A Project Report
on

Zomato Insights Hyderabad: Leveraging Sentiment Analysis and Clustering for Improved Restaurant Recommendations and Boost Business Growth

Work carried out for
Brightpoint Infotech Pvt. Ltd., Chennai, Tamilnadu.

Submitted to Pondicherry University in partial fulfillment of the requirement for the Award of the Degree of

Master of Business Administration

in

Data Analytics

By

RAMESH KUMAR M
(Reg. No. 21401031)

Under the Supervision of

Dr. B. RAJESWARI, ***Associate Professor***
Department of Management Studies, Pondicherry University

&

Mr. NAVIN MIRPURI, ***Founder***,
Brightpoint Infotech Pvt. Ltd.

Department of Management Studies
Pondicherry University, Pondicherry, INDIA – 605 014
Nov 2022 – Jan 2023

DEPARTMENT OF MANAGEMENT STUDIES
SCHOOL OF MANAGEMENT
PONDICHERRY UNIVERSITY
PONDICHERRY-605014

CERTIFICATE

This is to certify that this project report entitled **“Zomato Insights Hyderabad: Leveraging Sentiment Analysis and Clustering for Improved Restaurant Recommendations and Boost Business Growth”** done for **Brightpoint Infotech Private Limited** is submitted by **Ramesh kumar M (Reg.No:21401031),** II MBA (DA) to the **DEPARTMENT OF MANAGEMENT STUDIES, SCHOOL OF MANAGEMENT, PONDICHERRY UNIVERSITY** in partial fulfilment of the requirements for the award of the degree of **MASTER OF BUSINESS ADMINISTRATION IN DATA ANALYTICS** and is a record of an original and bonafide work done under the guidance of **Dr. B. Rajeswari**, Associate Professor, Department of Management Studies, Pondicherry University. This report has not formed the basis for the award of any degree, diploma, associateship, fellowship or other similar title to the candidate and that the report represents an independent and original work on the part of the candidate.

Dr. B. RAJESWARI
Associate Professor
Department of Management Studies

Dr. B. CHARUMATHI
Professor and Head
Department of Management Studies

Date:
Place: Pondicherry 605 014

DECLARATION

I hereby declare that the project titled, **“Zomato Insights Hyderabad: Leveraging Sentiment Analysis and Clustering for Improved Restaurant Recommendations and Boost Business Growth”** is original work done by me under the guidance of Dr. B. Rajeswari, Associate Professor, Department of Management Studies, Pondicherry University, and Mr. Ankur Arora, Senior Functional Consultant, Brightpoint Infotech Pvt. Ltd. This project or any part thereof has not been submitted for any Degree / Diploma / Associateship / Fellowship / any other similar title or recognition to this University or any other University.

I take full responsibility for the originality of this report. I am aware that I may have to forfeit the degree if plagiarism has been detected after the award of the degree. Notwithstanding the supervision provided to me by the Faculty Guide, I warrant that any alleged act(s) of plagiarism in this project report are entirely my responsibility. Pondicherry University and/or its employees shall under no circumstances whatsoever be under any liability of any kind in respect of the aforesaid act(s) of plagiarism.

Ramesh kumar m
21401031
II MBA
Pondicherry University

Place: Pondicherry 605 014

Date:

Acknowledgements

*I thank my research guide **Dr. B. Rajeswari**, Department of Management Studies, Pondicherry University, Pondicherry for his incessant encouragement and support extended throughout the research period.*

*Thanks to **Mr. Ankur Arora**, Brightpoint Infotech, for sharing and discussion of various research related matters.*

Many thanks to those faculty members who helped me in sharpening my thinking by cheerfully providing challenging comments and questions.

Non-teaching staff of the Department of Management Studies, Pondicherry University was extremely helpful to me during the research period and therefore, I am thankful to them.

I thank my classmates of Pondicherry University who had given me moral support.

Foremost, I am grateful to my family who had to tolerate late night work, and curtailed weekends & vacations. The support and encouragement of my parents gave me the energy, stamina and inspiration to complete the project.

Ramesh kumar m

BRIGHTPOINT
INFOTECH

05th January,2023

To,
Mr. Ramesh Kumar M

Certificate of Experience

This letter is to certify that **Ramesh Kumar M** has successfully completed his internship program for a period of **3 months** with our organization where he worked as a Business Analyst SCM Intern.

This internship tenure was from **October 2022** to **December 2022.** The said intern was working with our **SCM Department of Business Central** and was actively, diligently, and sincerely involved in the projects and tasks assigned to him. During this internship, we found the intern to be a punctual and hard-working person and wish him success in all his future endeavors.

Yours faithfully,
For **Brightpoint Infotech Private Ltd.,**

Chandiramani

Harsha Chandiramani
HR & Finance Director

Executive Summary

In the food industry, online platforms like Zomato provide customers with a wealth of information about local restaurants, but the sheer volume of data can be overwhelming. By leveraging clustering and sentiment analysis techniques, we can make sense of this data and provide more targeted restaurant recommendations to customers, as well as help businesses identify areas where they can improve their offerings and grow their customer base.

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 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.

Table of Contents

Chapter – I

INTRODUCTION

INTRODUCTION:

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.

Clustering and sentiment analysis are two powerful techniques in the field of data science that have become increasingly popular in recent years. Clustering is a machine learning technique used to group similar data points together based on their characteristics and attributes, while sentiment analysis is a natural language processing technique used to determine the underlying sentiment or emotional tone of a text or speech.

The combination of clustering and sentiment analysis can provide valuable insights into various areas of research, such as customer behavior, product development, and business strategies. By clustering similar data points, businesses can identify patterns and trends that are not immediately apparent, and by performing sentiment analysis on customer reviews and feedback, businesses can gain a deeper understanding of customer sentiment and satisfaction.

About the Company:

Brightpoint Infotech is a leading Enterprise & Business Solutions consulting firm headquartered at Fort Lauderdale, Florida. Brightpoint Infotech is an Asian-American minority-owned firm and, is a Microsoft Dynamics Gold Certified Partner & Direct Microsoft Cloud Solution Provider (CSP). Brightpoint Infotech is a prestigious member of the select few clubs of CSP and is authorized to sell Microsoft Dynamics & Cloud products & solutions across the globe. Brightpoint Infotech has a short but a rich legacy of over 15+ years in implementing Dynamics ERP and CRM solutions for small and medium enterprises. The exemplary leadership of Brightpoint Infotech has set a vision to be a Top 10 Microsoft Dynamics Consulting partner within 3 years.

At Brightpoint Infotech each of our service offerings is aligned to enable our client a successful journey of Digital transformation (ERP, CRM, Interactive Portals, Analytics). For more than a decade, Brightpoint Infotech has been expanding its global footprint with more than 100 Full-scale implementations, Upgrades and through providing enhancement support to its customers to scale up their operations. We have successfully completed implementation projects in North America, Middle East, South-East Asia, India, and Africa regions. With our operations spread all over the world, we can offer support services round the clock. Brightpoint Infotech has a global delivery center based out of India, which enables it to scale up quickly and deliver timely and cost-effective solutions and services to its customers across the globe.

Mission – “Our renowned coaching programs will allow you to: Work fewer hours — and generate more revenue.”

- Captivate and retain every esteemed customer(s)

- Manage your time so you’ll get more done in less time

- Hone sharp leadership skills to manage your team

- Cut expenses without sacrificing quality

- Automate your business, so you can leave for days, weeks, or even months at a time

Vision – “We are a team of passionate people whose goal is to improve everyone’s life through exceptional services and disruptive products. We provide professional services and build great products to solve your business problems. Along the way, we strive to empower our resources by incorporating our three Ss’: Safety, Stability, and Skillsets.

Chapter – II

Literature Review

Literature Review

In recent years, the rise of online platforms for restaurant reviews has revolutionized the way people make dining decisions. Zomato is one such platform that allows users to search for restaurants and read reviews written by other users. However, with millions of reviews on the platform, finding the right restaurant can be a daunting task. This is where sentiment analysis and clustering come in.

Clustering and sentiment analysis are two important techniques used in data science and machine learning. Clustering is a technique used to group similar data points together, while sentiment analysis is used to extract and classify opinions expressed in text data.

In recent years, clustering and sentiment analysis have been widely used in various fields such as social media analysis, customer segmentation, and product recommendations. Here are some examples of their applications:

- **Social Media Analysis:** Social media platforms generate a vast amount of data every day. Clustering is used to group similar users together based on their interests and behaviour, while sentiment analysis is used to extract opinions expressed in user-generated content. For example, sentiment analysis can be used to track customer sentiment towards a particular brand on social media.

- **Customer Segmentation:** Clustering is commonly used in marketing to segment customers based on their behaviour, demographics, and preferences. This enables companies to target specific customer groups with tailored marketing messages. For example, clustering can be used to group customers based on their purchase history, website behaviour, and other attributes.

- **Product Recommendations:** Clustering is also used in recommendation systems to group similar products together. This helps in making personalized recommendations to users based on their previous purchases and preferences. Sentiment analysis is used to extract feedback from customers on specific products and services, which can be used to improve the recommendation system.

- **Image and Video Analysis**: Clustering and sentiment analysis are not limited to text data. They can also be applied to image and video data. Clustering can be used to group similar images and videos together, while sentiment analysis can be used to extract emotions expressed in visual content.

Overall, clustering and sentiment analysis are powerful techniques that can be used in various applications to gain insights from data. With the increasing amount of data generated every day, these techniques are becoming more important for businesses to stay competitive and make data-driven decisions.

Chapter – III

OBJECTIVE

OBJECTIVE:

The main objective of this project is to leverage sentiment analysis and clustering techniques to provide improved restaurant recommendations to customers on the Zomato platform, as well as to help the company identify areas for improvement and growth in their business strategies. Specifically, the project aims to:

- Develop and apply sentiment analysis algorithms to extract and analyse the opinions, attitudes, and emotions expressed in Zomato user reviews.
- Use clustering algorithms to group Zomato restaurants based on shared characteristics, such as cuisine, price range, location, and customer ratings.
- Utilize the results of sentiment analysis and clustering to generate personalized restaurant recommendations for Zomato users, and to identify areas where the company can improve its offerings and business strategies.

Chapter – IV

OVERVIEW OF DATA

DATASETS DESCRIPTION:

In this project, the real-world datasets provided by Brightpoint Infotech Pt. Ltd.

Zomato Restaurant names and Metadata:

Name	: Name of Restaurants
Links	: URL Links of Restaurants
Cost	: Per person estimated Cost of dining
Collection	: Tagging of Restaurants with respect to Zomato categories
Cuisines	: Cuisines served by Restaurants
Timings	: Restaurant Timings

Zomato Restaurant 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

DATASETS:

There are two datasets given:

Restaurant Names and metadata:

- There are 105 records and 6 features in metadata.
- There are missing or null values in Collections and timings.
- There are no duplicated values.
- Cost must be int type but it contains comma (,), hence its datatype is object here.
- Timings represent the time from when the restaurant opens till end time when restaurants shut down, but it is given in the form of text, hence object datatype.
- There are total of 105 restaurant information available along with their websites.

Reviews dataset:

- There are 10000 records (or reviews) given with 7 features.
- Except Name of Restaurants and Number of pictures posted, there are null values.
- There are some of the duplicated values for restaurants which can be dropped (Since it contains null values for all the columns).
- Rating must be integer but it contains value 'like', hence it is object datatype.
- There is total 100 restaurants whose reviews are given.
- Total of 7446 reviewers have given their review on restaurants.
- Metadata contains number of followers and reviews on restaurants.
- Pictures posted by customers has 36 unique values.
- Rating may vary from 0 to 5. Let's check for its unique values.

Chapter – V

DATA WRANGLING

OBSERVATIONS:

For the Review dataset:

- Dropped the duplicate rows (since it contained null values)
- Changed the Rating - Like to numeric value and changed it datatype (Since it represents the ordinal data)
- Extracted Number of review and followers from Metadata column and filled the null values of followers with 0.
- Changed the time datatype to datetime and extracted Year and Hour from it.

For the Hotel dataset:

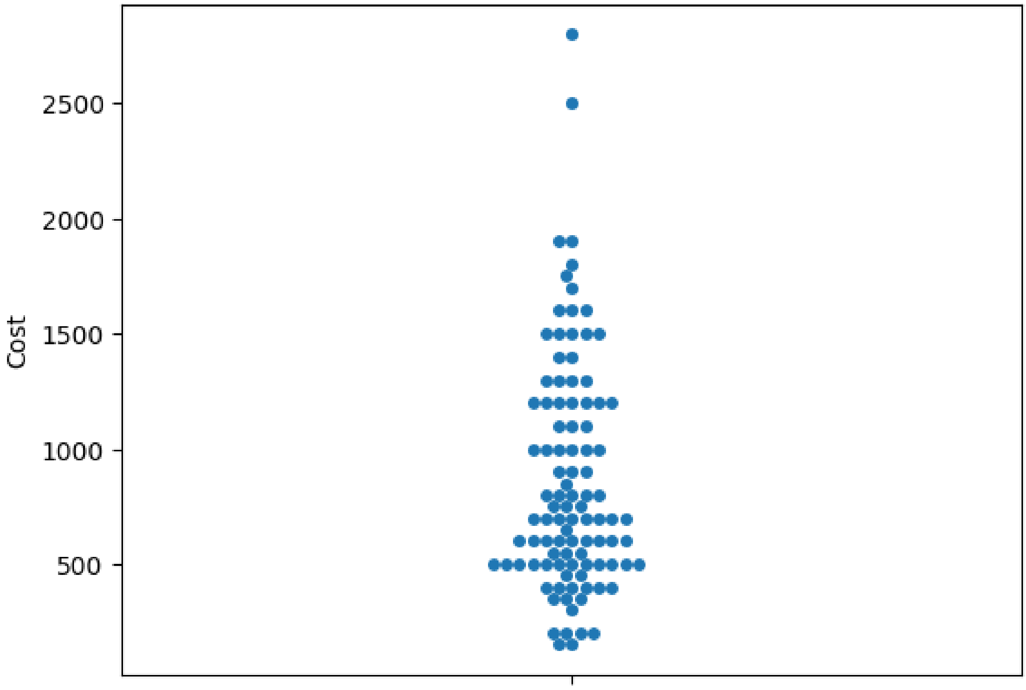
- Rename the Column 'Name' to 'Restaurant' for the sake of simplicity.
- Removed special character (,) from Cost and changed its datatype to integer.
- Get the number of cuisines.
- Merged the average rating in hotel dataset.

Chapter – VI

EXPLORATORY DATA ANALYSIS

Data Visualization, Storytelling & Experimenting with charts : Understand the relationships between variables.

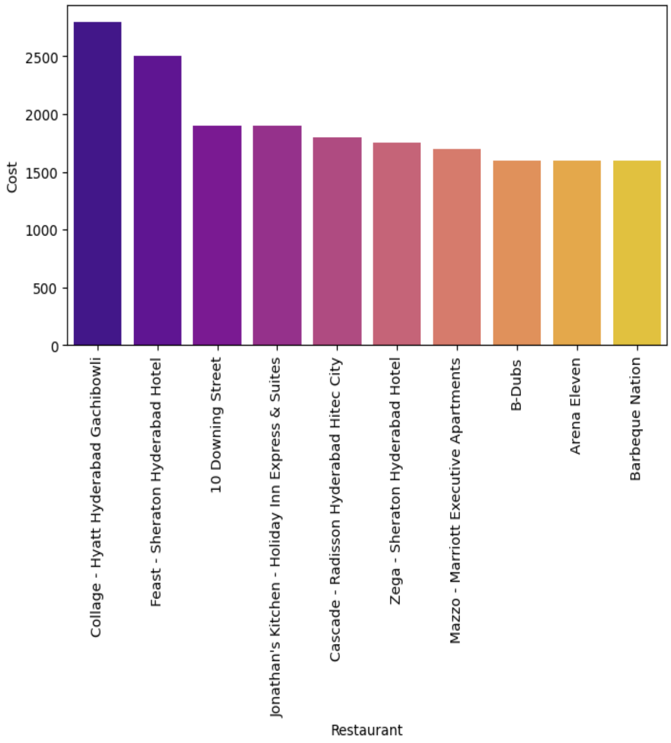
1. To find the cost of restaurants:

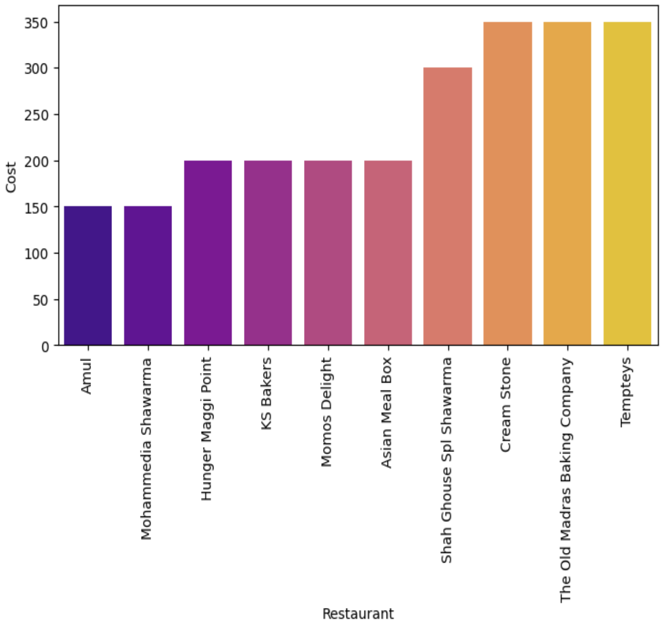


Observations:

It is clearly visible that average cost per person in restaurants varies from below 500 to more than 2500. But there are too few restaurants whose price is more than 2000. Let's find out more about Restaurant prices.

2. To visualize which are the expensive restaurants and which are the cheap restaurants on Zomato:





Observations:

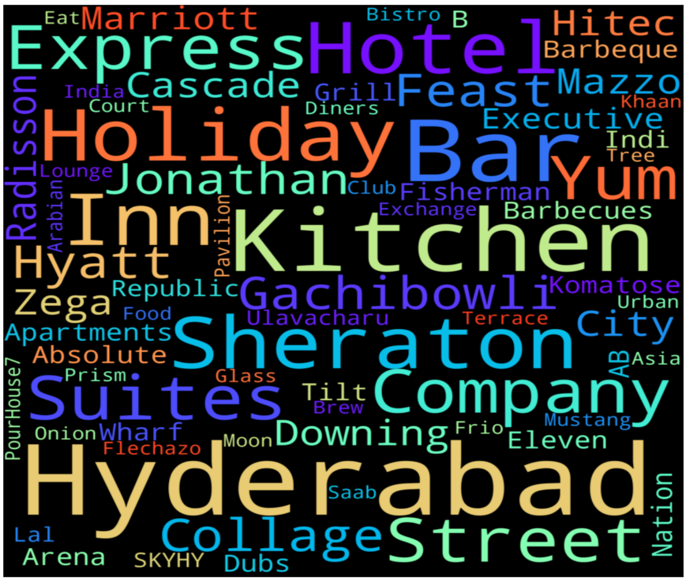
Expensive Restaurants: Here "Collage - Hyatt Hyderabad Gachibowli" is the most expensive restaurant whose price is rupees 2800 which is followed by "Feast - Sheraton Hyderabad Hotel" whose price is rupees 2500. Other expensive restaurants can be seen from the graph and table.

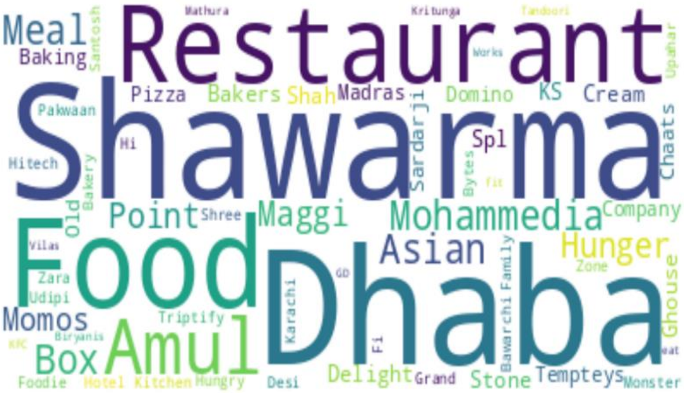
Cheap Restaurants: Here "Mohammedia Shawarma" and "Amul" is the cheapest restaurant where we can get the dish with the minimum price of rupees 150, which is followed by "Sweet Basket", "KS Bakers", "Momos Delight etc whose price is rupees 200.

From this insight we get to know about the restaurants which has dishes containing lower prices. So, a middle-waged person can afford it easily. Hence this can be beneficial for that particular restaurant as well as Zomato as more people will order food from Zomato.

Also, we get to know about the most expensive restaurant, which led to negative growth of that restaurants as fear of losing money if they do not get the taste they want by customers.

3. Word cloud is used because it shows all text and highlight the most frequent words (Restaurants):

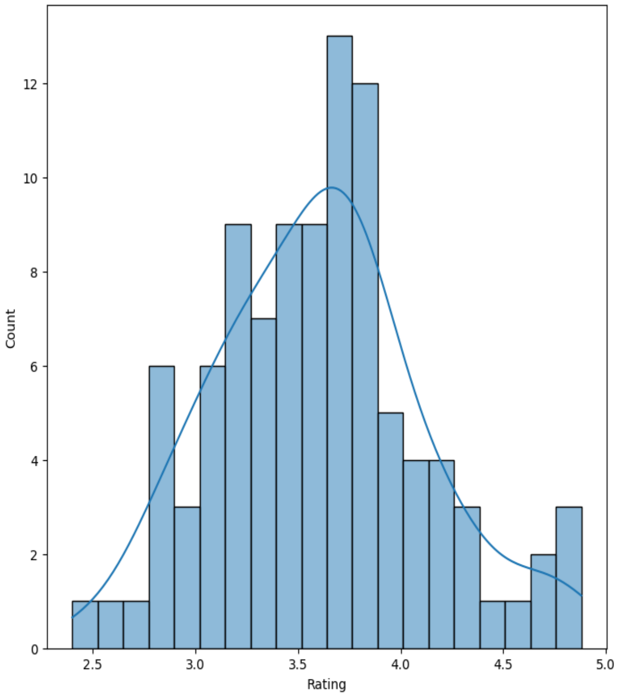


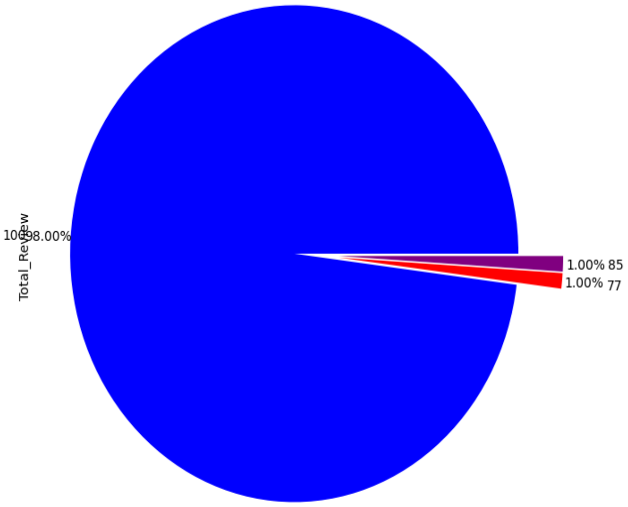


Observations:

From the above chart, HYDERABAD, HOTEL, BAR etc seems frequently repeating for expensive restaurant, while for cheap restaurants SHAWARMA, DHABA, RESTAURANTS seems frequently repeating. So, it can be inferring that Hotel and Bars of Hyderabad are expensive while Dhaba’s and Restaurants are cheaper.

4. histogram plot is used to see the distribution of average rating and pie chart is used to see review distribution:



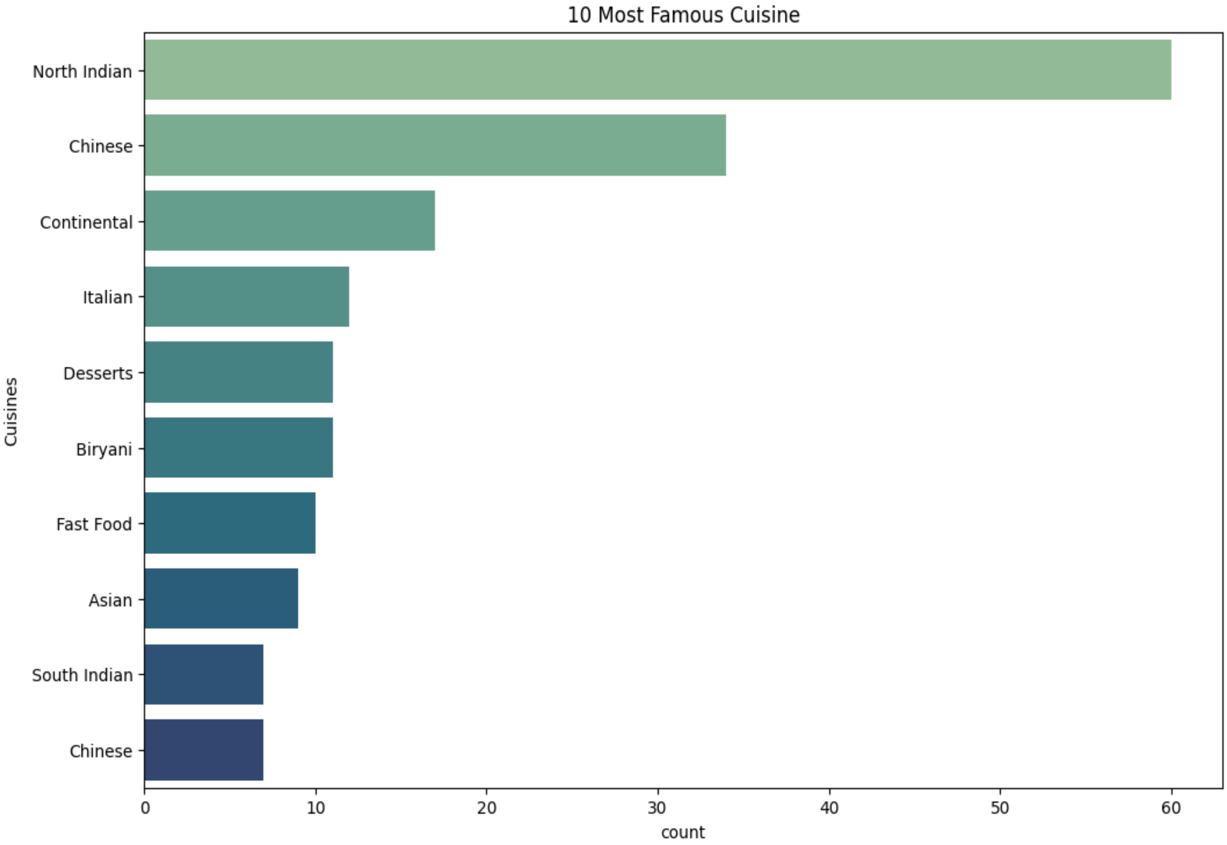


Observations:

Average Ratings are normally distributed for the restaurants.

100 reviews are given to all the restaurants except 2 restaurants whose reviews are 85 and 77 respectively.

5. Bar graph is used because of categorical features:

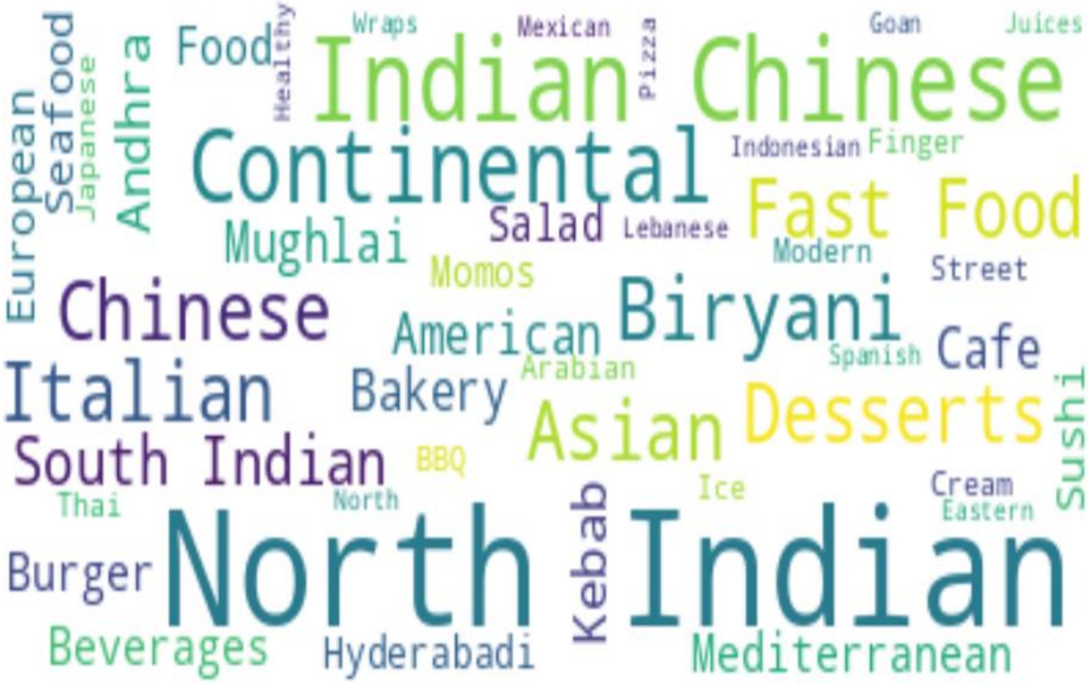


Observations:

It is clearly visible that North Indian is the most served cuisine in restaurants which is followed by Chinese and Continental.

It may be helpful for new entrepreneurs who wants to open new restaurants in their area so that they get know what people like mostly and keep that cuisine in their menu.

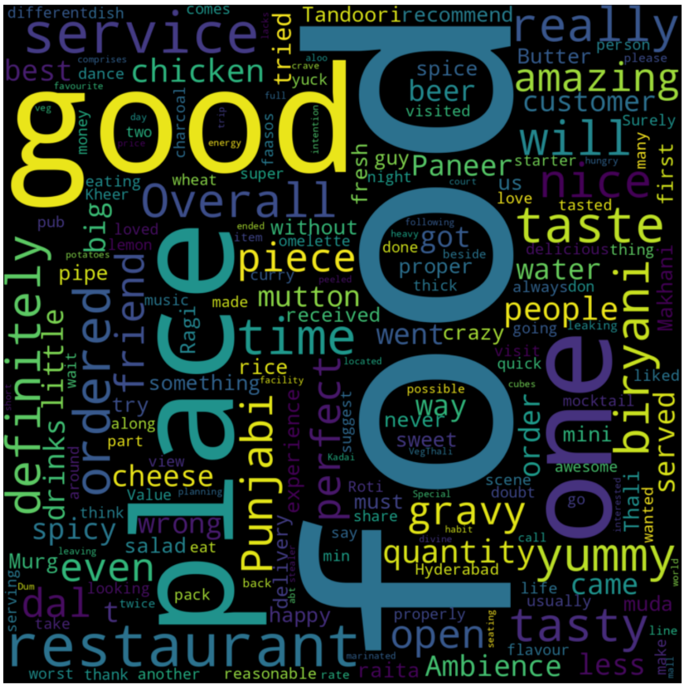
6. Word cloud is used because it shows all text and highlight the most frequent words (Food):



Observations:

From the above chart, North Indian is the most frequently used which is followed by Chinese and continental.

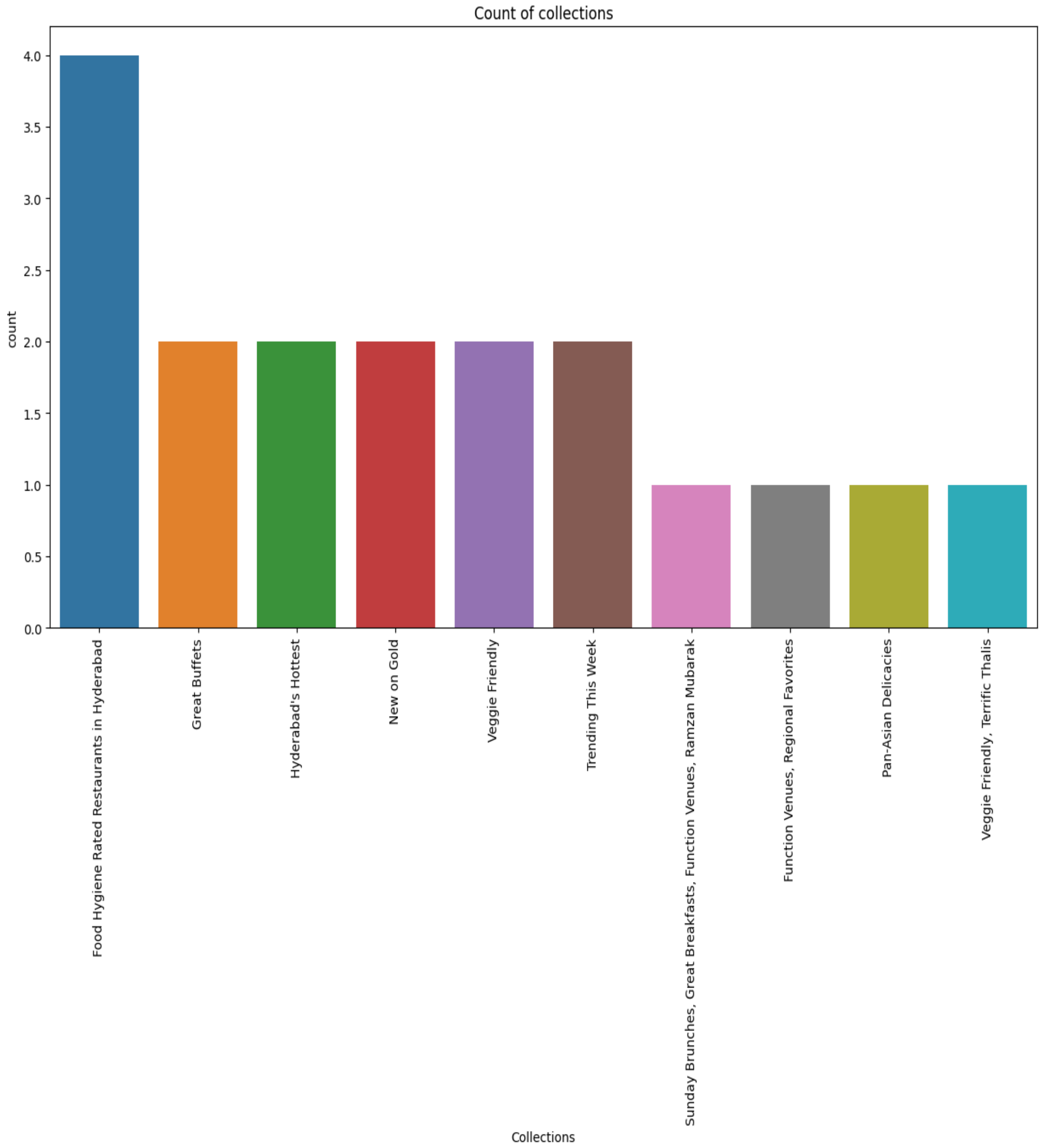
7. To see what word is frequently used by the reviewers:



Observations:

Most of the time customers liked the food because good is repeating most in reviews. Also, food is next most repeating word.

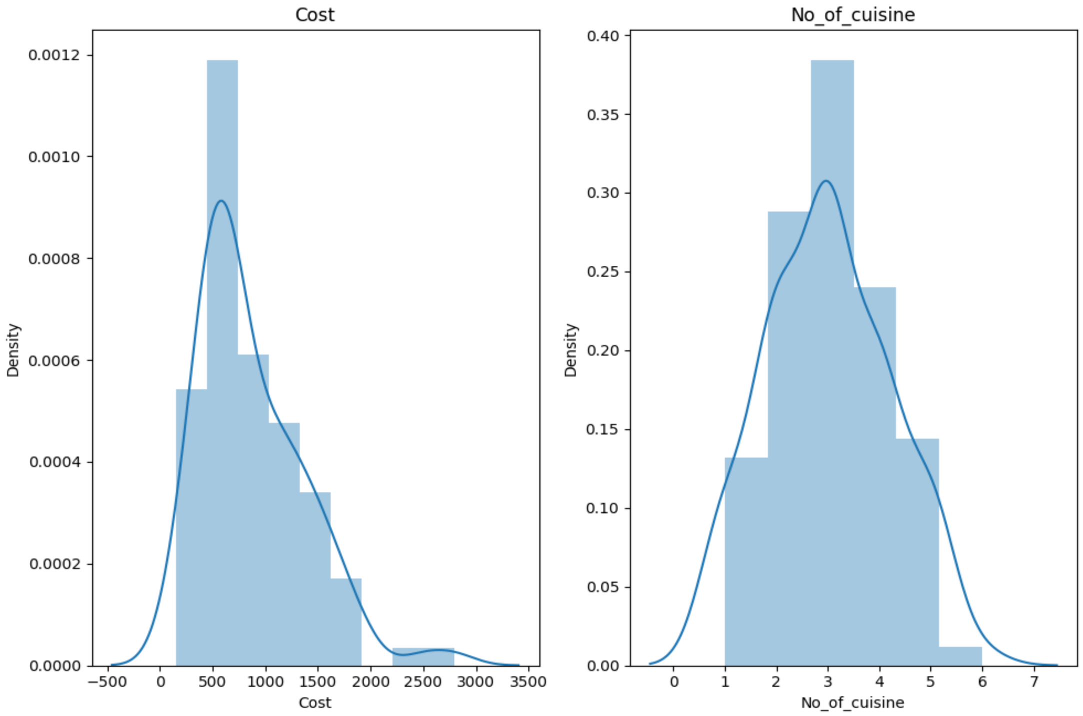
8. To see the count of each collection:

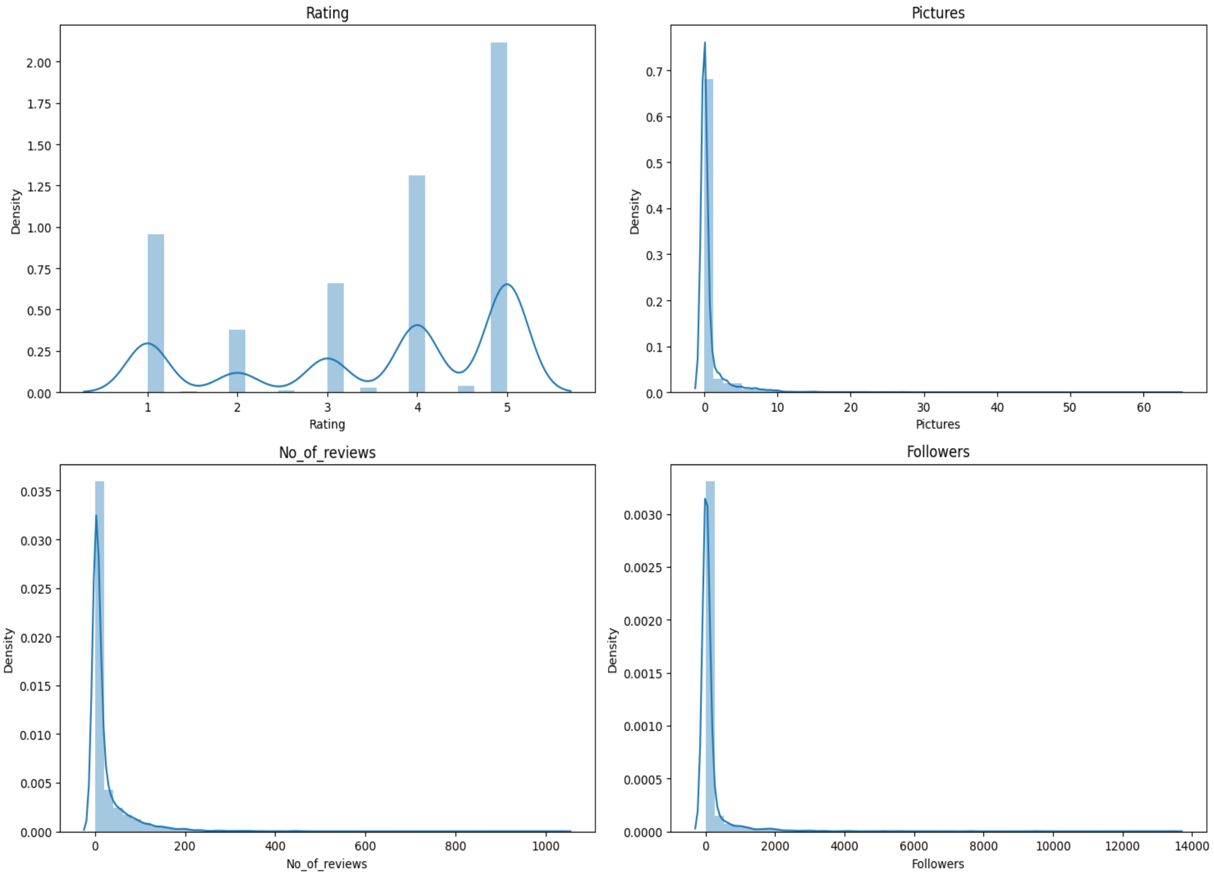


Observations:

Here Food Hygiene Rated Restaurants in Hyderabad has the maximum count of 4 which is followed by Great Buffets, Hyderabad Hottest etc.

9. To see the distribution of numerical columns:



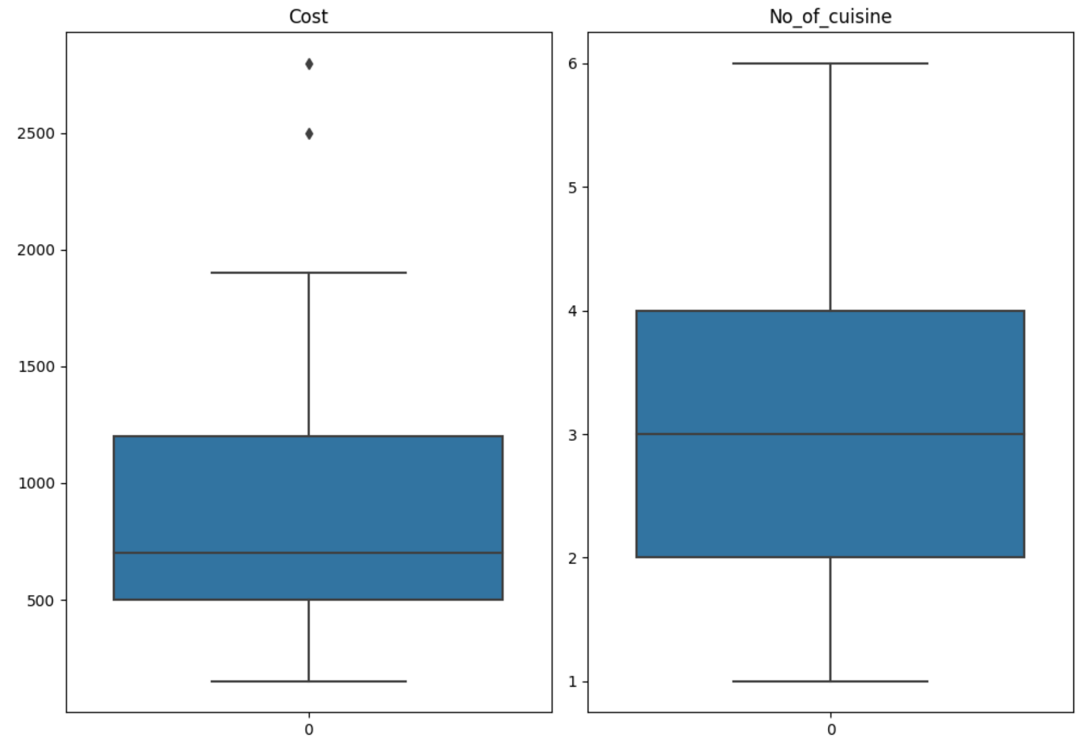


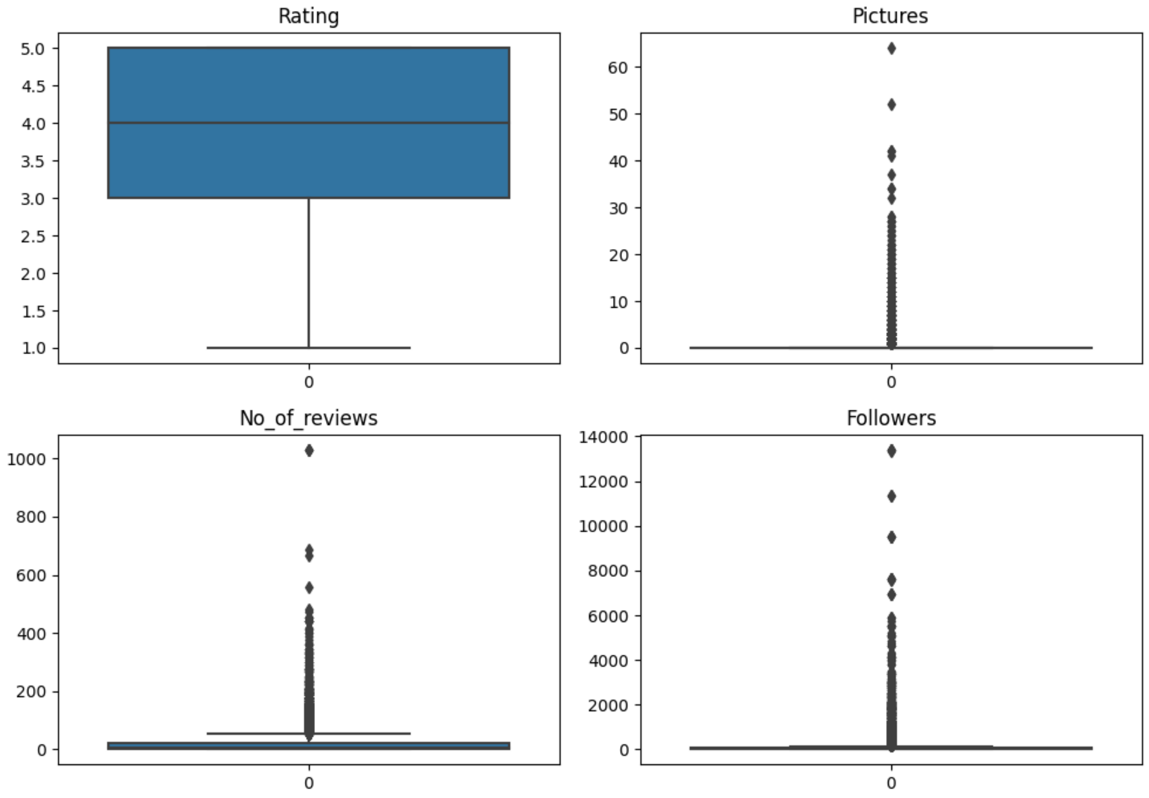
Observations:

Hotel Dataset: Cost is right skewed while Number of cuisines is normally distributed.

Review dataset: Pictures, Number of reviews and followers are right skewed.

10. Boxplot is used to check for outliers:



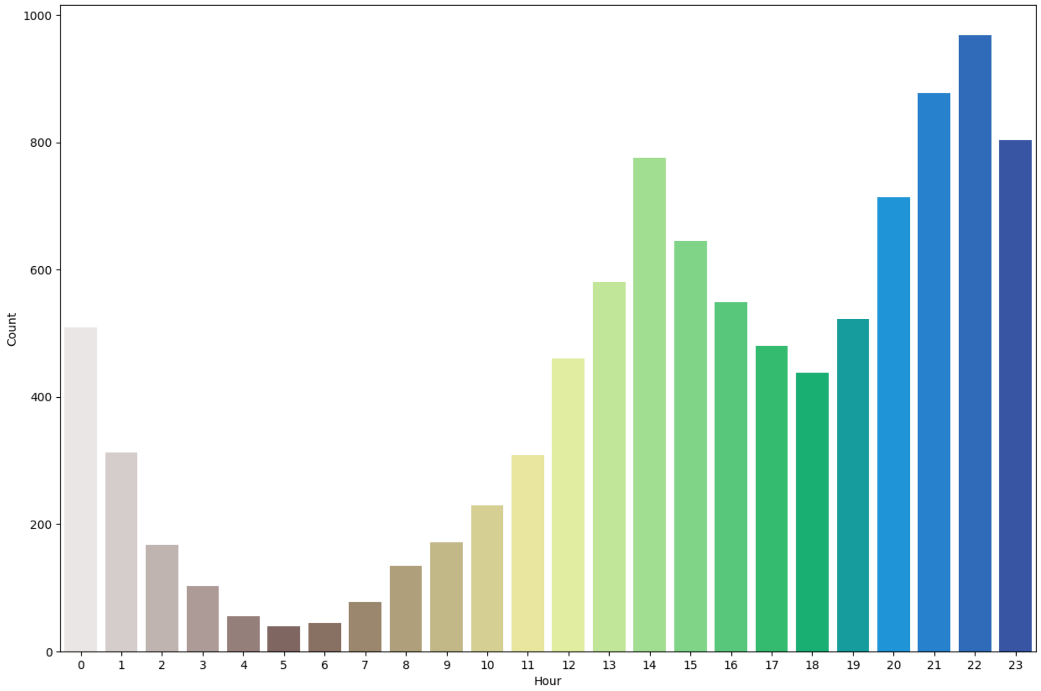


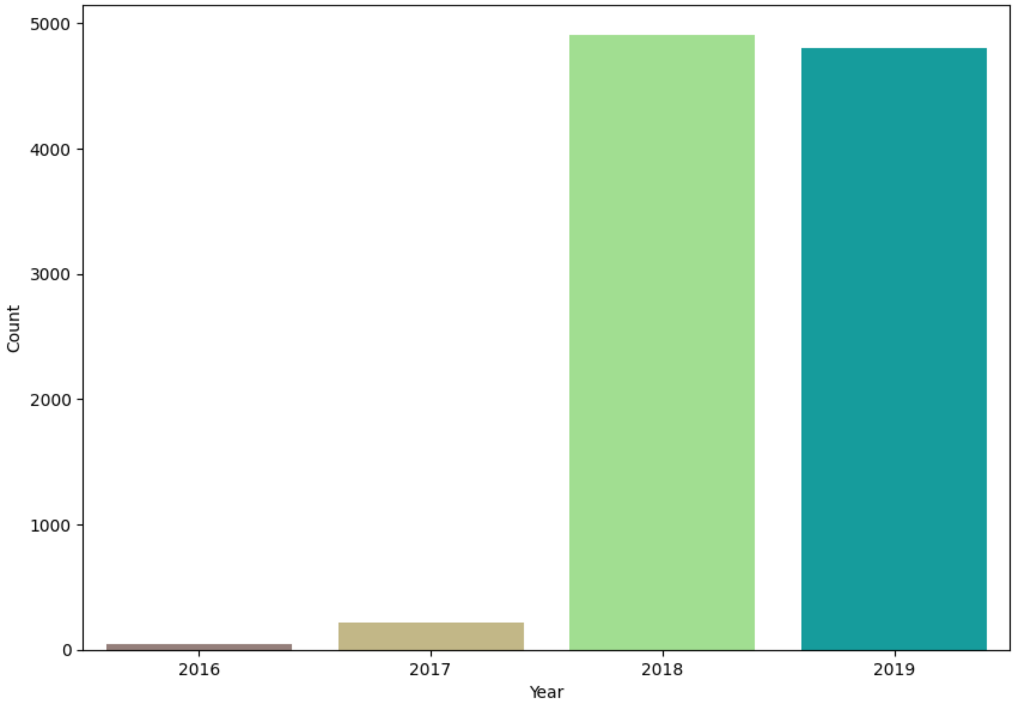
Observations:

Hotel Dataset: Not much of the outliers is seen in Cost and Number of cuisines.

Review Dataset: Pictures, Number of reviews and Followers seems to have outliers.

11. To see the count of review for restaurants given in each year and each hour:



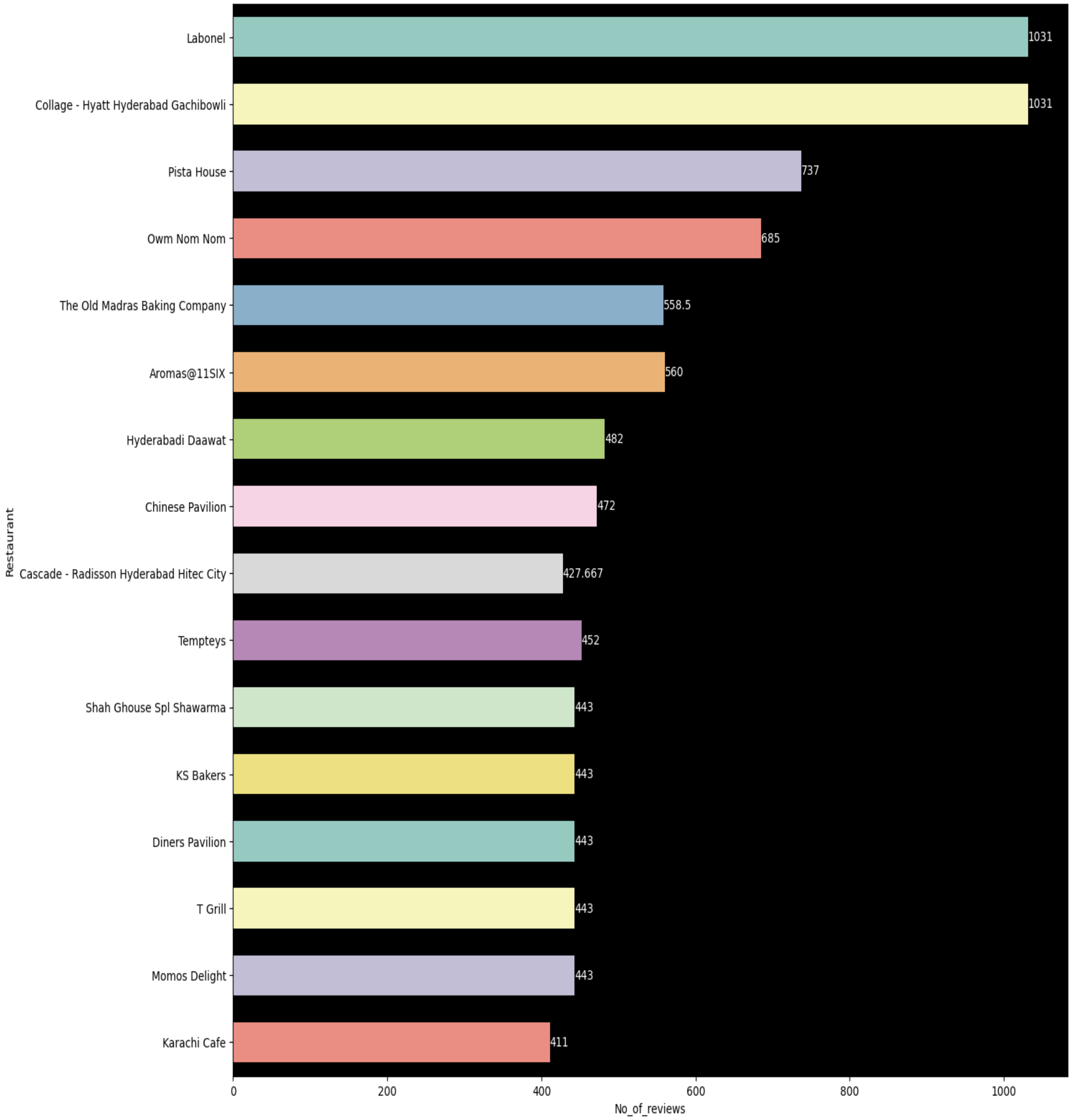


Observations:

Hour: The frequency is higher during the night time from hour 19 to 22, i.e., from 7:00 pm to 11:00 pm. Possibly because people mostly order food during these hours.

Year: The frequency is minimum in the year 2016 while its maximum in the year 2019. It is possible due to the fact that there is improvement in technology and people getting familiar with new applications and online system.

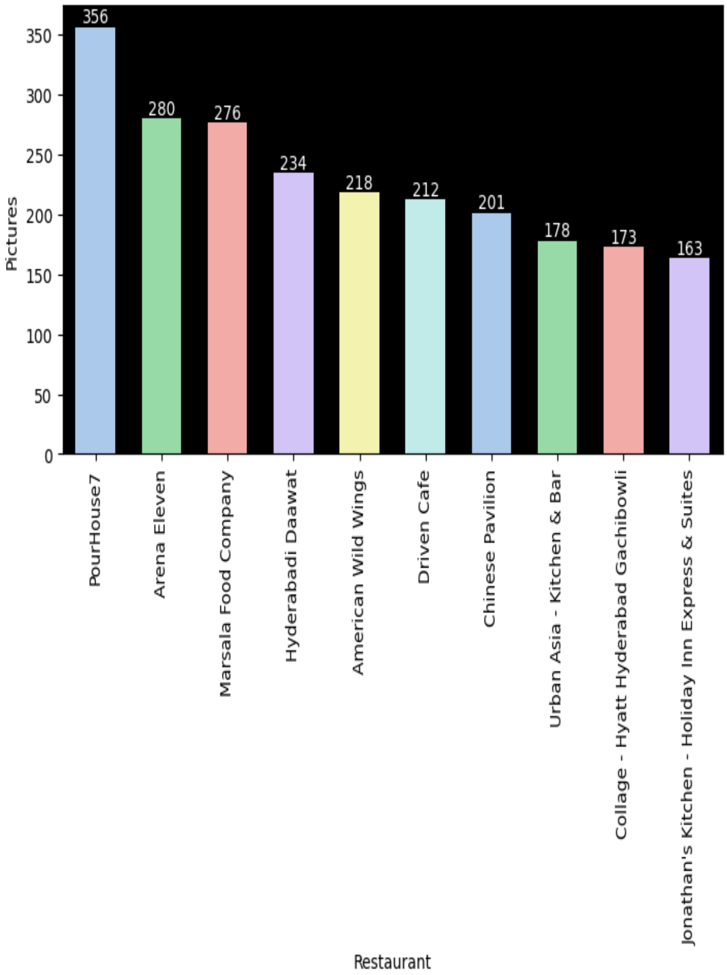
12. To see the count of review for restaurants given in each year and each hour:

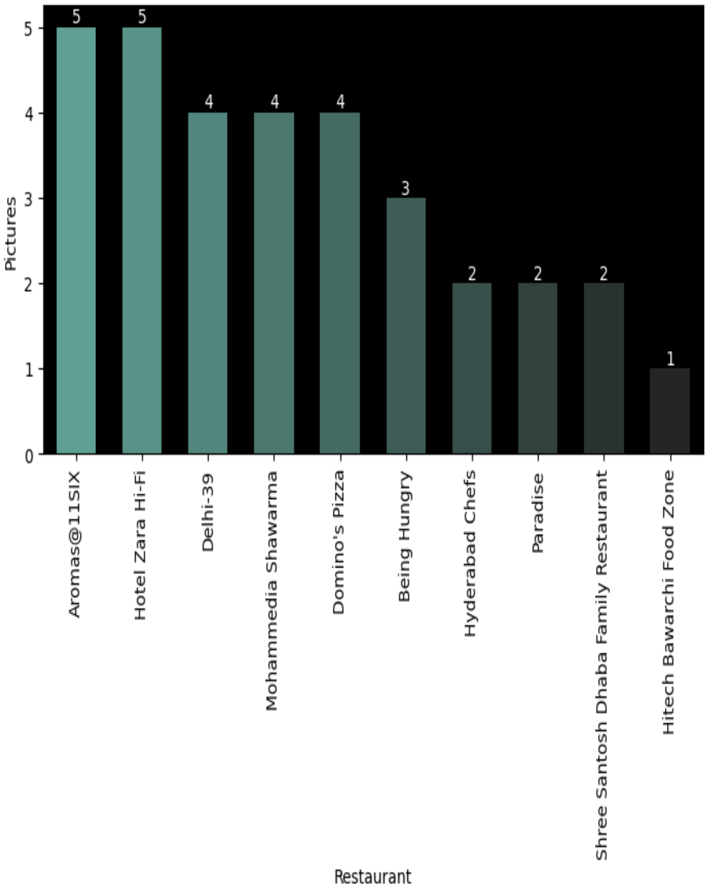


Observations:

Here Labonel and Collage - Hyatt Hyderabad Gachibowli (which is also the most expensive) are given he maximum number of reviews with the count of 1031 (which is really a good figure).

13. To see most and least number of pictures posted for the restaurant.:

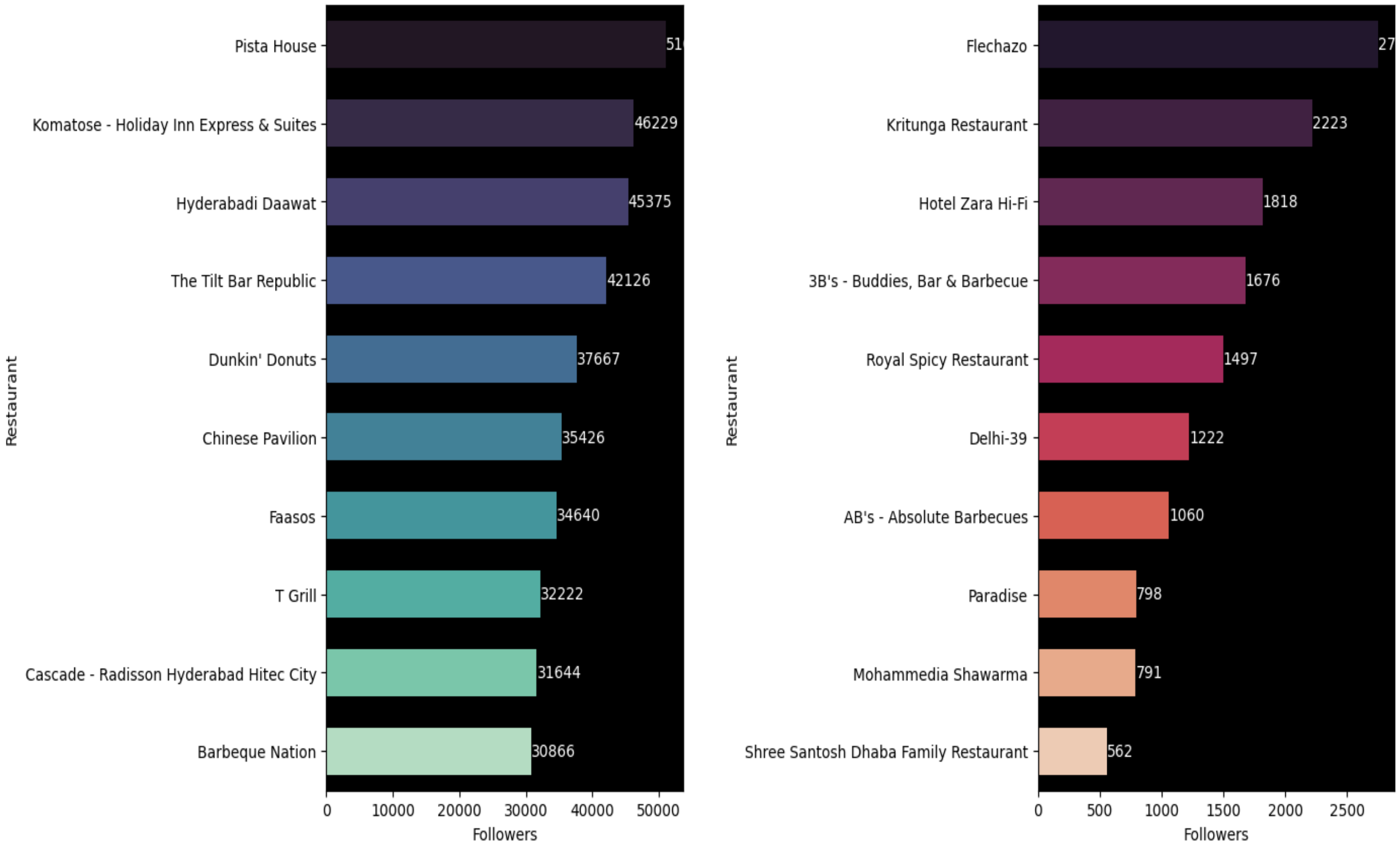




Observations:

Here PourHouse7 has the maximum number of pictures posted by the reviewers which is 356 followed by Arena Eleven.

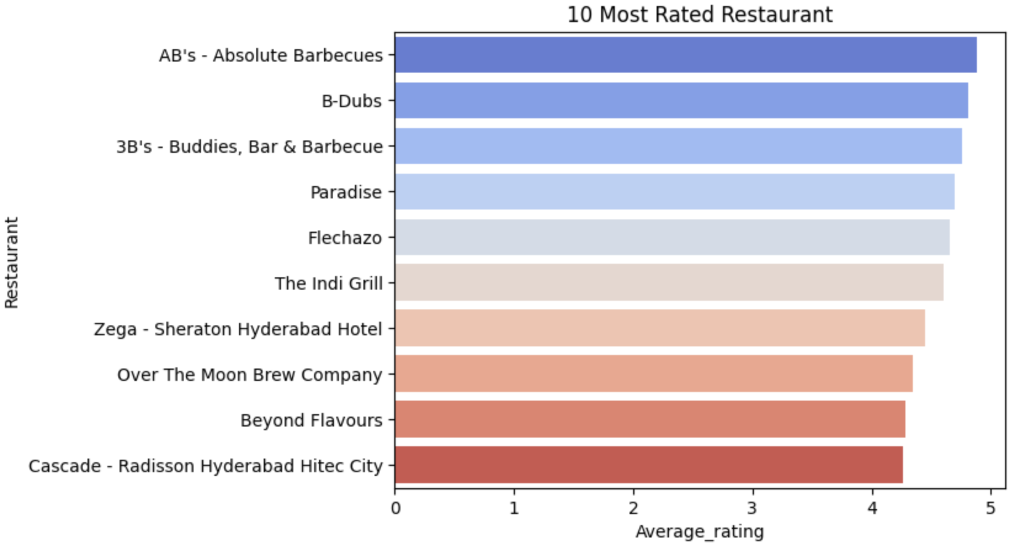
14. To see the least and most follower of reviewer of the restaurants:



Observations:

Here Reviewer of restaurant "Pista House" has the greatest number of followers, while Reviewers of restaurant "Shree Santosh Dhaba Family Restaurant" has the least number of followers.

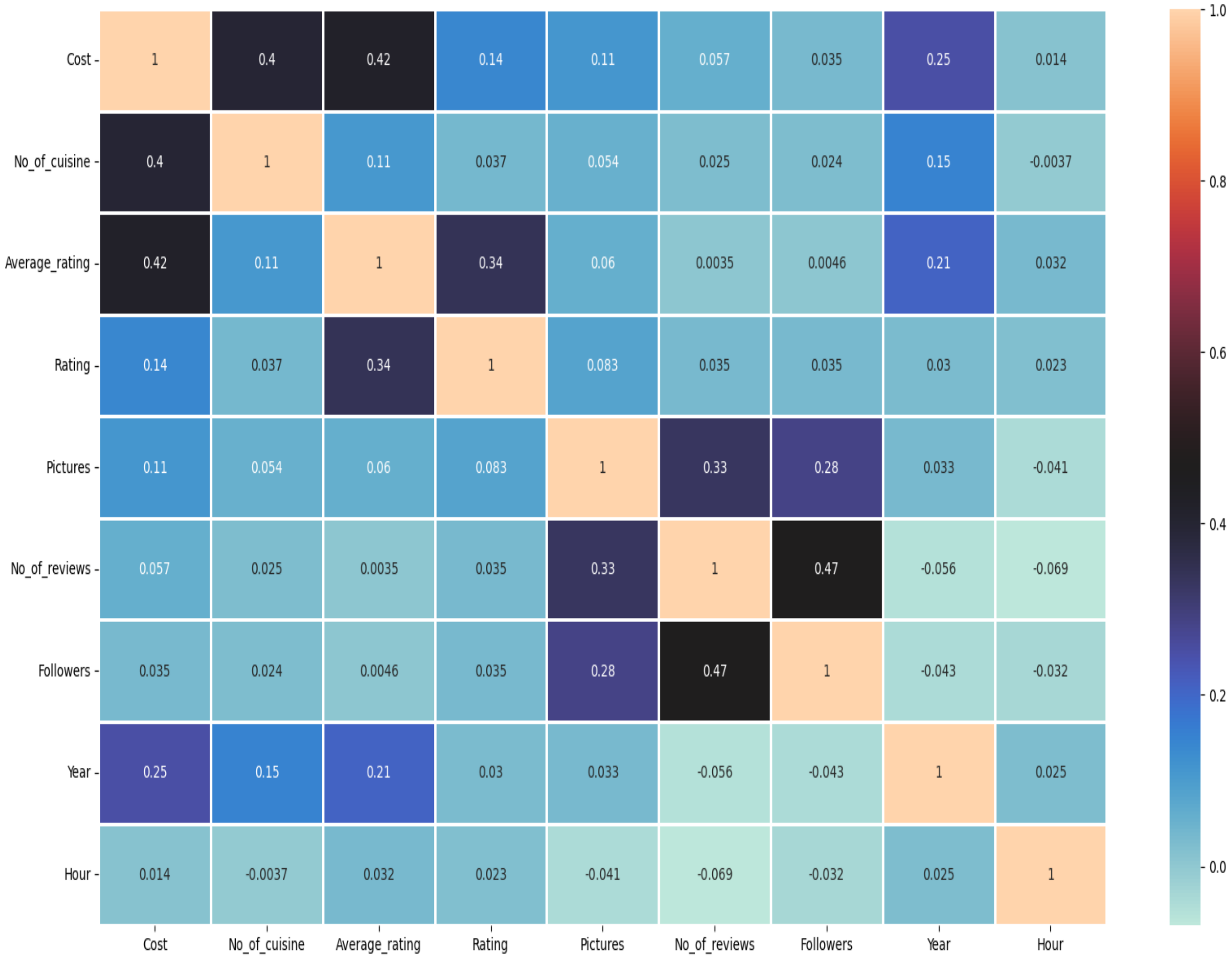
15. To see the top 10 restaurants having highest average rating:



Observations:

Here AB's - Absolute Barbecues is the top average rated restaurant followed by B-Dubs and 3B's - Buddies, Bar and Barbeque.

16. To see the correlation among numerical features:



Observations:

Number of reviews and followers has correlation of 0.47 which can be considered as moderate.

Similarly cost and number of cuisines has moderate correlation of 0.4.

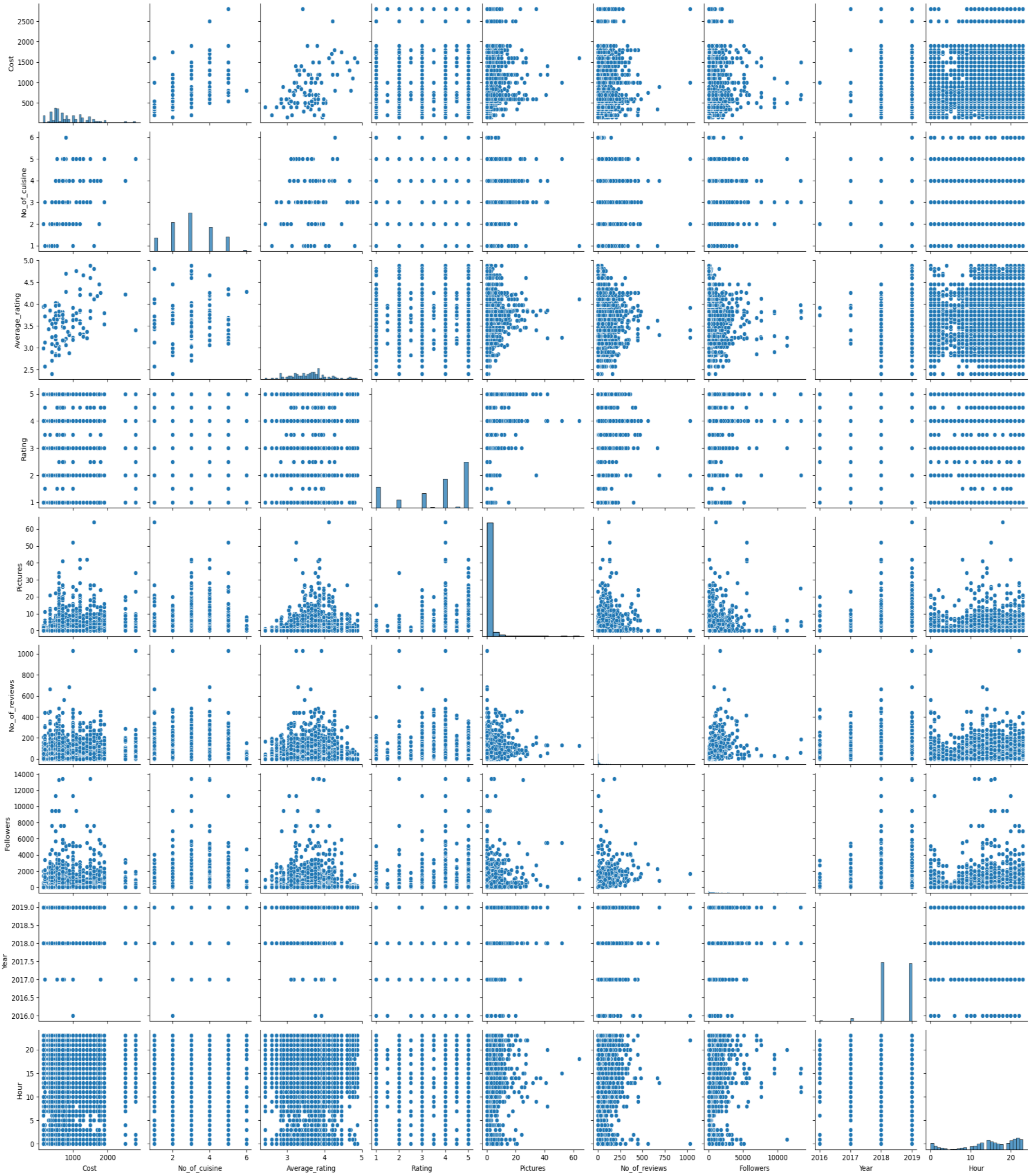
There is low correlation between:

- Pictures and Followers
- Pictures and No. of reviews
- Cost and year

Since these correlations are low, no case of multicollinearity arises.

Other features have very low correlation.

17. Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters:



Observations:

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

Chapter – VII

HYOTHESIS TESTING

Based on chart experiments, define three hypothetical statements from the dataset. In the next three questions, perform hypothesis testing to obtain final conclusion about the statements through code and statistical testing.

Hypothesis 1 : *Average rating by the customer is 3.5.*

Hypothesis 2 : *Restaurants which serves greater variety of cuisines are costly.*

Hypothesis 3 : *Cost is distributed normally.*

Hypothetical Statement - 1

1. State Your research hypothesis as a null hypothesis and alternate hypothesis.

Average rating by the customer is 3.5

Null Hypothesis H0: 𝓜 = 3.5, Mean rating is 3.5

Alternative Hypothesis H1: 𝓜! =3.5, Mean rating is not 3.5

Observations:

T-test for one sample (two-tailed test) is used to check if the average rating given by reviewers is 3.5 or not.

Since t test is used to test sample mean considering the population mean assuming that population parameters are unknown. The objective of this test is to compare the means of two related or unrelated sample groups. Hence considering the given data as sample and extracting Average rating as data and compared it with total average rating of 3.5.

INFERENCE: Since p_value obtained is less than level of significance (0.05). Hence, we reject the null hypothesis and conclude that average rating given by reviewers is not 3.5

Hypothetical Statement - 2

1. State Your research hypothesis as a null hypothesis and alternate hypothesis.

Hypothesis 2: Restaurants which serves greater variety of cuisines are costly.

Null Hypothesis H0: There is no relation between number of cuisines and cost.

Alternative Hypothesis H1: Restaurants which serve higher number of cuisines are more costly.

Observations:

Chi-square contingency test is used to check if cost and Number of cuisines have relationship among them or not.

The chi-square contingency test is used to test the independence of two events. It tells us whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

INFERENCE: Since p value obtained is 0.63 which is greater than level of significance, i.e., 0.05, hence we fail to reject the null hypothesis and conclude that there is no relation between number of cuisines and cost.

Hypothetical Statement - 3

1. State Your research hypothesis as a null hypothesis and alternate hypothesis.

Hypothesis 3: Cost is distributed normally.

Null Hypothesis H0: Cost has Gaussian distribution.

Alternative Hypothesis H1: cost does not have Gaussian distribution.

Observations:

Shapiro - Wilk test is used to check if Distribution of cost for restaurants is distributed normally or not.

The Shapiro-Wilk test is a way to tell if a random sample comes from a normal distribution. The test gives a p value; small values indicate sample is not normally distributed (can reject the null hypothesis that population is normally distributed if values are under a certain threshold). Also, it assumes that Observations in each sample are independent and identically distributed (iid).

INFERENCE: Since p value is 0.00 which is less than level of significance 0.05, hence we fail null hypothesis. And conclude that cost is not distributed normally.

Chapter – VIII

DATA PRE-PROCESSING

.

1. Handling Missing Values

Observations:

For Hotel dataset:

Imputed the one null value in Timings with mode of column because it is assumed that opening and closing time is similar for most of the restaurants.

Since Collections contain more than 50% null values, hence this column has been dropped.

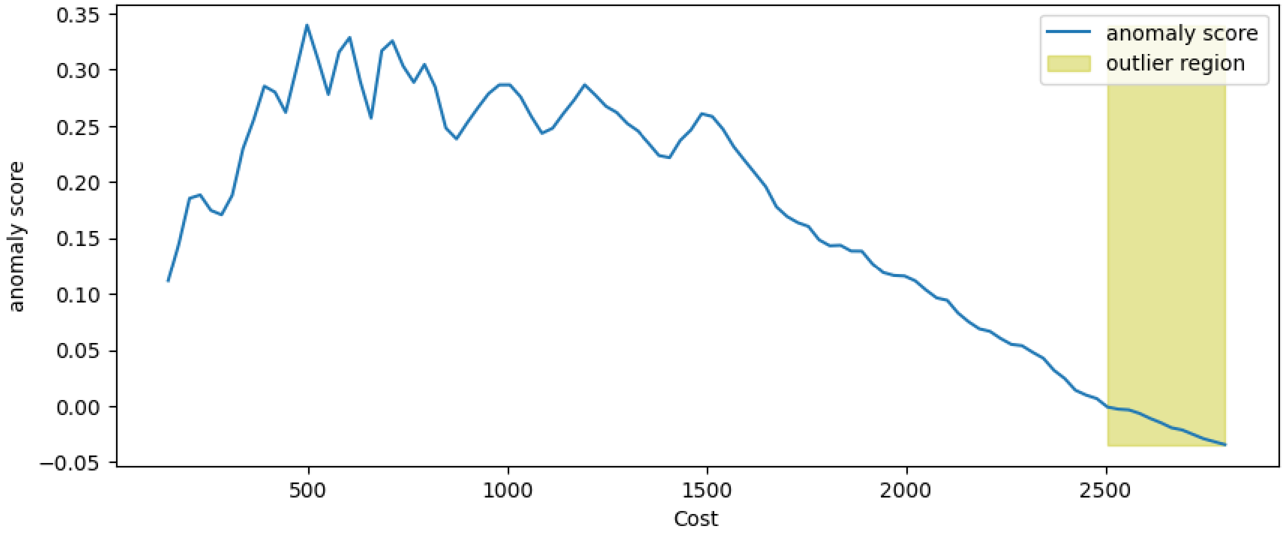
For review dataset:

There are missing values in review, hence filled it with 'No review'.

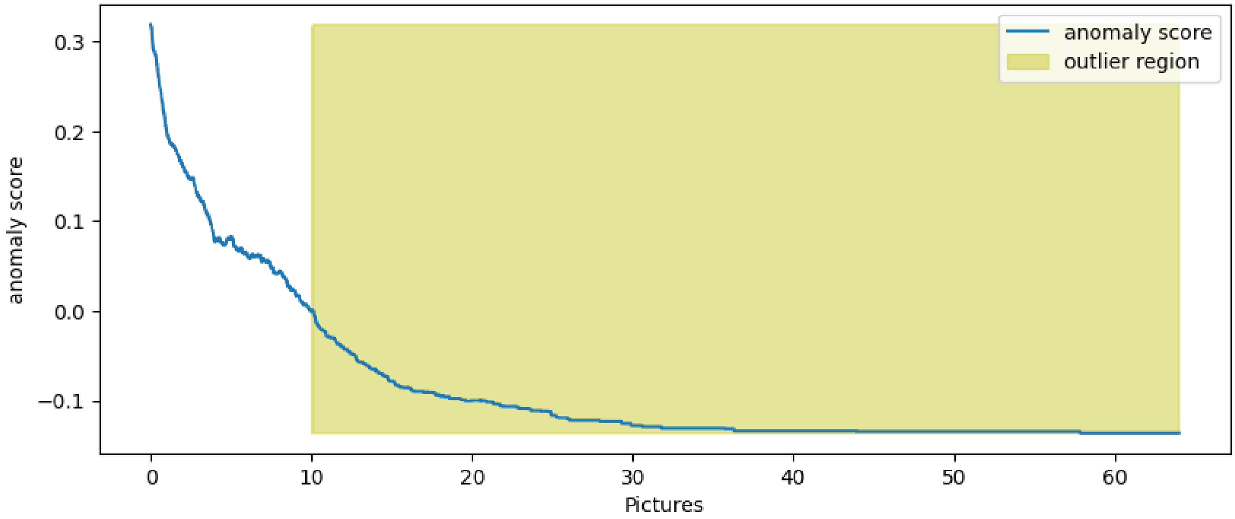
2. Handling Outliers

Anamoly Detection

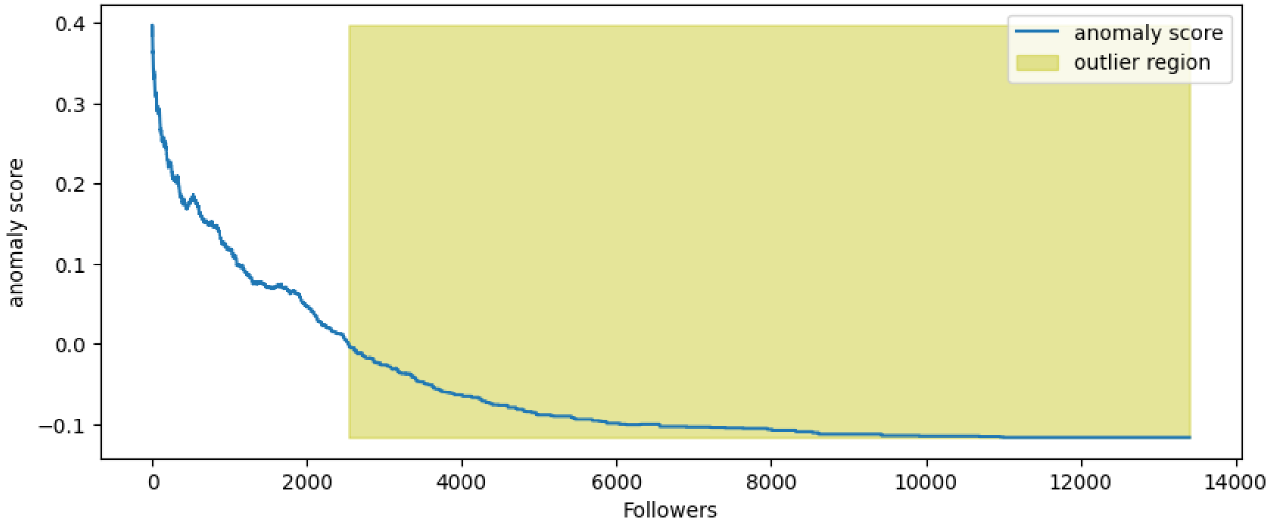
a. Visualizing outlier for cost:



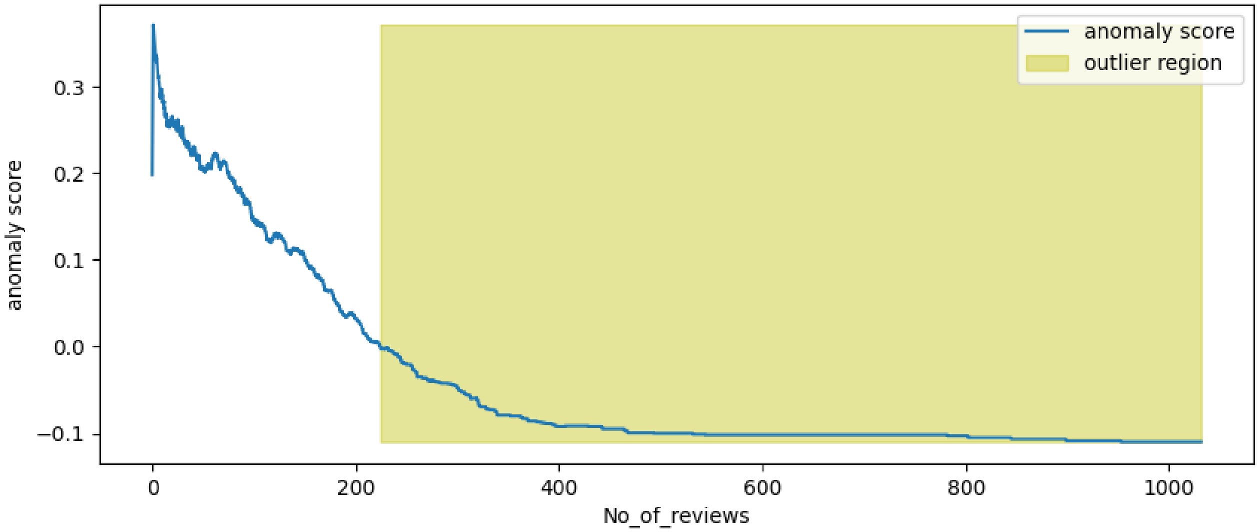
b. Visualizing outlier for Pictures:



c. Visualizing outlier for Followers:



a. Visualizing outlier for No. of Reviews:



Observations:

For the detection of outliers, isolation forest is used which is unsupervised technique to detect outliers. It assumes that points which end up in shorter branches indicate anomalies as it was easier for the tree to separate them from other observations while the points that travel deeper into the tree are less likely to be anomalies as they required more cuts to isolate them.

For the treatment of outliers, replaced upper outliers with upper bound and lower outliers with lower bound considering their maximum and minimum value cannot exceed these points.

3. Categorical Encoding

Observations:

For encoding of categorical feature which is 'Cuisines', First the cuisines have been splitted into a list and then created dummy variables for each of the cuisines and allotted to the restaurants.

4. Feature Manipulation & Selection

Observations:

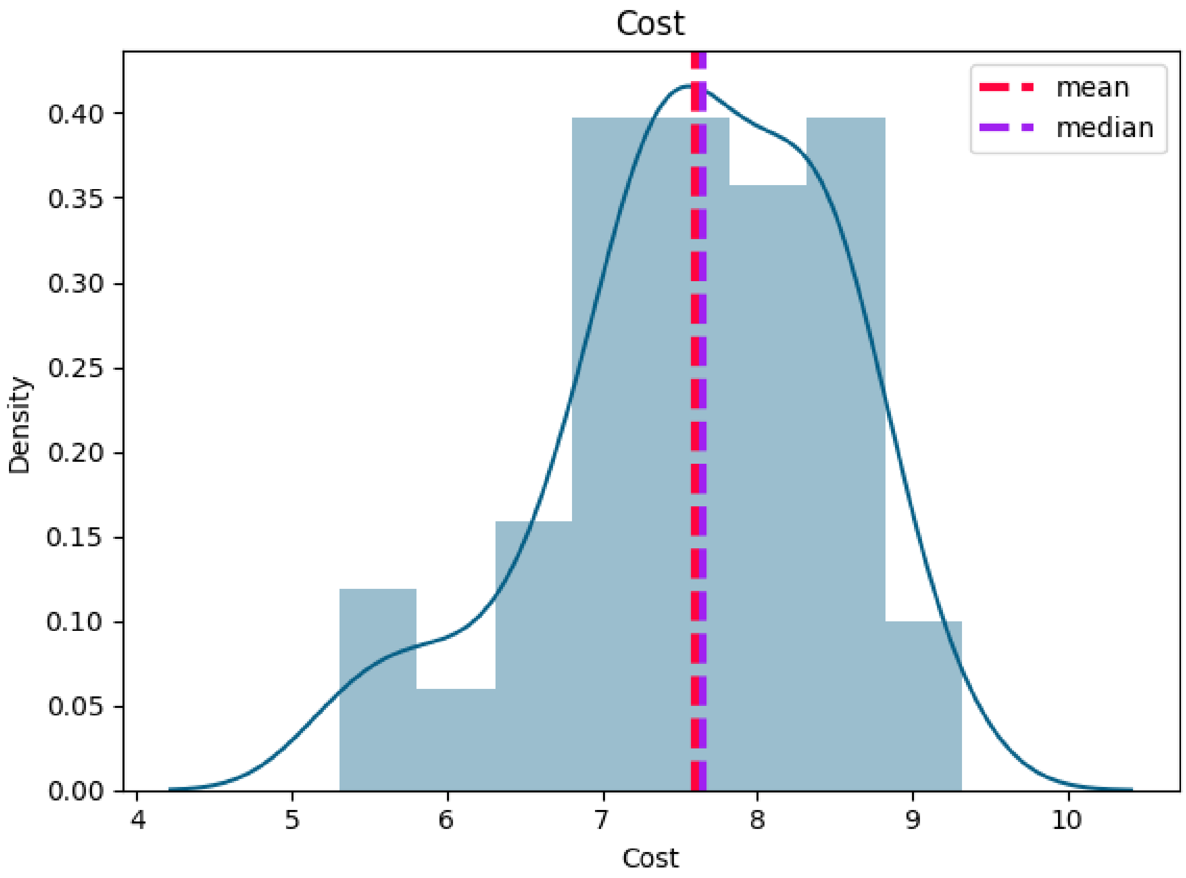
Feature are selected based on market experiences and consumer consumption. Some of the features will be selecting using Dimensionality reduction technique.

I have created two datasets for clustering and sentiment analysis.

For clustering: features include - 'Cost', 'No_of_cuisines', 'Average_rating', and all the cuisines count for each restaurant.

For sentiment analysis: Features include - 'Review' and 'Sentiments' where sentiment is extracted from rating.

5. Data Transformation



Observations:

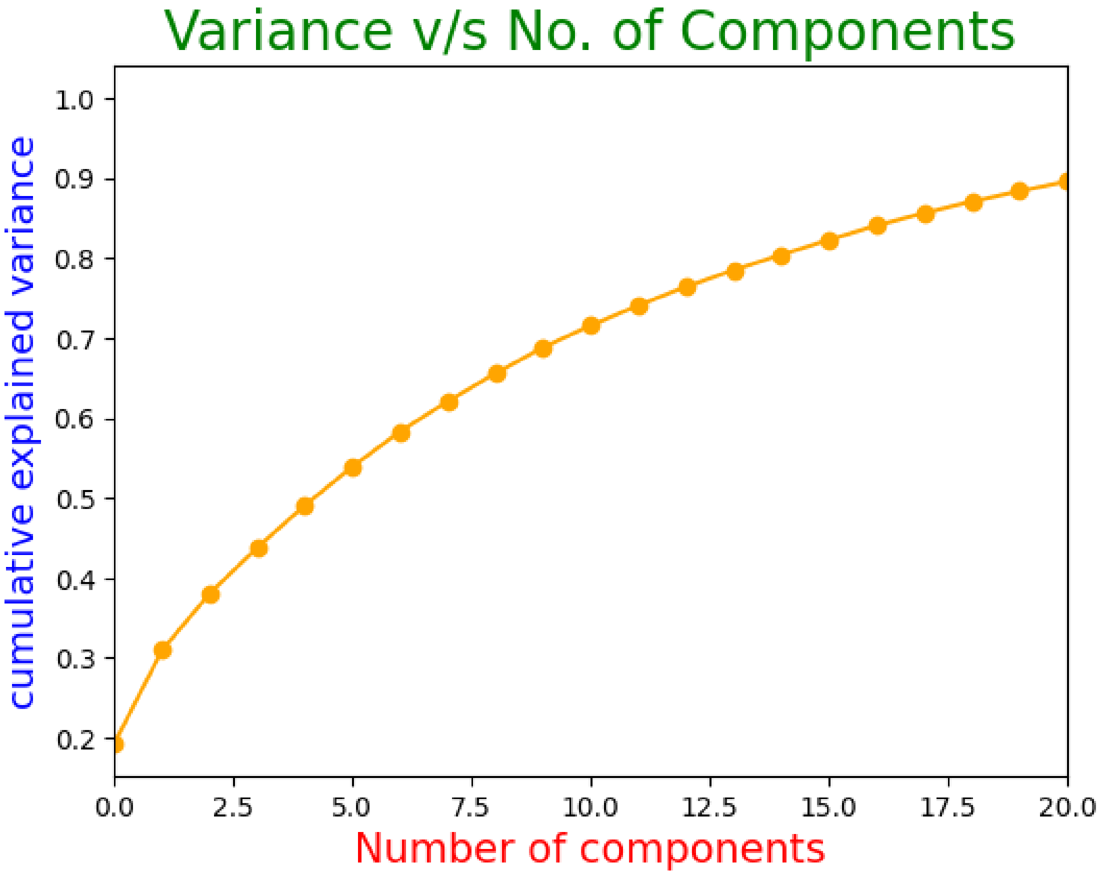
So now there is no skewness in cost. Hence, we can proceed further.

6. Data Scaling

Observations:

MinMax Scaler has been used to scale the data. The feature scaling is used to prevent the models from getting biased toward a specific range of values. Since the dummy variables created from cuisines contains the value 0 and 1 while other variables have different range of values.

7. Dimensionality Reduction (For clustering)



Observations:

For clustering dataset, there are more than 30 variables, i.e., there are 47 variables and higher number of features makes computational cost of clustering algorithms also higher. In addition, high dimensionality can lead to the "curse of dimensionality", where the data becomes sparse and the clusters become harder to identify.

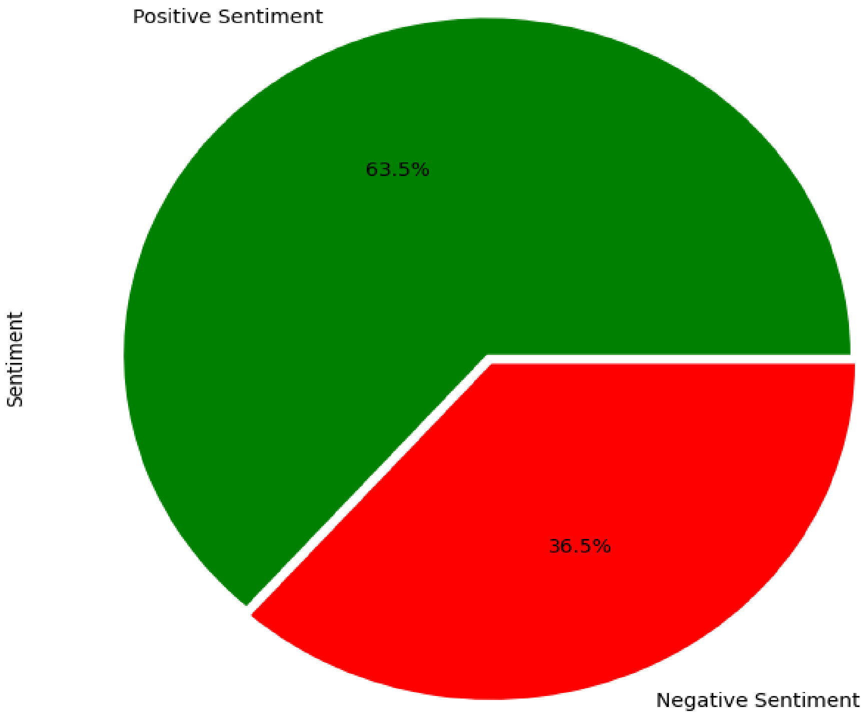
PCA is used as dimension reduction technique, because PCA (Principal Component Analysis) is a widely used dimensionality reduction technique because it is able to identify patterns in the data that are responsible for the most variation. Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.

8. Data Splitting (for sentimental analysis)

Observations:

80:20 split has been used which is one the most used split ratio. Since there were only 9961 data, therefore I have used more in training set.

9. Handling Imbalanced Dataset



Observations:

the dataset is imbalanced but since it is lightly imbalanced, hence handling is not necessary. So, we can proceed with same dataset.

Chapter – IX

TEXT DATA PRE-PROCESSING (USING NLP)

(It's mandatory for textual dataset i.e., NLP, Sentiment Analysis, Text Clustering etc.)

1. Expand Contraction

2. Lower Casing

3. Removing Punctuations

4. Removing URLs & Removing words and digits contain digits.

5. Removing Stop words & Removing White spaces

6. Rephrase Text

7. Tokenization

8. Text Normalization:

Lemmatization is the process of reducing words to their base or root form, similar to stemming. However, lemmatization uses a dictionary-based approach and considers the context of the word in order to determine its base form, while stemming uses simple heuristics and does not consider the context of the word. Lemmatization is a more accurate way of finding the root form of a word as it takes into account the context of the word as well as its grammatical structure.

Lemmatization is used because it is a more accurate way of reducing words to their base form than stemming. Lemmatization considers the context of the word and its grammatical structure to determine its base form, which can help to improve the performance of natural language processing models. Lemmatization is often used in tasks such as text classification and information retrieval, where the meaning of the words is important.

9. Part of speech tagging

10. Text Vectorization:

TF-IDF (term frequency-inverse document frequency) is a technique that assigns a weight to each word in a document. It is calculated as the product of the term frequency (tf) and the inverse document frequency (idf).

The term frequency (tf) is the number of times a word appears in a document, while the inverse document frequency (idf) is a measure of how rare a word is across all documents in a collection. The intuition behind tf-idf is that words that appear frequently in a document but not in many documents across the collection are more informative and thus should be given more weight.

The mathematical formula for tf-idf is as follows:

tf-idf(t, d, D) = tf(t, d) * idf(t, D)

where t is a term (word), d is a document, D is a collection of documents, tf(t, d) is the term frequency of t in d, and idf(t, D) is the inverse document frequency of t in D.

Chapter – X

MODELLING

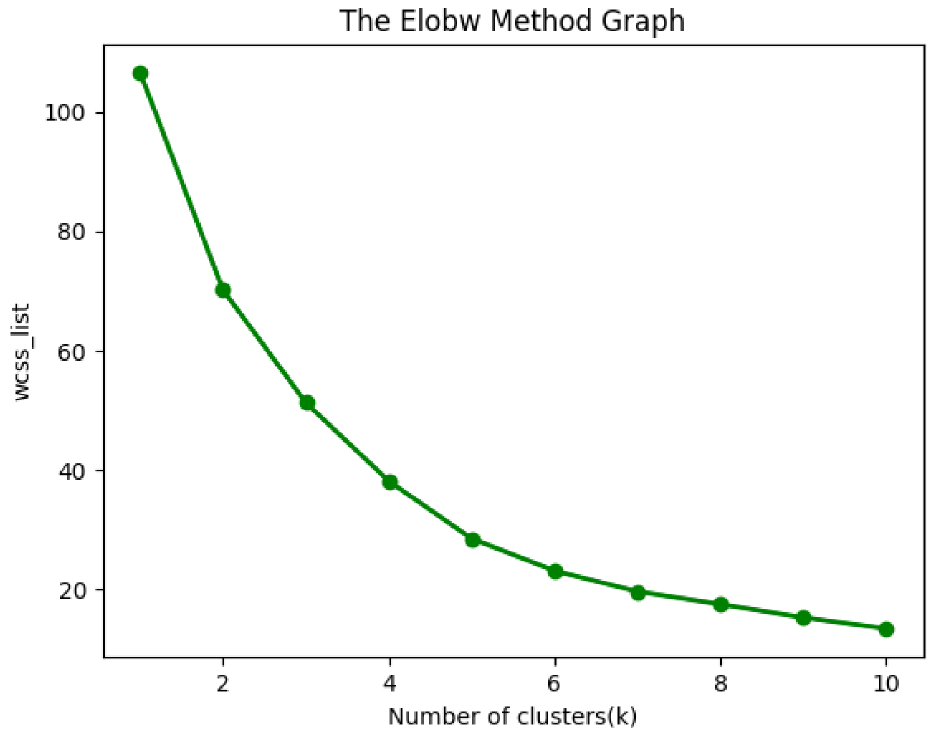
ML Model - 1: CLUSTERING

K Means clustering

K-Means Clustering is an Unsupervised Learning algorithm. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. he algorithm takes the unlabelled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm. There are 2 methods to determine k:

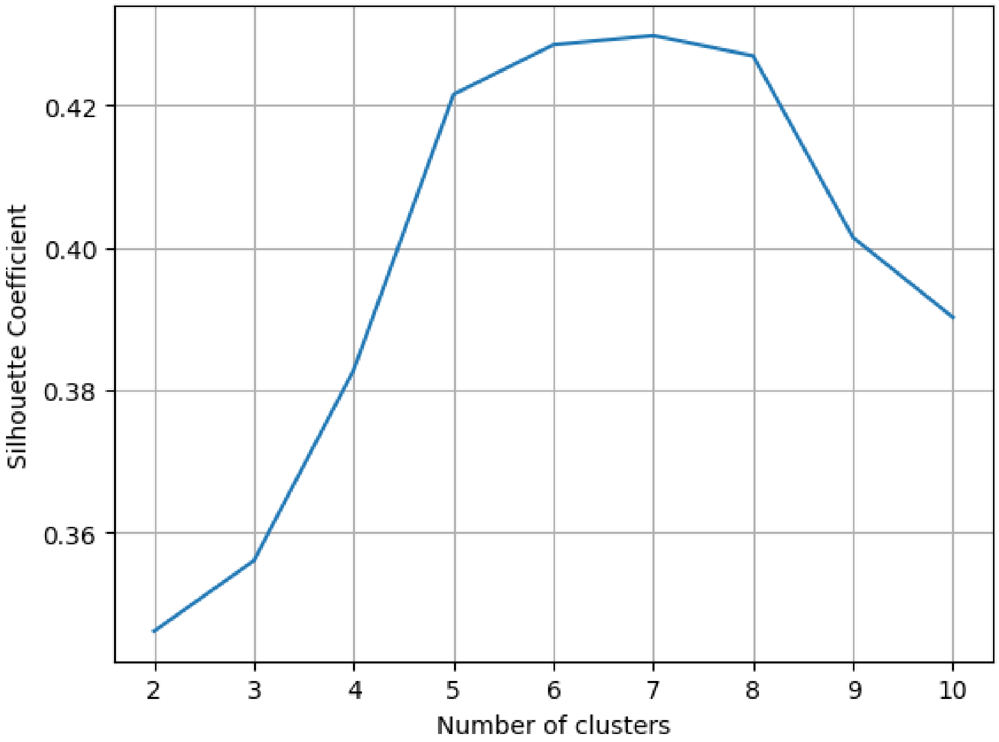
ELBOW METHOD

This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.



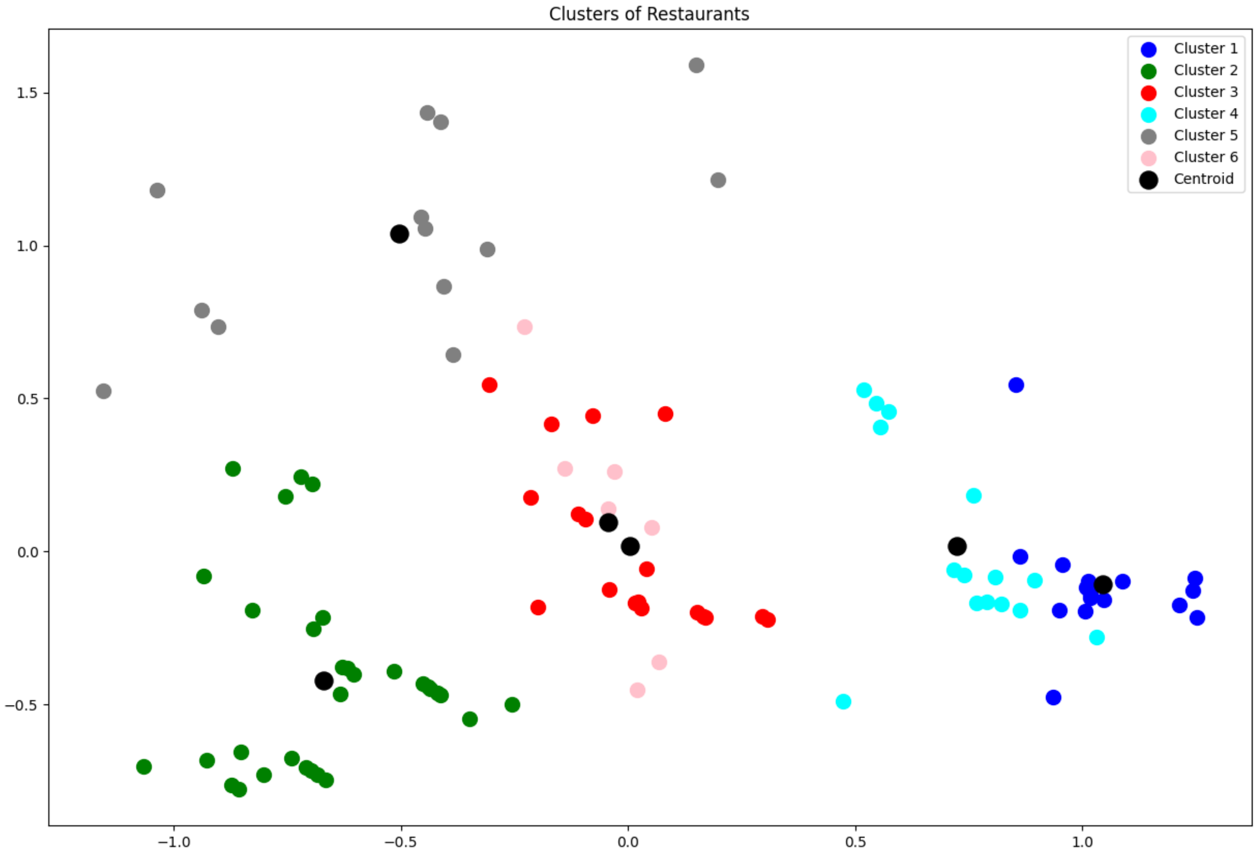
SILHOUETTE METHOD

The silhouette coefficient or silhouette score kmeans is a measure of how similar a data point is within-cluster (cohesion) compared to other clusters (separation).

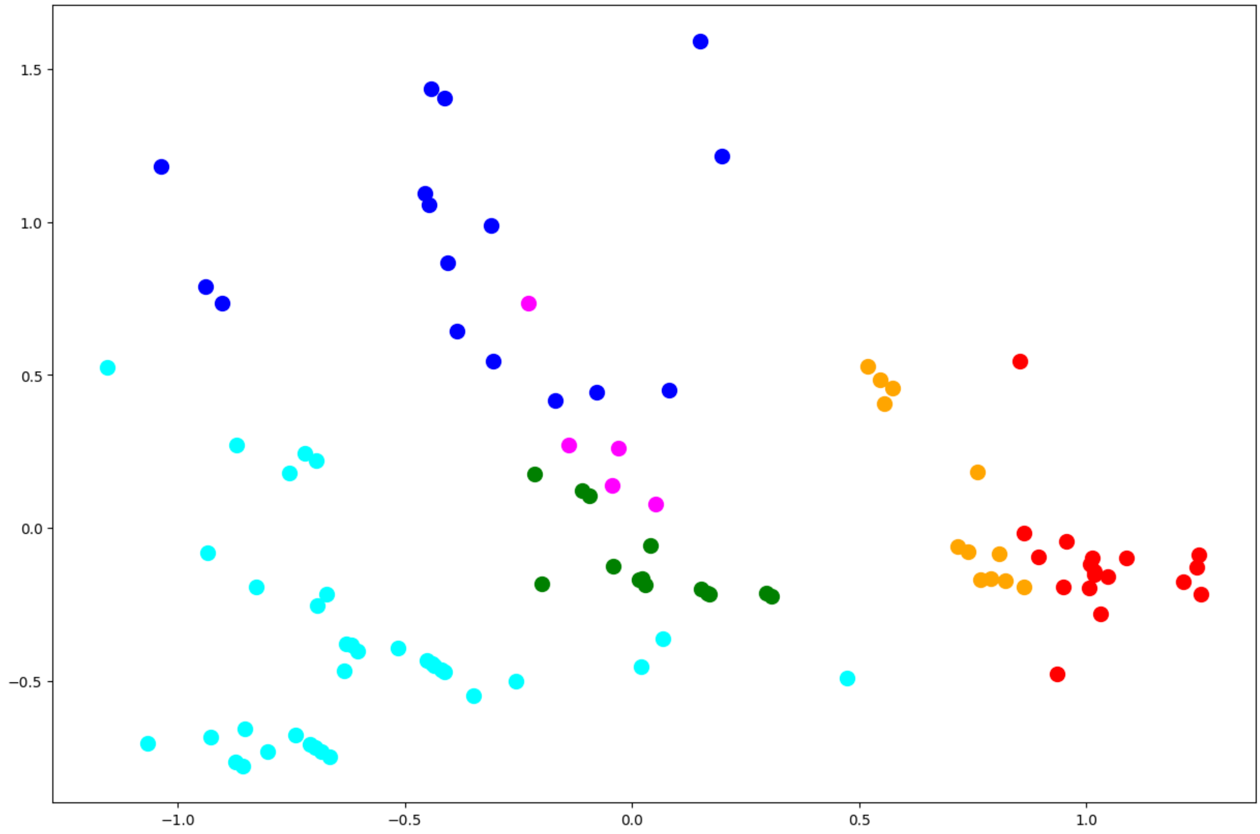


Here it is clearly visible that Silhouette coefficient is minimum for k=4. Hence optimal number of clusters must be 4.

K means clustering can be useful in customer segmentation based on demographic, psychographic and behavioural data as well as performance data to cluster your consumers for a particular product category. This can be eventually useful in making business strategies for different categories of customers and retain them for further growth of business.



ML Model - 2: HIERARCHICAL CLUSTERING



It creates groups so that objects within a group are similar to each other and different from objects in other groups. Clusters are visually represented in a hierarchical tree called a dendrogram.

There are two main types of hierarchical clustering:

Agglomerative: Initially, each object is considered to be its own cluster. According to a particular procedure, the clusters are then merged step by step until a single cluster remains. At the end of the cluster merging process, a cluster containing all the elements will be formed.

Divisive: The Divisive method is the opposite of the Agglomerative method. Initially, all objects are considered in a single cluster. Then the division process is performed step by step until each object forms a different cluster. The cluster division or splitting procedure is carried out according to some principles that maximum distance between neighbouring objects in the cluster.

Dendrogram in Hierarchical clustering

The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

Hierarchical Clustering can help an enterprise organize data into groups to identify similarities and, equally important, dissimilar groups and characteristics, so that the business can target pricing, products, services, marketing messages and more. Once the segments are identified, marketing messages and products can be customized for each segment. The better the segments chosen for targeting by a particular organization, the more successful the business will be in the market.

ML Model - 3: DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points. The most exciting feature of DBSCAN clustering is that it is robust to outliers. It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

ML Model - 4: SENTIMENTAL ANALYSIS

Sentiment analysis (or opinion mining) is a natural language processing (NLP) technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

Since we get a lot of reviews from the customers, it would be necessary to determine the sentiments of these reviews. We would be considered to build a supervised machine learning model to achieve the objective of determining the sentiments.

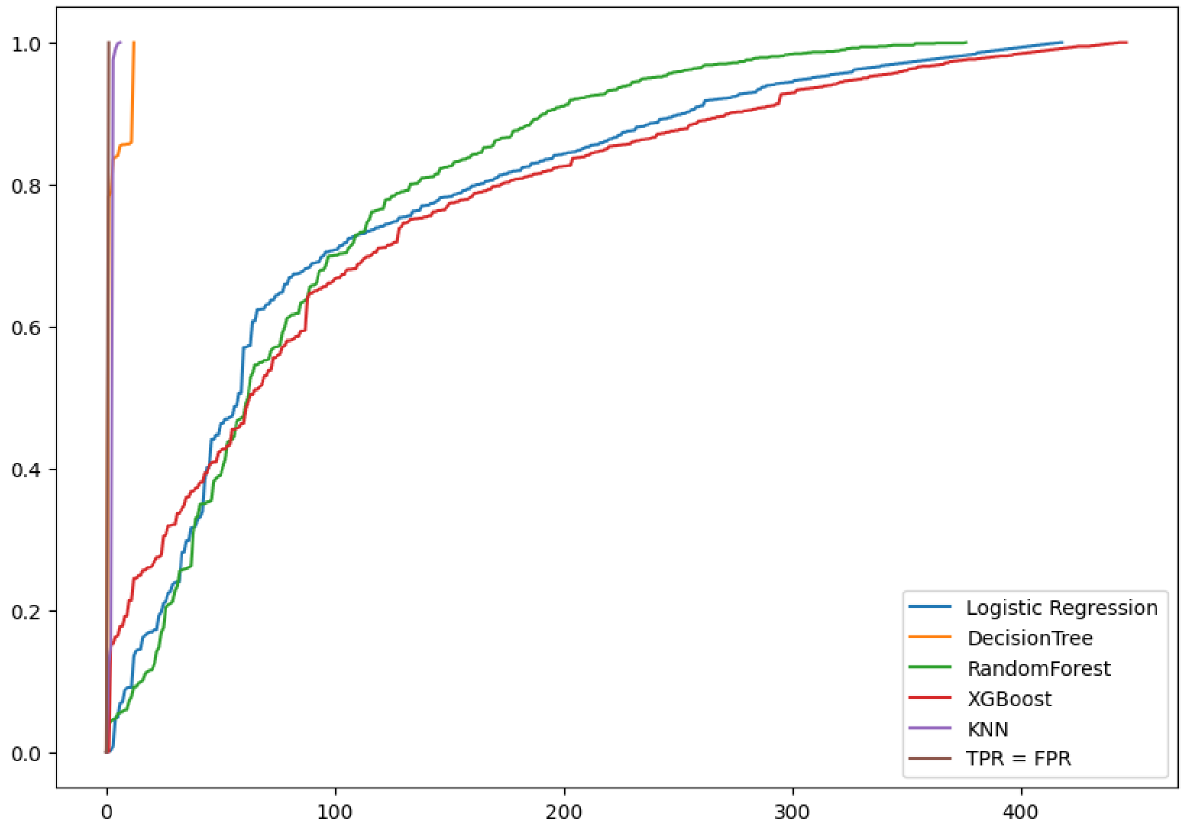
Model	: *Logistic Regression*
Train Accuracy	: 0.911281
Test Accuracy	: 0.850978
Test Precision	: 0.905405
Test Recall	: 0.852824
Test F1	: 0.878329
Test ROC_AUC Score	: 0.850325

Model	: *Decision Tree*
Train Accuracy	: 0.992471
Test Accuracy	: 0.784747
Test Precision	: 0.812689
Test Recall	: 0.856006
Test F1	: 0.833785
Test ROC_AUC Score	: 0.759525

Model	: *Random Forest*
Train Accuracy	: 0.992471
Test Accuracy	: 0.852484
Test Precision	: 0.844174
Test Recall	: 0.939539
Test F1	: 0.889307
Test ROC_AUC Score	: 0.821671

Model	: *XGBoost*
Train Accuracy	: 0.941649
Test Accuracy	: 0.861014
Test Precision	: 0.881026
Test Recall	: 0.901352
Test F1	: 0.891074
Test ROC_AUC Score	: 0.846736

Model	: *KNN*
Train Accuracy	: 0.977036
Test Accuracy	: 0.651279
Test Precision	: 0.648677
Test Recall	: 0.852824
Test F1	: 0.975338
Test ROC_AUC Score	: 0.536582



Observations:

From the above model metrics and ROC-AUC curve, we could consider that logistic regression is the best suitable model on this data, followed by XGBoost model and Random Forest.

KNN is the worst performing model for this data.

Hyperparameter tuning has to perform on logistic regression model to obtain a final model on this data

Hyperparameter tuning:

A) Logistic Regression

I have used Grid Search CV Hyperparameter optimization technique and tried to find the best values of C.I got best params 'C': 10. I have also used Cross validation with CV = 3.

After the hyperparameter tuning of Logistic Regression we observed the following improvements in the evaluation metrics.

Accuracy Before: 85.09% || Accuracy After: 86.00%

Precision Before: 90.54% || Precision After:85.89 %

Recall Before: 85.28% || Recall After: 86.00%

F1 Score Before: 87.84%|| F1 Score After: 85.89%

B) XGBoost

Taking much time than expected. Hence dropping it.

EVALUATION:

For sentiment analysis, evaluation metrics used were precision, recall, F1-score, and accuracy.

- Precision measures the proportion of true positive predictions among all positive predictions. It is a good metric to use when the cost of false positives is high.

- Recall (also known as sensitivity or true positive rate) measures the proportion of true positive predictions among all actual positive instances. It is a good metric to use when the cost of false negatives is high.

- F1-score is the harmonic mean of precision and recall, and is a good overall measure of a classifier's performance.

- Accuracy is the proportion of correctly classified instances among all instances.

The specific evaluation metric to use will depend on the specific use case and the relative costs of false positives and false negatives. For a positive business impact, F1-score can be considered as it balances the precision and recall to give an overall performance measure.

Logistic regression model has been chosen for final prediction because auc_roc score for logistic regression is highest among other models.

Chapter – XI

CONCLUSION

CONCLUSION:

Clustering and sentiment analysis were performed on a dataset of customer reviews for the food delivery service Zomato. The purpose of this analysis was to understand the customer's experience and gain insights about their feedback.

The proposed project has the potential to benefit both customers and the company. Customers can use the insights generated from sentiment analysis and clustering to make informed decisions about where to eat, while the company can use the insights to identify areas of improvement, tailor their business strategies, and ultimately improve their overall performance.

Overall, the proposed project can contribute to the field of restaurant recommendation systems and business analytics by leveraging advanced techniques such as sentiment analysis and clustering to provide insights that can drive improved decision-making.

Chapter – XII

REFERENCES

Here are some references:

- Huang, J., Rathod, P., Chen, S., & Patel, K. (2019). Analysis and visualization of restaurant reviews using supervised and unsupervised machine learning techniques. International Journal of Advanced Computer Science and Applications, 10(2), 184-189.

- Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

- Xu, J., Yu, Y., Zhang, Y., Xu, J., & Huang, H. (2019). A clustering-based approach to restaurant recommendation. In 2019 IEEE International Conference on Big Data (Big Data) (pp. 4335-4340). IEEE.

- Zhang, R., & Ye, Q. (2017). Analyzing online restaurant reviews using text mining approach. Journal of Hospitality Marketing & Management, 26(5), 523-545.

- Zhou, Y., Wu, X., Wu, L., & Huang, J. (2019). A clustering-based method for restaurant recommendation. IEEE Access, 7, 35145-35155.

- Zhu, J., Huang, J., & Hu, J. (2018). Using sentiment analysis to improve restaurant recommendations. Journal of Hospitality and Tourism Technology, 9(3), 387-398.

- Zomato API documentation. Retrieved from https://developers.zomato.com/api