**Clustering of cities based on their Neighborhoods**

**Asmita Khode**

**18 September, 2020**

1. **Introduction**

**1.1. Background**

India, one of the most fastest developing nations in the world. Today almost all of the big giants are eyeing to invest in India because of its huge market size and an incredible growth rate. To set up a business requires a well understanding of the place where one should establish. Location of any business plays an important role in it’s success and ultimately decides the path of its growth. Also different business has different location requirements and a different business plan adaptable to its locality and neighborhood and vice verse.Thus, choosing a perfect business location for a new set up or in expanding business to new locations is challenging task.

**1.2.Problem Description**

This project aims to group similar kind of cities in India based upon their locality and neighborhood (similarity between type of venues and its density) which will help to find most suitable location for a particular type of business and also in fetching the similar business plans for cities belonging to same cluster which will reduce overheads of searching and planning for the companies.

**1.3. Interests**

Businesses where location is primary concern like large cafes, hotel, gym chains, other food industry related companies, transportation businesses, electronics and fashion industry etc, would be most interested in any such exploration. Also these can benefit the government to understand similar kind of cities and help them in planning effectively. Understanding of a particular area and its neighborhood, venue density helps to determine the need, supply, target customers and their interests etc. Also venues like train station, airport etc will help in better planning as well.

1. **Data Description**

For this project we will require:

1. List of all cities of interest.
2. Coordinates of these cities for our exploration of locality and neighborhoods.
3. Information of locality and neighborhood on basis of which clustering of cities is performed.

**2.1.Data Sources**

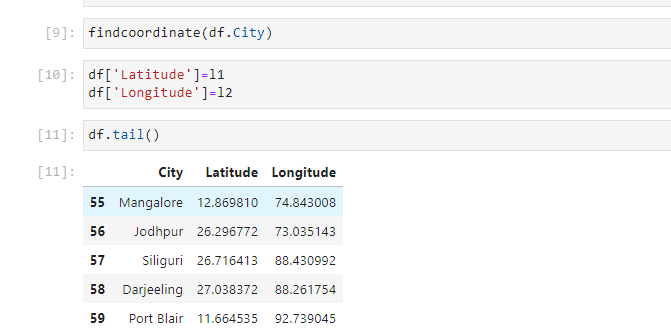
1. We will manually prepare a list of cities that we are interested to perform clustering on. For the purpose of this project we have taken sixty buzz Indian cities and stores them into using list data structure.
2. To find the coordinates of each city we have used the geolocator module of python. The geocode function returns the latitude and longitude of any given place.
3. To find venues and their density around any given set of coordinates we have used the FourSquare API. This API returns a list of venues set within limit(no.of venues to be returned) and radius(range of search).

**2.2.Data Preprocessing and transformations**

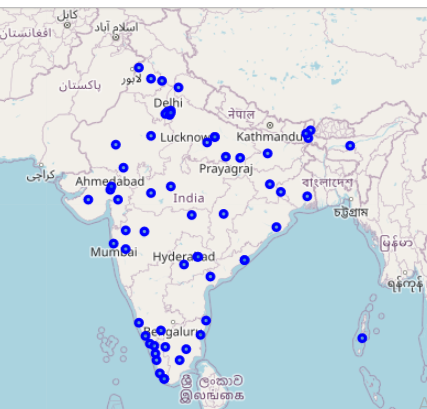
* First we have stored all the cities, their coordinates and venue, their category etc into a pandas dataframe.
* For preprocessing we have to transform the data so that it can be used for clustering. For this we have first used onehot transformation on venue category column to obtain all the venue category-wise.
* Now we will calculate the frequency of each category for each city. To do so we have grouped the data by city and calculated mean for each category and finally sorted them from high to low.
* Also, for now we will drop all the NAN rows for which we don’t have any data fetched from FourSquare API.

1. **Methodology**

First and most important step in any project is data collection. To gather all the relevant data first we have considered sixty most popular Indian cities and stored them using a list, then find their coordinates using geolocator module and store the city\_name, latitude and longitude in a pandas dataframe.



Now, we will first have a look at all our selected locations using Folium library to plot the map.



To explore the venues we will use Four Square API. Four Square API is a great platform for developers which allow to search for all the venues around a given coordinate along with details of venues like coordinates of venues, venue category, reviews, details about users who has given reviews to a particular venue etc.

To use the Four Square API we have to set two parameters:

1. Limit i.e.maximum number of results that is to be returned for particular query. Here, for this project we have set our limit to 50 venues.
2. Radius i.e. range of our search. We will set our range to 150kms.

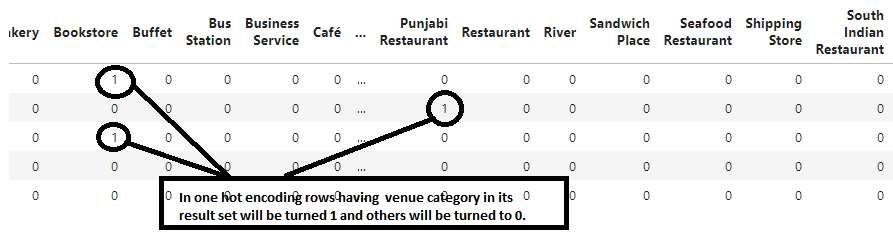
Screenshot (32)

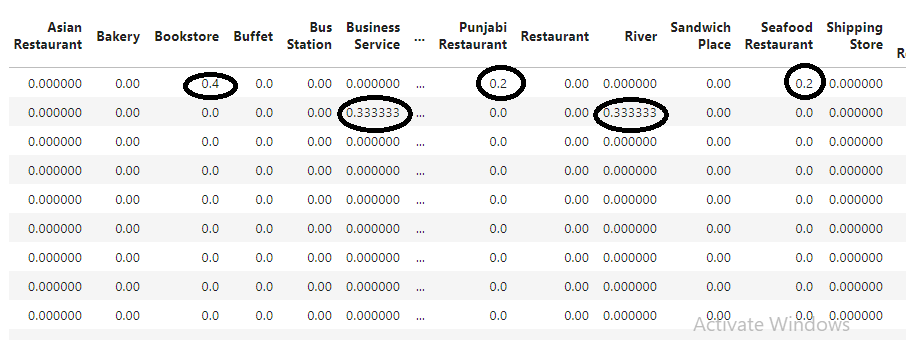
Now using our FourSquare API we will explore our locality, venue density and its neighborhood through passing coordinates of each city one by one.After this we will filter our search results and store only relevant details like venue name, it’s coordinates and venue category etc.

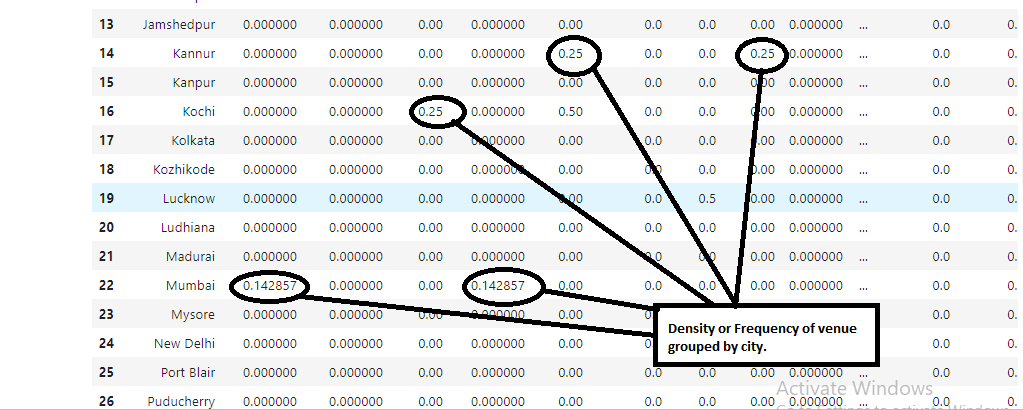
The result obtained will look like something:



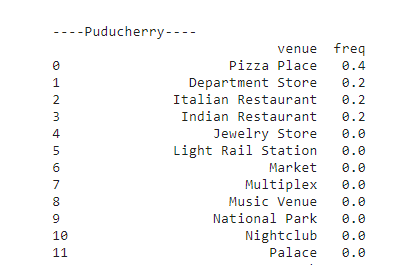
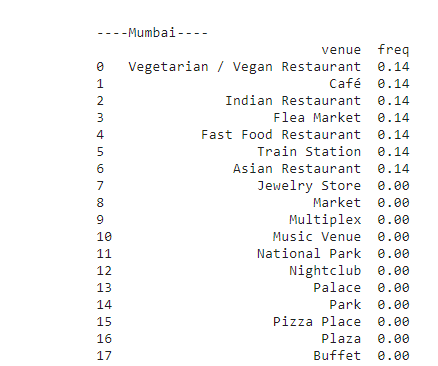
Here we have obtained 79 results in total. Now we will transform this data using onehot encoding technique so that it can be used for clustering.

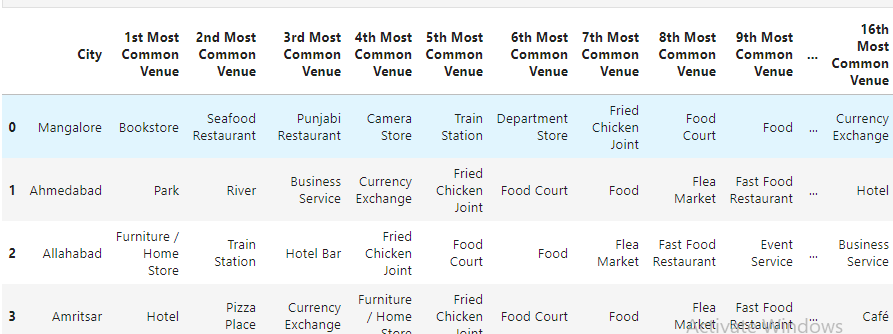
 Now to find density of particular venue category around a city we will first group our data by city and then calculate the mean of categories which will give us frequency of that category. After obtaining the density or frequency we will now arrange it from high to low so to get most common or popular venues of an area.





Note that, here we have considered only top 25 venues for our exploration and analysis.We have listed all the 25 venues according to their density.





Now our data is ready to perform clustering. We will use this prepared dataset to cluster our cities into different groups based upon similarity between their locality, venue density and neighborhood.

We have used K-means clustering technique to cluster our data. For now we have set our k to 5 because we are getting optimum result at k=5. We have used K-means clustering because it is most easy and most used technique. Also we find this technique most suitable for our dataset.

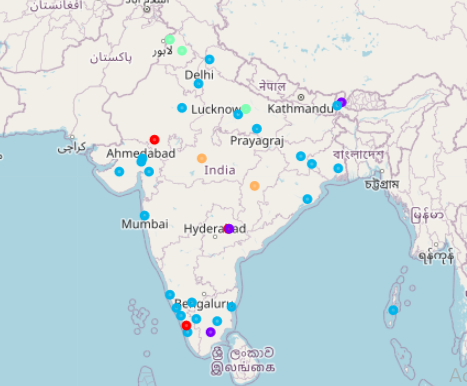
The result of k-means algorithm is an array of cluster number so we have to merge this with our data and then finally drop the nan values i.e. all the rows for which we have no result.

1. **Results**

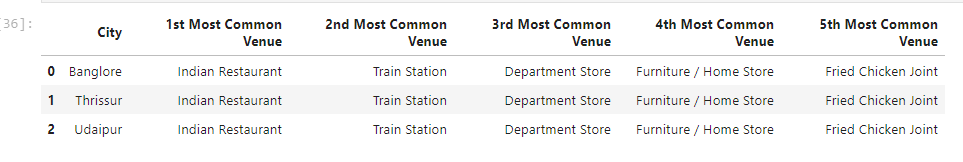
Merging the cluster number with our dataset:



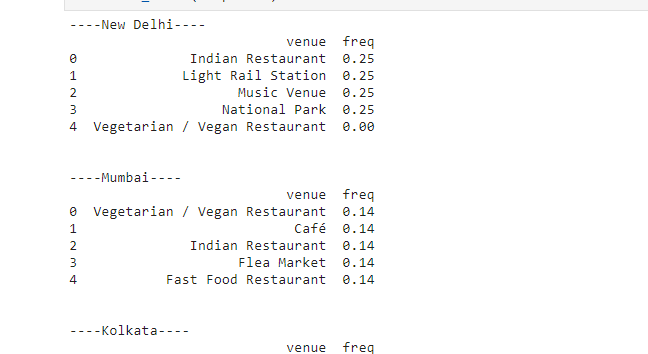
Let’s visulaize our clusters on the map.



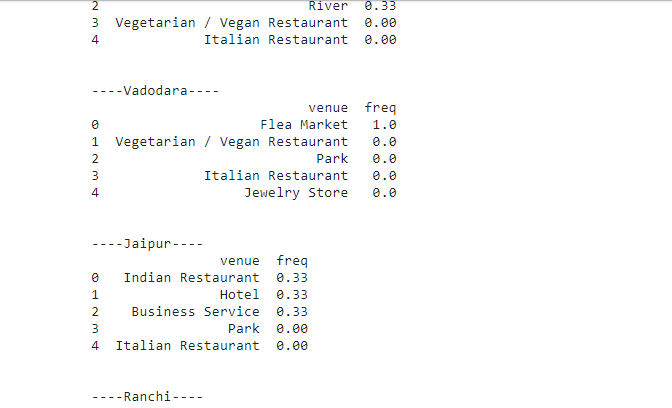
The different colors on the map shows the different clusters.For our analysis let us see individual frames based on clusters to get better understanding of our clusters and check top five most popular and common venues of a particular city.



**Cluster 1: Top five venues.**



**Cluster 3: Frequency of top 5 venues.**

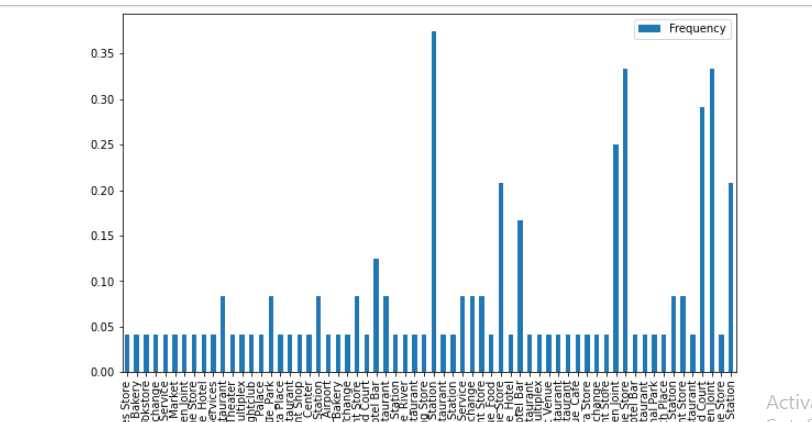
****

**Cluster 3**

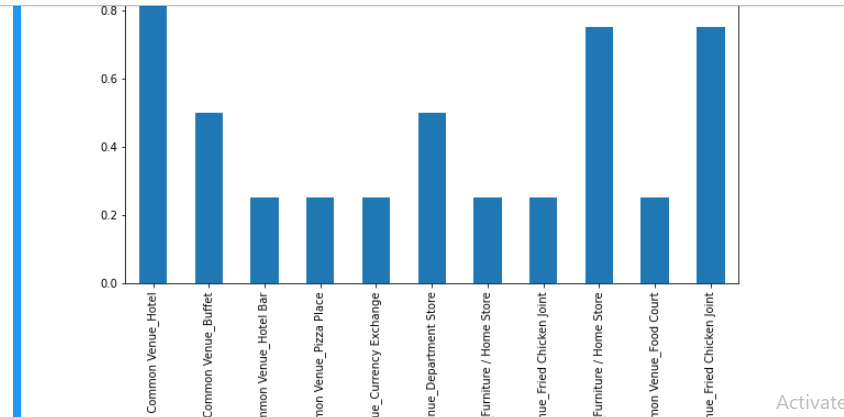
****

**Cluster 3: Top 5 venues for each city.**

Now let’s plot a bar graph for each cluster to see frequency of each venue.



**Cluster 3: frequency of each venue in cluster 3**

****

**Cluster 5: frequency of each venue in cluster 5.**

1. **Discussion**

Here we have initially taken 60 cities out of which we got results for only 36 cities. I.e only 79 venues. Hence for now we are able to perform clustering of these 36 cities. With the better search results for FourSquare API like with a premium account we can get more better and large dataset which will enhance our clustering results as well as our analysis of density of the venues in a particular city.

Also this method can be performed on a large number of big-small towns and cities for identifying most suitable locations and to form big clusters for establishing the business.

1. **Conclusion**

As business are growing faster their need to identify new profitable locations is a huge challenge, to find most suitable location for a new business or fetching a business plan as per location is also a crucial task that decides the growth of any business.

Any such platform or analysis which can easily identify similar locations as well as can give us brief study about the venues, their density and neighborhood ariund a city can reduce overheads of searching and planning for the locations and can contribute to better understanding of business requirements and can also help in ensuring success of a business in particular city or region.