Predictive Modeling Project

Balasubramaniyam, R.

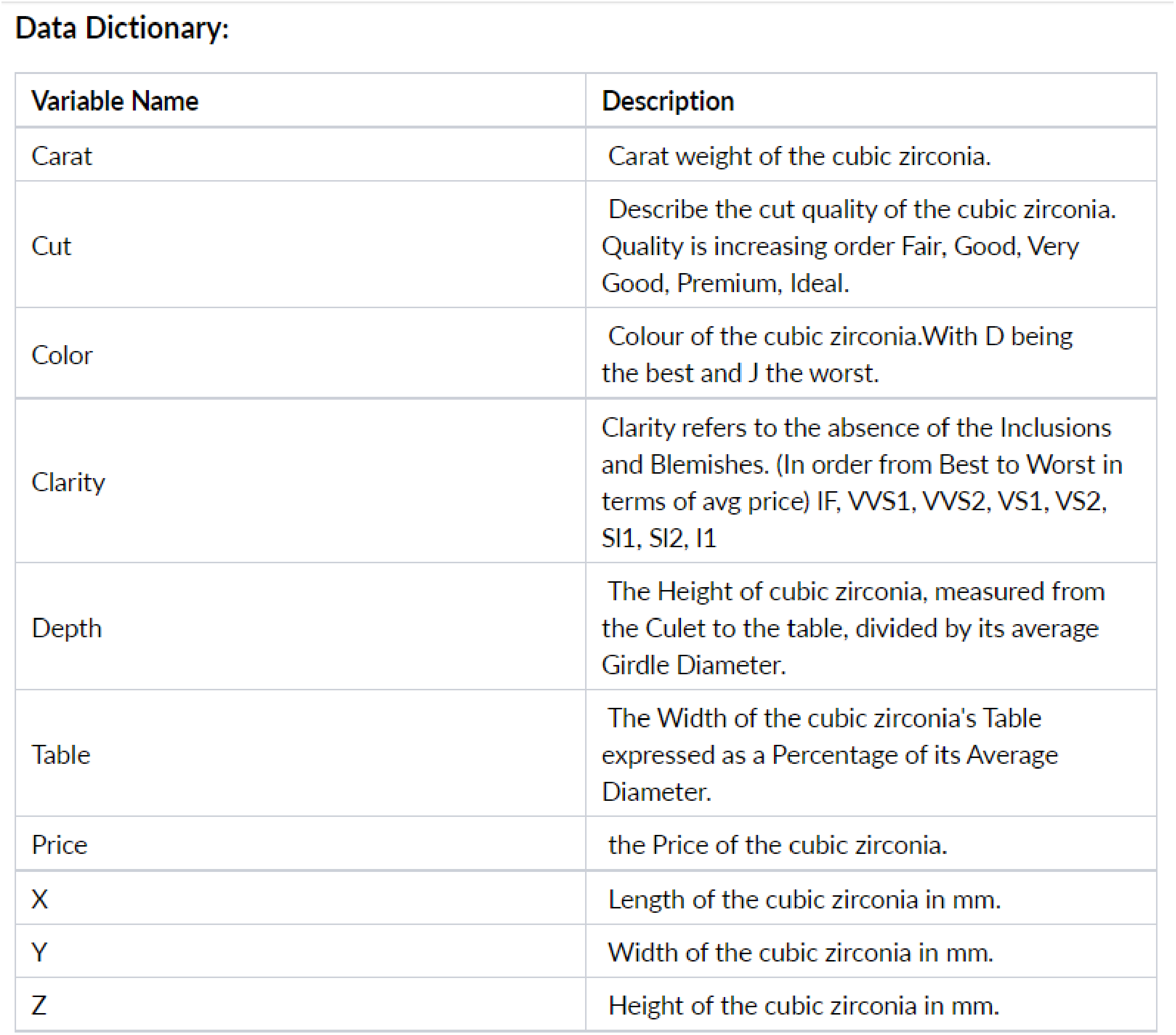
10/16/22

INDEX

Tabulation and figures

Problem 1: Linear Regression

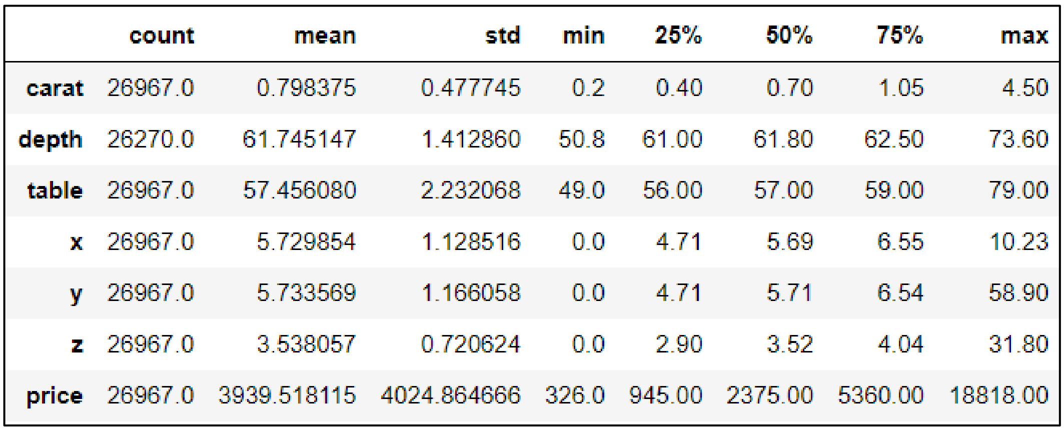
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is 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 best and J the worst.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, 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.

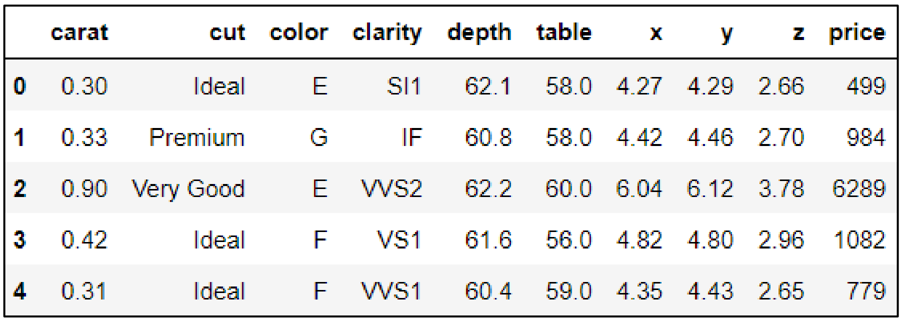


	count	mean	std	min	25%	50%	75%	max
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

Data summary

Describing the dataset

The data set contains 26967 row and 11 columns of independent variable. The column (Unnamed: 0) is not necessary for our prediction model, therefore we can drop that column during the data preparation stage. The data set contains one Integer type features, six Float type features and three Object type features. The 'Price' is the target variable and all other are predictor variable. The average weight of the cubic zirconia is 0.79 and from the summary the average price of the 3939.52, maximum price is 18818, minimum price is 326.



	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

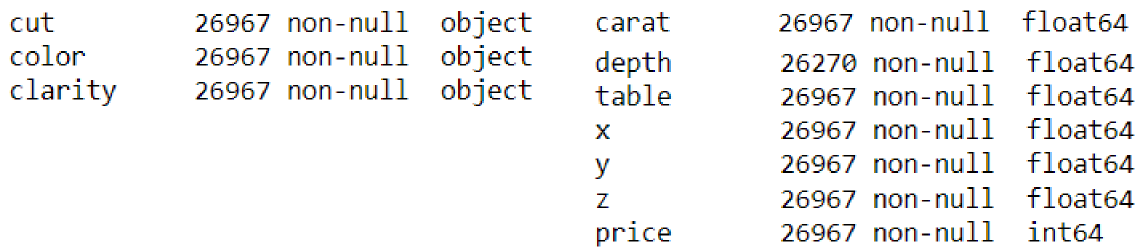
Loaded data

EXPLORATORY DATA ANALYSIS

After analysing the data for null and duplicate values, it is found that there are 697 null values in depth and 34 duplicate rows in the dataset.The null value in depth column are filled with the median value.

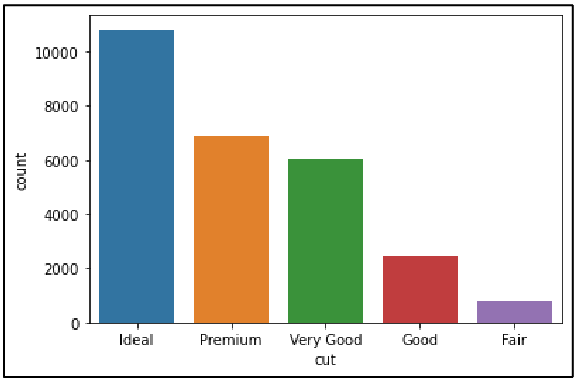
Shape of our dataset after removing (Unnamed: 0) is as follows,
No. of rows: 26967
No. of columns: 10

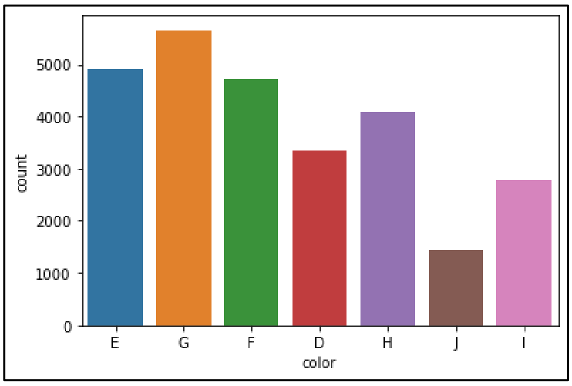
The univariate and bivariate analysis can be done on our dataset as per the datatype of the columns.

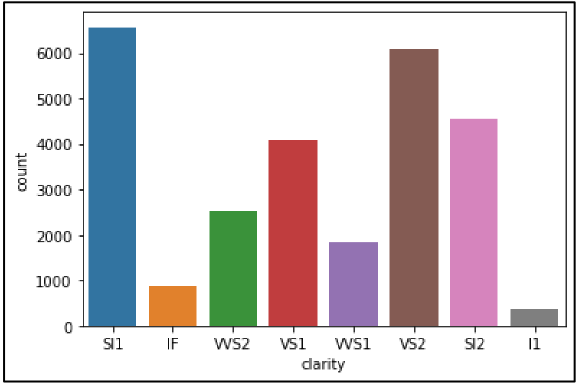


```
cut         26967 non-null  object     carat      26967 non-null  float64
color       26967 non-null  object     depth      26270 non-null  float64
clarity     26967 non-null  object     table      26967 non-null  float64
                                       x          26967 non-null  float64
                                       y          26967 non-null  float64
                                       z          26967 non-null  float64
                                       price      26967 non-null  int64
```

Datatype of Features



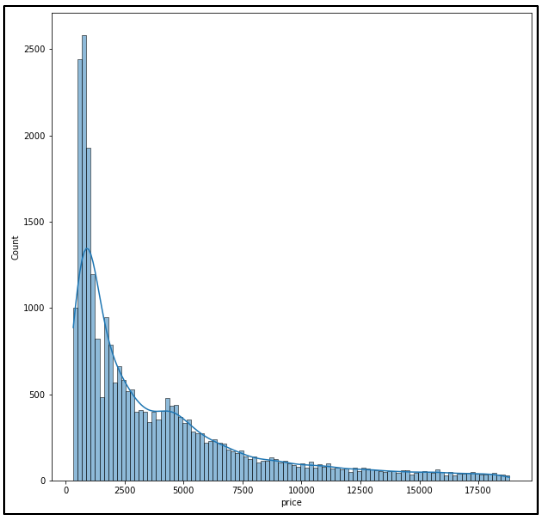




Count plots of object data

The Cut, Color, Clarity of cubic zirconia in our data are

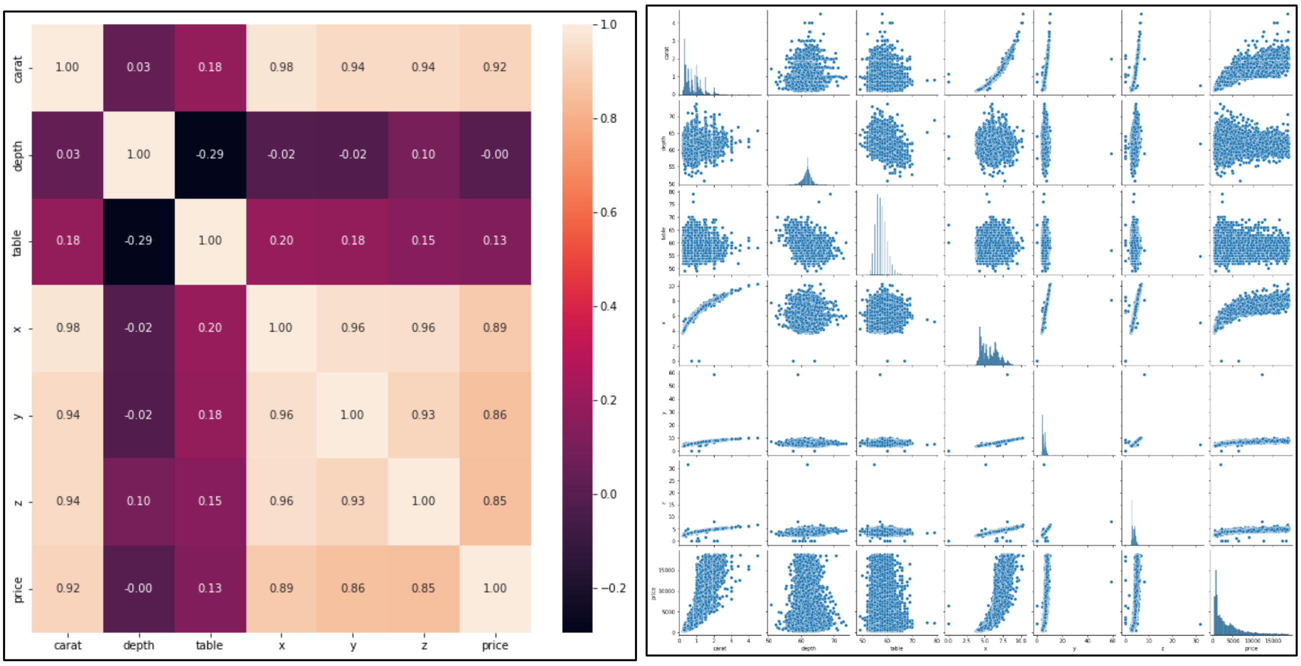
Cut	**Color**	**Clarity**
Ideal-10816	G-5661	SI1-6571
Premium-6899	E-4917	VS2-6099
Very Good-6030	F-4729	SI2-4575
Good-2441	H-4102	VVS2-2531
Fair-781	D-3344	VVS1-1839
	I-2771	IF-894
	J-1443	I1-365
		VS1-4093



Price Histogram

The Distribution of price data is right skewed and from the histogram we can see the greatest number of cubic zirconia sold are in the price range of (326 to 2500).

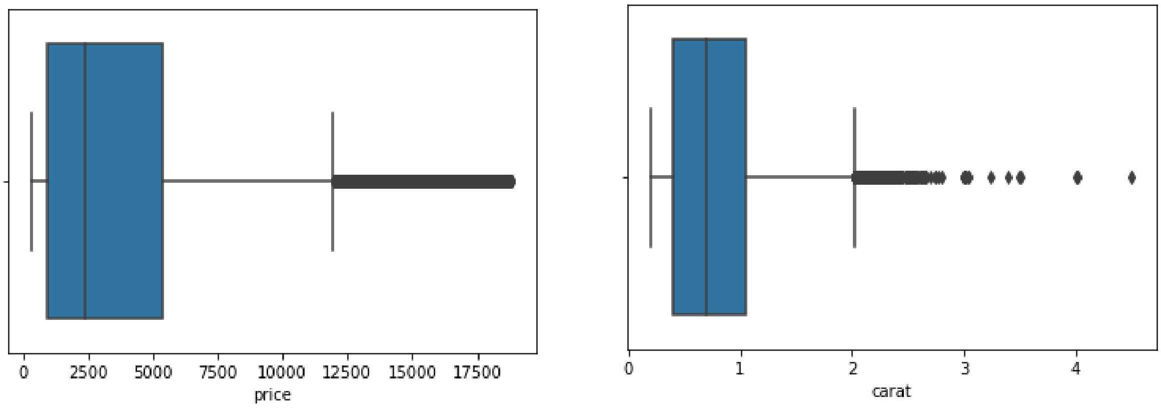
Now let’s study the relationship between the Target variable: Price and the other Predictor variables in our dataset.



Heatmap and Pair plot,

From the heatmap and pair plot, Carat and Price are 92% corelated, x, y, z is strongly corelated with price. It is observed that the depth and table are having a negative/weak corelation with price.

Outliers are present in the dataset, to maintain most data points we are proceeding without outlier treatment.

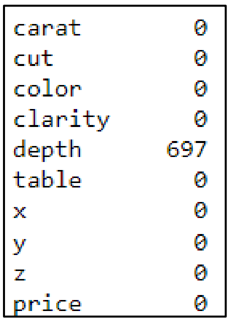


Outliers in dataset-Not treated

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.

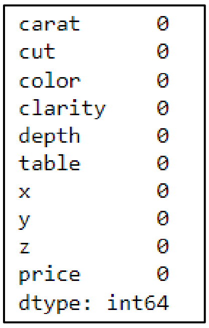
There were 697 null values present in depth column and they are imputed with the median values.

Before imputing null values



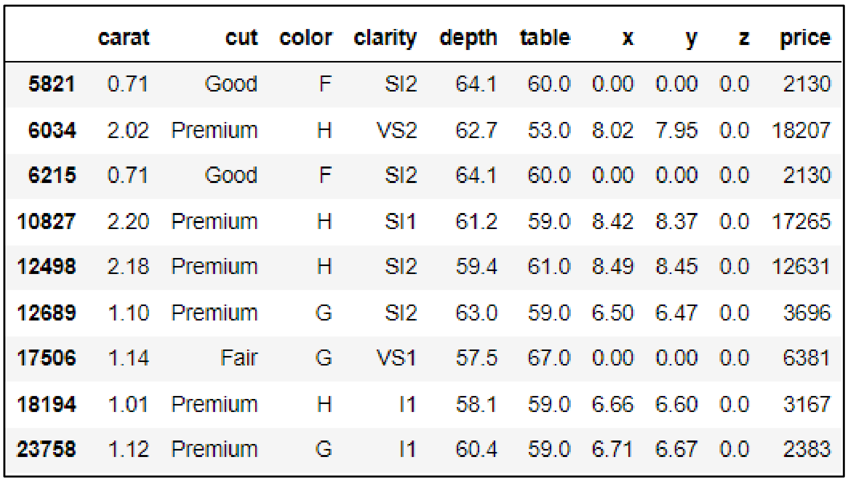
```
carat         0
cut           0
color         0
clarity       0
depth       697
table         0
x             0
y             0
z             0
price         0
```

After imputing null values



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

The dataset is checked for presence of any zero values, and it is found that the 'X, Y, Z' columns are having zero values. These are the length, width, height of the cubic zirconia, without these dimension values it would be impossible to manufacture the product. Therefore, these zero values can be considered as unavailable data and be removed from our dataset.



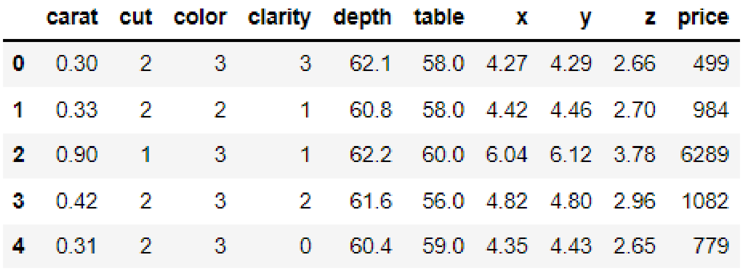
	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Zero values of the dataset

The sublevels of the (cut, clarity, color) categorical columns can be combined to reduce the number of categories in them and make it a well-defined categorical data. The data encoding is done as per the combination of the sublevels.

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.

The categorical columns are encoded with numerical values and saved to the dataset.



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

Encoded Data

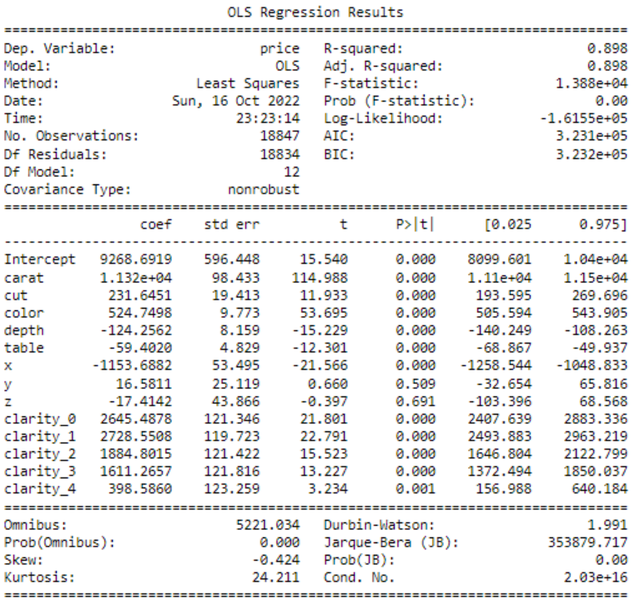
Splitting the data:

The data is split into X and Y variables by assigning all the predictor variables to X and assigning target variable to Y. Using the dependent variable, we split the X and Y data frames into training set and test set (70:30 split).

R square on training data = 0.898	RMSE on Training data=1277.61
R square on testing data = 0.899	RMSE on Testing data= 1284.8

Linear Regression using stats models

RMSE =1277.61



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.898
Model:                            OLS   Adj. R-squared:                  0.898
Method:                 Least Squares   F-statistic:                 1.388e+04
Date:                Sun, 16 Oct 2022   Prob (F-statistic):               0.00
Time:                        23:23:14   Log-Likelihood:            -1.6155e+05
No. Observations:               18847   AIC:                         3.231e+05
Df Residuals:                   18834   BIC:                         3.232e+05
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9268.6919    596.448     15.540      0.000    8099.601    1.04e+04
carat       1.132e+04     98.433    114.988      0.000    1.11e+04    1.15e+04
cut          231.6451     19.413     11.933      0.000     193.595     269.696
color        524.7498      9.773     53.695      0.000     505.594     543.905
depth       -124.2562      8.159    -15.229      0.000    -140.249    -108.263
table        -59.4020      4.829    -12.301      0.000     -68.867     -49.937
x          -1153.6882     53.495    -21.566      0.000   -1258.544   -1048.833
y             16.5811     25.119      0.660      0.509     -32.654      65.816
z            -17.4142     43.866     -0.397      0.691    -103.396      68.568
clarity_0   2645.4878    121.346     21.801      0.000    2407.639    2883.336
clarity_1   2728.5508    119.723     22.791      0.000    2493.883    2963.219
clarity_2   1884.8015    121.422     15.523      0.000    1646.804    2122.799
clarity_3   1611.2657    121.816     13.227      0.000    1372.494    1850.037
clarity_4    398.5860    123.259      3.234      0.001     156.988     640.184
==============================================================================
Omnibus:                     5221.034   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           353879.717
Skew:                          -0.424   Prob(JB):                         0.00
Kurtosis:                      24.211   Cond. No.                     2.03e+16
==============================================================================
```

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

Multi collinearity: It is observed that there is a very strong multi collinearity present in the data set. Dimension reducing techniques can be used to optimize our model further.

1. EXPLORATORY DATA ANALYSIS

- Checked and removed the duplicates in the dataset
- Checked and treated the missing values in the dataset
- Completed the Univariate Analysis and Bi-variate Analysis

2. Split of data into train and test data.

3. Linear Regression model built using ‘sklearn’ and ‘stats models’.

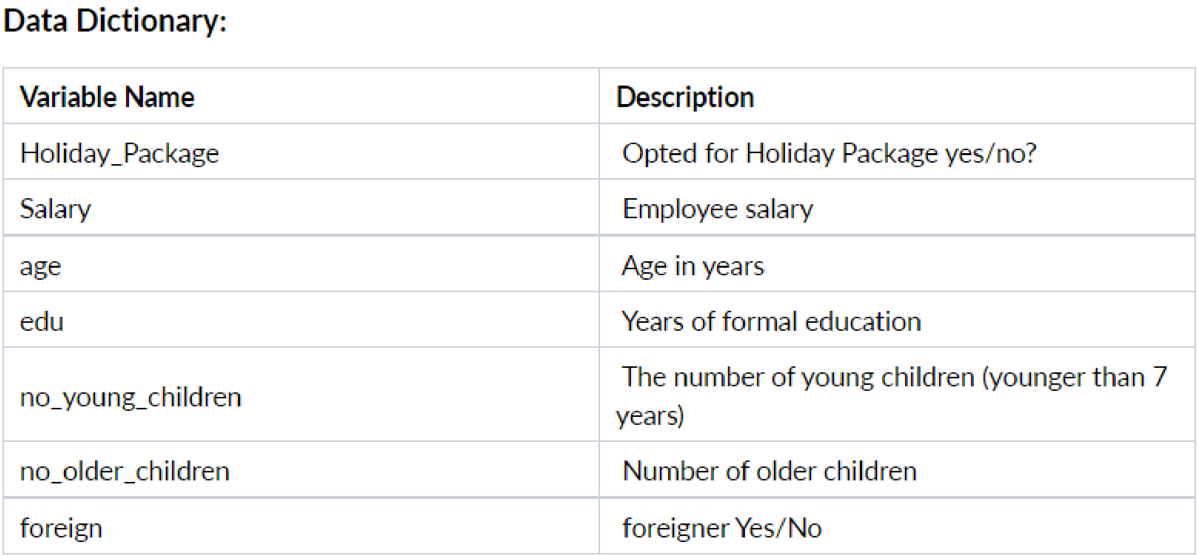
Insights:
The 'Price' is the target variable and all other are predictor variable. The average weight of the cubic zirconia is 0.79 and from the
summary the average price of the 3939.52, maximum price is 18818, minimum price is 326.

Recommendations:

- The company should consider the features like Carat, Cut, colour, clarity, and the dimensions of the stone. These features are highly correlated to price. These features may be helpful to distinguish between higher profitable stones and lower profitable stones.
- Stones with best clarity (SI1, SI2, VS1, VS2) are high profit stones.
- From pair plot, the dimension of the stone is having strong correlation with the price of the stone. The company can use this to their advantage by customising the stone as per customer requirements and charge them based on this data.
- Premium and Ideal cut stones are high profit stones, where as fair cut stones are low profit stones.

Problem 2: Logistic Regression and LDA

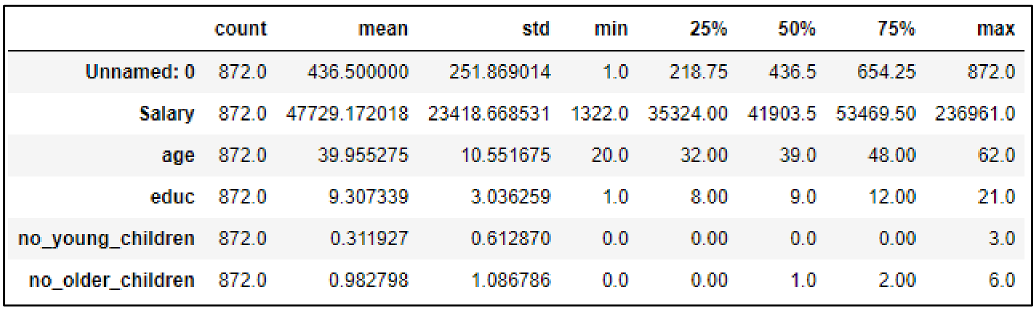
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not 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.



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

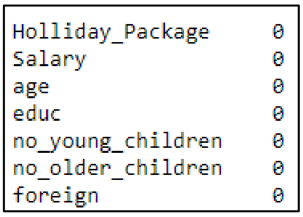
2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.



	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.0	436.500000	251.869014	1.0	218.75	436.5	654.25	872.0
Salary	872.0	47729.172018	23418.668531	1322.0	35324.00	41903.5	53469.50	236961.0
age	872.0	39.955275	10.551675	20.0	32.00	39.0	48.00	62.0
educ	872.0	9.307339	3.036259	1.0	8.00	9.0	12.00	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.00	0.0	0.00	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.00	1.0	2.00	6.0

Data Summary

The dataset is having 872 rows and 8 columns in it. The column 'Unnamed:0'is only the serial number and it is not going to help for model prediction. Therefore, we can drop the 'Unnamed:0' column. The average salary of the employee is 47729 and the max. is236961 in this company. The average age of the employee is 40 and maximum age is 62. There are only two object variables (Holliday Package, foreign), other five variables are of integer data type.

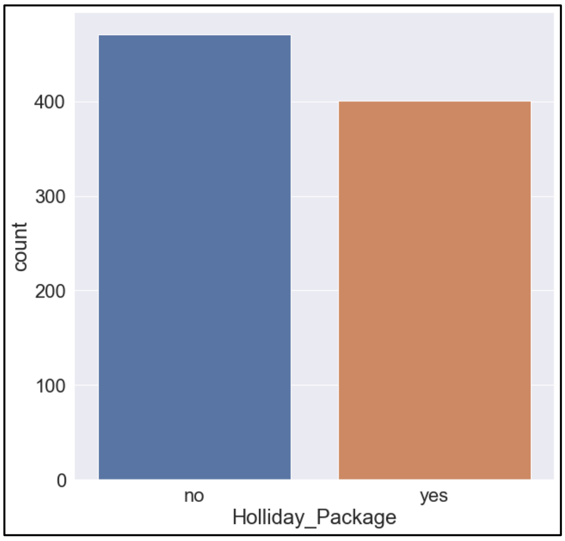


```
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
```

Null value check

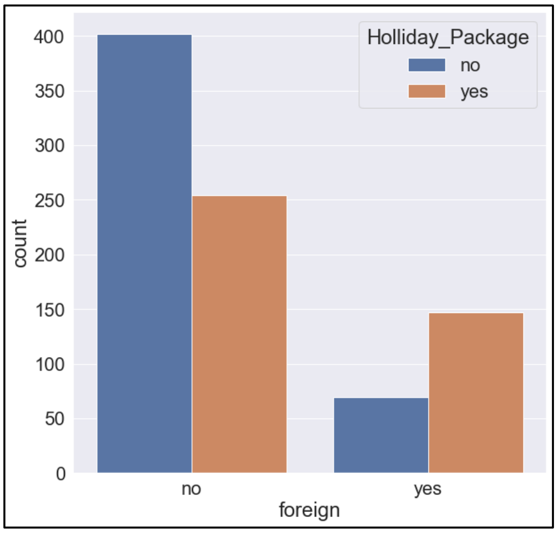
The dataset is checked for null values and there are no null values in the dataset. The dataset also doesn’t have any duplicated rows in it.

Shape of our dataset after removing (Unnamed: 0) is as follows,
No. of rows: 872
No. of columns: 7



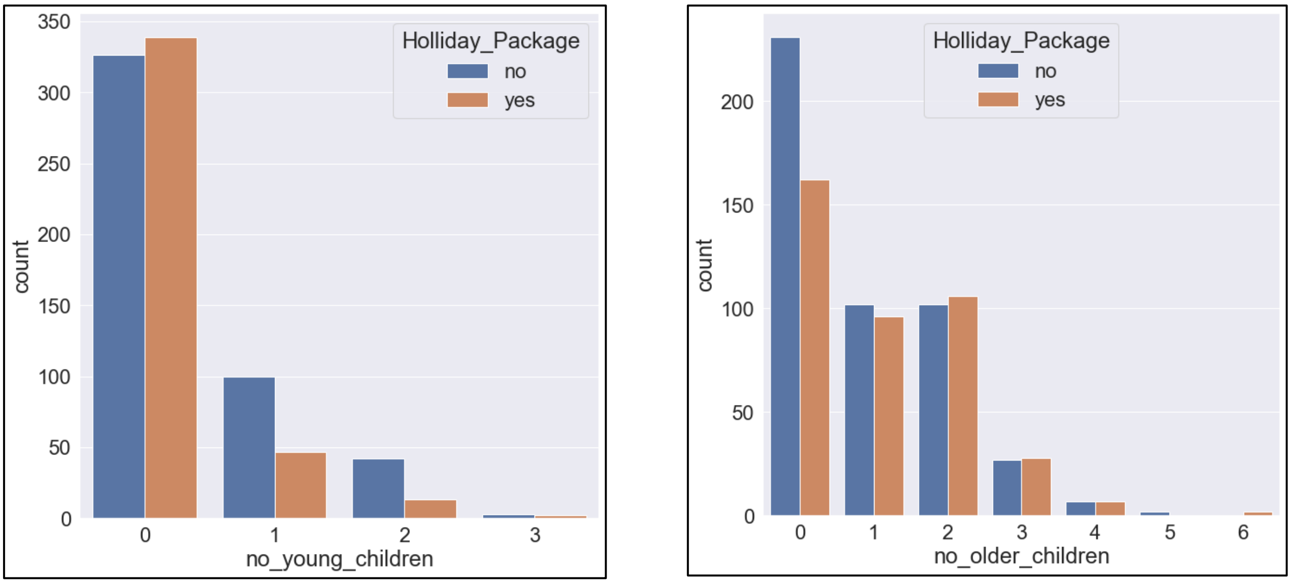
Holiday Package Status

About 471 employees have not taken the package and 401 employees have taken the package



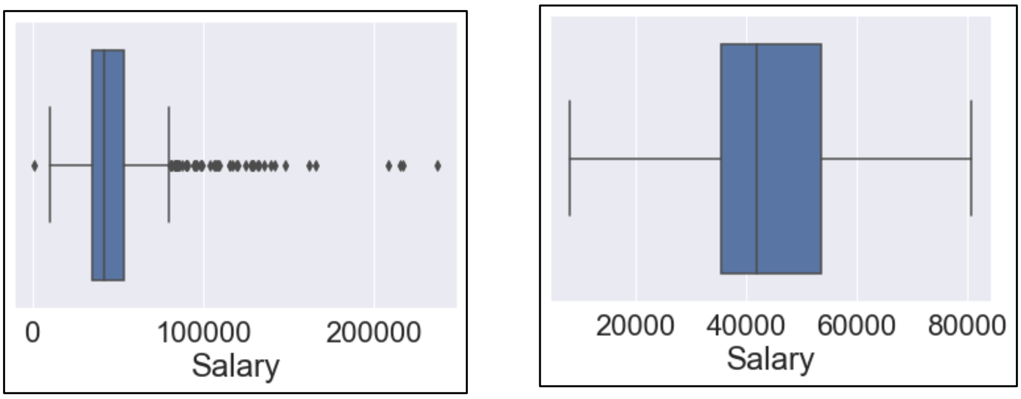
Foreign employee Status

About 150 foreigners have taken the holiday package out of 216 foreign employees.

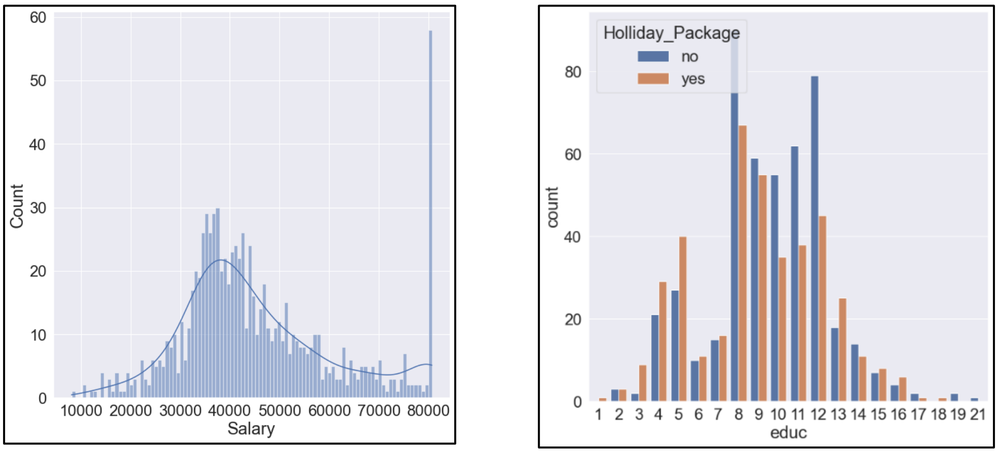


Employee Children status

From the graph the employees with no children are more interested in taking our holiday package in comparison to employees with children. Employees with younger children are not taking our holiday package.

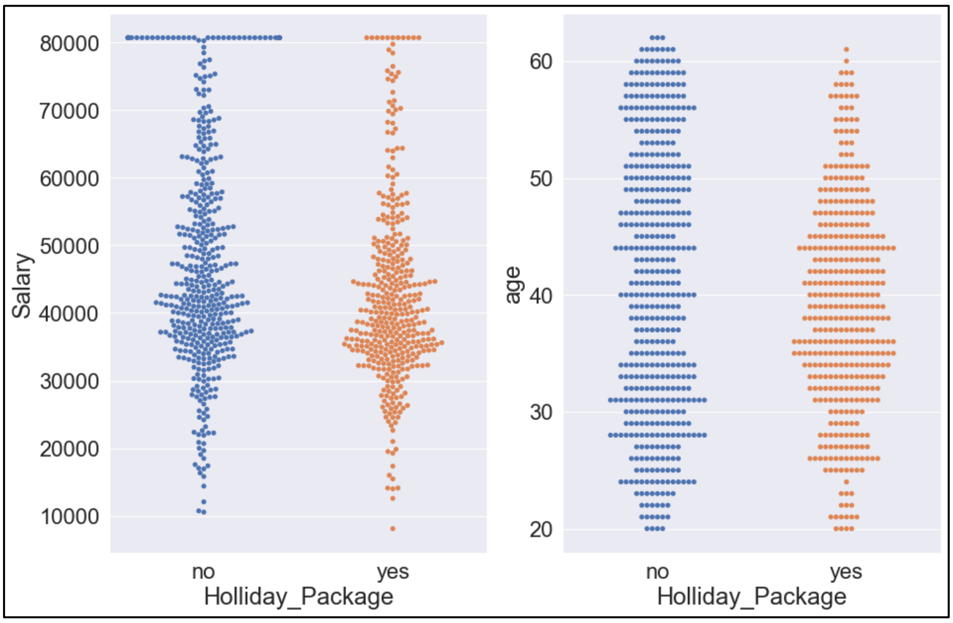


Outlier treatment of Salary data (Before and After boxplots)



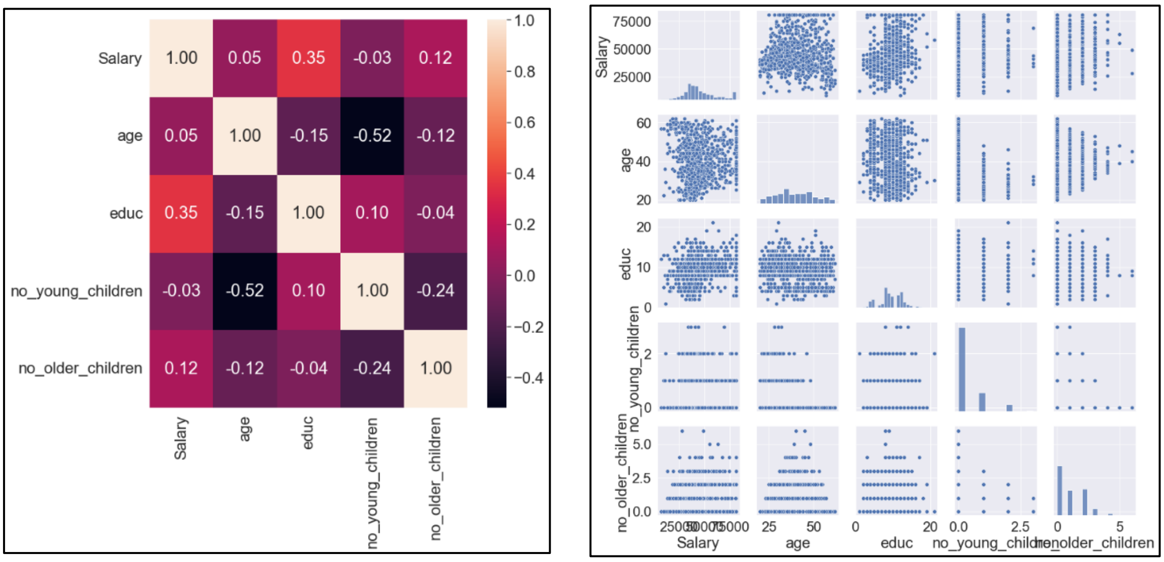
Employee Salary and education

The salary is distributed mostly in the range of 25k to 50k and most employees are having an education between 8 to 12 years, employee age group (30 to50). Employees within this range are more likely to take our holiday package.



Salary and Age of employees

The employees with age of 30 to 45 are also more likely to take our holiday package.

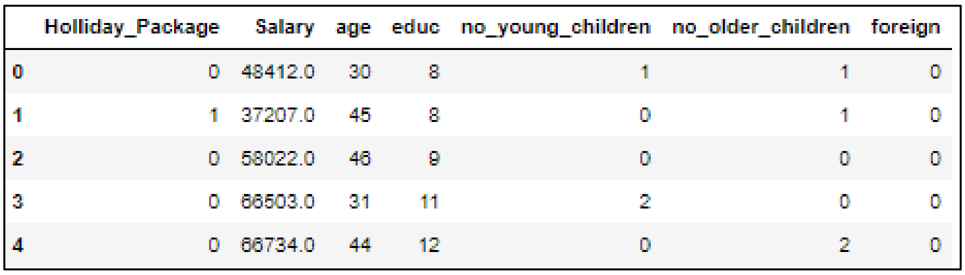


Heatmap and Pair plot of independent variables

There is no correlation between the data values in our dataset. All our independent variables are not having no correlation.

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).

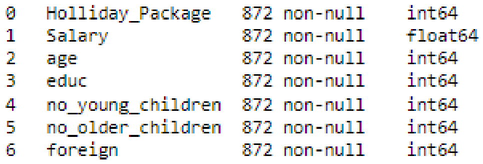
Encoding the data (having string variables)



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
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

Encoded Dataset

The Yes or No categories of the Holiday package and foreign columns are coded as 1 and 0 respectively. The data type for these columns are converted to int type.



```
0   Holliday_Package   872 non-null    int64
1   Salary             872 non-null    float64
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    int64
```

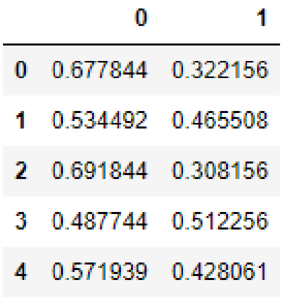
Data type of the dataset

Splitting of dataset.

The data is split into X and Y variables by assigning all the predictor variables to X and assigning target variable to Y. Using the dependent variable, we split the X and Y data frames into training set and test set (70:30 split).

Apply Logistic Regression and LDA

After Splitting the data into train and test sets the data is fit into the logistic regression model and the Predictions are made.



	0	1
0	0.677844	0.322156
1	0.534492	0.465508
2	0.691844	0.308156
3	0.487744	0.512256
4	0.571939	0.428061

The probabilities on the test set

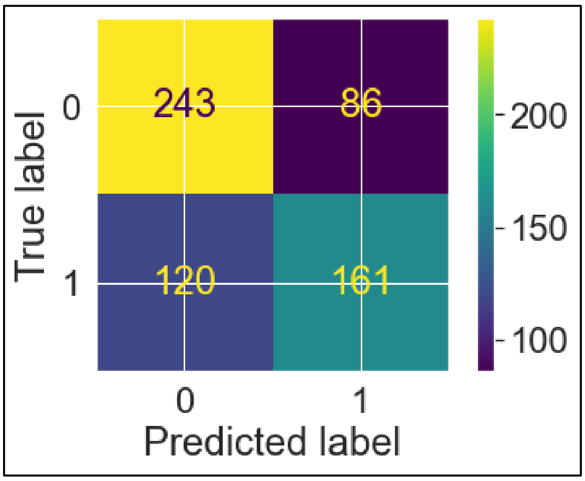
LDA (linear discriminant analysis)

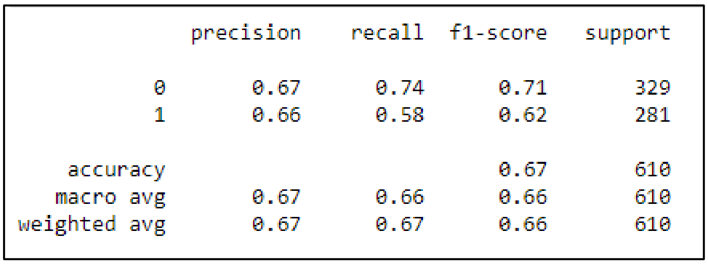
The same Split data is fit into the linear discriminant analysis model and the Predictions are made.

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 LOGISTIC REGRESSION

Confusion matrix and classification report on the training data

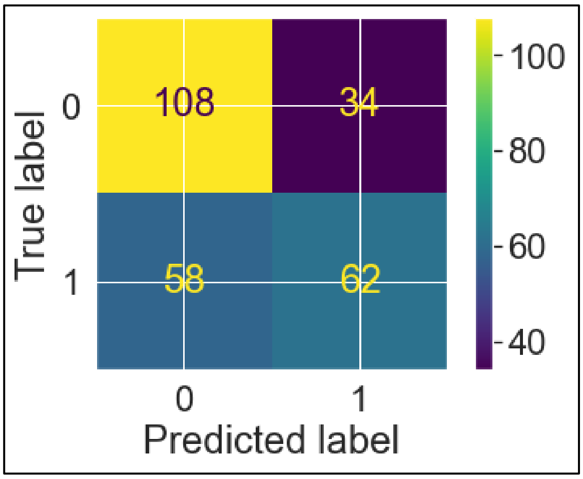


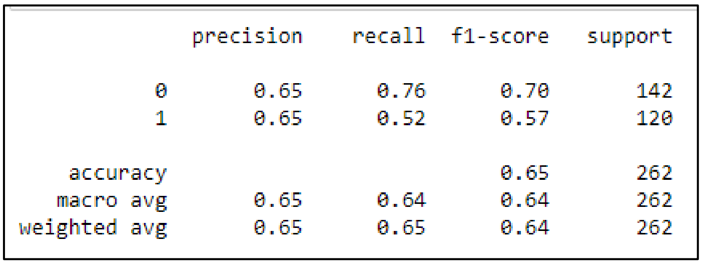


```
              precision    recall  f1-score   support

           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
```

Here we see that precision for 1 is 0.66, recall is 0.58 accuracy is 0.67 and f1 score is 0.62.

Confusion matrix and classification report on the test data





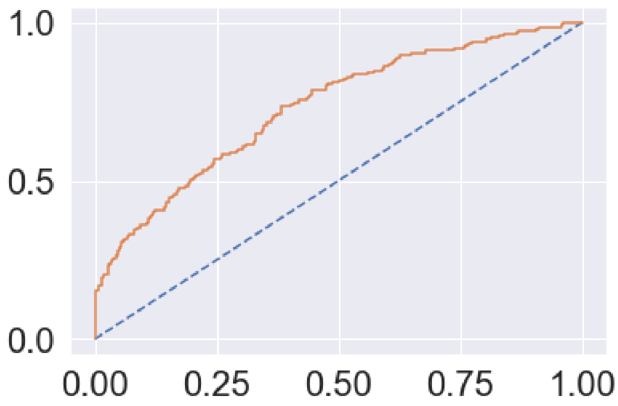
```
              precision    recall  f1-score   support

           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
```

Here we see that precision for 1 is 0.65, recall is 0.52 accuracy is 0.65 and f1 score is 0.57.

Accuracy of Training data- 0.66

AUC and ROC for the training data

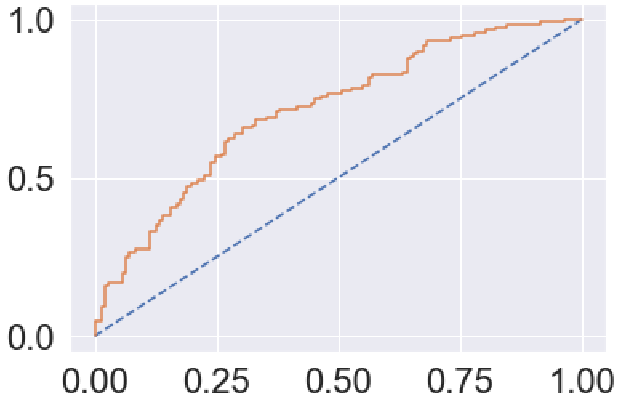
```
AUC: 0.731
```



Accuracy of Test data- 0.65

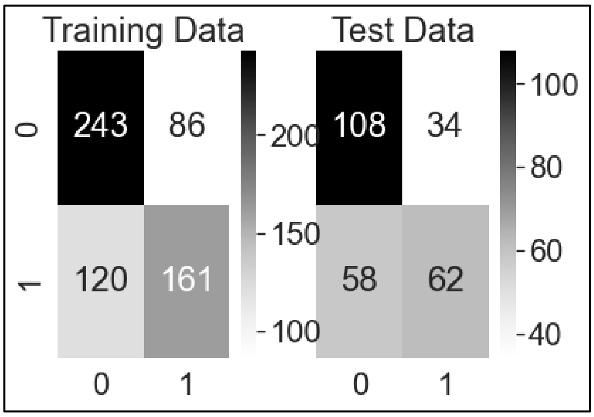
AUC and ROC for the test data

```
AUC: 0.731
```

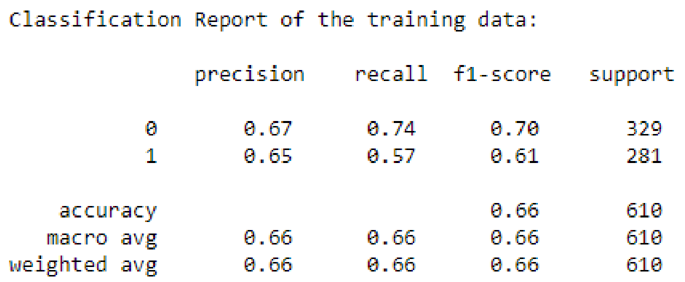


PERFORMANCE METRICS FOR LDA (linear discriminant analysis)

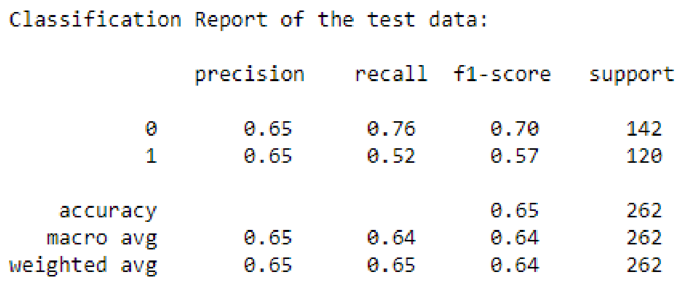
Confusion matrix (Training and Test data)



Classification report on the train and test data



```
Classification Report of the training data:

              precision    recall  f1-score   support

           0       0.67      0.74      0.70       329
           1       0.65      0.57      0.61       281

    accuracy                           0.66       610
   macro avg       0.66      0.66      0.66       610
weighted avg       0.66      0.66      0.66       610
```

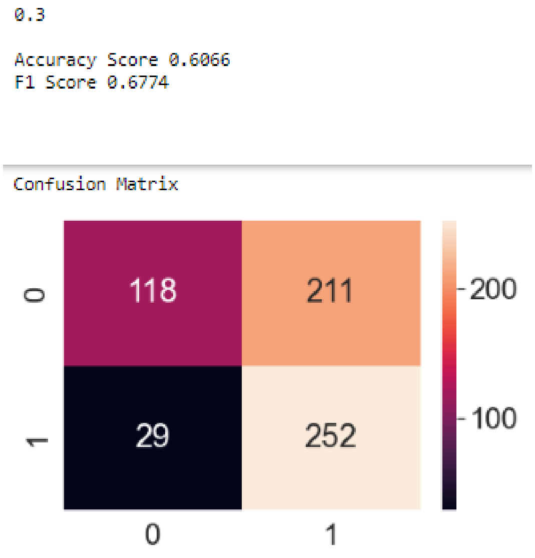


```
Classification Report of the test data:

              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
```

For train data, the precision for 1 is 0.65, recall is 0.57 accuracy is 0.66 and f1 score is 0.61.

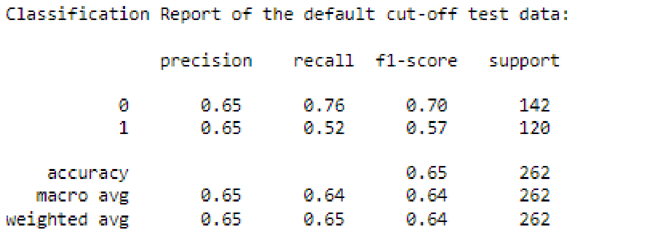
For test data, the precision for 1 is 0.65, recall is 0.52 accuracy is 0.65 and f1 score is 0.57.

CHANGING THE CUT-OFF VALUE TO CHECK OPTIMAL VALUE THAT GIVES BETTER ACCURACY AND F1 SCORE.

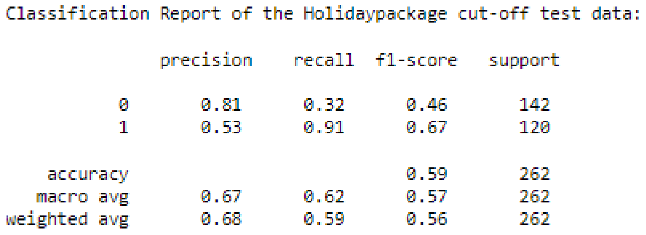


```
0.3

Accuracy Score 0.6066
F1 Score 0.6774
```

```
Confusion Matrix
```



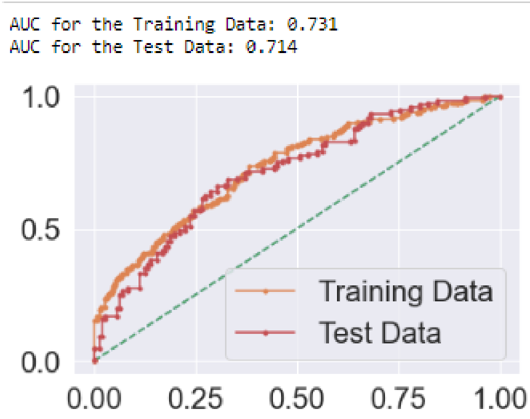
```
Classification Report of the default cut-off test data:

              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
```



```
Classification Report of the Holidaypackage cut-off test data:

              precision    recall  f1-score   support

           0       0.81      0.32      0.46       142
           1       0.53      0.91      0.67       120

    accuracy                           0.59       262
   macro avg       0.67      0.62      0.57       262
weighted avg       0.68      0.59      0.56       262
```

CONFUSION MATRIX AND CLASSIFICATION REPORT OF OPTIMAL CUTOFF

AUC and ROC for the train and test data



```
AUC for the Training Data: 0.731
AUC for the Test Data: 0.714
```

On Comparing both the LDA and logistic regression models, we find both results are same, but LDA works better. The AUC/ROC for the models are similar to each other. Both the models are working well. But, as the LDA works better with categorical values we can choose it for our prediction purpose.

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

We analysed the data using logistic regression and LDA to predict whether an employee will take the holiday package or not. For the given dataset there is no significant difference in the performance of both the models.

First, we loaded the data and performed univariate and bivariate analysis on it to understand about its distribution and its influence on our target variable. The dataset is then split into train and test dataset (70:30 split) to fit it in our prediction model. The predictions are made and the optimal model is selected for this application.

Data Insights:

- Employees who are in the age range between 30 to 50 are more interested in the holiday packages.
- Employees with salary less than 50k opt for holiday packages.
- Education also plays an important role in deciding the holiday packages.

Recommendations:

- Since older employees are not taking the holiday package, we can customize tours to relaxing locations and provide a tour guide or a personal assistant during the trip.
- Employees earning more than 50k may not have a suitable plan in our package. We can provide them with a holiday package in a cruise ship or packages with unique experiences like spa, personal space for dinner. Since they earn more, they may be ready to spend on a more attractive and rich experience.
- Employees with younger children can be given a package with amusement parks, movie studios.