**Unsupervised Capstone Project-4 Submission**

**ZOMATO RESTAURANT CLUSTERING AND SENTIMENTAL ANALYSIS**



GitHub Link: -

https://github.com/rskamit11/ZOMATO-UNSUPERVISED-CAPESTONE-PROJECT

**Introduction**

This project entailed the utilization of advanced data analytics techniques to gain a deeper understanding of the restaurants and customer feedback on the popular online food delivery platform, Zomato.

The data procured included information such as the restaurant's name, location, cuisines, average cost for two, ratings, and user reviews.

Subsequently, I embarked on the task of data cleaning and preprocessing, which involved the elimination of duplicate entries, addressing missing values, and transforming the data into a format amenable to analysis.

The next step in the project was the implementation of clustering on the restaurant data through the use of the k-means algorithm. The objective of the clustering was to group similar restaurants together and discern patterns within the data. The features employed for the clustering process included the restaurant's location, cuisines, and average cost for two. The number of clusters was determined by utilizing the elbow method.

I then proceeded to conduct sentiment analysis on the user reviews to gain a comprehensive understanding of the overall sentiment towards the restaurants. Certain libraries were utilized to classify the reviews as positive, negative, or neutral. Additionally, I extracted the most recurrent words utilized in the reviews and visualized them through the creation of word clouds.

The outcome of the analysis revealed that the restaurants within the city were grouped into five clusters based on their location, cuisines, and average cost for two. The sentiment analysis uncovered that, generally, customers held a positive sentiment towards the restaurants.

In conclusion, this project exemplifies the utility of clustering and sentiment analysis in gaining a more profound comprehension of restaurant data on Zomato. The insights procured from the analysis can be of immense benefit to both restaurants and customers in making informed decisions. Furthermore, the project can be extended to other cities or even countries to gain insight into the eating habits and preferences of individuals in different regions.

**Problem Statement**

The problem statement for this project is to analyze and understand the restaurant industry in India by utilizing data from the Indian restaurant aggregator and food delivery start-up, Zomato. The project aims to gain insights into the sentiments of customer reviews, cluster Zomato restaurants into different segments, and analyze the data to make useful conclusions in the form of visualizations. The data analyzed includes information on cuisine, costing, and customer reviews. The project aims to assist customers in finding the best restaurant in their locality and aid the company in identifying areas for growth and improvement in the industry. Additionally, the project aims to use the data for sentiment analysis and identifying critics in the industry through the metadata of reviewers.

**In this project, you are required to do**

1. Exploratory Data Analysis
2. Understanding what type content is available.
3. Clustering similar content by matching text-based features.

# **Data Description**

**Restaurant DataSet**

There are 105 total observation with 6 different features.

Feature like collection and timing has null values.

There is no duplicate values i.e., 105 unique data.

Feature cost represent amount but has object data type because these values are separated by comma ','.

Timing represent operational hour but as it is represented in the form of text has object data type.

**Review DataSet**

There are total 10000 observation and 7 features.

Except picture and restaurant feature all others have null values.

There are total of 36 duplicate values for two restaurant - American Wild Wings and Arena Eleven, where all these duplicate values generally have null values.

Rating represent ordinal data, has object data type should be integer.

Timing represent the time when review was posted but show object data time, it should be converted into date time.

**Approach**

As the problem statement says, understanding what type of content is available in our dataset both review and meta dataset. we have to do clustering on similar content by matching text-based features. For that we used Affinity Propagation, Agglomerative Clustering, and K-means Clustering.

**Tools Used**

The whole project was done using python, in google Collaboratory. Followinglibraries were used for analyzing the data and visualizing it and to build the modelto predict the Netflix clustring

* Pandas: Extensively used to load and wrangle with the dataset.
* Matplotlib: Used for visualization.
* Seaborn: Used for visualization.
* Nltk: It is a toolkit build for working with NLP.
* Datetime: Used for analyzing the date variable.
* Warnings: For filtering and ignoring the warnings.
* NumPy: For some math operations in predictions.
* Wordcloud: Visual representation of text data.
* Sklearn: For the purpose of analysis and prediction.

**Steps Involved**

The following steps are involved in the project

**1.Handling missing values:**

We will need to replace blank countries with the mode (most common) country. It would be better to keepdirector because it can be fascinating to look at a specific filmmaker's movie. As aresult, we substitute the null values with the word 'unknown' for further analysis.

There are very few null entries in thedate\_added fieldsthus we delete them.

**2. Duplicate Values Treatment:**

Duplicate values dose not contribute anything to accuracy of results.

Our dataset dose not contains any duplicate values.

**3.NaturalLanguage Processing (NLP) Model:**

For the NLP portion of this project, I will first convert all plot descriptions to word vectors so they can be processed by the NLP model. Then, the similarity between all word vectors will be calculated using cosine similarity (measures the angle between two vectors, resulting in a score

between -1 and 1, corresponding to complete opposites or perfectly similar vectors). Finally, I will extract the 5 movies or TV shows with the most similar plot description to a given movie or TV show.

**4. Exploratory Data Analysis**:

Exploratory Data Analysis (EDA) as the name suggests, is used to analyze and investigate datasets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. It also helps to understand the relationship between the variables (if any) and it will be useful for feature engineering. It helps to understand data well before making any assumptions, to identify obvious errors, as well as better understand patterns within data, detect

outliers, anomalous events, find interesting relations among the variables.

After mounting our drive and fetching and reading the dataset given, we performed the Exploratory Data Analysis for it.

**5.Missing or Null value treatment:**

In datasets, missing values arise due to numerous reasons such as errors, or handling errors in data.

We checked for null values present in our data and the dataset contains a null value.

In order to handle the null values, some columns and some of the null values are dropped.

**6.Hypothesis from the data visualized:**

Hypothesis testing is done to confirm our observation about the population using sample data, within the desired error level. Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.

We have performed hypothesis testing to get the insights on duration of movies and content with respect to different variables.

**7. Tfidf vectorization:**

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is a very common algorithm to transform text into a meaningful representation of numbers which is used to fit a machine learning algorithm for prediction.

We have also utilized the PCA because it can help us improve performance at a very low cost of model accuracy. Other benefits of PCA include reduction of noise in the data, feature selection (to a certain extent), and the ability to produce independent, uncorrelated features of the data.

So, it's essential to transform our text into tfidf vectorizer, then convert it into an array so that we can fit into our model.

* **Finding number of clusters**

The goal is to separate groups with similar characteristics and assign them to clusters.

We used the Elbow method and the Silhouette score to do so, and we have determined that 28 clusters should be an optimal number of clusters.

* **Fitting into model**

In this task, we have implemented a K means clustering algorithm. K-means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k).

**8. Data Preprocessing:**

**Removing Punctuation**: Punctuations does not carry any meaning in clustering, so removing punctuations helps to get rid of unhelpful parts of the data, or noise.

**Removing stop-words**: Stop-words are basically a set of commonly used words in any language, not just in English. If we remove the words that are very commonly used in a given language, we can focus on the important words instead.

**Stemming:** Stemming is the process of removing a part of a word, or reducing a word to its stem or root. Applying stemming to reduce words to their basic form or stem, which may or may not be a legitimate word in the language.

**9. Clustering:**

Clustering (also called cluster analysis) is a task of grouping similar instances into clusters. More formally, clustering is the task of grouping the population of unlabeled data points into clusters in a way that data points in the same cluster are more similar to each other than to data points in other clusters. The clustering task is probably the most important in unsupervised learning, since it has many applications.

for example:

**• Data analysis:** often a huge dataset contains several large clusters, analyzing which separately, you can come to interesting insights.

**• Anomaly detection:** as we saw before, data points located in the regions of low density can be considered as anomalies

**• Semi-supervised learning:** clustering approaches often helps you to automatically label partially labeled data for classification tasks.

**• Indirectly clustering tasks (tasks where clustering helps to gain good results):** recommender systems,search engines, etc.

• **Directly clustering tasks**: customer segmentation, image segmentation, etc.

**Building a clustering model**

Clustering models allow you to categorize records into a certain number of clusters. This can help you identify natural groups in your data.

Clustering models focus on identifying groups of similar records and labeling the records according to the group to which they belong. This is done without the benefit of prior knowledge about the groups and their characteristics. In fact, you may not even know exactly how many groups to look for.

This is whatdistinguishes clustering models from the other machine-learning techniques—there is no predefined output or target field for the model to predict.

These models are often referred to as **unsupervised learning** models, since there is no external standard by which to judge the model's classification performance.

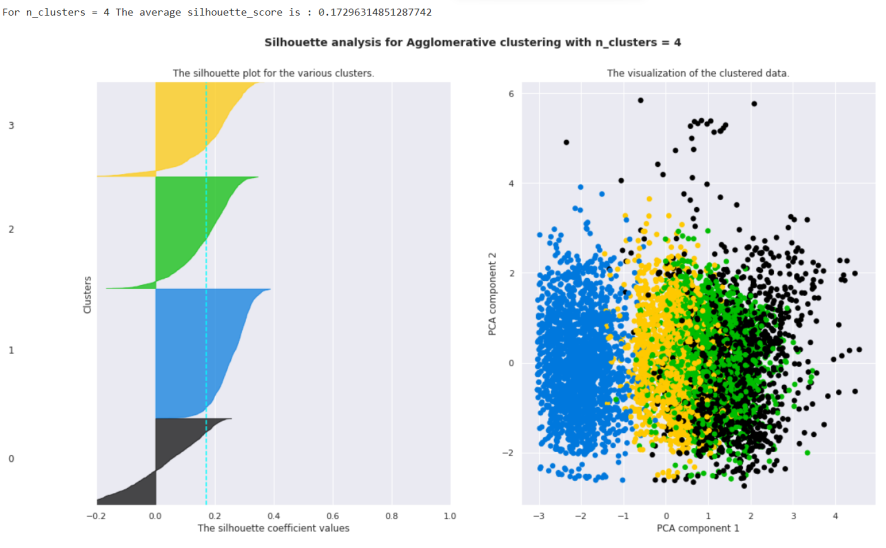
**10. Topic Modeling**  
**• Latent Dirichlet Allocation (LDA)**

LDA is a generative probabilistic model that assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.

# **11. Clusters Model Implementation**

1. **Agglomerative Clustering**
2. **K-means Clustering**
3. **Agglomerative Clustering**

The agglomerative clustering is the most common type of hierarchical clusteringused to group objects in clusters based on their similarity. ... Next, pairs ofclusters are successively merged until all clusters have been merged into one bigcluster containing all objects.

***Figure2.*Silhouette score and visualization**

**12.K-means Clustering**

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.

K-means algorithm works:

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then

performs iterative (repetitive) calculations to optimize the positions of the centroids. It halts creating and optimizing clusters when either:

• The centroids have stabilized — there is no change in their values because the clustering has been successful.

• The defined number of iterations has been achieved.

K-means algorithm is an iterative algorithm

that tries to partition the dataset into K pre-defined distinct non overlapping subgroupswhere each data point belongs to only onegroup.***Figure3.*Ideal clustering**

k-means clustering is a method of vector quantization, originally from signalprocessing, that aims to partition n observations into k clusters in which eachobservation belongs to the cluster with the nearest mean (cluster centers orcluster centroid), serving as a prototype of the cluster.

We created the sample data using build blobs and used rangen\_clusters to

specify the number of clusters we wanted to utilize in k means.

Silhouette score and visualization

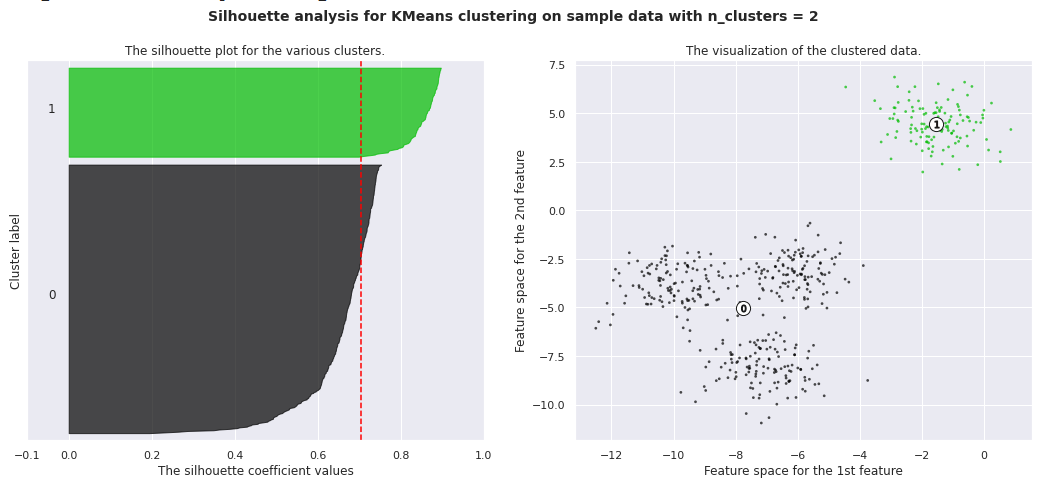
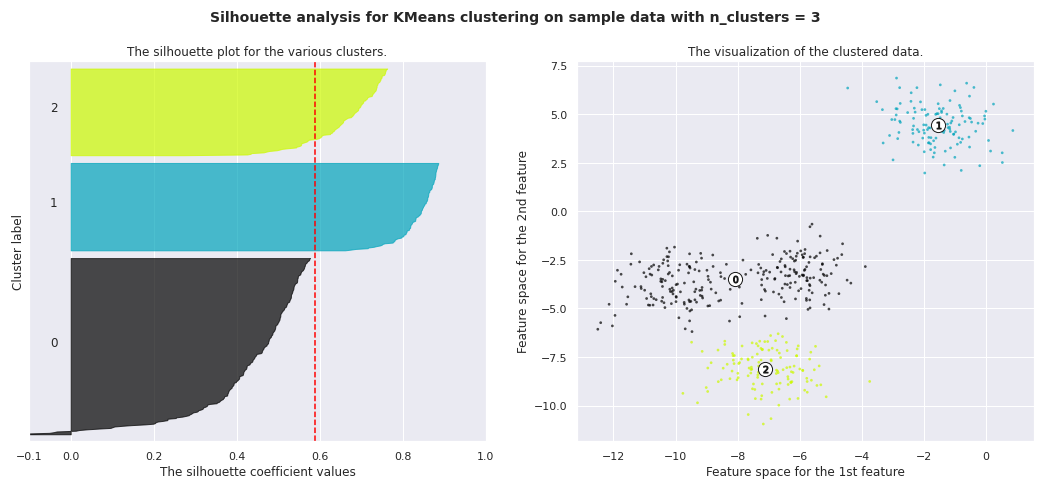
For clusters = 2 The average silhouette scoreis : 0.7049787496083262

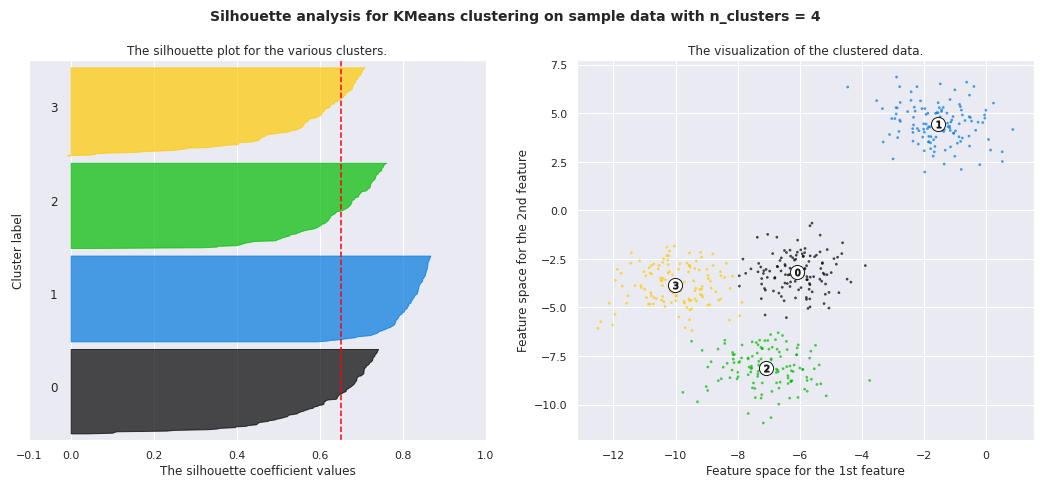
For clusters = 3 The average silhouette scoreis : 0.5882004012129721

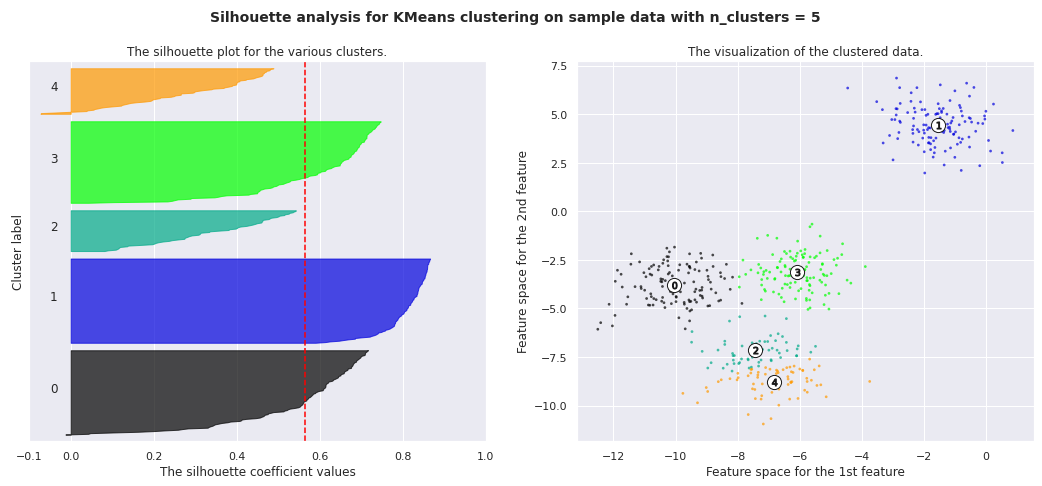
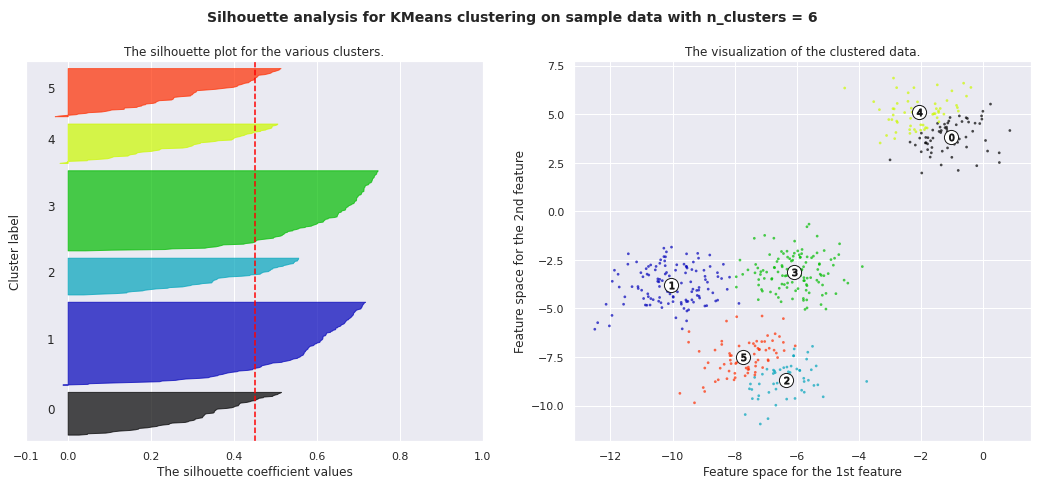
For clusters = 4 The average silhouette scoreis : 0.6505186632729437

For clusters = 5 The average silhouette scoreis : 0.56376469026194

For clusters = 6 The average silhouette scoreis : 0.4504666294372765







**13. Silhouette Coefficient or silhouette score(meaning)**

Silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1. 1: Means clusters are well apart from each other and clearly distinguished. ... a= average intra-cluster distance i.e., the average distance between each point within a cluster.

**1. Silhouette’s Coefficient**-

If the ground truth labels are not known, the evaluation must be performed utilizing the model itself. The Silhouette Coefficient is an example of such an evaluation, where a more increased Silhouette Coefficient score correlates to a model with better-defined clusters. The Silhouette Coefficient is determined for each sample and comprised of two scores

* 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.

The Silhouette Coefficient *s* for a single sample is then given as:

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:

**2. Elbow Curve:**

The Elbow Curve is one of the most popular methods to determine this optimal value of k.

The elbow curve uses the sum of squared distance (SSE) to choose an ideal value of k based on the distance between the data points and their assigned clusters.

**Conclusion**

**AB's - Absolute Barbecues, show maximum engagement and retention as it has maximum number of rating on average and Hotel Zara Hi-Fi show lowest engagement as has lowest average rating.**

**Price point for high rated hotel AB's= Absolute Barbecues is 1500 and price point for low rated restaurant Hotel Zara Hi-Fi is 400.**

**North Indian food followed by chinese are best or indeemand food as sold by most of the restaurants.**

**Great Buffets is the most frequently used tags and other tags like great, best, north, Hyderabad is also used in large quantity.**

**Satwinder singh is the most popular critic who has maximum number of follower and on an average he give 3.5 rating.**

**restaurant Collage - Hyatt Hyderabad Gachibowli is most expensive restaurant in the locality which has a price of 2800 for order and has 3.5 average rating. Hotels like Amul and Mohammedia Shawarma are least expensive with price of 150 and has 3.9 average rating.**

**REFERENCES:**

<https://colab.research.google.com/drive/11APbN2c9PfBm7rROZ9qmqdq6JJX0_Ye5?usp=sharing>

<https://machinelearningmastery.com/clustering-algorithms-with-python>

<https://towardsdatascience.com/introduction-to-machine-learning-algorith>