IBM Applied Data Science Capstone

'Exploring the Food Culture and Diversity of Gurgaon, India'

By: Karan Sud

May 2020

Introduction

Gurgaon, officially named Gurugram, is a city located in the Northern Indian state of Haryana with an estimated population of about 10 million as of 2020. It is situated about 30 kilometres southwest of the national capital New Delhi and is one of the leading Metropolitan cities of India, also sometimes referred to as the 'Job capital of North India'. Many people shift to this city for better job opportunities and hence, it becomes necessary to explore the Food culture and diversity of this city.

Business Problem

Undoubtedly, Food Diversity is an important part of an ethnically diverse metropolis. The idea of this project is to categorically segment the neighbourhoods of Gurgaon city into major clusters and examine their cuisines. A desirable intention is to examine the neighbourhood cluster's food habits and taste. This project will help to understand the diversity of a neighbourhood by leveraging venue data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm. Exploratory Data Analysis (EDA) will help to discover further about the culture and diversity of the neighbourhood.

Target Audience

This quantifiable analysis can be used to understand the distribution of different food cultures and cuisines over 'Job capital of North India'— Gurgaon, which can help any individual migrating to the city for work or education. Also, it can be utilized by a new food vendor who is willing to open his or her restaurant or by investors looking to invest in the food and beverage industry in the city.

Data Collection

To solve the problem, we will need the following data:

• List of neighbourhoods in Gurgaon: This defines the scope of this project which is confined to the city of Gurgaon, India.

• Latitude and longitude coordinates of those neighbourhoods: This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to restaurants and cuisines: We will use this data to perform clustering on the neighbourhoods. To begin with, we will take a look at Gurgaon on the Map using the folium library.

We will also fetch the data from two different APIs.

- Foursquare API: We will use the Foursquare API to fetch venues in Gurgaon starting from the middle up to 10 Kilometres in each direction.
- Zomato API: The Zomato API provides information about various venues including the food cuisines, user and restaurant ratings, price range and a lot more.

Data Source and Extraction

The page (https://www.mapsofindia.com/pincode/india/haryana/gurgaon/) contains a list of neighbourhoods in Gurgaon, with a total of 166. We will use web scraping techniques to extract the data from the page, with the help of Python requests and beautifulsoup package. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

After that, we will use Foursquare API to get the venue data for those neighbourhoods in order to help us to solve the business problem put forward and the Zomato API to get data on food cuisines and restaurants. Both these APIs are free and easy to use with a limitation on the number of calls. This is a project that will make use of many data science skills, from web scraping, working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Methodology

Exploring Data Set

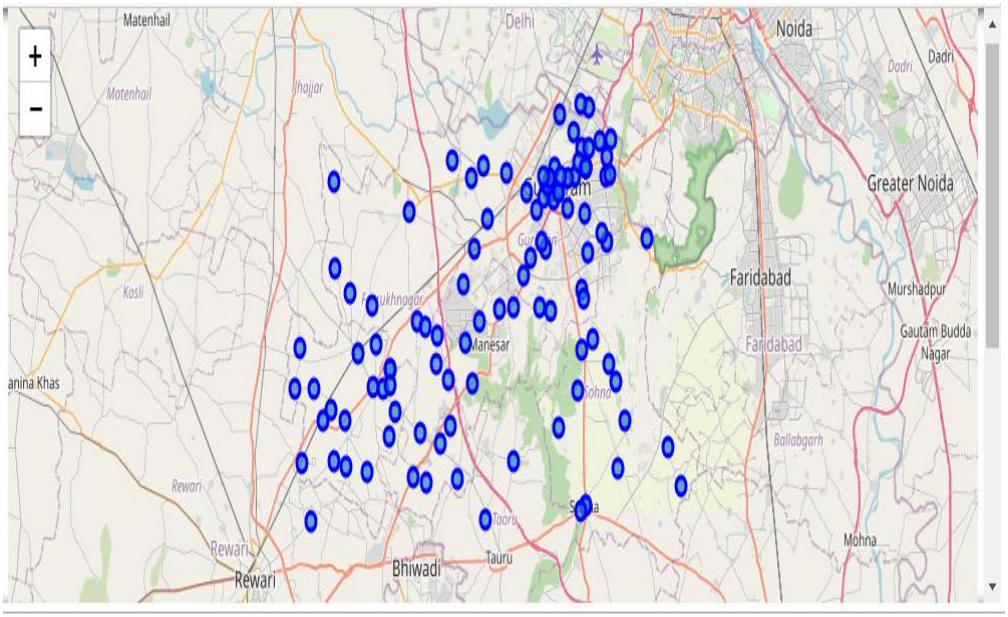
Firstly, we get the list of neighbourhoods in the city of Gurgaon which is available in the following page:

https://www.mapsofindia.com/pincode/india/haryana/gurgaon/)

We will do web scraping using Python requests and beautifulsoup packages to extract the list of neighbourhoods data. However, this is just a list of names and pin codes. We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame .

	PostalCode	Neighbourhood	Latitude	Longitude
0	122104	Agon	28.46219	77.02373
1	122105	AirForce	28.47762	77.06952
2	122107	Akhera	28.47762	77.06952
3	122001	ArjunNagar	28.45965	77.02004
4	121104	Baded	28.47762	77.06952

We then visualize the neighbourhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Gurgaon.

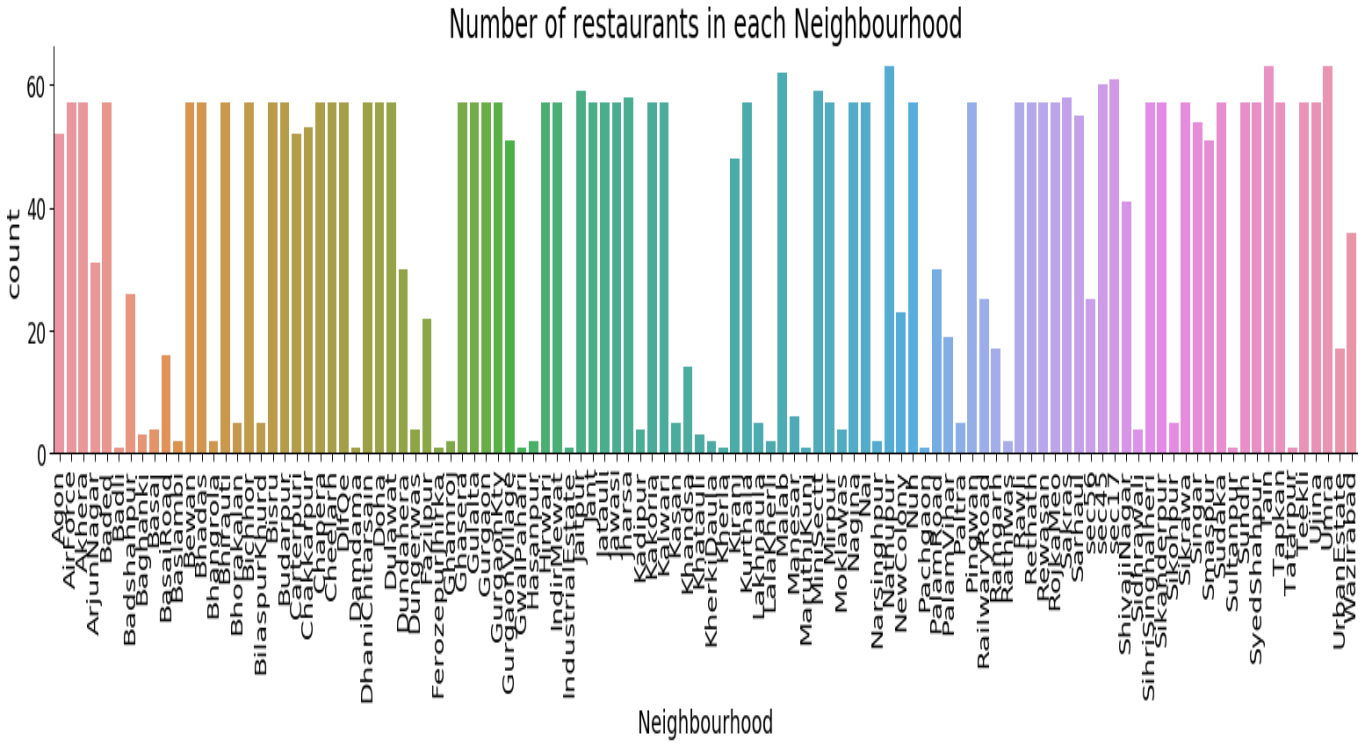


Next, we will use Foursquare API to get the top 100 venues that are within a radius of 4000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude.

	Neighbourhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Agon	28.46219	77.02373	Cocktails and Dreams Speakeasy	28.457921	77.044597	Bar
1	Agon	28.46219	77.02373	Bikanerwala	28.469620	77.036283	Indian Restaurant
2	Agon	28.46219	77.02373	The Monk	28.463142	77.049037	Chinese Restaurant
3	Agon	28.46219	77.02373	INOX	28.472378	77.017783	Multiplex
4	Agon	28.46219	77.02373	The World Spa	28.460302	77.058074	Residential Building (Apartment / Condo)

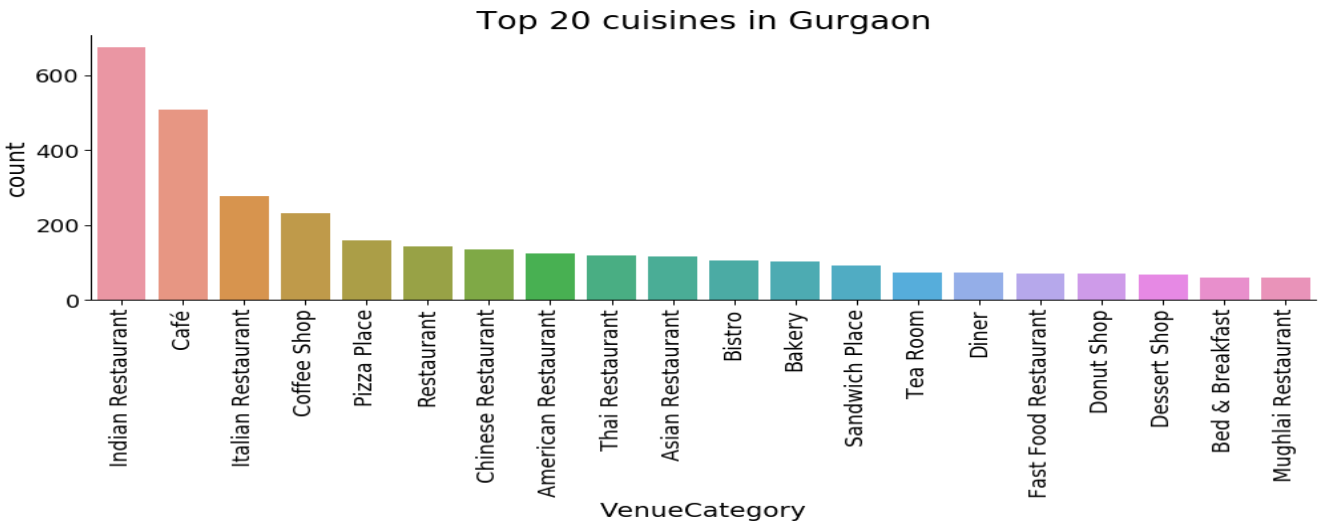
Exploratory Analysis

1. *Number of restaurants in each neighbourhood*:



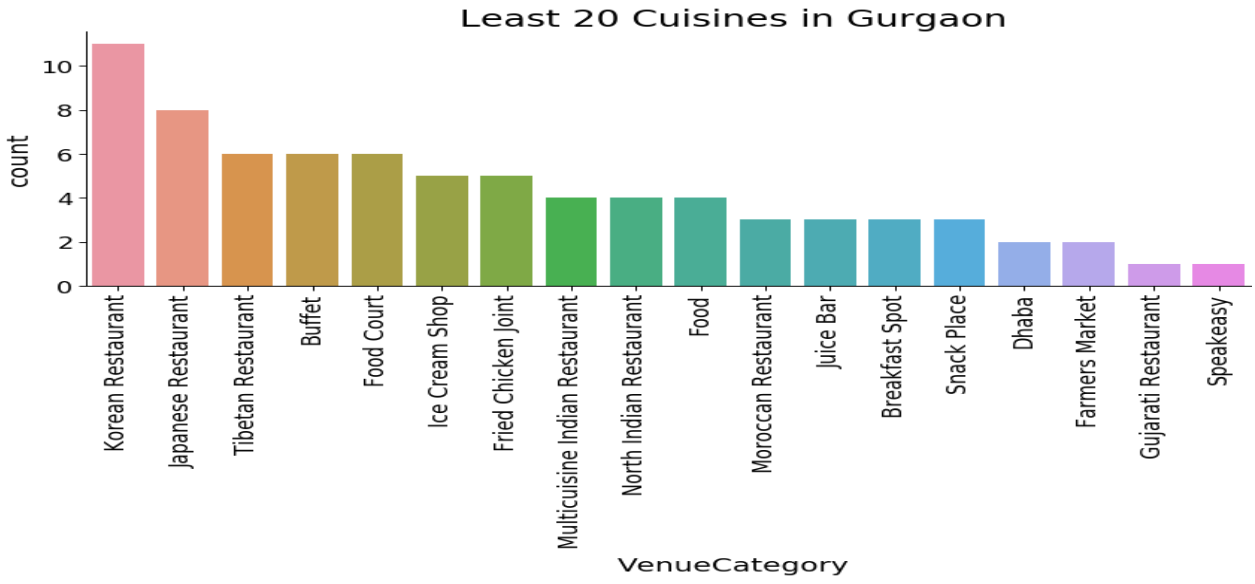
We see that there are varying trends for the number of restaurants in each neighbourhood. Neighbourhoods located in central Gurgaon have more venues compared to the outskirts which make sense as majority of the people are centrally located, including offices.

2. *Count plot of top 20 cuisines in Gurgaon*



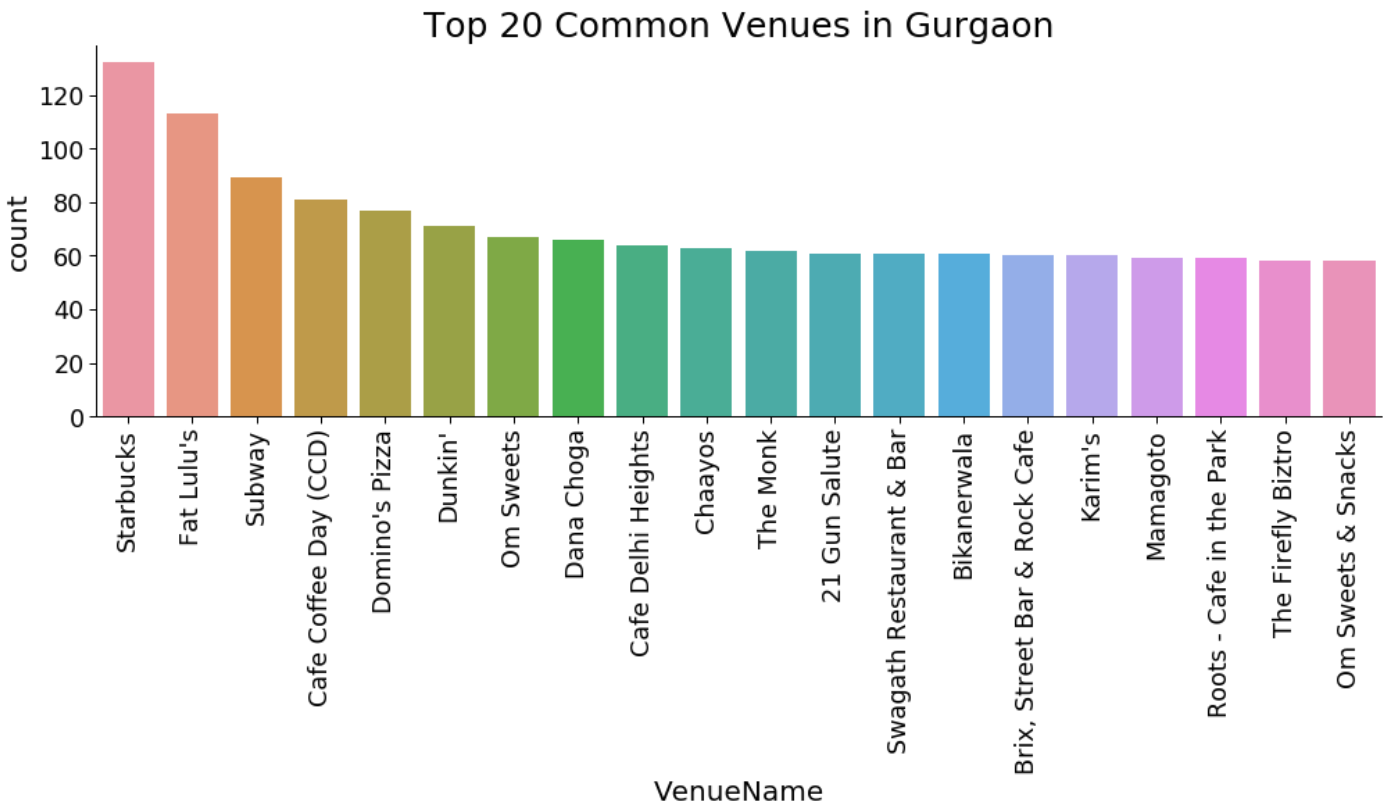
We see that there are a high number of Indian Restaurants in Gurgaon. This makes sense as people here prefer a North Indian cuisine.

3. *Count plot of Least 20 cuisines in Gurgaon*



These are venue categories that are not very famous in Gurgaon yet and will be hard to find.

4. Top 20 Common Venues



Starbucks is the most common venue in Gurgaon and you definitely won't have trouble getting a cup of coffee!

Note: Even though Indian cuisine is the most favourite, an Indian cuisine venue is not the most common. Maybe opening some more franchises is a good idea?

Feature Engineering

Now, each neighbourhood is analysed individually to understand the most common cuisine being served within its 4 kms of the vicinity.

The above process is taken forth by using 'one hot encoding' function of python 'pandas' library. One hot encoding converts the categorical variables (which are 'VenueCategory') into a form that could be provided to ML algorithms to do a better job in prediction.

Upon converting the categorical variables, 'Neighbourhood' column is added back. The size of the new dataframe 'ggn_onehot' is examined and it is found that there are 4,047 data points altogether.

The top 10 'Venue Categories' can also be found by counting their occurrences. This analysis shows that 'Indian Restaurant', 'Cafe, 'Pizza Place', 'Coffee shop', and 'Italian Restaurant' are among the top 5.

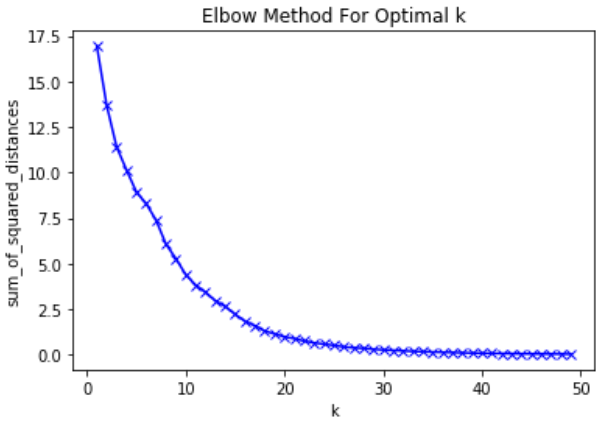
Machine Learning

'K-means' is an unsupervised machine learning algorithm which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighbourhoods for each cluster label for variable cluster size.

To implement this algorithm, it is very important to determine the optimal number of clusters (i.e. k). There are 2 most popular methods for the same, namely 'The Elbow Method' and 'The Silhouette Method'.

1. **The Elbow Method**

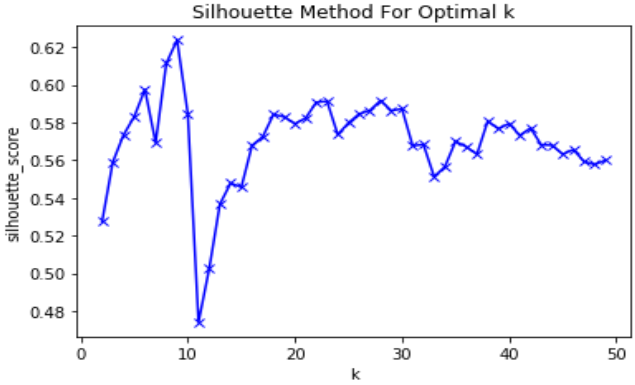
The Elbow Method calculates the sum of squared distances of samples to their closest cluster centre for different values of 'k'. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances.



Sometimes, Elbow method does not give the required result, which happened in this case. As there is a gradual decrease in the sum of squared distances, the optimal number of clusters cannot be determined. To counter this, another method can be implemented, as discussed below.

2. **The Silhouette Method**

The Silhouette Method measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). There is a peak at k = 8. Therefore, the number of clusters (i.e. 'k') is chosen to be 8.

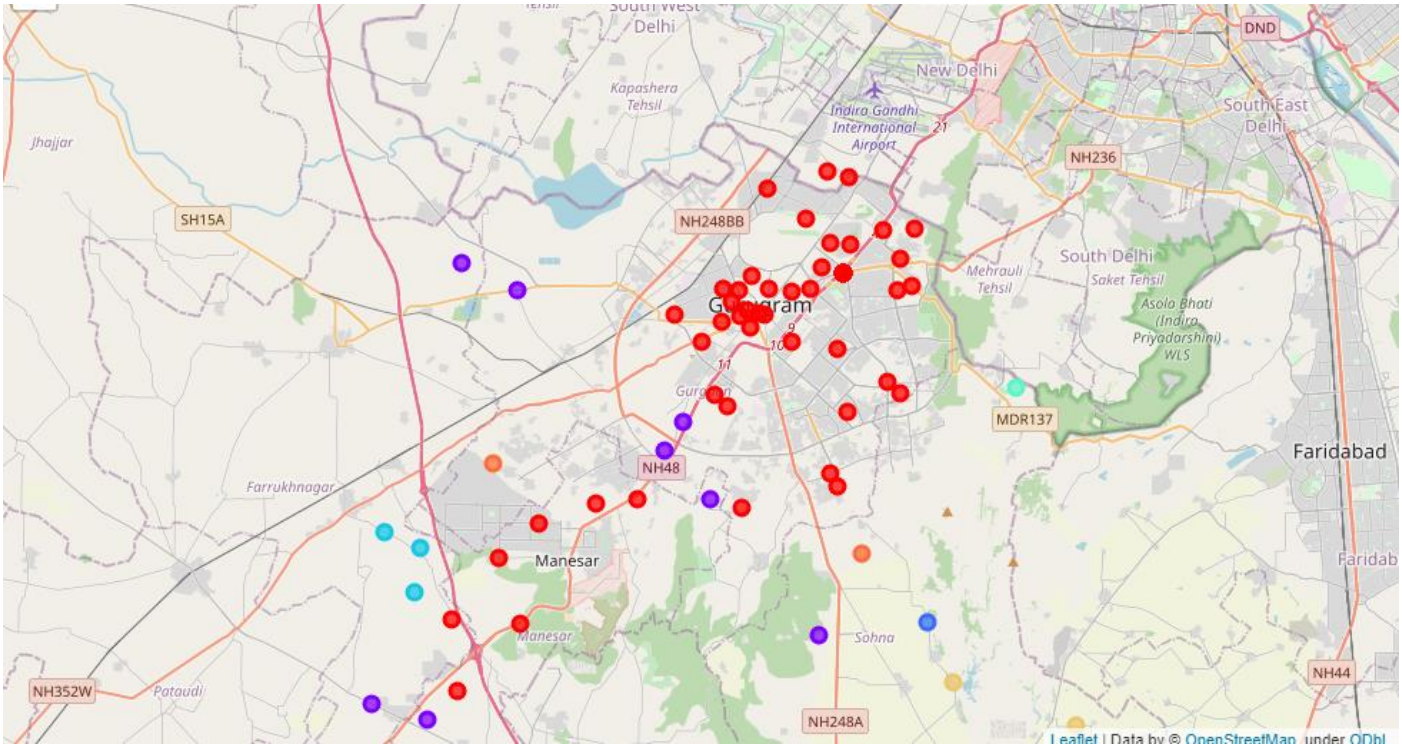


K-means Clustering Model

We will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the neighbourhoods into 8 clusters based on the frequency of occurrence for Food venues in the neighbourhood. The results will allow us to identify the food culture in Gurgaon and it's diversity in different neighbourhoods.

The clusters are visualized utilizing the python 'folium' library. Following map is generated which shows the desired segmentation of the Gurgaon's food culture and diversity:



Results

- **Cluster 0:**

```
Indian Restaurant       73
Café                     4
Fried Chicken Joint      3
Fast Food Restaurant     3
Food Truck               2
Pizza Place              2
American Restaurant      1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Café                    65
Indian Restaurant        6
Pizza Place              5
North Indian Restaurant  3
Fast Food Restaurant     3
Sandwich Place           2
Bakery                   1
Italian Restaurant       1
Japanese Restaurant      1
Coffee Shop              1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Italian Restaurant      50
Café                     7
Pizza Place              7
Coffee Shop              6
Indian Restaurant        4
Fast Food Restaurant     4
Bakery                   3
Sandwich Place           2
American Restaurant      2
Buffet                   2
Japanese Restaurant      1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------------
```

This is the biggest cluster of all, which means that majority of the neighbourhoods are clustered in it. Most neighbourhoods in this cluster have 'Indian Restaurant' as the 1st most common venue, 'Café' as the 2nd most common venue and 'Italian Restaurant' as the 3rd most common venue. This cluster includes the hub of Gurgaon where most people move to as it is closer to their workplace and hence, the most amount of restaurants have been opened in this area.

- **Cluster 1:**

```
Indian Restaurant    11
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Vegetarian / Vegan Restaurant    4
Buffet                           3
Breakfast Spot                   1
Restaurant                       1
Dhaba                            1
Fried Chicken Joint              1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Vegetarian / Vegan Restaurant    6
French Restaurant                3
Restaurant                       1
Dhaba                            1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------------
```

This cluster has Indian Restaurants as the 1st most common venue in all neighbourhoods followed by Vegetarian/Vegan restaurants as the 2nd and 3rd most common venues. This cluster lies on the outskirts of Gurgaon.

- **Cluster 2:**

```
Breakfast Spot    1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------
Vegetarian / Vegan Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------
Chinese Restaurant    1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------
```

This cluster contains only a single Neighbourhood ‘Kherla’ which is a village area. This neighbourhood can also be removed to improve the model.

- **Cluster 3:**

```
Snack Place    3
Name: 1st Most Common Venue, dtype: int64
----------------------------------------
Vegetarian / Vegan Restaurant    2
Japanese Restaurant              1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------
Chinese Restaurant               2
Vegetarian / Vegan Restaurant    1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------
```

This cluster is near the highway and hence the 1st most common venue is Snack place where people like to stop by while travelling.

- **Cluster 4:**

```
Café    2
Name: 1st Most Common Venue, dtype: int64
----------------------------------------
Chinese Restaurant    1
Restaurant            1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------
Chinese Restaurant    1
Food Court            1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------
```

This cluster has café as the 1st most common area. This makes sense as there are factories in this cluster and the workers usually prefer cafes around.

- **Cluster 5:**

```
Restaurant    1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Vegetarian / Vegan Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Chinese Restaurant    1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------------
```

This cluster also contains only a single neighbourhood ‘Industrial Estate’. As the name suggests it is an industrial area and opening a restaurant here will not make much sense from a business point of view.

- **Cluster 6:**

```
Chinese Restaurant    2
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Vegetarian / Vegan Restaurant    1
Gujarati Restaurant              1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Food Court           1
French Restaurant    1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------------
```

This cluster has Chinese restaurant as the 1st most common venue. This cluster is around cluster 2. If we reduced the number of clusters these might be grouped together.

- **Cluster 7:**

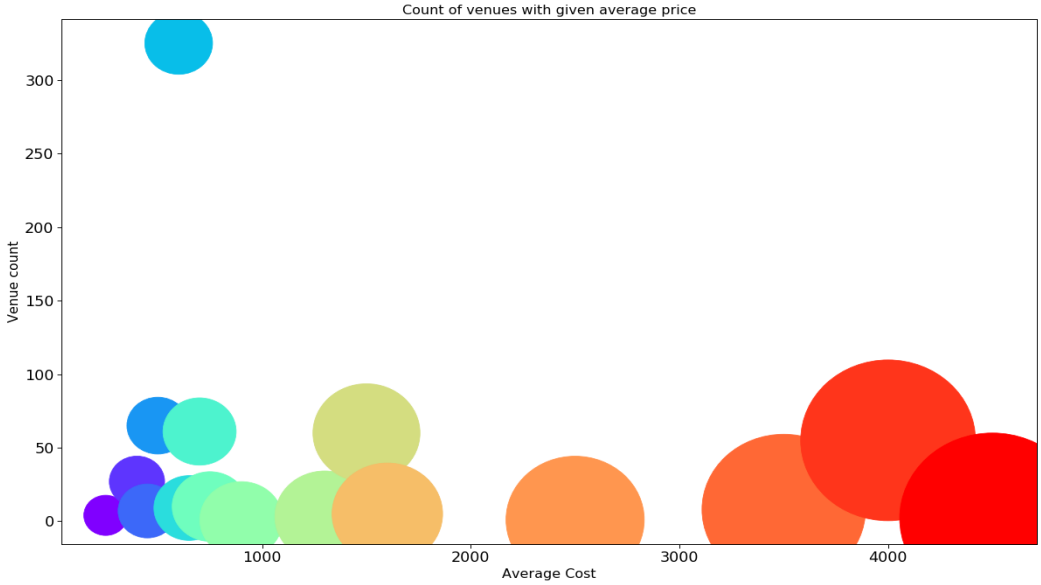
```
Pizza Place           1
Japanese Restaurant   1
Name: 1st Most Common Venue, dtype: int64
----------------------------------------------
Pizza Place                      1
Vegetarian / Vegan Restaurant    1
Name: 2nd Most Common Venue, dtype: int64
----------------------------------------------
Chinese Restaurant               1
Vegetarian / Vegan Restaurant    1
Name: 3rd Most Common Venue, dtype: int64
----------------------------------------------
```

This cluster has Pizza Place as the 1st most common venue. It is very close to the central Gurgaon hub (Cluster 0).

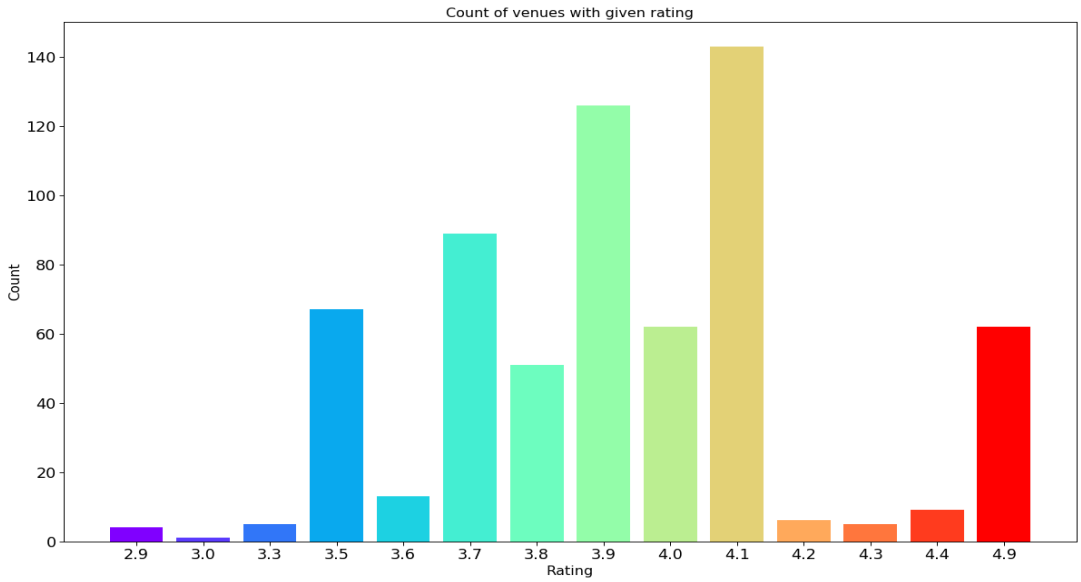
Use of Zomato API on Cluster 0

Zomato API will be used to get the Average Price and Ratings for Indian Restaurants in the neighbourhoods of this cluster. We can use the API to get data on other Venue Categories and other clusters as well, but due to the limitation of calls we can make, we will stick to just Indian restaurants in this cluster.

1) **Average Price:** We explore the average prices using a scatter plot along with the count of venues with that average price per person. The plot reveals that the majority venues have an average cost below 1000. Even though the maximum venues lie in that range, the actual range of prices is different. There are places with average price even as high as Rs 4000 for two people.



2) **Ratings:** We explore the ratings of Indian Restaurants in this cluster. We plot a bar chart with x-axis as the rating from 1 to 5 and the y-axis as the count of venues with that rating. From the plot we see that the average rating is spread across 3.7 and 4.1 with maximum number of venues with a rating of 4.1.



Discussion

As observations noted from the Results section, most of the restaurants are concentrated in the central area of Gurgaon, with the highest number in cluster 0 and moderate number in cluster 1. On the other hand, clusters 2 and 5 have a single neighbourhood very far from the central area. Cluster 7 is very close to the central area and has potential for future restaurants. From another perspective, the results also show that the oversupply of restaurants mostly happened in the central area of the city, with the suburb area still have very few restaurants. Therefore, this project recommends anyone looking to open new restaurants to capitalize on these findings to open new restaurants in neighbourhoods in cluster 1 and cluster 7 with little to no competition. Lastly, restaurant openers are advised to avoid neighbourhoods in cluster 2,4 and 5 which don't have enough potential for success of a restaurant.

Limitations and Suggestions for Future Research

In this project, we only consider one factor i.e. frequency of occurrence of restaurants, there are other factors such as population and income of residents that could influence the location decision of a new restaurant. However, to the best knowledge of this researcher such data are not available to the neighbourhood level required by this project. Future research could devise a methodology to estimate such data to be used in the clustering algorithm to determine the food culture and diversity of an area.

We may also choose to remove some neighbourhoods that act as outliers to improve the model and try playing around with different k values.

In addition, this project made use of the free Sandbox Tier Account of Foursquare API and Zomato API that came with limitations as to the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more results. For example: We can use the Zomato API to obtain average price and ratings for all Venue Categories instead of just Indian Restaurants!

Conclusion

We deduce some great results like which neighbourhoods have the most number of restaurants and hence highest number of foodies. We also saw what all cuisines are mostly preferred by people of Gurgaon. Our purpose was to determine the food culture and diversity of Gurgaon and we saw that ‘Indian Restaurant’, ‘Café’, ‘Pizza Place’, ‘Coffee shop’, and ‘Italian Restaurant’ are the top 5 cuisines in Gurgaon.

The ‘magic’ of data analysis is that it gives a new perspective towards a problem and can even help in discovering some amazing facts and figures just by looking at the data.

I hope this project helps anyone who looks into the Food culture and diversity of Gurgaon, not only from a data analysis point of view, but also as someone planning to move to the city.

Cheers!