Utilizing Restaurant Data to Predict Rating of Future Restaurants

Milestone: Project Report

Group 39

Yash Nema

Email ID: Nema.y@northeastern.edu

—

Submission Date: Apr 25, 2022

Problem Ideation

Problem Setting:

- The restaurant industry has tripled in last 25 years and with the advent of Online ordering apps like Zomato, Uber Eats etc., we see it will be crucial for every restaurant owner to use data to understand the ever-changing market.
- Restaurant owners will need to use ratings as a measure of their customer approval. Ratings will be highly dependent on the location in which they are located, the average cost a meal the customers must pay, customer service they provide and many more.
- New restaurant owners will need to take data driven decisions while deciding the locality, prices and cuisines offered.

Problem Definition:

- Data Source: Zomato App (Kaggle dataset) - ~10K restaurants from multiple countries and ~50K restaurants from City of Bangalore
- Goal 1: Understand what factors impact Rating of a restaurant
- Goal 2: Identify and cluster restaurants based on common attributes and see trends
- Goal 3: Predicting new restaurant rating depending on input parameters
- Goal 4: Building recommendation system to suggest restaurant based on customer preference.

Data Sources:

#1: Zomato API(Kaggle Dataset)

- Restaurant Search API to collect metadata of restaurants.
- Specific Categories: Cuisines, City, Country, Ratings etc.
- Data can be downloaded as .csv file.
- https://www.kaggle.com/shrutimehta/zomato-restaurants-data

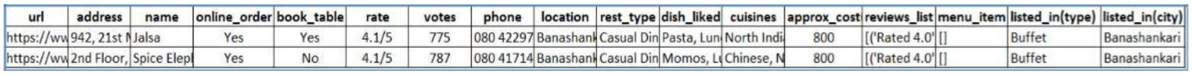
Data Description:

Table 1: Restaurant Data Dictionary (Dataset 1)

Sn	Column Name	Description	No. of rows
1	Restaurant ID	Restaurant id (Unique)	9551
2	Restaurant Name	Name of the restaurant	9551
3	Country Code	Country in which restaurant is located	9551
4	City	City in which restaurant is located	9551
5	Address	Address of the restaurant	9551
6	Locality	Location in the city	9551
7	Locality Verbose	Detailed description of the locality	9551
8	Longitude	Longitude coordinate of the restaurant's location	9551
9	Latitude	Latitude coordinate of the restaurant's Location	9551
10	Cuisines	Cuisines offered by the restaurant	9551
11	Average Cost of Two	Cost for two people in different Currencies	9551
12	Currency	Currency of the country	9551
13	Has Table Booking	Yes/No	9551
14	Has Online Delivery	Yes/No	9551
15	Is delivering Now	Yes/No	9551
16	Switch to Order Menu	Yes/No	9551
17	Price Range	Range of price of food	9551
18	Aggregate Rating	Average rating out of 5	9551
19	Rating Color	Depending upon the average rating Color	9551
20	Rating Text	Text on the basis of rating of rating	9551
21	Votes	Number of ratings casted by people	9551

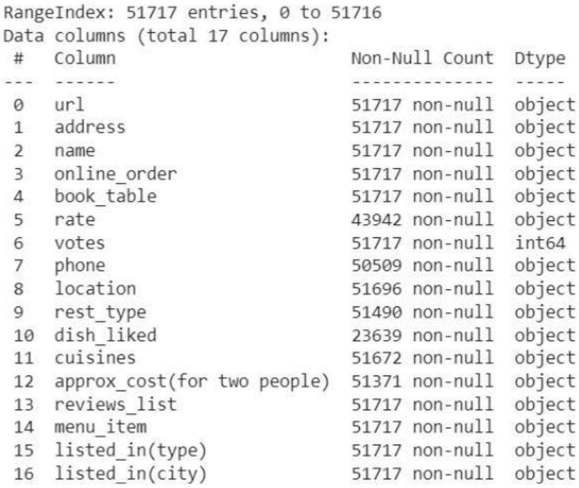
Data Collection and Processing

1. **Data Collection**: The Zomato restaurant review dataset consisted of two JSON files. One file, with aggregated ratings of restaurants mostly in the NCR region of India (50% dataset rows) and approximately 20 rows only for other Indian cities and other countries with Zomato. The second file consisted of Restaurant ratings based in Bangalore city in India. The first dataset consisted of ~10K restaurants and 21 columns. We will be focusing analysis more on the second file which contains ~56K rows and 17 variables (larger dataset and more depth for doing analysis). Next, while analyzing the data types, we noticed that the attributes were of both numeric and categorical types in this dataset. Categorical variables either have just two categories or have multiple categories. The target variable in the dataset is the "rate" attribute which consists of ratings from 0 to 5. (Sample attached below for the dataset before preprocessing) Further, based on the analysis in the exploration phase of the project (Milestone 3) we will explore if NCR region restaurants and Bangalore region datasets should/can be combined or kept separate for the entire analysis. Currently we would like to start with exploring the dataset with Bangalore focus.



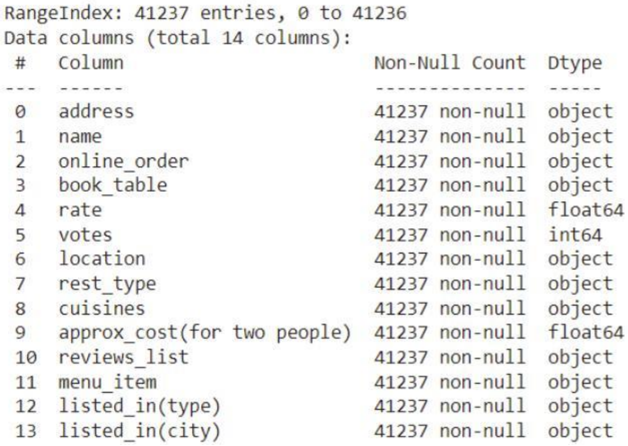
url	address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_cost	reviews_list	menu_item	listed_in(type)	listed_in(city)
https://ww	942, 21st M	Jalsa	Yes	Yes	4.1/5	775	080 42297	Banashank	Casual Din	Pasta, Lun	North Indi	800	[('Rated 4.0'	[]	Buffet	Banashankari
https://ww	2nd Floor,	Spice Elep	Yes	No	4.1/5	787	080 41714	Banashank	Casual Din	Momos, Lu	Chinese, N	800	[('Rated 4.0'	[]	Buffet	Banashankari

2. **Data Preprocessing:** The dataset was converted from JSON format to CSV format and downloaded. This process led to some parsing issues like data being put in incorrect columns or some columns being blank. So, additional processing is done to read the special characters in Python environment while loading the dataset. In the read_csv function, the parameter "engine" is used to read the CSV file correctly. Attributes "URL" and "phone" are dropped as they do not add additional information to the analysis. The attribute "dish_liked" consists of ~28K nulls, and so will not be used for predictive analysis. Only 4 other columns consisted of nulls. We can drop these rows for the exploration phase. We haven't filled any nulls with a specific statistic. Only rows with nulls in the target variable "rate" can be used as new data to predict the rate for those restaurants. Will take this decision based on the data exploration phase. The rate column consisted of data as "4.2/5.0", so had to be processed to extract just the rating and discard the "/5.0" Further, we noticed ~40 duplicate rows. These rows are dropped before going to the exploration phase. Also "votes" attribute is converted to int, "rate" and "approx_cost (for two people)" are converted to float format. Column encoding and scaling will be done based on the use case requirements in the next phases.



```
RangeIndex: 51717 entries, 0 to 51716
Data columns (total 17 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   url                          51717 non-null  object
 1   address                      51717 non-null  object
 2   name                         51717 non-null  object
 3   online_order                 51717 non-null  object
 4   book_table                   51717 non-null  object
 5   rate                         43942 non-null  object
 6   votes                        51717 non-null  int64
 7   phone                        50509 non-null  object
 8   location                     51696 non-null  object
 9   rest_type                    51490 non-null  object
 10  dish_liked                   23639 non-null  object
 11  cuisines                     51672 non-null  object
 12  approx_cost(for two people)  51371 non-null  object
 13  reviews_list                 51717 non-null  object
 14  menu_item                    51717 non-null  object
 15  listed_in(type)              51717 non-null  object
 16  listed_in(city)              51717 non-null  object
```

Before Preprocessing



```
RangeIndex: 41237 entries, 0 to 41236
Data columns (total 14 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   address                      41237 non-null  object
 1   name                         41237 non-null  object
 2   online_order                 41237 non-null  object
 3   book_table                   41237 non-null  object
 4   rate                         41237 non-null  float64
 5   votes                        41237 non-null  int64
 6   location                     41237 non-null  object
 7   rest_type                    41237 non-null  object
 8   cuisines                     41237 non-null  object
 9   approx_cost(for two people)  41237 non-null  float64
 10  reviews_list                 41237 non-null  object
 11  menu_item                    41237 non-null  object
 12  listed_in(type)              41237 non-null  object
 13  listed_in(city)              41237 non-null  object
```

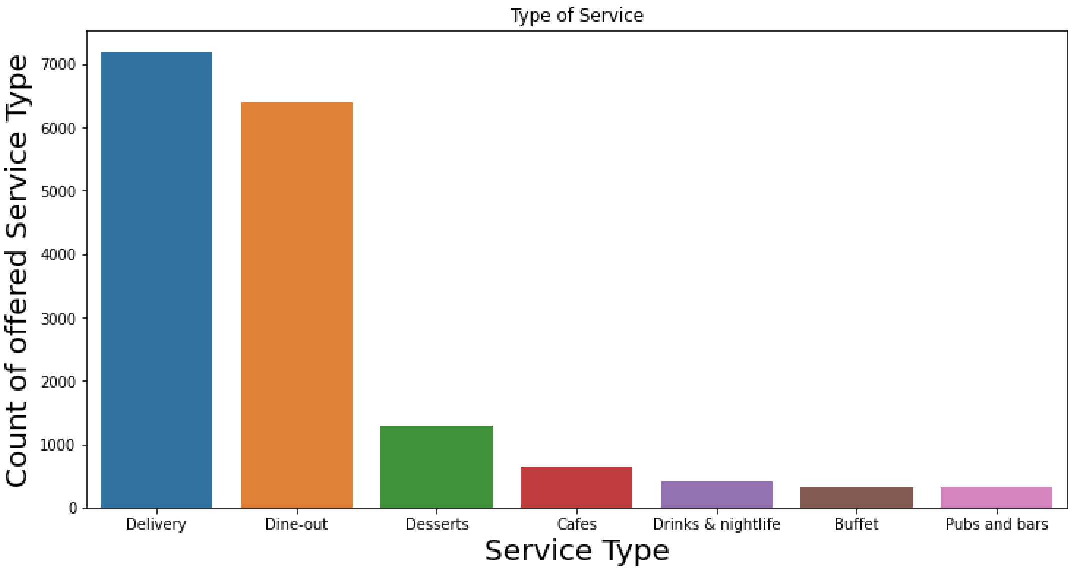
After Preprocessing

Data Exploration and Visualization

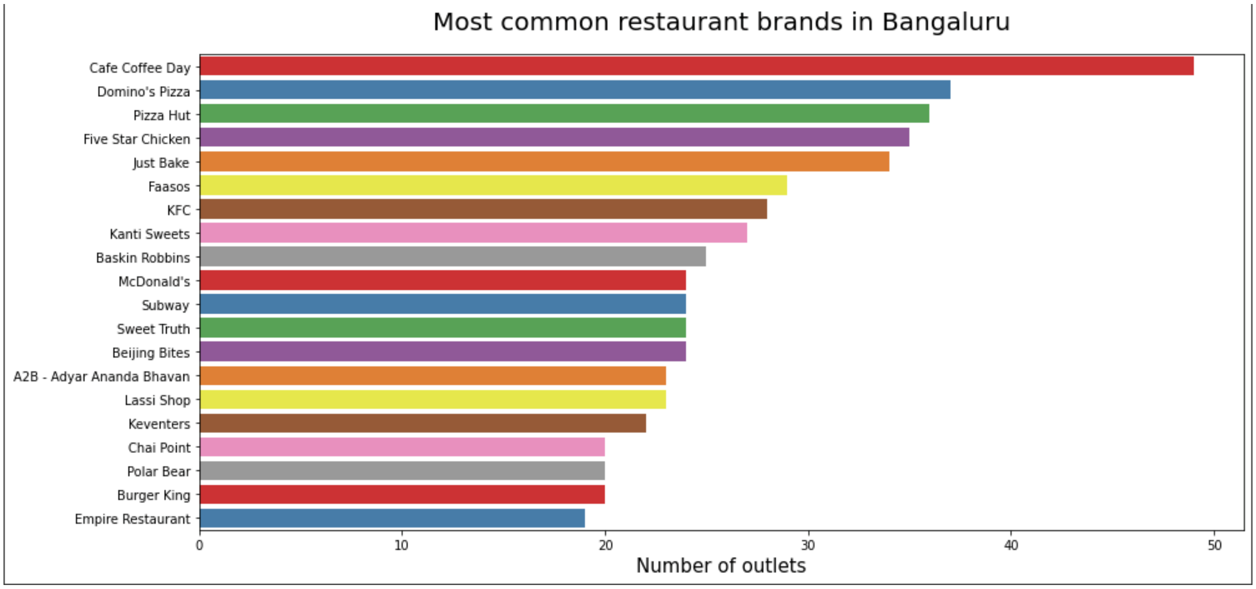
Data Exploration: · Data exploration is the first step in data analysis involving the use of data visualization tools and statistical techniques to uncover data set characteristics and initial patterns.

We find various trends, characteristics and statistics for the current dataset using data visualization. We have used different libraries such as pandas, matplotlib, numpy, seaborn etc.

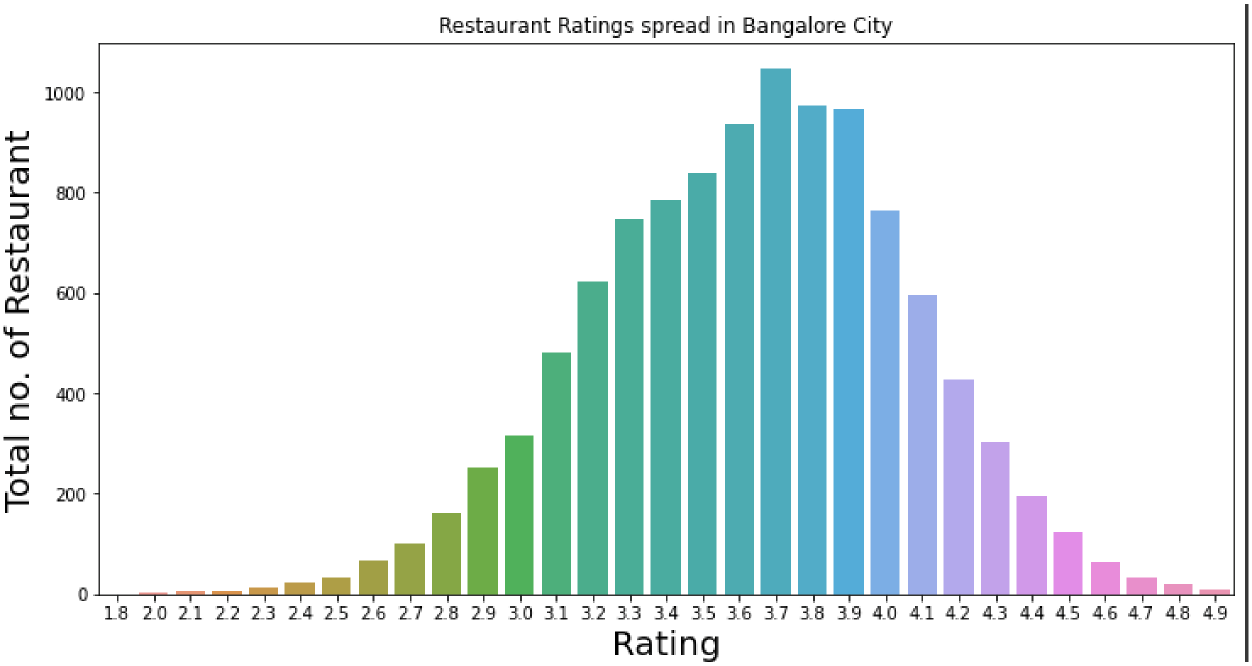
1) Bar Chart



Insights: - Restaurants which provide delivery is highest followed by Dine-out options



Insights: - **Cafe Coffee Day** and **Dominos** have the most branches. The top 5 restaurants are ones which people would go to on a regular basis. These restaurant brands could have low franchise cost



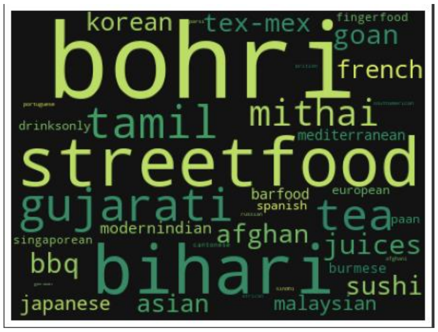
Insights: - Most restaurant have a rating of **3.7**

EDA (Explanatory Data Analysis):-It is a crucial part of any data science project because that's where you get to know more about the data. In this phase, we find hidden patterns in the data and generate insights from it.

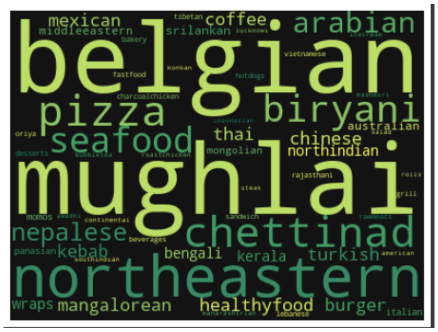
2) **Word cloud** is a great way to represent text data. The size and color of each word that appears in the word cloud indicate its frequency or importance.

Analysis based on the average cost vs Cuisines

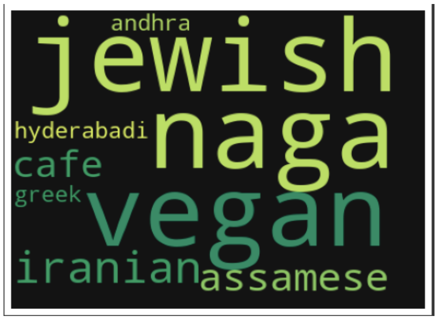
The Average Cost is <300 and preferred cuisines are Bohri and Street Food



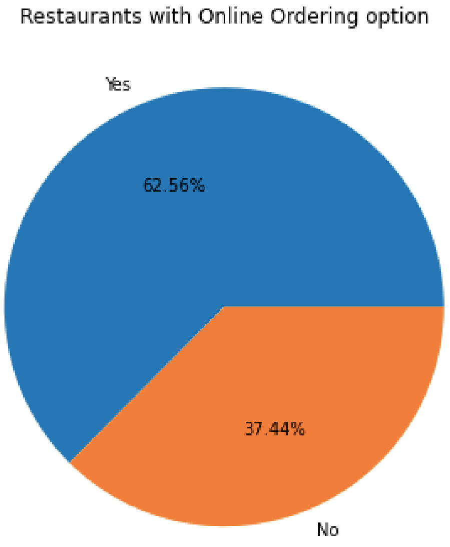
The Average Cost is 300 to 500 and preferred cuisines are Belgian and Muglai



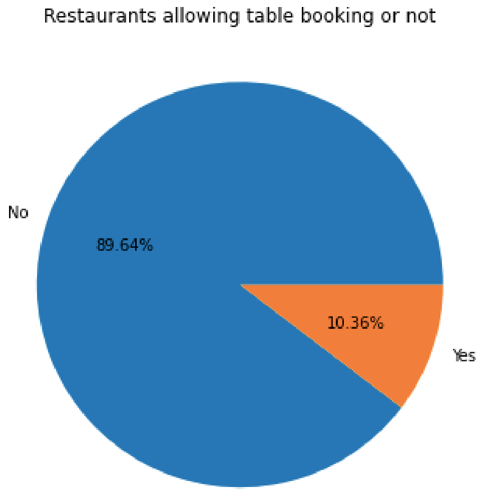
The Average Cost is 500+ and preferred cuisines are Jewish and Naga



3) Pie- Charts



Insights: - We got to know that about 63% of the Restaurants have the option of ordering online



Insights: - We got to know that only 10% of the Restaurants have the option of booking table that we most restaurant allow direct walk-ins

Model Exploration and Model Selection

Based on above EDA insights we want to further breakdown the analysis goals as follows:

Goal 1: Understand what factors impact Rating of a restaurant
Goal 2: Identify and cluster restaurants based on common attributes and see trends
Goal 3: Predicting new restaurant rating depending on input parameters
Goal 4: Building recommendation system to suggest restaurant based on customer preference. Model Selection

Methodology and Goals:

- **Goal 1**: Understand what factors impact Rating of a restaurant

To understand factors impacting restaurant rating we start by calculating correlation and then exploring variable importance using models like Linear Regression and Decision trees.

As correlation function and sklearn models take numeric inputs only we need to convert categorical variables into numeric variables.
We have two approaches:

1) One Hot Encoding
2) Encoding categories(each category as unique integer)

Advantages and Disadvantages of **Approach 1**

- If we have x categories in a column we will be left with x additional columns.
- As we have location like column with 30+ categories, we will have 30 additional column to One Hot Encode this column
- This will result in huge tables, high complexity, high number of 0s in rows, difficult to look through the data

Might lead to overfitting using Tree Based algorithms

Advantages and Disadvantages of **Approach 2**

- Not often used but we can add numerical encoding based on freqency of category or just adding a numerical value to identify the category
- This will result in just one column, so less complex, but with this approach higher numerical value may lead to bias
- That is it may show that higher the value higher the weight to be given which is not the case. So need to be careful
- Most tree-based models (SKLearn Random Forest, XGBoost, LightGBM) can handle number-labeled-columns very well.

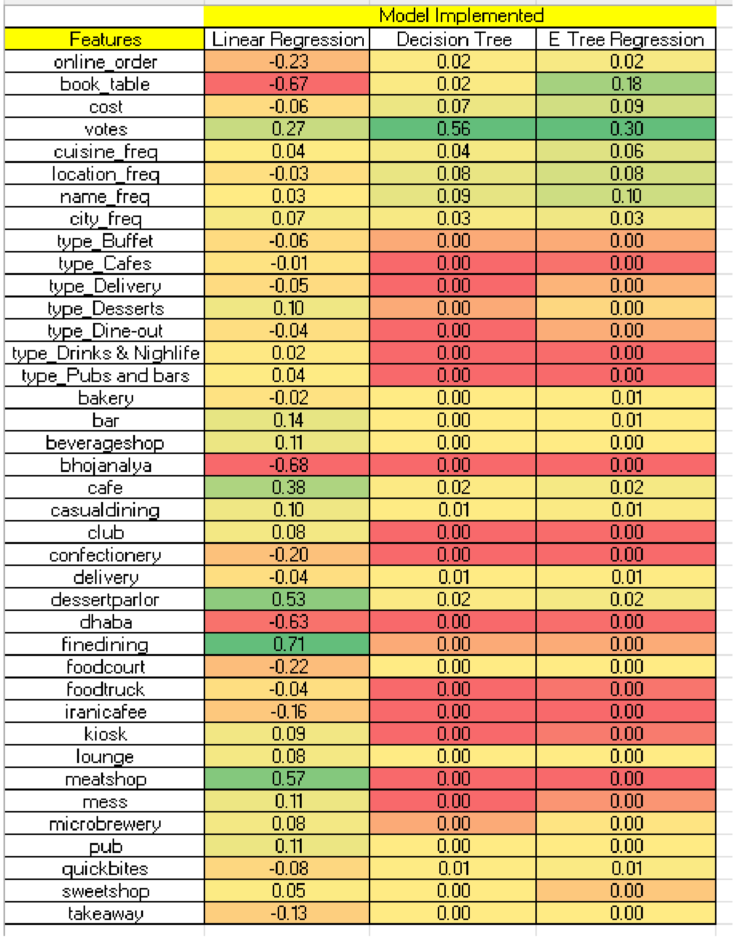
Model Approach
We have implemented both approaches and tested the models.

- We have evaluated approaches based on the Regression model(train and test data) R2 Scores, MAPE, RMSE scores and models were also compared with the Naive Model
- We will be imputing few of the categorical variables(more than 30 categories) by frequency of occurrence
- Categorical variables with binary categories will be label encoded to 1 or 0

- Categorical variables with less than 30 categories and greater than 2 will be one hot encoded numeric variable will be feature scaled.

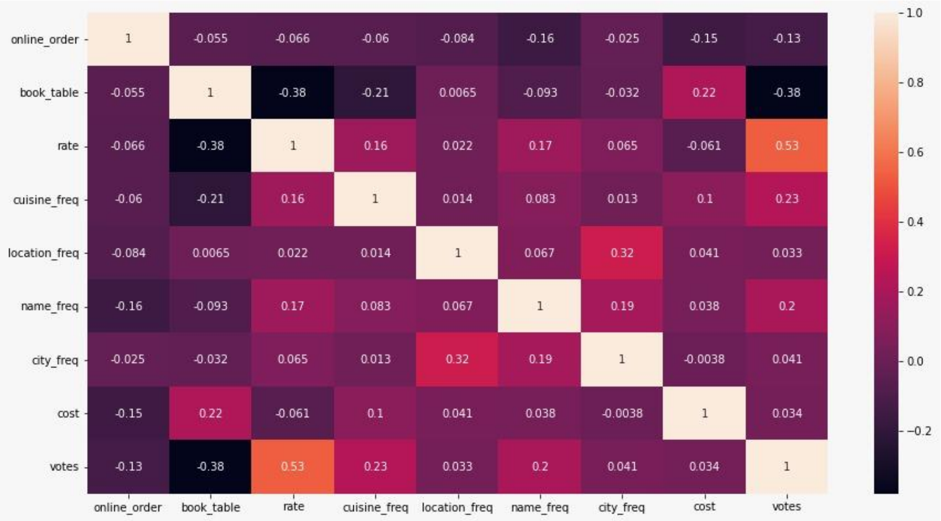
We can now use the columns for Model building and analysis

- Based on the EDA creating a final correlation chart. All columns are not highly correlated.
- Highest correlation between two columns was +0.53, so we don't need to drop any column specifically.
- To understand the importance of each variable and understand why a specific restaurant has a high rating or low rating we can build Multiple linear regression model or Decision tree or Random Forrest model. We can get the variable importance using the model coefficients. In the next Milestone we showcased the predictors and the importance score using each of the methods.
- We won't be able to use Naïve Bayes Model here as we have numeric predictors and binning them may remove the uniqueness and characteristic elements for individual restaurants
- We won't be implementing PCA as we want a explainable model to help understand which predictors impact ratings.



	Model Implemented		
Features	Linear Regression	Decision Tree	E Tree Regression
online_order	-0.23	0.02	0.02
book_table	-0.67	0.02	0.18
cost	-0.06	0.07	0.09
votes	0.27	0.56	0.30
cuisine_freq	0.04	0.04	0.06
location_freq	-0.03	0.08	0.08
name_freq	0.03	0.09	0.10
city_freq	0.07	0.03	0.03
type_Buffet	-0.06	0.00	0.00
type_Cafes	-0.01	0.00	0.00
type_Delivery	-0.05	0.00	0.00
type_Desserts	0.10	0.00	0.00
type_Dine-out	-0.04	0.00	0.00
type_Drinks & Nighlife	0.02	0.00	0.00
type_Pubs and bars	0.04	0.00	0.00
bakery	-0.02	0.00	0.01
bar	0.14	0.00	0.01
beverageshop	0.11	0.00	0.00
bhojanalya	-0.68	0.00	0.00
cafe	0.38	0.02	0.02
casualdining	0.10	0.01	0.01
club	0.08	0.00	0.00
confectionery	-0.20	0.00	0.00
delivery	-0.04	0.01	0.01
dessertparlor	0.53	0.02	0.02
dhaba	-0.63	0.00	0.00
finedining	0.71	0.00	0.00
foodcourt	-0.22	0.00	0.00
foodtruck	-0.04	0.00	0.00
iranicafee	-0.16	0.00	0.00
kiosk	0.09	0.00	0.00
lounge	0.08	0.00	0.00
meatshop	0.57	0.00	0.00
mess	0.11	0.00	0.00
microbrewery	0.08	0.00	0.00
pub	0.11	0.00	0.00
quickbites	-0.08	0.01	0.01
sweetshop	0.05	0.00	0.00
takeaway	-0.13	0.00	0.00

Table contains Feature Importance scores for each Models



Heat Map of Co-relation

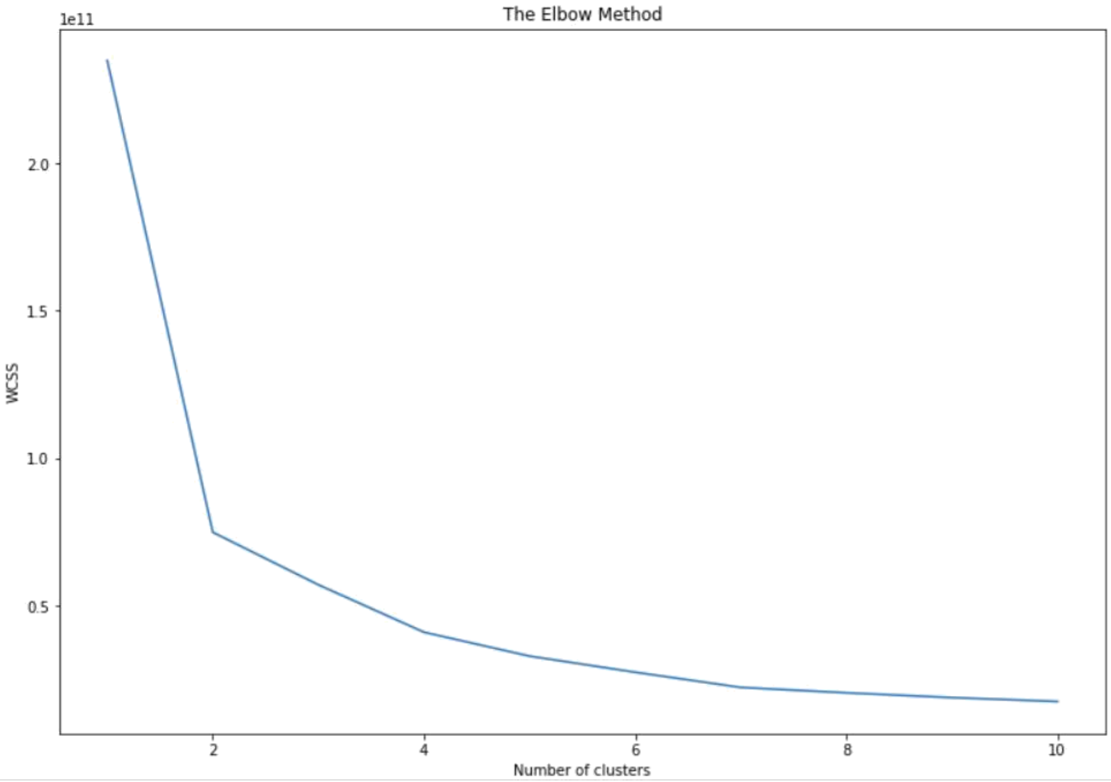
- **Goal 2**: Identify and cluster restaurants based on common attributes and see trends
  - Using Elbow curve, we will try identifying how many clusters can we split the model into.
  - Each cluster will help understand how restaurants relate to each other and what are the common characteristics among them
- **Goal 3**: Predicting new restaurant rating depending on input parameters
  - As most columns are encoded to numerical categories or contain numerical variables like costs we can use models like Multiple Linear Regression, Decision tree, Random Forest to predict new data points. We can even try a new model called Extra tree regression to test how accurately we are able to predict ratings for new restaurants
  - We will be using R2 to see measure model accuracy and how close we are too a good fitting model
- **Goal 4**: Building recommendation system to suggest restaurant based on customer preference.
  - We will be implementing a content-based recommendation using the rating information. So we can recommend a list of restaurants based on the choices of a user

Implementation of Selected Models

From the previous Milestone we have implemented the models explored and selected for each our goals.

- **Goal 1**: Understand what factors impact Rating of a restaurant
  - Based on the mode summary in below table we see that Restaurant Votes are an important factor for customers while choosing restaurants.
  - Restaurant Location and Cuisine offered, and cost for two-person meal are very important deciding factors too
  - Online Order, Booking Table and Menu items were not very important deciding factors

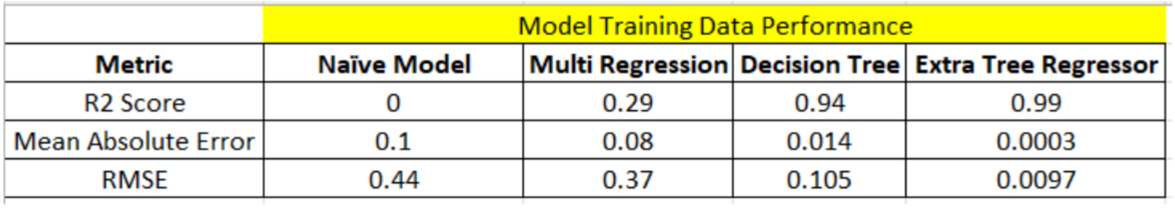
- **Goal 2**: Identify and cluster restaurants based on common attributes and see trends
  - Using Elbow curve, we found 3 clusters to be optimal.
  - These cluster can now be used to categorize new restaurant's and compare cluster wise trends. We can find average ratings, or common themes among the three clusters



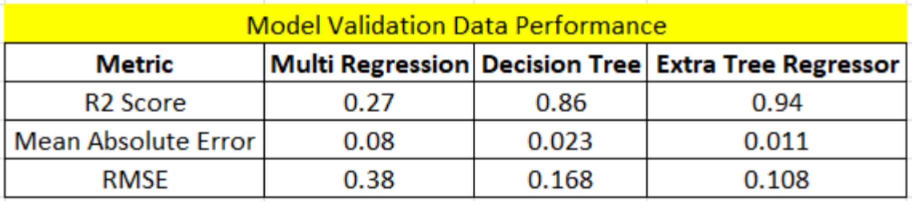
- **Goal 3**: Predicting new restaurant rating depending on input parameters

  - We start by analyzing understanding the Naïve Model (where predicted value is average rating for y_train dataset). As this is regression, we compute the R2 scores, Mean Absolute Error and the RMSE scores

  - We then compute the Performance Metrics at the Training and Validation dataset to evaluate performance of each model. It is expected that the training accuracy will be higher than the validation accuracy. But the accuracy scores(R2 Scores) should be close. If the training accuracy is very high and validation accuracy scores is low, this would mean overfitting. On the other hand, if the training and validation accuracy scores(R2 Scores) are low this would mean that the model is underfit and not complex enough. We will have to add new predictors



	Model Training Data Performance			
Metric	**Naïve Model**	**Multi Regression**	**Decision Tree**	**Extra Tree Regressor**
R2 Score	0	0.29	0.94	0.99
Mean Absolute Error	0.1	0.08	0.014	0.0003
RMSE	0.44	0.37	0.105	0.0097



Model Validation Data Performance			
Metric	**Multi Regression**	**Decision Tree**	**Extra Tree Regressor**
R2 Score	0.27	0.86	0.94
Mean Absolute Error	0.08	0.023	0.011
RMSE	0.38	0.168	0.108

  - From the training data performance, we see that the naïve model has a very poor R2 score when rating average is used as y_pred for the naïve model. This is a very poor model. We also compute the Mean absolute error and RMSE scores. Our other models should perform better on all three metrics (R2 Score, MAE, RMSE).

  - In the Training model, we start with Multi Regression model which has a very poor R2 Score. Closer the R2 Score to 1 the more accurate the model. This model has a high RMSE and MAE score as well. The model does not overfit as the training and validation R2 score and other metrics are close to each other. But overall looks like the Multi Regression model is underfit and is a very poor model to predict rating for new restaurants.

  - Next, we see Decision tree Model Performance, Training R2 Score is much higher 0.94 compared to previous Multi Linear regression model. The test R2 Score is 0.86, this means that mode is not overfit or underfit and is a good rating predictor. We also see that for MAE and RMSE scores are low compared to the Multi Linear Regression model. Lower the MAE and RMSE score more accurate will be Model in predicting rating for the new dataset.

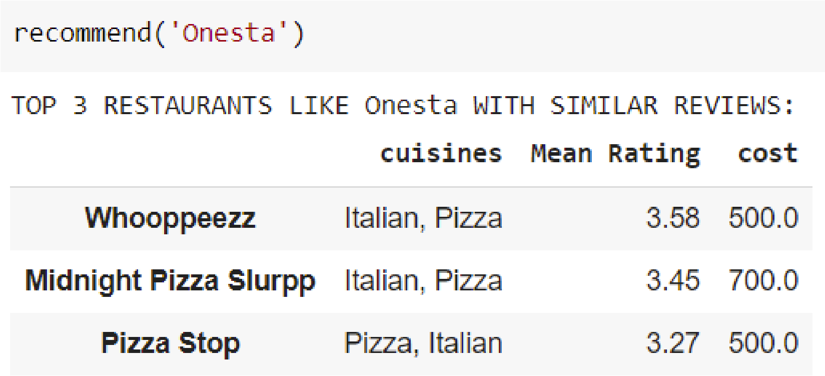
  - Finally, we use Extra Tree Regressor to evaluate the predictions. We see that this is the best model for the current dataset. It has the highest R2 score 0.99(almost a perfect predictor) for the Training data and high R2 Score 0.94 on the validation dataset too. So the predictions are very accurate. We also see that MAE and RMSE for both the train and test dataset are very low compared to the metrics for the other model. This proves that Extra tree regressor is the best model for our dataset and will give the most accurate new restaurant ratings predictions.

- Hyper Parameter tuning was not required as we already had high R2 scores and Restaurant rating need not be very precise small variations are fine. So the current base model optimizations are sufficient and so hyperparameter tuning was not done.

- **Goal 4**: Building recommendation system to suggest restaurant based on customer preference

  - After creation of the dataset in the required format we chose Term Frequency-Inverse Document Frequency while creating the Recommendation model vectors. The matrix is created after text cleaning the reviews list column. The model then uses cosine similarity to find similar restaurants and showcases the cuisines offered, Mean rating and cost for dining of the recommend restaurants.

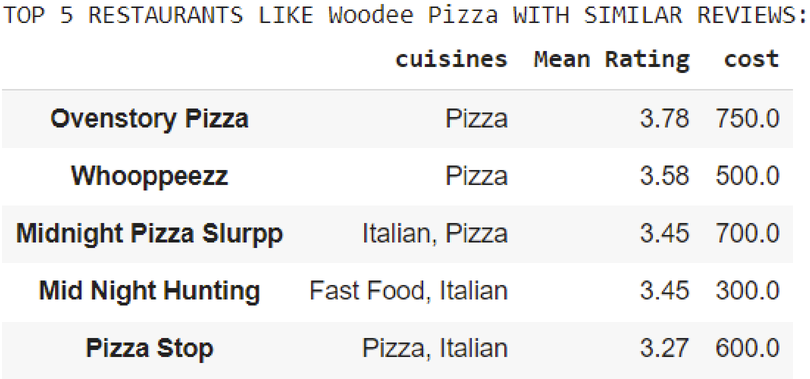
  - Below is an example of finding restaurants similar to "Onesta" a Pizza chain in Bangalore, India. To measure the performance, we can check the average cosine similarity score for the predicted restaurants and the original restaurant. We see based on this that predicted recommendations are relevant. Restaurants similar to Onesta which similarly offer Pizza, Italian cuisines, have comparable ratings, and costs are shown below



```
recommend('Onesta')
```

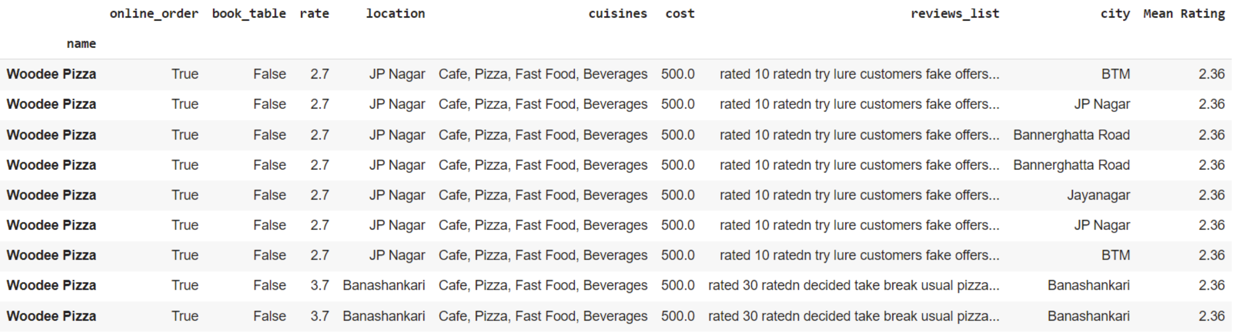
```
TOP 3 RESTAURANTS LIKE Onesta WITH SIMILAR REVIEWS:
```

	cuisines	Mean Rating	cost
Whooppeezz	Italian, Pizza	3.58	500.0
Midnight Pizza Slurpp	Italian, Pizza	3.45	700.0
Pizza Stop	Pizza, Italian	3.27	500.0



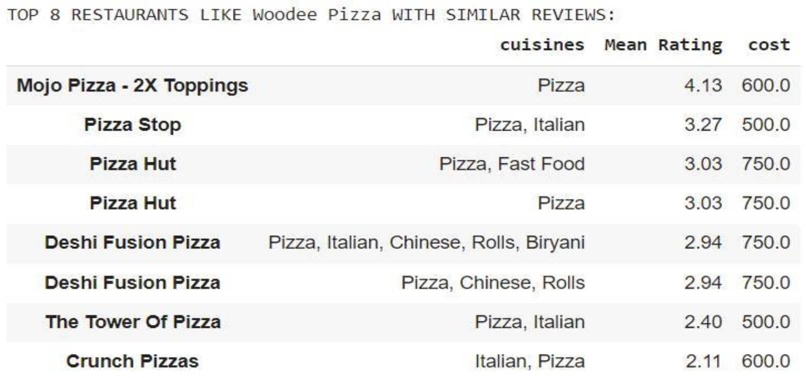
```
TOP 5 RESTAURANTS LIKE Woodee Pizza WITH SIMILAR REVIEWS:
```

	cuisines	Mean Rating	cost
Ovenstory Pizza	Pizza	3.78	750.0
Whooppeezz	Pizza	3.58	500.0
Midnight Pizza Slurpp	Italian, Pizza	3.45	700.0
Mid Night Hunting	Fast Food, Italian	3.45	300.0
Pizza Stop	Pizza, Italian	3.27	600.0



	online_order	book_table	rate	location	cuisines	cost	reviews_list	city	Mean Rating
name									
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	BTM	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	JP Nagar	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	Bannerghatta Road	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	Bannerghatta Road	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	Jayanagar	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	JP Nagar	2.36
Woodee Pizza	True	False	2.7	JP Nagar	Cafe, Pizza, Fast Food, Beverages	500.0	rated 10 ratedn try lure customers fake offers...	BTM	2.36
Woodee Pizza	True	False	3.7	Banashankari	Cafe, Pizza, Fast Food, Beverages	500.0	rated 30 ratedn decided take break usual pizza...	Banashankari	2.36
Woodee Pizza	True	False	3.7	Banashankari	Cafe, Pizza, Fast Food, Beverages	500.0	rated 30 ratedn decided take break usual pizza...	Banashankari	2.36

- Like above we check the recommendation for a low rated niche Pizza chain in Bangalore Woodee Pizza. As we see similarly rate Pizza chains are recommended. A tradeoff is done between Mean rating, cost in this recommendation. If a restaurant has low rating, we would also want the cost to be slightly lower, to ensure a similarity. Cost also reduces with increase in Mean rating. Example for Mojo Pizza as the Mean rating is high the cost is low and this results in a high similarity score between Mojo Pizza and Woodee Pizza and is the top recommendation.



TOP 8 RESTAURANTS LIKE Woodee Pizza WITH SIMILAR REVIEWS:

	cuisines	Mean Rating	cost
Mojo Pizza - 2X Toppings	Pizza	4.13	600.0
Pizza Stop	Pizza, Italian	3.27	500.0
Pizza Hut	Pizza, Fast Food	3.03	750.0
Pizza Hut	Pizza	3.03	750.0
Deshi Fusion Pizza	Pizza, Italian, Chinese, Rolls, Biryani	2.94	750.0
Deshi Fusion Pizza	Pizza, Chinese, Rolls	2.94	750.0
The Tower Of Pizza	Pizza, Italian	2.40	500.0
Crunch Pizzas	Italian, Pizza	2.11	600.0