BUSINESS REPORT

NITIN KUMAR

[COMPANY NAME] [Company address]

Contents

Problem 1: Linear Regression

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

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, SI1, SI2, I1
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.

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.

Summary of the dataset the data set contains 26967 row and 11 columns. In the given data set there are 2 Integer type features,6 Float type features and 3 Object type features. Where 'price' is the target variable and all other are predictor variable. The first column is an index ("Unnamed: 0") as this only serial no, we can remove it. Except for the column depth, the rest null count is 26967.

EXPLORATORY DATA ANALYSIS

Step 1: Check and remove any duplicates in the dataset.

Before: -

```
df.duplicated().sum()
```

```
34
```

There are 34 duplicates values and we have dropped all the duplicates

After: -

```
df.duplicated().sum()
```

```
0
```

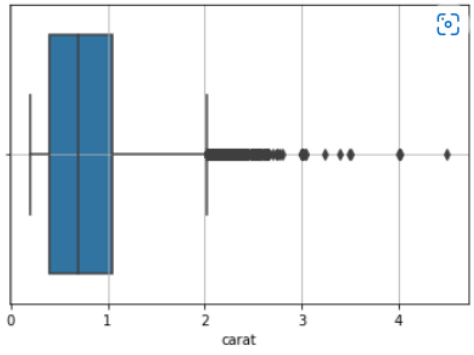
Step 2: Check and treat any missing values

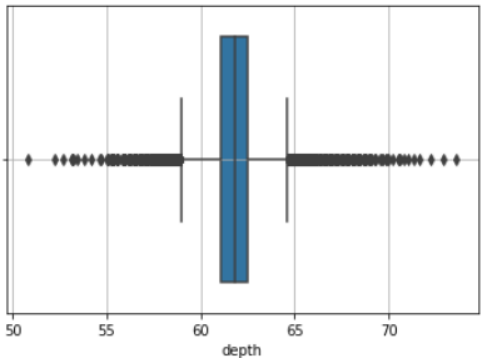
```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

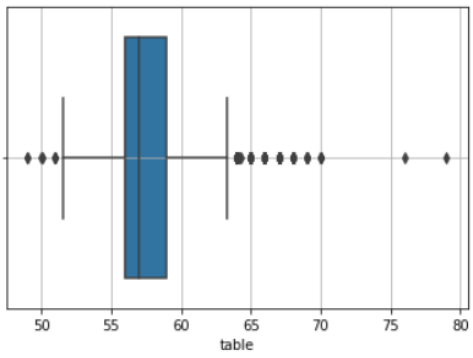
```
carat      0
cut        0
color      0
clarity    0
depth      0
table      0
x          0
y          0
z          0
price      0
dtype: int64
```

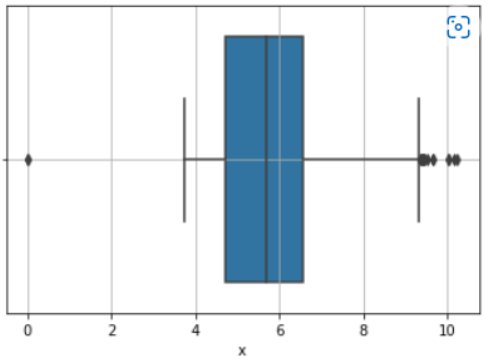
Step 3: Outlier Treatment Using the boxplot. There are many outliers in the data set and we are treating all the outliers

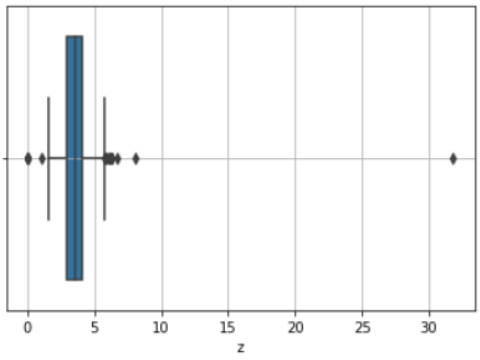
Before treating outliers:-

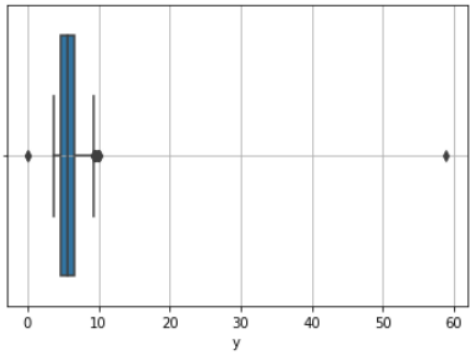


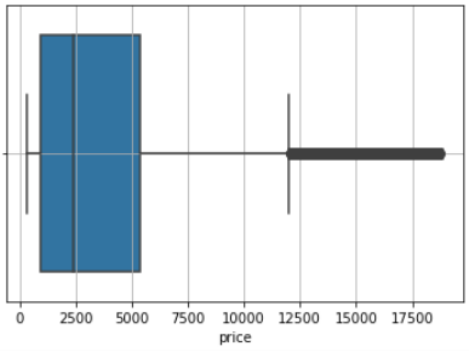




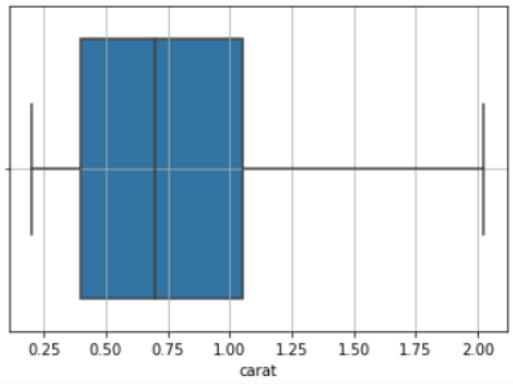


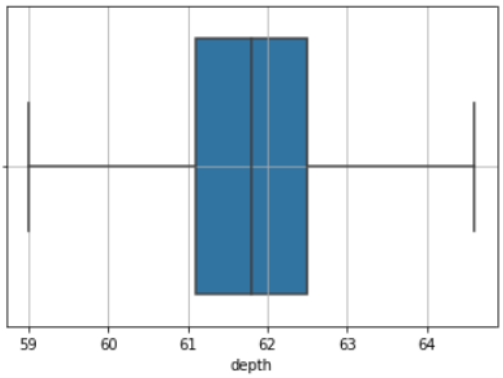


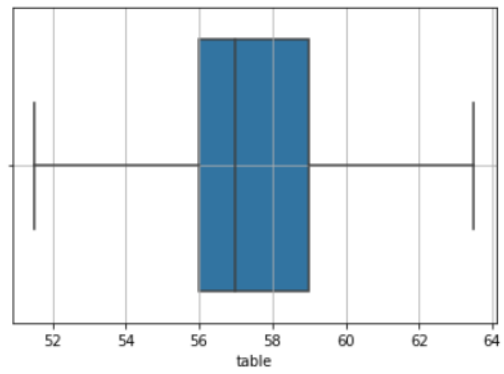


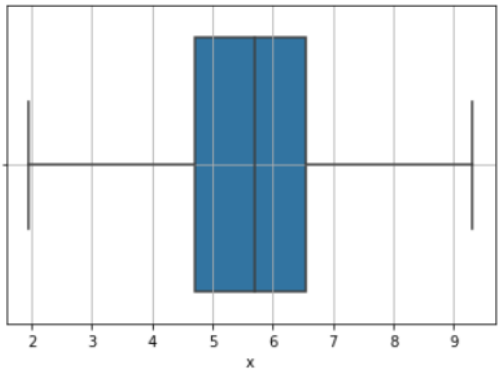


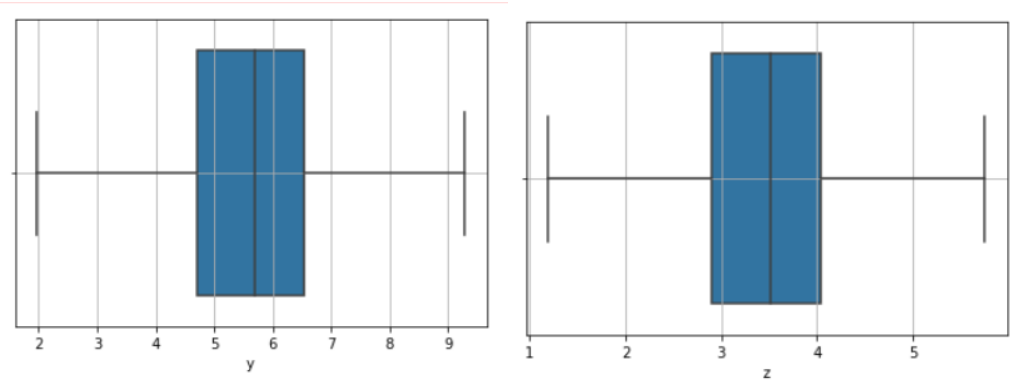
After treating outliers:-

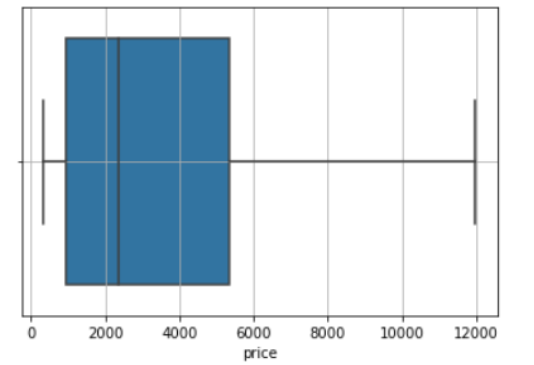










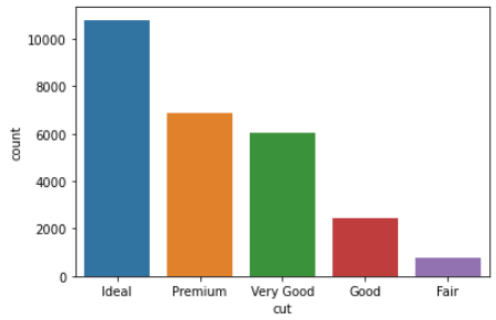


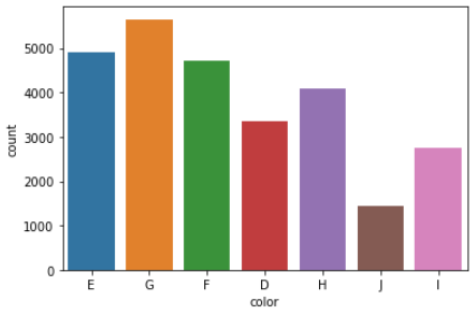
Step 4: Univariate Analysis: -

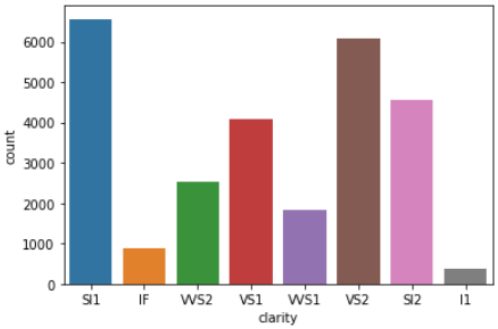
Observation-1: 'Price' is the target variable while all others are the predictors. In the given data set,3 Object type features. Where 'price' is the target variable and all other are predictor variable.
Observation-2: There are three object data type 'cut', 'colour' and 'clarity'.

Observation-3: We can observe there are 697 missing value in the depth column. There are some duplicate row presents. (33 duplicate rows out of 26958

Observation-4: There are 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.



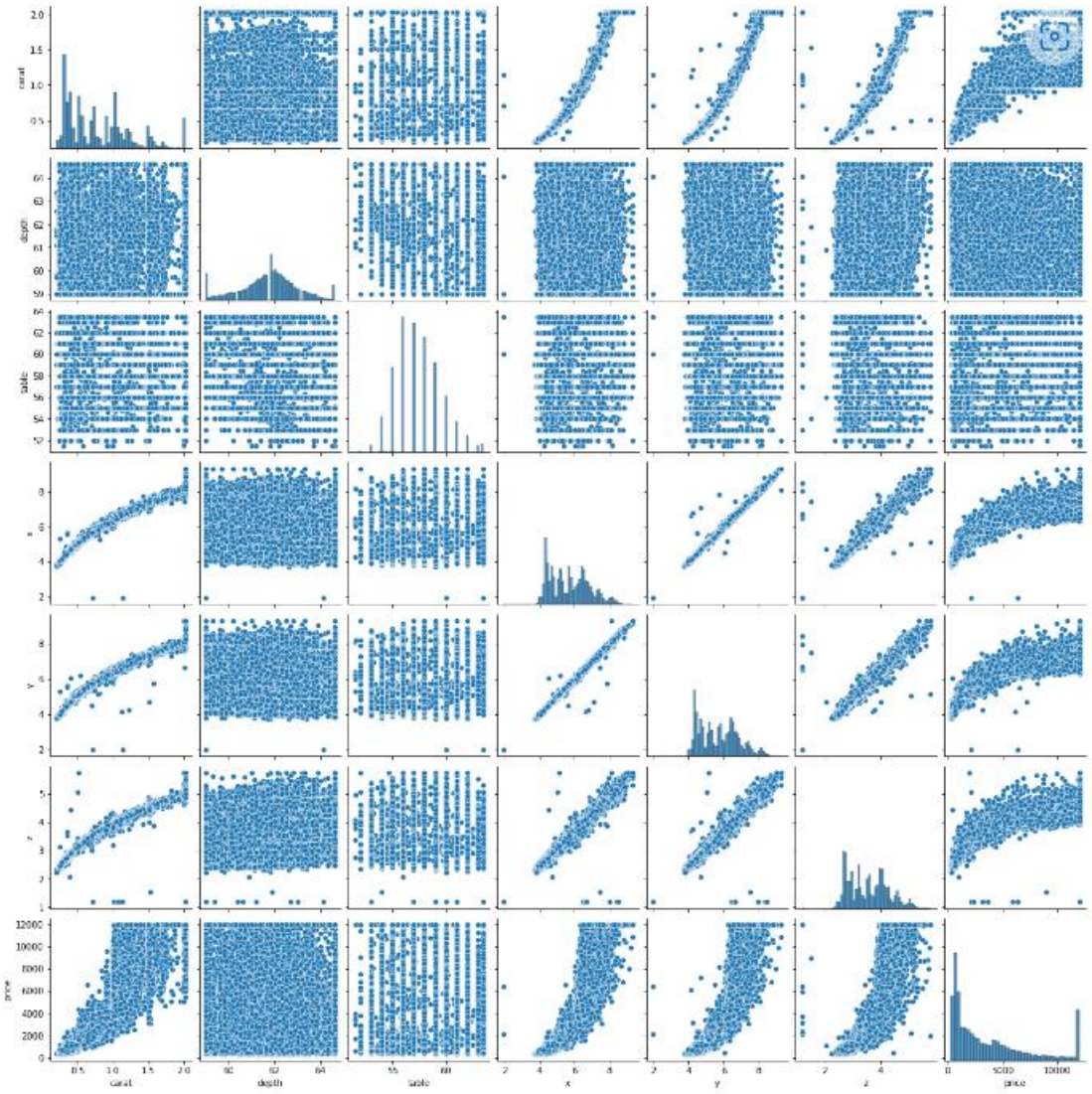




Step 5: Bi-variate Analysis

1>It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them.

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



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.

We start by checking through the dataset for any null values that are present as seen in Figure 8, it shows that there are a total of 697 null values in the depth column.

- Followed by which the median is computed for each attribute so that it can be used to replace the null values that are present in the dataset.
- In below given figure 9 we can see that the null values are replaced by the median that's computed.
- After the removing the null values the shape of the dataset becomes 26925 rows and 10 column

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	

Is scaling necessary in this case?

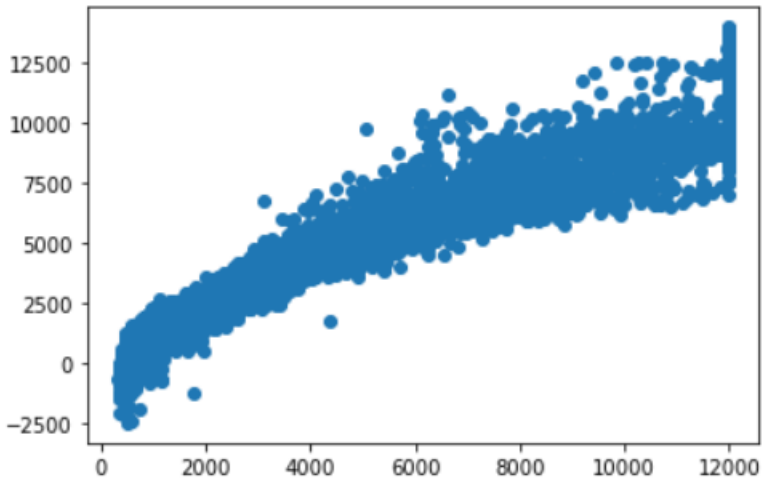
No, it is not necessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not

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

- Copy all the predictor variables into X data frame and copy target into the y data frame. Using the dependent variable, we split the X and Y data frames into training set and test set.
- For this we use the Sklearn package and then split X and Y in 70:30 ration and then invoke the linear regression function and find the best fit model on training data
- The intercept for our model is 1720.9060

- R square on training data: 0.9290035850988146
- R square on testing data: 0.9321174497713431
- R square is the percentage of the response variable variation that is explained by a linear mo del and computed by the formula as: R-square = Explained Variation / Total Variation
- It is always between 0 and 100%, in which 0% indicates that the model explains none of the variability of the response data around its mean and 100% indicates that the model explains all the variability of the response data around its mean.
- In the regression model we can see the R-square value on training and test data respectively as 0.931 1935886926559 and - 0.931543712584074.
- From the scatter plot, we see that it is a linear and there is very strong correlation present between th e predicted y and actual y
- RMSEV for training data set:- 923.4235946201566
- RMSEV for testing data set:- 923.7421306234233

- From the scatter plot, we see that it is a linear and there is very strong correlation present between th e predicted y and actual y.



-
- It also indicates that there's a lot spread which indicates some unexplained variances on the output.
- coefficients for each of the attributes

```
The coefficient for carat is 8775.226756064918
The coefficient for cut is 191.67829619067368
The coefficient for color is -393.0300288402921
The coefficient for depth is -17.34532907596946
The coefficient for table is -25.44500804357187
The coefficient for x is -1540.001830327264
The coefficient for y is 1293.8316218547159
The coefficient for z is -215.783500988086
The coefficient for clarity_0 is 1399.0178135123015
The coefficient for clarity_1 is 1143.6970745580632
The coefficient for clarity_2 is 517.8984727775139
The coefficient for clarity_3 is -448.13636676012345
The coefficient for clarity_4 is -2612.476994087755
```

Linear Regression using stats models

- Assuming the null hypothesis is true, i.e. price from that universe we have drawn coefficient for the va riable shown above.
- Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real w orld the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting HO and a ccepting Ha that at least 1 regression co-efficient is not '0'. Here all regression co-efficient are not '0'

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	2.054e+04
Date:	Sun, 20 Feb 2022	Prob (F-statistic):	0.00
Time:	12:58:19	Log-Likelihood:	-1.5548e+05
No. Observations:	18853	AIC:	3.110e+05
Df Residuals:	18840	BIC:	3.111e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1434.0884	575.776	2.491	0.013	305.517	2562.660
carat	8775.2268	82.510	106.353	0.000	8613.499	8936.955
cut	191.6783	14.311	13.393	0.000	163.627	219.730
color	-393.0300	6.615	-59.418	0.000	-405.995	-380.065
depth	-17.3453	9.126	-1.901	0.057	-35.233	0.542
table	-25.4450	3.647	-6.978	0.000	-32.593	-18.298
x	-1540.0018	123.613	-12.458	0.000	-1782.294	-1297.710
y	1293.8316	122.120	10.595	0.000	1054.466	1533.198
z	-215.7835	97.408	-2.215	0.027	-406.712	-24.855
clarity_0	1685.8355	117.864	14.303	0.000	1454.812	1916.859
clarity_1	1430.5148	115.886	12.344	0.000	1203.367	1657.662
clarity_2	804.7162	116.434	6.911	0.000	576.496	1032.937
clarity_3	-161.3187	117.326	-1.375	0.169	-391.288	68.650
clarity_4	-2325.6593	127.032	-18.308	0.000	-2574.654	-2076.665

Omnibus:	3909.882	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11687.907
Skew:	1.078	Prob(JB):	0.00
Kurtosis:	6.199	Cond. No.	5.50e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 4.48e-28. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

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

- We can see that the from the linear plot, very strong corelation 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
- The Gem Stones company should consider the features 'Carat', 'Cut', 'colour', '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 category 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 profitable 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.

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.

• We have no null values in the dataset.

• We have integer and object data

The data that we have is of integer and continuous data, here the holiday package is our target variable. Salary, age, educe and number young children, number older children of employee have the went to foreign, those are the given attributes we have to cross examine and help the company predict weather the person will opt for holiday package or not.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   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
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

NULL VALUES

```
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

There are no null values in the dataset CHECK FOR DUPLICATES IN THE GIVEN DATASET Number of duplicate rows = 0

```
dups = df.duplicated()
print('Number of duplicate rows = %d' % (dups.sum()))
```

```
Number of duplicate rows = 0
```

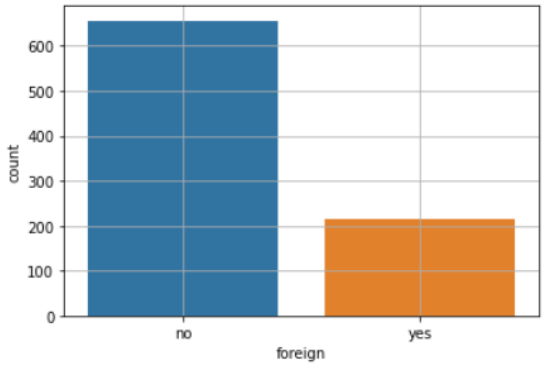
Unique values for categorical variables

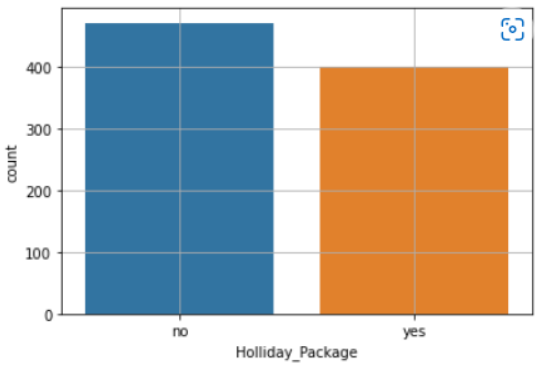
```
Holliday_Package
no     471
yes    401
Name: Holliday_Package, dtype: int64


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

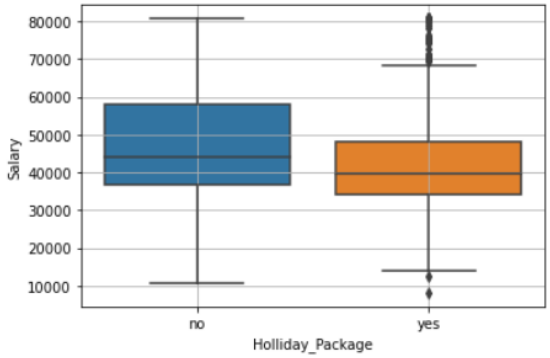
UNIVARIATE ANALYSIS:

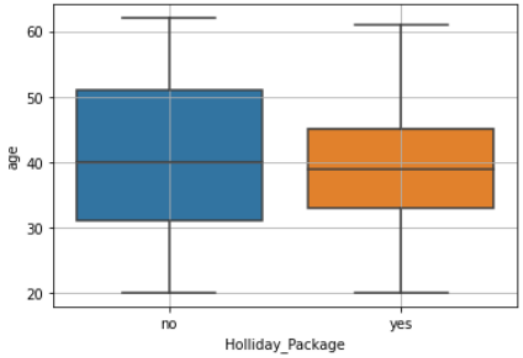
CATOGORICAL UNIVARIATE ANALYSIS

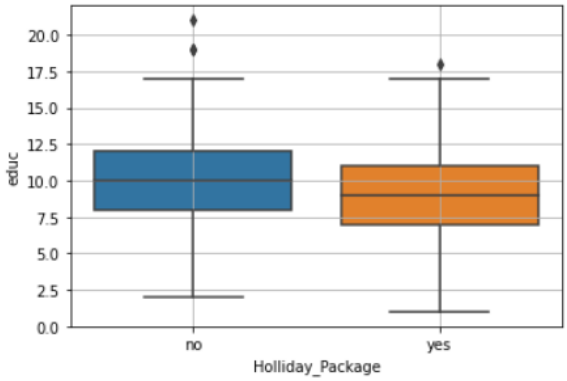
• As we can observe people with salaries below 150000 prefer holiday package. • Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package

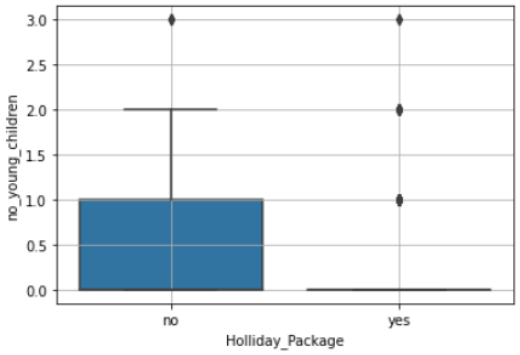




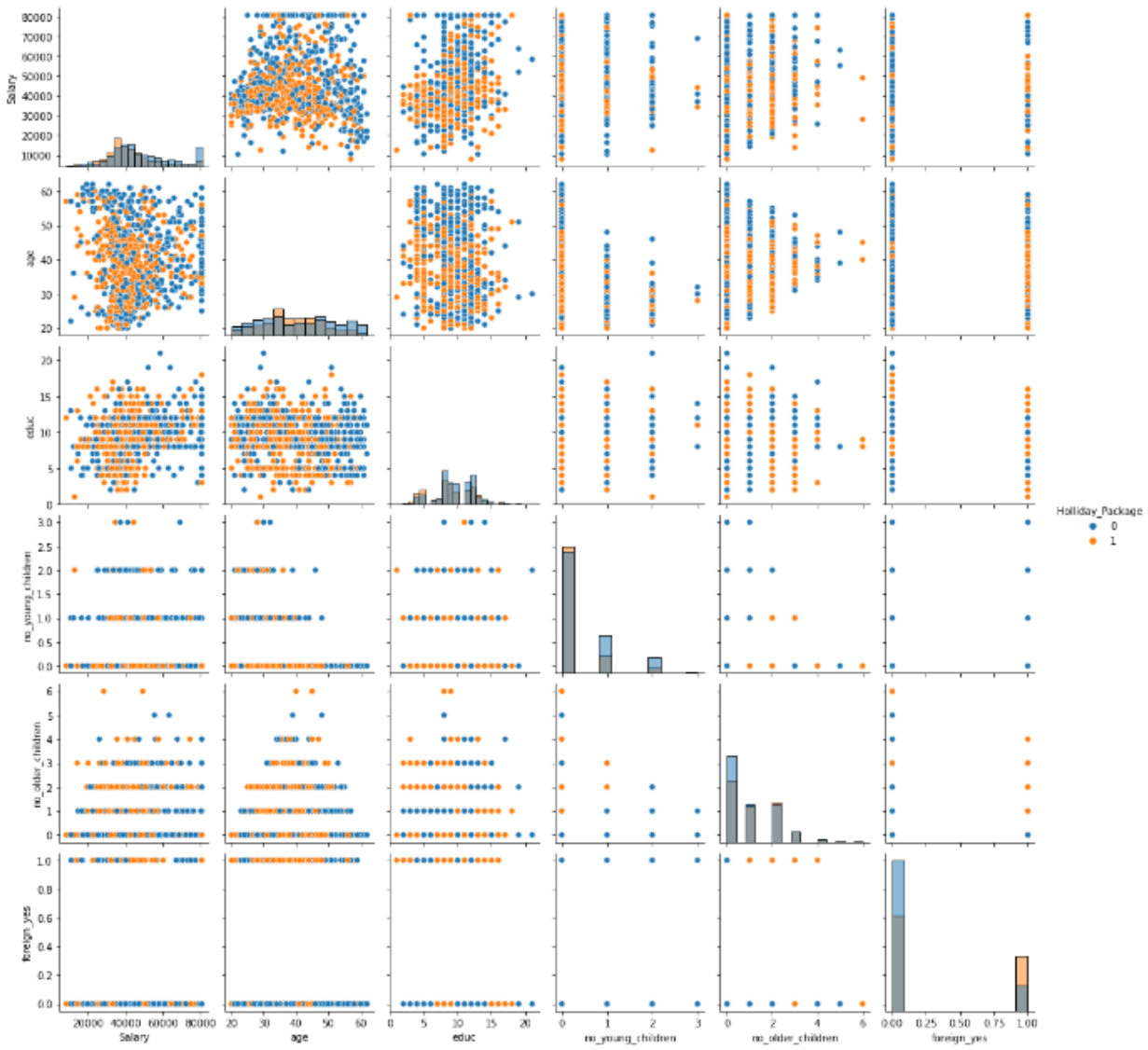






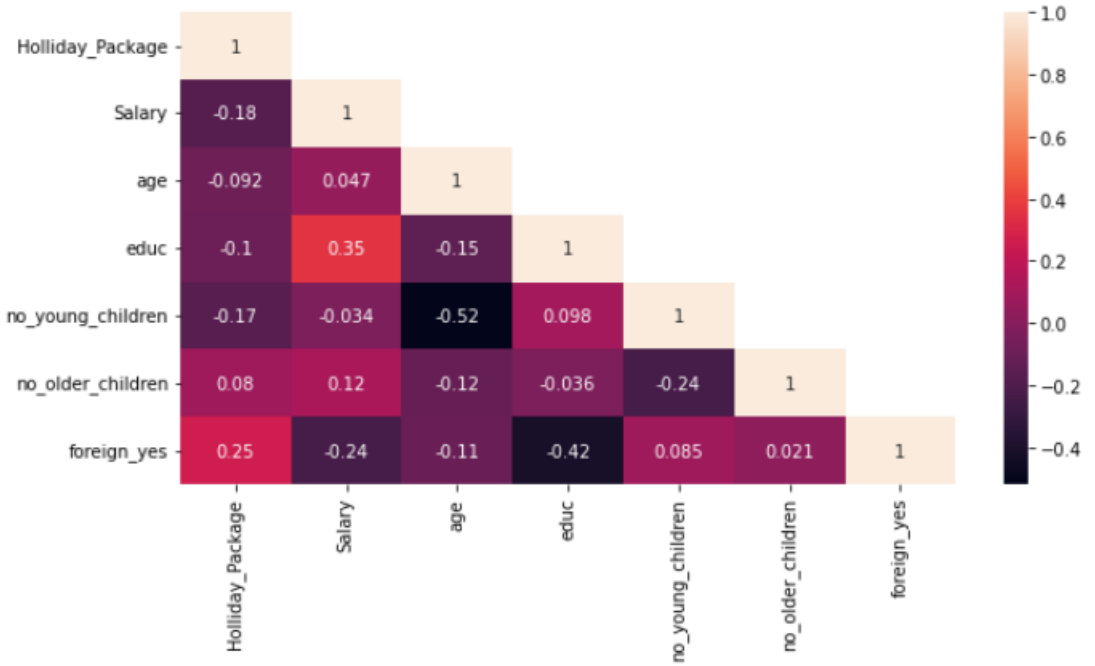


BIVARITE ANALYIS DATA DISTRIBUTION



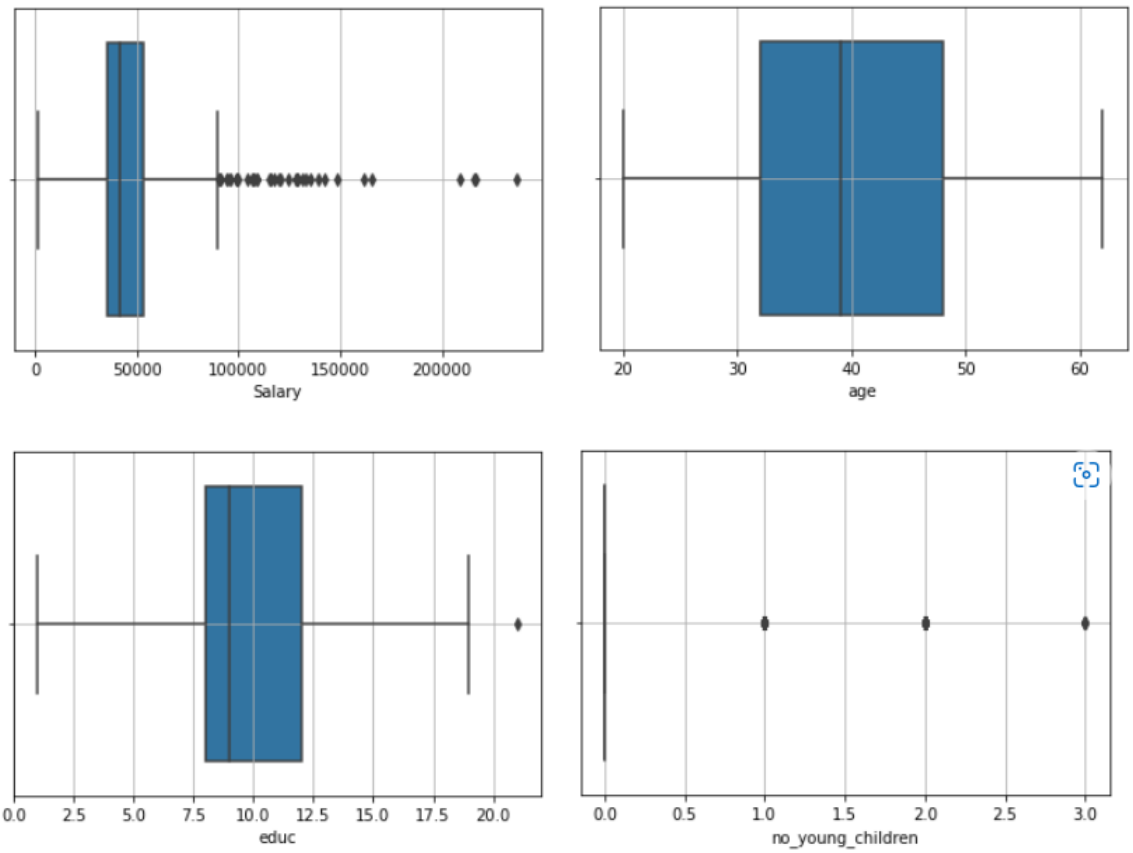
There is hardly any correlation between the data, the data seems to be normal. There is no huge difference in the data distribution among the holiday package, I don't see any clear two different distributions in the dataset provided

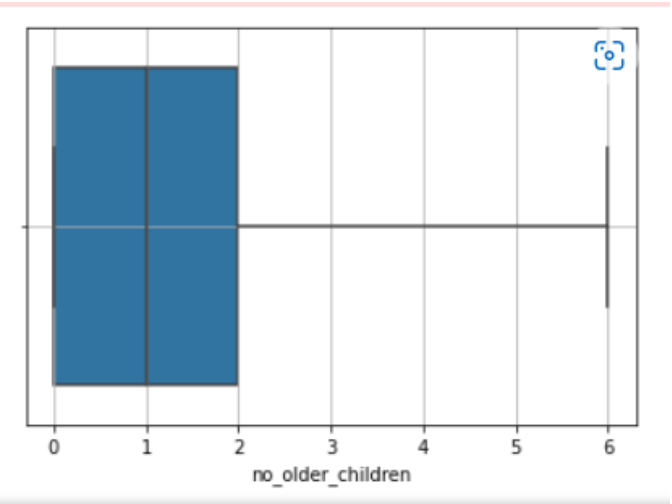
CHECKING FOR CORRELATION



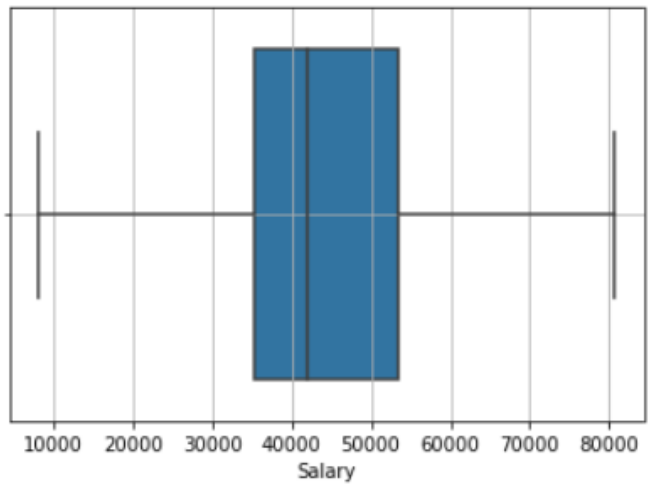
OUTLIER TREATMENT:

BEFORE:-





AFTER:-



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).

Here we have done ONE HOT ENCODING to create dummy variables and we can see all values for foreigners yes are 0. Better results are predicted by logistic regression model if encoding is done. Train/ Test split We will split the data in 70/30 ratio

Data before processing

	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

Data after processing:-

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign_yes
0	0	48412.0	30	8	1	1	0
1	1	37207.0	45	8	0	1	0
2	0	58022.0	46	9	0	0	0
3	0	66503.0	31	11	2	0	0
4	0	66734.0	44	12	0	2	0

Predicting on Training and Test dataset

```
Ytrain_predict = model.predict(X_train)
Ytest_predict = model.predict(X_test)
```

```
X_train
```

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
821	38974.0	47	12	0	2	1
805	40270.0	33	8	2	0	1
322	32573.0	30	11	1	0	0
701	43839.0	43	11	0	1	1
773	33060.0	40	5	1	1	1
...	...	...	...	...	...	...
594	42369.0	47	9	0	1	0
297	44207.0	45	12	0	2	0
76	50291.0	34	10	0	2	0
831	33434.0	44	7	0	1	1
187	36832.0	40	8	0	2	0

610 rows × 6 columns

Test:

	Salary	age	educ	no_young_children	no_older_children	foreign_yes
264	25118.0	58	8	0	0	0
189	40913.0	20	9	1	0	0
643	28446.0	58	8	0	0	0
65	36072.0	35	4	0	2	0
241	52736.0	40	10	0	3	0
...	...	...	...	...	...	...
165	34878.0	29	14	1	1	0
100	61159.0	38	10	0	3	0
503	41167.0	44	9	0	2	0
431	41769.0	43	9	0	0	0
119	46856.0	44	9	0	3	0

262 rows × 6 columns

Getting the Predicted Classes and Probs

```
Ytest_predict_prob=model.predict_proba(X_test)
pd.DataFrame(Ytest_predict_prob).head()
```

	0	1
0	0.677845	0.322155
1	0.534493	0.465507
2	0.691845	0.308155
3	0.487745	0.512255
4	0.571939	0.428061

Linear Discriminant Analysis

Building LDA model

```
Number of rows and columns of the training set for the independent variables: (610, 6)
Number of rows and columns of the training set for the dependent variable: (610,)
Number of rows and columns of the test set for the independent variables: (262, 6)
Number of rows and columns of the test set for the dependent variable: (262,)
```

```
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train,Y_train)
```

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.

PEFORMANCE METRICS FOR LINEAR REGRESSION

Confusion matrix on the training data

```
array([[244,  85],
       [118, 163]], dtype=int64)
```

Confusion matrix on the testing data

```
array([[108,  34],
       [ 58,  62]], dtype=int64)
```

Classification matrix on the training data

Accuracy - Training Data-67

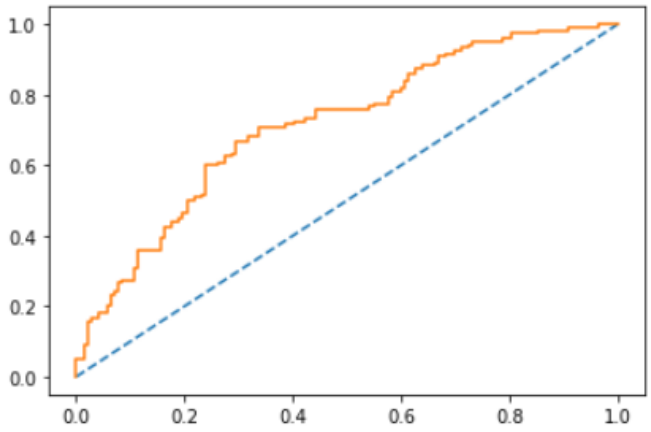
	precision	recall	f1-score	support
0	0.67	0.74	0.71	329
1	0.66	0.58	0.62	281
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Classification matrix on the testing data

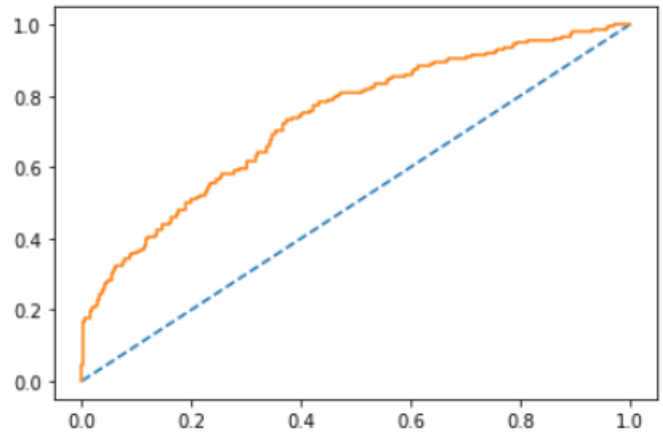
Accuracy - Testing Data-65

	precision	recall	f1-score	support
0	0.65	0.76	0.70	142
1	0.65	0.52	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.64	262

AUC and ROC for the testing data



AUC and ROC for the training data



Metrics for train data

lr_train_precision 0.65

lr_train_recall 0.52

lr_train_f1 0.57

Metrics for test data

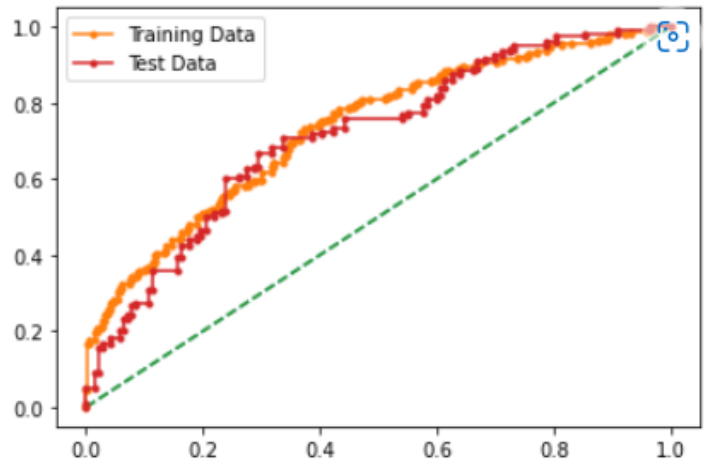
lr_test_precision 0.67

lr_test_recall 0.58

lr_test_f1 0.52

AUC and ROC for the training data and testing data

```
AUC for the Training Data: 0.733
AUC for the Test Data: 0.715
```



Comparing both these models, we find both results are same, but LDA works better when there is category target variable. As we can see the results for AUC/ROC for both the models are almost equivalent to each other So it is very difficult to differentiate between the two . The scores are also almost at par with each other . Both the models are working perfectly at par with each other. Since LDA works better with categorical values so we will pick it in this situation.

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

While doing EDA we found out that

• Most of the employees who are above 50 don't opt for holiday packages. It seems like they are not interested in holiday packages at all.

• Employees who are in the age gap of 30 to 50 opt for holiday packages. It seems like young people believe I spending on holiday packages so age here plays a very important role in deciding whether they will opt for package or not

• Also, people who have salary less than 50000 opt for holiday packages. So, salary is also a deciding factor for the holiday package.

• Education also plays an important role in deciding the holiday packages.

• To improve our customer base, we need to look into those factors

Recommendations

As we already have the customer base who are of the age of 30 to 50 so we need to look for the options and target the older people and the people who are earning more than 150000.

• As we know most of the people who are older prefer to visit religious places so it would be better if we target those places and provide them with packages where they can visit religious places.

• We can also look into the family dynamics of the people of the older people, if the older people have elder children e.g. 30 to 40 they can use the holiday packages so the deal should include the family package.

• People who earn more than 150000 don’t spend much on the holiday packages, they tend to go for lavish holidays and we can provide them with customized packages according to