

Capstone Project

Applied Data Science Capstone by IBM/Coursera

Han, Donghee

Table of contents

- Introduction: Business Problem
- Data
- Methodology
- Analysis
- Results and Discussion
- Conclusion

Introduction : Business Problem

In this project we will try to find an optimal location for delivery service company.

This project is targeting Korean student in KAIST(Korea Advanced Institute of Science and Technology). In Korea, many people use food delivering service. Most of restaurants used to hire their own delivery member, but these days, most of restaurants use delivery service applications.

We will assume that we are trying to make 'Drone Delivery Service' and looking for the optimal location for our 'Drone Base'. We are targeting student in KAIST so, our candidates will be the location near KAIST. But we also consider that there are many other restaurants in Daejeon city.

Data

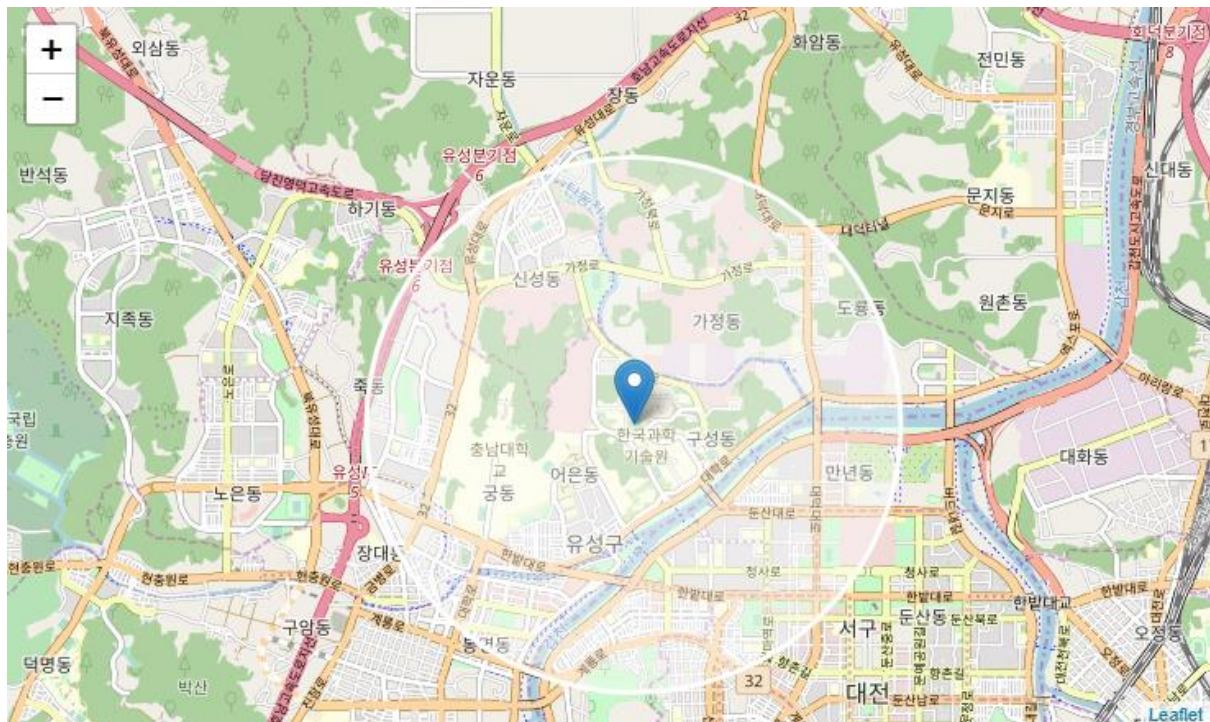
To solve our Business Problem we need to figure out the factors below

- number and location of restaurants clusters near KAIST
- distance of clusters from KAIST
- number of Chinese restaurants in the clusters (Chinese food is most popular delivery food)

We will use 'Google Maps API geocoding' and 'Foursquare API' to get locations and categories of the restaurants.

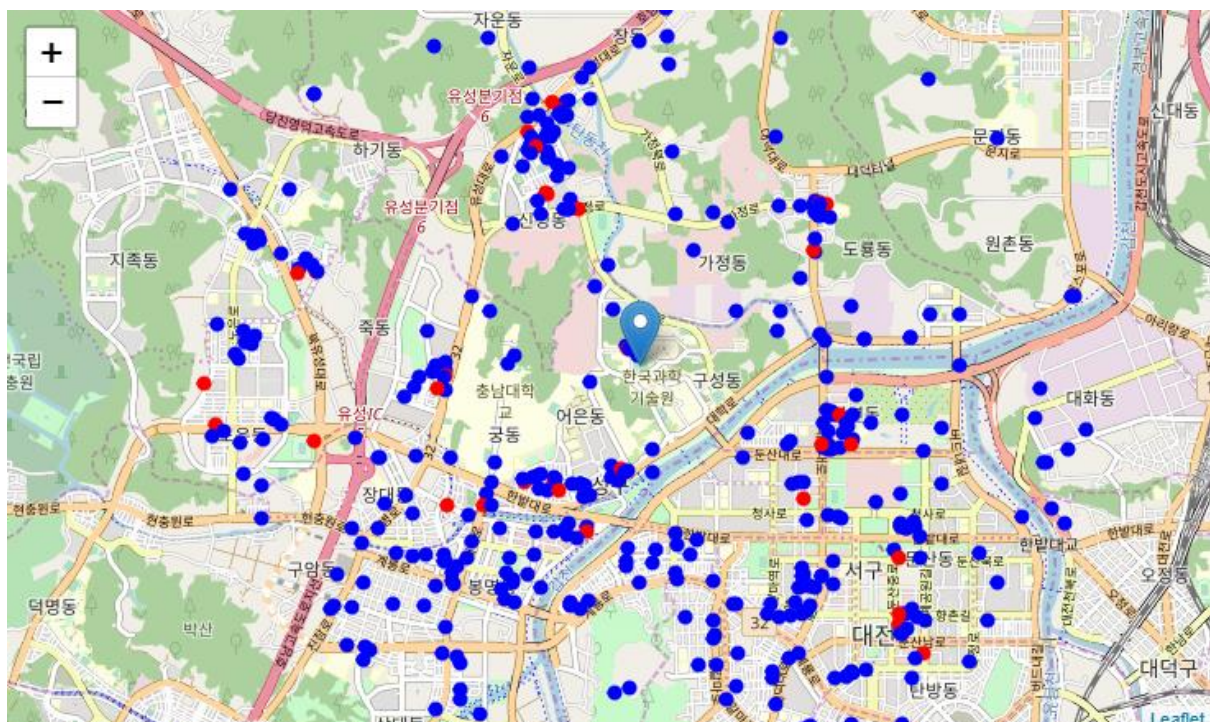
Get coordination data of KAIST from Google Maps API.

Draw map using folium and coordination data.



Get restaurant data from Foursquare. And label Chinese restaurant.

Red dots are Chinese restaurant and blue dots are other restaurant.



Methodology

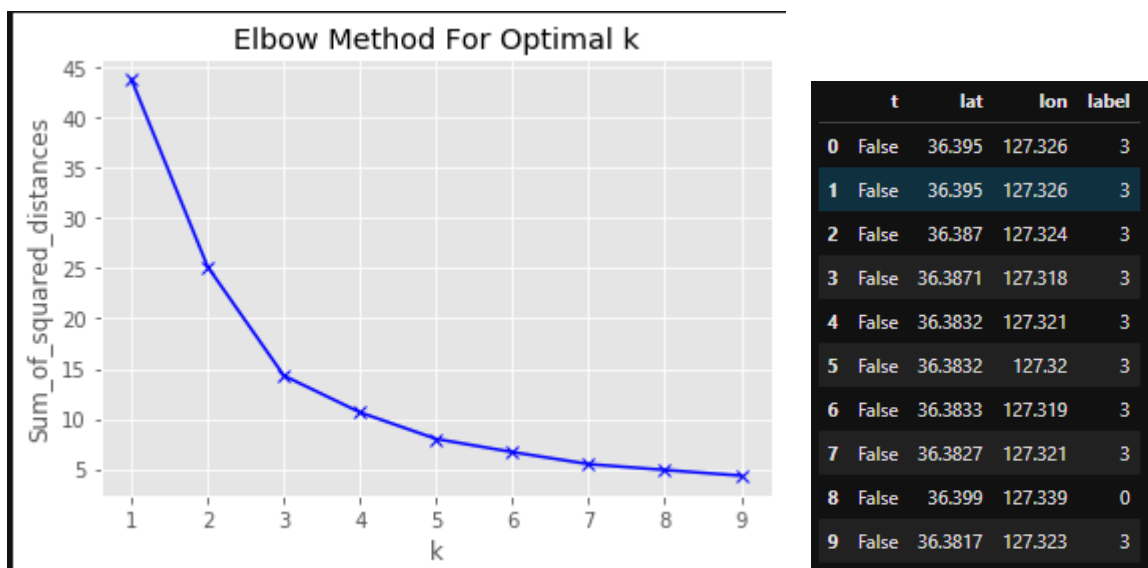
1. We will use k-means & DBSCAN to find clusters of restaurants
2. Using visualization optimize clustering.
4. find center location of the clusters and number of Chinese restaurants in the clusters
5. select important clusters
6. find best location for 'Drone Base' based on the distances from the clusters.

Analysis

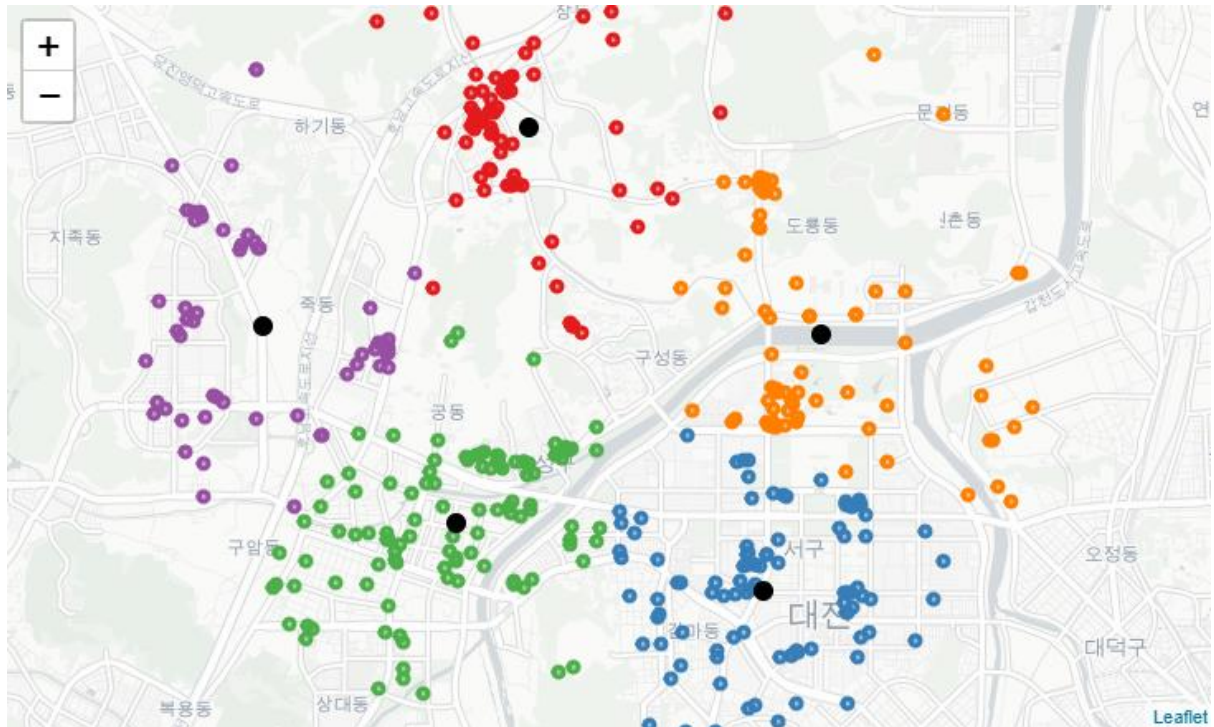
1. k-Means

Use Elbow Method for Optimal k. From this we can find the best k is 5.

Fit clusters to data on k=5 using *KMeans* in *sklearn.cluster* and label the cluster information to data.



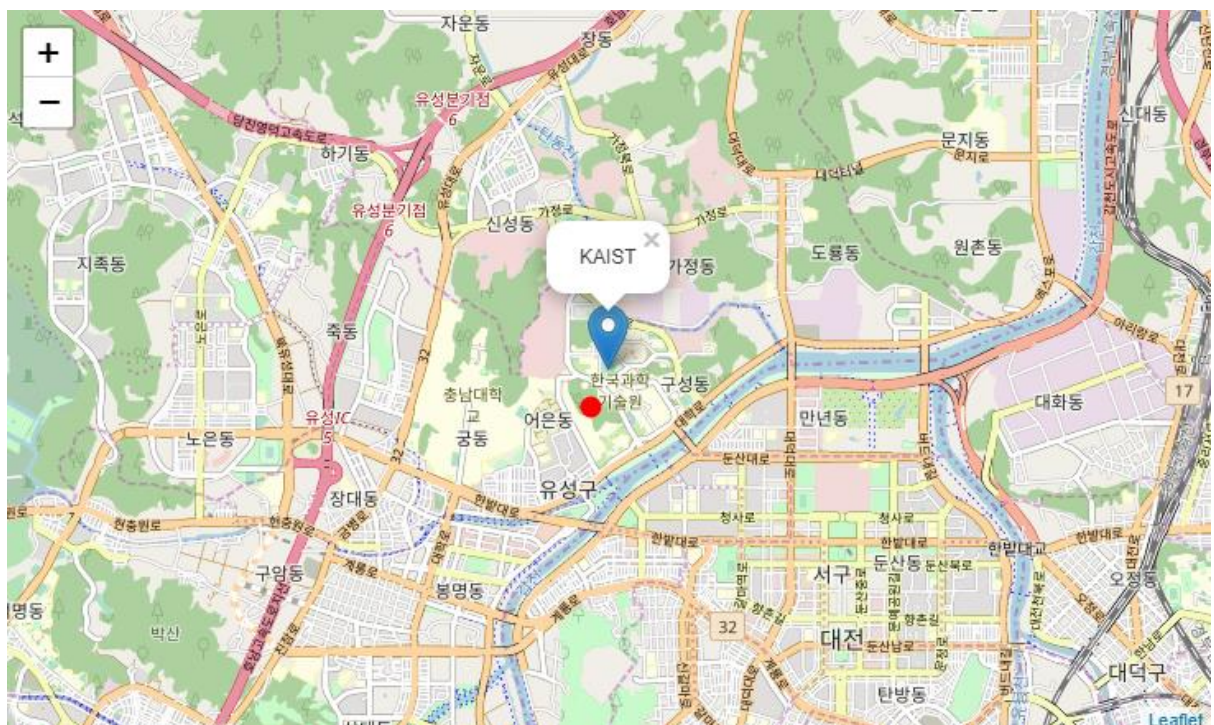
We can draw clusters on the map. Each clusters have different colors and block circle mark is the center of the cluster.



Check the number of Chinese restaurants in each clusters. It looks fine.

```
num of Chinese restaurants in 0 : 9
num of Chinese restaurants in 1 : 11
num of Chinese restaurants in 2 : 11
num of Chinese restaurants in 3 : 11
num of Chinese restaurants in 4 : 11
```

From the location data of each clusters, we can get center of the centers of the clusters.



2. DBSCAN

This time we will try DBSCAN for clustering

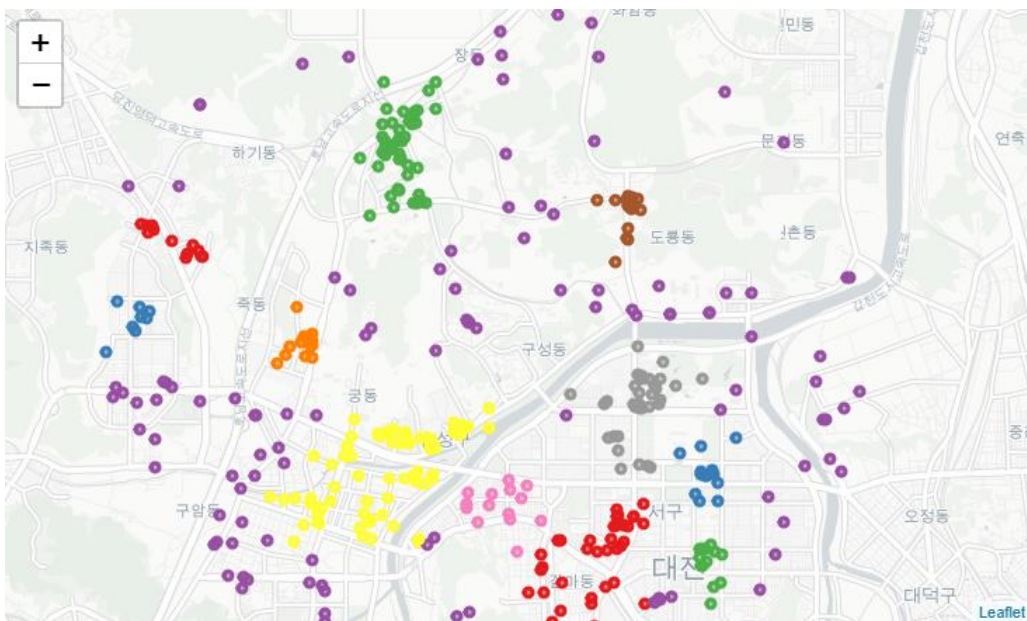
We used *DBSCAN* from *sklearn.cluster* and the parameters were like this.

eps = 0.22

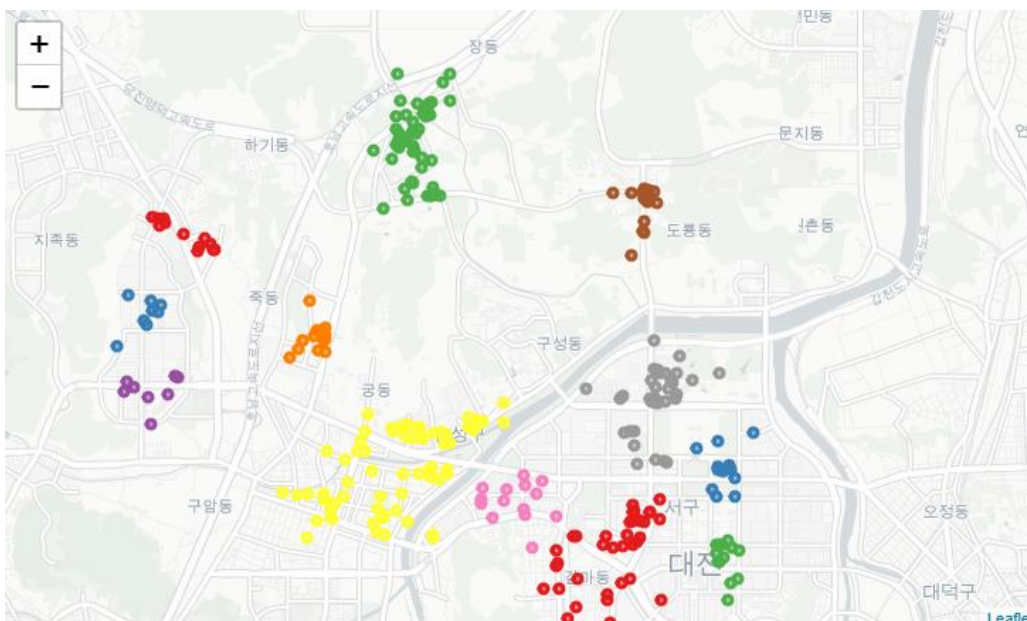
min_samples = 8

algorithm='kd_tree'

We setup parameters based on visualization analysis. This is the image of the clusters.



We remove outliers from the data.



We get 12 clusters and check the number of Chinese restaurant and total number of restaurant.

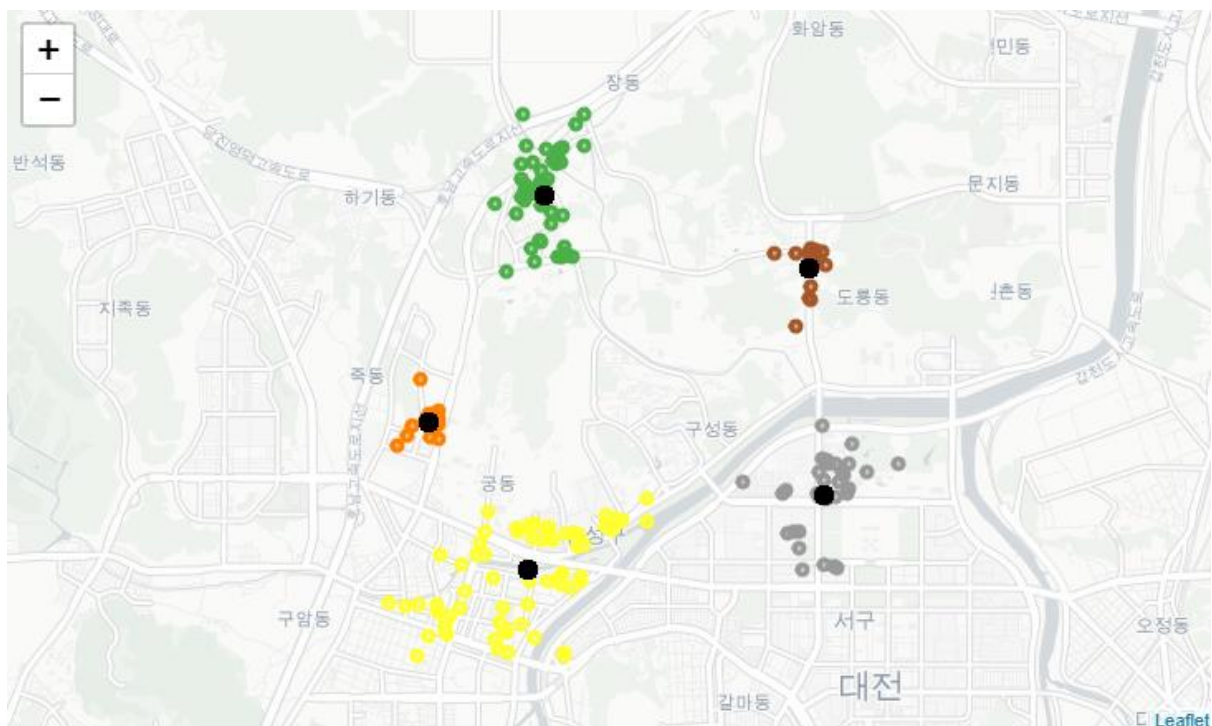
```
0: 2 / 12
1: 1 / 8
2: 7 / 47
3: 1 / 8
4: 2 / 15
5: 8 / 73
6: 3 / 16
7: 0 / 14
8: 4 / 34
9: 0 / 37
10: 1 / 14
11: 3 / 13
```

From this data we decided to remove group size under 15 and group without Chinese restaurant.

(Because, Chinese food is the best delivery food in Korea)

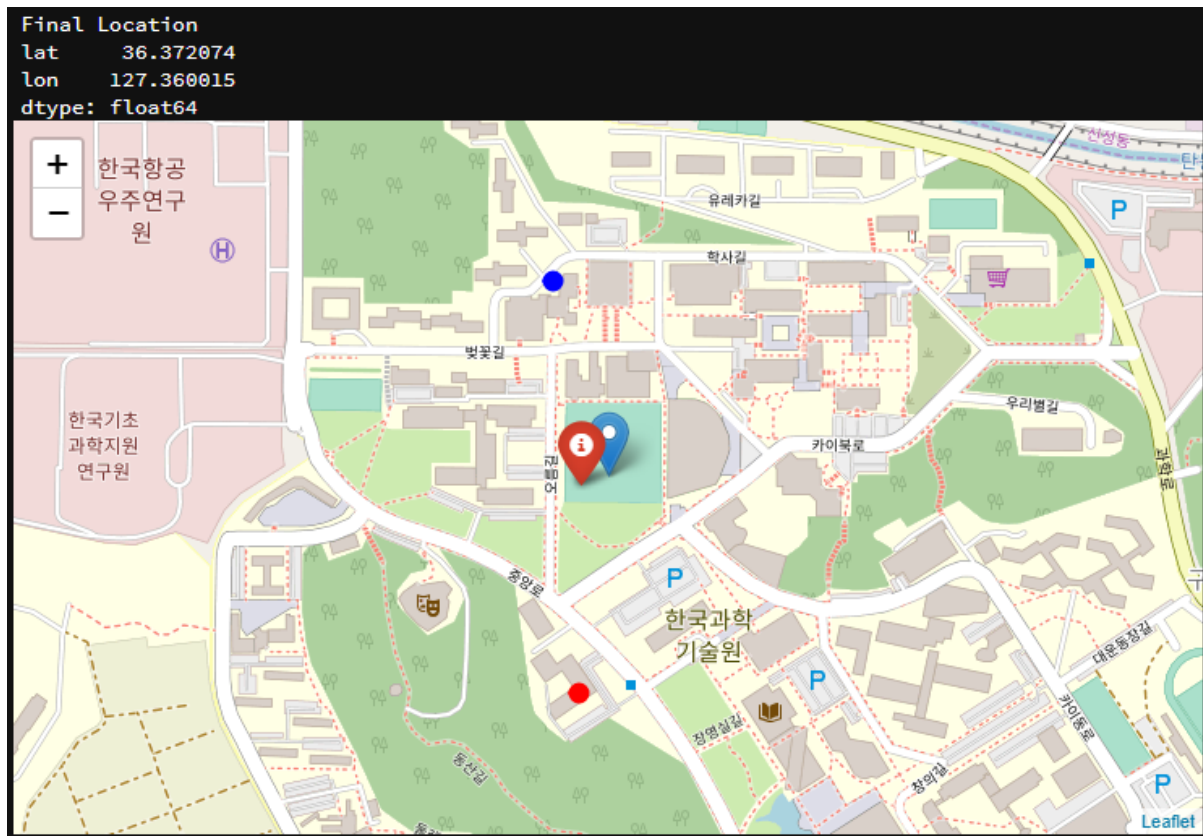
Now we have group 2,4,5,6,8.

You can see 5 clusters and each clusters center.



Now, get mean of the cluster centers.

Finally there are 4 marks on the map. Blue dot is location from DBSCAN and Red dot is From k-Means method. Blue mark is center of the KAIST and Red mark is the center of other 3location.



Results and Discussion

From this project we figure out two place for the Drone Base.

1. from k-means : [36.369084 , 127.358620]

2. from DBSCAN : [36.374233 , 127.359663]

We find restaurants clusters around KAIST based on data from Foursquare.

And also, mark Chinese restaurants to figure out the cluster is meaningful for us.

In case of k-means, we optimized the 'k' using elbow method. But, we couldn't remove outliers.

In case of DBSCAN, we remove outliers. But, it was hard to optimize the parameters. So, we changed the values for parameters based on visualization analysis.

We select the center of the centers of the clusters. We also have to consider the distance from KAIST. So, we get mean of two place and KAIST.

The red marker is the Final Location.

Final Location : [36.372074 , 127.360015]

Conclusion

From this project we figure out suitable place for the Food Delivery Service Drone Base.

We get data from Google API and Foursquare. From the data from Foursquare, we can get data of restaurants around KAIST.

Using k-means & DBSCAN, we find out main restaurants clusters around KAIST.

And count the number of Chinese restaurants in the clusters to make sure the cluster is important for our project.

and the result of k-means & DBSCAN was different but, the center of the clusters center was almost same. And optimizing in DBSCAN was based on visualization tool. So, it's hard to say both methods are perfect or mathematically right.

But, it true that from this project we get meaningful data for the business.

If we try to develop this project, we can consider about other data in real world.

Thank you