New York City Airbnb

Sara Kmair

10/8/2019

## R Markdown

#libraries  
library(corrplot)

## Warning: package 'corrplot' was built under R version 3.6.1

## corrplot 0.84 loaded

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.1

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.6.1

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidytext)

## Warning: package 'tidytext' was built under R version 3.6.1

library(tm) # for text mining

## Warning: package 'tm' was built under R version 3.6.1

## Loading required package: NLP

##   
## Attaching package: 'NLP'

## The following object is masked from 'package:ggplot2':  
##   
## annotate

library(SnowballC)  
library(wordcloud)

## Warning: package 'wordcloud' was built under R version 3.6.1

## Loading required package: RColorBrewer

library(RColorBrewer)  
library(bigmemory)

## Warning: package 'bigmemory' was built under R version 3.6.1

#checking the data and data structure  
airbnb <- read.csv("airbnb.csv")  
str(airbnb)

## 'data.frame': 48895 obs. of 16 variables:  
## $ id : int 2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...  
## $ name : Factor w/ 47897 levels "","'Fan'tastic",..: 12652 38163 45162 15693 19357 24992 8328 25039 15588 17673 ...  
## $ host\_id : int 2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...  
## $ host\_name : Factor w/ 11453 levels "","'Cil","#NAME?",..: 5051 4846 2962 6264 5982 1970 3601 9699 6935 1264 ...  
## $ neighbourhood\_group : Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...  
## $ neighbourhood : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42 62 138 14 96 203 36 ...  
## $ latitude : num 40.6 40.8 40.8 40.7 40.8 ...  
## $ longitude : num -74 -74 -73.9 -74 -73.9 ...  
## $ room\_type : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...  
## $ price : int 149 225 150 89 80 200 60 79 79 150 ...  
## $ minimum\_nights : int 1 1 3 1 10 3 45 2 2 1 ...  
## $ number\_of\_reviews : int 9 45 0 270 9 74 49 430 118 160 ...  
## $ last\_review : Factor w/ 1765 levels "","1/1/2013",..: 203 1059 1 1438 348 1234 277 1244 1383 1317 ...  
## $ reviews\_per\_month : num 0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...  
## $ calculated\_host\_listings\_count: int 6 2 1 1 1 1 1 1 1 4 ...  
## $ availability\_365 : int 365 355 365 194 0 129 0 220 0 188 ...

#changing id column from integer to char  
airbnb$host\_id <- as.character(airbnb$host\_id)  
#Delete id column and host\_name as it has no influence in our analysis   
airbnb <- subset(airbnb, select = -c(id, host\_name, availability\_365))  
#changing the name attribute to desc as it gives a small description about the listing  
colnames(airbnb)[which(names(airbnb) == "name")] <- "description"  
colnames(airbnb)[which(names(airbnb) == "calculated\_host\_listings\_count")] <- "listing\_cnt"  
str(airbnb)

## 'data.frame': 48895 obs. of 13 variables:  
## $ description : Factor w/ 47897 levels "","'Fan'tastic",..: 12652 38163 45162 15693 19357 24992 8328 25039 15588 17673 ...  
## $ host\_id : chr "2787" "2845" "4632" "4869" ...  
## $ neighbourhood\_group: Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...  
## $ neighbourhood : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42 62 138 14 96 203 36 ...  
## $ latitude : num 40.6 40.8 40.8 40.7 40.8 ...  
## $ longitude : num -74 -74 -73.9 -74 -73.9 ...  
## $ room\_type : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...  
## $ price : int 149 225 150 89 80 200 60 79 79 150 ...  
## $ minimum\_nights : int 1 1 3 1 10 3 45 2 2 1 ...  
## $ number\_of\_reviews : int 9 45 0 270 9 74 49 430 118 160 ...  
## $ last\_review : Factor w/ 1765 levels "","1/1/2013",..: 203 1059 1 1438 348 1234 277 1244 1383 1317 ...  
## $ reviews\_per\_month : num 0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...  
## $ listing\_cnt : int 6 2 1 1 1 1 1 1 1 4 ...

levels(airbnb$neighbourhood\_group)

## [1] "Bronx" "Brooklyn" "Manhattan" "Queens"   
## [5] "Staten Island"

#checking how many neighbourhood in each neighbourhood group  
airbnb %>%  
 group\_by(neighbourhood\_group) %>%  
 summarise(neighbourhood\_number = n\_distinct(neighbourhood))

## # A tibble: 5 x 2  
## neighbourhood\_group neighbourhood\_number  
## <fct> <int>  
## 1 Bronx 48  
## 2 Brooklyn 47  
## 3 Manhattan 32  
## 4 Queens 51  
## 5 Staten Island 43

#checking for NAs in the dataset  
names(airbnb)

## [1] "description" "host\_id" "neighbourhood\_group"  
## [4] "neighbourhood" "latitude" "longitude"   
## [7] "room\_type" "price" "minimum\_nights"   
## [10] "number\_of\_reviews" "last\_review" "reviews\_per\_month"   
## [13] "listing\_cnt"

sum(is.na(airbnb$price))

## [1] 0

sum(is.na(airbnb$minimum\_nights))

## [1] 0

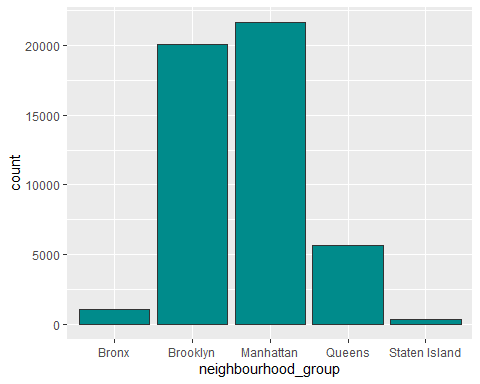
sum(is.na(airbnb$availability\_365))

## [1] 0

sum(is.na(airbnb$calculated\_host\_listings\_count))

## [1] 0

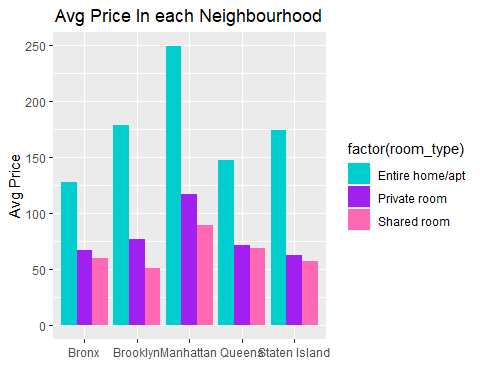
#Manhattan is the most popular neighborhood   
ggplot(airbnb, aes(x =neighbourhood\_group)) +  
 geom\_bar( fill = "cyan4", col = "grey20")



#renting entire home seems the most popular type   
ggplot(airbnb, aes(x =room\_type)) +  
 geom\_bar( fill = "pink", col = "grey20")



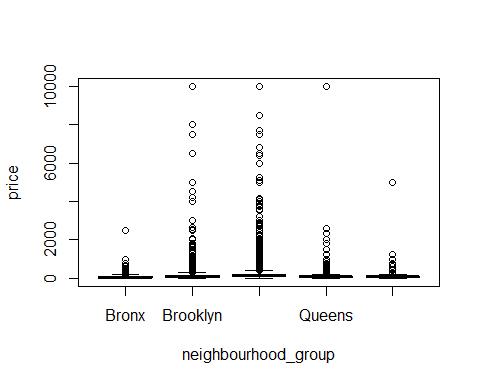
#visualizing Average price in each neighbourhood by room\_type  
airbnb %>%   
 group\_by(room\_type, neighbourhood\_group) %>%  
 summarise(AvgPrice = mean(price)) %>%  
 arrange(desc(AvgPrice)) %>%  
ggplot(., aes(x = neighbourhood\_group, y = AvgPrice, fill = factor(room\_type))) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 labs(x = "", y = "Avg Price") +  
 scale\_fill\_manual(values = c("cyan3", "purple", "hotpink")) +  
 ggtitle("Avg Price In each Neighbourhood") +  
 theme(plot.title = element\_text(hjust = 0.5))



#The average price for each neighborhood group   
airbnb %>%   
 group\_by(neighbourhood\_group) %>%  
 summarise(AvgPrice = mean(price)) %>%  
 arrange(desc(AvgPrice))

## # A tibble: 5 x 2  
## neighbourhood\_group AvgPrice  
## <fct> <dbl>  
## 1 Manhattan 197.   
## 2 Brooklyn 124.   
## 3 Staten Island 115.   
## 4 Queens 99.5  
## 5 Bronx 87.5

#In this step we will study outliers for each neighbourhood and how it is affecting the price   
#subsetting Bronx neighbourhood  
boxplot(price ~ neighbourhood\_group, data = airbnb)



Bronx\_sub <- airbnb[which(airbnb$neighbourhood\_group == "Bronx"), ]  
nrow(Bronx\_sub)

## [1] 1091

summary(Bronx\_sub$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 45.0 65.0 87.5 99.0 2500.0

#removing the ourliers from Bronx price   
out\_Bronx <- boxplot(Bronx\_sub$price, plot = FALSE)$out  
length(out\_Bronx)

## [1] 61

mean(out\_Bronx)

## [1] 366.7377

Bronx\_sub %>%  
 group\_by(room\_type) %>%  
 summarise(mean(price))

## # A tibble: 3 x 2  
## room\_type `mean(price)`  
## <fct> <dbl>  
## 1 Entire home/apt 128.   
## 2 Private room 66.8  
## 3 Shared room 59.8

Bronx\_sub\_no\_out <- Bronx\_sub[- which(Bronx\_sub$price %in% out\_Bronx), ]  
Bronx\_out\_only <- Bronx\_sub[ which(Bronx\_sub$price %in% out\_Bronx), ]  
Bronx\_out\_only %>%  
 group\_by(room\_type) %>%  
 summarise(length(room\_type))

## # A tibble: 3 x 2  
## room\_type `length(room\_type)`  
## <fct> <int>  
## 1 Entire home/apt 48  
## 2 Private room 12  
## 3 Shared room 1

#the outliers are going to affect out analysis since mean is increasing by 17  
#calculating the mean with and without outliers   
mean(Bronx\_sub$price)

## [1] 87.49679

mean(out\_Bronx)

## [1] 366.7377

mean(Bronx\_sub\_no\_out$price)

## [1] 70.95922

#subsetting Brooklyn neighbourhood   
Brooklyn\_sub <- airbnb[which(airbnb$neighbourhood\_group == "Brooklyn"), ]  
summary(Brooklyn\_sub$price)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 60.0 90.0 124.4 150.0 10000.0

nrow(Brooklyn\_sub)

## [1] 20104

#removing the ourliers from Brooklyn price   
out\_Brooklyn <- boxplot(Brooklyn\_sub$price, plot = FALSE)$out  
length(out\_Brooklyn)

## [1] 996

mean(out\_Bronx)

## [1] 366.7377

Brooklyn\_sub %>%  
 group\_by(room\_type) %>%  
 summarise(mean(price))

## # A tibble: 3 x 2  
## room\_type `mean(price)`  
## <fct> <dbl>  
## 1 Entire home/apt 178.   
## 2 Private room 76.5  
## 3 Shared room 50.5

Brooklyn\_sub\_no\_out <- Brooklyn\_sub[- which(Brooklyn\_sub$price %in% out\_Brooklyn), ]  
Bronoklyn\_out\_only <- Brooklyn\_sub[ which(Brooklyn\_sub$price %in% out\_Brooklyn), ]  
Bronoklyn\_out\_only %>%  
 group\_by(room\_type) %>%  
 summarise(length(room\_type))

## # A tibble: 3 x 2  
## room\_type `length(room\_type)`  
## <fct> <int>  
## 1 Entire home/apt 911  
## 2 Private room 82  
## 3 Shared room 3

#Calculating the mean with and without outliers   
mean(Brooklyn\_sub$price)

## [1] 124.3832

mean(out\_Brooklyn)

## [1] 543.9448

mean(Brooklyn\_sub\_no\_out$price)

## [1] 102.5137

#The most popular hosts and the number of times been rented   
top\_host <- airbnb %>%  
 group\_by(host\_id) %>%  
 summarise( mcount = length(host\_id), avgPrice = mean(price)) %>%  
 arrange(desc(mcount)) %>%  
 slice(1:10)  
   
#top ten host   
top\_host

## # A tibble: 10 x 3  
## host\_id mcount avgPrice  
## <chr> <int> <dbl>  
## 1 219517861 327 253.   
## 2 107434423 232 303.   
## 3 30283594 121 278.   
## 4 137358866 103 43.8  
## 5 12243051 96 213.   
## 6 16098958 96 209.   
## 7 61391963 91 146.   
## 8 22541573 87 215.   
## 9 200380610 65 290.   
## 10 1475015 52 103.

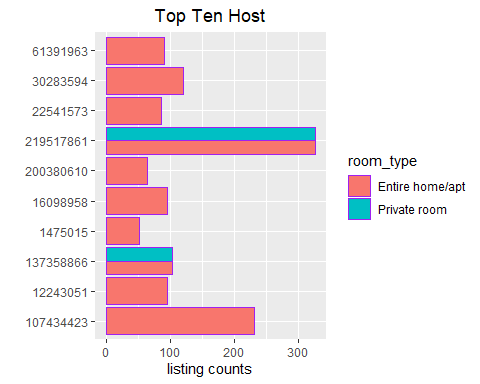
#extracting the listings of the top hosts   
pop\_host <- airbnb[which(airbnb$host\_id == top\_host$host\_id), ]

## Warning in airbnb$host\_id == top\_host$host\_id: longer object length is not  
## a multiple of shorter object length

head(pop\_host)

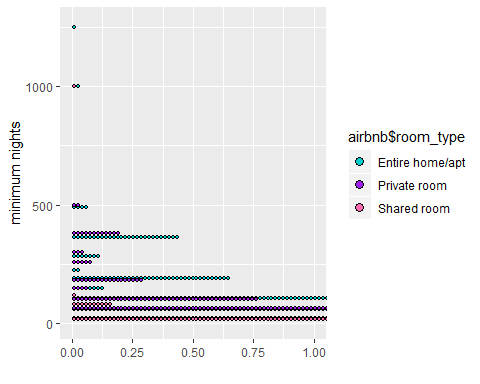
## description host\_id  
## 4450 Perfect 1BR / Doorman / 22 St & 2nd Ave 1475015  
## 4910 Amazing Renovated studio MurrayHill 1475015  
## 5270 Great Renovations. Columbus Circle 1475015  
## 5490 Prime Gramercy, Luxury 1BD DOORMAN 1475015  
## 5740 HUGE 1BD, RENOVATED, TIME SQUARE !! 1475015  
## 11573 Luxury Studio Apartment in NYC'S Chelsea District 30283594  
## neighbourhood\_group neighbourhood latitude longitude  
## 4450 Manhattan Murray Hill 40.74475 -73.97201  
## 4910 Manhattan Kips Bay 40.74318 -73.97858  
## 5270 Manhattan Hell's Kitchen 40.76867 -73.98541  
## 5490 Manhattan Gramercy 40.73771 -73.98058  
## 5740 Manhattan Hell's Kitchen 40.76239 -73.99227  
## 11573 Manhattan Chelsea 40.74490 -73.99167  
## room\_type price minimum\_nights number\_of\_reviews last\_review  
## 4450 Entire home/apt 150 30 2 8/31/2016  
## 4910 Entire home/apt 87 30 4 6/30/2019  
## 5270 Entire home/apt 87 30 6 4/21/2019  
## 5490 Entire home/apt 150 30 1 10/8/2014  
## 5740 Entire home/apt 125 30 0   
## 11573 Entire home/apt 109 30 1 10/31/2016  
## reviews\_per\_month listing\_cnt  
## 4450 0.04 52  
## 4910 0.07 52  
## 5270 0.11 52  
## 5490 0.02 52  
## 5740 NA 52  
## 11573 0.03 121

#visualizing top ten hosts and the number of listing counts for each with what room\_type it is  
#notice from the plot top ten most popular host hasonly entire home or private room that means shared room is not as popular   
ggplot(pop\_host, aes(x = host\_id, y = listing\_cnt, fill = room\_type)) +  
 geom\_bar(stat = "identity", position = "dodge", color = "purple" ) +  
 labs(x = "", y = "listing counts") +  
 coord\_flip() +  
 ggtitle("Top Ten Host") +  
 theme(plot.title = element\_text(hjust = 0.5))

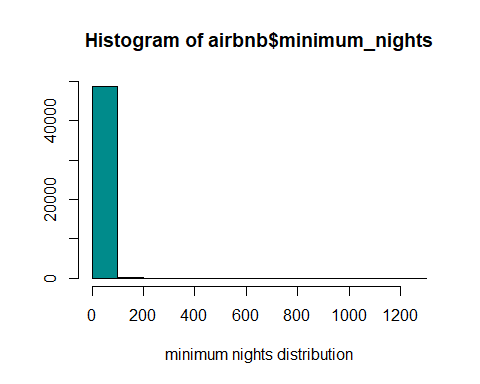


ggplot(airbnb) +  
 geom\_dotplot(dotsize = 0.4, aes(x = airbnb$minimum\_nights,fill = airbnb$room\_type)) +  
 coord\_flip() +  
 labs(y = "", x = "minimum nights") +  
 scale\_fill\_manual(values = c("cyan3", "purple", "hotpink"))

## `stat\_bindot()` using `bins = 30`. Pick better value with `binwidth`.



hist(airbnb$minimum\_nights, xlab = "minimum nights distribution", ylab = "", col = "cyan4", labels = FALSE)



#find the most expensive rent   
max\_price <- max(airbnb$price)  
airbnb[which(airbnb$price == max\_price), ]

## description host\_id  
## 9152 Furnished room in Astoria apartment 20582832  
## 17693 Luxury 1 bedroom apt. -stunning Manhattan views 5143901  
## 29239 1-BR Lincoln Center 72390391  
## neighbourhood\_group neighbourhood latitude longitude  
## 9152 Queens Astoria 40.76810 -73.91651  
## 17693 Brooklyn Greenpoint 40.73260 -73.95739  
## 29239 Manhattan Upper West Side 40.77213 -73.98665  
## room\_type price minimum\_nights number\_of\_reviews last\_review  
## 9152 Private room 10000 100 2 2/13/2016  
## 17693 Entire home/apt 10000 5 5 7/27/2017  
## 29239 Entire home/apt 10000 30 0   
## reviews\_per\_month listing\_cnt  
## 9152 0.04 1  
## 17693 0.16 1  
## 29239 NA 1

#let's study the most popular neighbourhood Manhattan  
Manhattan <- airbnb[airbnb$neighbourhood\_group == "Manhattan", ]  
head(Manhattan)

## description host\_id  
## 2 Skylit Midtown Castle 2845  
## 3 THE VILLAGE OF HARLEM....NEW YORK ! 4632  
## 5 Entire Apt: Spacious Studio/Loft by central park 7192  
## 6 Large Cozy 1 BR Apartment In Midtown East 7322  
## 8 Large Furnished Room Near B'way 8967  
## 9 Cozy Clean Guest Room - Family Apt 7490  
## neighbourhood\_group neighbourhood latitude longitude room\_type  
## 2 Manhattan Midtown 40.75362 -73.98377 Entire home/apt  
## 3 Manhattan Harlem 40.80902 -73.94190 Private room  
## 5 Manhattan East Harlem 40.79851 -73.94399 Entire home/apt  
## 6 Manhattan Murray Hill 40.74767 -73.97500 Entire home/apt  
## 8 Manhattan Hell's Kitchen 40.76489 -73.98493 Private room  
## 9 Manhattan Upper West Side 40.80178 -73.96723 Private room  
## price minimum\_nights number\_of\_reviews last\_review reviews\_per\_month  
## 2 225 1 45 5/21/2019 0.38  
## 3 150 3 0 NA  
## 5 80 10 9 11/19/2018 0.10  
## 6 200 3 74 6/22/2019 0.59  
## 8 79 2 430 6/24/2019 3.47  
## 9 79 2 118 7/21/2017 0.99  
## listing\_cnt  
## 2 2  
## 3 1  
## 5 1  
## 6 1  
## 8 1  
## 9 1

Manhattan %>%  
 group\_by (room\_type) %>%  
 summarise(AvgPrice = mean(price)) #average prices in Manhattan

## # A tibble: 3 x 2  
## room\_type AvgPrice  
## <fct> <dbl>  
## 1 Entire home/apt 249.   
## 2 Private room 117.   
## 3 Shared room 89.0

str(airbnb)

## 'data.frame': 48895 obs. of 13 variables:  
## $ description : Factor w/ 47897 levels "","'Fan'tastic",..: 12652 38163 45162 15693 19357 24992 8328 25039 15588 17673 ...  
## $ host\_id : chr "2787" "2845" "4632" "4869" ...  
## $ neighbourhood\_group: Factor w/ 5 levels "Bronx","Brooklyn",..: 2 3 3 2 3 3 2 3 3 3 ...  
## $ neighbourhood : Factor w/ 221 levels "Allerton","Arden Heights",..: 109 128 95 42 62 138 14 96 203 36 ...  
## $ latitude : num 40.6 40.8 40.8 40.7 40.8 ...  
## $ longitude : num -74 -74 -73.9 -74 -73.9 ...  
## $ room\_type : Factor w/ 3 levels "Entire home/apt",..: 2 1 2 1 1 1 2 2 2 1 ...  
## $ price : int 149 225 150 89 80 200 60 79 79 150 ...  
## $ minimum\_nights : int 1 1 3 1 10 3 45 2 2 1 ...  
## $ number\_of\_reviews : int 9 45 0 270 9 74 49 430 118 160 ...  
## $ last\_review : Factor w/ 1765 levels "","1/1/2013",..: 203 1059 1 1438 348 1234 277 1244 1383 1317 ...  
## $ reviews\_per\_month : num 0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...  
## $ listing\_cnt : int 6 2 1 1 1 1 1 1 1 4 ...