PREDICTIVE-MODELING BUSINESS REPORT

PGP-DSBA

Srikanthpr.27@gmail.com

12/6/20

Predictive-Modeling Project

Contents

1 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.

Dataset: 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 best and J the worst.
Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
Depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	The Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

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

Solution:

- Loaded the required packages and read the dataset.
- Remove the Unwanted column from the original data has it is a serial number.
- Dataset has 26,967 observations and 10 Features including the target variable.

	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
5	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706

Figure1: Head of the data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object
 2   color    26967 non-null  object
 3   clarity  26967 non-null  object
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

Figure2: Structure of the data

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.00000	26967	26967	26967	26270.00000	26967.00000	26967.00000	26967.00000	26967.00000	26967.00000
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.79838	NaN	NaN	NaN	61.74515	57.45608	5.72985	5.73357	3.53806	3939.51811
std	0.47775	NaN	NaN	NaN	1.41286	2.23207	1.12852	1.16606	0.72062	4024.86467
min	0.20000	NaN	NaN	NaN	50.80000	49.00000	0.00000	0.00000	0.00000	326.00000
25%	0.40000	NaN	NaN	NaN	61.00000	56.00000	4.71000	4.71000	2.90000	945.00000
50%	0.70000	NaN	NaN	NaN	61.80000	57.00000	5.69000	5.71000	3.52000	2375.00000
75%	1.05000	NaN	NaN	NaN	62.50000	59.00000	6.55000	6.54000	4.04000	5360.00000
max	4.50000	NaN	NaN	NaN	73.60000	79.00000	10.23000	58.90000	31.80000	18818.00000

Figure3: Summary Statistics of the data

Key takeaways from the descriptive statistics:

- Out of 9 independent variables, we have 7 continuous numerical variables and 3 categorical variables.
- 'Depth' variable contains few missing values.
- The dataset contains 7 float, 3 objects and one int data-type.
- Dataset pertain to 5 unique cut quality, wherein Ideal quality being the most occurring, there are 7 unique colours and 8 different clarity.
- Price has a high dispersion.
- Mean and 50th percentile is almost equal, seems the data is normally distributed.

```
carat     1.116481
depth    -0.028618
table     0.765758
x         0.387986
y         3.850189
z         2.568257
price     1.618550
dtype: float64
```

Figure4: Skewness of the data

- The variables carat, y, z and price is greater than 1, the data are highly skewed.
- The variable depth and x is between -0.5 and 0.5, the data are fairly symmetrical.
- The variable table is between 0.5 and 1, the data is moderately skewed.

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

Figure5: Missing Values

- Depth has 697 (2.6%) missing values.

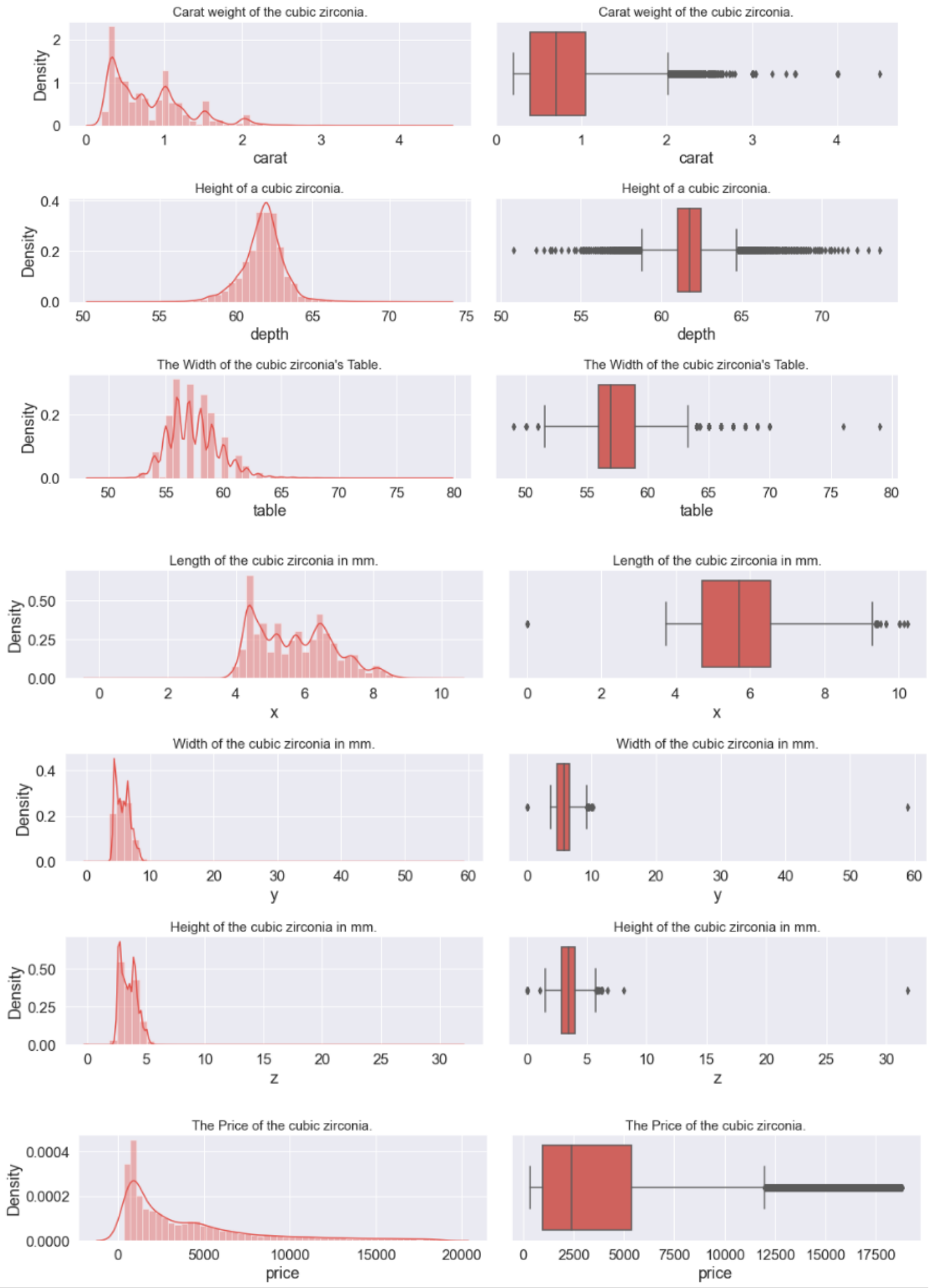


Figure6: Univariate Analysis on continuous variables

- Carat, depth and price column has huge number of outliers.
- Carat, table, x, y and z seems to have multiple modal values.

```
CUT :  5
Fair          781
Good         2441
Very Good    6030
Premium      6899
Ideal       10816
Name: cut, dtype: int64
```

```
COLOR :  7
J    1443
I    2771
D    3344
H    4102
F    4729
E    4917
G    5661
Name: color, dtype: int64
```

```
CLARITY :  8
I1        365
IF        894
VVS1     1839
VVS2     2531
VS1      4093
SI2      4575
VS2      6099
SI1      6571
Name: clarity, dtype: int64
```

Figure7: Unique Count of categorical variables

- Cut is of ordinal data-type with 5 category: Fair, Good, Very Good, Premium, and Ideal.
- Color is of ordinal data-type with 7 category: D, E, F, G, H, I, J with D being the best and j being the worst.
- Clarity is of ordinal data-type with 8 category: FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions).

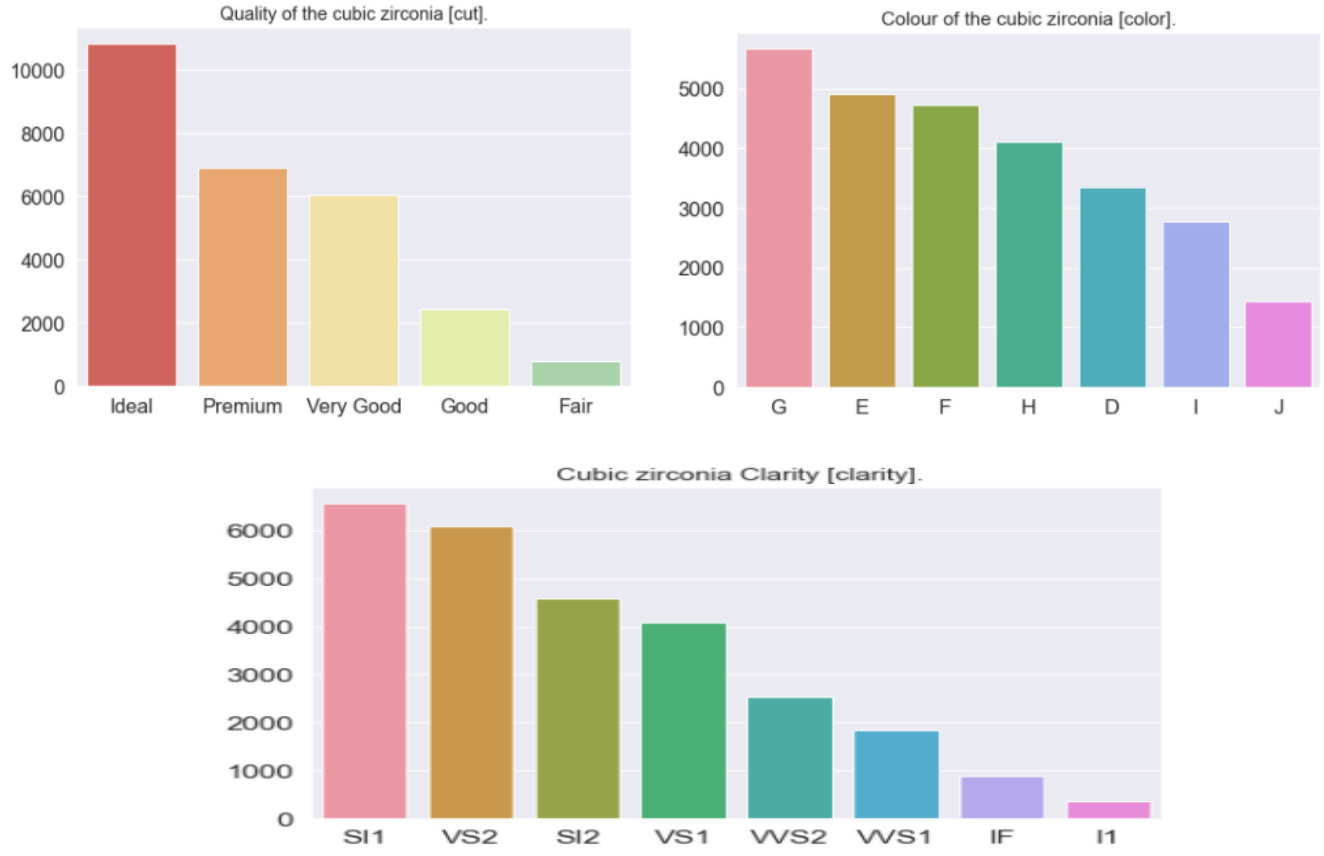


Figure8: Count plot of categorical variables

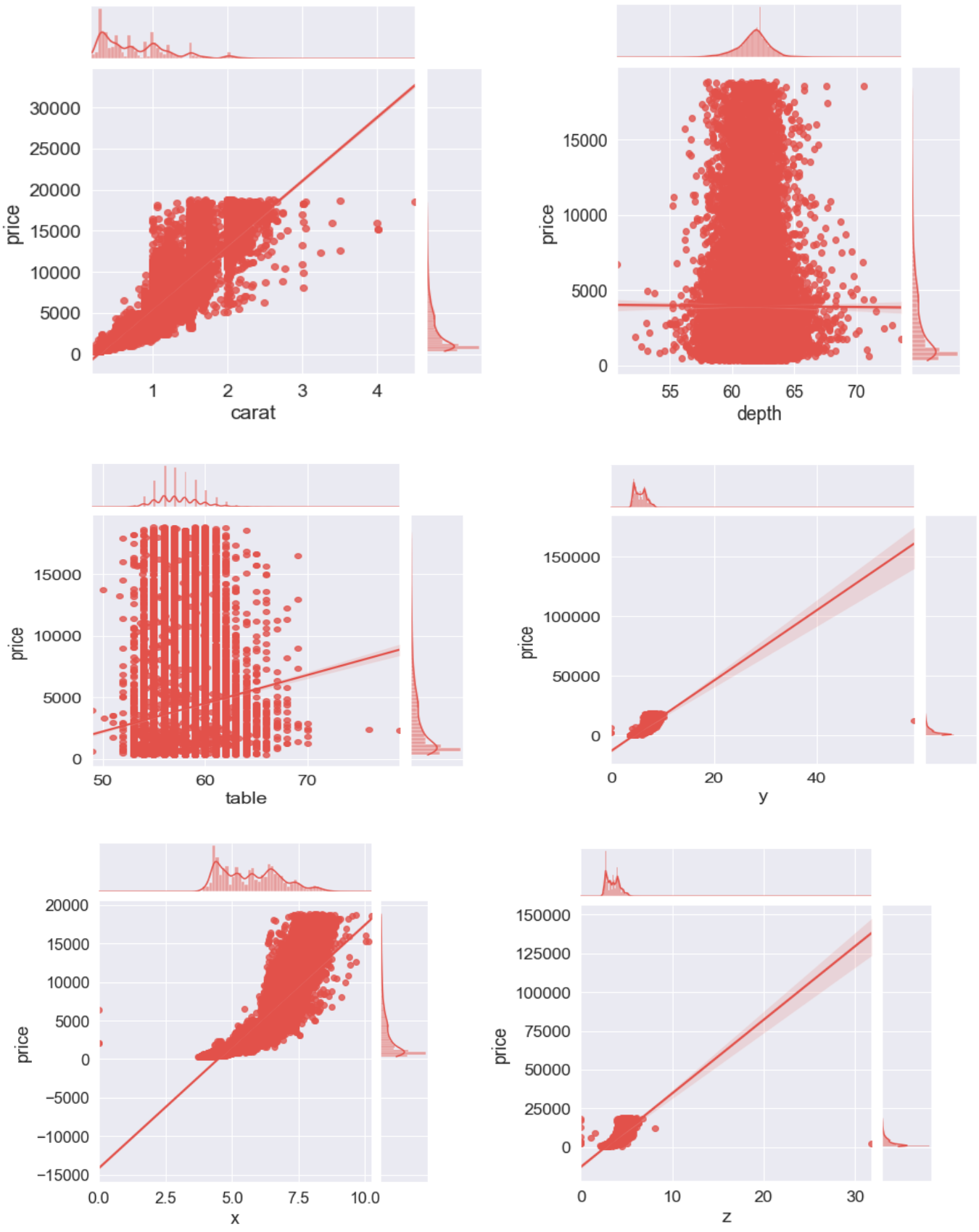


Figure9: Multivariate Analysis

From the above figure we can say that:

- 'Depth' has a very weak correlation to target variable price, which means weak predictor for a linear regression model.

- 'Carat', 'x', 'y' and 'z' has a very high correlation.
- 'Table' has a moderate linear relationship with the target variable 'price'.
- Variables 'x', 'y' and 'z' have an extreme outliers on both the lower and upper tail. Need to deal with outliers before building a model.

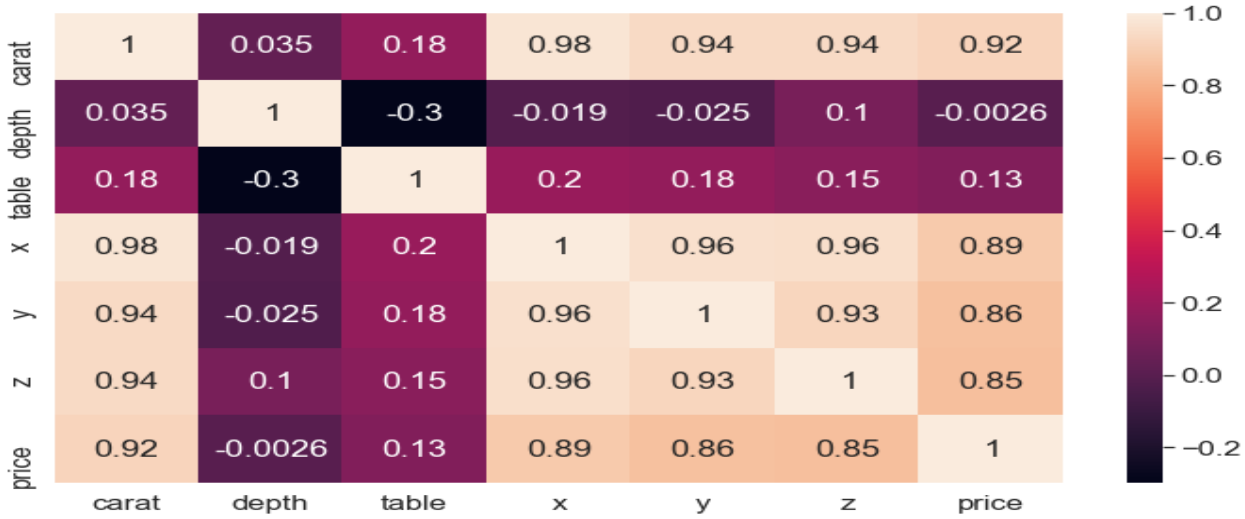


Figure-10: Heatmap to identify correlation between variables

- Carat is highly correlated with x, y, z and price. Multi-collinearity exist between independent variables.
- Depth and table has no correlation with the target variable price, also there is no linear relationship between the independent variables; hence multi-collinearity do not exist.
- 'x' is highly correlated with 'z', 'y' and 'price'. Multi-collinearity exist between independent variables.
- 'y' is highly correlated with 'z' and price.
- 'z' is highly correlated with 'price'.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

Solution:

	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

Figure-11: Display Values equal to zero

- x, y and z represent length, width and height of a cubic zirconia.
- So, for a shape of a diamond value of x, y and z having zero is meaningless. Hence convert all the zeros to missing values.

```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            3
y            3
z            9
price        0
dtype: int64
```

Figure-12: Missing values

- Since the missing values are less than 3% we can remove them from the original dataset.

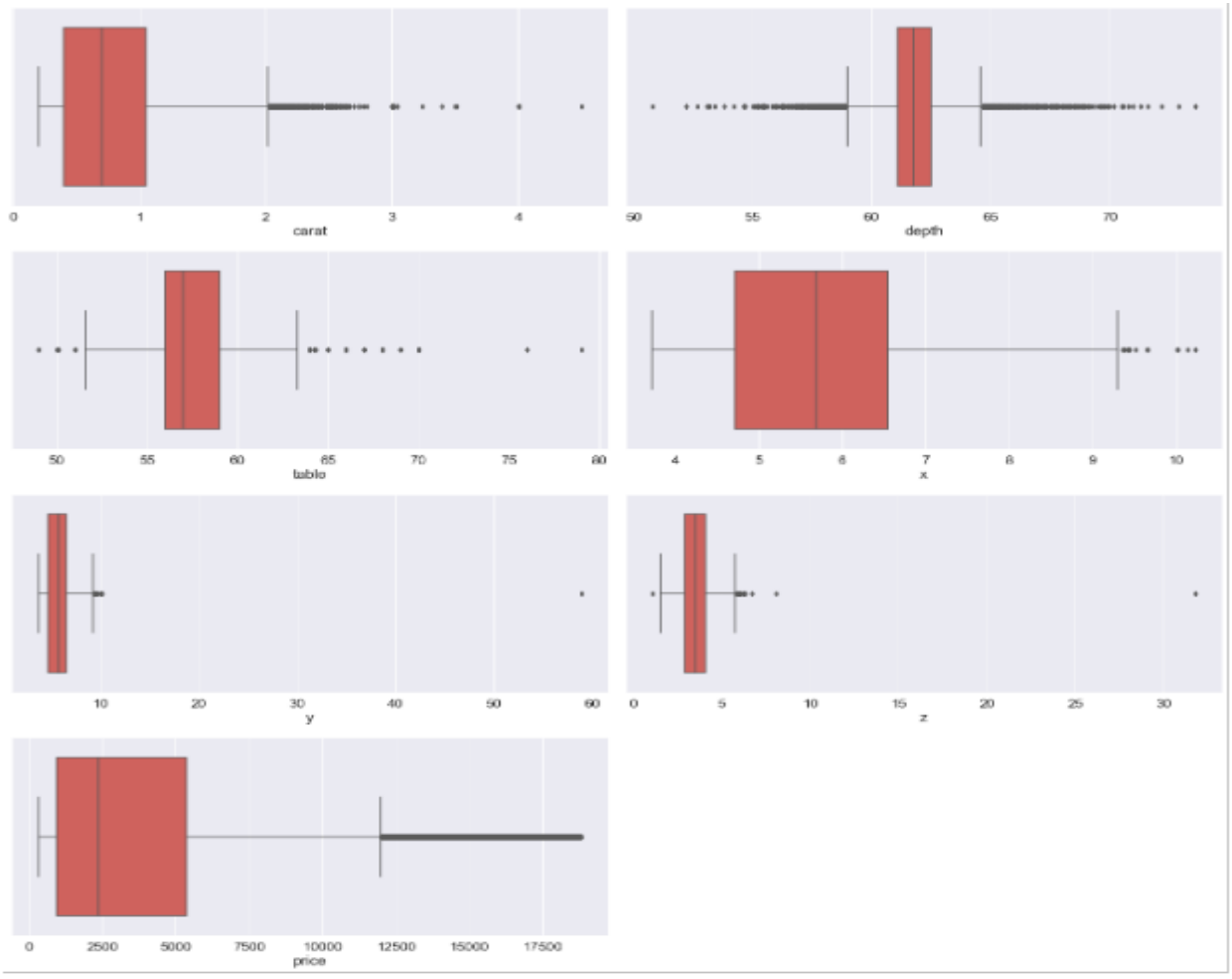


Figure-13a: Before Outliers Treatment

- Impute the outlier's value to upper limit and lower limit using ceiling and flooring technique.

Figure-13b: After Outliers Treatment

- In regression, it is better to centre the variables so that the predictors have mean 0. This makes it easier to interpret the intercept term as the expected value of Yi when the predictor values are set to their means. Otherwise, the intercept is interpreted as the expected value of Yi when the predictors are set to 0, which may not be a realistic or interpretable situation.
- Also, we have feature with different unit weight of cubic measured in carat whereas length or width measured in mm.

1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into test and train (70:30). Apply linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using Rsquare, RMSE?

Solution:

Encode the Data:

- Since there are 3 categorical variable or columns with the type as object, these variable were converted into categorical type.
- Initially categorical variable 'cut' were converted in to dummy variables using One hot encoding and the remaining two 'color' and 'clarity' category were converted using pd.categories to avoid higher dimensionality.
- After encoding the data, the target variable 'price' was captured in to separate vector for training and test data set.
- Then, the data was split into train and test in the ratio of 70:30.

```
Number of rows and columns of the training set for the independent variables: (18382, 13)
Number of rows and columns of the training set for the dependent variable: (18382, 1)
Number of rows and columns of the test set for the independent variables: (7879, 13)
Number of rows and columns of the test set for the dependent variable: (7879, 1)
```

Figure-14: Data-Split [70:30]

Apply Linear Regression:

- Initially linear regression model was built without scaling the data and without any specific parameter setting.
- The coefficients for each of the independent attributes were captured.

```
The coefficient for carat is 9305.030130375635
The coefficient for color is -226.90056110618627
The coefficient for clarity is 239.11467724683274
The coefficient for depth is 3.3424603361755265
The coefficient for table is -32.96115023521669
The coefficient for x is -1677.8771671728953
The coefficient for y is 1525.2860052518943
The coefficient for z is -917.0197363313135
The coefficient for cut_Fair is -562.3575327065515
The coefficient for cut_Good is -48.53211478120936
The coefficient for cut_Ideal is 316.2707706870597
The coefficient for cut_Premium is 182.95218574614586
The coefficient for cut_Very_Good is 111.66669105454564
```

Figure-15: Coefficients of Independent variable

- The intercept of our model is 1653.43.
- For training set, R^2 or coefficient of determinant or the model score is 91.15%
- The RMSE for training set is 1026.
- For testing set , R^2 or coefficient of determinant or the model score is 90.98%
- The RMSE for training set is 1052.

- The predicted y value vs actual y values for the test data were plotted.

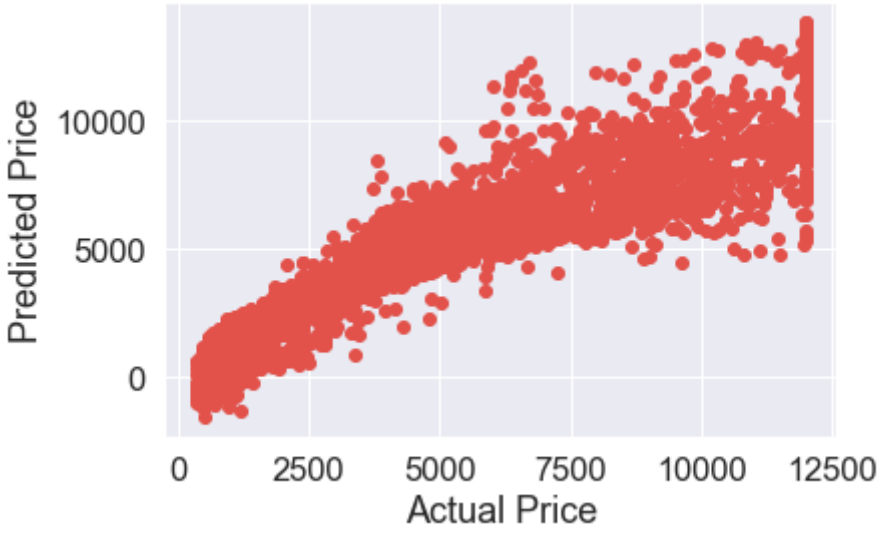


Figure-16: Actual price vs Predicted Price

- When all the predictors are 0, then the price of a cubic zirconia becomes equal to intercept value, which becomes meaningless.
- Once again linear regression is applied for the scaled data just to interpret the intercept value. However scaling do not affect the model score of a linear regression.
- The coefficients for each of the independent attributes were captured for a scaled data.

```
The coefficient for carat is 1.241155382417389
The coefficient for color is -0.11173293252334582
The coefficient for clarity is 0.11967342620149482
The coefficient for depth is 0.00121966495754702
The coefficient for table is -0.020689067653729715
The coefficient for x is -0.5462796701168283
The coefficient for y is 0.4930959806987278
The coefficient for z is -0.1843302445999642
The coefficient for cut_Fair is -0.03404693578417074
The coefficient for cut_Good is -0.01701516951930478
The coefficient for cut_Ideal is 0.022826480193777494
The coefficient for cut_Premium is 0.003445699948725065
The coefficient for cut_Very_Good is -0.005327947228240204
```

Figure-17: Coefficients of Scaled Independent variable

- The intercept of our model is almost 0.
- For training set, R^2 or coefficient of determinant or the model score is 91.15%
- The RMSE for training set is 0.298
- For testing set , R^2 or coefficient of determinant or the model score is 90.98%
- The RMSE for training set is 0.3
- The predicted y value vs actual y values for the test data were plotted.

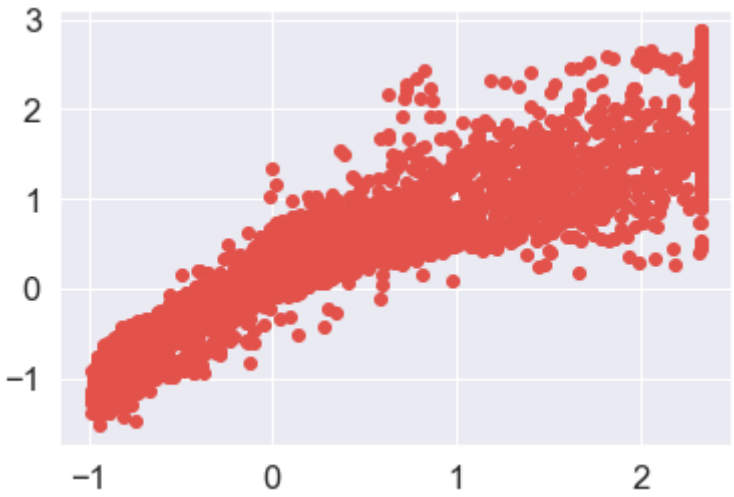


Figure-18: Actual price vs Predicted Price

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	1026.008047	1052.632064	0.911573	0.909852
Scaled Linear Regression	0.297367	0.300283	0.911573	0.909830

Figure-19: Final Output

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

Solution:

- The final Linear Regression equation is:

Price = (1653.42) * Intercept + (9305.04) * carat + (-226.90) * color + (239.12) * clarity + (3.35) * depth + (-32.97) * table + (-1677.88) * x + (1525.28) * y + (-917.02) * z + (-562.36) * cut_Fair + (-48.54) * cut_Good + (316.28) * cut_Ideal + (182.95) * cut_Premium + (111.67) * cut_Very_Good .

- When carat increases by 1 unit, price increases by 9305.04 units, keeping all other predictors constant.
- When cut quality is:
  1. fair, price decreases by -562.36
  2. Good, price decreases by -48.54
  3. Very_Good price increases by 111.67
  4. premium, price increases by 182.95
  5. Ideal, price increases by 316.28

So, as the quality of the cubic zirconia is varied from fair to Ideal quality, price also increases.

- When 'y' height of the cubic increases by 1 unit, price increases by 1525.28 units, keeping all other predictors constant.
- The best attributes that are most important were identified by using variance inflation factor technique.

```
carat ---> 3.6773673273394603
color ---> 3.37744981928417
clarity ---> 2.8684793717645363
cut_Fair ---> 1.0809553040342552
cut_Good ---> 1.18964360503517
cut_Premium ---> 1.578520020707355
cut_Very_Good ---> 1.4759636465422865
```

Figure-20: Best Attributes

- Inferential Statistics for Best attributes:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.909
Method:                 Least Squares   F-statistic:                 2.620e+04
Date:                Sun, 06 Dec 2020   Prob (F-statistic):               0.00
Time:                        17:28:11   Log-Likelihood:            -1.5380e+05
No. Observations:               18382   AIC:                         3.076e+05
Df Residuals:                   18374   BIC:                         3.077e+05
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2438.8512     28.021    -87.035      0.000   -2493.776   -2383.926
carat       7552.6578     18.133    416.524      0.000    7517.116    7588.199
color       -221.3600      4.732    -46.775      0.000    -230.636    -212.084
clarity      250.0318      4.584     54.548      0.000     241.047     259.016
cut_Fair   -1142.0301     48.335    -23.627      0.000   -1236.772   -1047.288
cut_Good    -466.0762     28.410    -16.405      0.000    -521.763    -410.390
cut_Very_Good  -240.4746     20.308    -11.841      0.000    -280.280    -200.669
cut_Premium    -271.6863     19.768    -13.744      0.000    -310.433    -232.939
==============================================================================
Omnibus:                     4496.937   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20188.186
Skew:                           1.129   Prob(JB):                         0.00
Kurtosis:                       7.610   Cond. No.                         32.6
==============================================================================
```

Figure-21: Inferential Statistics for best attributes

- The Linear Regression equation for best attributes is:

Price = (-2438.85) * Intercept + (7552.65) * carat + (-221.36) * color + (250.04) * clarity + (-1142) * cut_Fair + (-466.08) * cut_Good + (-271.68) * cut_Premium + (-240.48) * cut_Very_Good

- When carat increases by 1 unit, price increases by 7552.65 units, keeping all other predictors constant.
- When cut quality of the cubic zirconia which is having negative coefficients, is varied from fair to Ideal quality, price also increases slightly.
- For any category of Clarity of the cubic, price increases by 250.04 units, keeping all other predictors constant.

Recommendations and Business Insights:

- Overall, from the model we can recommend company to drop variables such as depth and table has they are weak predictors and has no impact on the predictions.
- We also noticed that, weight and length of the cubic zirconia has a very strong co-relation, which in turns lead to multicollinearity problem. Hence any one of them will not be a useful for predictions.

- Carat-weight of the cubic zirconia becomes one the most important feature for getting profitable sales of cubic zirconia also any clarity of the cubic zirconia gives profitable returns to the company.
- Although, the different variety of cut quality varied from fair to Ideal reduces loss, company still need to look into this particular variable.

2 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: 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.

Solution:

- Loaded the required packages and read the dataset.
- Remove the Unwanted column from the original data has it is a serial number.
- Dataset has 872 observations and 7 Features including the target variable.

	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
5	yes	61590	42	12	0	1	no
6	no	94344	51	8	0	0	no
7	yes	35987	32	8	0	2	no
8	no	41140	39	12	0	0	no
9	no	35826	43	11	0	2	no

Figure1: Head of the data

```
<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
```

Figure2: Structure of the data

- From the structure of the data, we could see there are no missing values.
- Variables 'educ', 'no_young_children' and 'no_older_children' is not a continuous, can be converted to categorical datatype.
- Variable Salary and age is of continuous variable.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.00000	872.00000	872.0	872.0	872.0	872
unique	2	NaN	NaN	20.0	4.0	7.0	2
top	no	NaN	NaN	8.0	0.0	0.0	no
freq	471	NaN	NaN	157.0	665.0	393.0	656
mean	NaN	47729.17202	39.95528	NaN	NaN	NaN	NaN
std	NaN	23418.66853	10.55167	NaN	NaN	NaN	NaN
min	NaN	1322.00000	20.00000	NaN	NaN	NaN	NaN
25%	NaN	35324.00000	32.00000	NaN	NaN	NaN	NaN
50%	NaN	41903.50000	39.00000	NaN	NaN	NaN	NaN
75%	NaN	53469.50000	48.00000	NaN	NaN	NaN	NaN
max	NaN	236961.00000	62.00000	NaN	NaN	NaN	NaN

Figure3: Summary Statistics of the data

Key takeaways from the descriptive statistics:

→ There 656 Non-Foreign employees out of 872 employees.
→ There are 665 employees and 393 employees with zero young children and zero older children respectively.
→ 471 employees did not opt for holiday package.
→ Salary variable has a very high dispersion.
→ Minimum age of an employee is 20 and maximum age of an employee is 62.
→ There are no missing values found in the data.

```
Salary    3.103216
age       0.146412
dtype: float64
```

Figure4: Skewness of the data

→ Salary variable is highly skewed.

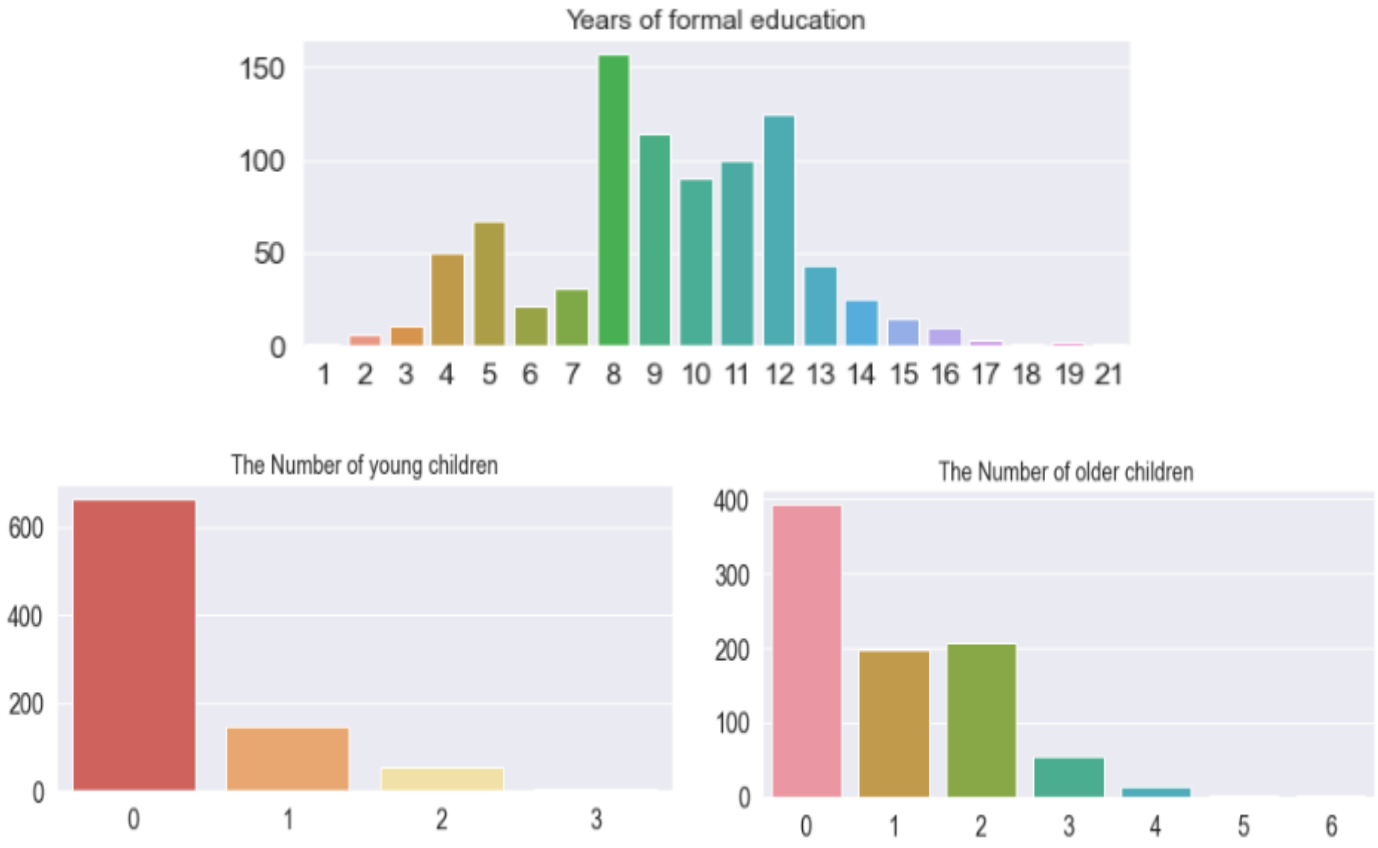
Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
no	471	471	471	471	471	471
yes	401	401	401	401	401	401

Figure5: Distribution of the Target Variable.

→ Most of the employee did not opt for holiday package. The ratio is almost 1:1 is in favour of both yes and no.
→ The model's ability to predict class no will be almost similar to predicting yes.

```
no     0.540138
yes    0.459862
Name: Holliday_Package, dtype: float64
```

Figure6: Data Balance check of a Target Variable.



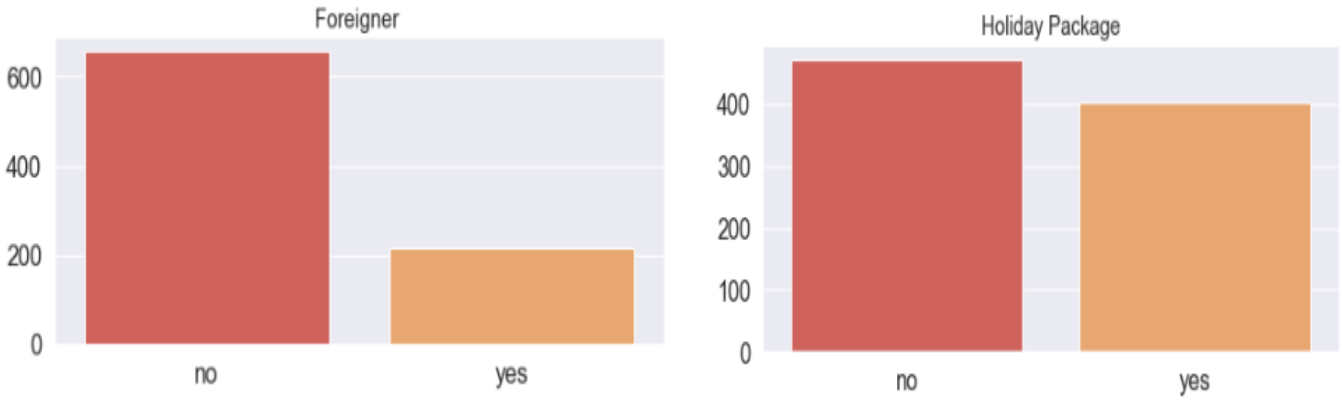


Figure7: Count-plot of categorical variables

```
EDUC :  20
1        1
21       1
18       1
19       2
17       3
2        6
16      10
3       11
15      15
6       21
14      25
7       31
13      43
4       50
5       67
10      90
11     100
9      114
12     124
8      157
Name: educ, dtype: int64
```

```
NO_YOUNG_CHILDREN :  4
3      5
2     55
1    147
0    665
Name: no_young_children, dtype: int64
```

```
NO_OLDER_CHILDREN :  7
6      2
5      2
4     14
3     55
1    198
2    208
0    393
Name: no_older_children, dtype: int64
```

```
FOREIGN :  2
yes    216
no     656
Name: foreign, dtype: int64
```

```
HOLLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holliday_Package, dtype: int64
```

Figure8: Unique Count of categorical variables

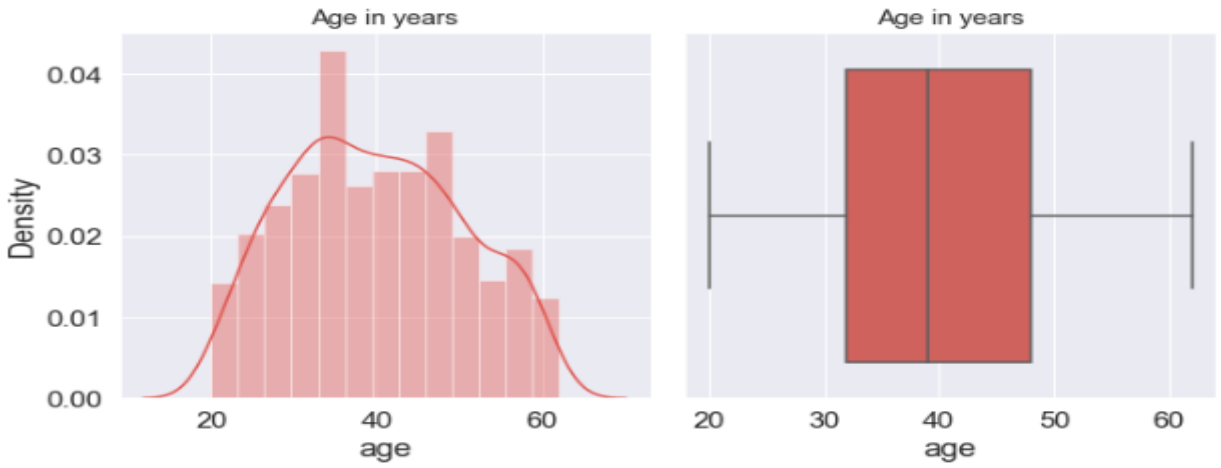


Figure 9: Univariate Analysis on continuous variables

→ Salary variable has a lot of outliers on the upper tail of the data.

Figure 10: Salary Variable after Outlier treatment

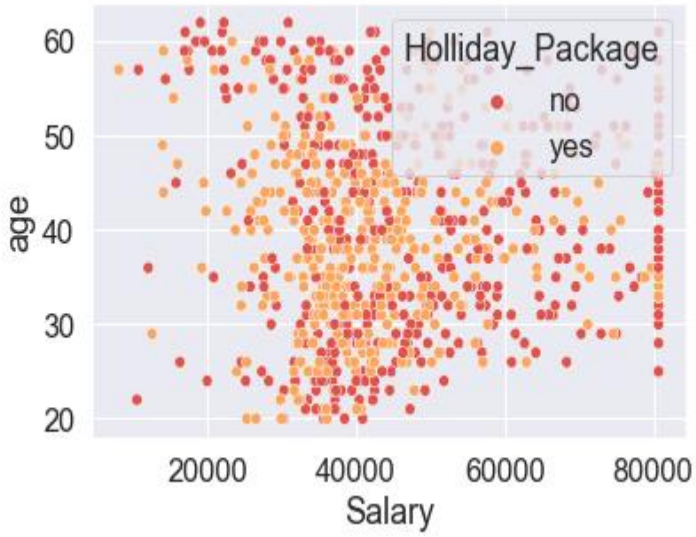


Figure 11: Age Vs Salary

- For a given dataset, 2 continuous numerical variable 'age' and 'Salary' do not have any correlation, seems to be a weak predictor.

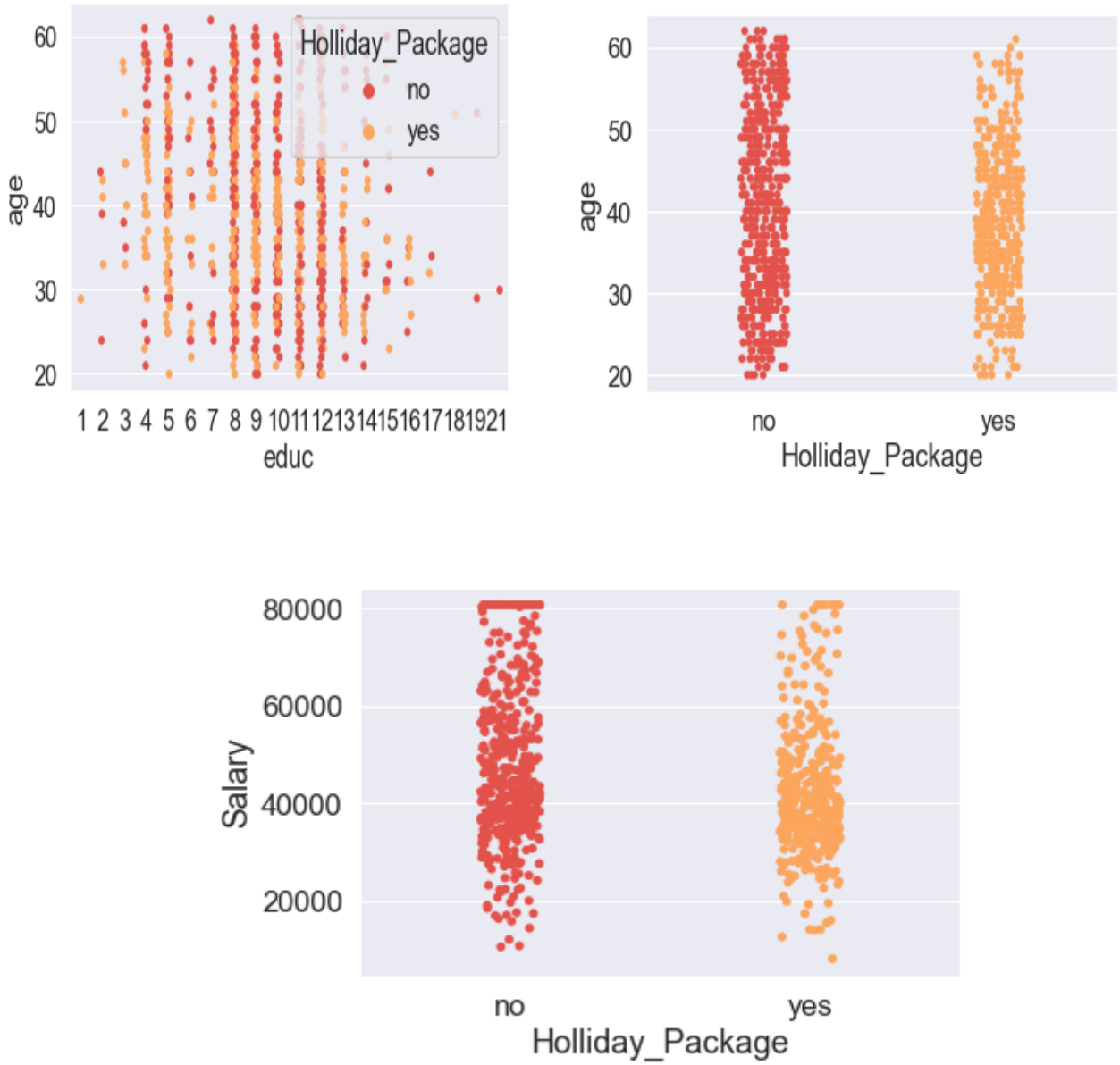


Figure 12: Bivariate Analysis

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

Solution:

Encode the Data:

- Since there are 5 categorical variable [including the target variable] or columns with the type as object, these variable were converted into categorical type using pd.categories to avoid higher dimensionality.
- After encoding the data, the target variable 'Holliday_Package was captured in to separate vector for training and test data set.
- Then, the data was split into train and test in the ratio of 70:30.

```
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,)
```

Figure-13: Data-Split [70:30]

Apply Logistic Regression:

Logistic regression model was built with a specific parameter setting in order to increase the model accuracy.

Below are the following parameter setting:

max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg', verbose=True

With the above parameter setting values were predicted for both training and test dataset. Also, probability prediction of classes were done on both train and test dataset.

Apply Linear Discriminant Analysis:

Linear Discriminant analysis was built without any specific parameter setting and then values were predicted on both training and test dataset. Also, probability prediction of classes were done on both train and test dataset.

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.

Solution:

1. Model Evaluation-Logistic Regression:

- Accuracy on Training set is 66.7% and 64.9% on testing set.
- **Classification report:**

Classification report for Logistic Regression model on Training set is

	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

Figure-14a: Classification Report on Training set

Classification report for Logistic Regression model on Testing set is

	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

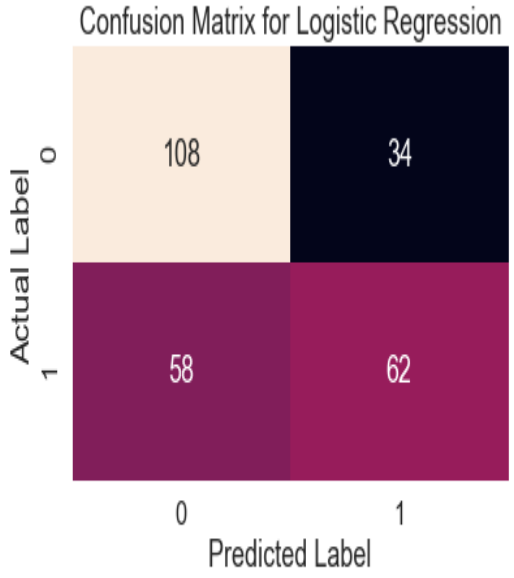
Figure-14b: Classification Report on Testing set

- **Confusion Matrix:**

Confusion Matrix for Logistic Regression model on Training set is

LR_train_precision 0.66
LR_train_recall 0.58
LR_train_f1 0.62

Confusion Matrix for Logistic Regression model on Testing set is

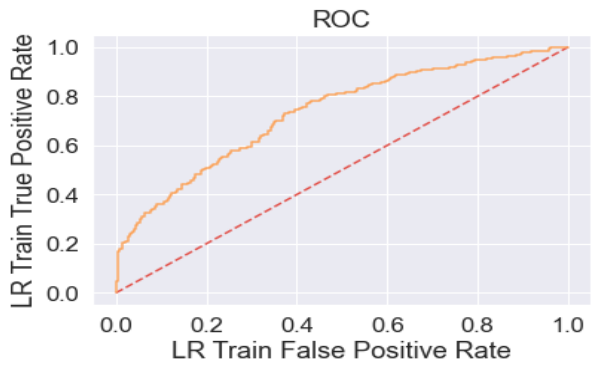


LR_test_precision 0.65
LR_test_recall 0.52
LR_test_f1 0.57

Figure-15: Confusion Matrix on Train and Test data

- AUC on the train dataset is 73.3% and 71.5%.

AUC: 0.733



AUC: 0.715

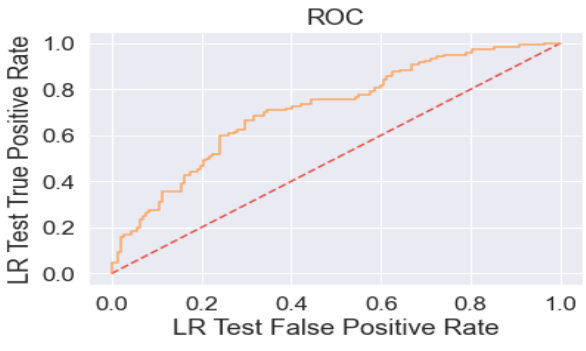


Figure-16: ROC Curve on Train and Test data

2. Model Evaluation-Linear Discriminant Analysis:

- Accuracy on Training set is 66.2% and 64.9% on testing set.
- **Classification report:**

Classification report for Linear Discriminant Analysis model on Training set is

	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

Figure-17a: Classification Report on Training set

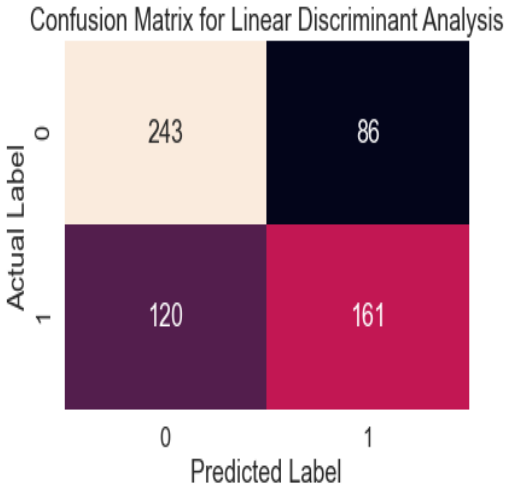
Classification report for Linear Discriminant Analysis model on Testing set is

	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

Figure-17b: Classification Report on Testing set

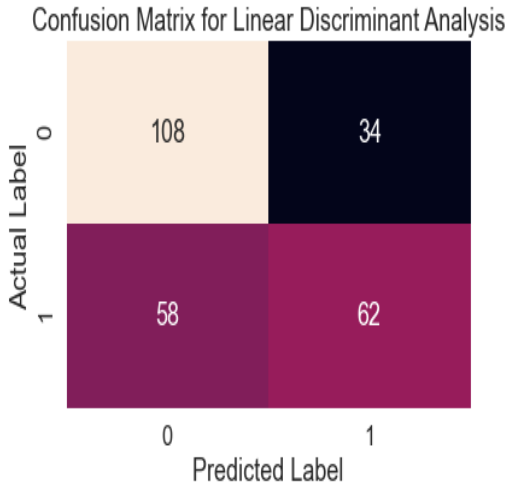
- **Confusion Matrix:**

Confusion Matrix for Linear Discriminant Analysis model on Training set is



LDA_train_precision 0.65
LDA_train_recall 0.57
LDA_train_f1 0.61

Confusion Matrix for Linear Discriminant Analysis model on Testing set is



LDA_test_precision 0.65
LDA_test_recall 0.52
LDA_test_f1 0.57

Figure-18: Confusion Matrix on Train and Test data

- AUC on the train dataset is 73.1% and 71.4%.

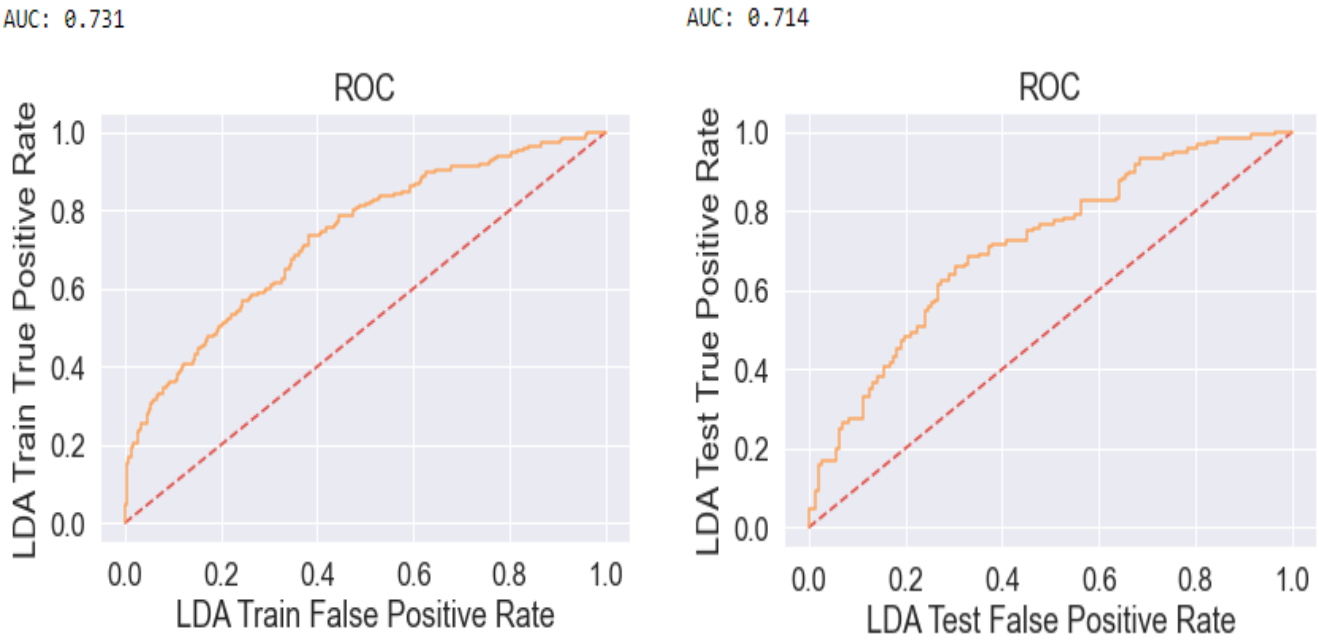


Figure-19: ROC Curve on Train and Test data

Final Model Comparison:

The model selection criterion would be the model which has more accurately predicted the class '1' which is an employee who opted for holiday package.

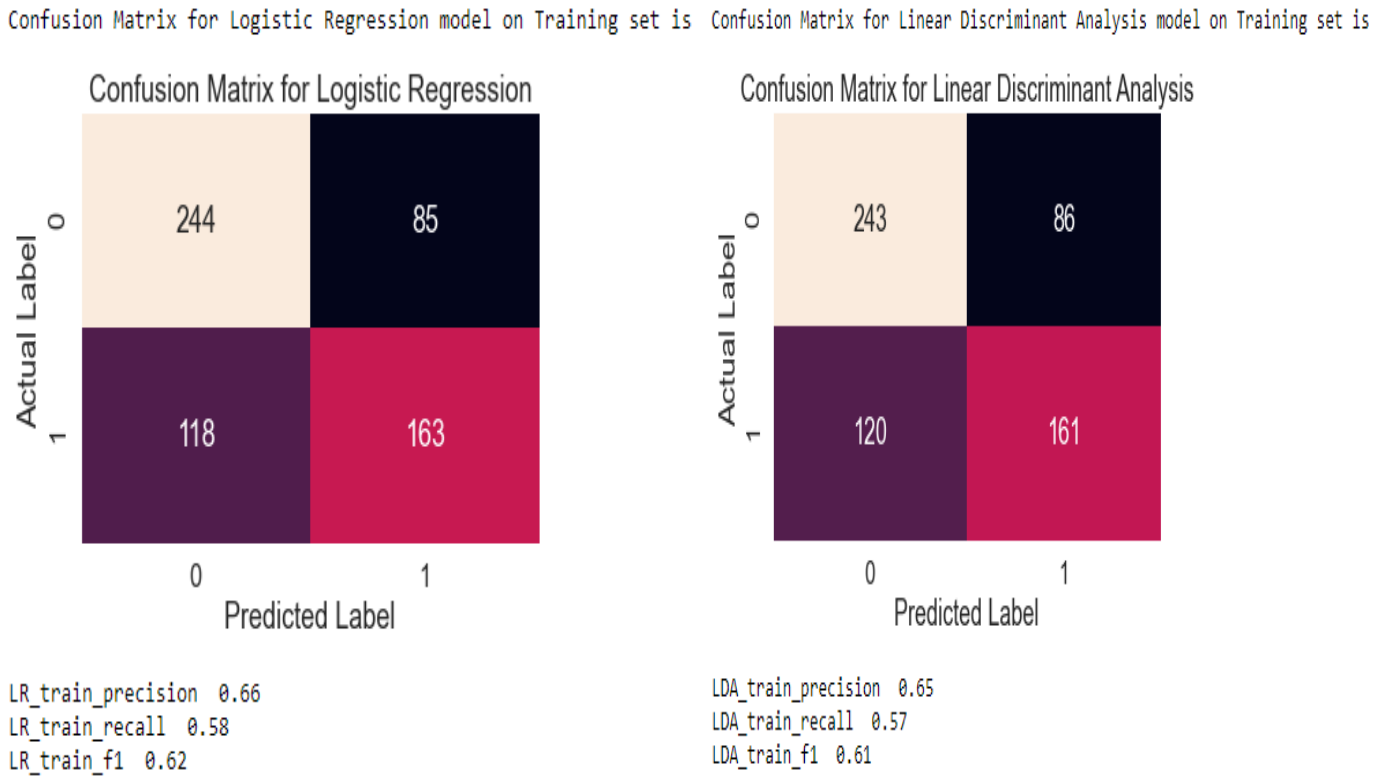
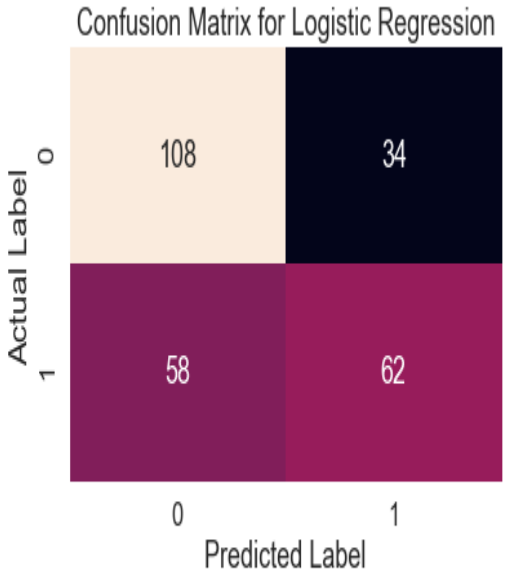


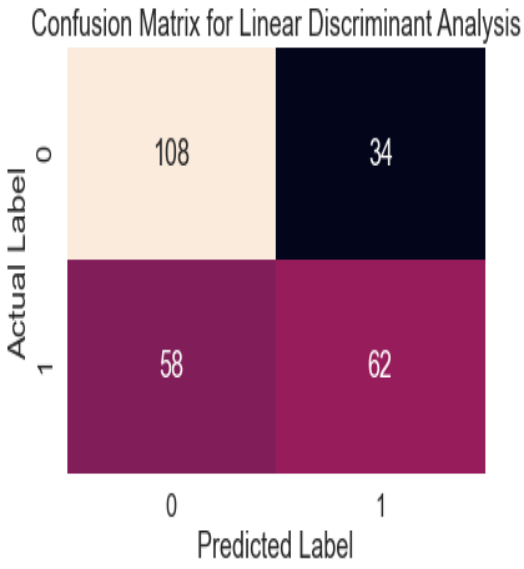
Figure-20: Comparison of Confusion Matrices for LR and LDA on Train data

Confusion Matrix for Logistic Regression model on Testing set is



LR_test_precision 0.65
LR_test_recall 0.52
LR_test_f1 0.57

Confusion Matrix for Linear Discriminant Analysis model on Testing set is



LDA_test_precision 0.65
LDA_test_recall 0.52
LDA_test_f1 0.57

Figure-21: Comparison of Confusion Matrices for LR and LDA on Test data

From the below consolidated table of performance metrics, since both the model has similar recall rate, we can conclude that any one of the model can be used for prediction.

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.67	0.65	0.66	0.65
AUC	0.73	0.72	0.73	0.71
Recall	0.58	0.52	0.57	0.52
Precision	0.66	0.65	0.65	0.65
F1 Score	0.62	0.57	0.61	0.57

Figure-22: Comparison of Performance Metrics for LR and LDA

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

Solution:

Recommendations and Business Insights:

- Overall, Logistic regression and LDA performed poorly with almost 65% accuracy.
- Years of formal education and foreigner variable has higher coefficient, both the variables becomes an important factor for class prediction.

- Since we have a fewer observations model accuracy was nearly low, travel agency should try to collect more samples in order to increase recall rate of a model.