Project on Unsupervised learning

Zomato clustering and sentiment analysis

Work done by- Sunil Kumar

Content

- Problem Statement.
- Project Steps.
- EDA-Data preparation and exploration.
- Visualization-Top 10 rated restaurants in 2019.
- Visualization-Top 10 rated restaurants in 2018.
- Visualization-Yearly reviews trend.
- Visualization-Popular cuisines restaurant in Hyderabad
- Visualization-Most Popular cuisines restaurant in Hyderabad.
- Visualization-10 most expensive restaurants.
- Visualization-10 least expensive restaurants
- Natural Language Processing(NLP).
- Clustering of restaurants.
- Most frequent word used and Popular cuisines.
- Modeling and Hyper parameter tuning.
- Techniques and accuracies
- Challenges.
- Conclusion.

Problem Statement

Zomato is an Indian restaurant aggregator and food delivery start-up 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. The data is visualized as it becomes easy to analyze data at instant. 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.

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.

Project Steps

- Data collection and data Understanding.
- Loading to tool and importing Libraries.
- EDA-Data Preparation and Data cleaning.
- Data Transformation.
- Modeling
- Model Ensembles.
- Model Evaluation.

Steps and challenges

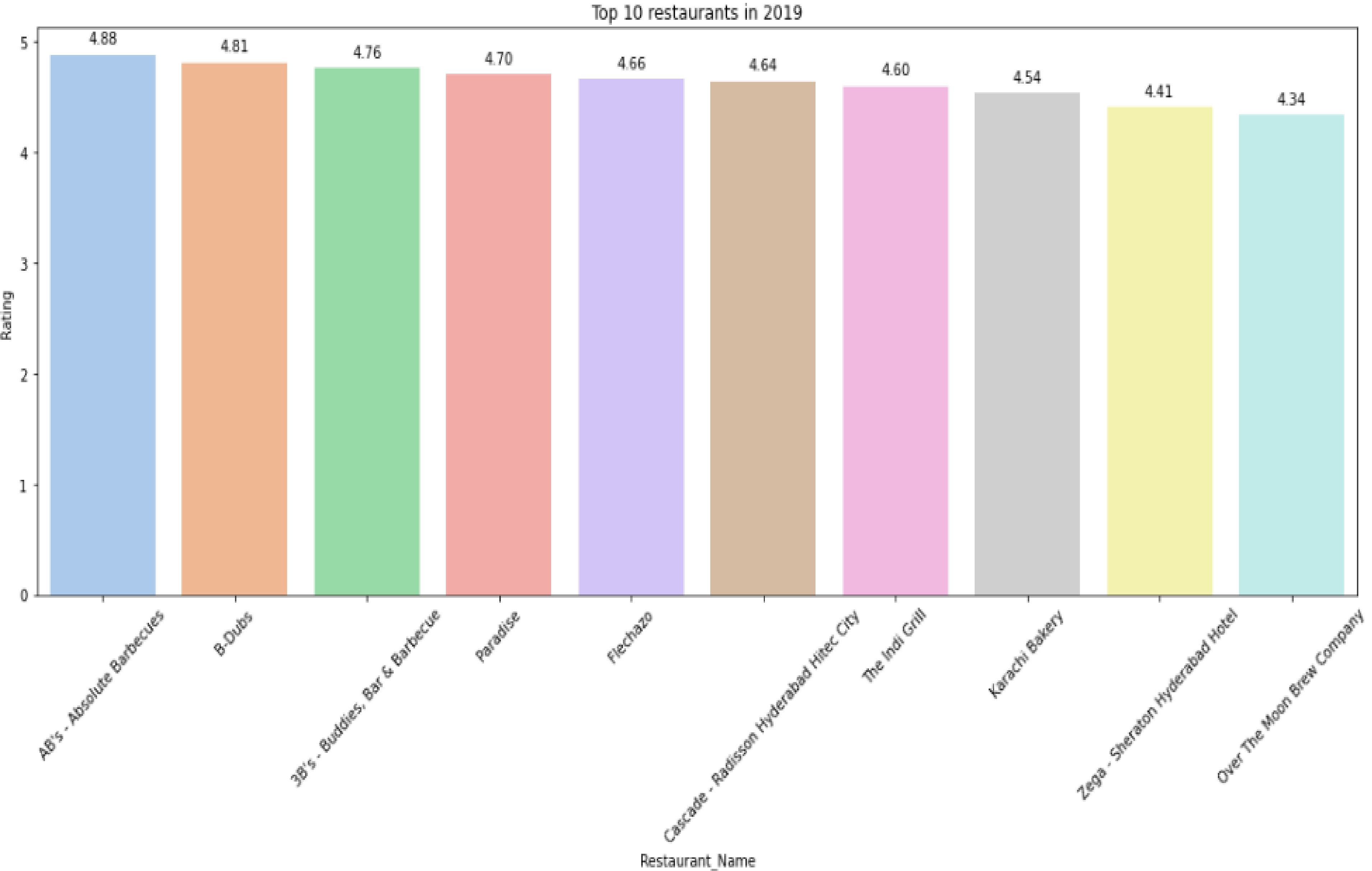
- Data collection and preparation.
- EDA.
- Finding out missing values.
- Finding out duplicates.
- Finding out outliers.
- Data Transformation.(One hot encoding or numerical coding).
- Standardization.
- Modeling.
- (i)- NLP-wag of words.
- (ii)-K-mean-clustering.
- Model ensembles.
- Hyper parameter tuning.
- Model evaluation.

Modeling and Hyper parameter tuning

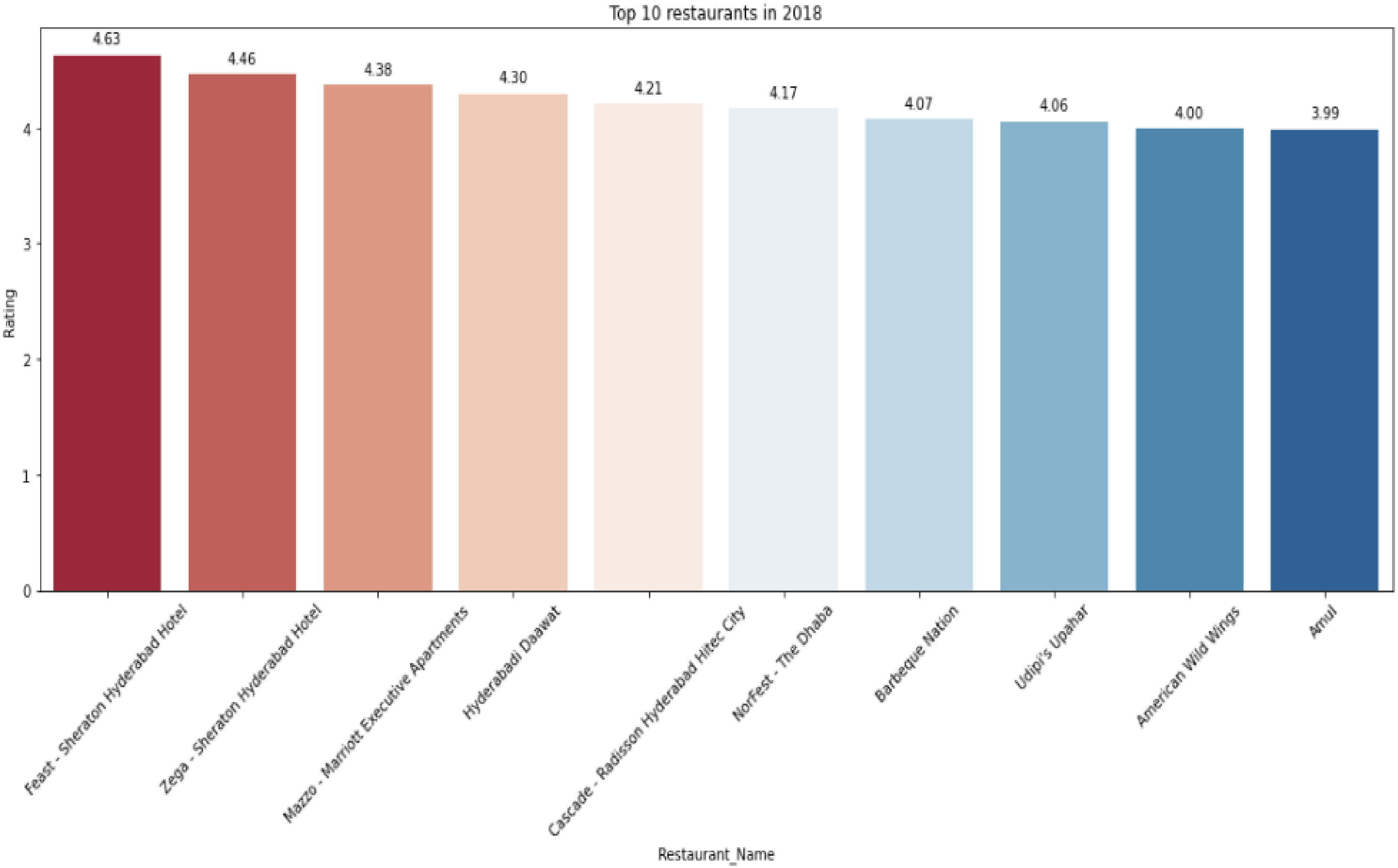
Model used

K-Mean clustering.
NLP-bag of words
Logistic regression.
Model Ensembles.
Decision Tree.
Naïve Bayes.
Hyper parameter tuning.
(i)-Cross Validation.
Iterating models using weights.

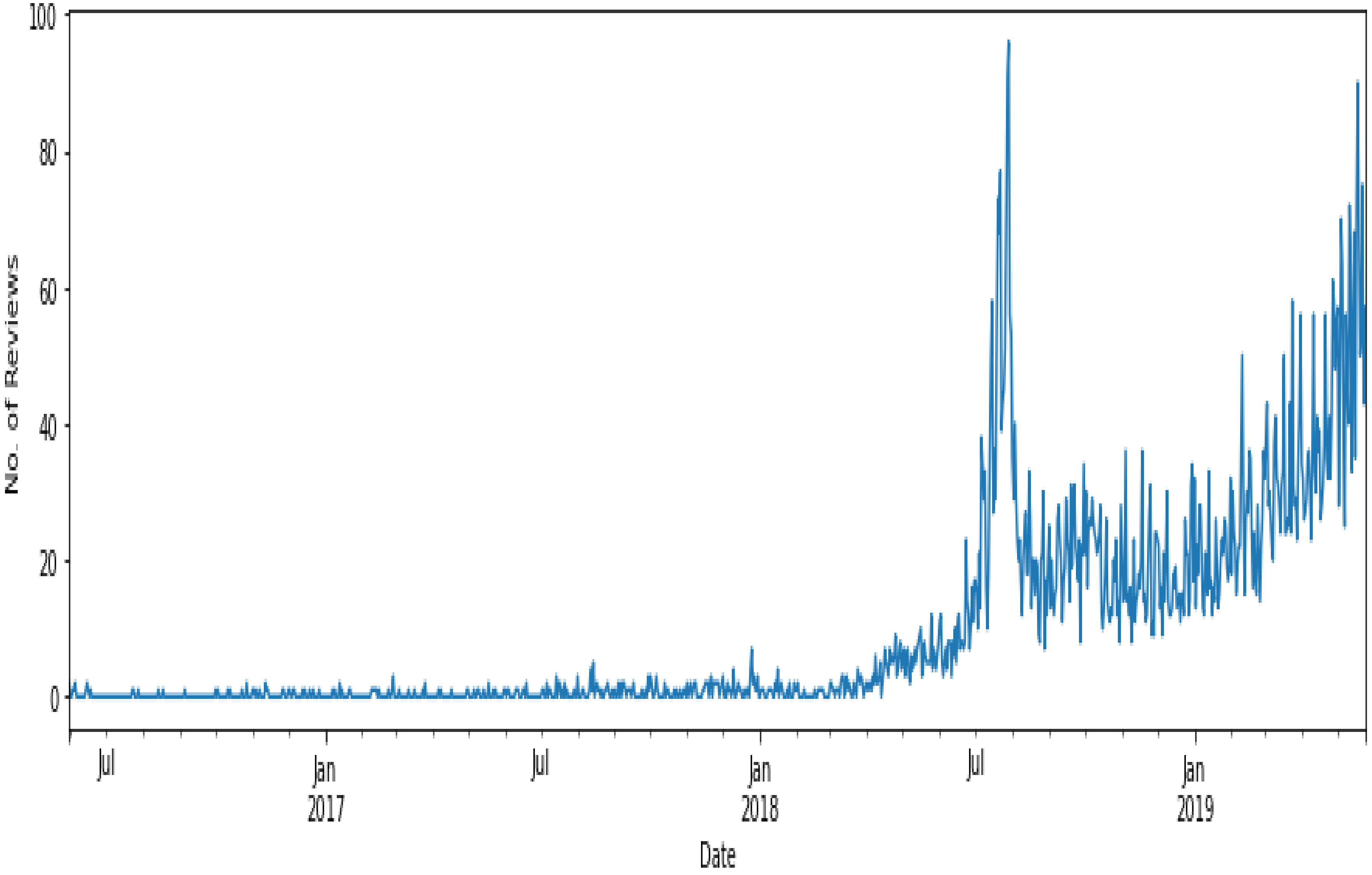
Visualization-Top rated restaurant in 2019



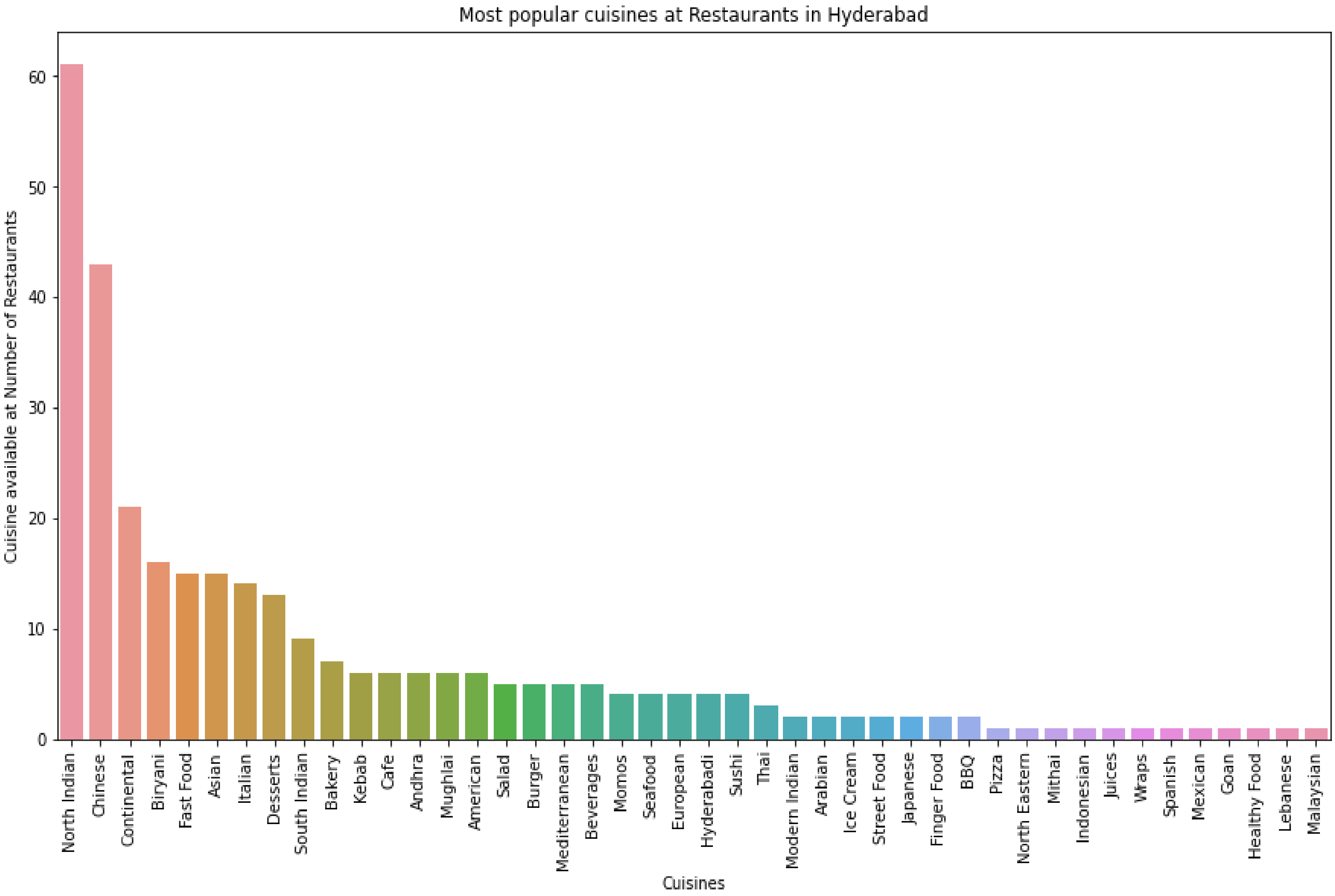
Visualization-Top Restaurant Rated in 2018



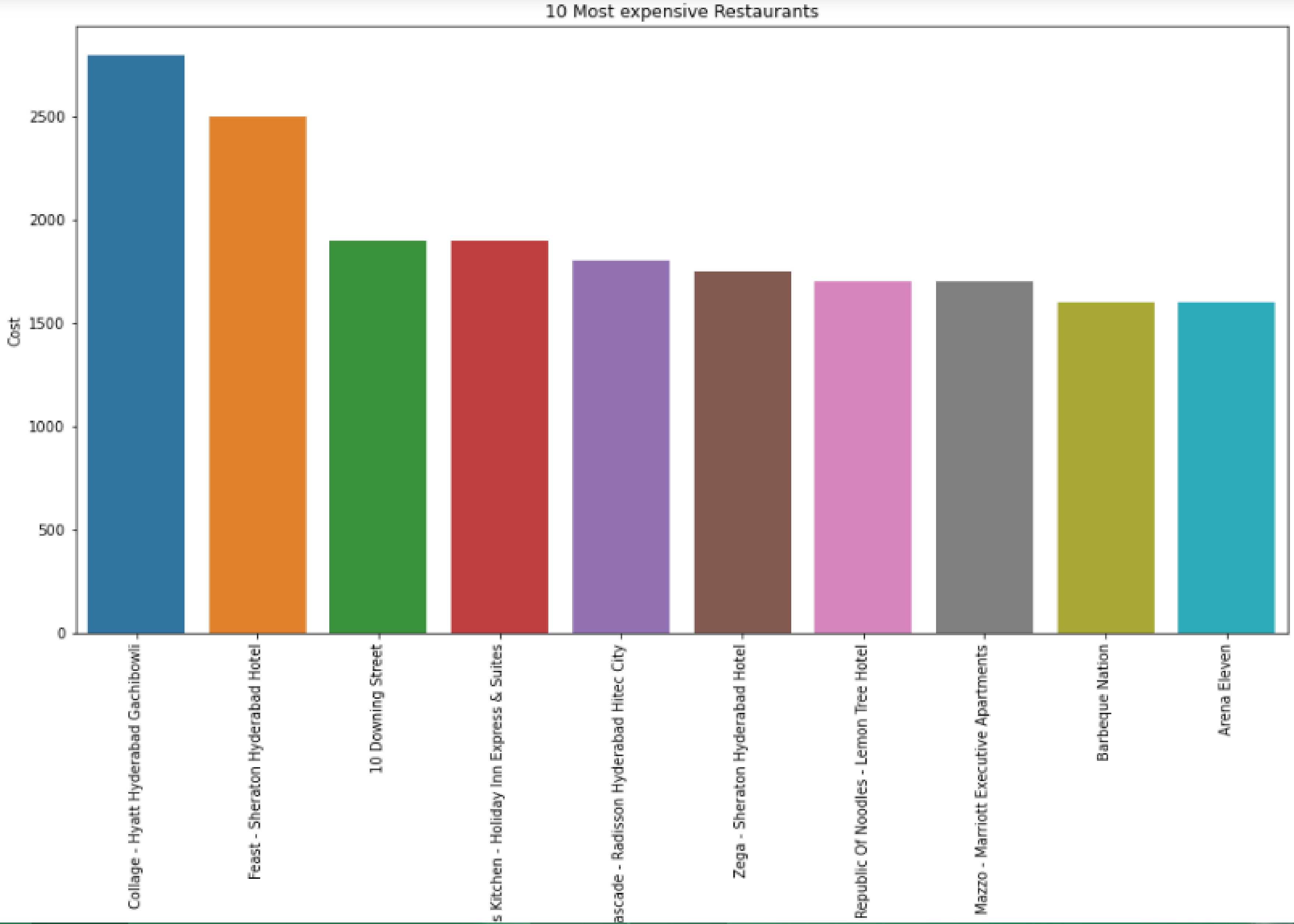
Visualization-Yearly reviews trend



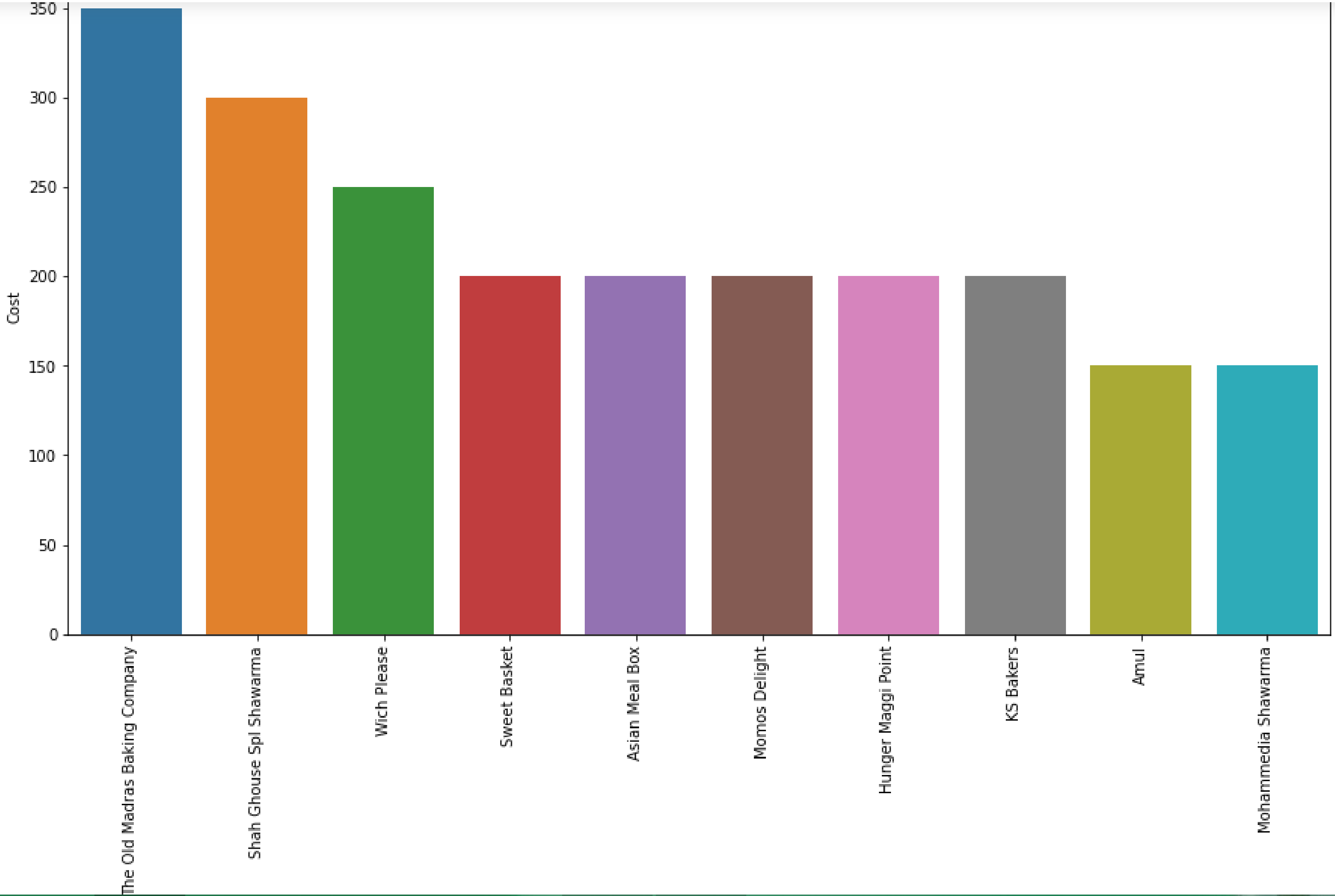
Visualization-Popular cuisines restaurant in Hyderabad



Visualization-10 most expensive restaurants

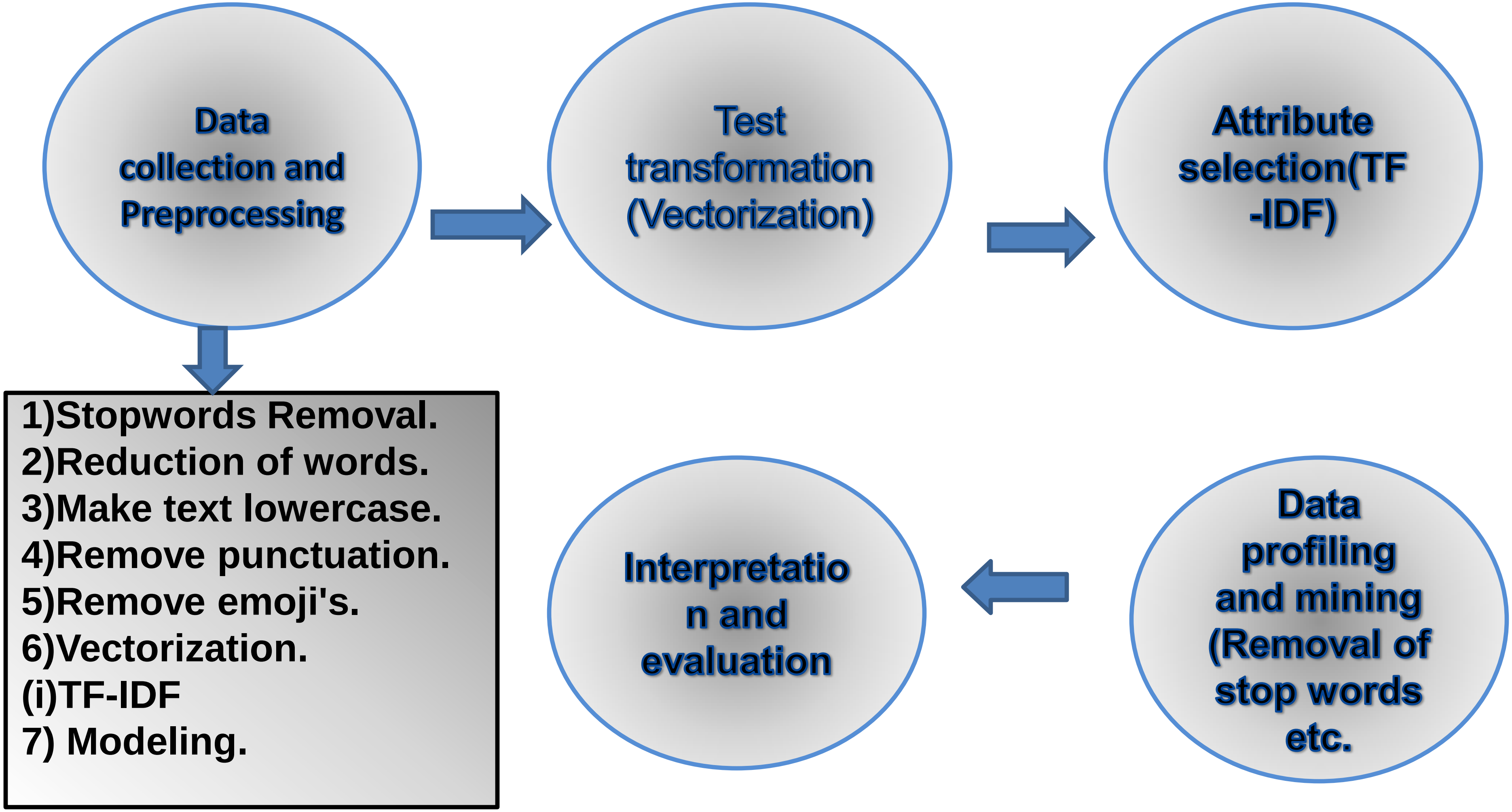


Visualization-10 least expensive restaurants



Natural Language Processing(NLP)

NLP part of sentimental analysis involves in below steps



Data collection and Preprocessing

Test transformation (Vectorization)

Attribute selection(TF-IDF)

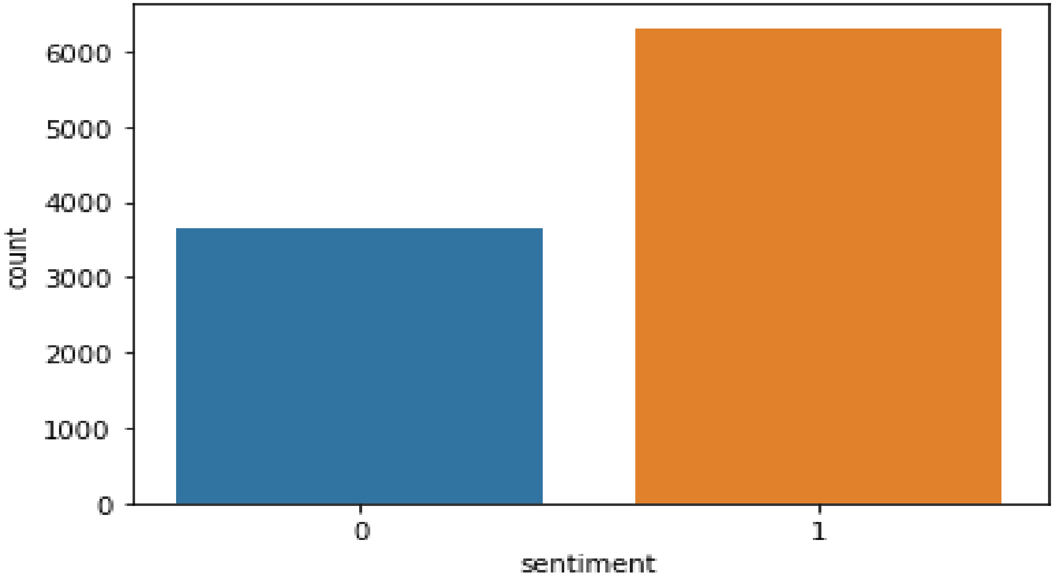
Data profiling and mining (Removal of stop words etc.

Interpretation and evaluation

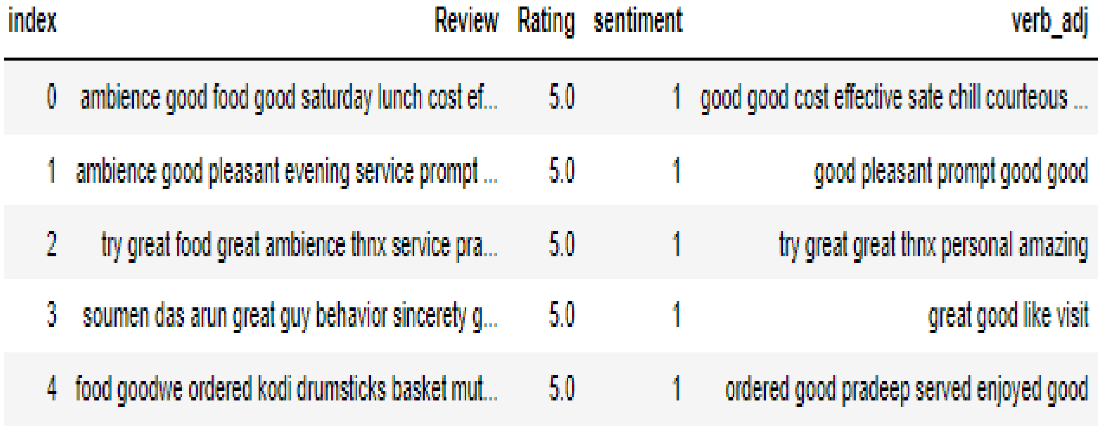
1)Stopwords Removal.
2)Reduction of words.
3)Make text lowercase.
4)Remove punctuation.
5)Remove emoji's.
6)Vectorization.
(i)TF-IDF
7) Modeling.

NLP(Sentimental analysis)

process of detecting positive or negative sentiment in text. It's often used by businesses to detect sentiment in social data, gauge brand reputation, and understand customers.



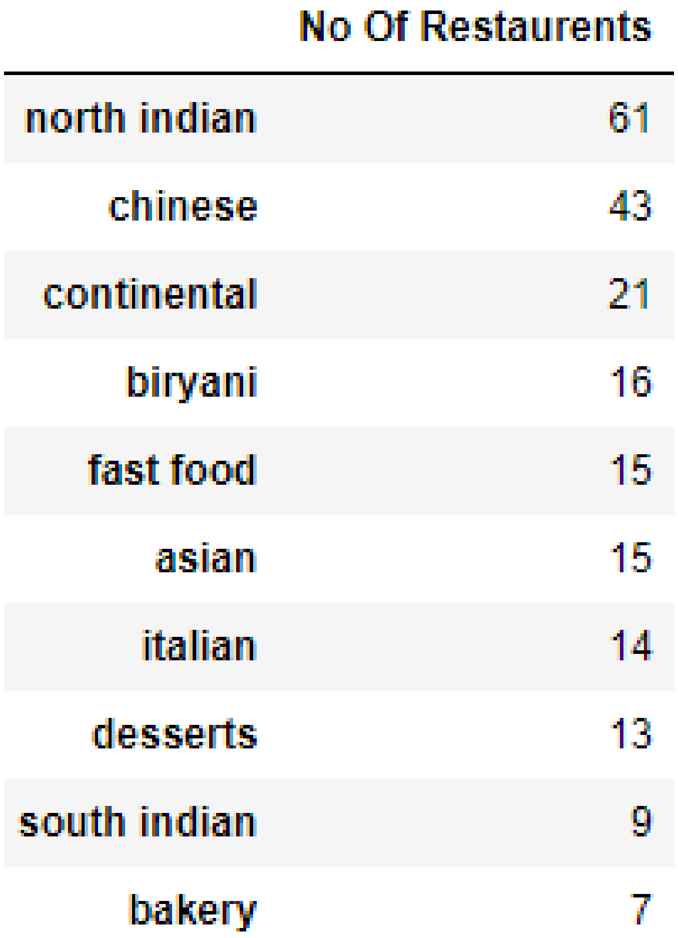
Sentiment is greater than 3.5 than map it to 1
Sentiment is less than 3.5 than map it to 0



index	Review	Rating	sentiment	verb_adj
0	ambience good food good saturday lunch cost ef...	5.0	1	good good cost effective sate chill courteous ...
1	ambience good pleasant evening service prompt ...	5.0	1	good pleasant prompt good good
2	try great food great ambience thnx service pra...	5.0	1	try great great thnx personal amazing
3	soumen das arun great guy behavior sincerety g...	5.0	1	great good like visit
4	food goodwe ordered kodi drumsticks basket mut...	5.0	1	ordered good pradeep served enjoyed good

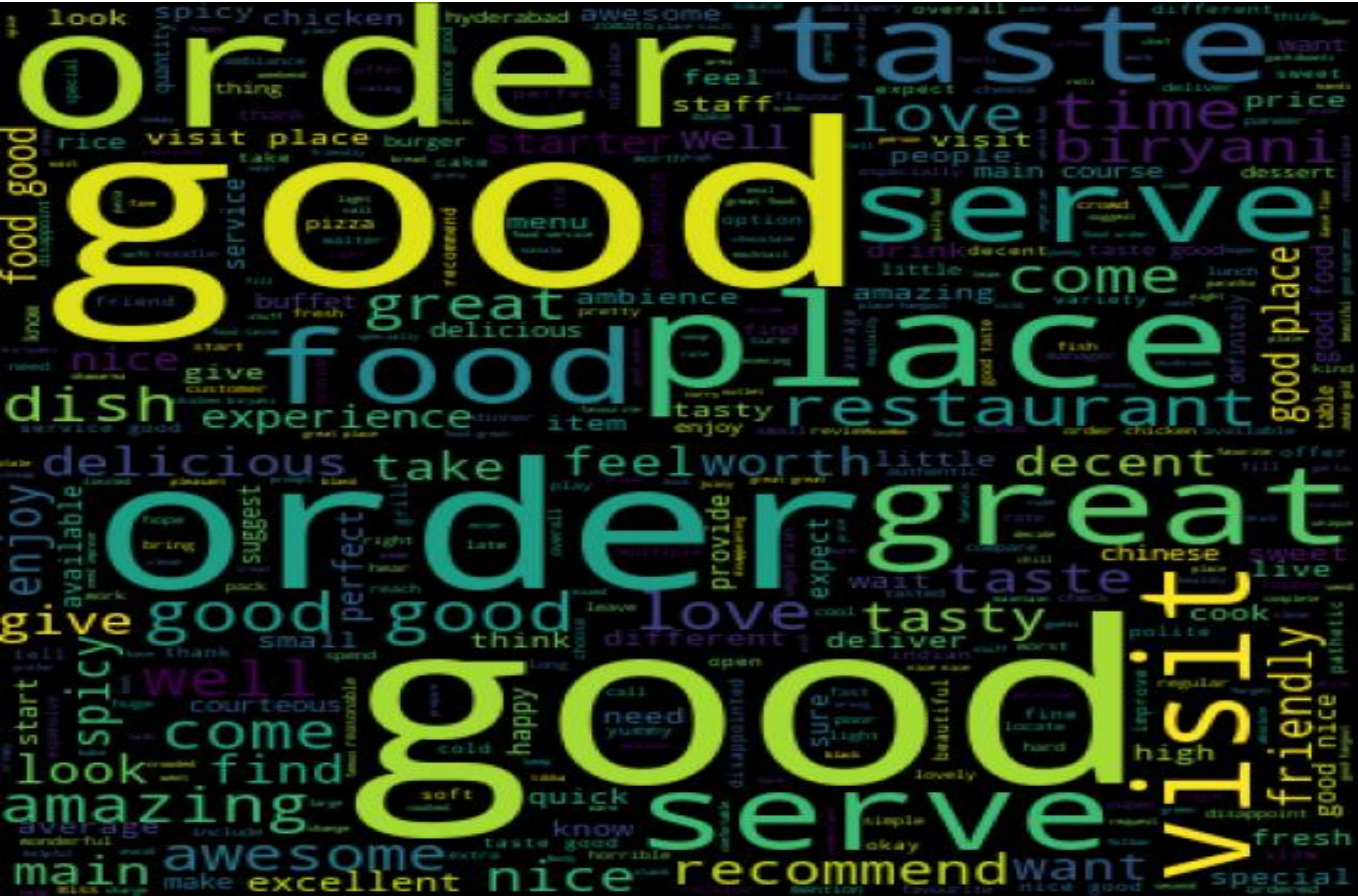
Clustering of restaurants

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.



	No Of Restaurents
north indian	61
chinese	43
continental	21
biryani	16
fast food	15
asian	15
italian	14
desserts	13
south indian	9
bakery	7

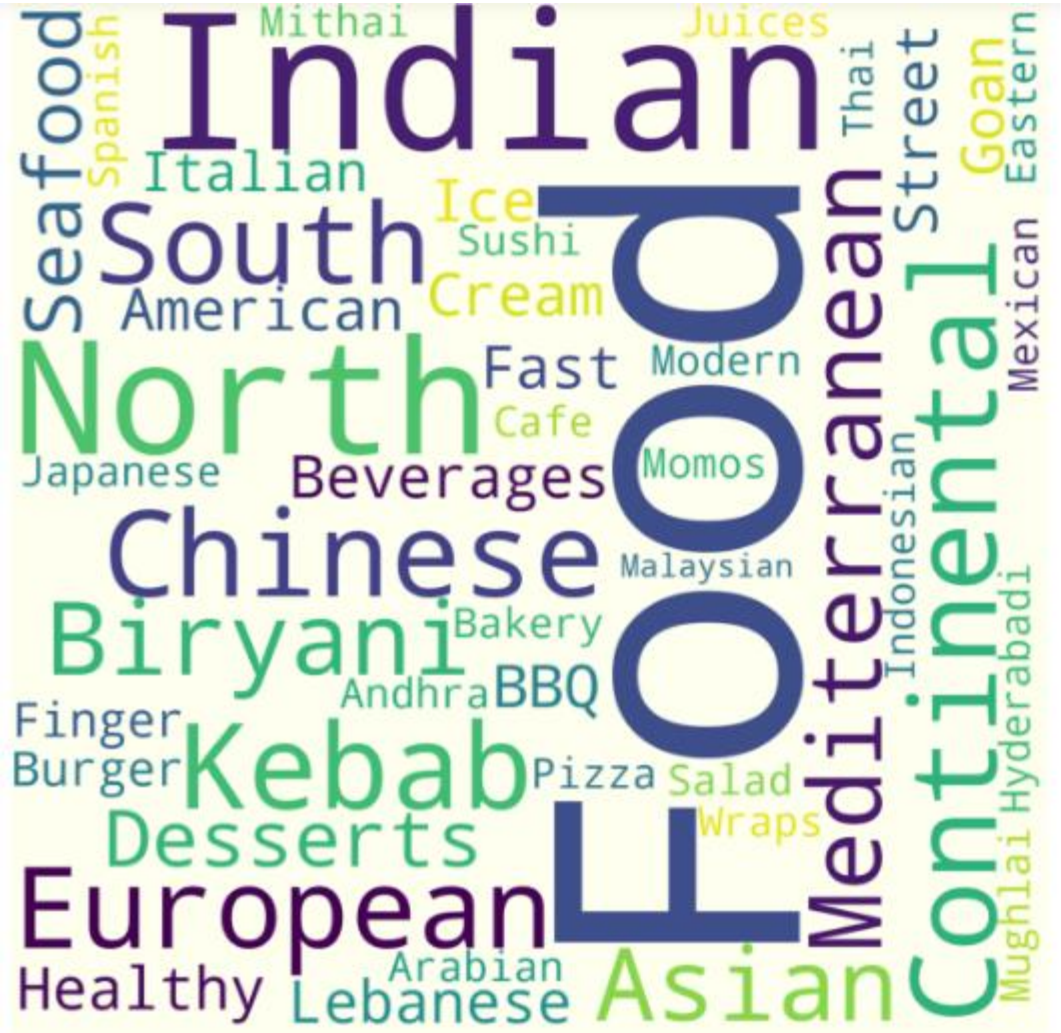
Most frequent used words in review



Popular cuisines

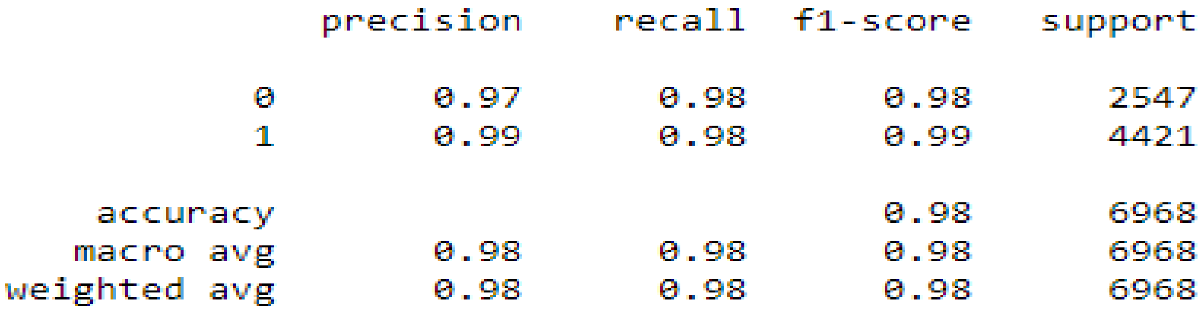
These are some of the highlighted frequencies of the cuisines which are popular and are repeated. Few populars are below.

Indian,Food,North,Chinese,European,Kebab,South,continental etc.

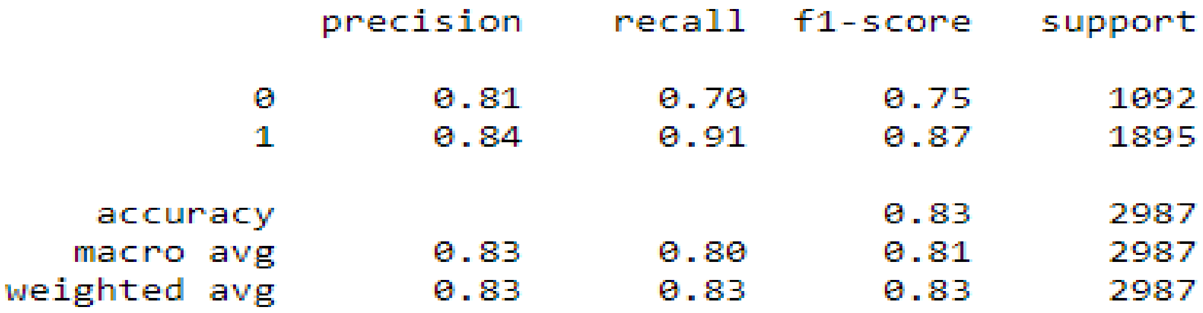


Techniques and accuracies

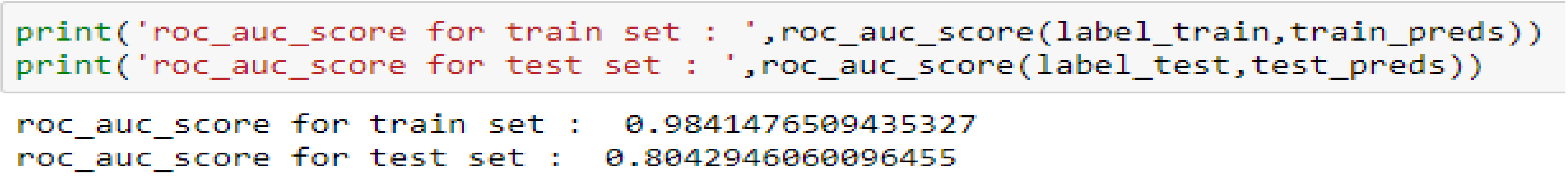
Random Forest with Cross Validation(class_weight={0: 2.0, 1: 1.0})



```
              precision    recall  f1-score   support

           0       0.97      0.98      0.98      2547
           1       0.99      0.98      0.99      4421

    accuracy                           0.98      6968
   macro avg       0.98      0.98      0.98      6968
weighted avg       0.98      0.98      0.98      6968
```



```
              precision    recall  f1-score   support

           0       0.81      0.70      0.75      1092
           1       0.84      0.91      0.87      1895

    accuracy                           0.83      2987
   macro avg       0.83      0.80      0.81      2987
weighted avg       0.83      0.83      0.83      2987
```

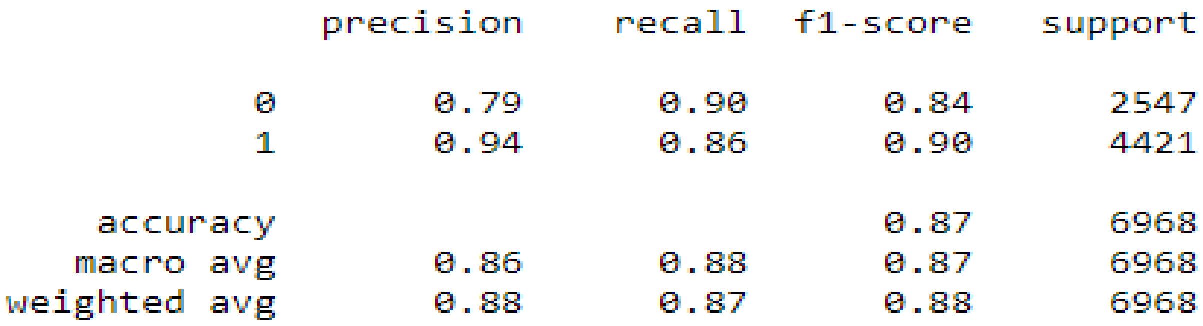


```
print('roc_auc_score for train set : ',roc_auc_score(label_train,train_preds))
print('roc_auc_score for test set : ',roc_auc_score(label_test,test_preds))
```

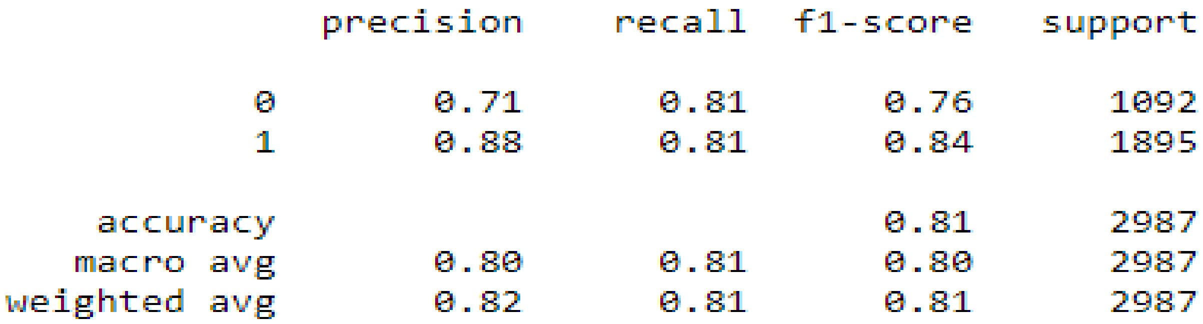
```
roc_auc_score for train set :  0.9841476509435327
roc_auc_score for test set :  0.8042946060096455
```

Techniques and accuracies continued

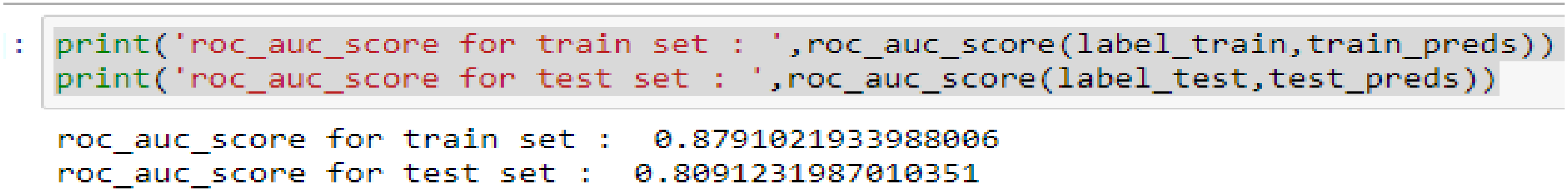
Logistic Regression (class_weight={0: 2.0, 1:1.0})



```
              precision    recall  f1-score   support

           0       0.79      0.90      0.84      2547
           1       0.94      0.86      0.90      4421

    accuracy                           0.87      6968
   macro avg       0.86      0.88      0.87      6968
weighted avg       0.88      0.87      0.88      6968
```



```
              precision    recall  f1-score   support

           0       0.71      0.81      0.76      1092
           1       0.88      0.81      0.84      1895

    accuracy                           0.81      2987
   macro avg       0.80      0.81      0.80      2987
weighted avg       0.82      0.81      0.81      2987
```



```
print('roc_auc_score for train set : ',roc_auc_score(label_train,train_preds))
print('roc_auc_score for test set : ',roc_auc_score(label_test,test_preds))
```

```
roc_auc_score for train set :  0.879102193988006
roc_auc_score for test set :  0.809123198701035l
```

Conclusion

Started with data loading and importing the libraries and then started with the exploring the data and done some of the visualization to see the common pattern in the data and looked into columns and rows. It was seen that there were few missing values present as per the ratio, imputed the values accordingly. I have seen that I have almost everything in the form of words which I converted accordingly. Found some of the string and removed. Found 4 years were unique. Explored the restaurant with respect to rating received. Found some of the top restaurants for the ratings and year wise popularity and saw the ratings trend with respect to year and seems to be growing over the year. Checked the restaurant expensive which were around 2800, 2500/- and being 1600/- top 10th number. Similarly explored the least costly restaurant around 350/- to 150 as top 10th.Also checked the overall cost. During the clustering of cuisines I have seen that North India, continental, chinese,sea food, European, fast food etc were popular.

Conclusion

Also, I have explored and seen restaurants weekly open time. I after all these exploration I have merged the dataframe of reviews. Found the north Indian restaurant being the highest numbers of 61 while top 10, 10th was bakery of 7 numbers. Finally started with natural language processing for removing the unwanted words or punctuations present in the data, tried removing emojis and some of the stopwords. I then divided the sentiments into 0 and 1 values where 0 being the negative comments or comments under 3.5 rating and 1 being good comments or comments above 3.5 rating for the restaurant. Converted the words to lower case, and removed the spaces and special character etc. Finally started with the Modeling made use of Bag of words, and Naïve bayes multinomial classification, Decision tree, Random Forest, K-mean clustering, and Logistic regression. However I have tried with different weight for the model at different times in an iteration but only I have good findings for the Random forest classfier and for logistic regression using cross validation hyper parameter tuning have given the better results with best parameters to use.

Challenges

The major challenges I have faced in this project are mentioned Below:

1)Analysing the data and exploring the data was making no much sense because data that I had was having objects datatypes and few strings and comments, restaurants names , reviews etc.

2)Removal of punctualtions and stopwords, symbols, emojis and some of the repeated words like (myyyyyyyyyyyyyyyyy).

3)I have faced challenged in merging the dataframes as it was not matching the length.

4)I have faced the challenged while importing the libraries. I had to install some of the libraries like spacy, contractions.

5)I have faced challenges while using K-Mean clustering could not use effectively. Scaling consumed more time and could not understand data for a while.

6)Hyperparameter tuning was another problem which I faced and consumed a lot more time for rendering.

Thank you