PROJECT REPORT

Predictive Modeling

PREEJA RAJESH

PGP – DSBA

Contents

1 Linear Regression

Problem Statement 1:

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

Dataset for Problem 1: cubic_zirconia.csv

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.**

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

Data set:

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Exploratory Data Analysis:

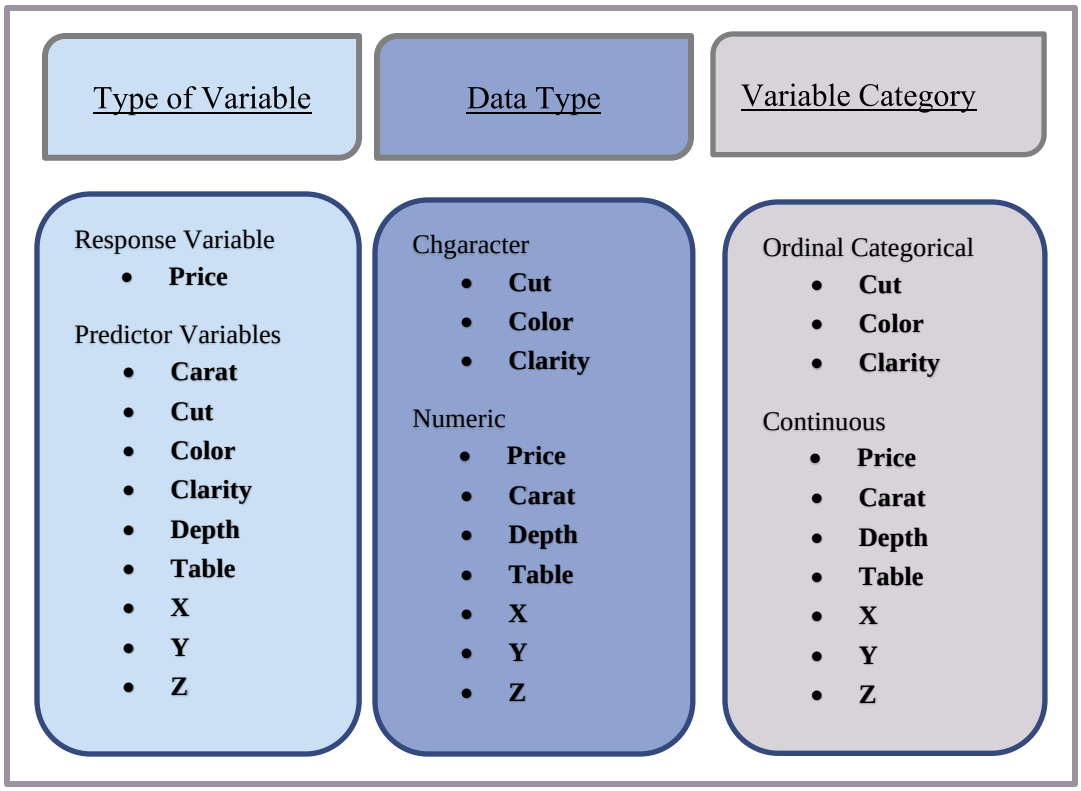
Dropped the 'Unnamed: 0' column as it is useless for the model

- **There are total 26967 rows and 10 columns in the dataset**
- **Data types of each attribute/variables are as follows:**

```
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):

 carat      26967 non-null float64
 cut        26967 non-null object
 color      26967 non-null object
 clarity    26967 non-null object
 depth      26270 non-null float64
 table      26967 non-null float64
 x          26967 non-null float64
 y          26967 non-null float64
 z          26967 non-null float64
 price      26967 non-null int64

dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```



- There are 697 null values present in the dataset. Hence we imputed the missing values wit h mean values.
- Number of duplicate rows = 34
  As it is of no use we have deleted all the 34 duplicates
  - Shape before deleting (26967, 10)
  - Shape after deleting (26933, 10)
- Getting unique counts of all Ordinal Variables

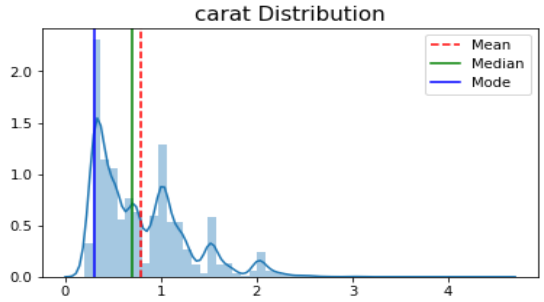
CUT : 5		COLOR : 7		CLARITY : 8	
FAIR	780	J	1440	I1	364
GOOD	2435	I	2765	IF	891
VERY GOOD	6027	D	3341	VVS1	1839
PREMIUM	6886	H	4095	VVS2	2530
IDEAL	10805	F	4723	VS1	4087
		E	4916	SI2	4564
		G	5653	VS2	6093
				SI1	6565

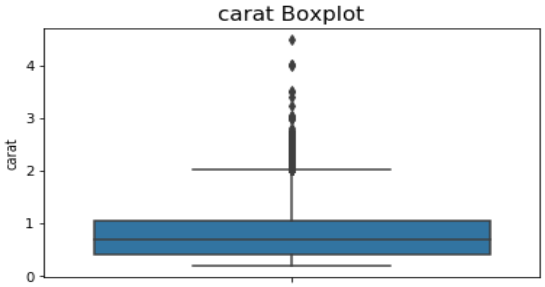
Observation:

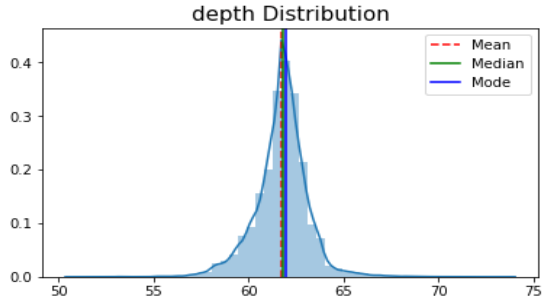
- There are no? or other character present.
- All nominal values have 5 to 8 categories which can be included in dataset for prediction.

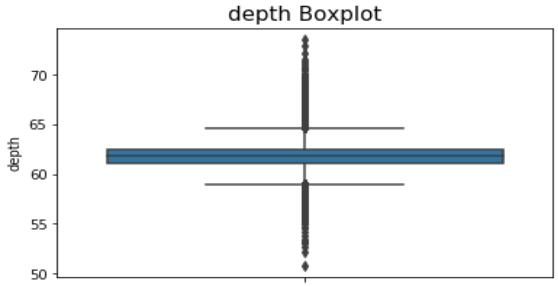
Univariate Analysis:

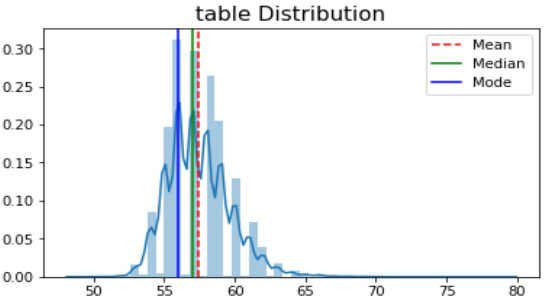
- Univariate analysis refers to the analysis of a single variable. The main purpose of univariate analysis is to summarize and find patterns in the data. The key point is that there is only one variable involved in the analysis.

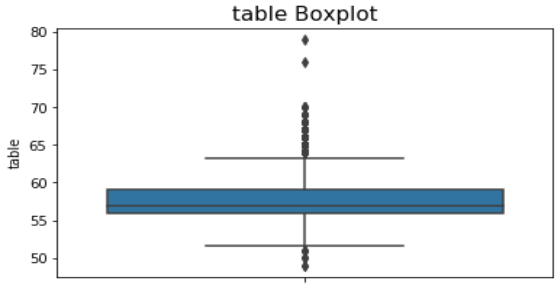


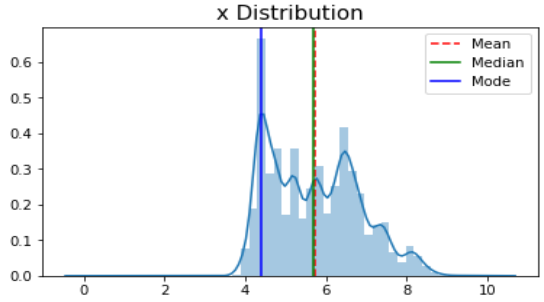


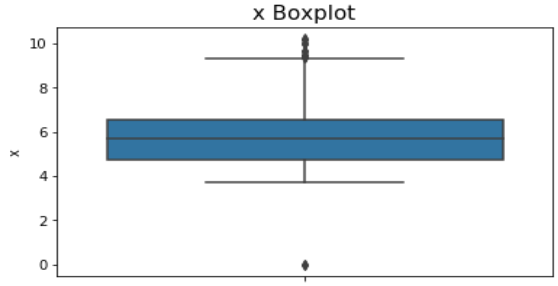


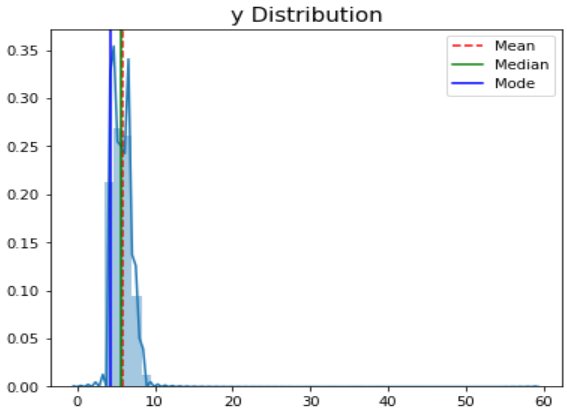


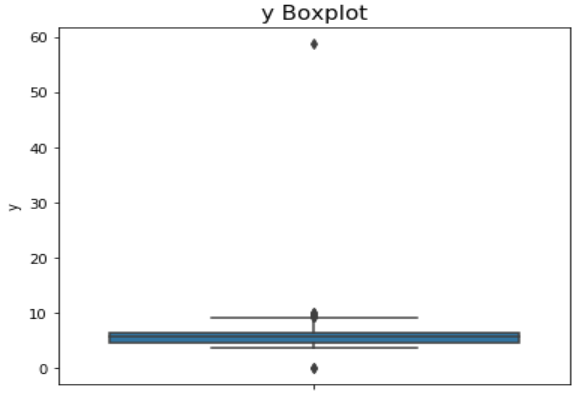


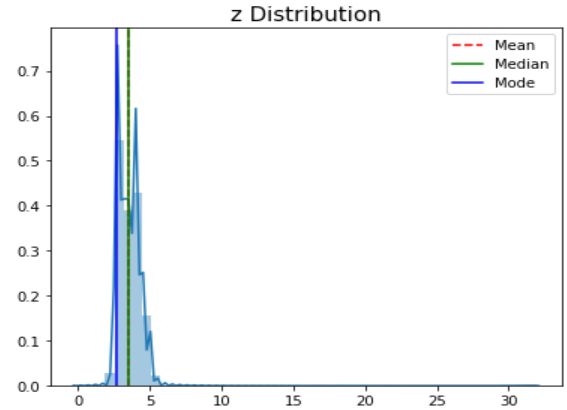


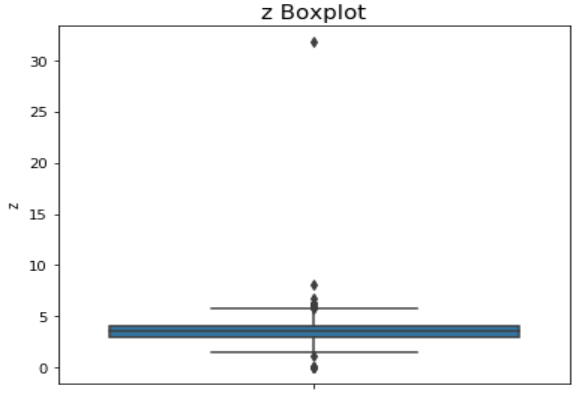


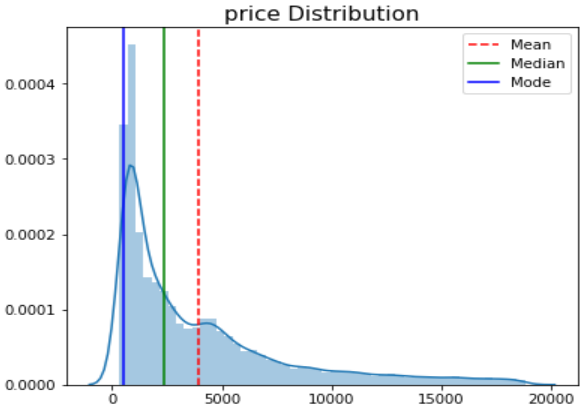


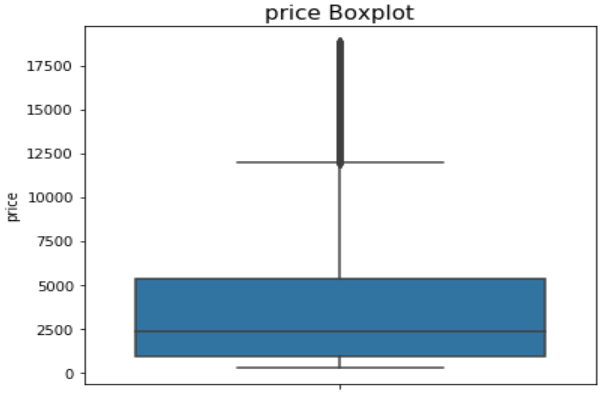












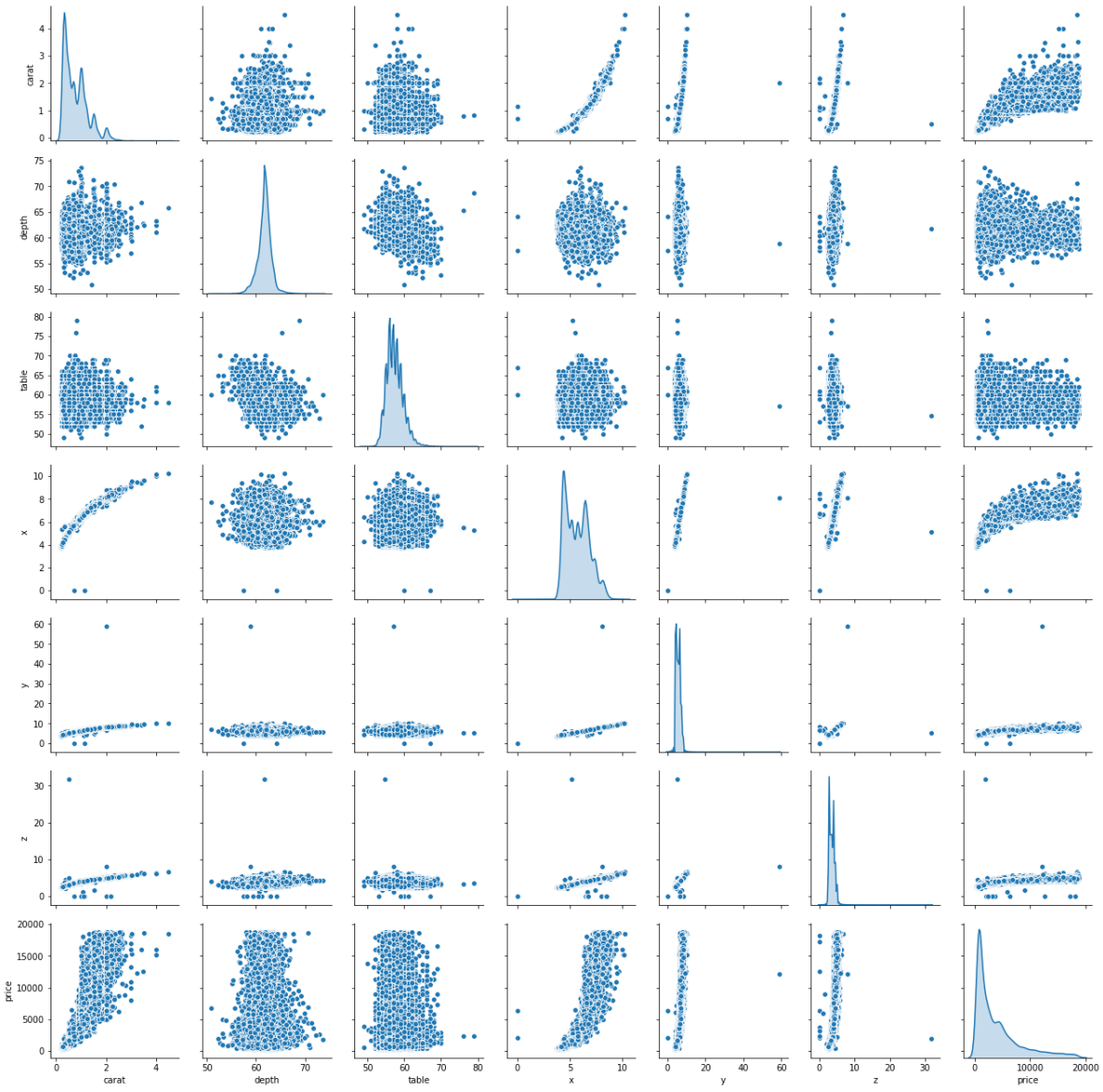
Skewness:

- carat 1.114789
- depth -0.026422
- table 0.765805
- x 0.392290
- y 3.867764
- z 2.580665
- price 1.619116

Inferences:

- **The skewness value of 3.8 shows that the variable 'y' has a right-skewed distribution, indicating the presence of extreme higher values. The maximum 'y' value of 58.9 proves this point.**
- **Variables 'depth', 'table' and 'x' seems to be normally distributed.**

Bivariate Analysis:



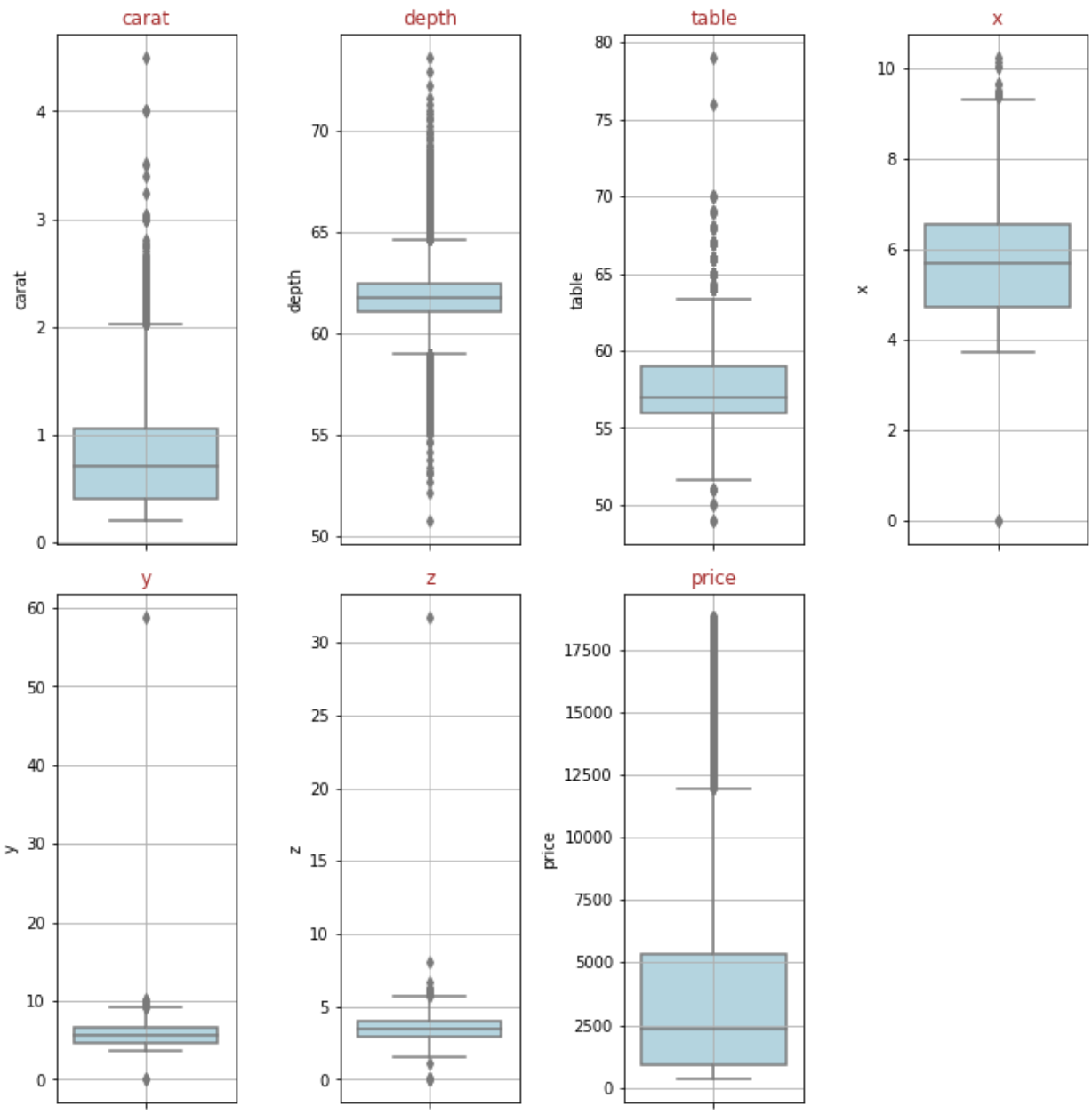
The pair plot only offers visual information about the degree of correlation. In order to obtain more precise information, we can use the inbuilt.corr() method in Pandas. This returns a table with all the correlations calculated for the numerical columns.

	carat	depth	table	x	y	z	price
carat	1.000000	0.034741	0.181539	0.976858	0.941442	0.940982	0.922409
depth	0.034741	1.000000	-0.293720	-0.018145	-0.024139	0.097659	-0.002855
table	0.181539	-0.293720	1.000000	0.196254	0.182352	0.148994	0.126844
x	0.976858	-0.018145	0.196254	1.000000	0.962601	0.956490	0.886554
y	0.941442	-0.024139	0.182352	0.962601	1.000000	0.928725	0.856441
z	0.940982	0.097659	0.148994	0.956490	0.928725	1.000000	0.850682
price	0.922409	-0.002855	0.126844	0.886554	0.856441	0.850682	1.000000

Inferences:

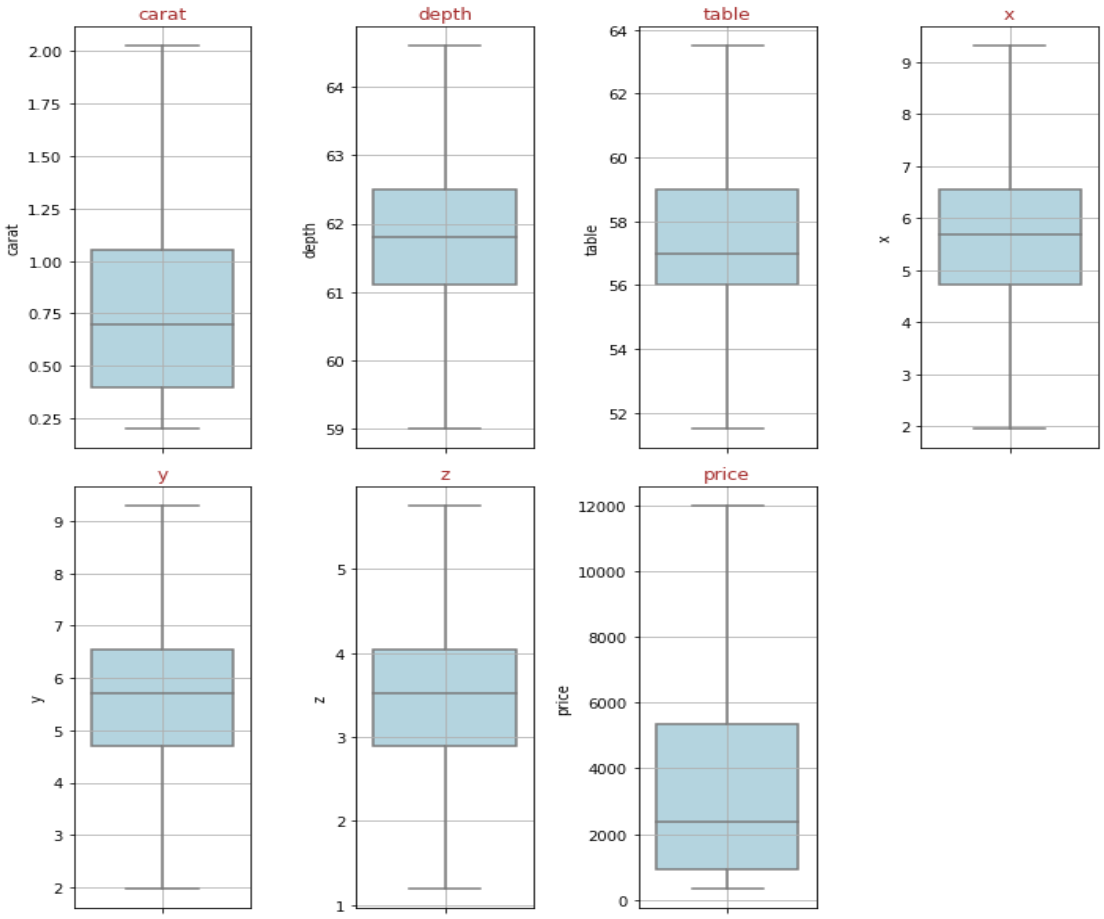
- Price' (response variable) is highly correlated to 'carat', 'x', 'y' and 'z' which are predictors. This indicates that independent variables influence the dependent variable.
- Also 'x', 'y' and 'z' are highly correlated to each other.
- Multicollinearity occurs when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.
- Multicollinearity is a problem because it undermines the statistical significance of an independent variable. Other things being equal, the larger the standard error of a regression coefficient, the less likely it is that this coefficient will be statistically significant.

Box plot for continuous variables before Outlier Removal:



All the outliers are replaced by the low and high value of IQR which is nothing but one type of outlier treatment.

Box plot for continuous variables after Outlier Removal:



Label encoding:

The cut, color and clarity variable are ordinal i.e. there is an order within the different categories hence, label encoding is preferred.

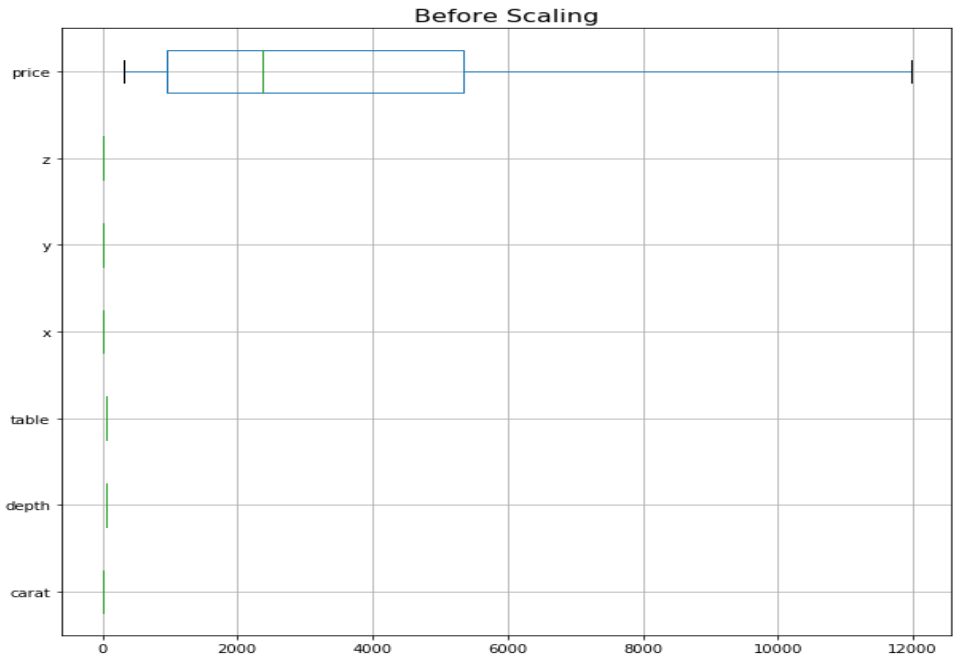
CUT: 5		COLOR: 7		CLARITY: 8	
FAIR	**0**	**J**	**0**	**I1**	**0**
GOOD	**1**	**I**	**1**	**SI2**	**1**
VERY GOOD	**2**	**H**	**2**	**SI1**	**2**
PREMIUM	**3**	**G**	**3**	**VS2**	**3**
IDEAL	**4**	**F**	**4**	**VS1**	**4**
		E	**5**	**VVS2**	**5**
		D	**6**	**VVS1**	**6**
				IF	**7**

Data after label encoding:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4	5	2	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	3	3	7	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	2	5	5	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	4	4	4	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	4	4	6	60.4	59.0	4.35	4.43	2.65	779.0

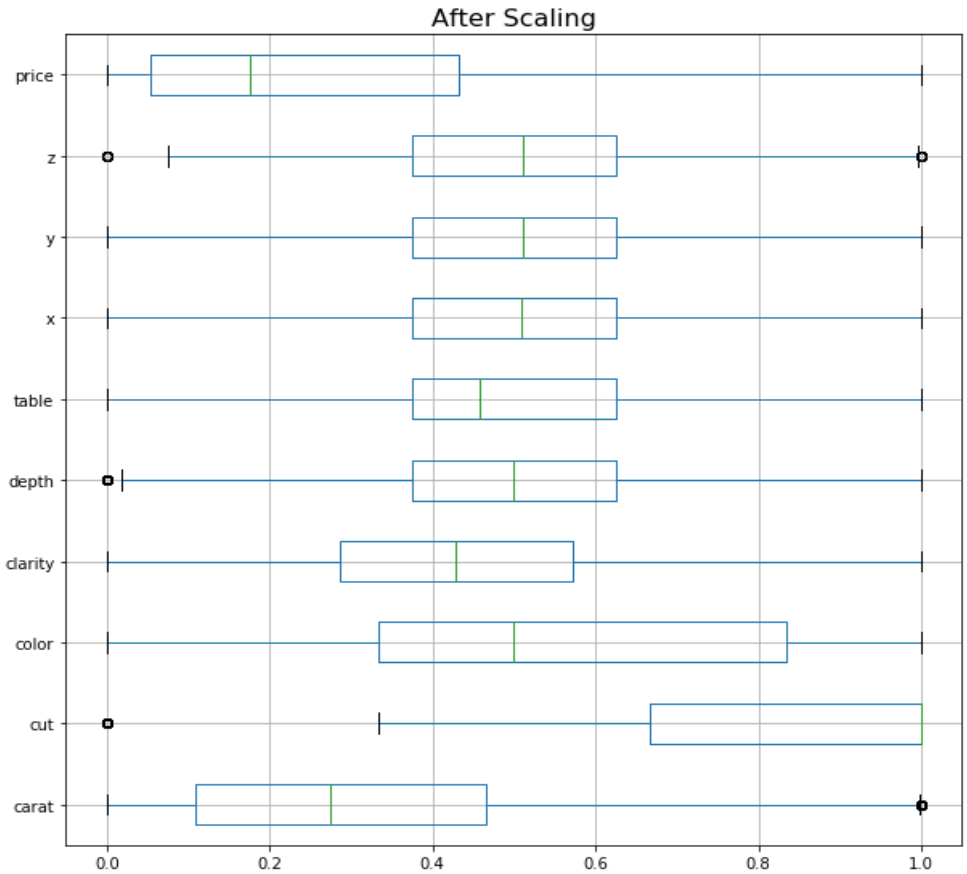
1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning, or do we need to change them or drop them? Do you think scaling is necessary in this case?

- There are 697 null values present in the dataset hence, there is a need to impute the missin g values by replacing it with the mean of the column.
- There are zero values in 'cut', 'color' and 'clarity' variable. But these are assigned while co nverting the categorical variable to numeric. Hence it has meaning and need not be delete d.
- Scaling is done only to make the intercept 0 in case where the intercept value has no mean ing. Here we have used Feature Scaling using MinMaxScaler that normalizes the data usi ng the formula (x - min)/(max - min)



Scaled dataset:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.054795	1.00	0.833333	0.285714	0.553571	0.541667	0.315217	0.317623	0.322368	0.014854
1	0.071233	0.75	0.500000	1.000000	0.321429	0.541667	0.335598	0.340847	0.331140	0.056498
2	0.383562	0.50	0.833333	0.714286	0.571429	0.708333	0.555707	0.567623	0.567982	0.511999
3	0.120548	1.00	0.666667	0.571429	0.464286	0.375000	0.389946	0.387295	0.388158	0.064912
4	0.060274	1.00	0.666667	0.857143	0.250000	0.625000	0.326087	0.336749	0.320175	0.038896



Inferences:

- From the above boxplots we can see that most of the outliers were removed after outlier treatment.
- After scaling it is much cleaner.
- If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.
- Scaling is a good practice and it improves the model.

1.3 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 R square, RMSE.

- Label encoding is done in previous step.

Train-Test Split

- X dataset:

	carat	cut	color	clarity	depth	table	x	y	z
0	0.054795	1.00	0.833333	0.285714	0.553571	0.541667	0.315217	0.317623	0.322368
1	0.071233	0.75	0.500000	1.000000	0.321429	0.541667	0.335598	0.340847	0.331140
2	0.383562	0.50	0.833333	0.714286	0.571429	0.708333	0.555707	0.567623	0.567982
3	0.120548	1.00	0.666667	0.571429	0.464286	0.375000	0.389946	0.387295	0.388158
4	0.060274	1.00	0.666667	0.857143	0.250000	0.625000	0.326087	0.336749	0.320175

	VIF Factor	features
0	33.3	carat
1	10.0	cut
2	5.4	color
3	5.0	clarity
4	13.2	depth
5	11.3	table
6	4635.3	x
7	4469.1	y
8	1279.0	z

Inference:

- VIF value for color is between 1 and 5 so, there is a moderate correlation, but it is not severe enough to warrant corrective measures.
- VIF value for rest all the variable is greater than 5 which, represents critical levels of multicollinearity where the coefficients are poorly estimated, and the p value are questionable.
- As expected, the x, y, z and the carat of the cubic zirconia have a high variance inflation factor because they "explain" the same variance within this dataset. We would need to discard few of these variables step by step before moving on to model building or risk building a model with high multicollinearity.

#Step 1: Discarding x

	VIF Factor	features
0	32.6	carat
1	9.7	cut
2	5.4	color
3	4.9	clarity
4	13.1	depth
5	10.8	table
6	1090.4	y
7	1208.6	z

#Step 2: Discarding z

	VIF Factor	features
0	30.6	carat
1	9.7	cut
2	5.4	color
3	4.9	clarity
4	5.8	depth
5	10.8	table
6	112.4	y

#Step 3: Discarding carat

	VIF Factor	features
0	7.2	cut
1	4.6	color
2	4.5	clarity
3	5.1	depth
4	8.2	table
5	11.7	y

Inference:

- Finally, all the attribute's VIF values are less or equal to 10. Hence, we can go forward with building the model using the remaining attributes

Linear Regression Model:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Coefficients for each of the independent attributes

- **The coefficient for cut is 0.04363802887493747**
- **The coefficient for color is 0.11963746843288983**
- **The coefficient for clarity is 0.2860697310114946**
- **The coefficient for depth is 0.07595309933802072**
- **The coefficient for table is 0.015747552383059277**
- **The coefficient for y is 2.0169263462517972**

The intercept for our model is -1.0143926868821223

1. **The coefficient of determination R^2 of the prediction on Train set 0.8876692651510796**

88.7% of the variation in the price is explained by the predictors in the model for train set.

2. **The coefficient of determination R^2 of the prediction on Test set 0.8880185460742231**

88.8% of the variation in the price is explained by the predictors in the model for test set

1. **The Root Mean Square Error (RMSE) of the model is for training set is 0.0997324818701998**
2. **The Root Mean Square Error (RMSE) of the model is for testing set is 0.09981734535378116**

Linear Regression using statsmodels:

- R^2 value for training and test set are equal to 88.8% which is good.
- R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable. Instead we use adjusted R^2 which removes the statistical chance that improves R^2.
- Scikit does not provide a facility for adjusted R^2... so we use statsmodel, a library that gives results similar to what you obtain in R language
- This library expects the X and Y to be given in one single dataframe

Coefficients for each of the independent attributes

- **Intercept -1.014393**
- **cut 0.043638**
- **color 0.119637**
- **clarity 0.286070**
- **depth 0.075953**
- **table 0.015748**
- **y 2.016926**

OLS Regression Results

Dep. Variable:	price	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.888
Method:	Least Squares	F-statistic:	2.482e+04
Date:	Sun, 05 Jul 2020	Prob (F-statistic):	0.00
Time:	18:19:51	Log-Likelihood:	16710.
No. Observations:	18853	AIC:	-3.341e+04
Df Residuals:	18846	BIC:	-3.335e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.0144	0.007	-146.171	0.000	-1.028	-1.001
cut	0.0436	0.003	13.607	0.000	0.037	0.050
color	0.1196	0.003	44.453	0.000	0.114	0.125
clarity	0.2861	0.003	84.345	0.000	0.279	0.293
depth	0.0760	0.004	19.678	0.000	0.068	0.084
table	0.0157	0.005	3.106	0.002	0.006	0.026
y	2.0169	0.005	371.177	0.000	2.006	2.028

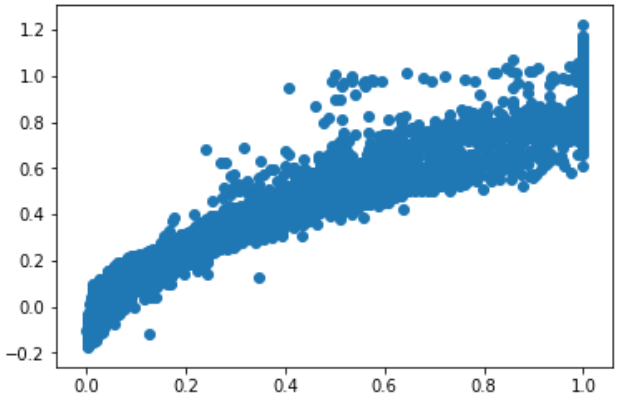
```
Omnibus:                      1874.189   Durbin-Watson:                   1.978
Prob(Omnibus):                   0.000   Jarque-Bera (JB):             3166.341
Skew:                            0.708   Prob(JB):                         0.00
Kurtosis:                        4.422   Cond. No.                         20.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is cor
rectly specified
```

Inference:

- The overall P value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression co-efficient is not 0. Here all regression co-efficient are not 0.

Prediction on Test data:



Inference:

- A good model's prediction will be close to actual leading to high R and R2 values

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

Inference:

- In other words, the p-value for a variable is less than the significance level, then the sample data provide enough evidence to reject the null hypothesis for the entire population. Hence the data favor the hypothesis that there is a non-zero correlation. Changes in the independent variable are associated with changes in the response at the population level. This variable is statistically significant and probably a worthwhile addition to the regression model.

The final Linear Regression equation is

- **price = b0 + b1 * cut + b2 * color + b3 * clarity + b4 * depth + b5 * table + b6 * y**
- **price = (-1.01) * Intercept + (0.04) * cut + (0.12) * color + (0.29) * clarity + (0.08) * depth + (0.02) * table + (2.02) * y**
- **All coefficients are positive.**
- **When 'y' increases by 1 unit, 'price' increases by 2.02 units, keeping all other predictors constant.**
- **Similarly, when 'depth' increases by 1 unit, 'price' increases by 0.08 units, keeping all other predictors constant.**

Recommendations:

- Higher profitable stones can be identified by predicting the price for the stone on the bases the 5 important attributes.

- The best 5 attributes that are most important are:
    - y (Width)
    - clarity
    - color
    - depth
    - cut
- All five attributes (y, clarity, color, depth and cut) will contribute positively to the price.
- If the width, clarity, color, depth and the cut style of the cubic zirconia is high then the price is high.
- If only the dependent/response variable is log-transformed.
- Exponentiating the coefficient, and subtracting one from this number, and multiplying by 100. This gives the percent increase (or decrease) in the response for every one-unit increase in the independent variable.

    Example:
    The coefficient is 0.29
    (exp(0.29) – 1) * 100 = 33.6 ~ 34%.
- For every one-unit increase in the independent variable (color), our dependent variable (price) increases by about 34%.

    The coefficient is 0.12
    (exp(0.12) – 1) * 100 = 12.7 ~ 13%.

For every one-unit increase in the independent variable (z), our dependent variable (price) decreases by about 13%.

2 Logistic Regression and LDA

Problem Statement 2:

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

Dataset for Problem 2: **Holiday_Package.csv**

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.

Data set:

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

Exploratory Data Analysis:

- **There are total 872 rows and 8 columns in the dataset**
- **Data types of each attribute/variables are as follows:**

RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):

- Unnamed: 0 872 non-null int64
- Holiday_Package 872 non-null object
- Salary 872 non-null int64
- age 872 non-null int64
- educ 872 non-null int64
- no_young_children 872 non-null int64
- no_older_children 872 non-null int64
- foreign 872 non-null object

dtypes: int 64(6), object(2)
memory usage: 54.6+ KB

- There are 0 null values present in the dataset
- Dropped the 'Unnamed: 0' column as it is useless for the model
- Number of duplicate rows = 0

	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

Getting unique counts of all Ordinal Variables

```
HOLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holiday_Package, dtype: int64

FOREIGN :  2
yes    216
no     656
Name: foreign, dtype: int64
```

Observation:

- There are no ? or other charecter present.
- All nominal values have 2 categories which can be included in dataset for prediction.

Label encoding:

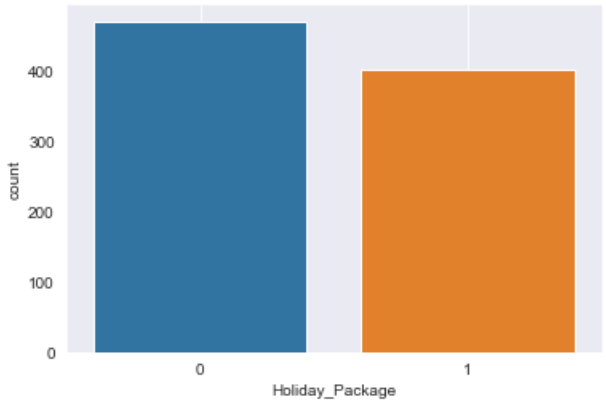
Convert Object Feature types for Linear Discriminant Analysis

```
feature: Holiday_Package
[no, yes]
Categories (2, object): [no, yes]
[0 1]


feature: foreign
[no, yes]
Categories (2, object): [no, yes]
[0 1]
```

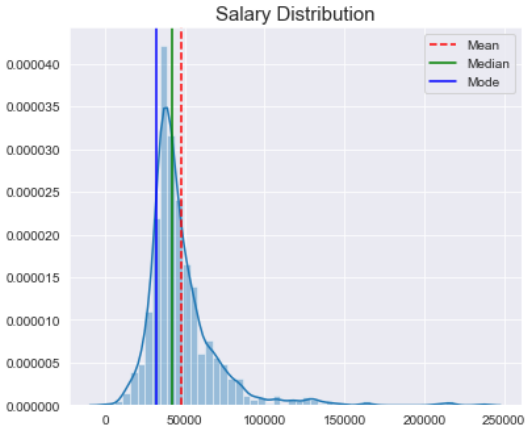
Proportion in the Target classes:

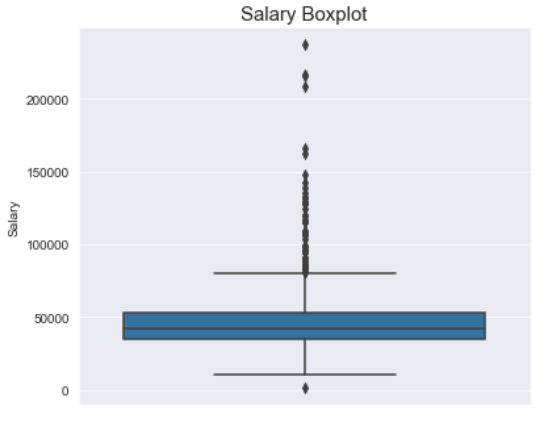
```
0    471
1    401
Name: Holiday_Package, dtype: int64
```

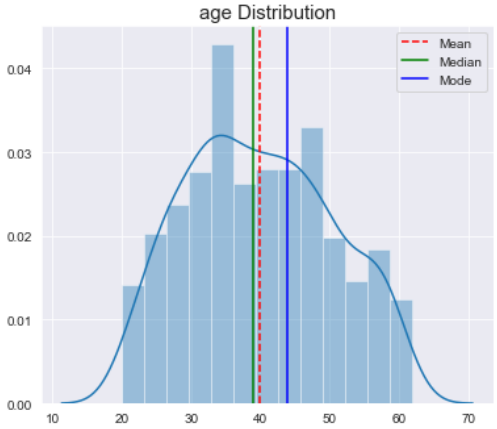


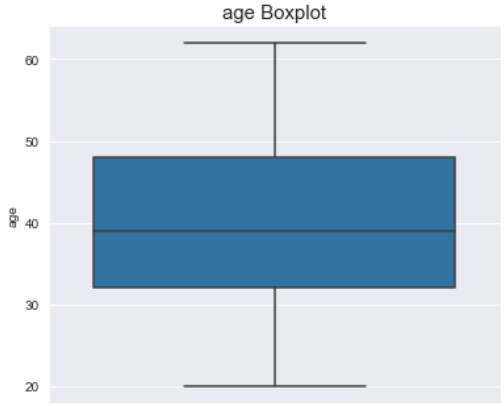
- **The percentage of zeroes in the Holiday_Package variable is 54.01376146788991**
- **and the percentage of ones in the Holiday_Package variable is 45.98623853211009**

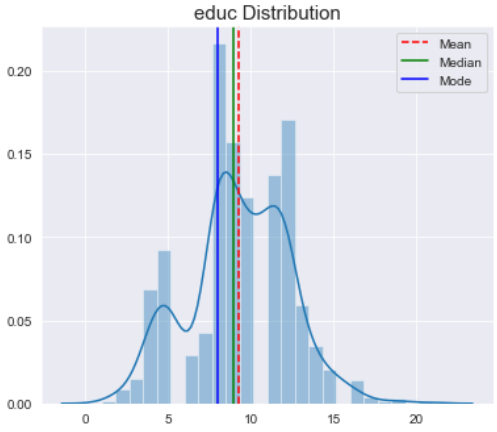
Univariate Analysis:

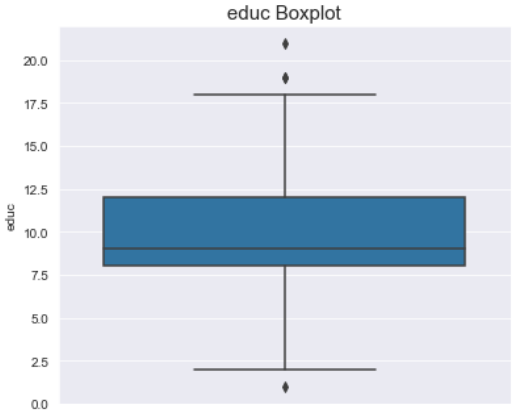












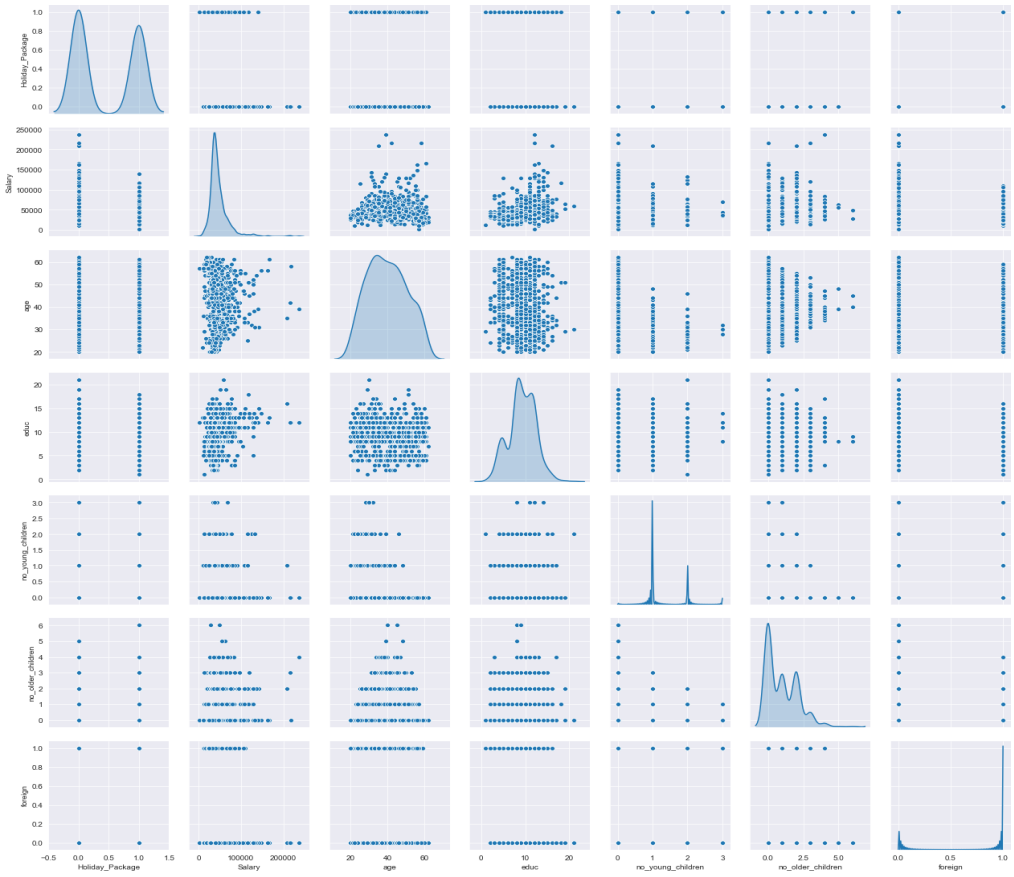
Skewness:

```
Holiday_Package      0.161348
Salary               3.103216
age                  0.146412
educ                -0.045501
no_young_children    1.946515
no_older_children    0.953951
foreign              1.170906
```

Inferences:

- The skewness value of 3.1 shows that the variable 'Salary' has a right-skewed distribution, indicating the presence of extreme higher values. The maximum 'Salary' value of 236961 proves this point.
- Variables 'age'and 'educ'seems to be normally distributed.
- Salary has too many outliers and educ has fe outliers.
- Age has no outliers atall.

Bivariate Analysis:

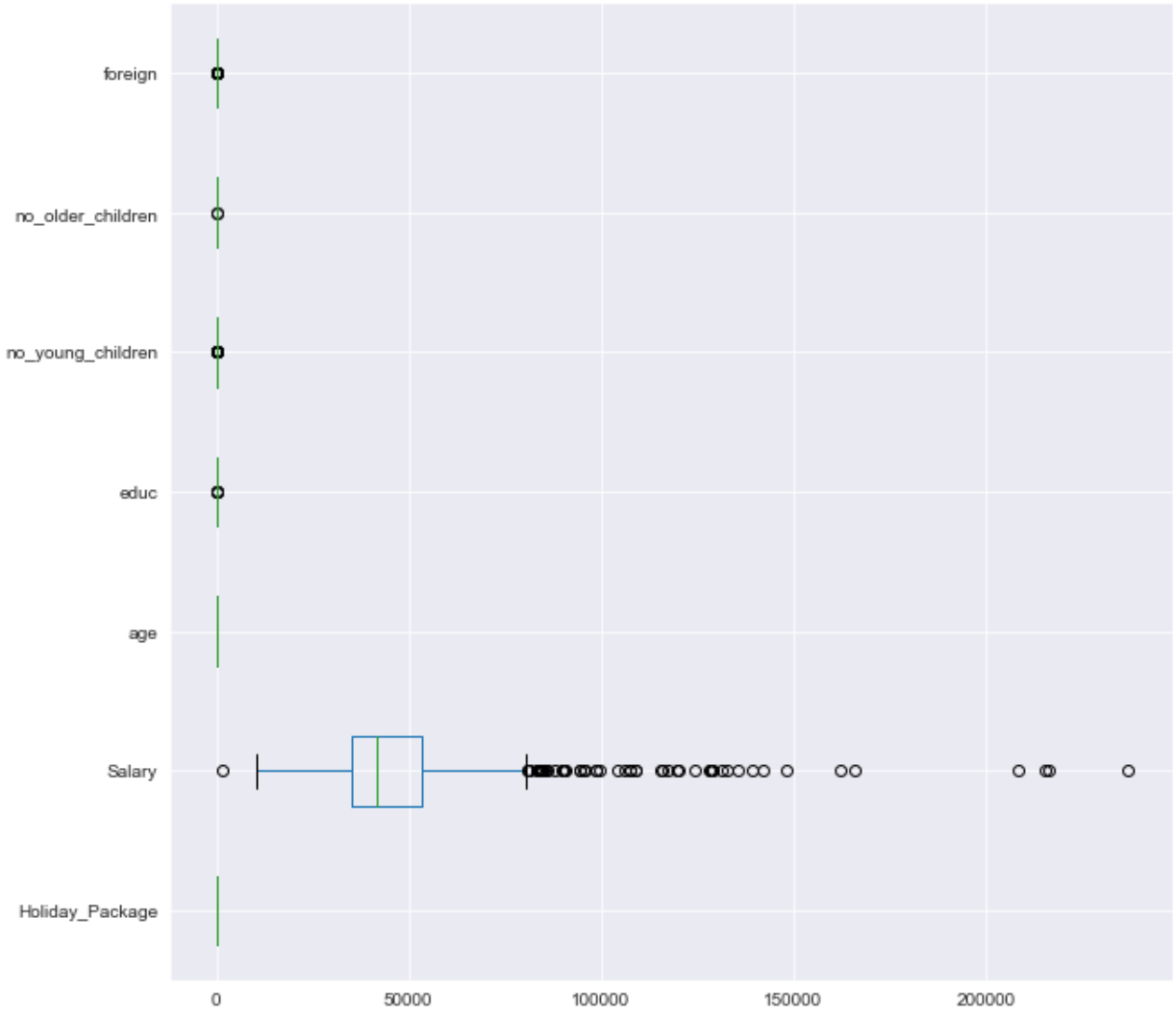


	Holiday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
Holiday_Package	1.000000	-0.185694	-0.092311	-0.102552	-0.173115	0.080286	0.254096
Salary	-0.185694	1.000000	0.071709	0.326540	-0.029664	0.113772	-0.201043
age	-0.092311	0.071709	1.000000	-0.149294	-0.519093	-0.116205	-0.107148
educ	-0.102552	0.326540	-0.149294	1.000000	0.098350	-0.036321	-0.419678
no_young_children	-0.173115	-0.029664	-0.519093	0.098350	1.000000	-0.238428	0.085111
no_older_children	0.080286	0.113772	-0.116205	-0.036321	-0.238428	1.000000	0.021317
foreign	0.254096	-0.201043	-0.107148	-0.419678	0.085111	0.021317	1.000000

5 point summary:

	count	mean	std	min	25%	50%	75%	max
Holiday_Package	872.0	0.459862	0.498672	0.0	0.0	0.0	1.0	1.0
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872.0	0.247706	0.431928	0.0	0.0	0.0	0.0	1.0

Outlier Checks:



Although outliers exists as per the boxplot, by looking at the data distribution in describe(), the values are not too far away. Treating the outliers by converting them to min/max values will cause most variables to have values to be the same. So, outliers are not treated in this case

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).

Code:

```
#capture the dependent column into separate vectors for training set and test set
copy_df_hp = df_hp.copy(deep=True)

X = copy_df_hp.drop("Holiday_Package" , axis=1)
y = copy_df_hp.pop("Holiday_Package")

X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=.30, random_state=1)
```

- Now since the data has been split in the 70:30 (train:test) ratio, let us check the distribution of Ones and Zeroes in the Y ('Holiday_Package') variable.
- Ones and Zeroes in the training and test set is the same as the proportion of Ones and Zeroes that were present in the whole dataset.

Linear Discriminant Analysis (LDA):

```
lda = LinearDiscriminantAnalysis()
model=lda.fit(X_train,Y_train)
ypred_train = model.predict(X_train)
ypred_test = model.predict(X_test)
```

Logistic Regression Model:

Logit Regression Results			
Dep. Variable:	**Holiday_Package**	No. Observations:	**610**
Model:	**Logit**	Df Residuals:	**604**
Method:	**MLE**	Df Model:	**5**
Date:	**Sun, 05 Jul 2020**	Pseudo R-squ.:	**0.1465**
Time:	**22:48:22**	Log-Likelihood:	**-359.62**
converged:	**True**	LL-Null:	**-421.37**
Covariance Type:	**nonrobust**	LLR p-value:	**5.687e-25**

	coef	std err	z	P>|z|	[0.025	0.975]

Intercept	**2.0997**	**0.638**	**3.290**	**0.001**	**0.849**	**3.351**
Salary	**-1.804e-05**	**5.37e-06**	**-3.361**	**0.001**	**-2.86e-05**	**-7.52e-06**
age	**-0.0522**	**0.010**	**-5.015**	**0.000**	**-0.073**	**-0.032**
educ	**0.0805**	**0.036**	**2.220**	**0.026**	**0.009**	**0.152**
no_young_children	**-1.4973**	**0.212**	**-7.068**	**0.000**	**-1.913**	**-1.082**
foreign	**1.5899**	**0.258**	**6.166**	**0.000**	**1.085**	**2.095**

Inference:

- The coefficient table showed that all attributes except one i.e. (no_older_children) has significant influence (p-values < 0.05) on Holiday_Package. The coefficients are in log-odds terms. The interpretation of the model coefficients could be as follows:
- Each one-unit change in foreign will increase the log odds of opting Holiday_Package by 1.5879, and its p-value indicates that it is significant in determining the opting of Holiday_Package. Similarly, with each unit increase in educ increases the log odds of opting Holiday_Package by 0.0781 and p-value is significant too.
- The interpretation of coefficients in the log-odds term does not make much sense if it needs to report it in any article or publication. That is why the concept of odds ratio was introduced.

ODDs Ratio

The ODDS is the ratio of the probability of an event occurring to the event not occurring. When we take a ratio of two such odds it called Odds Ratio**.**

- **Intercept** 8.164017
- **Salary** 0.999982
- **age** 0.949094
- **educ** 1.083855
- **no_young_children** 0.223725
- **foreign** 4.903171
- **dtype: float64**

Inference:

- In the above ODDS ratio table, you can observe that foreign has an ODDS Ratio of 4.8933, which indicates that one unit increase in no. of foreigners increases the odds of opting Holiday_Package by 4.8933 times.
- Even though the interpretation of ODDS ratio is far better than log-odds interpretation, still it is not as intuitive as linear regression coefficients; where one can directly interpret that how much a dependent variable will change if making one unit change in the independent variable, keeping all other variables constant.

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.

Linear Discriminant Analysis (LDA):

```
Accuracy Score for Train set is  0.6721311475409836
Accuracy Score for Test set is  0.6412213740458015
```

Accuracy score is the percentage of accuracy of the predictions made by the model. For our LDA model the accuracy score is 0.64, which is considerably quite accurate. But the more the accuracy score the efficient is you prediction model.

Model evaluation on test data set

In [218]:

```
Confusion Matrix
 [[103  42]
 [ 52  65]]

Classification Report
              precision    recall  f1-score   support

           0       0.66      0.71      0.69       145
           1       0.61      0.56      0.58       117

    accuracy                           0.64       262
   macro avg       0.64      0.63      0.63       262
weighted avg       0.64      0.64      0.64       262
```

Observations:

- True positive: 65 (We predicted a positive result and it was positive)
- True negative: 103 (We predicted a negative result and it was negative)
- False positive: 42 (We predicted a positive result and it was negative)
- False negative: 52 (We predicted a negative result and it was positive)

In the output, 103 and 65 are actual predictions, and 42 and 52 are incorrect predictions.

Recall: This Linear Discriminant Analysis model can identify the employees who will opt for the Holiday package 56% of the time.

Logistic Regression Model:

The target variable has only two possible outcomes such as Yes or No.

Model evaluation on test data set

Predicted	0	1	All
Actual			
0	**101**	**44**	**145**
1	**50**	**67**	**117**
All	**151**	**111**	**262**

Observations:

- True positive: 67 (We predicted a positive result and it was positive)
- True negative: 101 (We predicted a negative result and it was negative)
- False positive: 44 (We predicted a positive result and it was negative)
- False negative: 50 (We predicted a negative result and it was positive)

In the output, 101 and 67 are actual predictions, and 44 and 50 are incorrect predictions.

Recall: This Logistic Regression model can identify the employees who will opt for the Holiday package 57% of the time.

Classification accuracy

Accuracy: 0.64%

Accuracy score is the percentage of accuracy of the predictions made by the model. For our Bina ry Logistic Regression model the accuracy score is 0.64, which is considerably quite accurate. Bu t the more the accuracy score the efficient is the prediction model.

Classification report

	precision	recall	f1-score	support
0	0.67	0.70	0.68	145
1	0.60	0.57	0.59	117
accuracy			0.64	262
macro avg	0.64	0.63	0.64	262
weighted avg	0.64	0.64	0.64	262

Inference:

- The classification report revealed that the micro average of F1 score is about 0.63, which indicates that the trained model has a classification strength of 63%.

F1 Score

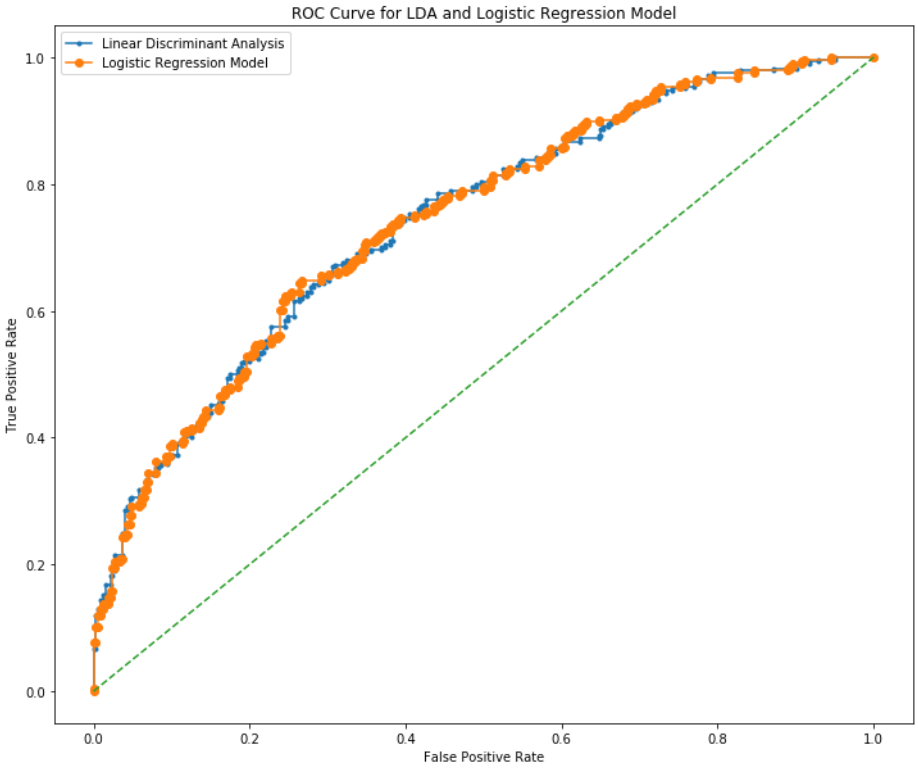
```
F1 Score: 0.5877192982456141
```

ROC_AUC estimation

```
AUC: 0.63%
```

Comparison of Two Models (LDA and Logistic Regression)

```
AUC for Linear Discriminant Analysis Train Model is 0.7421152682968979
AUC for Logistic Regression Model Train Model is 0.742720124427547
```



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

Inference:

- It appears that all models performed well for the majority class, with precision, recall metrics all above 0.7.
- Both models performance is almost the same for the minority class and arguably the "more important" classification of whether a customer was going to opt for the Holiday package or not.
- AUC for Linear Discriminant Analysis Train Model is 74.2%
- AUC for Logistic Regression Model Train Model is 74.27%

Recommendations:

- From the above ODDS ratio table, you can observe that foreign has an ODDS Ratio of 4.8933, which indicates that one unit increase in no. of foreigners increases the odds of opting Holiday_Package by 4.8933 times.
- Salary has an ODDS Ratio of 0.99, which indicates that one unit increase in Salary increases the odds of opting Holiday_Package by 0.99 times.
- educ has an ODDS Ratio of 1.083, which indicates that one unit increase in education increases the odds of opting Holiday_Package by 1.083 times.
- no_young_children has an ODDS Ratio of 0.223, which indicates that one unit increase in no of young children increases the odds of opting Holiday_Package by 0.223 times.

The END