Food-joints of any genre are a big hit, if placed at the right location. So, current assignment for me will focus on using Foursquare data to leverage or carve out a map of current locations of the available food-joints/restaurants to:- rightfully infer the spatial distribution of these food-joints/restaurants with their specific type/genre to spot dense and sparse distribution of restaurants type in a particular area.

Suitable theme for a Food-Joint/Restaurantaround IIT Delhi, India.

Applied Data Science Capstone

Sushant

Contents

1. Introduction

1.1 Backdrop

IIT Delhi happens to be one of the most premier engineering colleges of India. It attracts a plethora of students across the country with maximum possible cultural diversity. The population density of the area is very high as it is situated in an area called Hauz Khas, New Delhi, India which is one of the most prominent cultural melting pots of the country.

1.2 Problem

Food-joints of any genre are a big hit, if placed at the right location.

So, current assignment for me will focus on using Foursquare data to leverage or carve out a map of current locations of the available food-joints/restaurants to:- rightfully infer the spatial distribution of these food-joints/restaurants with their specific type/genre to spot dense and sparse distribution of restaurants type in a particular area. Eventually, a location and type of food-joints/restaurants can be concluded/ recommended as a prospective future site for a successful food-joint/restaurant. Target audience would be the final year undergraduates of IIT Delhi- who are willing to start their career on this front to launch their entrepreneurship endeavour on a right platform.

1.3 Interest

The scope of this assignment is not limited to areas around IIT Delhi, Hauz Khas, New Delhi, India. It can be replicated and improved upon for other cities across the world. Foursquare data for the place under consideration is a must for successful replication of this concept. Please see the Future Directions section of this paper for additional information that could be applied to situations such as that.

2. Data Acquisition and Cleaning

2.1 Data sources

This project relies entirely on Foursquare data. I limited my restrictions regarding restaurant type by using only the top-level category for “Food” within the Foursquare data. The starting point for data was IIT Delhi, India. The scope was left intentionally broad in an effort to avoid adding my own bias into the project.

2.2 Feature selection

There are only 4 fields within the results which we need to start our analysis with:-

1. The name of the venue,
2. the category of the venue,
3. and the latitude and longitude (location) of the venue.

“k-means” clustering based on the locations of each restaurant will be carried on the data, hence location parameter is of paramount importance. Other parameters of Foursquare data is not required for the current assignment

2.3 Data cleaning

We make API calls to collect Foursquare data and get the results in JSON format. These results have a lot of meta-data which needs to be removed before we can effectively use the data. Additionally, we want to get that information into a dataframe to leverage different python packages in the analysis.

The JSON data was casted into a data-frame with four columns. In the first step, relevant data were assigned to a variable. Then, remaining data was converted into a data-frame to have them in column based format as discussed in the previous section.

The name, latitude, and longitude fields were extracted. Category field was further converted into a data-frame for a better understanding and analysis.

Eventually, we have a data-frame with four columns, on which subsequent analysis will be done.

3. Exploratory Data Analysis

3.1 Understand the data

Since I approached this project with an intentionally broad scope regarding the outcome, I found it very valuable to visualize the data as I went through the process. Additionally, I collected some different statistics regarding the data. I spent a lot of time slicing and viewing this data in an effort to ‘gut-check’ each step as I went.

3.1.1 Can the data give us the answer we are looking for

Right at the onset, I could see that IIT Delhi has 100 different restaurants that fell into 32 categories. This was an important insight because it allowed me to see that there would be some good groupings among the restaurants. If it was 100 restaurants with 100 categories, this project approach would not have been successful because we would not see any overlap in restaurant type between whatever clusters we come up with. If it was the opposite (100 restaurants in 1 category) it would have proven equally difficult because we would see complete overlap between our clusters and therefore identifying opportunity would be difficult or impossible.

3.1.2 Visualize the data

The next step in my exploration was to visualize the 100 restaurants on a map. I performed this step to see if clustering still felt like the right approach. Though the data is pretty spread out, as I reviewed the map, I was able to see that there were definite areas where I could see the data clustering. In Figure 1, the red circle is the location of IIT Delhi which we used as the starting point. The blue circles are the locations of each restaurant returned by Foursquare.

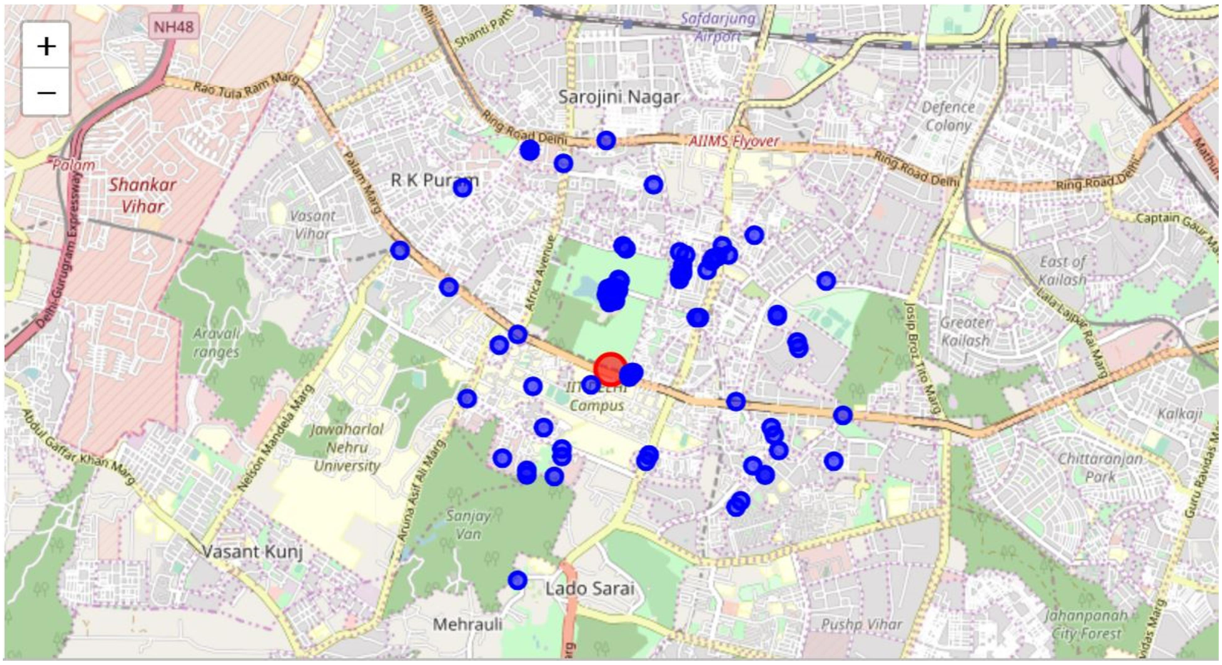


Figure 1 – Restaurants around IIT Delhi
(IIT Delhi is earmarked in red & restaurants are shown in blue).

3.2 How many clusters

I've established that the data seems suitable for clustering, but one challenge with k-means clustering is determining how many clusters to use. One approach here is called the “Elbow Method”. In this method, I visualize the sum of squared differences within the latitude and longitude over different values of k. The resulting chart should look like a bent arm and the location of the ‘elbow’ would represent the optimal k. The principle here is that the steeper portion of the graph has significant reductions in the sum of squared differences, but after the ‘elbow’ the improvements in reductions becomes less and less for each additional k. The best result is a very distinct elbow, but this was not the case in my data. (This can happen with data that does not have very distinct clusters.) Figure 2 shows the results from performing the Elbow Method.

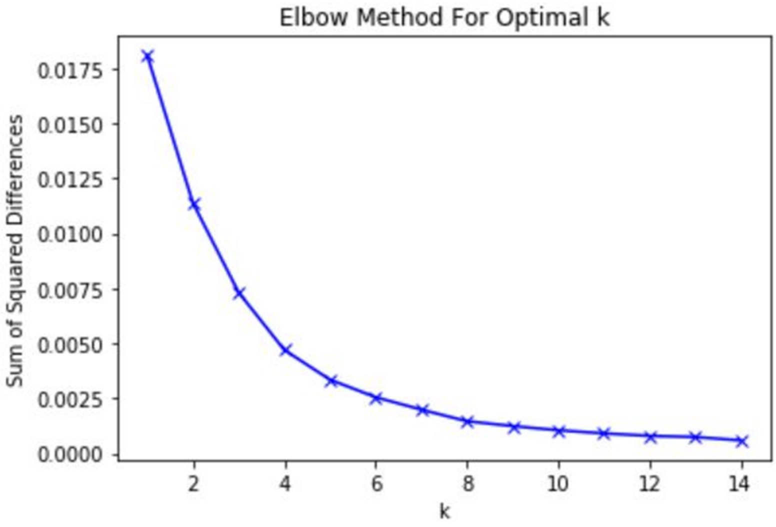


Figure 2 – Elbow Method for optimal k.

I was not satisfied with a distinct break within this chart. I see a break at k=2, but I also still see pretty significant reductions in the sum of squared differences for k = 4 and 5. So rather than looking at the raw data, I looked at the rate of change and determined that after k=4 there was a lot less improvement, and so I settled on using k=4 for further analysis.

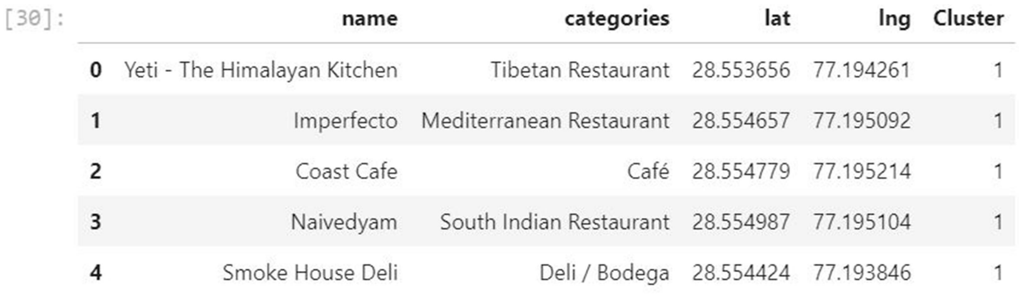
4. Predictive Modelling

4.1 Classification models

4.1.1 Setting up k-means clustering

Given the optimal k, I used the “scikitlearn”python package to perform 15 iterations of k-means clustering in an attempt to find the optimal centroids for each of my four clusters.

To perform this step, I isolated only the latitude and longitude of each restaurant and ran k-means clustering with k=4 and n_init = 15. The array that is returned by this process labels each of the restaurants with a category 0 through 3. From there, I add these labels back to my original dataframe to see which cluster each restaurant belongs in. Figure 3 shows my current dataframe with this information added.

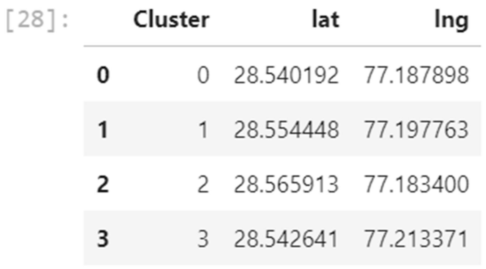


[30]:

	name	categories	lat	lng	Cluster
0	Yeti - The Himalayan Kitchen	Tibetan Restaurant	28.553656	77.194261	1
1	Imperfecto	Mediterranean Restaurant	28.554657	77.195092	1
2	Coast Cafe	Café	28.554779	77.195214	1
3	Naivedyam	South Indian Restaurant	28.554987	77.195104	1
4	Smoke House Deli	Deli / Bodega	28.554424	77.193846	1

Figure 3 – nearby_venues dataframe with the addition of “Cluster”

From here, I need to identify the centroids themselves. This process sets the centroid for each cluster at the center of each cluster, so the easiest way to identify the location of the centroids is just to look at the mean latitude and longitude values for each of the clusters. Figure 4 shows the results of the means of each cluster.



[28]:

	Cluster	lat	lng
0	0	28.540192	77.187898
1	1	28.554448	77.197763
2	2	28.565913	77.183400
3	3	28.542641	77.213371

Figure-4- Centroid locations

4.1.2 Visualizing the clusters

At this point, it’s important to validate that the steps I’ve taken up to this point make sense. The easiest way to do this is to visualize the previous map we produced and color-code by cluster. This allows me to validate that the clusters make sense and that I am not seeing the clusters intermingled. I additionally found it useful to plot the centroids. Figure 5 is the updated map. IIT Delhi and the centroids are not part of my data from Foursquare and have been coded black and pink to offset them from the rest of the data. IIT Delhi is pink circle, black fill. Centroids are black circle, pink fill. The venues themselves are rainbow colored, assigned dynamically in the plot. Cluster 0 is red, cluster 1 is purple, cluster 2 is teal, and cluster 3 is olive.

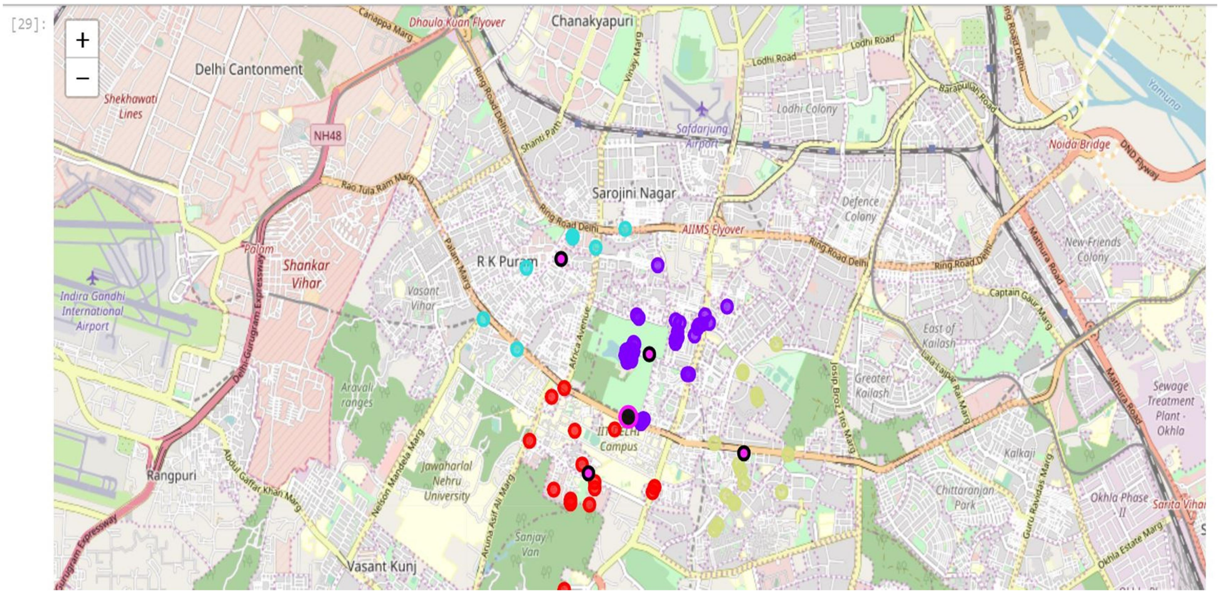
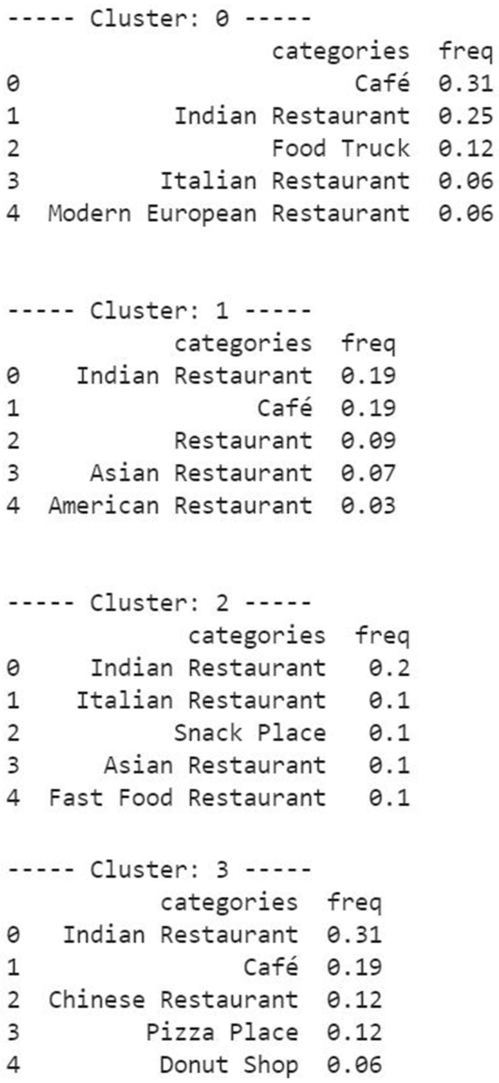


Figure 5 –clusters of restaurants around IIT Delhi, India.

4.1.3 Examining the clusters

The overall goal is to determine where a restaurant category is under-represented in one cluster, but highly represented in other clusters. From there, we can establish that X type of restaurant should be located in Y cluster. I need to examine the clusters more closely to properly determine the correct X and Y in this situation.

To do this, I need to 'one-hot' encode all the categories into dummy variables so I can look at the mean representation of each category across each of the clusters. From there, I am able to report on the top 5 categories for each of the clusters.



```
----- Cluster: 0 -----
                  categories  freq
0                       Café  0.31
1          Indian Restaurant  0.25
2                 Food Truck  0.12
3         Italian Restaurant  0.06
4  Modern European Restaurant  0.06


----- Cluster: 1 -----
            categories  freq
0    Indian Restaurant  0.19
1                 Café  0.19
2           Restaurant  0.09
3     Asian Restaurant  0.07
4  American Restaurant  0.03


----- Cluster: 2 -----
             categories  freq
0     Indian Restaurant   0.2
1    Italian Restaurant   0.1
2           Snack Place   0.1
3      Asian Restaurant   0.1
4  Fast Food Restaurant   0.1

----- Cluster: 3 -----
           categories  freq
0   Indian Restaurant  0.31
1                Café  0.19
2  Chinese Restaurant  0.12
3         Pizza Place  0.12
4          Donut Shop  0.06
```

Reviewing the frequencies of each category :-

1. **Café is the 2nd,2nd & 1st most common venue in cluster 0, cluster 1 & cluster 3 respectively,**
2. **& It doesn't even make the top 5 in Cluster 2.**
3. **Hence, the budding entrepreneur is recommended for a Café within the area defined as Cluster 2.**

5. Conclusions

1. Foursquare data was used to determine the profile of restaurants /food-joints in and around IIT Delhi.
2. **Café** is the **2nd,2nd & 1st most common venue** in **cluster 0, cluster 1 & cluster 2** respectively,

3. & **It doesn't even make the top 5 in Cluster 3.**
4. Hence, the budding entrepreneur is **recommended** for a **Café within the area defined as Cluster 3.**

6. Possible further refinements

Possible improvement avenues can be corroborated as per the following analysis cum suggestion.

This approach only capitalizes on location. Factors such as migration, demographic profile, crime rate etc. may be incorporated for a comprehensive analysis in the future.

Optimal k could be found out using a better approach as a better distinct break in the elbow might be deciphered, provided a distinct location with heavy clustering is available for interpretation.

References:

1. https://foursquare.com/developers/apps
2. https://en.wikipedia.org/wiki/Indian_Institute_of_Technology_Delhi