CONTRASTING AND COMPARING THE METROPOLITANS OF INDIA

MANSI PARASHAR

February 24, 2020

1.INTRODUCTION

1.1 Background

India is the 7th largest country in the world, and the 2nd most populated country.
India is a developing State and has 4 metropolitan cities so to speak of, Delhi which is the capital city, Mumbai(formerly Bombay), Chennai(formerly Madras) and Kolkata(formerly Calcutta).
In recent years, India has seen a tremendous boost in job opportunities and the general quality of life in these metros, leading to a large rate of migration to the aforementioned cities.
These 4 cities belong to different parts of the country, and hence vary significantly in the type of life they offer.

1.2 Problem Statement

Aim to analyse the 4 metropolitans of India ; explore the neighborhoods with relevant data and state the final findings in terms of the various venues and amenities provided at each city.
Contrast and compare this data retrieved for the 4 cities on the basis of different clusters formed by the neighborhoods and the services/amenities these clusters have to offer, and how they line with the interest of the stakeholders.
Also, compare the metropolitans on the basis of different unique amenities provided by them and the corresponding quantity.
Finally, identify the different regions in each metropolitan based on shared common venues to locate possible choices for potential residents.

1.3 Possible Applications

A family\An individual is looking forward to migrating to one of the metropolitans for better opportunities and for a better quality of life. Which neighborhood would be the best suited for them, and why?
A business owner wants to expand to one of the metros. What are the best regions for them to open up a business to conceivable success?
How similar are metropolitan cities, even when they are located in far corners of the country?
What are the differences between a city and a nearby metropolitan, on the basis of amenities offered?

2. DATA ACQUISATION AND PRE PROCESSING

2.1 Data Sources

We work on 3 sets of data :-

2.1.1 Data containing the neighborhoods, pincodes,latitude, longitude and for cities in India.This data is accessed from the official government of India site,https://www.india.gov.in/ which serves as a national portal for accessing different data concerning the country of India.
To access the data set, we need to fill in a form declaring we're using this data for academic purposes.
Unfortunately, all the latitude and longitude values were NaNs.

2.1.2 Furthermore, we found a data set containing Indian pincodes and their latitude and longitudes from the same site.

2.1.3 Based on the latitudes and longitudes in our database, we use the FourSquare API to access data about the neighborhood venues required for this study.
The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places.
Additionally, Foursquare allows developers to build audience segments for analysis and measurement.
JSON is the preferred response format which needs to be converted into the required dataframe
Accessing the API would require us to make an account on the Developer's Portal on the foursquare website,https://developer.foursquare.com/

2.2 Data Cleaning

From the second database, we delete the columns of city, neighborhood and accuracy to only get the pincodes for latitudes and longitudes, for a cleaner merge.

	pincode	latitude	longitude
0	110001	28.6333	77.2167
1	110002	28.6333	77.2500
2	110003	28.6500	77.2167
3	110004	28.6500	77.2167
4	110005	28.6500	77.2000

After merging on pincode, we get the following accurate dataset. We group the neighborhoods belonging to the same pincode and region.

Out[5]:

	pincode	latitude	longitude	regionname	neighborhood
0	110001	28.6333	77.2167	Delhi	Baroda House S.O,Bengali Market S.O,Bhagat Sin...
1	110002	28.6333	77.2500	Delhi	A.G.C.R. S.O,Ajmeri Gate Extn. S.O,Darya Ganj ...
2	110003	28.6500	77.2167	Delhi	Delhi High Court Extension Counter S.O,Delhi H...
3	110004	28.6500	77.2167	Delhi	Rashtrapati Bhawan S.O
4	110005	28.6500	77.2000	Delhi	Anand Parbat Indl. Area S.O,Anand Parbat S.O,B...

Because it is location data, there are no such outliers ; also, as the datasets have been merged on pincode which is a unique value for each region, the dataset is void of NaNs.

The foursquare data is returned as a json file. All the information is in the “items” key.
We extract the categories of the venue
We clean the json and convert it into a pandas dataframe

2.3 Feature Selection

From the dataset containing pincodes of all of India, we select the relevant ones;
That is, Pincodes corresponding to the cities :-

City	**number of rows**
Mumbai	197
Delhi	79
Kolkata	180
Chennai	141

Which corresponds to the size of the respective cities.

The features we’ve kept are
Pincode , Latitude, Longitude,City,Neighborhood.

Which are the only ones relevant to get the necessary foursquare data for exploratory analysis of neighborhoods in a city.

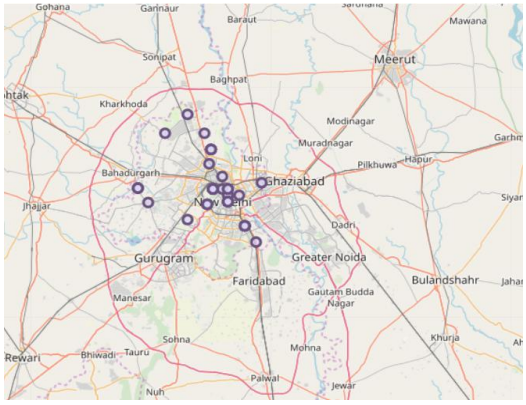
3. EXPLORATORY DATA ANALYSIS

3.1 Finding the neighborhoods in each of the metro cities,grouping them based on their pincodes and regionname and visualization the dataset obtained using using folium maps

Delhi

Out[5]:

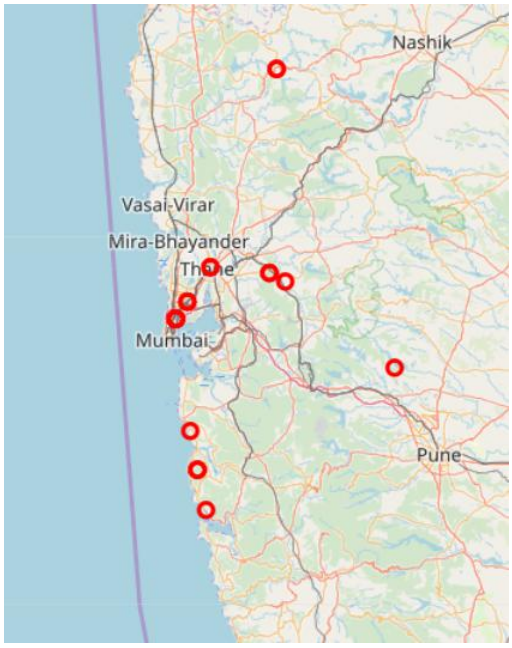
	pincode	latitude	longitude	regionname	neighborhood
0	110001	28.6333	77.2167	Delhi	Baroda House S.O,Bengali Market S.O,Bhagat Sin...
1	110002	28.6333	77.2500	Delhi	A.G.C.R. S.O,Ajmeri Gate Extn. S.O,Darya Ganj ...
2	110003	28.6500	77.2167	Delhi	Delhi High Court Extension Counter S.O,Delhi H...
3	110004	28.6500	77.2167	Delhi	Rashtrapati Bhawan S.O
4	110005	28.6500	77.2000	Delhi	Anand Parbat Indl. Area S.O,Anand Parbat S.O,B...



MUMBAI

Out[12]:

	pincode	latitude	longitude	regionname	neighborhood
0	400001	19.0167	72.8500	Mumbai	Bazargate S.O,Elephanta Caves Po B.O,M.P.T. S....
1	400002	19.0167	72.8500	Mumbai	Kalbadevi H.O,Ramwadi S.O,S. C. Court S.O,Thak...
2	400003	19.0167	72.8500	Mumbai	B.P.Lane S.O,Mandvi S.O (Mumbai),Masjid S.O,Nu...
3	400004	19.0167	72.8500	Mumbai	Ambewadi S.O (Mumbai),Charni Road S.O,Chaupati...
4	400005	19.0167	72.8500	Mumbai	Asvini S.O,Colaba Bazar S.O,Colaba S.O,Holiday...
...	...	...	...	...	...
92	410208	18.8389	73.6833	Mumbai	Devichapada B.O,Koynavale B.O,Navde B.O,Padghe...
93	421101	19.1833	73.2083	Mumbai	Padgha Extension Counter,Padgha S.O
94	421301	19.1833	73.2083	Mumbai	Kalyan City H.O,Kalyan D.C. S.O,Kalyan Rs S.O,...
95	421303	19.1833	73.2083	Mumbai	Abitghar B.O,Abje B.O,Alonde B.O,Baliwali B.O,...
96	421502	19.1500	73.2667	Mumbai	O.E.Ambernath S.O

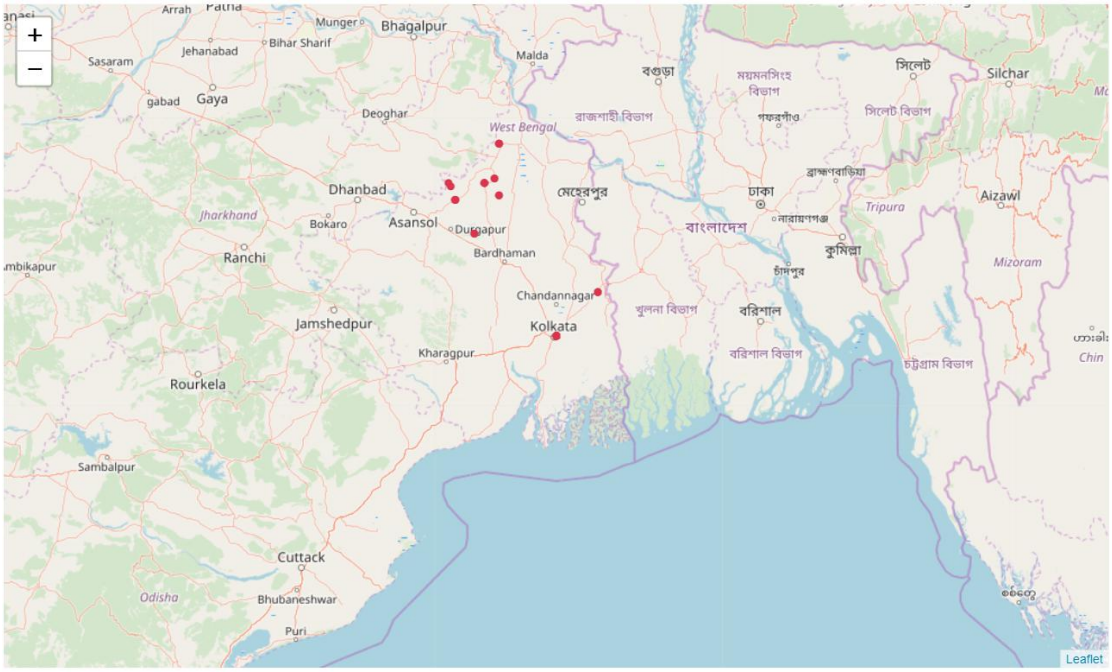


KOLKATA

Out[8]:

	pincode	latitude	longitude	regionname	neighborhood
0	700001	22.5690	88.3697	Calcutta	Council House Street S.O,Customs House S.O,Khe...
1	700002	22.5690	88.3697	Calcutta	Cossipore Gun Factory S.O,Cossipore H.O,Indian...
2	700003	22.5690	88.3697	Calcutta	Amrita Bazar Partika S.O,Baghbazar S.O,Girish ...
3	700004	22.5690	88.3697	Calcutta	R.G.Kar Medical College S.O,Shyambazar Mail S....
4	700005	22.5690	88.3697	Calcutta	Ahritola S.O,Hatkhola S.O
...	...	...	...	...	...
175	731241	22.9667	88.7833	Calcutta	Chandpara S.O,Dunigram B.O,Koyemba B.O,Nonadan...
176	731243	22.9667	88.7833	Calcutta	Ayas B.O,Banior B.O,Bautia B.O,Bhabanandapur B...
177	731301	23.8333	87.8167	Calcutta	Barha B.O,Belhati B.O,Ch.Nanoor S.O,Charkolgra...
178	731303	23.8333	87.8167	Calcutta	Abadanga B.O,Bhalas B.O,Bhatra B.O,Bipratikuri...
179	741101	13.9463	92.1467	Calcutta	Aminbazar S.O,Bowbazar S.O (Nadia),Golapati S....

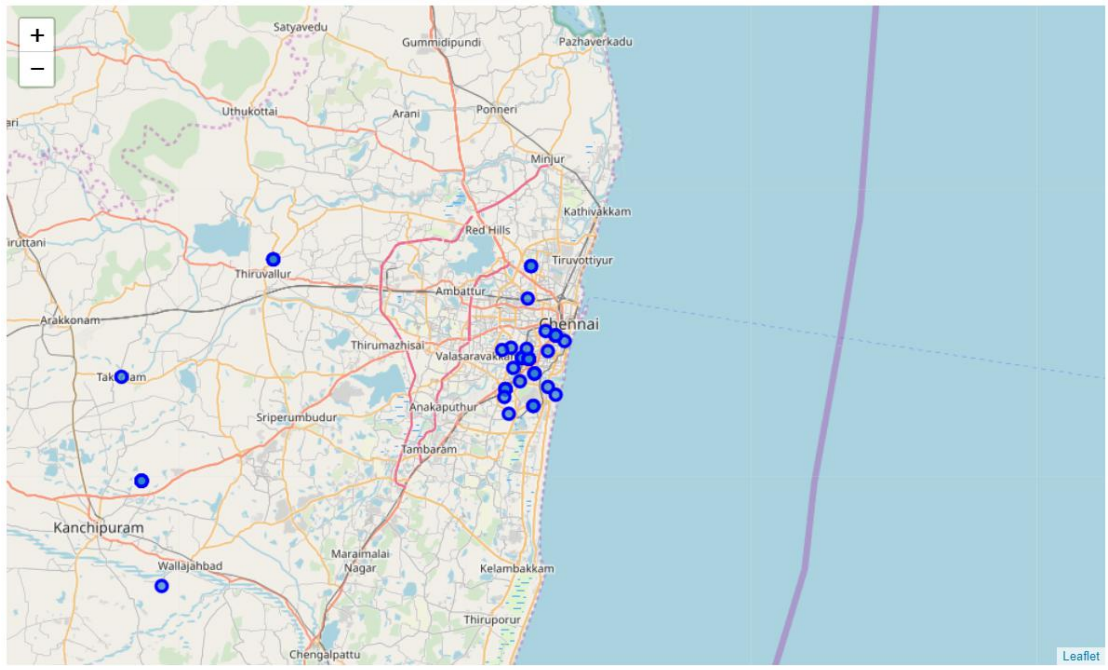
180 rows × 5 columns



CHENNAI

Out[9]:

	pincode	latitude	longitude	regionname	neighborhood
0	600001	13.0656	80.2672	Chennai Region	Flower Bazaar S.O,Govt Stanley Hospital S.O,Ma...
1	600002	13.0656	80.2672	Chennai Region	Anna Road H.O,Chintadripet S.O,Madras Electric...
2	600003	13.0656	80.2672	Chennai Region	Edapalayam S.O,Madras Medical College S.O,Park...
3	600004	13.0656	80.2672	Chennai Region	Mandaveli S.O,Mylapore H.O,Vivekananda College...
4	600005	13.0594	80.2789	Chennai Region	Chepauk S.O,Madras University S.O,Tiruvallikke...
...	...	...	...	...	...
136	632504	12.7667	79.4167	Chennai Region	Damarapakkam S.O,Valayathur B.O
137	632505	12.7667	79.4167	Chennai Region	Ayal B.O,Banavaram S.O,Govindacherrykuppam B.O...
138	632506	12.7667	79.4167	Chennai Region	Agaram B.O,Arumbakkam B.O,Chennasamudram B.O,G...
139	632511	12.9333	79.3833	Chennai Region	Ananthangal B.O,Brammadesam B.O,Esayanur B.O,J...
140	632513	12.9333	79.3833	Chennai Region	Anandhalai B.O,Gudimallur B.O,Poondi B.O,Satha...

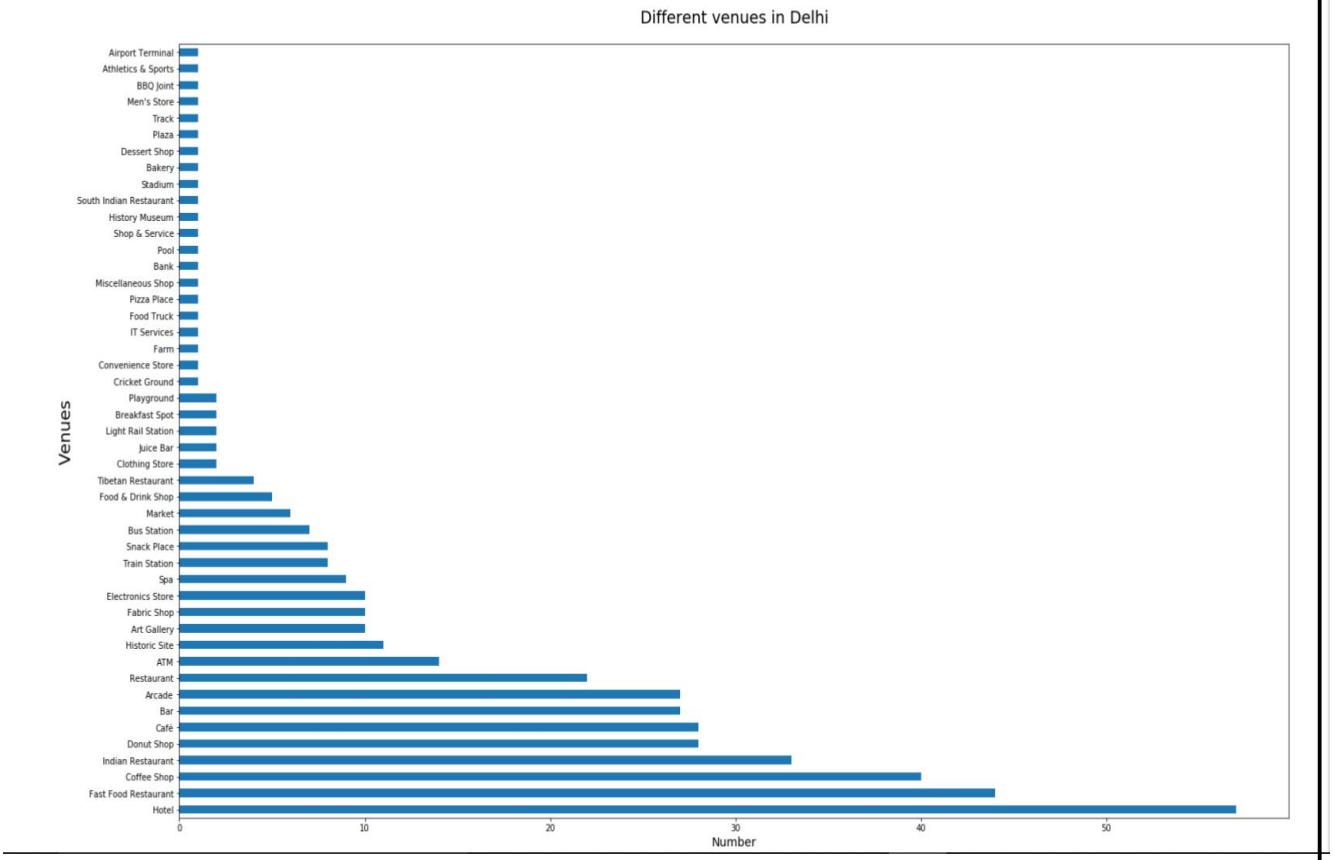


SEGMENTATION OF DATA

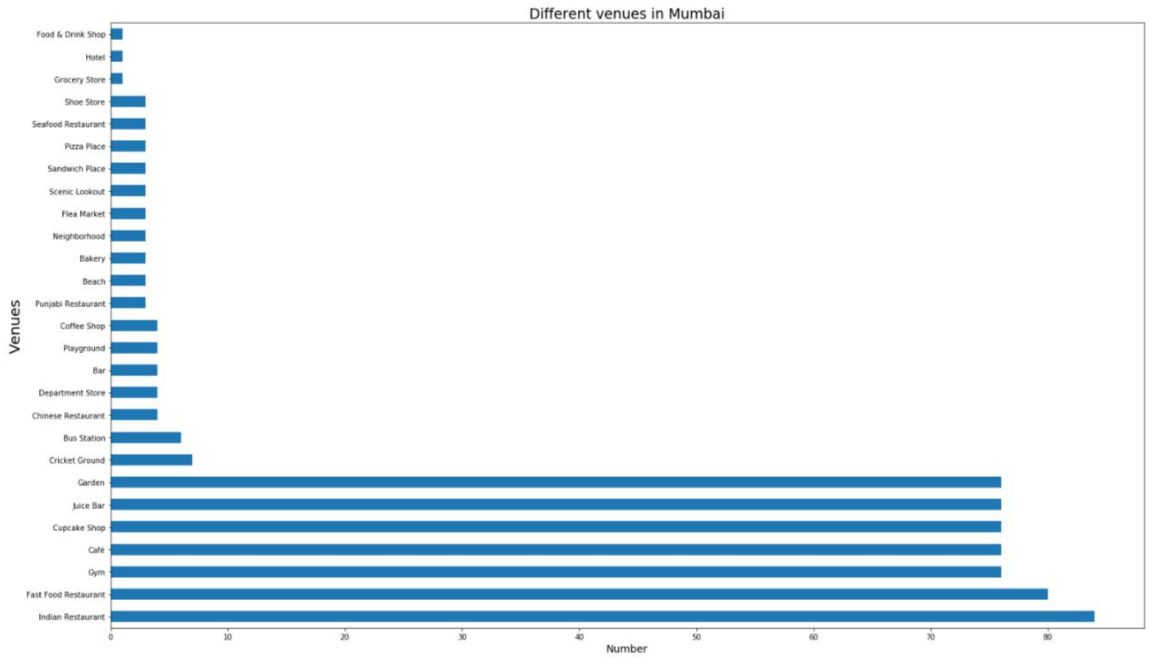
Data Segmentation is the process of taking the data you hold and dividing it up and grouping similar data together based on the chosen parameters so that you can use it more efficiently
We are going to utilize the Foursquare API to explore the neighborhoods and segment them.

We get the various venues corresponding to a neighborhood. We visualise the unique venues in order to get an idea of the different types of services/amenities provided.

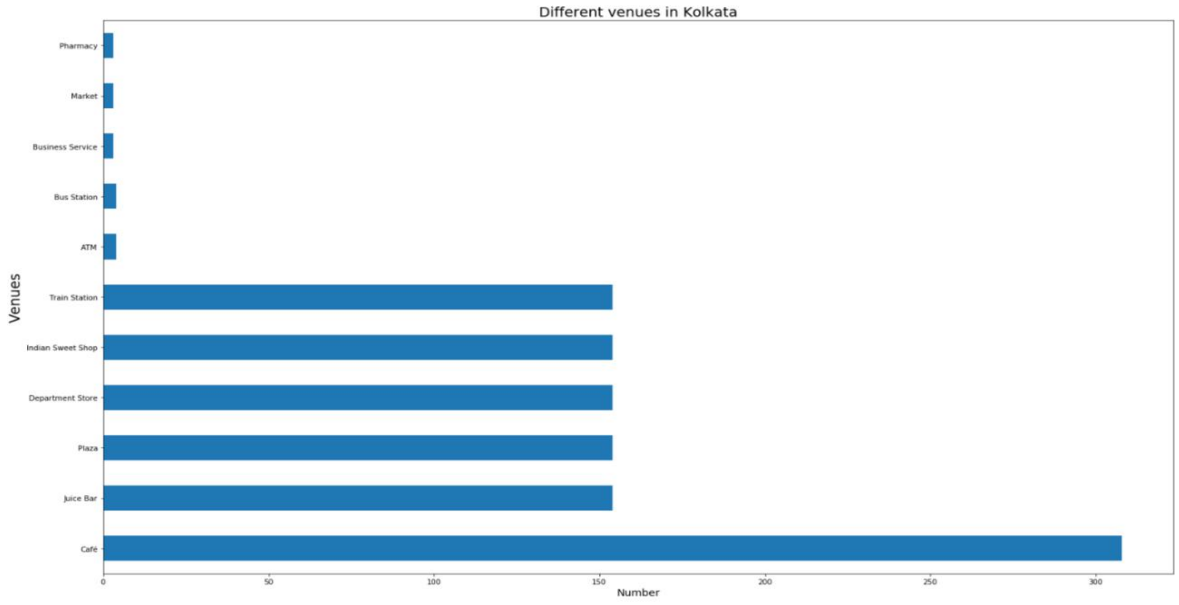
Delhi



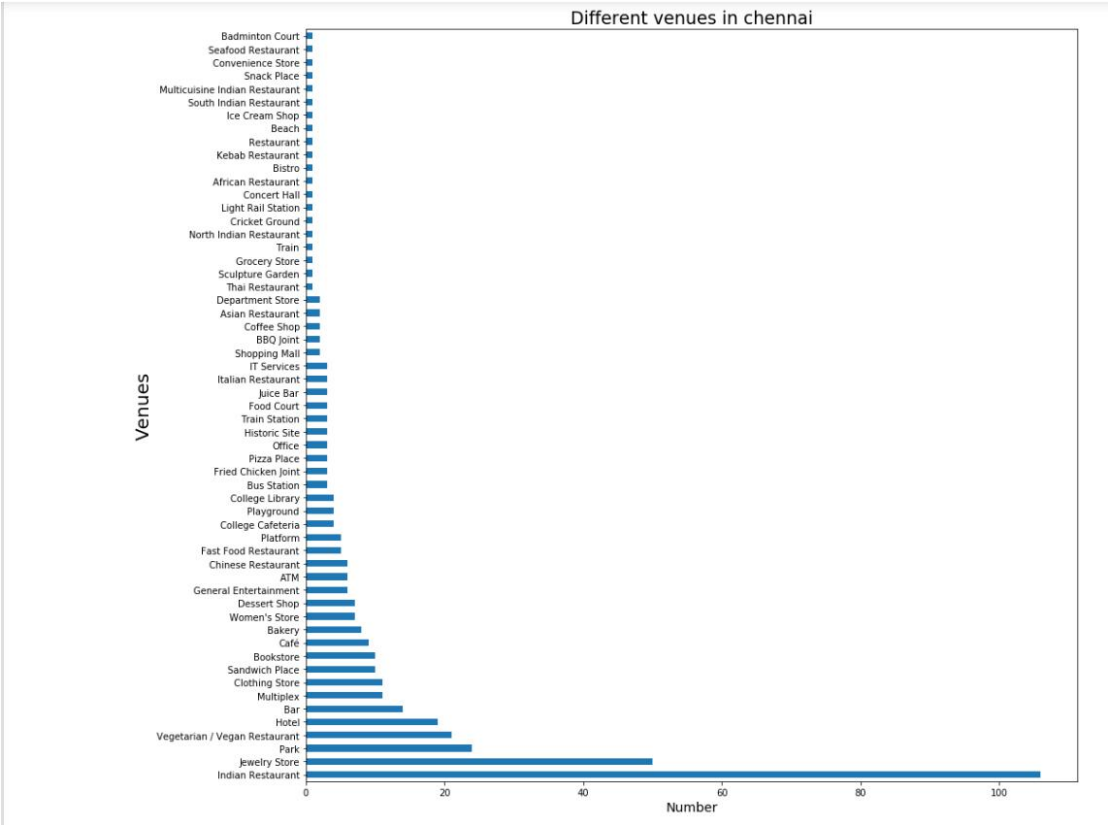
Mumbai



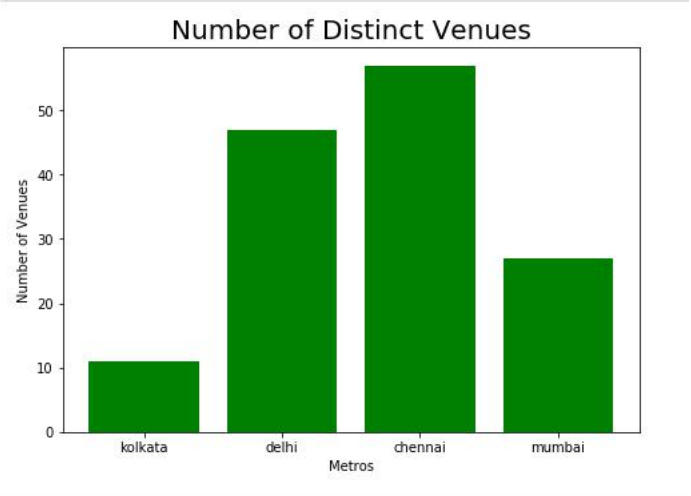
Kolkata



Chennai



We compare the different numbers of venues in the 4 metro cities in order to gauge the distinct activities one can engage in each of the cities



Next, we find out the top 10 most common venues for each neighborhood in each city. Then we cluster the neighborhoods based on the shared common venues.

The algorithm we use is **K Means clustering**

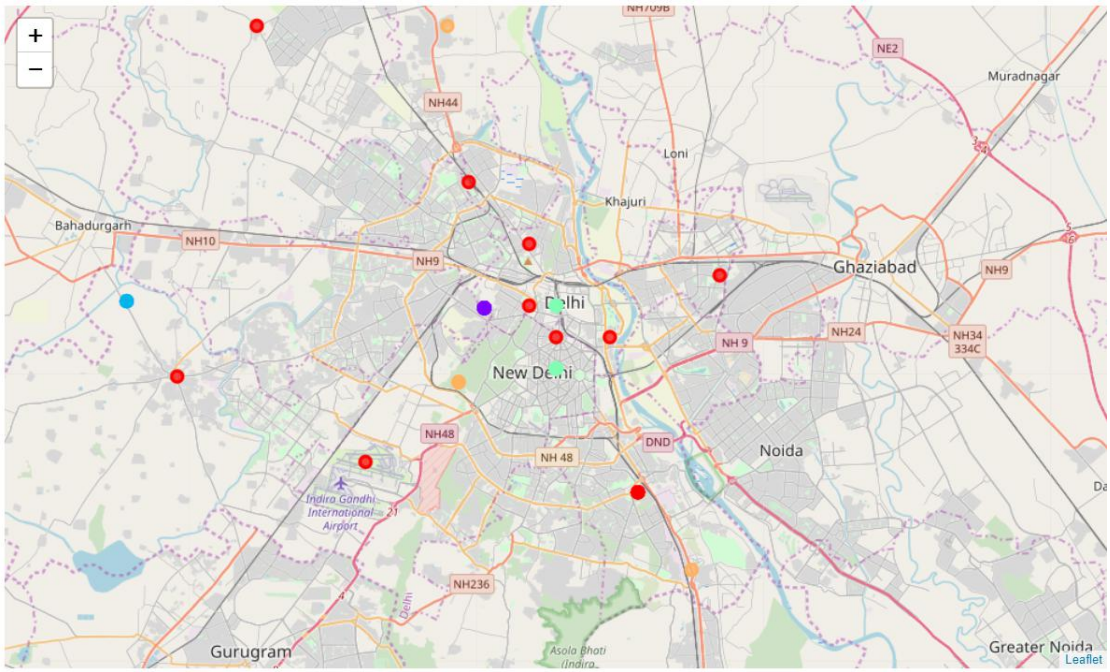
***k*-means clustering** is a method of unsupervised learning. *k*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*-Means minimizes within-cluster variances

This algorithm is to be run on numerical data. Hence, we have to run one hot encoding

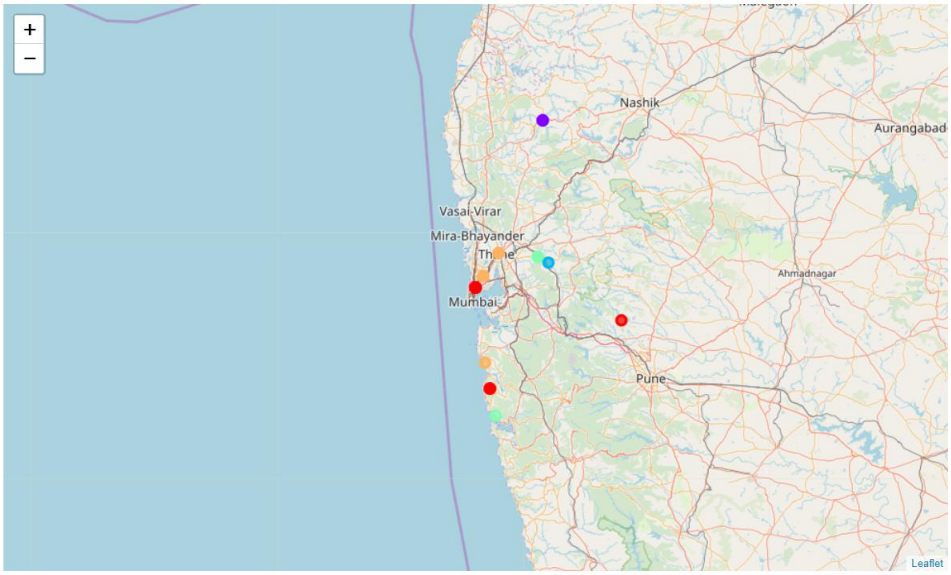
One-hot encoding is used in machine learning as a method to quantify categorical data. In short, this method produces a vector with length equal to the number of categories in the data set. If a data point belongs to the *i*th category then components of this vector are assigned the value 0 except for the *i*th component, which is assigned a value of 1. In this way one can keep track of the categories in a numerically meaningful way.

We create 5 clusters and visualize them

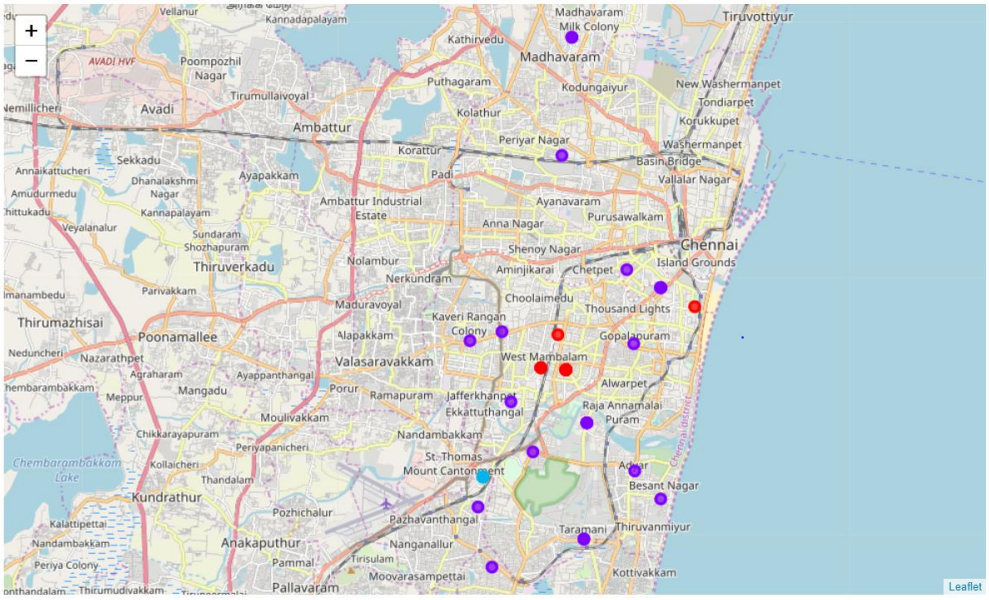
Delhi



Mumbai



CHENNAI



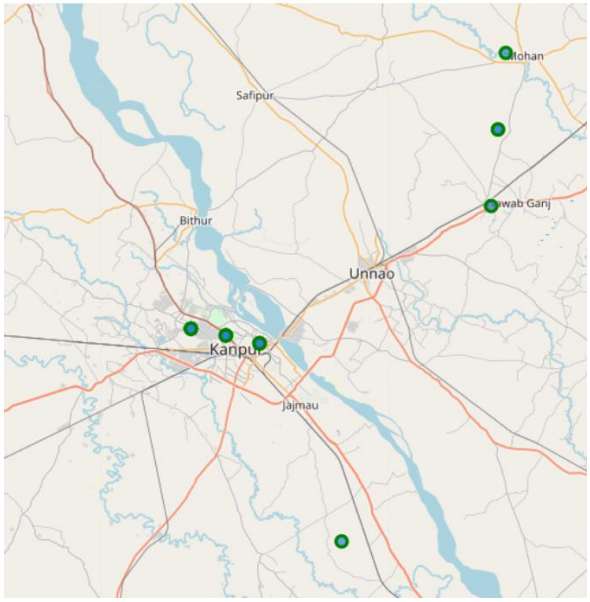
We analyse the clusters to obtain a better picture for comparing the metros.

Analysing Clusters

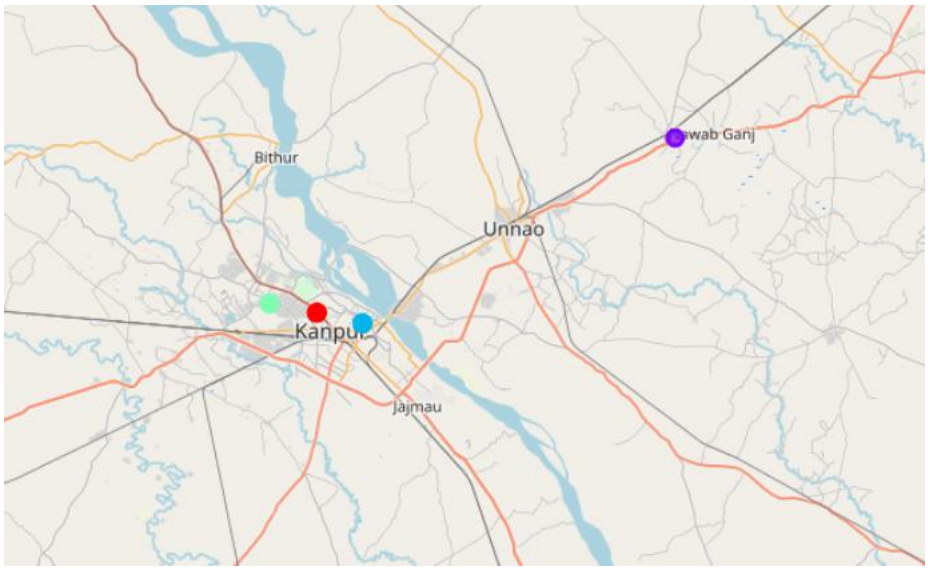
Cluster	Delhi	Chennai	Mumbai	Kolkata
0	Tourist places / Historic sites	Parks / Markets	Parks/ Cafe's	Train Stations, Cafe's
1	Bars / Hotels / cafe's	Restaurants / bars / entertainment	Scenic Spots/ bars	Train Stations, Bus Stations
2	Countryide	Hotels / bars / entertainment	Market/food joints	Markets, business services
3	Hotels / Restaurants	Food Joints	Restaurants, Flea Markets, Beaches	-
4	Train station / markets/ sports complexes	Hhistoric Places / food joints	Playgrounds / bars	-

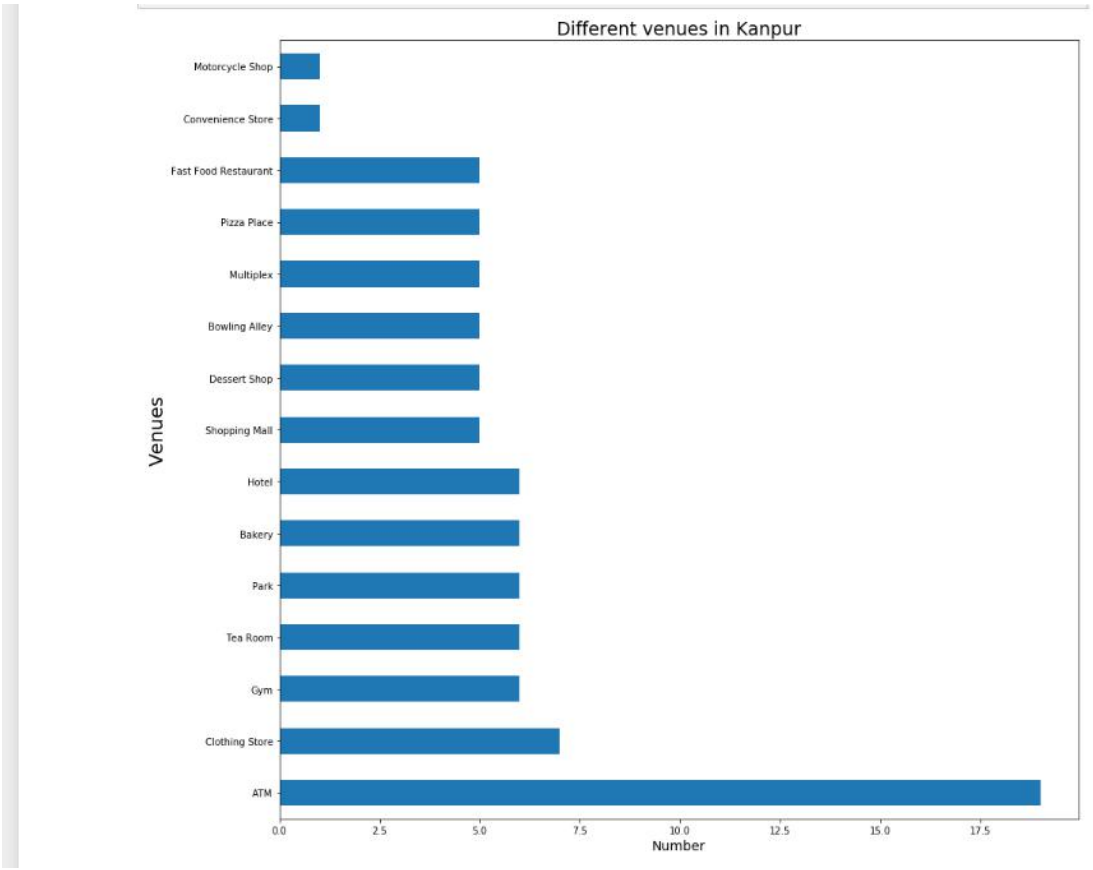
Now, we introduce a non metropolitan city, Kanpur, from the state of Uttar Pradesh

NEIGHBORHOODS



CLUSTERED NEIGHBORHOODS





COMPARISON OF CLUSTERS WITH METROPOLITAN CITIES

Cluster	Delhi	Chennai	Mumbai	Kolkata	Kanpur
0	Tourist places / Historic sites	Parks / Markets	Parks/ Cafe's	Train Stations, Cafe's	Tea Places, Parks
1	Bars / Hotels / cafe's	Restaurants / bars / entertainment	Scenic Spots/ bars	Train Stations, Bus Stations	Stores
2	Countryide	Hotels / bars / entertainment	Market/food joints	Markets, business services	Malls
3	Hotels / Restaurants	Food Joints	Restaurants, Flea Markets, Beaches	-	Bakeries
4	Train station / markets/ sports complexes	Hhistoric Places / food joints	Playgrounds / bars	-	Market

RESULTS:

->Chennai has the most number of distinct venues;ie, amenities and activities provided to its residents, followed by Delhi, Mumbai and Kolkata.

->Chennai has maximum number of family restaurants, jewellery stores,cafe's and open grounds.
->Kolkata has a decent number of cafe's
->Mumbai has maximum numbers of restaurants,bars, and general entertainment amenities
->Delhi has a plethora of fast food restaurants, hotels, bars, coffee shops and historic sites.

->The clusters can be analysed as the following

Cluster	Delhi	Chennai	Mumbai	Kolkata	Kanpur
0	Tourist places / Historic sites	Parks / Markets	Parks/ Cafe's	Train Stations, Cafe's	Tea Places, Parks
1	Bars / Hotels / cafe's	Restaurants / bars / entertainment	Scenic Spots/ bars	Train Stations, Bus Stations	Stores
2	Countryide	Hotels / bars / entertainment	Market/food joints	Markets, business services	Malls
3	Hotels / Restaurants	Food Joints	Restaurants, Flea Markets, Beaches	-	Bakeries
4	Train station / markets/ sports complexes	Hhistoric Places / food joints	Playgrounds / bars	-	Market

->Kanpur,a non-metropolitan city,had a measly 21 unique venues,centred around departmental stores and malls.

DISCUSSIONS and CONCLUSIONS:

->From the above results, we can conclude that Chennai ,Kolkata would be a good place to move in with families, whereas Mumbai and Delhi suit to bachelors more
->Chennai can be termed as the most versatile city, based on the amenities, and Kolkata, the least.

->For a person moving to Chennai, neighborhoods falling in clusters 0,1,3 are good for families to settle.
Clusters 1,2,3 are good for bachelors
Clusters 0,4 are good for business owners

->For a person moving to Mumbai, neighborhoods falling in clusters 0,2,3,4 are good for families to settle.
Clusters 1,3,4 are good for bachelors
Clusters 0,1,3 are good for business owners

->For a person moving to Delhi, neighborhoods falling in clusters 2,4 are good for families to settle.
Clusters 1,4 are good for bachelors
Clusters 0,3,4 are good for business owners

->For a person moving to Kolkata, neighborhoods falling in clusters 0,1 are good to settle.
Cluster 2 is good for business owners

->We see that clusters in a non metro city are restricted to mundane locations like malls and stores and parks, lacking the large number of hotels,business services and entertainment services provided by metro cities. It is safe to conclude that to be a metro city, the neighborhoods must be developed to cater to the demands of high populations and giving a better quality of living through high end venues like entertainment services, bars, restaurants, and numerous business services.Also, having tourist place or a well developed IT Park gives them a boost as they have a large income of foreign travelers too, giving a boost to local businesses and job seekers respectively.

FUTURE SCOPE:

We can further add data regarding the housing prices and extra costs incurred in living certain neighborhoods for each city, in order to classify neighborhoods based on their average cost of living. This would help people in choosing the neighborhood in a particular city based on their income.

Also, we can use matching algorithms to math the features of other cities to the 4 metropolitans, studying which city is most likely to become a metropolitan in coming years. This would be good in investing purposes.