Problem 1: Linear Regression

Problem Statement:

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.

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.
Colour	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.

Questions:

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

Answer:
The data has 26967 rows and 11 columns.

The first 5 rows in the data are:

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
2	3	0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779

The columns in the data set are:

1. Unnamed: 0
2. carat
3. cut
4. color
5. clarity
6. depth
7. table
8. x
9. y
10. z
11. price

The info of the raw data set is as shown below:

#	Column	Non-Null Count	Dtype
0	carat	26967 non-null	float64
1	cut	26967 non-null	object
2	color	26967 non-null	object
3	clarity	26967 non-null	object
4	depth	26270 non-null	float64
5	table	26967 non-null	float64
6	x	26967 non-null	float64
7	y	26967 non-null	float64
8	z	26967 non-null	float64
9	price	26967 non-null	int64

- The columns cut, color and clarity are categorical and are object data type
- The other columns are all numeric and hence are float and int data types

- The column depth has 697 blank entries or NaN which needs special attention

Five number summary of the data:

	count	mean	std	min	25%	50%	75%	max
carat	26967	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
depth	26270	61.745147	1.41286	50.8	61	61.8	62.5	73.6
table	26967	57.45608	2.232068	49	56	57	59	79
x	26967	5.729854	1.128516	0	4.71	5.69	6.55	10.23
y	26967	5.733569	1.166058	0	4.71	5.71	6.54	58.9
z	26967	3.538057	0.720624	0	2.9	3.52	4.04	31.8
price	26967	3939.518115	4024.864666	326	945	2375	5360	18818

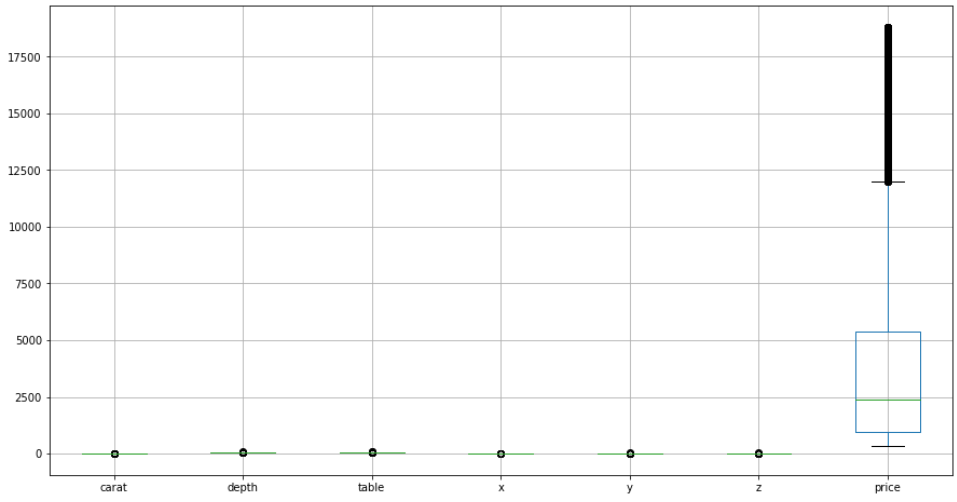
	count	unique	top	freq
cut	26967	5	Ideal	10816
color	26967	7	G	5661
clarity	26967	8	SI1	6571

- The x, y and z columns have 0 values. Since x, y and z are length, width and height of the gem respectively they cannot be 0mm. This too needs special attention.
- Columns depth, table and x are very close to a normal distribution
- Columns carat, y, z and price are right skewed
- There are 34 rows that have duplicate data

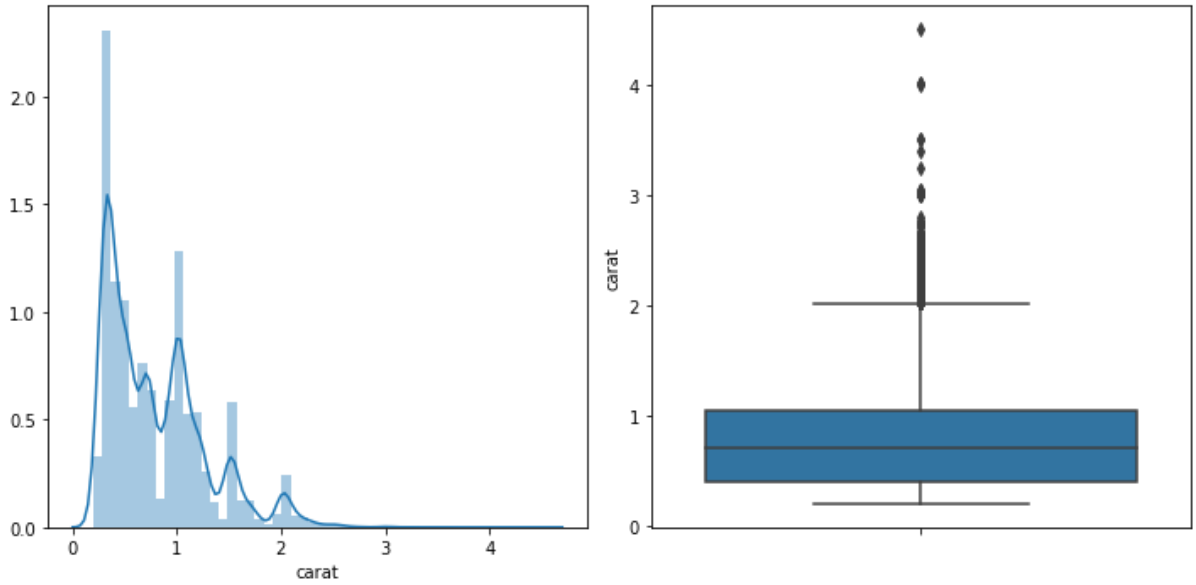
The price is the target variable in this data set.

Univariate analysis

Box plot of the entire data set:



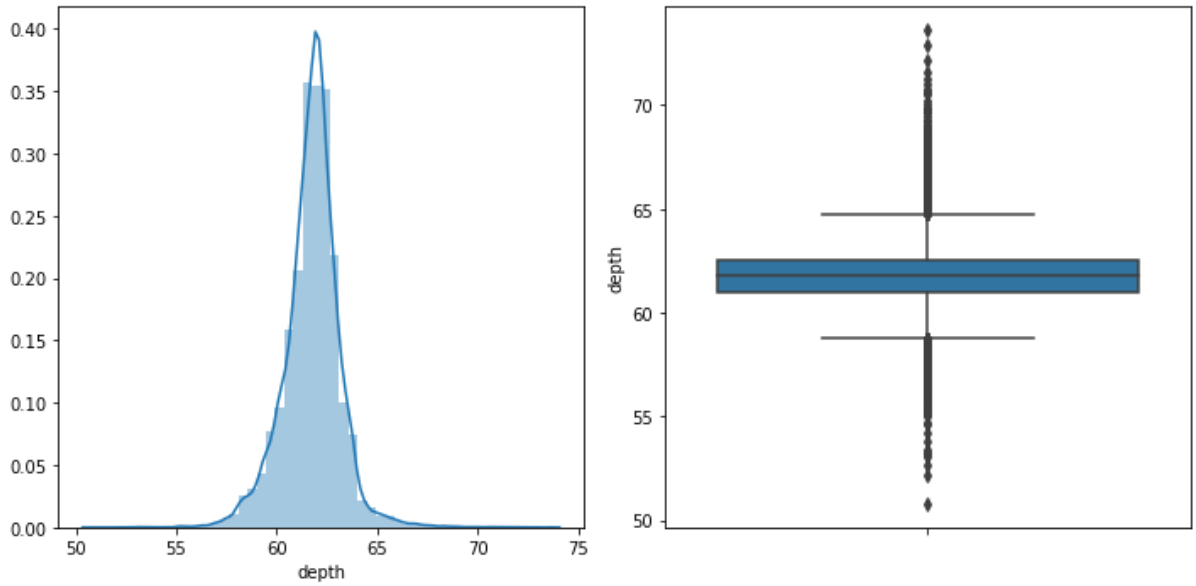
carat



carat is Positive or Right skewed.
The number of outliers in carat is 662

---

depth



depth is Negative or Left skewed.
The number of outliers in depth is 1225

---

table

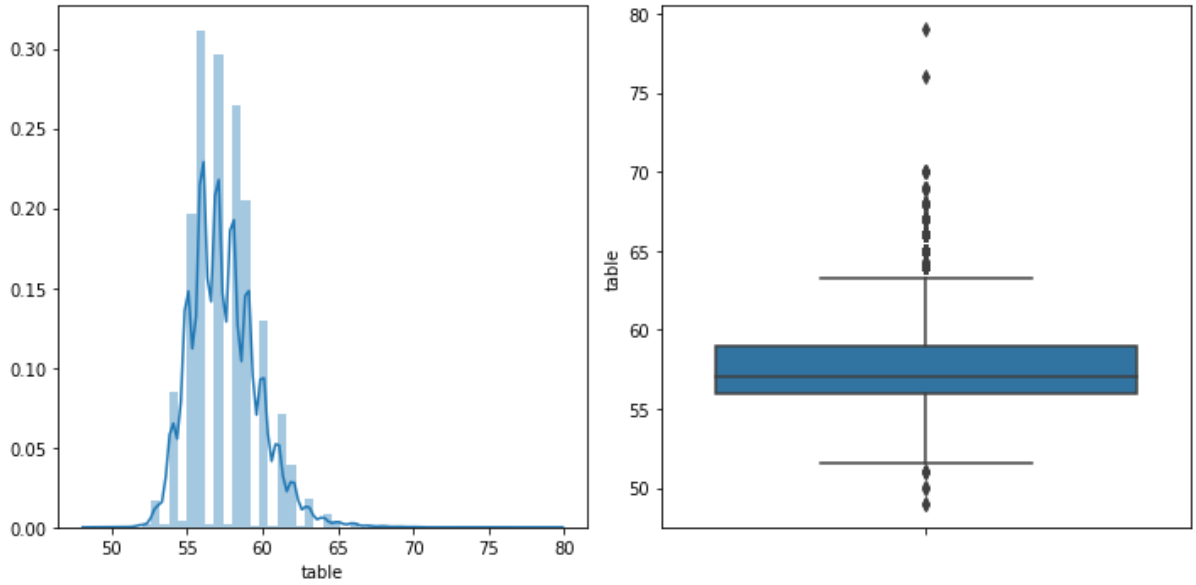
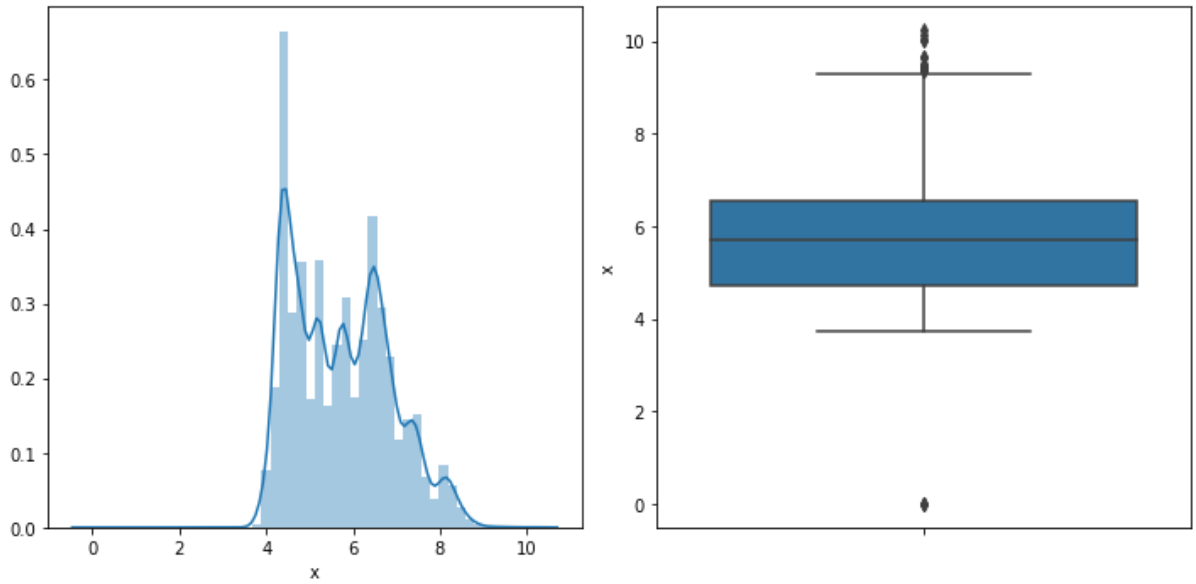


table is Positive or Right skewed.
The number of outliers in table is 318

---

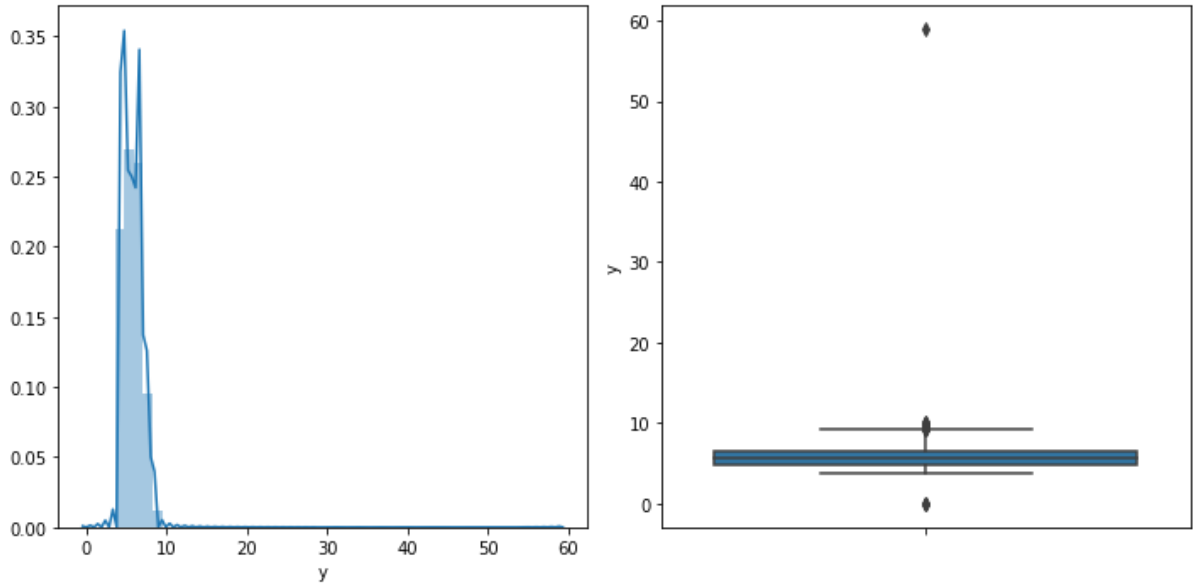
x



x is Positive or Right skewed.
The number of outliers in x is 15

---

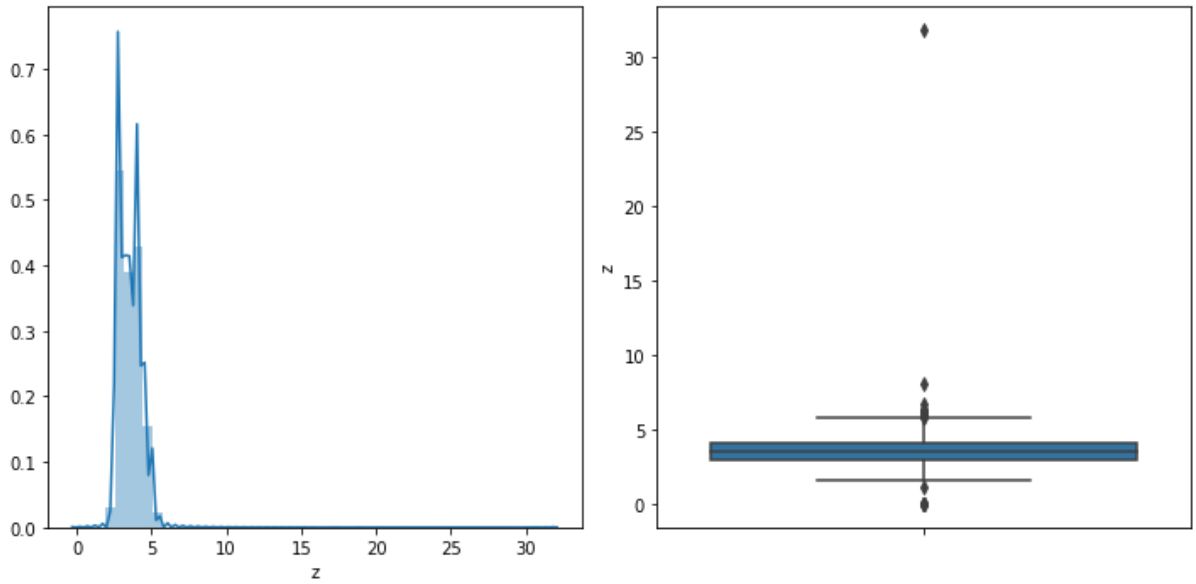
y



y is Positive or Right skewed.
The number of outliers in y is 15

---

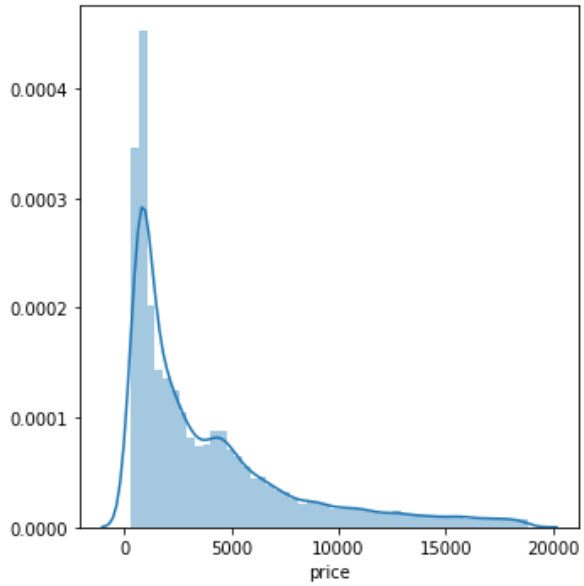
z

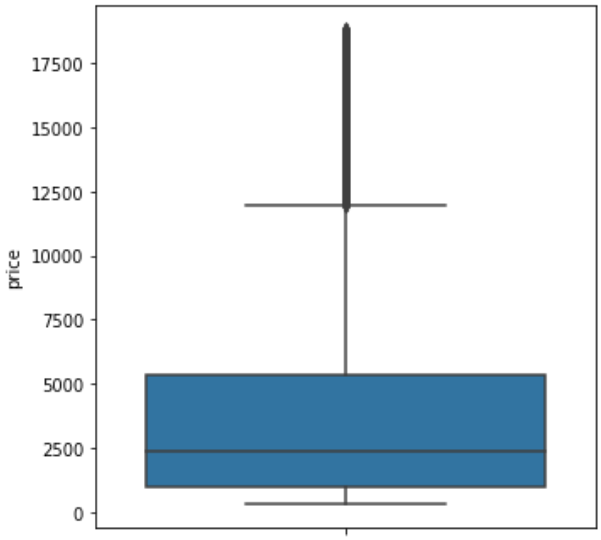


z is Positive or Right skewed.
The number of outliers in z is 23

---

price



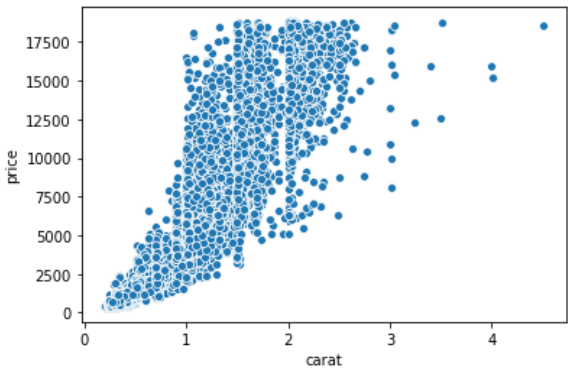


price is Positive or Right skewed.
The number of outliers in price is 1779

- All the variables, except depth, are multi-modal
- There are considerable number of outliers in each of the variables except

Bivariate analysis

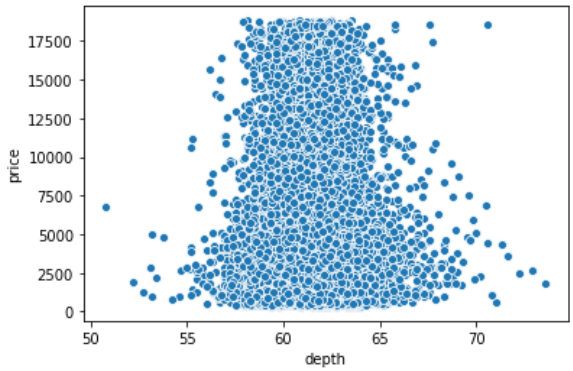
carat vs price



The correlation between the 2 variables are: 0.9224161094805412

---

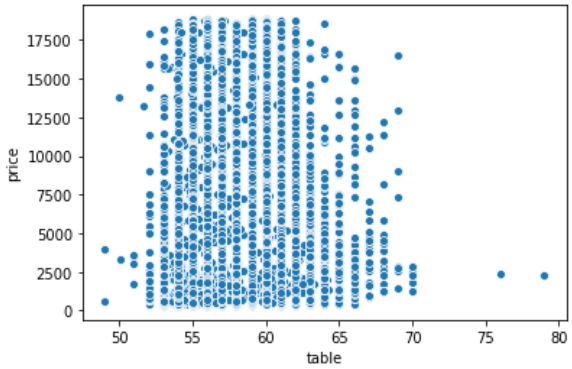
depth vs price



The correlation between the 2 variables are: -0.0025686163299551076

---

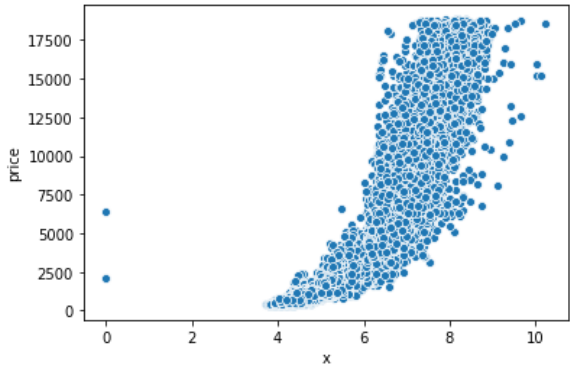
table vs price



The correlation between the 2 variables are: 0.12694223324168127

---

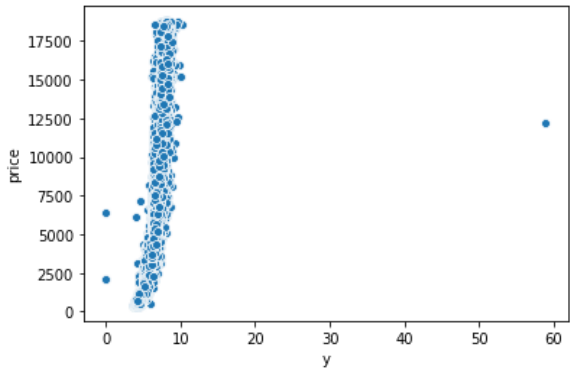
x vs price



The correlation between the 2 variables are: 0.8862471788154085

---

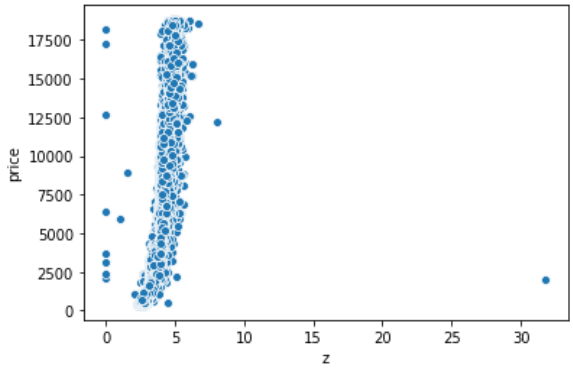
y vs price




```
The correlation between the 2 variables are:  0.856242540905528
```

---

z vs price

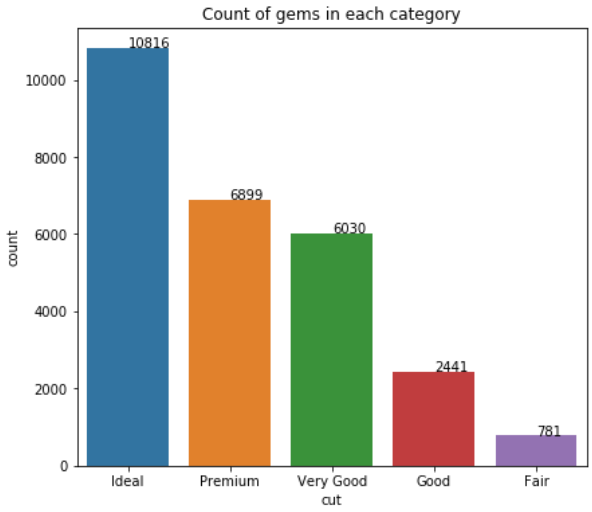


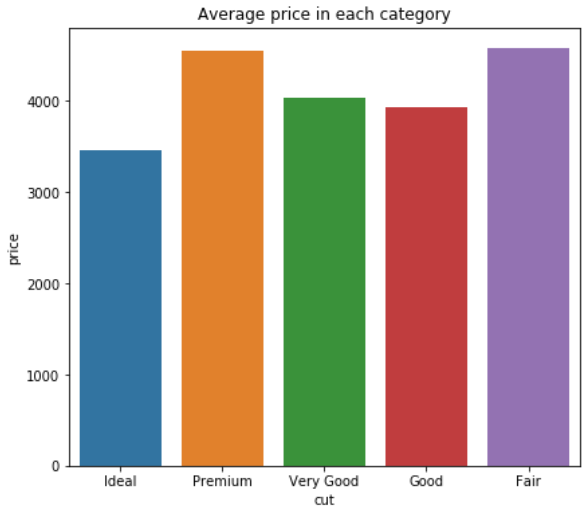

```
The correlation between the 2 variables are:  0.8505361306239169
```

- All the variables, except depth and table, have a good correlation with the target variable and prove to be important predictors.
- However, the depth and table form a cloud like plot which proves to be a weak predictor.

Cut vs price

```
The number of unique entries in the column cut : 5
The entry with the highest frequency in cut : Ideal
```

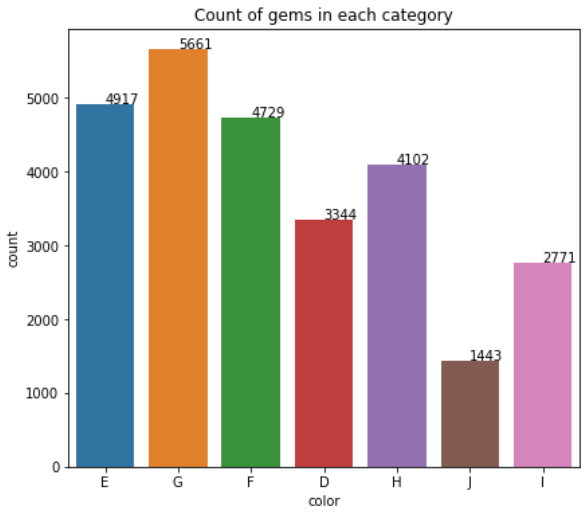


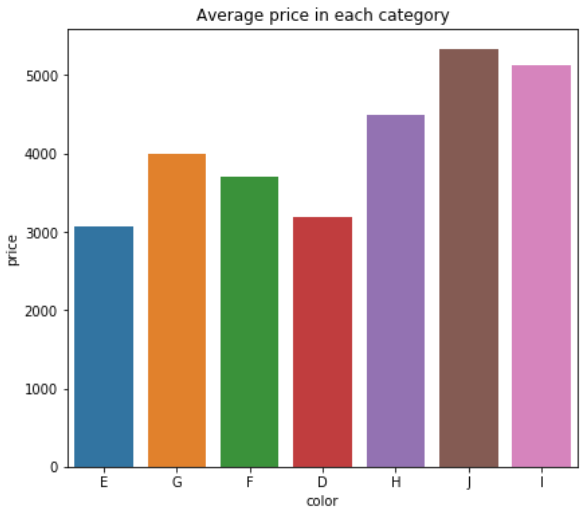



```
Percentage share:
Ideal        40.108280
Premium      25.583120
Very Good    22.360663
Good          9.051804
Fair          2.896132
Name: cut, dtype: float64
```

Color vs price

```
The number of unique entries in the column color : 7
The entry with the highest frequency in color : G
```





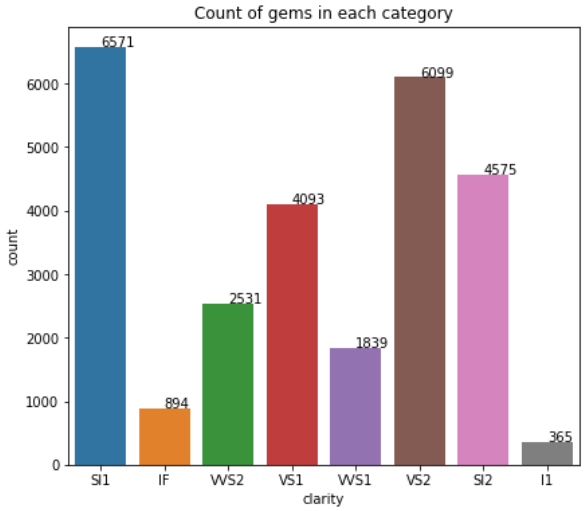

```
Percentage share:
G    20.992324
E    18.233396
F    17.536248
H    15.211184
```

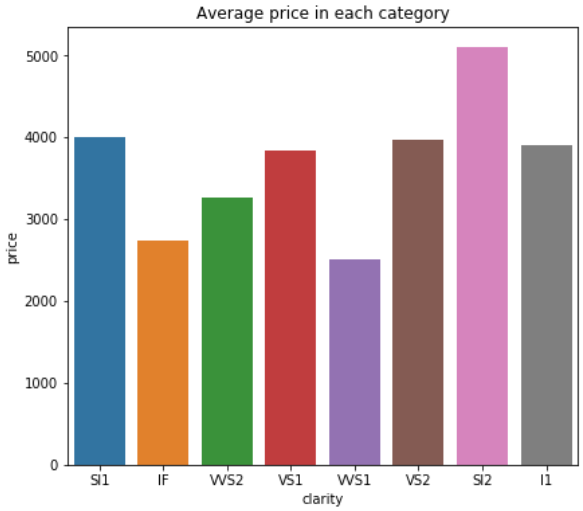
```
D    12.400341
I    10.275522
J     5.350985
Name: color, dtype: float64
```

---

clarity

```
The number of unique entries in the column clarity : 8
The entry with the highest frequency in clarity : SI1
```



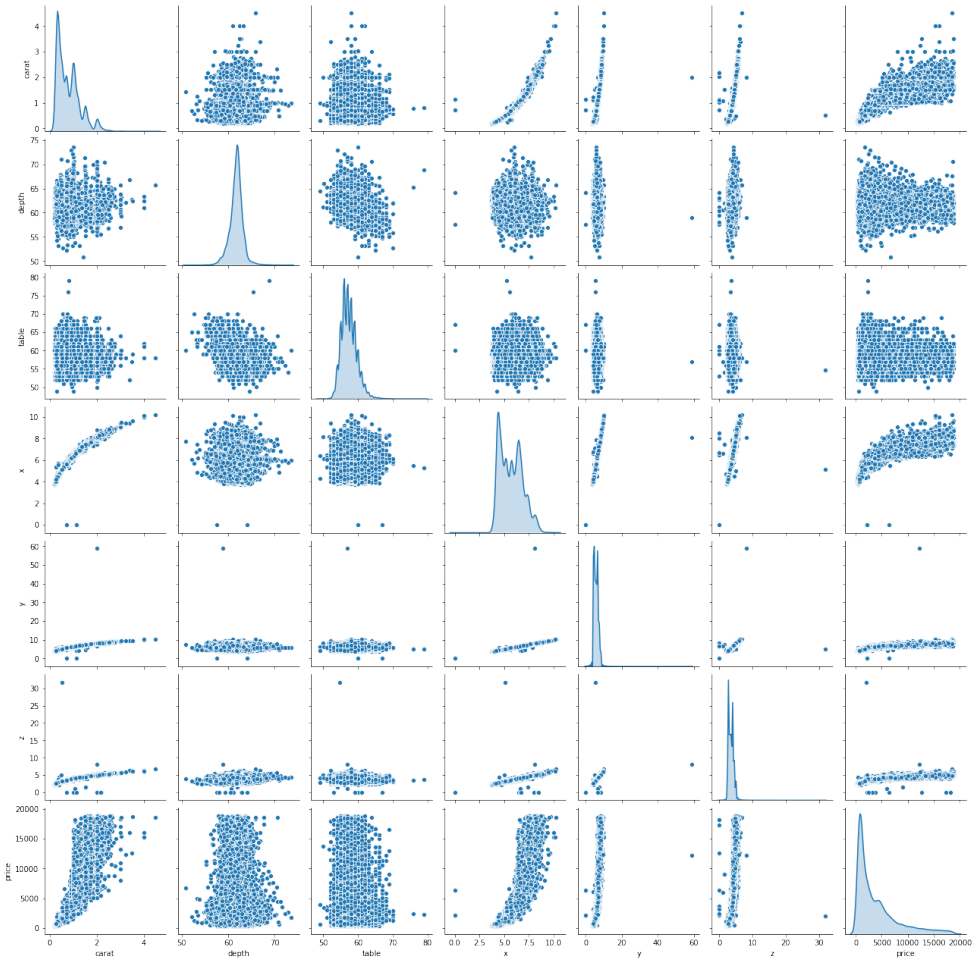



```
Percentage share:
SI1     24.366819
VS2     22.616531
SI2     16.965180
VS1     15.177810
VVS2     9.385545
VVS1     6.819446
IF       3.315163
I1       1.353506
Name: clarity, dtype: float64
```

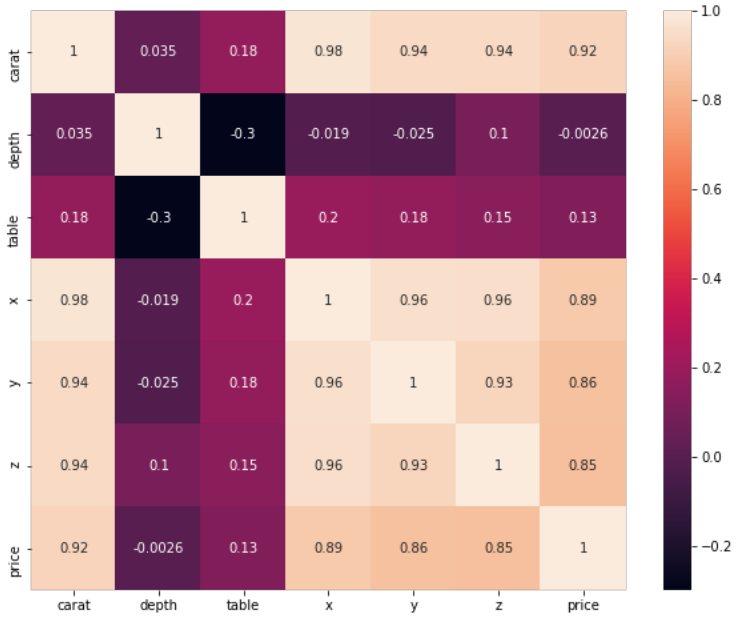
- Since the highest quality of cut is preferred the most we can say that people prefer a quality diamond.
- Since colours E, F, G and H are around the same count we can say that people prefer the mid-range colours. Even the prices are mid-ranged.
- Since SI2, VS2, SI2 and VS1 are all high in count mid-range in price they seem to be highly preferred by the customers. However the average price of the best clarity much lesser that the average price of the worst clarity.
- The average price of fair cut and premium cut are in par. The ideal cut as the lowest average price.
- Though they aren't good colours as per the rating, it seems like the availability of colour J is less and hence the most expensive amongst all categories. The next less availability is for colour I.
- Clarity of SI1 is the most preferred and the one in the mid-price range.

Multi-variate analysis

Pair plot:



Heatmap:



- The variables x, y and z are highly correlated. We can drop a few columns and check if the model performs better.

- There are highly correlated pairs with the price and other variable which is a good sign.
- Apart from x, y and z, there is no much correlation between the independent variables.

Imputation and outlier treatment

There are 34 duplicate values in the data. These will give weightage to the repetitive data points which avoid by dropping the duplicate rows.

Since the depth has many missing values which accounts to 2.58% of the number of rows in the data and also because the depth has many outliers we will impute the missing values with the mean.

Post missing value treatment we will proceed with treatment of 0 values in x, y and z. On checking the number of rows that have 0 as the value, we can see that there were 9 rows like so. However one row got removed while removing duplicate rows. Since cannot have length, width or height of 0mm we will proceed with dropping those rows from the data frame.

In univariate analysis it could be observed that there were many outliers. Since the linear regression analysis is very sensitive to outliers we will replace the outlier values with the closes whisker value.

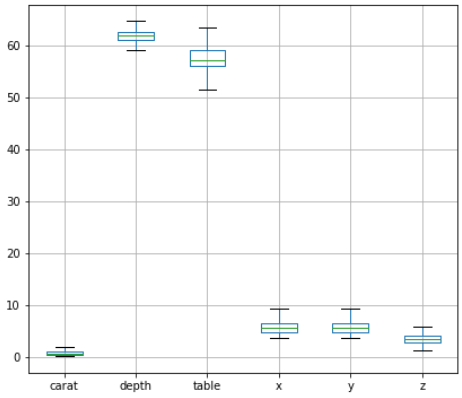
Post treatment of data:
5 number summary

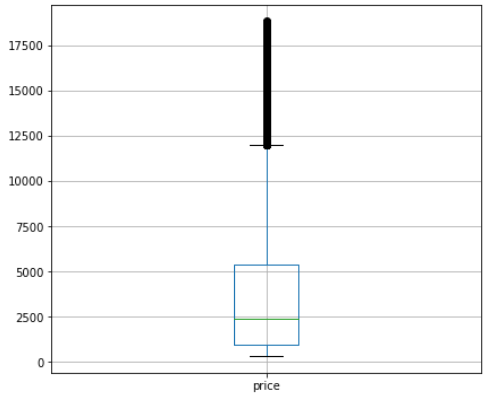
	count	mean	std	min	25%	50%	75%	max
carat	26925	0.793119	0.461998	0.2	0.4	0.7	1.05	2.025
depth	26925	61.749236	1.218216	59	61.1	61.8	62.5	64.6
table	26925	57.435023	2.156704	51.5	56	57	59	63.5
x	26925	5.729217	1.1255	3.73	4.71	5.69	6.55	9.31
y	26925	5.731159	1.117494	3.71	4.71	5.7	6.54	9.285
z	26925	3.537625	0.695681	1.19	2.9	3.52	4.04	5.75
price	26925	3936.249991	4020.983187	326	945	2373	5353	18818

Info of the treated data

#	Column	Non-Null Count	Dtype
0	carat	26925 non-null	float64
1	cut	26925 non-null	object
2	color	26925 non-null	object
3	clarity	26925 non-null	object
4	depth	26925 non-null	float64
5	table	26925 non-null	float64
6	x	26925 non-null	float64
7	y	26925 non-null	float64
8	z	26925 non-null	float64
9	price	26925 non-null	int64

The boxplot of treated data





Post dropping the duplicate rows and the rows with x, y and z as 0, the shape of the data is 26925 rows and 11 columns.

Scaling is not a mandatory step in linear regression. We will scale only if we want to pivot the intercept close to 0. In this case we will proceed without scaling. Scaling will only alter the coefficients but will not alter the performance of the model.

Encoding

Since the categorical are all ordinal types (they have an order in the clarity, cut or color), we will do a label encoding. In this we will number the least quality or preferred as 1 and the best quality or most preferred as n. This will add weightage to the preference to the numbers and then run the model. This will help add a penalty to lower quality classes.

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

	color_desc	color_code
0	J	1
1	I	2
2	H	3
3	G	4
4	F	5
5	E	6
6	D	7

	cut_desc	cut_code
0	Fair	1
1	Good	2
2	Very Good	3
3	Premium	4
4	Ideal	5

We will now merge this data on the respective description column to get the codes. This is similar to a vlook up in excel sheet. We will cross verify if the encoding is done right and then drop the columns containing the description and retain only the codes.

	carat	depth	table	x	y	z	price	cut_code	color_code	clarity_code
0	0.3	62.1	58	4.27	4.29	2.66	499	5	6	5
1	0.59	62	55	5.37	5.4	3.34	1664	5	6	5
2	0.54	62.1	56	5.24	5.19	3.24	1637	5	6	5
3	0.51	62.4	55	5.14	5.09	3.19	1443	5	6	5
4	0.43	61.745147	56	4.82	4.79	3.01	975	5	6	5
...	...	...	...	...	...	...	...	...	...	...
26920	0.9	64.6	60	5.96	5.88	3.98	1786	1	2	3
26921	0.73	64.6	58	5.63	5.5	3.68	1175	1	2	3
26922	1	64.6	57	6.15	6.03	4.09	1997	1	2	3
26923	1.2	64.6	55	6.65	6.59	4.31	2376	1	2	3
26924	0.73	64.6	55	5.51	5.34	3.84	1049	1	2	3

The data is now as shown above and all the columns are converted to int/float data types.

Modelling

We are splitting the data into train and test in the ration 7:3.
We will implement the regression model using stats model as it gives up more control and also gives us a detailed report.

Model 1:
The processed data is used in a simple linear regression model. The coefficients thus obtained are:

```
Intercept        1386.430438
carat           13811.506266
cut_code          141.775523
color_code        328.880503
clarity_code      477.178343
depth             -12.216172
table             -21.710030
x               -2565.491390
y                1592.704439
z               -1619.004643
```

The regression report is as shown below:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.917
Model:	OLS	Adj. R-squared:	0.917
Method:	Least Squares	F-statistic:	2.306e+04
Date:	Tue, 19 Jan 2021	Prob (F-statistic):	0.00
Time:	00:16:02	Log-Likelihood:	-1.5973e+05
No. Observations:	18847	AIC:	3.195e+05
Df Residuals:	18837	BIC:	3.196e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1386.4304	1014.996	1.366	0.172	-603.052	3375.913
carat	1.381e+04	104.839	131.740	0.000	1.36e+04	1.4e+04
cut_code	141.7755	9.318	15.215	0.000	123.511	160.040
color_code	328.8805	5.232	62.859	0.000	318.625	339.136
clarity_code	477.1783	5.695	83.787	0.000	466.015	488.341
depth	-12.2162	13.946	-0.876	0.381	-39.551	15.119
table	-21.7100	4.985	-4.355	0.000	-31.482	-11.938
x	-2565.4914	154.127	-16.645	0.000	-2867.594	-2263.389
y	1592.7044	154.168	10.331	0.000	1290.521	1894.888
z	-1619.0046	173.005	-9.358	0.000	-1958.110	-1279.900

Omnibus:	3487.536	Durbin-Watson:	1.992
Prob(Omnibus):	0.000	Jarque-Bera (JB):	36522.419
Skew:	0.583	Prob(JB):	0.00
Kurtosis:	9.719	Cond. No.	1.03e+04

```
The R-squared valued for the train data is 0.9167967711267376
The root mean squared value of the model on training data is 1159.84986
4
The root mean squared value of the model on tesing data is 1150.906593


The VIF values:
carat ---> 122.82142483907428
depth ---> 1266.0534160495022
table ---> 870.792818927523
x ---> 10679.18850534223
y ---> 9444.224919077704
z ---> 3351.134118738621
cut_code ---> 17.456889189268328
color_code ---> 8.532910645758758
clarity_code ---> 17.790060344085344
```

- The p_value for depth is more than 5% showing that the coefficient has no effect on the price.
- The prob of the model is 0. The model is performing at its best but this model may not be the best.
- The RMSE value is very high which means the error is high
- The coefficients are very large in numbers which means there is a high possibility of the model being overfit.

- The VIF's have very high values which show severe multicollinearity which will make interpretation of the model very hard.

Model 2:
Remedy for Model 1 short comings:

- We will drop depth to reduce the dimensionality of the data used for analysis.
- Since the VIF values are extremely high, we will remove a few columns form the data for modelling. From the correlation heatmap we can observe that x, y and z are highly correlated to carat an so these can be removed out by retaining only carat. The table variable also has a high VIF value and hence can be removed out.

The processed data is used in a simple linear regression model. The coefficients thus obtained are:

```
Intercept      -8540.301302
carat           9106.500966
cut_code         166.638982
color_code       320.138172
clarity_code     526.985397
```

The regression report is as shown below:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	4.582e+04
Date:	Tue, 19 Jan 2021	Prob (F-statistic):	0.00
Time:	09:14:18	Log-Likelihood:	-1.6080e+05
No. Observations:	18847	AIC:	3.216e+05
Df Residuals:	18842	BIC:	3.216e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-8540.3013	63.219	-135.091	0.000	-8664.216	-8416.387
carat	9106.5010	21.932	415.218	0.000	9063.513	9149.489
cut_code	166.6390	8.232	20.243	0.000	150.504	182.774
color_code	320.1382	5.526	57.934	0.000	309.307	330.969
clarity_code	526.9854	5.907	89.220	0.000	515.408	538.563

Omnibus:	3696.270	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16839.240
Skew:	0.891	Prob(JB):	0.00
Kurtosis:	7.274	Cond. No.	63.3

```
The R-squared valued for the train data is 0.9067707973902447
The root mean squared value of the model on training data is 1227.74357
6
The root mean squared value of the model on testing data is 1216.189305

The VIF values:
carat ---> 2.9771475046442455
cut_code ---> 11.146091243301237
color_code ---> 5.952711583834655
```

```
clarity_code ---> 10.043344198141916
```

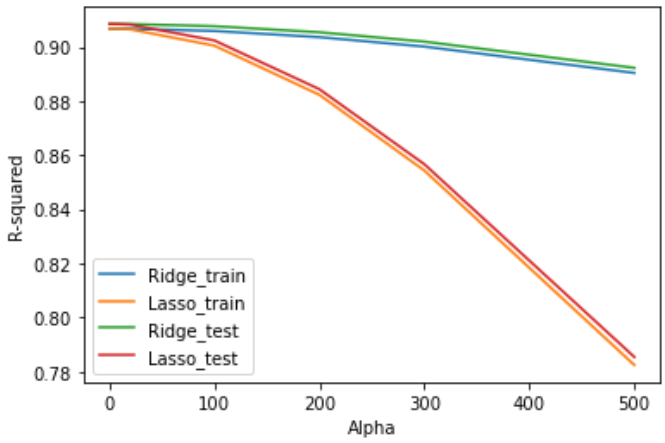
For a drop of 4 variables the Adjusted R^2 dropped by 0.01. This model is still performing equally well compared to the Model 1. The trade-off is that we are using much lesser data and there is a slight drop in R^2 and RMSE value has increased slightly.

- It can still be observed that the coefficients are quite high.

Model 3:
Remedy for Model 2 short comings:

- We will try various regularization techniques to reduce the magnitude of coefficients.

```
For alpha = 0.1, the root mean squared value of the model on train data is:
1227.7435820192222
For alpha = 0.5, the root mean squared value of the model on train data is:
1227.743721688514
For alpha = 1, the root mean squared value of the model on train data is:12
27.7441579708345
For alpha = 10, the root mean squared value of the model on train data is:1
227.801413783868
For alpha = 20, the root mean squared value of the model on train data is:1
227.97342054568
For alpha = 100, the root mean squared value of the model on train data is:
1233.1926824694183
For alpha = 200, the root mean squared value of the model on train data is:
1248.1080532645904
For alpha = 300, the root mean squared value of the model on train data is:
1270.5062438256546
For alpha = 500, the root mean squared value of the model on train data is:
1331.1221987116141
```

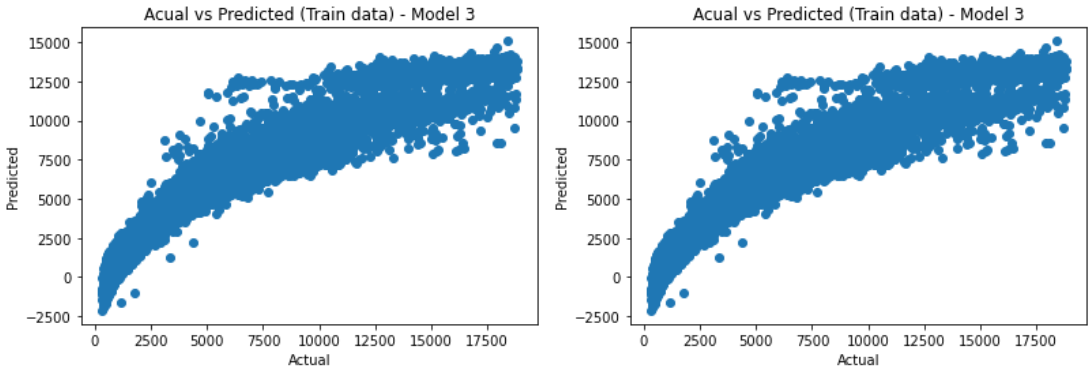


The lasso model reduces the R-squared value drastically when alpha increases. Since we will need large values of alpha without hampering the model parameters much we will go ahead with ridge model. From the above we can say that an alpha value of 300 will serve the purpose. It will not drop the R-squared vale as well the RMSE values are also not hampered a lot.

Ridge model with alpha of 300:

	Carat	Cut	Color	Clarity
Standard Linear Regression	9106.501	166.639	320.1382	526.9854
Ridge Regularized Regression	8305.138	140.5246	254.3325	447.1519

```
The root mean squared value of the model on training data is 1270.506244
The root mean squared value of the model on training data is 1258.70512
2
The R-squared valued for the train data is 0.9001632956964782
```



Model Comparison

Model parameters:

	R-Squared	RMSE (Train data)	RMSE (Test data)
Model 1	0.916797	1159.849864	1150.906593
Model 2	0.906771	1227.743576	1216.189305
Model 3	0.900163	1270.506244	1258.705122

Coefficients:

	Intercept	carat	cut_code	color_code	clarity_code	depth	table	x	y	z
Model 1	1386.430438	13811.50627	141.775523	328.880503	477.178343	-12.216172	-21.71003	-2565.49139	1592.704439	-1619.004643
Model 2	-8540.301302	9106.500966	166.638982	320.138172	526.985397	NaN	NaN	NaN	NaN	NaN
Model 3	-7028.980826	8305.138332	140.524598	254.332523	447.151899	NaN	NaN	NaN	NaN	NaN

Summary:

- Model 1 performs the best as it has the highest R-squared value and the lowest RMSE values on test and train data. However it makes use of 9 variables and it also has very high multi-collinearity which makes model interpretation difficult.
- One removing multi-colinear variables, and modelling we got Model 2. The model parameters have reduced in quality a little but this model uses only 4 variables against the originally used 9 variables.
- To regularize the model we have performed ridge model and we obtain Model 3. This model is very similar to Model 2 but is regularized and hence will be considered for modelling in production.

Business Insights

The variables that we will be using for predicting the price are:

1. Carat
2. Cut
3. Color
4. Clarity

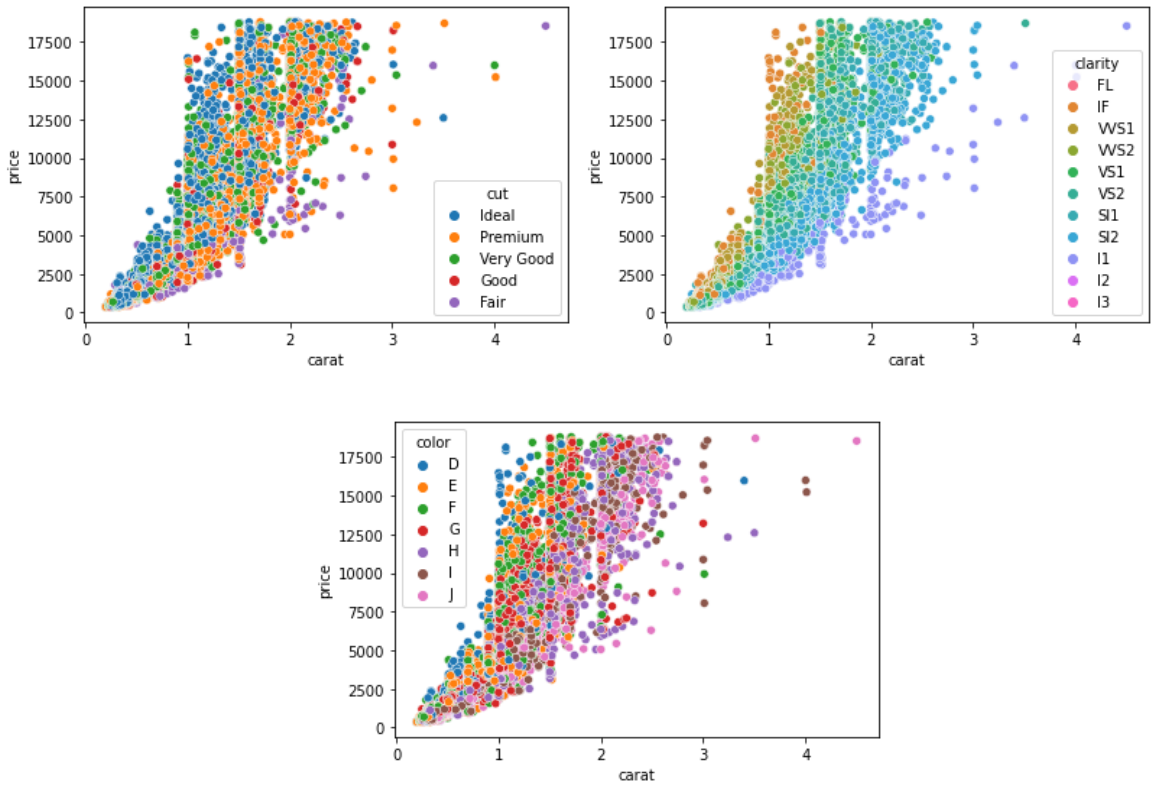
The relationship between price and the variable are as follows:
Price = 8305.14*Carat + 140.52*Cut + 254.33*Color + 447.15*Clarity + (-7028.98082609241)

Amongst the 4 variables we can say that Carat contributes to the price majorly. Clarity and Color also play a fairly important role in determining the price.

From the cost function we can say

- 8,305/carat
- 140/cut_code
- 254/color_code
- 447/clarity_code

For a unit increase in carat the cost of the diamond increases by 8,305.



As the quality of colour, cut and clarity increases, the price also increases. Carat proves to be a very strong predictor of the price.

Similar to diamond the qualities of zirconia are defined mainly by the 4C's i.e. Carat, Colour, Cut and Clarity. Other characters are basis of these variables.

Recommendations:

- Customers prefer to split their money into the 4C's. Though carat defines the resale value of the gem. There is huge clustering in the lower carats. The lower carats have higher quality of colour and cut. This shows that the customers want to purchase zirconia, more for the purpose of ornamentation and not as an investment.
- As the number of carats increases the quality of zirconia cut, colour and clarity decreases. The clustering in this space is very sparse.

- When purchasing zirconia from the vendor it is advised to focus more on the combination of good clarity and a considerable carat value that falls within the range of 5000. The ratio money spent on carat to clarity should approximately 18:1.

Problem 2: Logistic Regression and LDA

Problem Statement:

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.

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

Questions:

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.
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).
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.
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.

Answer:
The data has 872 rows and 8 columns. The first 5 entries in the data are as shown below:

	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

```
The columns in the data set are:
   1. Unnamed: 0
   2. Holliday_Package
   3. Salary
   4. age
   5. educ
   6. no_young_children
   7. no_older_children
   8. foreign
```

The info of the raw data set is as shown below:

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

- The columns Holiday_Pakage and foreign are nominal categories and so are object data type.
- The other columns are all integer data type
- All columns have 872 entries and so there are no NaN or blank value sin the data

Five number summary of the data:

	count	mean	std	min	25%	50%	75%	max
Salary	872	47729.17202	23418.66853	1322	35324	41903.5	53469.5	236961
age	872	39.955275	10.551675	20	32	39	48	62
educ	872	9.307339	3.036259	1	8	9	12	21
no_young_children	872	0.311927	0.61287	0	0	0	0	3
no_older_children	872	0.982798	1.086786	0	0	1	2	6

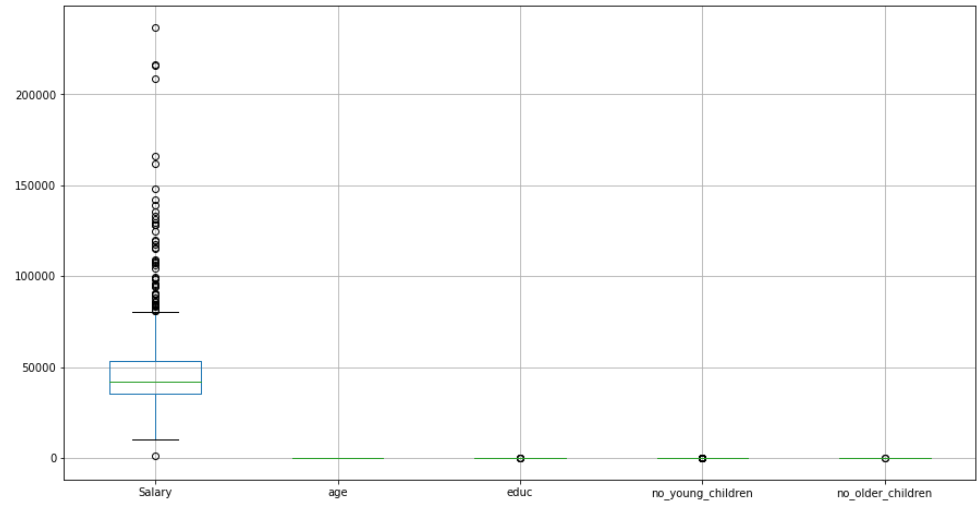
	count	unique	top	freq
Holliday_Package	872	2	no	471
foreign	872	2	no	656

- The Holiday_Package and foreign variables are mostly binary
- age and educ seems to be normally distributed but may have outliers
- no_young_children and no_older_children have very eneven distribution as they are natural numbers and not continuous

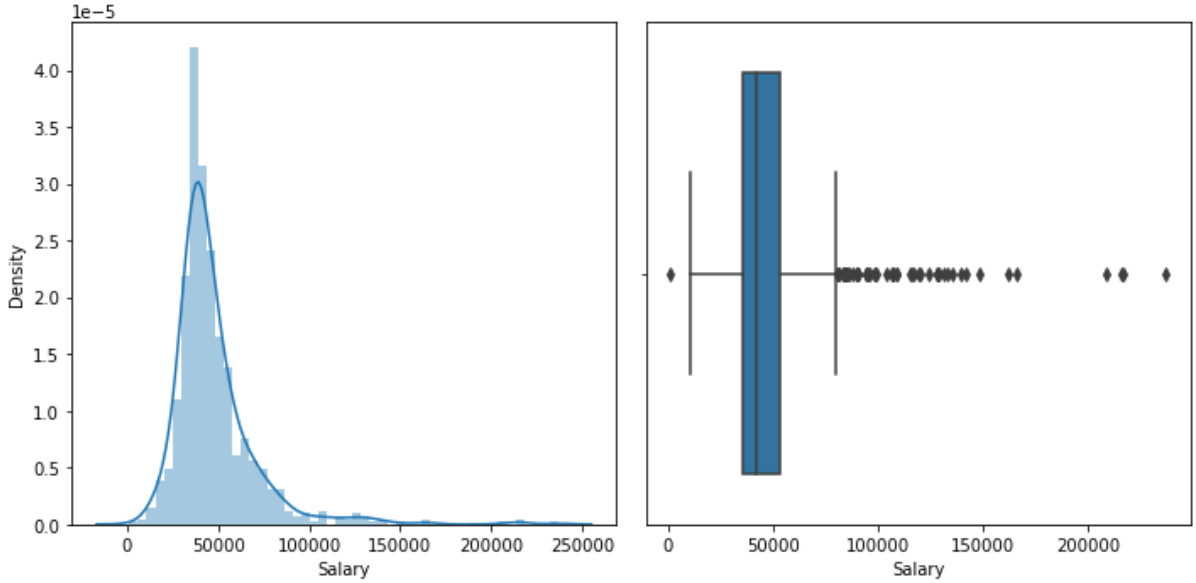
Here the Holiday_Package is the target variable.

Univariate analysis

Box plot of the entire data set:



Salary

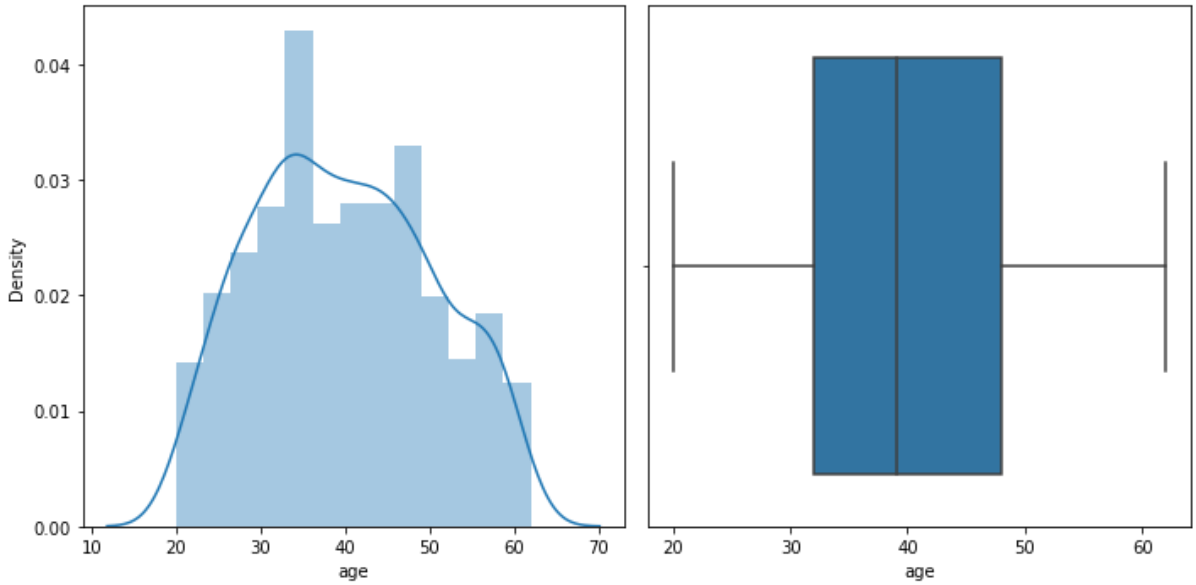


Salary is Positive or Right skewed.

The number of outliers in Salary is 57

---

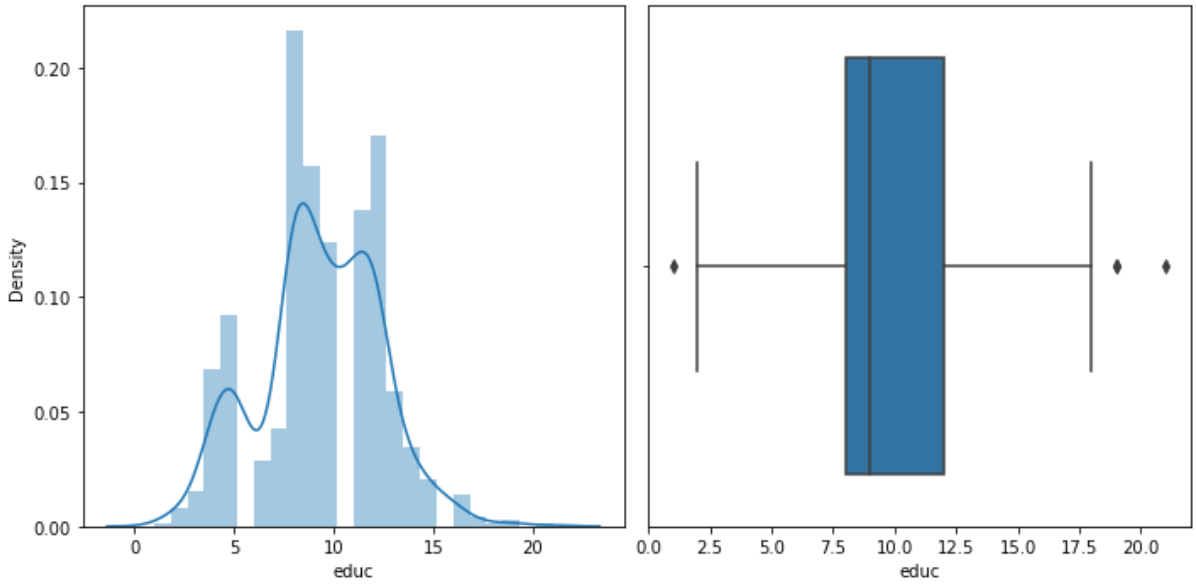
age



age is Positive or Right skewed.
The number of outliers in age is 0

---

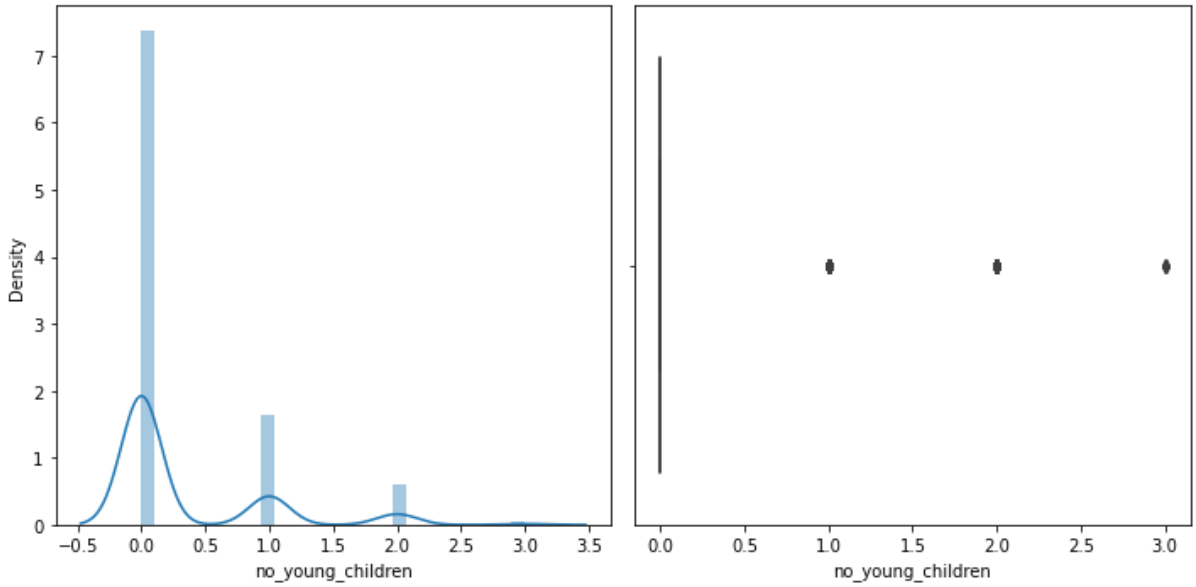
educ



educ is Negative or Left skewed.
The number of outliers in educ is 4

---

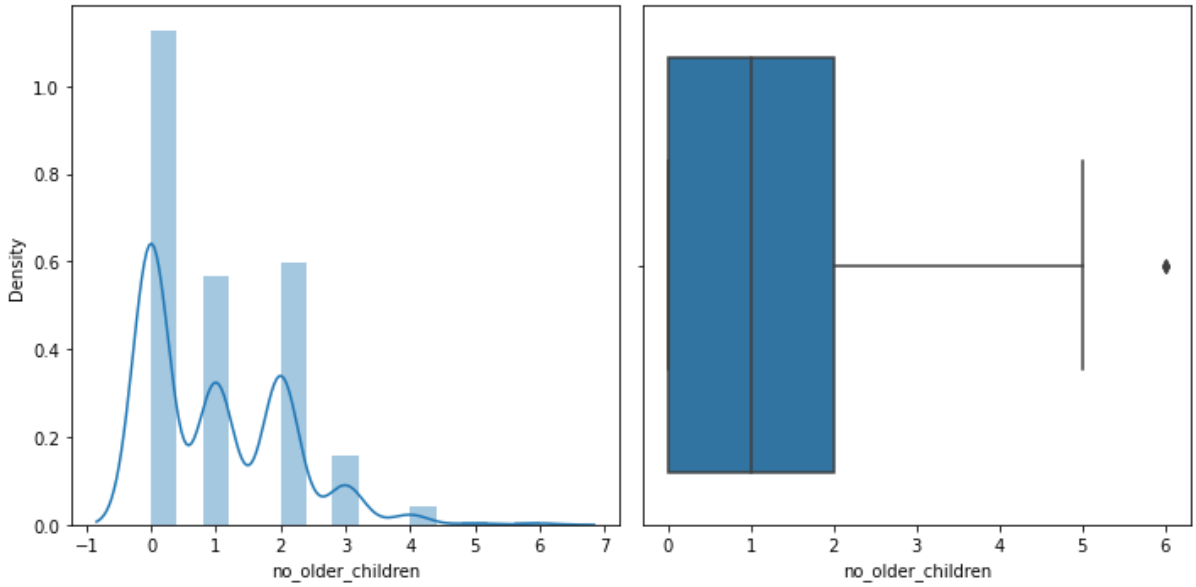
no_young_children



no_young_children is Positive or Right skewed.
The number of outliers in no_young_children is 207

---

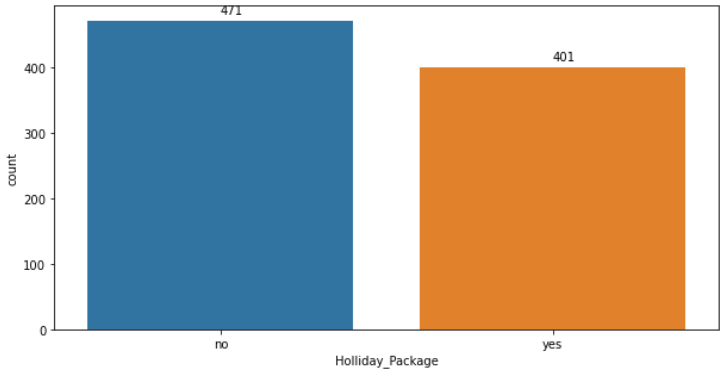
no_older_children



no_older_children is Positive or Right skewed.
The number of outliers in no_older_children is 2

Holliday_Package

```
The number of unique entries in the column Holliday_Package : 2
The entry with the highest frequency in Holliday_Package : no
```

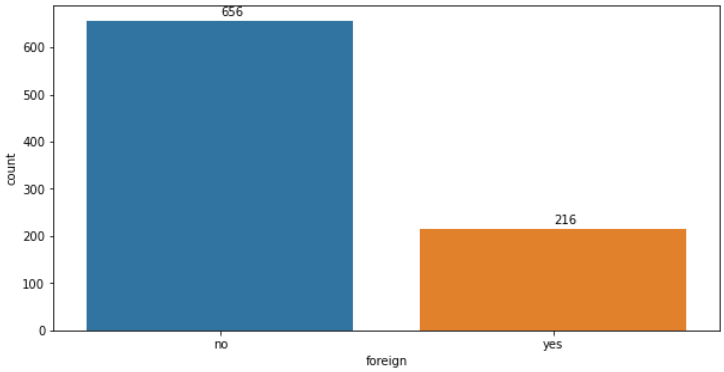



```
Percentage share:
no     54.013761
yes    45.986239
Name: Holliday_Package, dtype: float64
```

---

foreign

```
The number of unique entries in the column foreign : 2
The entry with the highest frequency in foreign : no
```

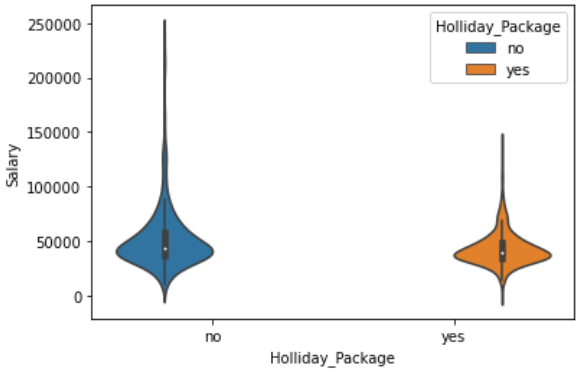



```
Percentage share:
no     75.229358
yes    24.770642
Name: foreign, dtype: float64
```

- Salary is very close to normal distribution
- All other continuous variables are multi-modal
- Except age all other variables have outliers
- The majority class in both the categorical variables is 'no'

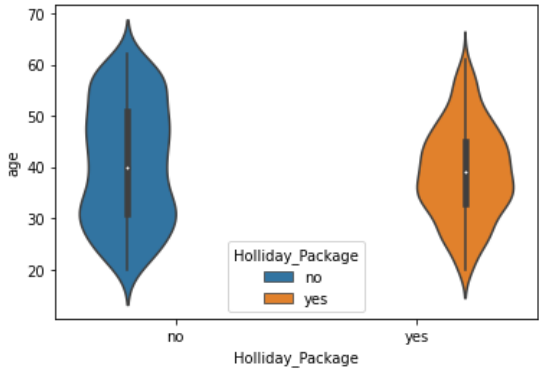
Bivariate analysis

Salary vs Holiday_Package



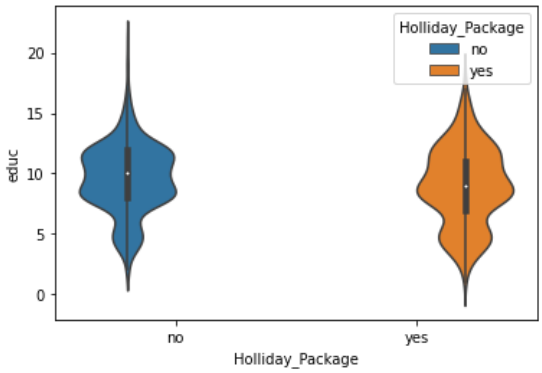
---

age vs Holiday_Package



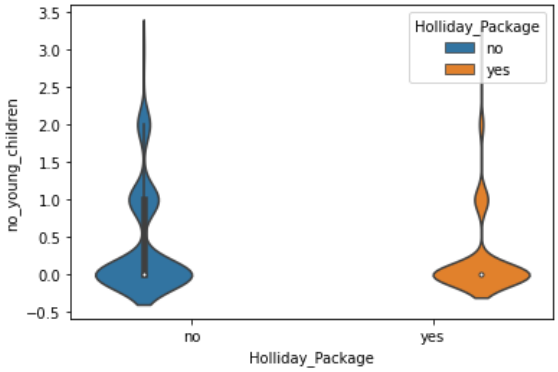
---

educ vs Holiday_Package



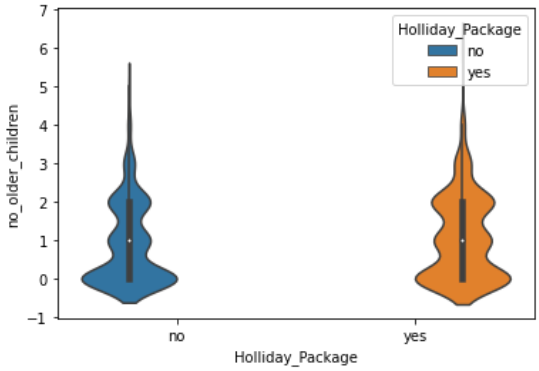
---

no_young_children vs Holiday_Package



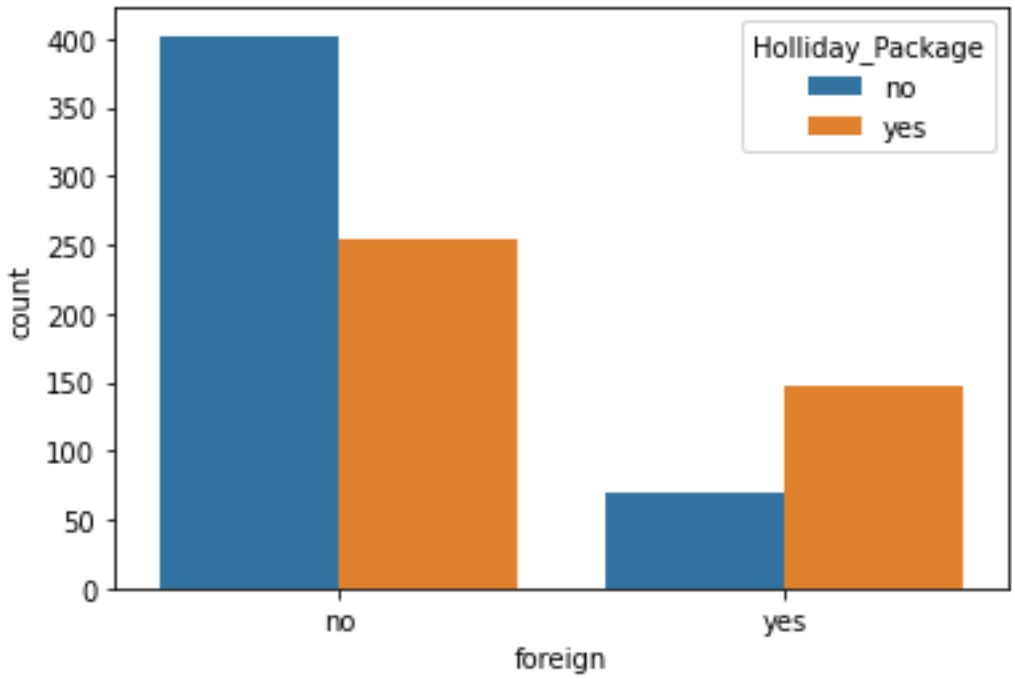
---

no_older_children vs Holiday_Package



---

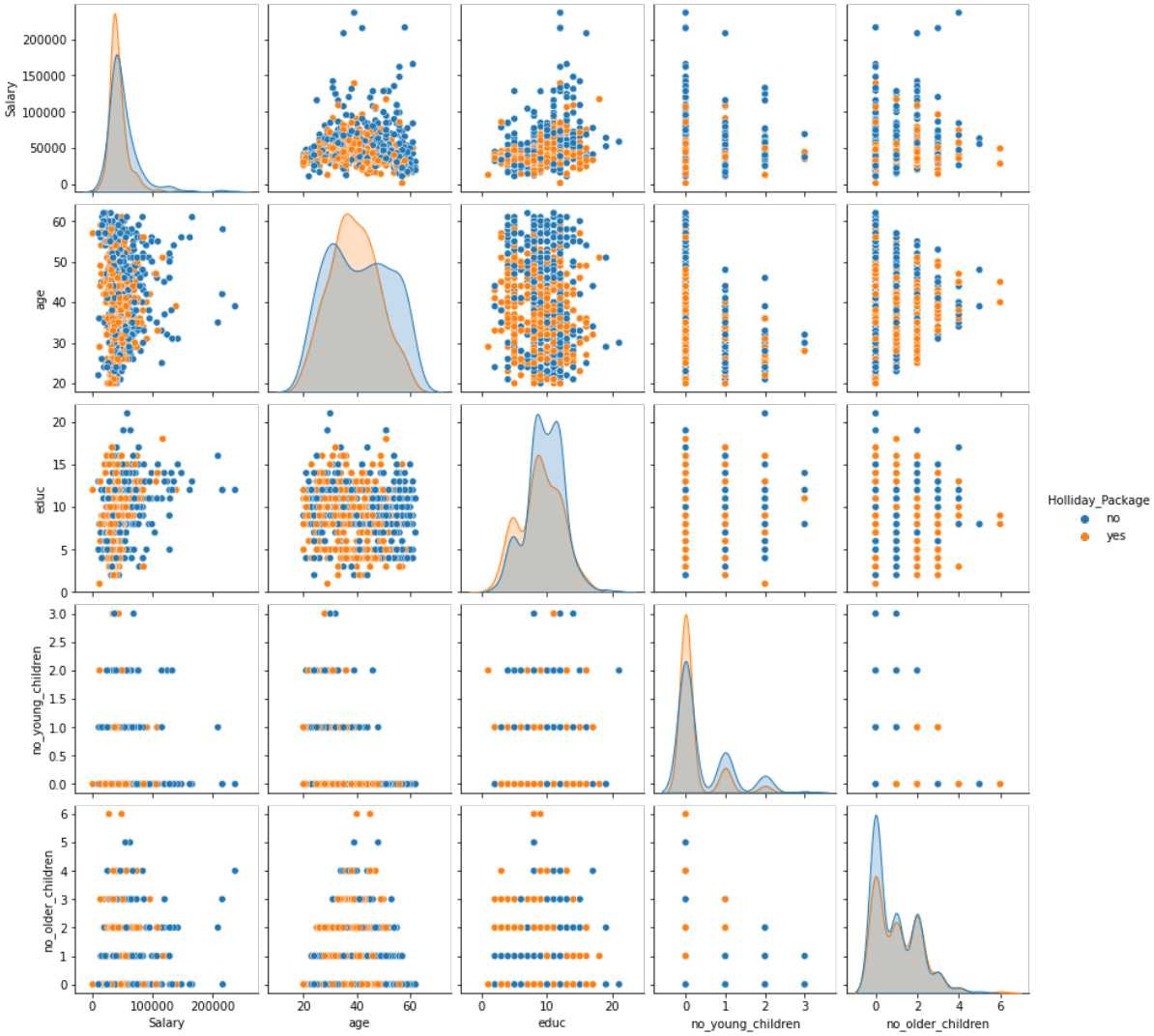
foreign vs Holiday_Package



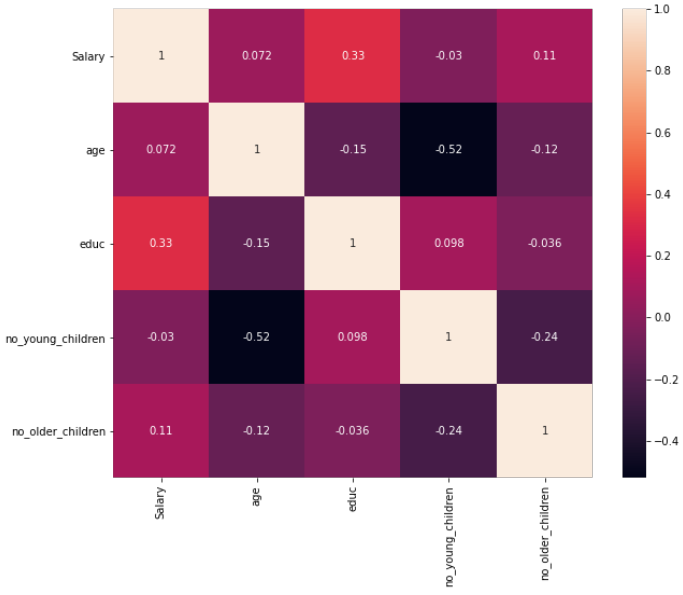
- Middle age group (30-50) tend to buy the holiday package than the age groups
- There is a greater chunk in far lower and far higher years of education groups tend to buy the holiday package than who do not.
- Foreign nationals have a higher tendency to buy the holiday package than non-foreign nationals.

Multi-variate analysis

Pair Plot:



Heatmap:



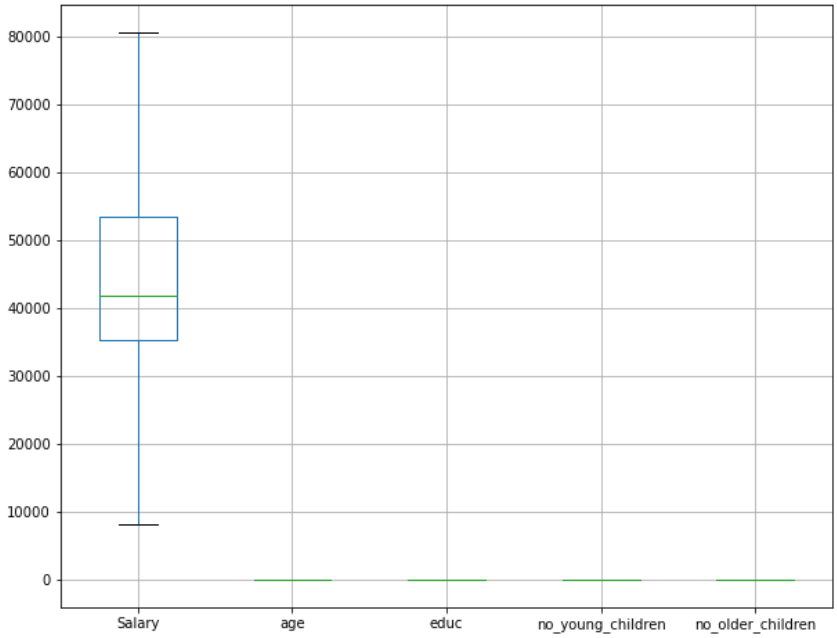
- In the diagonal distribution graphs, the graphs for both the classes are almost overlapping and we can see that all of them are week predictors of the target variable.
- There is very little correlation between the predictor variables.

Validation

The age must be greater than the number of year educated. All the data looks to be in line to this condition.

Outlier Treatment

We will replace all the outliers with the whisker value that is closest to it.
Post outlier treatment:



Encoding

The columns Holliday_Package and foreign are object data type we will convert them into 1 and 0 so that it is converted into numerical data type.

Since foreign is not ordinal we need to create dummy variables. However there are two classes, yes and no, hence it will not make any difference with or without dummy variables.

Modelling

We will proceed by splitting the provided data into train and test data in the ratio of 70:30.

Logistic Regression:

Upon grid search we found that the below are the best parameters thus obtained:

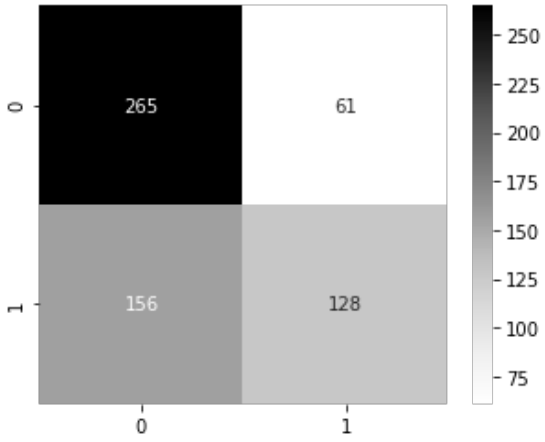
```
GridSearchCV(cv=3, estimator=LogisticRegression(n_jobs=2), n_jobs=-1,
             param_grid={'max_iter': [100, 1000, 10000],
                         'penalty': ['15', 'l2', '6'],
                         'solver': ['saga', 'sag', 'lbfgs', 'newton-cg'
,
                                    'liblinear'],
```

```
                    'tol': [0.0001, 1e-05]},
             scoring='f1')
```

The model was evaluated as shown below:

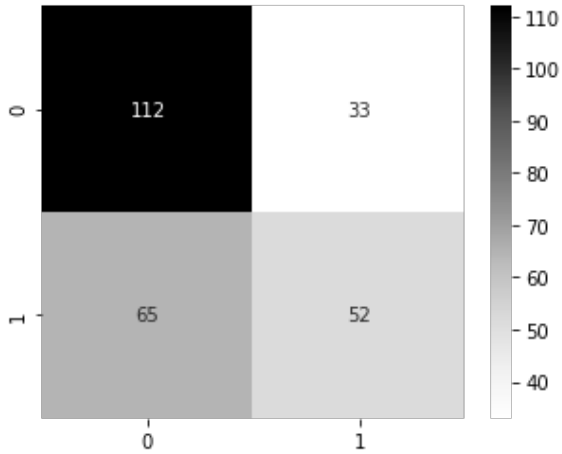
Logistic Regression Model (Train Data):

Confusion Matrix:

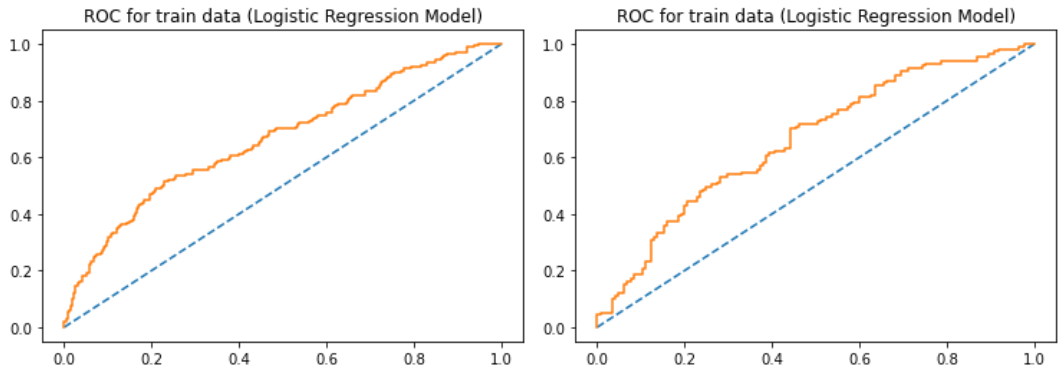


Logistic Regression Model (Test Data):

Confusion Matrix:



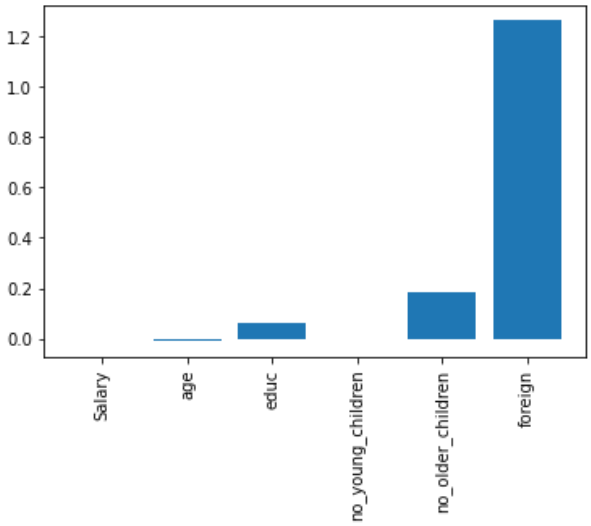
The ROC for the model is found to be:



The model isn't over fitted but the parameters are not very satisfactory. The F1 and accuracy are very low. Let us check if LDA performs better.

Feature importance

```
Feature importance for logistic regression:
Feature: Salary, Score: -0.00002
Feature: age, Score: -0.00957
Feature: educ, Score: 0.06151
Feature: no_young_children, Score: 0.00000
Feature: no_older_children, Score: 0.18378
Feature: foreign, Score: 1.26139
```

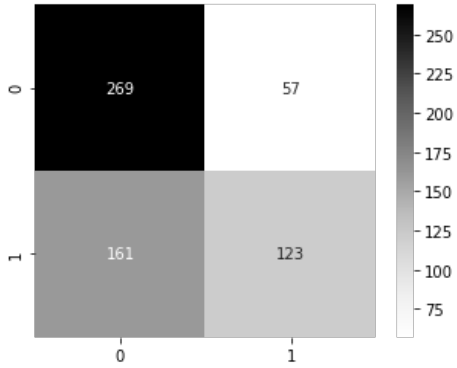


Linear Discriminant Analysis (LDA) :

For a simple model with the default threshold of 0.5, the below is the performance:

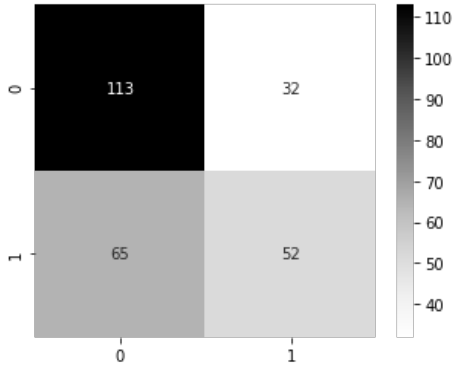
LDA Model (Train Data):

Confusion Matrix:

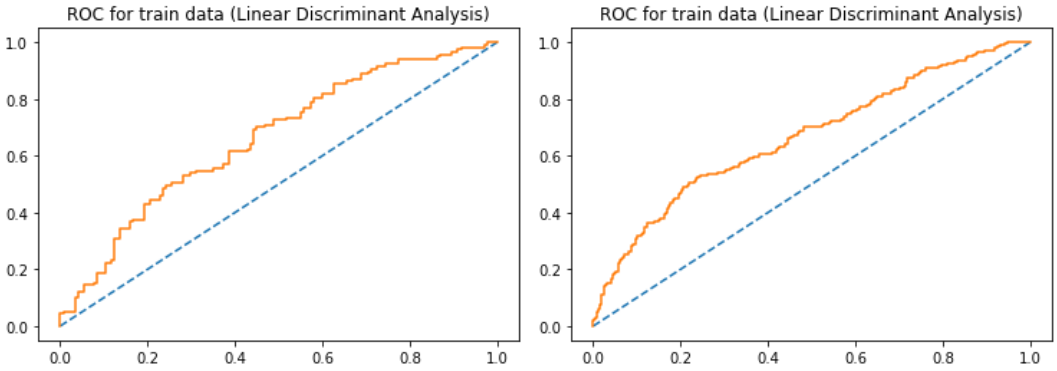


LDA Model (Test Data):

Confusion Matrix:



The ROC for the model is found to be:

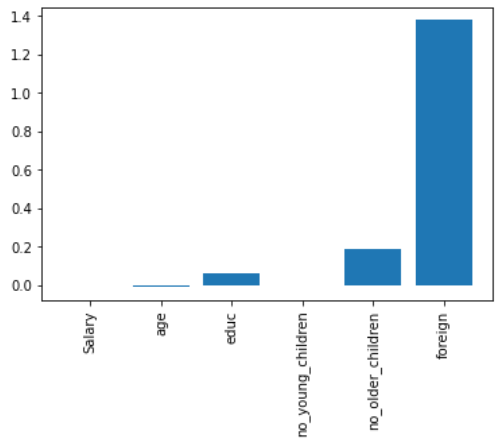


Model Performance for LDA:

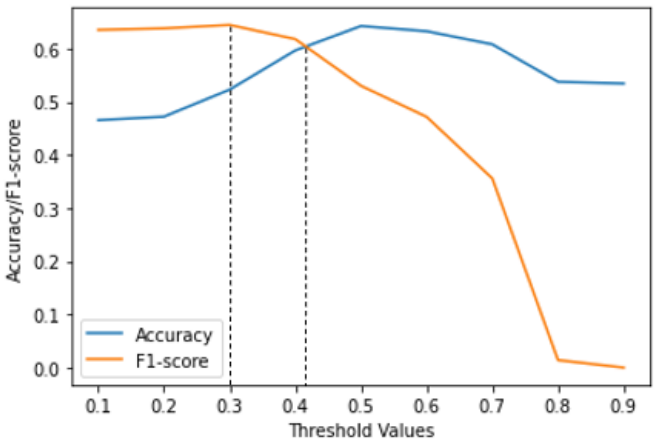
	Logistic Regression (Train)	Logistic Regression (Test)
Accuracy	64	63
AUC	67	66
Recall	45	44
Precision	68	61
F1 Score	54	51

The parameters are almost the same as logistic regression. There is a slight improvement in the scores.

```
Feature importance for logistic regression:
Feature: Salary, Score: -0.00002
Feature: age, Score: -0.00900
Feature: educ, Score: 0.06514
Feature: no_young_children, Score: -0.00000
Feature: no_older_children, Score: 0.18747
Feature: foreign, Score: 1.37651
```



Since the default value of threshold or max likelihood for LDA is set at 0.5, we will explore different values (from 0.1, 0.2,…., 1) and obtain the best out of them all. We then analysed for various values of threshold and found that around 0.4 is the threshold that needs to be maintained so as to get a good F1 score and a good precision.



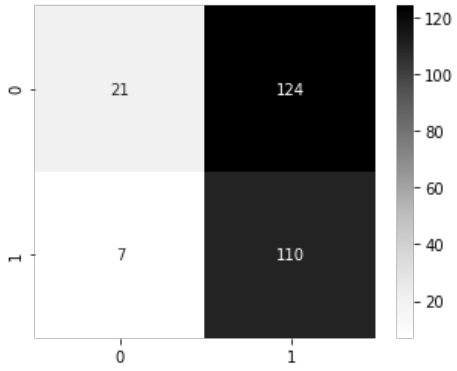
From the above graph we can take 0.4 as the threshold but at 0.4 the value of accuracy and F1-score is at a trade off point. However, there is a considerable drop in value at 0.4. So 0.3 is a much preferred threshold as there is a small peak at 0.3. The accuracy at 0.3 is less but the F1-score is high and is much preferred as we need to find class 1 predictions (people who take up the holiday package).

The model parameters obtained after custom threshold for classification of 0.3 are:

LDA Model Custom Cut-off=0.3 (Train Data):

When the cut-off probability is 0.3

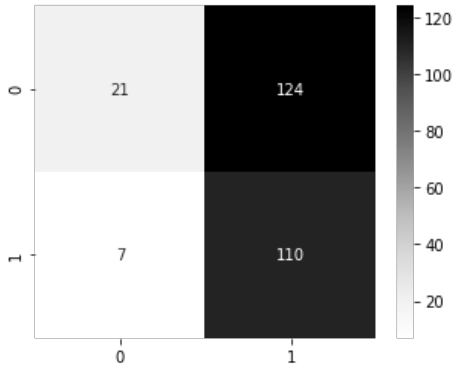
Confusion Matrix:



LDA Model Custom Cut-off=0.3 (Train Data):

When the cut-off probability is 0.3

Confusion Matrix:



Model Performance for LDA (threshold = 0.3):

	Linear Discriminant Analysis [Thresh>0.3] (Train)	Linear Discriminant Analysis [Thresh>0.3] (Test)
Accuracy	52	50
AUC	67	66
Recall	93	94
Precision	49	47
F1 Score	64	63

Conclusion:

The summary of all the models are:

	Logistic Regression (Train)	Linear Discriminant Analysis (Train)	Linear Discriminant Analysis [Thresh>0.3] (Train)
Accuracy	64	64	52
AUC	67	67	67
Recall	45	43	93
Precision	68	68	49
F1 Score	54	53	64

	Logistic Regression (Test)	Linear Discriminant Analysis (Test)	Linear Discriminant Analysis [Thresh>0.3] (Test)
Accuracy	63	63	50
AUC	66	66	66
Recall	44	44	94
Precision	61	62	47
F1 Score	51	52	63

All three models are neither overfitted nor are they underfitted.
When compared a Logistic regression model and a simple LDA model we can see that both the models are performing equally well.
When a customized threshold is used in LDA, the model is performing equally well in training and testing data set. This model has a drop in accuracy but the F1 score and recall has improved to a great extent. Amongst the three, LDA with a threshold of 0.3 is performing the best for the given data set.

Recommendations:

- The order of top 3 important features are:
  a. Foreign
  b. Number of older children
  c. Number of years education
- To target the foreign travellers we must increase the marketing in airports, railway stations and travel agencies.
- We also should target more on the middle aged working professions who are more likely to be educated and having an older child.
- Since the data has a no:yes ratio of 54:45, for $1000 of an investment and a customer acquisition cost of $10. We will get only 45 customers who will buy the package.
  If the model is used and the predicted set of customers were provided with the package offer, the churn will be 94 people given the same above situation.