***Exploration of Data***

After the first segment, the team decided to use a different, but similar, dataset. We switched to a dataset found on Kaggle, which was initially published on the Inside Airbnb website. The raw dataset was composed of 106 columns of different features, and 8111 rows of detailed Airbnb listings in San Francisco.

In order to understand what statistics were available within the dataset, as well as identifying which values would be ideal for the Machine Learning model, the team used Excel initially, to skim through the dataset and identify which characteristics of the dataset were preferred as well as locating null values. Jupyter Notebook was used to clean the dataset. The **listings.csv** file was read using Pandas and converted into a dataframe. We started by dropping the columns which did not contain any data. We then dropped columns we found irrelevant for our purpose of analysis. After which, we determined the columns and rows that had null values, and decided to handle those accordingly, by either dropping rows with categorical variables, or assigning 0 (zero) to numerical variables. Several columns were dropped because all rows had the same values (i.e. state & country). The last step was to convert data types as appropriate. A few columns were renamed accordingly, and an additional “total” column was added to the dataframe, which aggregated the “price” and “cleaning fee” columns. After the dataset was cleaned, there were 19 columns and 7815 rows. We saved it as a new csv file called **cleaned\_airbnb\_dataset2.csv**.

We start with looking at how many listings in each neighborhood. The top 3 neighborhoods with the most listings are (1) Mission – 744 listings, (2) Downtown/Civic Center – 660 listings, and (3) South of Market – 627 listings. Western Addition is not far behind with 595 listings.

Chart, bar chart, histogram

Description automatically generated

Chart, bubble chart

Description automatically generated

The most popular property types are apartments, condominiums and houses.

Chart, histogram

Description automatically generated

Chart, treemap chart

Description automatically generated

There are 4 room types, Entire home/apartment, Hotel room, Private room and Shared room, of which Entire home/apt has the most listings.Chart, bar chart

Description automatically generated

***Initial (Exploratory) Data Analysis***

Using the .describe() method, and the sns.distplot() function, we can see that the prices (total) are skewed to the right. This is probably due to some outliers, listings with very high prices.

Table

Description automatically generated

Chart

Description automatically generated

The following graphs shows the relationship of price (total) vs Property Type.

Chart, bar chart

Description automatically generated

A picture containing chart

Description automatically generated

The following graphs shows the relationship of price (total) vs the number of people the listing accommodates, the number of bathrooms and the number of bedrooms.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart

Description automatically generated

***Machine Learning Exploration & Analysis***

The data preprocessing began with a cleaned csv with null values filled in and columns dropped so only the necessary information would be presented. In order to continue with a Random Forest model, some values needed to be converted through Label Encoder.

We could then use all the values in this manipulated dataset to train a model. Features of the homes were split, and the values trained with the Train-Test split method. The trained values could make predictions partially using Linear Regression and a correlation matrix. The fitted data could also be used to find the relationship between each feature and the total price. Chart, bar chart, histogram

Description automatically generated

By  finding the feature vs. total relationship, the user could predict their price changes based on their home features. We chose to display these features in a heatmap to see where the correlations lay.

A picture containing text, screenshot

Description automatically generated

A picture containing graphical user interface

Description automatically generated

It also gives insight on how an Airbnb home could evolve their pricing in the future. For example, as an Airbnb property owner, we could make improvements to accommodate more people and therefore raise the home’s rental price and make a higher profit. Random Forest could potentially be limiting as it cannot make predictions outside of what is in the dataset. It wouldn’t be able to extrapolate from the values it is fed. A benefit is that it is very straightforward. We can see exactly which features are the most important predictors and create conclusions from this.