Capstone project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

BY – Ganesh Walimbe

Points of discussions

- ❑ Problem statement
- ❑ Introduction to data
- ❑ Data preprocessing
- ❑ Exploratory data analysis
- ❑ Clustering
  - K means clustering
  - Hierarchical clustering
- ❑ Sentimental analysis
  - Support vector machine classifier
  - Logistic regression
  - Random forest classifier
- ❑ Conclusions

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. 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 analysing the Zomato restaurant data for each city in India.

The Project focuses on Customers and Company, you have 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. The data is vizualized as it becomes easy to analyse data at instant.

The Analysis also solve some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing 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.

Introduction to data

I have given two dataset for this project

1. Zomato Restaurant names and Metadata

I have used this dataset for clustering part and it contains 105 rows and 5 variables those variables are as follows :

- 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

2. Zomato Restaurant reviews

I have Merged this dataset with Names and Matadata and then i have used for sentiment analysis part and this dataset contains10000 rows and 7 variables those variables are :

- 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

Data preprocessing

Zomato Restaurant names and Metadata

- Missing value treatment:

1. There were missing values present in two variables of Zomato Restaurant names and Metadata dataset.
2. Those variables are collections with almost 50% values missing and timings has only 1 value is missing
3. So i dropped the collections variable and after that i removed all the missing values.

Also

There were zero duplicates in names and Metadata dataset

```
#cheking for missing values for names metadata data
data.isnull().sum()
```

```
Name           0
Links          0
Cost           0
Collections    54
Cuisines       0
Timings        1
```

Data preprocessing

Zomato Restaurant reviews

- Missing values treatment :

1. There were missing values present in five variables of Zomato Restaurant reviews dataset.
2. Those variables are Reviewer , Review , Rating, metadata and time with very small amount of missing values.
3. So i dropped all the missing values from the dataset.

Also

There were zero duplicates in Zomato Restaurant reviews dataset.

```
#cheking for missing values present in
#Zomato Restaurant reviews datset
data1.isnull().sum()
```

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

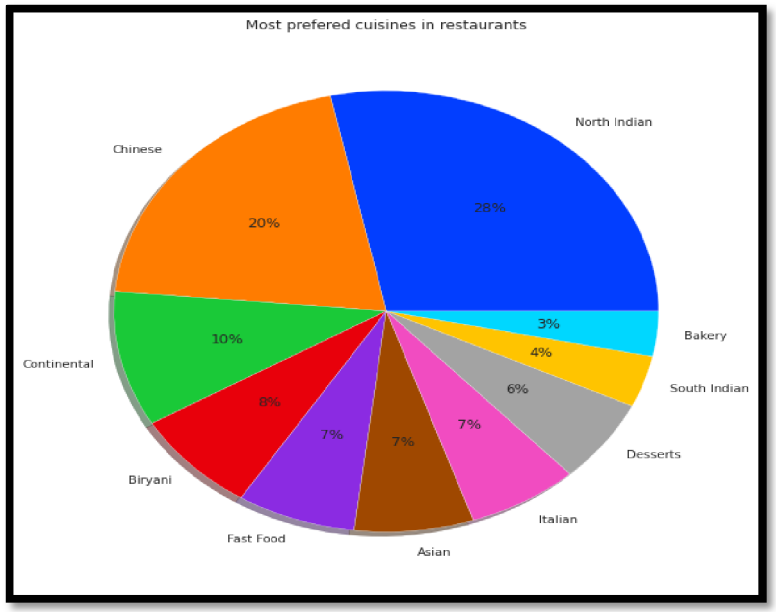
Exploratory data analysis

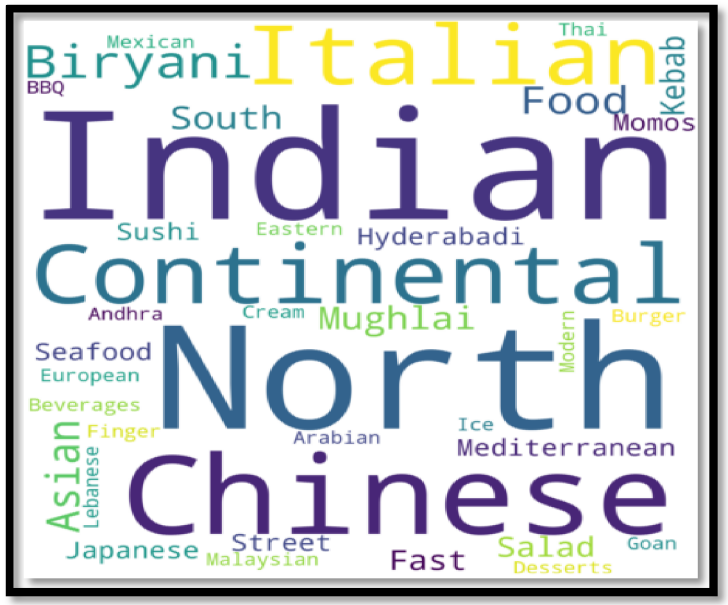
Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Objectives of EDA :

- Most preferred cuisines in restaurants
- Highest and lowest spending by customers in restaurants
- Highest rated restaurants by customers
- Busiest months for restaurants
- Ratings and reviews

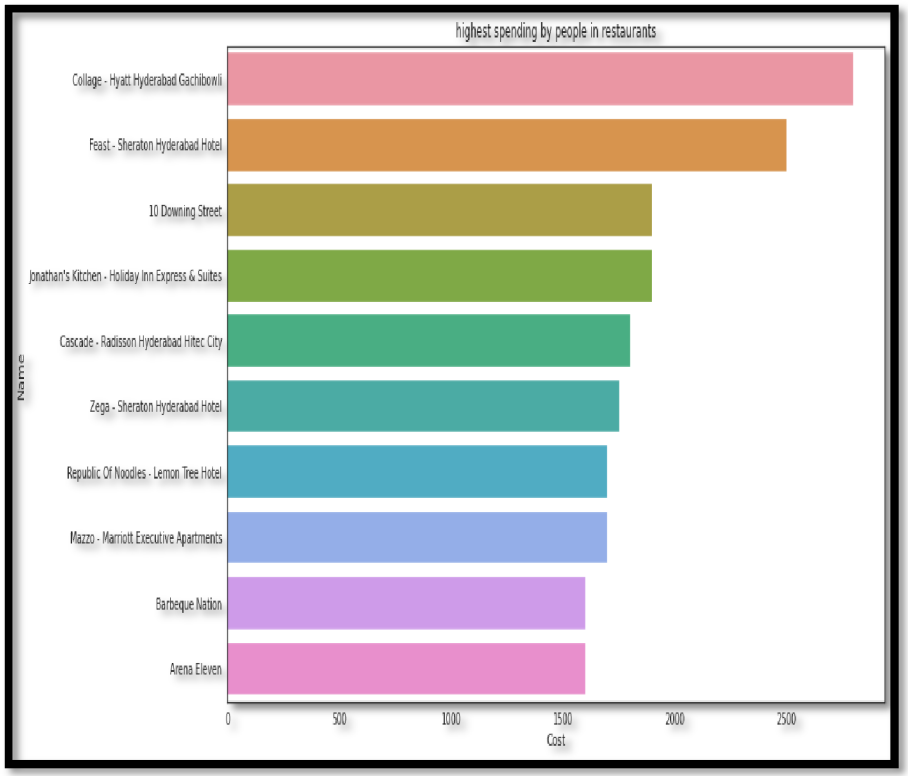
Most preferred cuisines in restaurants





From above pie chart and word cloud we can say that the most preferred cuisines in restaurants is North Indian, Chinese food, continental ,Biryani and fast food etc.

Highest spending by customers in restaurants

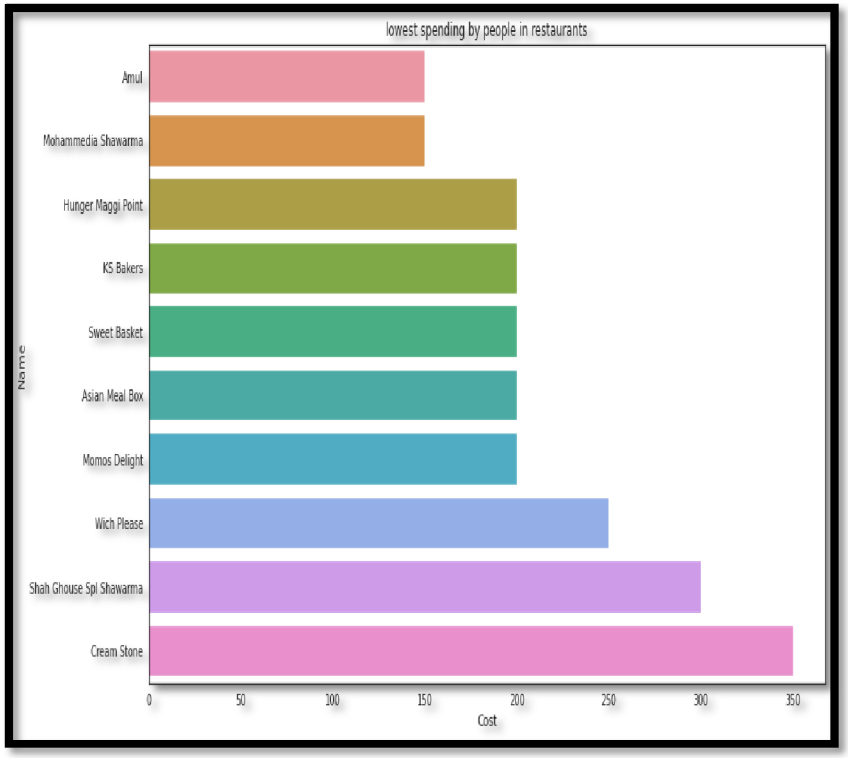


From barplot we can see that , Customers spending is highest at restaurants named :

1. Collage-Hyatt Hyderabad Ghachibowli
2. Feast Sheraton Hyderabad hotel
3. 10 Downing street
4. Jonathan’s kitchen-Holiday inn express and suites
5. Cascade –Radisson Hyderabad Hitec city

Customers must have gone to these restaurants for lunch ,dinner or for to party

Lowest spending by customers in restaurants

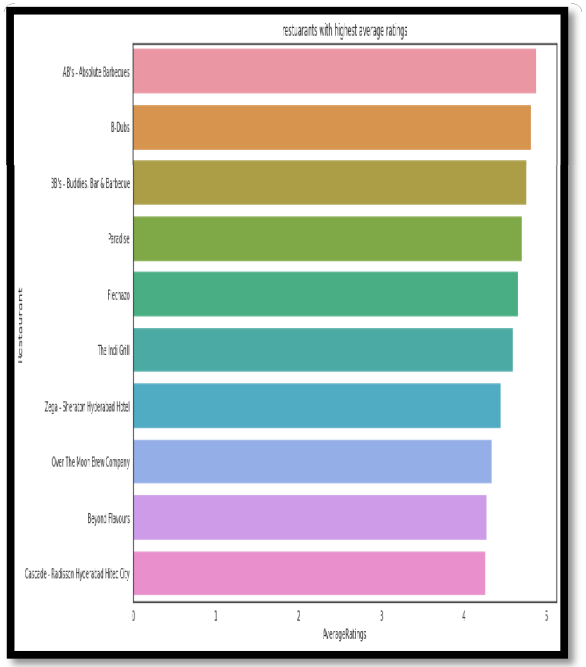


From barplot we can see that, Customers spending is lowest at restaurants named :

1. Amul
2. Mohammedia shawarma
3. Hunger maggi point
4. Ks bakers
5. Sweet baskets

Customers must have gone to these restaurants to eat breakfast, fast foods ,ice creams and to drink cold drinks.

Highest rated restaurants by customers

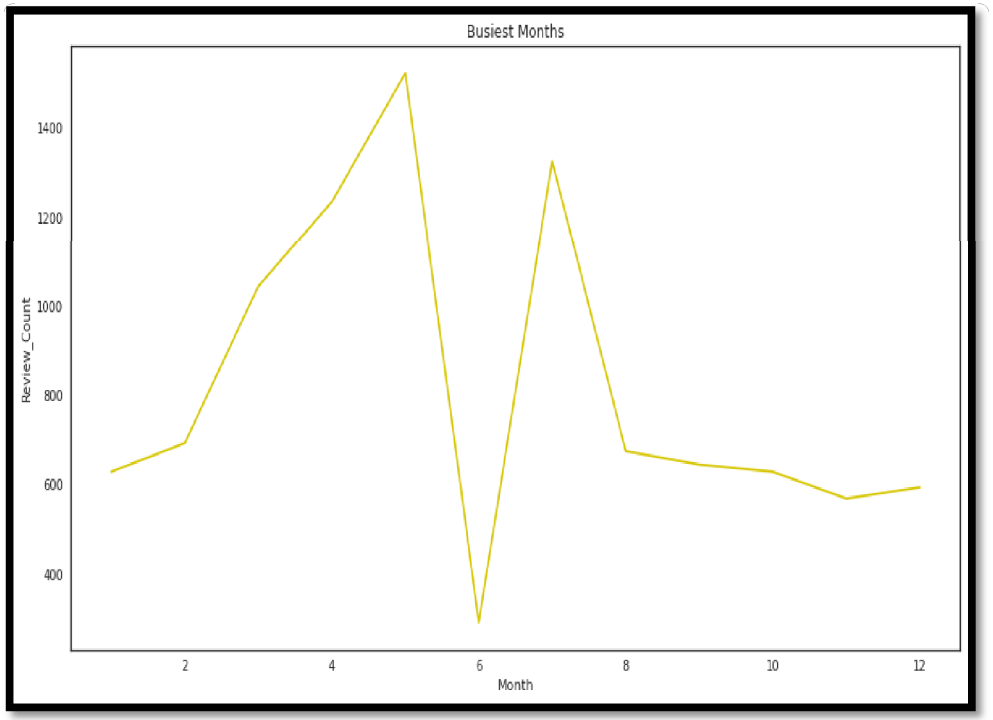


Restaurant	AverageRatings
AB's - Absolute Barbecues	4.88
B-Dubs	4.81
3B's - Buddies, Bar & Barbecue	4.76
Paradise	4.70
Flechazo	4.66
The Indi Grill	4.60
Zega - Sheraton Hyderabad Hotel	4.45
Over The Moon Brew Company	4.34
Beyond Flavours	4.28
Cascade - Radisson Hyderabad Hitec City	4.26

The highest rated restaurants according to average rating given to the restaurants by customers are as follows :

1. AB's –Absolute barbecues
2. B –Dubs
3. 3B's-Buddies,Bar & barbecues
4. Paradise
5. Flechazo

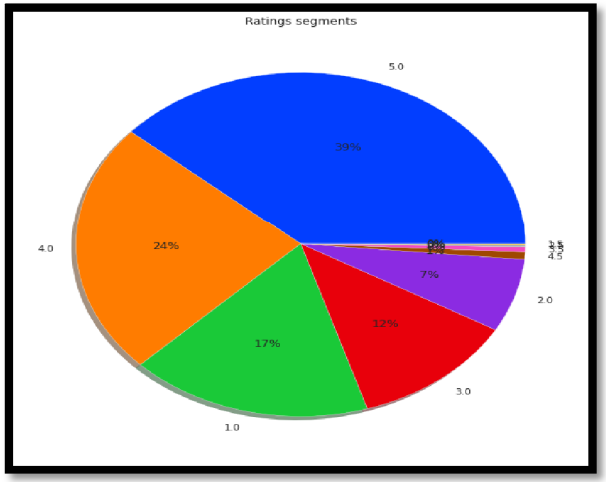
Busiest months for restaurants



From this lineplot we can see that ,

- The restaurants are busiest in the month of April ,May and July According to the reviewer count.
- And there are less customers in month of January, November and December.

Ratings and reviews



From above pie chart we can see that
39% of customers has given 5 stars rating to the and 24% of customer has given 4 stars and 17% of customers has given 1 stars to restaurants

From this word cloud we can see that ,customers are using words likes good ,place ,food ,service ,taste ,etc to give reviews to the restaurants

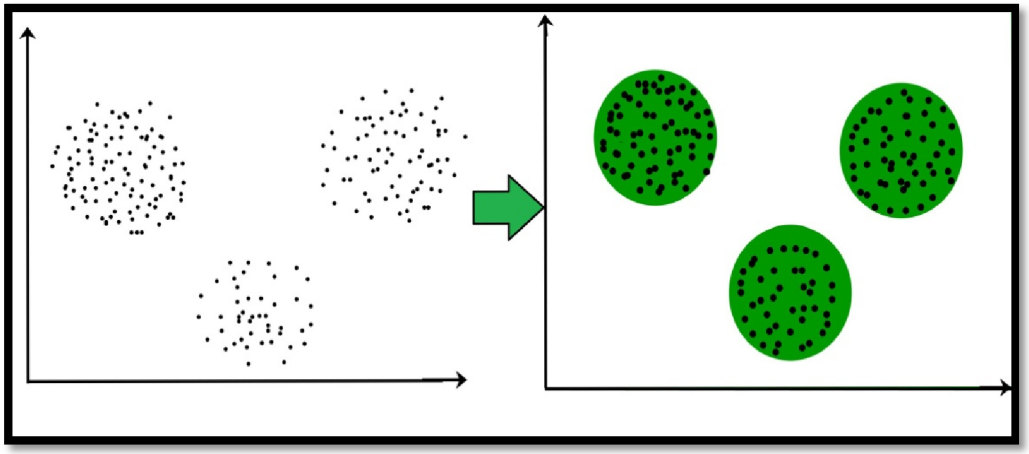
Clustering

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

I have used two unsupervised learning models for clustering those models are :

1.K means clustering

2.Hierarchical clustering



K means clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters.

Steps in k means clustering

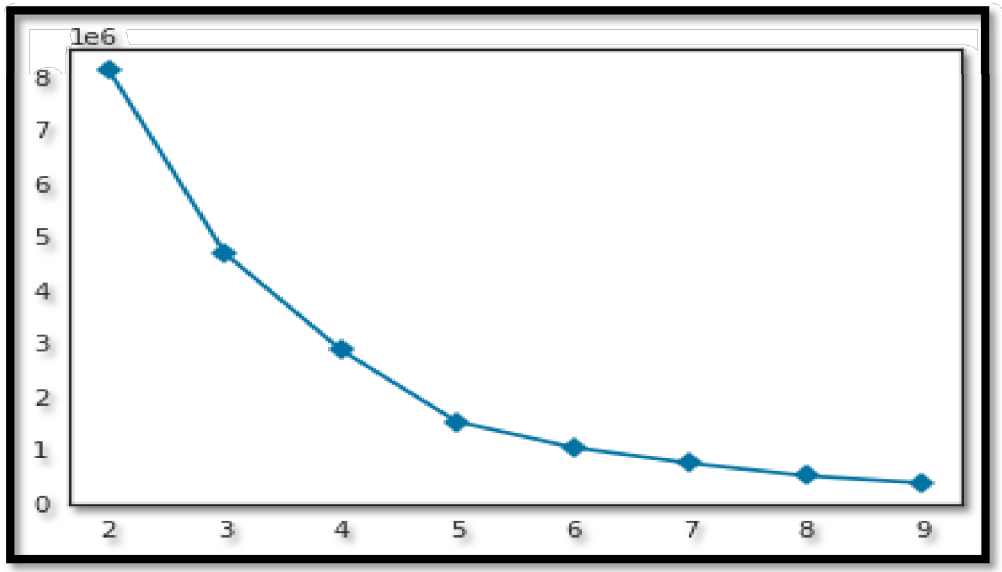
- Selection of k:

For selection of k i used elbow method.

Value of k =5

Fitting of the model

- Getting clusters from model

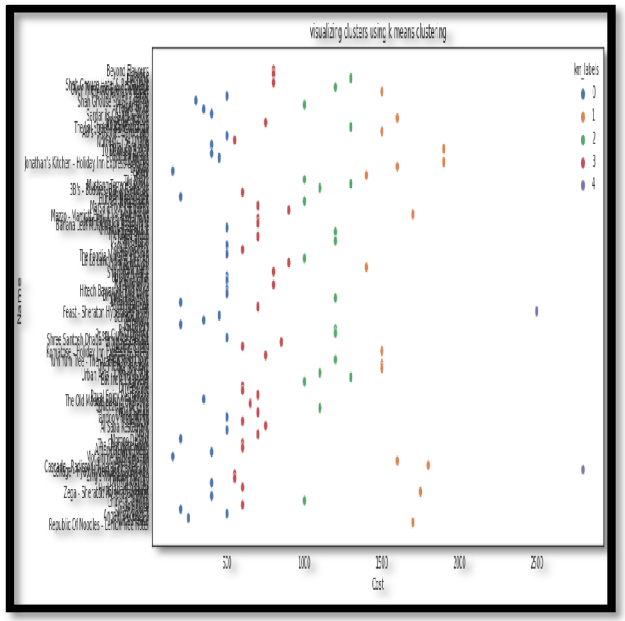


K means clustering

After applying k means on the dataset we can see that from the scatter that restaurants are divided into 5 clusters using k means clustering Based on the cuisines.

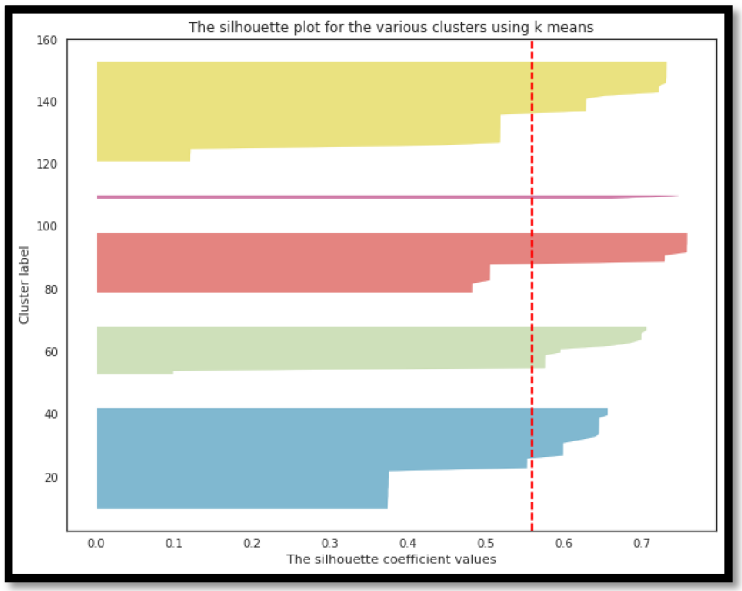
From all the clusters the we can see that North Indian ,Chinese cuisines are present in almost each cluster. so we can say that theses two cuisines are most preferred by customers.

```
# Top cuisines in each cluster according to k means
for i,df1 in enumerate(clusters_list):
  print(f'Top cuisines in cluster {i+1}\n', df1.dro

Top cuisines in cluster 1
 NorthIndian    16
Chinese         10
FastFood         8
Desserts         6
Biryani          5
dtype: int64

Top cuisines in cluster 2
 NorthIndian    11
Asian            6
Continental      6
Chinese          5
Italian          4
dtype: int64

Top cuisines in cluster 3
 NorthIndian    14
Chinese          9
Italian          7
Continental      7
Asian            4
dtype: int64

Top cuisines in cluster 4
 Asian           2
Italian          2
Continental      2
ModernIndian     1
Chinese          1
dtype: int64

Top cuisines in cluster 5
 NorthIndian    18
Chinese         18
Biryani         11
FastFood         7
Cafe             6
```



K means clustering

```
# Calculate Silhoutte Score
score = silhouette_score(x, kmeans.labels_, metric='euclidean')
print('Silhouette Score of k means: %.3f' % score)

Silhouetter Score of k means: 0.559
```



The silhouette score of k means clustering is 0.559 .

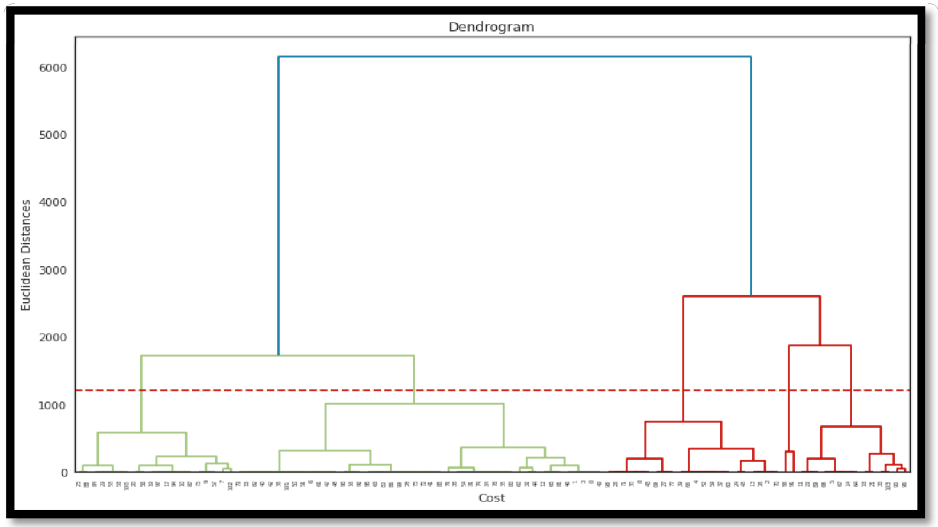
And from silhouette plot we can see that clusters are separated from each other

Hierarchical clustering

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis.

Agglomerative: Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

I have used the dendogram to find the optimal number of clusters for Hierarchical clustering From dendogram we can see that the optimal number of clusters are 5 clusters.



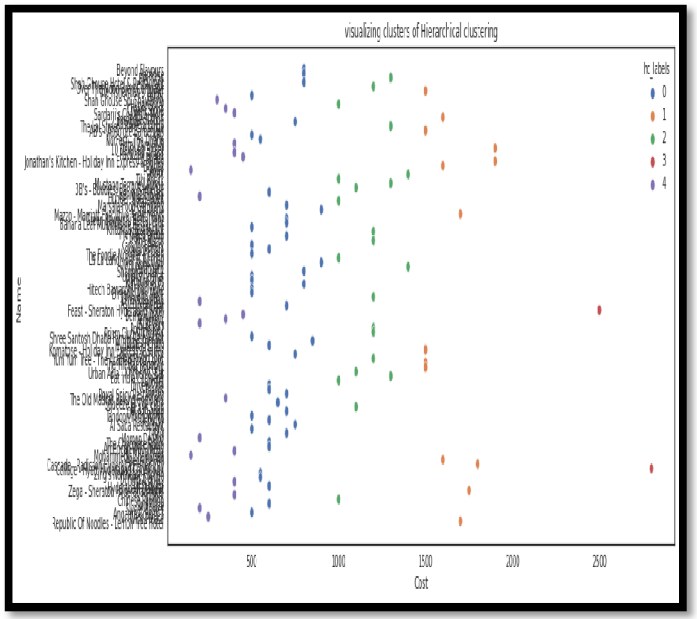
Hierarchical clustering

After applying hierarchical clustering on dataset we can see from scatter plot see that restaurants are divided into 5 clusters using Hierarchical clustering Based on the cuisines.

From all the clusters the we can see that North Indian ,Chinese cuisines are present in almost each cluster. so we can say that theses two cuisines are most preferred by customers.

The silhouette score of hierarchical clustering is 0.553 .

```
# Top cuisines in each cluster for Hierarchical clustering
for i,df2 in enumerate(clusters_list1):
  print(f'Top cuisines in cluster {i+1}\n', df2.drop(['inde

Top cuisines in cluster 1
 NorthIndian    27
Chinese         24
Biryani         15
FastFood        10
Desserts         7
dtype: int64

Top cuisines in cluster 2
 NorthIndian    9
Asian           6
Continental     5
Italian         4
Sushi           3
dtype: int64

Top cuisines in cluster 3
 NorthIndian    16
Chinese         11
Continental      8
Italian          7
Asian            4
dtype: int64

Top cuisines in cluster 4
 Asian           2
Italian          2
Continental      2
ModernIndian     1
Chinese          1
dtype: int64

Top cuisines in cluster 5
 NorthIndian    7
FastFood        5
Desserts        4
Chinese         4
Bakery          3
```



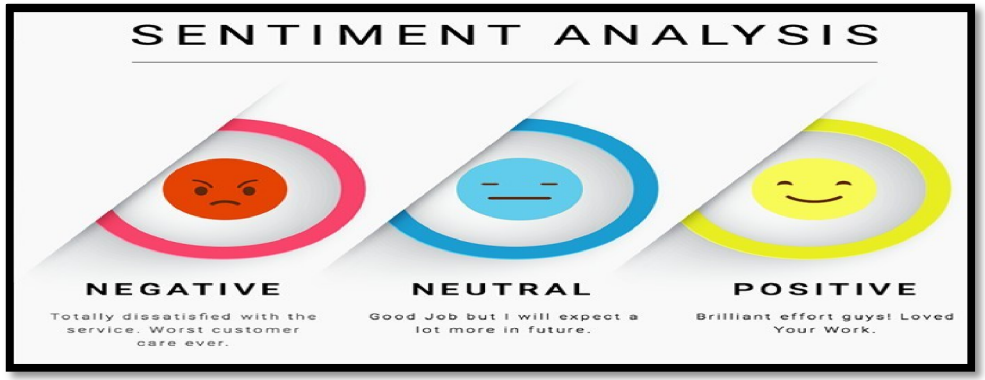

```
# Calculate Silhoutte Score
score = silhouette_score(x1, hc.labels_, metric='euclidean')
print('Silhouetter Score of hierachial clustering: %.3f' % score)

Silhouetter Score of hierachial clustering: 0.553
```

Sentimental analysis

Sentiment Analysis is the process of determining whether a piece of writing is positive, negative or neutral. A sentiment analysis system for text analysis combines natural language processing (NLP) and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

steps involved in sentimental analysis :

- Text preprocessing
- Assign a sentiment score to each phrase and component
- Performing sentimental analysis using machine learning models



Sentimental analysis

- Text preprocessing

1. Punctuation removal
2. Removing stopwords
3. Remove non letters

- Assign a sentiment score to each phrase and component

There are two types of sentiment score those are :

1. Polarity :Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement
2. Subjectivity: Subjective sentences generally refer to personal opinion, emotion or judgment whereas objective refers to factual information.

- Performing sentimental analysis using machine learning models

I have used 3 machine learning models for sentimental analysis those models are :

1. Support vector machine
2. Logistic regression
3. Random forest classifier

Support vector machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

- Before applying supervised learning models on the dataset i splited the independent variable and dependent variable into 80% for training and 20% for testing.
- SVM classifier is giving us 92% accuracy.
- Therefore SVM is performing very well on the dataset to predict the sentiments.

```
#Support vector machine classifier evaluation metrics
print(metrics.classification_report(y_test,predicted_result))
```

	precision	recall	f1-score	support
Negative	0.86	0.76	0.81	365
Neutral	0.83	0.72	0.77	94
Positive	0.94	0.97	0.95	1512
accuracy			0.92	1971
macro avg	0.88	0.82	0.85	1971
weighted avg	0.92	0.92	0.92	1971

Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables

- Logistic regression is given us 91% accuracy
- Therefore logistic regression is performing well on the dataset to predict the sentiments.

```
#evaluation metrics of logistic regression
print(metrics.classification_report(y_test,test_class_preds))
```

	precision	recall	f1-score	support
Negative	0.89	0.73	0.80	365
Neutral	0.84	0.38	0.53	94
Positive	0.91	0.98	0.95	1512
accuracy			0.91	1971
macro avg	0.88	0.70	0.76	1971
weighted avg	0.91	0.91	0.90	1971

Random forest classifier

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time

- Random forest classifier is given us 91% accuracy
- Therefore random forest classifier is performing well on the dataset to predict the sentiments
- I also performed hyperparameter tuning on random forest classifier.
- After hyperparameter tuning the accuracy did not improved.

```
#random forest classifier evaluation metrics
from sklearn import metrics
print(metrics.classification_report(test_preds_rf, y_test))
```

	precision	recall	f1-score	support
Negative	0.64	0.91	0.75	255
Neutral	0.79	0.77	0.78	96
Positive	0.98	0.91	0.95	1620
accuracy			0.91	1971
macro avg	0.80	0.87	0.83	1971
weighted avg	0.93	0.91	0.91	1971

```
#evaluation metrics of random forest classifier after hyperparameter tuning
print(metrics.classification_report(test_preds_ht_rf, y_test))
```

	precision	recall	f1-score	support
Negative	0.01	1.00	0.03	5
Neutral	0.00	0.00	0.00	0
Positive	1.00	0.77	0.87	1966
accuracy			0.77	1971
macro avg	0.34	0.59	0.30	1971
weighted avg	1.00	0.77	0.87	1971

Conclusions

❖ Conclusions for EDA :

- The most preferred cuisines in restaurants is North Indian, Chinese food, continental ,Biryani and fast food etc.
- Customers spending is highest at restaurants named Collage-Hyatt Hyderabad Ghachibowli,Feast Sheraton Hyderabad hotel and 10 Downing street etc.
- Customers spending is lowest at restaurants named Amul ,Mohammedia shawarma and Hunger maggi point ,etc.
- The highest rated restaurants are AB's –Absolute barbecues,B –Dubs and 3B's-Buddies,Bar & barbecues.
- The restaurants are busiest in the month of April ,May and July According to the reviewer count.And there are less number of customer in month of January, November and December.
- Customers are using words likes good ,place ,food ,service ,taste ,etc to give reviews to the restaurants.

Conclusions

❖ Conclusions from clustering

- k means clustering is giving us 5 clusters to divide restaurants on the basis of cuisines .k means is giving silhouette score of 0.559
- The north indian cuisine and chinses cuisine is present in all clusters of k means clusters. so we can say that theses two cuisines are most preferred by customers in all restaurants.
- Hierarchical clustering is also giving us 5 clusters to divide restaurants on the basis of cuisines .hierarchical clustering is giving silhouette score of 0.553
- The north Indian cuisine and Chinese cuisine is present in all clusters of hierarchical clusters. so we can say that theses two cuisines are most preferred by customers in all restaurants.

Conclusions

❖ Conclusions from sentimental analysis

I used three machine learning models for sentimental analysis those models are SVM classifier , logistic regression and random forest model.

- Logistic regression and random forest model are both giving us accuracy of 91%.
- Support vector machine classifier is giving highest among all the model i.e. 92%.Therefore we can use SVM for the predictions of sentiments

THANK YOU