Business Report

Predictive Modelling

By- Varun Kumar



**** -Great Learning

Contents

[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. 5](#_Toc96235110)

[1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis. 5](#_Toc96235111)

[Introduction: 5](#_Toc96235112)

[Univariate Analysis: 8](#_Toc96235113)

[Skewness: 12](#_Toc96235114)

[Bivariate Analysis: 13](#_Toc96235115)

[Multivariate Analysis 16](#_Toc96235116)

[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? Check for the possibilities of combining the sublevels of an ordinal numbers and take action accordingly. Explain why you are combining these sublevels with appropriate reasoning. 17](#_Toc96235117)

[Scaling: 21](#_Toc96235118)

[1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks foe significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE and Adj Rsquare. Compare these models and select the best one with appropriate reasoning. 22](#_Toc96235119)

[Linear Regression using statsmodels: 25](#_Toc96235120)

[The final Linear Regression equation is 26](#_Toc96235121)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 28](#_Toc96235122)

[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. 31](#_Toc96235123)

[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. 31](#_Toc96235124)

[Introduction 31](#_Toc96235125)

[Univariate & Bivariate Analysis: 33](#_Toc96235126)

[Skewness: 36](#_Toc96235127)

[Multivariate Analysis: 37](#_Toc96235128)

[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). 39](#_Toc96235129)

[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. 40](#_Toc96235130)

[Logistic Regression: 40](#_Toc96235131)

[Linear Discriminant Analysis: 42](#_Toc96235132)

[Comparison between Logit & LDA models: 44](#_Toc96235133)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. 45](#_Toc96235134)

[Conclusion: 46](#_Toc96235135)

[Figure 1 Distribution of Cut, Colour & Clarity of Gems 8](#_Toc96235136)

[Figure 2 Diamond Clarity Chart 9](file:///D:\GREAT%20LEARNING\MOD%205%20PREDECTIVE%20MODELLING\Project\project123.docx#_Toc96235137)

[Figure 3 Distribution of Depth, Carat, Table, Price 9](#_Toc96235138)

[Figure 4 Stone Sizes 10](file:///D:\GREAT%20LEARNING\MOD%205%20PREDECTIVE%20MODELLING\Project\project123.docx#_Toc96235139)

[Figure 5 Histogram 10](#_Toc96235140)

[Figure 6 Distribution of Dimensions 11](#_Toc96235141)

[Figure 7 Carat vs Price 13](#_Toc96235142)

[Figure 8 Cut, Colour, Clarity vs Price 14](#_Toc96235143)

[Figure 9 Heatmap 16](#_Toc96235144)

[Figure 10 Pairplot 17](#_Toc96235145)

[Figure 11 Boxplot to check Outliers 18](#_Toc96235146)

[Figure 12 Boxplot Post outliers treatment 19](#_Toc96235147)

[Figure 13 Scatter plot: Actual vs Predicted 23](#_Toc96235148)

[Figure 14 Scatter plot for STATSMODEL: ACTUAL vs PREDICTED 26](#_Toc96235149)

[Figure 15 Histogram 33](#_Toc96235150)

[Figure 16 Boxplot 34](#_Toc96235151)

[Figure 17 Countplot 35](#_Toc96235152)

[Figure 18 Countplot of Holiday Package & Foreign 35](#_Toc96235153)

[Figure 19 Stacked plot 36](#_Toc96235154)

[Figure 20 Heatmap 37](#_Toc96235155)

[Figure 21 Pairplot 38](#_Toc96235156)

[Figure 22 Boxplot Before Outlier Treatment 38](#_Toc96235157)

[Figure 23 Boxplot After outlier Treatment 39](#_Toc96235158)

[Figure 24 Confusion Matrix Training Dataset 40](#_Toc96235159)

[Figure 25 AUC curve for Training Dataset 41](#_Toc96235160)

[Figure 26 Confusion Matrix for Testing Dataset 41](#_Toc96235161)

[Figure 27 AUC curve for Testing Dataset 42](#_Toc96235162)

[Figure 28 Confusion Matrix for Training Dataset (LDA) 42](#_Toc96235163)

[Figure 29 AUC curve for Training Dataset (LDA) 43](#_Toc96235164)

[Figure 30 AUC curve for Test Dataset (LDA) 44](#_Toc96235165)

[Figure 31 AUC Curve comparison for Training Data 44](#_Toc96235166)

[Figure 32 AUC curve Comparison for Test Data 45](#_Toc96235167)

[Table 1 Sample Dataset 6](#_Toc96235168)

[Table 2 Summary Statistics 7](#_Toc96235169)

[Table 3 Anomalies 12](#_Toc96235170)

[Table 4 Skewness 12](#_Toc96235171)

[Table 5 Correlation 16](#_Toc96235172)

[Table 6 Feature Encoding for 'CUT' 20](#_Toc96235173)

[Table 7 Feature Encoding for 'COLOR' 20](#_Toc96235174)

[Table 8 Feature Encoding for 'CLARITY' 20](#_Toc96235175)

[Table 9 Sample Dataset 31](#_Toc96235176)

[Table 10 Summary Statistics 33](#_Toc96235177)

[Table 11 CrossTab 36](#_Toc96235178)

[Table 12 Skewness Of dataset 37](#_Toc96235179)

[Table 13 Classification Report for Training Dataset 40](#_Toc96235180)

[Table 14 Classification Report for Testing Dataset 41](#_Toc96235181)

[Table 15 Classification Report for Training Data (LDA) 42](#_Toc96235182)

[Table 16 Classification Report for Testing Dataset (LDA) 43](#_Toc96235183)

[Table 17 Confusion Matrix Testing Dataset (LDA) 43](#_Toc96235184)

[Table 18 Comparison between Logit & LDA models 44](#_Toc96235185)

**Problem 1**: Linear Regression

# 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.

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

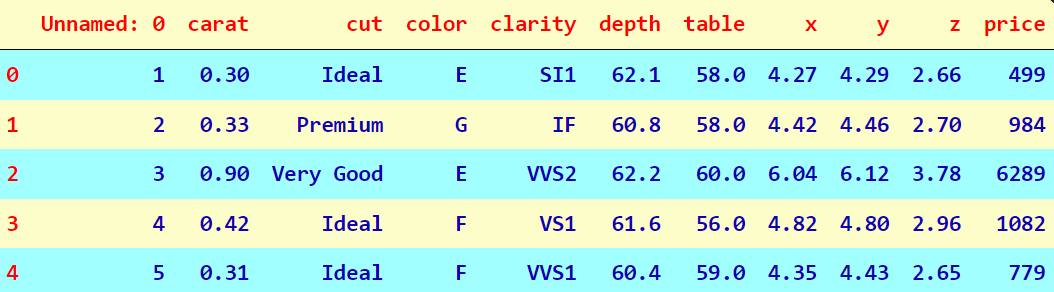
### Introduction:

Gem Stones co ltd, is a company which is a cubic zirconia manufacturer. The dataset contains 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 price slots. The data must be used to predict the price for the stone on the bases of the details given in the dataset to make it easier to distinguish between higher profitable stones and lower profitable stones to have better profit share.

The data first needs to be cleaned, followed by exploratory data analysis. Only after these two steps are completed can we use the data for modelling.

#### Sample Dataset:

Table 1 Sample Dataset



#### Data Description:

The dataset consists of data on sale of 26967 different zirconia. Each sale has 10 different features associated with it. The various features in the dataset are as follows:

* Carat – Carat weight of the cubic zirconia (26967 entries, datatype float64).
* Cut - Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal (26967 entries, datatype – object).
* Color – Color of the cubic zirconia. With D being the best and J the worst (26967 entries, datatype object).
* 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 (26967 entries, datatype – object).
* Depth - the Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter. (26270 entries, datatype -float64)
* Table – the width of the cubic zirconia’s table expressed as a percentage of its average diameter (26967 entries, datatype - float64)
* X – length of the cubic zirconia in mm (26967 entries, datatype - float64)
* Y – width of the cubic zirconia in mm (26967 entries, datatype - float64)
* Z – height of the cubic zirconia in mm (26967 entries, datatype - float64)
* Price – the price of the cubic zirconia (26967 entries, datatype - int64)

#### Data Cleaning:

Checking for missing values: There are 697 missing values for the feature depth. 50 percent of the data for the feature depth lies between the values 61 and 62.5 and has outliers. Therefore, a decision was taken to impute the missing values of depth with the median value of the feature i.e., 61.80.

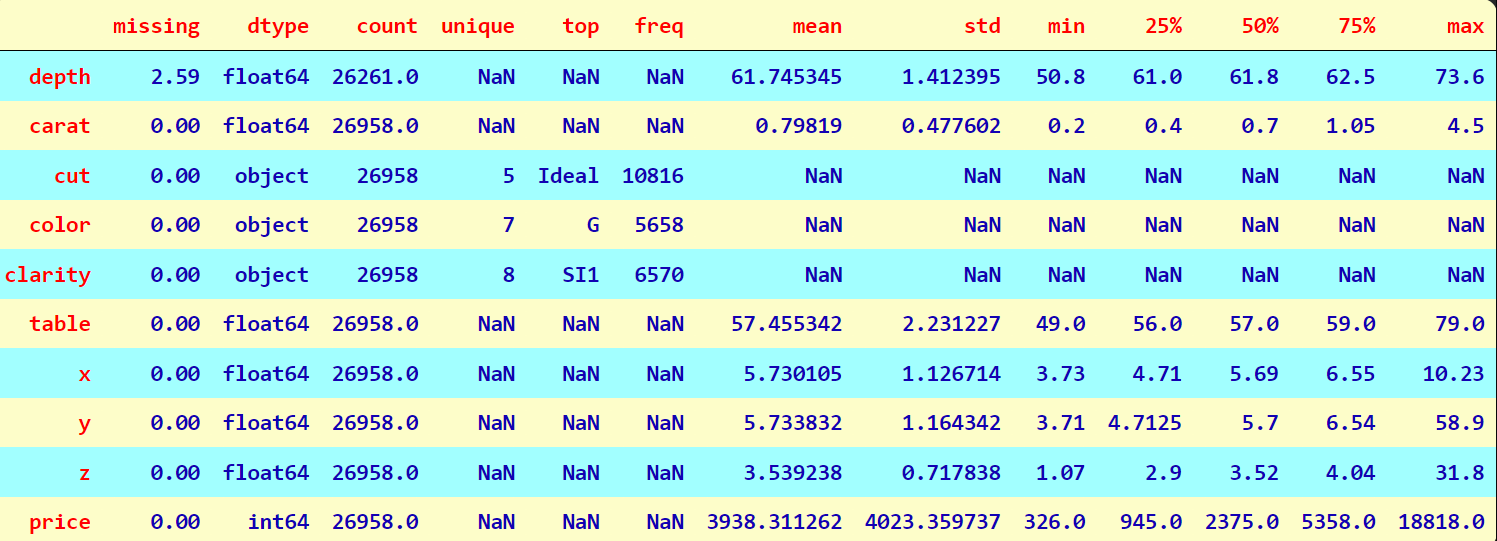
Checking for duplicates: There are zero duplicated rows.

Checking for anomalous data: the minimum values of x, y and z for some entries is 0. This is not physically possible to have a zirconia stone which has 0 length or 0 breath or 0 width. There are 9 such instances. The price range for these 9 instances varies to a great extent, as a result imputing with mean will not be the right solution. Therefore, decision has been taken to drop these values.

Outlier Treatment: The data has a high number of outliers across all the variables. Given that the price of the zirconia stone needs to be predicted, removing outliers will not make much business sense, as outliers can give meaningful insights. But for now, we are going to treat outliers to build an efficient model.

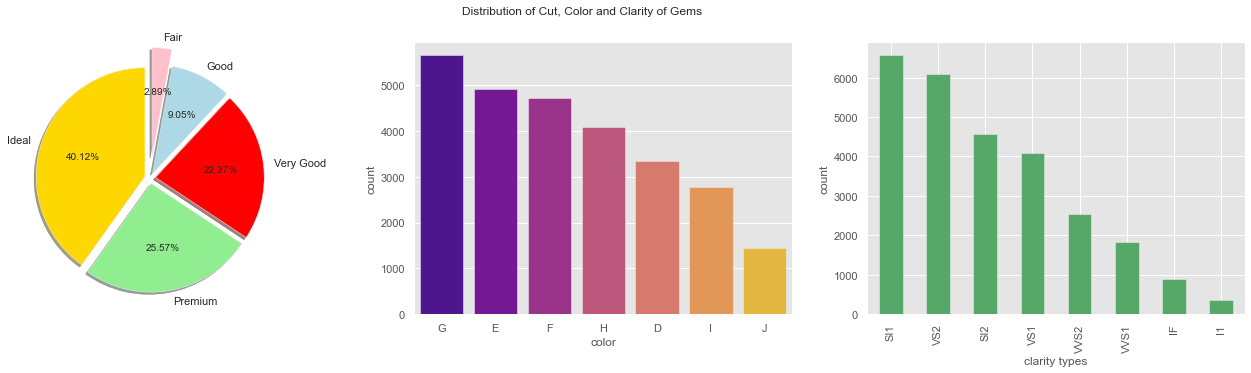
#### Summary Statistics

Table 2 Summary Statistics



### Univariate Analysis:

Figure 1 Distribution of Cut, Colour & Clarity of Gems



Cut: The categorical variable cut has 5 unique values. The most frequent cut in the dataset is the ‘Ideal’ one and the least frequent cut observed is the ‘Fair’ one. From the Swarovski website (Swarovski is the leading zirconia manufacturer and retailer in the world) we understand that cut refers to the proportions of symmetry of the gemstone, in descending order of value the order is:

Ideal > Premium > Very Good > Good > Fair

Color: The most frequent color is G and the least frequent one is J. from the Swarovski website (Swarovski is the leading zirconia manufacturer and retailer in the world) we understand that color which is uniform, vivid and saturated determines a higher value gemstone, D corresponds to the most valuable color and subsequent alphabets represent a reducing order with J being the least valuable.

This will help us in encoding the data to integer values before creating a Linear Regression Model.

Clarity: The most frequently observed entry is ‘SI1’ and the least frequently observed is ‘I1’. Further from the GIA (Gemological Institute of America) website, we get the following order on clarity of gemstones:

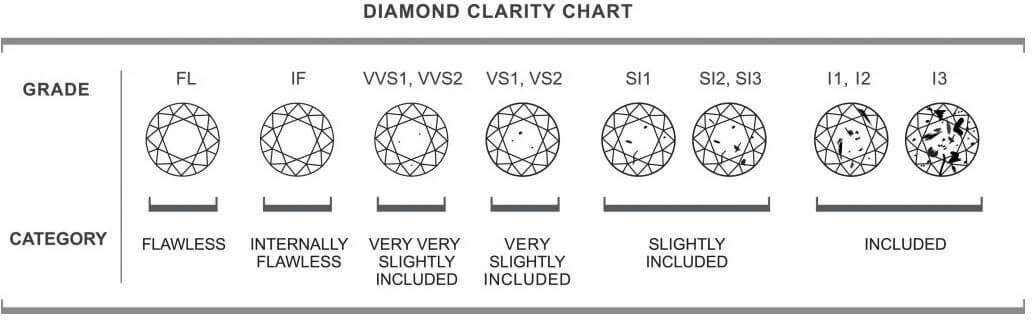
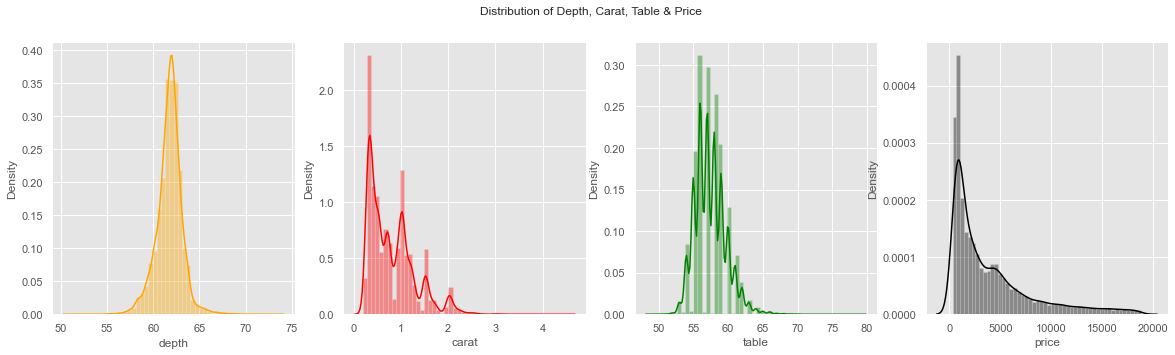


Figure Diamond Clarity Chart

Carat: The distribution spikes at 0.3,1, 1.5 and 2. The distribution is also right skewed. There is a presence of large number of outliers. 0.3 carat is a frequently occurring value followed by the 1.01 value. The minimum values are around 0.2 and maximum around 2 carats.

Figure 3 Distribution of Depth, Carat, Table, Price



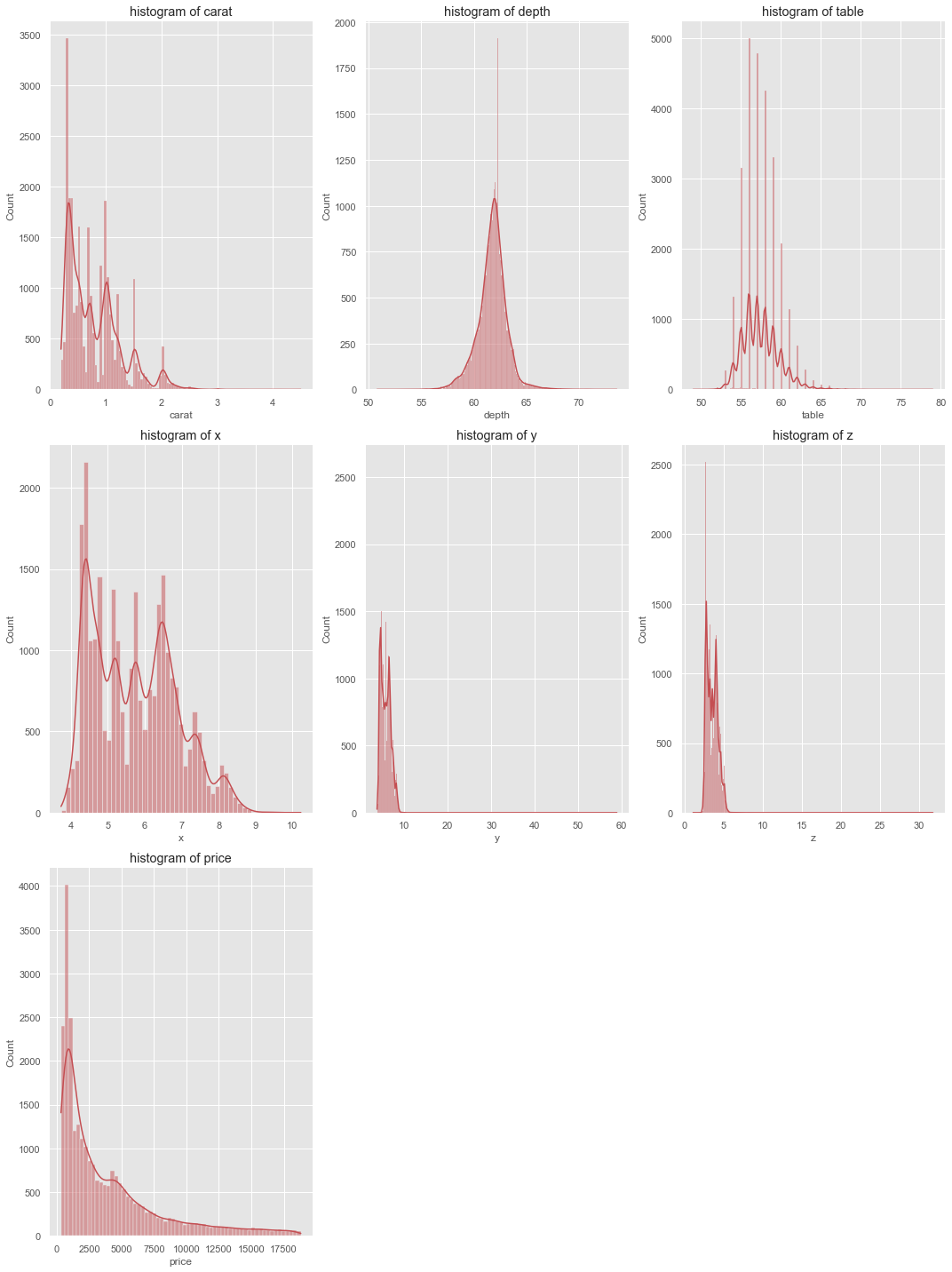
Further from the Swarovski website (Swarovski is the leading zirconia manufacturer and retailer in the world) we get the following understanding about carat, which is a measure of the size of the stone:



Figure Stone Sizes

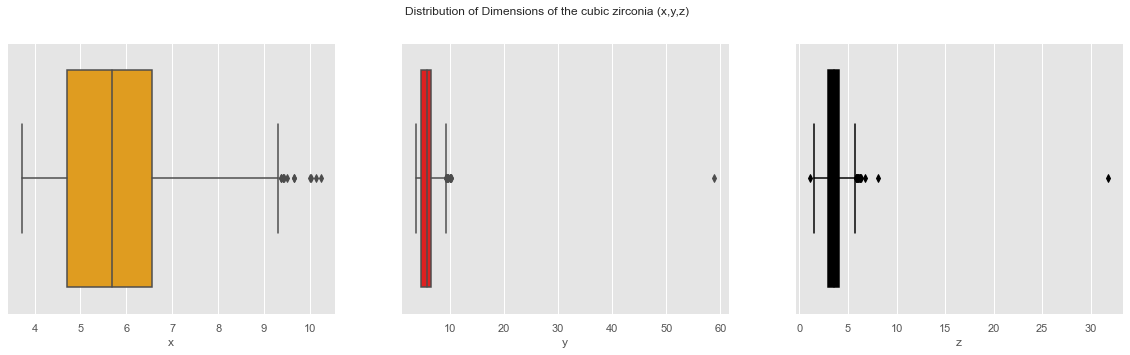
Depth: The distribution follows a near normal distribution with long tails both on the right side and the left side. The number of outliers is large. The value of depth in the dataset varies from 50.8 to 73.6. The common value for depth is in the 60 range, please see histogram below to understand the distribution.

Figure 5 Histogram

Table: The next variable “table” which represents the Width of the cubic zirconia expressed as a Percentage of its Average Diameter between 49 and 79. The most common value is 56. The distribution has multiple spikes at 53, 55,60 and 62.5. The distribution has a small left side tail and a long right-side tail.

X, Y & Z: The distribution of X too has various spikes, but the number of outliers is small compared to other features. “X” which stands for Length of the cubic zirconia in mm, ranges between 0 and 10.23.

Figure 6 Distribution of Dimensions

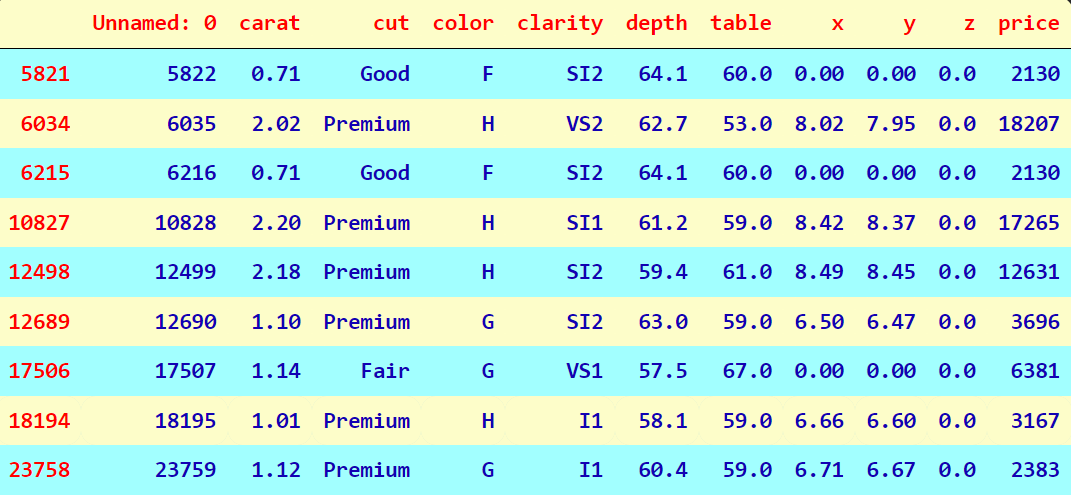


The distribution has an extremely long right-side tail because of one outlier. The distribution of Y is a compact normal distribution. Variable “y” which stands for width of the cubic zirconia in mm ranges between 0 and 58.9. The extreme outlier value is 58.9.

Like the Y distribution, this to, has a long right-sided tail because of one outlier. Rest of the distribution is a compact distribution in the range of 0 to 5. The variable “z” Height of the cubic zirconia in mm ranges between 1.07 and 31.8 but once again, 31.8 is observed to be an outlier. From the Histogram once again, we see most values clubbed in a narrow range.

As all 3 variables “x”, “y” and “z” are shown as 0 this appears to be an anomaly and as all three variables have outliers as shown in boxplot below, we drop these nine records with 0 values as dimensions cannot be 0 and dropping nine values doesn’t make any impact on our model. We call for the data with these values.

Table 3 Anomalies



Price: The distribution of the target feature is right skewed with large number of outliers. The last variable “price” our dependent variable has values ranging from 326 to 18818.

### Skewness:

Table 4 Skewness

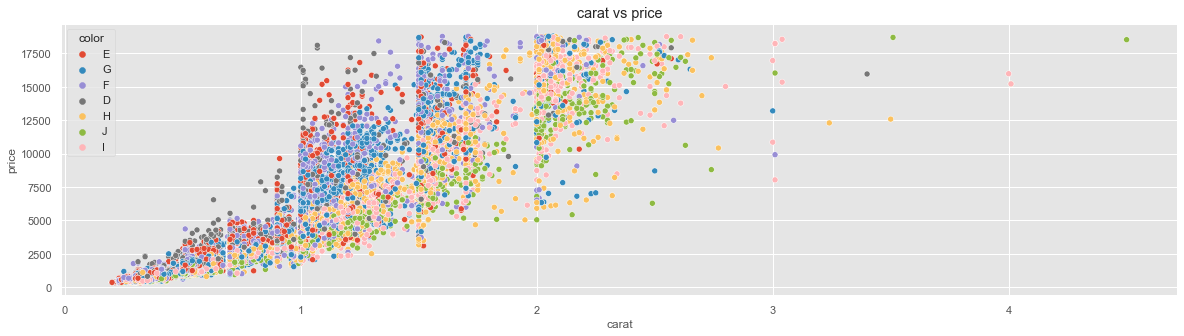
|  |  |
| --- | --- |
| Carat | 1.12 |
| Depth | -0.03 |
| Table | 0.76 |
| X | 0.40 |
| Y | 3.88 |
| Z | 2.63 |
| Price | 1.62 |

### Bivariate Analysis:

#### Carat vs Price:

If carat increases price is also increased.

Figure 7 Carat vs Price

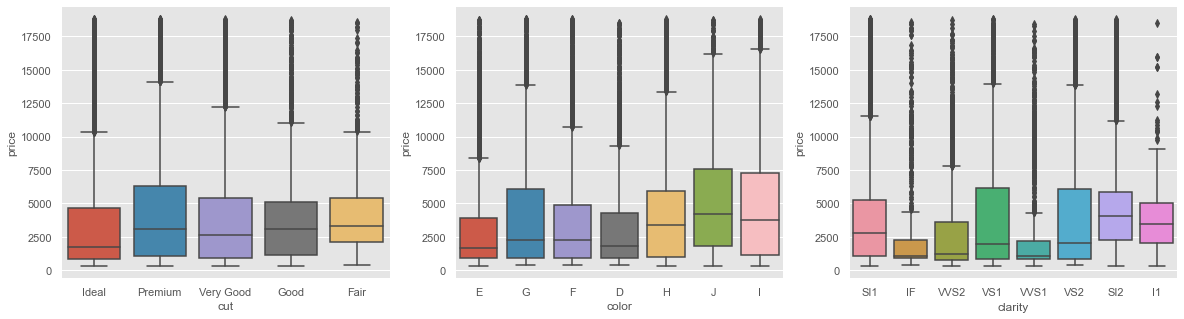


#### Dimensions vs Price:



#### Cut, Colour, Clarity vs Price:

Figure 8 Cut, Colour, Clarity vs Price



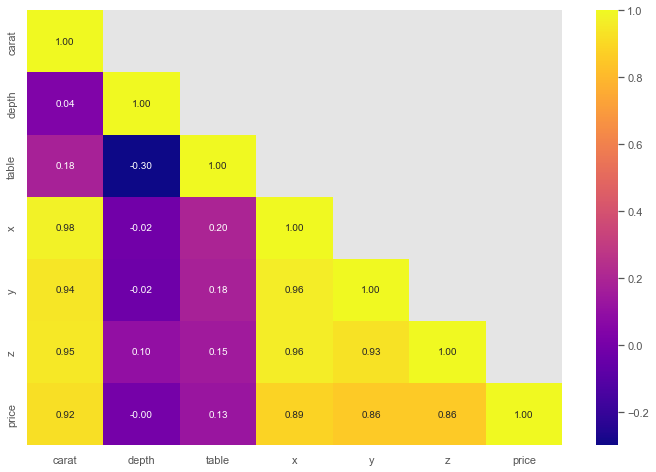
We create pivot tables to understand the three categorical variables viz z viz the dependent variable price better;

|  |  |
| --- | --- |
|  | We observe that the max price is for ‘Very Good’ cut followed by ‘Ideal’ this is interesting as literature identifies ‘Ideal’ as a more valuable cut than ‘Very Good’. This anomaly is consistent with the mean price of ‘Very Good’ zirconia also which is higher than the mean price of ‘Ideal’ zirconia. In fact, the highest mean price is fetched by ‘Fair’ zirconia.  Perhaps the stones of Fair, and very good cut were good on other attributes and hence scored a higher price. |
|  | The max & min price is by SI1 type.  The highest mean price is by SI2 type.  In the minimum price category, IF at number one position with 369 minimum price. |
|  | G color has the max price Whereas J color (the poorest as per literature) has the highest average price. |

These pivot tables show how the best in each individual attribute does not correlate with the highest price or even the highest average price, Perhaps, each gem is evaluated for all the attributes together and not individually hence the variation.

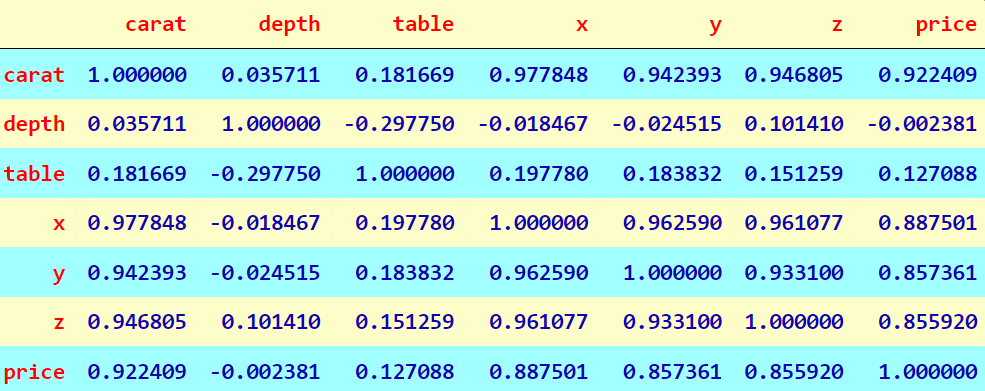
Multivariate Analysis:

Figure 9 Heatmap



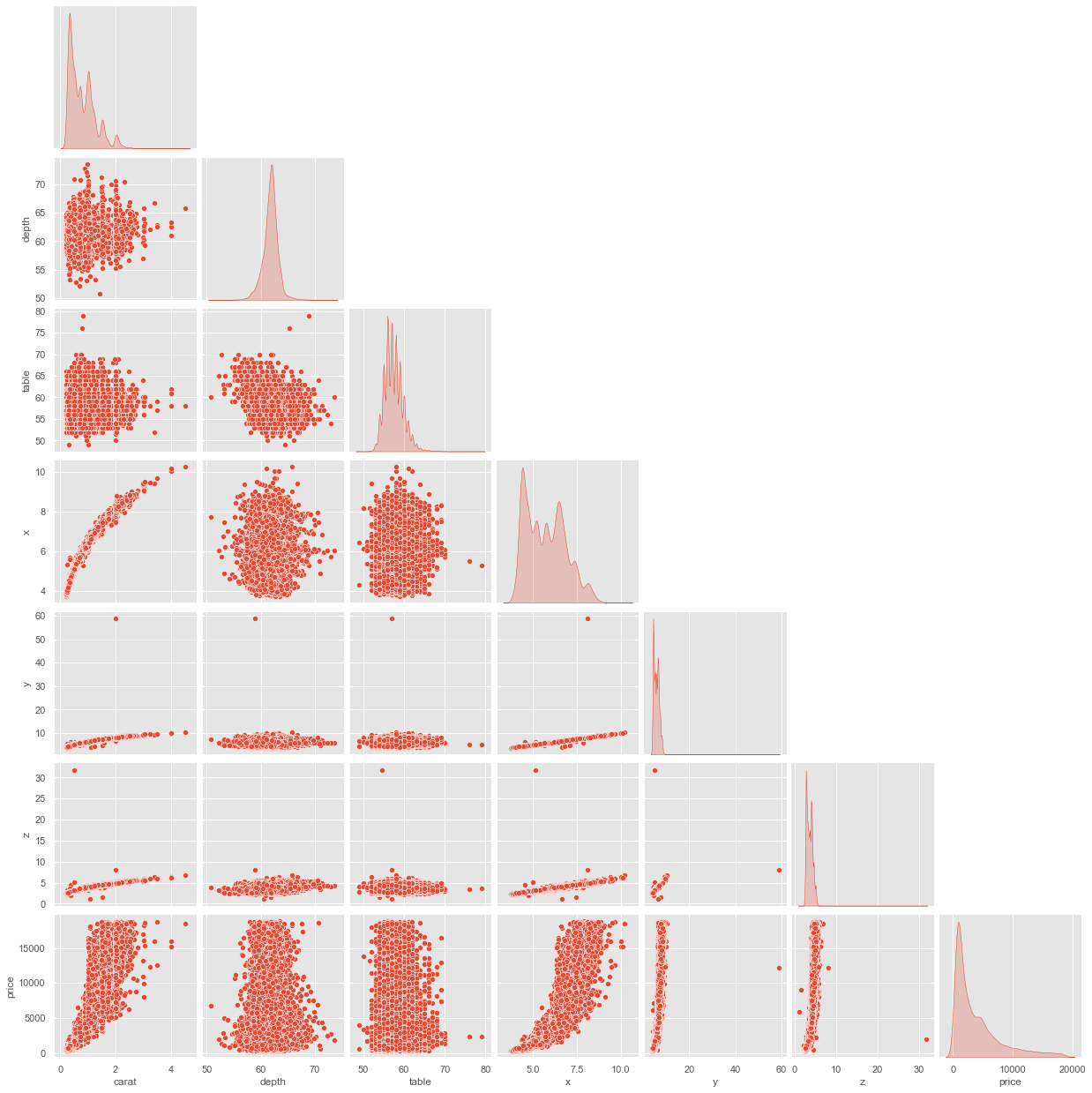
Correlation heatmap: From the heatmap one can observe that there is high correlation between the features x, y, z and between these variables and price as well as carat. Carat and price have a very low correlation with features such as cut, color, clarity, depth, and table. We can also analyze the correlations from the table given below:

Table 5 Correlation



Pairplot: The observations made above can be visualized with the Pairplot as well. The relationship of numerical variables with each other can also be visualized through the pair plot.

Figure 10 Pairplot



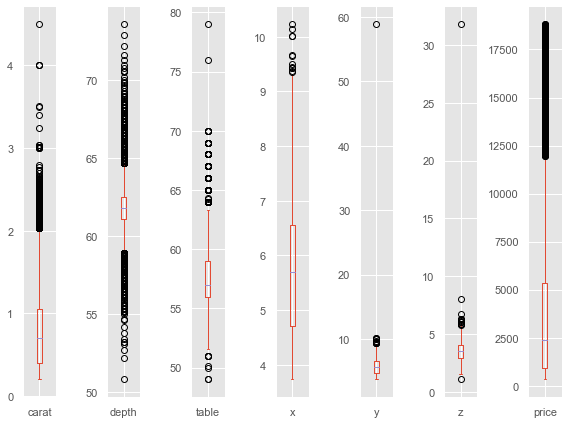
## 1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibilities of combining the sublevels of an ordinal numbers and take action accordingly. Explain why you are combining these sublevels with appropriate reasoning.

We find the “depth” variable to have 697 null values. All other variables do not show any null values. We fill these null values using median because depth has outliers, so, median seems to be appropriate to fill these null values.

This dataset has nine records with zero values in the dimensions. As we seen during EDA that all the three dimensions, i.e., x, y, z has zero values and we know that it’s not possible for a gem to be dimensionless. So, these are the faulty records and we choose to drop them as dropping of nine records out of almost 27000 records doesn’t impact our model.

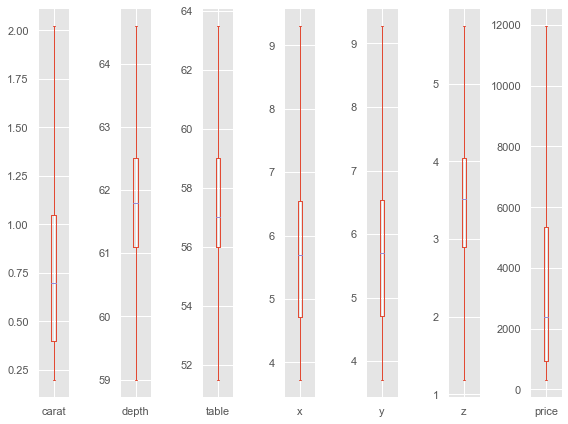
#### Checking for outliers:

Figure 11 Boxplot to check Outliers



As we can observe there is good number of outliers present in this data. Given that the price of the zirconia stone needs to be predicted, removing outliers will not make much business sense, as outliers can give meaningful insights. But for now, we are going to treat outliers to build an efficient model. I treat the outliers by replacing the higher side outliers with upper range (Q3+1.5(IQR)) and lower side outliers with lower range(Q1-1.5(IQR)). Boxplot post outlier treatment is below:

Figure 12 Boxplot Post outliers treatment



#### Ordinal numbers:

In this dataset, there is a proper ordered classes in the features. Few examples are following:

* Feature ‘Clarity’ has order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
* Feature ‘Cut’ has also proper pricing values on the bases of different cuts.

Ideal > Premium > Very Good > Good > Fair

* Gems pricing is also depending on feature color. With D being the best and J the worst.

We can combine few of the classes like in feature ‘Clarity’ we can combine SI1 & SI2 or in feature ‘color’ the worst considered colors I & J can be combined.

But during EDA we find that the price of gems is strongly correlated to these features and classes and we don’t have any expert knowledge on gems as well, therefore, we choose not to combine these classes. Instead, we are encoding them according to their importance for pricing as we mentioned during EDA as well. Encoded done for the different features are following:

Table 6 Feature Encoding for 'CUT'

|  |  |
| --- | --- |
| CUT | |
| Fair | 0 |
| Good | 1 |
| Very Good | 2 |
| Premium | 3 |
| Ideal | 4 |

Table 7 Feature Encoding for 'COLOR'

|  |  |
| --- | --- |
| COLOR | |
| J | 0 |
| I | 1 |
| H | 2 |
| G | 3 |
| F | 4 |
| E | 5 |
| D | 6 |

Table 8 Feature Encoding for 'CLARITY'

|  |  |
| --- | --- |
| CLARITY | |
| I1 | 0 |
| SI2 | 1 |
| SI1 | 2 |
| VS2 | 3 |
| VS1 | 4 |
| VVS2 | 5 |
| VVS1 | 6 |
| IF | 7 |

We need to convert the data types of these features from ‘object’ to ‘float64’ as well.

### Scaling:

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 is in 1000s unit and depth and table are in 100s unit, and carat is in 10s. So, it’s 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 its advisable to go with unscaled data because linear regression equation of unscaled data can give us an accurate picture.

We are going to build the model on both, Normal dataset and scaled dataset.

## 1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks foe significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE and Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

The data has been split into 70:30 ratio, with 70% of it being allocated for training the linear regression model and 30% of it is allocated for testing the model. We undertook model building with both scaled and unscaled data below are the finding from first the unscaled data and then the scaled data;

#### Unscaled Data Model:

We split the data into a 70:30 train test split and first run the linear Regression model on sklearn, we get the following coefficients:

* The coefficient for carat is 8768.12.
* The coefficient for cut is 111.21.
* The coefficient for color is 269.09.
* The coefficient for clarity is 431.10.
* The coefficient for depth is 42.99.
* The coefficient for table is -13.06.
* The coefficient for x is -1113.48.
* The coefficient for y is 1477.35.
* The coefficient for z is -1157.88.

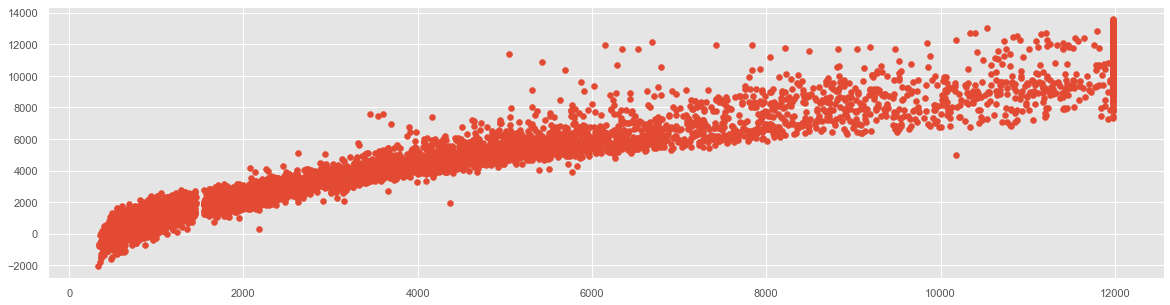
Observation:

Y=mx +c (m= m1, m2, m3...m9) here 9 different co-efficient along 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 8768.12.
* The one unit increase in cut increases price by 111.21.
* The one unit increase in color increases price by 269.09.
* The one unit increase in clarity increases price by 431.10.
* The one unit increase in depth increases price by 42.99,
* But the one unit increase in table decreases price by -13.06
* The one unit increase in x decreases price by -1113.48
* The one unit increase in y increases price by 1477.35.
* The one unit increase in z decreases price by -1157.88.

Figure 13 Scatter plot: Actual vs Predicted



#### Observation:

We can see that there 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: -5669.81.
* R square on training data: 0.93.
* R square on testing data: 0.93.
* RMSE on Training data: 907.42.
* RMSE on Testing data: 913.87.
* As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

#### Scaled Data Model:

We scaled this data using zscore. Below are the co-efficient from the model build on scaled data:

* The coefficient for carat is 1.17.
* The coefficient for cut is 0.04.
* The coefficient for color is 0.13.
* The coefficient for clarity is 0.20.
* The coefficient for depth is 0.02.
* The coefficient for table is -0.008.
* The coefficient for x is -0.36.
* The coefficient for y is 0.48.
* The coefficient for z is -0.23.

Linear regression Performance Metrics:

* Intercept for the model: 6.978190484948235e-16.
* R square on training data: 0.93.
* R square on testing data: 0.93.
* RMSE on Training data: 0.26.
* RMSE on Testing data: 0.26.
* As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Observation:

Now we can observe by applying z score the intercept became 6.978190484948235e-16. Earlier it was -5669.81. The co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

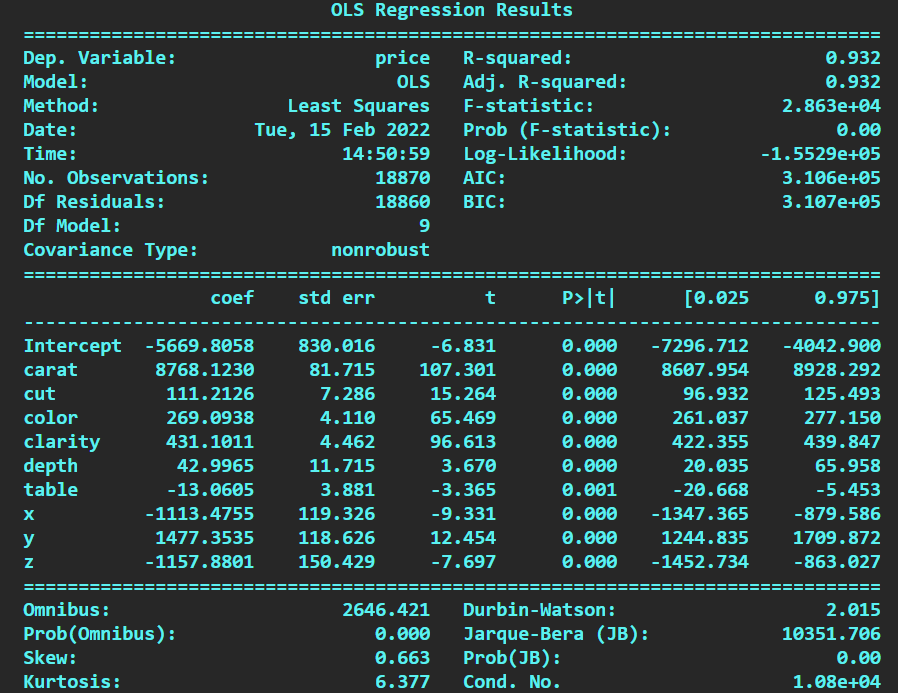
The variance inflation factor values for the various features are:

* carat ---> 122.76.
* cut ---> 10.30.
* color ---> 5.54.
* clarity ---> 5.46.
* depth ---> 1219.13.
* table ---> 874.31.
* x ---> 10685.46.
* y ---> 9435.55.
* z ---> 3315.68.

A value of 1 indicates that there is no correlation between the variables. A value between 1 and 5 suggest that there is a moderate correlation, small enough not to warrant any serious correction. Values greater than 5 represent critical levels of multicollinearity where the coefficients can be poorly estimated. For the purpose of this report, we are not going to treat the multicollinearity.

### Linear Regression using statsmodels:

While this does look like the model seems to be the right fit as scores are not dipping from training set to test set, we need to look at other statistical scores which we will get using stats model library, consider these statistical scores below:



Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 1.08e+04. This might indicate that there are strong multicollinearity or other numerical problems.

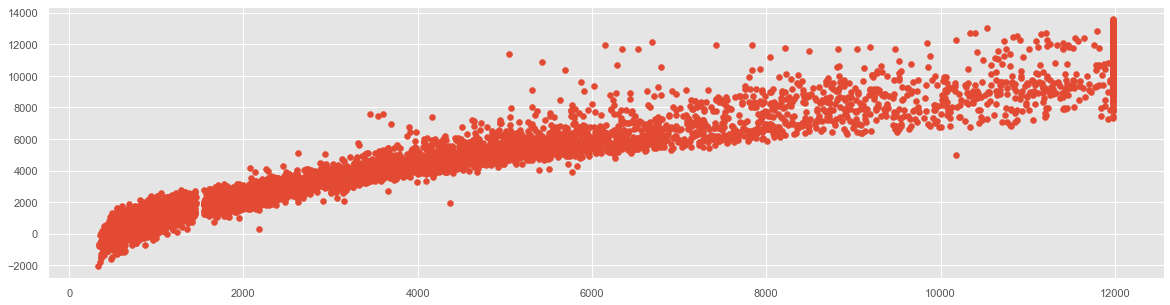
Looking at the P values, we can observe that all the p values less than 0.05. This means that we have sufficient evidence in the data to reject the null hypothesis (i.e., there is no relation between the independent variables and the target variable in the universe).

The R-squared and the adjusted R-squared values are same, 0.932. This signifies a good performance of the linear regression model.

The root mean squared error or RMSE for the train data is 907.42 and for the test data is 913.87.

The following scatter plot helps us to visualize the actual price values vs the predicted values of the price.

Figure 14 Scatter plot for STATSMODEL: ACTUAL vs PREDICTED



### The final Linear Regression equation is

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

price = (-5669.81) \* Intercept + (8768.12) \* carat + (111.21) \* cut + (269.09) \* color + (431.1) \* clarity + (43.0) \* depth + (-13.06) \* table + (-1113.48) \* x + (1477.35) \* y + (-1157.88) \* z

Based on the coefficients above we can conclude that the most important features are:

Carat (+ve effect on price)

Y (+ve effect on price)

Clarity (+ve effect on price)

Color (+ve effect on price)

Cut (+ve effect on price)

The following conclusion can be drawn from the above observations:

* When carat increases by 1 unit, diamond price increases by 8768.12 unit, keeping all other predictors constant.
* When cut increases by 1 unit, diamond price increases by 111.21 unit, keeping all other predictors constant.
* When color increases by 1 unit, diamond price increases by 269.09 unit, keeping all other predictors constant.
* When clarity increases by 1 unit, diamond price increases by 431.1 unit, keeping all other predictors constant.
* When depth increases by 1 unit, diamond price increases by 43.0 unit, keeping all other predictors constant.
* When y increases by 1 unit, diamond price increases by 1477.35 unit, 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 (-1113.48) for 'x’, (-1157.88) for z and (-13.06) 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 stone.
* Similarly, for the 'z' variable having negative co-efficient i.e., -1157.88. 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 stone.
* 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 stone.
* 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 positively.

## 1.4 Inference: Basis on these predictions, what are the business insights and 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 is 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.
* From this model building exercise “carat” emerges as the key variable that determines price. Hence the manufacturer should focus on this aspect of stone, the appropriate carat value identification deserves further data analysis; one aspect worth noting is that 1 carat and 0.3 carat are both more present in the dataset.
* Further we had noticed during EDA that the stone fetching the highest price was not necessarily the one with highest color, or clarity or cut; in fact, this area needs some exploration; perhaps it will be worthwhile to cluster the dataset and then build models for each cluster for more accurate predictions.
* From this dataset we can also conclude that the market values mid value chain rather than premium stones in color and clarity. While D is the highest rated color it is not the most popular one, similarly the S1 clarity is preferred over IF clarity.
* The size of the stone also plays a key role in determining the price hence it will be worthwhile to invest in stones of a larger dimension, after doing an exploratory analysis of customer sentiment.
* Further the following pivot tables shed some light.

|  |  |
| --- | --- |
|  | We observe that the max price is for ‘Very Good’ cut followed by ‘Ideal’ this is interesting as literature identifies ‘Ideal’ as a more valuable cut than ‘Very Good’  . This anomaly is consistent with the mean price of ‘Very Good’ zirconia also which is higher than the mean price of ‘Ideal’ zirconia. In fact the highest mean price is fetched by ‘Fair’ zirconia.  Perhaps the stones of Fair, and very good cut were good on other attributes and hence scored a higher price. |
|  | The max price is by SI1 type.  The highest mean price is by SI2 type.  The IF category gets the highest minimum price. |
|  | G color has the max price. Whereas J color (the poorest as per literature) has the highest average price. |

**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.

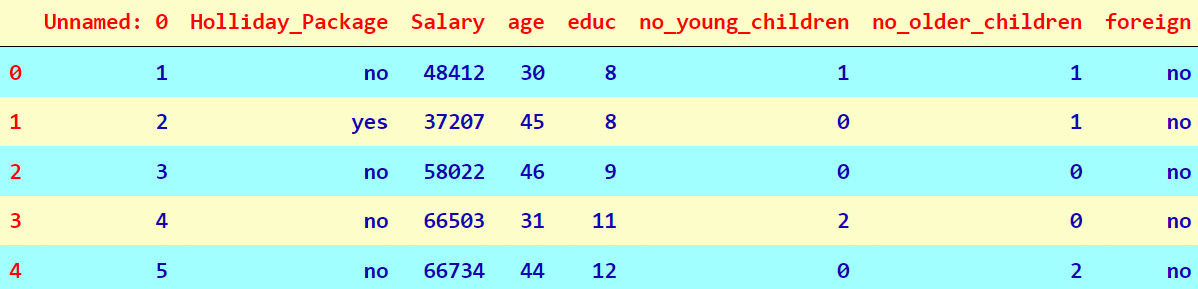
## 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

### Introduction

The data first needs to be cleaned, followed by exploratory data analysis. Only after these two steps are completed can we use the data for modelling.

#### Sample dataset:

Table 9 Sample Dataset



#### Data Description:

The dataset consists of data on 872 employees, some of them have opted for the holiday package and some haven’t. Each employee has 7 different features associated with them. The various features in the dataset are as follows:

1. Holliday\_Package – opted for holiday package yes/no? (872 entries, datatype - object)
2. Salary – Employees Salary (872 entries, datatype – int64)
3. age – age of employees (872 entries, datatype – int64)
4. educ – number of years of education of the employee (872 entries, datatype – int64)
5. no\_young\_children - The number of young children (younger than 7 years). (872 entries, datatype-int64)
6. no\_older\_children – number of older children (872 entries, datatype – int64)
7. foreign – foreigner (872 entries, datatype - object)

#### Data Cleaning:

Checking for missing values: there are no missing values in the dataset.

Checking for duplicates: there are no duplicated rows.

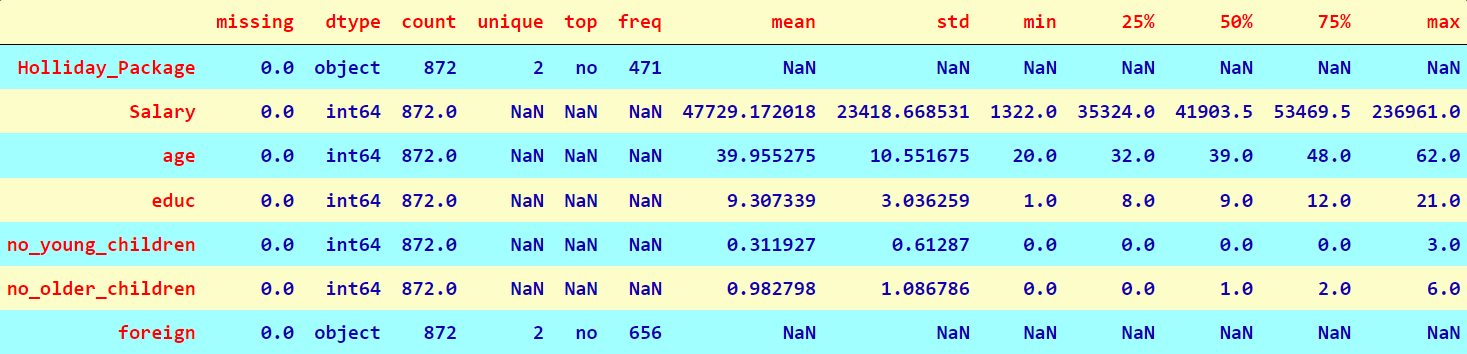
Checking for anomalous data: There are no anomalous values in the dataset.

Outlier Treatment: The data has a few outliers. The same will be treated later on. The feature values that are greater than Q3+1.5IQR will be substituted with the 100th percentile values. The feature values that are smaller than Q1– 1.5IQR will be substituted with the 0th percentile values.

#### Summary Statistics:

After dropping the column ‘Unnamed:0’, we get the following description:

Table 10 Summary Statistics



### Univariate & Bivariate Analysis:

Figure 15 Histogram

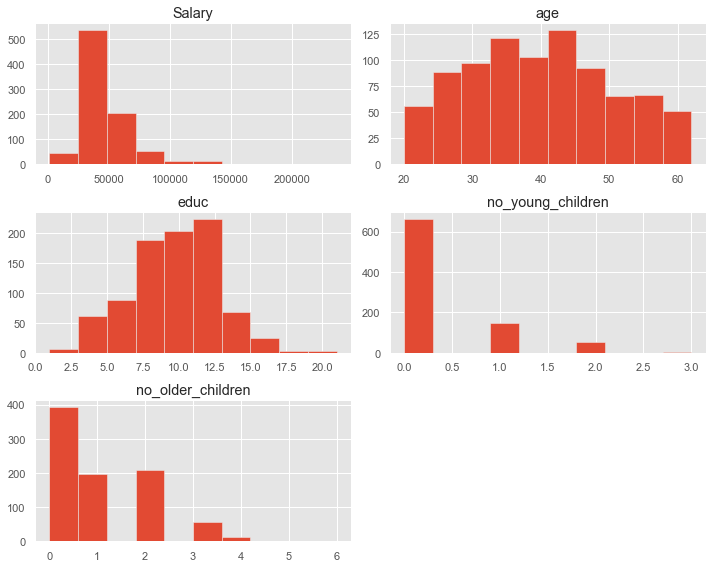
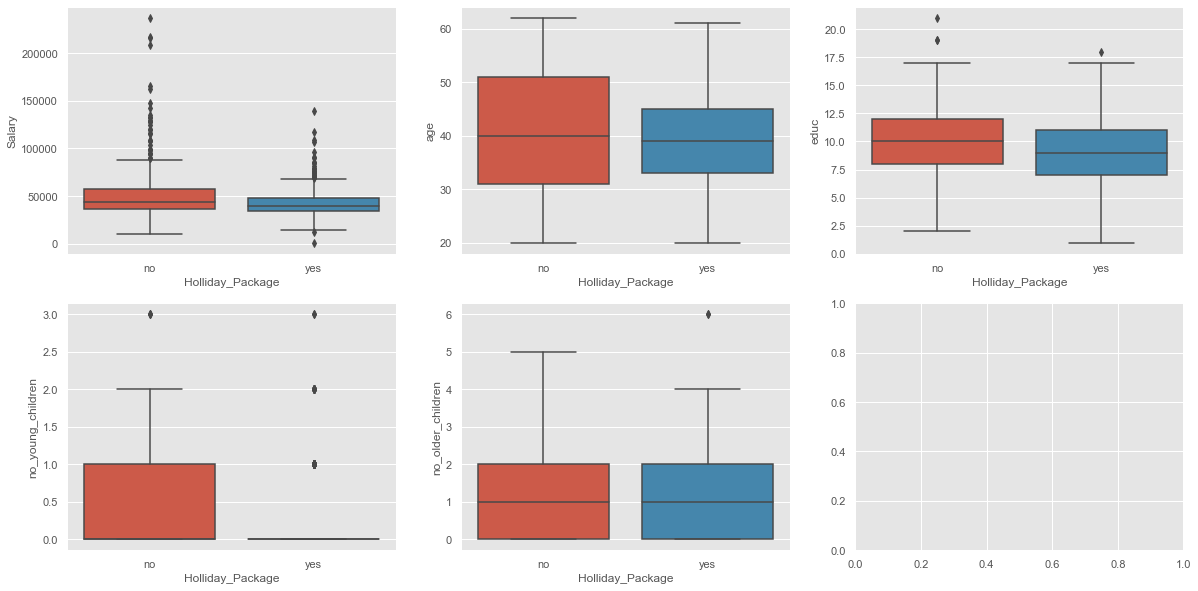


Figure 16 Boxplot



Salary: follows a near normal distribution with a long tail on the right sided. The data is right skewed and has quite a significant number of outliers. The salary is spread between 1322 to 236961.

Age: The minimum value is 20 and the maximum value is 62. The data spikes at around 37. The feature has no outliers.

Educ: The distribution has multiple spikes at 5, 8 and 12. The data has only 3 outliers. The variable names “educ” that denotes years of formal education is spread between 1 and 21; with most values being between 8-12.

No\_young\_children: The highest frequency is 0, with few employees having 1, few having 2 and one having 3.

No\_older\_children: The highest frequency is 0, with some employees having 1, some 2, 3 and so on. There is one outlier who has 6 children The variable “no\_older\_children” is between 0-6 and most values are 0.

Holiday Package Acceptance and Foreigner vs Numerical Variables: The boxplots and countplots help us visualize how the holiday package acceptance of employees varies across features of Salary, age, educ, number of young children, number of old children.

Looking at the 4th boxplot we can observe that employees who have less than 2 children below the age of 7 on a broad level do not opt for holiday packages. Employees opting for the holiday package have no children below the age of 7, void three outliers.

Foreigners across the dataset have lesser salary, lesser number of years in education compared to the natives across the data set. Foreigners accepting the holiday package have mean of years of formal education lesser than natives accepting the holiday package.

Figure 17 Countplot

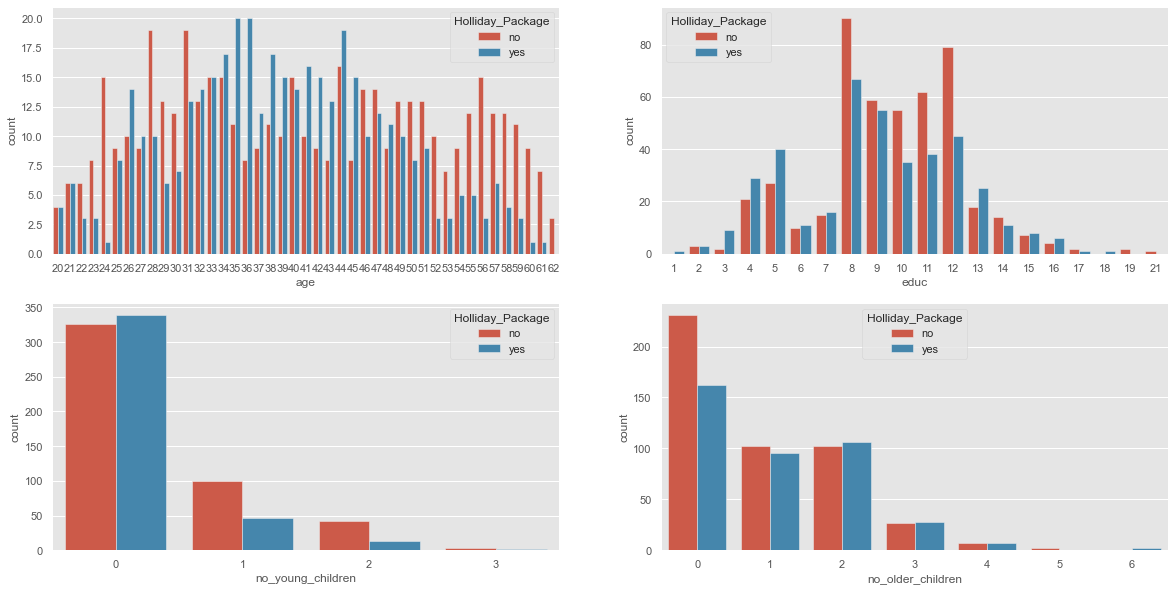


Figure 18 Countplot of Holiday Package & Foreign

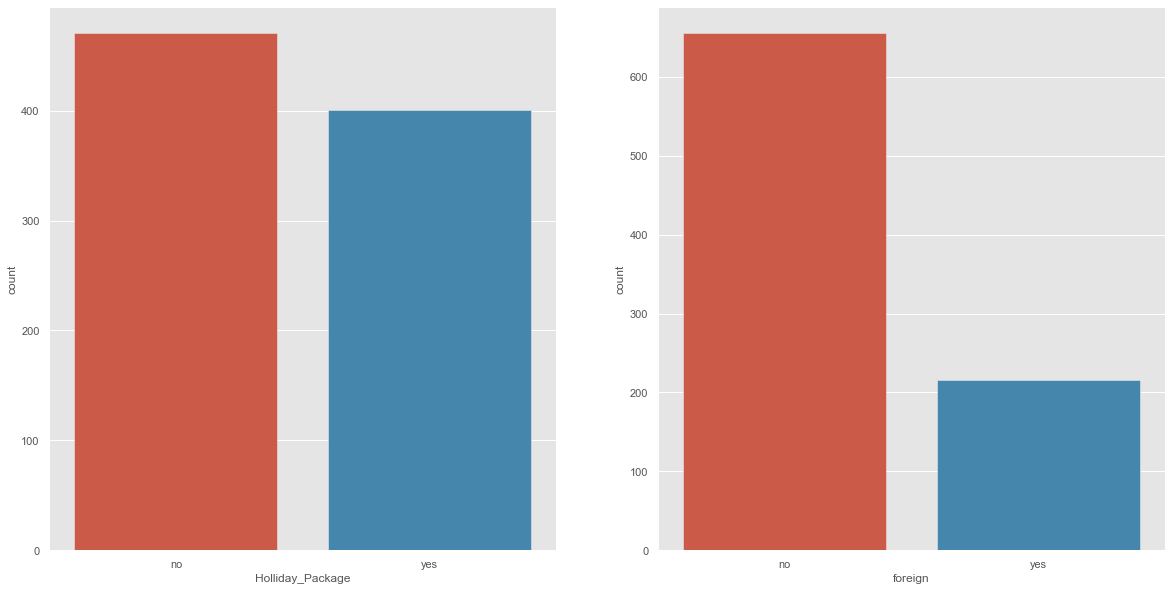
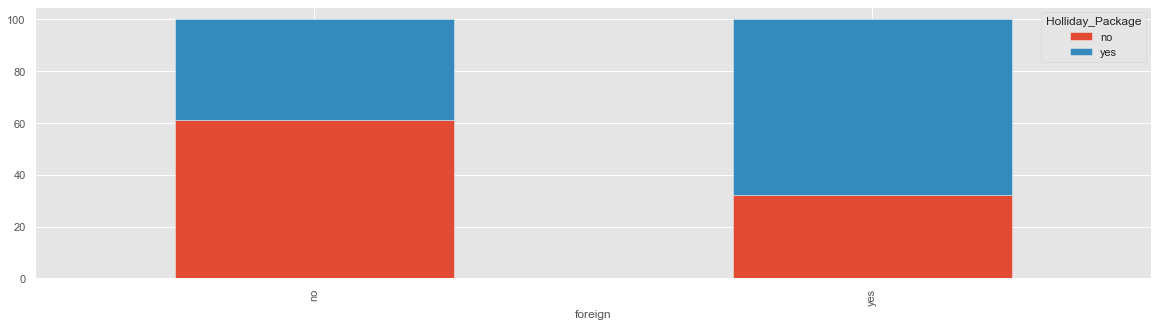


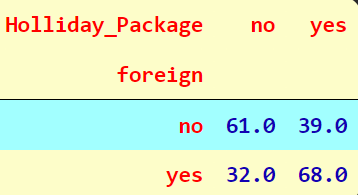
Figure 19 Stacked plot



Holliday\_Package: nearly half of the employees in the dataset have opted for the holiday package. The data is balanced in terms of dependent variable “Holliday\_Package” with 54% values as False (No) and 46% as True (Yes).

Foreign: less than 30% of the employees are foreigners. The number of foreign employees is 216. Out of which 69 (nearly 32%) choose holiday packages. From the countplot we can observe that the percentage of foreigners accepting the holiday package is substantially higher compared to the citizens.

Table 11 CrossTab



Holiday Package acceptance vs Foreigner: From the stacked countplot & Crosstab above we can observe that the percentage of foreigners accepting the holiday package is substantially higher compared to the citizens.

### Skewness:

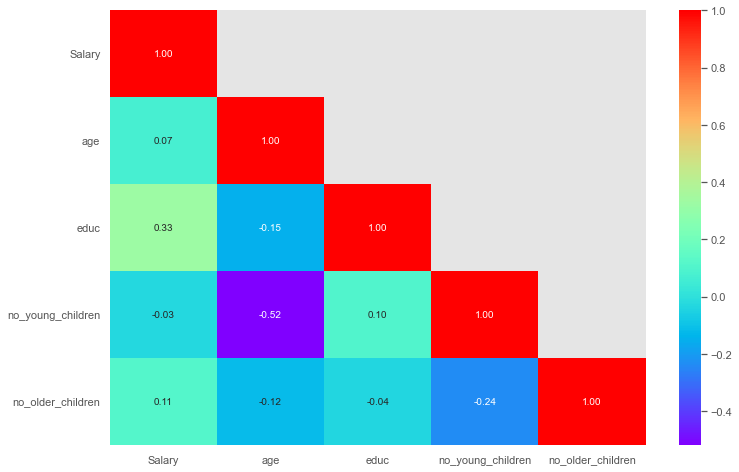
Table 12 Skewness Of dataset

|  |
| --- |
| Salary 3.10 |
| age 0.15 |
| educ -0.05 |
| no\_young\_children 1.95 |
| no\_older\_children 0.95 |

### Multivariate Analysis:

Correlation Heatmap: The heat map shows the degree of correlation between features. Observing the heatmap one can see that the highest positive correlation is among number of years of formal education and the salary received. The highest degree of negative correlation is among age and no of young children below age 7.

Figure 20 Heatmap



Pairplot: The Pairplot helps us to visualize how the features numerical in nature interact with each other.

Figure 21 Pairplot



#### Outlier:

All the numeric variables except “age” have outliers as seen in boxplot below:

Figure 22 Boxplot Before Outlier Treatment

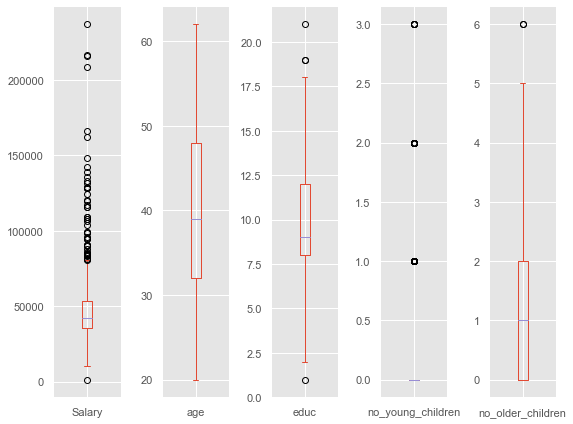
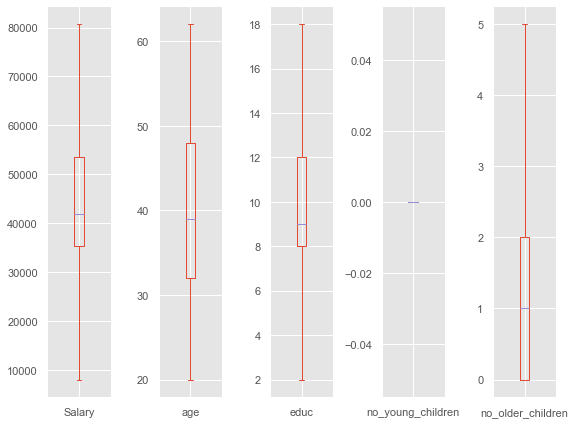


Figure 23 Boxplot After outlier Treatment



We remove outliers from the data by replacing the upper value outliers with the upper whisker value and the lower value outliers with the lower whisker values. The boxplot after removing outliers from the data.

## 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).

The features of holliday\_package and foreign have been encoded to values of 0 and 1, using one hot encoding. We encode the dependent variable Holiday Package where the choice is “No” as 0 and 1 where the choice is “Yes”. Similarly, the independent variable “foreign” we encode where the choice is “No” as 0 and 1 where the choice is “Yes”.

The data has been split into 70:30 ratio, with 70% of it being allocate for training the linear regression model and 30% of it is allocated for testing the model.

We run a Grid Search CV to identify the best model, which emerges to be Ridge Regression with liblinear solver.

## 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.

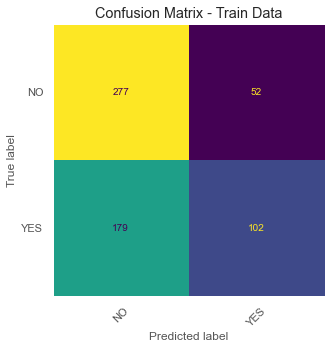
### Logistic Regression:

The model performance score for the training set is 0.62, and that for the testing set is 0.66. The classification report obtained for the training set is as follows:

Table 13 Classification Report for Training Dataset

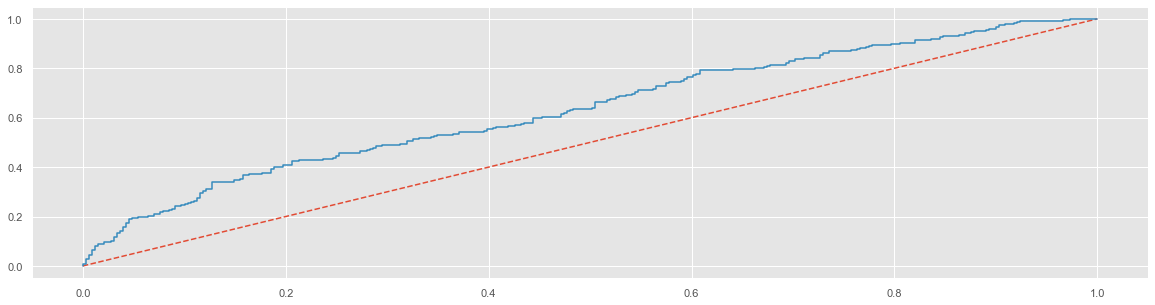
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.61 | 0.84 | 0.71 | 329 |
| 1 | 0.66 | 0.36 | 0.47 | 281 |
|  |  |  |  |  |
| Accuracy | 0.62 | | | 610 |
| Macro Avg. | 0.63 | 0.60 | 0.59 | 610 |
| Weighted Avg. | 0.63 | 0.62 | 0.60 | 610 |

Figure 24 Confusion Matrix Training Dataset



The AUC score for the training model is 0.636.

Figure 25 AUC curve for Training Dataset

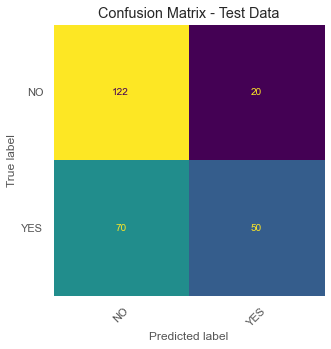


The classification report for the testing dataset is as follows:

Table 14 Classification Report for Testing Dataset

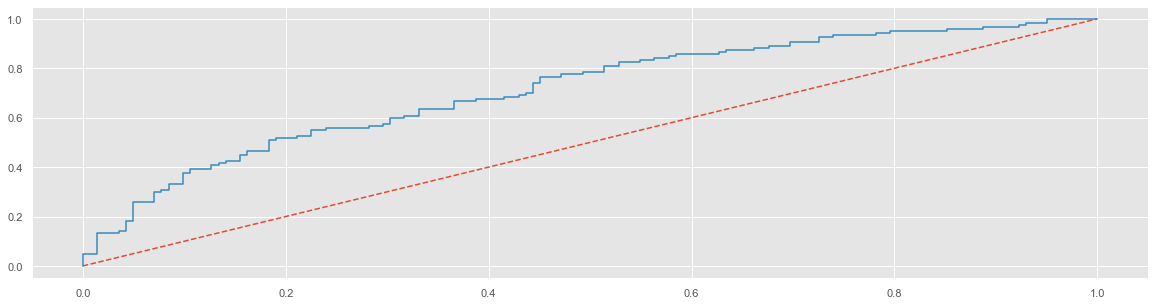
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.64 | 0.86 | 0.73 | 142 |
| 1 | 0.71 | 0.42 | 0.53 | 120 |
|  |  |  |  |  |
| Accuracy | 0.66 | | | 262 |
| Macro Avg. | 0.67 | 0.64 | 0.63 | 262 |
| Weighted Avg. | 0.67 | 0.66 | 0.64 | 262 |

Figure 26 Confusion Matrix for Testing Dataset



The AUC score for the testing model is 0.714.

Figure 27 AUC curve for Testing Dataset



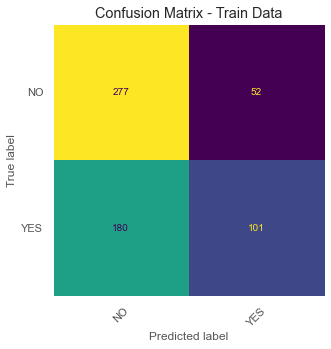
### Linear Discriminant Analysis:

The model performance score for the training set is 0.62, and that for the testing dataset is 0.65. The classification report for the training set is as follows:

Table 15 Classification Report for Training Data (LDA)

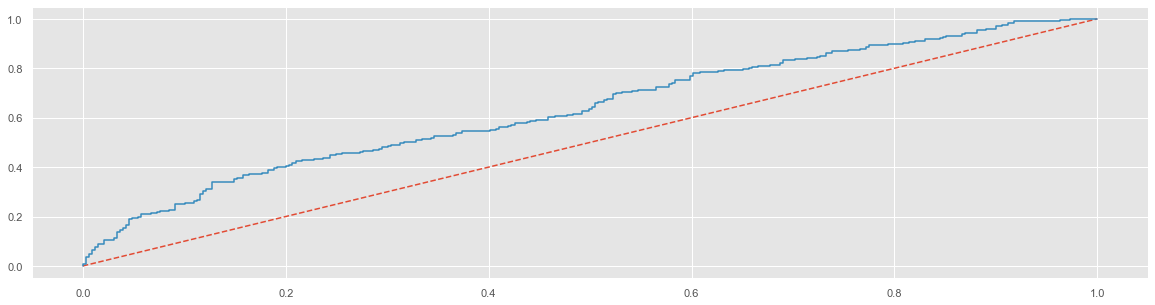
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.61 | 0.84 | 0.70 | 329 |
| 1 | 0.66 | 0.36 | 0.47 | 281 |
|  |  |  |  |  |
| Accuracy | 0.62 | | | 610 |
| Macro Avg. | 0.63 | 0.60 | 0.59 | 610 |
| Weighted Avg. | 0.63 | 0.62 | 0.59 | 610 |

Figure 28 Confusion Matrix for Training Dataset (LDA)



The AUC score for the training model is 0.636

Figure 29 AUC curve for Training Dataset (LDA)

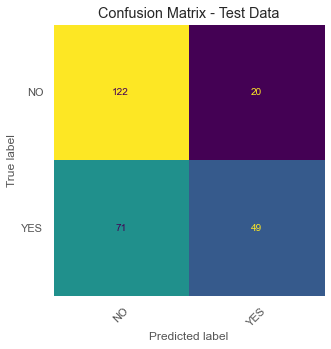


The classification report for the testing set is as follows:

Table 16 Classification Report for Testing Dataset (LDA)

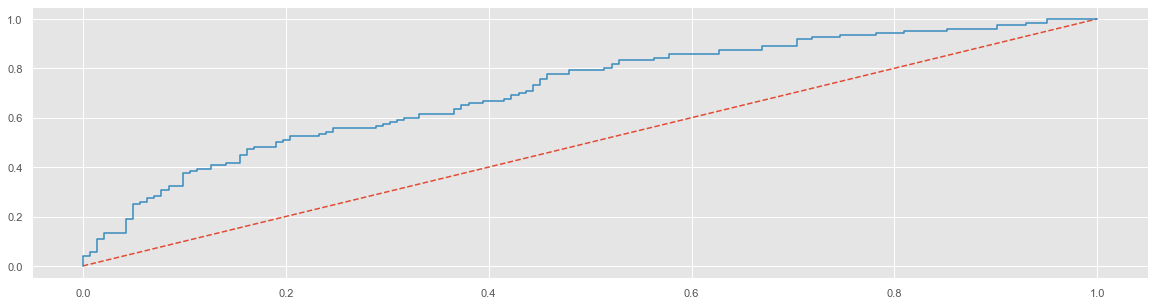
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1-Score | Support |
| 0 | 0.63 | 0.86 | 0.73 | 142 |
| 1 | 0.71 | 0.41 | 0.52 | 120 |
|  |  |  |  |  |
| Accuracy | 0.65 | | | 262 |
| Macro Avg. | 0.67 | 0.63 | 0.62 | 262 |
| Weighted Avg. | 0.67 | 0.65 | 0.63 | 262 |

Table 17 Confusion Matrix Testing Dataset (LDA)



The AUC score for the testing model is 0.712

Figure 30 AUC curve for Test Dataset (LDA)



### Comparison between Logit & LDA models:

Consider the following table comparing the models; as the Accuracy, Recall, Precision & F1 is almost equal for both the models, we can choose either one of them.

Table 18 Comparison between Logit & LDA models

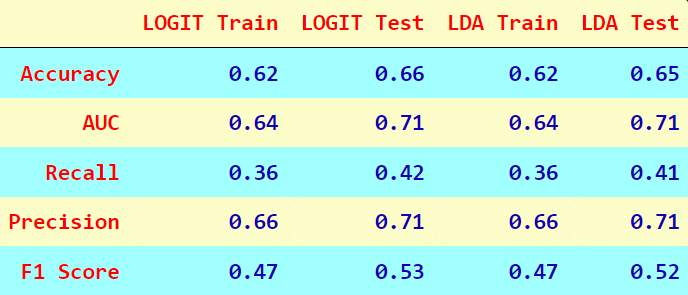


Figure 31 AUC Curve comparison for Training Data

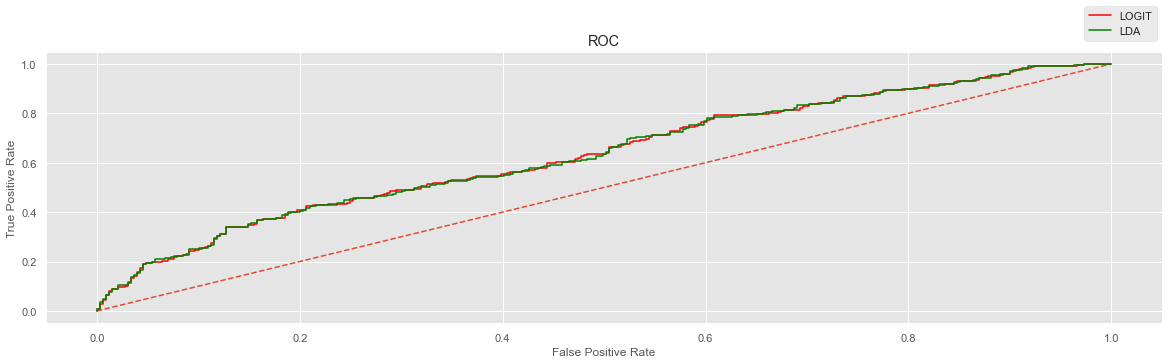
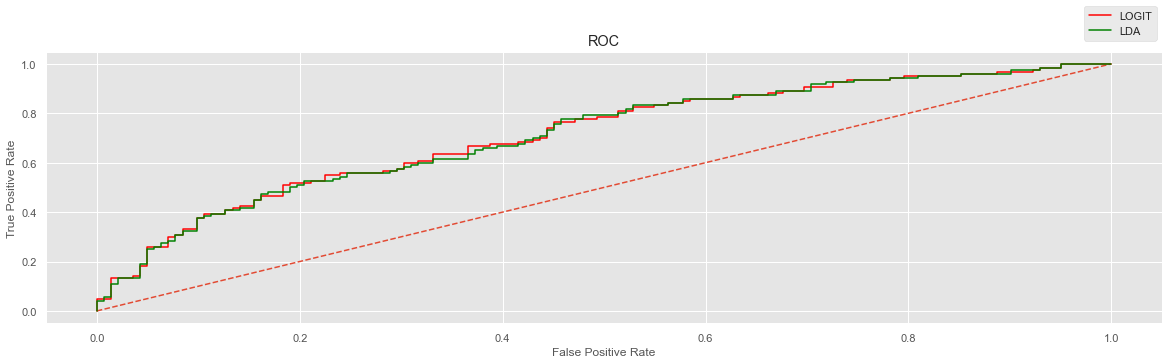


Figure 32 AUC curve Comparison for Test Data



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

In this problem we need to predict that whether an employee would opt holiday package or not, for this problem we had done predictions using both logistic regression (LR) and linear discriminant analysis (LDA). But we found that both results are same.

The EDA analysis clearly indicates certain criteria where we could find people 54% are not interested much in holiday packages and 46% are interested in a holiday packages.

• People ranging from the age 30 to 50 generally opt for holiday packages.

• Employee age over 50 to 60 have seems to be not taking the holiday package.

• Age 30 to 50 and salary less than 50000 people have opted more for holiday package.

The important factors deciding the predictions are salary, age and educ. People of a lower age even from a lower salary bracket choose holiday more than people from higher age group.

Perhaps older people are not finding holidays that appeal to their interest, the company should design packages for an older employee, as older people often have higher salaries and more disposable income.

### Conclusion:

The features that most affect the acceptance of the holiday package by the employees are number of formal years of education, number of young children below the age of 7, the age of employees and whether the employee is a native or a foreigner.

The business should implement the following strategies:

1. The business should not target any employees who have childern below the age of 7
2. The business should target foreigners who have formal education in the range of 5 to 9 years and native in the range of 8 to 12 years, as these are most likely to opt for the packages.
3. The business should target natives below the age of 45 and above the age of 35 and foreigners in the age range of 31 to 43.
4. If we have to improve holiday packages over the age, for above 50, we can add some religious destination packages.