Capstone Project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

UnSupervised Machine Learning Model

TITO VARGHESE

Content

Introduction

In this era of digital world,every business needs to have a online presence.As a result,online food business creates an opportunity to the businessman in expanding their service output with some additional pick-up in the business.

Restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. 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 each city in India.

Problem Statement

- To cluster the zomato restaurants into different segments and made some useful findings in the form of Visualization that can -

  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.

- To analyze the sentiments of the reviews given by the customer in the data

DATA SUMMARY

Zomato Restaurant names and Metadata (Clustering)

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 categories
5. Cuisines : Cuisines served by Restaurants
6. Timings : Restaurant Timings

DATA SUMMARY

Zomato Restaurant reviews (Sentiment Analysis)

1. Restaurant : 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

Pipeline

Understanding and Cleaning

- Null value analysis
- Missing value treatment
- Outlier Treatment

Graphical

- Univariate analysis with visualization
- Bivariate Analysis with visualization

Machine Learning

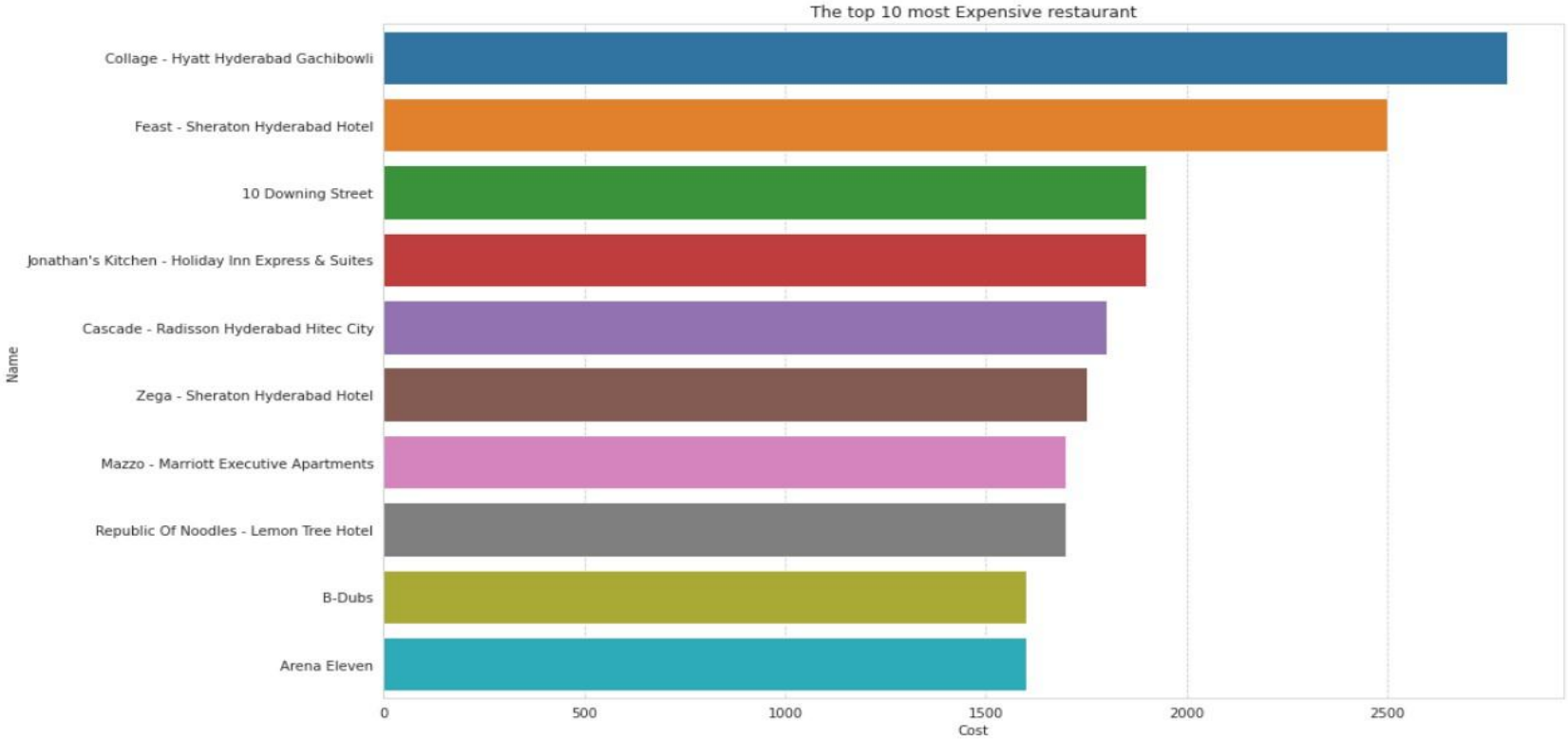
- Clustering
- Topic Modeling
- Classification

Data Inspection & Cleaning

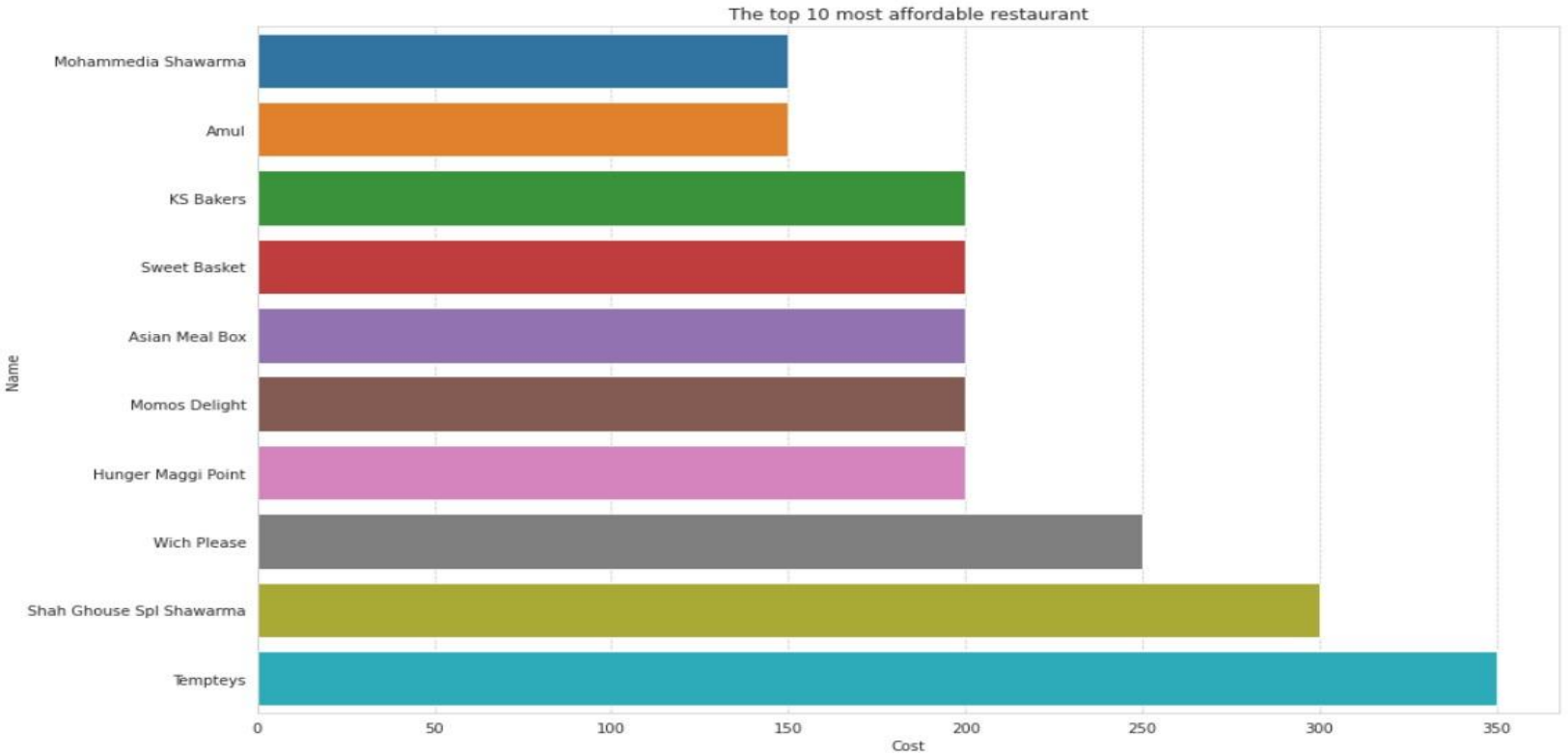
- The zomato restaurant dataset consist of 105 restaurant records.
- The zomato review dataset consist of 10,000 reviews.
- Two features in Zomato Restaurant dataset are having missing values i.e the Collections (54 null value) and timings(1 null value).
- We had 36 duplicate records in zomato review dataset.
- The most frequent working hours of the restaurant is between 11am to 11pm.
- We are having four year zomato review dataset between 2016 to 2019.
- We have done feature engineering on Metadata feature and Timing Feature and extracted few new features (Reviewer,Followers,Hour,Month,Week and Year).
- Followers feature has the highest null value count in the zomato review dataset.
- The average price range of cuisines in the restaurant lies between 200 to 2800 inr.

EXPLORATORY DATA ANALYSIS

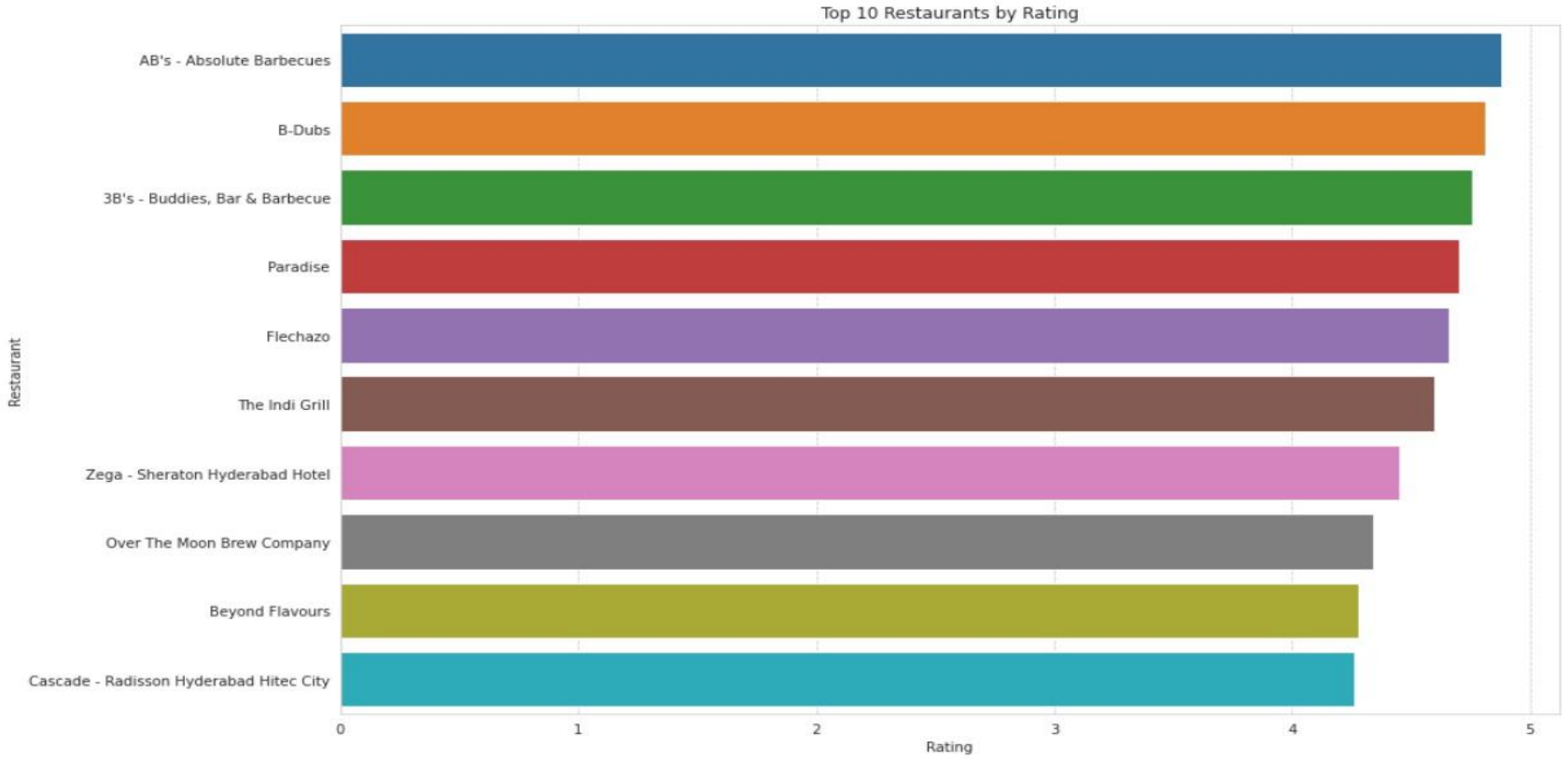
The top 10 most expensive restaurants



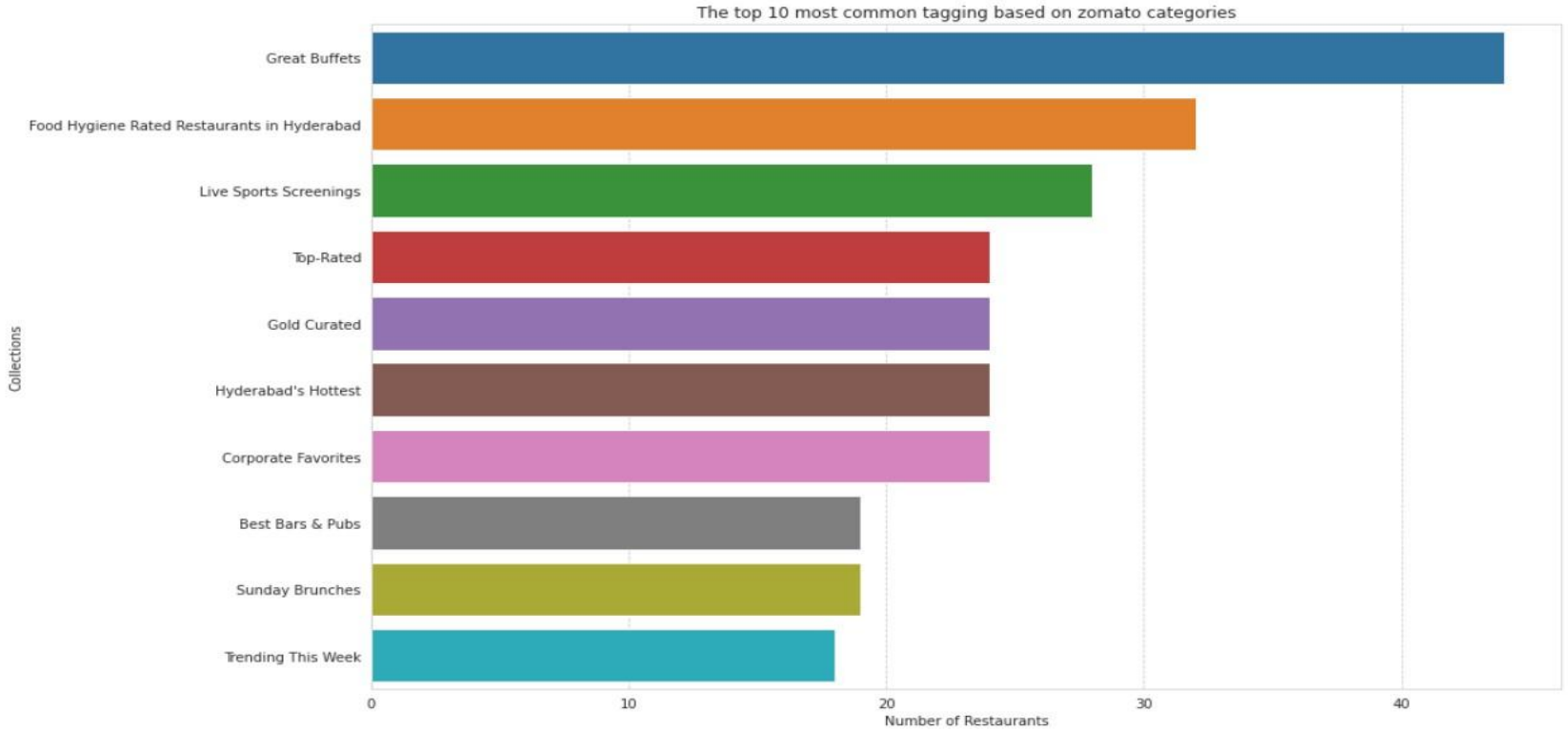
The top 10 most affordable restaurants



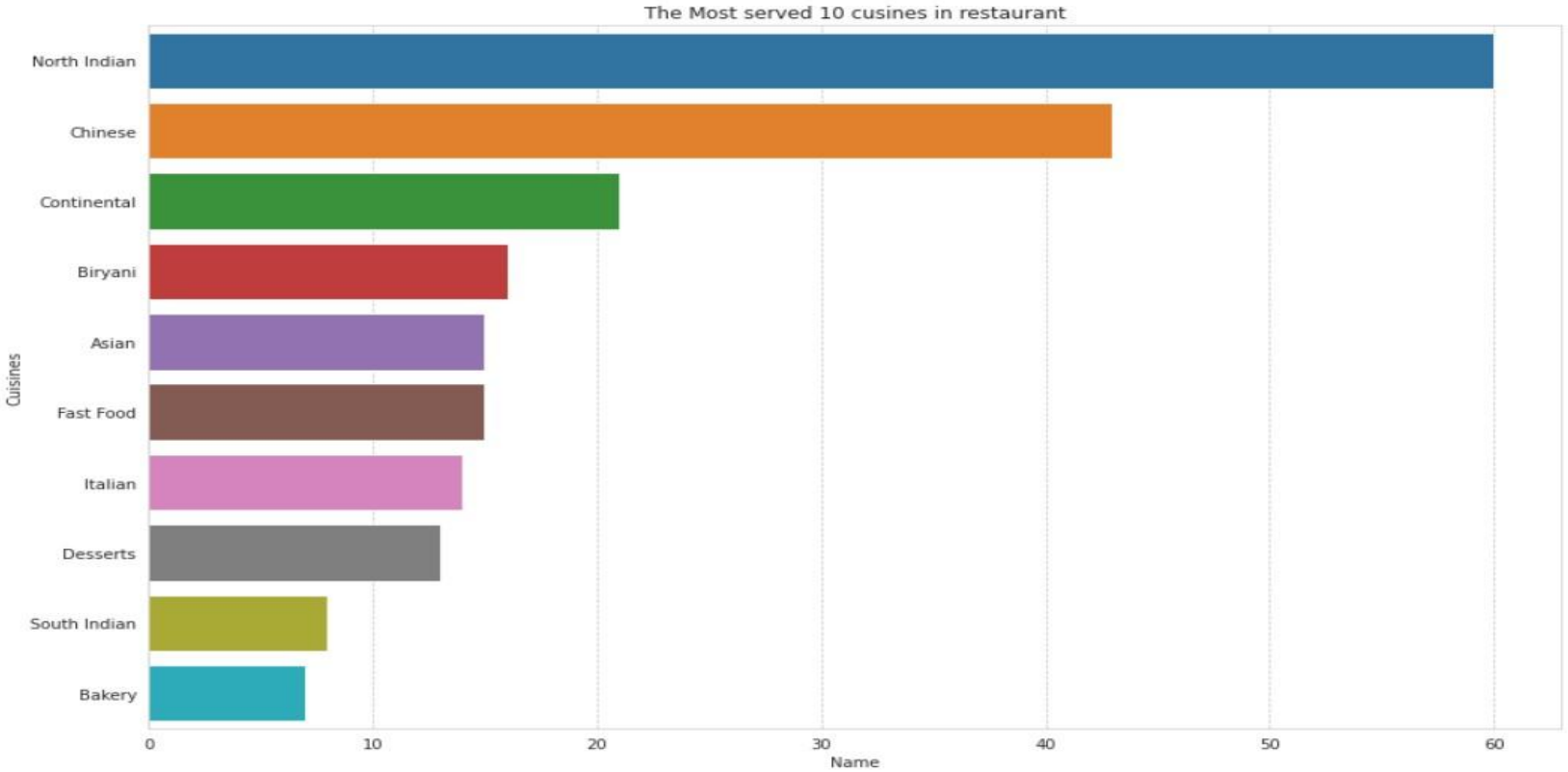
The top 10 best restaurants on Ratings



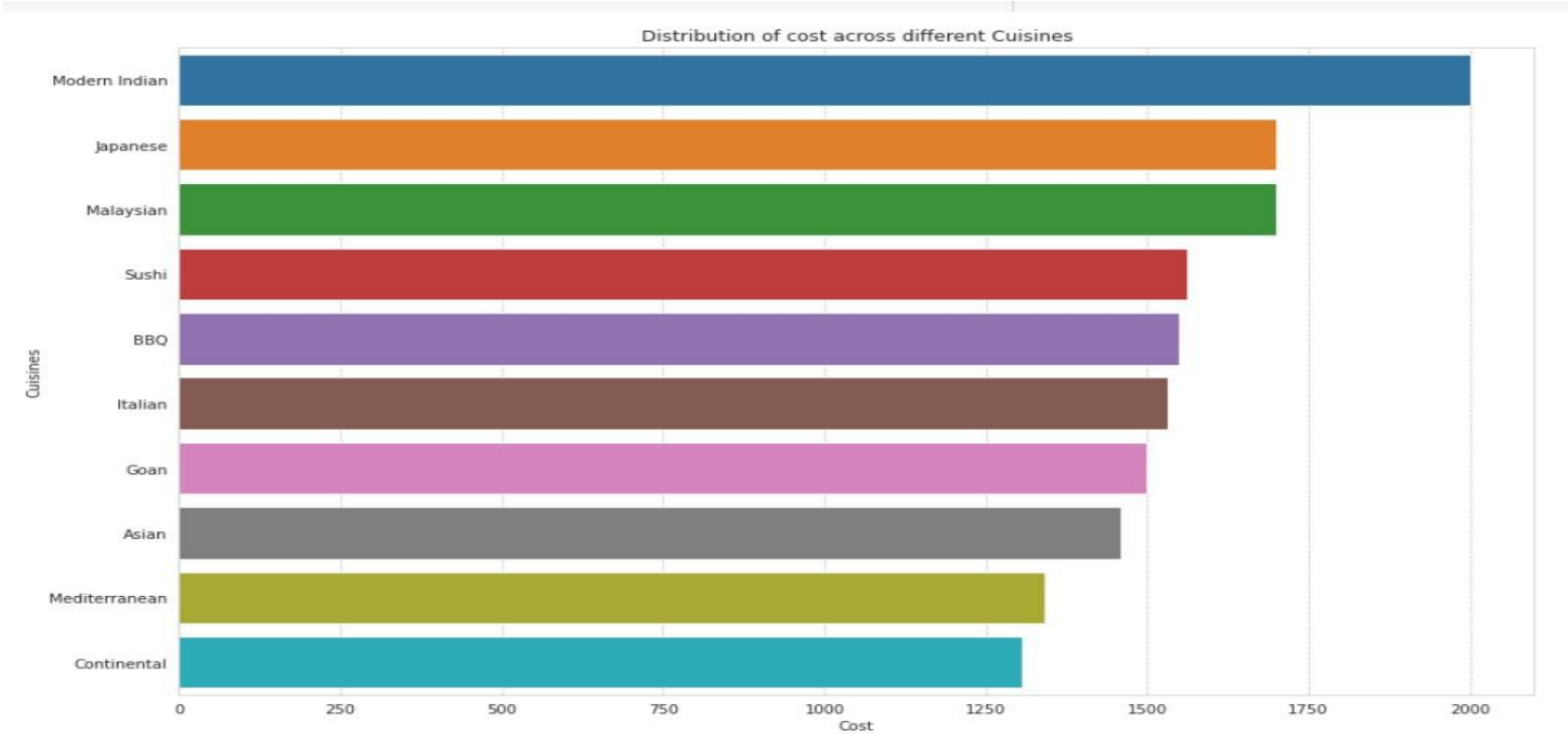
The top 10 most common tags of restaurants



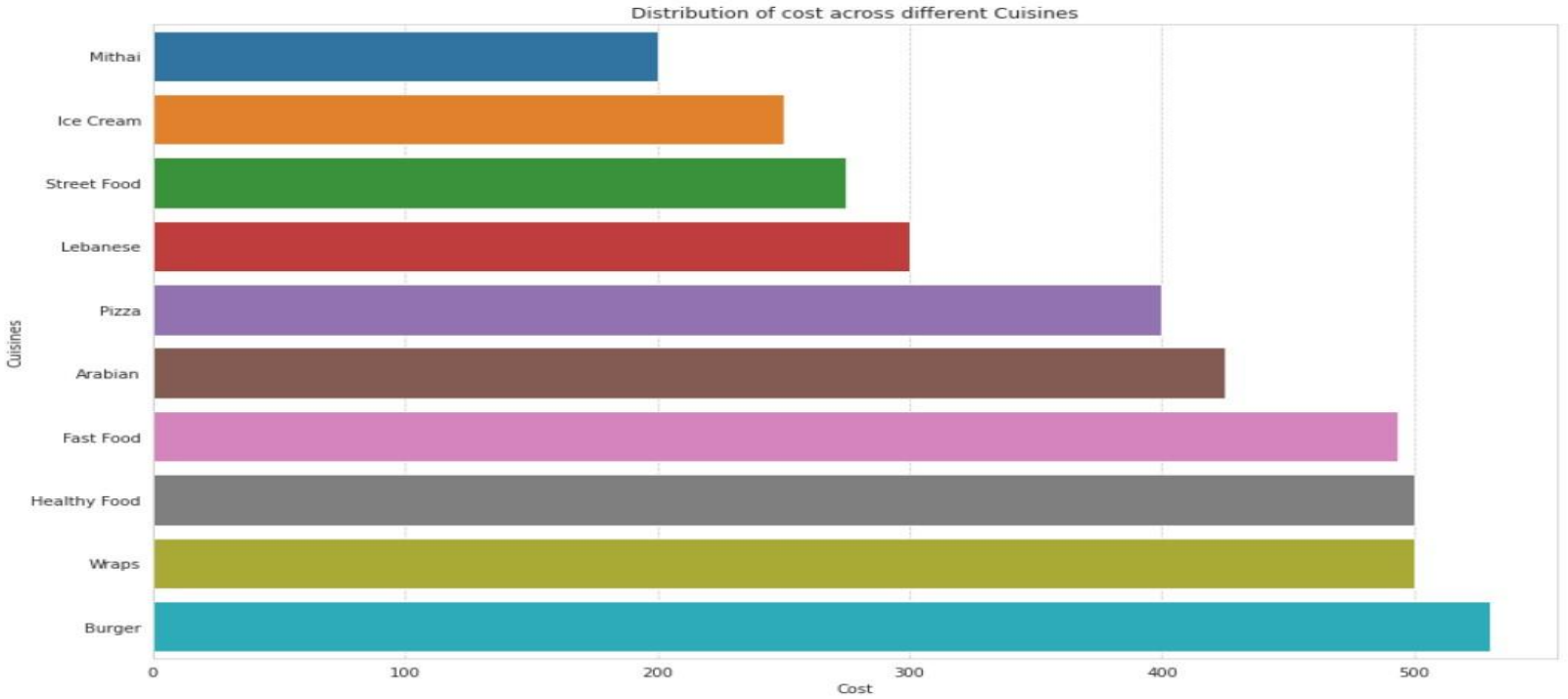
The top 10 most served cuisines in restaurants



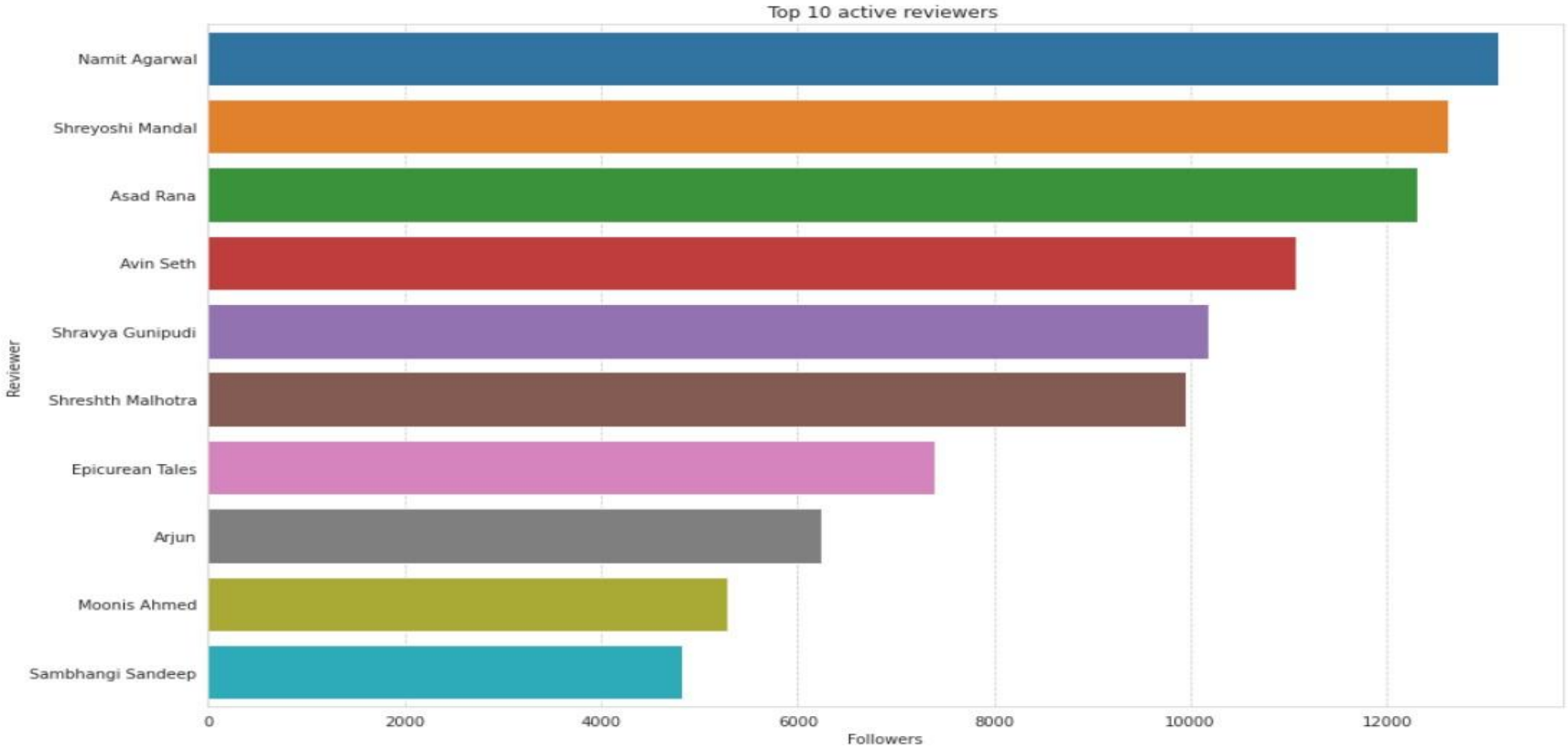
The top 10 most expensive cuisines in restaurants



The top 10 most affordable cuisines in restaurants



Food Critics



Modelling Overview

Supervised Classification Models used for sentiment analysis-

- Logistic Regression
- Decision Tree
- Random Forest
- Xgboost
- Lightgbm

Unsupervised Models used for Topic Modelling -

- Latent Dirichlet Allocation
- Non Negative Matrix Factorization

Unsupervised Models used for Clustering-

- K-Means Clustering
- Hierarchical Clustering
- Principal Component Analysis
- DBSCAN

Modelling Pipeline

- Feature selection
- Feature engineering
- Feature Extraction
- Train test data split(75%-25%)

- Start with default model parameters
- Hyperparameter tuning
- Measure scores on training & test data

- Model testing
- Compare models

LDA top 15 words of each topic

Latent Dirichlet allocation is one of the most popular methods for performing topic modeling. Each document consists of various words and each topic can be associated with some words. The aim behind the LDA to find topics that the document belongs to, on the basis of words contains in it. LDA depends on the frequency of words and the topics were selected according to how much these words were presented.

```
THE TOP 15 WORDS FOR TOPIC #0
['foodthe', 'hasebul', 'peace', 'scene', 'catering', 'bloody', 'cutlery', 'straw', 'yamuna', 'yuck', 'panini', 'wrost', 'salty', 'spicy', 'super']


THE TOP 15 WORDS FOR TOPIC #1
['gobind', 'thank', 'sabir', 'guy', 'govind', 'briyani', 'delivery', 'food', 'job', 'superb', 'service', 'tasty', 'awesome', 'excellent', 'nice']


THE TOP 15 WORDS FOR TOPIC #2
['ghouse', 'friend', 'experience', 'awesome', 'shah', 'visit', 'place', 'nice', 'great', 'taste', 'sizzler', 'thank', 'food', 'service', 'good']


THE TOP 15 WORDS FOR TOPIC #3
['money', 'taste', 'oily', 'worst', 'gud', 'time', 'food', 'maggi', 'waste', 'receive', 'late', 'order', 'quantity', 'deliver', 'bad']


THE TOP 15 WORDS FOR TOPIC #4
['nice', 'like', 'biryani', 'visit', 'try', 'ambience', 'time', 'great', 'chicken', 'taste', 'service', 'order', 'place', 'food', 'good']
```

NNMF top 15 words of each topic

Non-negative Matrix Factorization (NNMF) or the positive matrix analysis is another NLP technique for topic modeling. NNMF differs from LDA because it depends on creating two matrices from random numbers. The first matrix represents the relationship between words and topic while the second matrix represents the relationship between the topic and documents that forms the mathematical basis for categorizing texts as happened in LDA.NNMF is faster and more accurate than LDA because NNMF selects random correlation values between words and topics and training is run based in words exist or not which enable for adjusting weights as the training repeated. NNMF is more favorable for its dimension reduction.

```
THE TOP 15 WORDS FOR TOPIC #0
['boy', 'polite', 'test', 'quantity', 'price', 'quality', 'ambiance', 'spicy', 'ambience', 'burger', 'job', 'food', 'taste', 'service', 'good']


THE TOP 15 WORDS FOR TOPIC #1
['excellent', 'serve', 'try', 'friend', 'amazing', 'love', 'time', 'awesome', 'staff', 'visit', 'ambience', 'great', 'service', 'place', 'food']


THE TOP 15 WORDS FOR TOPIC #2
['music', 'sarvice', 'ambiance', 'service', 'overall', 'hangout', 'family', 'thank', 'enjoy', 'staff', 'ambience', 'place', 'friend', 'friendly', 'nice']


THE TOP 15 WORDS FOR TOPIC #3
['zomato', 'person', 'thank', 'awesome', 'guy', 'excellent', 'super', 'order', 'boy', 'quick', 'late', 'deliver', 'fast', 'time', 'delivery']


THE TOP 15 WORDS FOR TOPIC #4
['piece', 'spicy', 'try', 'paneer', 'veg', 'restaurant', 'like', 'quality', 'rice', 'quantity', 'bad', 'biryani', 'taste', 'order', 'chicken']
```

Logistic Regression

Logistic Regression utilizes the power of regression to do classification One of the main reasons for the model's success is its power of explainability.

```
test score
**********************************************
        The accuracy is  0.8288469264764966
        The precision is  0.8121277747698972
        The recall is  0.9499683343888538
        The f1 is  0.8756567425569177
        the auc  is  0.7843248265350862

                                classification report
**********************************************
              precision    recall  f1-score   support

           0       0.88      0.62      0.73       910
           1       0.81      0.95      0.88      1579

    accuracy                           0.83      2489
   macro avg       0.84      0.78      0.80      2489
weighted avg       0.84      0.83      0.82      2489
```

Parameters:

C: 0.1,

max_iter: 1000,

penalty: 'l2'

Decision Tree

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

```
test score
************************************************
        The accuracy is  0.7561269586179189
        The precision is  0.8301630434782609
        The recall is  0.7739075364154528
        The f1 is  0.8010488364470666
        the auc  is  0.749591130845089

                                classification report
************************************************
              precision    recall  f1-score   support

           0       0.65      0.73      0.69       910
           1       0.83      0.77      0.80      1579

    accuracy                           0.76      2489
   macro avg       0.74      0.75      0.74      2489
weighted avg       0.76      0.76      0.76      2489
```

Parameters:

max_depth=10,

max_leaf_nodes=50,

criterion='entropy'

Random Forest

Random forest is a supervised learning algorithm. The “forest” it builds is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

```
test score
************************************************
        The accuracy is  0.7687876758204957
        The precision is  0.7363379396984925
        The recall is  0.9900760135135135
        The f1 is  0.8445605187319885
        the auc  is  0.6874161672551078

                                classification report
 ************************************************
              precision    recall  f1-score   support

           0       0.96      0.38      0.55      2729
           1       0.74      0.99      0.84      4736

    accuracy                           0.77      7465
   macro avg       0.85      0.69      0.70      7465
weighted avg       0.82      0.77      0.74      7465
```

Parameters:

'criterion': 'entropy',

'max_depth': 7,

'n_estimators': 100

XGBoost

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.XGBoost works with large, complicated datasets. XGBoost is an ensemble modelling technique.

```
test score
************************************************
        The accuracy is  0.827641623141824
        The precision is  0.8438995215311005
        The recall is  0.8936035465484484
        The f1 is  0.868040602891418
        the auc  is  0.8033951798676309
```

classification report

	precision	recall	f1-score	support
0	0.79	0.71	0.75	910
1	0.84	0.89	0.87	1579
accuracy			0.83	2489
macro avg	0.82	0.80	0.81	2489
weighted avg	0.83	0.83	0.83	2489

Parameters:

'criterion': 'entropy',

'max_depth': 7,

'n_estimators': 125

LightGBM

LightGBM is a gradient boosting classifier in machine learning that uses tree-based learning algorithms. It is designed to be distributed and efficient with faster drive speed and higher efficiency, lower memory usage and better accuracy.

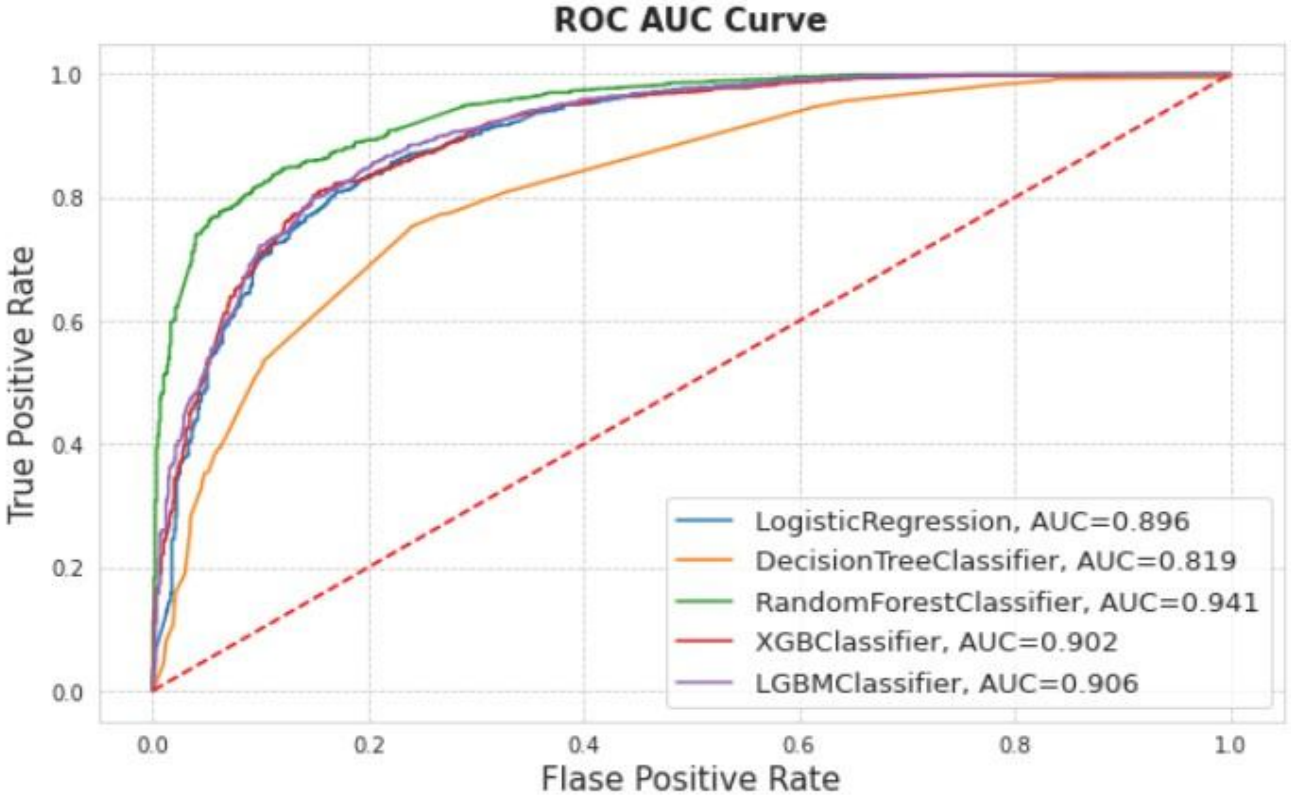
```
test score
**************************************************
        The accuracy is  0.8392928887103255
        The precision is  0.8549066827212523
        The recall is  0.8993033565547816
        The f1 is  0.8765432098765432
        the auc  is  0.8172340958598084

                                classification report
 **************************************************
               precision    recall  f1-score   support

           0       0.81      0.74      0.77       910
           1       0.85      0.90      0.88      1579

    accuracy                           0.84      2489
   macro avg       0.83      0.82      0.82      2489
weighted avg       0.84      0.84      0.84      2489
```

Parameters:

'max_depth': 20,

'n_estimators': 100

ROC-AUC Curve

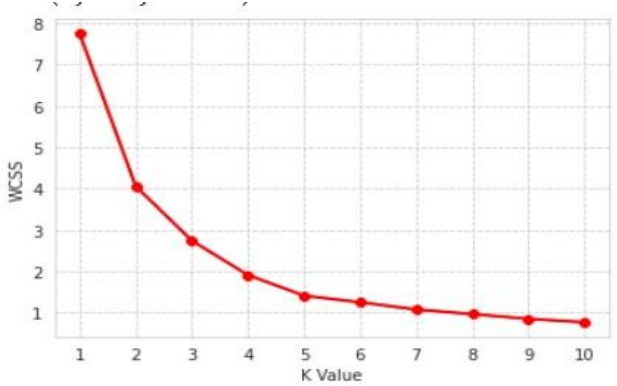


Performance Metrics Summary

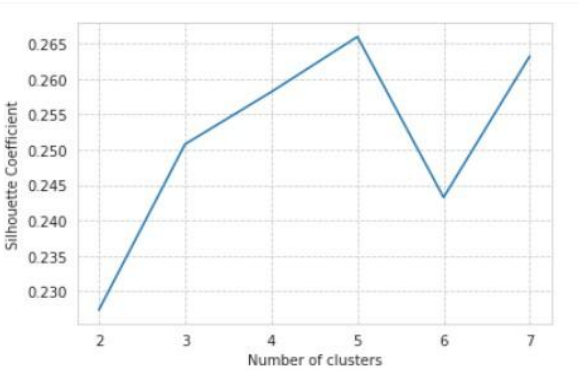
	Models	accuracy	precision	recall	f1	roc_auc
0	Logistic Regrestion	0.828847	0.812128	0.949968	0.875657	0.784325
1	Desision Tree	0.756127	0.830163	0.773908	0.801049	0.749591
2	Random forest	0.768788	0.736338	0.990076	0.844561	0.687416
3	XGboost	0.827642	0.843900	0.893604	0.868041	0.803395
4	LightGBM	0.839293	0.854907	0.899303	0.876543	0.817234

K-Means Clustering Plot

Sum of square elbow plot

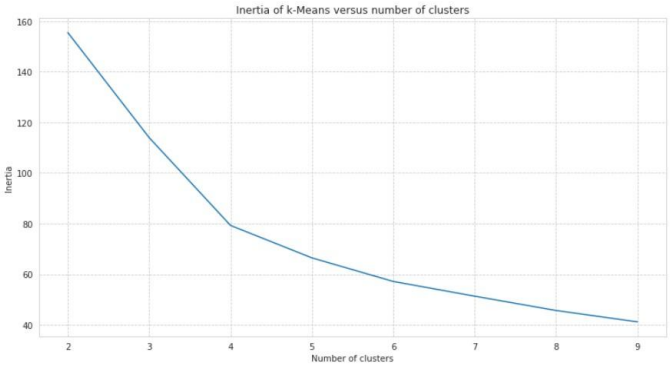


Silhouette Plot

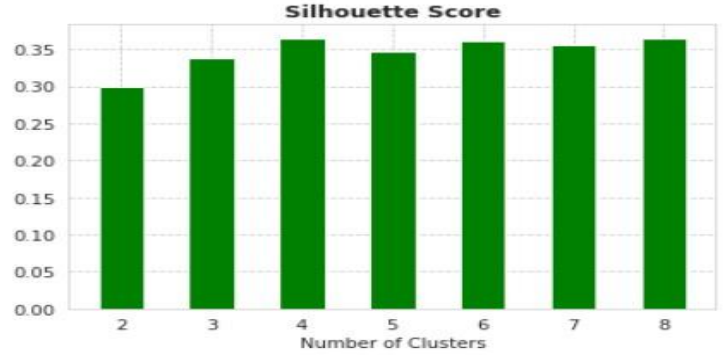


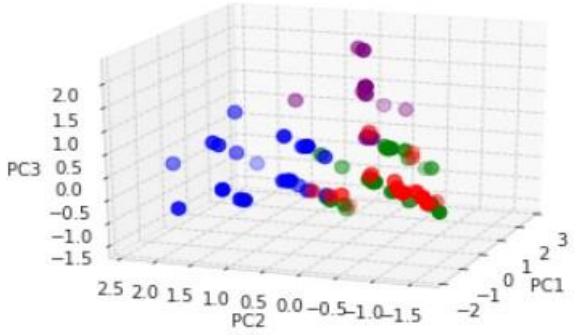
Principal Component Analysis

Sum of square elbow plot

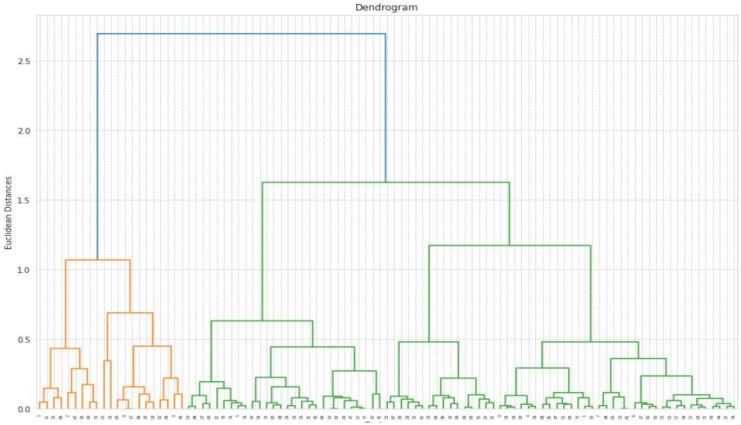


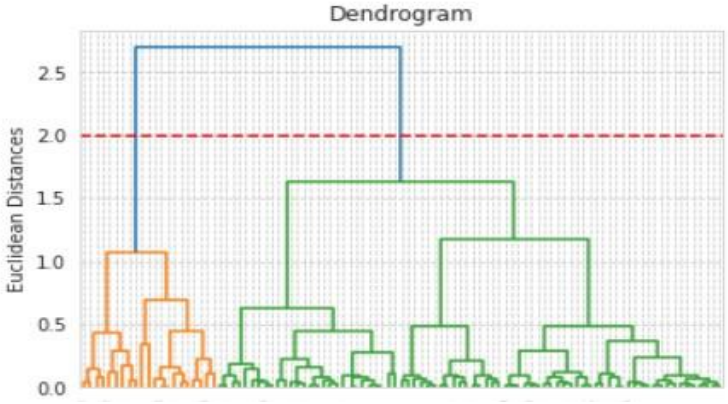

```
Parameter: {'n_clusters': 2} Score 0.2997951415660126
Parameter: {'n_clusters': 3} Score 0.3380779519988871
Parameter: {'n_clusters': 4} Score 0.3650580517654094
Parameter: {'n_clusters': 5} Score 0.34667741109516814
Parameter: {'n_clusters': 6} Score 0.3604063459503715
Parameter: {'n_clusters': 7} Score 0.3561844373513979
Parameter: {'n_clusters': 8} Score 0.3649019696274996
Text(0.5, 0, 'Number of Clusters')
```



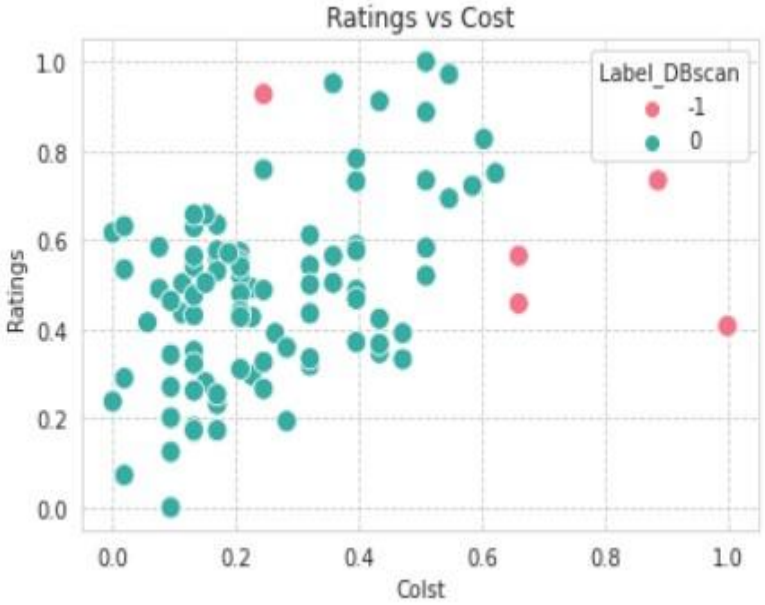


Hierarchical Clustering





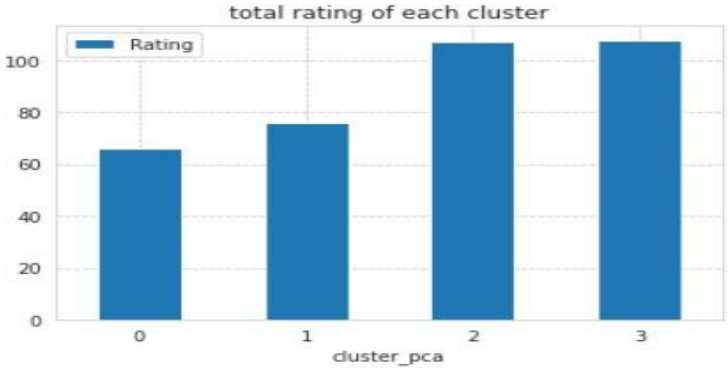
DBSCAN

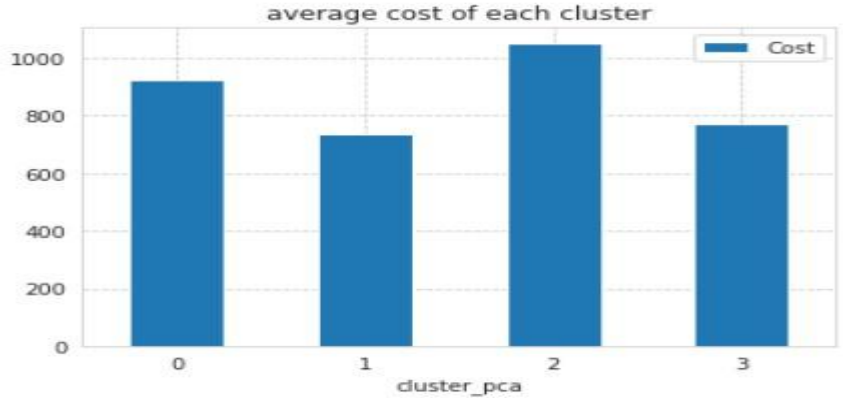


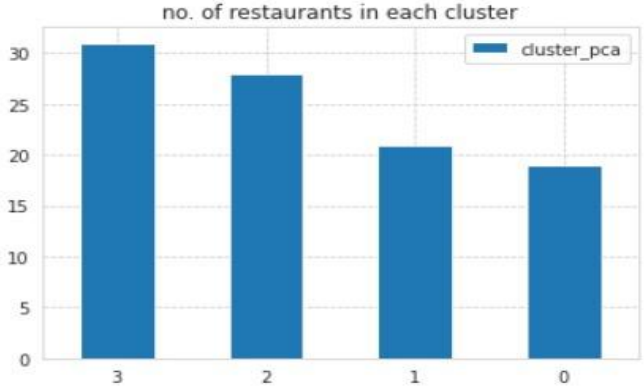
Clustering Summary Table

SL No.	Model_Name	Data	Optimal_Number_of_cluster
1	K-Means with elbow method	RC	range between 2 - 7
2	K-Means with silhouette_score	RC	5
3	PCA (3 components) with elbow method	RC	range between 3 - 7
4	PCA (3 components) with silhouette_score	RC	4
5	Hierarchical clustering with dendogram	RC	2
6	Hierarchical clustering with silhouette_score	RC	2
7	DBSCAN	RC	2

Cluster Analysis

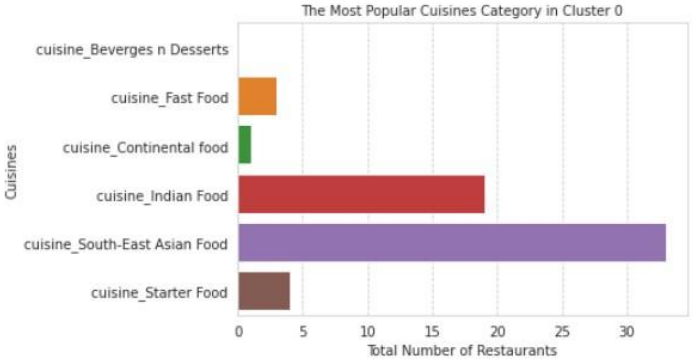


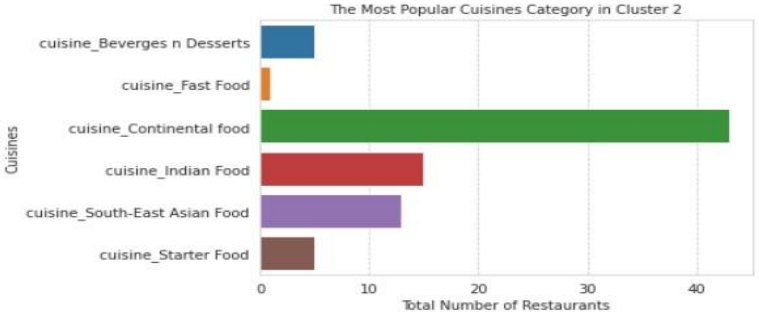


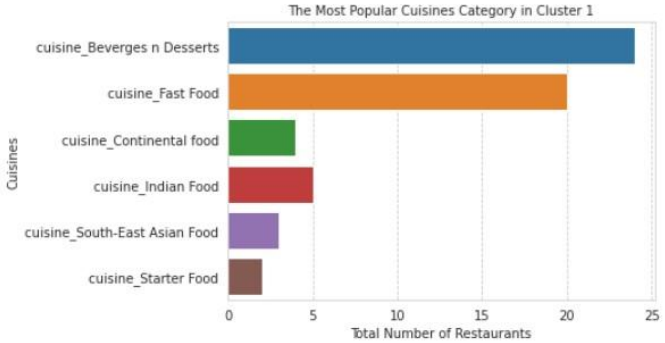


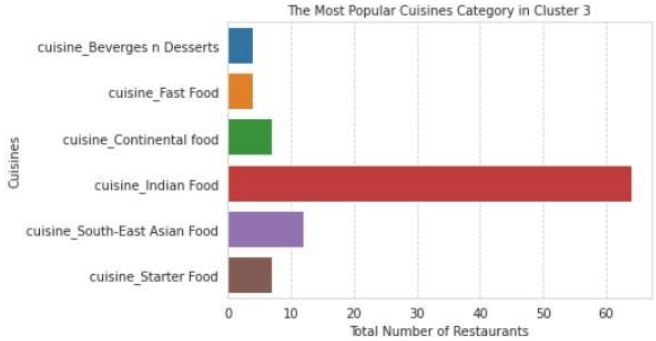
- Both the clusters 2 and 3 restaurants were given high ratings
- Clusters 2 restaurants were the most expensive ones. On the other hand, Cluster 1 restaurants were the most affordable ones.
- Clusters 3 were having pretty high numbers of restaurants in its cluster comparatively.

Cluster Analysis









Conclusion

- The most popular cuisines available in the most of the restaurant is the North Indian and Chinese.
- The Restaurant with the highest rating of nearly 4.8 and good reviews is the AB's Absolute Barbecues. On the contrary, the restaurant with worst reviews and rating is the Hotel Zara Hifi with a rating less than 2.5.
- Namit Agarwal is one of the active reviewer based on number of followers and ratings provided
- We have got optimal number of clusters as 4 clusters in PCA.
- Best no of cluster for sentiment analysis (unsupervised) is 2 i.e. for positive and negative reviews
- The best model accuracy for sentiment analysis we found out to be LightGBM and Logistic Regression model.
- The most frequent working hours of the restaurant is between 11am to 11pm.

Challenges

- Extracting new features from existing features like Metadata and Timings features in review dataset were a bit tedious job to do.
- Handling null values in both restaurant and review dataset and the text pre-processing in review feature was a challenging task.
- There were few undefined data records present in Collection feature and few duplicate records in review dataset.
- Finding Optimum number of clusters
- Selecting different parameter for hyperparameter tuning and finding the best parameters has to follow a trial n error technique,which is again a challenging job.

Thank You !!!