“The Battle of Neighborhoods” — Restaurant Problem

Author – Shreya Biswas

Applying the unsupervised machine learning algorithm of K-Means clustering to the problem of identifying a suitable location for opening a new restaurant in the city of Gurugram, India.

This project was completed as part of the *IBM Data Science Professional Certificate* and applies the unsupervised machine learning algorithm of *K-Means clustering* to the problem of identifying a suitable location for opening a new restaurant in the city of Gurugram, India.

Background

The *National Capital Region (NCR)* of India comprises the prominent urban cities of *New Delhi* (also the capital city), **Gurugram** (formerly Gurgaon), and *Noida*, along with several other adjoining districts. Of these cities, Gurugram, owing to its large multicultural local population as well as a steady tourist footfall — both domestic and international (at least during the pre-Covid era), represents a metropolitan milieu that is particularly conducive to opening a new themed restaurant or café, especially from the perspective of a small to medium-sized enterprise, or even a large and established restaurant chain.

Business Understanding and Problem Statement

Having discussed the eminence of the city of Gurugram, India as a lucrative location for opening a new themed eatery, the goal of this project was to assist businesses interested in opening a new restaurant or café within Gurugram, in identifying appropriate neighborhoods (or localities) for their venture. Recommendations regarding localities were based on several important factors such as popular types of eateries across the localities, customer traffic in each locality, and competition from pre-existing food joints in each locality.

Hence, the two-part problem that this project aimed to solve was defined:

- **Which neighborhoods (or localities) of Gurugram are ideal for opening a new restaurant or café?**
- **B. If a new restaurant or café is opened in the neighborhood (or locality) of choice, which cuisine would have the greatest chance of success?**

Data Requirements

With some clarity regarding the questions to be answered via this project, it was evident that the data required would comprise a *listing of the major localities within the city of Gurugram*, along with their *geospatial location data (i.e., latitude and longitude coordinates).* A reasonably comprehensive dataset for obtaining the aforementioned geospatial data was the Zomato API dataset available on **Kaggle**. Thereafter, the geospatial location data of the localities were utilized to explore the different neighborhoods, i.e., analyze the most popular cuisines of eateries in different neighborhoods, retrieve data regarding the eateries in geographical proximity to a particular neighborhood, etc., using **Foursquare API**.

Data Processing for Exploratory Analysis and Modeling

The Kaggle dataset containing restaurant data from the Zomato API (*zomato.csv*) is an extensive dataset consisting of 21 columns storing the geospatial location data of the restaurants within the localities of various cities and countries, along with other data such as the restaurant cuisines, average cost for two, whether a particular restaurant has table booking, or online delivery, etc. However, for this project, only geospatial data pertaining to the major localities within Gurgaon (now known as Gurugram) is required. Hence, the *zomato.csv* data was processed to extract only the relevant columns, for exploratory analysis and modeling.

The resultant dataframe is shown below:

	City	Locality	Latitude	Longitude
0	Gurgaon	Sohna Road	28.424831	77.039310
1	Gurgaon	Ambience Mall	28.503077	77.097118
2	Gurgaon	Palam Vihar	28.511416	77.042009
3	Gurgaon	Ardee City	28.440709	77.087851
4	Gurgaon	Sector 15	28.458088	77.034715

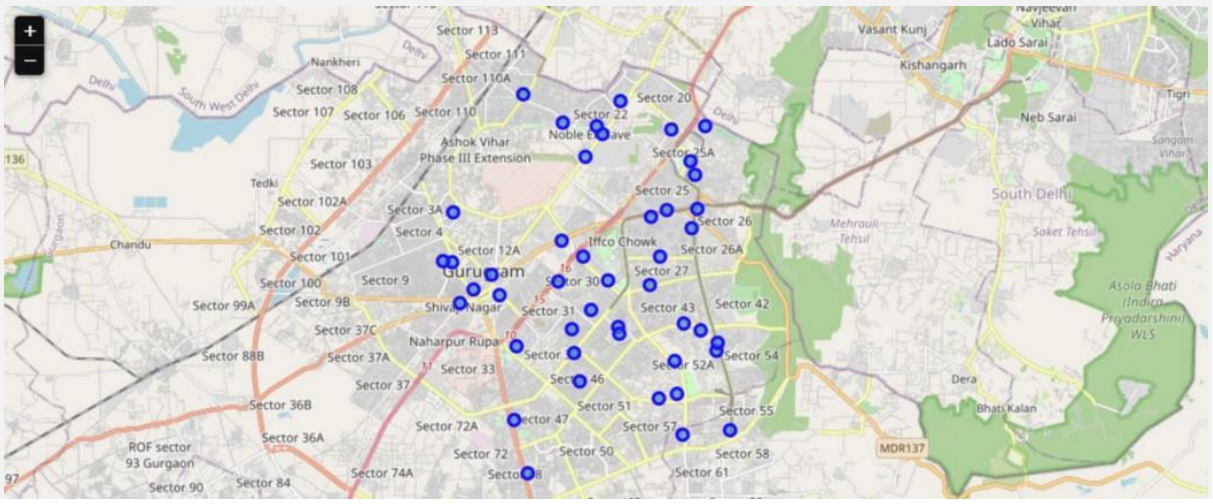
Zomato geospatial data for the localities in Gurgaon

Next, *missing values* and *outliers* (erroneous entries such as “0” for *Latitude* and/or *Longitude* values) in the zomato geospatial data were handled.

Methodology

A. Exploring the Different Localities in Gurugram

The *Geopy* library of Python was used to obtain the geospatial coordinates (latitude and longitude values) of Gurugram. Thereafter, all the unique localities within Gurugram were visualized on a *Folium* map.



Folium visualization of the unique localities in Gurugram

Following the Folium visualization, the *nearby venues* for each locality in Gurugram were obtained using **Foursquare API**. The limit for the maximum number of venues returned was set to *100* for each locality, and the radius to *500* meters. The data for the *nearby venues* was cast to a dataframe that also contained several venue categories irrelevant to this project, e.g. *Sporting Goods Shop*, *Movie Theater*, etc. Since this project is concerned with eateries only, different types of eateries, e.g. *Brewery*, *Fast Food Restaurant*, *Indian Restaurant* were filtered out from the dataframe to create a new dataframe pertaining to data for eateries only.

The resultant dataframe, filtered for eateries only is shown below, and was used for exploratory data analysis:

	Locality	Latitude	Longitude	Venue Name	Venue Latitude	Venue Longitude	Venue Category
0	Sohna Road	28.424831	77.03931	Brewer Street	28.424499	77.039134	Brewery
2	Sohna Road	28.424831	77.03931	KFC	28.424789	77.038722	Fast Food Restaurant
3	Sohna Road	28.424831	77.03931	Sagar Ratna	28.421042	77.038509	Indian Restaurant
5	Sohna Road	28.424831	77.03931	Torgauer Brewpub	28.423641	77.039426	Brewery
7	Sohna Road	28.424831	77.03931	Barbeque Nation	28.421888	77.040544	BBQ Joint

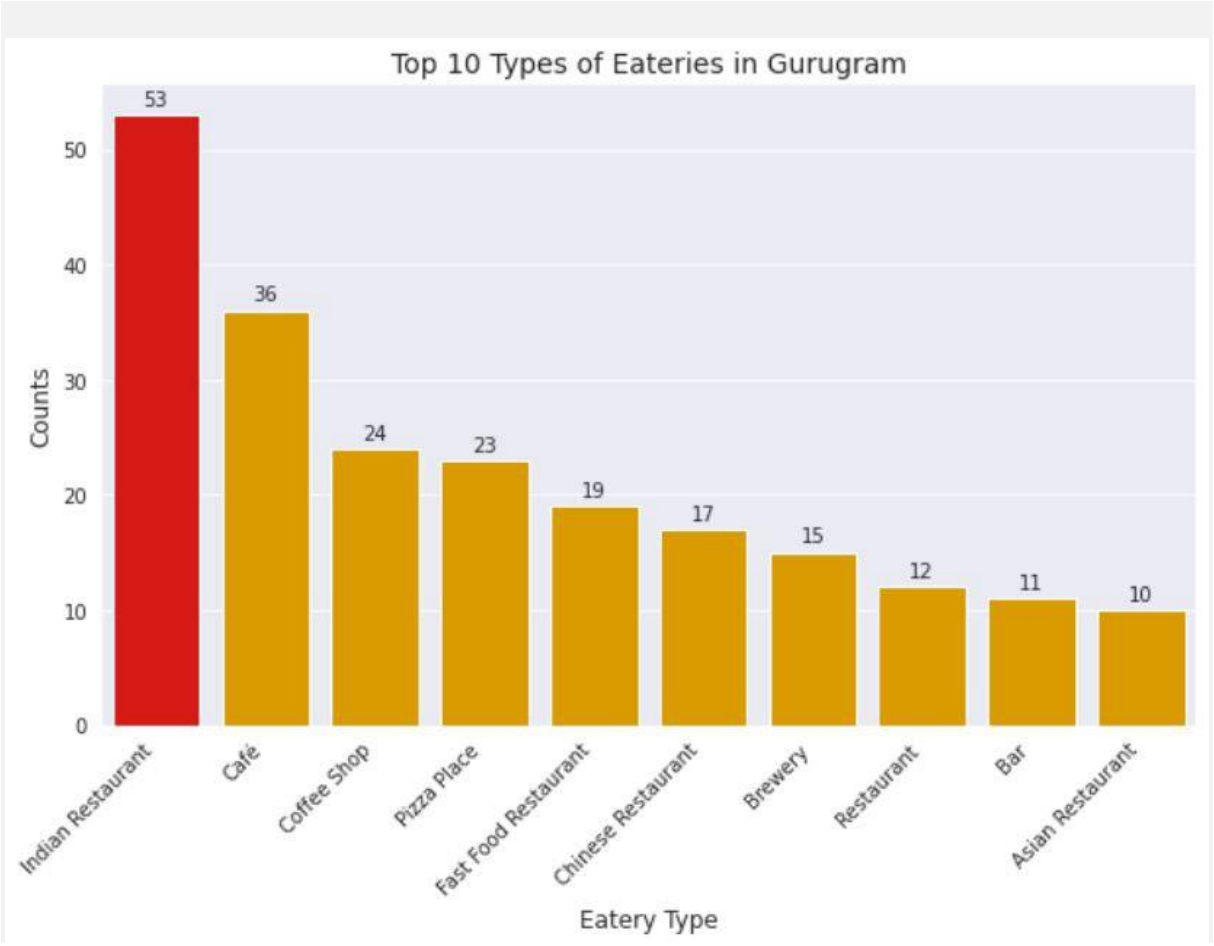
Foursquare API data for the venues near each Gurugram locality, filtered to contain only eateries as venues.

In order to identify the most common type of eatery across the localities of Gurugram, a new dataframe was created to display the different types of eateries and their respective counts.

	venue_category	counts
0	Indian Restaurant	53
1	Café	36
2	Coffee Shop	24
3	Pizza Place	23
4	Fast Food Restaurant	19
5	Chinese Restaurant	17
6	Brewery	15
7	Restaurant	12
8	Bar	11
9	Asian Restaurant	10

Top 10 types of eateries within the localities of Gurugram

The top 10 types of eateries in Gurugram (as per this dataframe) were then used to generate a bar chart using Python's *Seaborn* and *Matplotlib* libraries:



According to the bar chart, the most popular type of eatery in Gurugram was identified to be **Indian Restaurant**. Hence, opening a new Indian restaurant was decided upon as a good idea in general, ensuring a regular inflow of customers. Now, there was a need to identify locations within Gurugram with a steady customer footfall (or medium to high density of pre-existing eateries), wherein no Indian restaurants exist (minimizing competition from the same type of eatery) for opening the new Indian restaurant. This was done via further exploratory data analysis and *K-means clustering*.

First, the density of pre-existing eateries in the various localities of Gurugram was assessed.

The output for the above code snippet was as follows:

Locality	
DLF Cyber City	53
Sector 29	31
DLF Phase 4	28
Ambience Mall	26
MG Road	21
Sector 54	17
DLF Phase 5	15
Golf Course Road	14
DLF Phase 2	13
Sushant Lok	11
Woods Resort	11
DLF Phase 1	9
Omaxe Gurgaon Mall	7
Sector 21	7
Udyog Vihar	6
DLF Phase 3	6
Hyatt Place Gurgaon	5
Sector 23	5
Sohna Road	5
Sector 57	5
Sector 30	5
Sector 22	5
Ardee City	4
Country Inn & Suites by Carlson	4
Sikandarpur	4
Sector 56	3
Palam Vihar	3
Sector 7	3
Sector 39	3
Old Railway Road	3
Sector 12	2
Sadar Bazar	2
Sector 31	2
Sector 45	1
Sector 50	1
Sector 53	1
Sector 17	1
Sector 44	1
Sector 43	1
South City 2	1
Sector 15	1
Sector 14	1

Name: Venue Category, dtype: int64

Counts of pre-existing eateries in the various localities of Gurugram

From the output above, it was evident that *DLF Cyber City* and *Sector 29* are some of the most popular localities within Gurugram in terms of restaurant density, thus ensuring steady customer traffic. Other localities of Gurugram with a medium density of restaurants (and hence, still reasonably good customer traffic) are *DLF Phase 4*, *Ambience Mall*, and *MG Road*. However, the ideal location for opening a new *Indian restaurant* would be a **locality with a medium to high restaurant density but no pre-existing Indian restaurant**. Such a location was identified by creating a grouped dataframe from the *Gurgaon_eateries* dataframe via one-hot encoding and analyzing each locality within Gurugram via clustering and selecting localities within clusters wherein no Indian restaurants (or few) exist. But before clustering the Gurugram localities, the top 10 common types of eateries within each locality were identified.

The above code snippet produced the following output (truncated):

********** Ambience Mall **********

	Eatery type	Frequency
0	Indian Restaurant	0.23
1	Café	0.12
2	Fast Food Restaurant	0.12
3	Italian Restaurant	0.08
4	Asian Restaurant	0.08
5	American Restaurant	0.08
6	Bar	0.08
7	Diner	0.04
8	Food Court	0.04
9	Mediterranean Restaurant	0.04

********** Ardee City **********

	Eatery type	Frequency
0	Thai Restaurant	0.25
1	Indian Restaurant	0.25
2	Beer Garden	0.25
3	Chinese Restaurant	0.25
4	American Restaurant	0.00
5	New American Restaurant	0.00
6	Hookah Bar	0.00
7	Hotel Bar	0.00
8	Italian Restaurant	0.00
9	Japanese Restaurant	0.00

********** Country Inn & Suites by Carlson **********

	Eatery type	Frequency
0	Japanese Restaurant	0.50
1	Indian Restaurant	0.25
2	Café	0.25
3	American Restaurant	0.00
4	New American Restaurant	0.00
5	Gourmet Shop	0.00
6	Hookah Bar	0.00
7	Hotel Bar	0.00
8	Italian Restaurant	0.00
9	Kerala Restaurant	0.00

********** DLF Cyber City **********

	Eatery type	Frequency
0	Indian Restaurant	0.15
1	Coffee Shop	0.11
2	Café	0.08
3	Fast Food Restaurant	0.06
4	Asian Restaurant	0.06
5	Bar	0.06
6	Pizza Place	0.06
7	Donut Shop	0.04
8	Food Court	0.04
9	Mediterranean Restaurant	0.04

Top 10 common types of eateries within each Gurugram locality

B. Clustering the Different Localities in Gurugram

K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm used for partitioning *n* data points into *k* clusters, as per which each data point is assigned to the cluster with the nearest mean (cluster center or cluster centroid), serving as representative of the cluster. The optimized algorithm minimizes the intra-cluster *sum-of-squared-errors*. The squared error for each data

point is the square of the distance (Euclidean or Manhattan, for example) of the data point from its predicted cluster center/centroid (Source).

For this project, *K-means clustering* was used to cluster the localities in Gurugram according to the frequency of occurrence of different types of eateries within the city.

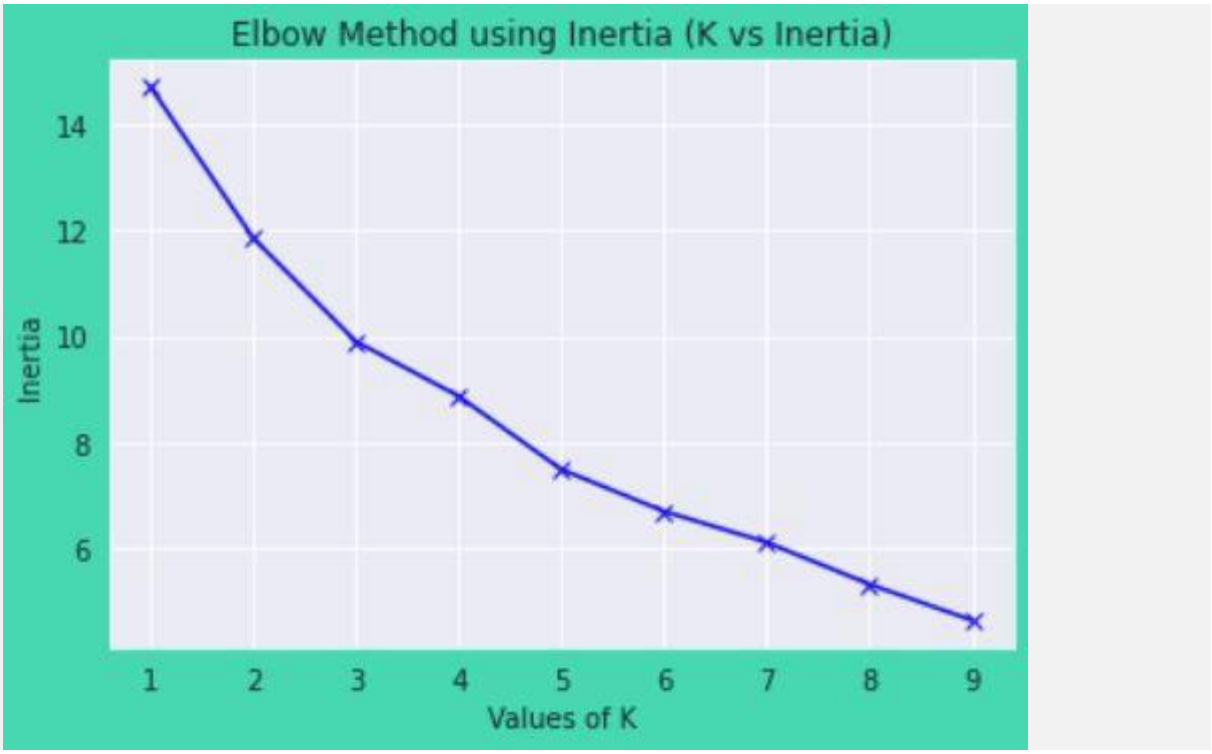
Selecting the Optimal Value of K for K-Means Clustering

Elbow *Method*
The Elbow method works by calculating the *Inertia*, or the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and selecting the value of k for which WSS first starts to stabilize. In the plot of *K vs Inertia*, this will be made evident by an "elbow" in the curve (Source).

Silhouette *Method*
The Silhouette Score measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette Score is between +1 and -1. A high value is desirable and indicates that the data point is assigned to the correct cluster. If many data points have a negative Silhouette Score, it may indicate that too many or too few clusters have been created. Generally, Euclidean distance is used as the distance metric. The Silhouette Score reaches its global maximum at the optimal k, appearing as a peak in the *K vs Silhouette Score* plot (Source).

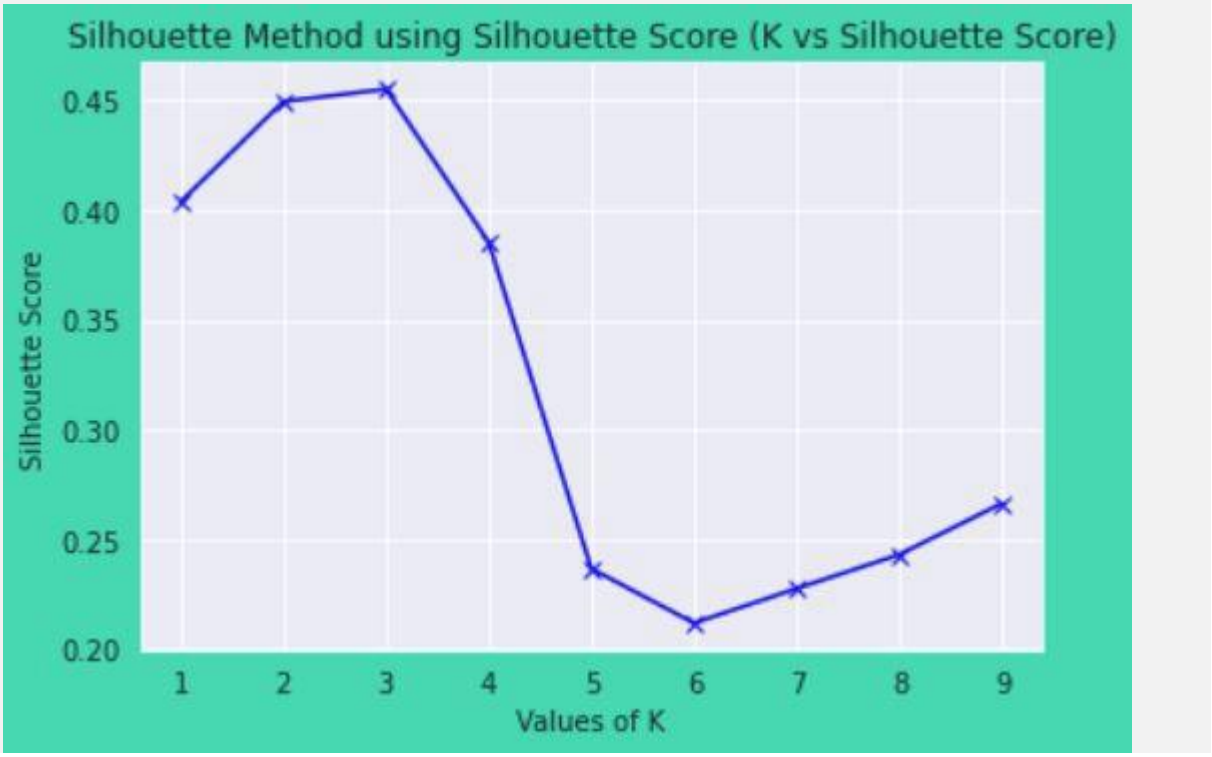
Both *Inertia* and *Silhouette Score* can be easily calculated using the *metrics* module of the *sklearn* library.

For this project, the *Elbow* method was used for selecting the optimal value of k for model building, whereas the *Silhouette* method was used for model evaluation and refinement.



Plot of K vs Inertia

From the plot of *k vs inertia*, it was evident that there was no clear "elbow". Hence, for the k-means clustering model built in this project, the *Elbow Method* was not a good metric for selecting the optimal value of k (or number of clusters). Hence, another metric was used for selecting the optimal value of k, namely, the *Silhouette Score*.



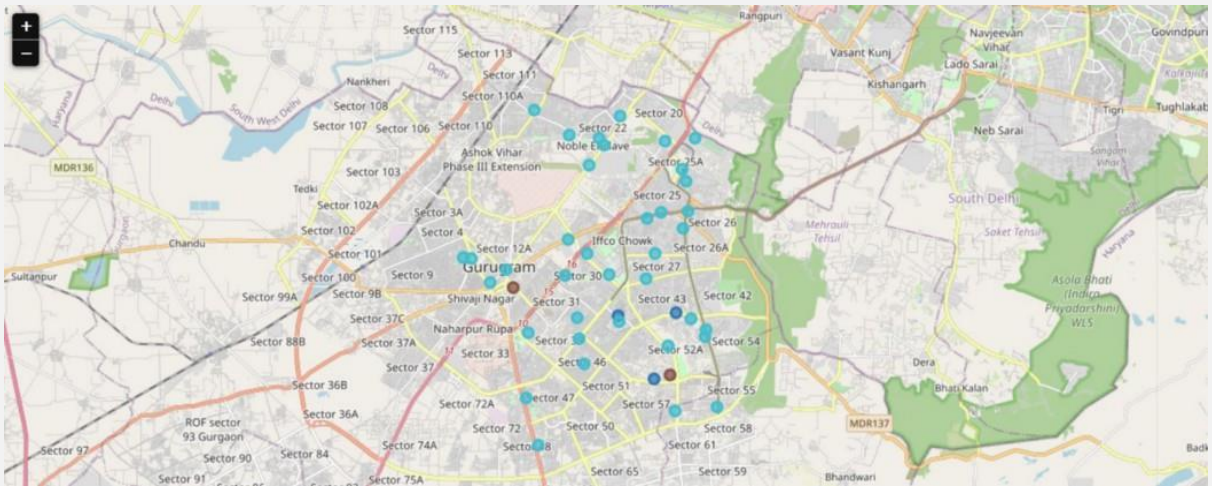
Plot of K vs Silhouette Score

From the plot of *k vs silhouette score*, it was clear that it is a much better metric for selecting the optimal value of k for the k-means clustering model built in this project, as there was a clear peak in the curve at **k = 3**. Therefore, the localities in Gurugram could potentially be optimally partitioned into *3 clusters*.

Next, the **K-Means clustering** algorithm was used to cluster the localities in Gurugram into *3 clusters*.

Results

After clustering the different localities in Gurugram using the *K-Means clustering* algorithm and generating the cluster labels, a final dataframe displaying the results of the clustering was created, by adding the cluster labels for each locality in Gurugram to the *Gurgaon_eateries_sorted* dataframe and merging with the dataframe containing the zomato geospatial data. Thereafter, the clustered localities in Gurugram were visualized on a map using Folium.



Folium visualization of the clustered localities in Gurugram

Finally, the distribution of the different localities in Gurugram by cluster labels, i.e., 0, 1, or 2, was displayed.

The above code snippet produced the output below:

Cluster Labels	Locality	
0	Ambience Mall	1
	Ardee City	1
	Country Inn & Suites by Carlson	1
	DLF Cyber City	1
	DLF Phase 1	1
	DLF Phase 2	1
	DLF Phase 3	1
	DLF Phase 4	1
	DLF Phase 5	1
	Golf Course Road	1
	Hyatt Place Gurgaon	1
	MG Road	1
	Old Railway Road	1
	Omaxe Gurgaon Mall	1
	Palam Vihar	1
	Sadar Bazar	1
	Sector 12	1
	Sector 14	1
	Sector 17	1
	Sector 21	1
	Sector 22	1
	Sector 23	1
	Sector 29	1
	Sector 30	1
	Sector 31	1
	Sector 39	1
	Sector 45	1
	Sector 50	1
	Sector 54	1
	Sector 56	1
	Sector 57	1
	Sector 7	1
	Sikandarpur	1
	Sohna Road	1
	Sushant Lok	1
	Udyog Vihar	1
	Woods Resort	1
1	Sector 43	1
	Sector 44	1
	South City 2	1
2	Sector 15	1
	Sector 53	1

Name: Locality, dtype: int64

Distribution of the different localities in Gurugram by cluster labels

From the Folium visualization of the clusters and the *Gurgaon_clusters_merged* dataframe grouped by 'Cluster Labels' and 'Locality' above, it was evident that *K-Means clustering* partitioned the localities in Gurgaon (or Gurugram) into **3 clusters** (labelled as **0**, **1**, and **2**).

- **Cluster 0,** indicated via *cyan* markers, was assigned the most number of localities (38).
- **Cluster 1,** indicated via *blue* markers, was assigned 3 localities (Sector 43, Sector 44, and South City 2).
- **Cluster 2,** indicated via *brown* markers, was assigned 2 localities (Sector 15 and Sector 53).

Conclusion

Analyzing the partioning of the different localities in Gurugram into the three clusters, i.e., 0, 1, and 2, with reference to the density of pre-existing eateries in the various localities within the city (indicated by the *Gurgaon_eateries* dataframe grouped by 'Locality', and then 'Venue Category') as well as the top 10 common types of eateries in each locality of the city (indicated by the *Gurgaon_eateries_grouped* dataframe grouped by 'Locality'), it was concluded that:

- *Cluster 1* and *Cluster 2* were assigned localities with the *least density of pre-existing eateries (indicative of low customer footfall)*. All these localities are listed as having only one restaurant within each of them. Hence, it was deduced that **the localities of Sector 43, Sector 44, South City 2, Sector 15, and Sector 53 in Gurugram are not ideal locations for opening a new restaurant of any kind, if a steady customer footfall is desired.**

- Within *Cluster 0*, the localities of DLF Cyber City, Sector 29, DLF Phases 1–5, Ambience Mall, MG Road, Sector 54, and Golf Course Road are all *densely populated with pre-existing eateries and can ensure regular customer traffic*. However, all **these localities already have at least one pre-existing Indian restaurant within them. Hence, it was inferred that opening a new Indian restaurant in these localities was not a good idea, keeping in mind competition from the same category of eatery**.

- Within *Cluster 0*, localities like *Woods Resort* (11 pre-existing eateries) and *Sector 21* (7 pre-existing eateries) have a medium density of established eateries — a reasonable measure of customer success. Further, these two localities do not have any Indian restaurant within them.

Thus, it was concluded that the localities of *Woods Resort* and *Sector 21* in Gurugram are potentially good choices for opening a new Indian restaurant, taking into account both customer traffic and competition from the same type of eatery.

Discussion and Future Scope

As discussed, Gurgaon (officially *Gurugram*) is one of the major satellite cities of the *National Capital Region* of India, serving as a flourishing location for opening a new themed restaurant or café. Hence, the goal of this project was two-fold:

- A. To analyze the various eateries in proximity to the localities within Gurugram in order to identify the most popular type of eatery across the localities.

- B. To identify appropriate localities within Gurugram for opening a new themed eatery, based on the most popular type of eatery across the city, customer footfall, and competition from pre-existing eateries.

Objective A was fulfilled via exploratory data analysis, whereas the unsupervised machine learning technique of *K-Means clustering* was used to address objective B. It was concluded that *Indian Restaurant* is the most common type of eatery across Gurugram, and the localities of *Woods Resort* and *Sector 21* in Gurugram would be ideal for opening a new Indian restaurant, considering customer traffic and competition from other Indian restaurants.

Although geospatial data pertaining to the different localities of Gurugram was used for this project, future iterations of the same can take into account other types of data, e.g. average per capita income of each of the localities, property rates, demographics, etc.

References:

1. *Course Materials for IBM Data Science Professional Certificate* (Source).
2. *Working With IBM Cloud Object Storage In Python* (Source).
3. *Kaggle: Zomato Restaurants Data* (Source).
4. *Foursquare API* (Source).
5. *How to Determine the Optimal K for K-Means?* (Source).
6. *K-Means Clustering algorithm* (Source).

The project code can be accessed here.