****

**SKILL DEVELOPMENT LABORATORY**

**DATA SCIENCE - R PROGRAMMING**

**Mini-Project report on**

**TOPIC: NEW-YORK CITY AIRBNB ANALYSIS**

**Submitted by:**

Raj Gupta | T194006 | BDS120

Aditya Ujalambkar | T194010 | BDS156

**Guided by:**

Prof. Jayvant Devare

Ms. Shobha Mourya

**School of Computer Engineering and Technology**

MIT Academy of Engineering

Alandi, Pune-412105

**CERTIFICATE**

This is to certify that the mini project entitled “**New-York City Airbnb Analysis**” has been carried out by Raj Gupta (T194006) and Aditya Ujalambkar (T194010), in the partial fulfilment of the requirements for the course Skill Development Laboratory (Data Science) in the Fifth Semester of Degree of Engineering in School of Computer Engineering and Technology, MIT Academy of Engineering.

Prof. Jayvant Devare

Course Instructor

**ACKNOWLEDGEMENTS**

We would like to express our gratitude to the School of Computer Engineering and Technology of MIT Academy of Engineering, who encouraged us to do this minor project in the course of Data Science, Skill Development Laboratory : **New-York city Airbnb analysis.**

We wish to express our profound thanks to our project guides, Prof. Jayvant H. Devare and Ms. Shobha Mourya for their full cooperation, guidance and advices for the betterment of our project.

We are grateful that such interesting courses are included in the curriculum and for the proper support and guidance.

Raj Gupta | T194006

Aditya Ujalambkar | T194010

**CONTENTS**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Topic** | | **Pg. No.** |
| **1.** | **Introduction** | |  |
|  | 1.1 | Synopsis |  |
|  | 1.2 | Background and Motivation |  |
|  | 1.3 | Problem definition & Objectives |  |
|  | 1.4 | Methodology & Process Workflow |  |
| **2.** | **Data set** | |  |
|  | 2.1 | Description |  |
|  | 2.2 | Attributes used |  |
| **3.** | **System Working** | |  |
|  | 3.1 | Block Diagram |  |
|  | 3.2 | Code |  |
|  | 3.3 | Results (Screenshots) |  |
| **4.** | **Conclusion** | |  |
| **5.** | **Future Work** | |  |
| **6.** | **References** | |  |

1. **INTRODUCTION**

**1.1] SYNOPSIS**

We have performed data exploration, manipulation, visualization on the New-York City Airbnb dataset, a database including information on rental landscape in NYC through Airbnb from 2008 through 2019. We have identified the crucial parameters from the vast dataset and produced the visual presentations that could be interpreted even by the layman.

As the job of an analyst, we have extracted specific information that would help identifying whether the growth of Airbnb has increased or decreased. Descriptive analysis of Airbnb will tell us the summary of the dataset and gives an overview by plotting graphs for different attributes, hence giving solution to many problems.

**1.2] MOTIVATION AND BACKGROUND**

Airbnb has seen a meteoric growth since its inception in 2008 with the number of rentals listed on its website growing exponentially each year. Airbnb has successfully disrupted the traditional hospitality industry as more and more travellers, not just the ones who are looking for a bang for their buck but also business travellers resort to Airbnb as their premier accommodation provider.

New York City has been one of the hottest markets for Airbnb, with over 45,000 listings as of November 2018. This means there are over 40 homes being rented out per square km. in NYC on Airbnb! One can perhaps attribute the success of Airbnb in NYC to the high rates charged by the hotels, which are primarily driven by the exorbitant rental prices in the city.

We were motivated to take up this project as we could perform an-in depth analysis on one of the most densely populated cities in the world. Descriptive analysis provides various insights and equips us for the future, by studying the history.

**1.3] PROBLEM DEFINITION & OBJECTIVES**

**Problem definition:**

To analyze the Airbnb dataset for New-York city by data exploration, manipulation and visualization using different libraries in R-programming and giving solution to many problems.

**Objectives:**

* To study the fundamentals of R programming and apply the various data manipulation techniques.
* To extract insights out of data set.
* To produce visualizations.
* To understand the more frequent regions where Airbnb is used and how it affects the life of the city.
* To make the huge data in the tabular format more meaningful.

**1.4] METHODOLOGY AND PROCESS WORKFLOW**

The input data is taken from Kaggle which is a dataset about Airbnb for New-York city from 2008 upto 2019 showing how guests and hosts have used Airbnb to expand on traveling possibilities and present more unique, personalized way of experiencing the world.

**Data Exploration and libraries:**

* Initially the required libraries are installed and loaded into the R studio. Libraries are: readr, ggplot2, dplyr
* Then the dataset is extracted by read command.
* The data is then explored by using the str, glimpse, ?(command) commands.
* The glimpse command gives the no of variables and observations, datatype of each variable and its values.
* The str command gives more detailed information about the variables that is it helps understanding the categorical variables easily by displaying its levels.
* The readr package is to provide a fast and friendly way to read rectangular data (like 'csv', 'tsv', and 'fwf').
* The ggplot2 is a plotting package that makes it simple to create complex plots from data in a data frame. It provides a more programmatic interface for specifying what variables to plot, how they are displayed, and general visual properties.
* The dplyr is a package for making data manipulation easier. The most common functions of this library are select(), mutate(), filter(),summarize(), etc.

**Data Manipulation:**

* Few variables are renamed for simplicity using rename function.
* The data is filtered on the basis of room type, price, availability of rooms and number of nights stayed.
* The data is also grouped by various variables and summarized for analysis using different plots.
* The data analysis is then done on the manipulated data.

**Data Analysis:**

* Data is analyzed using various plots like bar chart, histogram, density, jitter, cumulative distribution.
* By using latitude and longitude coordinates we plot a scatter graph w.r.t price
* A bar graph is drawn to show count of each room type in each neighbourhood group.
* A transformed price data is shown by using histogram.
* Cumulative distribution of property price is shown.
* A point distribution is shown for each room type by using jitter.
* A simple bar graph is used to show the number of property in each area.

1. **DATA SET**

**2.1] DESCRIPTION**

* The dataset has 16 variables and 48,895 records.
* Each record gives information about a particular home given on rent or available through Airbnb.
* The data includes the Name of the host, id, Area, Price, Number of reviews and more information about the house which is available on Airbnb.
* The data spans from 2008 to 2019 and has information about the property in the 5 neighbourhood group of New-York City.
* Maximum price is 10,000 USD for a Property
* There are 3 types of rooms available i.e. Shared, Private, Entire home.
* Brooklyn, Manhattan, Queens, Staten island, Bronx are the 5 neighbourhood groups.
* There are total 221 neighbourhood regions
* Total of 11453 people has used Airbnb uptil now.

**2.2] ATTRIBUTES USED**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| id | Listing id |
| Name | Name of the listing |
| Host\_id | Host id |
| Host\_name | Name of the host |
| Neighbourhood\_group | Location |
| Neighbourhood | Area |
| Latitude | Latitude coordinates |
| Longitude | Longitude coordinates |
| Price | Price in dollars |
| Room\_type | Listing space type |
| Minimum\_nights | No. of nights stayed |
| number\_of\_reviews | Number of reviews. |
| Last\_review | Latest review date |
| 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 |

**SYSTEM WORKING**

**3.1] BLOCK DIAGRAM**

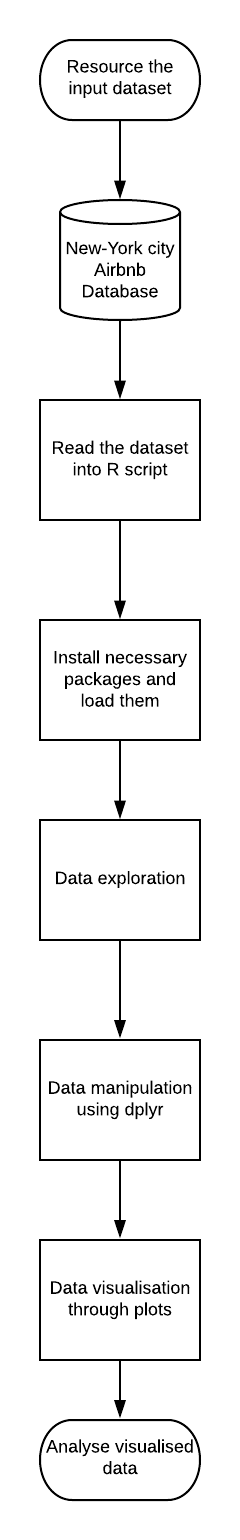
****

Fig. Block diagram showing the process

**3.2] CODE**

#install.packages("readr")

#install.packages("dplyr")

#install.packages("ggplot2")

#install.packages("na.tools")

library(na.tools)

library(ggplot2)

library(dplyr)

library(openintro)

library(readr)

#import data from a csv file

c="C:/Users/rajgu/Desktop/Talend Assignment/AB\_NYC\_2019.csv"

a=read.csv(c,header = TRUE)

#simple Operatioins on table to analyse the data

head(a)

str(a)

glimpse(a)

tail(a)

table(a$host\_name)

unique(a)

desc(a)

unique(a$name)

unique(a$host\_name)

unique(a$host\_id)

table(a$room\_type)

#Proportion of room type in each area

ggplot(a,aes(x=neighbourhood\_group,fill=room\_type))+geom\_bar()

#Cumulative distribution of Property price

ggplot(a, aes(price)) +

stat\_ecdf(geom = "step", color = '#fd5c63', lwd = 1.2) +

ylab("Proportion") + xlab("Price") + theme\_minimal(base\_size = 13) +

ggtitle("The Cumulative Distrubition of Property Price")

#First we produce the frequency of data to see how it will cost us to rent an accomadation in NYC.

a$price %>% hist(xlab="Price",main="frequency of prices", col = "dodgerblue3", breaks = 78)

#Becasue of the existing outliners depicted in figure below, it is not very clear that how much the prices are varying.

boxplot(a$price, main = "overal view of property prices")

#Hence by applying Log10 on prices we would be able to weaken the outliners and have a quick guess about the price range.

hist(log10(a$price),xlab="price", main = "Transformed frequency of the price data Airbnb NYC", col = "dodgerblue3", breaks = 78)

#Graph above depicts that most of the properties are around 10^2 = 100 USD

#Room type analysis

#We would like to see how much prices are varying dependant on whether it is a private room or it is the enitire house or apartment.

Privaterooms = a %>% filter(room\_type == "Private room")

Privaterooms$price %>% summary()

Entirehouse = a %>% filter(room\_type == "Entire home/apt")

Entirehouse$price %>% summary()

sharedroom = a %>% filter(room\_type == "Shared room")

sharedroom$price %>% summary()

#Since our data is extremely skewed because of the outliners, we would prefer to use median price as a more reliable prameter rather that mean value.

#Based on median values and our budget we are able to choose a suitable flat style that we wish to live in.

#Neighbourhood area is, also , an important factor in finding a nice place to accomodate.

table(a$room\_type,a$neighbourhood\_group)

#Table above depicts the number of entire apartment in each each neighbourhood based on the property type.

a %>% boxplot(price ~ neighbourhood\_group,data = ., main="Box Plot of neighbourhood vs price",

ylab="Price",xlab="neighbourhood",horizontal=FALSE, col = "violet")

#Box plot above shows that in Manhattan and Broklyn there are the most expensive properties comparing to other neighbourhoods.

#But yet we are not able to decide what are the median prices in each area by this plot.

#So we categorize each area versus each type of room to realize their median value.

Entirehouse %>% group\_by(neighbourhood\_group) %>% summarise(Min = min(price,na.rm = TRUE),

Q1 = quantile(price,probs = .25,na.rm = TRUE),

Median = median(price, na.rm = TRUE),

Q3 = quantile(price,probs = .75,na.rm = TRUE),

Max = max(price,na.rm = TRUE),

Mean = mean(price, na.rm = TRUE),

SD = sd(price, na.rm = TRUE),

n = n(),

Missing = sum(is.na(price)))

Privaterooms %>% group\_by(neighbourhood\_group) %>% summarise(Min = min(price,na.rm = TRUE),

Q1 = quantile(price,probs = .25,na.rm = TRUE),

Median = median(price, na.rm = TRUE),

Q3 = quantile(price,probs = .75,na.rm = TRUE),

Max = max(price,na.rm = TRUE),

Mean = mean(price, na.rm = TRUE),

SD = sd(price, na.rm = TRUE),

n = n(),

Missing = sum(is.na(price)))

sharedroom %>% group\_by(neighbourhood\_group) %>% summarise(Min = min(price,na.rm = TRUE),

Q1 = quantile(price,probs = .25,na.rm = TRUE),

Median = median(price, na.rm = TRUE),

Q3 = quantile(price,probs = .75,na.rm = TRUE),

Max = max(price,na.rm = TRUE),

Mean = mean(price, na.rm = TRUE),

SD = sd(price, na.rm = TRUE),

n = n(),

Missing = sum(is.na(price)))

#Number of property in each area

ggplot(a) + geom\_histogram(aes(neighbourhood\_group, fill = neighbourhood\_group), stat = "count",alpha = 0.85) +

theme\_minimal(base\_size=13) + xlab("") + ylab("") +theme(legend.position="none") +

ggtitle("The Number of Property in Each Area")

#Among roughly 45,000 properties, a majority of them are located in Manhattan and in Brooklyn. Manhattan with more than 20,000 properties has 50% of the properties in New York City, and Brooklyn has nearly 40%

#price less than 500$

l=a %>%

filter(price<500)

glimpse(l)

ggplot(l,aes(x=longitude,y=latitude,col=price))+geom\_point(size=1)

#Shows density graph of neighbourhood\_group with no.of reviews

ggplot(a,aes(neighbourhood\_group,fill=number\_of\_reviews))+geom\_density(alpha=0.3)

#Shows room availability for private room

ggplot(Privaterooms,aes(availability\_365,0))+geom\_jitter(col=my\_color)+scale\_y\_continuous(limits = c(-0.5,0.5))

#Shows room availability for private room

ggplot(Entirehouse,aes(availability\_365,0))+geom\_jitter(col="green")+scale\_y\_continuous(limits = c(-0.5,0.5))

#Shows room availability for private room

ggplot(sharedroom,aes(availability\_365,0))+geom\_jitter(col="violet")+scale\_y\_continuous(limits = c(-0.5,0.5))

#Plotting for minimum\_nights a customer has stayed

x=a %>%

filter(minimum\_nights<400)

ggplot(x,aes(minimum\_nights,0))+geom\_jitter()+scale\_y\_continuous(limits = c(-2,2))

review=a %>%

filter(number\_of\_reviews<300)

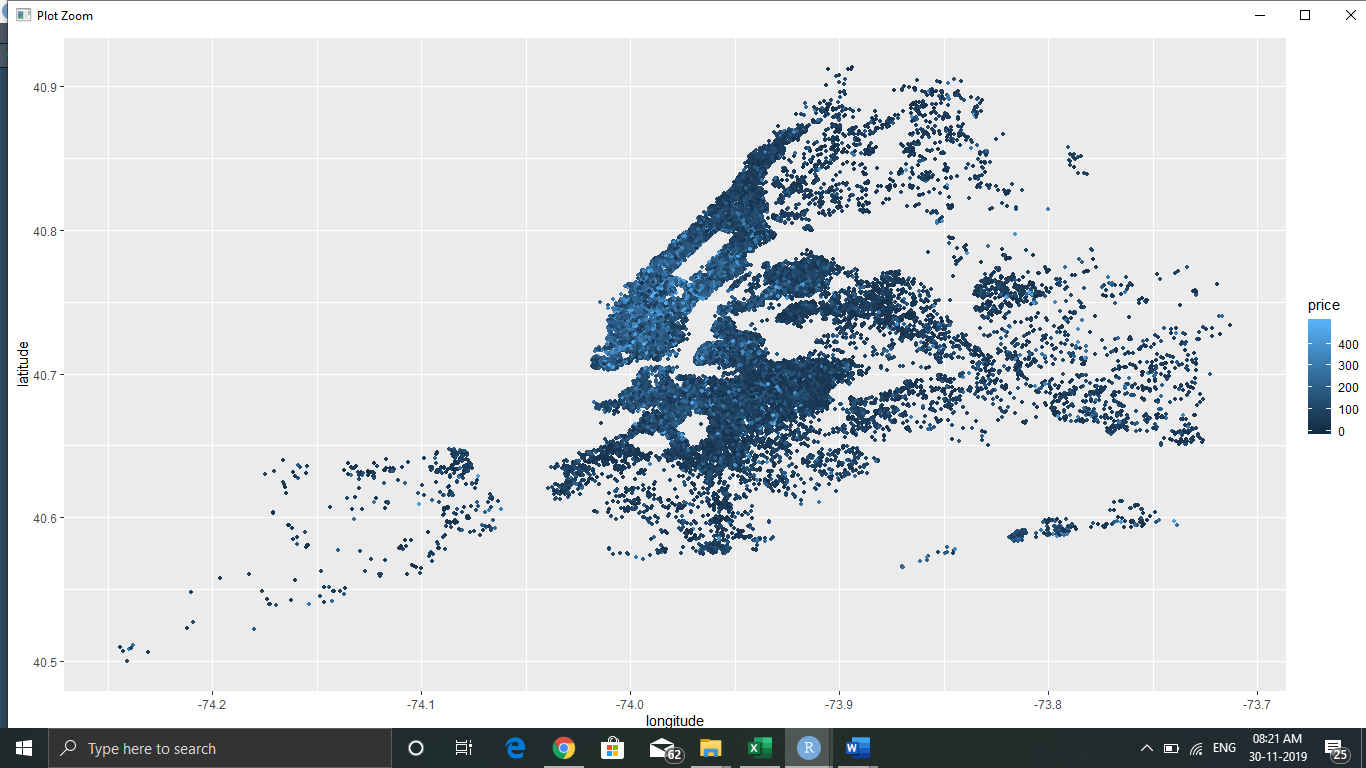
review\_sum=a %>% group\_by(neighbourhood\_group) %>% summarise(review\_sum1=sum(number\_of\_reviews))

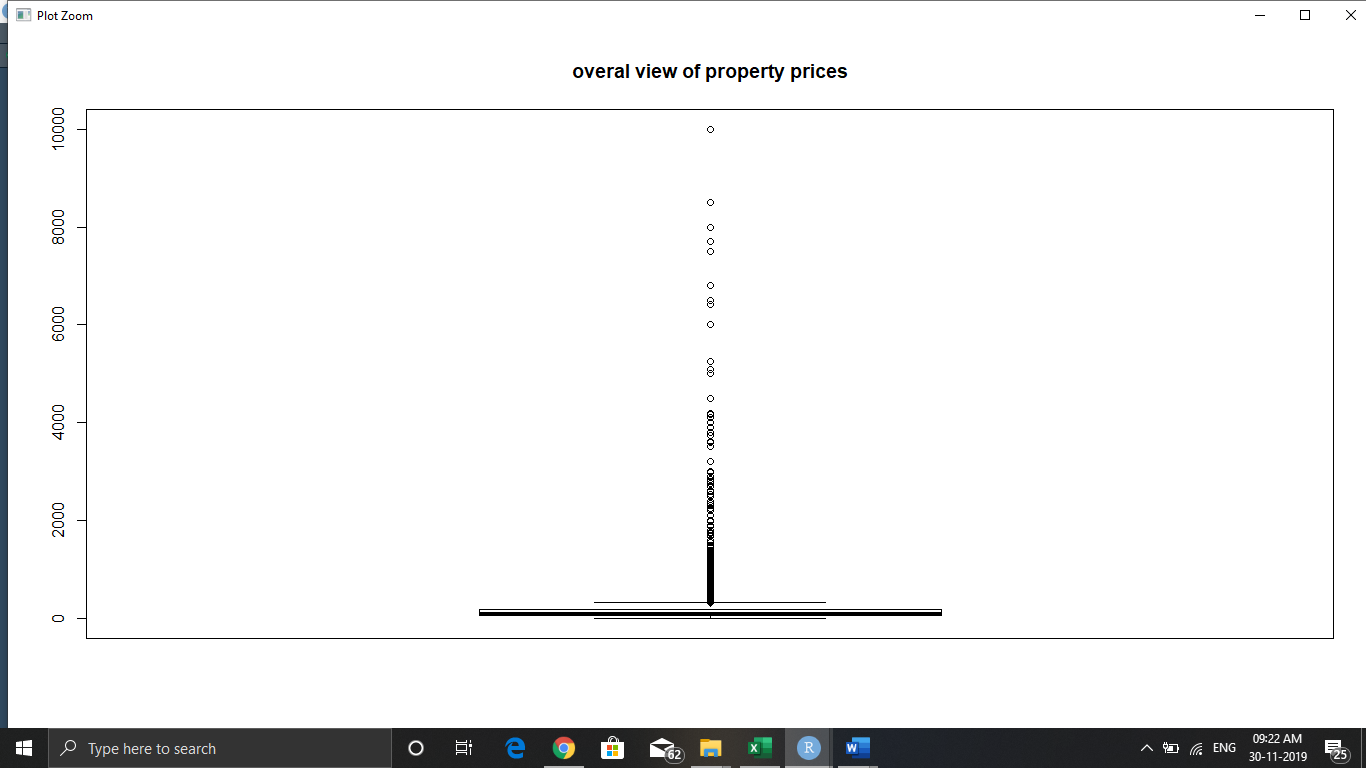
review\_sum

ggplot(review\_sum,aes(review\_sum1,neighbourhood\_group))+geom\_bar(position = "dodge")

**3.3] RESULTS (SCREENSHOTS)**

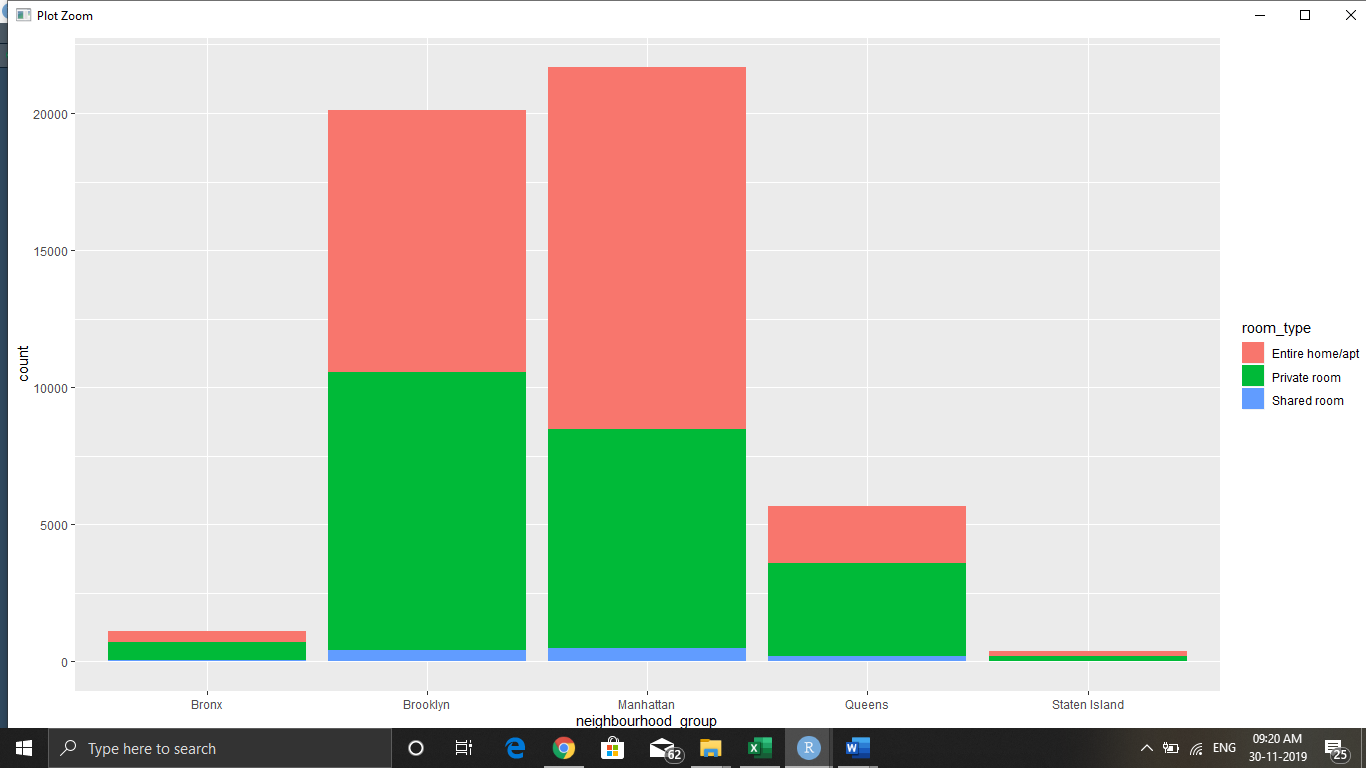
**Showing price less then 500 dollars by plotting a graph through latitude and longitude.**

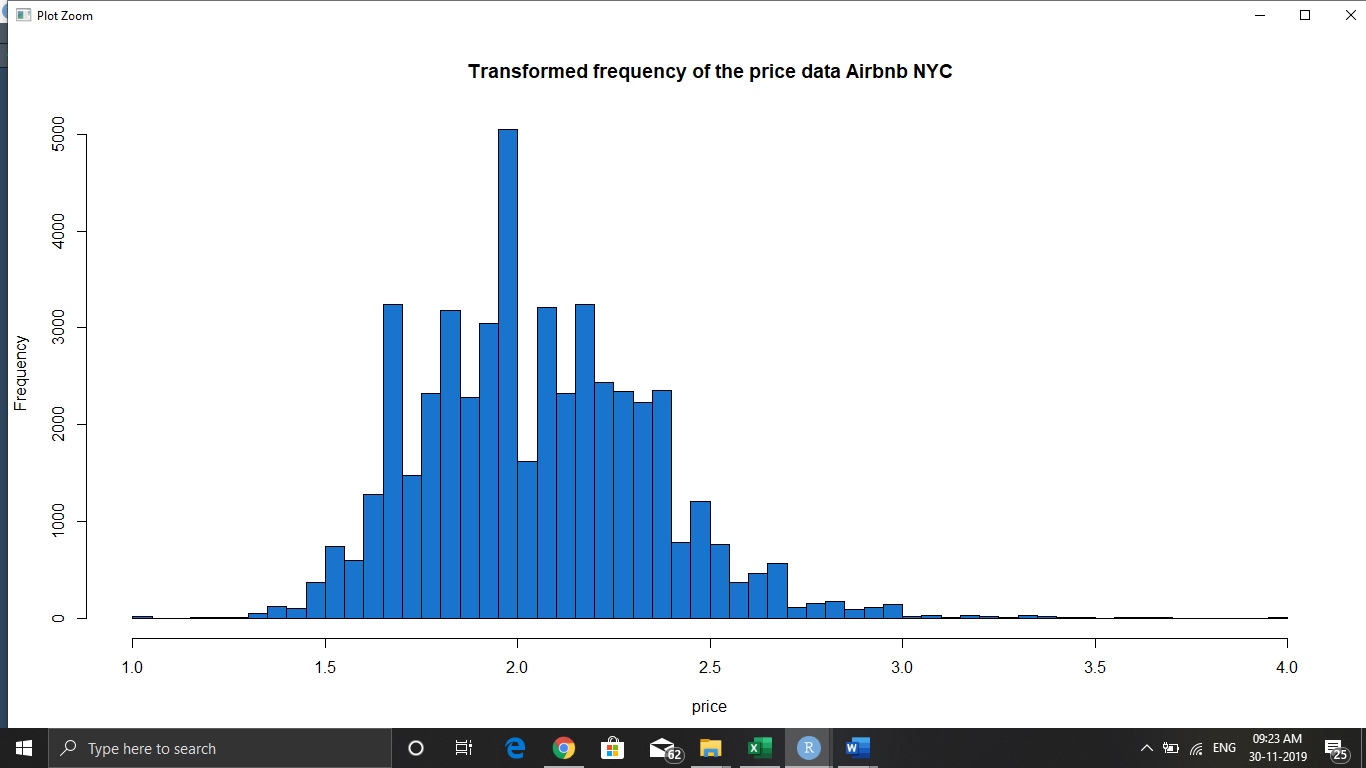




**Outliers are depicted in the figure**

**Proportion of room type in each area**





**Graph above depicts that most of the properties are around 10^2 = 100 USD**

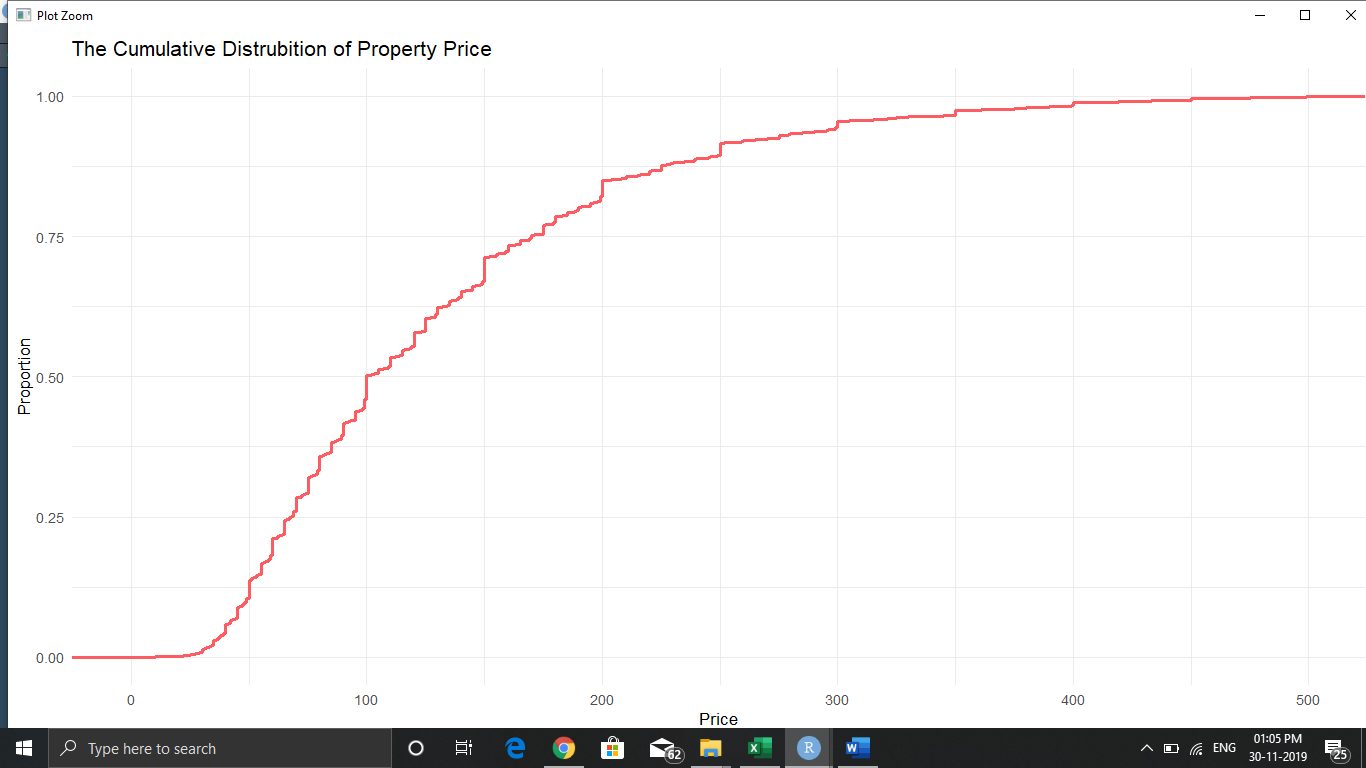


**Box plot above shows that in Manhattan and Broklyn there are the most expensive properties comparing to other neighbourhoods.**

**But yet we are not able to decide what are the median prices in each area by this plot.**

**So we categorize each area versus each type of room to realize their median value.**

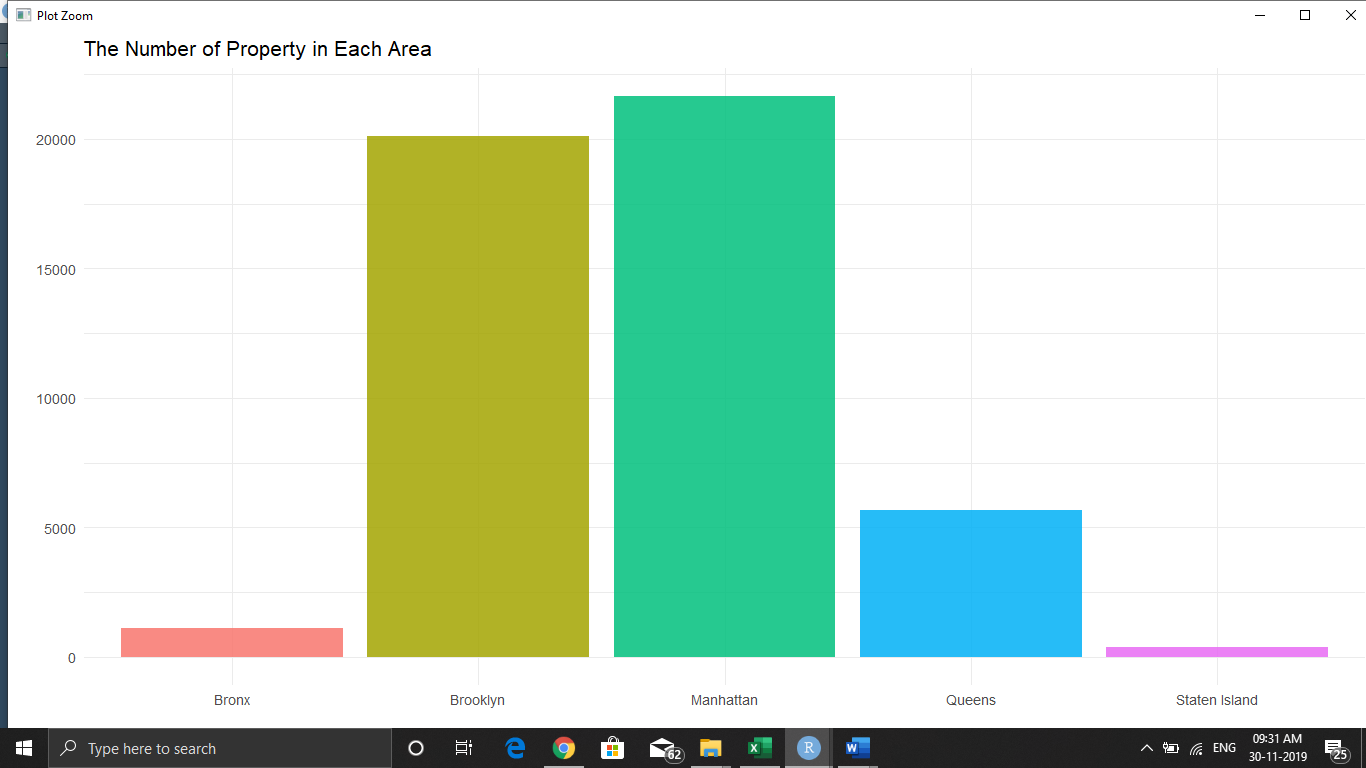
**Cumulative distribution of Property price**



**Median for price is around 100 dollars**

**It shows that 48% houses costs around less then 100 dollars**

**Number of property in each area**

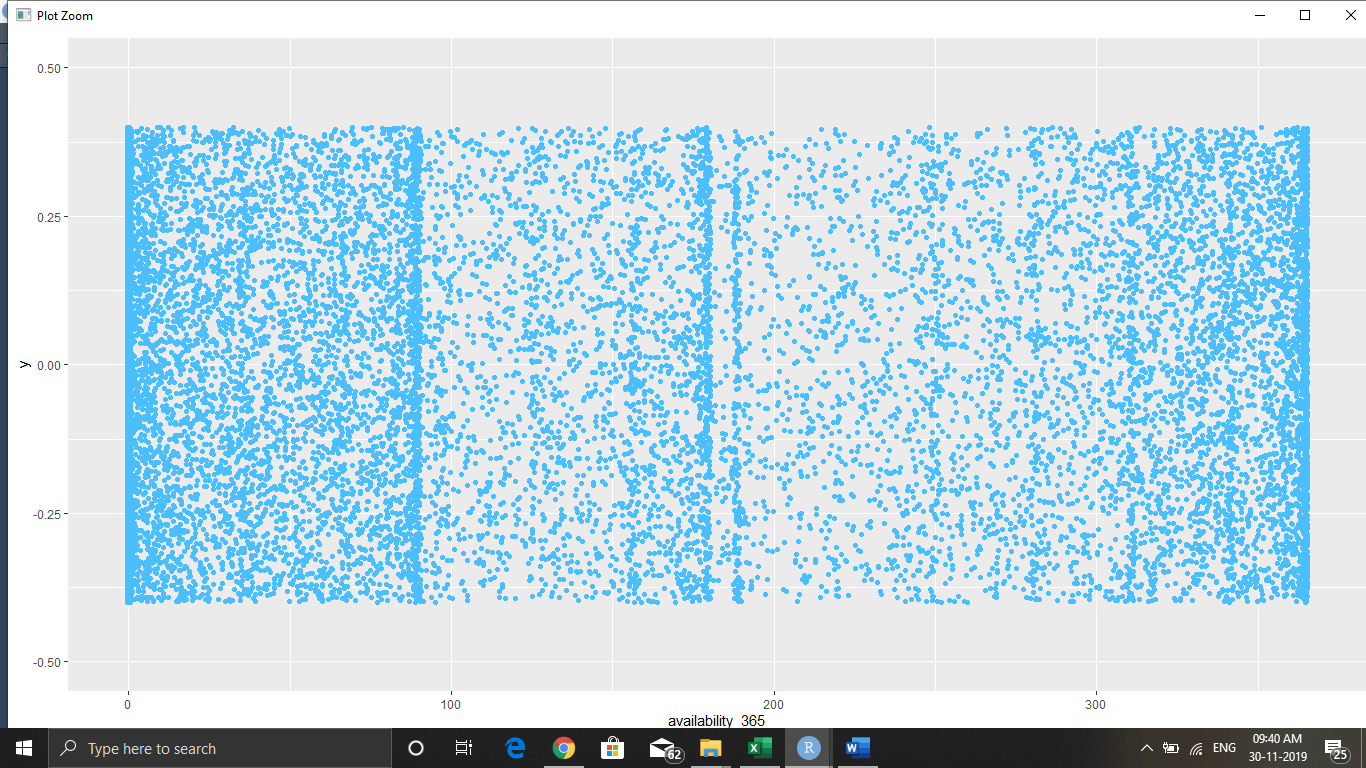


**Among roughly 45,000 properties, a majority of them are located in Manhattan and in Brooklyn. Manhattan with more than 20,000 properties has 50% of the properties in New York City, and Brooklyn has nearly 40%.**

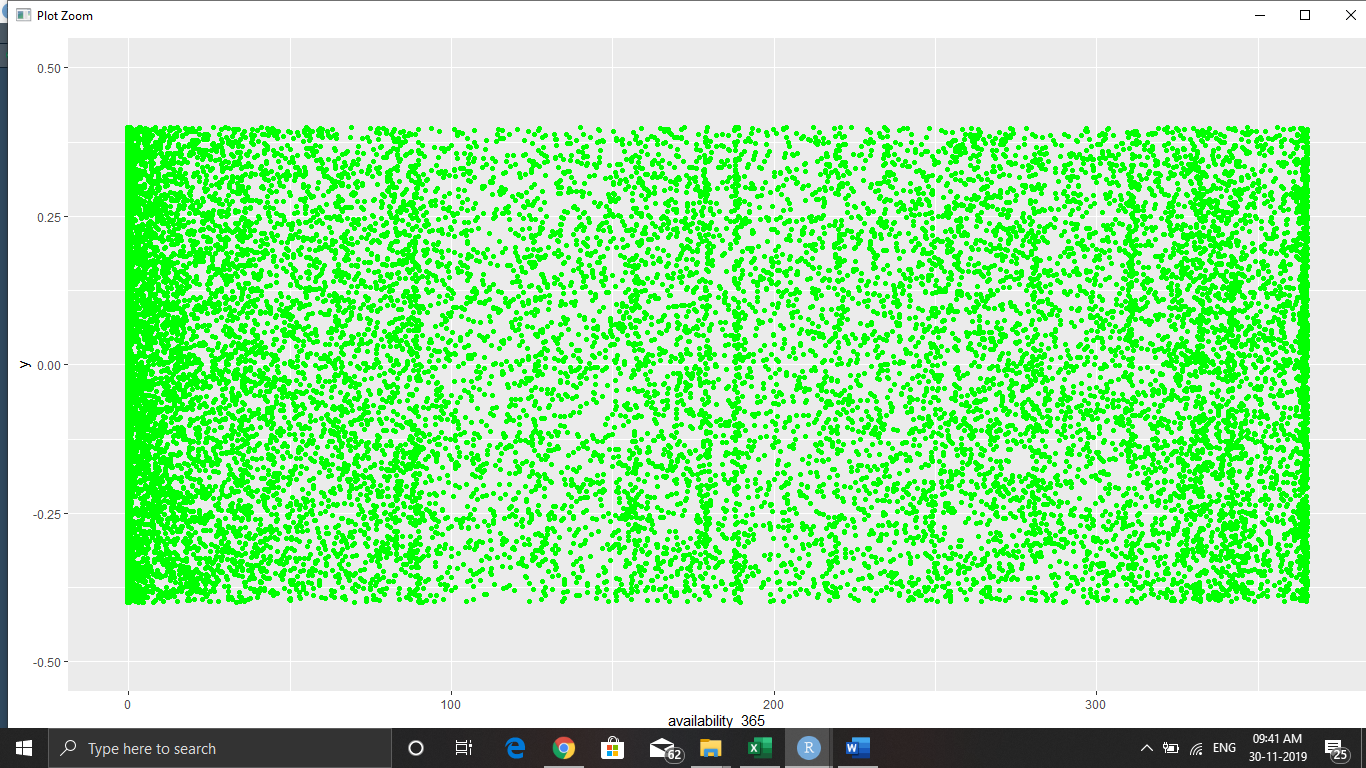
**Shows density graph of neighbourhood\_group with no.of reviews**



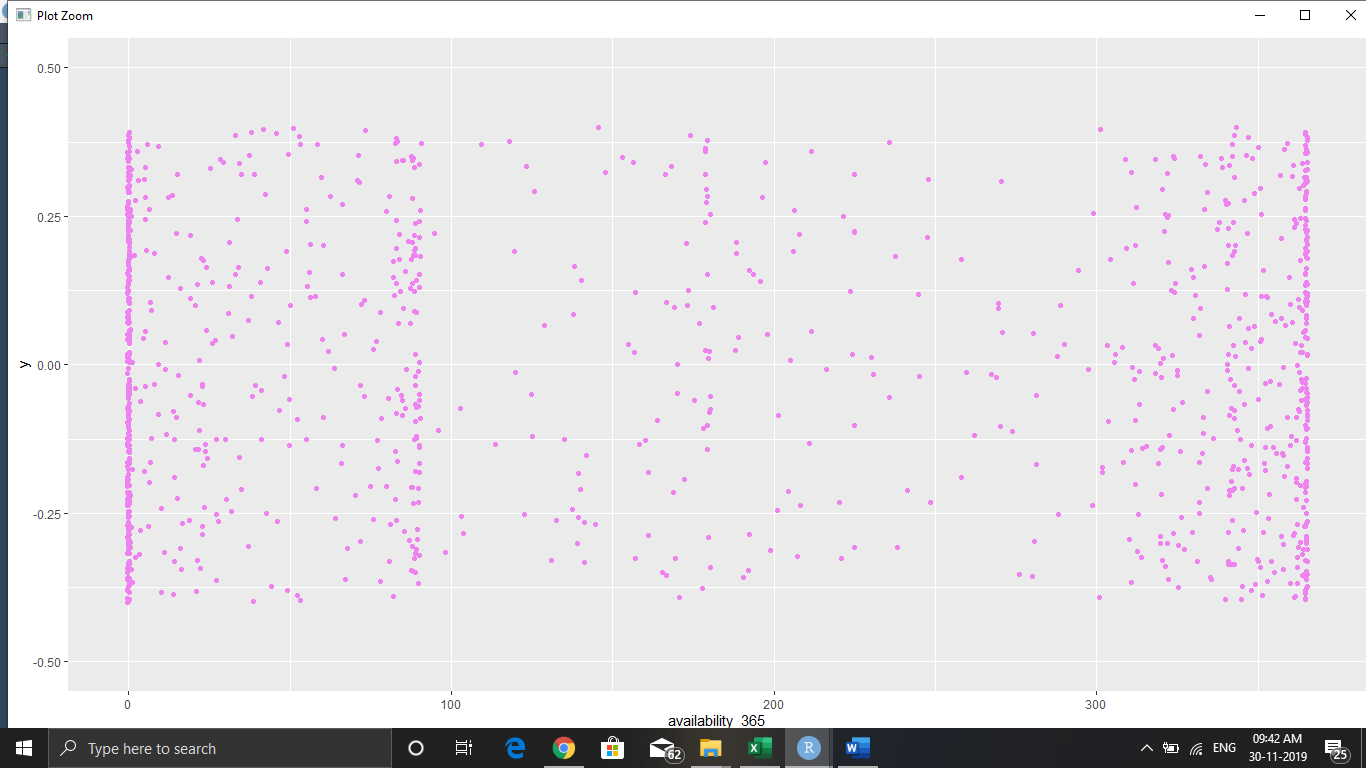
**Shows room availability for private room**



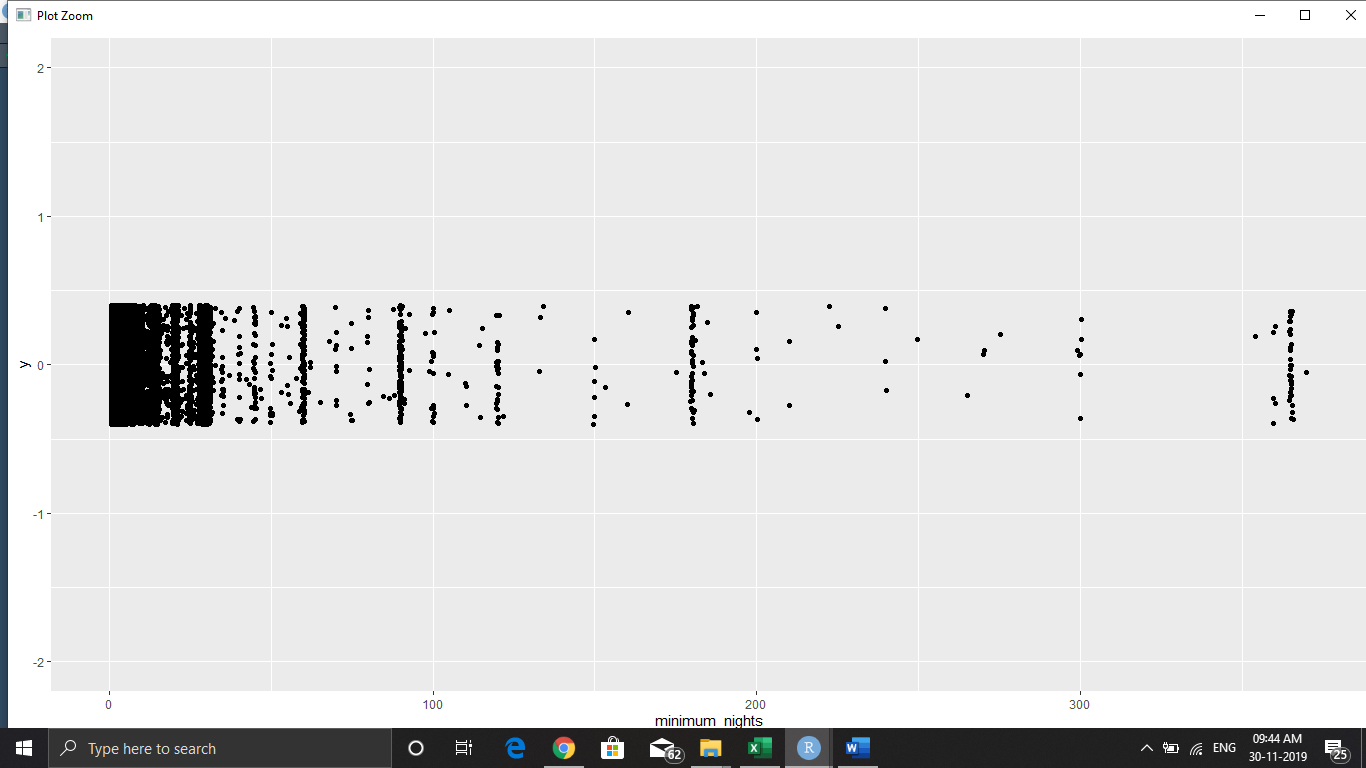
**Shows room availability for Entire house**

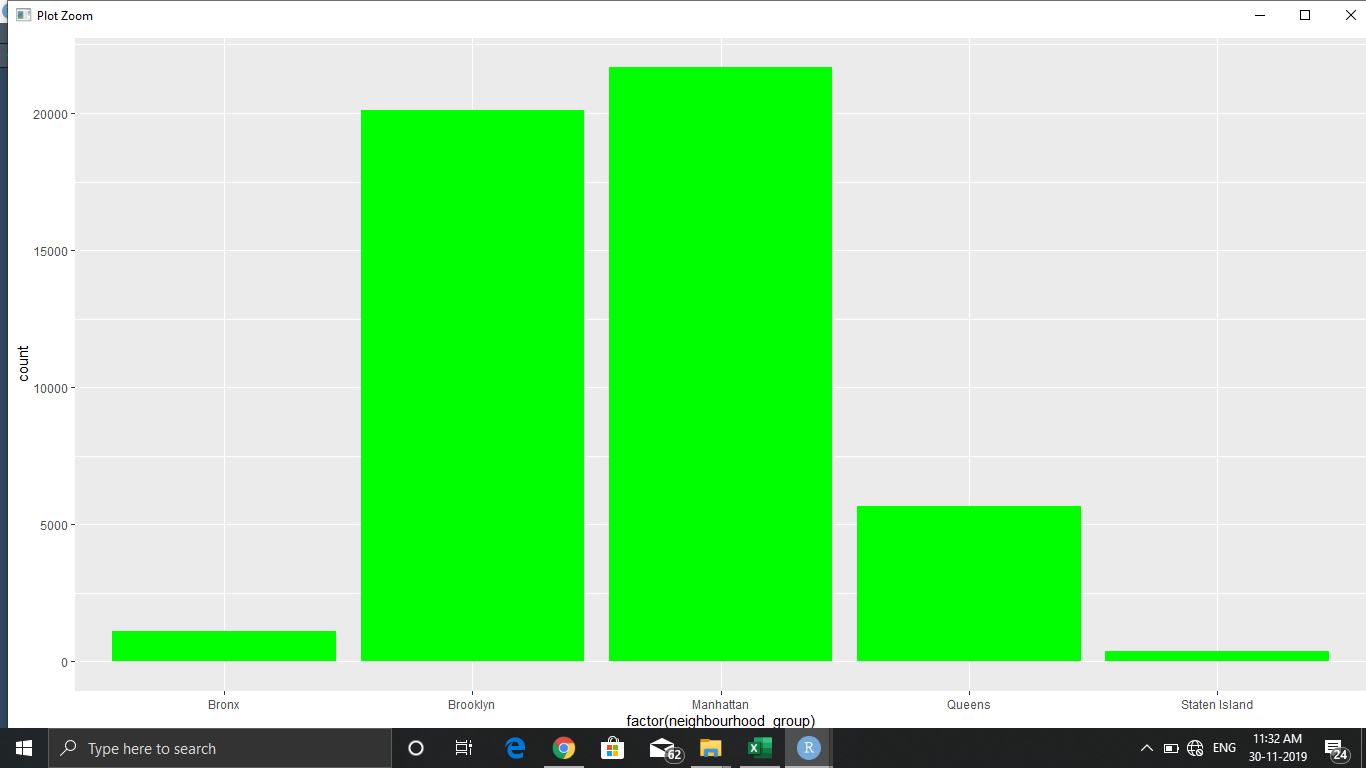


**Shows room availability for shared room**



**Plotting for minimum\_nights a customer has stayed**





**It shows that Manhattan and Brooklyn are more popular**

1. **CONCLUSION**

The Airbnb dataset provides us with a fantastic source to better understand New York’s bustling rental landscape. With over 50k listings registered in the last 9 years, New York has proven to be one of Airbnb’s fastest growing cities.

In General prices in Brooklyn and Manhattan tend to be more expensive other than other regions in all room type categories. In each neighbourhood, there were some outliners where the hosts were offering their property more than its normal prices and Manhattan and Brooklyn has more hosts who falls within that category. Staten Island and Bronx have been slow adopters.

From the data analysis and based on our adopted data, we also can derive the distance between the property and business district to have a better judgement on choosing an accomodation.

In addtion we can see which neighbourhood area had best properties in 2019 based on its total credited reviews.

1. **FUTURE WORK**

We want to expand our analysis to multiple cities and compare patterns and trends amongst these cities. From the insights we have derived, we would also like to build predictive models using different features from the dataset. Lastly, we hope to implement the visualizations and techniques used in this project to many other fields and datasets.

1. **REFERENCES**

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

<https://towardsdatascience.com/airbnb-rental-listings-dataset-mining-f972ed08ddec>