An Analytical Investigation into Airbnb NYC

Group 7: Sindhu Swaroop, Niraj Sai Prasad, Reema Yadav, Aditya Tilak

12/12/2021

```
library(magrittr)
library(tidyr)
library(date)
library(maps)
library(mapproj)
library(ggmap)
library(leaflet)
library(skimr)
library(DataExplorer)
library(plotly)
library(IRdisplay)
library(knitr)
library(ggplot2)
library(plotrix)
library(dplyr)
library(tidytext)
library(wordcloud2)
library(webshot)
library(htmlwidgets)
```

```
set_theme <- theme(panel.grid.major = element_blank(),
         panel.grid.minor = element_blank(),
         panel.border = element_blank(),
         plot.background = element_rect(fill = "white"),
         panel.background = element_rect(fill = "white"),
         legend.key = element_rect(fill = "white"),
         axis.line = element_line(colour = "black"),
         plot.title = element_text(hjust = .5, face="bold"),
         legend.text = element_text(face="bold"),
         legend.title = element_text(face="bold"),
         axis.text.x = element_text(face="bold"),
         axis.text.y = element_text(face="bold"),
         axis.title.x = element_text(face="bold"),
         axis.title.y = element_text(face="bold"))
```

Introduction

In this report, we will be investigating Airbnb data from New York City. Acquired from Kaggle, this data provides us with details such as price of the listing, which neighborhood it's in, and number of reviews it has.

We will be analyzing relationships between the different variables found in the dataset to help understand which neighborhoods may be cheapest to stay in as well as what words people use to search for Airbnbs.

```
df <- read.csv("AB_NYC_2019.csv", na.strings = "")
```

Data Wrangling

1. Discovering

Prior to beginning our analysis of the data, it is important that we have an understanding of the dataset. Below we are importing the data and then looking at the average price of each neighborhood. As expected, Manhattan is the most expensive of all and has the most number of listings, followed by Brooklyn. Staten Island has the lowest number of listings.

```
kable(df %>%
    group_by(neighbourhood_group) %>%
    summarise(total=n(), avg_price= mean(price))
    ,caption="Average Price of each Neighborhood")
```

Table 1: Average Price of each Neighborhood

neighbourhood_group	total	avg_price
Bronx	1091	87.49679
Brooklyn	20104	124.38321
Manhattan	21661	196.87581
Queens	5666	99.51765
Staten Island	373	114.81233

We also found the total number of records in the listing and the total unique records in the listing, which are equal. This means each listing is unique. As per this data set, there are 48895 Airbnbs in NYC.

```
kable(df %>%
        summarise(total_records=n())
      ,caption="Total Records in Dataset")
```

Table 2: Total Records in Dataset

total_records
48895

```
kable(df %>%
  distinct(id) %>% summarise(total_records=n())
  ,caption="Total Unique Apartments Listed")
```

Table 3: Total Unique Apartments Listed

total_records
48895

2. Cleaning

When exploring the dataset, we saw that the “minimum nights” column had values upto 1200, which is 4 years. This is an unusual amount of time to stay in an Airbnb, so we removed the rows with minimum number of nights greater than 365 days. The resulting dataset now has 48881 listings.

```
df <- df %>% filter(minimum_nights<=365)

kable(df %>%
        summarise(total_records=n())
      ,caption="Total Records in Dataset")
```

Table 4: Total Records in Dataset

total_records
48881

Some of the listings in the dataset had prices of zero. We cleaned the data of such listings and we now have a dataset of 48870 Airbnbs to work with.

```
df <- df %>% filter(price!=0)

kable(df %>%
        summarise(total_records=n())
      ,caption="Total Records in Dataset")
```

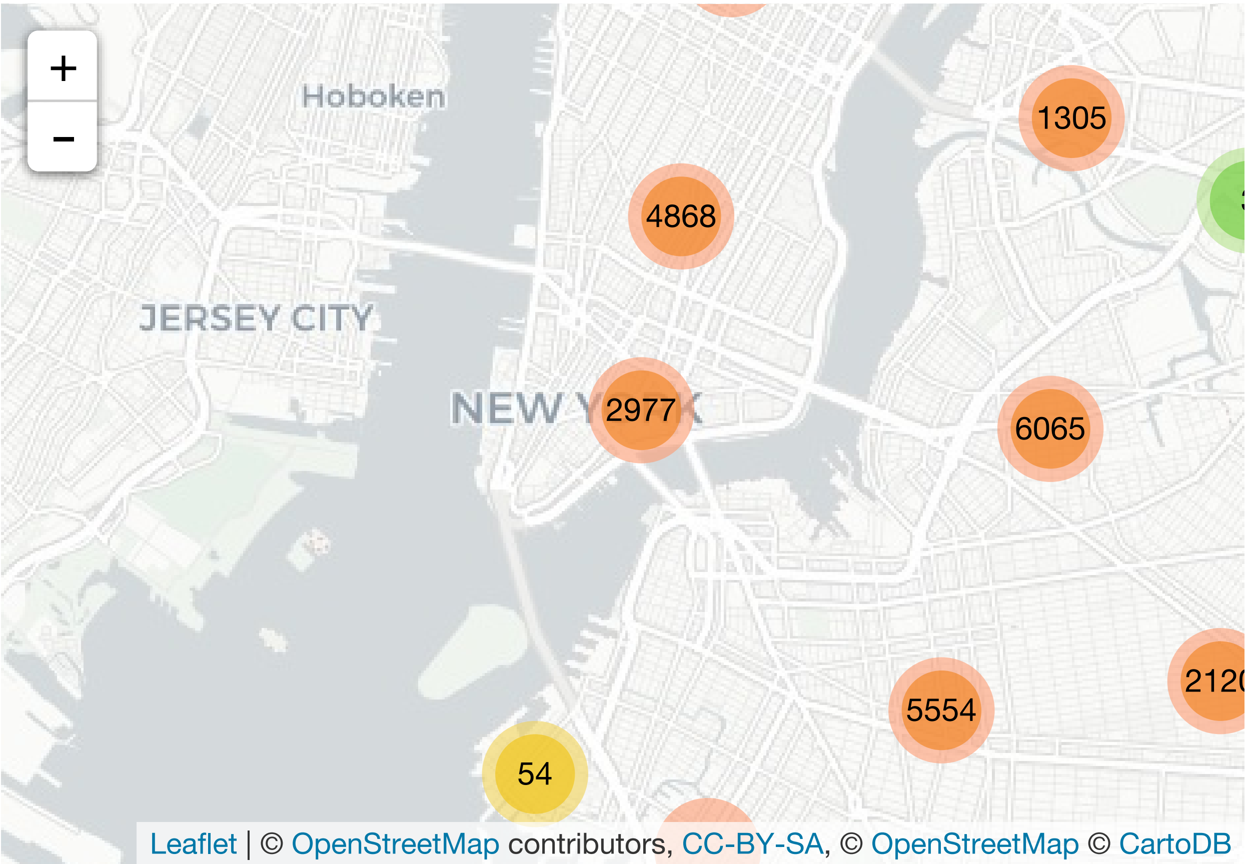
Table 5: Total Records in Dataset

total_records
48870

Business Question 1

When examining the dataset, we came across almost 49000 listings. Visualizing these listings on a map would be much easier for customers and Airbnb hosts alike. Plot Airbnbs by city/latitude/longitude to determine where Airbnbs are concentrated in New York City.

```
leaflet(df) %>%
  addTiles() %>%
  addMarkers(~longitude, ~latitude,labelOptions = labelOptions(noHide = F),clusterOptions = markerCluste
  setView(-74.00, 40.71, zoom = 12) %>%
  addProviderTiles("CartoDB.Positron")
```


+
−
Hoboken
JERSEY CITY
NEW YORK
1305
4868
2977
6065
5554
2120
54
Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap © CartoDB

Observations and Conclusions

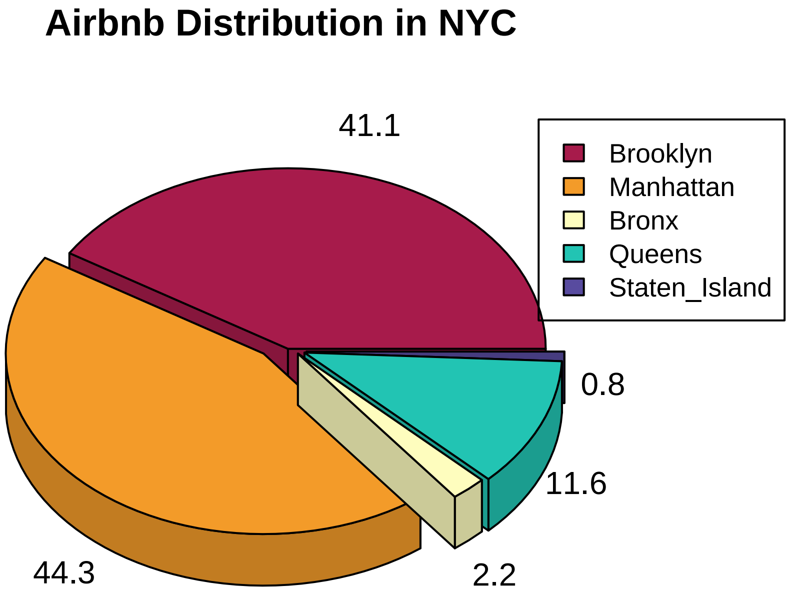
- New York City has a total of 48870 Airbnb listings as per our dataset.
- Using the above interactive map, we can see the location of each Airbnb in the city.

Business Question 2

New York City comprises of 5 major neighborhoods. It would be beneficial for travellers to know how the Airbnbs are spread out across the city. Plot a pie chart to visualize the distribution of Airbnbs among different areas by percentage.

```
Brooklyn = df %>% filter(neighbourhood_group == "Brooklyn") %>% summarize(n())
Manhattan = df %>% filter(neighbourhood_group == "Manhattan") %>% summarize(n())
Bronx = df %>% filter(neighbourhood_group == "Bronx") %>% summarize(n())
Queens = df %>% filter( neighbourhood_group == "Queens") %>% summarize(n())
Staten_Island = df %>% filter( neighbourhood_group == "Staten Island") %>% summarize(n())

Brooklyn = as.integer(Brooklyn)
Manhattan = as.integer(Manhattan)
Bronx = as.integer(Bronx)
Queens = as.integer(Queens)
Staten_Island = as.integer(Staten_Island)
```

```
count = c(Brooklyn,Manhattan,Bronx,Queens,Staten_Island)
labels_area = c('Brooklyn','Manhattan','Bronx','Queens','Staten_Island')
piepercent <- round(100*count/sum(count), 1)

fig = pie3D(count, labels = piepercent, explode=0.1, main = "Airbnb Distribution in NYC", col = hcl.colo
par(xpd=TRUE)
legend(1, 1, labels_area, cex = 0.8, fill = hcl.colors(length(count), "Spectral"))
```



Observations and Conclusions

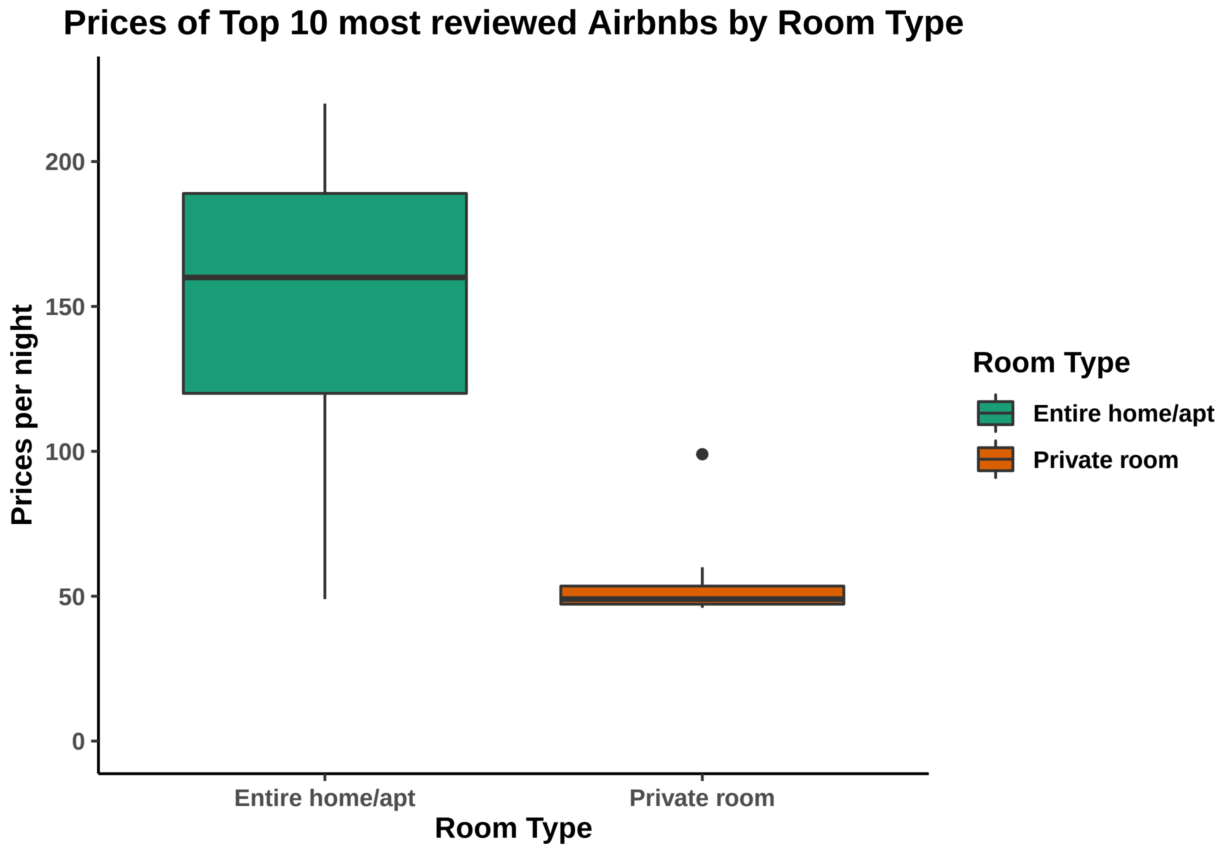
- 44.3% of all listings are located in Manhattan and 41.1% Airbnbs are situated in Brooklyn. Over 85% of Airbnbs are stationed in Manhattan and Brooklyn combined. So if a customer were to look for a place to stay in NYC, his/her best bet would be Manhattan or Brooklyn.
- Staten Island has the least number of Airbnbs (0.8%) i.e 373, in New York City.
- Queens, Bronx and Staten Island house less than 15% of listings. This is probably due to these neighborhoods not being well-connected to Manhattan, as compared to Brooklyn from where Manhattan is easily accessible through public transit.

Business Question 3

Different room types should be priced differently, but it would be advantageous to know how divergent the prices are. For the top 10 Airbnbs with most reviews in the dataset, what is the median, minimum and maximum price per night for each of the room types? Plot a boxplot for visualization.

```
top10private <- df %>%
  subset(room_type == "Private room") %>%
  arrange(desc(number_of_reviews)) %>%
  slice(1:10) %>%
  select(name, neighbourhood_group, neighbourhood, price, room_type)
```

```
top10home <- df %>%
  subset(room_type == "Entire home/apt") %>%
  arrange(desc(number_of_reviews)) %>%
  slice(1:10) %>%
  select(name, neighbourhood_group, neighbourhood, price, room_type)

top10 <- rbind(top10private, top10home)

ggplot(top10, aes(room_type, price))+
  geom_boxplot(aes(fill=room_type)) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  ylim(0,225)+
  labs(title="Prices of Top 10 most reviewed Airbnbs by Room Type",
       x="Room Type",
       y="Prices per night", fill="Room Type")+
  theme(axis.text.x = element_text(angle=0))+
  scale_fill_brewer(palette = "Dark2")+
  set_theme
```



Observations and Conclusions

We plotted a boxplot of Room Type vs. Prices per night, for the top 10 most reviewed listings in the dataset to visualize the median, maximum and minimum prices of each room type.

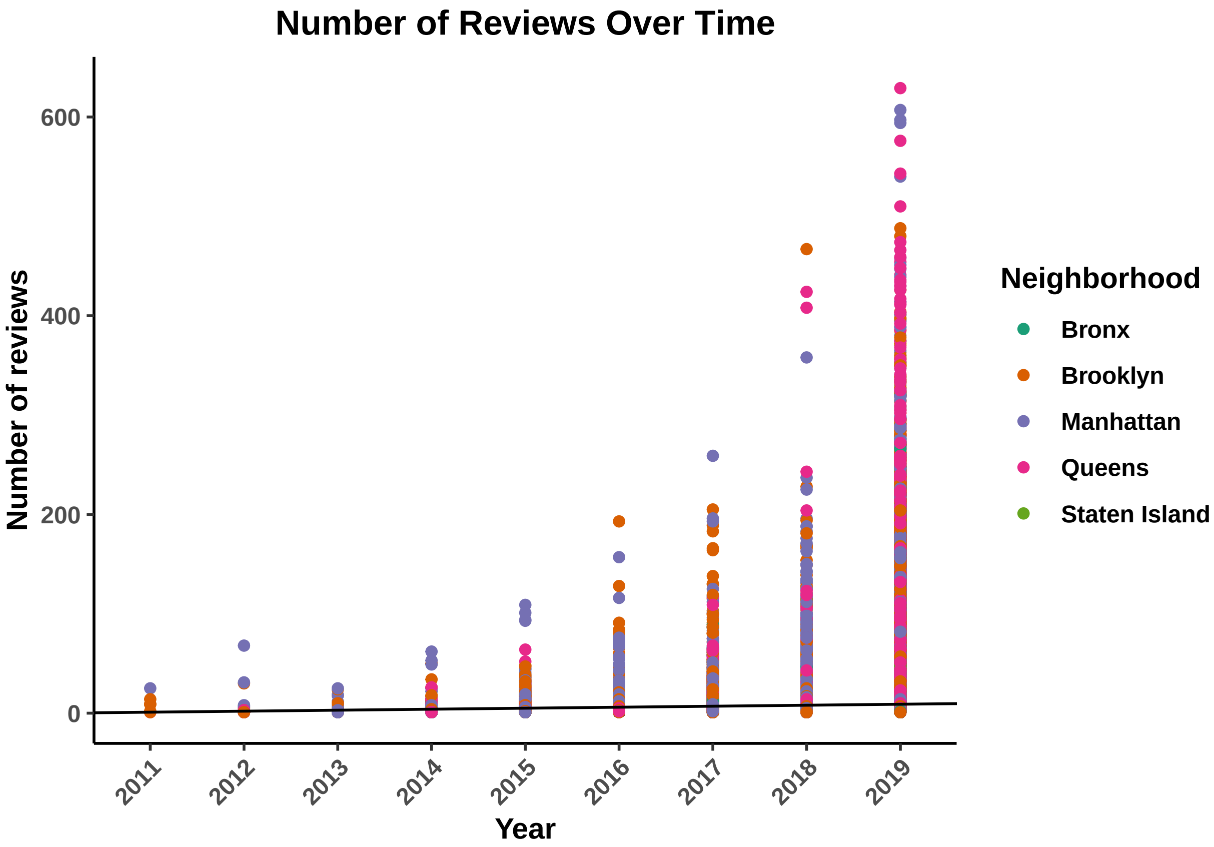
- The Room Type - Entire home/apt has a median price of 160, a minimum price of 49 and a maximum

price of 575 which is an outlier (not shown in graph). The outlier is “TriBeCa 2500 Sq Ft w/ Priv Elevator” and the high price is probably due to the huge area and the presence of a private elevator. The variability of prices of Entire home/apt is high.
- The Room Type - Private Room has a median price of 49, a minimum price of 46 and a maximum price of 99 which is an outlier. The variability of prices of Private rooms is extremely low.
- The minimum price of an entire home is equal to the median price of a private room.

Business Question 4

Airbnb was founded in 2008, and grew popular in 2011. As of today, it is one of the most preferred lodging rentals all over the world. In light of this, what is the trend of number of reviews over the years 2011 to 2019? Plot a scatterplot to envision this.

```r
df$last_review <- as.Date(df$last_review)

df$year <- format(as.Date(df$last_review, format="%d/%m/%Y"),"%Y")

df <- subset(df, year != 'NA')

ggplot(df, aes(x = year, y = number_of_reviews, color=neighbourhood_group))+
  geom_point()+
  geom_abline()+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(title="Number of Reviews Over Time", x="Year", y="Number of reviews", color = "Neighborhood") +
  scale_color_brewer(palette = "Dark2")+
  set_theme
```



Observations and Conclusions

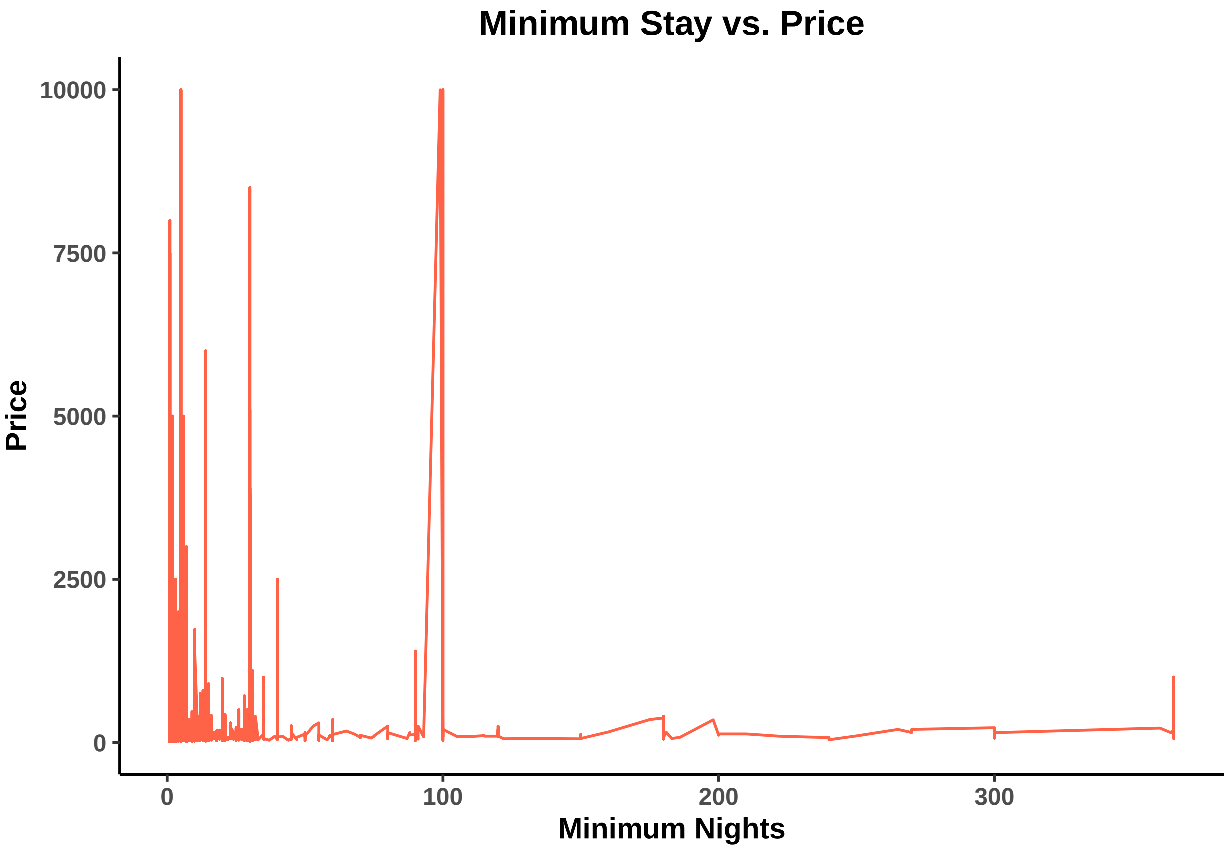
From the plot above, it is clear that the number of reviews over the years follows a positive trend. This indicates that more and more people have started reviewing their experiences at Airbnb.

- The number of reviews increased from 2011 to 2019 drastically. 2019 had the most number of reviews and 2011, the least.
- Staten Island and Bronx listings barely have any reviews owing to low number of Airbnbs in these two regions for NYC.
- In 2019, most number of reviews were gathered from the Queens region. 2018 saw highest number of reviews for Manhattan listings and in 2017, Brooklyn Airbnbs were most reviewed.
- Until 2016, number of reviews for a listing never crossed 200. On the contrary, maximum reviews gained in 2019 was over 600.

Business Question 5

Each Airbnb listing has a minimum length of stay, which requires customers to book the room for a minimum number of nights. On these lines, is there a relationship between price per night and minimum length of stay? Plot a line graph to visualize the trend.

```
ggplot(df, aes(x=minimum_nights, y=price)) +
  geom_line(color="tomato1")+
  labs(x="Minimum Nights" , y="Price", title = "Minimum Stay vs. Price")+
  set_theme
```



Observations and Conclusions

From the above line graph, we can see a decreasing trend for minimum number of nights vs. price.

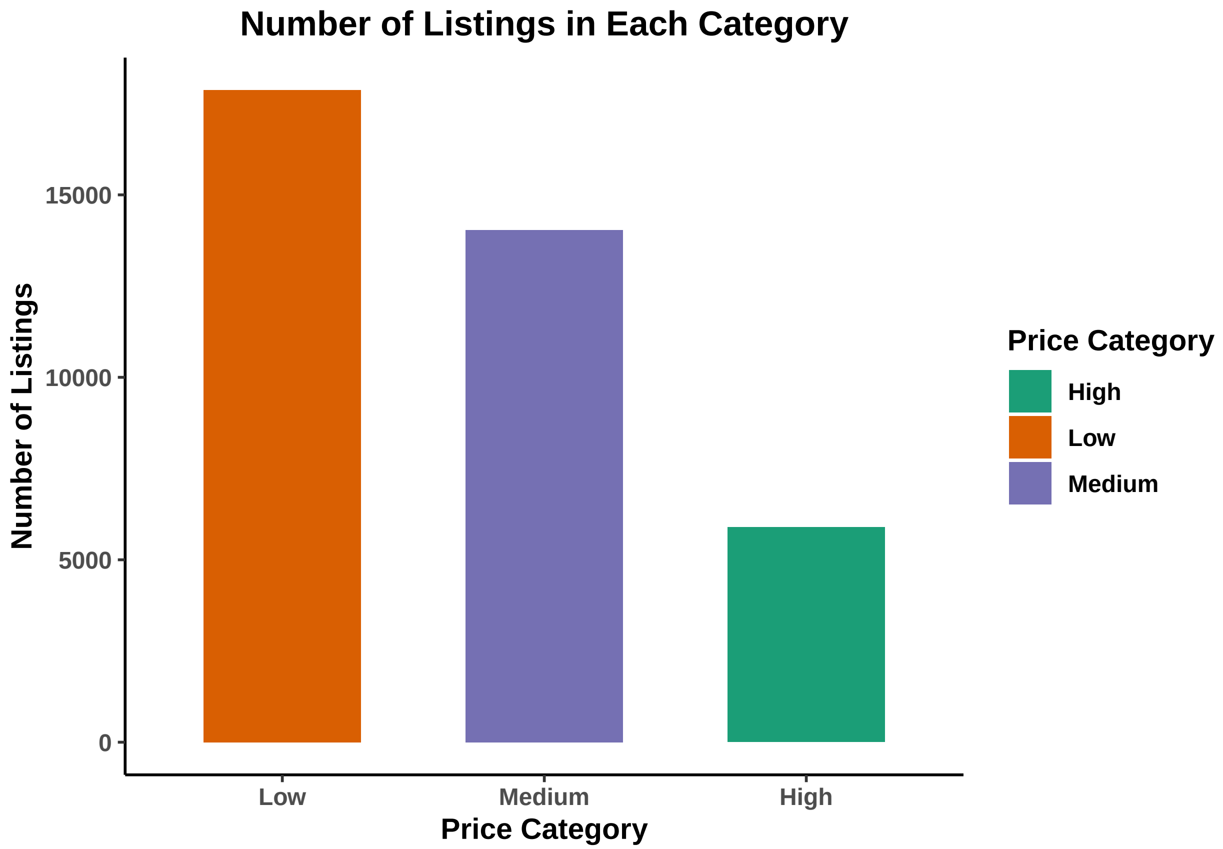
- The price is usually higher for listings with fewer minimum nights of stay.
- Most listings have minimum nights less than 100 days, which is convenient for tourists.
- As the minimum stay increases from ~120 days to ~365 days, the price reduces. This works in favor of those customers who have longer stays planned at NYC.

Business Question 6

Pricing of Airbnb listings varies from 10$ to 10000$. To enhance the usability of this data by owners, pricing information should be more accurately depicted. This way, they will know what price range is most common and how much to price their next listings. Divide the prices into three categories and plot a bar chart to visualize the number of listings in each price range.

```
df <- df %>%
  mutate(rent_category =
           case_when(
             price>=0 & price<100 ~ "Low",
             price>=100 & price<200 ~ "Medium",
             price>200 ~ "High")
```

```
         )
df %>%
  group_by(rent_category) %>%
  summarise(total= n()) %>%
  drop_na(rent_category) %>%
  ggplot(aes(x = reorder(`rent_category`, -total),  y = total, fill = rent_category)) +
  geom_bar(stat = "identity", width = 0.6)+
  labs(x="Price Category" , y="Number of Listings", title = "Number of Listings in Each Category", fill
  scale_fill_brewer(palette = "Dark2")+
  set_theme
```



Observations and Conclusions

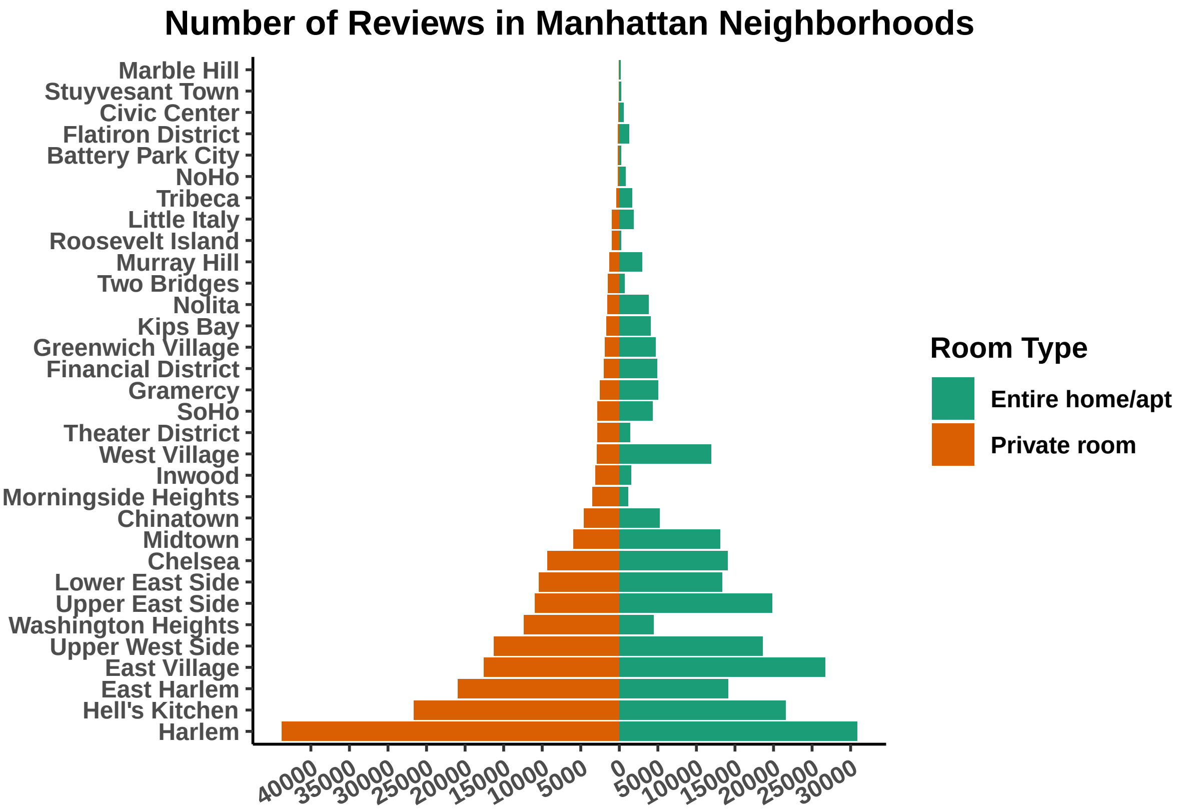
We divided the price range into 3 categories - Low (Rent<100), Medium (100<Rent<200) & High (Rent>200).

- From the graph above, we can see that 22000 listings have a price of less than 100$. Most listings belong to the Low price category. Owners would profit more by pricing their listings in the lower range as it attracts more customers, which is probably why the Low price category has so many listings.
- 18000 listings fall in the Medium price range, with a price range of 100 - 200$.
- Only a few Airbnbs (8000) cost more than 200$.

Business Question 7

Airbnb hosts can list anything from a private room to the entire home/apartment. Do travelers gravitate towards a specific room type? Plot a funnel chart for number of reviews in different neighborhoods in NYC's most popular town - Manhattan.

```
dfp <- df %>%
  subset(room_type == "Private room" & neighbourhood_group == "Manhattan") %>%
  mutate(number_of_reviews=-1*number_of_reviews) %>%
  group_by(neighbourhood, room_type) %>%
  summarize(num=sum(number_of_reviews))

dfh <- df %>%
  subset(room_type == "Entire home/apt" & neighbourhood_group == "Manhattan") %>%
  group_by(neighbourhood, room_type) %>%
  summarize(num=sum(number_of_reviews))

dff <- rbind(dfp, dfh)

dff$neighbourhood <- factor(dff$neighbourhood, levels=c("Harlem", "Hell's Kitchen", "East Harlem", "Eas

brks <- seq(-40000, 40000, 5000)
lbls <- as.character(c(seq(40000, 0, -5000), seq(5000, 40000, 5000)))

ggplot(dff, aes(x = neighbourhood, y = num, fill = room_type)) +
  geom_bar(stat = "identity", width = 0.9) +
  scale_y_continuous(breaks = brks, labels=lbls) +
  coord_flip() +
  labs(title="Number of Reviews in Manhattan Neighborhoods", x="", y="", fill="Room Type") +
  theme(axis.text.x = element_text(hjust=1, angle=30)) +
  scale_fill_brewer(palette = "Dark2") +
  set_theme
```



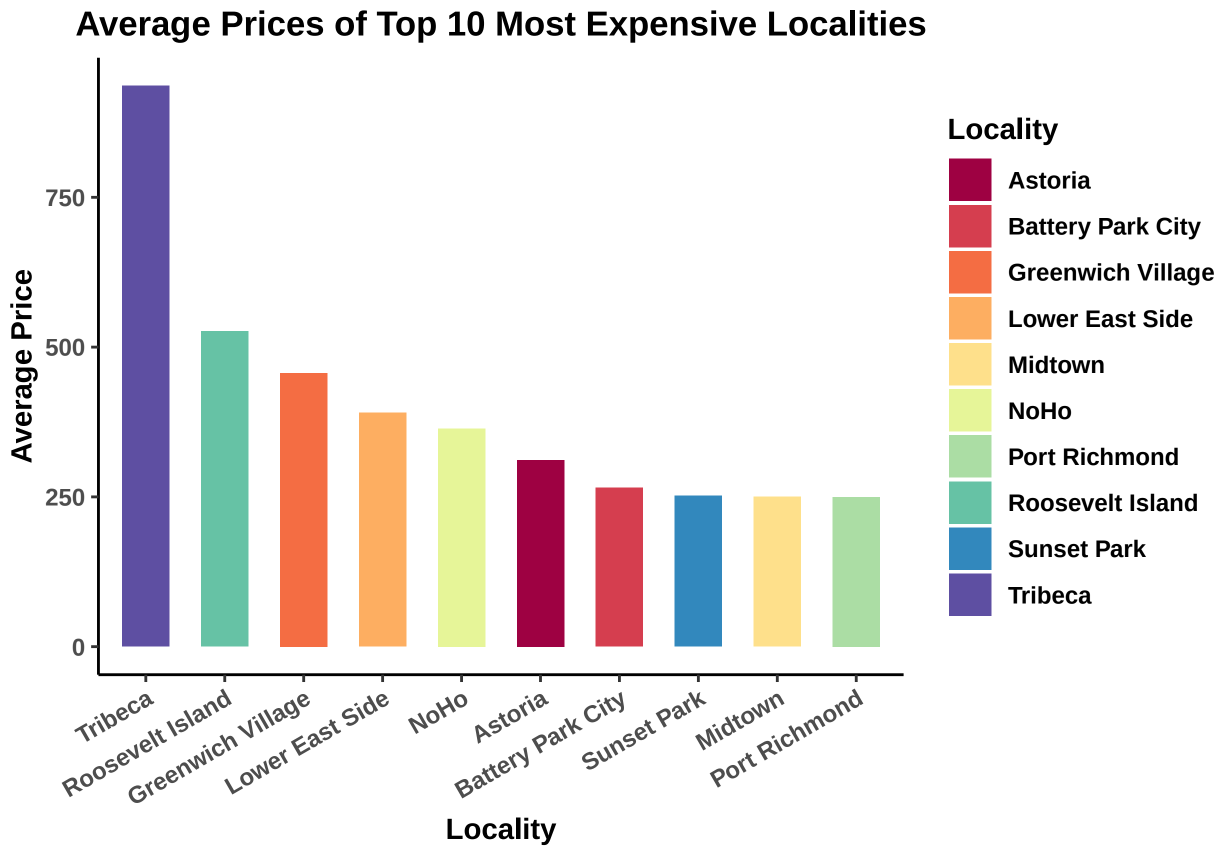
Observations and Conclusions

- Manhattan, the biggest town in NYC has 32 neighborhoods, the most popular in the Airbnb aspect being Harlem, and the least popular being Marble Hill.
- In most neighborhoods, we can see from the graph that travellers gravitate towards entire homes/apartments rather than renting out private rooms. This was an interesting result as at first glance, we would think private rooms, being the cheaper option, would be the go-to choice of travellers. From the Rentals policy in NYC, it should be noted that unhosted rentals for less than 30 days in buildings with multiple dwellings are illegal in the city. So families will rent out entire apartments and even individual travellers would prefer privacy rather than living with the host, and tend to rent out smaller apartments.
- However, there are a couple of neighborhoods like Harlem, Hell's Kitchen and Washington Heights which have more number of reviews for private rooms than entire homes. These listings are probably either >30-day listings or not in multiple-dwelling units (or they might be illegal listings!).

Business Question 8

Looking at New York City as a whole with all the neighborhoods included, what are the top 10 localities with the most expensive listings? Plot a bar graph to visualize this.

```
df_monthly <- df %>% filter(minimum_nights>=30)

df_monthly %>%
  group_by(neighbourhood) %>%
  summarise(avg_price=mean(price)) %>%
  arrange(desc(avg_price)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(`neighbourhood`, -avg_price), y = avg_price, fill = neighbourhood)) +
  geom_bar(stat = "identity", width = 0.6)+
  labs(x="Locality" , y="Average Price", fill = "Locality", title = "Average Prices of Top 10 Most Expe
  scale_fill_brewer(palette = "Spectral")+
  theme(axis.text.x = element_text(hjust=1, angle=30))+
  set_theme
```



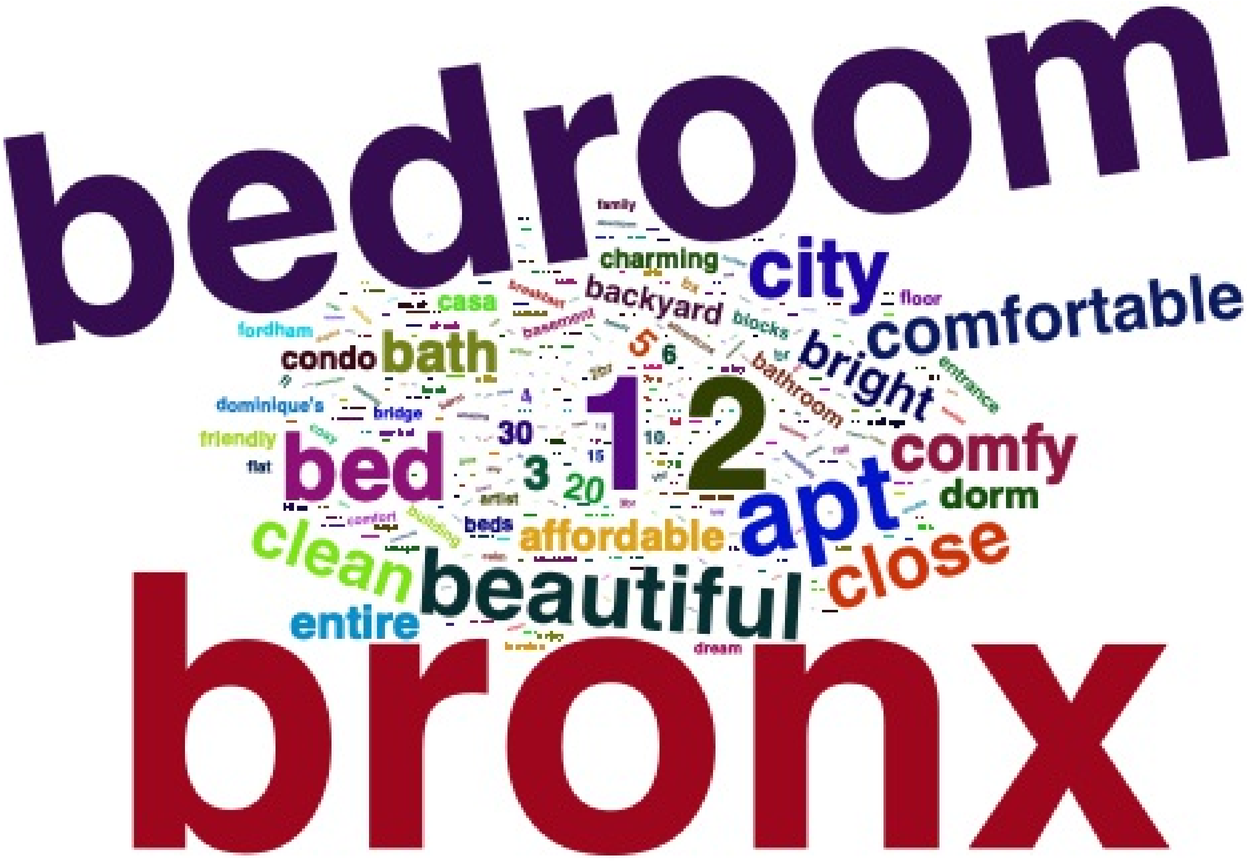
Observations and Conclusions

- From the above graph, we can see that 7 of the above 10 most expensive localities belong to the Manhattan neighborhood group.
- Brooklyn is home to 2 of the 10 most pricy listings and one of the listings is located in Queens.
- Bay Ridge and Forest Hills have an almost equal average rent, as do West Village and Astoria.
- Tribeca has the highest average rent, of above 600$.
- Travellers planning budget trips would do good to stay clear of these localities!

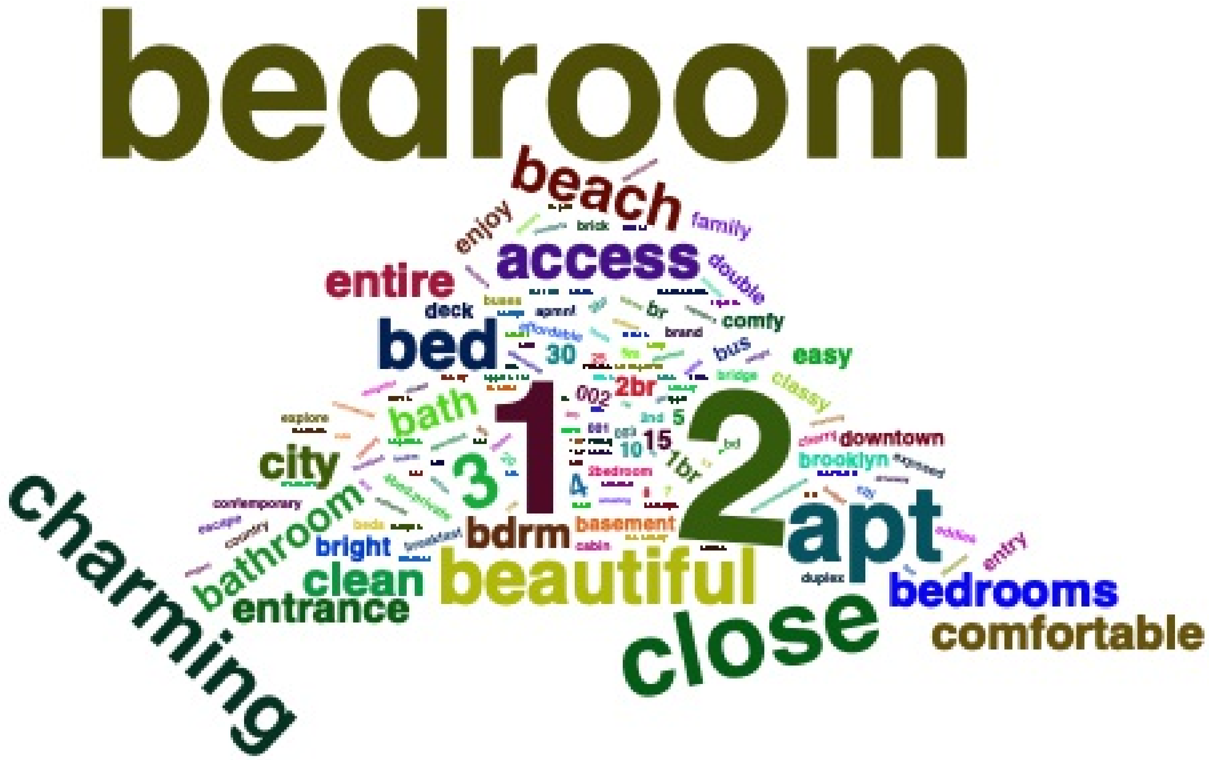
Business Question 9

Our dataset consists of unique names for each listing. We drew a Word Cloud in order to visualize the words that appear most frequently in listing names of a particular neighborhood group. These words could be useful to people in searching for an Airbnb on search engines.

```
data(stop_words)

# df_Manhattan <- df %>%
#   subset(neighbourhood_group == "Manhattan") %>%
#   unnest_tokens(word, name) %>%
#   anti_join(stop_words) %>%
#   count(word)
# wordcloud2(df_Manhattan, shape = "circle", shuffle=FALSE)
#
# df_Brooklyn <- df %>%
#   subset(neighbourhood_group == "Brooklyn") %>%
#   unnest_tokens(word, name) %>%
#   anti_join(stop_words) %>%
#   count(word)
# wordcloud2(df_Brooklyn, shape = "triangle", shuffle=FALSE)
#
# df_Queens <- df %>%
#   subset(neighbourhood_group == "Queens") %>%
#   unnest_tokens(word, name) %>%
#   anti_join(stop_words) %>%
#   count(word)
# wordcloud2(df_Queens, shape = "triangle-forward", shuffle=FALSE)

df_Bronx <- df %>%
  subset(neighbourhood_group == "Bronx") %>%
  unnest_tokens(word, name) %>%
  anti_join(stop_words) %>%
  count(word)
wordcloud2(df_Bronx, shape = "circle", shuffle=FALSE)
```



```
df_Staten <- df %>%
  subset(neighbourhood_group == "Staten Island") %>%
  unnest_tokens(word, name) %>%
  anti_join(stop_words) %>%
  count(word)
wordcloud2(df_Staten, shape = "triangle", shuffle=FALSE)
```



Observations and Conclusions

- The most common words unique to Manhattan listings are - midtown, central, park, harlem, times, square.
- If looking for an Airbnb in Brooklyn, people could type in the words - brownstone, bushwick, greenpoint.
- Queens listings mostly contain these words - queens, astoria, jfk.
- With the Yankees stadium located in Bronx, the most frequent words in Bronx listings are - bronx, yankee, stadium.
- Staten Island is famous for its ferry and its proximity to the beach. Tourists could look for listings using the words - staten, island, ferry, beach.
- Words which are repeated in all of NYC's Airbnb listings are - sunny, modern, cozy, bright, beautiful, apartment, bedroom, charming, loft, studio, subway, nyc, quiet, clean.
- Interestingly, another word which appears in most listings of all neighborhoods is - Manhattan, in the context of "close to Manhattan". Airbnbs attract customers by listing the description of a room as being close to Manhattan - the most happening place in NYC.