Springboard Capstone Project #1
Data Wrangling

For this project, I found all the datasets on insideairbnb.com. This website “is an independent, non-commercial set of tools and data that allows you to explore how Airbnb is really being used in cities around the world.” Since I am focusing on the Airbnb market in London (the second largest Airbnb city outside the US), I found the following csv files:

- listings.csv
  - information and metrics for listings in London
- reviews.csv
  - reviews left by the guests on the hosts’ listings
- neighbourhoods.csv
  - Lists all neighborhoods and neighborhood groups in London

I read these csv files into DataFrames using pandas.read_csv().

The listings data (listings.csv) contains 80,767 rows and 16 columns. Each row represents a unique listing in London. The columns give us information such as the host’s name, the listing’s location, room type, price and etc. When we get the info, we see that not every listing receives reviews since there are only 60,194 non-null values in the columns last_review and reviews_per_ month. Despite not having null values on the number_of_reviews column, we see that the minimum value is 0.

We get a better look at these null values when we study the reviews dataset (review.csv). It contains 1,249,466 rows and 6 columns--where each row represents a comment written on a specific listing’s page. When we apply .info() to the DataFrame, we see that there are 407 null values (i.e. empty reviews) on the “comments” column. Since this is only .03% of the reviews, we can drop these rows using .dropna().

Despite having 1,249,466 rows in the reviews data, there are only 59,270 unique listings in the dataset. This means that 21,497 listings (i.e. 26.62% of the listings in London) did not have any reviews. Not having reviews could mean two things--the listing is either not active or has just simply never cancelled on anyone.

Our outliers would be the inactive listings since giving them a zero probability of cancelling simply because they haven’t been active would be biased. In this study, we will define “inactive” listings as those who have 0 bookings for the year, i.e. where listings[‘availability_365’]==365. There are 2,693 of these listings. We then drop these rows from our listings data, leaving us with 78,074 rows.

The neighborhood data is simple and just gives us the 33 neighborhoods in London. There are no null values or outliers.