Coursera Capstone Project

IBM Applied-Data-Science-Capstone

Analyzing the places near office spaces to open new restaurants in Bangalore.

By Niteeka Prashant

April 2020

1.) Introduction

1.1 Background

This Coursera Capstone Project is for IBM Data Science Final Assignment. In this project we have taken a scenario where a person wants to open a new restaurant in the nearby space of an IT office in Bangalore. As opening the new restaurant in these areas is costly affair, it becomes important to analyze the places and existing restaurants around the offices to get a clear idea of the competition there. With the help of Machine Learning, we can not only determine the probable areas but the cuisines can also be analyzed. This project involves gathering location information of the IT offices spaces and using Foursquare APIs to get the location based details. Using the nearby venues we can divide the places in clusters to determine the best options.

1.2 Business Understanding/Problem Description

You need a spot that draws crowds, is easily accessible, and has the potential for growth. It makes sense to take your time, as you're looking for the right space. You might also want to do some research to find out spaces for restaurants that describe the market conditions which exist in the location. Whether you decide to rent space or build from scratch, selecting a location is one of the biggest decisions you'll make as an owner.

1.3 Target Audience

The aim of doing a target audience analysis is to know which portion of the population is most likely to come to your restaurant. As we have taken the place to be near and IT office our population we can explore the related options and cuisines. The target audience is the new business willing to open new restaurant.

2.) Data

In this assignment first we have to clean the data obtained from Web page (here) for the Bangalore city and using web scraping method. Then convert addresses into their equivalent latitude and longitude values. Use Foursquare API to explore neighborhoods in Bangalore City. Get the most common venue categories in each neighborhood, and then use this feature to group the neighborhoods into clusters. Use the k-means clustering algorithm to complete this task. Finally, use the Folium library to visualize the neighborhoods in Bangalore City and their emerging clusters. This project involves gathering location information of the IT offices spaces and using Foursquare APIs to get the location based details. Using the nearby venues we can divide the places in clusters to determine the best options.

Based on definition of our problem, factors that will influence our decision are:

- number of existing restaurants in the neighborhood (any type of restaurant)
- number of offices present in a particular neighborhood
- Preferable nearest places (Most Valued)

Following data sources will be needed to extract/generate the required information:

- The data is pulled from the site - here, which contains 300 IT company name along with their address. The data also contain PIN code for all the offices.
- Foursquare API are used to understand the data graphically and get the details of the nearby places

Figure: Dataframe created from the Web page

```
In [64]: bangalore_df.head()
```

Out[64]:

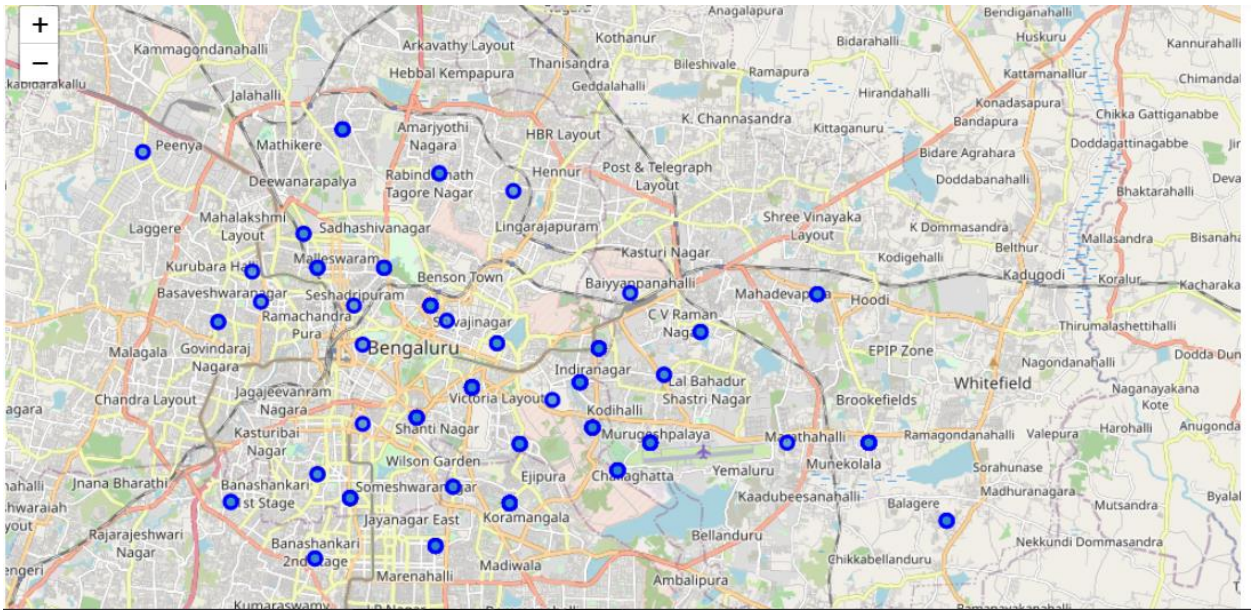
	Company_Name	Company_Address
1	24/7 Customer Pvt Ltd	Survey No 2/1, 2/2, 2/3, & 5/1, Challaghatta V...
2	247 Learning Solutions Pvt Ltd	No 20, Annaswamy Mudaliar Road, Ulsoor Lake, \...
3	Accenture Services Pvt Ltd	71 Cunningham Road\rBangalore - 560 052
4	Accord Software & Systems Pvt Ltd	# 37, K.R. Colony, Domlur Layout, \r\nBangalor...
5	Acme Insurance Services Pvt Ltd	3rd Floor, Monarch Chambers, \r122, Infantry R...

Figure2: Foursquare API is used get the details of the nearby places respectively.

Out[17]:

	Company_Name	Company_Address	Comapany_Code	latitude	longitude
3	Accenture Services Pvt Ltd	71 Cunningham Road\rBangalore - 560 052	560052	12.9902	77.596
4	Accord Software & Systems Pvt Ltd	# 37, K.R. Colony, Domlur Layout, \r\nBangalor...	560071	12.9576	77.6404
5	Acme Insurance Services Pvt Ltd	3rd Floor, Monarch Chambers, \r122, Infantry R...	560001	-33.0381	137.576
6	Adaptec (India) Pvt Ltd	No:5 , First Floor, \r\nSalarpuria Infinity\r\...	560029	12.9262	77.5974
7	Adea International Pvt Ltd	No.319/1, Bommanahalli\r\nHosur Main Road\r\nB...	560068	12.9003	77.6198
8	Aditi Technologies Pvt Ltd	224/16 Ramana Maharishi Rd\r\nBangalore 560 080	560080	13.0001	77.5833
9	Affiliated Computer Services of India (P) Ltd ...	Level 2, Creator Block International Tech Park...	560066	12.9536	77.7158
10	Ajax.com Pvt Ltd	#1, 3rd Floor Maruthi Complex, \r\nAbove Food ...	560032	13.0253	77.5984
11	Akamai Technologies India Pvt Ltd	Salarpuria Ascent\r\n#77, Jyothi Nivas College...	560095	12.9375	77.6179
12	Altair Engineering India Pvt Ltd	Mercury 2B Block, 5th Floor,\r\nPrestige Tech ...	560078	12.9005	77.5704

Figure3 Folium is used to understand the data graphically.



3.) Methodology Section

3.1 Data preparation

The data is extracted from a website using the web scrapping using BeautifulSoup Python Library to extract the data. The data is converted to Pandas dataframe for further processing.

Get the data set from the URL : https://www.naukri2000.com/careers/it_bangalore.php and then using BeautifulSoup package

In [2]:
```
dataset = requests.get('https://www.naukri2000.com/careers/it_bangalore.php').text
```

In [3]:
```
from bs4 import BeautifulSoup # library to parse HTML and XML documents
```

In [4]:
```
# parse data from the html into a beautifulsoup object
soup = BeautifulSoup(dataset, 'html.parser')
```

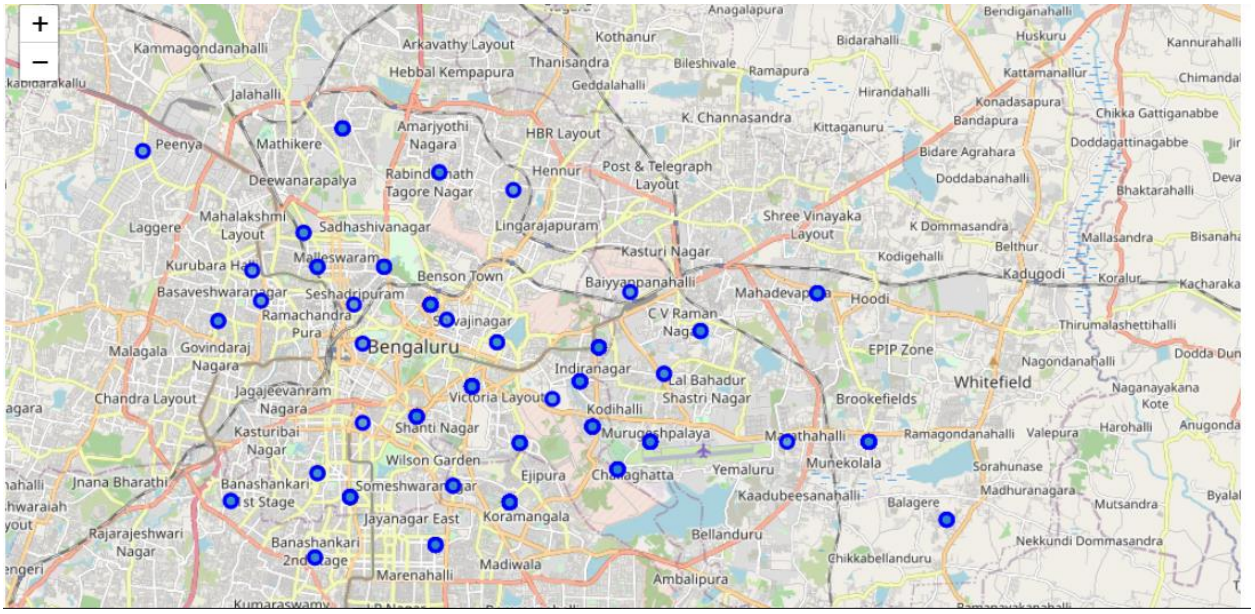
In [5]:
```
# create three lists to store table data
company_name = []
company_address = []
for row in soup.find_all('table'):
    cells = row.find_all('tr')
for companies in cells:
    company_name.append(companies.contents[2].text)
    company_address.append(companies.contents[4].text)
bangalore_df = pd.DataFrame({"Company_Name": company_name,
                             "Company_Address": company_address})
```

Then *Geopy* package is used to get the coordinates (latitude and longitude) for each and every location provided:

In [114]:
```
from geopy.geocoders import Nominatim
lat = []
long = []
for address in bangalore_df['Comapany_Code']:
    if address:
        geolocator = Nominatim()
        location = geolocator.geocode(address)
        if location:
            #print(location)
            print((location.latitude, location.longitude))
            lat.append(location.latitude)
            long.append(location.longitude)
        else:
            lat.append("")
            long.append("")
    else:
        lat.append("")
        long.append("")
```

3.2 Exploratory Data Analysis:

This is done using Folium Package to ensure the data we have picked have good enough spread throughout the city and we have picked up the data which covers all the area (majority) in Bangalore.



3.3 Picking up neighborhood data

Using the Foursquare API to get all the Venue details to further categorize it in different category.

```
In [138]: bangalore_venues.head(20)
```

Out[138]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Westside	12.982570	77.610713	Shopping Mall
1	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Mysore Saree Udyog	12.981433	77.610214	Women's Store
2	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Chaipatty ulsoor	12.976061	77.615338	Tea Room
3	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Vashi's House of Jeans	12.981449	77.610308	Clothing Store
4	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Bobby's Punjabi Dhaba	12.983826	77.613955	Indian Restaurant
5	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Cafe Coffee Day Ulsoor Lake	12.979498	77.618151	Café
6	247 Learning Solutions Pvt Ltd	12.980207	77.614153	Sreeraj Lassi Bar	12.982748	77.610739	Juice Bar

3.4 Onehot encoding

It is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. The screen shot of the data:

Out[145]:

	Neighborhood	ATM	American Restaurant	Andhra Restaurant	Arcade	Art Gallery	Arts & Crafts Store	Asian Restaurant	Auto Workshop	BBQ Joint	Badminton Court	Bakery	B
0	247 Learning Solutions Pvt Ltd	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
1	ANZ Information Technology Pvt. Ltd.	0.000000	0.000000	0.000000	0.041667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
2	ASM Technologies Ltd	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
3	Accenture Services Pvt Ltd	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.025641	0.025641	0.0000
4	Accord Software & Systems Pvt Ltd	0.000000	0.000000	0.000000	0.041667	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000
5	Acme Insurance Services Pvt Ltd	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0000

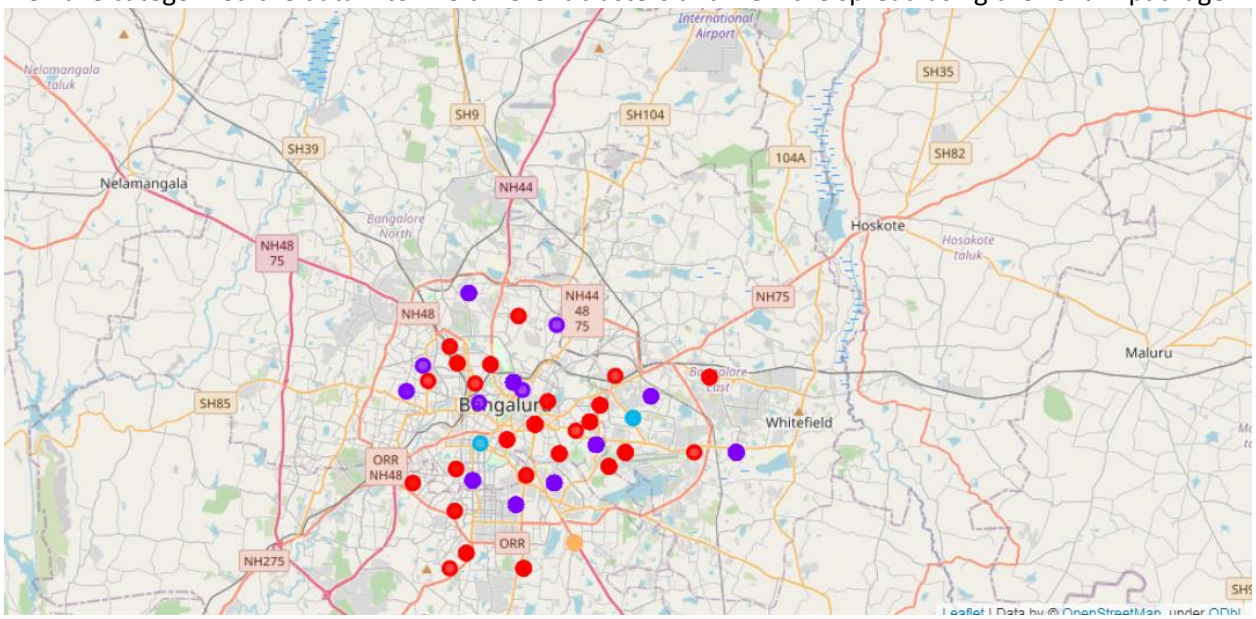
3.5 ML Algorithm

For the `K-means` Clustering Algorithm, all unique items under `Venue Category` are one-hot encoded and top 10 most common venues are figured out.

	Neighborhood	Company_Address	Comapany_Code	lattitude	longitude	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	247 Learning Solutions Pvt Ltd	No 20, Annaswamy Mudaliar Road, Ulsoor Lake, \...	560042	12.9802	77.6142	0	Clothing Store	Women's Store	Tea Room	Multiplex	Candy Store	Café
1	Accenture Services Pvt Ltd	71 Cunningham Road\rBangalore - 560 052	560052	12.9902	77.596	1	Indian Restaurant	Coffee Shop	Café	Italian Restaurant	Pizza Place	Tea Room
2	Accord Software & Systems Pvt Ltd	# 37, K.R. Colony, Domlur Layout, \r\nBangalor...	560071	12.9576	77.6404	1	Indian Restaurant	Hotel	Pizza Place	Irish Pub	Pub	Breakfast Spot
3	Acme Insurance Services Pvt Ltd	3rd Floor, Monarch Chambers, \r122, Infantry R...	560001	-33.0381	137.576	3	Bowling Alley	Grocery Store	Beach	Café	Discount Store	Department Store
4	Adaptec (India) Pvt Ltd	No:5 , First Floor, \r\nSalarpuria Infinity\r\...	560029	12.9262	77.5974	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Middle Eastern Restaurant	Women's Store	Discount Store

4.) Results

We have categorized the data into five different clusters and view the spread using the *Folium* package.



Cluster 1: We can see that it already have many options for food and catering. So that will not be profitable to open up the spaces there.

```
In [220]: bangalore_merged.loc[bangalore_merged['ClusterLabels'] == 0, bangalore_merged.columns[[1] + list(range(5, bangalore_merged.shape
```

Out[220]:

	Company_Address	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Com V
0	No 20, Annaswamy Mudaliar Road, Ulsoor Lake, \...	0	Clothing Store	Women's Store	Tea Room	Multiplex	Candy Store	Café	Juice Bar	South Indian Restaurant	Spo G
6	224/16 Ramana Maharishi Rd\r\nBangalore 560 080	0	Hotel	Ice Cream Shop	Indian Restaurant	Dance Studio	Café	Discount Store	Fast Food Restaurant	Electronics Store	Ea Euro Resta
8	#1, 3rd Floor Maruthi Complex, \r\nAbove Food ...	0	Park	Resort	Bus Station	Fast Food Restaurant	Cosmetics Shop	Cupcake Shop	Dance Studio	Department Store	D S

Cluster 2: We can see that it already have many options for food and catering. So that will not be profitable to open up the spaces there.

```
In [221]: bangalore_merged.loc[bangalore_merged['ClusterLabels'] == 1, bangalore_merged.columns[[1] + list(range(5, bangalore_merged.shape
```

30	Unit 2(B), Creator Building\r\nInternational T...	1	Indian Restaurant	Fast Food Restaurant	Café	Grocery Store	Ice Cream Shop	Shopping Mall	Chinese Restaurant	BBQ Joint	Pizza Place
36	No. 2/21, Shantha Complex, 1st Main, \r1st Cro...	1	Indian Restaurant	Park	Snack Place	Café	Movie Theater	Department Store	Dance Studio	Chinese Restaurant	Arts & Crafts Store
37	No. 111, 7 th Cross, \rRajmahal Vilas Extensio...	1	Indian Restaurant	Park	Snack Place	Café	Movie Theater	Department Store	Dance Studio	Chinese Restaurant	Arts & Crafts Store
38	Prestige Blue Chip, Block II\r\n9, Hosur Road,...	1	Indian Restaurant	Fast Food Restaurant	Asian Restaurant	Middle Eastern Restaurant	Women's Store	Discount Store	Electronics Store	Eastern European Restaurant	Dumpling Restauran

Cluster 3: We can see that it already have many options for food and catering. So that will not be profitable to open up the spaces there.

```
In [222]: bangalore_merged.loc[bangalore_merged['ClusterLabels'] == 2, bangalore_merged.columns[[1] + list(range(5, bangalore_merged.shape
```

Out[222]:

	Company_Address	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
169	R.K. Jalan / Saibal Mitra\r\nLakshmi Complex, ...	2	Indian Restaurant	Historic Site	Women's Store	Diner	Fast Food Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Discount Store
171	George Thangiah Complex (E), 2nd Floor\r\n80 F...	2	Indian Restaurant	Women's Store	Diner	Fast Food Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Discount Store	Dessert Shop
245	No. 7, Jeevan Bima Nagar Main Road, HAL 3rd St...	2	Indian Restaurant	Women's Store	Diner	Fast Food Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Discount Store	Dessert Shop

Cluster 4: This can be the target cluster as it doesn't have many restaurants and so this should be considered as our one of the options.

```
In [223]: bangalore_merged.loc[bangalore_merged['ClusterLabels'] == 3, bangalore_merged.columns[[1] + list(range(5, bangalore_merged.shape
```

Out[223]:

	Company_Address	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
3	3rd Floor, Monarch Chambers, \r122, Infantry R...	3	Bowling Alley	Grocery Store	Beach	Café	Discount Store	Department Store	Dance Studio	Design Studio	Dessert Shop	Diner
22	#30, Prestige Meridian II, \rUnit 701 & 702, \...	3	Bowling Alley	Grocery Store	Beach	Café	Discount Store	Department Store	Dance Studio	Design Studio	Dessert Shop	Diner
33	(A subsidiary of Canara Bank)\r\n7th Floor, Na...	3	Bowling Alley	Grocery Store	Beach	Café	Discount Store	Department Store	Dance Studio	Design Studio	Dessert Shop	Diner

Cluster 5: This can be the target cluster as it doesn't have many restaurants and so this should be considered as our one of the options.

```
In [225]: bangalore_merged.loc[bangalore_merged['ClusterLabels'] == 4, bangalore_merged.columns[[1] + list(range(5, bangalore_merged.shape
```

Out[225]:

	Company_Address	ClusterLabels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
5	No.319/1, Bommanahalli\r\nHosur Main Road\r\nB...	4	Bus Station	Indian Restaurant	Furniture / Home Store	Tea Room	Women's Store	Diner	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop
27	20-21, Pattalamma Layout, Singasandra, \rHosur...	4	Bus Station	Indian Restaurant	Furniture / Home Store	Tea Room	Women's Store	Diner	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop
53	"Shanthi Shree", \r\n#17/1, Hosur Road, \r\nBa...	4	Bus Station	Indian Restaurant	Furniture / Home Store	Tea Room	Women's Store	Diner	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop

5.) Discussion

So we see that how we can apply the ML Algorithm K-Means clustering algorithm to help in finding the solution and analyze the result. This would not have been possible with the only available data. The clusters help anyone see the data and the common venues all the cluster have near them. Whether the cluster posses any possibility to open up a restaurant business where there are not enough options or no options. Of course this is not the only factor to be considered before taking the final call, but this is one of the method by which one can refer for the location wise analysis. As we can clearly see that Cluster 4 and Cluster 5 offer a great opportunity for restaurants to be opened as these places are less crowded. Also, Cluster 1, and 2 are not so great options for a new business to be step up. Cluster 3 also have a handful of option and but can be kept in the consideration as of any cuisine which is not currently available there.

6.) Conclusion

In the above study, we explored the options of the opening of new restaurants based on the current situation of the IT offices and restaurants in Bangalore. We used a trusted website which contains all the data related to top MNC in Bangalore and using the various useful Python Packages and Machine Learning Clustering Algorithm to analyze the nearby venue data without much hassle.

This algorithm is based on limited number of data .i.e. limited number of IT companies in Bangalore, but it is shows the valuable aspect of the Model and usage and can be extended to include more companies name. Also, this can be useful for exploring the options in other cities too.