**Table of Contents**

**Project – Predictive Modelling Problem Statement:**

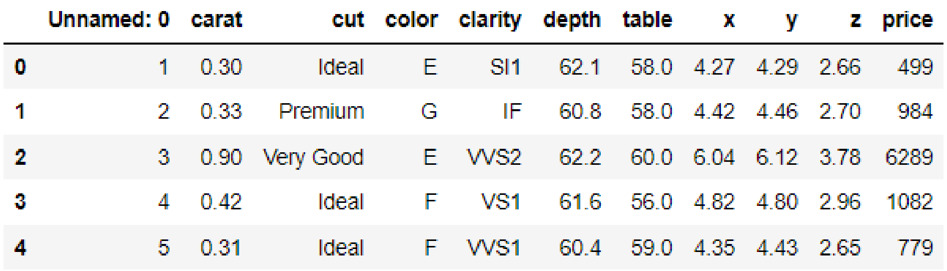
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is 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. The data set is **cubic_zirconia.csv**

**Data Dictionary:**

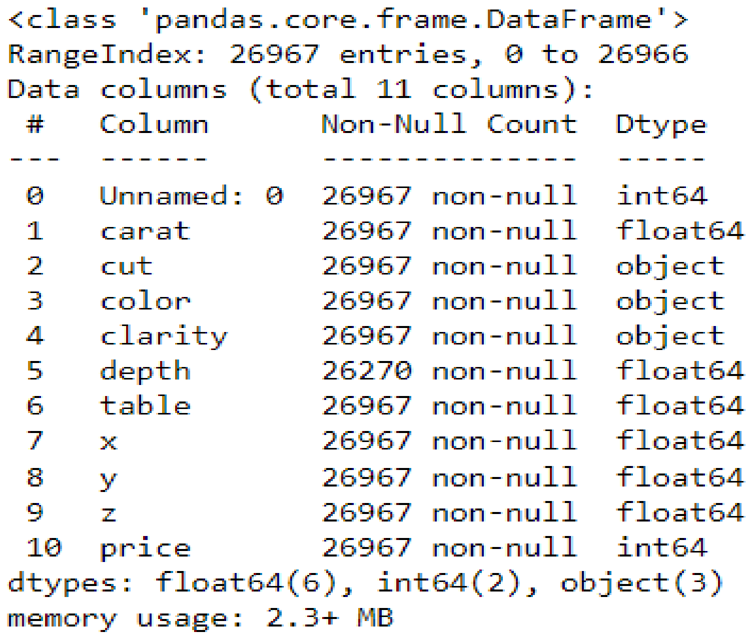
| **Variable Name** | **Description** |
|---|---|
| Carat | Carat weight of the cubic zirconia. |
| Cut | Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal. |
| Color | Colour of the cubic zirconia.With D being the worst and J the best. |
| Clarity | Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1 |
| Depth | The Height of 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. |

### Q1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.

**Exploratory Data Analysis or (EDA)** is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, Box plot, pair plot and many more. It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important. Imported the required libraries.



| | 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 |

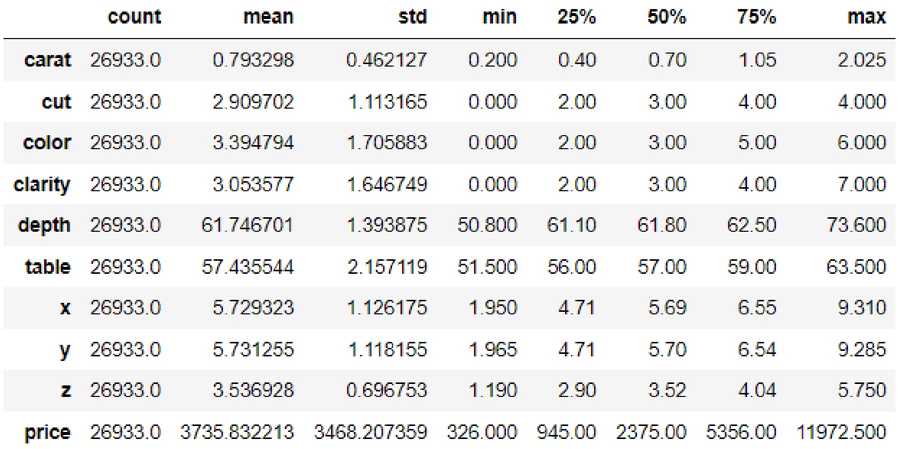


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   Unnamed: 0  26967 non-null  int64  
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  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  
dtypes: float64(6), int64(2), object(3)
memory usage: 2.3+ MB
```

1.The data set contains 26967 row, 11 columns. In the given data set, there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is target variable and all other are predictor variable.

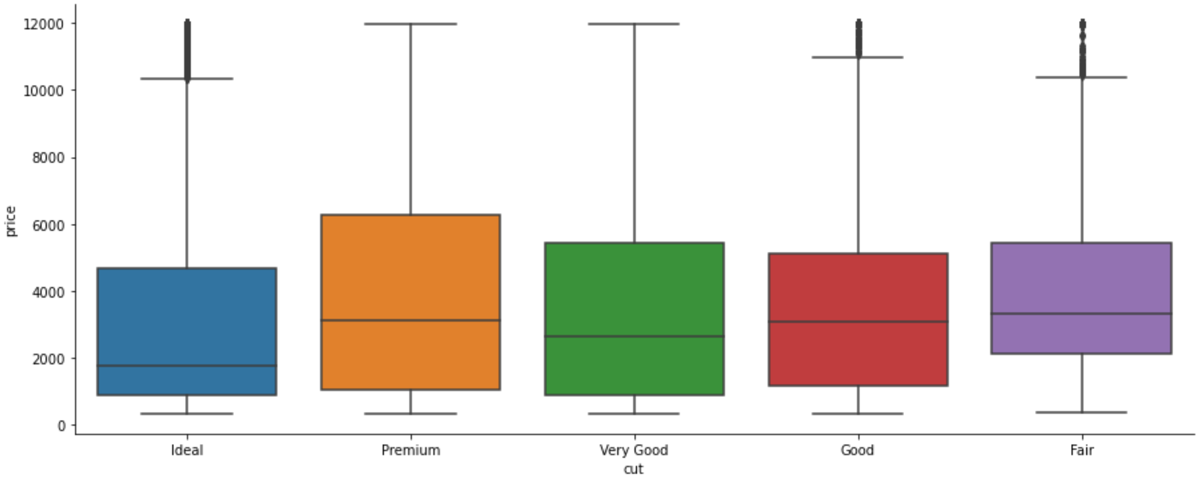
2.The first column is an index ("Unnamed: 0") as this only serial no, we can remove it.

3. Except depth, in all the column non null count is 26967.



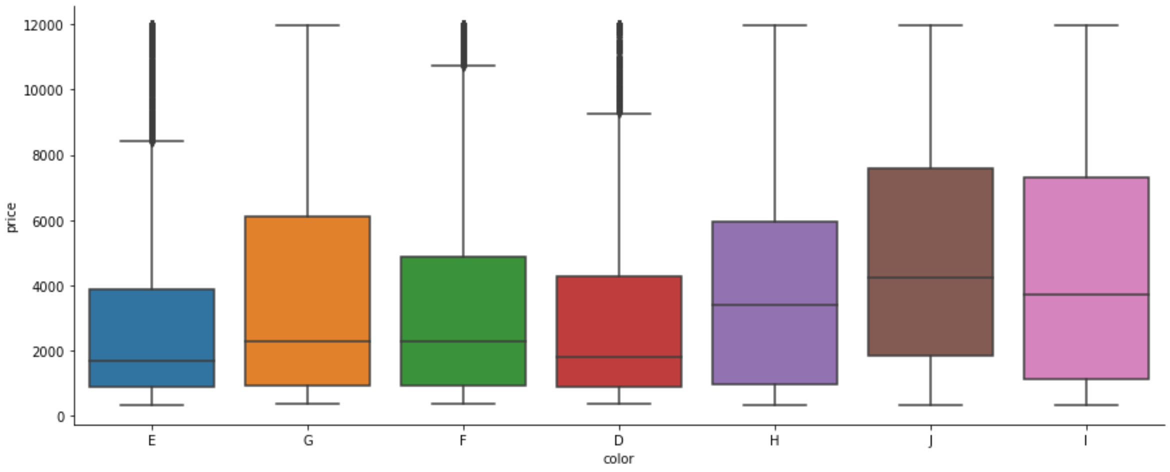
| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **carat** | 26933.0 | 0.793298 | 0.462127 | 0.200 | 0.40 | 0.70 | 1.05 | 2.025 |
| **cut** | 26933.0 | 2.909702 | 1.113165 | 0.000 | 2.00 | 3.00 | 4.00 | 4.000 |
| **color** | 26933.0 | 3.394794 | 1.705883 | 0.000 | 2.00 | 3.00 | 5.00 | 6.000 |
| **clarity** | 26933.0 | 3.053577 | 1.646749 | 0.000 | 2.00 | 3.00 | 4.00 | 7.000 |
| **depth** | 26933.0 | 61.746701 | 1.393875 | 50.800 | 61.10 | 61.80 | 62.50 | 73.600 |
| **table** | 26933.0 | 57.435544 | 2.157119 | 51.500 | 56.00 | 57.00 | 59.00 | 63.500 |
| **x** | 26933.0 | 5.729323 | 1.126175 | 1.950 | 4.71 | 5.69 | 6.55 | 9.310 |
| **y** | 26933.0 | 5.731255 | 1.118155 | 1.965 | 4.71 | 5.70 | 6.54 | 9.285 |
| **z** | 26933.0 | 3.536928 | 0.696753 | 1.190 | 2.90 | 3.52 | 4.04 | 5.750 |
| **price** | 26933.0 | 3735.832213 | 3468.207359 | 326.000 | 945.00 | 2375.00 | 5356.00 | 11972.500 |

## Price Distribution of Cut Variable: -



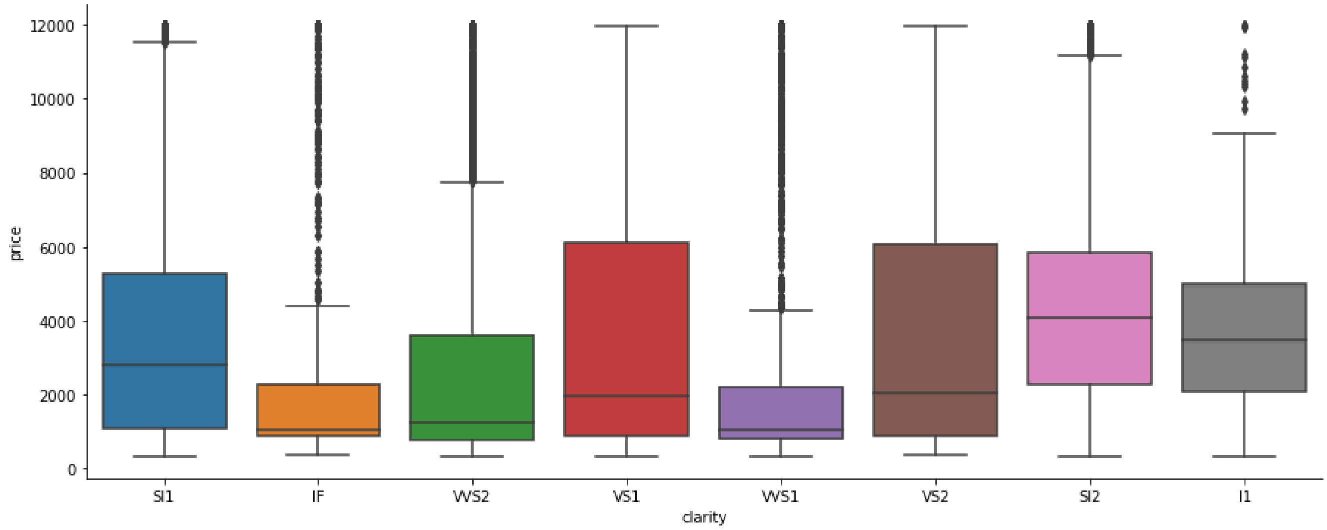
observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

## Price Distribution of Color Variable: -

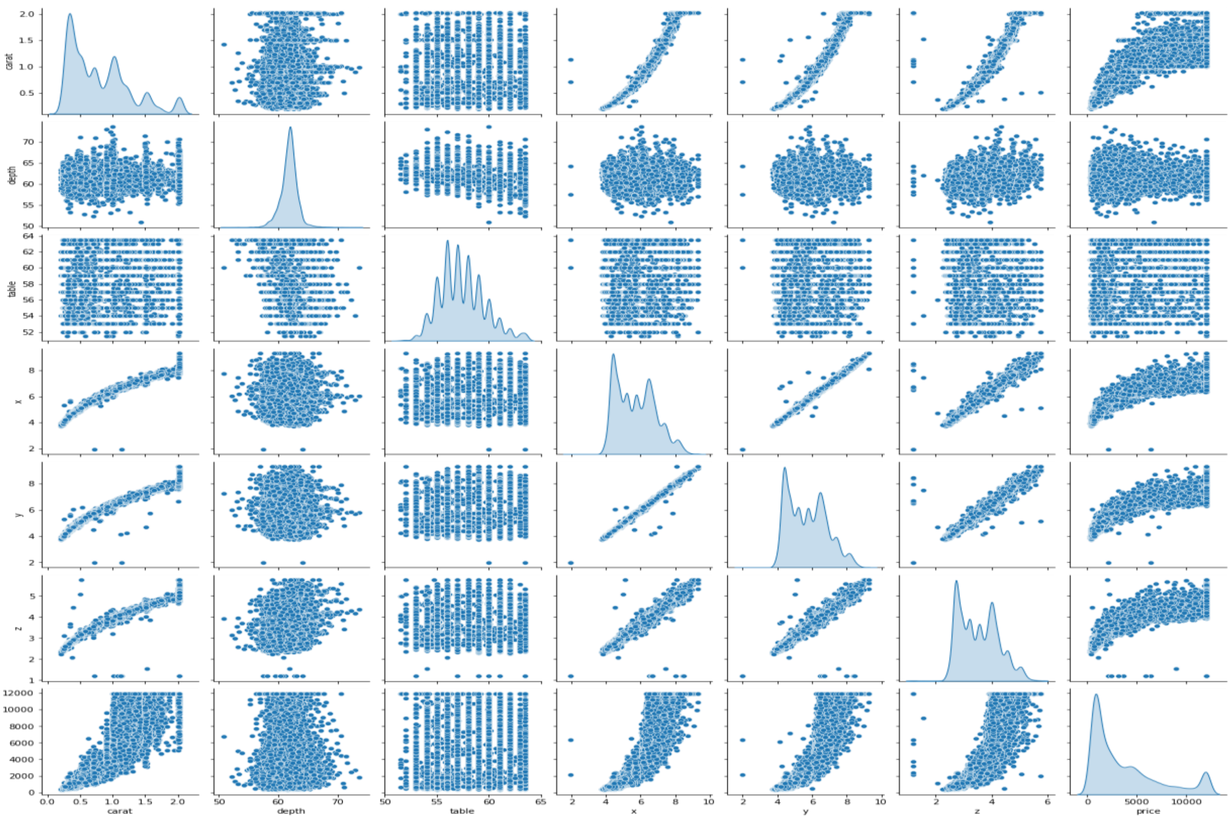


All color type gems E, F and D have outliers with respect to price.

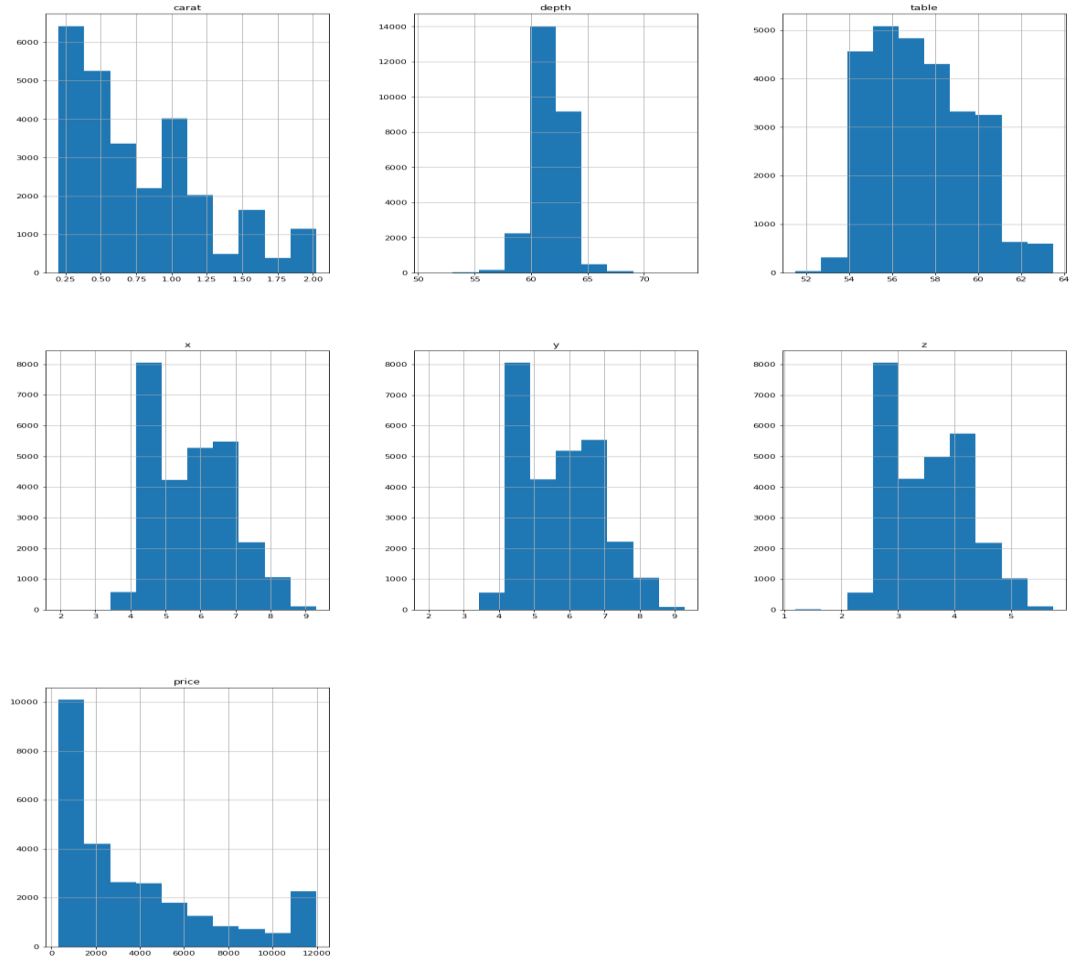
**Price Distribution of Clarity Variable: -**



**Pair plot:-**


carat
2.0
1.5
1.0
0.5
depth
70
65
60
55
50
table
64
62
60
58
56
54
52
x
8
6
4
2
y
8
6
4
2
z
5
4
3
2
1
price
12000
10000
8000
6000
4000
2000
0
0.0 0.5 1.0 1.5 2.0
carat
50 60 70
depth
50 55 60 65
table
2 4 6 8 10
x
2 4 6 8 10
y
2 4 6
z
0 5000 10000
price

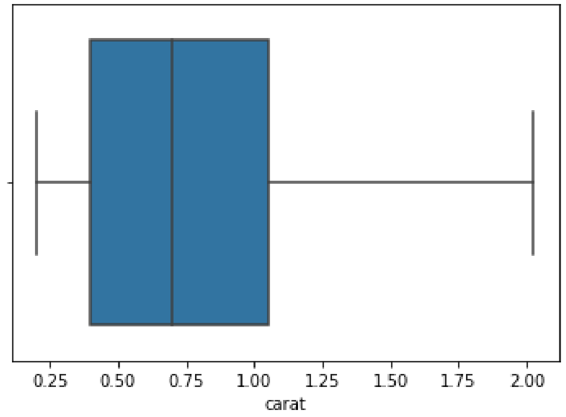
## Histogram:-

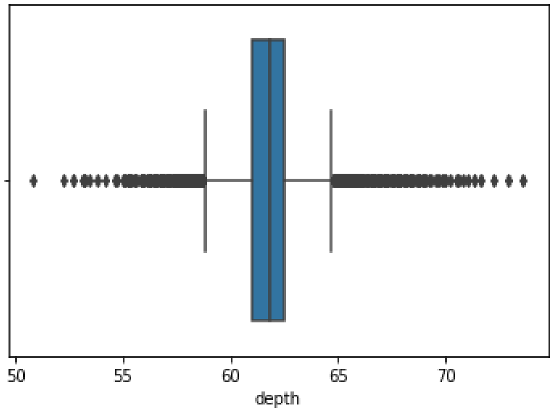


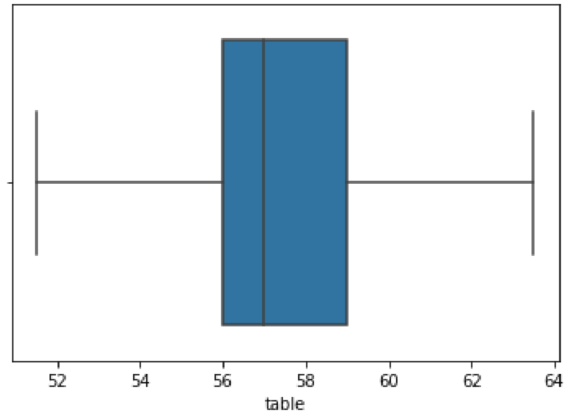
(1). There is significant amount of outlier present in some variable.

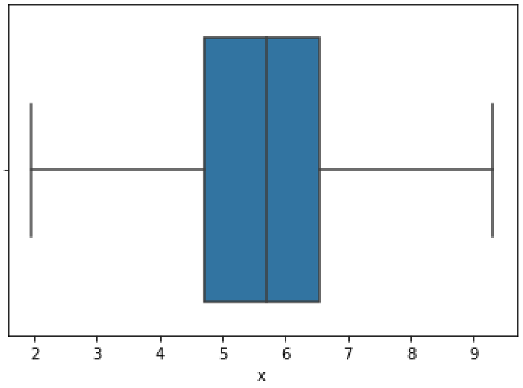
(2). We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

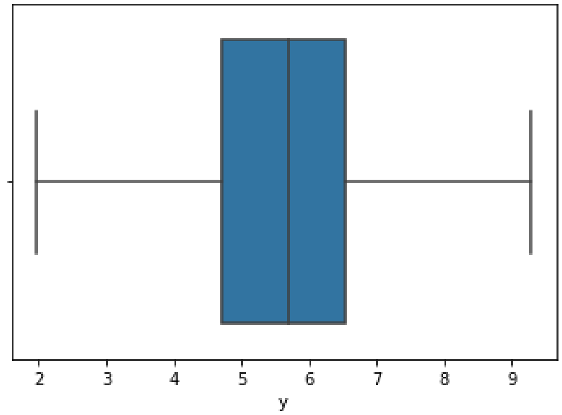
**Boxplot after outlier Treatment: -**

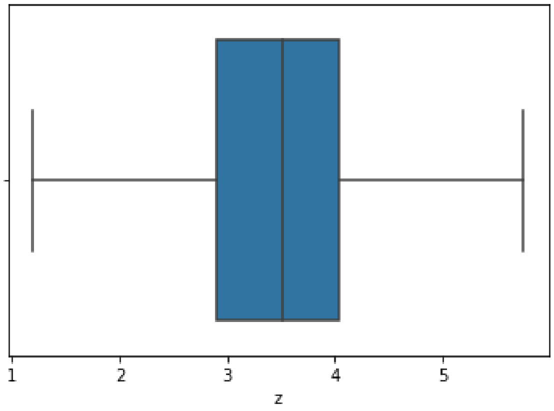

0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
carat

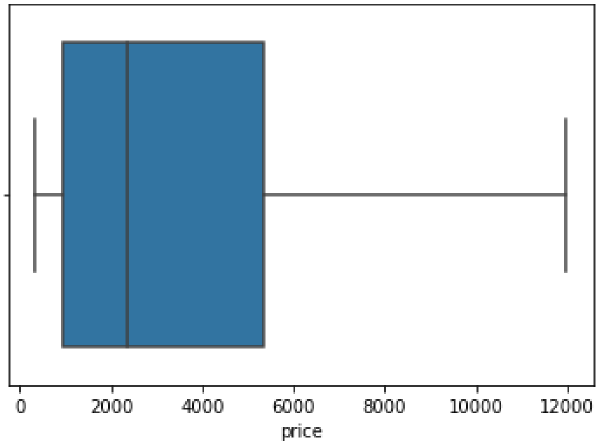

50
55
60
65
70
depth


52
54
56
58
60
62
64
table

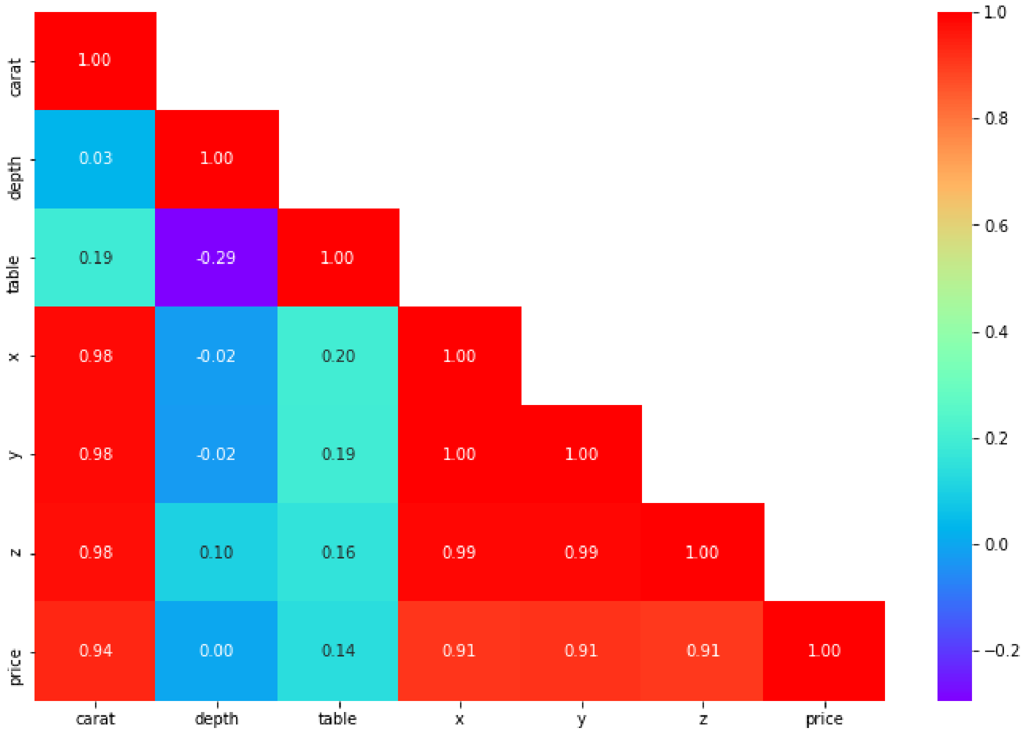

2
3
4
5
6
7
8
9
x


2
3
4
5
6
7
8
9
y


1
2
3
4
5
z



## Checking Correlation in the data using Heatmap: -



It can be inferred that most features correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (<1%).

**The inferences drawn from the above Exploratory Data analysis:**

**Observation-1:**

(1).'Price' is the target variable while all others are the predictors.

(2). The data set contains 26967 row, 11 column.

(3). 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 and all other are predictor variable.

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

**Observation-2:**

(1). On the given data set the the mean and median values does not have much difference.

(2). We can observe Min value of "x", "y", "z" is zero 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.

(3). There are three object data type 'cut', 'color' and 'clarity'.

**Observation-3:**

we can observe there are 697 missing values in the depth column. There are some duplicate row presents. (33 duplicate rows out of 26958). which is nearly 0.12 % of the total data. So, on this case we have dropped the duplicated row.

**Observation-4:**

There is significant amount of outlier present in some variable, the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So, we have treated the outlier. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

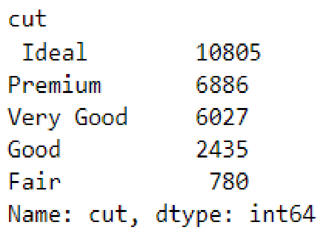
**Observation-5:**

It looks like most features do correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (~1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

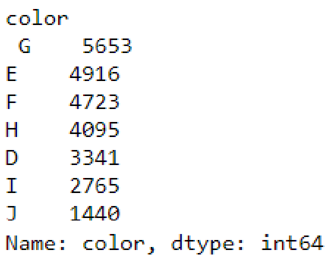
**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub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

It is important to check if the data has any missing value or gibberish data in it. We did check about the same for both object and numerical data types and can confirm the following:

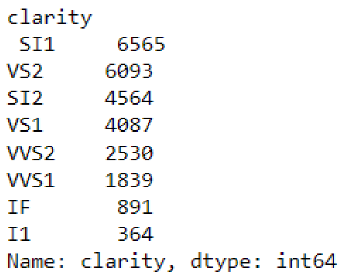
There is no gibberish or missing data in the object type data columns – Cut, color and clarity



```
cut
 Ideal        10805
Premium       6886
Very Good     6027
Good          2435
Fair           780
Name: cut, dtype: int64
```

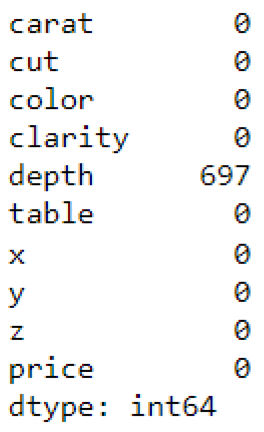


```
color
 G    5653
E    4916
F    4723
H    4095
D    3341
I    2765
J    1440
Name: color, dtype: int64
```

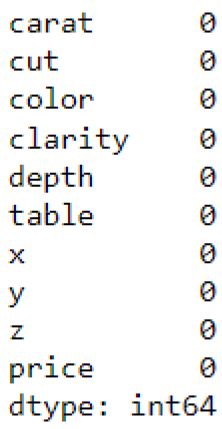


```
clarity
 SI1     6565
VS2     6093
SI2     4564
VS1     4087
VVS2    2530
VVS1    1839
IF       891
I1       364
Name: clarity, dtype: int64
```

There are missing values in the column "depth" – 697 cells or 2.6% of the total data set. We can choose to impute these values using a mean or median. We checked for both the values and the result for both is almost similar.



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



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

But is scaling necessary in this case? No, it is not necessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not. But recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in preprocessing.

For now, we will process the model without scaling and later we will check the output with scaled data of regression model output.

## 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 stats model. 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.

Observation-1:

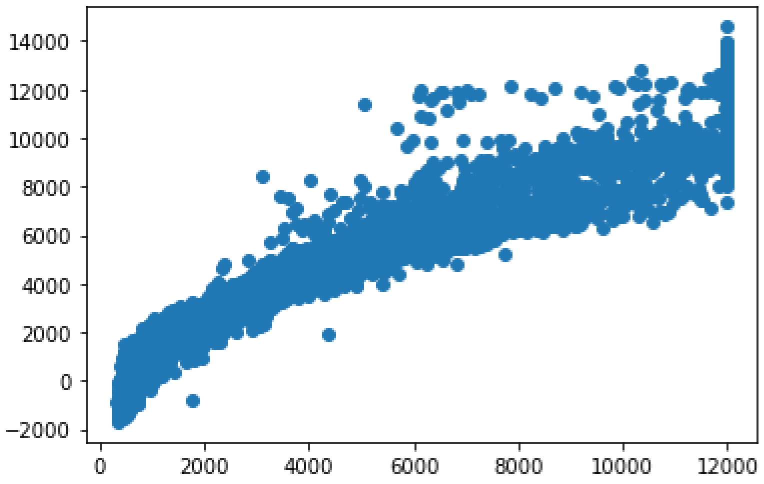
The intercept (often labelled the constant) is the expected mean value of Y when all X=0. If X never equals 0, then the intercept has no intrinsic meaning.

The intercept for our model is -3171.950447307667. In preset case when the other predictor variable is zero i.e like carat, cut, color, clarity all are zero then the C=-3172. (Y = m1X1 + m2X2+ …. + mnXn + C + e) that means price is -3172. which is meaningless. We can do Z score or scaling the data and make it nearly zero.

Observation-2: R-square is the percentage of the response variable variation that is explained by a linear model. Or:

R-square = Explained variation / Total variation

R-squared is always between 0 and 100%: 0% indicates that the model explains none of the variability of the response data around its mean.100% indicates that the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively **0.9312009582430333 and 0.9312386078792532**



Observation-2:-

we can see that the is a linear plot, very strong correlation between the predicted y and actual y. But there are lots of spread. That indicated some kind noise present on the data set i.e Unexplained variances on the output.

**Linear regression Performance Metrics**:

intercept for the model: -1275.7248878055789

R square on training data: 0.9312009582430333 (93.12%)

R square on testing data: 0.9312386078792532 (93.12%)

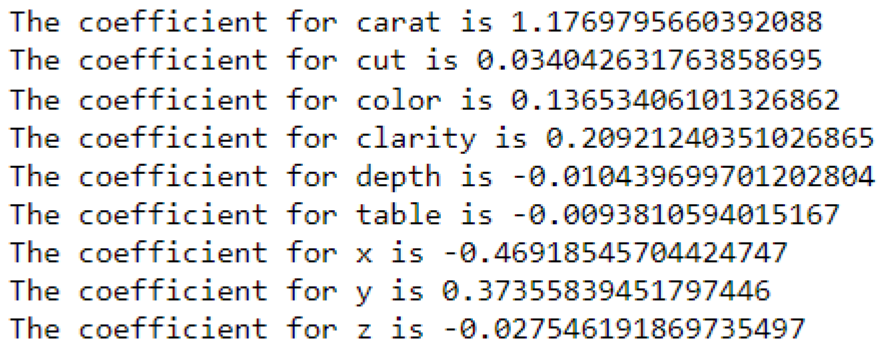
RMSE on Training data: 909.021075592446

RMSE on Testing data: 910.9629751292744

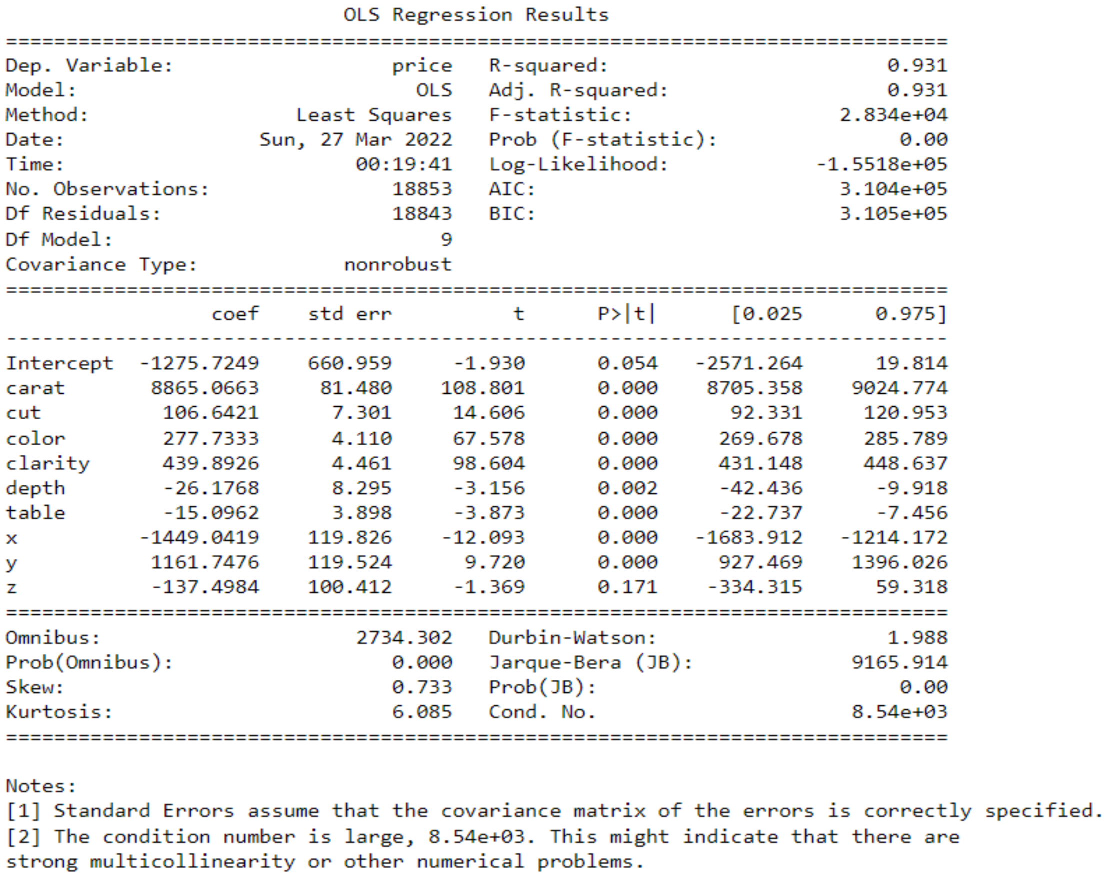
As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model

### Applying zscore stats models: -

The coefficients are mentioned below:



```
The coefficient for carat is 1.1769795660392088
The coefficient for cut is 0.034042631763858695
The coefficient for color is 0.13653406101326862
The coefficient for clarity is 0.20921240351026865
The coefficient for depth is -0.010439699701202804
The coefficient for table is -0.0093810594015167
The coefficient for x is -0.46918545704424747
The coefficient for y is 0.37355839451797446
The coefficient for z is -0.027546191869735497
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 2.834e+04
Date:                Sun, 27 Mar 2022   Prob (F-statistic):               0.00
Time:                        00:19:41   Log-Likelihood:            -1.5518e+05
No. Observations:               18853   AIC:                         3.104e+05
Df Residuals:                   18843   BIC:                         3.105e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept -1275.7249    660.959     -1.930      0.054   -2571.264      19.814
carat      8865.0663     81.480    108.801      0.000    8705.358    9024.774
cut         106.6421      7.301     14.606      0.000      92.331     120.953
color       277.7333      4.110     67.578      0.000     269.678     285.789
clarity     439.8926      4.461     98.604      0.000     431.148     448.637
depth       -26.1768      8.295     -3.156      0.002     -42.436      -9.918
table       -15.0962      3.898     -3.873      0.000     -22.737      -7.456
x         -1449.0419    119.826    -12.093      0.000   -1683.912   -1214.172
y          1161.7476    119.524      9.720      0.000     927.469    1396.026
z          -137.4984    100.412     -1.369      0.171    -334.315      59.318
==============================================================================
Omnibus:                     2734.302   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9165.914
Skew:                           0.733   Prob(JB):                         0.00
Kurtosis:                       6.085   Cond. No.                     8.54e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.54e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

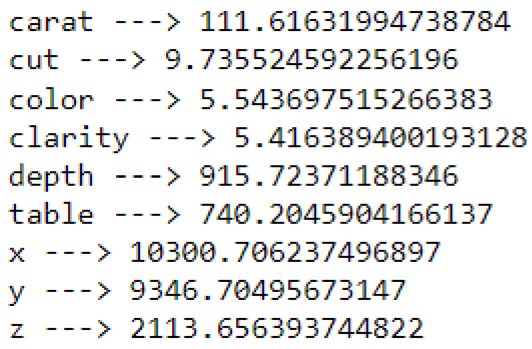
**Observation-3:** Assuming null hypothesis is true, i.e. there is no relationship between this variable with price. from that universe we have drawn the sample and, on this sample, we have found this co-efficient for the variable shown above.

Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression co-efficient is not '0'. Here all regression co-efficient are not '0'.

For an example: we can see the p value is showing 0.449 for 'depth' variable, which is much higher than 0.05. That means this dimension is useless. So, we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.

### Check Multi-collinearity using VIF: -

We can then look at the Variance Inflation Factor (VIF) score to check the multicollinearity scores. As per the industry standards or at least for this case study, any variable with a VIF score of greater than 10 has been accepted to indicate severe collinearity.



```
carat ---> 111.61631994738784
cut ---> 9.735524592256196
color ---> 5.543697515266383
clarity ---> 5.416389400193128
depth ---> 915.72371188346
table ---> 740.2045904166137
x ---> 10300.706237496897
y ---> 9346.70495673147
z ---> 2113.656393744822
```

**Observation-4:** On the given data set we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stone.

Similarly, for the 'z' variable having negative co-efficient i.e -137.49 And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stone.

Also, we can see the 'y' width in mm having positive co-efficient. And the p value is less than 0.05, so we can conclude that higher the width of the stone is a higher profitable stone.

Finally, we can conclude that best 5 attributes that are most important are 'Carat', 'Cut', 'color', 'clarity' and width i.e 'y' for predicting the price.

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

Inference: we can see that the from the linear plot, very strong correlation between the predicted y and actual y. But there are lots of spread. That indicates some kind noise present on the data set i.e Unexplained variances on the output.

Linear regression Performance Metrics:

intercept for the model: -1275.7248878055789 R square on training data: 0.9312009582430333 R square on testing data: 0.9312386078792532 RMSE on Training data: 909.021075592446 RMSE on Testing data: 910.9629751292744 As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Impact of scaling:

Now we can observe by applying z score the intercept became -1.5793957305153216e-16. Earlier it was - 1275.7248878055789. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

Multi collinearity: We can observe there are very strong multi collinearity present in the data set.

From stats models: we can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

Recommendations:

The Gem Stones company should consider the features 'Carat', 'Cut', 'color', 'clarity' and width i.e 'y' as most important for predicting the price. To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

As we can see from the model Higher the width('y') of the stone is higher the price.

So, the stones having higher width('y') should consider in higher profitable stones. The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.

The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So, these two categories also consider in higher profitable stones.

As we see for 'X' i.e Length of the stone, higher the length of the stone is lower the price.

So higher the Length('x') of the stone is lower is the profitability. higher the 'z' i.e Height of the stone is, lower the price. This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why

Stones with higher 'z' is also are lower in profitability.

**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. Dataset for Problem 2: Holiday_Package.csv

**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 |

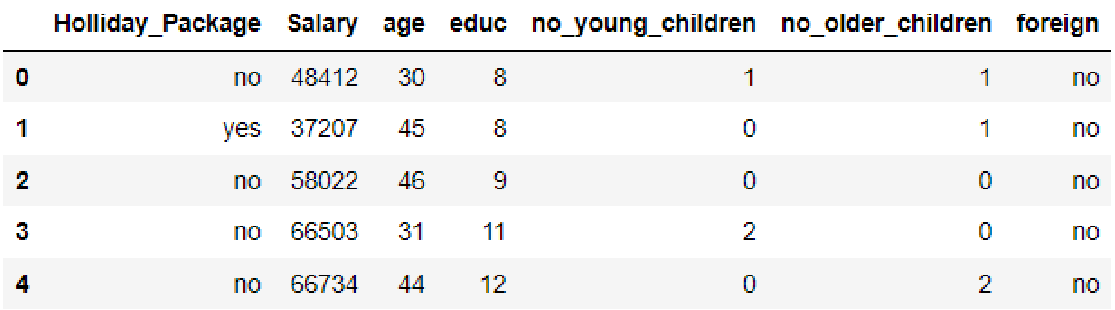
## 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

Exploratory Data Analysis or (EDA) is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, Box plot, pair plot and many more. It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.

Data Summary and Exploratory Data Analysis:

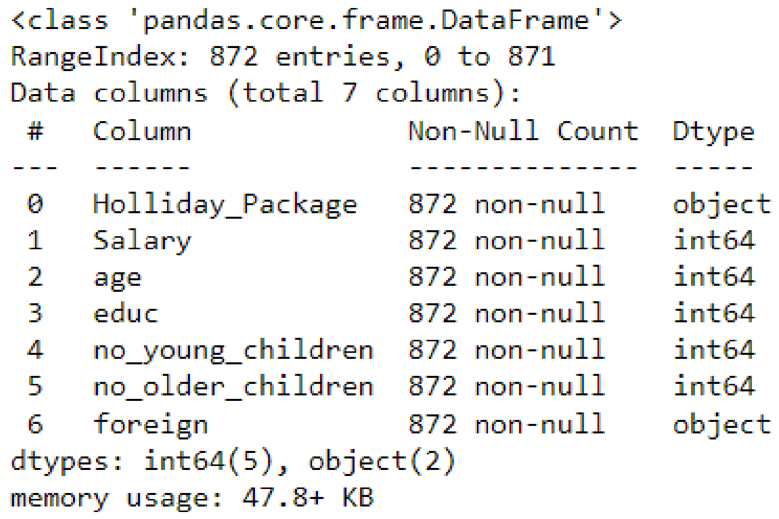
Checking if the data is being imported properly

Head: The top 5 rows of the dataset are viewed using head () function

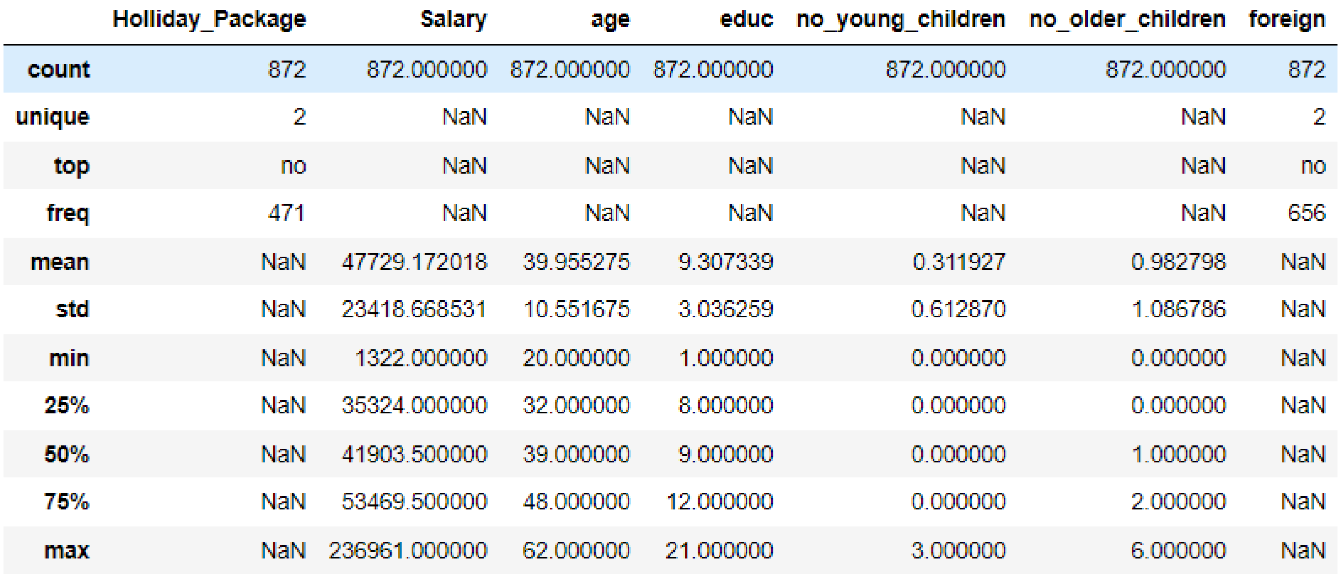


| | Holliday_Package | Salary | age | educ | no_young_children | no_older_children | foreign |
|---|---|---|---|---|---|---|---|
| **0** | no | 48412 | 30 | 8 | 1 | 1 | no |
| **1** | yes | 37207 | 45 | 8 | 0 | 1 | no |
| **2** | no | 58022 | 46 | 9 | 0 | 0 | no |
| **3** | no | 66503 | 31 | 11 | 2 |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

Dimension of the Dataset:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   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
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```



| | Holliday_Package | Salary | age | educ | no_young_children | no_older_children | foreign |
|---|---|---|---|---|---|---|---|
| **count** | 872 | 872.000000 | 872.000000 | 872.000000 | 872.000000 | 872.000000 | 872 |
| **unique** | 2 | NaN | NaN | NaN | NaN | NaN | 2 |
| **top** | no | NaN | NaN | NaN | NaN | NaN | no |
| **freq** | 471 | NaN | NaN | NaN | NaN | NaN | 656 |
| **mean** | NaN | 47729.172018 | 39.955275 | 9.307339 | 0.311927 | 0.982798 | NaN |
| **std** | NaN | 23418.668531 | 10.551675 | 3.036259 | 0.612870 | 1.086786 | NaN |
| **min** | NaN | 1322.000000 | 20.000000 | 1.000000 | 0.000000 | 0.000000 | NaN |
| **25%** | NaN | 35324.000000 | 32.000000 | 8.000000 | 0.000000 | 0.000000 | NaN |
| **50%** | NaN | 41903.500000 | 39.000000 | 9.000000 | 0.000000 | 1.000000 | NaN |
| **75%** | NaN | 53469.500000 | 48.000000 | 12.000000 | 0.000000 | 2.000000 | NaN |
| **max** | NaN | 236961.000000 | 62.000000 | 21.000000 | 3.000000 | 6.000000 | 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.

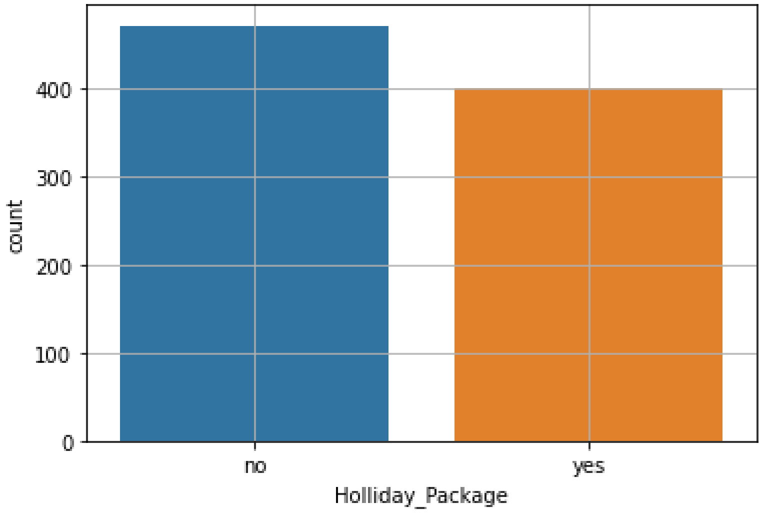
- Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the 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 analysis. The shape would be – 872 rows and 7 columns

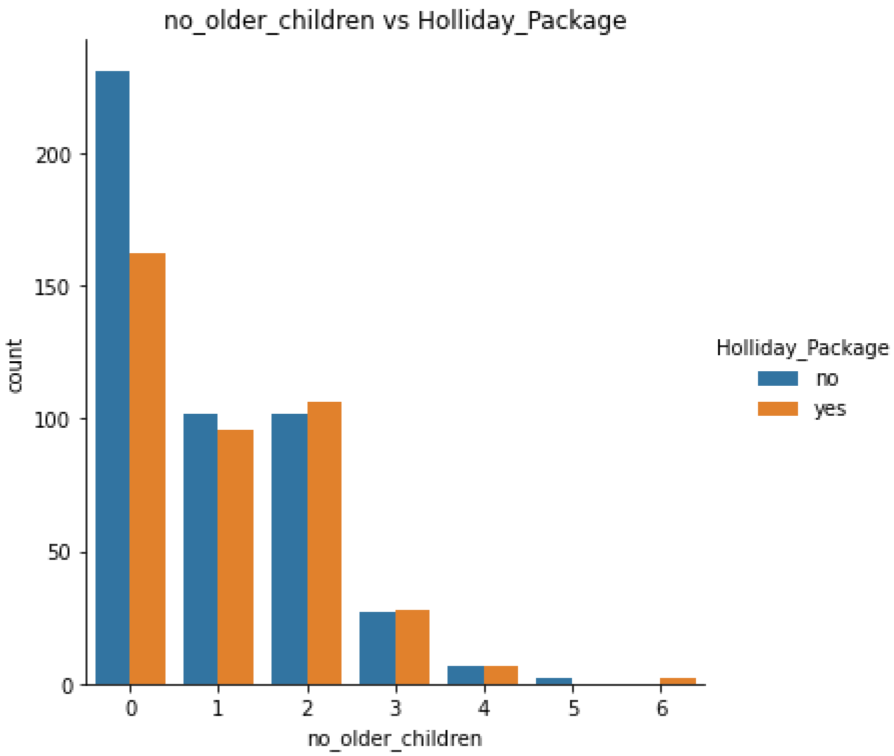
- There are no null values

- There are no duplicates

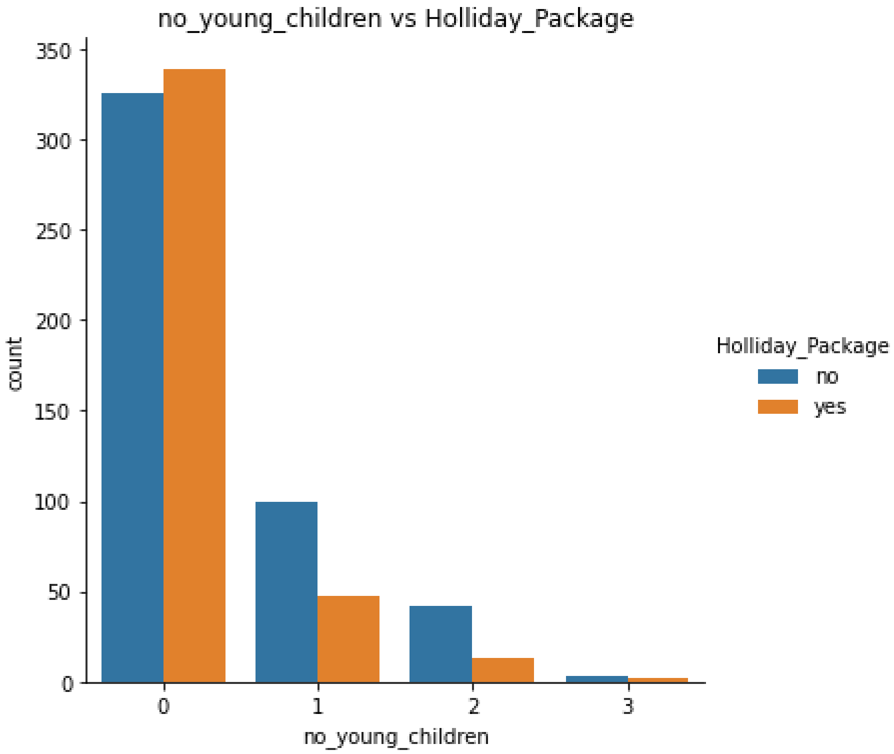


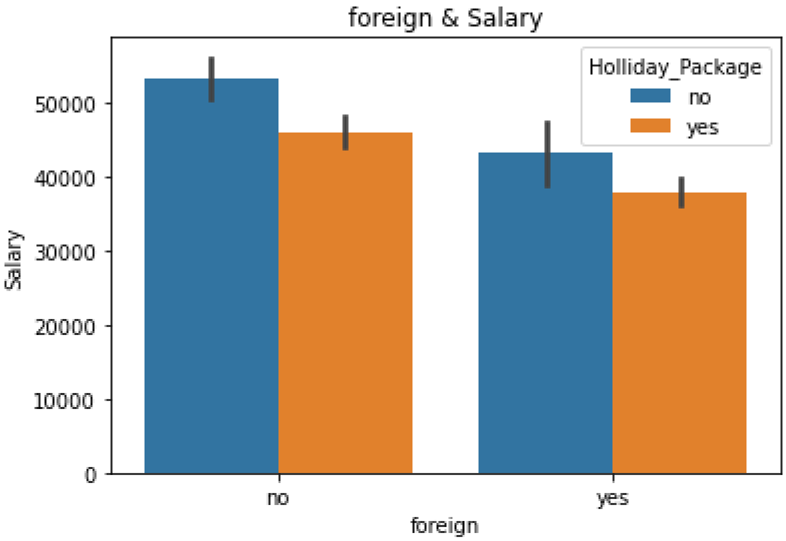
employees who opted for Holiday Package:401(45.9%)

employees who not opted for Holiday Package:471(54%)

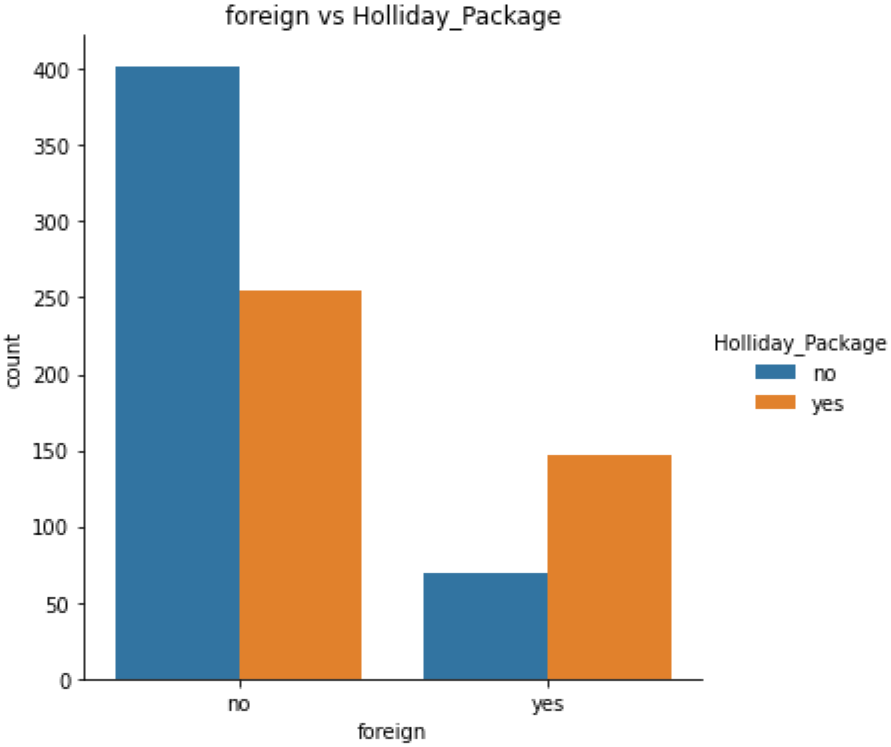


some Employees with no kids opted for Holiday_package but most of employees with no kids not opted the Holiday package.



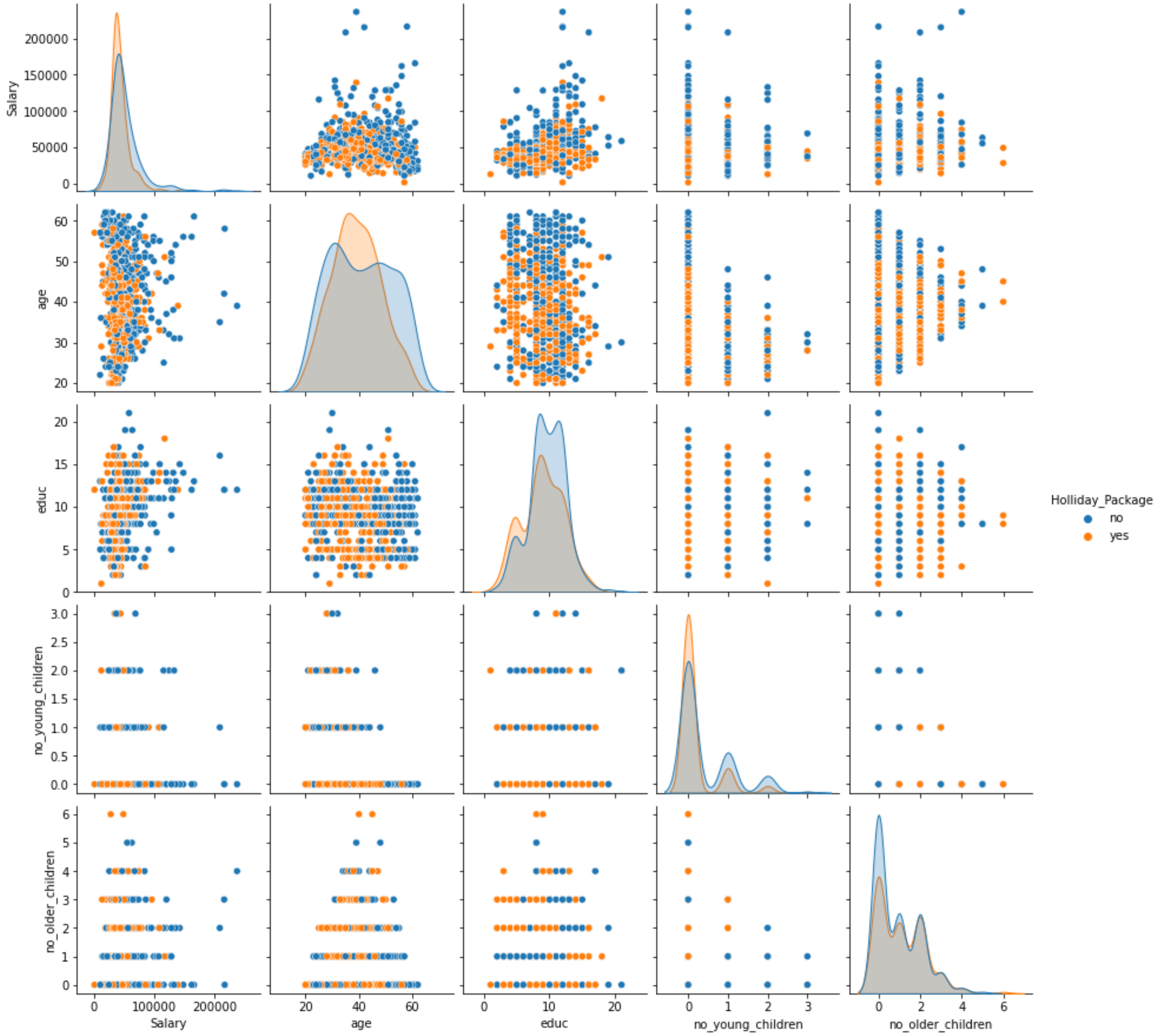


some foreign employers paid less salary than the other employees.



Most of foreign employees are not opted for holiday package.

**Pair plot: -**



Some of the attributes look like they may have an exponential distribution

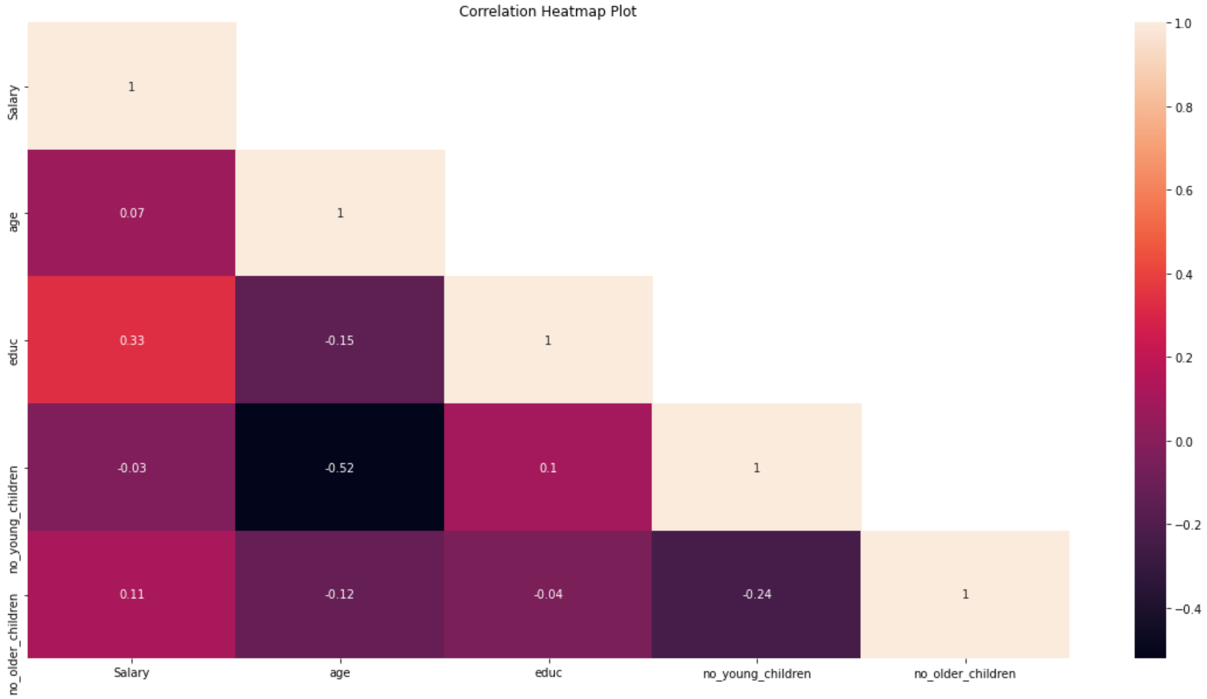
Salary should probably have a normal distribution, the constraints on the data collection may have skewed the distribution.

age and salary are correlated with each each other

educ and salary are correlated with each each other

There is no obvious relationship between no younger children and holiday package

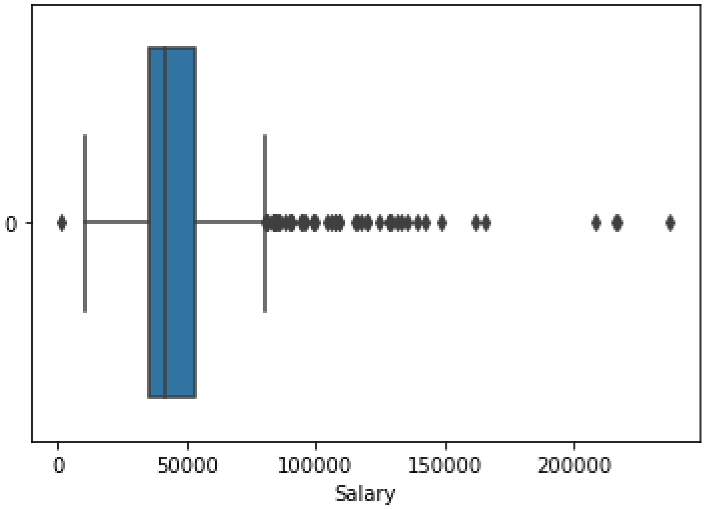
**Heat Map: -**

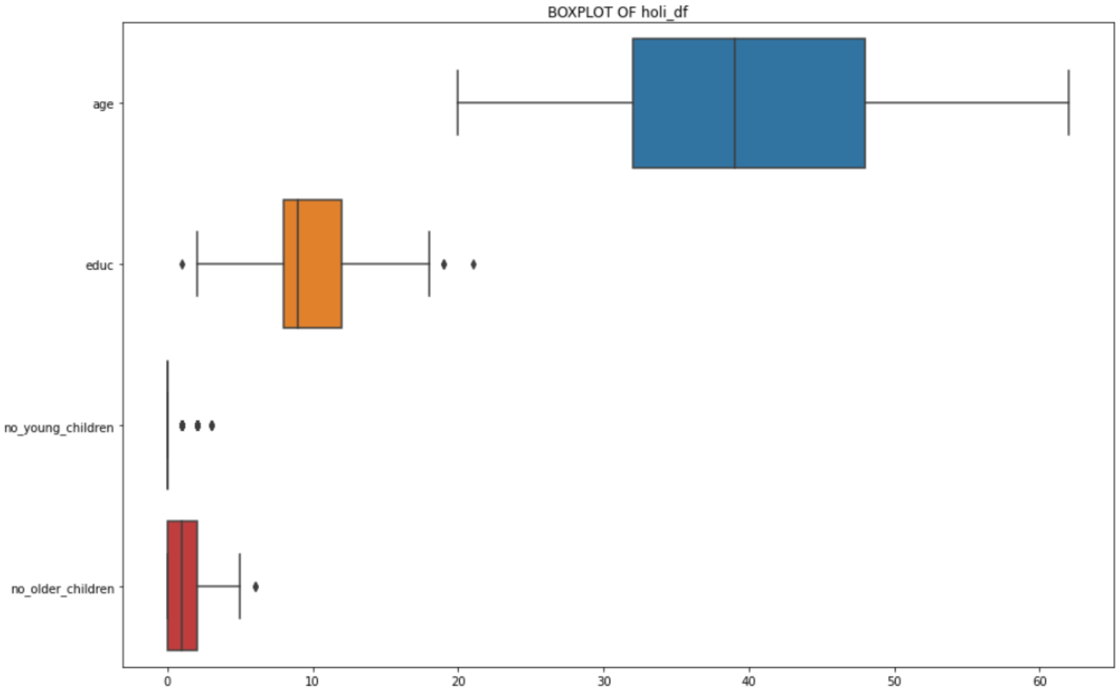


positive correlation between educ and salary

moderate negative correlation seen between age and no_young_children

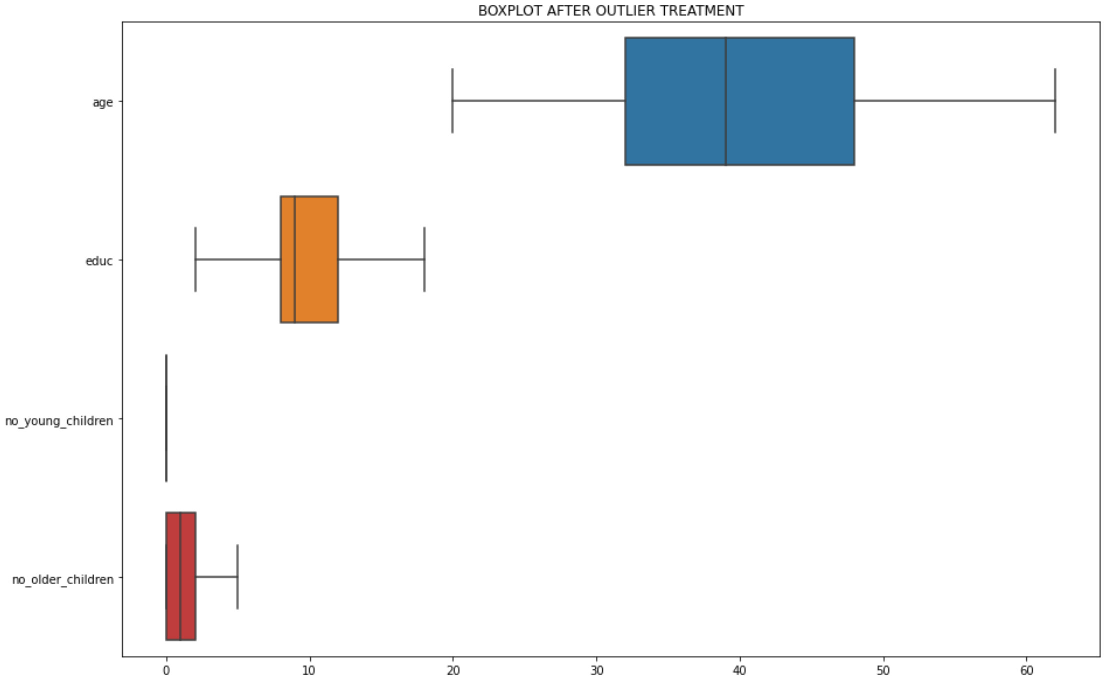
**Box plot for Outliers: -**

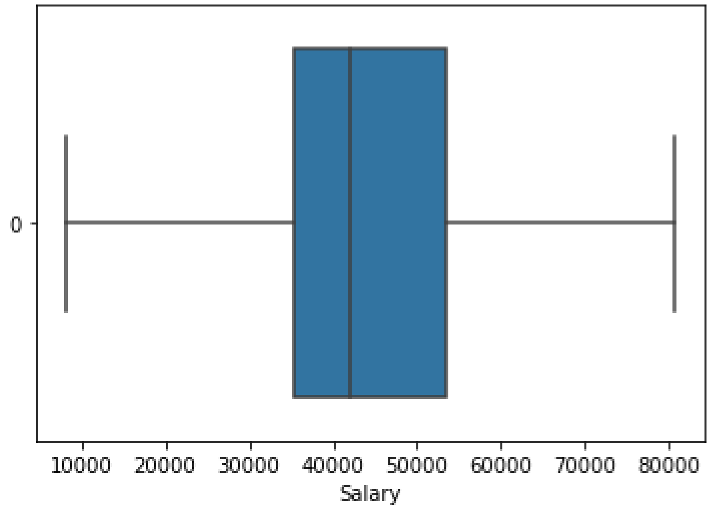




Except Age all variables contain outliers.

**Outlier treatment by IQR Method: -**





## 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).

### Logistic Regression:-

Using GridsearchCV, we input various parameters like 'max_iter', 'penalty', solver', 'tol' which will help us to find best grid for prediction of the better model

max_iter is an integer (100 by default) that defines the maximum number of iterations by the solver during model fitting.

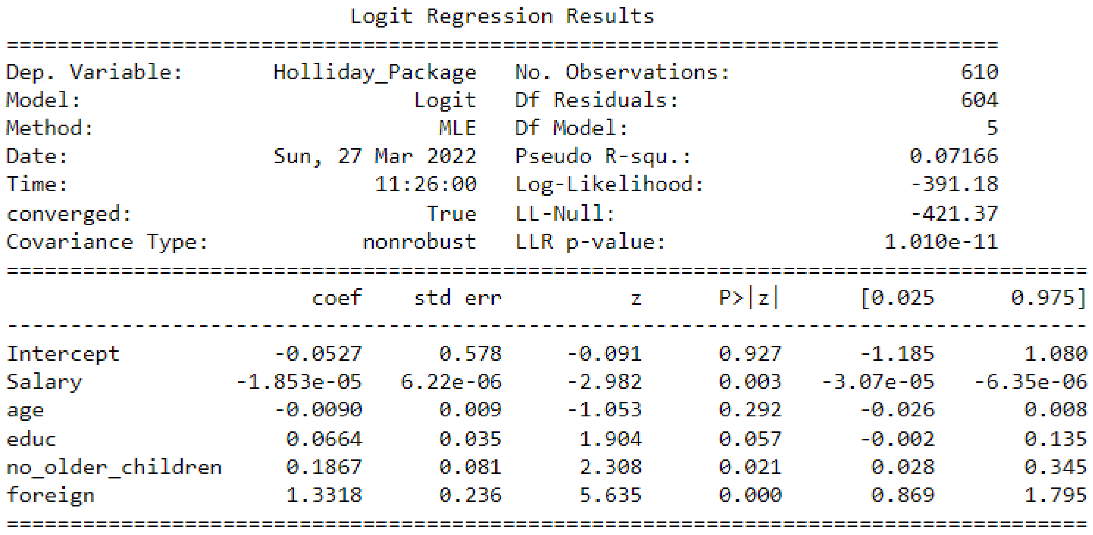
solver is a string ('liblinear' by default) that decides what solver to use for fitting the model. Other options are 'newton-cg', 'lbfgs', 'sag', and 'saga'.

penalty is a string ('l2' by default) that decides whether there is regularization and which approach to use. Other options are 'l1', 'elasticnet', and 'none'.

bestgrid:{'max_iter': 10000, 'penalty': 'none', 'solver': 'newton-cg', 'tol': 0.0001}

Accuracy score of training data:64%

Accuracy score of test data:62.9%



```
                           Logit Regression Results
==============================================================================
Dep. Variable:        Holliday_Package   No. Observations:                  610
Model:                           Logit   Df Residuals:                      604
Method:                            MLE   Df Model:                            5
Date:                 Sun, 27 Mar 2022   Pseudo R-squ.:                 0.07166
Time:                         11:26:00   Log-Likelihood:                -391.18
converged:                        True   LL-Null:                       -421.37
Covariance Type:             nonrobust   LLR p-value:                 1.010e-11
=====================================================================================
                        coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            -0.0527      0.578     -0.091      0.927      -1.185       1.080
Salary            -1.853e-05   6.22e-06     -2.982      0.003   -3.07e-05   -6.35e-06
age                  -0.0090      0.009     -1.053      0.292      -0.026       0.008
educ                  0.0664      0.035      1.904      0.057      -0.002       0.135
no_older_children     0.1867      0.081      2.308      0.021       0.028       0.345
foreign               1.3318      0.236      5.635      0.000       0.869       1.795
=====================================================================================
```

The summary table:

The summary table below, gives us a descriptive summary about the regression results.

foreign have coeffient of 1.3318 which shows foreign is important independent variable feature

This means that for a one-unit increase in foreign we expect a 1.3318 increase in the log-odds of the dependent variable holiday_package, holding all other independent variables constant.

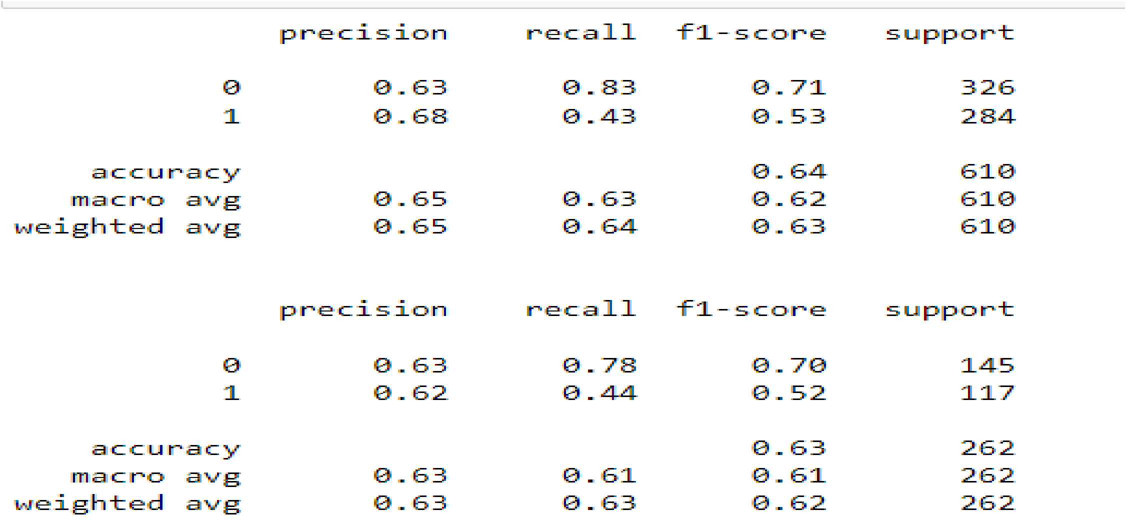
Std. Err. – These are the standard errors associated with the coefficients. Age has very low St. Err

foreign and salary having p-value <0.05. they are statistically significant.

## Linear Discriminant Analysis (LDA) algorithm: -

Accuracy score of training data:64.2%

Accuracy score of test data:62.9%



```
              precision    recall  f1-score   support

           0       0.63      0.83      0.71       326
           1       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

              precision    recall  f1-score   support

           0       0.63      0.78      0.70       145
           1       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
```

As shared above, the initial accuracy scores on the test data were 65%, which is also the same after the custom cut-off model but the F1 score improved from an initial value of 53% to 52% with the custom model. For all problems, accuracy might not be best possible metric based on which we can take a decision, so we need to be aware of other metrics such as precision, recall, f1- score, which becomes more relevant to choose the best LDA model.

### 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.

Confusion matrix cells are populated by the terms:

True Positive (TP)- The values which are predicted as True and are actually True.

True Negative (TN)- The values which are predicted as False and are actually False.

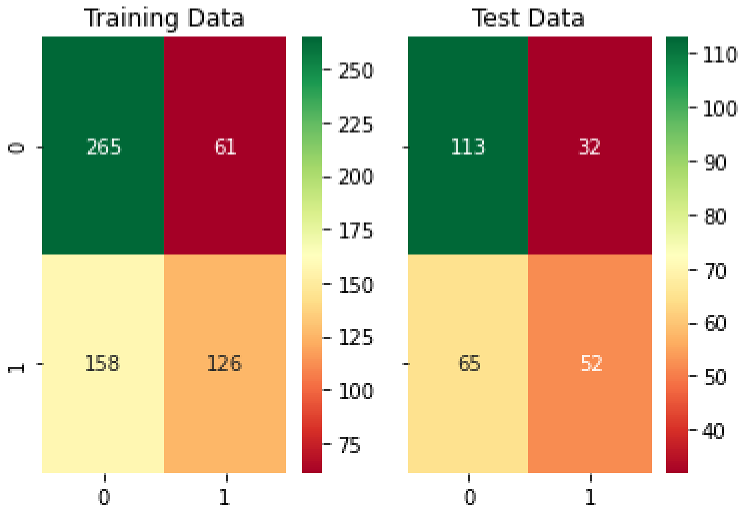
False Positive (FP)- The values which are predicted as True but are actually False.

False Negative (FN)- The values which are predicted as False but are actually True

ROC Curve- Receiver Operating Characteristic (ROC) measures the performance of models by evaluating the trade-offs between sensitivity (true positive rate) and false (1- specificity) or false positive rate.

AUC - The area under curve (AUC) is another measure for classification models is based on ROC. It is the measure of accuracy judged by the area under the curve for ROC

## Confusion matrix on the training and test data:-



Logistic regression Insights: -

Confusion matrix on the training and test data

Training data:

True Negative: 265 False Positive: 61

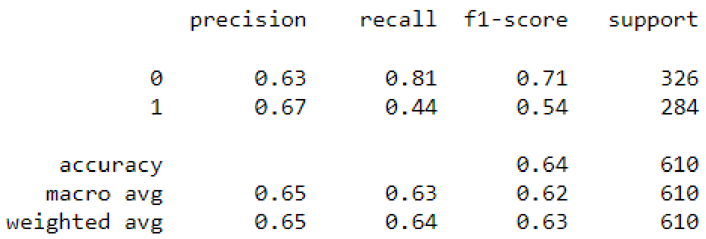
False Negative: 158 True Positive: 126

Test data:

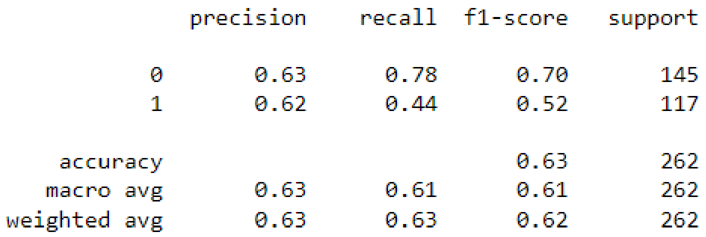
True Negative: 113 False Positive: 32

False Negative: 65 True Positive: 52

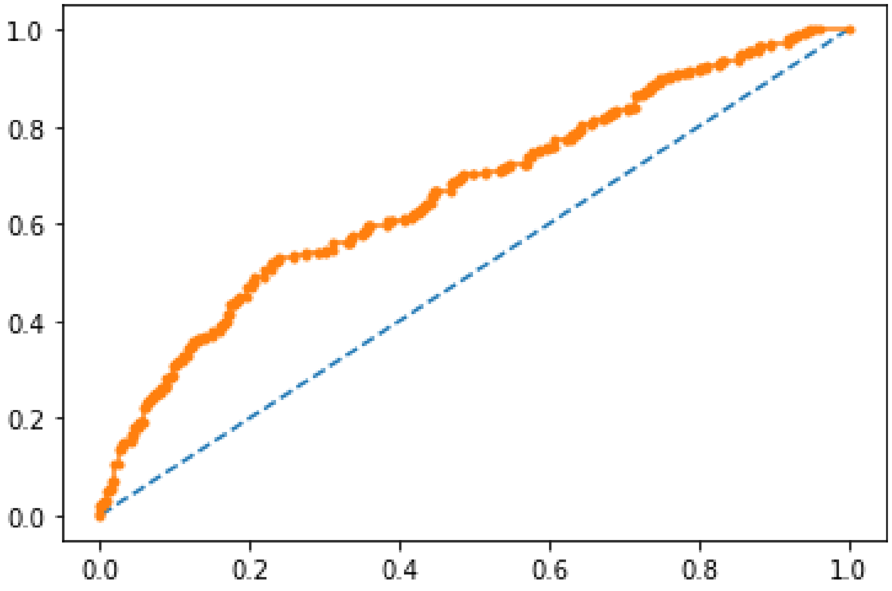
## Classification Report of training and test data: -



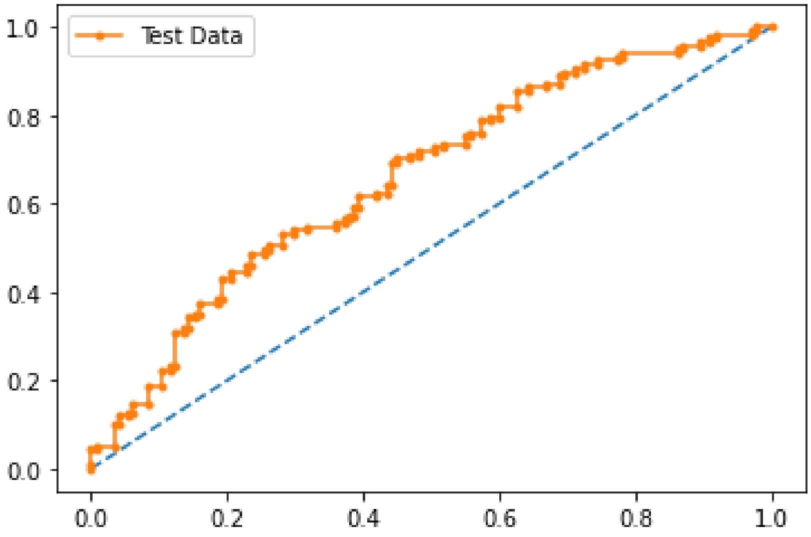
```
              precision    recall  f1-score   support

           0       0.63      0.81      0.71       326
           1       0.67      0.44      0.54       284

    accuracy                           0.64       610
   macro avg       0.65      0.63      0.62       610
weighted avg       0.65      0.64      0.63       610
```



```
              precision    recall  f1-score   support

           0       0.63      0.78      0.70       145
           1       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 for the Training Data: 0.667**



```
AUC for the Test Data: 0.661
```

Logistic regression

Train Data:

AUC: 66.7%

Accuracy: 64%

precision: 67%

recall: 44%

f1 :54%

Test Data:

AUC: 66.1%

Accuracy: 63%

Precision:  63

recall: 44%

f1: 52%

Training and Test set results are almost similar, this proves no overfitting or underfitting

The Precision and Recall metrics also almost similar between training and test set.

The coefficients for each of the independent attributes: -

The coefficient for Salary is -1.8534305725437244e-05
The coefficient for age is -0.008975252670392637
The coefficient for educ is 0.06639465493762979
The coefficient for no_young_children is 0.0
The coefficient for no_older_children is 0.18666772735970738
The coefficient for foreign is 1.331775458382739

Te sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable the dependent variable. A positive coefficient indicates that as the value of the independent variable increases, the mean of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

The coefficient for foreign is 1.33 which shows "foreign" variable is high positive correlation with "holiday_package" (dependent variables)

If more foreign employees opted for holiday package, no of employees opted for holiday_package will increase

There is more negative correlation between "holiday package " and salary, age variables

**Variance Inflation Factor: -**

```
Salary ---> 10.691328770445464
age ---> 7.883717090792149
educ ---> 9.289879244969368
no_young_children ---> nan
no_older_children ---> 1.8287587912872862
foreign ---> 1.3123489821120458
```

**INSIGHTS,**

Foreign and no older children are important.

LR_train_precision  0.67
LR_train_recall  0.44
LR_train_f1  0.54

LR_test_precision  0.62
LR_test_recall  0.44
LR_test_f1  0.52

LDA: -

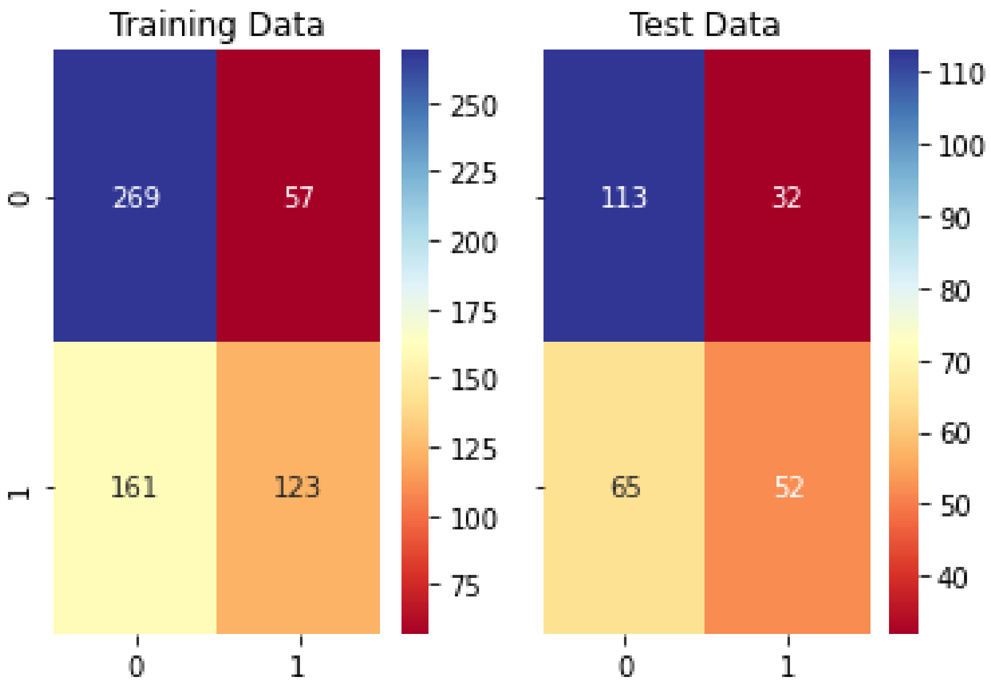
The coefficients for each of the independent attributes: -

```
The coefficient for Salary is -1.8319623552909e-05
The coefficient for age is -0.009003817204418353
The coefficient for educ is 0.06513802792962098
The coefficient for no_young_children is -3.784736982288265e-16
The coefficient for no_older_children is 0.18747103956268826
The coefficient for foreign is 1.3765145540963217
```

OUTPUT: -

```
LDA_train_precision  0.67
LDA_train_recall  0.44
LDA_train_f1  0.54


LDA_test_precision  0.62
LDA_test_recall  0.44
LDA_test_f1  0.52
```

CONFUSION MATRIX: -



LDA Inference: Confusion matrix on the training and test data

Training data:

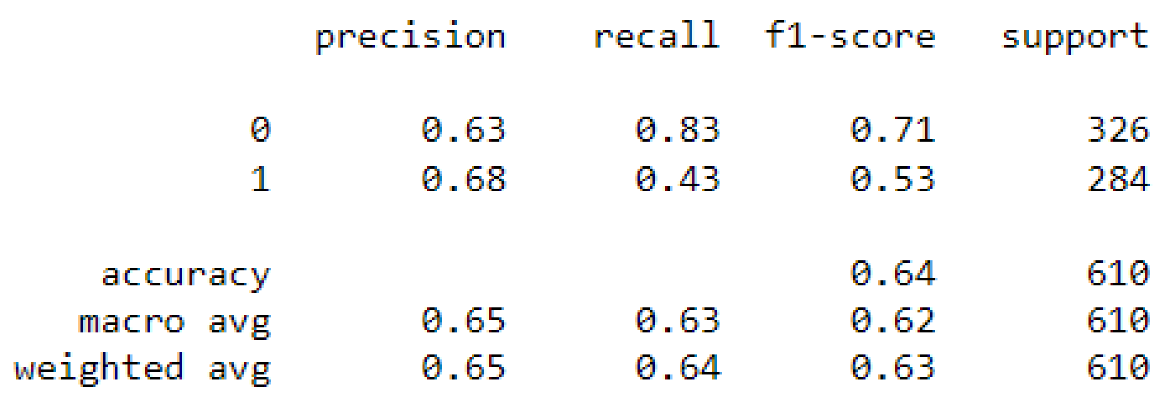
True Negative: 269 False Positive: 57

False Negative: 161 True Positive: 123

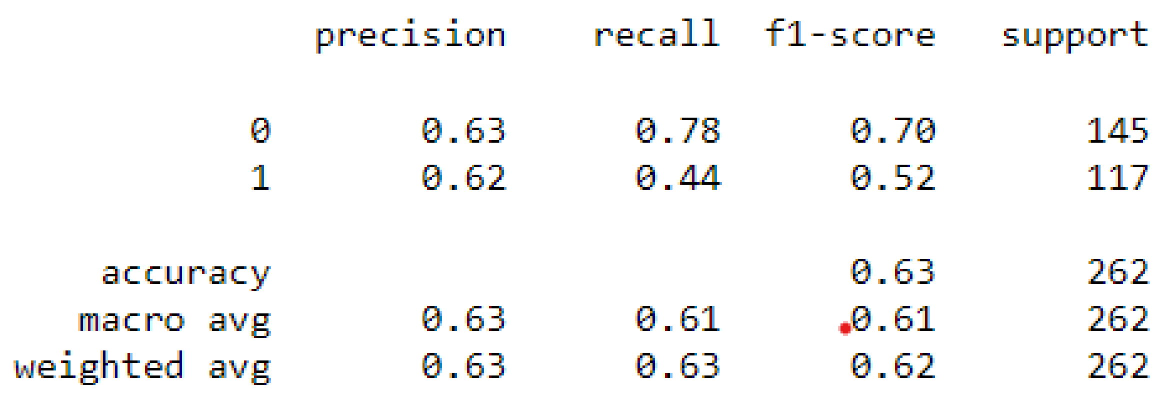
Test data:

True Negative: 113 False Positive: 32

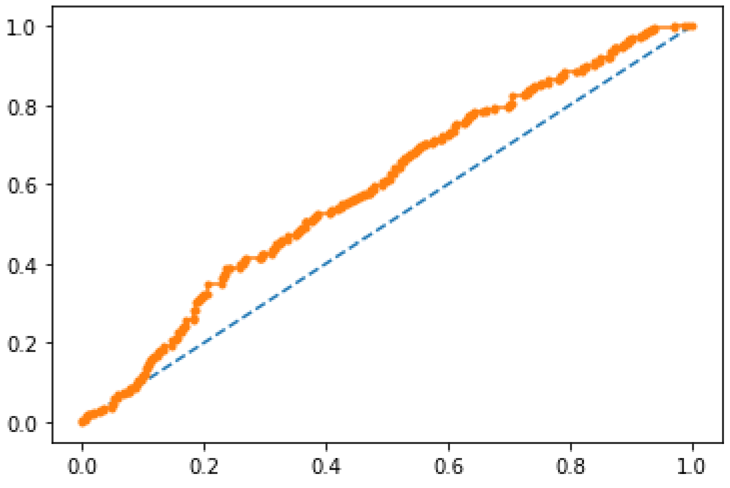
False Negative: 65 True Positive: 52



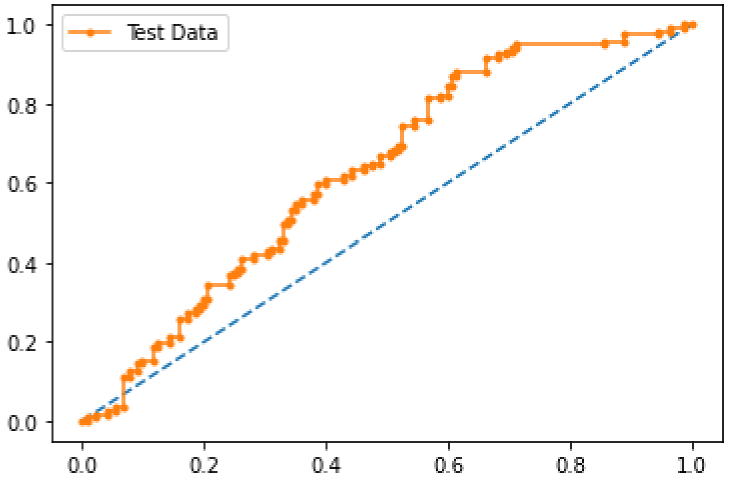
```
              precision    recall  f1-score   support

           0       0.63      0.83      0.71       326
           1       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
```



```
              precision    recall  f1-score   support

           0       0.63      0.78      0.70       145
           1       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
```

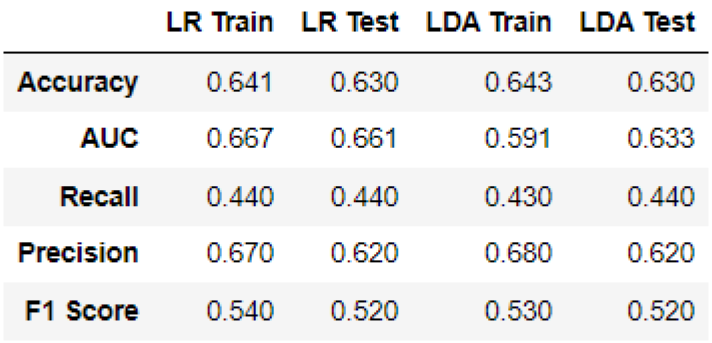


AUC for the Training Data: 0.591

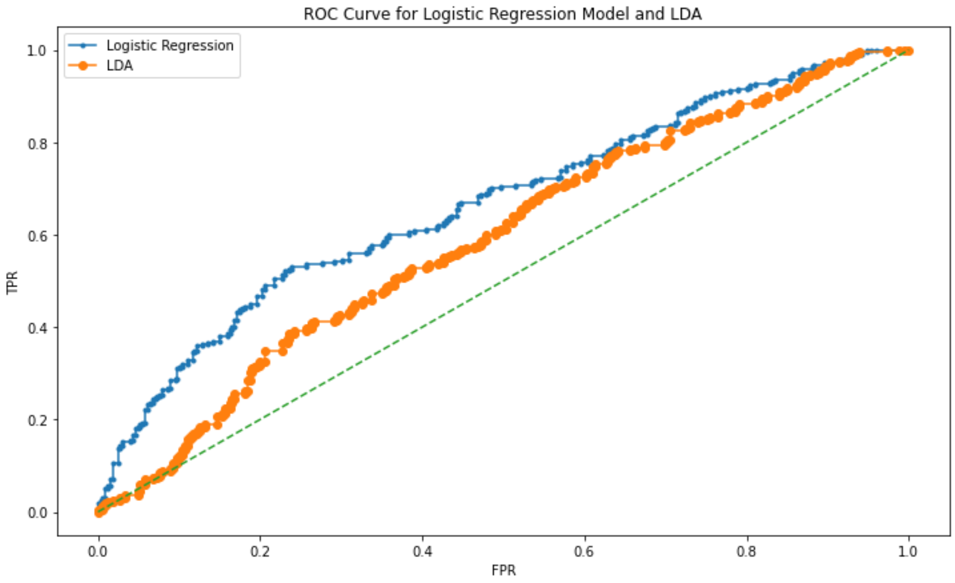


AUC for the Test Data: 0.633

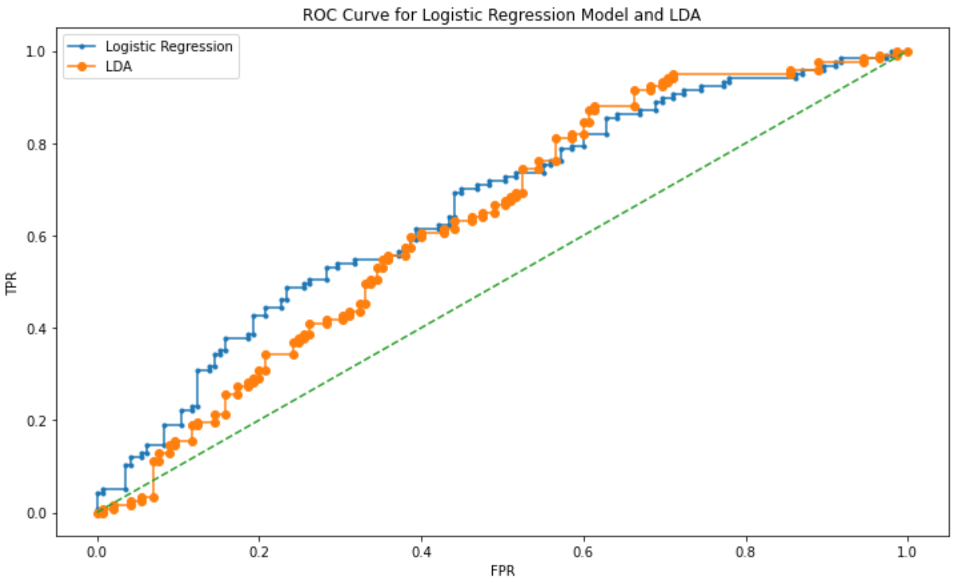
```
Logistic regression

Train Data: AUC: 59.1%, Accuracy: 64%, precision: 68%, recall: 43%, f1 :53%

Test Data: AUC: 63.3%, Accuracy: 63%, precision 63%, recall: 44%, f1: 52%

Training and Test set results are almost similar, this proves no overfitting or
underfitting The Precision and Recall metrics also almost similar between
training and test set.
```



| | LR Train | LR Test | LDA Train | LDA Test |
|---|---|---|---|---|
| **Accuracy** | 0.641 | 0.630 | 0.643 | 0.630 |
| **AUC** | 0.667 | 0.661 | 0.591 | 0.633 |
| **Recall** | 0.440 | 0.440 | 0.430 | 0.440 |
| **Precision** | 0.670 | 0.620 | 0.680 | 0.620 |
| **F1 Score** | 0.540 | 0.520 | 0.530 | 0.520 |



Area under the curve for Logistic Regression Model is 0.666864685042772
Area under the curve for LDA is 0.590976842650998



```
Area under the curve for Logistic Regression Model is 0.6610079575596816
Area under the curve for LDA is 0.6329501915708813
```

## 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.

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.

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.

**----- END OF THE REPORT------**