Case Study 2 EDA

Frances Hung

1/24/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(ggplot2)
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
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(tidyr)
```

```
AB<-read.csv("AB_NYC_2019.csv") %>% filter(grepl("2018",last_review)==TRUE)
colMeans(is.na(AB))
```

```
##                              id                            name
##                               0                               0
##                         host_id                       host_name
##                               0                               0
##             neighbourhood_group                   neighbourhood
##                               0                               0
##                        latitude                       longitude
##                               0                               0
##                       room_type                           price
##                               0                               0
##                  minimum_nights               number_of_reviews
```

```
##                                 0                              0
##                       last_review              reviews_per_month
##                                 0                              0
## calculated_host_listings_count                availability_365
##                                 0                              0
```

The only category with missing data is the reviews per month variable. There doesn't seem to be an obvious pattern to the missingness; the neighborhoods with more missing data are the neighborhoods which have more listings AND there is no missing data if the date of the last review is recent (i.e. more than 2018). We're interested in current trends anyways, so we can get rid of data where the last review is before 2018.

```r
# AB %>% filter(reviews_per_month %>% is.na()) %>% group_by(neighbourhood) %>% summarise(count=n(),med_

AB %>% group_by(neighbourhood) %>% summarise(count=n(),med_price=median(price)) %>% arrange(desc(count)
```

```
## # A tibble: 170 x 3
##    neighbourhood      count med_price
##    <fct>              <int>     <dbl>
##  1 Williamsburg         535     100
##  2 Bedford-Stuyvesant   433      75
##  3 Harlem               368      87.5
##  4 Bushwick             359      60
##  5 Upper West Side      257     145
##  6 Upper East Side      240     135
##  7 Crown Heights        224      85
##  8 East Village         217     149
##  9 Hell's Kitchen       199     146
## 10 Midtown              167     185
## # ... with 160 more rows
```

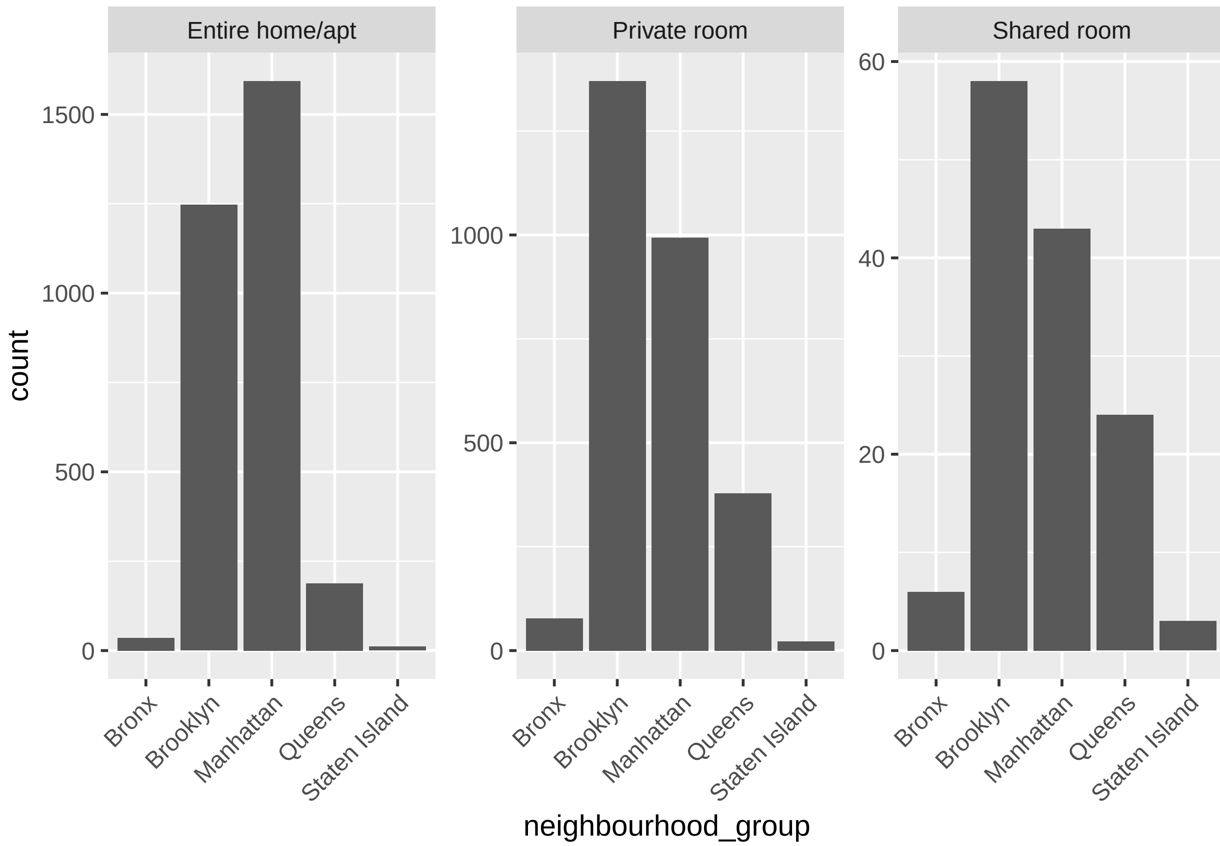
```r
# AB %>%
#   group_by(neighbourhood) %>%
#   summarise(n = count(is.na(reviews_per_month))) %>%
#   mutate(freq = n / sum(n))

# AB %>% group_by(neighbourhood) %>% group_by(neighbourhood) %>% summarise(y=count(is.na(reviews_per_mo
```

Including Plots

```r
ggplot(AB,aes(x=neighbourhood_group))+
  geom_histogram(stat = "count") +
  facet_wrap(.~room_type,scale="free") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
AB_w_couples<-AB %>% mutate(couples=ifelse(grepl("and |AND |And | \\& ",host_name),1,0))
```

Leased entire houses/apartments are the most common room type Airbnb offers in Manhattan, while in Brooklyn where living space tends to be larger, private rooms are also common offer. Queens offers mostly private rooms.

The median price of housing listed under couples is about the same as those listed under singles.

```
AB_w_couples %>%
  group_by(neighbourhood) %>%
  summarise(median_price=median(price), q25=quantile(price,.25),q75=quantile(price,.75),count=n()) %>%
  arrange(desc(median_price)) %>%
  filter(count>50)
```

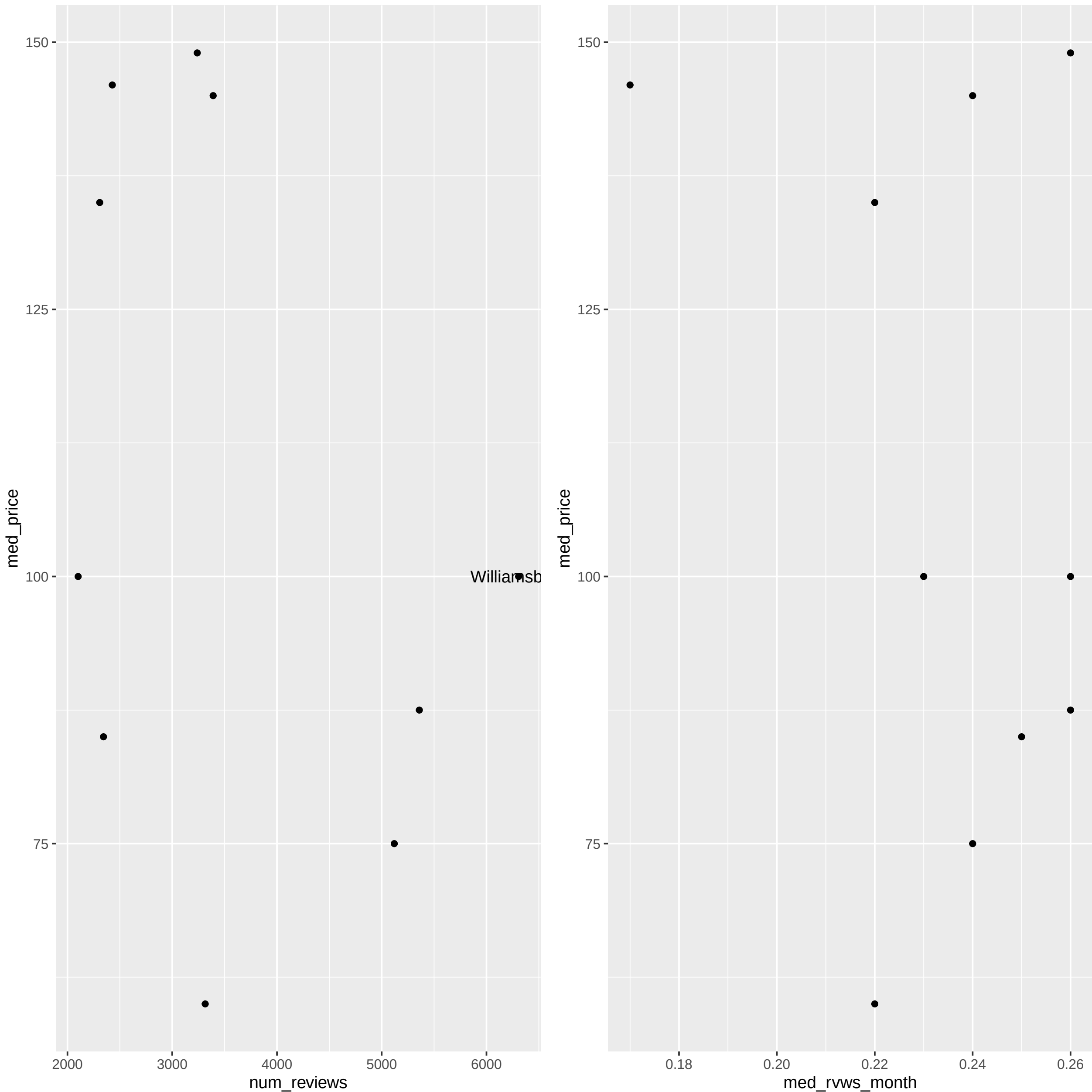
```
## # A tibble: 33 x 5
##    neighbourhood      median_price   q25   q75 count
##    <fct>                     <dbl> <dbl> <dbl> <int>
##  1 West Village                197  158.  250     92
##  2 Midtown                     185  138.  258.   167
##  3 Greenwich Village           180  125   250     57
##  4 Chelsea                     175  121.  250    110
##  5 Financial District          175  118.  240     62
##  6 Murray Hill                 162.  134.  200.    58
##  7 Kips Bay                    150  100   176.    70
##  8 East Village                149  103   199    217
##  9 Hell's Kitchen              146   99   200    199
## 10 Upper West Side             145   99   200    257
## # ... with 23 more rows
```

```r
most_pop_neighborhoods<-AB %>% drop_na() %>% group_by(neighbourhood) %>%
  summarise(num_reviews=sum(number_of_reviews),med_price=median(price),med_rvws_month=median(reviews_pe
  filter(num_reviews>2000)

total_rvws_plot<-most_pop_neighborhoods %>% ggplot(aes(x=num_reviews,y=med_price))+geom_point()+
  geom_text(data=subset(most_pop_neighborhoods, num_reviews>quantile(num_reviews,.9) | med_price>150),a

per_month_rvws_plot<-most_pop_neighborhoods %>% ggplot(aes(x=med_rvws_month,y=med_price))+geom_point()+
  geom_text(data=subset(most_pop_neighborhoods, med_rvws_month>quantile(med_rvws_month,.9) | med_price>

grid.arrange(total_rvws_plot,per_month_rvws_plot,ncol=2)
```



```
most_pop_neighborhoods %>% filter(num_reviews>quantile(num_reviews,.9))
```

```
## # A tibble: 1 x 6
##   neighbourhood num_reviews med_price med_rvws_month district available
##   <fct>               <int>     <dbl>          <dbl> <fct>        <dbl>
## 1 Williamsburg         6307       100           0.23 Brooklyn         0
```

```
most_pop_neighborhoods %>% filter(med_rvws_month>quantile(med_rvws_month,.9))
```

```
## # A tibble: 0 x 6
## # ... with 6 variables: neighbourhood <fct>, num_reviews <int>,
## #   med_price <dbl>, med_rvws_month <dbl>, district <fct>, available <dbl>
```

There seems to be a correlation between number of reviews per month and number of reviews, but it is not absolute. Perhaps the reviews per month is more indicative of up-and-coming neighborhoods than the total number (which may include Airbnbs which have been on the market for a long time). Looking at the total number of reviews versus median reviews per month, we can see that we have expensive rentals with relatively low numbers of reviews; these also unsurprisingly correspond to low numbers of reviews per month.

Things get interesting when we look at the neighborhoods with most total number of reviews (mostly in Brooklyn and Manhattan) and neighborhoods with the most reviews per month (mostly in Queens).

Manhattan/Brooklyn has quite a few renters who usually have available full-apartment space to rent for two or three months every year; we'd assume that they are likely people who rent out the spaces they live in while they're on vacation. Queens has quite a few renters who are renting private rooms or full apartments for a much larger portion of the year for cheaper; they probably have designated rooms for renting out. now does days available correspond to types of rooms? maybe a better profit metric is dollars per review per day available.

This is a hierarchical model: important metrics seem to be neighborhood_group, possibly the metric described above,