

**Unsupervised learning Project on**

**Zomato clustering and sentimental analysis**

**Project performed by**

**Sunil Kumar**



**Table of Contents**

* Summary............................................................................1)
* Problem statement..............................................................2)
* Introduction.........................................................................3)
* Technical Work...................................................................4)

1. Library used.
2. Function and Method used.

* Step involved in exploratory data analysis..........................5)
* Procedure performed and Observations.............................6)

**Summary:**

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. 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.

**Problem Statement**

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 conclusion in the form of Visualizations. Also, cluster the Zomato restaurants into different segments. The data is visualized as it becomes easy to analyse data at instant. The Analysis also solve 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.

.

**Technical Work**

* ***Library used***
* ***Functions and methods used.***

**Library used**

* **Pandas** This library is used in data analysis and manipulation and importing files.
* **NumPy** which stands for Numerical Python is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed. NumPy is a Python package. It stands for 'Numerical Python'.
* **Matplotlib** Used infor data visualization, graphs & plotting.
* **Seaborn** Used inmaking data visualization in a more colorful and meaningful way.
* **Sklearn:** useful and robust library for machine learning in Python.
* **os** Used in provides functions for creating and removing a directory folder.
* **Spicy -:-**Spicy is an open-source software python library used in advanced natural language processing and machine learning. It will be used to build information extraction, natural language understanding systems, and to pre-process text for deep learning.
* **Contractions-:** Contraction is the shortened form of a word like don’t stands for do not, aren’t stands for are not. Like this, we need to expand this contraction in the text data for better analysis. We can easily get the dictionary of contractions on Google or create your own and use the re module to map the contractions
* **Sys :-** The sys module in Python provides various functions and variables that are used to manipulate different parts of the Python runtime environment. It allows operating on the interpreter as it provides access to the variables and functions that interact strongly with the interpreter

**Common functions and methods used**

**pd.read\_csv ():** used in to import data in CSV format. This function has a number of arguments, but the only essential argument is file, which specifies the location and filename.

**%cd :** magical command finds present working directory.

**os.chdir():** Changes your directory to your data directory.

**head ():** Used in checking the first rows of the dataset.

**tail ():** Used in fetching the last rows of the dataset.

**shape ():** Used in getting the total no. Of columns and rows, a dataset contains.

**Type ():** Used in knowing the data structure of the dataset.

**df.dtypes:** Used in understanding the data types of the dataset.

**IsNull():** used in checking if any null values are available.

**Sum():** used in giving sum of the values.

**value\_counts() :** Used in providing counts of particular values present in data.

**sort\_values() :**function sorts a data frame in Ascending or Descending order of passed Column. It's different than the sorted Python function since it cannot sort a data frame and particular column cannot be selected.

**describe()** : This method is used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

.**corr():** aggregate function returns a coefficient of correlation between two numbers.

**unique():** function is used to find the unique elements of an array.

**re():**A regular expression (or RE) specifies a set of strings that matches it; the functions in this module let you check if a particular string matches a given regular expression (or if a given regular expression matches a particular string, which comes down to the same thing)

**GroupBy:** It allows you to split your data into separate groups to perform computations for better analysis

**Sum():** The Python sum() function calculates the total of all numerical values in an iterable.

**Mean():**It returns mean of the data set passed as parameters. Arithmetic mean is the sum of data divided by the number of data-points. It is a measure of the central location of data in a set of values which vary in range.

**Info():** The info() method prints information about the DataFrame. The information contains the number of columns, column labels, column data types, memory usage, range index, and the number of cells in each column (non-null values).

**astype():** The astype() function is used to cast a pandas object to a specified data type.

.

**Steps involved:**

* *Setting directory/path.*
* *Loading dataset.*
* *Text pre-processing and text cleaning*
* *Visualizations.*
* *Modeling.*
* *Performance improvement.*
* *Model selection and evaluation.*
* *Deployment.*

**Setting directory/path**

Before we proceed with python to understand the data, it is important to let python understand first that where our data available is so that, we can perform the operation on data using python. The setting directory is similar to that and helps in understanding the presence of the data.

**os.chdir(your path)**

**Loading dataset**

Once the directory is set now we need to load the dataset which is commonly done using:

**Pd.read\_csv(‘file\_name’)**

**Text pre-processing**

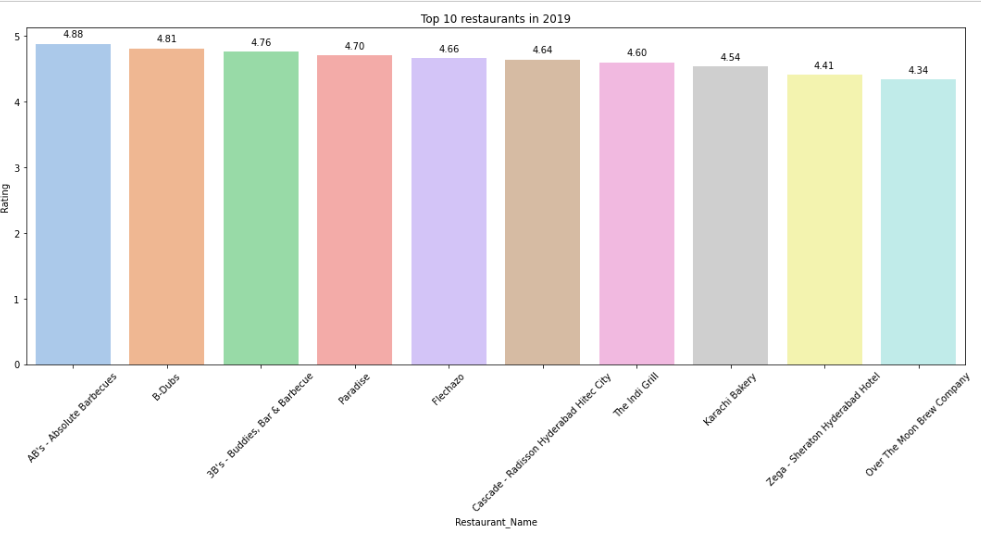
Text preprocessing is a method to clean the text data and make it ready to feed data to the model. Text data contains noise in various forms like emotions, punctuation, and text in a different case.

Essentially text-preprocessing has below operations to perform:

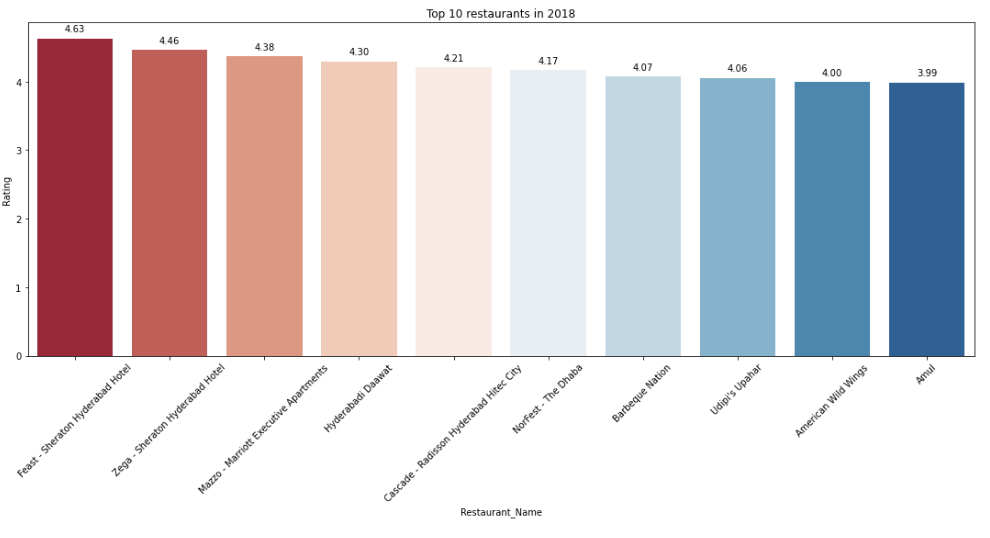
1. Loading the data.
2. Data cleaning (Stop word removal, punctuations, emoji’s,Expand Contractions, double space removal, lower case conversion etc).
3. Text reduction (stemming and lemmatization).
4. Vectorization.
5. Bag of words.
6. TF-IDF.
7. Modeling.
8. Performance evaluation.
9. Hyperparameter tuning.

**Data Visualization**

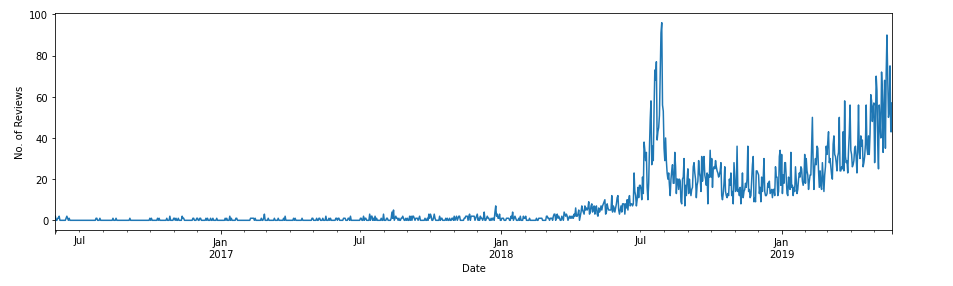
This step of visualization involves checking the data in pictorial form. Apart from this, visualization helps in easy understanding of the data and representation of most of the data information in right and quick fashion.



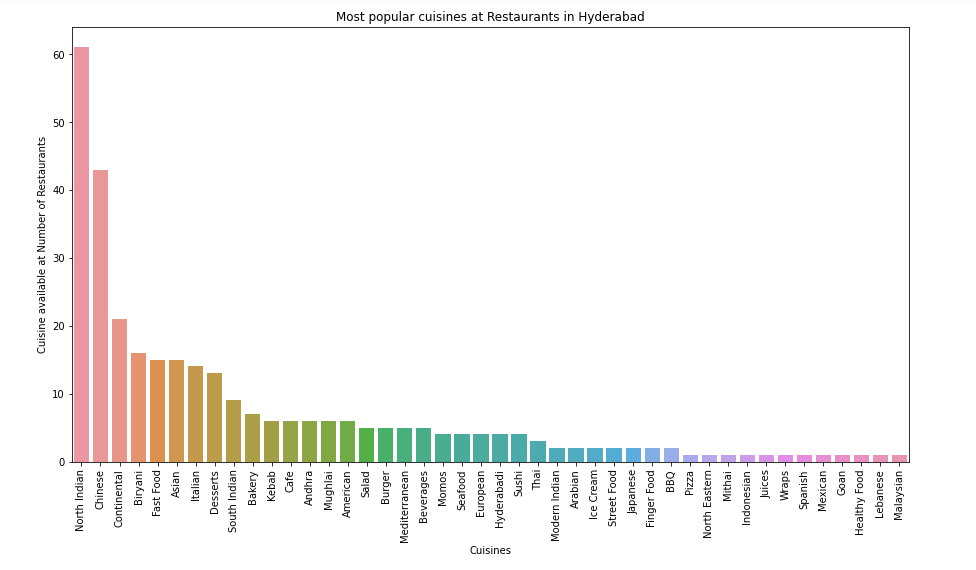
This pictorial representation is for top 10 restaurants in the year of 2019

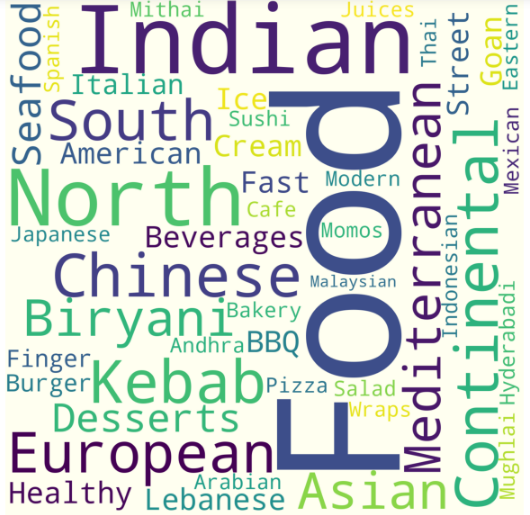


This pictorial representation is for top 10 restaurants in the year of 2018



This above pictorial representation is telling the review trend with years from 2017 to Jan of 2019 and it seems that frequencies of reviews has increased in 2018 and 2019 as the year progress.





These are some of the highlighted frequencies of the cuisines which are popular and are repeated.

Few populars are below.

Indian,Food,North,Chinese,European,Kebab,South,continental etc.

**Modelling**

Before we proceed with model building, it is important that we have the right amount of data and we understand the features/variables very well. Almost every model takes the numerical data for further processing in the model part so it is important to convert the entire data into numerical columns so that we can process it further.

In the case of text preprocessing we have to perform various steps depending upon the pre-processing required for the dataset. It can be word expending contractions where we have to convert don’t to do not and entire of word representation of its individual, and then removal of double spaces using regular expressions and removal of various punctuations present in the word which will not contribute much for the model and weightage, converting to the lower case all the words is also important because model may consider the same word with different case may consider differently, so it is important to convert the words into same case.

While, apart from this we have to reduce the words as well by using stemming and lemmatization and converting to base word.

It is equally important to classify your columns into dependent and independent variables if we are going fatherly process the same data for modeling to see the classes and predictions. In this case I have used bag of words which has converted the text into 0 and 1 but it provides the same score for the term so, in order to overcome this problem I have made use of TF-IDF which basically tells how important a word is actually by assigning so weightage based of occurring frequencies of words Once this task is performed we are ready for the next step in modeling which can be splitting the training and testing data.

 Next in this process, we would need some of the machine learning libraries and suitable models for the prediction After fitting the data look for the scores if we need to improve the scores, look for the scope of improvement if required, hyperparameter tuning or in this case of classification we can use ensembles techniques like boosting and bagging and Finally Model has to undergo a thorough evaluation which checks whether your model is doing good in terms of prediction in this case of classification evaluation parameter can be accuracy, precision, Recalls, F1 score. And for regression it can be RMSE, MAE, SSE etc.

**Model Used on data**

K-mean clustering.

Bad of words.

Logistic Regression.

MultinomialNB.

Decision tree.

Random forest.

**Procedure performed and Observations**

 Started with data loading and importing the libraries and then started with exploring the data and looking into columns and rows. Looked for the missing values in the dataset and imputed accordingly.

Did the EDA for the dataset and tried to find out the trend in the data. Got some of the questions like which are the cuisines which are popular and what are the weekly restaurant timings, and explored the cheap restaurants and costly restaurants in the given dataset. I have found that there were prices around 2500-2800 to maximum or costly prices and where 250 rupees to 150 rupees restaurants were the cheapest. I also have explored some commonly used words like good, order, food, India, Chinese etc.

Once data exploration was done I have started clustering using cuisines where 61 being the highest number for the restaurant with North Indian cuisines and having 7 as top 10th was bakery.

**Moved to the Natural Language Processing (NLP)**

I have performed some of the tasks for the text understanding:

Started with Data-loading and data preprocessing where data preprocessing has below steps to perform.

·         **Expand Contractions-(don’t to do not).**

**·         Lower Case-(Convert to lower case).**

**·         Remove Punctuations-(!?:).**

**·         Remove words and digits containing digits- (game5ts7 replace with empty string).**

**·         Remove Stopwords-(This, that, is, am are).**

**·         Rephrase Text(converting email ids to email address).**

**·         Stemming and Lemmatization. (Playing to play, and Methodology-Method)**

**·         Remove White spaces.( simply removal of extra spaces).**

Finally, started with converting the categorical data into numerical data with the vectorization. And tried sentimental analysis while saying that .5 above are positive sentiments and below 3.5 are negative sentiments also made use of bad of word also for the frequency improvement applied the TF-IDF it basically tells that how important your words are. Then I tried with the word weightage and used the models for classification and used random forest and logistic regression and multinomialNB, tried K-mean clustering. I also used the cross validation for the getting best hyperparameters. I have seen that logistic regression and random forest given great results over others.

**Challenges.**

The major challenges I have faced in this project are mentioned Below:

1)Analyzing the data and exploring the data was not making much sense because the data that I had was having objects, data types and few strings and comments, restaurant names , reviews etc.

2)Removal of punctuations and stopwords, symbols, emojis and some of the repeated words like (myyyyyyyyyyyyyyyyy).

3)I have faced challenges in merging the data frames as they were not matching the length.

4)I have faced the challenge while importing the libraries. I had to install some of the libraries like spacy, contractions