# **PROJECT TITLE : ZOMATO RESTAURANT CLUSTERING AND SENTIMENT ANALYSIS**

* **Lalit Ahirrao**
* **AniketGajmal**
* **Rushikesh Pawar**
* **Prasad Ghegade**
* **Samarth Gangurde**

Zomato is a fastest growing restaurant discovery website, established in 2008 by Deepinder Goyal and Pankaj Chaddah. Initially, it was named Foodiebay but in 2010, it was renamed as Zomato. It provides not only information related to nearby restaurants but also provides facilities such as online ordering, table reservations and management. Zomato currently serves 10,000 cities across 36 countries having 1.2 million popular restaurants with 80 million foodies every month. It is available in 10 different languages and has 10 million reviews with 18 million bookmarks. It gives a platform to the restaurant owners to serve a large number of users a good quality of food.

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 make some useful conclusions 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.

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.

### **Dataset-1: 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 w.r.t. Zomato categories
* Cuisines : Cuisines served by Restaurants
* Timings : Restaurant Timings

### **Dataset 1 :-information**

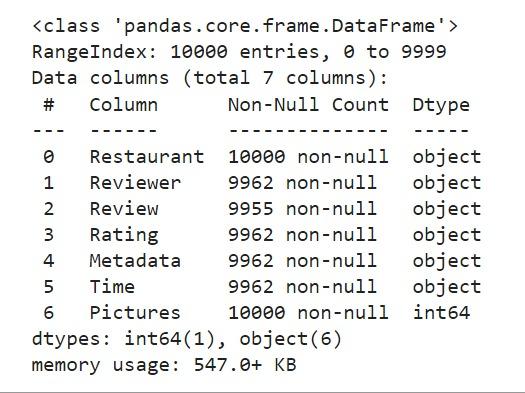
### 

### 

### 

### **Dataset-2: Zomato Restaurant reviews**

* 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



## **Preprocessing of dataset**

# In Preprocessing of dataset we improve the accuracy of our database. We remove any values that are wrong or missing as a consequence of human error or problems.

# Consistency should be improved. The accuracy of the results is harmed when there are data discrepancies or duplicates.

# Make the database as complete as possible. If necessary, we can fill up the missing properties.

# The data should be smooth. We make it easier to use and interpret this manne

### 

### 

### **Merging dataset**

Joins in Python is basically the merging of the two datasets with one or more common variables. Pandas provides various facilities for easily combining data frames together. We can use the simple merge () function for merging data in pandas.The different types of joins can be used to merge datasets, joints are left join, right join, inner join and outer join.

### **Left Join in Python**

Left Join returns all rows from the first dataset, even if there are no matches in the second dataset.

We’ve specified two data sets to be merged inside the pd.merge function as dataset 1 and dataset 2. The primary key to be used for merging is Name of Restaurants . The how argument is used to specify the type of join, in this case left

**DATA CLEANING:**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

❖ · Missing Values

❖ · Incorrect Format

❖ · Incorrect Headers

❖ · Anomalies/Outliers

**MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.

2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

**DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the isnull() and notnull() functions from the pandas library to determine null values

**EXPLORATORY DATA ANALYSIS:**

EDA means trying to understand the given data much better, so that we can make some sense out of it. Using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

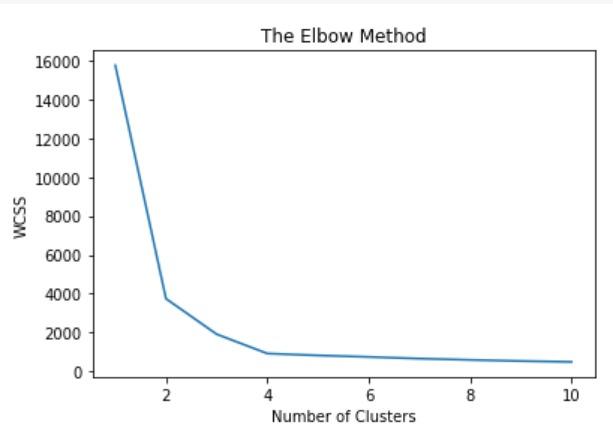
In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

**K-Means Clustering**

K Means algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

## **Elbow method**

* As we know we have to decide the value of k. But for deciding the value of k Elbow Method can help us to find the best value of k.
* It uses the sum of squared distance (SSE) between the data points and their respective assigned clusters, centroid or mean value. And We pick k value at where the point SSE starts to flatten out and form an elbow.
* This is how the method helps to find the good value of k (number of clusters for the dataset) and help in making the good clusters for the given dataset.



### **Silhouette method**

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 observations belonging to all the clusters:

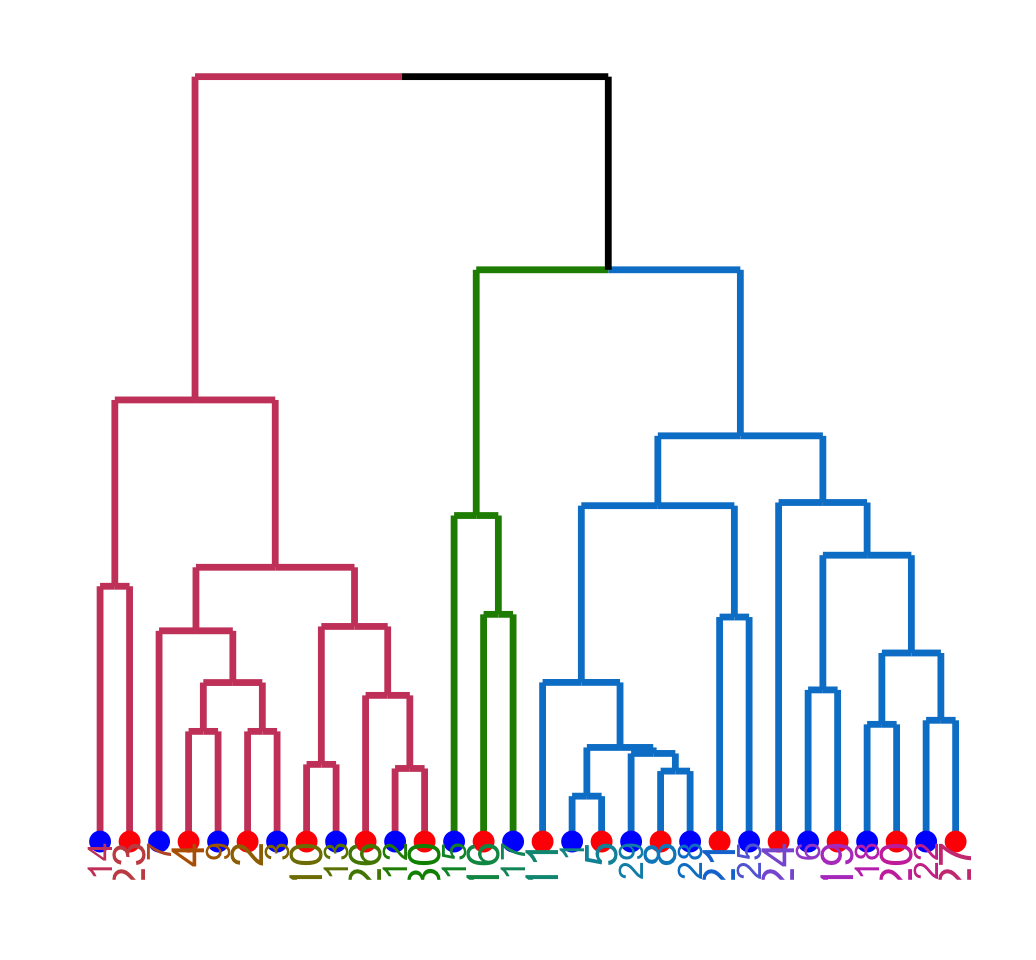
Mean distance between the observation and all other data points in the same cluster. This distance can also be called a mean intra-cluster distance. The mean distance is denoted by a. Mean distance between the observation and all other data points of the next nearest cluster. This distance can also be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

The Silhouette Coefficient for a sample is

**Silhouette score = (a − b) / m a x (a, b)**

### **Dendrogram diagram**

A Dendrogram is a diagram that represents the relationship between objects. The Dendrogram is used to display the distance between each pair of sequentially merged objects. These are commonly used in studying clusters before deciding the number of clusters significant to the dataset. The distance at which the two clusters combine is referred to as the dendrogram distance. The primary use of a dendrogram is to work out the best way to allocate objects to clusters. The key point to interpreting or implementing a dendrogram is to focus on the closest objects in the dataset. Note:-

* Distance between data points represents dissimilarities.
* Height of the blocks represents the distance between clusters.

## **TFIDF Vectorizer**

TF-IDF (term frequency-inverse document frequency) is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

It has many uses, most importantly in automated text analysis, and is very useful for scoring words in machine learning algorithms for Natural Language Processing (NLP).

TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as this, what, and if, rank low even though they may appear many times, since they don’t mean much to that document in particular.

# **Bag Of Words**

A bag-of-words model, or BOW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms.

The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents.

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

* A vocabulary of known words.
* A measure of the presence of known words.

It is called a “bag” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document

**REFERENCES:**

* [sklearn.feature\_extraction.text.TfidfVectorizer — scikit-learn 1.0.2 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)
* [Text Vectorization: Bag of Words (BoW) | by Vaibhav Jayaswal | Towards Data Science](https://towardsdatascience.com/text-vectorization-bag-of-words-bow-441d1bfce897#:~:text=Bag%20of%20words%20is%20a%20text%20vectorization%20technique,which%20can%20be%20overcome%20by%20using%20advanced%20techniques.)
* [Clustering Algorithms in Machine Learning that All Data Scientists Should Know (freecodecamp.org)](https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/)