Capstone Project 4

Zomato Restaurant Clustering and Sentiment Analysis

Individual Project:

Name : Mayank Ghai

Email:mayankghai1195@gmail.com

Content

- Problem Statement
- Business Problem Analysis
- Data Summary
- Methodology
- Exploratory Data Analysis
- Restaurant Clustering
- Sentiment Analysis
- Conclusion and Recommendations
- Challenges

Problem Statement

The Project focuses on analyzing the Zomato restaurant data. 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 Analysis also solves 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.

Business Problem Analysis

- To assure Zomato's success it is important for the company to analyze its datasets and make appropriate strategic decisions.

- The problem statement here asks us to cluster the restaurants to help customers find the best restaurants in their city and according to their taste and requirement. This will help Zomato in building a good recommendation system for their customers. Do a cost-benefit analysis using the cuisines and costs of the restaurants.

- It is important to do sentiment analysis to get an idea about how people really feel about a particular restaurant and understand the fields they are lagging in. To identify the industry critics and especially work on their reviews to build a reputation worth praising.

Data Summary

Restaurant Names and Metadata

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

Restaurant Reviews

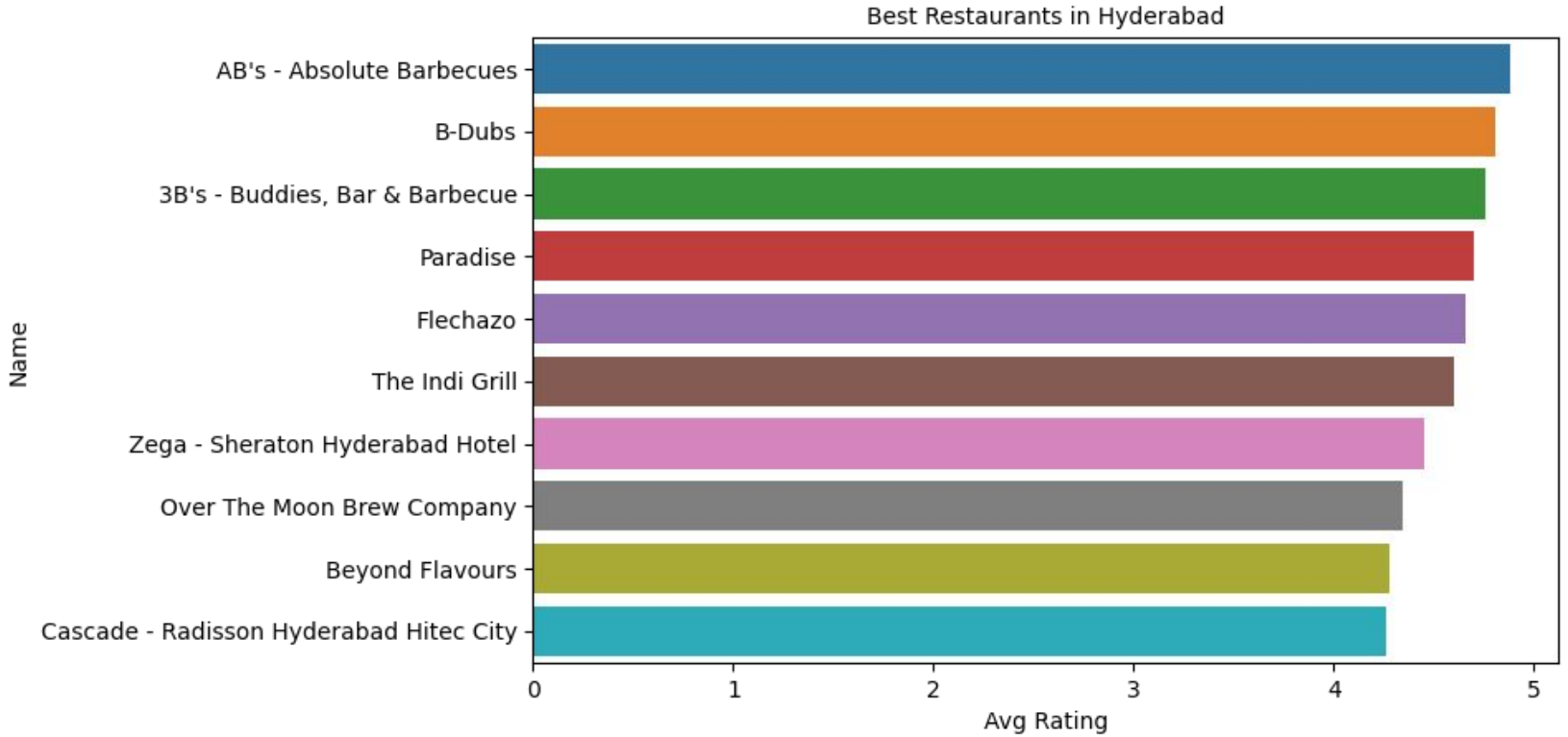
1. Restaurant : Name of the Restaurant
2. Reviewer : Name of the Reviewer
3. Review : Review Text
4. Rating : Rating Provided by Reviewer
5. MetaData : Reviewer Metadata - No. of Reviews and followers
6. Time: Date and Time of Review
7. Pictures : No. of pictures posted with review

Methodology

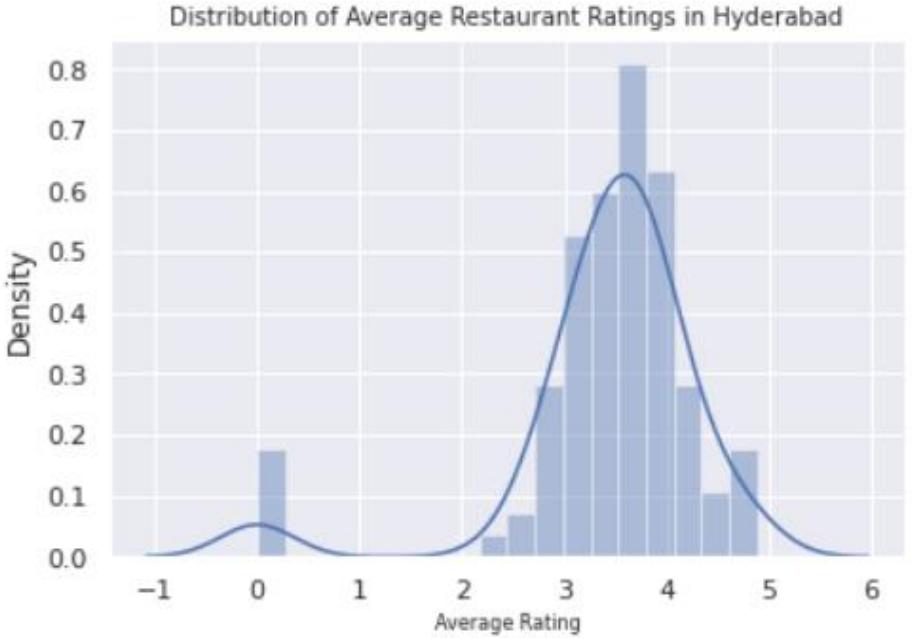
- Business Problem Analysis
- Data Collection
- Data Cleaning and Preprocessing
- Feature Engineering
- Exploratory Data Analysis
  - Best Restaurants in the City
  - The Most Popular Cuisines in Hyderabad
  - Restaurants and their Costs etc
  - Cost-Benefit Analysis
  - Hypotheses Generation on visualized data for Clustering
- Restaurant Clustering
  - K means Clustering on Cost and Ratings
  - Multi-Dimensional K means Restaurant Clustering
    - Principal Component Analysis
    - Silhouette Score
    - K Means Clustering
    - Cluster Exploration
- Sentiment Analysis
  - Exploratory Data Analysis
    - Critics in the Industry
  - Text Pre-Processing and Text Visualization
  - Modeling
- Conclusion

Exploratory Data Analysis

Best Restaurants in the City

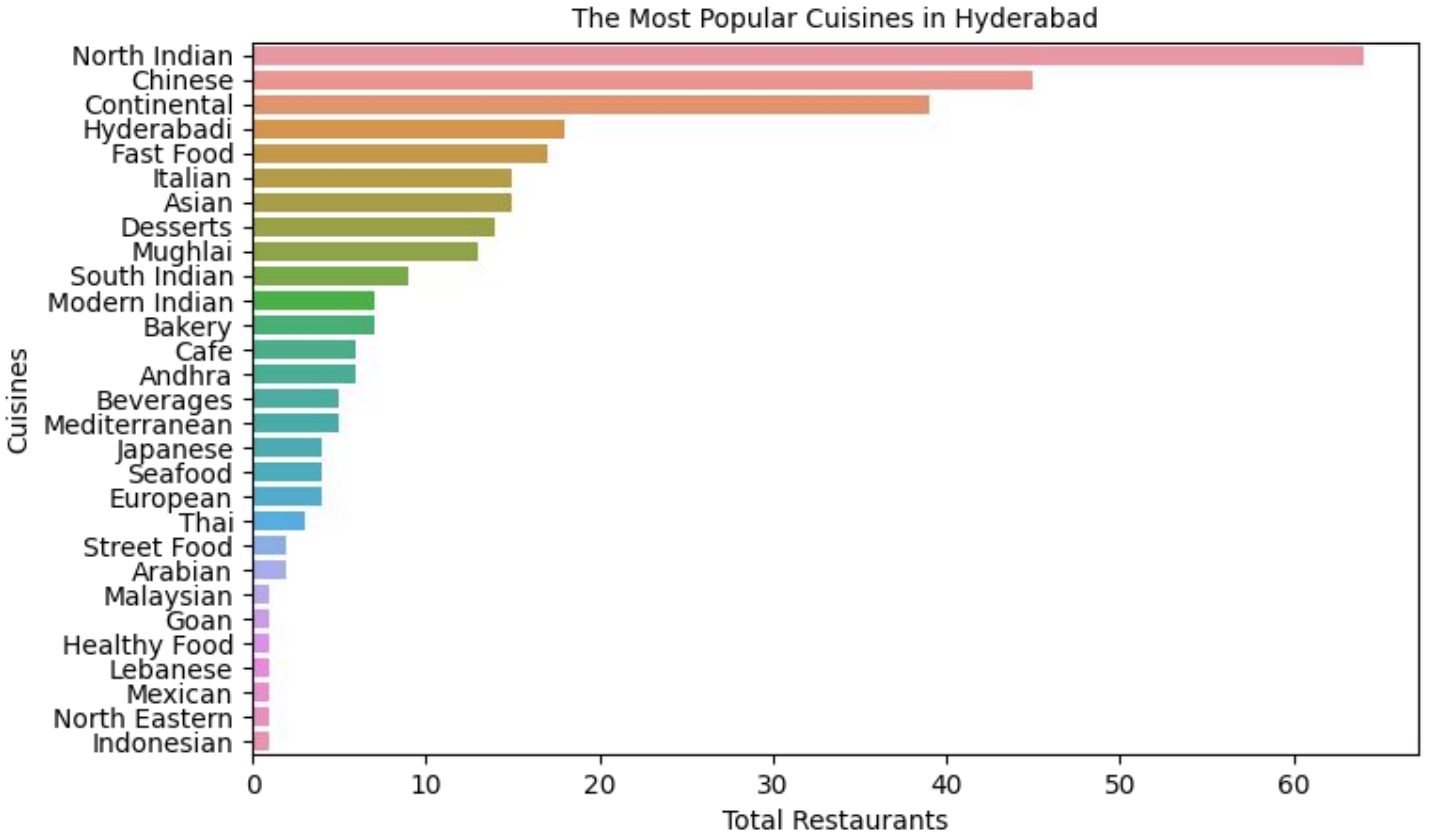


Distribution of Average Restaurant Rating in Hyderabad



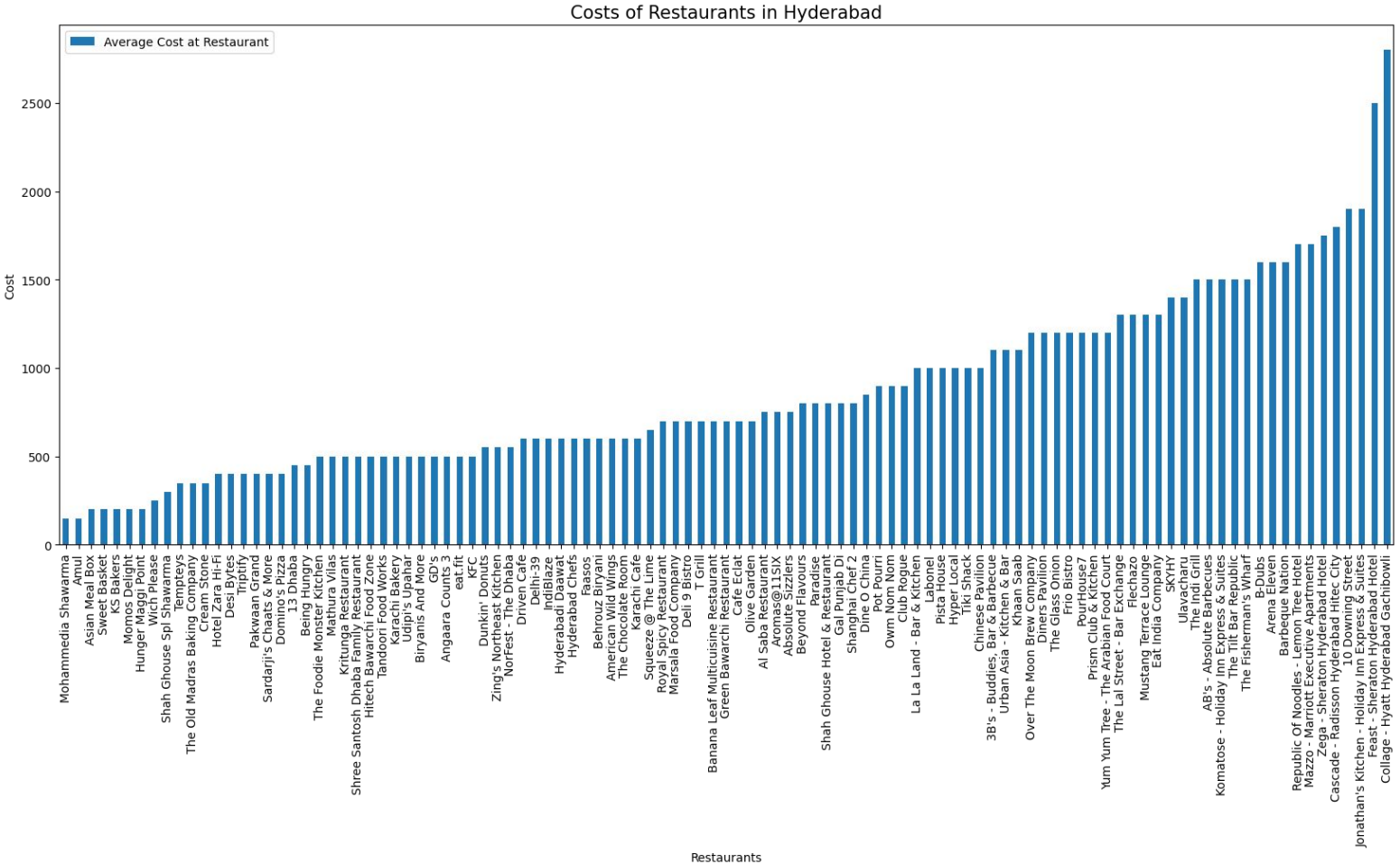
Few restaurants in the original restaurant dataset have not been rated by the people yet, most restaurants have ratings between 3.5 and 4. Efforts should be made by the company to improve the existing restaurants by pushing them to act on the reviews and to include restaurants with better services in the future to improve overall rating distribution.

The Most Popular Cuisines in Hyderabad

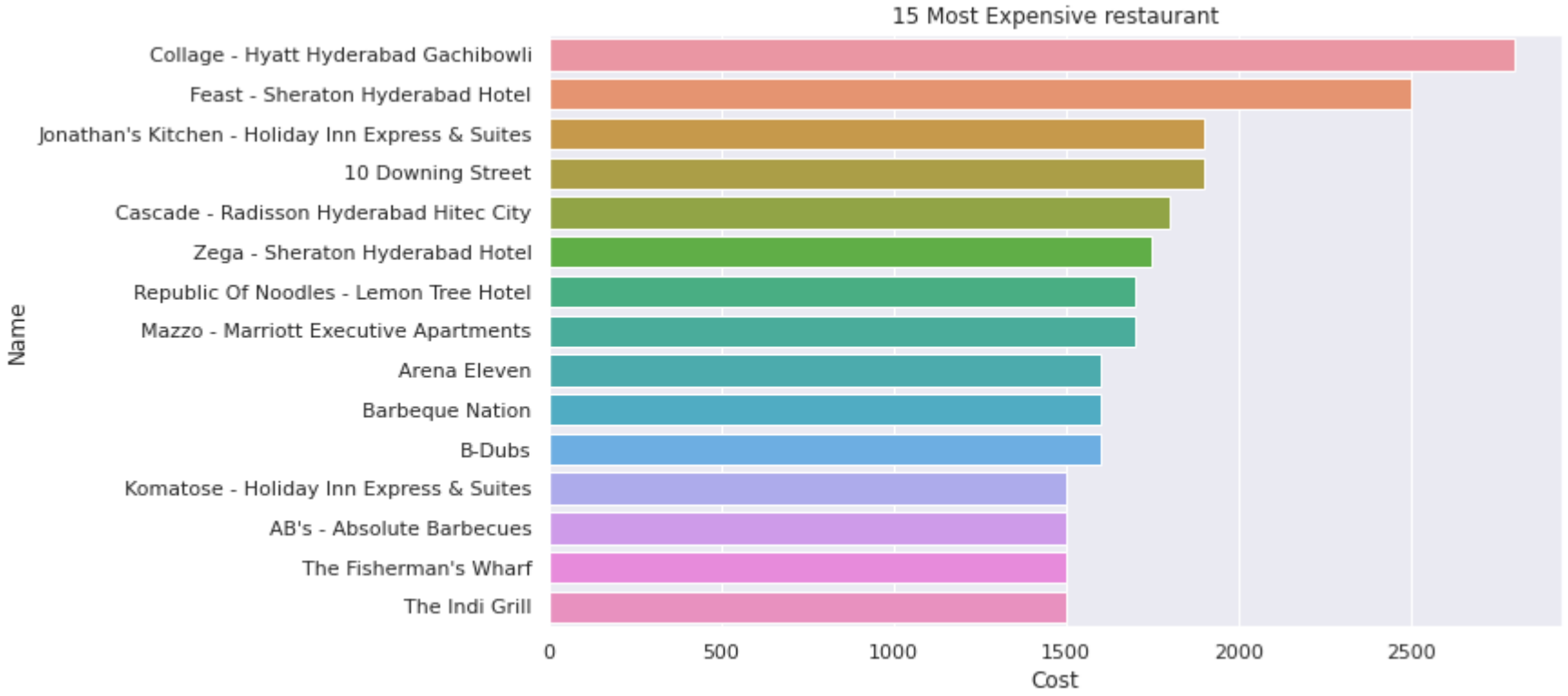


Although located in South India, North Indian food is dominating in the restaurants followed by Chinese, and Continental. The number of cuisines shows the diverse food options available in Hyderabad.

Restaurants and their Costs

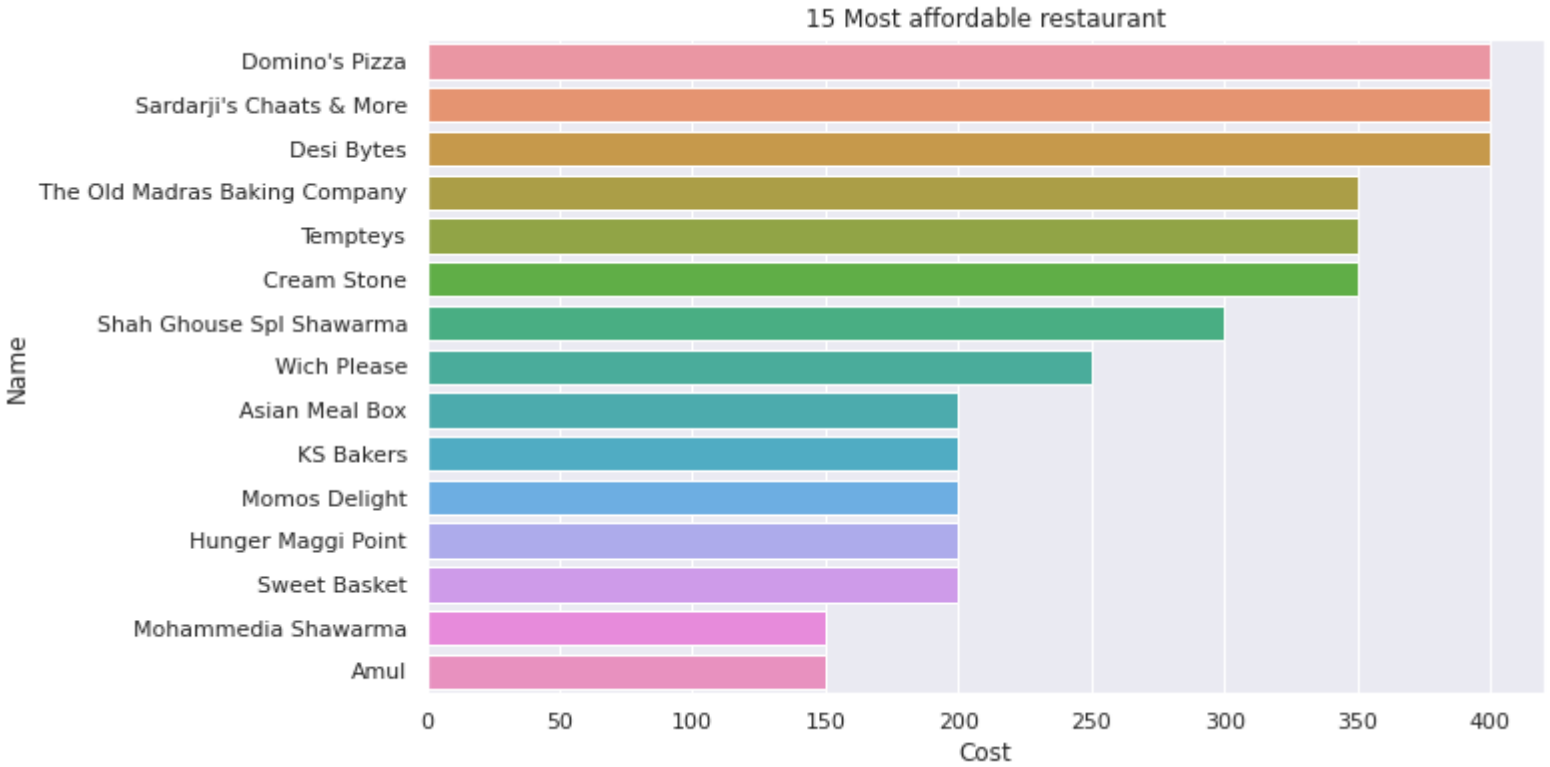


15 Most expensive Restaurants



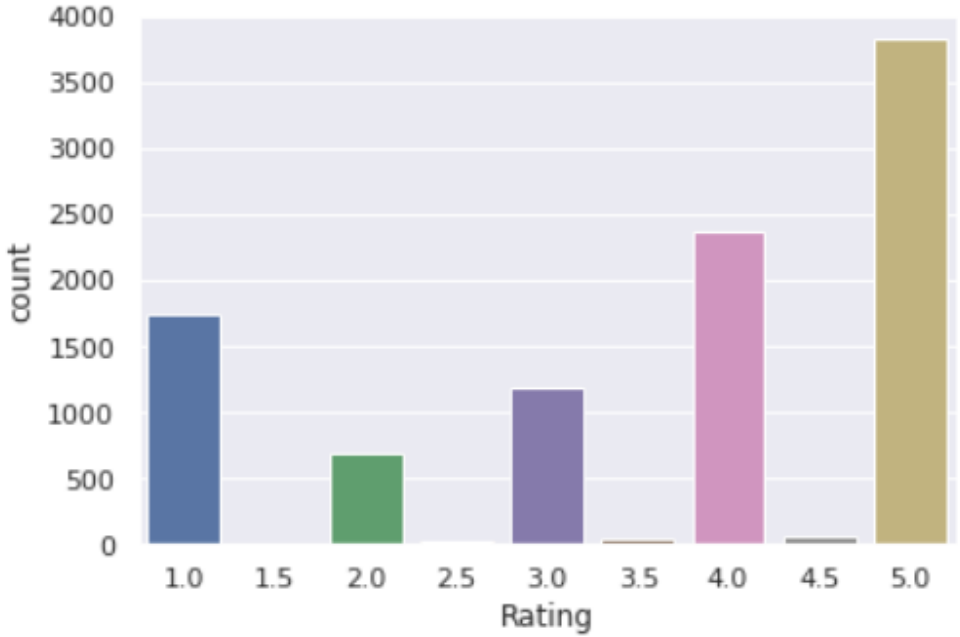
The most expensive restaurants in the dataset are restaurants by 4 star above hotels

15 most Affordable Restuarents



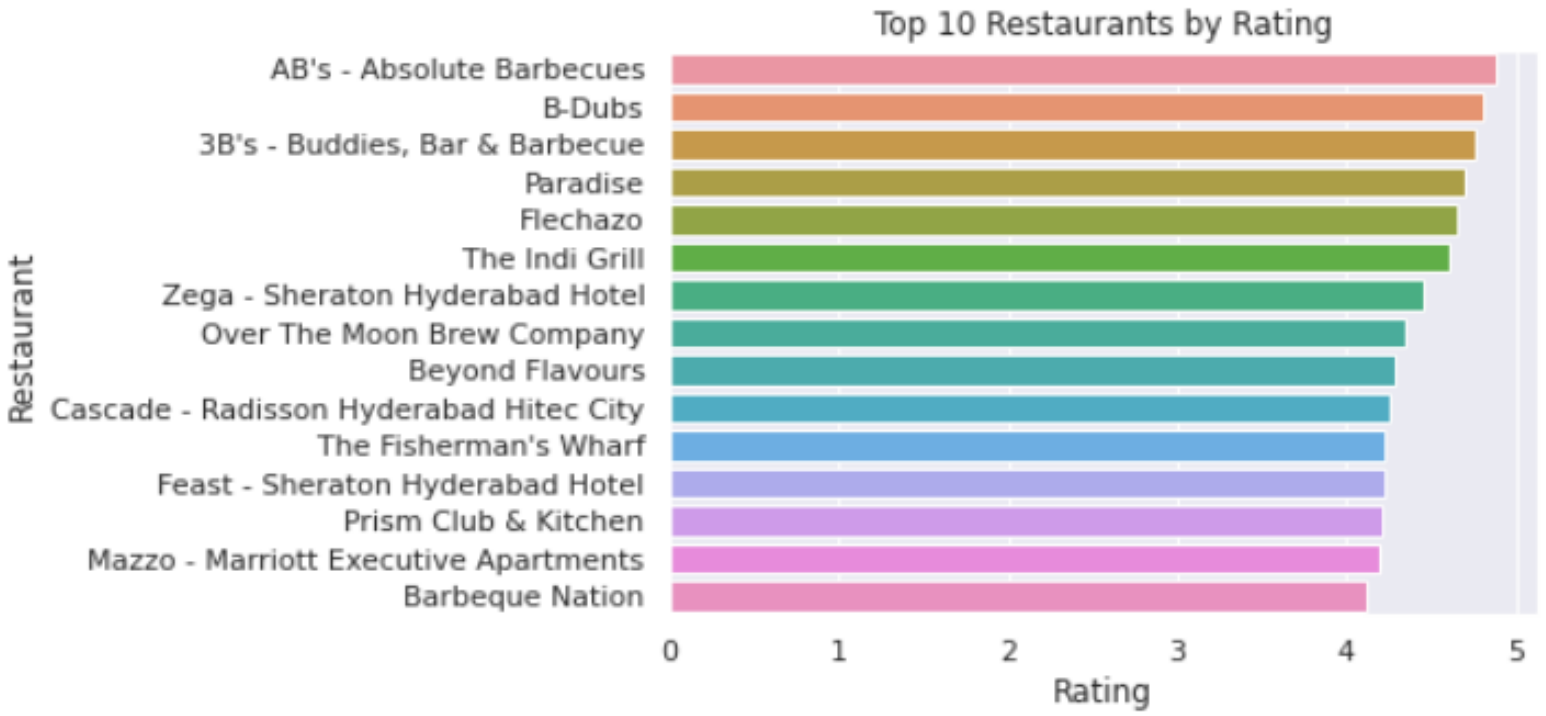
The cheapest restaurants in the dataset are basically small food joints and bakeries.

Rating

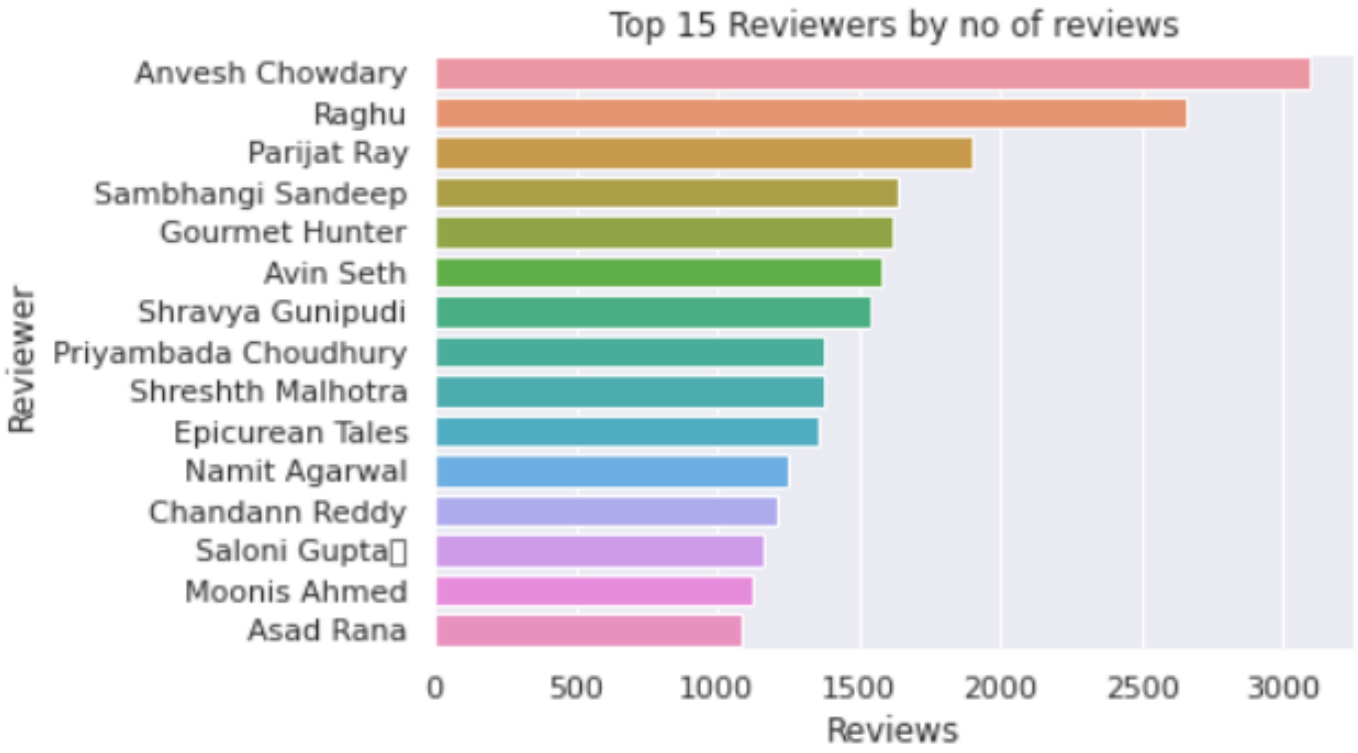


Even if majority ratings are good, we still have considerable count of poor ratings.

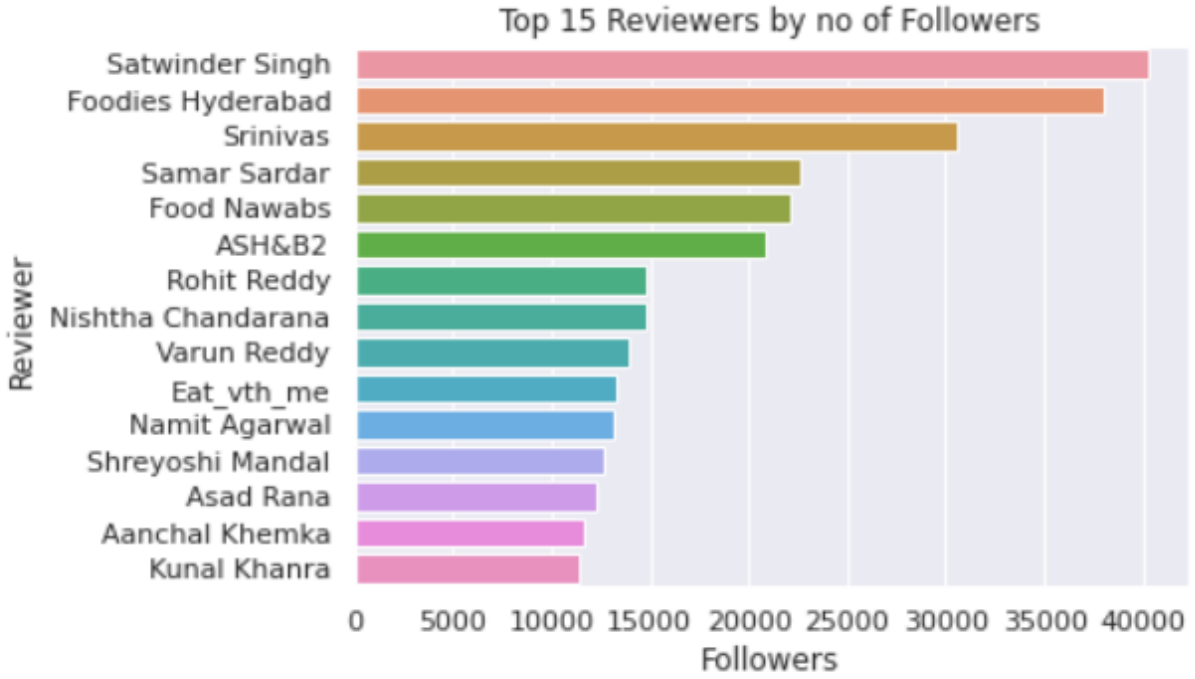
Restaurant by rating



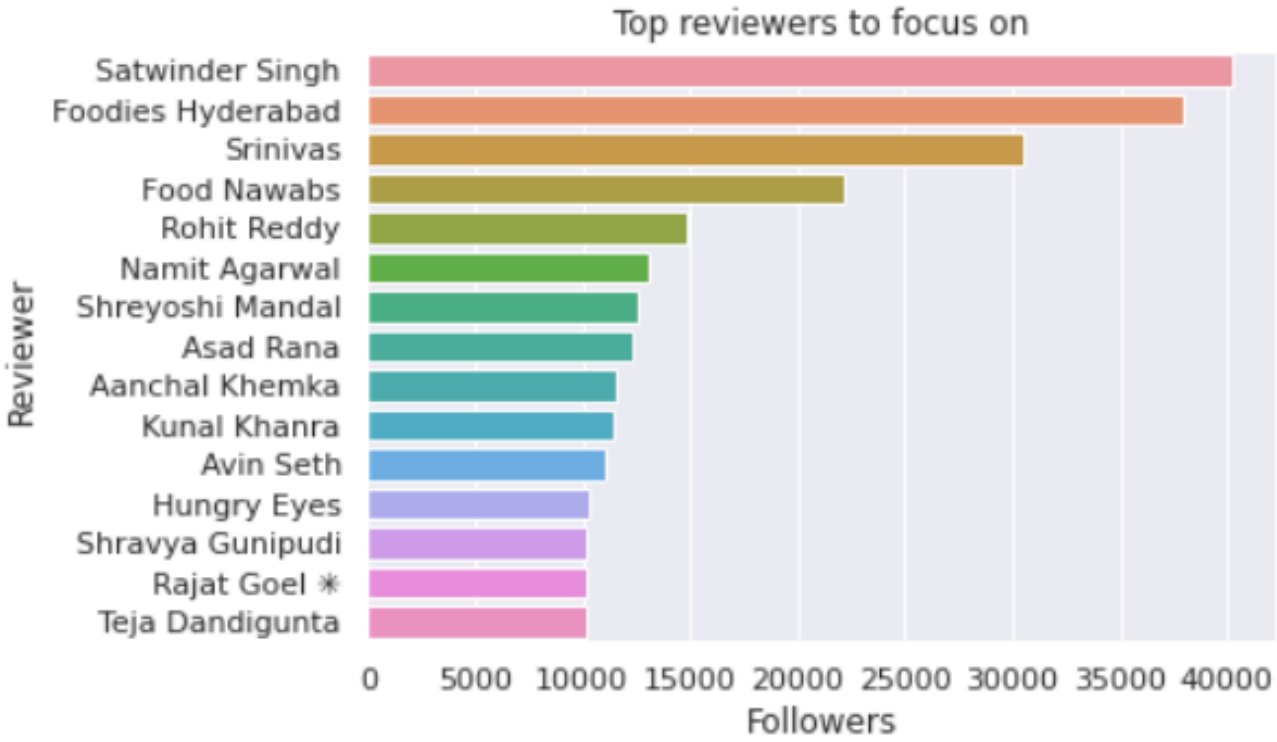
Reviewers by number of reviews



Reviewers by number of Followers



Reviewers to focus on

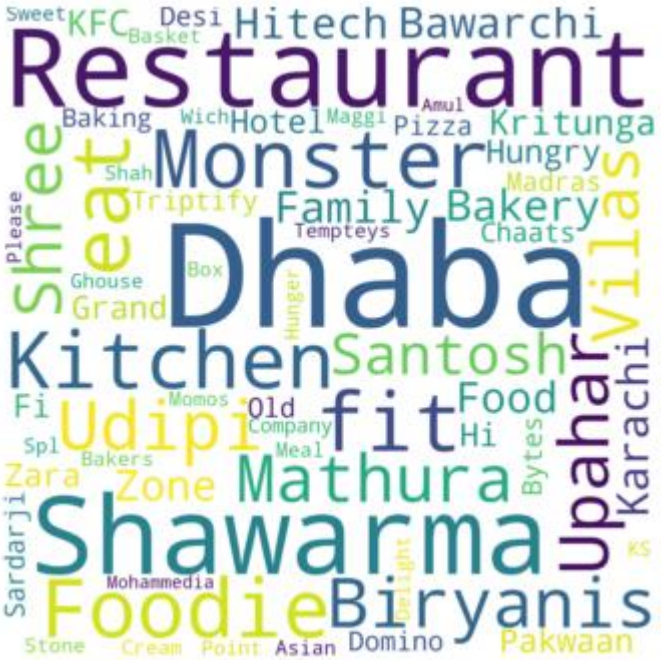


These are the reviewer a restaurant should focus on who have reviewed more than 100 restaurants and have followers greater than 10000 with an average rating above 3.5

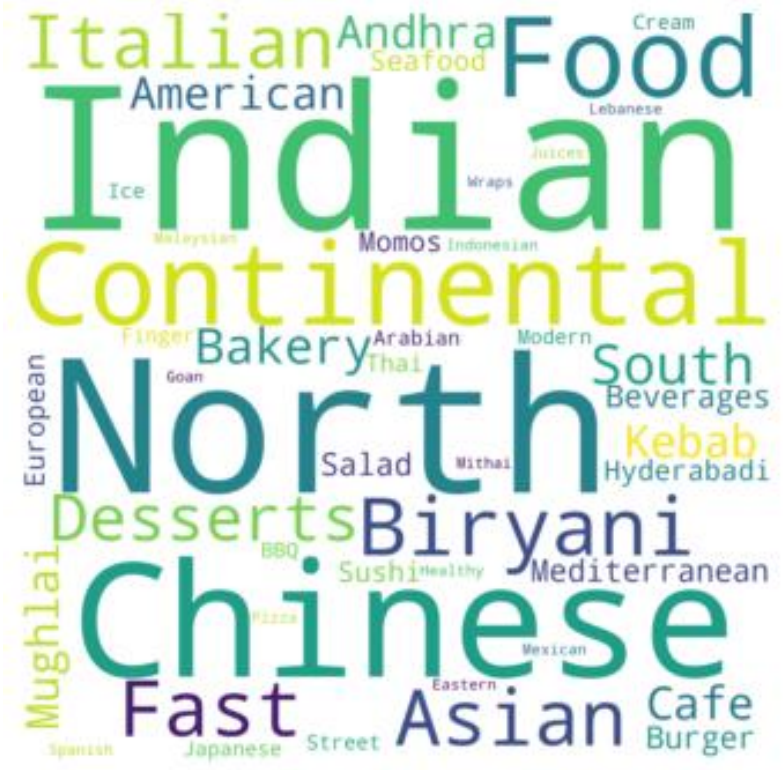
Frequent Keywords Used for Restaurant

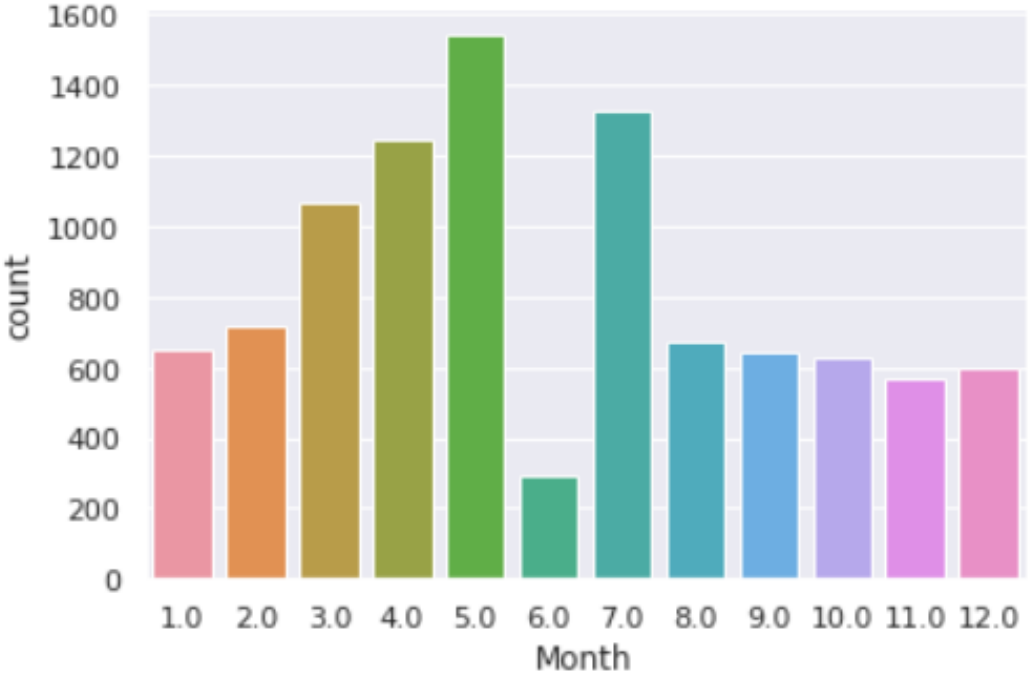
Most Expensive

Most Affordable



Frequent Keyword Used for cuisine





Observations: Most of the reviews are in the month of 5 and 7 months of year

Cost-Benefit Analysis

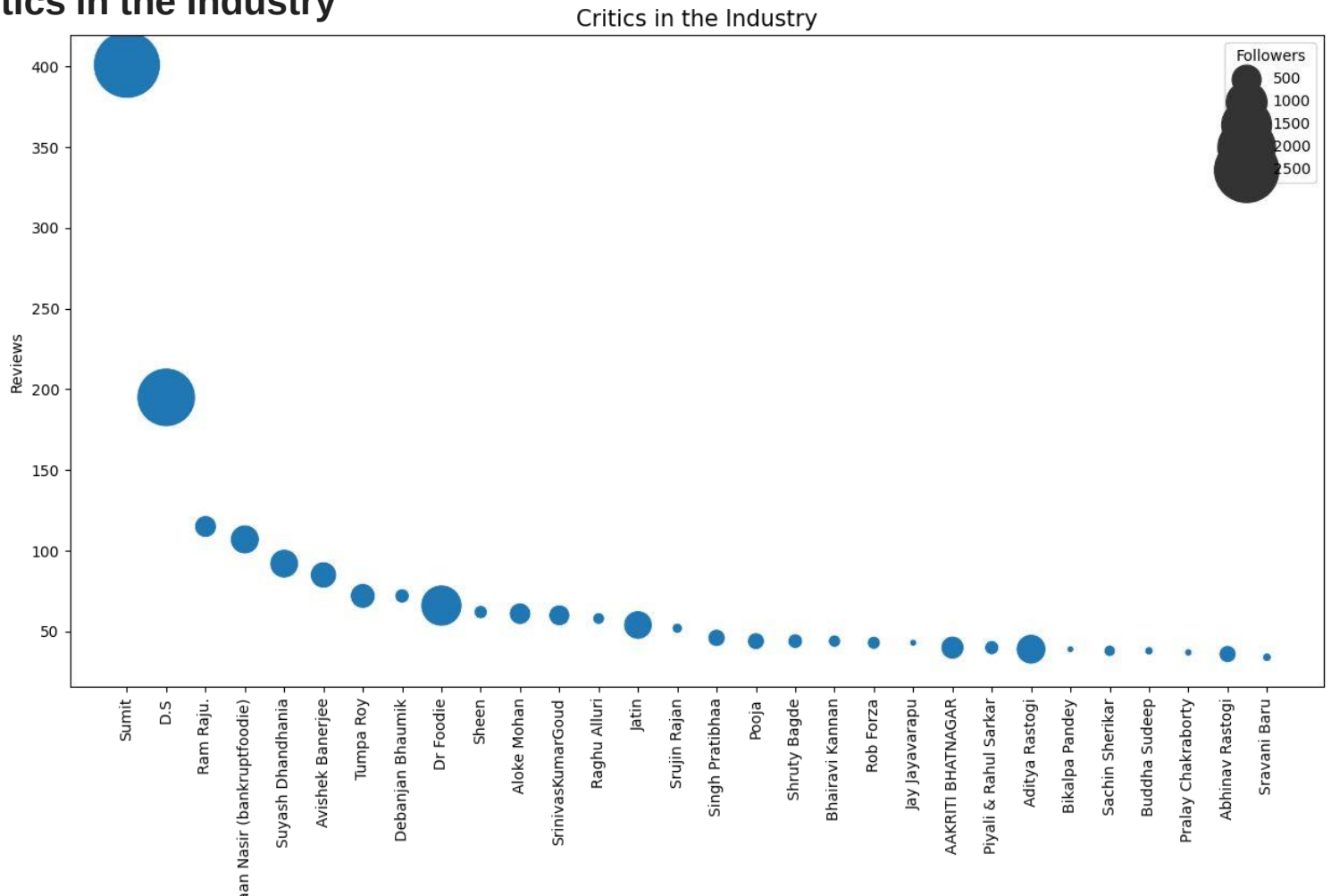
- A Cost-Benefit Analysis is a process of analyzing the worth of a decision by estimating the costs incurred in implementing that decision and comparing them with the benefits of that decision. If the projected benefits outweigh the costs, you'll be making money out of that decision and if not, it's important to strategize a better plan.

- The data that we have consists of per-person cost, cuisines available at the restaurant, and an average rating of the restaurant. If a restaurant isn't performing well in terms of rating and has a high per-person cost and a low number of popular cuisines, this is going to be a problem for Zomato. Since negative reviews would be an intangible cost to the company and with that the company will start to lose daily application users. The application users are an asset to the company, Zomato gets advertising by different restaurants because of the large audience they have.

- All in all, it is important to separate out the restaurants that Zomato needs to work on in order to improve its overall customer experience and if improvement strategies don't work out, they need to delist those restaurants themselves.

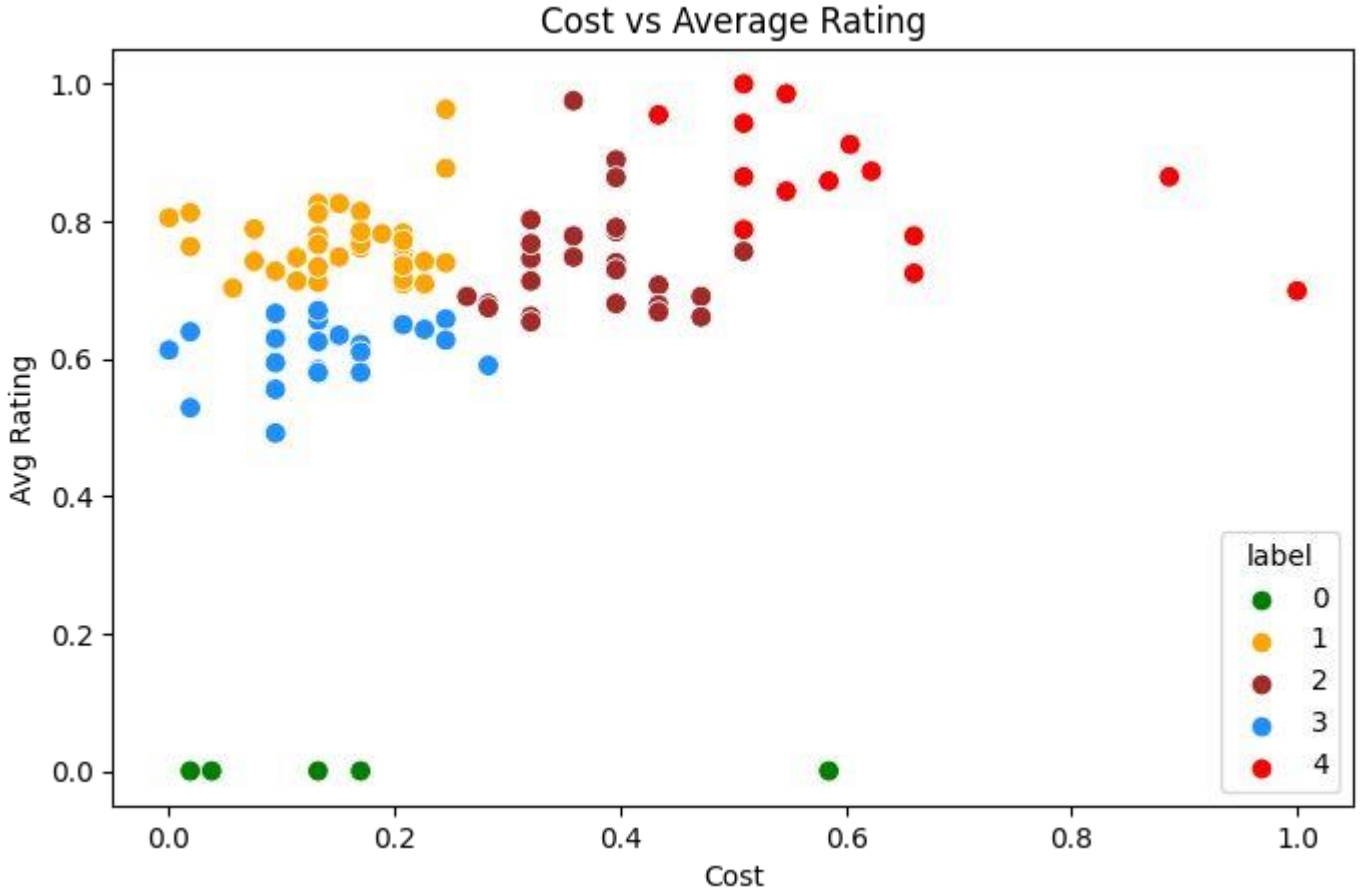
These restaurants are basically small food joints or restaurants with high prices according to the food they are serving. Efforts should be made to advertise more and analyze the reviews, especially for these restaurants, and work on them. Mohammedia Shawarma has the highest rating with the lowest cost. It seems it is doing well in its capacity.

Critics in the Industry



Positive and Negative Word Clouds

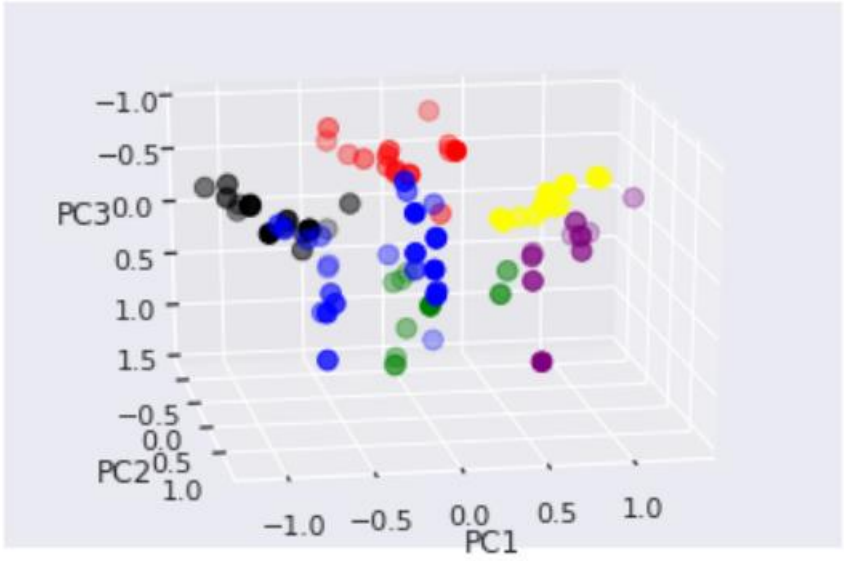
Restaurant Clustering



Cluster 0:
Color: Purple
Cuisines: Fast food and Continental
Average Rating: 3.42
Average Cost: 942 INR
Median Cost: 600 INR

Cluster 1:
Color: Red
Cuisines: North Indian and Complimentary
Average Rating: 3.63
Average Cost: 823 INR

Cluster 2:
Color: Blue
Cuisines: North Indian, Chinese and Continental
Average Rating: 3.77
Average Cost: 1331 INR



Cluster 3:
Color: Green
Cuisines: Chinese, Thai, Asian, Malaysian etc
Average Rating: 3.18
Average Cost: 890 INR

Custer 4:
Color: Yellow
Cuisines: Cafe, Bakeries, Desserts, etc
Average Rating: 3.14
Average Cost: 406 INR

Custer 5:
Color: Black
Cuisines: North Indian, Chine
Hyderabadi
Average Rating: 3.24
Average Cost: 674 INR

Sentiment Analysis:

Evaluation:

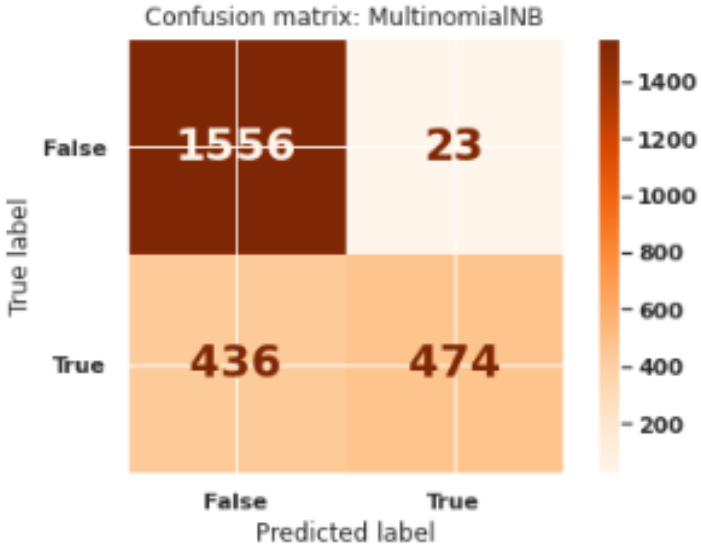
- In the business problem, predicting the negative sentiments correctly is really important but is more important for the models to reduce the number of false positives.

- False positives indicate that the reviews were actually negative but they were categorized as positive and this will lead to missing a complaint to work on.

Models performed

- Multinomial Naïve Bayes
- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost Modeling
- Light GBM

Multinomial Naïve Bayes Metrics

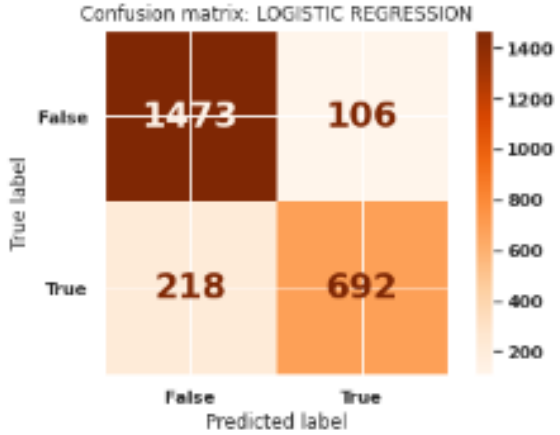



```
Training time: 0.0001min
                                score matrix for train
**************************************************************************************
        The accuracy is  0.8557267247153383
        The precision is  0.9752589182968929
        The recall is  0.6211066324661048
        The f1 is  0.7588985896574882
        the auc  is  0.8060136202871064

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

           0       0.82      0.99      0.90      4736
           1       0.98      0.62      0.76      2729

    accuracy                           0.86      7465
   macro avg       0.90      0.81      0.83      7465
weighted avg       0.88      0.86      0.85      7465



                                score matrix for test
**************************************************************************************
        The accuracy is  0.8155885897950984
        The precision is  0.9537223340040242
        The recall is  0.5208791208791209
        The f1 is  0.673773987206823
        the auc  is  0.7531564698759124

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

           0       0.78      0.99      0.87      1579
           1       0.95      0.52      0.67       910

    accuracy                           0.82      2489
   macro avg       0.87      0.75      0.77      2489
weighted avg       0.84      0.82      0.80      2489
```

Logistic Regression

Parameters :

- C = 10
- Max_iter = 1000
- Penalty = L2

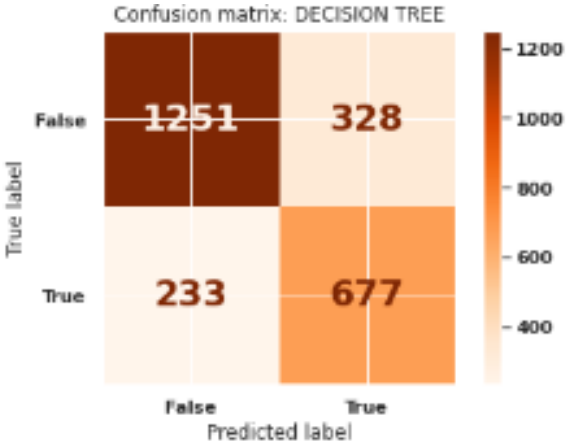



```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
Training time: 0.2493min
The best parameters found out to be : {'C': 10, 'max_iter': 1000, 'penalty': 'l2'}

where negative mean squared error is:  0.7607110931881573

                                score matrix for train
********************************************************************************
        The accuracy is  0.959410582719357
        The precision is  0.954307116104869
        The recall is  0.9336753389519971
        The f1 is  0.9438784960177811
        the auc  is  0.953957601908431

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

           0       0.96      0.97      0.97      4736
           1       0.95      0.93      0.94      2729

    accuracy                           0.96      7465
   macro avg       0.96      0.95      0.96      7465
weighted avg       0.96      0.96      0.96      7465



                                score matrix for test
********************************************************************************
        The accuracy is  0.8581759742868622
        The precision is  0.829585798816568
        The recall is  0.7703296703296704
        The f1 is  0.7988603988603987
        the auc  is  0.8395663551141702

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

           0       0.87      0.91      0.89      1579
           1       0.83      0.77      0.80       910

    accuracy                           0.86      2489
   macro avg       0.85      0.84      0.84      2489
weighted avg       0.86      0.86      0.86      2489
```

Decision Tree

Parameters :

- max_depth = 10
- max_leaf_nodes = 45
- Criterion = 'entropy'

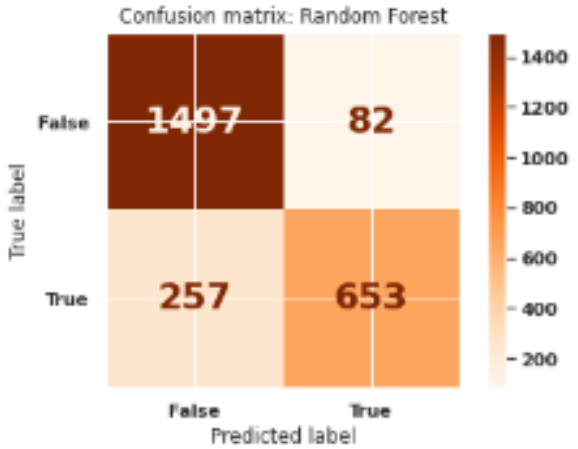



```
Training time: 0.0122min
                              score matrix for train
*********************************************************************************
        The accuracy is  0.7959812458137977
        The precision is  0.695906432748538
        The recall is  0.7849028948332722
        The f1 is  0.7377303254692612
        the auc  is  0.7936338798490685

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

           0       0.87      0.80      0.83      4736
           1       0.70      0.78      0.74      2729

    accuracy                           0.80      7465
   macro avg       0.78      0.79      0.79      7465
weighted avg       0.80      0.80      0.80      7465



                              score matrix for test
*********************************************************************************
        The accuracy is  0.7746082764162314
        The precision is  0.6736318407960199
        The recall is  0.743956043956044
        The f1 is  0.7070496083550915
        the auc  is  0.768114817418174

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

           0       0.84      0.79      0.82      1579
           1       0.67      0.74      0.71       910

    accuracy                           0.77      2489
   macro avg       0.76      0.77      0.76      2489
weighted avg       0.78      0.77      0.78      2489
```

Random Forest

Parameters :

- max_depth = 15
- N_estimators = 150
- Criterion = 'entropy'

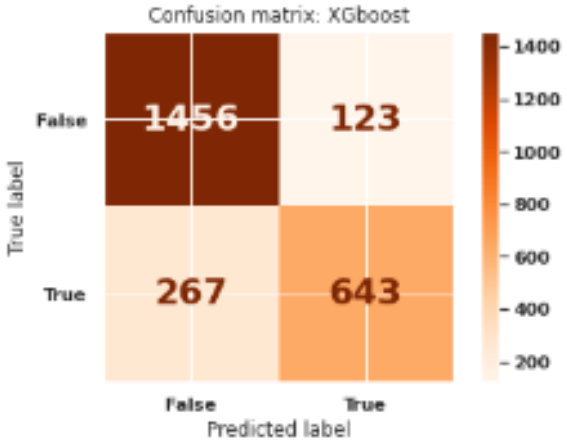



```
Fitting 5 folds for each of 9 candidates, totalling 45 fits
Training time: 0.6742min
The best parameters found out to be : {'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 150}

where negative mean squared error is:  0.24285714285714283

                                score matrix for train
**********************************************************************************
        The accuracy is  0.7657693852952994
        The precision is  1.0
        The recall is  0.35934065934065934
        The f1 is  0.5286984640258691
        the auc  is  0.6796703296703297

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

           0       0.73      1.00      0.84      1579
           1       1.00      0.36      0.53       910

    accuracy                           0.77      2489
   macro avg       0.87      0.68      0.69      2489
weighted avg       0.83      0.77      0.73      2489



                                score matrix for test
**********************************************************************************
        The accuracy is  0.7148024112525118
        The precision is  0.9838709677419355
        The recall is  0.22352510076951265
        The f1 is  0.3642878471185428
        the auc  is  0.6107068071415132

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

           0       0.69      1.00      0.82      4736
           1       0.98      0.22      0.36      2729

    accuracy                           0.71      7465
   macro avg       0.84      0.61      0.59      7465
weighted avg       0.80      0.71      0.65      7465
```

XGBoost Modeling

Parameters :

- max_depth = 15
- N_estimators = 1
- Criterion = 'entropy'

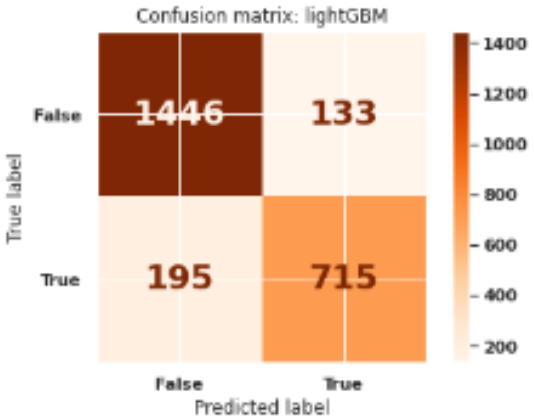



```
Fitting 3 folds for each of 9 candidates, totalling 27 fits
Training time: 4.9473min
The best parameters found out to be : {'criterion': 'entropy', 'max_depth': 15, 'n_estimators': 1

where negative mean squared error is:  0.7449634707060451

                                score matrix for train
*********************************************************************************
        The accuracy is  0.9596784996651038
        The precision is  0.9404934687953556
        The recall is  0.949798460974716
        The f1 is  0.945123062898815
        the auc  is  0.9575850412981688

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

           0       0.97      0.97      0.97      4736
           1       0.94      0.95      0.95      2729

    accuracy                           0.96      7465
   macro avg       0.96      0.96      0.96      7465
weighted avg       0.96      0.96      0.96      7465


                                score matrix for test
*********************************************************************************
        The accuracy is  0.8670148654077943
        The precision is  0.8475390156062425
        The recall is  0.7758241758241758
        The f1 is  0.8100975329890994
        the auc  is  0.847696761756293

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

           0       0.88      0.92      0.90      1579
           1       0.85      0.78      0.81       910

    accuracy                           0.87      2489
   macro avg       0.86      0.85      0.85      2489
weighted avg       0.87      0.87      0.87      2489
```

Light GBM

Parameters :

- max_depth = 25
- N_estimators = 150




```
Fitting 3 folds for each of 9 candidates, totalling 27 fits
Training time: 1.1545min
The best parameters found out to be : {'max_depth': 25, 'n_estimators': 150}

where negative mean squared error is:  0.7665860725266666

                                score matrix for train
********************************************************************************
       The accuracy is  0.9509711989283323
       The precision is  0.9380793474230626
       The recall is  0.9270795163063393
       The f1 is  0.932546995945448
       the auc  is  0.9459088459910074

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

           0       0.96      0.96      0.96      4736
           1       0.94      0.93      0.93      2729

    accuracy                           0.95      7465
   macro avg       0.95      0.95      0.95      7465
weighted avg       0.95      0.95      0.95      7465



                                score matrix for test
********************************************************************************
       The accuracy is  0.8678184009642427
       The precision is  0.8421672555948174
       The recall is  0.7857142857142857
       The f1 is  0.8129619101762364
       the auc  is  0.8504252239211073

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

           0       0.88      0.92      0.90      1579
           1       0.84      0.79      0.81       910

    accuracy                           0.87      2489
   macro avg       0.86      0.85      0.86      2489
weighted avg       0.87      0.87      0.87      2489
```

Evaluation Metrics Table

	Models	accuracy	precision	recall	f1	roc_auc	train_time
0	MultinomialNB	0.815589	0.953722	0.520879	0.673774	0.753156	0.0001
1	Logestic Regrestion	0.858176	0.829586	0.770330	0.798860	0.839566	0.2493
2	Desision Tree	0.774608	0.673632	0.743956	0.707050	0.768115	0.0122
3	Random forest	0.714802	0.983871	0.223525	0.364288	0.610707	0.6742
4	XGboost	0.867015	0.847539	0.775824	0.810098	0.847697	4.9473
5	lightGBM	0.867818	0.842167	0.785714	0.812962	0.850425	1.1545

Conclusion and Recommendations:

Some important conclusions drawn from the analysis are as follows:

- The best restaurants in Hyderabad are AB's - Absolute Barbecues, B-Dubs, and 3B's - Buddies, Bar & Barbecue..
- The most popular cuisines are the cuisines which most of the restaurants are willing to provide. The most popular cuisines in Hyderabad are North Indian, Chinese, Continental, and Hyderabadi.
- The restaurants in Hyderabadi have a flexible per person cost of 150 INR to 2800 INR. The cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage - Hyatt Hyderabad Gachibowli.
- Restaurant Clustering was done in two approaches. First with just two features and then with all of them. K means Clustering worked well in the first approach but as we increase the dimensions, it isn't able to distinguish the clusters hence principal component analysis was done and then clustered into 6 clusters. The similarities in the data points within the clusters were pretty great.
- Even though the number of false negatives is Lower in the case of Multinomial NB and Logistic Regression than Light GBM, it is performing better in terms of reducing False positives. This indicates that Mutltinomial NB and Logistic Regression is penalizing False positives more just as we want.

Recommendations:

- Restaurants with negative reviews should be worked with in order to arrive at a win-win situation.
- Ratings should be collected on a category basis such as rating for packaging, delivery, taste, quality, quantity, service, etc. This would help in targeting specific fields that are lagging.

Challenges

- **Feature engineering.**
- **Finding optimum number of Cluster**
- **Text preprocessing**

Thank You