LINEAR REGRESSION & LOGISTIC REGRESSION & LINEAR DISCRIMINANT ANALYSIS PROJECT

NAME – VIVEK AUGUSTINE

DATE – 7-31-22

Table of Contents :

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.

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

Checking Null Values :

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

Datatypes:

```
Unnamed: 0      int64
carat         float64
cut            object
color          object
clarity        object
depth         float64
table         float64
x             float64
y             float64
z             float64
price           int64
```

Dataframe 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
```

Observation: 1.The data set contains 26967 row, 11 columns .

2.In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable.

3.The first column is an index ("Unnamed: 0")as this only serial no, we can remove it.

4.Except depth, in all the column the count is 26967.

Dropped “Unnamed: 0” Column because it will be of no use to us

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

Describing data for insights:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

The first three observations that caught my eye were the x,y,z variables their mean value is 0 which is bit strange to see as we know dimensionless or 2-dimensional diamonds are not possible. So,we are going to drop these values.Carat and price have a slightly distinct nature in terms of their mean and median value resulting in slight skweness also.

Number of zeros present in rows “X”, “Y”, “Z” :

```
Number of rows with x == 0: 3
Number of rows with y == 0: 3
Number of rows with z == 0: 9
```

No of Duplicate present :

```
No of Duplicate present is {} = 33
```

Shape after dropping duplicates :

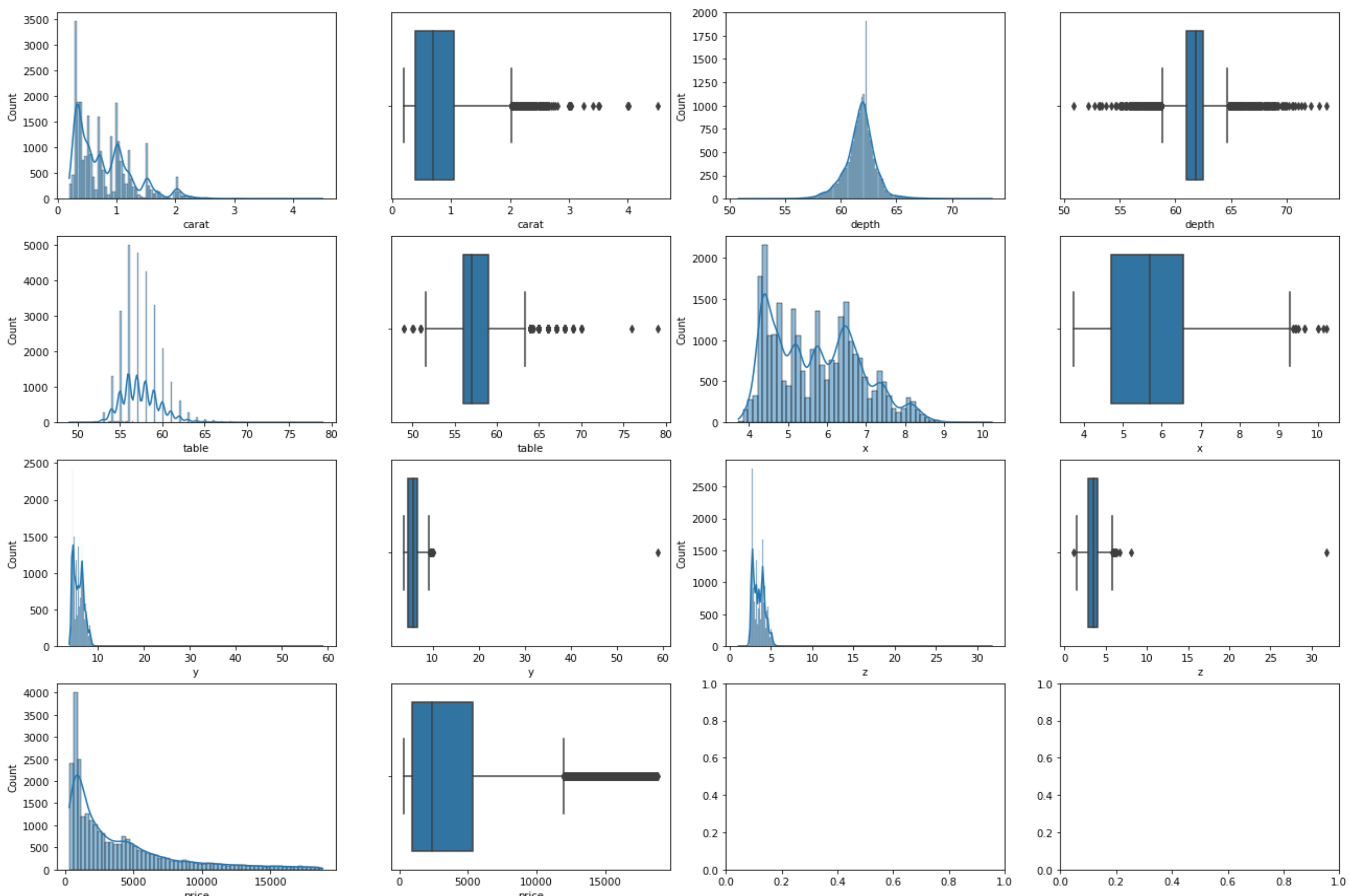
```
(26925, 10)
```

Unique Counts of our categorical variable :

```
CUT :  5
Fair          779
Good         2434
Very Good    6027
Premium      6880
```

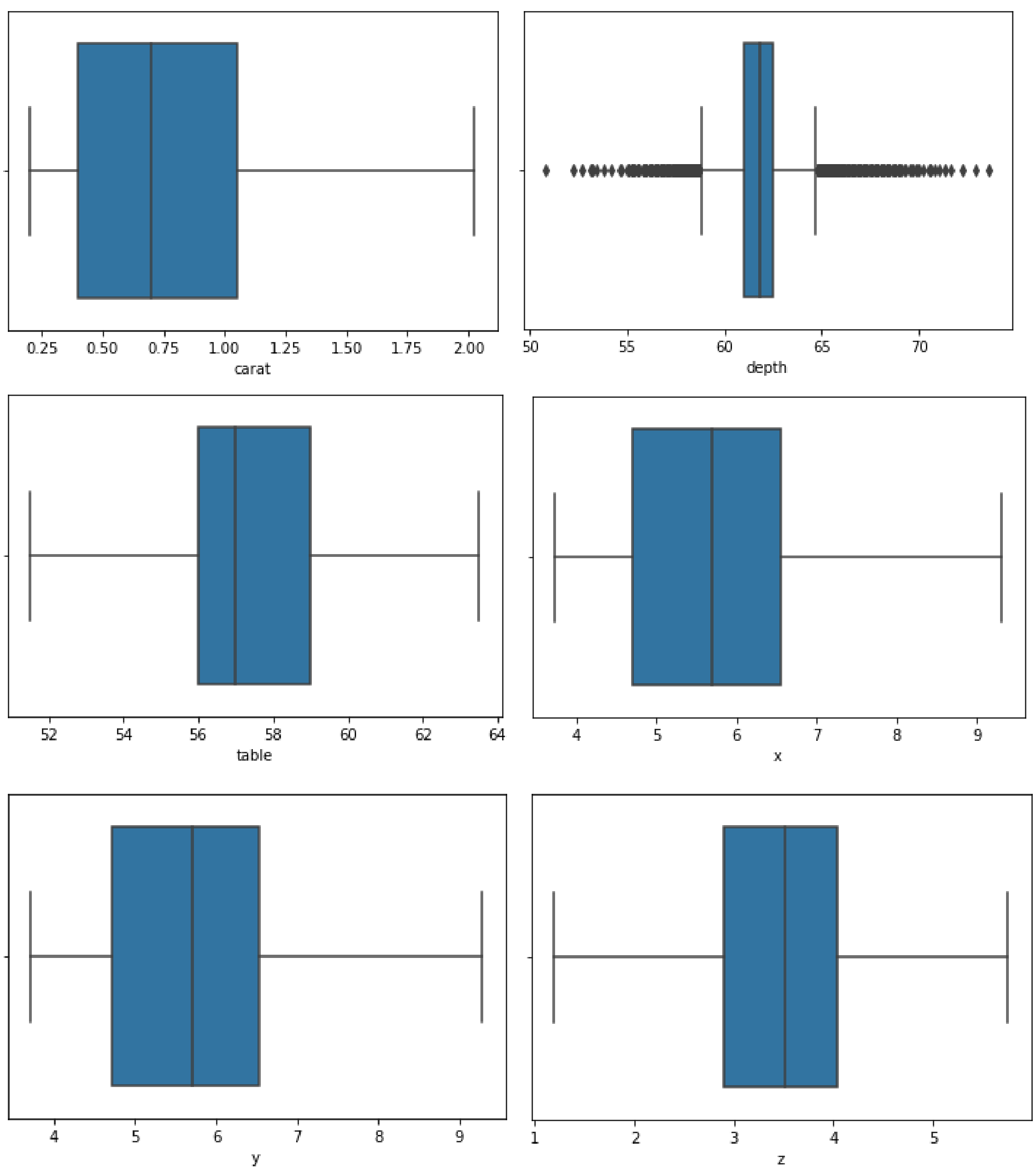
```
Ideal        10805


COLOR :  7
J     1440
I     2765
D     3341
H     4091
F     4722
E     4916
G     5650


CLARITY :  8
I1        362
IF        891
VVS1     1839
VVS2     2530
VS1      4086
SI2      4561
VS2      6092
SI1      6564
```

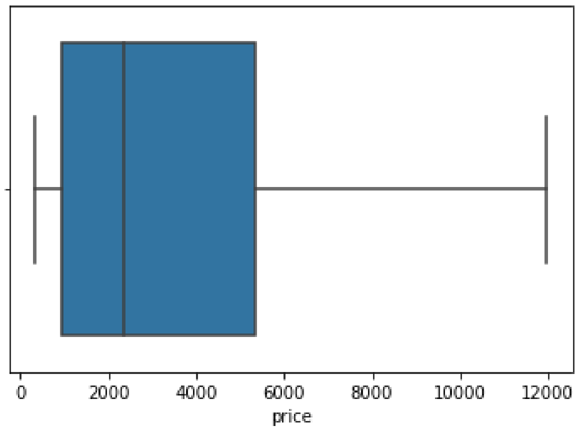
Univariate and Bivariate Analysis:



Here, we can see every independent variables have outliers in them and also we can see that the 2 variables have a distinct mean and median value from each other i.e they are not close to each other those variable are carrat and price, for the other independent variables we can see that their mean and median values are almost close with each other because the two variable i.e carat and price their mean and median value are distinct we can get an idea that there must be skweness present in both variables.

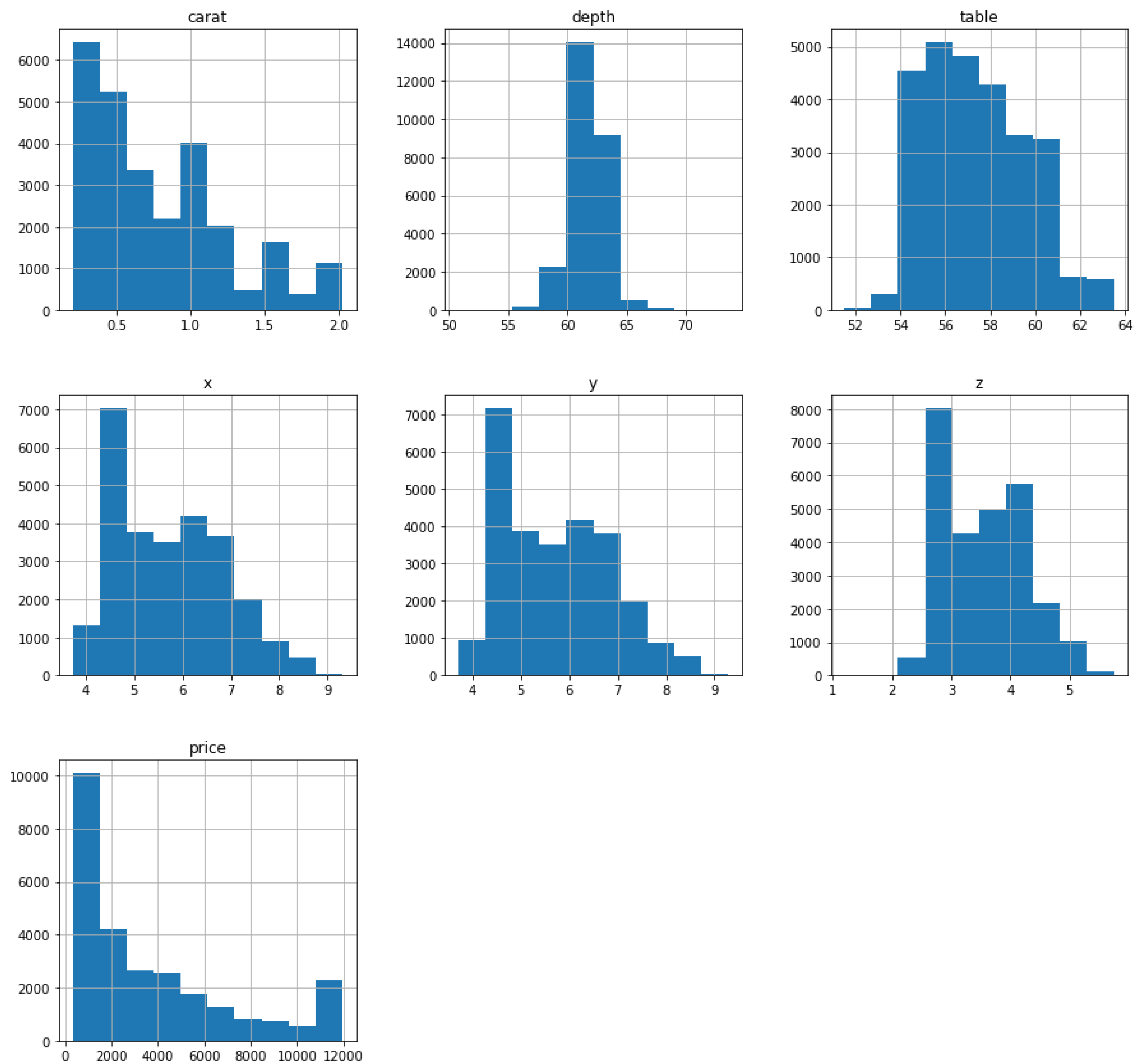
Treating Outliers :





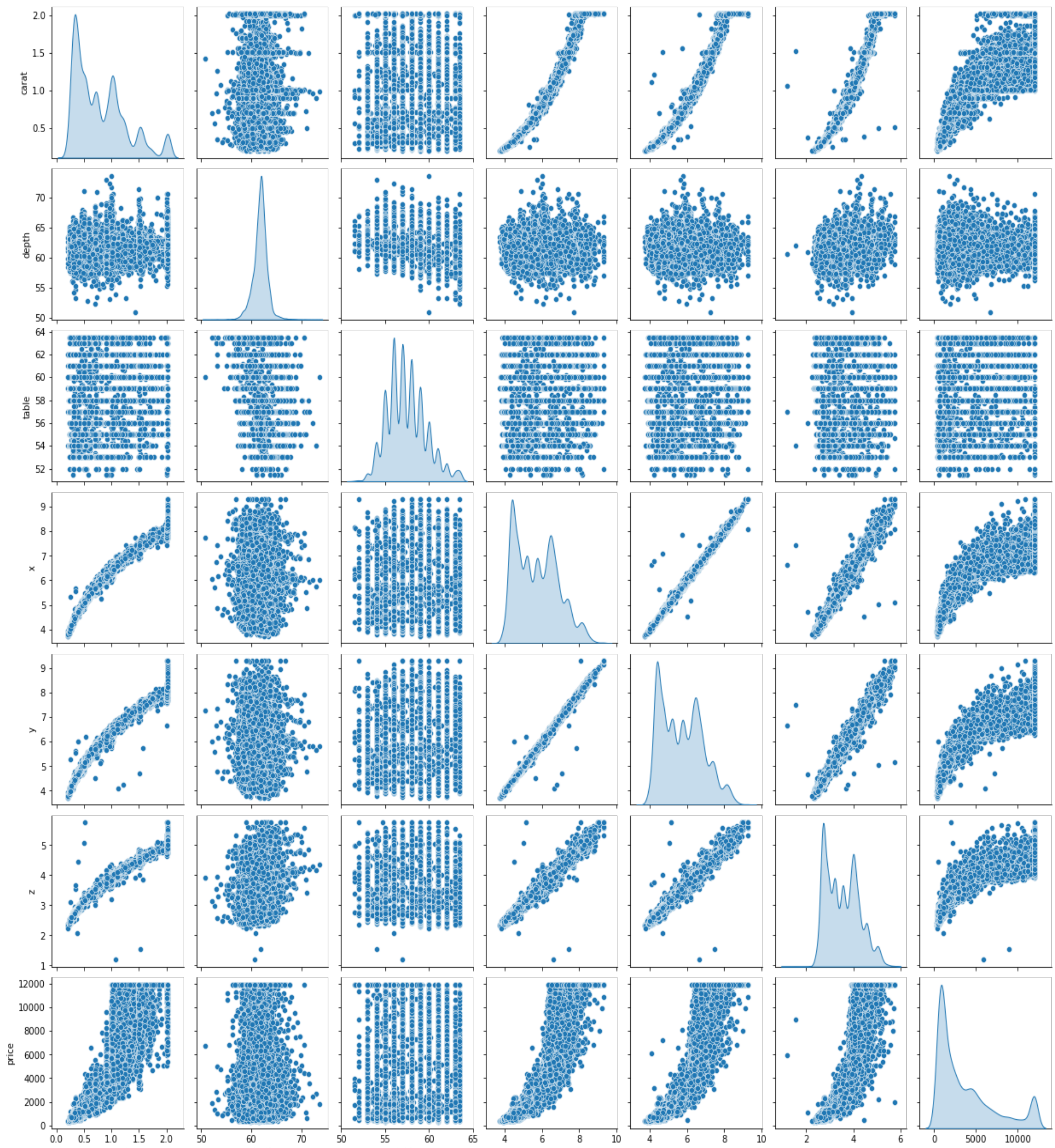
Only depth variable has outlier in it otherwise all the variables have been treated.

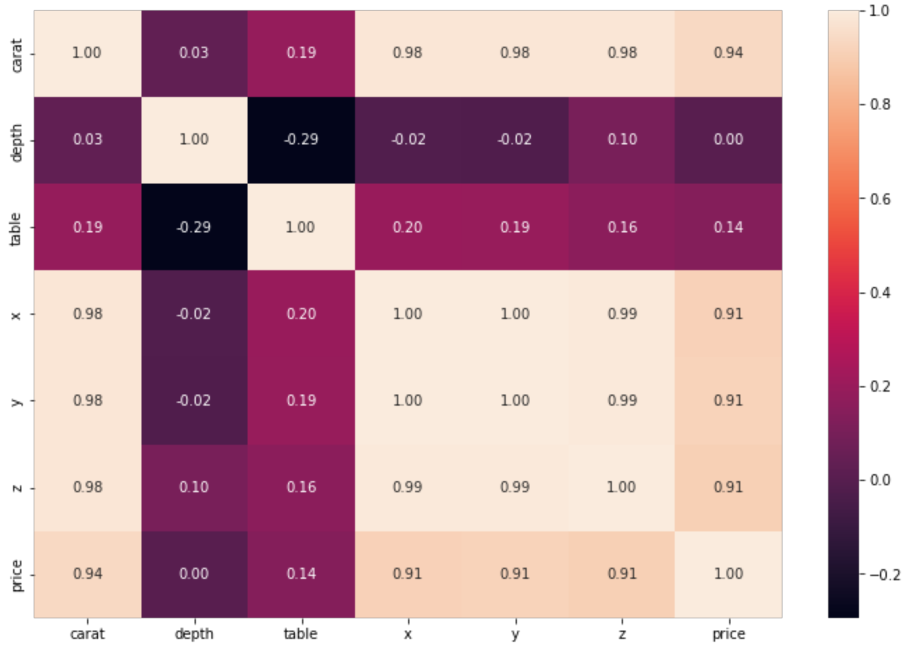
Histograms :



Here, we can see two variable are skewed i.e. Carat and Price.

Pairplot :





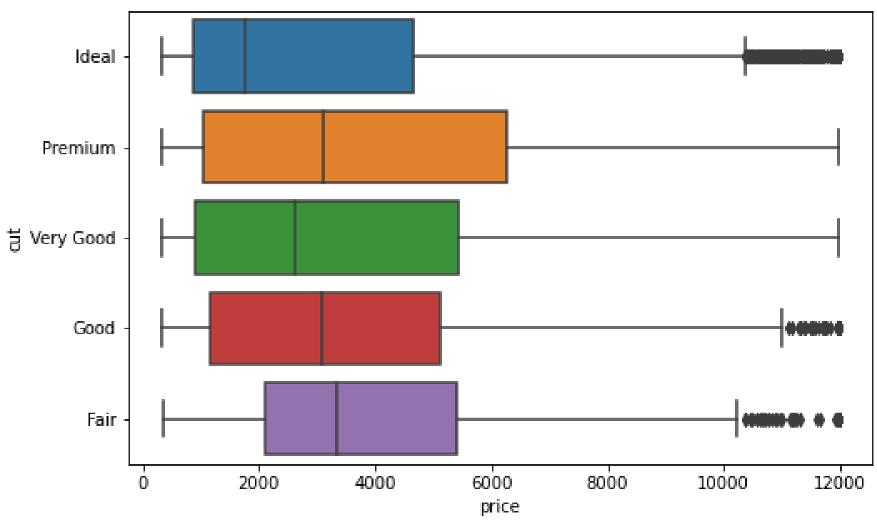
Observations:

➢ High correlation between the different features like carat, x, y, z and price.

➢ Less correlation between table with the other features.

➢ Depth is negatively correlated with most the other features except for carat.

EDA for Categorical variable:



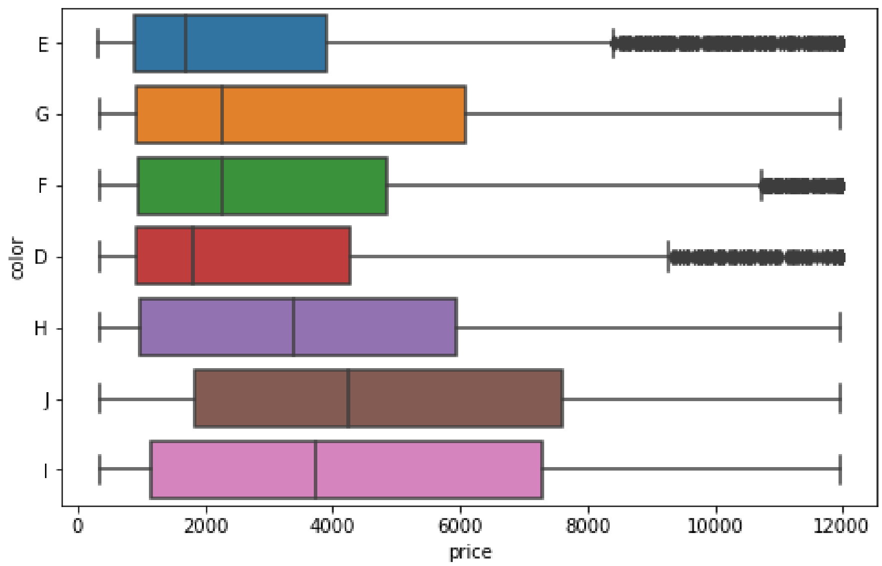
Price with respect to cut.

1. Here, we can know the price of every cut type gem and can know which one is cheap and which one is expensive.

2. We can see here that ideal cut type gem is the cheapest one and Premium cut type gem is the most expensive one.

3. Every cut type gems have outliers with them with respect to price leaving Premium & Very Good.

4. Each cut type gems price is decided by their quality or segment i.e. Ideal, Premium, Very Good, Good, Fair.

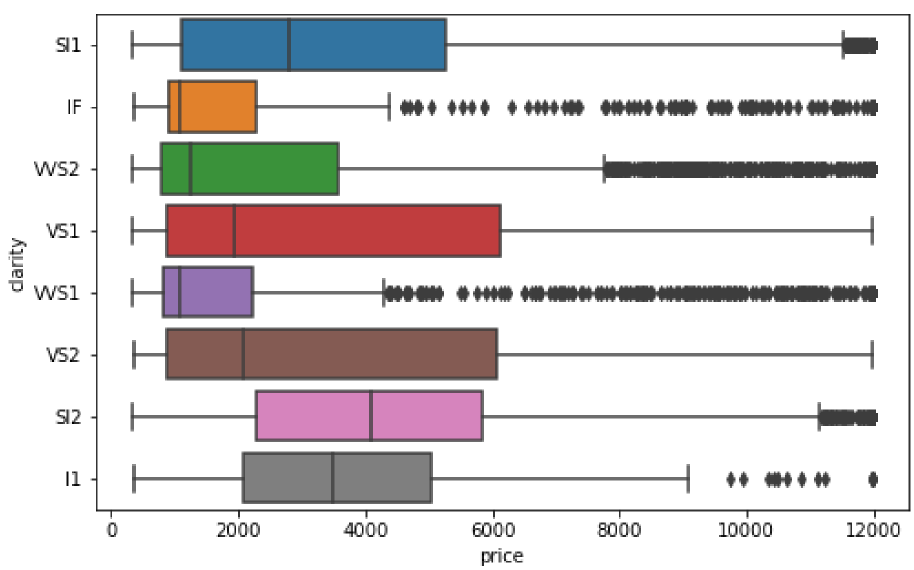


Price with respect to color.

1.Here we can see that E coloured gem is the cheapest one and J & I colored gem are the most expensive one.

2.Three colored gems have outliers i.e E, F, D leaving the rest.This insight is with respect to price.

3.Color is having some kind of influence or can say indirect influence on prices of the gems.



Price with respect to clarity.

1.Each segment have its own price according to the quality or demand of it.

2.IF is the cheapest one here while VS1 and VS2 are the most expensive one.

3.Leaving VS1 and VS2 each category have outliers with it.

CONCLUSION OF EDA:

• Price – This variable gives the continuous output with the price of the cubic zirconia stones. This will be our Target Variable.

• Carat, depth, table, x, y, z variables are numerical or continuous variables.

• x, y, z variables had mean value 0 which was a faulty value so we dropped them and our row came as 26958 & 10 columns.

• Cut, Clarity and colour are categorical variables.

• We will drop the first column 'Unnamed: 0' column as this is not important for our study.

• Only in 'depth 697 missing values are present which we will impute by its median values.

• There are total of 33 duplicate rows as computed using. Duplicated () function. We will drop the duplicates.

• Upon dropping the duplicates – The shape of the data set is – 26925 rows & 10 columns.

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.

Filling the 697 NA values by median values :

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

Checking for the values which are equal to zero. We have alrady checked for 'Zero' values and we can observe there are some amount of 'Zero' value present on the data set on variable 'x = 3', 'y = 3','z = 9'.

This indicates that they are faulty values.

As we know dimensionless or 2-dimensional diamonds are not possible. So we have filter out those as it clearly faulty data entries.

Value counts of categorical variable :

```
cut
 Ideal        10805
Premium       6880
Very Good     6027
Good          2434
Fair           779


color
 G    5650
E    4916
F    4722
H    4091
D    3341
I    2765
J    1440


clarity
 SI1     6564
VS2     6092
SI2     4561
VS1     4086
VVS2    2530
VVS1    1839
IF       891
I1       362
```

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.

Encoded the variables and changed their types for further calculations :

```
carat       float64
cut         float64
color       float64
clarity     float64
depth       float64
table       float64
x           float64
y           float64
z           float64
price       float64
```

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4.0	1.0	2.0	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	3.0	3.0	7.0	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	2.0	1.0	5.0	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	4.0	2.0	4.0	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	4.0	2.0	6.0	60.4	59.0	4.35	4.43	2.65	779.0

Splitted the data :

	carat	cut	color	clarity	depth	table	x	y	z
0	0.30	4.0	1.0	2.0	62.1	58.0	4.27	4.29	2.66
1	0.33	3.0	3.0	7.0	60.8	58.0	4.42	4.46	2.70
2	0.90	2.0	1.0	5.0	62.2	60.0	6.04	6.12	3.78
3	0.42	4.0	2.0	4.0	61.6	56.0	4.82	4.80	2.96
4	0.31	4.0	2.0	6.0	60.4	59.0	4.35	4.43	2.65

Fitted the linear-regression model and took out co-efficients :

```
The coefficient for carat is 8901.94122507089
The coefficient for cut is 109.18812485149377
The coefficient for color is -272.92132964490315
The coefficient for clarity is 436.4411042154908
The coefficient for depth is 8.236971791613918
The coefficient for table is -17.345170384368316
The coefficient for x is -1417.9089304449476
The coefficient for y is 1464.827270146809
The coefficient for z is -711.225032681408
```

Observation :

Y=mx +c (m= m1,m2,m3...m9) here 9 different co-efficients will learn aling with the intercept which is "c" from the model.

From the above coefficients for each of the independent attributes we can conclude:

The one unit increase in carat increases price by 8901.941.

The one unit increase in cut increases price by 109.188.

The one unit increase in clarity increases price by 436.441.

The one unit increase in depth increases price by 8.236,

The one unit increase in y increases price by 1464.827.

But The one unit increase in table decreases price by -17.345,

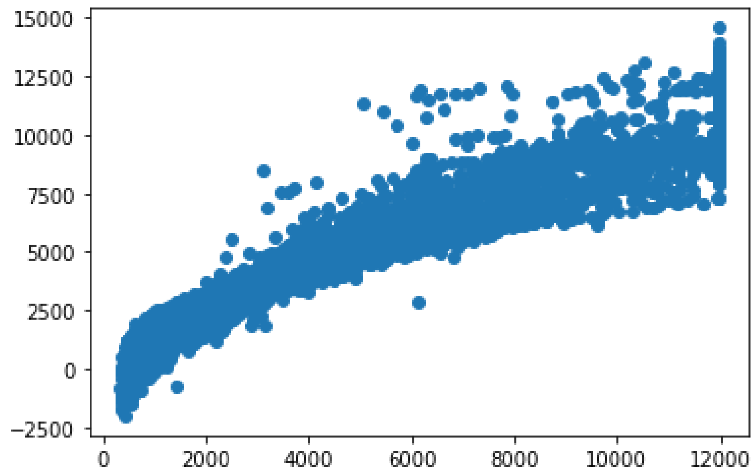
The one unit increase in color decreases price by -272.921,

The one unit increase in x decreases price by -1417.908,

The one unit increase in z decreases price by -711.225.

Linear regression Performance Metrics:

1. intercept for the model: -1534.4224694382478
2. R square on training data: 0.9311935886926559
3. R square on testing data: 0.931543712584074
4. RMSE on Training data: 907.1312415459143
5. RMSE on Testing data: 911.8447345328436
6. Our adjusted R-Squared is 0.9312771758547017



Interaction of values in our predicted Y.

Finding Co-efficients with statsmodels :

```
Intercept   -2334.381945
carat        8832.855317
cut           104.669525
color        -273.825428
clarity       434.952436
depth          18.369628
table         -15.235849
x           -1220.409362
y            1410.269805
z            -906.253902
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 4.055e+04
Date:                Sun, 31 Jul 2022   Prob (F-statistic):               0.00
Time:                        19:46:40   Log-Likelihood:            -2.2161e+05
No. Observations:               26925   AIC:                         4.432e+05
Df Residuals:                   26915   BIC:                         4.433e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2334.3819    676.918     -3.449      0.001   -3661.176   -1007.588
carat       8832.8553     68.850    128.292      0.000    8697.906    8967.805
cut          104.6695      6.070     17.243      0.000      92.771     116.568
color       -273.8254      3.434    -79.736      0.000    -280.557    -267.094
clarity      434.9524      3.747    116.066      0.000     427.607     442.298
depth         18.3696      9.477      1.938      0.053      -0.205      36.944
table        -15.2358      3.256     -4.679      0.000     -21.618      -8.854
x          -1220.4094    101.732    -11.996      0.000   -1419.810   -1021.009
y           1410.2698    102.178     13.802      0.000    1209.995    1610.545
z           -906.2539    138.009     -6.567      0.000   -1176.758    -635.750
==============================================================================
```

```
Omnibus:                     3700.849   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14669.878
Skew:                           0.646   Prob(JB):                         0.00
Kurtosis:                       6.377   Cond. No.                     1.05e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
[2] The condition number is large, 1.05e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Applying zscore statsmodels

With this specific dataset, I don't think we need to scale the data, however, to see its impact, lets quickly view the results post scaling the data. I have used Z score to scale the data. Z-Scores become comparable by measuring the observations in multiples of the standard deviation of that sample. The mean of a z-transformed sample is always zero. from scipy.stats import zscore

x_train_scaled = x_train.apply(zscore)

x_test_scaled = x_test.apply(zscore)

y_train_scaled = y_train.apply(zscore)

y_test_scaled = y_test.apply(zscore)

Co-efficients :

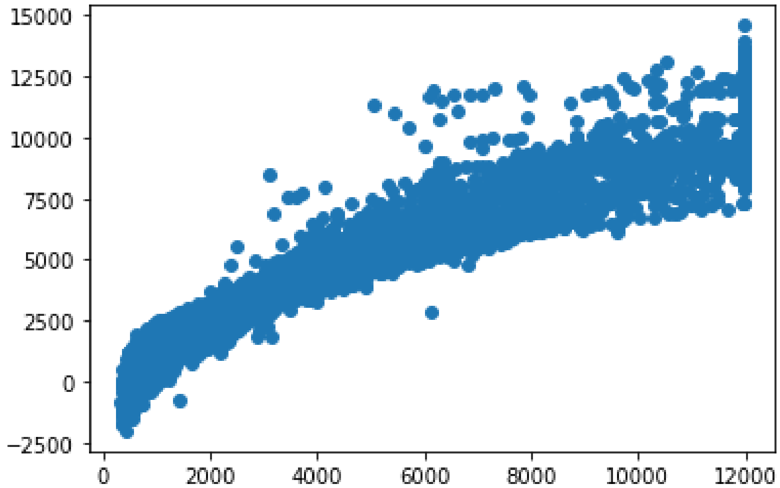
```
The coefficient for carat is 1.1837737061779434
The coefficient for cut is 0.035125000655529742
The coefficient for color is -0.13449269287641508
The coefficient for clarity is 0.20809779325621863
The coefficient for depth is 0.0033262937188390045
The coefficient for table is -0.010815851633643205
The coefficient for x is -0.459689842412527
The coefficient for y is 0.4716627091792411
The coefficient for z is -0.14249737973827153
```

Table :

	carat	cut	color	clarity	depth	table	x	y	z	price
5030	1.10	1.0	1.0	1.0	63.3	56.0	6.53	6.58	4.15	4065.0

	carat	cut	color	clarity	depth	table	x	y	z	price
12108	1.01	2.0	0.0	1.0	64.0	56.0	6.30	6.38	4.06	5166.0
20181	0.67	1.0	5.0	3.0	60.7	61.4	5.60	5.64	3.41	1708.0
4712	0.76	1.0	3.0	2.0	57.7	63.0	6.05	5.97	3.47	2447.0
2548	1.01	3.0	3.0	4.0	62.8	59.0	6.37	6.34	3.99	6618.0

```
OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 2.833e+04
Date:                Sun, 31 Jul 2022   Prob (F-statistic):               0.00
Time:                        19:46:40   Log-Likelihood:            -1.5510e+05
No. Observations:               18847   AIC:                         3.102e+05
Df Residuals:                   18837   BIC:                         3.103e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1534.4225    787.066     -1.950      0.051   -3077.143       8.298
carat       8901.9412     82.792    107.521      0.000    8739.661    9064.222
cut          109.1881      7.268     15.024      0.000      94.943     123.433
color       -272.9213      4.105    -66.478      0.000    -280.968    -264.874
clarity      436.4411      4.473     97.581      0.000     427.674     445.208
depth          8.2370     10.876      0.757      0.449     -13.080      29.554
table        -17.3452      3.904     -4.443      0.000     -24.998      -9.693
x          -1417.9089    136.590    -10.381      0.000   -1685.637   -1150.181
y           1464.8273    136.068     10.765      0.000    1198.122    1731.533
z           -711.2250    156.187     -4.554      0.000   -1017.366    -405.084
==============================================================================
Omnibus:                     2652.028   Durbin-Watson:                   2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9642.429
Skew:                           0.687   Prob(JB):                         0.00
Kurtosis:                       6.223   Cond. No.                     1.03e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly spec
ified.
[2] The condition number is large, 1.03e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```



Interaction of values in our predicted Y.

Mean Squared Error(Training) = 907.1312415459142

Mean Squared Error(Testing) = 911.8447345328435

R-squared: 0.931

Adj. R-squared: 0.931

CHECK MULTI-COLLINEARITY USING VIF

```
carat ---> 121.75332708925261
cut ---> 10.38788625226542
color ---> 3.729750905495899
clarity ---> 5.460420380299075
depth ---> 1219.3950498545585
table ---> 877.9704845924091
x ---> 10743.99485978316
y ---> 9475.980399634736
z ---> 3693.953245513562
```

1. Variance inflation factor (VIF) is used to detect the severity of multicollinearity in the ordinary least square (OLS) regression analysis.
2. Multicollinearity inflates the variance and type II error. It makes the coefficient of a variable consistent but unreliable.
3. VIF measures the number of inflated variances caused by multicollinearity.

VIF measures the intercorrelation among independent variables in a multiple regression model. In mathematical terms, the variance inflation factor for a regression model variable would be the ratio of the overall model variance to the variance of the model with a single independent variable. As an example, the VIF value for Carat in the table above is the intercorrelation with other independent variables in the dataset and so on for other variables

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

We have a database which have strong correlation between independent variables and hence we need to tackle with the issue of multicollinearity which can hinder the results of the model performance. Multicollinearity makes it difficult to understand how one variable influence the target variable. However, it does not affect the accuracy of the model. As a result while creating the model, I had dropped a lot of independent variables displaying multicollinearity or the ones with no direct relation with the target variable.

While we looked at the data during univariate analysis, we were able to establish that Carat is strongly related with the price variable, and also with a lot of other independent variables - x, y, and z, and low correlation with variables such as table and cut as well. It can be established that Carat will be a strong predictor in our model creation. The same trend was displayed even after the object columns were encoded. The carat variable continues to display strong to low correlation with most of the variables, making its claim to be the most important predictor firm.

Recommendations:

As expected Carat is a strong predictor of the overall price of the stone. Clarity refers to the absence of the Inclusions and Blemishes and has emerged as a strong predictor of price as well. Clarity of stone types IF, VVS_1, VVS_2 and vs1 are helping the firm put an expensive price cap on the stones. Color of the stones such H, I and J won't be helping the firm put an expensive price cap on such stones. The company should instead focus on stones of color D, E and F to command relative higher price points and support sales. This also can indicate that company should be looking to come up with new color stones like clear stones or a different color/unique color that helps impact the price positively. The company should focus on the stone's carat and clarity so as to increase their prices. Ideal customers will also contribute to more profits. The marketing efforts can make use of educating customers about the importance of a better carat score and importance of clarity index. Post this, the company can make segments, and target the customer based on their income/paying capacity etc, which can be further studied.

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.

Table :

	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

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
```

There are 0 null values and duplicates :

```
Unnamed: 0           0
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
```

Description of data :

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872.0	NaN	NaN	NaN	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Summary of the Dataset

• Holiday Package – This variable is a categorical Variable. output with the This will be our Target Variable.

• Salary, age, educ, no_young_children, no_older_children, variables are numerical or continuous variables.

• Salary ranges from 1322 to 236961. Average salary of employees is around 47729 with a standard deviation of 23418. Standard deviation indicates that the data is not normally distributed. skew of 0.71 indicates that the data is right skewed and there are few employees earning more than an average of 47729. 75% of the employees are earning below 53469 while 255 of the employees are earning 35324.

• Age of the employee ranges from 20 to 62. Median is around 39. 25% of the employees are below 32 and 25% of the employees are above 48. Standard deviation is around 10. Standard deviation indicates almost normal distribution.

• Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the data

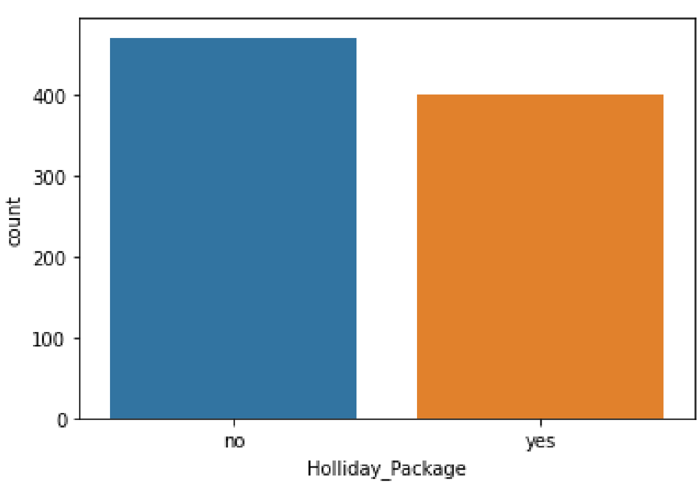
• Foreign is a categorical variable

• We have dropped the first column 'Unnamed: 0' column as this is not important for our study. Unnamed is a variable which has serial numbers so may not be required and thus it can be dropped for further analysisThe shape would be – 872 rows and 7 columns

• There are no null values

• There are no duplicates

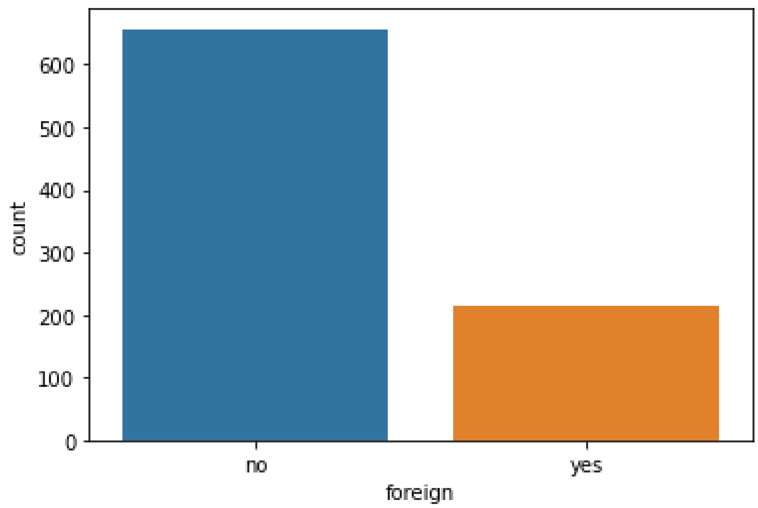
Checking the unique values for the target Variables 'Holiday Package":



We can observe that 54% of the employees are not opting for the holiday package and 46% are interested in the package. This implies we have a dataset which is fairly balanced

```
no      471
yes     401
```

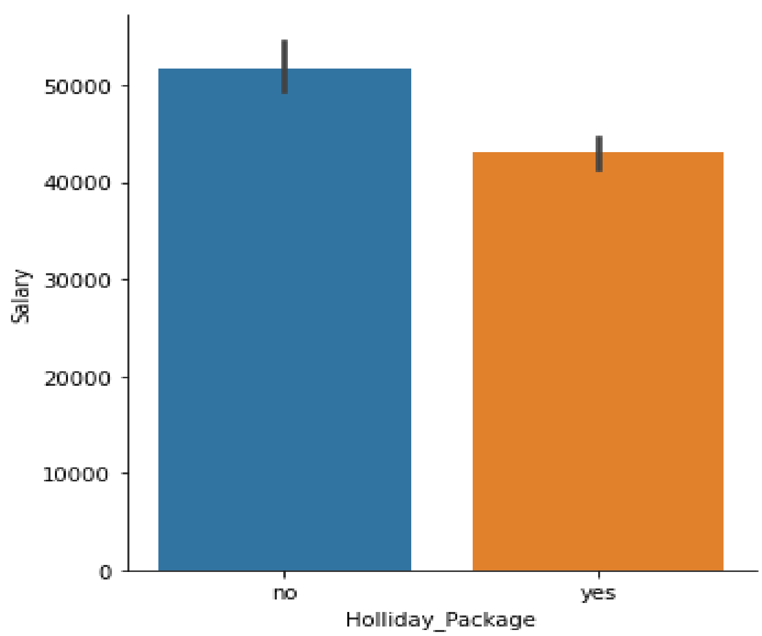
Checking the unique values of the Foreign Variables as it is categorical:



We can observe that 75% of the emp loyees are not Foreigners and 25% ar e foreigners

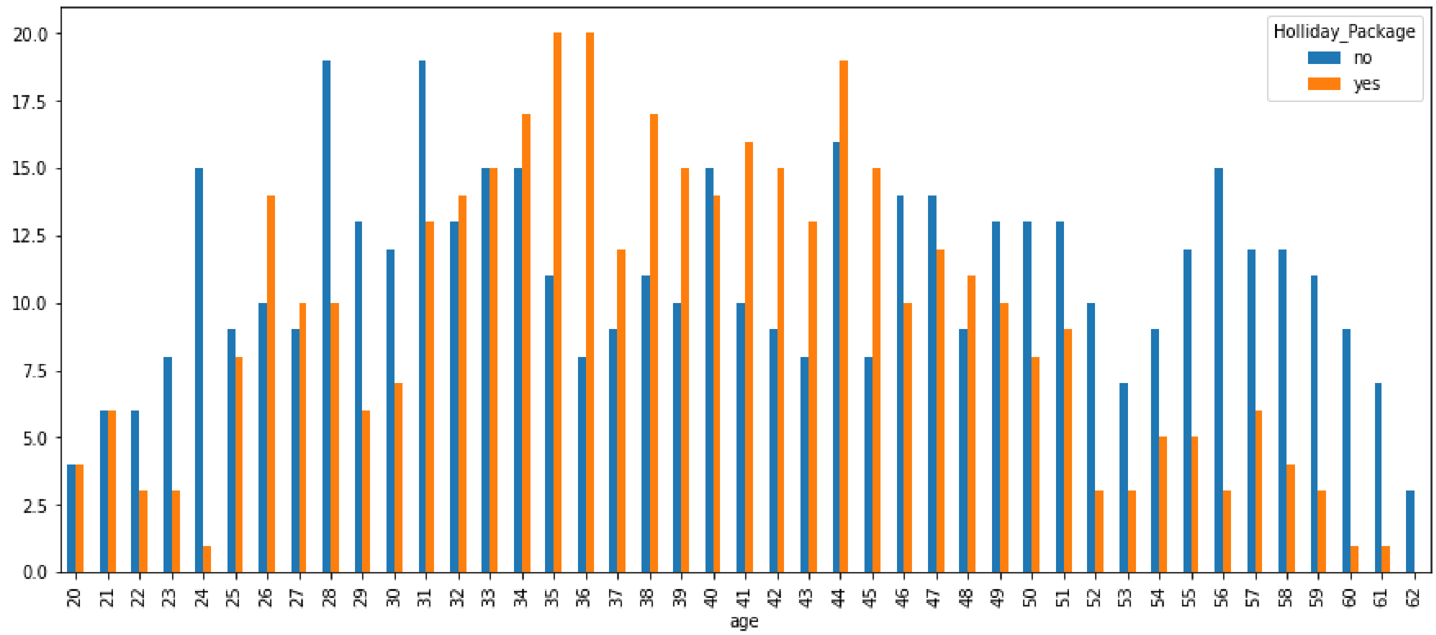
```
no      656
yes     216
```

SALARY :



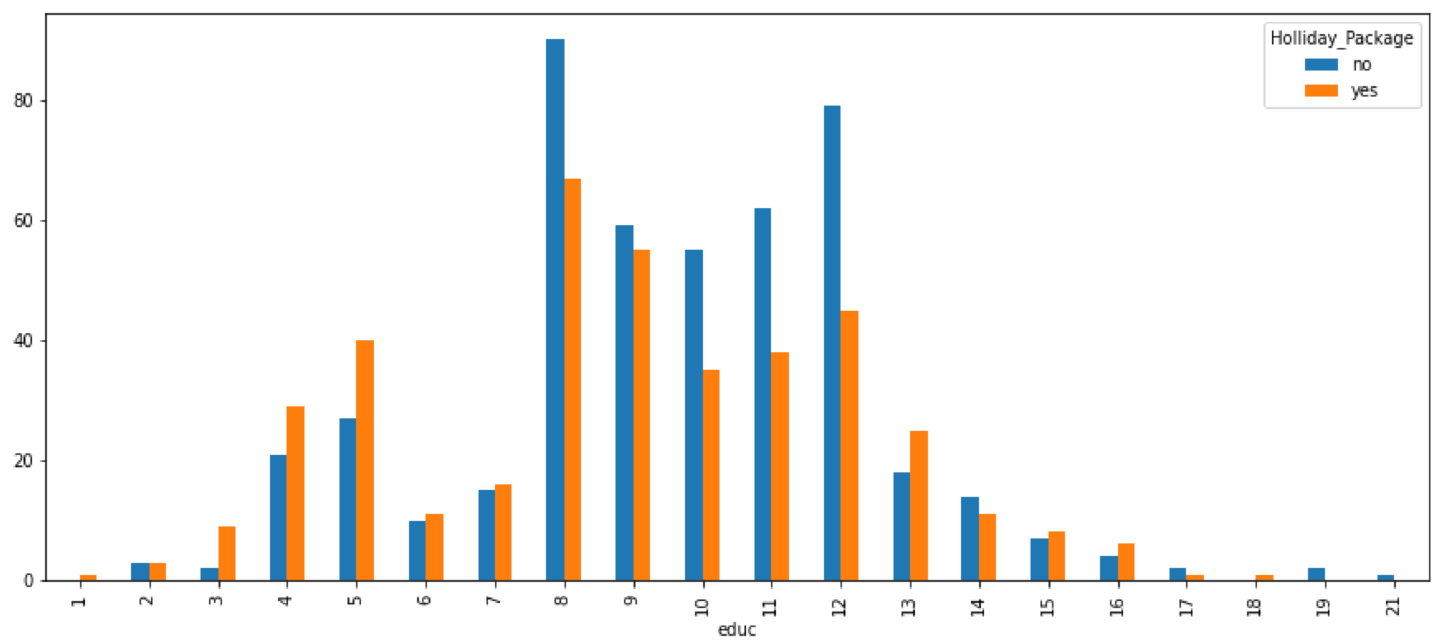
We observe that Salary for employees opting for holiday package and for not opting for holiday package is similar in nature. However, the distribution is fairly spread out for people not opting for holiday packages.

AGE :



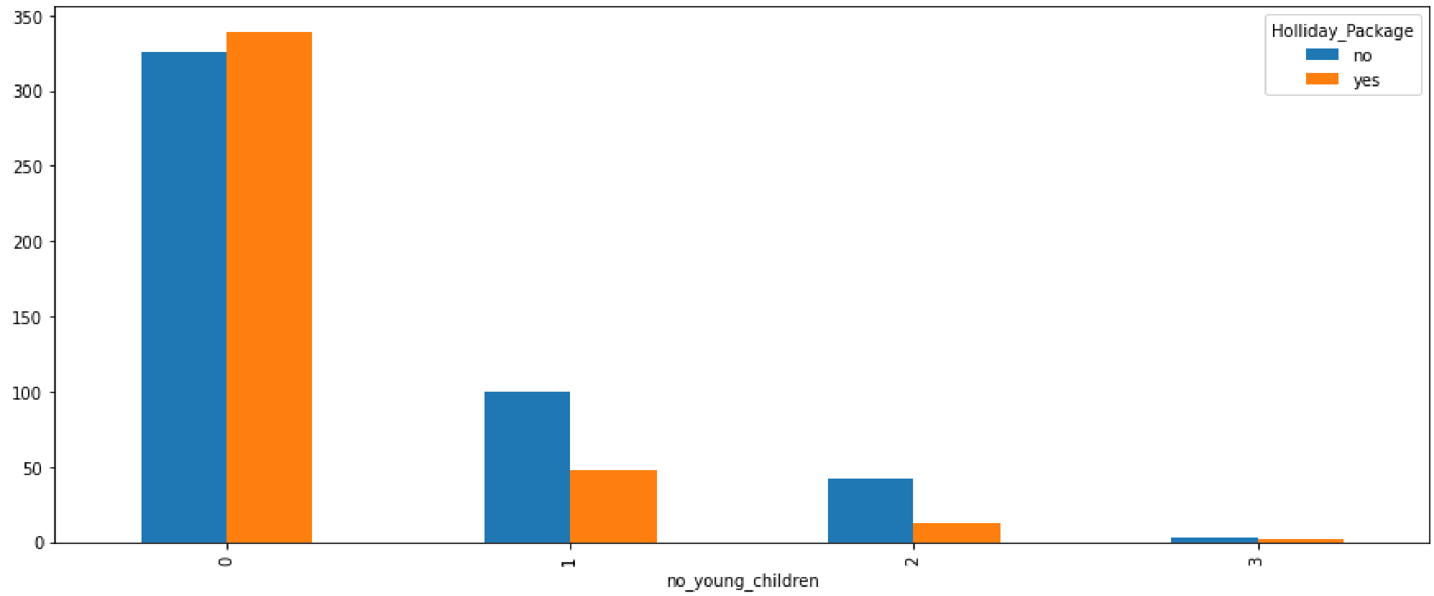
We can clearly see that employees in middle range (34 to 45 years) are going for holiday package as compared to older and younger employees

Education :



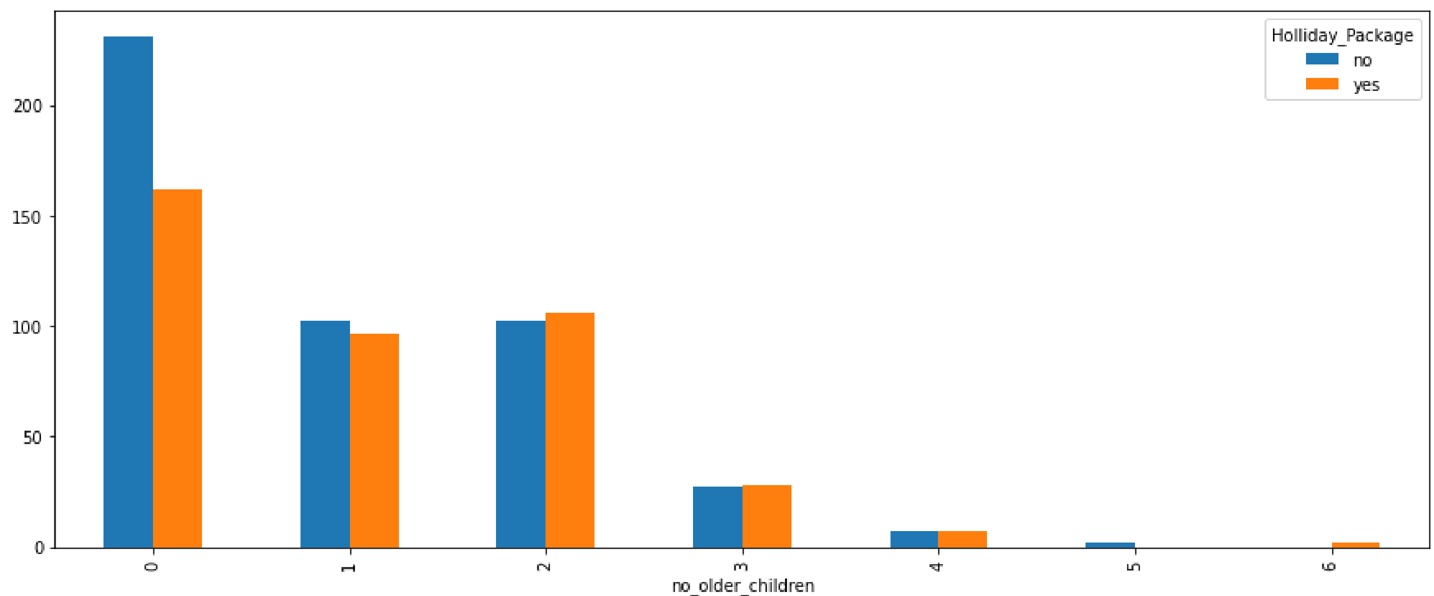
We observe that employees with less years of formal education(1 to 7 years) and higher education are not opting for the Holiday package as compared to employees with formal education of 8 year to 12 years.

No of young Children :



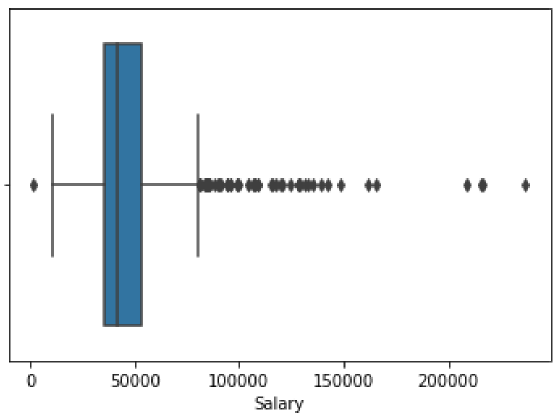
We can clearly see that people with younger children are not opting for holiday packages.

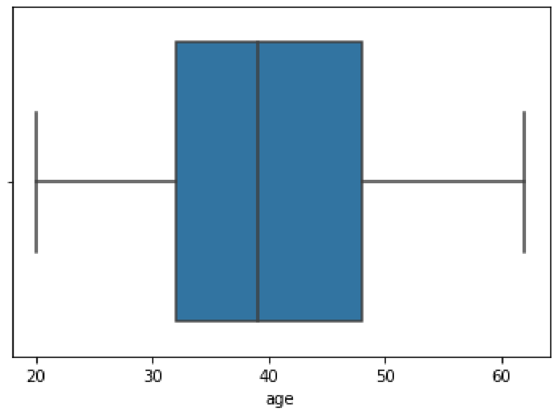
No of old children :

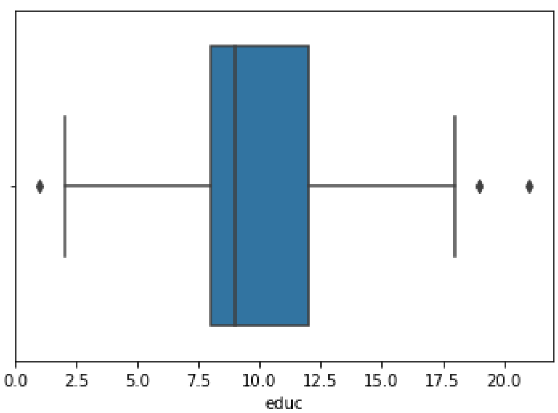


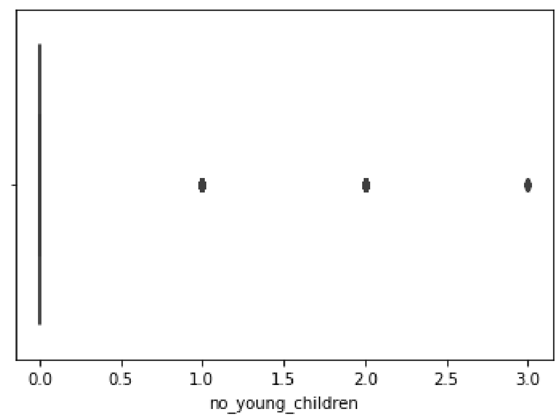
Almost same distribution for both the scenarios when dealing with employees with older children

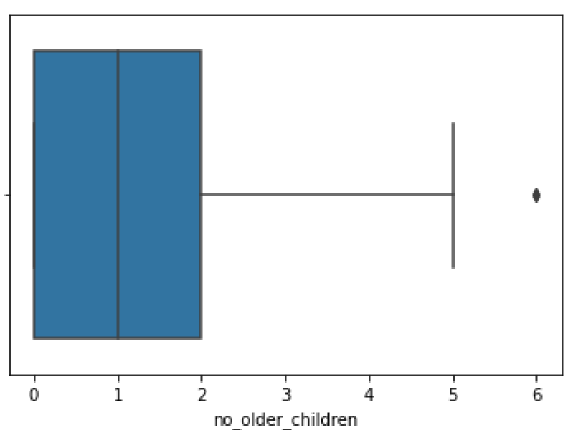
Checking Outliers :





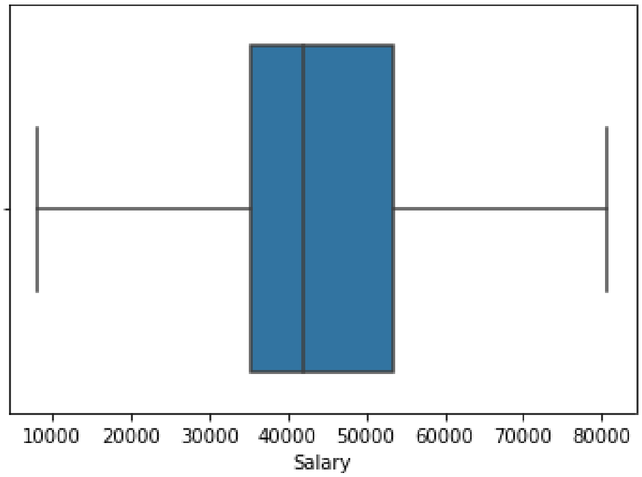


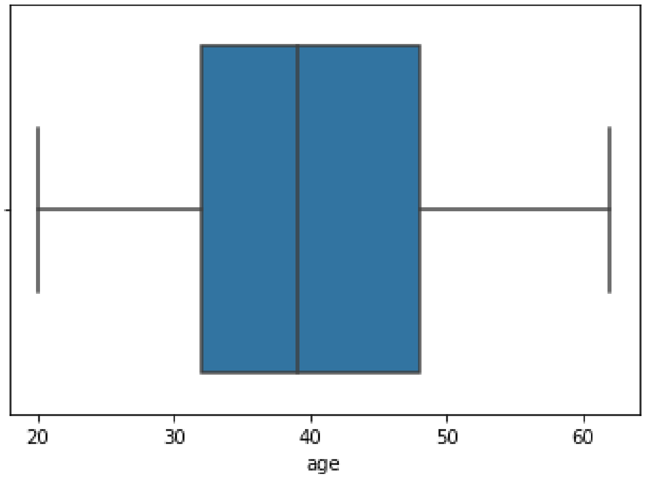


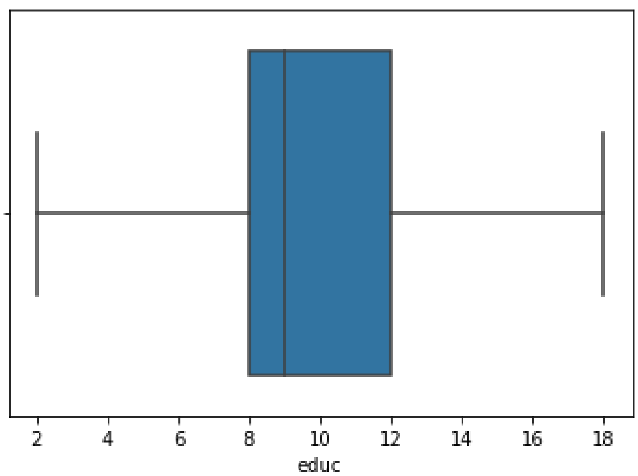


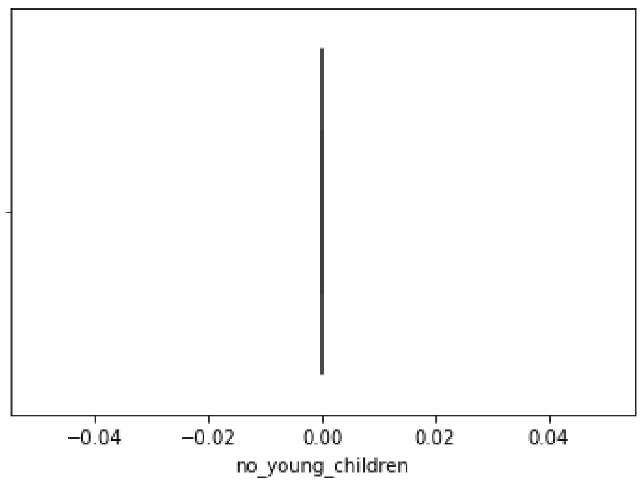
We can observer that there are significant outliers present in variable “ Salary”, however there are minimal outliers in other variables like ‘educ’, ‘no. of young children’ & ‘no. of older children’. There are no outliers in variable ‘age’. For Interpretation purpose we would need to study the variables.

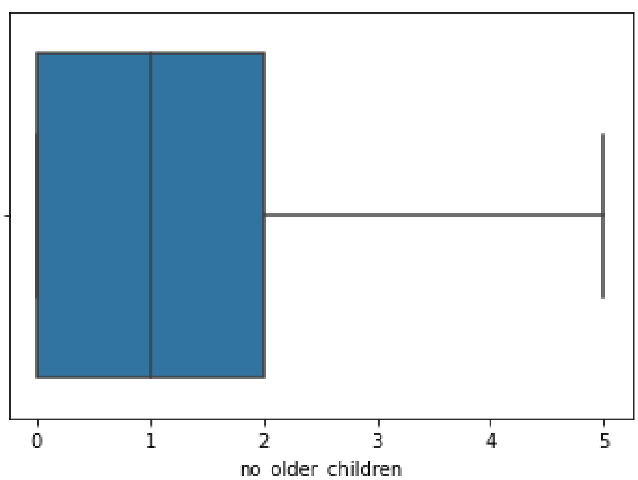
Treating Outliers :



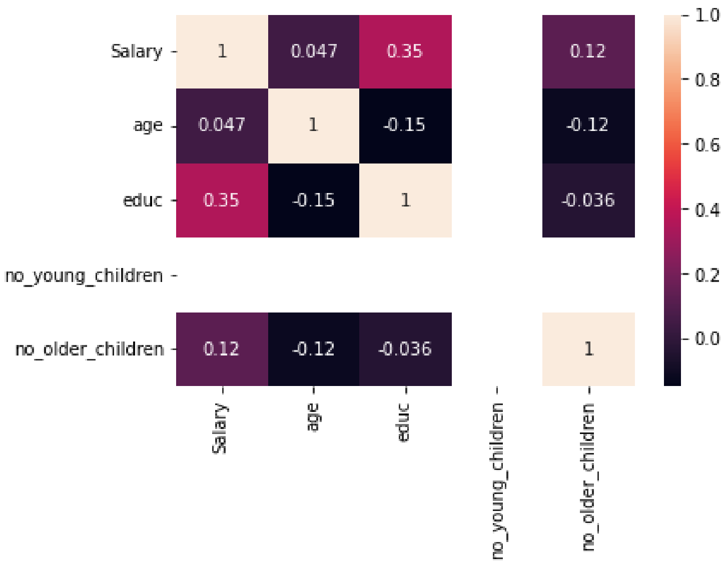






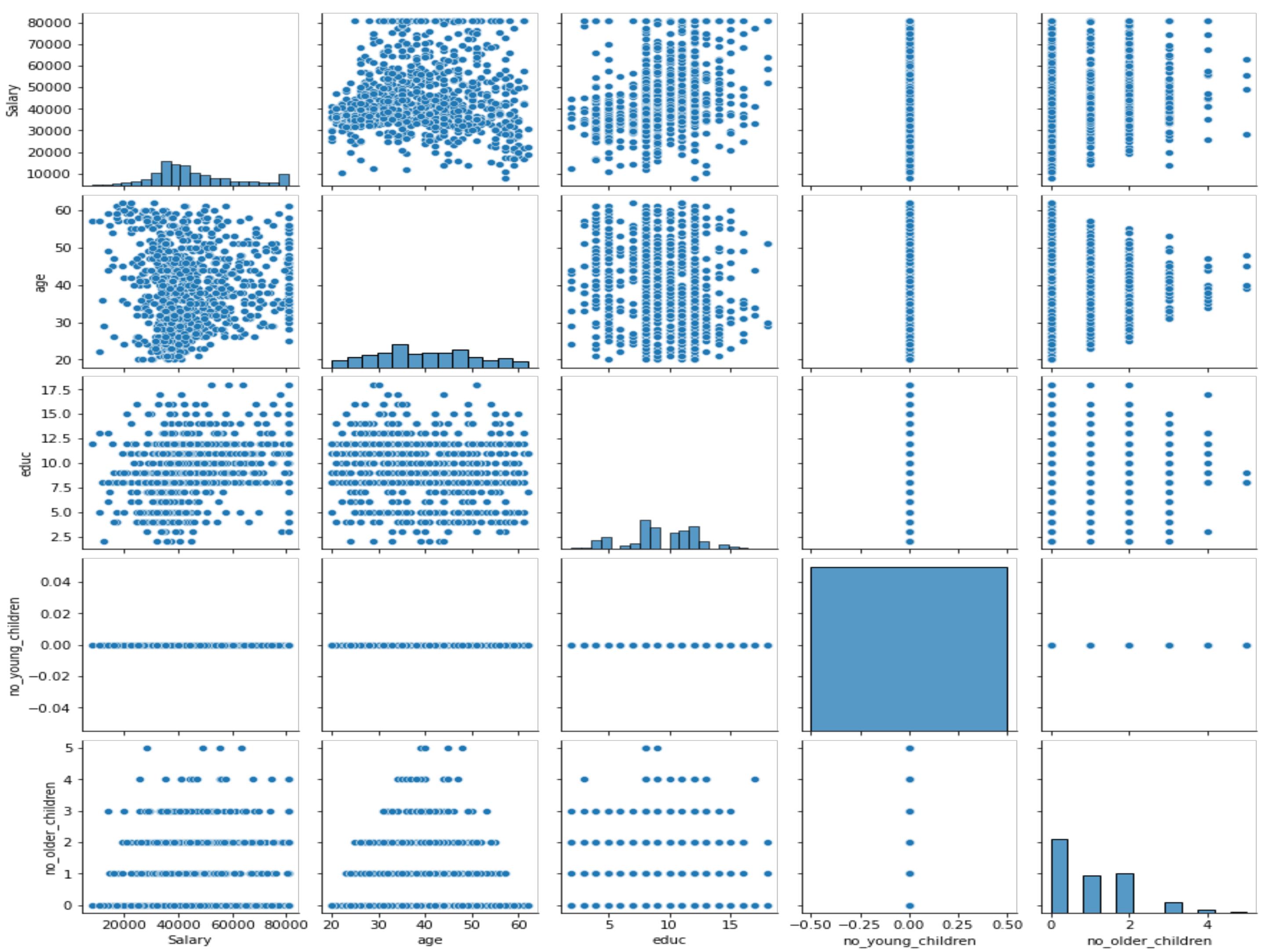


Heatmap :



We can relate there isn't any strong correlation between any variables. Age education display a moderate relationship.

Pairplot :



1. Checked for data Correlation.
2. We will see correlation between independent variables to see which factors might influence choice of holiday package.

2.2Do 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).

Both the variables can be encoded into numerical values for model creation analytical purposes.

Holiday package = no = 0

Holiday package =yes=1

Foreign=no=0

Foreign=yes=1

Table :

Splitted data :

	Salary	age	educ	no_young_children	no_older_children	foreign
0	48412.0	30.0	8.0	0.0	1.0	0.0
1	37207.0	45.0	8.0	0.0	1.0	0.0
2	58022.0	46.0	9.0	0.0	0.0	0.0
3	66503.0	31.0	11.0	0.0	0.0	0.0
4	66734.0	44.0	12.0	0.0	2.0	0.0

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.

Train-Test Score , Confusion matrix , Classification report :

Train- 0.5344262295081967
Test - 0.5534351145038168

Train :

```
([[326,   0],
  [284,   0]])
```

Test :

```
([[145,   0],
  [117,   0]])
```

Train :

	precision	recall	f1-score	support
0.0	0.53	1.00	0.70	326
1.0	0.00	0.00	0.00	284
accuracy			0.53	610
macro avg	0.27	0.50	0.35	610
weighted avg	0.29	0.53	0.37	610

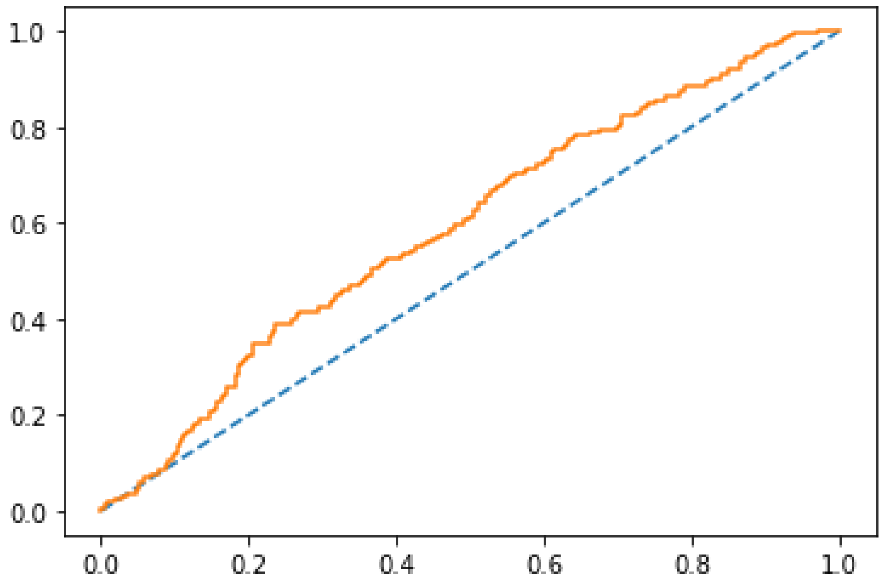
Test :

	precision	recall	f1-score	support
0.0	0.55	1.00	0.71	145
1.0	0.00	0.00	0.00	117
accuracy			0.55	262

macro avg	0.28	0.50	0.36	262
weighted avg	0.31	0.55	0.39	262

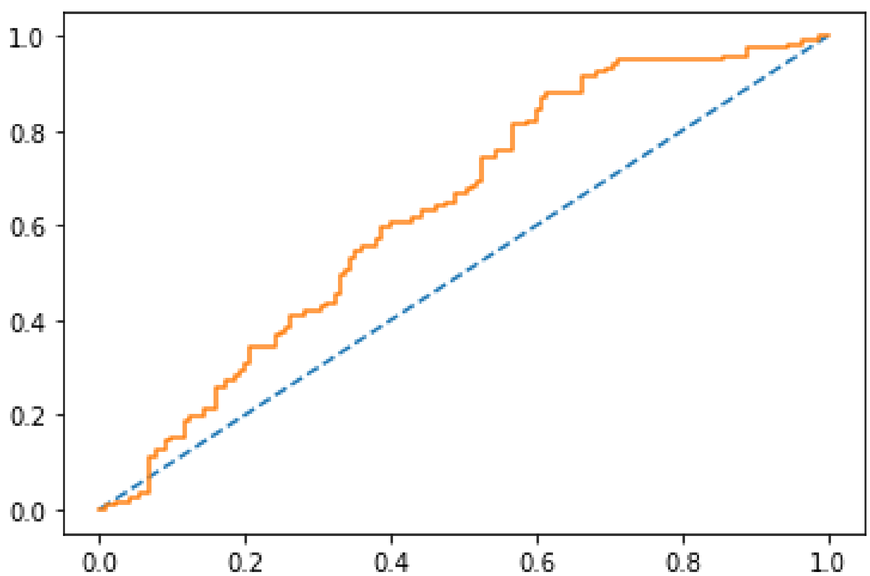
AUC, ROC-Curve (Training) :

AUC- 0.591



AUC, ROC-Curve (Testing) :

AUC- 0.591



Linear discriminant analysis :

Train-Test Score , Confusion matrix , Classification report :

Train-0.6426229508196721
Test -0.6297709923664122

Train :

```
([[269,  57],
  [161, 123]])
```

Test –

```
([[113,  32],
  [ 65,  52]])
```

Train :

	precision	recall	f1-score	support
0.0	0.63	0.83	0.71	326
1.0	0.68	0.43	0.53	284
accuracy			0.64	610
macro avg	0.65	0.63	0.62	610
weighted avg	0.65	0.64	0.63	610

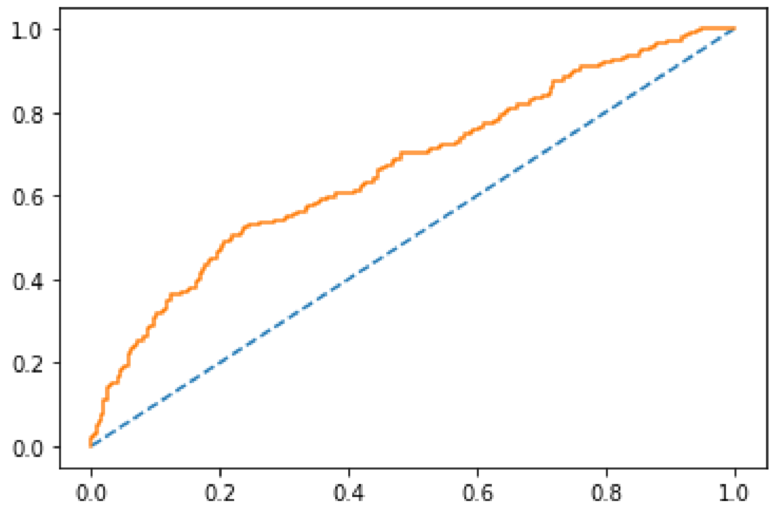
Test –

	precision	recall	f1-score	support
0.0	0.63	0.78	0.70	145
1.0	0.62	0.44	0.52	117

accuracy			0.63	262
macro avg	0.63	0.61	0.61	262
weighted avg	0.63	0.63	0.62	262

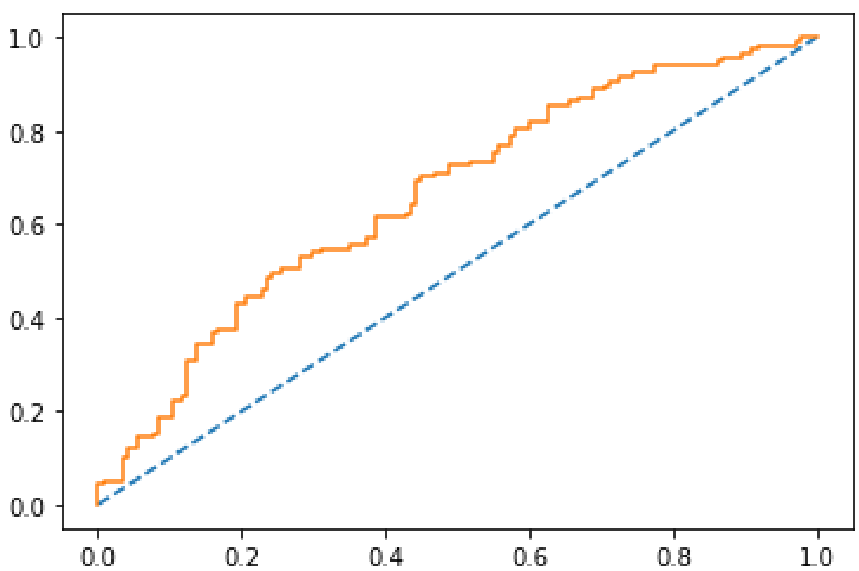
AUC, ROC-Curve (Training) :

AUC - 0.667



AUC, ROC-Curve (Testing) :

AUC - 0.667



I will be choosing LDA model because as we can see its giving better accuracy and overall values of the different variables so I would be going forward with LDA and also it's a very flexible model to use so on the basis of auc roc curve and value we can see its clear that LDA is the better one then the Logistic regression model.

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

As interpretation,

1) There is no plausible effect of salary, age, and education on the prediction for Holliday_packages. These variables don't seem to impact the decision to opt for holiday packages as we couldn't establish a strong relation of these variables with the target variable

2) Foreign has emerged as a strong predictor with a positive coefficient value. The log likelihood or likelihood of a foreigner opting for a holiday package is high.

3) no_young_children variable is negating the probability for opting for holiday packages, especially for couple with number of young children at 2.

The company can try to bin salary ranges to see if they can derive some more meaningful interpretations out of that variable. May be club the salary or age in different buckets and see if there is some plausible impact on the predictor variable. OR else, the business can use some different model techniques to do a deep dive.

Recommendation:

1) The company should really focus on foreigners to drive the sales of their holiday packages as that's where the majority of conversions are going to come in.

2) The company can try to direct their marketing efforts or offers toward foreigners for a better conversion opting for holiday packages

3) The company should also stay away from targeting parents with younger children. The chances of selling to parents with 2 younger children is probably the lowest. This also gels with the fact that parents try and avoid visiting with younger children.

4) If the firm wants to target parents with older children, that still might end up giving favorable return for their marketing efforts then spent on couples with younger children.