Food Cost Prediction

# 

# Problem Definition

We all love food. And we all crave for one of our favourite foods from a special restaurant that we all love to have at least once in a month. However, one strong factor that make us reconsider going back to that special restaurant is **cost**.

Data has been collected from various sources, which includes the price information of thousands of restaurants across India. We will predict the cost of a meal for different restaurants across the country based on various features.

The data consists of the following features.

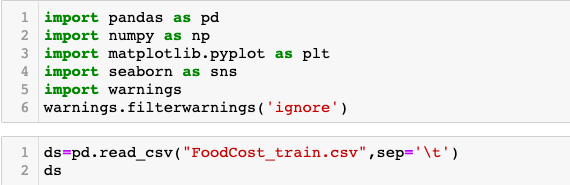
|  |  |
| --- | --- |
| **Description** | **Records held** |
| Size of training set | 12,690 |
| Size of test set | 4,231 |

Dataset has the following Columns/Features :

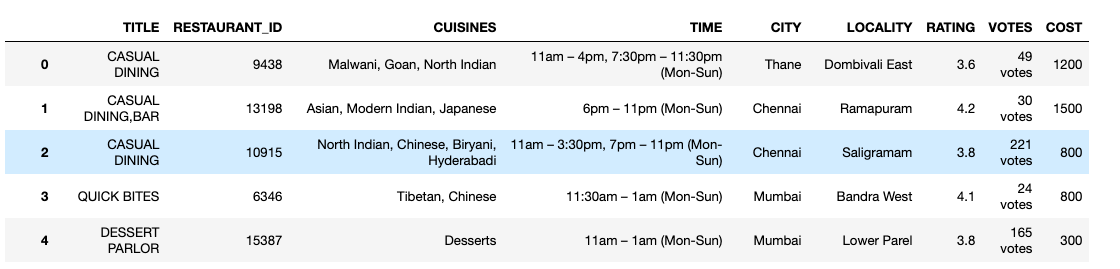
| **Column Heading** | **Significance** |
| --- | --- |
| TITLE | The feature of the restaurant which can help identify what and for whom it is suitable for. |
| RESTAURANT\_ID | A unique ID for each restaurant |
| CUISINES | The variety of cuisines that the restaurant offers |
| TIME | The open hours of the restaurant |
| CITY | The city in which the restaurant is located |
| LOCALITY | The locality of the restaurant |
| RATING | The average rating of the restaurant by customers |
| VOTES | The overall votes received by the restaurant |
| COST | The average cost of a two-person meal |

1. **Explanatory Data Analysis**

To start with, importing all dependencies and the dataset to get the information about the subject.

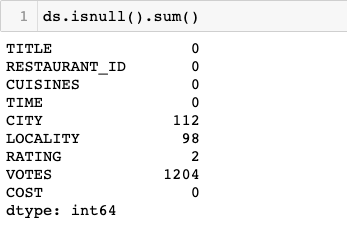


By using the above code, following data will appear on screen to have a sense of the information as per records held.



It is clearly visible that some data cleaning needs to be done prior to any modelling process.

Let’s start by looking at the number of missing values in this training dataset. The details of missing values in the records with respect to different entities is as appended below.

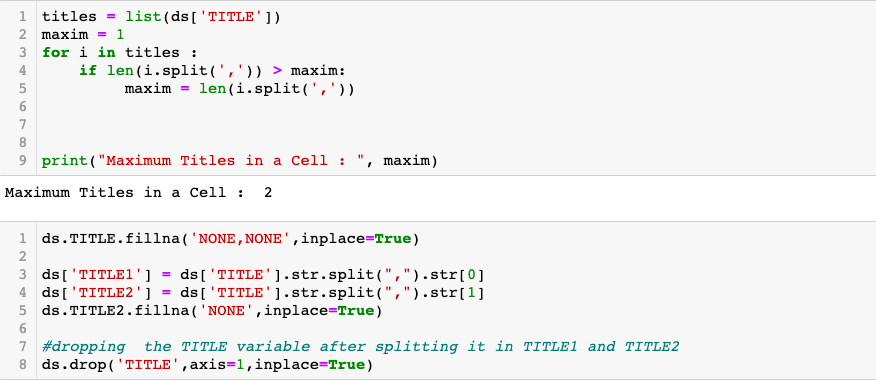


The image above showed the missing values such as CITY, LOCALITY, RATING and VOTES column. I am now going to deal with these missing values individually starting from column 1.

In addition, I will do data cleaning for removing/correcting erroneous entries entered in different columns as given below:-

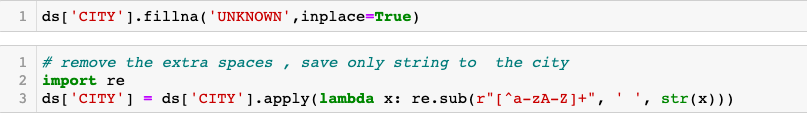
# TITLE

This column has comma separated values. So, I will split those values by making two distinct columns for TITLE and will remove the earlier column with erroneous value. In addition, I will replace all null values with NONE.



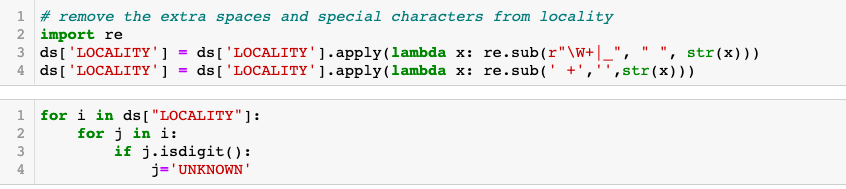
# CITY

This column has extra spaces in string. Also, It contains numerical values like pin code. So, I will only extract string values of the column.



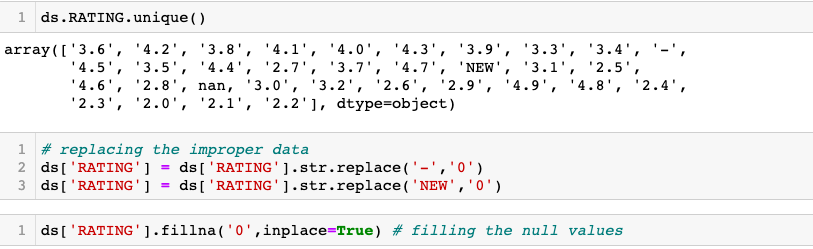
1. **LOCALITY**

This column has extra spaces and special characters in the values. I will remove all of them and will replace null values and numerical values with ‘UNKNOWN’



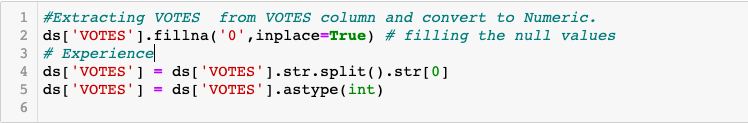
# RATING

This column has 3 irrelevant entries. I will explicitly remove them and will replace null values with 0.



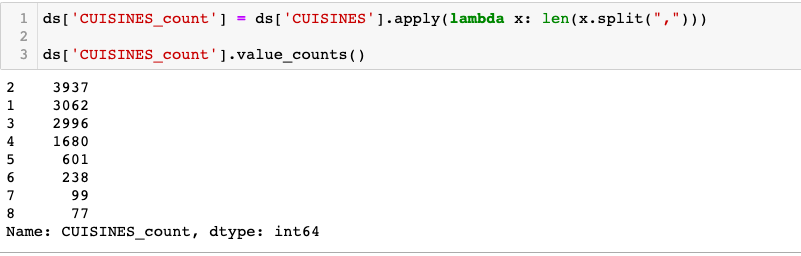
# VOTES

This column has alphanumeric values. So, I will replace null values with 0 and extract votes count by leaving the string ‘Votes’.

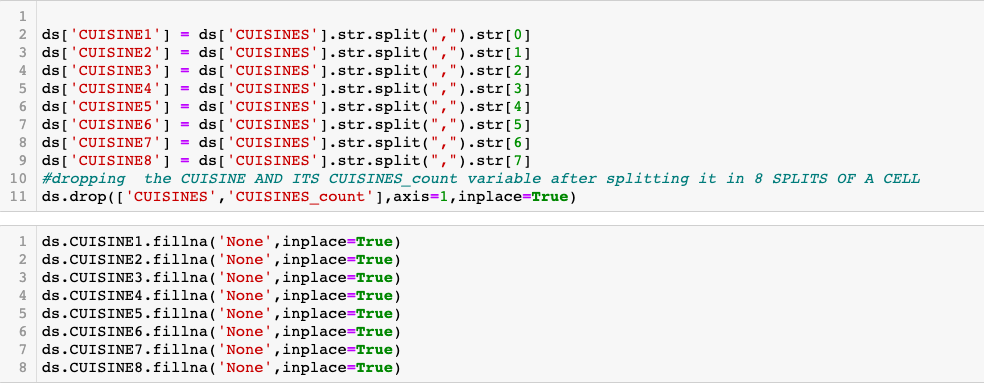


# CUISINES

This column has max 8 cell in a string i.e count of unique CUISINES .



For that, I will make separate 8 columns for all cuisine names. I will replace the null values with ‘NONE’. And then will drop the original column CUISINES.



1. **TIME**

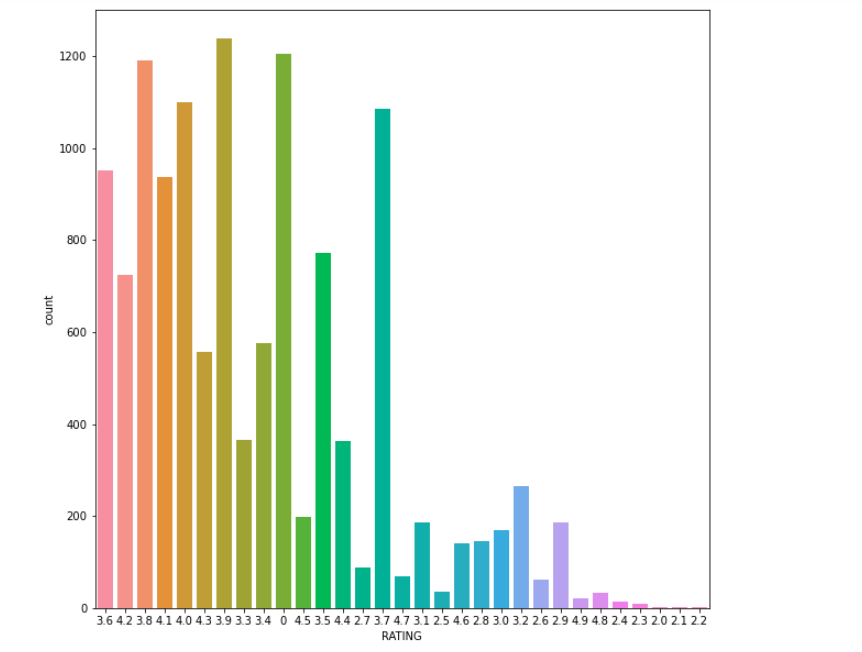
This column has not much relevance in predicting the cost. So, I will drop this column.



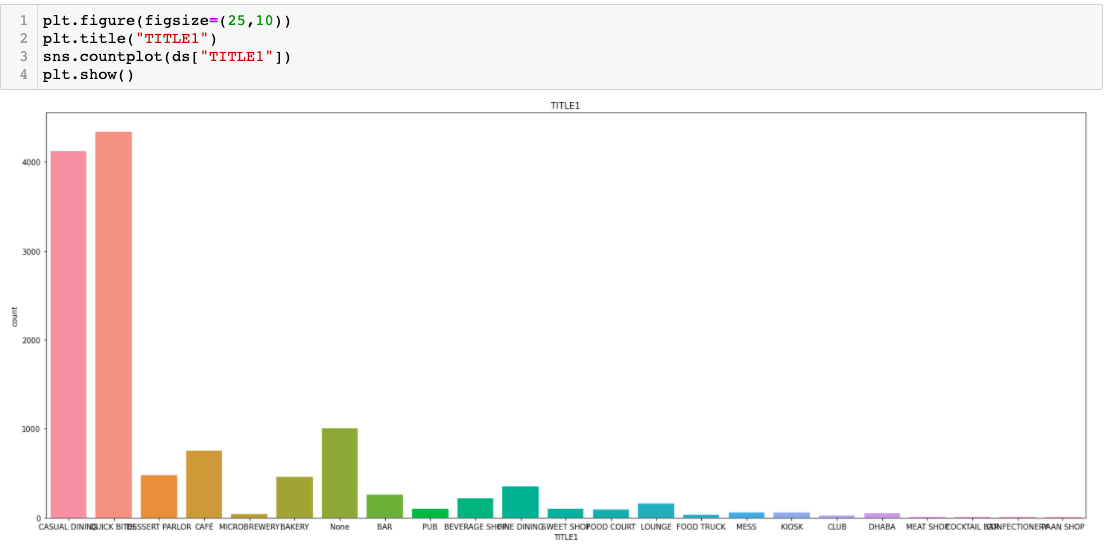
# C. Univariate and Bivariate Analysis

Below shown count plot will display the count of all given ratings

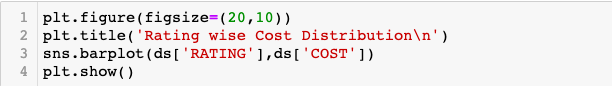


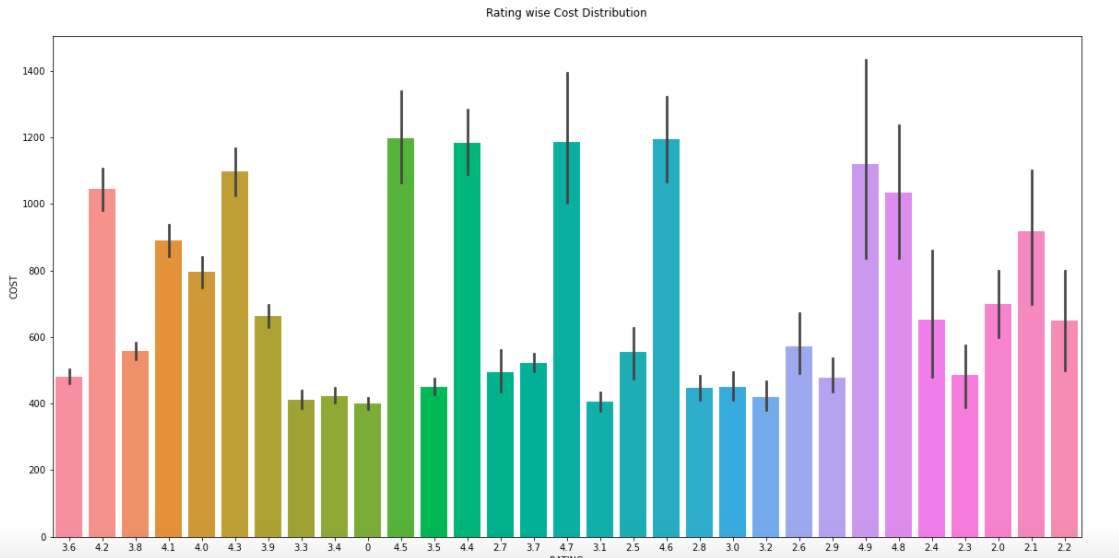


Count plot that will help in knowing the favourite TITLE among all are as shown below:-



Bar plot which will give us a better view of Rating wise Cost distribution is as shown below:-





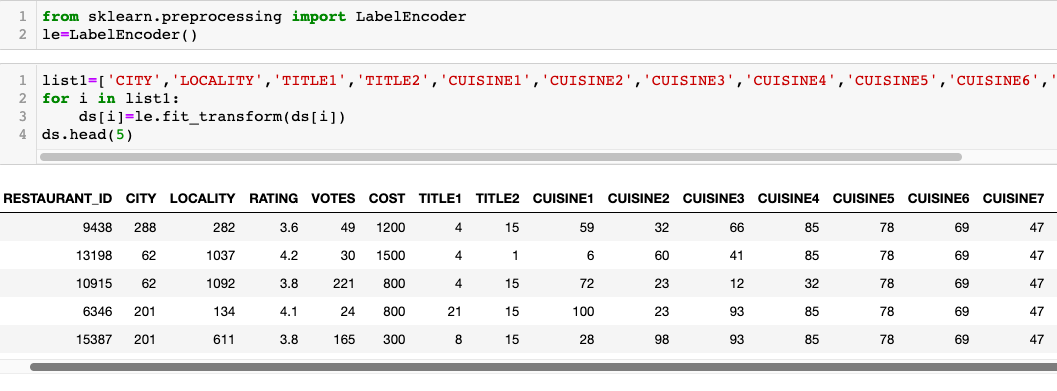
**D.**  **EDA Concluding Remarks**

1. Cost is high where the ratings are above 4.0
2. Max count is concentrated in Casual Dining and Chick Bites in TITLE1
3. Max of ratings given by the customers is in the range of 3.5 - 4.0

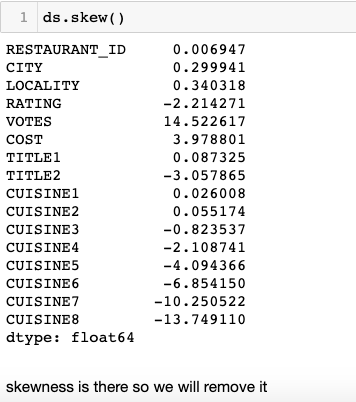
**E. Pre-Processing Pipeline.**

As it is shown in the dataset, only votes and cost are two numerical fields. Both may vary in all the restaurants. So, I decided not to remove the outliers.

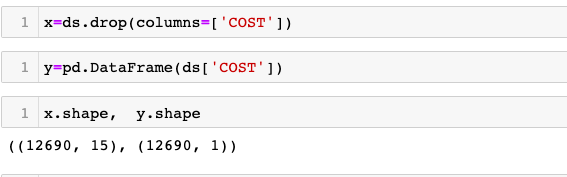
Also, Maximum columns have the nominal values. For moving further for model training, we need to transform the nominal values into numerical values by encoding the data.



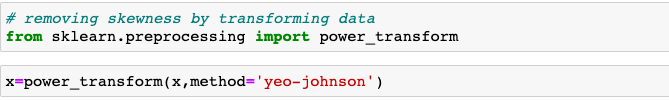
Now, the dataset is all numerical. I will check the distribution of skewness in the data. If it is there, I will remove it.



As we remove the skewness from x variables only, first I will split the x and y variable from dataset.



After splitting , we will remove the skewness from the data.



Once skewness removal has been done, I will do scaling of the dataset. **Scaling** is a technique to standardize the independent features present in the data in a fixed range. If feature **scaling** is not done, then a **machine learning** algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.



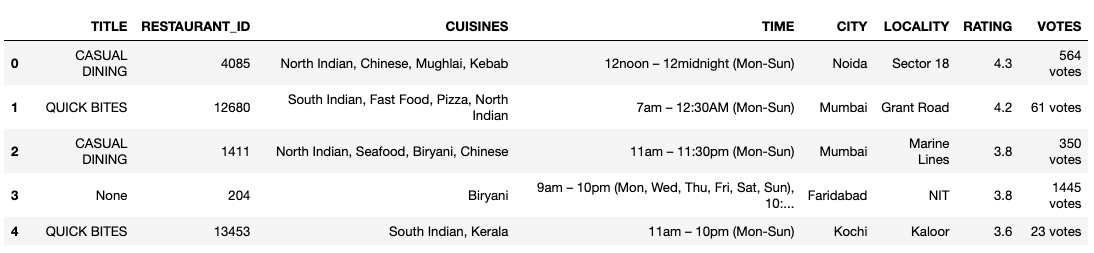
Now, all the **data cleaning and pre-processing** has been done on training dataset. Same process we will follow for testing dataset.

1. **Explanatory Data Analysis**

To start with, importing the dataset to get the information about the subject.

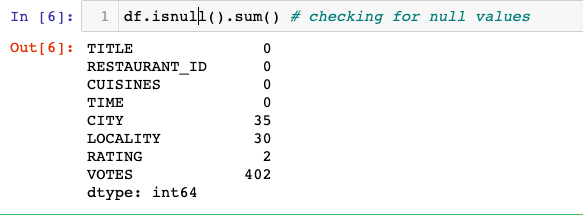


By using the above code, following data will appear on screen to have a sense of the information as per records held.



It is clearly visible that some data cleaning needs to be done prior to any modelling process.

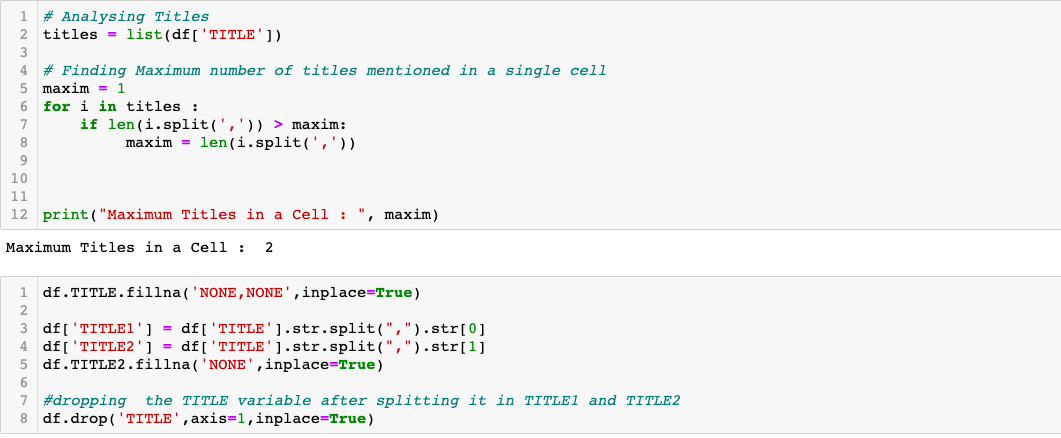
Let’s start by looking at the number of missing values in this training dataset. The details of missing values in the records with respect to different entities is as appended below.



The image above showed the missing values such as CITY, LOCALITY, RATING and VOTES column. I am now going to deal with these missing values individually starting from column 1. In addition, I will do data cleaning for removing/correcting erroneous entries entered in different columns as given below :-

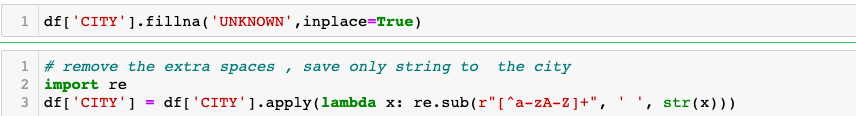
# TITLE

This column has comma separated values. So, I will split those values by making two distinct columns for TITLE and will remove the earlier column with erroneous value. In addition, I will replace all null values with NONE.



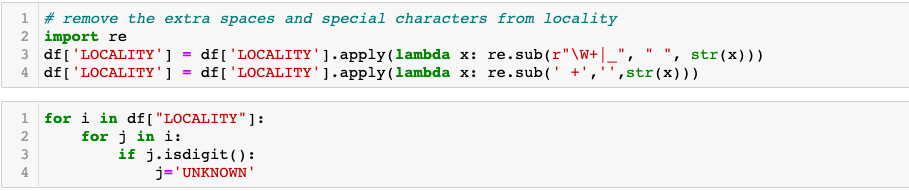
# CITY

This column has extra spaces in string. Also, It contains numerical values like pin code. So, I will only extract string values of the column.



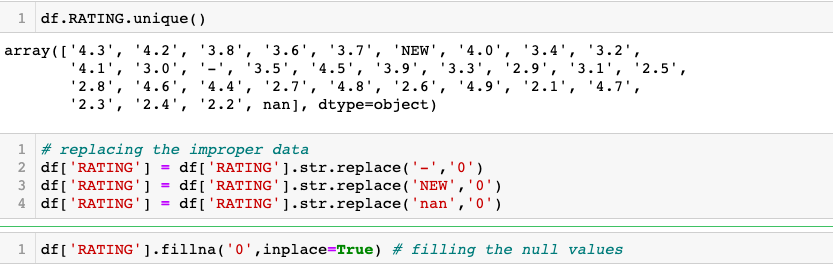
1. **LOCALITY**

This column has extra spaces and special characters in the values. I will remove all of them and will replace null values and numerical values with ‘UNKNOWN’



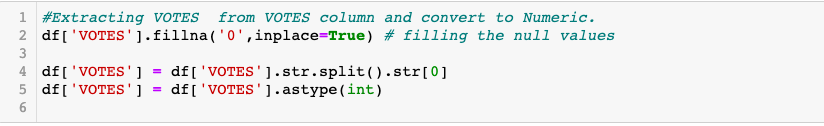
# RATING

This column has 3 irrelevant entries. I will explicitly remove them and will replace null values with 0



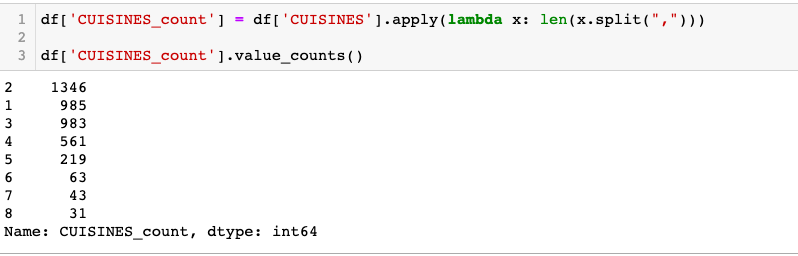
# VOTES

This column has alphanumeric values. So, I will replace null values with 0 and extract votes count by leaving the string ‘Votes’

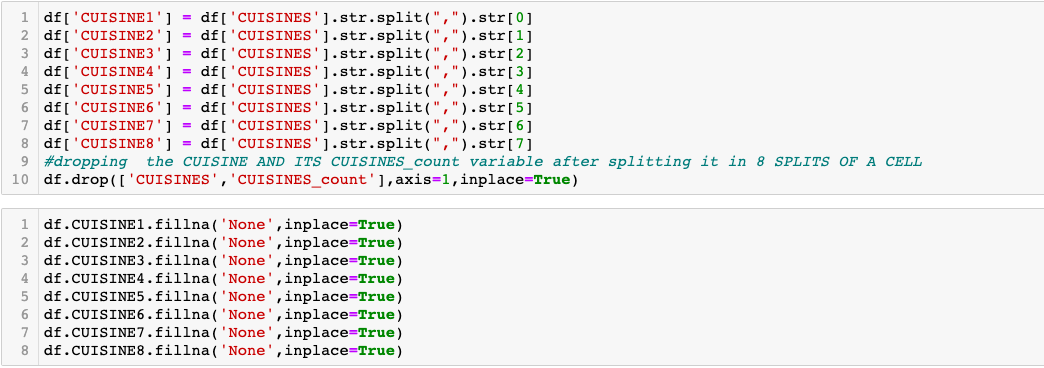


# CUISINES

This column has max 8 cell in a string i.e count of unique CUISINES.



For that, I will make separate 8 columns for all cuisine names. I will replace the null values with ‘NONE’. And then will drop the original column CUISINES.



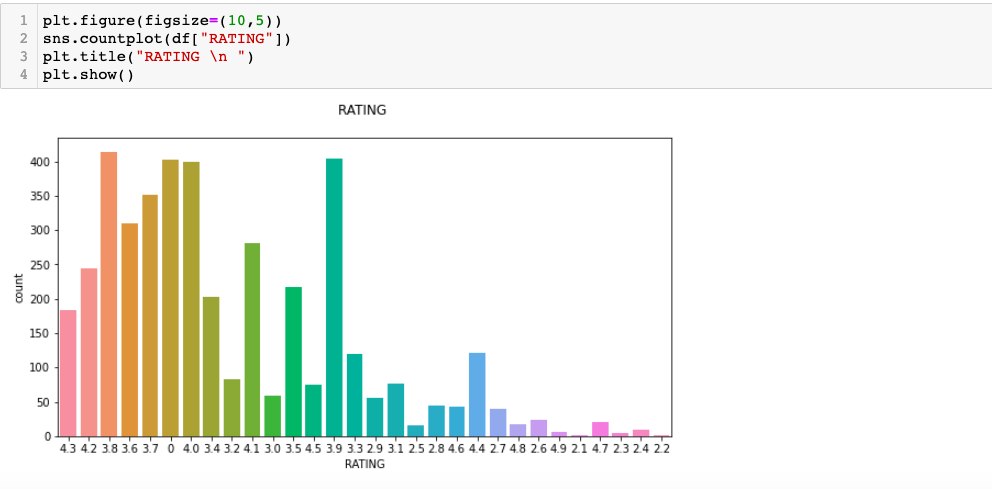
1. **TIME**

This column has not much relevance in predicting the cost. So, I will drop this column.

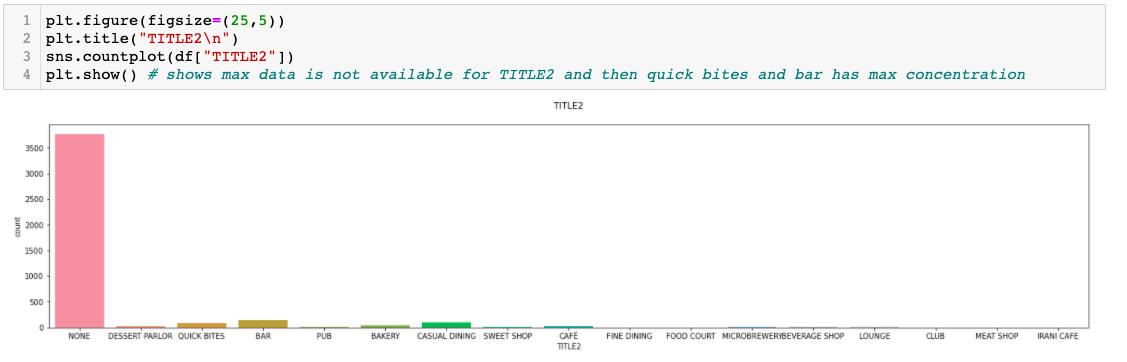


# Univariate and Bivariate Analysis

Below shown count plot will display the count of all given ratings



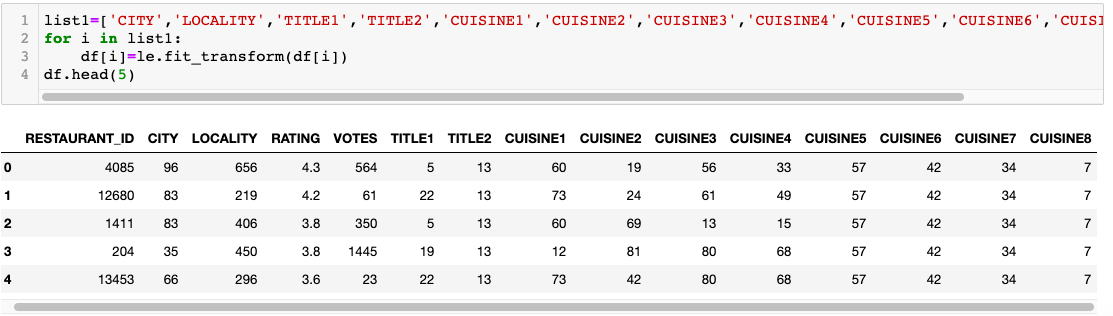
Below shown count plot shows that the maximum number of TITLE2 is not available. Quick Bites , Bar and Casual Dining have max. concentration of data.



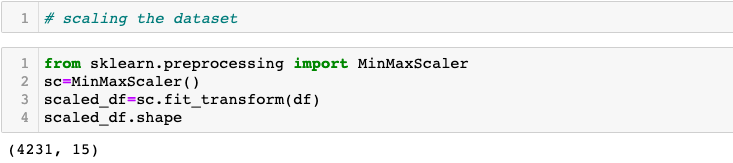
1. **EDA Concluding Remarks**
2. Cost is high where the ratings are above 4.0
3. Max count is concentrated in Casual Dining and Chick Bites in TITLE1
4. Max of ratings given by the customers is in the range of 3.5 - 4.0
5. **Preprocessing Pipeline**

As it is shown in the dataset , only votes is a numerical field. Votes may vary in all the restaurants . So I decided not to remove the outliers. Before implementing the algorithms, we have to encode the categorical variables and scale its features.

Also, Maximum columns have the nominal values . For moving further for model training, we need to encode the data .



For moving further , I will do scaling of the data.

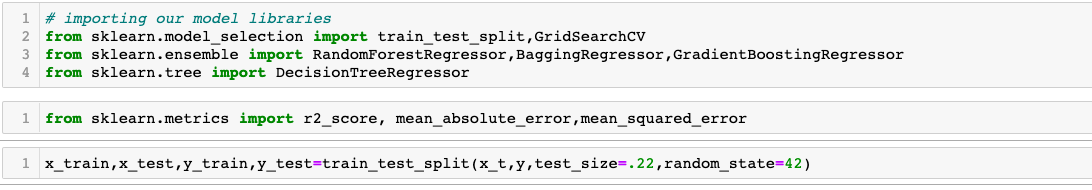


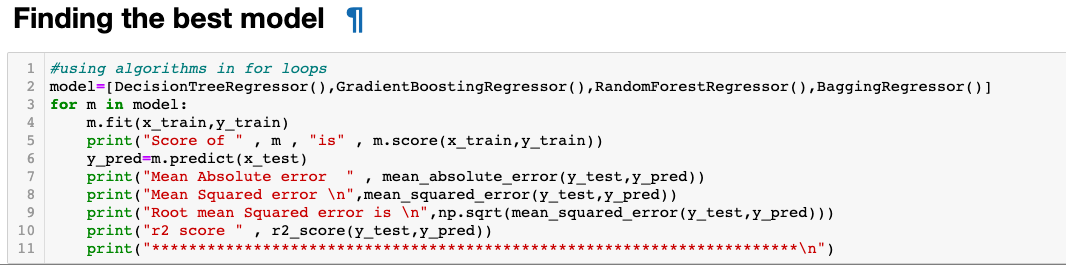
Now , all the data cleaning and preprocessing has been completed . Datasets are now ready for model training .

1. **Building Machine Learning Models**

**Let’s start with importing the dependencies.**

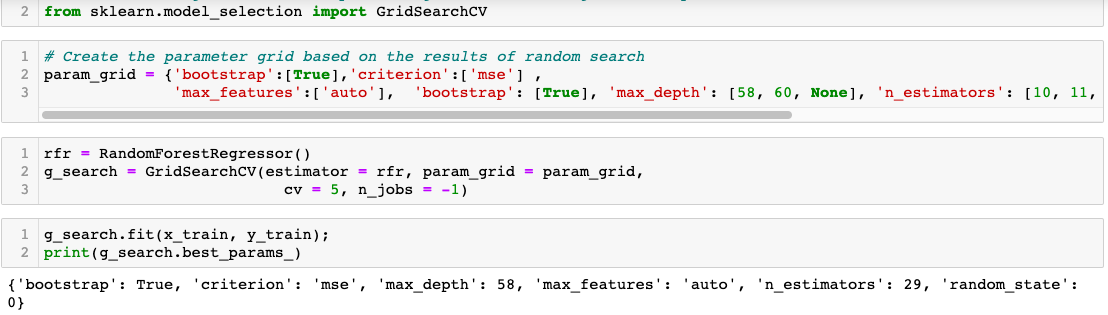
And then let us divide the data in the test and train set.  
In this project, I have divided the data into an 78: 22 ratio. That is, the training size is 78% and testing size is 22% of the whole data.



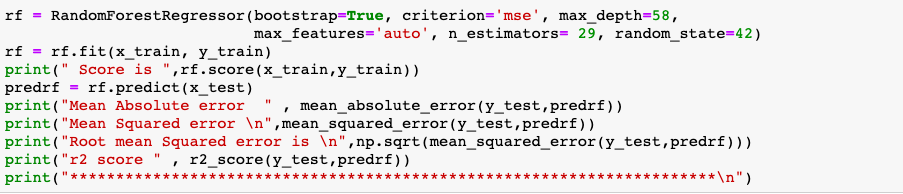


Here, RandomForestRegressor performed the best with score 0.954607657325165

Doing a GridSearchCV is a great way to do hyperparameters tunning

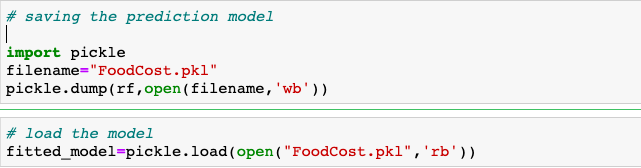


Finally using the best parameters .



The score obtained here is  0.9579788855799336

A slight improvement but an improvement nonetheless. Next step will be to save the best fit model.For that we will do serialization using pickle.dump method and will load the fitted model .



Using the fitted model we will find prediction over the test dataset.



1. **Concluding Remarks**

In this article, I have tried to explain the flow of how we perform machine learning on real time datasets and apply the best algorithm to predict the target. Also, we have seen the correlation between predicted and actual results. I hope you find this article useful. Thanks

All the code can be found in my GitHub [here](https://github.com/piplani-bhavna/ProjectsDSR/blob/DynamicsDSR/Project14_FoodCostPrediction_Modelling%20(1).ipynb)