Lab 1: Airbnbs in NYC

Alex Ptacek

Airbnb in NYC (or your city)

Airbnb has had a disruptive effect on the hotel, rental home, and vacation industry throughout the world. The success of Airbnb has not come without controversies, with critics arguing that Airbnb has adverse impacts on housing and rental prices and also on the daily lives of people living in neighborhoods where Airbnb is popular. This controversy has been particularly intense in NYC, where the debate been Airbnb proponents and detractors eventually led to the city imposing strong restrictions on the use of Airbnb. If you find this issue interesting and want to go deeper, there is the potential for an interesting project that brings in hotels (which have interesting regulations in NYC), hotel price data, and rental data and looks at these things together.

Because Airbnb listings are available online through their website and app, it is possible for us to acquire and visualize the impacts of Airbnb on different cities, including New York City. This is possible through the work of an organization called inside airbnb

Github Instructions

Before we introduce the data and the main assignment, let's begin with a few key steps to configure the file and create a github repository for your first assignment. This is optional but I think it is a good idea to start getting familiar with github tools. These steps come from the happygitwithr tutorial.

- Start a new github repository in your account, clone it to your computer (using RStudio to start a new project from a repository or any other way)
- Update the YAML, changing the author name to your name, and **Render** the document.
- Commit your changes with a meaningful commit message.
- Push your changes to GitHub.
- Go to your repo on GitHub and confirm that your changes are visible in your Qmd **and** md files. If anything is missing, commit and push again.

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualisation, and the **ggridges** package to make a ridge plot in the last exercise. You may need to install **ggridges** if you haven't already, you can do that using:

```
install.packages("ggridges")
```

Then make sure to load both packages:

Data

The data for this assignment can be found on my github page by clicking here and downloading nycbnb.csv

If you are adventurous and want to perform this assignment for a different city you can choose one from inside airbnb). If you go that route, make sure to download the file listings.csv.gz for the city you selected (gz is an archive format which you should be able to expand), and you will only need to keep the following columns for this analysis:

```
nycbnb = nycbnb |>
  select(
    id,
    price,
    neighbourhood_cleansed,
    neighbourhood_group_cleansed,
    accommodates,
    bathrooms,
    bedrooms,
    beds,
    review_scores_rating,
    number_of_reviews,
    listing_url )
```

You will also need to do some post-processing, including changing the price column from a string to a numerical variable. If you decide to go this custom route let me know and make sure to share your data, but I recommend sticking with the data I provided.

You can read the data into R using the command:

```
nycbnb = read_csv("/Users/alex/Downloads/nycbnb.csv")
```

where you should replace `/home/georgehagstrom/work/Teaching/DATA607/website/data/nycbnb.csv"nycbr` with the local path to your file.

Important note: It is generally not wise to include datasets in github repositories, especially if they are large and can change frequently.

You can view the dataset as a spreadsheet using the `View()` function. Note that you should not put this function in your R Markdown document, but instead type it directly in the Console, as it pops open a new window (and the concept of popping open a window in a static document doesn't really make sense…). When you run this in the console, you'll see the following **data viewer** window pop up.

Exercises

Preliminary Step 1. Load Packages

```
library(tidyverse)
library(ggridges)
```

Preliminary Step 2. Read csv file from local computer

```
nycbnb = read_csv("/Users/alex/Downloads/nycbnb.csv")
```

Problem 1. How many observations (rows) does the dataset have? Instead of hard coding the number in your answer, use inline code.

The nycbnb dataset has 37765 observations.

Problem 2. Run `View(nycbnb)` in your Console to view the data in the data viewer. What does each row in the dataset represent?

Each row in the data represents an observation.

Each column represents a variable. We can get a list of the variables in the data frame using the `names()` function.

```
names(nycbnb)
```

```
 [1] "id"                  "price"               "neighborhood"
 [4] "borough"             "accommodates"        "bathrooms"
 [7] "bedrooms"            "beds"                "review_scores_rating"
[10] "number_of_reviews"   "listing_url"
```

You can find descriptions of each of the variables in the help file for the dataset, which you can find online at the inside airbnb data dictionary

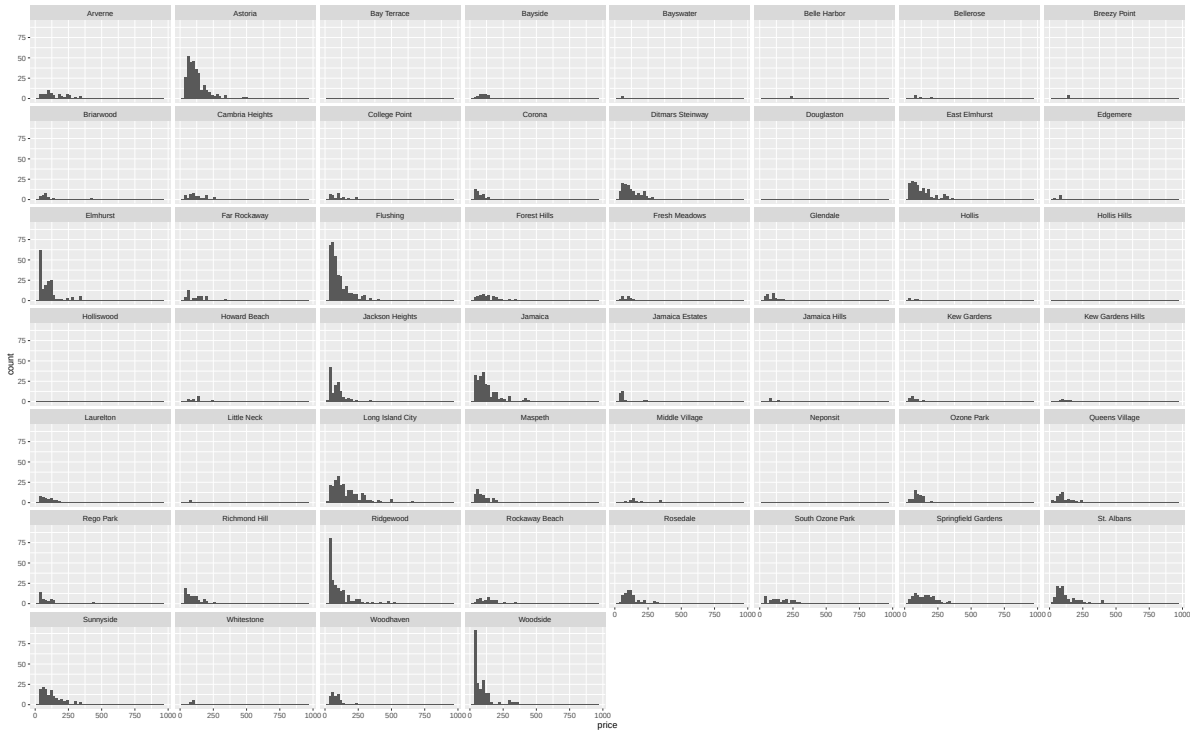
Problem 3. Pick one of the five boroughs of NYC (Manhattan, Queens, Brooklyn, the Bronx, or Staten Island), and create a faceted histogram where each facet represents a neighborhood in your chosen borough and displays the distribution of Airbnb prices in that neighborhood. Think critically about whether it makes more sense to stack the facets on top of each other in a column, lay them out in a row, or wrap them around. Along with your visualization, include your reasoning for the layout you chose for your facets.

1. Create a new dataset by filtering the 'borough' column.

```
queens <- nycbnb %>%
  filter(borough == "Queens")
```

2. Create a histogram to display price distribution for each neighborhood. It makes the most sense to wrap these graphs around, instead of laying them out in one row or column, because the large number of graphs is easier to view this way.

```
ggplot(queens, aes(x = price)) +
  geom_histogram(binwidth = 20) +
  facet_wrap(~neighborhood)
```



Problem 4. Use a single pipeline to identity the neighborhoods city-wide with the top five median listing prices that have a minimum of 50 listings. Then, in another pipeline filter the data for these five neighborhoods and make ridge plots of the distributions of listing prices in these five neighborhoods. In a third pipeline calculate the minimum, mean, median, standard deviation, IQR, and maximum listing price in each of these neighborhoods. Use the visualization and the summary statistics to describe the distribution of listing prices in the neighborhoods. (Your answer will include three pipelines, one of which ends in a visualization, and a narrative.)

1. Create a pivot table of NYC neighborhoods, filter to >50 listings, order by median price, and only keep the first 5 values, leaving us with the 5 highest median prices. Interestingly (but not too surprising), all of these neighborhoods are in Manhattan. This chunk also creates a new dataframe, 'top5', which will be used later. Note: null values must first be removed to prevent errors from executing summary functions on price.

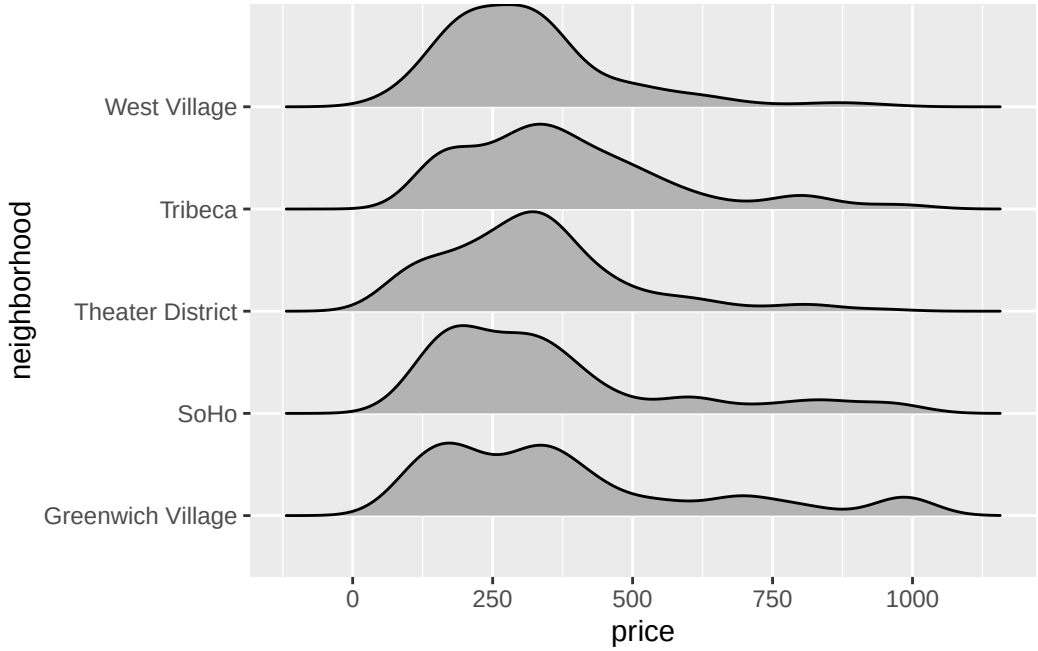
```
top5 <- nycbnb[complete.cases(nycbnb$price),] %>%
  group_by(neighborhood, borough) %>%
  summarize(count_by_hood = n(),
            median_price = median(price)) %>%
  filter(count_by_hood > 50) %>%
  arrange(desc(median_price)) %>%
```

```
  head(n=5) %>%
print()
```

```
# A tibble: 5 x 4
# Groups:   neighborhood [5]
  neighborhood      borough   count_by_hood median_price
  <chr>             <chr>             <int>        <dbl>
1 Tribeca           Manhattan           129         341
2 Greenwich Village Manhattan            81         330
3 Theater District  Manhattan           268         310
4 SoHo              Manhattan           117         299
5 West Village      Manhattan           232         292.
```

2. Create ridge plots of the top 5 neighborhoods by filtering 'nycbnb' to only neighborhoods found in 'top5'. By observing the plot, we can see that Tribeca and the Theater District have high percentages of their listings priced around $300, which seems to be the highest priced peak among the graphs. Greenwich Village appears to have the highest concentration of listings around the $1000 price and has possibly the highest deviation. The West Village's listing prices are the most evenly distributed.

```
nycbnb[complete.cases(nycbnb$price),] %>%
  filter(neighborhood %in% top5$neighborhood) %>%
  ggplot(aes(x = price, y = neighborhood)) +
    geom_density_ridges(scale = 1)
```



3. Create a pivot table with the top 5 median prices by doing the same filter as the previous step. Based on the summary statistics below, Greenwich Village has the highest average listing price and the highest deviation. Tribeca has the second-highest average listing price, but much lower deviation than Greenwich Village, and th highest minimum price for a listing at $150. The West Village has the lowest minimum price for a listing. the lowest average listing price, and the lowest deviation.

```
nycbnb[complete.cases(nycbnb$price),] %>%
  filter(neighborhood %in% top5$neighborhood) %>%
  group_by(neighborhood, borough) %>%
  summarise(min = min(price), mean = mean(price), median = median(price),
            st_dev = sd(price), IQR = IQR(price), max = max(price) )
```

```
# A tibble: 5 x 8
# Groups:   neighborhood [5]
  neighborhood      borough     min  mean median st_dev   IQR   max
  <chr>             <chr>     <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>
1 Greenwich Village Manhattan    73  385.    330    251.  285    999
2 SoHo              Manhattan    89  356.    299    223.  205    995
3 Theater District  Manhattan    57  321.    310    165.  175.   946
4 Tribeca           Manhattan   150  378.    341    188.  217    999
5 West Village      Manhattan    39  312.    292.   153.  163.   950
```

Problem 5. Create a visualization that will help you compare the distribution of review scores (`review_scores_rating`) across neighborhoods. You get to decide what type of visualization to create and which neighborhoods are most interesting to you, and there is more than one correct answer! In your answer, include a brief interpretation of how Airbnb guests rate properties in general and how the neighborhoods compare to each other in terms of their ratings.

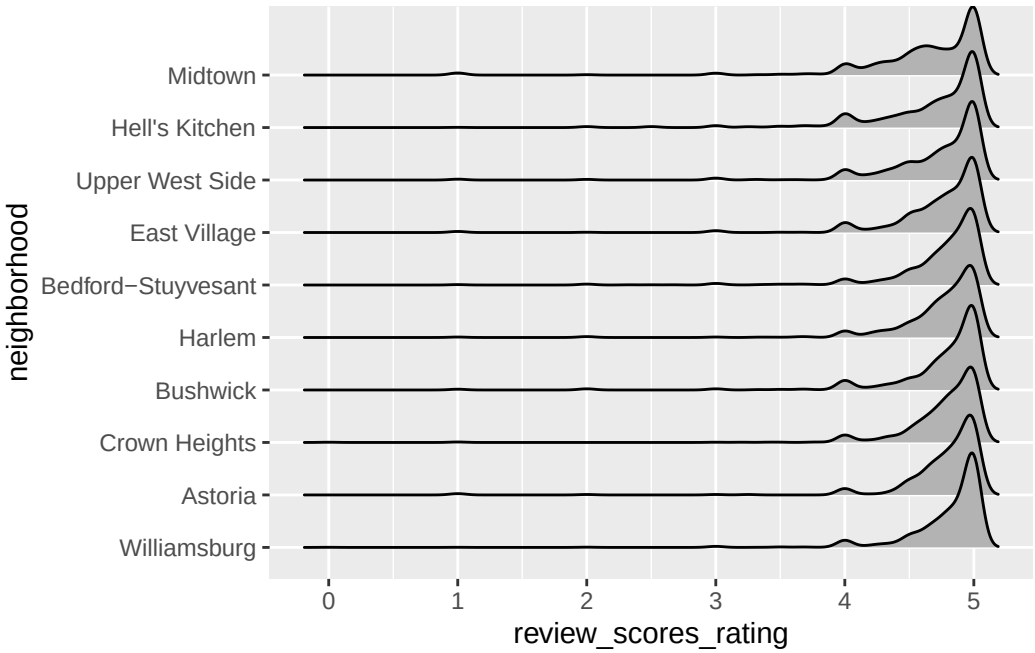
1. The neighborhoods with the highest number of reviews were chosen, in order to have the highest sample size possible to analyze review scores.

```
topreviews <- nycbnb[complete.cases(nycbnb$review_scores_rating),] %>%
  group_by(neighborhood, borough) %>%
  summarise(total_reviews = sum(number_of_reviews), avg_score = mean(review_scores_rating)) %
  arrange(desc(total_reviews)) %>%
  head(10) %>%
  arrange(desc(avg_score)) %>%
  print()
```

```
# A tibble: 10 x 4
# Groups:   neighborhood [10]
   neighborhood       borough   total_reviews avg_score
   <chr>              <chr>             <dbl>     <dbl>
 1 Williamsburg       Brooklyn          47619      4.79
 2 Astoria            Queens            20505      4.76
 3 Crown Heights      Brooklyn          31917      4.76
 4 Bushwick           Brooklyn          32009      4.74
 5 Harlem             Manhattan         56669      4.74
 6 Bedford-Stuyvesant Brooklyn          96650      4.73
 7 East Village       Manhattan         23023      4.72
 8 Upper West Side    Manhattan         21638      4.70
 9 Hell's Kitchen     Manhattan         31168      4.66
10 Midtown            Manhattan         28365      4.62
```

2. The graph below is ordered by average review score ascending from top to bottom. It becomes clear from looking at these graphs that airbnb reviewers tend to give 5 stars. The amount of 4 star reviews has the biggest effect on decreasing a listing's average score from 5 stars, because the second most noticeable bump for each graph (after 5 stars) is 4 stars.

```
nycbnb %>%
  filter(neighborhood %in% topreviews$neighborhood) %>%
  mutate(across(neighborhood, ~factor(., levels = c(topreviews$neighborhood)))) %>%
  ggplot(aes(x = review_scores_rating, y = neighborhood)) +
  geom_density_ridges()
```


neighborhood
Midtown
Hell's Kitchen
Upper West Side
East Village
Bedford−Stuyvesant
Harlem
Bushwick
Crown Heights
Astoria
Williamsburg
0
1
2
3
4
5
review_scores_rating