ZOMATO RESTAURANTS AND DETAILS:

INTRODUCTION ABOUT THE COMPANY:

Zomato is a multinational food delivery and restaurant discovery platform that was founded in 2008 by Deepinder Goyal and Pankaj Chaddah in India. The company started as an online restaurant directory, but has since expanded to include food delivery, table reservations, and online ordering.

Zomato operates in over 25 countries and has a presence in more than 10,000 cities worldwide. It offers a comprehensive database of restaurants, menus, and reviews that can help users find their ideal dining experience.

The company has a user-friendly mobile app that allows users to search for restaurants by location, cuisine, price, and user reviews. The app also offers features like online ordering, table reservations, and food delivery.

In addition to its food delivery and restaurant discovery services, Zomato has also launched a grocery delivery service, Zomato Market, in select cities.Zomato has received multiple rounds of funding from various investors, including Alibaba Group, Temasek Holdings, and Sequoia Capital. In 2021, the company went public and listed on the Indian stock exchange.

Overall, Zomato has become a leading player in the online food delivery and restaurant discovery market, with a strong presence in several countries around the world.

OBJECTIVE OF THE ZOMATO DATASET:

The objective of a project using the Zomato restaurants dataset could be to gain insights into the restaurant industry, specifically related to cuisine, timings, and cost. By analyzing the dataset, we could potentially answer questions such as:

What are the most popular cuisines in different regions?
How do the costs of restaurants vary across different regions?
What are the busiest times for restaurants?
Are there any correlations between cuisine, and cost?

Overall, the objective of the project could be to use data analysis to gain a better understanding of the restaurant industry and potentially provide recommendations for restaurants or other stakeholders in the food industry.

DATA CLEANING AND ANALYSIS :

Drop irrelevant columns: We dropped columns such as 'phone number' and 'address' which were not necessary for our analysis.

Remove duplicates: We removed any duplicate rows from the dataset.

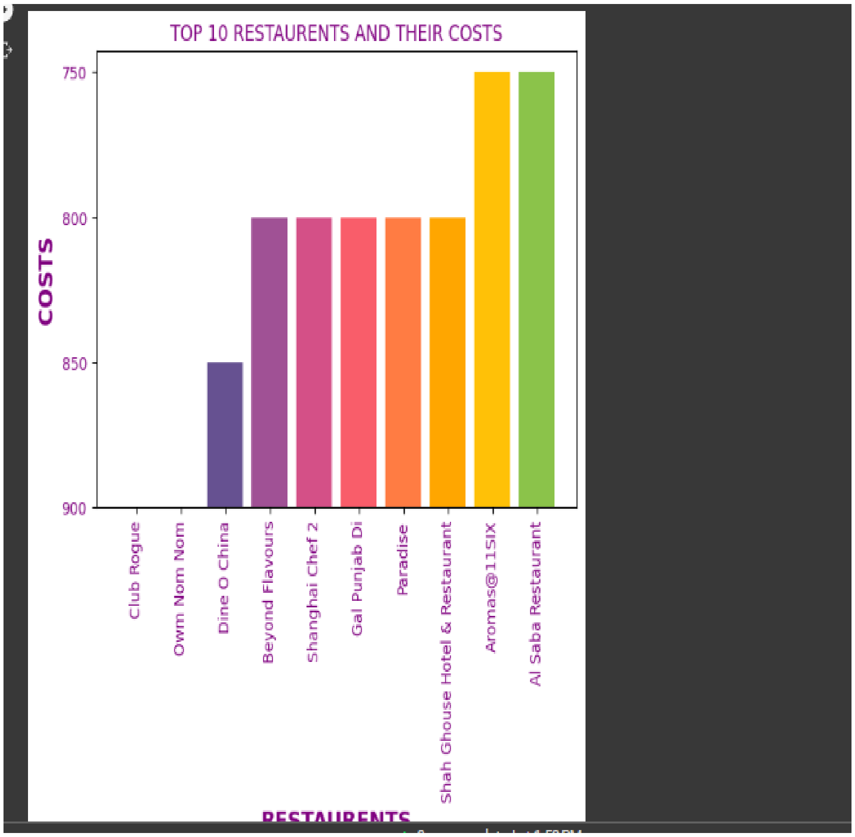
Handle missing values: We checked for any missing values and imputed them where possible.

Convert data types: We converted any date/time columns to the appropriate format.For analyze the busiest timing.

After cleaning and processing the data we have 104 rows and 6 columns

ANALYZING THE DATA:

Visualizing the data of top 10 restaurants and their costs by bar graph



1.Above the bar graph shows the top 10 restaurants based on their costs .

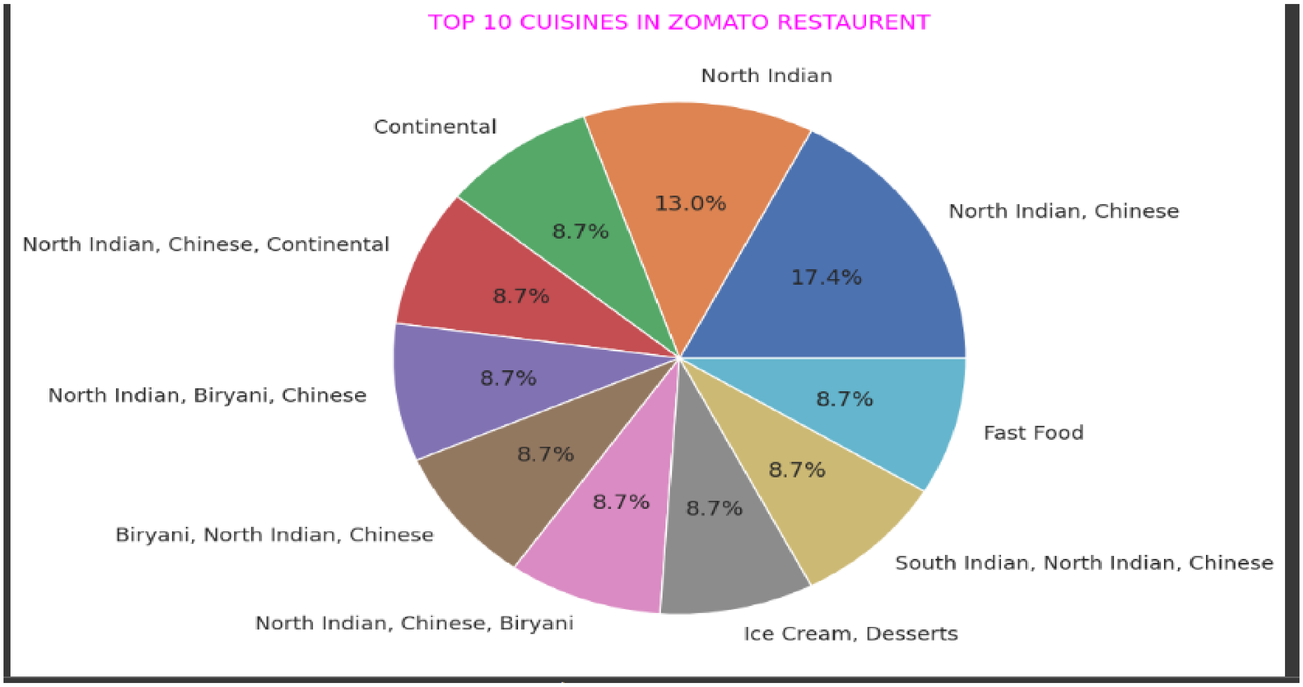
2.Here it show the restaurants names on the x axis and the costs on the y axis .

3.It is plotted in such a way that descending order of the costs and restaurants based on that.

4.The top 10 restaurants of costs vary from the 850 to 750 rupees.

5.Bases on this information we can say that there are most of the restaurants have the costs vary from the 850 and 750 .

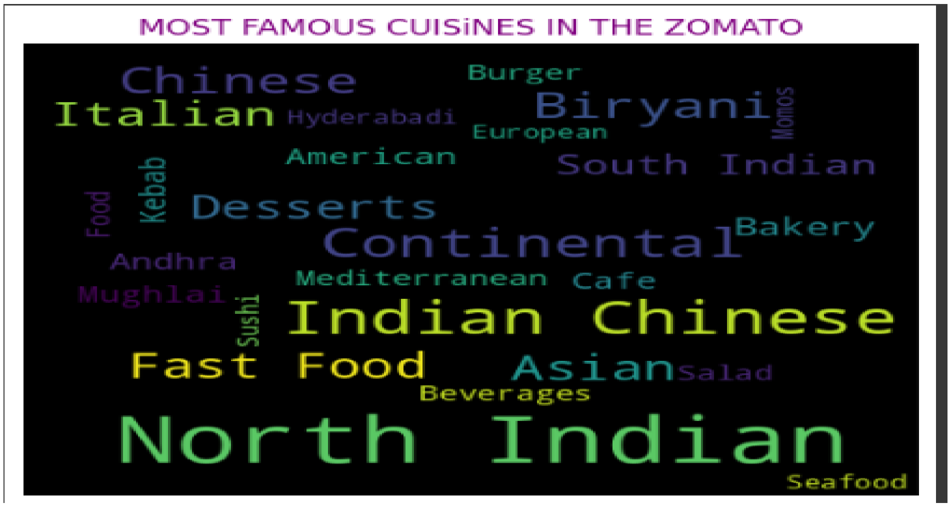
VISUALIZING THE TOP 10 CUISINES IN THE ZOMATO RESTAURANT:



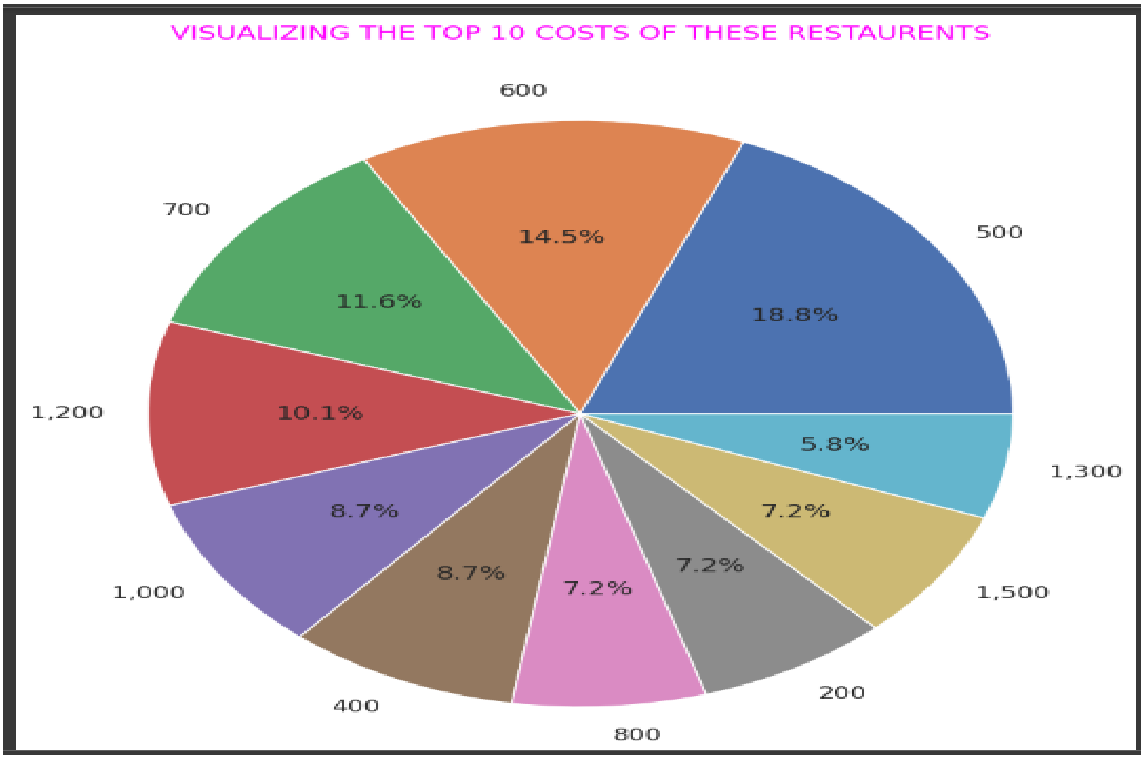
1.The pie chart shows about the top 10 cuisines famous in the zomato restaurant and their Percentages.

2.Based on the information we can say that cuisine names called North india and North indian,chinese are the top most ones among all other cuisines.

VISUALIZING THE MOST FAMOUS CUISINES IN THE RESTAURANTS BY WORD CLOUD:



VISUALIZING THE COSTS IN THE RESTAURANTS :



1.Above the piechart shows about the information about the top 10 costs of the restaurants .

2.From the above information we can analyze that the costs of the restaurants contain 500 and 600 and also 700.

3.Totally the costs of the restaurants are vary from the 500 to 1500 .

CONCLUSION BASED ON THE DATA VISUALIZATION:

Here are some insights that gained from the dataset are:

1.Popular cuisines: You may be able to identify which cuisines are most popular in the area covered by the dataset. This could help inform decisions about what types of restaurants to open in the area.

2.Price ranges: You may be able to see which price ranges are most common in the area covered by the dataset. This could help inform decisions about setting prices for new restaurants.

3.Restaurant names: You may be able to identify which restaurants have the highest customer ratings or are the most popular in the area covered by the dataset. This could help inform decisions about marketing and branding for new restaurants.

USING THE KMEANS CLUSTERING ALGORITHM FOR THE DATASET:

REASONS WHY I USED KMEANS CLUSTERING ALGORITHM FOR THE ZOMATO DATASET:

There are many different clustering algorithms that can be used for a dataset like Zomato, and the choice of which algorithm to use depends on the specific characteristics of the data and the goals of the analysis. However, K-means is a commonly used clustering algorithm that is well-suited for datasets like Zomato. Here are some reasons why K-means may be a good choice for clustering the Zomato dataset:

1.K-means is computationally efficient: K-means is computationally efficient and can handle large datasets, making it a good choice for situations where there are many observations or features.

2.K-means is suitable for datasets with continuous variables: K-means works well with continuous variables, which is often the case in datasets like Zomato that involve prices, ratings, and other quantitative features.

3.K-means is suitable for identifying natural clusters: K-means can be effective in identifying natural clusters in the data, which can be useful for segmentation or identifying patterns and trends.

4.K-means mainly used for the grouping of the similar data into a cluster based on the information provided to it.

5.It is the type of the unsupervised learning and helps in grouping the data on the similarities between them and make them into a clusters and also assign a centroid for each cluster based on the information we provided it contains the nearest data points to it.

6.By this KMeans clustering we can personalized recommendations based on the clustering of restaurants, Zomato can improve the user experience on their platform. Users will be more likely to use their platform again if they receive personalized recommendations.

7.This are the some reasons for using the kmeans clustering for the zomato dataset.

ANALYSIS ON DATA USING THE CLUSTERING:

1.Before using of the clustering algorithm we have to do cleaning and processing of the data.

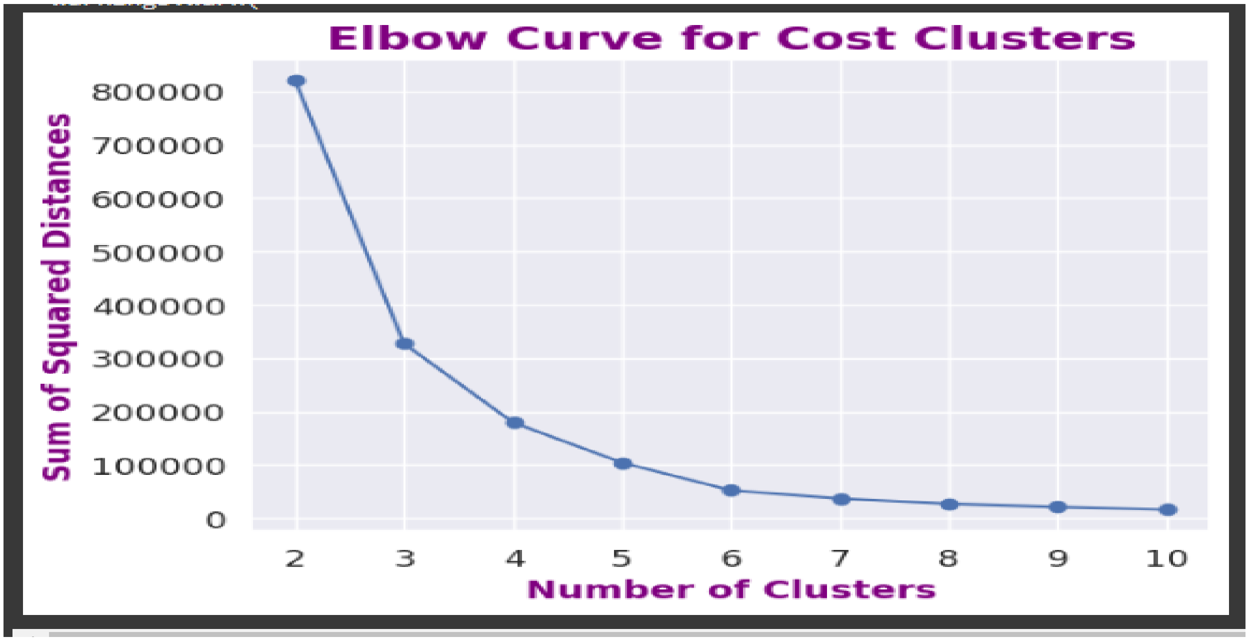
2.After the we have to import the important libraries used for analysis on the data .

Example: from sklearn.cluster import KMeans.

3.To make the cost clusters we have to use the elbow method:it is a heuristic used to determine the optimal number of clusters for a given dataset. It works by plotting the within-cluster sum of squares (WCSS) against the number of clusters, and identifying the "elbow" point in the plot where the rate of decrease in WCSS starts to level off. This is taken as an indication of the optimal number of clusters.

4.For making the cost clusters we have to check the all the missing values in the cost column and also any nan values or infinity values in the dataset .if there we have to clean the data and Replace the missing values with the true values and also for nan values before using the algorithm and also for elbow method in the analysis.

ELBOW METHOD GRAPH:

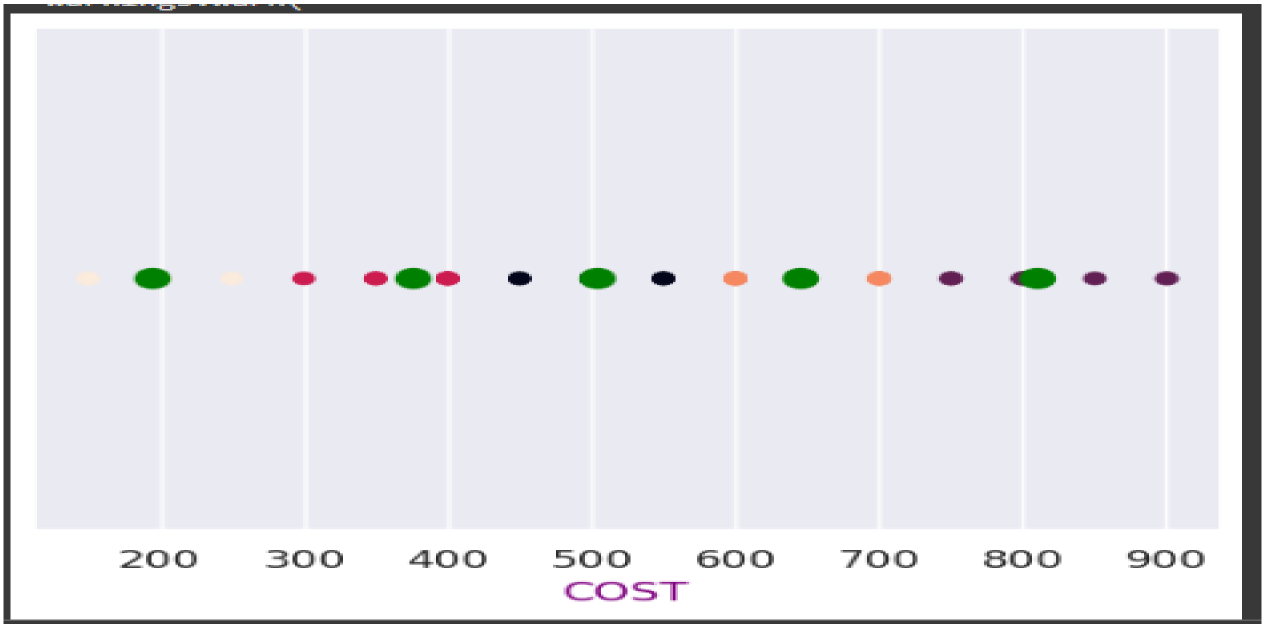


5.Based on the above graph we can make the cost column into the 5 clusters.

6.Now we have fit the data according to the optimal number of clusters.

7.After fitting the data we have to plot the data .

GRAPH OF COST CLUSTERS:



8.Based on the graph we can see the cost clusters that is 5 clusters green colour dots indicates

the centroids of the each clusters .

9.Now for further analysis avarage of the each cluster in the above graph are:

Cluster 1:   502.777778
Cluster2: 809.090909
Cluster 3:   375.000000
Cluster 4:   644.736842
Cluster 5:   193.750000

10.we can observe the average rate of cost is high for the second cluster we can say that most

of the costs are in range of 300 to 400 which common people can afford the expenses.

11.Least average is for the cost ranges between the 800 to 900 .Least restaurents are offering

such prices .

USING OF THE HIERARCHICAL CLUSTERING FOR THE ANALYSIS:

Hierarchical clustering is another type of clustering algorithm that works by creating a hierarchy of clusters. In this method, the algorithm starts by considering each point as an individual cluster and then iteratively merges the closest pair of clusters until only one cluster remains.

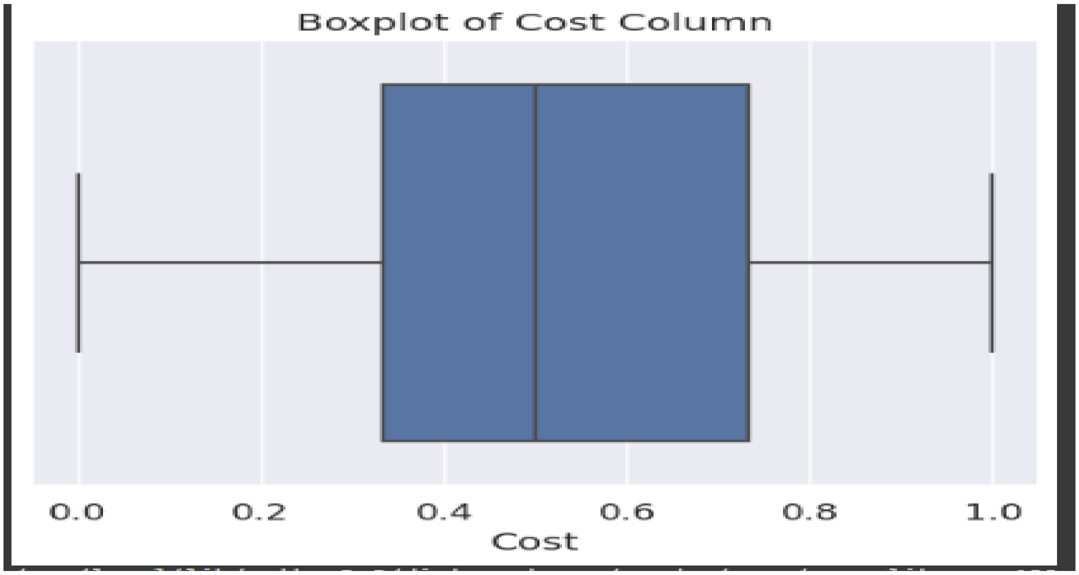
1.Like earlier we have to process the data before using the hierarchical clustering for the dataset.

2.For more accuracy of the output we have remove the outliers whIch may cause the inappropriate results if they were not removed in the data.

3.For removing the outliers we have to check for the outliers in the data by using the box plotting By importing seaborn module and plotting the boxplot for outliers.

4.Then, outliers are removed from the Cost column using the IQR method. Finally, the Cost column is transformed using a logarithmic transformation.

GRAPH OF THE OUTLIERS:



5.The above graph indicates the line inside the box represents the median value of the data.

The median is the middle value when the data is sorted in ascending or descending order. It is a measure of central tendency that is less sensitive to outliers compared to the mean.

6.The two lines extended from the box plot are called whiskers. They extend from the box to the

smallest and largest observations within 1.5 times the interquartile range (IQR) of the box. Any observations that fall outside the whiskers are considered outliers.

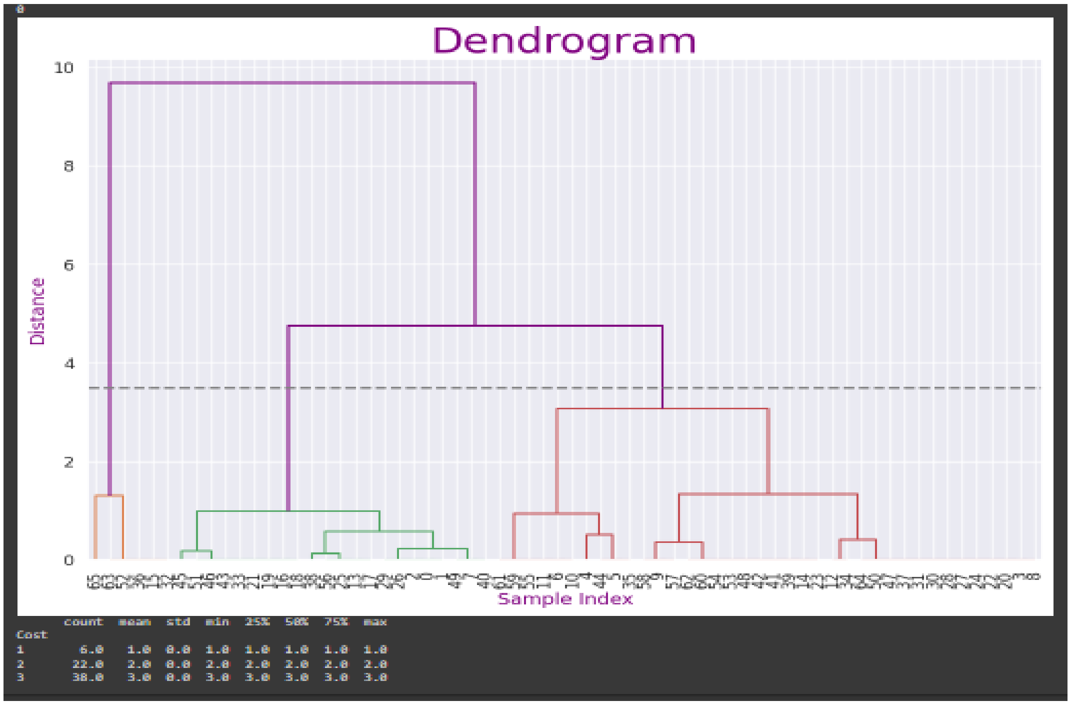
7.After removing the outliers we have to apply the algorithm.

8.For using the hierarchical clustering using we have to import the important libraries .

eg:from scipy.cluster.hierarchy import dendrogram,linkage,fcluster

from sklearn.preprocessing import StandardScaler as ss

9.We use the dendrogram plot for the data analysis :.



```
        count  mean  std  min  25%  50%  75%  max
Cost
1         6.0   1.0  0.0  1.0  1.0  1.0  1.0  1.0
2        22.0   2.0  0.0  2.0  2.0  2.0  2.0  2.0
3        38.0   3.0  0.0  3.0  3.0  3.0  3.0  3.0
```

EXPLANATION:

1**.The different colors of branches in a dendrogram represent different clusters or groups that are formed during the clustering process. When the algorithm groups similar data points together, it creates a new branch or cluster, which is represented by a different color in the dendrogram. The different colors help to differentiate between different clusters in the** dendrogram.

2.A big branch in a dendrogram indicates a larger distance between clusters. This means that the data points in those clusters are more dissimilar than the data points in other clusters that are connected by smaller branches. The bigger the branch, the greater the distance between the clusters.

3.Based on the dendrogram there divided the clusters into the 5 as kmeans it indicates the correct results.

USING THE ASSOCIATION RULE MINING ALGORITHM FOR THE DATA:

REASONS:

Association rule mining is a technique used to find patterns in data sets that show relationships between variables. It is commonly used in market basket analysis to identify items that are frequently purchased together. It works by identifying frequent itemsets, which are sets of items that appear together in a certain percentage of transactions, and generating association rules.

which are statements that describe the relationships between those items.

1.Before using this algorithm we have to do data cleaning and processing as mentioned earlier .

2.After processing the data we have to import the important libraries for the processing the data .

3.After that create a binary matrix for the restaurants offer each cuisine.

OUTPUT:

```
                          antecedents  \
1585     ( Continental,  North Indian,  European)
1063              ( South Indian,  European)
277             ( Hyderabadi,  North Indian)
1076                ( Continental, Chinese)
1071      ( Continental,  North Indian, Chinese)
```

```
1068                              ( European)
1553  ( Continental,  South Indian,  Kebab, Chinese)
1555       ( Continental,  South Indian,  European)
1556           ( South Indian,  European, Chinese)
1557            ( South Indian,  Kebab,  European)

                         consequents  antecedent support  \
1585                   ( Kebab, Chinese)           0.01087
1063              ( Continental, Chinese)           0.01087
277                              (Andhra)           0.01087
1076           ( North Indian,  European)           0.01087
1071                         ( European)           0.01087
1068  ( Continental,  South Indian, Chinese)           0.01087
1553                         ( European)           0.01087
1555                   ( Kebab, Chinese)           0.01087
1556               ( Continental,  Kebab)           0.01087
1557              ( Continental, Chinese)           0.01087

     consequent support  support  confidence  lift  leverage  conviction
1585           0.01087  0.01087        1.0  92.0  0.010751        inf
1063           0.01087  0.01087        1.0  92.0  0.010751        inf
277            0.01087  0.01087        1.0  92.0  0.010751        inf
1076           0.01087  0.01087        1.0  92.0  0.010751        inf
1071           0.01087  0.01087        1.0  92.0  0.010751        inf
1068           0.01087  0.01087        1.0  92.0  0.010751        inf
1553           0.01087  0.01087        1.0  92.0  0.010751        inf
1555           0.01087  0.01087        1.0  92.0  0.010751        inf
1556           0.01087  0.01087        1.0  92.0  0.010751        inf
1557           0.01087  0.01087        1.0  92.0  0.010751        inf
```

EXPLANATION:

This output is the result of association rule mining, which is a technique used in data mining to discover relationships between variables in large datasets. In this particular case, the output shows ten association rules, each with an antecedent (a set of items that occur together in the dataset) and a consequent (another set of items that also occur together in the dataset).

The output also shows several measures of the strength of these association rules. The

antecedent support is the proportion of transactions (or observations) in the dataset that contain

the items in the antecedent. The consequent support is the proportion of transactions in the

dataset that contain the items in the consequent. The support is the proportion of transactions in

the dataset that contain both the antecedent and consequent.

The confidence is the conditional probability that the consequent occurs given that the

antecedent also occurs. The lift is a measure of the strength of the association between the

antecedent and consequent, compared to what would be expected if they were independent.

The leverage is another measure of the strength of association, which compares the observed

frequency of both the antecedent and consequent to what would be expected if they were

independent. The conviction is a measure of how much the antecedent implies the

consequent In this output, all the association rules have a confidence of 1.0, meaning that the

consequent always occurs when the antecedent occurs. The lift is also very high, with a value of 92.0 for

all rules, indicating a very strong association between the antecedent and consequent. The

leverage and conviction measures are also very high, indicating that the antecedent is a

strong predictor of the consequent.

CONCLUSION:

1.From the above observations we can conclude that clustering based on cost can help to

identify pricing trends in different regions or neighborhoods, which can inform business

decisions such as setting prices or expanding to new locations.

2.Another possible conclusion is that there may be distinct clusters of restaurants at different price points, which may have different customer preferences and behaviors.

3.From association rule mining we can conclude the relationship between the different attributes of the restaurants,

Association between different cuisine types: Association rule mining can reveal which cuisines are often associated with each other in the dataset. For example, the analysis might reveal that restaurants serving North Indian cuisine are often associated with South Indian cuisine or that restaurants serving Continental cuisine are often associated with European cuisine.

THANK YOU