PREDICTIVE MODELLING

DINESH YADAV MEKALA

Table of Contents

Table of Figures

Tables list

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, duplicate values). Perform Univariate and Bivariate Analysis.

DataFrame Info

Table 1 DF Info

#	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

We have float, int, object data types in the data

Table 2 DF Head

	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

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
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

Table 3 DF Description

	Unnamed: 0	carat	depth	table	x	y	z	price
count	26967.000000	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	13484.000000	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	7784.846691	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	1.000000	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	6742.500000	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000

	Unnamed: 0	carat	depth	table	x	y	z	price
max	26967.000000	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

- We have both categorical and continuous data. In categorical we have cut, colour and clarity.
- In continuous data we have carat, depth, table, x, y, z, price.
- Price will be the target variable.

Table 4 Null Values

carat	0
cut	0
color	0
clarity	0
depth	697
table	0
x	0
y	0
z	0
price	0

Unique Values

Table 5 ***Unique Values for cut***

CUT : 5

Fair	781
Good	2441
Very Good	6030
Premium	6899
Ideal	10816

We have 5 cuts and Ideal seems to be the most preferred cut.

Table 6 ***Unique Values for color***

COLOR : 7

J	1443
I	2771
D	3344
H	4102
F	4729
E	4917
G	5661

We have 7 colors.

Table 7 ***Unique Values for clarity***

CLARITY : 8

I1	365
IF	894
VVS1	1839
VVS2	2531
VS1	4093
SI2	4575
VS2	6099
SI1	6571

We have 8 types of clarity

Univariate Analysis

Carat

Figure 1 Boxplot Carat

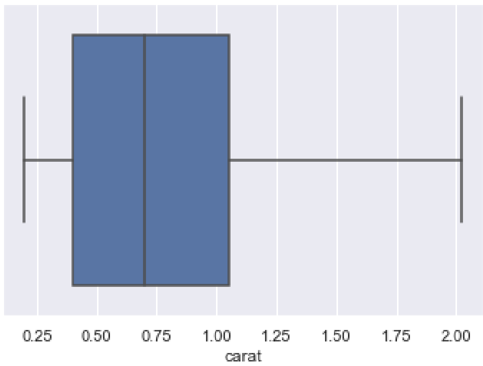
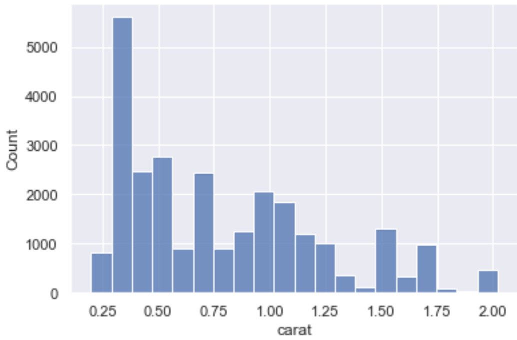


Figure 2 Histogram Carat



The distribution of data in carat seems to be positively skewed and there are multiple peak points in the distribution. In the range of 0 to 1 maximum of the data lies.

Depth

Figure 3 Boxplot Depth

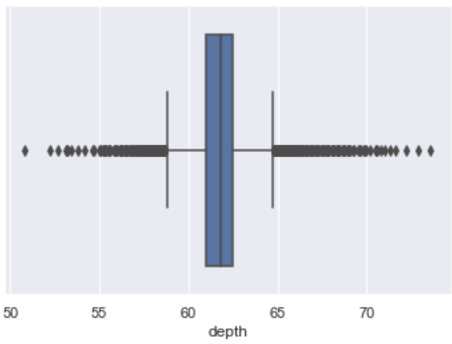
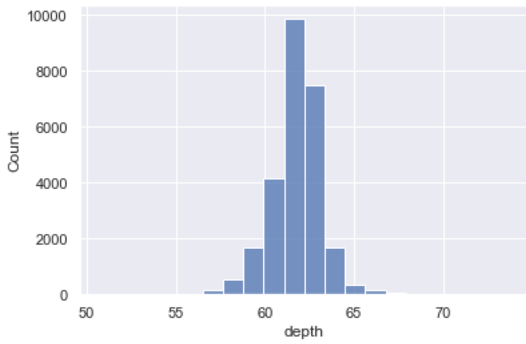


Figure 4 Histogram Depth



Distribution of depth seems to be normal. Depth ranges from 55 to 65.
The boxplot has many outliers.

Table

Figure 5 Boxplot Table

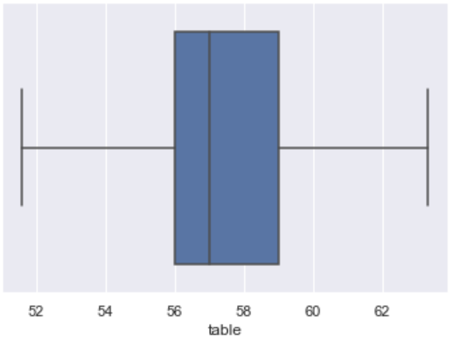
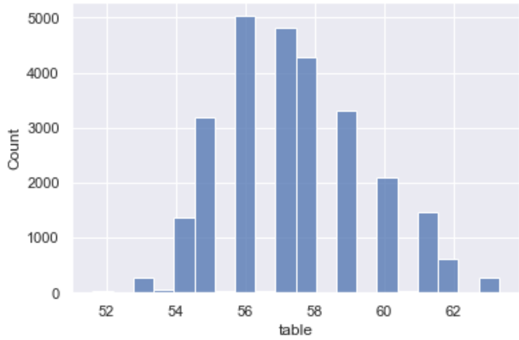


Figure 6 Histogram Table



The distribution seems to be positively skewed. Data distribution is between 55 to 65 in table. It has many outliers in it.

X

Figure 7 Boxplot X

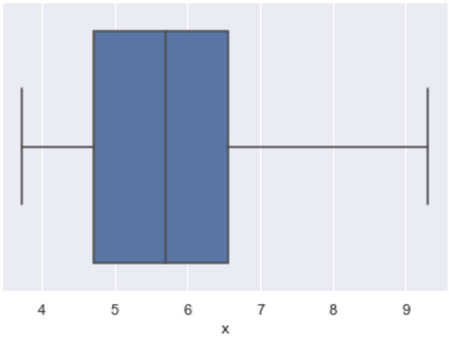
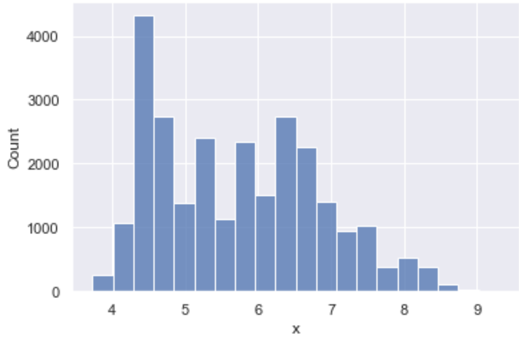


Figure 8 Histogram X



Distribution of X is positively skewed and ranges from 4 to 8. Boxplot of X has many outliers.

Y

Figure 9 Boxplot Y

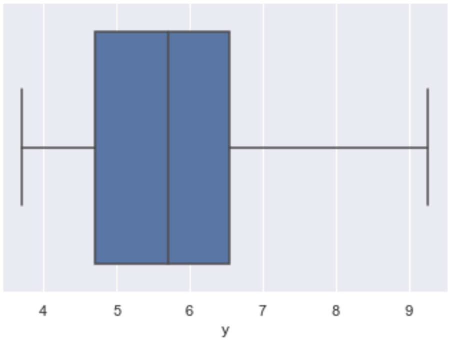
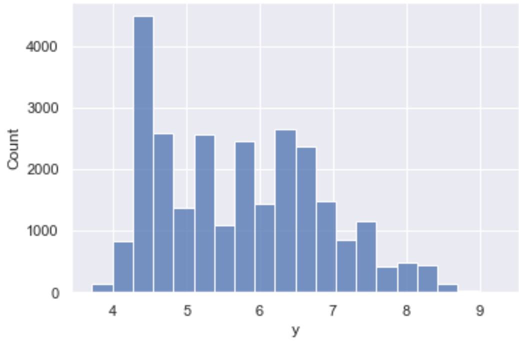


Figure 10 Histogram Y



Distribution of Y is positively skewed and ranges from 4 to 7. Boxplot of Y has outliers.

Z

Figure 11Boxplot Z

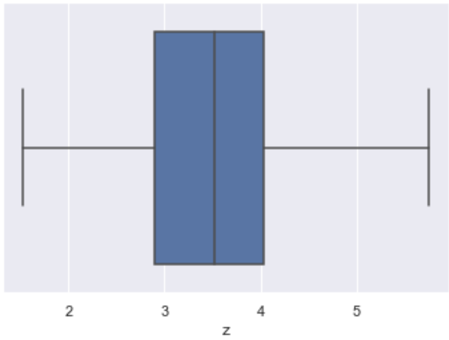
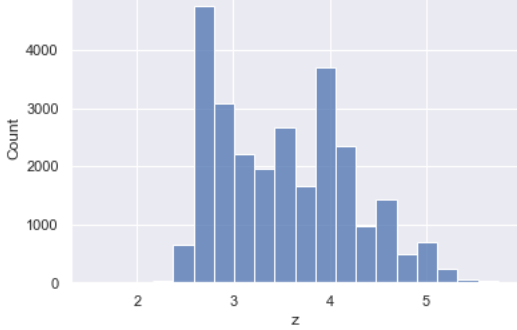


Figure 12 Histogram Z



Distribution of Z is positively skewed, skewness may be due to diamonds are always made in specific shape . Boxplot has outliers,

Price

Figure 13Boxplot Price

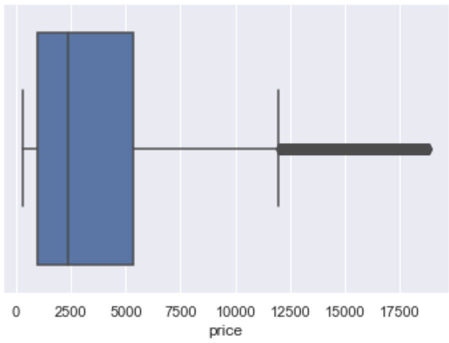
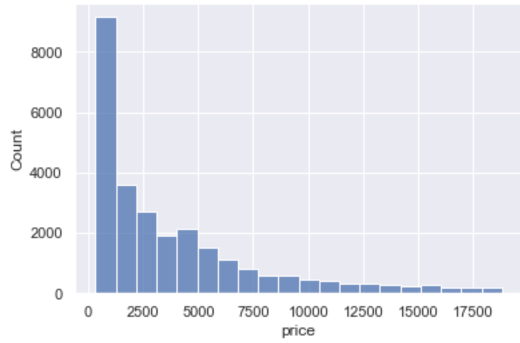


Figure 14 Histogram Price

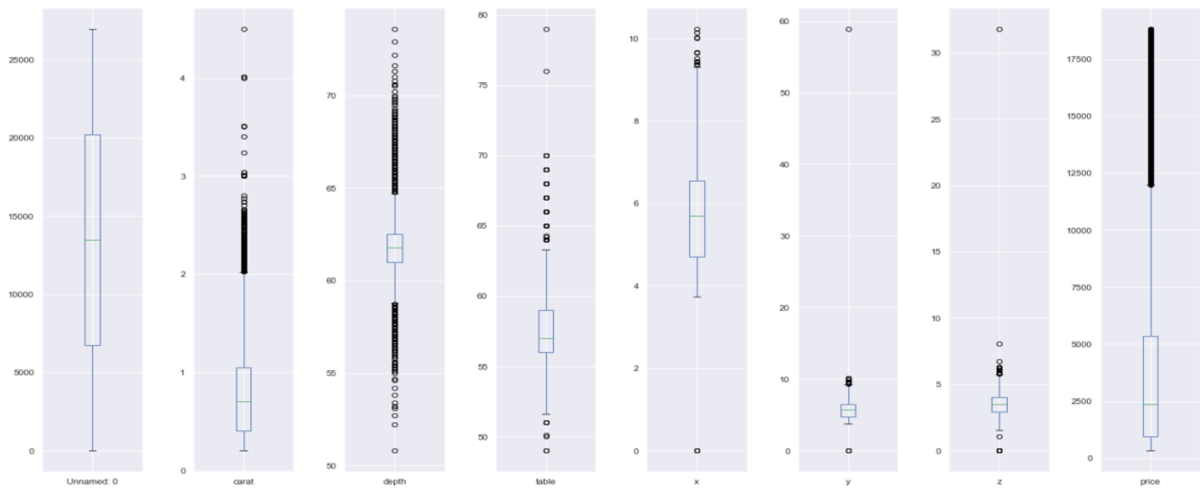


Distribution of Price is positively skewed, and distribution is in the range of 100 to 8000. Boxplot has outliers.

Outliers

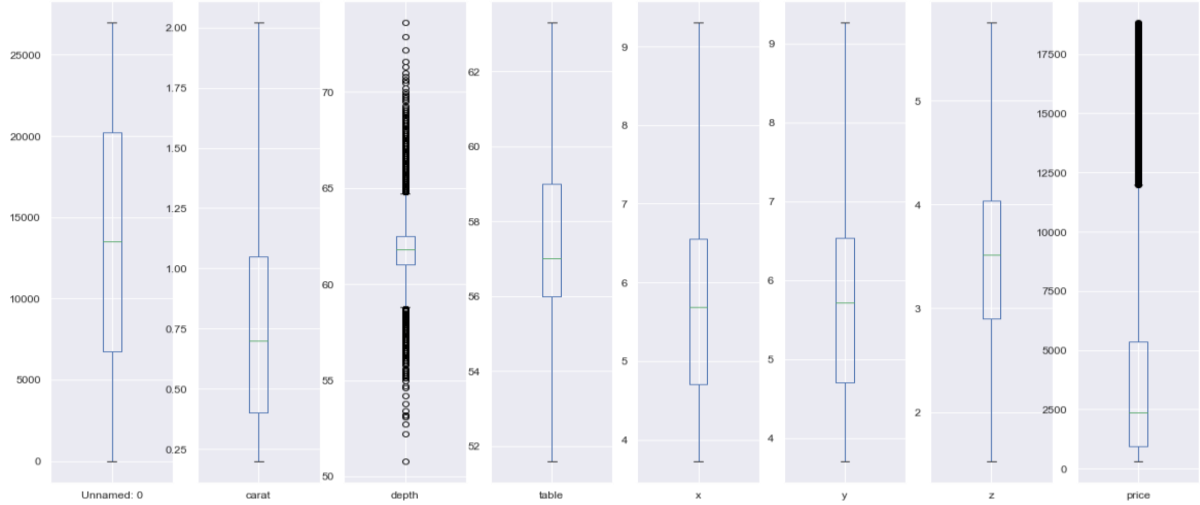
Before treating outliers

Figure 15 Before treating outliers



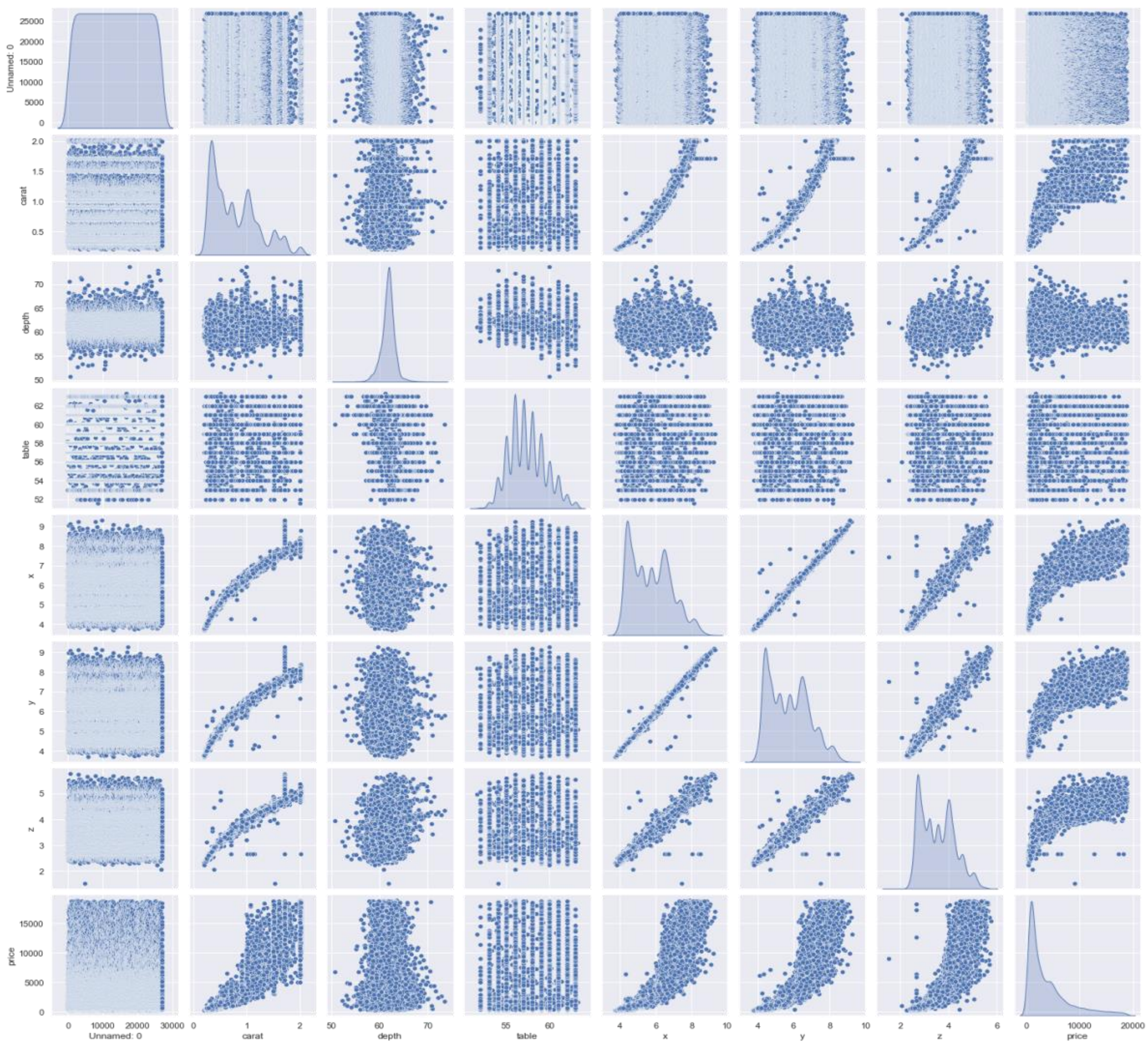
After treating outliers

Figure 16 After treating outliers



Multivariate Analysis

Figure 17 Multivariate Analysis



Heatmap

Figure 18 Heatmap

The matrix shows the presence of multi collinearity in the dataset.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

Table 8 Values equal to zero

carat	False
cut	False
color	False
clarity	False
depth	False
table	False
x	False
y	False
z	False
price	False

Table 9 Null Values before imputing

carat	0
cut	0
color	0
clarity	0
depth	697
table	0

x	0
y	0
z	0
price	0

Table 10 DataFrame head after imputing

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.3	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	2.7	984
2	3	0.9	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	4.8	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 11 Null Values after imputing

carat	0
cut	0
color	0
clarity	0
depth	0
table	0
x	0
y	0
z	0
price	0

Table 12 Description after imputing

	Unnamed: 0	carat	depth	table	x	y	z	price
count	26967.000000	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	13484.000000	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115

	Unnamed: 0	carat	depth	table	x	y	z	price
std	7784.846691	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	1.000000	0.200000	50.800000	51.600000	3.730000	3.710000	1.530000	326.000000
25%	6742.500000	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	26967.000000	2.020000	73.600000	63.300000	9.300000	9.260000	5.750000	18818.000000

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.

Changing the data types of cut, color, and clarity

Table 13 cut data type changed to int

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.3	5	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	4	G	IF	60.8	58.0	4.42	4.46	2.7	984
2	3	0.9	3	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
3	4	0.42	5	F	VS1	61.6	56.0	4.82	4.8	2.96	1082
4	5	0.31	5	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 14 color data type changed to int

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.3	5	6	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	4	4	IF	60.8	58.0	4.42	4.46	2.7	984
2	3	0.9	3	6	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	5	5	VS1	61.6	56.0	4.82	4.8	2.96	1082
4	5	0.31	5	5	VVS1	60.4	59.0	4.35	4.43	2.65	779

Table 15 clarity data type changed to float

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	5	6	-1.0	62.1	58.0	4.27	4.29	2.66	499.0
1	2	0.33	4	4	-1.0	60.8	58.0	4.42	4.46	2.70	984.0
2	3	0.90	3	6	-1.0	62.2	60.0	6.04	6.12	3.78	6289.0
3	4	0.42	5	5	-1.0	61.6	56.0	4.82	4.80	2.96	1082.0
4	5	0.31	5	5	-1.0	60.4	59.0	4.35	4.43	2.65	779.0

Table 16 DF head after scaling

	count	mean	std	min	25%	50%	75%	max
carat	26967.0	-1.186263e-16	1.000019	-1.319405	-0.868988	-0.193364	0.594864	2.779383
cut	26967.0	4.798569e-16	1.000019	-2.613667	-0.817058	0.081246	0.979550	0.979550
color	26967.0	1.487790e-16	1.000019	-1.989430	-0.817070	-0.230890	0.941470	1.527650
clarity	26967.0	6.182452e-17	1.000019	-1.853722	-0.639402	-0.032241	0.574919	2.396400
depth	26967.0	-4.311599e-16	1.000019	-7.850470	-0.467179	0.106281	0.536376	8.493127
table	26967.0	-7.827470e-16	1.000019	-2.778656	-0.673506	-0.195062	0.761824	2.819130
x	26967.0	-2.734161e-16	1.000019	-1.777884	-0.906476	-0.035068	0.729637	3.174915
y	26967.0	-2.663514e-16	1.000019	-1.810303	-0.914705	-0.019107	0.724240	3.160267
z	26967.0	-7.779919e-16	1.000019	-2.889003	-0.917248	-0.024921	0.723482	3.184577
price	26967.0	-2.910285e-17	1.000019	-0.897815	-0.744018	-0.388720	0.352933	3.696710

Coefficient for the following columns are

Table 17 Table of coefficients

`The coefficient for carat`	`1.2801213328224796`

The coefficient for cut	0.044061306493186514
The coefficient for color	0.1233528534141017
The coefficient for clarity	0.19240675413742578
The coefficient for depth	-0.003832957799618385
The coefficient for table	-0.015416741736581297
The coefficient for x	-0.5361037488818746
The coefficient for y	0.44081340476733166
The coefficient for z	-0.16420841159037086

The intercept for our model is 0.0015672526389941405
Regression model score for train data is 0.8886993336877839
Regression model score for test data is 0.883659588050507
RMSE for training data is 0.33336543663305496
RMSE for training data is 0.34168755937542916

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

- Carat is the dominant factor in deciding the price of diamond. Higher the Carat higher the price of diamond.
- Carat is measure of weight which has direct correlation with physical dimensions (x,y,z).
- Diamond with clarify IF, and colour D has higher price.
- Clarity VVS1, VVS2, VS1, VS2 and colour E, F, G also have positive effect on price of the diamond.
- In terms of cut, Ideal, Premium Very Good would fetch better price.
- It advisable to avoid diamonds of cut ‘Fair’, & Good. Regarding Colour J, H and J will have less price, clarity I1, SI2 and SI1 will have lower price and should be avoided.
- Using these parameter diamonds of higher price can be selected and avoid lower price
- for better marketability and profit.

Problem 2

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.

Table 18 Dataframe head

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

Table 19 Data info

#	Column	Non-Null Count	Dtype
0	Unnamed: 0	872 non-null	int64
1	Holliday_Package	872 non-null	object
2	Salary	872 non-null	int64
3	age	872 non-null	int64
4	educ	872 non-null	int64
5	no_young_children	872 non-null	int64
6	no_older_children	872 non-null	int64
7	foreign	872 non-null	object

- No null values in the dataset,
- We have integer and object data

Table 20 Data Frame Description

	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

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
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

- We have integer and continuous data,
- Holiday package is our target variable
- Salary, age, educ and number young children, number older children of employee have the went to foreign, these are the attributes we have to cross examine and help the company predict weather the person will opt for holiday package or not.

Table 21 Null Values

Holliday_Package	0
Salary	0
age	0
educ	0
no_young_children	0
no_older_children	0
foreign	0

Univariate Analysis

Salary

Figure 19

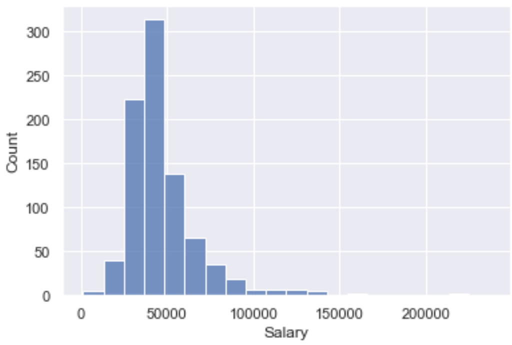
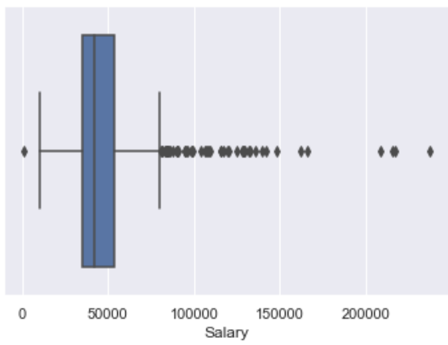


Figure 20



Distribution of salary seems to be normal. The boxplot has many outliers.

Age

Figure 21

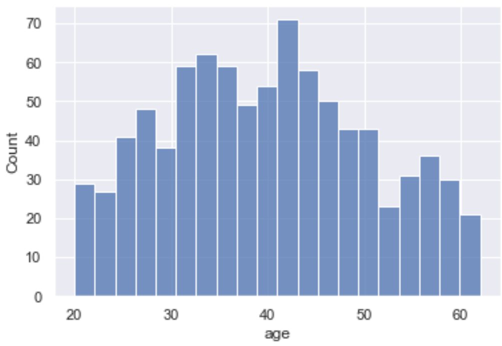
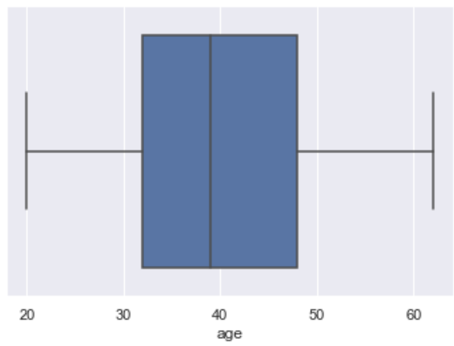


Figure 22



Distribution appears to be normal and range is between 30 to 50.

Educ

Figure 23

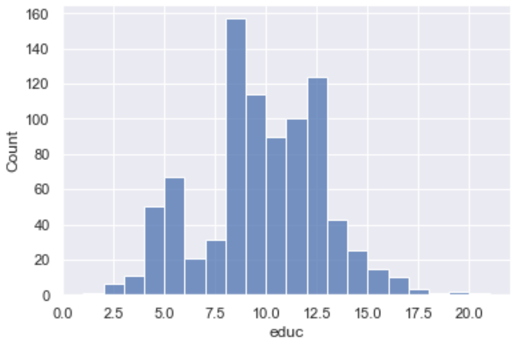
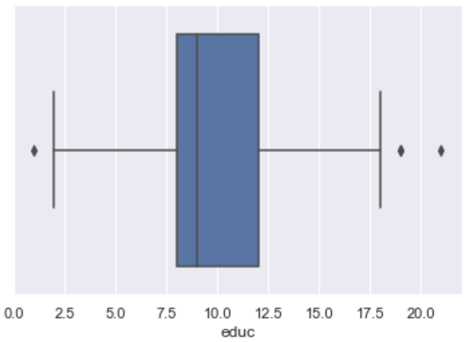


Figure 24



Distribution appears to be normal. . The boxplot has outliers

No young children

Figure 25

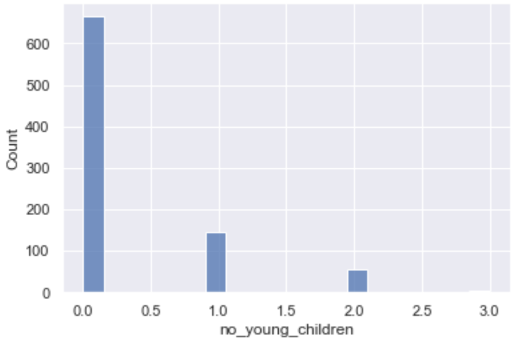
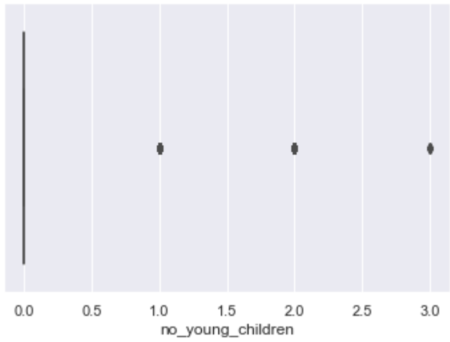


Figure 26



Distribution is right skewed and the boxplot has outliers.

No older children

Figure 27

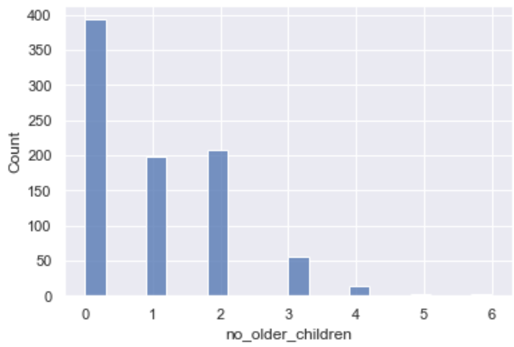
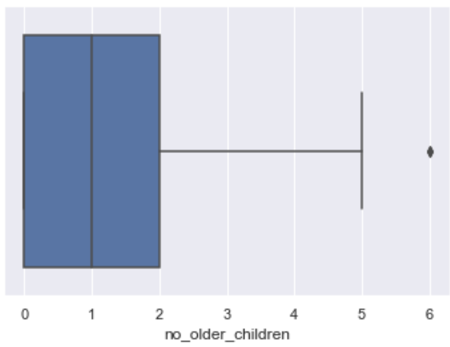


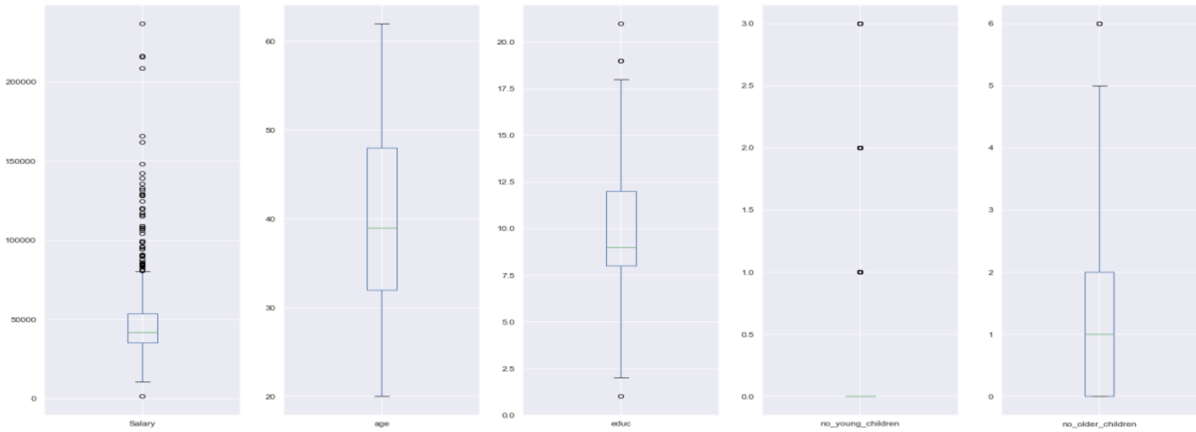
Figure 28



Distribution is right skewed and the boxplot has outliers.

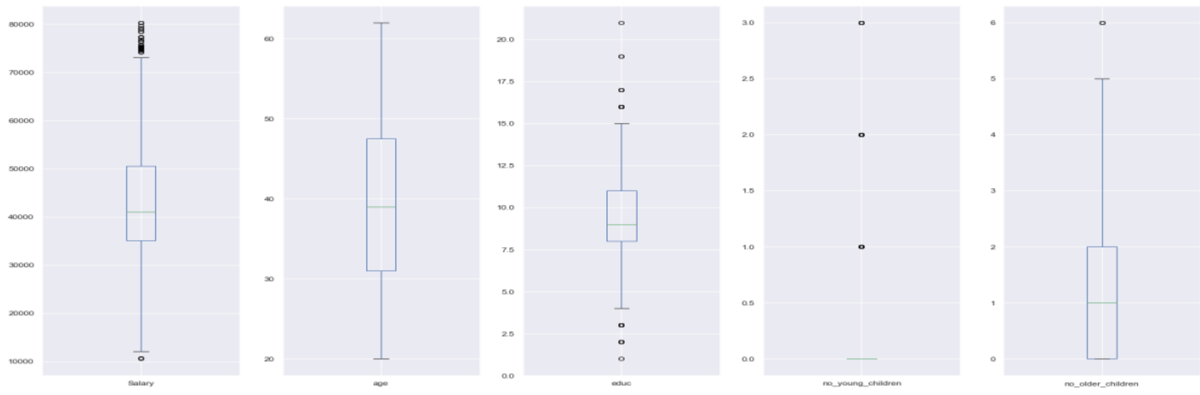
Boxplot before treating outliers

Figure 29



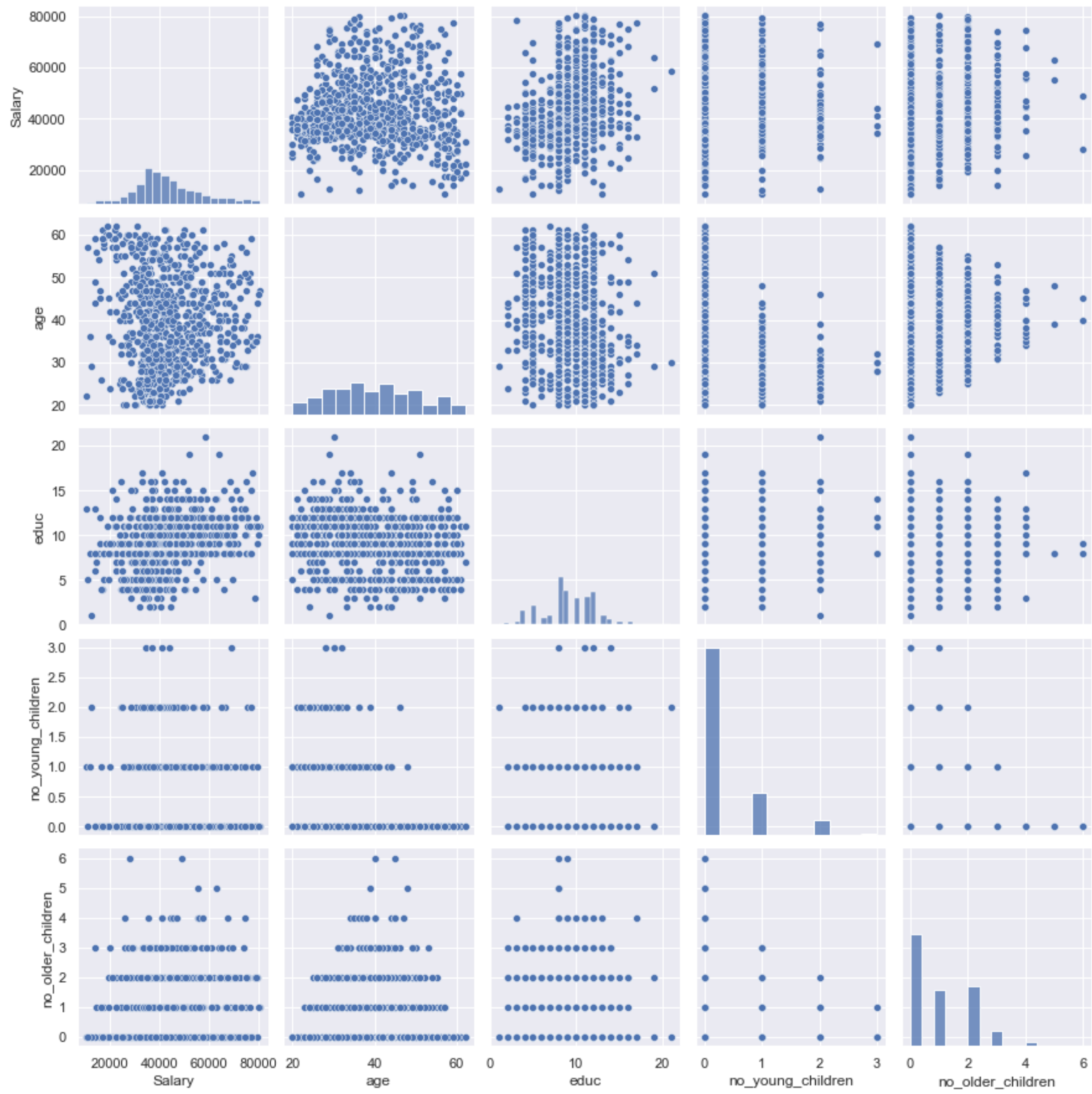
Boxplot after treating outliers

Figure 30



Multivariate Analysis

Figure 31 Multivariate Analysis



There is no correlation between the data, the data seems to be normal. There is

no huge difference in the data distribution among the holiday package, I don't
see any clear two different distribution in the data. No multi collinearity in the data.

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

The encoding helps the logistic regression model predict better results

Table 22 Encoded table

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

The grid search method is used for logistic regression to find the optimal solving and the parameters for solving.
The grid search method gives, liblinear solver which is suitable for small datasets.
Tolerance and penalty has been found using grid search method.

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

Accuracy score for Logistic regression train variables is 0.6508771929824562
Accuracy score for Logistic regression test variables is 0.6204081632653061

Figure 32 confusion matrix Train variables for logistic regression

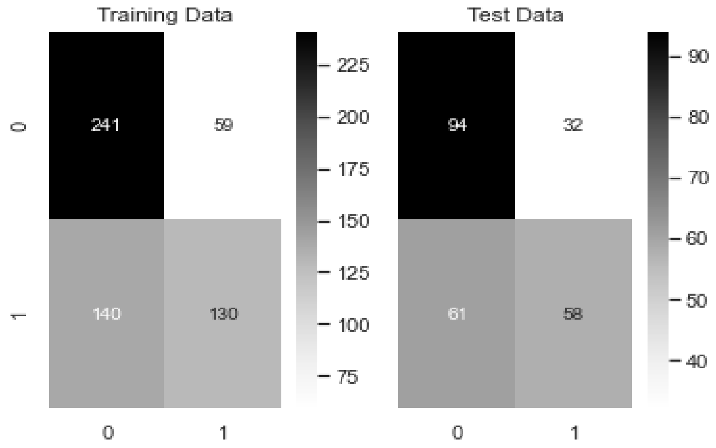


Table 23 Logistic regression Classification report for train data

	precision	recall	f1-score	support
no	0.63	0.80	0.71	300
yes	0.69	0.48	0.57	270
accuracy			0.65	570
macro avg	0.66	0.64	0.64	570
weighted avg	0.66	0.65	0.64	570

Table 24 Logistic regression Classification report for test data

	precision	recall	f1-score	support
no	0.61	0.75	0.67	126
yes	0.64	0.49	0.56	119
accuracy			0.62	245
macro avg	0.63	0.62	0.61	245
weighted avg	0.62	0.62	0.61	245

AUC and ROC FOR Logistic regression
AUC for the Training Data: 0.738
AUC for the Test Data: 0.665

Figure 32 AUC and ROC FOR Logistic regression

LDA

Accuracy score for LDA train variables is 0.6754385964912281
Accuracy score for LDA test variables is 0.6204081632653061

Figure 33 confusion matrix Train variables for LDA

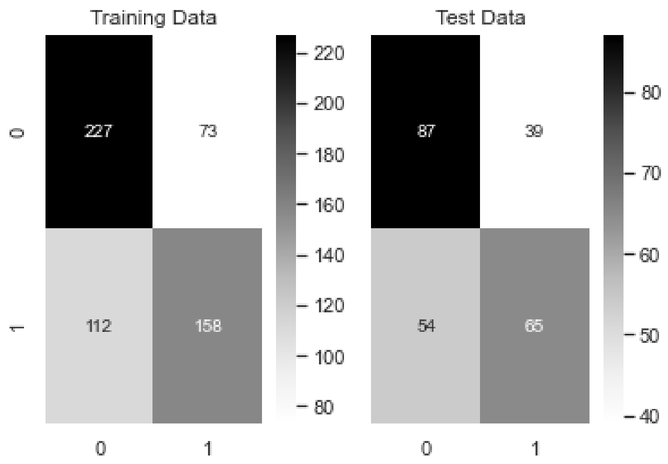


Table 25 LDA Classification report for train data

	precision	recall	f1-score	support
no	0.67	0.76	0.71	300
yes	0.68	0.59	0.63	270
accuracy			0.68	570
macro avg	0.68	0.67	0.67	570
weighted avg	0.68	0.68	0.67	570

Table 26 LDA Classification report for test data

	precision	recall	f1-score	support
no	0.62	0.69	0.65	126
yes	0.62	0.55	0.58	119
accuracy			0.62	245
macro avg	0.62	0.62	0.62	245
weighted avg	0.62	0.62	0.62	245

Figure 34 AUC and ROC FOR LDA

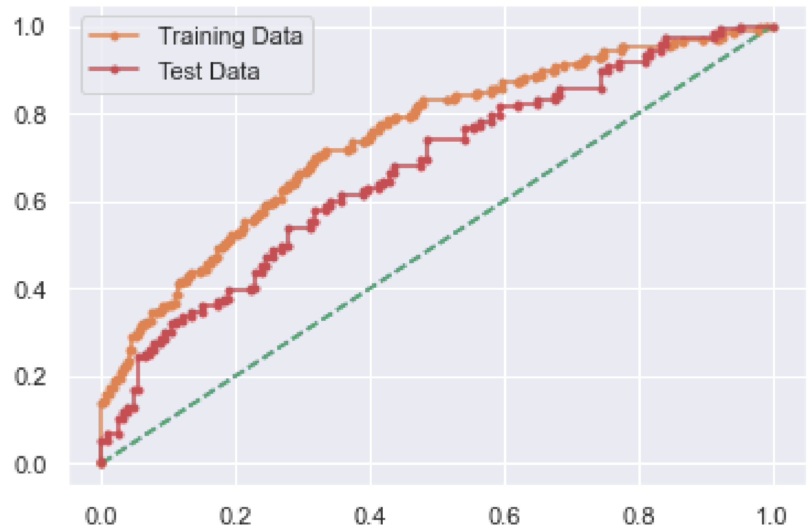


Table 27 LDA and logistic regression Train and Test data

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test
Accuracy	0.65	0.62	0.68	0.62
AUC	0.74	0.67	0.74	0.67
Recall	0.48	0.49	0.59	0.55
Precision	0.69	0.64	0.68	0.62
F1 Score	0.57	0.56	0.63	0.58

Figure 35 LDA and Linear Regression train data

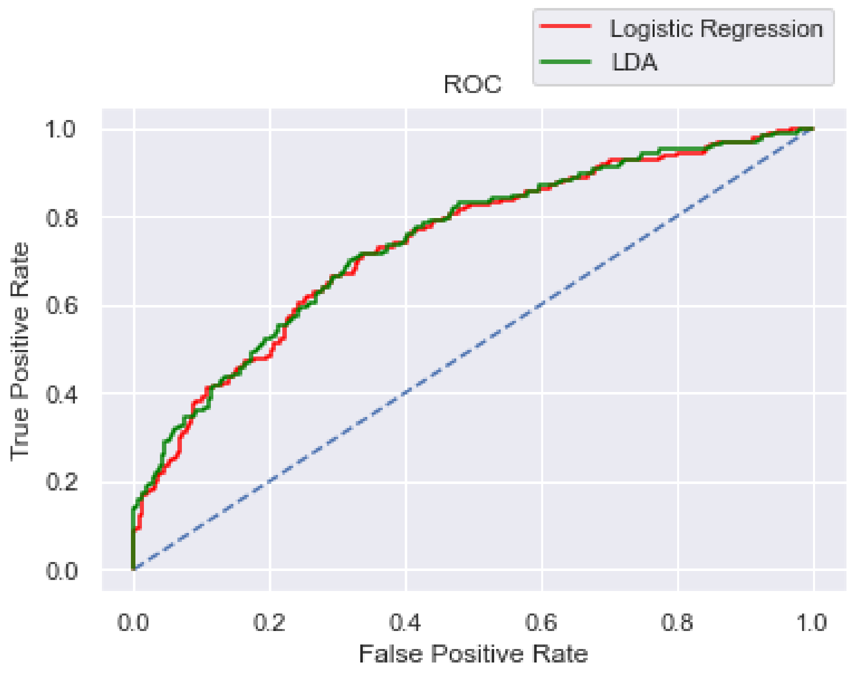
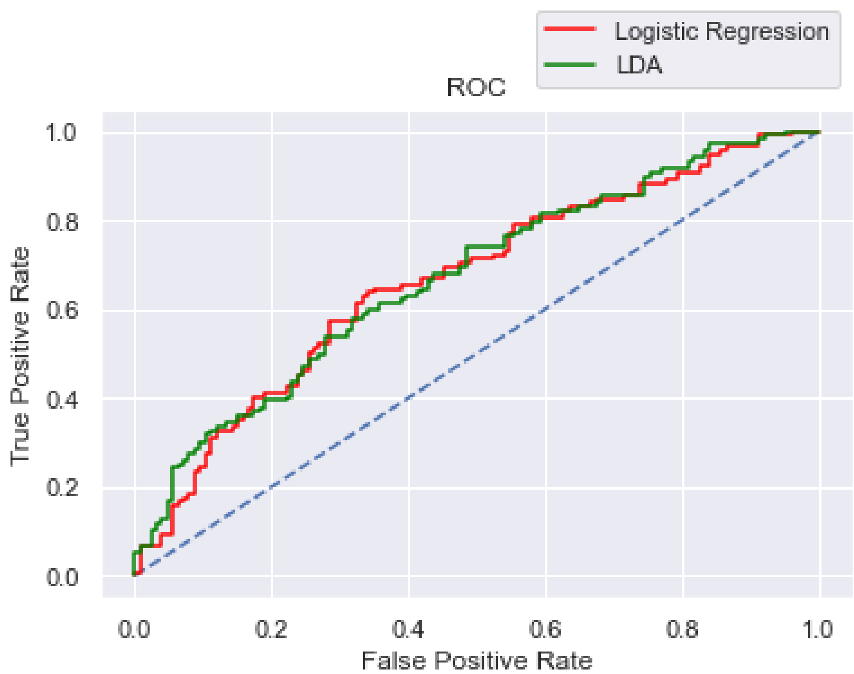


Figure 36 LDA and Linear Regression test data



Comparing both these models, we find both results are same, but LDA works better when there is category target variable.

2.4 Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

- We had a business problem where we need predict whether an employee would opt for a holiday package or not, for this problem we had done predictions both logistic regression and linear discriminant analysis. Since both are results are same.
- The EDA analysis clearly indicates certain criteria where we could find people aged above 50 are not interested much in holiday packages. So this is one of the we find aged people not opting for holiday packages.
- People ranging from the age 30 to 50 generally opt for holiday packages.
- Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package. The important factors deciding the predictions are salary, age and educ.

Recommendations

- To improve holiday packages over the age above 50 we can provide religious destination places.
- For people earning more than 150000 we can provide vacation holiday packages.
- For employee having more than number of older children we can provide packages in holiday vacation places.