Capstone project Data Wrangling Steps

1. **Data Acquisition and Cleaning**
   1. The Airbnb Data was obtained from a Kaggle Dataset at the following address

<https://www.kaggle.com/navaneesh/airbnb#train.csv>

* 1. There isn’t a need for additional data mining or web scraping.
  2. The data provided is in the form of CSV files and are listed below:
     1. **train.csv**​: The training set of users
     2. **test.csv**​: The test set of users.
  3. We Drop the ‘**thumbnail\_url** ‘ as it doesn’t seem relevant.
  4. We will do some basic modifications like converting the string(“t”/”f”) categorical column to numerical (0 & 1) and other categorical columns into numerical types
  5. Further we will calculate Missing values and handle them as described in next section
  6. Finally, we will see any outliers if any.
  7. No additional cleaning steps required at this stage

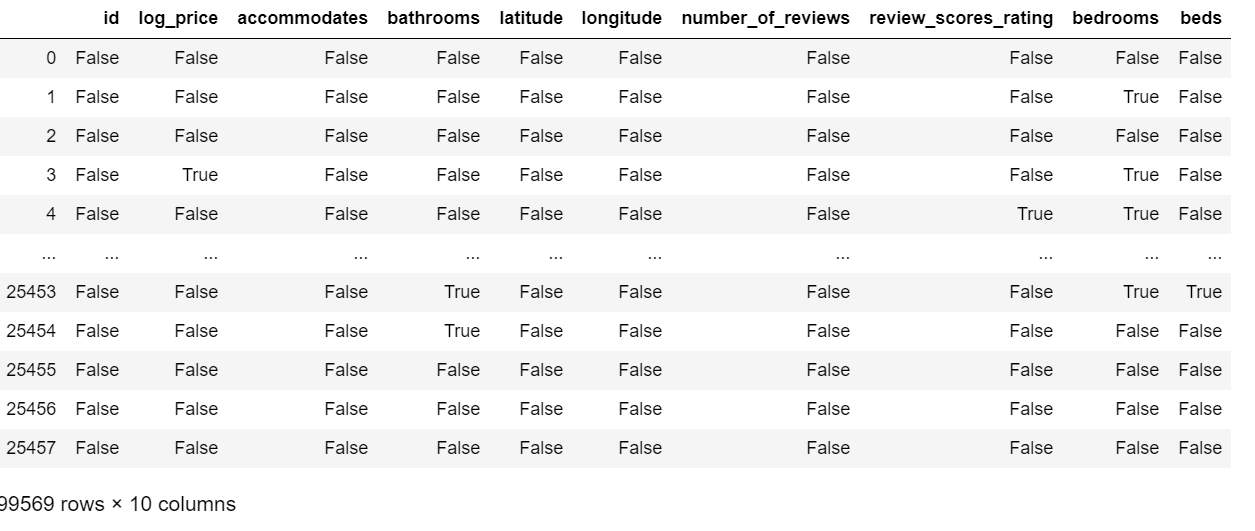
1. **Missing Values** 
   1. We calculate the percentage of missing values of every column. Below is the result of the same

A screenshot of a cell phone

Description automatically generated

* 1. We will fill all the NAN values (numerical columns as 0)

1. **Outliers**
   1. We visualise the outlier for numerical columns using box plot as well as see outlier rows wth IQR method (<Q1 - 1.5 \* IQR & > Q3 +1.5 \* IQR)
   2. This is only for visualization and checking. We may keep the dataset as is t check effects of model later on.
   3. Below are the sample rows with outliers marked. We will remove them later from the dataset on analysis with the algorithm

****

* 1. Following method used to identify the outlier rows

**A screenshot of a cell phone

Description automatically generated**