Capstone Project

Zomato Restaurant Clustering and Sentiment Analysis

Anuj Menaria

Points to Discuss:

- Defining the problem statement
- Data Description
- Dataset
- Handling Outliers
- Data Visualization & EDA
- Clustering
- Applying Model
- Hyperparameter Tuning
- Conclusion

Problem Statement:

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.

India is quite famous for its diverse multi-cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. The 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 analyzing the Zomato restaurant data for each city in India.

The Project focuses on Customers and companies, you have to analyze the sentiments of the reviews given by the customer in the data and made some useful conclusions in the form of Visualizations. Also, cluster the Zomato restaurants into different segments. The data is visualized as it becomes easy to analyze data in an instant. The Analysis also solve some of the business cases that can directly help the customers find the Best restaurant in their locality and for the company to grow up and work in the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also, the data has valuable information about cuisine and cost 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 DESCRIPTION:

❑ Zomato Restaurant names and Metadata:

Use this dataset for the clustering part.

- **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

DATA DESCRIPTION(Contd.):

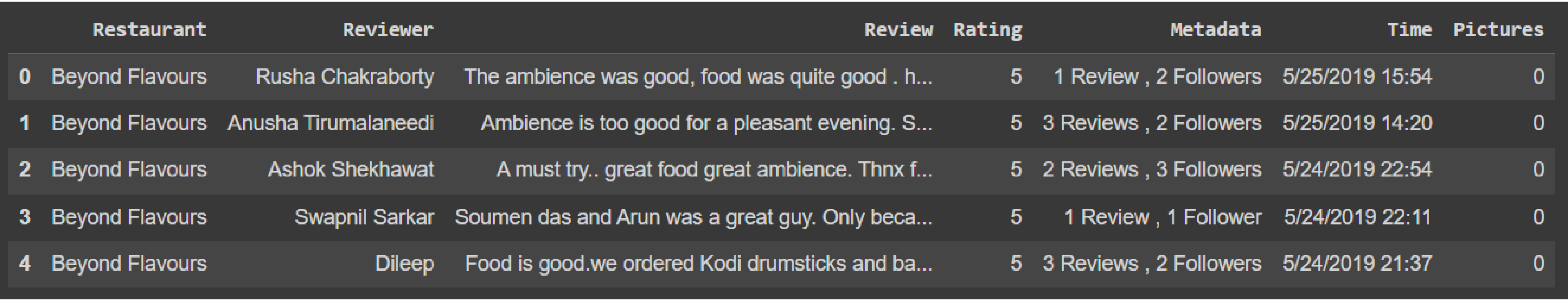
❑ Zomato Restaurant reviews:

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

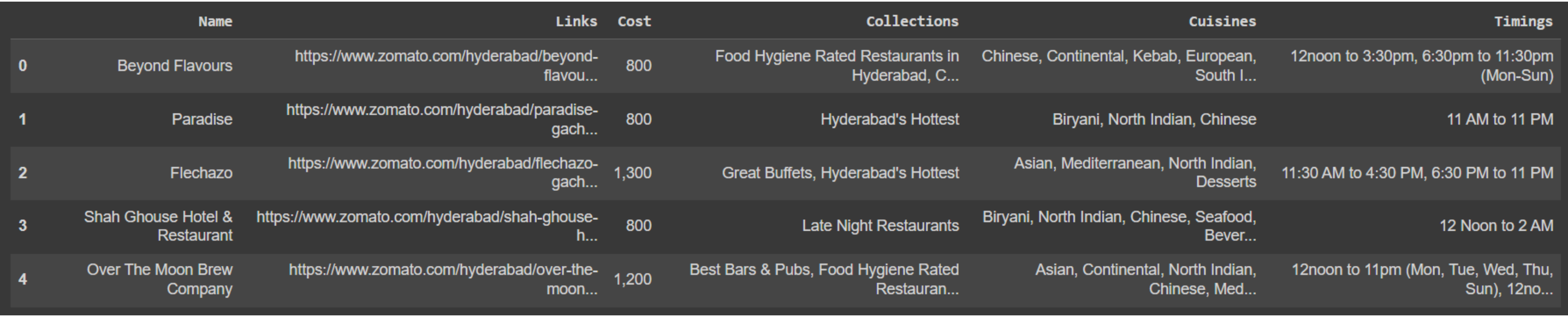
- **Restaurant:** Name of the Restaurant
- **Reviewer:** Name of the Reviewer
- **Review:** Review Text
- **Rating:** Rating Provided by the Reviewer
- **Metadata:** Reviewer Metadata - No. of Reviews and followers
- **Time:** Date and Time of Review
- **Pictures:** No. of pictures posted with review

Dataset:

- After Loading the Metadata dataset we can observe that it has : Rows: **105,** Columns: **6**
- After Loading the Reviews dataset we can observe that it has : Rows: **10000,** Columns: **7**

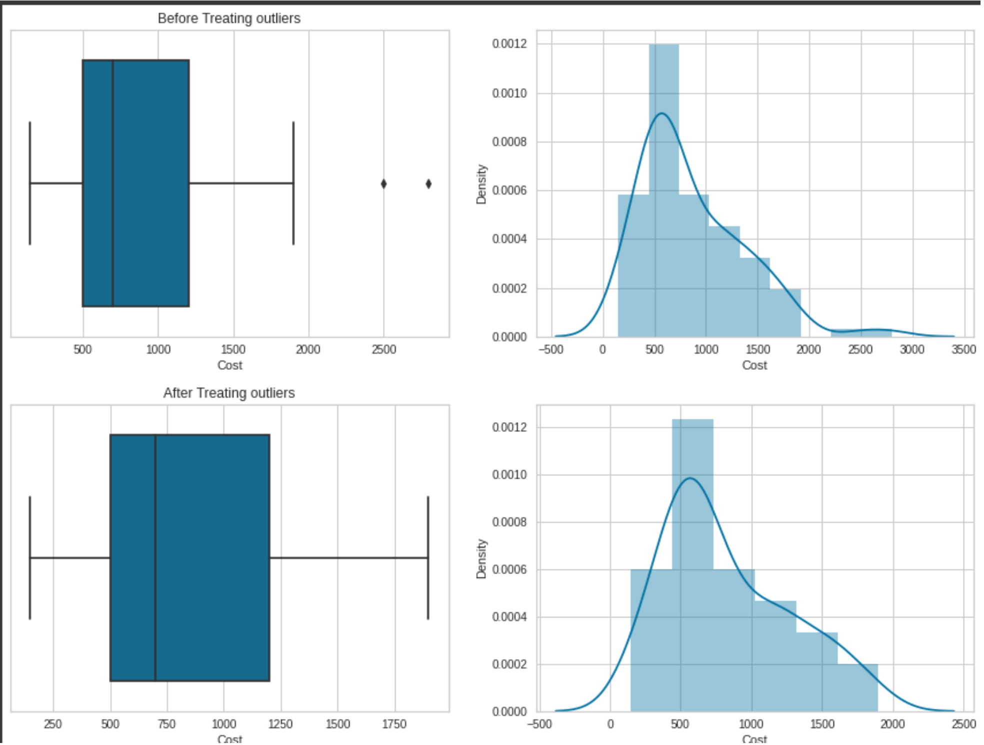


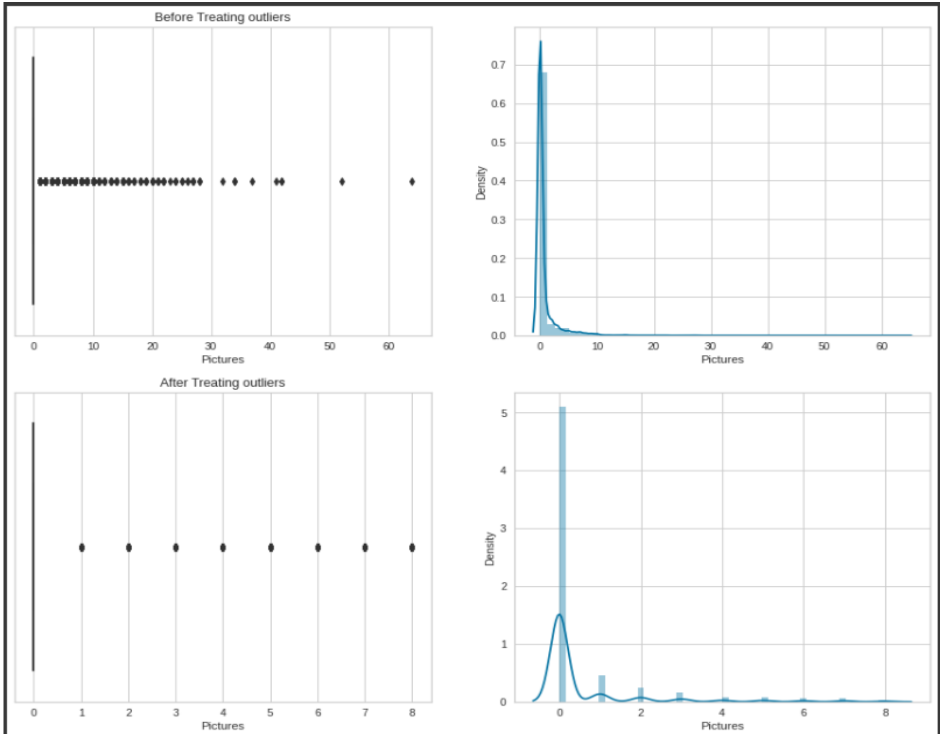
	Restaurant	Reviewer	Review	Rating	Metadata	Time	Pictures
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . h...	5	1 Review , 2 Followers	5/25/2019 15:54	0
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. S...	5	3 Reviews , 2 Followers	5/25/2019 14:20	0
2	Beyond Flavours	Ashok Shekhawat	A must try.. great food great ambience. Thnx f...	5	2 Reviews , 3 Followers	5/24/2019 22:54	0
3	Beyond Flavours	Swapnil Sarkar	Soumen das and Arun was a great guy. Only beca...	5	1 Review , 1 Follower	5/24/2019 22:11	0
4	Beyond Flavours	Dileep	Food is good.we ordered Kodi drumsticks and ba...	5	3 Reviews , 2 Followers	5/24/2019 21:37	0



	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	800	Hyderabad's Hottest	Biryani, North Indian, Chinese	11 AM to 11 PM
2	Flechazo	https://www.zomato.com/hyderabad/flechazo-gach...	1,300	Great Buffets, Hyderabad's Hottest	Asian, Mediterranean, North Indian, Desserts	11:30 AM to 4:30 PM, 6:30 PM to 11 PM
3	Shah Ghouse Hotel & Restaurant	https://www.zomato.com/hyderabad/shah-ghouse-h...	800	Late Night Restaurants	Biryani, North Indian, Chinese, Seafood, Bever...	12 Noon to 2 AM
4	Over The Moon Brew Company	https://www.zomato.com/hyderabad/over-the-moon...	1,200	Best Bars & Pubs, Food Hygiene Rated Restauran...	Asian, Continental, North Indian, Chinese, Med...	12noon to 11pm (Mon, Tue, Wed, Thu, Sun), 12no...

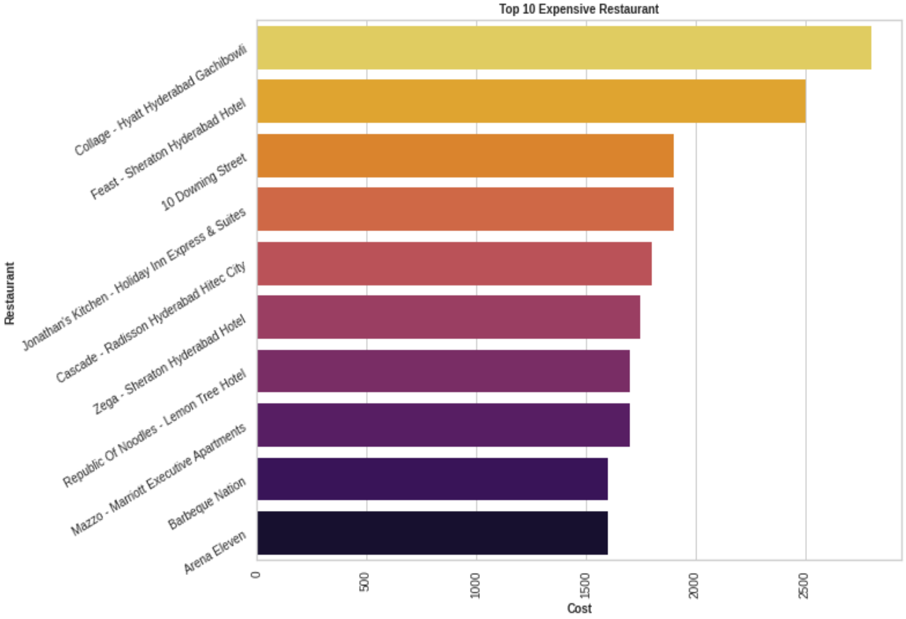
Handling Outliers:

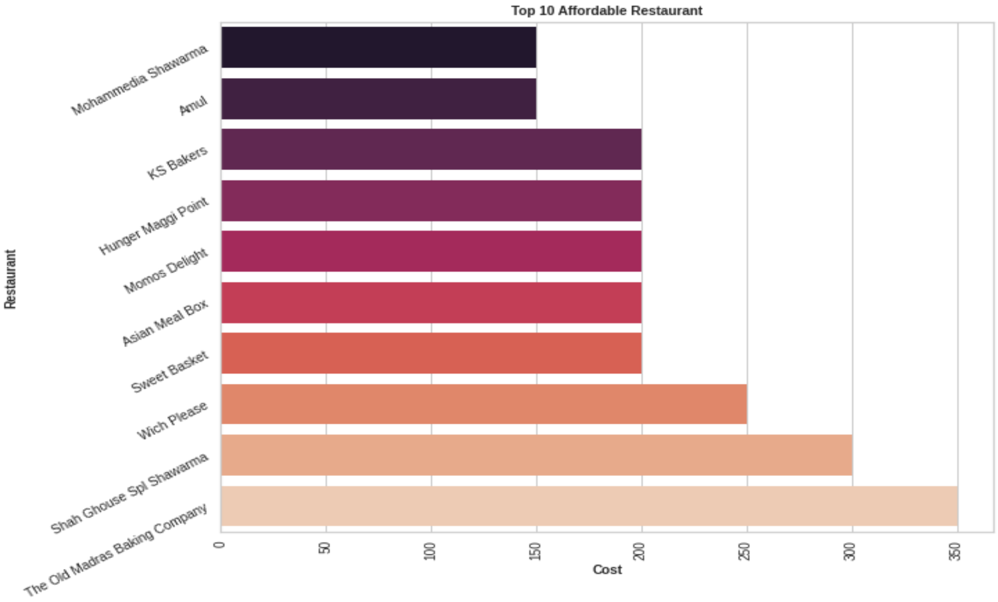




- As we can see from the metadata, the "Cost" Column includes outliers, but following treatment, the data is regularly distributed.
- The "Pictures" column in the reviews data frame contains the majority of the outliers that have been well addressed, but some outliers are still visible.

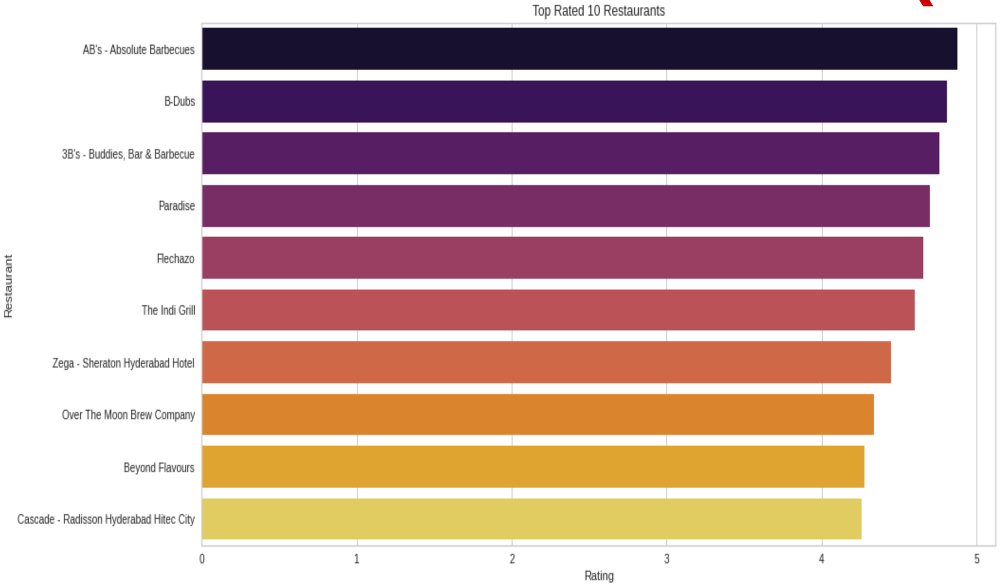
Data Visualization & EDA:

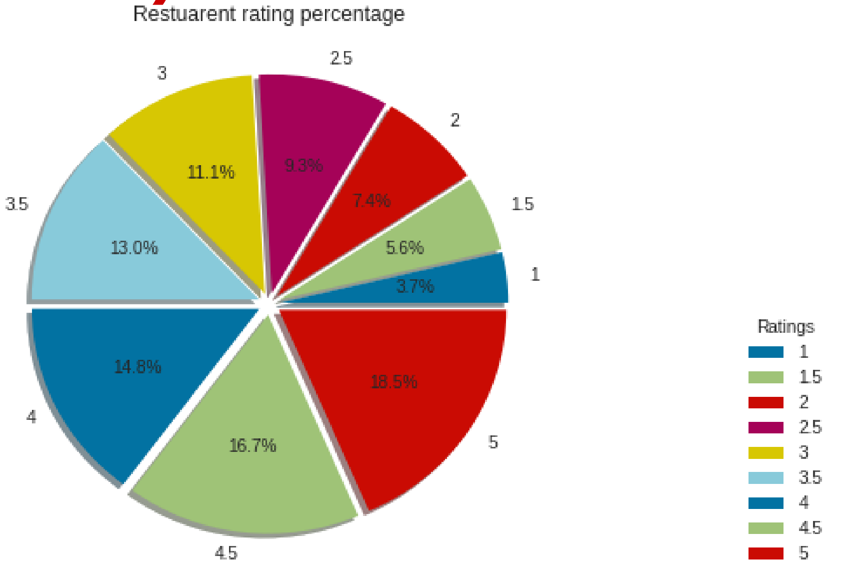




- The most expensive restaurant is Collage - Hyatt Hyderabad Gachibowli.
- The price of the most expensive Restaurant is around Rs. 2800 approx.
- The most economical dining establishments are Amul and Mohammedia Shawarma.
- The price of the most Affordable Restaurant is around Rs. 150 approx.

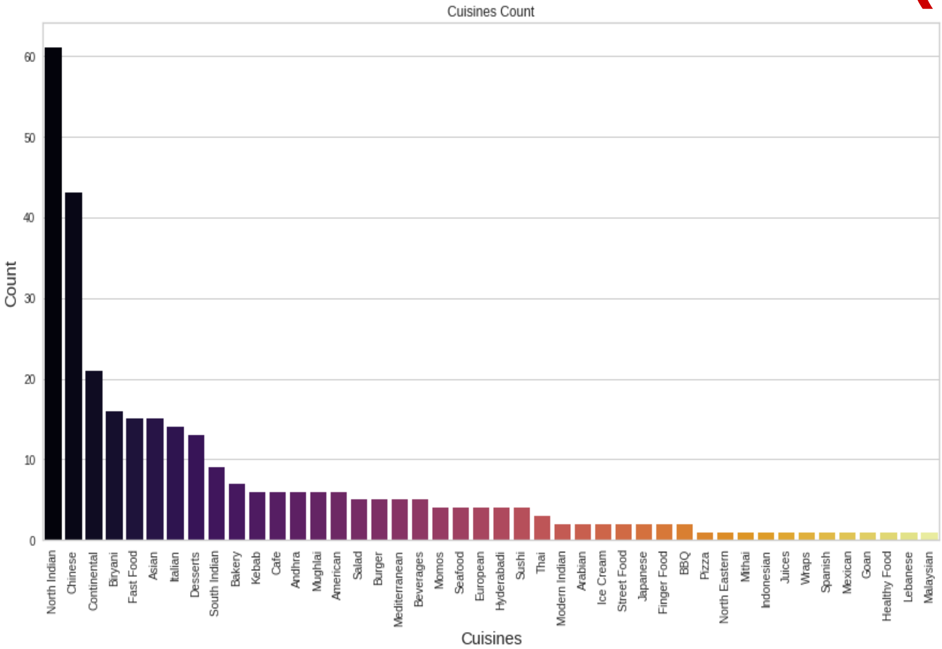
Data Visualization & EDA(Contd.):

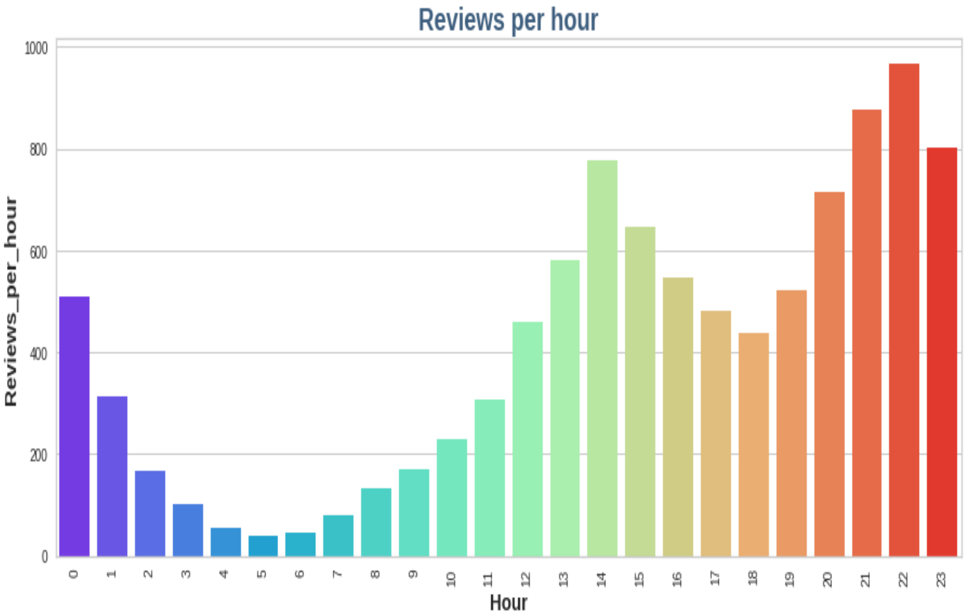




- The top eateries include Bar & BBQ, B-Dubs, and **AB's - Absolute Barbecues**.
- According to ratings, the top-rated restaurants are highly recommended and have ratings of more than 4.8 on average.
- The pie chart of ratings shows that restaurants with 4.5 and 5 stars have the highest percentage of ratings.

Data Visualization & EDA(Contd.):



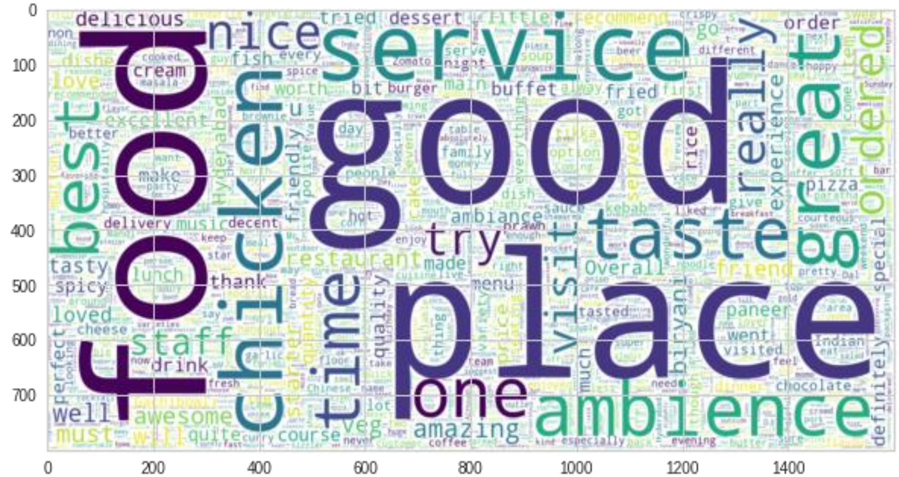


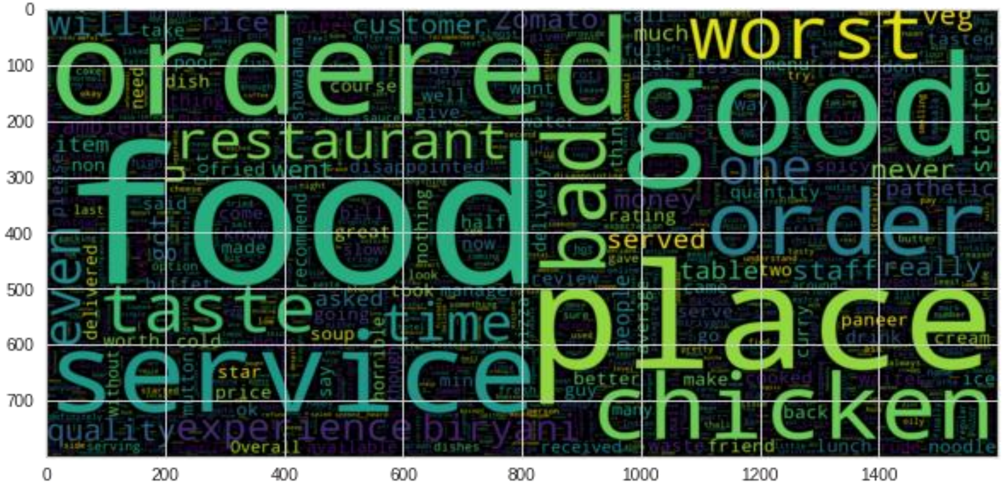
- We can notice that practically all eateries provide North Indian as their most popular cuisine. And "Malaysian" food is a delicacy.
- After north Indian cuisine, we can notice that Chinese and continental cuisines are also in high demand.
- The majority of the reviews take place between 9:00 and 10:00 at night.
- Fewer reviews are issued between 5:00 and 6:00 in the morning.

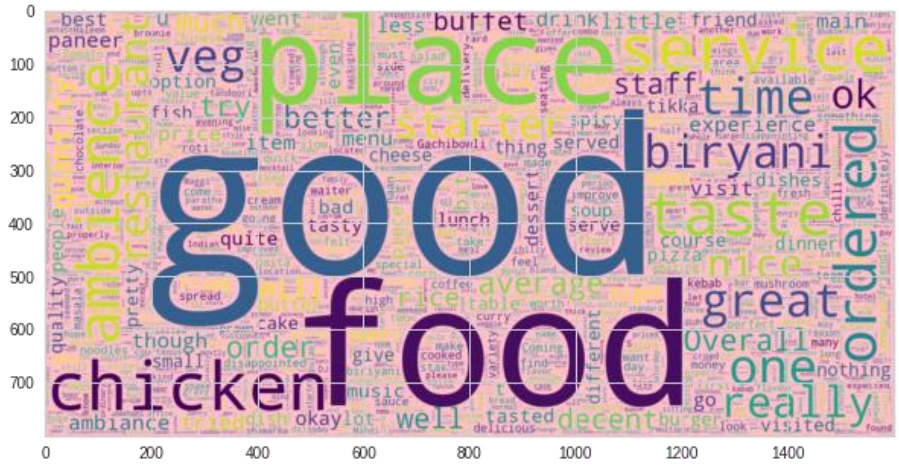
Text Preprocessing:

- **Lower Casing:** Converting a word to lowercase (NLP -> nlp). Words like Book and book mean the same but when not converted to lowercase 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, teor xt into a list of tokens. One can think of tokens 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 words data and data! are treated equally after the process of removal of punctuation
- **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 the process s of grouping together different forms of the same word and converting words into base or root form.

Data Visualization & EDA(Contd.):

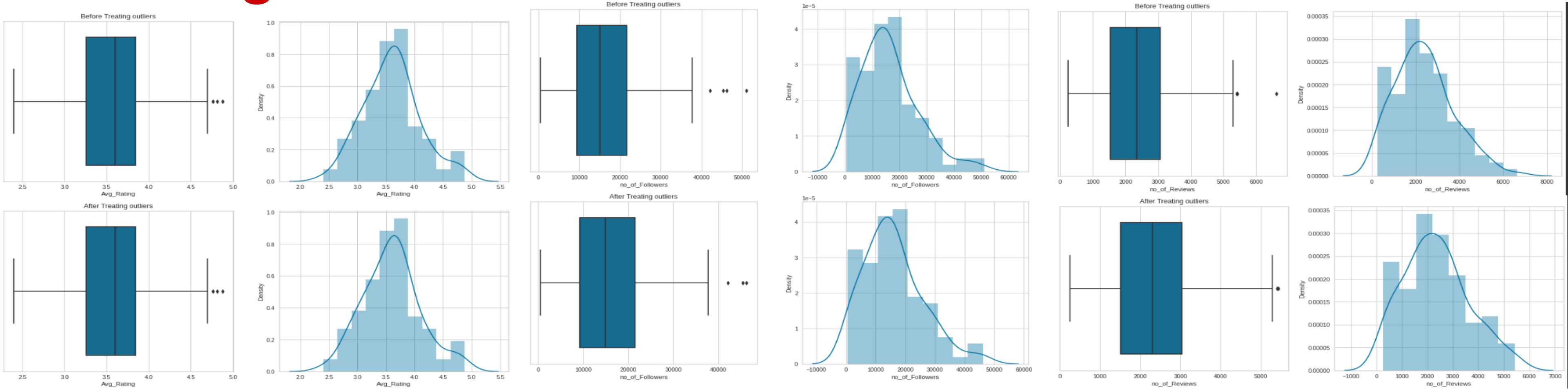


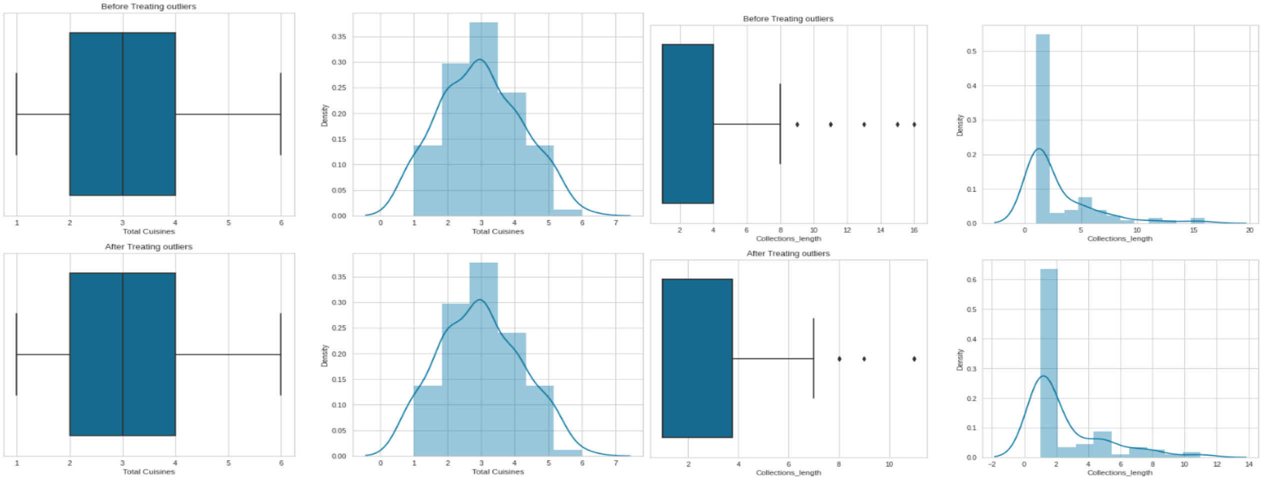




- To determine which word groups are most frequently used in the evaluations, a word cloud was produced.
- The words "good place" and "service" are utilized, as is evident from the Good reviews.
- We can notice that the place, chicken, ordered is utilized in negative reviews.
- In average reviews food, the chicken, place is used.

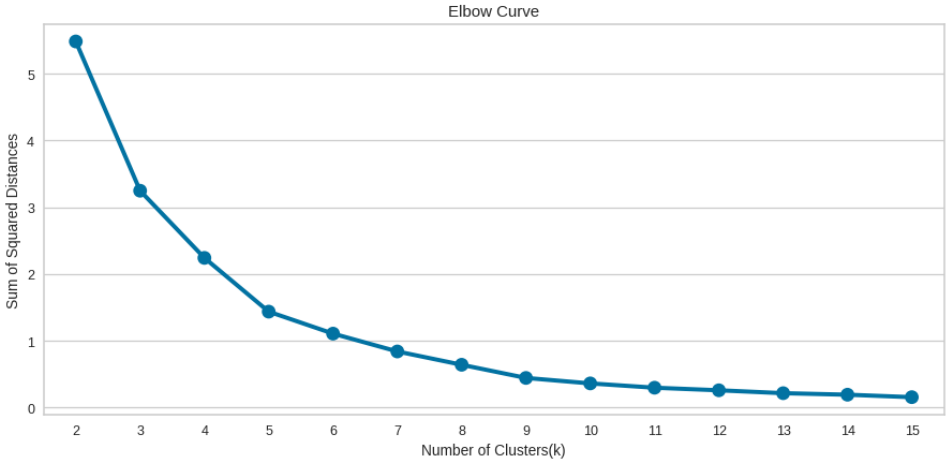
Handling Outliers:

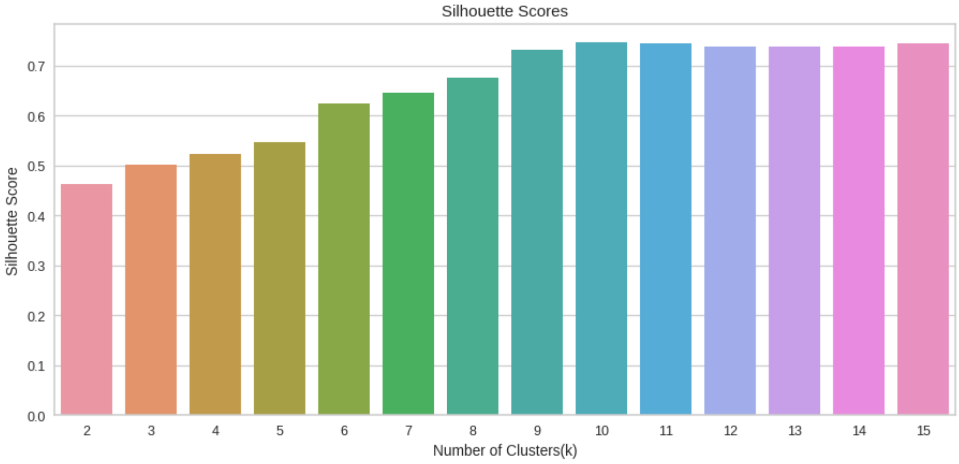




- We added additional variables for future processing, one of which was a number column including all other numerical columns.
- We handled all the outliers present in the numerical data frame.

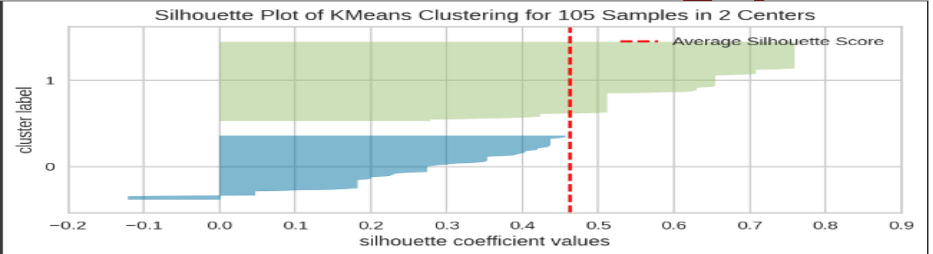
K-Means Clustering:

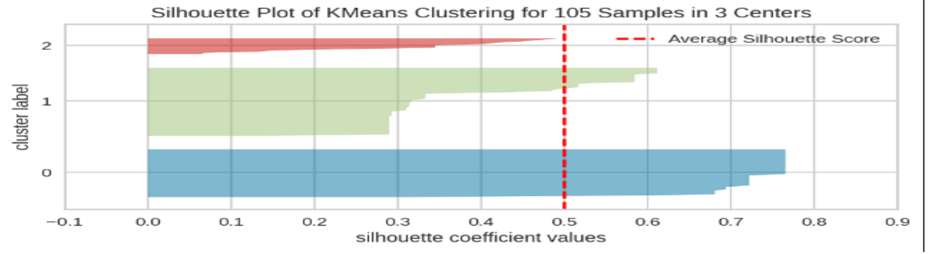


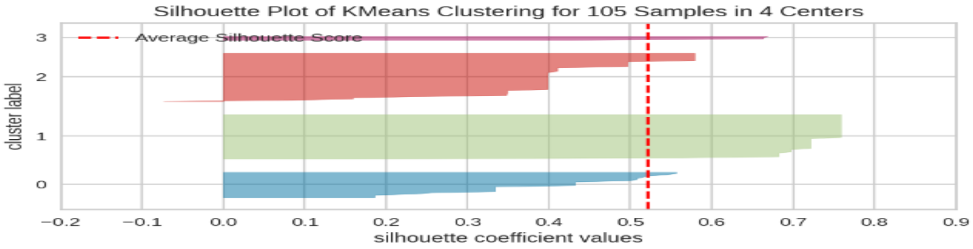


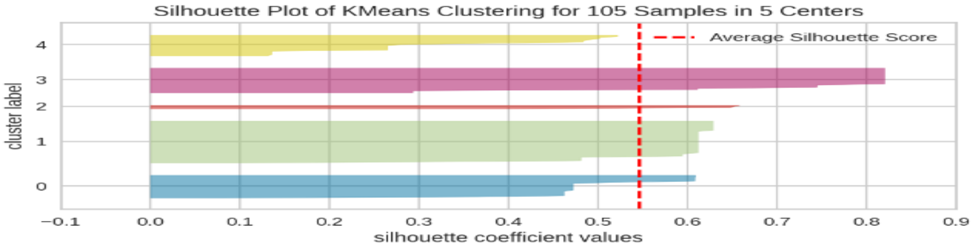
- Unsupervised learning algorithm K-Means Clustering divides the unlabeled dataset into several clusters. Here, K specifies how many pre-defined clusters must be produced as part of the process; for example, if K=2, there will be two clusters, if K=3, there will be three clusters, and so on.
- The elbow technique applies k-means clustering to the dataset using a range of k values (for example, 1-10) and computes the average score for each value of k. By default, the distortion score the sum of the square distances between each point and the center to which it is assigned—is calculated.

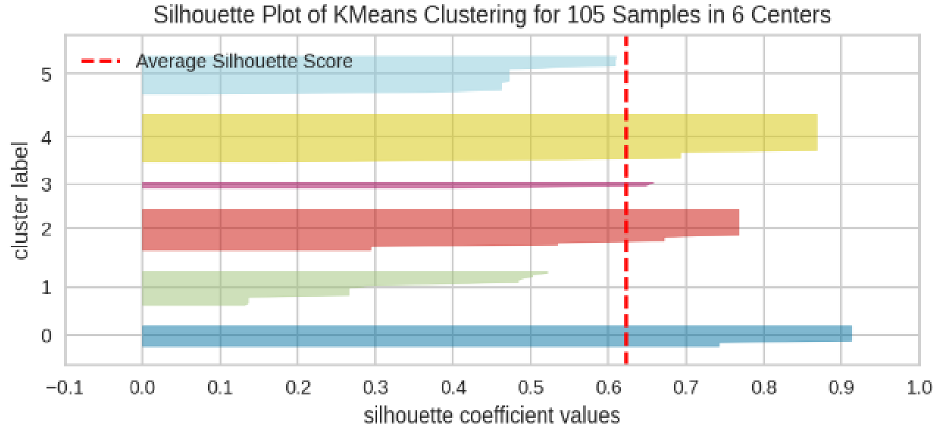
K-Means Clustering (Contd.):


Silhouette Plot of KMeans Clustering for 105 Samples in 2 Centers


Silhouette Plot of KMeans Clustering for 105 Samples in 3 Centers

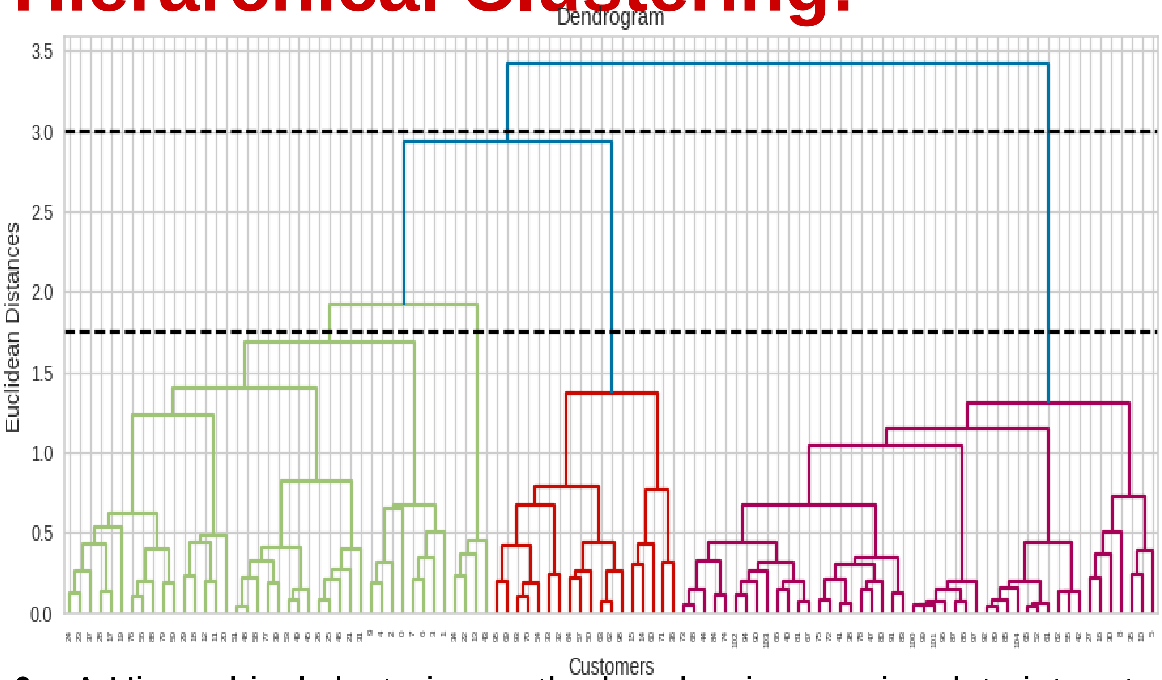

Silhouette Plot of KMeans Clustering for 105 Samples in 4 Centers

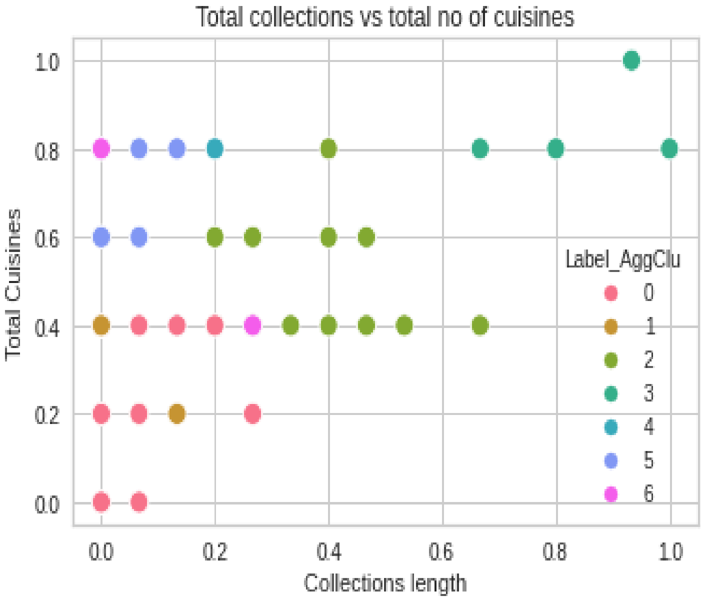

Silhouette Plot of KMeans Clustering for 105 Samples in 5 Centers


Silhouette Plot of KMeans Clustering for 105 Samples in 6 Centers

- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters.

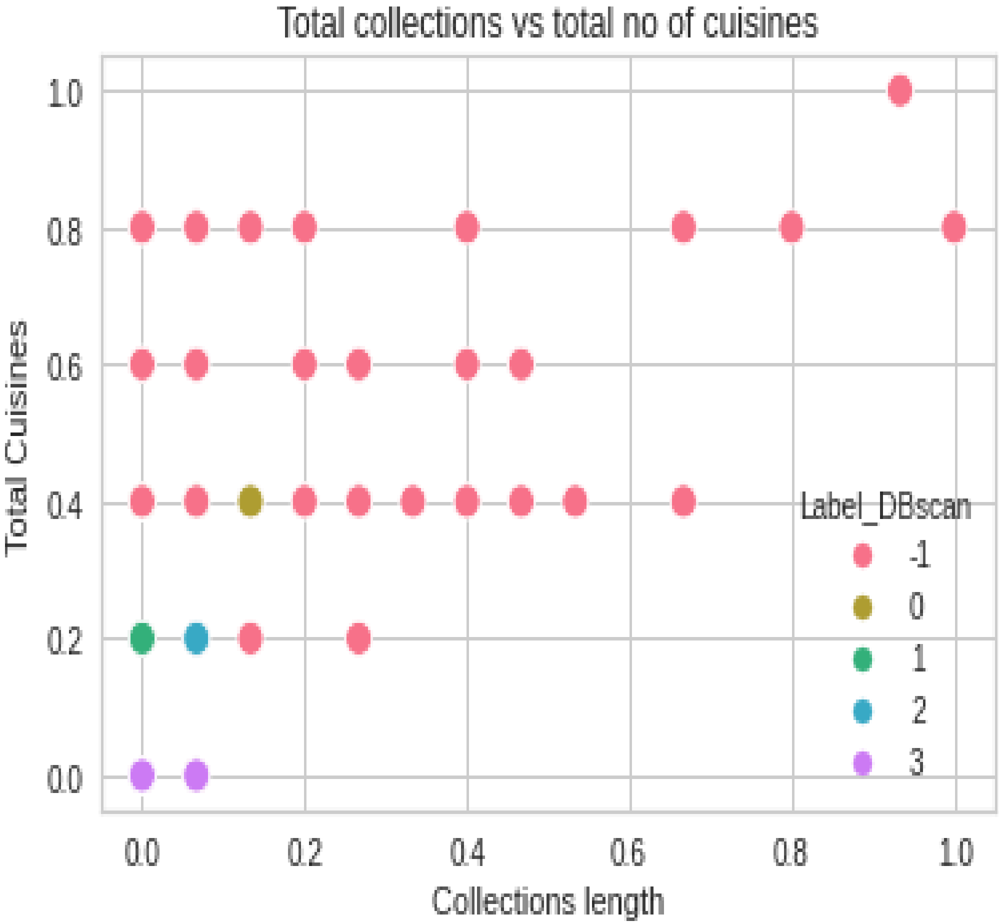
Hierarchical Clustering:





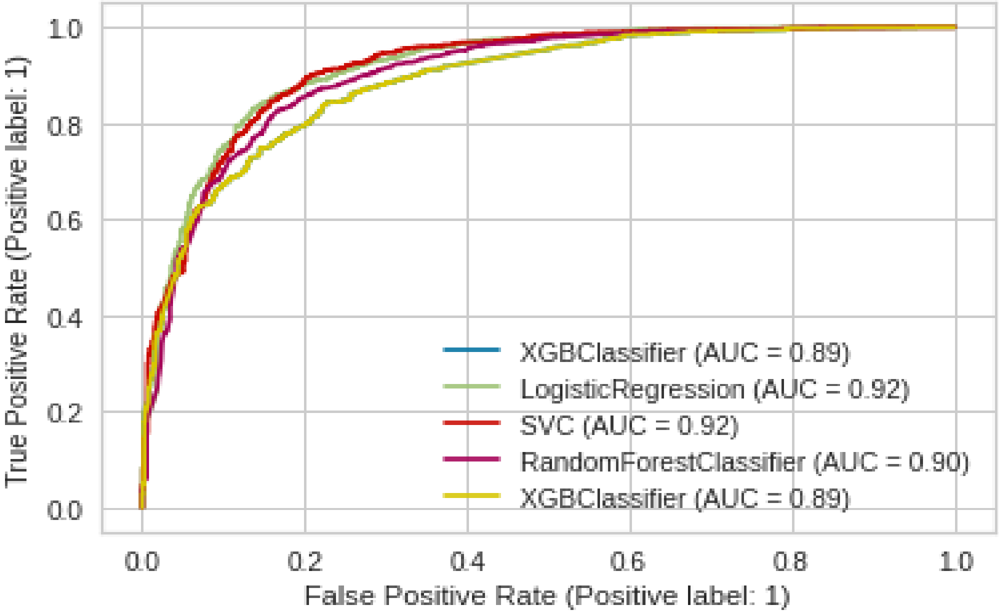
- A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps: Identify the 2 clusters which can be closest together, and. Merge the 2 maximum comparable clusters.
- 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 if split along the line where Euclidean distance is around 1000 we get 6 clusters(as we cut 6 vertical lines)

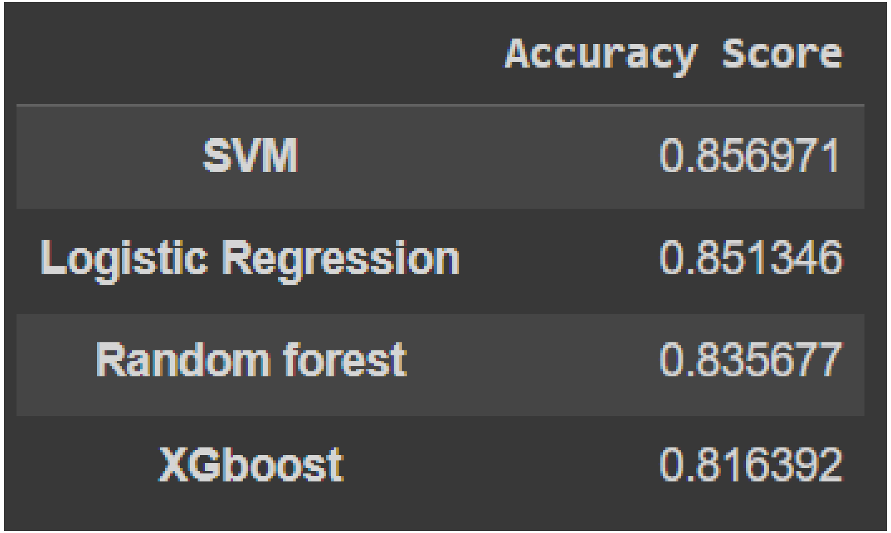
Dbscan Clustering:



- A cluster, according to DBSCAN, is defined as an area with a high point density that is isolated from other clusters by regions with a low point density.

ROC Curve & Accuracy Score:

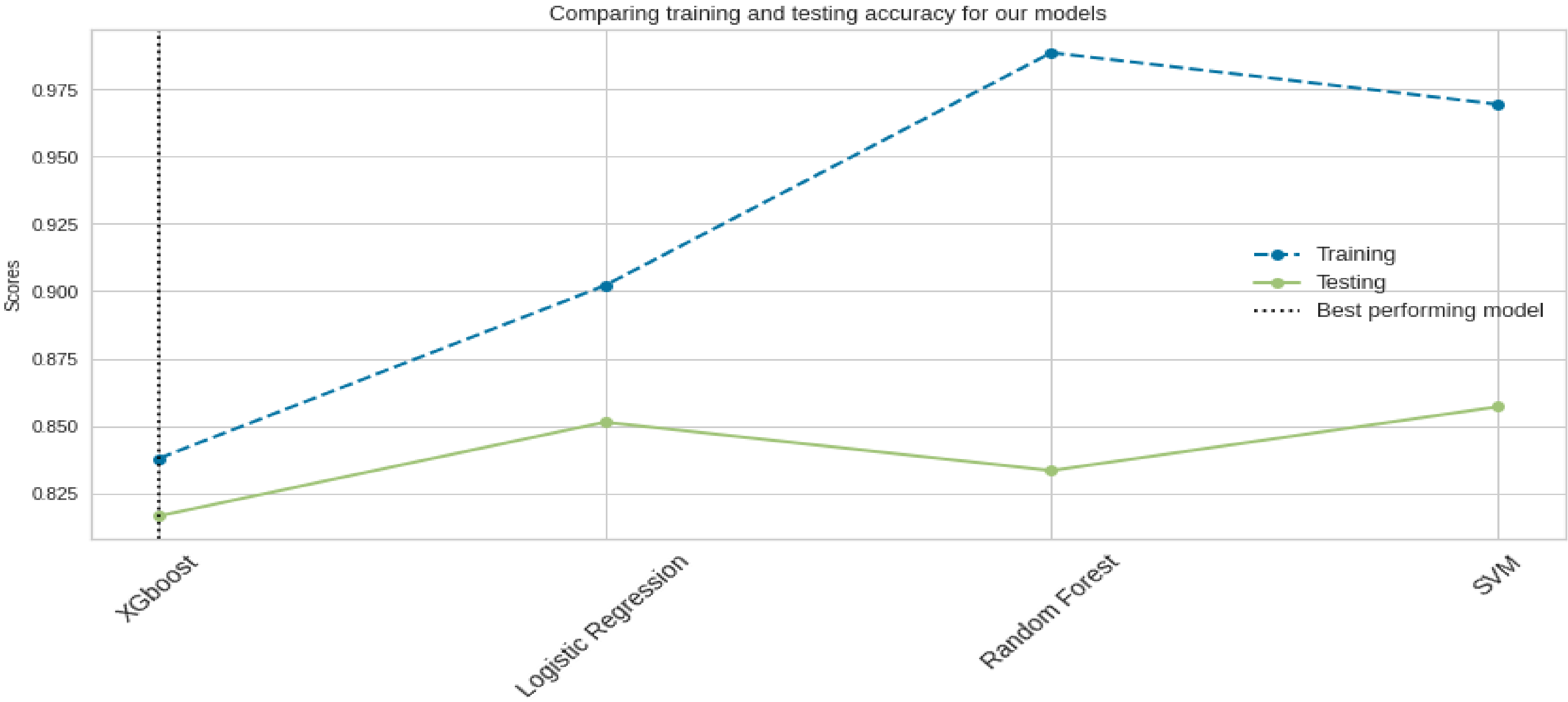




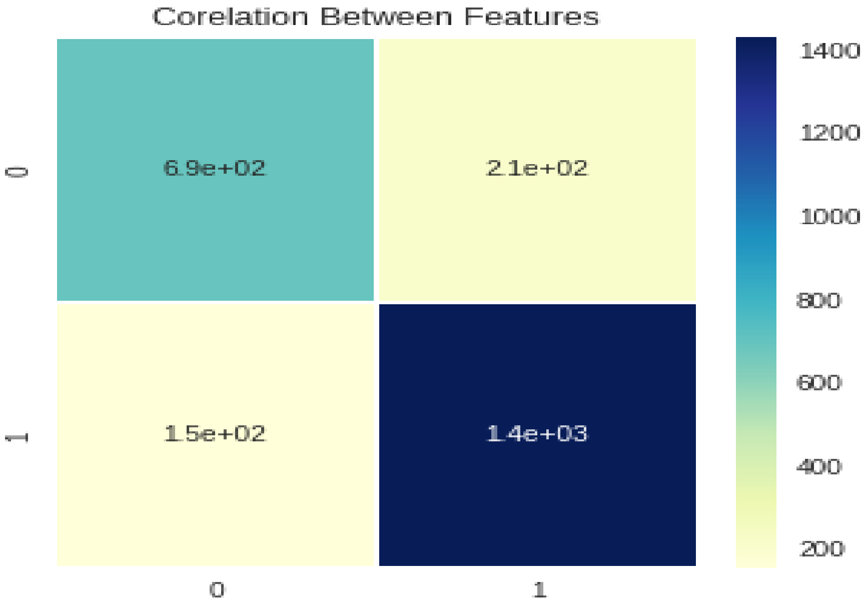
	Accuracy Score
SVM	0.856971
Logistic Regression	0.851346
Random forest	0.835677
XGboost	0.816392

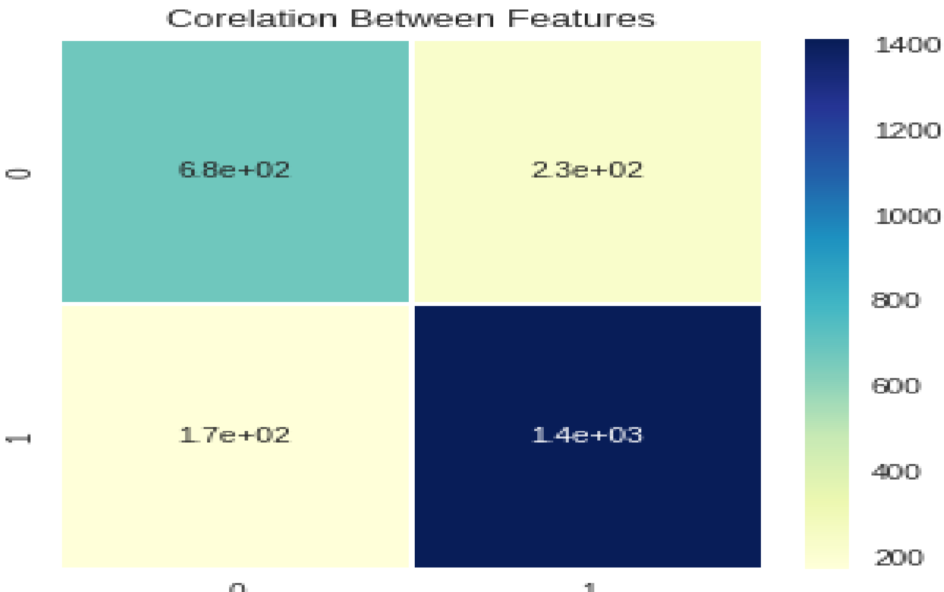
- We have used XGB, SVC, Logistic Regression, Random Forest, and Regression Algorithms in Regression.
- After executing each method, we can observe that SVC and logistic regression have the highest accuracy scores in the training dataset.
- With SVM, the best outcome may be observed after predicting the score.

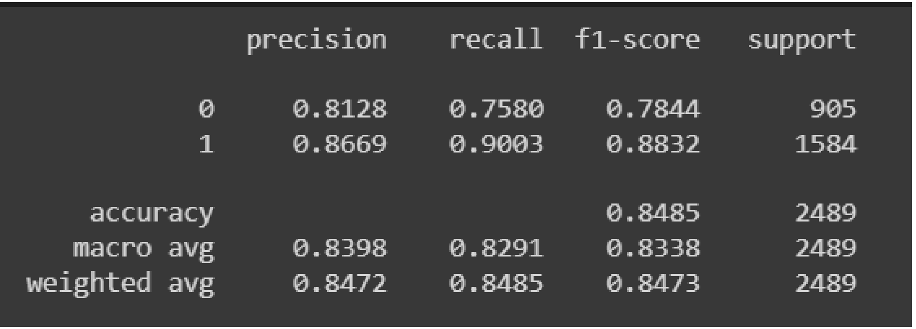
Accuracy Comparison:



Confusion Matrix:



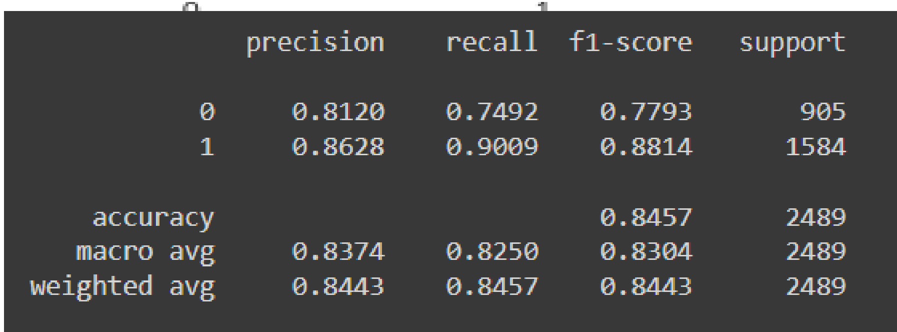




```
              precision    recall  f1-score   support

           0     0.8128    0.7580    0.7844       905
           1     0.8669    0.9003    0.8832      1584

    accuracy                         0.8485      2489
   macro avg     0.8398    0.8291    0.8338      2489
weighted avg     0.8472    0.8485    0.8473      2489
```



```
              precision    recall  f1-score   support

           0     0.8120    0.7492    0.7793       905
           1     0.8628    0.9009    0.8814      1584

    accuracy                         0.8457      2489
   macro avg     0.8374    0.8250    0.8304      2489
weighted avg     0.8443    0.8457    0.8443      2489
```

Conclusion:

- The most typical food to be found in restaurants is North Indian cuisine.
- The most expensive restaurant is Collage - Hyatt Hyderabad Gachibowli.
- The most economical dining establishments are Amul and Mohammedia Shawarma.
- The top eateries include Buddies, Bar & BBQ, B-Dubs, and AB's - Absolute Barbecues.
- The most frequent word in sentiments of extreme positivity is good.
- The most frequent word in sentiments of extreme Negatively is worst.
- Restaurants Arena Elven and Banana Leaf Multicuisine have received the most negative comments.
- It is crucial to separate the restaurants with the lowest rating toorder to enhance the overall customer experience, according to the results of a simple cost-benefit analysis on Zomato conducted with a few assumptions as the foundation for the little business expertise that could be acquired. These establishments were little eateries or ones that charged a lot for the meals they served. More effort should be put into advertising, and reviews, particularly for these eateries, should be examined and improved. It appears that Mohammedia Shawarma is lucrative.

Conclusion(Contd.):

- The reviews were subjected to sentiment analysis, and a model was developed to distinguish between good and negative attitudes. Logistic regression performs better in terms of lowering False positives, although having a greater rate of false negatives. As a result, Logistic Regression appears to be punishing False Positives more harshly as desired.

- Ratings have to be gathered according to categories, such as packaging, delivery, taste, excellence, amount, and service. This would aid in focusing on particular fields that are falling behind.

- The score of the XGB Classifier after hyperparameter tuning is 83%. Conversely, logistic regression is effective, scoring 84%.

Thank You.