NYC Airbnb Listing data – 2019

Viraj Kothari

Dt: March 4, 2020

Understanding the characteristics of Airbnb listing in New York City.

❖ Overview of Data and Data cleaning

This data set is known as NYC Airbnb open data, this data is open for public. You can acquire this dataset by going to NYC open data website or you can also find it on www.kaggle.com . I took this data set from Kaggle, below is the direct link to the data,

https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data

File name is AB_NYC_2019, it is the summary information about the Airbnb listing specifically in NYC. Originally data had 48895 rows and there were 16 columns and they were as shown in fig, but since we did not need some of the columns.

COLUMN NAMES	DESCRIPTION
id	Listing ID
name	name of the listing
host_id	host ID
host_name	name of the host
neighbourhood_group	NYC borough
neighbourhood	area
latitude	latitude coordinates
longitude	longitude coordinates
room_type	listing space type, private/shared room or full apartment
price	price in dollars per night
minimum_nights	amount of nights minimum
number_of_reviews	number of reviews
last_review	latest review
reviews_per_month	number of reviews per month
calculated_host_listings_count	amount of listing per host
availability_365	number of days when listing is available for booking

As we can see that while analyzing data, what columns will be most useful and interesting to study. According to me, our main focus will be on looking "neighborhood_group", "room_type", "price", "number_of_reviews", "reviews_per_month", "neighborhood" and "precise location". Hence our final data set consisted 48895 rows and 12 columns, which were,

{"name", "host_id", "host_name", "neighbourhood_group", "neighbourhood", "latitude", "longitude", "room_type", "price", "minimum_nights", "number_of_reviews", "reviews_per_month"}

Since we figured our columns, we need to make sure that there is no NA value in our data which was done in R by running some code which as below, we did find one column that had NA's but which would have been computed easily, which was also done in R.

```
> df = read.csv("AB_NYC_2019.csv")
> col2keep = c("name","host_id","host_name","neighbourhood_group",
             "neighbourhood","latitude","longitude","room_type","price",
             "minimum_nights","number_of_reviews","reviews_per_month")
> df = df[,col2keep]
> na.cols = c()
> # This is to check weather we have any columns with NA
> for (i in colnames(df)) {
+   if(any(is.na(df[i]))){
+     na.cols=c(i)
+     }
+ }
> na.cols
[1] "reviews_per_month"
> # It gives one col that has NA's in it and it is reviews_per_month,
which we can compute it
> #   from reviews by dividing with 12
> df$reviews_per_month = df$number_of_reviews/12
> any(is.na(df$reviews_per_month))
[1] FALSE
```

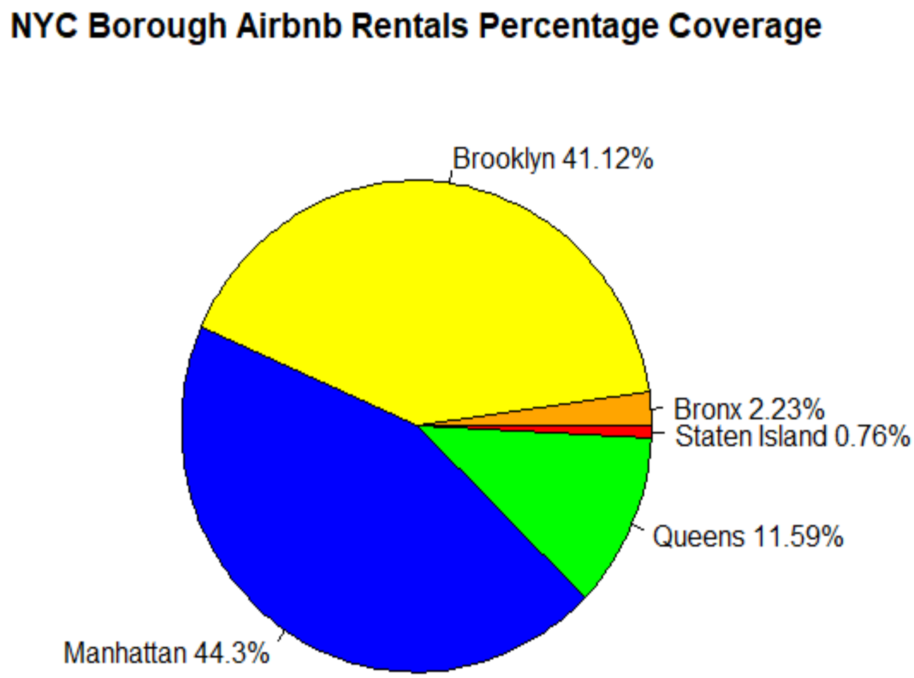
Hence, we have finished our data preparation and data cleaning process, we can proceed to analyze our data for further understanding

❖ Analysis of Neighborhood group and Room type in Dataset

We know that NYC is divided into five boroughs, which are Manhattan, Queens, Brooklyn, Staten Island, and Bronx. Although other boroughs are famous, tourists always come for Manhattan, and that is fact. I won't be surprise if Airbnb have most of the Ad listing in Manhattan. However, Brooklyn and Queens might not be lagging because of expensive Manhattan. Let's have look at what our data has to say about 2019 Airbnb listing,

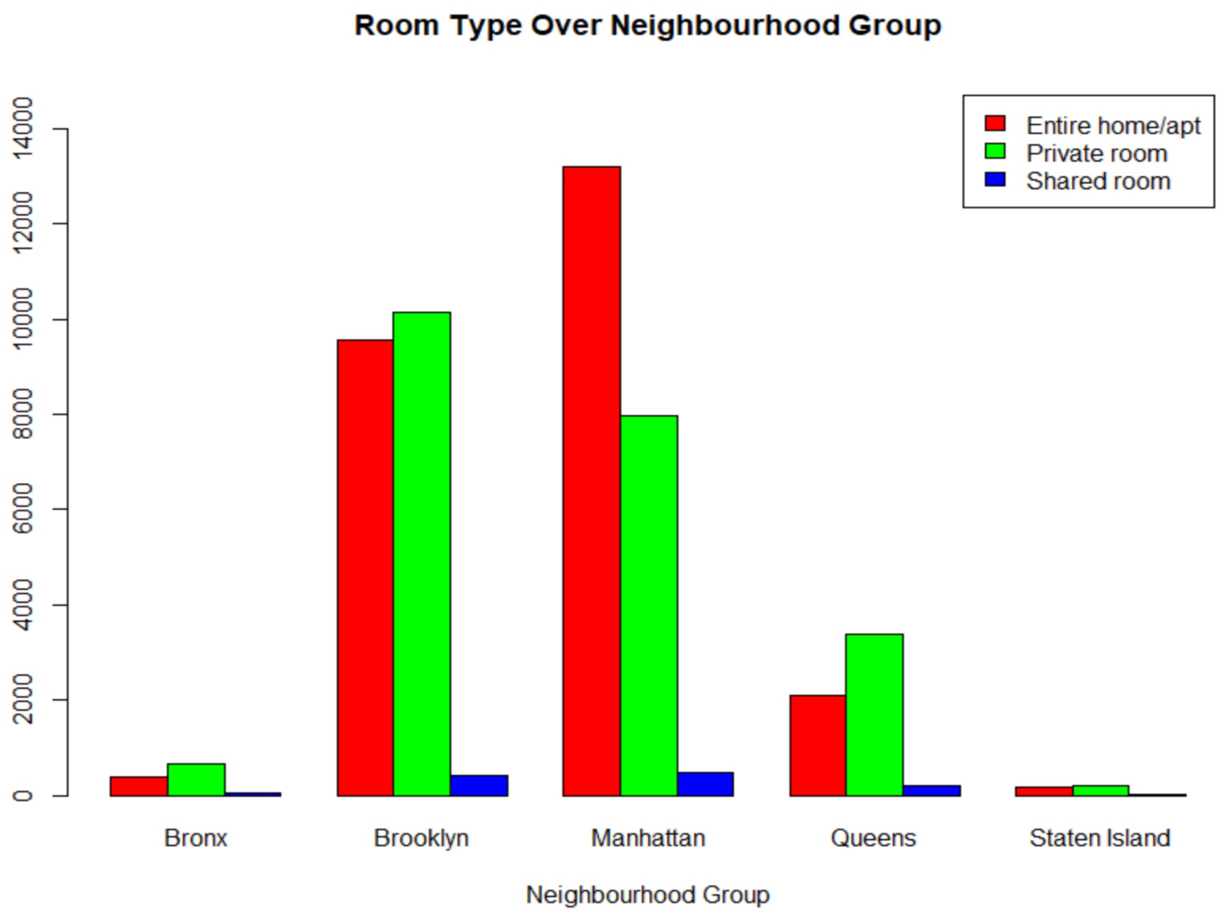
```
> x = table(df$neighbourhood_group)
> x
   Bronx    Brooklyn    Manhattan    Queens    Staten Island
   1091     20104       21661        5666      373
```

As it was expected that Manhattan will lead the listing, Brooklyn did surprise me. Looking at Pie chart of their coverage below we can see that Manhattan is 44% whereas Brooklyn is 41%, does that mean people will choose to rent in Brooklyn since they can save some money for renting Airbnb but have more travel time to city since NYC is also famous for its traffic.



Low percentage of Staten Island and Bronx can be understood since they are the farthest from Manhattan but having this low ad posting is not giving enough options for Airbnb users to research their trip specifically to Staten Island. There are also certain tourist spots in Staten Island also their beaches.

It will also be interesting to look up what type of place is offered on listing at respective boroughs. We know since there are three types of space offered shared room, private room and full apartment/house. Below is shown their room type for their respective borough by their frequency of listing. R code shows the percentage of their contribution as well,




```
> # Frequency and their respective total
> addmargins(y)
                 Bronx Brooklyn Manhattan Queens Staten Island   Sum
  Entire home/apt  379     9559     13199   2096           176 25409
  Private room     652    10132      7982   3372           188 22326
  Shared room       60      413       480    198             9  1160
  Sum             1091    20104     21661   5666           373 48895

> # Percentage of contribution in data
> addmargins(y.prop)*100

                    Bronx Brooklyn Manhattan  Queens   Staten     Sum

  Entire home/apt  0.7751  19.5500  26.9945   4.2867  0.3599    51.96
  Private room     1.3334  20.7219  16.3247   6.8964  0.3844    45.66
  Shared room      0.1227   0.8446   0.9816   0.4049  0.0184     2.37
  Sum              2.2313  41.1166  44.3010  11.5880  0.7628   100.00
```

As it seems that Airbnb listing is mostly for Entire home/apt and very few for shared room, and highest percentage of listing is Entire home/apt in Manhattan. We can see the probability of listing being private room if it is in Queens. Although we cannot say anything about

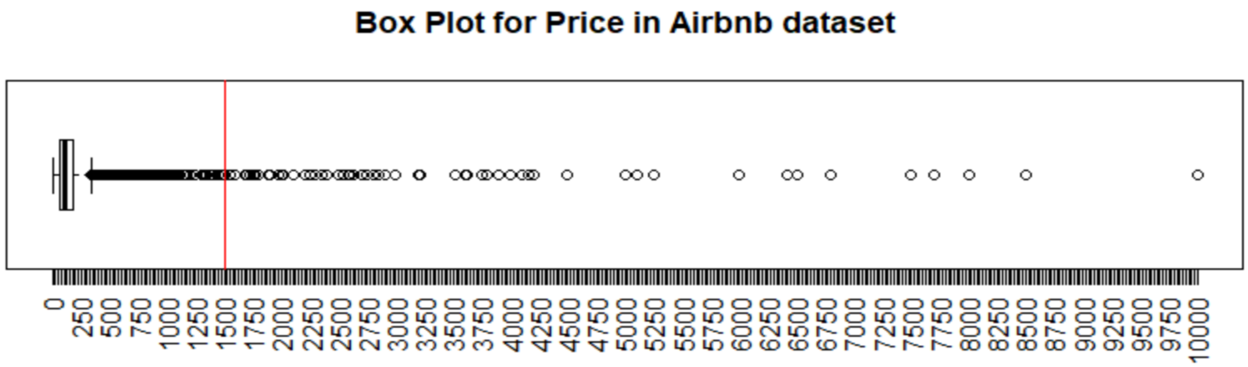
Brooklyn, we can say that if listing is private room then it is most likely to be in Brooklyn. I am mostly surprised to see the pattern of private rooms, as it is seen that in all borough, private room listing is higher but quite opposite when it comes to Manhattan as we would expect to be more of room rental because of Airbnb model was based on renting extra room. I think it’s maybe of corporate rentals or investment rentals in Manhattan. Brooklyn will be second option over Queens for Airbnb users since Brooklyn has second most Airbnb listings.

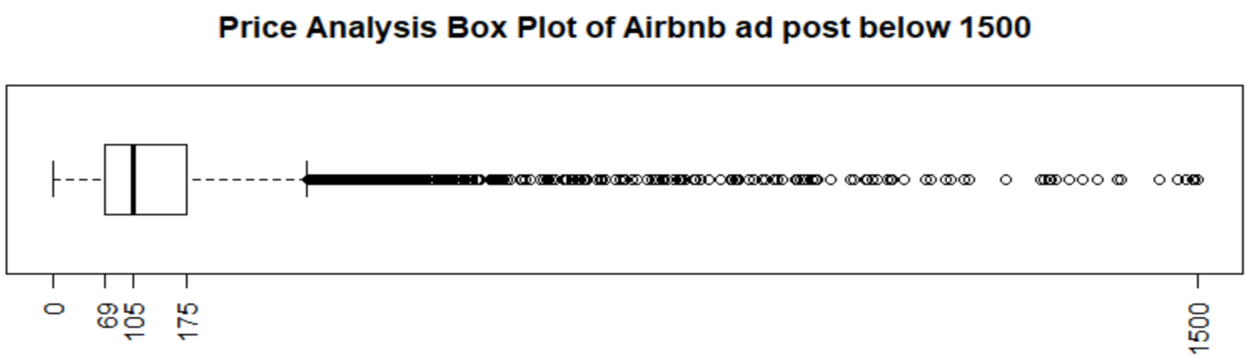
❖ Analysis of Price Variable in our Dataset

We are talking about Airbnb in NYC and it wouldn’t be interesting we don’t look into price of listing. Therefore first thing we do is to see what is the summary for our price, which is mean, median, min and max also Inter Quartile Range.

```
> summary(df$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0    69.0   106.0   152.7   175.0 10000.0 
> f = fivenum(df$price)
> IQR = f[4]-f[2]
> IQR
[1] 106
```

As we can see that Maximum price for Airbnb is listed to 10000 and Minimum is 0 which is definitely wrong, so I ran code to see how many 0 are there which came to be 11 which is very low number hence we can neglect that much error. After running Boxplot on our data, it gave some light on our data. That gave us much more outliers, which you can see in fig below, the first boxplot is the one with whole data set. After seeing boxplot can be concluded that our data might not have as much outlier because we can see good amount of data points till some point. Hence, I divided data to price of $1500, which is acceptable amount of rents considering Manhattan data. Second Boxplot is for the data points below price 1500.





Removing data points over $1500 price still does not affect mean or median, only thing affected is Maximum. Although plot shows us outliers, there are so many points that we cannot considered them outliers. However, it would nice to study points over $1500 price. Let's look their respective borough after making subset data frame, presented is the R code and output,

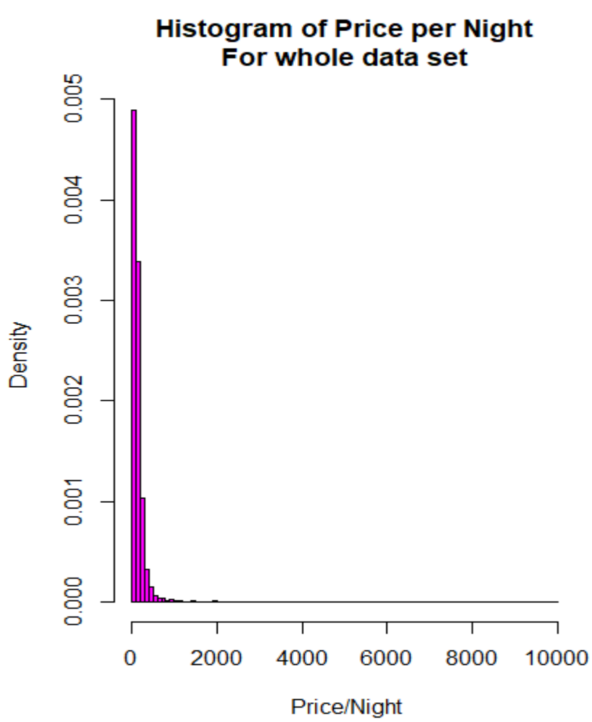
```
> df_price_1500 = subset(df,df$price>1500)
> nrow(df_price_1500)
[1] 139
> table(df_price_1500$neighbourhood_group)
     Bronx      Brooklyn     Manhattan          Queens Staten Island
         1            34            97               6             1
```

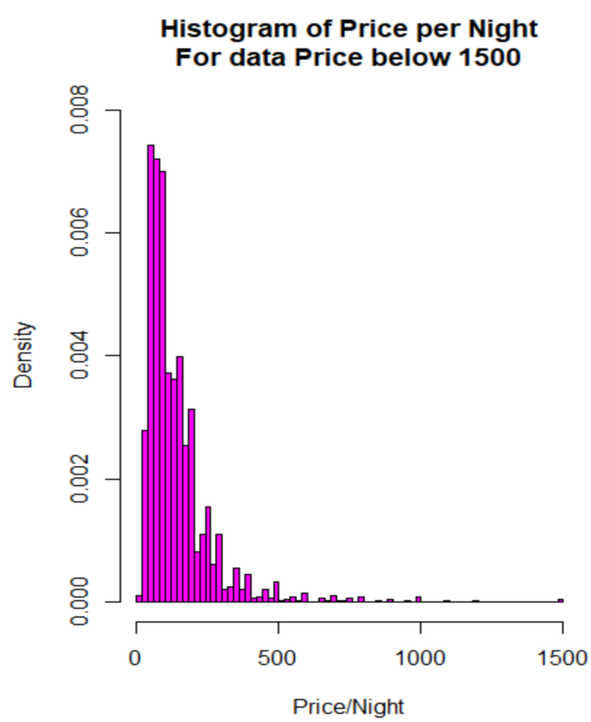
Someone pays over 1500 for place is not impossible for Manhattan and Brooklyn because for their high-class parties. Though it is highly doubtable for Bronx, Staten Island and Queens because we already seen their listing frequency which was not high. Let's look up the top 10 highest priced listing of Airbnb in NYC.

```
> df_price_1500[order(df_price_1500$price,decreasing = TRUE)[1:10],]
                                                  name   host_id host_name neighbourhood_group     neighbourhood
9152                 Furnished room in Astoria apartment  20582832  Kathrine              Queens           Astoria
17693    Luxury 1 bedroom apt. -stunning Manhattan views   5143901      Erin            Brooklyn        Greenpoint
29239                             1-BR Lincoln Center  72390391    Jelena           Manhattan   Upper West Side
6531                              Spanish Harlem Apt   1235070     Olson           Manhattan       East Harlem
12343                 Quiet, Clean, Lit @ LES & Chinatown   3906464       Amy           Manhattan   Lower East Side
40434 2br - The Heart of NYC: Manhattans Lower East Side   4382127      Matt           Manhattan   Lower East Side
30269  Beautiful/Spacious 1 bed luxury flat-TriBeCa/Soho  18128455       Rum           Manhattan           Tribeca
4378                                    Film Location   1177497   Jessica            Brooklyn      Clinton Hill
29663           East 72nd Townhouse by (Hidden by Airbnb) 156158778     Sally           Manhattan   Upper East Side
42524                   70' Luxury MotorYacht on the Hudson   7407743      Jack           Manhattan Battery Park City
      latitude longitude       room_type price minimum_nights number_of_reviews reviews_per_month
9152  40.76810 -73.91651    Private room 10000             100                 2        0.16666667
17693 40.73260 -73.95739 Entire home/apt 10000               5                 5        0.41666667
29239 40.77213 -73.98665 Entire home/apt 10000              30                 0        0.00000000
6531  40.79264 -73.93898 Entire home/apt  9999               5                 1        0.08333333
12343 40.71355 -73.98507    Private room  9999              99                 6        0.50000000
40434 40.71980 -73.98566 Entire home/apt  9999              30                 0        0.00000000
30269 40.72197 -74.00633 Entire home/apt  8500              30                 2        0.16666667
4378  40.69137 -73.96723 Entire home/apt  8000               1                 1        0.08333333
29663 40.76824 -73.95989 Entire home/apt  7703               1                 0        0.00000000
42524 40.71162 -74.01693 Entire home/apt  7500               1                 0        0.00000000
```

Queens being in over 1500$ was doubtable enough and we are seeing it being top one, but it does seem like it's an error in posting. It was supposed to be price per night and minimum nights, the person might have posted total price of stay but when you go down you see 8000$ and night is one that is not error.

❖ Examine the Distribution of Listing Price





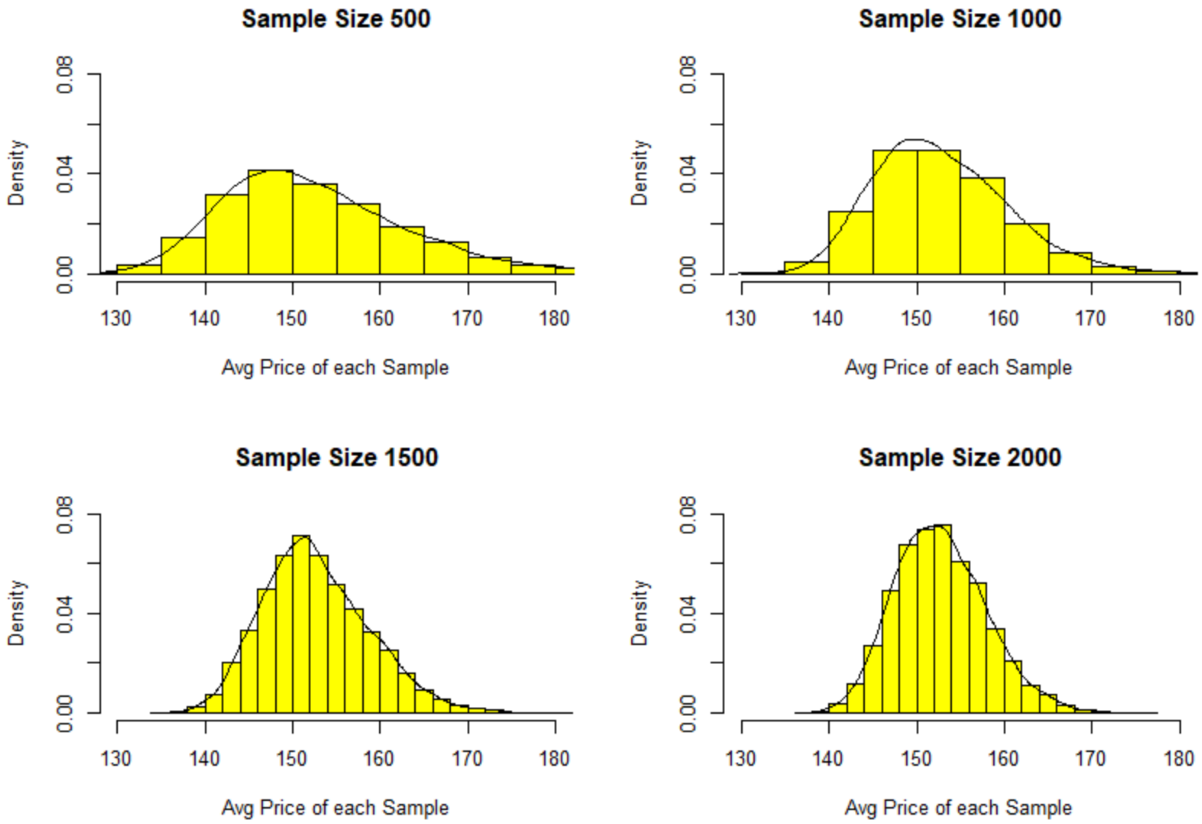
```
> summary(df$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    69.0   106.0   152.7   175.0 10000.0
> sd.price
[1] 240.1542
> summary(price_below_1500$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    69.0   105.0   143.6   175.0  1500.0
> sd(price_below_1500$price)
[1] 127.8256
```

Here is distribution of price variable for different data set, first histogram shows that it is skewed right with very high percentage. That's all we can say about that distribution, in other case we can see that it looks like little bit normal distribution but it is also skewed right, we can say price variable in our data set is right skewed distribution. In both cases mean will be higher than median and median will be appropriate for calculation. Standard deviation is huge different because of range of data points.

At the end we can conclude, looking at all the calculations, I think there is quite missing or incorrect observation in the data. Although R calculation might say they are outliers, we do not have any evidence to tell for sure. Since living in NY, I know it is not impossible for someone to rent at 3000 or 5000 in Manhattan for some occasion because someone may pay huge amount during New Year's Eve. If person really wanted to spend specific occasion in NYC. Hence let's consider this data is not corrupt for now and all the calculation as valid.

❖ Applying Central Limit Theorem on price variable in our dataset

Central Limit Theorem states that distribution of sample means of several samples will always have normal distribution regards of actual distribution of original data. Here we have right skewed distribution for price variable, so to show CLT we will be taking 10000 samples of size 500, 1000, 1500 and 2000. Another thing about CLT is that mean of this sample will most likely be around actual mean.



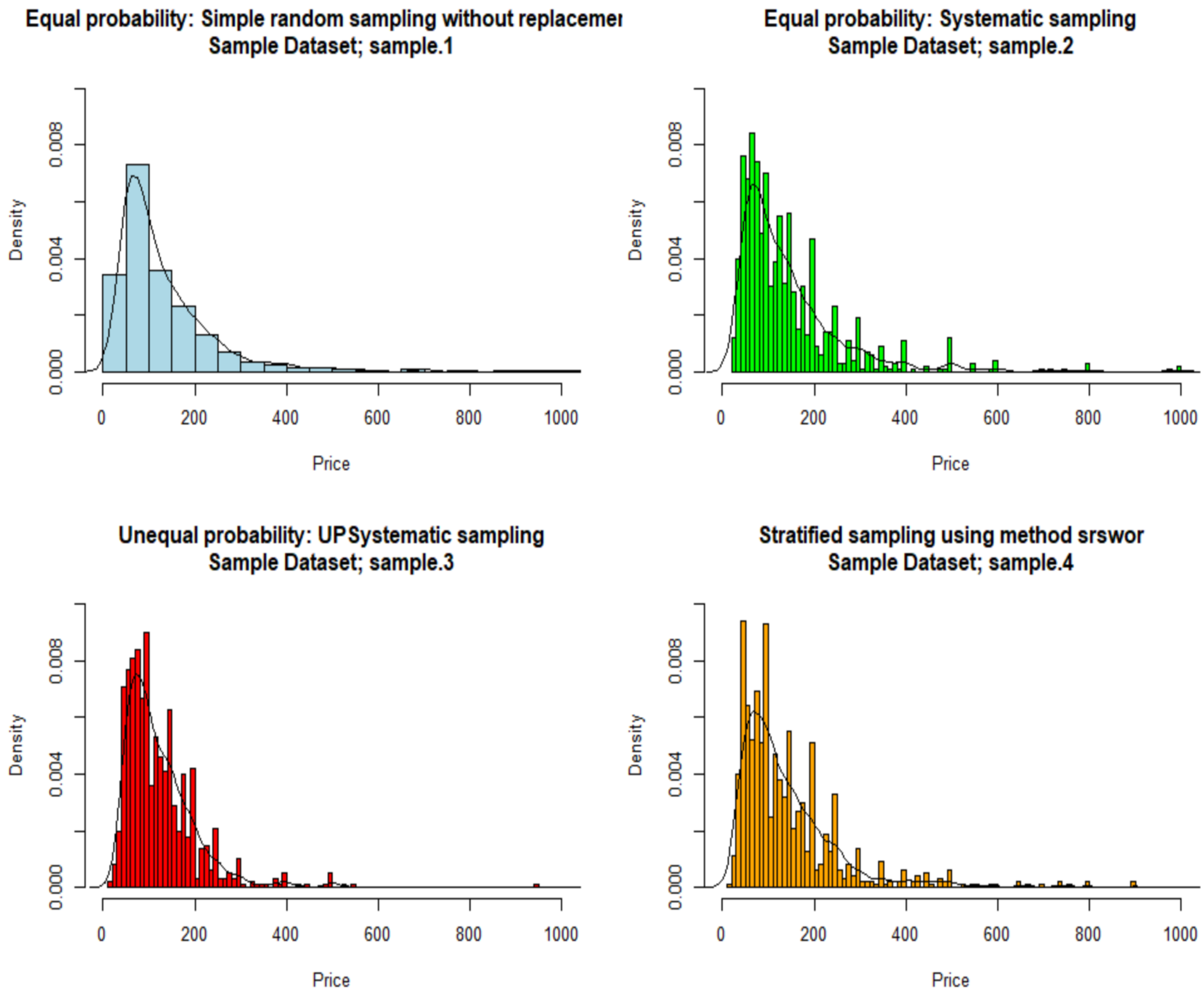

```
    Actual Mean = 152.7  Actual SD = 240.1542
Sample Size =   500  Mean =  152.6708   SD =  10.70101
Sample Size =  1000  Mean =  152.767    SD =  7.61398
Sample Size =  1500  Mean =  152.69     SD =  6.085857
Sample Size =  2000  Mean =  152.7408   SD =  5.212373
```

As we can see that all means are around 152.7 which is our actual mean, as the samples start increasing, it is getting more and more normal, that is characteristics of CLT. As it can be noticed that as more sample is going to give us more normal distribution and we can also see that our Standard deviation is also changed very much and that is because of CLT. SD of sample will be around the SD of actual dataset over square root of sample size. For example, our sample is 1000 then root will be 31.63, So 240.15 divided by 31.63 will be 7.59, which is quite near to 7.62 that is SD we got for 1000 sample. Hence, Central Limit Theorem holds true for our price variable in this dataset.

❖ Applying different Sampling Methods

There are different sampling methods that you can apply to get smaller dataset, which we could easily able to understand dataset behavior. There is Simple random sampling method, Systematic sampling, and Stratified sampling. For first sample I am using simple random sampling method without replacement, in this every member(row) has equal probability so 1/n, here we will have 1/48895. Second set of sample dataset is done by using systematic sampling with equal probability, so in here it will make groups and number of groups will be same to size of sample. The first element will be randomly chosen from first set and then getting every element systematically from every group. For example, I am taking sample of 1000, since our data members are around 49000, so first element will be chosen randomly from first 49 and then after every 49^{th} element will be selected. Third data sample is done using systematic sampling but with unequal probability, so here getting selected is using the probability of a specific variable. Hence you don't have equal probability getting selected for each member and sum of these probabilities will be sample size. Last and fourth sample dataset is done by using stratified sampling with equal probability, here data is divided into subgroup of some categorial variable, and then sample is selected from those subgroups but making sure of having same proportion of categorial variable as actual dataset. For our stratified sample dataset, we will be using categorial variable neighborhood group, that is our NYC borough in our data. Below is histograms different sample and their Mean and SD,

"Sample Dataset; Original	Mean = 152.7	SD = 240.15	Min = 0	Max = 10000"
"Sample Dataset; sample.1	Mean = 153.5	SD = 271.28	Min = 0	Max = 6500"
"Sample Dataset; sample.2	Mean = 147.0	SD = 128.00	Min = 20	Max = 1500"
"Sample Dataset; sample.3	Mean = 129.3	SD = 112.80	Min = 16	Max = 2000"
"Sample Dataset; sample.4	Mean = 151.7	SD = 165.70	Min = 10	Max = 2545"

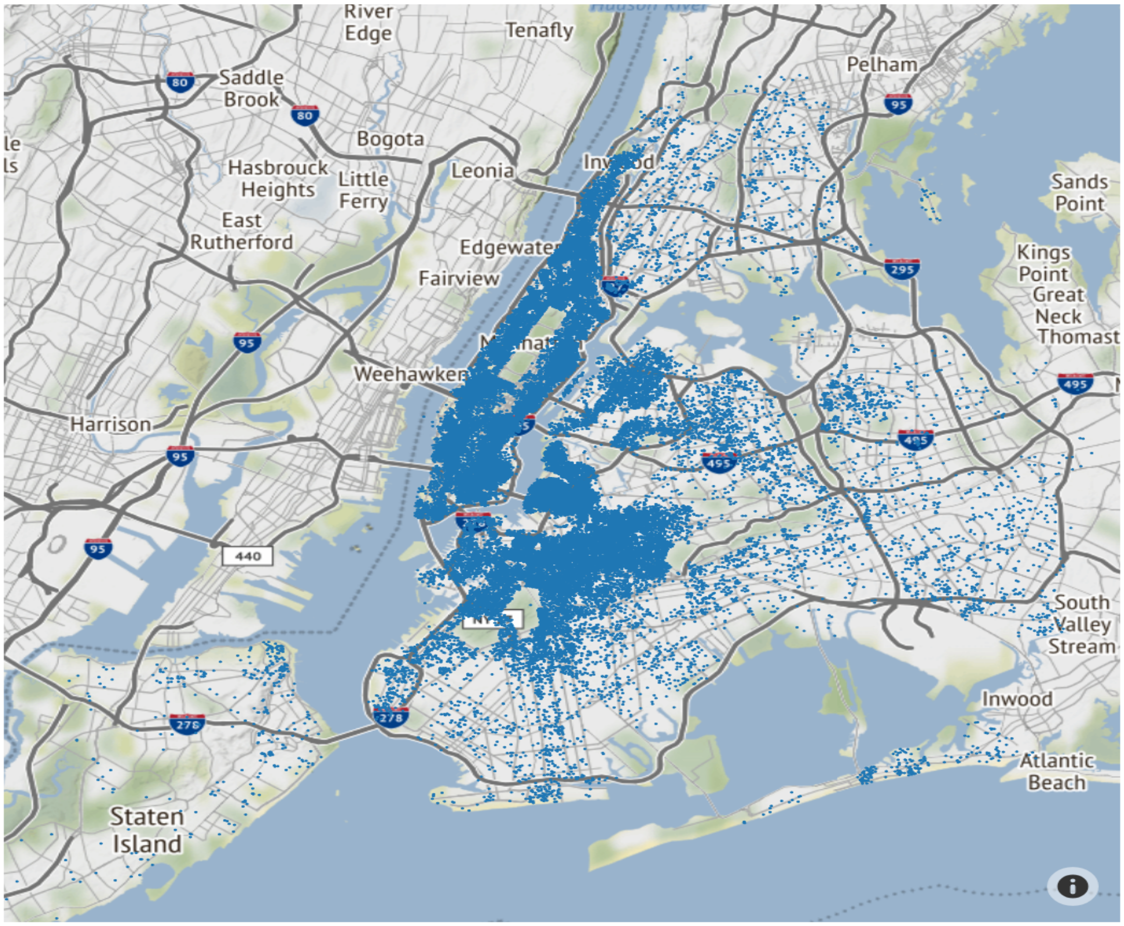


Here, our data is looks normal but still right skewed, from graph and values we get it is quite often if we use study sample dataset. It will be quite easy to understand and mean of sample is also around our actual mean. Hence this shows us that studying sample dataset can also be useful rather than studying whole population, since sample will have same characteristics as population.

❖ Using Plotly package in R to showcase points over NYC map

➢ <u>First plotly we will be having is basic scatterplot over NYC map</u>

2019 Airbnb Listing's rental precise location with Price



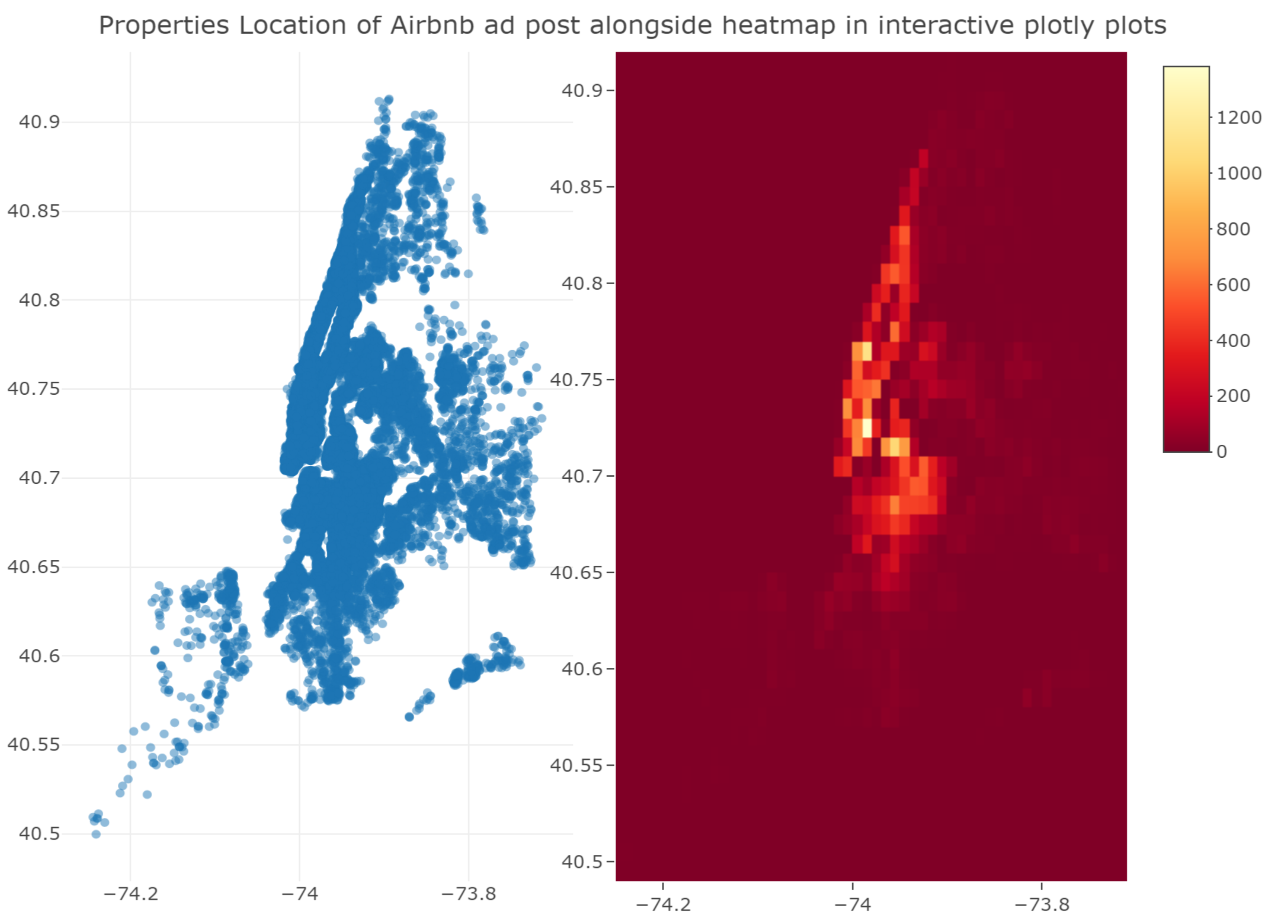
It is actual longitude and latitude of listing space provided in data set, as we can see that blank space in Manhattan is famous Central Park and around is maximum listing posted. I also have attached another html file of this map data, which should be more interactive environment which will also show price for each point, or you can run the code as below in R.

```
library(plotly)
library(mvtnorm)
fig <- plot_ly( fill = "toself", lon = df$longitude, lat = df$latitude, type = 'scattermapbox',
```

```
                    text = paste("Price : ",df$price), marker = list(size = 2,color = "Light Blue"),
                    fillcolor = 'color'  )
fig <- fig %>%
layout( title = "2019 Airbnb Listing's rental precise location with Price", mapbox =
                    list( style = "stamen-terrain", center = list(lon = mean(df$longitude), lat =
                    mean(df$latitude)), zoom = 9.5), showlegend = TRUE)
fig
```

- <u>Another Plotly graph it is heat map alongside regular scatterplot</u>

This is just to get knowledge of how location is a huge factor in our data



As conclusion for whole analysis, Airbnb rental are mostly listed around Downtown Manhattan or Brooklyn. At averagely cost of around 150-200$ for Airbnb rental in NYC.