Zomato Restaurant Clustering & Sentiment Analysis (Unsupervised)

Team Presentation

Vinayak Marathe
Data Science Enthusiast

Riya Patel
Data Science Enthusiast

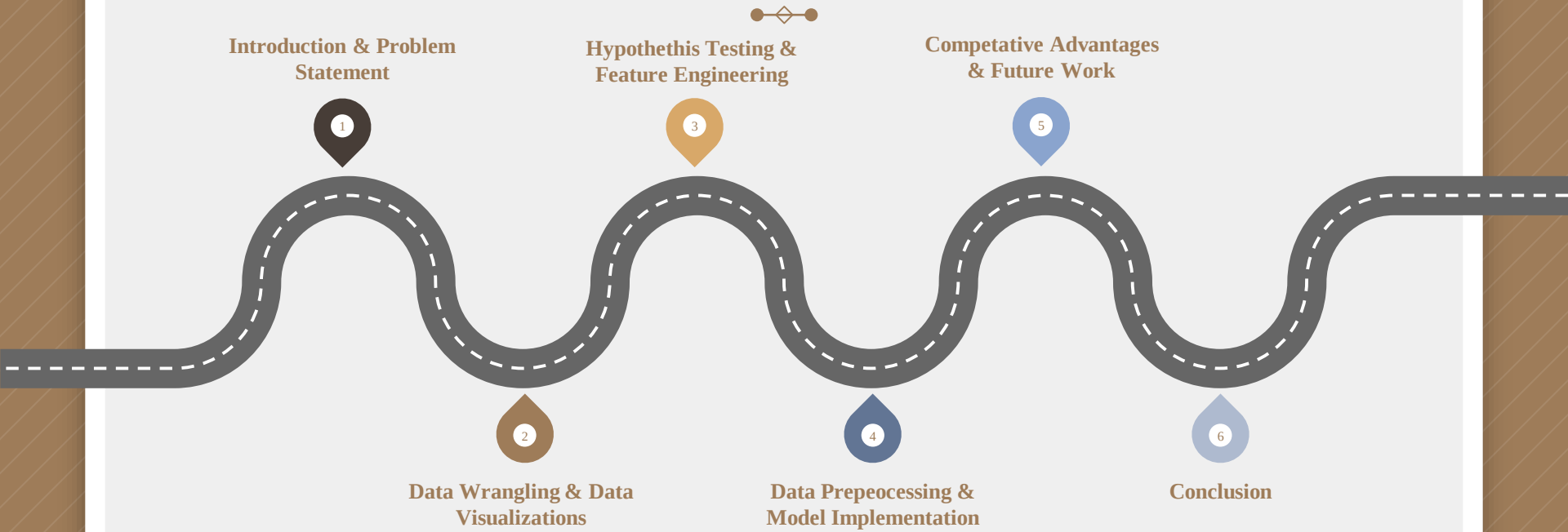
We are as thrilled as you to present our project !!!

Hello!

Our project aims clustering of the restaurants based on their features and performing sentiment analysis on customer reviews. The project uses unsupervised learning techniques to group similar restaurants and identifies patterns and similarities among them.

Let's Start !!!

Roadmap


Introduction & Problem Statement
1
Data Wrangling & Data Visualizations
2
Hypothethis Testing & Feature Engineering
3
Data Prepeocessing & Model Implementation
4
Competative Advantages & Future Work
5
Conclusion
6

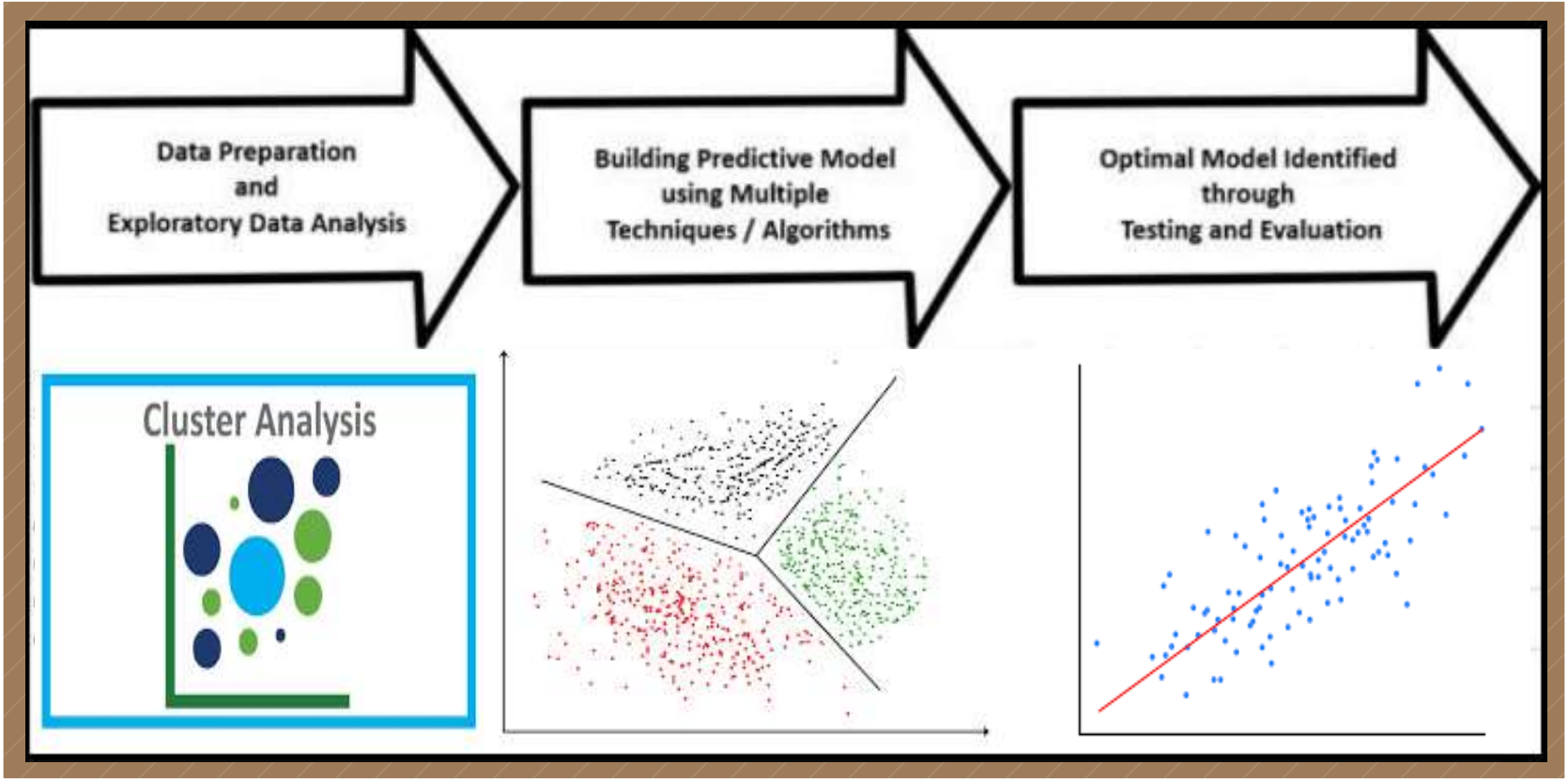
I. Introduction

Let’s start with the information about zomato dataset and used techniques

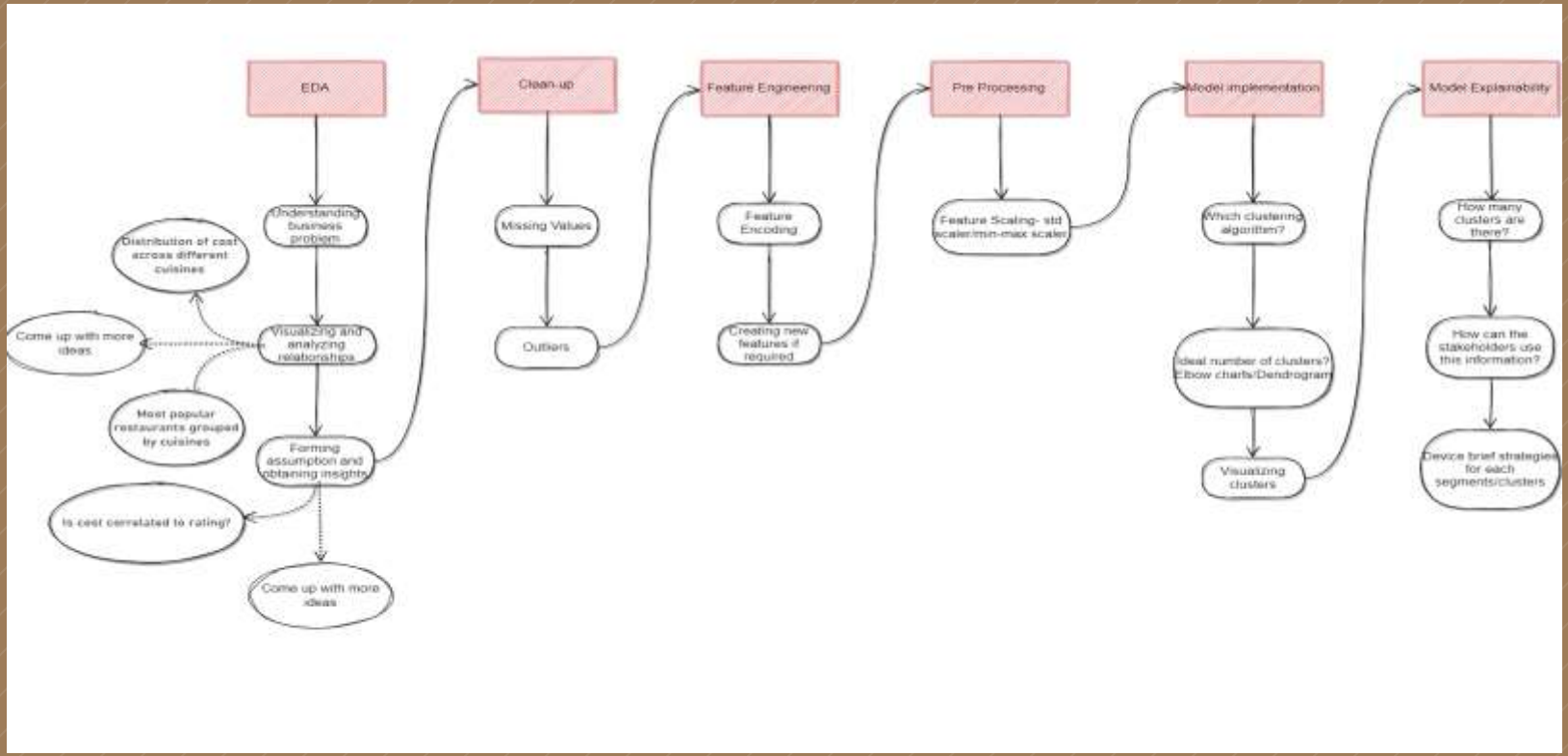
Project Summary

- 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. Zomato provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities.
- Our task here is to make a project that involves clustering restaurants based on their features and performing sentiment analysis on customer reviews. The project uses unsupervised learning techniques to group similar restaurants and identifies patterns and similarities among them.
- Our Project uses data from customer reviews to create clusters of restaurants using their ratings, and then predicts the sentiment of their reviews. This is helpful for restaurants to understand how customers feel about their offerings, and how they can improve the quality of their services.
- The Analysis also solves 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.

Approach


Data Preparation
and
Exploratory Data Analysis
Building Predictive Model
using Multiple
Techniques / Algorithms
Optimal Model Identified
through
Testing and Evaluation
Cluster Analysis

Project Architecture



II. Problem Statement

Let’s start with the challenge of understanding customer preferences

BUSINESS PROBLEM OVERVIEW

The restaurant industry is highly competitive and dynamic, with new restaurants constantly entering the market. Restaurant owners face the challenge of understanding customer preferences and meeting their expectations to stay relevant in the market. In this context, there is a need for a data-driven approach that can help restaurant owners gain insights into customer preferences and improve their services accordingly.

The problem statement of Zomato Restaurant clustering and sentiment analysis is to use Machine Learning algorithms to cluster restaurants into groups according to certain criteria and to analyze customer reviews to determine the sentiment and opinion about each restaurant. The purpose of this project is to provide insights about customer preferences, trends in restaurants and the overall satisfaction of customers with dining experiences. The goal is to improve the customer experience by providing better recommendations and improved customer service. The Zomato Restaurant Clustering and Sentiment Analysis Project aims to address this problem by leveraging machine learning techniques to cluster restaurants based on their features and perform sentiment analysis on customer reviews to identify areas of improvement.

III. Understanding Variables

Let’s understand features, its data type and its description

Undertanding Datasets

We have two datasets in our project - "Zomato Restaurant names and Metadata.csv" and "Zomato Restaurant reviews.csv".

The **"Zomato Restaurant names and Metadata.csv"** dataset contains six columns: "Name", "Links", "Cost", "Collections", "Cuisines", and "Timings". There are 105 non-null values in each column except for "Collections" and "Timings", which have 51 and 104 non-null values, respectively.

The **"Zomato Restaurant reviews.csv"** dataset contains seven columns: "Restaurant", "Reviewer", "Review", "Rating", "Metadata", "Time", and "Pictures". There are 10,000 non-null values in the "Restaurant" and "Pictures" columns, while the "Reviewer", "Review", "Rating", "Metadata", and "Time" columns have 9962 non-null values.

Overall, these datasets contain valuable information about restaurants such as their names, cuisines, cost, and customer reviews. The information can be used to perform various analysis tasks such as sentiment analysis and clustering to gain insights into customer preferences and improve restaurant services.

Variable Description

Variable Description of Zomato Restaurant Names and Metadata :

- Name : Name of Restaurant
- Links : URL Links of Restaurants
- Cost : Per person estimated cost of dining
- Collections : Tagging of Restaurants with respect to Zomato Category
- Cuisines : Cuisines served by restaurants
- Timings : Restaurant timings

```
# Dataset Info
restaurant_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 105 entries, 0 to 104
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   Name         105 non-null    object
 1   Links        105 non-null    object
 2   Cost         105 non-null    object
 3   Collections  51 non-null     object
 4   Cuisines     105 non-null    object
 5   Timings      104 non-null    object
dtypes: object(6)
memory usage: 5.0+ KB
```

Variable Description (Cont.)

Variable Description of Zomato Restaurants Reviews :

- Restaurant **:** Name of Restaurant
- Reviewer : Name of the reviewer
- Review : Review text provided
- Rating : Rating provided
- Metadata : Reviewer metadata i.e. No of reviews and followers
- Time : Date and Time of Review
- Pictures : Number of pictures posted with reviews

```
reviews_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Restaurant  10000 non-null  object
 1   Reviewer    9962 non-null   object
 2   Review      9955 non-null   object
 3   Rating      9962 non-null   object
 4   Metadata    9962 non-null   object
 5   Time        9962 non-null   object
 6   Pictures    10000 non-null  int64 
dtypes: int64(1), object(6)
memory usage: 547.0+ KB
```

IV. Data Wrangling & EDA

Let’s find meainingful insights and analysis

Overall Insights From Zomato Restaurant Names and Metadata.csv

- This dataset has a shape of **(105, 6)** with 54 null values in ‘Collection’ column and 1 null values in ‘Timing’ column.
- There are **44 different/unique Cuisines** present in our dataset.
- There are 33 restaurants with 3 different cuisines, 26 restaurants with 2 different cuisines, 21 restaurants with 4 different cuisines, 12 restaurants with 5 different cuisines, 12 restaurants with only 1 cuisine and only 1 restaurant with 6 different cuisines.
- **North Indian** is the mostly available food on **61 number of different restaurants** followed by Chinese and Continental food which is available on 43 and 21 restaurants. It means mostly people demands for this food as compared to other food like pizza, juices, malaysian food which is available on less number of restaurant.
- We found that, **Beyond Flavors restaurant** has a maximum number of different type of cuisine followed by Shah Ghouse Hotel & Restaurant. While **Republic Of Noodles - Lemon Tree Hotel** has only 4 types of cuisine available.
- We also found that **North Indian, Chinese, Continental food** are more in demand in Hyderabad city where there are **63 restaurants with Indian cuisine** is available and **43 Restaurants with Chinese** cuisine is available..

Overall Insights From Zomato Restaurant Names and Metadata.csv

- **Collage - Hyatt Hyderabad Gachibowli and Feast - Sheraton Hyderabad Hotel** are the two most expensive restaurants with the **cost of 2800 and 2500 rupees** which is a average per person estimated cost of dining.
- **Mohammedia Shawarma, Amul and Asian Meal Box Hotel** are the most affordable Restaurants where **the cost of 150 and 200 rupees** which is a average per person estimated cost of dining.

Overall Insights From Zomato Reviews Data.csv

- From this dataset, we found that **5 star** is the highest rating given by **3832 number of customer**. At the same time, Lowest rating given by the customers is 1.5 star.
- We analyses the Top 20 Popular customers (With Highest number of followers) and their order timings. Here we got to know that **Satwinder Singh** has the highest number of followers with **13410** followers, followed by Eat_vth_me who has slightly less number of 13320.
- We found Top 20 Popular customers and restaurants (With Highest number of Review_count). **Anvesh Chowdary** has posted a high number of reviews i.e. **1031 reviews**, while the hotel named **Collage - Hyatt Hyderabad Gachibowli, Pista House and Labonel** has **1031** number of reviews count.
- There are **968 orders** on zomato application between the time frame of **10PM to 11PM** which is highest among all. The time frame between **7PM to 12AM is the peak hour** for the restaurants as well as zomato. Zomato can throw more offer in this time frame so that one can use more number of offers and this is how zomato can increase their customers.

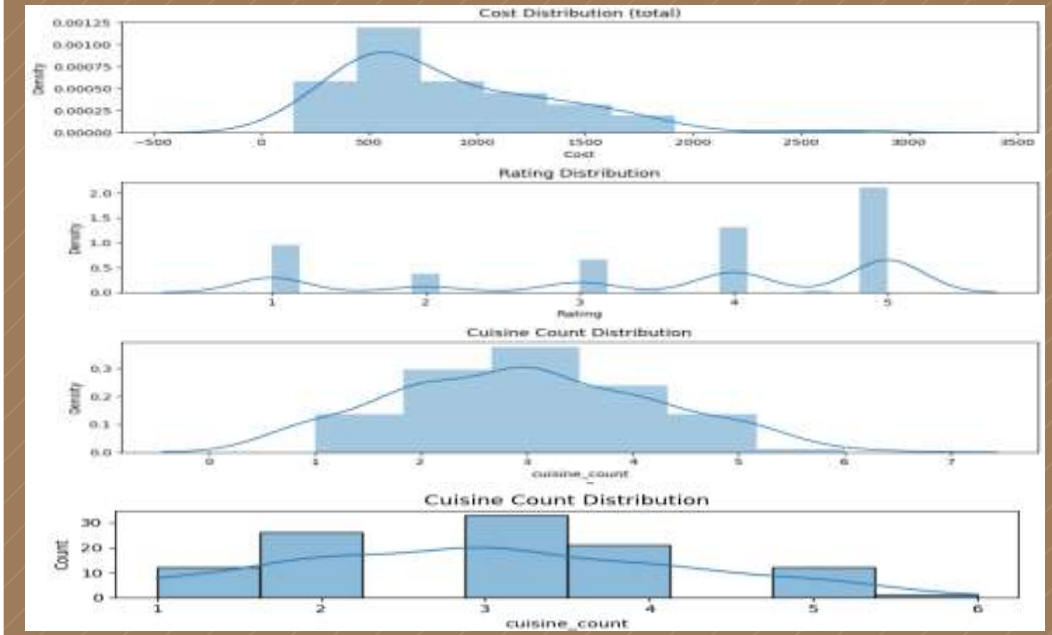
Overall Insights From Zomato Reviews Data.csv

- The **first order** is placed on Zomato is on: **2016-05-31 16:41:00**. While the **last order** is placed on Zomato is on: **2019-05-25 20:23:00**.
- Using Time column present in our dataset, we found that our overall dataset **containing data of 1089 Days Three Hours and 42 Minutes**.
- There are **7447 Different Customer** who have ordered food From Zomato from which **6105 Customers** Ordered food items **only once** on Zomato. While **1342** are the **repeated customers or can say regular customer** on Zomato.
- **'Kiran', 'Ankita', 'Parijat Ray', 'Vedant Killa', 'Jay Mehta‘** are the most valuable customers who have ordered food from Zomato **more than 10 times**.

V. Data Visualization

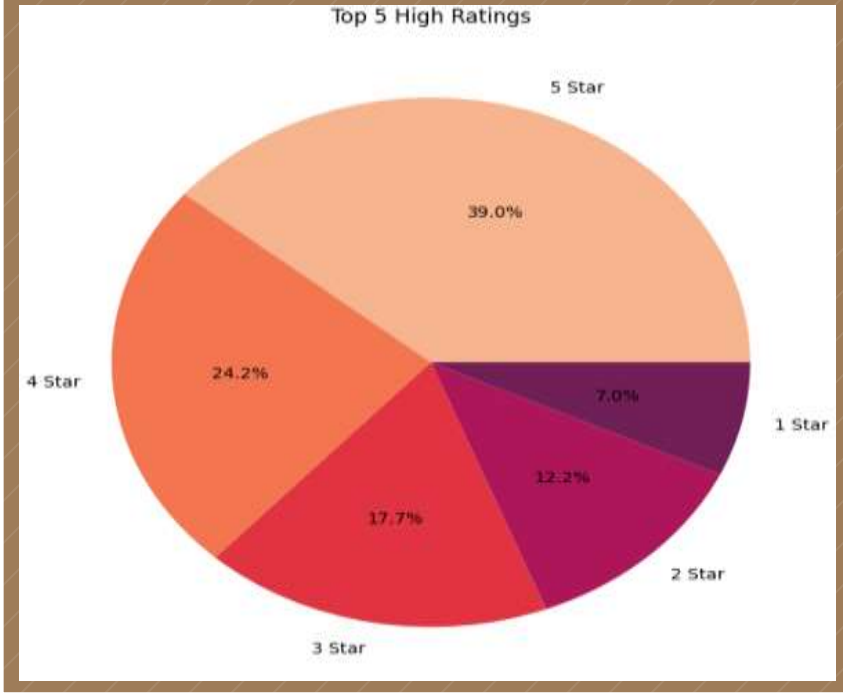
Let’s check storytelling of the charts and understands the relationship b/w variables

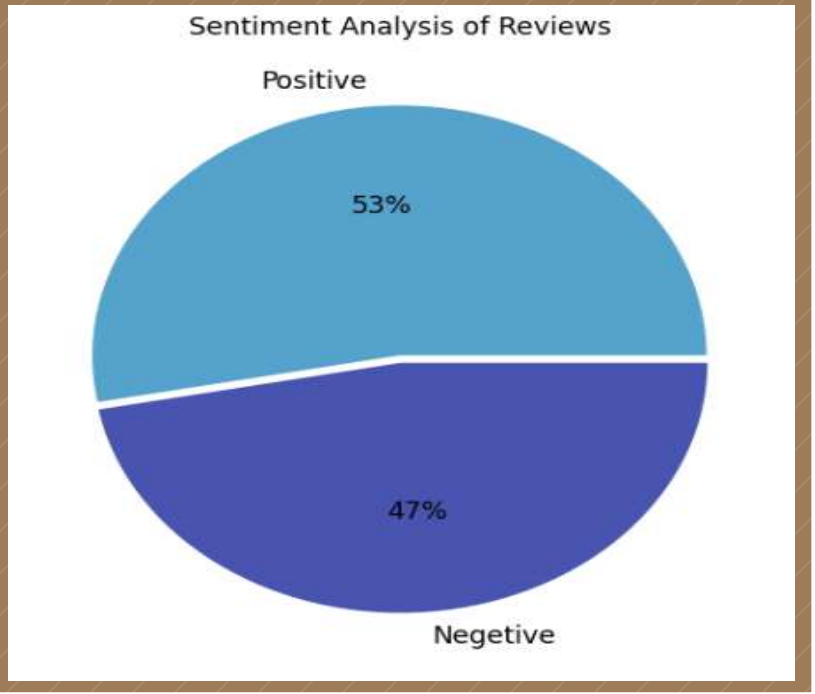
Distribution of Rating, Cuisine Count and Total Cost



- Cost column shows skewness.
- Maximum restaurant has a price range of 450 to 800 rupees.
- Most of the restaurants got 5 star ratings.
- Most of the restaurants has 3 types of cuisines.
- Most of customers posts 1 or 2 pictures with their reviews.

Ratings Percentages

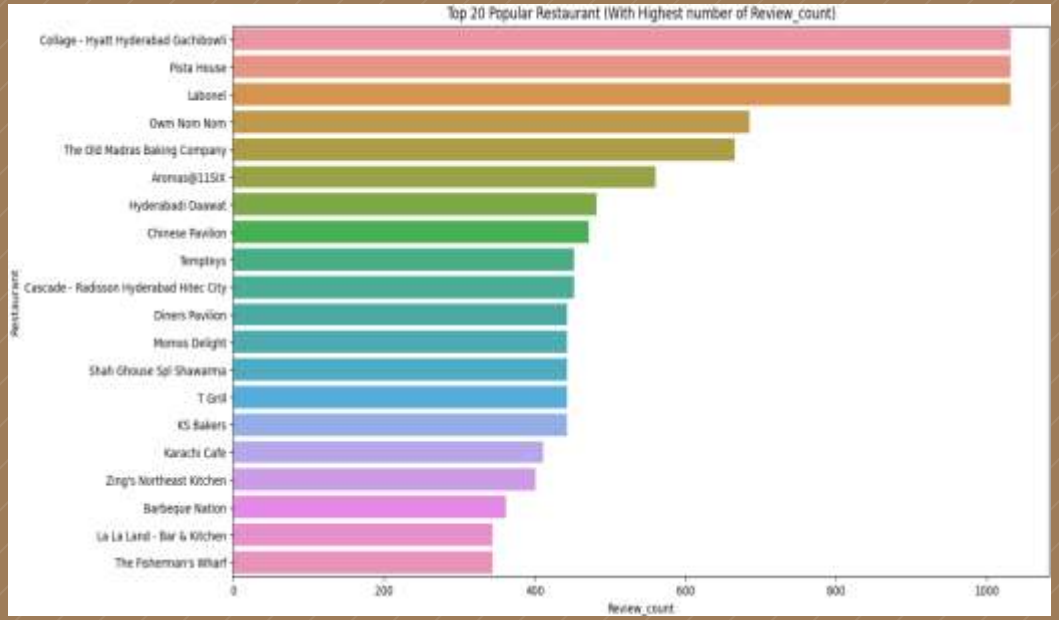




Word cloud of Reviews Sentiment

Customers gave positive as well as negative feedback using this keywords. So that we can easily examine that what exactly the behavior and views of our customers.

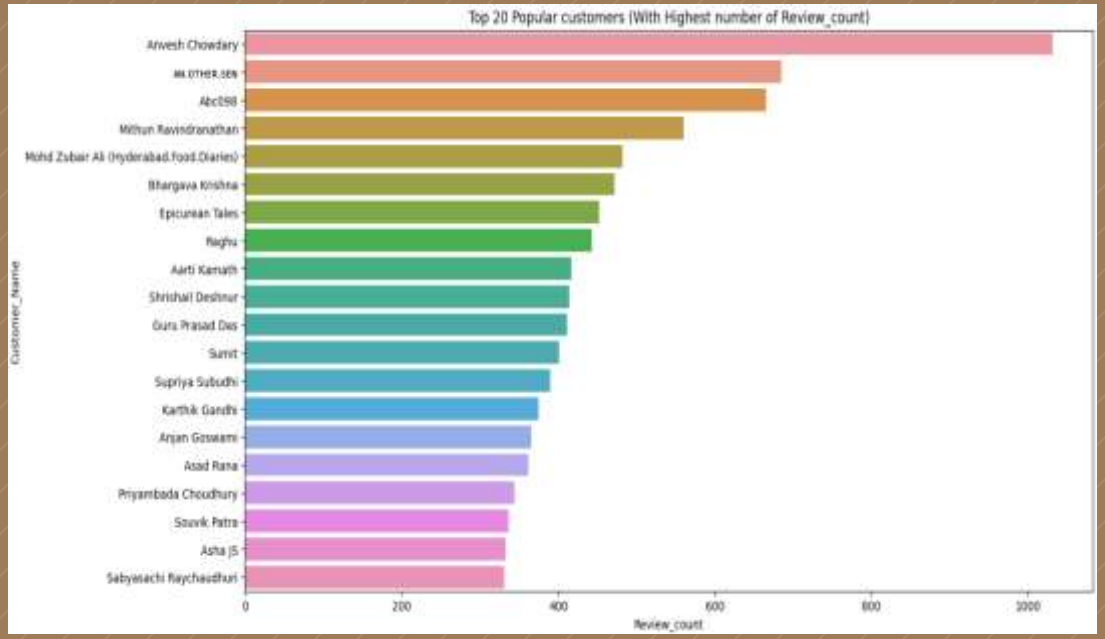
Bar Plot Analysis



Popular Restaurants Based on Review Count

Collage - Hyatt Hyderabad Gachibowli, Pista House and Labonel is the most reviewed restaurants having 1031 reviews. While La La Land - Bar & Kitchen and The Fisherman's Wharf has the least reviews.

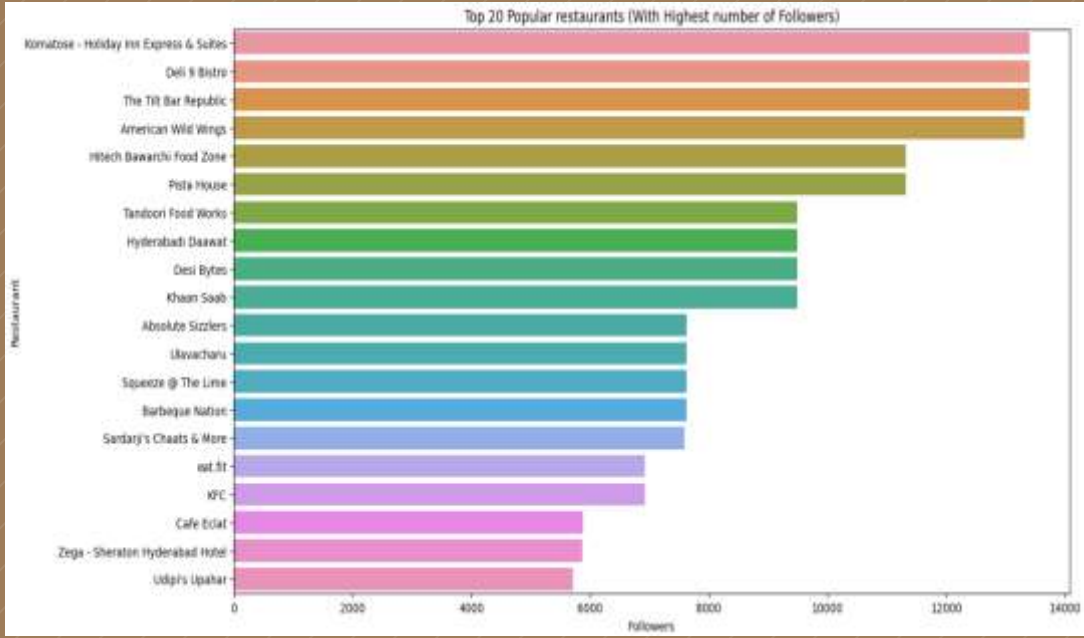
Bar Plot Analysis (Cont.)



Popular Customer Based on Review Count

Anvesh Chowdary has posted the most number of reviews nearly 1031 reviews.

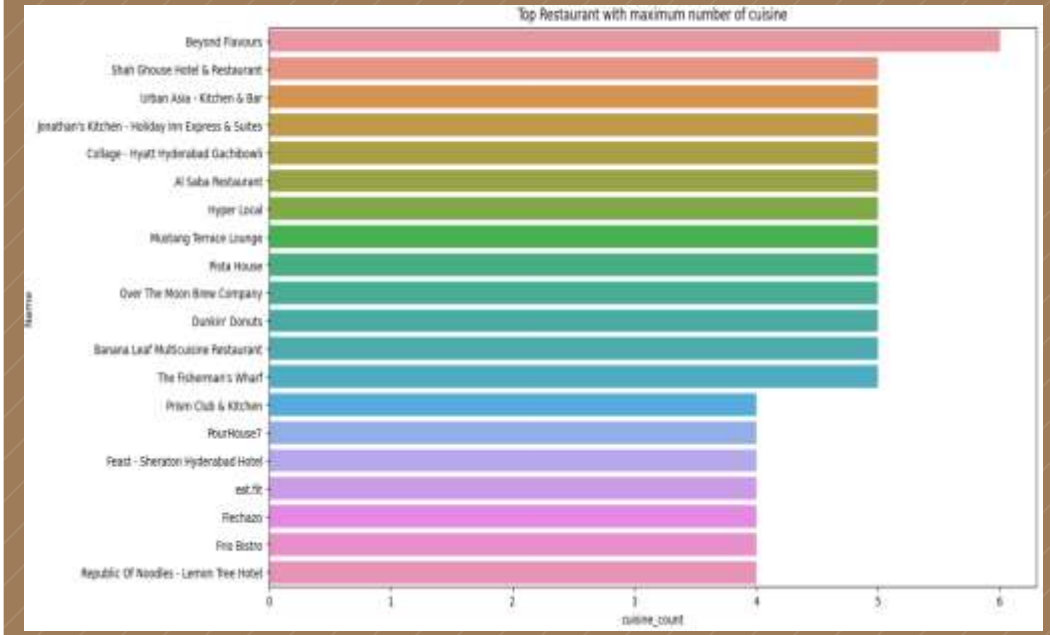
Bar Plot Analysis (Cont.)



Popular Restaurants Based on Followers

Komatose - Holiday Inn Express & Suites, Deli 9 Bistro and The Tilt Bar Republic has highest number of followers nearly of 13410, it means we can conclude that may be this restaurants will have great performance.

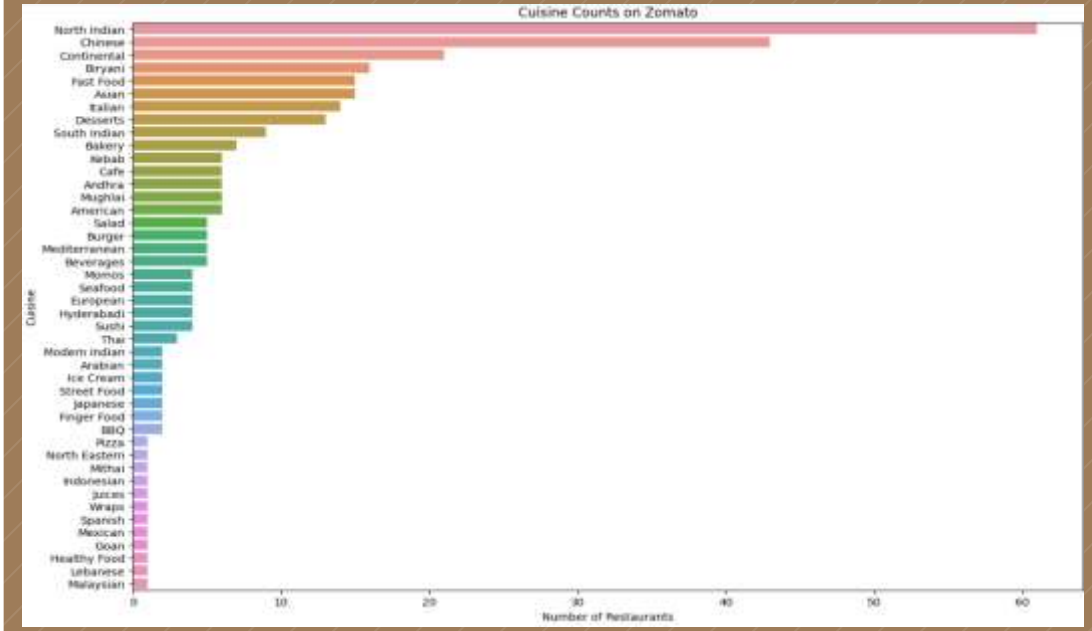
Bar Plot Analysis (Cont.)



Popular Restaurants Based on Maximum Number of Cuisine

Only Beyond Flavour has maximum numbers of cuisines available so that customers can get any type of food which they want. Other restaurant has mostly 4 and 5 types of different cuisines.

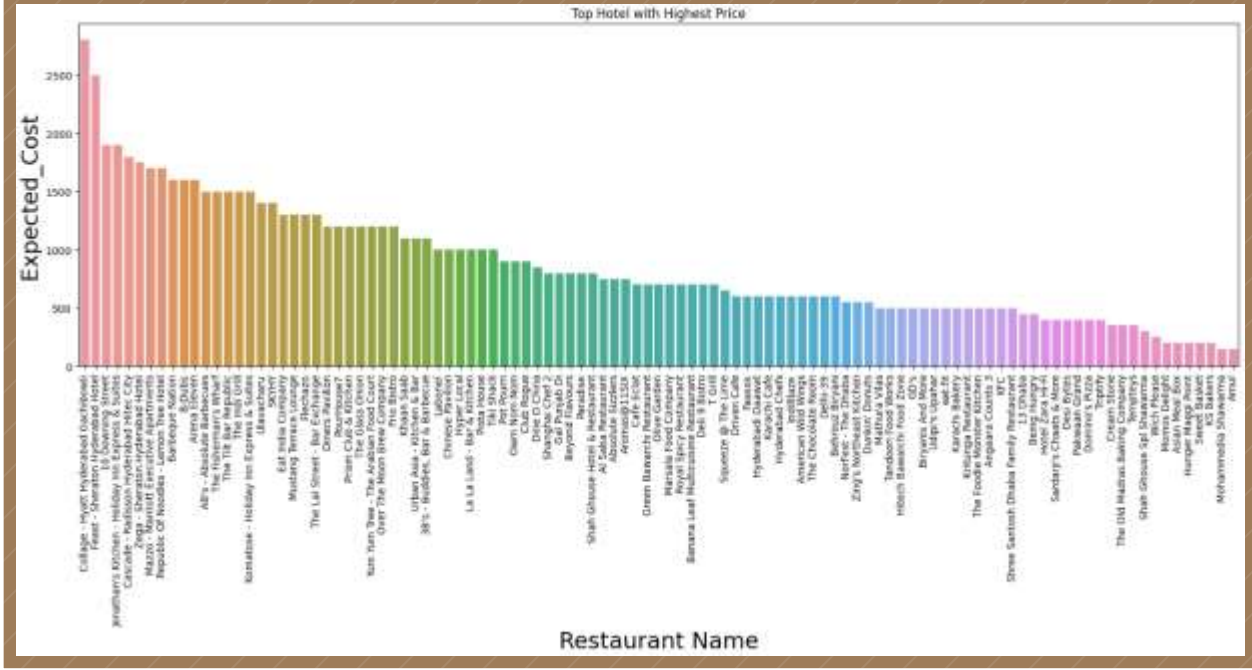
Bar Plot Analysis (Cont.)



Most Popular Cuisines on Zomato

North Indian Food is mostly available on nearly 60 number of restaurants. So we can conclude that north Indian food is mostly ordered and demanded food by the customers followed by Chinese food. while malaysian, pizza etc are the less ordered and demanded cuisines.

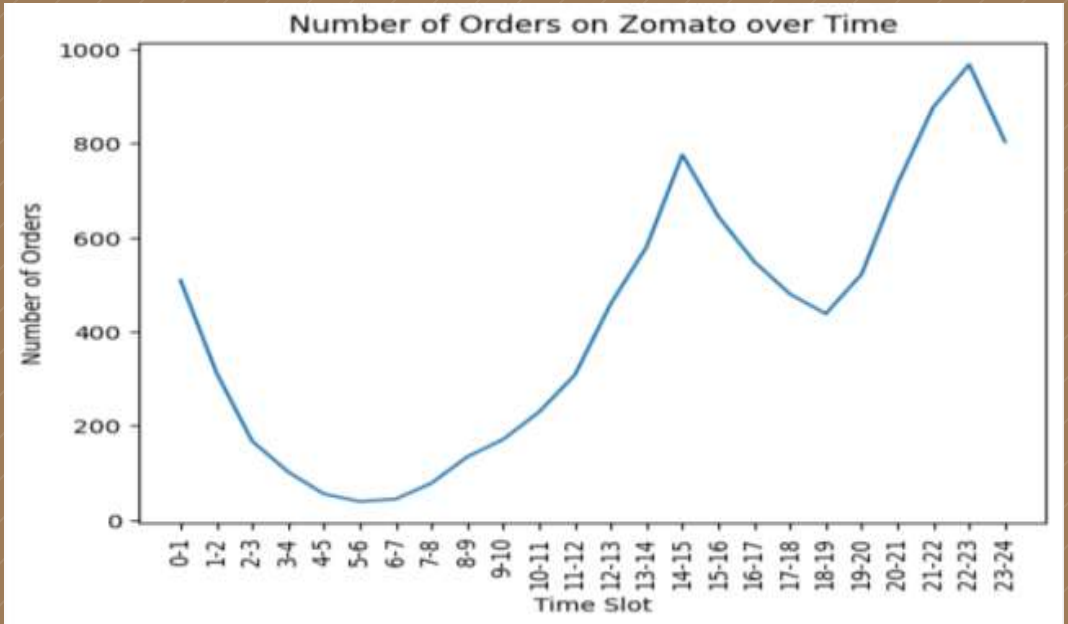
Bar Plot Analysis (Cont.)



Top Restaurants With Highest Price

Collage - Hyatt Hyderabad Gachibowli is the most expensive restaurant among all the restaurants. While Amul has the least price range.

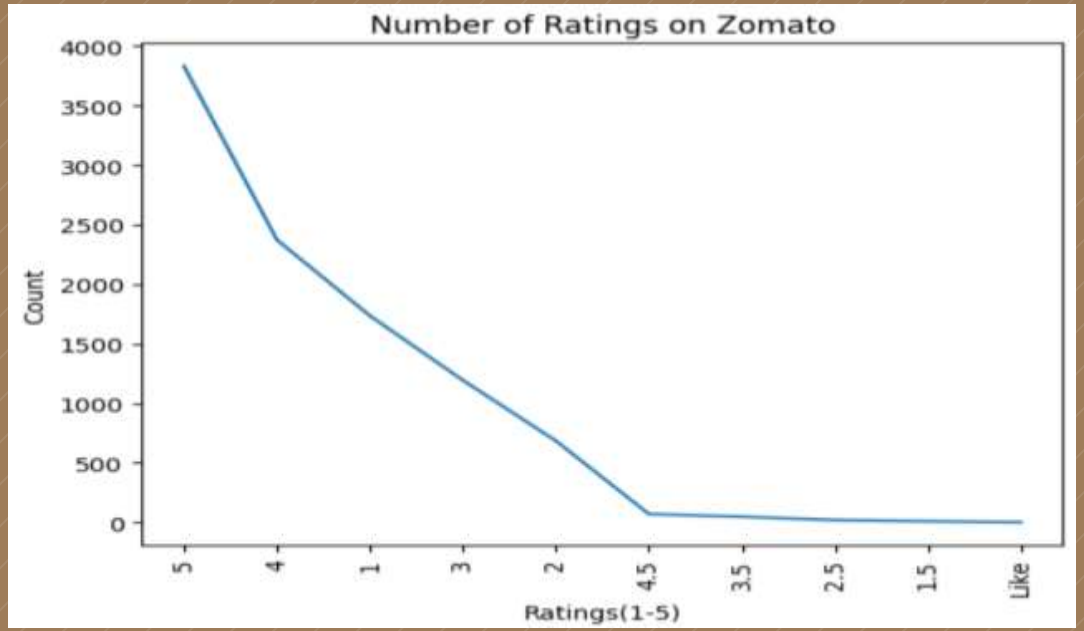
Line Plot Analysis



Orders on Zomato Over a Different Time Frame

most of the people orders food between 3PM to 4PM and 10PM to 11PM. Least number of order is between 3AM to 7AM may be because of very early morning timing.

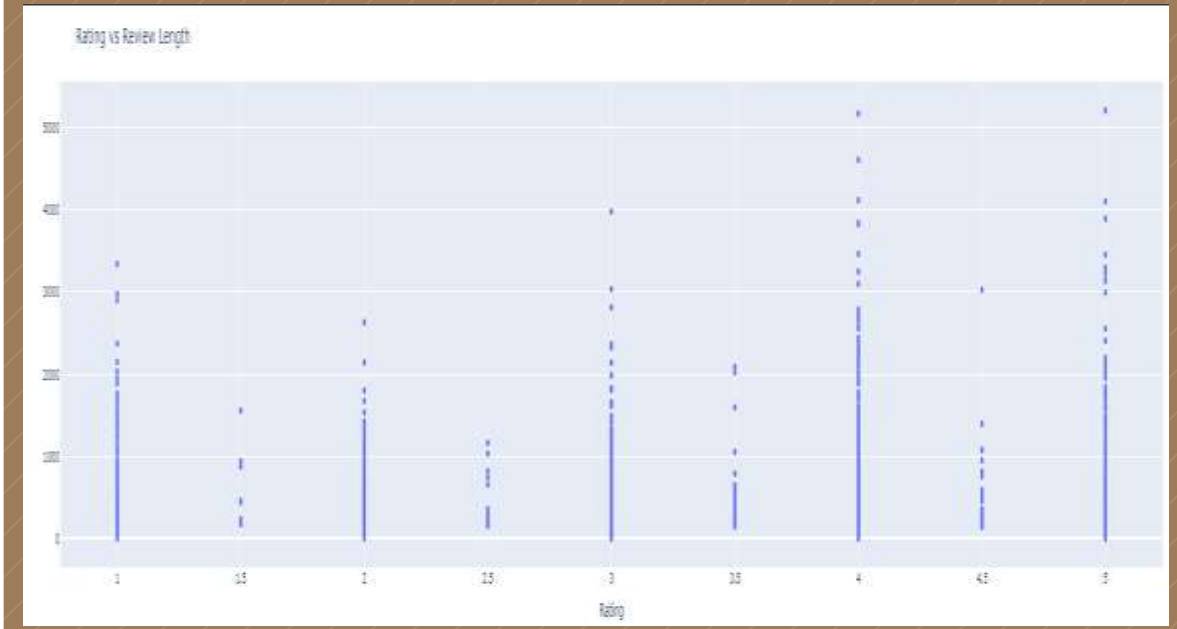
Line Plot Analysis (Cont.)



Ratings on Zomato

We can see that 5 star rating has more number of count than 4 star. Ratings 1.5, 2.5, 3.5 and 4.5 star has least and similar number of count.

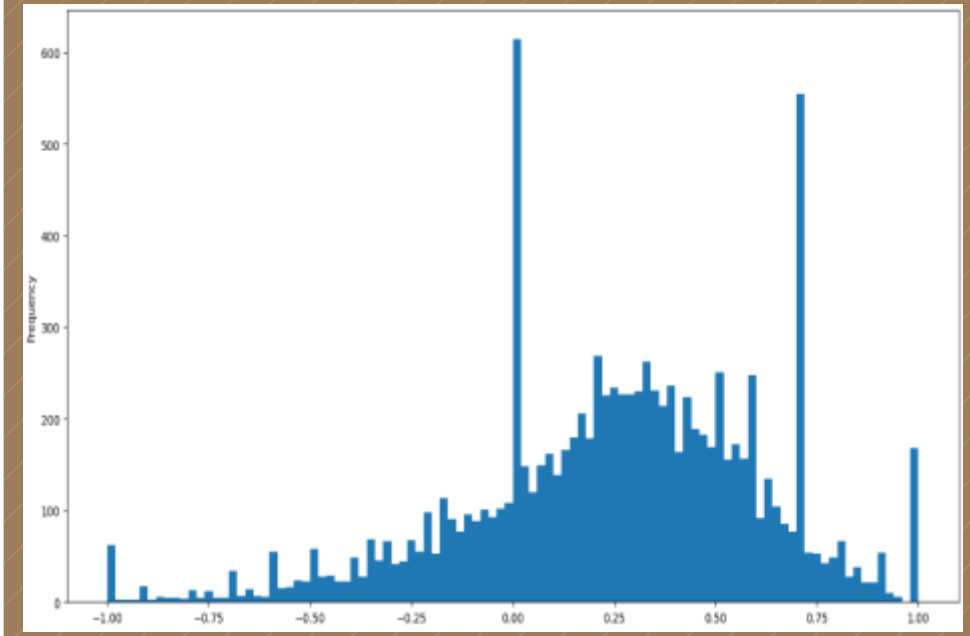
Scatter Plot Analysis



Ratings v/s Review Length

From the graph, it shows us that length of review doesn't impact ratings of the restaurants.

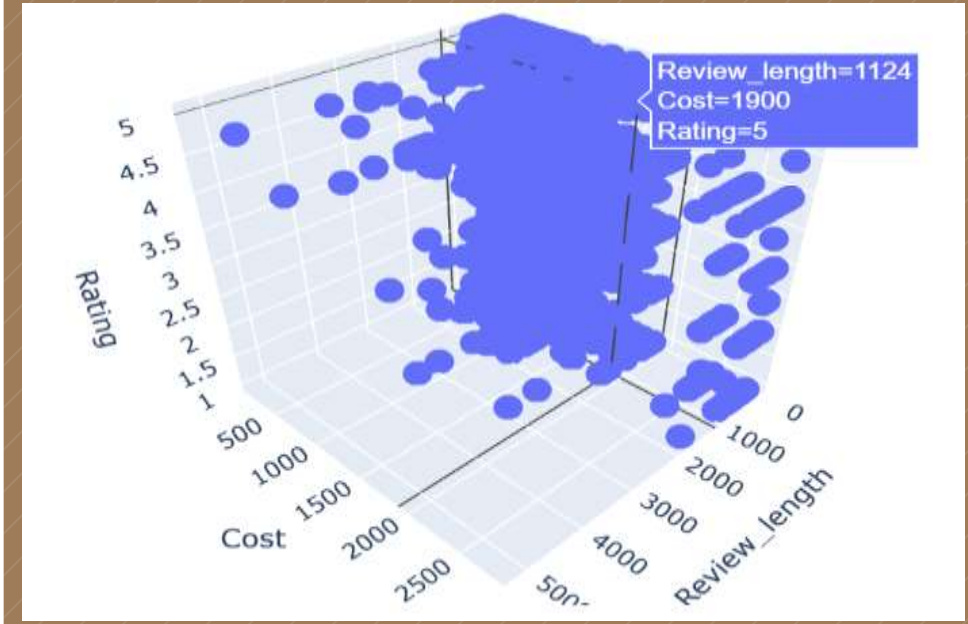
Scatter Plot Analysis (Cont.)



Sentiment Analysis on Review based on Polarity Value

- These graph shows us the majority of reviews are neutral 0, probably suggesting mixture of bad and good words in reviews.
- Also the number of positive reviews (greater than 0) are higher than negative reviews.
- More than 200 odd reviews have very high positive sentiments.

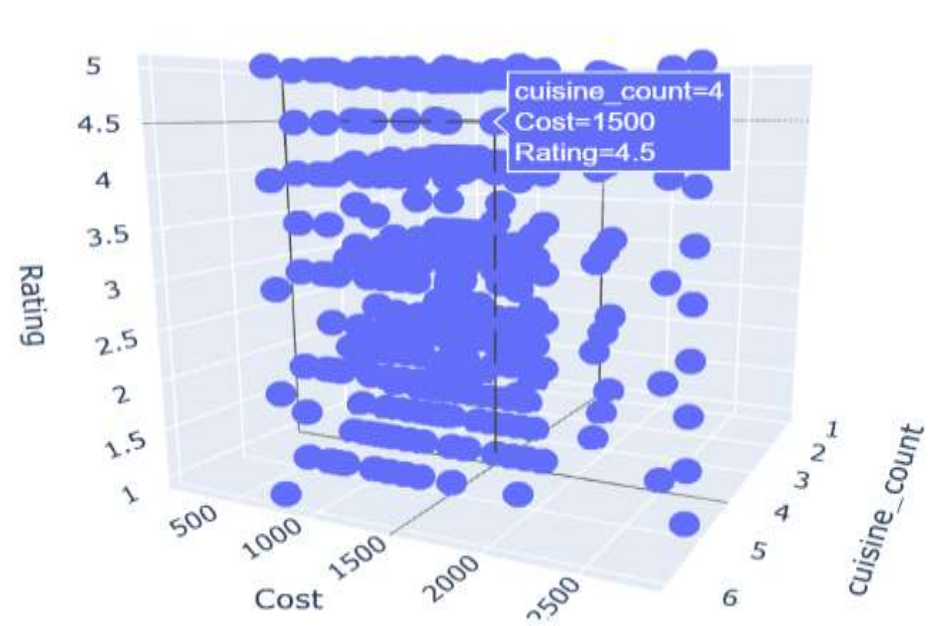
3-D Scatter Plot Analysis (Cont.)



Sentiment Analysis on Review length v/s Cost v/s Ratings

- The 3D plot says that having high number of reviews length also has high number of ratings and high number of cost.
- Average ratings has very small length of review. Number of reviews do not show much impact on ratings.
- Low rated restaurants has low price.

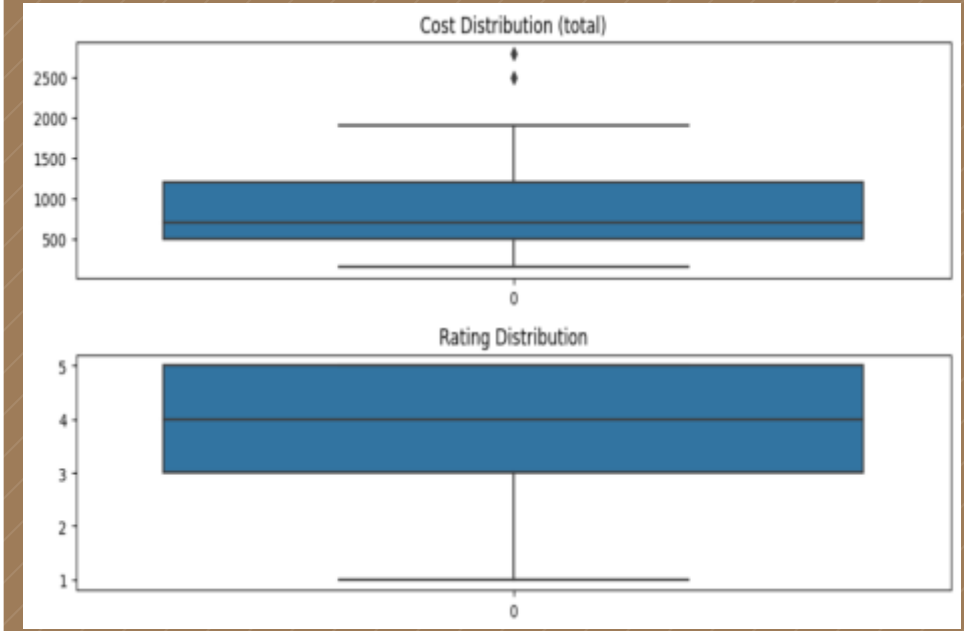
3-D Scatter Plot Analysis



Sentiment Analysis on Review length v/s Cost v/s Ratings

- These 3D plot says that cuisine count neither affecting the cost nor the ratings.
- Average cuisine count has average cost and average ratings. While low rated restaurants also has high cuisine count. This means that having more number of cuisines doesn't means the restaurant is good.

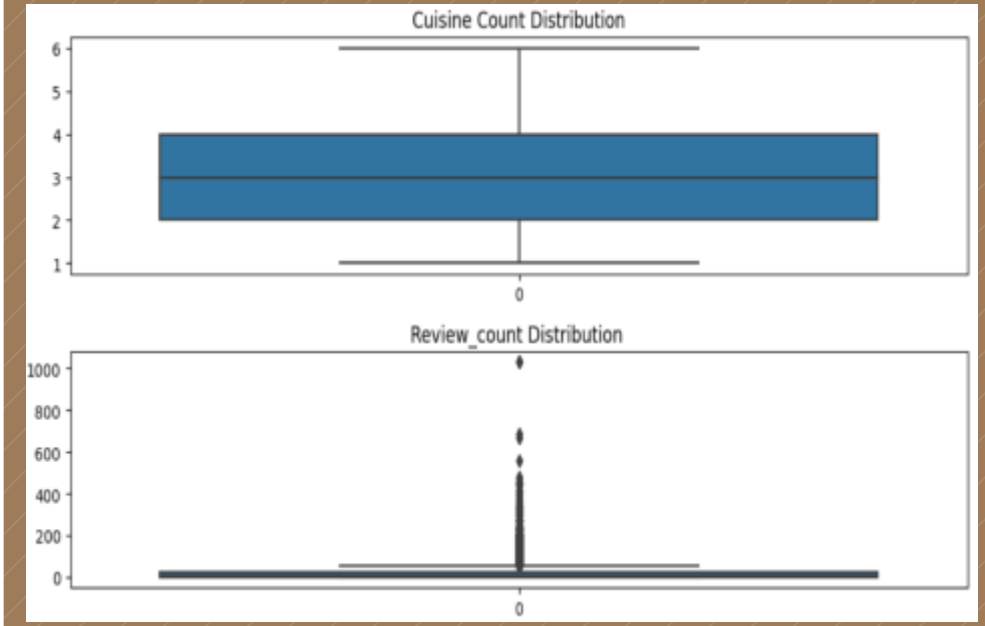
Box Plot Analysis



Outlier Detection of Cost and Rating Feature

- We found that Cost column has some values with high cost of 2500 or more for per person which shows us that there are some most expensive restaurants.
- Ratings has no outlier and has average ratings of 4 star and maximum ratings of 5 star.

Box Plot Analysis (Cont.)

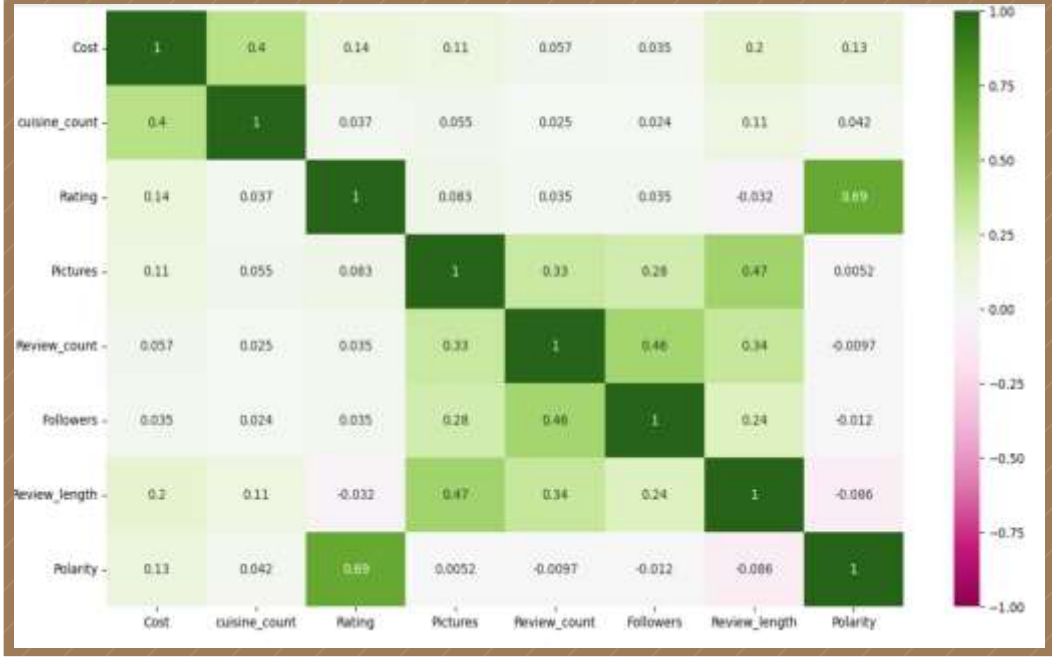


Outlier Detection of Cuisine count and Review count Feature

Maximum restaurant has a 6 types of cuisines. Average cuisine count is 3 means restaurants has on an average 3 cuisines.

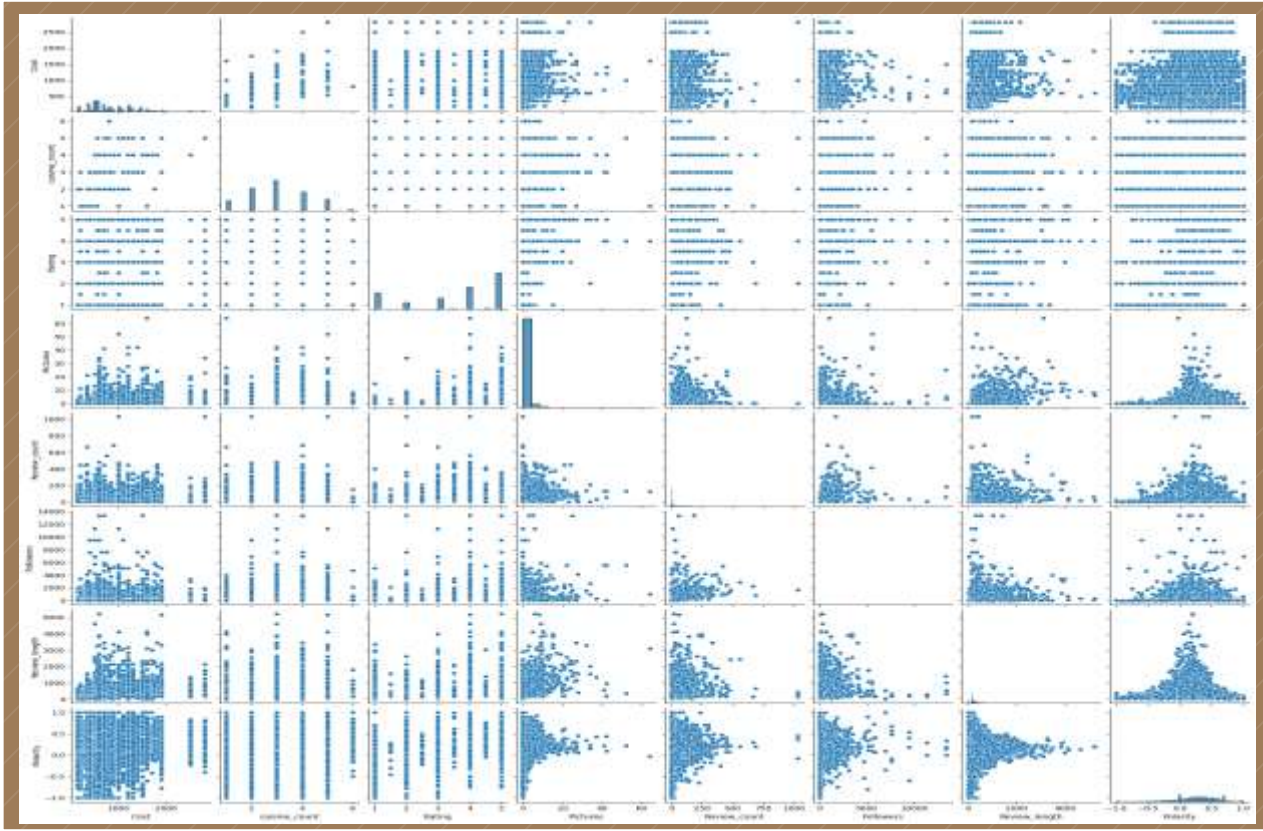
Reviews counts distribution is looking abnormal i.e. some restaurants has a more than 1000 number of reviews means may be it is a outlier.

Correlation Matrix



Relationship Between Numerical Features

- There are 46% of positive correlation b/w Review_count and Followers.
- There are positive correlation between Cuisine_count and cost i.e. the restaurant with more number of cuisine has higher cost as compared to the restaurant with less number of cuisine.
- There are also 33% of positive correlation b/w review_count and pictures.
- Polarity has highest correlation with ratings i.e.70%.



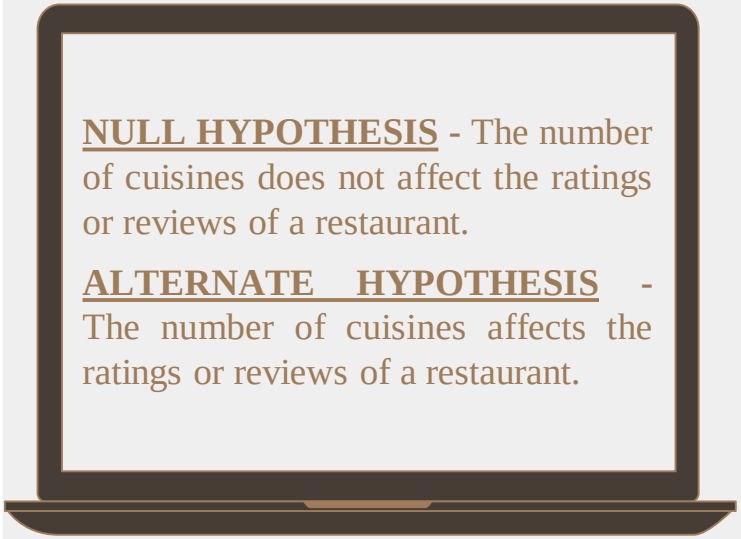
Pair Plot Representation

It can be seen that there is no significant correlation between the given features in the merged data frame.

VI. Hypothesis Testing

Let's make some assumptions which validates claim about a population based on sample data.

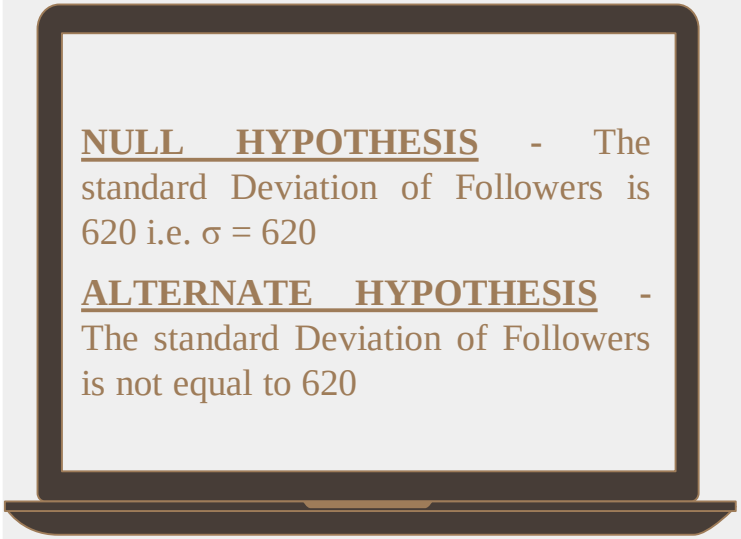
Hypothesis Test - 1



RESULT

- The one-way ANOVA test results show an F-value of 9.82 and a very small p-value of 5.51e-05.
- This means that there is evidence to reject the null hypothesis that the number of cuisines does not affect the ratings or reviews of a restaurant. So, we can conclude that the number of cuisines offered by a restaurant does have a significant effect on its ratings or reviews.

Hypothesis Test - 2



RESULT

- The calculated p-value is 0.5052, which is greater than the significance level of 0.05.
- Therefore, we fail to reject the null hypothesis that the standard deviation of followers is equal to 620. This means that there is not enough evidence to suggest that the standard deviation of followers is significantly different from 620.

Hypothesis Test - 3

NULL HYPOTHESIS - There is no significant correlation between the length of the review and the rating provided by the customer.

ALTERNATE HYPOTHESIS - There is a significant positive correlation between the length of the review and the rating provided by the customer.

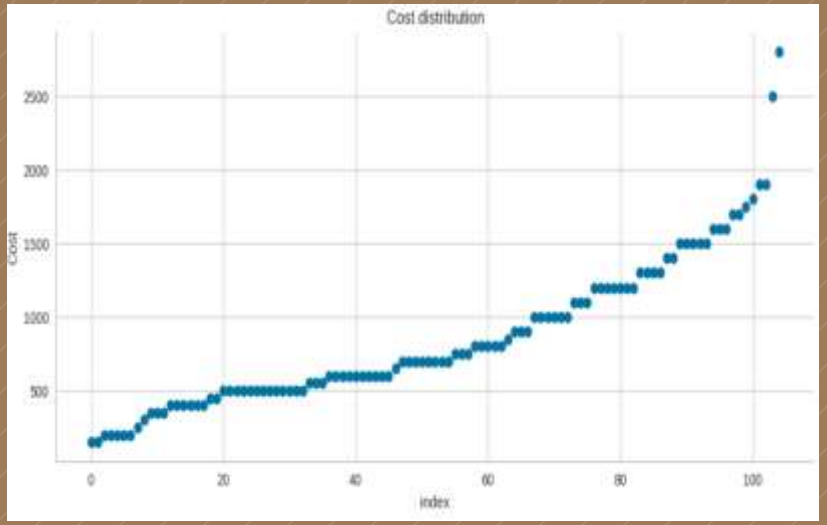
RESULT

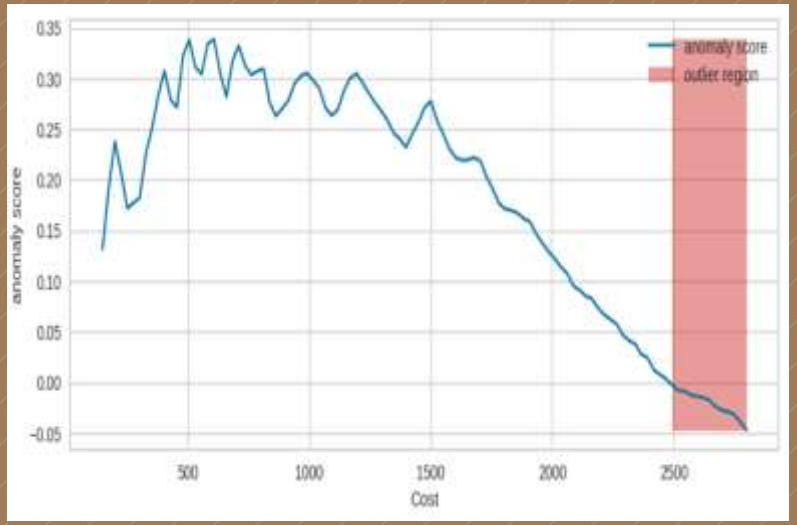
- The p-value associated with the Pearson correlation coefficient is 0.0022, which is less than the conventional threshold of 0.05. This suggests that the correlation coefficient is statistically significant, indicating that it is unlikely to have occurred by chance.
- This suggests that there is a weak, negative relationship between the length of a review and the rating of the restaurant. That is, as the length of the review increases, the rating of the restaurant tends to decrease slightly. However, the strength of this relationship is quite weak, as indicated by the small absolute value of the correlation coefficient.

VII. Feature Engineering & Textual Data Preprocessing

Let's play with some categorical features and analyze our data for further splitting and clustering

Anomaly Detection (Handling Outliers)





Cost variable is highly positively skewed with Skewness: 1.143450 and Kurtosis: 1.534478 and having some outliers also. So we have visualize the outliers using Isolation Forest Algorithm and handled it by using IQR method(Quantile based flooring and capping).

Textual Data Preprocessing

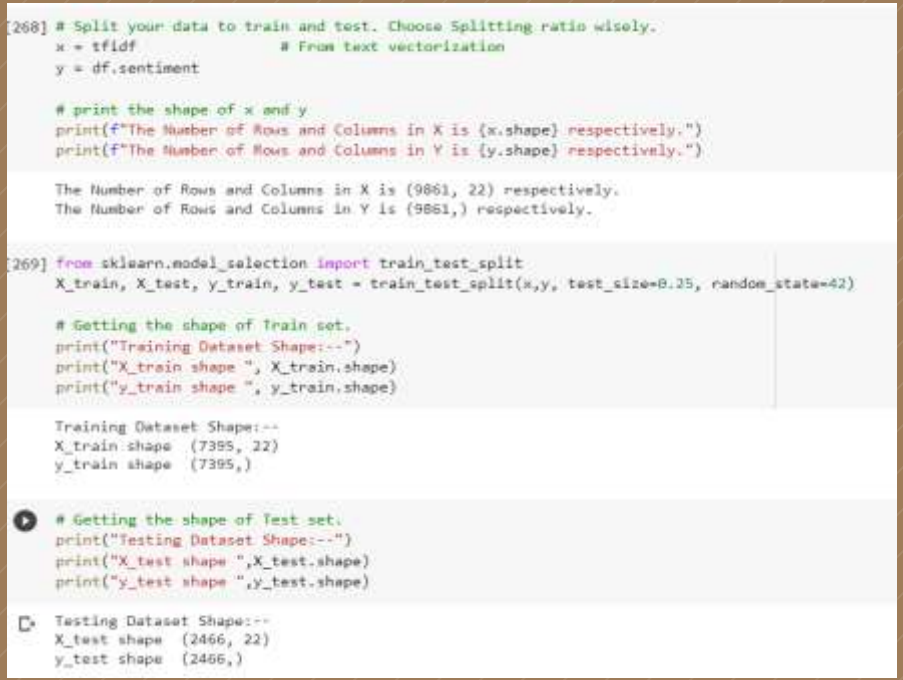
```
[166] # creating new column named polarity for the sentiment analysis
      from textblob import TextBlob
      reviews_df['Polarity'] = reviews_df['Review'].astype(str).apply(lambda x: TextBlob(x).sentiment.polarity)
      reviews_df.head(2)
```

	Restaurant	Customer_Name	Review	Rating	Time	Pictures	Review_count	Followers	Review_length	Polarity
0	Beyond Flavours	Rusha Chakraborty	The ambience was good, food was quite good . had Saturday lunch , which was cost effective .\nGood place for a sate brunch. One can also chill with friends and or parents.\nWaiter Soumen Das was really courteous and helpful.	5.0	2019-05-25 15:54:00	0	1	2	222	0.660000
1	Beyond Flavours	Anusha Tirumalaneedi	Ambience is too good for a pleasant evening. Service is very prompt. Food is good. Over all a good experience. Soumen Das - kudos to the service	5.0	2019-05-25 14:20:00	0	3	2	144	0.606667

- After completing the necessary text processing part, which includes Expanding Contraction, Removing Punctuations, Removing StopWords, Tokenization, Text Normalization (Lemmatization), Text Vectorization (Tfidf Vectorizer) and POS tagging. We moved towards **Sentiment Analysis**.
- We have created new variable called **Polarity** through which we can just determine the reviews as positive, negative and neutral.
- Polarity score in sentiment analysis is nothing but a float value which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement.

Data Splitting

We have made the use of 75% and 25% Training and Testing data respectively. The data should be divided in such a way that neither of them is too high, which is more dependent on the amount of data we have. If the data is too small, then no split will give satisfactory variance so we'll have to do cross-validation, but if the data is huge then it doesn't really matter whether we choose an 80:20 split or a 75:25 split (indeed we may choose to use less training data as otherwise, it might be more computationally intensive).

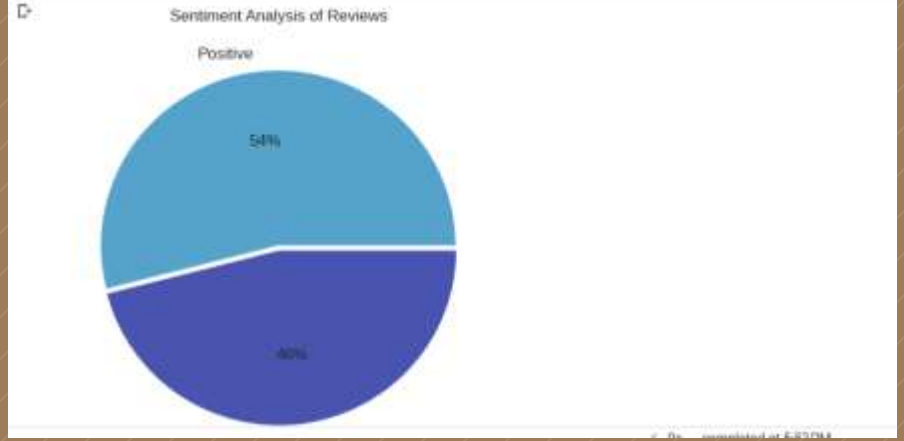


```
[268] # Split your data to train and test. Choose Splitting ratio wisely.
      x = tfidf                    # From text vectorization
      y = df.sentiment

      # print the shape of x and y
      print(f"The Number of Rows and Columns in X is {x.shape} respectively.")
      print(f"The Number of Rows and Columns in Y is {y.shape} respectively.")

      The Number of Rows and Columns in X is (9861, 22) respectively.
      The Number of Rows and Columns in Y is (9861,) respectively.

[269] from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(x,y, test_size=0.25, random_state=42)

      # Getting the shape of Train set.
      print("Training Dataset Shape:--")
      print("X_train shape ", X_train.shape)
      print("y_train shape ", y_train.shape)

      Training Dataset Shape:--
      X_train shape  (7395, 22)
      y_train shape  (7395,)

      # Getting the shape of Test set.
      print("Testing Dataset Shape:--")
      print("X_test shape ",X_test.shape)
      print("y_test shape ",y_test.shape)

      Testing Dataset Shape:--
      X_test shape  (2466, 22)
      y_test shape  (2466,)
```

Data Imbalanced

As our most important variable is in the 50:50 ratio. However it is said to be very slightly imbalanced. therefore not applying any technique (over sampling & undersampling).


```
# Handling Imbalanced Dataset (If needed)
fig = plt.figure(1, figsize=(8,6))
keys = ['Positive','Negative']
data= df['sentiment'].value_counts()
explode = [0.01, 0.02]
palette_color = sns.color_palette('icefire')

plt.pie(data, labels=keys, colors=palette_color, explode=explode, autopct='%.0f%%',textprops={'fontsize': 12})
plt.title('Sentiment Analysis of Reviews')
plt.show()
```



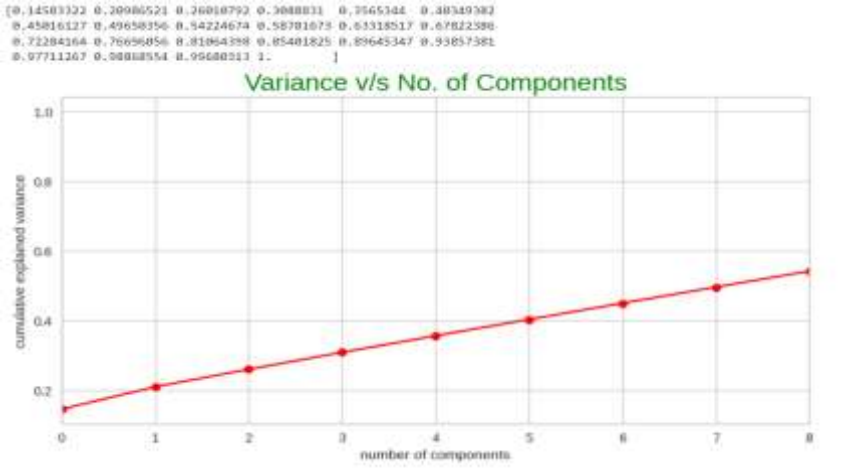
VII. Model Implementation

Let’s implement the model and find the best model for the data

Model Performed

- Principal Component Analysis
- K-Means Clustering
- Hierarchical Clustering (Agglomerative)
- DBSCAN
- Content Based Recommendation System

Model 1 - Principal Component Analysis




```
print(pca.explained_variance_ratio_)
```

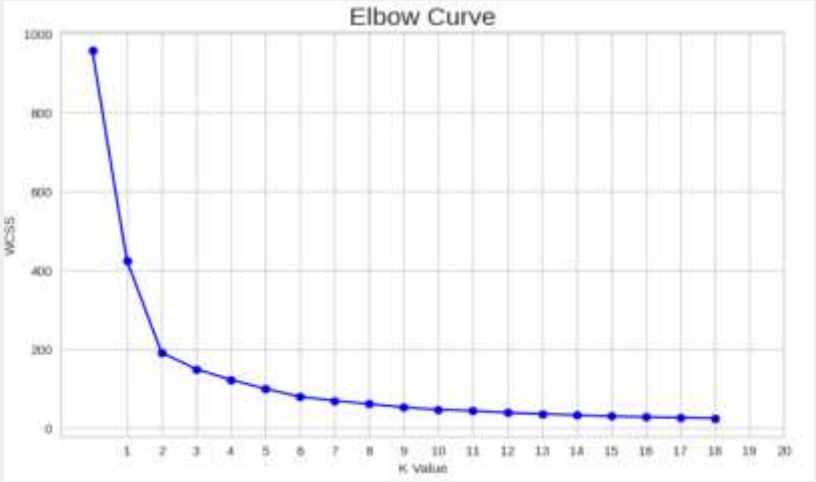
```
[0.14583093 0.06371601 0.04967964]
```

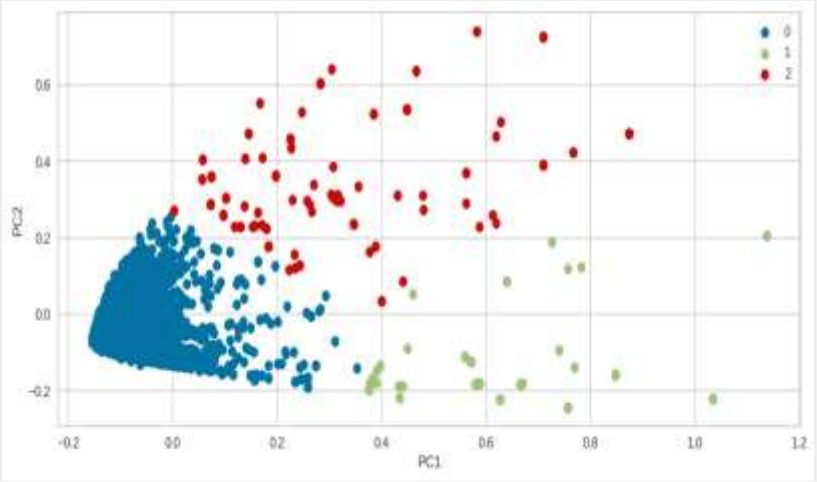
```
[266] print("original shape: ", df.shape)
      print("PCA transformed shape:", principalComponents.shape)
```

```
original shape:  (9861, 55)
PCA transformed shape: (9861, 3)
```

As we can see above graph, it is showing little abrupt change at point 1 then constant line as well as we can see the explained_variance_ratio_, it is showing the relationship b/w variance and no. of components. So here we can form n_components = 1 or 2 or 3 principal components here.

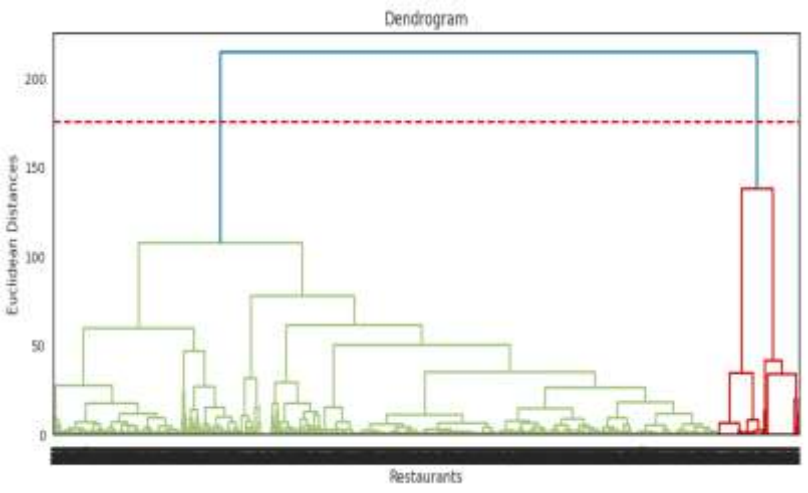
Model 2 - K-Means Clustering

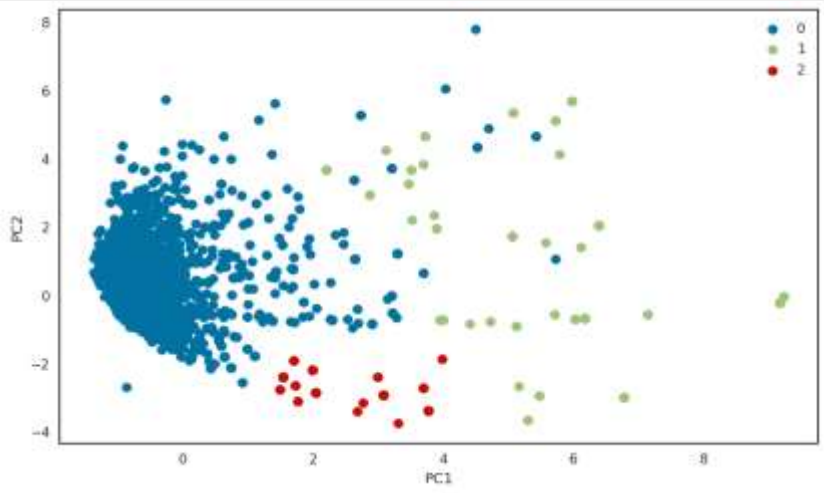




- According to the elbow curve, we should have 2 cluster because 2nd cluster is forming the elbow. However, the values does not reduce linearly until 6th cluster.
- After k-means clustering, we found the result with the score of Silhouette Coefficient: 0.640 and Distortion Score: 20978.197

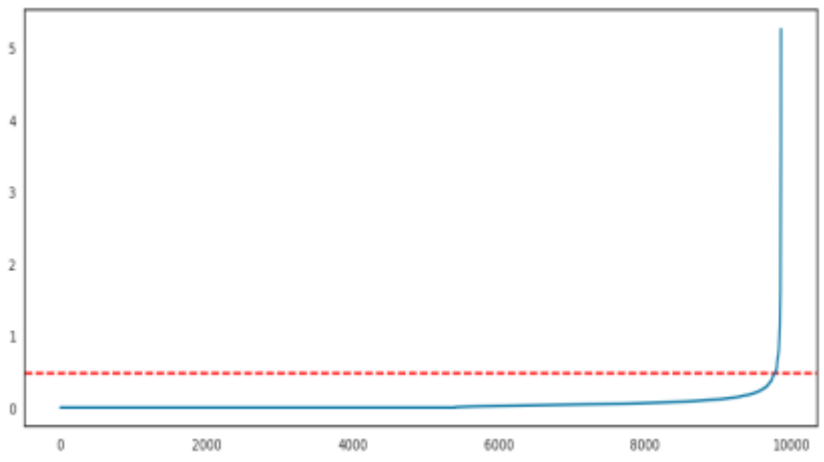
Model 3 - Hierarchical Clustering (Agglomerative)

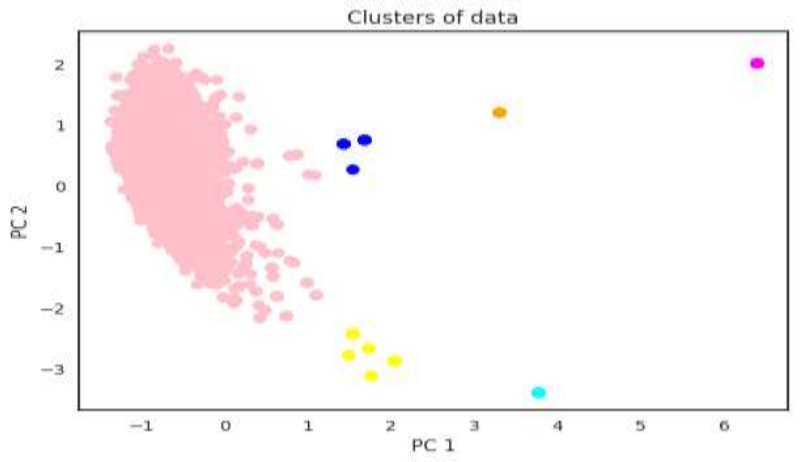




- According to the dendrogram, we should have 2 cluster because we try to set the threshold in such a way that it cuts the tallest vertical line. However, we got the 2 point that cuts the line. So we got n_cluster = 2.
- Another graph is showing the performance of the 3 clusters on the dataset using agglomerative clustering.

Model 4 - DBSCAN Clustering





- According to the line plot, the optimal value for epsilon will be found at the point of maximum curvature. So we train our model, selecting 0.48 for eps using the NearestNeighbors and setting min_samples to 50. And found the number of clusters to be 6.
- Another graph is showing the performance of the 6 clusters on the dataset using DBSCAN clustering method.

Model 5 – Content Based Recommendation System

```
[262] give_recommendation('Karachi Cafe')
```

TOP 10 RESTAURANTS LIKE Karachi Cafe WITH SIMILAR REVIEWS:

	Cuisines	Rating	Cost
The Chocolate Room	Cafe, Desserts	5.0	600
Shah Ghouse Spl Shawarma	Lebanese	5.0	300
Hotel Zara Hi-Fi	Chinese, North Indian	5.0	400
Owm Nom Nom	Chinese, Biryani, Andhra, North Indian	5.0	900
Delhi-39	North Indian, Chinese	5.0	600
The Foodie Monster Kitchen	American, Wraps, Desserts	5.0	500
KS Bakers	Bakery, Desserts, Fast Food	4.5	200
T Grill	North Indian, Andhra, Biryani, Chinese	4.5	700
Driven Cafe	Cafe, Fast Food, Beverages	4.0	600
The Lal Street - Bar Exchange	North Indian, Italian, Finger Food	4.0	1300

```
[276] give_recommendation('chinese Pavilion')
```

TOP 10 RESTAURANTS LIKE 10 WITH SIMILAR REVIEWS: Chinese Pavilion

	Cuisines	Rating	Cost
Flechazo	Asian, Mediterranean, North Indian, Desserts	5.0	1300
Urban Asia - Kitchen & Bar	Asian, Thai, Chinese, Sushi, Momos	5.0	1100
Kritunga Restaurant	Andhra, Biryani, Hyderabadi, North Indian	5.0	500
Absolute Sizzlers	Continental, American, Chinese	5.0	750
Beyond Flavours	Chinese, Continental, Kebab, European, South Indian, North Indian	5.0	800
Marsala Food Company	Arabian, North Indian, Chinese	5.0	700
Feast - Sheraton Hyderabad Hotel	Modern Indian, Asian, Continental, Italian	4.0	2250
La La Land - Bar & Kitchen	Finger Food, North Indian, Kebab, Chinese	4.0	1000
The Fisherman's Wharf	Seafood, Goan, North Indian, Continental, Asian	4.0	1500
Shanghai Chef 2	Chinese, Thai, Indonesian, Asian	3.0	800

A content-based recommendation system is an AI-driven approach to personalized recommendations that filters items based on their characteristics or content. It recommends items to users based on what they have previously consumed, or what items are similar. It makes use of natural language processing (NLP) and machine learning algorithms to analyze the content of items and recommend new items to users.

IF's & BUT's

Que1. Which Evaluation metrics did we consider for a positive business impact and why ?
Ans. Silhouette score will be the best evaluation metrics to select the model. Because it measures how well each data point is matched to its own cluster compared to other clusters, providing a range from -1 to 1 with a score of 1 indicating a perfect match and -1 indicating the worst possible match. A higher score indicates a better clustering.

Que2. Which ML model did we choose from the above created models as our final prediction model and why ?
Ans. DBSCAN (Density-Based Spatial Clustering of Applications and Noise) is the most suitable algorithm for our project because restaurants in a given area may have varying densities, and the number of restaurants may not be known in advance. DBSCAN can discover the optimal number of clusters by detecting regions of high density and identifying clusters as regions with a high density of points.

VIII. Advantages & Future Work

Let’s see some excited competitive advantages and future work

Advantages

- AI technology can be leveraged to analyze customer reviews and ratings, helping to quickly identify patterns regarding customer satisfaction, dissatisfaction and specific customer trends.
- Automated customer review analysis can provide quick and accurate feedback from customers, helping to identify potential problems in the restaurant's service delivery or customer experience.
- AI-driven analysis of customer reviews can be further used to make improvements to menus, pricing, services and quality of food.
- Restaurateurs can get valuable insights into which dishes and ingredients are most popular and this in turn can help create more appealing menus and offer better value for customers.
- AI also helps restaurants target customers with offers and discounts based on their priorities.

Future Work

- **AI-driven algorithms** can be used to automatically generate summary reports of restaurant reviews in various languages and identify common trends in customer feedback.
- **Sentiment analysis** can be used to identify what people really think about a restaurant based on reviews. This can help customers determine which restaurants are worth their time, and which ones should be avoided.
- Utilize **computer vision** techniques to identify objects and classify food items in restaurant photos.
- Use **deep learning** algorithms to compare reviews between two different restaurants and generate comparison results.
- **Online Learning**: Currently, the system is trained on a batch of historical data. Future work can involve developing an online learning system that can adapt to the changing preferences of the users in real-time.
- **Multi-language support**: Zomato is a global platform and supports multiple languages. Future work can involve developing a multi-language content-based recommendation system that can handle different languages and provide recommendations in the user's preferred language.

IX. Conclusion

Let’s gets end of the project

Conclusion

The conclusion of this Zomato restaurant clustering and metadata sentiment analysis project is that it is possible to use natural language processing and machine learning algorithms to build a model that can accurately cluster restaurants based on their reviews and sentiments. This project has helped identify customer preferences and restaurant ratings in order to better understand the impacts of customer feedback on the restaurant industry. This model can then be used to improve the decision-making process of a restaurant owner or manager in terms of advertising, pricing, customer acquisition, and other important business decisions. With this data, business owners can make more informed decisions about the quality of their restaurants and better understand the customer experience. We have seen that AI-based solutions provide a powerful tool for business owners to gain insight into the performance of their restaurants. After that sentiment analysis can be used to gain insights into customer preferences, providing data-driven understanding into how customers perceive different aspects of service quality. Furthermore, the insights provided by this project can be used for further business growth strategies.

Thanks!

Any questions?

You can find us at:
@https://github.com/riyapatelrp
@https://github.com/v1git12