DATA SCIENCE CAPSTONE PROJECT

*Sakshi Chahande*

**Requirements of Project Report**

1. Introduction where you discuss the business problem and who would be interested in this project.

2. Data section where you describe the data that will be used to solve the problem and the source of the data.

3. Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

4. Results section where you discuss the results.

5. Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.

6. Conclusion section where you conclude the report.

**Introduction**

I am a student pursuing my B.Tech in Computer Engineering in Mumbai, India. My idea for this project is to identify the most visited places in a particular city within India, and compare them, in order to determine which city is more suitable to live in. I believe this project is useful for people looking to move to a new city to expand their career, for furthering their education or to improve their lifestyle.

**Interested Audience:**

The idea behind this project is to provide a streamlined analysis for students and working people alike, to determine what which area, in which city would be most suited for their needs if they have to live there.

**Data Selection**

I am comparing India's two largest cities, Mumbai and Delhi.

I have obtained the data-set from the following link:

https://data.gov.in/resources/all-india-pincode-directory-contact-details-along-latitude-and-longitude/api

This dataset contains the postcodes of the post offices of each locality in India. I have extracted two data frames from this dataset and applied the GeoPy API to find and integrate the latitude and longitude for each area. This has helped me create the map using Folium API. I further plan to find the different venues in a particular area and compare the two using the FourSquare API.

**Methodology**

1. I have looked for a dataset that contains the pin codes of different areas in India.
2. Filter the dataset into two dataframes based on the relevant cities
3. Apply GeoPy API in order to find the latitude and longitude for a particular area. In case they are not available, drop the columns for further implementation
4. I have displayed my remaining data on a Folium map
5. Using the FourSquare database, I have found the nearby venues in each area
6. Using frequency calculations, I have found the top 10 categories of venues visited in each area.
7. Using this data, we can further filter which areas are best suited in a particular city.

**Results**

I have found the top 10 most visited venue categories in all areas. This data can help the users identify the areas in the selected cities and compare the two. For example, if a person has been offered a job in Mumbai, and wants to live in an area where a departmental store is easily available, they will be able to search for that particular area whose most visited category is ‘Department Store’

**Discussions**

As part of a future scope, I would like to compare the cities based on the total number of venues in each category to best determine the living areas.

**Conclusion**

I have successfully filtered the data set to identify the relevant cities, found the necessary latitude and longitude for each, and applied the foursquare API to find the most relevant places in each city.