**Zomato Restaurant Clustering and Sentiment Analysis**

**Jayanth V, Gowthaam Kumarasamy**

**Data Science Trainees,**

**AlmaBetter, Bangalore.**

**Abstract:**

This project contains two datasets, one of which contains information about Zomato restaurant names and Metadata and the other contains Zomato restaurant names, ratings, and reviews. This project contains Exploratory data analysis, text pre-processing, choosing appropriate features, clustering algorithms, model validation, dataset merging, sentiment analysis, model for sentiment analysis, and analysis of model performance. The clustering helps to group the restaurants based on cuisines which benefit the customers to choose between similar restaurants. Another point of view is to do sentiment analysis which helps the restaurant’s owner to know the sort of reputation their restaurant has earned. With KMeans Clustering and DBSCAN Clustering restaurants are clustered with the silhouette score of 0.195 and 0.107 respectively.

In this experiment by using Clustering, EDA, and Sentiment analysis, we are analyzing and visualizing the different perspectives. Which gives us different insights like

* Similar set of cuisines in various restaurants.
* Similar group of restaurants based on cuisines
* Most affordable and expensive restaurants
* Commonly used words by reviewers for expressing a positive or negative sentiment
* Restaurantst with a greater number of positive sentiments
* Do expensive restaurants always get higher ratings?
* Is there any relationship between ratings and review length?
* Distribution plot for polarity.

**Problem statement:**

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. The 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 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 an instant. The Analysis also solve some of the business cases that can directly help the customers find the best restaurant in their locality and for the company to grow up and work in the fields they are currently lagging in.

This could help in clustering the restaurants into segments. Also, the data has valuable information about cuisine and cost 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.

By using the given two datasets perform clustering and sentiment analysis to gain insights that would help improve the customer experience and solve some of the business cases.

**Features:**

## **Zomato Restaurant names and Metadata**

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 reviews**

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 the review.

**Introduction:**

Zomato is one of the most comprehensive and user-friendly apps where people can search for nearby restaurants and cafes, order food online, and get it delivered at their doorstep in no time. Moreover, you can also get accurate information about restaurants as it provides menus, reviews, and ratings. Based on that, users can place orders and enjoy lip-smacking food at their homes.

Zomato was founded by Deepinder Goyal and Pankaj Chaddah, two Delhi IIT graduates, in 2008. Till November 2010, Zomato was known as “Foodiebay.” Once they saw their colleagues who were seeking menus of different restaurants to order food. That’s when the idea took birth, and they thought of converting these manual menus into a digital format. In the year 2012, Zomato had spread its wings across the globe and started to list out the number of restaurants in the market.

Zomato’s Business Model is aimed at providing quality food services, information related to restaurants, their menus and user reviews. The Business Model of Zomato consists of providing food delivery services, information, user reviews and menus of partner restaurants. It has created a revolution in industries doing food business by including different restaurants and facilitating people to look for restaurants more conveniently.

The main work of Zomato is to suggest local and nearby restaurants to users and receive orders from them. Users can place orders from their favorite restaurant based on ratings and reviews shared by previous customers.

**Step 1:** From the desiccated app solution or website, users can explore various restaurants and order meals.

**Step 2:** Particular restaurant owners receive an order request and start preparing a meal.

**Step 3:** Once the food is ready to dispatch, it will be handed over to delivery providers.

**Step 4:** Delivery providers deliver the meal to the customer’s preferred location.

**Step 5:** From the given payment options, customers can make payments and share reviews based on their experience.

**Steps Involved:**

**Treating Null Values:**

Our dataset has missing values with many features which could lead us to loss of information and can mislead to wrong results so addressing Null values are important Null values are imputed with closest mean value as possible.

**Exploratory Data Analysis:**

Exploratory data analysis refers to the critical process of performing initial investigations on data so as to discover new patterns, to find anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

After loading the dataset, we performed EDA to get insights and also better understanding of the dataset. This process helped us figure out various aspects and relationships between the features. It gave us a better idea of which feature behaves in which manner compared to the other features.

**Building structured multi-plot grids and graphs**

When exploring multidimensional data, a useful approach is to draw multiple instances of the same plot on different subsets of your dataset. It allows a viewer to quickly extract a large amount of information about a complex dataset. Matplotlib offers good support for making figures with multiple axes; seaborn builds on top of this to directly link the structure of the plot to the structure of your dataset.

We have used count plot, box plot, distplot and pie chart in multi-plots for various features and the graphs provide a high level of convenience for comparison Multicollinearity for all the variables has been plotted and some highly correlated variables have been detected.

**Text Pre-processing:**

Natural Language Processing (NLP) is a branch of Data Science which deals with Text data. Apart from numerical data, Text data is available to a great extent which is used to analyze and solve business problems. But before using the data for analysis or prediction, processing the data is important.

Developers use it in almost all-natural language processing (NLP) pipelines, including voice recognition software, search engine lookup, and machine learning model training. It is an essential step because text data can vary.

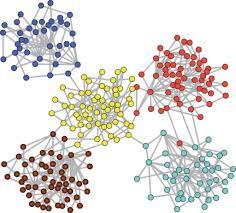
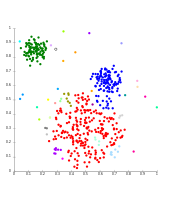
To prepare the text data for the model building we perform text preprocessing. It is the very first step of NLP projects. Some of the preprocessing steps are:

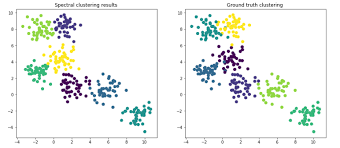
* Removing punctuations like. ,! $( ) \* % @
* Removing URLs
* Removing Stop words
* Lower casing
* Tokenization
* Stemming
* Lemmatization

**Clustering:**

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields

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.



**KMeans Clustering**

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

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.

**Working of K-Means Algorithm**

We can understand the working of K-Means clustering algorithm with the help of following steps −

**Step 1** − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

**Step 2** − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

**Step 3** − Now it will compute the cluster centroids.

**Step 4** − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more

**DBSCAN Clustering**

Clusters are dense regions in the data space, separated by regions of the lower density of points. The ***DBSCAN algorithm*** is based on this intuitive notion of “clusters” and “noise”. The key idea is that for each point of a cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Partitioning methods (K-means, PAM clustering) and hierarchical clustering work for finding spherical-shaped clusters or convex clusters. In other words, they are suitable only for compact and well-separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data.

Real life data may contain irregularities, like:

1. Clusters can be of arbitrary shape such as those shown in the figure below.
2. Data may contain noise.

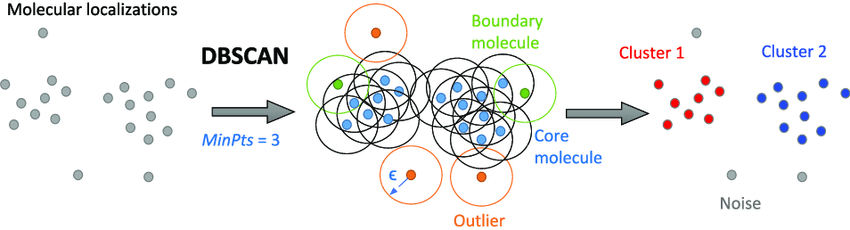
The following are the DBSCAN clustering algorithmic steps:

Step 1: Initially, the algorithms start by selecting a point (x) randomly from the data set and finding all the neighbor points within Eps from it. If the number of Eps-neighbours is greater than or equal to MinPoints, we consider x a core point. Then, with its Eps-neighbours, x forms the first cluster. After creating the first cluster, we examine all its member points and find their respective Eps -neighbors. If a member has at least MinPoints Eps-neighbors, we expand the initial cluster by adding those Eps-neighbours to the cluster. This continues until there are no more points to add to this cluster.

Step 2: For any other core point not assigned to cluster, create a new cluster.

Step 3: To the core point cluster, find and assign all points that are recursively connected to it.

Step 4: Iterate through all unattended points in the dataset and assign them to the nearest cluster at Eps distance from themselves. If a point does not fit any available clusters, locate it as a noise point.



**Vader Model**

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied directly to unlabeled text data.

VADER sentimental analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text. The steps involve:

* First, consider the text being analyzed. A model trained on paragraph-long reviews might not be effective. Make sure to use an appropriate model for the task at hand.
* Next, decide the type of analysis to perform. Some rudimentary sentiment analysis models go one step further, and consider two-word combinations, or *bigrams*. We will be going to work on complete sentences, and for this we’re going to import a trained NLTK lexicon called *VADER*.

**Text Blob**

TextBlob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more

**Libraries Used:**

**Pandas**

Pandas is mainly used for data analysis and associated manipulation of tabular data in Data Frames. Pandas allows importing data from various file formats such as comma-separated values, JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features. The development of pandas introduced into Python many comparable features of working with Data frames that were established in the R programming language.

**NumPy**

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It is open-source software. It contains various features including these important ones:

* A powerful N-dimensional array object
* Sophisticated (broadcasting) functions
* Tools for integrating C/C++ and Fortran code
* Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using NumPy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

**Matplotlib**

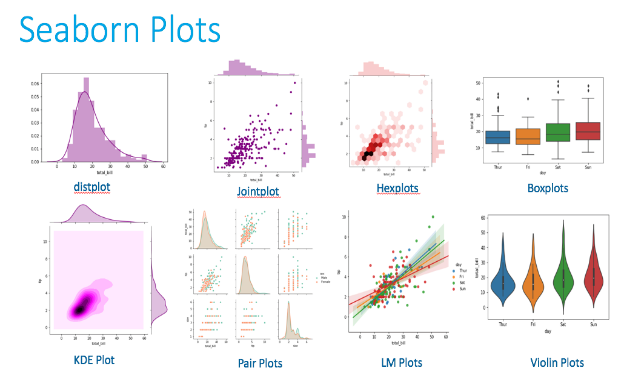
Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.



**Seaborn**

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.  
Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.



**Conclusion**

* **North Indian cuisine** is most common cuisine found in the restaurants.
* **Collage - Hyatt Hyderabad Gachibowli** is most expensive restaurant.
* **Amul and Mohammedia Shawarma** are the most affordable restaurants.
* The Restaurants are clustered on cuisines into **15** clusters by using KMeans clustering algorithm with the Silhouette score of **0.195**.
* DBSCAN algorithm is also used to cluster the restaurants into **15** clusters and also helps us to detect the outliers with the Silhouette score of **0.107**.
* **Anvesh Chowdary** has given the greatest number of reviews.
* **AB’s – Absolute Barbecues** is the top-rated restaurant.
* Almost **79 percent** of the observations have Positive sentiment and **14** and **7** percent of the observations have Neutral and Negative sentiments respectively
* **Good** is the most common word in the Highly positive sentiment.
* **Worst** is the most common word in the Highly negative sentiment.
* **AB’s – Absolute Barbecues, The Indi grill and B-Dubs** are the restaurants with the greatest number of positive reviews.
* **Arena Eleven and Banana leaf** Multicuisine restaurant are the restaurants with the greatest number of negative reviews.