# PES University, EC Campus
# Data Analytics (UE19CS312)

# Phase 1 Report

**Team Name:** Insight Strategists

**Team Members**:

| Name: | SRN: |
|---|---|
| Aarushi Agarwal | PES2UG19CS004 |
| Anwesha Kelkar | PES2UG19CS058 |
| Arjun Harish | PES2UG19CS062 |
| Mohnish Srikanth | PES2UG19CS241 |

**Problem Statement:**
Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants.

**Description of Problem Statement:**
Reviews on Zomato are still in the form of text and can be classified with positive, negative, or neutral ratings. Zomato doesn't have an analysis of how users interact with the reviews and which words indicate that the customers enjoyed their restaurant experience or not. We need to extract the words from the reviews and analyse them so we know how users interact in the Zomato review section.

**Link to the Dataset:**
https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

**Brief Description of the Dataset:**
Bengaluru has more than 50,000 restaurants serving dishes from all over the world. The basic idea of the Zomato dataset is to summarise all the

details of these restaurants such as name, address, phone number, website url etc in a tabular form. It also contains customer feedback in the form of ratings, most liked dish, customer reviews etc.
The data is in a csv format and is approximately 547MB in size. The dataset contains 17 variables all of which were scraped from the Zomato website. The dataset contains details of 51,717 restaurants in Bengaluru.

# Literature Review Report

**Food Reviews Classification using multi-label convolutional neural network text classifier**

Research conducted by Krutuja S Lasne, Sejal S Nandrekar, Ashraf A Khan,Tushar Ghorpade from Ramrao Adik Institute of Technology named "Food Reviews Classification using multi-label convolutional neural network text classifier" aims to create an automatic text-based classification model that can forecast the usefulness of Zomato, Swiggy, and Reddit feedback accurately. The project uses SpaCy, which is a powerful tool to detect sentiment analysis on the review. This project has a precision of 86.40%, recall of 87% and F-score of 86.70%.

**Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier**

"Sentiment Analysis of Customer Reviews in Zomato Bangalore Restaurants Using Random Forest Classifier" by Bern Jonathan , Jay Idoan Sihotang and Stanley Martin uses a method to analyze user's sentiment of Zomato Restaurants and focusing review in Bangalore for study cases. A Random Forest Classifier is used to classify the sentiments of users based on their review. In addition to this, the specific words that affect the model classifier are also found. This project deals with imbalanced data in positive, negative, and neutral data. Imbalanced dataset algorithms can be used to improve these results.

**Sentiment Classification on Twitter Dataset Using Supervised Learning Algorithms**

"Sentiment Classification on Twitter Dataset Using Supervised Learning Algorithms" uses Coronavirus tweets dataset from IEEEDataPort for their research. In this work, tweets are extracted from Twitter associated to COVID-19 and different sentiment analysis techniques are performed. After calculating sentiment score, Machine learning algorithms have been applied to calculate accuracy of each classifier and completed comparative analysis. It was found that Random Forest classifiers had the highest percentage of accurate results of 94.9% and Naive Bayes had the lowest accuracy of 45%.

### An Analysis of Online Food Ordering Applications in India: Zomato and Swiggy

"An Analysis of Online Food Ordering Applications in India: Zomato and Swiggy" research paper by Anupriya Saxena aims to get an insight on the emerging innovative technologies in the restaurant industry and strategies followed by online food startups like Zomato, Swiggy. From this research paper we can understand drivers of online food sites. Different services given by application that makes consumers happy and satisfied .Comfort and Convenience which makes consumers more inclined towards online food ordering.

### Innovation resistance theory perspective on the use of food delivery applications

"Innovation resistance theory perspective on the use of food delivery applications" research paper uses Innovative Resistance Theory (IRT) and helps to understand the resistance toward food delivery applications (FDAs). This study has adapted the existing criteria to measure different consumer barriers toward FDAs. It also examined the relationships between various consumer barriers, intention to use FDAs and word-of-mouth.

### Decoding the effect of restaurant reviews on customer choice: Insights from Zomato

"Decoding the effect of restaurant reviews on customer choice: insights from zomato" research paper is written by Vaishnavi Vajjhala and Munmun Ghosh.

In this study we try to understand the impact of online reviews and the nature of reviews on consumer's purchase intention on the platform of Zomato. We use a mixed-method approach using Sequential Explanatory Research design. The results confirm that online reviews and star ratings available on Zomato significantly impact the willingness to purchase.

### Zomato: a shining armour in the food tech sector

In the case study, "Zomato: a shining armour in the food tech sector" by Prashant Raman, we try to understand the life cycle of Zomato right from scratch to its growth and consolidation phase. The case uses a secondary research method. The information, interview excerpts and data related to the company are gathered from different sources such as online databases, magazines, personal blogs of founders and company websites.This case identifies the different factors that were instrumental to the success of Zomato, its various revenue sources, its development through new products and services and growth expansion through acquisitions.

### The effects of pictures, review credibility and personalization on users satisfaction of using restaurant recommender apps: Case study: Zomato

In the study, "The effects of pictures, review credibility and personalization on users satisfaction of using restaurant recommender apps: Case study: Zomato" by Dan Qraved, we aim to determine the effects of pictures, reviews credibility and

personalization feature on users' satisfaction in accessing restaurant recommender apps. This study uses a quantitative approach with a total of 419 collected respondents. The data gathered within this study is being analyzed using multiple regressions using SPSS 22.0. The results of this study indicate that pictures and reviews credibility are affecting user satisfaction on using restaurant recommendation apps. Additionally, providing personalization on a restaurant recommender app may also improve its user satisfaction in accessing the apps.

## Popularity of online food ordering and delivery services- a comparative study between zomato, swiggy and uber eats in Ludhiana

"Popularity of online food ordering and delivery services- a comparative study between zomato, swiggy and uber eats in Ludhiana" by Ashish Raina , Varinder Singh Rana , Dr. Arun Singh Thakur, aims to gauze the customer reviews and satisfaction towards the available online food ordering and delivery services in Ludhiana. The study further compares various aspects of the three available food delivery services in the area. Based on the the findings of the study, service providers can meet customer expectations in a better way.

## Location based restaurant preferences in Bangalore

"Location based restaurant preferences in Bangalore" is written by Dr. Anupam Bhatia and Ms. Sneha from CRSU, Jind. Hadoop ecosystem integrated with HIVE data-warehouse is used in this work. A restaurant is chosen and is empirically investigated to analyze how the success of this restaurant can be affected by its location and address.

## Analyzing Scientific Papers Based on Sentiment Analysis

"Analyzing Scientific Papers Based on Sentiment Analysis" by Doaa Mohey El-Din Mohamed Hussein uses a SAOOP model that uses natural language processing, text analysis and opinion mining in the sentiment analysis evaluation. It presents an enhancement of the bag-of-words model and creates a new miniature lexicon. The proposed technique uses a bag-of-words model for fitting the review structure which is short and relevant to a scientific topic domain. The proposed technique solves two essential Bag-of-words weaknesses to improve accuracy.

## BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" uses BERT which is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5%, MultiNLI accuracy to 86.7%, SQuAD v1.1 question answering Test F1 to 93.2 and SQuAD v2.0 Test F1 to 83.1.

# Exploratory Data Analysis & Visualization

**Link to EDA & Visualization code:**
https://github.com/AnweshaKelkar/Zomato-Bangalore-Restaurants-Analysis

1. Importing the required libraries: For our EDA & visualization, we required the following libraries.

```
#Importing libraries

import numpy as nps
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
#import plotly.express as px
import re
import warnings
warnings.filterwarnings('ignore')
```

2. Displaying the initial number of rows and columns in our dataset: We initially have 51717 rows and 17 columns.

```
Out[4]: (51717, 17)
```

The following are the column names:

```
Out[6]: Index(['url', 'address', 'name', 'online_order', 'book_table', 'rate', 'votes',
       'phone', 'location', 'rest_type', 'dish_liked', 'cuisines',
       'approx_cost(for two people)', 'reviews_list', 'menu_item',
       'listed_in(type)', 'listed_in(city)'],
      dtype='object')
```

3. Dimensionality reduction : we remove certain columns that are not required for our study such as: 'url', 'address', 'phone', 'dish_liked', 'menu_item'. We also change the names of certain columns to more meaningful ones: approx_cost(for two people)' to 'approx_cost', 'listed_in(type)' to 'type' and 'listed_in(city)' to 'city'.

```
Out[9]: Index(['name', 'online_order', 'book_table', 'rate', 'votes', 'location',
               'rest_type', 'cuisines', 'approx_cost', 'reviews_list', 'type', 'city'],
              dtype='object')
```

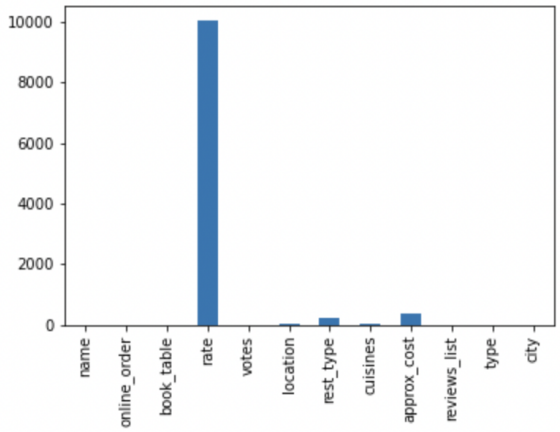
4. Duplicate Data: We check for duplicate values using the duplicated() function and drop them using the drop_duplicates() function. After that, we are left with 51077 rows and 13 columns.

```
Out[41]: (51077, 13)
```

5. Missing values: We check for number of missing values in each column using isnull().sum() function

```
Out[19]: name                0
         online_order        0
         book_table          0
         rate            10023
         votes               0
         location           21
         rest_type         227
         cuisines           45
         approx_cost       345
         reviews_list        0
         type                0
         city                0
         dtype: int64
```

Using a bar plot to help visualise this:



We find that ‘rate’ has the highest number of missing values. We replace all the null values with the mean of all of the rate values as removing all of them will cause data loss. For this we use the fillna() function. We drop all the remaining rows with missing values using the dropna() function.

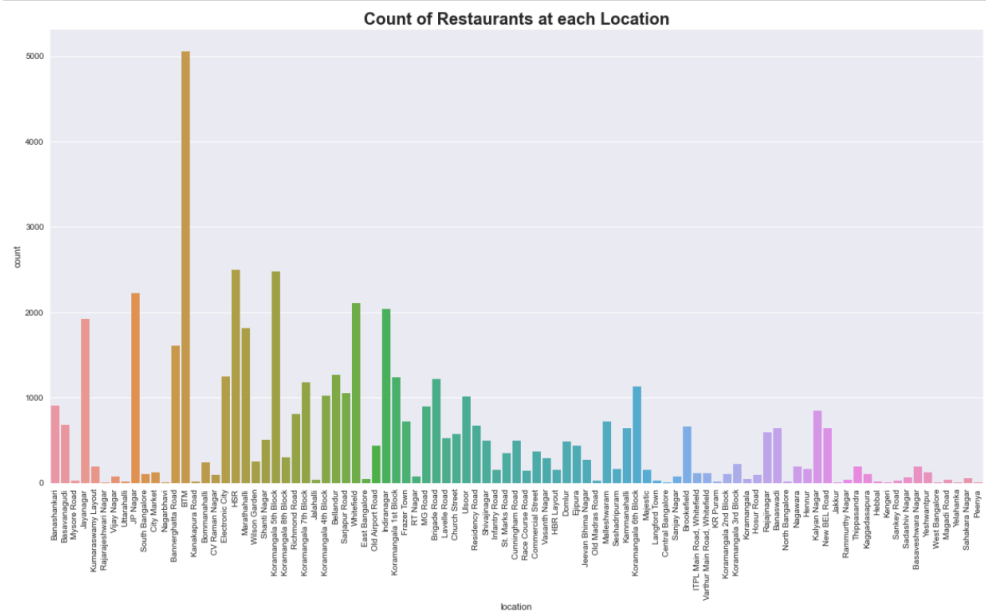
```
Out[23]: name            0
         online_order    0
         book_table      0
         rate            0
         votes           0
         location        0
         rest_type       0
         cuisines        0
         approx_cost     0
         reviews_list    0
         type            0
         city            0
         dtype: int64
```

6. Adding some meaningful information : Making a new column for total number of cuisines offered by each restaurant name ‘Number_of_cuisines_offered’

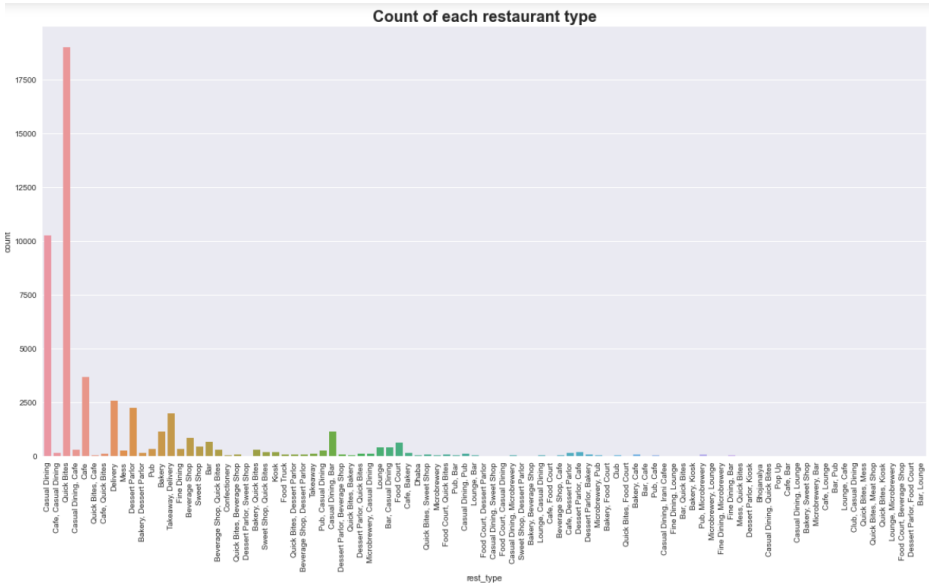
```
Out[42]: Index(['name', 'online_order', 'book_table', 'rate', 'votes', 'location',
                'rest_type', 'cuisines', 'approx_cost', 'reviews_list', 'type', 'city',
                'Number_of_cuisines_offered'],
               dtype='object')
```

7. Visualization: Visualizing the data to understand clearly what the dataset is trying to convey.

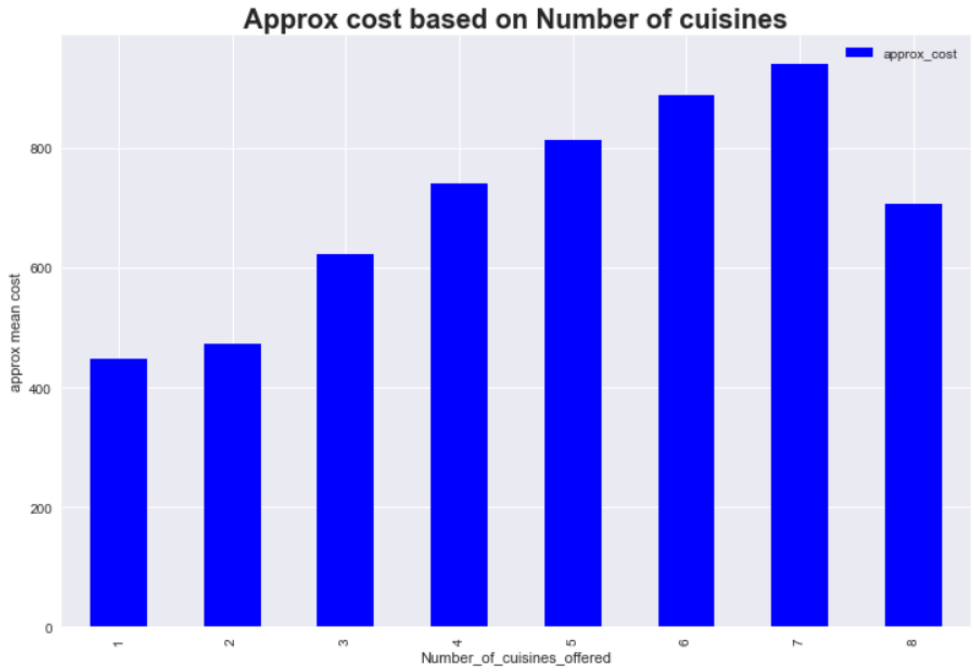
(1) Count of Restaurants plotted against Location



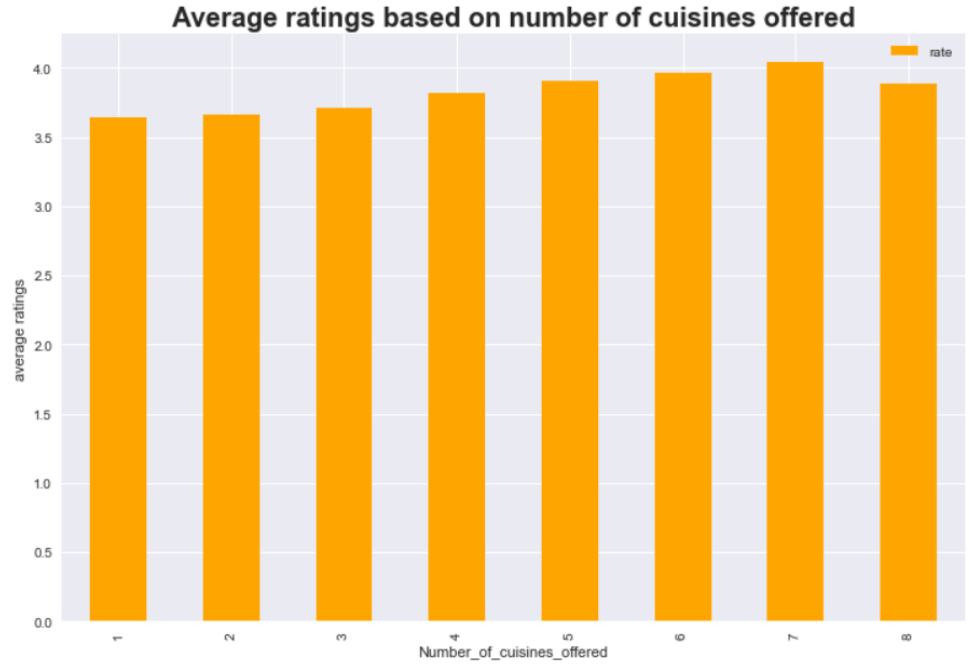

(2) Plotting restaurants by their type



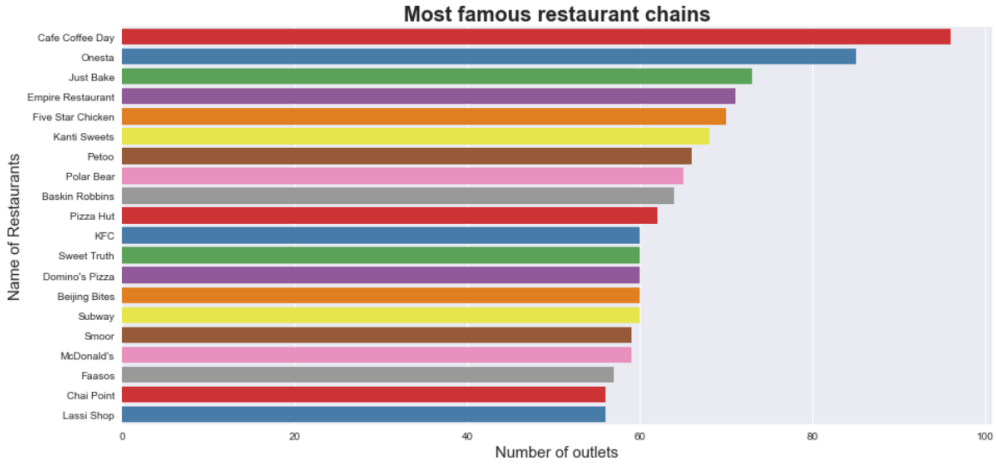
(3) Approx cost plotted against the Number of cuisines offered by the restaurant




(4) Average rating against the number of cuisines offered



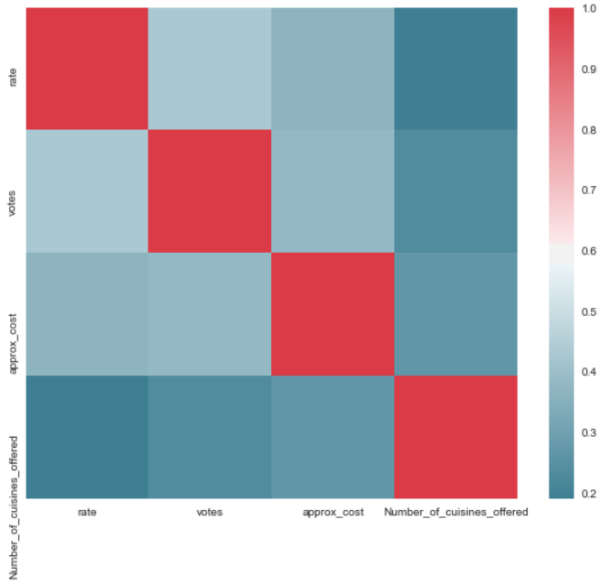
(5) Most Famous restaurant chains and the number of outlets they have



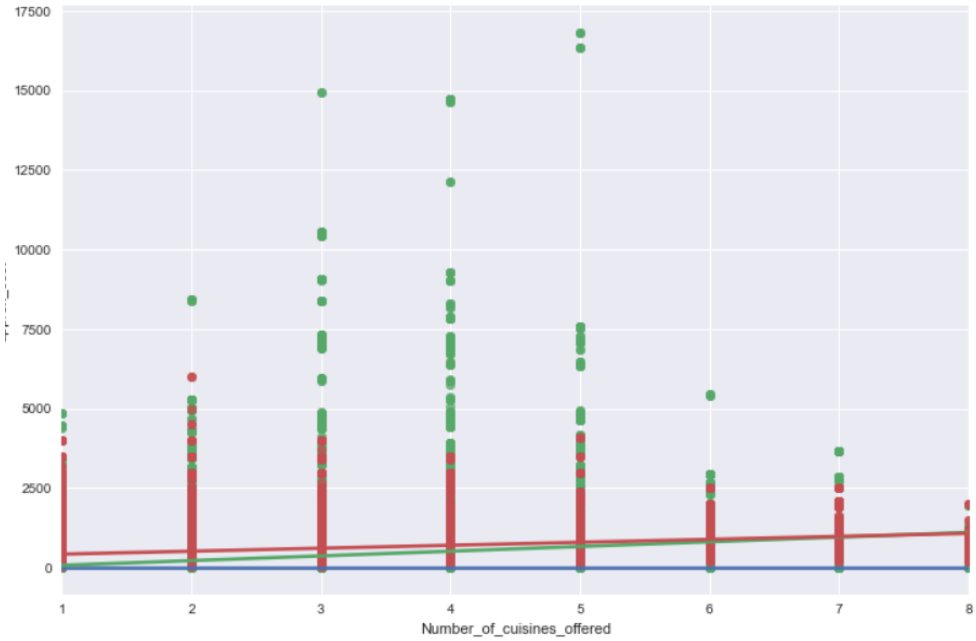

8. Correlation: Finding the correlation between all numeric attributes in the Dataframe

| | rate | votes | approx_cost | Number_of_cuisines_offered |
|---|---|---|---|---|
| **rate** | 1.000000 | 0.428843 | 0.366382 | 0.190368 |
| **votes** | 0.428843 | 1.000000 | 0.381648 | 0.232068 |
| **approx_cost** | 0.366382 | 0.381648 | 1.000000 | 0.268574 |
| **Number_of_cuisines_offered** | 0.190368 | 0.232068 | 0.268574 | 1.000000 |

Plotting a Heatmap for the same



Correlation plots showing the correlation of Rate, Votes and Approx Cost respectively  against number of cuisines offered




# References:


- https://scholar.google.com/
- https://www.kaggle.com/datasets
- https://jurnal.unai.edu/index.php/isc/article/view/1003