Capstone Project-4

Zomato Restaurant Clustering and Sentiments Analysis

Team members

KAMALUDDIN SHAIKH
AMOL RASAM
SHUBHAM JHA
PRETESH

Content

- Problem Statement
- Data Summary

Clustering

- EDA
- Optimal No of Clusters
- Restaurants in Different Clusters
- Cuisines in Different Clusters

Sentiment Analysis

- EDA
- Data Preprocessing
- Top Words before and after Stemming
- ML Model & Metrics
- Challenges
- Conclusion

Problem Statement

- The Project focuses on Customers and Company, we have tried 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 visualized as it becomes easy to analyses data at instant.
- 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 Summary

For Clustering

- 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 Restaurants reviews

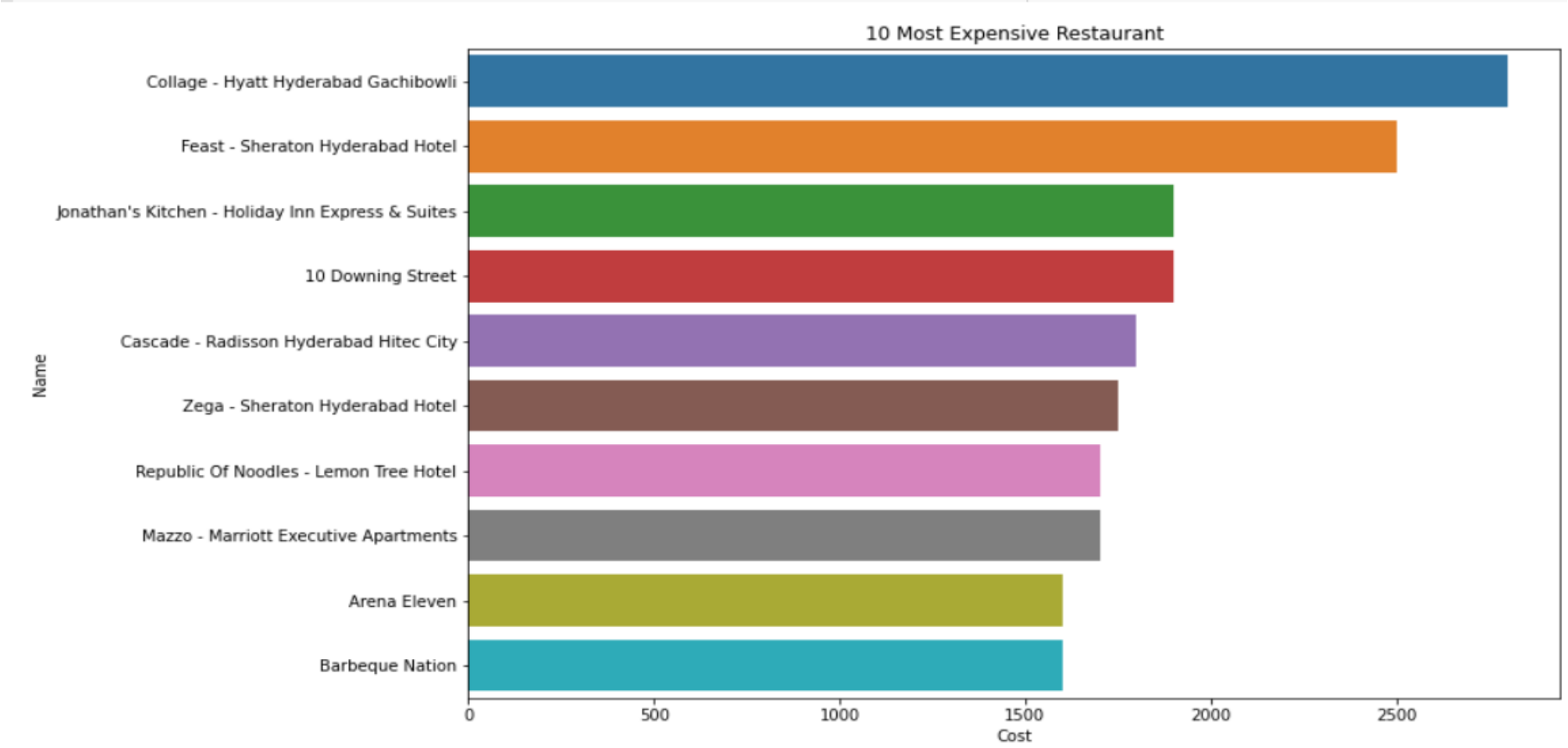
- Restaurant : Name of the Restaurant
- Reviewer : Name of the Reviewer
- Review : Review Text
- Rating : Rating Provided by Reviewer
- Metadata : Reviewer Metadata - No. of Reviews and followers
- Time: Date and Time of Review
- Pictures : No. of pictures posted with review

Basic Exploration

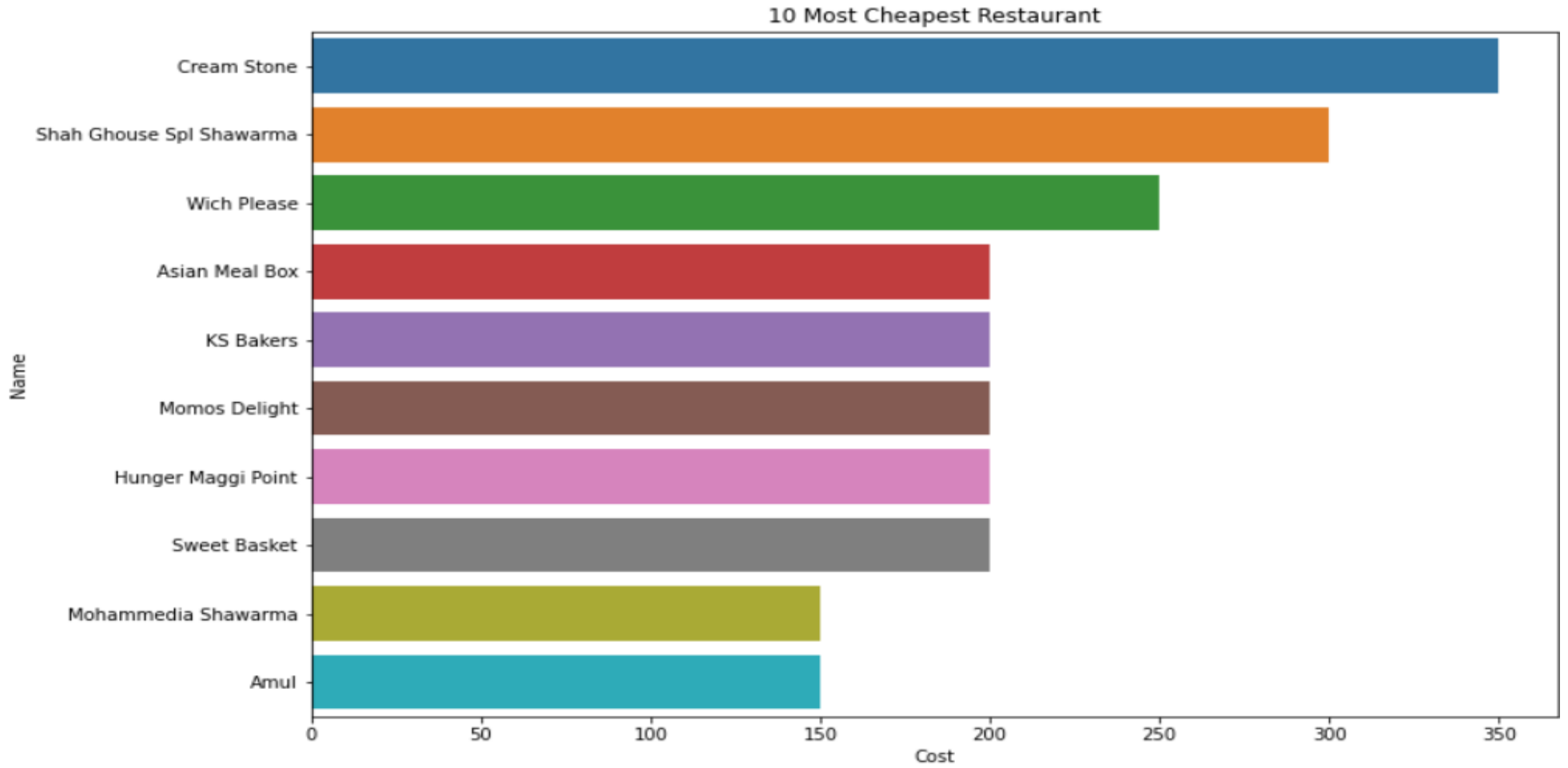
- Data of 105 restaurants.
- Data of 9000 reviews.
- 3 years of customer’s reviews.
- 0.36 percent null values were present.
- 50 percent of collection data is missing.
- Average price of a Restaurant ranges from 150 to 2800.

EDA- Clustering

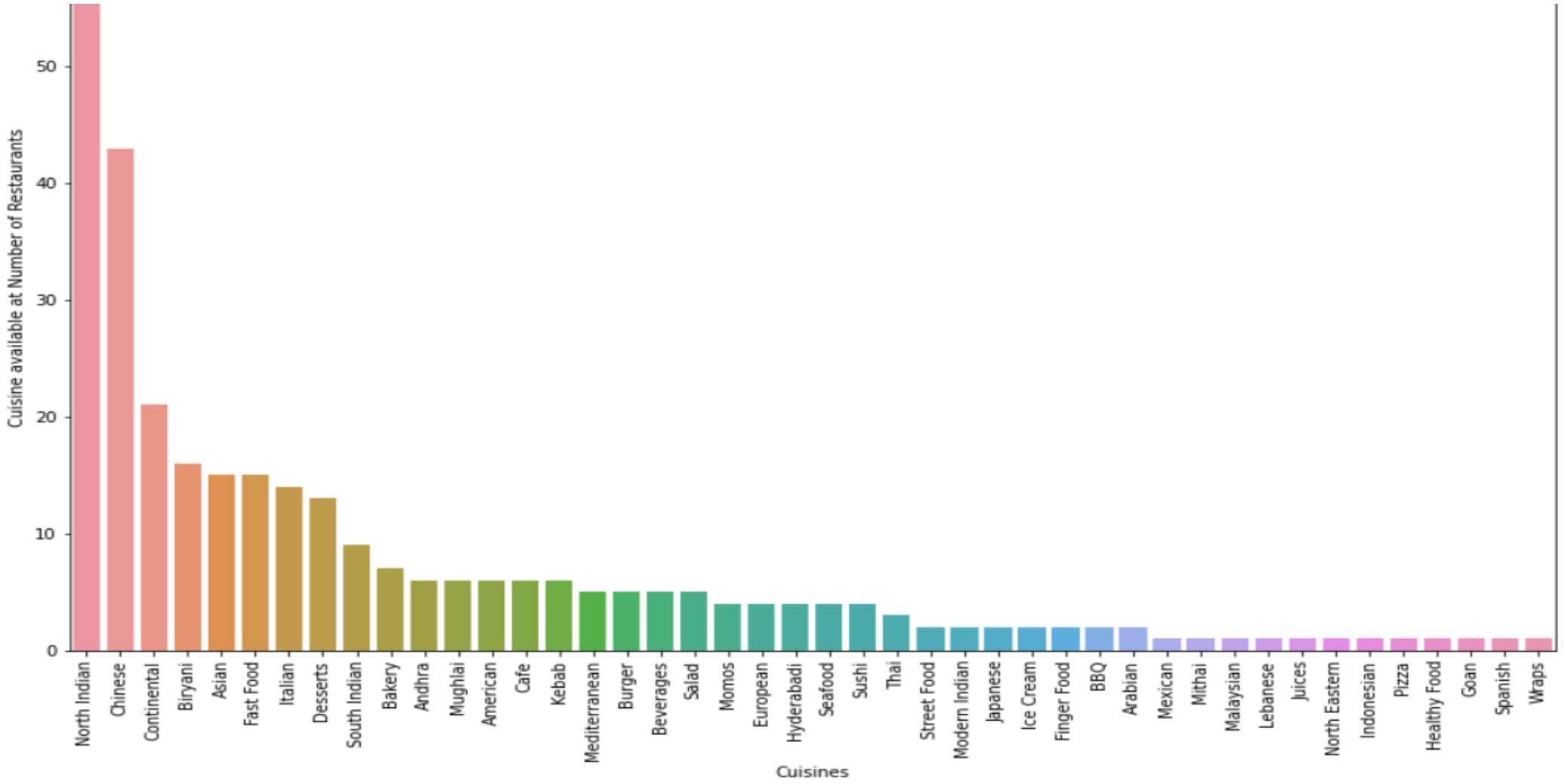
Top 10 Most Expensive Restaurant



Top 10 Budget Friendly Restaurants

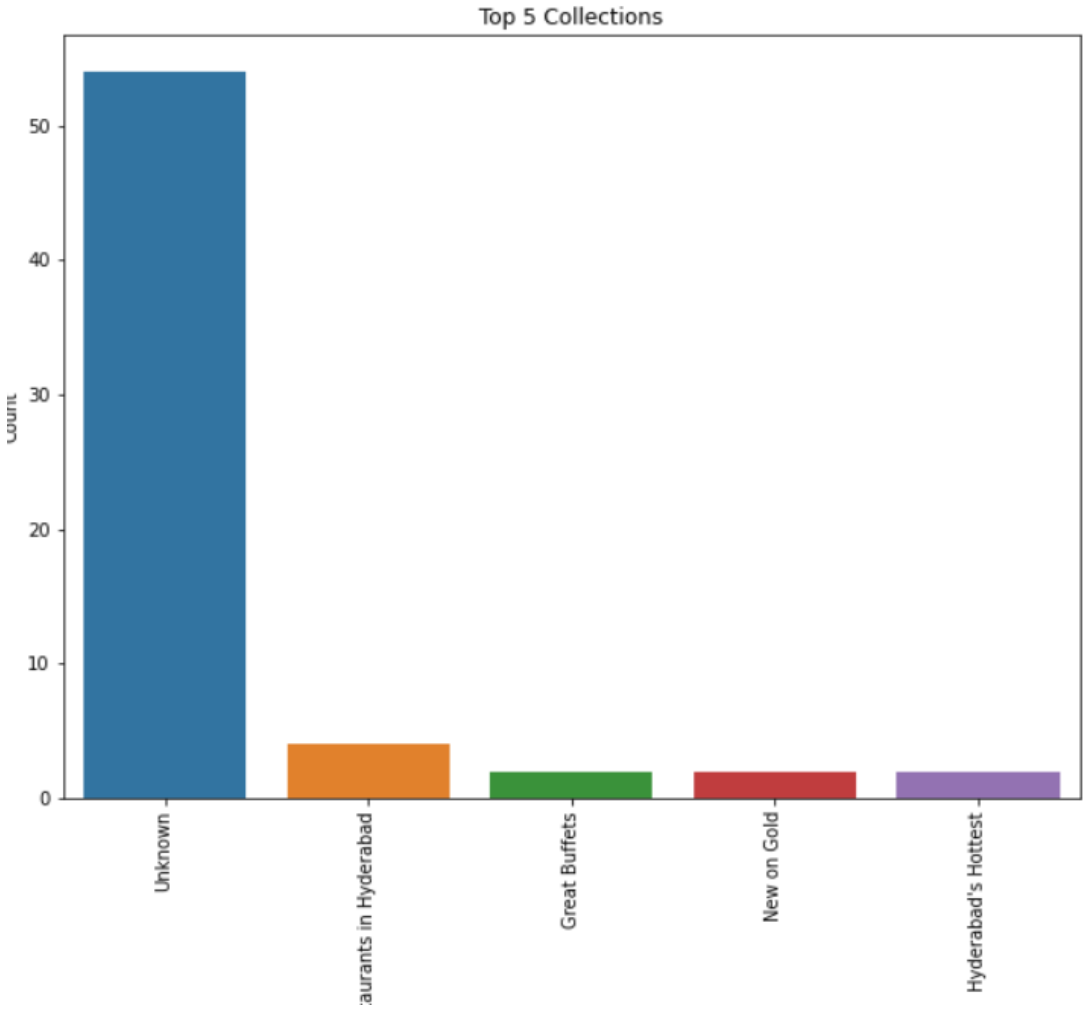


Most Popular Cuisines at Restaurants

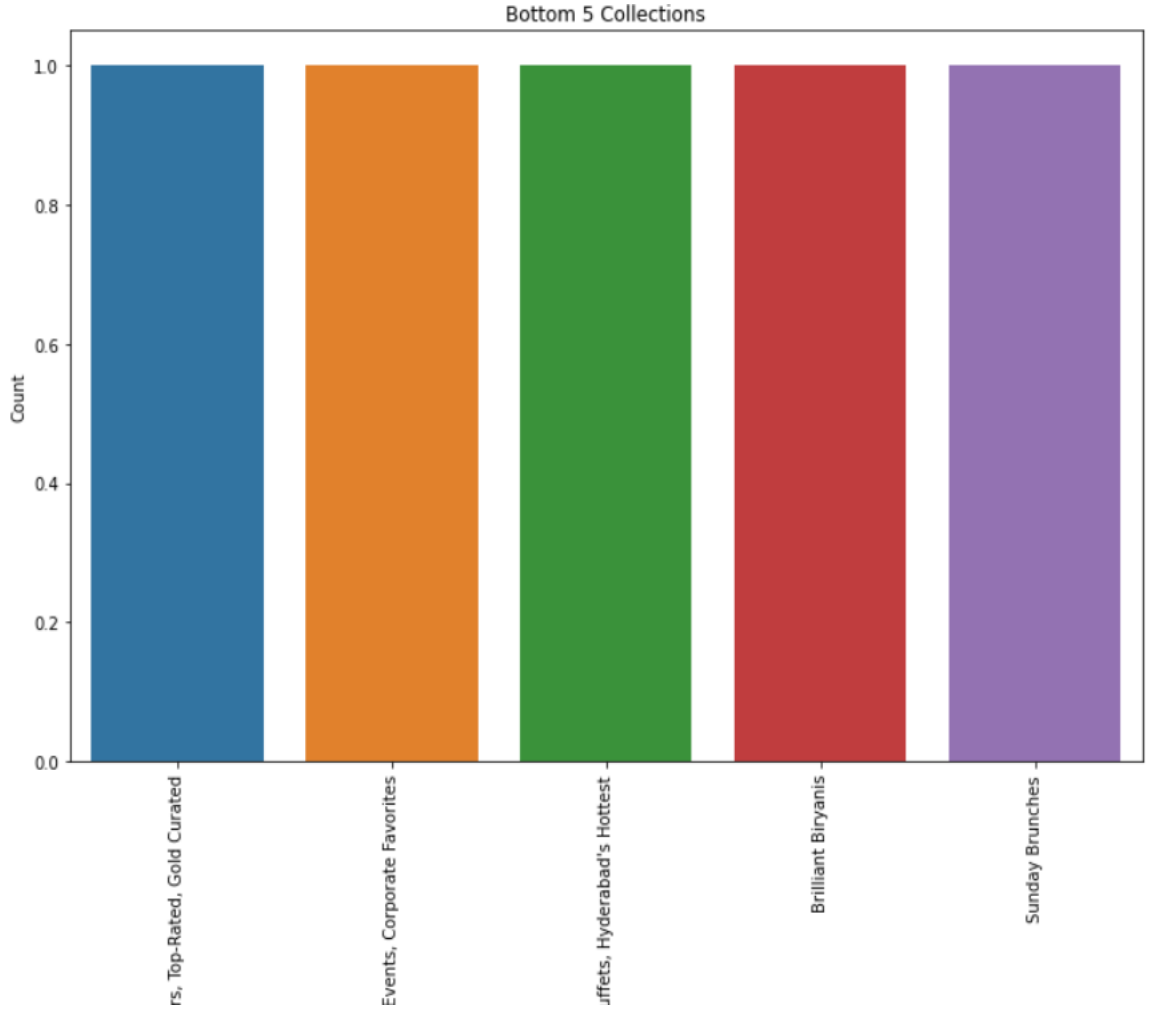


Insights From Collections

Top 5 Collections

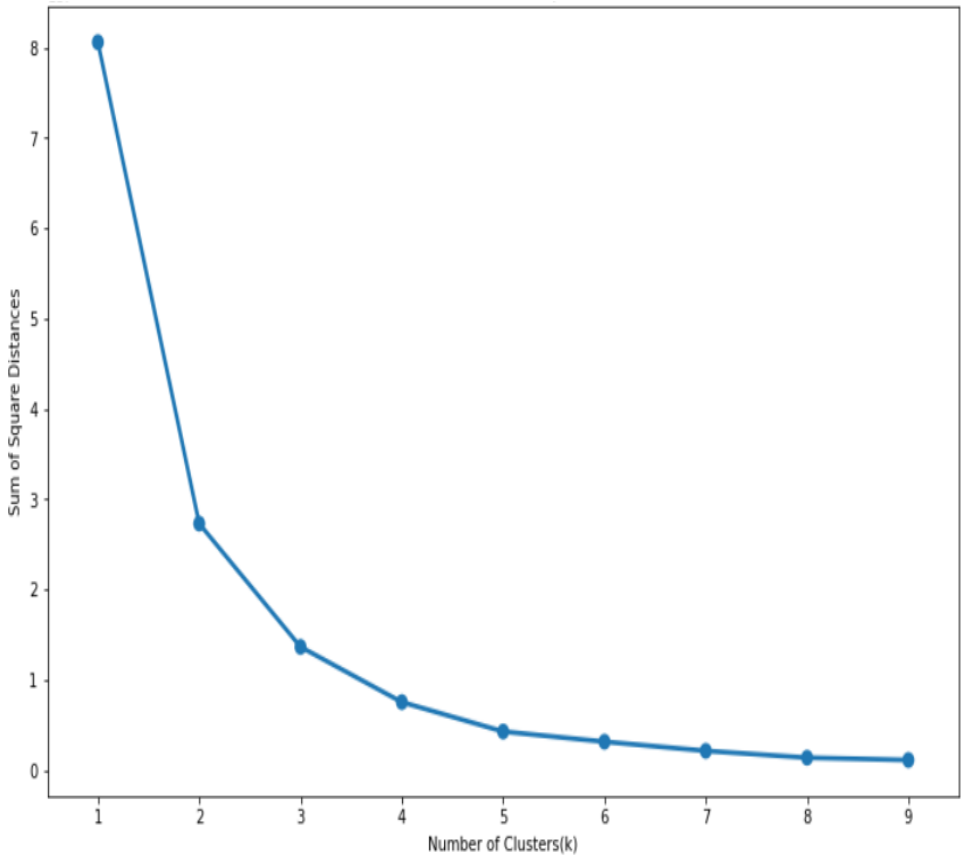


Bottom 5 Collections

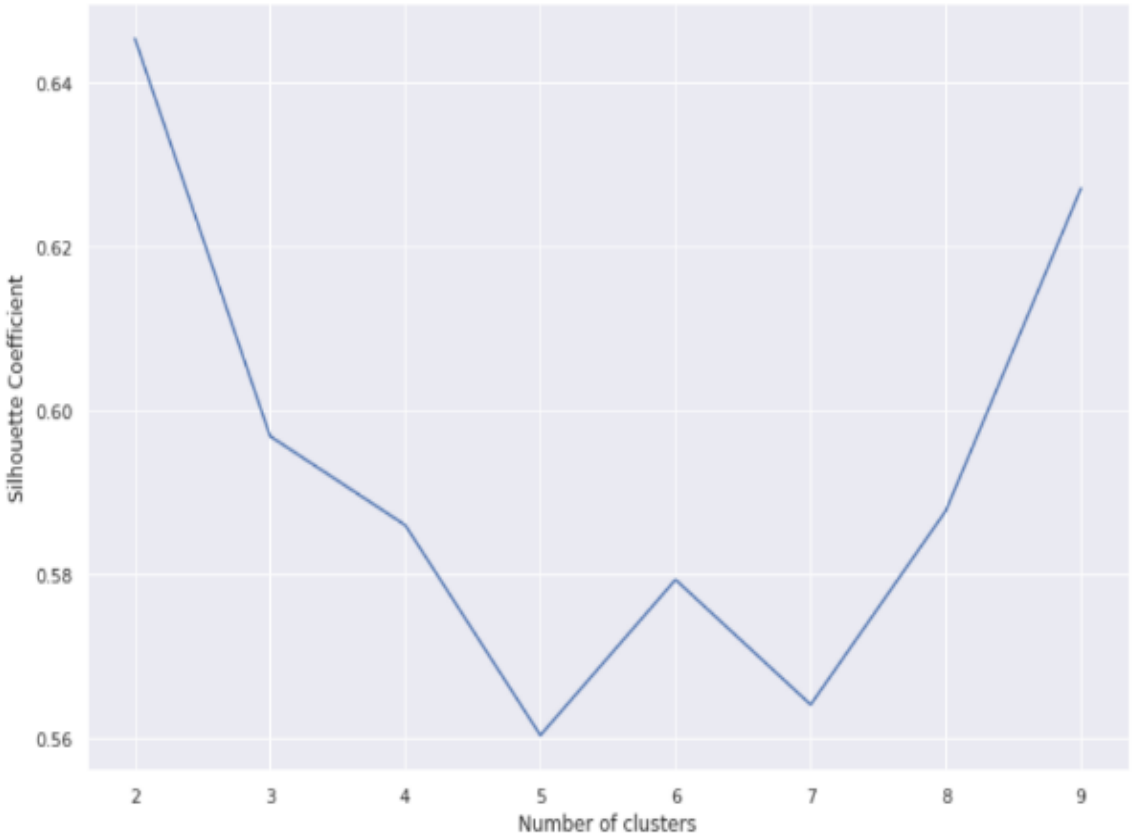


Optimal Number of Clusters

Elbow Method for optimal K



Silhouette Coefficient



Restaurants and Ratings in different Clusters

Cluster 1

```
cluster_1 = cluster_dataset.loc[cluster_dataset['labels']==0]
cluster_1[['Name','Rating','labels']].head()
```

	Name	Rating	labels
1	13 Dhaba	3.48	0
7	Amul	3.94	0
10	Asian Meal Box	2.58	0
15	Being Hungry	3.66	0
17	Biryanis And More	3.74	0

Cluster 2

```
cluster_1 = cluster_dataset.loc[cluster_dataset['labels']==0]
cluster_1[['Name','Rating','labels']].head()
```

	Name	Rating	labels
1	13 Dhaba	3.48	0
7	Amul	3.94	0
10	Asian Meal Box	2.58	0
15	Being Hungry	3.66	0
17	Biryanis And More	3.74	0

Restaurants and Ratings in different Clusters

Cluster 3

```
] cluster_3 = cluster_dataset.loc[cluster_dataset['labels']==2]
  cluster_3[['Name','Rating','labels']].head()
```

	Name	Rating	labels
4	Absolute Sizzlers	3.620000	2
5	Al Saba Restaurant	3.155000	2
6	American Wild Wings	3.974026	2
9	Aromas@11SIX	3.460000	2
12	Banana Leaf Multicuisine Restaurant	3.690000	2

Cluster 4

```
[ ] cluster_4 = cluster_dataset.loc[cluster_dataset['labels']==3]
    cluster_4[['Name','Rating','labels']].head()
```

	Name	Rating	labels
22	Collage - Hyatt Hyderabad Gachibowli	3.41	3
34	Feast - Sheraton Hyderabad Hotel	4.22	3

Cluster 5

```
[ ] cluster_5 = cluster_dataset.loc[cluster_dataset['labels']==4]
    cluster_5[['Name','Rating','labels']].head()
```

	Name	Rating	labels
2	3B's - Buddies, Bar & Barbecue	4.760	4
20	Chinese Pavilion	3.745	4
28	Diners Pavilion	3.320	4
32	Eat India Company	3.260	4
35	Flechazo	4.660	4

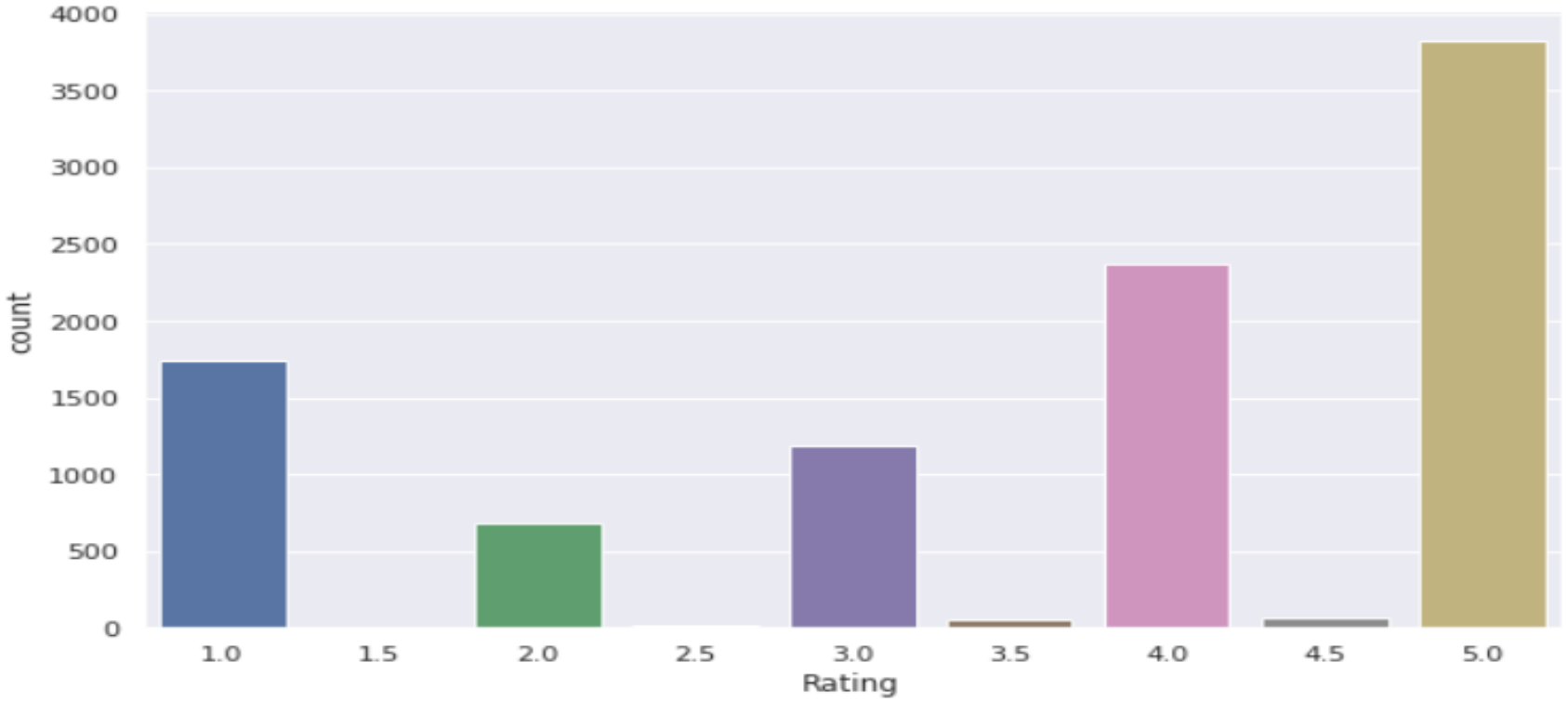
Cuisines in Clusters

```
cluster no 0
['north indian' 'ice cream' 'desserts' 'asian' 'chinese' 'biryani' 'pizza'
 'fast food' 'cafe' 'beverages' 'burger' 'american' 'bakery' 'andhra'
 'hyderabadi' 'south indian' 'arabian' 'street food' 'momos' 'lebanese'
 'kebab' 'wraps' 'north eastern' 'healthy food' 'continental']


cluster no 1
['chinese' 'north indian' 'continental' 'european' 'mediterranean'
 'american' 'bbq' 'kebab' 'italian' 'asian' 'salad' 'japanese' 'sushi'
 'south indian' 'seafood' 'goan' 'modern indian' 'andhra']


cluster no 2
['chinese' 'continental' 'american' 'north indian' 'seafood' 'biryani'
 'hyderabadi' 'salad' 'burger' 'fast food' 'mughlai' 'andhra'
 'south indian' 'kebab' 'european' 'cafe' 'bakery' 'desserts' 'momos'
 'asian' 'beverages' 'arabian' 'thai' 'indonesian' 'italian']


cluster no 3
['north indian' 'italian' 'chinese' 'continental' 'asian' 'modern indian']


cluster no 4
['european' 'north indian' 'mediterranean' 'chinese' 'seafood' 'italian'
 'continental' 'mughlai' 'beverages' 'asian' 'desserts' 'spanish'
 'american' 'kebab' 'south indian' 'finger food' 'bakery' 'salad'
 'mexican' 'juices' 'sushi' 'thai' 'momos' 'hyderabadi']
```

North Indian, Continental and Chinese are available mostly in every Cluster.

EDA- Sentiment Analysis

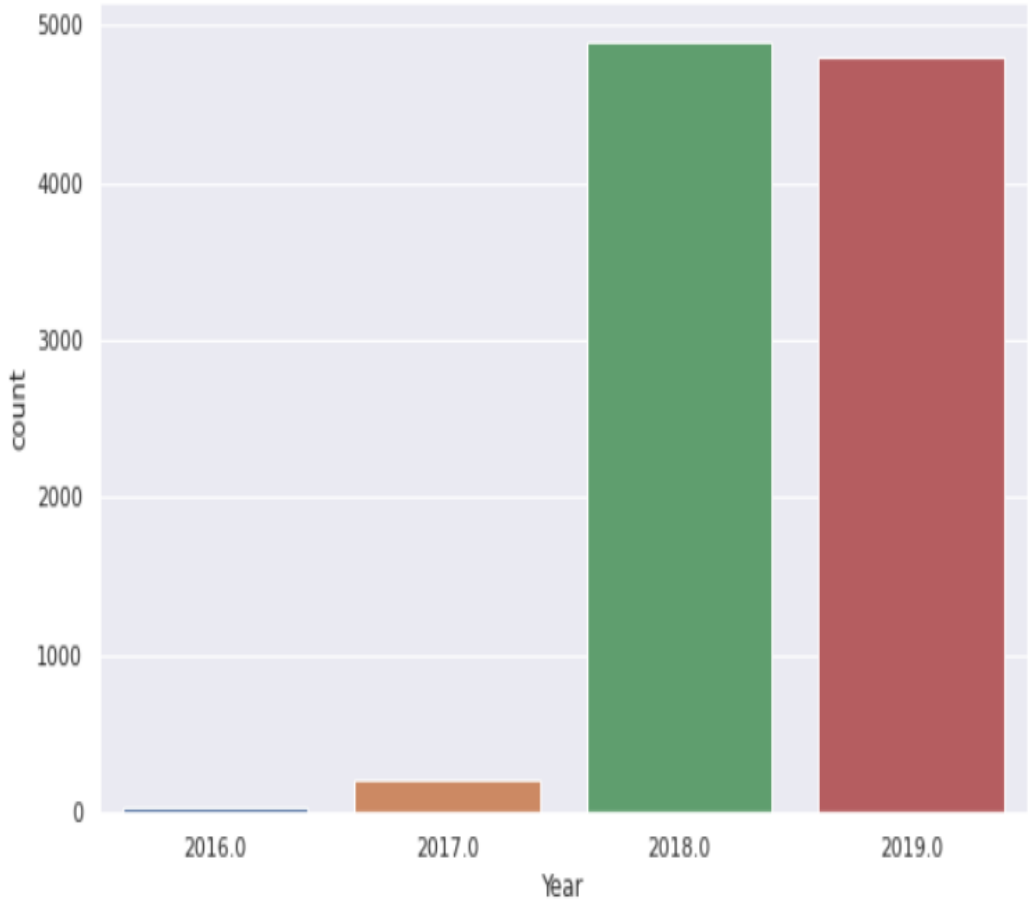
Ratings Analysis



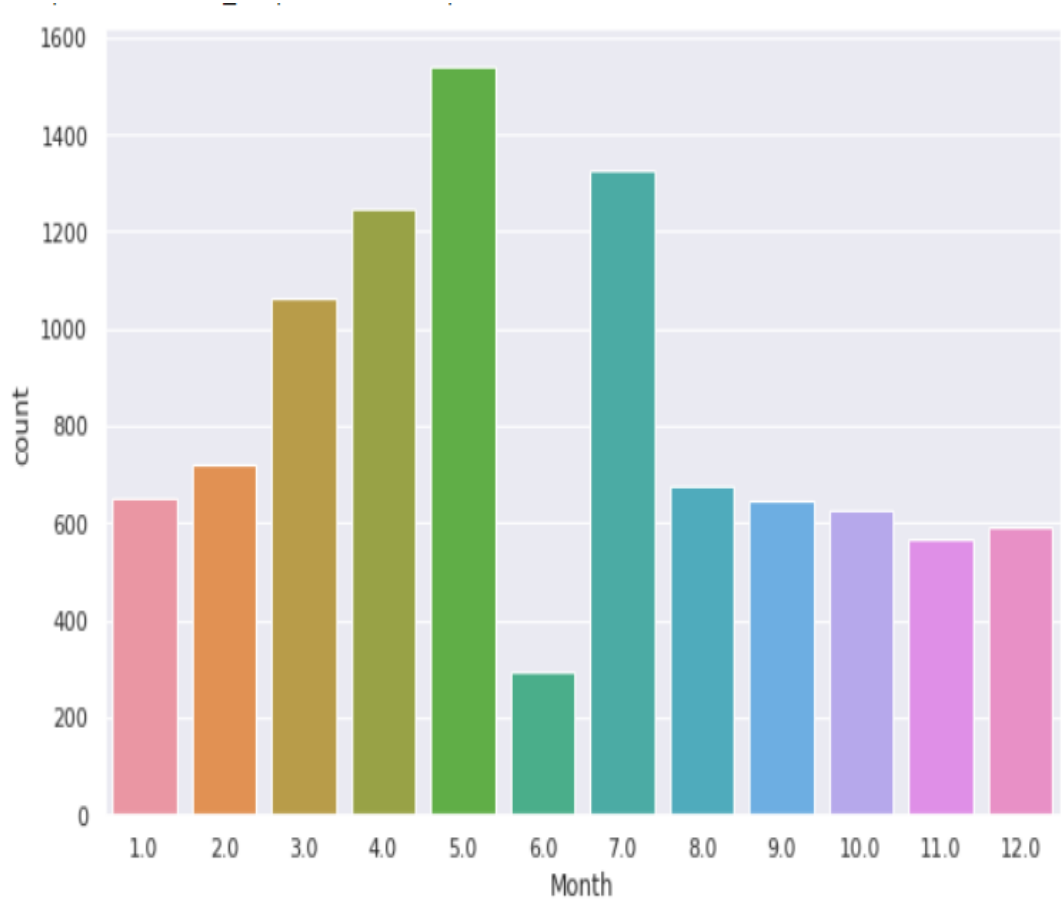
We can see most number of people game rating 5 followed by 4 and 1 respectively.

Reviews Analysis

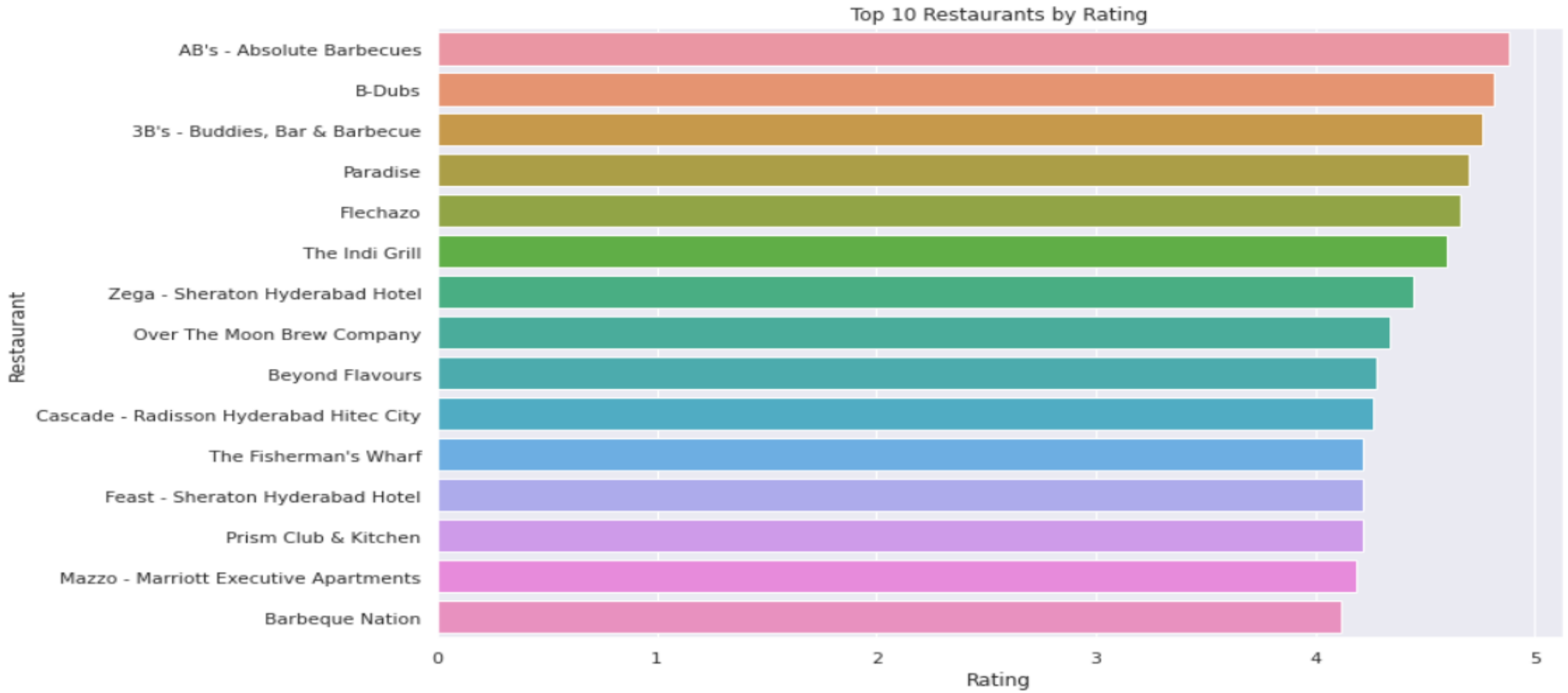
Yearly Review Analysis



Monthly Review Analysis

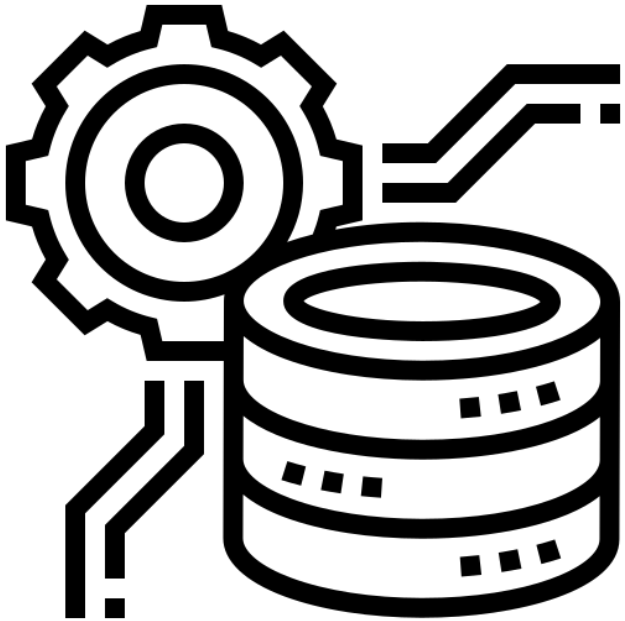


Top 10 Restaurants by Rating



Data Pre-Processing

- **Steps taken to prepare the data**
  - Removing Punctuations
  - Removing Stop words
  - Removing # symbols and retaining the tags
  - Removing Emoji’s
  - Removing short words
  - Tokenization and stemming



Stemming

Top words before Stemming


```
[('good', 7007),
 ('food', 6488),
 ('place', 5663),
 ('service', 3157),
 ('chicken', 3045),
 ('taste', 2214),
 ('ordered', 2163),
 ('ambience', 2044),
 ('great', 1936),
 ('one', 1885),
 ('really', 1659),
 ('time', 1581),
 ('also', 1577),
 ('nice', 1464),
 ('like', 1450),
 ('best', 1432),
 ('biryani', 1305),
 ('staff', 1295),
 ('try', 1276),
 ('visit', 1255)]
```

Stemming

Top words after Stemming


```
[('good', 7011),
 ('food', 6546),
 ('place', 6007),
 ('order', 3600),
 ('servic', 3246),
 ('chicken', 3046),
 ('tast', 2982),
 ('ambienc', 2045),
 ('time', 2040),
 ('tri', 2035),
 ('one', 1993),
 ('great', 1940),
 ('visit', 1875),
 ('like', 1815),
 ('realli', 1659),
 ('serv', 1591),
 ('also', 1577),
 ('nice', 1510),
 ('best', 1433),
 ('restaur', 1430)]
```

Comparison Of performance of all models

index	Model	Test accuracy	Precision	Recall	Auc - Roc Score
1	Logistic Regression	0.862404	0.769848	0.849409	0.859256
2	Random Forest	0.857717	0.706512	0.891892	0.867575
3	CatBoost Classifier	0.849012	0.700268	0.872222	0.855613
4	XGBoost	0.821895	0.597681	0.892144	0.845222
5	K Nearest Neighbour	0.753264	0.497770	0.762295	0.756314
6	Decision Tree	0.723468	0.518287	0.670127	0.707705

Challenges

- Feature engineering.
- Finding optimum number of Cluster
- Text preprocessing
- ML Model and Metrics

Conclusion

- That's it! We reached the end of our exercise. Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection, and then model building.
- For clustering, we have decided on 5 clusters after the Silhouette score plot and elbow plot where we used K means clustering algorithm.
- For Sentiment Analysis we have implemented six different models to predict the sentiment of Reviews. Logistic Regression, Random Forest Classifier, Decision Tree, CatBoost Classifier, K Nearest Neighbor and Xgboost Classifier.
- Logistic Regression and Random Forest model performed the best among them. In this way, we can explore more from various textual data and Reviews. Our models will try to predict the various sentiments correctly.

Thank You