Nirav Bariya

Achievement 6

Exercise 6.1

Date: 04/28/2024

## 1. Summary of Airbnb listings for Toronto

**Data Source:** This data source is external. It was collected by insideairbnb.com and not published by Airbnb itself. The website however is not endorsing nor associated with Airbnb or its’ competition. The website claims that the data compiled is for public analysis, discussion, and community website. In my opinion, I think we can consider this data set as trustworthy since it is collected from Airbnb and due transparency is shown by Inside Airbnb.

**Collection Method:** This dataset can be considered as usage data as this can be utilized for improving pricing of Airbnb listings. The collection method used is web scrapping as the data is collected from the Airbnb website. This is an automatic method. The data set is relatively new as it includes listings from the month of February 2024. Therefore, the time lag is minimal in this case.

**Data Content:** The data set has 75 features including id, listing\_url, last\_scrapped, bathrooms, accommodates, bedrooms, price, minimum\_nights and so on. It also information about reviews including features such as review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, and review\_scores\_checkin.

**Data Relevance:** This data is the primary data set that I will be working with as it not only has information about pricing and features of the house, but it also has information about location and reviews.

**Why this Data set?**

This data set is one of the new data sets in the public domain and is relevant to the country and locality in which I am currently residing. It provides great value as decisions about how to price an Airbnb listing can be made. Moreover, the factors which impact success of the Airbnb listing can also be studied. Lastly, this data set gives ample opportunity to learn machine learning algorithms including regression and classification. It also provides time series data that may be used for time series analysis.

## 2. Data Profile

**Cleaning the Data:**

1. Missing Data:

A screenshot of a computer

Description automatically generated

**Results:**

A close-up of a text

Description automatically generated

2. Handling Duplicates:



3. Checking for mixed type data:

A screenshot of a computer code

Description automatically generated

The columns were made consistent in the Jupyter notebook.

**Data Summary:**

A screenshot of a computer

Description automatically generated

A table of data types

Description automatically generated

**Limitations and ethics:**

The data is collected from the Airbnb website itself using. Web scrapping may be considered unethical and falls under grey area. The Inside Airbnb website states that no private information is used as the Name, photographs, and listings are available publicly. There is a high degree of transparency about the way the data is collected. Some of the variables are calculated by Inside Airbnb. These may not be used in the analysis and hence is not of concern to us.

Steps have been taken to anonymize location information by Airbnb itself. Considering all these, I think this data set offers great value in terms of learning and can be considered unbiased.

**Questions to explore:**

In a third section of your project document, define a list of questions to explore with your analysis.

1. What is the distribution of super host?
2. How are the listings distributed in Toronto? Are certain neighbourhoods have more listings than others?
3. What is the distribution of property type and room type in Toronto?
4. What are the common amenities offered in the listings?
5. Is there any relationship between bedrooms and price?
6. Are certain neighbourhoods priced more than others?
7. Is there any relationship between property type and price?
8. Is there a relationship between accommodations and price?
9. Which property types have higher price?
10. Is there a certain relationship between number of reviews and price?
11. Do customers prefer lower minimum nights or higher minimum nights?
12. Is there a relationship between price and average number of reviews per month?