Exploratory Data Analysis and Visualization of Airbnb Dataset

Introduction

Below is our exploratory data analysis and visualizations and some of interesting insights into the Airbnb data. We have focused on Newyork's airbnb data covering five boroughs, Manhattan, Brooklyn, Bronx, Queens and Staten Island to analyse the most expensive and popular locality in new York city.

Below are questions that we aim to answer through our analysis:

- How do prices of property vary by location?
- Are the demand and prices of the properties related anyway?
- What are the different types of properties in NYC? Do they vary by neighborhood?
- What localities in NYC are rated highly by guests?

Dealing with Missing Values

The data also had some missing values. To preserve all the information, we imputed or dropped the rows and columns containing null values while conducting exploratory.

For checking missing values in our data, we have used ***is.na()*** function as below:

```
table(is.na(airbnbOriginalDF))

table(is.na(airbnbOriginalDF$reviews_per_month))
```

With this we see that all our missing values are present in reviews per month. So we will be removing rows which have missing values..

```
airbnbNoNADT = airbnbOriginalDF[airbnbOriginalDF$reviews_per_month != 'NA']
```

Now when we check the table, we will not get any missing value.

```
table(is.na(airbnbNoNADT))
```

Converting Data Types

The column “last_review” is character type. We will convert its data type into appropriate date format.

```
airbnbNoNADT[,last_review:=as.Date(last_review, '%m/%d/%Y')]
```

Factors:

Lets try to further analyze our data by analysing data types.

```
str(airbnbNoNADT)
```

```
unique(airbnbNoNADT$neighbourhood_group)
```

As the neighbourhood_group column has 5 categorical values, we can factor it, and convert our char data type.

```
airbnbNoNADT[,neighbourhood_group:= factor(neighbourhood_group)]
```

Similarly, for neighbourhood, we get 217 unique values. Here to reduce storage we can covert all similar type to lower case and also trim white spaces, so that each name is unique.

```
unique(airbnbNoNADT $neighbourhood)
```

```
airbnbNoNADT [,neighbourhood:=tolower(neighbourhood)]
```

```
airbnbNoNADT [,neighbourhood:=trimws(neighbourhood)]
```

We will refrain ourselves from factoring it at this stage, as we can do it later when required.

Similarly, for room type, we get 3 unique categorical values. We can factor it, and convert our string datatype.

```
unique(airbnbNoNADT $room_type)
```

```
airbnbNoNADT [,room_type:= factor(room_type)]
```

Analyzing Outliers

For this we have first tried finding summary of our numerical value columns.

Summary for NUMERIC column will give us six values:

Minimum Value

1st Quartile

2nd Quartile

Median

Mean

3rd Quartile

Maximum

The results will give us insight about data values deviation from mean.

1.Latitude and longitude data

When we check summary for latitude and longitude, we see that data is evenly spread with no outlier.

summary(airbnbNoNADT$longitude)

str(airbnbNoNADT)

2. Analysing availability data. The data is fair and no extreme values.

summary(airbnbNoNADT$availability_365)

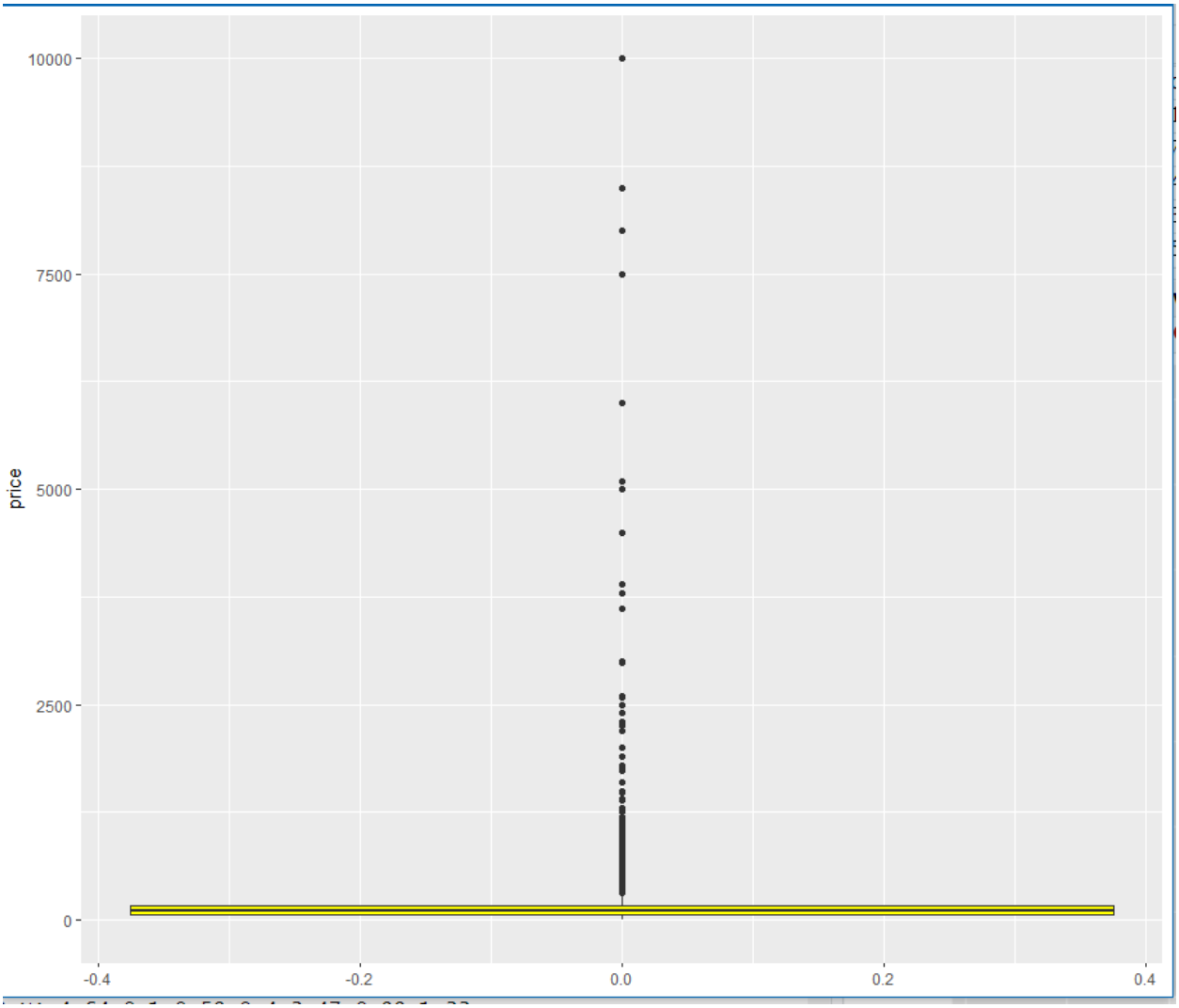
3. Analyze price data.

We could see extremely large values. Lets draw a plot to see the distribution.

summary(airbnbNoNADT$price)

ggplot(airbnbNoNADT,aes(y=price))+geom_boxplot(fill='yellow')

In plot we can see some outliers. Lets run below and see how many are such properties that have price greater than 2500.



nrow(airbnbNoNADT[price>2500])

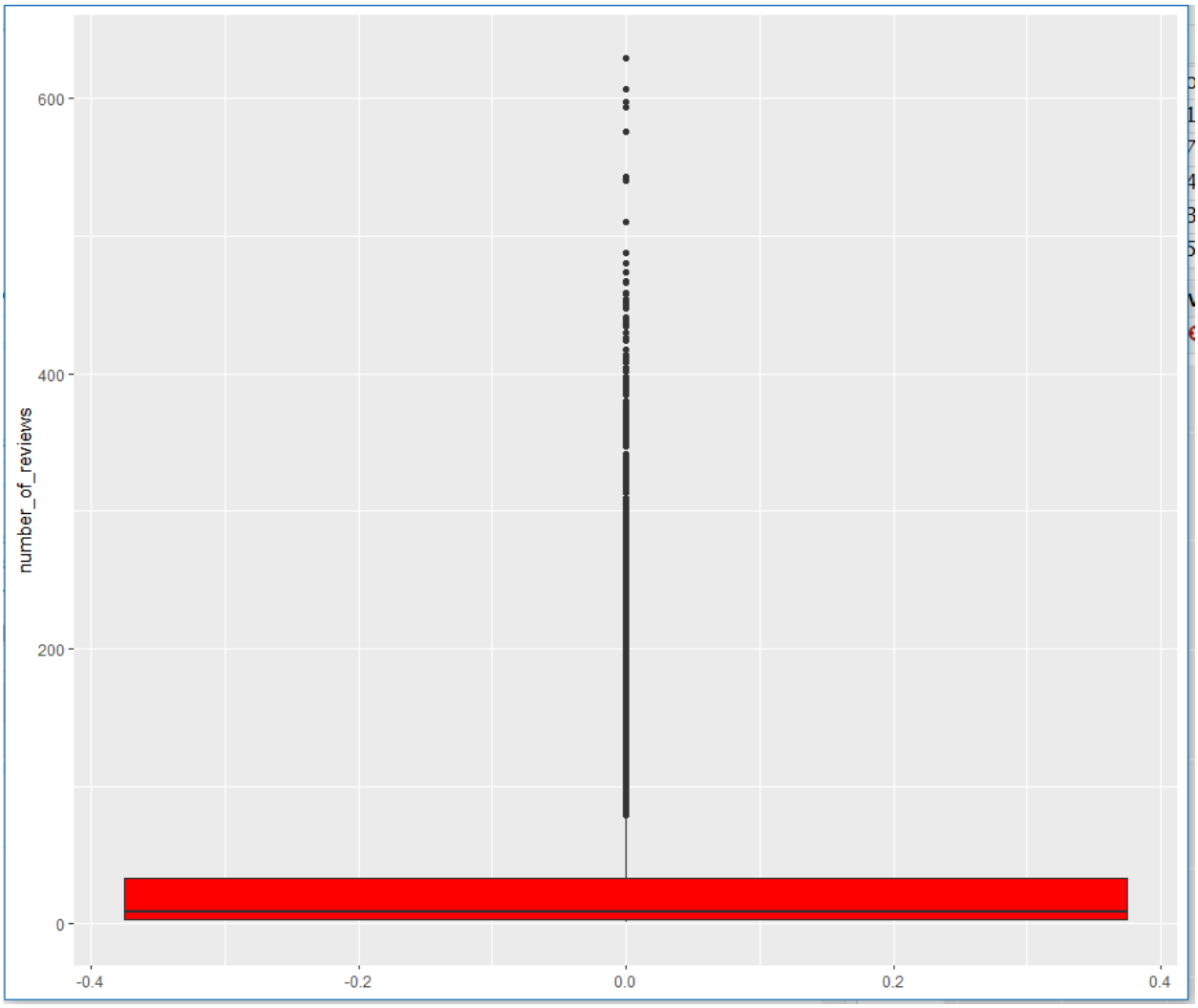
By runing this, we find only 25 such properties. This can be dropped as we 38k plus data

4. Analysing number of reviews data.

We could see extremely large values. Lets draw a plot to see the distribution.

summary(airbnbNoNADT$number_of_reviews)

ggplot(airbnbNoNADT,aes(y=number_of_reviews))+geom_boxplot()



In plot we can see some outliers. Lets run below and see how many are such properties that have number of reviews greater than 400. Such a huge review for one or two property seems to be some spam or fake. We shall how many such rows are there in our data.

nrow(airbnbNoNADT[number_of_reviews>400])

We found 39 rows which have number of reviews greater than 400.

airbnbNoNADT[number_of_reviews>400,unique(neighbourhood_group)]

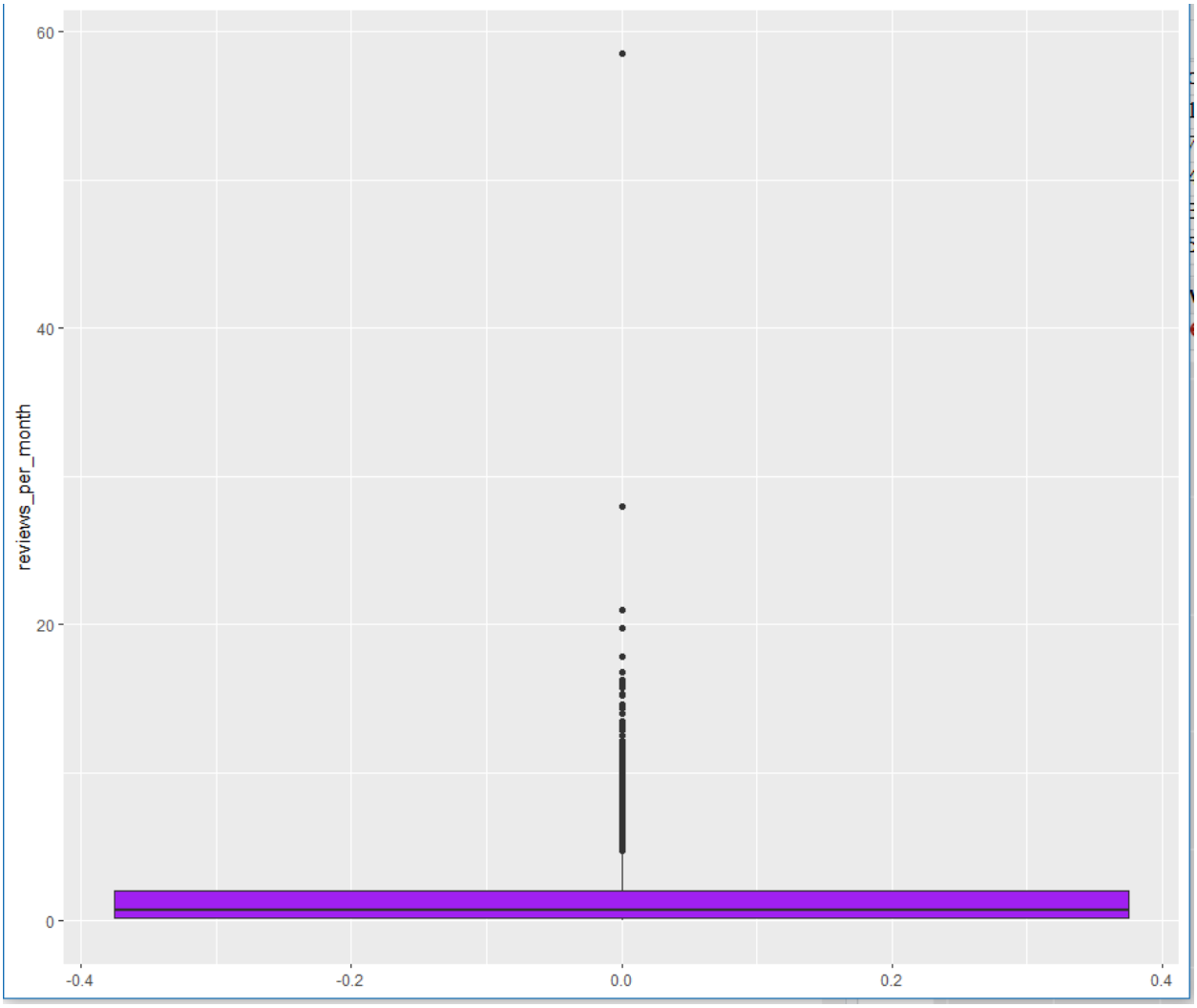
When we checked for which areas this spam review is , it shows Manhattan, Brooklyn and Queens. So there is no clear indication by this data, we will drop this to further clean our data and remove outliers.

5. Analysing reviews per month

Could see extremely large values. Lets draw a plot to see the distribution.

summary(airbnbNoNADT$reviews_per_month)

ggplot(airbnbNoNADT,aes(y=reviews_per_month))+geom_boxplot()



In plot we can see some outliers. lets run below and see how many are such properties that have reviews per month greater than 10.

Most of the data is located below 5. We shall how many such rows are there in our data which have review per month greater than 10

nrow(airbnbNoNADT[reviews_per_month>10])

airbnbNoNADT[reviews_per_month>10,unique(neighbourhood_group)]

When we tried checking if any particular locality has more reviews, it does not give any indication. The result is spread out for all localities. We can drop this rows, as it will not yield anything peculiar.

With above summary and plot we found few ouliers, therefore we have dropped that data, assuring it is not impacting our main dataset.

airbnbCleaned = airbnbNoNADT[price<2500 & number_of_reviews<400 & reviews_per_month<10]

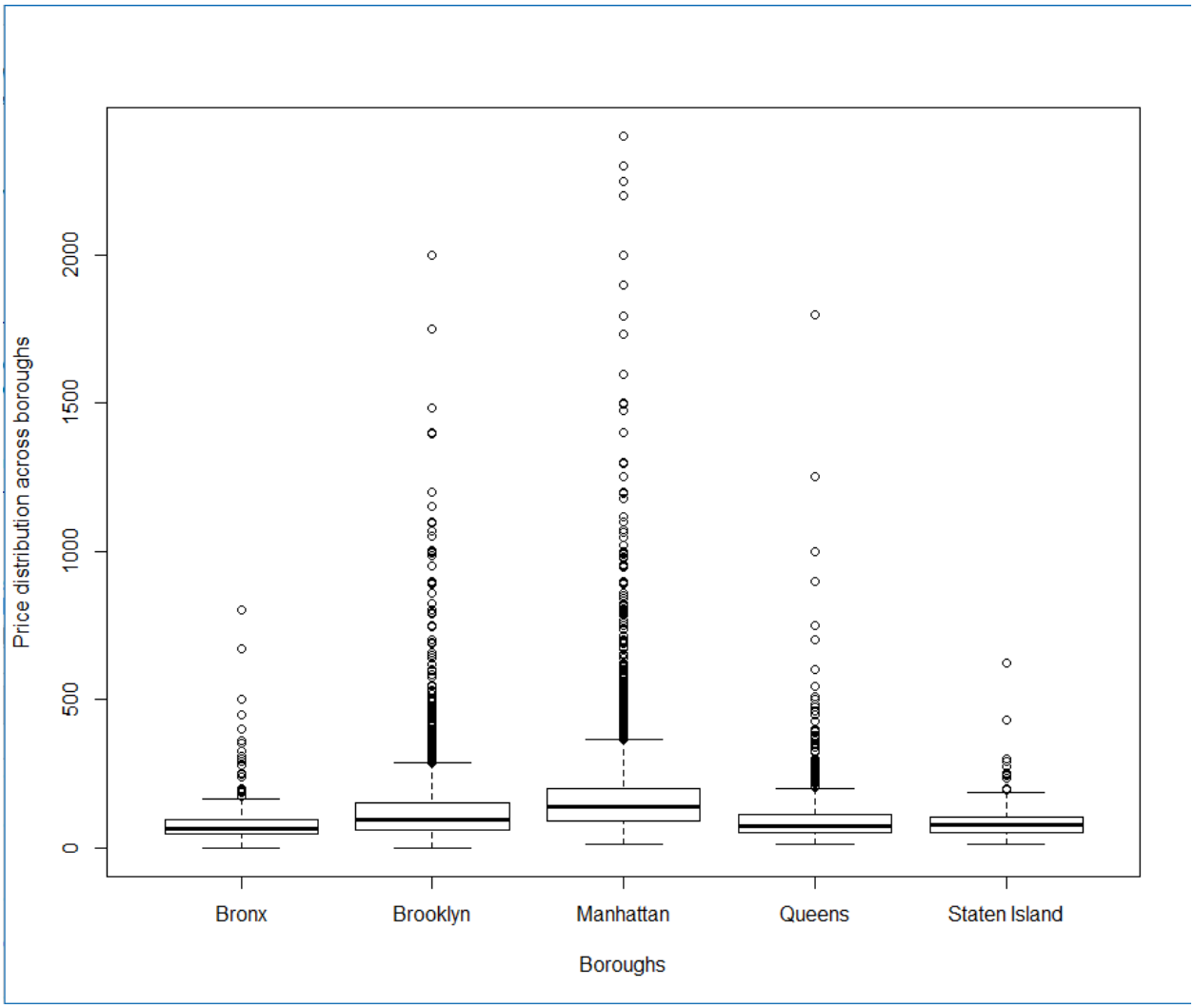
airbnbCleaned is our Final cleaned dataset.

Demand and Price Analysis

- **How do prices of property vary by location?**

Analysing the price distribution based on location

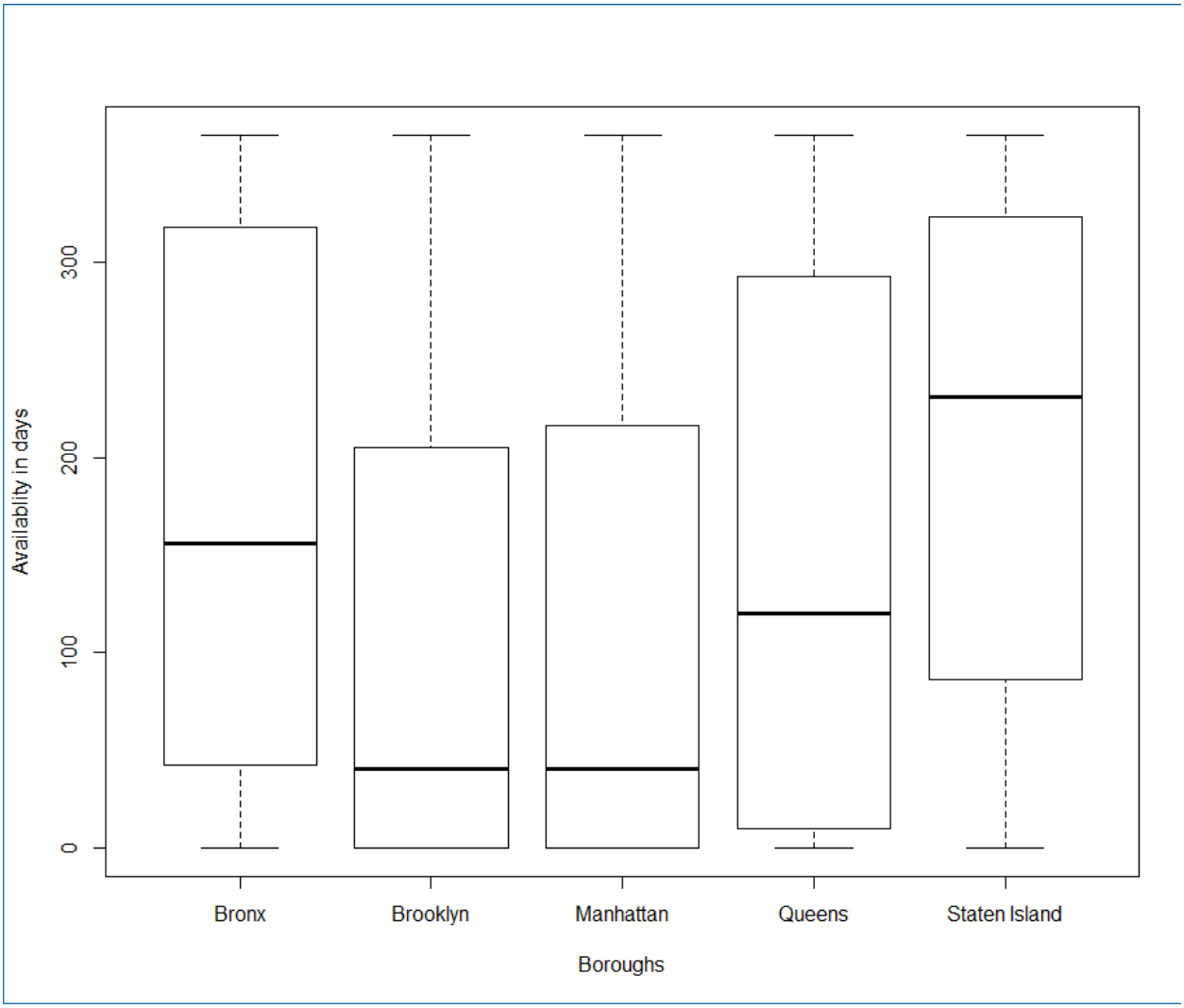
plot(neighbourhood_group,price, xlab= 'Boroughs', ylab='Price distribution across boroughs')



This gives us insight that Manhattan has the most highly prices properties in NYC, and that's true since Manhattan is expensive among all, followed by Brooklyn.

Analysing the availability across borouhgs

plot(airbnbCleaned$neighbourhood_group, airbnbCleaned$availability_365



- **Are the demand and prices of the properties related anyway?**

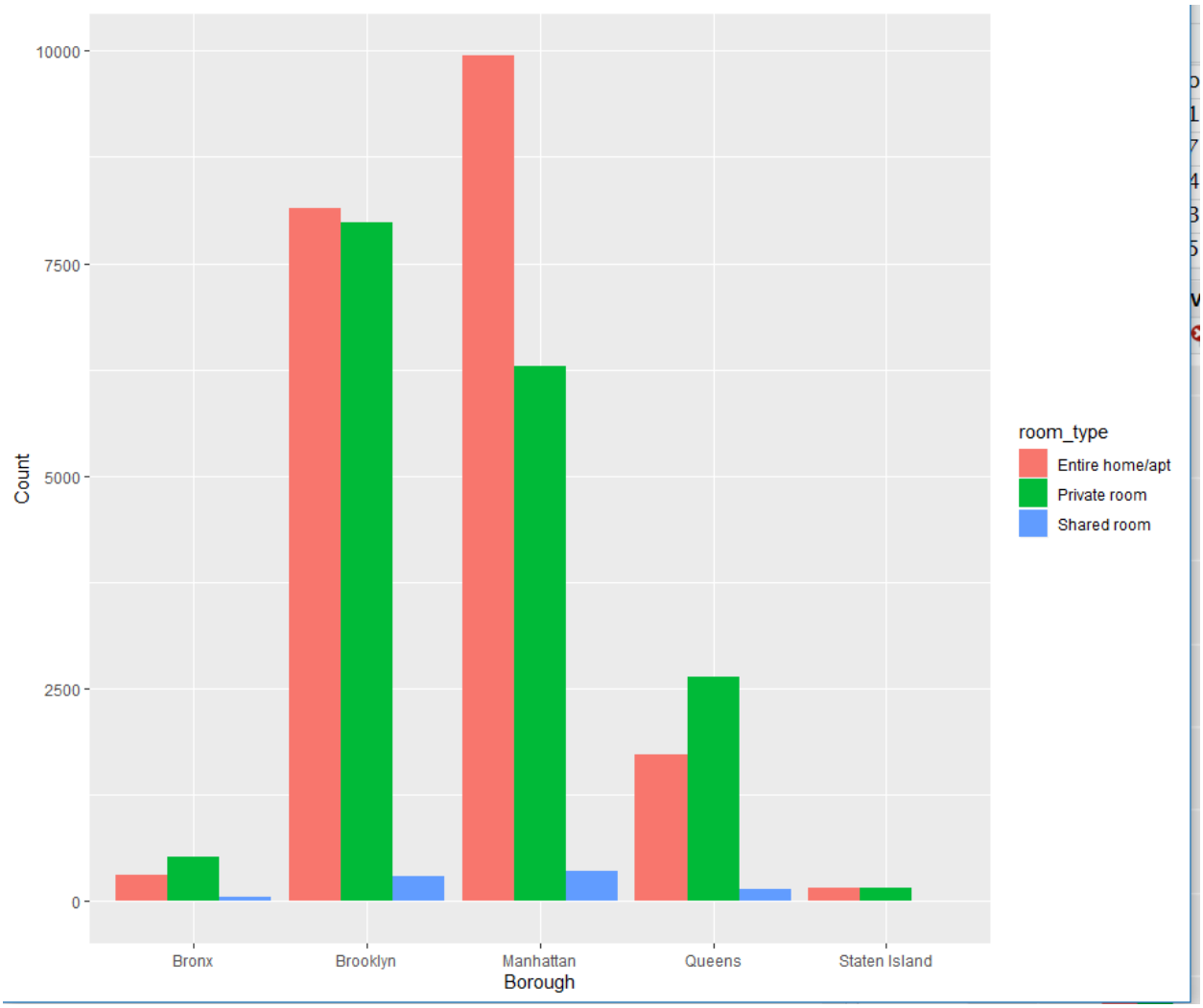
This plot gives us insight that the least available properties are located in Manhattan and Brooklyn.

This shows that the areas where demand is more (less availability) have higher prices as seen in above plot.

- **What are the different types of properties in NYC? Do they vary by neighborhood?**

  Analysing the room types which are prominent among boroughs.

  ggplot(airbnbCleaned, aes(x=neighbourhood_group, fill = room_type))+geom_bar(position = "dodge") + xlab("Borough") + ylab("Count")

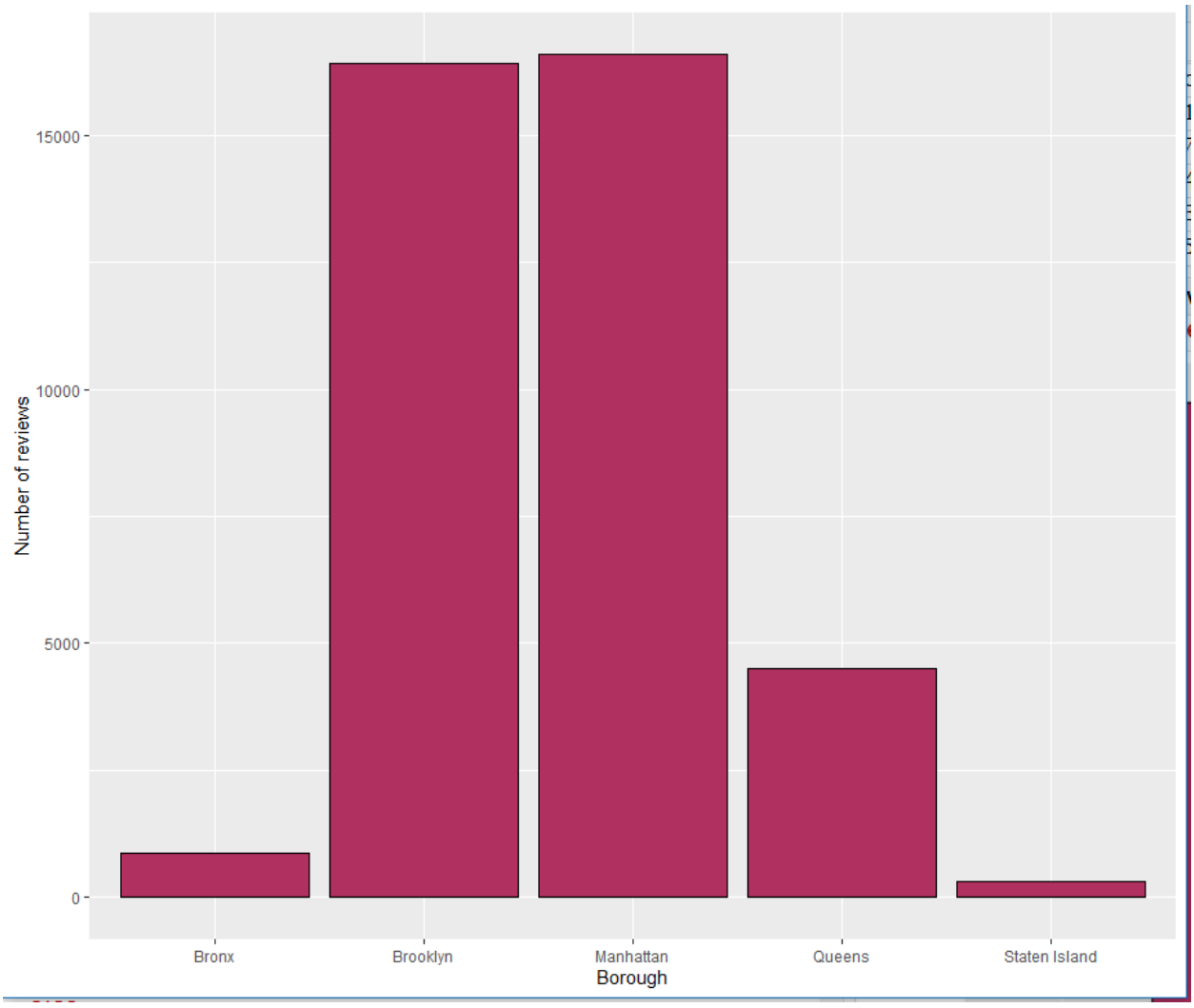


From the above,We can see that Entire home apartment listings are highest in number except Queens and Bronx. Queens has more 'Private' style property than 'Apartments'.

The maximum apartment style listings are located in Manhattan, constituting 90% of all properties in that neighborhood. Next is Brooklyn with 75% Apartment style listing.

- **What localities in NYC are rated highly by guests?**

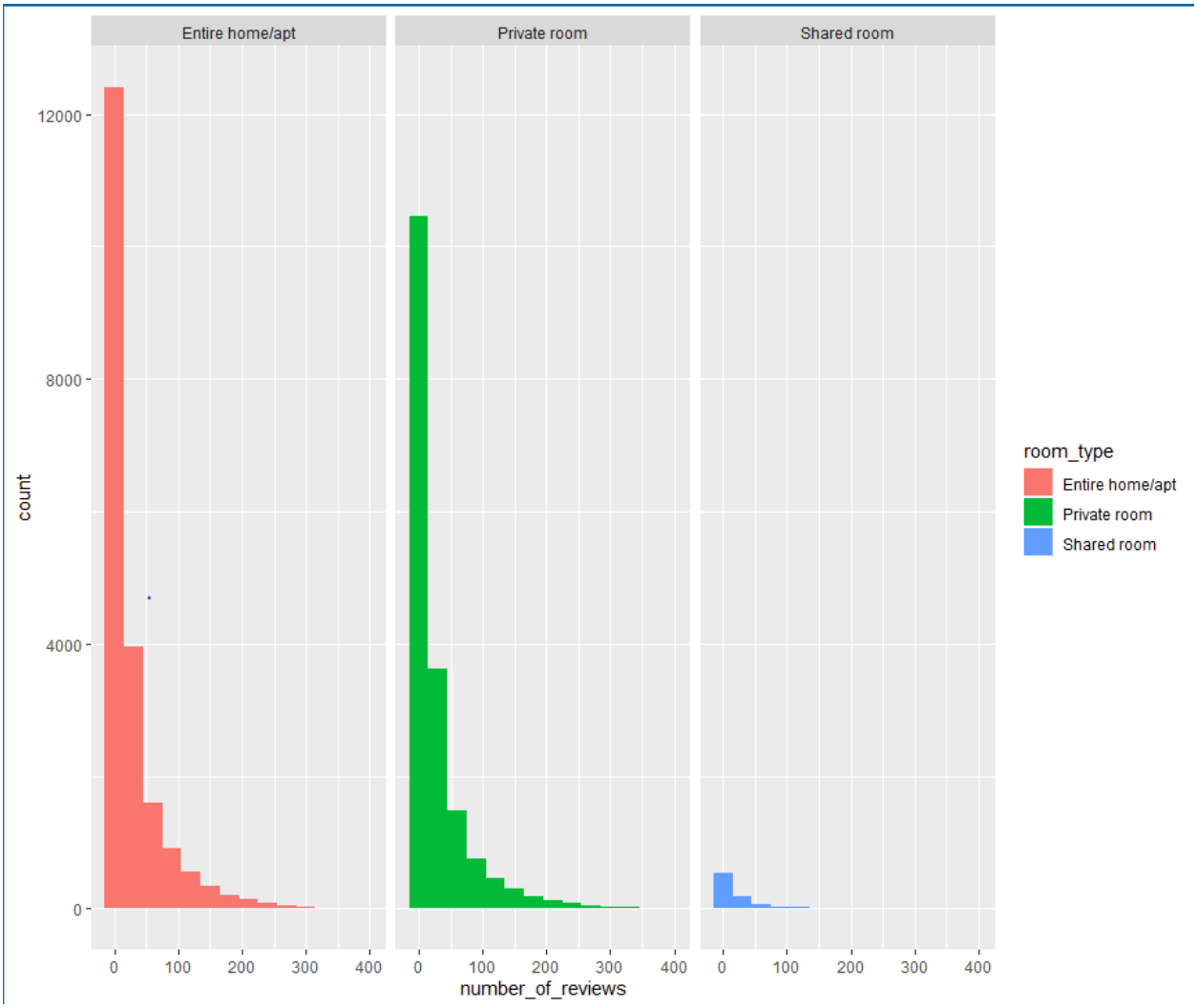
```
ggplot (airbnbCleaned, aes(x= neighbourhood_group, fill =
number_of_reviews))+geom_bar(color='black', fill='maroon') + xlab("Borough") +
ylab("Number of reviews")
```



This gives us clear insight that most preferred location as per review received is for properties listed in Manhattan, followed by Brooklyn, and then Queens, and then Bronx. Staten Island has least nuber of reviews.

- **What type of property in NYC are preferred or highly rated by guests?**

  ggplot(airbnbCleaned, aes(x= number_of_reviews, fill= room_type)) +
  geom_histogram(binwidth = 30)+facet_wrap(room_type)



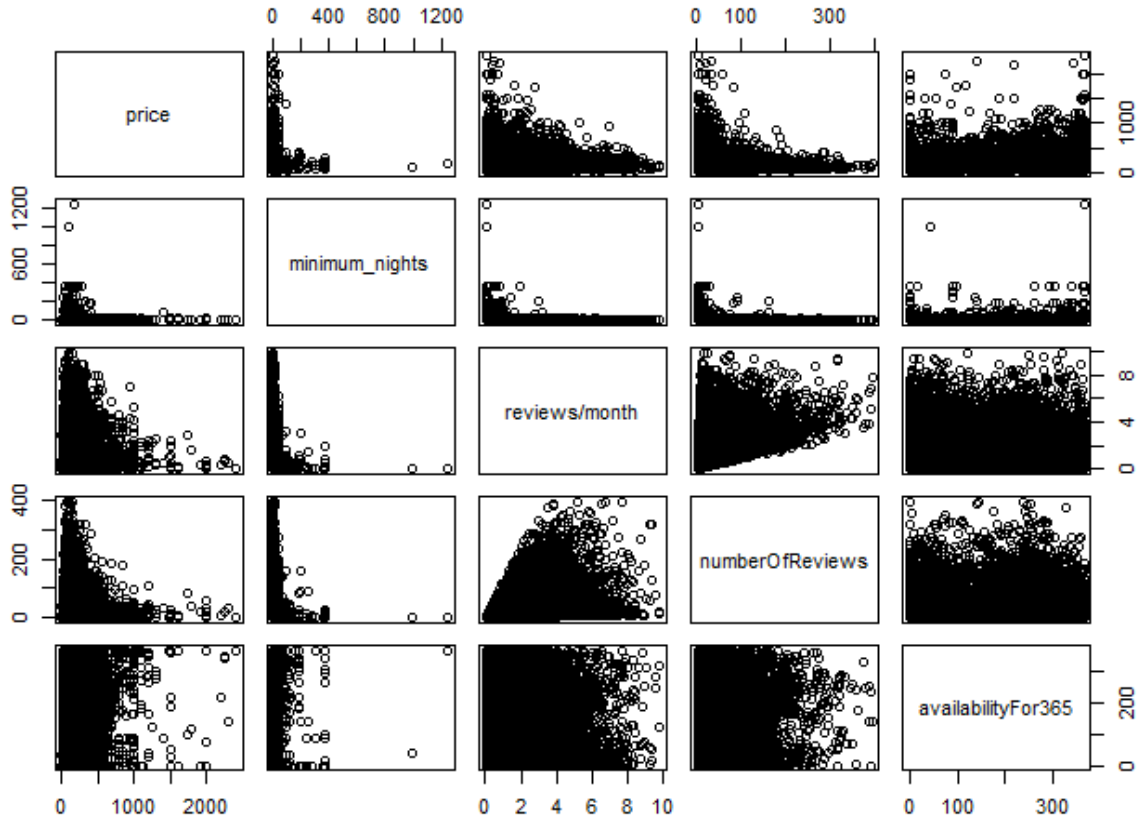
With above data, we can see that Apartment type properties are mostly preferred, since they are the ones receiving maximum ratings. After which people prefer private rooms. Shared rooms have received very few rating. This would be helpful for other business to avoid providing shared rooms.

Correlations

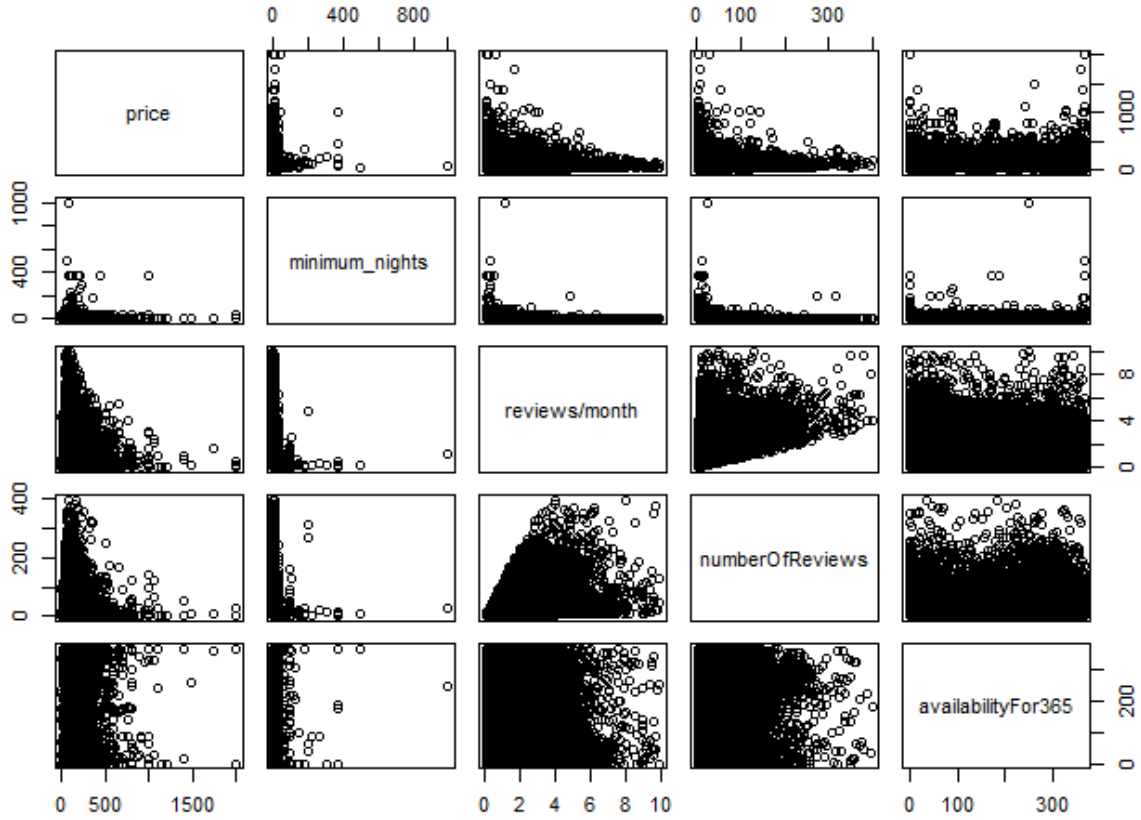
After cleaning the data and putting the outliers off we did some initial analysis on the different variables and how they can potentially be related. Also, we tried to come up with some initial facts just to highlight the potential things we can do by analyzing the dataset. Here we show the correlations across

various variables and how they can be related. For this correlation we have separated the data across boroughs since we think different factors might affect different variables in localities. The plots below show the relationship between some significantly important variables.

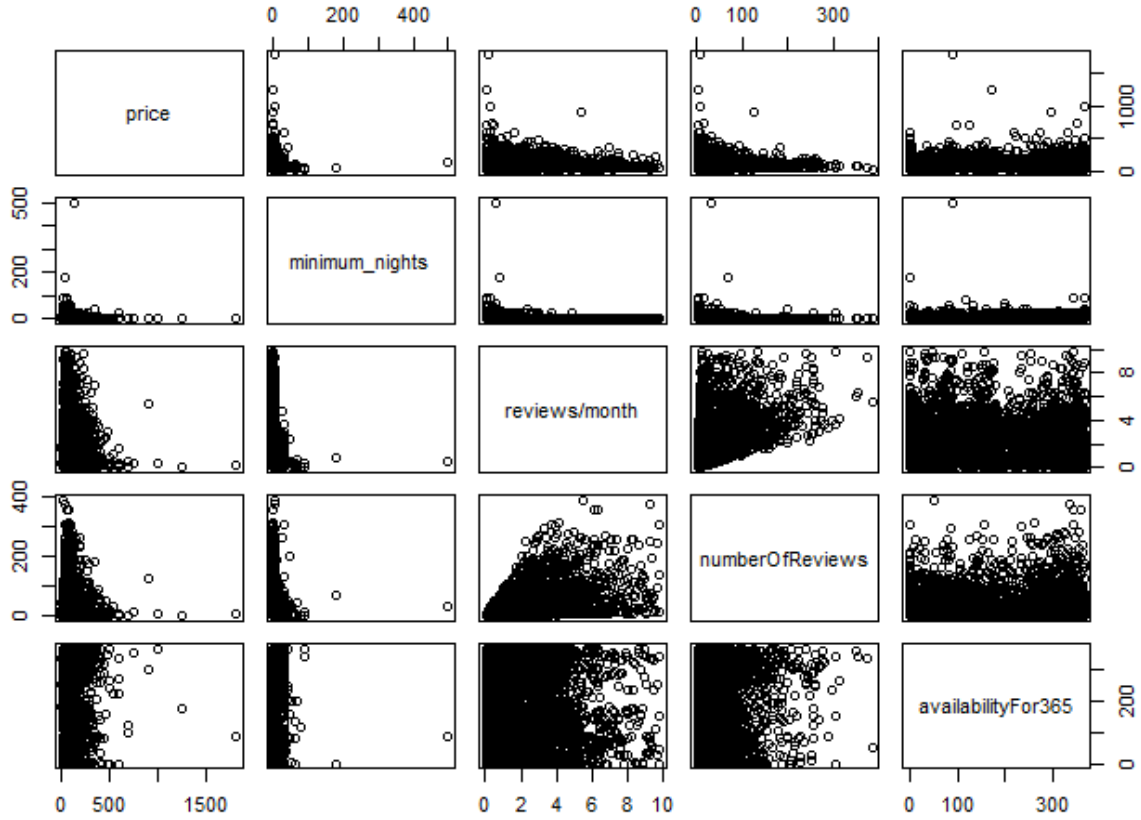
- **Manhattan**



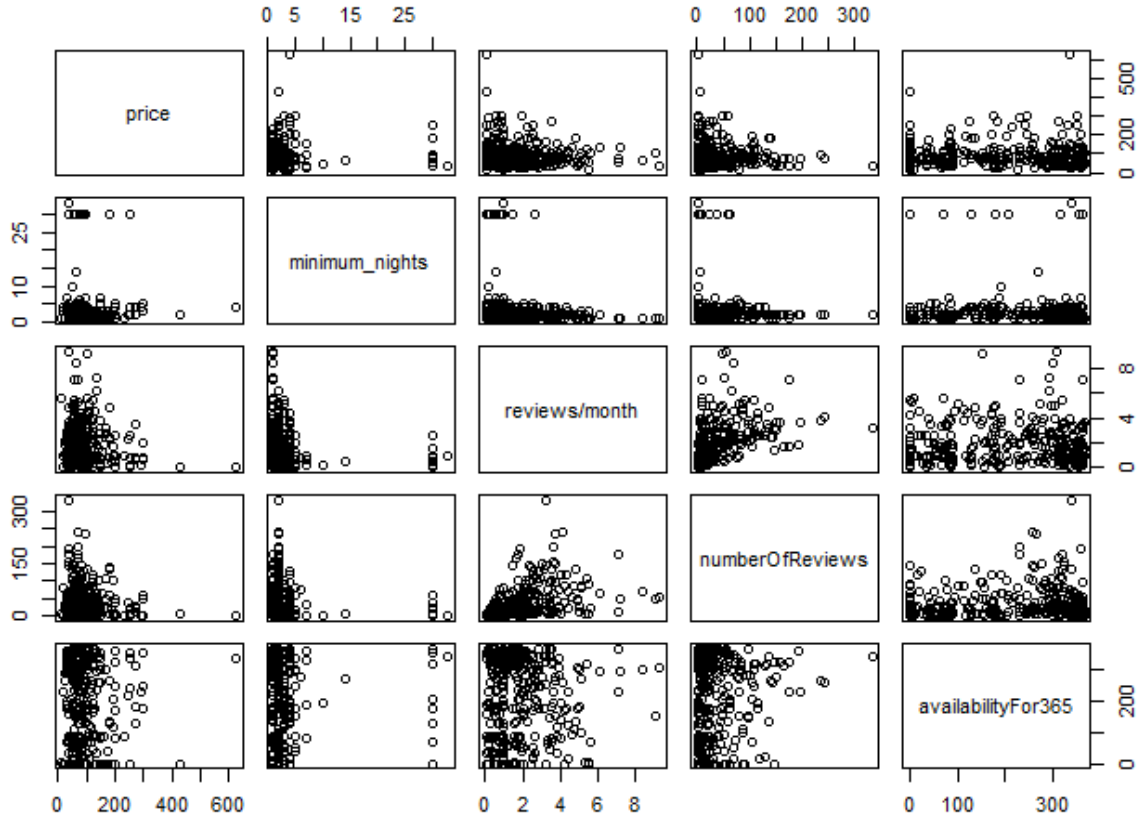
- **Brooklyn**


price
minimum_nights
reviews/month
numberOfReviews
availabilityFor365

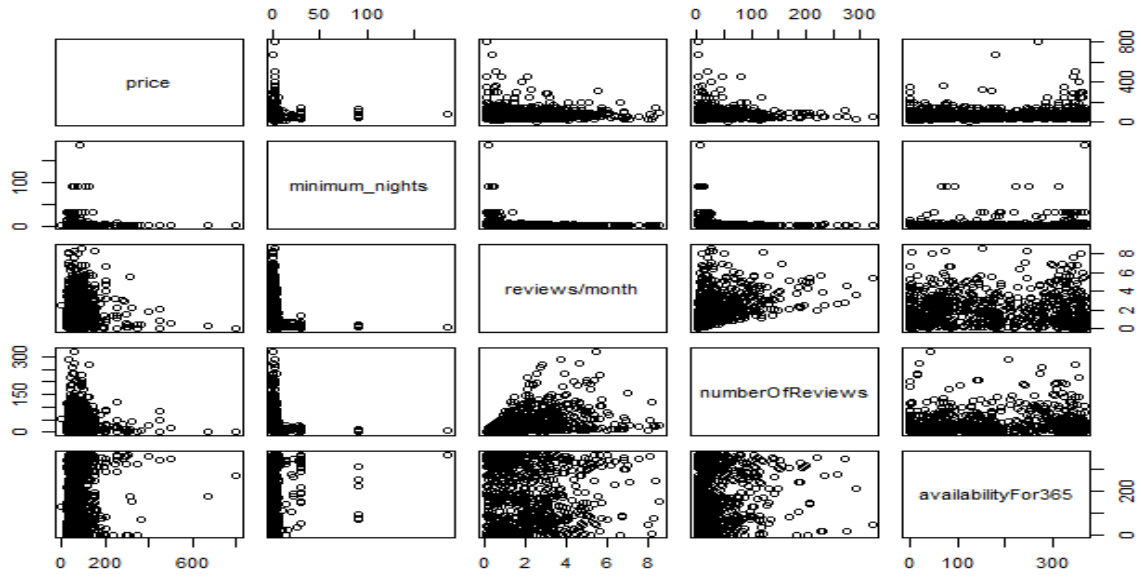
- **Queens**


price
minimum_nights
reviews/month
numberOfReviews
availabilityFor365
0 200 400
0 100 300
0 500 1500
0 2 4 6 8 10
0 100 300
0 200 400
0 200 500
0 4 8
0 200 1000
0 200

- **Staten Island**



- **Bronx**



- Let's also see some other interesting correlations
- Correlation between

```
> cor(airbnbCleaned[,c("host_id","reviews_per_month","availability_365")])
                  host_id reviews_per_month availability_365
host_id         1.0000000         0.3004072       0.1551920
reviews_per_month 0.3004072       1.0000000       0.1912579
availability_365  0.1551920       0.1912579       1.0000000
```

SUMMARY

```
> summary(airbnbCleaned)
       id                 name             host_id           host_name
 Min.   :    2539   Length:38697       Min.   :     2438   Length:38697
 1st Qu.: 8720865   Class :character   1st Qu.:  7033731   Class :character
 Median :18857134   Mode  :character   Median : 28307663   Mode  :character
 Mean   :18088961                      Mean   : 64067108
 3rd Qu.:27543590                      3rd Qu.:101493347
 Max.   :36455809                      Max.   :273841667
   neighbourhood_group neighbourhood        latitude       longitude
 Bronx        :  875    Length:38697       Min.   :40.51   Min.   :-74.24
 Brooklyn     :16421    Class :character   1st Qu.:40.69   1st Qu.:-73.98
 Manhattan    :16584    Mode  :character   Median :40.72   Median :-73.95
 Queens       : 4504                       Mean   :40.73   Mean   :-73.95
 Staten Island:  313                       3rd Qu.:40.76   3rd Qu.:-73.94
                                           Max.   :40.91   Max.   :-73.71
          room_type         price        minimum_nights     number_of_reviews
 Entire home/apt:20269   Min.   :   0.0   Min.   :   1.000   Min.   :  1.00
 Private room   :17586   1st Qu.:  69.0   1st Qu.:   1.000   1st Qu.:  3.00
 Shared room    :  842   Median : 102.0   Median :   2.000   Median :  9.00
                         Mean   : 138.7   Mean   :   5.872   Mean   : 28.71
                         3rd Qu.: 170.0   3rd Qu.:   4.000   3rd Qu.: 33.00
                         Max.   :2400.0   Max.   :1250.000   Max.   :398.00
  last_review         reviews_per_month calculated_host_listings_count availability_365
 Min.   :2011-03-28   Min.   :0.010     Min.   :  1.000                 Min.   :  0.0
 1st Qu.:2018-07-07   1st Qu.:0.190     1st Qu.:  1.000                 1st Qu.:  0.0
 Median :2019-05-19   Median :0.710     Median :  1.000                 Median : 55.0
 Mean   :2018-10-03   Mean   :1.344     Mean   :  5.173                 Mean   :114.6
 3rd Qu.:2019-06-23   3rd Qu.:2.000     3rd Qu.:  2.000                 3rd Qu.:229.0
 Max.   :2019-07-08   Max.   :9.970     Max.   :327.000                 Max.   :365.0
```

Above picture shows a quick summary of various variables and the quality of data after cleaning. We know that airbnb is a big player when it comes to renting and tourism business. Therefore a quick EDA on this dataset shows the quality of data. Also, it shows how different variables are distributed and how airbnb data can be used to come up with more interesting insights which can be useful to both airbnb to expand its business in NYC and to customers to get an insight on airbnb properties. Such insight can be used by different customers to help them in their choices. Below figure shows, airbnb's presence in NYC.

TESTS:

Here we have performed two tests , T test and Levene test for process against different borughs.

T –Test for prices across different boroughs:

1. **Manhattan vs Brooklyn**

with(data=airbnbCleaned,t.test(price[neighbourhood_group=="Manhattan"],price[neighbourhood_group=="Brooklyn"],var.equal=TRUE))

Result:

```
Two Sample t-test

data:  price[neighbourhood_group == "Manhattan"] and price[neighbourhood_grou
p == "Brooklyn"]
t = 39.869, df = 33003, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 53.53904 59.07535
sample estimates:
```

```
mean of x mean of y
 174.9481  118.6409
```

2. **Queens vs Bronx**

```
with(data=airbnbCleaned,t.test(price[neighbourhood_group=="Queens"],price[neighbourhood_group==
"Bronx"],var.equal=TRUE))
```

```
        Two Sample t-test

data:  price[neighbourhood_group == "Queens"] and price[neighbourhood_group =
= "Bronx"]
t = 5.1808, df = 5377, p-value = 2.291e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.690078 19.270338
sample estimates:
mean of x mean of y
 93.51621  79.53600
```

Levene test

Performed levene test for prices with neighbourhood group

```
leveneTest(price ~ neighbourhood_group, data=airbnbCleaned)
```

Result:

```
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     4  235.07 < 2.2e-16 ***
      38692
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```