Lab1 - Airbnbs in NYC

Yana Rabkova

```
#loading packages
library(readr)
library(ggridges)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v purrr     1.0.2
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
nycbnb = read_csv("nycbnb.csv")
```

```
Rows: 37765 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): neighborhood, borough, listing_url
dbl (8): id, price, accommodates, bathrooms, bedrooms, beds, review_scores_r...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
nrows <- nycbnb %>% nrow()
print(paste("The dataset contains",nrows,"observations"))
```

```
[1] "The dataset contains 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 column represents a variable. We can get a list of the variables in the data frame using the `names()` function.

```
names(nycbnb)
```

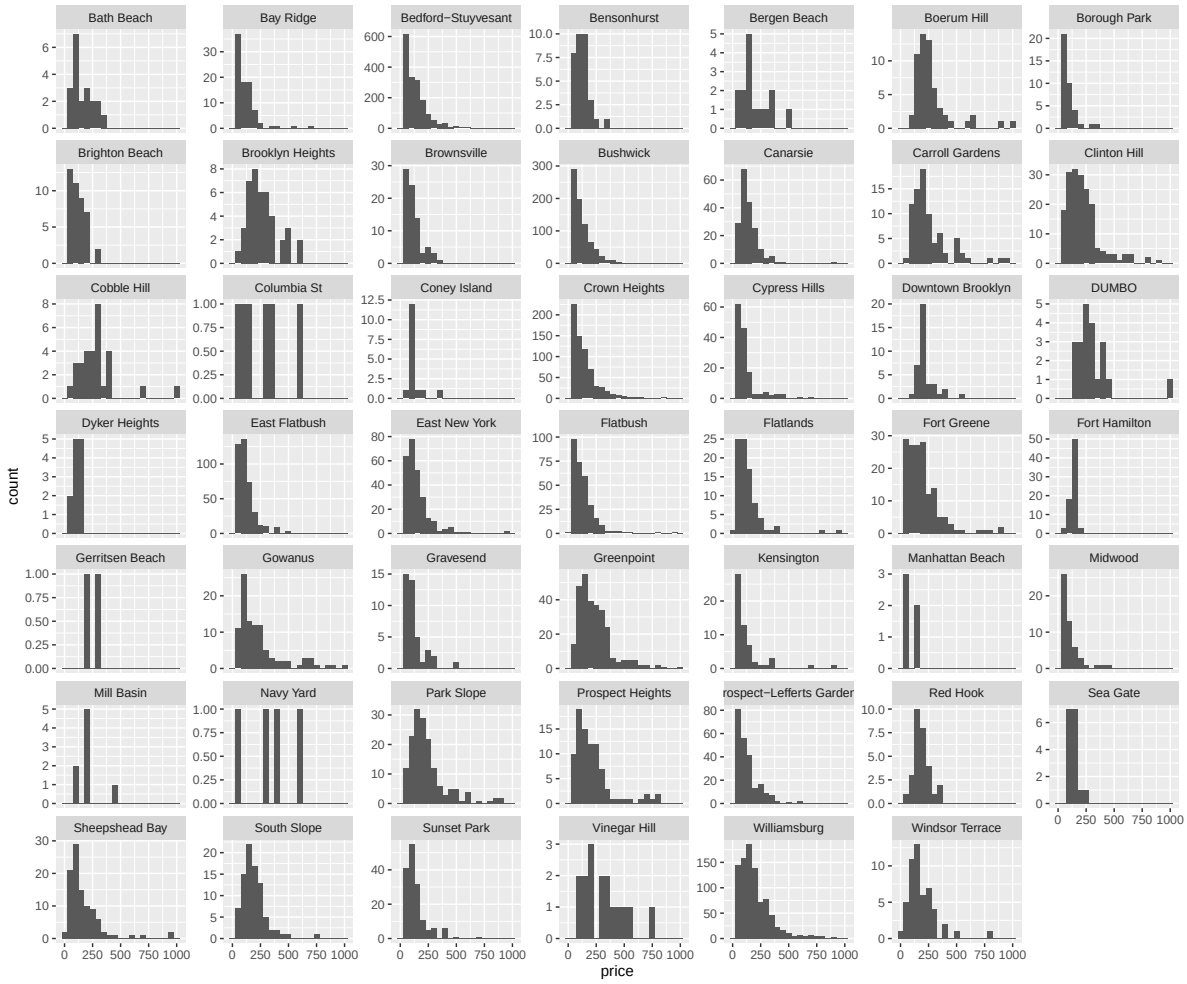
```
 [1] "id"                  "price"               "neighborhood"
 [4] "borough"             "accommodates"        "bathrooms"
 [7] "bedrooms"            "beds"                "review_scores_rating"
[10] "number_of_reviews"   "listing_url"
```

Each row represents an individual listing.

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.

```
brooklynbnb <- nycbnb %>% filter(borough == "Brooklyn")
ggplot(data = brooklynbnb,
       aes(x = price)) +
  geom_histogram(binwidth = 50)+
  facet_wrap(~ neighborhood, scales = "free_y")
```

```
Warning: Removed 5928 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Since there are a lot of neighborhoods I have decided to wrap them around which I believe makes it easier for comparison.

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.)

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.