Capstone Project

ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS

By

Dinesh A Wagh.

Points To Discuss

- Introduction
- Abstract
- Problem Statement
- Data Description
- Exploratory Data Analysis
- Models used
- Performance metrics for models
- Challenges
- Conclusion

Introduction

- 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 a given city in India. There are two sets of data. The first ones provides the data on the restaurants and the other is about the reviews for these restaurants. The main objective of the project is clustering of Zomato restaurants into different segments. The Project also focuses on analyzing the sentiments of the reviews given by the customer in the data and to make some useful conclusion in the form of visualizations. 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.

Abstract

- In today's digital world, food apps like Zomato are widely used because it provides a platform for people to share their opinion about the restaurants and cafes they have visited. This paper includes analysis of client ratings and reviews in Zomato utilizing content mining. Utilizing content mining, break down the content audits/reviews from the client with a specific end goal to create productive results and legit surveys. The rating has a review of the restaurant which can be used for sentiment analysis. Based on this, writers want to discuss the sentiment of the review to be predicted. The method used for preprocessing the review is to make all words lowercase, tokenization, remove numbers and punctuation, stop words, and lemmatization. Then after that, we create a word to vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 10,000 reviews. After that, we make positive reviews that have a rating of 3.5 and above, negative with reviews that have a rating of 3 and below. We have used Split Test, 80% Data Training and 20% Data Testing. The metrics used to determine random forest classifiers are precision, recall, accuracy, F1 score.

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 analyzing 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 visualized as it becomes easy to analyze 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.

Data Description

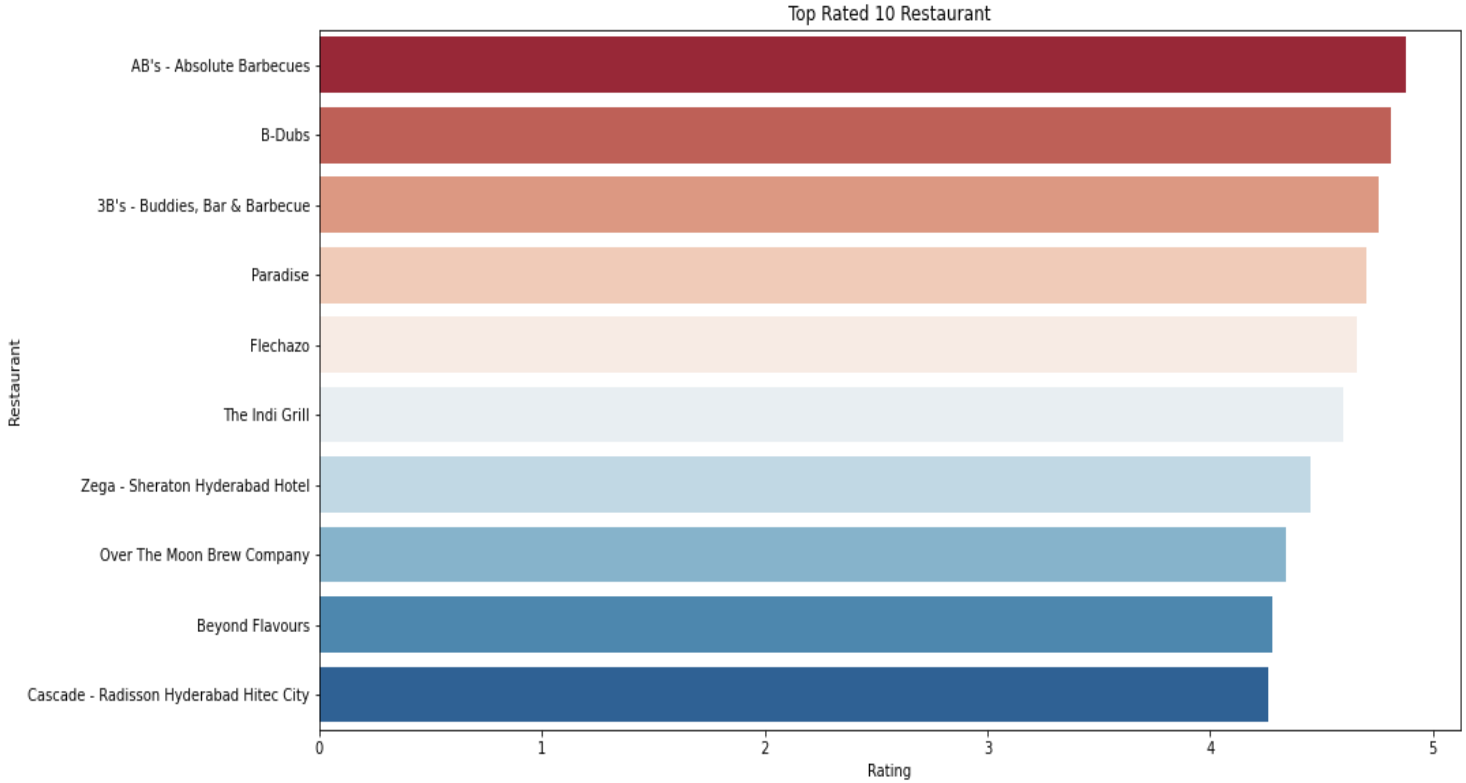
For Clustering

- 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

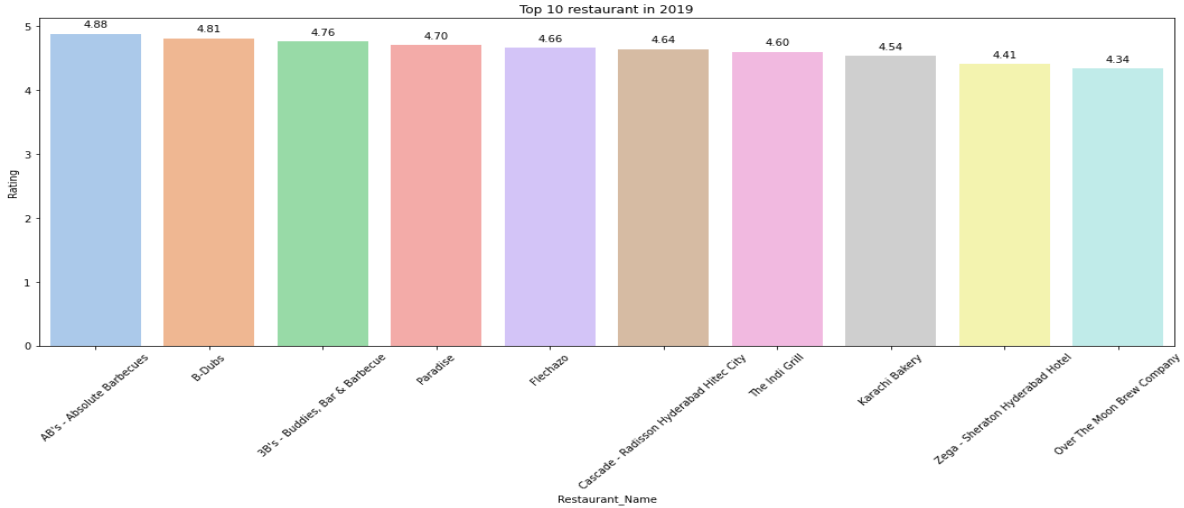
For Sentiment Analysis

- 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

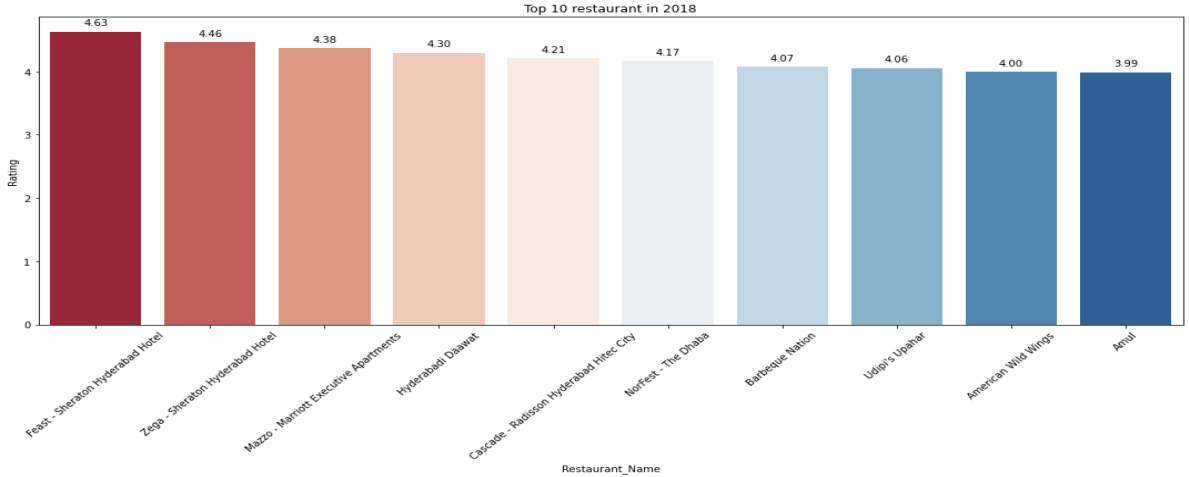
EXPLORATORY DATA ANALYSIS



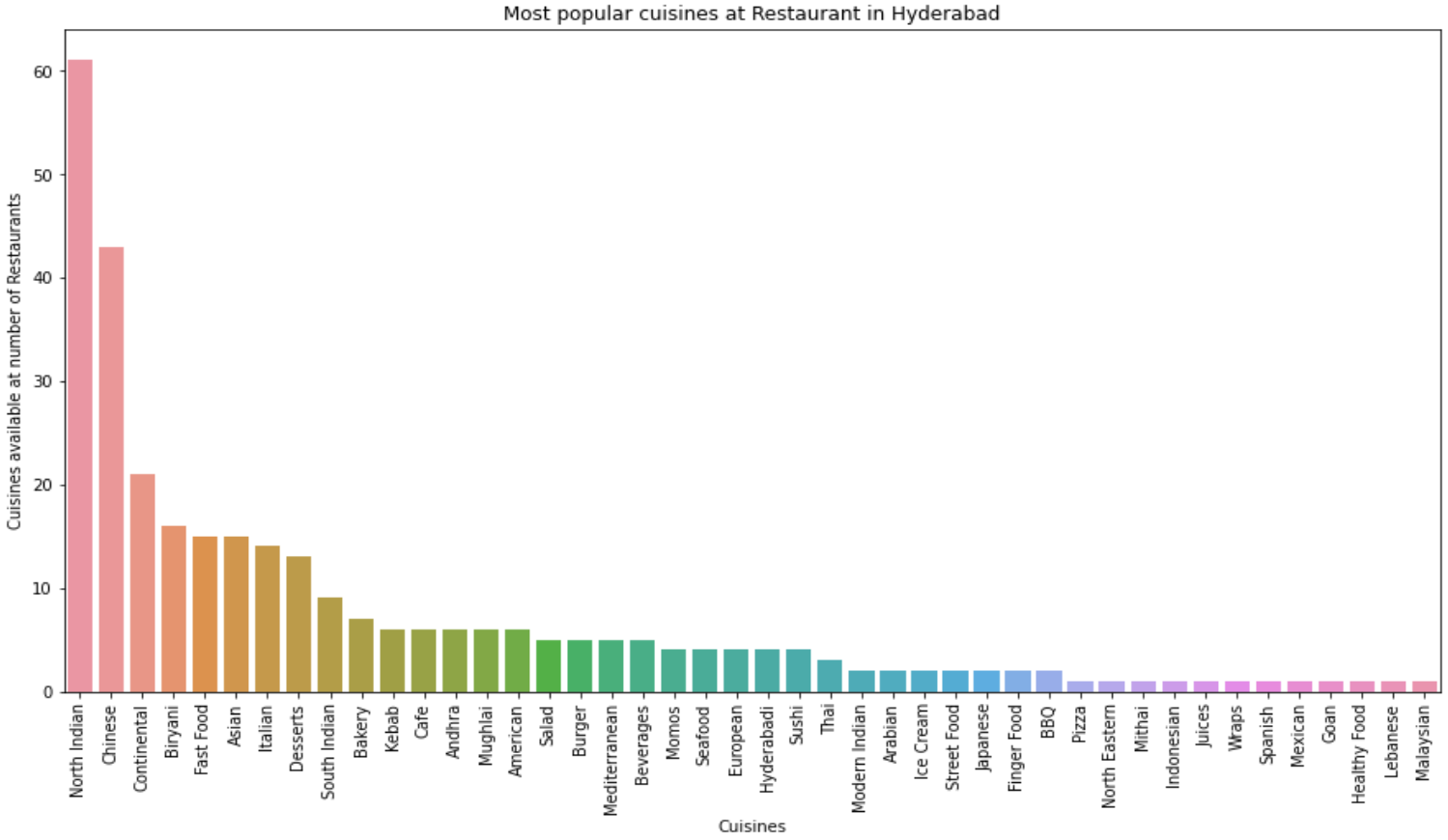
Top 10 Restaurants based on ratings. All having ratings between 4.2 to 4.8.



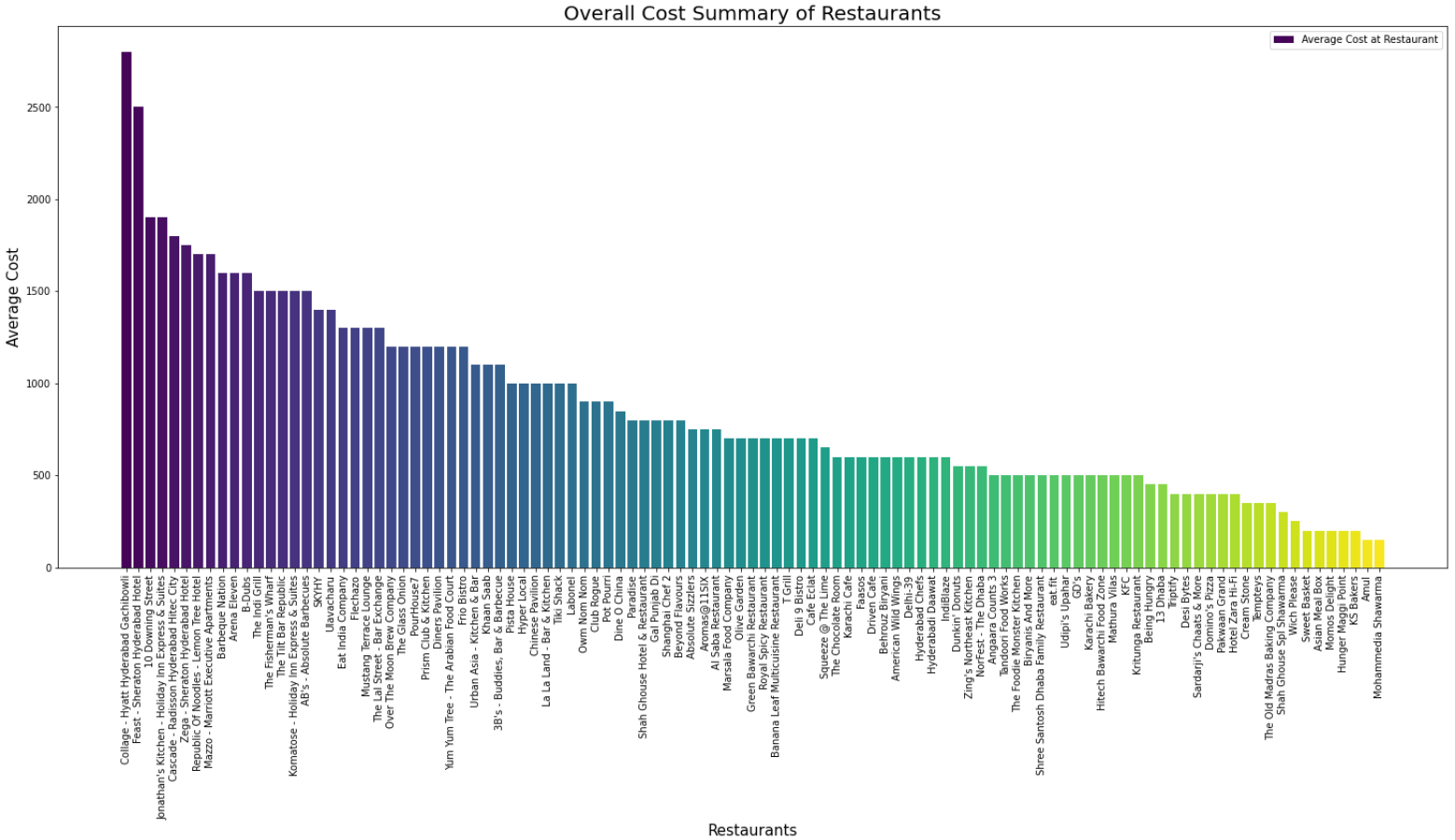
Top 10 Restaurants based on ratings in year 2019.



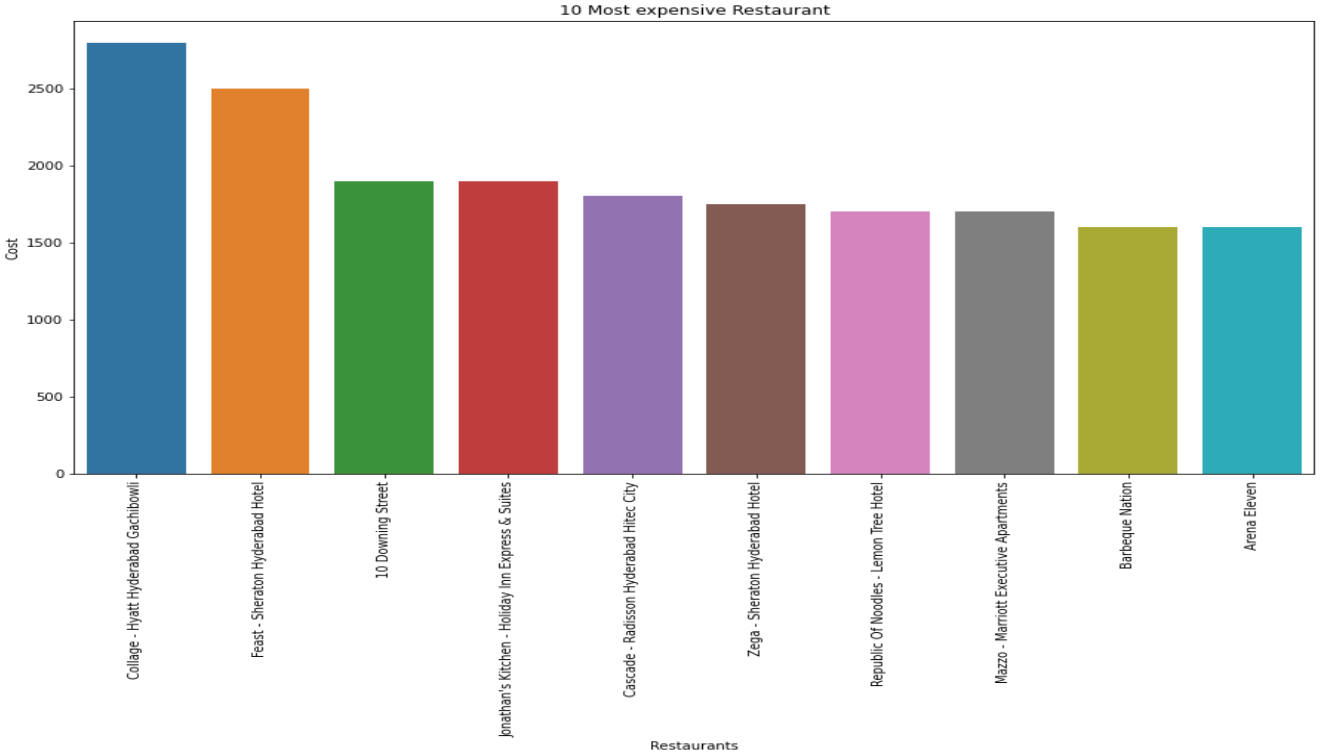
Top 10 Restaurants based on ratings in year 2018.



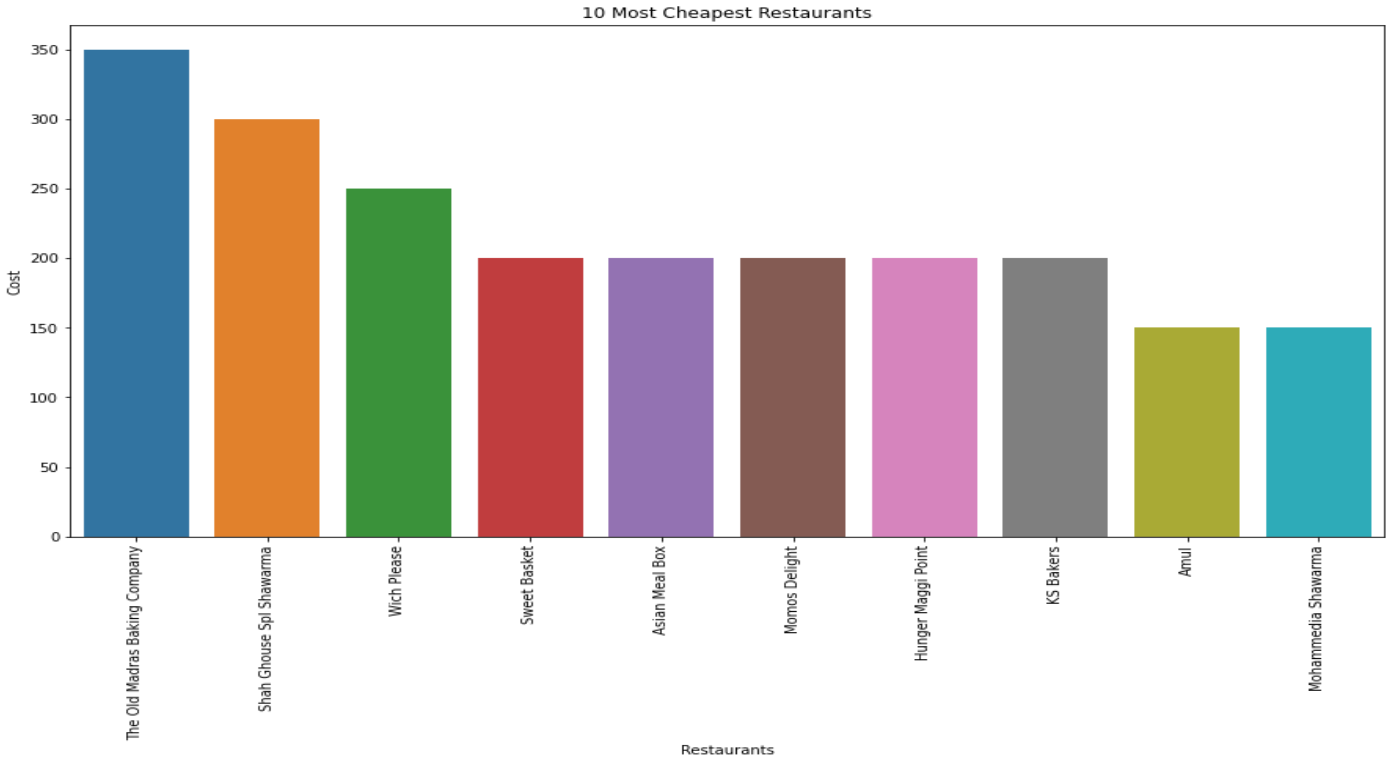
Among Cuisines, North Indian and Chinese are dominant across more than 50% of the restaurants.



Cost Summary of the Restaurants.

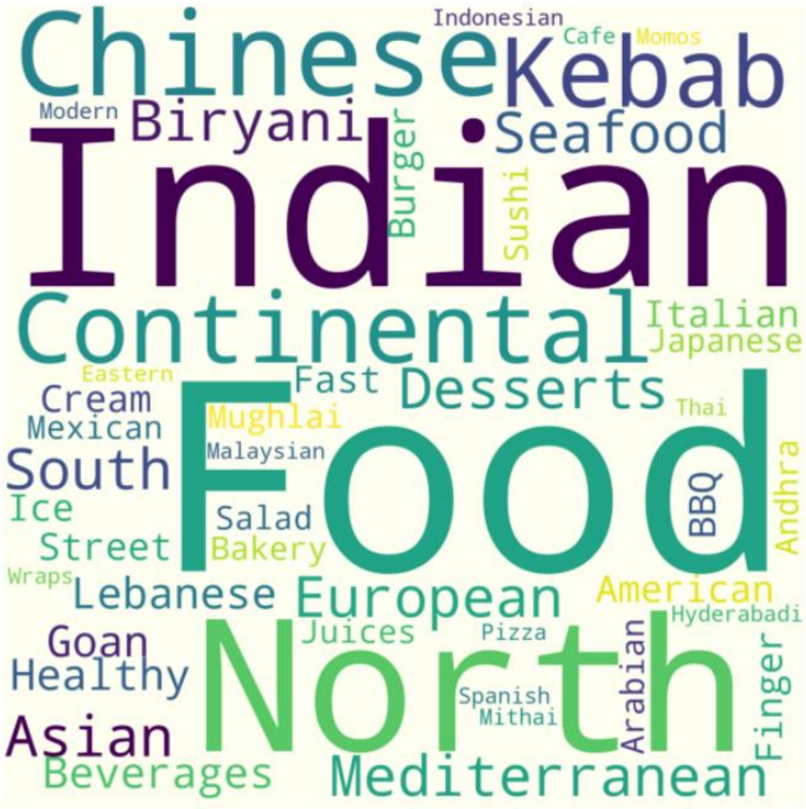


Top 10 Most Expensive Restaurants.



Top 10 Most Cheapest Restaurants.

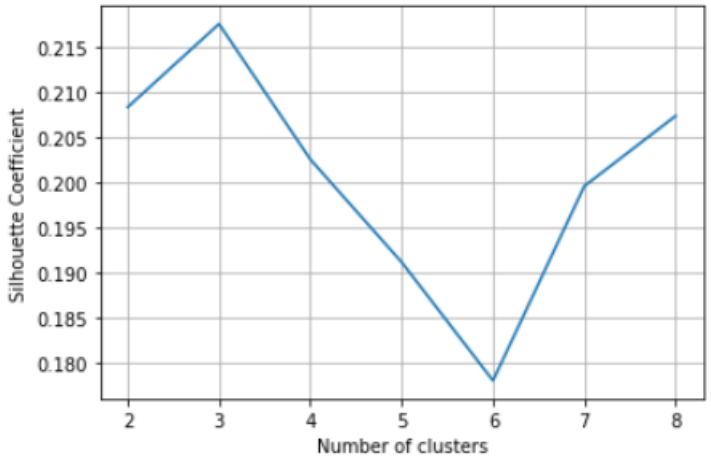
Wordcloud


Chinese
Kebab
Indonesian
Cafe
Momos
Modern
Biryani
Seafood
Indian
Burger
Sushi
Continental
Italian
Japanese
Eastern
Fast
Desserts
Cream
Mexican
Mughlai
Thai
Malaysian
South
Food
BBQ
Andhra
Ice
Salad
Street
Bakery
Wraps
Lebanese
European
American
Hyderabadi
Goan
Juices
Pizza
Healthy
North
Arabian
Spanish
Mithai
Finger
Asian
Beverages
Mediterranean

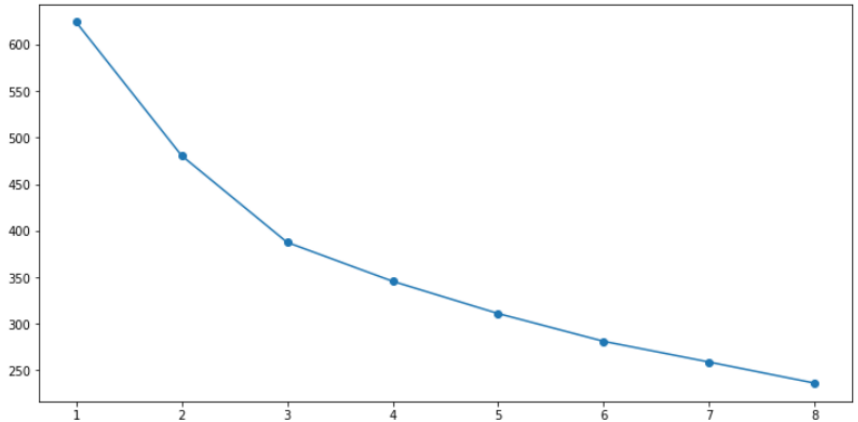

Wings
Monster
Prism
Terrace
Bakers
Asian
Beyond
KFC
Bar
Tree
Dhaba
Mazzo
eat
Wharf
Radisson
Lal
Vilas
Spl
Pakwaan
City
Sizzlers
Royal
Aromas
Onion
Spicy
Glass
Suites
Company
Gal
Zara
Driven
fit
Shanghai
Express
Tempteys
Local
Santosh
Fi
Paradise
Jonathan
Pista
Republic
Biryani
Inn
Bawarchi
Olive
Apartments
Family
Ulavacharu
Feast
NorFest
Desi
Restaurant
Delight
Hitec
Eclat
Leaf
Faasos
Lime
Garden
SKYHY
Frio
Court
Pot
Squeeze
American
Old
3B
Exchange
Pourri
Hyper
Works
Barbecue
Madras
Street
Hyatt
Yum
Nation
Saab
Bytes
PourHouse7
Barbeque
Mathura
Owm
Chaats
Kritunga
Bakery
Chinese
Club
Kitchen
Upahar
China
Asia
Hunger
Box
Room
Udipi
Indi
Delhi
Fisherman
Shah
Mustang
Ghouse
Shack
Grill
Rogue
Point
Chocolate
Dubs
Cafe
Holiday
Deli
Punjab
La
India
Flechazo
Zega
Brew
Multicuisine
Stone
Amul
Pizza
Lounge
Downing
Arena
Domino
Marsala
Hotel
Nom
Tandoori
Food
Chef
Hi
Zone
Diners
Zing
Cascade
Land
Dine
Tilt
AB
Hyderabadi
Urban
Cream
Pavilion
Komatose
Arabian
Moon
House
Bistro
Hyderabad
Donuts
Momos
Meal
Saba
Khaan
Northeast
Eleven
Buddies
Karachi
Foodie
Sheraton
Hungry
Tiki
Shree
Sardarji
Flavours
Behrouz
Labonel
Hitech
Grand
Banana
Baking
Absolute
Shawarma
Green
Executive
Marriott
Maggi
Triptify
Angaara

K Means Clustering plots

Silhouette score

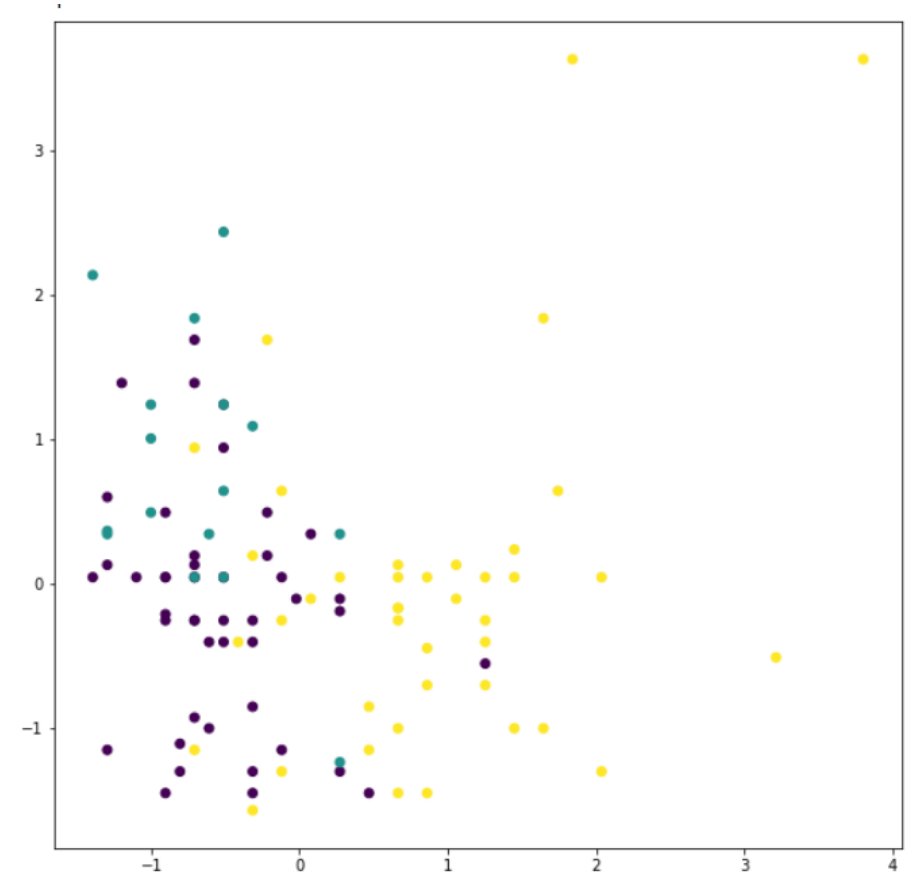


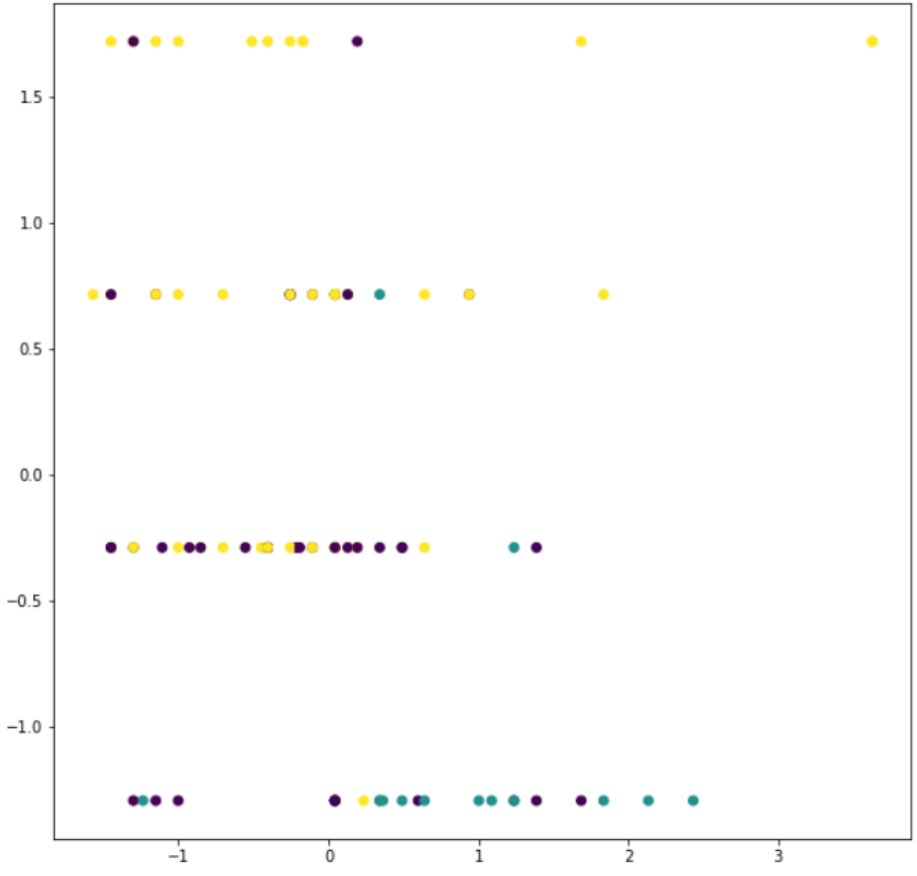
Sum of squares elbow plot



From the Silhouette score plot, we can see that the optimum number of clusters is 3, and from the elbow plot we can see that the optimum number of clusters is 3.

Cluster Visualization





Cuisines in different clusters

Cluster 0

'north indian', ' chinese', ' continental', ' mediterranean', ' european', 'european', ' north indian', ' seafood', ' biryani', ' hyderabadi', 'continental', 'american', 'biryani', ' south indian', ' andhra', 'mediterranean', ' kebab', ' bbq', 'chinese', ' italian', ' asian', ' mughlai', ' beverages', 'modern indian', 'asian', ' desserts', ' spanish', ' japanese', ' salad', ' sushi', 'andhra', 'italian', 'mexican', 'kebab', 'thai', ' malaysian', ' thai', ' indonesian', 'seafood', ' goan', 'bbq', ' modern indian', ' finger food', 'healthy food'

Cluster 1

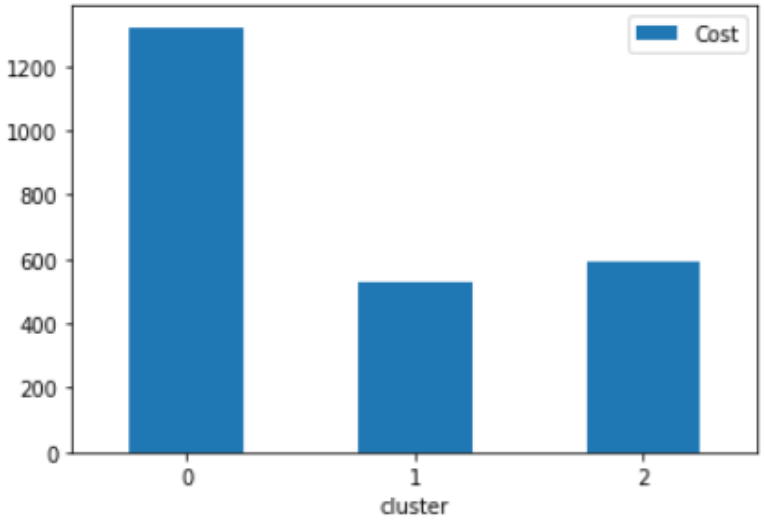
'ice cream', ' desserts', 'cafe', ' bakery', ' continental', ' fast food', ' beverages', 'desserts', ' cafe', ' burger', 'fast food', ' biryani', 'bakery', ' north indian', ' mughlai', ' juices', ' chinese', ' mithai', 'american', ' wraps'

Cluster 2

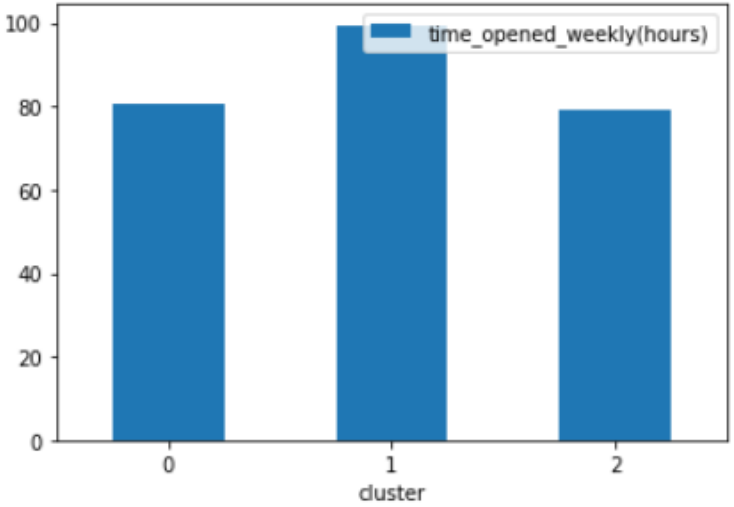
'north indian', 'continental', ' american', ' chinese', 'american', ' fast food', ' salad', ' burger', ' biryani', ' mughlai', 'asian', 'mughlai', 'chinese', ' seafood', ' asian', ' momos', 'fast food', ' pizza', 'burger', ' continental', 'biryani', ' north indian', 'hyderabadi', ' japanese', ' sushi', 'finger food', ' kebab', 'arabian', 'south indian', 'street food', ' arabian', 'momos', ' south indian', 'lebanese', ' andhra', ' thai', 'north eastern'

Average Cost and Time in three different clusters

Cost Distribution



Time Distribution



Cuisines in different clusters

Cluster 0

'north indian', ' biryani', ' chinese', ' mughlai', 'asian', 'mughlai', ' fast food', 'cafe', ' continental', ' desserts', 'chinese', ' asian', ' momos', 'fast food', ' pizza', 'biryani', ' north indian', 'hyderabadi', 'burger', ' japanese', ' sushi', 'finger food', ' kebab', 'momos', 'street food', ' burger', 'continental'

Cluster 1

'north indian', ' chinese', ' continental', ' seafood', ' biryani', ' hyderabadi', 'continental', 'mediterranean', ' north indian', ' kebab', ' bbq', ' italian', ' asian', ' mughlai', ' beverages', 'modern indian', 'asian', ' mediterranean', ' desserts', 'andhra', 'italian', ' south indian', 'kebab', 'thai', ' malaysian', 'seafood', ' goan', 'bbq', ' modern indian', ' sushi'

Cluster 2

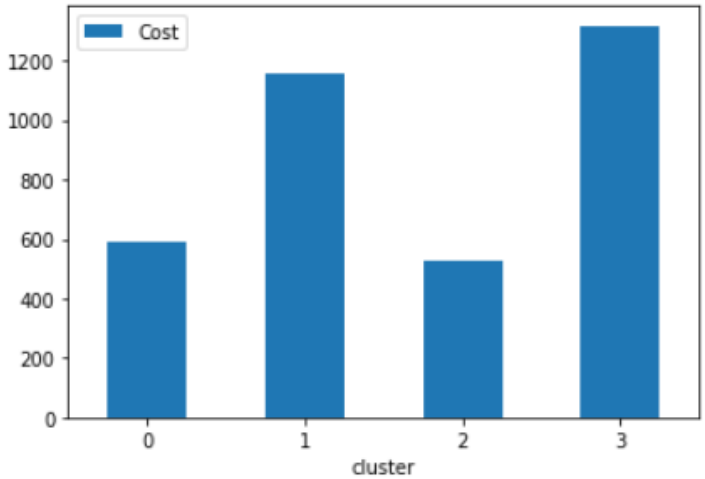
'ice cream', ' desserts', 'cafe', ' bakery', ' fast food', ' beverages', 'desserts', ' cafe', ' burger', 'fast food', ' biryani', 'bakery', ' north indian', ' mughlai', ' juices', ' chinese', ' mithai', 'american', ' wraps'

Cluster 3

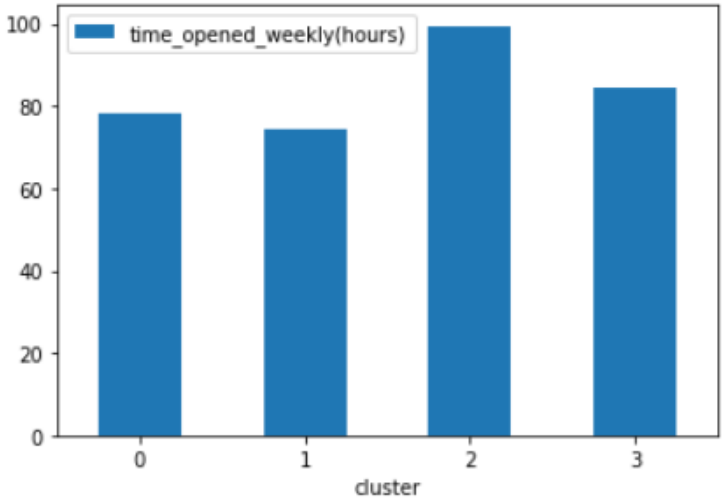
'north indian', ' mediterranean', ' european', 'european', ' north indian', 'continental', ' american', ' chinese', 'american', ' fast food', ' salad', ' burger', 'biryani', ' south indian', ' andhra', 'chinese', ' continental', ' kebab', ' seafood', ' italian', ' spanish', 'burger', 'fast food', ' japanese', ' sushi', 'arabian', 'south indian', 'street food', ' arabian', 'mexican', ' biryani', ' beverages', 'lebanese', ' thai', ' indonesian', ' asian', 'italian', ' finger food', 'andhra', 'asian', ' momos', 'north eastern', 'healthy food'

Average Cost and Time in four different clusters

Cost Distribution

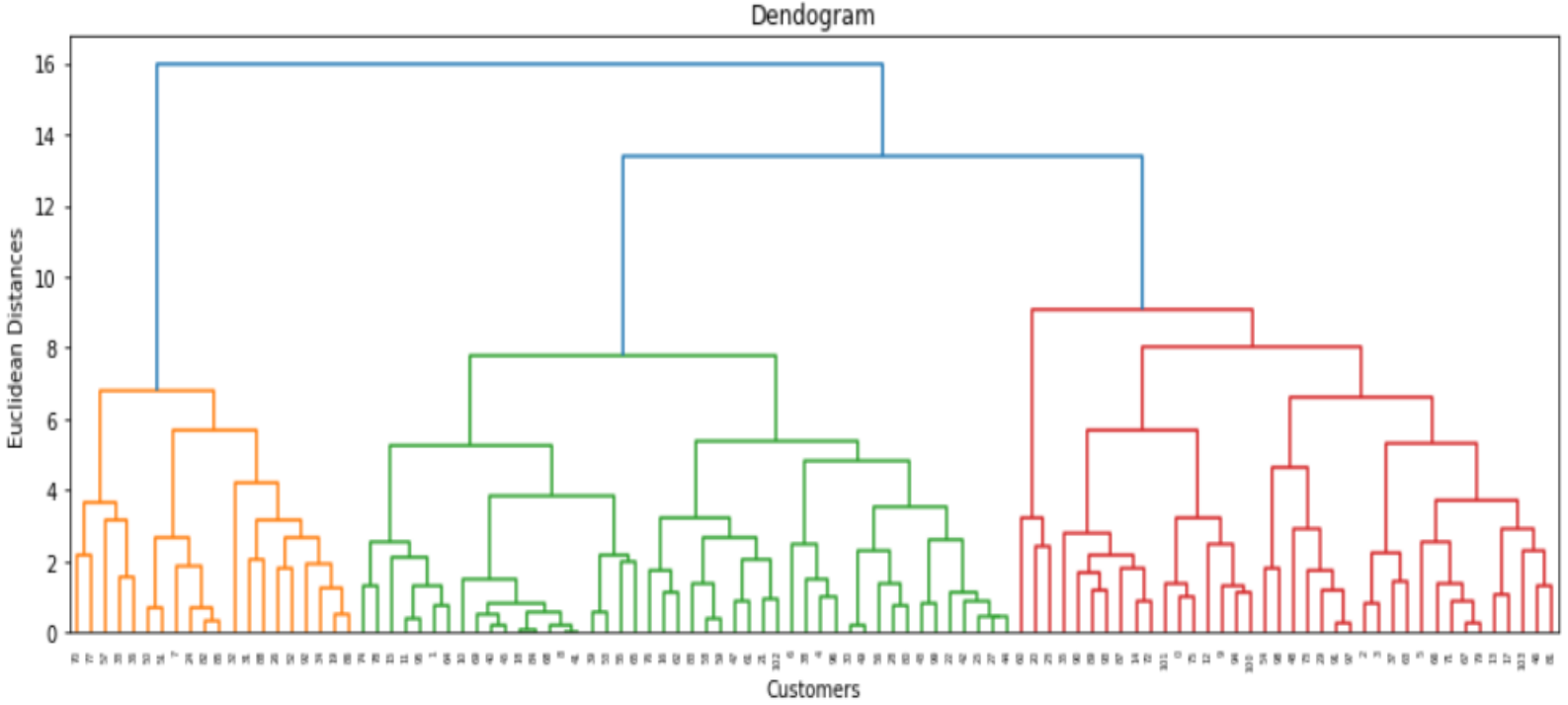


Time Distribution

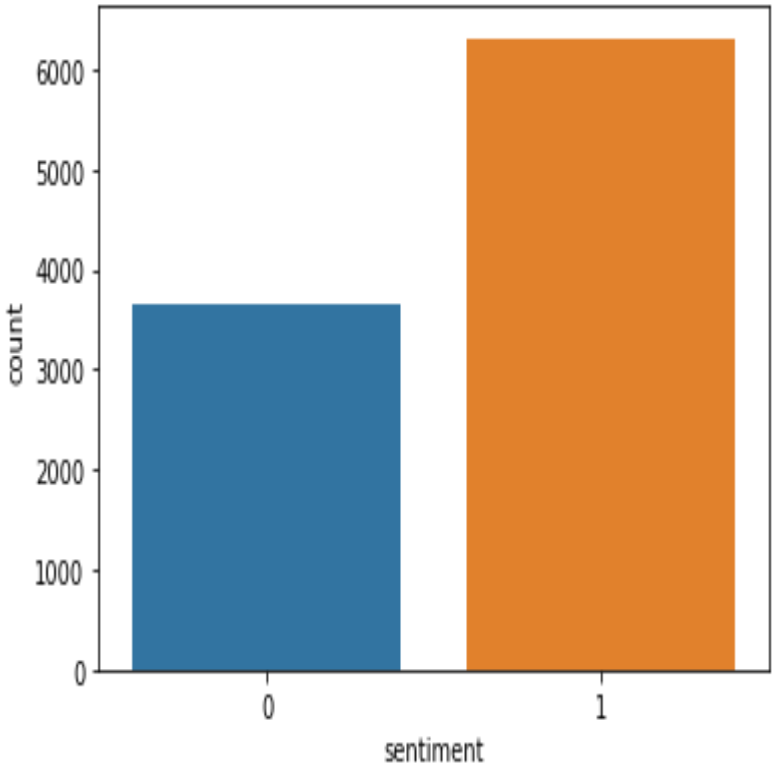


We can say that restaurants that are open most of the time are relatively costlier.

Dendrogram for Hierarchical clustering



Sentiment Analysis



Ratings greater than or equal to 3.5 are classified as 1 (Positive sentiment), others as 0 (Negative sentiment)

Sentiment Analysis

AI

Short review: Decent breads and starters, bad Shaam Savera, a little screwed up daal makhni (but should be a one-off case)\nBetter ambience at night... And importantly good quantity!\nI wanted to give a 2.5 but that's not an option 🙈\nExtensive...\nAmbience: It's better at night. Friends who had been here during the day didn't like it much. So if you prefer ambience, go at night!\nFood! Ordered a Hara Bhara Kebab, Shaam Savera, Daal Makhni, rotis and raita (Yes! I'm a vegetarian 🙈)\nThe good part! Quantity 😍\nI was excited to see Shaam Savera on their menu. But take it from me, you do not want to order it. It is no where close to what it should be. They messed it up real bad, especially the colors.\nThe daal makhni had big chunks of garlic cloves which spoiled it for me. It wasn't creamy too :(\nThe breads and raita were decent!\nThe Hara Bhara Kebabs were decent too, soft! They had an extra outer coating of peanuts. So if you're not a peanut fan on your kebabs, ask them to skip that step 🙈\nConclusion: It looked like they had made food in a hurry.\nWould like to give this place another try hoping that mine was one of the odd cases. If things change, you'll know!

The reason for giving only a 3 star is because of the longlong time wait for the tables. We went on Friday afternoon. We waited almost more than 30mins to get a table. We lost our patience and left for another restaurant as v didn't get a table even aftr waiting for such a long time.I think restaurant management should take this into consideration and try to decrease the waiting time of customer.

We went for lunch buffet yesterday and took a chance after seeing some good and some average reviews in Zomato.\n\nThe spread in buffet is too less, just 3 veg and 3 non veg starters, 2 main course in veg and non veg, just one fried rice and chicken Biryani.\nDesserts are ice cream and jamun.\n\nThe taste was ok and not to expect too much. Little disappointing when starters are not hot.\n\nFish fry was decent, Schezwan chicken was ok.\nVeg spring rolls were too good.\nBig surprise for me is Onion rings as starters 🤣😍\n\nKadai ghost in main course was good but missing that masala flavour.\nChicken was ok. Biryani is average, egg fried rice was good.\n\nFor the money we pay, we expect some good food. They use much of colour in food item's.\n\nSuggestion to management is use less coating and less food colouring in starters.\n\nOverall experience is ok, but they can be much better for the amount you pay.

We happened to go to this place on last sunday and it was mothers day and they had flat 50 on the buffet meal for the mother which i really liked..Needless to say the Ambience and music was really good .We had a reservation but had to wait fir 10 mins to get a seat..They served us mango juice as wrlcome drink which tasted really good.The buffet spread was one of a kind:\nStrters : Veg : paneer tikka, veg spring roll and veg cutlet.They were lil cold but the taste wad good\nNon veg : chicken kebab , chicken manchurian and fish .Fish was very tasty though it coupd taste better if it was hot.Chicken was fair enuf.\nSingapore Noodles,Egg salad and pasta : They were good but it was cold.\nMain course: mutton rogan josh, chicken tikka masala,palak paneer All were equally good\nBiriyani: Egg frued rice and chicken biriyani were really great.\nDesserts : moong dal halwa tasted really yumm, ice cream,gulab jamun,cut fruits,banana cake\n\nThe staff were too slow in service and the food was refilles very slow.We had to wait for quite long time to actually eat to stomach full.\n\nThey could improve the above points to stand better!\nOverall :3.2

I have been to this place twice and had 2 different experiences.\n1st time it was with colleagues and I loved the food and ambience and service and everything except the prices as it was too costly. Their special baked biriyani is work a try, it's unique and tasty.\nBecause of the good experience I revisited this place with my sister. She orders a mocktail and it was terrible.they replace it but it was still not good. And then I asked them about a starter, if it was with bone or boneless and they told me it was boneless, later on the arrival they told me that it comes with both. It's disappointing to know that the staff doesn't know about the menu.

Special characters, emojis and stop words in the data were removed

Word cloud for Reviews

Overall word cloud of review features

Adjectives and verbs

For reducing the dimensionality, adjectives and verbs were included as features

Performance metrics for diff. models (training set) - (TF-IDF vectorizer)

Algorithm	Class label	Performance parameters			
		Accuracy	Precision	Recall	F1 score
Multinomial NB	0	0.81	0.91	0.67	0.77
	1		0.83	0.96	0.89
Logistic Regression	0	0.88	0.78	0.92	0.84
	1		0.94	0.85	0.90
Decision trees	0	0.75	0.72	0.66	0.69
	1		0.81	0.85	0.83
Random Forest	0	0.98	0.97	0.99	0.98
	1		0.99	0.98	0.99

Performance metrics for diff. models (test set) - (TF-IDF vectorizer)

Algorithm	Class label	Performance parameters			
		Accuracy	Precision	Recall	F1 score
Multinomial NB	0	0.79	0.86	0.64	0.74
	1		0.82	0.94	0.87
Logistic Regression	0	0.82	0.70	0.87	0.78
	1		0.91	0.78	0.84
Decision trees	0	0.73	0.70	0.64	0.67
	1		0.80	0.84	0.82
Random Forest	0	0.81	0.80	0.74	0.77
	1		0.85	0.89	0.87

Performance metrics for diff. models (training set) - (Bag of words)

Algorithm	Class label	Performance parameters			
		Accuracy	Precision	Recall	F1 score
Multinomial NB	0	0.83	0.85	0.75	0.80
	1		0.86	0.92	0.89
Logistic Regression	0	0.92	0.86	0.94	0.90
	1		0.96	0.91	0.93
Decision trees	0	0.76	0.68	0.73	0.71
	1		0.84	0.80	0.82
Random Forest	0	0.81	0.64	0.93	0.76
	1		0.94	0.70	0.80

Performance metrics for diff. models (test set) - (Bag of words)

Algorithm	Class label	Performance parameters			
		Accuracy	Precision	Recall	F1 score
Multinomial NB	0	0.81	0.80	0.73	0.76
	1		0.85	0.89	0.87
Logistic Regression	0	0.81	0.73	0.81	0.77
	1		0.88	0.82	0.85
Decision trees	0	0.74	0.66	0.72	0.69
	1		0.82	0.78	0.80
Random Forest	0	0.75	0.60	0.86	0.70
	1		0.89	0.66	0.76

Challenges

1) Number of data points as well as the number of features were low for the clustering dataset, because of which we have only 3 clusters out of it.

2) Converting Timings feature in metadata for clustering is challenging as the same restaurant is open for different timings for different days of a week, we somehow converted it into the total number of hours.

3) Collection feature from metadata has 54 null values and converting it into a useful feature for clustering is a big task, we thought of getting information using the link features, though for every restaurant we have to do it manually one by one, so we dropped the idea.

4) In the Sentiment Analysis we tried by reducing features using regularization though it's causing overfitting, we did a different experiment, though the result remain the same.

Conclusion

- **The best model we found out is Logistic regression for sentiment analysis.**
- **We can say that restaurants that are open most of the time are relatively costlier.**
- **Getting 3 optimum number of clusters by using elbow analysis and 3 number of optimum clusters by using silhouette coefficients.**
- **In our case 4 clusters were best to cluster the data.**

Thank You