Project Proposal

Team lol - Tamsin Connerly, Hannah Lee, Jasmine Xiang

```
library(tidyverse)
library(tidymodels)

airbnb <- read.csv("data/NYC-Airbnb-2023.csv")
```

Introduction

The rise of short-term rental platforms, particularly Airbnb, has significantly disrupted the traditional hospitality industry and transformed urban housing markets worldwide. In New York City, one of the world's most popular tourist destinations, the impact of Airbnb has been particularly pronounced, raising questions about its effects on local communities, housing affordability, and the broader urban economy.

Previous research has identified several factors that impact Airbnb pricing. Wang and Nicolau (2017) found that host attributes, site and property attributes, amenities and services, rental rules, and online review ratings all play significant roles in determining listing prices (Wang and Nicolau 2017). Furthermore, recent studies have provided evidence of Airbnb's influence on housing markets. Barron et al. (2020) found that a 1% increase in Airbnb listings leads to a 0.018% increase in rents and a 0.026% increase in house prices(Barron, Kung, and Proserpio 2018). This effect is more pronounced in areas with a lower share of owner-occupiers, suggesting that non-owner-occupiers are more likely to reallocate their properties from long-term to short-term rentals.

Our research question is: "How do various factors, such as bedroom/bathroom number, accommodation capacity, room type, review scores, and distance from city center, influence the price of Airbnb listings in New York City?"

Understanding the determinants of Airbnb pricing in New York City is crucial for several reasons. Firstly, it can provide valuable insights for policymakers grappling with the challenges posed by the growth of short-term rentals, including potential impacts on housing affordability and neighborhood character(Toader et al. 2021). Secondly, it can help hosts make more informed pricing decisions, potentially leading to more efficient market outcomes.

Based on existing literature and our understanding of the New York City housing market, we hypothesize that:

- Listings with more bedrooms and bathrooms will command higher prices, reflecting the premium placed on space in urban environments.
- The number of guests a property can accommodate will positively correlate with price, as larger groups often have higher budgets.
- The type of room (entire home/apartment vs. private room) will significantly impact pricing, with entire homes/apartments having a higher price.
- Higher review scores will be associated with higher prices, as positive feedback may justify premium pricing.
- Properties closer to the city center will be priced higher due to their convenient location and proximity to attractions.
- The impact of these factors on price may vary across different boroughs or neighborhoods.

Data description

The Airbnb dataset that we are utilizing can be found on Kaggle, but the author of the dataset obtained the data from Inside Airbnb (https://insideairbnb.com/). Inside Airbnb has collected data on dozens of countries and cities, but we decided to focus on New York City. The data was sourced from publicly available data on the Airbnb website in 2023.

Each row in the dataset represents a unique Airbnb listing in New York City. Each of these correspond to individual properties available for rental on the platform and have many (18) variables such as name of the listing, latitude and longitude, room type, price, minimum number of nights required for booking, total number of reviews the listing has reviewed, and more. We are particularly interested in the following explanatory variables that we believe could impact the price of an Airbnb listing:

- Number of bedrooms and bathrooms
  - These could give insights into the size and comfort level of the Airbnb, likely affecting the price.
- Accommodation capacity (number of guests)
  - This could influence pricing as larger accommodations can host more guests.
- Room type

  - Type of room (whether the listing is an entire home/apartment or a private room) can impact pricing

- Review scores (overall rating)
  - Listings with better or higher number of reviews could be priced higher because of higher perceived value and trustworthiness

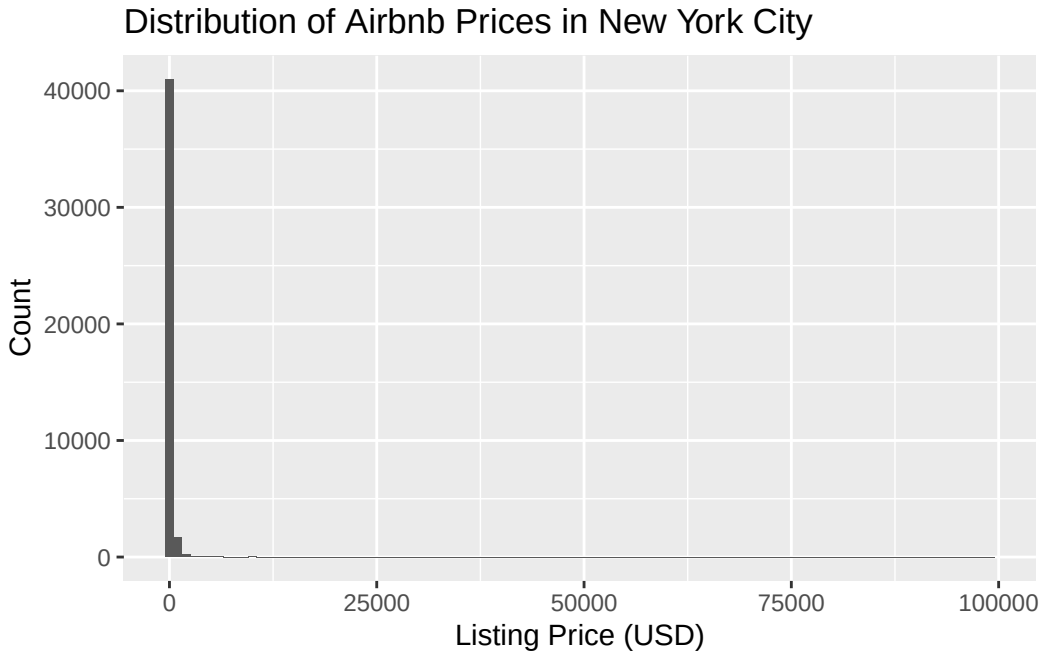
- Distance from city center
  - Proximity to central locations can impact pricing (since it is more convenient and desirable)

Exploratory data analysis

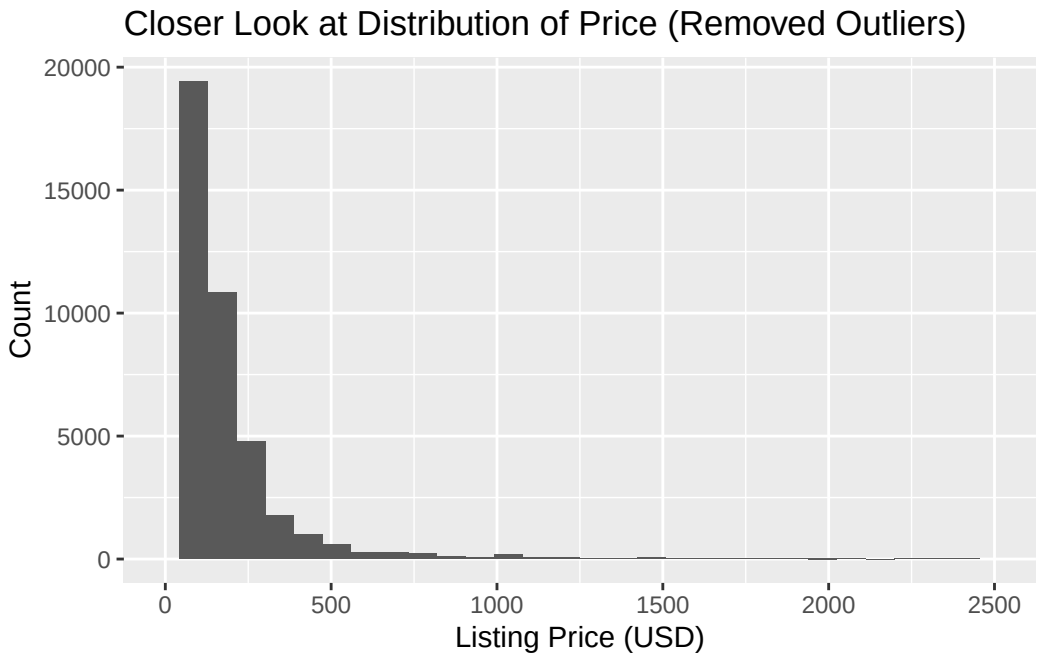
For data processing, we will look into the number of NaN values for price in the data set. If there are not that many, then we will drop the listings with them, but if needed, we will impute the missing data for numerical columns. We will also look to see if there are any extreme outliers present, and remove them if there are any. Additionally, since room_type is a categorical variable, we plan to utilize one hot encoding so that it is in numerical form that we can use as part of our analysis.

```
# histogram looks small because of the one outlier at 99k

ggplot(data = airbnb, aes(x = price)) +
  geom_histogram(bins = 100) +
  labs(title = "Distribution of Airbnb Prices in New York City", x = "Listing Price (USD)", y
```




```
# set x axis ranges to get closer look at dist
ggplot(airbnb, aes(x = price)) +
  geom_histogram(bins = 30) +
  labs(title = "Closer Look at Distribution of Price (Removed Outliers)", x = "Listing Price
  xlim(0, 2500)
```




```
summary_stats <- summary(airbnb$price)

summary_tibble <- tibble(
  Statistic = names(summary_stats),
  Value = summary_stats
)


summary_tibble
```

```
# A tibble: 6 x 2
  Statistic Value
  <chr>     <table>
1 Min.          0.0000
2 1st Qu.      75.0000
3 Median      125.0000
4 Mean        200.3072
5 3rd Qu.     200.0000
6 Max.      99000.0000
```

The distribution is pretty heavily right skewed, as can be seen from both of the histograms. It is difficult to analyze the distribution in the first histogram because there is an outlier at

$99,000 and makes the bins and binwidth very narrow and zoomed out (since the range of the data is too large). It is also clear that this outlier impacts the mean, since the median of $125 is quite a bit less than the mean of $200.37, and the mean is roughly equal to the 3rd quartile which is also around $200. We plan to to remove this outlier as a result when doing our analysis, and we will go into more depth later and check for additional outliers.

From the initial histogram, since the majority of the listings seem to fall between 0 and 2500 dollars, we visualized the distribution of the listings between this price range to get a better view of it. We can see that the distribution is still right skewed, and the vast majority of the listings seem to cost between $50-$200. There are a few more points visible that were not in the initial histogram that appear to be quite far from where the majority of the listings are clustered, so we will go back and check to see if they are greater than 1.5 x IQR and can be classified as outliers.

Analysis approach

Potential Predictor Variables of Interest:

Our potential predictors capture many key aspects of Airbnb listings — space, amenities, perceived quality, and location—that are likely to have significant influence on nightly rates.

- Number of Bedrooms (quantitative): Reflects the amount of sleeping space in a listing. We anticipate more bedrooms generally commanding higher prices.
- Number of Bathrooms (quantitative): Indicates the level of comfort/convenience. Listings with more bathrooms are typically priced higher.
- Accommodation Capacity (quantitative): Represents how many guests can stay in the listing. We expect larger capacity to have higher nightly rates.
- Room Type (categorical): Classified primarily as "Entire home/apt," "Private room," or "Shared room." We plan to use one-hot encoding, as room type strongly affects price (entire units usually cost more).
- Review Scores (quantitative): Represents overall rating (on a typical 1–5 scale, or 1–100 in some datasets). Higher-rated listings may set premium prices.
- Distance from City Center (quantitative): Measures proximity to major landmarks or downtown areas. We expect closer listings to charge higher prices.

Regression model technique:

Our responsible variable is price, which is a continuous variable, and we will employ a Multiple Linear Regression model to estimate the effect of each predictor on price while controlling for the others.

Model specification:

$$Price = \beta_0 + \beta_1(Bedrooms) + \beta_2(Bathrooms) + ... + \epsilon$$

Assumption checking:

- Linearity: The relationship between predictors and our response variable price should be approximately linear.
- **I**ndependence of Residuals: We assume each listing is an independent observation.
- Homoscedasticity: Residuals should have constant variance across all fitted values.
- Normality of Residuals: Residuals should be roughly normally distributed.

Model evaluation:

In evaluating our MLR model, we will begin by examining the R^2 and Adjusted R^2 value. They can help us understand how well the model explains the variability in price while adjusting for the number of predictors used. We will also look at p-values for each predictor, which will indicate which variables have a significant impact on the listing's price.

Feature engineering:

Depending on the distribution of our predictors and response variable, we may consider examining the potential for interaction terms, particularly if we hypothesize that the effect of one variable depends on another. For example, the impact of accommodating additional guests may differ significantly if the listing is an entire home versus a private room.

In addition, we plan to carefully encode categorical variables, such as Room Type, using dummy variables. This will allow the model to compare and contrast distinct categories effectively. We plan to capture the detailed relationships among our features, thus producing more robust, interpretable insights into Airbnb pricing.

Data dictionary

The data dictionary can be found here.

Barron, Kyle, Edward Kung, and Davide Proserpio. 2018. "The Sharing Economy and Housing Affordability: Evidence from Airbnb." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3006832.

Toader, Valentin, Adina Letiţia Negruşa, Oana Ruxandra Bode, and Rozalia Veronica Rus. 2021. "Analysis of Price Determinants in the Case of Airbnb Listings." *Economic Research-Ekonomska Istraživanja* 35 (1): 2493–2509. https://doi.org/10.1080/1331677x.2021.1962380.

Wang, Dan, and Juan L. Nicolau. 2017. "Price Determinants of Sharing Economy Based Accommodation Rental: A Study of Listings from 33 Cities on Airbnb.com." *International Journal of Hospitality Management* 62 (April): 120–31. https://doi.org/10.1016/j.ijhm.2016.12.007.