Capstone Project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

Vilas Sonawane
Nandeesh Umesha
Bhavika Gaurkar
Soumya Ranjan Dash

Introduction

- Zomato is an Indian restaurant aggregator and food delivery start-up founded 2008.
- Customers use Zomato platform to search restaurants, read/ write customer reviews, order food online or book a table for dining.
- It also provides restaurant partners with industry-specific marketing tools to enable them to engage and acquire customers.
- The growing number of restaurants going online 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.
- Inspired by the same idea, this project analyses the data related to restaurants in Gachibowli, Hyderabad and the customer reviews they obtained to find actionable insights for the business to improve in the areas they are currently lagging in.

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 - Name, cost per person, cuisines, collections, timings of each restaurant.

- **Zomato Restaurant Reviews.csv:**
  - **Rows:** 10000 rows and each row corresponds to a customer review. We had customer reviews pertaining to 100 restaurants given by ~7400 unique customers.
  - **Columns:** 7 columns - restaurant, reviewer, rating, metadata of reviewer, review(text), time of review, number of pictures in the review

Project Description

The project focuses on three broad objectives:

1. Analyze the sentiments of the reviews given by the customers and make some useful conclusions.
2. **For the Company:** Cluster the zomato restaurants into different segments. 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. **For the customer:** Provide a methodology to find the best restaurants in Gachibowli.

Data cleaning - Restaurants data

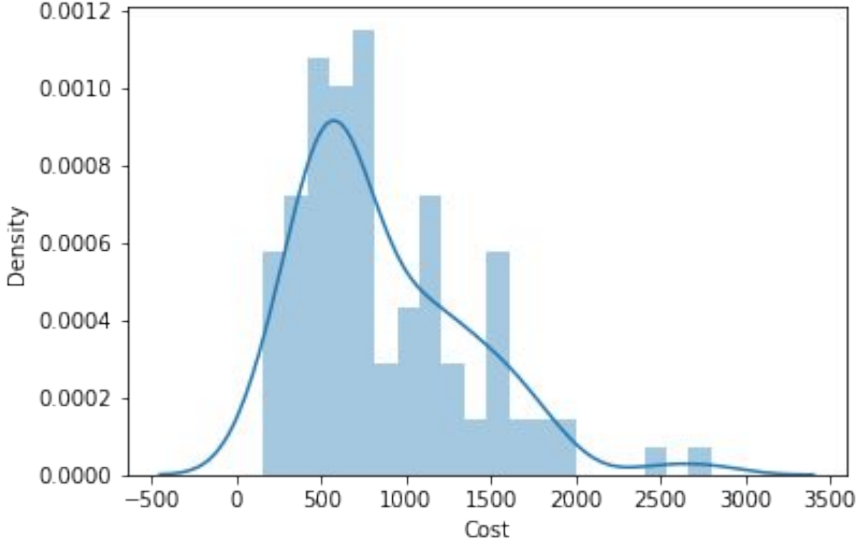
- **Web scraping** for extracting additional data from the web links of restaurants. **(Addition of New Features)**
    - Latitude | Longitude | additional services | Has Featured
    - Using Scraping Bee services and parsing html strings using regex
- Cleaning of features like - Collections, Cuisines, 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.

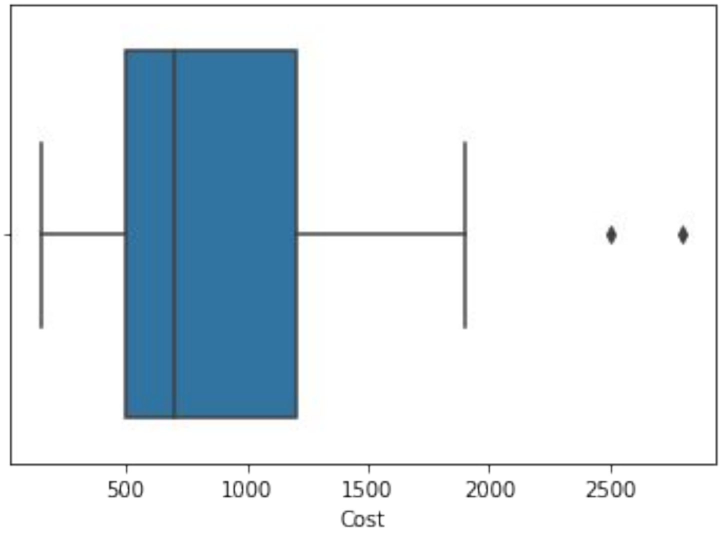
Data cleaning - Reviews data

- **Handling Missing Values:** (Drop the rows as no is very Small)
- **Rating:** One rating was given as 'Like', so replaced 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

Per person cost

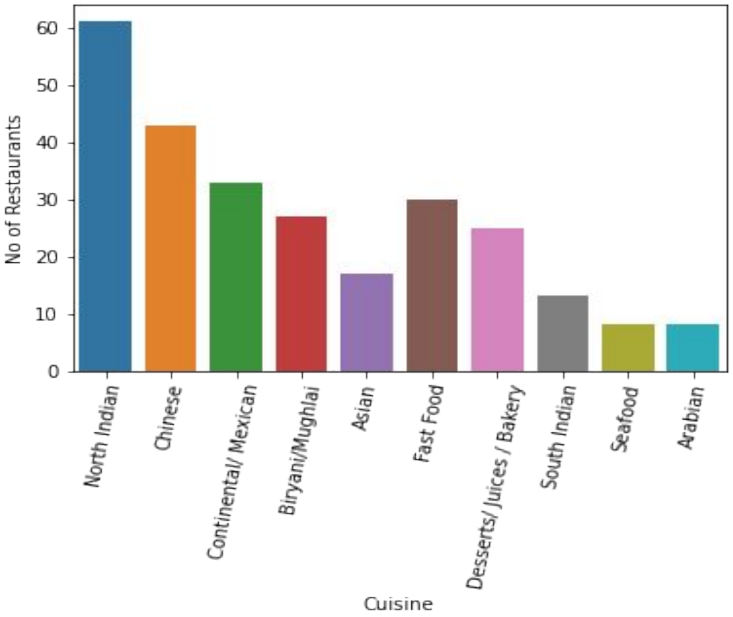




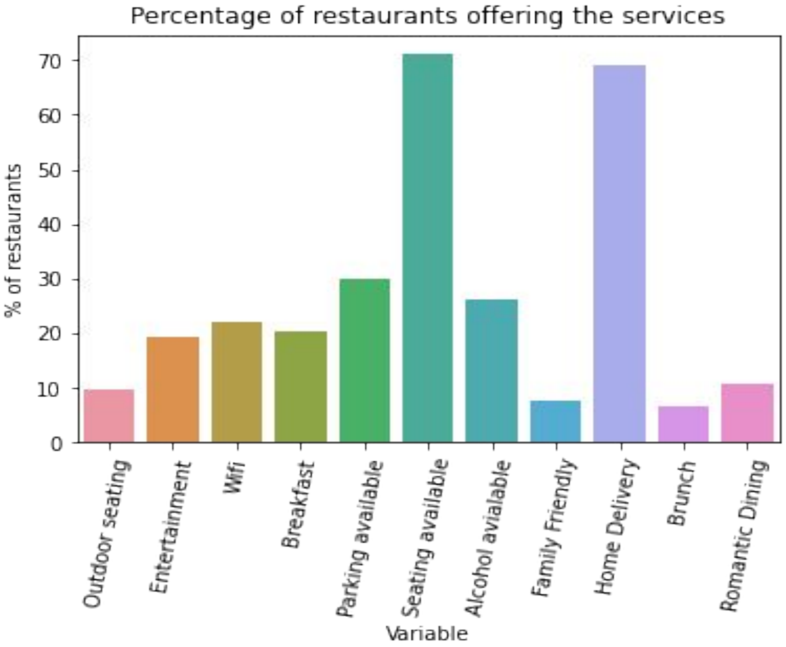
Cost of food at the restaurants are in the range of 150 to 2800 rupees per person with a median of ~700 Rs.

Univariate Analysis - Restaurant Data

Cuisines



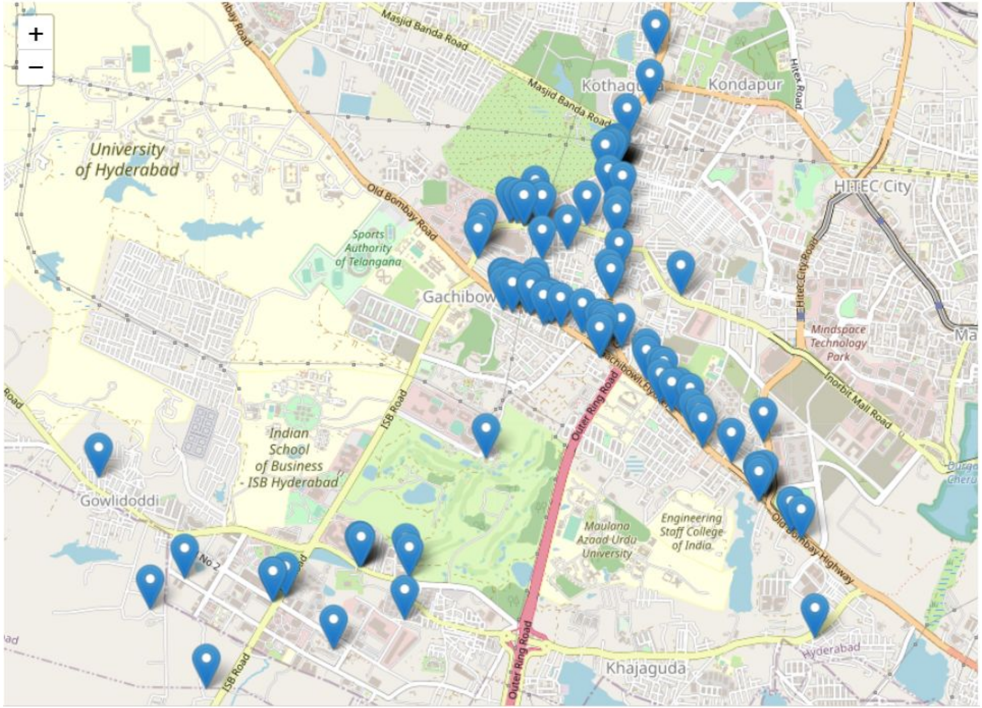
Services offered



North Indian is the predominant cuisine

Univariate Analysis - Restaurant Data

Latitude & Longitude

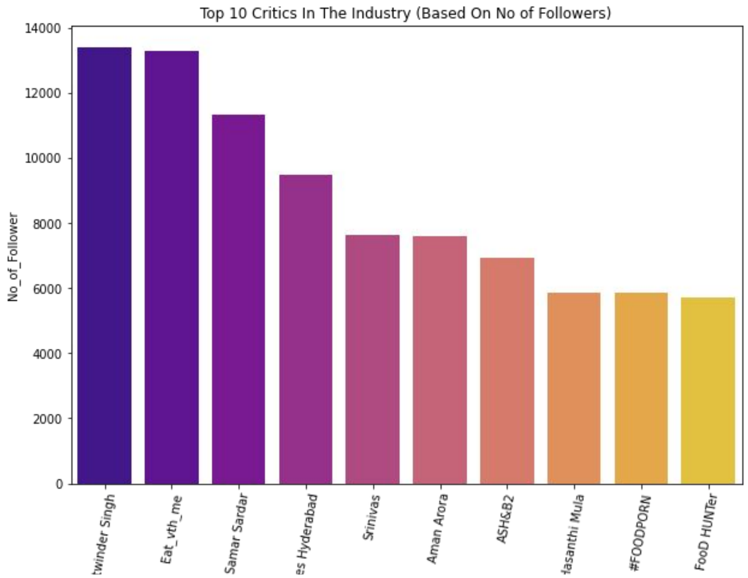


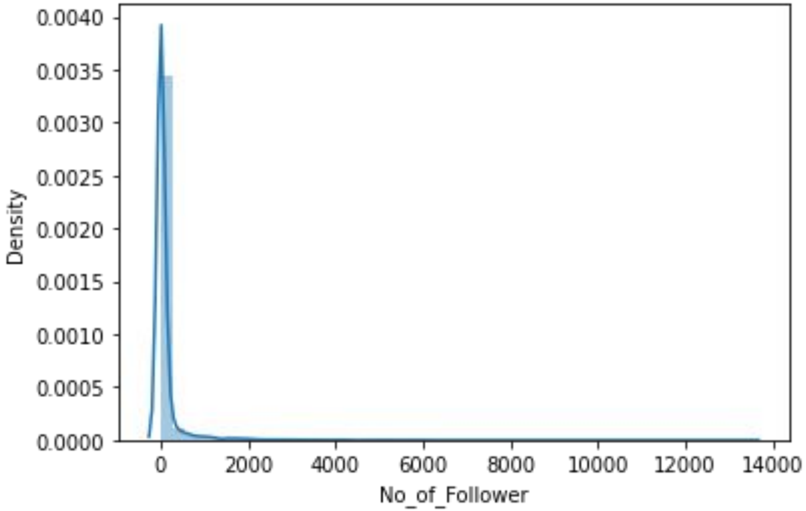
Majority of the restaurants seem to be located on the old - Bombay highway. It is interesting to note that three clear clusters are visible clearly:

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

Univariate analysis - Reviews data

Top 10 Critics in The Market

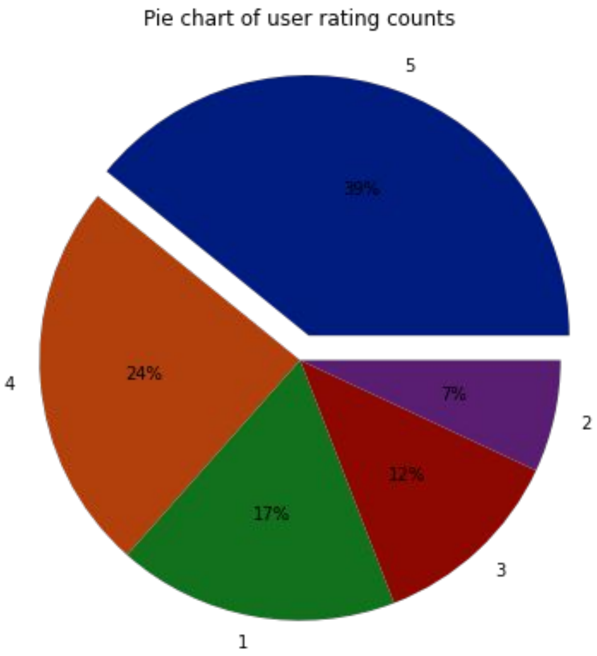




200 followers was chosen as the cutoff above which a reviewer will be called a critic since there is significant increase from 90th percentile to 95th. And also it is commonsensical that a person having ~200 followers on zomato should be a critic.

Univariate analysis - Reviews data

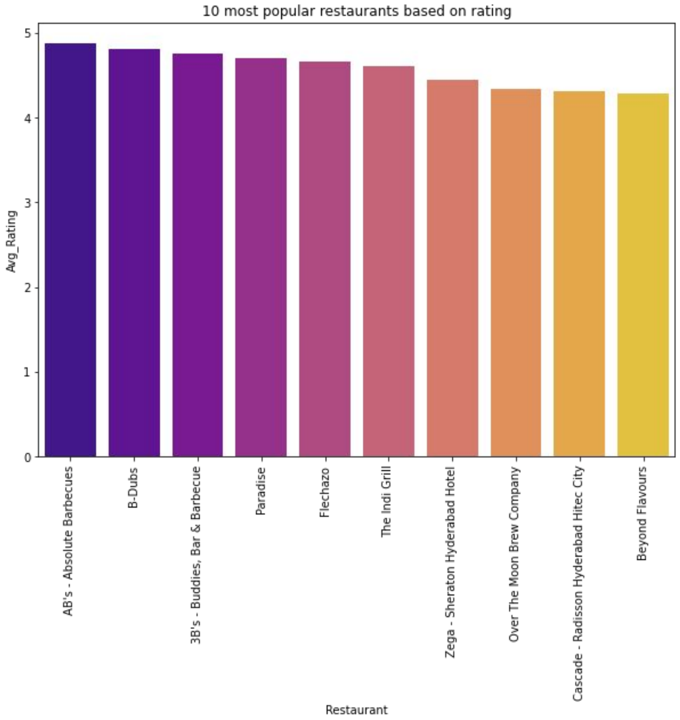
Ratings


Pie chart of user rating counts

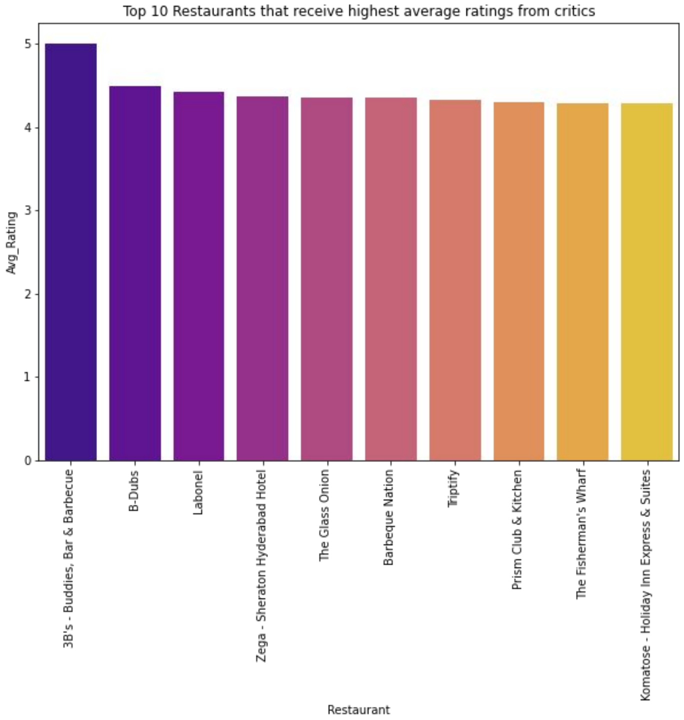
- 39% of the reviews represent 5 star rating
- Another 24% are 4 star ratings.
- So the reviews should largely be skewed towards positive opinions.

Multivariate Analysis

10 most popular restaurants (overall)

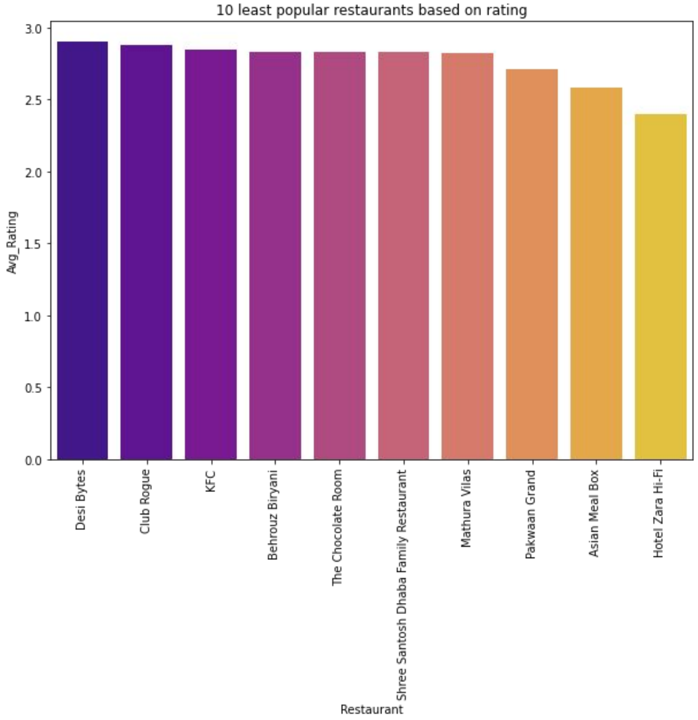

10 most popular restaurants based on rating

10 most popular restaurants (based on critics only)

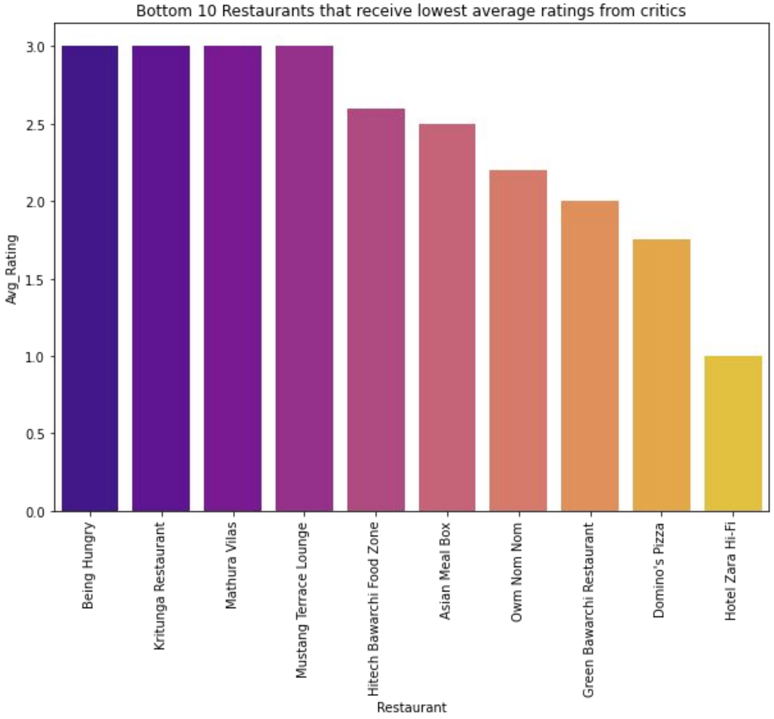

Top 10 Restaurants that receive highest average ratings from critics

Multivariate Analysis

10 least popular restaurants (overall)

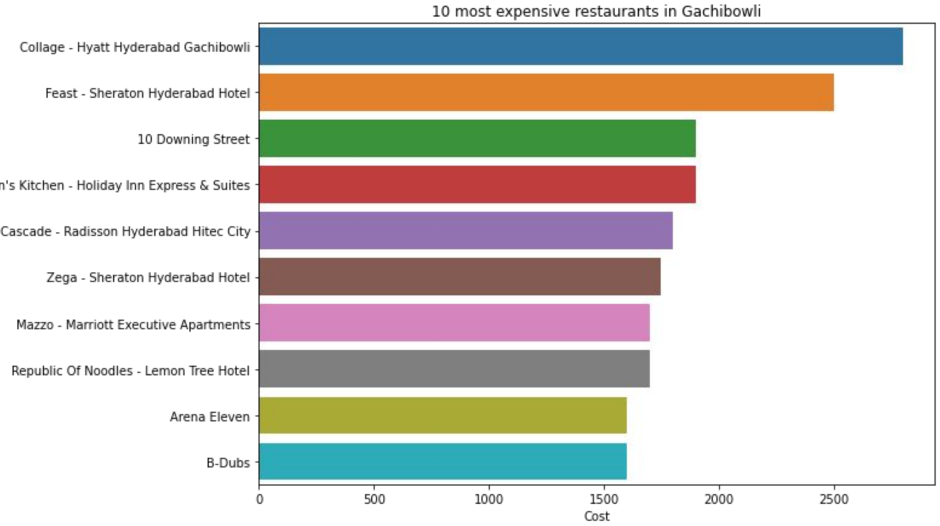


10 least popular restaurants (based on critics only)

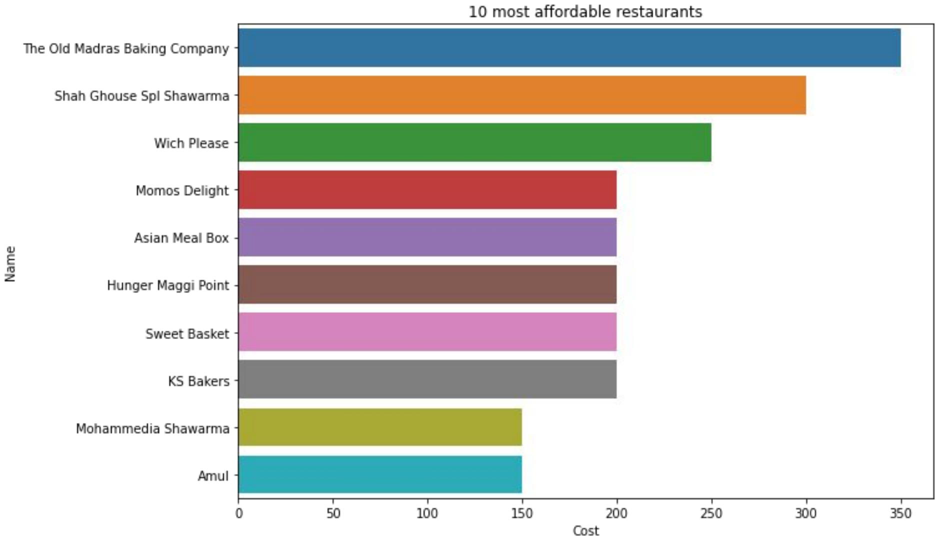


Multivariate Analysis

10 Most expensive restaurants



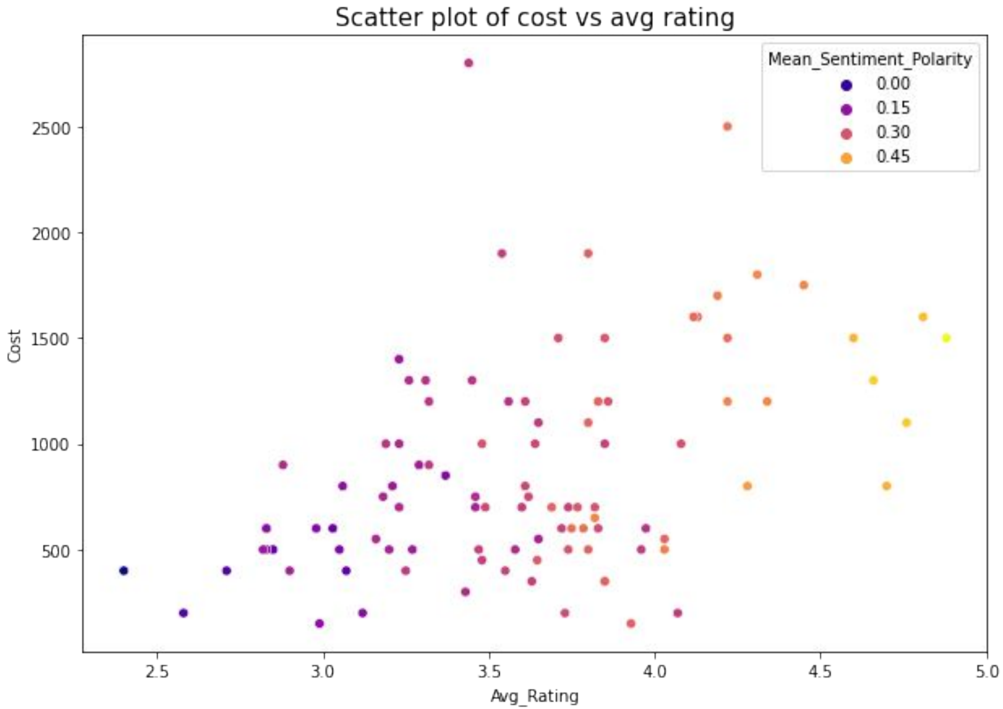
10 least expensive restaurants



All of the 10 least expensive restaurants cost less than 350 Rs per person. According to Indian food market standards, the restaurants listed in Zomato in Gachibowli are relatively costly options. Inclusion of cheaper restaurants may help improve the user base of Zomato.

Multivariate Analysis

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 customers are more satisfied with the costlier restaurants than with the cheaper restaurants.

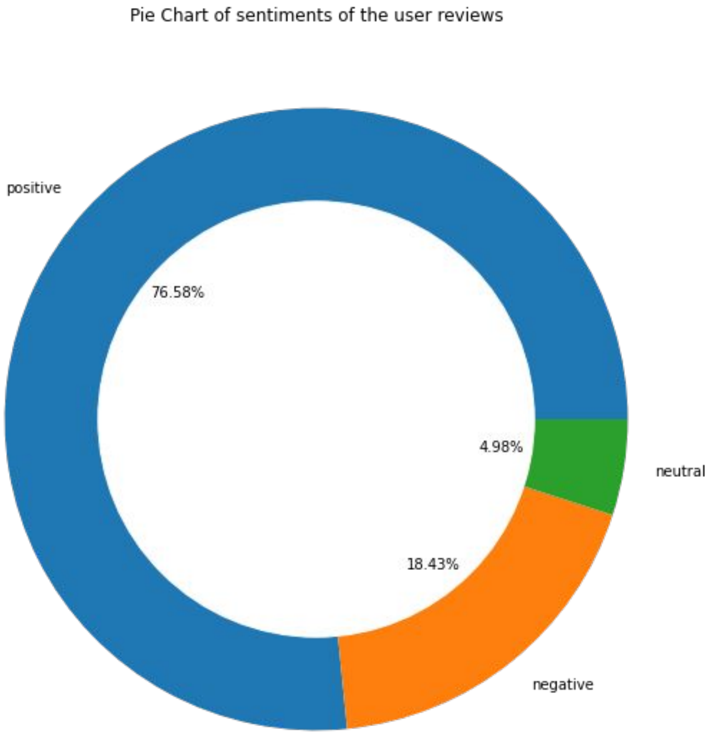
Further study and analysis needs to be done to find out the reason behind the poor performance of the affordable restaurants. We will use clustering techniques in the following section to perform the analysis.

Sentiment Analysis of user reviews

Steps:

- Cleaning of text of reviews
  - ○ Removing punctuation marks, hyperlinks, special characters using regular expressions
  - ○ Removal of stop words
  - ○ Part of speech tagging - identifying the part of speech of the words
  - ○ Lemmatization - to convert the words into their root words
- Implementation of sentiment analysis using Python’s textblob module
- Exploratory data analysis of the results

Sentiment Analysis of user reviews

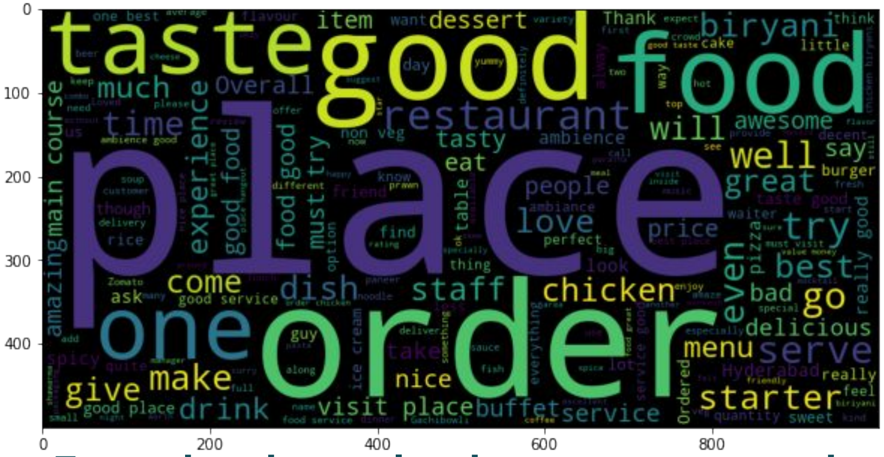


- 77% of the reviews represent positive sentiments
- 18% of them represent negative sentiments. This is in accordance with our earlier results on univariate analysis on ratings column where we saw that 63% of the reviews had a rating of 4 or higher.
- Only 5% of them represent neutral sentiments

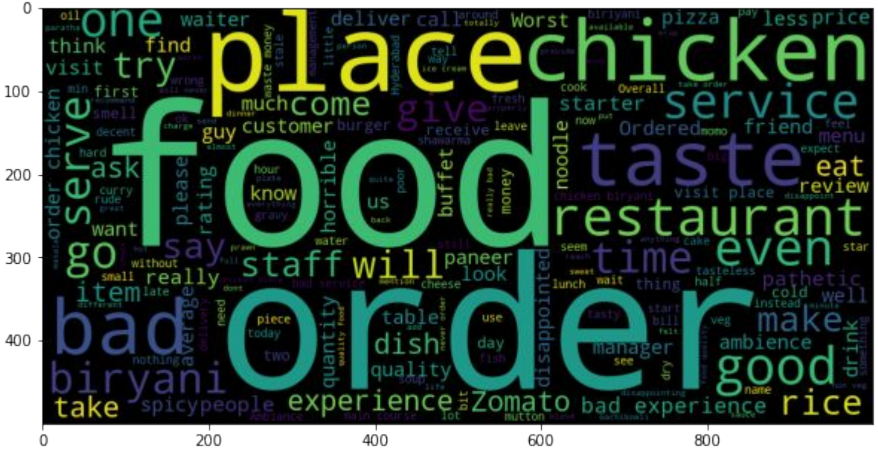
Sentiment Analysis of user reviews

Most frequent words used by reviewers

Positive reviews



Negative reviews

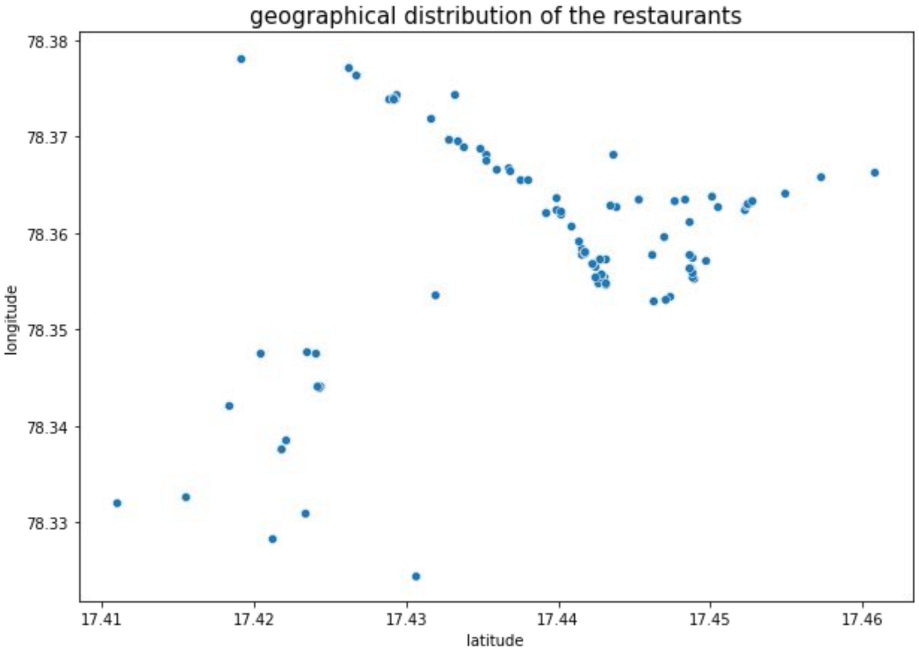


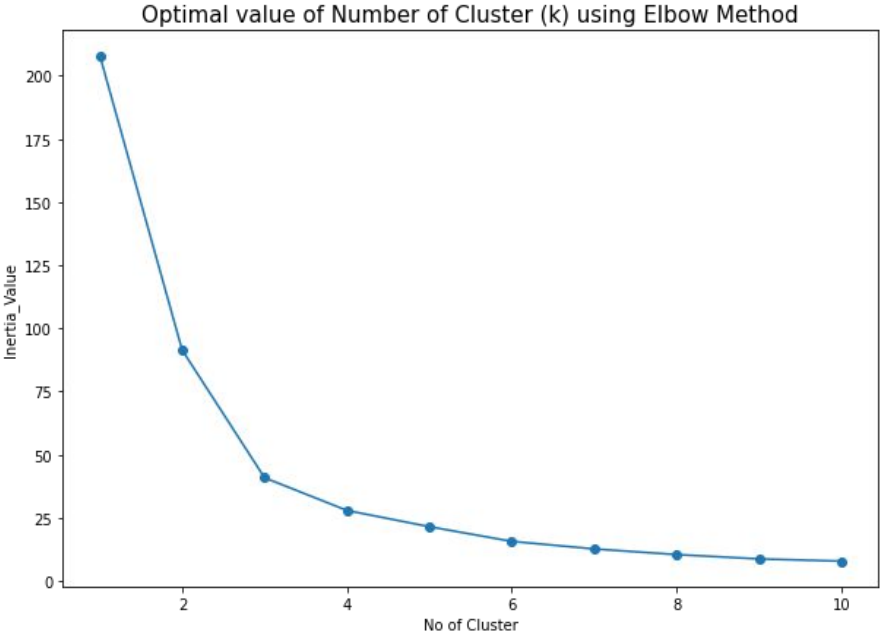
From the above cloud we can suggest that the following areas need to be improved

- 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

Clustering Based on the Location of Restaurants

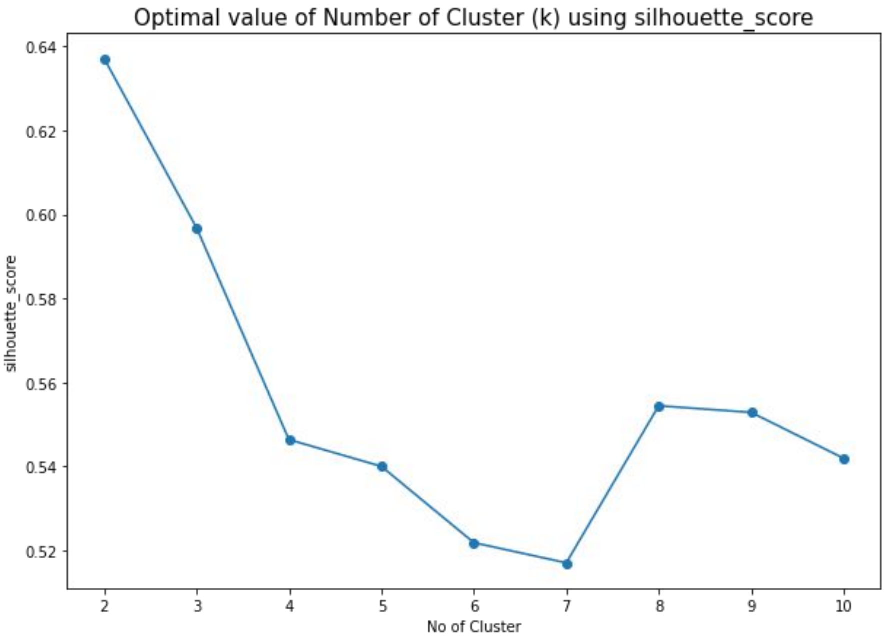


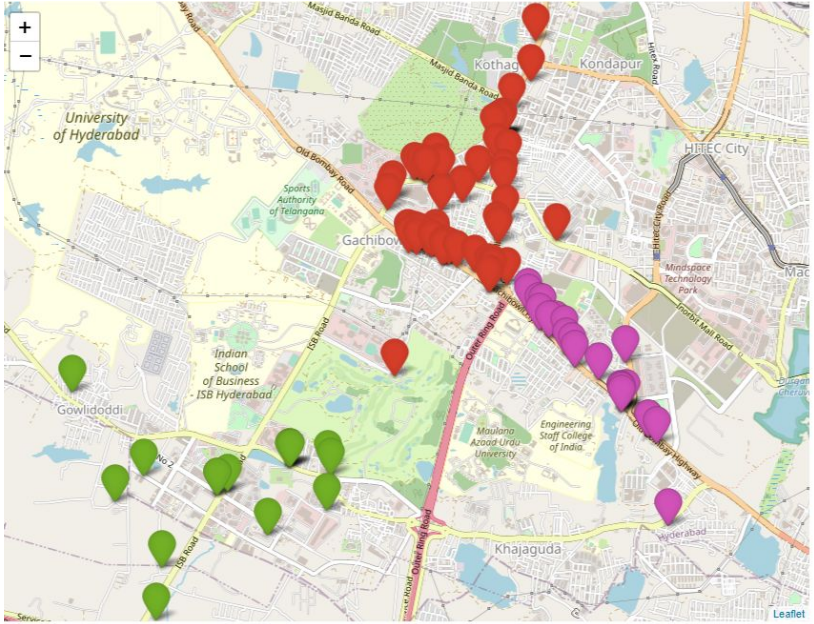


So, according Elbow method, k=2 or k=3 might be a good choice

Clustering of Restaurants

Clustering Based on the Location of Restaurants




+
−
University of Hyderabad
Kothag
Kondapur
HITEC City
Gachibow
Sports Authority of Telangana
Indian School of Business - ISB Hyderabad
Gowlidoddi
Mindspace Technology Park
Maulana Azaad Urdu University
Engineering Staff College of India
Khajaguda
Masjid Banda Road
Old Bombay Road
ISB Road
Outer Ring Road
Leaflet

k=3 will be chosen since it has decent silhouette score and provides better clusters for our use case

Clustering of Restaurants

Clustering Based on the Location of Restaurants

Looking at the clusters on the map, we named the three clusters as below.

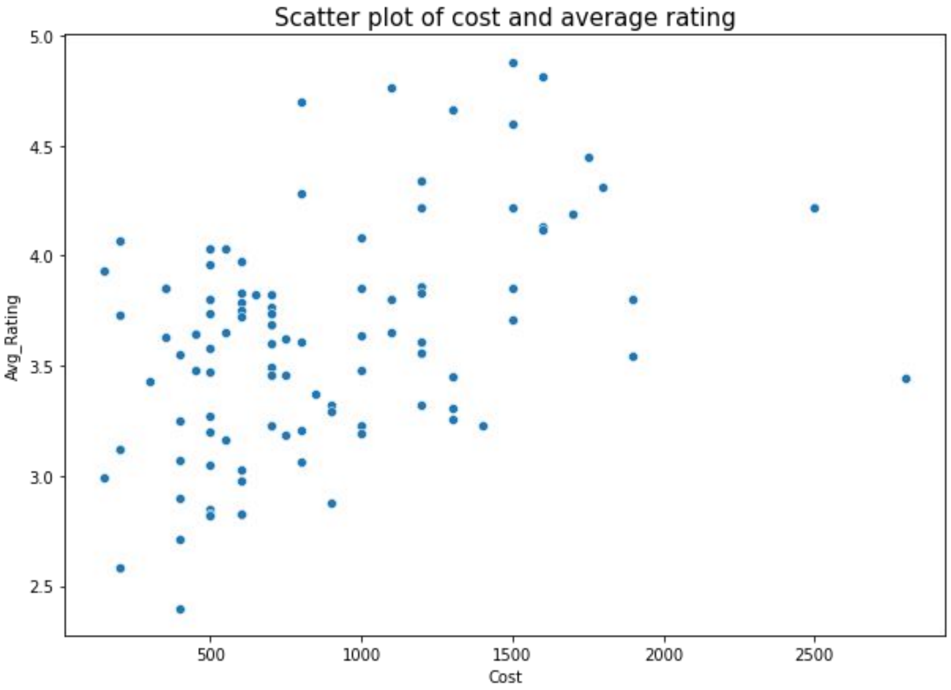
Cluster Profile

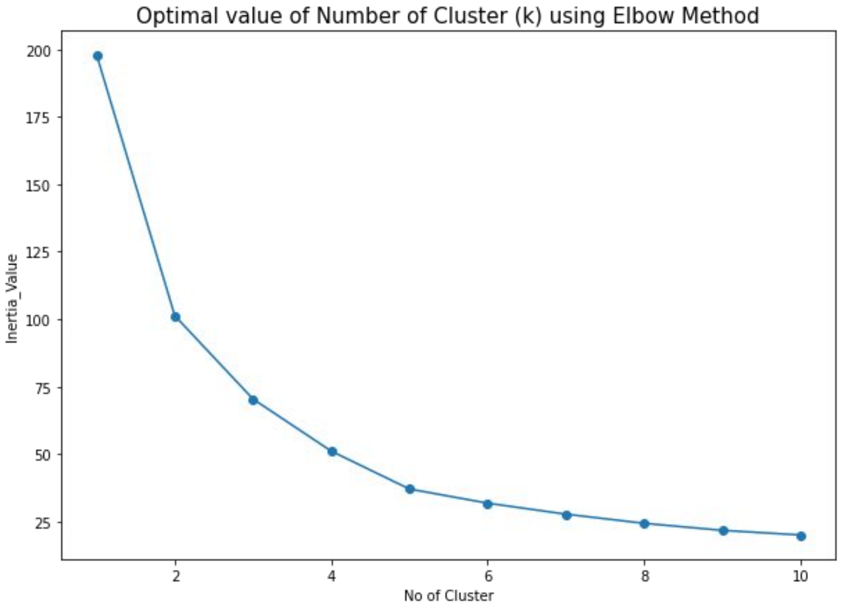
Geo Cluster Name	No of restaurants	Avg Cost	Min Cost	Max Cost	Avg Rating
Near Wipro and ISB	14	1342	400	2800	3.8
Old Bombay Rd(between Khajaguda Jn and ORR Jn)	21	914	300	1600	3.6
Old Bombay Rd(between ORR junction and ISB Jn	64	746	150	1900	3.6

- Restaurants near Wipro and ISB are relatively costlier than others in Gachibowli.
- Most of the restaurants are located in the Old Bombay Rd(between ORR jn and ISB jn). So users can find a lot of options within this small stretch.

Clustering of Restaurants

Clustering Based Average Rating and Cost per person

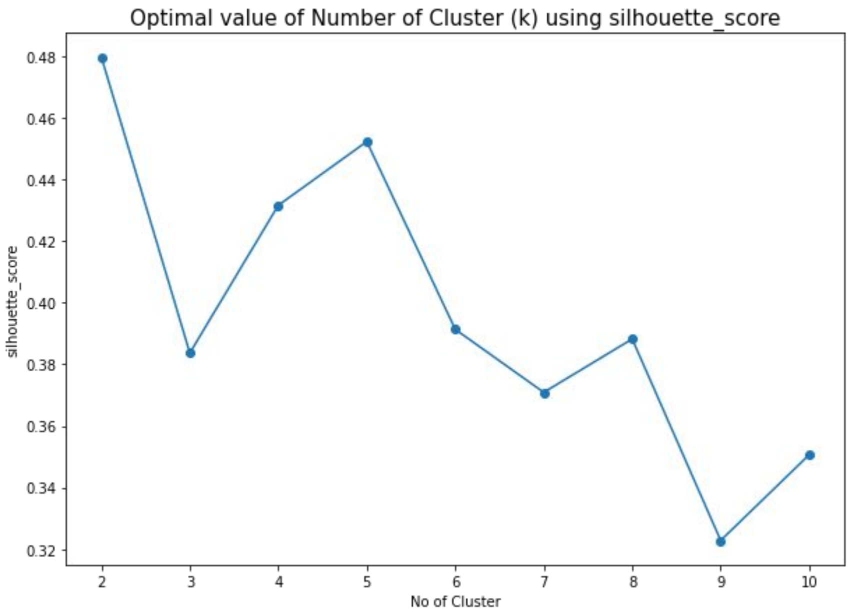




So, according Elbow method, k=2 to k=5 might be a good choice

Clustering of Restaurants

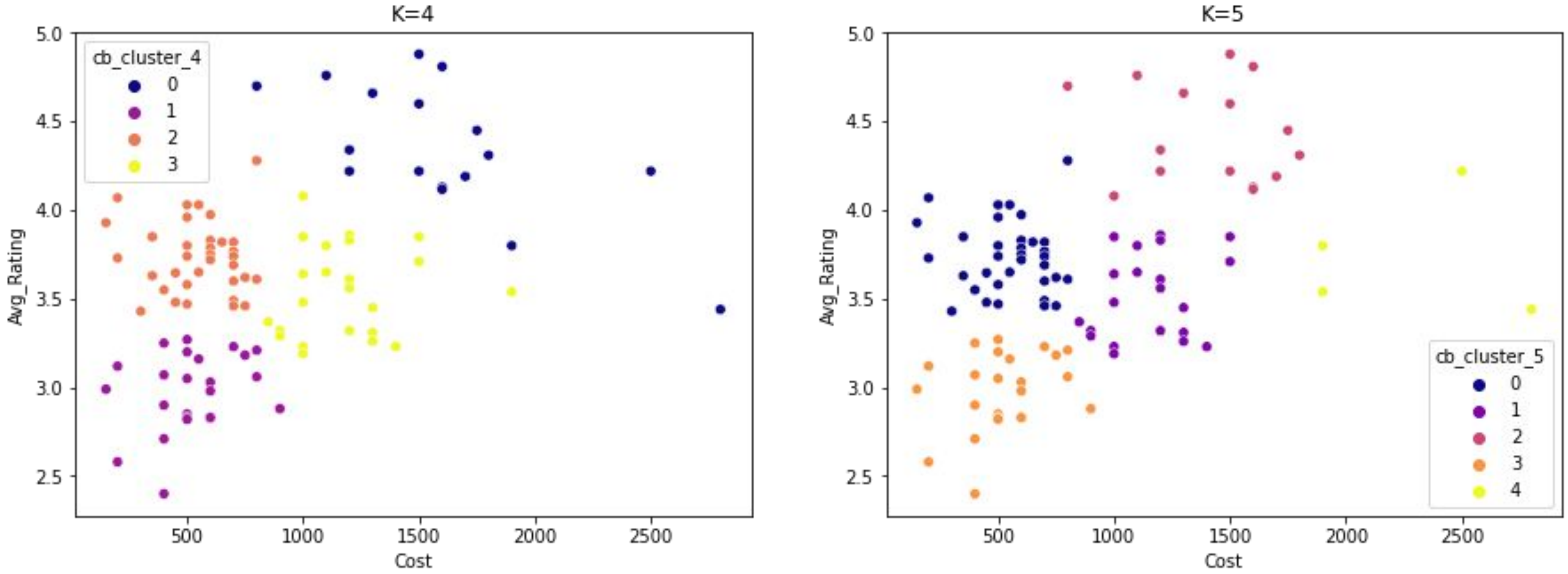
Clustering Based Average Rating and Cost per person



- Silhouette score is reasonably high for K=2,4,5.
- Since K=2 means two clusters which would classify the restaurants into two custers only, it will not be useful for a good cost to benefit analysis.
- So let us look at the visualizations for k=4 or 5 to choose optimal number of clusters for this problem.

Clustering of Restaurants

Clustering Based Average Rating and Cost per person

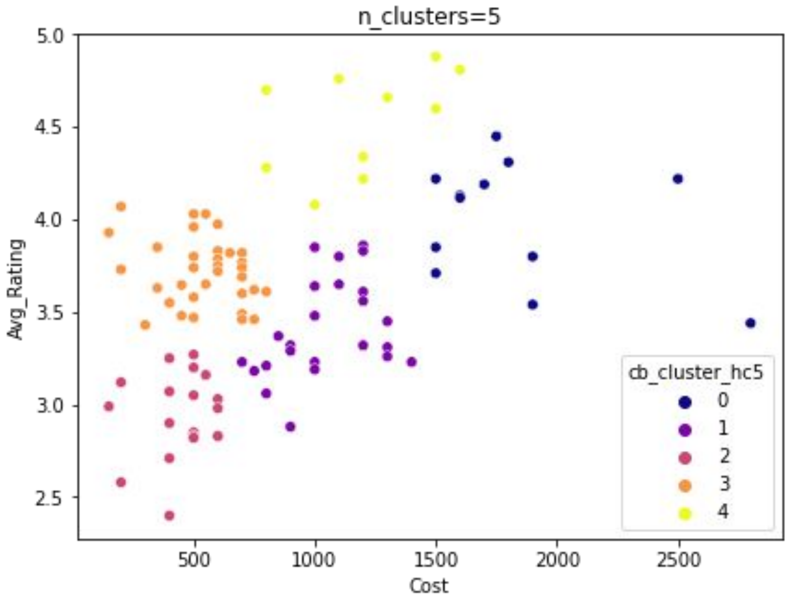


- For k=5 the algorithm tried to fit the outlier points on the super high cost side and make them a separate category.
- In reality, these are not actually outliers but we have very few restaurants that are very costly(>2000/-). Since this is a profile of interest for our study, we will consider number of clusters as 5.

Clustering of Restaurants

Clustering Based Average Rating and Cost per person

- Since we see hierarchical clusters, let us implement hierarchical clustering algorithm to explore if we get better clusters.
- Since our data has slightly globular clusters and the gaps between the clusters is not so large, we will use 'Wards' method instead of single linkage.



- Looking at the visualization of clusters, we can see that the clusters provided by the k-means algorithm have lesser variance than those from the hierarchical clustering.
- So we will select the cluster results we obtained from the k-means clustering algorithm with K=5 only.

Clustering of Restaurants

Clustering Based Average Rating and Cost per person

Cluster Profile

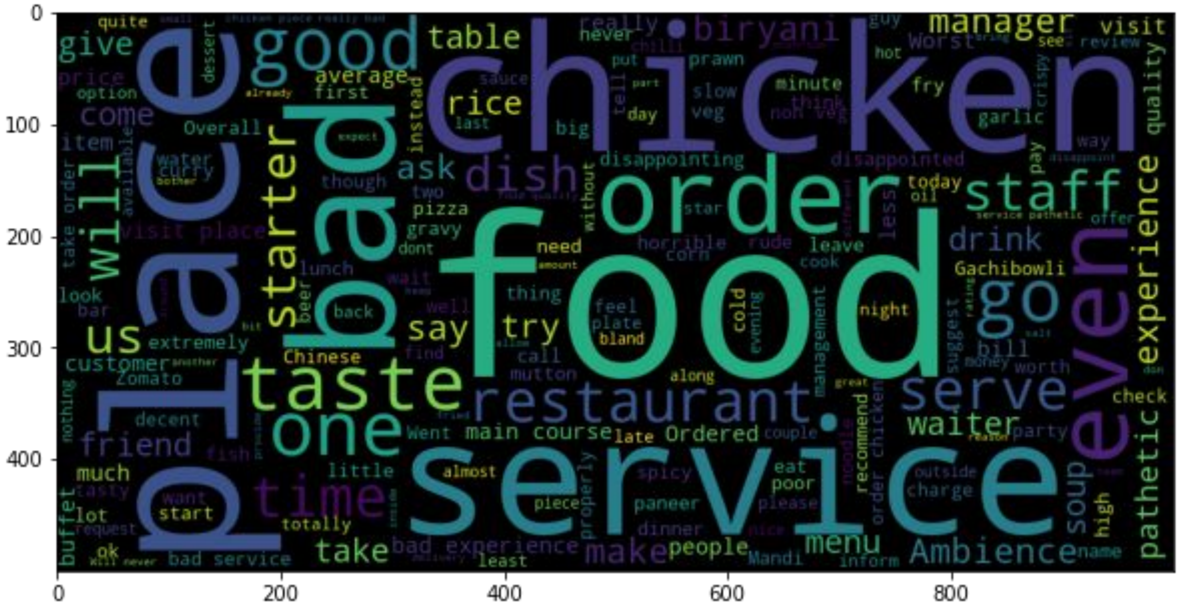
Cluster Name	No of restaurants	Median Cost	Min Cost	Max Cost	Avg Rating
Low Cost and Good Rating	35	550	150	800	3.7
High Cost and Poor Rating	21	1200	850	1500	3.5
High Cost and Good Rating	15	1500	800	1800	4.4
Low Cost and Poor Rating	24	500	150	900	3.0
Super Costly	4	2200	1900	2800	3.8

We saw earlier that high cost restaurants have better ratings on average. So cluster 2(which is high cost) having a mean rating of 3.5 can be considered as having a poor rating. Similarly, Cluster 0(Low cost) which has an average rating of 3.7 can be considered to be having a good rating for low cost restaurants.

Clustering of Restaurants

Analysis of High Cost and Poor Rating Cluster

Word Cloud of Negative Reviews

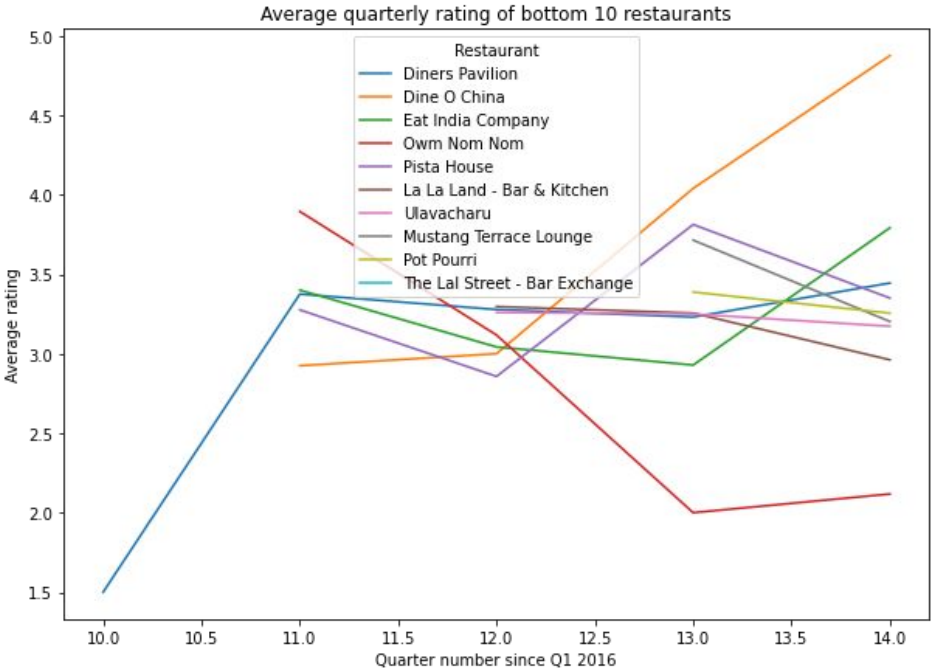


- Words like service, time, ambience, staff, manager, serve, management, experience indicate that many customers are unhappy with the dining experience at the restaurants. This is the predominant theme in the customers' negative reviews.
- Taste of Chicken items needs attention.

Clustering of Restaurants

Analysis of High Cost and Poor Rating Cluster

Quarter-wise performance of the 10 restaurants with the least average rating

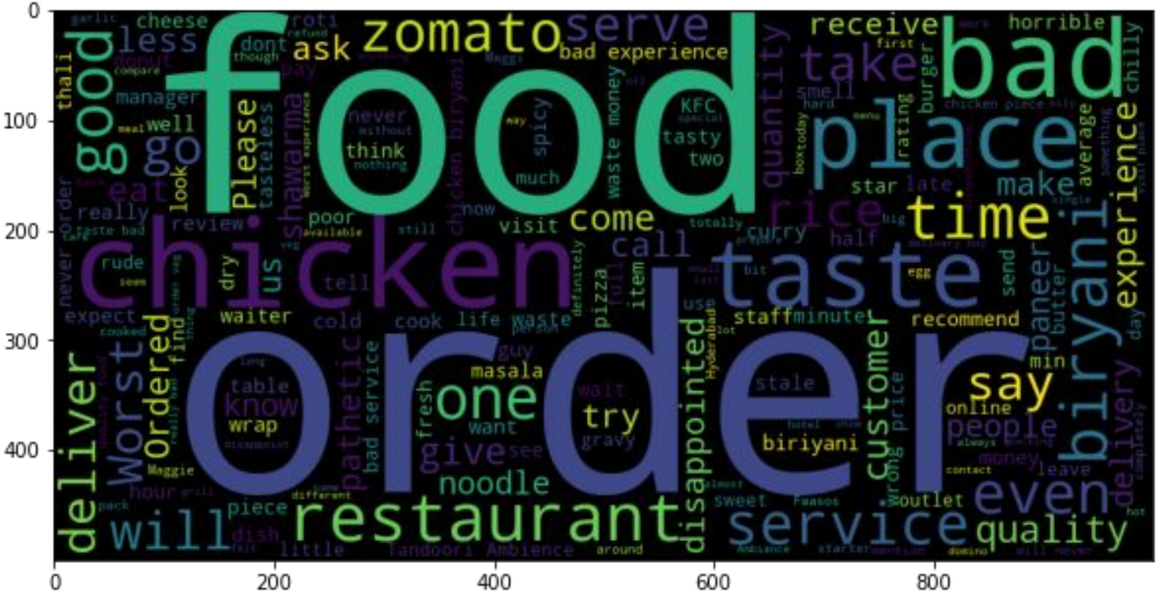


- Except for Dine O China, all other restaurants have continued with their average performance for the last 3 quarters.
- Though there is no drop in rating over time, this may indicate that the restaurants did not act on the negative reviews to improve over time. Only, Dine O China has acted positively to improve its rating from 3 to 5 over the latest 3 quarters.

Clustering of Restaurants

Analysis of Low Cost and Poor Rating Cluster

Word Cloud of Negative Reviews

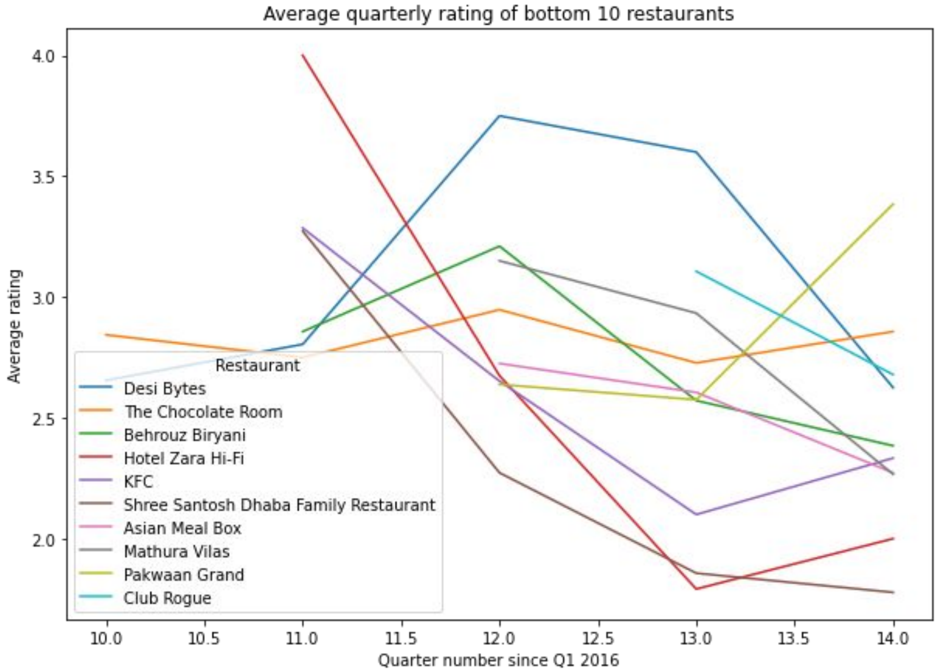


- Users highlighted the issues related to delivery and delivery time by zomato. This may indicate a potential logistical issue from Zomato in delivery of orders from these restaurants.
- The taste and quality of Chicken items and Biryani needs attention.
- Unlike for cluster 3, the predominant theme in this cluster is about delivery of food.

Clustering of Restaurants

Analysis of Low Cost and Poor Rating Cluster

Quarter-wise performance of the 10 restaurants with the least average rating

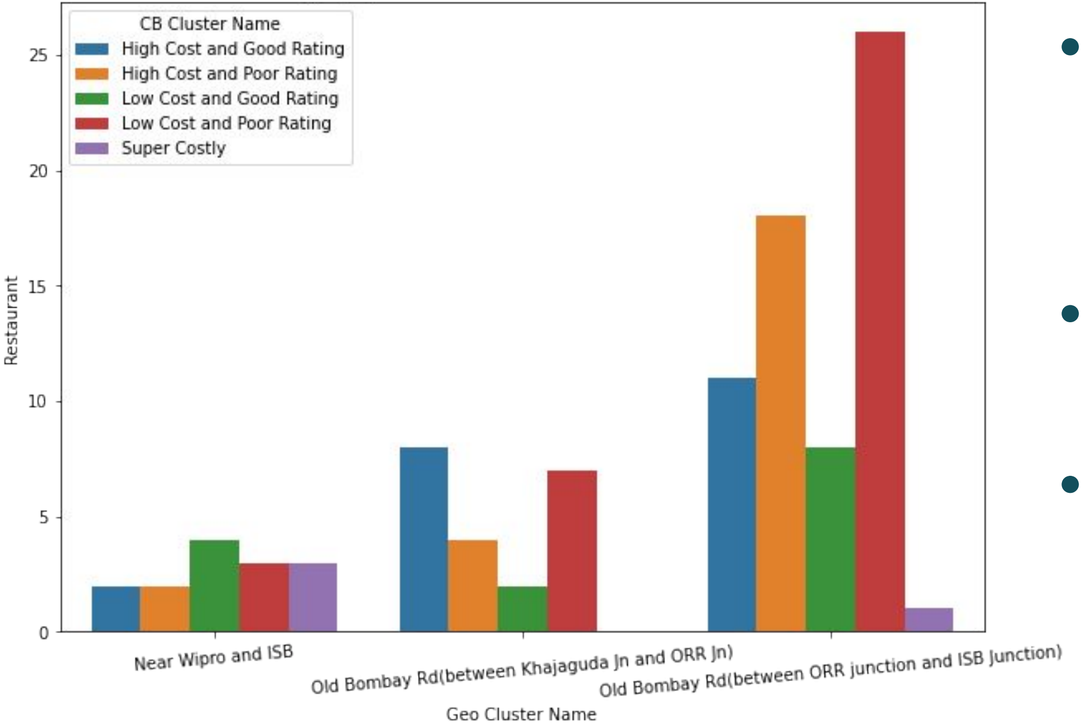


- Except for Pakwaan Grand, which made a slight improvement from 2.5 to 3.5, all the restaurants have either continued with their poor performance or have become even worse in terms of user rating.
- Especially, Shree Santosh Dhaba, Hotel Zara Hi-Fi, Mathura Vilas, Aromas@11SIX and KFC had a steep decline in their ratings.

Recommendation of restaurants for customers

- To recommend the best restaurants for the customers, we used the geographical clustering and cost benefit clustering we performed in the previous section.
- We built an interactive filtering system, which displays the top restaurants based on average user rating for the choices selected by the customer.
- The system allows the user to choose the following filters to get the best restaurants according to his/her needs:
  - Area - 3 choices from three clusters
  - Cost Category - 3 choices - High cost/ Low cost/ Very luxurious
  - Cuisines from the list of available cuisines
  - Services from the list of major services available
  - Should have featured - to select only those restaurants that have featured in Hyderabad's most popular places list

Recommendation of restaurants for customers

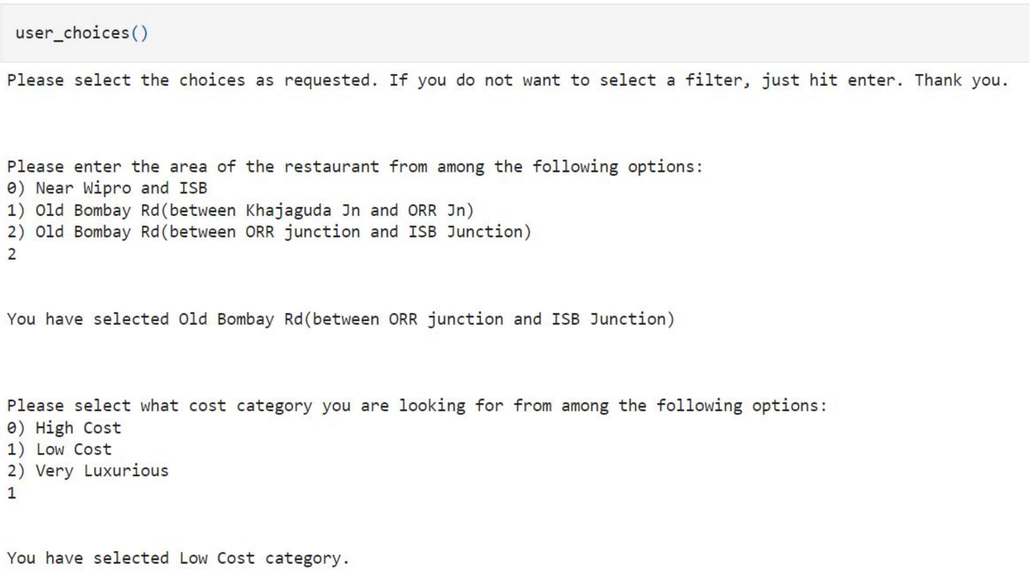


Number of restaurants in each combination of geographical cluster and cost benefit cluster

- As expected, 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.

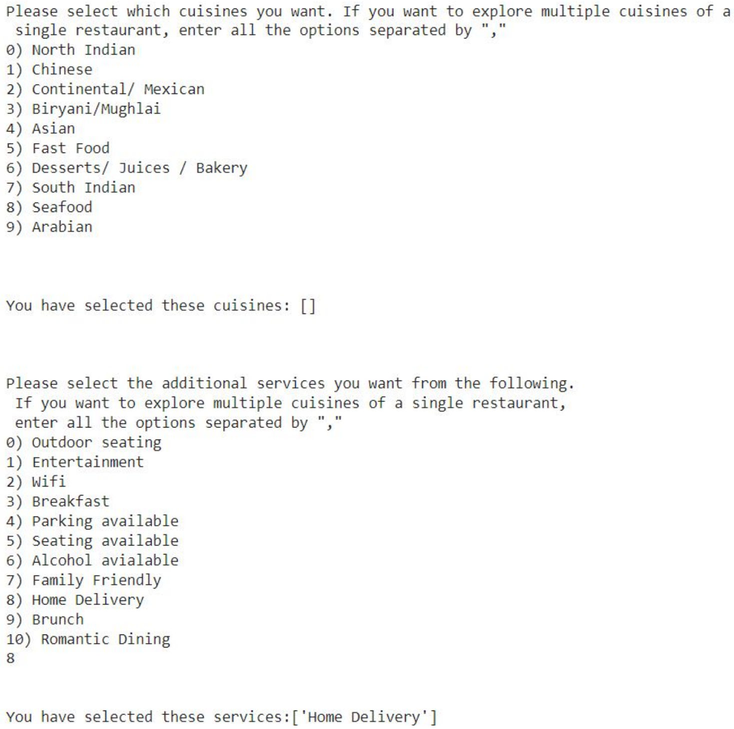
Recommendation of restaurants for customers

Demonstration of Customer Recommendation System



```
user_choices()
```

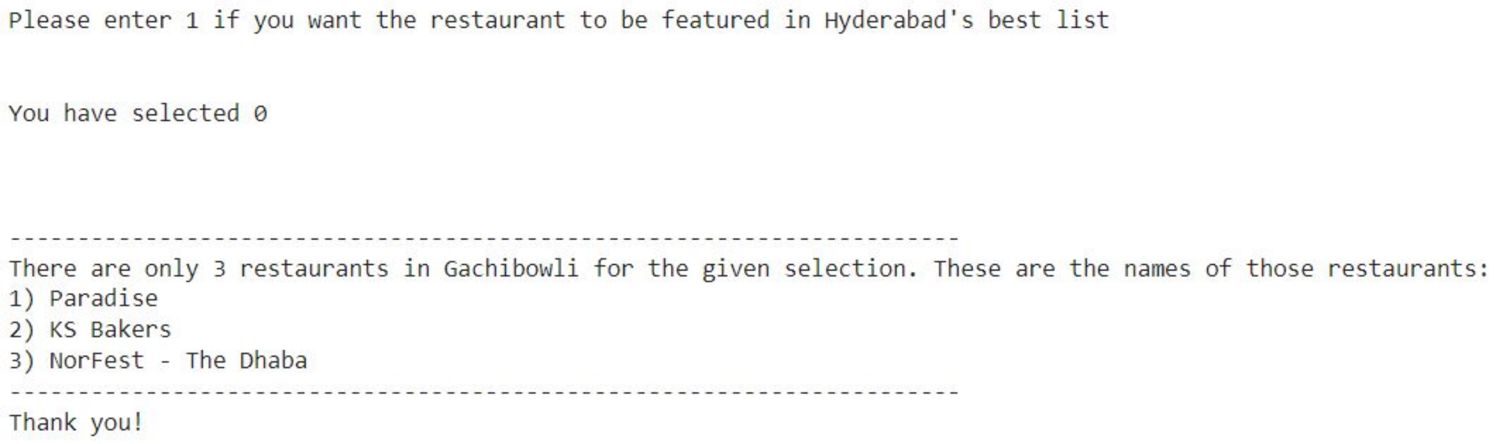
```
Please select the choices as requested. If you do not want to select a filter, just hit enter. Thank you.



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)
2


You have selected Old Bombay Rd(between ORR junction and ISB Junction)



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


You have selected Low Cost 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



You have selected these cuisines: []



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
8


You have selected these services:['Home Delivery']
```

Recommendation of restaurants for customers

Demonstration of Customer Recommendation System



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


You have selected 0



----------------------------------------------------------------------
There are only 3 restaurants in Gachibowli for the given selection. These are the names of those restaurants:
1) Paradise
2) KS Bakers
3) NorFest - The Dhaba
----------------------------------------------------------------------
Thank you!
```

Conclusion

- Through this project we have demonstrated our ability to process and explore an unlabelled 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 and we learned to use multiple libraries like numpy, pandas, sklearn, folium maps, matplotlib, seaborn, nltk.

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