A Tale of Two Cities!

An analysis for comparison of cities using Foursquare data and Machine Learning

- Rohit Sinha

Jul 2, 2021

Introduction

Description & Discussion of the Background

Picking a city, when it comes to **Delhi** and **Mumbai** is always a hard decision as both these cities are truly multicultural, and cosmopolitan cities found in one of fastest developing Nation, India. Along with being two of India's most important Financial and political centres, they are major centres for commerce, sciences, fashion, arts, culture and gastronomy. Both Delhi (officially the National Capital Territory (NCT) of Delhi) and Mumbai (capital city of the Indian state of Maharashtra) have a rich history and are two of the most visited and sought-after cities in India. Mumbai is the second-most populous city in the country after Delhi (11 million) and the seventh-most populous city in the world with a population of roughly 20 million. Mumbai lies on the Konkan coast on the west coast of India and has a deep natural harbour. Delhi, is a city and a union territory of India containing New Delhi, the capital of India. It is bordered by the state of Haryana on three sides and by Uttar Pradesh to the east.

Our goal is to perform a comparison of the two cities to see how similar or dissimilar they are. Such techniques allow users to identify similar neighbourhoods among cities based on amenities or services being offered locally, and thus can help in understanding the local area activities, what are the hubs of different activities, how citizens are experiencing the city, and how they are utilising its resources.

What kind of clientele would benefit from such an analysis?

A potential job seeker with transferable skills may wish to search for jobs in selective cities which provide the most suitable match for their

qualifications and experience in terms of salaries, social benefits, or even in terms of a culture fit for expats.

- Further, a person buying or renting a home in a new city may want to look for recommendations for locations in the city similar to other cities known to them.
- Similarly, a large corporation looking to expand its locations to other cities might benefit from such an analysis.
- Many within-city urban planning computations might also benefit from modelling a city's relationship to other cities.

Data Description

To consider the problem we can list the datas as below:

- I found the location data on Mumbai and Delhi's neighborhood using Wikipedia and Geosquare library.
- I used **Forsquare API** to get the most common venues of given Area of the cities.
- I used Google Map, 'Search Nearby' option to get the center coordinates of the each location which was incorrect.

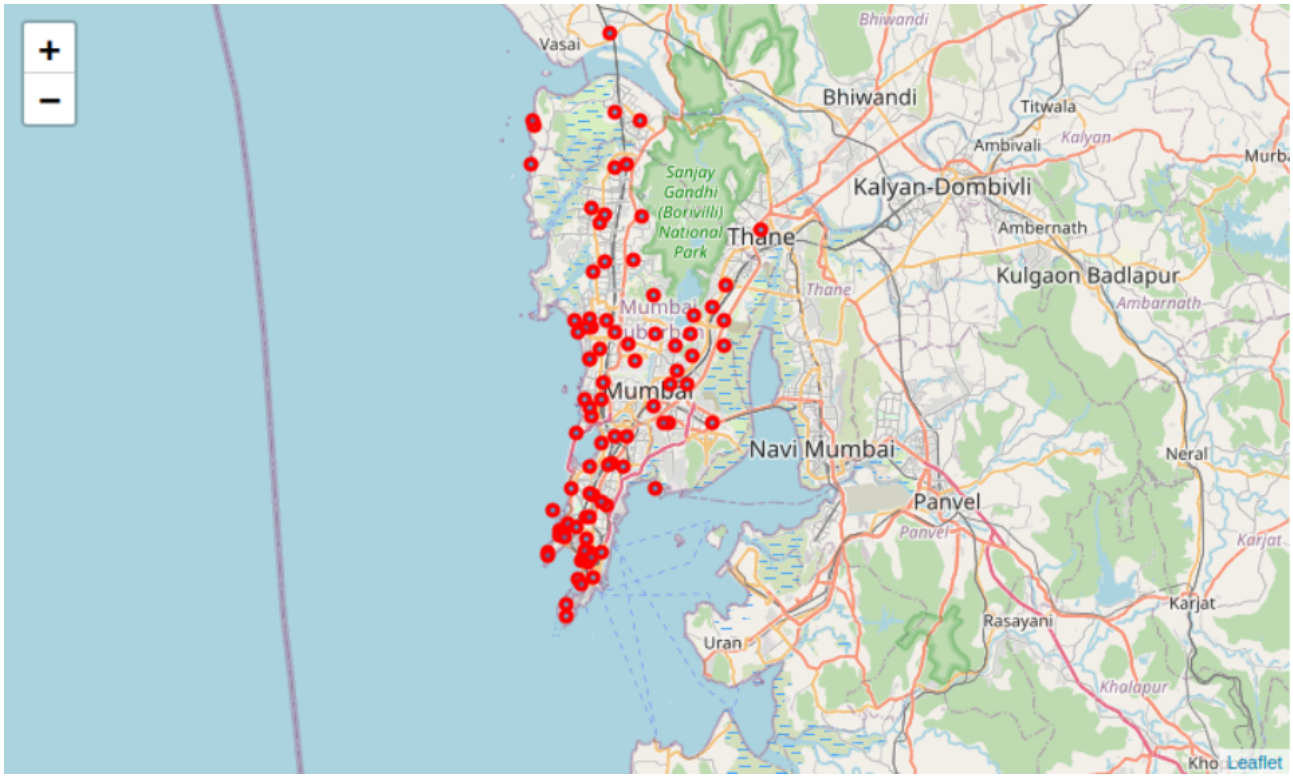
Methadology

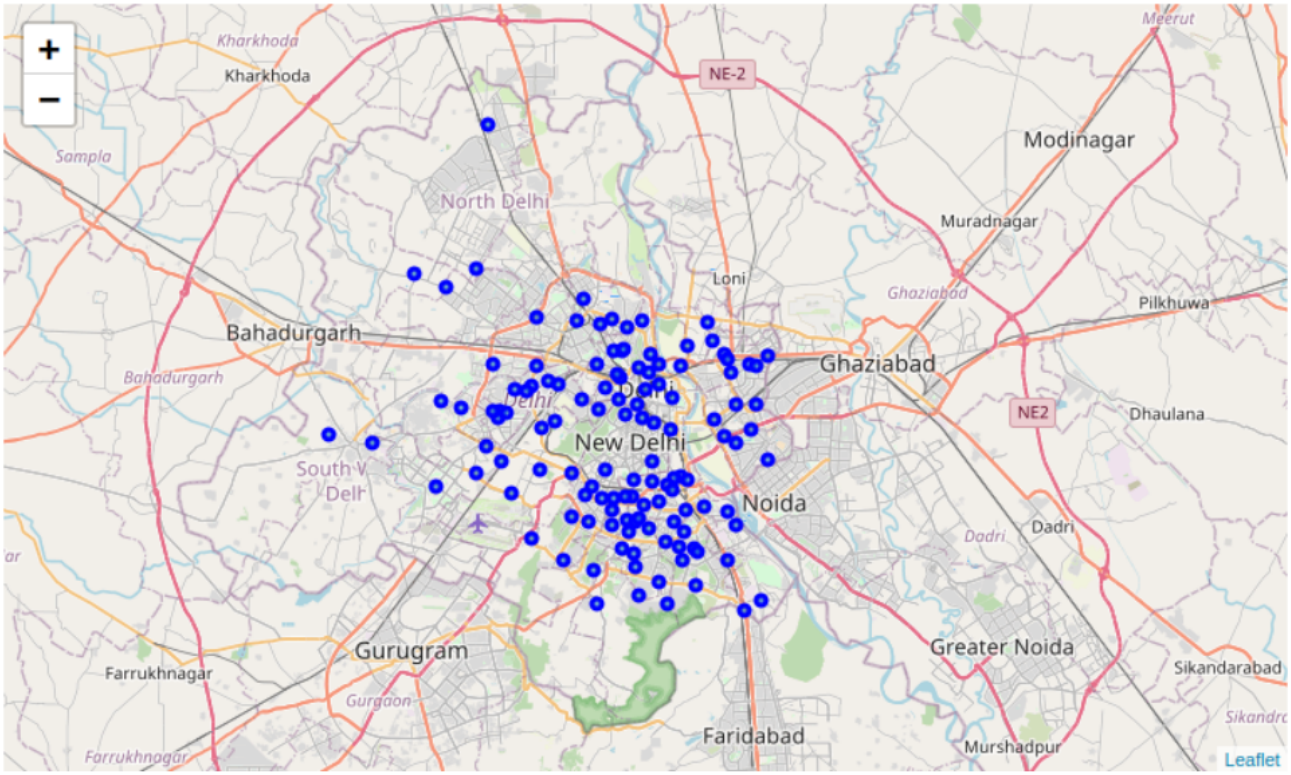
As a database, I used GitHub repository in my study. My master data which has the main components Area, *Location, Latitude* and *Longitude* informations of the city.

	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

	Area	Location	Latitude	Longitude
0	Adarsh Nagar	North West Delhi	28.716580	77.170422
1	Ashok Vihar	North West Delhi	28.699453	77.184826
2	Begum Pur	North West Delhi	28.725503	77.058371
3	Karala	North West Delhi	28.735140	77.032511
4	Narela	North West Delhi	28.842610	77.091835

I used python **folium** library to visualize geographic details of Mumbai and Delhi and its areas and I created a map of of the cities with areas superimposed on top. I used latitude and longitude values to get the visual as below:





Foursquare location data gave information about the list of venues within a 1 km radius of each borough. This is a reasonable distance to understand the characteristics of the neighbourhood.

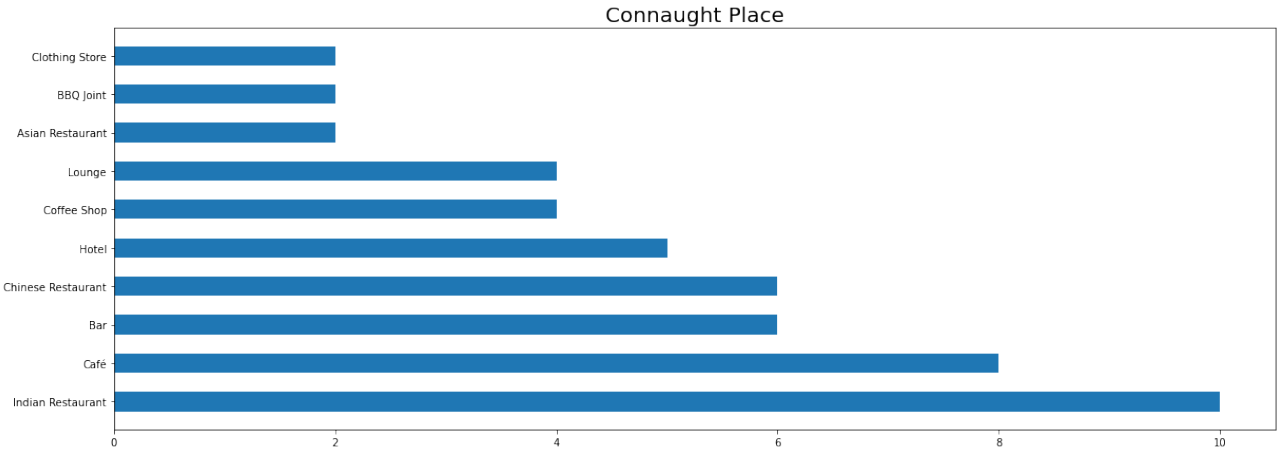
	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Amboli	19.1293	72.8434	Cafe Arfa	19.128930	72.847140	Indian Restaurant
1	Amboli	19.1293	72.8434	Jaffer Bhai's Delhi Darbar	19.137714	72.845909	Mughlai Restaurant
2	Amboli	19.1293	72.8434	5 Spice , Bandra	19.130421	72.847206	Chinese Restaurant
3	Amboli	19.1293	72.8434	Domino's Pizza	19.131000	72.848000	Pizza Place
4	Amboli	19.1293	72.8434	Shetty's Corner	19.124845	72.837858	Chinese Restaurant

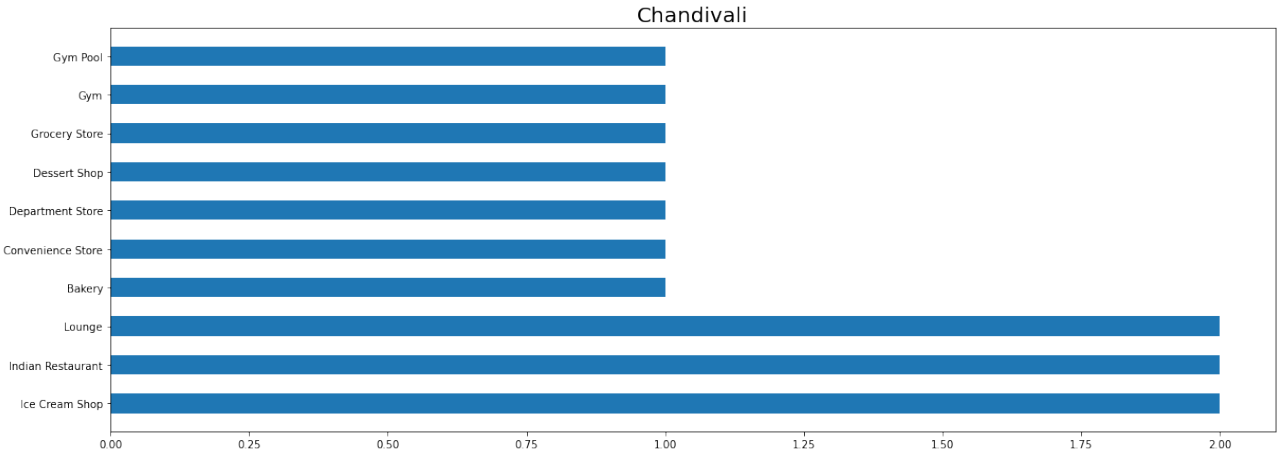
	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adarsh Nagar	28.716580	77.170422	Vishyavidyalaya Metro Station@Entry gate #1 n ...	28.715596	77.170981	Train Station
1	Adarsh Nagar	28.716580	77.170422	Adarsh Nagar Metro Station	28.716598	77.170436	Light Rail Station
2	Adarsh Nagar	28.716580	77.170422	Pahalwan Dhaba	28.714594	77.172155	Indian Restaurant
3	Adarsh Nagar	28.716580	77.170422	Giani's	28.717900	77.173907	Ice Cream Shop
4	Adarsh Nagar	28.716580	77.170422	Cues N Ball	28.711260	77.176680	Pool Hall

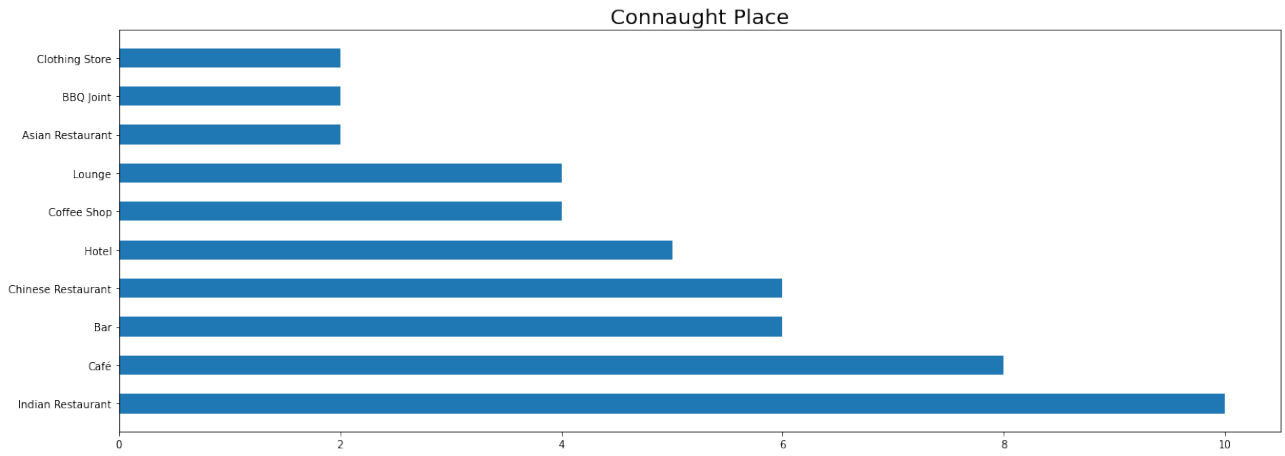
In summary of this data **1865** venues were returned by Foursquare for Mumbai and **1348** Venues for Delhi.

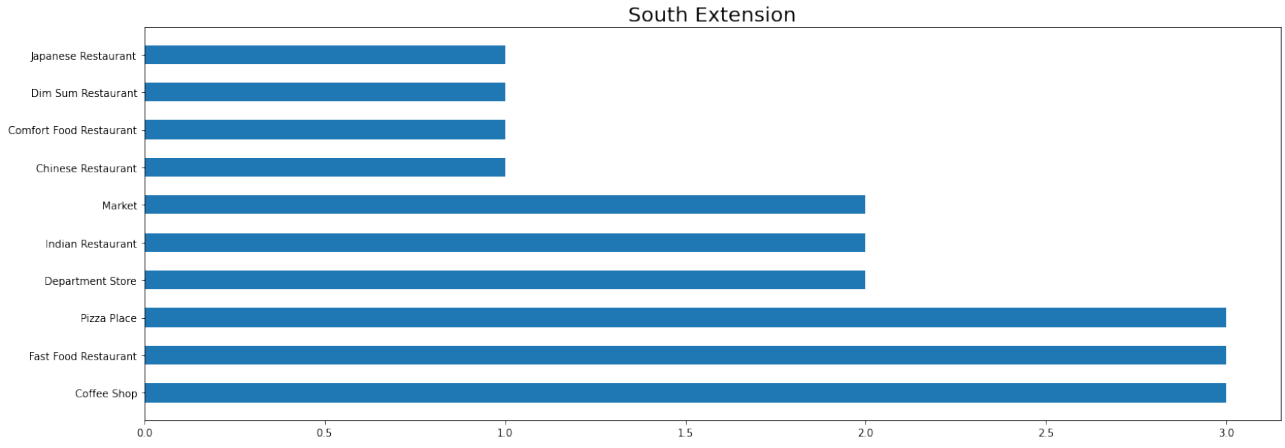
We can see that some areas how reached the **100** limit of venues. On the other hand; some areas are below **20** venues in our given coordinates with Latitude and Longitude, in below graph.

The result doesn’t mean that inquiry run all the possible results in boroughs. Actually, it depends on given Latitude and Longitude informations and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighborhood informations with more Latitude and Longitude informations.


Connaught Place
Clothing Store
BBQ Joint
Asian Restaurant
Lounge
Coffee Shop
Hotel
Chinese Restaurant
Bar
Café
Indian Restaurant
0
2
4
6
8
10


Chandivali
Gym Pool
Gym
Grocery Store
Dessert Shop
Department Store
Convenience Store
Bakery
Lounge
Indian Restaurant
Ice Cream Shop
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00





In order to explore the venue data in a more comprehensive way and further use it for analysis, foursquare venue data was arranged into pandas data frame as follows:

- First, create a data-frame with pandas one hot encoding for each of the venue categories
- Obtain the mean of each one-hot encoded venue categories using pandas groupby method on the borough column
- Use the venue category mean to obtain a venue based data frame for each city giving the ten most common venues for each borough

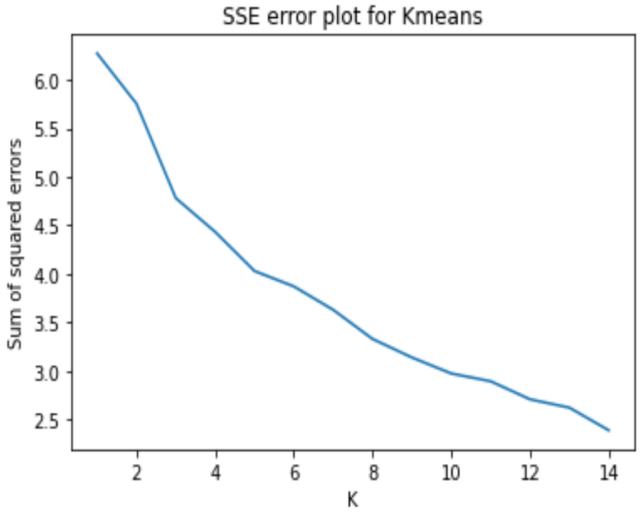
In summary of this graph **221** unique categories for Mumbai and **205** unique categories were returned by Foursquare, then I created a table which shows list of top 10 venue category for each Area in below table.

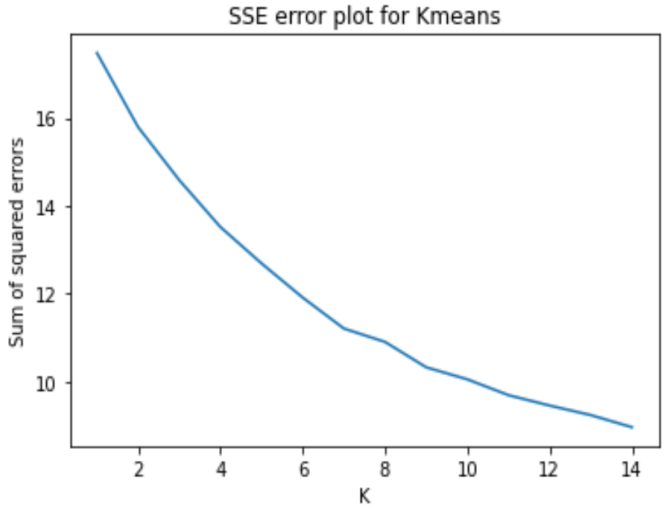
	Cluster Labels	Area	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	2	Aarey Milk Colony	Indian Restaurant	Gym / Fitness Center	Restaurant	Resort	Golf Course	Hotel	Farm	Café	Zoo	Dim Sum Restaurant
1	0	Agripada	Indian Restaurant	Bakery	Coffee Shop	Gym	Zoo	Golf Course	Pizza Place	Nightclub	Movie Theater	Cupcake Shop
2	2	Altamount Road	Bakery	Chinese Restaurant	Bar	Indian Restaurant	Café	Fast Food Restaurant	Coffee Shop	Sandwich Place	Pizza Place	Snack Place
3	2	Amboli	Indian Restaurant	Bar	Asian Restaurant	Pizza Place	Coffee Shop	Chinese Restaurant	Pub	Falafel Restaurant	Bowling Alley	Smoke Shop
4	2	Amrut Nagar	Café	Lounge	Indian Restaurant	Clothing Store	Fast Food Restaurant	Pizza Place	Electronics Store	Vegetarian / Vegan Restaurant	Diner	Chinese Restaurant

	Cluster Labels	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 Common Venue	10th Most Common Venue
0	0	Adarsh Nagar	Pool Hall	Mobile Phone Shop	Train Station	Ice Cream Shop	Convenience Store	Indian Restaurant	Light Rail Station	Fried Chicken Joint	French Restaurant	Frozen Yogurt Shop
1	0	Alaknanda	Coffee Shop	Restaurant	Indian Restaurant	Market	BBQ Joint	Convenience Store	Dessert Shop	Plaza	Sandwich Place	Park
2	0	Anand Vihar	Clothing Store	Hotel	Multiplex	Pizza Place	Wine Shop	Movie Theater	Bus Station	Shopping Mall	Café	Garden
3	0	Ashok Nagar	Fast Food Restaurant	Indian Restaurant	Pizza Place	Donut Shop	Coffee Shop	Restaurant	Café	Shopping Mall	Multiplex	Clothing Store
4	0	Ashok Vihar	Donut Shop	Asian Restaurant	Snack Place	Frozen Yogurt Shop	Sandwich Place	South Indian Restaurant	Coffee Shop	Fast Food Restaurant	Indian Restaurant	Pizza Place

I performed a clustering analysis using the ‘k-means’ algorithm in order to categorise similar areas into clusters based on the similarities provided by the venue categories. To gain some understanding, I decided to do some investigation into the number of clusters (k) to be used as follows:

Elbow method: I tried to determine the effective number of clusters (k) using the elbow method for Mumbai clustering analysis and saw a small kink around k = 8 (although not clear and sharp). The elbow method uses Within-Cluster-Sum of Squared Errors (WSS) for different values of k and one can choose the value of k for which WSS starts to diminish and can be seen as an elbow in the WSS-versus-k plot. Similarly for Delhi, the kink can be noticed at k=5. I decided to categorise Mumbai areas into 8 set of clusters and Delhi neighbourhoods into 5 set of clusters for the purpose of our analysis. It might be useful to look into a more detailed analysis to optimise k in the future for such studies.

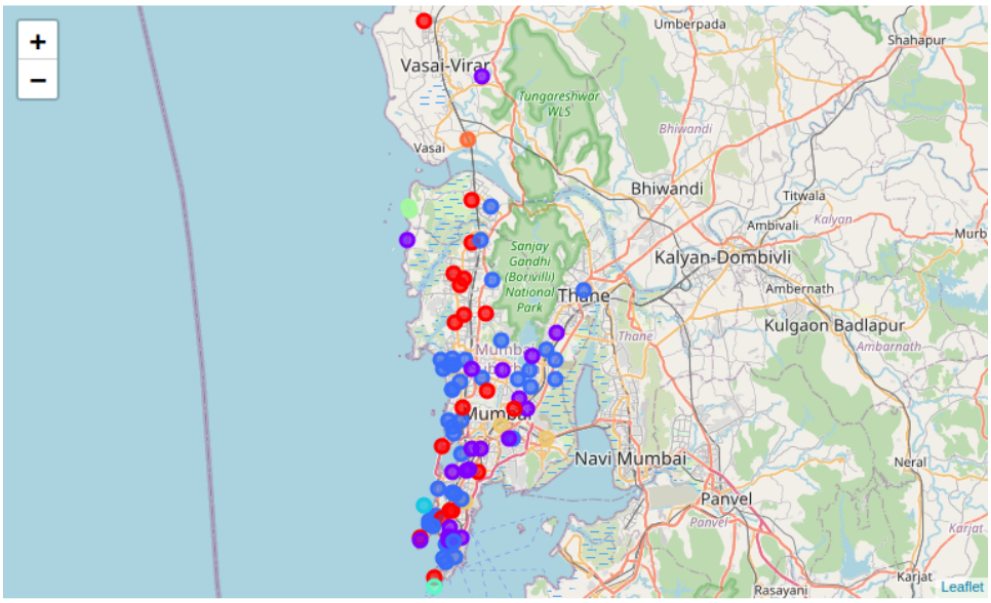


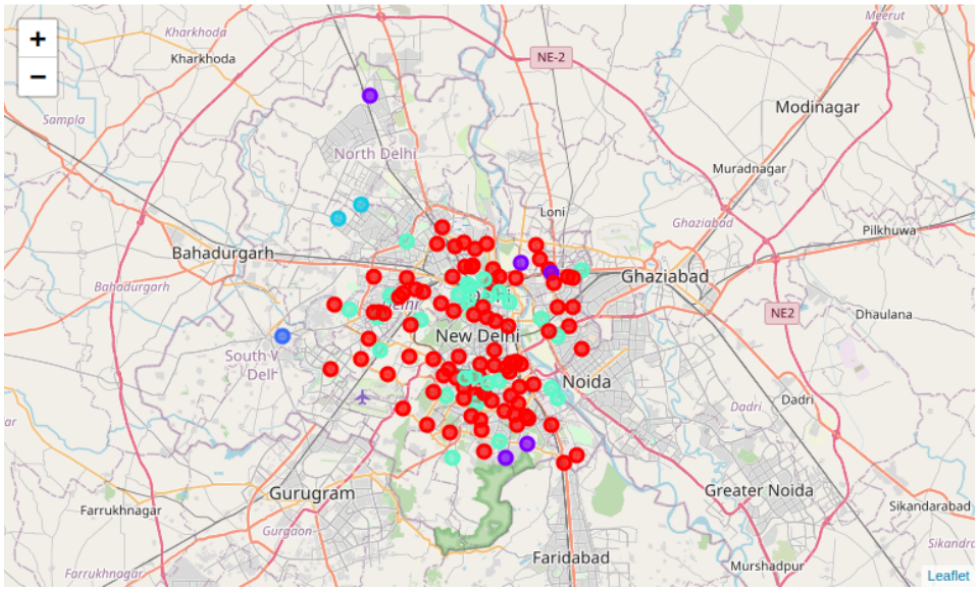


Here is my merged table with cluster labels for each Location.

	Area	Location	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6 C
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400	2	Indian Restaurant	Bar	Asian Restaurant	Pizza Place	Coffee Shop	Re
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833	2	Indian Restaurant	Hotel	Seafood Restaurant	Vegetarian / Vegan Restaurant	Café	
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373	2	Bar	Pizza Place	Pub	Gym / Fitness Center	Vegetarian / Vegan Restaurant	
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210	2	Pub	Café	Vegetarian / Vegan Restaurant	Bar	Lounge	
4	Lokhandwala	Andheri,Western Suburbs	19.130815	72.829270	2	Bar	Pub	Coffee Shop	Pizza Place	Indian Restaurant	

Lets Visualize our inital map with thr cluster Labels in place.





Discussions of Results

Some of the inferences which were drawn from the explanatory analysis are:

Initial exploration of the Foursquare venue data revealed that Indian Restaurants, cafes, pubs and juice bars are the most common venues in main areas in Mumbai. Similarly Clothing, food restaurants and hotels were the most common venues seen in main areas of Delhi.

Further, machine learning analysis of the venue based data revealed most of the areas of Mumbai can be grouped together into one cluster. The most common venues in such boroughs were always cafes, pubs, shops or restaurants followed by some kind of clothing, convenience stores or pharmacies. Delhi was categorised into five separate clusters in total with two of its clusters amounting for the majority. Although the most common venue in both the clusters was always a Indian restaurant, it was followed by a high number of Italian restaurants, hotels, and cafes in the first cluster and variations of other cuisine restaurants, bars, bistros, clothing stores or supermarkets in the subsequent cluster.

The most common type of venues in either of the cities are mostly restaurants, cafes, hotels, pubs/bars, clothing stores or parks. This in a way highlights that how similar the cities of Mumbai and Delhi are in terms of services being offered.

> One can further use the venue data to compare the cities is a more comprehensive way where one can also explore different levels of spatial aggregation, namely grids, neighbourhoods, and the city as a whole. The level of spatial aggregation can be an important factor when characterising a city in terms of its venues.

Some of the questions one can answer with different levels of spatial aggregation could be:

- How are the venue categories distributed inside an area, i.e., is the area more of a residential or a commercial one.
- Which city has the highest number of each of the amenities (bars, restaurants, parks, universities, libraries, shopping centres, etc.)

Conclusion

To summarise, analysing cities using venue based data from Foursquare lead to an overall understanding of the type of venues in each area and presented some of the key features of the cities but the level of data is not adequate to provide a comprehensive analysis for a city-to-city comparison. For a potential interested person (job-seeker or person deciding to move to either of the cities) or a bigger clientele like a business corporation or city planners, one would need to do a more detailed analysis adding features such as rents, salaries, transportation, cost of living, growth rate, economy, etc.

The capstone project provided a medium to understand in depth about how real life data science projects work and what all steps go in building a data science methodology. All steps from understanding the business problem, data understanding to data preparation, and model building were discussed in detail here. Many drawbacks of the current analysis and further ways to improve the analysis were also mentioned. This was an initial attempt to understand and solve the business problem at hand. However, there still exists a huge potential to extend this project in real life scenarios.

References

List of neighbourhoods in Mumbai - Wiki

List of neighbourhoods in Delhi - Wiki

Foursquare

Google Maps