UTKRISHT INTERNSHIP REPORT

Classifying Geolocational data

Using

K-means clustering

Submitted to

Ms. Ruchi Sehrawat

USICT

Submitted by

Parth Bisht(04216403220)

INDEX

Objective :–

This project involves the use of K-Means Clustering to find the best accommodation for students in Delhi (or any other city of your choice) by classifying accommodation for incoming students on the basis of their preferences on amenities, budget and proximity to the location.

In the fast-moving, effort-intense environment that the average person inhabits, It's a frequent occurrence that one is too tired to fix oneself a home-cooked meal. And of course, even if one gets home-cooked meals every day, it is not unusual to want to go out for a good meal every once in a while for social/recreational purposes. Either way, it's a commonly understood idea that regardless of where one lives, the food one eats is an important aspect of the lifestyle one leads.

Now, imagine a scenario where a person has newly moved into a new location. They already have certain preferences, certain tastes. It would save both the student and the food providers a lot of hassle if the student lived close to their preferred outlets. Convenience means better sales, and saved time for the customer.

Food delivery apps aside, managers of restaurant chains and hotels can also leverage this information. For example, if a manager of a restaurant already knows the demographic of his current customers, they'd ideally want to open at a location where this demographic is at its highest concentration, ensuring short commute times to the location and more customers served.If potential hotel locations are being evaluated, a site that caters to a wide variety of tastes would be ideal, since one would want every guest to have something to their liking.

Introduction :-

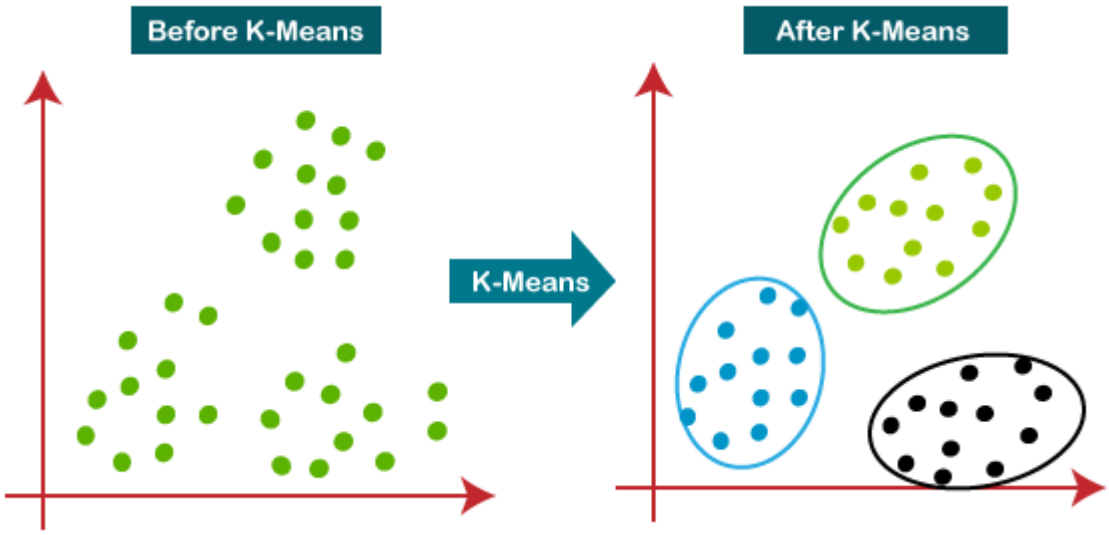
K-means clustering –

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

"It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties."

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



In our case, we will be classifying geolocational data on the basis of their distance to amenities using the k-means clustering algorithm.

	lat	lng	Restaurants	Fruits,Vegetables,Groceries	Cluster
1	28.596122	77.045318	12	6	0
2	28.603650	77.040910	12	4	0
3	28.604821	77.042990	12	4	0
4	28.593309	77.047081	12	5	0
5	28.603310	77.051254	12	4	0
7	28.577158	77.047214	8	6	1
8	28.588090	77.063409	12	6	0
9	28.560323	77.054060	12	7	0
10	28.562868	77.051080	7	6	1
11	28.562260	77.055954	7	7	1
12	28.583254	77.066298	16	7	2
13	28.579831	77.075499	16	7	2

Here each location has been clustered regarding how many restaurants and groceries are in their vicinity.

GGSIPU Locational analysis overview –

Finding apartments for students near GGSIPU according to their wants and then classifying the apartments most rich in amenities to those which are not using k-means clustering.

Planning :-

Software requirements –

IDE –

1. Jupyter Notebook

Libraries –

1. Pandas : For data analysis and manipulation
2. Matplotlib : For creating different type of visualizations
3. Sklearn : For machine learning tools such as k-means and confusion matrix

Etc.

Foursquare API : for geolocational data obtained regarding query

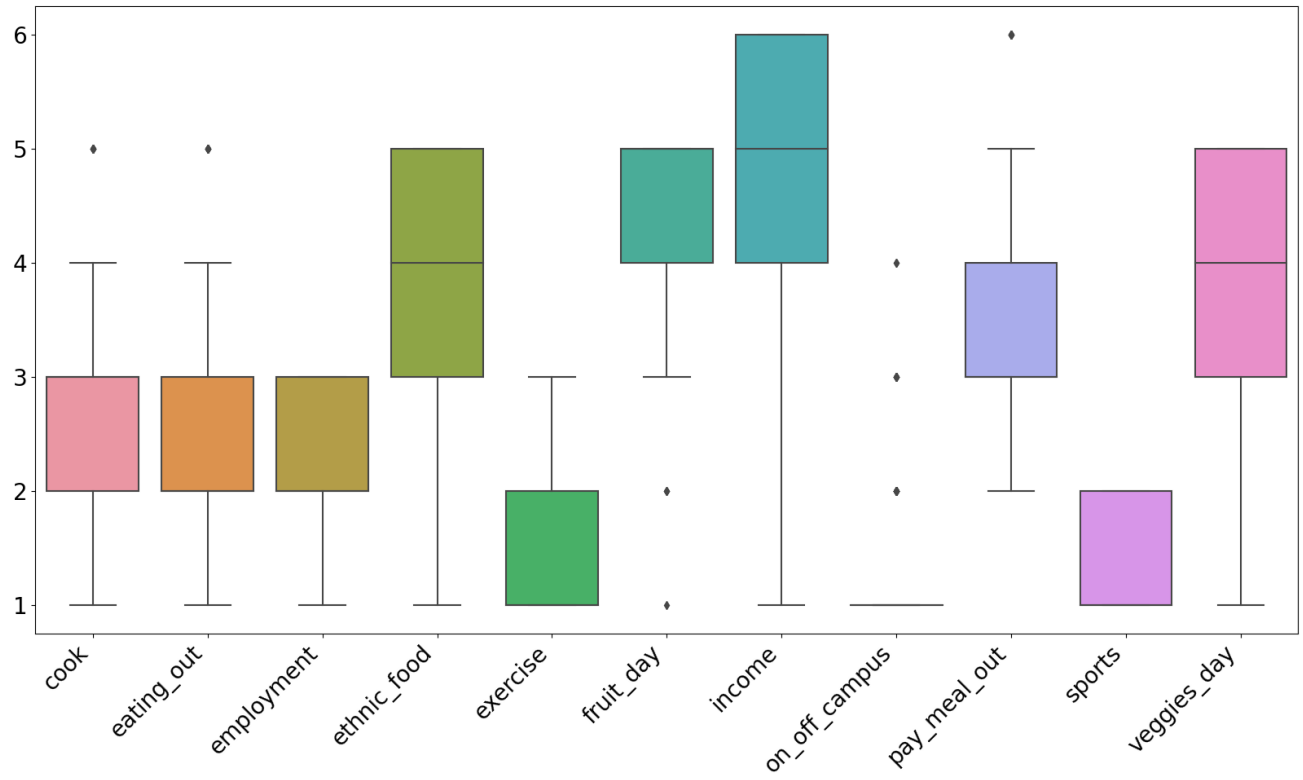
Methodology –

1. Fetching and cleaning data

   a. Head over to this link to get a dataset we'll be using.
   b. Explore the codebook_food file. There are around 70 parameters. Not all of them are relevant. Think carefully about which ones are the most useful. Which ones can be used to quantifiably

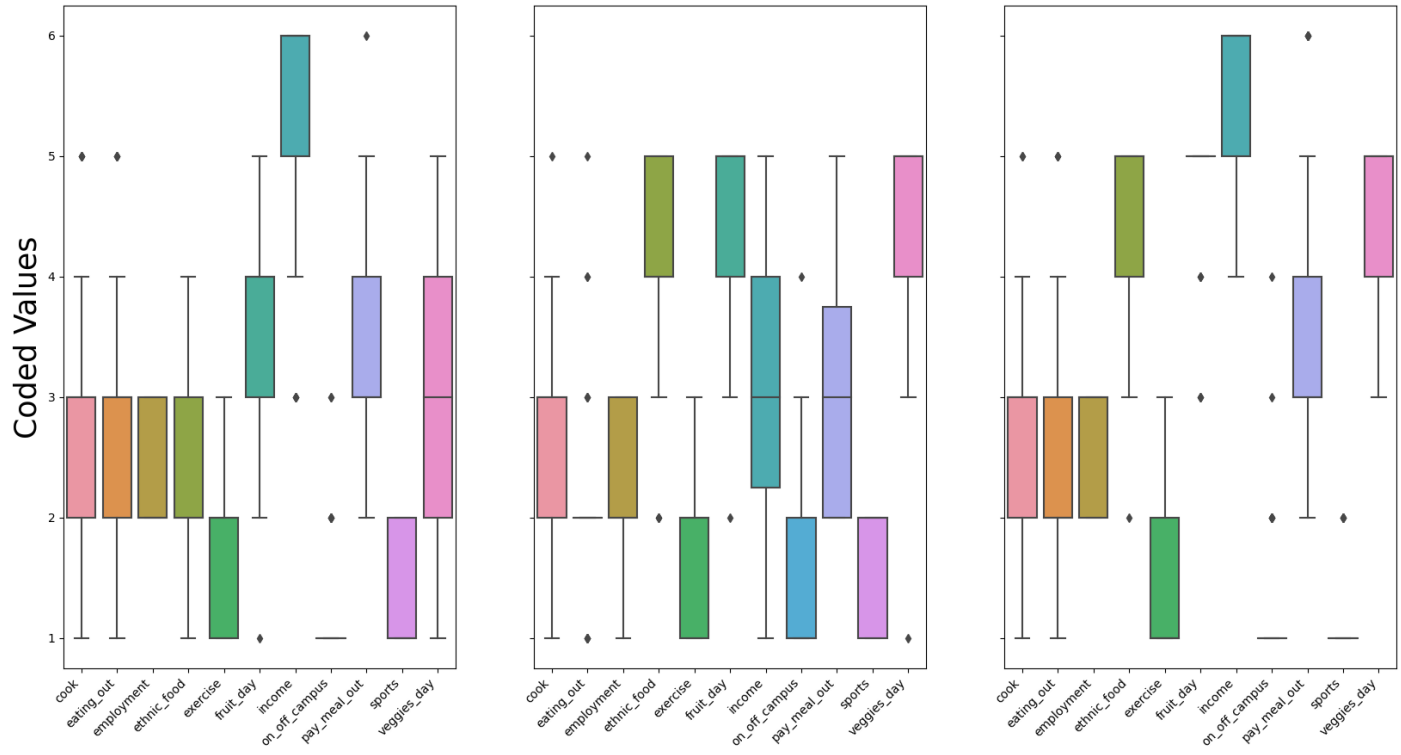
differentiate students? A good example of this is income. A more qualitative parameter like "How much they like vegetables" on a 1-5 scale might not be as useful.

c. Extract the most relevant features into a pandas dataframe.
d. Visualize the data via a boxplot.



2. Run KMeans Clustering on the data

Plot a boxplot of the clustered data so that we can observe difference and draw insight among students and the amenities they desire.



High income students in general are:

1.>Much more likely to stay on campus.

2.>Less likely to cook.

3.>Eat out more often.

In addition, a subset of higher income students: 1.>Eat out less often.

2.>Have a healthier, more varied diet in general. (Eat ethnic food, fruits, vegetables etc.)

Lower income students are more likely to:

1.>Eat more vegetables

2.>Eat more fruits.

3.>More likely to cook, so more likely to stay off campus.

4.>Obviously, pay less for meals out.
Essentially, we need to think about accomodation for 2 types of students, low income and high income(varied diet).

The high income(fixed diet) students will ideally stay on campus, so they don't have to be taken into account.

3. Get Geolocational Data from Foursquare API

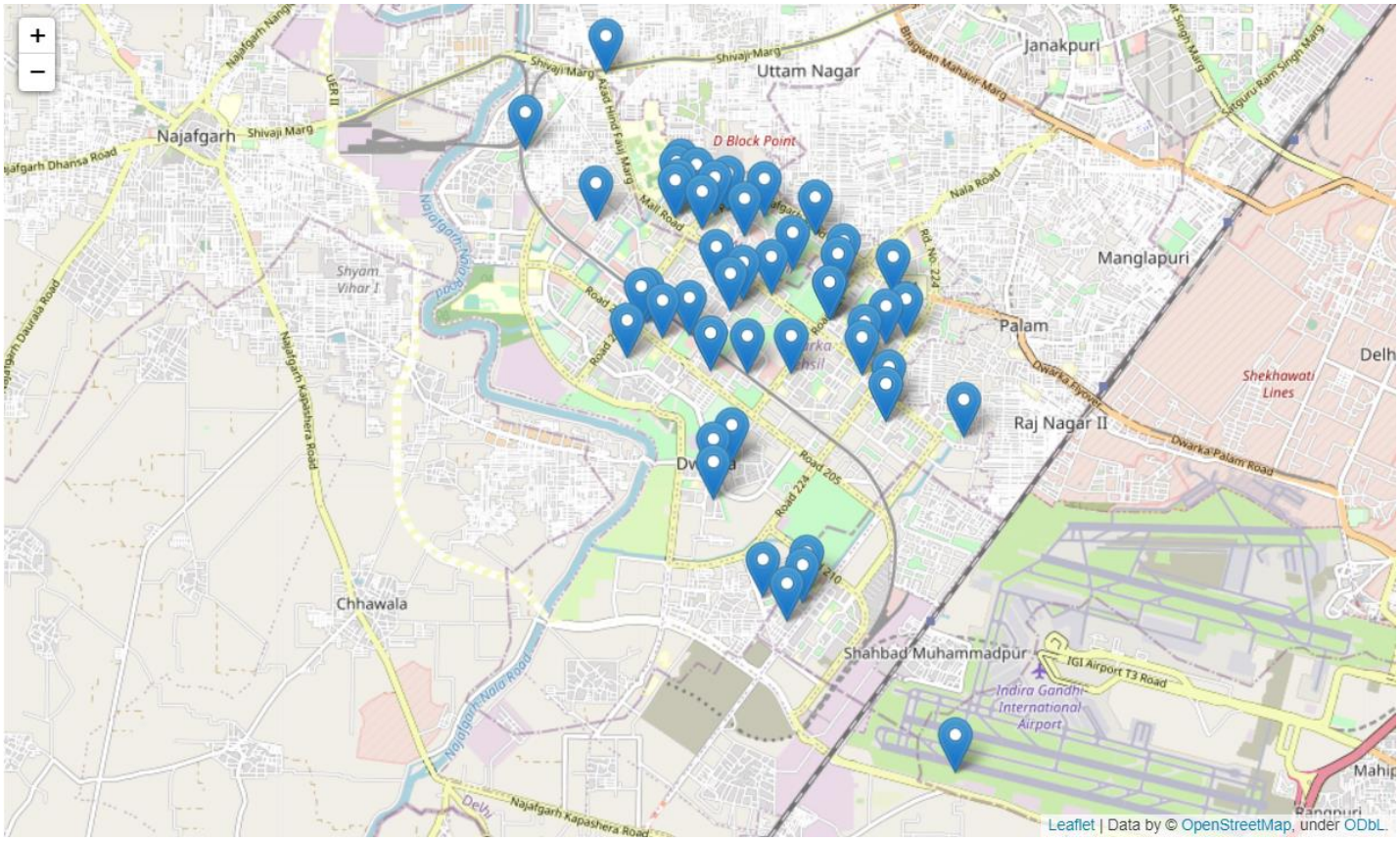
a. Make a free foursquare account and get your API credentials set up. Set up your query in such a way that you can check for residential locations in a fixed radius around a point of your choosing. For example, you can pick (28.594899, 77.021327) location of GGSIPU. Here's how the API response might look:

	id	name	categories	referralId	hasPerk	location.address	location.lat	location.lng
0	4edb45ac0aaf49e02aa30cc3	Ispatika Apartment	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	v-1663246177	False	Sec 4, Dwarka	28.601269	77.048880
1	5184977d498ec7f6416fdcba	Shivam,Apartment, Sec 12	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	v-1663246177	False	Dwarka	28.596122	77.045318
2	4efe8845b6346feaeba89dc0	National Apartment Sector 3, Near sec3/13 traf...	[]	v-1663246177	False	Sector 3 Dwarka	28.603650	77.040910
3	4efe88f3775bec6b42df0632	national apartment sector 3 dwarka	[]	v-1663246177	False	NaN	28.604821	77.042990
4	516a4e8ee4b01de80d736d8e	gokul apartment sector 11	[{'id': '4d954b06a243a5684965b473', 'name': 'R...	v-1663246177	False	NaN	28.593309	77.047081

b. Hit the endpoint, and parse the response data into a usable dataframe. There is a lot of information you don't need, so apply the same data cleaning principles you used earlier.

c. Similarly search for amenities such as restaurants and groceries regarding those locations.

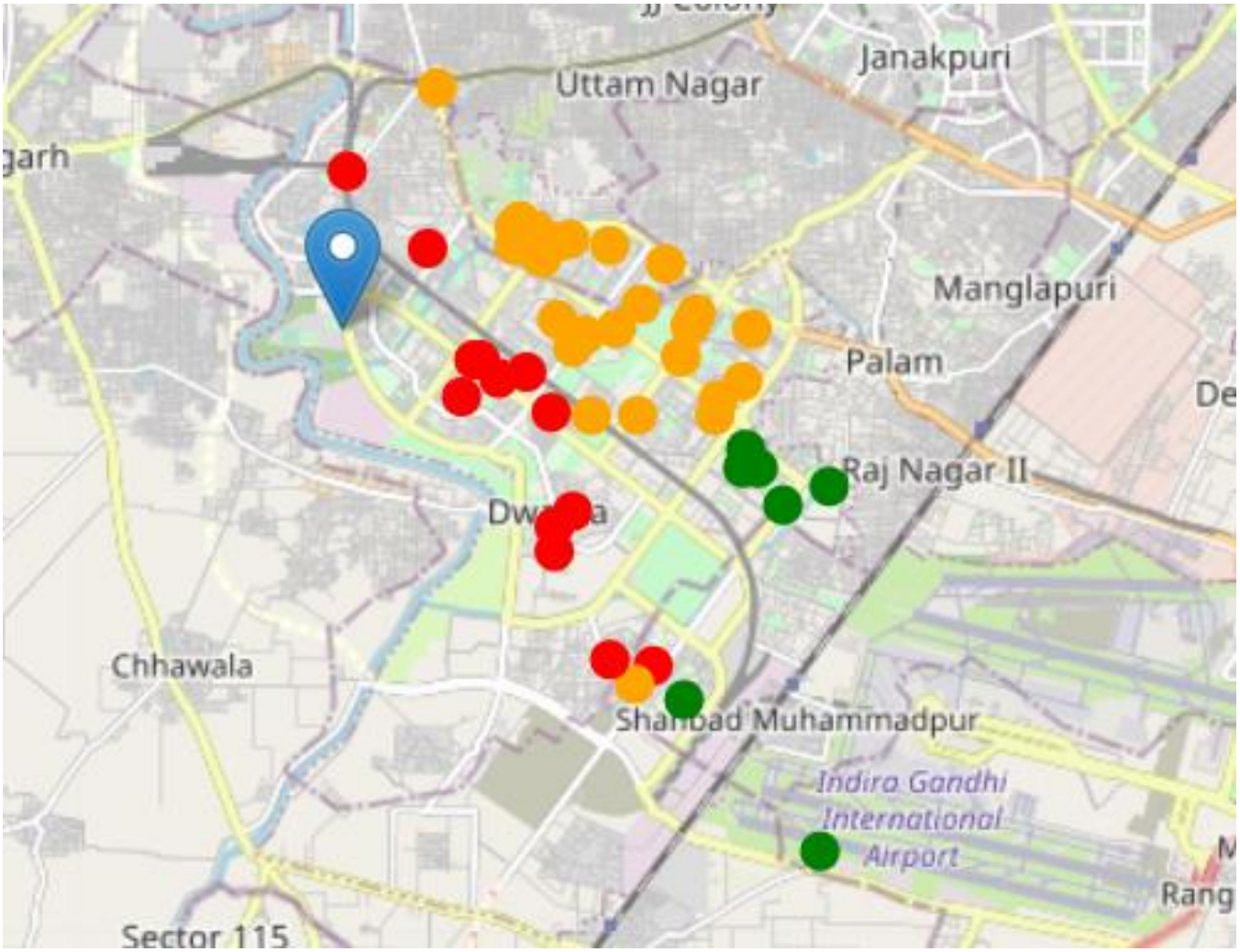
d. Clean up the data and deal with the null values and plot on a map



4. Clustering the data and ploting on map

a. Apply k-means clustering on the geolocational data and plot on map.

	lat	lng	Restaurants	Fruits,Vegetables,Groceries	Cluster
1	28.596122	77.045318	12	6	0
2	28.603650	77.040910	12	4	0
3	28.604821	77.042990	12	4	0
4	28.593309	77.047081	12	5	0
5	28.603310	77.051254	12	4	0
7	28.577158	77.047214	8	6	1
8	28.588090	77.063409	12	6	0
9	28.560323	77.054060	12	7	0
10	28.562868	77.051080	7	6	1
11	28.562260	77.055954	7	7	1



Applying K-Means, we find 3 prominent clusters:

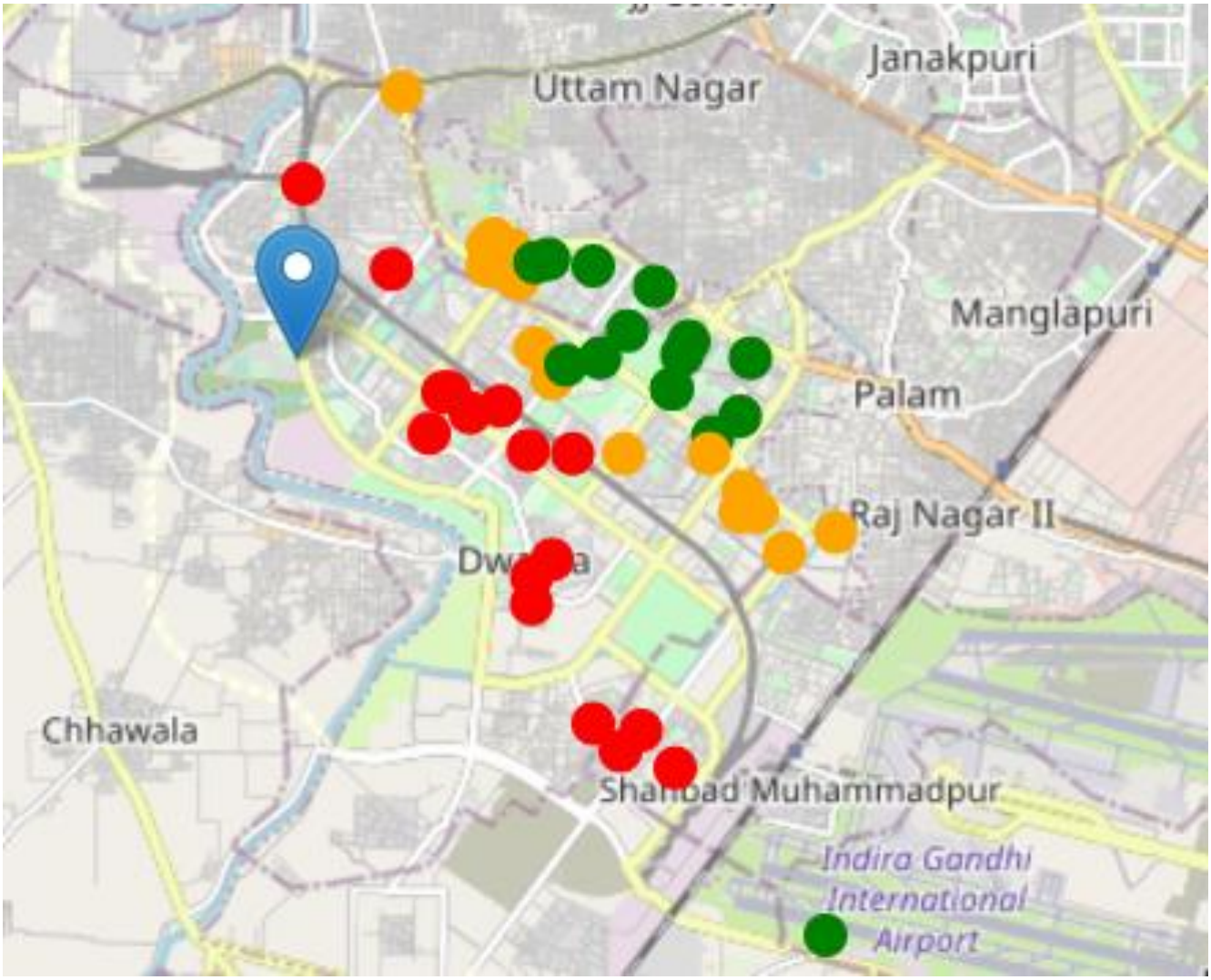
Cluster 0(Green) Where both (fruits and vegetables) and (restaurants) are abundant

Cluster 1(Yellow): Restaurants are plentiful, but groceries less so.

Cluster 2(Red): Restaurants and groceries are relatively hard to find.

Blue marker – GGSIPU

b. Cleaning the foursquare api response and removing results outside of the distance limit yields us different clusters.



Applying K-Means, we find 3 prominent clusters:

Cluster 1(Green) Where both (fruits and vegetables) and (restaurants) are abundant

Cluster 0(Yellow): Restaurants are plentiful, but groceries less so.

Cluster 2(Red): Restaurants and groceries are relatively hard to find.

Blue marker – GGSIPU

These are the clusters obtained after we clean up responses from the foursquare api and exclude results outside the specified range.

Code

We have our code at github repository :
https://github.com/Parthbisht16/GGSIPU-locational-analysis

Conclusion

One can easily notice, the further away from the college, the more options one finds for food.

The same can be said about other amenities as well.

One thing I would like to note is that the foursquare data seems incomplete and at maximum gives us only 50 search results; Many locations seem to be missing or ill-classified leading to poor clustering which is only 53% accurate if results beyond the distance limit are not excluded.

```
ConfusionMatrix
 [[ 1 14  0]
 [ 5 11  0]
 [ 1  2 13]]
              precision    recall  f1-score   support

           0       0.14      0.07      0.09        15
           1       0.41      0.69      0.51        16
           2       1.00      0.81      0.90        16

    accuracy                           0.53        47
   macro avg       0.52      0.52      0.50        47
weighted avg       0.52      0.53      0.51        47
```

Reference

food_coded.csv - Values regarding habits of students regarding various parameters https://github.com/Parthbisht16/GGSIPU-locational-analysis/blob/main/References/food_coded.csv

codebook_food.docx - Document explaining the meaning of values in food_coded.csv https://github.com/Parthbisht16/GGSIPU-locational-analysis/blob/main/References/codebook_food.docx

K-means clustering https://towardsdatascience.com/machine-learning-algorithms-part-9-k-means-example-in-python-f2ad05ed5203

Foursquare API Overview (foursquare.com)