Zomato Restaurant Clustering and Sentiment Analysis

Lakshmi Keerthana

Data Science Trainee, AlmaBetter

Abstract

Zomato is one of the successful on-demand food delivery platforms that helps users discover food places and get it delivered at their doorstep. We can also get accurate information about restaurants as it provides menus, reviews, and ratings. Based on that, users can place orders and enjoy food at their homes.In this project we utilize data analytics techniques to gain a deeper understanding of the restaurants and customer feedback on the popular online food delivery platform, Zomato.

The data contains the information about the restaurant's name, location, cuisines, average cost for two, ratings, and user reviews. I did the task of data cleaning and preprocessing. Handling missing values and transforming the data into a form that can be used for analysis and visualization.

Then comes the implementation of clustering on the restaurant data using K means Clustering and Agglomerative Clustering. The objective of this clustering is about grouping the similar restaurants together. The features that were used to cluster the data includes, restaurants, cuisines, Average cost of each restaurant and the rating.

To get the optimum number of clusters, elbow method and silhouette scores were used. Later conducted sentiment analysis on the customers reviews to understand the sentiment towards the restaurants.

The information gained from this analysis can be very useful for the restaurants and customers to make better decisions.

Data Fields

Restaurant Data

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

Review Data

1.Reviewer: Name of the reviewer

2.Review: Review text

3.Rating: Rating provided

4.MetaData: Reviewer metadata-No of reviews and followers

5.Time: Date and Time of Review

6.Pictures: No of pictures posted with reviews

Introduction

This is our Unsupervised Machine Learning Capstone Project, hence we will be looking into multiple unsupervised models and some sentiment analysis. We are only focussing on all that algorithm which has been taught to us till now in our class. , K Means Clustering, Agglomerative Clustering and Sentiment Analysis have been implemented in this capstone project.

ML pipeline to be followed

- Dataset Inspection
- Data Cleaning
- Exploratory Data Analysis
- Data Wrangling
- Data Preprocessing
- Handling Outliers
- Categorical Encoding
- Textual Data Preprocessing
- Feature Selection
- Model Implementation
- Sentiment Analysis

Dataset Inspection

We loaded two datasets. One is a restaurant dataset and another is a review dataset using the given csv files. We checked the general information about data.

Restaurant DataSet

- There are 105 total observations with 6 different features.
- Features like collection and timing have null values.
- There are no duplicate values i.e., 105 unique data.
- Feature cost represents amount but has object data type because these values are separated by comma ','.
- Timing represents operational hours but as it is represented in the form of text has object data type.

Review DataSet

- There are a total of 10000 observations and 7 features.
- Except pictures and restaurant features all others have null values.
- Rating represents ordinal data, and the object data type should be integer.
- Timing represents the time when review was posted but shows object data time, it should be converted into date time.

Data Cleaning

Data cleaning is one of the important parts. We remove the unwanted observations, fix the structural errors, manage the unwanted outliers and handle the missing data.

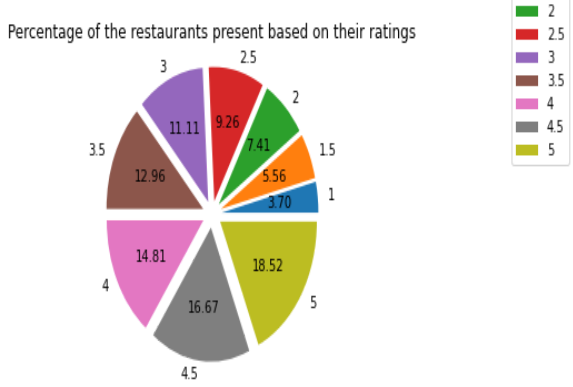
We first checked the datatypes of all the features, checked for null values. Handled missing values for Rating column and converted the datatype of Rating to float. Converted the datatype of the Time column to datetime.

Ohe most important steps as part of data preprocessing is detecting and treating the outliers as they can negatively affect the analysis and the training process of a machine learning algorithm resulting in lower accuracy.

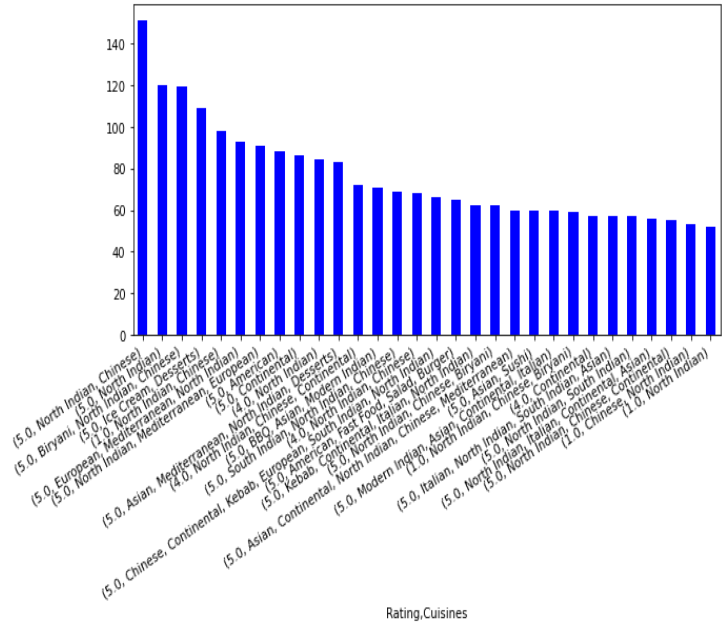
An outlier may occur due to the variability in the data, or due to experimental error/human error.Box plots are a visual method to identify outliers. It is a data visualization plotting function. It shows the min, max, median, first quartile, and third quartile.

Exploratory Data Analysis:

Restaurant Analysis: Percentage of restaurants present based on their ratings visualization through a pie chart. Engagement and retention for any business is very much important as profit and scalability for any business depend upon retention of customers. Maximum retention means people prefer to use the same brand over others.Some restaurant show less rating which can show negative growth if not monitored why they recieve less order.



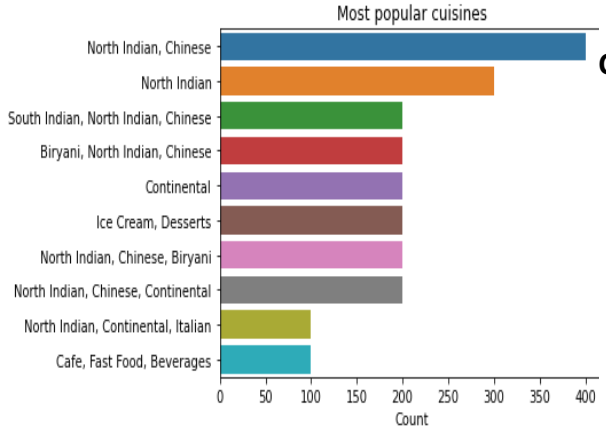
Cuisines:



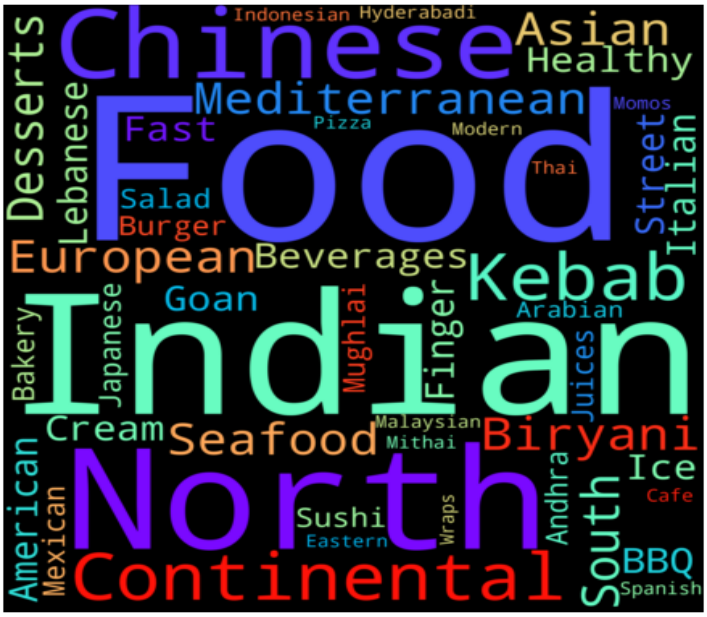
Identifying the Commoditized Cuisine plays an important role as it helps in identifying the challenge or Competitive Advantage i.e., Knowing which cuisines are commoditized allows a restaurant or food business to differentiate themselves from their competitors by offering unique and non-commoditized options.

If a cuisine is commoditized, the prices for ingredients and labor for that cuisine may be higher than for non-commoditized cuisines.

Most Popular Cuisines:



Cuisines through wordcloud:



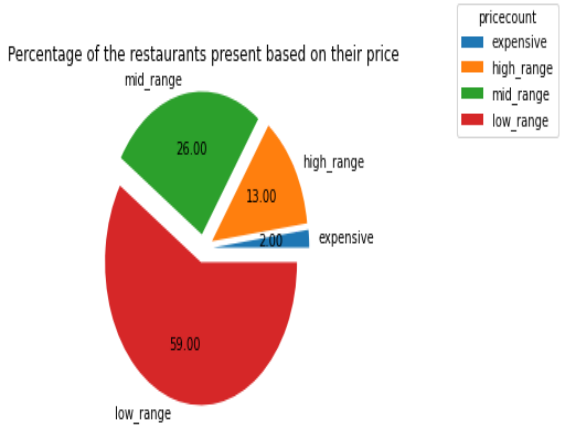
Cuisines can provide insight into consumer preferences, which can be used to make informed decisions about menu offerings, pricing, and promotions.

Identifying these commoditized cuisines can help a business to control costs by focusing on non-commoditized options or finding ways to lower the cost of commoditized items.

A word cloud of tags used to describe food can help Zomato identify the most frequently mentioned food attributes in customer reviews. This can provide insight into which attributes are most popular and well-regarded among customers, and which attributes may need improvement.

based on these hotels.

Cost Analysis:

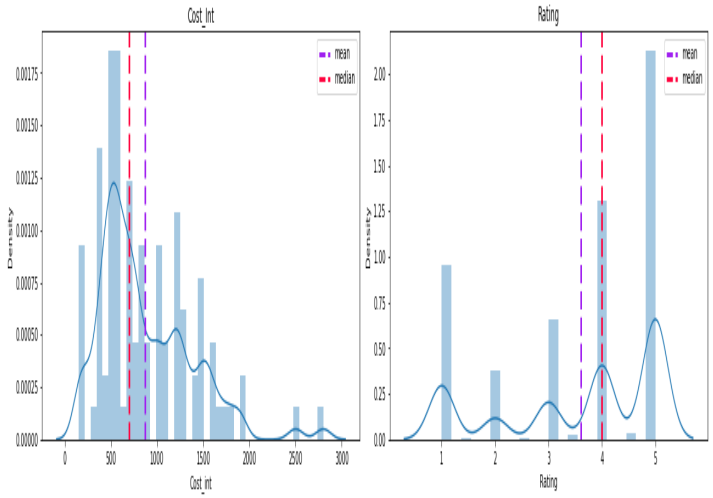


Visualizing the percentage of restaurants present based on their price. Most expensive product are always center of attraction for a niche market (subset of the market on which a specific product is focused) at the same time for a business purpose, this product are prefered to be most revenue generating market.

Definitely for food delivery platform Zomato, it is very important to focus and improve sales

Cost and Rating:

Visualizing Cost and Rating through distribution plots



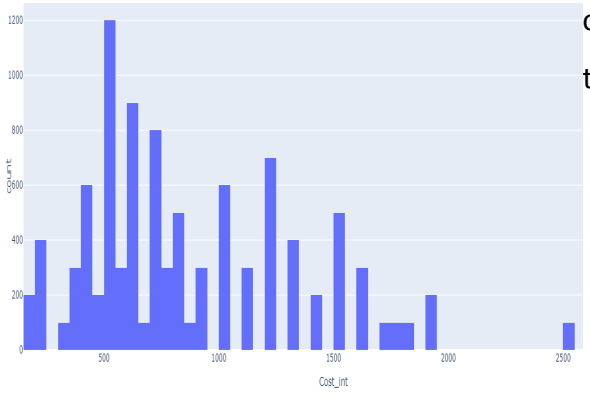
Distplot is helpful in understanding the distribution of the feature.Price always place important role in any business alongwith rating which show how much engagement are made for the product.But in this chart it is unable to figure any impact on business when plotted all alone.

Here most liked restaurant has a price point of 1500 which is even though a little high than average but as this business is all about food quality and taste it show maximum engagement which means it serve best quality of food, however deep dive on analysing review text can exactly give why this price point is prefered most.

Some restaurant with lowest rating even with low price point is not making engagement, this may create a negative impact on business.

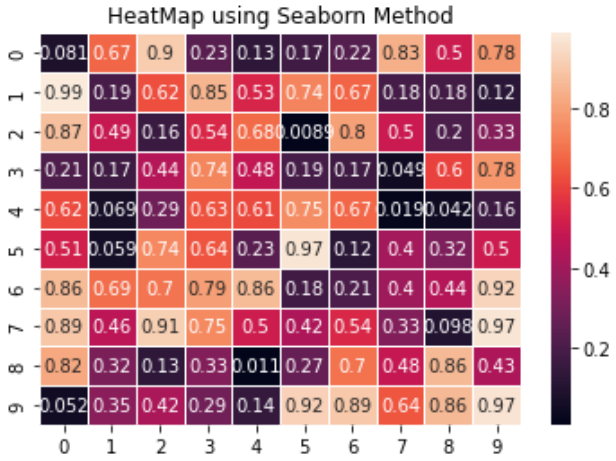
However it can not be finalized that this hotel should unlisted as there may be chance of different cuisine they both serve and it also depend upon the locality they both serve, therefore based on that small promotional offers can also be given for low rated restaurant to increase sales.

Cost and Restaurants:



Since it is customer centered business i.e., direct to consumer it is important to understand price point which makes this business more affordable for evryone, therefore it is important for business to crack the price point.

Heatmap:



A correlation matrix is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. A correlation matrix is used to summarize data, as an input into a more advanced analysis, and as a diagnostic for advanced analyses. The range of correlation is [-1,1].

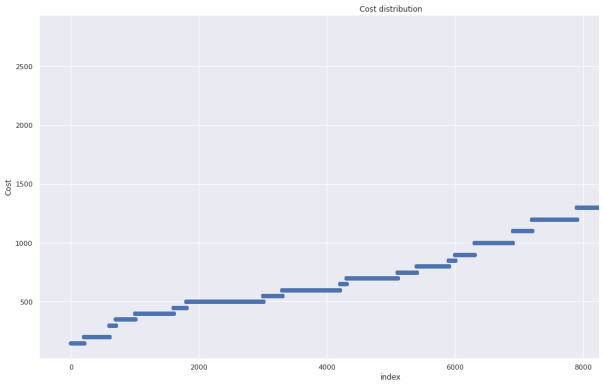
Thus to know the correlation between all the variables along with the correlation coeficients, i used correlation heatmap.

From the above correlation heatmap, it can be depicted that few features are correlated, like reviewer total review is related to reviewer follower and again reviewer total review is related to pictures.

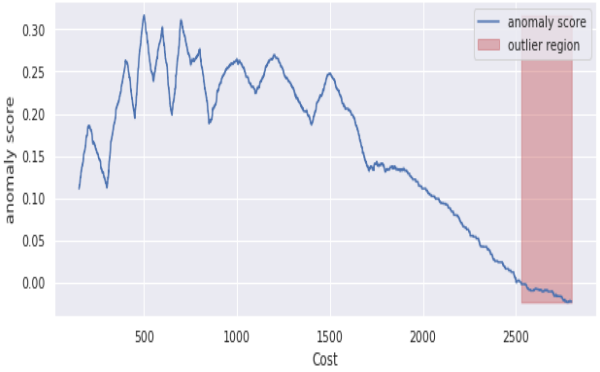
Rest all correlation can be depicted from the above chart.

Handling Outliers:

Cost distribution:



Outlier Region:



Since cost and reviewer follower feature or column show positive skewed distribution and using isolation forest found they have outliers, hence using the capping technique instead of removing the outliers, capped outliers with the highest and lowest limit using IQR method.

Clustering Data Preprocessing:

Data Preprocessing or Data Preparation is a data mining technique that transforms raw data into an understandable format for ML algorithms.

K-means input data requirements:

- **Numerical variables only.** K-means uses distance-based measurements to determine the similarity between data points. If you have categorical data, use K-modes clustering, if data is mixed, use K-prototype clustering.
- **Data has no noises or outliers.** K-means is very sensitive to outliers and noisy data.
- **Data has a symmetric distribution of variables (it isn't skewed).** Real data always has outliers and noise, and it's difficult to get rid of it. Transformation data to normal distribution helps to reduce the impact of these issues. In this way, it's much easier for the algorithm to identify clusters.
- **Variables on the same scale** — have the same mean and variance, usually in a range -1.0 to 1.0 (standardized data) or 0.0 to 1.0 (normalized data). For the ML algorithm to consider all attributes as equal, they must all have the same scale.
- **There is no collinearity** (a high level of correlation between two variables). Correlated variables are not useful for ML segmentation algorithms because they represent the same characteristic of a segment. So correlated variables are nothing but noise.
- **Few numbers of dimensions.** As the number of dimensions (variables) increases, a distance-based similarity measure converges to a constant value between any given examples. The more variables the more difficult to find strict differences between instances.

In our project, we have binned all the cuisines into their respective cuisine categories and created a cuisine category list.

Merging the cuisine category dataframe in our cluster data frame.

We have used the get_dummies method in order to generate binary values for our cuisines.

We have created a final data frame that was used for the clustering process.

Standard Scaler:

We have used Standard Scaler in order to scale the data. StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature.

Textual Data Preprocessing:

Text data derived from natural language is unstructured and noisy. Text preprocessing involves transforming text into a clean and consistent format that can then be fed into a model for further analysis and learning.

Text preprocessing techniques may be general so that they are applicable to many types of applications.

Natural Language Processing system for textual data reads, processes, analyses and interprets data. NLP pipeline include steps such as sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal and spelling correction.

- Segmentation involves breaking up text into sentences.
- Tokenization stage involves converting sentences into a stream of words also called tokens. Tokens are basic building blocks upon which analysis and other methods are built.
- Spell Correction includes correcting the spelling of all words in the text.
- Stop Words are frequently occurring words used to construct sentences. The words that are redundant are removed at the preprocessing stage.
- Stemming is the process of converting all words to their base form or stem. Stemming is also used at the preprocessing stage for applications such as emotion identification and text classification.
- Lemmatization is a more advanced form of stemming and involves converting all words to their root form. It requires more processing power and time than a stemmer to generate output.
- 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:

1. A vocabulary of known words.
2. A measure of the presence of known words.

Text preprocessing improves the performance of an NLP system. For tasks such as sentiment analysis, document categorization, document retrieval based upon user queries and adding a text preprocessing layer provides more accuracy.

Stages such as stemming, lemmatization and text normalization make the vocabulary size more manageable and transform the text into more standard form across a variety of documents acquired from different sources.

Feature Manipulation and Feature Selection:

It is necessary to provide a pre-processed and good input dataset in order to get better outcomes. We collect a huge amount of data to train our model and help it to learn better. Generally, the dataset consists of noisy data, irrelevant data, and some part of useful data.

It is very necessary to remove noises and less-important data from the dataset and to do this, and Feature selection techniques are used.

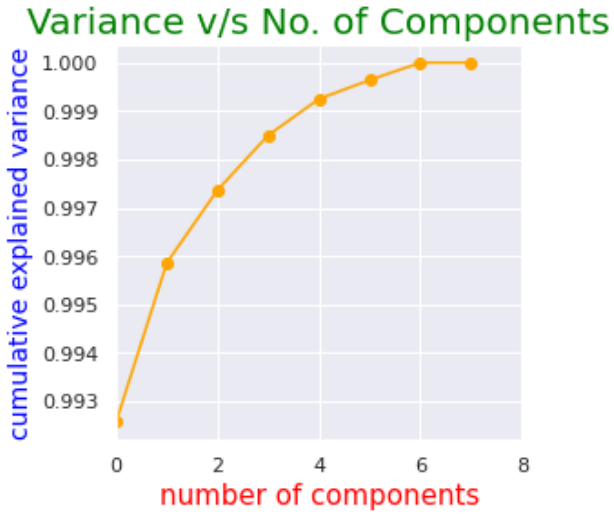
- It helps in avoiding the curse of dimensionality.
- It helps in the simplification of the model so that it can be easily interpreted by the researchers.
- It reduces the training time.
- It reduces overfitting hence enhances the generalization.

Principal Component Analysis:

Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

The steps in PCA involve getting the dataset, representing data into a structure, Standardizing the data and calculating the new features Or Principal Components.

The new features extracted from PCA are then used in the algorithms.



original shape: (100, 8)

transformed shape: (100, 3)

Machine Learning Model Implementation:

KMeans Clustering:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. K-Means Clustering groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

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.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Implementation of K-means:

Let's see the steps on how the K-means machine learning algorithm works using the Python programming language.We'll use the Scikit-learn library and data to illustrate K-means clustering.

We'll import the following libraries in our project:Pandas for reading and writing spreadsheets.Numpy for carrying out efficient computations.Matplotlib for visualization of data

K-Means Clustering Elbow Method:

It is the simplest and commonly used iterative type unsupervised learning algorithm. In this, we randomly initialize the K number of centroids in the data (the number of k is found using the Elbow method which will be discussed later in this article) and iterate these centroids until no change happens to the position of the centroid.

1) Select the number of clusters for the dataset (K)

2) Select K number of centroids

3) By calculating the Euclidean distance or Manhattan distance assign the points to the nearest centroid, thus creating K groups

4) Now find the original centroid in each group

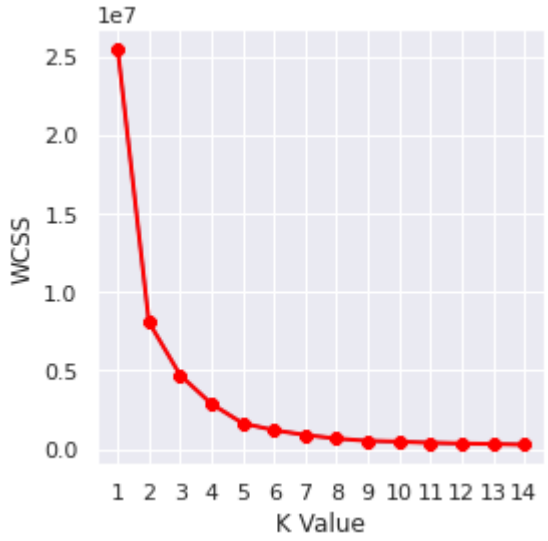
5) Again reassign the whole data point based on this new centroid, then repeat step 4 until the position of the centroid doesn't change.

Finding the optimal number of clusters is an important part of this algorithm. A commonly used method for finding optimal K value is the Elbow Method.

In the Elbow method, we are varying the number of clusters. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster. When we plot the WCSS with the K value, the plot looks like an Elbow. As the number of clusters increases, the WCSS value will start to decrease.

WCSS value is largest when K = 1. When we analyze the graph we can see that the graph will rapidly change at a point and thus creating an elbow shape. From this point, the graph starts

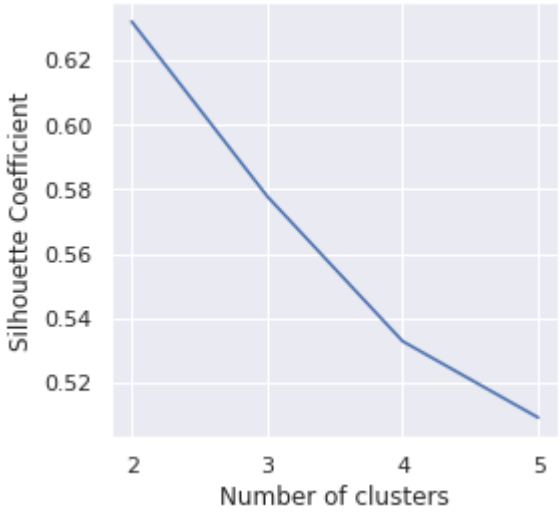
to move almost parallel to the X-axis. The K value corresponding to this point is the optimal K value or an optimal number of clusters.



Silhouette Method:

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters by providing a succinct graphical representation of how well each object has been classified.

Similar to the previous Elbow method, we pick a range of candidate values of k (number of clusters), then train K-Means clustering for each of the values of k. For each k-Means clustering model represents the silhouette coefficients in a plot and observe the fluctuations and outliers of each cluster.



The silhouette analysis is used to choose an optimal value for n_clusters. The silhouette plot shows that the n_clusters values of 5 was a good pick for the given data. Silhouette analysis is more ambivalent in deciding between 2 and 4.

Visualizing the clusters:

The last step is to visualize the clusters. As we have 5 clusters for our model, we will visualize each cluster one by one.

The output image is clearly showing the five different clusters with different colors.

Agglomerative Clustering:

Agglomerative Clustering is a type of hierarchical clustering algorithm. It is an unsupervised machine learning technique that divides the population into several clusters such that data points in the same cluster are more similar and data points in different clusters are dissimilar.

Points in the same cluster are closer to each other.

Points in the different clusters are far apart.

The intuition behind Agglomerative Clustering:

Agglomerative Clustering is a bottom-up approach, initially, each data point is a cluster of its own, further pairs of clusters are merged as one moves up the hierarchy.

To obtain the desired number of clusters, the number of clusters needs to be reduced from initially being n clusters (n equals the total number of data-points). Two clusters are combined by computing the similarity between them.

There are some methods which are used to calculate the similarity between two clusters:

- Distance between two closest points in two clusters.
- Distance between two farthest points in two clusters.
- The average distance between all points in the two clusters.

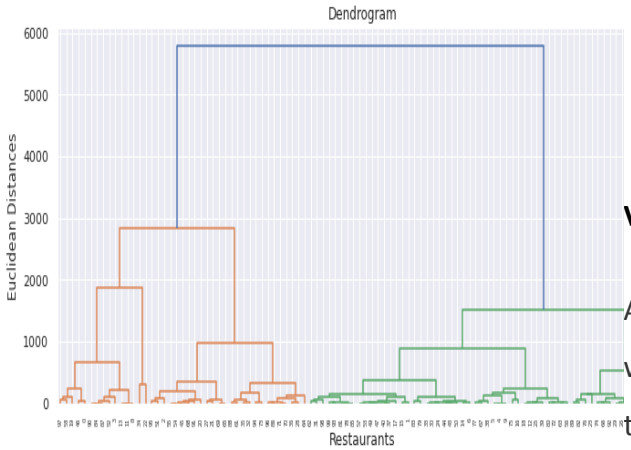
- Distance between centroids of two clusters.

There are several pros and cons of choosing any of the above similarity metrics.

To find the optimal number of clusters by observing the dendrograms:

From the above dendrogram plot, find a horizontal rectangle with max-height that does not cross any horizontal vertical dendrogram line.

The portion in the dendrogram in which a rectangle having the max-height can be cut, and the optimal number of clusters will be 3 as observed in the right part of the above image. Max height rectangle is chosen because it represents the maximum Euclidean distance between the optimal number of clusters.



Training the hierarchical clustering model:

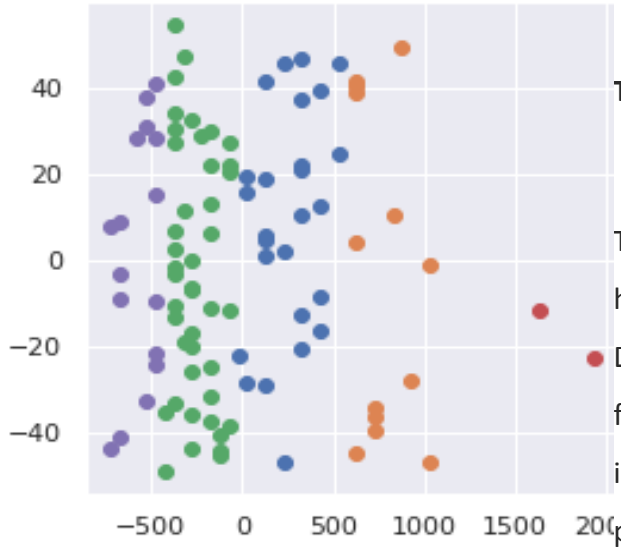
After that, we have imported the **AgglomerativeClustering** class from the cluster module of scikit learn library. The AgglomerativeClustering class takes the following parameters:

- **n_clusters=5**: It defines the number of clusters, and we have taken here 5 because it is the optimal number of clusters.
- **affinity='euclidean'**: It is a metric used to compute the linkage.
- **linkage='ward'**: It defines the linkage criteria, here we have used the "ward" linkage. This method is the popular linkage method that we have already used for creating the Dendrogram. It reduces the variance in each cluster.

Visualizing the clusters:

As we have trained our model successfully, now we can visualize the clusters corresponding to the dataset.

Here we will use the same lines of code as we did in k-means clustering, except one change. Here we will not plot the centroid that we did in k-means, because here we have used dendrograms to determine the optimal number of clusters.



Sentiment Analysis:

Text Analytics is the process of converting unstructured text data into meaningful insights to measure customer opinion, product reviews, sentiment analysis, and customer feedback. Our domain of expertise in this solution is driven towards sentiment analysis.

A sentiment analysis system for text analysis combines natural language processing and machine learning techniques to assign weighted sentiment scores to the entities, topics, themes and categories within a sentence or phrase.

Topic Modeling with Gensim:

Topic Modeling is a technique to extract the hidden topics from large volumes of text. Latent Dirichlet Allocation(LDA) is a popular algorithm for topic modeling with excellent implementations in the Python's Gensim package. The challenge, however, is how to extract good quality topics that are clear, segregated and meaningful. This depends heavily on the quality of text preprocessing and the strategy of finding the optimal number of topics. This tutorial attempts to tackle both of these problems.

We needed stopwords from NLTK and spacy's en model for text preprocessing.The core packages used in this tutorial are re, gensim, spacy and pyLDAvis. Besides this matplotlib, numpy and pandas for data handling and visualization were used.

LDA's approach to topic modeling is that it considers each document as a collection of topics in a certain proportion. And each topic is a collection of keywords, again, in a certain proportion.

The following are key factors to obtaining good segregation topics:

1. The quality of text processing.
2. The variety of topics the text talks about.
3. The choice of topic modeling algorithm.
4. The number of topics fed to the algorithm.
5. The algorithm's tuning parameters.

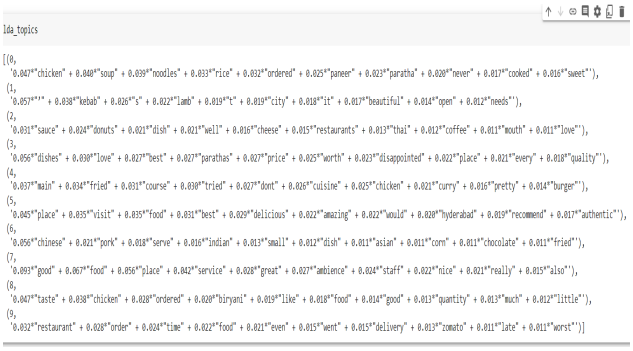
Gensim's Phrases model can build and implement the bigrams, trigrams, quad grams and more. The two important arguments to Phrases are min_count and threshold. The higher the values of these params, the harder it is for words to be combined into bigrams.

We have everything required to train the LDA model. In addition to the corpus and dictionary, you need to provide the number of topics as well.

Apart from that, alpha and eta are hyperparameters that affect sparsity of the topics. According to the Gensim docs, both default to 1.0/num_topics prior.

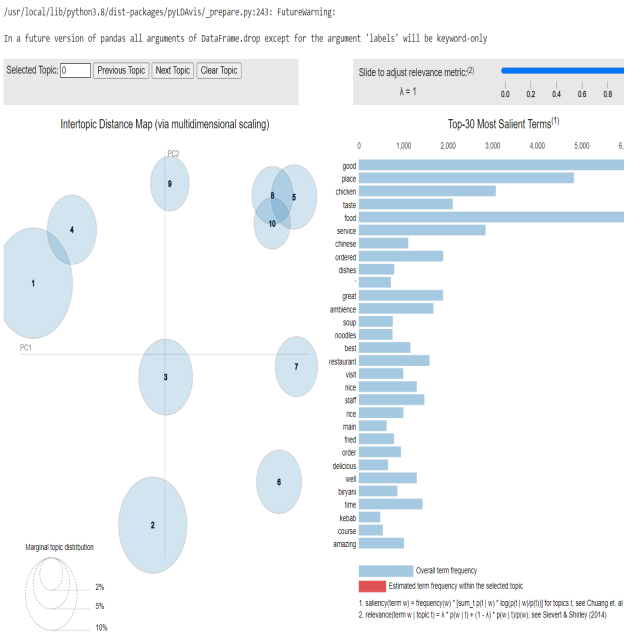
chunksize is the number of documents to be used in each training chunk. update_every determines how often the model parameters should be updated and passes is the total number of training passes.

You can see the keywords for each topic and the weightage(importance) of each keyword using lda_model.print_topics()



Now that the LDA model is built, the next step is to examine the produced topics and the associated keywords. There is no better tool

than pyLDAvis package's interactive chart and is update. These words are the salient keywords designed to work well with jupyter notebooks. that form the selected topic.



Each bubble on the left-hand side plot represents a topic. The larger the bubble, the more prevalent that topic is.

A good topic model will have fairly big, non-overlapping bubbles scattered throughout the chart instead of being clustered in one quadrant.

A model with too many topics, will typically have many overlaps, small sized bubbles clustered in one region of the chart.

If you move the cursor over one of the bubbles, the words and bars on the right-hand side will

Conclusion:

1. In this unsupervised machine learning project, we looked into multiple unsupervised models and sentiment analysis. We started with loading the data, inspecting the data and Data Cleaning

2. Through Exploratory Data Analysis, we visualized the data through bar plot, pie chart, wordcloud and distplots.

3. We have seen a correlation matrix is a table showing correlation coefficients between variables through heatmap.

4. We have the capped outliers with the highest and lowest limit using IQR method.

5. During data pre-processing, we have transformed raw data into an understandable format for ML algorithms.

6. For textual data pre-processing, NLP pipeline include steps such as sentence segmentation, word tokenization, lowercasing, stemming or lemmatization, stop word removal and spelling correction.

7. Principal Component Analysis was used to reduce the dimensionality of large datasets.

7. Models like K-means was used for grouping the data into clusters and visualizing them in clusters. Elbow method and Silhouette score was used to find the optimum number of clusters.

8. Agglomerative Clustering is a hierarchical clustering algorithm that was used to cluster the data points. Dendrogram plot was used to find the optimal number of clusters.

9. For Sentiment Analysis, topic modeling was a technique to extract the hidden topics from large volumes of text. Latent Dirichlet Allocation(LDA) is the algorithm that was used for topic modeling and extracted good quality topics that are clear, segregated and meaningful.