Project

Predictive Modeling

# Megha R

# PGP- DSBA.O.MAR22.C

# Table of Contents

# [Q1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null](#_bookmark0) [values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis………..….…](#_bookmark0)4

[Q1.2. Impute null values if present, also check for the values which are equal to zero. Do they have](#_bookmark1) [any meaning, or do we need to change them or drop them? Do you think scaling is necessary in this](#_bookmark1) [case?...................................................................................................1](#_bookmark1)8

[Q1.3. Encode the data (having string values) for Modelling. Data Split: Split the data into test and](#_bookmark2) [train (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions](#_bookmark2) ………………………………………………………………………………………... 21

[Q1.4 Inference: Basis on these predictions, what are the business insights and recommendations. …………………………………………………………………….. 2](#_bookmark3)7

Q2.1 [Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check,](#_bookmark4)write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.. ………………………………………………………………………..29

Q2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis)……………………………………………………………….…………….. 38

Q2.3 [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 Compare Both the models and write inference which model is best/optimized.](#_bookmark6)....................42

Q2.4 [Inference: Basis on these predictions, what are the insights and recommendations.](#_bookmark7) 47

## Problem Statement:

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.

#### Data Dictionary

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| **Carat** | Carat weight of the cubic zirconia. |
| **Cut** | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| **Color** | Colour of the cubic zirconia.With D being the best and J the worst. |
| **Clarity** | cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3 |
| **Depth** | The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. |
| **Table** | The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter. |
| **Price** | the Price of the cubic zirconia. |
| **X** | Length of the cubic zirconia in mm. |
| **Y** | Width of the cubic zirconia in mm. |
| **Z** | Height of the cubic zirconia in mm. |

# 

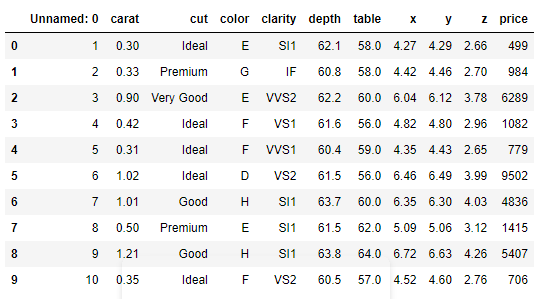
# Q1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

**Exploratory Data Analysis or (EDA)** is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, Box plot, pairplot and many more. It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.

Imported the required libraries.

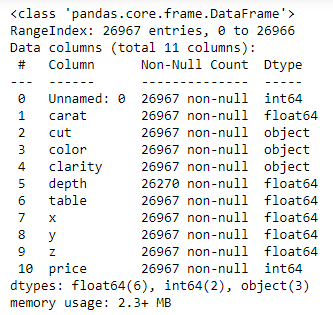
To build the Linear Regression Model on our dataset we need to import the following packages:

* + - Importing the dataset.
    - Data Summary and Exploratory Data Analysis:
      * Checking if the data is being imported properly
      * Head: The top 10 rows of the dataset are viewed.

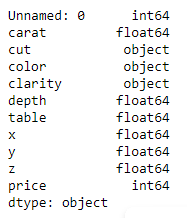


* + - * Dimension of the Dataset: The Dimension or shape of the dataset is as follows. It shows that the dataset given to us has 26967 rows and 11 columns or variables.

Structure of the Dataset : Structure of the dataset is computed.

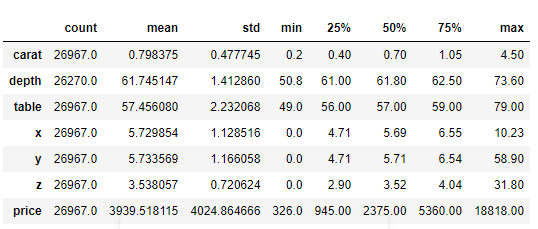


* Data Type is – Integer/Float/Object



* Checking for Duplicates: - There are 34 duplicate rows in the dataset as computed.
* We will drop the first column ‘**Unnamed: 0**’ column as this is not important for our study.

## Descriptive Statistics for the dataset :

****

**Carat:-** This is an independent variable, and it ranges from 0.2 to 4.5. mean value is around 0.8 and 75% of the stones are of 1.05 carat value. Standard deviation is around 0.477 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower carat. There are very few stones above 1.05 carat.

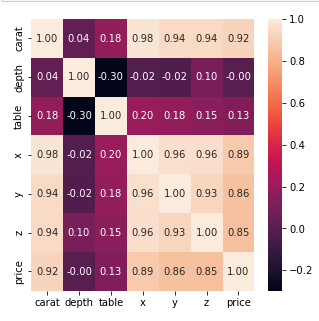
**Depth :-** The percentage height of cubic zirconia stones is in the range of 50.80 to 73.60. Average height of the stones is 61.80 25% of the stones are 61 and 75% of the stones are 62.5. Standard deviation of the height of the stones is 1.4. Standard deviation is indicating a normal distribution

**Table**:- The percentage width of cubic Zirconia is in the range of 49 to 79. Average is around 57. 25% of stones are below 56 and 75% of the stones have a width of less than 59. Standard deviation is 2.24. Thus the data does not show normal distribution and is similar to carat with most of the stones having less width also this shows outliers are present in the variable.

**Price**:- Price is the Predicted variable. Prices are in the range of 3938 to 18818. Median price of stones is 2375, while 25% of the stones are priced below 945. 75% of the stones are in the price range of 5356. Standard deviation of the price is 4022. Indicating prices of majority of the stones are in lower range as the distribution is right skewed.

Variables x, y, and z seems to follow a normal distribution with a few outliers.

## Checking Correlation in the data using Heatmap

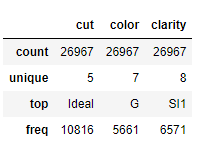
****

Observations:

* + - High correlation between the different features like carat, x, y, z and price.
    - Less correlation between table with the other features.
    - Depth is negatively correlated with most the other features except for carat

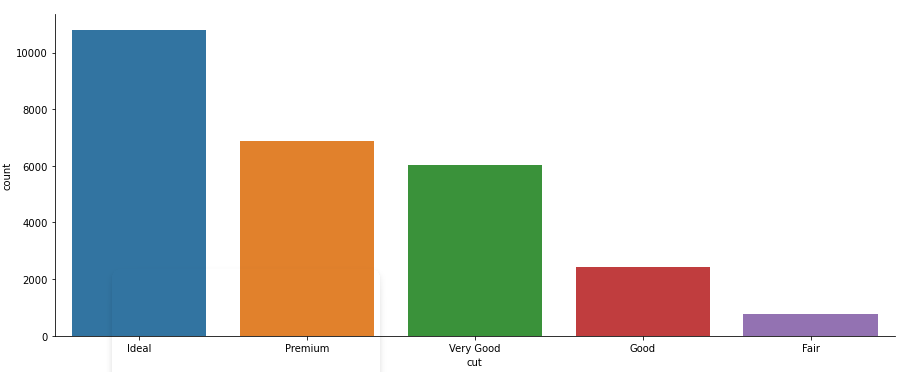
## Univariate & Bivariate Analysis

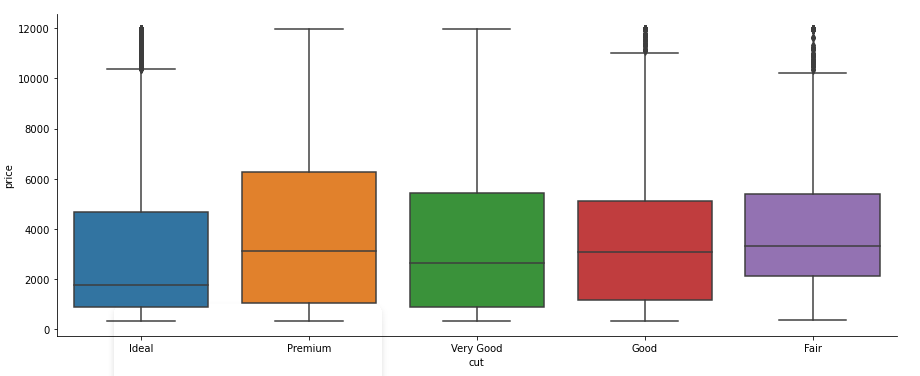
#### Getting unique counts of Categorical Variables



Looking at the above unique values for variable “Cut “ we see the ranking given for each unique value like “ **Fair, Good, Ideal, Premium, Very Good “**

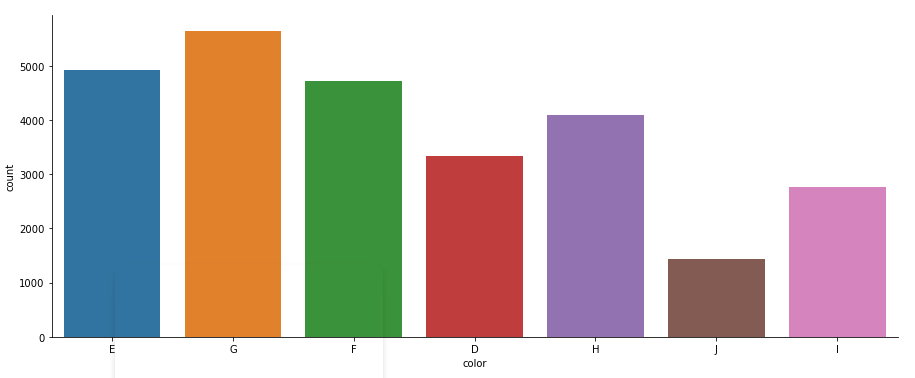
#### Price Distribution of Cut Variable

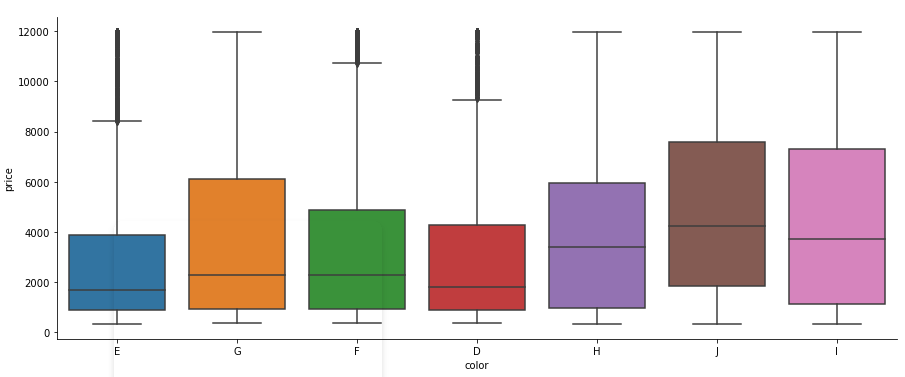
****

****

* + - For the cut variable we see the most sold is Ideal cut type gems and least sold is Fair cut gems
    - All cut type gems have outliers with respect to price
    - Slightly less priced seems to be Ideal type and premium cut type to be slightly more expensive

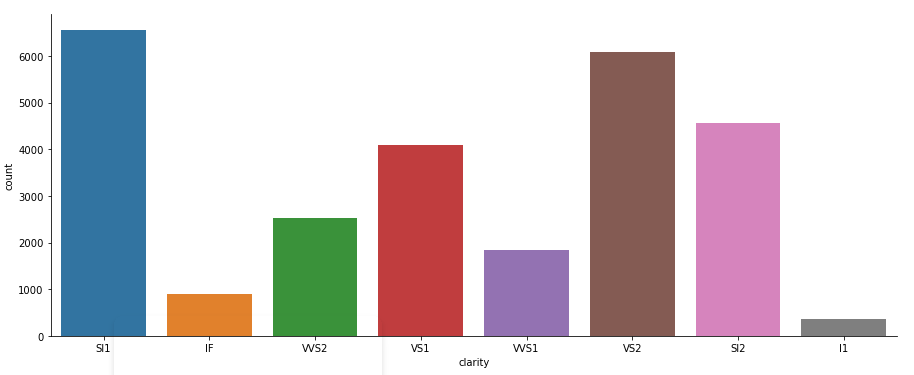
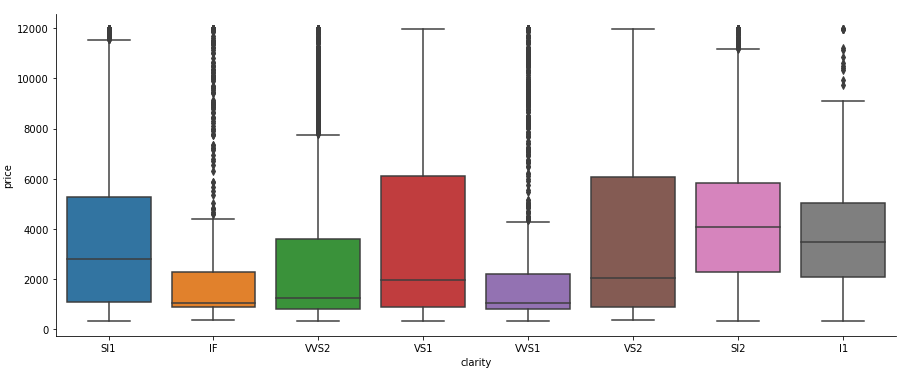
#### Price Distribution of Color Variable





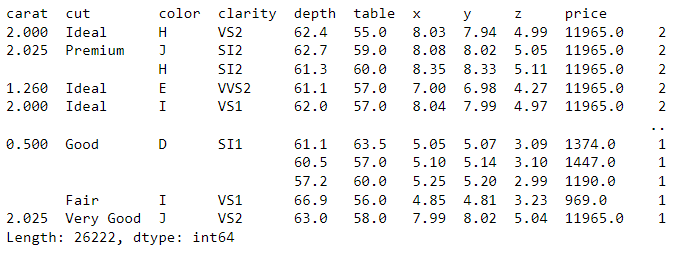
* + - For the color variable we see the most sold is G colored gems and least is J colored gems
    - All color type gems have outliers with respect to price
    - However, the least priced seems to be E type; J and I colored gems seems to be more expensive

#### Price Distribution of Clarity Variable

** **

* + - For the clarity variable we see the most sold is SI1 clarity gems and least is I1 clarity gems
    - All clarity type gems have outliers with respect to price
    - Slightly less priced seems to be SI1 type; VS2 and SI2 clarity stones seems to be more expensive

#### Getting unique counts of Numeric Variables



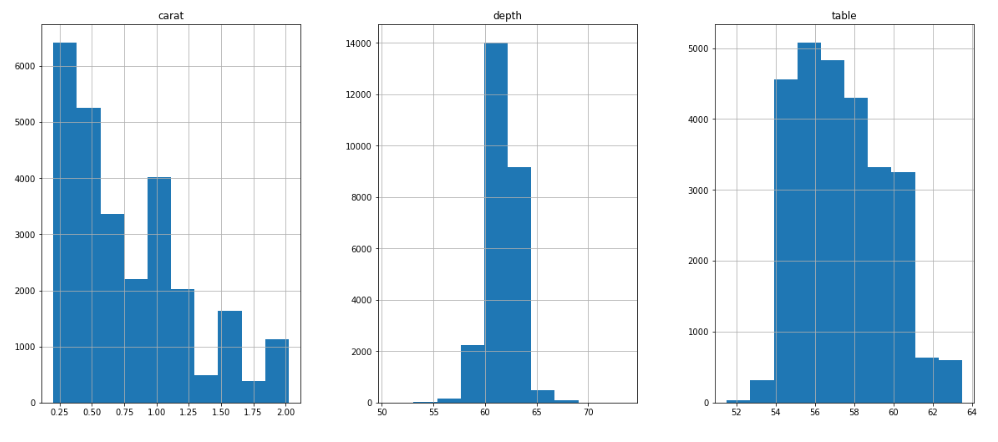
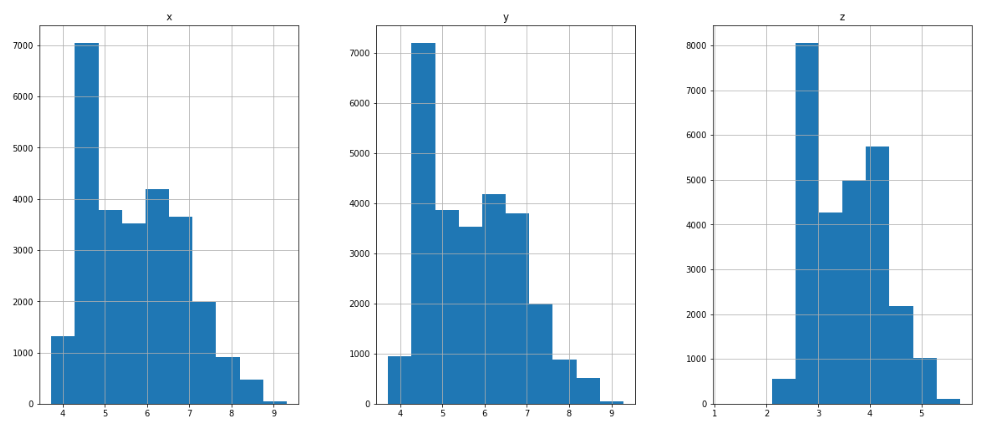
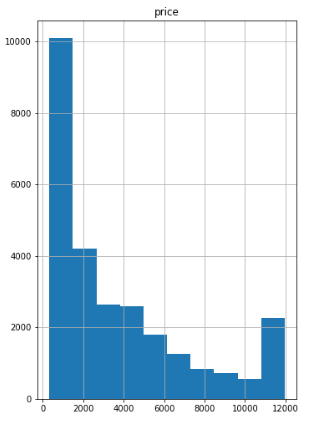
Histograms and Boxplot for each variable to check the data distribution Observations:

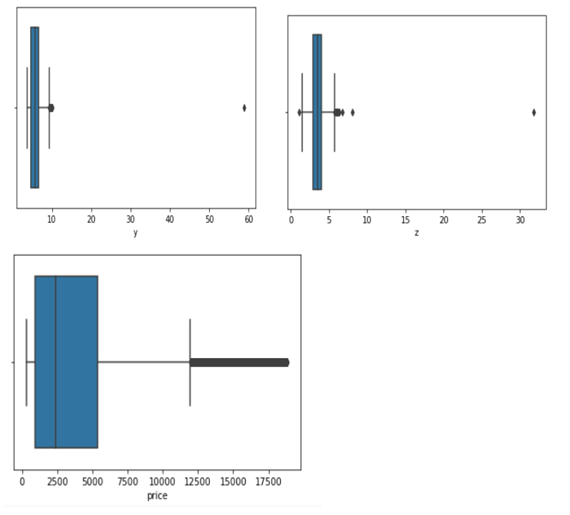
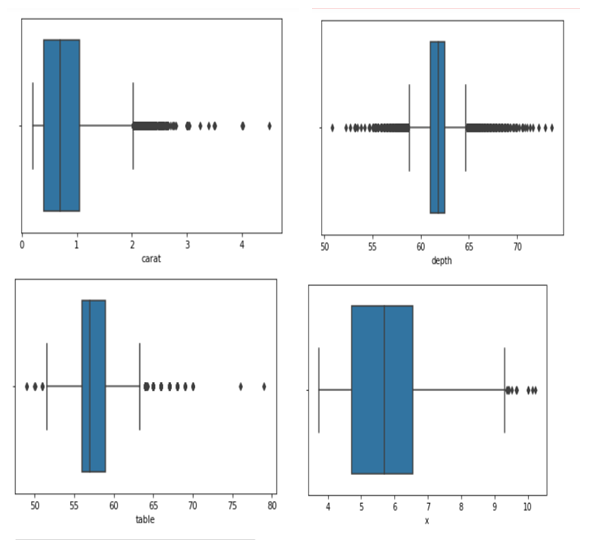
#### Independent Variables

* + - Depth is the only variable which can be considered as normal distribution
    - Carat, Table, x, y, z these variables have multiple modes with the spread of data
    - Outliers: Large number of outliers are present in all the variables (Carat, Depth, Table, x, y, z)

#### Price will be the target variable or dependent variable

* + - It is right skewed with large range of outliers



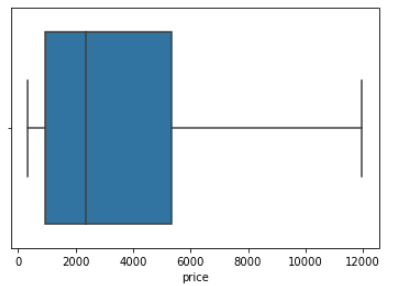
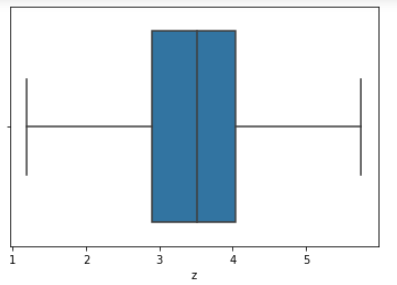
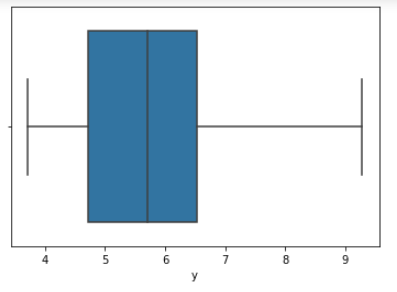
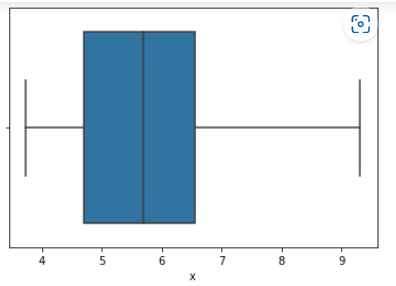
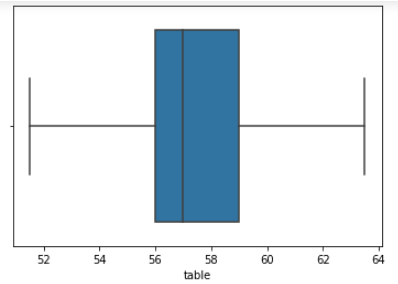
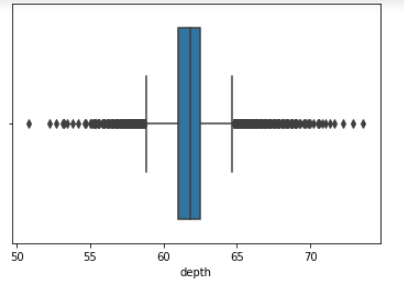
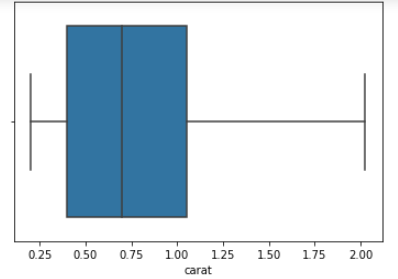


There are outliers present in all the variables as per the above plot

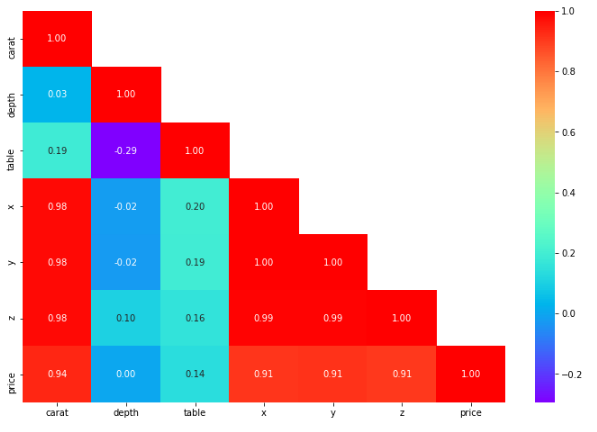
From above data it is seen that except for carat and price variable, all other variables have mean and median values very close to each other, seems like there is no skewness in these variables. Whereas for carat and price we see some difference in value of mean and median, which slightly indicates existence of some skewness in the data

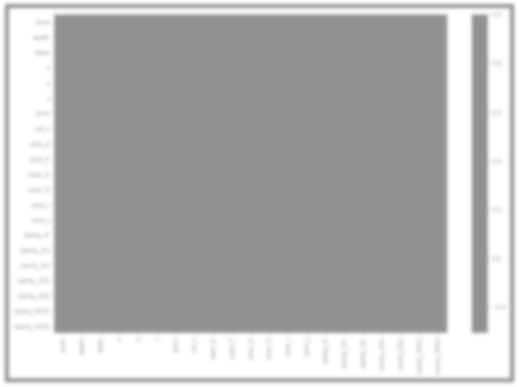
#### Treatment of outliers by IQR method

Box Plots after outliers’ treatment



* + - Checked for data Correlation via heatmap:
    - Heatmap showing correlation between variables

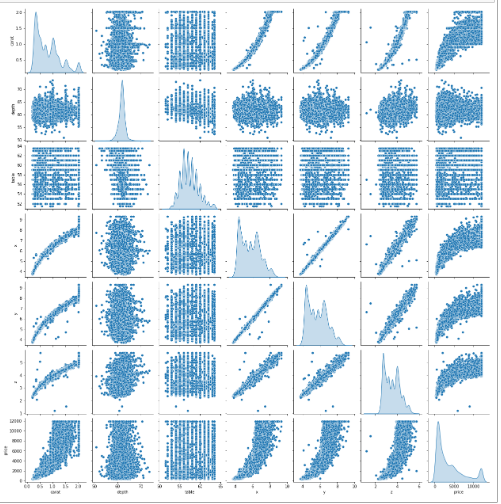




* + - We see strong correlation between Carat, x,y, and z that are demonstrating strong correlation or multicollinearity

Bivariate Analysis :

Pair Plot :



Observations:

* + - Pair plot allows us to see both distribution of single variable and relationships between two variables.

## Conclusion of EDA:

### **Price** – This variable gives the continuous output with the price of the cubic zirconia stones. This will be our **Target Variable**.

* **Carat, depth, table, x, y, z** variables are numerical or continuous variables.
* **Cut, Clarity and colour** are categorical variables.

### We will drop the first column ‘**Unnamed: 0**’ column as this is not important for our study which leaves the shape of the dataset with 26967 rows & 10 Columns

* Only in ‘**depth** 697missing values are present which we will impute by its median values.
* There are total of 34 duplicate rows as computed using. Duplicated () function. We will drop the duplicates
* Upon dropping the duplicates – The shape of the data set is – **26933 rows & 10 columns**

#### Observation-1:

(1).'Price' is the target variable while all others are the predictors. (2).The data set contains 26967 row, 11 column. (3).In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable and all other are predector variable. (4)The first column is an index ("Unnamed: 0")as this only serial no, we can remove it.

* **Observation-2:**

(1).On the given data set the the mean and median values does not have much differenc. (2).We can observe Min value of "x", "y", "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So we have filter out those as it clearly faulty data entries. (3).There are three object data type 'cut', 'color' and 'clarity'.

* **Observation-3:**

We can observe there are 697 missing value in the depth column. There are some duplicate row present. (33 duplicate rows out of 26958). which is nearly 0.12 % of the total data. So on this case we have dropped the duplicated row.

#### Observation-4:

There are significant amount of outlier present in some variable,the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So we have treat the outliar. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

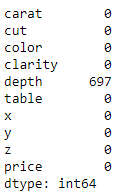
#### Observation-5:

It looks like most features do correlate with the price of Diamond. The notable exception is "depth" which has a negligble correlation (~1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

# Q1.2. 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?

It is important to check if the data has any missing value or gibberish data in it. We did check about the same for both object and numerical data types and can confirm the following:

* There is no gibberish or missing data in the object type data columns – Cut, color and clarity.



* There are missing values in the column “depth” – 697 cells or 2.6% of the total data set. We can choose to impute these values using a mean or median. We checked for both the values and the result for both is almost similar.
* For this case study, I have used median to impute the missing values. Table 1.2.1 – Checking the missing values

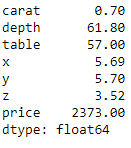


Table 1.2.2. – checking for missing values after imputing the values

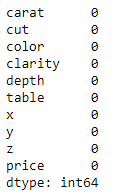
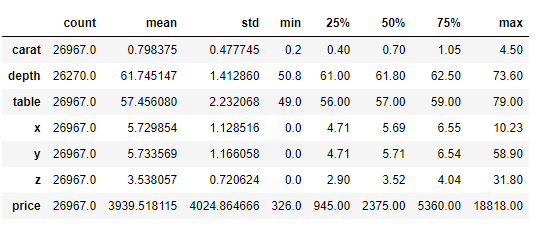
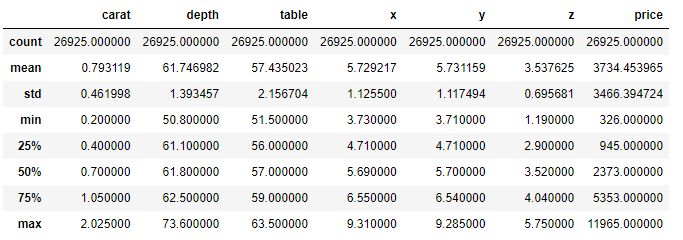


Table: 1.2.3: Showing presence of 0 values in x, y, and z columns



* While there are no missing values in the numerical columns, there are a few 0 in columns – x (3 in count), y(3 in count) and z (9 in count) – in the database. A single row was dealt with during checking for duplicates and the other eight rows were taken care here. Since the total number of rows that had 0 value in them was 8 only, it accounts for a negligible number and for this case study we could have avoided them or dropped. Also, when I checked the correlation values, it seems there is a strong multicollinearity between all three columns. There is a most likely case that I won’t even use them in creating my Linear Regression model. I have chosen to drop those rows as it represented an insignificant number when compared to the overall dataset and it won’t add much value to the analysis here.

Table: 1.2.4: Confirming there are no 0 values in x, y, and z columns



We now have a dataset with 26,925 rows as supposed to 26,933 after treating duplicate entries.

**Do you think scaling is necessary in this case?**

Scaling or Standardizing the features around the centre and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.

For example, A variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization will give the variable with the larger range weight of 1000 in the analysis. Transforming the data to comparable scales can prevent this problem.

In this data set we can see the all the variable are in different scale i.e price are in 1000s unit and depth and table are in 100s unit, and carat is in 10s. So its necessary to scale or standardise the data to allow each variable to be compared on a common scale. With data measured in different "units" or on different scales (as here with different means and variances) this is an important data processing step if the results are to be meaningful or not dominated by the variables that have large variances.

**But is scaling necessary in this case?**

No, it is not necessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not. But recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in preprocessing.

**For now we will process the model without scaling and later we will check the output with scaled data of regression model output.**

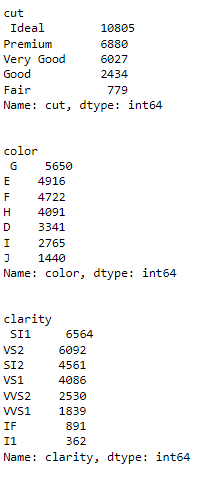
# 

# Q1.3. Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply Linear regression.

Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE.

Answer:

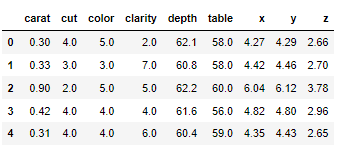
Let us read the data in brief before deciding on what kind of encoding technique needs to be used.



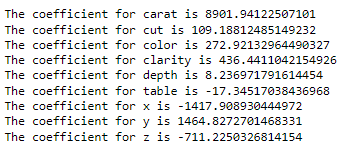
### Now let’s convert the objects to categorical codes.

### 

**Train-Test Split:**



**Coefficient**

****

Y=mx +c (m= m1,m2,m3...m9) here 9 diferent co-efficients will learn aling with the intercept which is "c" from the model.

From the above coefficients for each of the independent attributes we can conclude

The one unit increase in carat increases price by 8901.941.

The one unit increase in cut increases price by 109.188.

The one unit increase in color increases price by 272.921.

The one unit increase in clarity increases price by 436.441.

The one unit increase in y increases price by 1464.827.

The one unit increase in depth increases price by 8.236,

But The one unit increase in table decreases price by -17.345,

The one unit increase in x decreases price by -1417.908,

The one unit increase in z decreases price by -711.225.



The intercept (often labelled the constant) is the expected mean value of Y when all X=0. If X never equals 0, then the intercept has no intrinsic meaning.

The intercept for our model is -3171.950447307667. In preset case when the other predictor variable are zero i.e like carat,cut, color, clarity all are zero then the

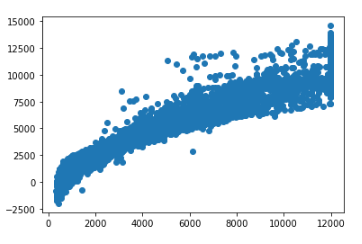
C=-3172. ( Y = m1X1 + m2X2+ ….. + mnXn + C + e) that means price is -3172. which is meaningless. We can do Z score or scaling the data and make it nearly zero.

R-square is the percentage of the response variable variation that is explained by a linear model. Or:

R-square = Explained variation / Total variation

R-squared is always between 0 and 100%: 0% indicates that the model explains none of the variability of the response data around its mean.100% indicates that the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively 0.9311935886926559 and 0.931543712584074.

Scatterplot on test data between dependent variable – price - and Independent variables



RMSE on Training data:



RMSE on Testing data:



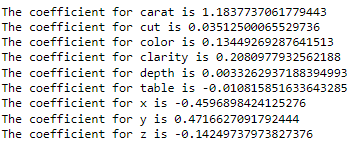
We can see that the is a linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicated some kind noise present on the data set i.e Unexplained variances on the output.

**Linear regression Performance Metrics:**

* intercept for the model: -3171.950447307667
* R square on training data: 0.9311935886926559
* R square on testing data: 0.931543712584074
* RMSE on Training data: 907.1312415459143
* RMSE on Testing data: 911.8447345328436

As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

**Applying zscore statsmodels**



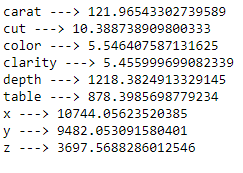


Model score - R2 or coeff of determinant:



Now we can observe by applying z score the intercept became -5.87961525130473e-16. Earlier it was -3171.950447307667. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

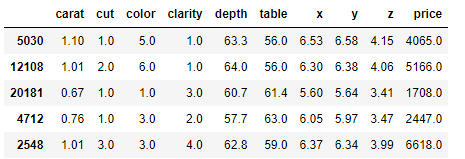
**Check Multi-collinearity using VIF**

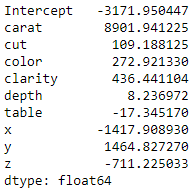


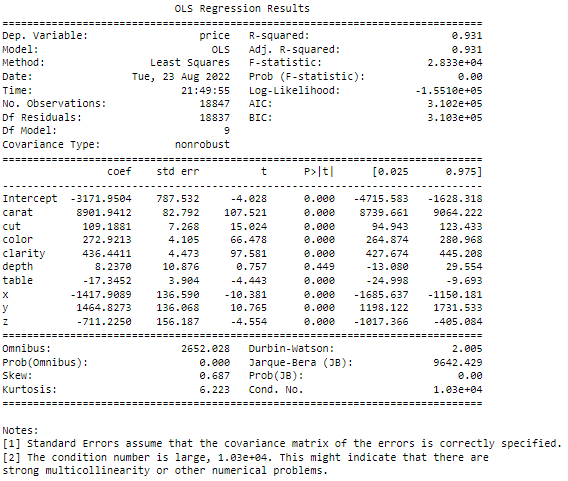
We can observe there are very strong multi collinearity present in the data set. Ideally it should be within 1 to 5.

We are exploring the Linear Regression using statsmodels as we are interested in some more statistical metrics of the model.

**Linear Regression using statsmodels.**







Assuming null hypothesis is true, i.e there is no relationship between this variable with price. from that universe we have drawn the sample and on this sample we have found this co-efficient for the variable shown above.

Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not '0'. Here all regression co-efficients are not '0'.

For an example: we can see the p value is showing 0.449 for 'depth' variable, which is much higher than 0.05. That means this dimension is useless. So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.



The final Linear Regression equation is

price = b0 + b1 \*carat[T.True] + b2 \* cut + b3 \* color + b4 \* clarity+ b5 \* depth + b6 \* table + b7 \* x + b8 \* y + b9 \*z True

price = (-3171.95) \* Intercept + (8901.94) \* carat + (109.19) \* cut + (272.92) \* color + (436.44) \* clarity + (8.24) \* depth + (-17.35) \* table + (-1417.91)) \* x + (1464.83) \* y + (-711.23) \* z \_True

1. When carat increases by 1 unit, diamond price increases by 8901.94 units, keeping all other predictors constant.
2. When cut increases by 1 unit, diamond price increases by 109.19 units, keeping all other predictors constant.
3. When color increases by 1 unit, diamond price increases by 272.92 units, keeping all other predictors constant.
4. When clarity increases by 1 unit, diamond price increases by 436.44 units, keeping all other predictors constant.
5. When y increases by 1 unit, diamond price increases by 1464.83 units, keeping all other predictors constant.

As per model these five attributes that are most important attributes 'Carat', 'Cut', 'color','clarity' and width i.e 'y' for predicting the price.

There are also some negative co-efficient values, for instance, corresponding co-efficient (-1417.91) for 'x',(-711.23) for z and (-17.35) for table This implies, these are inversely proportional with diamond price.

* On the given data set we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stones.
* Similarly for the 'z' variable having negative co-efficient i.e -711.23. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stones.
* Also we can see the 'y' width in mm having positive co-efficient. And the p value is less than 0.05, so we can conclude that higher the width of the stone is a higher profitable stones.
* Finally we can conclude that best 5 attributes that are most important are 'Carat', 'Cut', 'color','clarity' and width i.e 'y' for predicting the price.

# 

# Q1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

**Inference:**

we can see that the from the linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

**Linear regression Performance Metrics:**

* intercept for the model: -3171.950447307667
* R square on training data: 0.9311935886926559
* R square on testing data: 0.931543712584074
* RMSE on Training data: 907.1312415459143
* RMSE on Testing data: 911.8447345328436

**As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.**

**Impact of scaling:**

Now we can observe by applying z score the intercept became -5.87961525130473e-16. Earlier it was -3171.950447307667. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

**Multi collinearity:**

We can observe there are very strong multi collinearity present in the data set.

**From statsmodels:**

We can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

Finally we can conclude that **Best 5 attributes** that are most important are **'Carat', 'Cut', 'color','clarity' and width i.e 'y' for predicting the price.**

1. When 'carat' increases by 1 unit, diamond price increases by 8901.94 units, keeping all other predictors constant.
2. When 'cut' increases by 1 unit, diamond price increases by 109.19 units, keeping all other predictors constant.
3. When 'color' increases by 1 unit, diamond price increases by 272.92 units, keeping all other predictors constant.
4. When 'clarity' increases by 1 unit, diamond price increases by 436.44 units, keeping all other predictors constant.
5. When 'y' increases by 1 unit, diamond price increases by 1464.83 units, keeping all other predictors constant.
6. we can see the p value is showing 0.449 for depth variable, which is much greater than 0.05. That means this attribute is useless.
7. There are also some negative co-efficient values, we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient -1417.9089. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stones.
8. Similarly for the 'z' variable having negative co-efficient i.e -711.23. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stones.

**Recommendations:**

**The Gem Stones company should consider the features 'Carat', 'Cut', 'color','clarity' and width i.e 'y' as most important for predicting the price.**

To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

As we can see from the model Higher the width('y') of the stone is higher the price.

**So the stones having higher width ('y') should consider in higher profitable stones.**

The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.

The Diamonds clarity with 'VS1' &'VS2' are the most Expensive.So these two category also consider in higher profitable stones.

As we see for 'X' i.e Length of the stone, higher the length of the stone is lower the price.

**So higher the Length('x') of the stone are lower is the profitability.**

higher the 'z' i.e Height of the stone is, lower the price.This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why **Stones with higher 'z' is also are lower in profitability.**

The marketing efforts can make use of educating customers about the importance of a better carat score and importance of clarity index. Post this, the company can make segments, and target the customer based on their income/paying capacity etc, which can be further studied.

**Problem 2:** Logistic Regression and LDA

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.

**Data Dictionary:**

|  |  |
| --- | --- |
| Variable Name | Description |
| **Holiday\_Package** | Opted for Holiday Package yes/no? |
| **Salary** | Employee salary |
| **age** | Age in years |
| **edu** | Years of formal education |
| **no\_young\_children** | The number of young children (younger than 7 years) |
| **no\_older\_children** | Number of older children |
| **foreign** | foreigner Yes/No |

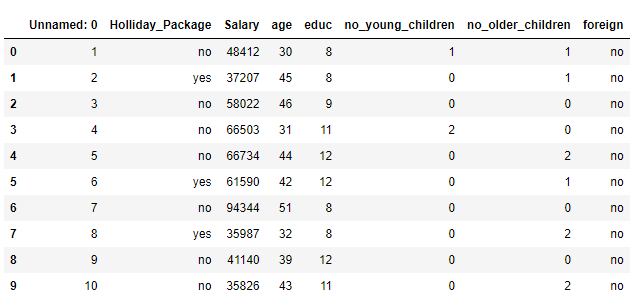
# 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

**Exploratory Data Analysis or (EDA)** is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, Box plot, pairplot and many more. It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.

Imported the required libraries.

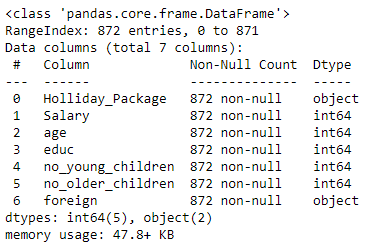
To build the Logistic Regression Model and LDA on our dataset we need to import the packages

* + - Importing the dataset.
    - Data Summary and Exploratory Data Analysis:
      * Checking if the data is being imported properly
      * Head: The top 10 rows of the dataset are viewed.



* + - * Dimension of the Dataset: The Dimension or shape of the dataset is found out. It shows that the dataset given to us has 872 rows and 8 columns or variables

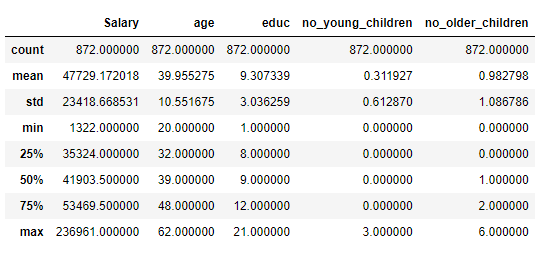
Structure of the Dataset : Structure of the dataset is computed.



* Data Type is – Integer/Object
* Checking for Duplicates: - There are No duplicate rows in the dataset.
* We have dropped the first column ‘**Unnamed: 0**’ column as this is not important for our study. The shape would be – 872 rows and 7 columns

**Univariate Analysis**

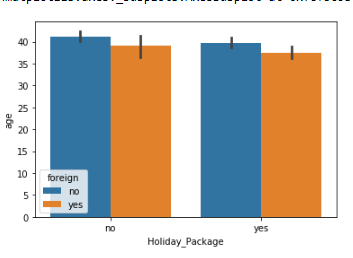
#### Descriptive Statistics for the dataset:



Summary of the Dataset

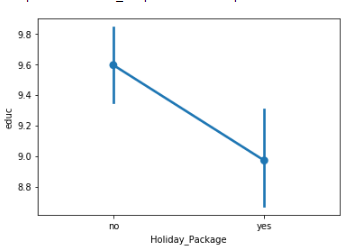
* Holiday Package – This variable is a categorical Variable. output with the this will be our **Target Variable**.
* Salary, age, educ, no\_young\_children, no\_older\_children, variables are numerical or continuous variables.
* Salary ranges from 1322 to 236961. Average salary of employees is around 47729 with a standard deviation of 23418. Standard deviation indicates that the data is not normally distributed. skew of 0.71 indicates that the data is right skewed and there are few employees earning more than an average of 47729. 75% of the employees are earning below 53469 while 255 of the employees are earning 35324.
* Age of the employee ranges from 20 to 62. Median is around 39. 25% of the employees are below 32 and 25% of the employees are above 48. Standard deviation is around 10. Standard deviation indicates almost normal distribution.
* Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the data
* Foreign is a categorical variable
* We have dropped the first column ‘**Unnamed: 0**’ column as this is not important for our study. Unnamed is a variable which has serial numbers so may not be required and thus it can be dropped for further analysisThe shape would be – 872 rows and 7 columns
* There are no null values
* There are no duplicates
  + - Checked for data distribution by bar plot and point plot

**Holiday package and age:**

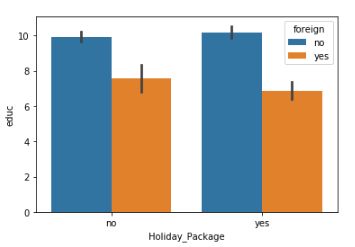


From the above graph, the Indian individuals have opted for holiday package more as compared to foreign individuals. Although the difference is very small.

**Holiday package and educ:**

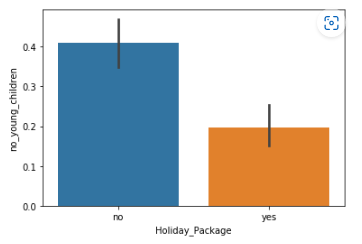


From the above plot we can say that, the guys whose education is little less have opted for package more as compared to those who have higher education.



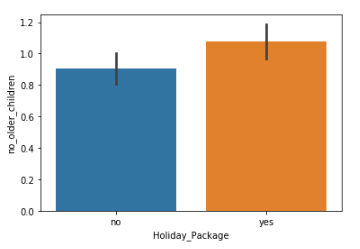
Out of all the customers, whose who have chosen the holiday package, most of them are non-foreigners.

**Holiday package and no\_young\_children:**



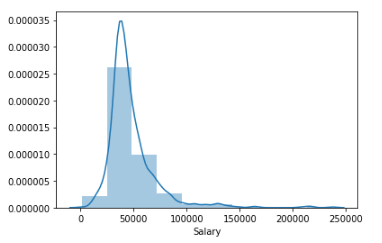
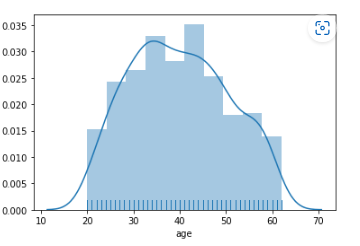
People which have lower number of young children has opted for package more compared to those having greater number of children.

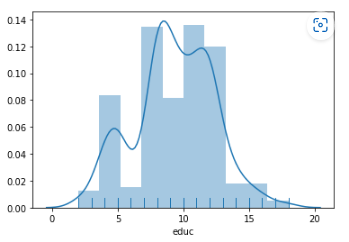
**Holiday package and no\_older\_children:**



People which have higher number of older children has opted for package more compared to those having smaller number of young children.

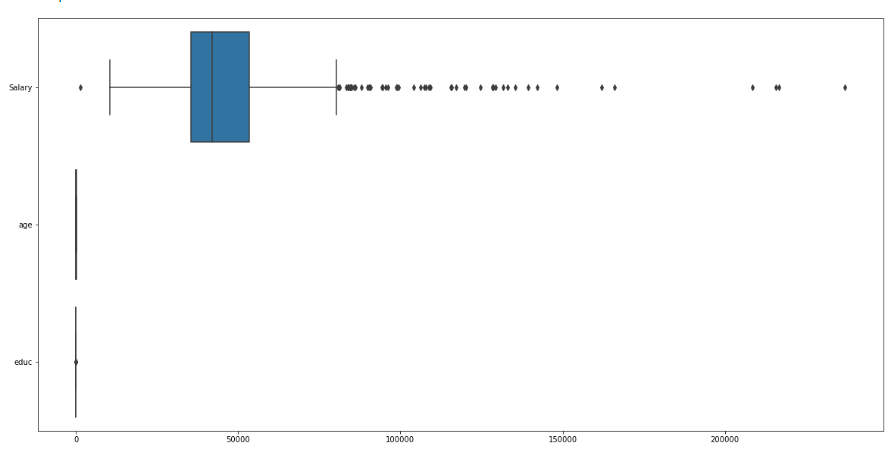
* + - Checked for data distribution by plotting histogram:



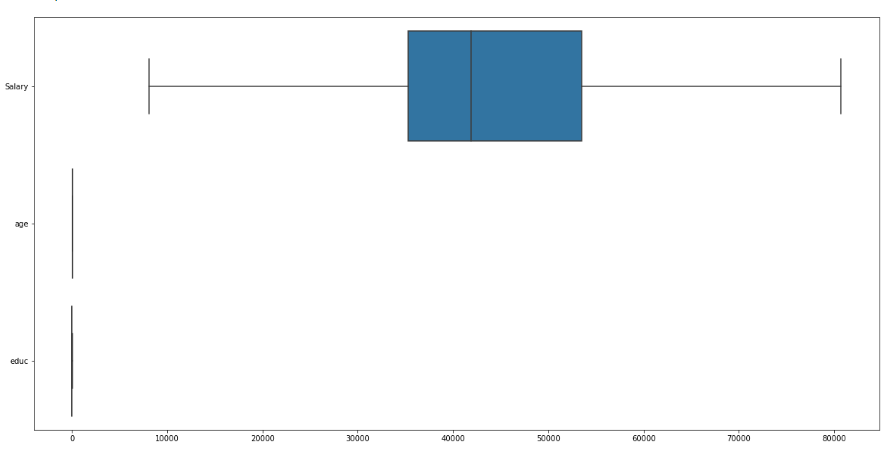
* + - For the Histogram of salary the data look little skewed.
    - For age and educ the data looks to be normally distributed.

Let us check for outliers by plotting the box plot



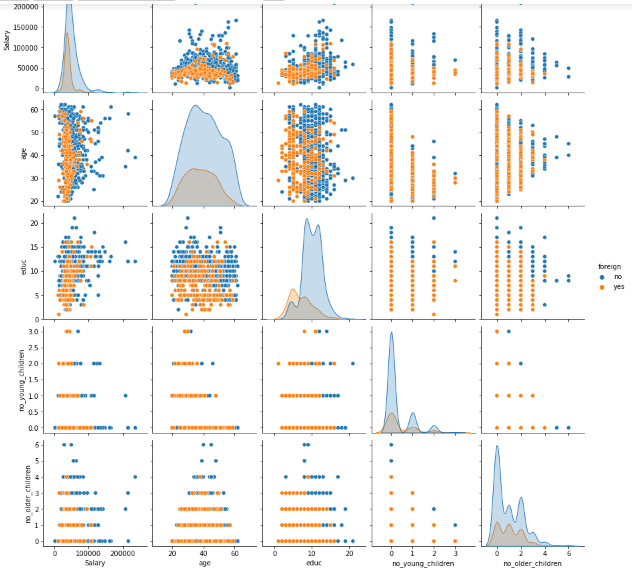
We can observer that there are significant outliers present in variable “ Salary”, however there are minimal outliers in other variables like ‘educ’, ‘no. of young children’ & ‘no. of older children’. There are no outliers in variable ‘age’. For Interpretation purpose we would need to study the variables such as no. of young children and no. of older children before outlier treatment. For this case study we have done outlier treatment for only salary & educ.

**Treatment of outliers by IQR method** (placed it here for comparison sake only) Box Plots after outliers’ treatment –

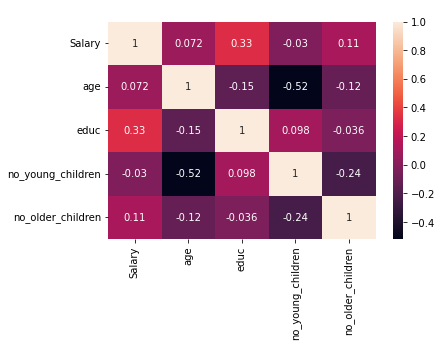


Bi-Variate Analysis:

Checking pairwise distribution of the continuous variables: [Salary, age, educ, no. of young children, 'no of older children]



* + - Checked for data Correlation
    - We will see correlation between independent variables to see which factors might influence choice of holiday package.
    - Heatmap showing correlation between variables



We can relate there isn’t any strong correlation between any variables. Salary and education display moderate corelation and no\_older\_children is somewhat correlated with salary variable. However, there are no strong correlation in the data set.

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

Answer:

While Linear Regression helps us predicting continuous target variable, Logistic Regression helps us for predicting a discrete a target variable. Logistic Regression is one of the “white- box” algorithms which helps us in determining the probability values and the corresponding cut-offs. Logistic regression is used to solve such problem which gives us the corresponding probability outputs and then we can decide the appropriate cut-off points to get the target class outputs.

Precisely Logistic Regression is defined as a statistical approach, for calculating the probability outputs for the target labels. In its basic form, it is used to classify binary data. Logistic regression is very much similar to linear regression where the explanatory variables(X) are combined with weights to predict a target variable of binary class(y).

Evaluation of Logistic regression model- Performance measurement of classification algorithms are judge by confusion matrix which comprise the classification count values of actual and predicted labels.

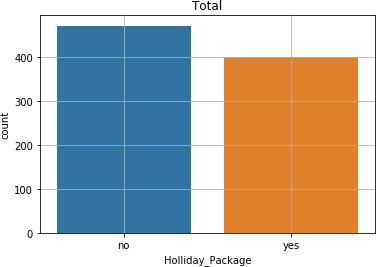
Pros and cons of Logistics Regression:

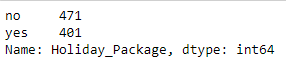
Pros- Logistic regression classification model is simple and easily scalable for multiple classes.

Cons- Classifier constructs linear boundaries and the interpretation of coefficients value is difficult.

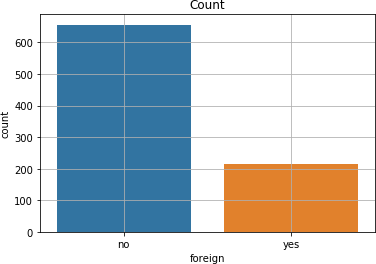
In the given dataset, the target variable – Holiday Package and an independent variable – Foreign are object variables. Let us study them one at a time.

Holiday\_Package: The distribution seems to be fine, with 54% for no and 46% for yes.





Foreign: The data is imbalanced with more skewed towards no and relatively a smaller shared for yes.





Both the variables can be encoded into numerical values for model creation analytical purposes.

Table 2.2.1: Encoding the string values

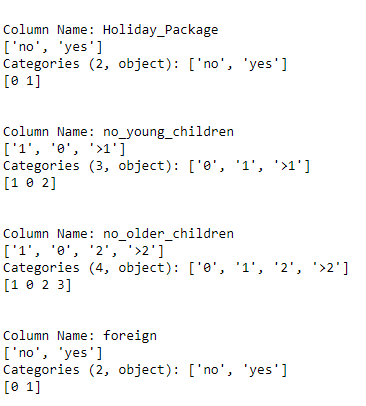
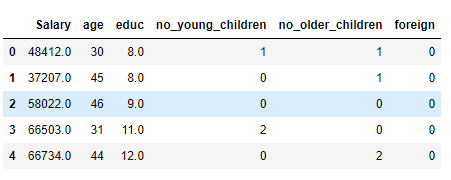
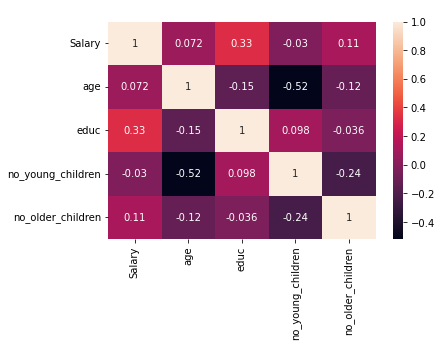


Table 2.2.2: Checking if the values are converted into numeric.



Before we proceed with the model creation, let us read the other part of the data to see how the numerical data also impacts the model. I will first look at the data correlation to quickly identify the variable importance using the heatmap.



We can relate there isn’t any strong correlation between any variables. Salary and education display moderate correlation and no\_older\_children is somewhat correlated with salary variable. However, there are no strong correlation in the data set.

# 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.

Confusion matrix cells are populated by the terms:

True Positive(TP)- The values which are predicted as True and are actually True. True Negative(TN)- The values which are predicted as False and are actually False. False Positive(FP)- The values which are predicted as True but are actually False. False Negative(FN)- The values which are predicted as False but are actually True.

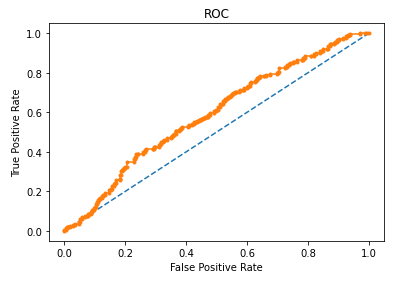
ROC Curve- Receiver Operating Characteristic(ROC) measures the performance of models by evaluating the trade-offs between sensitivity (true positive rate) and false (1- specificity) or false positive rate.

AUC - The area under curve (AUC) is another measure for classification models is based on ROC. It is the measure of accuracy judged by the area under the curve for ROC.

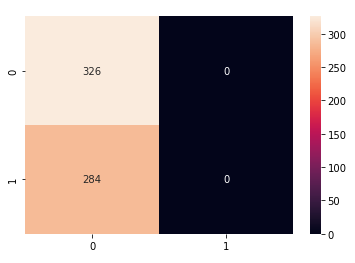
#### Performance Matrix of Logistics Regression model:

* + - Train data

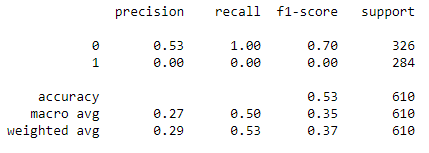
AUC score is 0.590 or 59.0%



Confusion Matrix for Train data:

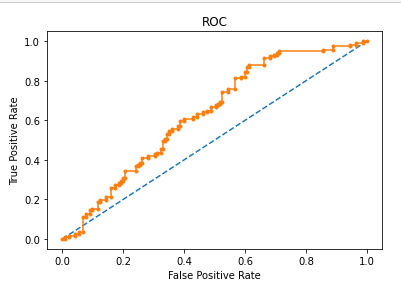


Classification report for Train data:

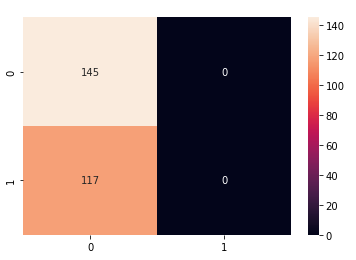


* Test Data:

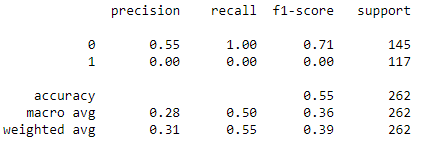
AUC score is 0.6329 or 63.29%



Confusion Matrix for Test Data



Classification report for Test Data



While looking the metrices for both training and the test data, it seems the accuracy scores are same on both models at 66%. Our model is close enough to be treated as a right fit model. The current model is not struggling with being a over fit model or an under fit model.

The AUC scores for both the training data is 59% and test data 63.29%

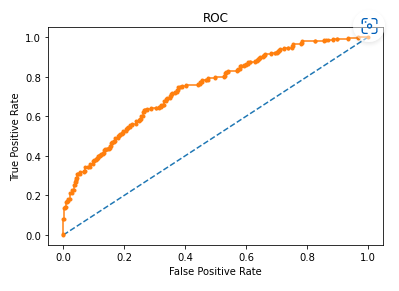
The model performance is good on F1 score as well with training data performing better at 70% while the test data gave a F1 score of 71%.

**LDA Model:**

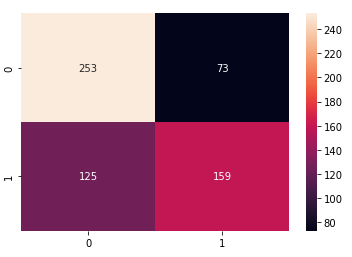
* Train data:

AUC score is 0.744 or 74.4%

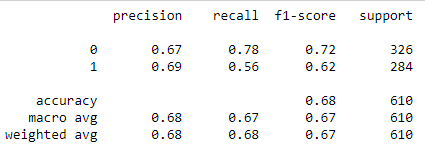
ROC Curve of train data:



Confusion Matrix on Training data:



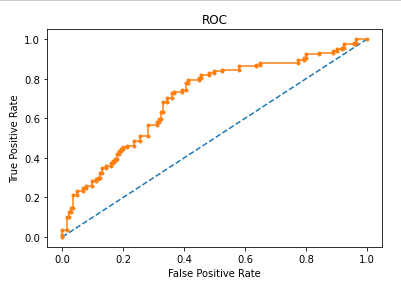
Classification report for training data:



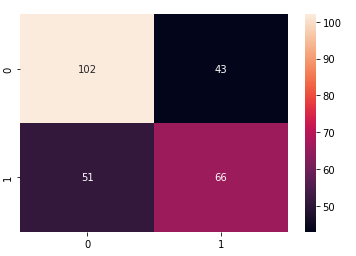
* Test data:

AUC score is 0.704 or 70.4%

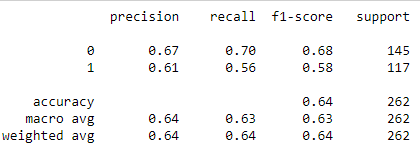
ROC Curve of test data:



Confusion Matrix on Testing data:



Classification report for testing data:



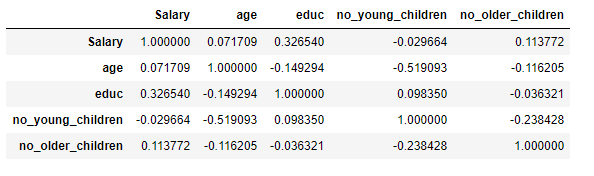
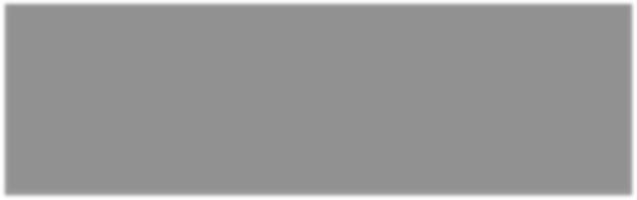
The accuracy score of the training data is 74.4% and test data is 70.4%. This is almost similar to the Logistic Regression model result so far. The AUC scores are marginally lower for the test data, else they are also almost similar to the Logistic Regression model. F1 scores are 62% and 58% for train and test data, respectively, which again is almost close to the logistic regression model.

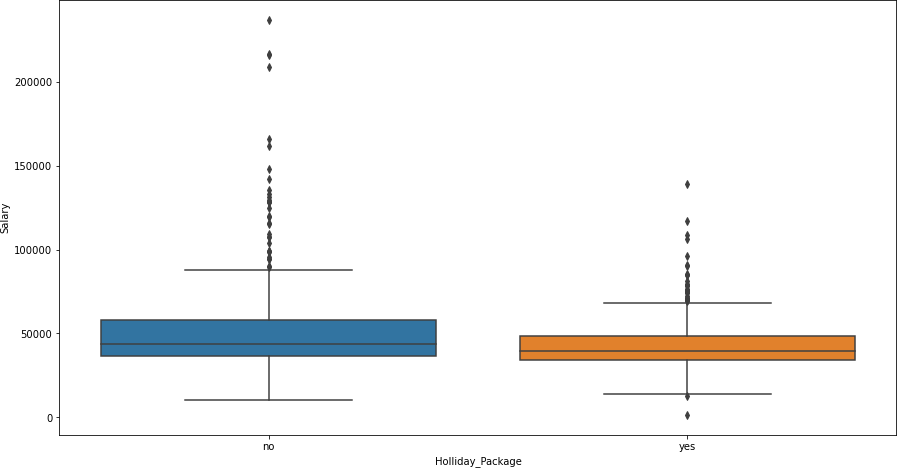
Overall, the model seems to be a right fit model and is staying away from being referred as under fit or over fit model. Let us see if we can refine the results further and improve on the F1 score of the test data specifically.

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

We started this test case with looking at the data correlation to identify early trends and patterns. At one stage, Salary and education seems to be important parameters which might have played out as an important predictor.

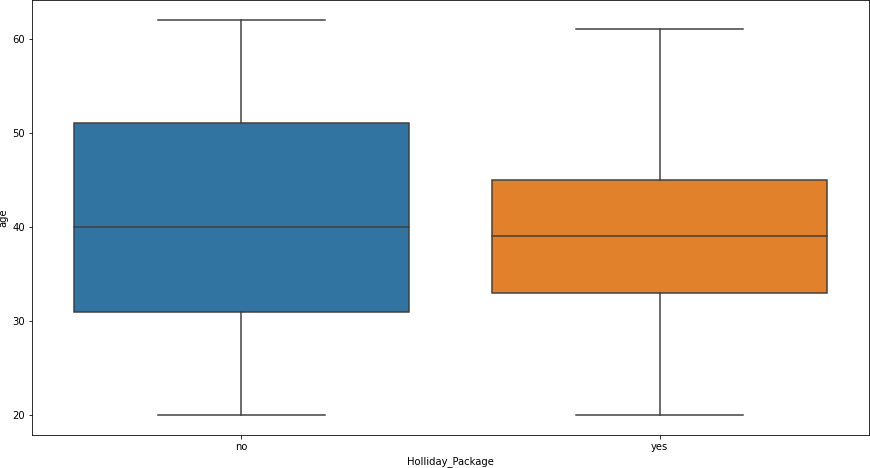
**Holiday Package & Salary**





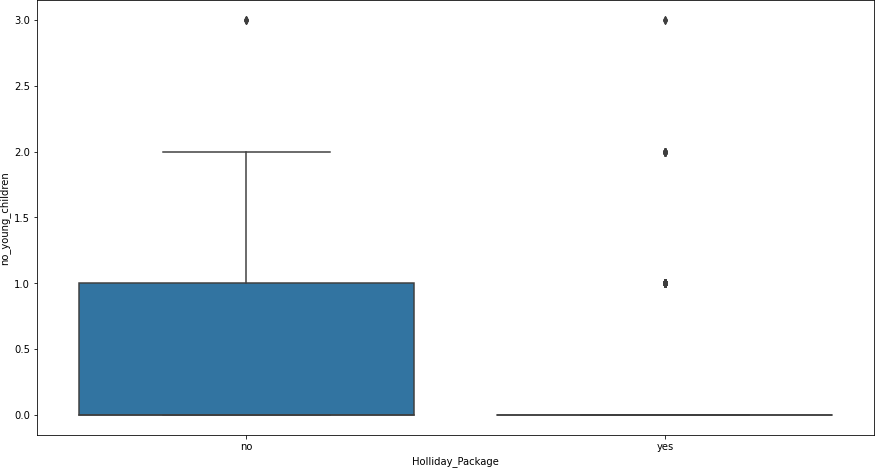
While performing the bivariate analysis we observe that Salary for employees opting for holiday package and for not opting for holiday package is similar in nature. However, the distribution is fairly spread out for people not opting for holiday packages.

**Holiday Package & age**



There are no outliers present in age

However, almost similar distribution here for salary and age is indicating that they might not come out as strong predictors after the model is created. Lets carry on with more data exploration and check.



There is a significant difference in employees with younger children who are opting for holiday package and employees who are not opting for holiday package

Interestingly and as expected by my, Salary and age didn’t turn out to be an important predictor for my model. Also, number of young children has emerged as a strong predictor (likelihood ) in not opting for holiday packages.

**As interpretation,**

1. There is no plausible effect of salary, age, and education on the prediction for Holliday\_packages. These variables don’t seem to impact the decision to opt for holiday packages as we couldn’t establish a strong relation of these variables with the target variable
2. Foreign has emerged as a strong predictor with a positive coefficient value. The log likelihood or likelihood of a foreigner opting for a holiday package is high.
3. no\_young\_children variable is negating the probability for opting for holiday packages, especially for couple with number of young children at 2.

The company can try to bin salary ranges to see if they can derive some more meaningful interpretations out of that variable. May be club the salary or age in different buckets and see if there is some plausible impact on the predictor variable. OR else, the business can use some different model techniques to do a deep dive.

**Insights:**

1. Looking into all the important parameters such as Accuracy, Recall, AUC score and ROC curve, LDA is performing better as compared to Logistic Regression mode.
2. But the accuracy that we are getting is still not good to make any predictions. Hence we should try some more models such as neural networks, random forests to choose the optimum model for our predictions.
3. We should try to gather some more data to make the model better and more robust.
4. We can change outlier treatment techniques such as scaling the data and treating those values which are above and below +3 & -3 SD respectively. It was not done here as we were not asked to scale the data

**Recommendation:**

1. The company should really focus on foreigners to drive the sales of their holiday packages as that’s where the majority of conversions are going to come in.
2. The company can try to direct their marketing efforts or offers toward foreigners for a better conversion opting for holiday packages
3. The company should also stay away from targeting parents with younger children. The chances of selling to parents with 2 younger children is probably the lowest. This also gels with the fact that parents try and avoid visiting with younger children.
4. If the firm wants to target parents with older children, that still might end up giving favorable return for their marketing efforts then spent on couples with younger children.