Project Proposal

The Four-mula: Hellen Han, Lila Rogers, Amy Duan, Eva Aggarwal

```
library(tidyverse)
library(tidymodels)

data <- read_csv("data/AB_NYC_2019.csv")
```

Introduction - Lila

- An introduction to the subject matter you're investigating (citing any relevant literature)
- Statement of a well-developed research question.
- The motivation for your research question and why it is important
- Your team's hypotheses regarding the research question
  - This is a narrative about what you think regarding the research question, not formal statistical hypotheses.Data description

Exploratory data analysis - Eva

- The source of the data set
- A description of when and how the data were originally collected (by the original data curator, not necessarily how you found the data)
- A description of the observations and general characteristics being measured

Analysis approach

- Description of data processing you need to do to prepare for analysis, such as joining multiple data sets, handling missing data, etc.

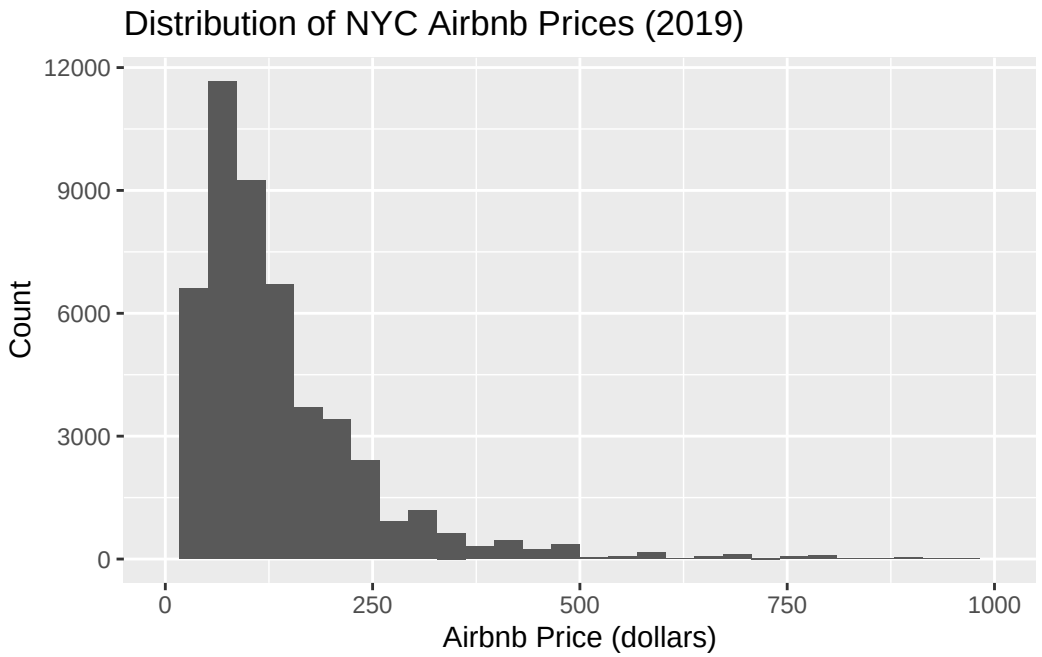
For the following New York City Airbnb dataset, to prepare the dataset for analysis, first, we need to consider missing data values within the dataset, eliminating observations that are incomplete across our predictor/response variables. The majority of these missing values appear in the last_review and reviews_per_month variables. Furthermore, we need to check for data inconsistencies in the quantitative variables by analyzing the summary statistics for the predictor variables and note whether there are extreme outliers that are improbable (ex. Airbnb prices of 0, unreasonably long minimum night Airbnb rentals). In addition, our categorical predictors, such as neighbourhood_group and room_type need to be converted to factors to allow us to use them in our regression analysis. Lastly, since our dataset describes all NYC Airbnb listings in 2019, there is a possibility of hosts with multiple Airbnb listings, and thus due to host dependency, we need to create a unique set of Airbnb listings by host_id to analyze the relationship between price and possible predictor variables.

- Visualizations, summary statistics, and narrative to describe the distribution of the **price** variable.

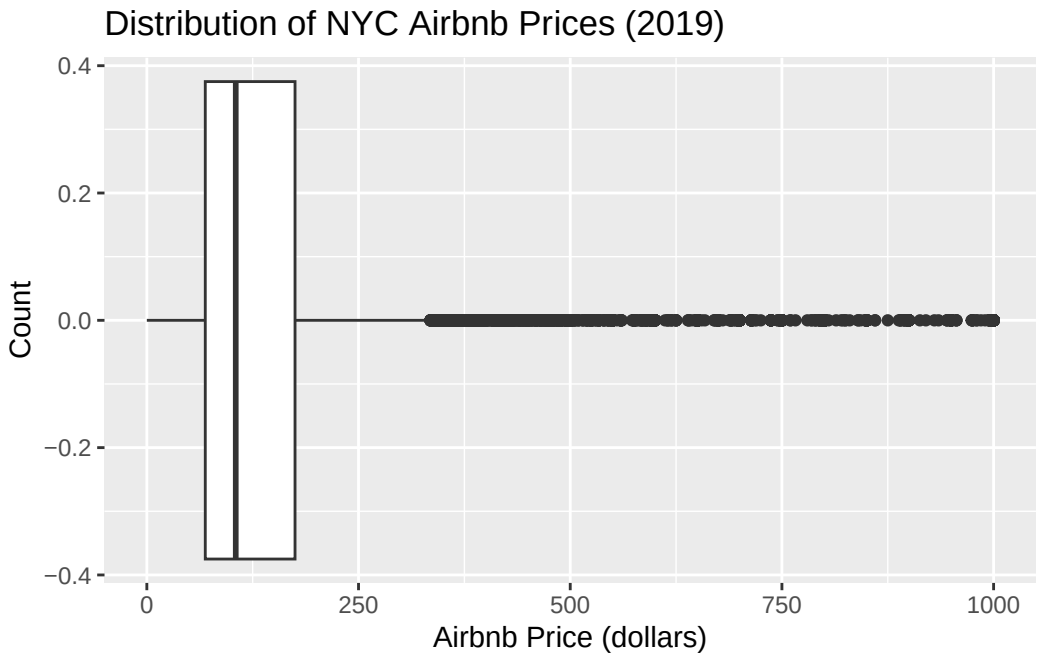
```
summary(data$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    69.0   106.0   152.7   175.0 10000.0
```

```
data %>%
  ggplot(mapping = aes(x = price)) +
  geom_histogram() +
  labs(title = "Distribution of NYC Airbnb Prices (2019)",
       x = "Airbnb Price (dollars)",
       y = "Count") +
  scale_x_continuous(limits = c(0, 1000))
```




```
data %>%
  ggplot(mapping = aes(x = price)) +
  geom_boxplot() +
  labs(title = "Distribution of NYC Airbnb Prices (2019)",
       x = "Airbnb Price (dollars)",
       y = "Count") +
  scale_x_continuous(limits = c(0, 1000))
```



The visualizations describe that the distribution of NYC Airbnb prices in 2019 is heavily right-skewed and unimodal, indicating that although most NYC Airbnb prices are typically priced below 200 dollars, there are various outliers in Airbnb prices that are significantly more expensive. For the purpose of creating meaningful visualizations, we omitted Airbnb prices that are greater than 1000 in the figures. Furthermore, the distribution of Airbnb prices has a center of approximately 106.0 dollars, described by the median, and a spread of approximately 106.0 dollars, described by the interquartile range. There are multiple outliers in the dataset for prices greater than approximately 350 dollars, as shown in the boxplot visualization, as more high-end NYC Airbnbs have a significantly greater price. The majority of NYC Airbnb prices tend to be between approximately 50 and 200 dollars.

Data dictionary - Hellen

The data dictionary can be found here [Update the link and remove this note!]