Stat 405 Project Proposal

Title: Investigating Factors Influencing Airbnb Listing Prices
Group members: Jon Karanezi, Jingzhe Zhang, Jacob Stugelmeyer, Max Gehred, Eigene Ohba

Airbnb has become a significant player in the accommodation industry, offering a wide range of listings across various locations. Understanding the factors that influence Airbnb listing prices is crucial for both hosts and guests. In this project, we aim to explore the determinants of Airbnb listing prices.

Dataset:
Our dataset is homestay listings on airbnb for different cities. Each city has an individual dataset in the format of csv.

Useful variables:
Host_since: date that the host of the listing first joined Airbnb
Host_location: city and country that the host of the listing is from (might be different from listing itself)
Host_is_superhost: True or false. Whether the host is classified as superhost by airbnb.
Latitude and **longitude**: geological location of the listing, which enables us to visualize the location of listing on map
Room_type: whether the listing is private room or entire home/apartment
Price: price per night in US dollar
Minimum_minimum_nights: minimum number of nights to book the listing
Accommodates: number of guests that the listing can hold
Bathrooms: number of bathrooms in the listing
Beds: number of beds in the listing
Review_scores_rating: the average review score out of 5
City (added by us): city that the listing is located. (same for listings in same csv file)

Computation steps:

1. Cleaning the dataset by handling missing values, outliers, and data inconsistencies.
2. train regression models on the large Airbnb dataset. Parallel processing techniques will be employed to expedite model training and evaluation.
3. Assessing model performance using R-squared. Cross-validation will be conducted to validate the models' generalizability. Change variables used if necessary.
4. Interpreting model coefficients and feature importance scores to understand the impact of different variables on listing prices.

Link to Github repository: https://github.com/Joni003/Project_405

Bash code to download all individual datasets:

```
#!/bin/bash

# Get the webpage
wget https://insideairbnb.com/get-the-data
mv get-the-data input.html

# Remove any preexisting data directory
rm -rf data/

# Create a data directory
mkdir -p data

# Extract hyperlinks associated with listings.csv.gz, calendar.csv.gz, reviews.csv.gz
grep -Eo 'href="([^"]*\/(listings|calendar|reviews)\.csv\.gz)"' input.html | \
awk -F '/' '{print $4,$5,$6}' | \
awk -F "" '{print $1 " " $2}' | \
sort -u > data/locations.txt

# Read each line in locations.txt to create directories
while IFS= read -r location; do
    # Extract country, region, and city names from the location
    country=$(echo "$location" | awk '{print $1}')
    region=$(echo "$location" | awk '{print $2}')
    city=$(echo "$location" | awk '{print $3}')

    # Create directory if it doesn't exist
    mkdir -p "data/$country/$region/$city"
done < data/locations.txt

# Extract hyperlinks associated with listings.csv.gz, calendar.csv.gz, reviews.csv.gz
grep -Eo 'href="([^"]*\/(listings|calendar|reviews)\.csv\.gz)"' input.html | \
awk -F "" '{print $2}' > data/file.txt

# Read each line in file.txt to download files
while IFS= read -r line; do
    # Extract country, region, and city names from the URL
    country=$(echo "$line" | awk -F '/' '{print $4}')
```

```
    region=$(echo "$line" | awk -F '/' '{print $5}')
    city=$(echo "$line" | awk -F '/' '{print $6}')

    # Download the file using wget
    wget "$line" -P "data/$country/$region/$city"
done < data/file.txt
```