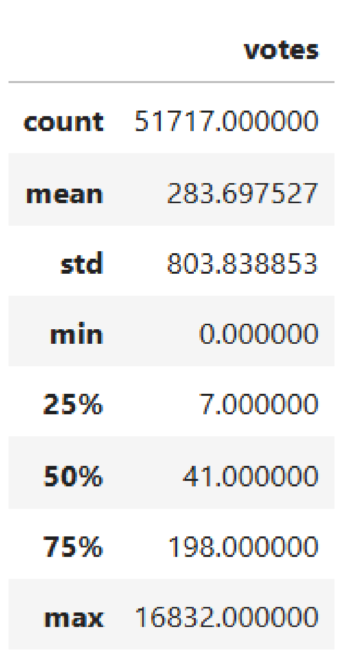
Data Collection and Preprocessing Phase

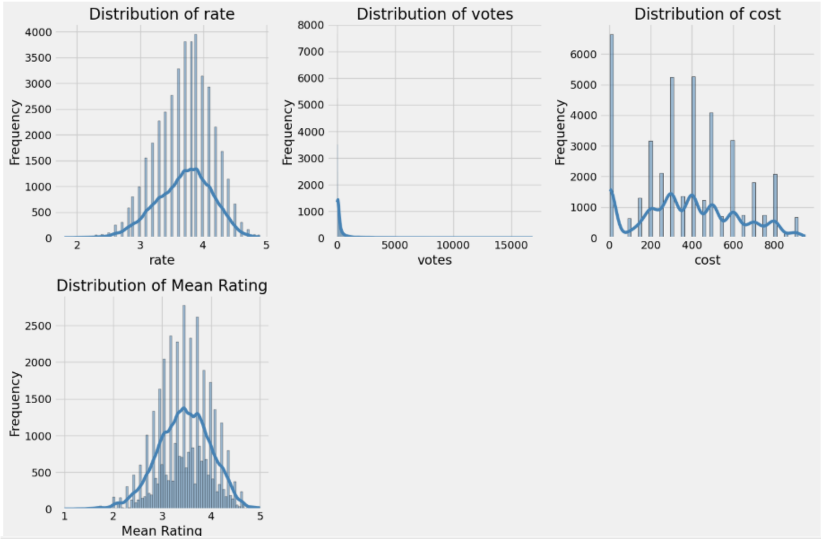
Date	18 June 2025
Team ID	xxxxxx
Project Title	Restaurant Recommendation System
Maximum Marks	6 Marks

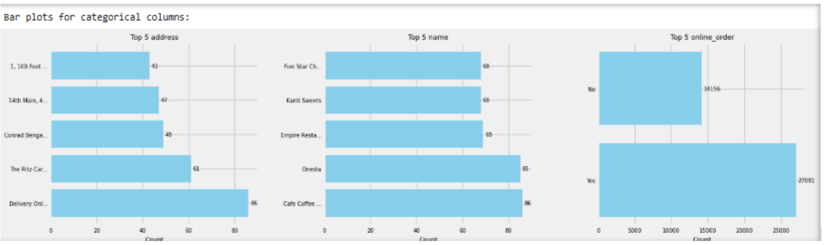
Data Exploration and Preprocessing Report

Data set variables will be statistically analyzed to identify patterns and outliers, with Python Employed for pre-processing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

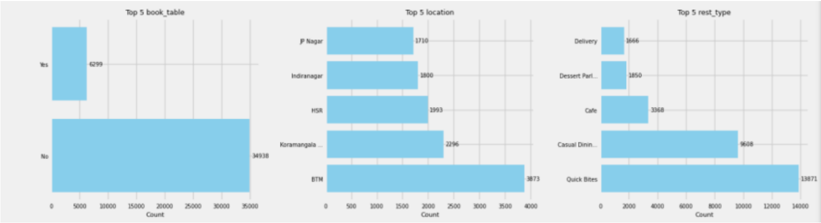
Section	Description
Data Overview	**Dimensions:** 51717 rows x 17 columns **Descriptive statistics:**  votes count 51717.000000 mean 283.697527 std 803.838853 min 0.000000 25% 7.000000 50% 41.000000 75% 198.000000 max 16832.000000

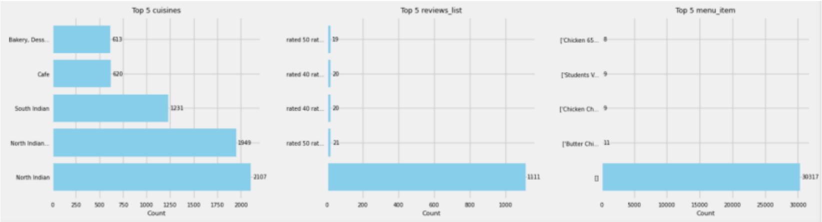
Univariate Analysis

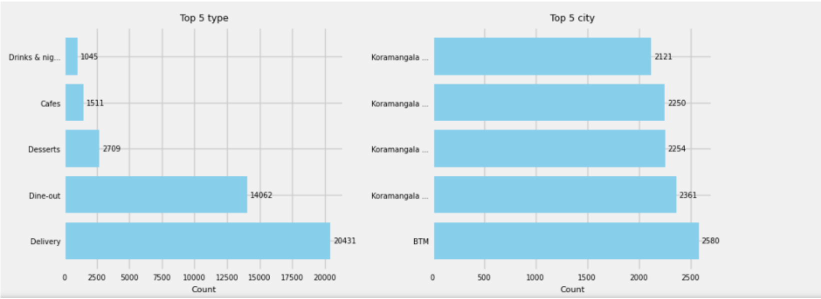




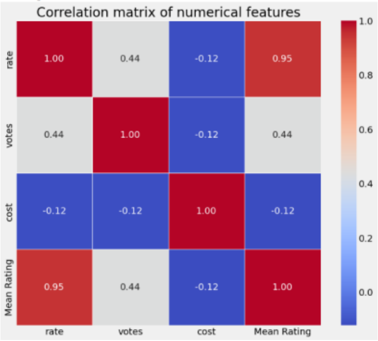
```
Bar plots for categorical columns:
```

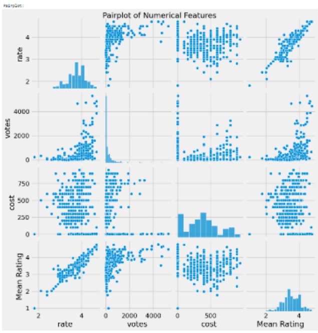




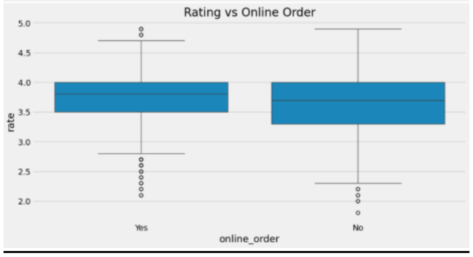


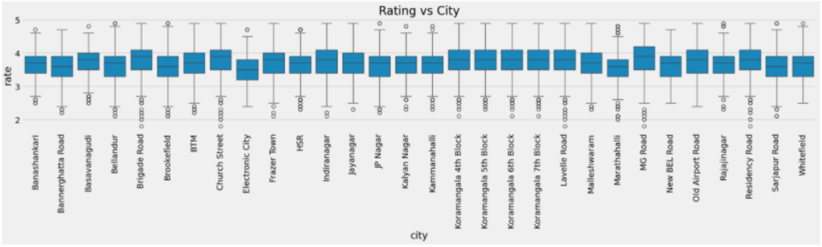
Bivariate Analysis



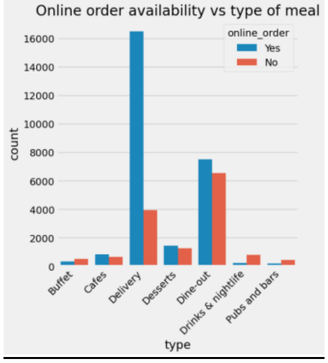


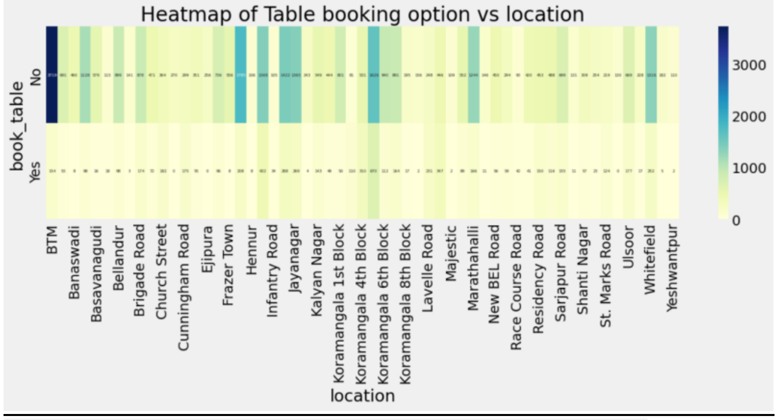
Numerical vs Categorical Analysis:

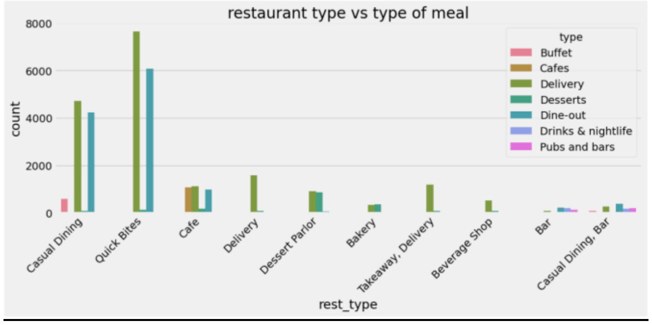




Categorical vs Categorical Analysis:

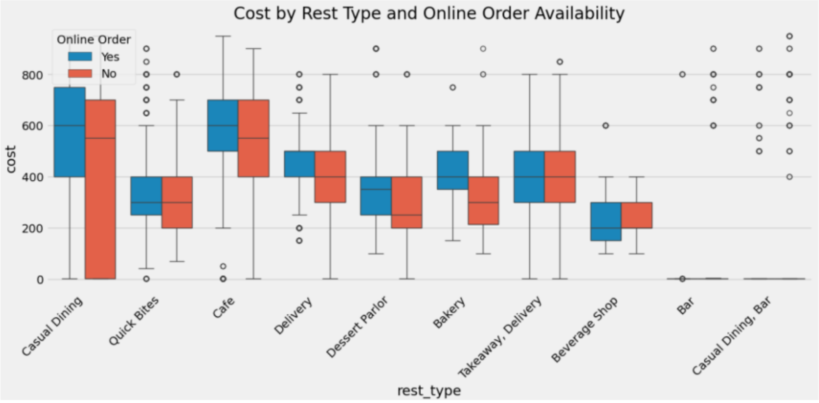




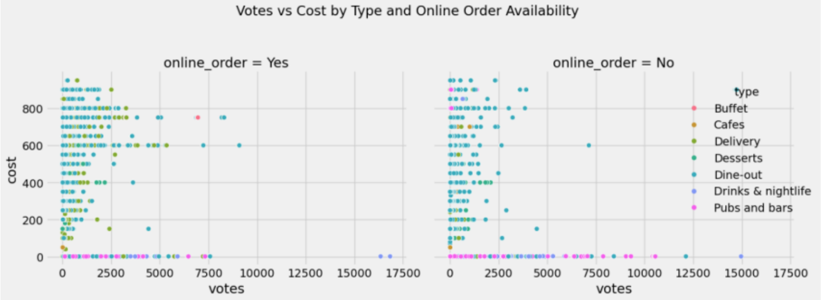


Multivariate Analysis

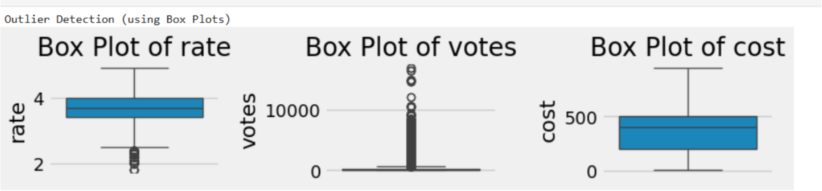
Cost by Rest Type, broken down by Online Order:



Votes vs Cost, separated by Online Order and colored by Type:

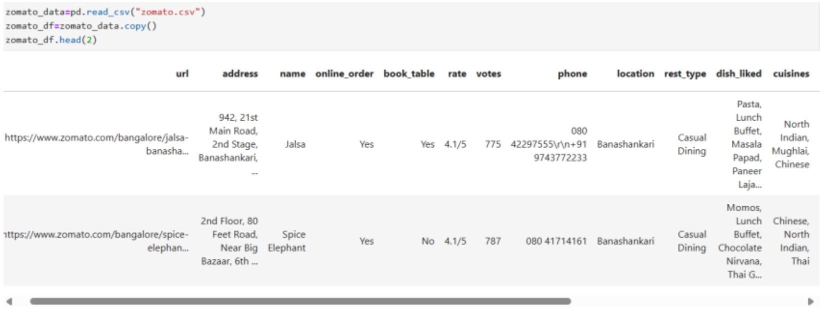


Outliers and Anomalies


Outlier Detection (using Box Plots)

Data Preprocessing Code Screenshots

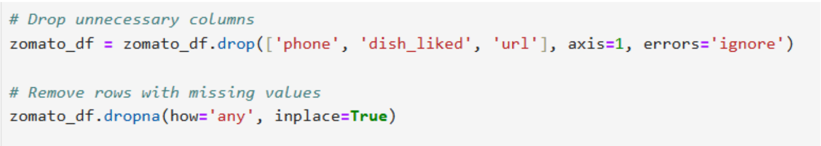
Loading Data



```
zomato_data=pd.read_csv("zomato.csv")
zomato_df=zomato_data.copy()
zomato_df.head(2)
```

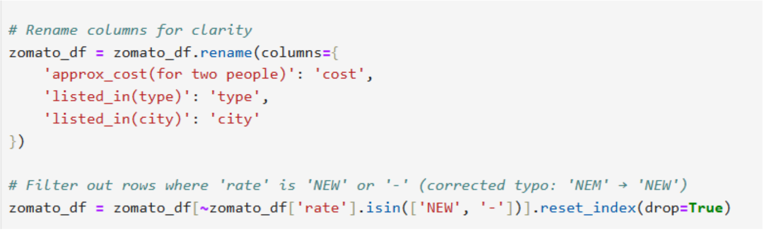
	url	address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines
	https://www.zomato.com/bangalore/jalsa-banasha...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese
	ittps://www.zomato.com/bangalore/spice-elephan...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai

Handling Missing Data

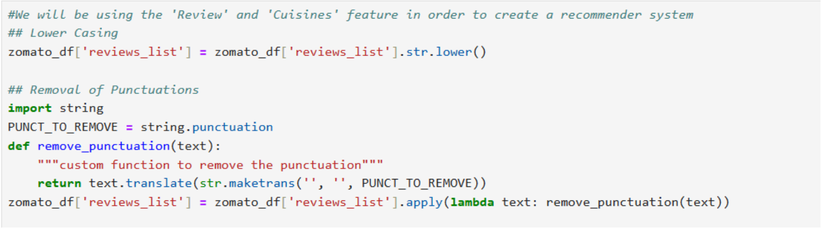


```
# Drop unnecessary columns
zomato_df = zomato_df.drop(['phone', 'dish_liked', 'url'], axis=1, errors='ignore')

# Remove rows with missing values
zomato_df.dropna(how='any', inplace=True)
```

Data Transformation

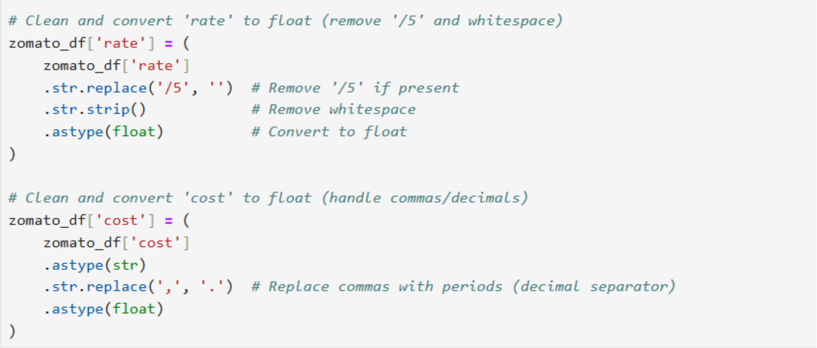


```
# Rename columns for clarity
zomato_df = zomato_df.rename(columns={
    'approx_cost(for two people)': 'cost',
    'listed_in(type)': 'type',
    'listed_in(city)': 'city'
})

# Filter out rows where 'rate' is 'NEW' or '-' (corrected typo: 'NEM' → 'NEW')
zomato_df = zomato_df[~zomato_df['rate'].isin(['NEW', '-'])].reset_index(drop=True)
```



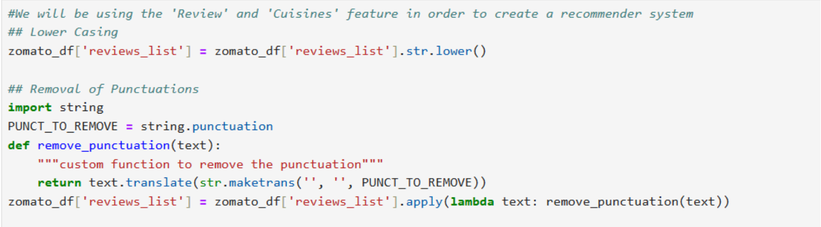
```
#We will be using the 'Review' and 'Cuisines' feature in order to create a recommender system
## Lower Casing
zomato_df['reviews_list'] = zomato_df['reviews_list'].str.lower()

## Removal of Punctuations
import string
PUNCT_TO_REMOVE = string.punctuation
def remove_punctuation(text):
    """custom function to remove the punctuation"""
    return text.translate(str.maketrans('', '', PUNCT_TO_REMOVE))
zomato_df['reviews_list'] = zomato_df['reviews_list'].apply(lambda text: remove_punctuation(text))
```

<table>
<tr>
<td>Feature Engineering</td>
<td>



```python
# Clean and convert 'rate' to float (remove '/5' and whitespace)
zomato_df['rate'] = (
    zomato_df['rate']
    .str.replace('/5', '')  # Remove '/5' if present
    .str.strip()            # Remove whitespace
    .astype(float)          # Convert to float
)

# Clean and convert 'cost' to float (handle commas/decimals)
zomato_df['cost'] = (
    zomato_df['cost']
    .astype(str)
    .str.replace(',', '.')  # Replace commas with periods (decimal separator)
    .astype(float)
)
```

prepares **reviews_list** to be used as a textual feature for TF-IDF vectorization



```python
#We will be using the 'Review' and 'Cuisines' feature in order to create a recommender system
## Lower Casing
zomato_df['reviews_list'] = zomato_df['reviews_list'].str.lower()

## Removal of Punctuations
import string
PUNCT_TO_REMOVE = string.punctuation
def remove_punctuation(text):
    """custom function to remove the punctuation"""
    return text.translate(str.maketrans('', '', PUNCT_TO_REMOVE))
zomato_df['reviews_list'] = zomato_df['reviews_list'].apply(lambda text: remove_punctuation(text))
```

</td>
</tr>
<tr>
<td>Save Processed Data</td>
<td>-</td>
</tr>
</table>