Capstone Project

Zomato Restaurant Clustering and Sentiment Analysis

By:- Om Prakash Pradhan & Ruchika Nayak

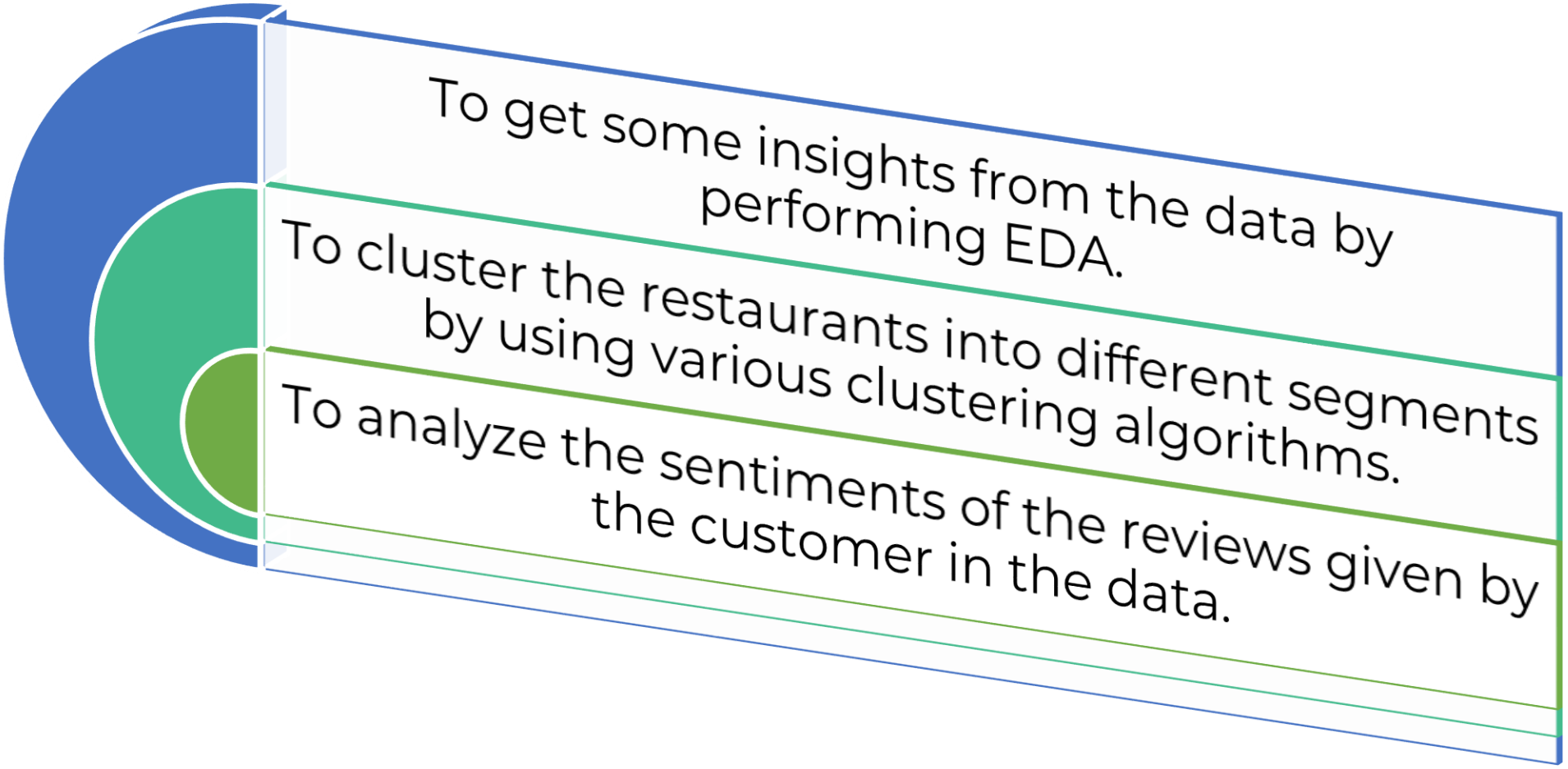
Key Points

- ❑ **Introduction**
- ❑ **Project Objectives**
- ❑ **Attribute Information**
- ❑ **Insights from EDA**
- ❑ **Clustering**
  - Methodology for clustering
  - Dataset for clustering
  - Results from clustering
- ❑ **Sentiment Analysis**
  - ❑ Methodology for sentiment analysis
  - ❑ Dataset for sentiment analysis
  - ❑ Results from sentiment analysis
- ❑ **Conclusions**
- ❑ **Challenges**

Introduction

- Zomato is an Indian restaurant aggregator and food delivery start-up. It provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities.
- This project mainly focuses on analyzing the Zomato restaurants data to get some insights by performing clustering and sentiment analysis.

Project Objectives



Attribute Information

Zomato Restaurant names and Metadata

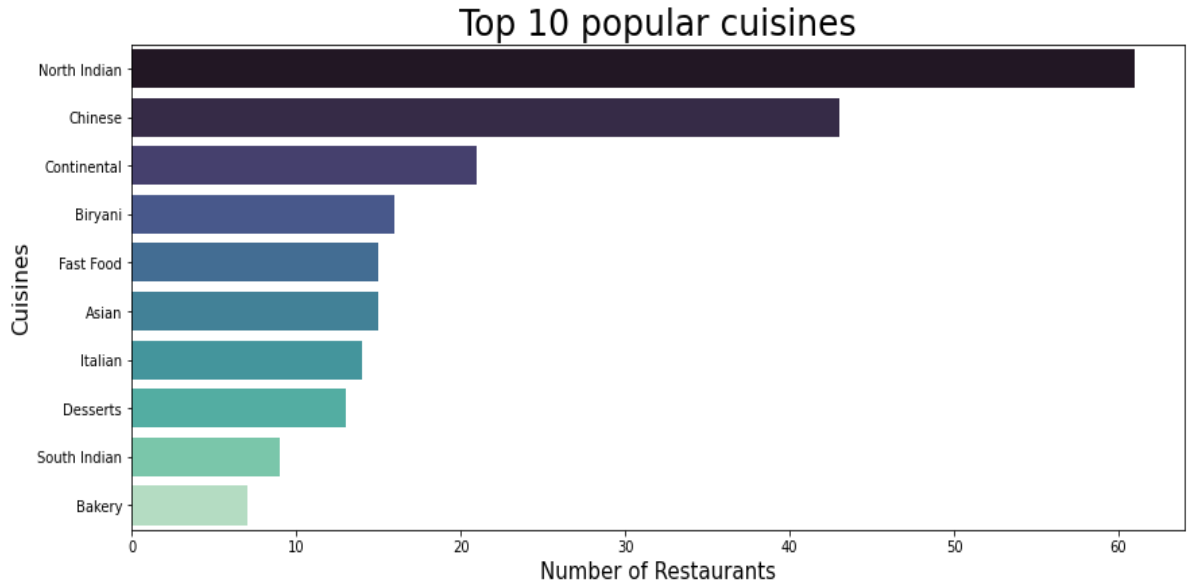
SL No.	Name	Description
1	**Name**	Name of restaurants
2	**Links**	URL links of restaurants
3	**Cost**	Per person estimated cost of dining
4	**Collection**	Tagging of restaurants w.r.t. Zomato category
5	**Cuisines**	Cuisines served by restaurants
6	**Timings**	Restaurants timing

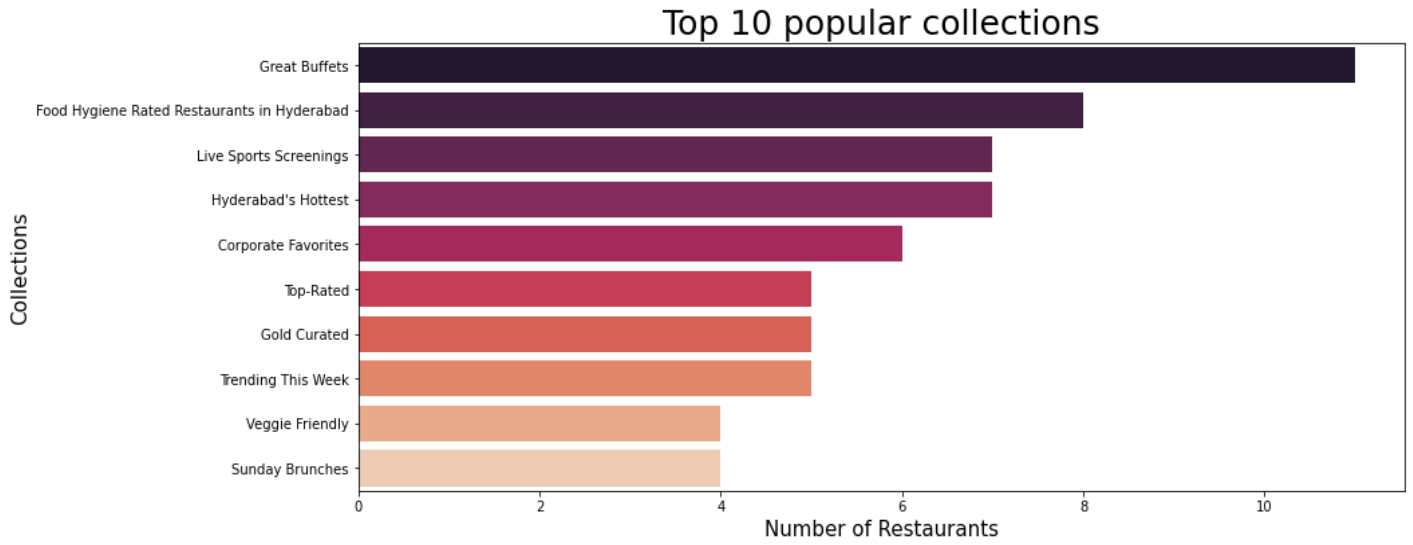
Zomato Restaurant reviews

SL No.	Name	Description
1	**Restaurants**	Name of the restaurant
2	**Reviewer**	Name of the reviewer
3	**Review**	Review text
4	**Rating**	Rating provided by Reviewer
5	**MetaData**	Reviewer MetaData - No. of Reviews and Followers
6	**Time**	Date and Time of Review
7	**Pictures**	No. of pictures posted with review

Insights from EDA

These horizontal bar plots show the top 10 popular cuisines and collections.





Insights from EDA

Top 10 costliest restaurants

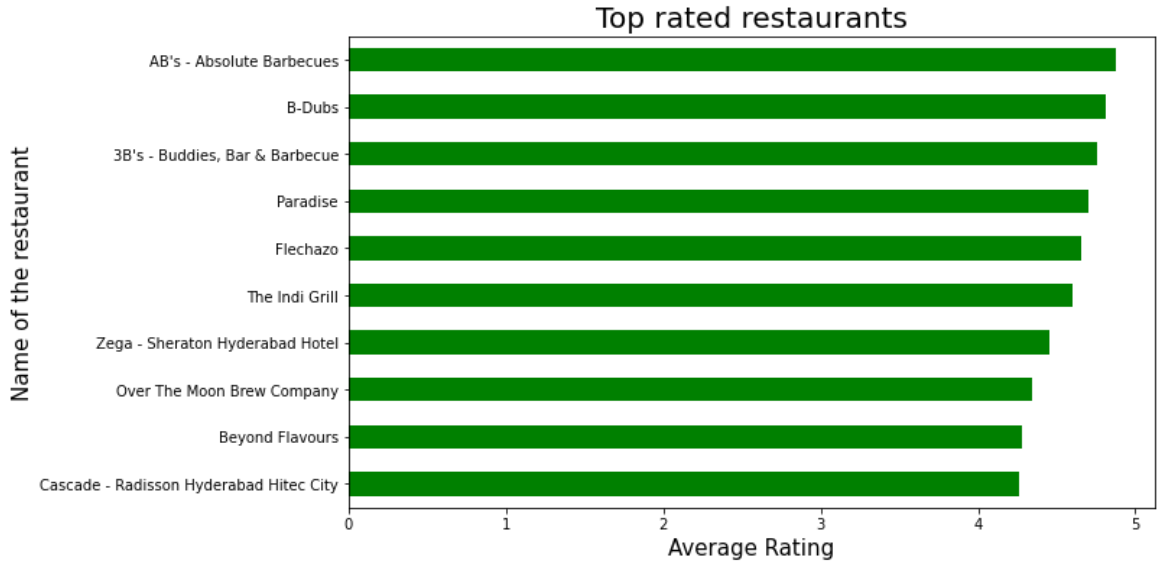
Name	Cuisines	Cost
Collage - Hyatt Hyderabad Gachibowli	Continental, Italian, North Indian, Chinese, A...	2800
Feast - Sheraton Hyderabad Hotel	Modern Indian, Asian, Continental, Italian	2500
Jonathan's Kitchen - Holiday Inn Express & Suites	North Indian, Japanese, Italian, Salad, Sushi	1900
10 Downing Street	North Indian, Chinese, Continental	1900
Cascade - Radisson Hyderabad Hitec City	North Indian, Italian, Continental, Asian	1800
Zega - Sheraton Hyderabad Hotel	Asian, Sushi	1750
Republic Of Noodles - Lemon Tree Hotel	Thai, Asian, Chinese, Malaysian	1700
Mazzo - Marriott Executive Apartments	Italian, North Indian, South Indian, Asian	1700
Arena Eleven	Continental	1600
Barbeque Nation	Mediterranean, North Indian, Kebab, BBQ	1600

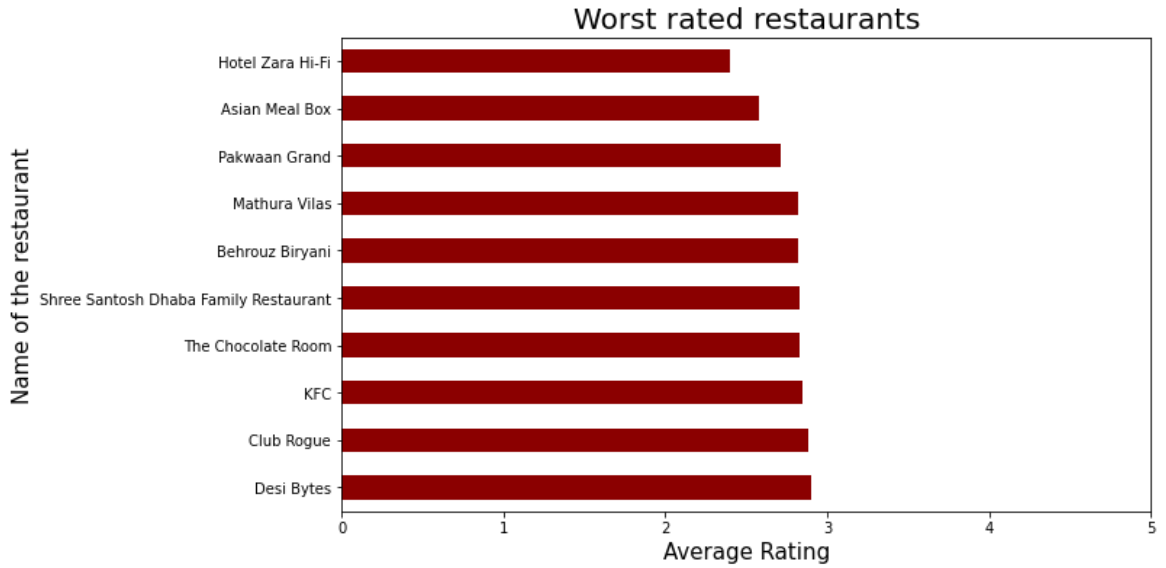
Top 10 cheapest restaurants

Name	Cuisines	Cost
Mohammedia Shawarma	Street Food, Arabian	150
Amul	Ice Cream, Desserts	150
Asian Meal Box	Asian	200
KS Bakers	Bakery, Desserts, Fast Food	200
Momos Delight	Momos	200
Hunger Maggi Point	Fast Food	200
Sweet Basket	Bakery, Mithai	200
Wich Please	Fast Food	250
Shah Ghouse Spl Shawarma	Lebanese	300
Cream Stone	Ice Cream, Desserts	350

Insights from EDA

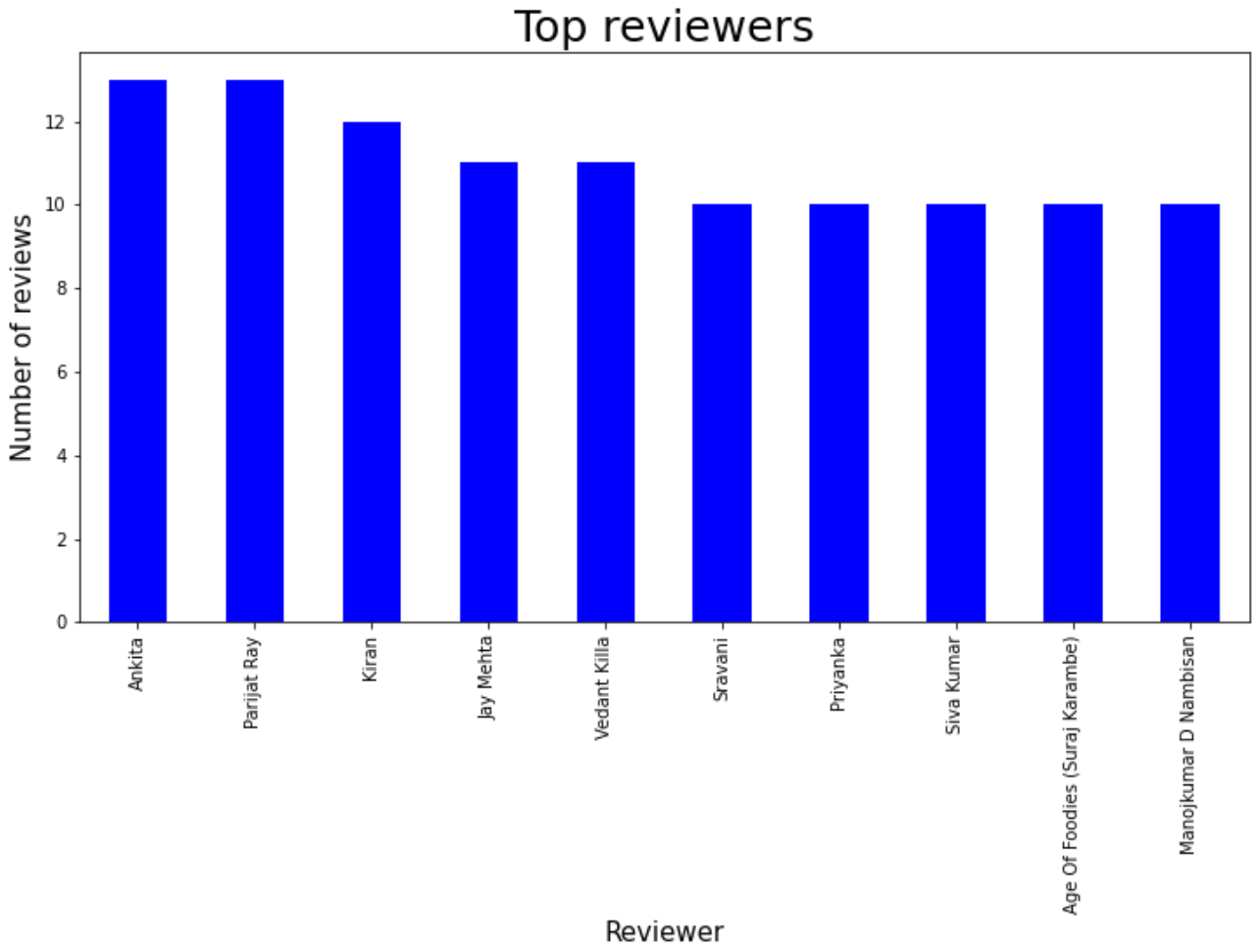
These horizontal bar plots show the top rated and worst rated restaurants.





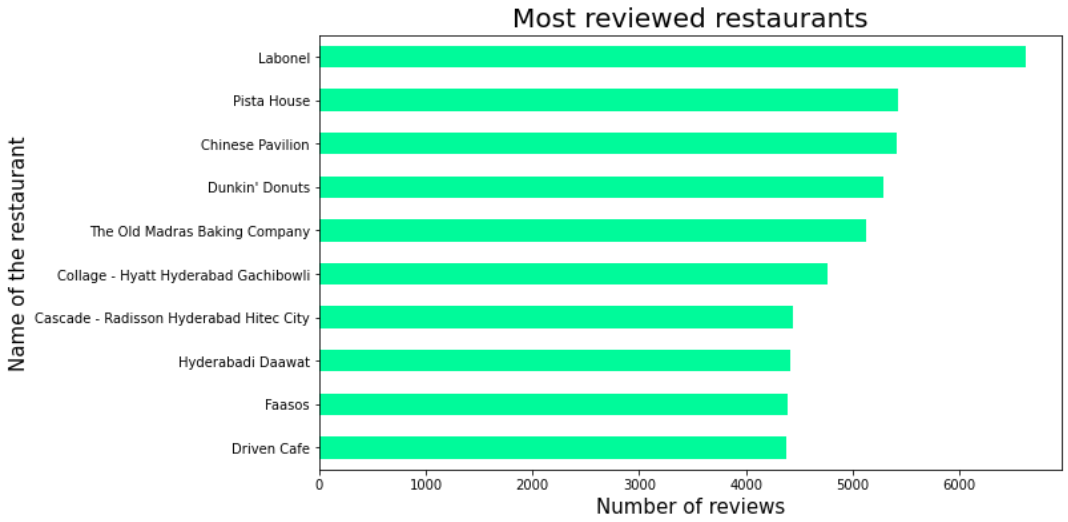
Insights from EDA

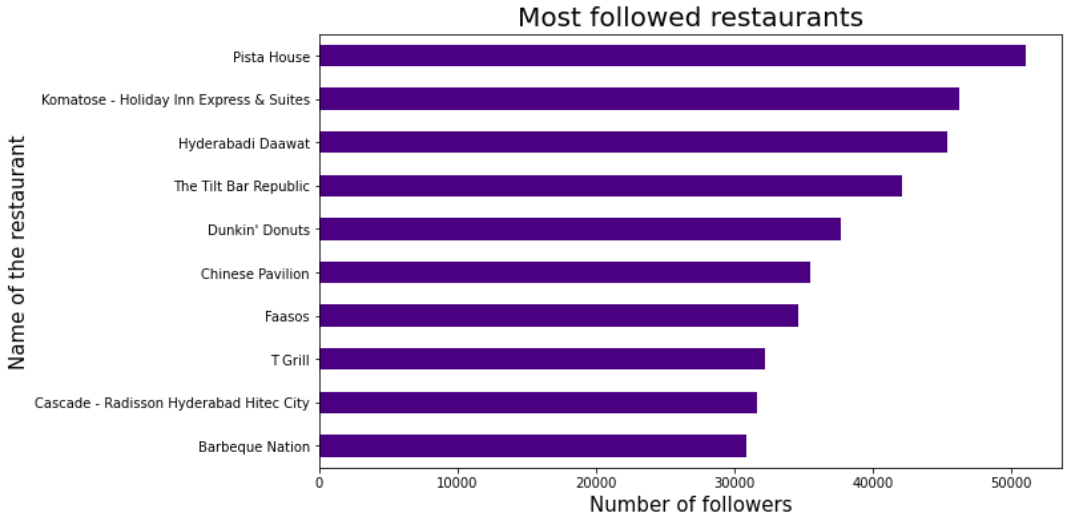
This bar plot shows the reviewers with most number of reviews.



Insights from EDA

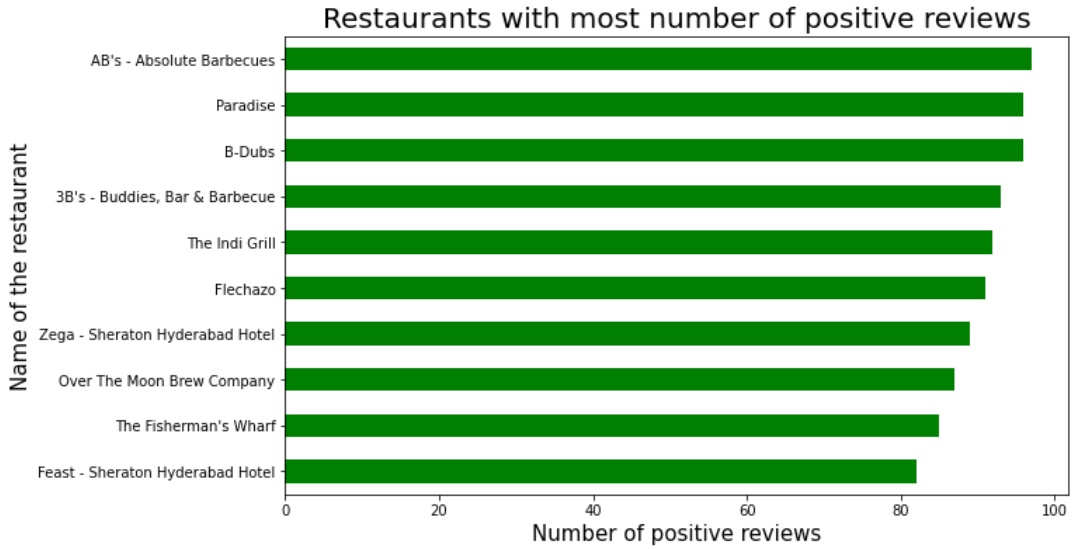
These horizontal bar plots show the most reviewed and most followed restaurants.

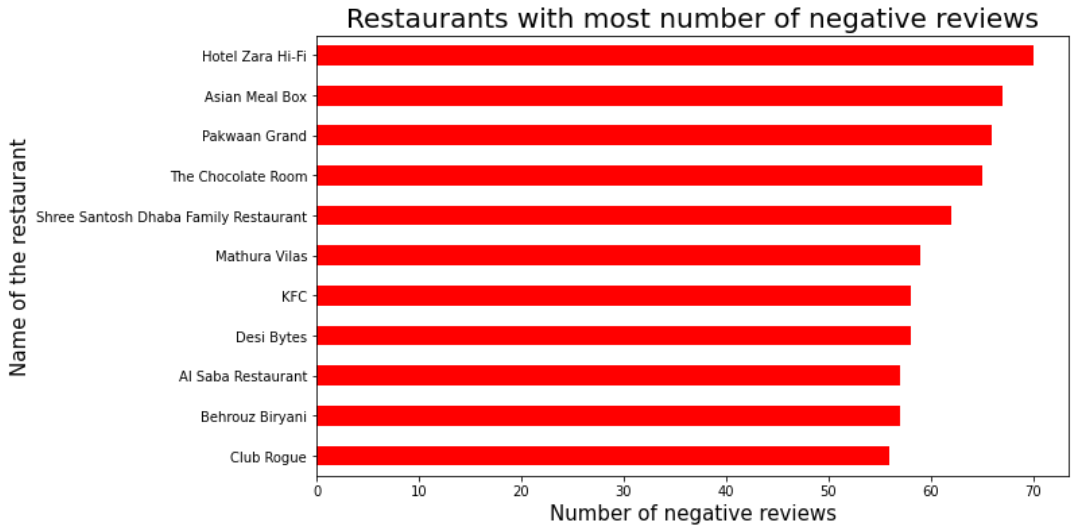




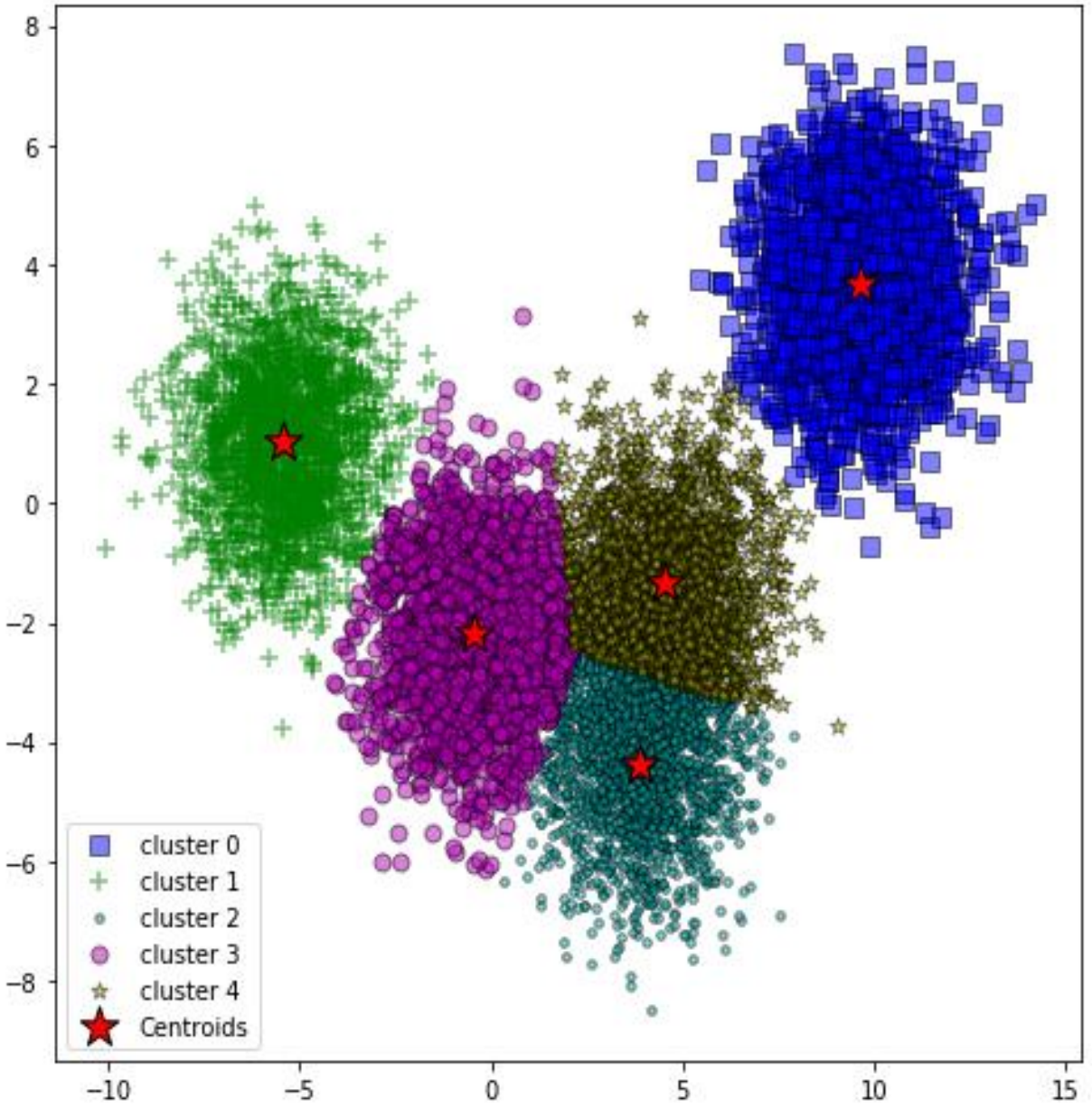
Insights from EDA

These horizontal bar plots show the restaurants with most positive and most negative reviews.





Clustering

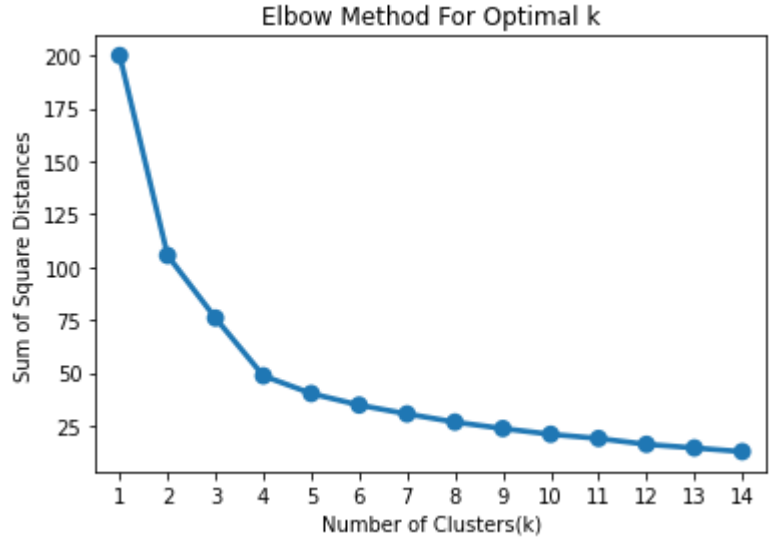

8
6
4
2
0
−2
−4
−6
−8
−10
−5
0
5
10
15
cluster 0
cluster 1
cluster 2
cluster 3
cluster 4
Centroids

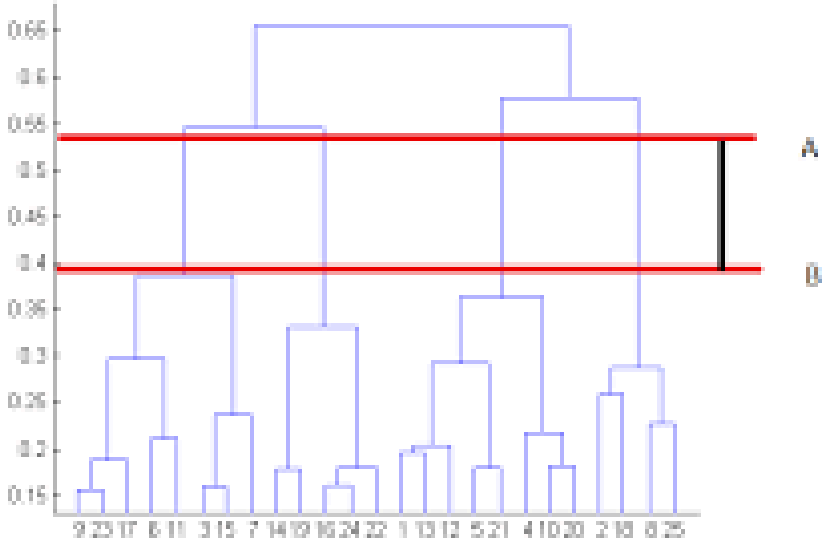
Clustering

- Clustering or cluster analysis is a machine learning technique, which groups the un labelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points.
- **K-Means** is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
- **Silhouette score** is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other.

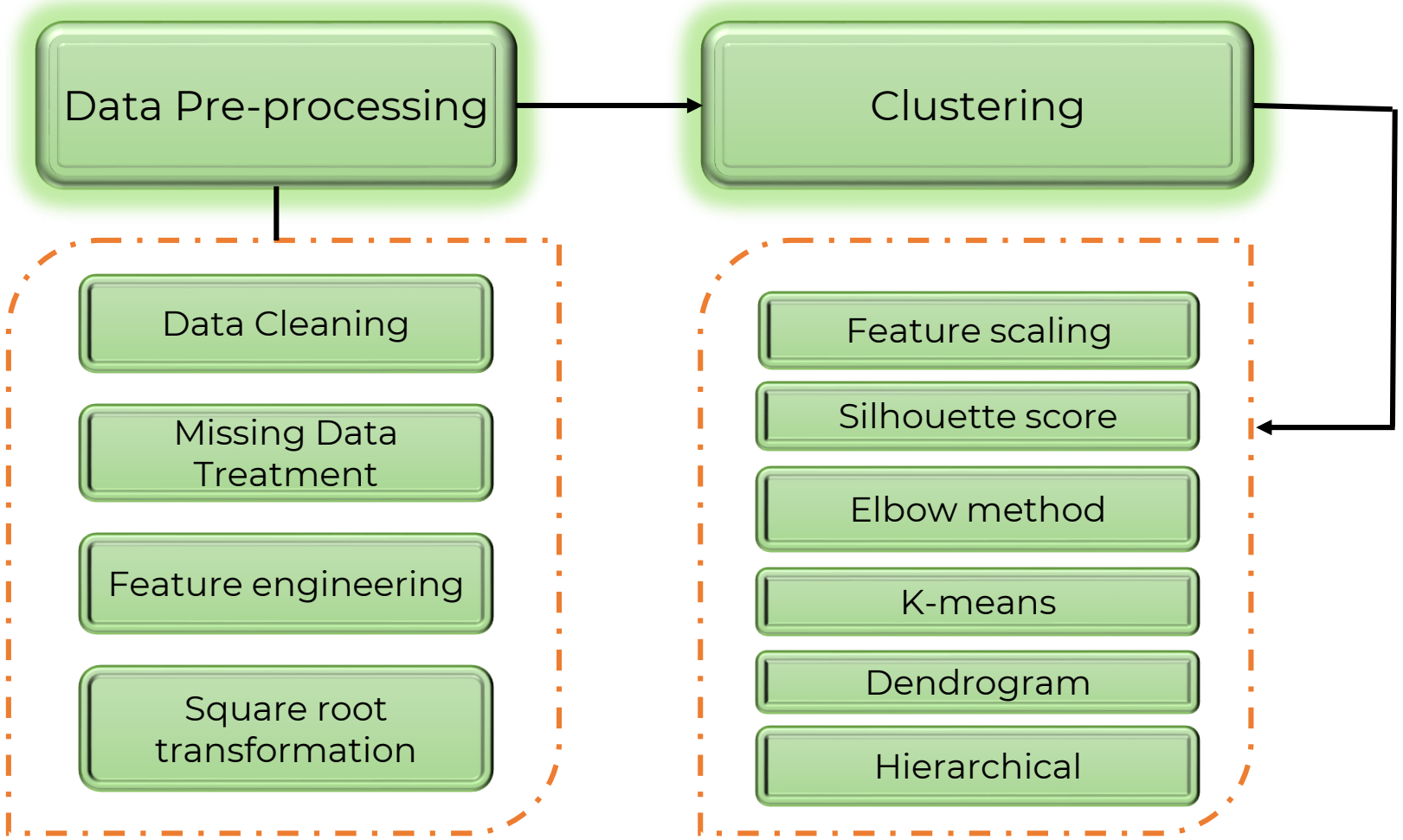
Clustering

- In cluster analysis, the **elbow method** is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.
- The sole concept of hierarchical clustering lies in just the construction and analysis of a **dendrogram**. A dendrogram is a tree-like structure that explains the relationship between all the data points in the system.





Methodology for clustering



Dataset for clustering

From the metadata column, we have separated the reviews and followers. For this we made 2 separate column to store those values.

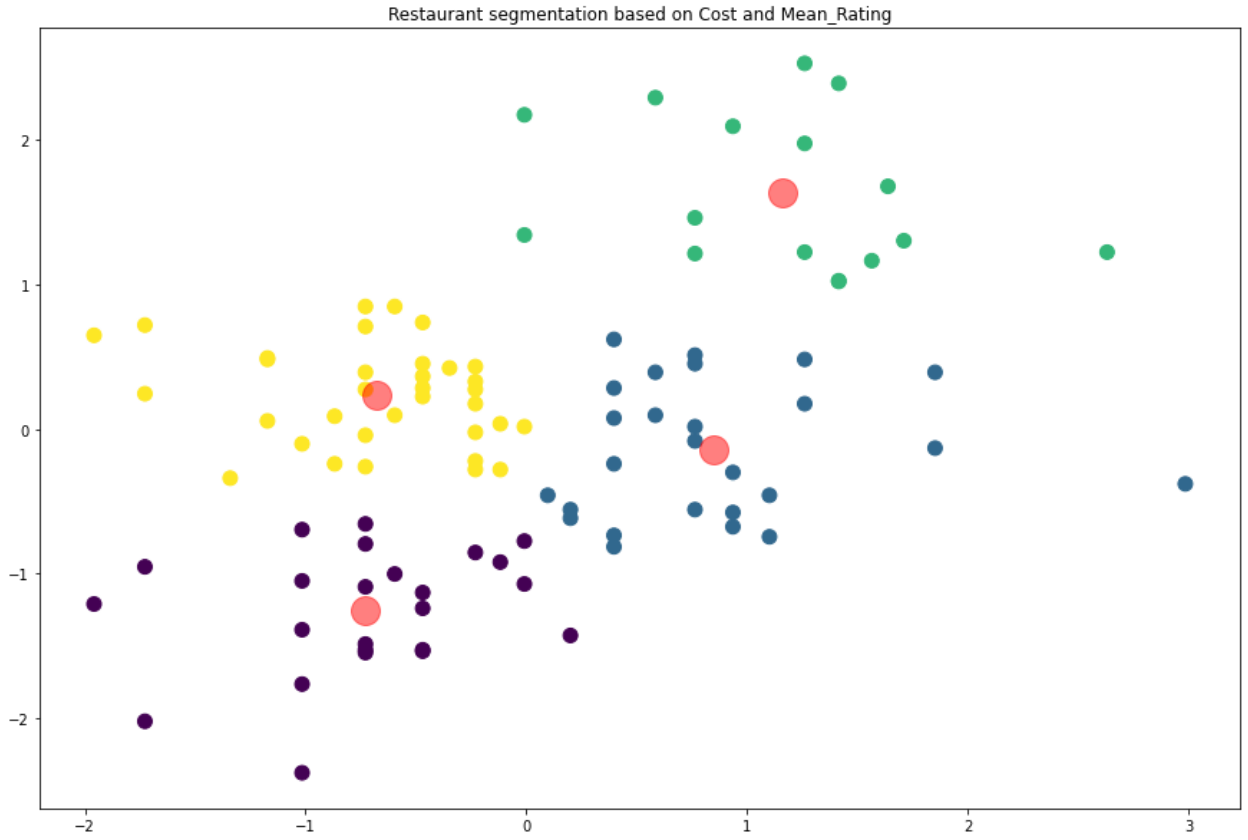
Then we aggregated the 'Rating‘ and 'No_of_Followers‘ columns to make it a single value for each restaurant. For this we took the average of each column by grouping.

	Name	Cost	Mean_Rating	Mean_Followers
0	Beyond Flavours	28.284271	4.280	11.582170
1	Paradise	28.284271	4.700	3.046157
2	Flechazo	36.055513	4.660	6.357025
3	Shah Ghouse Hotel & Restaurant	28.284271	3.210	12.483147
4	Over The Moon Brew Company	34.641016	4.340	8.545223
5	The Fisherman's Wharf	38.729833	4.220	12.048432
6	eat.fit	22.360680	3.200	15.873679
7	Shah Ghouse Spl Shawarma	17.320508	3.430	12.232453
8	Hyper Local	31.622777	3.640	13.051330
9	Cream Stone	18.708287	3.845	12.610953

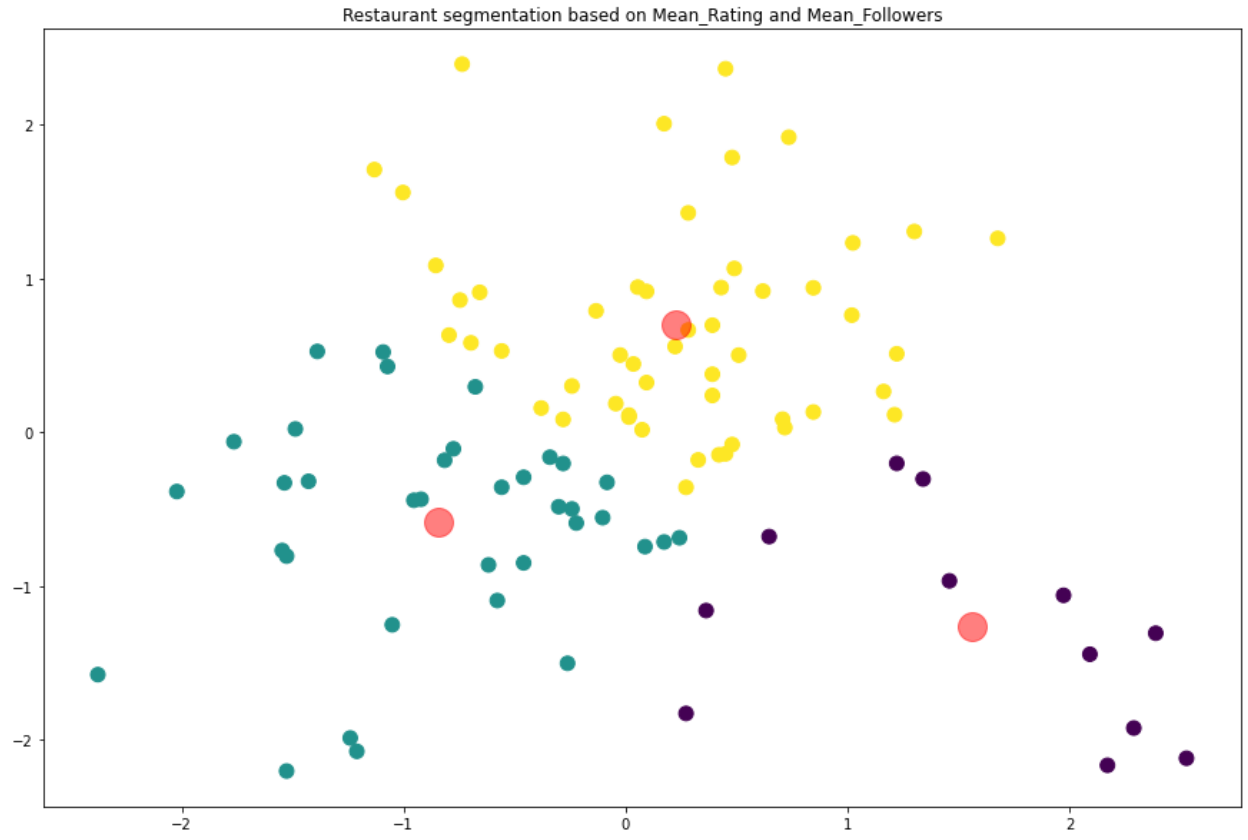
Results from clustering

SL No.	Model_Name	Data	Optimal_Number_of_clusters
1	K-Means with silhouette_score	Cost and Mean_Rating	4
2	K-Means with Elbow method	Cost and Mean_Rating	4
3	Hirarchical Clustering	Cost and Mean_Rating	2
4	K-Means with silhouette_score	Mean_Rating and Mean_Followers	3
5	K-Means with Elbow method	Mean_Rating and Mean_Followers	3
6	Hierarchical Clustering	Mean_Rating and Mean_Followers	3
7	K-Means with silhouette_score	Cost and Mean_Followers	9
8	K-Means with Elbow method	Cost and Mean_Followers	3
9	Hierarchical clustering	Cost and Mean_Followers	2
10	K-Means with silhouette_score	Cost, Mean_Rating and Mean_Followers	6
11	K-Means with Elbow method	Cost, Mean_Rating and Mean_Followers	4
12	Hierarchical clustering	Cost, Mean_Rating and Mean_Followers	2

Clusters

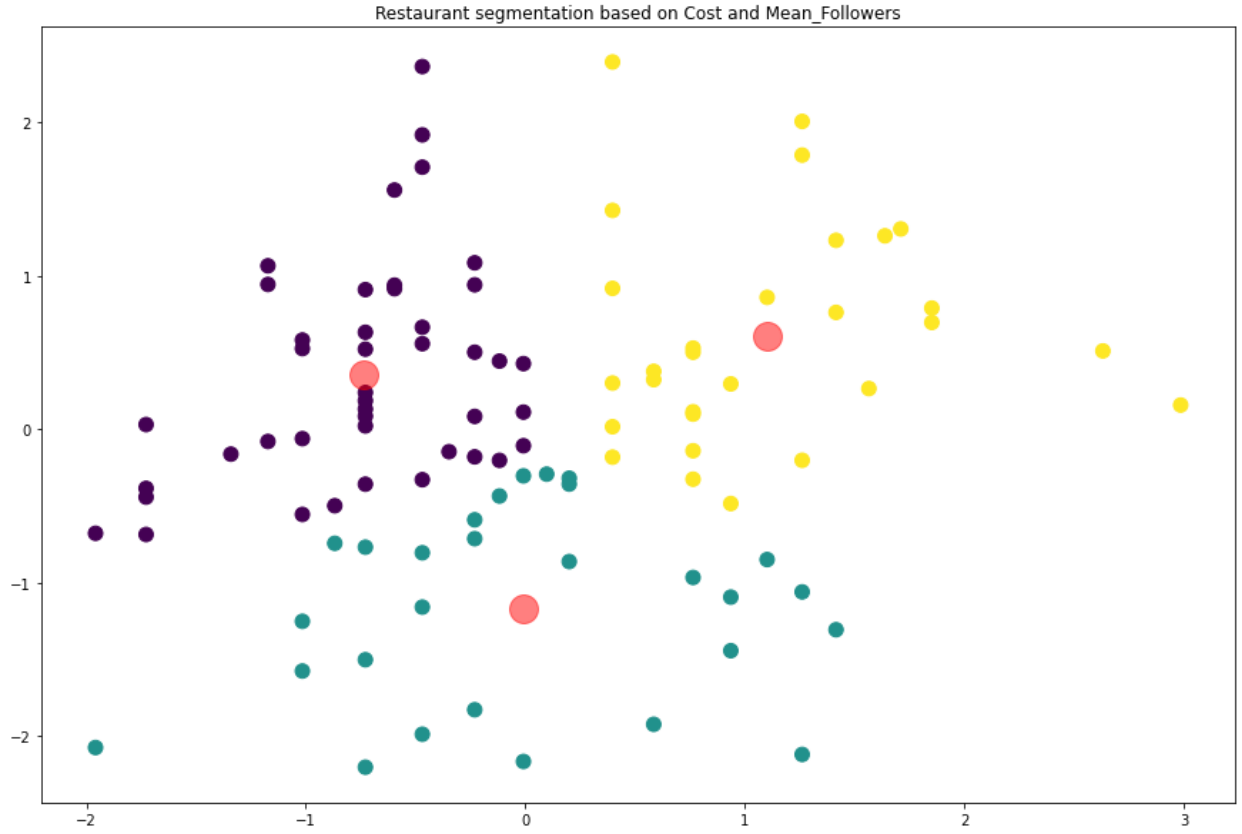


Cost and Mean_Rating

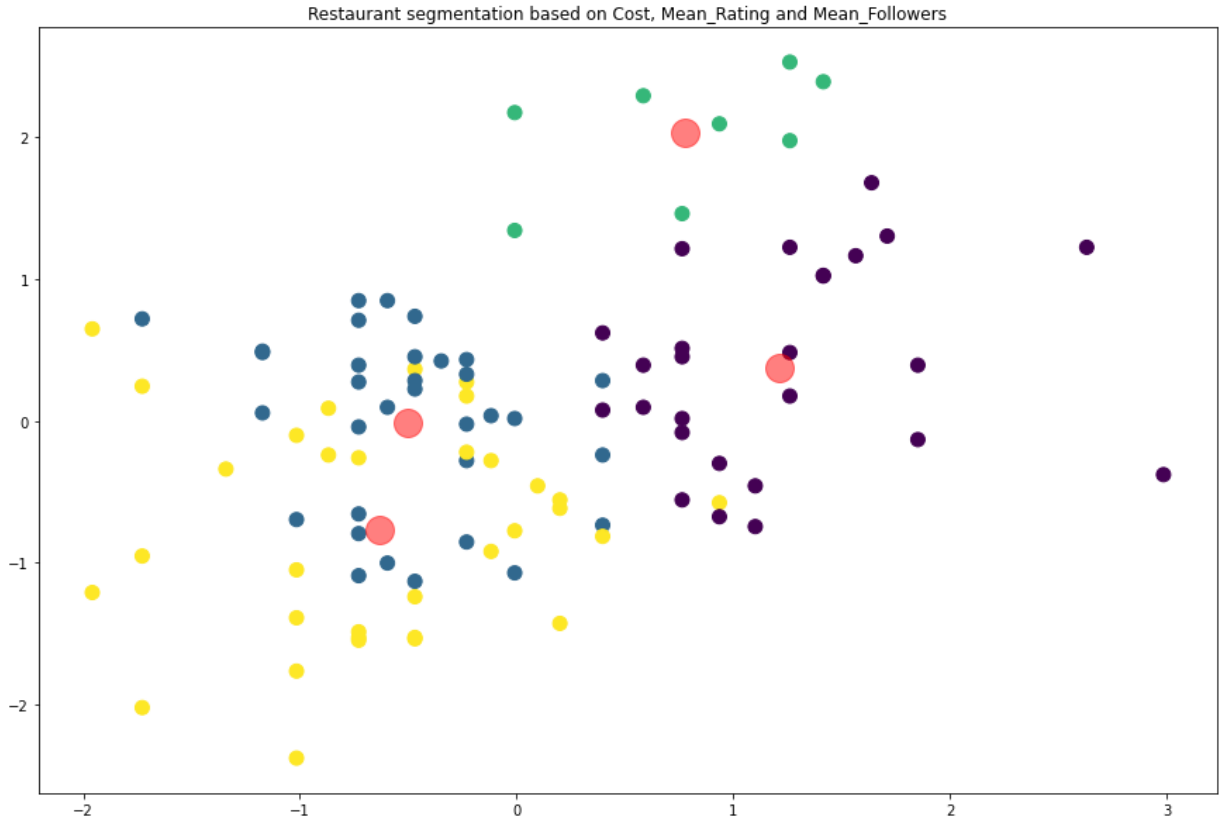


Mean_Rating and Mean_Followers

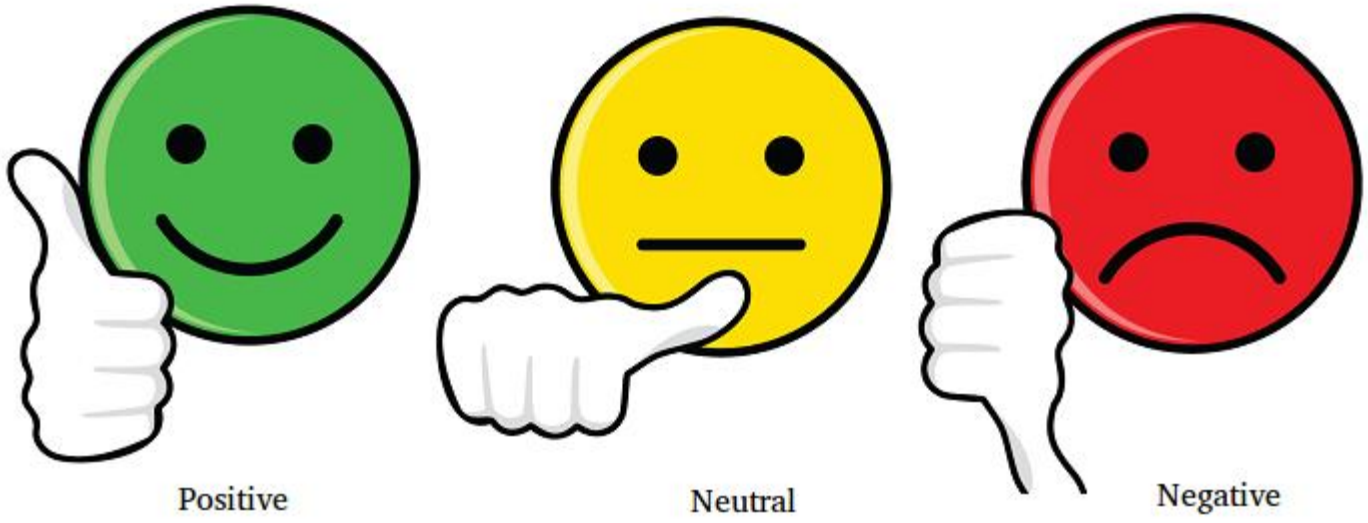
Clusters



Cost and Mean_Followers



Cost, Mean_Rating and Mean_Followers

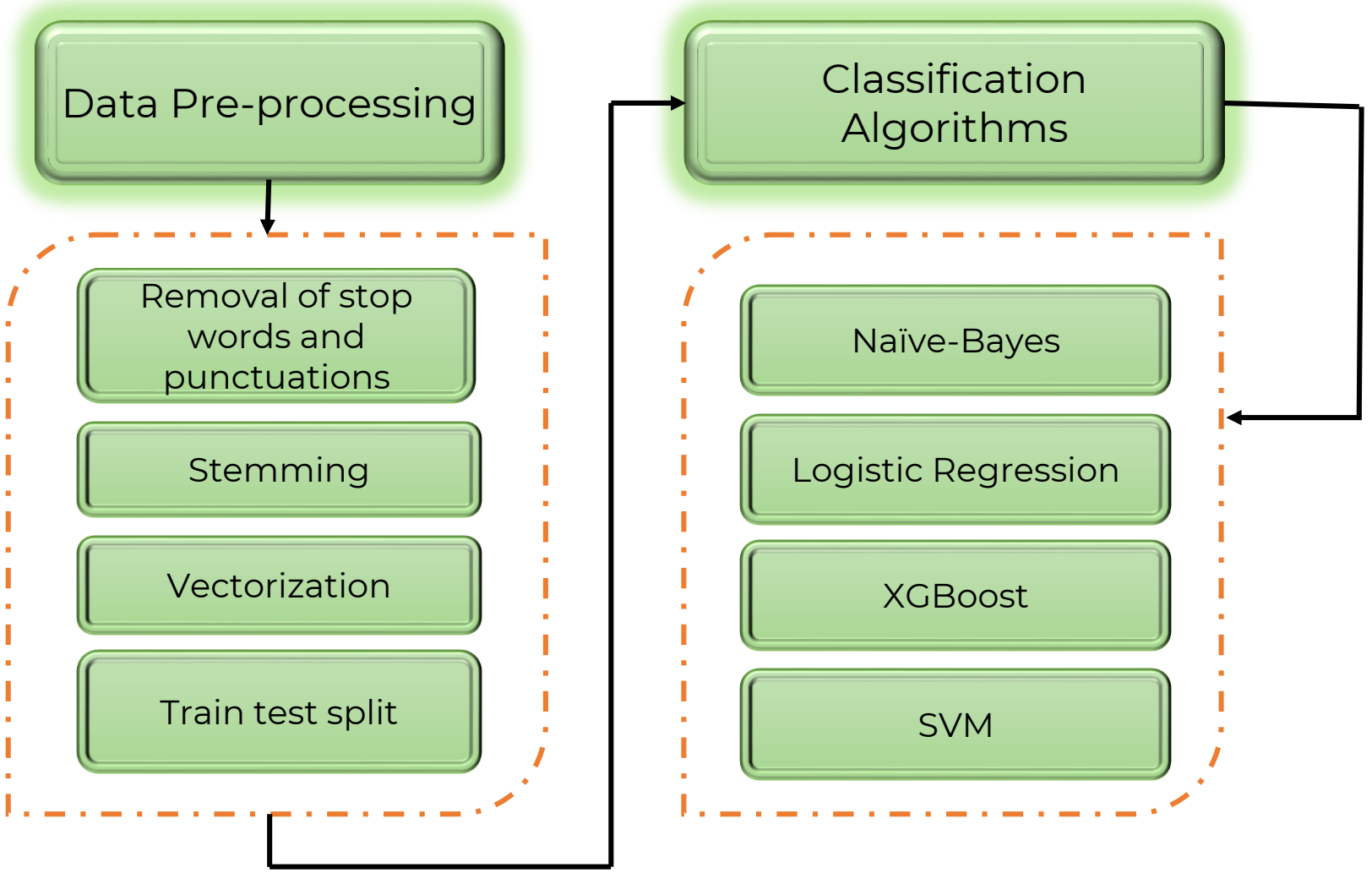


Sentiment Analysis

Sentiment Analysis

- Sentiment analysis (or opinion mining) is a **natural language processing (NLP)** technique used to determine whether data is positive, negative or neutral.
- Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

Methodology for sentiment analysis



Dataset for sentiment analysis

The ‘Rating’ column is grouped into two parts {rating 0 to 3 as negative(0) and rating 4 to 5 as positive(1) class}

The stop words and punctuations are removed from the ‘Review’ column followed by stemming and vectorization.

	Name	Review	Rating
0	Beyond Flavours	The ambience was good, food was quite good . h...	1
1	Beyond Flavours	Ambience is too good for a pleasant evening. S...	1
2	Beyond Flavours	A must try.. great food great ambience. Thnx f...	1
3	Beyond Flavours	Soumen das and Arun was a great guy. Only beca...	1
4	Beyond Flavours	Food is good.we ordered Kodi drumsticks and ba...	1
5	Beyond Flavours	Ambiance is good, service is good, food is aPr...	1
6	Beyond Flavours	Its a very nice place, ambience is different, ...	1
7	Beyond Flavours	Well after reading so many reviews finally vis...	1
8	Beyond Flavours	Excellent food , specially if you like spicy f...	1
9	Beyond Flavours	Came for the birthday treat of a close friend....	1

Results from Sentiment Analysis

SL No.	Model_Name	Train accuracy	Test accuracy
1	Naive-Bayes	0.84	0.83
2	Logistic Regression	0.88	0.86
3	XGBoost	0.86	0.84
4	SVM	0.92	0.87

- Logistic regression gives good accuracy without overfitting the data.
- SVM also gives good accuracy but it over fits the data.

Conclusions

- In the EDA we have explored popular cuisines & collections, costliest & cheapest restaurants, top rated & worst rated restaurants, most reviewed & most followed restaurants, and restaurants with most positive & most negative reviews.
- The optimal number of clusters by taking 2 variables at a time are either 3 or 4. And optimal number of clusters by taking all at a time is 4.
- For sentiment analysis logistic regression and SVM are the two most appropriate models with accuracy of nearly 87%.

Challenges

- Selecting the appropriate evaluation methods for clustering.
- Selecting the optimal number of clusters.
- Selecting the optimal number of features for sentiment analysis.


Thank you