Comparative view of 2 large metropolitan cities of India: Chennai & Mumbai

Shubhashish Banerjee, 21/June/2021

Index

1. *Introduction*

Chennai and Mumbai - both are large coastal metropolitan cities in India with lot of history. Having spent time in both of these cities, it's a pleasure to use data science techniques to illustrate similarities and differences within these cities. Both the cities have their unique lifestyle, types of industries, economical drivers and lifestyle of people.

2. *Business problem description*

The comparison will show how these two large metropolitan cities are different and what are the similarities. We will use available data from the web to identify major clusters in the city and use that information to create a point of view about the kind of economy may flourish or be better suited for these two cities.

3. *Data handling*

There are 2 sources of data that would be used for this study:

a. **Location, Area and their geographical coordinates:** This data would be sourced from Wiki location pages of both cities. This information is present in tabular forms on both wiki pages. We will scrape it and use that to join with the 2nd data item. We will use the location information to generate map visual to visualize the spread and then to show clusters within each city.
   i. Chennai Areas
   ii. Mumbai Areas

b. **FourSquare venue locations:** This data would be sourced from FourSquare by querying thru a developer API. This will be joined with the data item #1 for analysis. After joining, the collected data would be transformed and encoded so that k-Means (from sklearn library) algorithm can be used to bifurcate clusters within each city.
   i. Developer API of FourSquare

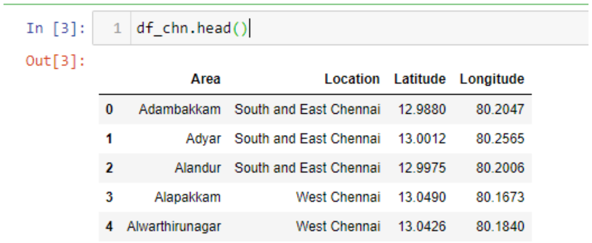
4. *Methodology*

a) Data ETL – Extraction, Transformation and Loading is done through **Requests**, **NumPy** & **Pandas**.

Dataframe were made from Wikipedia links which had area and geo coordinates in a tabular structure.

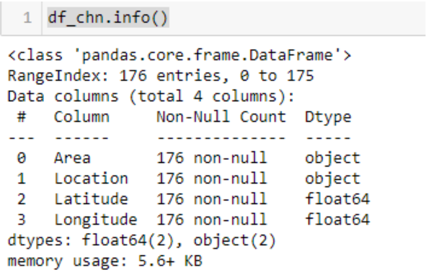
Data import from the wiki pages came out easily and in clean format as wiki had the area/neighborhoods information already tabulated. Geographical coordinates were also present in the same table so it saved additional steps of using geocoders to get spatial coordinates.

Chennai dataframe:



```
In [3]: df_chn.head()
Out[3]:
```

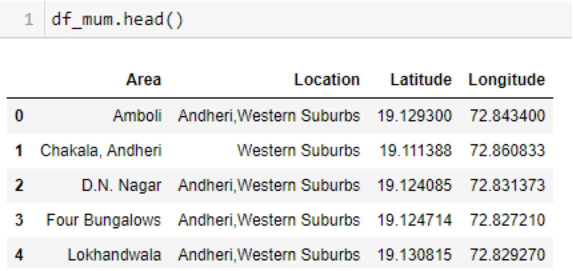
	Area	Location	Latitude	Longitude
0	Adambakkam	South and East Chennai	12.9880	80.2047
1	Adyar	South and East Chennai	13.0012	80.2565
2	Alandur	South and East Chennai	12.9975	80.2006
3	Alapakkam	West Chennai	13.0490	80.1673
4	Alwarthirunagar	West Chennai	13.0426	80.1840



```
df_chn.info()
```

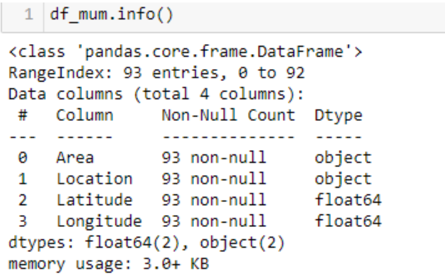
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 176 entries, 0 to 175
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Area       176 non-null    object 
 1   Location   176 non-null    object 
 2   Latitude   176 non-null    float64
 3   Longitude  176 non-null    float64
dtypes: float64(2), object(2)
memory usage: 5.6+ KB
```

Mumbai dataframe



```
df_mum.head()
```

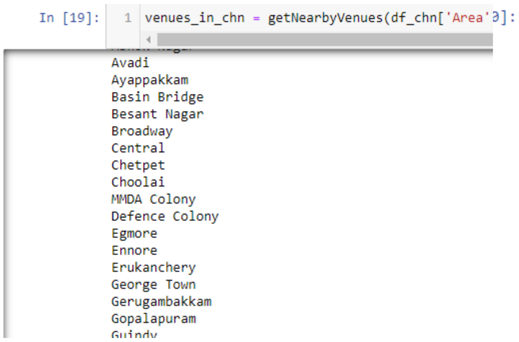
	Area	Location	Latitude	Longitude
0	Amboli	Andheri,Western Suburbs	19.129300	72.843400
1	Chakala, Andheri	Western Suburbs	19.111388	72.860833
2	D.N. Nagar	Andheri,Western Suburbs	19.124085	72.831373
3	Four Bungalows	Andheri,Western Suburbs	19.124714	72.827210
4	Lokhandwala	Andheri,Western Suburbs	19.130815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



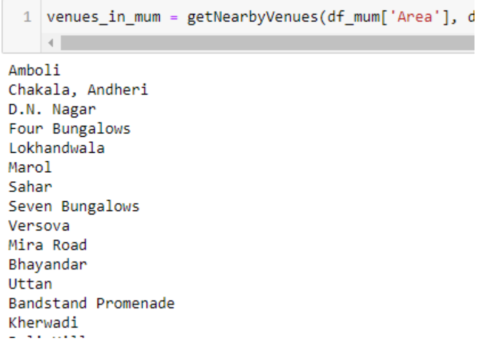
```
df_mum.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93 entries, 0 to 92
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Area       93 non-null     object 
 1   Location   93 non-null     object 
 2   Latitude   93 non-null     float64
 3   Longitude  93 non-null     float64
dtypes: float64(2), object(2)
memory usage: 3.0+ KB
```

FourSquare venue information could be fetched easily since both these cities are large metropolitans and have been charted on FourSquare service. Venue information had relevant detailing for us to proceed with the study.



```
In [19]:   venues_in_chn = getNearbyVenues(df_chn['Area'0]:

Avadi
Ayappakkam
Basin Bridge
Besant Nagar
Broadway
Central
Chetpet
Choolai
MMDA Colony
Defence Colony
Egmore
Ennore
Erukanchery
George Town
Gerugambakkam
Gopalapuram
Guindy
```



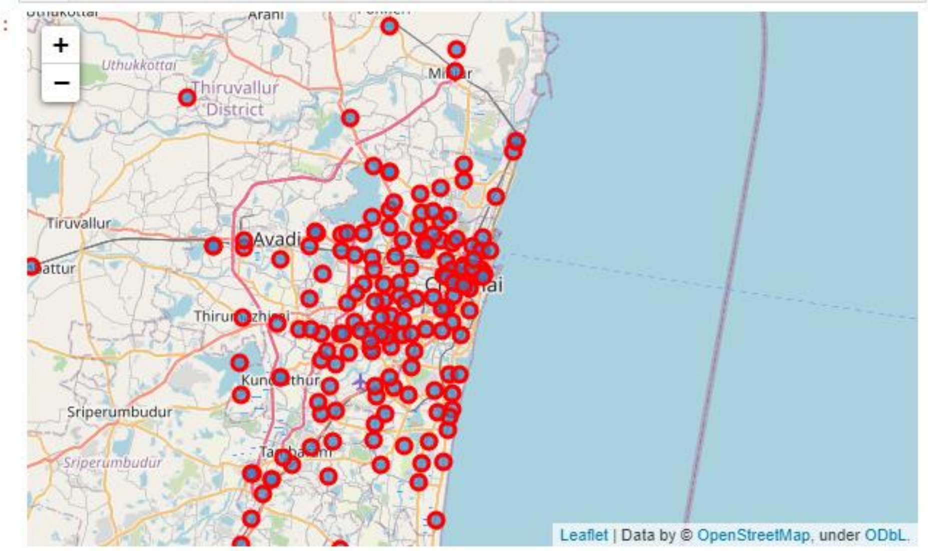
```
venues_in_mum = getNearbyVenues(df_mum['Area'], d

Amboli
Chakala, Andheri
D.N. Nagar
Four Bungalows
Lokhandwala
Marol
Sahar
Seven Bungalows
Versova
Mira Road
Bhayandar
Uttan
Bandstand Promenade
Kherwadi
```

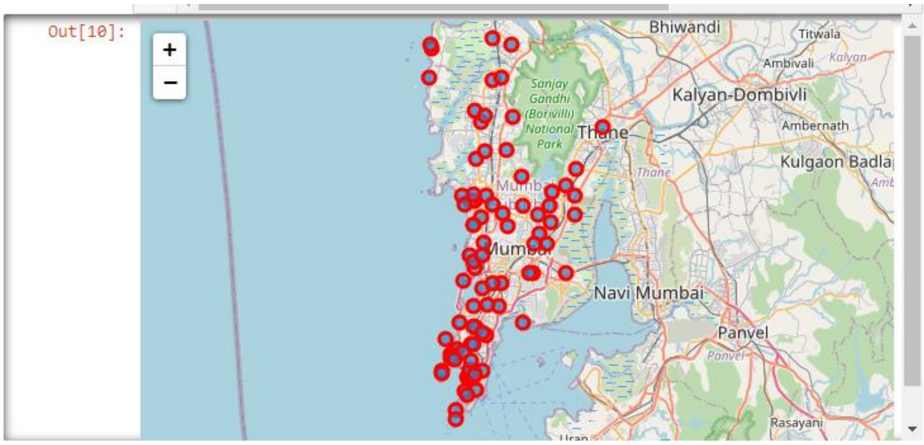
b) Modelling has been done with **Sklearn**.

Why k-means algorithm?
K-means is an unsupervised machine learning algorithm, while studying similarities and patterns of the datasets of two cities k-means would not require any explicit training in clustering and we would be able to make the clusters without any bias involved, since the chosen cities would have their own uniqueness. Furthermore, by using k-means, this code can be made to compare any cities across the world by own changing the input dataframe(s) - area/location information & geographical coordinates information.

c) Map visualization is done with **Folium** and **Matplotlib**.

Map visualization was used in the beginning to check data accuracy and spread of neighborhood as well as in the end after the venues were clustered.





Out[10]:

The sequenced steps of processing are as follows:

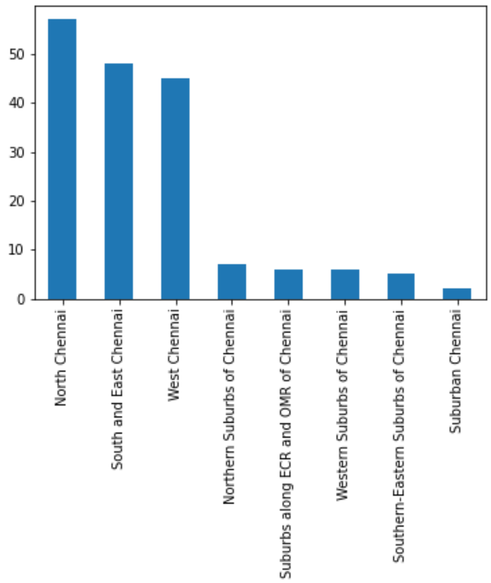
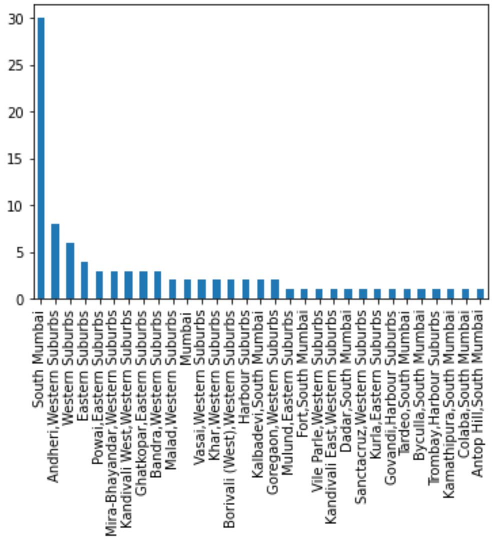
1. **Data Collection**
   - i) Importing dependent libraries for the study
   - ii) Simultaneous data gathering and process - Each step will be executed for both the cities - Chennai & Mumbai
   - iii) Using Pandas to read data from Wikipedia city pages
   - iv) Initial visual observation of scraped data from Wiki links
   - v) Map visualization to check accuracy of lat/long data
   - vi) Query venue data from FourSquare using a developer API
2. **Data Treatment**
   - vii) Grouping Venues by categories
   - viii) Transforming collected data using One Hot Encoding
   - ix) Combining neighborhoods names to venue data
   - x) Segregating most common and then top venues
3. **Data Modelling**
   - xi) Building model using K-Means
4. **Data Visualization**
   - xii) Visualization of various clusters in the city
   - xiii) Listing of 5 clusters for comparison of each city

5. *Results*

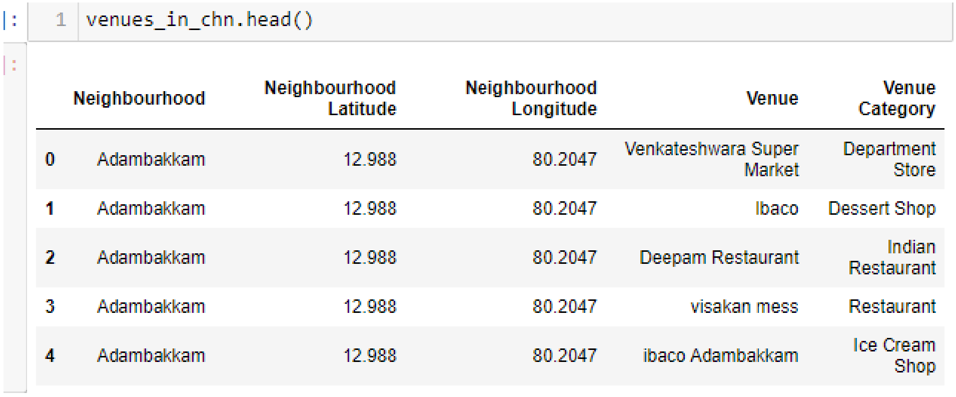
The results of the 4 stages (Data Collection Data Treatment, Data Modelling, & Data Visualization) of this study are given below:

1. Data Collection

Area/Neighborhood information from Wikipedia:

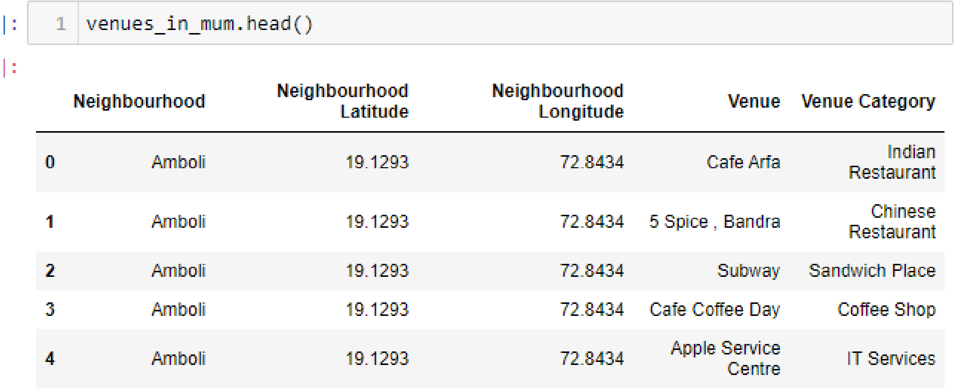
<table>
<tr><th>Mumbai</th><th>Chennai</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Venue information from FourSquare:



```
venues_in_chn.head()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Adambakkam	12.988	80.2047	Venkateshwara Super Market	Department Store
1	Adambakkam	12.988	80.2047	Ibaco	Dessert Shop
2	Adambakkam	12.988	80.2047	Deepam Restaurant	Indian Restaurant
3	Adambakkam	12.988	80.2047	visakan mess	Restaurant
4	Adambakkam	12.988	80.2047	ibaco Adambakkam	Ice Cream Shop

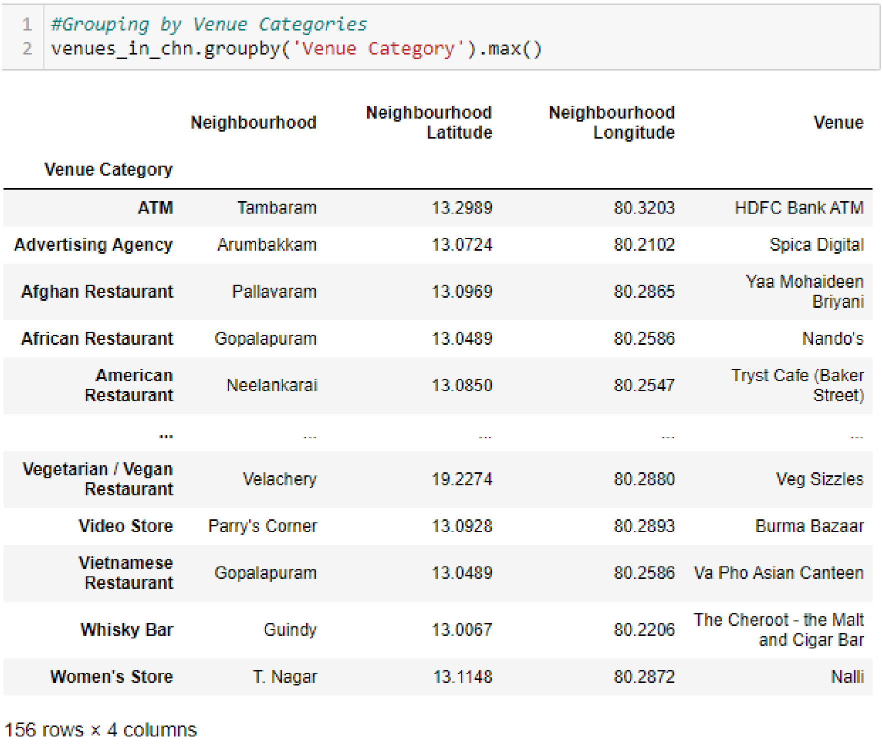


```
venues_in_mum.head()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Category
0	Amboli	19.1293	72.8434	Cafe Arfa	Indian Restaurant
1	Amboli	19.1293	72.8434	5 Spice , Bandra	Chinese Restaurant
2	Amboli	19.1293	72.8434	Subway	Sandwich Place
3	Amboli	19.1293	72.8434	Cafe Coffee Day	Coffee Shop
4	Amboli	19.1293	72.8434	Apple Service Centre	IT Services

2. Data Treatment

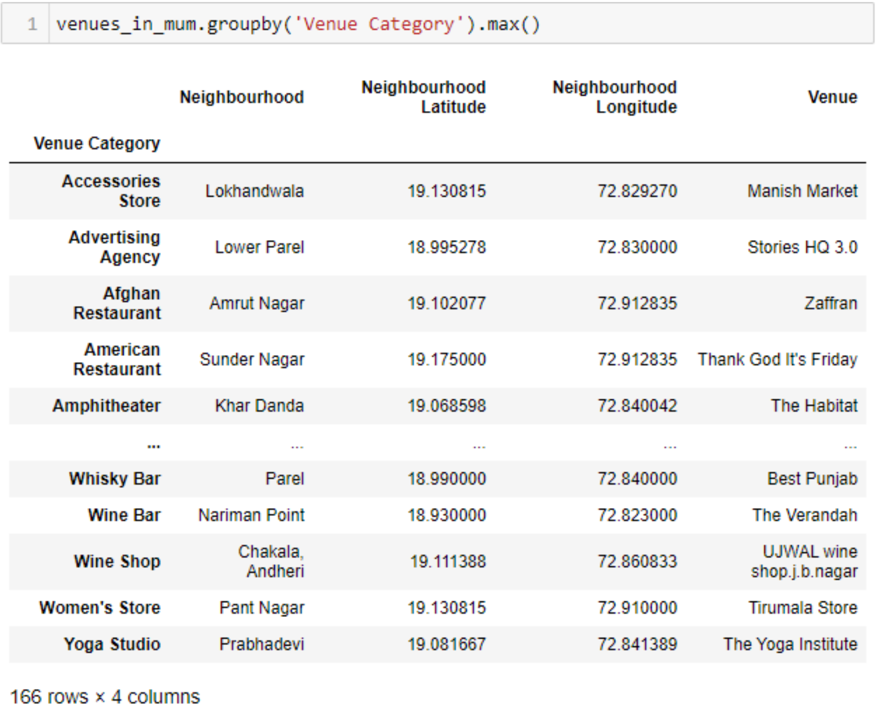
Both the cities came back with similar sized venues lists: Chennai had 156 venue categories and Mumbai had 166



```
#Grouping by Venue Categories
venues_in_chn.groupby('Venue Category').max()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue
Venue Category				
ATM	Tambaram	13.2989	80.3203	HDFC Bank ATM
Advertising Agency	Arumbakkam	13.0724	80.2102	Spica Digital
Afghan Restaurant	Pallavaram	13.0969	80.2865	Yaa Mohaideen Briyani
African Restaurant	Gopalapuram	13.0489	80.2586	Nando's
American Restaurant	Neelankarai	13.0850	80.2547	Tryst Cafe (Baker Street)
...	...	...	...	...
Vegetarian / Vegan Restaurant	Velachery	19.2274	80.2880	Veg Sizzles
Video Store	Parry's Corner	13.0928	80.2893	Burma Bazaar
Vietnamese Restaurant	Gopalapuram	13.0489	80.2586	Va Pho Asian Canteen
Whisky Bar	Guindy	13.0067	80.2206	The Cheroot - the Malt and Cigar Bar
Women's Store	T. Nagar	13.1148	80.2872	Nalli

156 rows × 4 columns

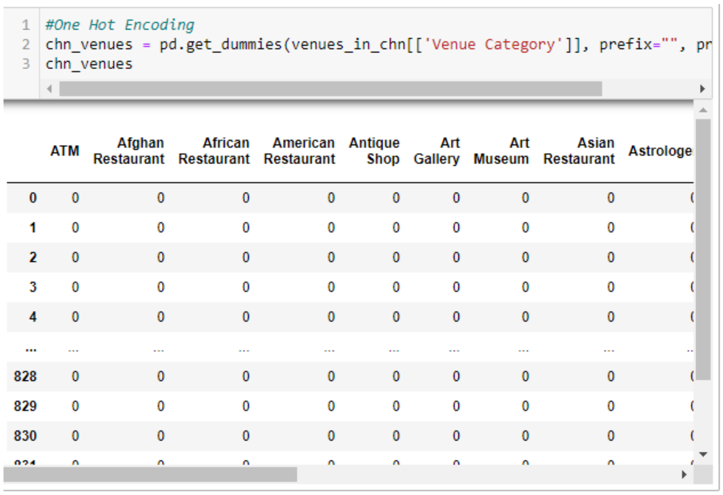


```
venues_in_mum.groupby('Venue Category').max()
```

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue
Venue Category				
Accessories Store	Lokhandwala	19.130815	72.829270	Manish Market
Advertising Agency	Lower Parel	18.995278	72.830000	Stories HQ 3.0
Afghan Restaurant	Amrut Nagar	19.102077	72.912835	Zaffran
American Restaurant	Sunder Nagar	19.175000	72.912835	Thank God It's Friday
Amphitheater	Khar Danda	19.068598	72.840042	The Habitat
...	...	...	...	...
Whisky Bar	Parel	18.990000	72.840000	Best Punjab
Wine Bar	Nariman Point	18.930000	72.823000	The Verandah
Wine Shop	Chakala, Andheri	19.111388	72.860833	UJWAL wine shop.j.b.nagar
Women's Store	Pant Nagar	19.130815	72.910000	Tirumala Store
Yoga Studio	Prabhadevi	19.081667	72.841389	The Yoga Institute

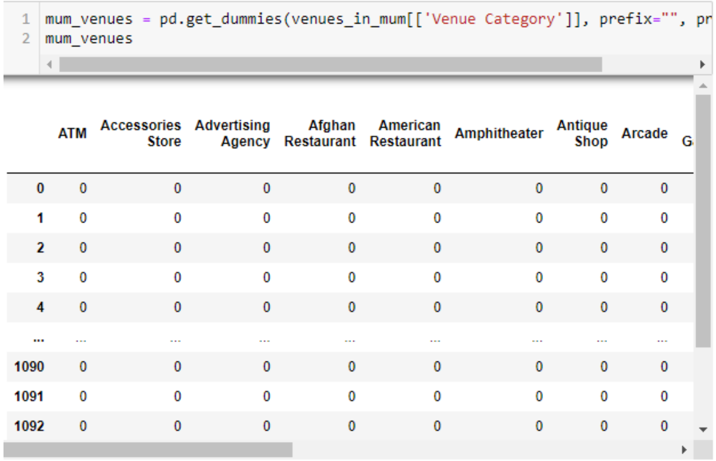
166 rows × 4 columns

One hot encoding on both data frames gave expected output -



```
#One Hot Encoding
chn_venues = pd.get_dummies(venues_in_chn[['Venue Category']], prefix="", pr
chn_venues
```

	ATM	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Art Gallery	Art Museum	Asian Restaurant	Astrologe
0	0	0	0	0	0	0	0	0	(
1	0	0	0	0	0	0	0	0	(
2	0	0	0	0	0	0	0	0	(
3	0	0	0	0	0	0	0	0	(
4	0	0	0	0	0	0	0	0	(
...	...	...	...	...	...	...	...	...	..
828	0	0	0	0	0	0	0	0	(
829	0	0	0	0	0	0	0	0	(
830	0	0	0	0	0	0	0	0	(

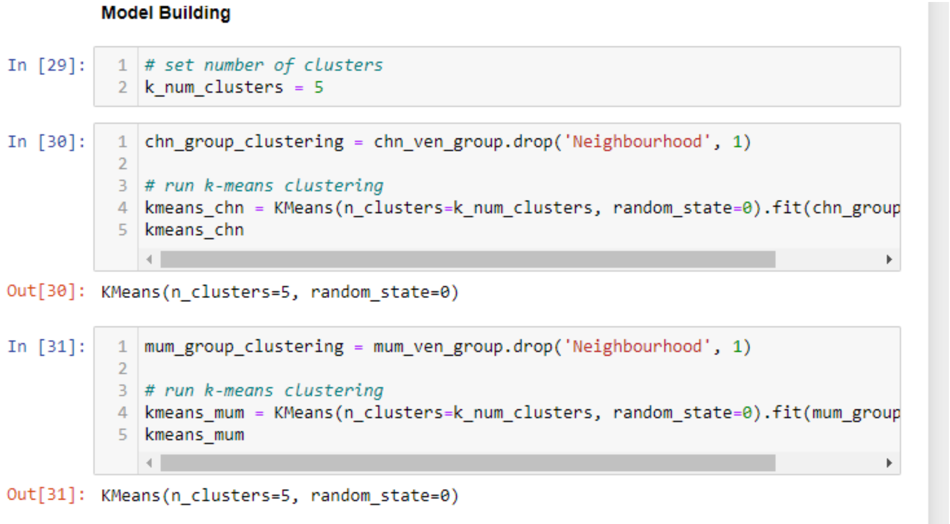


```
mum_venues = pd.get_dummies(venues_in_mum[['Venue Category']], prefix="", pr
mum_venues
```

	ATM	Accessories Store	Advertising Agency	Afghan Restaurant	American Restaurant	Amphitheater	Antique Shop	Arcade	G
0	0	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	
2	0	0	0	0	0	0	0	0	
3	0	0	0	0	0	0	0	0	
4	0	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	
1090	0	0	0	0	0	0	0	0	
1091	0	0	0	0	0	0	0	0	
1092	0	0	0	0	0	0	0	0	

3. Data Modelling

K-means clustering was done with k = 5, model built without any issues.



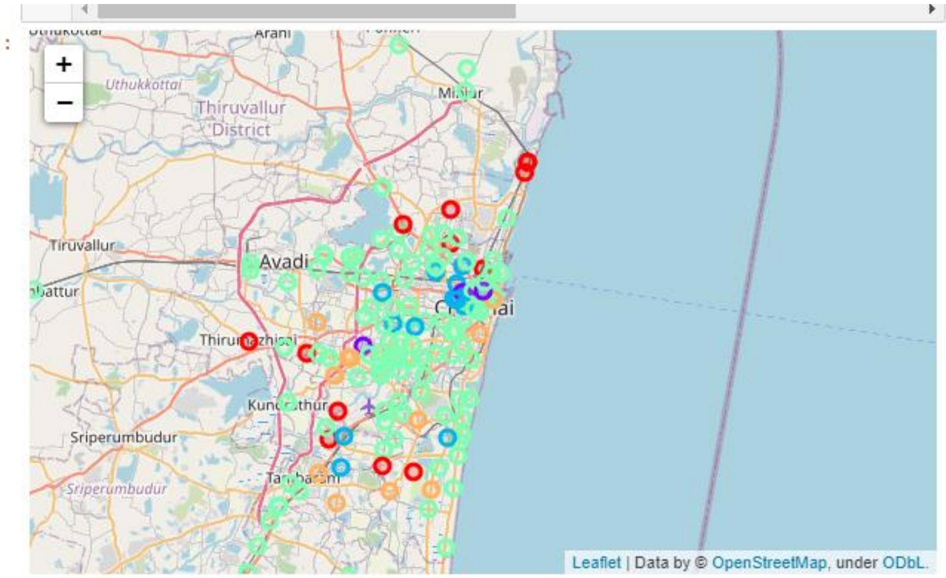
Model Building

```
In [29]: # set number of clusters
         k_num_clusters = 5

In [30]: chn_group_clustering = chn_ven_group.drop('Neighbourhood', 1)

         # run k-means clustering
         kmeans_chn = KMeans(n_clusters=k_num_clusters, random_state=0).fit(chn_group
         kmeans_chn

Out[30]: KMeans(n_clusters=5, random_state=0)

In [31]: mum_group_clustering = mum_ven_group.drop('Neighbourhood', 1)

         # run k-means clustering
         kmeans_mum = KMeans(n_clusters=k_num_clusters, random_state=0).fit(mum_group
         kmeans_mum

Out[31]: KMeans(n_clusters=5, random_state=0)
```

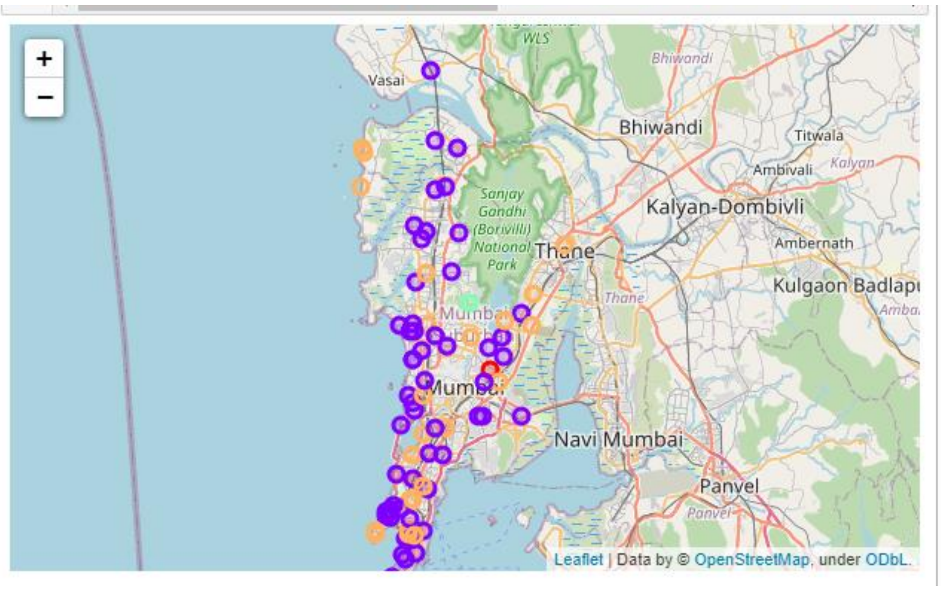
4. Data Visualization

Folium output of maps plotted the clusters properly.

Chennai Clusters:



Mumbai Clusters:



6. *Observations*

Similarities

1. Venues are of similar types/categories - as you would expect the needs of large size cities
2. City seemed to have started around the main beach/coast and then developed progressively inward
3. Points of interest and venues are clustered towards the main/old city and are scattered scantily inwards
4. Train stations scattered throughout the city in both cities

Dissimilarities

1. Chennai 1st preference seemed to be developing around fast food whereas Mumbai has been a mix of restaurants, markets and hotels
2. Multiple and more references of vegetarian & vegan food styles in Chennai compared to Mumbai
3. More outdoor lifestyle options in Mumbai compared to Chennai
4. Chennai seems to have developed in concentric circular circles whereas Mumbai has developed around a linear development of settlements

7. *Conclusions*

Chennai and Mumbai are 2 of the 4 large metropolitan cities in India, both have rich history and have developed over time. The cities started small and concentrated inwards closer to the coast/beach and have developed expanding inwards. These cities continue to grow and expand, on account of increasing economic activity as shown based on the venue listing from FourSquare. There are subtle lifestyle differences as observed by presence/absence of certain features. Travel options seemed to be common with inland train station/network within the cities.

With this trend or growth, both the cities can be expected to become more significant in area and more points of interest/venues coming up around the outer periphery of the cities which continues to expand. Both the cities offer idea ground for newer businesses and economic growth.