DS412LabFinal

Name: Md. Nasif Sarwar

Student ID: 203-35-3133

```
library(ggplot2)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.3.2
```

```
library(geomtextpath)
```

```
## Warning: package 'geomtextpath' was built under R version 4.3.2
```

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.3.2
```

Answer to the question "UNDERSTANDING AND ANALYSIS" I

By analyzing the data we can see that this dataset is not proper for visulization. Most of the feature has outliers and uneven data for instance we can take price mean is 142 and median is 101 but the maximum value is 10000 most of the features are same filed with outliers few of them are okay to use.

```
dataSet <-read.csv("NYCAirBnb.csv")
dataSet <- na.omit(dataSet)
head(dataSet)
```

```
##     id                                              name host_id   host_name
## 1 2539               Clean & quiet apt home by the park    2787        John
## 2 2595                           Skylit Midtown Castle    2845    Jennifer
## 4 3831                 Cozy Entire Floor of Brownstone    4869 LisaRoxanne
## 5 5022 Entire Apt: Spacious Studio/Loft by central park    7192       Laura
## 6 5099        Large Cozy 1 BR Apartment In Midtown East    7322       Chris
## 7 5121                                  BlissArtsSpace!    7356       Garon
##   neighbourhood_group      neighbourhood latitude longitude       room_type
## 1            Brooklyn         Kensington 40.64749 -73.97237    Private room
## 2           Manhattan            Midtown 40.75362 -73.98377 Entire home/apt
## 4            Brooklyn       Clinton Hill 40.68514 -73.95976 Entire home/apt
## 5           Manhattan        East Harlem 40.79851 -73.94399 Entire home/apt
## 6           Manhattan        Murray Hill 40.74767 -73.97500 Entire home/apt
## 7            Brooklyn Bedford-Stuyvesant 40.68688 -73.95596    Private room
##   price minimum_nights number_of_reviews last_review reviews_per_month
## 1   149              1                 9  10/19/2018              0.21
```

```
## 2   225             1               45   5/21/2019             0.38
## 4    89             1              270    7/5/2019             4.64
## 5    80            10                9  11/19/2018             0.10
## 6   200             3               74   6/22/2019             0.59
## 7    60            45               49   10/5/2017             0.40
##   calculated_host_listings_count availability_365
## 1                              6              365
## 2                              2              355
## 4                              1              194
## 5                              1                0
## 6                              1              129
## 7                              1                0
```

```
numeric_df<- dataSet[, !(names(dataSet) %in% c("id", "name","host_id","host_name","neighbourhood_group"
```

```
print("For Numeric Features")
```

```
## [1] "For Numeric Features"
```

```
summary(numeric_df)
```

```
##     latitude       longitude          price         minimum_nights
##  Min.   :40.51   Min.   :-74.24   Min.   :    0.0   Min.   :   1.000
##  1st Qu.:40.69   1st Qu.:-73.98   1st Qu.:   69.0   1st Qu.:   1.000
##  Median :40.72   Median :-73.95   Median :  101.0   Median :   2.000
##  Mean   :40.73   Mean   :-73.95   Mean   :  142.3   Mean   :   5.868
##  3rd Qu.:40.76   3rd Qu.:-73.94   3rd Qu.:  170.0   3rd Qu.:   4.000
##  Max.   :40.91   Max.   :-73.71   Max.   :10000.0   Max.   :1250.000
##  number_of_reviews reviews_per_month calculated_host_listings_count
##  Min.   :  1.0     Min.   : 0.010    Min.   :  1.000
##  1st Qu.:  3.0     1st Qu.: 0.190    1st Qu.:  1.000
##  Median :  9.0     Median : 0.720    Median :  1.000
##  Mean   : 29.3     Mean   : 1.373    Mean   :  5.165
##  3rd Qu.: 33.0     3rd Qu.: 2.020    3rd Qu.:  2.000
##  Max.   :629.0     Max.   :58.500    Max.   :327.000
##  availability_365
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median : 55.0
##  Mean   :114.9
##  3rd Qu.:229.0
##  Max.   :365.0
```

II

Created one function to compute mode of categorical data.

infoCategoricalData - This function takes column name as a parameter and simply print mode and frequency of that feature.

```
infoCategoricalData<- function(columnName){
 cat(sprintf("\n Information of: %s\n", columnName))
print(DescTools::Mode(dataSet[,columnName]))
}
```

```r
print("For Categorical Features")
```

```
## [1] "For Categorical Features"
```

```r
infoCategoricalData('name')
```

```
##
##  Information of: name
## [1] "Home away from home"
## attr(,"freq")
## [1] 12
```

```r
infoCategoricalData('host_name')
```

```
##
##  Information of: host_name
## [1] "Michael"
## attr(,"freq")
## [1] 335
```

```r
infoCategoricalData('neighbourhood_group')
```

```
##
##  Information of: neighbourhood_group
## [1] "Manhattan"
## attr(,"freq")
## [1] 16632
```

```r
infoCategoricalData('neighbourhood')
```

```
##
##  Information of: neighbourhood
## [1] "Williamsburg"
## attr(,"freq")
## [1] 3163
```

```r
infoCategoricalData('room_type')
```

```
##
##  Information of: room_type
## [1] "Entire home/apt"
## attr(,"freq")
## [1] 20332
```

```r
infoCategoricalData('last_review')
```

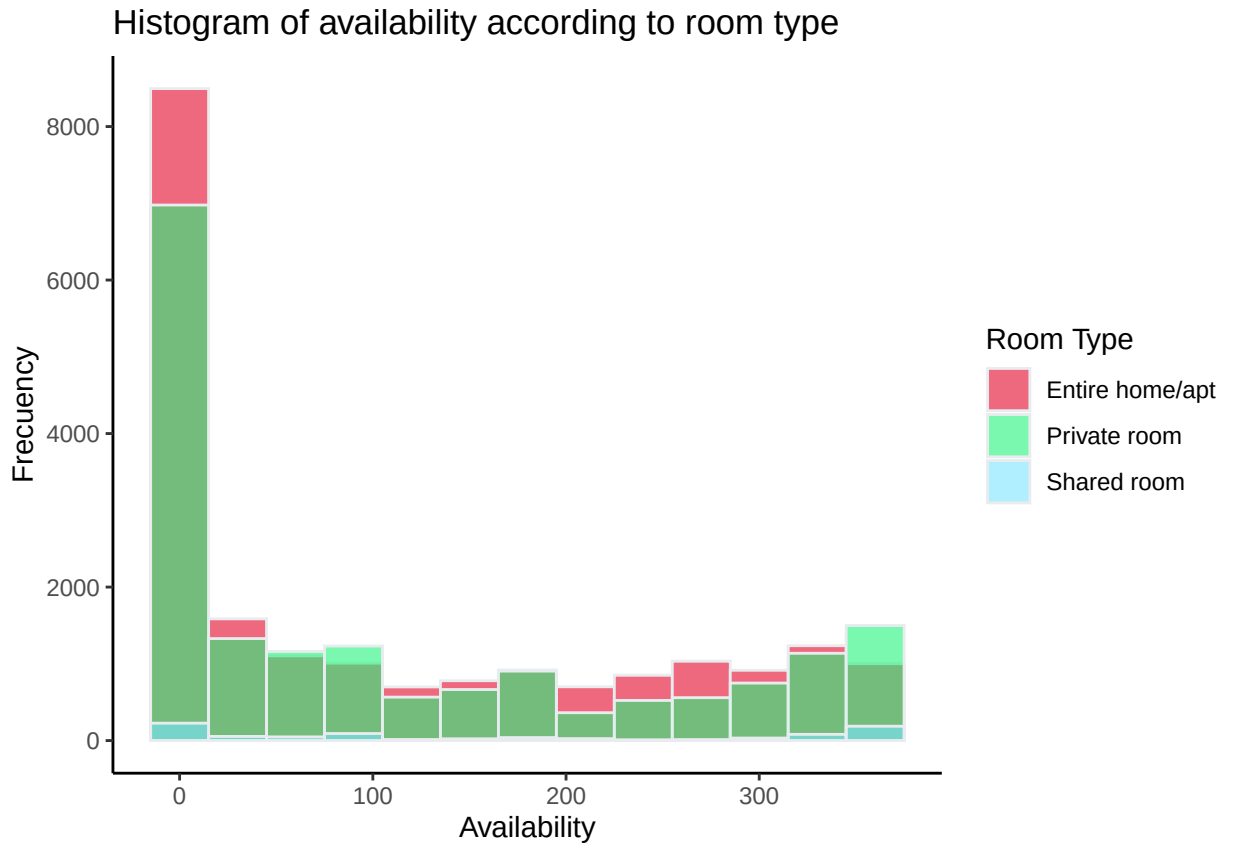
```
##
##  Information of: last_review
## [1] "6/23/2019"
## attr(,"freq")
## [1] 1413
```

Answer to the Question "IMPLEMENTATION" I & II

(a)

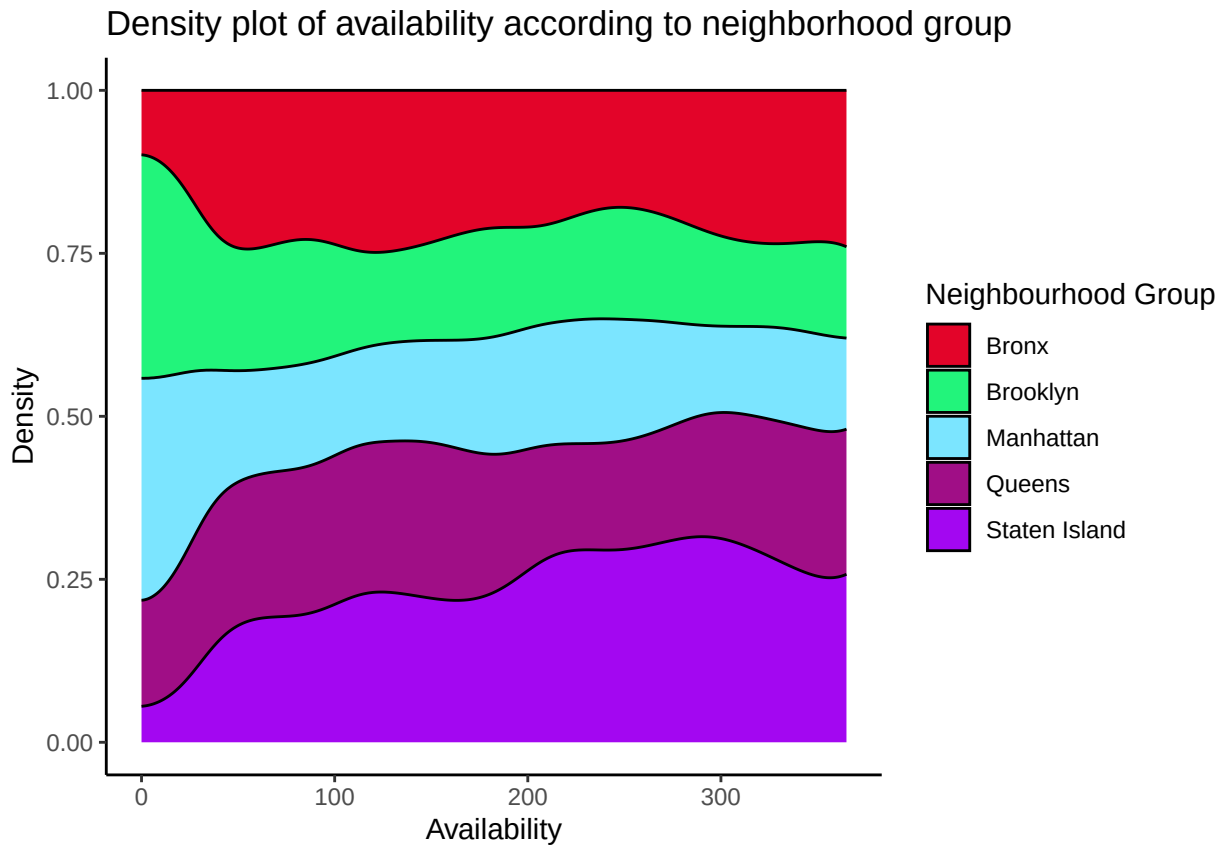
Histogram: This is showing us the number of available rooms in a year a accordind to room types 2 features used 1. availability_365 2. room_type x-axis shows the availability y-axis shows the count or frequency Histogram is grouped into 3 room types

```
ggplot( dataSet,aes(x=availability_365, fill=room_type)) +
  geom_histogram(binwidth= 30,color="#e9ecef", alpha=0.6, position = 'identity') +
  scale_fill_manual(name= "Room Type",values=c("#E30429", "#22F47B","#7BE5FF")) +
  labs(title = "Histogram of availability according to room type", x="Availability", y= "Frecuency")+
  theme_classic()
```



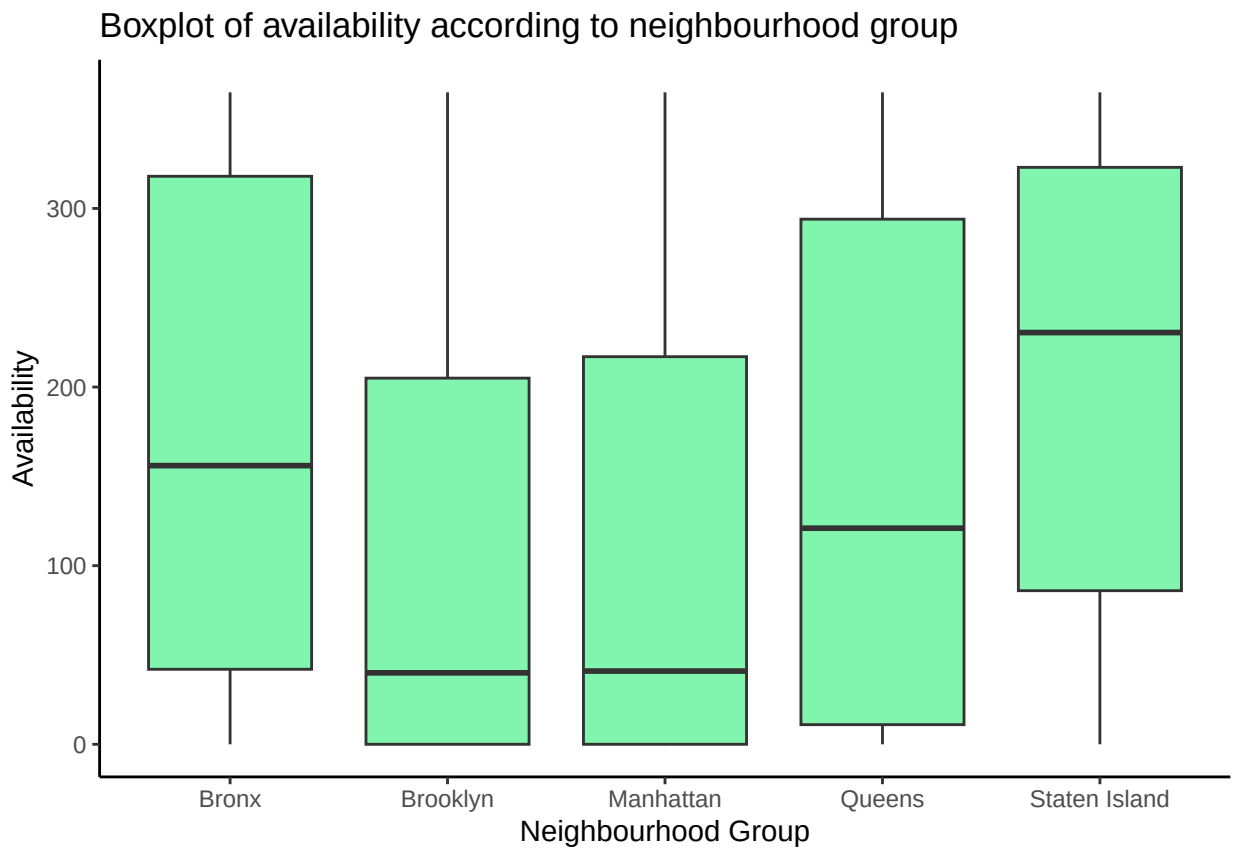
(b) Density Plot: This is showing us the density of available room in each neighborhood group. 2 features used 1. availability_365 2. neighbourhood_group x-axis shows the availability y-axis shows the density Density plot is grouped into 5 neighborhood groups.

```
ggplot(dataSet,aes(x=availability_365, group=neighbourhood_group, fill=neighbourhood_group)) +
    geom_density(position="fill")+
  scale_fill_manual(name= "Neighbourhood Group",values=c("#E30429", "#22F47B","#7BE5FF","#A00E86","#A30
  labs(title = "Density plot of availability according to neighborhood group", x="Availability", y= "De
   theme_classic()
```



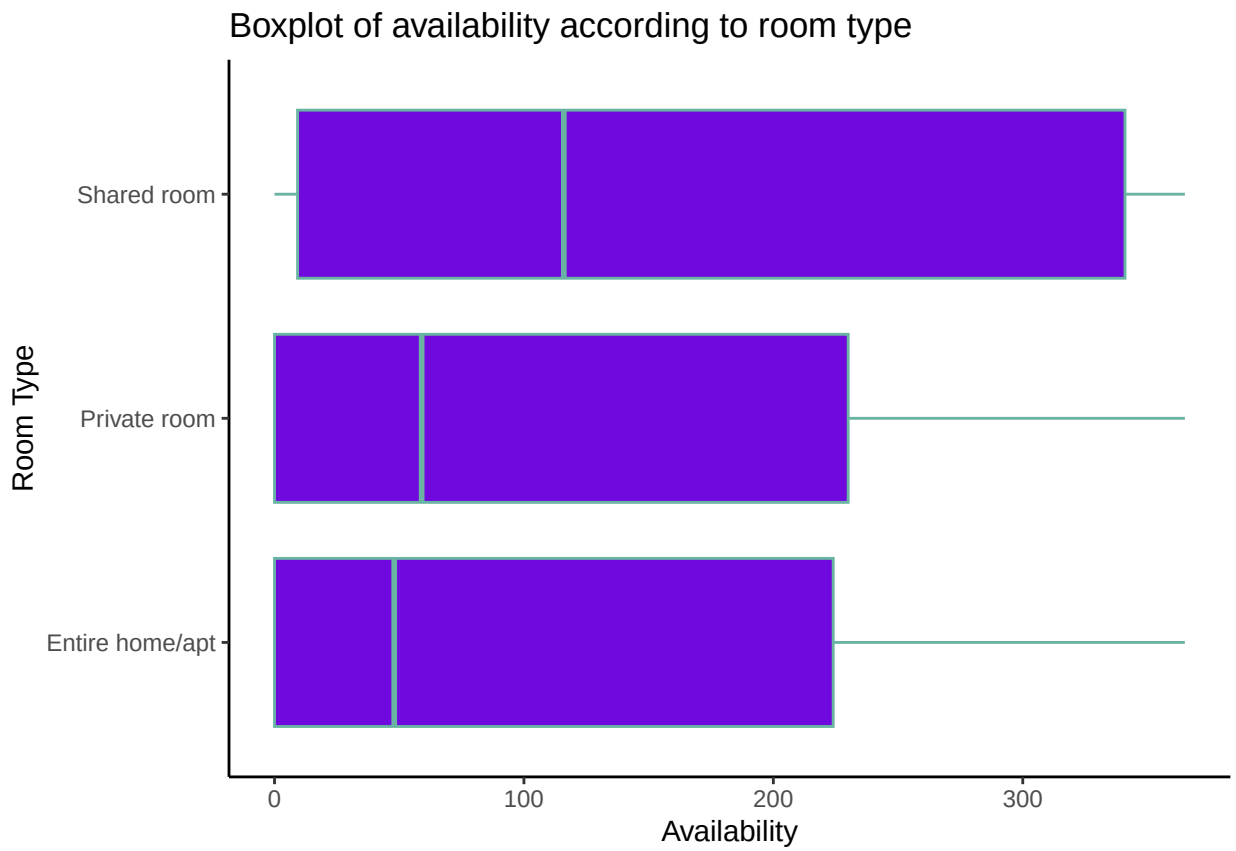
(c) Box plot-This is showing us the data summery of each neighborhood group according to availability. 2 features used 1. availability_365 2. neighbourhood_group x-axis shows shows the density y-axis the availability Box plot is grouped into 5 neighborhood groups.

```
ggplot(dataSet, aes( neighbourhood_group ,  availability_365)) +
  geom_boxplot(fill= "#81F5AD")+
    labs(title = "Boxplot of availability according to neighbourhood group", x="Neighbourhood Group", y=
   theme_classic()
```



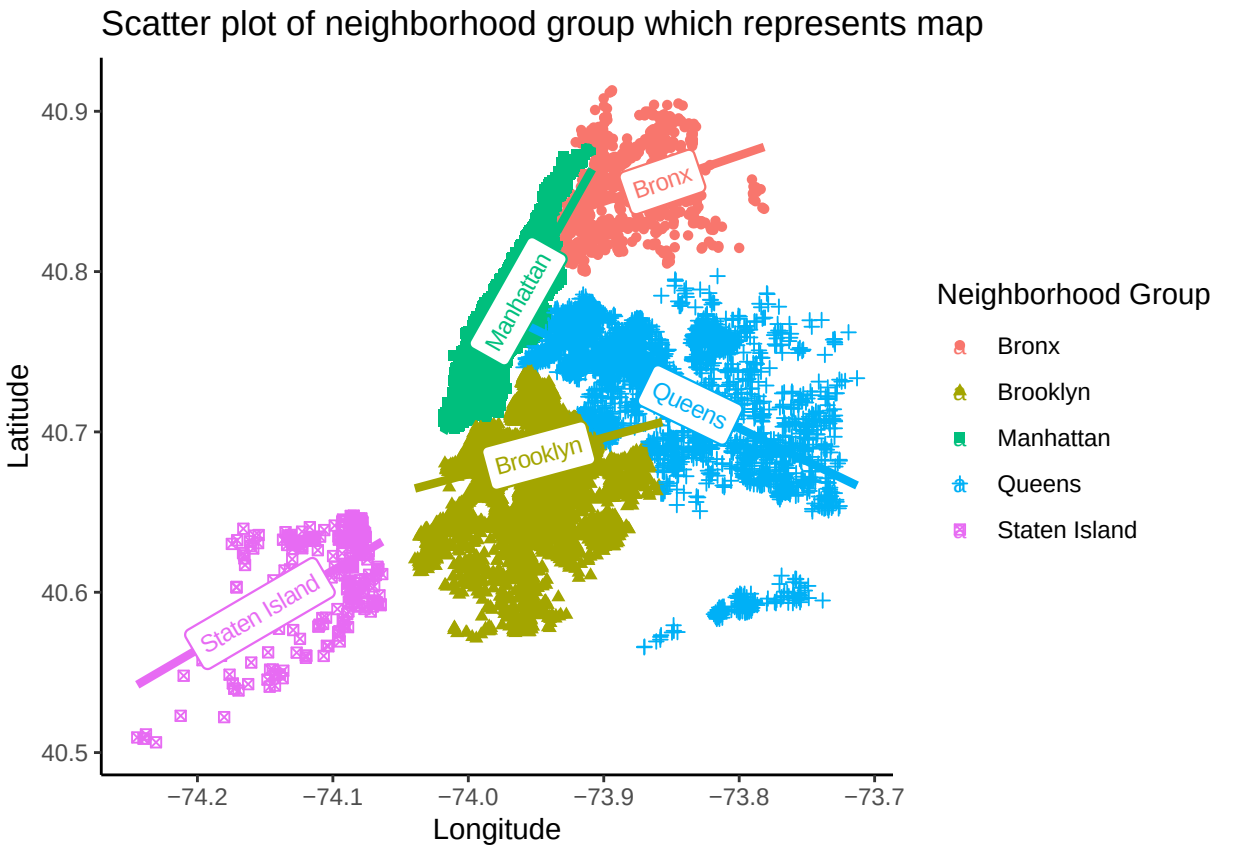
Box plot-This is showing us the data summery of each room type according to availability. 2 features used 1. availability_365 2. room_type x-axis shows the availability y-axis shows the room type Box plot is grouped into 3 room types

```
#boxplot(data$availability_365 ~ data$room_type , ylab="sickness" , col="#69b3a2", boxwex=0.5 , main="",
ggplot(dataSet, aes( availability_365 ,  room_type)) +
  geom_boxplot(fill= "#7009DE", col="#69b3a2")+
  labs(title = "Boxplot of availability according to room type", x="Availability", y= "Room Type") +
   theme_classic()
```



(d) Scatter Plot- This shows the longitude and latitude according to neighborhood group. This represents like a map which is a fun fact because longitude and latitude represents location. 3 features used 1. longitude 2. latitude 3. neighbourhood_group x-axis shows the longitude y-axis shows the latitude Scatter plot is grouped into 5 neighborhood groups which is clearly representend and looks like a map.

```
ggplot(dataSet , aes(x=longitude, y=latitude, shape=neighbourhood_group, color=neighbourhood_group)) +
    geom_point()  +
  geom_labelsmooth(aes(label = neighbourhood_group), fill = "white",
                method = "lm", formula = y ~ x,
                size = 3, linewidth = 1.5, boxlinewidth = 0.4)+
  labs(title = "Scatter plot of neighborhood group which represents map", x="Longitude", y= "Latitude",
       shape = "Neighborhood Group", color = "Neighborhood Group") +
   theme_classic()
```

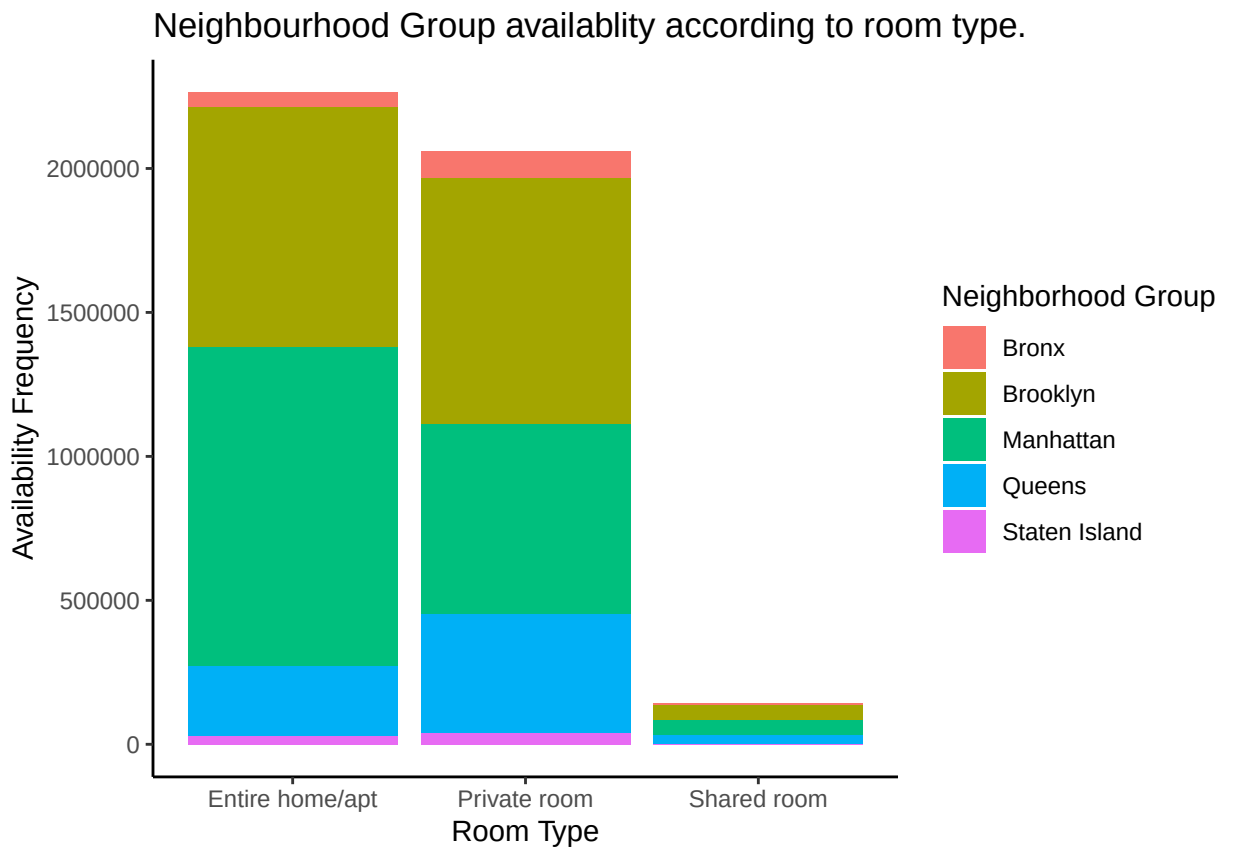


(e) Heat map-Room which are more avalaible or stays empty does not gets enough reviews 3 features used 1. neighbourhood_group 2. calculated_host_listings_count 3. reviews_per_month x-axis shows the Neighborhood y-axis shows the Total host listings

```
ggplot(dataSet, aes(neighbourhood_group, calculated_host_listings_count  )) +
  geom_tile(aes(fill= reviews_per_month))+
  labs(title = "Heat map of neighborhood and total host listing which has reviews per month",
       x="Neighborhood Group", y= "Total Host listing", fill="Reviews Per Month")+
  theme_classic()
```

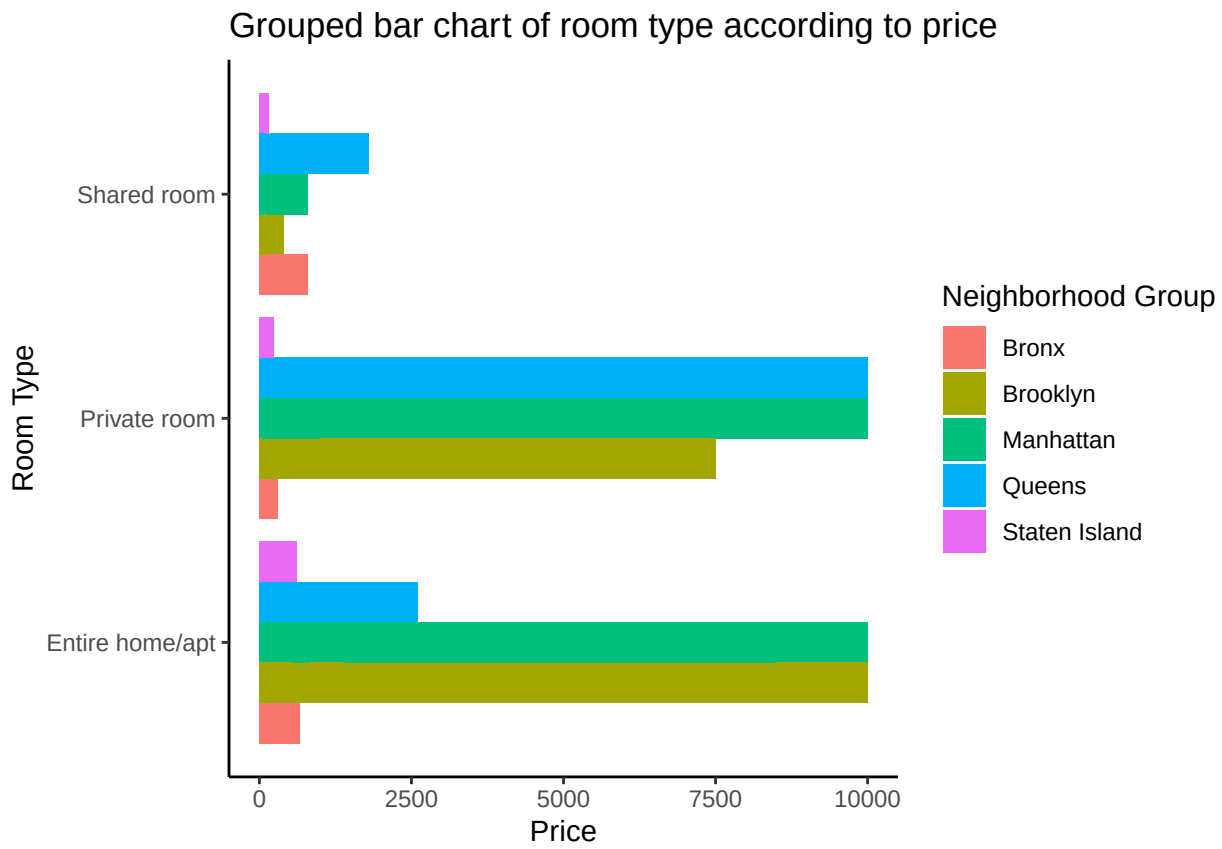
(f) Stacked Bar Chart- Neighbourhood Group availablity according to room type 3 features used 1. neighbourhood_group 2. availability_365 3. room_type x-axis shows the Room type y-axis shows the frequency of availability stacks are grouped using neighborhood group

```
ggplot(dataSet, aes(fill=neighbourhood_group, y=availability_365, x=room_type)) +
    geom_bar(position="stack", stat="identity")+
    labs(title = "Neighbourhood Group availablity according to room type.",
       x="Room Type", y= "Availability Frequency", fill= "Neighborhood Group")+
     theme_classic()
```



(g) Grouped Bar Chart- Price in each neighborhood group based on room type 3 features used 1. neighbourhood_group 2. price 3. room_type x-axis shows the price y-axis shows the Room type

```
ggplot(dataSet, aes(fill= neighbourhood_group, price,room_type  )) +
  geom_bar(position = "dodge", stat= "identity")+
  labs(title = "Grouped bar chart of room type according to price",
       x="Price", y= "Room Type", fill= "Neighborhood Group")+
    theme_classic()
```



Answer to the Question "Accuracy" Showing strong correlation from this dataset

Here I took the dataset and inserted to a new one named cor_df where I have all the numeric data

```
cor_df<- dataSet[, !(names(dataSet) %in% c("id", "name","host_id","host_name","neighbourhood_group","ne
```

Later we calculated all the correlation of the numeric features

```
cor_df<- round(cor(cor_df),3)
cor_df
```

```
##                                 latitude longitude  price minimum_nights
## latitude                           1.000     0.088  0.031          0.025
## longitude                          0.088     1.000 -0.155         -0.055
## price                              0.031    -0.155  1.000          0.026
## minimum_nights                     0.025    -0.055  0.026          1.000
## number_of_reviews                 -0.009     0.055 -0.036         -0.069
## reviews_per_month                 -0.010     0.146 -0.031         -0.122
## calculated_host_listings_count     0.004    -0.093  0.053          0.073
## availability_365                  -0.022     0.103  0.078          0.102
##                                 number_of_reviews reviews_per_month
## latitude                                   -0.009            -0.010
## longitude                                   0.055             0.146
## price                                      -0.036            -0.031
## minimum_nights                             -0.069            -0.122
## number_of_reviews                           1.000             0.550
```

```
## reviews_per_month                          0.550           1.000
## calculated_host_listings_count            -0.060          -0.009
## availability_365                           0.194           0.186
##                                 calculated_host_listings_count availability_365
## latitude                                                 0.004          -0.022
## longitude                                               -0.093           0.103
## price                                                    0.053           0.078
## minimum_nights                                           0.073           0.102
## number_of_reviews                                       -0.060           0.194
## reviews_per_month                                       -0.009           0.186
## calculated_host_listings_count                           1.000           0.183
## availability_365                                         0.183           1.000
```

Using ggcorrplot we plotted all the correlation with values inside generation a correlation matrix Where top 3 correlation are: 1. Number of reviews-Reviews per month 2. Availability 365-Number of reviews 3. Availability 365-Number of host listing

```
ggcorrplot(cor_df,hc.order = TRUE, title="", lab=TRUE, lab_size =2.5 )
```