IBM APPLIED DATA SCIENCE CAPSTONE

Generating Area-based recommendations regarding the category of Restaurants to open in Mumbai

Kaustav Sadhukhan

Introduction:-

Nowadays since the market is very competitive it is very important to analyze a particular area very carefully and find out the taste of the local people, their preferences and their inclination towards a particular cuisine before starting a new business venture, to open a particular restaurant, in this case. This project is aimed towards generating Area-based recommendations regarding the category of Restaurants to open in Mumbai.

Business Problem:-

Opening the right-type of restaurant serving the particular category of cuisine is very important for anyone who wants to start a new business venture in terms of restaurants. The decision to open a restaurant which specializes in a particular cuisine depends on many important factors, few of which are listed below:-

1. The taste of the local people staying in that particular Area, their preferences and inclinations towards a particular food item/category/cuisine.

2. The number of similar type of restaurants serving the same kind of food in the nearby locations of that Area.

The target audience for this project includes any **new food chain** who is interested in starting a **new business venture** in the **city of Mumbai** and wishes to have recommendations regarding **which particular cuisine or which particular food item** they should serve **in their outlets** based on the **particular Area in Mumbai** where they are planning **to set up their new outlets**.

Here the project scope is limited to **South of Mumbai** , this scope can be extended to other Locations/Sectors of Mumbai by similar kind of Analytics.

Data Anatomization

To tackle the above mentioned problem, we need to have the dataset that contains:-

1. All the Areas and Locations in the City Of Mumbai.
2. Latitude and longitudes of all the Areas.
3. The most common venues and their categories in a particular Location of Mumbai (eg :- South Mumbai, Andheri; Western Suburbs; Bandra, Western Suburbs etc.)

- The Wikipedia page https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai is the major source of data that is being used to obtain all the Areas and Locations of Mumbai. We then use **beautifulsoup4 package, a Python module that helps to scrape** information from the web Pages to extract all the tables from this Wikipedia page and convert it into a pandas Data frame.
- Then we use **Python's geopy package** to obtain the latitude and longitude of all the Areas present in the Data frame.
- Finally we use **Four Square API** to generate **nearby venues and their categories** for a particular Area of Mumbai.

Description of the data:-

The output shows the different datasets used to complete the project. Description of each processed dataset is given at the top of the Dataframe.

- Data scraped from https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai showing the Area, locations and Latitude/Longitude values.

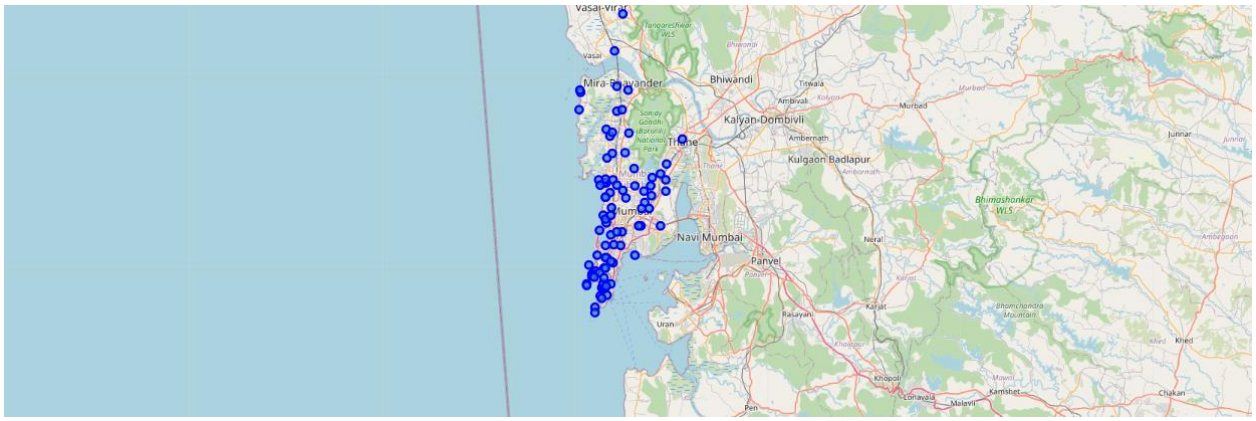
<title>List of neighbourhoods in Mumbai - Wikipedia</title>

Area	Location	Latitude	Longitude
Amboli	Andheri,Western Suburbs	19.1293	72.8434
Chakala, Andheri	Western Suburbs	19.111388	72.860833
D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
Four Bungalows	Andheri,Western Suburbs	19.124714	72.82721
Lokhandwala	Andheri,Western Suburbs	19.130815	72.82927
Marol	Andheri,Western Suburbs	19.119219	72.882743
Sahar	Andheri,Western Suburbs	19.098889	72.867222
Seven Bungalows	Andheri,Western Suburbs	19.129052	72.817018
Versova	Andheri,Western Suburbs	19.12	72.82
Mira Road	Mira-Bhayandar,Western Suburbs	19.284167	72.871111
Bhayandar	Mira-Bhayandar,Western Suburbs	19.29	72.85
Uttan	Mira-Bhayandar,Western Suburbs	19.28	72.785
Bandstand Promenade	Bandra,Western Suburbs	19.042718	72.819132
Kherwadi	Bandra,Western Suburbs	19.0553	72.8314
Pali Hill	Bandra,Western Suburbs	19.068	72.826
I.C. Colony	Borivali (West),Western Suburbs	19.247039	72.84983
Gorai	Borivali (West),Western Suburbs	19.250057	72.782021
Dahisa	Western Suburbs	19.250069	72.859347
Aarey Milk Colony	Goregaon,Western Suburbs	19.148493	72.881756
Bangur Nagar	Goregaon,Western Suburbs	19.167362	72.832252
Jogeshwari West	Western Suburbs	19.12	72.85
Juhu	Western Suburbs	19.1	72.83
Charkop	Kandivali West,Western Suburbs	19.216182	72.830575
Poisar	Kandivali West,Western Suburbs	19.204511	72.837639
Mahavir Nagar	Kandivali West,Western Suburbs	19.211319	72.842737
Thakur village	Kandivali East,Western Suburbs	19.210206	72.87298
Pali Naka	Khar,Western Suburbs	19.062742	72.829396

- Then, we transform the data into a pandas data frame. The data frame obtained is given below:-

	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270

- Use geopy library to get the latitude and longitude values of Mumbai. Then we use folium to plot the neighborhood locations of Mumbai.



- After this it is time to use the Four Square API to explore the different areas of South Mumbai and segment them. Now since we are interested in only the areas in South Mumbai, so let's slice the original data frame and create a new data frame of the South Mumbai data.

```
southmumbai_data = neighbourhoods[neighbourhoods['Location'] == 'South Mumbai'].reset_index(drop=True)
southmumbai_data.head()
```

	Area	Location	Latitude	Longitude
0	Agripada	South Mumbai	18.9777	72.8273
1	Altamount Road	South Mumbai	18.9681	72.8095
2	Bhuleshwar	South Mumbai	18.9500	72.8300
3	Breach Candy	South Mumbai	18.9670	72.8050
4	Carmichael Road	South Mumbai	18.9722	72.8113

- Once this is done we pick-up a specific area “Agripada” in South Mumbai, get its latitude and longitude using geopy/from the data frame we already scraped from Wikipedia and find the top 100 venues that are within the radius of 500 mts. from Agripada. We also get the category name for each venue and the latitude and longitude values. This is done using the Four Square API and the final dataset looks like below:-

	name	categories	lat	lng
0	Celejor	Bakery	18.975844	72.823679
1	cafe coffee day	Coffee Shop	18.976988	72.824051
2	Gold Gym IndiaBulls, Lower Parel	Gym	18.981424	72.824900
3	Warden bakery	Bakery	18.980180	72.826118
4	Vedanta Veg Restaurant	Indian Restaurant	18.980548	72.828777

- Finally we extend this process to all nearby areas in South Mumbai:-

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agripada	18.977700	72.827300	Celejor	18.975844	72.823679	Bakery
1	Agripada	18.977700	72.827300	cafe coffee day	18.976988	72.824051	Coffee Shop
2	Agripada	18.977700	72.827300	Gold Gym IndiaBulls, Lower Parel	18.981424	72.824900	Gym
3	Agripada	18.977700	72.827300	Warden bakery	18.980180	72.826118	Bakery
4	Agripada	18.977700	72.827300	Vedanta Veg Restaurant	18.980548	72.828777	Indian Restaurant
5	Agripada	18.977700	72.827300	Blue Waters	18.981119	72.827098	Restaurant
6	Altamount Road	18.968100	72.809500	Theobroma	18.970735	72.809816	Bakery
7	Altamount Road	18.968100	72.809500	Swati Snacks	18.966442	72.813531	Indian Restaurant
8	Altamount Road	18.968100	72.809500	Sophia Bhabha Hall	18.969819	72.807082	Theater
9	Altamount Road	18.968100	72.809500	Under The Banyan Tree	18.969328	72.808344	Café
10	Altamount Road	18.968100	72.809500	Cafe Coffee Day	18.966999	72.813983	Coffee Shop

- Data Pre-processing

Now we perform one-hot encoding with respect to the venue-categories of each area, then *group rows by Area and by taking the mean of the frequency of occurrence of each category and finally print each Area along with the top 5 most common venues.*

```
----Agripada----
              venue  freq
0            Bakery  0.33
1  Indian Restaurant  0.17
2       Coffee Shop  0.17
3        Restaurant  0.17
4               Gym  0.17


----Altamount Road----
              venue  freq
0             Café  0.29
1  Indian Restaurant  0.14
2            Bakery  0.14
3       Coffee Shop  0.14
4           Theater  0.14


----Bhuleshwar----
              venue  freq
0  Indian Restaurant  0.50
1       Cheese Shop  0.07
2        Restaurant  0.07
3            Market  0.07
4    Ice Cream Shop  0.07
```

- Finally processed data obtained from Four Square API showing the different Areas and the 10 most common venues for each Area corresponding to **South Mumbai.** This data is then used for clustering purposes.

	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most
0	Agripada	Bakery	Indian Restaurant	Coffee Shop	Gym	Restaurant	Food Court	Diner	Donut Shop	Dumpling R
1	Altamount Road	Café	Indian Restaurant	Coffee Shop	Bakery	Theater	Sandwich Place	Food	Diner	Do
2	Bhuleshwar	Indian Restaurant	Restaurant	Cheese Shop	Fast Food Restaurant	Food	Ice Cream Shop	Market	American Restaurant	T
3	Breach Candy	Café	Coffee Shop	Bakery	Sandwich Place	Fast Food Restaurant	Department Store	Dessert Shop	Men's Store	Japanese R
4	Carmichael Road	Ice Cream Shop	Chinese Restaurant	Fast Food Restaurant	Food Truck	Snack Place	Outlet Mall	Café	Sandwich Place	Shop