A..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 to 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.

B..DATA ANALYSIS

First, we have to import necessary libraries like pandas, NumPy, matplotlib, seaborn,

Then we will import datasets

In this dataset there are 9551rows and 21 columns.

After merging the country code data set now there are 22 columns in the dataset.

C..EDA Concluding Remarks

No duplicates values are present in the dataset.

There are numbers of unique values present.

No null values are present.

We will show the no null values on heatmap for better understanding.

While doing univariate analysis Whose are having 1 as country code they are having more count.

We can see, Price range of 1 has maximum price range followed by 2.

All Price range are decreasing in order.

We can see, Indian Rupees is most used currency. It means that Zomato has main business in India followed by Dollar.

Remaining all has no such used, it means Zomato not functioning in that country as Zomato does in India.

We can see Zomato has not provided table booking. But in some places, it provided table booking. May be their Zomato own its Restaurant.

Most places have no online delivery available.

But in few places has available online delivery.

we can see, India has 98.9% online delivery available, but UAE has only 1.1% online delivery.

But the rest of the countries don’t have online delivery.

Most of the places do not have an available delivery Now system.

Most of the customers have given 0 aggregate rating than others.

* 1. and 3.2 is the most given rating after a 0-aggregate rating.

We can see, orange is the most used color for rating followed by white.

Red colors are less used for rating. It means no one dislikes the food delivered by Zomato.

In rating text, Average is the most used test followed by good.

But some of the customers are not rated the Zomato service.

Very few customers give poor ratings.

As we observed before that Indian currency is most used. It because of in India is the most of Zomato business do follow by USA.

Rest of all countries Zomato not gives service as in India they give.

68.7% Zomato works in New Delhi followed by Gurgaon.

In this we can observed here these cities are comes in Delhi NCR Region. It means that Zomato mostly works in Delhi NCR region.

We can see, 26.58% people or customers like to eat north Indian cuisine followed by Chinese.

These are the top 10 Cuisine like by customers.

These are the top 10 places where most of Zomato alliance restaurants are located.

In which Connaught Place is top among the others it has 13.74% restaurant in it place followed by Rajouri Graden.

We can see 50% of the restaurants are affordable.

We can see, Indonesia has high Average cost for two. It means that Indonesia has high cost of foos than others.

Rest of all having low average cost for two.

These are the top 10 restaurant names where the average cost for two is lower than others.

In which Satoo hotel has high average cost for two followed by Skye.

These are top cities where high average cost has for two in which Jakarta is in to followed by Tangerang.

You can observe that the median number of votes for both categories vary.

Restaurants accepting online orders tend to get more votes from customers as there is a rating option popping up after each order through Zomato application.

These are the top 10 restaurant names where most of the customers give good aggregate ratings.

In which Atlanta is top among the others.

The average cost for two has not affected by online delivery system.

We can see, top 10 restaurants based on the customers’ votes.

Toit is the top restaurant top among the others whose have top votes followed by Hauz Khans.

It does not have any linear correlation of both features.

Most of the restaurants are above 50 Longitude.

We can see, both features is not have any linear relationship with each other’s.

* When rating is between 4.5 to 4.9-----> Excellent
* When rating is between 4.0 to 3.4-----> Very good
* When rating is between 3.5 to 3.9-----> Good
* When rating is between 2.5 to 2.9-----> Average
* When rating is between 2.0 to 2.4-----> Poor
* And a 0 rating was given by many people.

We can see, by latitude The Kitchen has highest latitude. It means at high latitude criteria, it is top.

Yes, you can observe that median number of votes for both categories vary. Restaurants accepting online orders tend to get more votes from customers as there is a rating option popping up after each order through Zomato application.

We can see all top 10 restaurant names, where they are generally lies in nearly same latitude.

Many restaurants have as price range 3 it means that those are also having high average price range.

it may be directly proportional to each other’s.

Whose have as rating text Excellent those restaurant are having high average cost for two.

We can see, most of the places have no online delivery option. or online order system. But Malviya Nagar is top among the restaurants whose having online ordered or online delivery.

We can see Zomato has not provided the table booking. But in some place it provided table booking. May be their Zomato own its Restaurant.

Most places has no online delivery available. But in few places has available online delivery.

We can see those are not providing online delivery they also have less table booking but it doe not have any impact on business because most of the restaurant has not provided online delivery.

According to the rating test the average cost for to is dependent.

We can see, many restaurants are having Excellent as rating text and their Aggrege=ate rating also high it mean more than 4.5.

Similarly, Rating text is depending upon the aggregate rating. Those are having good rating their rating text also good.

We can see, most of the Indian customers are gives the 0 rating.

It means that may be service is bad or customers are not interested in gives any rating.

We can see, most of the places have no online delivery option. or online order system.

But Malviya Nagar is top among the restaurants whose having online ordered or online delivery.

We can see, those are not providing online delivery they also have less table booking but it doe not have any impact on business because most of the restaurant has not provided online delivery.

We can see, As per price range the rating color has change. In Which the Orange rating color has more count than others.

Votes has no linear relationship with Average cost for two.

But up to 2000 Votes Average cost has some linear relationship with votes.

We can see, Indonesia has high Average cost for two. It means that Indonesia has high cost of food than others. Rest of all having low average cost for two.

There is no direct linear relationship with each others.

But After 2 rating there is little linear relationship.

As we observed before that Indian currency is most used. It because of in India is the most of Zomato business do followed by USA. Rest of all countries Zomato not gives service as in India they gives.

68.7% Zomato works in New Delhi followed by Gurgaon. In this we can observed here these cities are comes in Delhi NCR Region. It means that Zomato mostly works in Delhi NCR region.

We can see, 26.58% people or customers like to eat north Indian cuisine followed by Chinese.

Those restaurant not accepted online ordered they also have good rating.

we can see, Indonesia has high average cost for two and they also not available of online delivery. I means they are not having online order system. rest of all are having low average cost foe two and they may not have any online system or may have online delivery.

We can see, very few restaurant has accepted the online delivery it mean online orders.

here those restaurant not accepted online ordered they also have good rating.

Not Rated count is very high.

Maximum number of rating are between 2.5 to 3.4.

we can see, Indonesia has high average cost for two and they also not available of online delivery. I means they are not having online order system.

The rest of all are having low average cost foe two and they may not have any online system or may have online delivery.

I observe there is no linear relation between price and rating. For instance, Restaurants with good rating (like 4–5) have restaurants with all the price range and spread across the entire X axis.

We can see, many restaurant are having more than 4 rating and they also have online delivery.

But many restaurant are don’t have online delivery and they have good rating.

Difference between 75% and max in Average Cost for two, Evaporation and Votes is high, columns is considerable indicating presence of outliers.

In some features Mean to Standard of deviation difference is close.

In Average cost for two has mean and std difference is more.

Price range has 8% positive correlation with target column.

Votes has 7% positive correlation with target column.

Country code has 97% positive correlation with country.

we can see, Indonesia has high average cost for two and they also not available of online delivery. I means they are not having online order system. rest of all are having low average cost foe two and they may not have any online system or may have online delivery.

I observe there is no linear relation between price and rating. For instance, Restaurants with good rating (like 4–5) have restaurants with all the price range and spread across the entire X axis

Longitude has 77% negative correlation with country.

Has Table Booking has 50% positive correlation with Price range.

Price range has 44% positive correlation with Aggregate rating.

Country code has 77% positive correlation with Longitude.

Outcome of the skewness: Skewness threshold is taken is +/-0.65.

Almost all features are having some skewness.

It is observed that Price range, Votes and Aggregate rating have the highest positive correlation with Average cost for two.

Switch to order menu, Longitude and rating color have the highest negative correlation with Average cost for two.

Restaurant Name, Address, Rating color, Rating text, Has online delivery, Cuisine and Locality are not having Outliers.

All features are having some skewness.

Average cost for two and votes are seem to be present a outliers.

Some feature are having outliers but those are categorical features.

We use z score method for removing outliers.

After that we will check multicollinearity.

If VIF > 10, It means multicollinearity is present.

Multicollinearity exists in Country Code, Switch to order menu and Country Based on ANOVA F scores.

Now we will transform the dataset through power transform.

Then we will scale the training data.

D..BUILDING MACHINE LEARNING MODELS.

For building machine learning models first we have to split dataset into test and train data and after that we will import models like linear regression, decision tree regressor, random forest regressor, gradient boosting regressor ,SVR , kneighbors regressor , randomized search CV, import metrics like mean absolute error, mean squared error.

**Based on comparing Accuracy Score results with Cross Validation results, it is determined Decision Tree Regressor is the best model. It has least difference between accuracy score and cross validation.**

**After determining the best model we will do hyper parameter tuning to find out the best parameters. And then apply these parameter on data .**

**After do all the analysis we will dump the data into pkl file.**