Capstone Project

Unsupervised ML
ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

Subham Choudhary

Contents

- ❑ Introduction
- ❑ Problem Statement
- ❑ Data Overview
- ❑ Data cleaning and processing
- ❑ Exploratory data analysis
- ❑ Clustering
- ❑ Sentiment analysis
- ❑ Modelling
- ❑ Conclusion

Introduction

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008.

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity.

Problem Statement

The Project focuses on Customers and Company, you have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of Visualizations. Also, cluster the zomato restaurants into different segments.

This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis

Data could be used for sentiment analysis. Also the metadata of reviewers can be used for identifying the critics in the industry.

Data Overview

The problem statement here has two datasets for us to work on:

- Zomato Restaurant Names and Metadata: This dataset was used for clustering.
  - Name: Name of Restaurants
  - Links: URL Links of Restaurants
  - Cost: Per person estimated Cost of dining
  - Collection: Tagging of Restaurants w.r.t. Zomato categories
  - Cuisines: Cuisines served by Restaurants
  - Timings: Restaurant Timings

Data Overview (continued)

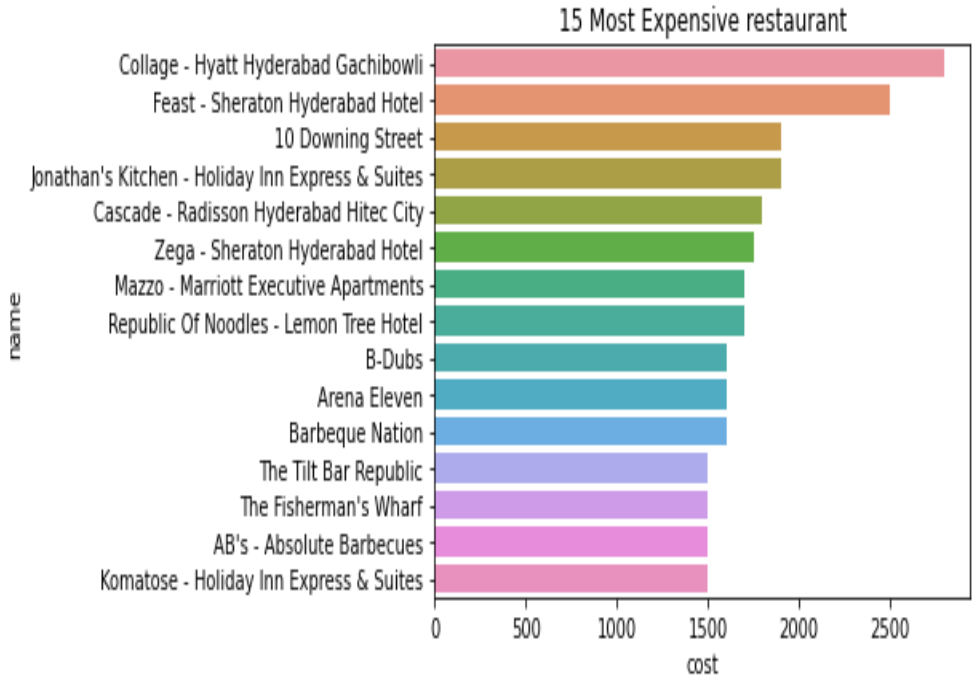
- Zomato Restaurant Reviews: This dataset was merged with Names and Metadata and then used for sentiment analysis.
  - Restaurant: Name of the Restaurant
  - Reviewer: Name of the Reviewer
  - Review: Review Text
  - Rating: Rating Provided by Reviewer
  - MetaData: Reviewer Metadata - No. of Reviews and followers
  - Time: Date and Time of Review
  - Pictures: No. of pictures posted with review

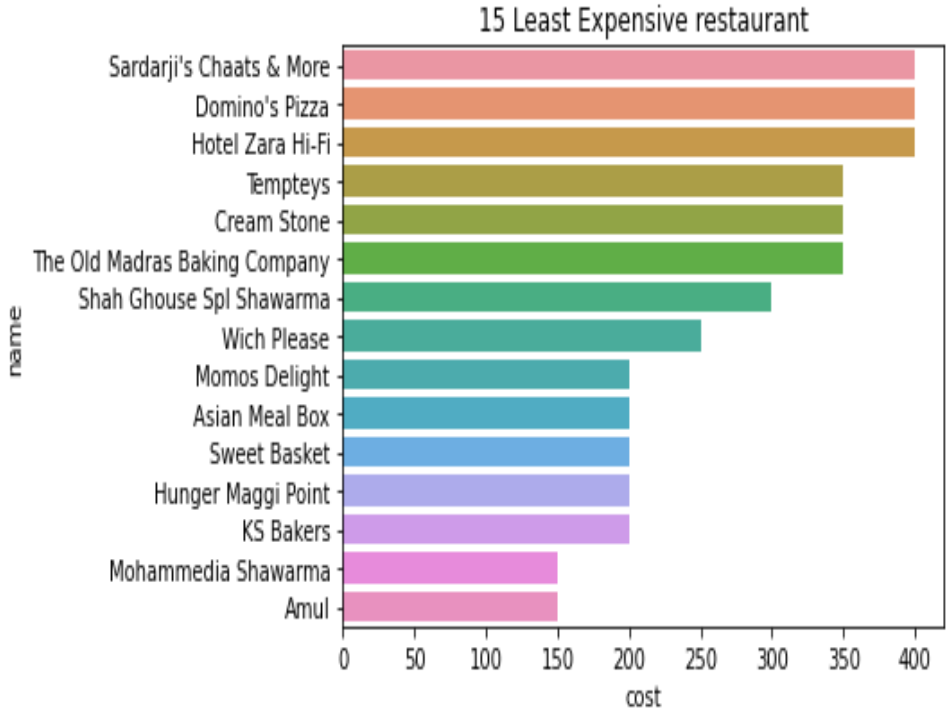
Data cleaning and processing

- Null and duplicate values removed from both dataset.
- Columns is converted to lowercase.
- Cost column is changed to numeric datatype.
- Some features like links, collections and timings as it didn’t contribute much to cluster the dataset were removed.
- Rating column had one string value (‘Like’) which changed to numeric value (4).
- The time column datatype was changed to date time. The time column was exploded into year, month, day and hour columns.
- The metadata column was also exploded into two columns: review_number and followers. Time and metadata columns were dropped.

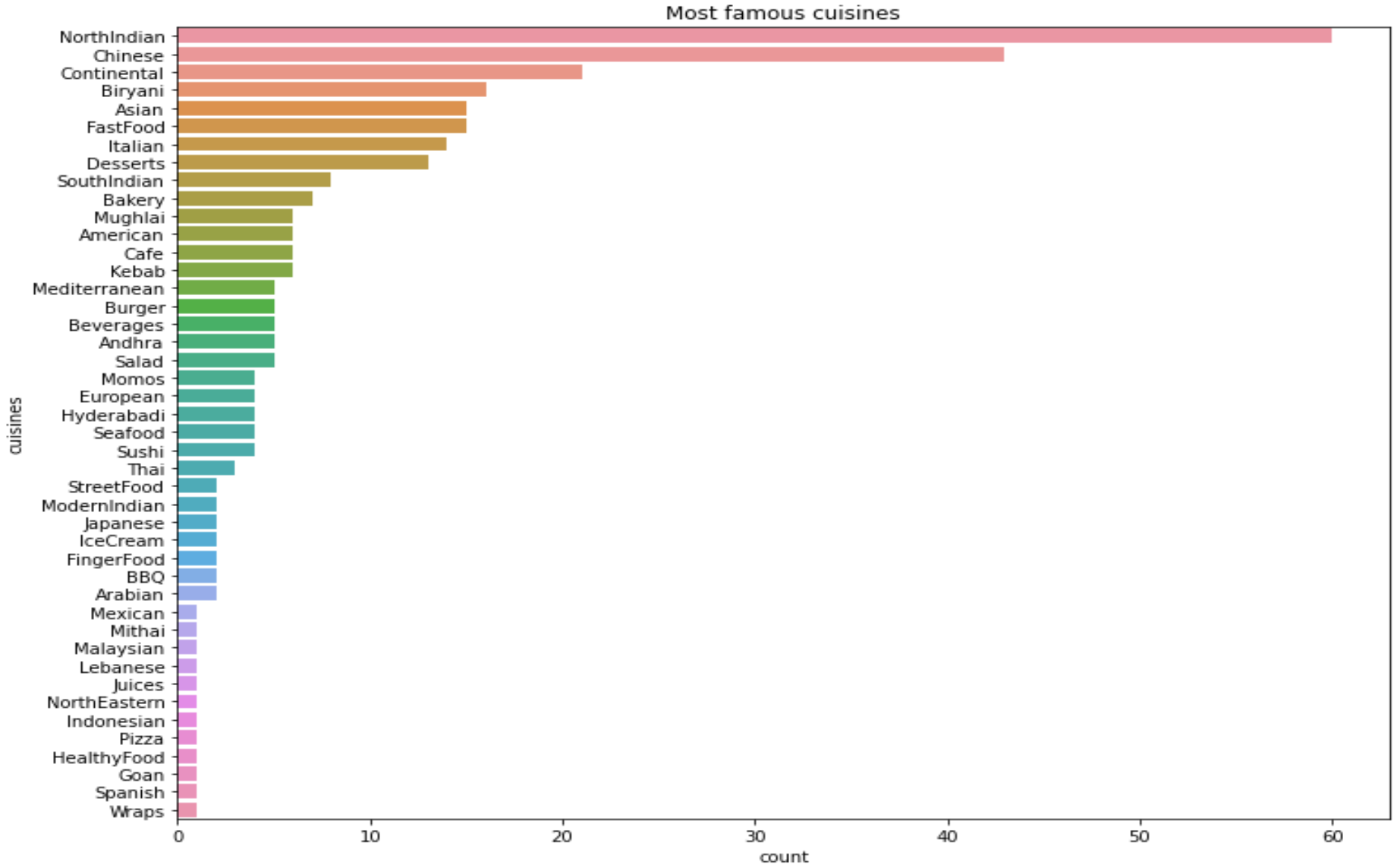
Exploratory Data Analysis

Meta_df



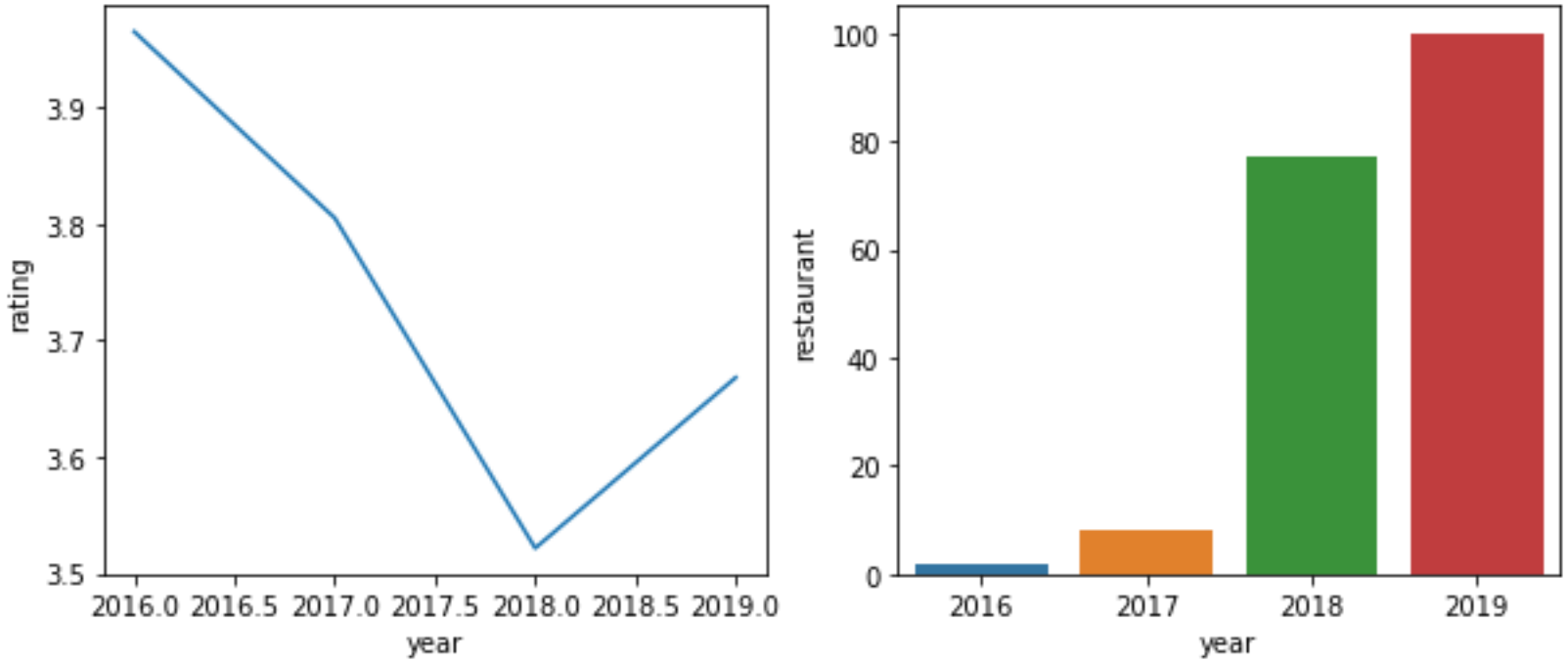


Collage – Hyatt Hyderabad Gachibawli and Feast – Sheraton Hyderabad hotel are most expensive restaurants with cost around 2600 INR. Amul and Mohammedia Shawarma are least expensive restaurants with cost around 150 INR.

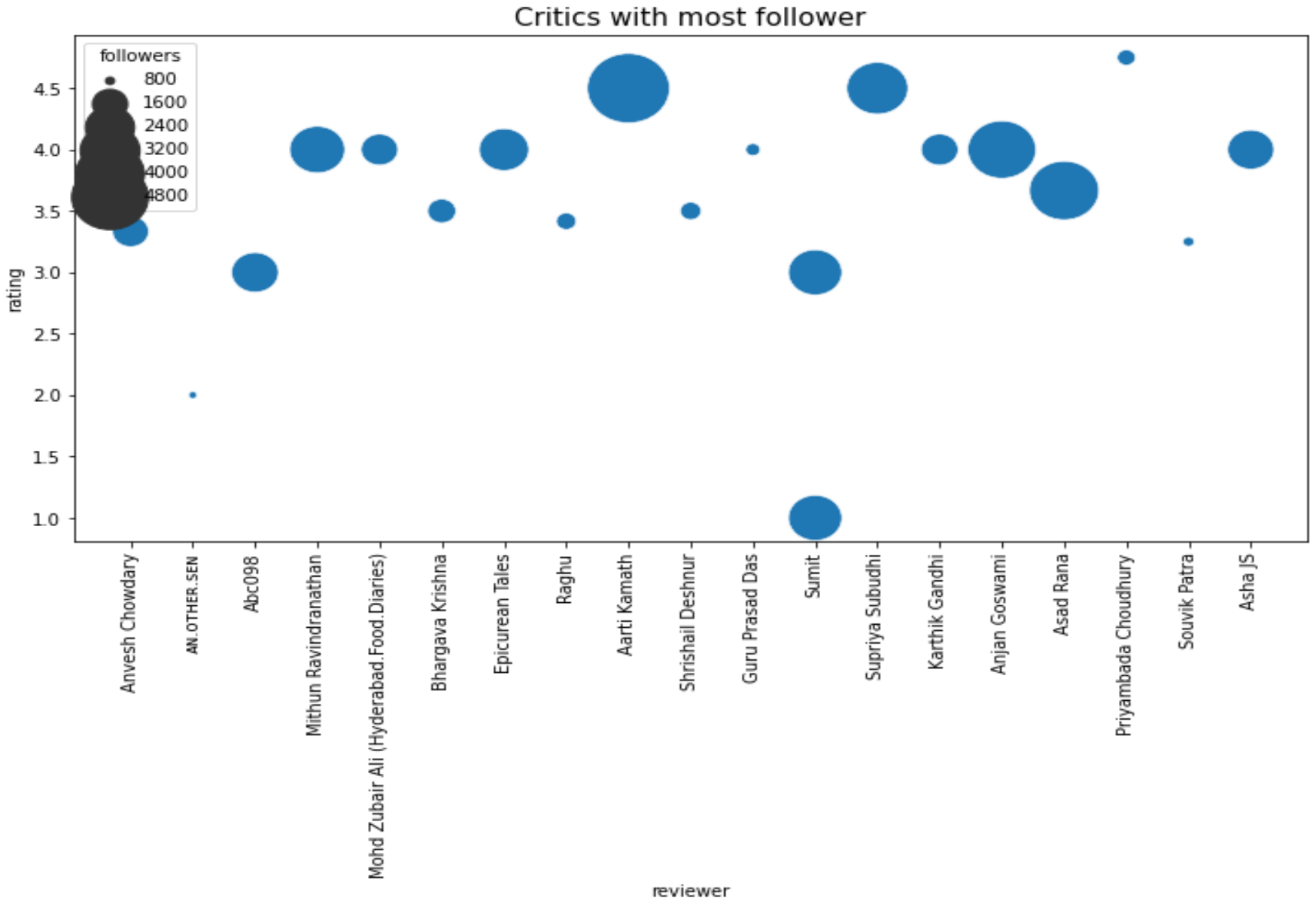


North Indian, Chinese and Continental cuisines are most preferred choice of customers.

Review_df

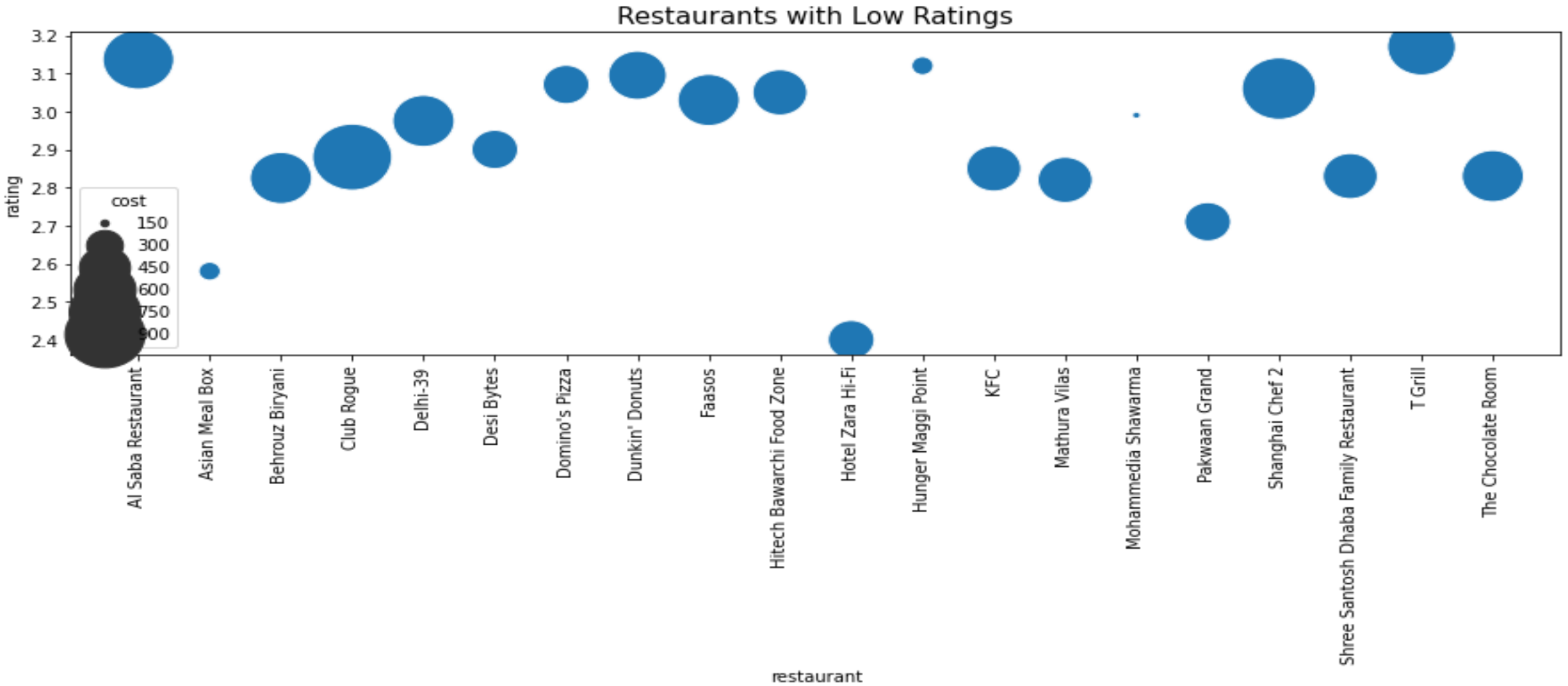


- Since 2016, the ratings of the restaurants had significantly decreased till 2018 but there was improvement after 2018 to 2019.
- High rating in 2016 can be adjudged to low number of restaurants. Improved rating from 2018 to 2019 with even more restaurants suggests focus given on customer service and satisfaction.



Aarti Kamath, Asad Rana and Anjan Goswami are some reviewers with huge followings and have also given high ratings to the restaurants whereas Sumit which have significant following have given very low ratings.

Merge_df

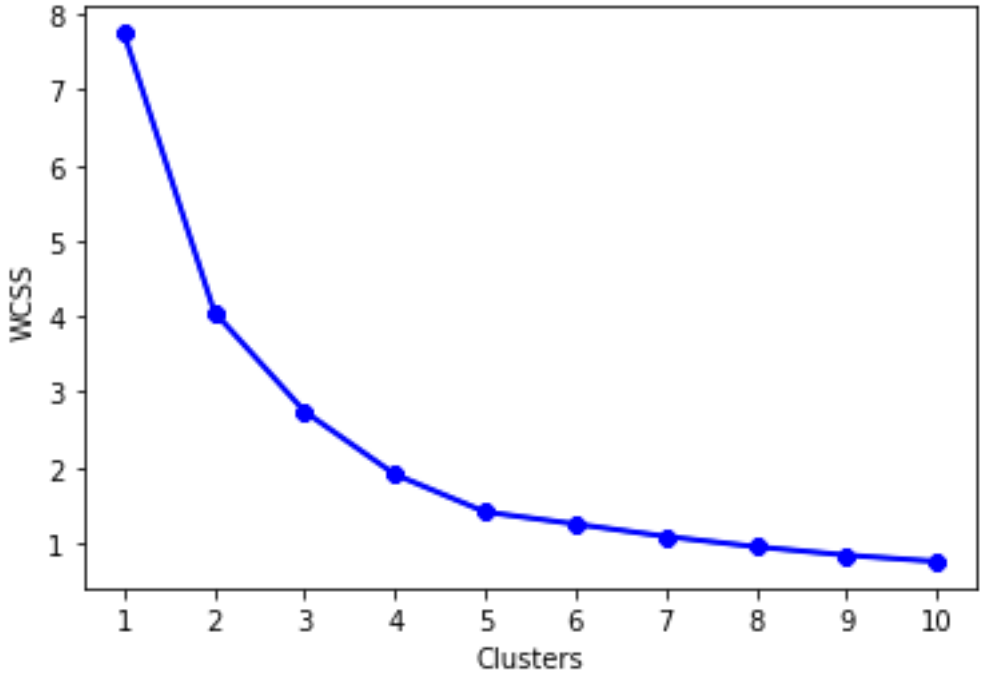


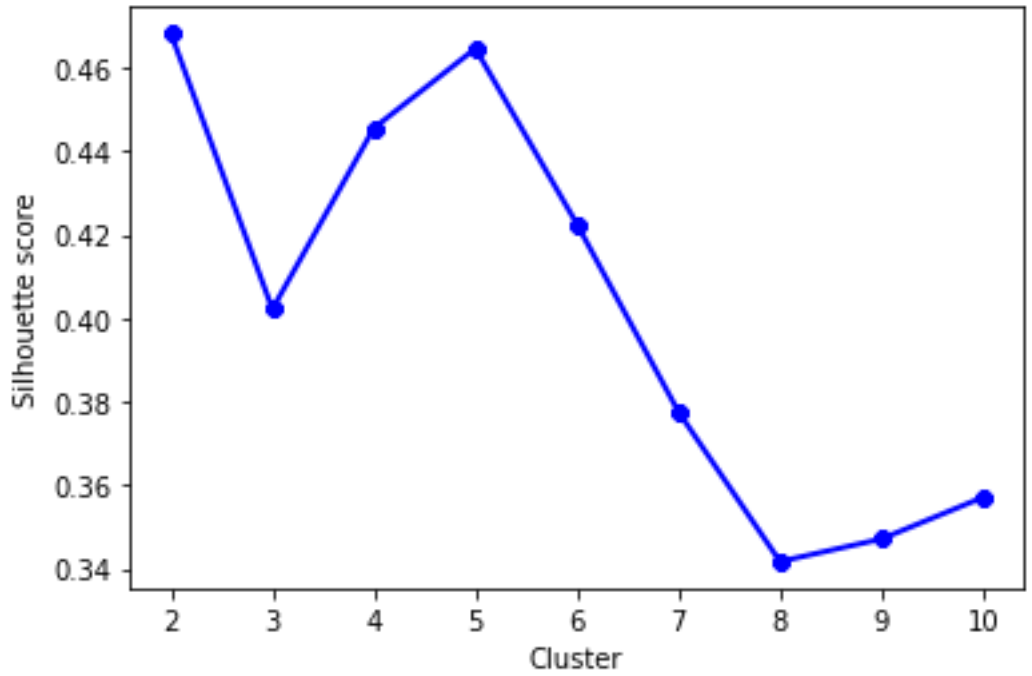
The plot shows the restaurants with low ratings and the size of the dot represents the cost. Some restaurants have very high cost but the service is poor.

```
Average rating of Amul
3.93
Average rating of Mohammedia Shawarma
2.99
```

- Though the cost of both Amul and Mohammedia Shawarma are similar, but the average rating of Amul is much higher than Mohammedia Shawarma.
- This shows that Amul is doing much better in terms of customer service and satisfaction in very less cost than most other restaurants.
- Amul can be a model restaurant for the restaurants with poor ratings.

Clustering





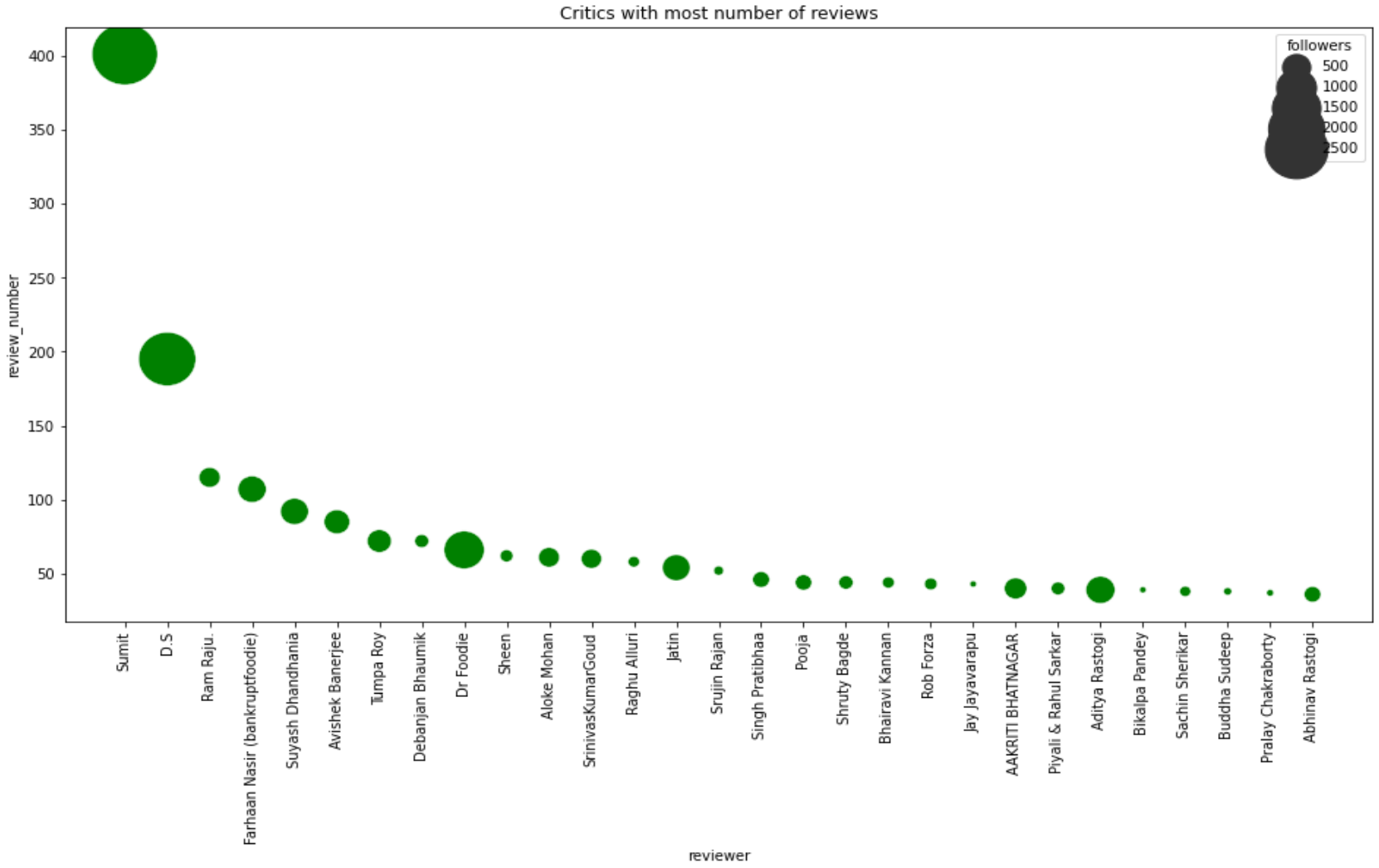
- Kmeans clustering method used for clustering of dataset.
- The number of clusters was ascertained by the elbow curve and silhouette score.

```
Average Cost and Rating across cluster 0
 cost      1029.166667
rating       3.651887
dtype: float64

Average Cost and Rating across cluster 1
 cost      972.222222
rating      3.724167
dtype: float64

Average Cost and Rating across cluster 2
 cost      513.461538
rating      3.566693
dtype: float64

Average Cost and Rating across cluster 3
 cost      790.625000
rating      3.415038
dtype: float64

Average Cost and Rating across cluster 4
 cost      1618.181818
rating       4.000000
dtype: float64
```

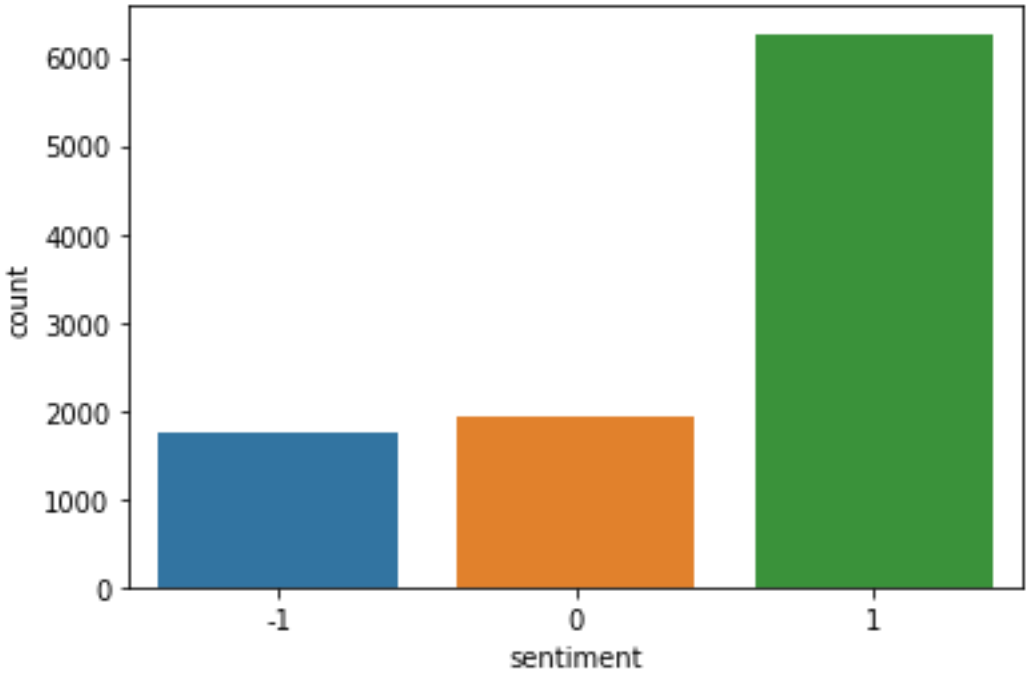
```
Top Cuisines in Cluster 0
 index       765
Chinese       7
Asian         6
dtype: int64

Top Cuisines in Cluster 1
 index             684
NorthIndian        18
Mediterranean       4
dtype: int64

Top Cuisines in Cluster 2
 index        1451
Desserts       12
FastFood       10
dtype: int64

Top Cuisines in Cluster 3
 index           1411
NorthIndian       32
Chinese           31
dtype: int64

Top Cuisines in Cluster 4
 index           540
Continental      10
NorthIndian       9
dtype: int64
```

Average cost, average ratings and top cuisines distribution in each cluster.

Sentiment Analysis



- The most experienced critics based on their number of reviews given in order. Size represents their followers.
- Sumit and D.S has given highest number of reviews and also has most followers.



- Positive sentiment is much higher than the neutral and negative sentiment.
- The phrases like food really good, service also good for positive sentiment and worst experience ever, worst food ever for negative sentiment suggest that the grouping is done very nicely.

Negative Words	Positive Words	Neutral Words
worst experience ever	must visit place	veg non veg
ordered chicken biryani	place hangout friends	paneer butter masala
veg non veg	north indian food	starters main course
veg fried rice	veg non veg	overall good experience
starters main course	good place hangout	food good service
worst food ever	food really good	egg fried rice
went team lunch	overall good experience	yum yum tree
worst service ever	food good service	thai green curry
pan fried noodles	one best places	double ka meetha
overall bad experience	good food good	food 35 service
please dont order	place good food	raju gari kodi
worst biryani ever	place hang friends	main course ordered
waste money time	service also good	quantity less compared
biryani taste good	would love visit	eat india company
dont waste money	one best place	good quantity less

Modelling

```
Confusion matrix for Logistic Regression

[[ 274   66   10]
 [  78  224   86]
 [  41  171 1042]]


Classification report
              precision    recall  f1-score   support

          -1       0.70      0.78      0.74       350
           0       0.49      0.58      0.53       388
           1       0.92      0.83      0.87      1254

    accuracy                           0.77      1992
   macro avg       0.70      0.73      0.71      1992
weighted avg       0.79      0.77      0.78      1992
```

Logistic Regression

```
confusion matrix for Random Forest

[[ 262   24   64]
 [  53   60  275]
 [  12   26 1216]]


Classification report
              precision    recall  f1-score   support

          -1       0.80      0.75      0.77       350
           0       0.55      0.15      0.24       388
           1       0.78      0.97      0.87      1254

    accuracy                           0.77      1992
   macro avg       0.71      0.62      0.63      1992
weighted avg       0.74      0.77      0.73      1992
```

Random Forest

confusion matrix for KNN

```
[[167  37 146]
 [165  37 186]
 [454  69 731]]
```

Classification report

	precision	recall	f1-score	support
-1	0.21	0.48	0.29	350
0	0.26	0.10	0.14	388
1	0.69	0.58	0.63	1254
accuracy			0.47	1992
macro avg	0.39	0.39	0.35	1992
weighted avg	0.52	0.47	0.48	1992

K-nearest neighbor

confusion matrix for SVM

```
[[ 260   49   41]
 [  53  139  196]
 [  10   43 1201]]
```

Classification report

	precision	recall	f1-score	support
-1	0.80	0.74	0.77	350
0	0.60	0.36	0.45	388
1	0.84	0.96	0.89	1254
accuracy			0.80	1992
macro avg	0.75	0.69	0.70	1992
weighted avg	0.78	0.80	0.78	1992

Support Vector Machine

Random Forest and SVM has given the least false positives, but SVM has performed significantly better than all other models, as it has the least false positive and false negative for both negative and neutral statements.

Conclusion

- ✓ The best restaurants are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue.
- ✓ The most popular cuisines are North Indian, Chinese, Continental, and Biryani.
- ✓ The cheapest food joint is Mohammedia Shawarma and Amul and the costliest restaurant is Collage - Hyatt Hyderabad Gachibowli.
- ✓ Rating decreased from year 2016 to 2018 but there is increase in rating from the year 2018 to 2019.
- ✓ Amul has higher ratings with very low cost. It can be model restaurant for the restaurants with poor ratings.
- ✓ Aarti Kamath and Supriya Subudhi are few reviewers with huge followings and gives mostly positive ratings. Sumit and D.S are few critics with huge followings.
- ✓ SVM perform better than all other models, as it has the least false positive

Thank You