NYC\_MarkdownFile

Elaine Franta

1/15/2022

##Read in NYC AirBnB Data

rm(list = ls())  
library(plyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:plyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

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

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

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

ABNB<-read.csv("AB\_NYC\_2019.csv")  
attach(ABNB)  
  
dim(ABNB)

## [1] 48895 16

names(ABNB)

## [1] "id" "name"   
## [3] "host\_id" "host\_name"   
## [5] "neighbourhood\_group" "neighbourhood"   
## [7] "latitude" "longitude"   
## [9] "room\_type" "price"   
## [11] "minimum\_nights" "number\_of\_reviews"   
## [13] "last\_review" "reviews\_per\_month"   
## [15] "calculated\_host\_listings\_count" "availability\_365"

head(ABNB)

## id name host\_id host\_name  
## 1 2539 Clean & quiet apt home by the park 2787 John  
## 2 2595 Skylit Midtown Castle 2845 Jennifer  
## 3 3647 THE VILLAGE OF HARLEM....NEW YORK ! 4632 Elisabeth  
## 4 3831 Cozy Entire Floor of Brownstone 4869 LisaRoxanne  
## 5 5022 Entire Apt: Spacious Studio/Loft by central park 7192 Laura  
## 6 5099 Large Cozy 1 BR Apartment In Midtown East 7322 Chris  
## neighbourhood\_group neighbourhood latitude longitude room\_type price  
## 1 Brooklyn Kensington 40.64749 -73.97237 Private room 149  
## 2 Manhattan Midtown 40.75362 -73.98377 Entire home/apt 225  
## 3 Manhattan Harlem 40.80902 -73.94190 Private room 150  
## 4 Brooklyn Clinton Hill 40.68514 -73.95976 Entire home/apt 89  
## 5 Manhattan East Harlem 40.79851 -73.94399 Entire home/apt 80  
## 6 Manhattan Murray Hill 40.74767 -73.97500 Entire home/apt 200  
## minimum\_nights number\_of\_reviews last\_review reviews\_per\_month  
## 1 1 9 10/19/2018 0.21  
## 2 1 45 5/21/2019 0.38  
## 3 3 0 NA  
## 4 1 270 7/5/2019 4.64  
## 5 10 9 11/19/2018 0.10  
## 6 3 74 6/22/2019 0.59  
## calculated\_host\_listings\_count availability\_365  
## 1 6 365  
## 2 2 355  
## 3 1 365  
## 4 1 194  
## 5 1 0  
## 6 1 129

##Check for missing values

n\_distinct(id) #48895 unique ids

## [1] 48895

n\_distinct(host\_id) #37457 unique host ids

## [1] 37457

summary(ABNB) #shows 10052 NA values for reviews\_per\_month

## id name host\_id host\_name   
## Min. : 2539 Length:48895 Min. : 2438 Length:48895   
## 1st Qu.: 9471945 Class :character 1st Qu.: 7822033 Class :character   
## Median :19677284 Mode :character Median : 30793816 Mode :character   
## Mean :19017143 Mean : 67620011   
## 3rd Qu.:29152178 3rd Qu.:107434423   
## Max. :36487245 Max. :274321313   
##   
## neighbourhood\_group neighbourhood latitude longitude   
## Length:48895 Length:48895 Min. :40.50 Min. :-74.24   
## Class :character Class :character 1st Qu.:40.69 1st Qu.:-73.98   
## Mode :character Mode :character Median :40.72 Median :-73.96   
## Mean :40.73 Mean :-73.95   
## 3rd Qu.:40.76 3rd Qu.:-73.94   
## Max. :40.91 Max. :-73.71   
##   
## room\_type price minimum\_nights number\_of\_reviews  
## Length:48895 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## Class :character 1st Qu.: 69.0 1st Qu.: 1.00 1st Qu.: 1.00   
## Mode :character Median : 106.0 Median : 3.00 Median : 5.00   
## Mean : 152.7 Mean : 7.03 Mean : 23.27   
## 3rd Qu.: 175.0 3rd Qu.: 5.00 3rd Qu.: 24.00   
## Max. :10000.0 Max. :1250.00 Max. :629.00   
##   
## last\_review reviews\_per\_month calculated\_host\_listings\_count  
## Length:48895 Min. : 0.010 Min. : 1.000   
## Class :character 1st Qu.: 0.190 1st Qu.: 1.000   
## Mode :character Median : 0.720 Median : 1.000   
## Mean : 1.373 Mean : 7.144   
## 3rd Qu.: 2.020 3rd Qu.: 2.000   
## Max. :58.500 Max. :327.000   
## NA's :10052   
## availability\_365  
## Min. : 0.0   
## 1st Qu.: 0.0   
## Median : 45.0   
## Mean :112.8   
## 3rd Qu.:227.0   
## Max. :365.0   
##

##Cleaning

null\_val<-filter(ABNB, is.na(reviews\_per\_month) == TRUE)  
head(null\_val) #null when number of reviews is 0

## id name host\_id  
## 1 3647 THE VILLAGE OF HARLEM....NEW YORK ! 4632  
## 2 7750 Huge 2 BR Upper East Cental Park 17985  
## 3 8700 Magnifique Suite au N de Manhattan - vue Cloitres 26394  
## 4 11452 Clean and Quiet in Brooklyn 7355  
## 5 11943 Country space in the city 45445  
## 6 51438 1 Bedroom in 2 Bdrm Apt- Upper East 236421  
## host\_name neighbourhood\_group neighbourhood latitude longitude  
## 1 Elisabeth Manhattan Harlem 40.80902 -73.94190  
## 2 Sing Manhattan East Harlem 40.79685 -73.94872  
## 3 Claude & Sophie Manhattan Inwood 40.86754 -73.92639  
## 4 Vt Brooklyn Bedford-Stuyvesant 40.68876 -73.94312  
## 5 Harriet Brooklyn Flatbush 40.63702 -73.96327  
## 6 Jessica Manhattan Upper East Side 40.77333 -73.95199  
## room\_type price minimum\_nights number\_of\_reviews last\_review  
## 1 Private room 150 3 0   
## 2 Entire home/apt 190 7 0   
## 3 Private room 80 4 0   
## 4 Private room 35 60 0   
## 5 Private room 150 1 0   
## 6 Private room 130 14 0   
## reviews\_per\_month calculated\_host\_listings\_count availability\_365  
## 1 NA 1 365  
## 2 NA 2 249  
## 3 NA 1 0  
## 4 NA 1 365  
## 5 NA 1 365  
## 6 NA 2 0

null\_lastreview<-filter(ABNB, last\_review =='')  
head(null\_lastreview)

## id name host\_id  
## 1 3647 THE VILLAGE OF HARLEM....NEW YORK ! 4632  
## 2 7750 Huge 2 BR Upper East Cental Park 17985  
## 3 8700 Magnifique Suite au N de Manhattan - vue Cloitres 26394  
## 4 11452 Clean and Quiet in Brooklyn 7355  
## 5 11943 Country space in the city 45445  
## 6 51438 1 Bedroom in 2 Bdrm Apt- Upper East 236421  
## host\_name neighbourhood\_group neighbourhood latitude longitude  
## 1 Elisabeth Manhattan Harlem 40.80902 -73.94190  
## 2 Sing Manhattan East Harlem 40.79685 -73.94872  
## 3 Claude & Sophie Manhattan Inwood 40.86754 -73.92639  
## 4 Vt Brooklyn Bedford-Stuyvesant 40.68876 -73.94312  
## 5 Harriet Brooklyn Flatbush 40.63702 -73.96327  
## 6 Jessica Manhattan Upper East Side 40.77333 -73.95199  
## room\_type price minimum\_nights number\_of\_reviews last\_review  
## 1 Private room 150 3 0   
## 2 Entire home/apt 190 7 0   
## 3 Private room 80 4 0   
## 4 Private room 35 60 0   
## 5 Private room 150 1 0   
## 6 Private room 130 14 0   
## reviews\_per\_month calculated\_host\_listings\_count availability\_365  
## 1 NA 1 365  
## 2 NA 2 249  
## 3 NA 1 0  
## 4 NA 1 365  
## 5 NA 1 365  
## 6 NA 2 0

sum(last\_review == '') #10052 empty values of last\_review

## [1] 10052

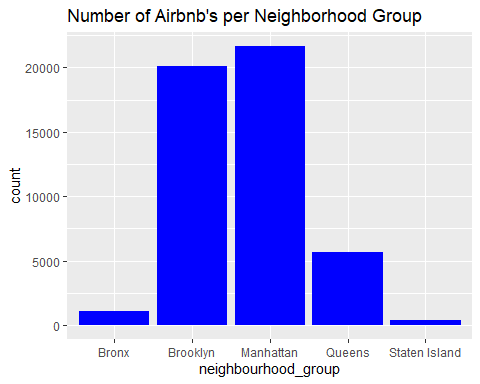
#When number of reviews is 0, last\_review is empty and reviews\_per\_month is NA  
  
#replace NA values with 0   
ABNB$reviews\_per\_month[is.na(reviews\_per\_month) == TRUE] = 0  
sum(is.na(reviews\_per\_month)) #Check that replacement worked

## [1] 10052

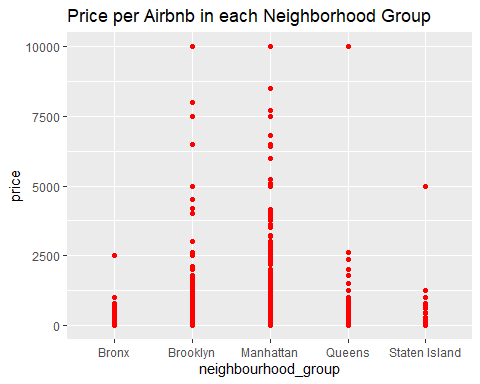
#Make factors  
ABNB$neighbourhood\_group<-factor(neighbourhood\_group)  
ABNB$room\_type <-factor(room\_type)

##Data Analysis

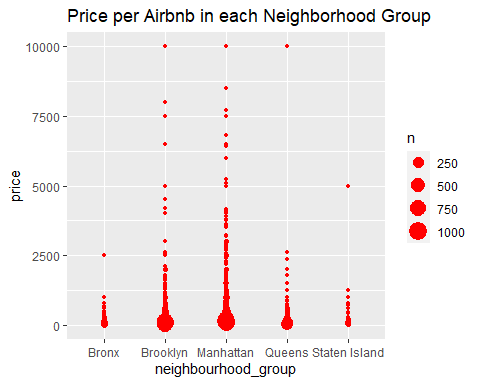
ggplot(data = ABNB)+ ggtitle("Number of Airbnb's per Neighborhood Group")+  
 geom\_bar(mapping = aes(x = neighbourhood\_group),fill = "blue")



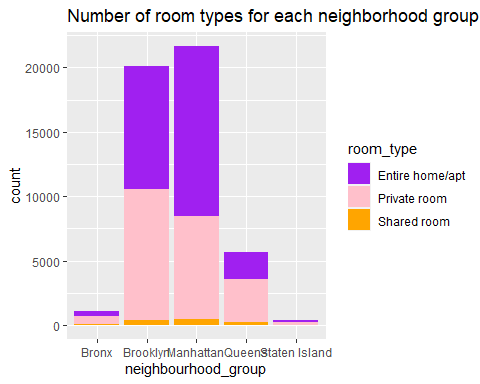
ggplot(data = ABNB) + ggtitle("Price per Airbnb in each Neighborhood Group")+  
 geom\_point(mapping = aes(x = neighbourhood\_group, y = price), color = "red")



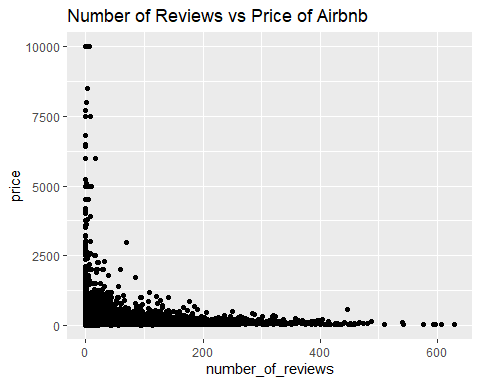
ggplot(data = ABNB) + ggtitle("Price per Airbnb in each Neighborhood Group")+  
 geom\_count(mapping = aes(x = neighbourhood\_group, y = price), color = "red")



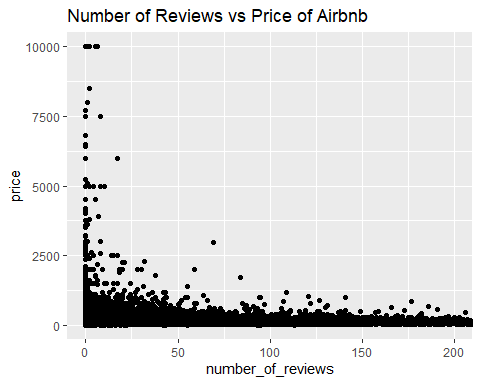
ggplot(data = ABNB, aes(x = neighbourhood\_group, fill = room\_type))+geom\_bar()+  
 ggtitle("Number of room types for each neighborhood group")+  
 scale\_fill\_manual(values = c("Entire home/apt" = "purple",  
 "Private room" = "pink",  
 "Shared room" = "orange"))



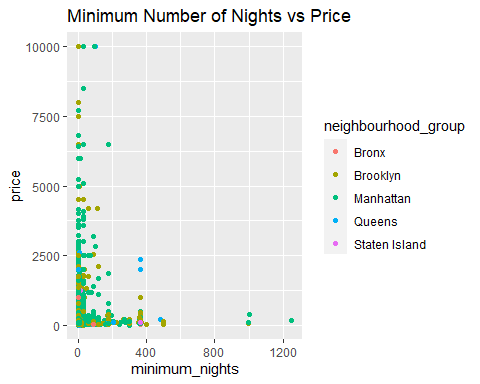
ggplot(data = ABNB, aes(x = number\_of\_reviews, y = price)) +   
 ggtitle("Number of Reviews vs Price of Airbnb")+ geom\_point()



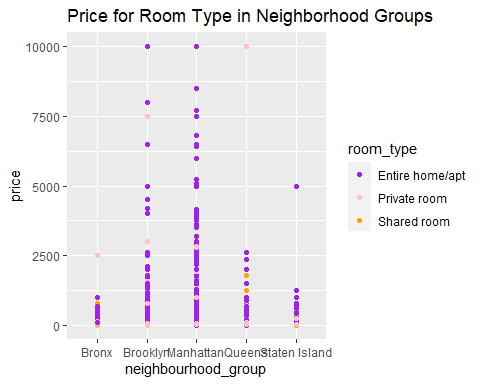
ggplot(data = ABNB, aes(x = number\_of\_reviews, y = price)) +   
 ggtitle("Number of Reviews vs Price of Airbnb")+  
 geom\_point() +   
 coord\_cartesian(xlim = c(0,200))



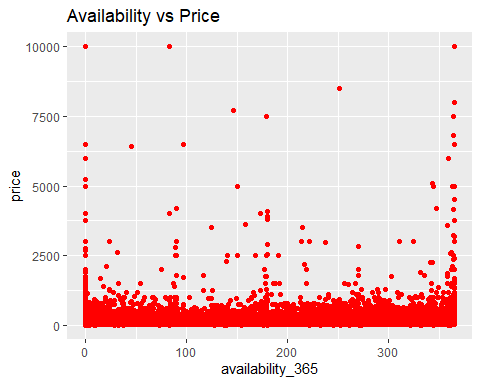
ggplot(data = ABNB) + ggtitle("Minimum Number of Nights vs Price")+  
 geom\_point(mapping = aes(x = minimum\_nights, y = price, color = neighbourhood\_group))



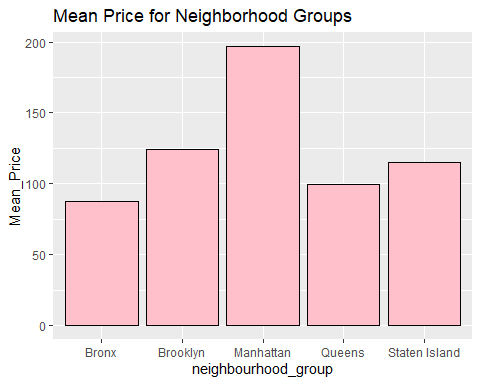
ggplot(data = ABNB, aes(x = neighbourhood\_group, y =price, color = room\_type)) +   
 geom\_point() + ggtitle("Price for Room Type in Neighborhood Groups")+  
 scale\_color\_manual(values = c("Entire home/apt" = "purple",  
 "Private room" = "pink",  
 "Shared room" = "orange"))



ggplot(data = ABNB, aes(x = availability\_365, y = price)) +   
 ggtitle("Availability vs Price")+  
 geom\_point(color = "red")



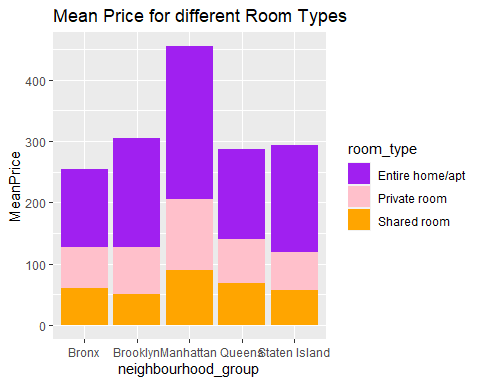
#Using mean price  
NGanalysis <- ABNB %>%  
 group\_by(neighbourhood\_group) %>% summarise(Mean\_Price = mean(price))  
  
ggplot(NGanalysis, aes(x = neighbourhood\_group, y = Mean\_Price)) +   
 ggtitle("Mean Price for Neighborhood Groups")+  
 geom\_col(fill = "pink", color = "black")



#Examine room\_type and mean prices for neighborhood groups  
rt\_price<- ABNB %>%  
 group\_by(neighbourhood\_group, room\_type) %>% summarise(MeanPrice = mean(price))

## `summarise()` has grouped output by 'neighbourhood\_group'. You can override using the `.groups` argument.

ggplot(rt\_price, aes(x = neighbourhood\_group, y = MeanPrice, fill = room\_type)) +   
 geom\_col() + ggtitle("Mean Price for different Room Types")+  
 scale\_fill\_manual(values = c("Entire home/apt" = "purple",  
 "Private room" = "pink",  
 "Shared room" = "orange"))



#Most expensive neighborhoods  
Neighbour\_analysis<- ABNB %>%  
 group\_by(neighbourhood) %>% summarise(MeanP = mean(price))  
  
highest <- Neighbour\_analysis[with(Neighbour\_analysis, order(-MeanP)),]  
  
top<-highest[1:10,]  
  
ggplot(top, aes(y = neighbourhood, x = MeanP)) +   
 ggtitle("Top 10 Most expensive Neighborhoods for Airbnbs")+  
 geom\_col(fill = "purple")

