Project\_NY\_Airbnb

Pradeep Paladugula

6/8/2020

# Loading the libraries  
library('dplyr')

##   
## 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

#Dataset  
airbnb\_ny = read.csv('AB\_NYC\_2019.csv')  
head(airbnb\_ny)

## 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 2018-10-19 0.21  
## 2 1 45 2019-05-21 0.38  
## 3 3 0 NA  
## 4 1 270 2019-07-05 4.64  
## 5 10 9 2018-11-19 0.10  
## 6 3 74 2019-06-22 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

## 'data.frame': 48895 obs. of 16 variables:  
## $ id : int 2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...  
## $ name : Factor w/ 47906 levels ""," 1 Bed Apt in Utopic Williamsburg ",..: 12573 38016 45018 15591 19219 24849 8257 24896 15486 17573 ...  
## $ host\_id : int 2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...  
## $ host\_name : Factor w/ 11453 levels "","​ Valéria",..: 4997 4791 2913 6210 5929 1938 3549 9649 6880 1235 ...  
## $ 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 "","2011-03-28",..: 1503 1717 1 1762 1534 1749 1124 1751 1048 1736 ...  
## $ 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 ...

# Checking summary  
summary(airbnb\_ny)

## id name   
## Min. : 2539 Hillside Hotel : 18   
## 1st Qu.: 9471945 Home away from home : 17   
## Median :19677284 : 16   
## Mean :19017143 New york Multi-unit building : 16   
## 3rd Qu.:29152178 Brooklyn Apartment : 12   
## Max. :36487245 Loft Suite @ The Box House Hotel: 11   
## (Other) :48805   
## host\_id host\_name neighbourhood\_group  
## Min. : 2438 Michael : 417 Bronx : 1091   
## 1st Qu.: 7822033 David : 403 Brooklyn :20104   
## Median : 30793816 Sonder (NYC): 327 Manhattan :21661   
## Mean : 67620011 John : 294 Queens : 5666   
## 3rd Qu.:107434423 Alex : 279 Staten Island: 373   
## Max. :274321313 Blueground : 232   
## (Other) :46943   
## neighbourhood latitude longitude   
## Williamsburg : 3920 Min. :40.50 Min. :-74.24   
## Bedford-Stuyvesant: 3714 1st Qu.:40.69 1st Qu.:-73.98   
## Harlem : 2658 Median :40.72 Median :-73.96   
## Bushwick : 2465 Mean :40.73 Mean :-73.95   
## Upper West Side : 1971 3rd Qu.:40.76 3rd Qu.:-73.94   
## Hell's Kitchen : 1958 Max. :40.91 Max. :-73.71   
## (Other) :32209   
## room\_type price minimum\_nights number\_of\_reviews  
## Entire home/apt:25409 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## Private room :22326 1st Qu.: 69.0 1st Qu.: 1.00 1st Qu.: 1.00   
## Shared room : 1160 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  
## :10052 Min. : 0.010 Min. : 1.000   
## 2019-06-23: 1413 1st Qu.: 0.190 1st Qu.: 1.000   
## 2019-07-01: 1359 Median : 0.720 Median : 1.000   
## 2019-06-30: 1341 Mean : 1.373 Mean : 7.144   
## 2019-06-24: 875 3rd Qu.: 2.020 3rd Qu.: 2.000   
## 2019-07-07: 718 Max. :58.500 Max. :327.000   
## (Other) :33137 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   
##

# Replacing the N/A's with 0s  
clean\_airbnb\_ny <- replace(airbnb\_ny,is.na(airbnb\_ny),0)  
  
# Checking summary again  
summary(clean\_airbnb\_ny)

## id name   
## Min. : 2539 Hillside Hotel : 18   
## 1st Qu.: 9471945 Home away from home : 17   
## Median :19677284 : 16   
## Mean :19017143 New york Multi-unit building : 16   
## 3rd Qu.:29152178 Brooklyn Apartment : 12   
## Max. :36487245 Loft Suite @ The Box House Hotel: 11   
## (Other) :48805   
## host\_id host\_name neighbourhood\_group  
## Min. : 2438 Michael : 417 Bronx : 1091   
## 1st Qu.: 7822033 David : 403 Brooklyn :20104   
## Median : 30793816 Sonder (NYC): 327 Manhattan :21661   
## Mean : 67620011 John : 294 Queens : 5666   
## 3rd Qu.:107434423 Alex : 279 Staten Island: 373   
## Max. :274321313 Blueground : 232   
## (Other) :46943   
## neighbourhood latitude longitude   
## Williamsburg : 3920 Min. :40.50 Min. :-74.24   
## Bedford-Stuyvesant: 3714 1st Qu.:40.69 1st Qu.:-73.98   
## Harlem : 2658 Median :40.72 Median :-73.96   
## Bushwick : 2465 Mean :40.73 Mean :-73.95   
## Upper West Side : 1971 3rd Qu.:40.76 3rd Qu.:-73.94   
## Hell's Kitchen : 1958 Max. :40.91 Max. :-73.71   
## (Other) :32209   
## room\_type price minimum\_nights number\_of\_reviews  
## Entire home/apt:25409 Min. : 0.0 Min. : 1.00 Min. : 0.00   
## Private room :22326 1st Qu.: 69.0 1st Qu.: 1.00 1st Qu.: 1.00   
## Shared room : 1160 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  
## :10052 Min. : 0.000 Min. : 1.000   
## 2019-06-23: 1413 1st Qu.: 0.040 1st Qu.: 1.000   
## 2019-07-01: 1359 Median : 0.370 Median : 1.000   
## 2019-06-30: 1341 Mean : 1.091 Mean : 7.144   
## 2019-06-24: 875 3rd Qu.: 1.580 3rd Qu.: 2.000   
## 2019-07-07: 718 Max. :58.500 Max. :327.000   
## (Other) :33137   
## availability\_365  
## Min. : 0.0   
## 1st Qu.: 0.0   
## Median : 45.0   
## Mean :112.8   
## 3rd Qu.:227.0   
## Max. :365.0   
##

# Checking whether any Null values / bad data in the dataframe after cleansing  
sapply(clean\_airbnb\_ny, function(x)all(any(is.na(x))))

## id name   
## FALSE FALSE   
## host\_id host\_name   
## FALSE FALSE   
## neighbourhood\_group neighbourhood   
## FALSE FALSE   
## latitude longitude   
## FALSE FALSE   
## room\_type price   
## FALSE FALSE   
## minimum\_nights number\_of\_reviews   
## FALSE FALSE   
## last\_review reviews\_per\_month   
## FALSE FALSE   
## calculated\_host\_listings\_count availability\_365   
## FALSE FALSE