Zomato Restaurant Clustering and Sentiment Analysis

Rahul
Chauhan
AlmaBetter, Bangalore

Abstract: This paper is intended to study clustering and sentiment analysis. Therefore, the unfolding of knowledge in texts is selected as the proper methodology to be followed and the steps are explained in order to reach the unsupervised clustering problem. After conducting an experiment with the most known methods of unsupervised clustering problem and the assessment of the results with the Silhouette index, it could be observed that the better grouping was with four groups. Natural Language Processing is one part of Artificial Intelligence and Machine Learning to make an understanding of the interactions between computers and human (natural) languages. Sentiment analysis is one part of Natural Language Processing, that often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not.

Keywords: Clustering, Elbow, Silhouette, Sentiment Analysis, Machine Learning, Natural Language Processing (NLP).

1. PROBLEM STATEMENT

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepender 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.

Attribute Information:

Zomato restaurant names and metadata

Use this dataset for clustering part

---

1. Name: Name of Restaurants
2. Links: URL Links of Restaurants

3. Cost: Per person estimated Cost of dining
4. Collection: Tagging of Restaurants w.r.t. Zomato categories
5. Cuisines: Cuisines served by Restaurants
6. Timings: Restaurant Timings

Zomato Restaurant review

Merge this dataset with Names and Metadata and then use for sentiment analysis part

---

1. Restaurant: Name of the Restaurant
2. Reviewer: Name of the Reviewer
3. Review: Review Text
4. Rating: Rating Provided by Reviewer
5. Metadata: Reviewer Metadata - No. of Reviews and followers
6. Time: Date and Time of Review
7. Pictures: No. of pictures posted with review

2. INTRODUCTION

Clustering analysis is unsupervised machine learning technique. What is the means of clustering? Clustering is **the task of dividing the population or data points into a number of groups such** that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. To cluster data points we have used k-means clustering. K-mean clustering uses “centroids”, k different randomly – initiated points in the data, and after every point has been assigned, the centroid is moved to the average of all of the points assigned to it. After clustering its important to finalize the number of clusters. To get perfect number of clusters there are different techniques available. One off name as elbow method. The elbow method runs k-means clustering on the dataset for a range of values for k and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to assigned center. Second method is silhouette score. Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. Another method for clustering is hierarchical clustering. In hierarchical clustering there are two types of technique one is agglomerative clustering (bottom-up approach) another is divisive (top-down approach). Hierarchical clustering, also known as hierarchical cluster analysis, is **an algorithm that groups similar objects into groups called clusters**. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Sentiment Analysis is the process of computationally determining whether a piece of writing is positive, negative or neutral. It’s also known as opinion mining, deriving the opinion or attitude of a speaker. Natural Language Processing is one part of Artificial Intelligence and Machine Learning to make an understanding of the interactions between computers and human (natural) languages. Sentiment analysis is one part of Natural Language Processing, that often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not. Zomato is an application for rating restaurants. 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 word to vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 150,000 reviews. After that make positive with

reviews that have a rating of 3 and above, negative with reviews that have a rating of 3 and below, and neutral who have a rating of 3. The author uses Split Test, 80% Data Training and 20% Data Testing. The metrics used to determine logistic regression are precision, recall, and accuracy. The accuracy of this research is 82%. The 10 words that affect the results are: “bad”, “good”, “average”, “best”, “place”, “love”, “order”, “food”, “try”, and “nice”.

3. EDA

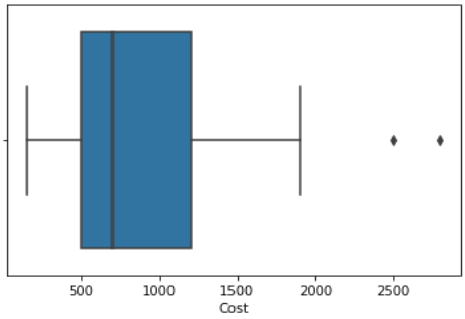
Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for **seeing what the data can tell us before the modeling task**. EDA is the **process of investigating the dataset to discover patterns, and anomalies (outliers)**, and form hypotheses based on our understanding of the dataset. EDA involves generating summary statistics for numerical data in the dataset and creating various graphical representations to understand the data better.

Our dataset contains 105 rows and 6 columns. There are 6 more columns in clustering dataset like name, links, cost, collection, cuisines, timings. To check null data, we use pandas library. We will fill the NAN values of collections with unknown. We will drop the names, links, timings column because we don’t have any use in our analysis.

Column Name	Null Values
Collections	54

After this we will check for duplicate data using duplicated method. It is important part of our data because many times same data is repeated so it will lead to wrong direction. It’s better to drop those data. we have only 2 duplicate data in our dataframe.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst to decide what will be considered abnormal. Before abnormal observation can be singled out, it is necessary to characterize normal observation. To identify outlier, we used box plot.



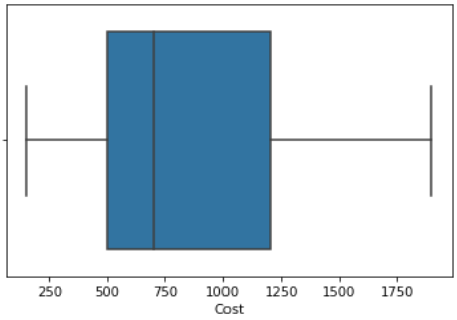


Figure 1: Cost (before imputing outlier) Figure 1.1: Cost (After imputing outlier)

We applied log transformation on skewed data. Log Transformation is pretty awesome. It makes our skewed original data i.e., High, Low, Open, Close columns more normal. It improves linearity between our dependent and independent variables. It boosts validity of our statistical analyses. Now our all columns are almost normally distributed.

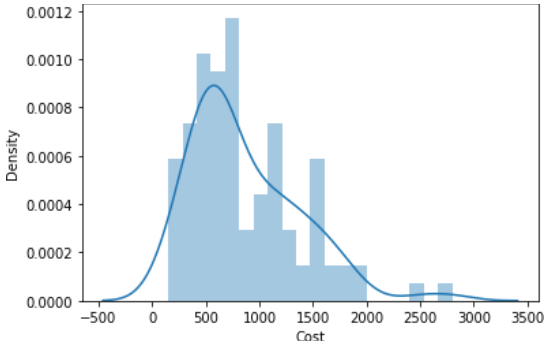


Figure 2: Cost (Skewed data)

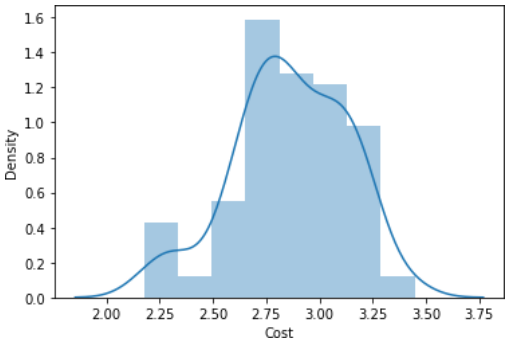


Figure 2.2: Cost (Normalized data)

After data cleaning we performed some tasks like five top cuisine and collections. Bottom five collection and cuisine.

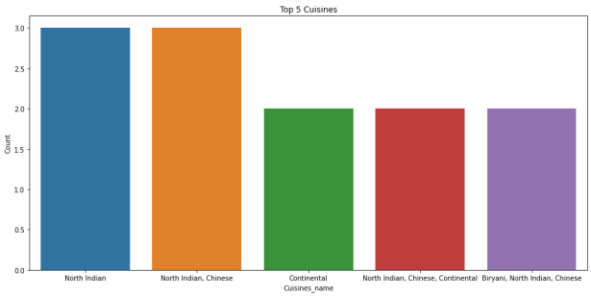


Figure 3: Top 5 Cuisines

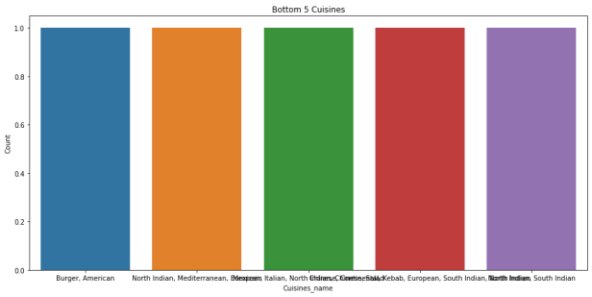


Figure 3.1: Bottom 5 cuisines

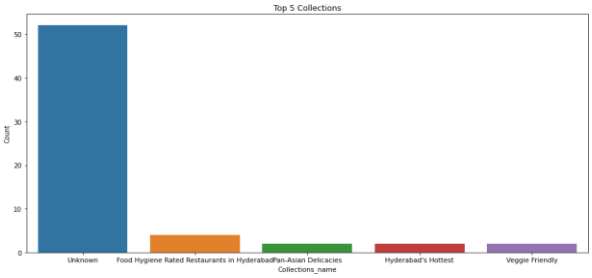


Figure 4: Top 5 collections

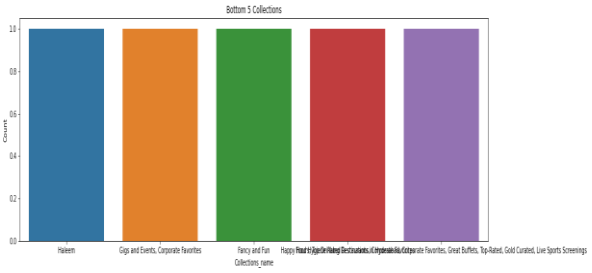


Figure 4.1: Bottom 5 collections

Next, we are trying to retrieve words which are used in our data set most of the time. We can see in left image, that words like venu, veggie, unknown, shawarma etc. used in Collection. In right side of image sushi, Spanish, south, seafood, salad, pizza, north, Momo, Mexican, kebab etc. words used in cuisines.

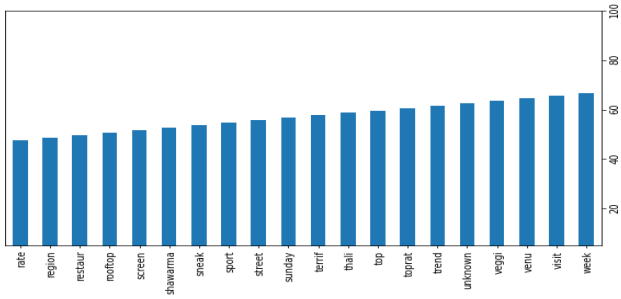


Figure 5 : Top Collection vocab

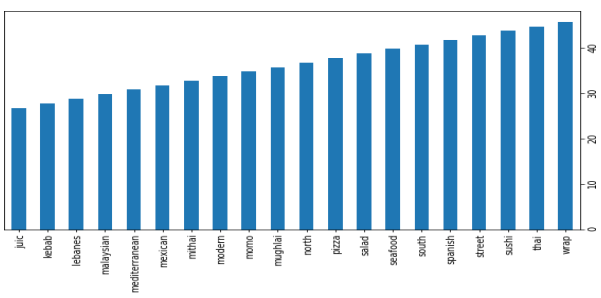


Figure 5.1: Top Cuisines vocab

4. Clustering Algorithms

Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observation belonging to all the clusters:

Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a.

Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be called a mean nearest-cluster distance. The mean distance is denoted by b.

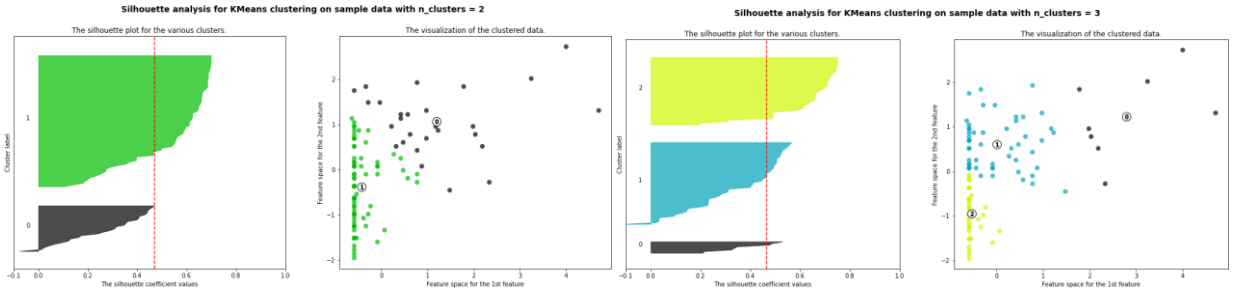


Figure 6: Silhouette score (n_cluster=2)

Figure 6.1: Silhouette score (n_cluster=3)

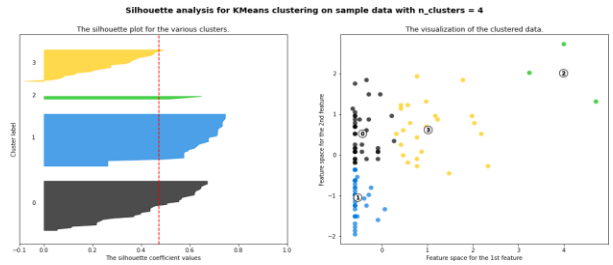


Figure 6.2: Silhouette score (n_cluster=4)

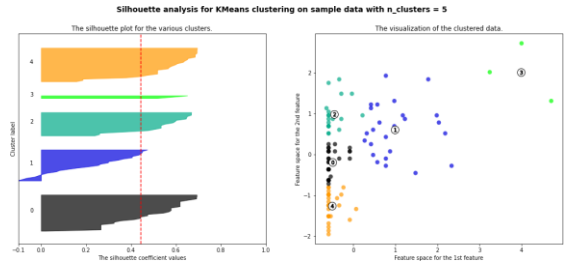


Figure 6.3: Silhouette score (n_cluster=5)

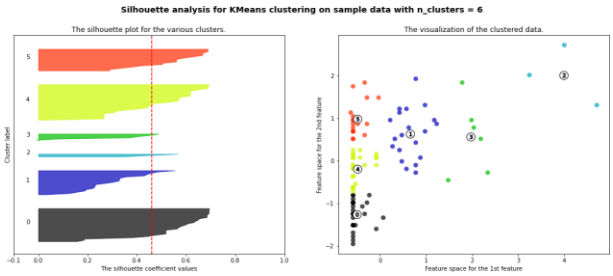


Figure 6.4: Silhouette score (n_cluster=6)

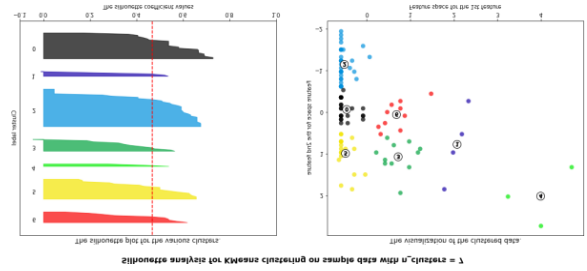


Figure 6.5: Silhouette score (n_cluster=7)

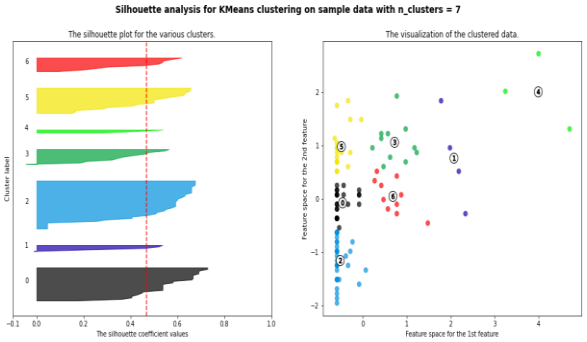


Figure 6.6: Silhouette score (n_cluster=7)

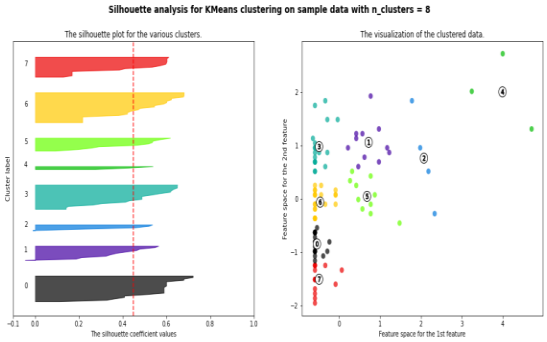


Figure 6.7: Silhouette score (n_cluster=8)

We have also used Elbow method to get perfect number of cluster. we got 4 number of cluster best suites to our dataframe.

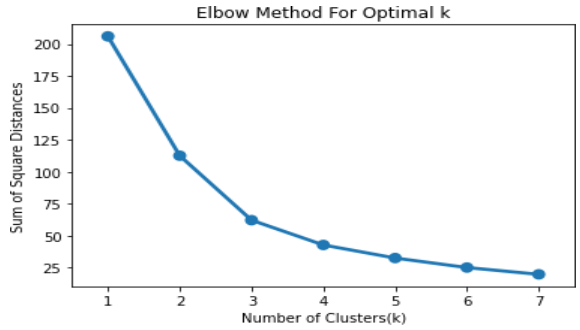


Figure 7: Elbow curve

K = 4, we draw k means clustering. Here we can see 4 black dots as centroid of clusters.

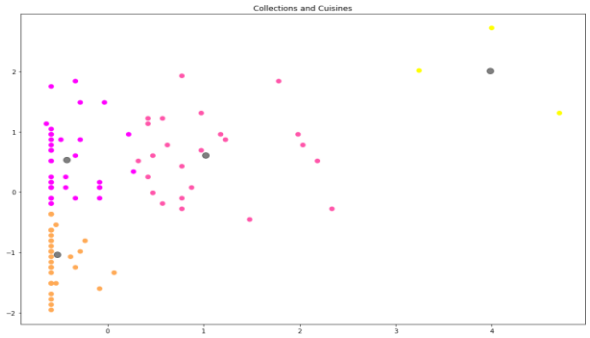


Figure 8: K-Mean Cluster

A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is **to work out the best way to allocate objects to clusters.**

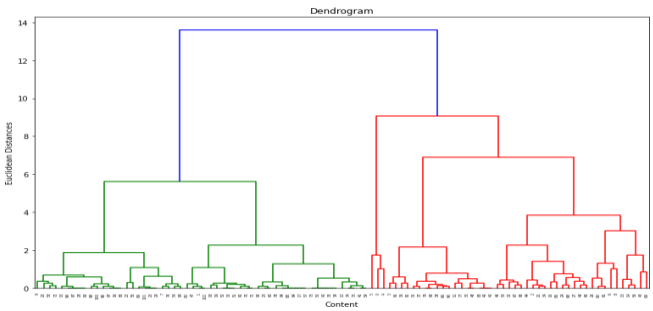


Figure 9: Dendrogram

The agglomerative clustering is **the most common type of hierarchical clustering used to group objects in clusters based on their similarity**.

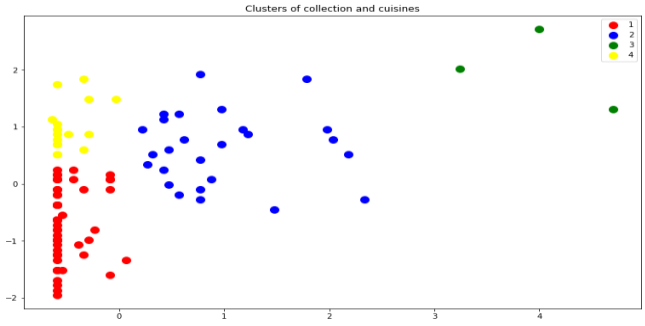


Figure 10: Hierarchical clustering

5. Sentiment Analysis

Sentiment Analysis is the process of computationally determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker.

Natural Language Processing is one part of Artificial Intelligence and Machine Learning to make an understanding of the interactions between computers and human (natural) languages.

Sentiment analysis is one part of Natural Language Processing, that often used to analyze words based on the patterns of people in writing to find positive, negative, or neutral sentiments. Sentiment analysis is useful for knowing how users like something or not.

Zomato is an application for rating restaurants. 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 word to vector with the term frequency-inverse document frequency (TF-IDF). The data that we process are 10,000 reviews. After that make positive with reviews that have a rating of 3 and above, negative with reviews that have a rating of 3 and below, and neutral who have a rating of 3. we use Split Test, 75% Data Training and 25% Data Testing. We have applied logistic regression on reviews dataset. getting 82% Accuracy on model. The 10 words that affect the results are: "bad", "good", "average", "best", "place", "love", "order", "food", "try", and "nice".

In sentiment analysis we have 7 features. Restaurant, reviewer, review, rating, metadata, time, pictures. We calculated the percentage of restaurants ratings.

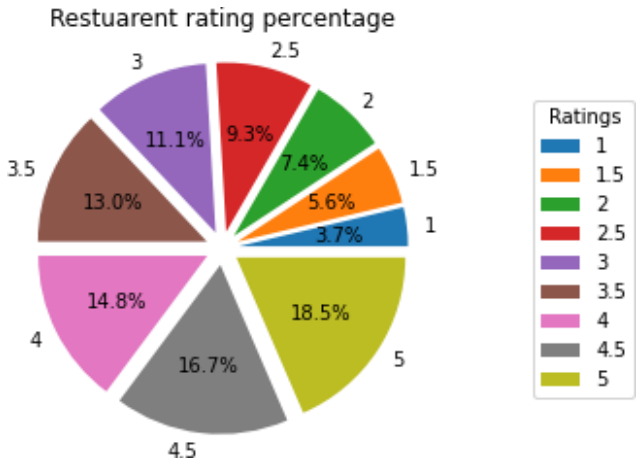


Figure 11: Restaurant rating percentage

- **Logistic Regression**

Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. There are 4 types of logistic regression BINARY, ORDINAL, NOMINAL, POISSON logistic regression. We used ordinal logistic regression. Number of categories 3 or more. Characteristic are at natural ordering of the levels. Like in our case we categories our ratings in 3 types i.e., bad, average and good. We used logistic regression to classify the ratings.

Accuracy: 0.8224186420249096

Precision: 0.8929315433850494

Recall: 0.8224186420249096

- **Word Cloud**

Word Clouds are **visual displays of text data – simple text analysis**. Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language (“a”, “an”, “the” etc.).

Figure 12: word cloud (good)

Figure 12.1: word cloud (average)

Figure 12.2: Word Cloud (bad)

6. Conclusions

- Top 5 cuisines are

  i) North Indian, Chinese ii) North Indian iii) Continental iv) Ice Cream, Desserts v) Fast Food

- Bottom 5 cuisines are

  i) American, Fast Food, Salad, Burger ii) Continental, Italian, North Indian, Chinese iii) North Indian, Italian, Continental, Asian iv) Mexican, Italian, North Indian, Chinese, Salad v) Momos

- Top 5 collections are

  i) Unknown ii) Food Hygiene Rated Restaurants in Hyderabad iii) Great Buffets iv) Trending This Week v) Hyderabad's Hottest

- Bottom 5 collections are

  i) Sneak Peek Hyderabad ii) Best Milkshakes iii) Happy Hours, Top-Rated, Gold Curated iv) Best Bakeries v) Great Breakfasts, Late Night Restaurants

- In collection top 3 vocab used is week, visit and veggie.
- In cuisines top 3 vocab used is wrap, Thai and sushi.
- For n_clusters = 4, we get highest silhouette score is 0.4722959202437076
- From elbow method we get 4 number of clusters is best among all.
- Used dendrogram to find optimal number of clusters
- Applied agglomerative hierarchical clustering from this we find 4 number of cluster good fit our model.
- By applying different clustering algorithm to our dataset. we get the optimal number of clusters is equal to 4.
- we have categorized rating in 3 types i.e., good, bad and average. 4500+ good, 1700+ bad and 900+ average ratings given by customer.
- We have applied logistic regression on reviews dataset. getting 82% Accuracy on model.