midterm-proj

Yudong Wang

4/18/2021

```r
library(tidyverse)
```

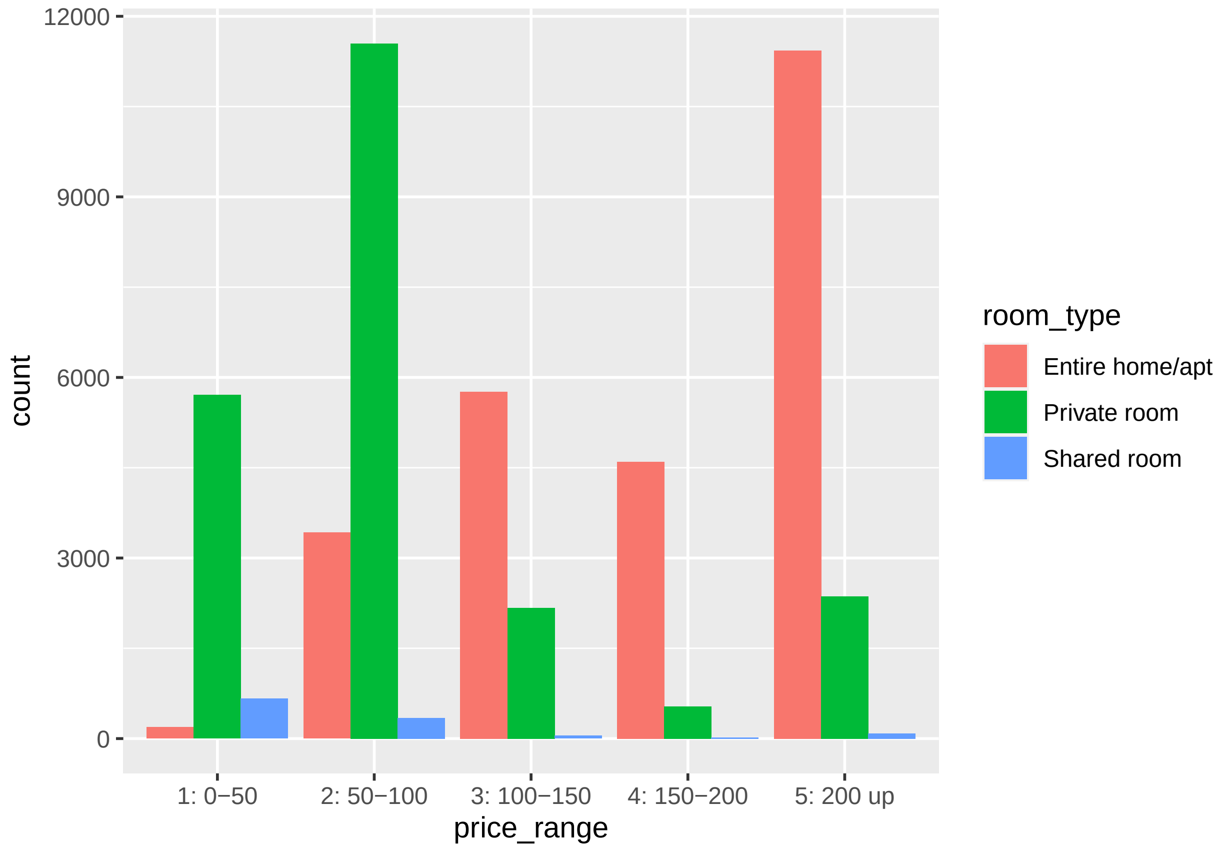
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(ggplot2)
# read data
# setwd("C:/Users/wyd98/Desktop/homework")
mydata = read.csv("Airbnb_NYC_2019.csv")
# remove unnecessary columns
data1 = mydata[,c(5:6, 9:16)]
# turn price into price ranges (categorical)
price_range = c()
for (i in 1:48895) {
  if (data1$price[i] <= 50){
    price_range = c(price_range, "1: 0-50")
  } else if (data1$price[i] >50 & data1$price[i] < 100) {
    price_range = c(price_range, "2: 50-100")
  } else if (data1$price[i] >100 & data1$price[i] < 150) {
    price_range = c(price_range, "3: 100-150")
  } else if (data1$price[i] >150 & data1$price[i] < 200) {
    price_range = c(price_range, "4: 150-200")
  } else {
    price_range = c(price_range, "5: 200 up")
  }
}
data1 = cbind(data1, price_range)
```

```r
ggplot(data1, aes(price_range)) + geom_bar(aes(fill = room_type), position = "dodge")
```

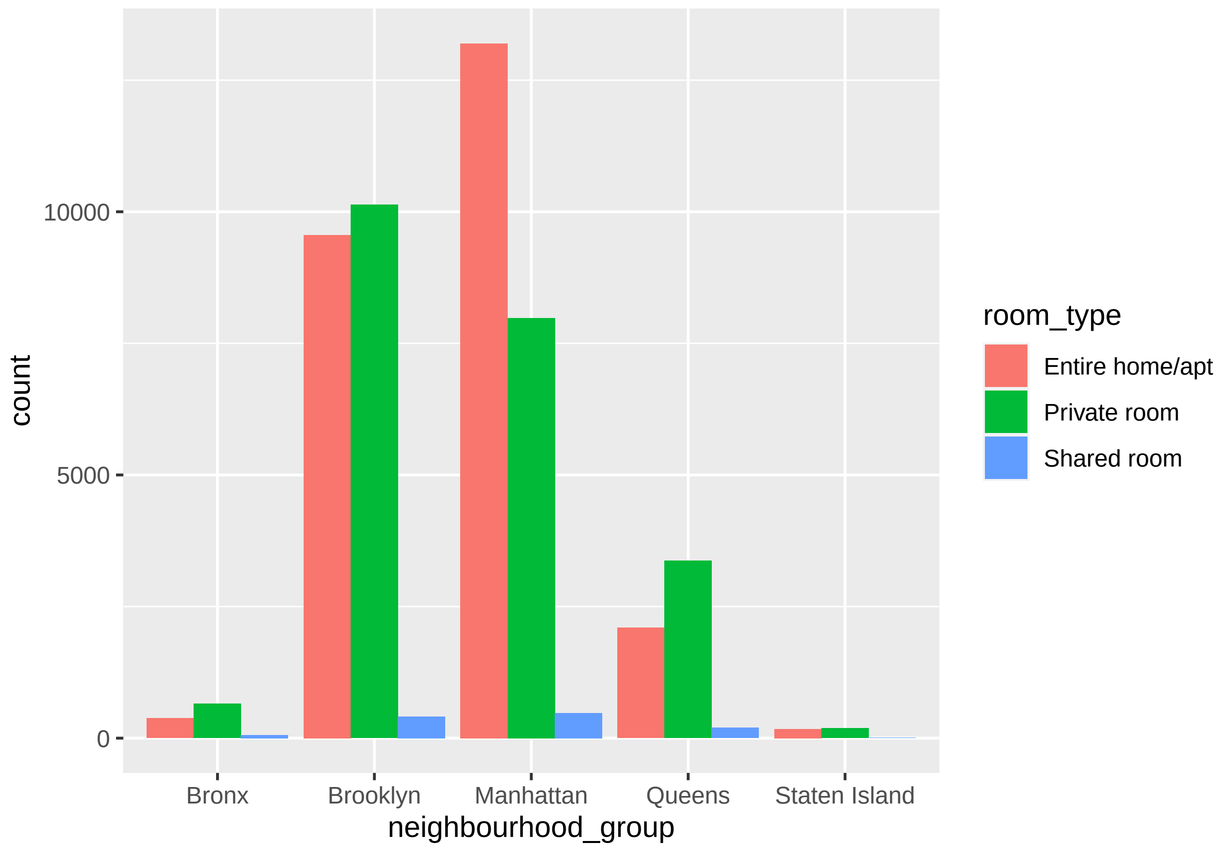


```
# room type and their avg price
data1 %>%
  group_by(room_type) %>%
  summarise(u = mean(price))
```

```
## # A tibble: 3 x 2
##   room_type           u
## * <chr>           <dbl>
## 1 Entire home/apt 212.
## 2 Private room     89.8
## 3 Shared room      70.1
```

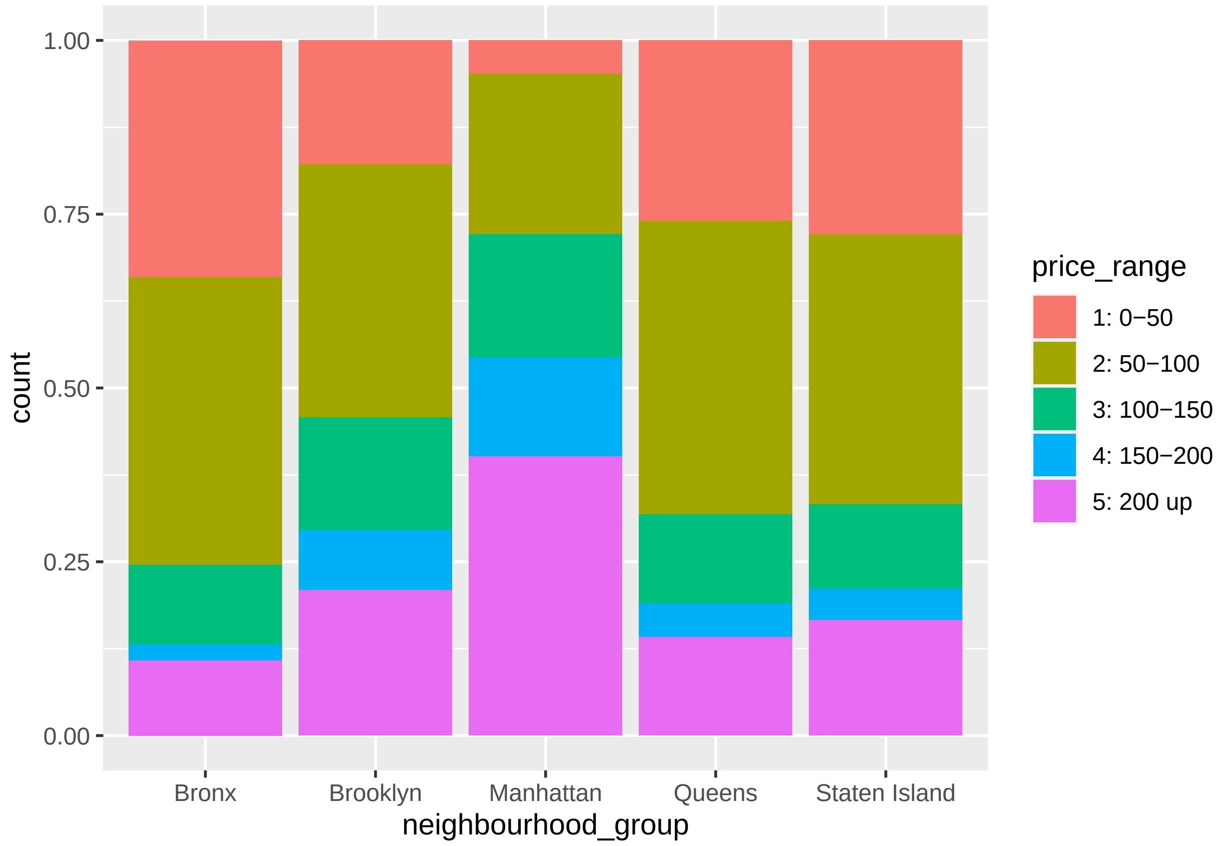
proof of the entire room is more expensive. (graph + avg price)

```
# histogram of city and room type
ggplot(data1, aes(neighbourhood_group)) + geom_bar(aes(fill = room_type), position = "dodge")
```



### Entire home and private room much more popular than shared room. People want privacy. Most boroughts have more private rooms than entire room, but not in Manhattan. Reasonable because Manhattan is the richest boroughs in NYC ### https://nypost.com/2019/12/12/gdp-in-nycs-outer-boroughs-leads-state-in-economic-output/

```
ggplot(data1, aes(neighbourhood_group)) + geom_bar(aes(fill = price_range), position = "fill")
```

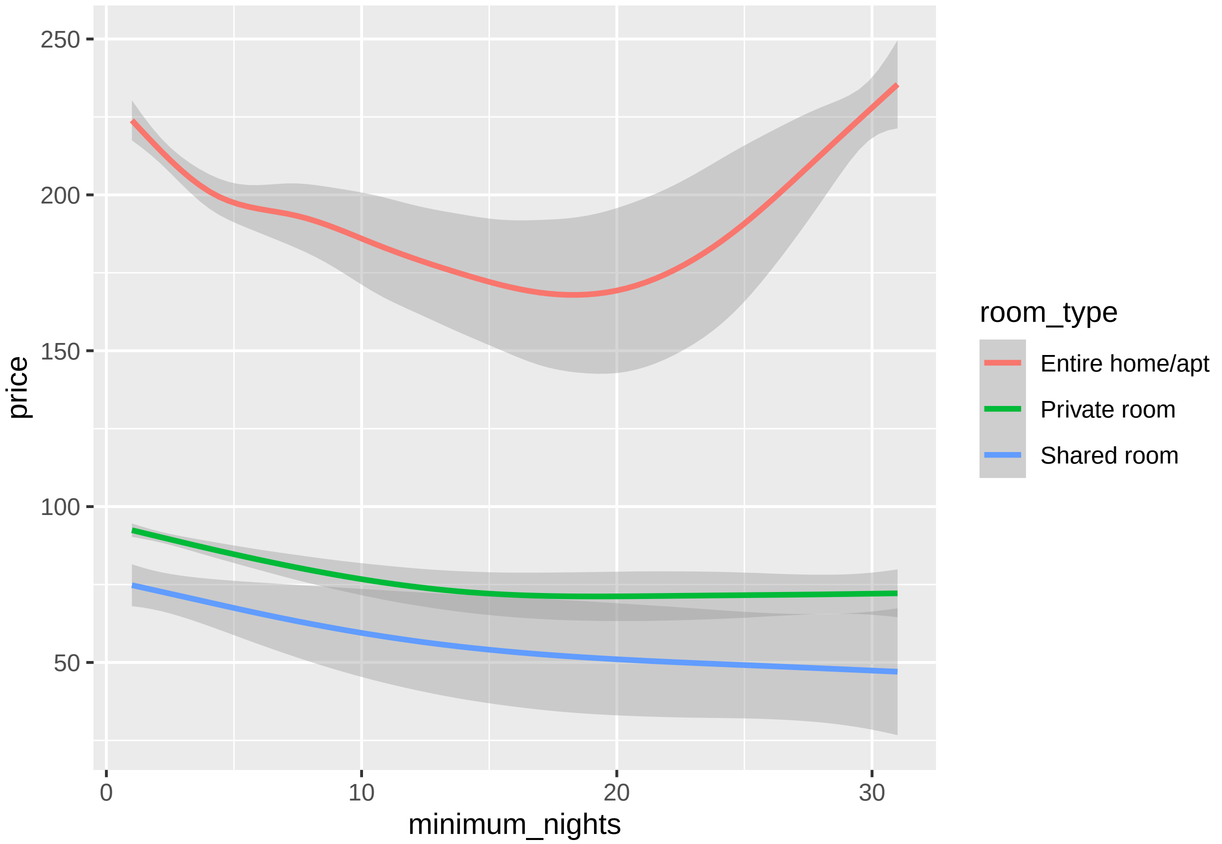


### Combine previous 2 graph's information, get distribution

what affects price?

```
# get short term rental data
# minimum nights <= 31 (one month)
mn31 = data1 %>%
  filter(minimum_nights <= 31)
# plot against price and room type
ggplot(mn31, aes(minimum_nights, price)) + geom_smooth(aes(color = room_type))
```

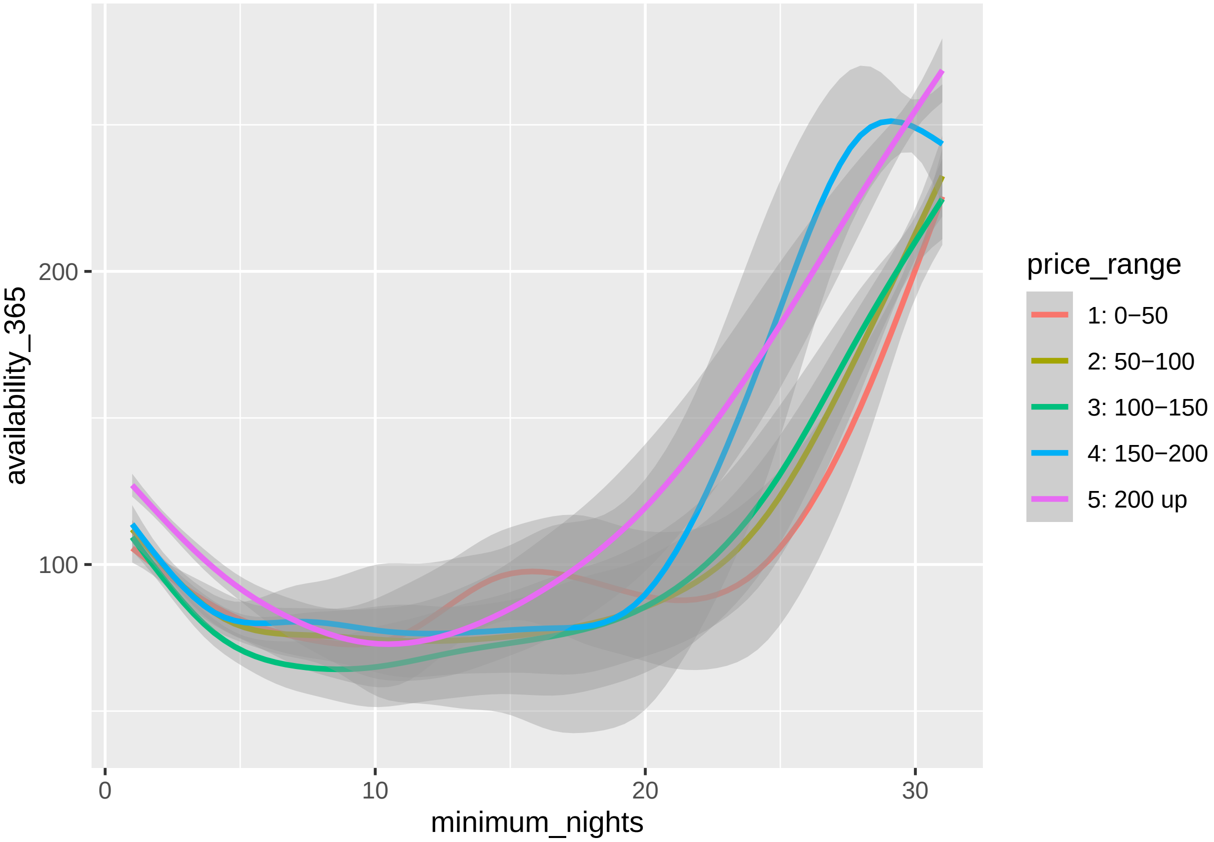
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



### Private and shared rooms' prices drop as minimum night increases. This fits economic rule (like risk premium?), the longer stay, the lower per night price. ### Entire room's price decrease first, then start increasing at mn=20. Maybe longer stay means the room is better? I'm not sure.

```
a50 = data1 %>%
  filter(availability_365 <= 50 & minimum_nights <= 31)
ggplot(mn31, aes(minimum_nights, availability_365)) + geom_smooth(aes(col = price_range))
```

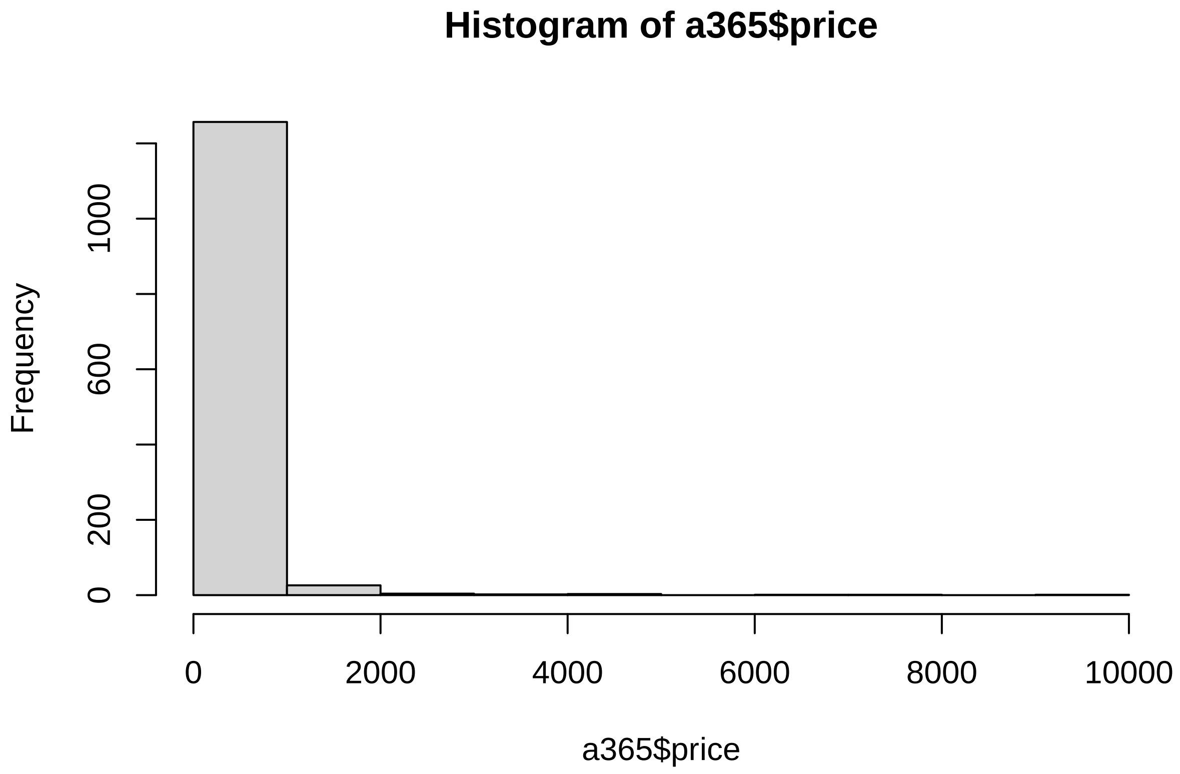
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



### Overall, availability would increase as minimum nights increase. This fits our expectation because people who rent Airbnb prefer stay short in a room (people can just rent a house/apartment if they have to stay long). Or people tend to rent Airbnb during vacation, which is around a week or two.

Just some sketches, you can ignore it.

```
a365 = data1 %>%
  filter(availability_365 == 365)
hist(a365$price)
```



```
data1 %>%
  group_by(neighbourhood_group, price_range) %>%
  summarize(u = mean(availability_365))
```

```
## `summarise()` has grouped output by 'neighbourhood_group'. You can override using the `.groups` argu
```

```
## # A tibble: 25 x 3
## # Groups:   neighbourhood_group [5]
##    neighbourhood_group price_range     u
##    <chr>               <chr>       <dbl>
##  1 Bronx               1: 0-50     152.
##  2 Bronx               2: 50-100   171.
##  3 Bronx               3: 100-150  174.
##  4 Bronx               4: 150-200  153.
##  5 Bronx               5: 200 up   184.
##  6 Brooklyn            1: 0-50      93.4
##  7 Brooklyn            2: 50-100    96.2
##  8 Brooklyn            3: 100-150  100.
##  9 Brooklyn            4: 150-200  115.
## 10 Brooklyn            5: 200 up   107.
## # ... with 15 more rows
```

```
data1 %>%
  filter(availability_365 <= 50) %>%
  group_by(neighbourhood_group, price_range) %>%
  count()
```

```
## # A tibble: 25 x 3
## # Groups:   neighbourhood_group, price_range [25]
##    neighbourhood_group price_range     n
##    <chr>               <chr>       <int>
##  1 Bronx               1: 0-50       116
##  2 Bronx               2: 50-100     120
##  3 Bronx               3: 100-150     34
##  4 Bronx               4: 150-200      8
##  5 Bronx               5: 200 up      30
##  6 Brooklyn            1: 0-50      2166
##  7 Brooklyn            2: 50-100    4106
##  8 Brooklyn            3: 100-150   1743
##  9 Brooklyn            4: 150-200    843
## 10 Brooklyn            5: 200 up    2212
## # ... with 15 more rows
```