(PM)
Predictive Modelling
PROJECT

Business Report

Done By

Hariharan Manickam

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.

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.

Out[3]:

	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

Out[4]: (26967, 11)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  26967 non-null  int64
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  object
 3   color       26967 non-null  object
 4   clarity     26967 non-null  object
 5   depth       26270 non-null  float64
 6   table       26967 non-null  float64
 7   x           26967 non-null  float64
 8   y           26967 non-null  float64
 9   z           26967 non-null  float64
 10  price       26967 non-null  int64
dtypes: float64(6), int64(2), object(3)
memory usage: 2.3+ MB
```

Out[7]:

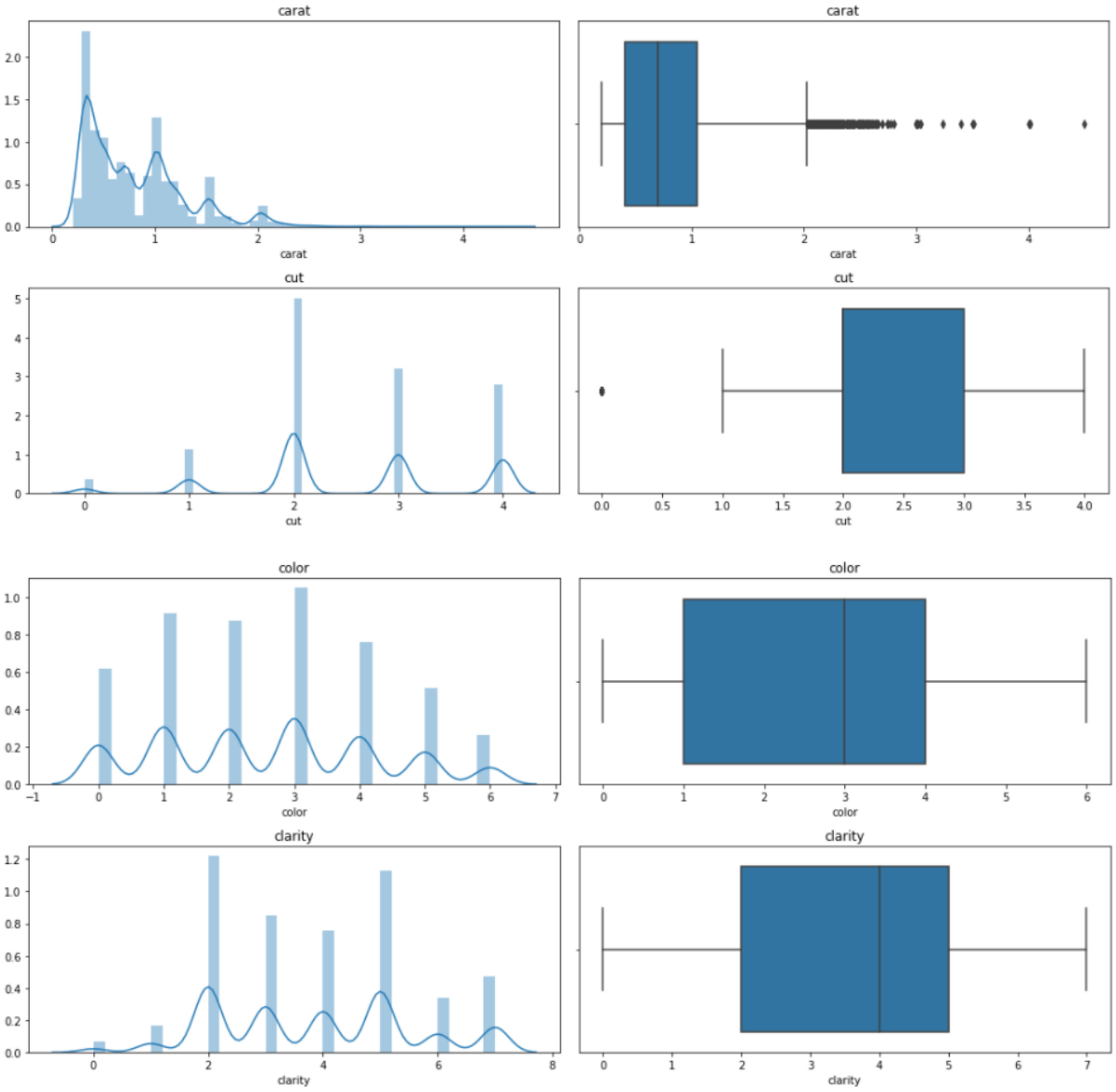
	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	13484.000000	7784.846691	1.0	6742.50	13484.00	20225.50	26967.00
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

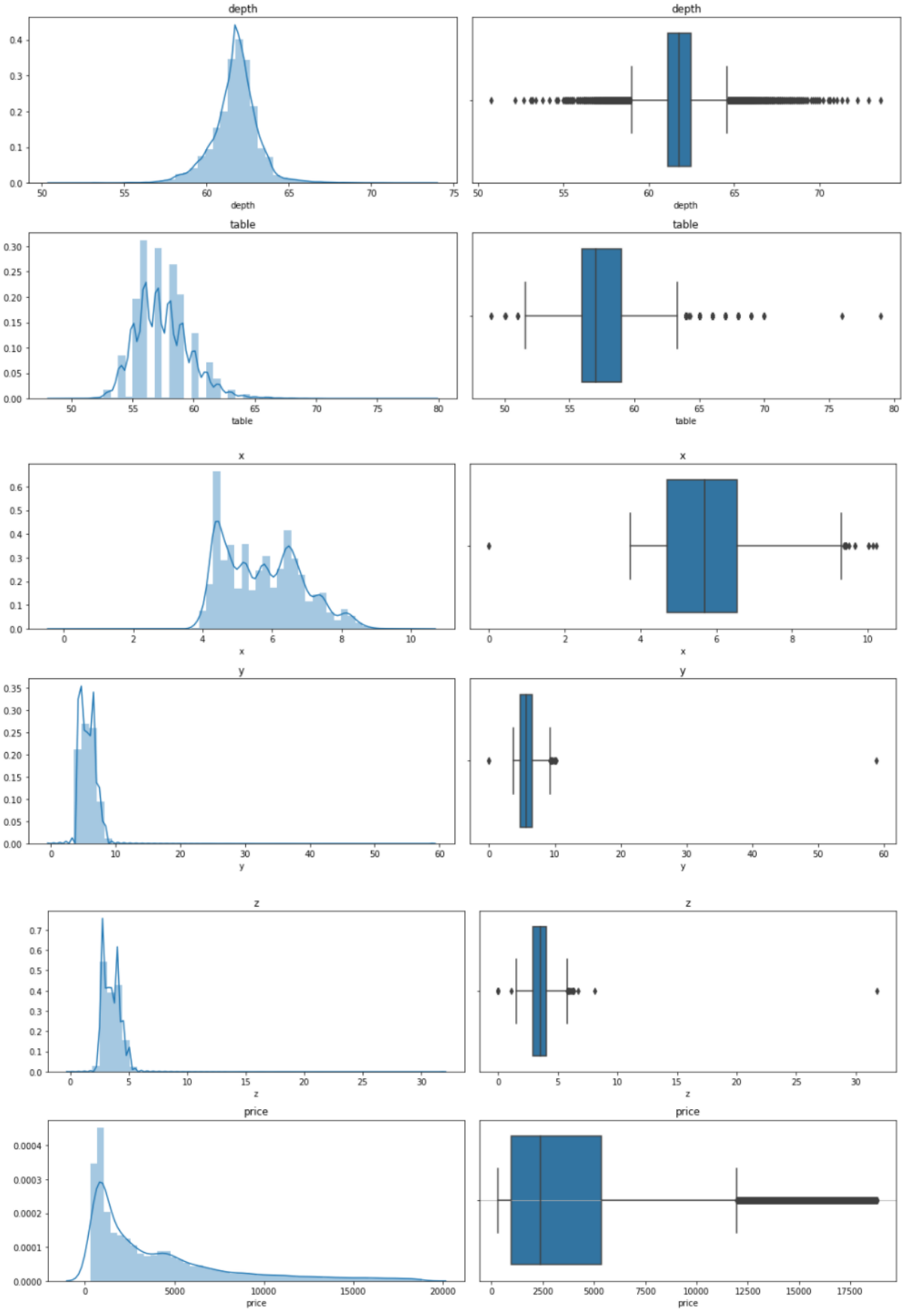
```
CUT :  5
Fair            781
Good           2441
Very Good      6030
Premium        6899
Ideal         10816
Name: cut, dtype: int64
COLOR :  7
J     1443
I     2771
D     3344
H     4102
F     4729
E     4917
G     5661
Name: color, dtype: int64
CLARITY :  8
I1        365
IF        894
VVS1     1839
VVS2     2531
VS1      4093
SI2      4575
VS2      6099
SI1      6571
Name: clarity, dtype: int64
```

Null Values (below)

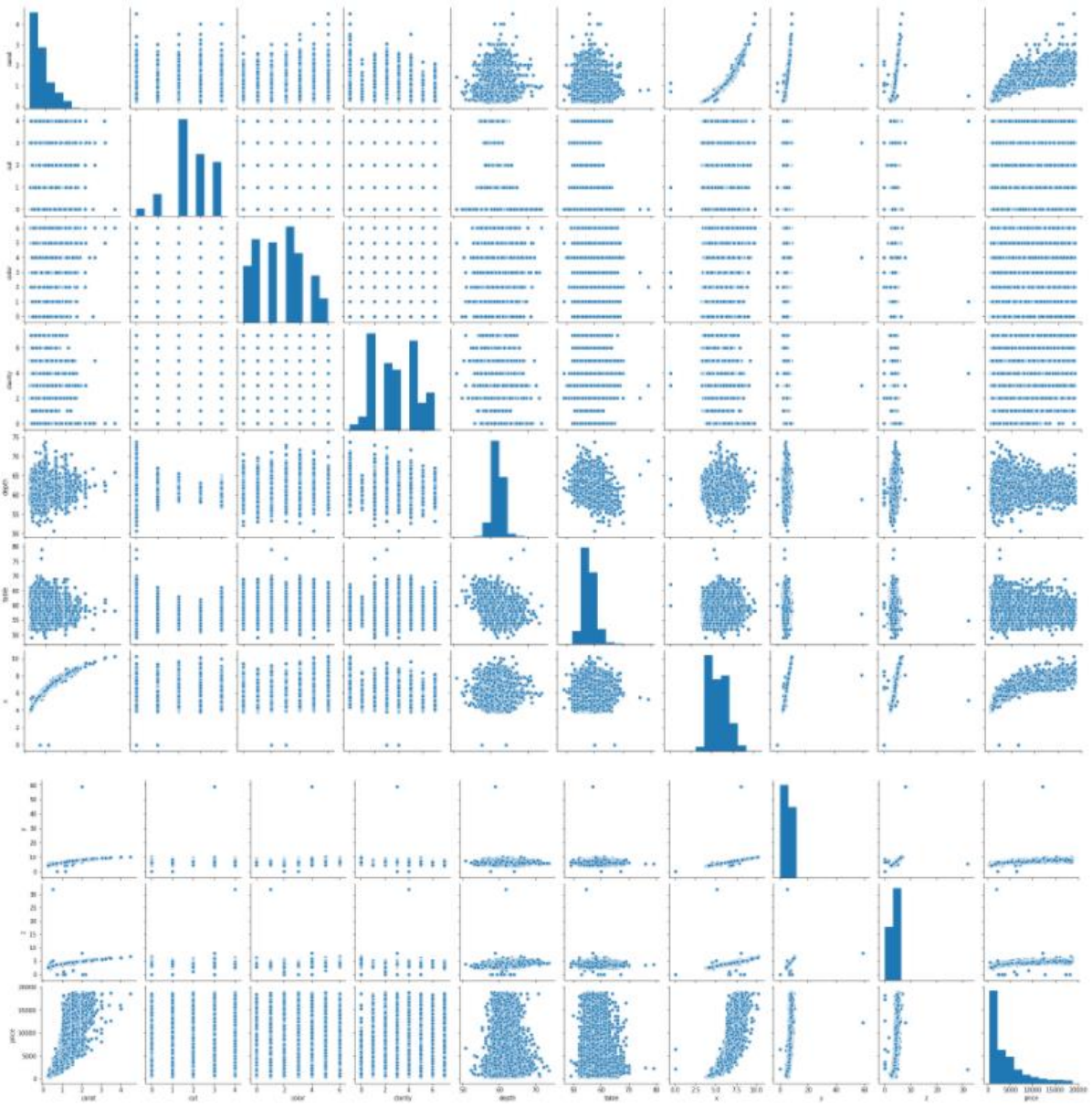
```
Out[10]: carat        0
         cut          0
         color        0
         clarity      0
         depth      697
         table        0
         x            0
         y            0
         z            0
         price        0
         dtype: int64
```

Univariate Analysis

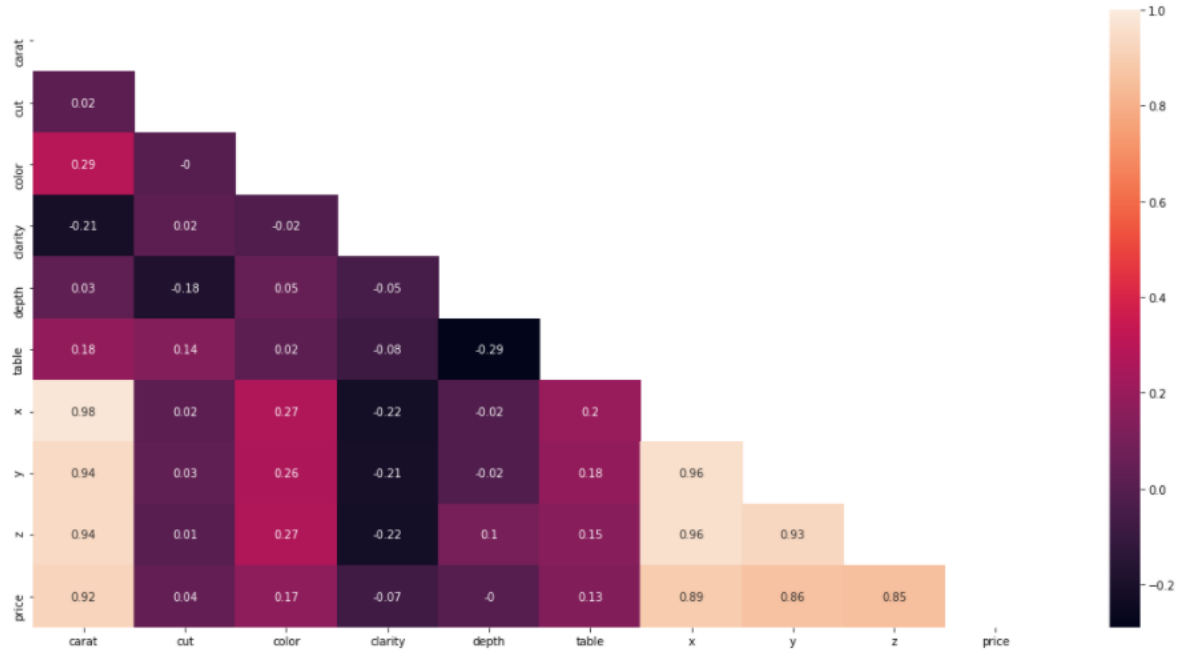



depth
depth
table
table
x
x
y
y
z
z
price
price

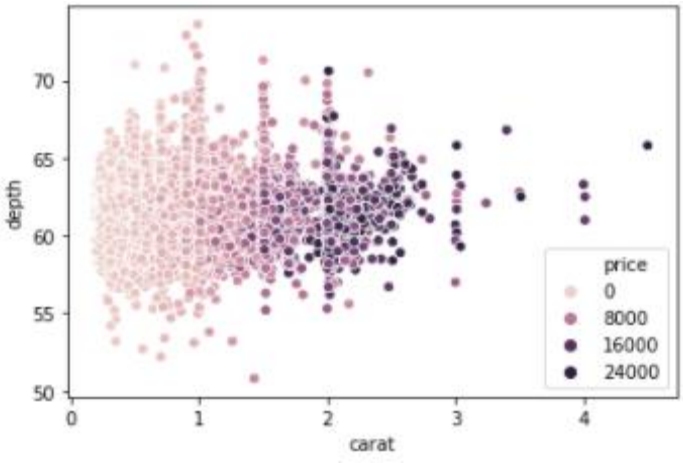
Bivariate Analysis

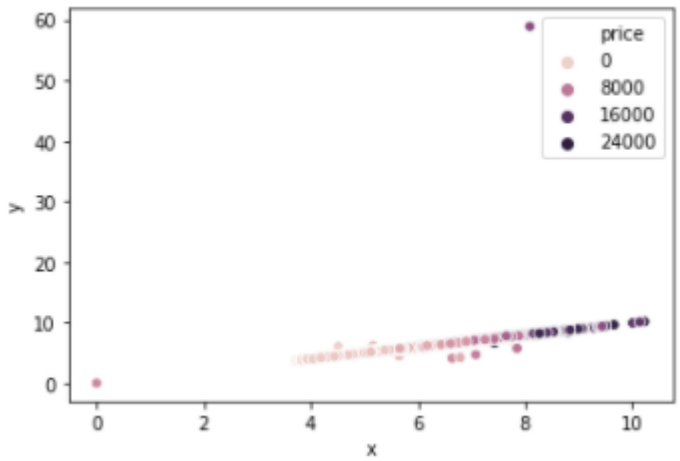


Correlation



Multivariate Analysis





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?

- The "zero" values indicates the length, width and height of zirconia in millimeter.
- Other parameters such as price and carrot are associated with the zero values.
- Hence scaling is not to be done as the attributes are in dimnesions which do not need to be standardized.

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

```
Out[32]: LinearRegression()
```

```
Out[36]: Intercept    0.001432
         carat        1.294935
         cut          0.011243
         color       -0.115210
         clarity      0.124223
         depth       -0.051908
         table       -0.046864
         x           -0.788456
         y            0.513132
         z           -0.014394
         dtype: float64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.885
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                 1.618e+04
Date:                Tue, 29 Jun 2021   Prob (F-statistic):               0.00
Time:                        18:14:55   Log-Likelihood:                -6352.3
No. Observations:               18853   AIC:                         1.272e+04
Df Residuals:                   18843   BIC:                         1.280e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0014      0.002      0.580      0.562      -0.003       0.006
carat          1.2949      0.012    109.821      0.000       1.272       1.318
cut            0.0112      0.003      4.433      0.000       0.006       0.016
color         -0.1152      0.003    -44.331      0.000      -0.120      -0.110
clarity        0.1242      0.003     48.865      0.000       0.119       0.129
depth         -0.0519      0.003    -18.042      0.000      -0.058      -0.046
table         -0.0469      0.003    -17.459      0.000      -0.052      -0.042
x             -0.7885      0.047    -16.897      0.000      -0.880      -0.697
y              0.5131      0.048     10.776      0.000       0.420       0.606
z             -0.0144      0.008     -1.811      0.070      -0.030       0.001
==============================================================================
Omnibus:                     4432.251   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           163854.651
Skew:                           0.397   Prob(JB):                         0.00
Kurtosis:                      17.421   Cond. No.                         53.7
==============================================================================
```

Performance Score (Training)

```
Out[39]: 0.8854038589924862
```

Performance Score (Testing)

```
Out[40]: 0.8294698627154848
```

RMSE (Testing)

```
Out[41]: 0.338916610866983
```

RMSE (Training)

```
Out[42]: 0.4118232692058547
```

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

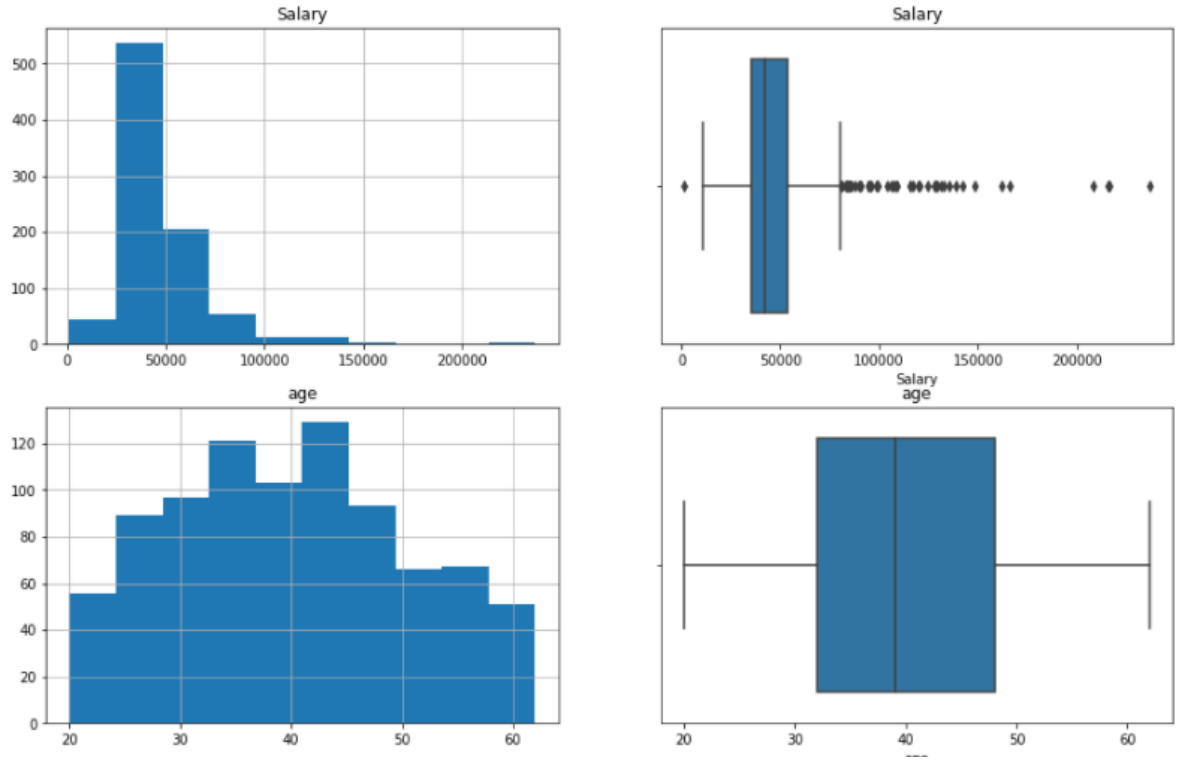
1. Classifying the stones based on quality cut, color and clarity is not seeming to help the company.
2. Focussing on creating stones of cubic zirconia with higher carat weight(avg weight of more than 2 carats) for a higher profit margin.
3. Less profitable stones should be of less than average size of 5mm in length, width and breadth of the stone so girdle diameter will reduce. This has an impact on the weight of the cubic zirconia stone.
4. Price depends more on the carat, length, width and height of the stone more than anything else. (carat > length > width > height)

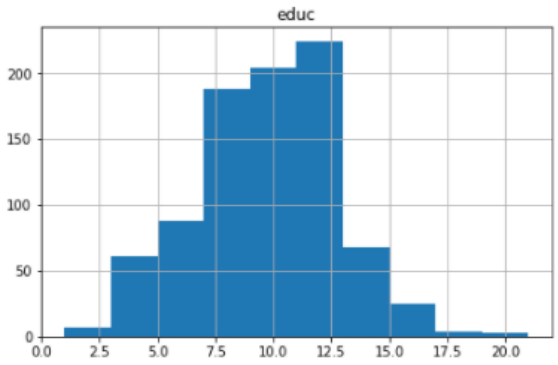
Problem 2: Logistic Regression and LDA

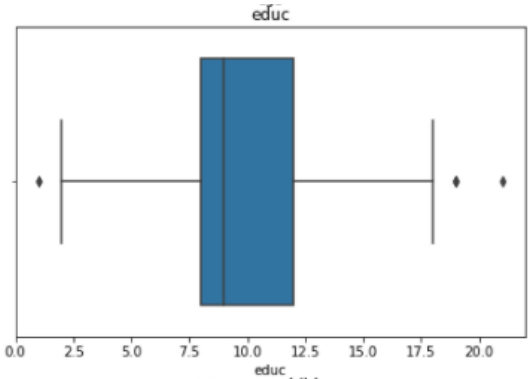
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not 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.

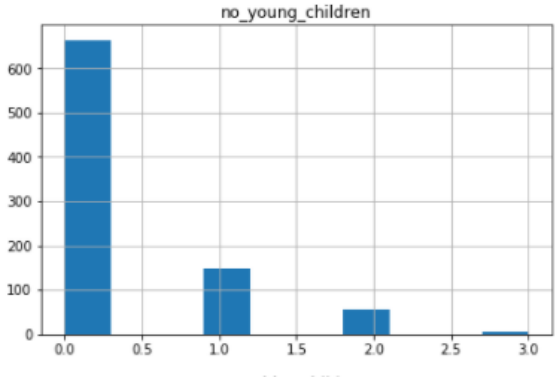
2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

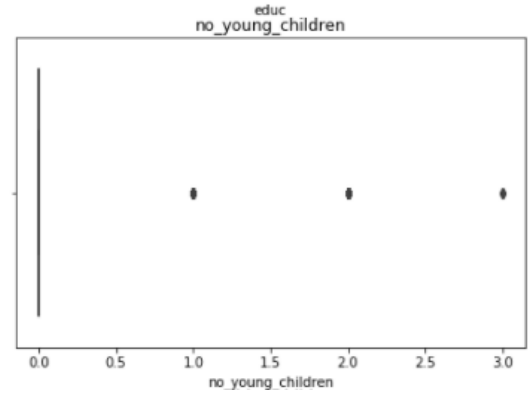
Univariate Analysis

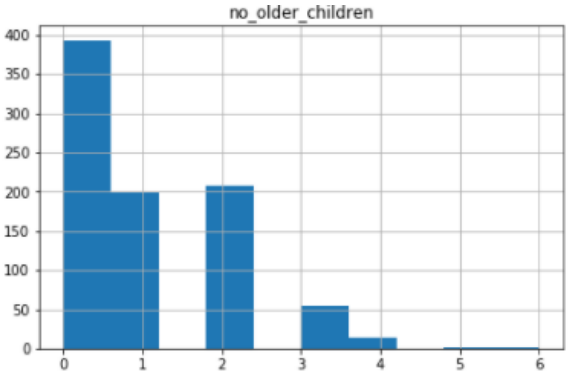


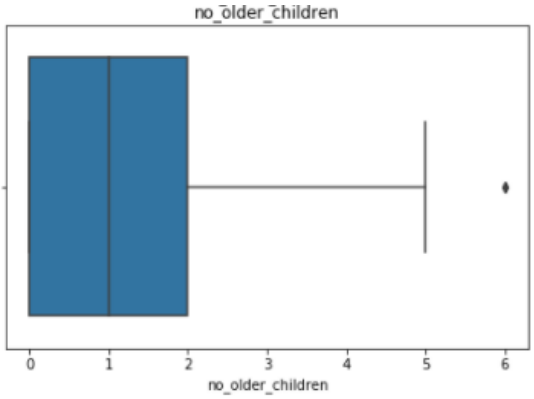

educ
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
0 50 100 150 200


educ
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
educ

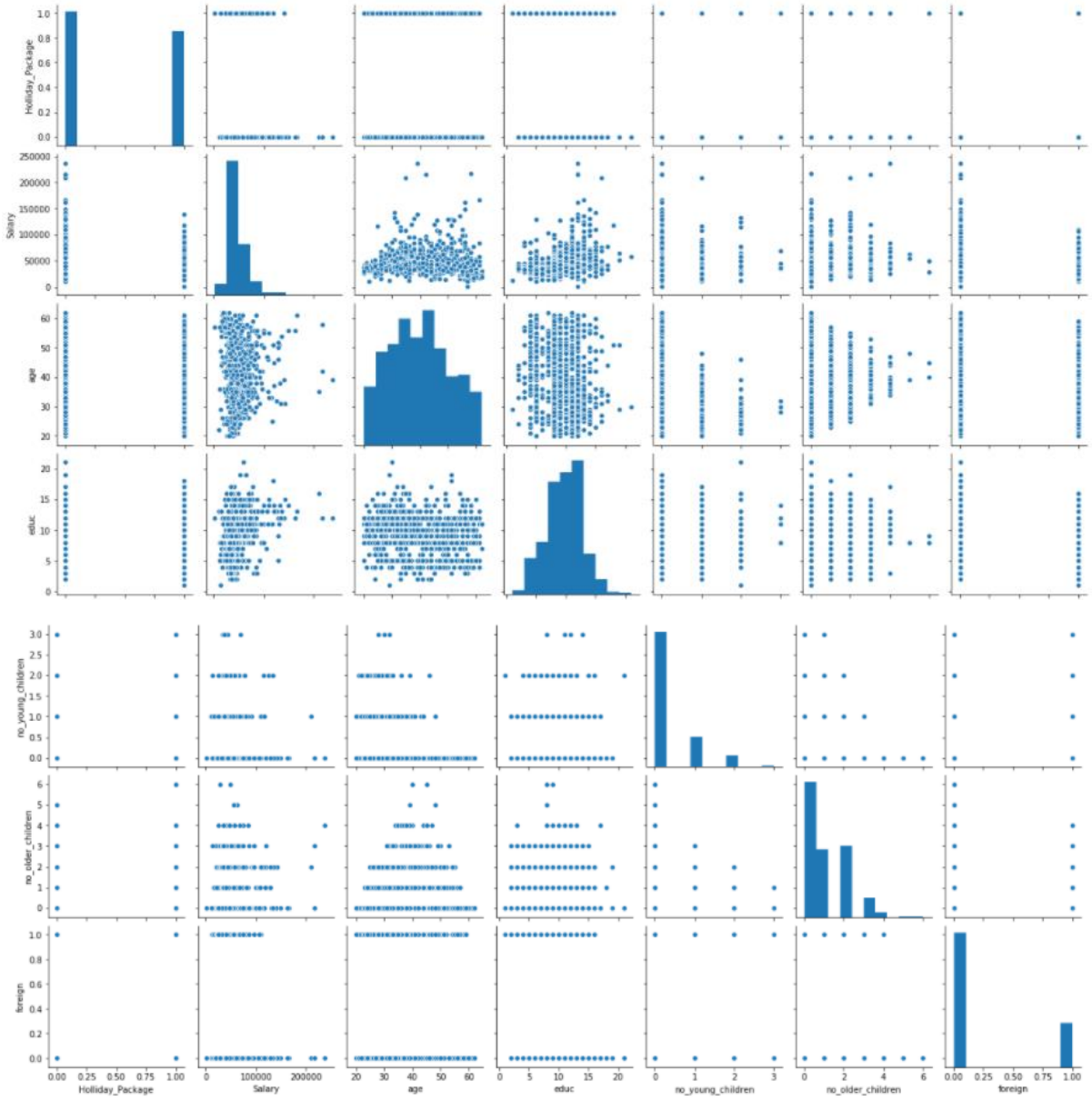

no_young_children
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 100 200 300 400 500 600


no_young_children
0.0 0.5 1.0 1.5 2.0 2.5 3.0
no_young_children

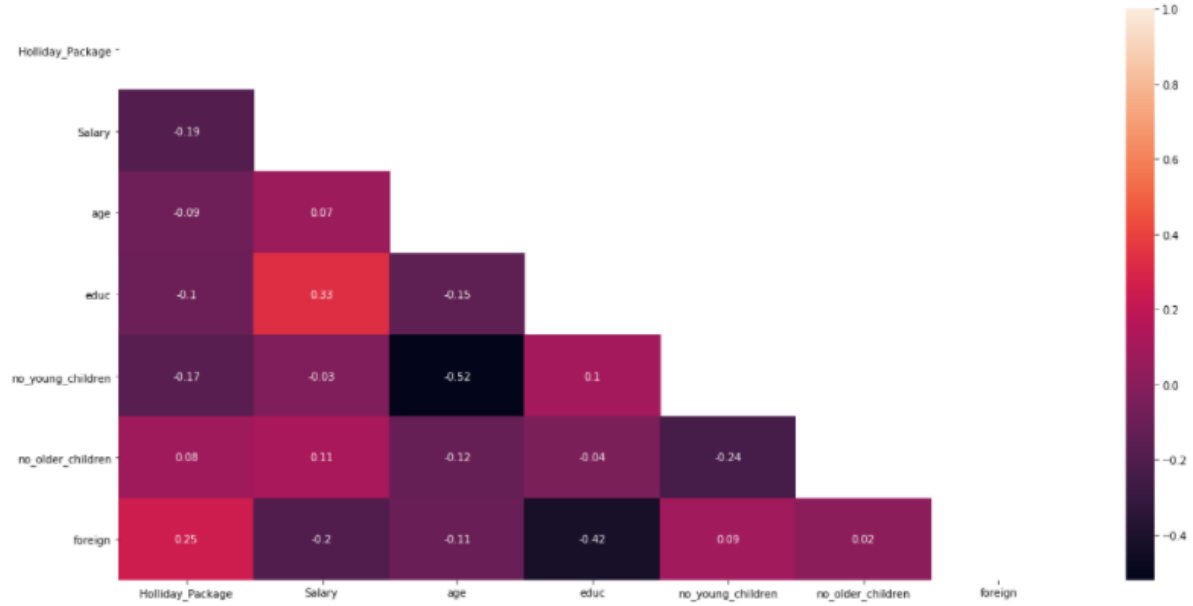

no_older_children
0 1 2 3 4 5 6
0 50 100 150 200 250 300 350 400


no_older_children
0 1 2 3 4 5 6
no_older_children

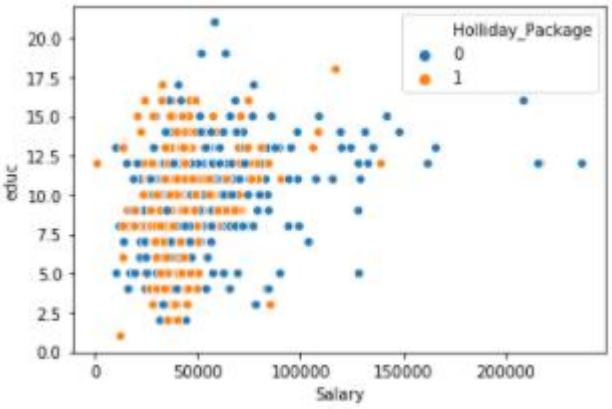
Bivariate Analysis

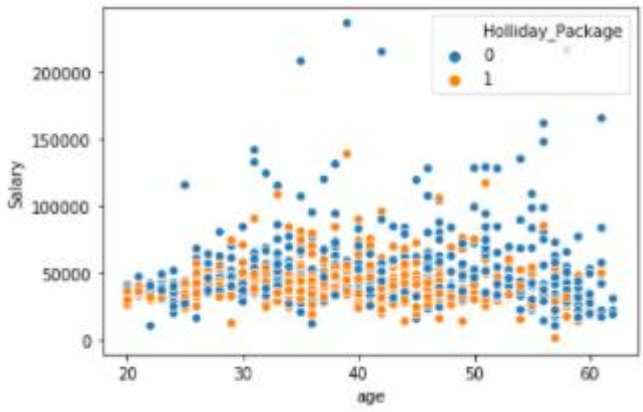


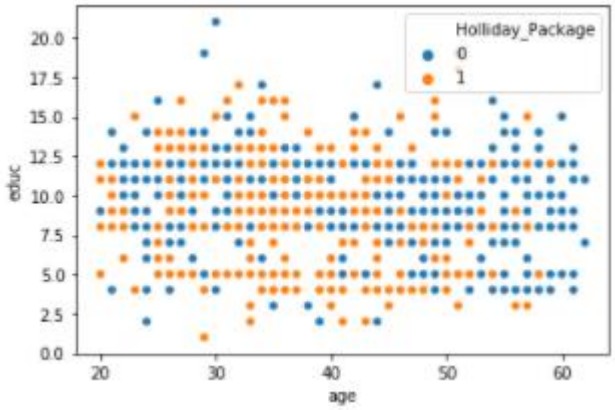
Correlation



Multivariate Analysis







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).

```
X_train (610, 6)
X_test (262, 6)
y_train (610,)
y_test (262,)
```

```
Out[74]: LogisticRegression(C=1, intercept_scaling=1.0, solver='liblinear', tol=1e-05)
```

```
Out[83]: LinearDiscriminantAnalysis(shrinkage=0.0, solver='lsqr')
```

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

Performance Score (Training)

```
Out[76]: 0.6754098360655738
```

Performance Score (Tresting)

```
Out[77]: 0.6755725190839694
```

Confusion Matrix (Training)

```
Out[78]: array([[254,  72],
                [126, 158]], dtype=int64)
```

Confusion Matrix (Testing)

```
Out[79]: array([[111,  34],
                [ 51,  66]], dtype=int64)
```

Classification Report (Training)

	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.56	0.61	284
accuracy			0.68	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.68	0.67	610

Classification Report (Testing)

	precision	recall	f1-score	support
0	0.69	0.77	0.72	145
1	0.66	0.56	0.61	117
accuracy			0.68	262
macro avg	0.67	0.66	0.67	262
weighted avg	0.67	0.68	0.67	262

AUC ROC (Training)

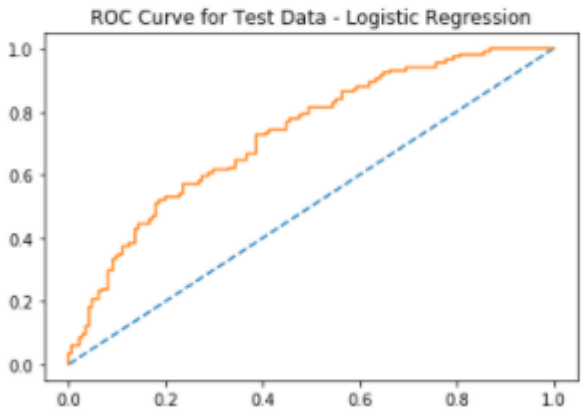
```
AUC: 0.724
```

```
Out[86]: Text(0.5, 1.0, 'ROC Curve for Training data - Logistic Regression')
```

AUC ROC (Testing)

```
AUC: 0.728
```

```
Out[87]: Text(0.5, 1.0, 'ROC Curve for Test Data - Logistic Regression')
```



Linear Discriminant Analysis

Performance Score (Training)

```
Out[90]: 0.6688524590163935
```

Performance Score (Training)

```
Out[91]: 0.6603053435114504
```

Confusion Matrix (Training)

```
Out[93]: array([[107,  38],
                [ 51,  66]], dtype=int64)
```

Confusion Matrix (Testing)

```
Out[93]: array([[107,  38],
                [ 51,  66]], dtype=int64)
```

Classification Report (Training)

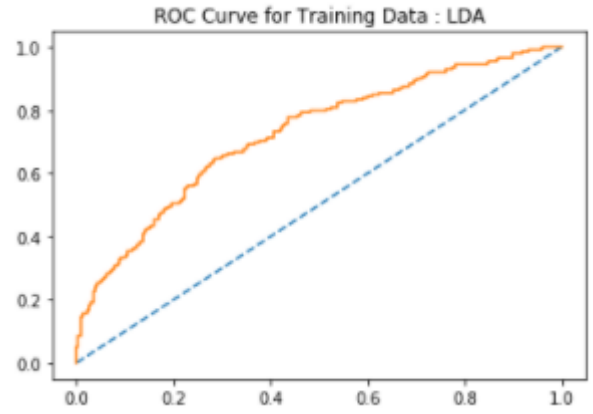
	precision	recall	f1-score	support
0	0.67	0.76	0.71	326
1	0.67	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Classification Report (Testing)

	precision	recall	f1-score	support
0	0.68	0.74	0.71	145
1	0.63	0.56	0.60	117
accuracy			0.66	262
macro avg	0.66	0.65	0.65	262
weighted avg	0.66	0.66	0.66	262

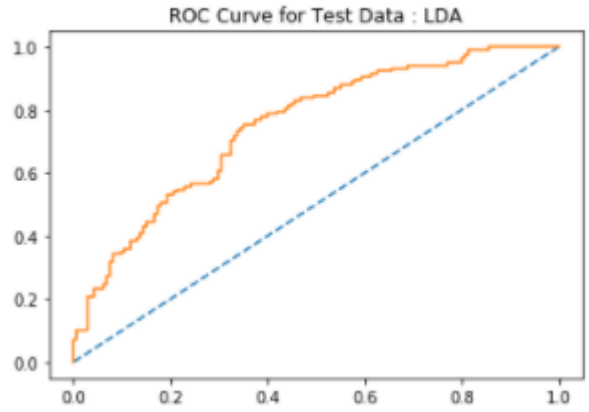
AUC ROC (Training)

```
AUC: 0.725
```



AUC ROC (testing)

```
AUC: 0.748
```



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

Both models have a very similar accuracy. Linear Discriminant Analysis has a slightly better AUC Score than Logistic Regression. Hence we will choose Linear Discriminant Analysis.

1. Employees with a salary of less than 50000-55000 choose the holiday package. This makes sense as people with less money to spend will want the better deal/package.
2. The increase in salary directly reduces the probability of the employee choosing the holiday package. We can observe this as there is a drastic decline in the number of employees opting for the holiday package post the salary range of 55000.
3. Customers above the age of 40-45 tend to not opt for holiday package.
4. The tour and travel agency should try to target employees with the following profile. Salary range : 50k to 55k, Age : Less than 45