**Capstone Project Submission**

**Instructions:**

i) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| RAHUL SINGH WALDIA  [rahulwaldia2@gmail.com](mailto:rahulwaldia2@gmail.com)   1. Acquire and loading data 2. Cleaning dataset 3. Exploring and Visualizing Data |
| **Please paste the GitHub Repo link.** |
| Github Link:- https://github.com/RahulSinghWaldia/Airbnb-data-analysis.git |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |
| Since 2008, guests and hosts have used Airbnb to expand on traveling possibilities and present a more unique, personalized way of experiencing the world. Today, Airbnb became one of a kind service that is used and recognized by the whole world. Data analysis on millions of listings provided through Airbnb is a crucial factor for the company. These millions of listings generate a lot of data — data that can be analyzed.For this project, I am using google collab with a python programming language to write our script. Google collab is a inbuilt software application used for software development.To get the data, i am using Airbnb data that that is shared by the almabetter. Before we are able to load the data into our IDE, first we need to import various external libraries/modules that needed for visualization and analysis.After loading the dataset, we need to understand the dataset by using various techniques. First, we need to look for information on how big is our dataset. By using **shape** attributes, we get to know our data size from a number of rows which consist of listing index, and the number of columns with the content of every features related to the index.Then we check all the data type of every column if it already matches our requirement. For instance, we need a numerical data type (integer and float) on the longitude and latitude, for listing names we need to make sure the data is using string/object data type.The next step is cleaning up the data, oftentimes the data we load have various faults, such as typo, missing value, incomplete data, etc. By doing cleaning up, the data quality will have better quality to be used for further analysis.In my case, the missing values that are observed do not need too much treatment. Looking into our dataset, we can state columns ‘ name’ and ‘host\_name’, ‘last\_review’ are irrelevant and unethical for further data exploration analysis. Therefore, we can get rid of those columns.After we clean up the data, the next step is exploring the data by visualizing and analyzing the values of the features, explaining the process and the results.For our case, we will look up a various listing category consisting of each biggest value, visualize the listing distribution using a map, create a room type proportion for each area, looking for selling value from their listing name, and finding the average price of the most popular listing.Finally I reached the conclusion that the properties have large differences in prices. Separating the dataset by price categories is useful for the analysis. The most interesting variables regarding price prediction are: Location Room type calculated\_host\_listings\_count Number of review Price prediction models are not performing well Best score is 0.55 Prediction are nore accurate for price under $175 (75% of the dataset) Using categorical ecnoded data did not improve the model Possible next steps The next step could be to transform continuous variables into categorical variables as it can help capture non-linear relations. I doubt a neural network would be useful here because the number of observations is limited. |