# Capstone Project Submission

**Instructions:**

i) Please fill in all the required information.<br>
ii) Avoid errors.

**Team Member's Name, Email and Contribution:**

ABDUL QUADIR
abdulec1002@gmail.com

1. EDA and Code
    a. Data Wrangling
    b. Code Comments, Code Quality Check
    c. Data Visualizations
    d. Price distributions
    e. Revenue analysis
    f. Notebook summarization
2. Documentation
    a. Presentation
    b. Technical Documentation

Mohammad Atique Najmuddin Shaikh
atiqueshaikh0141@gmail.com

3. EDA and Code
    a. Data Wrangling
    b. Code Comments, Code Quality Check
    c. Distributions on availability
    d. Finding busiest hosts
    e. Analyzing reviews
    f. Code modulation
4. Documentation
    a. Technical Documentation
    b. Technical Summary

Please paste the GitHub Repo link.

https://github.com/Abdul1328431002/Air-bnb

**Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)**

About the Dataset:
The dataset that we have chosen is Airbnb (NYC 2019) dataset and has 49895 and 16 columns. these columns are mixed between numerical and categorical columns. In this dataset we will analyze all the possible columns which we can use in analysis such as- host id, host name, revenue, price, listings, availability and so on.

Problem Statement:
Some of our problem statements are given as. but we are not limited to these problems:
1. What we can learn about hosts and areas?
2. Difference of Traffic among different Neighborhood.
3. Prediction on different columns of dataset ( exp:-price, availability etc.)
4. Which hosts are the busiest and why?

Approach taken:
Our approach can be explained in these steps-

- Acquire and loading data
- Understanding and summarizing the variables
- Cleaning dataset
- Exploring and Visualizing data
- Analyzing relationships between variables
- Summarizing the whole work

Challenges faced:

- We have faced challenges in selecting the columns for the analysis.
- Although there were only few columns which had missing values but it was a bit difficulty in treating these missing values.
- We also have face some challenges in choosing the right plot as visualization is the most important part of the project hence selecting the appropriate features was very important.

Overall Summary: -
This Airbnb-NYC (2019) dataset is very informative dataset having 48895 rows and 16 columns. we found that our top host has 327 listings. Then we have seen that Manhattan has highest number of listings followed by Brooklyn and so on. After that, we proceeded with analyzing neighborhood groups, neighborhoods and found that in top 10 neighborhoods only Manhattan and Brooklyn groups take part. Then we moved to analyze price and observed that Price is more distributed across the dataset in a specified range (40,200). also there are some outliers. Another very interesting insights we found by analyzing revenue made by top hosts (top revenue makers) that top1000 hosts occupy approx. 25% of all revenue and top 10000 occupy approx. 61% of all revenue. In final section we observed that we have noticed that the minimum avg. availability is in Brooklyn followed by Manhattan and the maximum availability is in Staten. then we have found top 10 busiest hosts across NYC and noticed that Micheal is the busiest host followed by David and Sunder(NYC). This order is the as we had seen in section 2.1 for most listings. This can be due to

high number of reviews in the neighborhood.