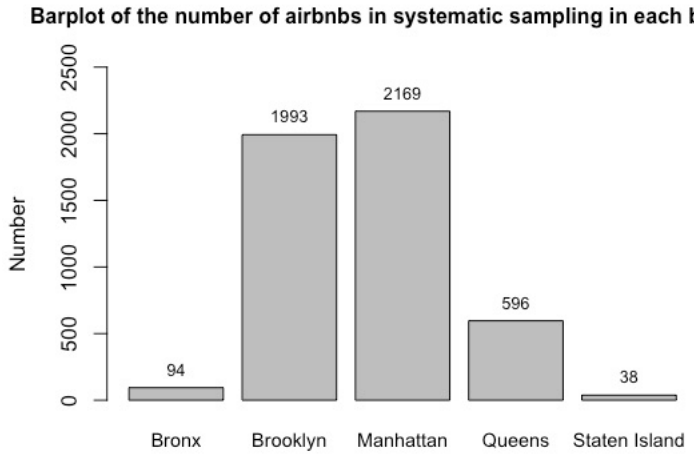
```
r <- sample(numItems,1)
NYC_sampling2 <- seq(r, by = numItems, length = size5000)
ss_NYC <- airbnb[NYC_sampling2, ]
```

```
> table(ss_NYC$neighbourhood_group)

       Bronx      Brooklyn     Manhattan        Queens Staten Island
          94          1993          2169           596            38
```

By using systematic sampling to draw 5000 samples form airbnb data set. The table shows 94 samples from Bronx, 1993 from Brooklyn, 2169 from Manhattan which is the biggest, 596 samples from Queens, and only 38 sample from Staten Island, which is the least one.

- Barplot of systematic sampling with size = 5000 for each borough

```
numberOfss <- as.numeric(table(ss_NYC$neighbourhood_group))

NYC_ss<-barplot(table(ss_NYC$neighbourhood_group),
                main = "Barplot of the number of airbnbs in systematic sampling in each borough",
                ylab = "Number", ylim = c(0,2500), cex.main = 1, cex.names = 0.9)

text(x = NYC_ss, y = numberOfss, label = numberOfss,
     pos = 3, cex = 0.8)
```



Using barplot to show the number of samples drew from each borough could also show Manhattan has the most samples and Staten Island has the least samples.

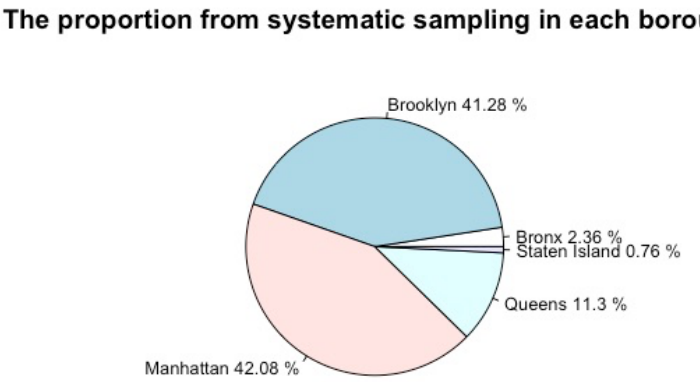
- Table of the proportion of systematic sampling with size = 5000 for each borough

```
> table(ss_NYC$neighbourhood_group)/size5000

        Bronx      Brooklyn     Manhattan        Queens Staten Island
       0.0188        0.3986        0.4338        0.1192        0.0076
```

Calculating the proportion of the samples drew by systematic sampling of each borough. Moreover, doing pie chart again to see the clearly proportion of samples in each borough.

- Pie chart for the proportion of systematic sampling with size = 5000 for each borough

```
table(ss_NYC$neighbourhood_group)/size5000

NYC_ss_prob <- table(ss_NYC$neighbourhood_group)/size5000

NYC_ss_prob_label <- names(NYC_ss_prob)
percent_ss <- paste(round(NYC_ss_prob,4)*100,"%")
NYC_ss_prob_label<-paste(NYC_ss_prob_label, percent_ss)


NYC_prob_ss<-pie(NYC_ss_prob, labels = NYC_ss_prob_label,
                 main = "The proportion from systematic sampling in each borough",
                 cex = 0.8)
```



From the pie chart above, it shows Manhattan has 42.08% of 5000 samples, which has the most samples. Brooklyn has the second the greatest number of samples and occupied 41.28%. Queens is the next has 11.3%. Bronx has 2.36%. Staten Island has the least number of samples out of 5000, it only has 0.76%.

4. Inclusion probability

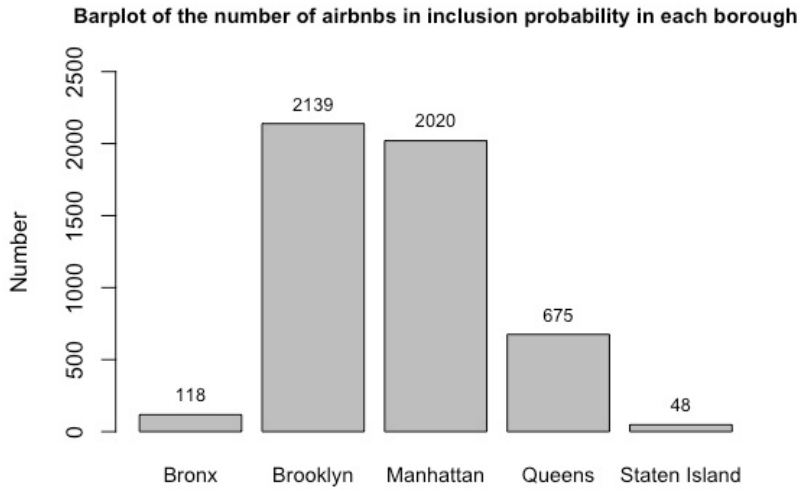
- Table of inclusion probability

```
inclusion <- inclusionprobabilities(airbnb$number_of_reviews, size5000)

NYC_sampling3 <- UPsystematic(inclusion)

ip_NYC <- airbnb[NYC_sampling3 != 0,]

> table(ip_NYC$neighbourhood_group)

        Bronx      Brooklyn     Manhattan        Queens Staten Island
          118          2139          2020           675            48
```

For inclusion probability, drawing 5000 samples. 118 samples from Bronx, 2139 sample from Brooklyn, 2020 samples from Manhattan, 675 samples from Queens, and 48 samples from Staten Island.

- Barplot of inclusion probability

```
numberOfip <- as.numeric(table(ip_NYC$neighbourhood_group))

NYC_ip<-barplot(table(ip_NYC$neighbourhood_group),
                main = "Barplot of the number of airbnbs in inclusion probability in each borough",
                ylab = "Number", ylim = c(0,2500), cex.main = 0.9, cex.names = 0.9)

text(x = NYC_ip, y = numberOfip, label = numberOfip,
     pos = 3, cex = 0.8)
```



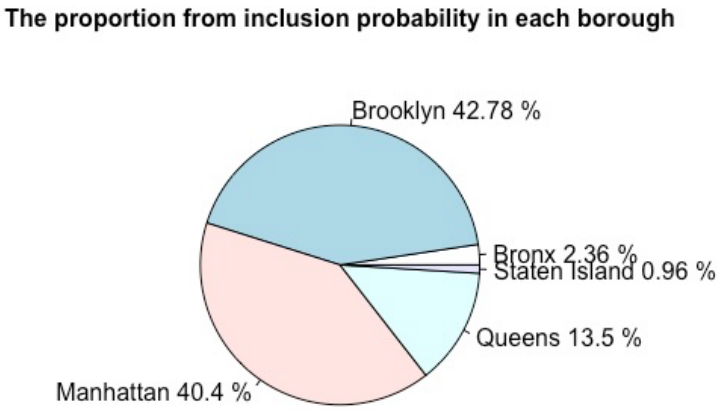
From the diagram, Brooklyn has the most samples this time, Manhattan has the second the greatest number of samples. Staten Island still has the least number of samples.

- Table of the proportion of inclusion probability of the number of airbnbs

```
> table(ip_NYC$neighbourhood_group)/size5000

       Bronx      Brooklyn     Manhattan        Queens Staten Island
      0.0236        0.4278        0.4040        0.1350        0.0096
```

- Pie chart for the proportion of inclusion probability of the number of airbnbs

```
NYC_ip_prob <- table(ip_NYC$neighbourhood_group)/size5000

NYC_ip_prob_label <- names(NYC_ip_prob)
percent_ip <- paste(round(NYC_ip_prob,4)*100,"%")
NYC_ip_prob_label<-paste(NYC_ip_prob_label, percent_ip)


NYC_prob_ip<-pie(NYC_ip_prob, labels = NYC_ip_prob_label,
                 main = "The proportion from inclusion probability in each borough",
                 cex.main = 1)
```



From the pie chart above, it shows Manhattan has 40.4% of 5000 samples, which has the second most samples this time. Brooklyn has the second the greatest number of samples and occupied 42.78%, which has the greatest number of samples this time.

Queens is the next has 13.5%. Bronx has 2.36%. Staten Island still has the least number of samples out of 5000, it only has 0.96%.

5. Stratified

- Setting data

```
boroughs <- rep(names(sort(table(airbnb$neighbourhood_group), each = nrow(airbnb))))

NYC_rows <- round(runif(500, 1, nrow(airbnb)))

NYC_data <- data.frame(
  Boroughs = boroughs,
  Row = NYC_rows)
```

- Table and description of stratified data

```
> table(NYC_data$Boroughs)

        Bronx      Brooklyn     Manhattan        Queens Staten Island
          100           100           100           100           100
```

- Strata function for airbnb open data with 500 sample size

```
> NYC_sampling4 <- strata(NYC_data, stratanames = c("Boroughs"),
+                   size = rep(100, 5), method = "srswor",
+                   description = TRUE)
Stratum 1

Population total and number of selected units: 100 100
Stratum 2

Population total and number of selected units: 100 100
Stratum 3

Population total and number of selected units: 100 100
Stratum 4

Population total and number of selected units: 100 100
Stratum 5

Population total and number of selected units: 100 100
Number of strata  5
Total number of selected units 500
```

These sampling could help researchers to save their times and obtain a reliable result. Without sampling, researchers need to spend a lot of time to get the data from real-

time data. By using sampling, researchers could do the data analysis and expect the future result. In this project, the sample size is 10000 and 5000 which provide the idea that if there are 5000 or 10000 customers are looking for an airbnb to stay, which price interval they probably choose. And it could help the company gets an accurate result because random variables simulated the number of real customers. The company could clearly see the mean, standard deviation and based on different variable such as price or number of reviews which one they choose the most. Applying these samples to the whole dataset will not provide an expectation, it shows the regular analysis.

6. Means comparison of price

```
> mean(rs_NYC$price)
[1] 155.3012
> summary(ss_NYC$price)["Mean"]
    Mean
149.5936
> mean(ip_NYC$price)
[1] 136.807
```

Conclusion

The goal of this project is to show and determine three information, three relationships and sampling for the data. Three information that need to be determined is:

1. How many airbnbs in each borough, which has the most and which has the least?
2. Which borough has the highest price and which has the lowest?
3. What are their proportions?

For the first one, from first part of the project, we can know that there are 1091 airbnbs in Bronx, 20104 airbnbs in Brooklyn, 21661 airbnbs in Manhattan, 5666 airbnbs

in Queens, and 373 airbnbs in Staten Islands. Secondly, Manhattan has the largest number of airbnbs which is 21661. Staten Island has the least number of airbnbs which is 373. Eventually, Manhattan occupied the most percentage of the number of airbnbs which is 44.3%. Brooklyn is NO.2 which has 41.12%. Queens has 11.59%. Bronx has 2.23%. And Staten Island has the least percentage, which is 2.23%.

The three relationships that this project want to explore is:

1. The relationship between price and number of reviews.
2. The relationship between price and minimum nights
3. The relationship between number of reviews and minimum nights.

For these three relationships, from pairs diagram. More minimum nights need, lower price will be offered. And converse to be true that higher price gives less need of minimum nights. In addition, a greater number of reviews, then the price gets lower. Conversely, higher price gets a smaller number of reviews. Lastly, for minimum_nights and number_of_reviews. More number of reviews implies fewer minimum nights. Moreover, more minimum nights needed, a smaller number of reviews provided.

From distribution analysis, it is less possible to get more reviews. The probability is decreasing. From the diagram of the probability density function, it implies that more probabilities concentrated on [0,300], and the density is increasing from 0 to 600. As a result, an airbnb has a higher probability to receive fewer reviews. From cumulative density function, the increasing number of reviews, the probability is decreasing declining significantly. With the decreasing of the slope, the probability is decreasing. Also, the density is decreasing with the increasing number of reviews.

Applying central limit theorem to the whole data set and each borough, it could be determined that the price column is a left-skewed diagram. From the large data set airbnb, the sample means are normal distributed. And it is normal distributed in all four sample sizes. Next step, applying central limit theorem to each borough and doing data visualization for each one.

Four sampling methods used in this project, sample random sampling, systematic sampling, inclusion probability, and strata. Generating 5000 samples from the whole airbnb data set. By comparing the mean of the first three sampling methods, the means are decreasing a little.

References

D. (2019, August 12). *New York City Airbnb Open Data*. Kaggle. https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data?select=AB_NYC_2019.csv

Kalathur, S. (2020, October 7). *CS 544 Module 5*. BU CS544. https://onlinecampus.bu.edu/bbcswebdav/pid-7953809-dt-content-rid-40368175_1/courses/20fallmetcs544_o1/course/module5/allpages.htm

Kalathur, S. (2020, September 30). *CS 544 Module 4*. BU CS544. https://onlinecampus.bu.edu/bbcswebdav/pid-7953808-dt-content-rid-40368145_1/courses/20fallmetcs544_o1/course/module4/allpages.htm

Kalathur, S. (2020, September 23). *CS 544 Module 3*. BU CS544. https://onlinecampus.bu.edu/bbcswebdav/pid-7953807-dt-content-rid-40368122_1/courses/20fallmetcs544_o1/course/module3/allpages.htm