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 ones 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

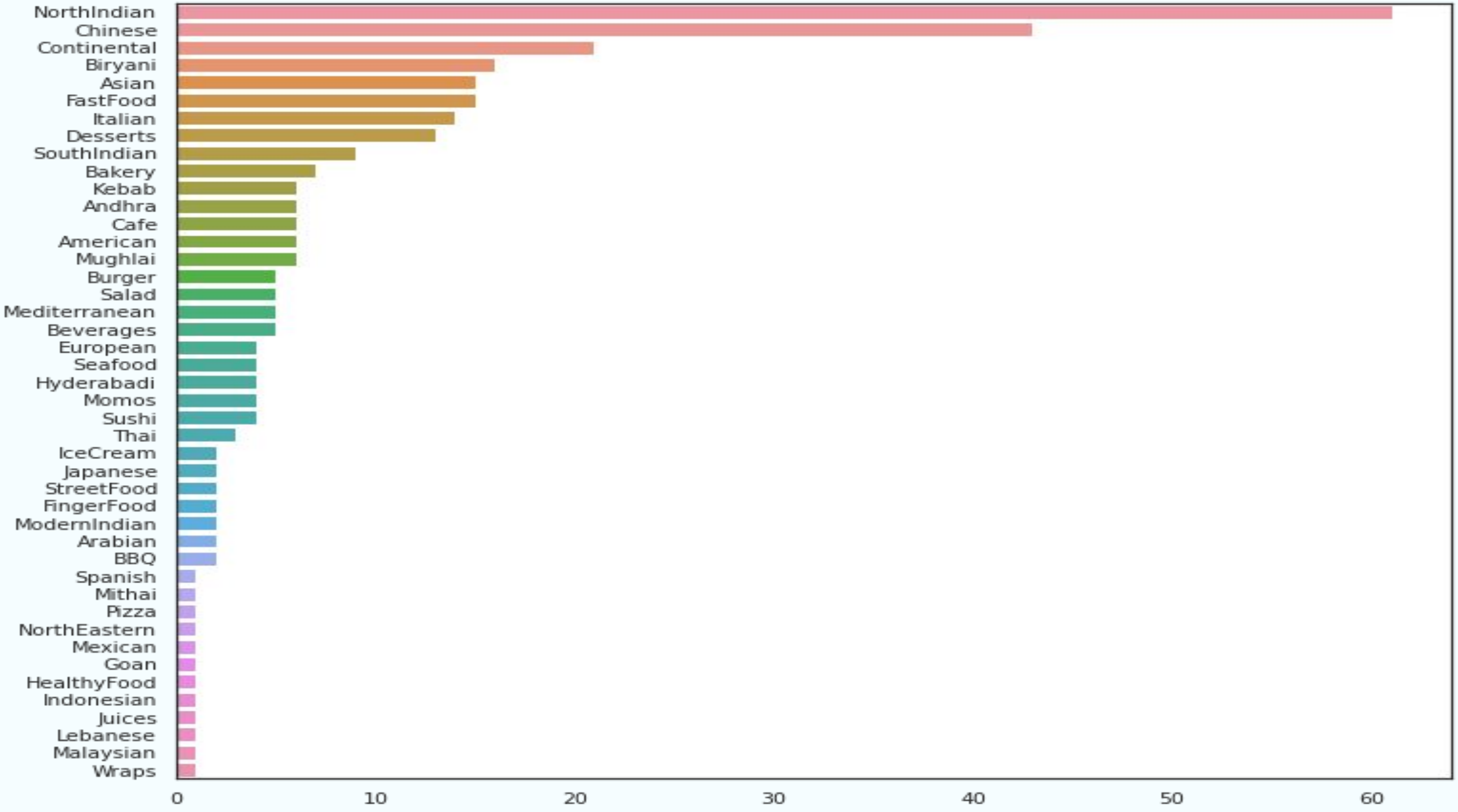
- Dropped the ‘Collections’ column as it had more than 50% missing values.
- ‘Cuisines’ column was a string containing different cuisines. It was converted to list of cuisines and was further one hot encoded using MultilabelBinarizer.
- ‘Links’ & ‘Timings’ Column are not relevant for the analysis, as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

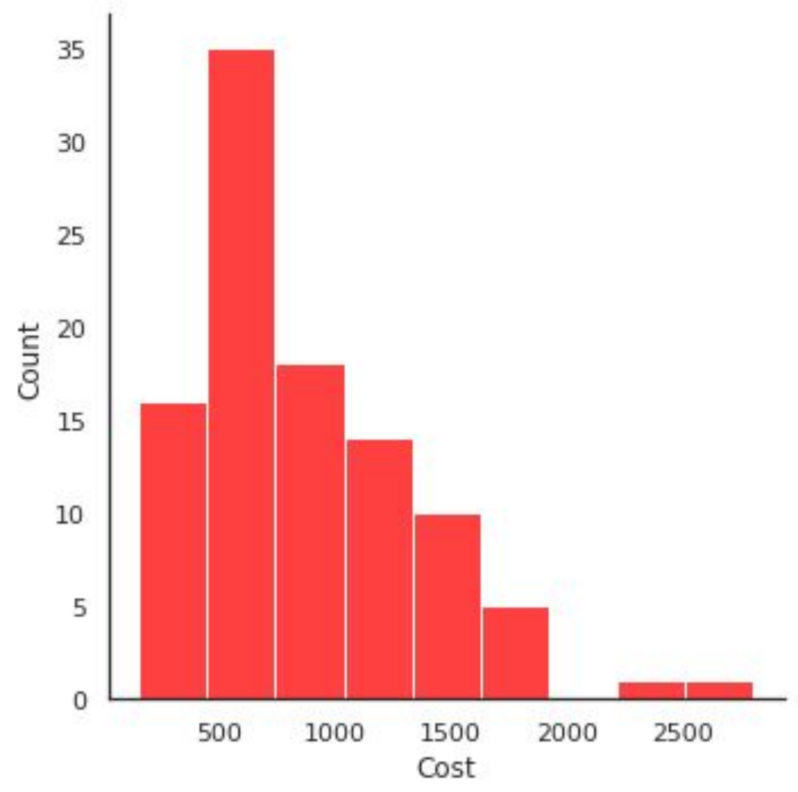
EXPLORATORY DATA ANALYSIS

Frequency chart of the cuisines



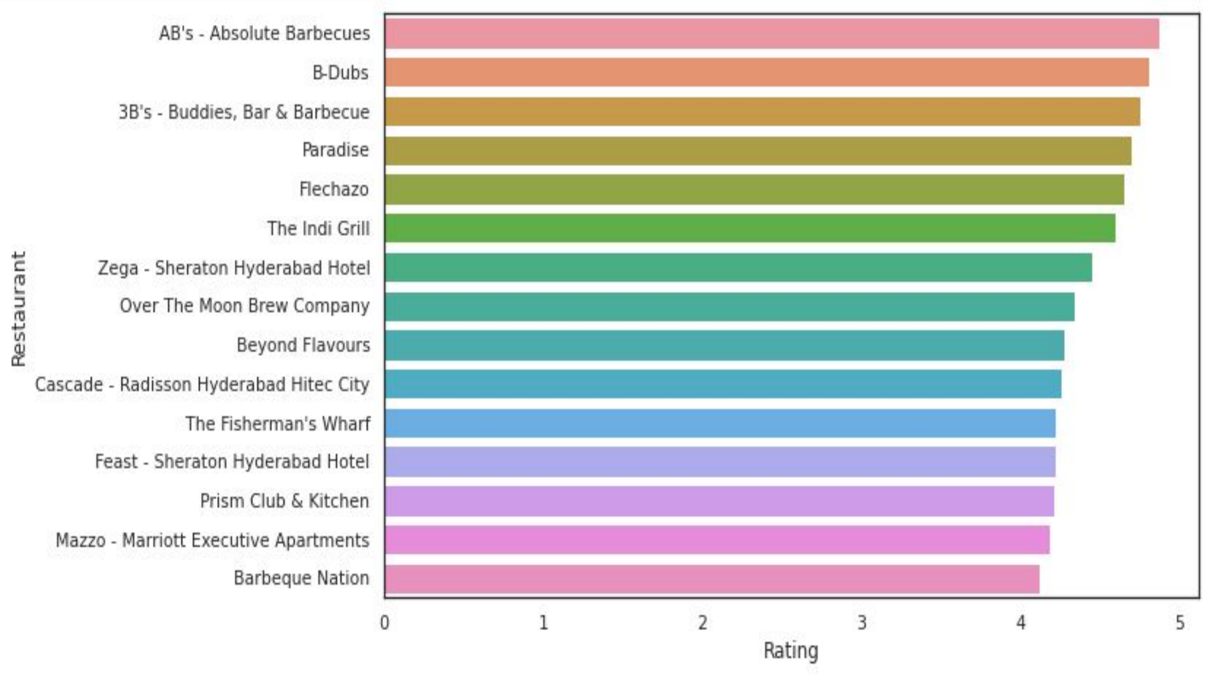
Among Cuisines, North Indian and Chinese are dominant across more than 50% of the restaurants.

Distribution of cost of food in the restaurants



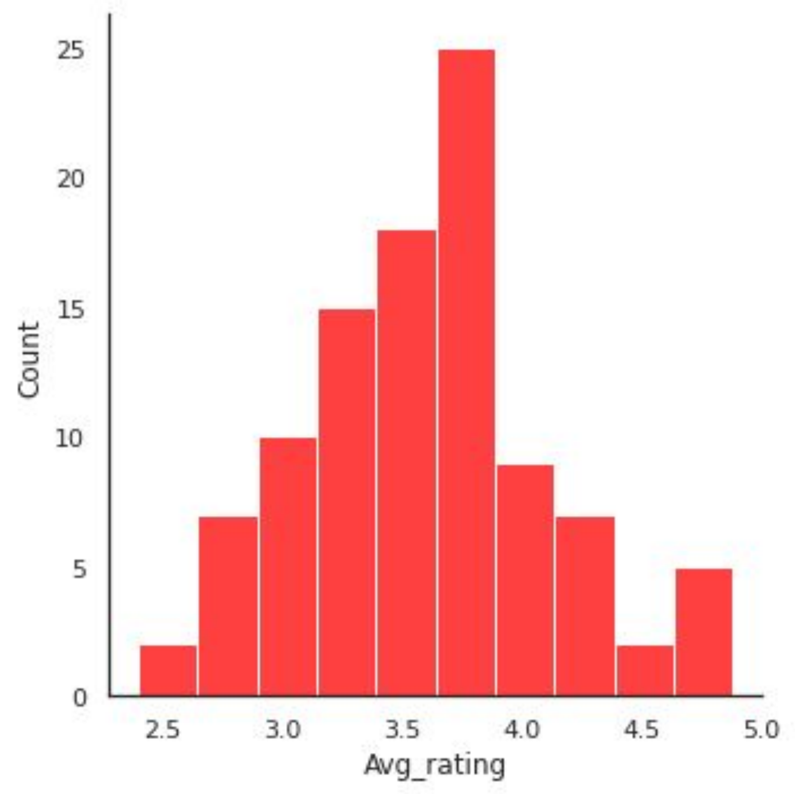
Most of the restaurant’s cost is under Rs-1000 , but some restaurants are as costly as Rs-2500.

Top 15 Restaurants Based on Avg Rating

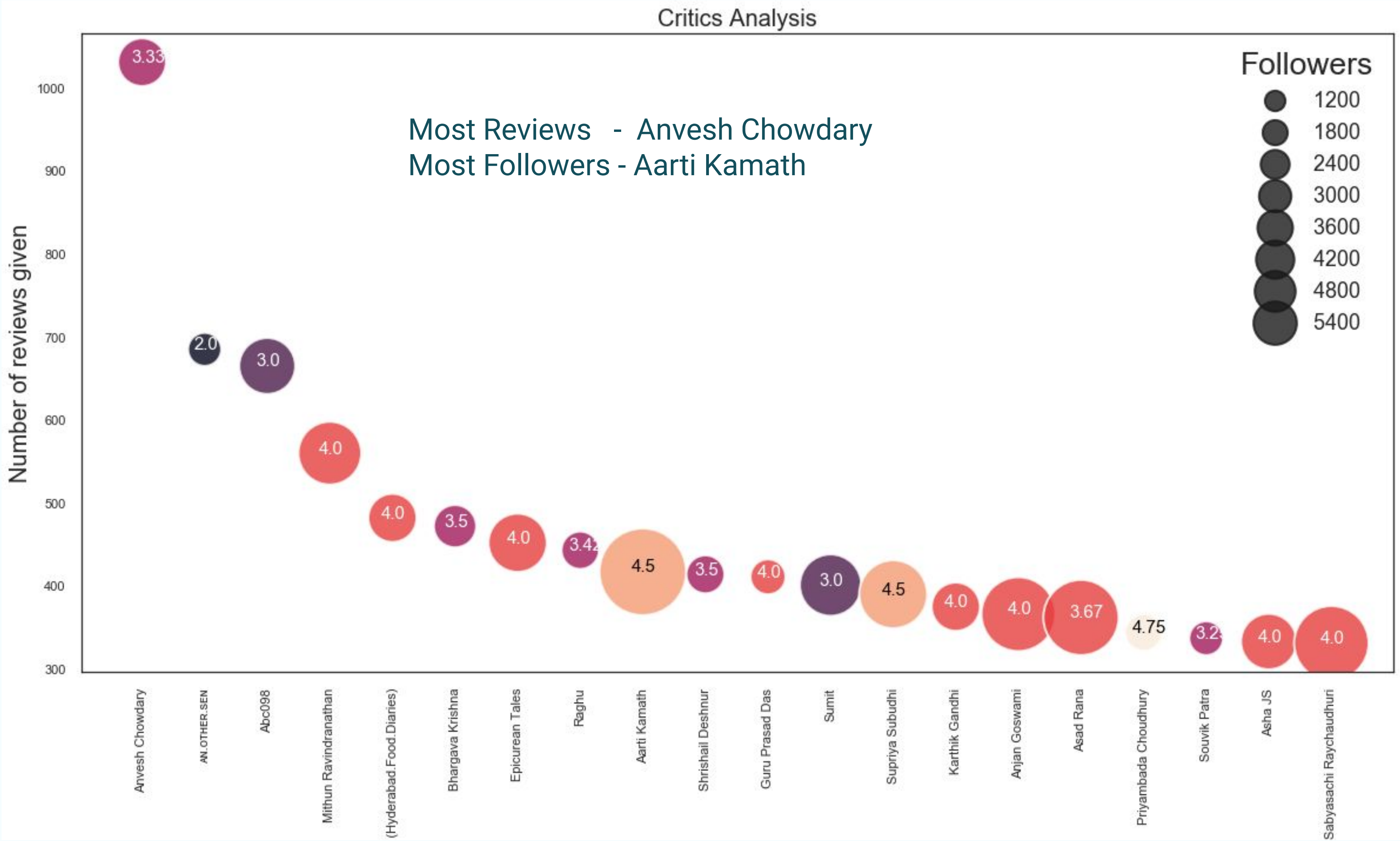


The top 20 restaurants are having ratings between 4.2 - 4.8.

Distribution of Ratings across all restaurants



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

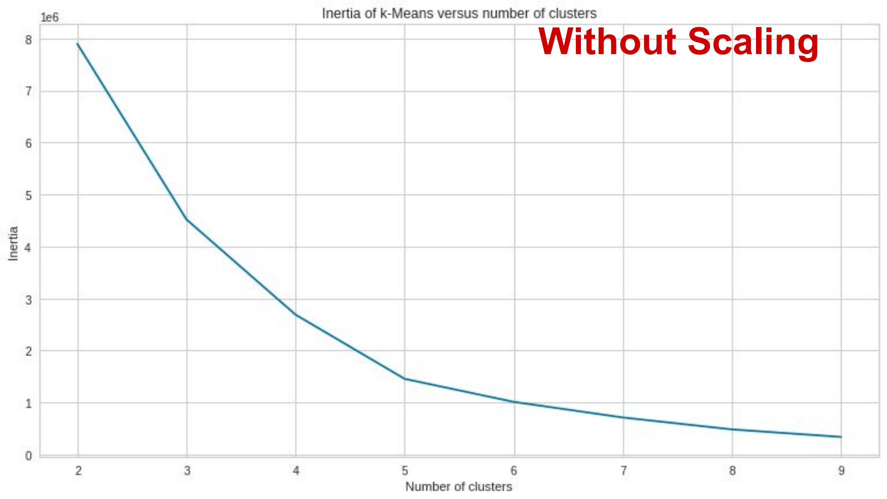


Most Reviews - Anvesh Chowdary
Most Followers - Aarti Kamath

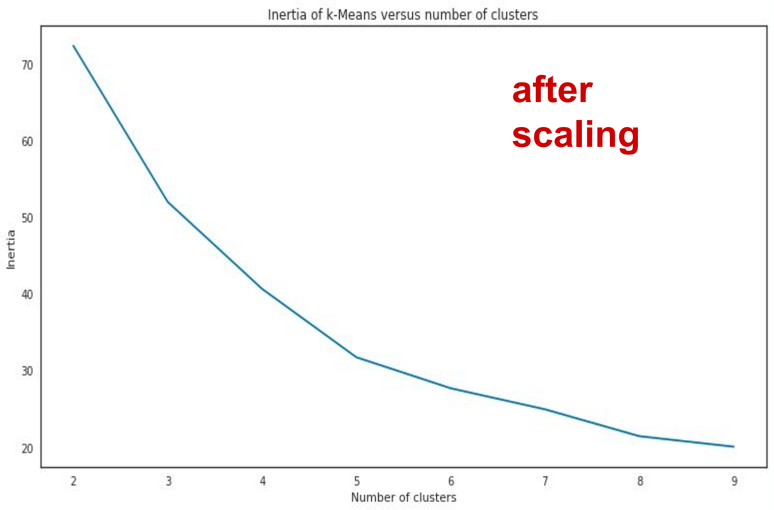
Modeling

- 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

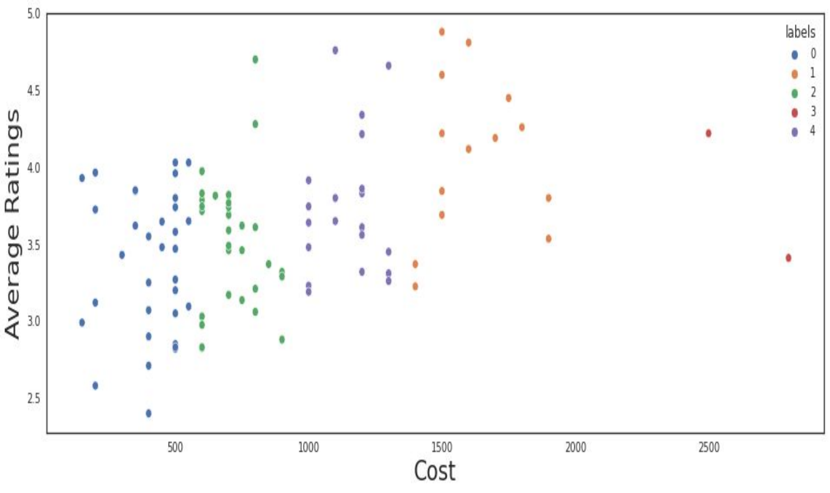


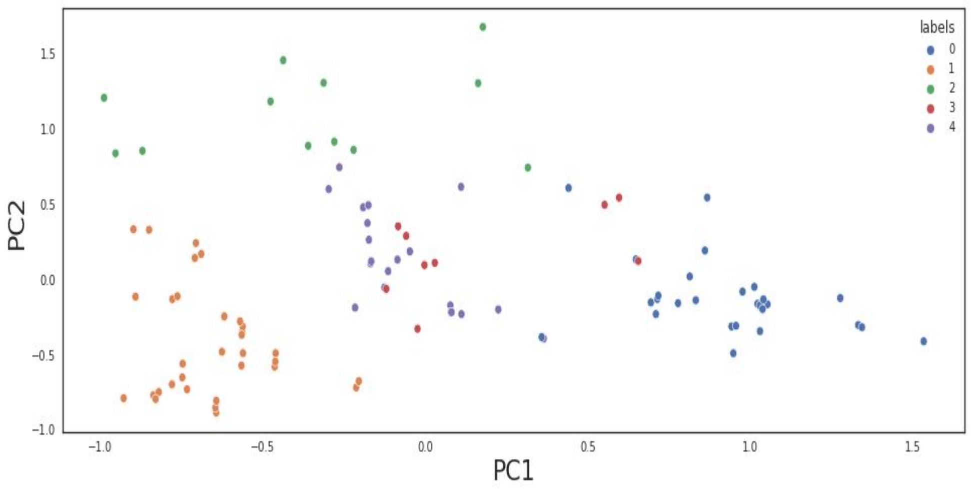
Without Scaling



after scaling

Visualisation of Clusters





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

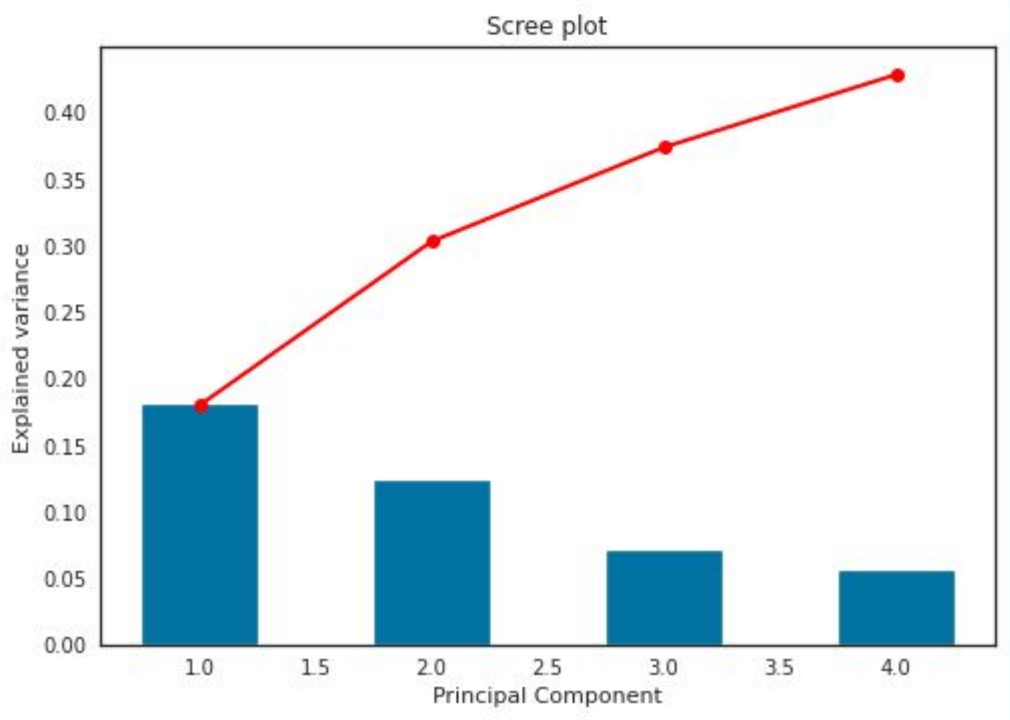
AFTER SCALING				
Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Desserts	NorthIndian	NorthIndian	Asian	NorthIndian
FastFood	Chinese	Continental	Chinese	SouthIndian
Cafe	Biryani	Italian	Sushi	Mediterranean

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

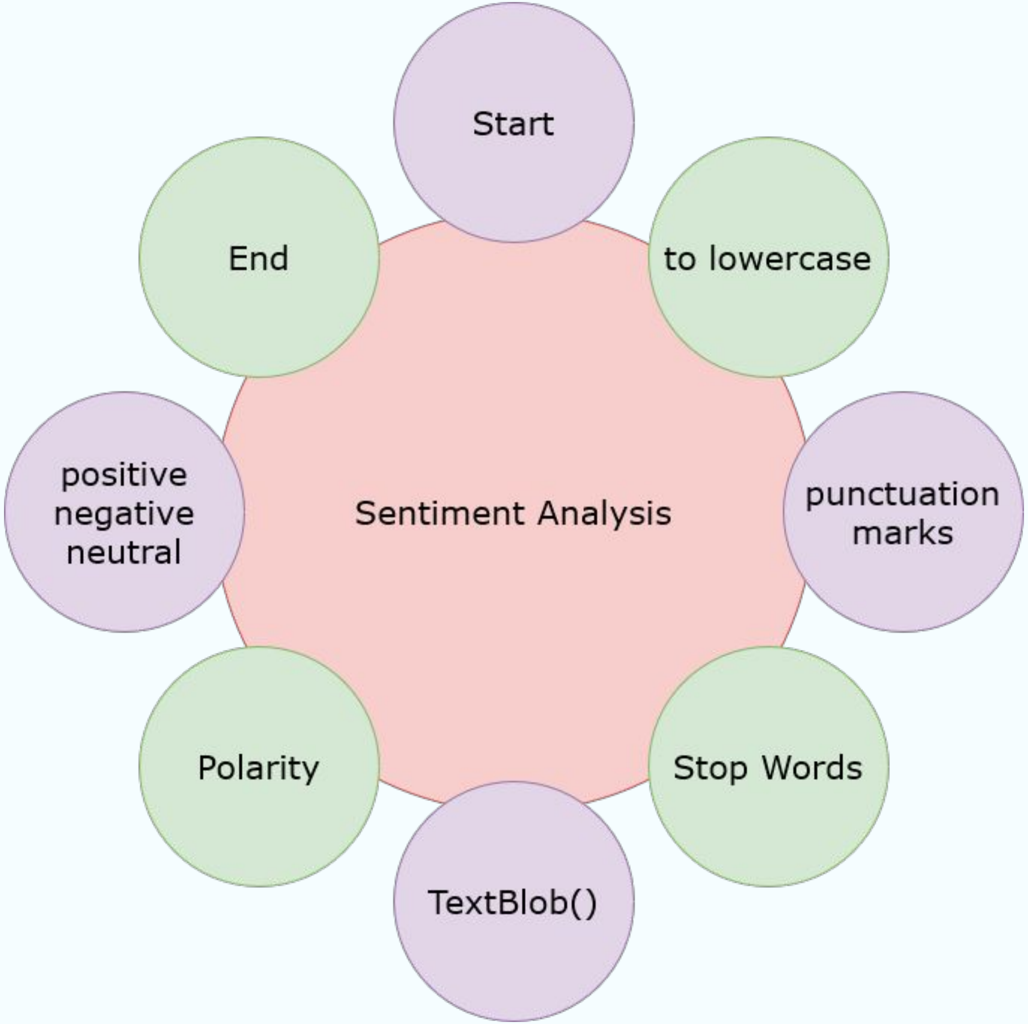
AVERAGE COST AND RATING		
	Avg_Cost	Avg_Rating
Cluster 0	553	3.58
Cluster 1	790	3.41
Cluster 2	1525	3.88
Cluster 3	950	3.65
Cluster 4	1005	3.72


Scree plot
Explained variance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Principal Component

Sentiment analysis

Libraries used :

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



Sentiment Analysis

Start

to lowercase

punctuation marks

Stop Words

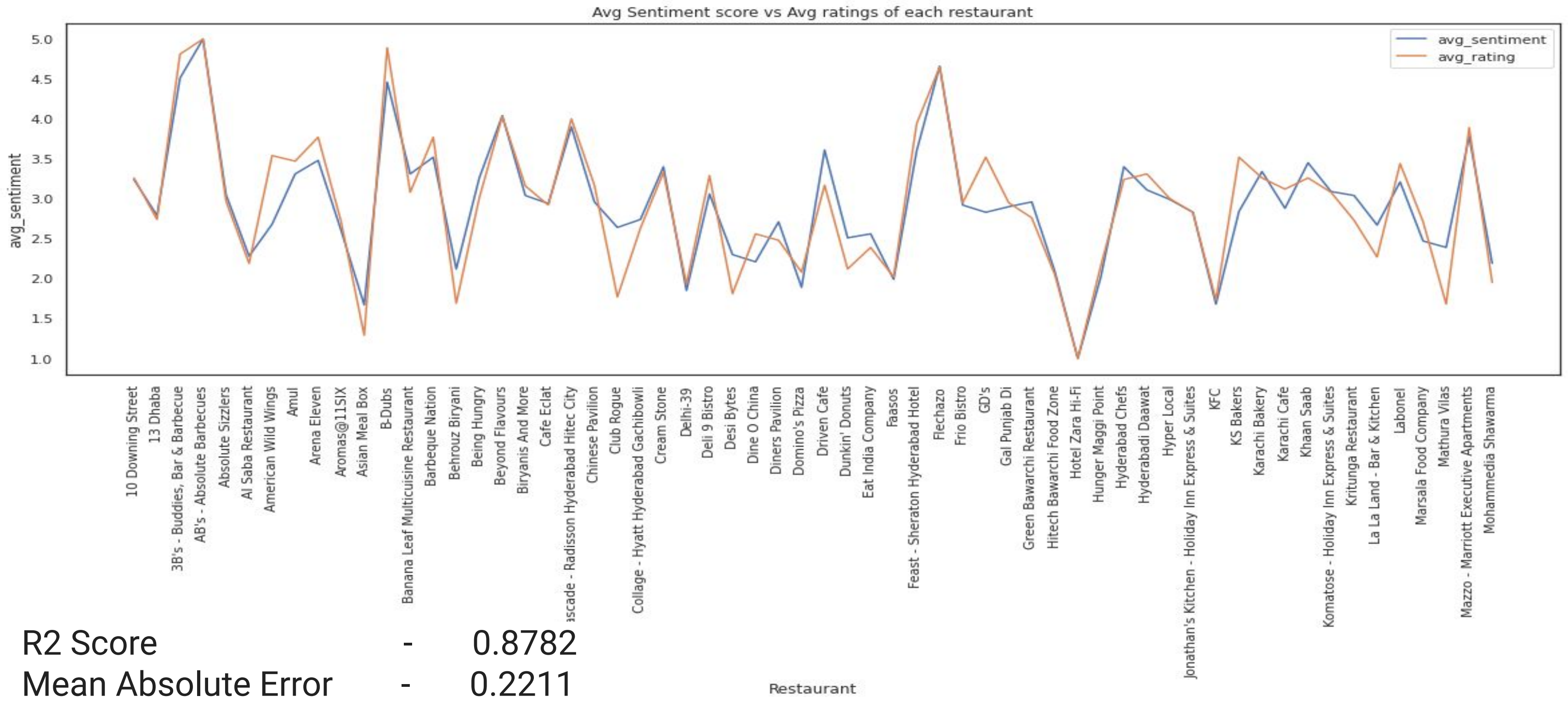
TextBlob()

Polarity

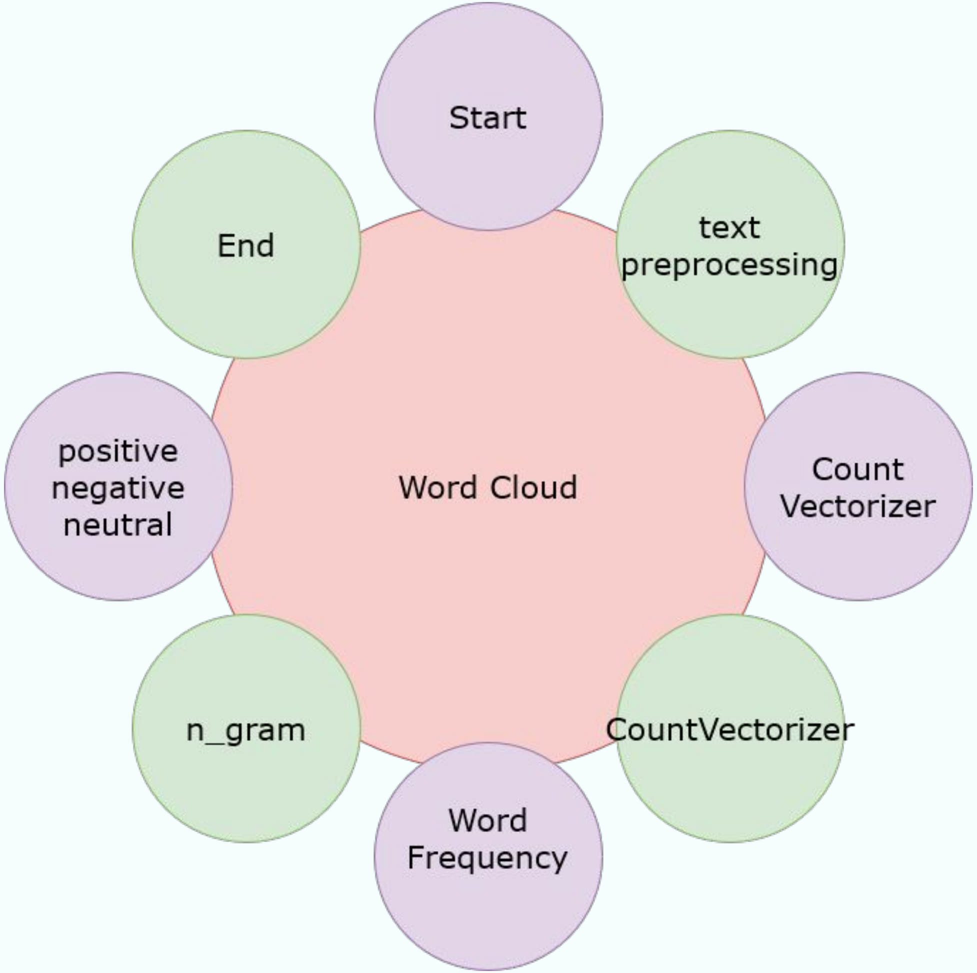
positive
negative
neutral

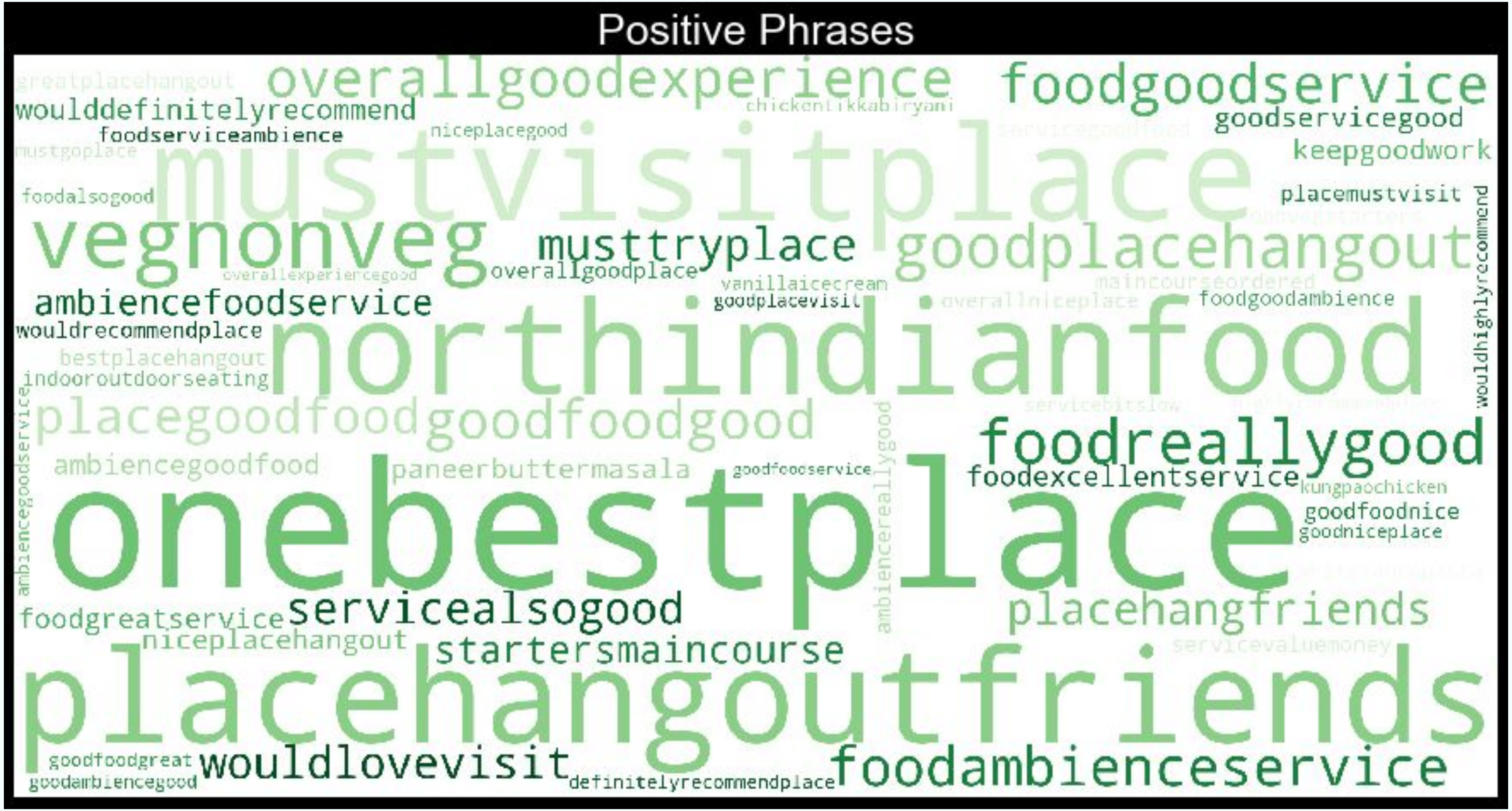
End

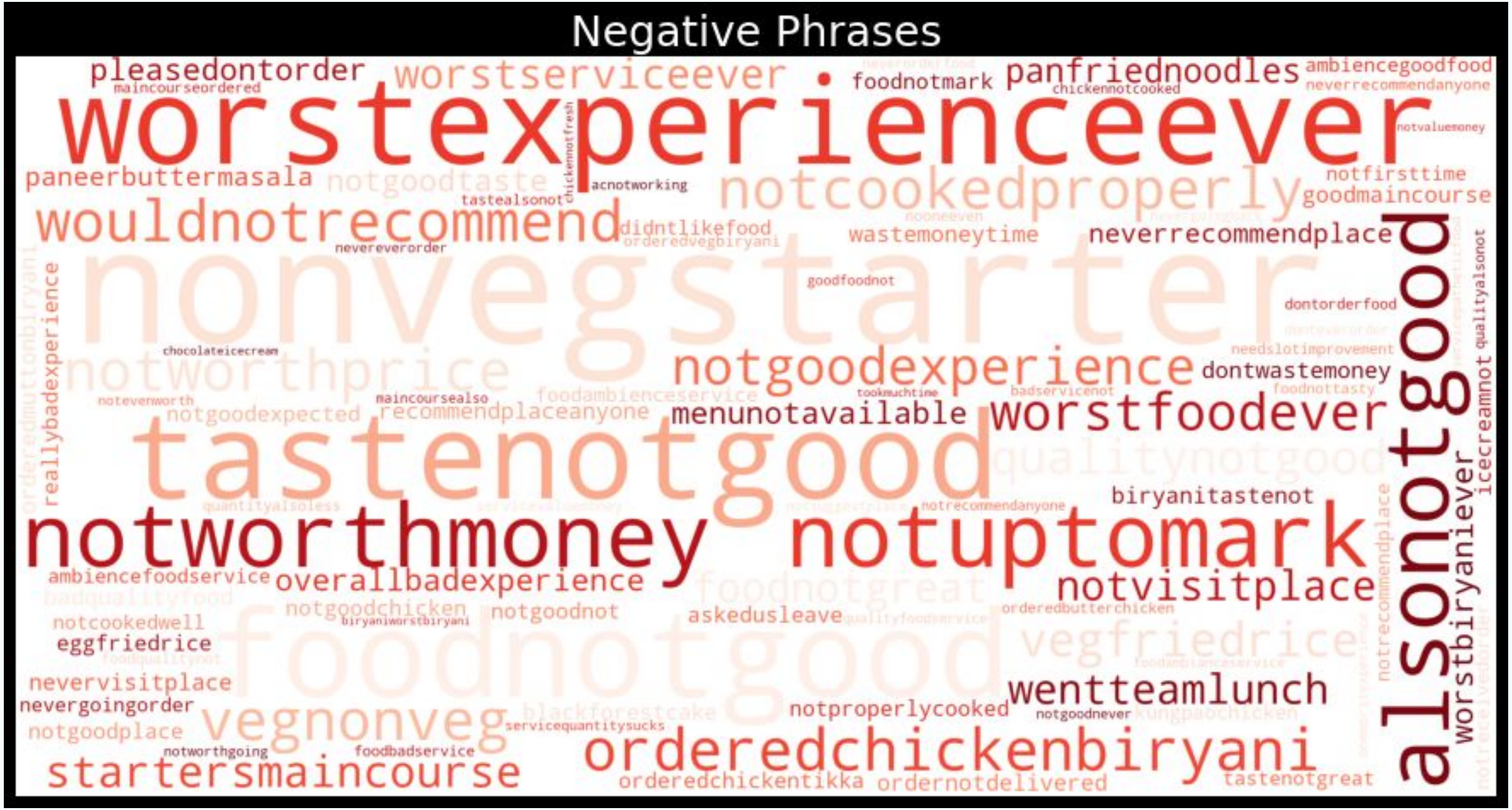
The sentiment score from our function and Rating from given dataset follow closely.



R2 Score - 0.8782
Mean Absolute Error - 0.2211


Start
text preprocessing
Count Vectorizer
CountVectorizer
Word Frequency
n_gram
positive negative neutral
End
Word Cloud


Positive Phrases
greatplacehangout
overallgoodexperience
foodgoodservice
wouldddefinitelyrecommend
chickentikkabiryani
foodserviceambience
niceplacegood
servicegoodfood
goodservicegood
mustgoplace
mustvisitplace
keepgoodwork
foodalsogood
placemustvisit
vegnonveg
musttryplace
goodplacehangout
overallexperiencegood
overallgoodplace
vanillaicecream
maincourseordered
ambiencefoodservice
goodplacevisit
overallniceplace
foodgoodambience
wouldrecommendplace
northindianfood
wouldhighlyrecommend
bestplacehangout
indooroutdoorseating
placegoodfood
goodfoodgood
servicebitslow
ambiencegoodservice
ambiencereallygood
foodreallygood
ambiencegoodfood
paneerbuttermasala
goodfoodservice
foodexcellentservice
kungpaochicken
onebestplace
goodfoodnice
goodniceplace
foodgreatservice
servicealsogood
placehangfriends
niceplacehangout
startersmaincourse
servicevaluemoney
placehangoutfriends
goodfoodgreat
wouldlovevisit
goodambiencegood
definitelyrecommendplace
foodambienceservice


Negative Phrases
pleasedontorder
maincourseordered
worstserviceever
foodnotmark
panfriednoodles
ambiencegoodfood
chickennotcooked
neverrecommendanyone
worstexperienceever
notvaluemoney
chickennotfresh
paneerbuttermasala
notgoodtaste
tastealsonot
acnotworking
notcookedproperly
notfirsttime
goodmaincourse
wouldnotrecommend
didntlikefood
orderedvegbiryani
nevereverorder
wastemoneytime
neverrecommendplace
nonvegstarter
goodfoodnot
dontorderfood
alsonotgood
qualityalsonot
reallybadexperience
orderedmuttonbiryani
chocolateicecream
notworthprice
needslotimprovement
notgoodexperience
dontwastemoney
notevenworth
maincoursealso
foodambienceservice
tookmuchtime
badservicenot
foodnottasty
notgoodexpected
recommendplaceanyone
menunotavailable
worstfoodever
icecreamnot
tastenotgood
qualitynotgood
worstbiryaniever
quantityalsoless
notrecommendanyone
biryanitastenot
notworthmoney
notuptomark
notrecommendplace
ambiencefoodservice
overallbadexperience
foodnotgreat
notvisitplace
badqualityfood
notgoodchicken
notgoodnot
askedusleave
qualityfoodservice
orderedbutterchicken
notcookedwell
biryaniworstbiryani
eggfriedrice
foodnotgood
vegfriedrice
foodqualitynot
foodambienceservice
nevervisitplace
wentteamlunch
nevergoingorder
vegnonveg
blackforestcake
notproperlycooked
notgoodnever
kungpaochicken
notgoodplace
servicequantitysucks
notworthgoing
foodbadservice
notreceivedorder
startersmaincourse
orderedchickenbiryani
orderedchickentikka
ordernotdelivered
tastenotgreat

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.