

Capstone Project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

Team Members :-

Sidhartha Patel
Anirudh Upadhyay
Abdul Wadood
Ankit Patel
Bhanu Prakash
Md. Akram Raza

INTRODUCTION



The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city. So, this project focuses on analysing the Zomato restaurant data for a given city in India.

There are two sets of data. The first one provides the data on the restaurants and the other is about the reviews for these restaurants.

The main objective of the project is **clustering of zomato restaurants** into different segments. The Project also focuses on analyzing the **sentiments of the reviews** given by the customer in the data and to make some useful conclusion in the form of visualizations. The Analysis also solve some of the business cases that can directly help the customers finding the best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

DATA SUMMARY

Restaurant Dataset (105, 6)

Columns	Description
Name	Name of Restaurants
Links	URL Links of Restaurants
Cost	Per person estimated Cost of dining (string)
Collections	Tagging of Restaurants w.r.t. Zomato categories
Cuisines	Cuisines served by Restaurants (string)
Timings	Restaurant Timings

Reviews Dataset (10k, 7)

Columns	Description
Restaurant	Name of the Restaurant
Reviewer	Name of the Reviewer
Review	Review Text
Rating	Rating Provided by Reviewer
Meta Data	Reviewer Metadata - No. of Reviews and followers
Time	Date and Time of Review
Pictures	No. of pictures posted with review

DATA CLEANING AND ENGINEERING

Restaurant Dataset

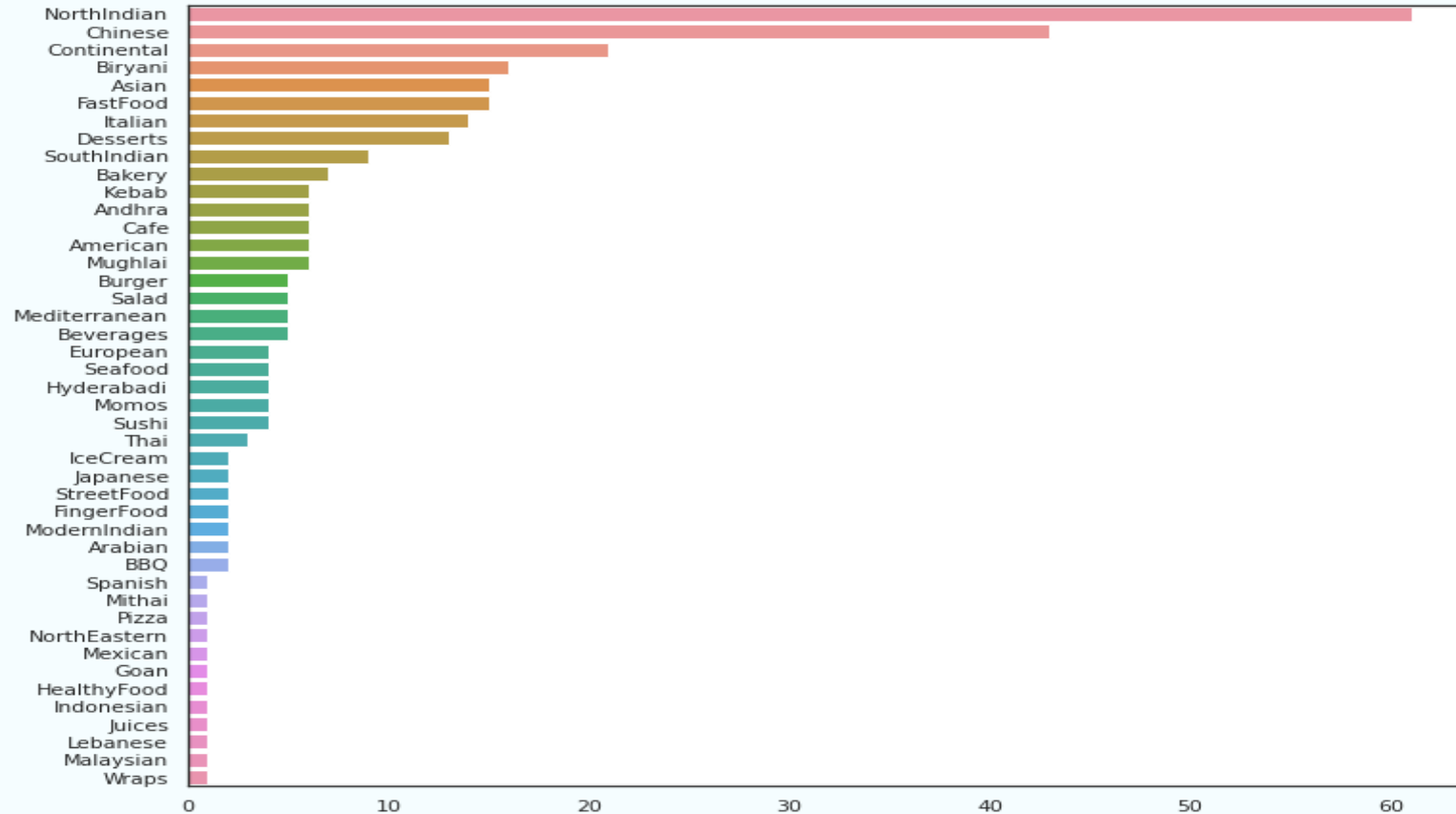
- Dropped the 'Collections' column as it had more than 50% missing values.
- 'Cuisines' column was a string containing different cuisines. It was converted to list of cuisines and was further one hot encoded using MultilabelBinarizer.
- 'Links' & 'Timings' Column are not relevant for the analysis, as 'Timings' were almost same for all restaurants.

Reviews Dataset

- Around 45 observations with Null values across multiple columns were removed.
- The cols 'Metadata' about reviewers were split into 2 cols : Reviews & Followers.
- Also, In the 'Ratings' cols, some values labelled as 'Like' were changed to numerical 4 for simplicity.
- **Finally, both Datasets were merged on 'Restaurant' and 'Name' columns.**

EXPLORATORY DATA ANALYSIS

Frequency chart of the cuisines



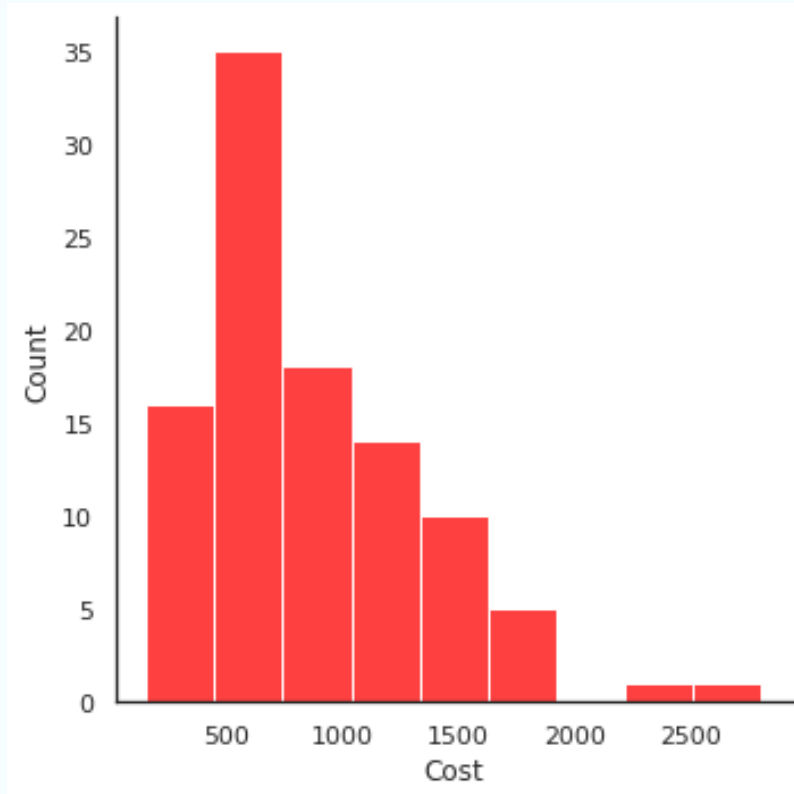
Among Cuisines, North Indian and Chinese are dominant across more than 50% of the restaurants.

Top 15 Restaurants Based on Avg Rating



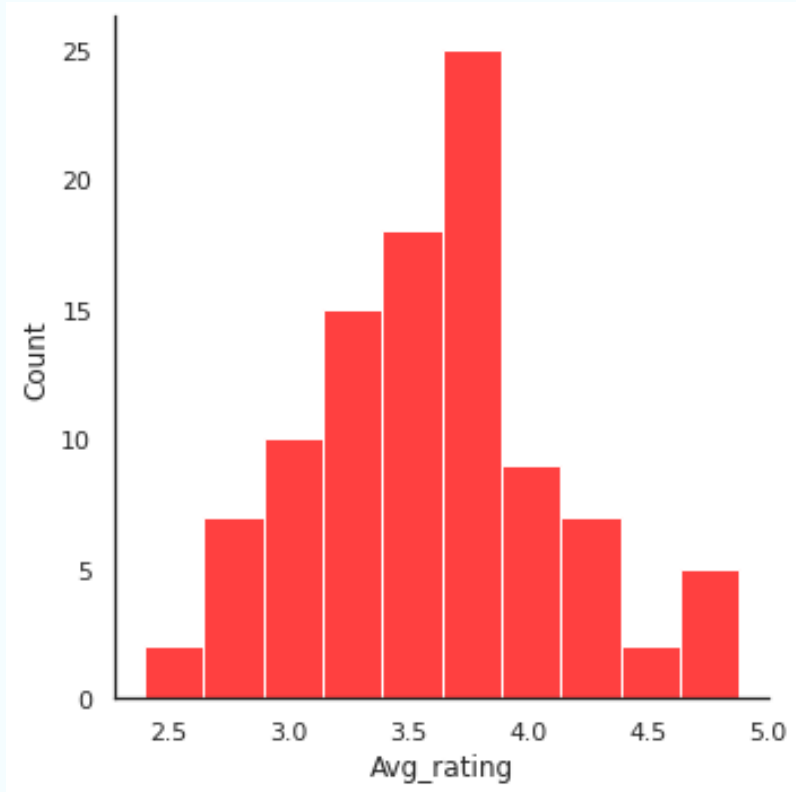
The top 20 restaurants are having ratings between 4.2 - 4.8.

Distribution of cost of food in the restaurants



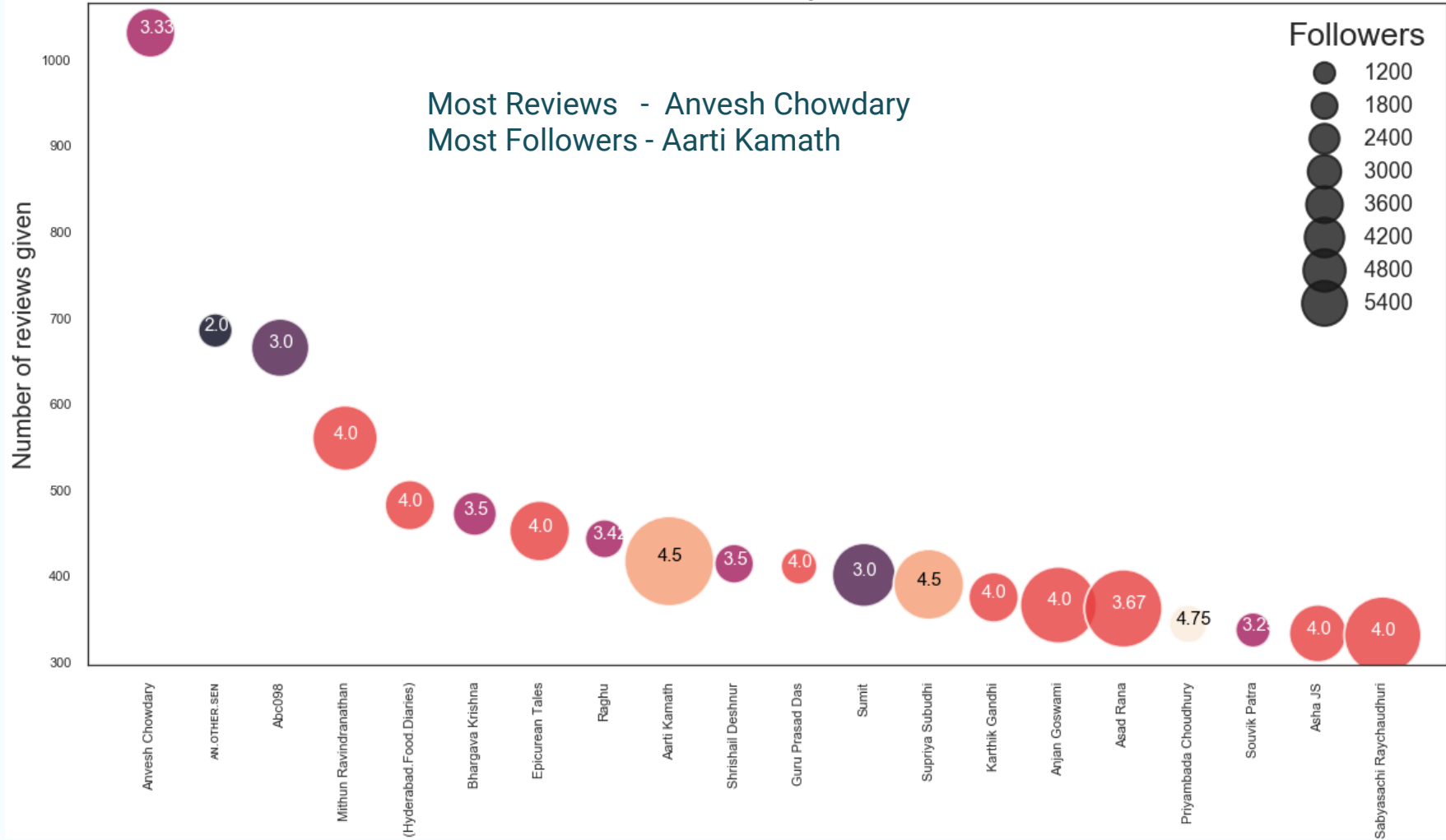
Most of the restaurant's cost is under Rs-1000 , but some restaurants are as costly as Rs-2500.

Distribution of Ratings across all restaurants



This variable is almost normally distributed with majority of the restaurants rated between 3 to 4.

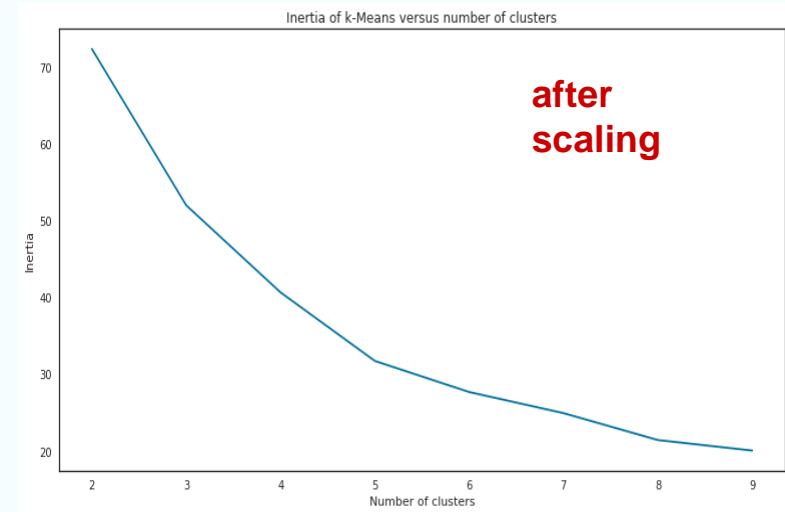
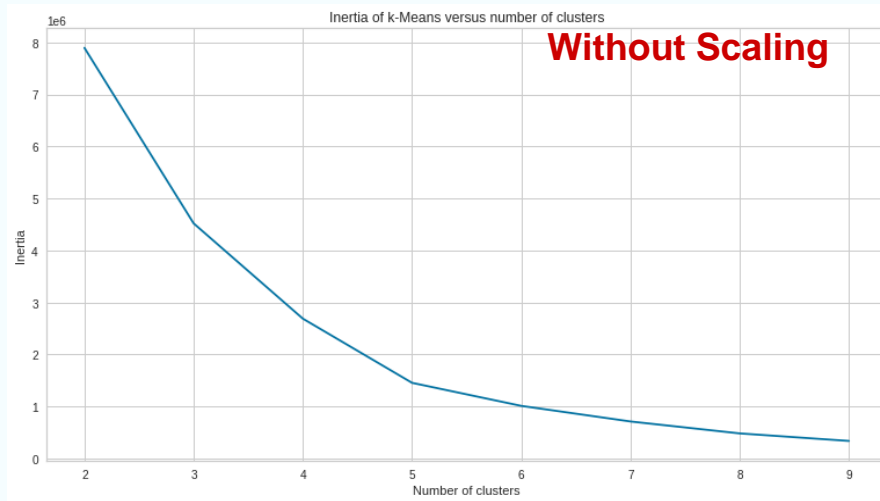
Critics Analysis



Modeling

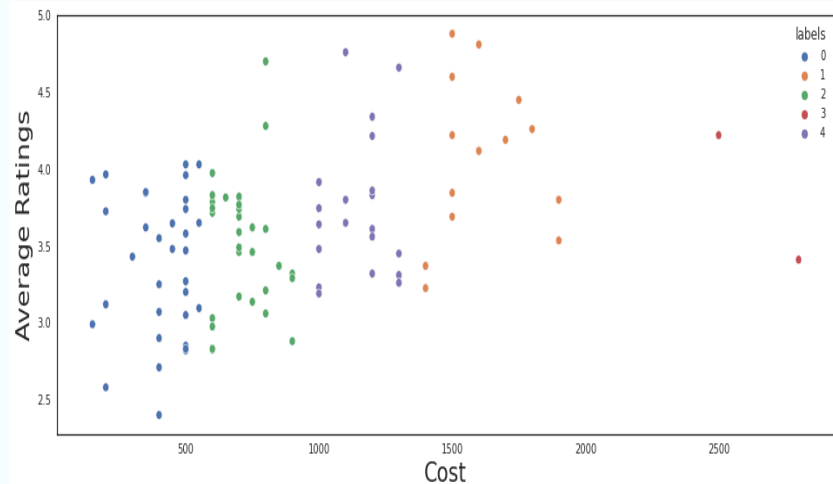
- Used KMeans Clustering to cluster the restaurants.
- Performed clustering with and without scaling to give weightage to different features.
- Applied PCA after scaling to visualise the clusters in 2-Dimensions.
- Used Elbow Method to find the optimal number of clusters ($k=5$).

Clustering Analysis

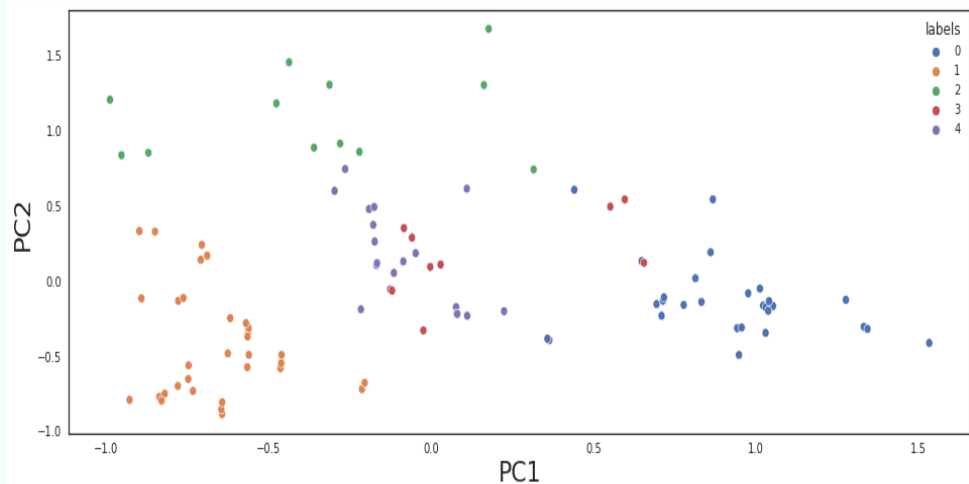


Visualization of Clusters

Before Scaling



After Scaling



Analysis on clusters

BEFORE SCALING				
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
NorthIndian	NorthIndian	Chinese	Italian	NorthIndian
Chinese	Continental	NorthIndian	Asian	Chinese
FastFood	Asian	Biryani	Continental	Italian

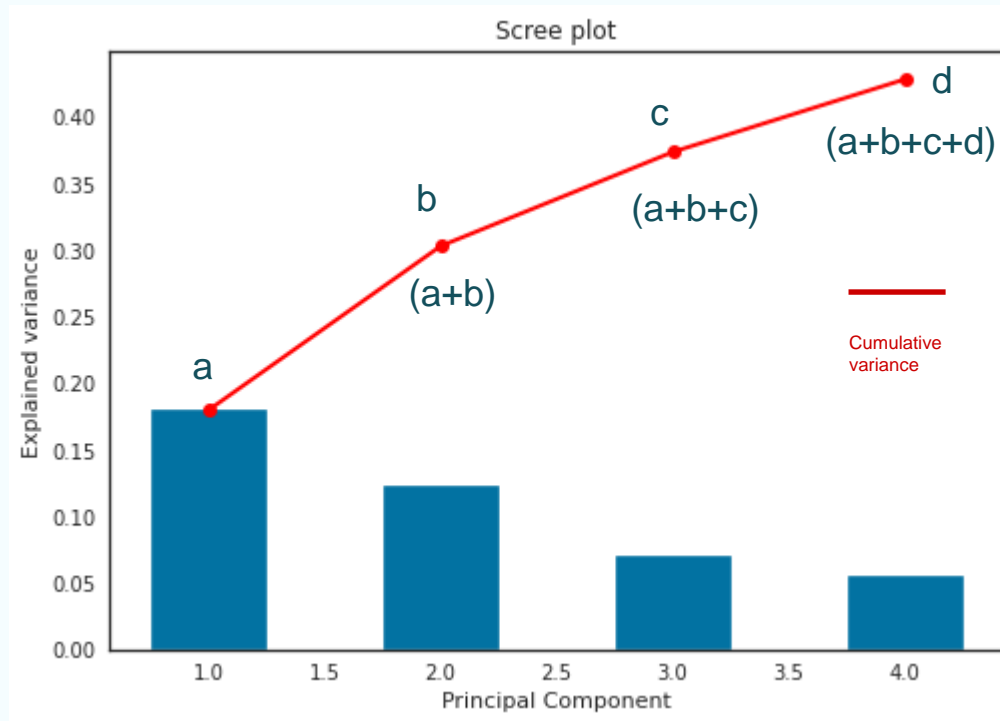
AFTER SCALING				
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Desserts	NorthIndian	NorthIndian	Asian	NorthIndian
FastFood	Chinese	Continental	Chinese	SouthIndian
Cafe	Biryani	Italian	Sushi	Mediterranean

Before Scaling

AVERAGE COST AND RATING		
	Avg_Cost	Avg_Rating
Cluster 0	406	3.37
Cluster 1	1610	4.07
Cluster 2	715	3.50
Cluster 3	2650	3.81
Cluster 4	1145	3.74

After Scaling

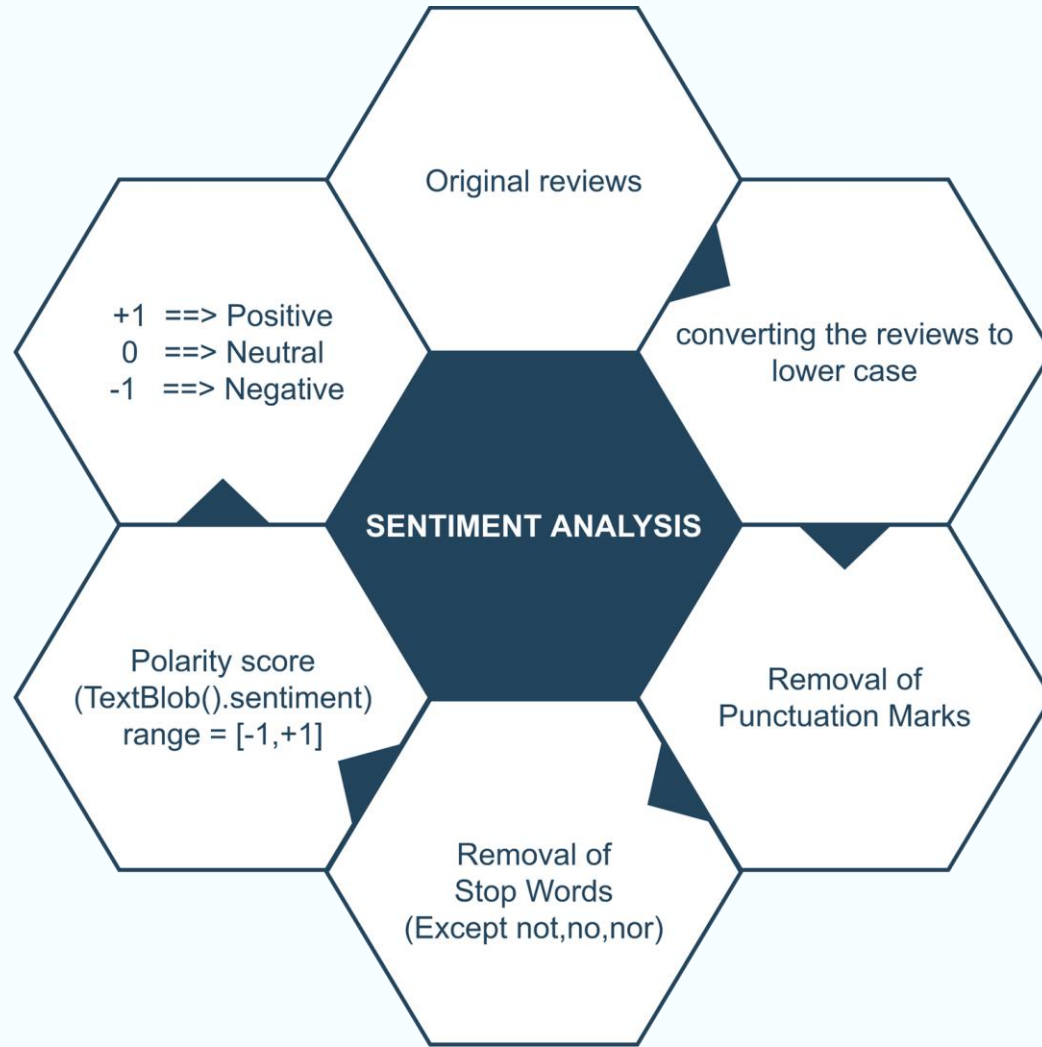
AVERAGE COST AND RATING		
	Avg_Cost	Avg_Rating
Cluster 0	553	3.58
Cluster 1	790	3.41
Cluster 2	1525	3.88
Cluster 3	950	3.65
Cluster 4	1005	3.72



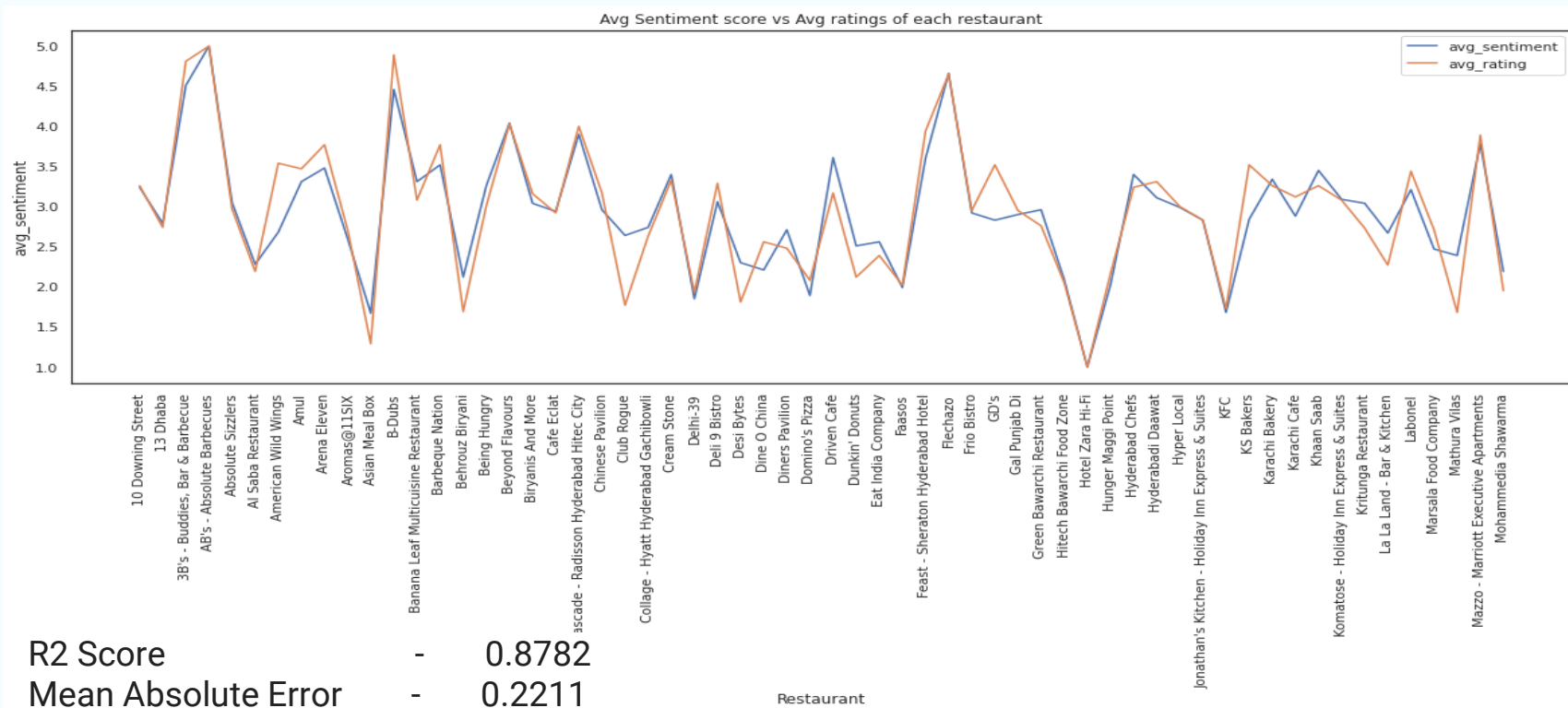
Sentiment analysis

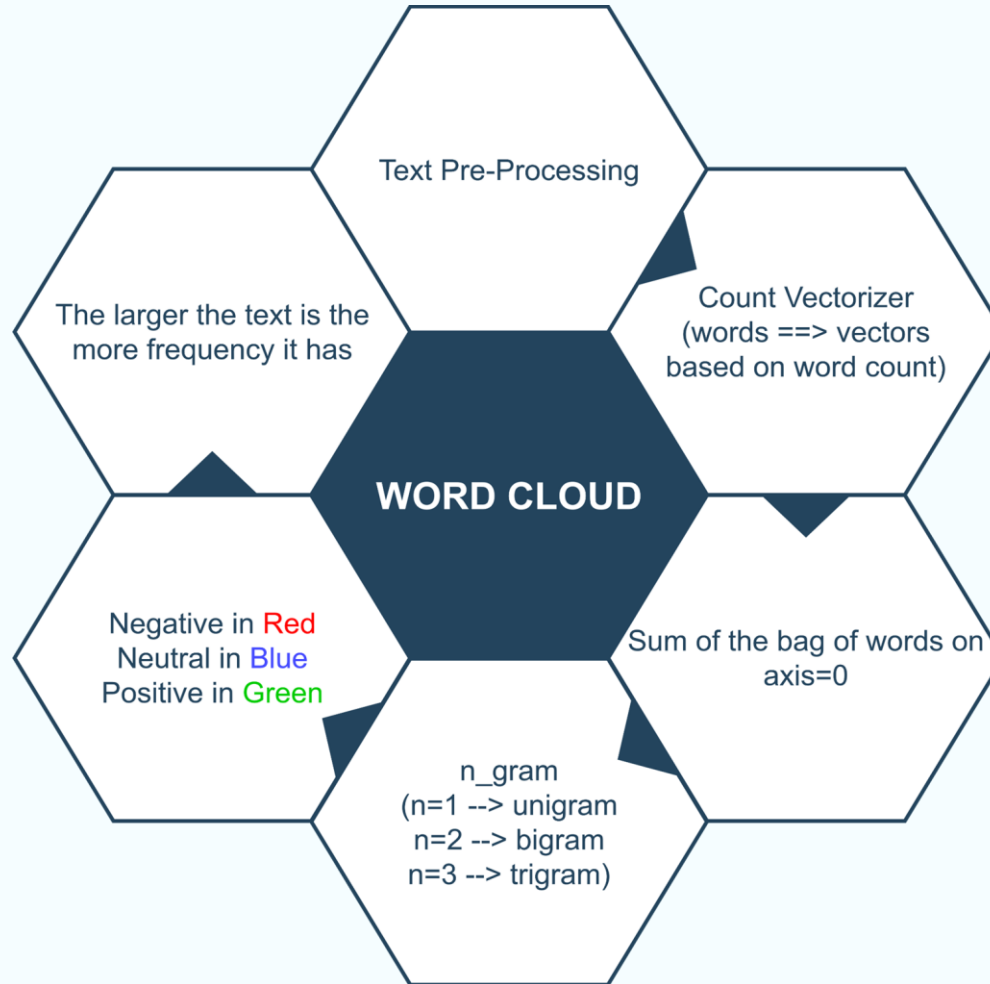
Libraries used :

- 1. CountVectorizer**
- 2. TextBlob**
- 3. WordCloud**



The sentiment score from our function and Rating from given dataset follow closely.





Positive Phrases

overall good experience food good service
 would definitely recommend nice place good
 food service ambience must go place
 food also good must visit place
 veg non veg must try place good place hangout
 ambience food service overall nice place
 would recommend place good place visit
 best place hangout indoor outdoor seating
 place good food good food good
 ambience good food paneer butter masala
 one best place food really good
 food excellent service food nice place
 service also good place hang friends
 nice place hangout starters main course
 place hangout friends
 good food great would love visit
 good ambience definitely recommend place food ambience service

CONCLUSION

Observations from Clustering

- Although Cuisines had impact on clustering, Cost was most dominant feature in explaining the clusters.

Observations from Sentiment Analysis

- Here we can observe that some of the critics gave low ratings but positive reviews.
- Also 3 star rating given by reviewers could be for a good experience for some and for others it may be for 'not upto the mark' experience.

Business Suggestion

- Depending on the business scenario we can choose the features such as cuisines and group restaurants by locality, type of restaurant, operating hours and not just cost.
- People can give low ratings because of delivery/ packing issues etc in addition to food quality.
- So it is better to store ratings for individual category like delivery, packaging, food etc.

THANK YOU