Zomato Restaurant Project with Machine Learning

How to make the most out of your Zomato Dining Experience with Machine Learning techniques.

Introduction:-

Zomato is a food delivery and restaurant discovery platform that operates in 24 countries around the world. The company was founded in 2008 in India and has since grown to become a leading player in the online food delivery and restaurant discovery space. Zomato allows users to search for and discover restaurants, view menus and ratings, and place orders for delivery or pickup. The company also offers a range of services for restaurants, including marketing and advertising, menu management, and point-of-sale systems. In addition to food delivery, Zomato also operates Zomato Gold, a subscription-based service that offers members discounts and perks at participating restaurants.

Benefits Zomato has over traditional ways of finding and ordering food

* Convenience: With Zomato, users can easily search for and discover restaurants, view menus, and place orders from the comfort of their own home or office. This eliminates the need to travel to a restaurant or make phone calls to place orders.
* Wide selection: Zomato offers a wide selection of restaurants and cuisines, allowing users to easily find a restaurant that fits their preferences and budget.
* Ratings and reviews: Zomato allows users to view ratings and reviews from other customers, which can help them make informed decisions about where to eat.
* Discounts and perks: Zomato Gold, a subscription-based service offered by the company, provides members with discounts and perks at participating restaurants.
* Integration with other services: Zomato often integrates with other services such as payment apps and loyalty programs, making it easier for users to pay and earn rewards.
* Contactless delivery: With contactless delivery, users can have their food delivered to their doorstep without the need for face-to-face interaction with the delivery person. This is especially beneficial during times when social distancing is necessary

SWOT Analysis for Zomato:-

**Strengths:**

* Strong brand recognition and reputation
* Large customer base and wide coverage area
* Extensive selection of restaurants and cuisines
* Strong partnerships with restaurants and food delivery partners

**Weaknesses:**

* Dependence on third-party delivery partners
* High commission fees for restaurants
* Competition from other food delivery and restaurant discovery platforms
* Limited control over the quality of food and service at partner restaurants

**Opportunities:**

* Expansion into new markets and regions
* Diversification of services, such as grocery delivery or meal kit delivery
* Partnerships with third-party companies and organizations
* Use of data science and machine learning to optimize operations and improve customer experience

**Threats:**

* Intense competition from other food delivery and restaurant discovery platforms
* Changes in consumer preferences and behaviour
* Economic downturns and recessions
* Regulatory and legal challenges
* Potential security and data privacy breaches

PROBLEM DEFINITION:-

Zomato Data Analysis is one of the most useful analyses for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines. Additionally, this analysis caters the needs of people who are striving to get the best cuisine of the country and which locality of that country serves those cuisines with maximum number of

Restaurants.

There are a few key ways in which Zomato data analysis can be used to achieve these goals:

* Identifying the most popular cuisines: By analyzing the data on Zomato, it is possible to identify the most popular cuisines in a specific area or country. This can be useful for foodies looking to try the best local dishes or for people looking to open a restaurant in an area with a high demand for a particular cuisine.
* Finding the best value for money restaurants: Zomato data analysis can also be used to identify restaurants that offer good value for money. This can be done by analyzing the data on the average cost of a meal at different restaurants, as well as the ratings and reviews of those restaurants.
* Identifying the best localities for specific cuisines: Zomato data analysis can also be used to identify the localities in a country that have the highest number of restaurants serving a particular cuisine. This can be useful for people looking to find the best places to eat a specific type of food.
* Improving the dining experience: By identifying the best value for money restaurants, foodies can get the most out of their dining experience, while also staying within their budget.

Therefore, the goal is to create an algorithm which will be able to accurately predict the price range as well the average cost of two based on the preferences of the user with the help of the features in the dataset .

**Dataset Details**

**Total 9551 rows & 22 columns**

* Restaurant Id: Unique id of every restaurant across various cities of the world
* Restaurant Name: Name of the restaurant
* Country Code: Country in which restaurant is located
* City: City in which restaurant is located
* Address: Address of the restaurant
* Locality: Location in the city
* Locality Verbose: Detailed description of the locality - Latitude: Longitude coordinate of the restaurant’s location
* Latitude: Latitude coordinate of the restaurant’s location
* Cuisines: Cuisines offered by the restaurant
* Average Cost for two: Cost for two people in different currencies
* Currency: Currency of the country
* Has Table booking: yes/no
* Has Online delivery: yes/ no
* Is delivering: yes/ no
* Switch to order menu: yes/no
* Price range: range of price of food
* Aggregate Rating: Average rating out of 5
* Rating colour: depending upon the average rating colour
* Rating text: text on the basis of rating of rating
* Votes: Number of ratings casted by people

DATA ANALYSIS:-

As the first step we would like to know the nature of data within each of the features as well as the target columns in the data set.

**Checking the unique values in the columns:**

* Restaurant Id: total 9551 ids as its different for each restaurant
* Restaurant Name: 7446 unique names , some are repeated
* Country Code: 15 codes so 15 countries
* City: 141 cities
* Address: 8918 adresses
* Locality: 1208 localities like the landmarks or surrounding area names
* Locality Verbose: 1265 same as locality
* Longitude: 8120
* Latitude: 8677
* Cuisines: 1825 different cuisines offered by the restaurant
* Average Cost for two: 140 different values
* Currency: Botswana Pula(P), 'Brazilian Real(R$), 'Dollar($),Emirati Diram(AED), Indian Rupees(Rs.),Indonesian Rupiah(IDR), NewZealand($), Pounds(\x8c£),Qatari Rial(QR), Rand(R), Sri Lankan Rupee(LKR),Turkish Lira(TL)
* Has Table booking: yes/no
* Has Online delivery: yes/ no
* Is delivering: yes/ no
* Switch to order menu: yes/no
* Price range: 3, 4, 2, 1, higher the costlier
* Aggregate Rating: Average rating out of 5 and total 33 unique ratings
* Rating colour: Dark Green, Green, Yellow, Orange, White, Red
* Rating text: Excellent, Very Good, Good, Average, Not rated, Poor
* Votes: Number of ratings casted by people total of 1012
* Country : Phillipines, Brazil, United States, Australia, Canada, Singapore, UAE, India, Indonesia, New Zealand, United Kingdom, Qatar, South Africa, Sri Lanka & Turkey

# Check the datatypes of the columns:

Restaurant Name object

City object

Locality Verbose object

Longitude float64

Latitude float64

Cuisines object

Average Cost for two int64

Currency object

Has Table booking object

Has Online delivery object

Is delivering now object

Price range int64

Aggregate rating float64

Rating color object

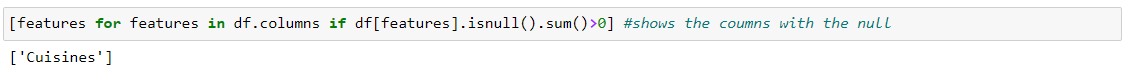
Rating text object

Votes int64

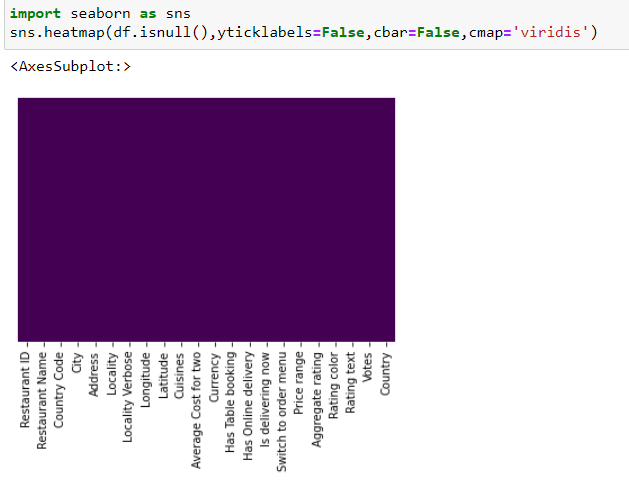
Country object

We see that we have 6 numerical columns and 11 categorical or object columns

Checking for NULL Values:-

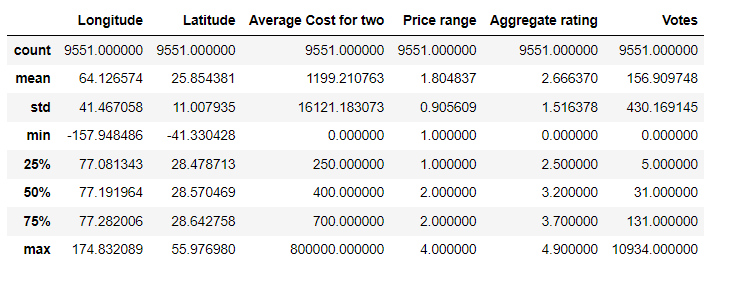


We see that that Cuisines has a 9 Null values which need to be treated , as the column has string values we will treat them by removing them or by replacing them with the mode of the column , In this case we replace them with mode.



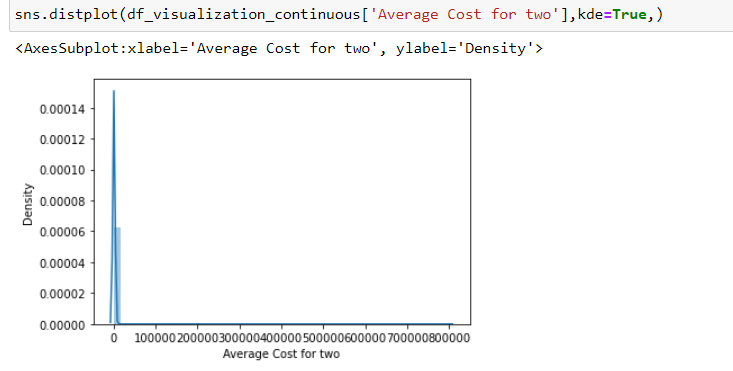
After treatment we see that there are no NULL Values in the column.

Describing the numerical features:

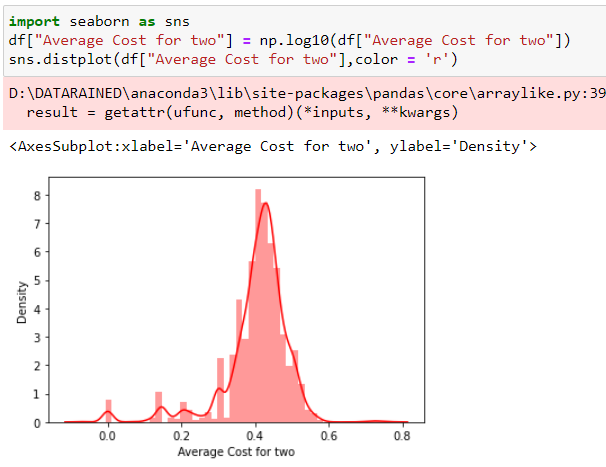


We see we have some outliers in Average cost of two as well as Votes , we see that cost is 0 in Average price of two which is not possible , so we need to treat it with log transformation techniques and Votes will be treated with Zscore.

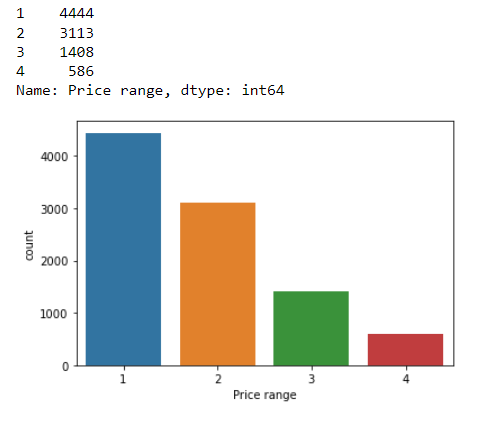
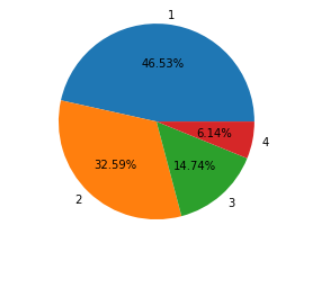
Before Transformation:



After Transformation:



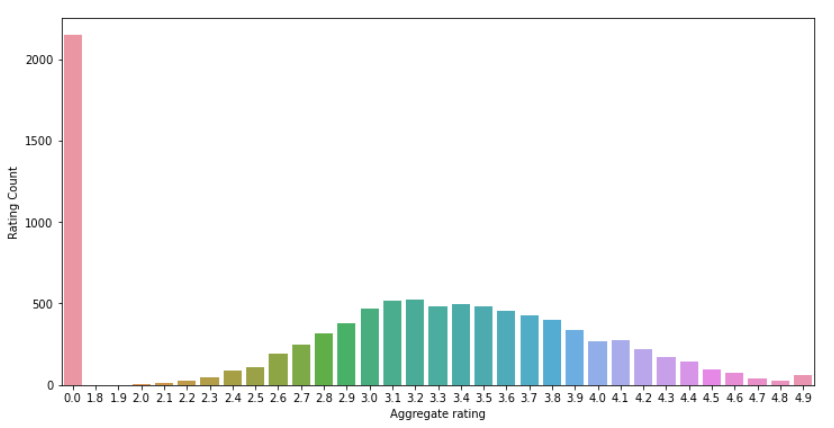
Visualization of the Target Variable – Price Range

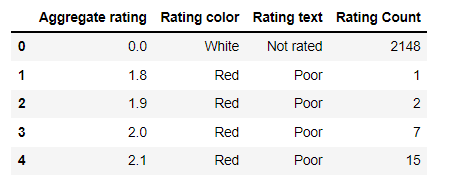
 

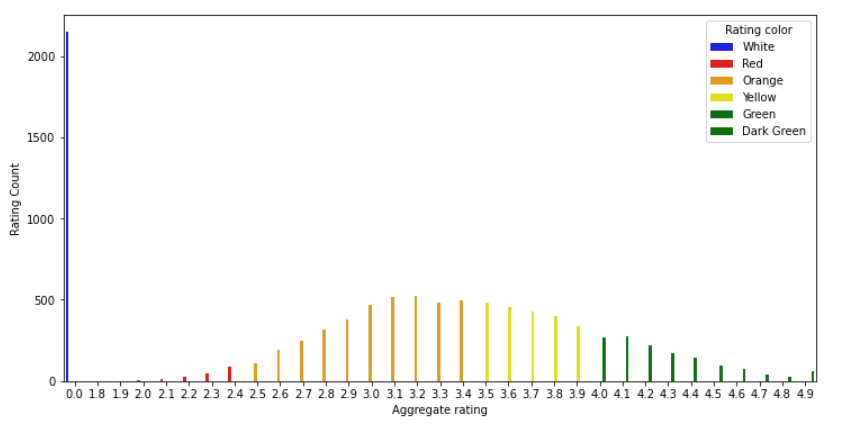
We see that the Price range category 1 is the highest at 4444 count and covers 46.53% of the data. Category 4 is the lowest at 586 and covers only 6.14% of the data.

Visualization of the numerical columns:

Aggregate rating:



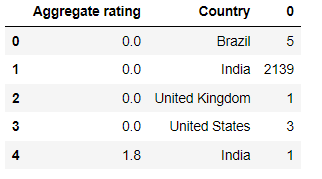




We see that the maximum count is no ratings at 2148 which means the users have not rated the experience. Then the maximum count for users who did vote lies between 2.5 to 3.4 as seen.

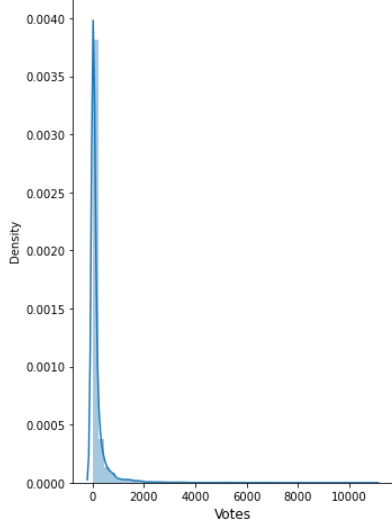
* Rating 0 — White — Not rated
* Rating 1.8 to 2.4 — Red — Poor
* Rating 2.5 to 3.4 — Orange — Average
* Rating 3.5 to 3.9 — Yellow — Good
* Rating 4.0 to 4.4 — Green — Very Good
* Rating 4.5 to 4.9 — Dark Green — Excellent

**Finding which country has given 0 rating:**



We see that India has the highest no of users who have not given ratings at 2139

Votes:

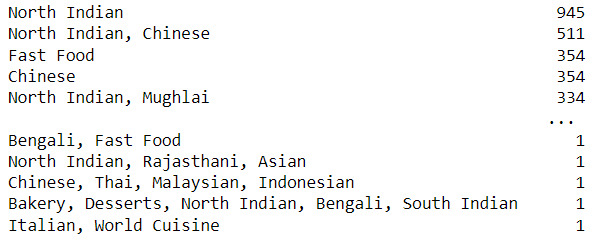


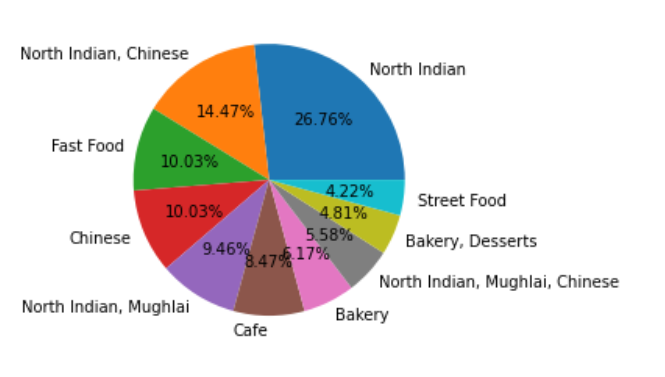
We see that Votes is also right skewed and the majority of the users have not voted their experience which shows as 0 on the distplot.

As we saw in aggregate rating so also here in Votes the same trend is reflecting as well.

Visualization of Categorical features:

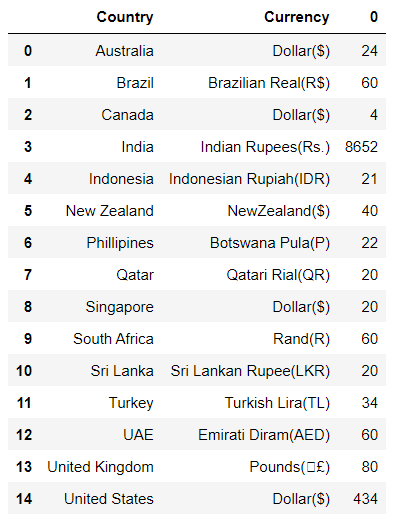
Cuisines:

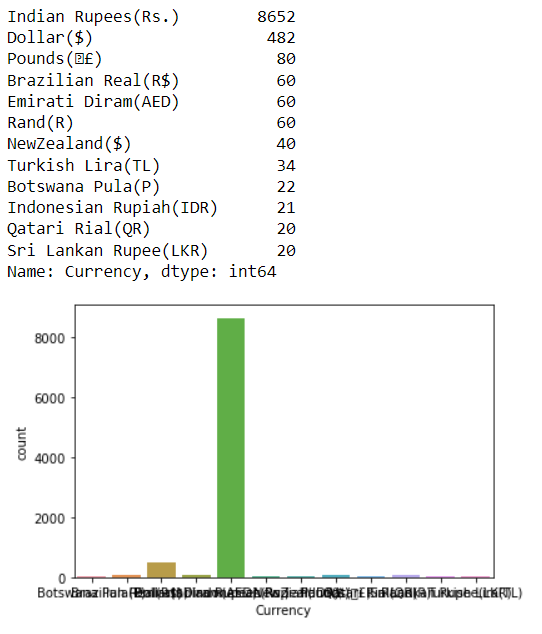


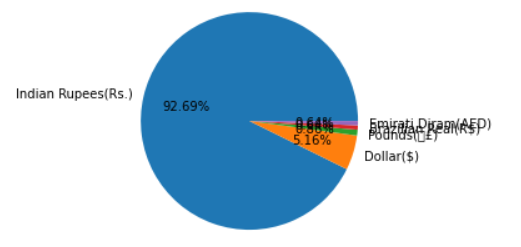


As there are too many cuisines we see the top 5 in the count table, But we see the top 10 in the pie chart which shows that North Indian is the highest at 26.76% and then, North Indian, Chinese is 2nd highest cuisine at 14.47% coverage

Currency:

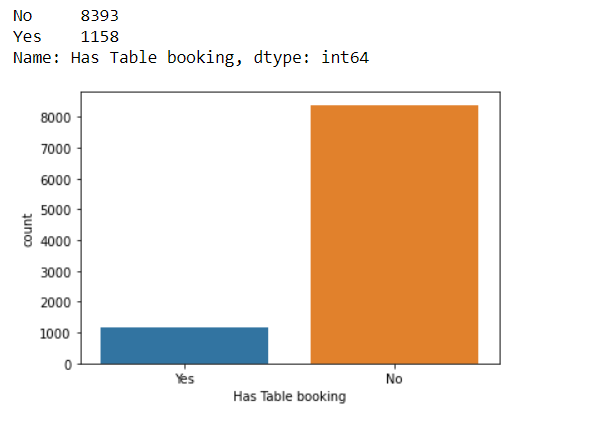






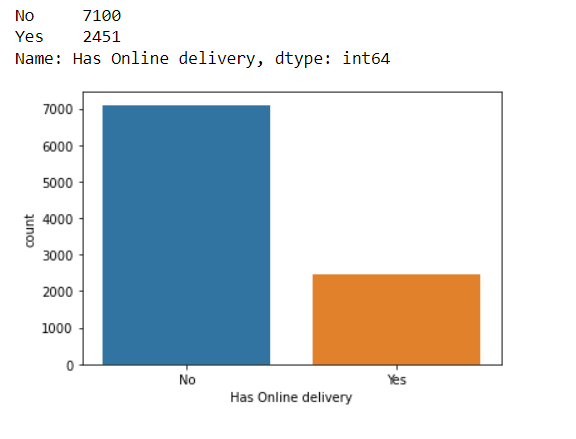
We see that 8852 user transactions have occurred in Indian Rupee and its coverage is 92.69% which is huge, Second is Dollar at 482(total US, Canada,Australia and Singapore) and 5.16% coverage and the rest are in double digits only , which shows that Zomato presence in India is prominent.

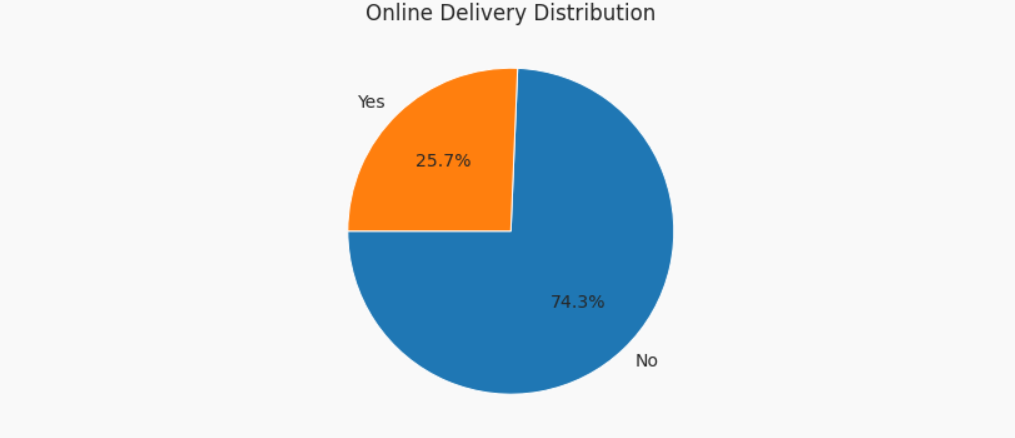
Has Table Booking(YES/NO):



We see that the majority of restaurants don’t have reservations and people can come in anytime.

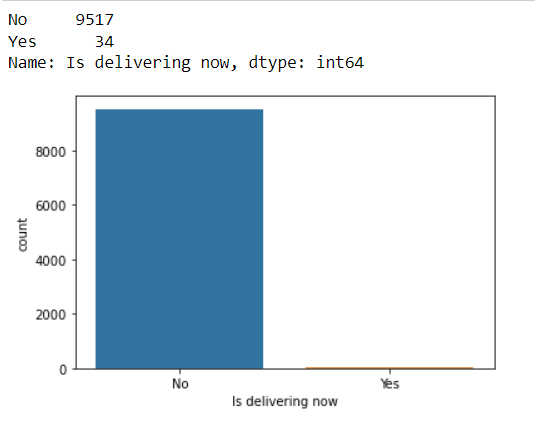
Has Online Delivery(YES/NO):





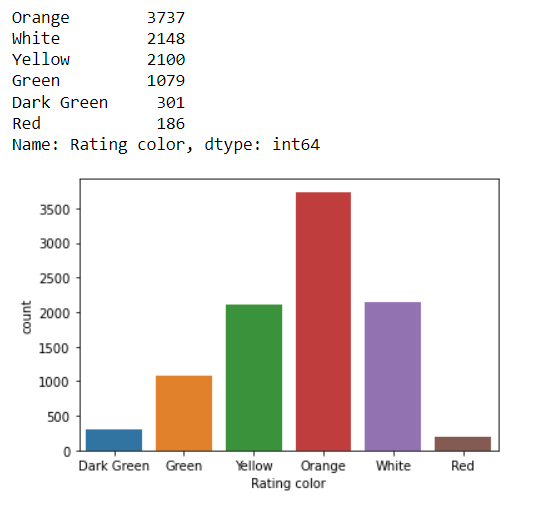
We see that the majority of restaurants don’t have online delivery

Is Delivering Now(YES/NO):



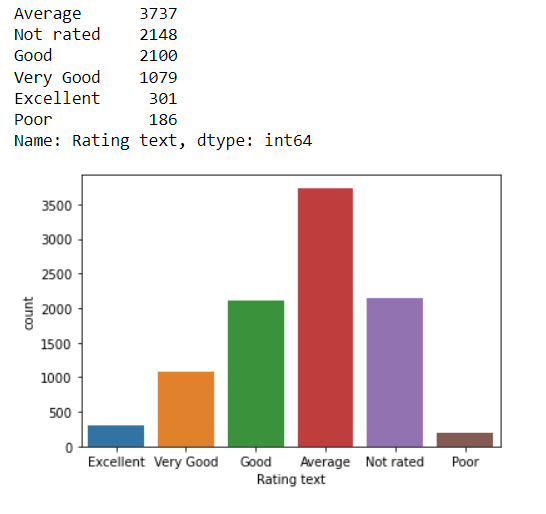
We see that the maximum restaurants are not delivering now and delivering now is only 34.

Rating Colour:



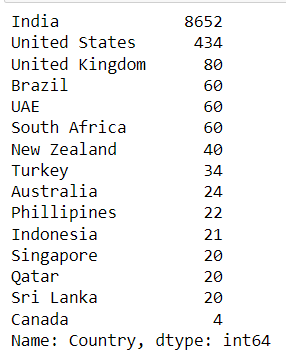
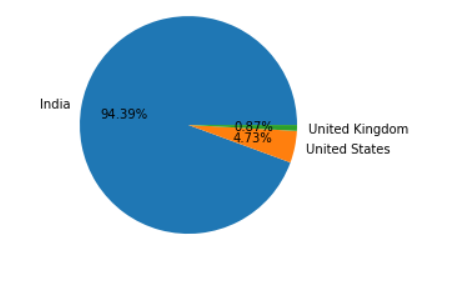
We see that orange rating which is Average is the highest at 3737, followed by white 2148 which is not rated and good i.e. Yellow is 3rd highest.

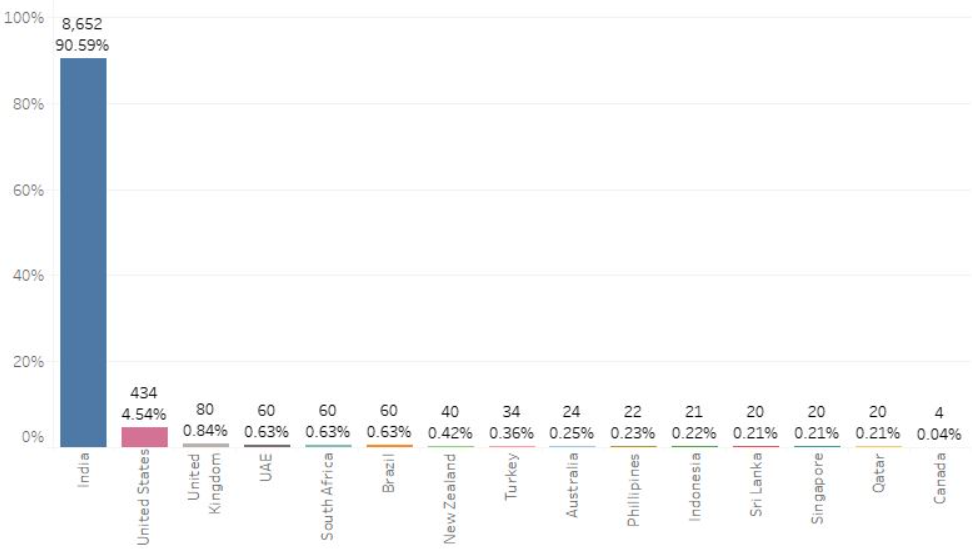
Rating Text:



As seen in the previous column, we see that Average is highest followed by not Rated and Good.

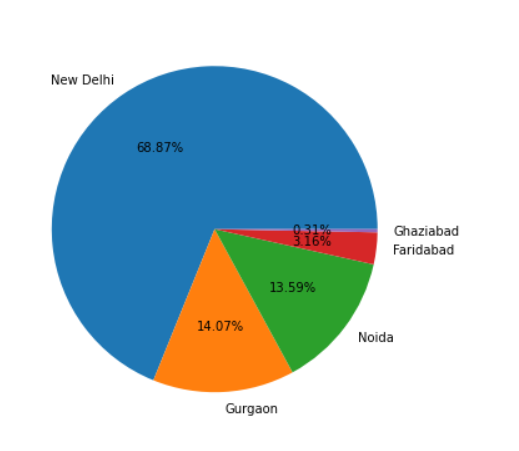
Country:



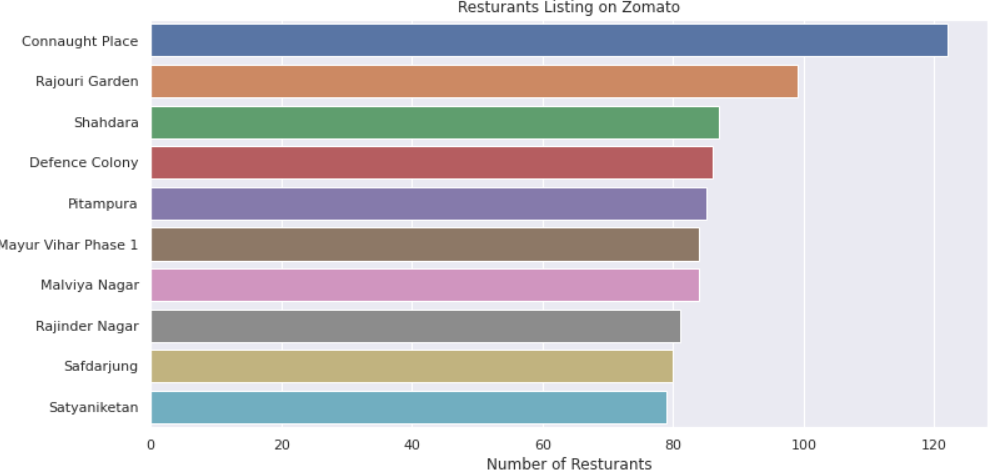
We see that India accounts for over 90% of the total dataset transactions which again shows the prominence of Zomato in the country , Second highest is USA and UK.

City:



We see that the top 5 cities in the dataset is in India and New Delhi is the highest, Gurgaon is the 2nd highest as well.

Locality:

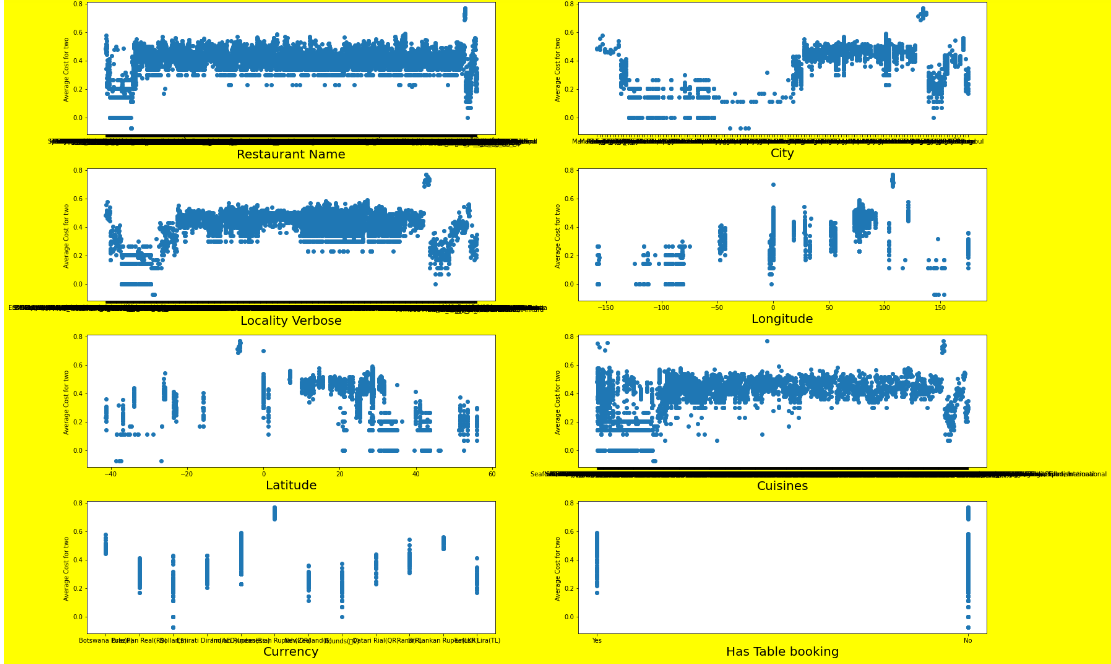


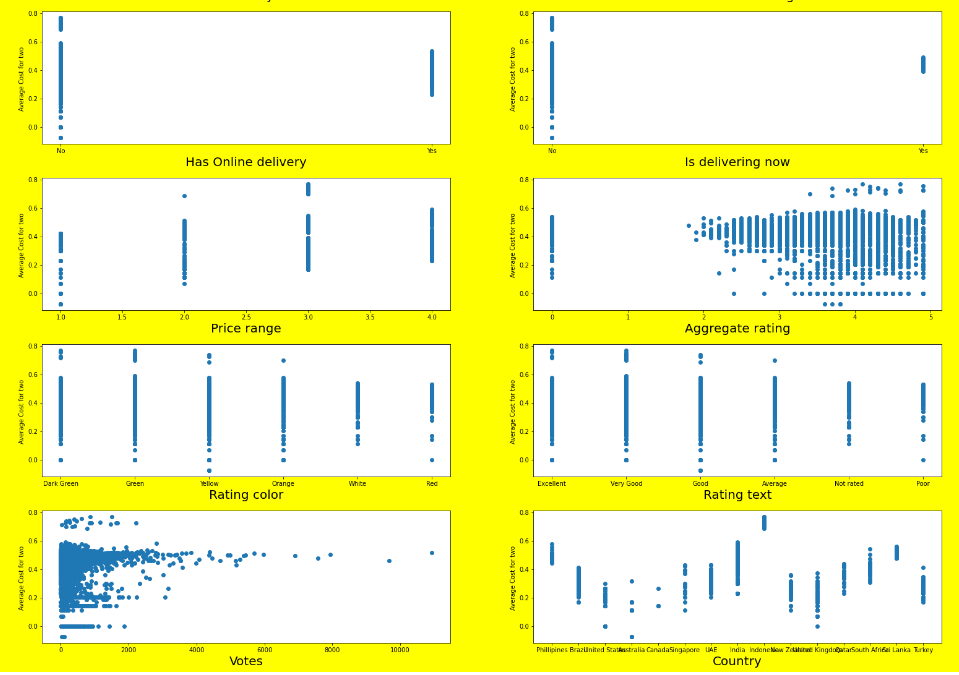
We see that New Delhi’s localities have the highest localities with the highest no of restaurants; In the graph we see that Connaught Place is the highest.

Relationship of features and target Bivariate Analysis:

With 1st Target- Average Cost for two:

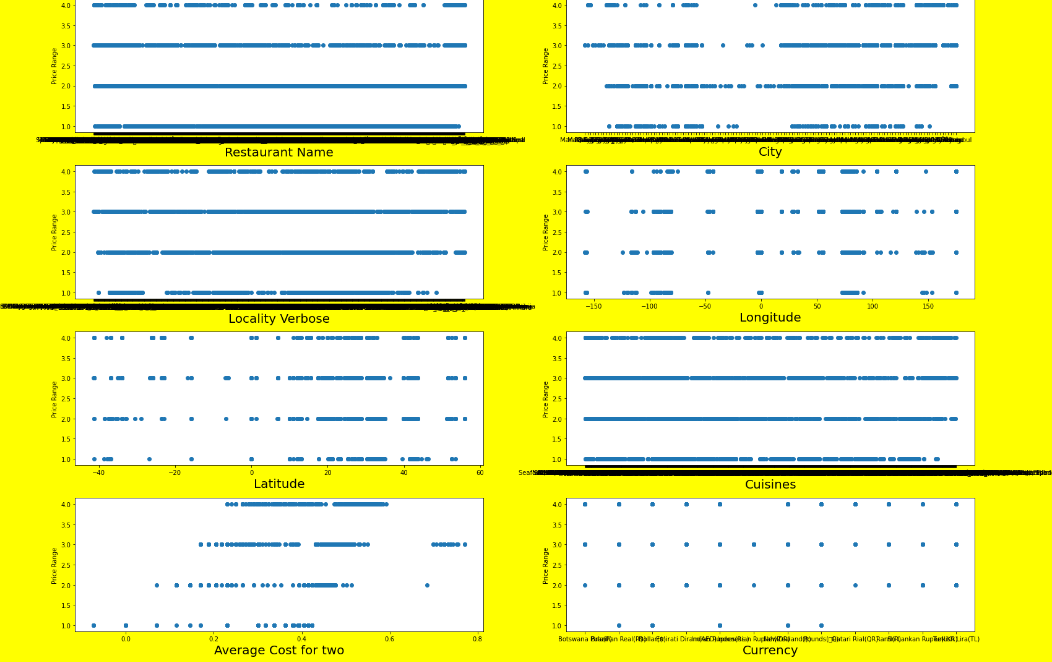
ScatterPlot:

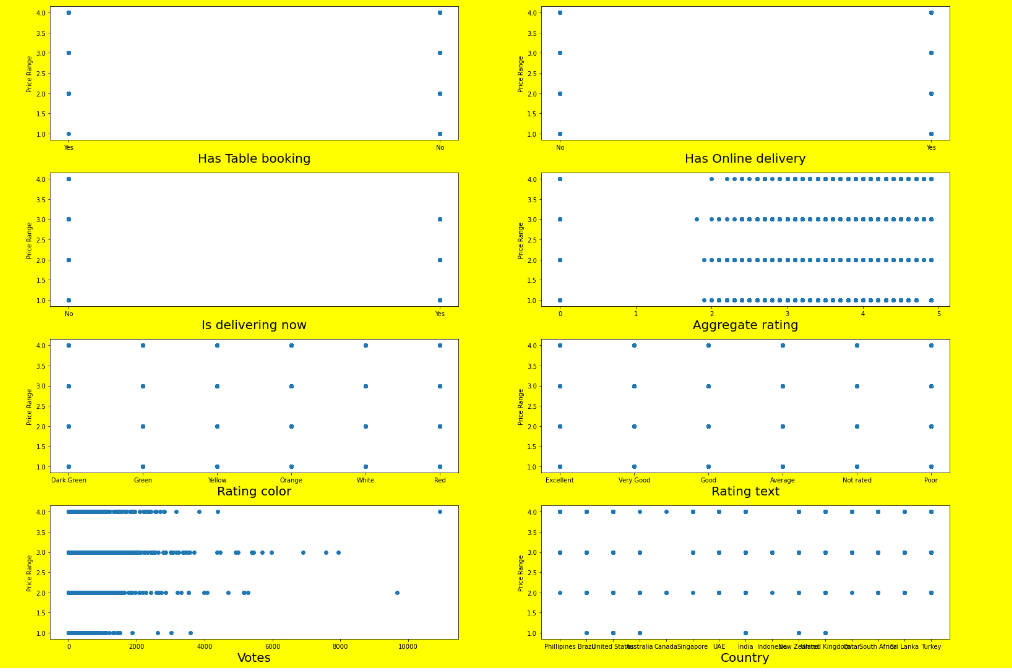




We don’t see any significant patterns or trends with the scatter plot , but we see that votes has max data in range 0-1000 approx and similar trend seen with aggregate rating as we saw individual dist plot

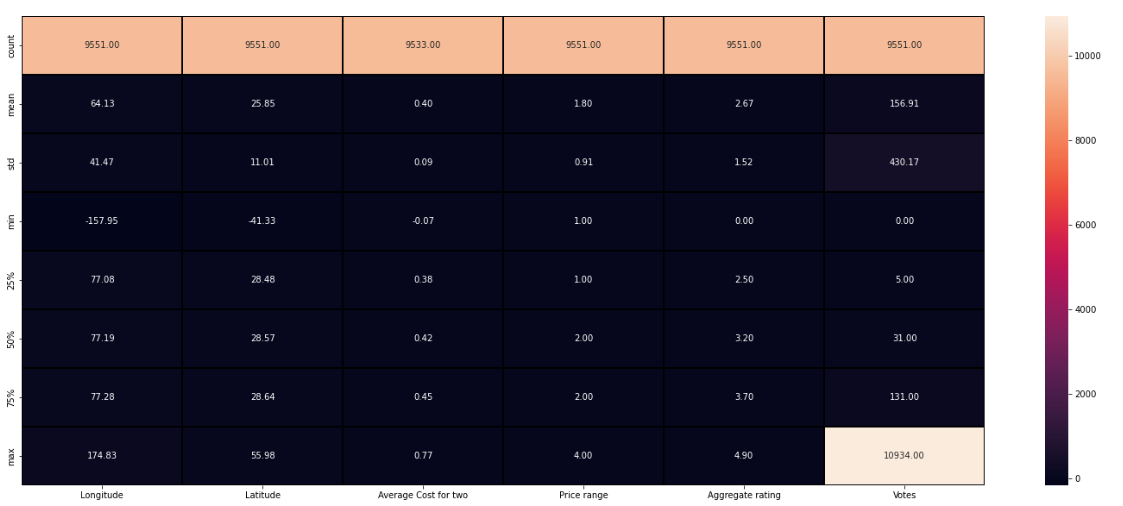
With 2nd Target- Price Range:





We see that since the target is categorical in nature it is difficult to ascertain the relationship here. See same trends with votes and Aggregate rating like the previous graph

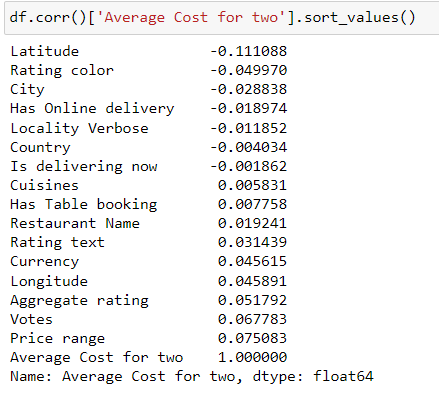
Describing the dataset with visualization:

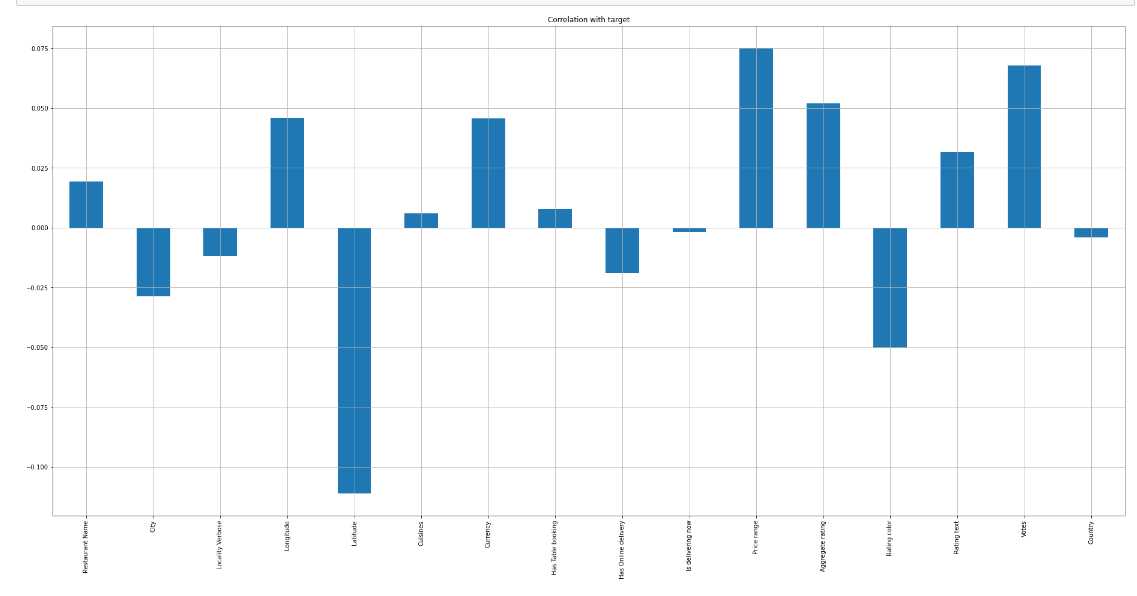


As we saw earlier we have treated the average cost of two column with log transform , now we see some outliers in votes and that std is greater than mean which is not correct which we will treat with zscore

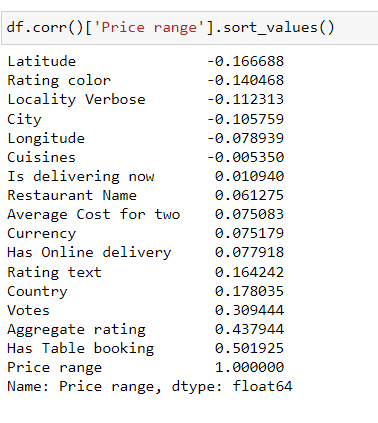
Multivariate Analysis:

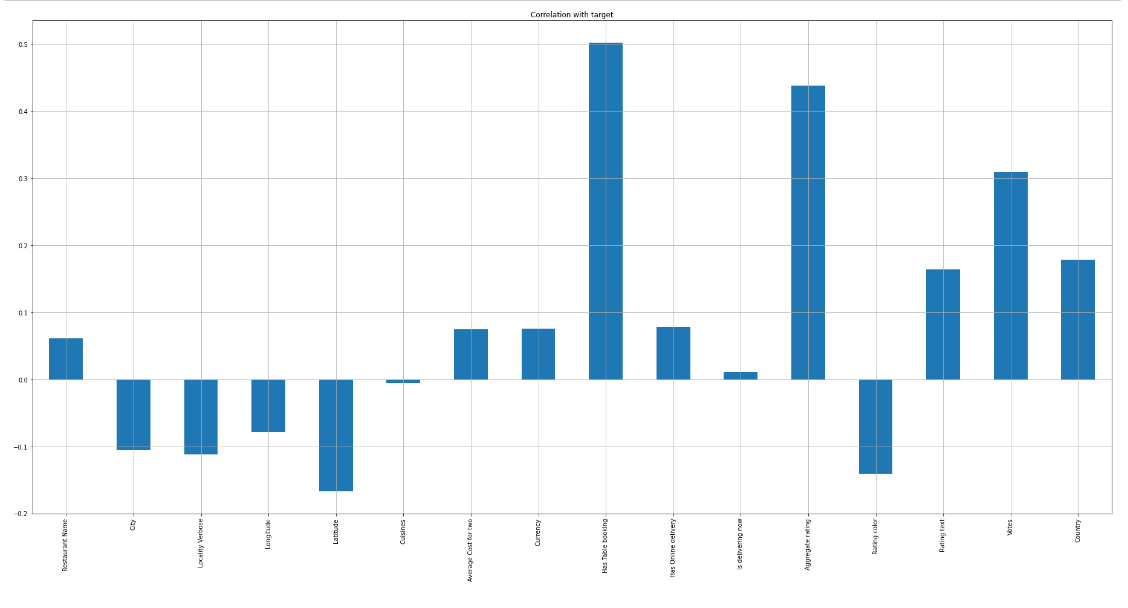
Firstly correlation with target 1 and 2



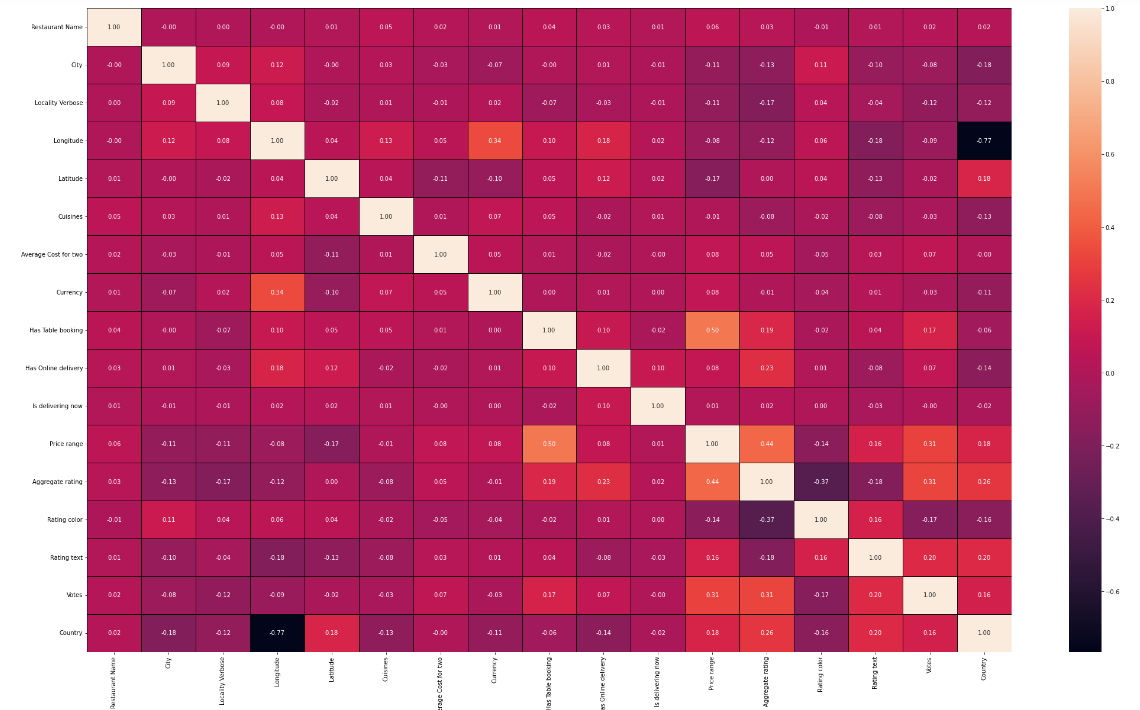


We see that price range, Votes , Aggregate rating are having the highest correlation with the Average Price for two label 1, but the overall impact is very less where the correlation is only 5-7 % which is not great ,



We see for the 2nd Label Price range Has Table booking, Votes and Aggregate Rating have a very good correlation score between 30% - 50%, these features will have the highest impact with the prediction

HEATMAP – Checking for Multicollinearity Problem



We don’t see any significant correlation between feature to feature but some of the relationships with the highest as per this dataset are:-

Currency and Longitude has 34%

Aggregate rating and votes 31%

Aggregate rating and Country 36%

Finally we see highest is Price range and Has table booking as we saw earlier with 50% correlation.

EDA Concluding Remarks

The dataset is skewed towards India having more than 90% transactions and doesn't represent the complete data of restaurants worldwide. This may have been so due to Zomato not having prominent presence in the other countries or we may have not got significant transactions in this dataset that we have analysed.

We also see that the majority of Indian users have not rated the dining experience , Whereas almost 97-98% of users in US and UK have rated the experience either in store or online , So we need to increase the customer engagement in India so that we can have more user give ratings.

Due to the sheer magnitude of the Indian transactions we are not able to get any relationship between the average cost and the ratings which will affect the score.

We also see that this dataset states that 90-95% of restaurants don’t have online delivery but the actual real world this is not the case and Zomato is literally having their maximum business from online orders , we can see in this comment by the CEO Deepinder Goyal on New year’s stating the sheer volume of online orders and lack of infra to support the demand which goes to show their main business is online delivery of food.

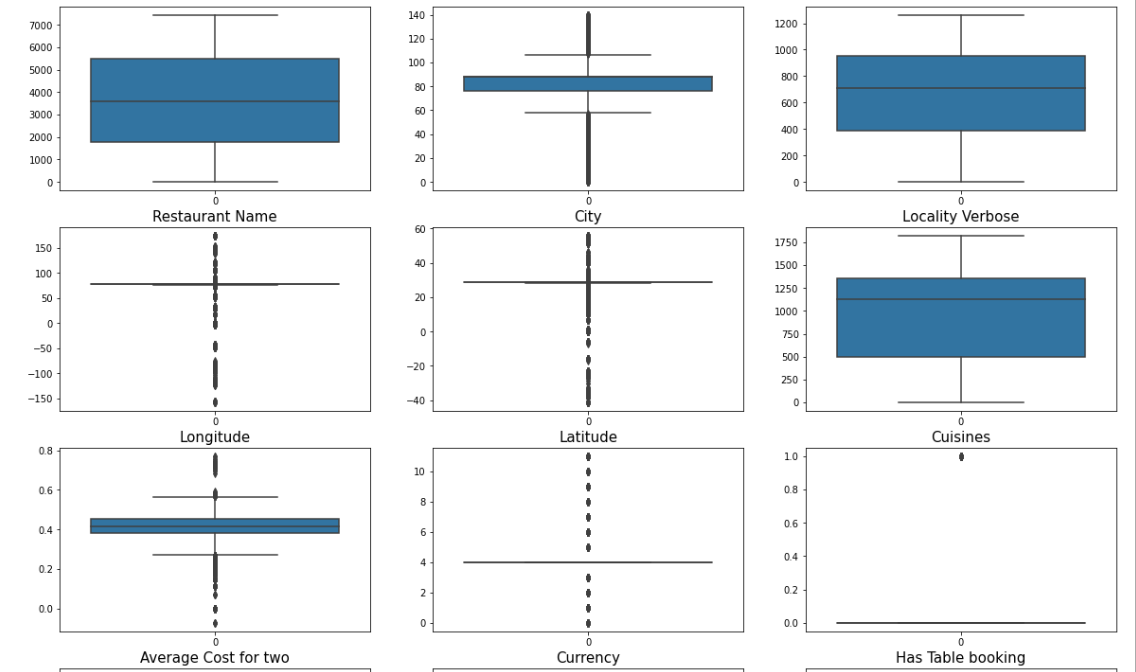


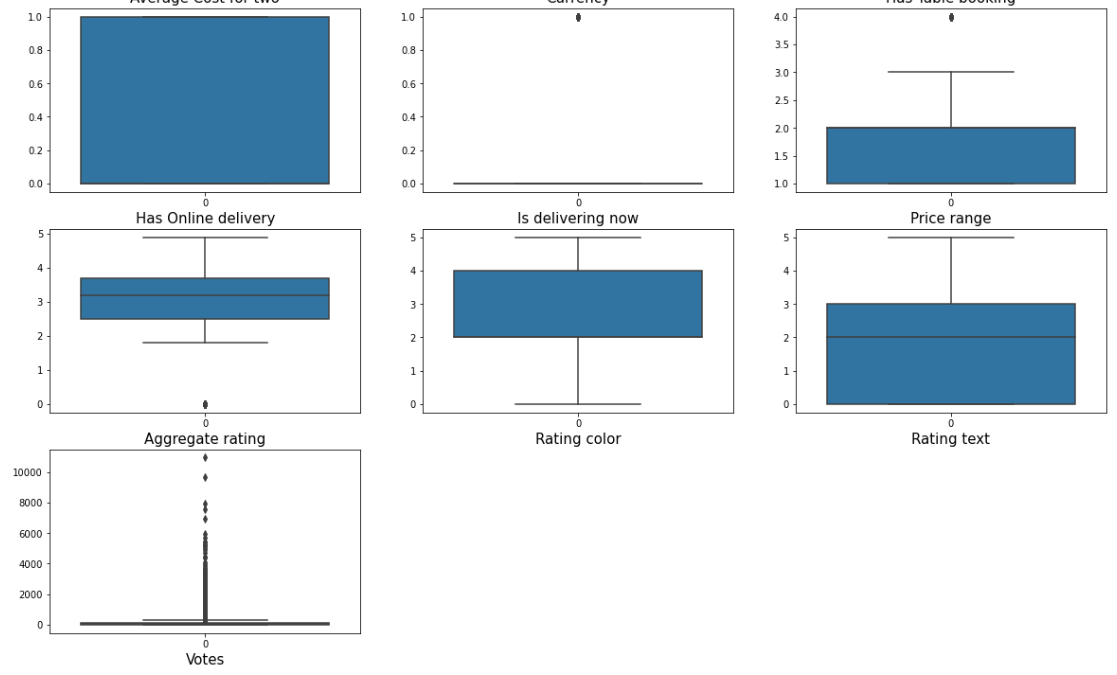
And to finally talk about the heat map of correlation we don’t see very many features making an impact of affecting the change in the label Average Price for two so we will need to do some feature selection techniques which we will see in the next part of this end to end project.

Pre-Processing Pipeline

Checking for Outliers:

Outliers are observations in a dataset that are significantly different from the majority of the other data points. In the context of data science and machine learning, outliers can have a significant impact on the results of an analysis or model. It is important to identify and handle outliers in a dataset because they can skew the results of an analysis or model and lead to incorrect or misleading conclusions

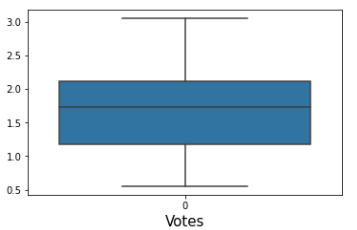
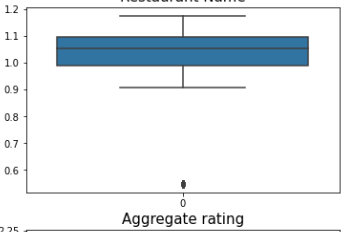




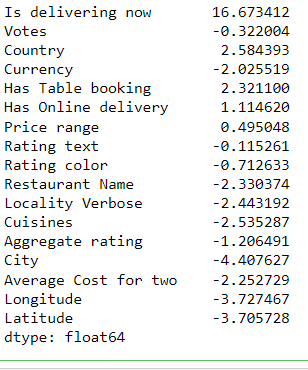
We see that Longitude , latitude, average cost for two (label),Votes are continuous variable which have outliers . We cannot treat the outliers of the label but the rest we will treat them with Zscore or Power transformation.

The rest of the columns are encoded categorical columns which we can avoid treating

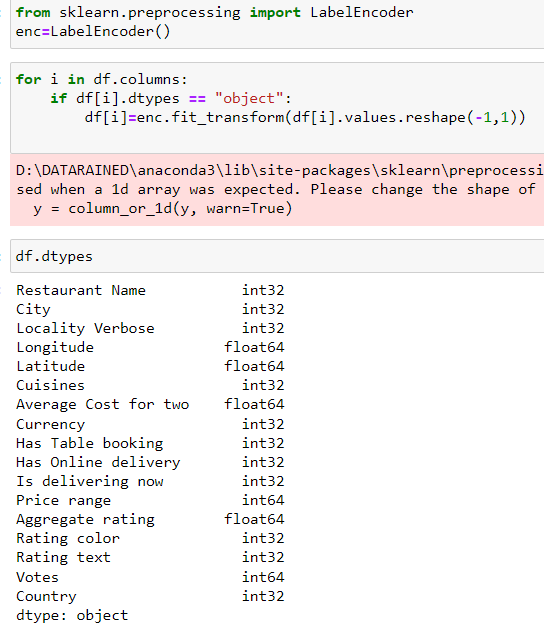
After treating with Box cox Power transformation method

The new skewness values:

 We see that we have reduced the skewness in all the columns and we have got rid of most of the outliers from continuous variables

**Encoding the categorical Features to numerical features**

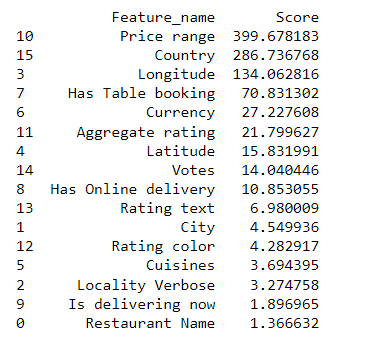


We have use label encoder to convert the categorical columns to numerical columns as the model will only understand numerical values.

**Using SelectKBest to do Feature selection using Classif**

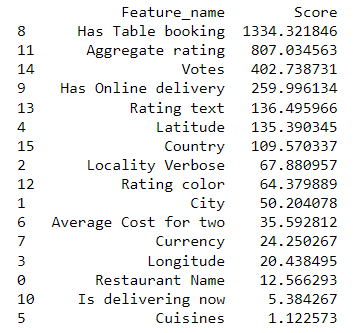
The SelectKBest function in scikit-learn is a feature selection method that can be used to select a specific number of the best features from a dataset based on the univariate statistical test used. The f\_classif function is a statistical test that can be used as part of the SelectKBest function to identify the features that are most associated with a target variable.

TARGET 1: Average Cost for Two



We see that Price range, Country, Longitude and Has Table booking show high numbers and constitute to the best features, Then we see a significant drop in the score to 27 and lesser. SO we can safely say these are the top influencers for the 1st target.

TARGET 2: Price range



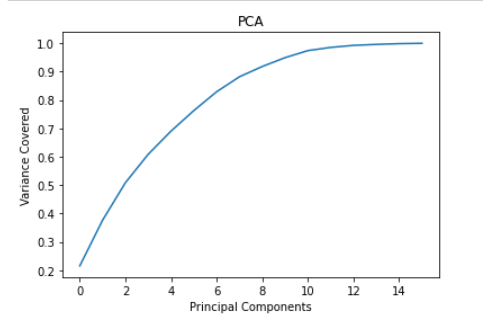
In this table for the Price Range label we see that Has table booking, Aggregate rating, Votes, Has online delivery, rating text, latitude, country, locality verbose, rating colour and City are have good scores the top 7 having phenomenal scores and will have the highest impact in the prediction of the target 2.

**Principal Component Analysis (PCA)**

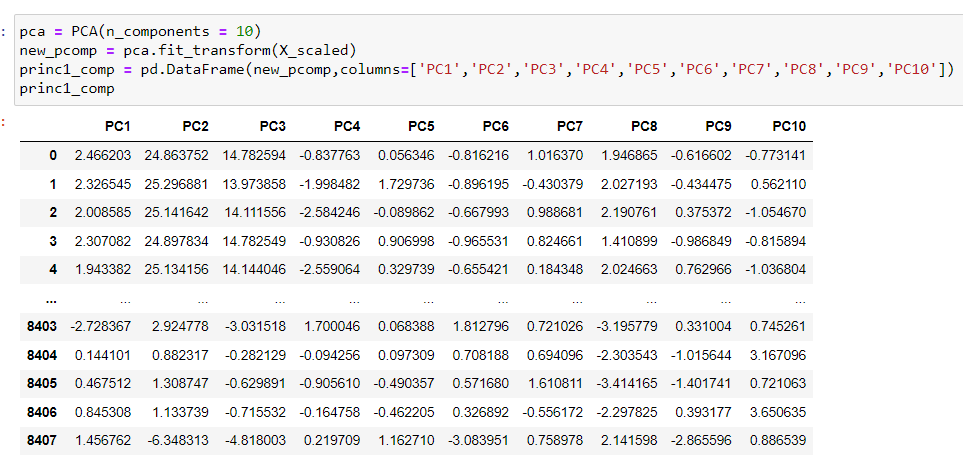
It is a dimension reduction technique and not a feature selection one to help us compute the problem quicker and get rid or non impactful features .

The reason we are going to apply PCA is as we saw in the Kbest feature selection , we have many features which are not having a big impact on the target . This will be applied on the features only , it is mainly used if there are too many features and no correlation with the target.

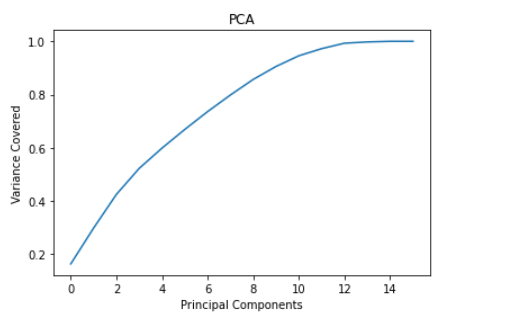
TARGET 1: Average Price for two

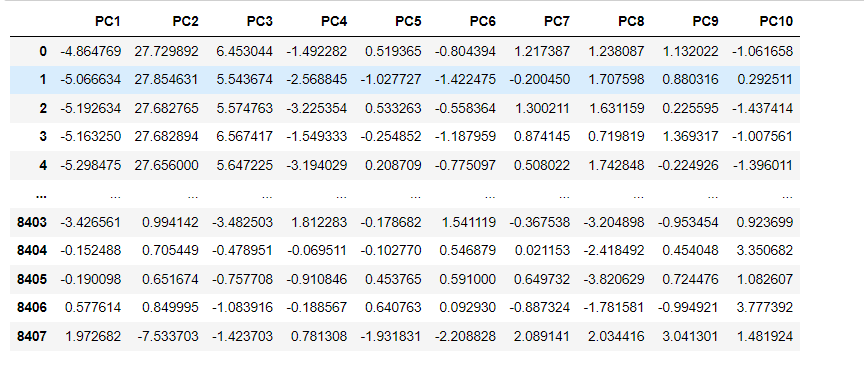


We see rom the Scree plot above that to cover 95% to 100% of the data only 10 features is enough for the 1st target and therefore we have done the needful and only taken those features in the data frame displayed below



TARGET 2: PRICE RANGE





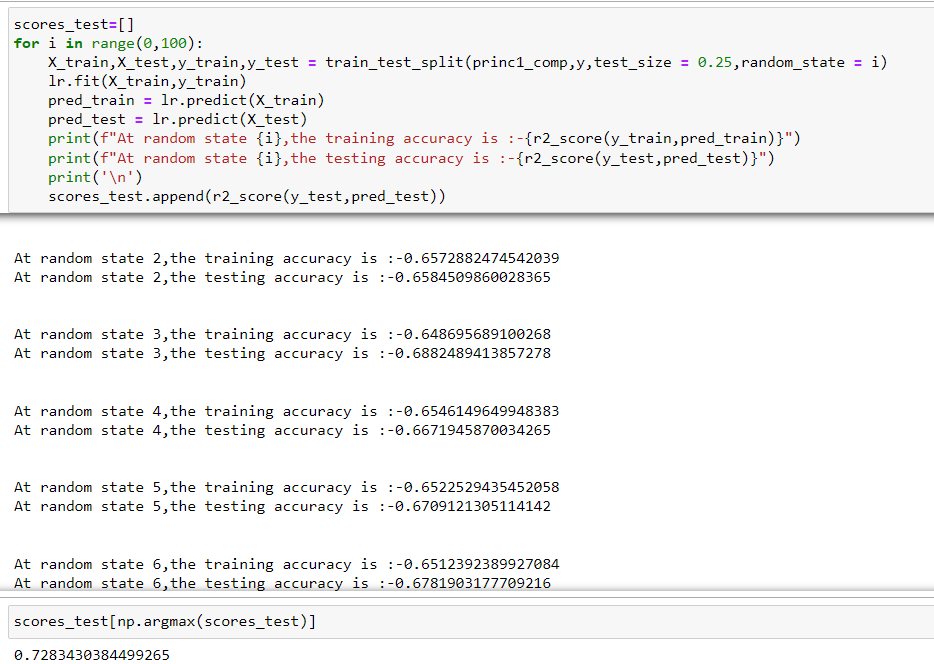
Here as well we see that 10 features are enough again , and we can omit the rest of the features and take only the Principal components for model building , Another advantage of doing this is in case the model is not having a good score , we can tweak around by increasing or reducing the features and see which combination will give the best score overall.

Building Machine Learning Models

TARGET 1: MODEL BUILDING FOR

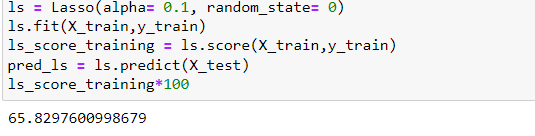
Average Cost For Two

**Linear Regression**

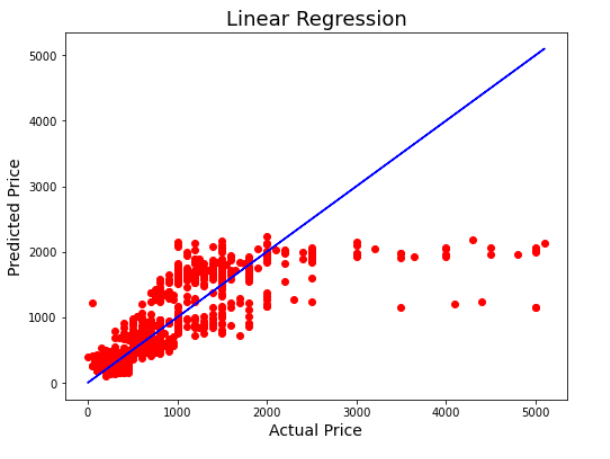


We have use train test split to split the training and Testing data, then used a for loop to check the best random state from 0 to 100 and we use Argmax function to derive the best score which is 72.8% for testing r2 score.

Regularization of the Linear Model

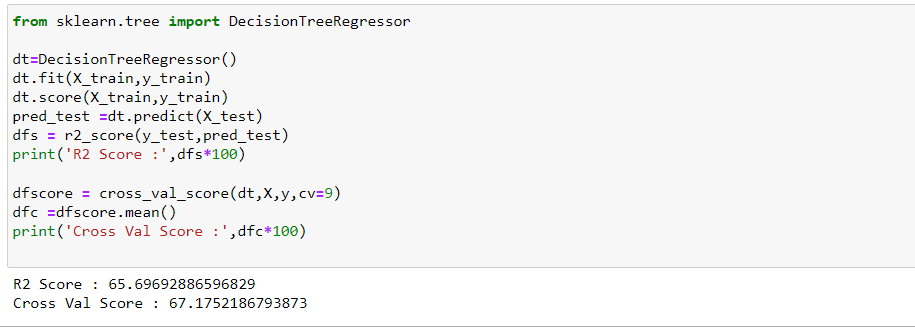


We are getting a lower score of 65.83 % appox using lasso regression and parameters tuned.



The model is not doing a good job as the predictions are not accurate and the model score is not up to the mark ,

**Decision Tree Regressor**

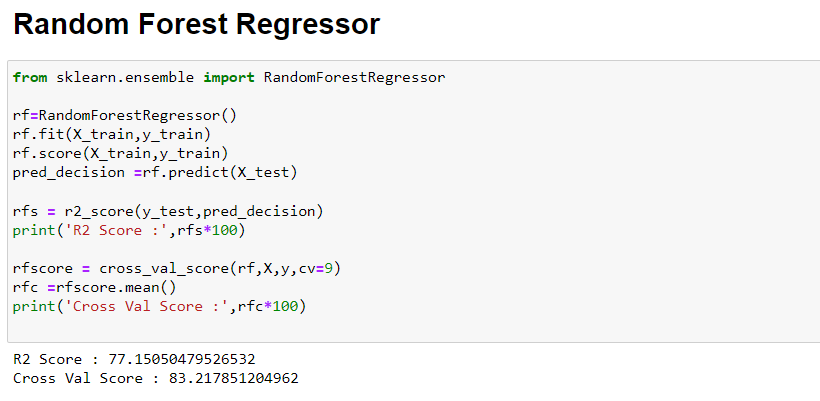


The model is at R2 Score of 65.70 %approx which is way better than linear regression, We also see that the cross val score is much much better than the previous model at 67.18% approx is really good as its close to the r2 score, There is very slight difference between the r2 score and cross val score , We see that the score is much much better than Linear regression model of 65.2% approx..

**K- Nearest Neighbors**



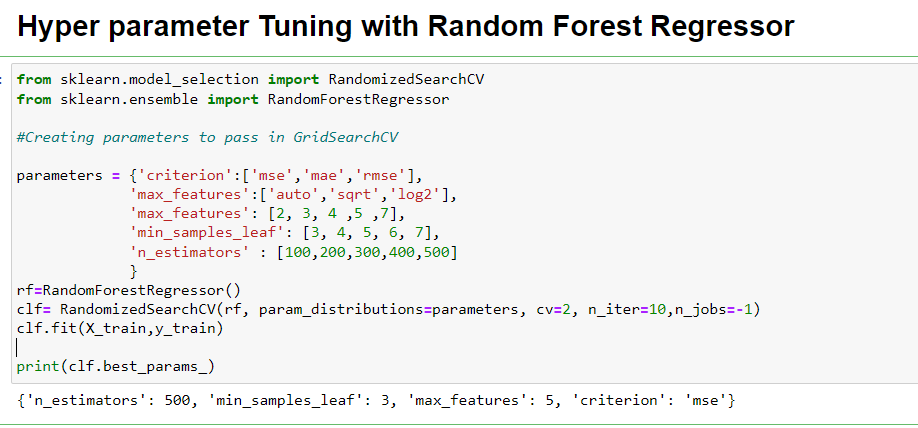
The model is not at all working well for the data set and we see that the score is much better than linear regression model & Decision tree @ 72.89% approx.,But the model has a very bad cv score , the model shows overfitting , we have cv score @ best state 9 which is 6.53% which is bad. There is very very high difference between the r2 score and cross val score

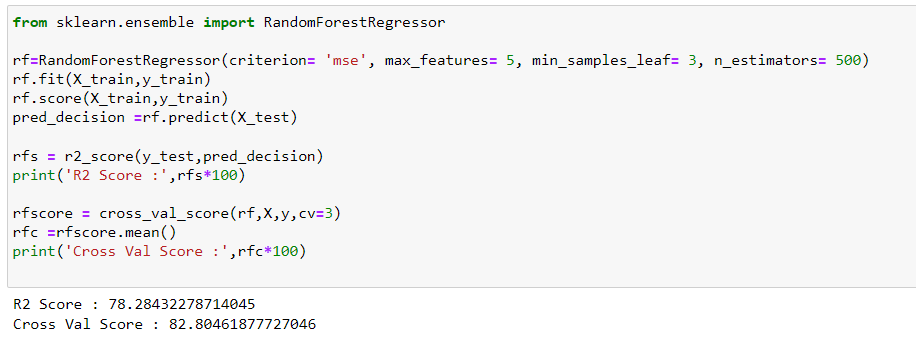


The model is doing much better than the 3 models we test before , the R2 score is @ 77% approx which is much better than the rest, decision tree was only giving 65% approx and our cross val is higher than the r2 score but still better @ 83.21% approx. We also see that the cross val score is a lot better compared to linear and tree, we see the Difference between the R2 score and cv score is good.



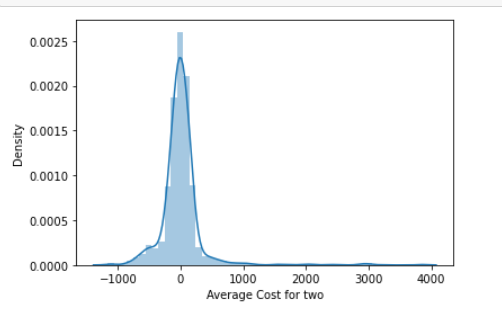
The model is giving a slightly better score than the random forest model @ 77.44% and forest gave 77.15%, The cv score is very different compared to forest @ 60.60 but the random forest gave 83% which means we can increase the score with tuning,



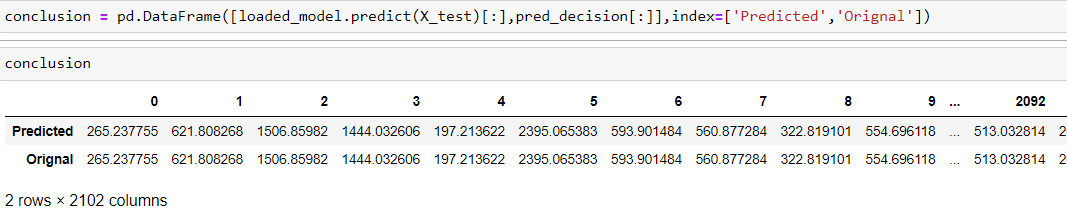


We see that we have increased the r2 score to 78% from 77% and have brought the cv score to 82% as the best model is the one which has the least difference between the r2 score and cv score.

Plotting the graphs with Random forest Regressor



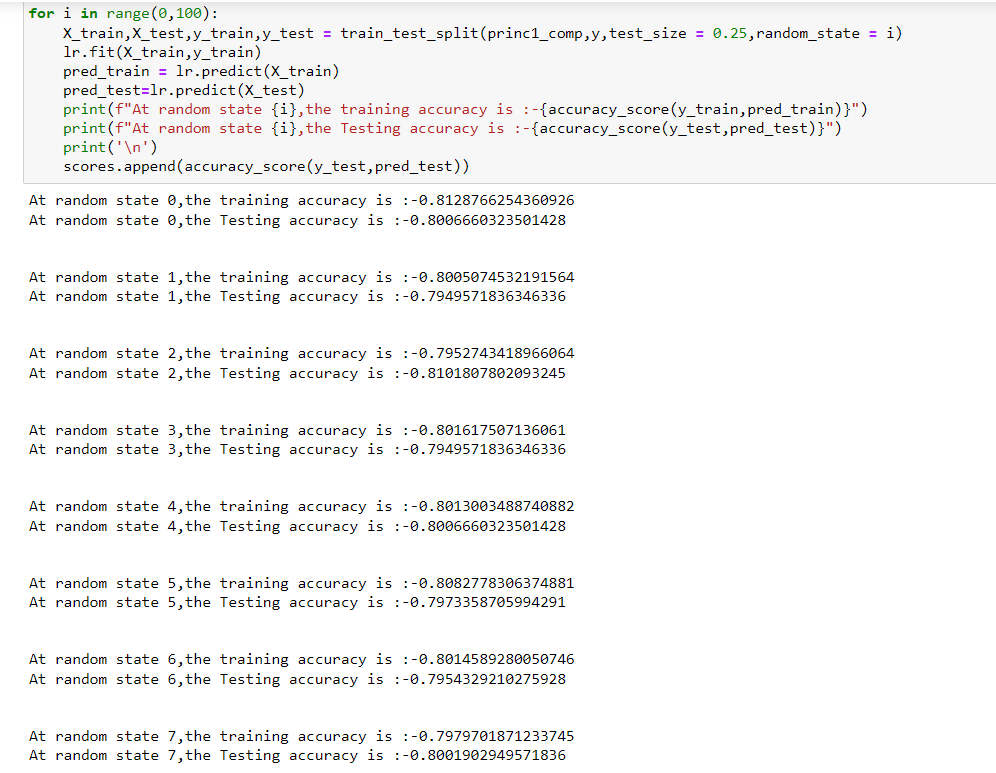
Predictions

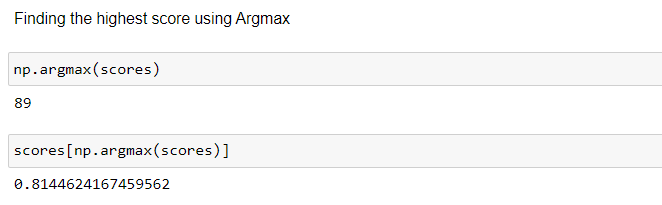


TARGET 2: MODEL BUILDING FOR

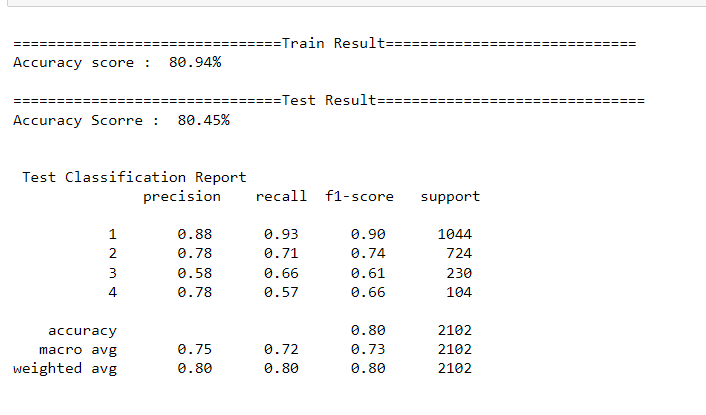
PRICE RANGE

**Logistic Regression**



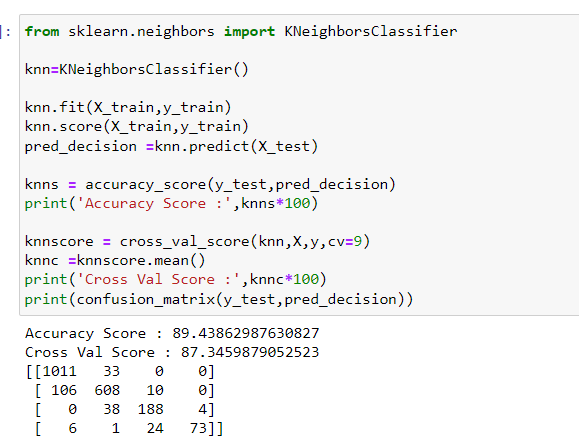


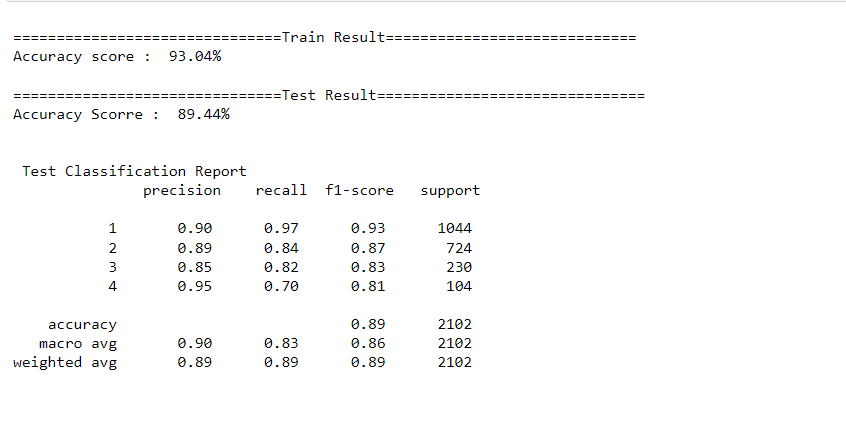
We have split the data into train and test and this time are solving a classification problem as the label is categorical in nature and we have used the same argmax to get the best accuracy score which is 81.44% for Logistic regression , But now we will choose the best random state and will create a classification report.



We have got a very good score where training is 80.94% and test is 80.45% , there are really close which means the model is really able to give the same accuracy as with the learning it has done on the training data.

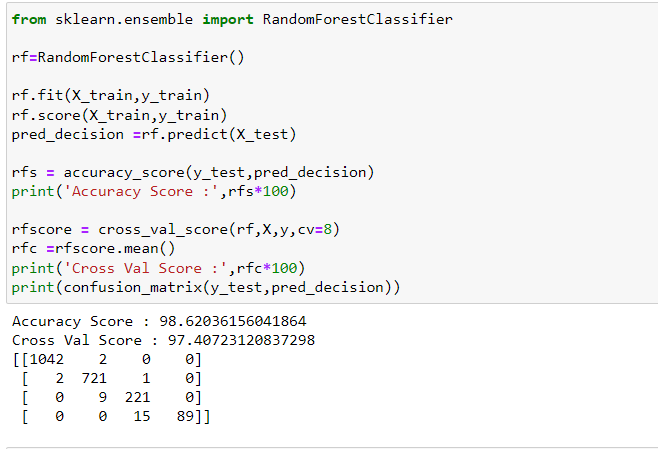
**KNN Classifier**

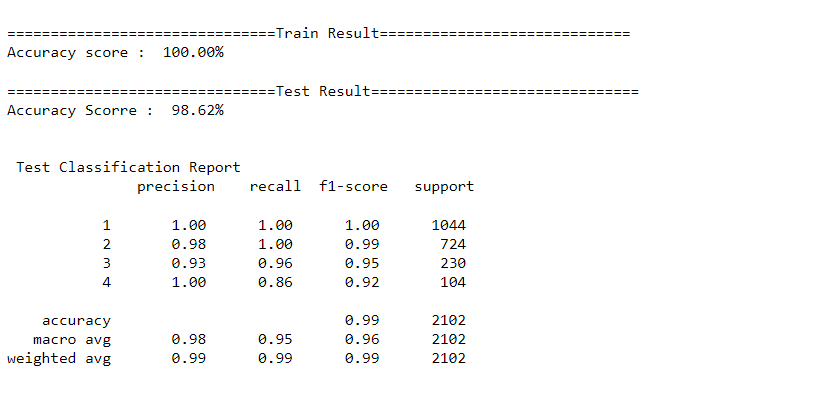




We see that the training score is lower than Decision tree @ 89.44% but the train score is closer to test @ 93.04% The CV score is good though and very similar to the test accuracy @ 87% highest among the 3 models we tested , so overall the model is ok ,but decision tree is the highest in number as off now We see the confusion matrix where the typ 1 and typ 2 error is much better than the previous models and we see that many of the error are 0 so its overall a good contender for the best model.

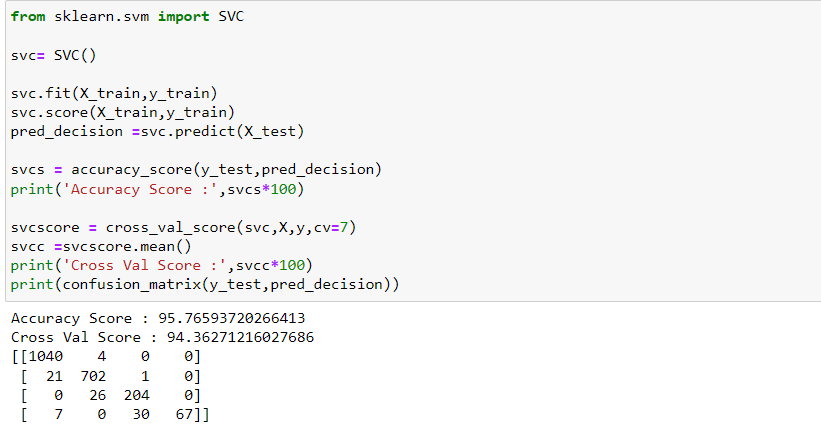
**Random Forest Classifier**

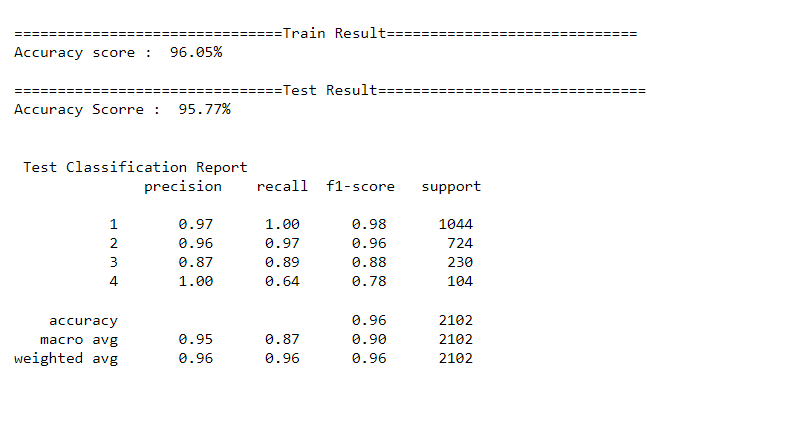




We see that like the decision tree the train score is at the max at 100% and we have test score equal to tree at 98.62% , the F1 score is at 99% and precision @ 99% which is a good score. We have an almost equal Cv score of 97.40% approx. so we are getting a good cv score as well which is on par with the test sore which is what we need the model to do and highest among the models we have seen. The model has much lower errors in the confusion matrix as all the models but can be avoided.

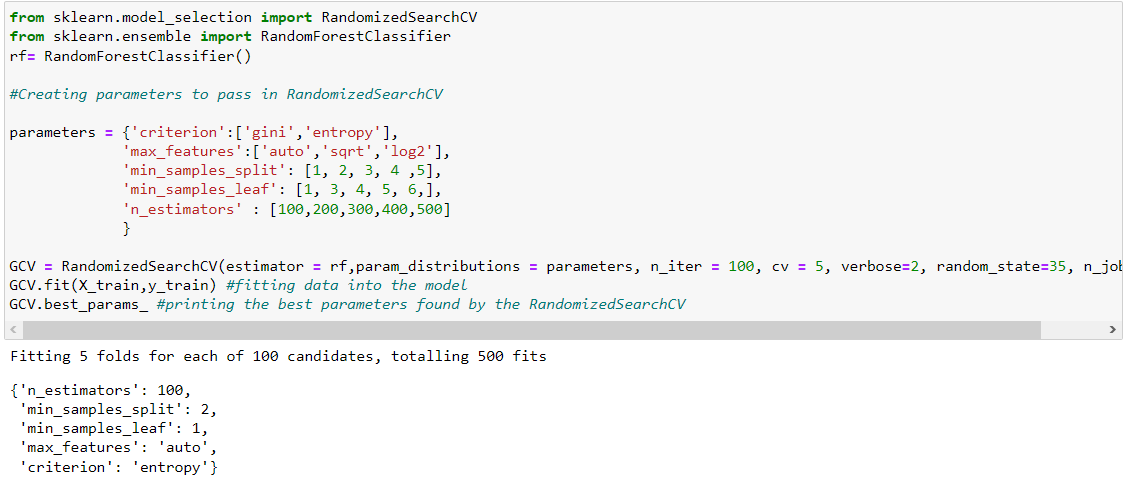
**SVC**

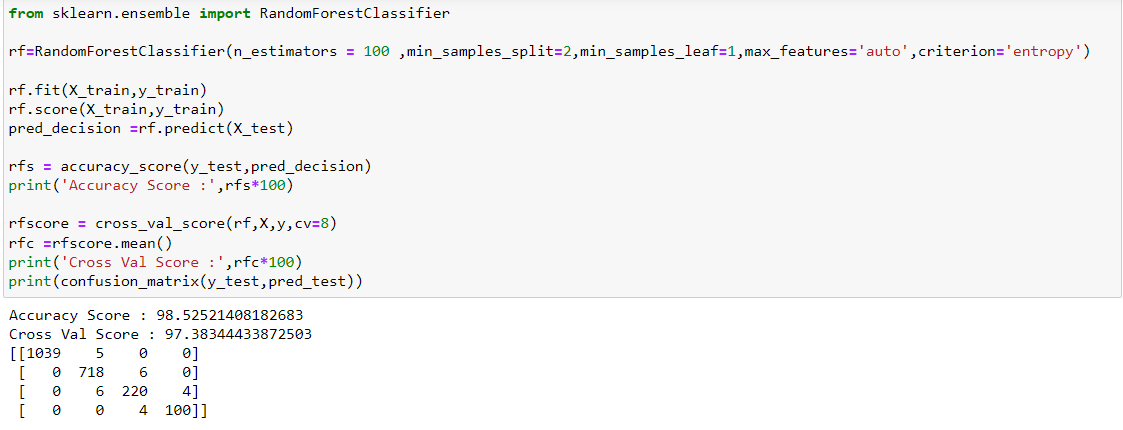




We see that this model is really good like the other models we tested where the train score is 96% and the test is 95.77% F1 score is good and precision is 96% which is really good , the cv score is also almost identical to the test score , and the train test is really close which makes this a contender for the best model.

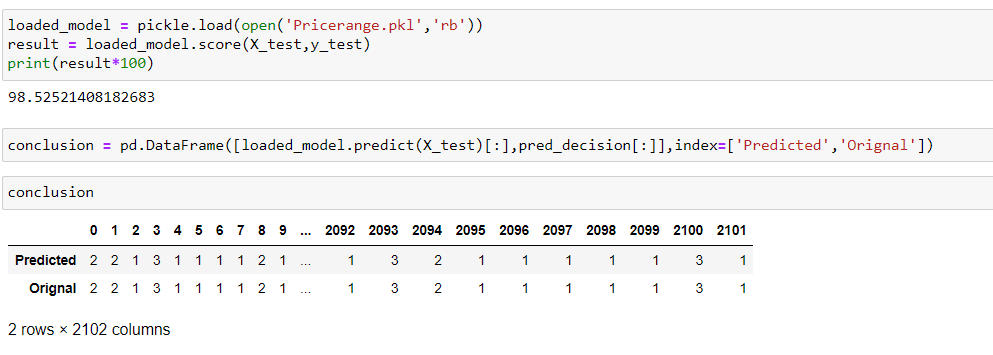
**Hyper Parameter Tuning with Random Forest Classifier**





We see that we are getting similar accuracy score with and without tuning @ 98.53% and the Cv score has actually become better and we have brought it as close as it can to the accuracy score @ 97.38%.

Predictions :



We have saved the model and then used it to predict on the test data creating 2102 entries.

Concluding Remarks

In conclusion, the analysis of Zomato data provided valuable insights into the restaurant industry. By exploring the relationships between various features such as aggregate rating, cuisine type, and price range and average price and so on, we were able to identify patterns and trends that can be useful for both restaurants and customers.

One key finding was the strong influence of the features Has Table booking , Votes and Aggregate rating on Price range, which gave us a very good model score of 98% with Random Forest Classifier,

From the Data Analysis we see that Indian users are not rating as much as the rest of the world as per the data and if this increases by increasing customer engagement, The restaurants in certain areas consistently receiving higher ratings and more bookings will definitely get the value they deserve and the users or customers can have the best dining experience based on their preferences.

We also saw the sheer dominance of Indian transactions in the dataset which suggests that the location of a restaurant can be a crucial factor in its success, and that restaurants in prime locations may have a competitive advantage.

Overall, this project demonstrates the value of data science in the restaurant industry and highlights the potential for using machine learning techniques to uncover patterns and trends in large datasets. By applying these techniques, restaurants can make more informed decisions about their operations and better understand the needs and preferences of their customers.