Capstone Project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

1. Vilas Sonawane
2. Nandeesh Umesha
3. Bhavika Gaurkar
4. Soumya Ranjan Dash

Point for Discussion

- Introduction
- Purpose of The Project
- Data Summary
- Data Cleaning
    - Data cleaning of restaurants data
    - Data cleaning of reviews data
- Univariate analysis
    - Univariate analysis of Restaurants data
    - Univariate analysis of reviews data
- Sentiment Analysis of user reviews
- Multivariate Analysis
- Clustering of Restaurants
- Clustering Based on the Location of Restaurants
    - K Means Clustering Algorithm & Hierarchical Clustering
- Clustering Based on the Rating & Dining cost at Restaurants
    - K Means Clustering Algorithm & Hierarchical Clustering
- Cost to Benefit Analysis of restaurants (Clusters):
- Recommendation of restaurants for customers
- Conclusion

Introduction

Zomato is an Indian restaurant aggregator and food delivery start-up founded by **Deepinder Goyal** and **Pankaj Chaddah in 2008**.

- Zomato connects customers, restaurant partners and delivery partners, to serve their needs.
- Customers use Zomato platform to search and discover restaurants, read and write customer generated reviews and upload photos, order food delivery, book a table and make payments while dining-out at restaurants.

- Along with this its also provide restaurant partners with industry-specific marketing tools which enable them to engage and acquire customers to grow their business while also providing a reliable and efficient last mile delivery service.

- Restaurant business in India is always evolving. 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 inspired by the same idea, In this project we have analyzed the Zomato restaurants data of Hyderabad to solve some business cases for the company to grow up and work on the fields they are currently lagging in.

Purpose of The Project

This project mainly focuses on three broad objectives:

1. Analyze the sentiments of the reviews given by the customers in the data and make some useful conclusions.

2. Cluster the Zomato restaurants into different segments and use the clustering to solve some business cases for the company to grow up and work on the fields they are currently lagging in.

3. Design Recommendation system which can directly help the customers to find the best restaurants in their locality.

Data Summary

The dataset included two csv files:

- **Zomato Restaurant names and Metadata.csv:**
  - **Rows:** 105 instances. Each row has info corresponding to Restaurants
  - **Columns:** 6 columns

- **Zomato Restaurant Reviews.csv:**
  - **Rows:** 10000 instances. Each row has info corresponding to customer review regarding to restaurant.
  - **Columns:** 7 columns

Data Cleaning
I. Data cleaning of restaurants data

- **Web scraping for extracting additional data from the web links of restaurants. (Addition of New Features)**
  1. Latitude 2. Longitude 3. additional services 4. Known for
- **Cleaning of existing features like**
  'Collections', 'Cuisines', additional services etc.
- **Cost:** String format converted to integer
- **Timings :** extracted days column (How many days restaurant is open in week?)
- **Handling missing value / null value :** Only one value in longitude & latitude was missing so dropped it.

- drop them.

II. Data cleaning of reviews data

- **Handling Missing Values :** (Drop the rows as no is very Small)
  No of missing values : Reviewer : 38 ,Review : 45 , Rating :38 ,Metadata :38 ,Time :38

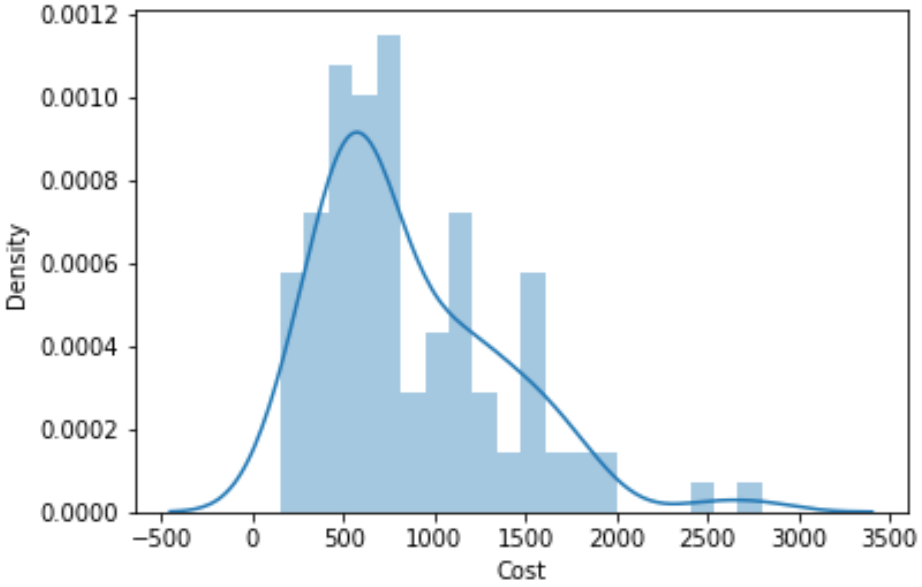
- **Rating :** One value ratings was given as 'Like', Just one occurrence, so replace it by a score of 5.

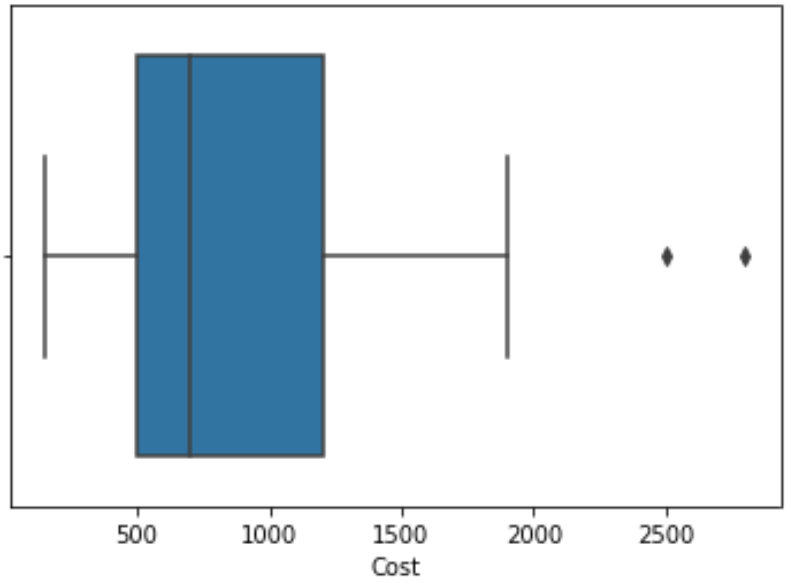
- **Metadata :** Split the metadata column into No. of reviews and No. of followers.

- **Time :** String format converted to date time format & extracted Date, month, year from date time.

Univariate analysis Restaurant Data

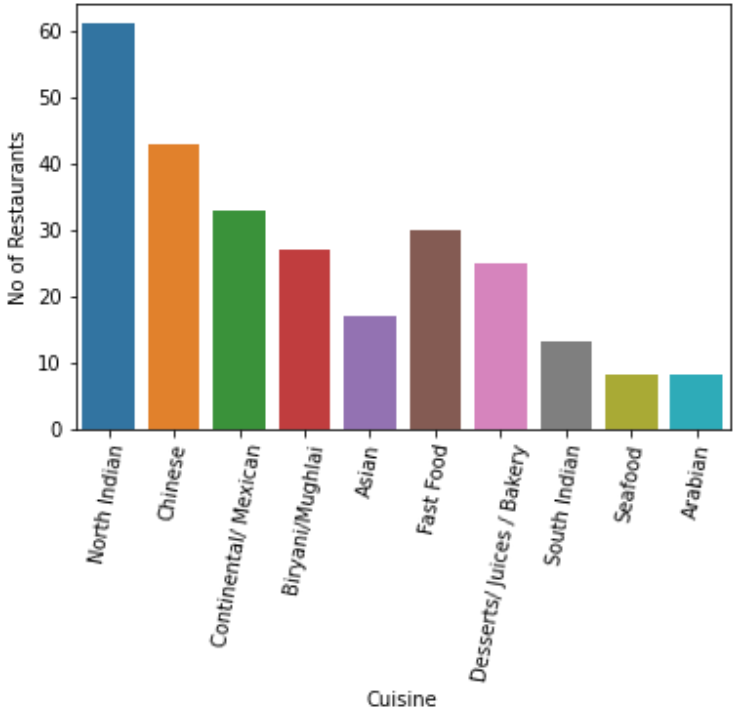
I. Cost





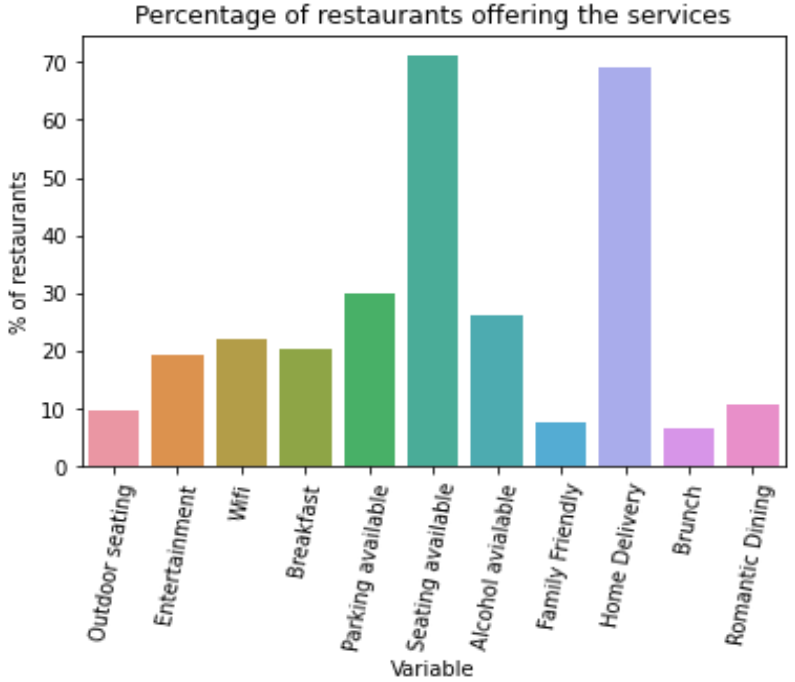
Dining cost of the restaurants are mostly in the range of 150 to 2800 rupees per person.

2. Cuisines

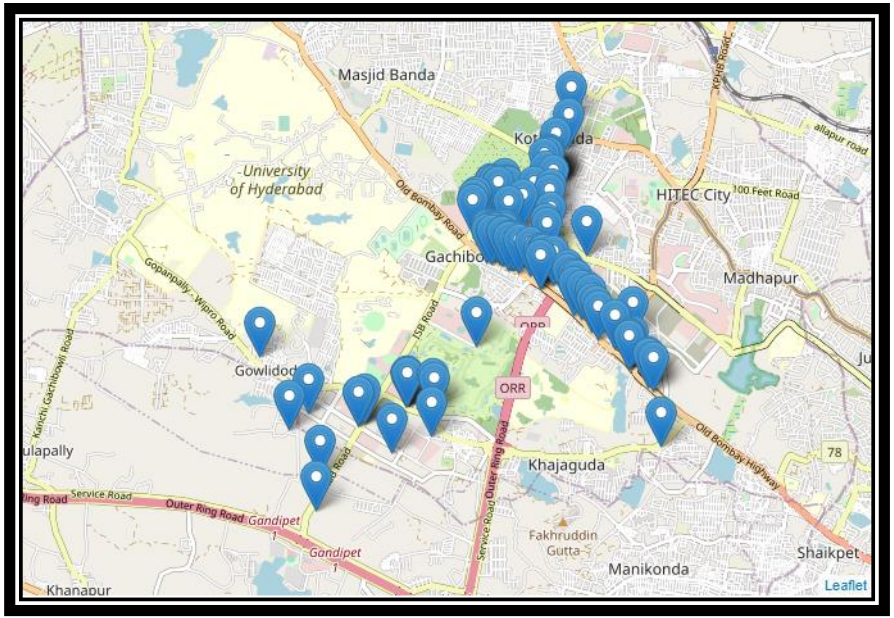


North Indian is the predominant cuisine in Gachibowli Hyderabad

3. Services offered



4. Latitude & Longitude (Location of Restaurants)



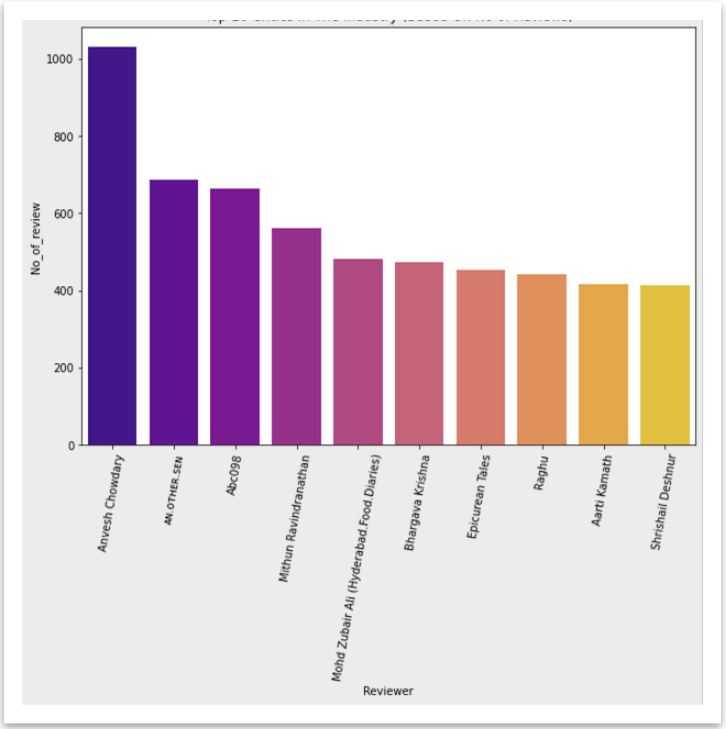
Majority of the restaurants seem to be located on the old - Bombay highway. It is interesting to note that three clear clusters are visible clearly:

1. Along the old Bombay highway
2. Below the highway(towards ISB)
3. Above the highway(towards Botanical gardens)

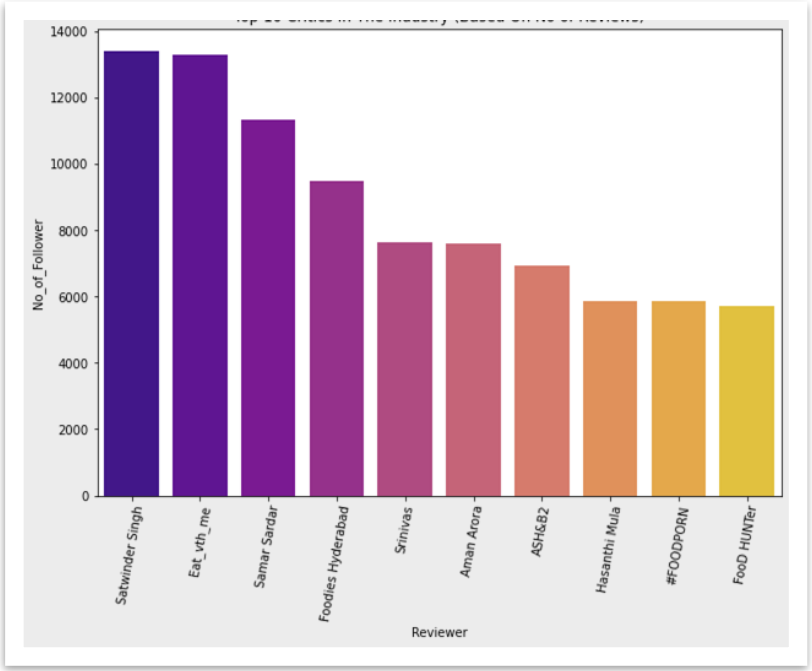
Univariate analysis of reviews data

I .Top 10 Critics in The Market

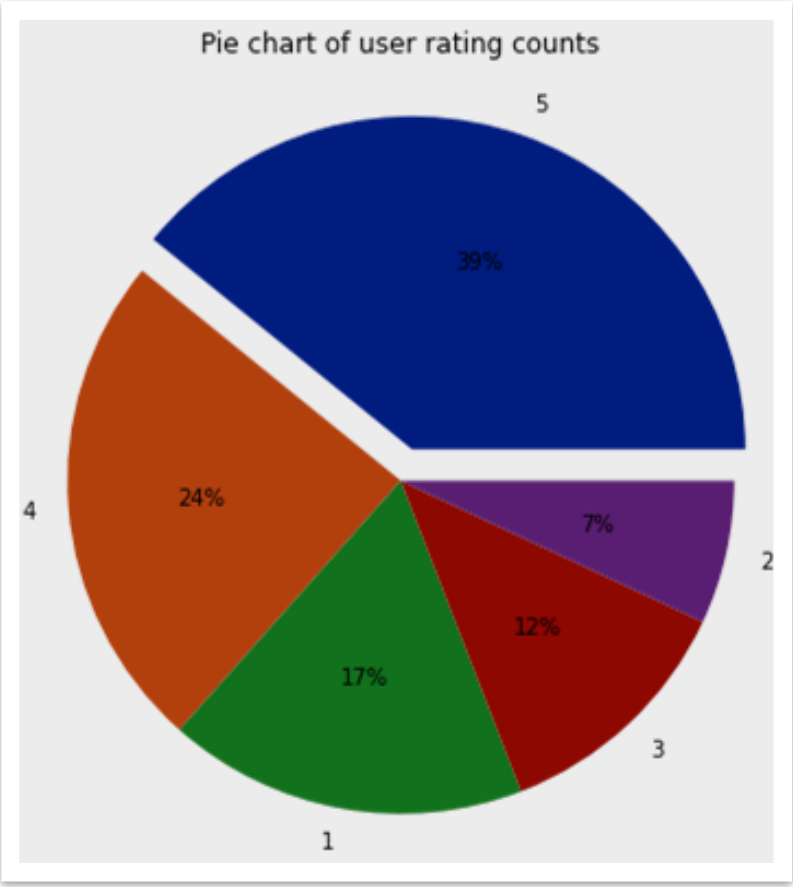
(Based on No of Reviews Given)



(Based on No of Followers)



Rating



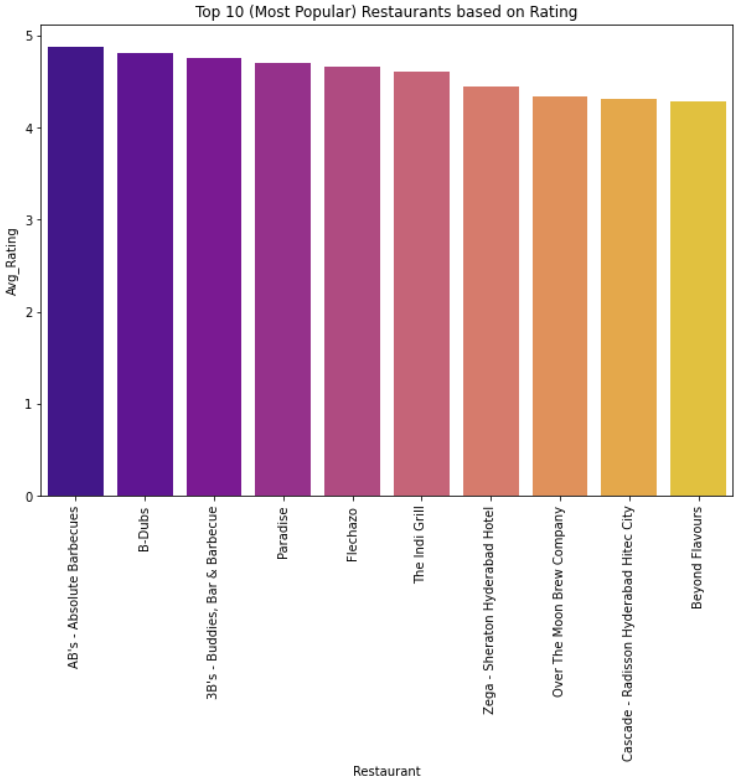
So based on the above observations:

- ❑ **63% of the restaurant has got high rating (Above 4)**
- ❑ **12% Restaurant have got average rating (Around 3)**
- ❑ **24 % Restaurant have got Low rating (Below 2)**

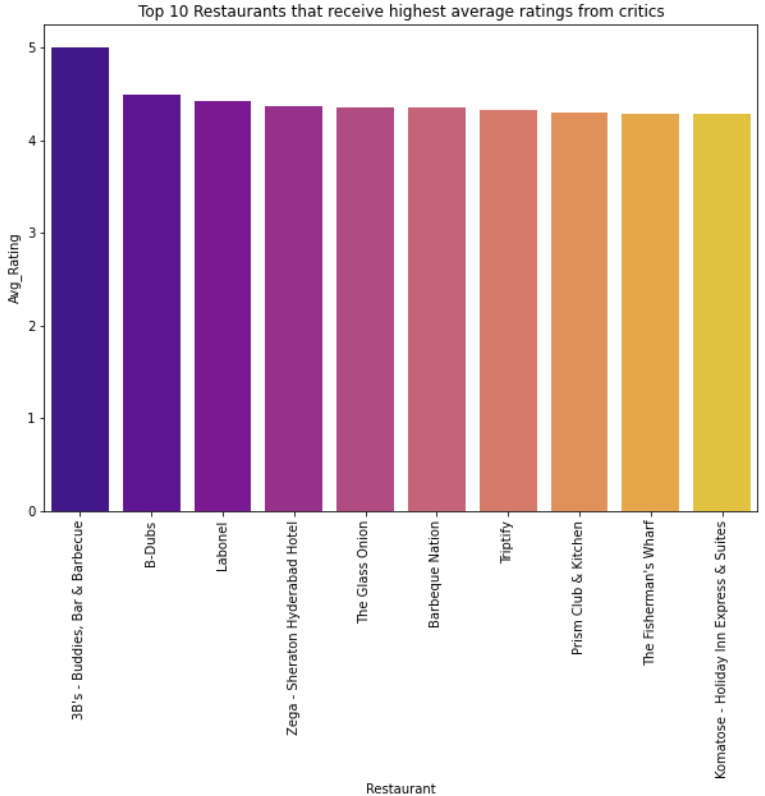
Multivariate Analysis

1. Top 10 (Most Popular) Restaurants

1. Based on Rating

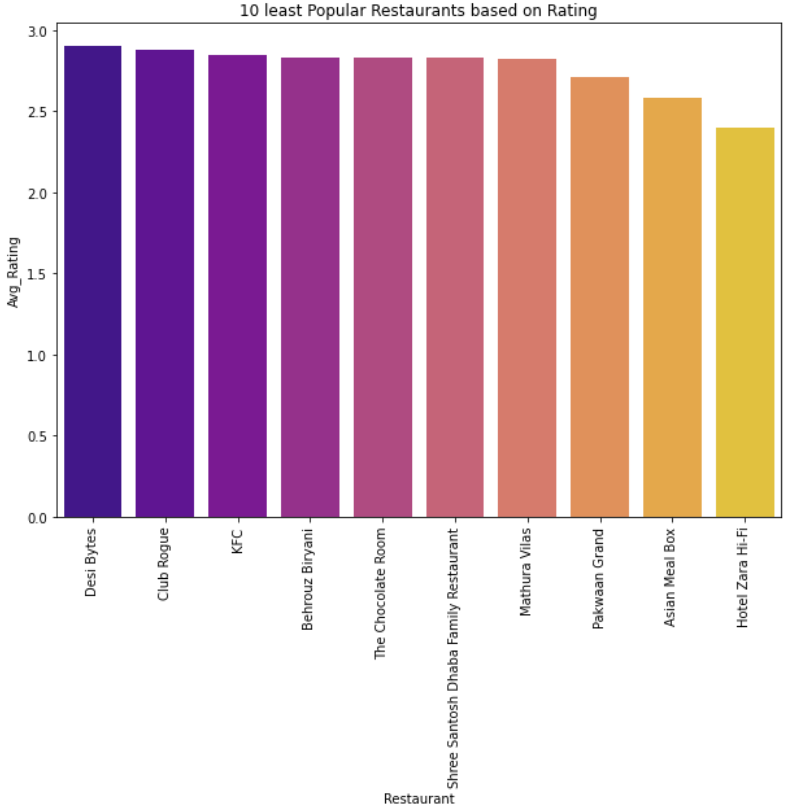


2 . Based on average Rating of Critics

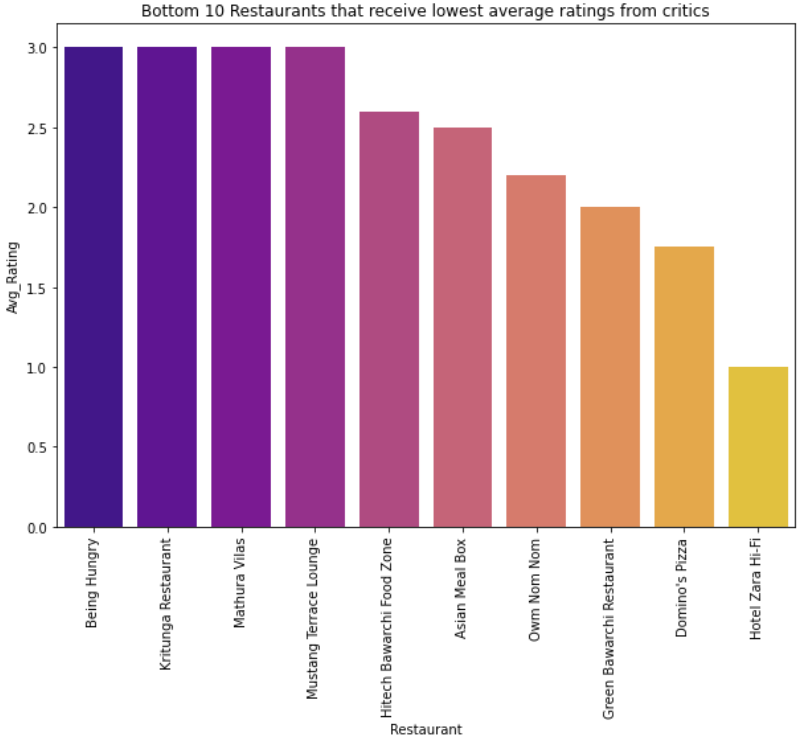


2. Bottom 10 (Least Popular) Restaurants

1. Based on Rating

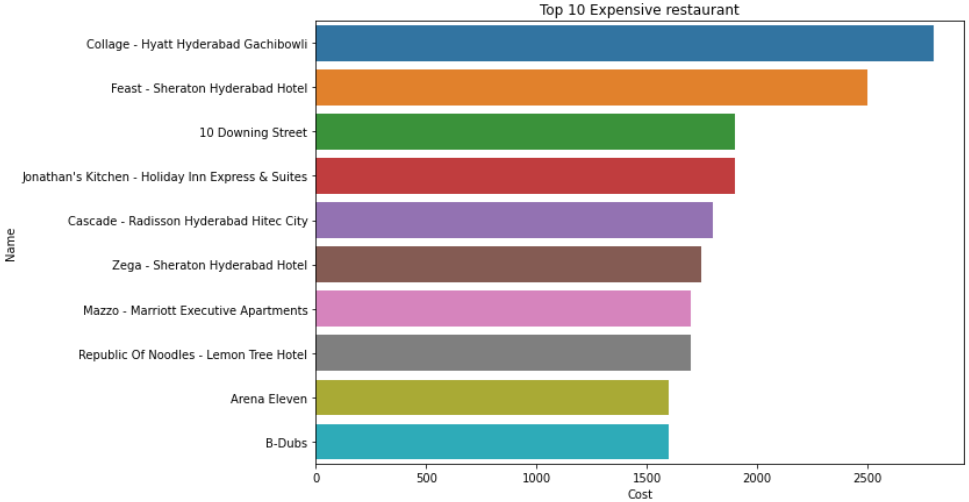


2 . Based on average Rating of Critics

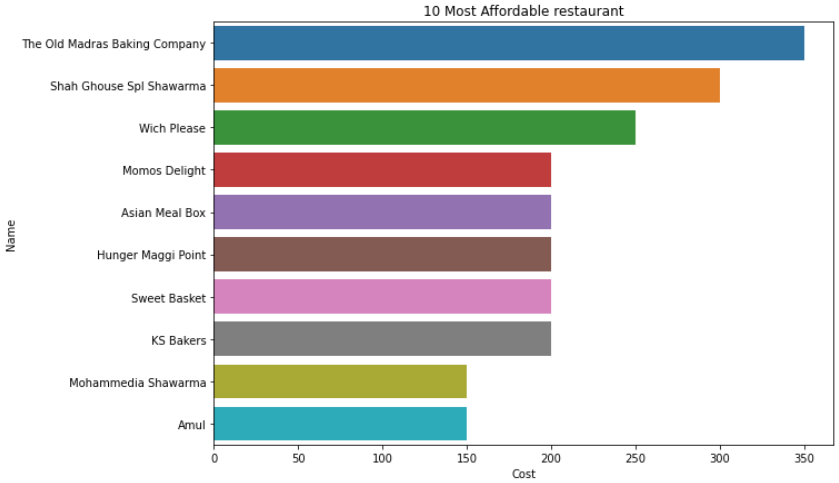


3. Most Expensive / Most Affordable Restaurant

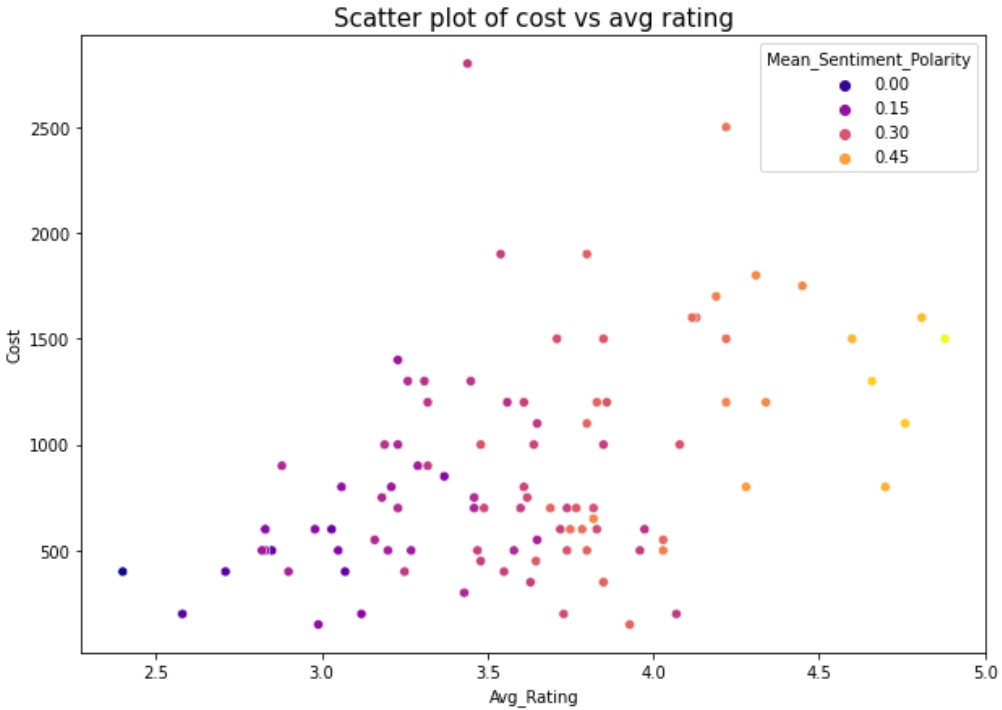
1. Top 10 Expensive Restaurant



2 . 10 Most Affordable Restaurant



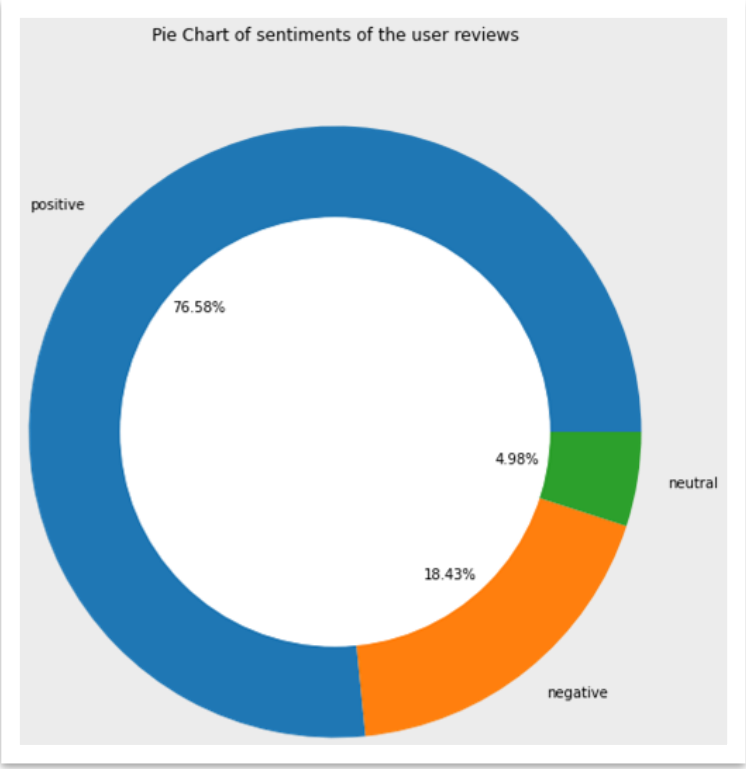
4. Dining Cost per Person vs Average Rating



There is a slightly positive correlation between average rating and cost per person of a restaurant.

It means costly restaurants are performing better than cheaper restaurants.

Sentiment Analysis of user reviews

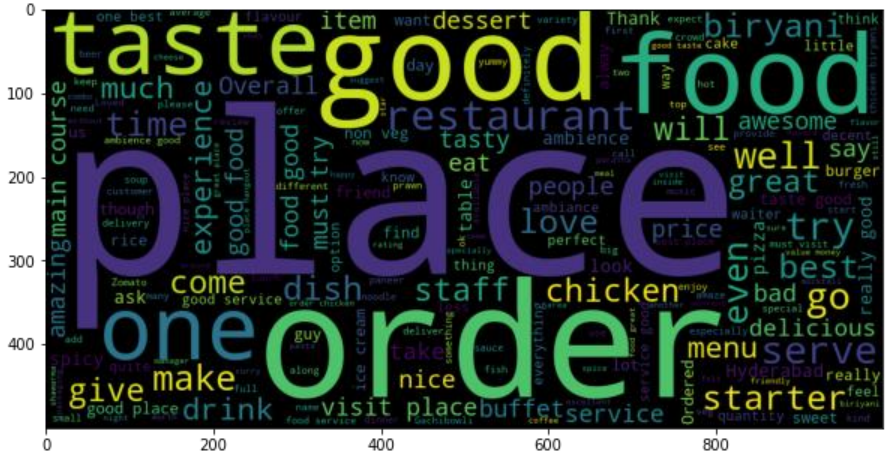


So based on sentiment analysis of the review given by customer

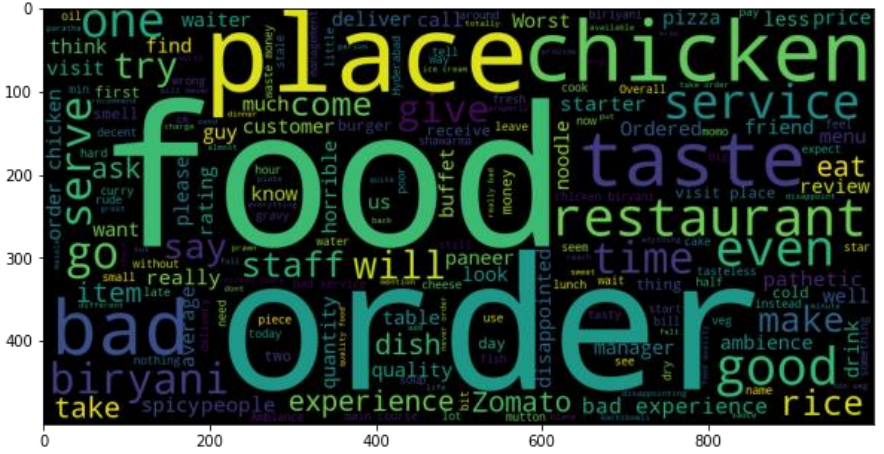
1. Overall 76 % Reviewer have given Positive reviews.
2. 18 % Reviewer have given Negative reviews.
3. 4.99 % Reviewer have given Neutral reviews.

Words used by customer while expressing their Sentiments

1. Words Used for Positive Reviews



2. Words Used for Negative Reviews



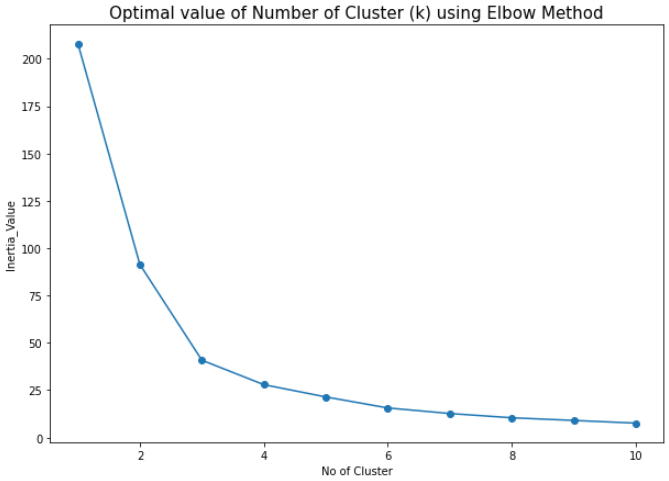
From the above cloud we can suggest that the following areas need to be improved

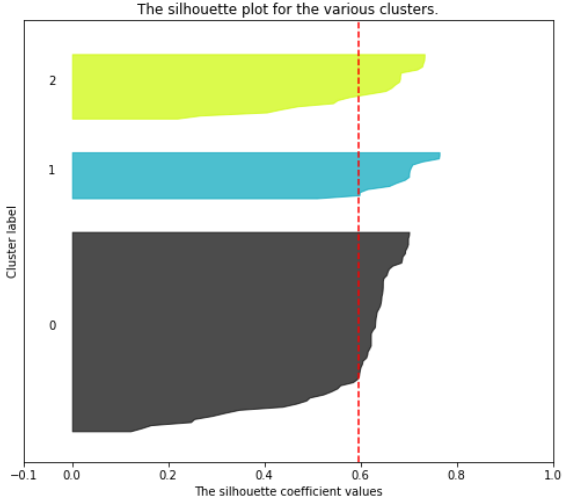
- Taste of the food at restaurants
- Improve staff services at dining restaurants
- Chicken items appear predominantly in negative reviews, quality and taste of these items need to improve.

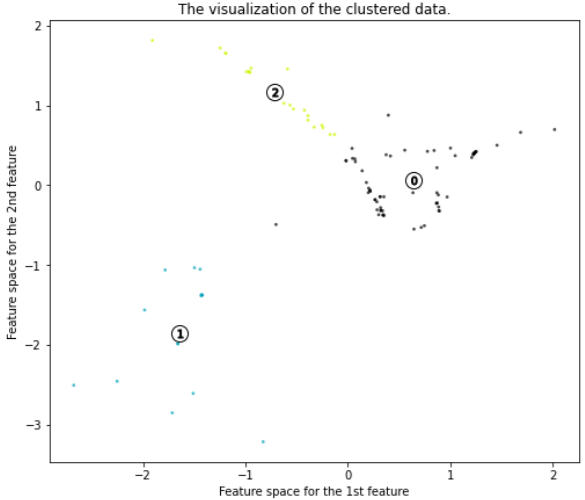
Clustering of Restaurants

I. Clustering Based on the Location of Restaurants

1. K Means Clustering Algorithm

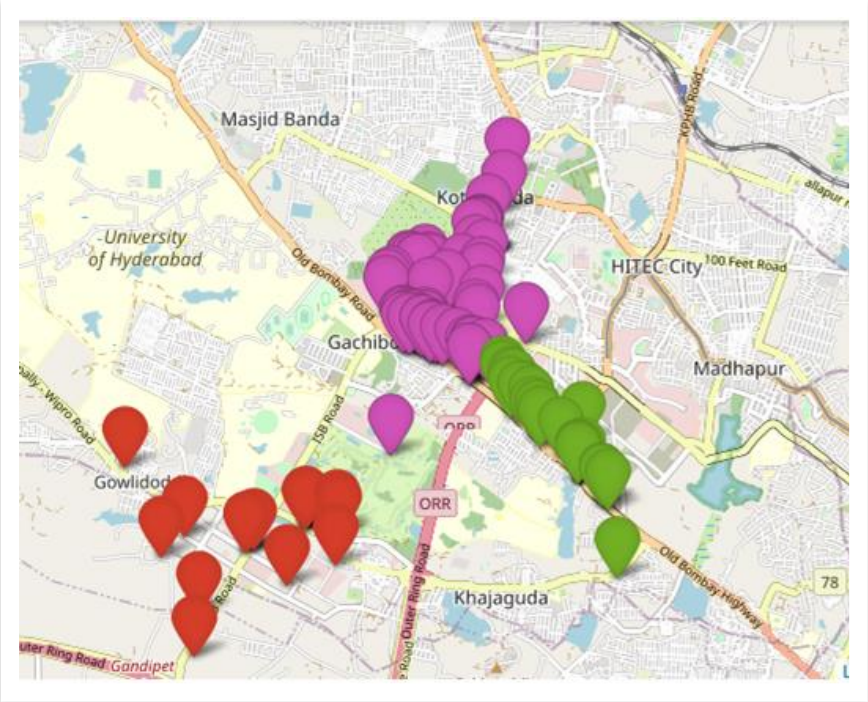


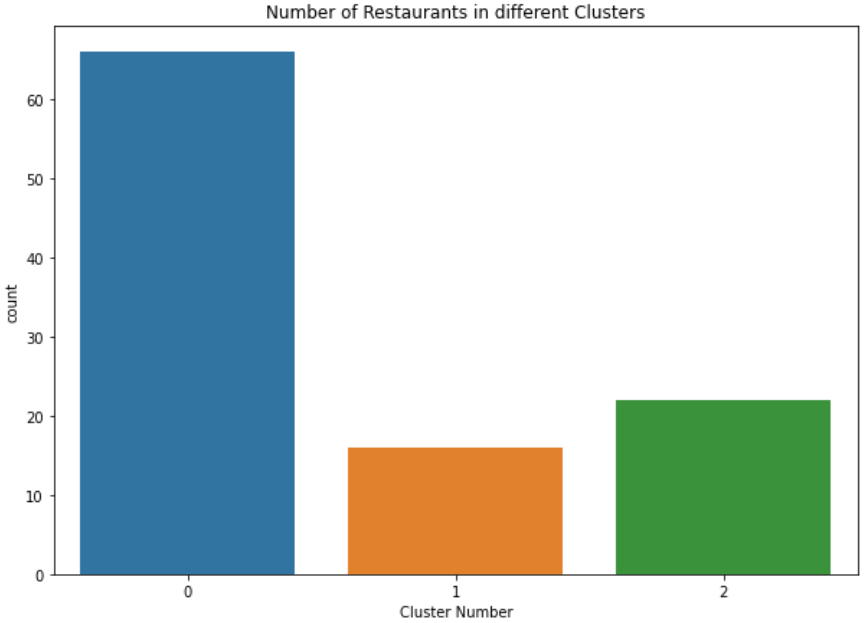




So, according Elbow method, & silhouette score optimum number of cluster(k) will be 3 since the curve almost tapers out after k = 3 & silhouette score is maximum at k =3 .

Clustering of Restaurants

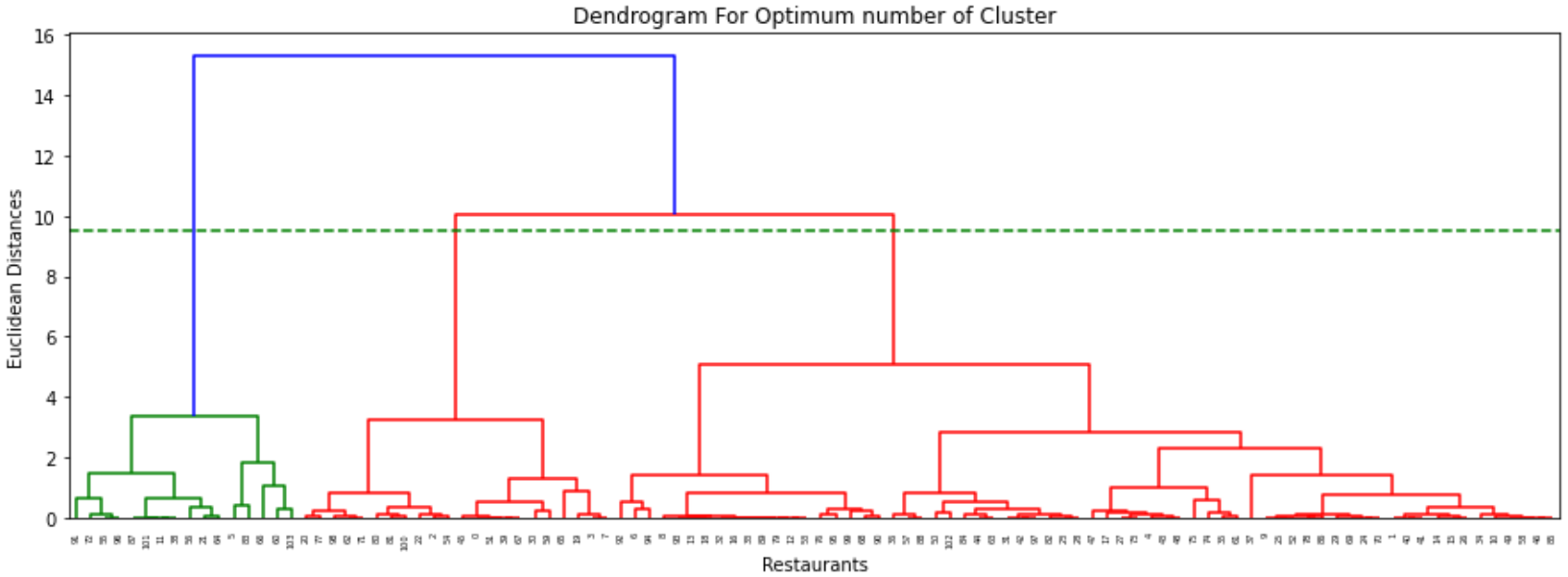




Looking at the above map, we can name the three geographical clusters we obtained as follows:

- Cluster 0: Old Bombay Rd(between ORR junction and ISB Junction)
- Cluster 1: Near Wipro and ISB
- Cluster 2: Old Bombay Rd(between Khajaguda Jn and ORR Jn)

2. Hierarchical Clustering



So as per the dendrogram (Hierarchical Clustering algorithm) we can chose the optimal number of clusters will be equal to 3.

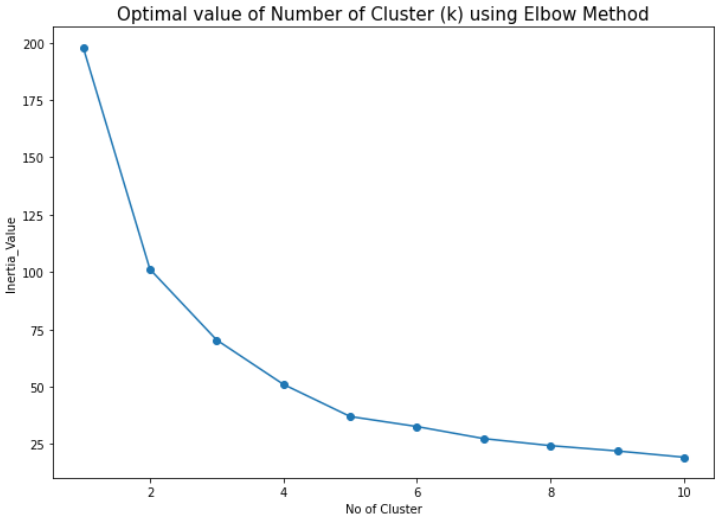
Comparison of Cluster

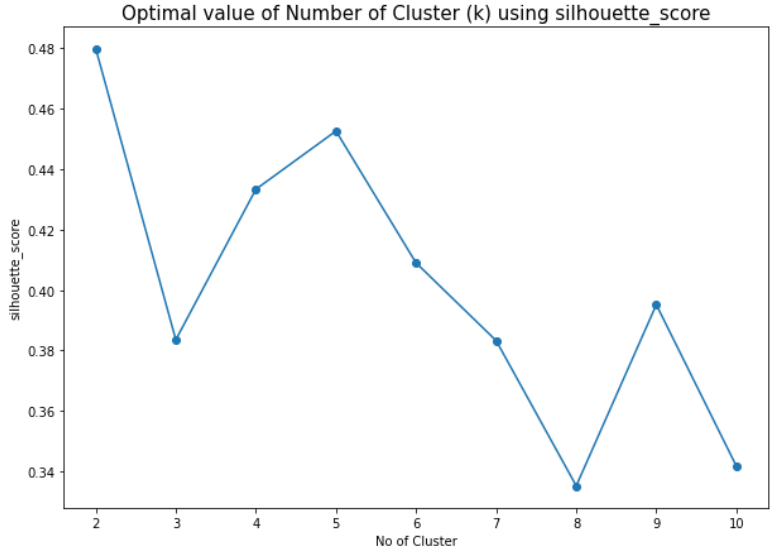
Cluster obtained by the K Means Clustering Algorithm & Hierarchical Clustering Algorithm are almost same. But specifically for clustering the restaurants based on the coordinates K means Clustering will be the best method because it is distance based algorithm.

Clustering of Restaurants

II. Clustering Based on the Rating & Dining Cost per person at restaurants.

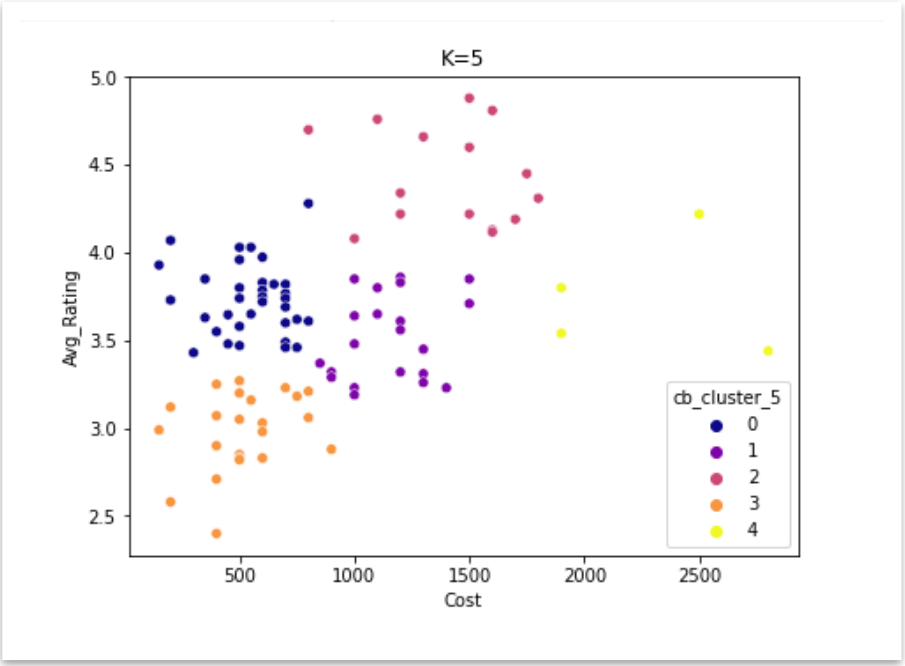
1. K Means Clustering Algorithm





So, according Elbow method, & silhouette score optimum number of cluster(k) will be [2,5] since the curve almost tapers out after k = 2 to5 & silhouette score is maximum at k =2 & 5. So lets select K=5

II. Clustering Based on the Rating & Dining Cost per person



cb_cluster_5	No of restaurants	Median Cost	Min Cost	Max Cost	Avg Rating
0	35	550.0	150	800	3.738376
1	21	1200.0	850	1500	3.514762
2	15	1500.0	800	1800	4.431176
3	24	500.0	150	900	2.976355
4	4	2200.0	1900	2800	3.750000

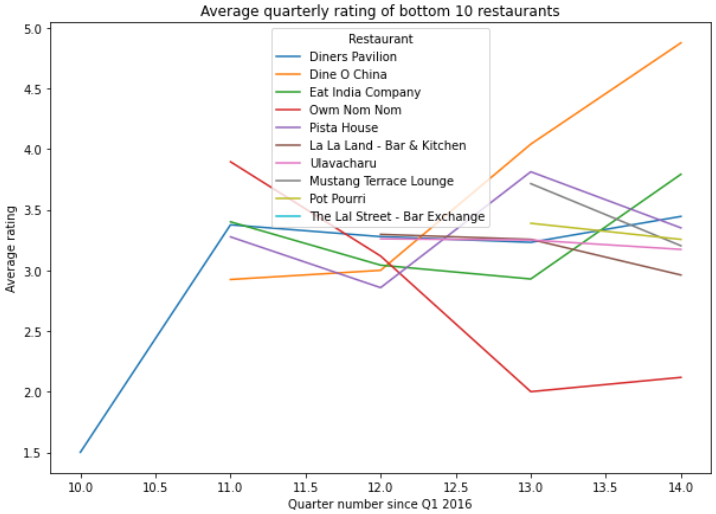
So we have 5 clusters based on Cost & Rating :

1. Cluster: 0 ---> Low Cost and Good Rating
2. Cluster: 1 ---> High Cost and Poor Rating
3. Cluster: 2 ---> High Cost and Good Rating
4. Cluster: 3 ---> Low Cost and Poor Rating
5. Cluster: 4 ---> Super Costly

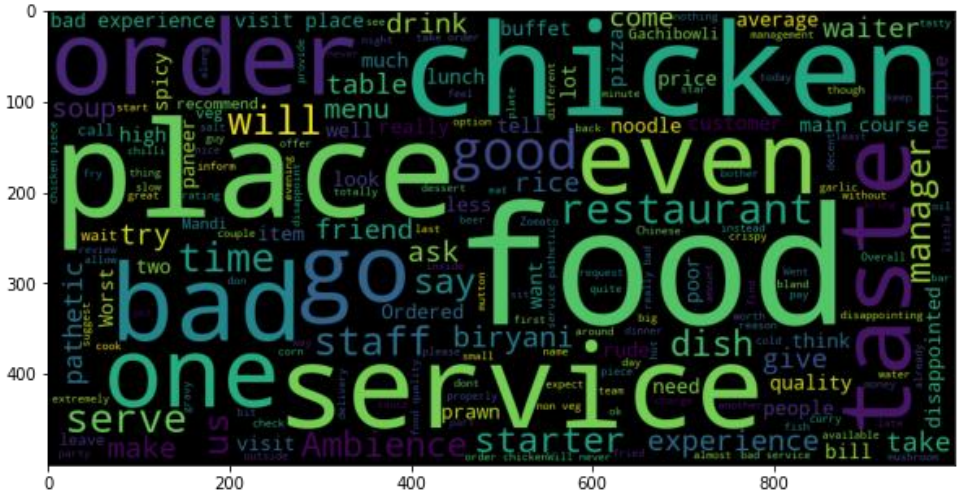
Cost Benefit Analysis of Restaurants

High Cost and Poor Rating Restaurants

Average Quarterly Rating



Word Cloud of Negative Reviews



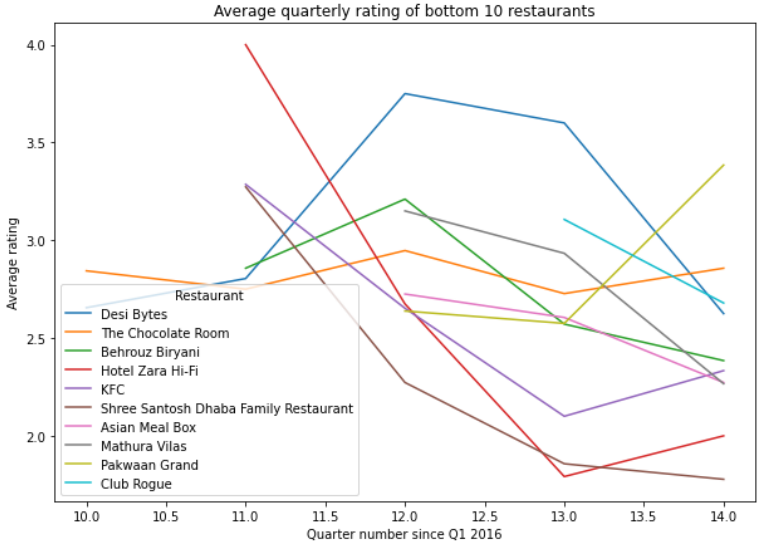
Dine O China is continually improving its performance except it all other restaurants have continued with their average performance for the last 3 quarters.

Customers are very unhappy with respect to :
Taste of Chicken items, service, time, ambience, staff, manager, service, management, experience indicate that many customers are unhappy with the dining experience at the restaurants.

Cost Benefit Analysis of Restaurants

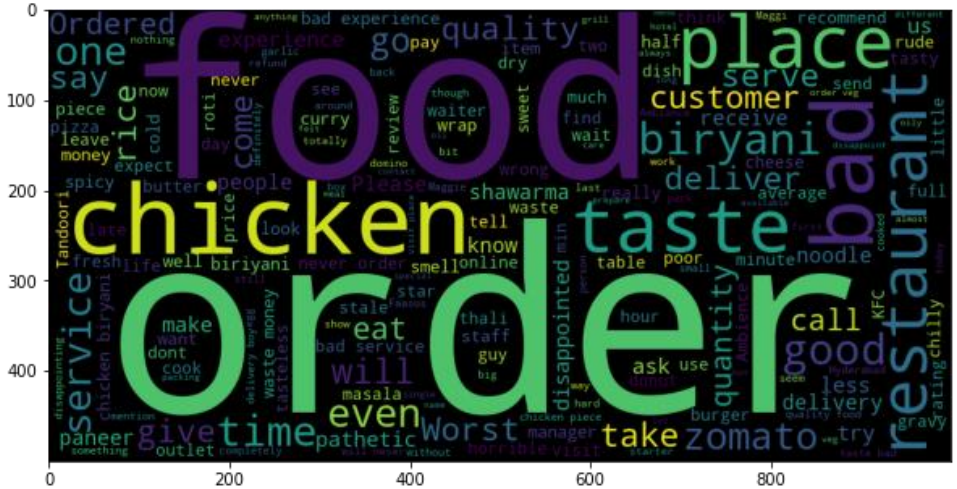
Low Cost and Poor Rating Restaurants

Average Quarterly Rating



Except for Pakwaan Grand, which made an improvement from 2.5 to 3.5, all restaurants have either continued with their poor performance or have become even worse in terms of user rating

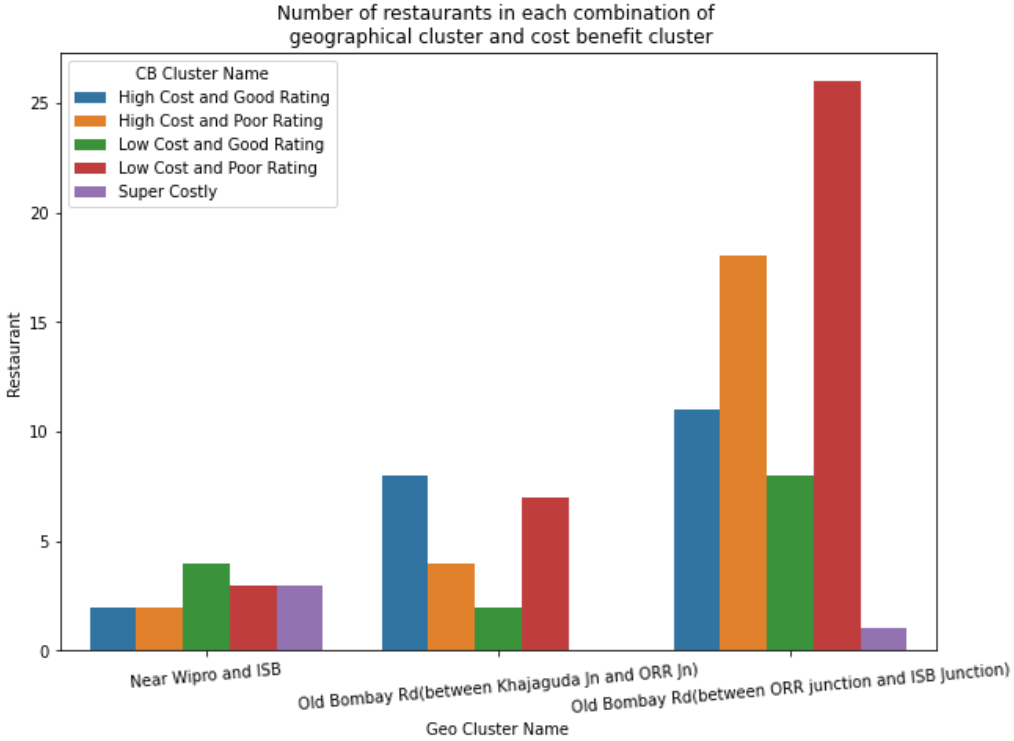
Word Cloud of Negative Reviews



Customers are very unhappy with respect to:

1. Chicken items and Biryani. The taste and quality of these items need immediate attention.
2. Customers also highlighted the issues related to delivery and delivery time.

Recommendation of restaurants for customers



- Most of the restaurants are located in the ORR junction and ISB junction area. This area contains the most affordable restaurants.
- Except one, all the super costly restaurants are located near Wipro and ISB.
- High cost restaurants with good rating are available in all the areas of Gachibowli

Demonstration of Customer Recommendation System

1. Customer who want to find best Restaurant in Wipro & ISB Area, Very Luxurious Category, Continental Cuisines

```
user_choices()

Please select the choices as requested. If you do not want to select a filter, just hit enter. Thank you.


Please enter the area of the restaurant from among the following options:
0) Near Wipro and ISB
1) Old Bombay Rd(between Khajaguda Jn and ORR Jn)
2) Old Bombay Rd(between ORR junction and ISB Junction)
0


You have selected Near Wipro and ISB


Please select what cost category you are looking for from among the following options:
0) High Cost
1) Low Cost
2) Very Luxurious
2


You have selected Very Luxurious category.


Please select which cuisines you want. If you want to explore multiple cuisines of a
 single restaurant, enter all the options separated by ","
0) North Indian
1) Chinese
2) Continental/ Mexican
3) Biryani/Mughlai
4) Asian
5) Fast Food
6) Desserts/ Juices / Bakery
7) South Indian
8) Seafood
9) Arabian
2


You have selected these cuisines: ['Continental/ Mexican']


Please select the additional services you want from the following.
 If you want to explore multiple cuisines of a single restaurant,
 enter all the options separated by ","
0) Outdoor seating
1) Entertainment
2) Wifi
3) Breakfast
4) Parking available
5) Seating available
6) Alcohol avialable
7) Family Friendly
8) Home Delivery
9) Brunch
10) Romantic Dining
2


You have selected these services:['Wifi']


Please enter 1 if you want the restaurant to be featured in Hyderabad's best list0


You have selected 1



---------------------------------------------------------------------
There are only 2 restaurants in Gachibowli for the given selection. These are the names of those restaurants:
1) Feast - Sheraton Hyderabad Hotel
2) Jonathan's Kitchen - Holiday Inn Express & Suites
---------------------------------------------------------------------
Thank you!
```

So in this way according to need of customer we can recommend best restaurant in there area.

Conclusion

- Through this project we have demonstrated our ability to effectively process and explore an unlabeled dataset and implement unsupervised algorithms like sentiment analysis and k-means clustering, Hierarchical clustering algorithm in Python.

- We were able to obtain actionable insights from an extensive technical analysis of the given dataset to improve the areas where the business is currently lagging in.

- This project also helped us to hone our python programming skills where we learned to use multiple libraries like Numpy, pandas, sklearn, folium maps, matplotlib, seaborn, nltk, NLP, TextBlob & Folium etc.

- We have gained a thorough understanding of how to question the data by using appropriate data visualization techniques throughout the project.


thank
YOU