Capstone Project 1: New York City AirBnb Data Wrangling

1. Data Collection and Importing

The dataset was loaded in as a .csv file. This dataset is publicly accessible and was found from this website. The file contains summary information and metrics for listings in New York City. What is good about the data is that it is good for exploratory data analysis, visualization, and prediction models. After loading the data into the dataframe, the size was verified to have **48,895 rows** and **16 columns**.

2. Data Cleaning

Group the columns into column categories for a better understanding of the dataset

Host descriptors:

- host_id: host ID
- host_name: host name
- calculated_host_listings_count

Listing descriptors:

- id: ID
- name: name of listing
- room_type: listing space type
- minimum_nights: amount of nights minimum
- availability_365: number of days when listing is available for booking
- price: price in dollars

Review descriptors:

- number_of_reviews: number of reviews
- last_review: latest review
- Reviews_per_month: number of reviews per month

Location descriptors:

- neighbourhood_group: location
- neighbourhood: area
- latitude: latitude coordinates

- longitude: longitude coordinates

Before I dive into the data, it is worth considering situational factors in what drives the listing prices in this dataset. For example, stays during a major holiday possibly being more expensive than non-holiday stays.

3. Missing Values

Replacing missing values

Listings without any reviews have missing values for last_review and reviews_per_month. For these listings, missing values will be replaced by 0. There are some listings that are missing a name or where the host name is missing. These will be replaced with none.

4. Outliers

The one major outlier I have seen so far in the dataset is a stay that is over 1000 nights. If I set the duration of stay limits to not exceed 31 days, this outlier will not skew the data too much.