Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

INTRODUCTION

- New Delhi is the capital city of India, and one of the most densely populated ones.
- Historically, the city has served as the seat of various empires and ruling regimes dating back at least two thousand years, and as a result has a unique diversity in people, architectural design, languages and above all, food!!!
- Indians are known to be big time foodies, and saying New Delhi is the hub of restaurants, will not be far from truth. It was, therefore, an obvious choice for my Capstone project, which I selected to be about, of course, “food”!!!

INTRODUCTION

- The idea behind this project is to categorize the neighborhoods of a borough called South Delhi, into clusters and examine their cuisines, to identify the clusters' food habits.
- The project leverages data from Foursquare's 'Places API' and 'k-means clustering' unsupervised machine learning algorithm.
- This analysis can be used to understand the distribution of different cultures and popular restaurant cuisines in a posh neighborhood teeming with diversity. It can be utilized for choosing a location to open a restaurant of choice, by an investor, based on the cuisine. This study can also be used as an indicator of the neighborhood's cultural and ethnic diversity.

Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

Data Source

- ***Kaggle***
- Following data source was used:
- Link: https://www.kaggle.com/shaswatd673/delhi-neighborhood-data
- This data set was freely available on Kaggle.
- The excel file with the information Boroughs and Neighborhoods in New Delhi was downloaded from the link above.
- ***Foursquare API***
- Link: https://developer.foursquare.com/docs.
- Foursquare API, a location data provider, was used to make API calls to retrieve data about venues in different neighborhoods.

Data Cleaning

- The excel file was read into a dataframe and QA/ QC on the data set was performed.
- Rows with missing information or incorrect information were removed.
- The cleaned data set had 9 Boroughs and 153 n3ighborhoods.
- The dataframe was filered for South Delhi borough for analysis.

Data Filtering

[12]:

	Borough	Neighborhood	latitude	longitude
0	South Delhi	Alaknanda	28.529336	77.251632
1	South Delhi	Chhattarpur	28.507007	77.175417
2	South Delhi	Chittaranjan Park	28.538752	77.249249
3	South Delhi	Dayanand Colony	28.562200	77.247613
4	South Delhi	Defence Colony	28.571791	77.232010
5	South Delhi	East of Kailash	28.557032	77.244614
6	South Delhi	Friends Colony	28.566751	77.261918
7	South Delhi	Greater Kailash	28.554633	77.228570
8	South Delhi	Green Park	28.558002	77.206821
9	South Delhi	Gulmohar Park	28.557101	77.213006
10	South Delhi	Hauz Khas	28.544256	77.206707
11	South Delhi	Hauz Khas Village	28.553855	77.194713
12	South Delhi	Jaitpur	28.500477	77.316192
13	South Delhi	Jangpura	28.582457	77.241500
14	South Delhi	Jasola	28.542233	77.294386

- The dataframe was filtered for the South Delhi borough.
- The ‘geopy’ library was used to get the latitude and longitude values of New Delhi, which was returned to be Latitude: 28.61, Longitude: 77.21.
- The curated dataframe was then used to generate a map of New Delhi with neighborhoods superimposed on top using the python ‘folium’ library.

Data Filtering

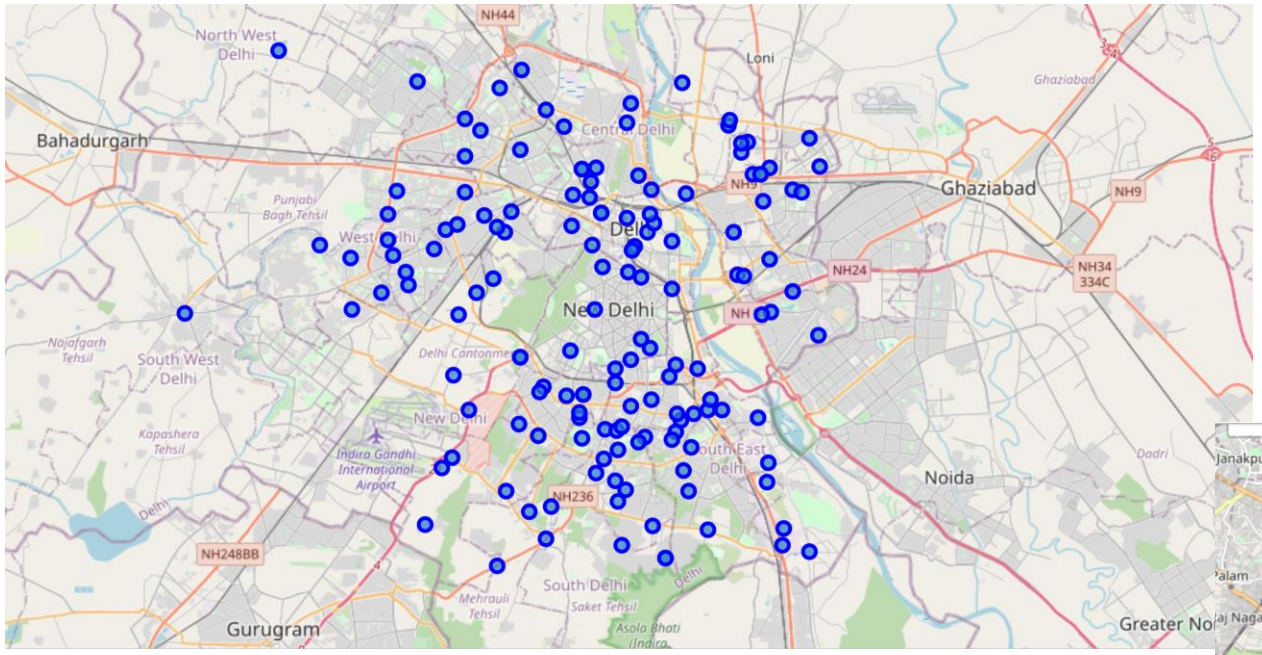

Fig: Map of New Delhi showing all the neighborhoods

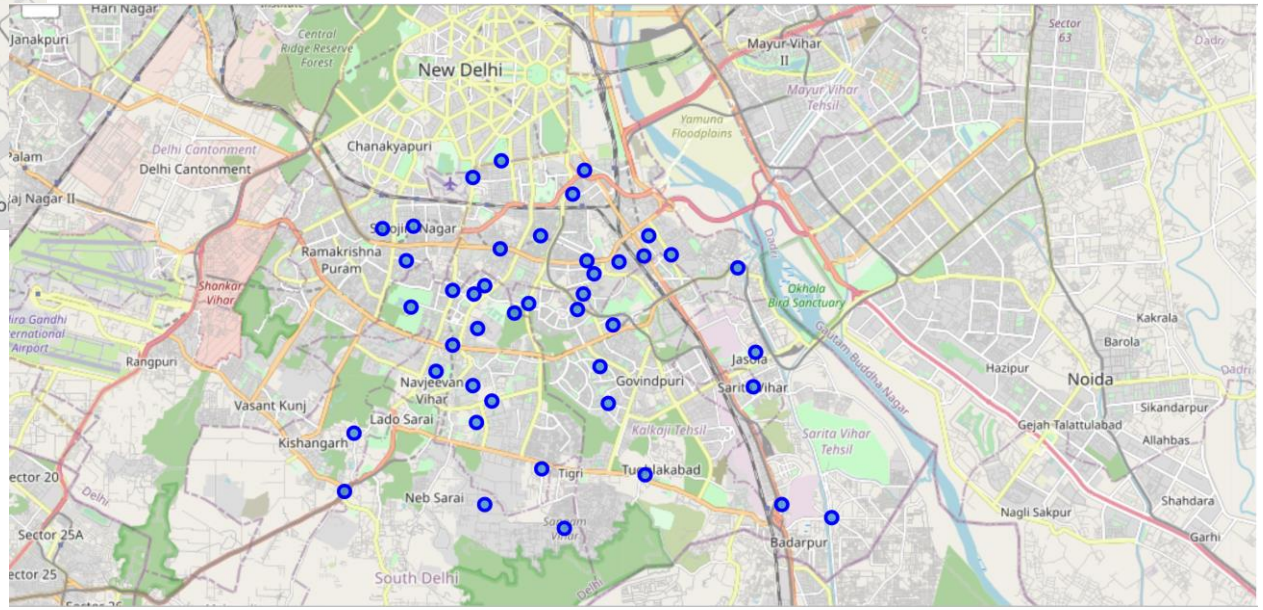

Fig: Map of South Delhi Borough with neighborhoods

Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

Foursquare API Calls

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Alaknanda	28.529336	77.251632	City Of Joy	28.532685	77.253003	Restaurant
1	Alaknanda	28.529336	77.251632	Starbucks	28.534053	77.243059	Coffee Shop
2	Alaknanda	28.529336	77.251632	Culinaire	28.530777	77.245816	Thai Restaurant
3	Alaknanda	28.529336	77.251632	Yeti - The Himalayan Kitchen	28.533562	77.242361	Tibetan Restaurant
4	Alaknanda	28.529336	77.251632	CR Park Market No. 2	28.536463	77.253386	Market
5	Alaknanda	28.529336	77.251632	Costa Coffee	28.533811	77.243336	Coffee Shop
6	Alaknanda	28.529336	77.251632	Chocolateria San Churro	28.534612	77.243642	Dessert Shop
7	Alaknanda	28.529336	77.251632	China Garden	28.532931	77.243214	Chinese Restaurant
8	Alaknanda	28.529336	77.251632	Artusi Ristorante e Bar	28.533452	77.242032	Italian Restaurant
9	Alaknanda	28.529336	77.251632	Olympia Gymnasium & Spa	28.527990	77.245713	Gym
10	Alaknanda	28.529336	77.251632	Amalfi	28.532367	77.242859	Italian Restaurant
11	Alaknanda	28.529336	77.251632	CR Park Market No. 1	28.540075	77.248847	Market
12	Alaknanda	28.529336	77.251632	Smoke House Grill	28.537651	77.238286	Restaurant

Fig: Dataframe showing the venues for each neighborhood

- The Foursquare API was used to explore the neighborhoods and segment them using the API, ‘CLIENT_ID’ and ‘CLIENT_SECRET’

Foursquare API Calls- Data Preparation

```
DelhiNW_venues= DelhiNW_venues[(DelhiNW_venues['Venue Category'].str.contains('Restaurant')) & (DelhiNW_venues['Venue Category'] !=('Restaurant'))]
DelhiNW_venues
```

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Alaknanda	28.529336	77.251632	Culinaire	28.530777	77.245816	Thai Restaurant
3	Alaknanda	28.529336	77.251632	Yeti - The Himalayan Kitchen	28.533562	77.242361	Tibetan Restaurant
7	Alaknanda	28.529336	77.251632	China Garden	28.532931	77.243214	Chinese Restaurant
8	Alaknanda	28.529336	77.251632	Artusi Ristorante e Bar	28.533452	77.242032	Italian Restaurant
10	Alaknanda	28.529336	77.251632	Amalfi	28.532367	77.242859	Italian Restaurant
16	Alaknanda	28.529336	77.251632	Naivedyam	28.541933	77.254269	Indian Restaurant
17	Alaknanda	28.529336	77.251632	Chungwa	28.534002	77.243135	Asian Restaurant
19	Alaknanda	28.529336	77.251632	Not Just Paranthas	28.532235	77.242645	Indian Restaurant
22	Alaknanda	28.529336	77.251632	Bikanervala	28.537367	77.239015	Indian Restaurant

Fig: Filtered dataframe showing venues categorized as restaurants

```
array(['Thai Restaurant', 'Tibetan Restaurant', 'Chinese Restaurant',
       'Italian Restaurant', 'Indian Restaurant', 'Asian Restaurant',
       'French Restaurant', 'Fast Food Restaurant',
       'Mediterranean Restaurant', 'Japanese Restaurant',
       'Bengali Restaurant', 'Australian Restaurant',
       'South Indian Restaurant', 'English Restaurant',
       'Modern European Restaurant', 'Eastern European Restaurant',
       'Middle Eastern Restaurant', 'Vegetarian / Vegan Restaurant',
       'Korean Restaurant', 'American Restaurant', 'Tapas Restaurant',
       'Turkish Restaurant', 'Scandinavian Restaurant',
       'Mexican Restaurant', 'Burmese Restaurant', 'Mughlai Restaurant',
       'North Indian Restaurant', 'Comfort Food Restaurant',
       'Falafel Restaurant', 'Tex-Mex Restaurant',
       'Northeast Indian Restaurant', 'Seafood Restaurant',
       'New American Restaurant'], dtype=object)
```

Fig: Final categories for the data frame for feature engineering

- The dataframe was filtered so as to contain only those venues which had restaurants marked as a category. Also, the venues for which the type of restaurant was not available was dropped from the data frame.
- Thirty-Four unique restaurant categories were finally captured in the data frame.

Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

Methodology

- Next, each neighborhood was analyzed individually to understand the most common cuisine being served within its vicinity.

- The above process was implemented using ‘one hot encoding’ function of python ‘pandas’ library. One hot encoding converts the categorical variables (which are ‘Venue Category’) to binary form that can be provided to Machine Learning algorithms for classification, clustering etc.

- After one hot encoding, the venues were sorted by the most common occurrence, as shown below.

Methodology- One Hot Encoding & Sorting

	Neighborhood	Afghan Restaurant	American Restaurant	Asian Restaurant	Australian Restaurant	Bengali Restaurant	Burmese Restaurant	Chinese Restaurant
0	Alaknanda	0.000000	0.0	0.157895	0.000000	0.000000	0.0	0.157895
1	Chhattarpur	0.000000	0.0	0.000000	0.000000	0.000000	0.0	0.111111
2	Chittaranjan Park	0.000000	0.0	0.111111	0.037037	0.037037	0.0	0.148148
3	Dayanand Colony	0.000000	0.0	0.000000	0.000000	0.033333	0.0	0.066667
4	Defence Colony	0.028571	0.0	0.028571	0.000000	0.000000	0.0	0.085714

Fig: One hot encoding output

Fig: Venues sorted by most commonly visited to least for each neighborhood

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Alaknanda	Indian Restaurant	Chinese Restaurant	Asian Restaurant	Fast Food Restaurant	Italian Restaurant
1	Chhattarpur	Indian Restaurant	Italian Restaurant	Fast Food Restaurant	Japanese Restaurant	Thai Restaurant
2	Chittaranjan Park	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant	Asian Restaurant	Italian Restaurant
3	Dayanand Colony	Indian Restaurant	Fast Food Restaurant	Italian Restaurant	Chinese Restaurant	French Restaurant
4	Defence Colony	Indian Restaurant	Italian Restaurant	Fast Food Restaurant	Chinese Restaurant	Japanese Restaurant

Methodology- Clustering and Optimal K

- 'k-means' is an unsupervised machine learning algorithm which creates clusters of data points aggregated together based on certain similarities.

- This algorithm was used to divide South Delhi into clusters of neighborhoods with similar 'taste'. To implement this algorithm, its vital to determine the optimal number of clusters (i.e. k). 'The Elbow Method' was used for this.

- The Elbow Method calculates the sum of squared distances of samples to their closest cluster center 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.

Methodology- Optimal K and Clustered Dataframe

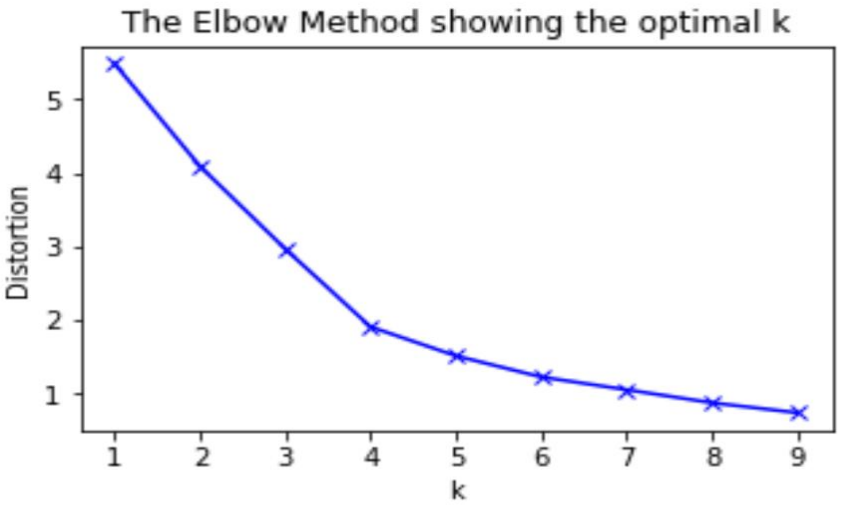

Fig: Elbow method to determine optimal K

- The Elbow Method calculates the sum of squared distances of samples to their closest cluster center 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. K was selected as 5.

Borough	Neighborhood	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
South Delhi	Alaknanda	28.529336	77.251632	0	Indian Restaurant	Chinese Restaurant	Asian Restaurant
South Delhi	Chhattarpur	28.507007	77.175417	3	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
South Delhi	Chittaranjan Park	28.538752	77.249249	0	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant
South Delhi	Dayanand Colony	28.562200	77.247613	3	Indian Restaurant	Fast Food Restaurant	Italian Restaurant
South Delhi	Defence Colony	28.571791	77.232010	3	Indian Restaurant	Italian Restaurant	Fast Food Restaurant

Fig: Dataframe showing clusters of neighborhoods

Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

Results- Clustered Neighborhoods of S Delhi

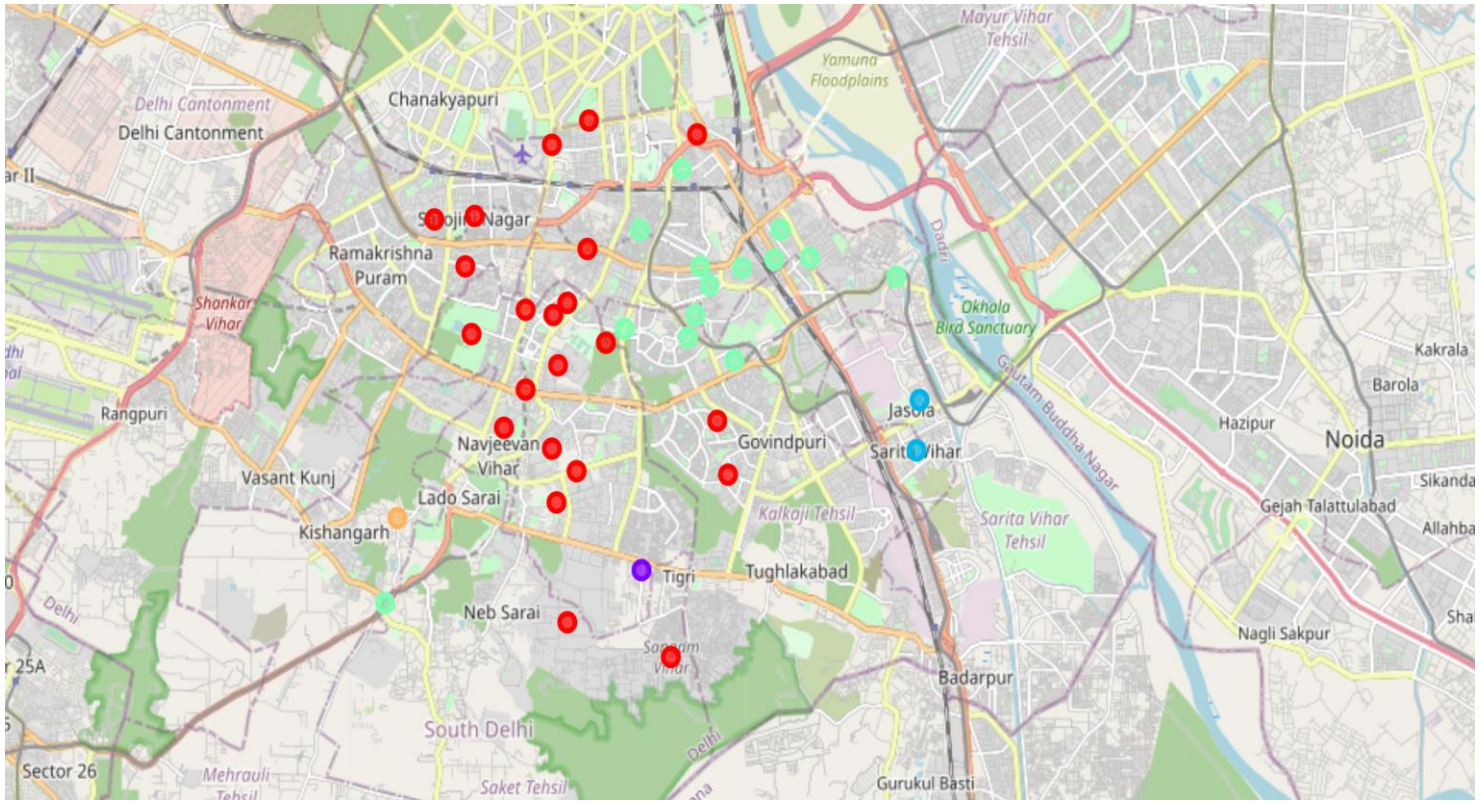

Fig: Map of South Delhi showing clustering of neighborhoods

- The results from clustering were displayed on the map for South Delhi, shown to the left.
- A total of 5 clusters with varying number of neighborhoods were identified.
- In the next section, the results from this clustering are discussed

Results- Cluster 0

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Alaknanda	Indian Restaurant	Chinese Restaurant	Asian Restaurant
Chittaranjan Park	Indian Restaurant	Fast Food Restaurant	Chinese Restaurant
Green Park	Indian Restaurant	Chinese Restaurant	Italian Restaurant
Gulmohar Park	Indian Restaurant	Chinese Restaurant	Mediterranean Restaurant
Hauz Khas	Indian Restaurant	Asian Restaurant	Chinese Restaurant
Hauz Khas Village	Indian Restaurant	Asian Restaurant	Chinese Restaurant
Jor Bagh	Indian Restaurant	Eastern European Restaurant	Chinese Restaurant
Khirki Village	Indian Restaurant	Chinese Restaurant	Asian Restaurant
Lodi Colony	Indian Restaurant	Chinese Restaurant	Italian Restaurant
Malviya Nagar	Indian Restaurant	Chinese Restaurant	Asian Restaurant
Neeti Bagh	Indian Restaurant	Chinese Restaurant	Italian Restaurant

Fig: Cluster-0, showing most commonly visited restaurant types

- Following are the results of the Cluster — 0. The Indian Restaurant is the 1st most visited restaurant type in this cluster. All the other clusters, which had more than one neighborhood, had the Indian restaurant as the 1st most visited type of restaurant.
- It is important to observe the 2nd and 3rd most visited restaurant types. As we can see, the Chinese restaurant type is the 2nd most visited, also very popular in India. If somewhere to open a Chinese restaurant, this would be good choice for location, this cluster of neighborhoods.

Results- Cluster 3

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Chhattarpur	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
Dayanand Colony	Indian Restaurant	Fast Food Restaurant	Italian Restaurant
Defence Colony	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
East of Kailash	Indian Restaurant	Fast Food Restaurant	Italian Restaurant
Friends Colony	Fast Food Restaurant	Indian Restaurant	Vegetarian / Vegan Restaurant
Greater Kailash	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
Jangpura	Indian Restaurant	Italian Restaurant	Fast Food Restaurant
Kailash Colony	Indian Restaurant	Fast Food Restaurant	Italian Restaurant
Lajpat Nagar	Indian Restaurant	Fast Food Restaurant	Italian Restaurant

Fig: Cluster-3, showing most commonly visited restaurant types

- Following are the results of the Cluster -3.
- The 1st most commonly visited restaurant type is Indian, followed by Fast Food and Italian restaurants
- Fast Food and Italian restaurants were equally popular after Indian Restaurants

Results- Clusters 1,2 and 4

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Khanpur	Thai Restaurant	Vegetarian / Vegan Restaurant	Comfort Food Restaurant

Fig: Cluster-1, showing most commonly visited restaurant types

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
14	Jasola	Indian Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant
34	Sarita Vihar	Indian Restaurant	Vegetarian / Vegan Restaurant	Dumpling Restaurant

Fig: Cluster-2, showing most commonly visited restaurant types

Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
Mehrauli	Thai Restaurant	Italian Restaurant	Japanese Restaurant

Fig: Cluster-4, showing most commonly visited restaurant types

- Cluster -1. As we can notice, the 1st most commonly visited restaurant type is Thai, followed by Vegetarian or Vegan. This cluster returned with only one neighborhood, which kind of makes it stand out in the clusters.
- Cluster -2. The 1st most commonly visited restaurant type is Indian, followed by Vegetarian or Vegan. This cluster also returned with only two neighborhoods.
- Cluster -4. The 1st most commonly visited restaurant type is Thai, followed Italian and Japanese. This cluster also returned with only one neighborhood.

Capstone Project- IBM Data Science Professional Certificate

- Introduction
- Data Source/ Data Cleaning
- API Calls to Foursquare
- Methodology
  - Feature Engineering
  - Unsupervised machine learning
  - K means clustering- optimal K
- Results
- Conclusions

Discussion

- The unsupervised machine learning algorithm does a good job at clustering the restaurants, which upon closer examination clearly reveal the 1st, 2nd and 3rd most commonly visited type of restaurants.
- What was also considered was the total number of neighborhoods returned in a cluster, for added confidence in the analysis.
- Based on this, there were two clusters which were identified. It must be noted, that in both these clusters, Indian Restaurant is the most common type, which is very akin to North India where New Delhi is. This analysis clearly reveals competitive advantage for an entrepreneur looking to invest in restaurants in South Delhi.
- *Cluster 0 — Indian/ Chinese*
- *Cluster 3 — Indian/ Fast Food/ Italian*

Conclusion

- Data analytics was put to good use in identifying which regions of South Delhi in India had which restaurant types as the most popular. Results were then refined to identify two clusters which had at least five neighborhoods clustered t0ogether, and clearly identified the popular choices of cuisines.
- Another observation was that, Indian Restaurant was the most commonly visited in most neighborhoods, followed by Chinese cuisine in one cluster and Fast-Food/ Italian in another.
- All of this information can be used to make an investment decision for the choice of restaurant type and also the location.