RESTAURANT VENUES CLUSTERING IN CHENNAI CITY USING FOURSQUARE & POPULATION DATA

IBM Data Science Professional Certificate – Final Capstone project (The Battle of Neighborhoods)

By

Prabhakaran Elangovan

Contents

1. Introduction

1.1 Business Case

Chennai is one of the emerging market places in India next to Delhi and Mumbai. It is also a hub for industries, due to the availability of port for easy exports and import. The city's economy is supported majorly by urban population with high per capita income working in various sectors. It ranks 2nd in GDP contribution among cities and 10th in GDP per capita. This provides ample opportunities for budding entrepreneurs, especially in organized food retail sector with the recent gigantic growth of food delivery partners like Swiggy, Zomato, etc. As the city is developing for a really long period, there comes a concern for entrepreneurs in finding out the right place for their restaurant business. The objective of the project is to harness the power of data & Data Science and to derive insights about various locations. The project is aimed at reducing the burden of entrepreneurs to make better and profitable decisions while finalizing the venue to invest.

2. Data Requirements

This section will discuss about the publicly available data that are required in building a solution to our business case. The business case requires the public venues listed in each area along with the population data for each area. The data needs are met by accessing data from two popular public data sources.

- Population data – Primary Census Abstract (PCA) data released by Government of India partnered with the Chennai Metro Corporation department.
- Venue master list – Venue list for each area in Chennai as available in Foursquare application.

2.1 Chennai city Population data

The population data provided by Government of India includes multiple details that are irrelevant to the task at hand. Hence, the following details were utilized for the project:

- Total Households – Total number of households/houses.
- Total Population – Total population
- Male Population – Total male population
- Female Population – Total female population
- Kids Population – Total population of kids aged between 0-6
- Educated Population – Total number of educated persons
- Working Population – Total number of working persons

Data Source: http://censusindia.gov.in/pca/pcadata/Houselisting-housing-TM.html

The data is available for download in the above given link. As stated earlier, the population data contained multiple other details which are either irrelevant or not of much help in this context.

Also, the data does not cover complete demography of the population; few important details like age-wise classification, sector-wise workers cover only Agricultural, Household industries and does not cover other sector workers. Though, if available these details can help in improving our end solution it is not mandatory in this context. Snapshot of the excel file read into a pandas DataFrame is provided below for reference.

```
In [9]: ch_population = pd.read_csv(body)

        ch_population.rename(columns={'Area':'Location'}, inplace=True)
        ch_population.head()
```

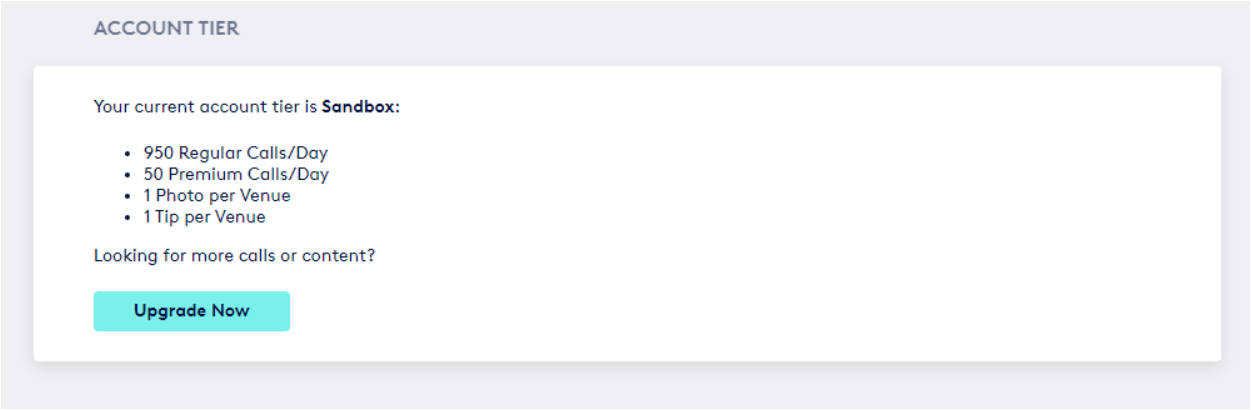
Out[9]:

	Ward No	Location	Total Houeholds	Population	Male Population	Female Population	Kids Population	Educated Population	Working Population
0	1	Kodungaiyur (West)	18900	76760	38805	37955	8209	63354	29282
1	2	Kodungaiyur (East)	16713	66897	33781	33116	7196	54439	25368
2	3	Dr.Radhakrishnan Nagar (North)	13248	52995	26804	26191	6326	40119	19429
3	4	Cherian Nagar (North)	3634	15186	7506	7680	1564	11579	5176
4	5	Jeeva Nagar (North)	11147	45204	22583	22621	4707	34811	17155

2.2 Venue Master Data

The list of venues available in each neighborhood is gathered from Foursquare via API available in their Developer Portal. Foursquare provides geographical data service to its users and the data creation and maintenance of valid information is crowd-sourced. Hence, their venues database available under a specific city is exhaustive and useful for business cases such as these.

Foursquare enables developers to access their data by touching different endpoints using API. This is free of cost; however there is a limit on the number of API calls that can be made in a single day for the free account (screenshot below).



2.3. Additional Data

The Foursquare API call requires the latitude and longitude of the area for which the venue data is required. It is to be noted that the population data does not contain the location co-ordinates; hence for this

purpose a ‘geocoder’ python library is used. Though multiple other libraries were available for the same purpose, geocoder was more effective.

****End of Document****