***Trends Prevalent in NYC Airbnb***

*An Analysis of Trends in Airbnb Listings in New York City*

***Submitted by****: Seorin Chae, Bhargavi Kashyap, Asil Qrareya, Kaushik*

***Submitted to****: Eli T. Brown*

***Date****: November 25, 2019*

**Table of Contents**

[Abstract 3](#_Toc25606221)

[Introduction 3](#_Toc25606222)

[Methodology 3](#_Toc25606223)

[Analysis, Result and Findings 3](#_Toc25606224)

[Future Work 9](#_Toc25606225)

[Appendix 10](#_Toc25606226)

[References 46](#_Toc25606227)

# 

# Abstract

The goal of this analysis was to identify the general trends within New York City’s Airbnb offerings between the years 2011 to 2019.

Questions the team sought to answer are as follows:

1. In which neighborhood groups are Airbnb costs the highest and lowest?
2. Which neighborhood groups have the highest density?
3. What does the pricing look like based on property type and location?
4. What is the relationship between availability, reviews, and pricing?
5. How have the pricing of Airbnb listings changed over the years?
6. Are there any additional underlying themes and relationships to be observed? For example, are there any common themes in the names of Airbnb listings that hosts emphasize in the hopes of gaining visitors?

# Introduction

The group consisted of four members, Seorin Chae, Bhargavi Kashyap, Kaushik, and Asil. The team conducted an exploratory and exploratory analysis of the New York City’s Airbnb dataset to seek out trends along with other potentially useful insights that could help both users and hosts better understand the environment. As the gig economy grows, we believe it would be interesting to see the changes that have occurred in this specific market over the span of the past four years.

# Methodology

The dataset was obtained from [Kaggle](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data) and is titled AB\_NYC\_2019. The csv file contains the price, minimum number of nights stayed (used as an indicator of popularity), and neighborhood groups along with various other variables collected between the years 2011 to 2019. The dataset was used to identify key trends in Airbnb listings across New York City.

As a part of data-preprocessing, missing values were checked. The dataset contained numerous missing values, which were eliminated using listwise deletion. taking the dataset from 48,896 observations with 16 variables to 38,843 with 16 variables. The dataset contained both quantitative and qualitative variables. Column names such as ID, host ID, host name was not considered for analysis.

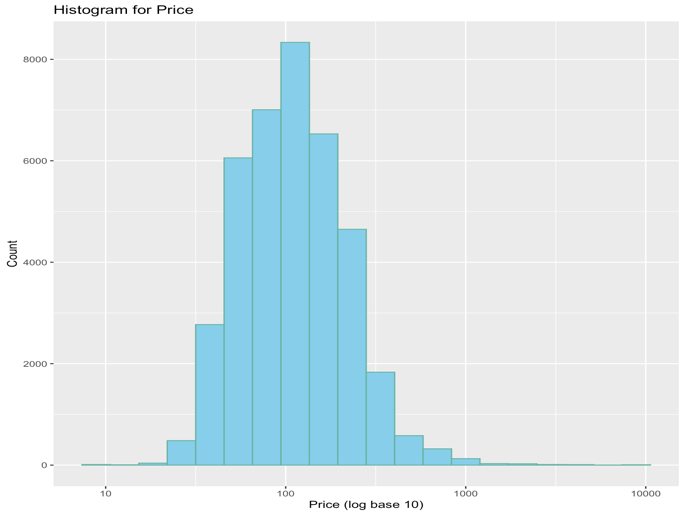
For analysis, various methods such as histogram, scatterplot, heat map and its variation was utilized.

**Tool Used**: Both R and Tableau were used in the creation of the charts and graphs.

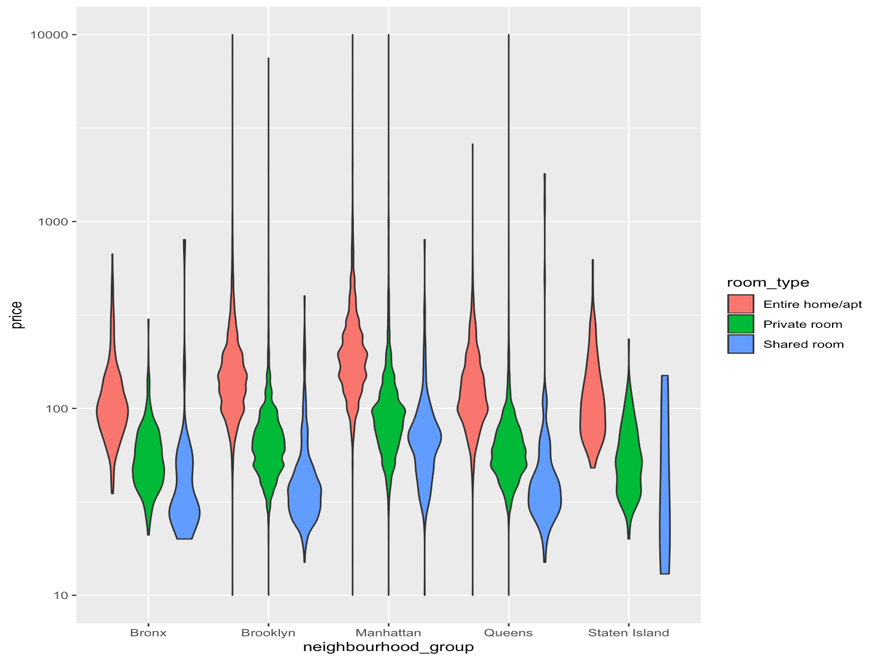
# Analysis, Result and Findings

**Exploratory Analysis:**

* Distribution of price, overall and spread across listing and neighborhood was evaluated
  + Histogram of price
    - As the descriptive statistics indicated, the mean of the price was $142.32, standard deviation of $196.95 and kurtosis of 953.73.
    - Being a highly skewed data, initial histogram was not easy to read. For better clarity, the x-axis was log transformed with log base 10

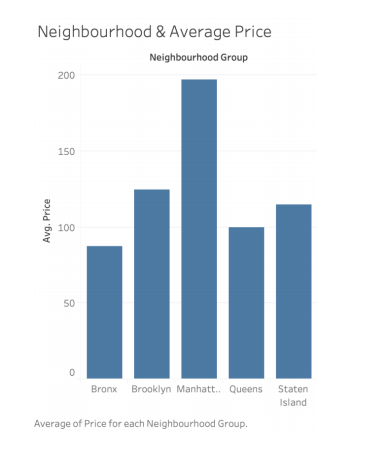


* + Relationship of price across neighborhoods and room types using Violin Plot
    - Distribution of the price in NYC neighborhoods, across room types and its [probability density](https://en.wikipedia.org/wiki/Probability_density_function)



**Explanatory Analysis:**

* **Bar chart**



This plot displays the average price by neighborhood groups. Through this bar chart, we’re able to identify Manhattan as the most expensive neighborhood group, followed by Brooklyn, Staten Island, Queens, and Bronx.

* **Leaflet – Spatial Analysis**



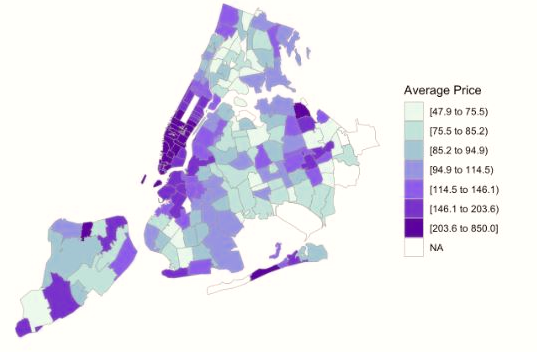
Interactive link: <http://rpubs.com/schae4/551667>

The interactive graph of New York City provides all Airbnb listings in clusters based on sub-regions. If you click on the clusters, you are able to see the listing information in detail – Listing Name, Host Name, Price, and Room Type. On a high level, the clustering allows the viewer to see the distribution of listings and differentiate between high-density and low-density areas for Airbnbs in New York City. The map here states that the maximum listings are in or around Manhattan, Brooklyn, followed by Queens and Bronx. Staten Island region has the least number of listings during 2011-2019.

Furthermore, a closeup of the leaflet can provide an in-depth assessment into the price distribution – minimum and maximum – as well as room type offerings within specific regions. The combination of this information, the listing distribution, along with additional variables (transportation offerings, sentiment of reviews, etc.) will allow future hosts, along with the company, to see areas of opportunity and oversaturation.

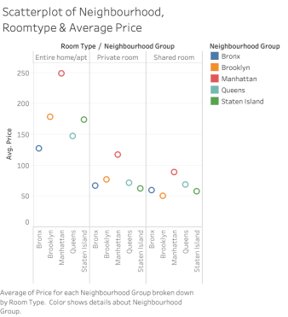
* **Choropleth**

Choropleth displaying the pricing in various sub neighborhoods on a map of New York.



The map confirms our hypothesis further with Manhattan listing costs being the highest, with Brooklyn following as second. Given more information such as amenities, sentiments of review, etc. it would be interesting to study the relationship between the popularity of cumulative listings per area and the costs in that region to better understand the determinants of listing price. This, in combination with the leaflet and other visualizations listed above could help users, hosts, and Airbnb identify areas of opportunity by identifying high rating areas with low rent spots and vice versa.

* **Scatterplot**



After understanding the pricing of each neighborhood groups, we decided to examine the average price of the neighborhood groups by room type. Some key observations that can be noted in this scatterplot is as follows:

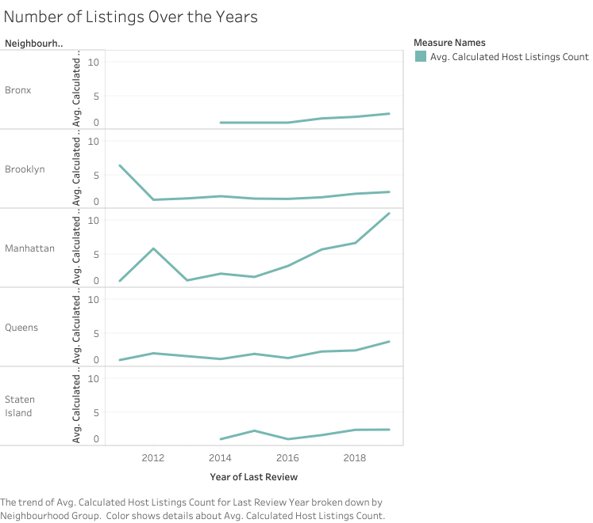
1. All five neighborhood groups offer primarily two room offerings – private room and shared room, with few offerings for entire home/apt
   1. This potentially can be explained by the affluence of the neighborhood & is indicative of the wallet size of those who chose to rent/live there (even for short duration)
2. Manhattan is the most expensive across room types – Entire home/apt, Private room, and Shared room.
3. Brooklyn has the second most expensive offerings in terms of Entire home/apt and Private room; however, a shared room in Queens is more costly than a shared room in Brooklyn.
4. Queens has the third most expensive private room offerings.
5. Brooklyn has the cheapest Shared room offerings
6. Staten Island has the cheapest Private room offerings

* **Time series**



Next, we tried to understand the fluctuations of price in various neighborhood groups between the years 2011 to 2019. Some key observations that can be noted in this scatterplot is as follows:

1. Listings in Bronx became available beginning in 2014 and has seen minimal changes in price since then.
2. Listings in Brooklyn averaged the $200’s but saw a drop in 2013 and has plateaued over the years.
3. Manhattan listings averaged around $100’s in 2011 and saw a sharp incline in 2013, dropped approximately half its growth, and has remained steadily between the $150-$200 mark.
4. Listings in Queens saw a minor peak in 2016 but dropped and normalized quickly thereafter and has remained steady.
5. Staten Island offerings became available in 2014 with prices seeking the fastest growth amongst the neighborhood groups. In 2015, they even exceeded Manhattan’s average price, but quickly fell to the lowest cost range by 2016. This may imply the novelty of the services in the area in combination with the new hosts lack of knowledge of the market demands led to a high price mark, but the eventual regulation of the market.



Lastly, we tried to understand the trends in the number of listings in various neighborhood groups between the years 2011 to 2019. Some key observations that can be noted in this scatterplot is as follows:

1. Listings in Bronx became available beginning in 2014 and has seen a gradual increase in offerings beginning in 2016.
2. Brooklyn listing availabilities dropped significantly in 2012 and has maintained stability over the years.
3. Manhattan listing availabilities fluctuate the most over the span of nine years, seeing a peak in 2012 and a steep drop in 2013, with persistent increase all the way up to 2019.
4. Listings in Queens have seen a gradual increase over the years in listing. The neighborhood has a reputation of not being as affluent. The multiple peaks display how the reputation of the neighborhood has changed and thus, popularity and listings.
5. Listings in Staten Island were apparent in 2014, growth curve similar to price, but deteriorated towards 2016. The listings increased gradually implying the popularity of the neighborhood and also quality of properties offered.

# Future Work

As a next step, the team recommends further collection of observations and variable additions to the dataset in order to build a predictive model. Further information on listings such as when certain Airbnb listings were brought to market and an average of reviews to the days they have been available to users will assist to regulate the analysis of individual listings. Additional specifications on the sentiments of reviews – positive vs. negative vs. neutral – will provide opportunities for the team to test the hypothesis on the relevance of reviews to bookings. Lastly, research into factors such as amenity offerings, transportation options near or around the location, safety, etc. would greatly add to the value of the model as it would assist hosts to understand consumer needs and desires. Overall, this continued analysis of the New York City Airbnb could be valuable to the company by providing consumer insights to the hosts, who in turn can work to enhance the desirability of their listings on a feature or service-level as well as adjust price accordingly, leading to potential increases in profit for Airbnb.

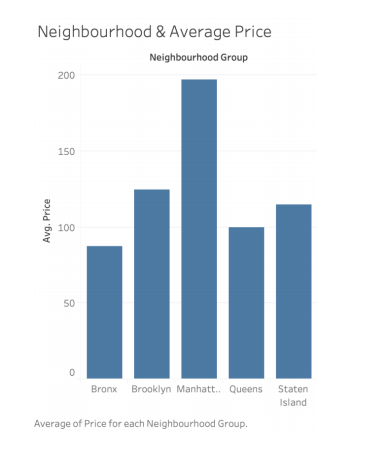
# Appendix

**Individual Report – Seorin Chae**

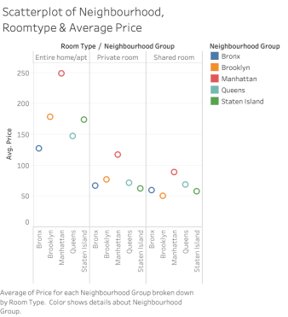
I initially contributed to the group project by helping to assess and seek out suitable datasets for the report. Once the NYC Airbnb dataset was selected, I researched the individual variables provided in the spreadsheet to determine which variables to explore in our research and the limitations of others as determinants for trends. I used this to create hypotheses and research questions. Additionally, I researched through the *Kernel* section of the *Kaggle* website, along with additional research online to understand how others’ have chosen to approach this dataset. References are shared below.

Once the template was created by Bhargavi, I built out the *Abstract, Introduction, Methodology*, and *Future Works* portion of the paper, along with my individual report. Bhargavi and I co-created *Analysis, Results, and Findings* portion of the paper along with the entire presentation. This was then presented together along with Kaushik’s choropleth.

Provided below are some of the charts I created for the group project along with explanations on how they were created and what information regarding the data they each reveal. A combination of R (codes provided) and Tableau were used.



For the initial assessment, I created an *Average Price vs. Neighborhood Group* bar chart. This plot displays the average price by neighborhood groups. Through this bar chart, we’re able to identify Manhattan as the most expensive neighborhood group, followed by Brooklyn, Staten Island, Queens, and Bronx.



After understanding the pricing of each neighborhood groups, we decided to examine the average price of the neighborhood groups by room type. Some key observations that can be noted in this scatterplot is as follows:

1. All five neighborhood groups offer primarily two room offerings – private room and shared room, with few offerings for entire home/apt
   1. This potentially can be explained by the affluence of the neighborhood & is indicative of the wallet size of those who chose to rent/live there (even for short duration)
2. Manhattan is the most expensive across room types – Entire home/apt, Private room, and Shared room.
3. Brooklyn has the second most expensive offerings in terms of Entire home/apt and Private room; however, a shared room in Queens is more costly than a shared room in Brooklyn.
4. Queens has the third most expensive private room offerings.
5. Brooklyn has the cheapest Shared room offerings
6. Staten Island has the cheapest Private room offerings



A free text analysis was conducted on the names of Airbnb listings through the creation of a word cloud. Above, the most frequently utilized terms are highlighted when hosts create their offerings.



Interactive link: <http://rpubs.com/schae4/551667>

The interactive graph of New York City provides all Airbnb listings in clusters based on sub-regions. If you click on the clusters, you are able to see the listing information in detail – Listing Name, Host Name, Price, and Room Type. On a high level, the clustering allows the viewer to see the distribution of listings and differentiate between high-density and low-density areas for Airbnbs in New York City. The map here states that the maximum listings are in or around Manhattan, Brooklyn, followed by Queens and Bronx. Staten Island region has the least number of listings during 2011-2019.

Furthermore, a closeup of the leaflet can provide an in-depth assessment into the price distribution – minimum and maximum – as well as room type offerings within specific regions. The combination of this information, the listing distribution, along with additional variables (transportation offerings, sentiment of reviews, etc.) will allow future hosts, along with the company, to see areas of opportunity and oversaturation.

**Code:**

#set working directory

AB\_NYC\_2019 <- read.csv("~/Desktop/AB\_NYC\_2019.csv", header=TRUE)

#rename dataset

nycair <- AB\_NYC\_2019

#check sample size and number of variables

dim(nycair)

#sample size of 48896 with 16 variables

#read the first and last 6 lines of data

head(nycair)

tail(nycair)

#check for missing values

sum(is.na(nycair))

#10052 missing values

#apply listwise deletion to eliminate all missing values in dataset

nycair <- na.omit(nycair)

#check number of observations and variables after listwise deletion

dim(nycair)

#38843 observations and 16 variables

library(dplyr)

library(ggplot2)

library(choroplethr)

library(choroplethrMaps)

library(leaflet)

install.packages("leaflet")

#Creating Listings across NYC

leaflet(nycair) %>%

addTiles() %>%

addMarkers(~longitude, ~latitude,labelOptions = labelOptions(noHide = F),clusterOptions = markerClusterOptions(),popup = paste0("<b> Name: </b>", nycair$name , "<br/><b> Host Name: </b>", nycair$host\_name, "<br> <b> Price: </b>", nycair$price, "<br/><b> Room Type: </b>", nycair$room\_type, "<br/><b> Room Type: </b>", nycair$room\_type

)) %>%

setView(-74.00, 40.71, zoom = 12) %>%

addProviderTiles("CartoDB.Positron")

##########

library(readr)

library(dplyr)

library(ggplot2)

library(choroplethr)

library(choroplethrMaps)

library(choroplethrZip)

library(GGally)

library(lubridate)

library(zoo)

library(scales)

library(ggmap)

library(scales)

library(stringr)

library(zipcode)

library(leaflet)

library(extracat)

library(gridExtra)

install.packages("zipcode")

update.packages()

Yes

install.packages('knitr', repos = c('http://rforge.net', '<http://cran.rstudio.org>'),

type = 'source')

# install choroplethrZip

library(devtools)

[install\_github('arilamstein/choroplethrZip@v1.5.0](mailto:install_github('arilamstein/choroplethrZip@v1.5.0)')

# load choroplethrZip

library(choroplethrZip)

#Which area is more expensive?

nycair$price <- as.numeric(gsub(",", "", substring(nycair$price, 2)))

zipPrices <- nycair %>% group\_by(neighbourhood\_group = neighbourhood\_group) %>% summarise(avg\_price = mean(price, na.rm = TRUE))

colnames(zipPrices) <- c("region","value")

zipPrices$region <- as.character(zipPrices$region)

nyc\_fips = c(36005, 36047, 36061, 36081, 36085)

g\_price\_location <- zip\_choropleth(zipPrices,

county\_zoom = nyc\_fips,

title = "Average Price by Region",

legend = "Average Score") + ggtitle("Which area is expensive?",

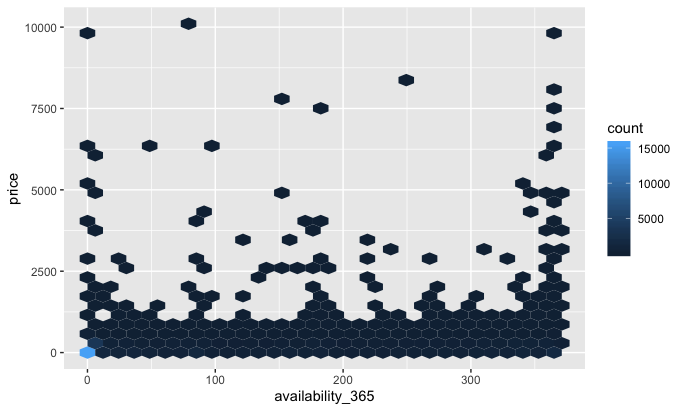
subtitle = "Map showing Average Price by Area") +

theme(plot.title = element\_text(face = "bold")) +

theme(plot.subtitle = element\_text(face = "bold", color = "grey35")) +

theme(plot.caption = element\_text(color = "grey68"))+scale\_color\_gradient(low="#d3cbcb", high="#852eaa")+ scale\_fill\_brewer("Average Price",palette=4)

g\_price\_location



A hex plot was created to display the relationship between pricing and availability of each Airbnb listings. As this plot did not bring forth valuable information regarding our desired objectives, it was eliminated from the final assessment.

**Code:**

#Seorin Chae

#DSC 465 NYC AirBnB Analysis

#Conducting initial data check, descriptives and checking multicollinearity

library(corrplot)

library(car)

library(QuantPsyc)

library(leaps)

library(psych)

library(ggplot2)

library(GGally)

library(stats)

library(DescTools)

#set working directory

AB\_NYC\_2019 <- read.csv("~/Desktop/AB\_NYC\_2019.csv", header=TRUE)

#rename dataset

nycair <- AB\_NYC\_2019

data("nycair")

str(nycair)

library(ggplot2)

library(hexbin)

Plot1 <- ggplot(nycair, aes(number\_of\_reviews, availability\_365))

Plot1 + geom\_hex()

#Number bins in each direction:

Plot1 + geom\_hex(bins = 10)

#specifying width

Plot1 + geom\_hex(binwidth = c(1, 10000))

Plot2 <- ggplot(nycair, aes(availability\_365, price))

Plot2 + geom\_hex(bins = 30)

**Key Learnings:**

Through this project and coursework, I was able to go deeper into the decision-making process of creating audience and message-friendly visualizations. In previous DSC courses, much of the course material focused on the technical functionalities of tools such as R and SAS and the various predictive and descriptive models to create. This class took it a step further by introducing the ways in which your intended message can be emphasized or misrepresented through the application of vizualization rules. I learned various new techniques to visualize data – such as hex plot, rose plot, cartogram, etc. - but also gained an understanding of when to apply certain methods over others. The coursework taught me to think critically about every detail of a chart/graph as a slight variation in line, text, and color can irreparably skew audience perception. Furthermore, I was able to learn a new tool (Tableau) while discovering new aspects of R that I had never explored before. Overall, the coursework and project was challenging and required repetition and experimentation, but has been immensely rewarding due to the knowledge acquired.

**Individual Report - Bhargavi Kashyap**

I contributed towards finding dataset for the report and understanding the usability of the dataset to perform the tasks required for the visualization project. Along with this, I did the initial data pre-checks, outlined the potential variables for exploration (initial hypothesis based on [research](http://www.columbia.edu/~sg3637/airbnb_final_analysis.html) & references to [past work](https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data/kernels)). I also looked at [Airbnb database](http://insideairbnb.com/new-york-city/) to understand how they categorize and analyze data.

Additionally, I created the templates for submission and worked on the *Abstract, Introduction, Methodology and Analysis* section of the paper with Seorin Chae, along with my analysis. For the overall project, I contributed towards maps (NYC map based on neighborhood and price, year-on-year price growth, data distribution plots (i.e. boxplots, density plot, scatterplot, violin plot), rose plot.

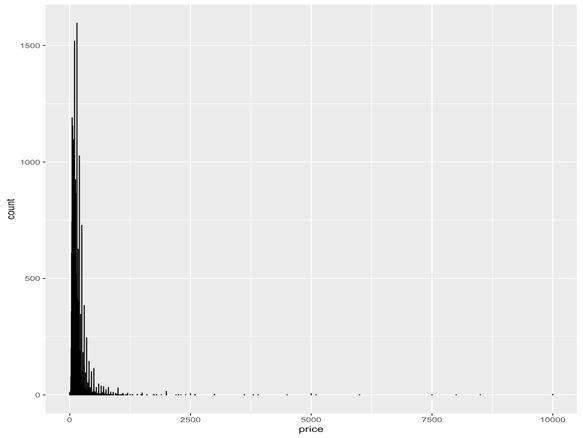
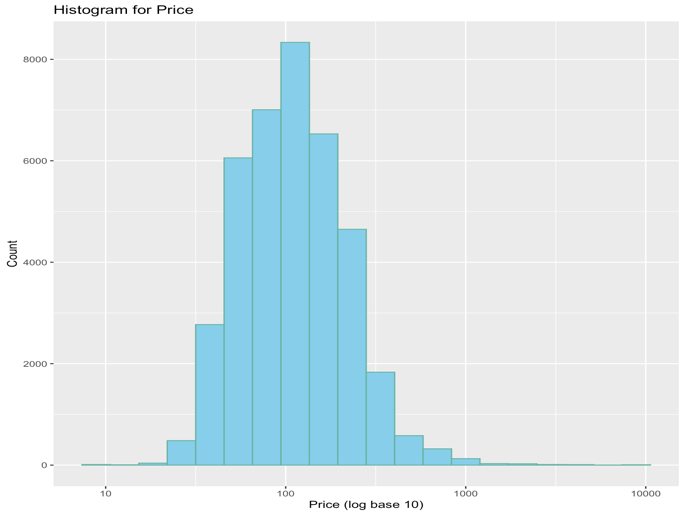
Provided below is a summary of the steps taken to understand the trends in the NYC Airbnb. Both R and Tableau were used to create graphs.

**Data Cleaning/ Pre-processing**

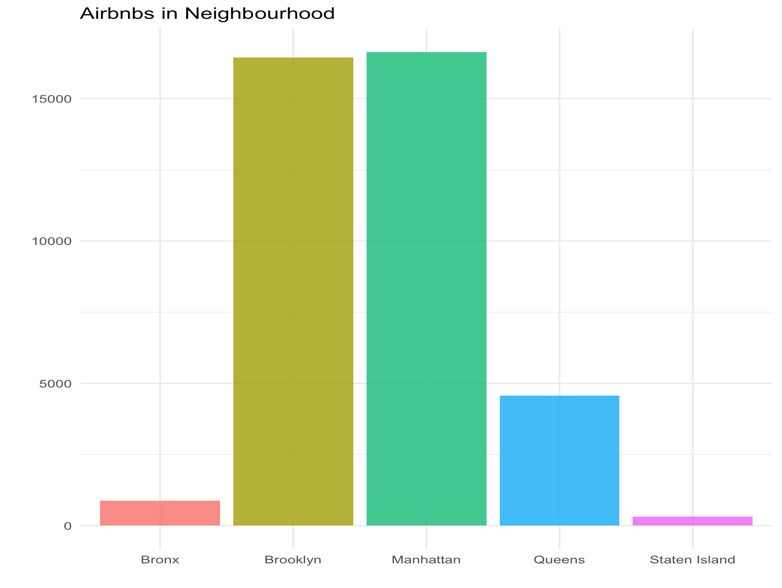
* Overall objective was to understand the trends prevalent in NYC Airbnb i.e. popular neighborhoods, room-type, pricing based on minimum nights of stay, number of reviews, availability, listings
* Few columns such as ID, host ID, host name was removed as they were not relevant for analysis
* Descriptive statistics (i.e. mean, standard deviation etc.) was evaluated to understand the distribution of data
* Dataset was checked for missing data, especially in variable of interest such as price.
* Missing data was addressed using listwise deletion
  + The skewness and kurtosis don’t change & hence listwise deletion of data is effective & does not affect the integrity of data

**Data Visualization**

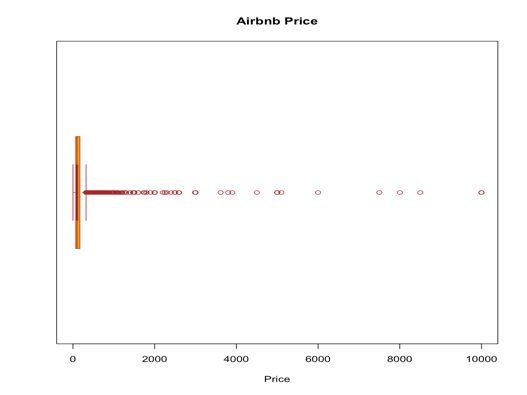
* ***Exploratory Analysis***
  + Histogram of price
    - As the descriptive statistics indicated, the mean of the price was $142.32, standard deviation of $196.95 and kurtosis of 953.73.
    - Being a highly skewed data, initial histogram was not easy to read. For better clarity, the x-axis was log transformed with log base 10.

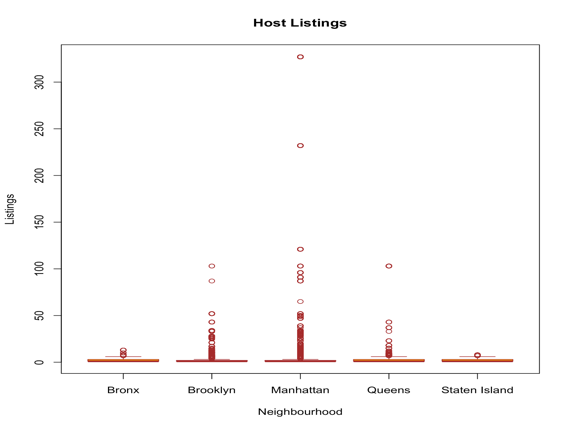
* + Distribution of Airbnb’s in neighborhood
    - Manhattan has the greatest number of properties, closely followed by Brooklyn. This could be because of the popularity of the neighborhood, both for tourists and among the working class. (Both locations are classified as central business districts). Both these neighborhoods have more than 80% of the Airbnb properties in all of NYC.
    - Queens, Bronx and Staten Island has the least number of properties, in that order



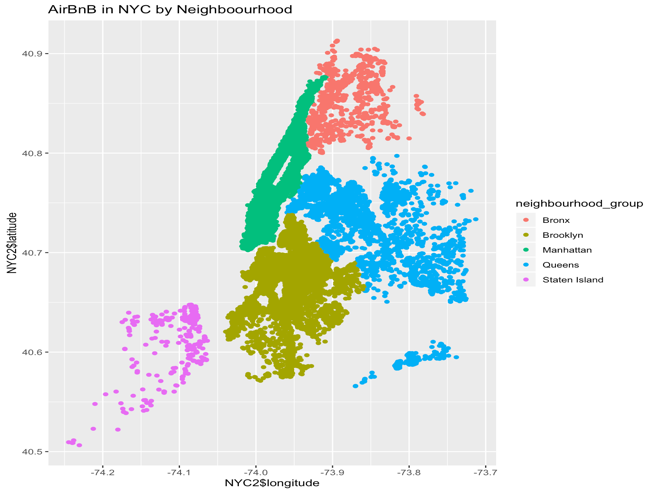
* + Boxplot of price distribution – To better understand the outlier and distribution. Both vertical & horizontal box plots are shown below.

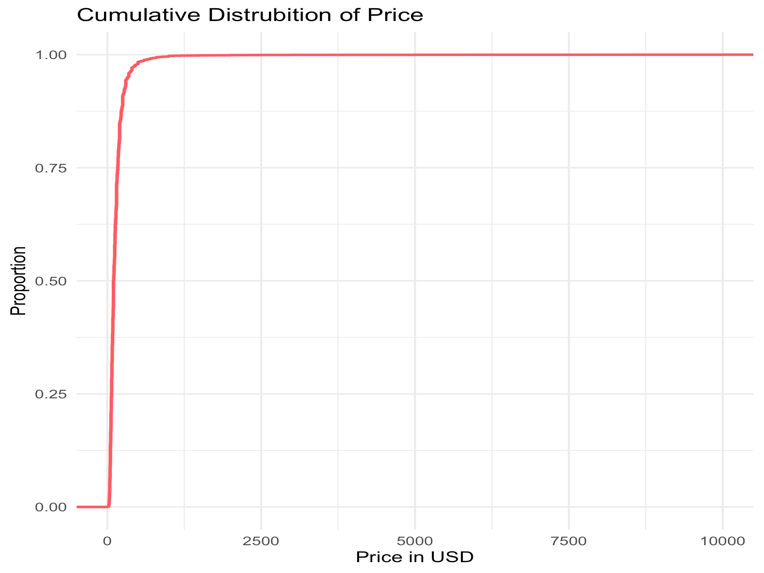
* + Boxplot of neighborhoods and host listing – To better understand the outlier and distribution. Both vertical & horizontal box plots are shown above
  + Listings across neighborhoods – Manhattan had the most listings (as shown by the boxplot below), followed by Brooklyn, Queens, Bronx and Staten Island



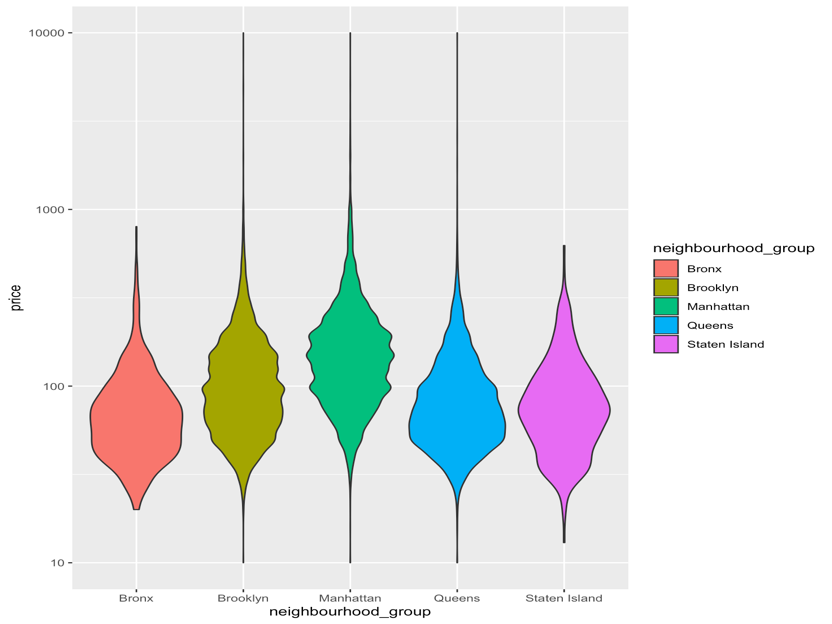
* + Correlation among variables via Scatterplot:
    - Scatterplot of price, minimum nights and number of reviews, across room type was created.
    - Click [here](http://rpubs.com/BKashyao/552882) to view the graph. The graph produced is *unable* to differentiate the relationship among the variable and is *not* used for further analysis
* ***Explanatory Analysis***
  + Distribution of properties across neighborhoods
    - NYC neighborhood data was mapped using the longitude and latitude information in the dataset
    - The graph below is the representation of density of properties across NYC. As understood during exploratory analysis, Manhattan and Brooklyn have the most density and Staten Island has the least
    - No colorblind save color combination was available with five factors on Color Brewer, however, hue and saturation was experimented with for the best clarity



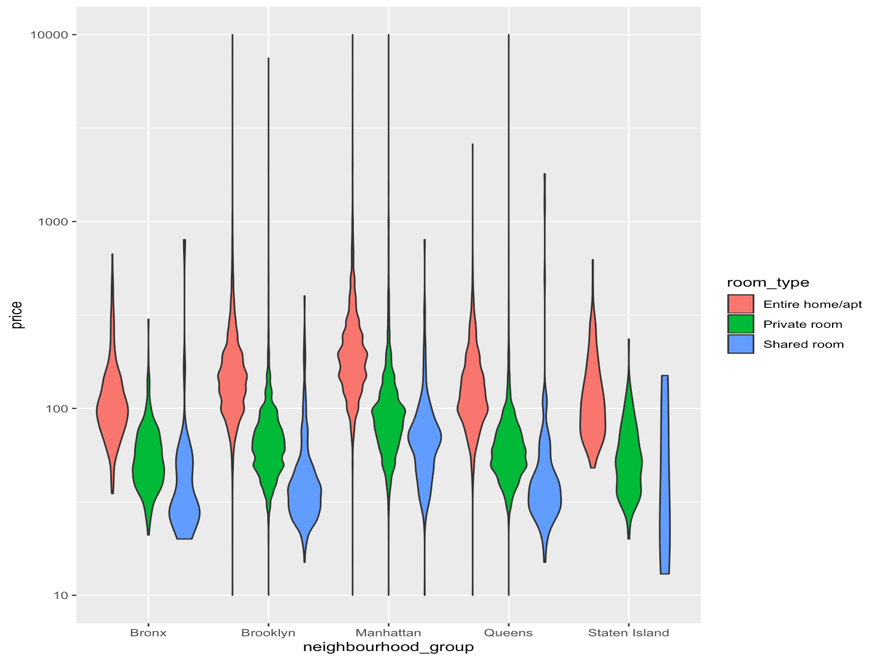
* + Distribution of Price
    - The cumulative distribution plot indicates that consumers can chose most of the Airbnb properties in NYC under $1250



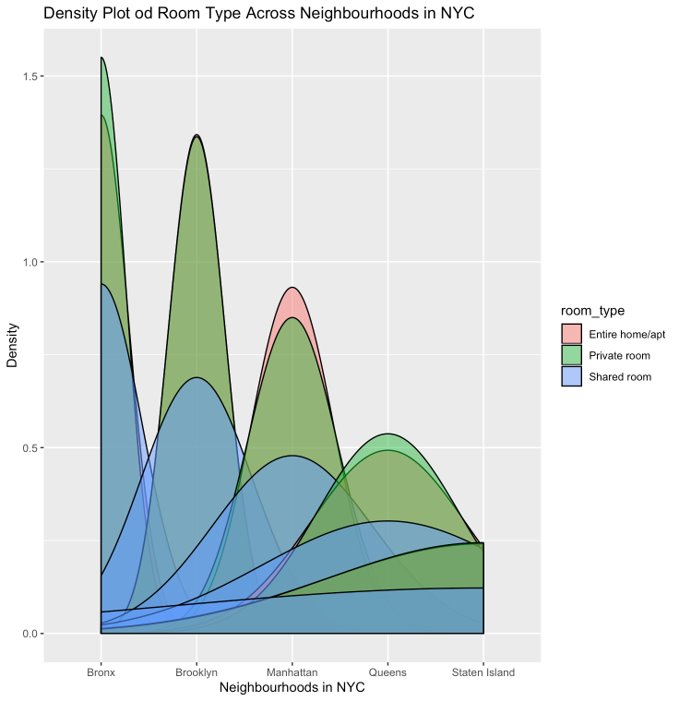
* + Relationship of price across neighborhoods and room types using Violin Plots
    - Two violin plots were created (price across neighborhood and price across neighborhood, distributed by type of room) to visualize the distribution of the data and its [probability density](https://en.wikipedia.org/wiki/Probability_density_function)
      * Price across neighborhood



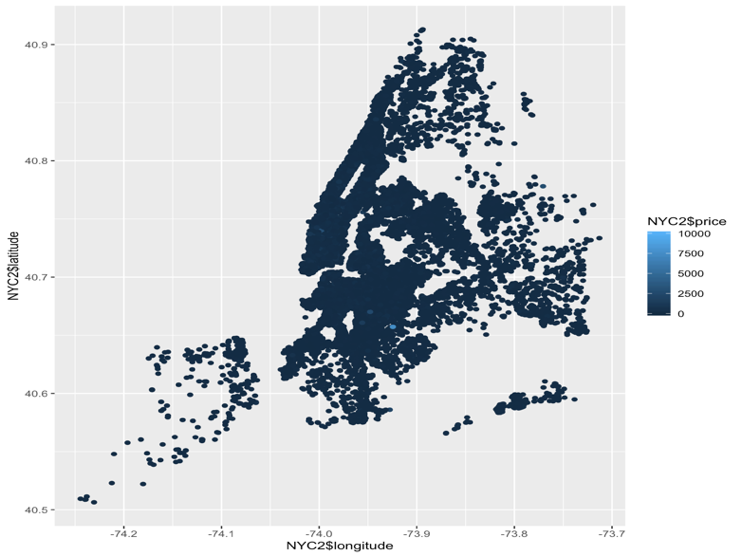
* + - * Price across neighborhood with room-type differentiation



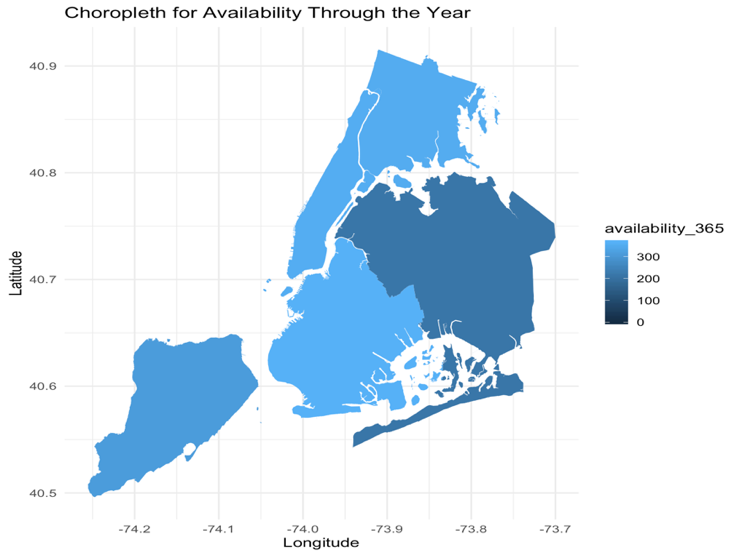
* + Room type & neighborhoods – Density of



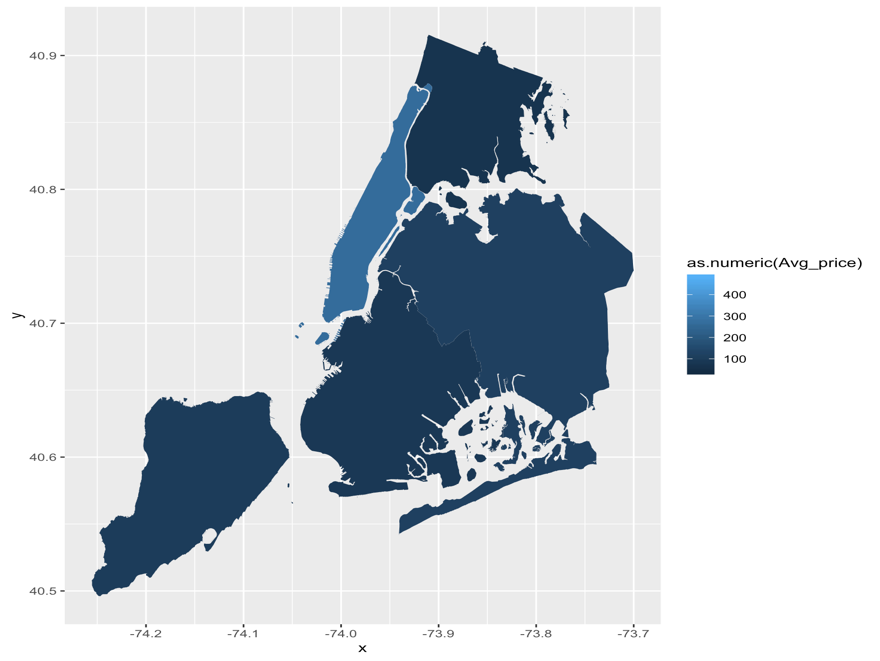
* + Price across neighborhoods
    - An [interactive map](https://maildepaul-my.sharepoint.com/personal/bkashyap_mail_depaul_edu/Documents/DSC%20465%20-%20Group%20Project/(http:/rpubs.com/BKashyao/552316) of NYC was created using longitude and latitude information and pricing was mapped
    - Colorblind safe color palette was used



* + Choropleth
    - Availability across the year in NYC neighborhood groups



* + - Average pricing NYC neighborhood groups based on neighborhoods within

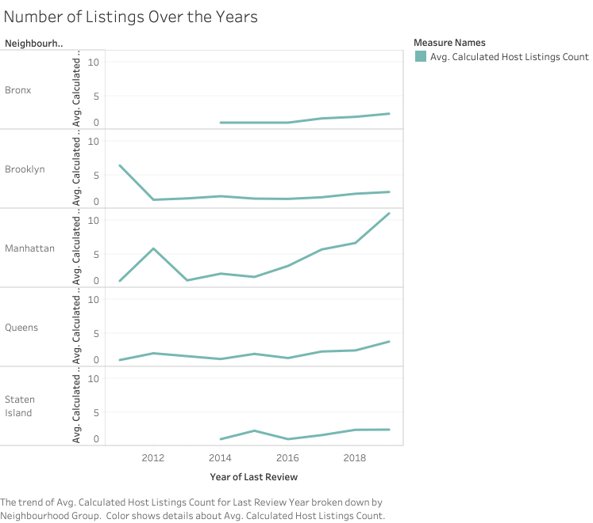


* Pricing trend over the years
  + Listings in Bronx became available beginning in 2014 and has seen minimal changes in price since then.
  + Listings in Brooklyn averaged the $200’s but saw a drop in 2013 and has plateaued over the years.
  + Manhattan listings averaged around $100’s in 2011 and saw a sharp incline in 2013, dropped approximately half its growth, and has remained steadily between the $150-$200 mark.
  + Listings in Queens saw a minor peak in 2016 but dropped and normalized quickly thereafter and has remained steady.
  + Staten Island offerings became available in 2014 with prices seeking the fastest growth amongst the neighborhood groups. In 2015, they even exceeded Manhattan’s average price, but quickly fell to the lowest cost range by 2016. This may imply the novelty of the services in the area in combination with the new hosts lack of knowledge of the market demands led to a high price mark, but the eventual regulation of the market.

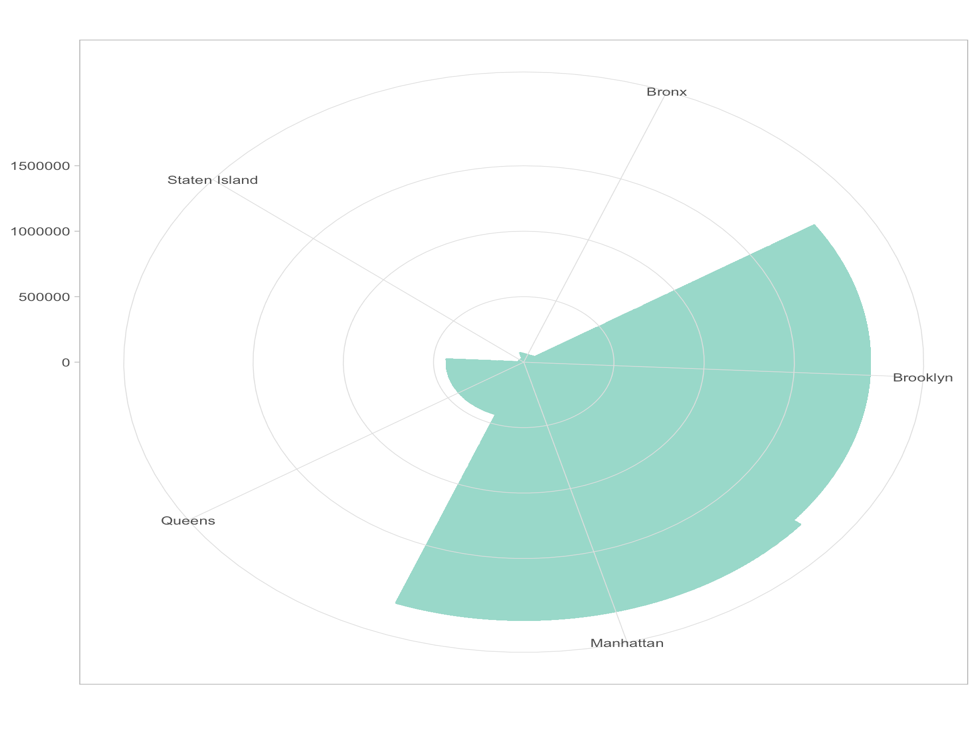


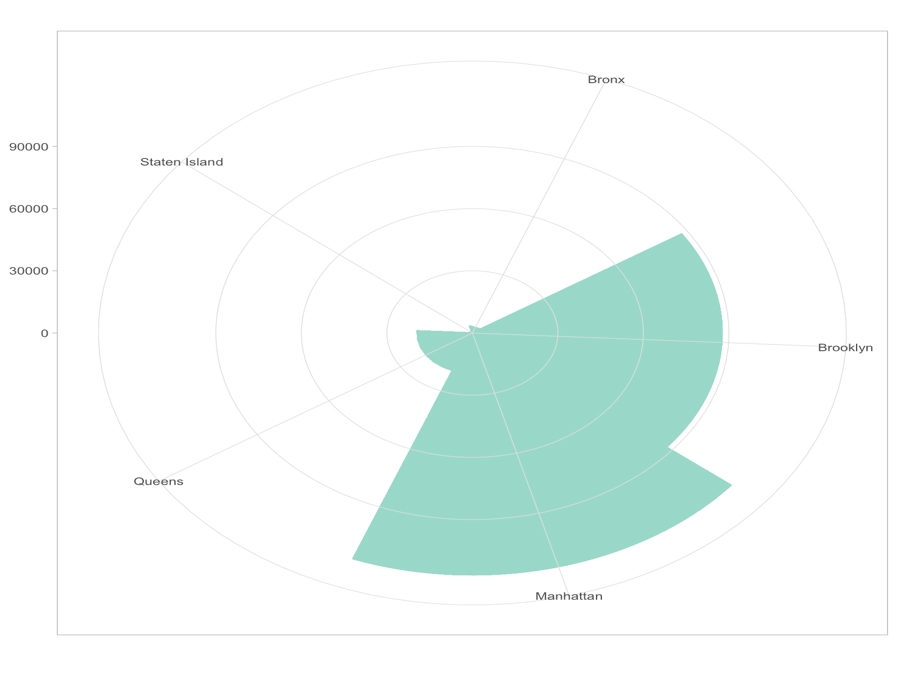
Lastly, we tried to understand the trends in the number of listings in various neighborhood groups between the years 2011 to 2019. Some key observations that can be noted in this scatterplot is as follows:

1. Listings in Bronx became available beginning in 2014 and has seen a gradual increase in offerings beginning in 2016.
2. Brooklyn listing availabilities dropped significantly in 2012 and has maintained stability over the years.
3. Manhattan listing availabilities fluctuate the most over the span of nine years, seeing a peak in 2012 and a steep drop in 2013, with persistent increase all the way up to 2019.
4. Listings in Queens have seen a gradual increase over the years in listing. The neighborhood has a reputation of not being as affluent. The multiple peaks display how the reputation of the neighborhood has changed and thus, popularity and listings.
5. Listings in Staten Island were apparent in 2014, growth curve similar to price, but deteriorated towards 2016. The listings increased gradually implying the popularity of the neighborhood and also quality of properties offered.



* Price across neighborhood groups using rose plot





**Key Learnings:**

* Better understanding of techniques learned in earlier classes
  + Importance of understanding message and audience and then applying techniques to deliver the intended message and not just blindly following techniques
* Better understanding of R syntax and its usage
* New techniques as visual representations (beeswarm plot, geographical mapping, cartograms and choropleths)
* Developed acumen to critique visual representation of data and how audience can be misinformed
  + Refreshed my knowledge of color and representation
  + Usage of mathematical tools to fine tune visualizations
* Knowledge of a new tool (tableau) and intense practice of R
  + Techniques such as mapping, using various colors and aesthetics in R were tough to learn but felt extremely rewarding once achieved
  + Basic understanding of Tableau, though I am certain there is a lot of that I haven’t yet explored (I was intrigued by the timeseries mapping and I am hoping to replicate that in R)

**R Code:**

*(Please do note that the code segment provided below is a fairly sieved through and all trails and iterations are not included. Also, not all the graphs created were used for analysis)*

#Title: Exploratory and Explanatory Data Analysis and Visualization of NYC Airbnb Dataset

#Author: Bhargavi Kashyap

#Student ID: 1946134

#import dataset & set directory

setwd("~/Documents/DePaul University - Marketing Analytics /Quarter 3/DSC 465 - Data Visualization/Group Project")

#read dataset

NYC <- read.csv(file="AB\_NYC\_2019.csv", header=TRUE, sep=",")

#understanding daraset

dim(NYC)

#48895 observation in 16 variables

View(NYC)

head(NYC)

tail(NYC)

str(NYC)

#remove columns ID, host ID, host name

NYC1 <- NYC [,c(2,5:16)]

View(NYC1)

#Show descriptive statistics

library(psych)

describe(NYC1)

#check for missing data

sum(is.na(NYC1))

#10052 missing values

#addressing missing variables

#listwise deletion method; creating a new dataset

NYC2 <- na.omit(NYC1)

#checking for missing data after listwise deletion

sum(is.na(NYC2))

#no missing value

#Check Sample Size and Number of Variables

dim(NYC2)

#38843 observations in 13 variables

describe(NYC2)

#basic descriptive of the dataset has not changed. Minimal chnages that does not imapct the data integrity and visulization

#Show for first and last 6 rows of data

head(NYC2)

tail(NYC2)

#shownames of variable

names(NYC2)

View(NYC2)

#structure of dataset

str(NYC2)

#Exploratory analysis graphs for each subsetted variable

library(GGally)

library(stats)

#Exploratory Analysis

#Histogram - distributiom of price

library(dplyr)

library(ggplot2)

#simple histogram

ggplot(NYC2) +

geom\_histogram(aes(x = price),

binwidth = 0.05, fill = "blue", color = "black")

#histogram with log-transformed x-axis

NYC2 %>%

ggplot(aes(x= NYC2$price)) +

geom\_histogram(bins=20, fill='skyblue', color='#69b3a2') + scale\_x\_log10() + labs(title="Histogram for Price", x="Price (log base 10)", y="Count")

#histogram - neighbourhood grooups - to understand no. of airbnbs in each neighbourhood

ggplot(NYC2) + 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("Airbnbs in Neighbourhood")

#violin plot

ggplot(NYC2, aes(x=neighbourhood\_group, y=price, fill=neighbourhood\_group)) +

geom\_violin() + scale\_y\_log10()

ggplot(NYC2, aes(x=neighbourhood\_group, y=price, fill=room\_type)) +

geom\_violin() + scale\_y\_log10()

#boxplot of neighbourgood groups and price distribution

boxplot(NYC2$price,

main = "Airbnb Price",

xlab = "Price",

col = "orange",

border = "brown",

horizontal = TRUE,

notch = TRUE)

boxplot(NYC2$price,

main = "Airbnb Price",

xlab = "Price",

col = "orange",

border = "brown")

boxplot(price~neighbourhood\_group,

data=NYC2,

main="Price Distribution Across Neighbourhood",

xlab="Neighbourhood",

ylab="Price",

col="orange",

border="brown")

boxplot(calculated\_host\_listings\_count~neighbourhood\_group,

data=NYC2,

main="Host Listings",

xlab="Neighbourhood",

ylab="Listings",

col="orange",

border="brown")

ggplot(NYC2, aes(price, neighbourhood\_group, color=neighbourhood\_group)) +

geom\_point() +

ggtitle("Cost by Neighbourhood") +

xlab("Price") + ylab("Neighbourhoods")

#scatterplot

library(car)

library(RColorBrewer)

Scatter\_color <- brewer.pal(nlevels(as.factor(NYC2$room\_type)), "Set2")

scatterplotMatrix(~price+minimum\_nights+number\_of\_reviews|room\_type, data=NYC2 ,

reg.line="" , smoother="", col=Scatter\_color ,

smoother.args=list(col="grey") , cex=1.5 ,

pch=c(15,16,17) ,

main="Scatter plot")

scatterplot(price ~ calculated\_host\_listings\_count, data = NYC2)

scatterplot(price ~ number\_of\_reviews, data = NYC2) + scale\_x\_log10() + scale\_y\_log10() #this comes to null (no chan ge in graph)

#Distribution of Price

ggplot(NYC2, aes(price)) +

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

ylab("Proportion") + xlab("Price in USD") + theme\_minimal(base\_size = 15) +

ggtitle("Cumulative Distrubition of Price")

#Properties Across the Year based on last review date

ggplot(NYC2) +

geom\_histogram(aes(last\_review), stat = "count", fill = '#fd5c63',alpha = 0.85) +

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

ggtitle("Properties Across the Year ")

#Small Multiple

#basic scatterplot

library(ggplot2)

ggplot(data = NYC2, aes(x = price, y = number\_of\_reviews)) +

geom\_point()

#breaking data by room type

ggplot(data = NYC2, aes(x = price, y = number\_of\_reviews)) +

geom\_point() +

facet\_wrap(~room\_type)

#transformed for ease of reading

ggplot(data = NYC2, aes(x = price, y = number\_of\_reviews)) +

geom\_point() +

facet\_wrap(~room\_type) + scale\_x\_log10()

ggplot(data = NYC2, aes(x = price, y = number\_of\_reviews)) +

geom\_point() +

facet\_wrap(~neighbourhood\_group)

#transformed for ease of reading

ggplot(data = NYC2, aes(x = price, y = number\_of\_reviews)) +

geom\_point() +

facet\_wrap(~neighbourhood\_group) + scale\_x\_log10()

#breaking data by neighbourhood group

ggplot(data = NYC2, aes(x = price, y = calculated\_host\_listings\_count)) +

geom\_point() +

facet\_wrap(~neighbourhood\_group)

ggplot(data = NYC2, aes(x = price, y = calculated\_host\_listings\_count)) +

geom\_point() +

facet\_wrap(~neighbourhood\_group) + scale\_x\_log10()

#Hex Plot

Plot1 <- ggplot(NYC2, aes(number\_of\_reviews, price))

Plot1 + geom\_hex()

#log transforming

Plot1 + geom\_hex() + scale\_x\_log10() + scale\_y\_log10()

#Number bins in each direction:

Plot1 + geom\_hex(bins = 10)

Plot1 + geom\_hex(bins = 10) + scale\_x\_log10() + scale\_y\_log10()

#specifying width

Plot1 + geom\_hex(binwidth = c(1, 10000))

Plot1 + geom\_hex(binwidth = c(1, 10000))+ scale\_x\_log10() + scale\_y\_log10()

Plot2 <- ggplot(NYC, aes(availability\_365, price))

Plot2 + geom\_hex(bins = 30) + scale\_x\_log10() + scale\_y\_log10()

#Kernel Density Plot

#neighbourhood groups and room types

qplot(neighbourhood\_group, data=NYC2, geom="density", fill=room\_type,

alpha=I(.5),

main="Density Plot od Room Type Across Neighbourhoods in NYC",

xlab="Neighbourhoods in NYC",

ylab="Density")

#

qplot(neighbourhood\_group, minimum\_nights, data=NYC2, geom=c("boxplot"),

color=calculated\_host\_listings\_count,

main="Q Plott",

xlab="Neighbourhood\_group", ylab="Minimum Nights") +

geom\_smooth()

qplot(neighbourhood\_group, minimum\_nights, data=NYC2, geom=c("boxplot"),

color=calculated\_host\_listings\_count,

main="With Linear Regression Line",

xlab="Neighbourhood\_group", ylab="Minimum Nights") +

geom\_smooth(method="lm", formula=y~x)

library(dplyr)

library(ggplot2)

library(choroplethr)

library(choroplethrMaps)

library(leaflet)

#Creating Listings across NYC

leaflet(NYC) %>%

addTiles() %>%

addMarkers(~longitude, ~latitude,labelOptions = labelOptions(noHide = F),clusterOptions = markerClusterOptions(),popup = paste0("<b> Name: </b>", NYC$name , "<br/><b> Host Name: </b>", NYC$host\_name, "<br> <b> Price: </b>", NYC$price, "<br/><b> Room Type: </b>", NYC$room\_type, "<br/><b> Room Type: </b>", NYC$room\_type

)) %>%

setView(-74.00, 40.71, zoom = 12) %>%

addProviderTiles("CartoDB.Positron")

#Understanding availability based on listing and annual availability

ggplot(data = NYC2, aes(x = availability\_365, y = calculated\_host\_listings\_count)) +

geom\_point()

#Neighbourhood Group & Neighbourhoods within

library(MASS)

NYC\_Neighbourhood<- ggplot(NYC2, aes(x=price, y=calculated\_host\_listings\_count, label=rownames(NYC2))) +

geom\_text(size=3)

NYC\_Neighbourhood

NYC\_Neighbourhood + scale\_x\_log10()

NYC\_Neighbourhood1<- ggplot(NYC2, aes(x=neighbourhood\_group, y=neighbourhood, label=rownames(NYC2))) +

geom\_text(size=3)

NYC\_Neighbourhood1

#mapping overlayed on NYC

library(ggplot2)

plot(data=NYC2, x=NYC2$longitude, y=NYC2$latitude, color=NYC2$neighbourhood\_group)

#scale\_color\_hue(l=40, c=35)

#SCATTERPLOT OF NYC BY NEIGHBOURHOOD

NYC\_MapScatter <- ggplot(data=NYC2, aes(x=NYC2$longitude, y=NYC2$latitude, color=neighbourhood\_group)) + geom\_point() + ggtitle("AirBnB in NYC by Neighboourhood")

NYC\_MapScatter

#Mapping

#CHANGING COLOR

library(RColorBrewer)

brewer.pal(n = 5, name = "Set3")

NYC\_MapScatter + scale\_fill\_brewer(palette="Set3")

NYC\_MapScatter + scale\_color\_hue(l=40, c=35)

#changing tone

NYC\_MapScatter + scale\_color\_hue(l=55, c=35)

#SCATTERPLOT BY NEIGHBORTHOOD + PRICE

NYCMap\_Price <- ggplot(data=NYC2, aes(x=NYC2$longitude, y=NYC2$latitude, color=NYC2$price)) + geom\_point() + scale\_fill\_brewer(palette = "Blues")

NYCMap\_Price

library(plotly)

NYCMap\_Interactive <- ggplotly(NYCMap\_Price)

NYCMap\_Interactive

#Roseplot

NYC2$neighbourhood\_group <- factor(NYC2$neighbourhood\_group, levels = NYC2$neighbourhood\_group)

ggplot(NYC2,

aes(x = neighbourhood\_group, y = minimum\_nights \* number\_of\_reviews)) +

geom\_col(width = 1, fill = "#99d8c9", color = "#99d8c9") +

coord\_polar(start = -pi/12) + # change start value if you want a different orientation

theme\_light() +

theme(axis.title = element\_blank(),

panel.ontop = TRUE, # change to FALSE for grid lines below the wind rose

panel.background = element\_blank())

ggplot(NYC2,

aes(x = neighbourhood\_group, y = minimum\_nights)) +

geom\_col(width = 1, fill = "#99d8c9", color = "#99d8c9") +

coord\_polar(start = -pi/12) + # change start value if you want a different orientation

theme\_light() +

theme(axis.title = element\_blank(),

panel.ontop = TRUE, # change to FALSE for grid lines below the wind rose

panel.background = element\_blank())

#Choropleth

library(rgeos)

library(maptools)

library(gpclib)# may be needed, may not be

library(rgdal)

library(raster)

str(NYC2)

# MAP

NYC\_Map<- shapefile("~/Documents/DePaul University - Marketing Analytics /Quarter 3/DSC 465 - Data Visualization/Group Project/Borough Boundaries/geo\_export\_b2d118a7-29ab-4b99-b992-b608bc19aef8.shp")

# VERIFY IT LOADED PROPERLY

plot(NYC\_Map)

library(ggplot2)

NYC\_Map1 <- fortify(NYC\_Map, region = "boro\_name")

#NYC\_Map1$id <- toupper(NYC\_Map1$id) #change ids to uppercase

ggplot() + geom\_map(data = NYC2, aes(map\_id = neighbourhood\_group , fill = price),

map = NYC\_Map1) + expand\_limits(x = NYC\_Map1$long, y = NYC\_Map1$lat)

#plotting average price

ggplot() + geom\_map(data = NYC2, aes(map\_id = neighbourhood\_group , fill = as.numeric(Avg\_price)),

map = NYC\_Map1) + expand\_limits(x = NYC\_Map1$long, y = NYC\_Map1$lat)

#by availability around the year

ggplot() + geom\_map(data = NYC2, aes(map\_id = neighbourhood\_group , fill = availability\_365),

map = NYC\_Map1) + expand\_limits(x = NYC\_Map1$long, y = NYC\_Map1$lat) +

theme\_minimal(base\_size=13)+xlab("Longitude")+ylab("Latitude") +

ggtitle("Choropleth for Availability Through the Year")

#by listing

ggplot() + geom\_map(data = NYC2, aes(map\_id = neighbourhood\_group , fill = calculated\_host\_listings\_count),

map = NYC\_Map1) + expand\_limits(x = NYC\_Map1$long, y = NYC\_Map1$lat)

#neighbourhoods in overall neighbourhood groups

# MAP using NYC Zillow data

NYC\_NMap<- shapefile("~/Documents/DePaul University - Marketing Analytics /Quarter 3/DSC 465 - Data Visualization/Group Project/zillow-neighborhoods/zillow-neighborhoods.shp")

# VERIFY IT LOADED PROPERLY

plot(NYC\_NMap)

str(NYC\_NMap)

View(NYC2)

library(ggplot2)

library(ggrepel)

#average price

Avg\_price<- ave(NYC2$price, NYC2$neighbourhood)

#mapping average price

NYC\_NMap1 <- fortify(NYC\_NMap, region = "name")

ggplot() + geom\_map(data = NYC2, aes(map\_id = neighbourhood , fill = as.numeric(Avg\_price)),

map = NYC\_NMap1) + expand\_limits(x = NYC\_NMap1$long, y = NYC\_NMap1$lat) +

geom\_path(color = "white") + scale\_color\_gradient() + ggtitle("Neighbourhood Pricing Heat Map") +

coord\_equal()

Avg\_price<- ave(NYC2$price, NYC2$neighbourhood)

view(NYC2)

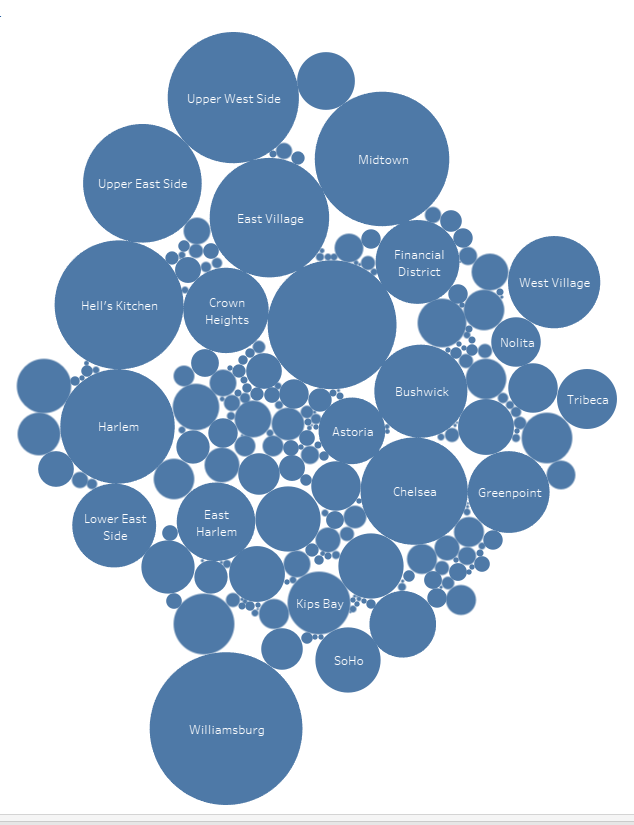
library(RColorBrewer)

Map\_color <- brewer.pal(nlevels(as.numeric(NYC2$Avg\_price)), "Set2")

Kaushik’s Analysis:

I focused on building different kind of visualizations displaying the price in various neighborhoods such as bubble chart and choropleth. I also tried to compare how the price varies with room type and what room types are abundant in what locations. I also tried to explore the number of listings in different sub neighborhoods which gives us a brief idea why price might be higher in areas with less airbnbs and so.

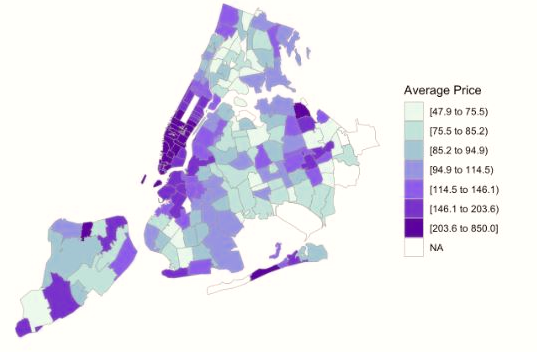
**Bubble chart displaying the prices of various neighborhoods**



We can see the pricing for various neighborhoods here like Williamsburg , the hell’s kitchen has higher prices compared to other areas.

**Choropleth**

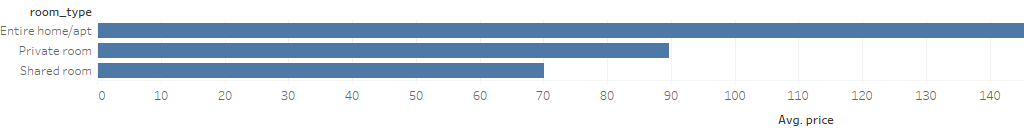
Choropleth displaying the pricing in various sub neighborhoods on a map of New York.



Downtown Manhattan is the highest when it comes to high prices and the same is true for the neighborhoods of Brooklyn close to Manhattan.

Built using <https://www.displayr.com/create-a-geographic-map/> which is a very easy web tool to make choropleths and already has many maps preloaded.

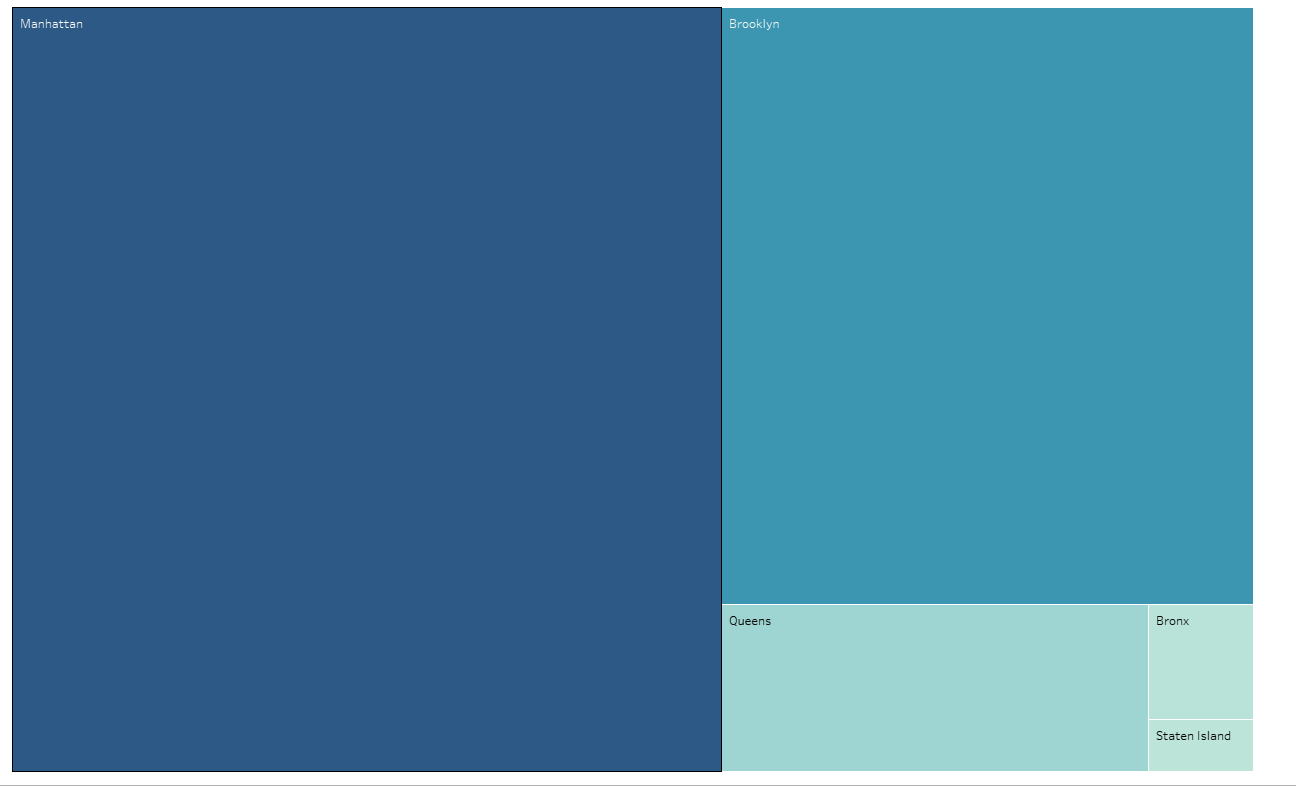
**Avg price versus room type .**



As expected, price for an entire house is greater than a private room which is in turn more than a shared room price.

**Tree map displaying pricing .**

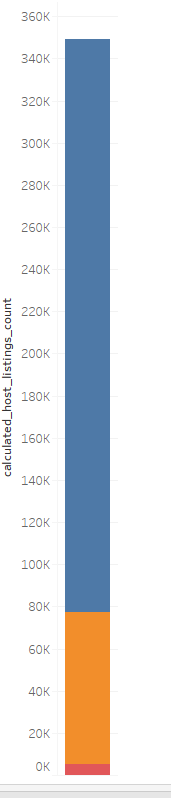
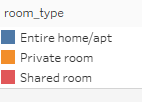
Very easy and clear chart which shows that Manhattan has the highest prices.



Tree map of neighbourhood groups and their pricing.

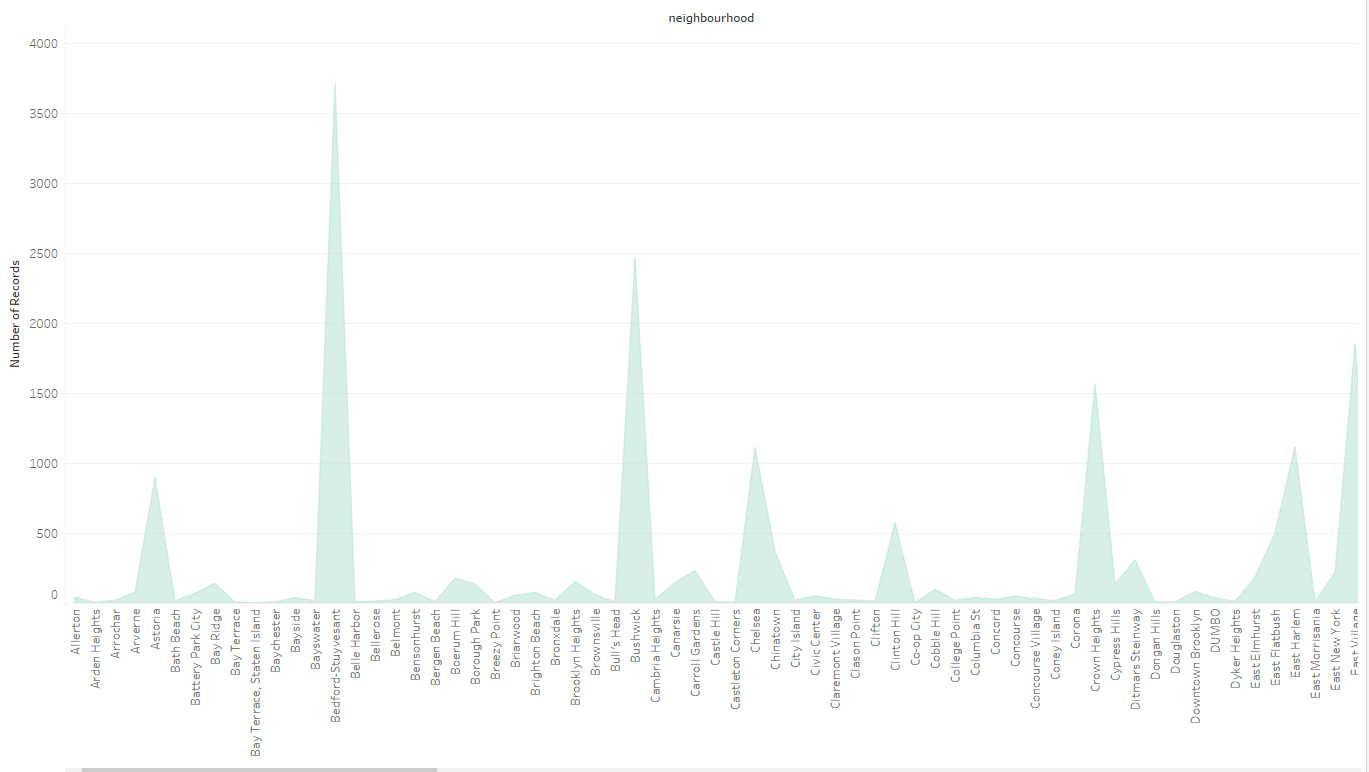
Manhattan has the highest price and straten island has the least.

**Stacked bar showing number of listings based on room type.**

We can see here that highest number of listings are entire homes or apartments followed by one or more private room and shared room listings are the least.

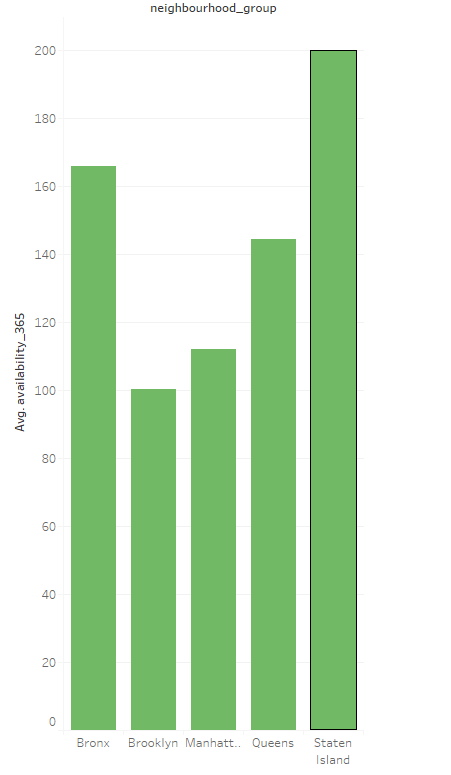
**Number of listings in various neighborhoods:**



You can see the total number of listings in various neighborhoods. This graph allows us to see how many airbnb providers are present in different sub neighborhoods.

This lets us determine the most popular or most dense airbnb zone in NYC .

**Average availability in different neighborhood groups.**



You can see here that staten island has the highest availability of listings and It is followed by bronx.

Takeaways:

I learnt that colors and human perception are very important to communicate through charts.

I also understood that the audience and their background is very important while deciding what kind of visualizations must be used. I have known that same data or trend can be represented using different charts, but we must choose the most easily understandable and affective visuals to give the message.

I’ve always wanted to learn the tableau and work on it. This class helped me to understand the tableau and gave me some valuable experience to talk about tableau. I also used some other web tools to create choropleth which was way easier than the tableau to create choropleths.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

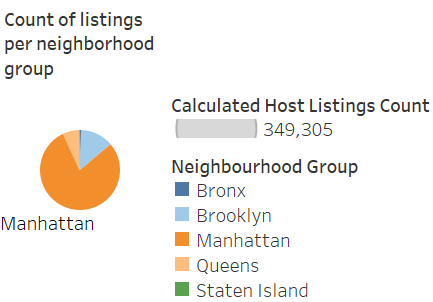
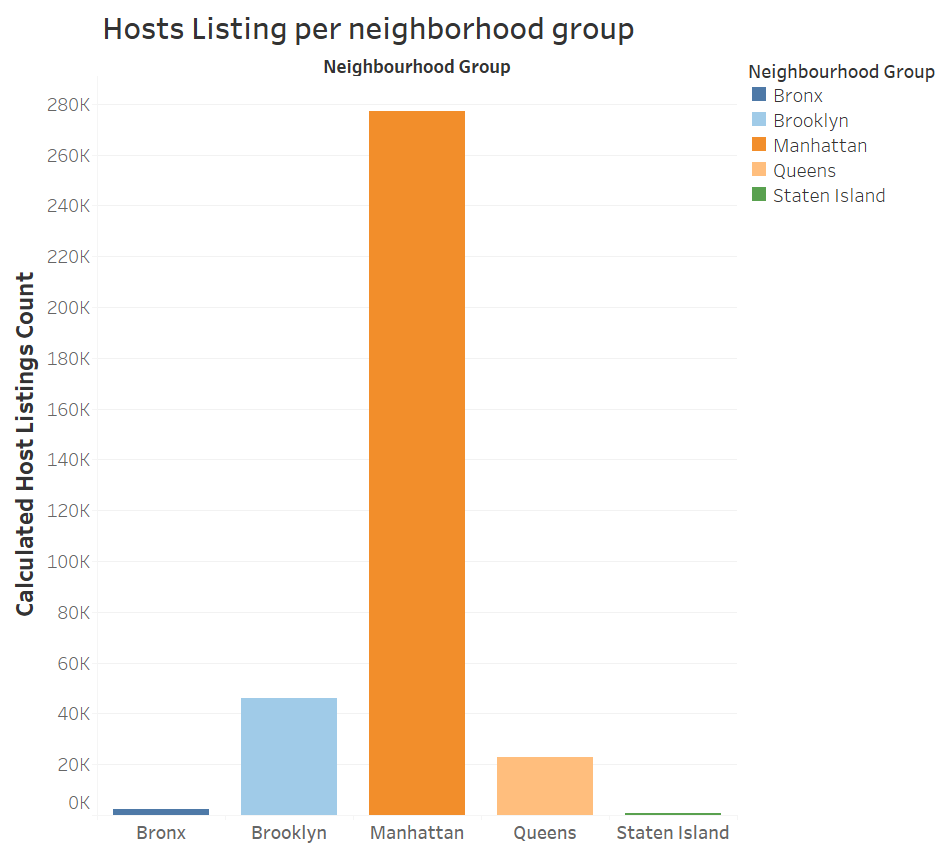
**Individual Report – Asil Qrareya**

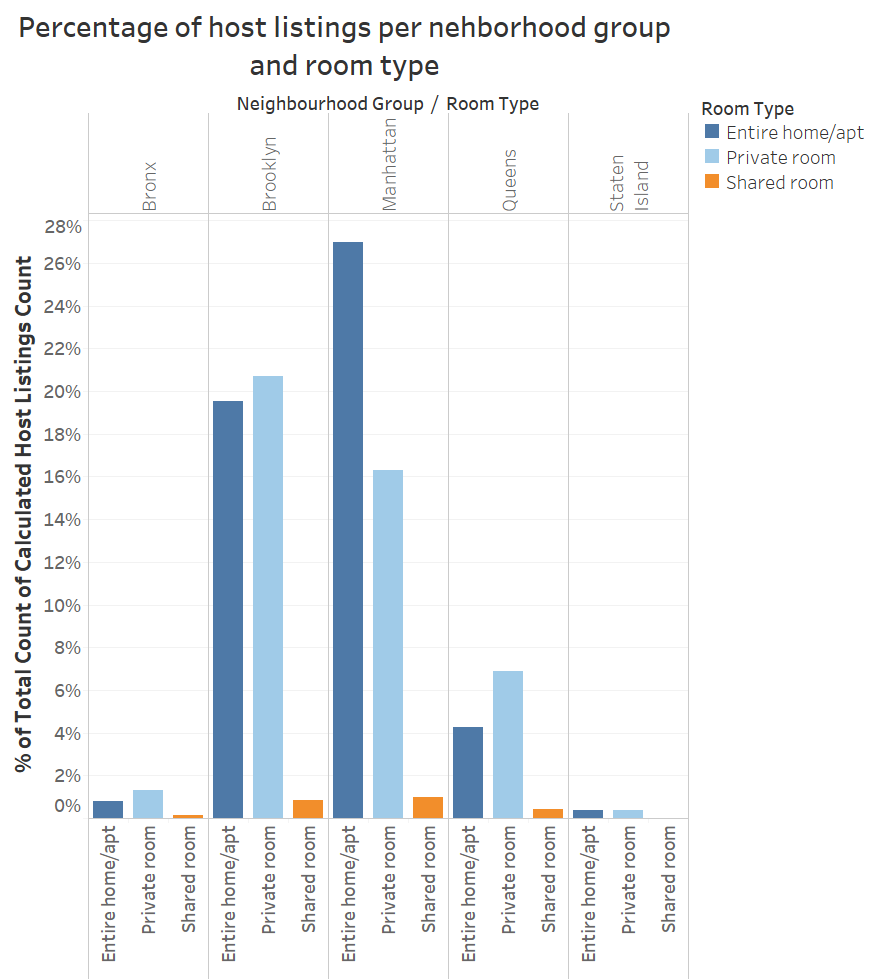
I focused on exploratory analysis in my part. After discussing the story line with the group, I decided to focus on summarizing the characteristics of the data and variables. This allowed me to check on trends and patterns in the data, what variables to focus on and in what direction to data is leading us. I used tableau in my analysis as it gave me the flexibility to test multiple variables together at the same time and play with the data to check if there is any latent variables here or there.

Analysis included the following:

1. **Overview of the count of listings per neighborhood groups and room types**

Graphs show that Manhattan neighborhood has the highest number of listings of airbnb. Graphs show also that mor than %50 of listings in NY were for entire apartments.

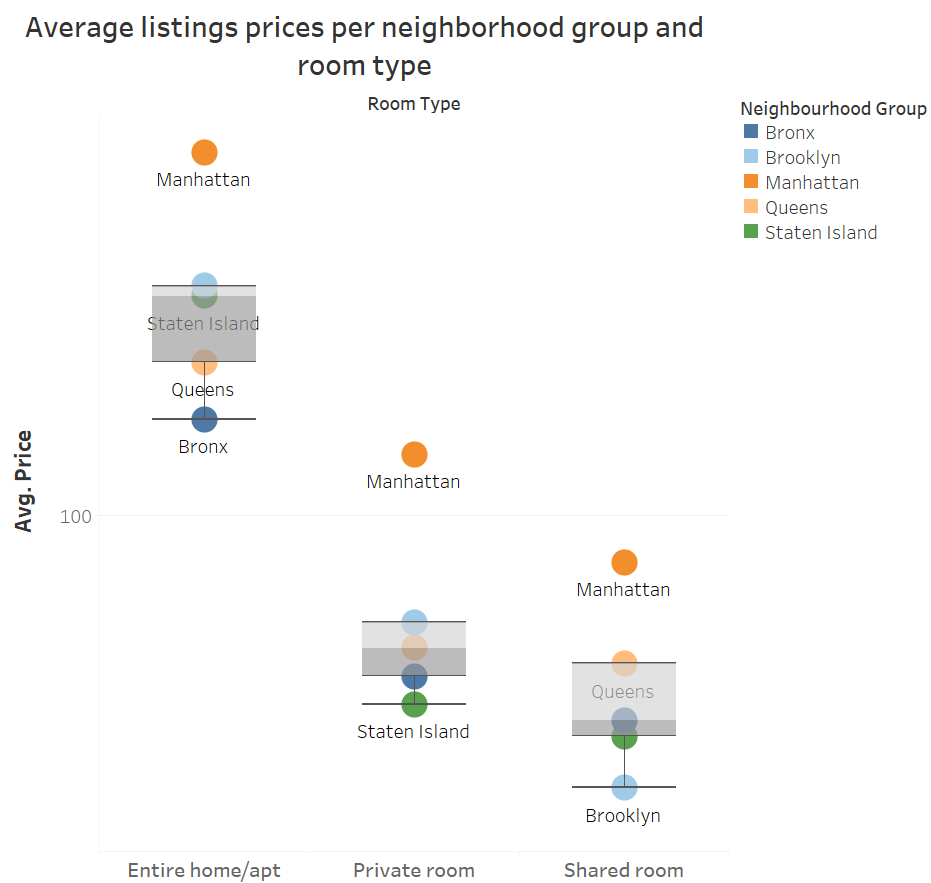


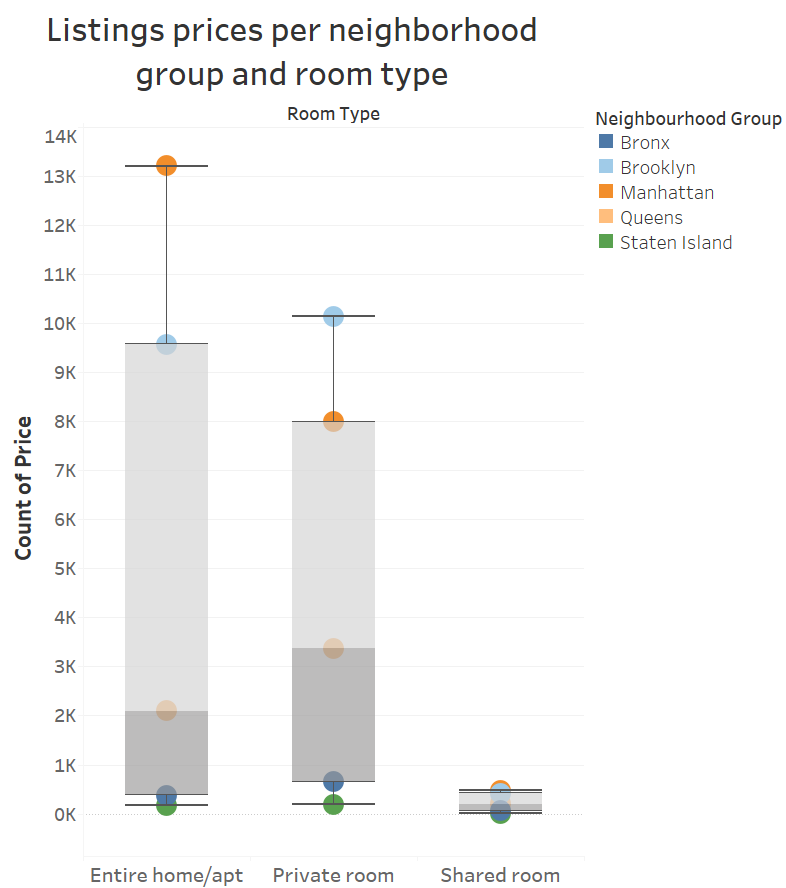


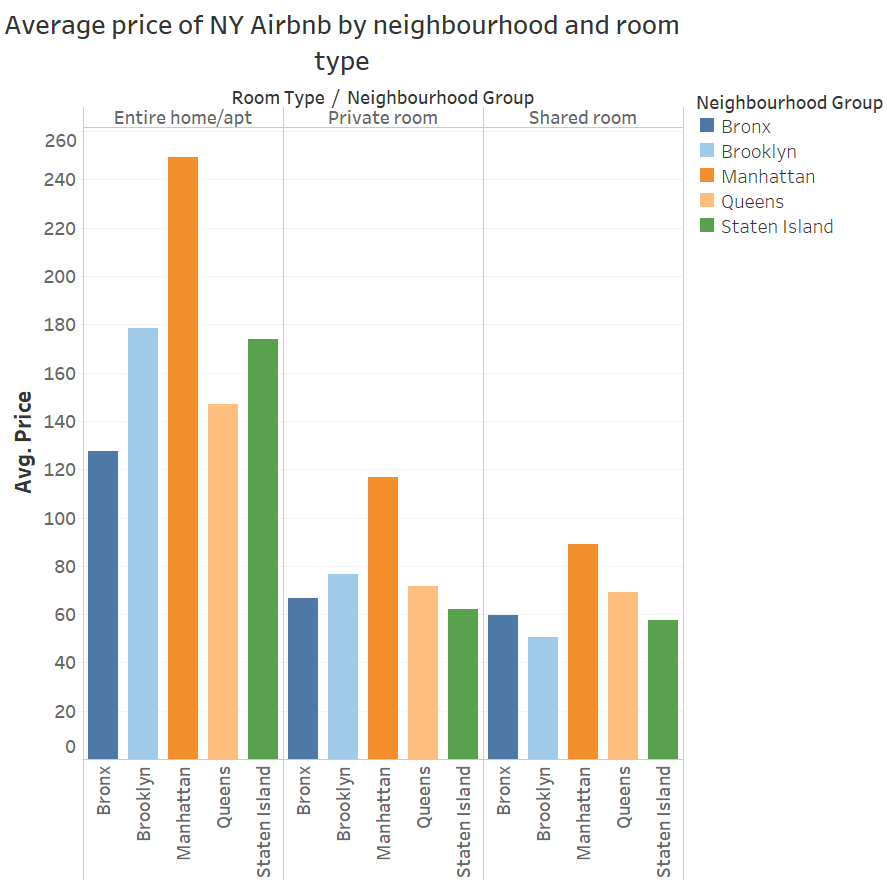
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Overview of price per neighborhood groups and room types**

I graphed the price variable using different metrics; count and average. First graph below shows that, on average, Manhattan prices tend to be the highest among other neighborhood groups within the three type of room we have; entire apartment, private room and shared room. Graph of count of prices shows that Manhattan continues to have the highest prices of entire apartments and even way higher than the average count. As for private and shared rooms, Manhattan prices falls within the average. This can be an indicator for customers who is searching for a room in manhattan, that there are options that fall within the average of overall prices, if he/she is searching for a private or shared room.



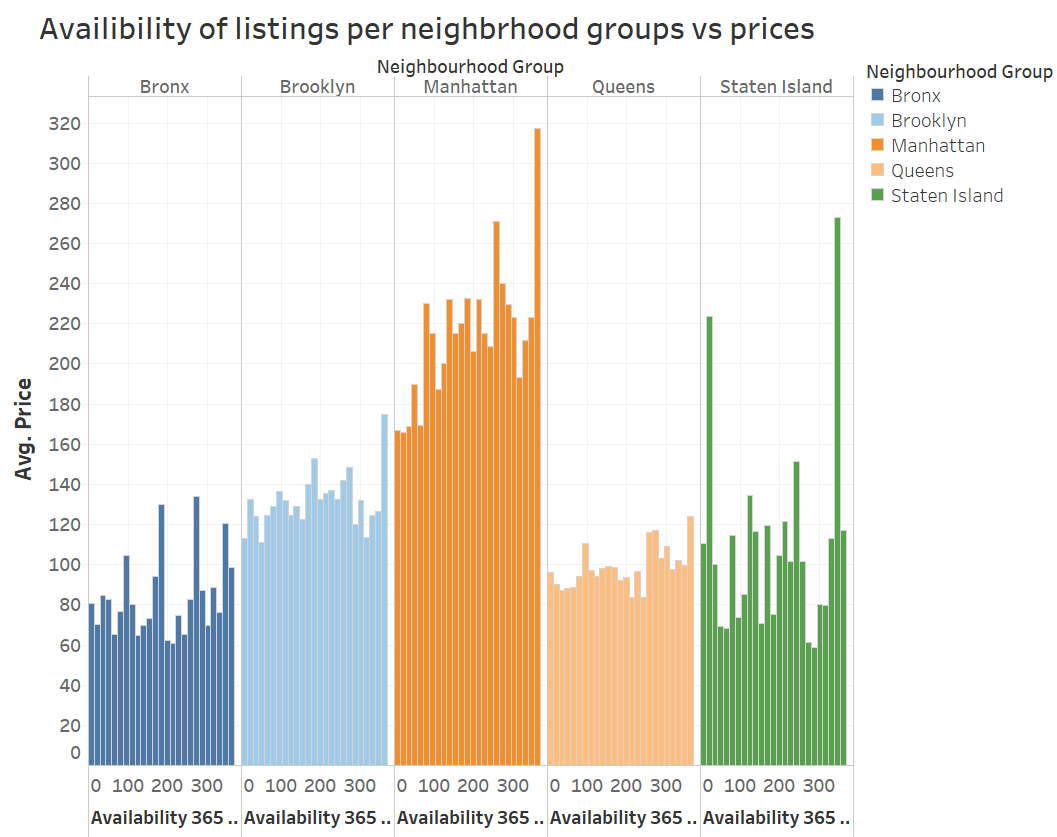




\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Availability of listings**

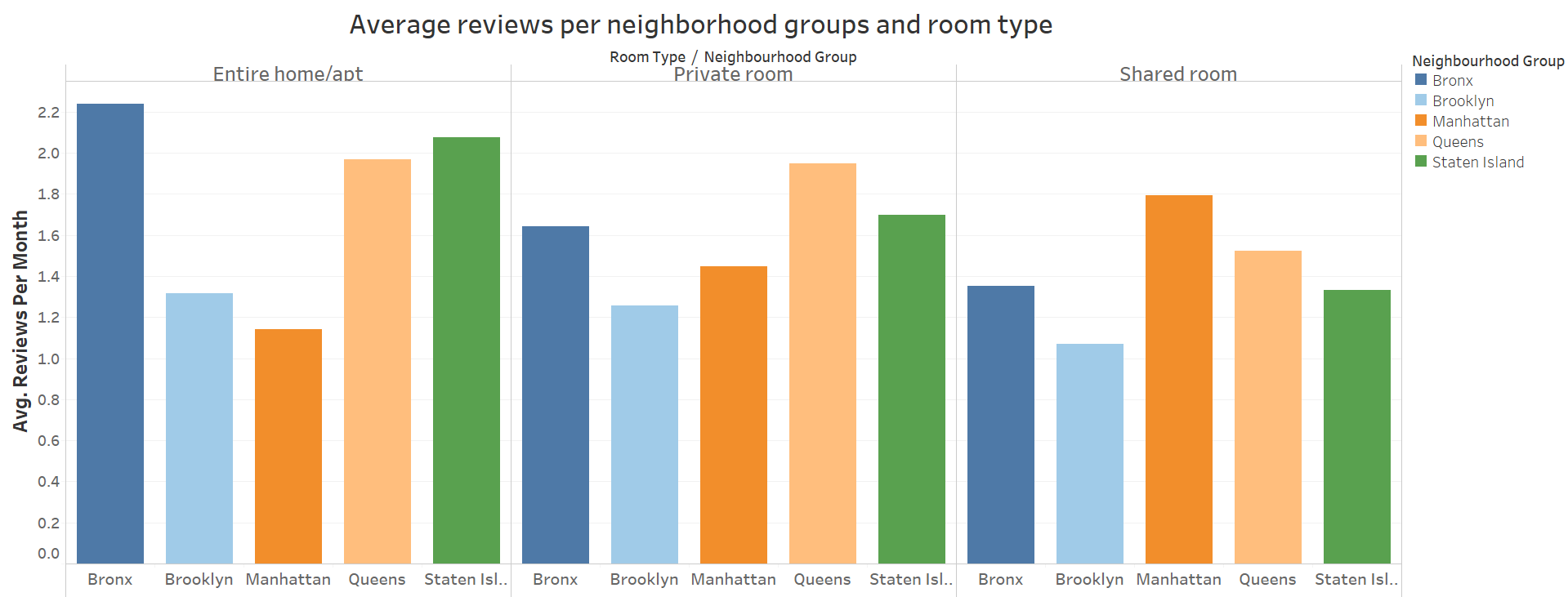
Graph below shows an interesting association between price of the room and its availability throughout the year. It seems that rooms which are %100 available along the year have the highest prices. This case is obvious in the longest spike we see in bar graphs below in Manhattan, Staten island, Brooklyn and Queens.

\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

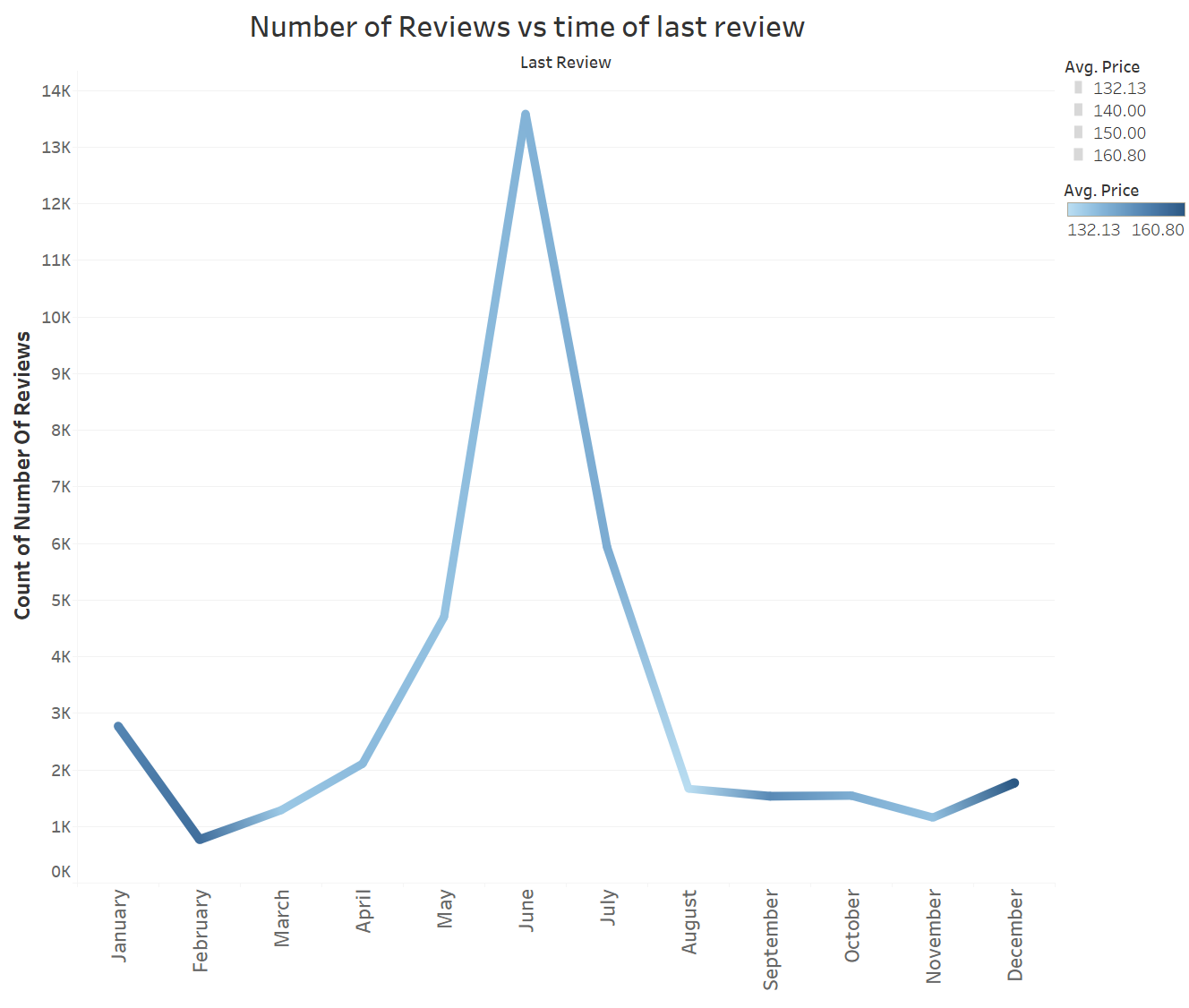
1. **Overview of Reviews**

Graphs below might be misleading in a way. We can’t judge here that Manhattan for example has the lowest reviews for entire home – room type wise, since Manhattan already has the highest number of listings and with that comes number of reviews weather good or bad.

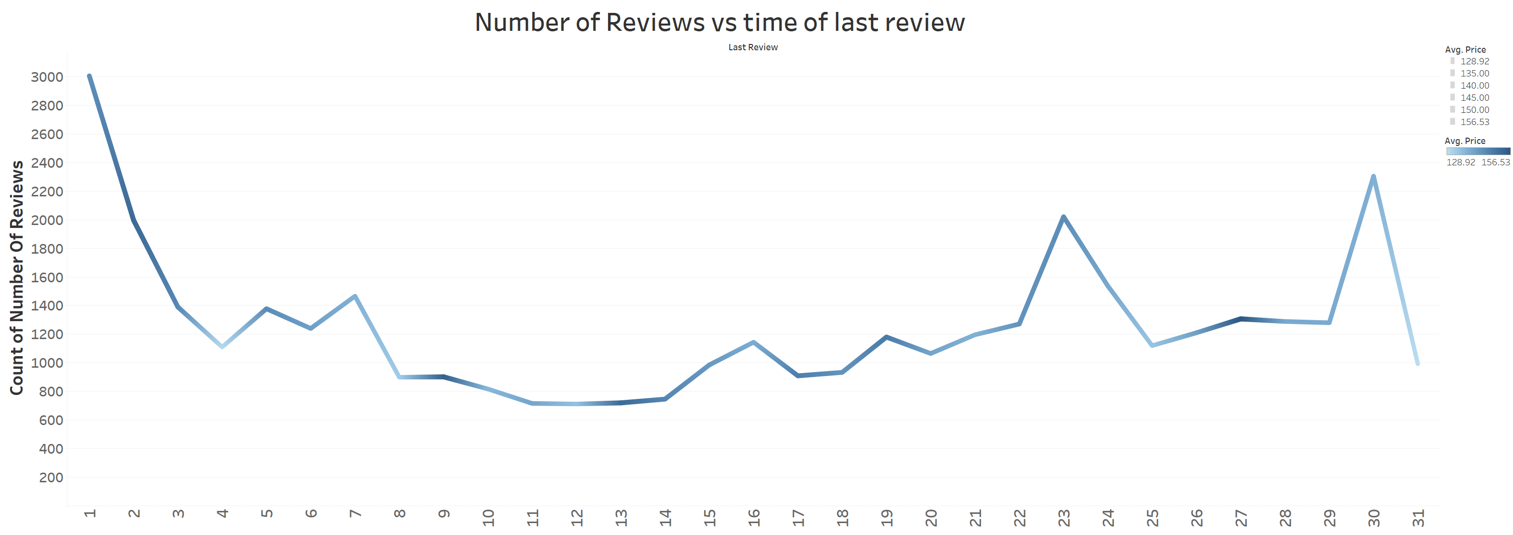


Graph below can be an indicator of the hot season for Airbnb in NY (June). We can’t use price to make conclusions of trends in prices here, since there is no proof that prices stayed the same over the years. But in general, we can say that prices seem to be highest in the beginning and end of the year (holidays season). As NY is already second top destination for tourists in New Year's Eve, this is when prices go the highest.

This graph below can be used by Airbnb to check on users tend to write reviews the most. This will them to draw some strategies to encourage users to write reviews.



Graph below shows that number of reviews on average, tend to be the highest in the first days of the month.



End of the summary of work

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Summary of Takeaways**

This class has helped me to develop critical skills for judging graphs and visualizations, based on theories and organized rules of building a graph. Although, I noticed it’s feeding my perfectionist side and causing delays in delivering the work, but I’m aware that this can be overcome by practice. Also, it taught me that a good graph is not just a graph that is beautiful and has harmonic colors. A good graph is a graph that delivers a message easily and effectively with the least clutter, taking into consideration the audience we’re delivering the message to. Yet, beautiful with harmonic colors that doesn’t disturb the message.

Project wise, working with this dataset taught me that, understanding your variables and making sure there is a connection between them is important before doing anything. Drafting an expected and projected story line before starting is an essential step, so it gives you direction.

# References

<http://www.columbia.edu/~sg3637/airbnb_final_analysis.html>