Capstone Project-4

Zomato Restaurant Clustering and Sentiment-Analysis

Nakul Pradeep

Introduction

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities. We have to analyze the sentiments of the reviews given by the customer in the data and also make some useful conclusion in the form of Visualizations. Further we have to cluster the zomato restaurants into different segments.

What is sentiment analysis ?

- Sentiment Analysis is a use case of Natural Language Processing (NLP) and comes under the category of text classification.
- Sentiment Analysis involves classifying a text into various sentiments, such as positive or negative, Happy, Sad or Neutral, etc. Thus, the ultimate goal of sentiment analysis is to decipher the underlying mood, emotion, or sentiment of a text.

Project Objective

- The given datasets are ‘zomato restaurant and metadata’ and ‘zomato restaurant reviews’.
- Our aim is to assess the sentiment of various critics using their reviews and also cluster restaurants in to different segments and draw useful conclusions with the help of visualisation.
- The sentiment analysis is important as it is used to find the product feedback and also can be used to reveal key information into whether the company is doing right or wrong. Companies may use sentiment analysis to assess the effectiveness of a new product, ad campaign, or other marketing initiatives.

Variable Breakdown

Zomato restaurant names and metadata

We use this data for clustering

- **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

Variable Breakdown(Continued)

Zomato Restaurant reviews

Merge this dataset with Names and Metadata and then use for sentiment analysis part

- **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

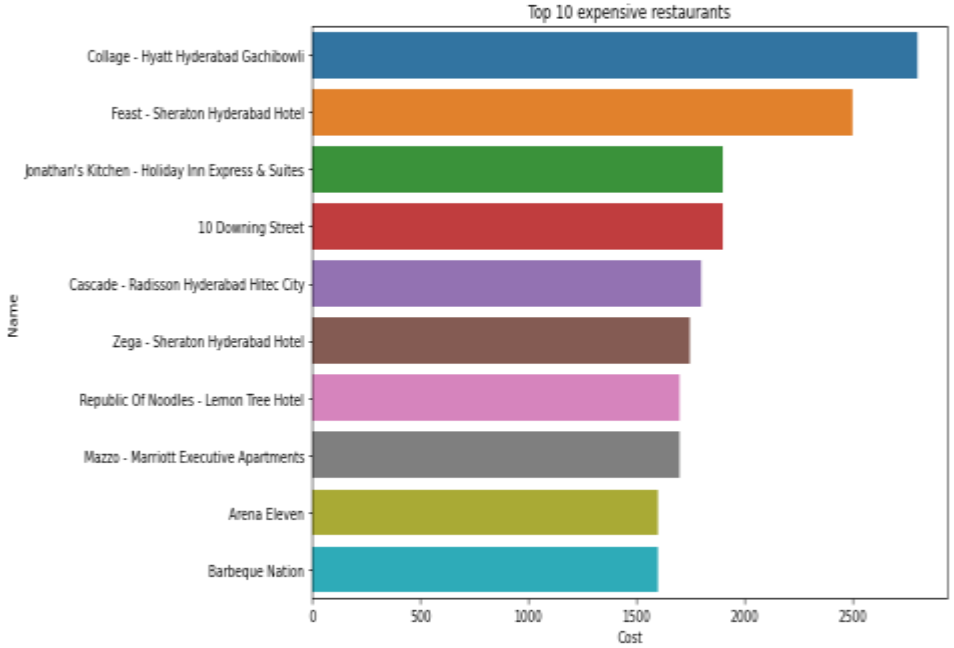
Steps of our Data Analysis

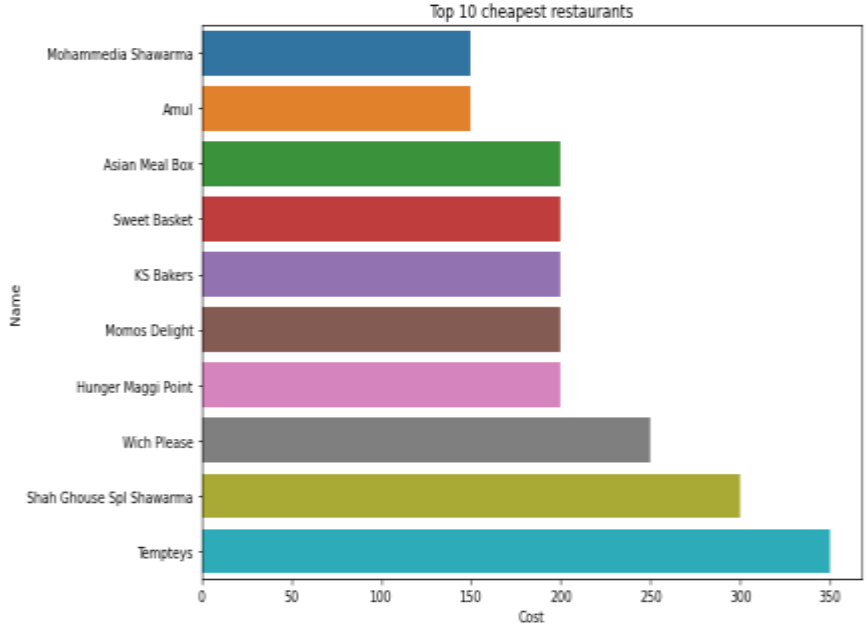
1. Problem Description and Data Description
2. Import the Libraries and Dataset
3. Handling Null value and Checking for duplicate values
4. Conduct exploratory data analysis with Visualisation
5. Clustering with K mean clustering, Hierarchical clustering .
6. Text processing like cleaning of data, removal of stop words, punctuations, tags etc.
7. Sentiment analysis using Count vectorizer, Text blob function for analyzing polarity and subjectivity.
8. Evaluating various models and metrics for sentimental analysis.
9. Summaries and valuable insights

Exploratory Data Analysis

- For the zomato restaurant metadata there are 105 rows and 6 columns or features, while in review dataset there were 10,000 rows and 7 columns.
- Both the data set does not contain duplicate values.
- There were null values present in both dataset , In the metadata the null values present in collection dataset was converted to 'Not rated', while in the review dataset the dataset containing null values were very few but important features, so in order to avoid inaccuracy these were removed.
- The mean cost of food for the restaurants in zomato is 861.42 rupees.
- Most Common timing of restaurants in zomato are from 11 AM to 11PM.

Top 10 expensive and cheap restaurants



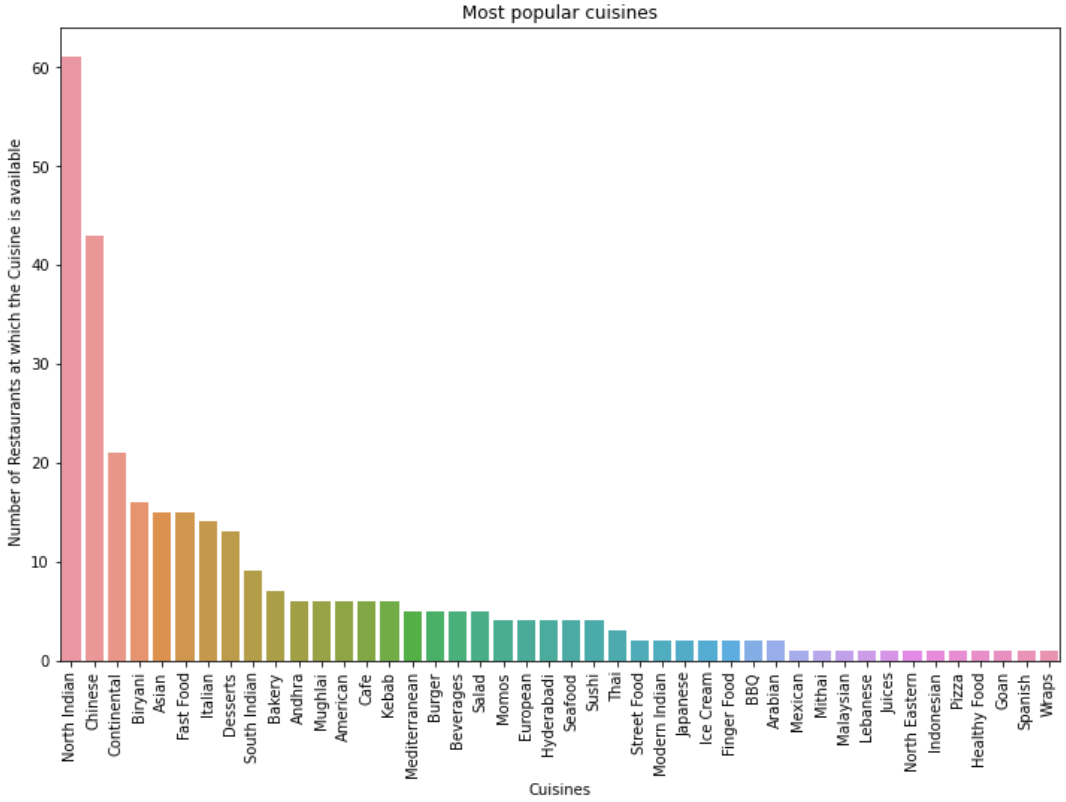


Collage-Hyatt Hyderabad Gachiboli is the most expensive restaurant.

Amul and Mohammedia Shawarma are the cheapest restaurant

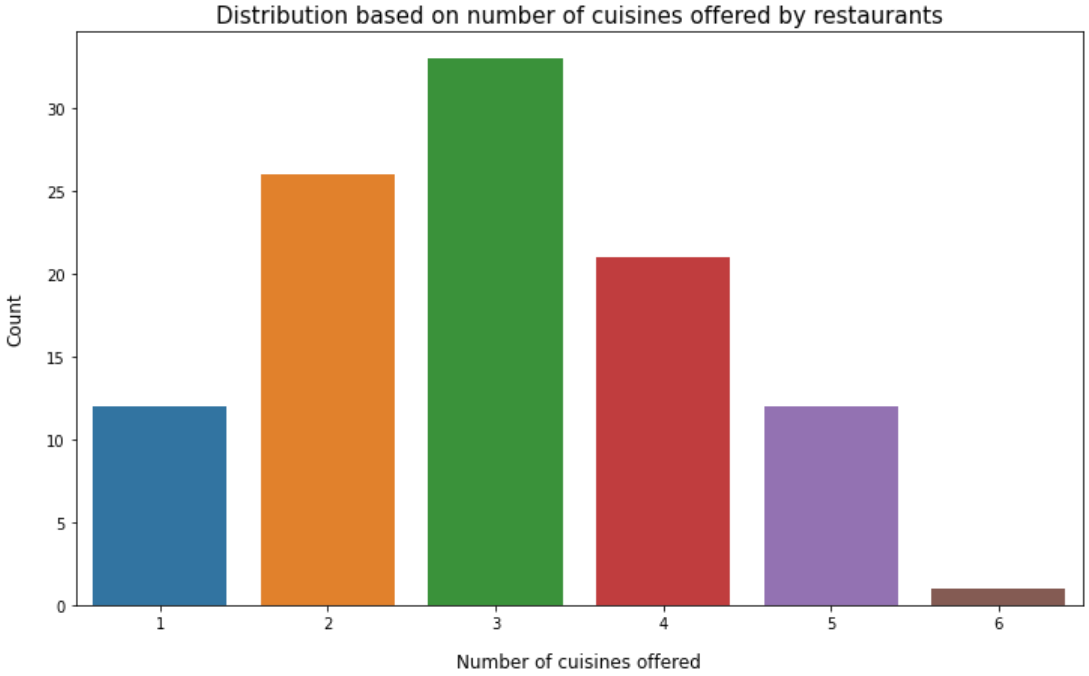
Top 10 most popular cuisine

- Most popular cuisine of our dataset is North Indian with 61 restaurants offering it.
- Chinese is the second most popular cuisine with 43 restaurants offering it.

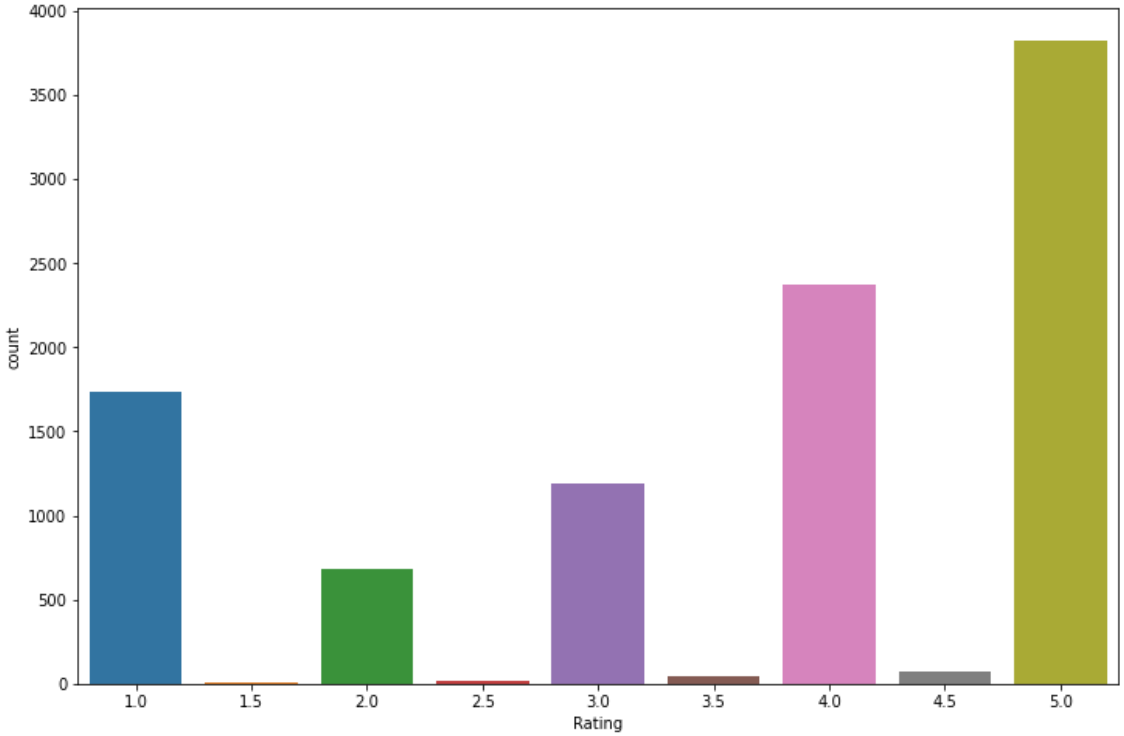


Distribution based on the number of cuisines offered by restaurants

- Majority of restaurants provide three different types of cuisines
- Only one restaurant provides six different types of cuisines and that restaurant is “Beyond flavours”.



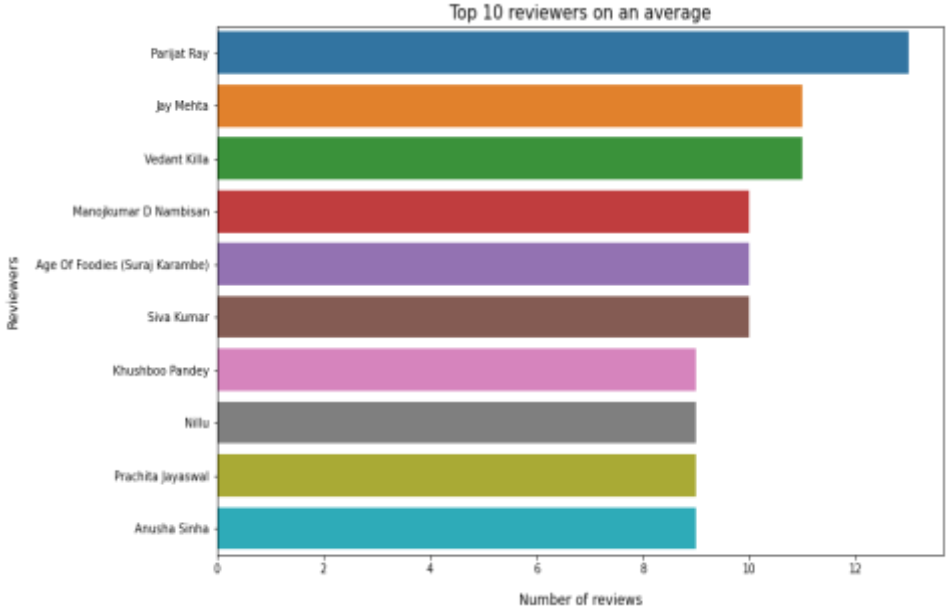
Distribution of rating in our dataset

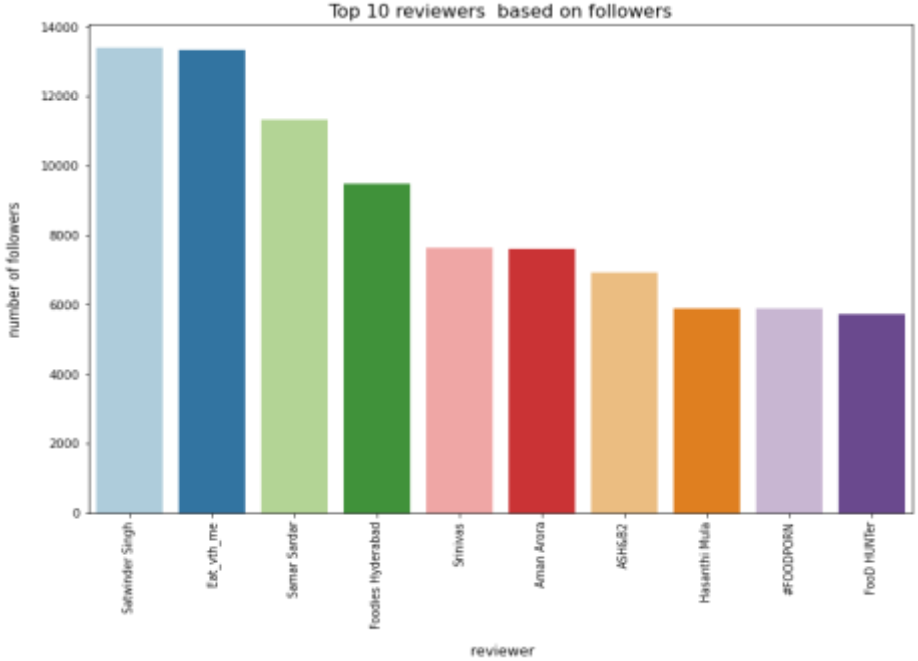


	Rating	Pictures
count	9955.000000	9955.000000
mean	3.600402	0.751984
std	1.483565	2.575691
min	1.000000	0.000000
25%	3.000000	0.000000
50%	4.000000	0.000000
75%	5.000000	0.000000
max	5.000000	64.000000

- Average rating of the restaurants in zomato is 3.6, while median rating is 4 implying our rating column is negatively skewed.
- Maximum pictures uploaded by a person is 64.
- More than 75 percentile of people do not upload pictures.

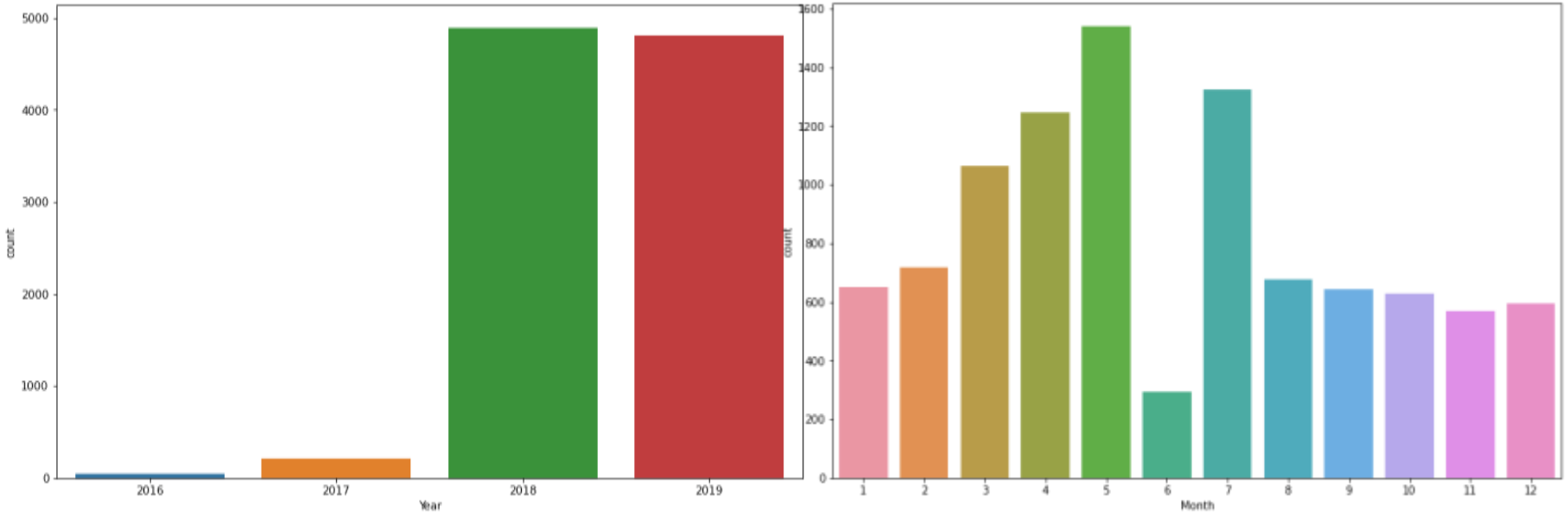
Top 10 Reviewers based on number of reviews and followers





- Out of all the reviewers, Parijat Ray with 13 reviews was the largest reviewer(in our dataset).
- Satwinder Singh was the reviewer with most number of followers, he has 13,410 followers, Eat_vth_me was a close second with 13,320 followers.

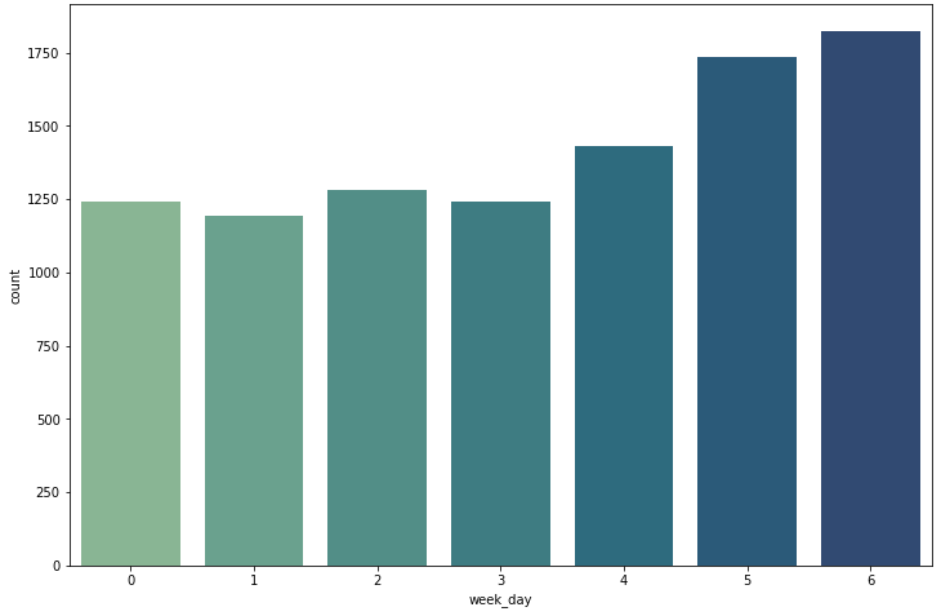
Count plot of Year and month of reviews

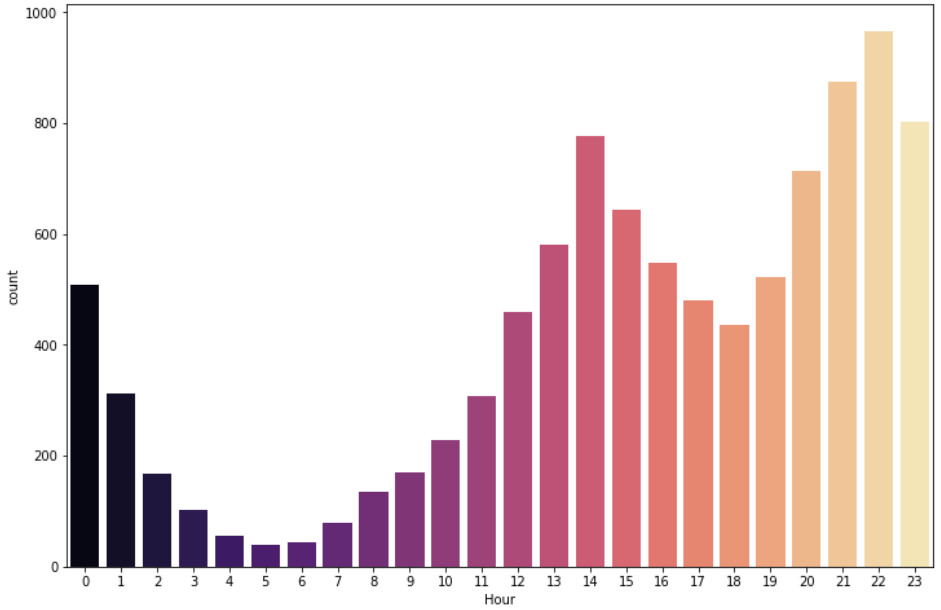


The reviews till 2017 were very less ,from 2018 and 2019 the number of reviews increased exponentially.

The count of reviews were higher during the months of May and July and least during the month of June

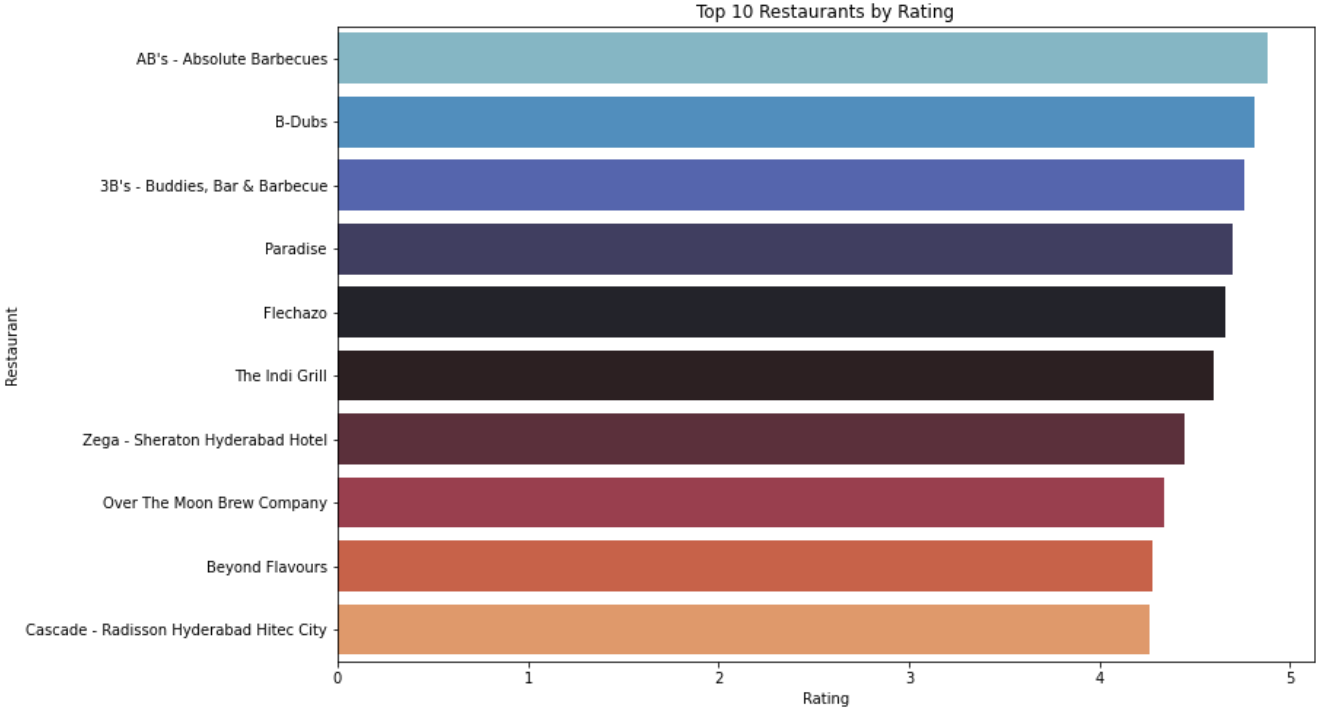
Distribution of reviews based on week day and hour





Most of the reviews are posted during weekends(i.e. Saturday and Sunday)
The majority of reviews are posted at night from 9 to 11pm.

Top 10 Restaurants based on rating



The restaurant with highest average rating from the given dataset is AB's – Absolute Barbecues with a rating of 4.88

Word Cloud

Word cloud of the reviews in our dataset

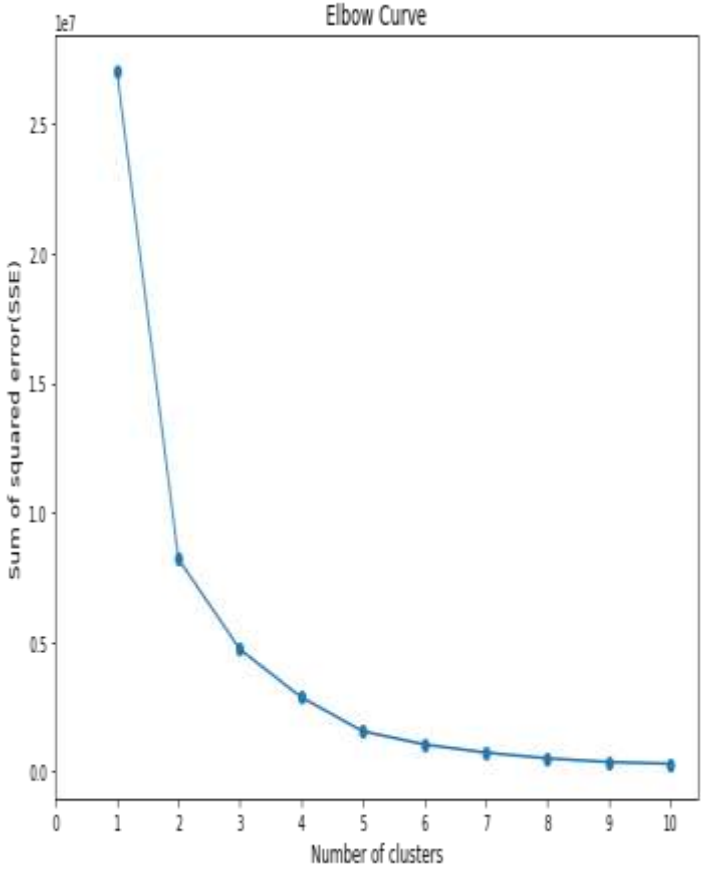
Clustering

Clustering is the process of segregating groups with similar traits and assigning them to clusters. It is an unsupervised learning method. It is important as it determines inherent grouping of the unlabelled data present. There is no criteria for good clustering. It depends upon the user and criteria they need to satisfy. We have used ‘K means clustering ‘ and ‘hierarchical clustering’ for the given data.

K means clustering is a type of partitioning method, It partitions the object in to k clusters. The number of clusters we have to partition it in to is to be provided by us. To get the optimal number of clusters we need, we use elbow curve method and silhouette coefficient analysis.

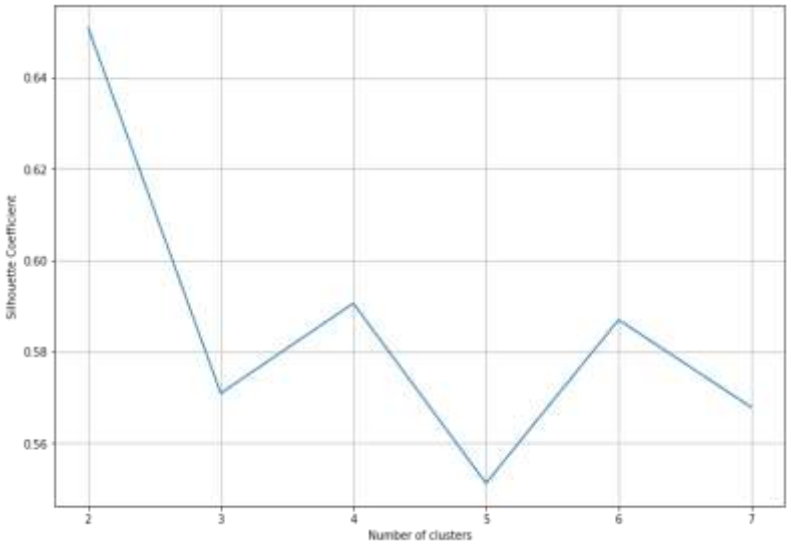
Clustering(Continued)

- The number of clusters was decided with the help of an elbow curve
- Sum of squared error is defined as the sum of squared distance between the centroids and each point in the cluster
- From the elbow method we could see that the optimum number of cluster could be 5 or 6.
- We can use silhouette coefficient to decide between them



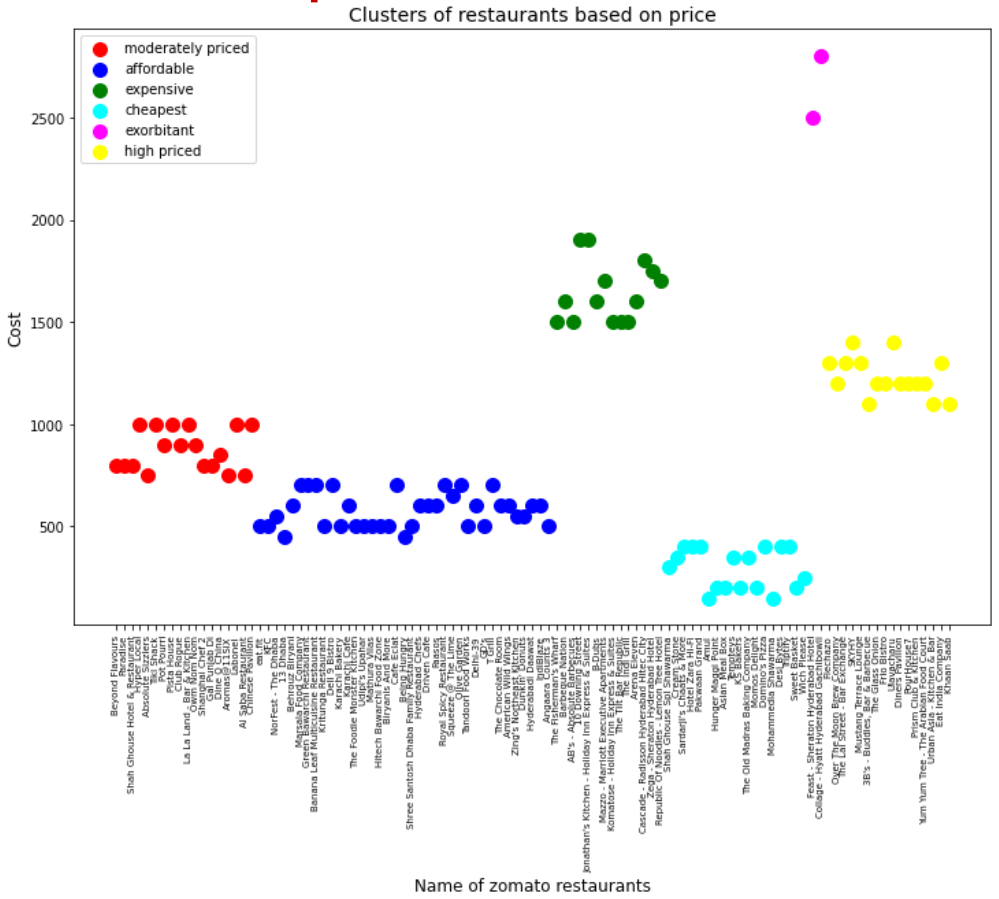
Silhouette Analysis

- Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.
- Higher the coefficient better the clustering(as higher value implies each samples are far away from neighbouring clusters).
- So we can select 6 as the better number of cluster over 5 as the silhouette score of 6 is higher.



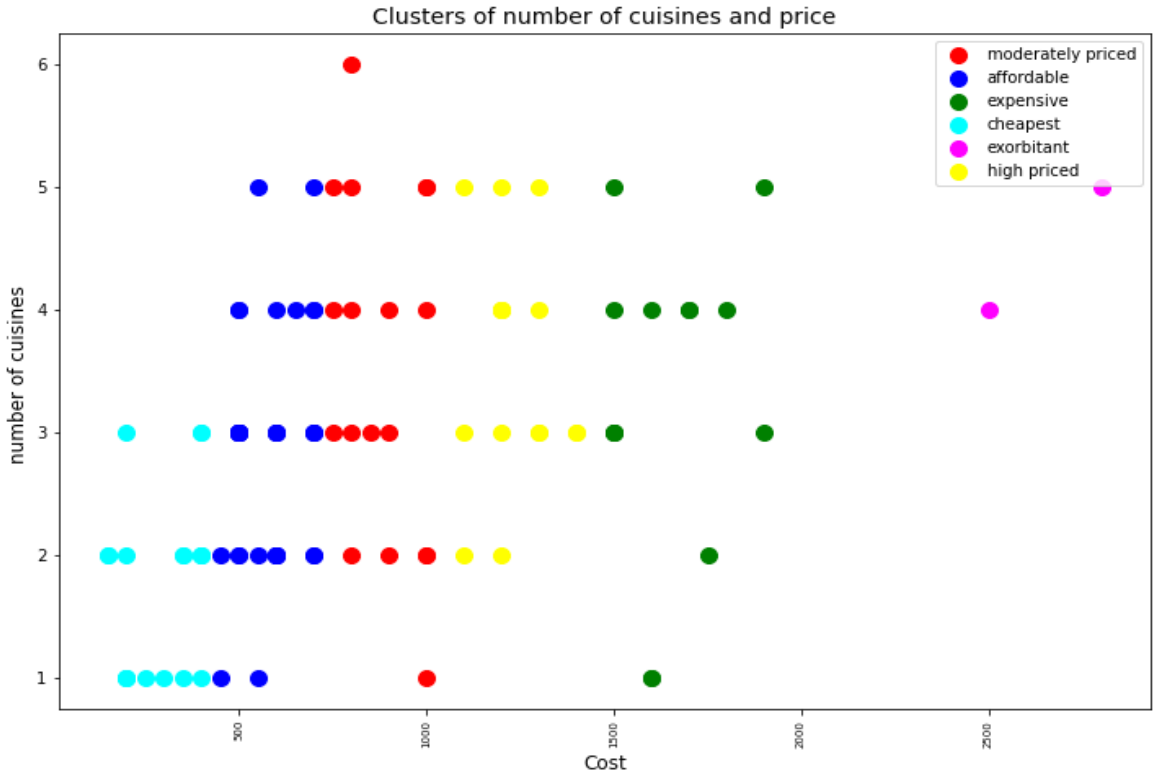
Cluster of restaurants based on price

- From the cluster distribution we can see that the biggest cluster is for the restaurants having cost between 400 and 700.
- The smallest cluster is having only two restaurants where food is heavily priced.



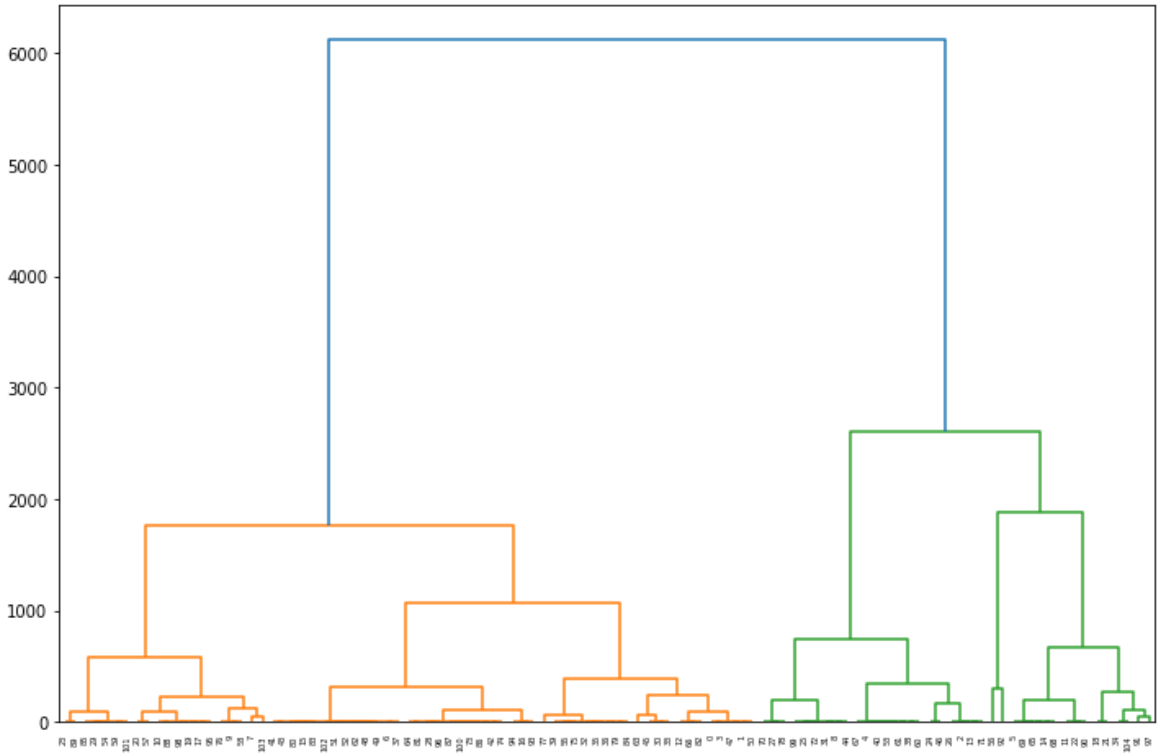
Cluster of number of cuisines and cost

- Cheaper restaurants have lesser number of cuisines
- Restaurants having 4 or 5 cuisines offer a larger range of prices



Hierarchical clustering : Dendogram

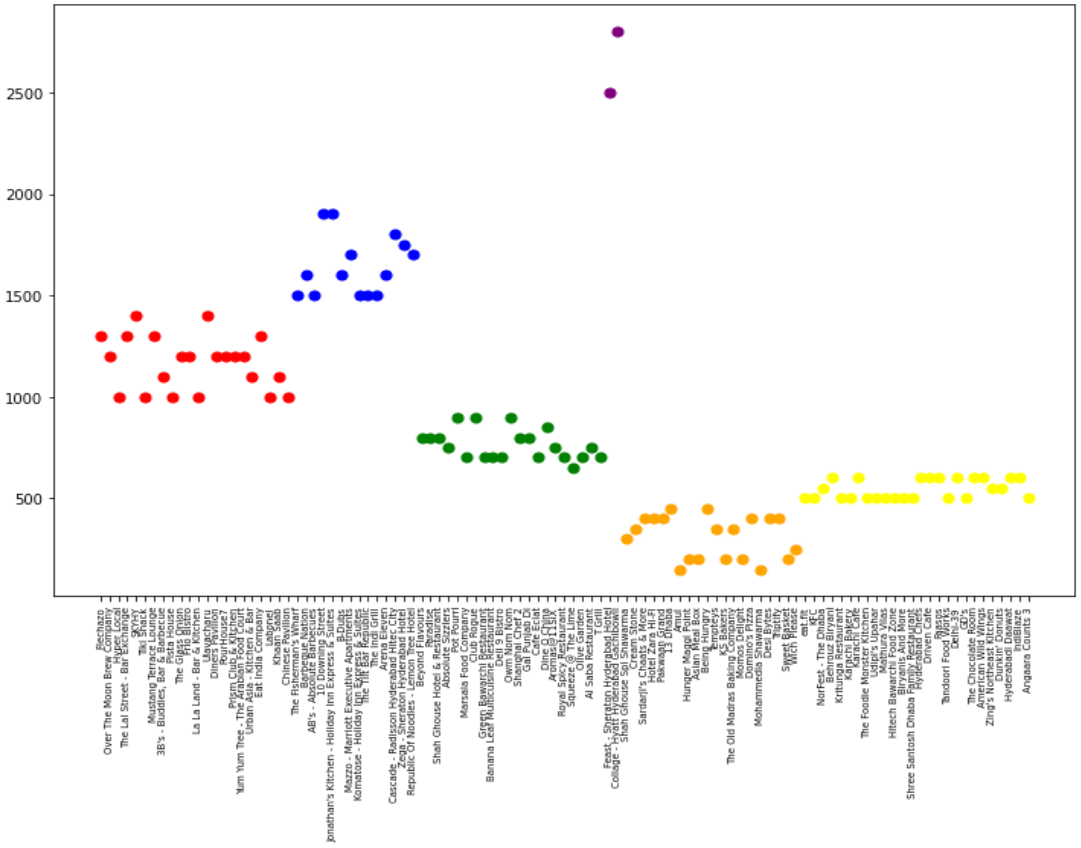
- A dendrogram is a diagram that shows the hierarchical relationship between objects.
- The main use of a dendrogram is to work out the best way to allocate objects to clusters.
- Here we if split along the line where Euclidean distance is around 1000 we get 6 clusters(as we cut 6 vertical lines)



Cluster distribution after doing hierarchical clustering

- The cluster distribution as we can see is similar to that what we get after doing k mean clustering

- Here also the smallest cluster is having only two restaurants where food is heavily priced.



Sentiment Analysis

Review Cleaning steps

- Removing urls from dataset
- Removing punctuations and special character
- Removing emoji pattern from given string
- Removing words with length less than or equal to 2.
- Removing stopwords
- Lemmatization

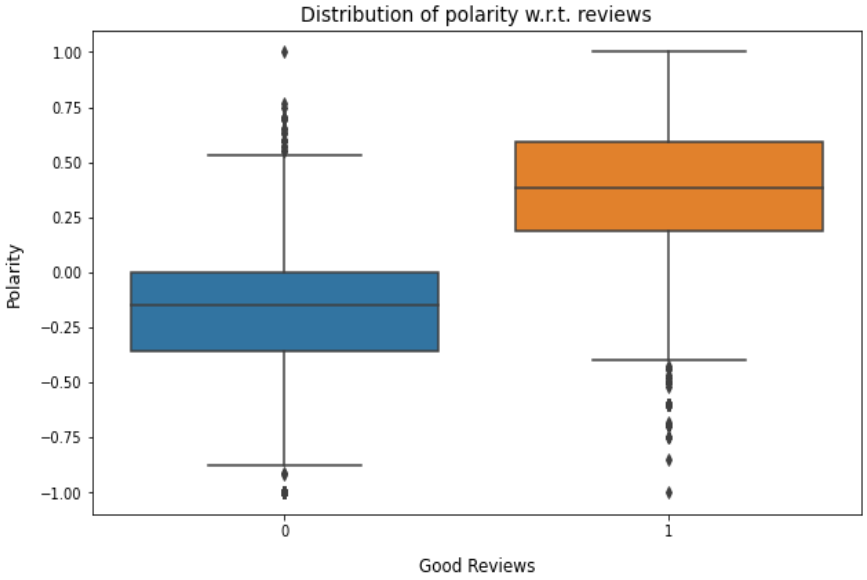
Polarity and Subjectivity

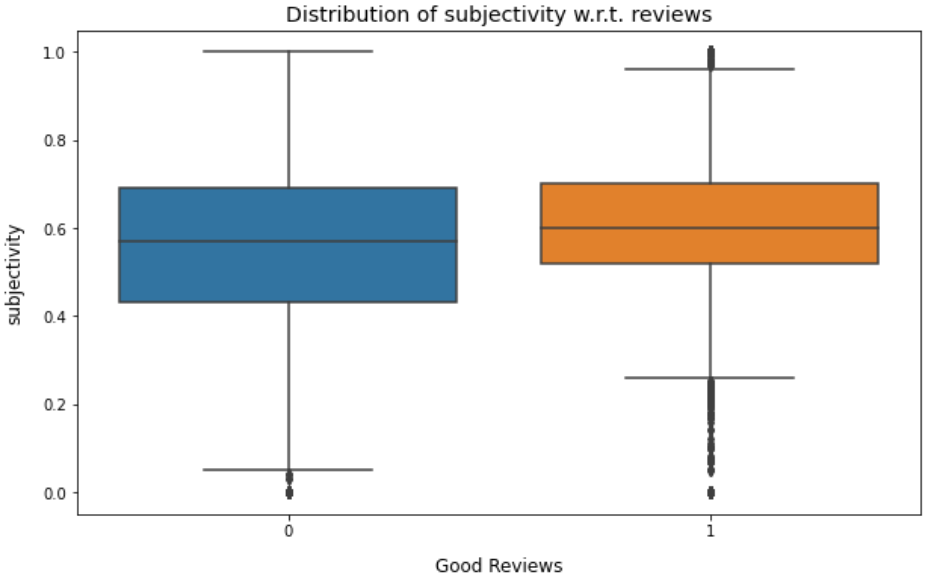
- Textblob sentiment analyzer returns two properties for a given input sentence:

- Polarity is a float that lies between [-1,1], -1 indicates negative sentiment and +1 indicates positive sentiments.

- Subjectivity is also a float which lies in the range of [0,1].
- Subjective sentences generally refer to personal opinion, emotion, or judgment.

Distribution of Polarity and Subjectivity





- We can observe that there are bad reviews with high polarity and good reviews with low polarity
- We can observe that the median of good reviews is higher than the median of bad reviews for subjectivity.

Pre-Processing data for fitting machine learning models

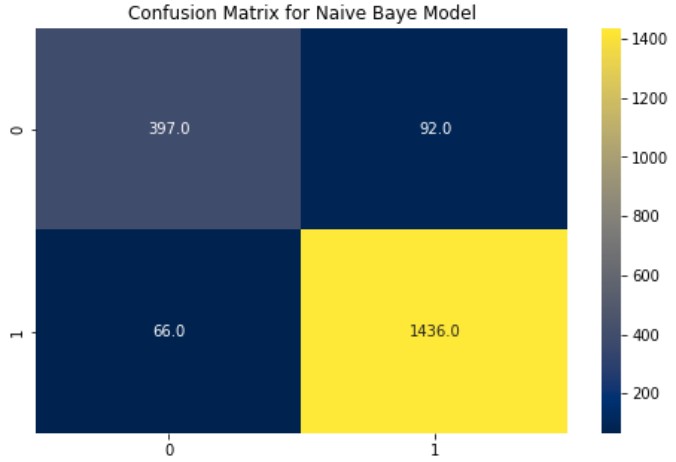
- Vectorization : It is the process of converting messages or text in to vectors that algorithms can work with

- The methods we used here is Count Vectorizer

- Count Vectorizer :It will transform the clean data to columns that represent a word or pair of words in the training dataset and the row count the frequency of each word/pair of words in each sentence.

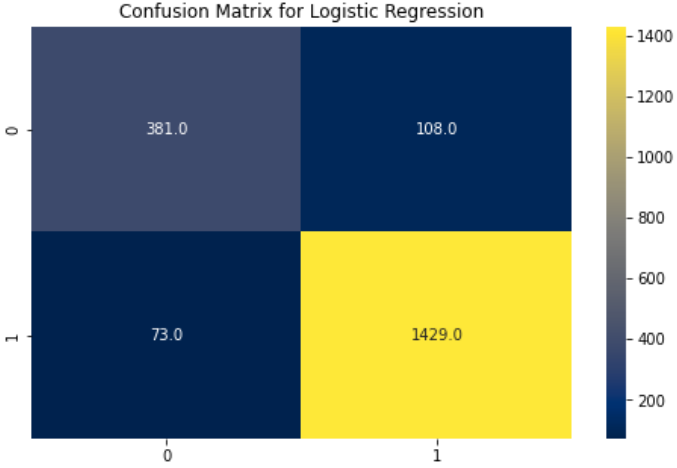
Naïve Bayes Model

- Accuracy:0.92
- Recall:0.92
- Precision:0.919
- F1-score:0.92
- ROC AUC score:0.88



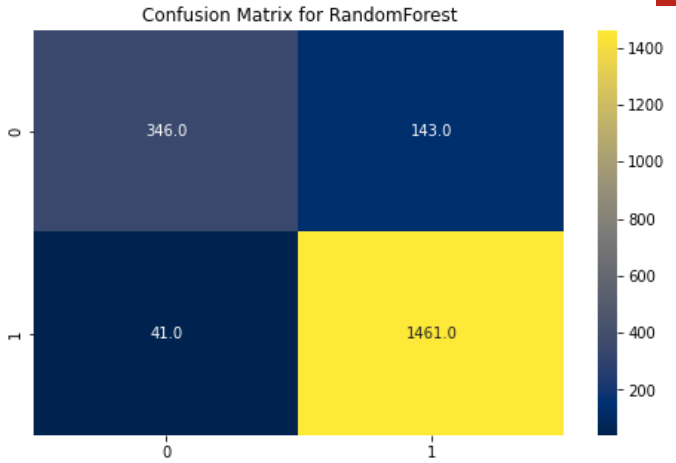
Logistic Regression

- Accuracy:0.909
- Recall:0.909
- Precision:0.907
- F1-score:0.907
- ROC AUC score:0.86



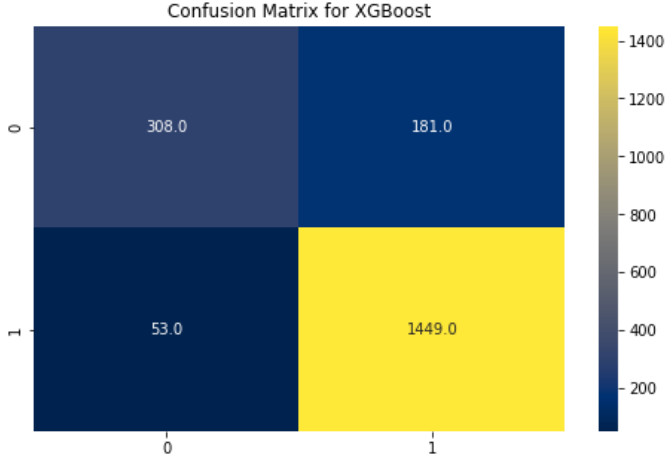
Random forest

- Accuracy:0.907
- Recall:0.907
- Precision:0.907
- F1-score:0.904
- ROC AUC score:0.84



XG Boost

- Accuracy:0.882
- Recall:0.882
- Precision:0.880
- F1-score:0.876
- ROC AUC score:0.797



Data frame of sentiment analysis prediction models with metrics

	MODEL NAME	ACCURACY	RECALL	PRECISION	F1-SCORE	ROC AUC SCORE
0	Naive_Baye(Multinomial)	0.920643	0.920643	0.919568	0.919900	0.883960
1	Logistic_Regression	0.909091	0.909091	0.907500	0.907927	0.865270
2	Random Forest-CountVectorizer	0.907584	0.907584	0.906724	0.903721	0.840135
3	XG Boost	0.882471	0.882471	0.880171	0.876023	0.797285

Conclusion

- The clustering analysis is done and the restaurants are divided in to 6 different clusters based on it.

- Sometimes people give good rating even though if they are giving bad reviews.

- The Naïve Bayes model gives the best result for sentimental analysis of the given dataset.
- It gives an accuracy, recall, precision and F1 score of 0.92 and an ROC-AUC score of 0.88.

THANK YOU