CAPSTONE PROJECT

Trip Advisor- Restaurant Improvement And New Restaurant Advice

GROUP -2 Final Report

MENTORED BY: ANKUSH BANSAL

SUBMITTED BY : Melvin Jharge SR
Mohammed Jaffer Atuf
Rashul
Sujith M
Yamini E

Table of Contents:

CHAPTER 1

Introduction to the project

1.1 Background

Tripadvisor, Inc. is an American online travel company that runs a comparison shopping website, a mobile app, and websites with user-generated content. Additionally, it offers restaurant reservations, hotel reservations, and bookings for trip activities, housing, and transportation online. Needham, Massachusetts serves as its corporate headquarters.

1.2 Problem Statement

Restaurant Rating has become the most commonly used parameter for judging a restaurant for any individual. A lot of research has been done on different restaurants and the quality of food it serves. Rating of a restaurant depends on factors like reviews, area situated, average cost for two people, votes, cuisines and the type of restaurant.The reviews of the restaurants are what either makes or breaks it. Based on positive reviews restaurants expect a higher footfall.

1.3 Impact on business

The main objective of this project is two fold. First is to understand the reason why already existing restaurants are being rated poorly and provide insights to increase their rating. The second objective of the project is to advise new entrants in the restaurant field with the ideal elements that will make their business successful.

Dataset and Domain

2.1 Dataset

The dataset includes all European restaurants available on TripAdvisor - more than a million records. Given that the dataset includes all records and there was no initial filter applied during the scraping script, there are several fields that have most of the values blank.

Restaurant ratings are rounded on TripAdvisor to the nearest 0.5/5 value.This means that the distribution of ratings for all restaurants have gaps of at least 0.5 and the level of accuracy of ratings is not particularly satisfying for a deep analysis on the restaurant ratings.

From the dataset, "Italy" has been chosen as the subset. Italy is a famous tourist destination. Italy is renowned not just for its magnificent landscapes and vibrant culture but also for its exquisite food, which is ideal for everyone. One of the primary reasons that such a large number of people travel to Italy is for the cuisine.

Traveling in late spring and early autumn allows one to avoid the inflated prices and overbooked hotels, while an extensive rail network, conveniently connecting all the major cities of the country, prides itself on great service and reasonable fares.

Data Source:[TripAdvisor European restaurants | Kaggleaset]

2.2. Data Dictionary

- **Restaurant_link -** TripAdvisor restaurant link
- **Restaurant_name -** Name of the restaurant on TripAdvisor
- **Location -** Location displayed on TripAdvisor

- **Country -** Country name retrieved from original_location
- **Region -** Region name retrieved from original_location
- **Province -** Province name retrieved from original_location
- **City -** City name retrieved from original_location
- **Address -** Address mentioned on Trip Advisor
- **Latitude -** Latitude coordinate
- **Longitude -** Longitude coordinate
- **Claimed -** Restaurant business claimed on TripAdvisor
- **Awards -** Awards received by the Hotel
- **Popularity_detailed -** Popularity detailed ranking
- **Popularity_generic -** Popularity generic ranking (among all places to eat in the area)
- **Top_tags -** Top tag names
- **Price_level -** Level of prices in current currency (data scraped from www.tripadvisor.ie - so currency is €)
- **Price_range -** Range of prices in current currency (ranges displayed on TripAdvisor seem to be quite unreliable)
- **Meals** - Types of meal
- **Cuisines -** Type of cuisine
- **Special_diet -** Types of special diets
- **Features -** Restaurant features
- **Vegetarian_friendly -** Is the restaurant vegetarian friendly?(Y or N)
- **Vegan_options -** Does the restaurant offer vegan options?(Y or N)
- **Gluten_free -** Does the restaurant have gluten-free options?(Y or N)
- **Original_open_hours -** Original open hours on TripAdvisor
- **Open_days_per_week -** Number of days open per week retrieved from original_open_hours
- **Open_hours_per_week -** Number of open hours per week retrieved from original_open_hours
- **Working_shifts_per_week -** Number of working shifts per week retrieved from original_open_hours

- **Avg_rating -** Average restaurant rating
- **Total_reviews_count -** The total reviews count
- **Default_language -** The default language displayed while scraping
- **Reviews_count_in_default_language -** Total reviews count in default language
- **Excellent -** Excellent reviews count in default language
- **Very_good -** Very_good reviews count in default language
- **Average -** Average reviews count in default language
- **Poor -** Poor reviews count in default language
- **Terrible -** Terrible reviews count in default language
- **Food -** Food rating
- **Service -** Service rating
- **Value -** Value rating
- **Atmosphere -** Atmosphere rating
- **Keywords -** Popular Keywords

Variable categorization (count of numeric and categorical) & null value percentage

Column Name	Datatype	Non-Null Values	Null Values (%)
Restaurant_link	Int64	63758	0
Restaurant_name	object	63758	0
Original_location	object	63758	0
Country	object	63758	0
Region	object	63758	0
Province	object	56771	10.95
City	object	63758	0
Address	object	63758	0
Latitude	float64	63499	0.40
Claimed	object	63758	0
Awards	object	26781	57.99
Popularity_detailed	object	60800	4.63
Popularity_generic	object	60692	4.80
top_tags	object	63748	0.01
Price_level	object	63634	0.19
Price_range	object	63758	0
Meals	object	24690	61.27
Cuisines	object	62443	2.06
Special_diets	object	31417	50.72

Features	object	3701	94.19

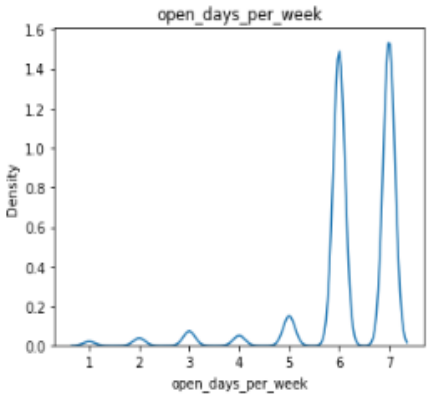
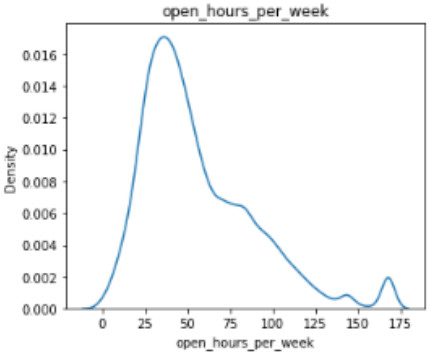
Vegetarian_friendly	object	63758	0
Vegan_options	object	63758	0
Gluten_free	object	63758	0
Original_open_hours	object	48665	23.67
Open_days_per_week	float64	48665	23.67
Open_hours_per_week	float64	48665	23.67
working_shifts_per_week	float64	48665	23.67
avg_rating	float64	60687	4.81
total_reviews_count	float64	62484	1.99
default_language	object	60818	4.61
reviews_count_in_default_language	float64	60818	4.61
Excellent	float64	60818	4.61
Very_good	float64	60818	4.61
Average	float64	60818	4.61
Poor	float64	60818	4.61
Terrible	float64	60818	4.61
Food	float64	43722	31.42
Service	float64	43801	31.30
Value	float64	43734	31.40

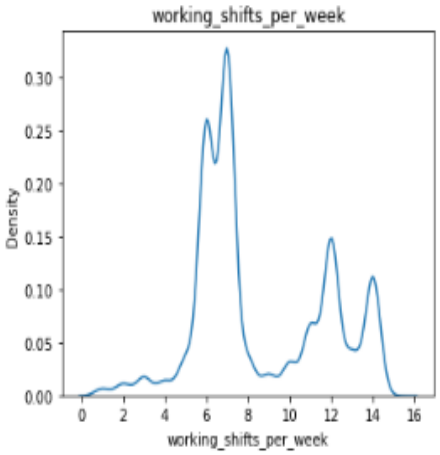
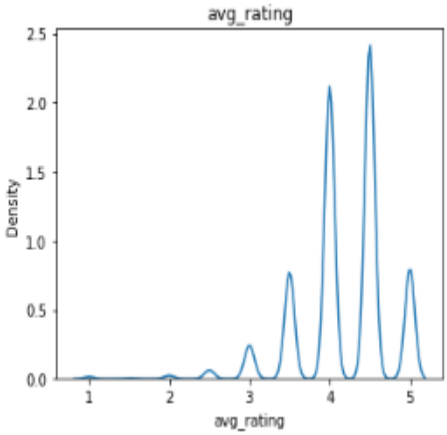
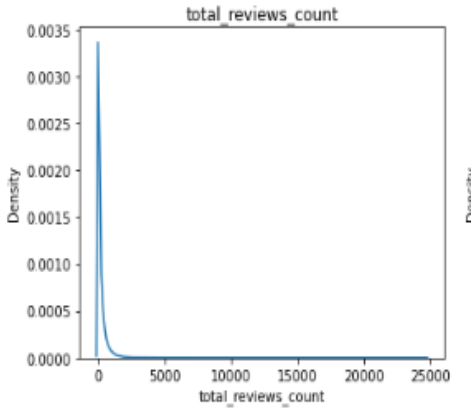
Longitude	float64	63499	0.40

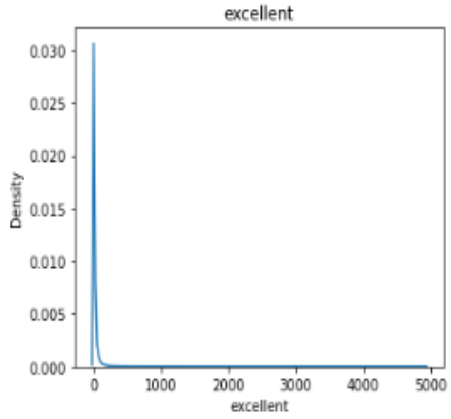
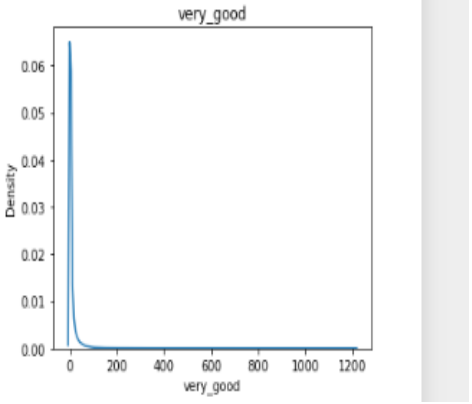
Atmosphere	float64	22636	64.49
Keywords	object	2827	95.56

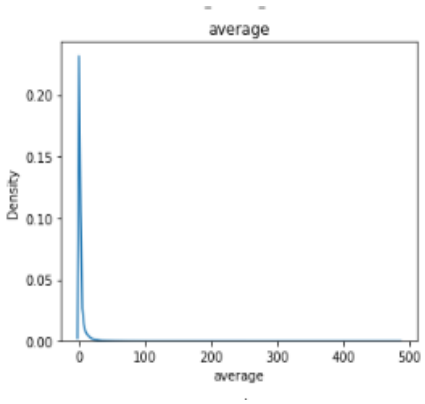
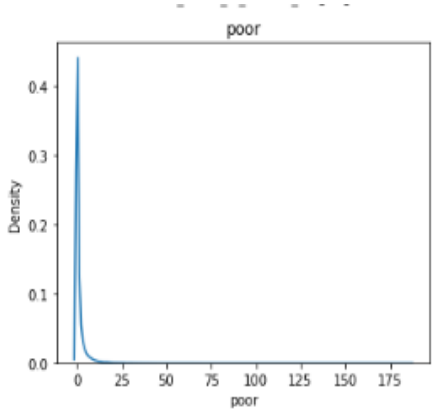
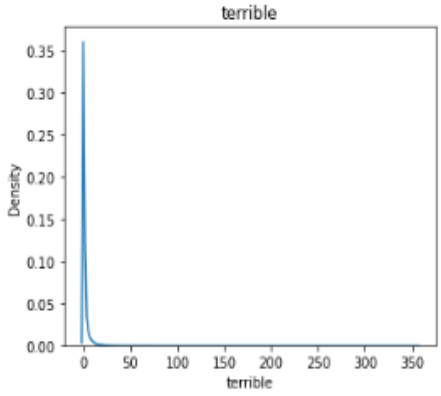
Univariate Analysis:

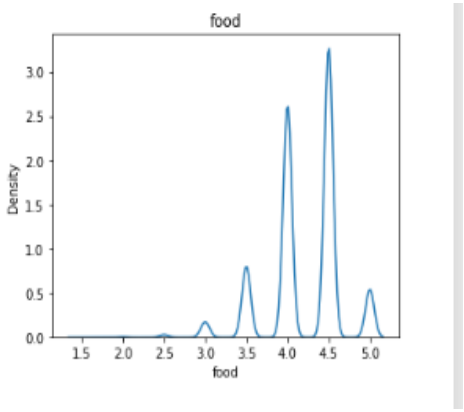
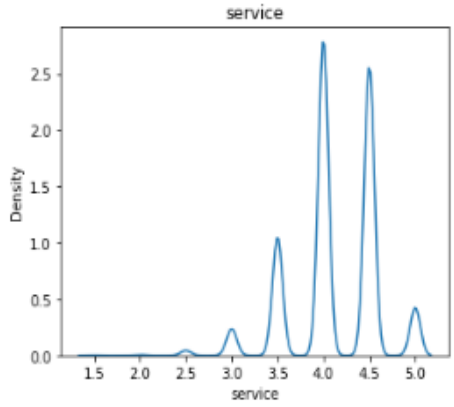
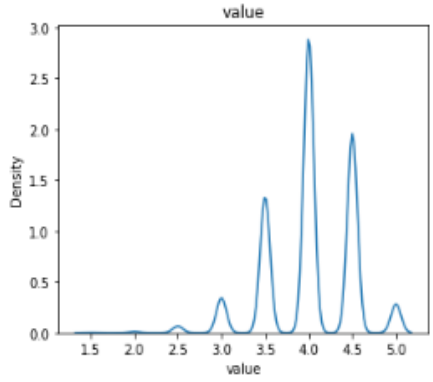
NUMERICAL COLUMNS:

COLUMN NAME	GRAPH	INFERENCE
open_days_per_week		• Open days per week - Most restaurants are open for 7 days or 6 days a week. • There are also restaurants which are open for only 1 day.
open_hours_per_ week		• Open hours per week on average most restaurants are open for 40 hours a week. • There are restaurants which are open 24 hours for entire week .

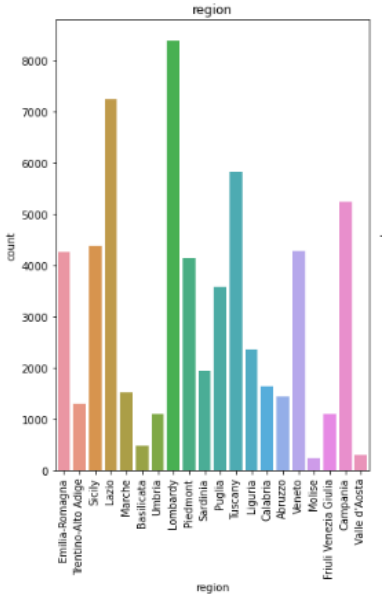
Working_shifts_per_ week		• Mostly there are 8 shifts per week. Every shift is for 3 hours to 4 hours .
avg_rating		• avg rating - The rating ranges from 0.5-5. The highest value is 4.5.
total_reviews_count		• total_review_counts is positively skewed.

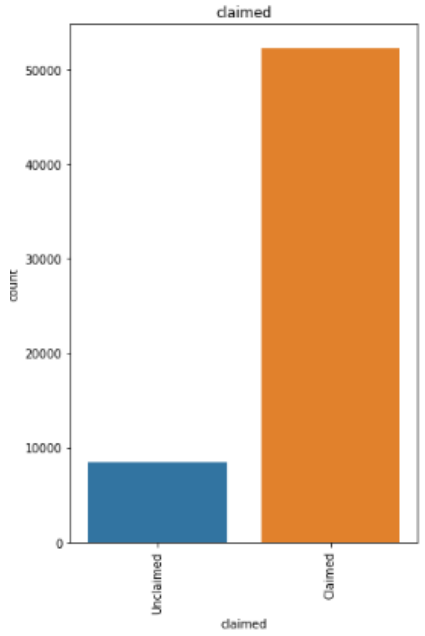
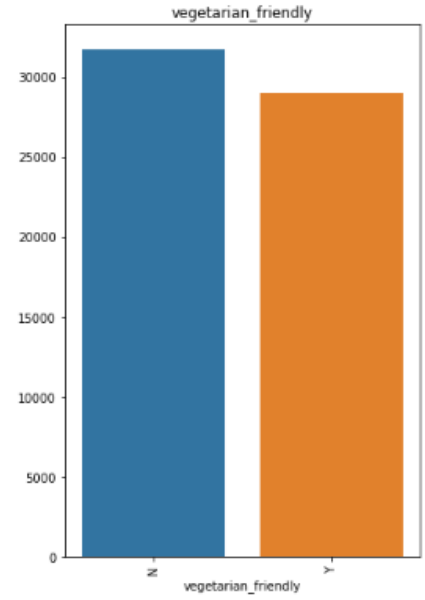
Reviews_count_in_ default_language		• Reviews_count_in _default_language is positively skewed.
excellent		• Excellent column is positively skewed.
very_good		• Very_good column is positively skewed.

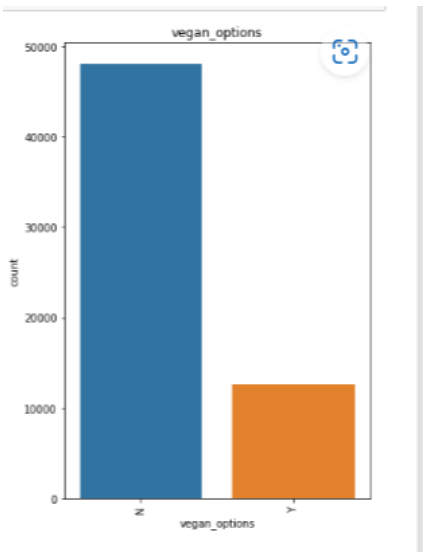
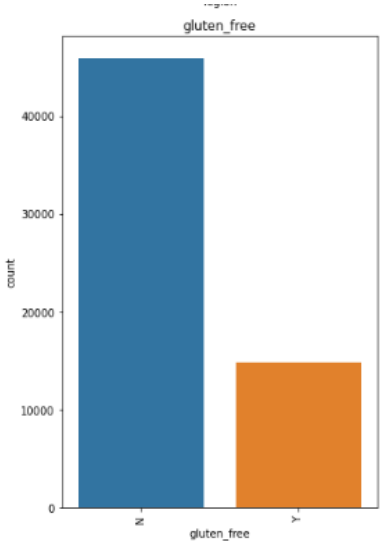
average		• Average column is positively skewed.
poor		• Poor column is positively skewed.
terrible		• Terrible column is positively skewed.

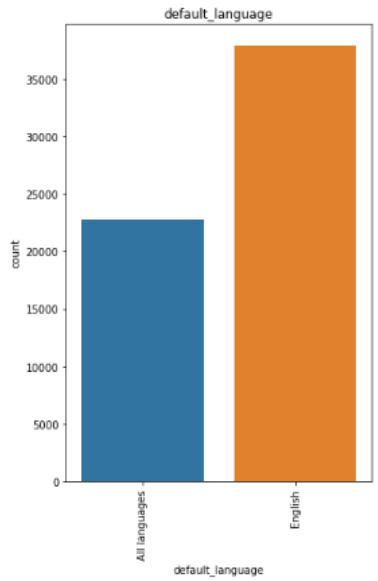
food		• Food has 4.5 as its highest rating .
service		• Service has highest rating as 4.0.
value		• Value has highest rating as 4.0 .

CATEGORICAL COLUMNS:

COLUMN NAME	GRAPH	INFERENCE
region		• There are regions which are highly populated and those which have low population. • Lombardy has the highest population • Most of the restaurants are in Lombardy region

claimed		• Most of the restaurants are claimed on Tripadvisor (verified or not verified)
Vegetarian_friendly		• The vegetarian friendly options are available but they are less than non-veg options

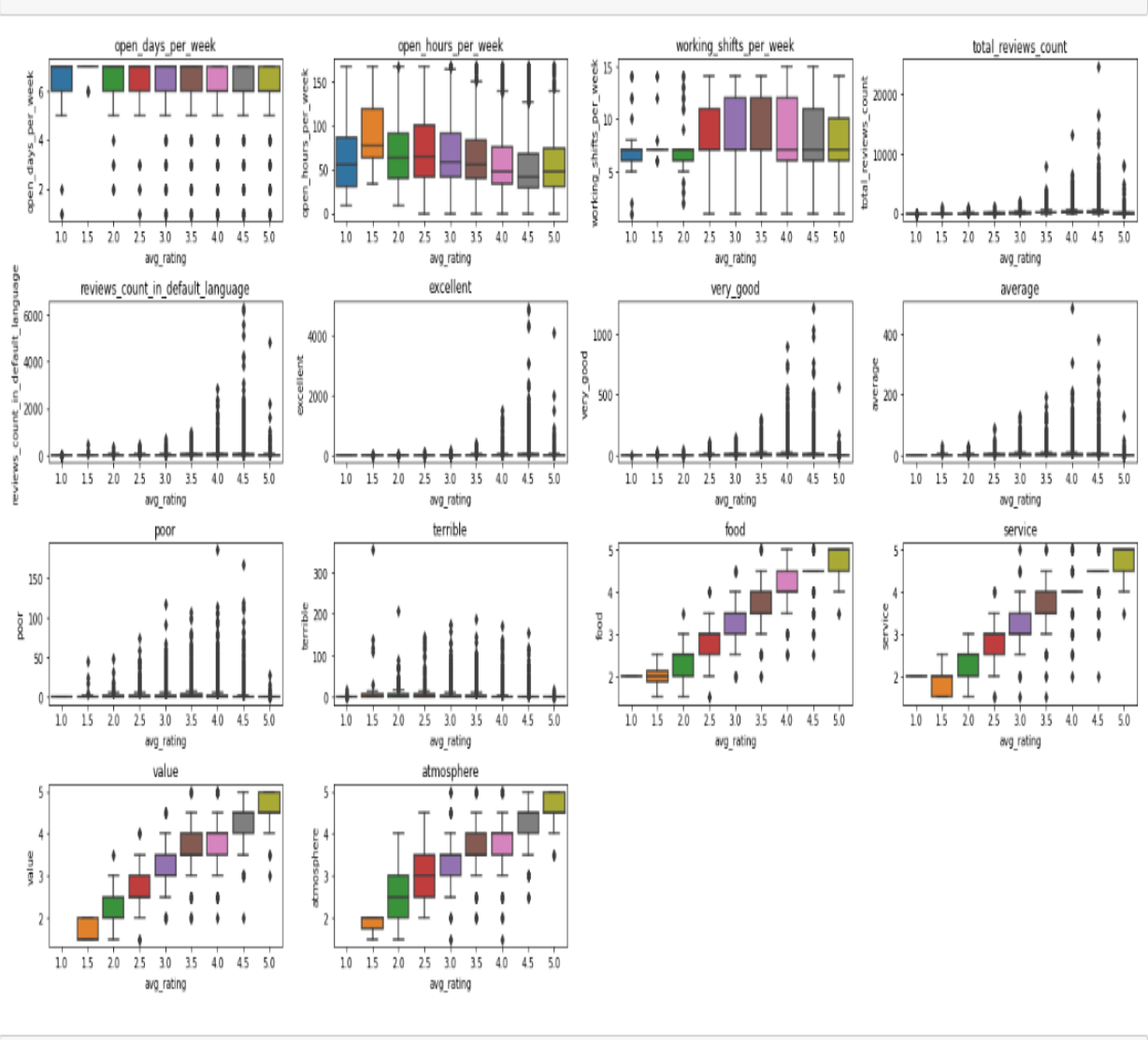
vegan_options		• Very limited vegan options are available
gluten_free		• Gluten free options are very low

default_languages		• English is the default language in most restaurants

- Numerous distinct values are contained in the columns restaurant link, restaurant name, original location, country, province, city, address, awards, popularity generic, popularity detailed, top tags, price level, price range, meals, cuisines, special diets, features, and keywords. Therefore, after univariate analysis, we are unable to draw any inferences from these columns. So we don't show the results of the univariate analysis.

Bivariate Analysis:

NUMERICAL VS TARGET:



- Since these are all predictors for the average rating we cannot remove the outliers hence we are keeping the outliers in the data.

Feature Engineering:

<table>
<tr><th>BEFORE FEATURE ENGINEERING</th><th>AFTER FEATURE ENGINEERING</th><th>INFERENCE</th></tr>
<tr>
<td>
<table>
<tr><th>province</th><th>city</th><th>address</th></tr>
<tr><td>Province of Bologna</td><td>NaN</td><td>Piazza Liberta' 14, 40060 Toscanella, Dozza Italy</td></tr>
<tr><td>Province of Bologna</td><td>NaN</td><td>Via Allende Parco Ruggi, 40060 Toscanella, Doz...</td></tr>
<tr><td>Province of Bologna</td><td>NaN</td><td>Piazza Antonio Gramsci 2, 40060 Toscanella, Do...</td></tr>
</table>
</td>
<td>
<table>
<tr><th>province</th><th>city</th><th>address</th></tr>
<tr><td>Province of Bologna</td><td>Toscanella</td><td>Piazza Liberta' 14, 40060 Toscanella, Dozza Italy</td></tr>
<tr><td>Province of Bologna</td><td>Toscanella</td><td>Via Allende Parco Ruggi, 40060 Toscanella, Doz...</td></tr>
<tr><td>Province of Bologna</td><td>Toscanella</td><td>Piazza Antonio Gramsci 2, 40060 Toscanella, Do...</td></tr>
</table>
</td>
<td>

The city column has more null values.So we are fetching the city from the original location column using split function.

</td>
</tr>
<tr>
<td>
<table>
<tr><th>price_level</th><th>price_range</th><th>meals</th></tr>
<tr><td>€</td><td>€5-€10</td><td>NaN</td></tr>
<tr><td>€€-€€€</td><td>€4-€16</td><td>NaN</td></tr>
<tr><td>€€-€€€</td><td>€12-€22</td><td>Lunch</td></tr>
</table>
</td>
<td>
<table>
<tr><th>price_level</th><th>price_range</th><th>meals</th></tr>
<tr><td>€</td><td>5,10</td><td>NaN</td></tr>
<tr><td>€€-€€€</td><td>4,16</td><td>NaN</td></tr>
<tr><td>€€-€€€</td><td>12,22</td><td>Lunch</td></tr>
</table>
</td>
<td>

Here the price range column has Euro and hyphen symbols. So We are removing those symbols using the function replace.

</td>
</tr>
</table>

latitude	longitude	claimed	awards
45.964397	12.434187	Claimed	Travellers' Choice, Certificate of Excellence ...
45.662094	8.087583	Claimed	Certificate of Excellence 2019
45.200523	10.762072	Claimed	Travellers' Choice, Certificate of Excellence ...

min_price	max_price	avg_price	total_COE_awards	total_tc_awards
5	10	7.5	0	0
4	16	10.0	0	0
12	22	17.0	0	0
3	5	4.0	0	0
6	26	16.0	1	0

- We take the total number of awards that the specific restaurant has received from the information in the awards column and add them to the new column total_tc_award and total_coe_awards.

region	province	city	address
Emilia-Romagna	Province of Bologna	Toscanella	Piazza Liberta' 14, 40060 Toscanella, Dozza Italy
Emilia-Romagna	Province of Bologna	Toscanella	Via Allende Parco Ruggi, 40060 Toscanella, Doz...
Emilia-Romagna	Province of Bologna	Toscanella	Piazza Antonio Gramsci 2, 40060 Toscanella, Do...

region	province	city	address
Emilia-Romagna	Province of Bologna	tier2	Piazza Liberta' 14, 40060 Toscanella, Dozza Italy
Emilia-Romagna	Province of Bologna	tier2	Via Allende Parco Ruggi, 40060 Toscanella, Doz...
Emilia-Romagna	Province of Bologna	tier2	Piazza Antonio Gramsci 2, 40060 Toscanella, Do...

- We are segregating the city column into three tiers based on the population of the corresponding city.

province	city	address	latitude	longitude	claimed	
Province of Pordenone	NaN	Via Vittorio Veneto 5, 33070 Stevena, Caneva I...	45.964397	12.434187	Claimed	
Province of Biella	NaN	Localita Bielmonte SNC, 13835 Bielmonte, Piatt...	45.662094	8.087583	Claimed	
Province of Mantua	NaN	Strada Mantova 21 presso Corte Barco, 46045 Ma...	45.200523	10.762072	Claimed	

min_price	max_price	avg_price	total_COE_awards	total_tc_awards	zipcode
5	10	7.5	0	0	40060
4	16	10.0	0	0	40060
12	22	17.0	0	0	40060

- Extracting zipcode of the particular restaurant from the address column.

min_price	max_price	avg_price	total_COE_awards	total_tc_awards	zipcode
5	10	7.5	0	0	40060
4	16	10.0	0	0	40060
12	22	17.0	0	0	40060

total_COE_awards	total_tc_awards	zipcode	restaurant_count_zip_code_wise
0	0	40060	14
0	0	40060	14
0	0	40060	14

- We are creating the new column restaurant_count_zip_code_wise and adding the count of restaurants in the particular zip code from zipcode.

country	region	province	city	address
Italy	Emilia-Romagna	Province of Bologna	Toscanella	Piazza Liberta' 14, 40060 Toscanella, Dozza Italy
Italy	Emilia-Romagna	Province of Bologna	Toscanella	Via Allende Parco Ruggi, 40060 Toscanella, Doz...
Italy	Emilia-Romagna	Province of Bologna	Toscanella	Piazza Antonio Gramsci 2, 40060 Toscanella, Do...

country	region	province	city
Italy	6	Province of Bologna	tier2
Italy	6	Province of Bologna	tier2
Italy	6	Province of Bologna	tier2

- We are having 20 regions in the region column. We have ranked those 20 regions with respect to their population.

country	region	province	city
Italy	6	Province of Bologna	tier2
Italy	6	Province of Bologna	tier2
Italy	6	Province of Bologna	tier2

country	region	province	city
Italy	Highly_populated	Province of Bologna	tier2
Italy	Highly_populated	Province of Bologna	tier2
Italy	Highly_populated	Province of Bologna	tier2

- We are binning the region column into 3 bins based on the population of the region as “Highly populated, Averagely populated and Sparsely populated”.

popularity_detailed	popularity_generic	top_tags	p
#5 of 5 Restaurants in Toscanella	#6 of 7 places to eat in Toscanella	Cheap Eats, Italian, Pizza	
#1 of 5 Restaurants in Toscanella	#1 of 7 places to eat in Toscanella	Mid-range, Italian, Fast food, Vegetarian Frie...	
#3 of 5 Restaurants in Toscanella	#3 of 7 places to eat in Toscanella	Mid-range, Chinese, Japanese, Sushi	

popularity_detailed	popularity_generic	top_tags
1.0	1.0	Cheap Eats, Italian, Pizza
4.0	4.0	Mid-range, Italian, Fast food, Vegetarian Frie...
3.0	3.0	Mid-range, Chinese, Japanese, Sushi

- Extracting the rank and no.of restaurants in a particular city from Popularity_detailed and Popularity_generic .
- And also ranking the restaurants of the particular city based on the percentiles into 5 categories.

price_level	price_range	meals	cuisines
€	4,10	NaN	Italian, Pizza, Fast food
€€-€€€	5,25	Lunch, Dinner	Pizza, Pub, Italian
€€-€€€	18,25	Lunch	Bar, Seafood, Mediterranean

restaurant_count_zip_code_wise	lunch	Breakfast	Dinner	Brunch	Drinks
14	0	0	0	0	0
14	0	0	0	0	0
14	1	0	0	0	0

- Creating the individual features for the different types of meals present in each restaurant.

meals	cuisines	special_diets
NaN	Italian, Pizza, Fast food	Vegetarian Friendly, Vegan Options
Lunch, Dinner	Pizza, Pub, Italian	NaN
Lunch	Bar, Seafood, Mediterranean	NaN

Italian	Pizza	Mediterranean	Seafood	European	Barbecue	Fast_food	Bar
1	1	0	0	0	0	0	0
1	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0

- Creating different features for the different types of cuisines present in each restaurant

price_range	meals	cuisines	special_diets	features
5,10	unknown	Italian, Pizza	NaN	Reservations, Seating
4,16	unknown	Italian, Fast food	Vegetarian Friendly	NaN
12,22	Lunch	Chinese, Japanese, Sushi, Asian, Middle Eastern	NaN	NaN

Reservations	Wheelchair Accessible	Takeout	Seating
1	0	0	1
0	0	0	0
0	0	0	0

- We have created new categories for the features available in the particular restaurant.

avg_rating	total_reviews_count
3.0	13.0
4.5	22.0
3.5	56.0

avg_rating	total_reviews_count
average	13.0
good	22.0
average	56.0

- We are binning the average rating column into 4 categories such as “Poor, Average, Good, Excellent”.

Encoding:

BEFORE ENCODING	AFTER ENCODING	INFERENCE

vegetarian_friendly	vegan_options	gluten_free
N	N	N
Y	N	N
N	N	N

vegetarian_friendly	vegan_options	gluten_free	c
0	0	0	
1	0	0	
0	0	0	

Encoding the categorical variables into 0 and 1 for machine learning purposes .

latitude	longitude	claimed
44.383064	11.639176	Unclaimed
44.384930	11.635023	Unclaimed
44.383890	11.636492	Unclaimed

latitude	longitude	claimed
44.383064	11.639176	0
44.384930	11.635023	0
44.383890	11.636492	0

- We have encoded the categorical column claimed as 1 for Unclaimed and 0 for claimed.

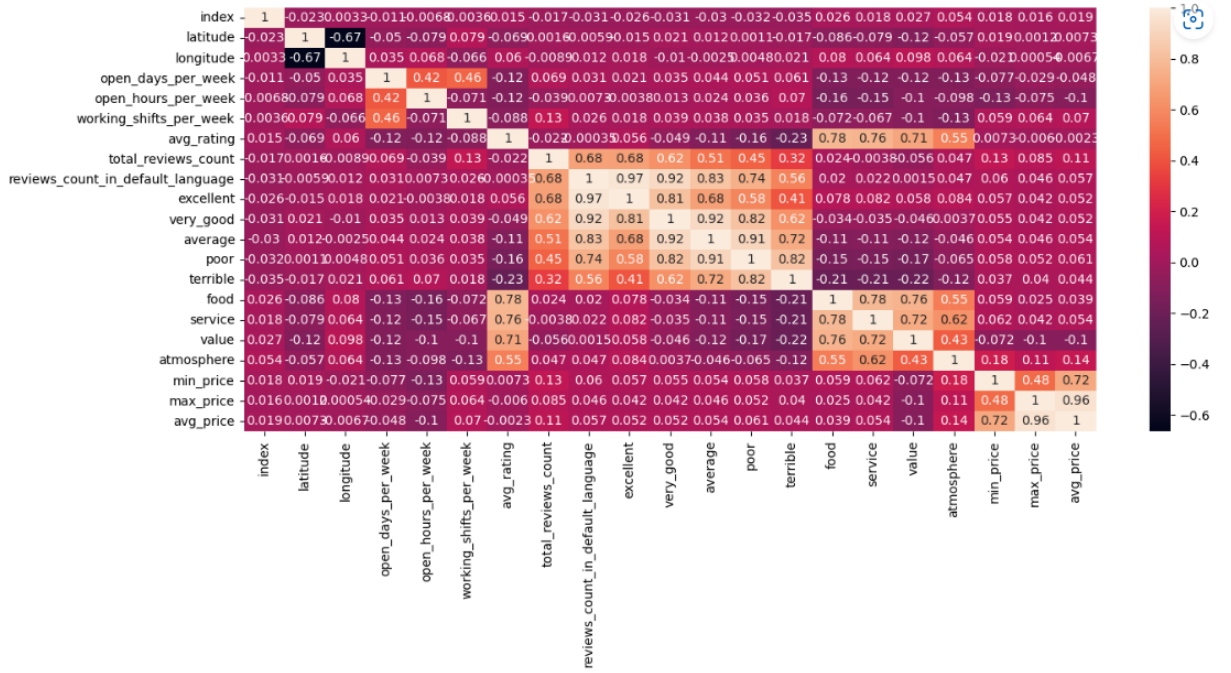
total_reviews_count	default_language
13.0	All languages
22.0	All languages
56.0	All languages

total_reviews_count	default_language
13.0	1
22.0	1
56.0	1

- We have encoded the column default_languages as 0 for English and 1 for All languages.

HEAT MAP:

```
sns.heatmap(df_italy.corr(),annot=True)
plt.show()
```



- The variables food, services and values have a strong correlation with the target variable.
- Atmosphere is moderately correlated with the target variable.

Statistical Tests:

Chi-Square Test:

The chi-square test is used to compare actual outcomes with predictions. The goal of this test is to establish whether a discrepancy between observed and expected data is the result of chance or a correlation between the variables you are researching. To examine the association between two categorical variables, perform the chi-square test.

Ho(Null hypothesis): that the predictor is independent of target.In other words, predictor and target are not related

Ha(Alternate hypothesis): Ho is false

```
cols = ['region', 'city','claimed','popularity_detailed', 'popularity_generic'
       ,'vegetarian_friendly', 'vegan_options', 'gluten_free']
```

```
for i in cols:
    tbl = pd.crosstab(df_italy.loc[:,i],df_italy.avg_rating)
    teststats, pvalue,df,exp_feeq = stats.chi2_contingency(tbl)
    print('Pvalue for',i,'is',np.round(pvalue,5))
```

```
Pvalue for region is 0.0
Pvalue for city is 0.0
Pvalue for claimed is 0.0
Pvalue for popularity_detailed is 0.0
Pvalue for popularity_generic is 0.0
Pvalue for vegetarian_friendly is 0.0
Pvalue for vegan_options is 0.0
Pvalue for gluten_free is 0.0
```

- Since the Pvalue is less than 0.05 we reject the null hypothesis.
- The above attributes are significant .

One-Way-ANOVA:

- In order to determine whether there is statistical support that the associated population means are statistically substantially different, one-way ANOVA ("analysis of variance") examines the means of two or more independent groups.One-Way ANOVA is a parametric test. This test is also known as: One-Factor ANOVA
- When examining if variations or different levels of a single independent variable, or factor, have a measurable impact on a dependent variable, one-way ANOVA is frequently used.

```
num_cols = ['total_reviews_count',
       'default_language', 'reviews_count_in_default_language', 'excellent',
       'very_good', 'average', 'poor', 'terrible', 'food', 'service', 'value'
           ,'avg_price', 'total_COE_awards', 'total_tc_awards',
       'open_hours_per_day', 'lunch', 'Breakfast', 'Dinner', 'Brunch',
       'Drinks']
```

```
signif_feats =[]
for i in num_cols:
    zero = df_italy.loc[df_italy.avg_rating=='excellent',i]
    one = df_italy.loc[df_italy.avg_rating=='good',i]
    two = df_italy.loc[df_italy.avg_rating=='average',i]
    three = df_italy.loc[df_italy.avg_rating=='poor',i]

    teststsats,pvalue = stats.f_oneway(zero,one,two,three)
    if pvalue<0.05:
        signif_feats.append(i)
```

```
signif_feats
```

```
['total_reviews_count',
 'default_language',
 'reviews_count_in_default_language',
 'excellent',
 'very_good',
 'average',
 'poor',
 'terrible',
 'avg_price',
 'total_COE_awards',
 'total_tc_awards',
 'open_hours_per_day',
 'lunch',
 'Breakfast',
 'Dinner',
 'Brunch',
 'Drinks']
```

- After performing the one way anova, we have identified the significant numerical variables.

Model Building:

Decision Tree:

- A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes.Decision tree is one of the predictive modeling approaches used in statistics, data mining and machine learning. • Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. • Tree models where the target variable can take a discrete set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Classification And Regression Tree (CART) is general term for this

	precision	recall	f1-score	support
0	0.86	0.97	0.91	79
1	0.81	0.82	0.82	1984
2	0.94	0.92	0.93	8570
3	0.83	0.87	0.85	1505
accuracy			0.90	12138
macro avg	0.86	0.90	0.88	12138
weighted avg	0.90	0.90	0.90	12138

XgBoost :

XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.

Before Hyperparameter Tuning:

Testing:

```
In [1358]: ypred_xgb = xgb.predict(xtest1)

In [1359]: print(classification_report(ytest1,ypred_xgb))
```

	precision	recall	f1-score	support
0	0.38	0.14	0.20	79
1	0.73	0.67	0.70	1984
2	0.85	0.91	0.88	8570
3	0.63	0.44	0.52	1505
accuracy			0.81	12138
macro avg	0.65	0.54	0.57	12138
weighted avg	0.80	0.81	0.80	12138

Training:

Training Accuracy

```
In [1364]: ypred_train=xgb.predict(xtrain1)
           print(classification_report(ytrain1,ypred_train))
```

```
              precision    recall  f1-score   support

           0       0.94      0.75      0.83       327
           1       0.84      0.78      0.81      8123
           2       0.89      0.95      0.92     34100
           3       0.82      0.61      0.70      5999

    accuracy                           0.88     48549
   macro avg       0.87      0.77      0.81     48549
weighted avg       0.87      0.88      0.87     48549
```

After HyperParametric Tuning:

Hyper Parametric Tuning

```
# Grid Search
```

```
params = [{'n_estimators':[100, 200, 500, 1000],'max_depth':[2, 4, 6, 8]}]

xgb = XGBClassifier()
grid = GridSearchCV(estimator = xgb, param_grid = params, cv = 5)

grid.fit(xtrain1,ytrain1)
grid.best_params_
```

```
{'max_depth': 2, 'n_estimators': 1000}
```

Testing:

```
In [ ]: # Test Data
```

```
In [1406]: ypred_xgb = xgb.predict(xtest1)
           print(classification_report(ytest1,ypred_xgb))
```

```
              precision    recall  f1-score   support

           0       0.33      0.14      0.20        79
           1       0.73      0.67      0.70      1984
           2       0.85      0.91      0.88      8570
           3       0.63      0.46      0.53      1505

    accuracy                           0.81     12138
   macro avg       0.64      0.54      0.58     12138
weighted avg       0.80      0.81      0.80     12138
```

Training:

```
In [ ]: # Train Data

In [1407]: ypred_train=xgb.predict(xtrain1)
           print(classification_report(ytrain1,ypred_train))
```

	precision	recall	f1-score	support
0	0.78	0.50	0.61	327
1	0.78	0.72	0.75	8123
2	0.86	0.93	0.89	34100
3	0.69	0.49	0.57	5999
accuracy			0.84	48549
macro avg	0.78	0.66	0.71	48549
weighted avg	0.83	0.84	0.83	48549

Business Interpretation:

From the above models we can come to the conclusion that the following are the important features that contribute to the average rating of the restaurant:

- The Restaurants in the following zip codes have a higher average rating:
    - 00186,00187,00185 - The city is Rome and it is a city that attracts a huge number of tourists.
    - 96100 - It is the city on the island of Sicily which is again a popular tourist destination notable for its rich Greek culture and history.
    - 20121, 20123 and 20124 - The city of Milano is another popular city known

for its art galleries, Catholic churches and Castles.
    - 50123 - The City of Florence is famous for its religious monuments which makes it a popular location to open up a restaurant.
    - 73100 - Leece, is famous for its historical architecture.
- The Restaurants serving the following cuisines have a higher average rating:
    - Pizza - When we say the name Italy, one of the first dishes that comes into our minds is Pizza. Hence we know the availability of Pizza in restaurants plays a vital role in its rating.
    - European - Italy is a country in the European continent and the travelers from other countries in the continent prefer to have the food that they are used to and hence European cuisine attracts the customers to the restaurant.
    - Seafood - Seafood is a major food source since Italy is a peninsular country. Hence serving Seafood in the restaurant will attract many customers.
- The Restaurants which have been claimed or verified by Tripadvisor see a higher footfall compared to the ones which are unclaimed. This plays a vital role in the rating of the restaurant
- The Restaurants having the following features attract more customers:
    - Takeout - Takeout/Takeaway is a popular feature seen in restaurants that have higher ratings as it is a touriSt destination. There are a lot of parks which are good picnic spots where people can enjoy their meals.
    - Wheelchair Accessibility - Restaurants that are Wheelchair Friendly also boost the average rating.
    - Reservations - Restaurants which take reservations are popular among the crowd since they can pre book tables.
- Price range of the meals in the restaurant plays a vital role in the footfall of customers. Customers generally prefer restaurants having a price range between

€10 - €30

- Restaurants serving vegan friendly and gluten free meals also help in gaining higher footfalls since Italy is a tourist destination and people following different diets visit the country.
- The Open Hours of a restaurant per day plays an important role in the customers footfall which directly also affects its average rating. As Italy is a tourist destination, people prefer restaurants which stay open for long hours.
- Wine & Dine is popular in European culture and hence restaurants serving drinks or attached with bars have a greater footfall
- The restaurants that have won awards should display it, this attracts customers and also helps in increasing the average rating