Predictive Modeling

Business Report

June 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Thakur Arun Singh

Contents

[Problem 1: 2](#_Toc75725474)

[Problem 1.1 2](#_Toc75725475)

[Problem 1.2 13](#_Toc75725476)

[Problem 1.3 20](#_Toc75725477)

[Problem 1.4 23](#_Toc75725478)

[Problem 2: 24](#_Toc75725479)

[Problem 2.1 24](#_Toc75725480)

[Problem 2.2 38](#_Toc75725481)

[Problem 2.3 42](#_Toc75725482)

[Problem 2.4 48](#_Toc75725483)

## Problem 1:

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

### Problem 1.1

Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

**Resolution:**

**Describing the data:**

* First we import all the necessary libraries in Python, and then import the data file which is ‘cubic\_zirconia’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
* Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
* To check the data type of all the columns and also to check the null values, ‘info’ function. Has been used.
* To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,

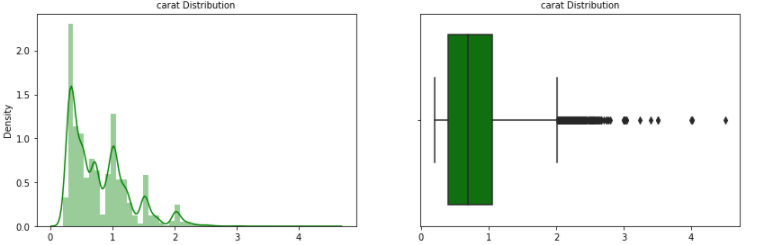
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Unnamed: 0** | 26967 | 13484 | 7784.847 | 1 | 6742.5 | 13484 | 20225.5 | 26967 |
| **carat** | 26967 | 0.798375 | 0.477745 | 0.2 | 0.4 | 0.7 | 1.05 | 4.5 |
| **depth** | 26270 | 61.74515 | 1.41286 | 50.8 | 61 | 61.8 | 62.5 | 73.6 |
| **table** | 26967 | 57.45608 | 2.232068 | 49 | 56 | 57 | 59 | 79 |
| **x** | 26967 | 5.729854 | 1.128516 | 0 | 4.71 | 5.69 | 6.55 | 10.23 |
| **y** | 26967 | 5.733569 | 1.166058 | 0 | 4.71 | 5.71 | 6.54 | 58.9 |
| **z** | 26967 | 3.538057 | 0.720624 | 0 | 2.9 | 3.52 | 4.04 | 31.8 |
| **price** | 26967 | 3939.518 | 4024.865 | 326 | 945 | 2375 | 5360 | 18818 |

* Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
* Using the ‘dups’ function we check for the duplicates and there were no duplicate values.
* We have both categorical and continuous data. For categorical data, we have we have cut, color and clarity. For continuous data, we have carat, depth, table and price.
* We also identified the unique values in categorical data.

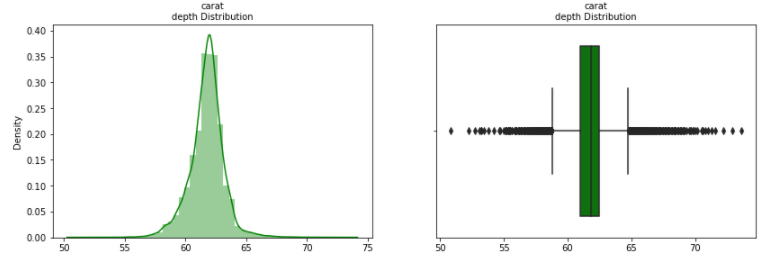
After reviewing the data thoroughly, and based on the above analysis we can say that, we have seven variables, Mean and Median values are almost equal, and Standard deviation for ‘Spending’ is higher than other variables. There are no duplicates in the data set.

**Exploratory data analysis**

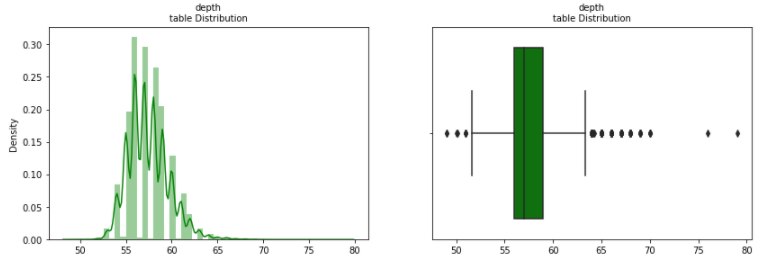
**Univariate and multivariate analysis**



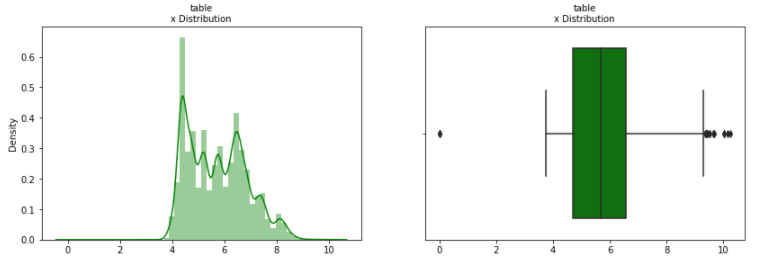
The distribution of data seems to be positively skewed as there are multiple peak points in the distribution there could be multimode and the box plot of carat seems to have large number of outliers .In the range of 0 to 1, where majority of data resides.



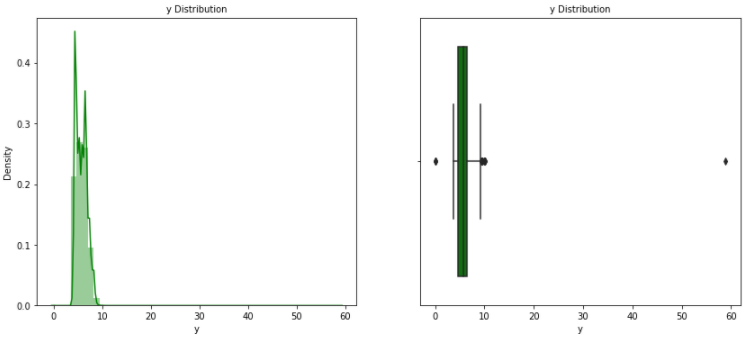
The above dist plot shows the normal distribution of data from 55 – 65. Boxplot shows that there are lot of outliers.



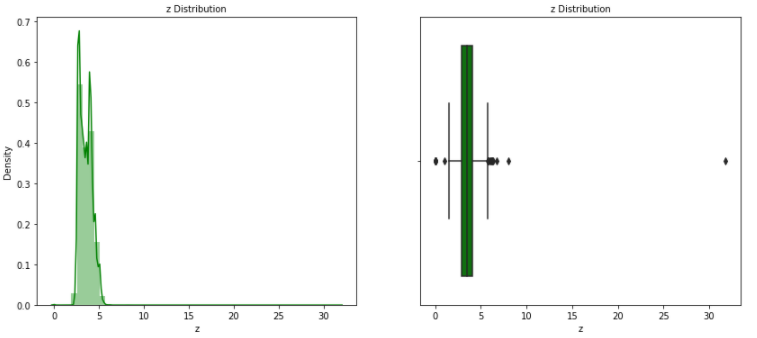
The above dist plot shows the distribution of data from 55 – 65 and is positively skewed. Boxplot shows that there are a few outliers.



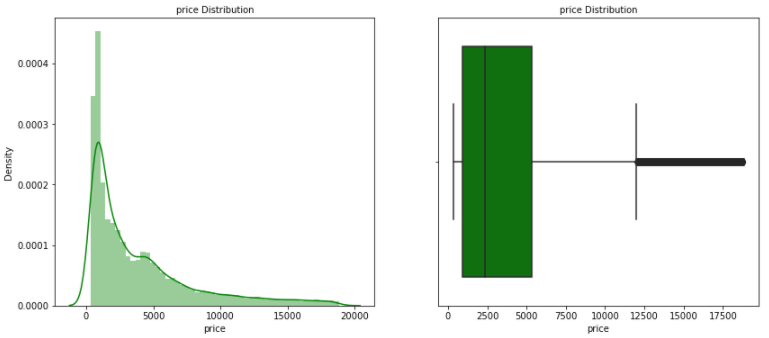
The above dist plot shows the distribution of data from 4 – 8 and is positively skewed. Boxplot shows that there are lot of outliers.



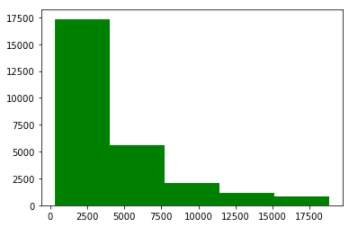
The above dist plot shows the distribution of data from 5 – 10 and is positively skewed. Boxplot shows that there are few outliers. The distribution is too much positively skewed. The skewness maybe due to diamonds are always made in specific shape. There might not be too many sizes in the market.



The above dist plot shows the distribution of data from 2 – 5 and is positively skewed. Boxplot shows that there are a few outliers. The skewness may be due to diamonds that are made in specific shape. There may not be too many sizes in the market.



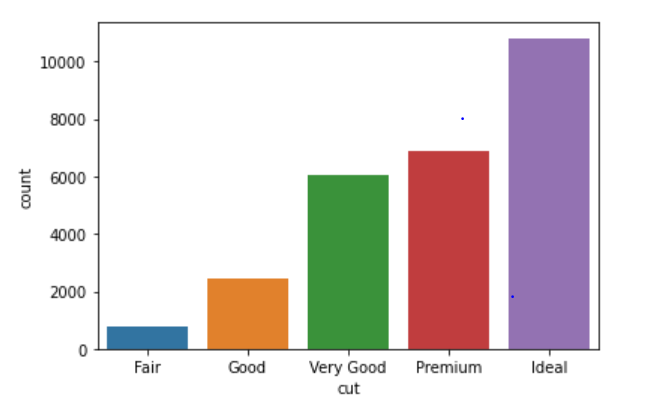
The above dist plot shows the distribution of data from 100 – 8000 and is positively skewed. Boxplot shows that there are a lot of outliers.



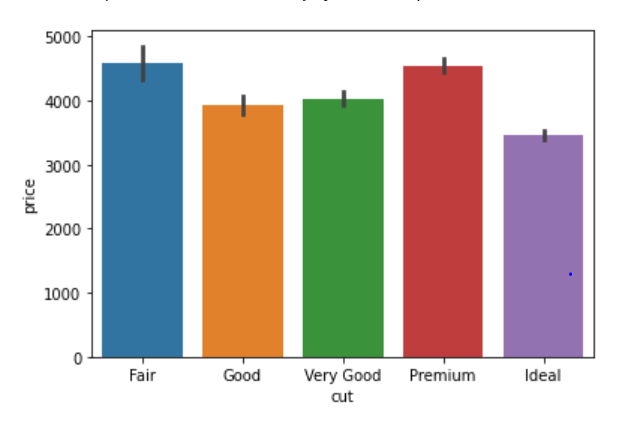
The above histogram shows the price.

**Bi – Variate Analysis:**

Quality is increasing order fair, good , very good , premium, ideal.

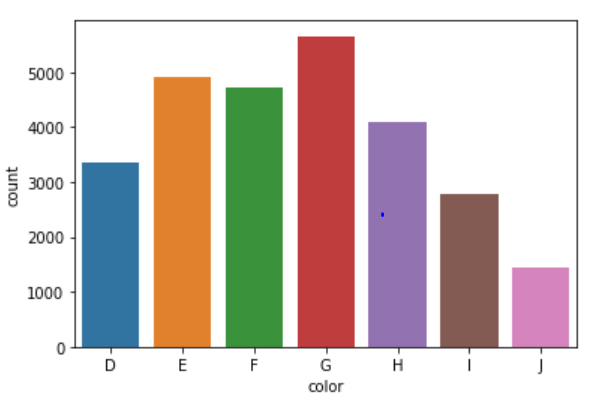


The most preferred cut seems to be ideal for diamonds



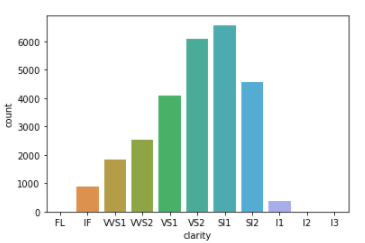
The reason for the most preferred cut ideal is because those diamonds are priced lower than the other cuts.

‘D’ being the best and J the worst.

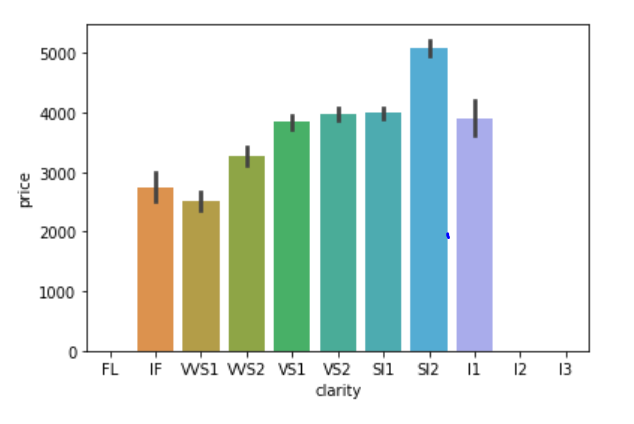


We see the G is priced in the middle of the seven colours, where J being the worst colour price seems to be high.

**Best to worst, FL-> flawless, l3-> level 3 inclusions)FL,IF,VVS1,VVS2,VS1, VS2,SI1,SI2,I1,I2,I3**



the clarity VS2 seems to be preferred by people.

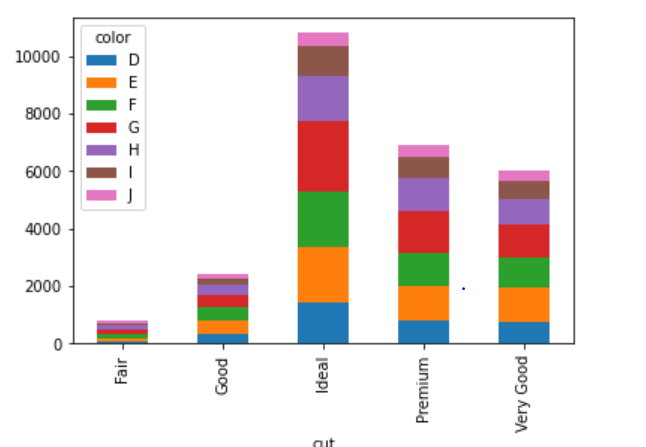


The clarity VS2 seems to be preferred by people.

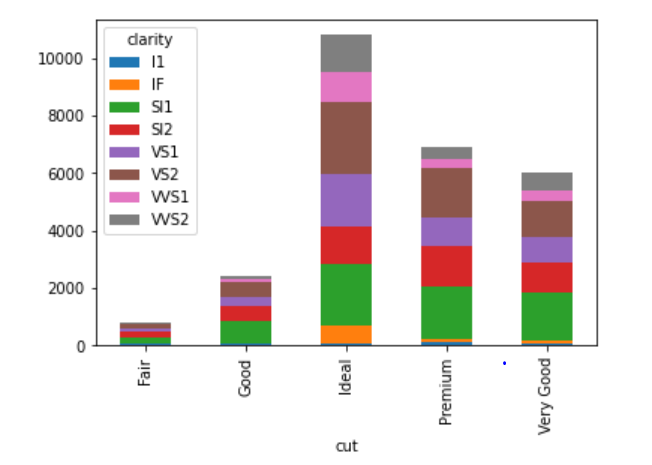
The data has no FL diamonds from this we can determine the flawless diamonds are not bringing any profits to the store.

Relationship between categorical variables.

Cut and color:

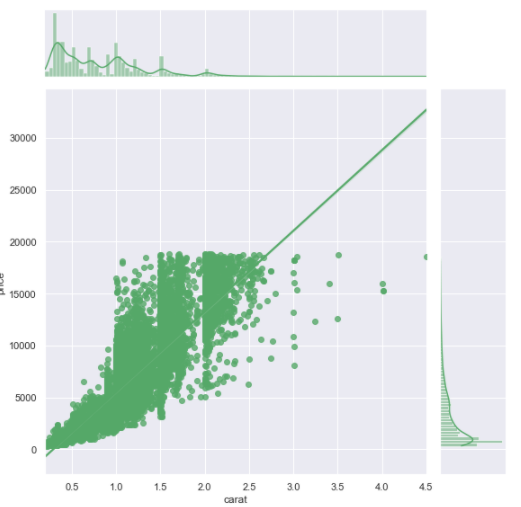


Cut and Clarity

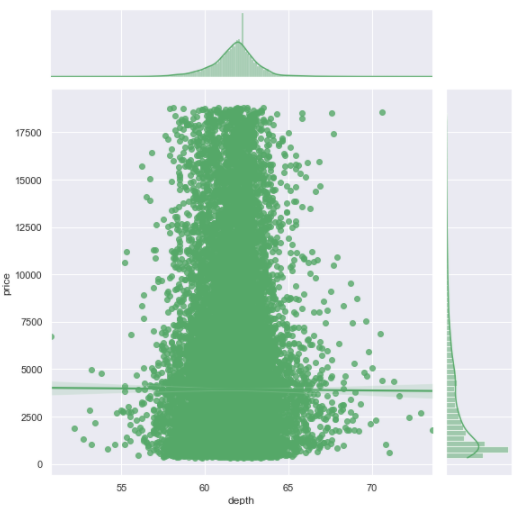


Correlation:

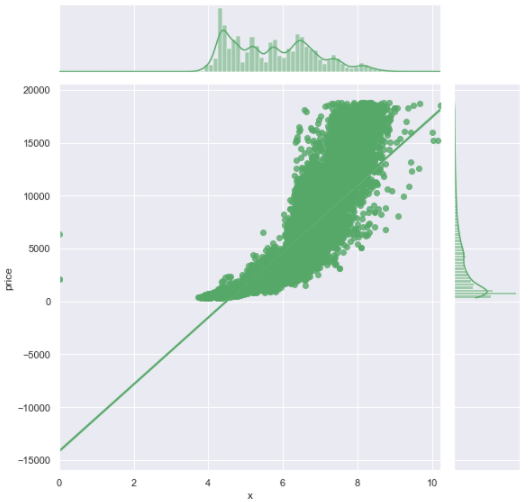
CARAT V/S Price:



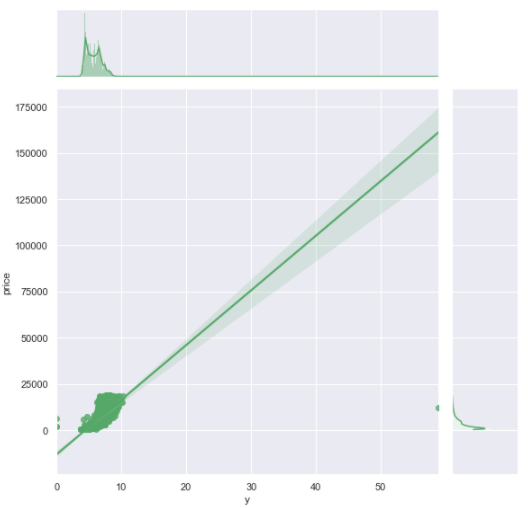
Depth V/S Price



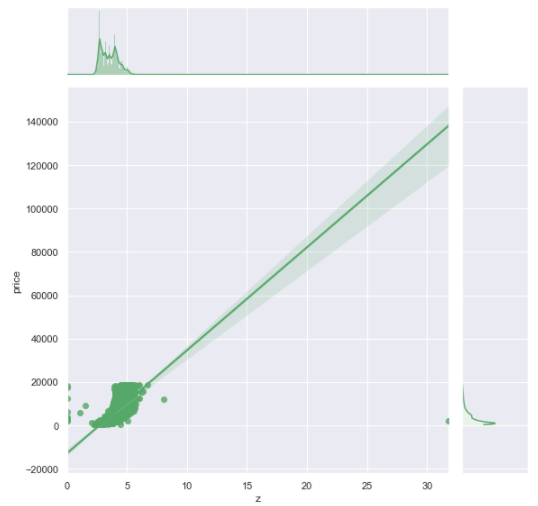
X V/S Price



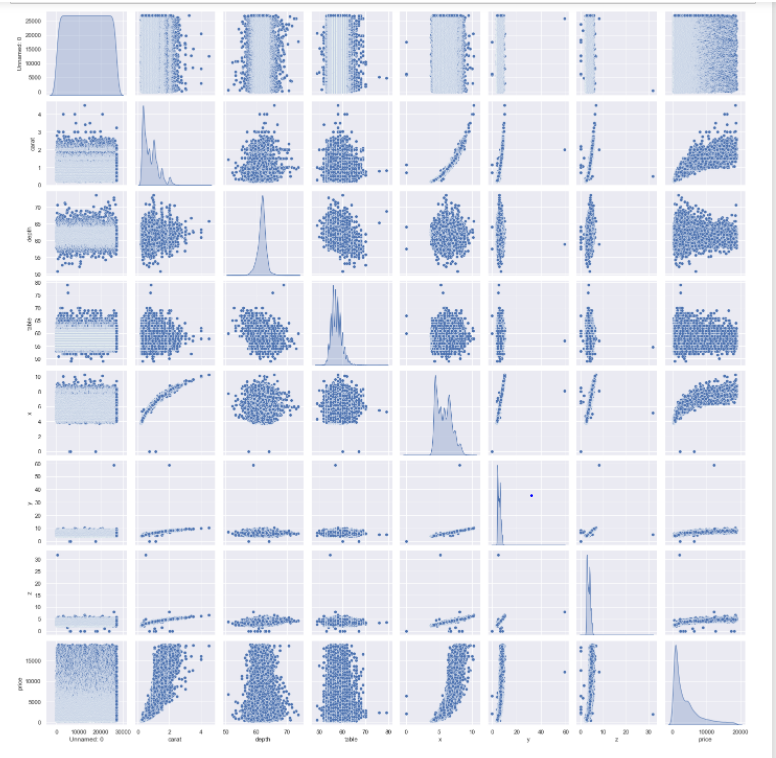
Y V/S Price



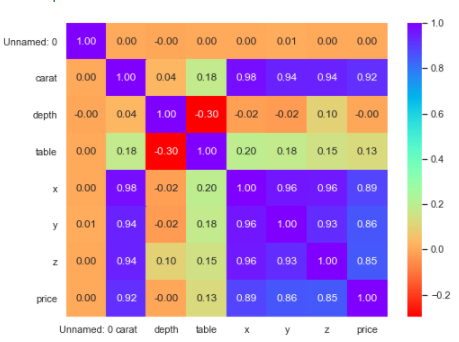
Z V/S Price



Data Distribution



Correlation Matrix



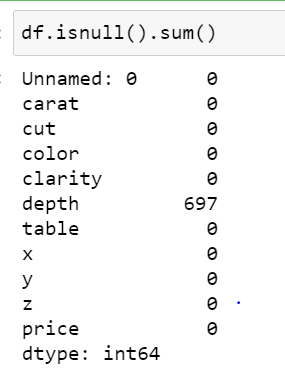
The matrix shows the presence of multi-collinearity in the dataset.

From both the analysis we can see that there is strong positive correlation between all the variables.

### Problem 1.2

Impute null values if present; also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

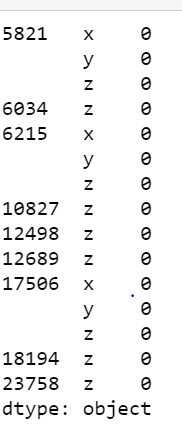
**Resolution:**



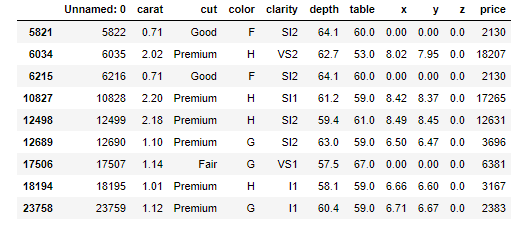
As per the above screen shot, there are null values since depth being continuous, carriable mean or median imputation can be done.

The percentage of null values is less than 5%, we can also drop these if we want.

After median imputation, we don’t have any null values in the dataset.



Checking value that is 0



We have certain rows which has zero the x,y and z are the dimensions of a diamond of a diamond so this can’t take into model as these are very less rows.

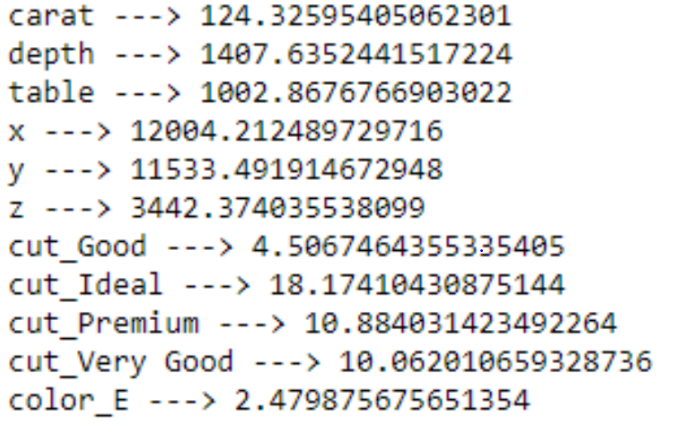
We can drop these rows since they don’t have any meaning in model building.

Scaling:

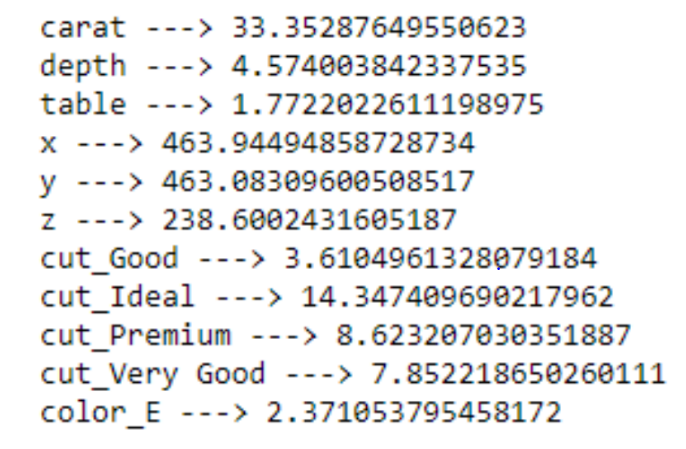
Scaling can be reduced to check the multi-collinearity in the data.so if scaling is not applied we find the VIF- variance inflation factor values which are high, which indicates the presence of multi-collinearity.

These values are calculated after building the model of linear regression. To understand the multi-colliearity in the model. The scaling has no impact in model score or coeffients of neither attributes nor the intercept.

Before scaling:

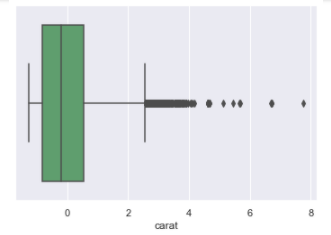


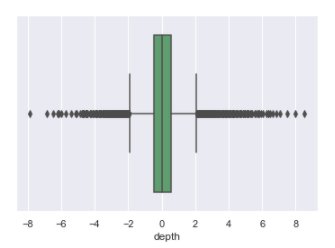
After Scaling

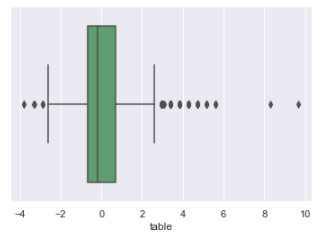


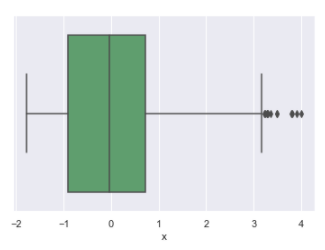
Checking for outliers in the data.

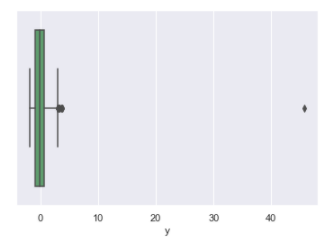
Before treating outliers

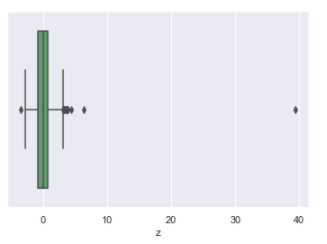


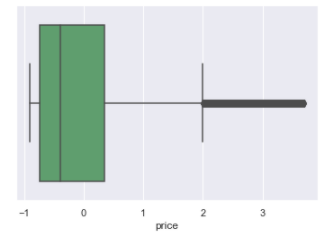




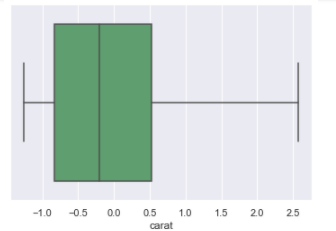


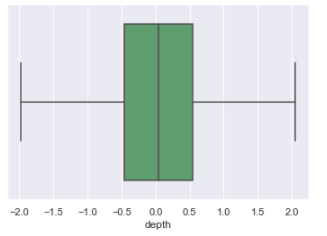


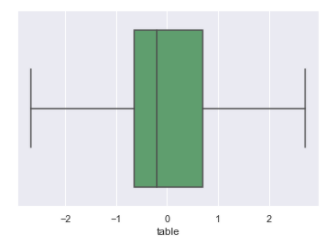


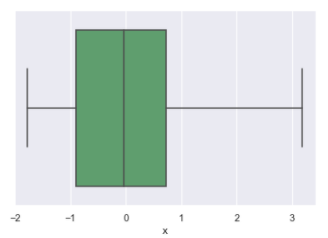


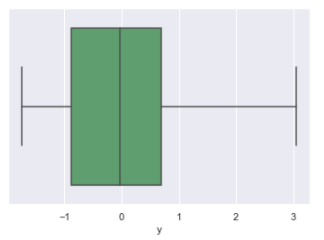
After Treating Outliers

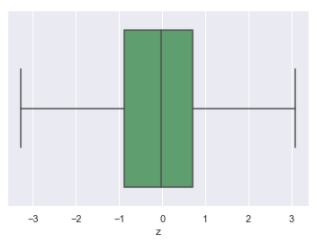


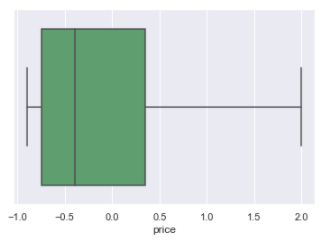




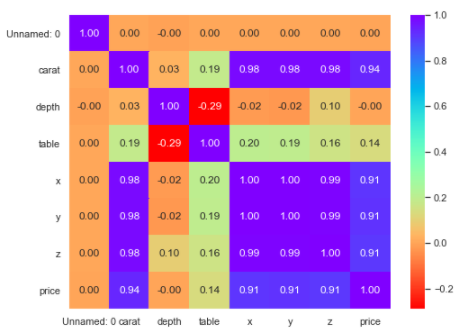








**Correlation**

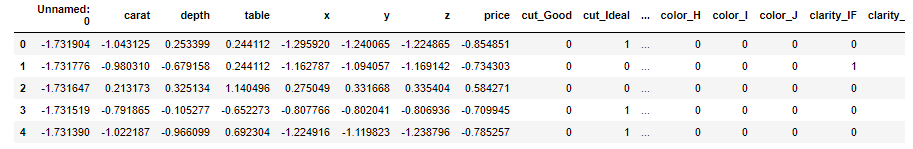


### Problem 1.3

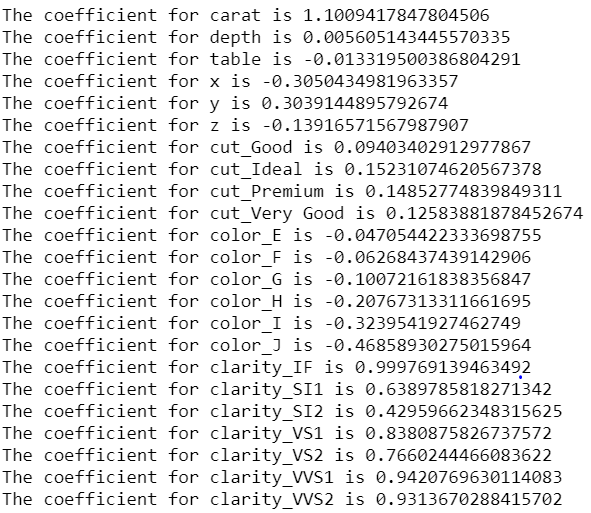
Encode the data (having string values) for Modeling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

**Resolution:**

Linear regression model does not take categorical values so that we have encoded categorical values to integer for better results.

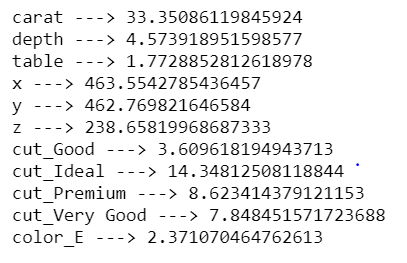


We drop all the unwanted columns and then create Linear regression model.





VIF Values

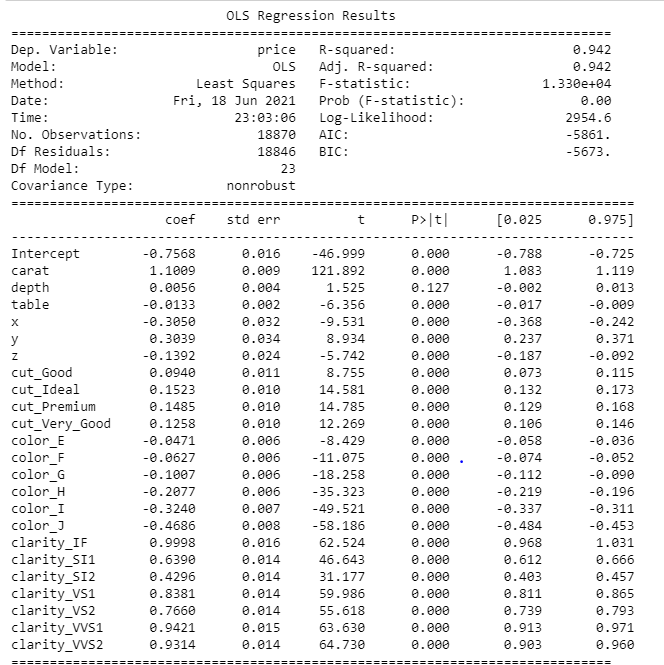


We still have to find multi-collinearity in the dataset, to drop these values to lower values we can drop columns after doing stats model.

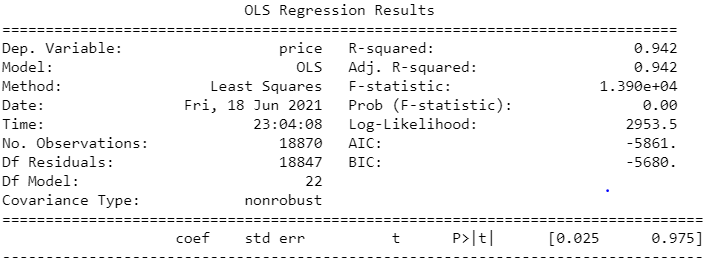
From the stat model, we can understand the feature that do not contribute to the model. We can remove the features after the VIF values be reduced. Ideal value of VIF is less than 5%

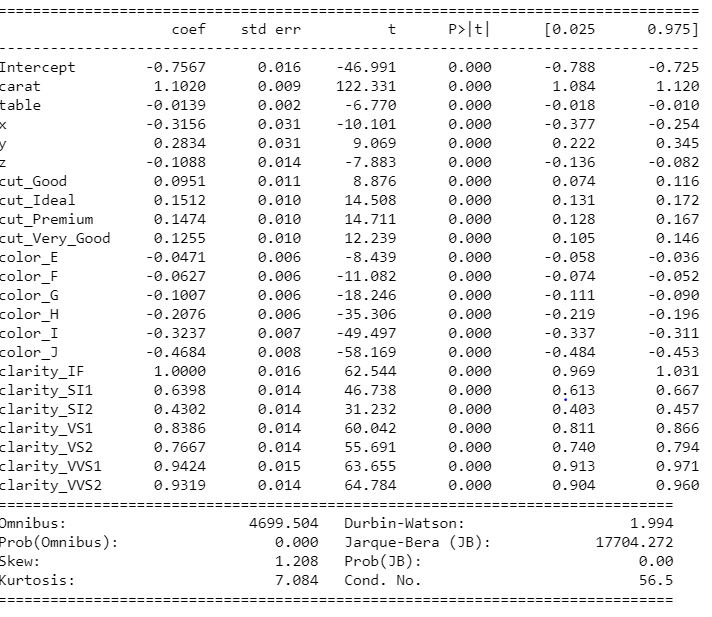
**STATSMODEL**

Best param summary:



Best param summary after dropping the depth variable:





To ideally bring down the values to lower levels we can drop one of the variables that are highly correlated.

Dropping variables would bring down the multi-collinearity level down.

### Problem 1.4

Inference: Basis on these predictions, what are the business insights and recommendations.

**Resolution:**

To predict the price of the stone and provide insights and provide insights for the company on the profits on different prize slots.

* From the EDA we could understand the cut; ideal cut had number profits to the company. The colors H, I, J have bought profits to the company.
* In clarity, if we could see there were no flawless stones and no profits coming from L1, L2, and L3 stones.
* The ideal, premium and very good types of cut were bringing profits where as fair are not bringing profits.

The predictions were able to capture 95% variation in the price and it is explained by the predictors in the training set.

Using stats model, if we could run the model again we can have P values and coefficients which will give us better understanding of the relationship. So that values more than 0.05 we can drop those variables and re-run the model again for better results.

* The ideal, premium, very good cut types are the ones which are bringing profits so that we could use marketing for these to bring more profits.
* The clarity of the diamond is the next important attributes the more the clear is the stone the profits are more.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## Problem 2:

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

### Problem 2.1

Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

**Resolution:**

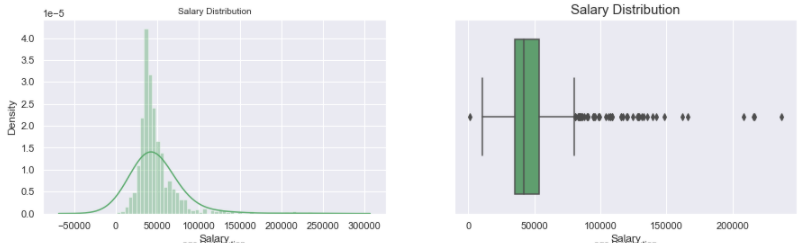
* First we import all the necessary libraries in Python, and then import the data file which is ‘Holiday\_Package’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
* Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
* To check the data type of all the columns and also to check the null values, ‘info’ functions. Has been used.
* To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Unnamed: 0** | 872 | 436.5 | 251.869 | 1 | 218.75 | 436.5 | 654.25 | 872 |
| **Salary** | 872 | 47729.17 | 23418.67 | 1322 | 35324 | 41903.5 | 53469.5 | 236961 |
| **age** | 872 | 39.95528 | 10.55168 | 20 | 32 | 39 | 48 | 62 |
| **educ** | 872 | 9.307339 | 3.036259 | 1 | 8 | 9 | 12 | 21 |
| **no\_young\_children** | 872 | 0.311927 | 0.61287 | 0 | 0 | 0 | 0 | 3 |
| **no\_older\_children** | 872 | 0.982798 | 1.086786 | 0 | 0 | 1 | 2 | 6 |

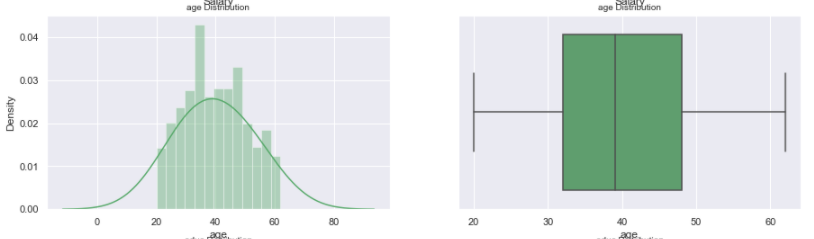
* Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
* Using the ‘dups’ function we check for the duplicates and there were few duplicate values which are noted.
* Using the ‘drop\_duplicates’ function, we can exclude the duplicate values. Then check for the data.

The split indicates that 45% of the employees are interested in the holiday package

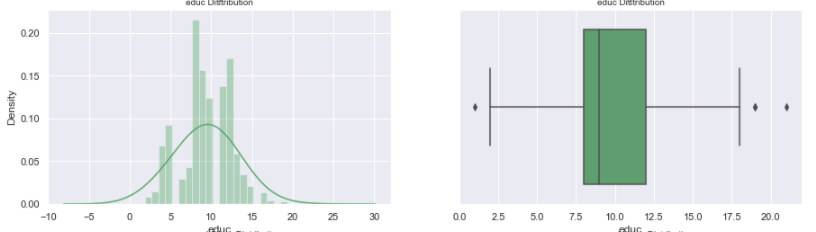
Categorical univariate analysis:



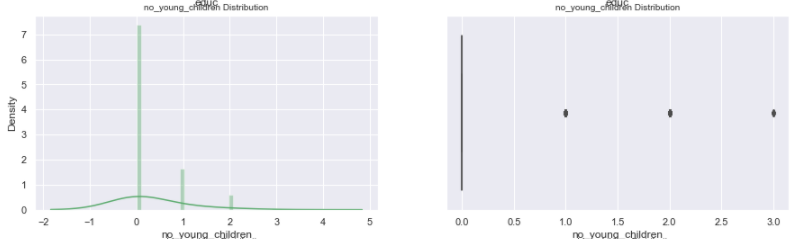
The above dist plot shows the distribution of data from 1000 – 150000 and is positively skewed. Boxplot shows that there are lot of outliers. Majority of the distribution lies in the range of 0 – 100000.



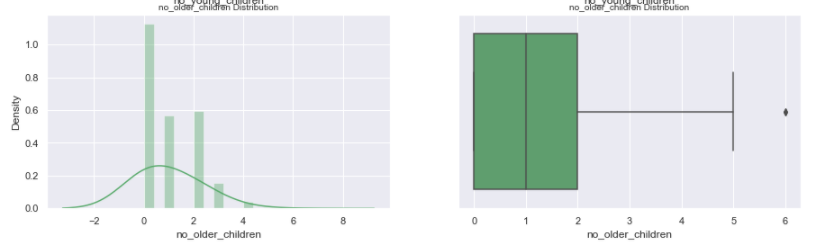
The above dist plot shows the distribution of data from 20 – 60 and is positively skewed. Boxplot shows that there are no outliers.



The above dist plot shows the distribution of data from 0 – 20 and is positively skewed. Boxplot shows that there are a few outliers.

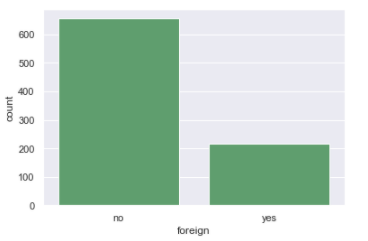


The above dist plot shows the distribution of data from 0 – 2 and is positively skewed. Boxplot shows that there are a few outliers.

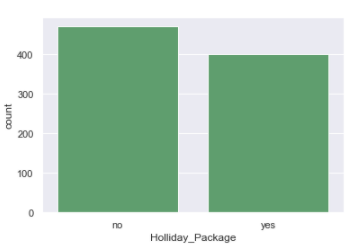


The above dist plot shows the distribution of data from 0 – 4 and is positively skewed. Boxplot shows that there are a few outliers.

Categorical variables



The distribution of the foreign ‘no’ shows maximum frequency.



The distribution of the Holiday Package ‘no’ shows maximum frequency.

**Holiday Package V/S Salary**

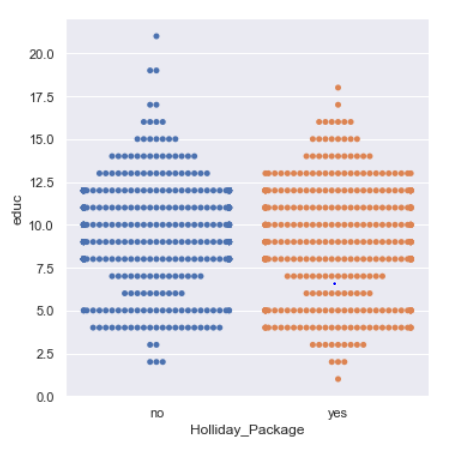


We can see employee below salary 150000 have always opted for holiday package

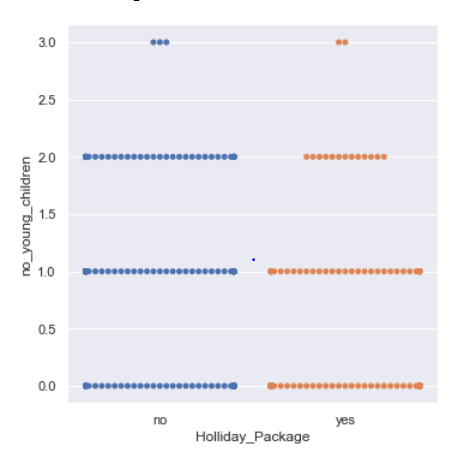
**Holiday Package V/S Age**



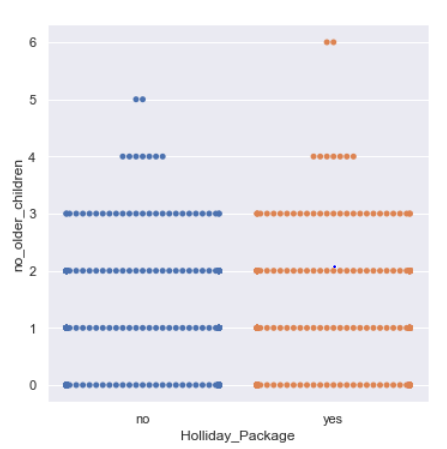
**Holiday Package V/S EDUC**



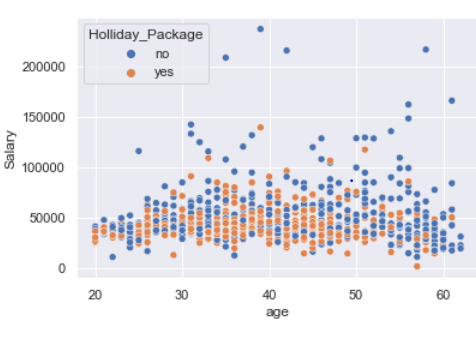
**Holiday Package V/S no. of Young Children**

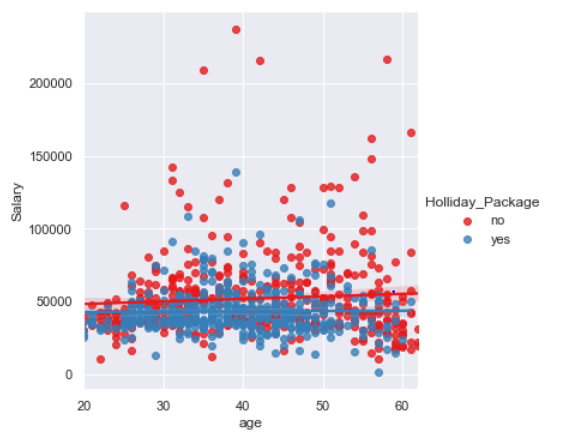


**Holiday Package V/S no. of Older Children**



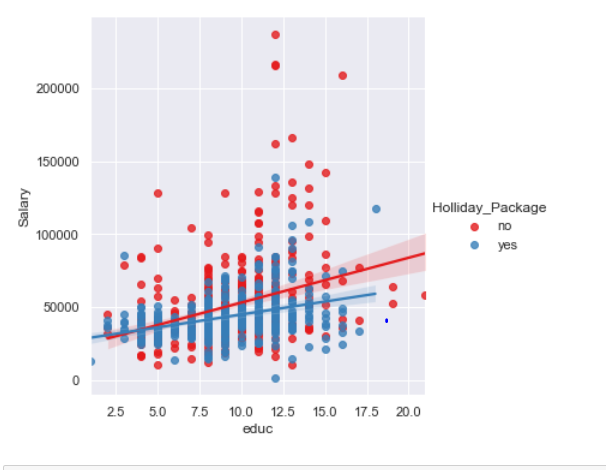
**Holiday Package V/S Age V/S Salary**

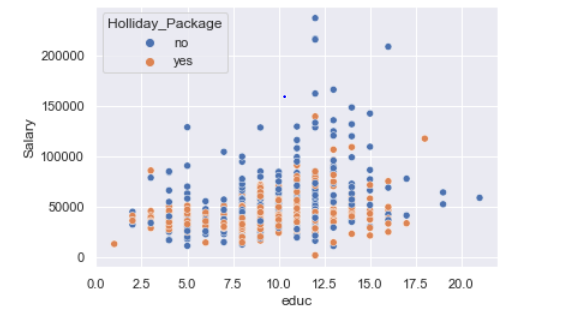




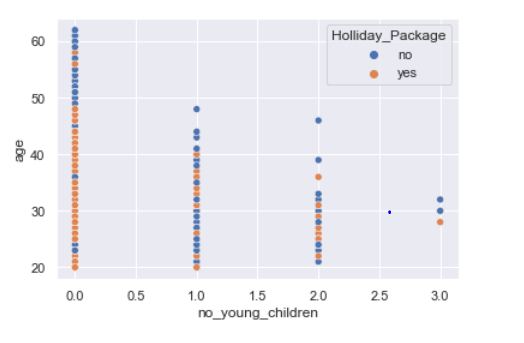
Employee age over 50 to 60 have seems not taking the holiday package. Whereas in the age 30 to 50 group salary less than 50000 people have opted for holiday package.

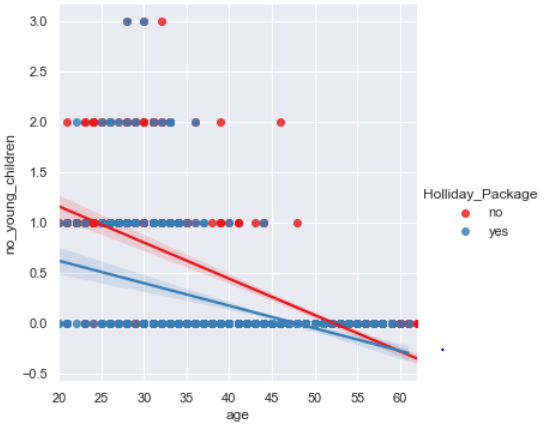
**Holiday Package V/S EDUC V/S Salary**



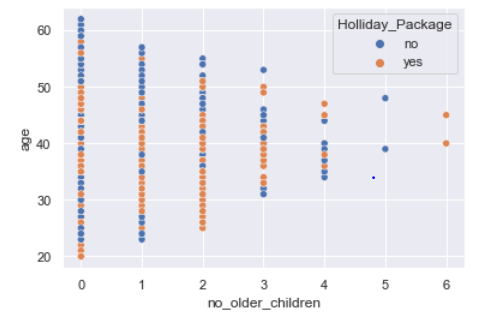


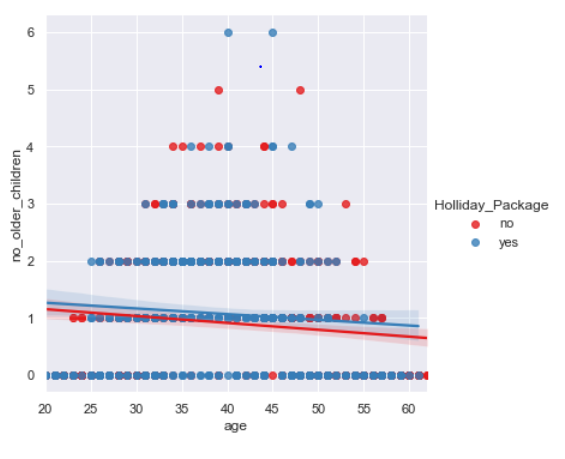
**Holiday Package V/S Age V/S Young Children**





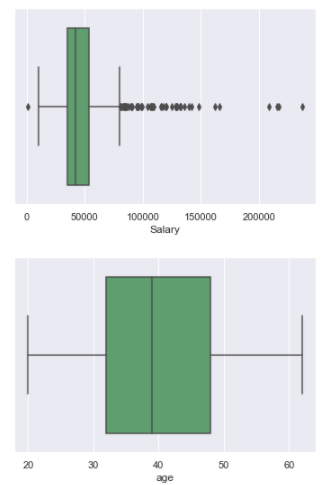
**Holiday Package V/S Age V/S Older Children**

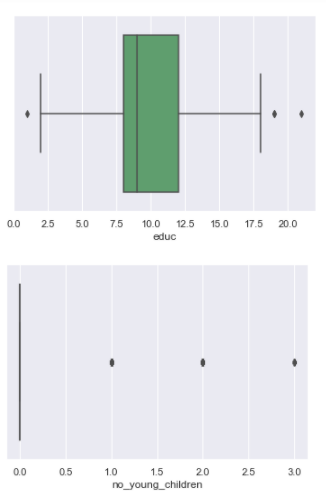


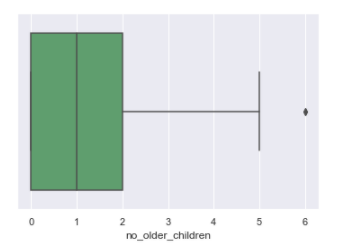


Performing and identifying outliers for Salary ,age , educ , no\_young\_children, no\_older\_children

**Salary and Age**



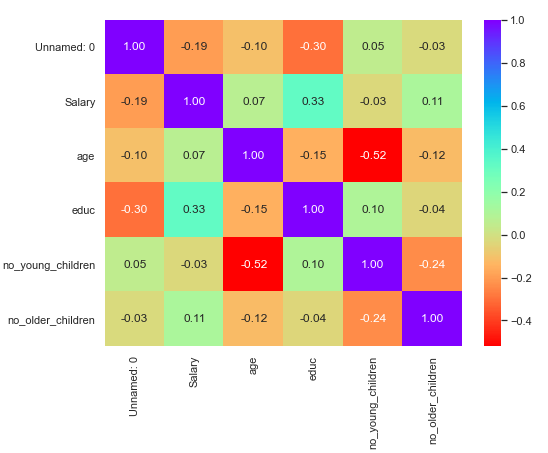




**Bi Variate analysis data distribution**

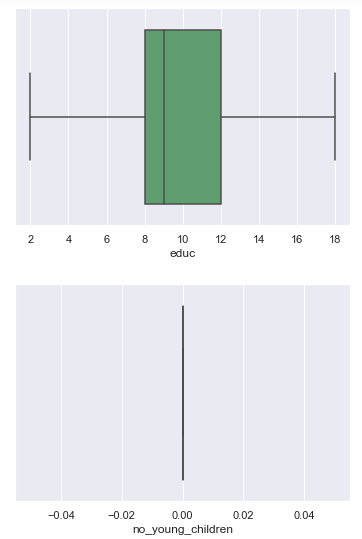


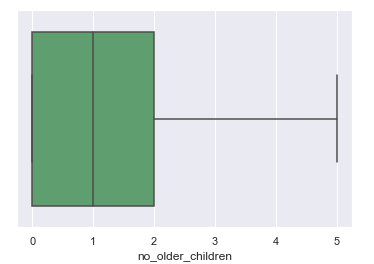
There is no correlation between the data; the data seems to be normal. There is no huge difference in the data distribution among the holiday package.



**After Outlier Treatment**







### Problem 2.2

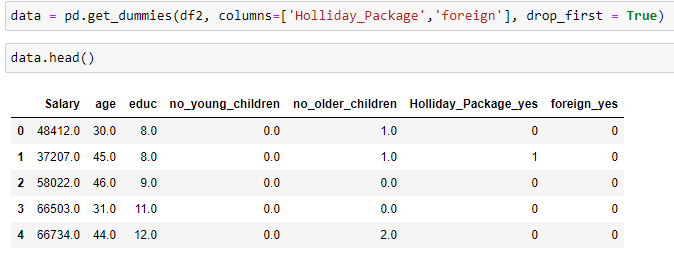
Do not scale the data. Encode the data (having string values) for Modeling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

**Resolution:**

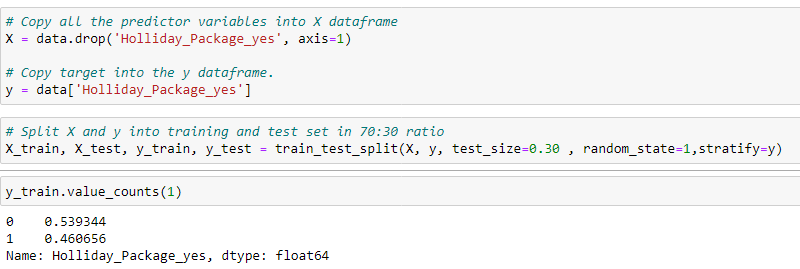
For training and testing purpose we are splitting the dataset into train and test data in the ratio 70:30.

We have divided the dataset into train and test.

Encoding the categorical variables:

\

Train and Test Split



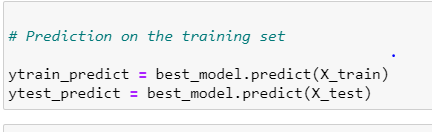
Grid search method:

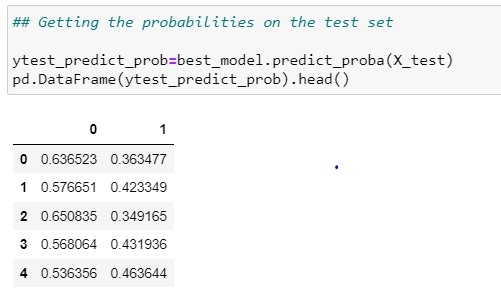
The grid search method is used for logistic regression to find the optimal solving and the parameters for solving:



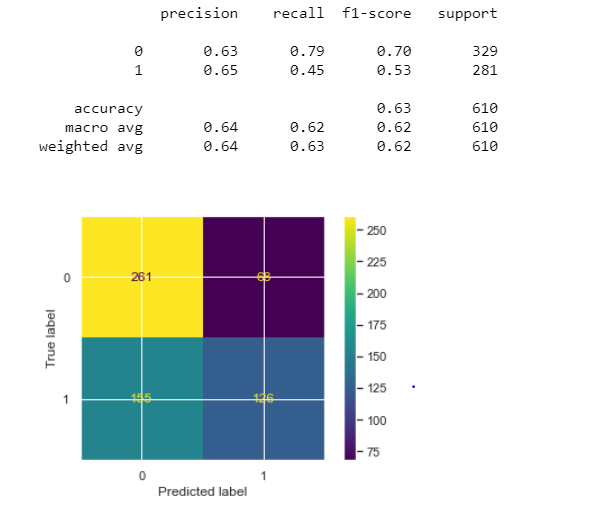
The grid search method gives,linear solver which is suitable for small datasets.tolerance and penalty has been found using grid search method

Predicting the training data:

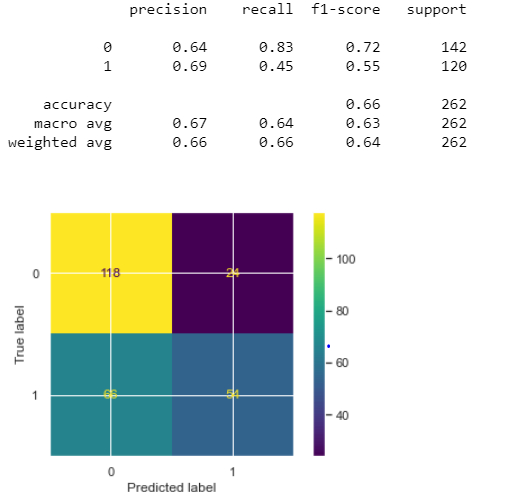




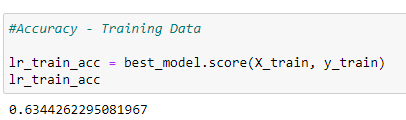
Confusion Matrix Data



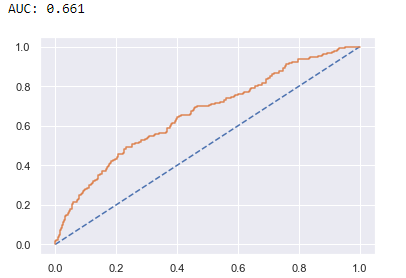
Confusion Matrix Test Data



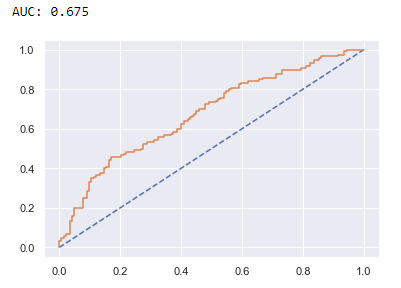
Accuracy



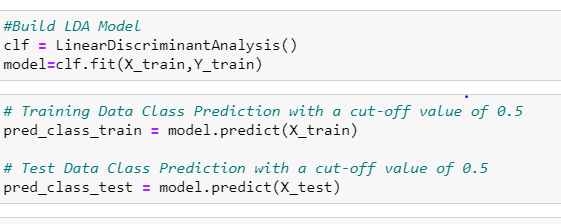
AUC and ROC for the training data



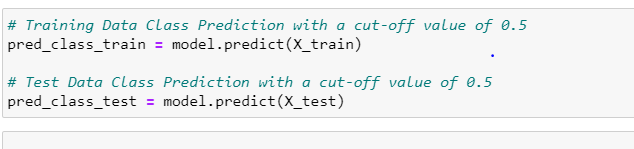
AUC and ROC Curve from the Test data



LDA –



Predicting the variable –

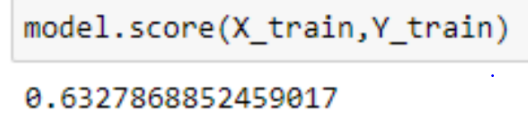


### Problem 2.3

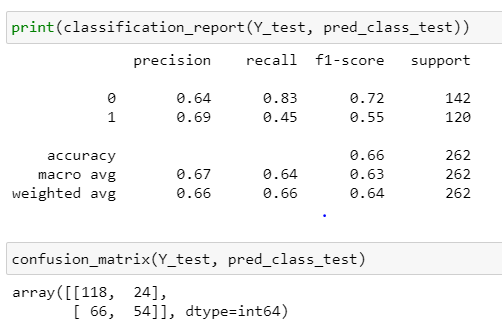
Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

**Resolution:**

Model Score:



Classification report for the test data:



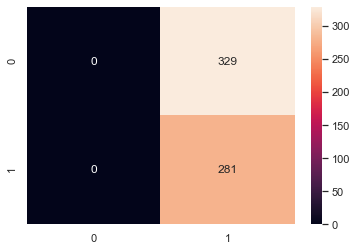
Changing the cut off value to check optimal value that gives better accuracy and f1 score

**0.1**

**Accuracy Score 0.4607**

**F1 Score 0.6308**

**Confusion Matrix**

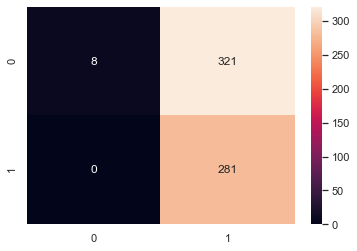
****

**0.2**

**Accuracy Score 0.4738**

**F1 Score 0.6365**

**Confusion Matrix**

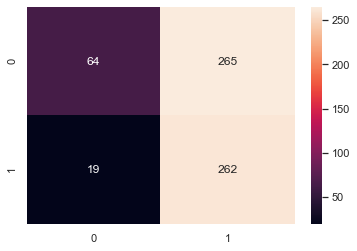
****

**0.3**

**Accuracy Score 0.5344**

**F1 Score 0.6485**

**Confusion Matrix**

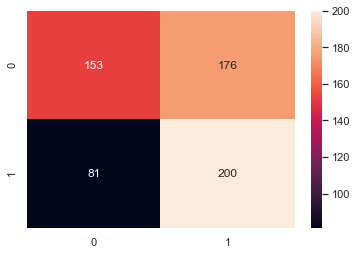
****

**0.4**

**Accuracy Score 0.5787**

**F1 Score 0.6088**

**Confusion Matrix**

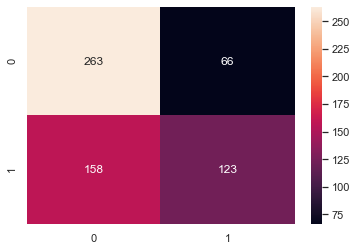
****

**0.5**

**Accuracy Score 0.6328**

**F1 Score 0.5234**

**Confusion Matrix**

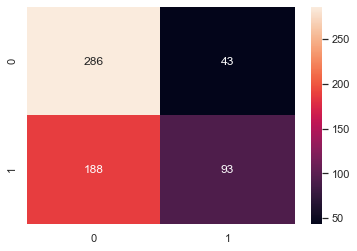
****

**0.6**

**Accuracy Score 0.6213**

**F1 Score 0.446**

**Confusion Matrix**

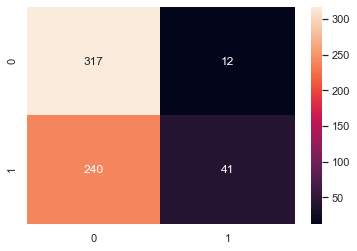
****

**0.7**

**Accuracy Score 0.5869**

**F1 Score 0.2455**

**Confusion Matrix**

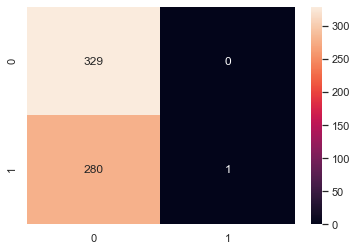
****

**0.8**

**Accuracy Score 0.541**

**F1 Score 0.0071**

**Confusion Matrix**

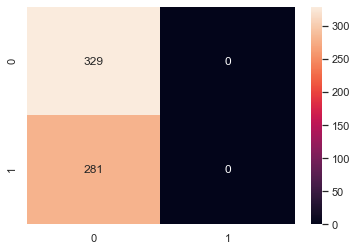
****

**0.9**

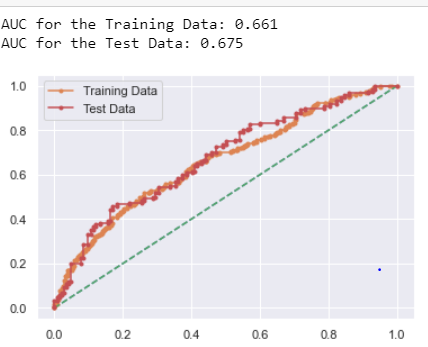
**Accuracy Score 0.5393**

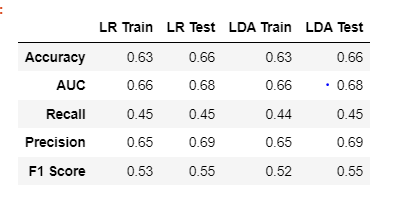
**F1 Score 0.0**

**Confusion Matrix**

****

**AUC and ROC curve for the train data:**





Comparing the above models, the results are pretty much the same, but LDA works better when there is categorical variable.

### Problem 2.4

Basis on these predictions, what are the insights and recommendations.

Please explain and summaries the various steps performed in this project. There should be proper business interpretation and actionable insights present.

**Resolution:**

We had a business report where we need to predict whether an employee would opt for a holiday or not, in order to arrive at this prediction both logistic regression and LDA, since both the results are pretty much the same.

The EDA analysis clearly indicates certain criteria we could find people aged above 50 are not interested much in holiday packages.

Employees ranging from 30 to 50 generally opt for holiday packages, while in the age bracket of 30 to 50 and salary less than 50k, people have opted for more holiday packages.

The important factors for predictions are salary, age and educ.

**Recommendations:**

* To improve holiday packages above 50 , we can provide religious packages.
* For people earning more than 150k , we can provide vacation packages.

The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*