PREDICTIVE MODELING PROJECT
BUSINESS REPORT

BY

RITUSRI MOHAN

CONTENTS

Problem 1– Zirconia Data Analysis 3
Problem 1.1 3
Problem 1.2 8
Problem 1.3 10
Problem 1.4 12

Problem 2– Holiday Package Data Analysis 13
Problem 2.1 13
Problem 2.2 18
Problem 2.3 19
Problem 2.4 23

PROBLEM 1

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

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**.**

The head of the dataset can be seen below. There are 11 columns. The first column, i.e., 'Unnamed: 0' is dropped from the dataset as it does not contribute to further analysis.

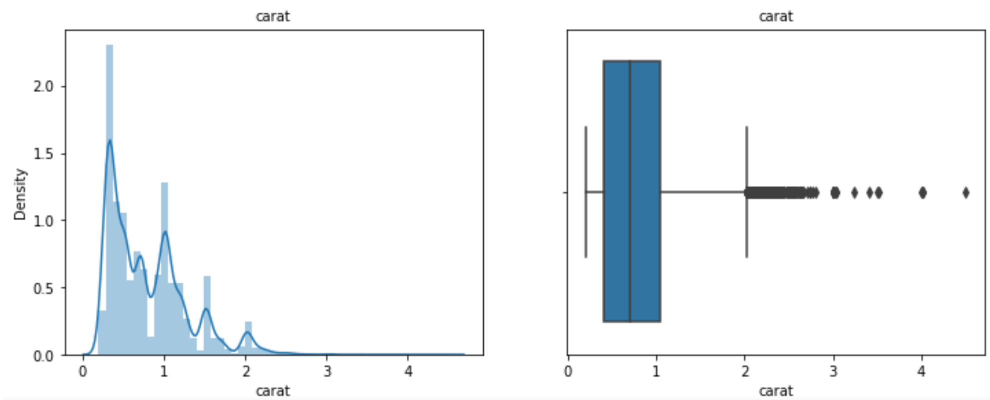
	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35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

From the summary of the dataset below, it is seen that there is difference in the scale of the values across the columns. There is a chance for the outliers to be present as there is a huge difference between the 75% value and max value for some variables.

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

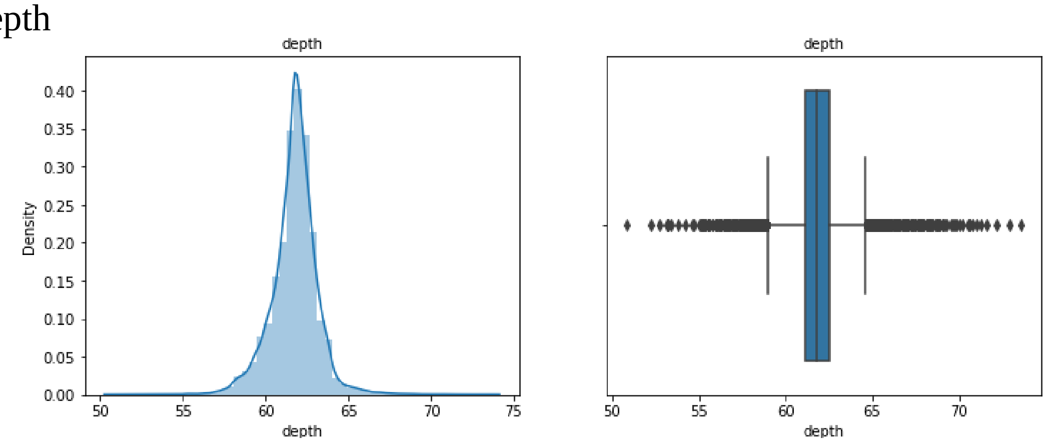
Univariate Analysis

- Carat



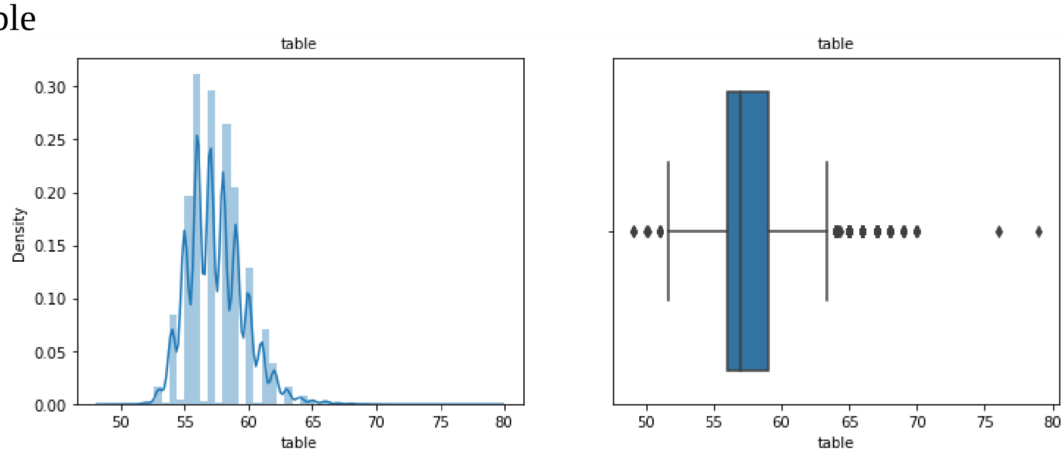
The distribution seems to be random and there are many outliers present.

- Depth



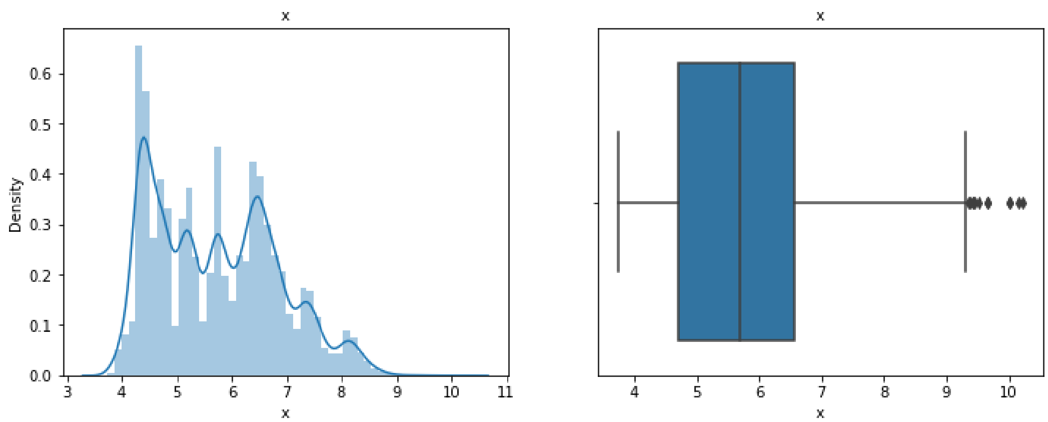
The distribution seems to be normal and there are many outliers present.

- Table



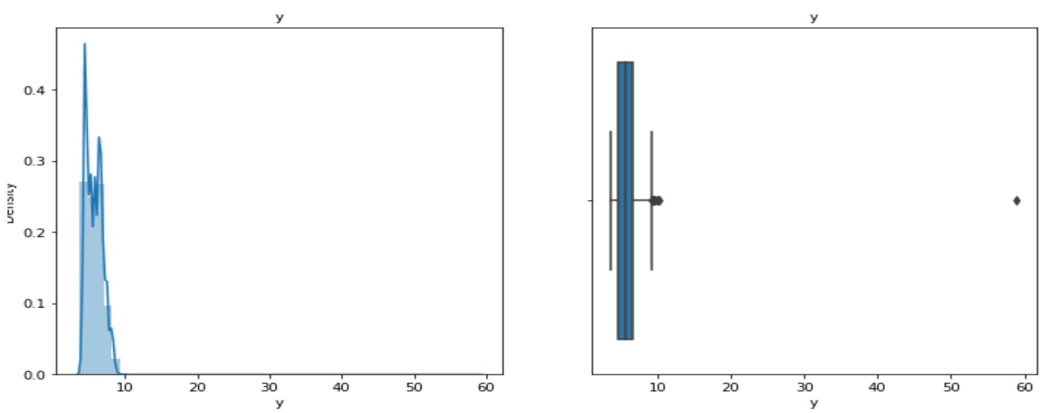
The distribution seems to be random and there are many outliers present.

- x



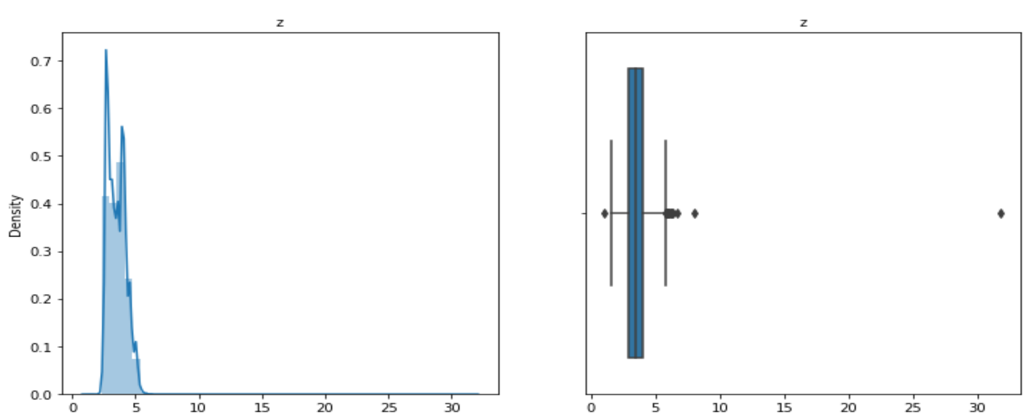
The distribution seems to be random and there are some outliers present.

- y



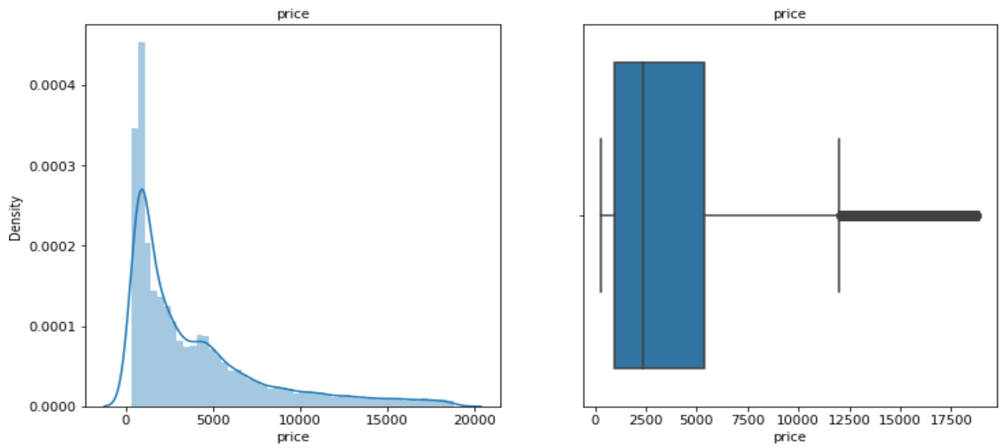
The distribution seems to be random and there are few outliers present.

- z



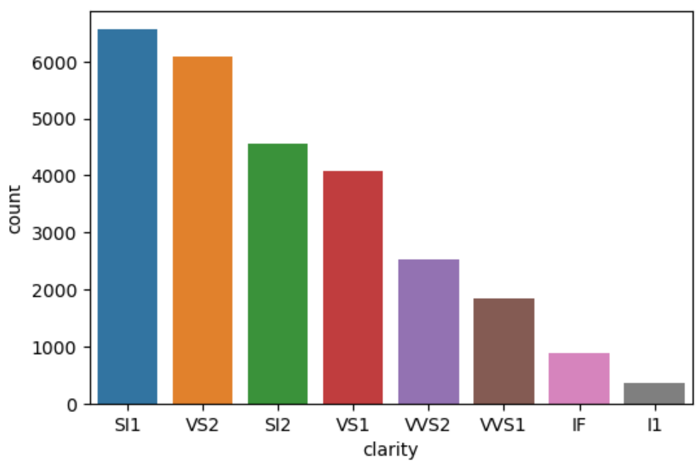
The distribution seems to be random and there are few outliers present.

• Price



The distribution is positively skewed and there are plenty outliers present.

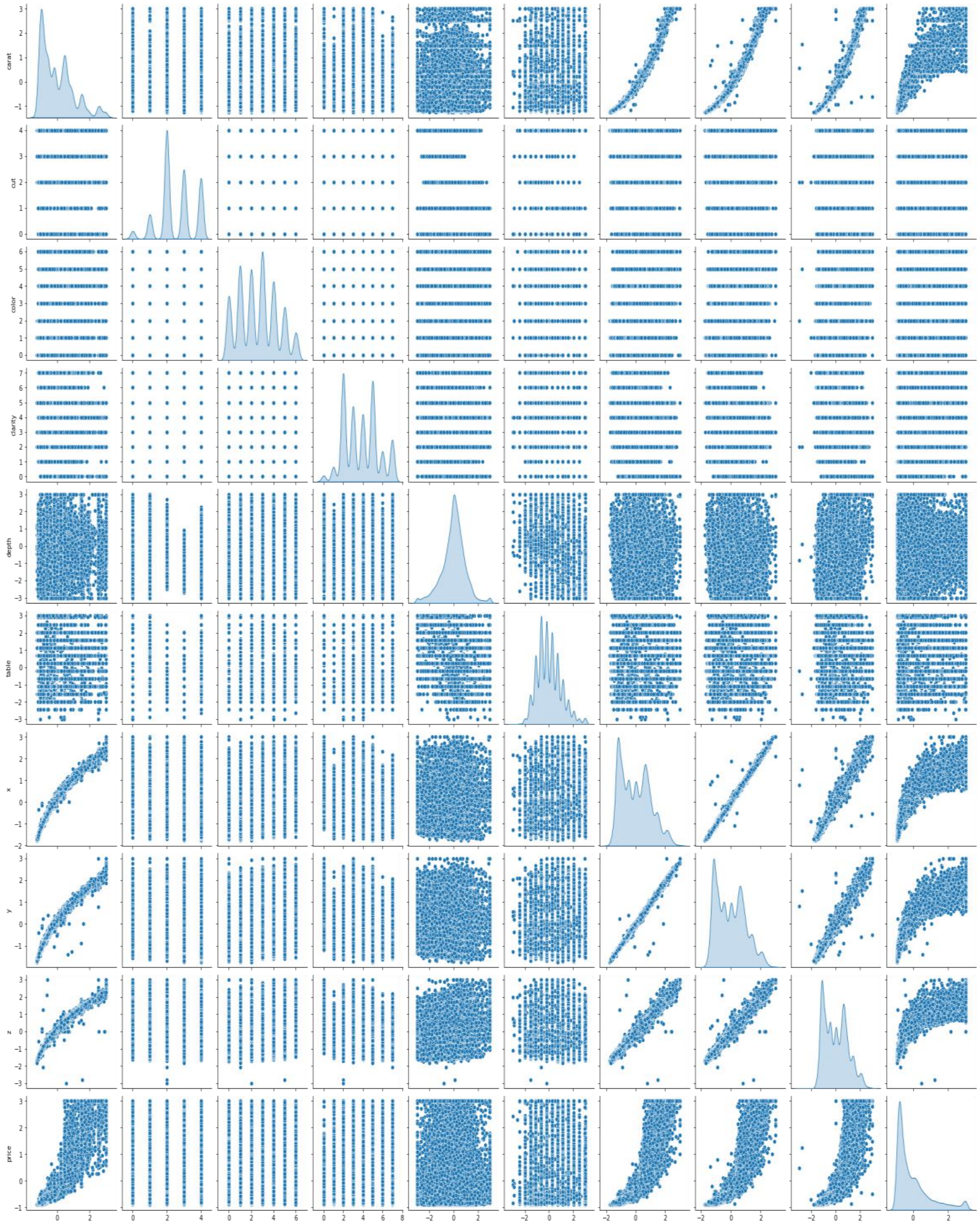
• Clarity



It is seen that count of Sl1 clarity zirconia is the most.

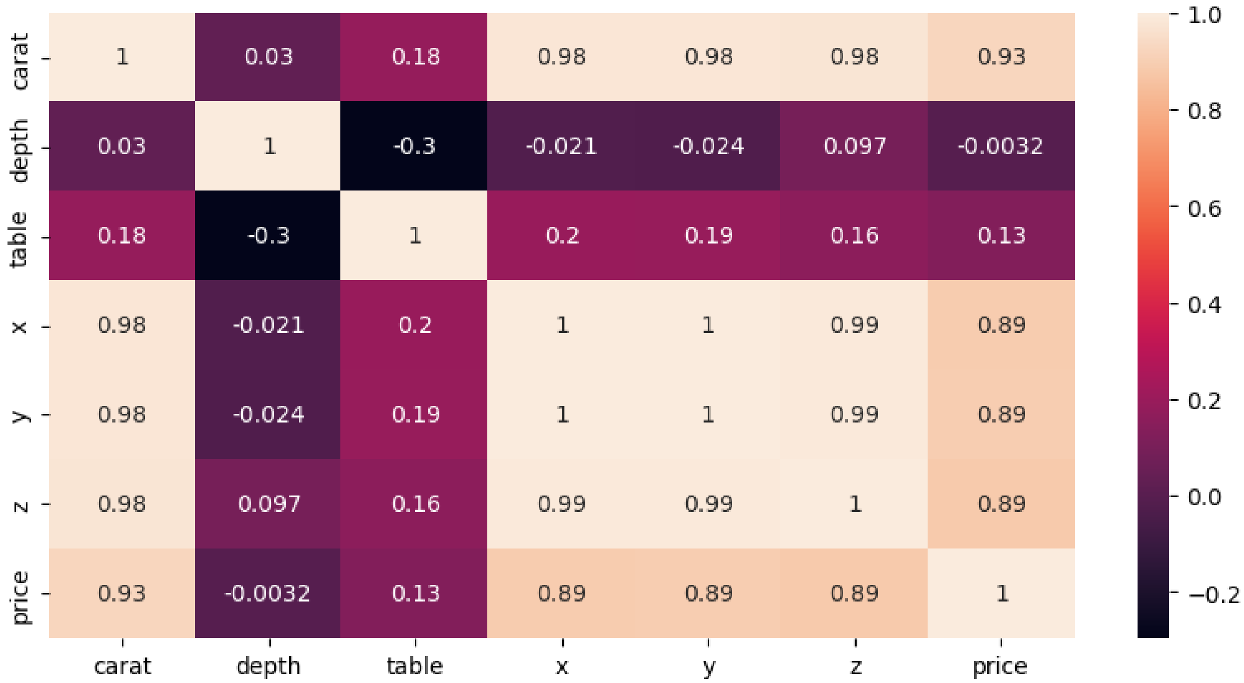
Multivariate Analysis

- Pairplot



This plot helps us to understand the relationship between all the numerical values in the dataset and establish the trends in the dataset.

• Heatmap



- There is a strong positive correlation between ‘carat’ and the dimensions, i.e., ‘x’, ‘y’, ‘z’.
- There is a strong positive correlation between ‘carat’ and ‘price’ which indicates that higher the carat weight of the cubic zirconia, higher is its price.
- There is a strong negative correlation between ‘price’ and ‘depth’/ ‘table’.

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?

There are 697 null values in ‘depth’. The 26th record showing the null value is shown below.

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

	carat	cut	color	clarity	depth	table	x	y	z	price
26	0.34	Ideal	D	SI1	NaN	57.0	4.50	4.44	2.74	803

These 'NaN' values were imputed with mean value of 'depth'. The 26th record is shown after imputation.

	carat	cut	color	clarity	depth	table	x	y	z	price
26	0.34	Ideal	D	SI1	61.745147	57.0	4.5	4.44	2.74	803

The number of 0 values in the dataset is shown below:

```
Number of zero values for the carat is 0
Number of zero values for the cut is 0
Number of zero values for the color is 0
Number of zero values for the clarity is 0
Number of zero values for the depth is 0
Number of zero values for the table is 0
Number of zero values for the x is 3
Number of zero values for the y is 3
Number of zero values for the z is 9
Number of zero values for the price is 0
```

The 0 values of 'x', 'y' and 'z' are imputed with their respective mean values. After imputation, the number of 0 values in the dataset are shown below:

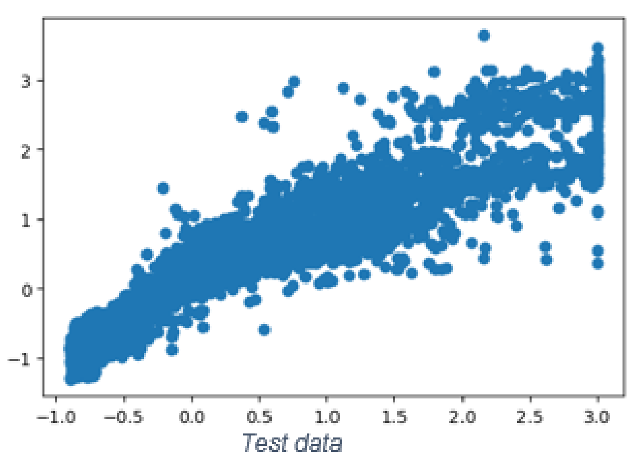
```
Number of zero values for the carat is 0
Number of zero values for the cut is 0
Number of zero values for the color is 0
Number of zero values for the clarity is 0
Number of zero values for the depth is 0
Number of zero values for the table is 0
Number of zero values for the x is 0
Number of zero values for the y is 0
Number of zero values for the z is 0
Number of zero values for the price is 0
```

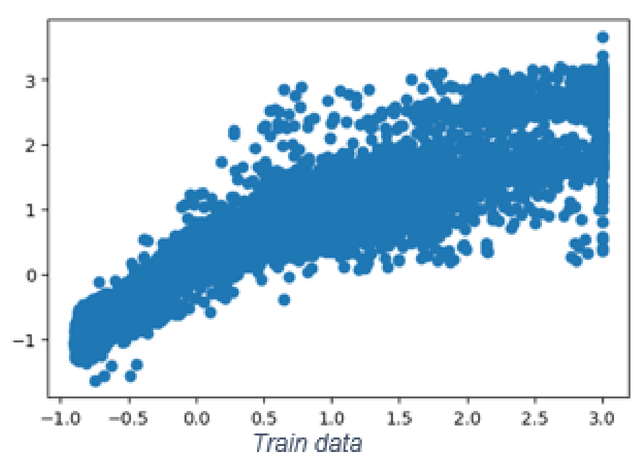
1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train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

Encoding, splitting of data and applying linear regression are done in the Jupyter code file.

<u>Model 1</u>

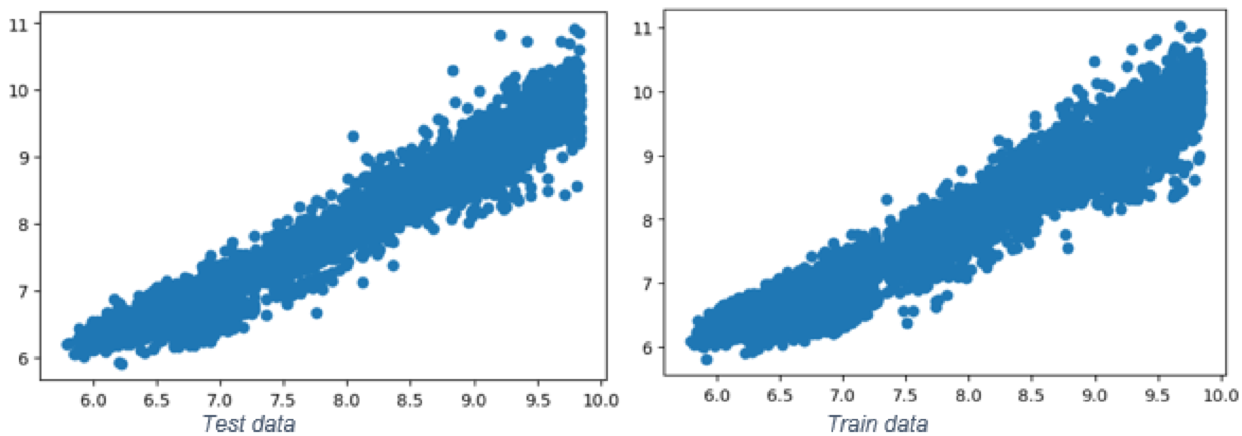
```
Intercept    -0.111082
carat         1.481283
cut           0.010084
color        -0.068422
clarity       0.068152
depth        -0.023274
table        -0.049422
x            -0.743269
y             0.586980
z            -0.324482
dtype: float64

R-squared:0.902
MSE(training):2.717818610753163e-34
MSE(testing): 0.09273158174263757
RMSE(training): 1.6485807868446007e-17
RMSE(testing):  0.304518606562288
```





Model 2

```
Intercept     2.452438
color        -0.070634
clarity       0.061969
x             0.920610
dtype: float64

R-squared: 0.947
MSE(training): 0.054735294847755196
MSE(testing): 0.05556891869202576
RMSE(training): 0.23395575403856858
RMSE(testing): 0.23573060618431743
```



The intercept in Model 1 was meaningless as it was a negative value. Model 2 has a positive intercept value and also a higher R-squared value when compared to Model 1.
Therefore, Model 2 is chosen.
The linear equation formed using Model 2 is:

log (Price) = (-0.0706* color) + (0.0619* clarity) + (0.9206* x) + 2.4524

- When Color increases by 1 unit, price of zirconia decreases by 0.0706 units, keeping all other predictors constant.

- When X(Length.) increases by 1 unit, prices increases by 0.9206 units, keeping all other predictors constant.

- When clarity increases by 1 unit, prices increase by 0.0619 units, keeping all other predictors constant.

.

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

- As the price of zirconia is hugely dependent on x(length) and, x and carat are highly correlated, the company should try to manufacture high carat zirconia to get better pricing

- The price of Premium and Fair cut is maximum. Hence we should try to maximize the manufacturing of zirconia with such cuts.

- From the count plot of ‘clarity’, it is seen that l1(which is best) is the least manufactured. Higher clarity zirconia needs to be manufactured to attain better pricing.

- ‘Good’ and ‘Fair’ category of zirconia cut must be reduced.

PROBLEM 2

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

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.

The head of the dataset can be seen below. There are 8 columns. The first column, i.e., 'Unnamed: 0' is dropped from the dataset as it does not contribute to further analysis.

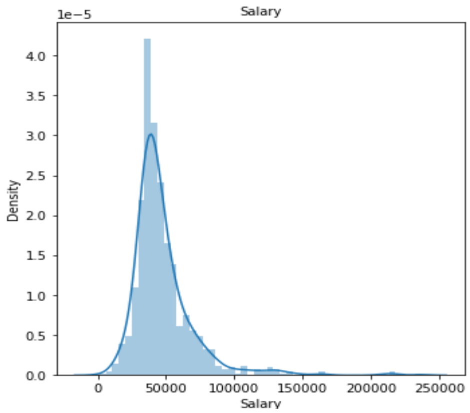
	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

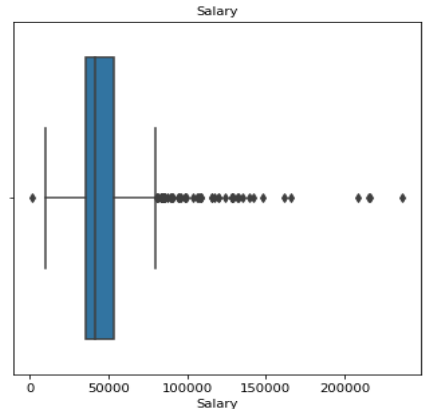
From the summary of the dataset below, it is seen that there is difference in the scale of the values across the columns. There is a chance for the outliers to be present as there is a huge difference between the 75% value and max value for some variables

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.000000	62.000000	21.000000	3.000000	6.000000

Univariate Analysis & Bivariate Analysis

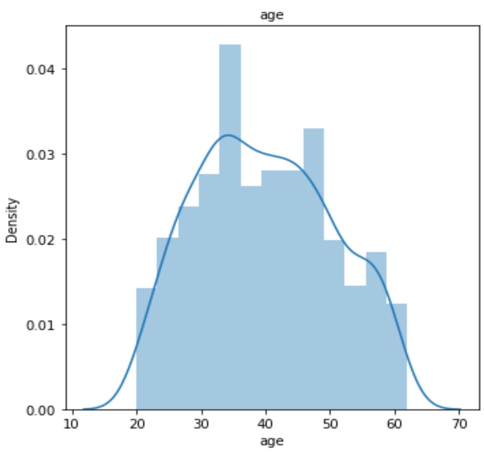
- Salary

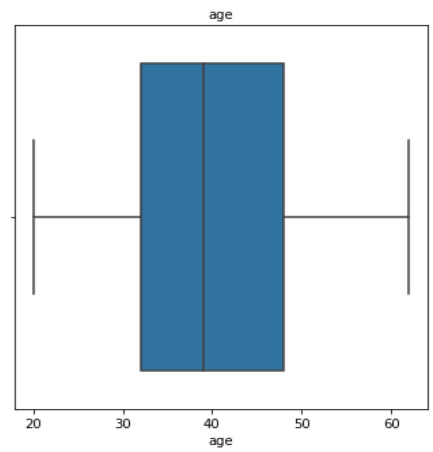




The distribution is positively skewed and there are many outliers.

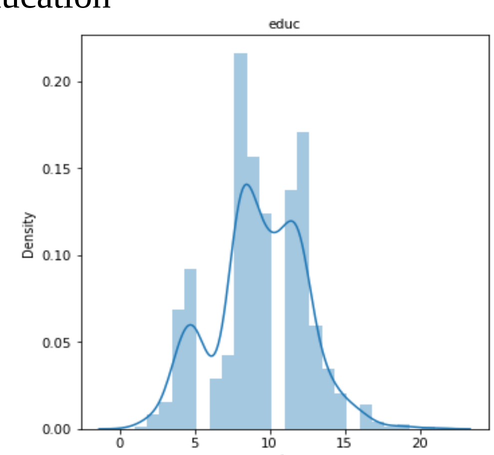
- Age

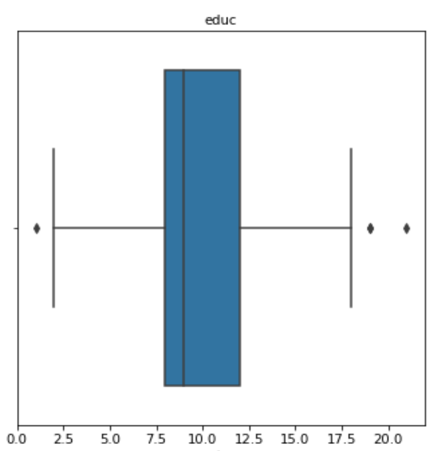




The distribution seems to be normal without any outliers.

- Education

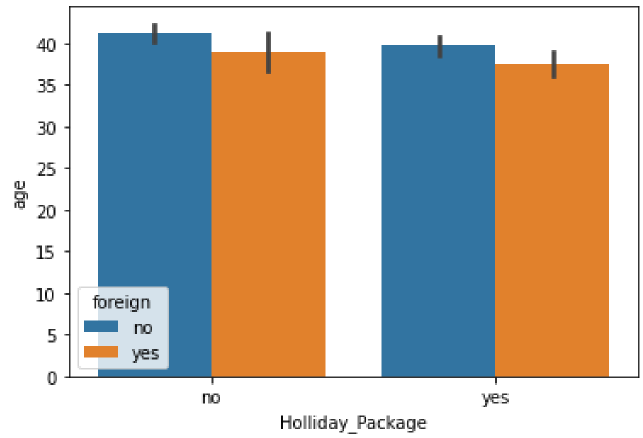




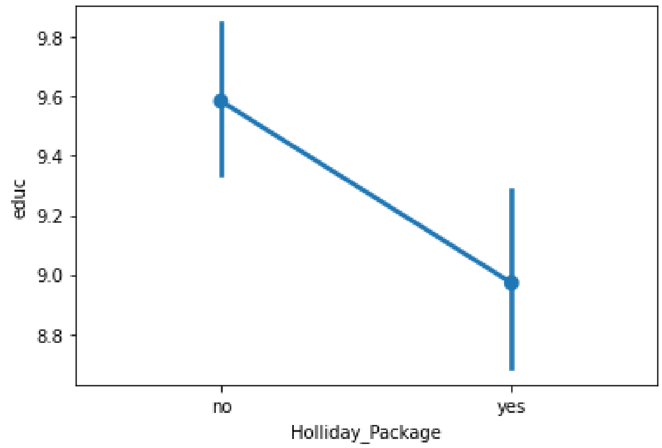
The distribution seems to be normal with very few outliers.

- Holiday Package

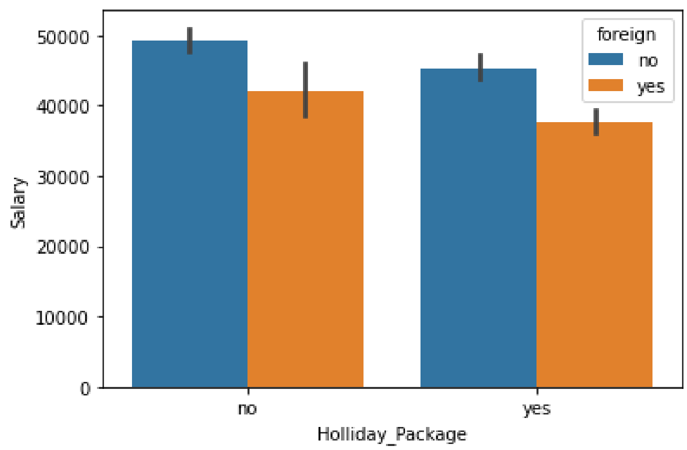
  - Non-foreigners have opted for holiday package more as compared to foreign individuals.



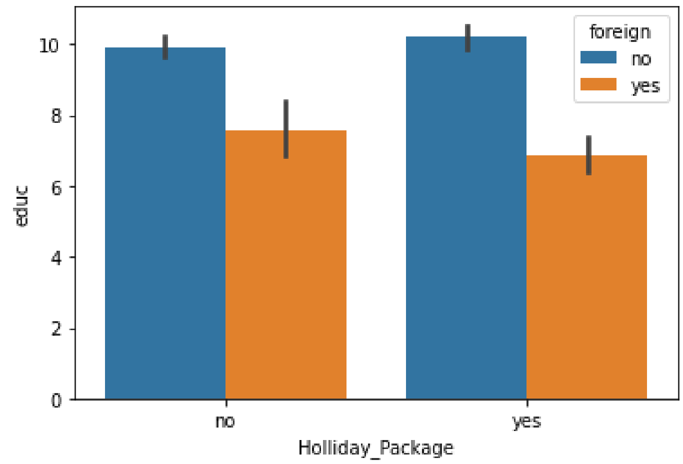
  - People with lower level of education have opted for holiday package when more compared to those with higher education.



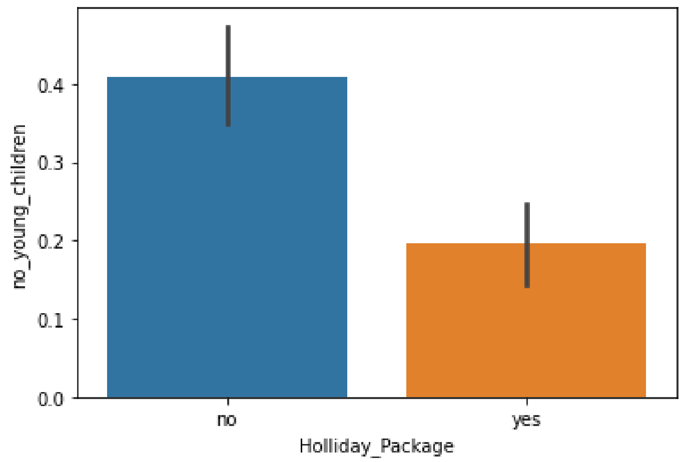
  - Most of the non-foreigners who have chosen the holiday package have a salary between 35,000 and 40,000.



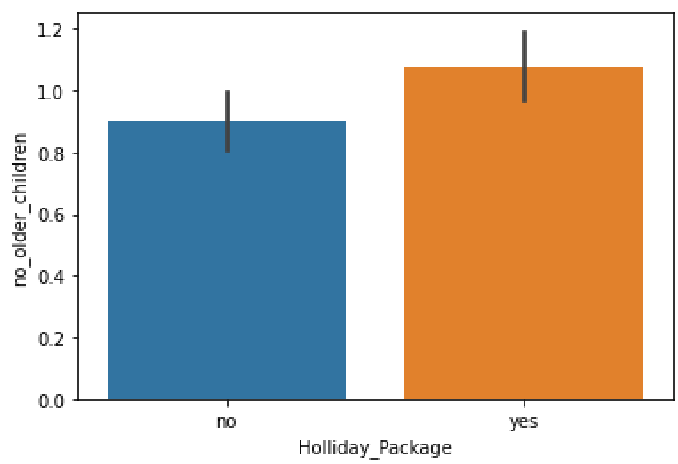
- The number of customers who have not chosen the holiday package are non-foreigners.



- More number of customers who have lower number of young children have opted for holiday package when compared to those having more number of young children.



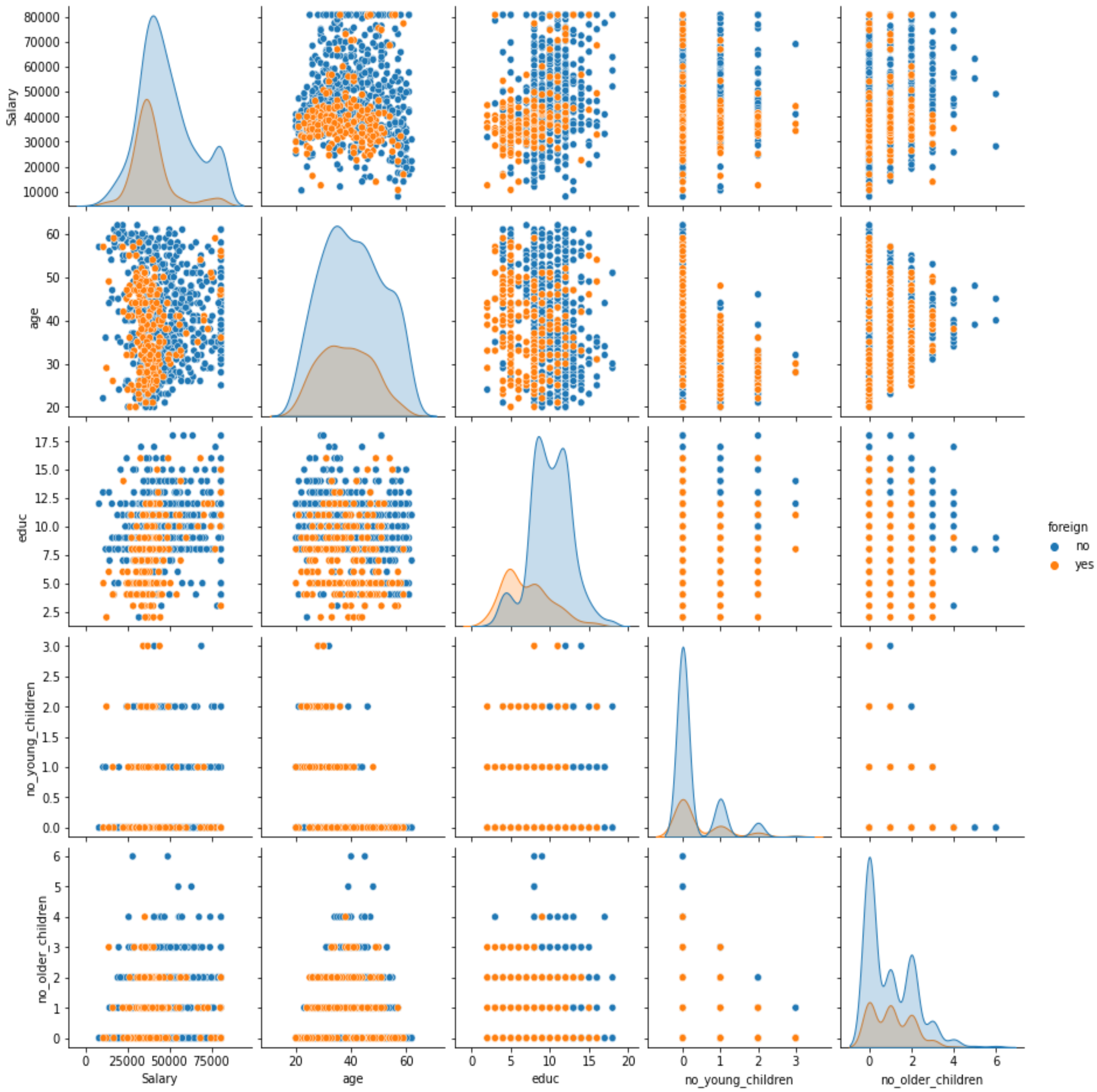
- More number of customers who have higher number of older children have opted for holiday package when compared to those having lower number of old children.



.

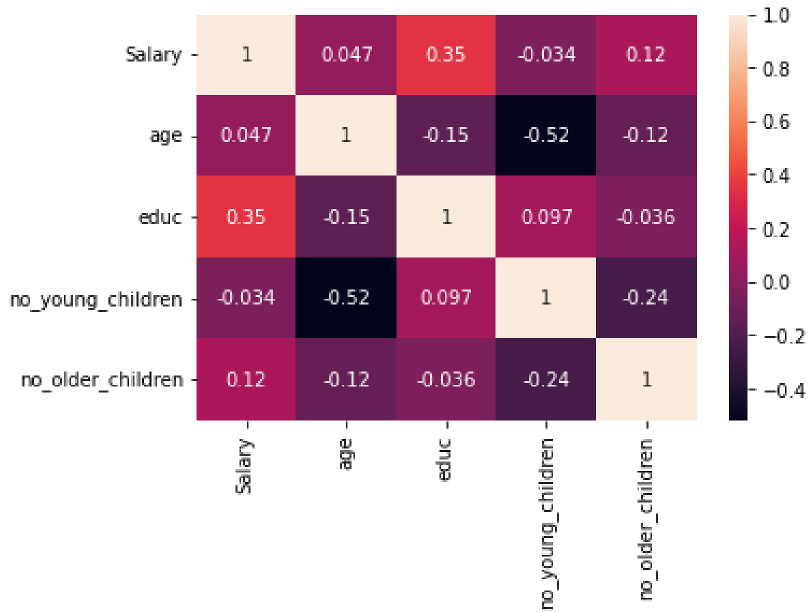
<u>Multivariate Analysis</u>

• Pairplot



This plot helps us to understand the relationship between all the numerical values in the dataset and establish the trends in the dataset.

• Heatmap



- No strong positive correlations are seen in the heatmap.

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 data was encoded in the Jupyter file and its headset is shown below:

	Salary	age	educ	no_young_children	no_older_children	foreign
0	48412.0	30	8.0	1	1	0
1	37207.0	45	8.0	0	1	0
2	58022.0	46	9.0	0	0	0

The below code was written and executed to split the data.

```
from sklearn.model_selection import train_test_split
X_train,X_test,y_train,y_test=train_test_split(X,Y,test_size=0.30,random_state=1)
```

Logistic regression and Linear discriminant analysis were applied to the split data.

```
#LOGISTIC REGRESSION
lgmodel = LogisticRegression()
lgmodel=lgmodel.fit(X_train, y_train)
lgmodel
```

```
LogisticRegression()
```

```
#LDA
clf = LinearDiscriminantAnalysis()
ldamodel=clf.fit(X_train, y_train)
ldamodel
```

```
LinearDiscriminantAnalysis()
```

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.

<u>Logistic Regression</u>

- Classification Report

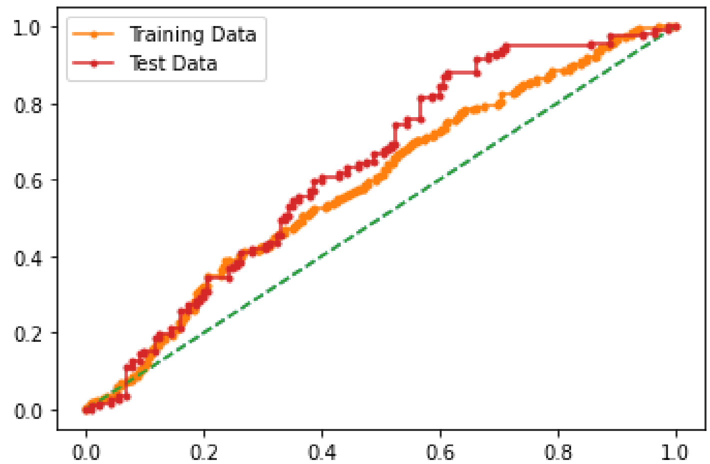
```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.54      0.94      0.68       326
           1       0.49      0.07      0.12       284

    accuracy                           0.53       610
   macro avg       0.51      0.50      0.40       610
weighted avg       0.51      0.53      0.42       610
```

```
Classification Report of the test data:

              precision    recall  f1-score   support

           0       0.56      0.94      0.70       145
           1       0.53      0.09      0.15       117

    accuracy                           0.56       262
   macro avg       0.54      0.51      0.42       262
weighted avg       0.54      0.56      0.45       262
```

- Confusion Matrix

• AUC & ROC

```
AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.591
AUC for the Test Data: 0.633
```



Linear Discriminant Analysis

• Classification Report

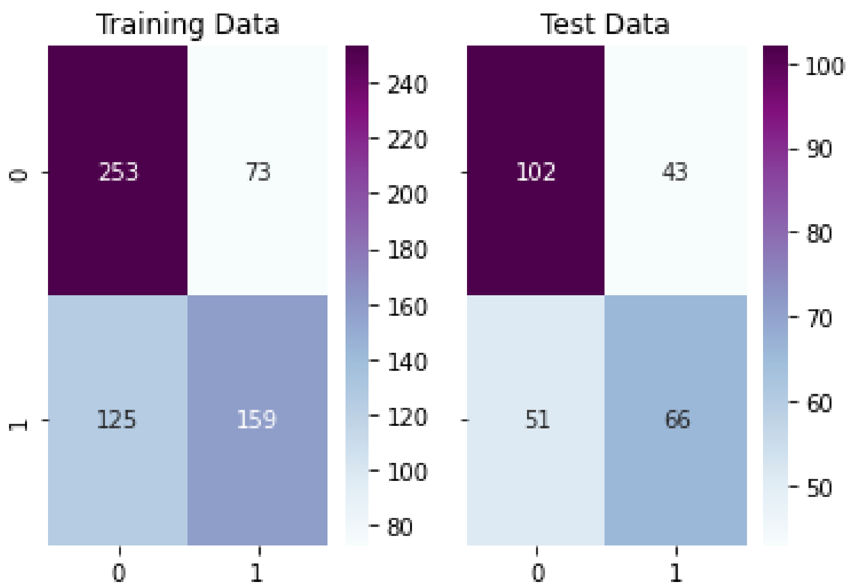
Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.56	0.62	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 of the test data:

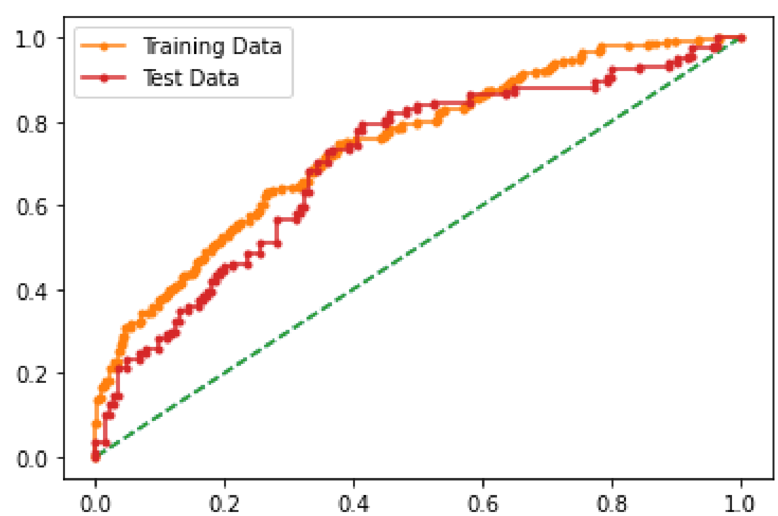
	precision	recall	f1-score	support
0	0.67	0.70	0.68	145
1	0.61	0.56	0.58	117
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

• Confusion Matrix



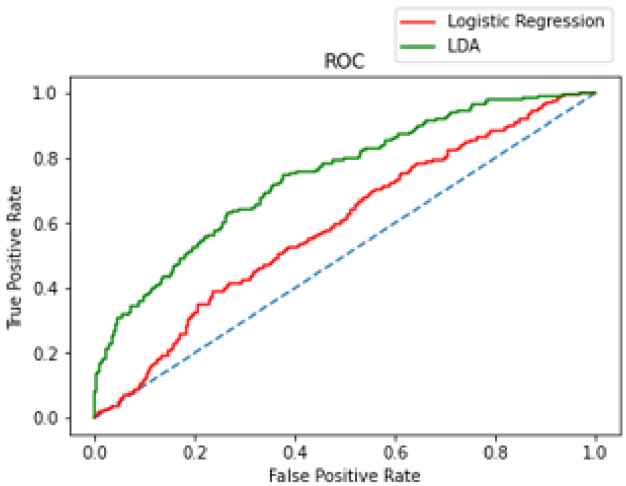
• AUC & ROC

```
AUC and ROC FOR LDA
AUC for the Training Data: 0.744
AUC for the Test Data: 0.704
```

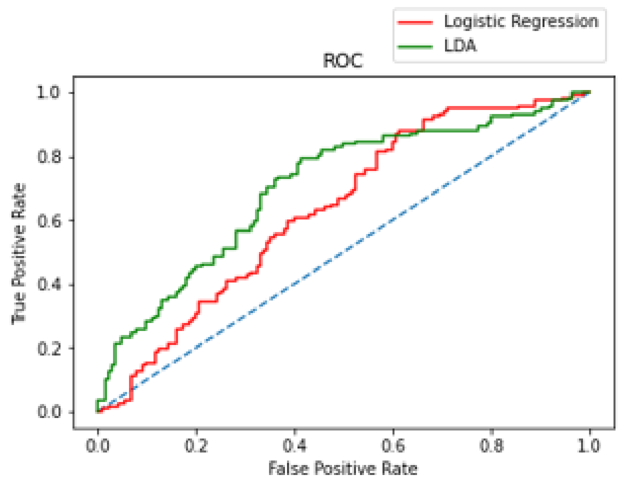


MODEL COMPARISON: Logistic Regression Vs Linear Discriminant Analysis

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test
Accuracy	0.53	0.55	0.68	0.64
AUC	0.59	0.63	0.74	0.70
Recall	0.00	0.00	0.56	0.56
Precision	0.00	0.00	0.69	0.61
F1 Score	0.00	0.00	0.62	0.58



Train data-ROC curve



Test data-ROC curve

From the table above, it is seen that the LDA model has higher accuracy, AUC, recall value, precision and F1 score. Therefore, LDA model is chosen.

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

- The greatest number of people who are opting in for the package has a salary of range between 30,000 to 40,000 which says that the package is of average price with basic to medium level facilities. Therefore, addition of luxury facilities like can attract more customers of higher income group which will increase sales.

- More number of customers who have lower number of young children have opted for holiday package when compared to those having more number of young children. This problem can be overcome by including places of attraction for younger kids to enjoy.

- Many people with high salary are not opting for holiday packages. One of the reasons can be that their budget for vacation is more and they expect more or something different from the offered package. This can be done by providing luxurious facilities or by also giving them the option of customizing a holiday package according to their requirements.

THE END