The Battle of Neighborhoods

Introduction:

Hyderabad is one of the most densely populated area and MetroPolitan City in India. Being the land of Opportunity, it brings in a variety of people from different ethnic backgrounds to the core city, Hyderabad. Being the largest city in India, there is no doubt about the diversity of the population. The multiculturalism is seen through the various neighborhoods including; Andhra Pradesh, Telangana, Maharashtra,Orissa, TamilNadu and many more. Hyderabad being the hub of interactions between ethnicities, brings many opportunities for entrepreneurs to start or grow their business. It is a place where people can try the best of each culture, either while they work or just passing through. Hyderabad is well known for its great food. Hyderabad is well known for Hyderabadi Dhum Biryani and Irani Chai.

The objective of this project is to use Foursquare location data and regional clustering of venue information to determine what might be the 'best' neighborhood in Hyderabad to open a restaurant. Gobi Manchurian and Noodles are one of the most bought dishes in Hyderabad originating from China. Hyderabad being the Metropolitan City with people from various countries and states, there are numerous opportunities to open a new Chinese restaurant. Through this project we will find the most suitable location for an entrepreneur to open a new Chinese restaurant in Hyderabad, India.

Target Audience: • Entrepreneurs who want to open a Chinese Restaurant in Hyderabad

Data Overview:

The data that will be required will be a combination of CSV files that have been prepared for the purposes of the analysis from multiple sources which will provide the list of neighborhoods in Hyderabad (via GHMC website), the Geographical location of the neighborhoods (via Geocoder package) and Venue data pertaining to Chinese restaurants (via Foursquare). The Venue data will help find which neighborhood is best suitable to open a Chinese restaurant.

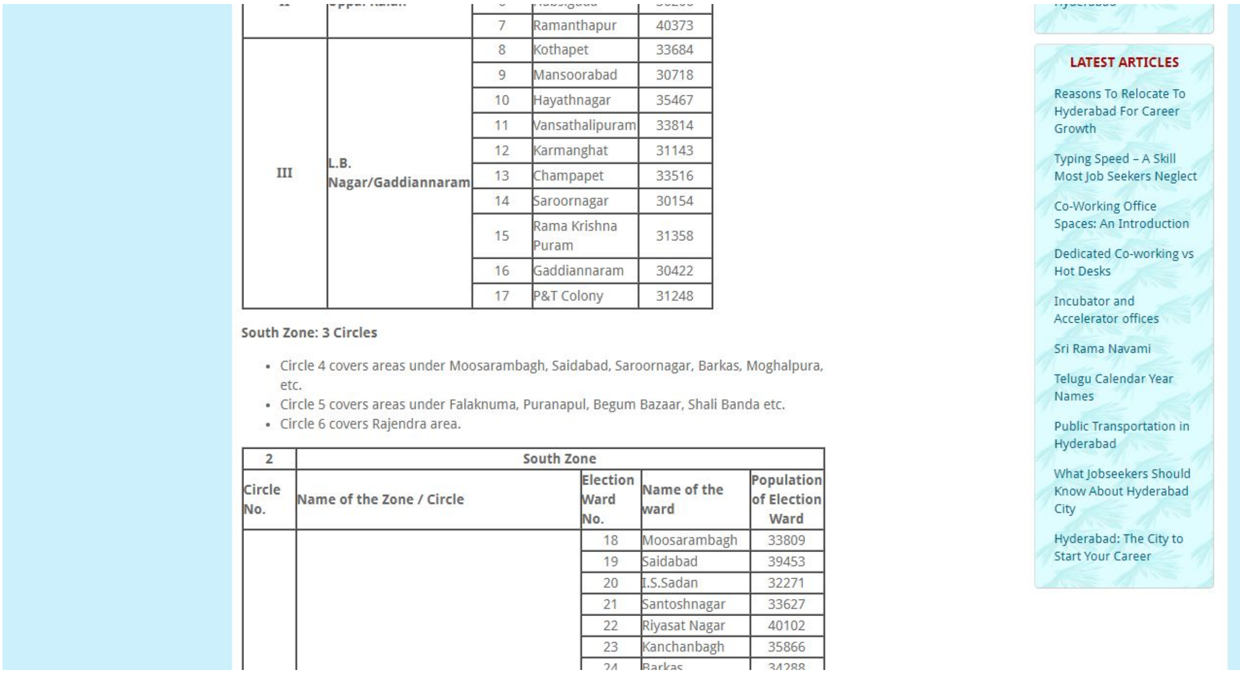
Methodology:

First, we will need to extract the data from the data sources:

Source 1: Hyderabad Neighborhoods via Hyderabad India Online Site

The site (http://hyderabad-india-online.com/2009/12/ghmc-zones-circles-and-wards/) shown above, provided almost all the information about the neighborhoods. It included the ward number, Zone and the name of the neighborhoods present in Hyderabad.

Since the data is not in a format that is suitable for analysis, Cleaning and preprocessing of the data from this site is done



		7	Ramanthapur	40373
III	L.B. Nagar/Gaddiannaram	8	Kothapet	33684
		9	Mansoorabad	30718
		10	Hayathnagar	35467
		11	Vansathalipuram	33814
		12	Karmanghat	31143
		13	Champapet	33516
		14	Saroornagar	30154
		15	Rama Krishna Puram	31358
		16	Gaddiannaram	30422
		17	P&T Colony	31248

South Zone: 3 Circles

- Circle 4 covers areas under Moosarambagh, Saidabad, Saroornagar, Barkas, Moghalpura, etc.
- Circle 5 covers areas under Falaknuma, Puranapul, Begum Bazaar, Shali Banda etc.
- Circle 6 covers Rajendra area.

2	South Zone			
Circle No.	Name of the Zone / Circle	Election Ward No.	Name of the ward	Population of Election Ward
		18	Moosarambagh	33809
		19	Saidabad	39453
		20	I.S.Sadan	32271
		21	Santoshnagar	33627
		22	Riyasat Nagar	40102
		23	Kanchanbagh	35866

LATEST ARTICLES

- Reasons To Relocate To Hyderabad For Career Growth
- Typing Speed – A Skill Most Job Seekers Neglect
- Co-Working Office Spaces: An Introduction
- Dedicated Co-working vs Hot Desks
- Incubator and Accelerator offices
- Sri Rama Navami
- Telugu Calendar Year Names
- Public Transportation in Hyderabad
- What Jobseekers Should Know About Hyderabad City
- Hyderabad: The City to Start Your Career

Figure : Hyderabad Online portal showing List of Neighborhoods in Hyderabad with respective ward numbers,zones and neighbourhoods.

```
In [9]: hyd_data.head()
```

Out[9]:

Area Code	Zone	Neighborhood	Population
1	Kapra	Kapra	32108
2	Kapra	Cherlapalli	34729
3	Kapra	Mallapur	39842
4	Kapra	Nacharam	39793
5	Uppal Kalan	Uppal	40638

Figure : Data that was obtained from a site and preprocessed into Pandas data frame.

Source2: Geographical Location data using Geopy Package

The second source of data is by using the Python package geopy to get the Geographical coordinates of the neighborhoods with the respective Neighbourhoods.

Area Code	Zone	Neighborhood	Population	latitude	longitude
1	Kapra	Kapra	32108	17.4846	78.5610
2	Kapra	Cherlapally	34729	17.4687	78.6025
3	Kapra	Mallapur	39842	17.4405	78.5789
4	Kapra	Nacharam	39793	17.4285	78.5528
5	Uppal Kalan	Uppal	40638	17.4025	78.5613

Figure : Geographical data of Neighborhoods in Hyderabad

Source3: Venue Data using Foursquare The retrieval of the location, name and category about the various venues in Hyderabad was collected through the Foursquare explore API. To obtain the data, it was required to make an account where it would provide a ‘Secret Key’ as well as a ‘Client ID’ which would allow me to pull any data.

```
venues_list
```

```
1]: [('Kapra',
      17.4846,
      78.561,
      'The Coffee Cup',
      17.48318023870212,
      78.55210435802412,
      'Café'),
     ('Kapra',
      17.4846,
      78.561,
      'Cafe Coffee Day',
      17.481262412602668,
      78.55507673229836,
      'Café'),
     ('Kapra',
      17.4846,
      78.561,
      'Parivaar Restaurant',
      17.476849799609987,
```

Figure : Venue data pulled from Foursquare explore API

It is seen through figure 4 (above) that the neighborhoods are grouped by the neighborhood, so data clustering is made easier later on.

After all the data was collected and put into data frames, cleansing and merging of the data was required to start the process of analysis.

After Data cleaning and preprocessing, the venue data pulled from the Foursquare API was merged with the table above providing us with the local venue within a 1000-meter radius shown below.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Kapra	17.4846	78.561	The Coffee Cup	17.483180	78.552104	Café
1	Kapra	17.4846	78.561	Cafe Coffee Day	17.481262	78.555077	Café
2	Kapra	17.4846	78.561	Parivaar Restaurant	17.476850	78.563525	Indian Restaurant
3	Kapra	17.4846	78.561	McDonald's	17.476961	78.564754	Fast Food Restaurant
4	Kapra	17.4846	78.561	Swagath Grand	17.482022	78.553261	Indian Restaurant

Figure : Local Venues near the respective Neighborhood

Now after cleansing the data, the next step was to analyze it. We then created a map using folium and color coded each Neighborhood depending on what Zone it was located in.

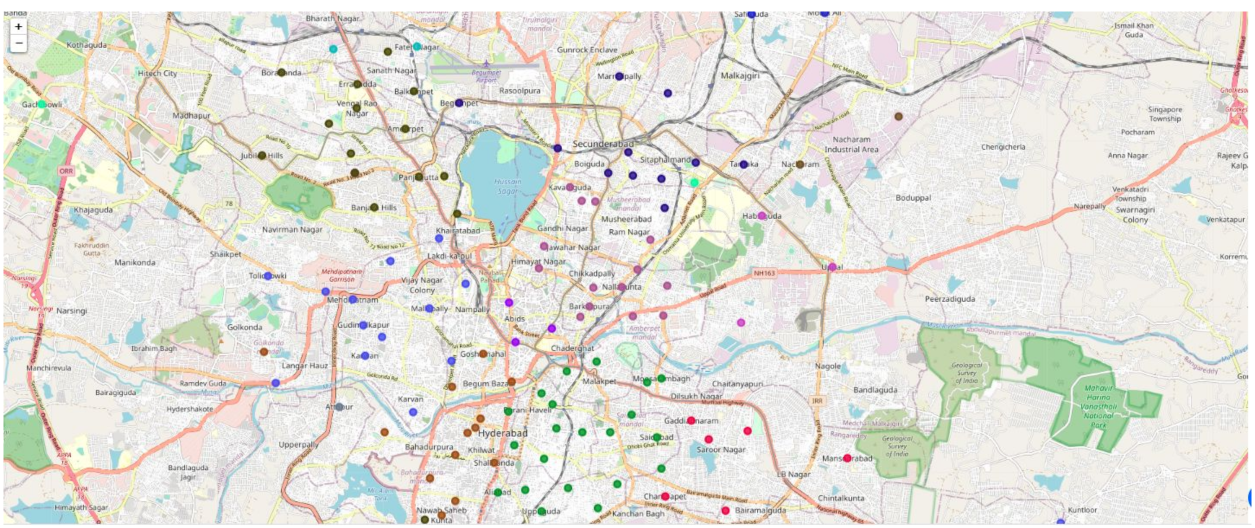


Figure: Visualize the Hyderabad Map

Now after cleansing the data, the next step was to analyze it. We then created a map using folium and color coded each Neighborhood depending on what Zone it was located in.
Next, we used the Foursquare API to get a list of all the Venues in Hyderabad which included Parks, Schools, Café Shops, Asian Restaurants etc. Getting this data was crucial to analyzing the number of Chinese Restaurants all over Hyderabad. We then merged the Foursquare Venue data with the Neighborhood data which then gave us the nearest Venue for each of the Neighborhoods.
Then to analyze the data we performed a technique in which Categorical Data is transformed into Numerical Data for Machine Learning algorithms. This technique is called One hot encoding.

	Neighbourhoods	ATM	Accessories Store	Afghan Restaurant	American Restaurant	Arcade	Asian Restaurant	Athletics & Sports	Auto Garage	BBQ Joint	Bakery	Bank	Bar	Basketball Court	Bengali Restaurant	Bistro	Boat or Ferry	Bookstore	Boutique	Bc
0	Kapra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	Kapra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	Kapra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3	Kapra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4	Kapra	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure: Onehot Encoding

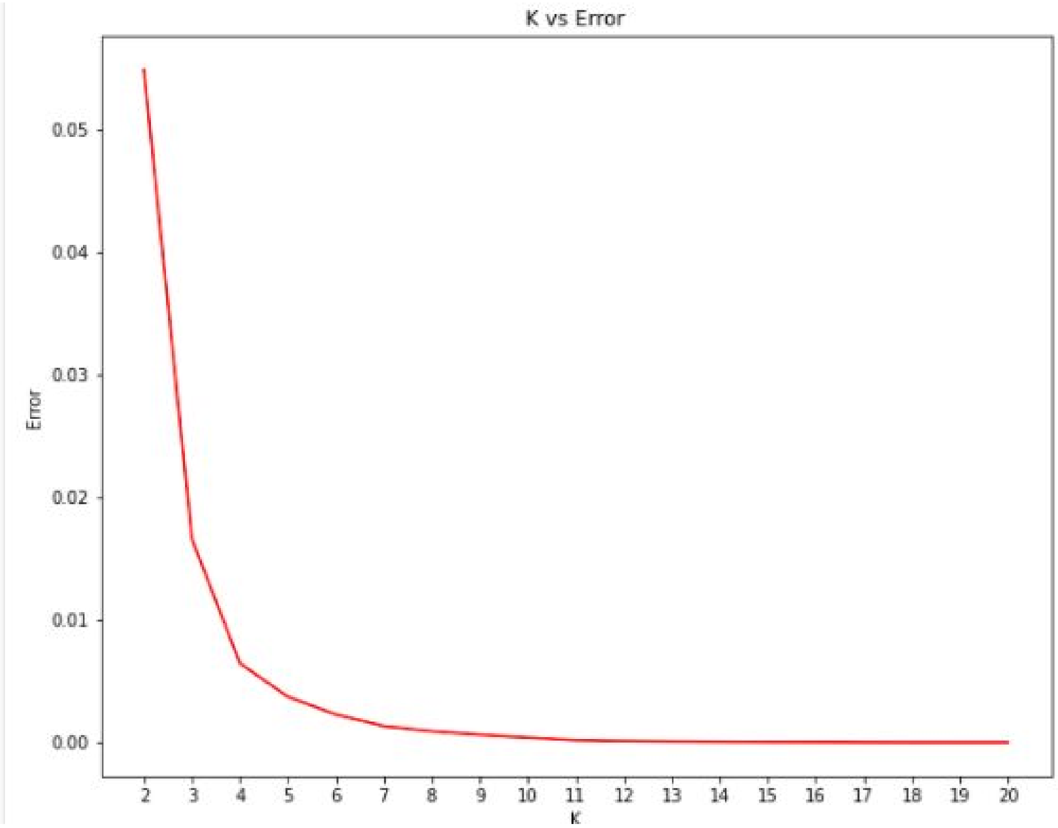
For each of the neighborhoods, individual venues were turned into the frequency at how many of those Venues were located in each neighborhood. Then we grouped those rows by Neighborhood and by taking the Average of the frequency of occurrence of each Venue Category. After, we created a new data frame which only stored the Neighborhood names as well as the mean frequency of Chinese Restaurants in that Neighborhood. This allowed the data to be summarized based on each individual Neighborhood and made the data much simpler to analyze.

	Neighbourhoods	Chinese Restaurant
0	Addagutta	0.000000
1	Adikmet	0.000000
2	Ahmed Nagar	0.050000
3	Akbarbagh	0.000000
4	Aliabad	0.111111

Figure : Extracted data of Neighbourhoods and Chinese Restaurant

Clustering:

To make the analysis more interesting, we wanted to cluster the neighborhoods based on the neighborhoods that had similar averages of Chinese Restaurants in that Neighborhood. To do this we used K-Means clustering. To get our optimum K value that was neither overfitting or underfitting the model, we used the Elbow Point Technique. In this technique we ran a test with different number of K values and measured the accuracy and then chose the best K value. The best K value is chosen at the point in which the line has a sharpest turn. In our case we had the Elbow Point at K = 4. That means we will have a total of 4 clusters.



Finding the K vs Error Values

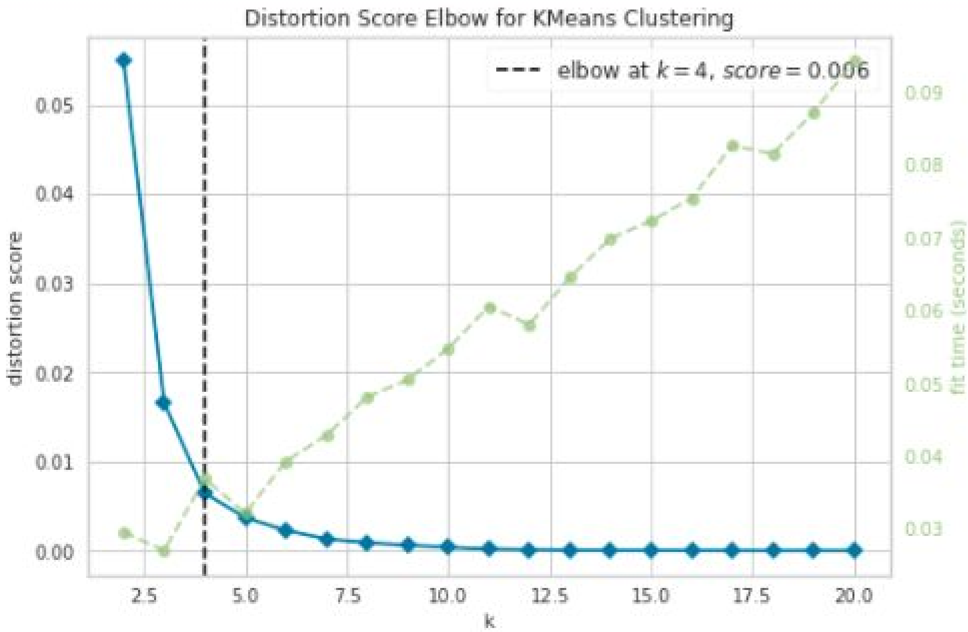


Figure: Finding right k using the Elbow Point

New data frame storing Neighborhoods and the average Chinese Restaurant in that Neighborhood. We integrated a model which would fit the error and calculate the distortion score. From the dotted line, we see that the Elbow is at K=4. Moreover, in K-Means clustering, objects that are similar based on a certain variable are put into the same cluster. Neighborhoods that had similar mean frequency of Italian Restaurants were divided into 4 clusters. Each of these clusters were labelled from 0 to 3 as the indexing of labels begin with 0 instead of 1.

	Neighbourhood	Chinese Restaurant	Cluster Labels
0	Addagutta	0.000000	1
1	Adikmet	0.000000	1
2	Ahmed Nagar	0.050000	3
3	Akbarbagh	0.000000	1
4	Aliabad	0.111111	0

Figure: Appropriate Cluster labels were added

After, we merged the venue data with the table above creating a new table which would be the basis for analyzing new opportunities for opening a new Chinese Restaurant in Hyderabad.

Then we created a map using the Folium package in Python and each neighborhood was colored based on the cluster label.
For example, cluster 2 was purple and cluster 3 was blue. The map above shows the different clusters that had similar mean frequency of Italian restaurants.

Analysis:

We have a total of 4 clusters (0,1,2,3). Before we analyze them one by one lets check the total amount of neighborhoods in each cluster and the average Italian Restaurants in that cluster.

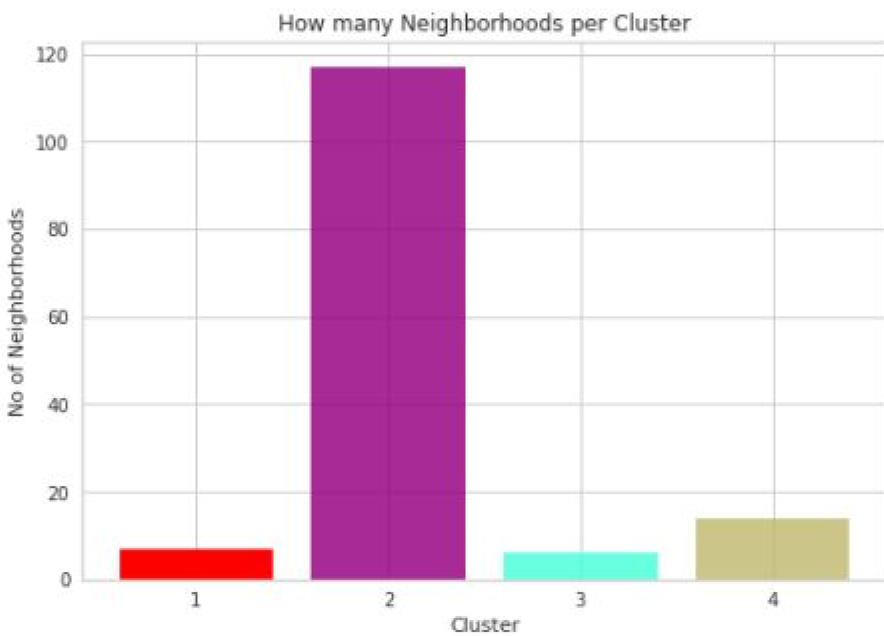


Figure : Number of Neighborhoods per cluster

From the above bar graph that was made using Matplotlib , we can compare the number of Neighborhoods per Cluster.

We see that Cluster 3 has the least neighborhoods(6), while cluster 2 has the most(118). Cluster 4 has 14 neighborhoods and cluster 1 has only 7. Then we compared the average Chinese Restaurants per cluster.

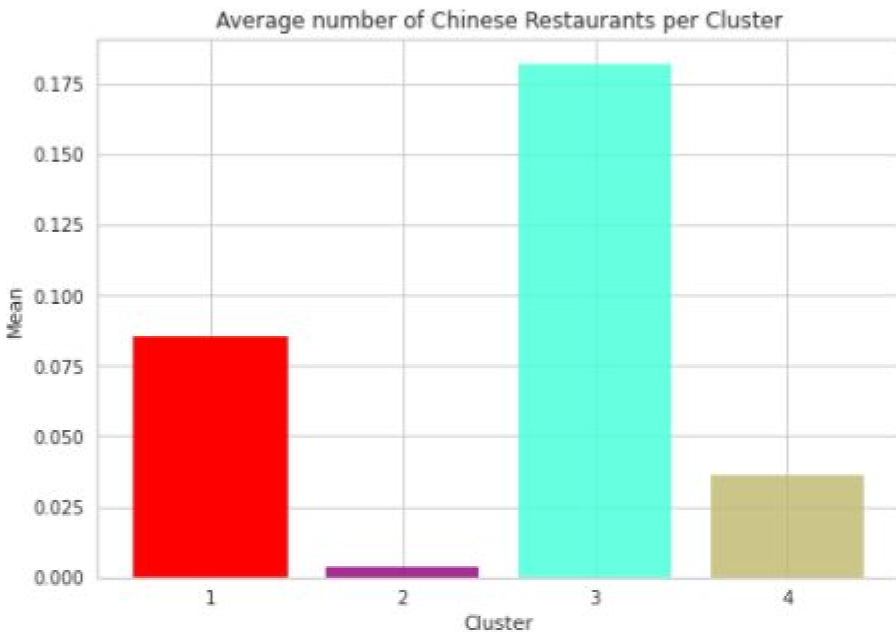


Figure : Average Italian restaurant in each neighborhood

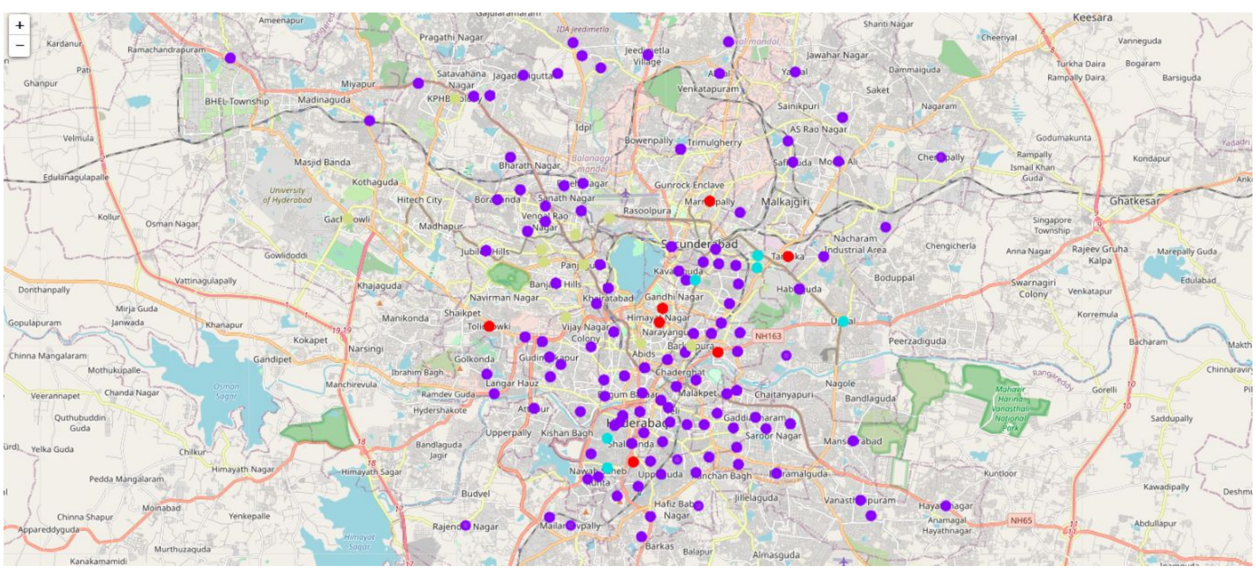


Figure : Map with different Clusters

This information is crucial as we can see that even through there are only few neighborhoods in Cluster 3, it has the highest number of Chinese Restaurants, while Cluster 2 has the most neighborhoods but has the least average of Chinese Restaurants Also, from the map, we can see that neighborhoods in Cluster 2 are the most densely populated.

Analysis of each cluster:

Now let's analyze the Clusters individually (Note: these are just snippets of the data).

Cluster 1 (Red):

Cluster 1 was in the North side of hyderabad. Himayathnagar and Domalguda were the two Neighborhoods that were in that cluster. Cluster 1 had very few neighbourhoods,but a good number of chinese restaurants. Cluster 1 had the 2nd highest average of Chinese Restaurants equating to 0.080. The reason why the average of Chinese Restaurants is the high is because all these Restaurants are in two neighborhoods, Himayathnagar and Domalguda.

Cluster 2 (Blue) :

There was a total of 118 neighborhoods. Therefore, the average amount of Chinese Restaurants that were near the venues in Cluster 2 is the lowest,but highest number of chinese restaurants is concentrated only in few places like Jubilee Hills,Chintalbasti,Somajiguda,Banjara Hills,Red Hills.Even though these places have highest count,all the other places have very few ,so the average is very low in cluster 2.

Cluster 3(Aquamarine):

Cluster 3 has the highest average number of Chinese Restaurants,but not many neighbourhoods.

In the map we can see that nodes of Cluster 3 were dispersed all throughout Hyderabad making it one of the most sparsely populated cluster.Neighborhoods such Uppal,Jahanuma,Musheerabad and many more were included in this cluster.

Cluster 4 (Dark Khaki):

Cluster 4 venues were located in Begumpet,Panjagutta,Kothapet,KPHB Colony and so on.This made up the second least average of Chinese Restaurants in that cluster which was approximately 0.035.

Therefore, the ordering of the average Italian Restaurant in each cluster goes as follows: 1. Cluster 3 (≈0.180),Cluster 1 (≈0.080) ,Cluster 4 (≈0.035),Cluster 2 (≈0.010) ,

Discussion:

Most of the Chinese Restaurants are in cluster 3 represented by the Aquamarine clusters. The Neighborhoods located in this cluster,that have the highest average of Chinese Restaurants are Himayathnagar and Domalguda . Even though there is a huge number of Neighborhoods in cluster 2, there is little to no Italian Restaurant. We see that in the Ameerpet,Panjagutta area (cluster 4) has the second least average of Chinese Restaurants. Looking at the nearby venues, the optimum place to put a new Chinese Restaurant is in Central Hyderabad as there are many Neighborhoods in the area but little to no Chinese Restaurants therefore, eliminating any competition. The second-best Neighborhoods that have a great opportunity would be in areas such as Ameerpet,Panjagutta,KPHB colony which is in Cluster 4. Having 118 neighborhoods in the area with no Chinese Restaurants gives a good opportunity for opening a new restaurant. Some of the drawback of this analysis are – the clustering is completely based on data obtained from Foursquare API. Also, the analysis does not take into consideration the Chinese population across neighborhoods as this can play a huge factor while choosing which place to open a new Chinese restaurant. This concludes the optimal findings for this project and recommends the entrepreneur to open an authentic Chinese restaurant in these locations with little to no competition.

Conclusion:

In conclusion, to end this project, we had an opportunity on a business problem, and it was tackled in a way that was similar to how a genuine data scientist would do. We utilized numerous Python libraries to fetch the information , to control the content and to break down and visualize those datasets. We have utilized Foursquare API to investigate the settings in neighborhoods of Hyderabad, get great measure of data from online web pages. We also visualized utilizing different plots present in seaborn and matplotlib libraries. Similarly, we applied AI strategy to anticipate the error given the information and utilized Folium to picture it on a map. Places that have room for improvement or certain drawbacks gives us that this project can be additionally improved with the assistance of more information and distinctive Machine Learning strategies. Additionally, we can utilize this venture to investigate any situation, for example, opening an alternate cuisine or opening of a Movie Theater and so forth. Ideally, this task acts as an initial direction to tackle more complex real-life problems using data-science