Capstone Project - 4

Zomato Restaurant Clustering and Sentiment Analysis

(Unsupervised Machine Learning)

Team Members

Nihal Habeeb

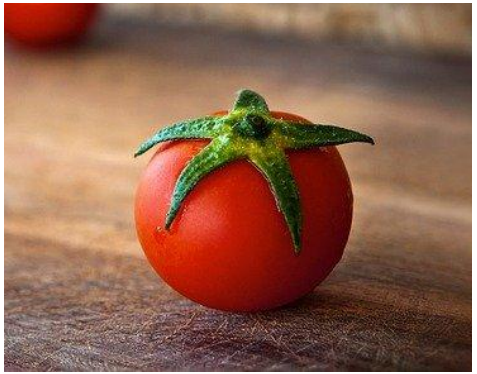
Parvez Makandar

Content

- Problem Description
- Objective
- Data Summary
- Exploratory Data Analysis
- Clustering
- Sentiment Analysis
- Conclusion

The Nature of Indian Restaurant Business

- People are getting more and more interested in dining outside and ordering food online.
- Thus services like Zomato became very popular.
- To thrive in this market it is important to understand its characteristics in detail.



How to analyze the market?

We can study the nature of the restaurant business in detail using the available data in different ways

- **Exploratory Data Analysis:** Relationship between cost per person, cuisines, number of reviews etc. Trends of cost, total cuisines, collections.
- **Cluster Analysis:** Categorizing the restaurants to clusters based on their properties. Understand the distribution and variety of the market as a whole. Recommending similar restaurants, Understanding which restaurant to prioritize etc.
- **Sentiment Analysis:** Study customer response to the services.

Problem Description:

We have two datasets:

1. Details and names of restaurants associated with Zomato.
2. Reviews and reviewer metadata

We want to use this information to understand the trends and relationships within the restaurant business, cluster the restaurants into categories based on their characters and analyze the reviews to understand the sentiment of the reviewer regarding the restaurant and evaluate customer satisfaction.

Objective:

- Use exploratory data analysis to gain insights from the datasets and visualize trends.
- Use text processing methods and text vectorization to extract features from the text based data.
- Cluster the restaurants based on their features using various algorithms.
- Extract features from reviews to build classification models to study and predict the sentiment classes.

Data Summary

The first dataset contains information related to the restaurants such as their name, cost per person, names of cuisines available there etc. The second dataset contains reviews, reviewer information and ratings.

Features:

1. Zomato Restaurant names and Metadata

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

2. Zomato restaurant reviews

Restaurant : Name of the Restaurant

Reviewer : Name of the Reviewer

Review : Review Text

Rating : Rating Provided by Reviewer

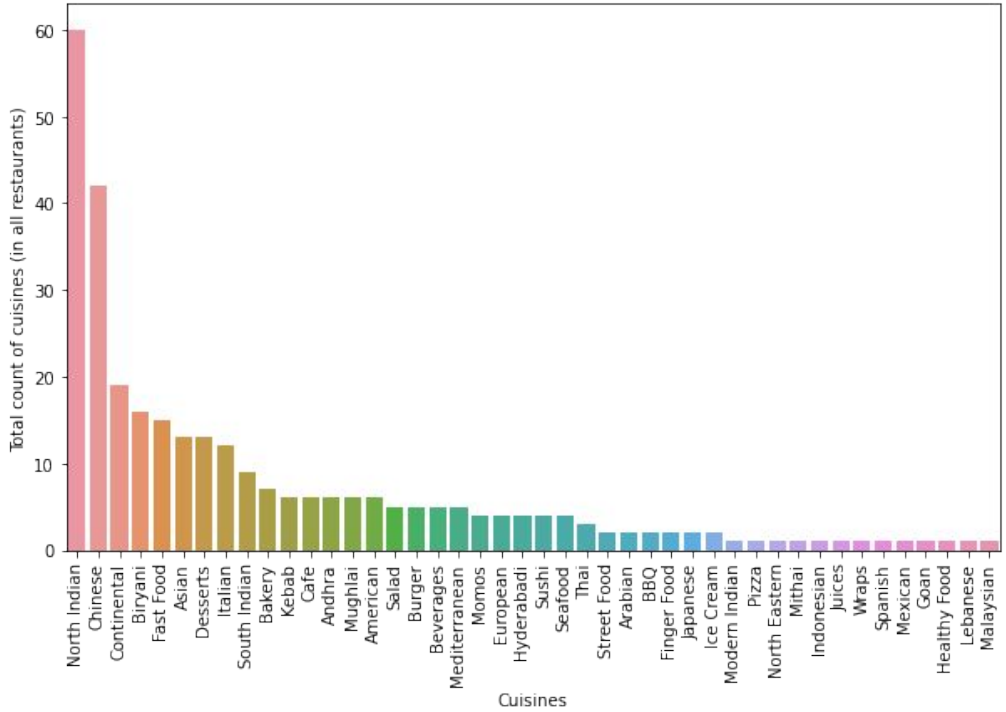
MetaData : Reviewer Metadata - No. of Reviews and followers

Time: Date and Time of Review

Pictures : No. of pictures posted with review

Exploratory Data Analysis

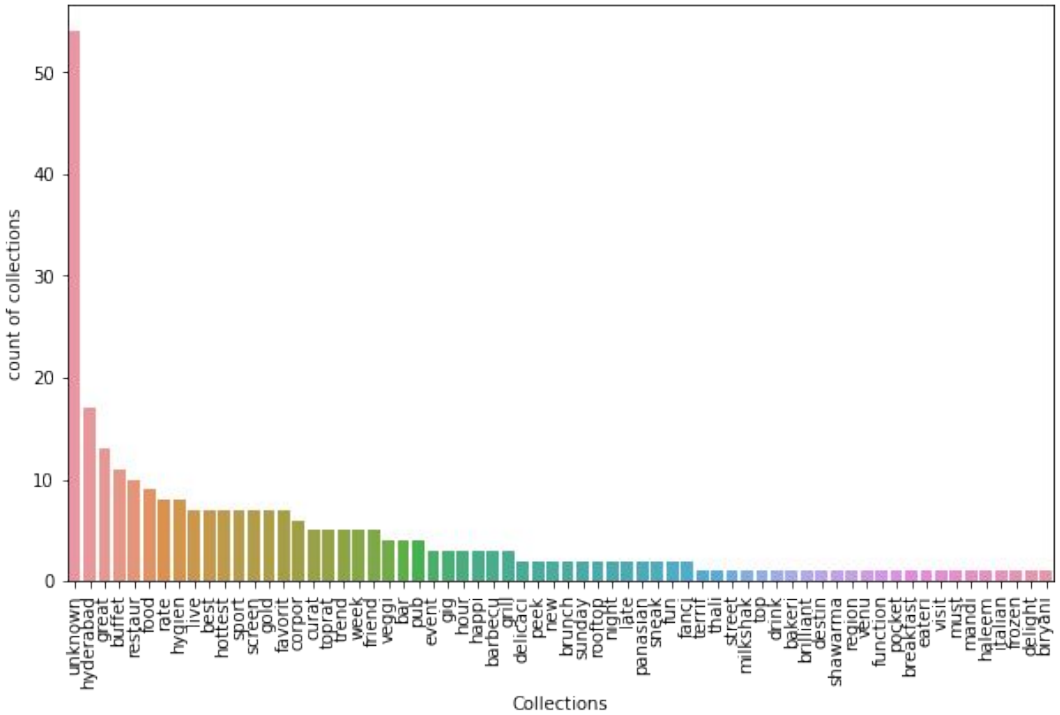
Count of Cuisines



North Indian is the most common cuisine followed by Chinese and Continental. Malaysian is one of the least common.

Exploratory Data Analysis

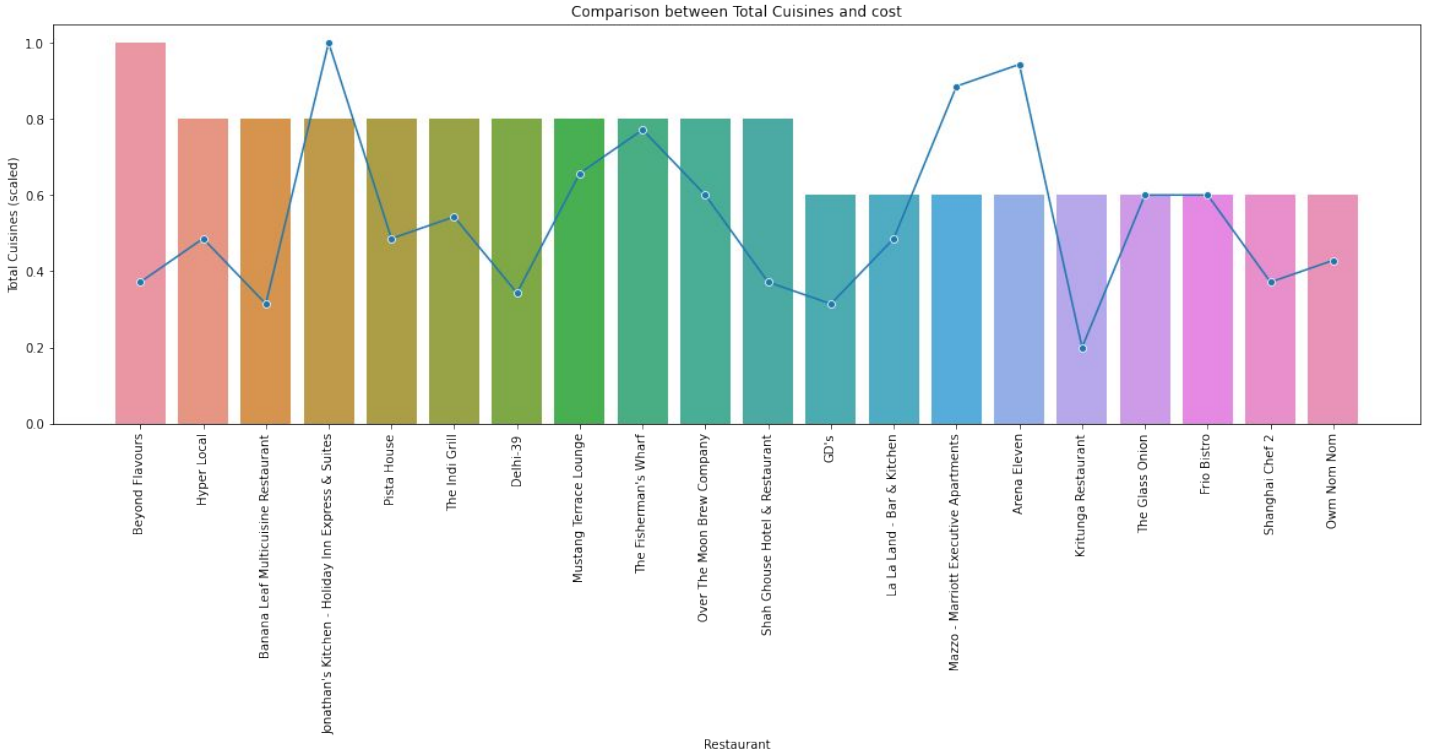
Count of Collections



- “hyderabad” followed by “great” is the most common collection tag.
- “unknown” refers to missing values.
- Collection column has around 51% missing values.

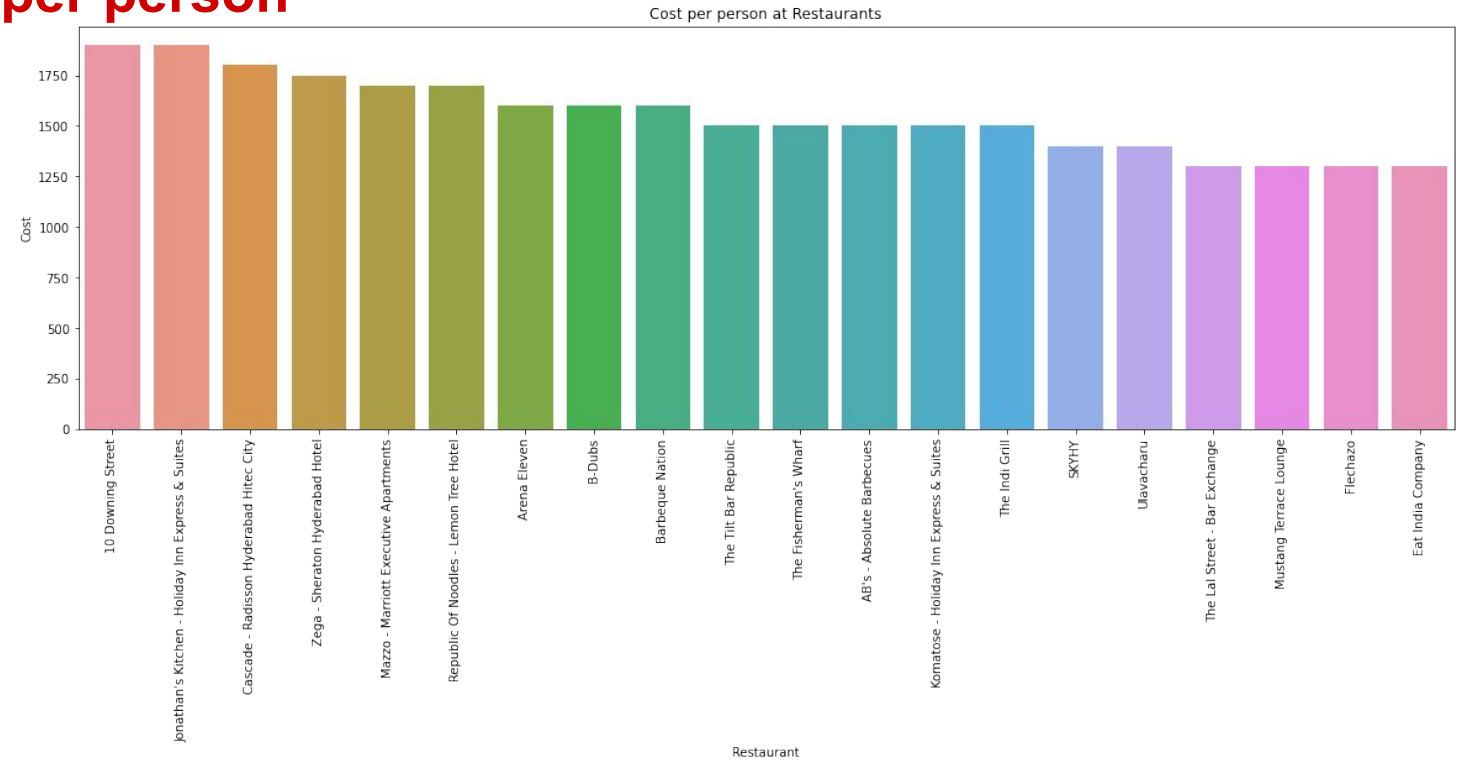
Exploratory Data Analysis

Cuisine and Cost Comparison



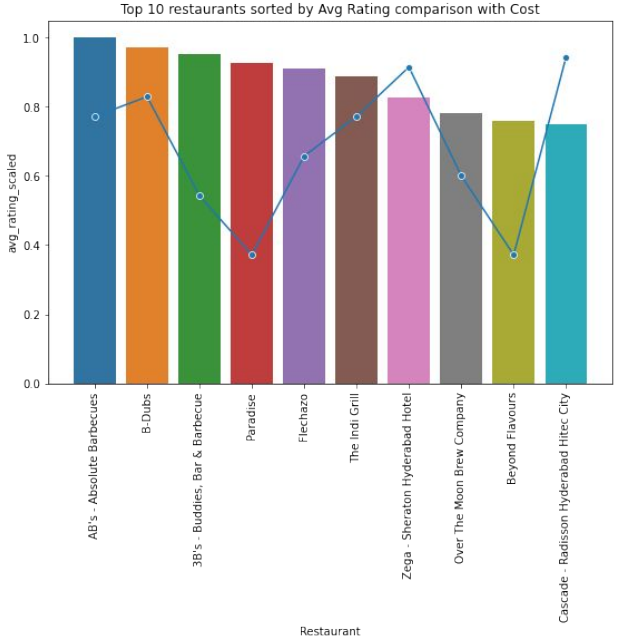
Exploratory Data Analysis

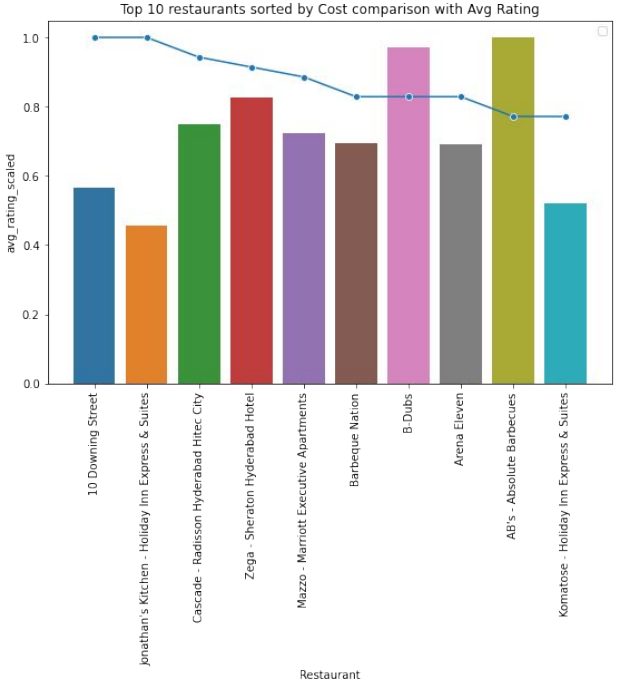
Cost per person



Exploratory Data Analysis

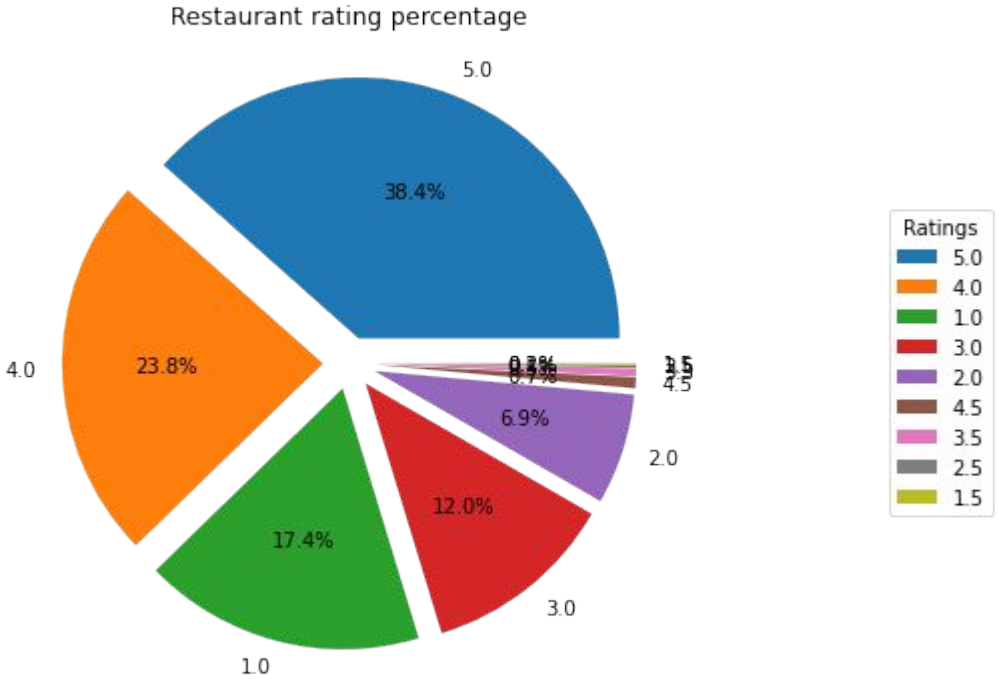
Comparison between Average Ratings and Cost





Exploratory Data Analysis

Distribution of Ratings



A large percentage of the restaurants have 5 rating followed by 4 and 1.

Exploratory Data Analysis

Word Cloud

Good

Bad

Clustering

- Restaurants can be differentiated to groups based on their features using clustering algorithms.
- Include aggregated features like average rating and total followers from reviews dataset along with data from restaurant dataset.
- Use KNN Imputer to handle missing values in the new dataset.
- Extract features from cuisines column:
  - Get list of cuisines.
  - Remove space in case of double worded cuisines.
  - Convert to lowercase.
  - TF-IDF vectorizer to extract features from the text data.

TF-IDF Vectorizer

- Convert each document to a vector.
- All the unique words in the corpus become features (after text processing methods like removal of stopwords, punctuations, lemmatization/stemming etc.)
- Each element of the vector is a TF-IDF weight of a word in the vocabulary.

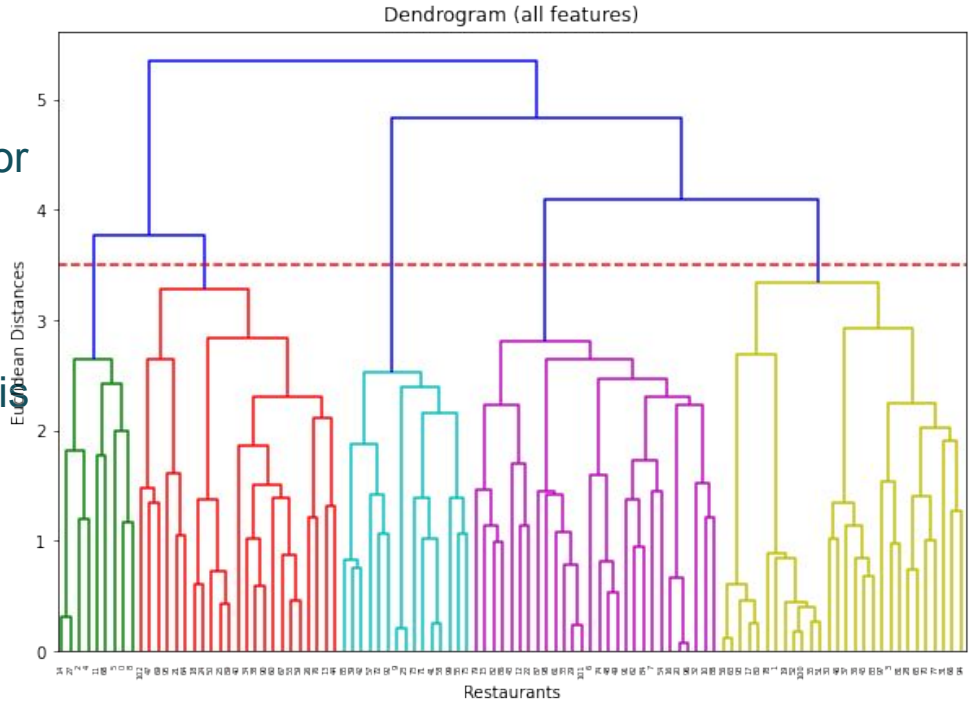
TF-IDF = (normalized Term Frequency)*(Inverse Document Frequency)

TF = (Number of times the term appears in document)/(total terms in the document)

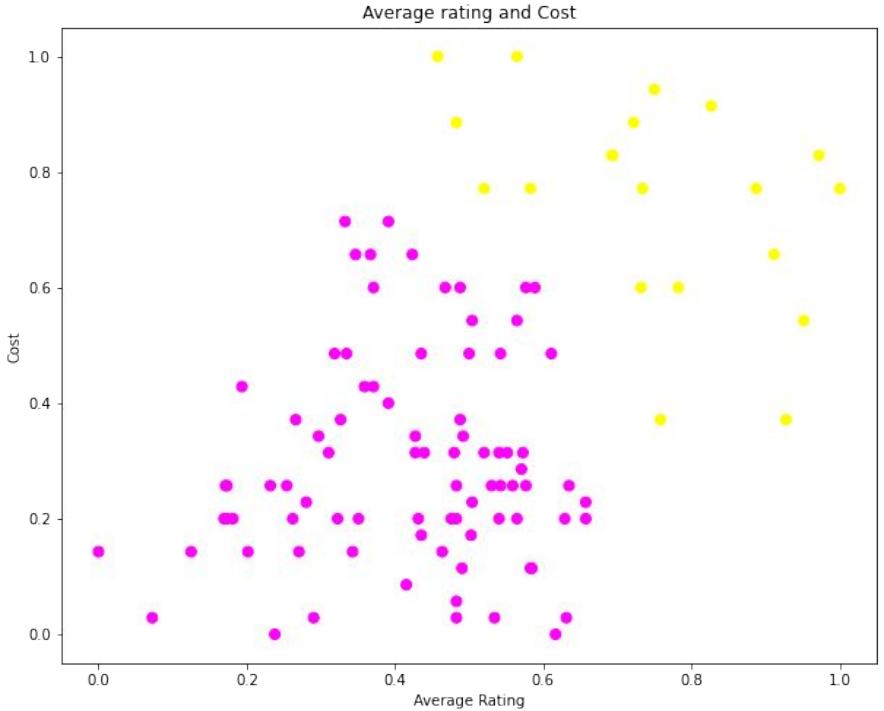
IDF = $\log_e$(Total number of documents/Number of documents with the term in it)

Hierarchical Clustering

- There is no predetermined number of clusters.
- We sequentially build (agglomerative) or split clusters (divisive) from each data point as one cluster or all the points together as a cluster.
- Use dendrogram and silhouette analysis to decide the optimal cluster number
- Use agglomerative clustering to build clusters from single data point clusters to K number of clusters (optimum number).

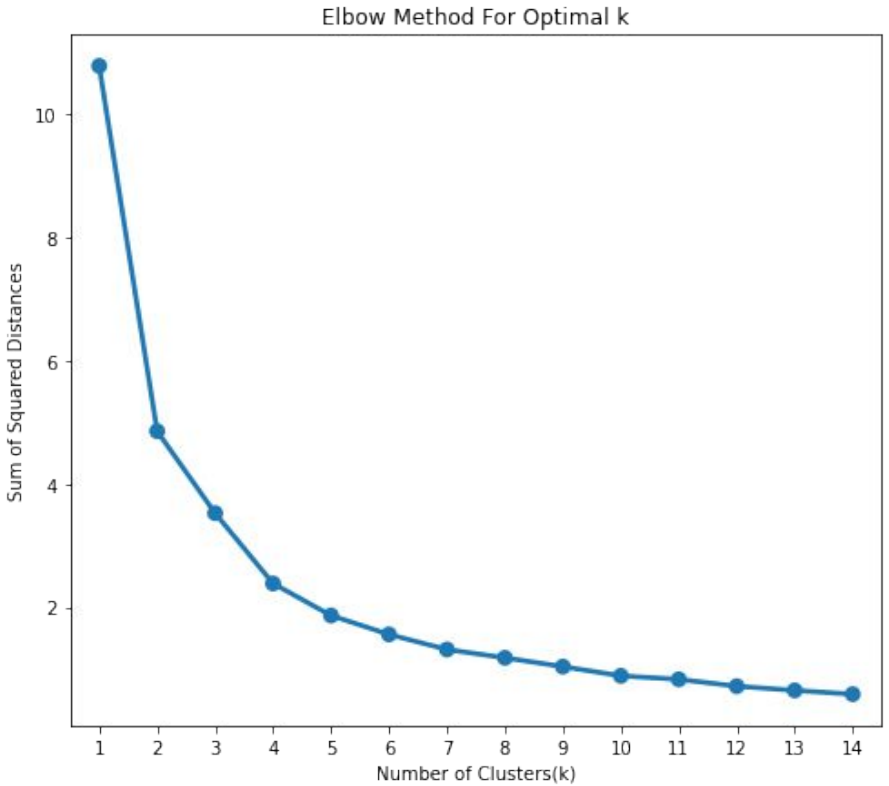


Hierarchical Clustering

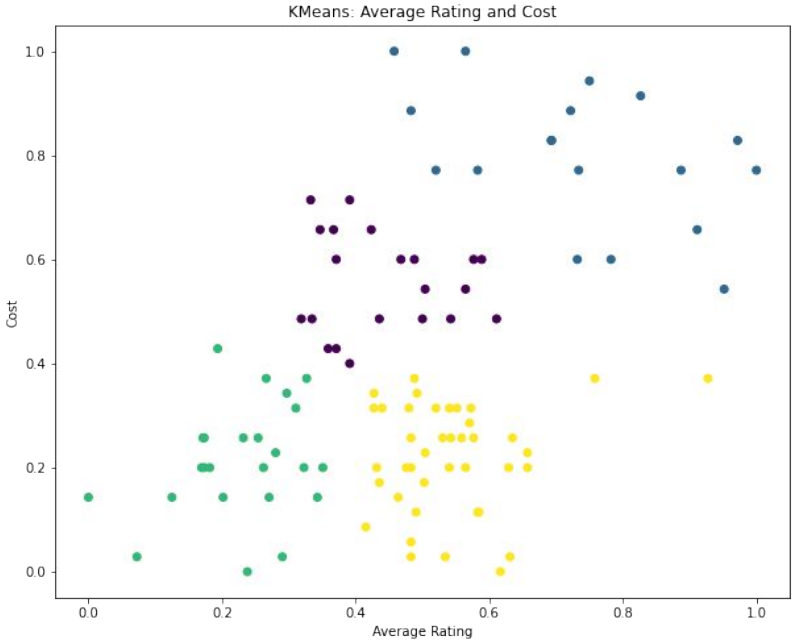


K Means Clustering

- K Means algorithm uses **Expectation-Maximization** to create clusters.
  - Guess the predetermined number of cluster centres
  - Assign points to the nearest cluster centers
  - Set the mean as the cluster centers
  - Repeat steps 2 and 3 until convergence
- Elbow method used to decide the optimal number of clusters.



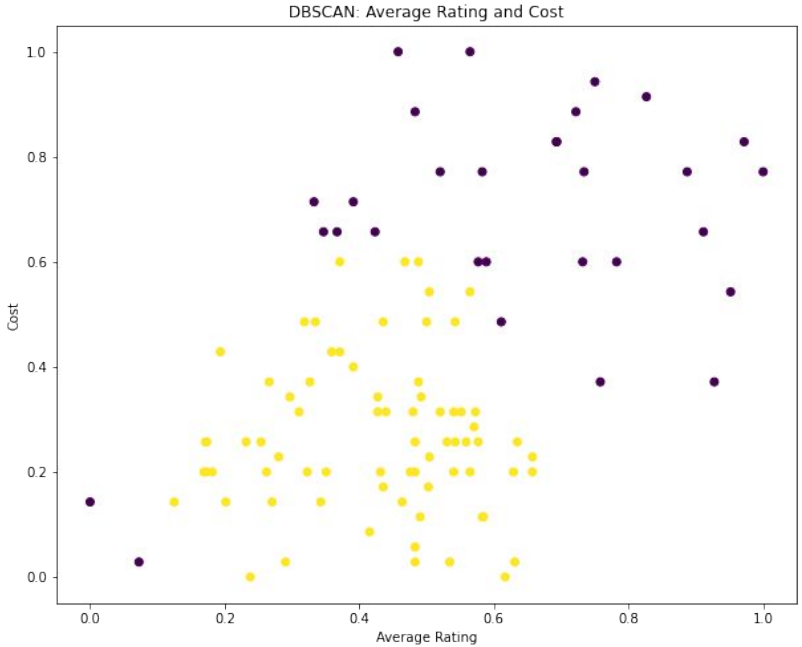
K Means Clustering



DBSCAN Clustering

- DBSCAN is based on the idea that a cluster is a region of high point density separated from other clusters by regions of low point density.
- Parameters
  - epsilon
  - Min Samples

DBSCAN Clustering



Silhouette Analysis

- The silhouette score,
  S = (b-a)/max(a,b)
- a is the mean distance between the observation and all other data points in the same cluster
- b is the mean distance between the observation and all the other data points in the nearest neighboring cluster
- Silhouette Analysis is an effective method to evaluate the clustering algorithm performances

Silhouette Analysis

All features

Model	Optimal Clusters	Silhouette Scores
Agglomerative	8	0.122
K Means	11	0.14

Average Rating and Cost

Model	Optimal Clusters	Silhouette Scores
Agglomerative	2	0.488
K Means	4	0.435
DBSCAN	2	0.443

Sentiment Analysis

- Study the relationship between reviews and customers sentiment towards the restaurant.
- Steps:
  - Text processing
  - Text vectorization
  - Sentiment feature creation
  - Classification models

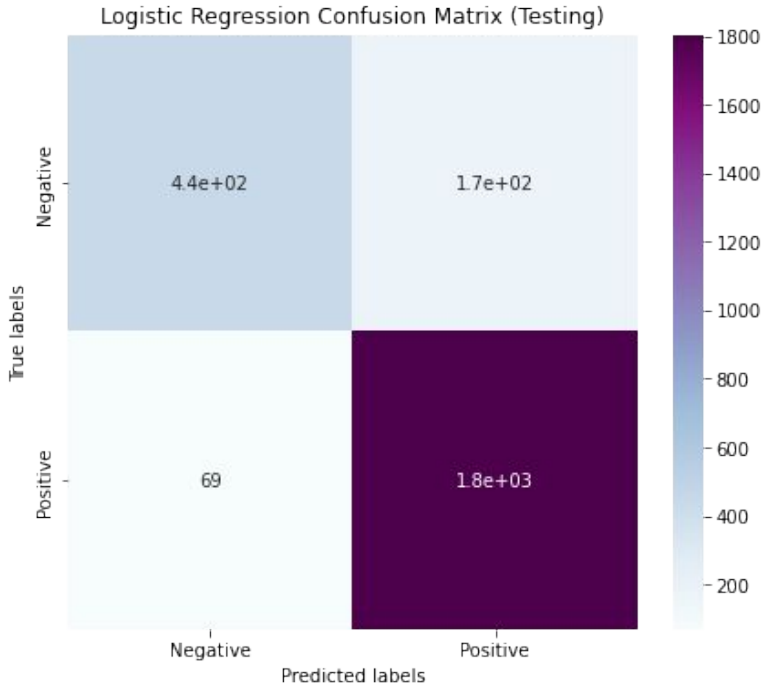
Sentiment Analysis

Text processing

- Uppercase to lowercase conversion
- Remove numbers, symbols, emojis, punctuation and other special characters
- Lemmatization (conversion of words into their root form “lemma” using morphological analysis).
- Remove rarely used words

TF-IDF vectorizer

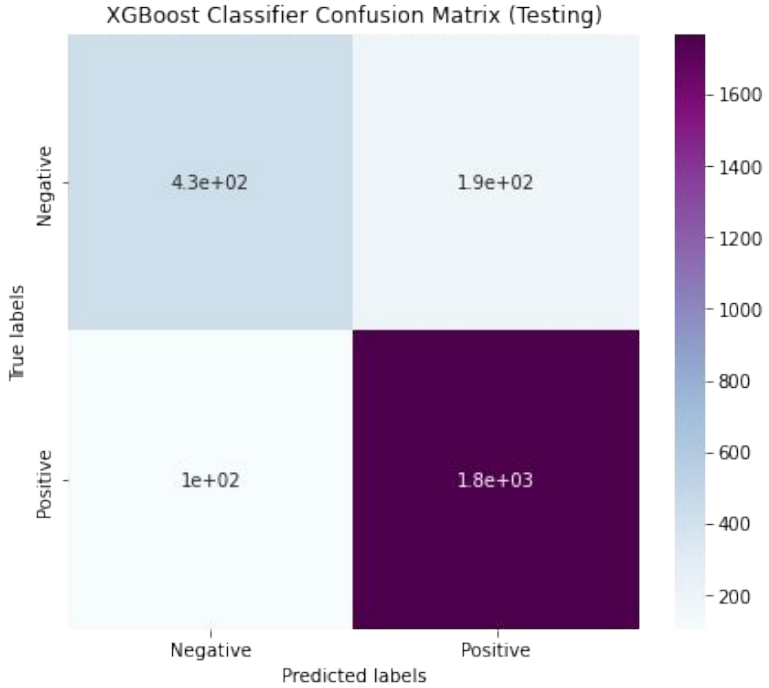
Sentiment Analysis
Logistic Regression



Testing Data Performance
Accuracy : 0.904
Precision: 0.919
Recall: 0.955
F1-Score: 0.937
Area Under ROC Curve: 0.952

Training Data Performance
Accuracy : 0.953
Precision: 0.958
Recall: 0.981
F1-Score: 0.969
Area Under ROC Curve: 0.983

Sentiment Analysis
XGBoost Classifier

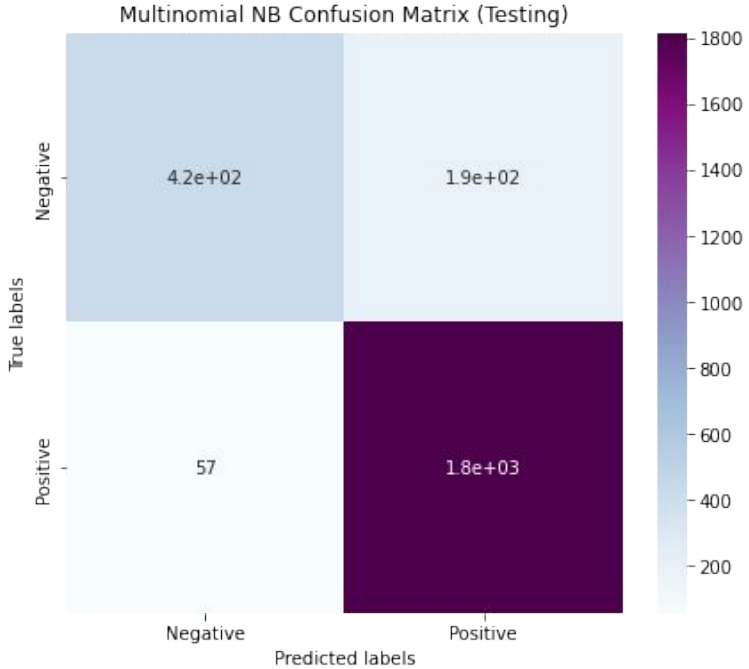


Testing Data Performance
Accuracy : 0.883
Precision: 0.905
Recall: 0.944
F1-Score: 0.924
Area Under the ROC Curve: 0.947

Training Data Performance
Accuracy : 0.957
Precision: 0.961
Recall: 0.984
F1-Score: 0.972
Area Under the ROC Curve: 0.989

Sentiment Analysis

Multinomial Naive Bayes Classifier



Testing Data Performance

Accuracy : 0.9
Precision: 0.904
Recall: 0.97
F1-Score: 0.936
Area Under the ROC Curve: 0.952

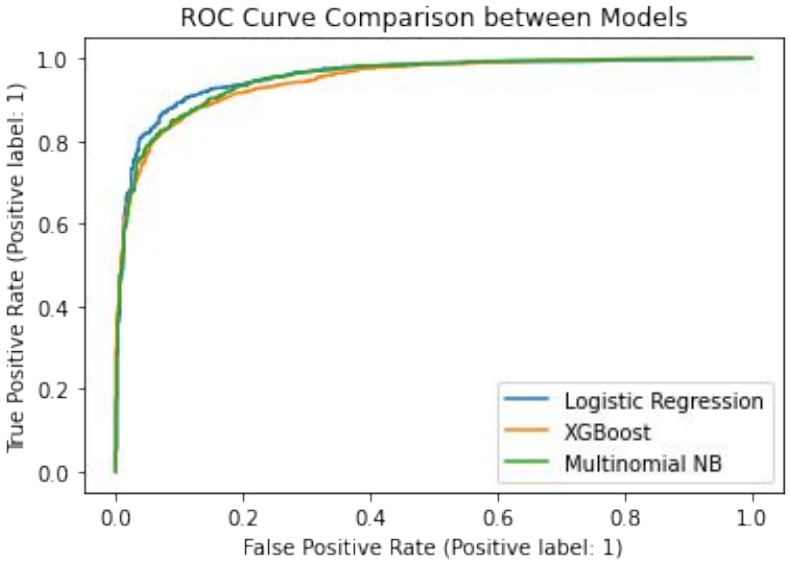
Training Data Performance

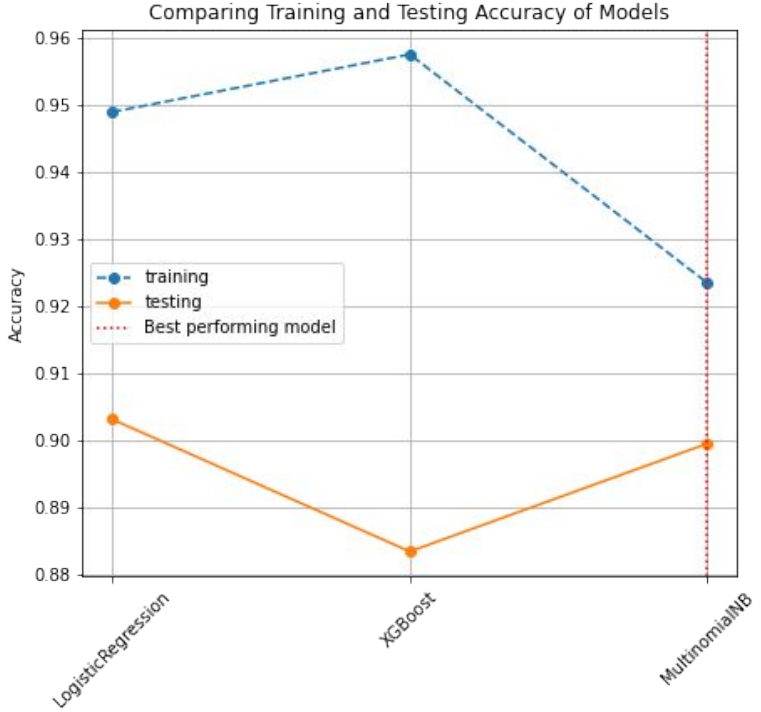
Accuracy : 0.923
Precision: 0.927
Recall: 0.975
F1-Score: 0.951
Area Under the ROC Curve: 0.977

Sentiment Analysis
Model Performance Comparison

Model	Accuracy	F1 Score	ROC AUC Score
Logistic Regression	0.903	0.937	0.956
XGBoost	0.883	0.924	0.947
Multinomial NB	0.899	0.935	0.951

Sentiment Analysis
Model Performance Comparison





Conclusion

- **EDA:**
  - North Indian is the most common cuisine in the data. Malaysian is the least common.
  - ‘Hyderabad’ is the most common collection tag.
  - Cuisines and cost weren’t found to be correlated.
  - 10 Downing street followed by Jonathan’s Kitchen has the highest cost.
  - Ratings and cost weren’t directly correlated.
  - Absolute Barbeque has the highest average rating.
- We were able to successfully gain a lot of insights about the restaurant business and the market that is associated with Zomato through EDA.
- We used a lot of text processing methods to clean the text data which were then converted to vectors to extract features from text. This was done on cuisines (cluster analysis) and reviews (sentiment analysis).

Conclusion

- Used algorithms like agglomerative, K Means and DBSCAN for successfully clustering restaurants to groups based on the extracted features.
- The silhouette scores ranged around 0.1 (when all features were used) and around 0.5 (when only pairs of features were used).
- Visualized clusters in 2D using pairs of features.
- Created a ‘sentiment’ feature from the ratings and used the extracted features from the reviews to build classification models.
- Built logistic regression, XGBoost and Multinomial Naive Bayes classifiers and compared their performances with accuracies varying around 90%.
- Multinomial NB has the smallest difference between testing and training data performance. Logistic regression model performed best on testing data. XGBoost performed best on training data.

Thank You!!