# Capstone Project

## Zomato Restaurant Clustering and Sentiment Analysis

by

Mohammed Haseebuddin

&

Pavan Potnuru

# Points to be Discussed

- Problem statement
- Data summary
- EDA
- Feature engineering
- Sentiment Analysis
- Machine learning models
- Clustering
- Conclusion

# Problem Statement

The Project focuses on analyzing the Zomato restaurant data. You must 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.

# Data Summary

## Zomato 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

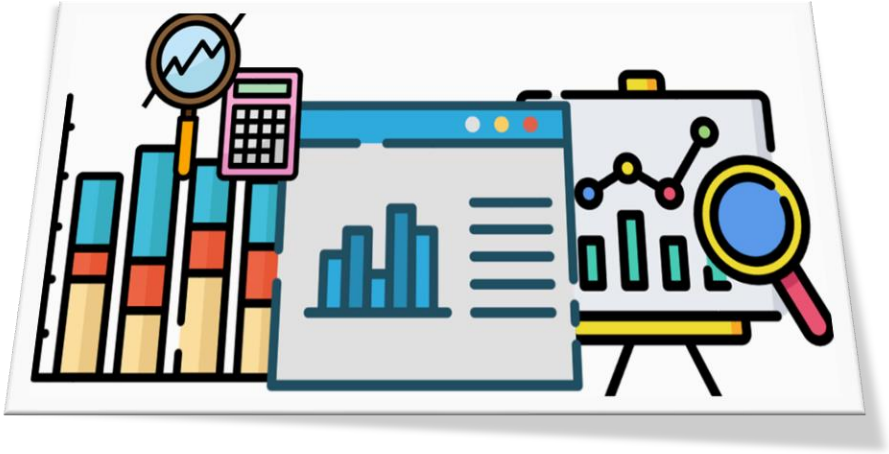
# Data Summary

## Zomato Restaurant Reviews

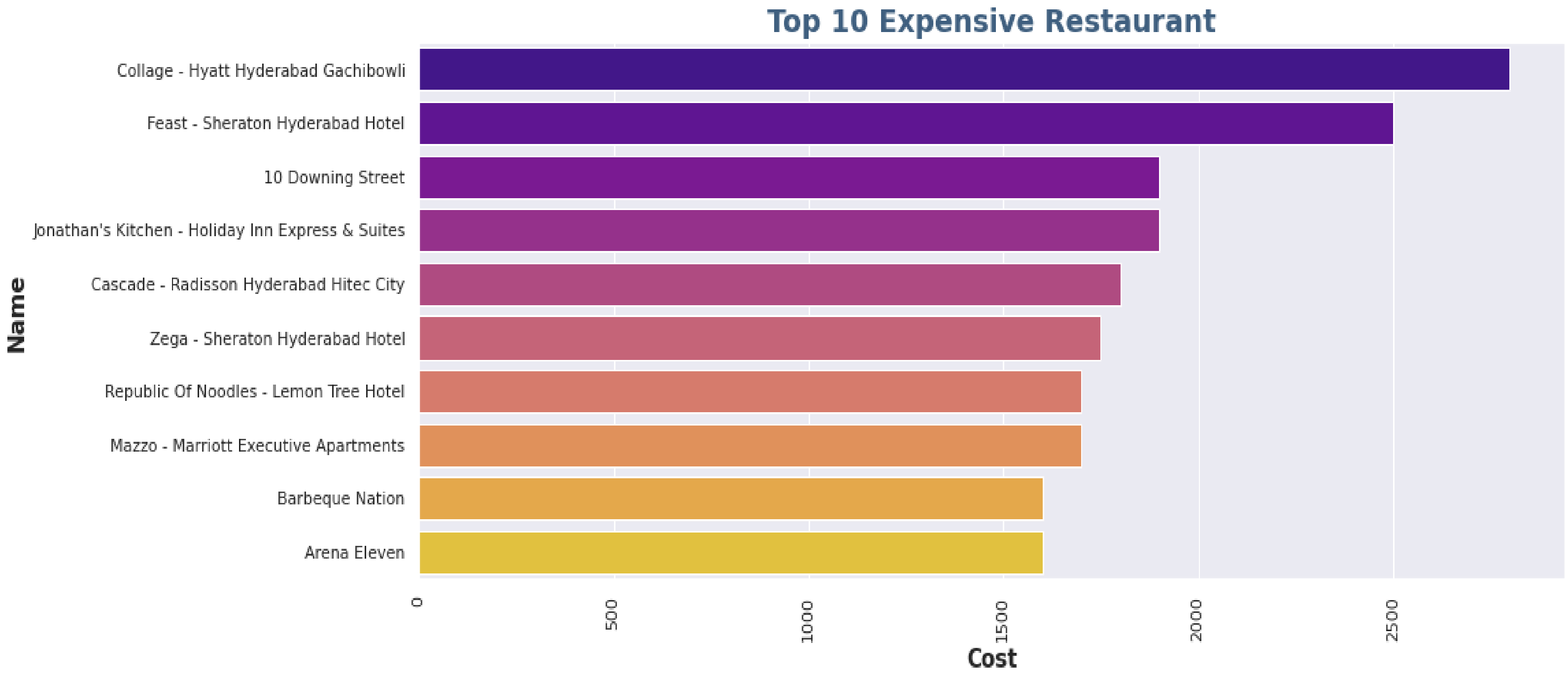
1. **Restaurant :** Name of the Restaurant
2. **Reviewer :** Name of the Reviewer
3. **Review :** Review Text
4. **Rating :** Rating Provided by Reviewer
5. **MetaData :** Reviewer Metadata - No. of Reviews and followers
6. **Time:** Date and Time of Review
7. **Pictures :** No. of pictures posted with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

# EXPLORATORY DATA ANALYSIS

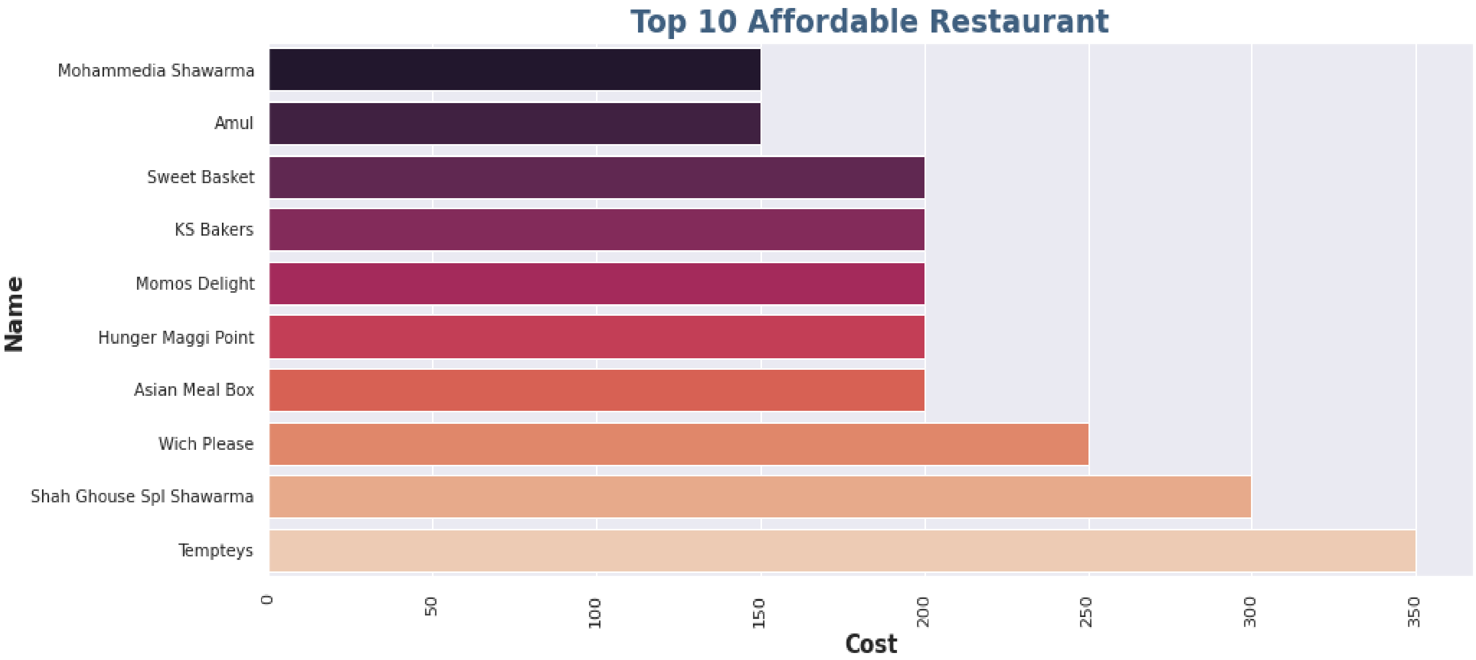
Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.



# EXPLORATORY DATA ANALYSIS

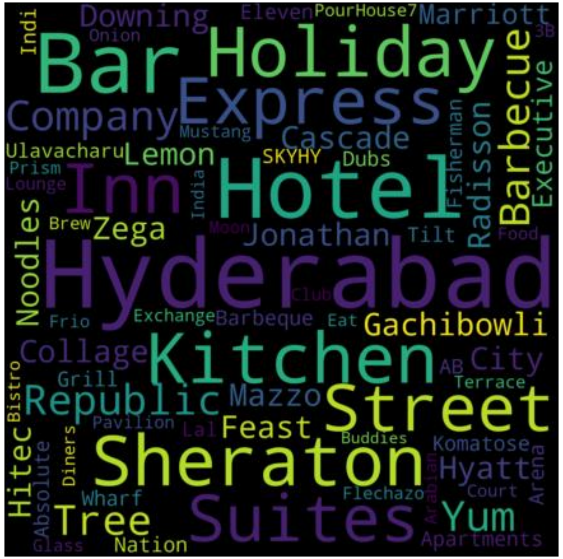


# EXPLORATORY DATA ANALYSIS

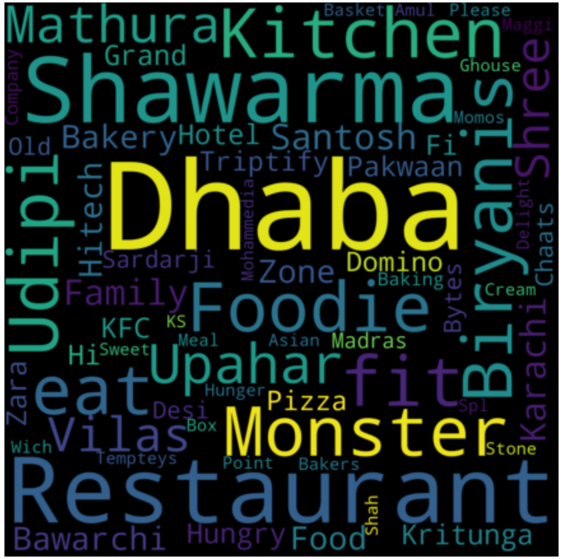


# EXPLORATORY DATA ANALYSIS

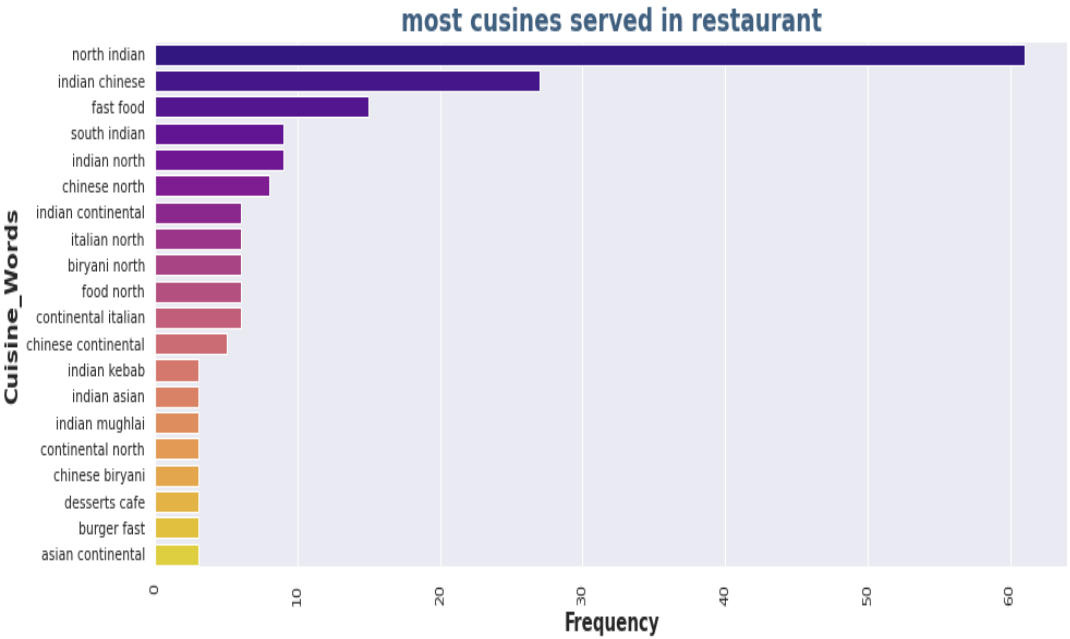
### Word Cloud for Expensive Restaurants

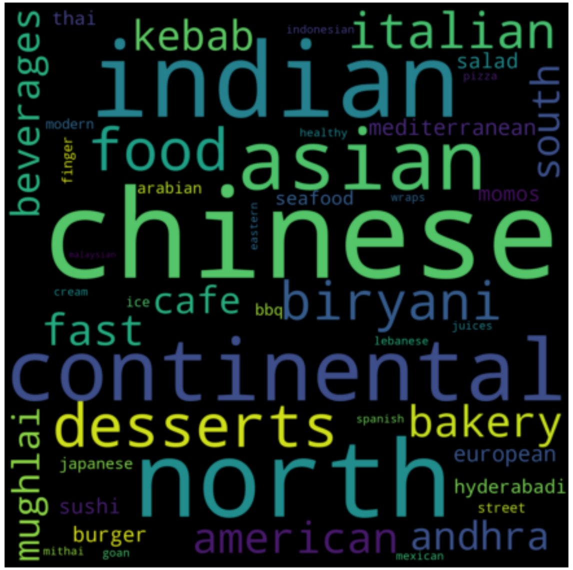


### Word Cloud for Affordable Restaurants

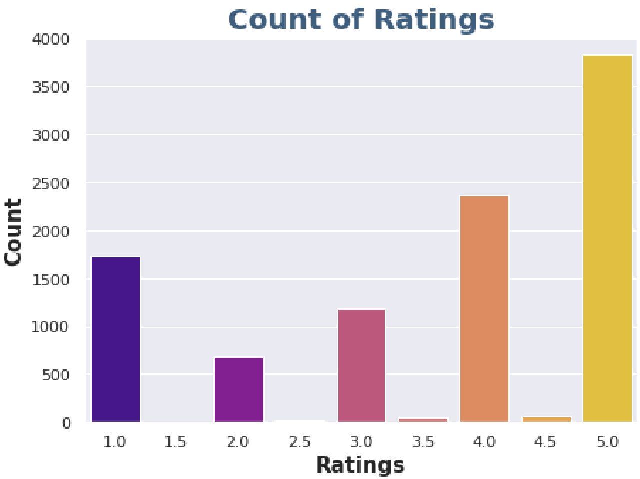


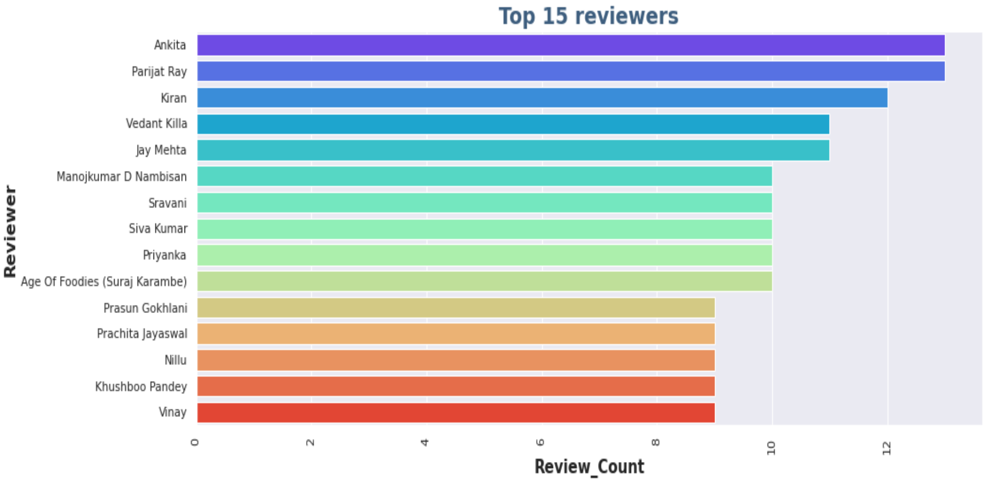
# EXPLORATORY DATA ANALYSIS



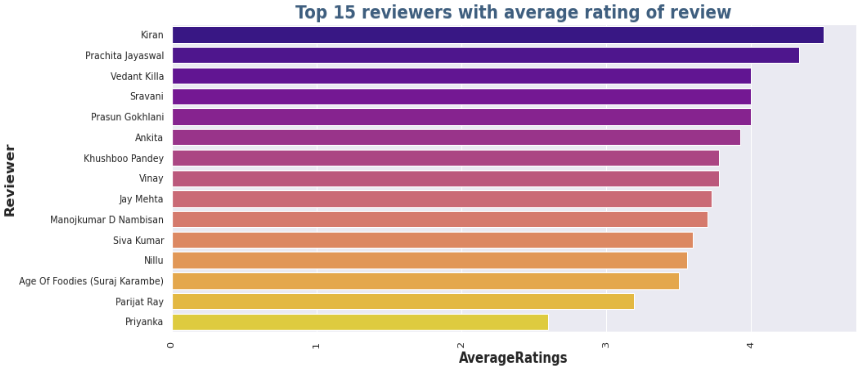


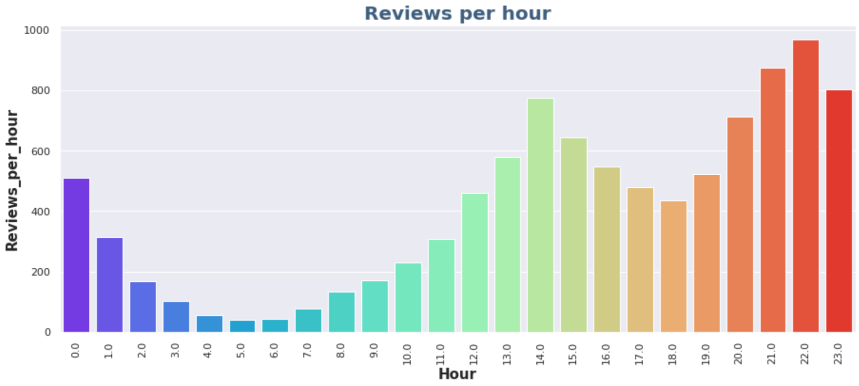
# EXPLORATORY DATA ANALYSIS





# EXPLORATORY DATA ANALYSIS


Top 15 reviewers with average rating of review


Reviews per hour

# FEATURE ENGINEERING

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning. In order to make machine learning work well on new tasks, it might be necessary to design and train better features.

# Text Processing

Data Preprocessing is the most essential step for any Machine Learning model. How well the raw data has been cleaned and preprocessed plays a major role in the performance of the model. Likewise in the case of NLP, the very first step is Text Processing.

The various preprocessing steps that are involved are :

1. Lower Casing
2. Tokenization
3. Punctuation Mark Removal
4. Stop Word Removal
5. Stemming
6. Lemmatization

# Text Processing

**Lower Casing:** Converting a word to lower case (NLP -> nlp). Words like Book and book mean the same but when not converted to the lower case those two are represented as two different words in the vector space model (resulting in more dimensions)

**Tokenization :** It is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph

**Punctuation Mark Removal:** The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations

**Stop Word Removal :** The idea is simply removing the words that occur commonly across all the documents in the corpus. Typically, articles and pronouns are generally classified as stop words.

**Stemming :** This is the process of reducing a word to its word stem that affixes to suffixes and prefixes

**Lemmatization :** This is process of the grouping together of different forms of the same word and converting words into base or root form.

# Sentiment Analysis

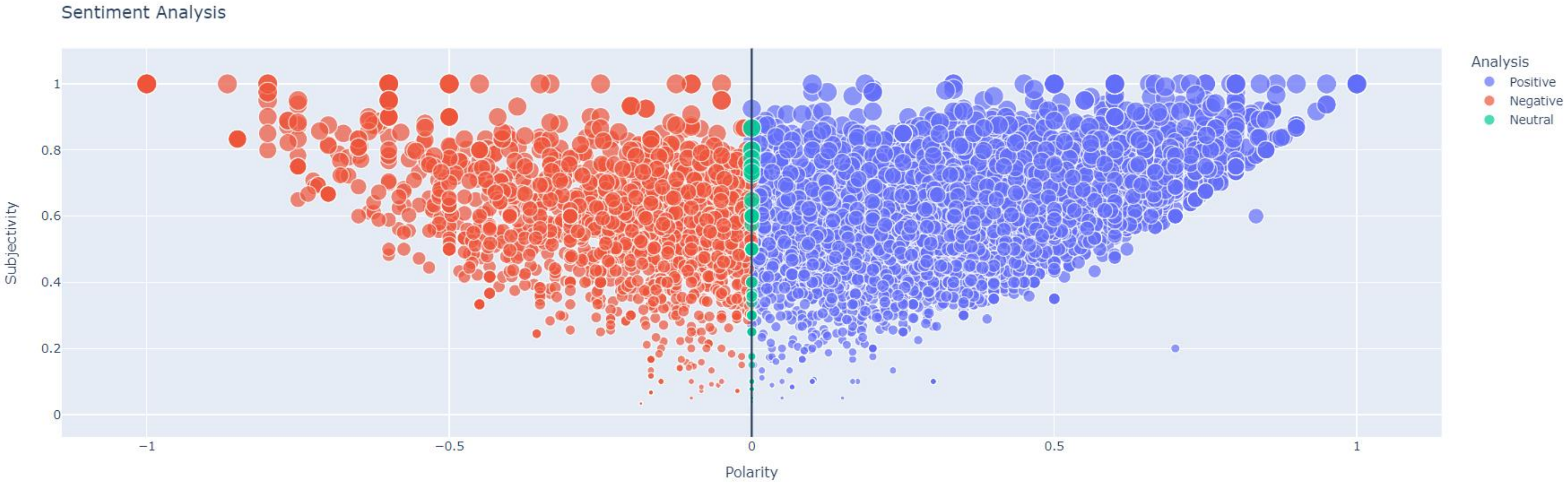
Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral.

In this project we will try to see the sentiment of the review given by the customers for this we will be using following scores:

1. **Subjectivity score :** This score tell us how subjective or opinionated the text is it ranges between (0 – 1)
2. **Polarity score :** This score tells us how positive or negative the text is, it ranges between (- 1 to +1) if score in –ve it means sentiment is negative, if score is positive, it means sentiment is positive

By using these two scores we try to categorize the given review in Positive , Negative or Neutral Category

# Sentiment Analysis



**Out of 9,954 reviews :**

- **7,478 reviews are positive (i.e., 75% of total reviews)**
- **1,887 reviews are Negative ( i.e., 19% of total reviews)**
- **589 reviews are Neutral (i.e., 6% of total reviews)**

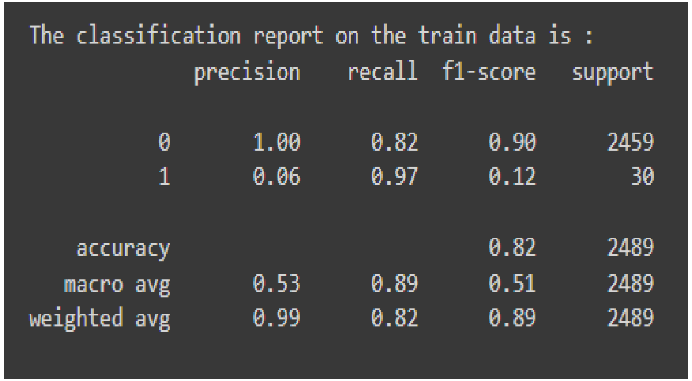
# Applying Models

- Naive Bayes (Multinomial)
- Random Forest Classifier
- XGB Classifier
- Support Vector Classifier

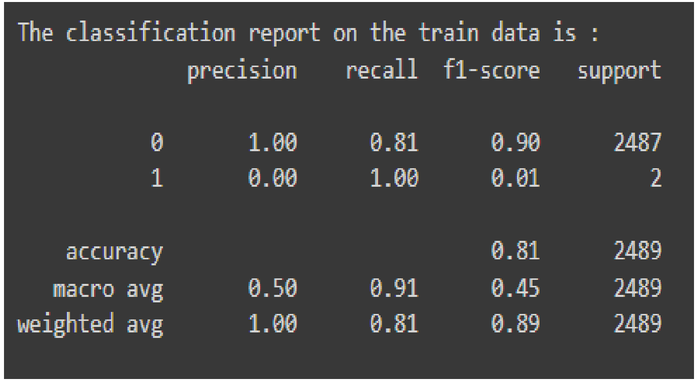
# Machine Leaning Model

| Naive Bayes (Multinomial) | |
|---|---|
| Train Accuracy | 0.836 |
| Test Accuracy | 0.825 |

| Random Forest Classifier | |
|---|---|
| Train Accuracy | 0.813 |
| Test Accuracy | 0.811 |



```
The classification report on the train data is :
              precision    recall  f1-score   support

           0       1.00      0.82      0.90      2459
           1       0.06      0.97      0.12        30

    accuracy                           0.82      2489
   macro avg       0.53      0.89      0.51      2489
weighted avg       0.99      0.82      0.89      2489
```

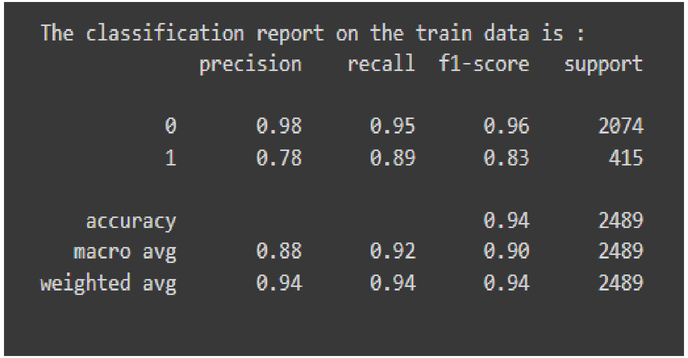


```
The classification report on the train data is :
              precision    recall  f1-score   support

           0       1.00      0.81      0.90      2487
           1       0.00      1.00      0.01         2

    accuracy                           0.81      2489
   macro avg       0.50      0.91      0.45      2489
weighted avg       1.00      0.81      0.89      2489
```

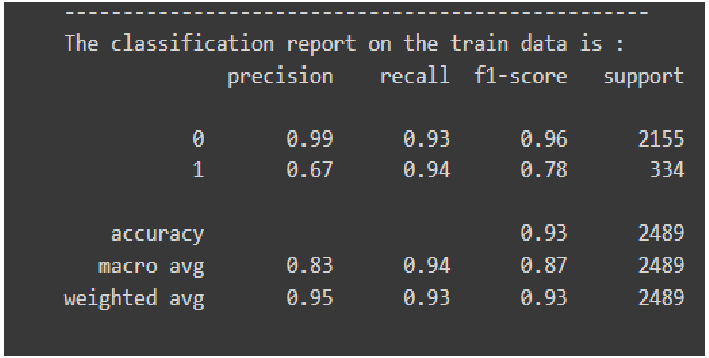
# Machine Leaning Model

| XGB Classifier | |
|---|---|
| Train Accuracy | 0.995 |
| Test Accuracy | 0.939 |

| Support Vector Machine | |
|---|---|
| Train Accuracy | 0.997 |
| Test Accuracy | 0.929 |



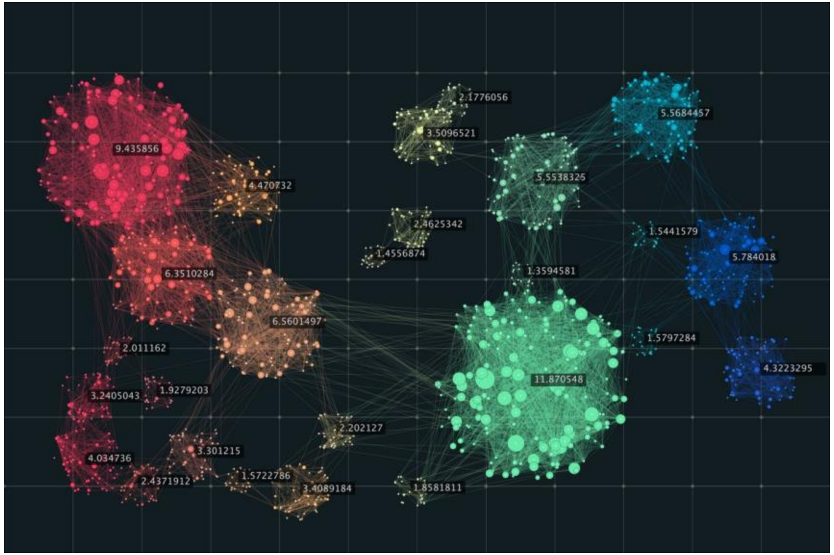
```
The classification report on the train data is :
              precision    recall  f1-score   support

           0       0.98      0.95      0.96      2074
           1       0.78      0.89      0.83       415

    accuracy                           0.94      2489
   macro avg       0.88      0.92      0.90      2489
weighted avg       0.94      0.94      0.94      2489
```



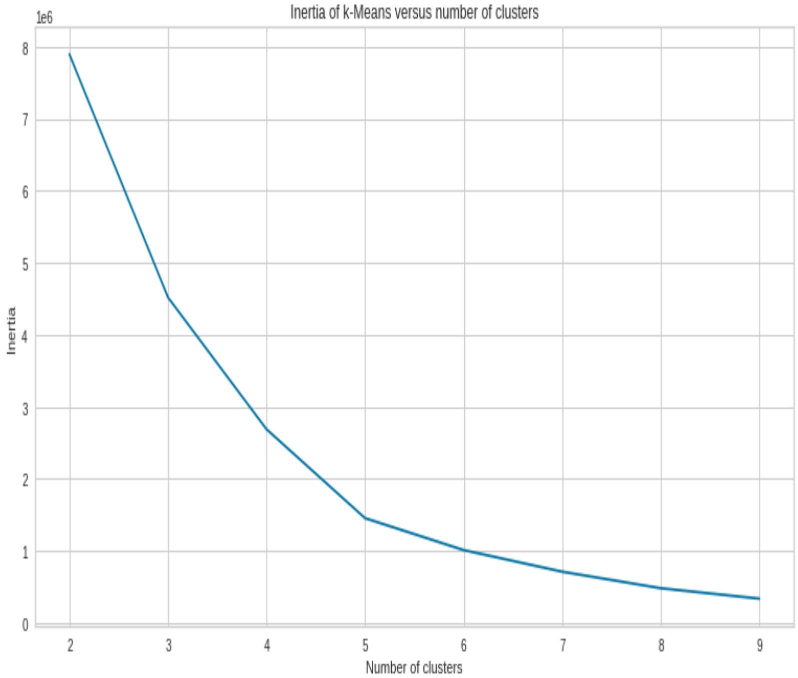
```
--------------------------------------------------
The classification report on the train data is :
              precision    recall  f1-score   support

           0       0.99      0.93      0.96      2155
           1       0.67      0.94      0.78       334

    accuracy                           0.93      2489
   macro avg       0.83      0.94      0.87      2489
weighted avg       0.95      0.93      0.93      2489
```

# Clustering

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.



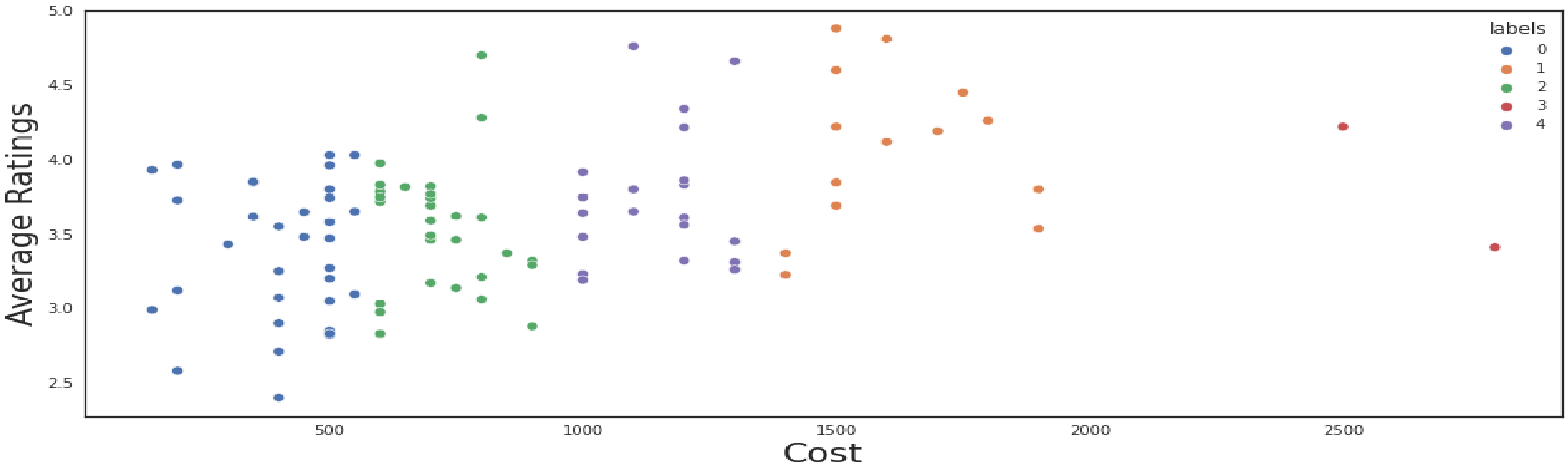
# K-Means Clustering



K-means clustering is a simple unsupervised learning algorithm that is used to solve clustering problems. It follows a simple procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed beforehand.

To get the optimal value of 'k' we use elbow method. From the graph we can see optimal cluster value is 5

# K-Means Clustering



**We can be here that clusters were divided in terms of the cost, there are total 5 clusters.**

# Top 3 Cuisines in each Cluster

| Top Cuisines in Cluster 0 | |
|---|---|
| North Indian | 16 |
| Chinese | 9 |
| Fast food | 8 |

| Top Cuisines in Cluster 1 | |
|---|---|
| North Indian | 11 |
| Continental | 6 |
| Asian | 5 |

| Top Cuisines in Cluster 2 | |
|---|---|
| North Indian | 18 |
| Continental | 18 |
| Biryani | 11 |

| Top Cuisines in Cluster 3 | |
|---|---|
| Asian | 2 |
| Italian | 2 |
| Continental | 2 |

| Top Cuisines in Cluster 4 | |
|---|---|
| North Indian | 14 |
| Chinese | 9 |
| Italian | 7 |

# Conclusion

- 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 cheapest is the food joint called Mohammedia Shawarma and the costliest restaurant is Collage – Hyatt Hyderabad Gachibowli.
- Sentiment Analysis was done on the reviews and a model was trained in order to identify negative and positive sentiments.
- SVM and XGB both performed well, and we can choose any one them.
- SVM and XGB are having 0.921 and 0.981 of testing accuracy, respectively.
- We got best cluster as 5 in K-Means and Principal Component Analysis(PCA).