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.

Exploratory Data Analysis

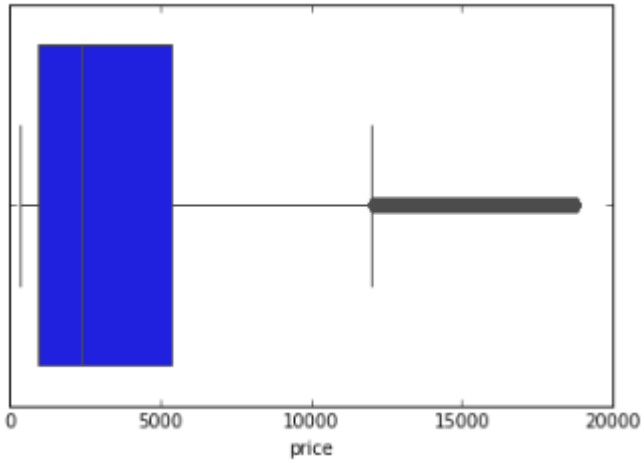
- Dataset contains 26967 rows and 11 columns
- **Columns Names** : Unnamed: 0 ,carat, cut, color, clarity, depth, table, x, y, z, price
- "Unnamed: 0" is not a significant column. Therefore, we drop this column
- After dropping Unnamed: 0, three are of type "object" (cut, color and clarity) , six float and one integer.
- Variable "Depth" has 697 null values.

Descriptive Statistics for the dataset

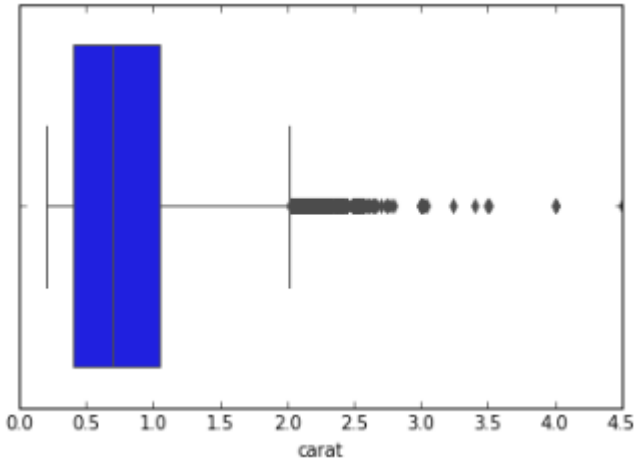
	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	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

Univariate Analysis

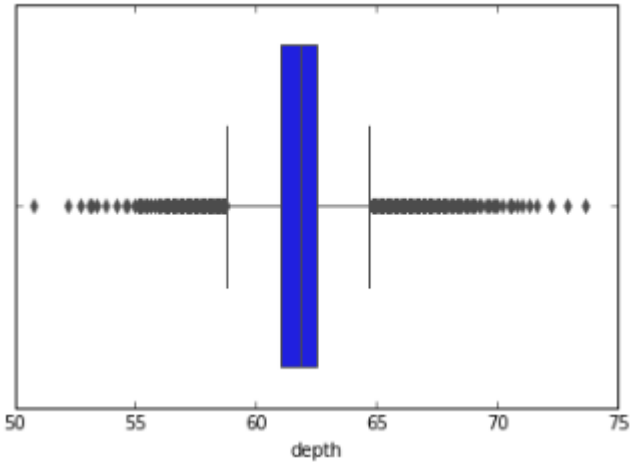
Price



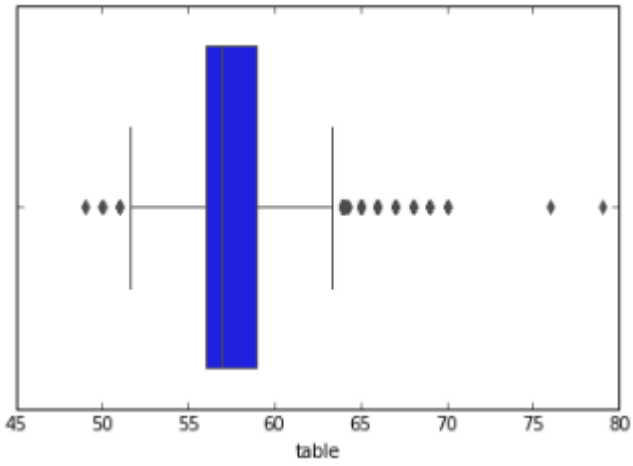
Carat



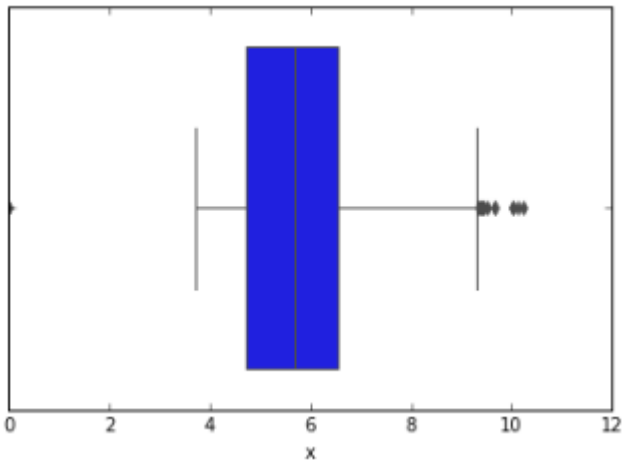
Depth



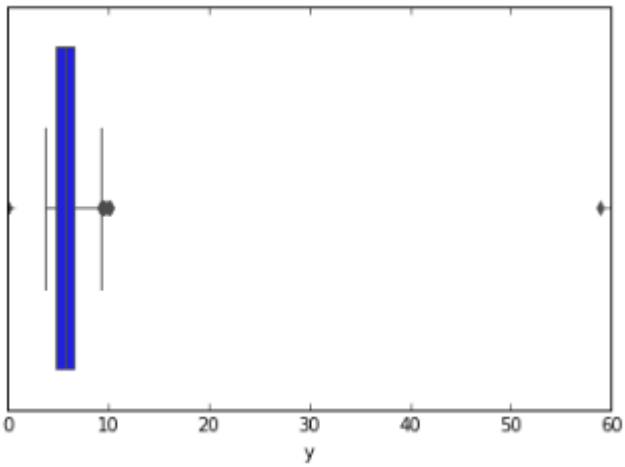
Table



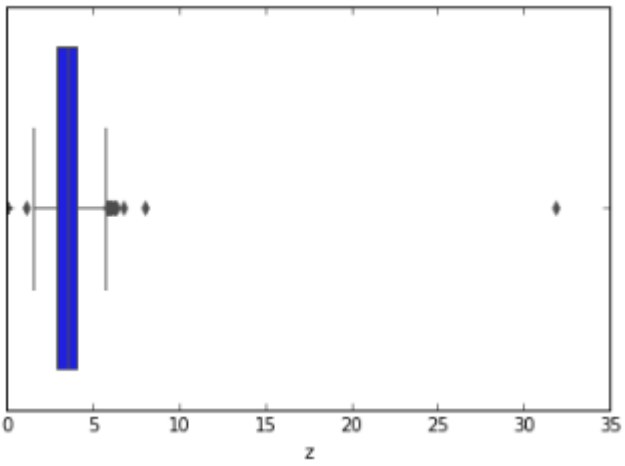
X



Y



Z



Outliers are present and need to be treated.

Values and Distribution of the Categorical Data

Cut

```
Ideal        10816
Premium       6893
Very Good     6030
Good          2439
Fair           780
Name: cut, dtype: int64
```

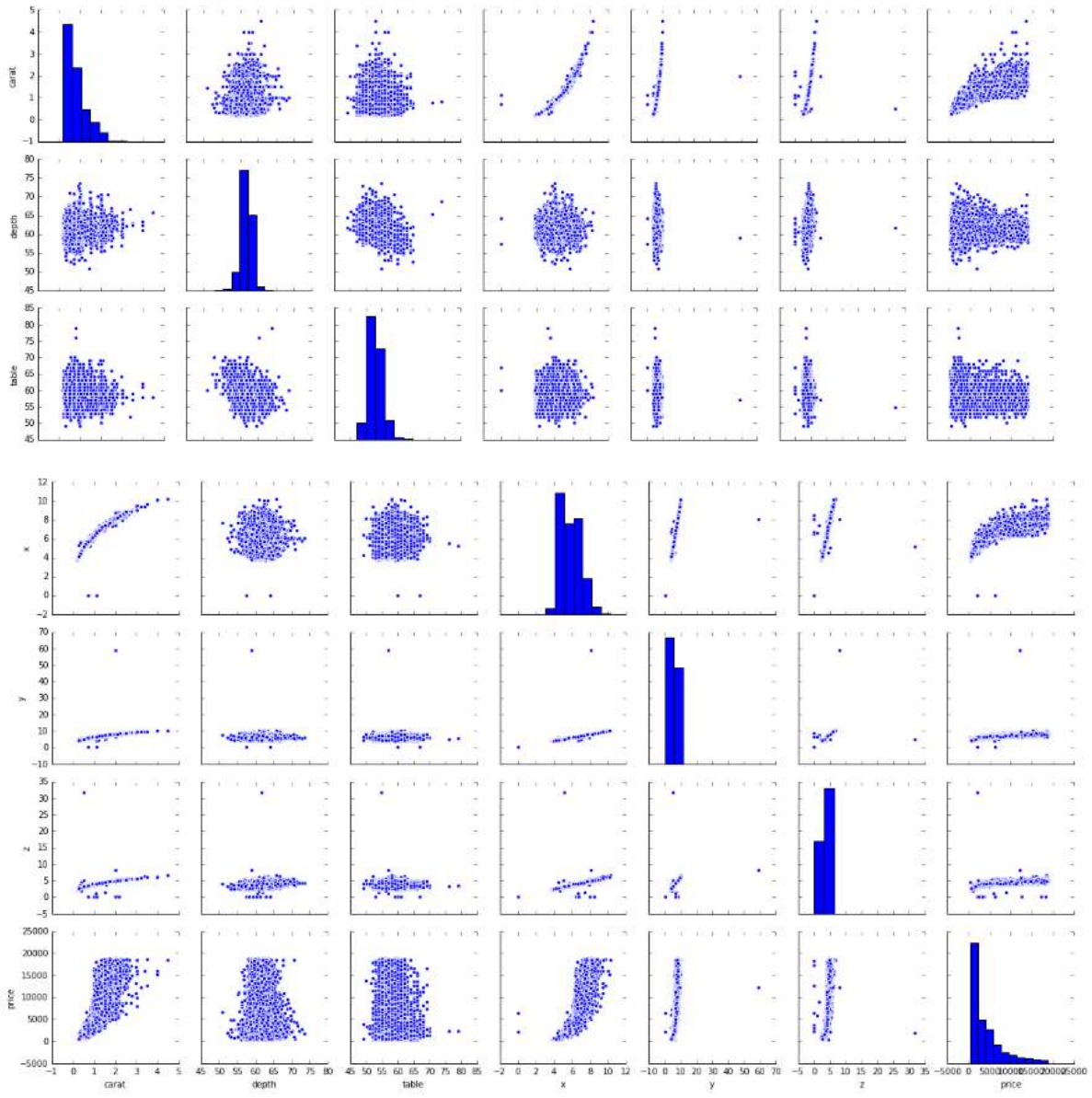
Color

```
G    5658
E    4917
F    4727
H    4098
D    3344
I    2771
J    1443
Name: color, dtype: int64
```

Clarity

```
SI1     6570
VS2     6098
SI2     4571
VS1     4092
VVS2    2531
VVS1    1839
IF       894
I1       363
Name: clarity, dtype: int64
```

Bi-Variate Analysis


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

Correlation

Price is highly correlated with Carat, x, y and z
Price is negatively correlated with Depth
Carat is highly correlated with x, y and z. Also, X, Y and Z are highly correlated. Therefore, we can drop x, y and z for our model

1.2 Impute null value 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?

"Depth" Variable has 697 null values. We have imputed these null values with the median.

x and y have three records which have value as 0.

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
6215	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

There are 9 records where z has value 0.

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

But, because we know that x, y and z are highly correlated with carat and we would not need these values in our model building. We can ignore zero values as we will drop these three columns itself.

Scaling is useful in this case as the variables are in different units and scaling will bring uniformity to the data for better understanding. However, it is not mandatory.

In our case, I have scaled the independent variables.

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 Rsquare, RMSE.

Since we know that the variables cut, color and clarity are ordinal in nature. Therefore, the below codes are assigned accordingly.

For Cut:

Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Therefore, we will assign the following codes:

Fair: 1
Good: 2
Very Good: 3
Premium: 4
Ideal: 5

For Color:
D being the best and J the worst

J: 1
I: 2
H: 3
G: 4
F: 5
E: 6
D: 7

For Clarity:

In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3

I3 = 1, I2 = 2, I1 = 3, SI2 = 4, SI1 = 5, VS2 = 6, VS1=7, VVS2 = 8, VVS1 = 9, IF = 10, FL =11

Depth and Table do not contribute much to the model. Hence, we have dropped these variables from the model.

The coefficient for carat is 3638.462432876167

The coefficient for cut is 133.03479539540825

The coefficient for color is 464.5308411435908

The coefficient for clarity is 740.4260286453052

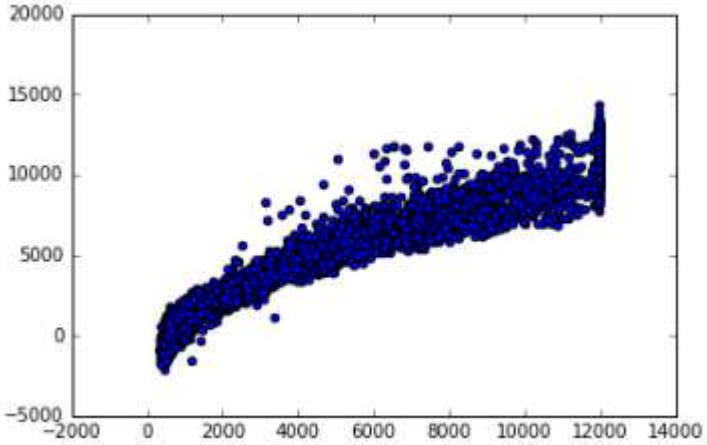
The coefficient for depth is -9.313134040767547

The **intercept** for the model is 3723.632350735075

R square on training data: 0.9291767255027183

R square on testing data: 0.9296181762468217

RMSE: 914.5294389960341



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

Final Conclusion

From the above model, we can conclude that Price is highly dependent on the below five attributes:

1. **Carat**
2. **Clarity**
3. **Color**
4. **Cut**
5. **Depth**

Price = 3723.632350735075 + 3638.462432876167*carat + 740.4260286453052*clarity + 464.5308411435908***color + 133.03479539540825*cut + (-9.313134040767547)*depth**

Price of the cubic zirconia will be more for higher value of "carat", better clarity, color and cut.

Value of Price decreases with the increase in the value of "Depth".

For better profit shares, the company must focus on producing stones with higher carat value and maximum clarity, color and cut.

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.

Exploratory Data Analysis

The dataset has 872 rows and 8 columns

Column Names: Unnamed: 0, Holliday_Package, Salary, age, educ, no_young_children, no_older_children, foreign

"Unnamed: 0" is of no significance as it is a serial number. We will drop this column

No Null values are present

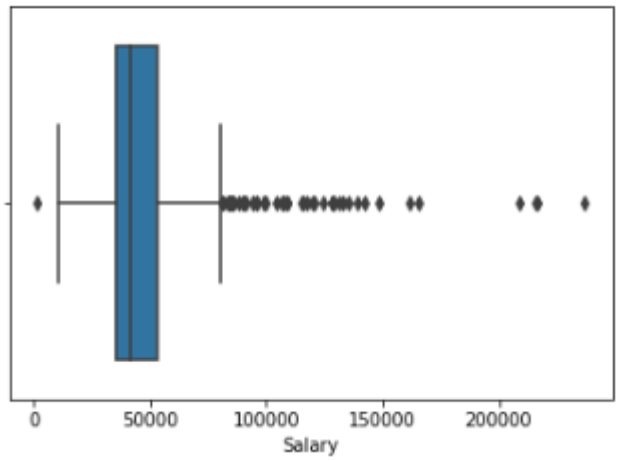
Descriptive Statistics for the Dataset

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

From the above statistics, we can say that there is skewness in some variables and outliers are present. Let's plot the histogram to get more insights

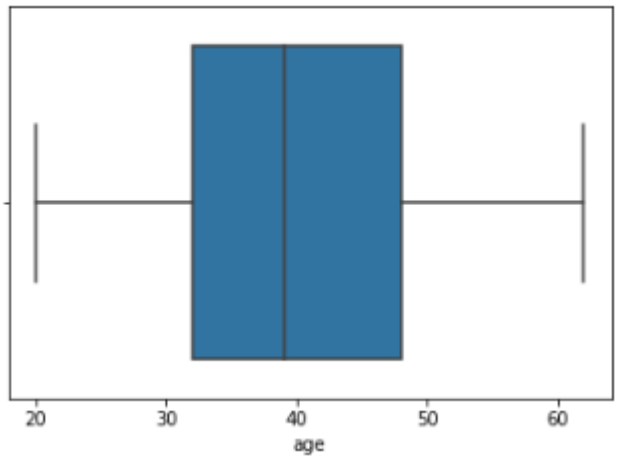
Univariate Analysis

Salary



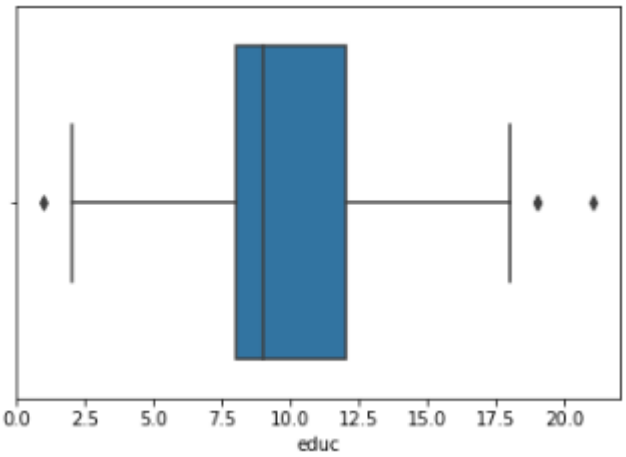
Salary is right skewed and Outliers are present that need to be treated

Age



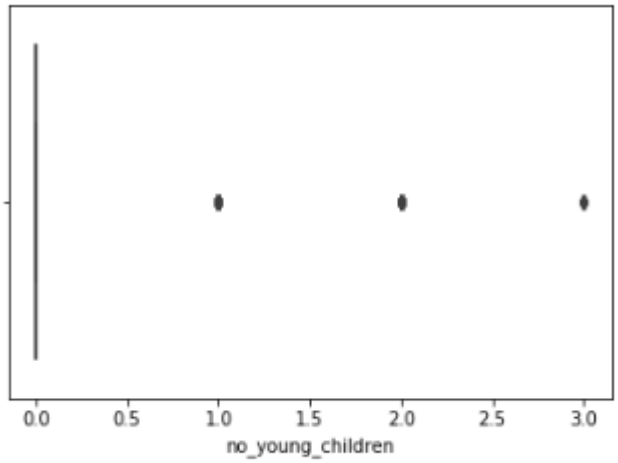
No outliers present and data is approximately normal

Educ



Very few outliers present which can be ignored and data is right skewed

No_young_children



Number of young children has only 4 values 0, 1, 2 and 3. Therefore this can be treated a categorical variable.

Same is with number of older children. It can be treated as a categorical variable and has 7 values (0, 1, 2, 3, 4, 5, and 6)

These two variables are converted to Categorical.

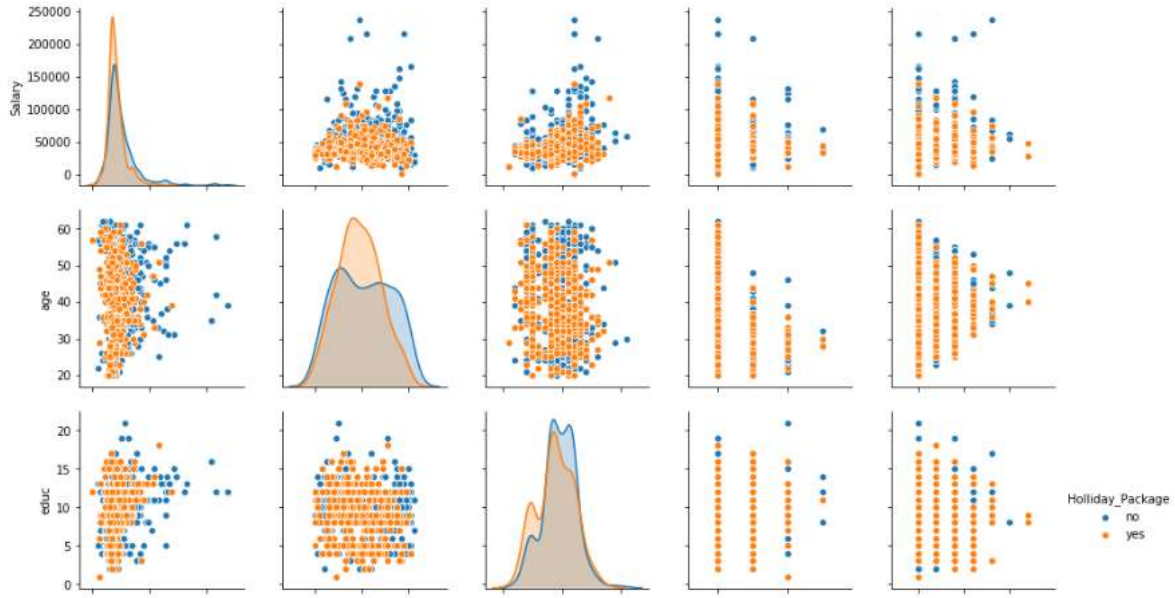
Values of other categorical Variable:

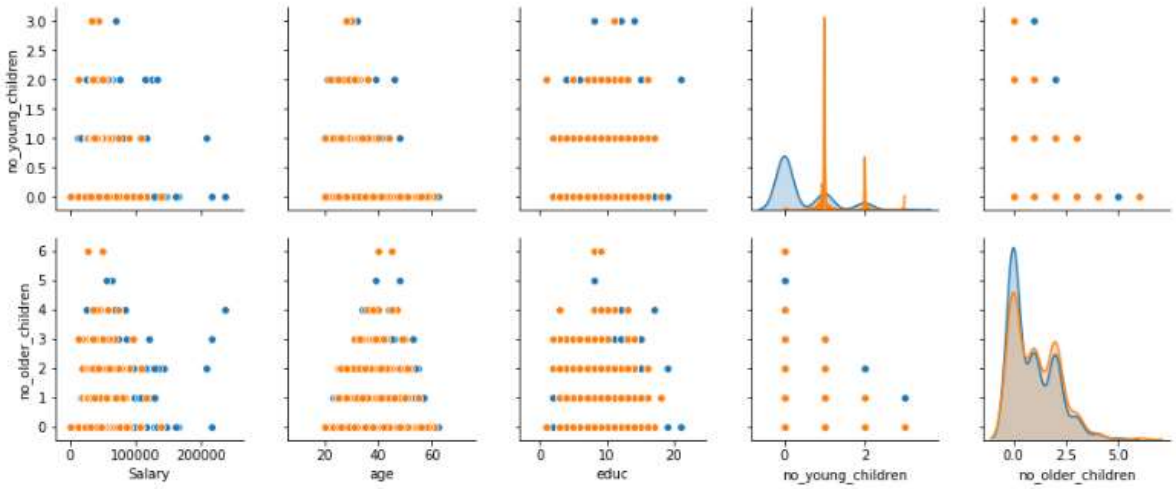
```
Holliday_Package
no     471
yes    401
Name: Holliday_Package, dtype: int64


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

Proportion of Target Variable (Claimed)

No	0.540138
Yes	0.459862

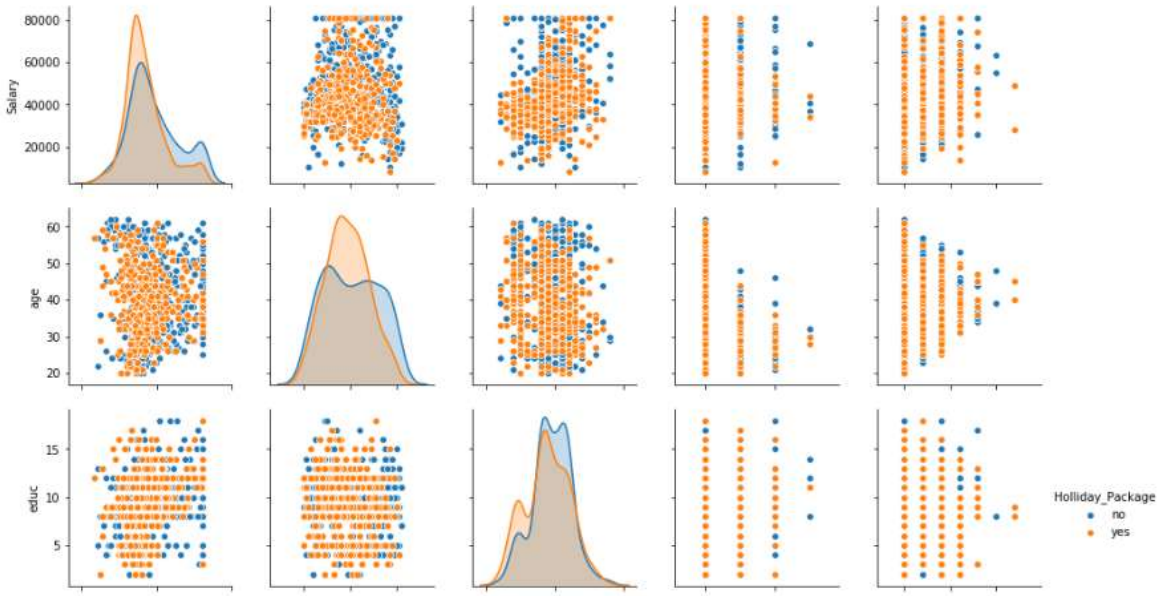
Bi-Variate Analysis

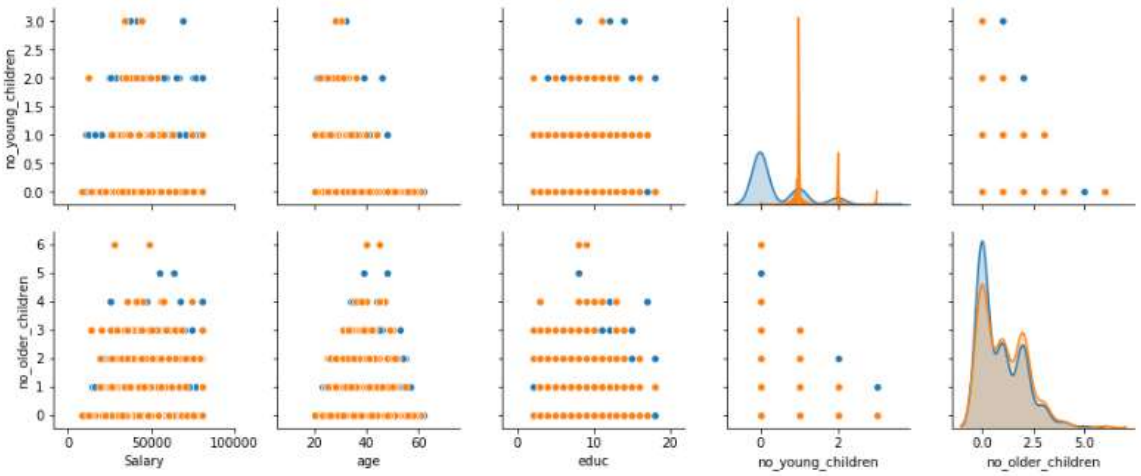




From the pair plot we can see that "age" and "no_young_children" are good predictors for Holliday_Package but the rest show similar graph for different values of “Holiday_Package” (Target Variable).

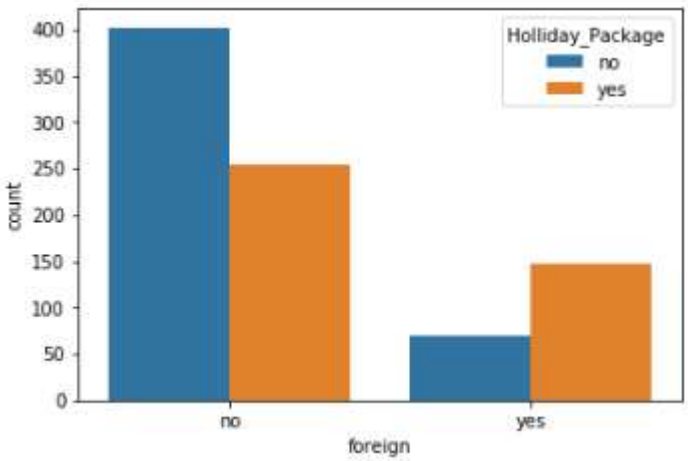
After Treating the outliers, we get the below plot





After treating the outliers, we can see that “Salary” can be considered as a differentiator.

Relation between Foreign and Holiday_Package



Correlation

There is no correlation between the numerical attributes

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).

Encoding the data with String Values.

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


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

Model is built using variables: 'age', 'no_young_children', 'no_older_children', 'foreign'

If we include "educ" the difference between accuracy of train and test data increases by 5 % which can be considered as an example of overfitting

Also, "Salary" reduces the performance of the model significantly.

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 M odel: Compare Both the models and write inference which model is best/optimized.

Performance Evaluation

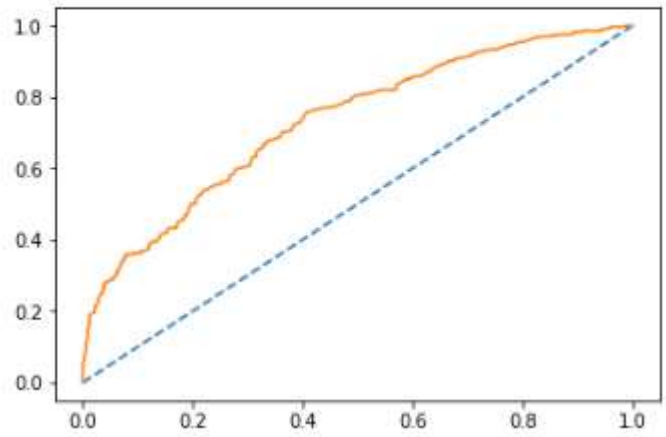
Logistic Regression

Accuracy Train Data: 0.6655737704918033
Accuracy Test Data: 0.6564885496183206

Train Data:

AUC: 0.731

```
AUC: 0.731
[<matplotlib.lines.Line2D at 0x1a8a01650b8>]
```



Confusion Matrix

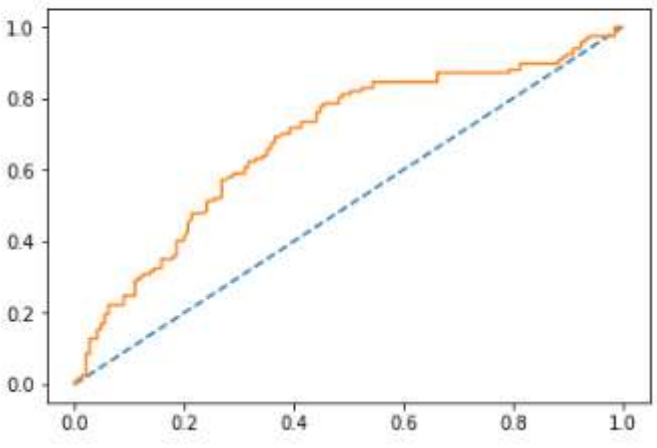
```
array([[256,  70],
       [134, 150]], dtype=int64)
```

Classification Report

	precision	recall	f1-score	support
0	0.66	0.79	0.72	326
1	0.68	0.53	0.60	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Test Data:
AUC: 0.627

```
AUC: 0.627
[<matplotlib.lines.Line2D at 0x2284951a198>]
```



Confusion Matrix

```
array([[107,  38],
       [ 52,  65]], dtype=int64)
```

Classification Report

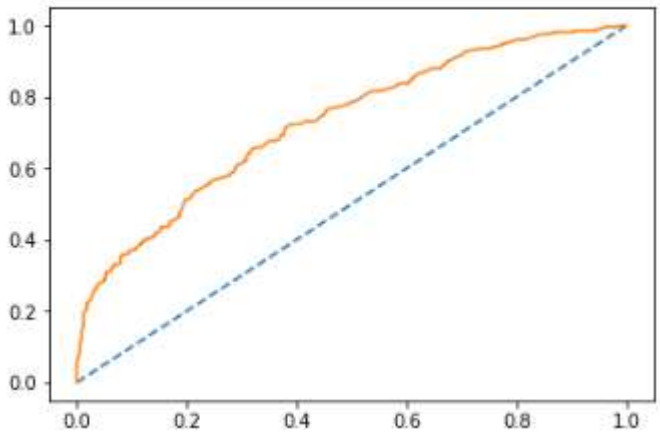
	precision	recall	f1-score	support
0	0.67	0.74	0.70	145
1	0.63	0.56	0.59	117
accuracy			0.66	262
macro avg	0.65	0.65	0.65	262
weighted avg	0.65	0.66	0.65	262

LDA

Train Data:

Accuracy: 0.6672131147540984
AUC: 0.730

```
[<matplotlib.lines.Line2D at 0x1a8a447b518>]
```



Confusion Matrix:

```
array([[254,  72],
       [131, 153]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.66	0.78	0.71	326
1	0.68	0.54	0.60	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Test Data:

Accuracy: 0.6603053435114504

Confusion Matrix:

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

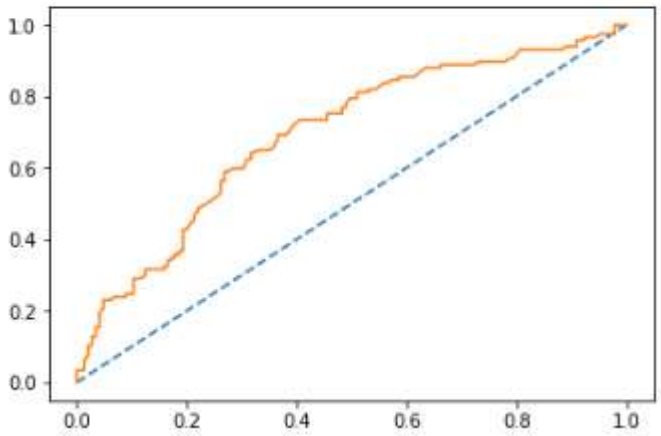
Classification Report:

	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: 0.694

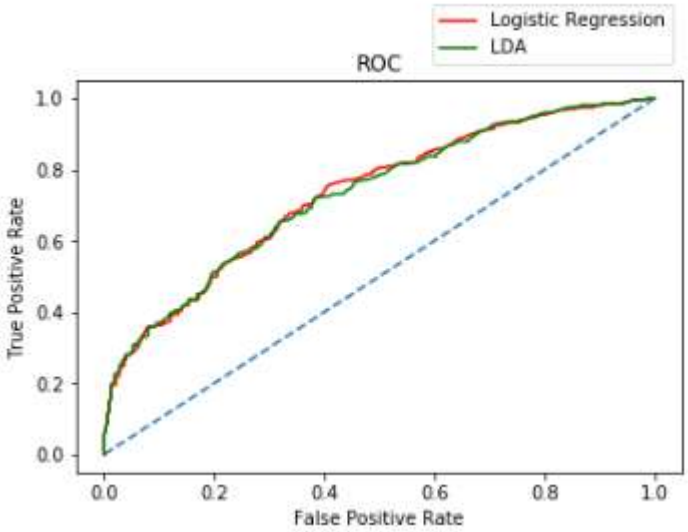
```
AUC: 0.694
```

```
[<matplotlib.lines.Line2D at 0x1a8a3cfc898>]
```

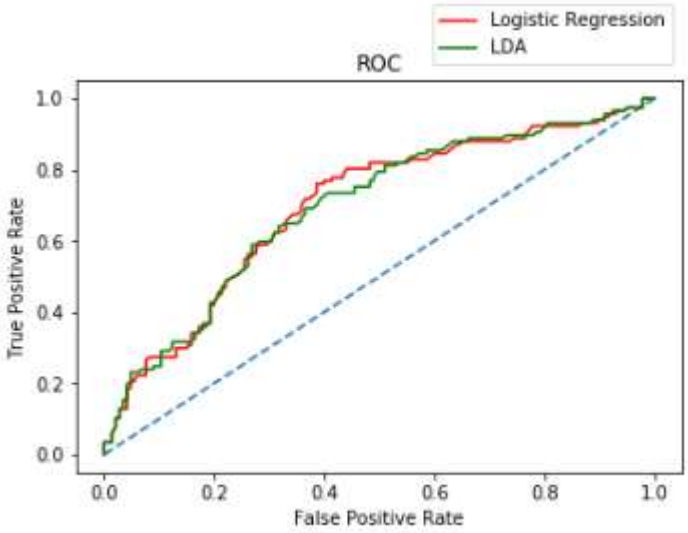


Comparing Logistic Regression and LDA

TRAIN DATA



TEST DATA



Both train and test data show similar behaviour. We can choose any model out of the two.

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

Inference:

The model is not very good for making prediction. More data should be gathered for better prediction.

Especially data for varied salaries should be obtained in order to improve accuracy for prediction.

If we include "educ" the difference between accuracy of train and test data increases by 5 % which can be considered as an example of overfitting

Also, "Salary" reduces the performance of the model significantly.

From the model coefficients we can infer that:

- Young foreign nationals are more likely to buy a package.
- People with older kids are more likely to buy a package other than the ones with kids younger than 7 years old.

The coefficient for foreign is 1.4101597858838493

The coefficient for age is -0.04851604659830628

The coefficient for no_young_children is -1.3268405454803827

The coefficient for no_older_children is -0.06010949287673012