Airbnb project

Fengshou Liang

2020-9-23

This data file contains information on Airbnb in New York City, including all needed information about hosts, availability, geography and etc. The aim of this project is cleaning and exploring the data file, and make sure the following questions are answered.

1. How many observations can be made if we adjust the data size according to last review date and availability?
2. which neighbourhood group has the highest average price in 2019?
3. What can we learn about different hosts?
4. Which hosts and areas are the busiest and why?

First load relevant packages needed for inspection.

library(tidyverse)  
library(dplyr)  
library(readr)

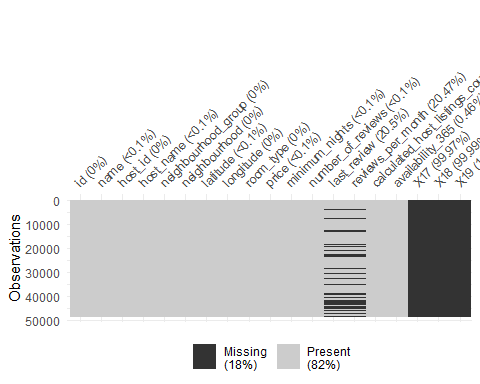
Load data and take a look at the first 6 observations.

head(AB\_NYC)

## # A tibble: 6 x 19  
## id name host\_id host\_name neighbourhood\_g~ neighbourhood latitude  
## <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 2539 Clea~ 2787 John Brooklyn Kensington 40.6  
## 2 2595 Skyl~ 2845 Jennifer Manhattan Midtown 40.8  
## 3 3647 THE ~ 4632 Elisabeth Manhattan Harlem 40.8  
## 4 3831 Cozy~ 4869 LisaRoxa~ Brooklyn Clinton Hill 40.7  
## 5 5022 Enti~ 7192 Laura Manhattan East Harlem 40.8  
## 6 5099 Larg~ 7322 Chris Manhattan Murray Hill 40.7  
## # ... with 12 more variables: longitude <chr>, room\_type <chr>, price <dbl>,  
## # minimum\_nights <dbl>, number\_of\_reviews <chr>, last\_review <chr>,  
## # reviews\_per\_month <dbl>, calculated\_host\_listings\_count <dbl>,  
## # availability\_365 <dbl>, X17 <lgl>, X18 <lgl>, X19 <lgl>

Now we want to see the distribution of missing values, and this can be done in an interesting way by using the visdat package, but it is better not to use this method when the data size is too big.

library(visdat)  
vis\_miss(AB\_NYC, warn\_large\_data = FALSE)



AB\_NYC <- AB\_NYC %>% select(-c(17, 18, 19))

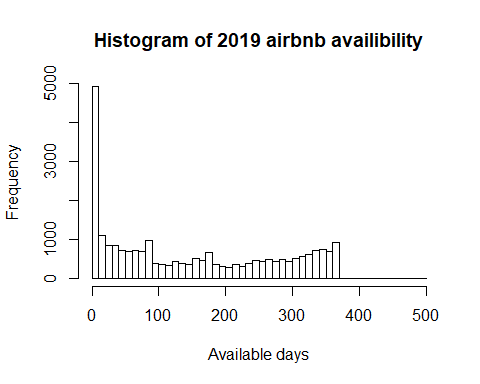
As seen from the plot, most missing values occur in the last three columns as they are empty and some occur in the column "last review". As a result, we are removing columns 17-19.

Choose data recording to the last review date that is in 2019.

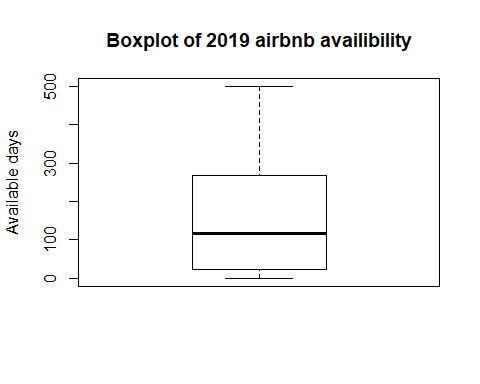
AB\_NYC\_2019 <- AB\_NYC %>%   
 drop\_na(last\_review) %>%  
 #remove na in column last\_review  
 filter(last\_review >= "2019-1-01" & last\_review < "2020-1-01")  
  
##make sure there is no missing date in 2019  
library(assertive)  
assert\_all\_are\_not\_na(AB\_NYC\_2019$last\_review)

Now let's take a look at availibilty of airbnb and make sure there is no strange value.

hist(AB\_NYC\_2019$availability\_365, breaks = 50, main = "Histogram of 2019 airbnb availibility", xlab="Available days")



boxplot(AB\_NYC\_2019$availability\_365, main="Boxplot of 2019 airbnb availibility", ylab = "Available days")



Although the histogram appears to be acceptable, the boxplot shows there might be some obsurdities. In fact, there are observations which have availibilty more than 365 days.

AB\_NYC\_2019 %>%  
 drop\_na(availability\_365) %>%  
 filter(availability\_365 > 365)

## # A tibble: 2 x 16  
## id name host\_id host\_name neighbourhood\_g~ neighbourhood latitude  
## <dbl> <chr> <chr> <chr> <chr> <chr> <dbl>  
## 1 9783 back~ 32294 Ssameer ~ Manhattan Harlem 40.8  
## 2 15711 2 be~ 61491 D Manhattan Upper East S~ 40.8  
## # ... with 9 more variables: longitude <chr>, room\_type <chr>, price <dbl>,  
## # minimum\_nights <dbl>, number\_of\_reviews <chr>, last\_review <chr>,  
## # reviews\_per\_month <dbl>, calculated\_host\_listings\_count <dbl>,  
## # availability\_365 <dbl>

After excluding the two observations outside of range (0-365), we have a new dataset.

AB\_NYC\_2019\_fixed <-  
 AB\_NYC\_2019 %>%   
 filter(availability\_365 <= 365)

Thus, in the end, we have reduced the data size from 48576 to 24864. Note that in this scenario, the two outliers can be deleted because the data size still remains large after such procedure, so that the outcome won't change much. However, if the data size is limited, it is best to take an educated guess at the two outliers of what they could possibly be. For example, the availabilities are 400 and 500 days, but it is plausible that these two pieces of data were mistyped and should have been 40 and 50 days.

To answer question 2), let's first take a look at all the neighbourhood group we have in hand.

AB\_NYC\_2019\_fixed %>%  
 count(neighbourhood\_group)

## # A tibble: 8 x 2  
## neighbourhood\_group n  
## <chr> <int>  
## 1 Broncx 1  
## 2 Bronx 692  
## 3 Brooklyn 10351  
## 4 Manhatan 1  
## 5 Manhattan 10178  
## 6 Queen 1  
## 7 Queens 3377  
## 8 Staten Island 263

we find some typos in the names of neighbourhood group, thus we need to change them.

AB\_NYC\_2019\_updated <-   
AB\_NYC\_2019\_fixed %>%   
 mutate(neighbourhood\_group\_collapsed =   
 fct\_collapse(neighbourhood\_group,   
 Bronx = "Broncx", Manhattan = "Manhatan", Queens = "Queen"))   
AB\_NYC\_2019\_updated %>%   
 filter(price > 0) %>%  
 group\_by(neighbourhood\_group\_collapsed) %>%  
 summarise(avg\_price = mean(price)) %>%  
 arrange(desc(avg\_price))

## # A tibble: 5 x 2  
## neighbourhood\_group\_collapsed avg\_price  
## <fct> <dbl>  
## 1 Manhattan 184.   
## 2 Brooklyn 122.   
## 3 Queens 93.0  
## 4 Staten Island 86.3  
## 5 Bronx 84.4