1. A full report consisting of all of the following components (**15 marks**):

* Introduction where you discuss the business problem and who would be interested in this project.
* Data where you describe the data that will be used to solve the problem and the source of the data.
* Methodology section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.
* Results section where you discuss the results.
* Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.
* Conclusion section where you conclude the report



Capstone Project - The Battle of Neighborhoods

**Report: IBM Capstone**

Coursera IBM Professional Data Science Project

**27 – 04 – 2020**

**Data Science Authored by:**

**Samir Mulani**

**(BE-comp.Sci)**

**Introduction:**

**A.1:Business Problem:**

* **Finding best new location for coffee shop in Mumbai city.**

**A.2:Description & Discussion of the Background:**

The city of Mumbai consists of a large number of coffee shop, restaurants, but still there is always scope for new ones. The total area of Mumbai is 603.4 km2 (233 sq mi).Of this, the island city spans 67.79 km2 (26 sq mi).and Mumbai was the second most populous city in India after Delhi and the seventh most populous city in the world with a population of 19.98 million. As per Indian government population census of 2011. The most recent census was conducted in India during 2018, which put Mumbai's Urban Agglomeration at 20,748,395, while the city itself was recorded at 12,478,447.

The rapid population growth is attributed to migration from other regions in the country, with migrants seeking business and employment opportunities. On an average 25000 person come to Mumbai daily for work.

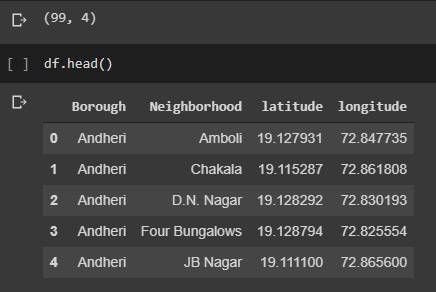
The Goal of this problem is to find a location that suits the below criteria:

1) A location that has many restaurants in the vicinity like (Indian, South Indian).

2) A location that has no or few café coffee, as this will ensure that there very little competition with other competitors.

**A.3. Data Description:**

The data that will be used in these projects is a csv file having data related to all neighborhoods in the city of Mumbai. File data collected from [https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Mumbai](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai%20) using this URL.



**Data that will be used to solve problem of finding best location using neighborhoods and location.**

We explore the neighborhoods using Foursquare API to find the avenues within 500 meters of each neighborhood.

The Foursquare API that will be used to explore the neighborhoods is https://api.foursquare.com/v2/venues/explore. This API returns json response which will be transformed into a Data Frame, taking only the required details into consideration.

**A.4: Target Audience**

To recommend the correct location, XYZ Company Ltd has appointed me to lead of the Data Science team. The objective is to locate and recommend to the management which neighborhood of Mumbai city will be best choice to start a Coffee Shop. The Management also expects to understand the rationale of the recommendations made. This would interest anyone who wants to start a new café in Mumbai city.

**A.5: Success Criteria:**

The success criteria of the project will be a good recommendation of borough/Neighborhood choice to XYZ Company Ltd based on Lack of such café in that location and nearest sources of customer.

**B: Methodology:**

As version control and hosting of files Github was used. Below is the link to the repository.

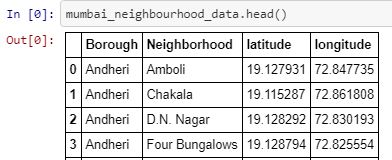
<https://github.com/samir-cell/fantastic-datascience-Coursera_Capstone>

1.Using the Foursquare API venue information is obtained in nearby vicinity locations in Mumbai.

Started with scrapping the List of neighborhoods in Mumbai . The data was wrangled to get it set up in a desired format and get the neighborhoods from following link:

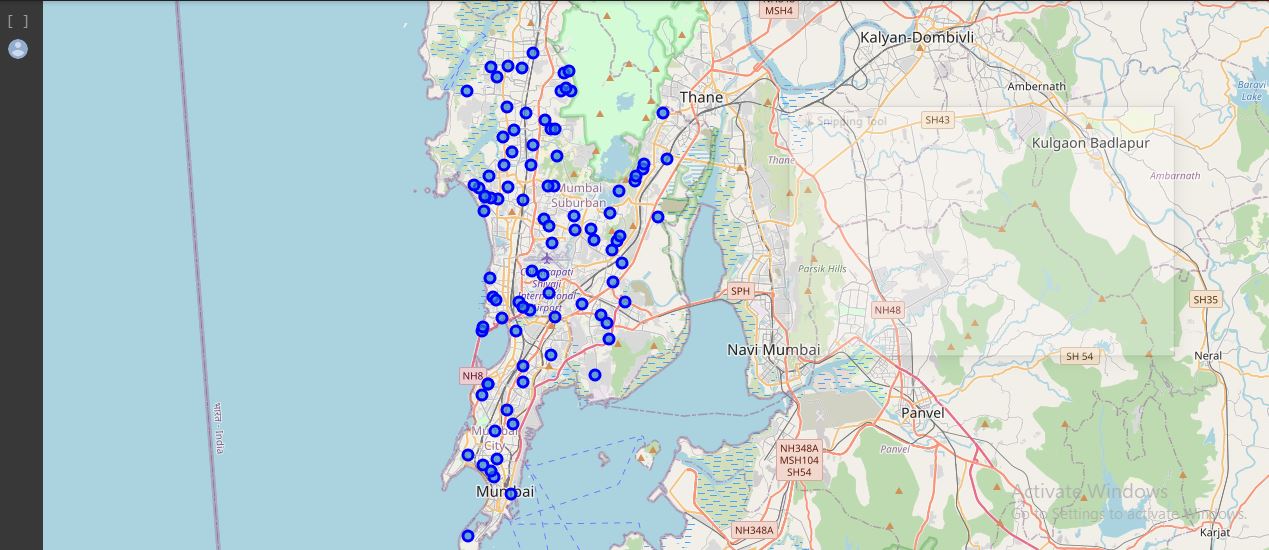
<https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai>

2.Using geolocator API scrapping the list of latitude and longitude .



The geographical coordinates were fetched using **Nominatim open street API**. Coordinates for all the localities could not be fetched. So I plotted the fetched co-ordinates using **Folium** library and found that without the missing co-ordinates, there is still a good distribution of the localities could be gathered.

**Distribution of Sample Data**



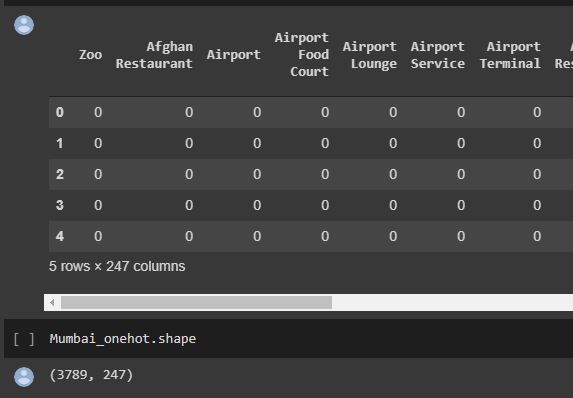
A radius is set to cover large neighborhoods in a particular area in Mumbai

3.Getting Nearby Venues using Four-Square API:

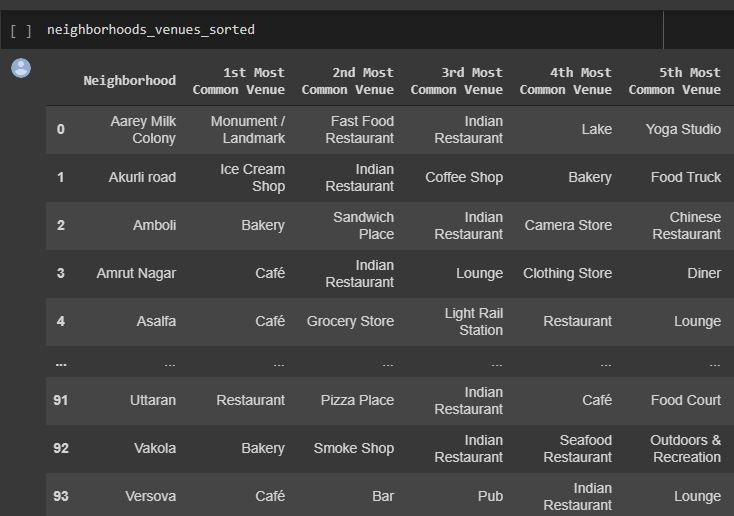
The total types of venues received from the Four-Square API was 3789. These types are often overlapping and similar in terms of their ability to contribute to clustering of Neighbourhoods. It made more sense to converge the types of venues by grouping them.

*There are 247 unique categories.*

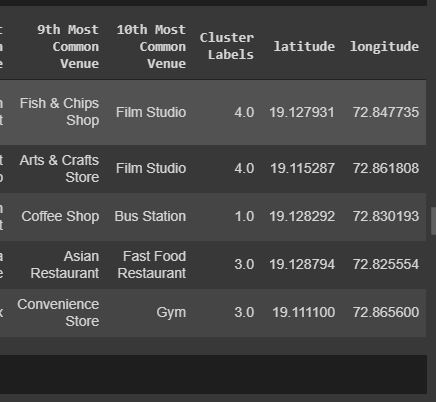
*4.* *One-hot Encoding* was used to assign dummy variables to each “venue\_types”. The sum of venues for each locality was multiplied with the “*importance in clustering”.* Hence the data was prepared for clustering.



5.we get most common top 10 venues of all neighborhoods.



6. K-Means Algorithm was used to cluster the neighbourhoods. The algorithm was iterated with cluster.



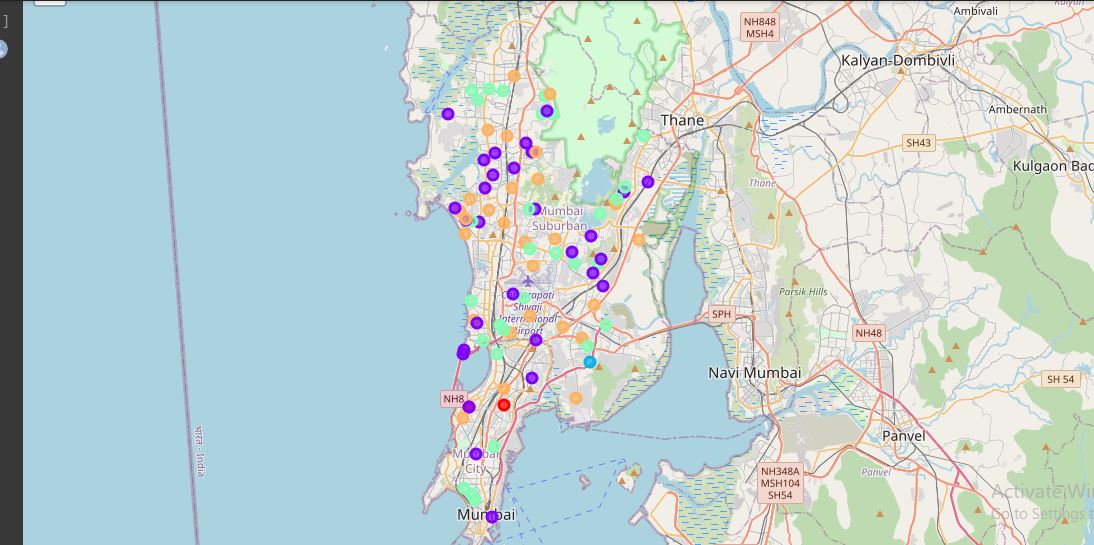
With appending longitude, latitude and cluster label we just created using k- means.

The **K**-**means clustering** algorithm is **used** to find groups which have not been explicitly labeled in the data. This can be **used** to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets.

Other clustering algorithms with **better** features tend to be more expensive. In this case, **k**-**means** becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. **K**-**means** is the simplest

**C: DATA Analysis:**

**1.We Create Folium Map of Cluster we just define for better understanding of cluster location and intensity.**

****

1. **Analysis the Data of common venue from each cluster**

****

**Resulting :**

1.Indian Restaurant are present in each cluster (1st common venue )

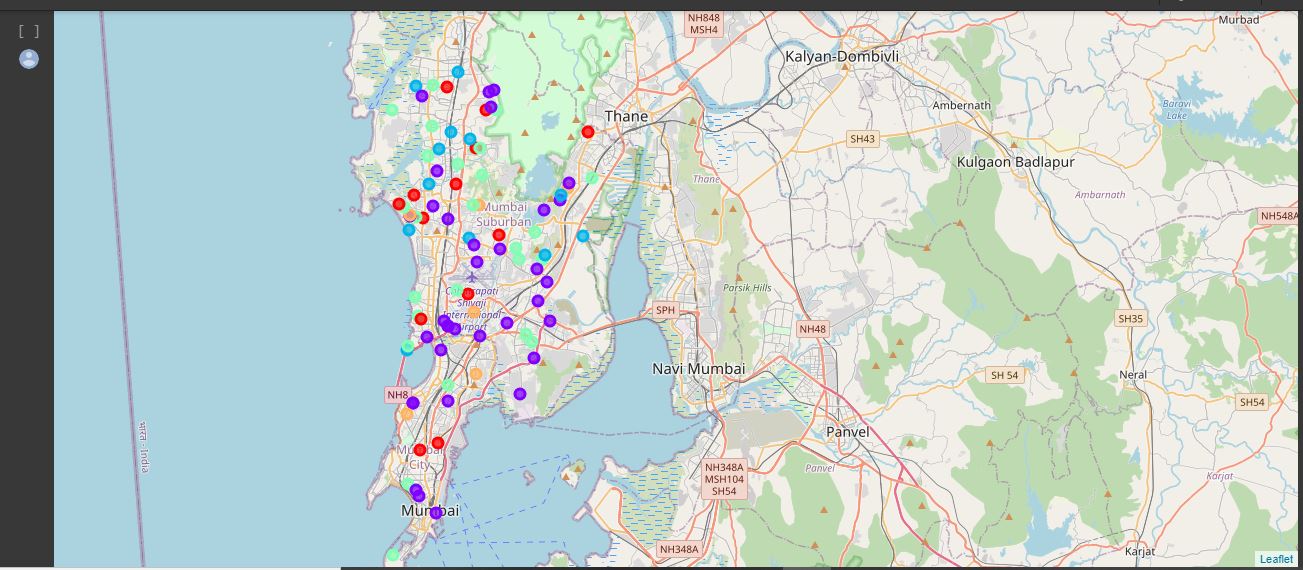
2.coffe shop is popular among Mumbai people. (1st common venue in some neighborhood)

we can locate coffee shop location for opening new coffee shop.

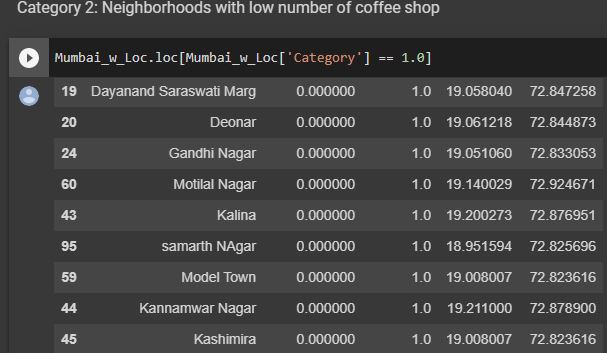
3.Lets Find out the Best location or Neighborhood for new coffee shop:

4.We cluster venue again and Get Mean for each Neighborhoods of coffee shop.

5.And Created Folium Map to analyze cluster just defined. Along with cluster label.



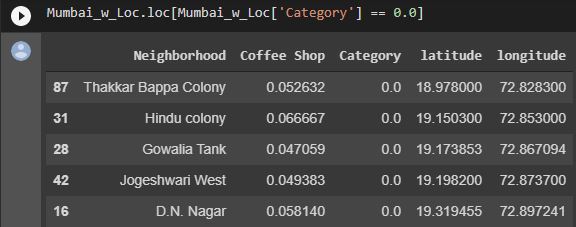
6. Cluster are sorted with its number and Neighborhoods with very low, low, medium high and very high number of Coffee Shop. Like this.



**C. Discussion and Conclusion:**

**1:** **Results:**

By exploring the requirements we found only two neighborhoods that match the requirements (Many restaurants in the vicinity & only a few coffee shops.) refer above fig.



**2: Discussion:**

According to results we observe that most common venues (Top 10) come out to be restaurants and café, coffee shape and like Snack places, which means any new chef/business man can start a restaurant provided that they need to compete with existing restaurants (Other Types) but only if he provides top class facilities to get to top.

Result was more beneficiary having dataset of rating of shop and population of neighborhood.

People all over the world are turning to big cities to start a business or for work. This model can be used to further decide which can recommend most favoured locations as per the preferences of a user.

**3: Conclusion:**

Based on the both Clusters formed it would be a good idea to open a new coffee shop in Clusters 0 or 1 & 2 since the other clusters already have coffee shop and Restaurants in their vicinities. cluster 0 and "NAN" value having no competition has possibility is there people are not interested at coffee shop.

cluster 1 & 2 have many restaurants in the vicinity (Pizza restaurants, European, Intercontinental, Indian) so one will be able to attract a good crowd.

(Note) These results have limitations - The venue data obtained is of top 10 venues in each neighbourhood, where we might neglect Coffee shop with less frequency or another venue name of same category.

References: Wikipedia pages :

1. [https://en.wikipedia.org/wiki/List\_of\_neighbourhoods\_in\_Mumbai](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai%20)
2. [https://en.wikipedia.org/wiki/ Mumbai](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Mumbai%20)