The Battle of the Neighbourhoods

Finding Cafes and Pizzeria in Major Cities of India

Sachin Namdeo

21st September 2019

Table of Contents

1. Introduction / Business Problem

A Startup XYZ in India wants to set up Cafes and Pizzeria in top 10 Major Cities in India. The Startup wants to know the Specific Location in the Major Cities of India where setting up of Cafes or Pizzeria in those cities would profit it. This would also help in getting to know the taste of the people living in those cities. The Startup **aims at opening Cafe or Pizzeria** depending upon the **availability of the Pizzeria or Cafe at that place**, is **no Cafes or Pizzeria in vicinity**, and is **closer to the city center**.

In Summation of the Problem set we need:

1. Availability of the Cafe or Pizzeria
2. No Cafes or Pizzeria in vicinity
3. Should be Closer to the City Center

If we can figure out the above problems then, it would not also help the Startup but also help other startups that require necessary information about the types of restaurants that would be needed so as to introduce new taste to the people and would also help in knowing the specific taste of the people in those major cities. This information would also help businesses that are running already to expand their food options in either the same city in different vicinity or in some other cities.

Top 10 Major Cities in India by Population are:

1. **Mumbai (Maharashtra)**
2. **Delhi (Delhi)**
3. **Bangalore (Karnataka)**
4. **Hyderabad (Telangana)**

5. Ahmedabad (Gujarat)
6. Chennai (Tamil Nadu)
7. Kolkata (West Bengal)
8. Surat (Gujarat)
9. Pune (Maharashtra)
10. Jaipur (Rajasthan)

2 Dataset Description

2.1 Requirements

The Raw Data Required for the Problem set are:

1. Area Wise Postal Code of India
   : https://data.gov.in/sites/default/files/all india PO list without APS offices ver2.csv
2. Latitudes and Longitudes of the Cities according to Postal Codes
   : http://download.geonames.org/export/zip/IN.zip

2.2 Processing

The Raw Data needs to be processed on so as to generate a meaningful dataset that will help us in out in our problem set.

We will do the following necessary operations:

- Generating the Data Frames on both the Datasets :

- Extracting the information from csv where major cities are there
- Extracting the information from zip where major cities are there
- We can also use pgeocode library instead of the zip

- Combining both the Datasets so as to get the preferred Dataset
- Mapping the data set according to the latitudes and longitudes of the major cities
- Getting the necessary restaurants in the cities according to categories using Foursquare api
- Exploring the Cities based on the categories and latitude and longitude

```
data
```

	officename	pincode	officeType	Deliverystatus	divisionname	regionname	circlename	Taluk	Districtname	statename	Telephone	Su
0	Achalapur B.O	504273	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Asifabad	Adilabad	TELANGANA	NaN	Rec
1	Ada B.O	504293	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Asifabad	Adilabad	TELANGANA	NaN	A
2	Adegaon B.O	504307	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Boath	Adilabad	TELANGANA	NaN	Echo
3	Adilabad Collectorate S.O	504001	S.O	Non-Delivery	Adilabad	Hyderabad	Andhra Pradesh	Adilabad	Adilabad	TELANGANA	08732-226703	
4	Adilabad H.O	504001	H.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Adilabad	Adilabad	TELANGANA	08732-226738	
...	...	...	...	...	...	...	...	...	...	...	...	
154792	Uttar Sautanchak B.O	721649	B.O	Delivery	Tamluk	South Bengal	West Bengal	Nandakumar	East Midnapore	WEST BENGAL	NaN	Mirik
154793	Uttarjianda B.O	721151	B.O	Delivery	Tamluk	South Bengal	West Bengal	Panskura-i	East Midnapore	WEST BENGAL	NaN	E S. Mid
154794	Uttarkotebarh B.O	721626	B.O	Delivery	Tamluk	South Bengal	West Bengal	Bhagawanpur	East Midnapore	WEST BENGAL	NaN	Ka
154795	Uttarmechogram B.O	721139	B.O	Delivery	Tamluk	South Bengal	West Bengal	Panskura-i	East Midnapore	WEST BENGAL	NaN	Pa
154796	Uttarsonamui B.O	721648	B.O	Delivery	Tamluk	South Bengal	West Bengal	Nandakumar	East Midnapore	WEST BENGAL	NaN	Byab

154797 rows × 13 columns

Figure 1

2.3. Description of Data

The Table above consists of 13 Columns. This data is a raw data and needs to be processed.

The Columns are:

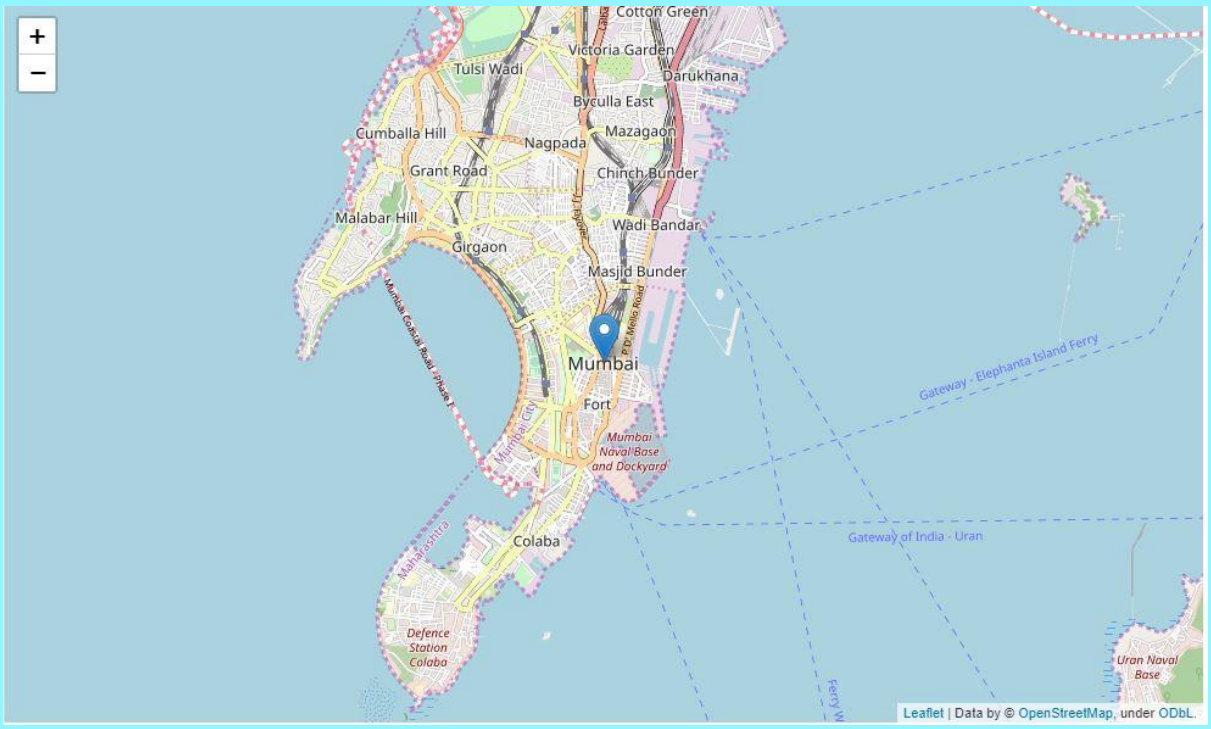
- Office Name of the Post Office
- Pincode of the Area of Post Office
- The Type of Office
- Delivery Status
- Division Name
- Region Name
- Circle Name
- Taluk(or Administrative Centers)
- District Name
- State Name
- Telephone
- Related Suboffice
- Related Headoffice

Necessary Columns Are:

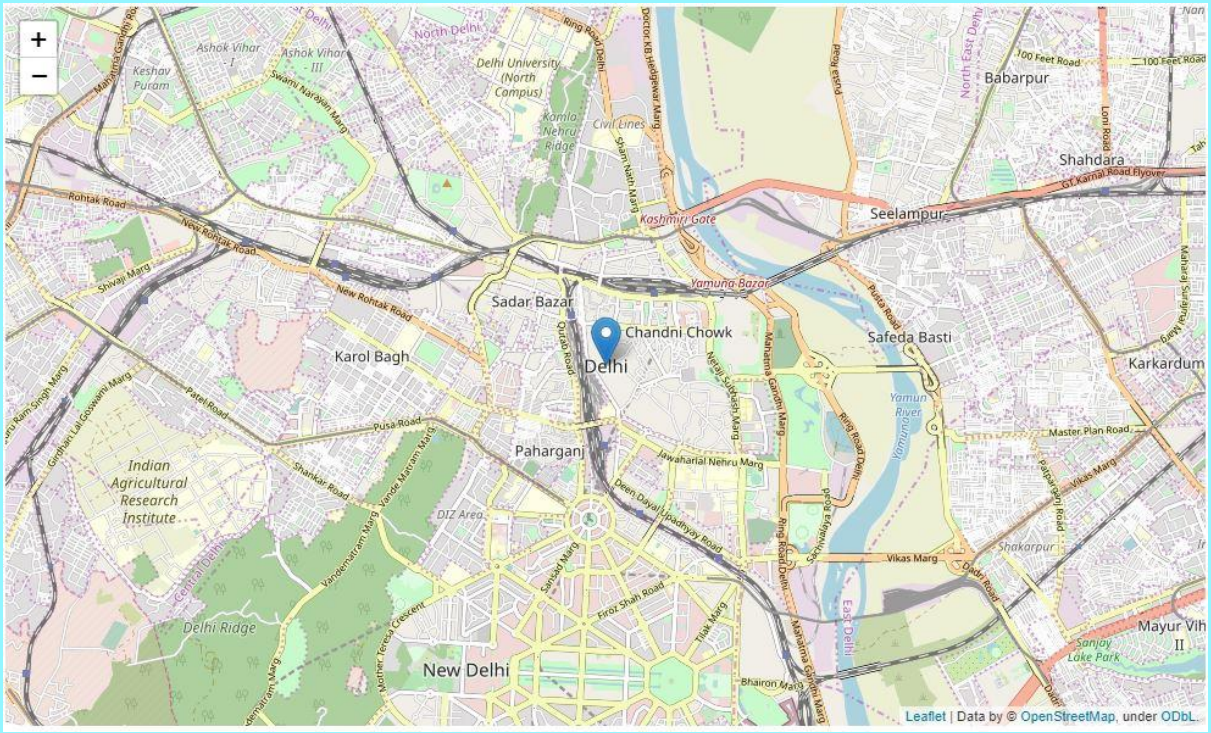
- Pincode of the Area of Post Office
- Taluk(or Administrative Centers)
- District Name

2.4 Maps of Each City

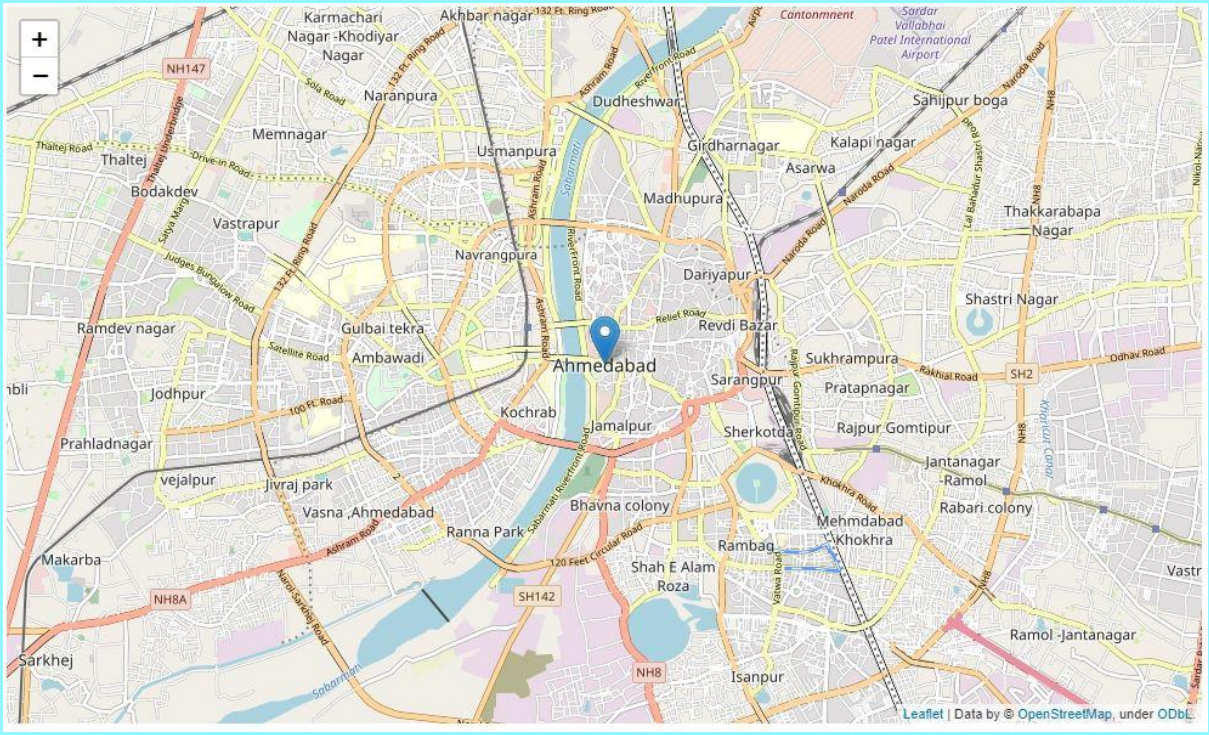
Mumbai



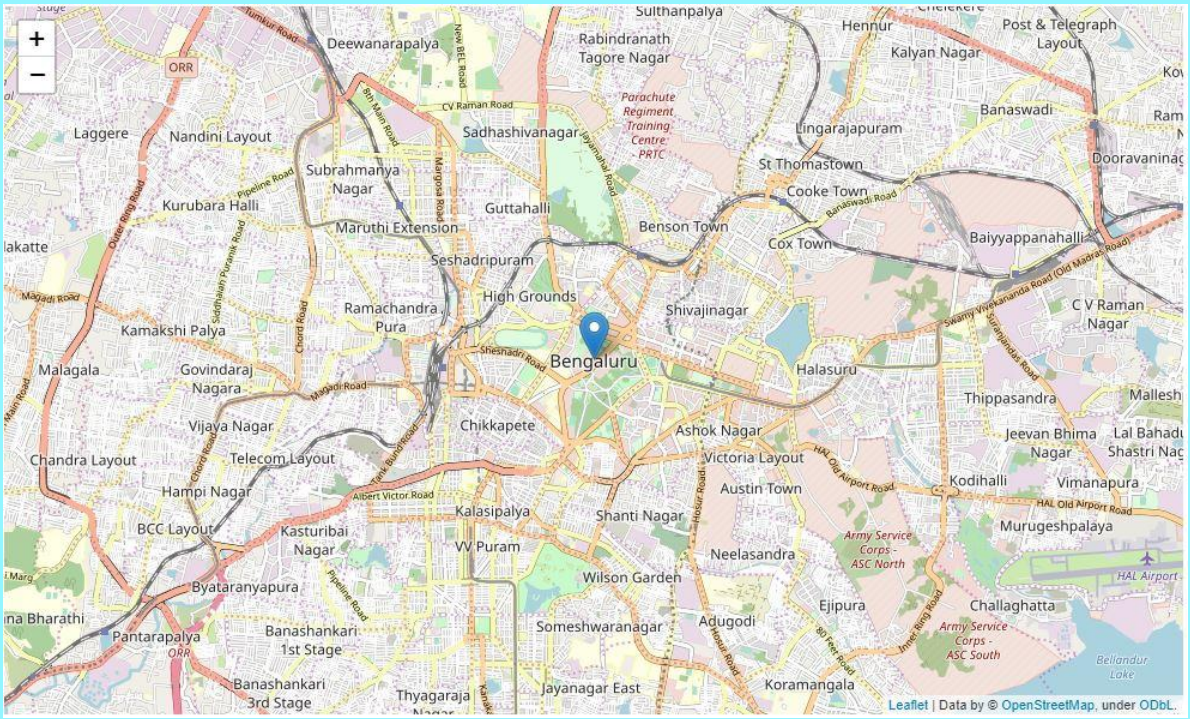
Delhi



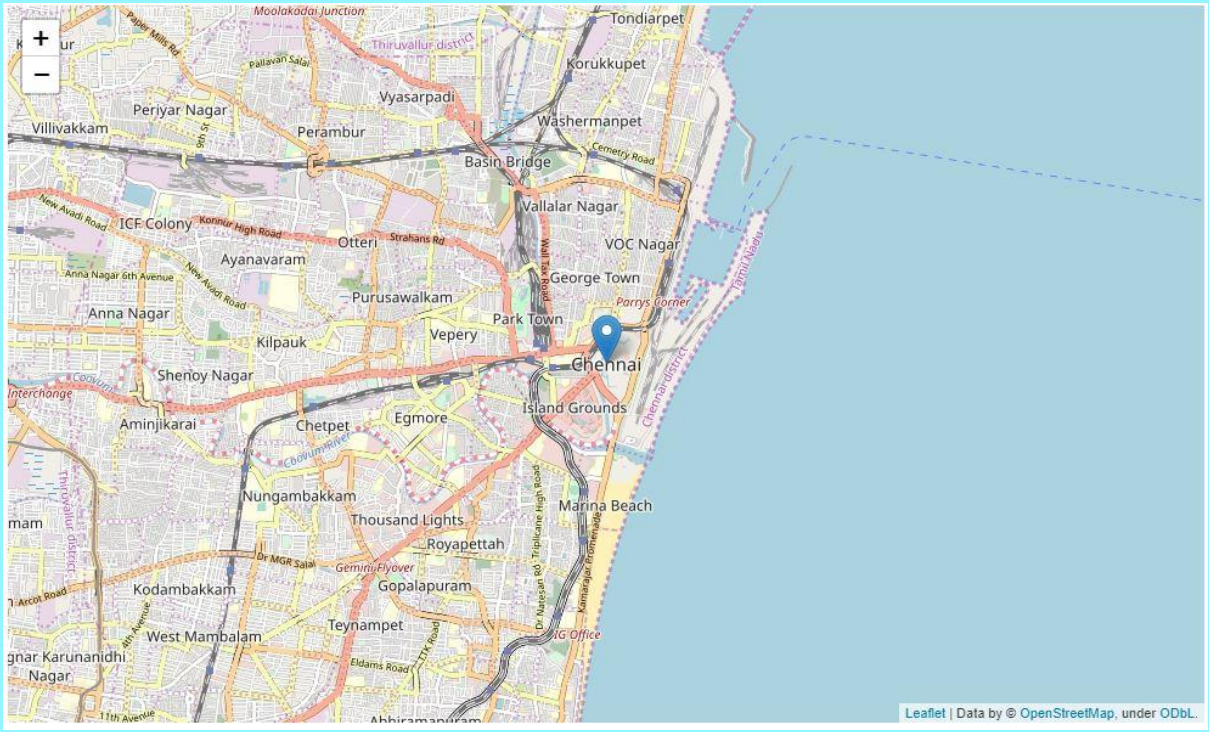
Ahmedabad



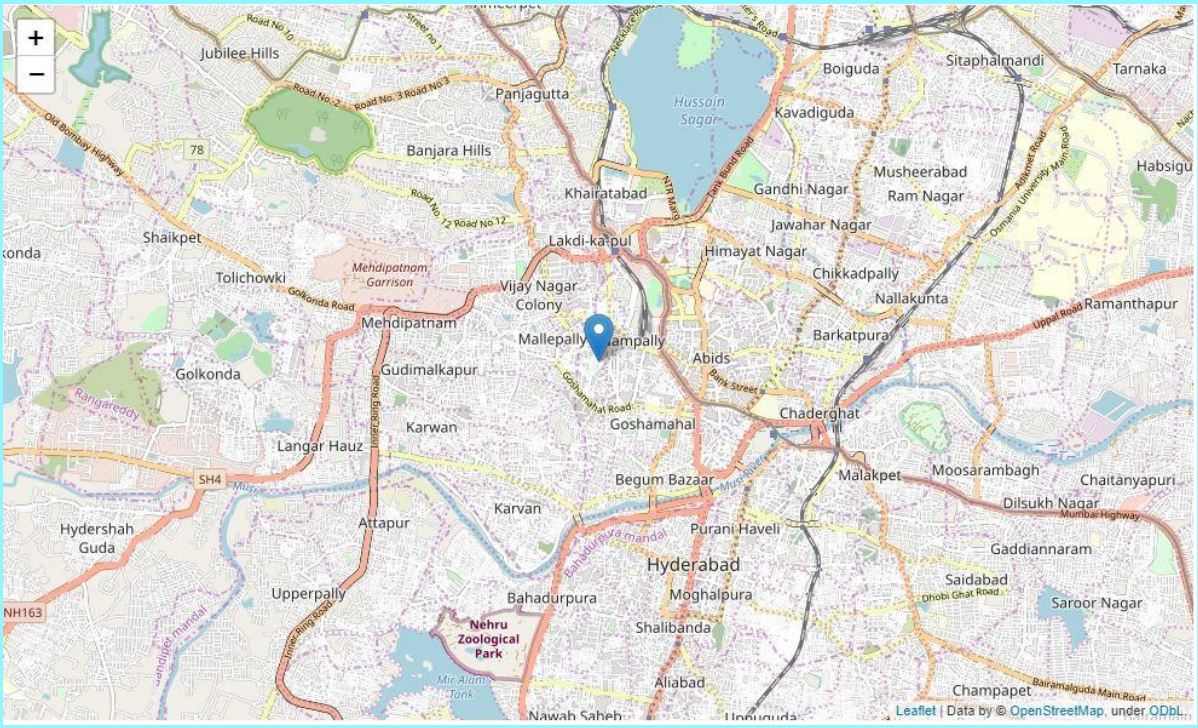
Bangalore



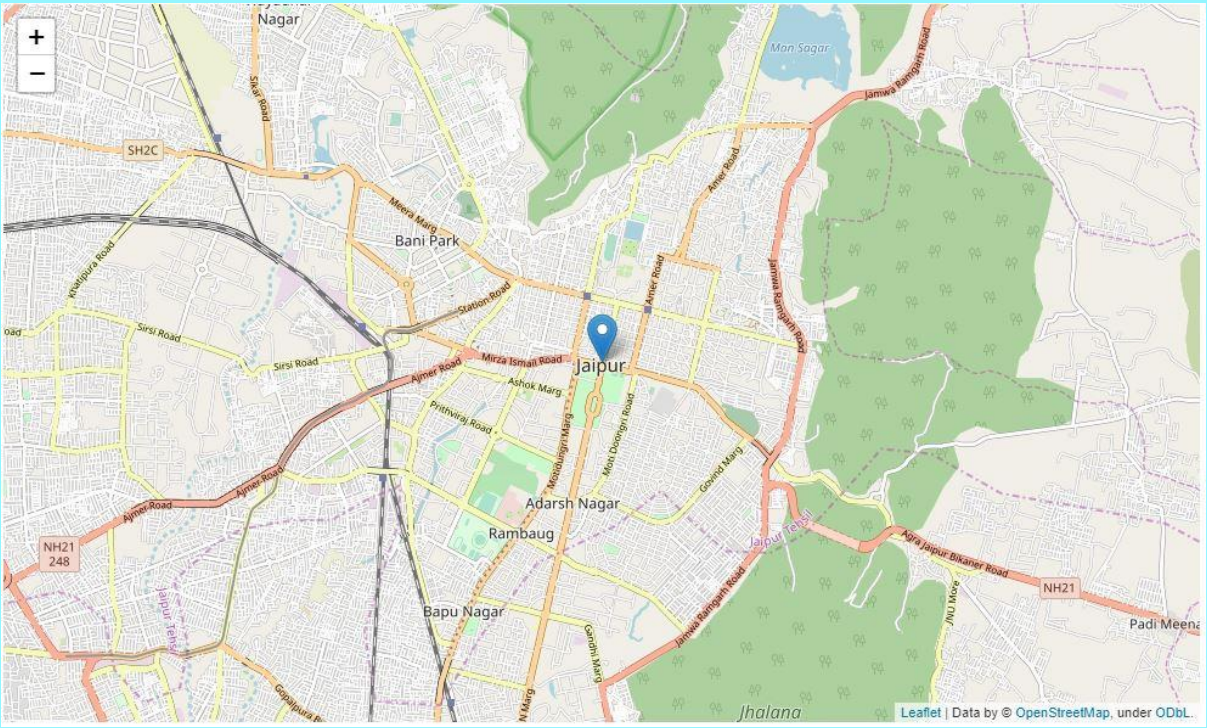
Chennai



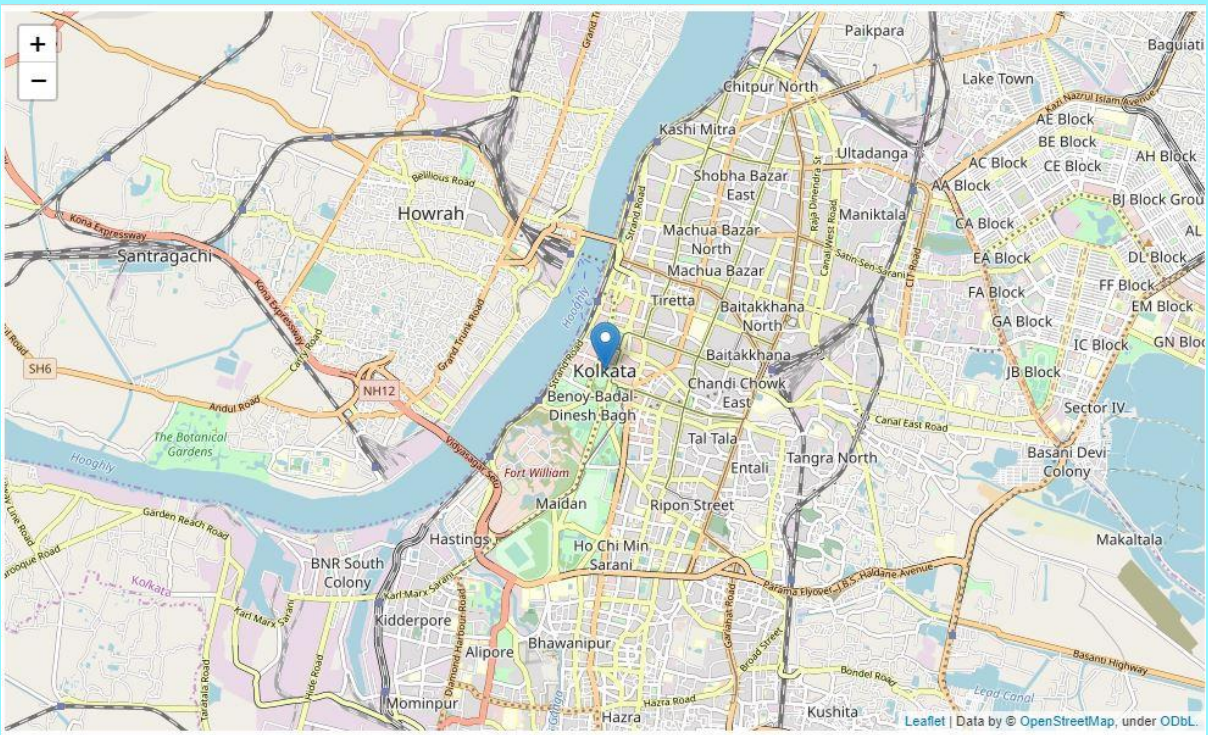
Hyderabad



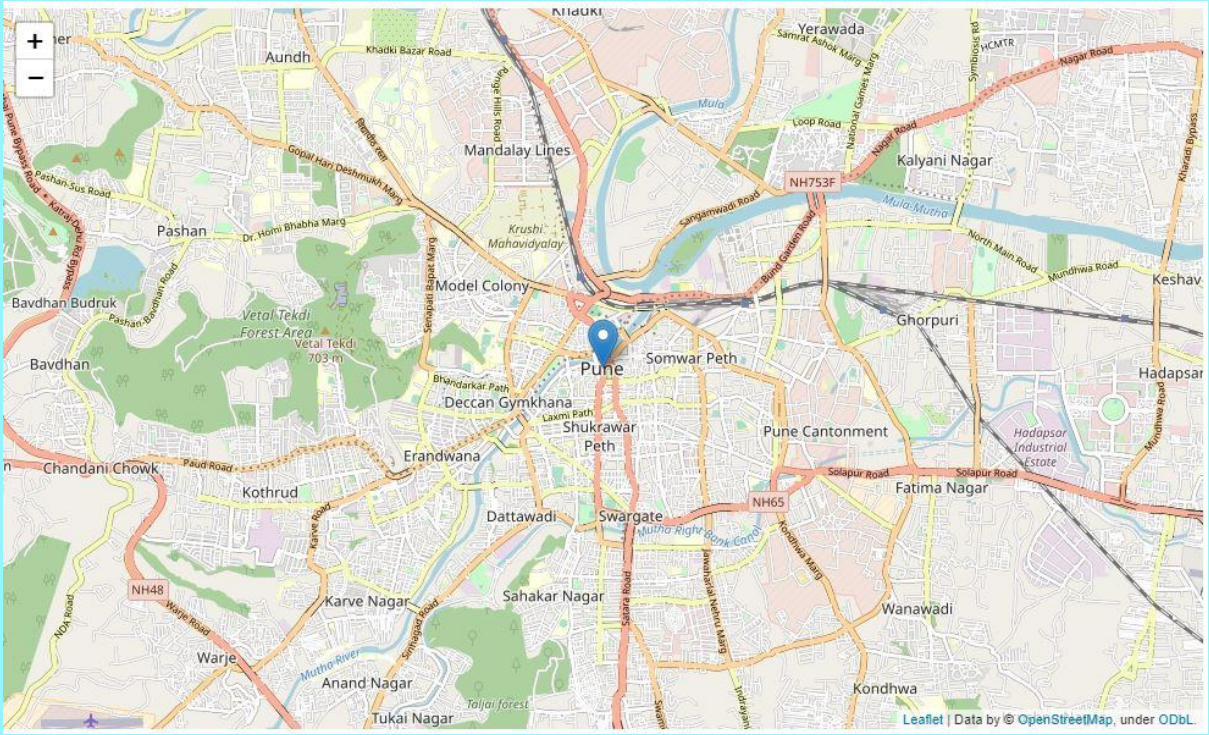
Jaipur



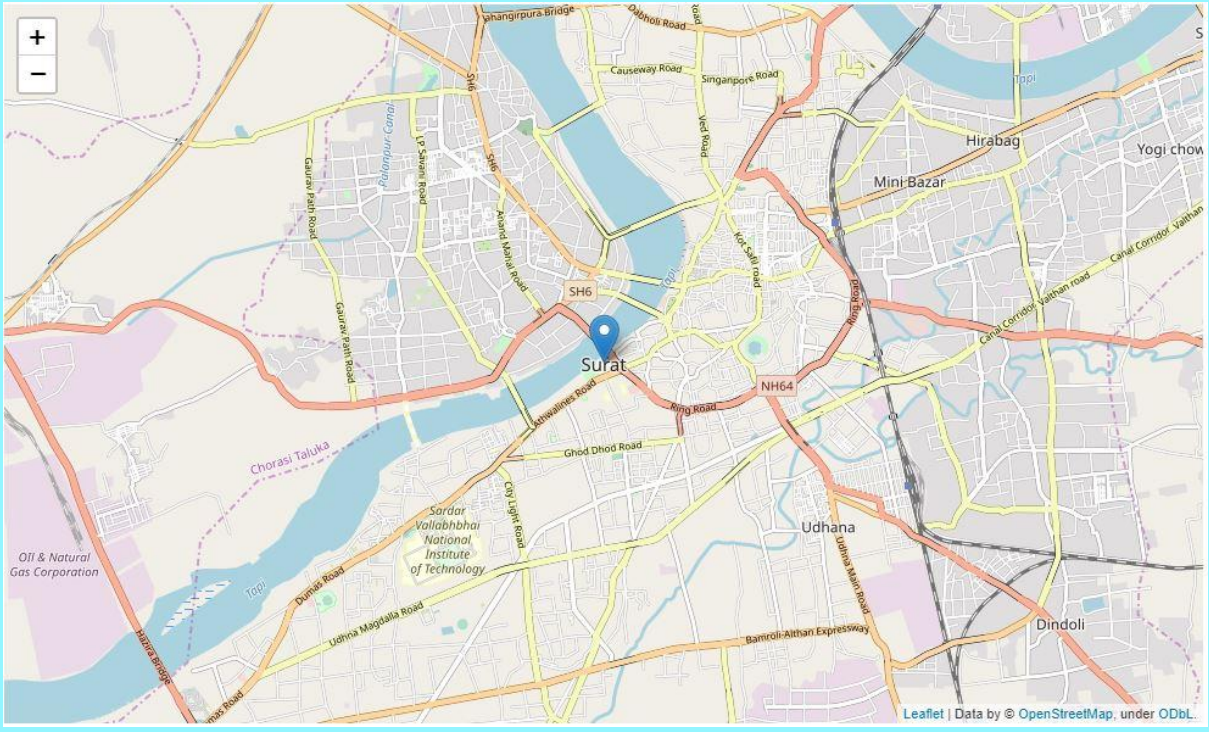
Kolkata



Pune


Aundh
Khadki Bazar Road
Yerawada
Samrat Ashok Marg
Nagar Road
Mula
Loop Road
National Games Marg
Mandalay Lines
Kalyani Nagar
NH753F
Mula-Mutha
Pashan
Dr. Homi Bhabha Marg
Krushi Mahavidyalay
Sangamwadi Road
Bund Garden Road
North Main Road
Mundhwa Road
Keshav
Bavdhan Budruk
Model Colony
Ghorpuri
Vetal Tekdi Forest Area
Vetal Tekdi 703 m
Bavdhan
Pune
Somwar Peth
Hadapsar
Deccan Gymkhana
Shukrawar Peth
Pune Cantonment
Hadapsar Industrial Estate
Chandani Chowk
Erandwana
Solapur Road
Kothrud
Fatima Nagar
NH65
Dattawadi
Swargate
Mutha Right Bank Canal
NH48
Karve Nagar
Sahakar Nagar
Wanawadi
Warje
Anand Nagar
Mutha River
Kondhwa
Tukai Nagar
Taljai forest
Leaflet | Data by © OpenStreetMap, under ODbL.

Surat


Jahangirpura Bridge
Causeway Road
Singanpore Road
Hirabag
Mini Bazar
Yogi chow
Gaurav Path Road
Malonjpur Canal
Anand Mahal Road
SH6
Surat
NH64
Ring Road
Ghod Dhod Road
Chorasi Taluka
Sardar Vallabhbhai National Institute of Technology
City Light Road
Udhana
Oil & Natural Gas Corporation
Dumas Road
Udhna Magdalla Road
Bamroli-Althan Expressway
Dindoli
Leaflet | Data by © OpenStreetMap, under ODbL.

3. Methodology

In this project we will direct our efforts on detecting areas of Major Cities in India that have low restaurant density, particularly those with low number of Cafes and Pizzeria. We will limit our analysis to area ~6km around city center.

In first step we have collected the required data: location and type (category) of every restaurant within 6km from City Center. We have also identified Cafes and Pizzeria (according to foursquare categorization).

Second step in our analysis will be calculation and exploration of 'restaurant density' across different areas of those cities - we will use heatmaps to identify a few promising areas close to center with low number of restaurants in general (and no Cafes and Pizzeria in vicinity) and focus our attention on those areas.

In third and final step we will focus on most promising areas and within those create clusters of locations that meet some basic requirements established in discussion with stakeholders: we will take into consideration locations with no more than three restaurants in radius of **350** meters, and we want locations without Cafes and Pizzeria in radius of **500** meters. We will present map of all such locations but also create clusters (using k-means clustering) of those locations to identify general zones / neighborhoods / addresses which should be a starting point for final 'street level' exploration and search for optimal venue location by stakeholders.

3.1. Exploratory Analysis

First and Foremost was the data cleaning. The data was cleaned by checking for any 'Null ' Values of Taluk. There were **8309** rows that had 'Null' values of Taluk. These Null values were removed and then the data remained to about **146488** rows and looked like the Dataframe below:

	officename	pincode	officeType	Deliverystatus	divisionname	regionname	circlename	Taluk	Districtname	statename	Telephone	Su
0	Achalapur B.O	504273	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Asifabad	Adilabad	TELANGANA	NaN	Rec
1	Ada B.O	504293	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Asifabad	Adilabad	TELANGANA	NaN	A
2	Adegaon B.O	504307	B.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Boath	Adilabad	TELANGANA	NaN	Echo
3	Adilabad Collectorate S.O	504001	S.O	Non-Delivery	Adilabad	Hyderabad	Andhra Pradesh	Adilabad	Adilabad	TELANGANA	08732-226703	
4	Adilabad H.O	504001	H.O	Delivery	Adilabad	Hyderabad	Andhra Pradesh	Adilabad	Adilabad	TELANGANA	08732-226738	
...	...	...	...	...	...	...	...	...	...	...	...	
146483	Uttar Sautanchak B.O	721649	B.O	Delivery	Tamluk	South Bengal	West Bengal	Nandakumar	East Midnapore	WEST BENGAL	NaN	Mirik
146484	Uttarjianda B.O	721151	B.O	Delivery	Tamluk	South Bengal	West Bengal	Panskura-i	East Midnapore	WEST BENGAL	NaN	E S. Mid
146485	Uttarkotebarh B.O	721626	B.O	Delivery	Tamluk	South Bengal	West Bengal	Bhagawanpur	East Midnapore	WEST BENGAL	NaN	Ka
146486	Uttarmechogram B.O	721139	B.O	Delivery	Tamluk	South Bengal	West Bengal	Panskura-i	East Midnapore	WEST BENGAL	NaN	Pa
146487	Uttarsonamui B.O	721648	B.O	Delivery	Tamluk	South Bengal	West Bengal	Nandakumar	East Midnapore	WEST BENGAL	NaN	Byab

146488 rows × 13 columns

Figure 2

After that taking the Necessary Columns '**Pincode**', '**Taluk**', '**District Name**' as our necessary columns and dropping the duplicates the dataframe reduced to **18710** values only

As shown below:

	pincode	Taluk	Districtname
0	504273	Asifabad	Adilabad
1	504293	Asifabad	Adilabad
2	504307	Boath	Adilabad
3	504001	Adilabad	Adilabad
4	504251	Bellampalle	Adilabad
...	...	...	...
18705	721604	Haldia Municipality	East Midnapore
18706	721625	Reapara	East Midnapore
18707	721153	Panskura	West Midnapore
18708	721655	Kajlagarh	East Midnapore
18709	721171	Sahid Matangini	East Midnapore

18710 rows × 3 columns

Figure 3

Using pgeocode library got the latitudes and longitudes according to the pincodes of the places and dropped null values of rows that had null values of latitudes and longitudes.

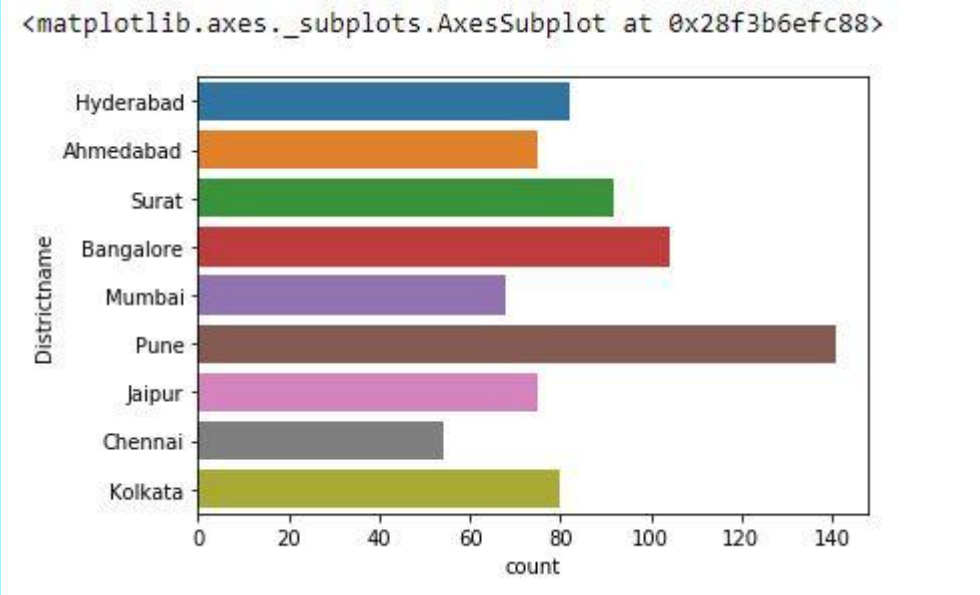
After this took only the major cities of **India** that are **'Mumbai'** ,**'Delhi'** ,**'Bangalore'** ,**'Hyderabad'** ,**'Ahmedabad'** ,**'Chennai'** ,**'Kolkata'** ,**'Surat'** ,**'Pune'** & **'Jaipur'**. This further reduced the Dataframe to **771** rows.

	pincode	Taluk	Districtname	latitude	longitude
0	500004	Khairatabad	Hyderabad	17.3872	78.4621
1	500045	Khairatabad	Hyderabad	17.5290	78.6839
2	500091	Rajendra Nagar	Hyderabad	17.3535	78.2402
3	500034	Khairatabad	Hyderabad	17.4168	78.4384
4	500022	Khairatabad	Hyderabad	17.4560	78.6476
...	...	...	...	...	...
766	700052	Kolkata	Kolkata	23.3945	88.2559
767	700067	Kolkata	Kolkata	22.4839	88.3038
768	700090	Kolkata	Kolkata	22.4485	88.3883
769	700004	Kolkata	Kolkata	22.7101	88.3167
770	700050	Kolkata	Kolkata	23.3945	88.2559

771 rows × 5 columns

Figure 4

The countplot of the dataframe gave us this figure:

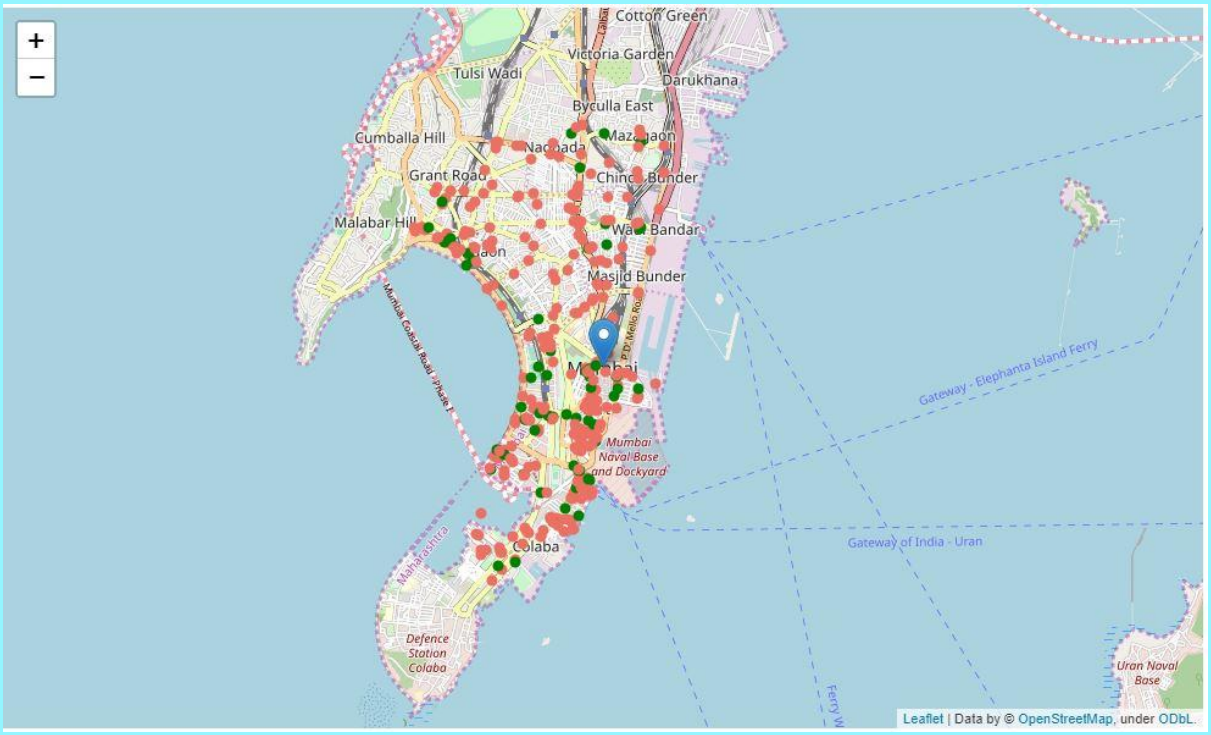


This figure showed that the number of areas of Pune is the most while the number of areas of Chennai is the least. Then after Transforming those latitudes and longitudes after using Geopy Library to Get the Coordinates of the centers of each city.

OpenCage Geocoder was used to reverse geocode those latitudes and longitudes of the centers of each city to get their addresses. OpenCage geocoder uses and API key to either forward or reverse geocode a location. Then addresses of other locations were also found out using OpenCage.

3.2 Foursquare API

Foursquare was then used on to the locations and areas were generated using Food Category '**4d4b7105d754a06374d81259**' and the Cafes '**4bf58dd8d48988d16d941735**' Irani Cafes '**54135bf5e4b08f3d2429dfe7**' and Pizza Places '**4bf58dd8d48988d1ca941735**'. Then Maps were generated to visualize the data.



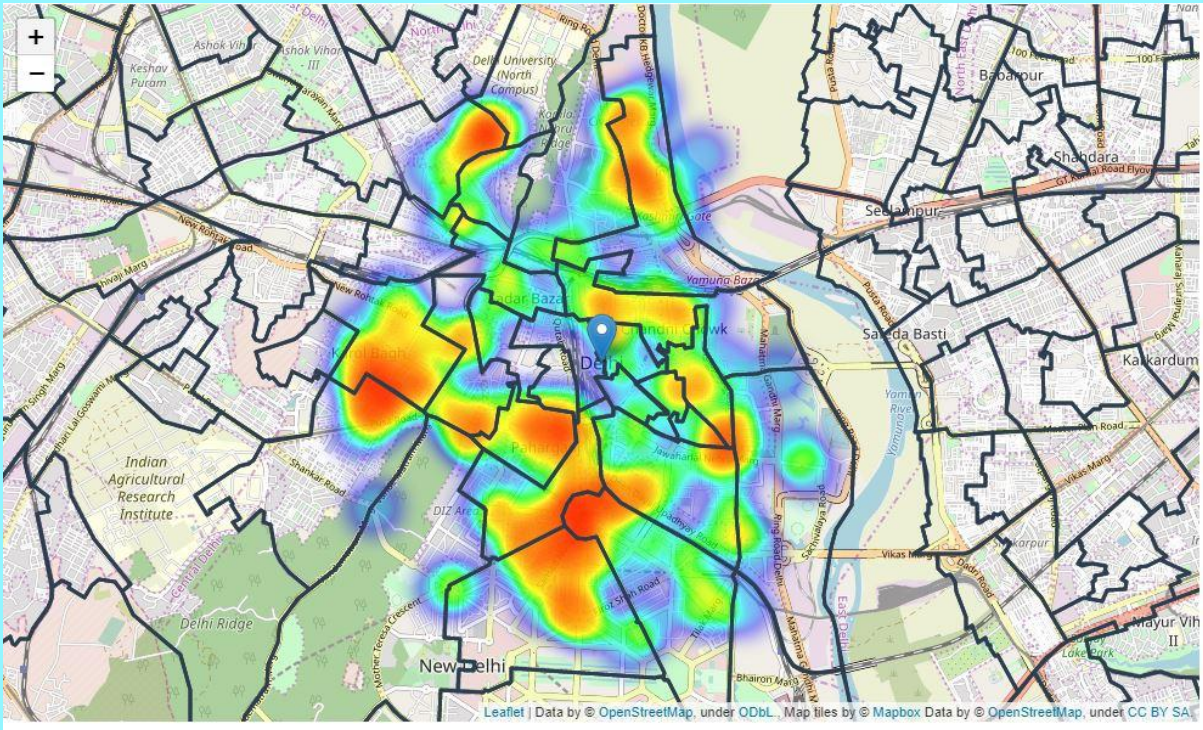
This is a Map of Mumbai where green dots denote Cafes and Pizzeria and red dots denote other restaurants.

A city dataframe was created that had address latitudes and longitudes along with restaurants in that area and distance from cafes and pizzeria.

	address	latitude	longitude	distance from center	x	y	Restaurants in area	Distance to Cafe & Pizzeria
0	A Ward, Mumbai - 400005, Maharashtra, India	18.916231	72.812756	5768.882041	7.523144e+06	3.641538e+06	0	1548.175044
1	Dalit Nagar, A Ward, Mumbai - 400005, Maharash...	18.914237	72.816781	5600.000000	7.523944e+06	3.641538e+06	4	748.378030
2	Sea Wind, Captain Prakash Pethe Marg, A Ward, ...	18.912242	72.820805	5542.562584	7.524744e+06	3.641538e+06	11	57.388028
3	Sassoon Docks, Dumayne Marg, A Ward, Mumbai - ...	18.910248	72.824829	5600.000000	7.525544e+06	3.641538e+06	0	603.370064
4	Sassoon Docks, Sassoon Dock Road, A Ward, Mumb...	18.908254	72.828853	5768.882041	7.526344e+06	3.641538e+06	0	1355.960576
...	...	...	...	...	...	...	...	...
1985	Bani Park, Jaipur Municipal Corporation - 3020...	26.946224	75.834696	5768.882041	7.144371e+06	5.133750e+06	0	6060.231126
1986	Bani Park, Jaipur Municipal Corporation - 3020...	26.943354	75.838607	5600.000000	7.145171e+06	5.133750e+06	5	6012.802918
1987	Bani Park, Jaipur Municipal Corporation - 3020...	26.940484	75.842518	5542.562584	7.145971e+06	5.133750e+06	2	5853.742464
1988	Jamwa Ramgarh Road, Adarsh Nagar, Jaipur Munic...	26.937614	75.846428	5600.000000	7.146771e+06	5.133750e+06	0	5528.040888
1989	Jaipur, Jaipur Municipal Corporation - 302 002...	26.934745	75.850338	5768.882041	7.147571e+06	5.133750e+06	0	5303.976932

1990 rows × 8 columns

Heatmaps were then generated for restaurants using foursquare API and used GeoJsons of each city to denote the boundaries of each neighbours in the city.



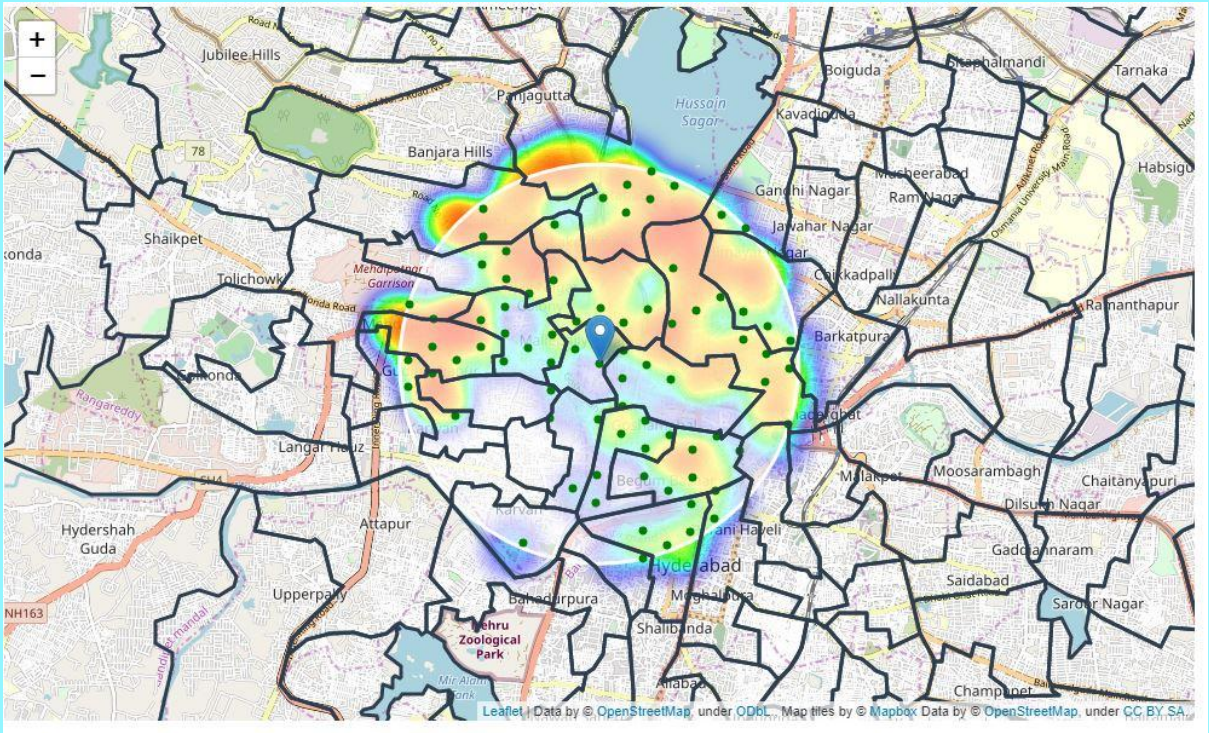
This is a Restaurant Heatmap of Delhi.

Then after that areas that had More than 1 restaurant and had a distance of more than 500 meters from Cafes and Pizzeria were taken and put into a new dataframe.

	address	latitude	longitude	distance from center	x	y	Restaurants in area	Distance to Cafe & Pizzeria
0	A Ward, Mumbai - 400005, Maharashtra, India	18.918551	72.816584	4995.998399	7.523544e+06	3.642231e+06	3	1328.481188
1	Mumbai Coastal Road - Phase I, A Ward, Mumbai ...	18.920871	72.820411	4233.202098	7.523944e+06	3.642924e+06	1	825.694455
2	Lala Ningam Road, A Ward, Mumbai - 400005, Mah...	18.914885	72.832482	4454.211490	7.526344e+06	3.642924e+06	2	778.982768
3	Cooperage, Maharshi Karve Road, A Ward, Mumbai...	18.925508	72.828067	2771.281292	7.524744e+06	3.644310e+06	3	584.865228
4	Mumbai Naval Base and Dockyard, Shahid Bhagat ...	18.925830	72.835919	2400.000000	7.525944e+06	3.645002e+06	2	563.144493
...	...	...	...	...	...	...	...	...
583	Jamwa Ramgarh Road, Adarsh Nagar, Jaipur Munic...	26.928676	75.844794	4454.211490	7.147571e+06	5.132364e+06	1	3960.171534
584	Bani Park, Jaipur Municipal Corporation - 3020...	26.941753	75.833879	4995.998399	7.144771e+06	5.133057e+06	2	5331.859005
585	Jamwa Ramgarh Road, Adarsh Nagar, Jaipur Munic...	26.933145	75.845611	4995.998399	7.147171e+06	5.133057e+06	2	4741.423554
586	Adarsh Nagar, Jaipur Municipal Corporation - 3...	26.930276	75.849520	5245.950819	7.147971e+06	5.133057e+06	1	4549.069415
587	Bani Park, Jaipur Municipal Corporation - 3020...	26.940484	75.842518	5542.562584	7.145971e+06	5.133750e+06	2	5853.742464

588 rows × 8 columns

Then the dataframe was split into 10 dataframes that had addresses corresponding to each city. These dataframes were then visualized into Heatmaps where Heatmap showed the restaurants and Points showed the location of Cafes and Pizzeria.



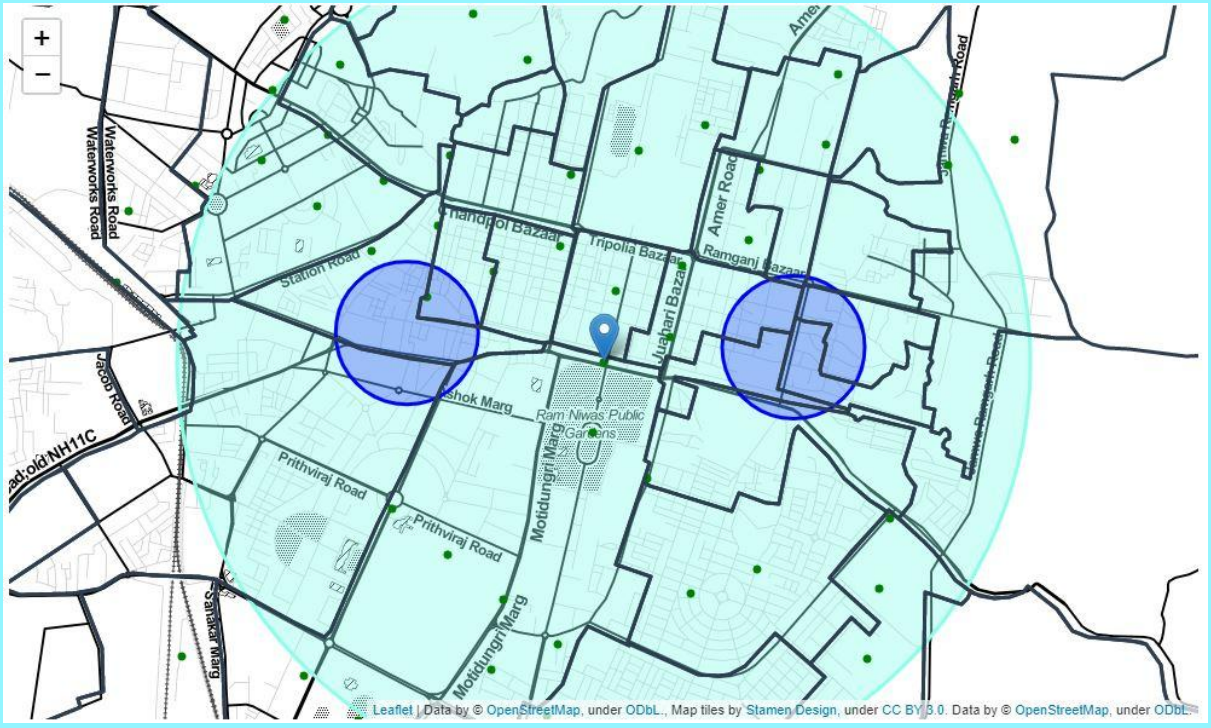
This is a restaurant heatmap of Hyderabad where green dots denote Cafes and Pizzeria.

The same was done with Cafes and Pizzeria Heatmap where the heatmaps of Cafes and Pizzeria was shown.

3.3. KMeans Clustering

Then using sklearn we imported KMeans and then we fitted the dataframe Need_DF's X and Y and then generated cluster Neighbourhoods with n_cluster of KMeans taken as **20**. These cluster centers were then put into Maps of each city to identify the neighbourhoods that are the best suitable for the opening of Cafes and Pizzeria.

These addresses of these cluster centers were then taken by reverse geocoding of the latitudes and longitudes using OpenCage geocoding and **21** addresses combined were found



This is a Cluster center map of Jaipur. The blue circles denote the cluster centers.

4.Results and Discussion

Our Analysis on the top 10 Major Cities of India showed us that there are more **2514 restaurants** combined in these cities out of which only **499 are Cafes & Pizzeria**. The Points of Interests are the Locations in Need_DF Dataframe which has more than 1 Restaurants in its vicinity and has no Cafes or Pizzeria in 500 meters. The Heatmaps generated shows that there are less Cafes or Pizzeria in 6Km radius of the city center than there are restaurants. The **Center of Interests** are the Places that Fulfil our Analysis of those areas which **are 21 in total**. These are Very Less Compared to the Total Number of Restaurants but provide us the **Most Suitable Locations** that can be used to put up Cafes or Pizzeria without any competition from other Cafes and Pizzeria that have been there before the setup of Start-up XYZ's Cafes and Pizzeria.

These Locations Were Found out Using **KMeans** Clustering Algorithms. The locations were filtered out that didn't contain any Cafes or Pizzeria in 500m Radius and was less than 6Km from the City Centers of Respective Cities. Those location candidates were then clustered to create zones of interest which contain greatest number of location candidates. Addresses of centers of those zones were also generated using reverse geocoding to be used as markers/starting points for more detailed local analysis based on other factors.

5. Conclusion

Purpose of this project was to identify areas close to City Center with low number of restaurants (particularly Cafes & Pizzeria) in order to aid stakeholders in narrowing down the search for optimal location for a new Cafe or Pizzeria restaurant. By calculating restaurant density distribution from Foursquare data we have first identified general Areas that justify further analysis, and then generated extensive collection of locations which satisfy some basic requirements regarding existing nearby restaurants. Clustering of those locations was then performed in order to create major zones of interest (containing greatest number of potential locations) and addresses of those zone centers were created to be used as starting points for final exploration by stakeholders.

Final decision on optimal restaurant location will be made by stakeholders based on specific characteristics of neighbourhoods and locations in every recommended zone, taking into consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighbourhood etc.