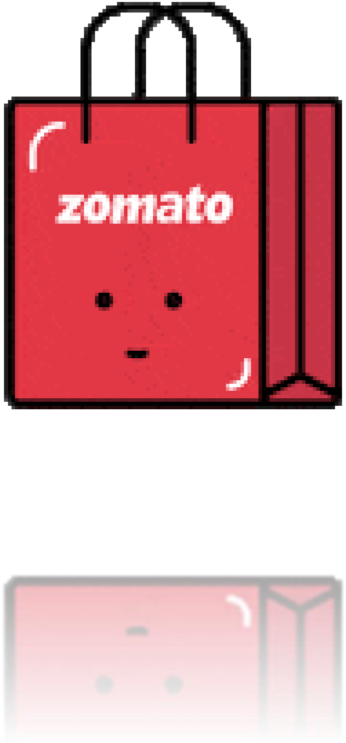
Capstone Project-4

Unsupervised ML-
Zomato Restaurant Clustering and Sentiment Analysis

Presented by:
Vishal Raul

Content

1. Introduction
2. Defining Problem Statement
3. Data Summary
4. Checking and imputing outliers
5. Exploratory Data Analysis(EDA)
6. ML Model-Clustering
7. Sentiment Analysis
8. Conclusion



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.

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.

This Project focuses on Customers and Company, 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.

Defining Problem Statement

- Our task is 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 solve some of the business cases that can directly help the customers finding the Best restaurant in their locality

Data Summary:

For Clustering

Data Set Name :- Zomato Restaurant names and Metadata.csv

Statistics : –

- Rows - 105
- Features - 6

Data Fields:-

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 Summary:

For Sentiment Analysis

Data Set Name :- Zomato Restaurant reviews.csv

Statistics : –

- Rows - 10000
- Features - 7

Data Fields:-

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

Approach Overview

Data Cleaning

Data Exploration

Predictive Modeling

Understand and Clean

- Find information on undocumented columns values
- Clean data to get it ready for analysis
- Null values treatment
- Outlier Treatment

Graphical and Statistical

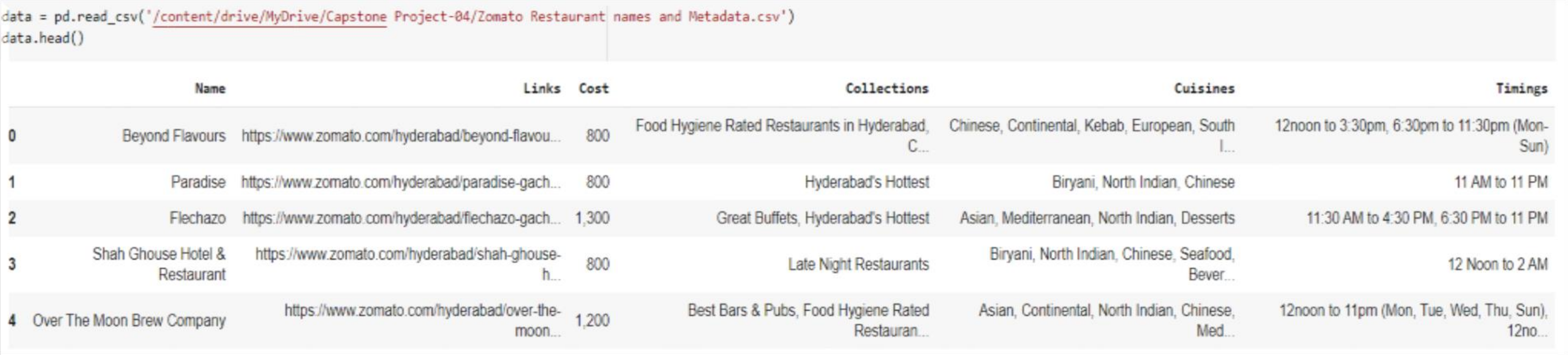
- Univariate analysis with visualization
- Bivariate Analysis with visualization

Machine Learning

- Clustering
- Topic Modeling
- Classification

Restaurant Clustering

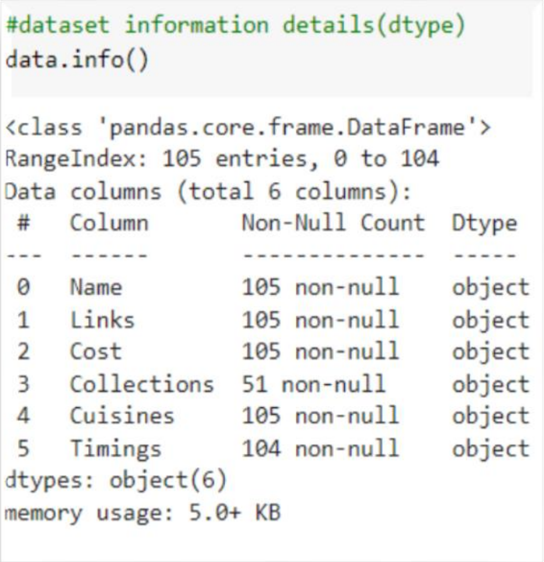
Loading the Dataset



```
data = pd.read_csv('/content/drive/MyDrive/Capstone Project-04/Zomato Restaurant names and Metadata.csv')
data.head()
```

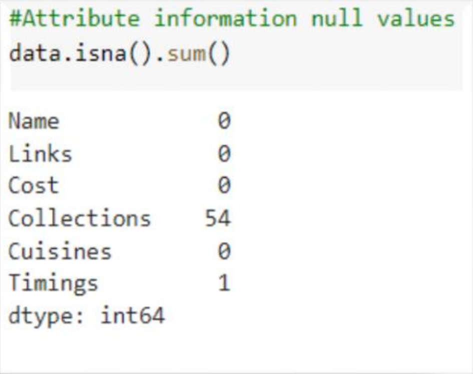
	Name	Links	Cost	Collections	Cuisines	Timings
0	Beyond Flavours	https://www.zomato.com/hyderabad/beyond-flavou...	800	Food Hygiene Rated Restaurants in Hyderabad, C...	Chinese, Continental, Kebab, European, South I...	12noon to 3:30pm, 6:30pm to 11:30pm (Mon-Sun)
1	Paradise	https://www.zomato.com/hyderabad/paradise-gach...	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...

Attribute Information:- Dtypes and Null values



```
#dataset information details(dtype)
data.info()
```

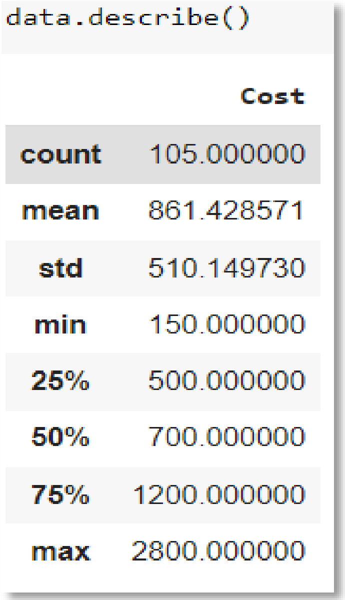
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105 entries, 0 to 104
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Name         105 non-null    object
 1   Links        105 non-null    object
 2   Cost         105 non-null    object
 3   Collections  51 non-null     object
 4   Cuisines     105 non-null    object
 5   Timings      104 non-null    object
dtypes: object(6)
memory usage: 5.0+ KB
```



```
#Attribute information null values
data.isna().sum()
```

```
Name            0
Links           0
Cost            0
Collections    54
Cuisines        0
Timings         1
dtype: int64
```

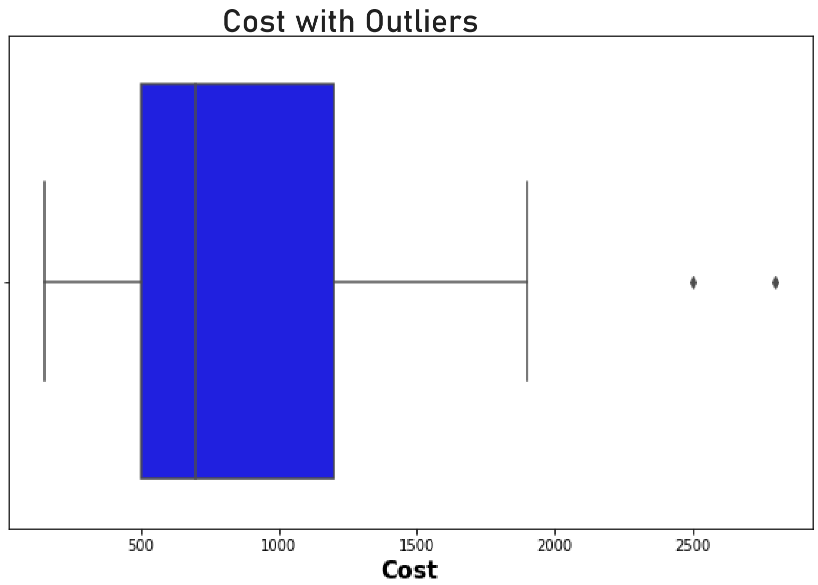
Descriptive Statistics

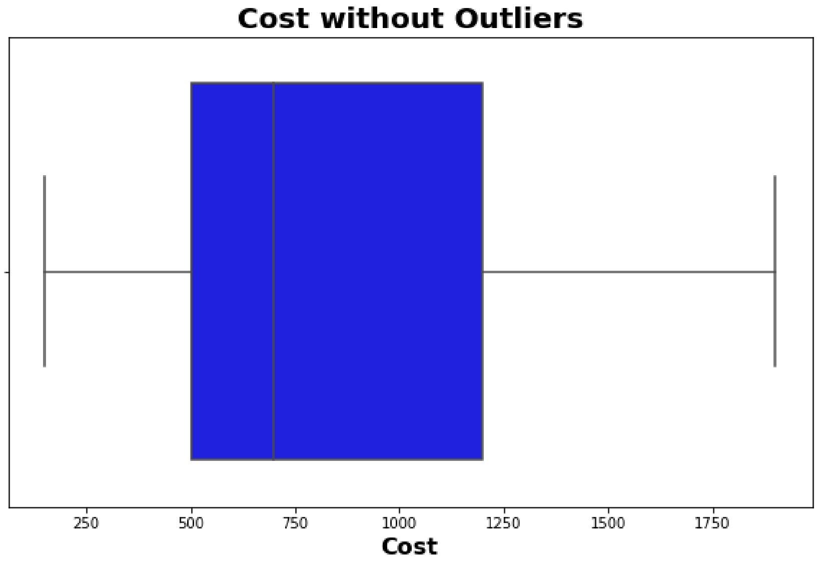


```
data.describe()
```

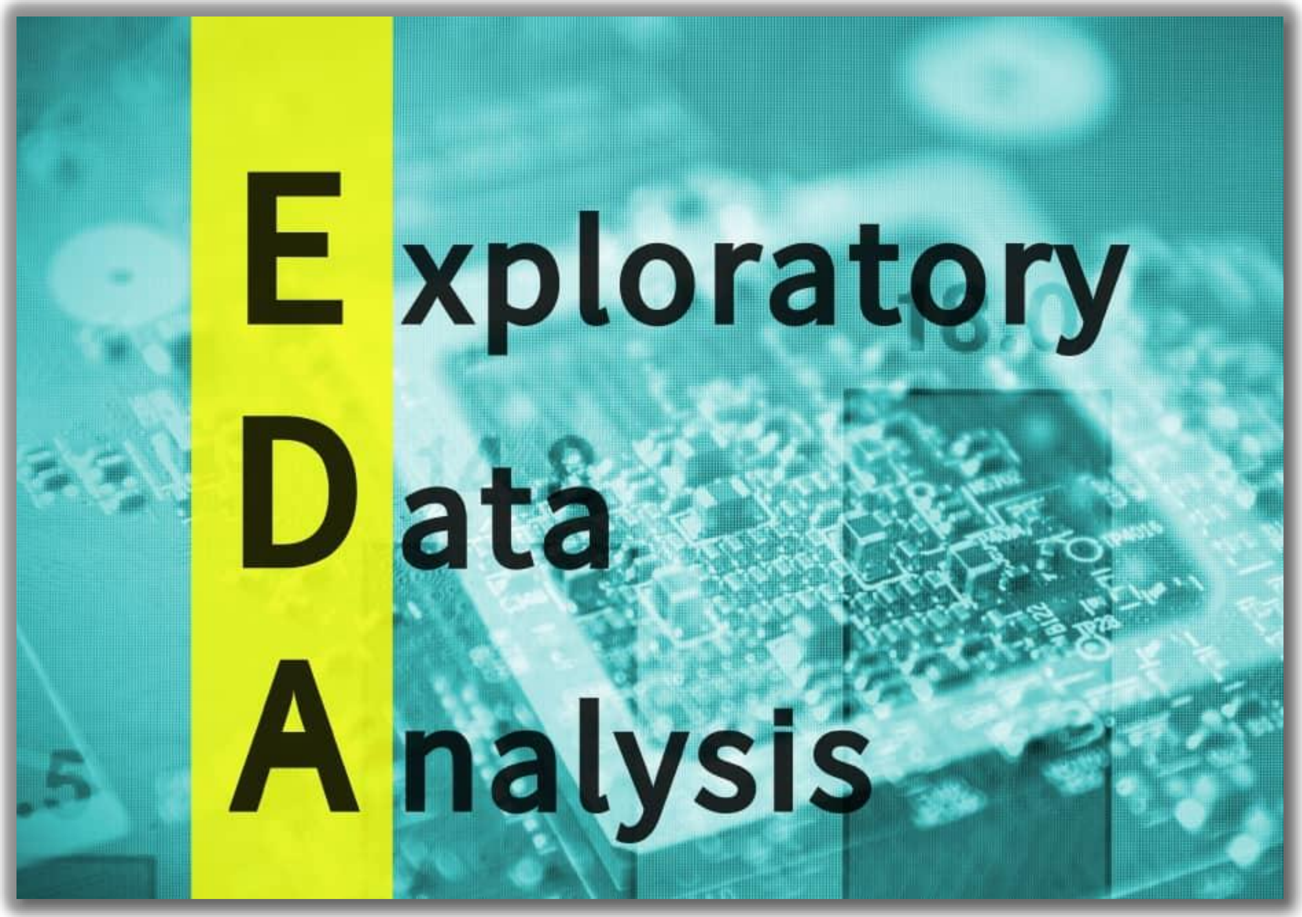
	Cost
count	105.000000
mean	861.428571
std	510.149730
min	150.000000
25%	500.000000
50%	700.000000
75%	1200.000000
max	2800.000000

Implementing Outliers

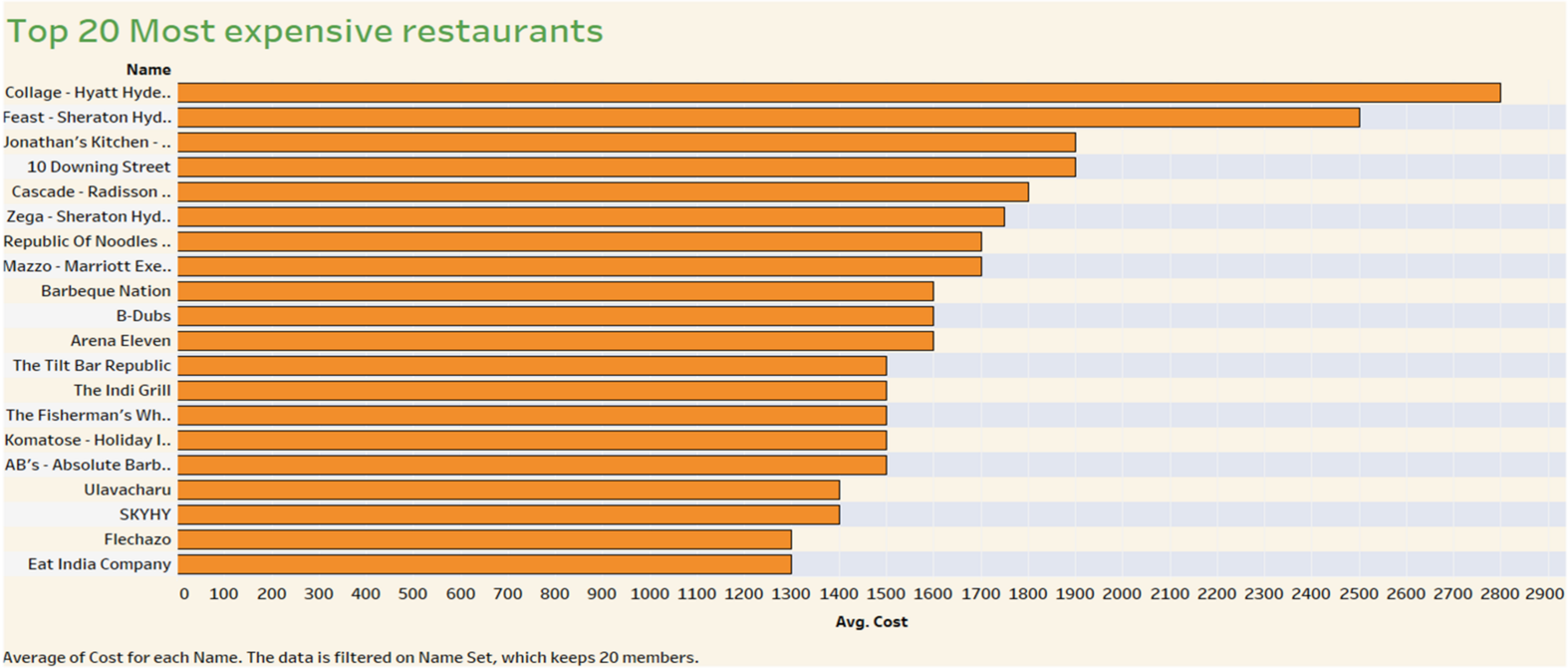




- In above boxplot-1, detects some outliers and in boxplot-2 outliers we removed the outliers.


Exploratory
Data
Analysis

EDA



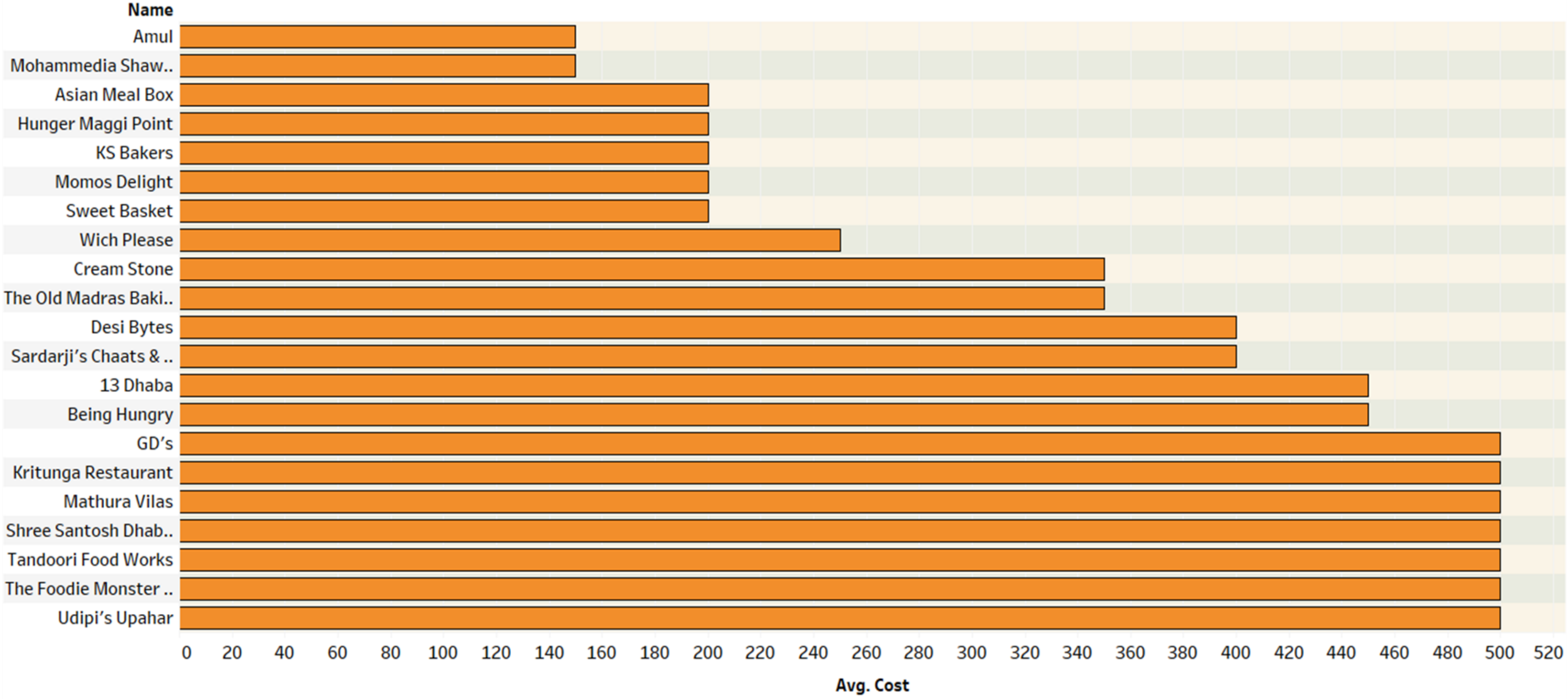
Top 20 Most expensive restaurants

Average of Cost for each Name. The data is filtered on Name Set, which keeps 20 members.

- The cost per person in restaurants ranges from 150 INR to 2800 INR. The cheapest restaurant is and Mohammedia Shawarma, while the most expensive is Collage - Hyatt Hyderabad Gachibowli.

EDA

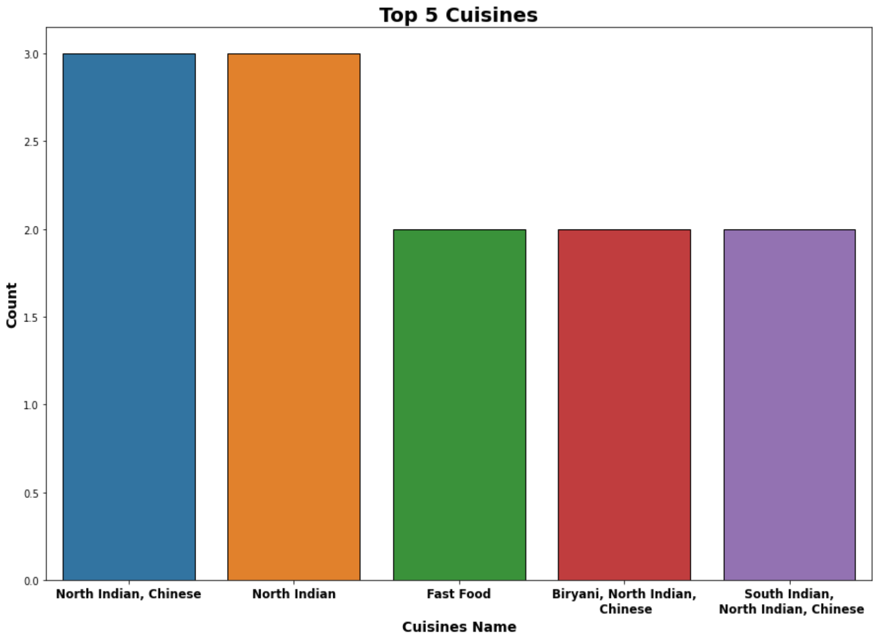
Top 20 Affortable restaurants

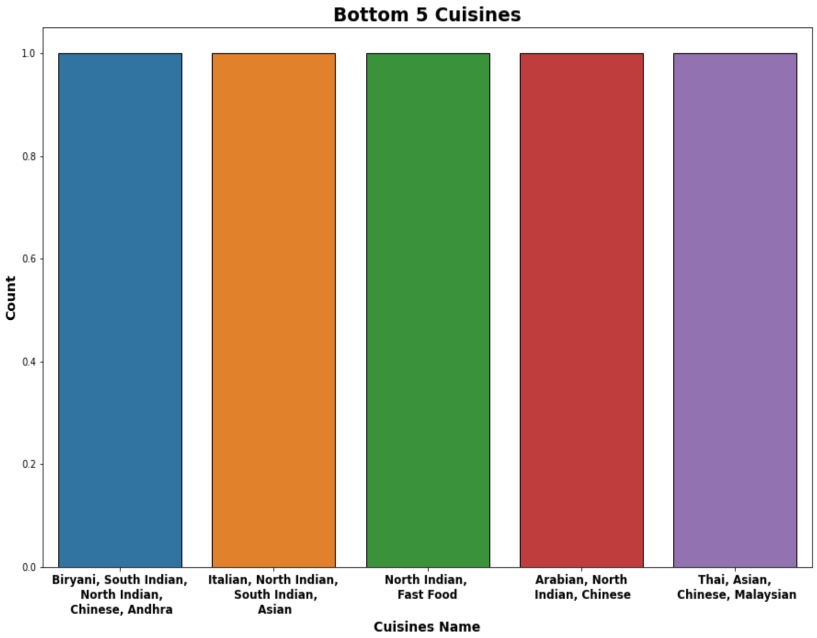


Average of Cost for each Name. The data is filtered on Timings Set, which keeps 20 members.

EDA

Cuisines


Top 5 Cuisines

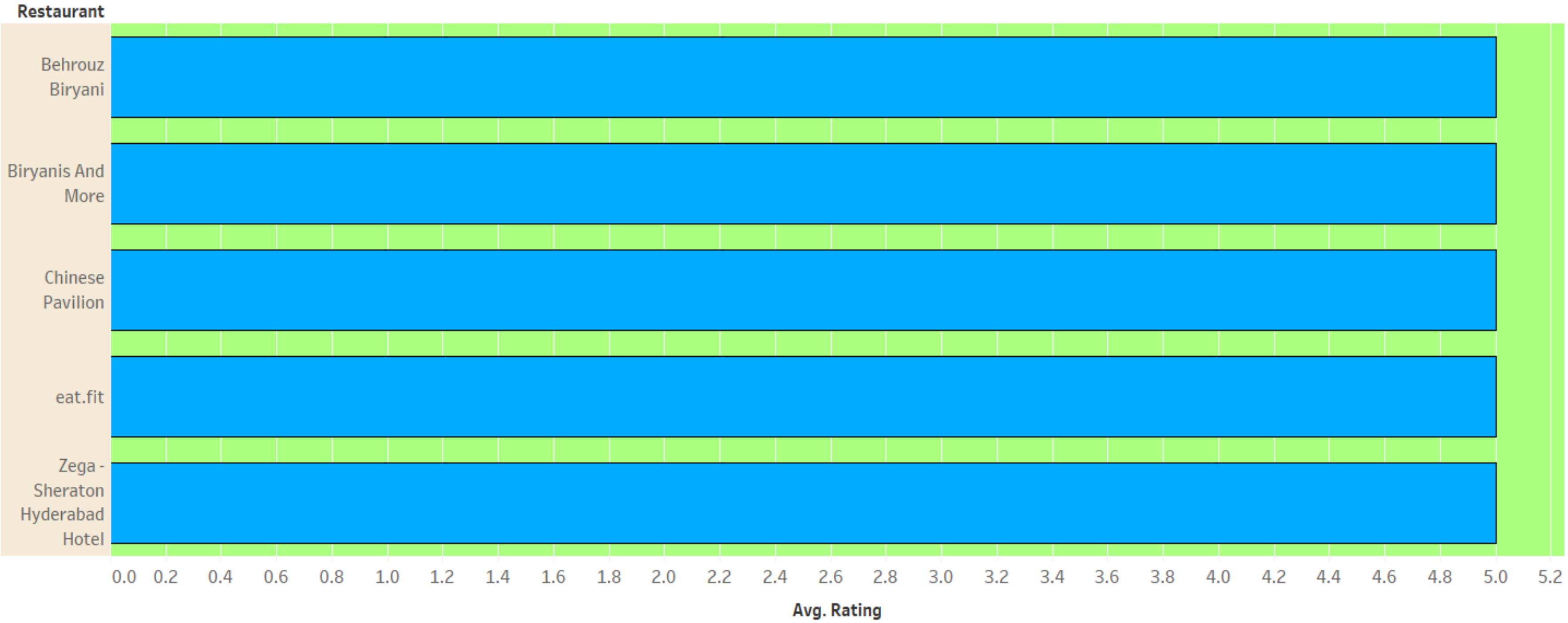

Bottom 5 Cuisines

- Indian cuisine consists of a variety of regional and traditional cuisines native to the Indian subcontinent. North Indian, cuisine is the most popular in restaurants, followed by fast food and Biryani.

EDA

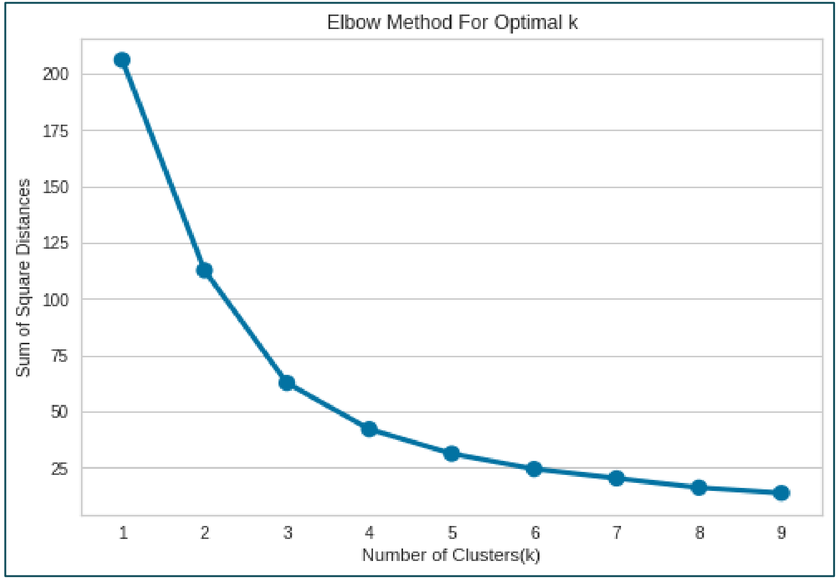
Best Restaurants

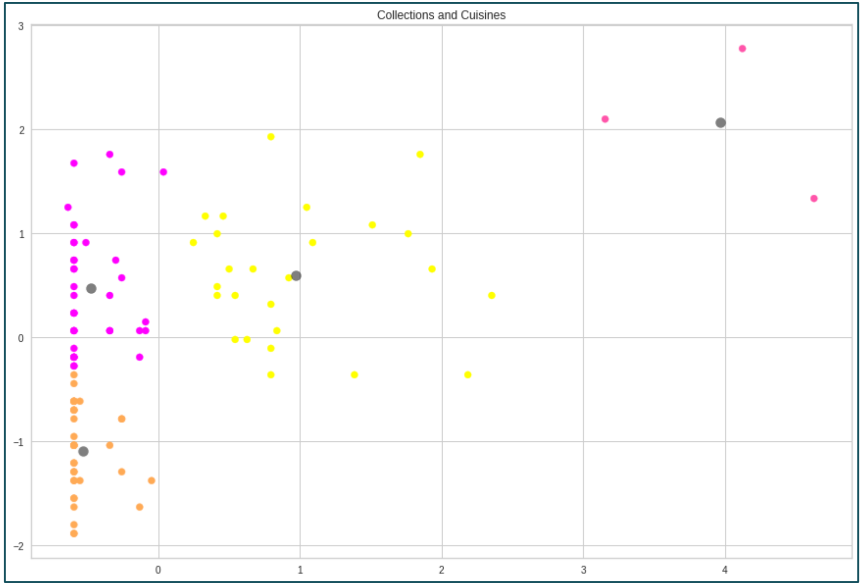
Sheet 1



Average of Rating for each Restaurant. The data is filtered on Reviewer Set, which keeps 5 members.

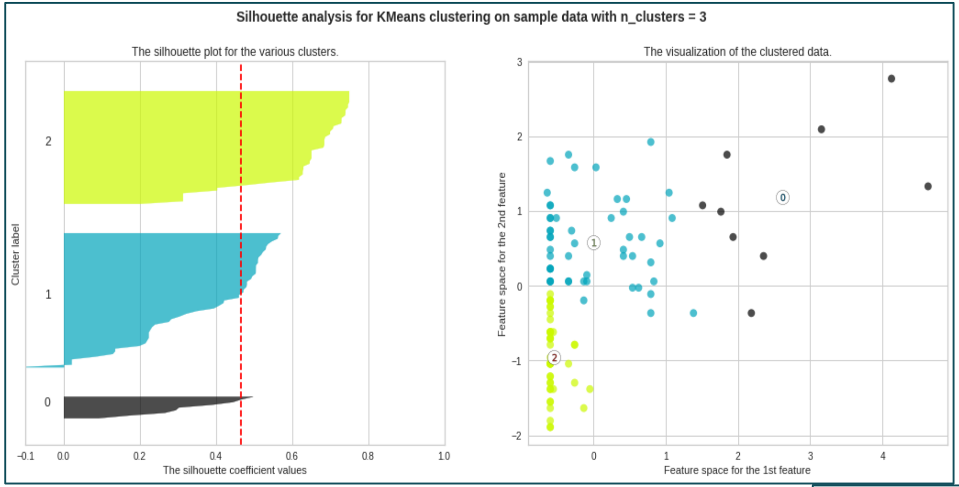
KMeans Clustering





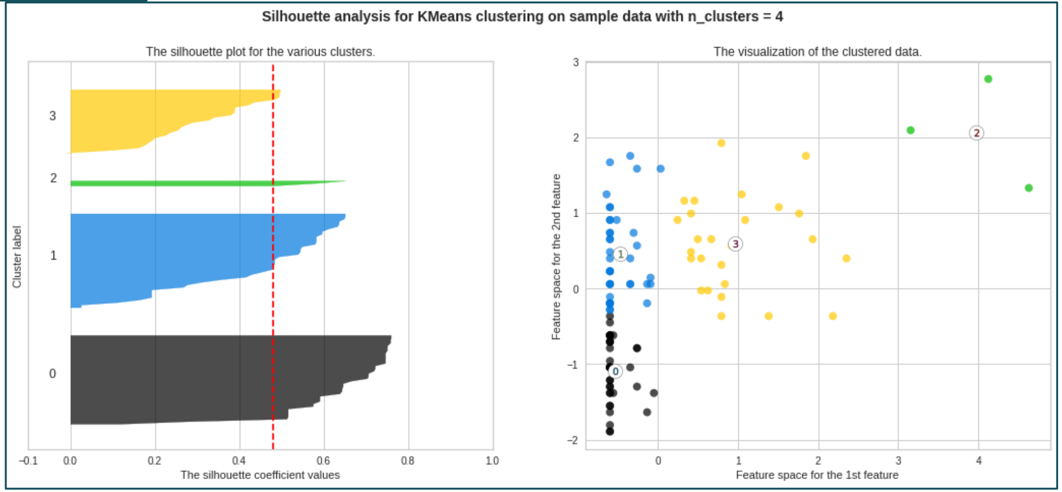
- The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found.
- We got best cluster as n_clusters=4 in KMeans.

Silhouette Score Method



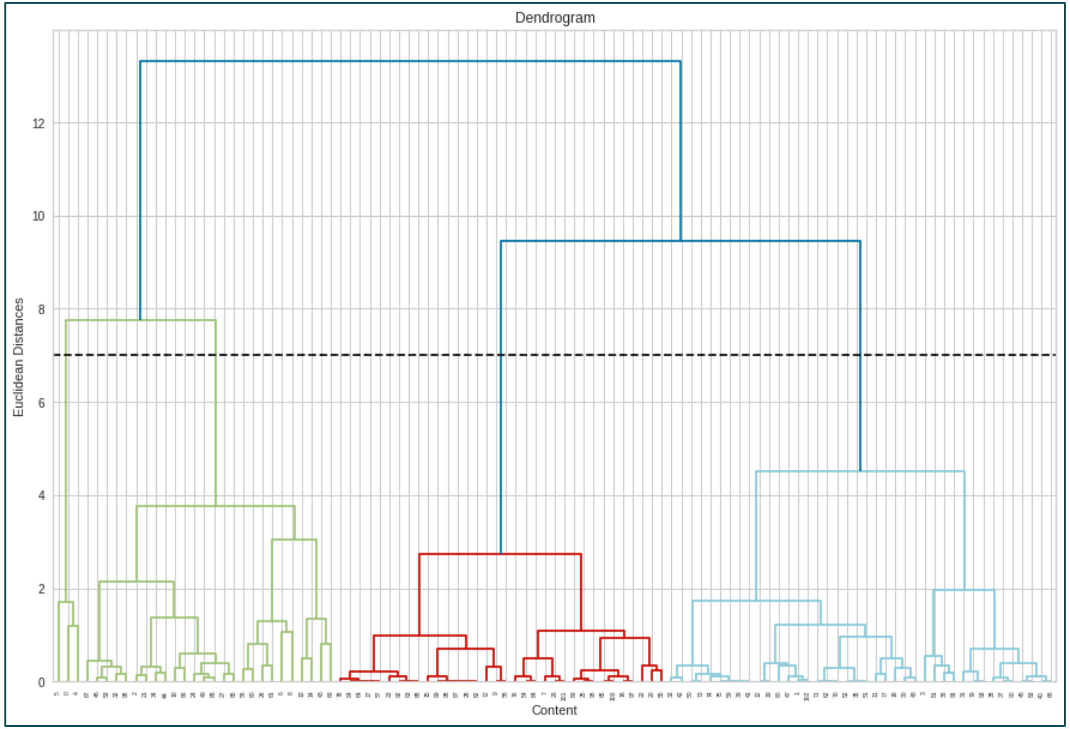
For **n_clusters = 3**
The average silhouette score is : 0.4655

silhouette score value ranges from **-1 to 1.**
1: Means clusters are well apart from each other and clearly distinguished.
0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
-1: Means clusters are assigned in the wrong way.



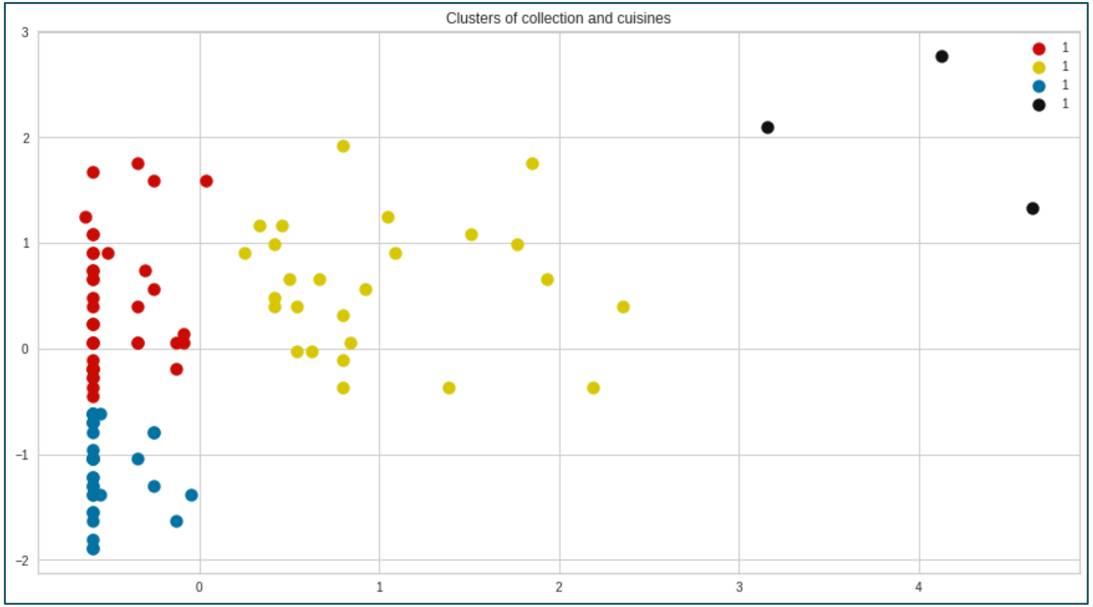
For **n_clusters = 4**
The average silhouette score is : 0.4799

Dendrogram to find optimal number of cluster



- Hierarchical clustering can be represented by a dendrogram. Cutting a dendrogram at a certain level gives a set of clusters.
- From abow dendrogram cutting at y=7 its gives **n_cluster=4**.

Agglomerative hierarchical Clustering



- Agglomerative Clustering is a bottom-up strategy in which each data point is originally a cluster of its own, and as one travels up the hierarchy, more pairs of clusters are combined.

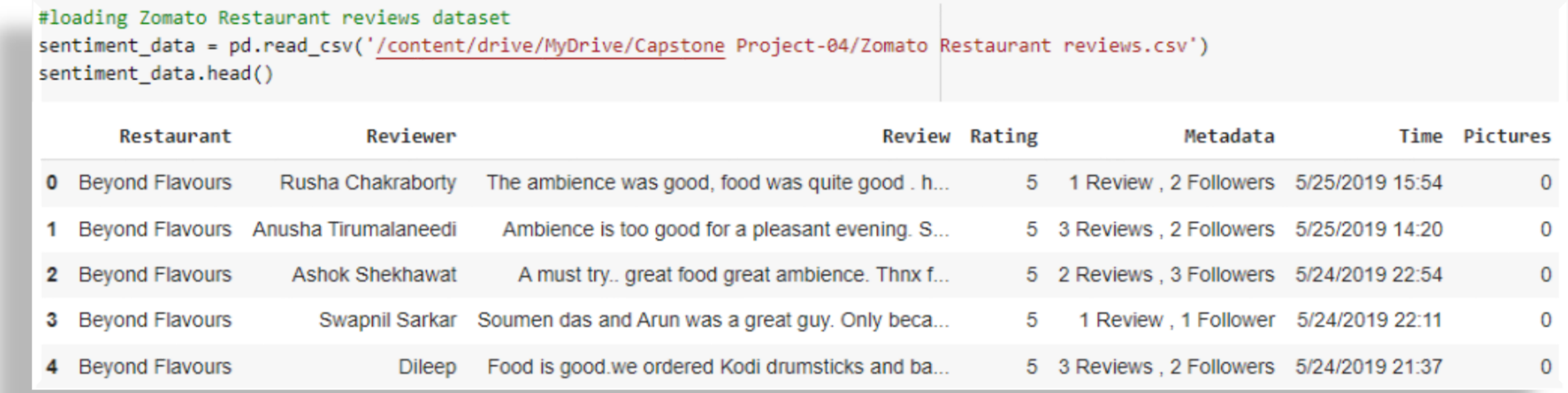
Sentiment Analysis

Sentiment Analysis

Sentiment Analysis is the most common text classification tool that analyses an incoming message and tells whether the underlying sentiment is positive, negative our neutral. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations.

Loading the Dataset

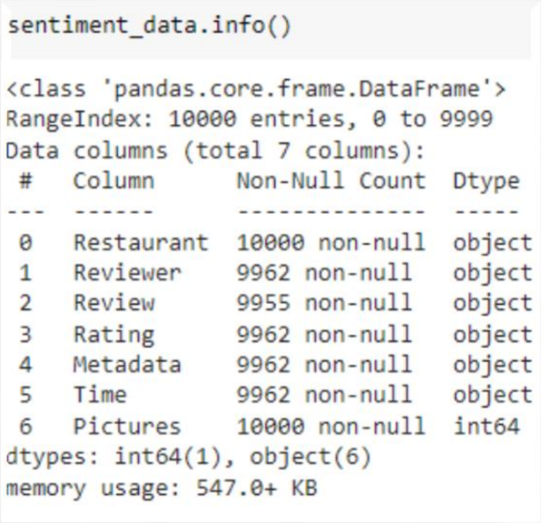
Zomato Restaurant reviews dataset



```
#loading Zomato Restaurant reviews dataset
sentiment_data = pd.read_csv('/content/drive/MyDrive/Capstone Project-04/Zomato Restaurant reviews.csv')
sentiment_data.head()
```

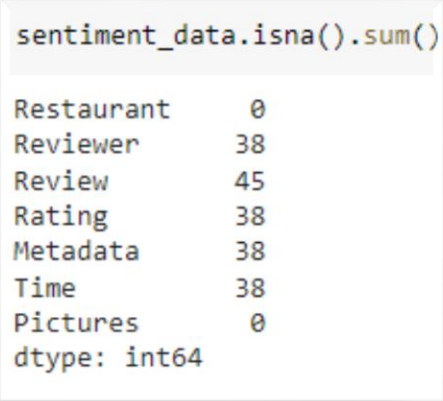
```
        Restaurant             Reviewer                                          Review  Rating               Metadata             Time  Pictures
0  Beyond Flavours      Rusha Chakraborty  The ambience was good, food was quite good . h...       5    1 Review , 2 Followers  5/25/2019 15:54         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
```

Attribute Information:- Dtypes and Null values



```
sentiment_data.info()
```

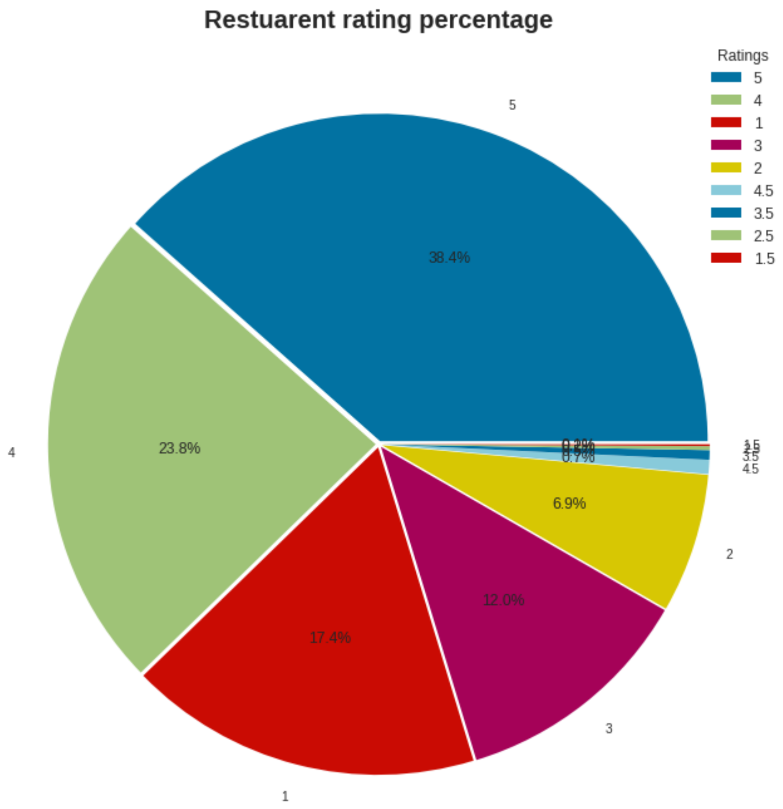
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Restaurant  10000 non-null  object
 1   Reviewer    9962 non-null   object
 2   Review      9955 non-null   object
 3   Rating      9962 non-null   object
 4   Metadata    9962 non-null   object
 5   Time        9962 non-null   object
 6   Pictures    10000 non-null  int64 
dtypes: int64(1), object(6)
memory usage: 547.0+ KB
```



```
sentiment_data.isna().sum()
```

```
Restaurant     0
Reviewer      38
Review        45
Rating        38
Metadata      38
Time          38
Pictures       0
dtype: int64
```

Restaurant Rating Percentage



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

- The customers with a good number of followers who have given more reviews with constantly low ratings to understand the fields that need to be worked on.

Text Preprocessing

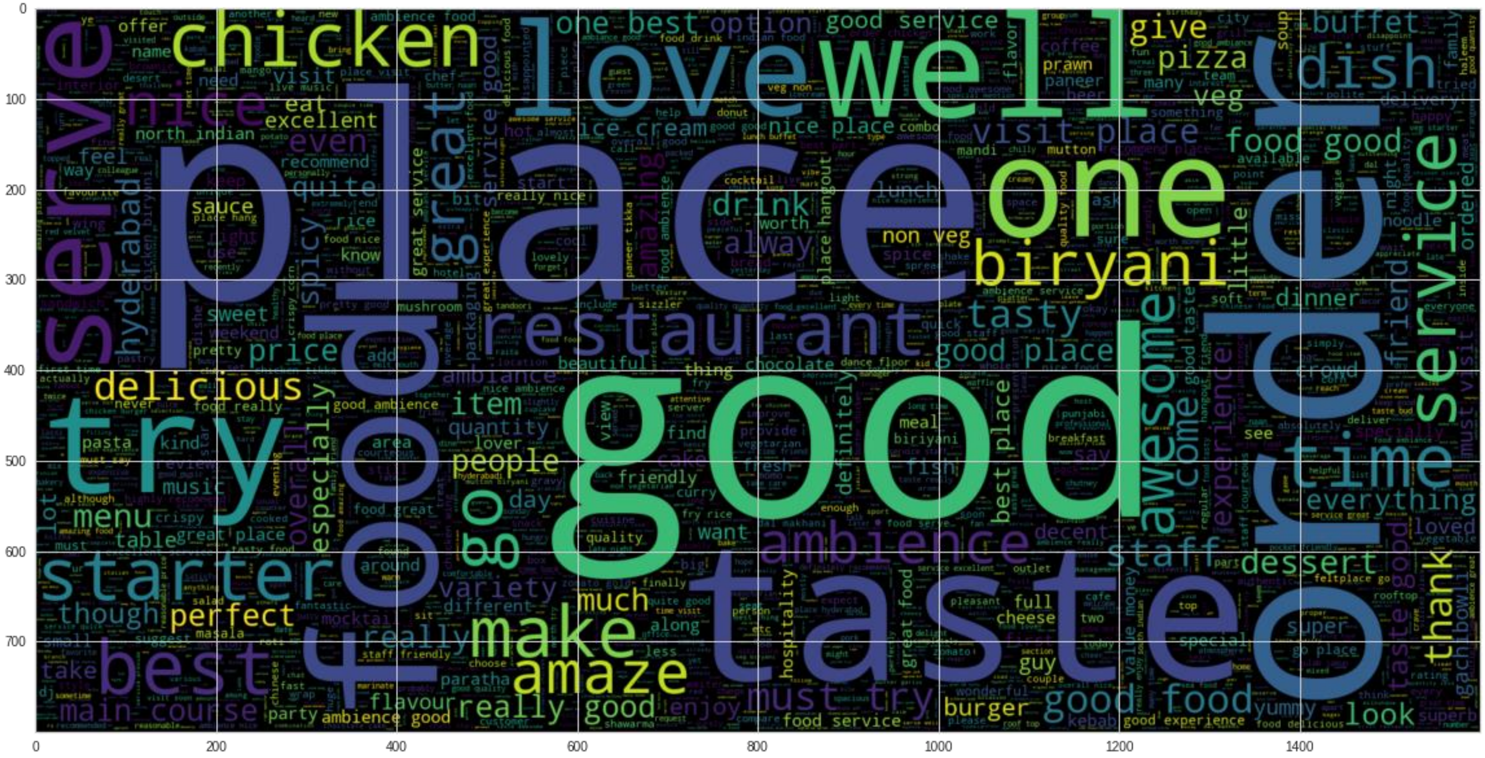
- **Stemming**

Stemming is a process to reduce the word to its root stem for example run, running, runs, runed derived from the same word as run. basically stemming do is remove the prefix or suffix from word like ing, s, es, etc. NLTK library is used to stem the words. The stemming technique is not used for production purposes because it is not so efficient technique and most of the time it stems the unwanted words.

- **Lemmatization**

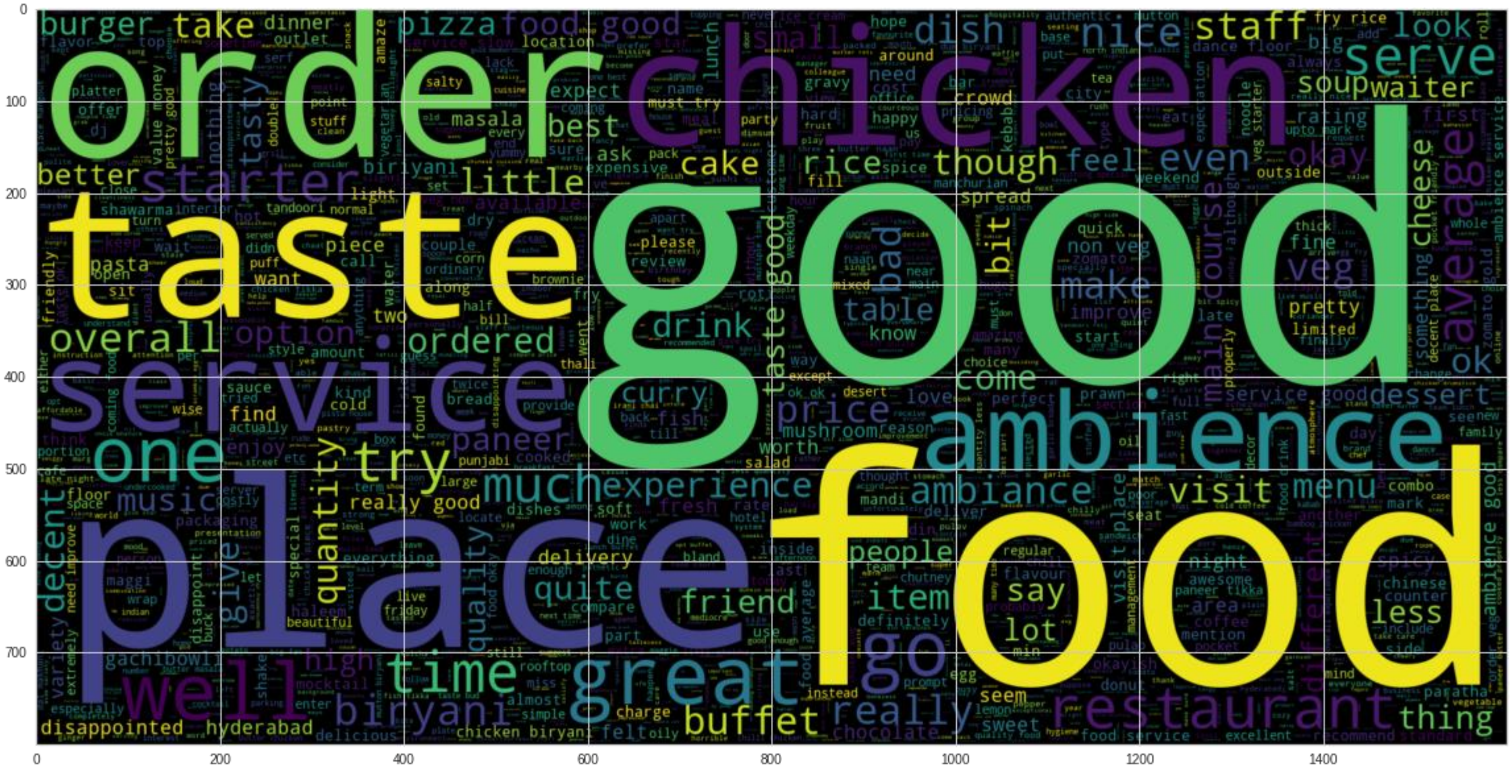
Lemmatization is similar to stemming, used to stem the words into root word but differs in working. Actually, Lemmatization is a systematic way to reduce the words into their lemma by matching them with a language dictionary.

Frequent Keywords Used for good reviews

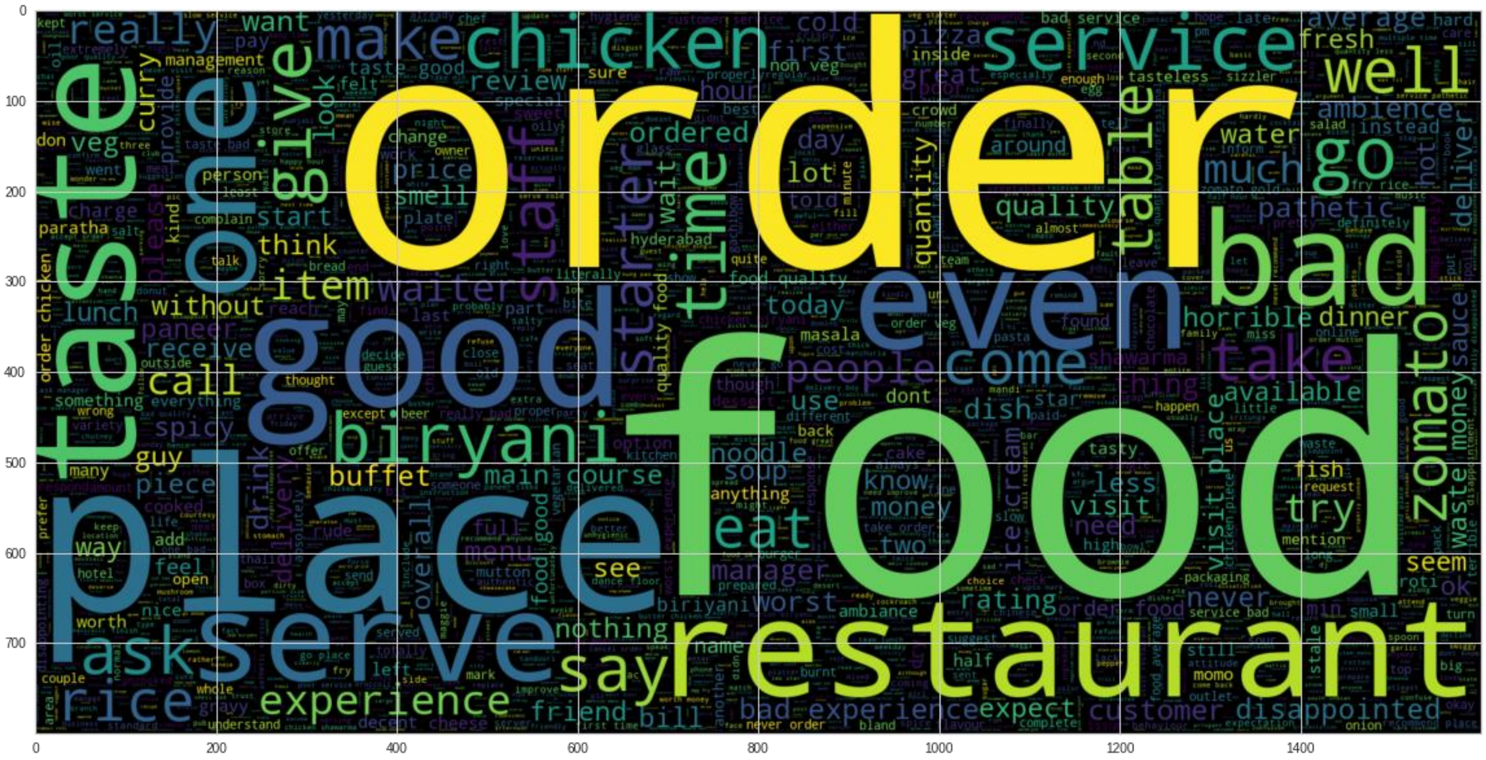


Word Clouds are visual displays of text data – simple text analysis. Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language (“a”, “an”, “the” etc.)

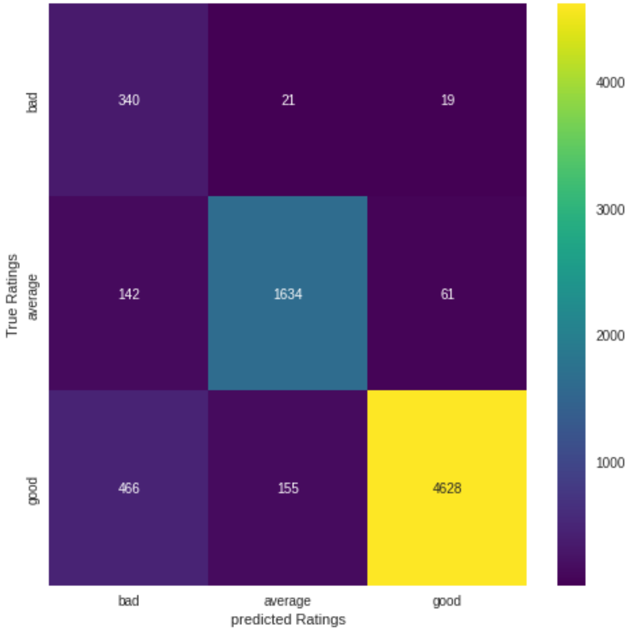
Frequent Keywords Used for Average reviews

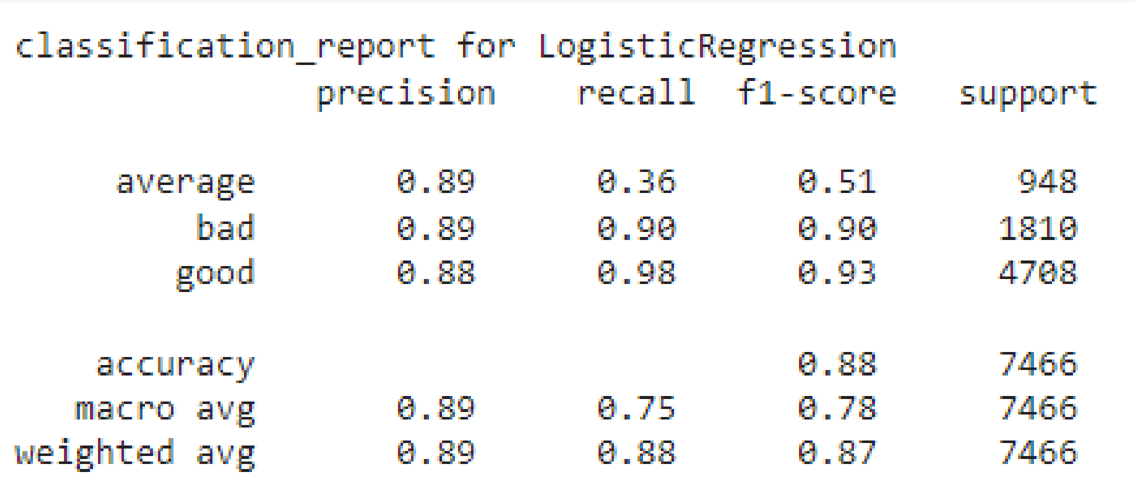


Frequent Keywords Used for bad reviews



Logistic Regression



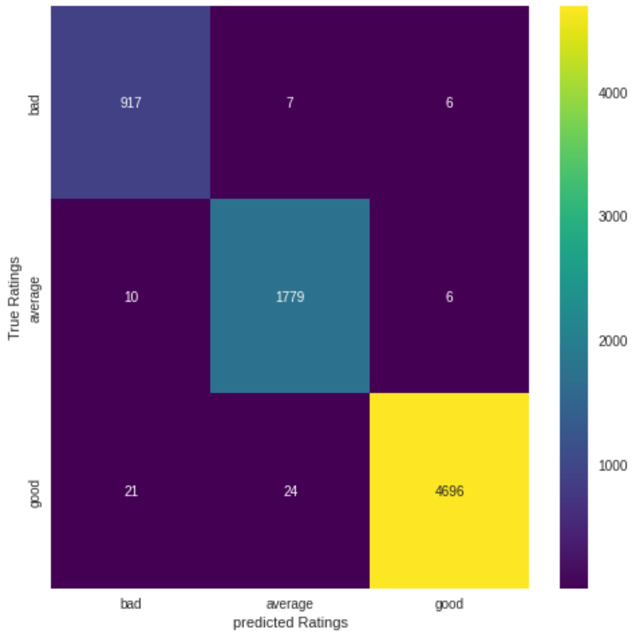


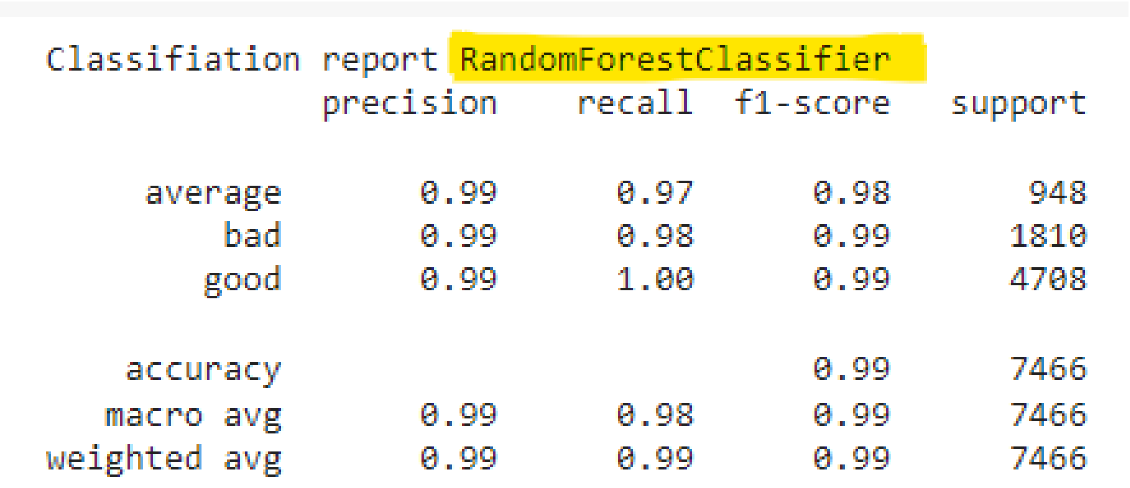
```
classification_report for LogisticRegression
              precision    recall  f1-score   support

     average       0.89      0.36      0.51       948
         bad       0.89      0.90      0.90      1810
        good       0.88      0.98      0.93      4708

    accuracy                           0.88      7466
   macro avg       0.89      0.75      0.78      7466
weighted avg       0.89      0.88      0.87      7466
```

Random Forest Regression





```
Classifiation report RandomForestClassifier
              precision    recall  f1-score   support

     average       0.99      0.97      0.98       948
         bad       0.99      0.98      0.99      1810
        good       0.99      1.00      0.99      4708

    accuracy                           0.99      7466
   macro avg       0.99      0.98      0.99      7466
weighted avg       0.99      0.99      0.99      7466
```

Conclusion

Clustering Analysis

- silhouette score at **n_clusters = 4**, we get highest silhouette score is 0.47229.
- From elbow method we get 4 number of cluster is best among all.
- Applied agglomerative hierarchical clustering from this we find 4 number of cluster good fit our model.
- By applying different clustering algorithm to our dataset. we get the optimal number of cluster is equal to 4.

Sentiment Analysis

- Categorize rating in 3 types i.e. good, bad and average. 4500+ good, 1700+ bad and 900+ average ratings given by customer.
- By using logistic regression and random forest regression model on reviews dataset, we get 89% accuracy at logistic regression & 99% accuracy at random forest regression.

References

1. https://www.kaggle.com/
2. https://www.analyticsvidhya.com/
3. https://www.geeksforgeeks.org/
4. https://learn.almabetter.com/