Predictive modeling

27.06.2021

MANDIRA ROY
BURDWAN,WEST BENGAL
713101

CONTENT

Goals

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	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.

Dataset for Problem 1: cubic_zirconia.csv

Problem 2: Logistic Regression and LDA

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and

some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

Dataset for Problem 2: Holiday_Package.csv

Data Dictionary:

Variable Name	Description
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Criteria

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.

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?

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

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

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.

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).

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.

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

ANSWER

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.

Before doing analysis on the data lets first take a look at the data.

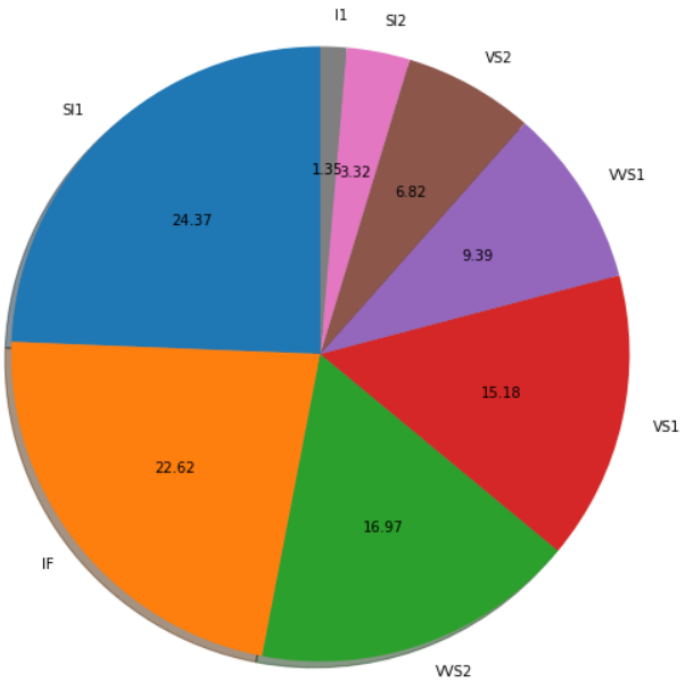
	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779
5	6	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	7	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	8	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	9	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	10	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706
10	11	0.32	Ideal	E	VS2	61.6	56.0	4.40	4.43	2.72	637
11	12	1.10	Premium	D	SI1	60.7	55.0	6.74	6.71	4.08	6468
12	13	0.50	Good	E	VS1	61.1	58.2	5.08	5.12	3.11	1932
13	14	0.71	Ideal	D	SI2	61.6	55.0	5.74	5.76	3.54	2767
14	15	1.50	Fair	G	VS2	66.2	53.0	7.12	7.08	4.70	10644
15	16	0.31	Ideal	G	VS2	61.6	55.0	4.37	4.39	2.70	544
16	17	0.34	Ideal	G	SI1	61.2	57.0	4.56	4.53	2.78	650
17	18	1.01	Ideal	D	VS2	59.8	56.0	6.52	6.49	3.89	7127
18	19	0.90	Good	D	SI1	61.9	64.0	6.00	6.09	3.74	3567
19	20	0.54	Premium	G	VS2	60.0	59.0	5.42	5.22	3.19	1637
20	21	1.04	Premium	D	VVS2	61.1	60.0	6.54	6.51	3.99	10984
21	22	0.40	Ideal	F	VS2	62.9	57.0	4.72	4.69	2.96	1080
22	23	1.52	Ideal	D	SI2	62.7	56.0	7.35	7.28	4.59	8631
23	24	1.19	Ideal	J	SI2	61.7	56.0	6.80	6.85	4.21	4508

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  26967 non-null  int64
 1   carat       26967 non-null  float64
 2   cut         26967 non-null  object
 3   color       26967 non-null  object
 4   clarity     26967 non-null  object
 5   depth       26270 non-null  float64
 6   table       26967 non-null  float64
 7   x           26967 non-null  float64
 8   y           26967 non-null  float64
 9   z           26967 non-null  float64
 10  price       26967 non-null  int64
dtypes: float64(6), int64(2), object(3)
memory usage: 2.3+ MB
```

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

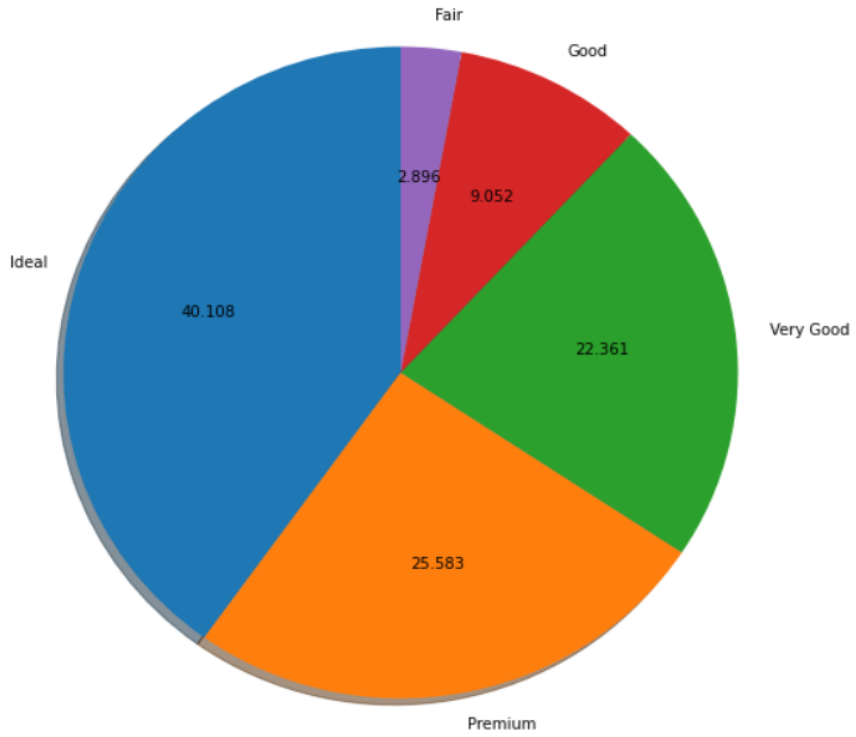
Inference:As we can see there are a total of 11 columns and 26967 rows where 697 entries are missing in the depth column . We are going to impute the median value in the null places of the depth column . We have values of object integer and float data type . Here our target variable is the price column. There are no duplicate values .

UNIVARIATE ANALYSIS:



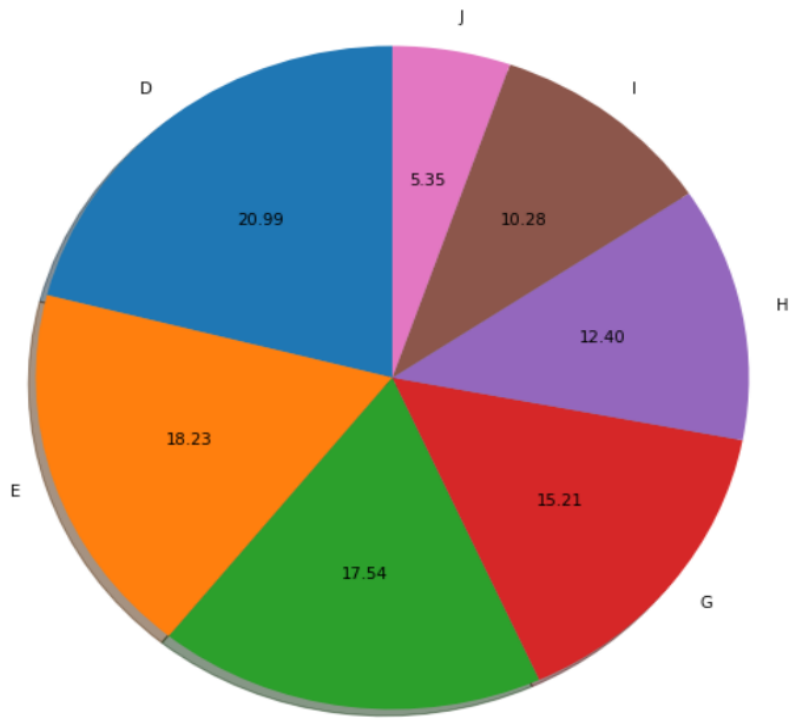
CLARITY PLOT:

INFERENCE:Here the SI1 and IF covers the maximum plot with the percentage of 24% and 23%.



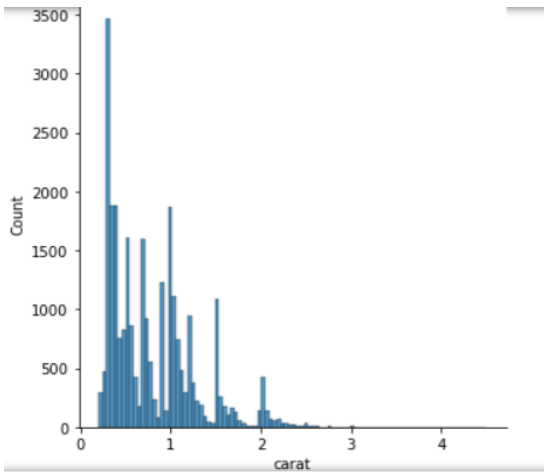
CUT PLOT:

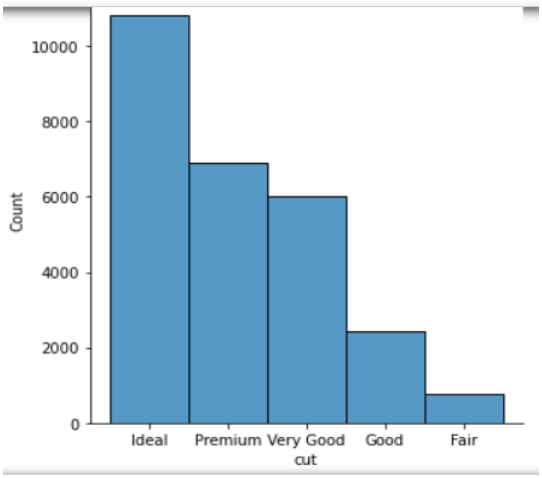
INFERENCE: Here the Ideal and Premium covers the maximum area with a percentage of 40% and 25.5%.

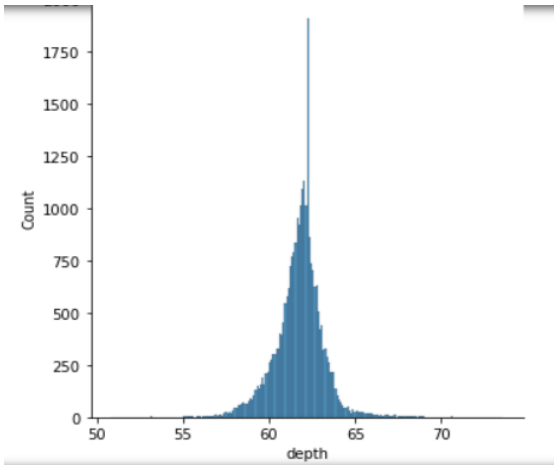


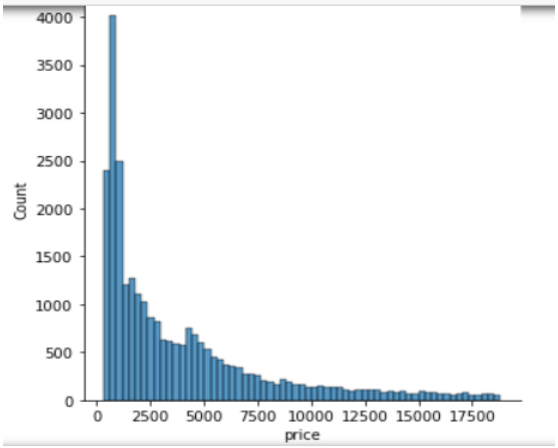
COLOR PLOT:

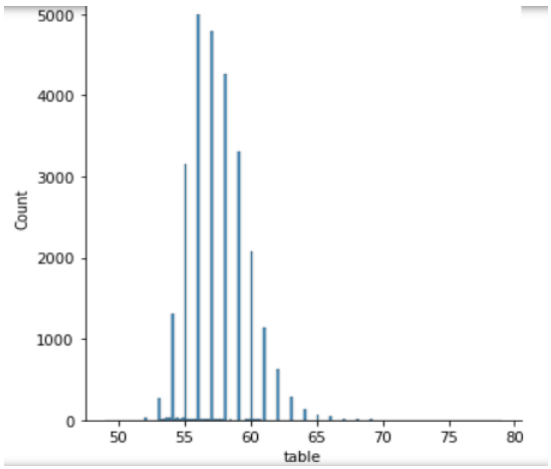
INFERENCE: Here the D and E color covers the maximum area and thus has the maximum popularity in the market.

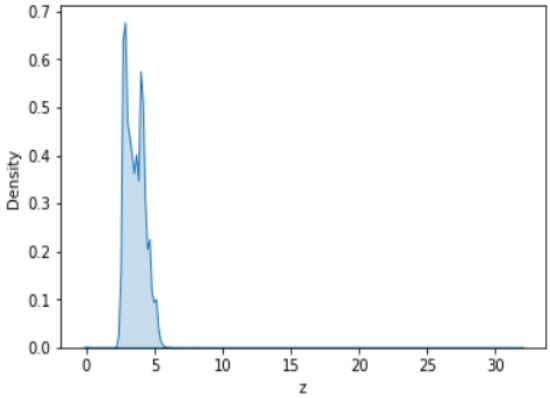

Count
carat

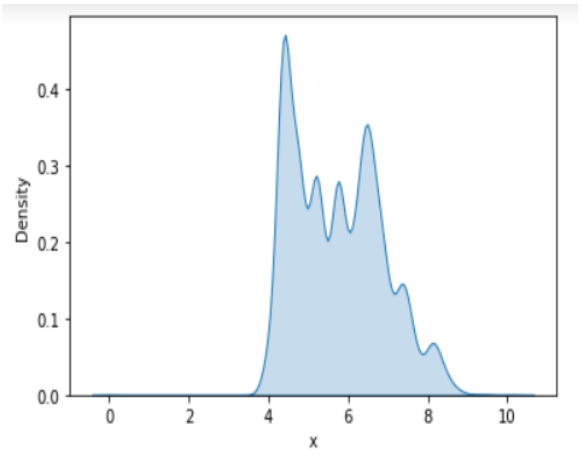

Count
Ideal
Premium
Very Good
Good
Fair
cut

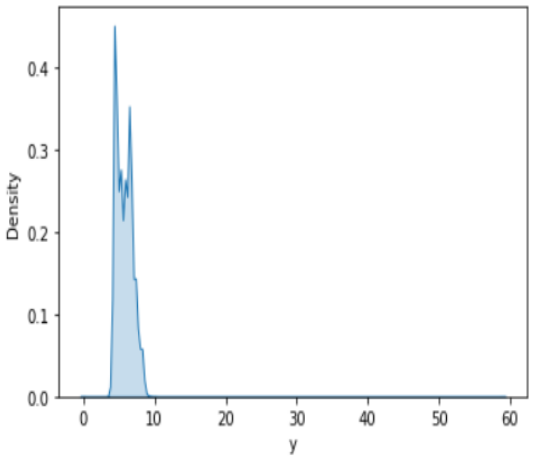

Count
depth


Count
price

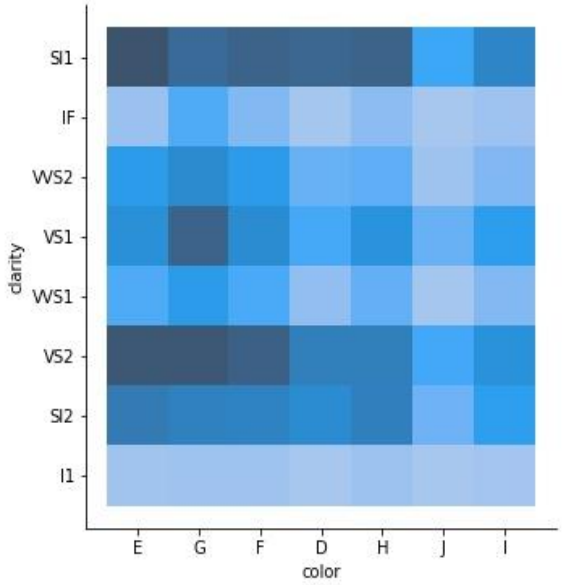


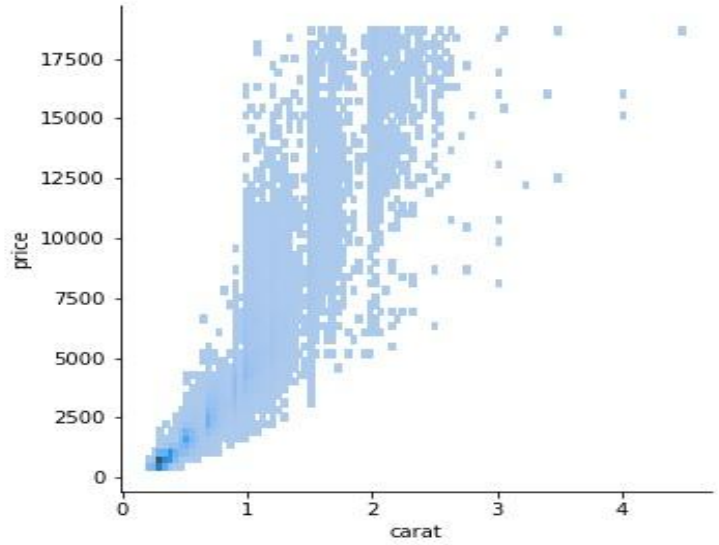


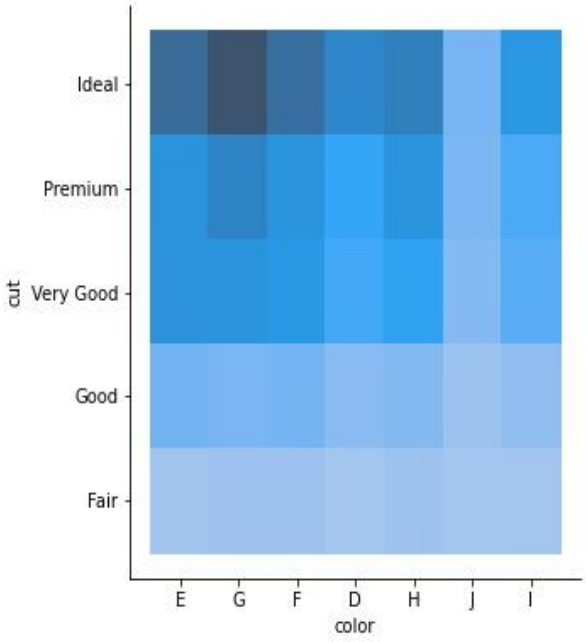


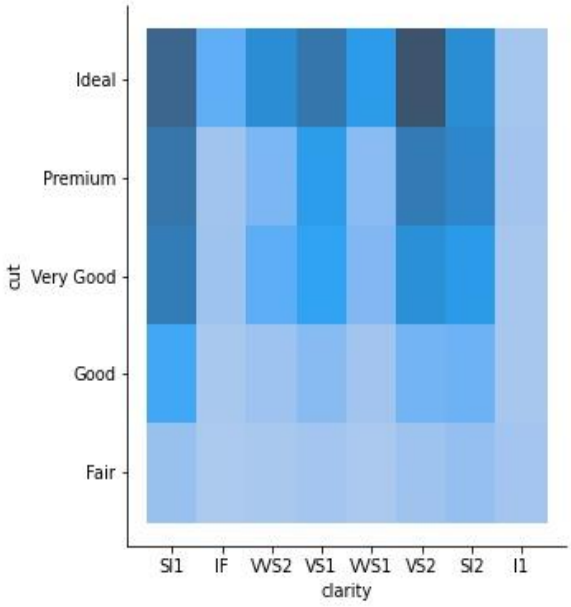


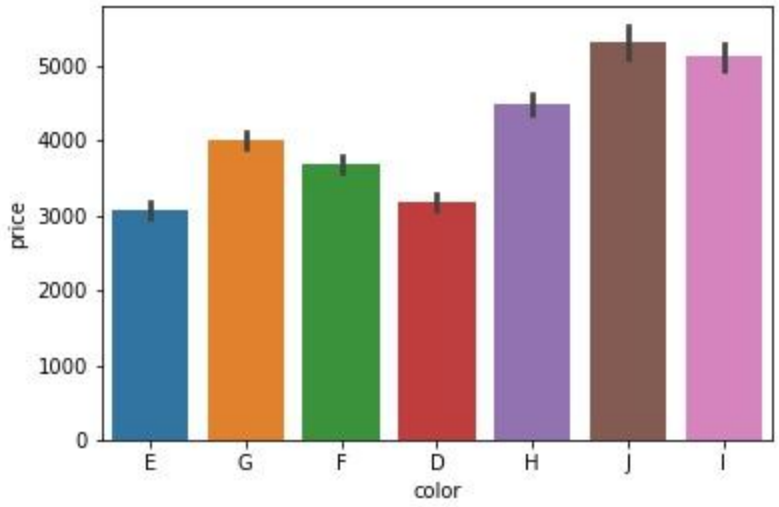
BIVARIATE ANALYSIS:

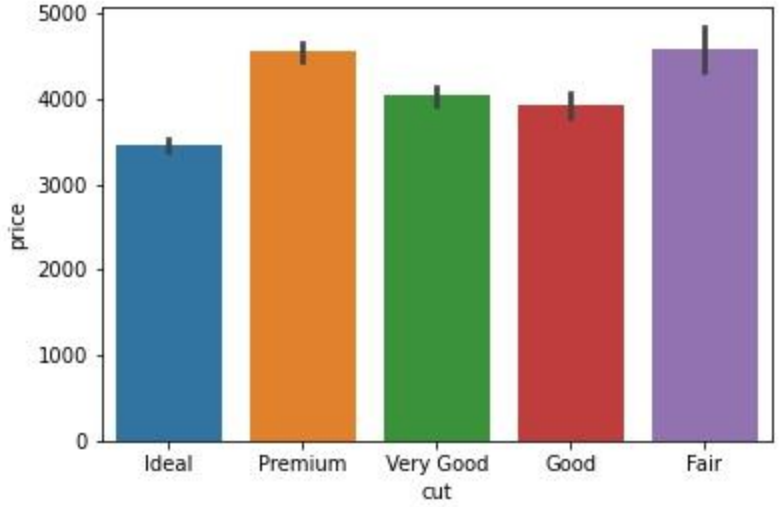



price
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
carat


cut
Ideal
Premium
Very Good
Good
Fair
E
G
F
D
H
J
I
color


cut
Ideal
Premium
Very Good
Good
Fair
SI1
IF
VVS2
VS1
VVS1
VS2
SI2
I1
clarity





INFERENCE: Seems like price for the fair an premium cut is high in the market an in case o color 'J' and 'I' is quite high rated.There is a good correlation between the ideal cut and G color , ideal cut and VS2 clarity,SI1 clarity vs E color,VS2 clarity vs E color,VS2 clarity vs G color.

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?

INFERENCE:There are some null values which have been imputed with the median values or the mode values in the continuous and categorical columns . There are no values which are below zero according to the given data. In this case the scaling is not necessary but should be done as the mean,min,max value are't overlapping and there is a little difference in the scale of the data and moreover it is a good practice to do .Though the scaling won't be affecting the results very much.

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

```
==============================================================================
Dep. Variable:                  price   R-squared:                       0.917
Model:                            OLS   Adj. R-squared:                  0.917
Method:                 Least Squares   F-statistic:                 1.986e+04
Date:                Wed, 30 Jun 2021   Prob (F-statistic):               0.00
Time:                        01:35:27   Log-Likelihood:            -2.2853e+05
No. Observations:               26967   AIC:                         4.571e+05
Df Residuals:                   26951   BIC:                         4.572e+05
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     8350.5761    442.844     18.857      0.000    7482.579    9218.573
carat         1.121e+04     71.666    156.433      0.000    1.11e+04    1.14e+04
cut             21.3897      6.077      3.520      0.000       9.479      33.300
color          327.5319      4.392     74.567      0.000     318.923     336.141
depth         -106.7385      5.960    -17.909      0.000    -118.420     -95.057
table          -64.2374      3.688    -17.418      0.000     -71.466     -57.009
x             -983.7812     42.162    -23.333      0.000   -1066.422    -901.141
y               15.1845     22.581      0.672      0.501     -29.075      59.444
z              -73.1399     36.891     -1.983      0.047    -145.448      -0.831
clarity_I1   -2876.6814     81.812    -35.162      0.000   -3037.036   -2716.326
clarity_IF    2628.7917     62.801     41.859      0.000    2505.699    2751.885
clarity_SI1    859.9506     59.059     14.561      0.000     744.192     975.709
clarity_SI2   -111.8569     59.786     -1.871      0.061    -229.040       5.326
clarity_VS1   1819.2726     57.936     31.401      0.000    1705.714    1932.831
clarity_VS2   1494.4039     57.809     25.851      0.000    1381.094    1607.713
clarity_VVS1  2308.8647     59.043     39.105      0.000    2193.138    2424.591
clarity_VVS2  2227.8309     58.515     38.073      0.000    2113.138    2342.524
==============================================================================
Omnibus:                     6446.582   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           182307.870
Skew:                           0.530   Prob(JB):                         0.00
Kurtosis:                      15.693   Cond. No.                     3.11e+17
==============================================================================
```

INFERENCE:

For train data ,results by manual calculation:

MAE: 770.7919231487668

MSE: 1356307.8434150126

RMSE: 1164.606304042277
R-Squared: 0.9161451087135447

For test data results by manual calculation:

MAE: 770.6578567021251

MSE: 1317848.8909343185

RMSE: 1147.9759975427703
R-Squared: 0.9650756787096231

Clearly from the above manual calculations we are getting an under fitted model.

But for the entire data we are getting an adjusted r value as 91% which is a good score comparatively .

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

INFERENCE AND RECOMMENDATIONS: The model is showing the r_squared and adjusted r_squared value to be 91.7% for both the cases and the total p_value is less than 0.05.

Intercept	8350.576095
carat	11210.882385
cut	21.389660
color	327.531921
depth	-106.738533
table	-64.237389
x	-983.781231
y	15.184484
z	-73.139891
clarity_I1	-2876.681409
clarity_IF	2628.791679

clarity_SI1 859.950649

clarity_SI2 -111.856930

clarity_VS1 1819.272647

clarity_VS2 1494.403941

clarity_VVS1 2308.864657

clarity_VVS2 2227.830862

From the above table an observations we can see that the most significant features are the carat ,color,cut,clarity are the good attributes to take in observation but the p-value for width ie., y value is above 0.05 and the clarity of S_I2 is greater than 0.05 proving the attributes to be useless. We can also observe the in every one unit increase in color cut carat width(ie., y) ,clarity there is good or some increase in the price like when the carat increases by 1 unit there is an increase in the price by 11210.88 unit keep other predictors constant. There are some negative intercepts in the above table which states that they are inversely proportional to the diamond price.We observe the x coefficient to be -983.78 from which we can conclude the increase in length the decrease in the dimon price similarly with the z height of the cube increase the price decrease facing less profit . From the analysis and scoring we can see that surely the model is a good predictor model with a 91% accuracy rate and very strong correlation between the predicted and actual (90.64%train data,91.6% test data)but there is a presence of some kind of noise in the data.I would like to recommend for better prediction the diamond company should consider these features-'Carat','Cut','Color','Clarity',and width(in come cases) .The company should be aware of the fact the higher the height(z) an length(x) more the lesser the price as that will lead to a diamond with a dark appearance as more the height less it will reflect light.

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.

Before jumping into the analysis let's look into the data first.

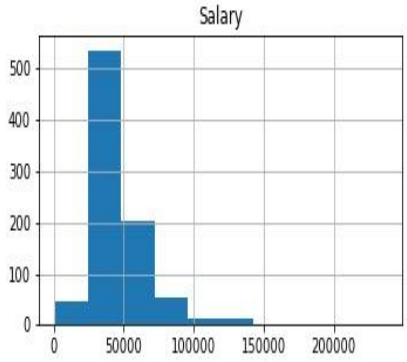
	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no
5	6	yes	61590	42	12	0	1	no
6	7	no	94344	51	8	0	0	no
7	8	yes	35987	32	8	0	2	no
8	9	no	41140	39	12	0	0	no
9	10	no	35826	43	11	0	2	no
10	11	no	42643	45	11	0	2	no
11	12	no	35157	60	12	0	0	no
12	13	no	75327	33	11	2	0	no
13	14	no	148221	56	14	0	0	no
14	15	no	98870	56	11	0	0	no
15	16	no	80297	47	11	0	1	no
16	17	no	52117	50	8	0	0	no
17	18	yes	139253	39	12	0	0	no
18	19	no	62858	47	8	0	1	no
19	20	yes	57400	53	11	0	0	no
20	21	no	52059	29	19	0	0	no
21	22	yes	66711	46	11	0	1	no
22	23	no	51463	44	8	0	2	no
23	24	no	35682	20	12	1	0	no

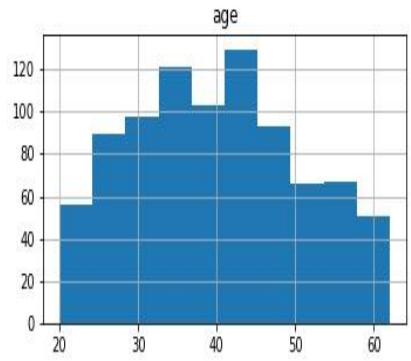
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Salary                872 non-null    int64
 1   age                   872 non-null    int64
 2   educ                  872 non-null    int64
 3   no_young_children     872 non-null    int64
 4   no_older_children     872 non-null    int64
 5   Holliday_Package_yes  872 non-null    uint8
 6   foreign_yes           872 non-null    uint8
dtypes: int64(5), uint8(2)
memory usage: 35.9 KB
```

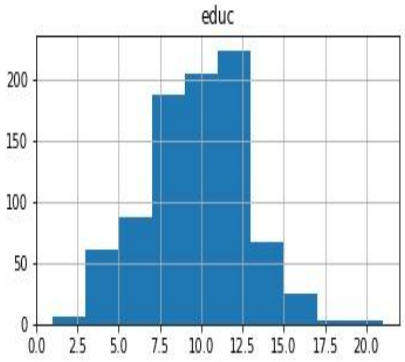
	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	47729.172018	39.955275	9.307339	0.311927	0.982798	0.459862	0.247706
std	23418.668531	10.551675	3.036259	0.612870	1.086786	0.498672	0.431928
min	1322.000000	20.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	35324.000000	32.000000	8.000000	0.000000	0.000000	0.000000	0.000000
50%	41903.500000	39.000000	9.000000	0.000000	1.000000	0.000000	0.000000
75%	53469.500000	48.000000	12.000000	0.000000	2.000000	1.000000	0.000000
max	236961.000000	62.000000	21.000000	3.000000	6.000000	1.000000	1.000000

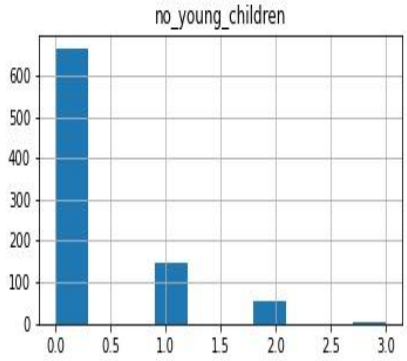
After looking into the data we can see there are 872 entries with 8 columns including the index column(Unnamed:0). The mean and standard deviation of salary and age column is scale higher than other columns .There are 4 integer variables and 2 categorical variables (holiday package and foreign column).There are no null values or duplicate values.

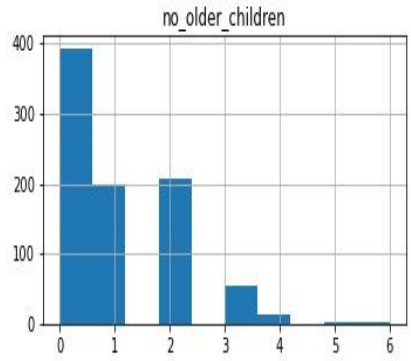
UNIVARIATE ANALYSIS:

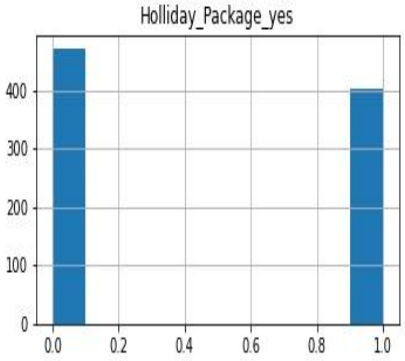

Salary
0
100
200
300
400
500
0
50000
100000
150000
200000

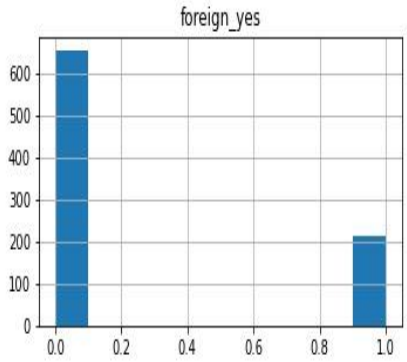

age
0
20
40
60
80
100
120
20
30
40
50
60

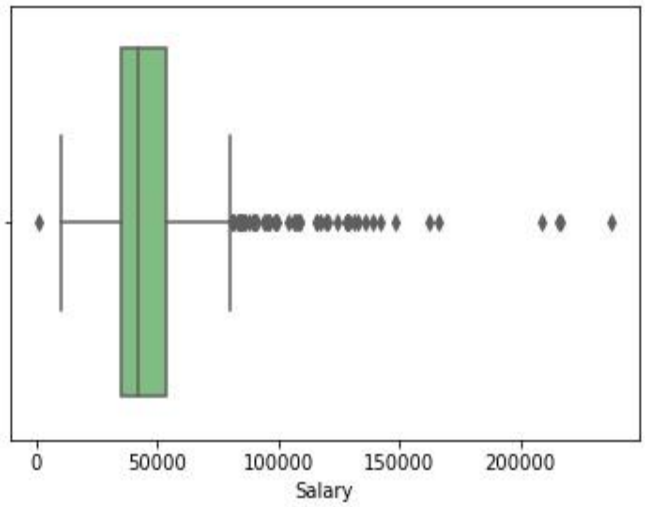

educ
0
50
100
150
200
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

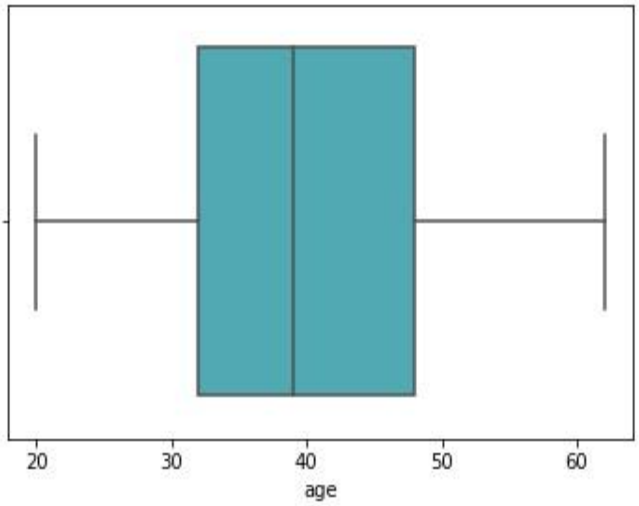

no_young_children
0
100
200
300
400
500
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0

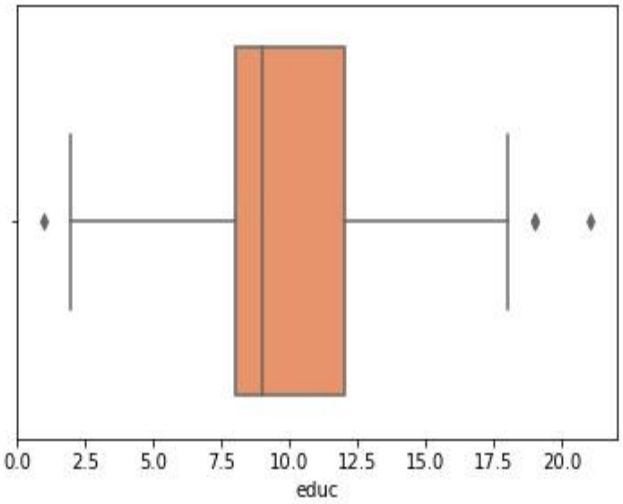

no_older_children
0
100
200
300
400
0
1
2
3
4
5
6

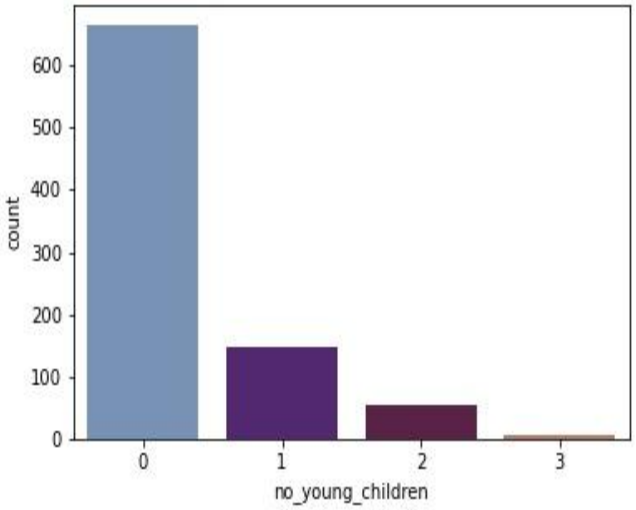

Holliday_Package_yes
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0

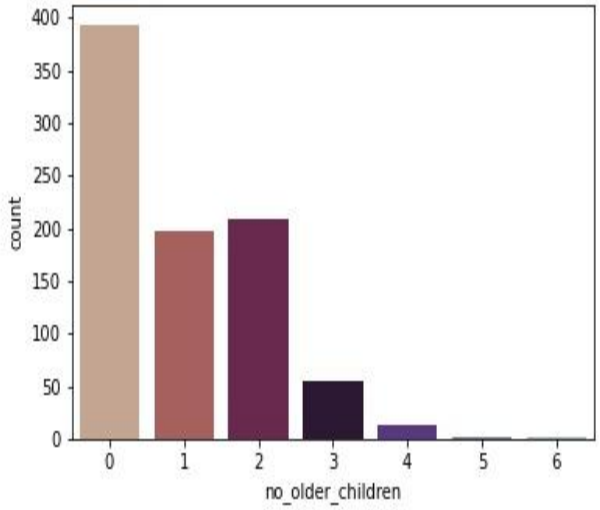

foreign_yes
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0

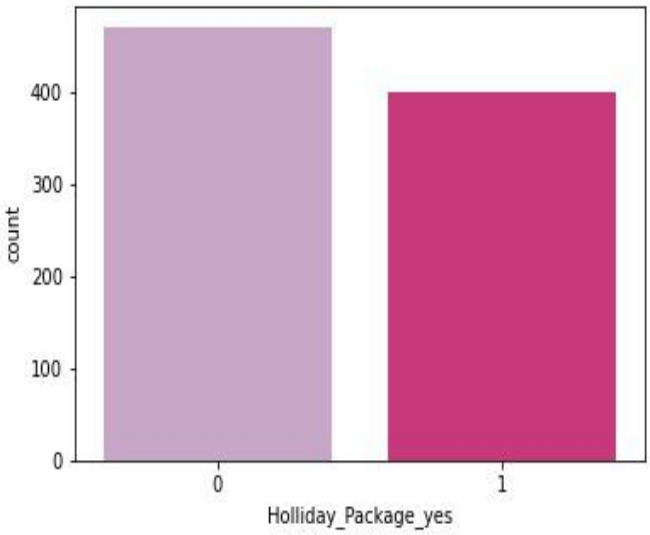

0
50000
100000
150000
200000
Salary

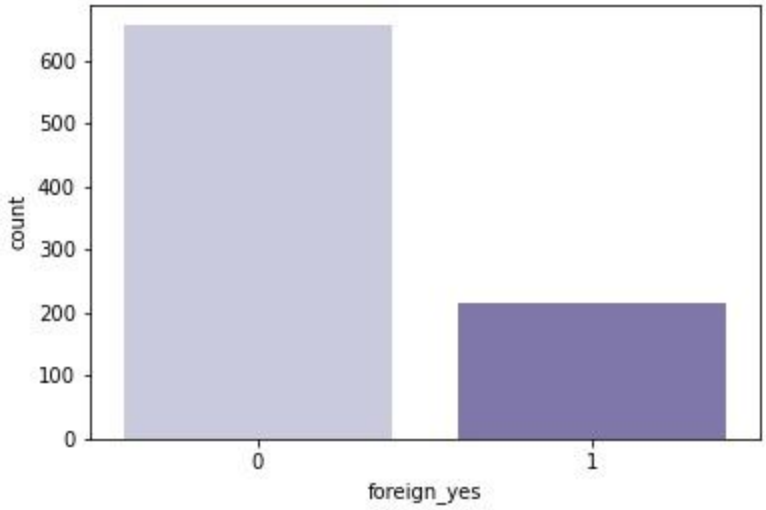

20
30
40
50
60
age


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ


count
0
100
200
300
400
500
600
0
1
2
3
no_young_children

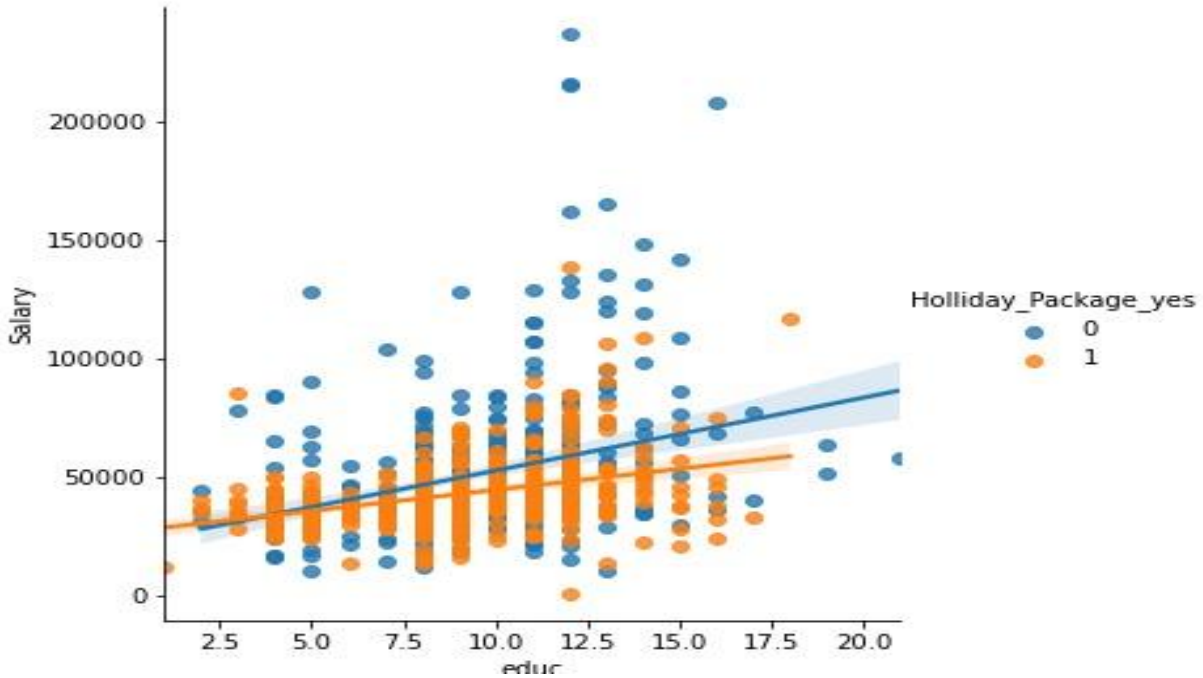

count
0
50
100
150
200
250
300
350
400
0
1
2
3
4
5
6
no_older_children

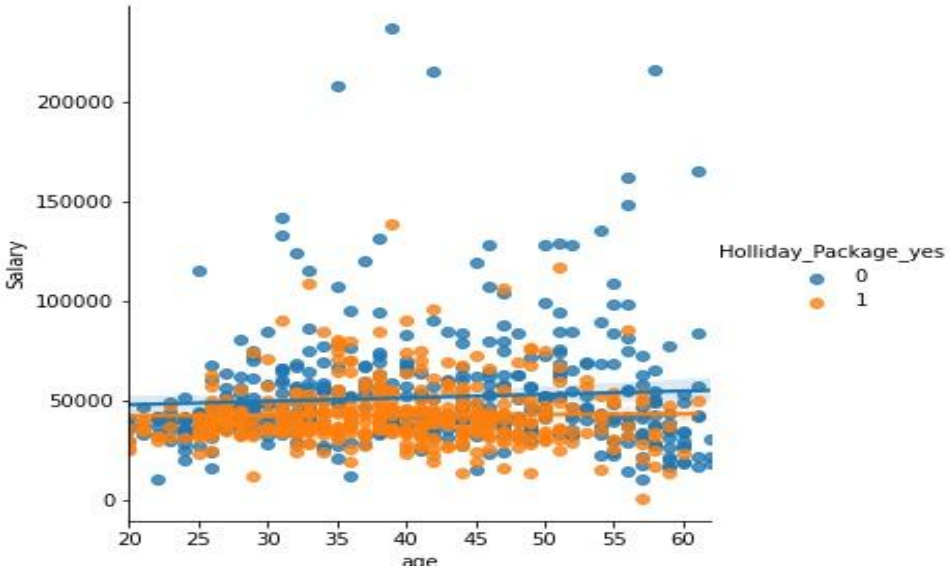

count
0
100
200
300
400
0
1
Holliday_Package_yes

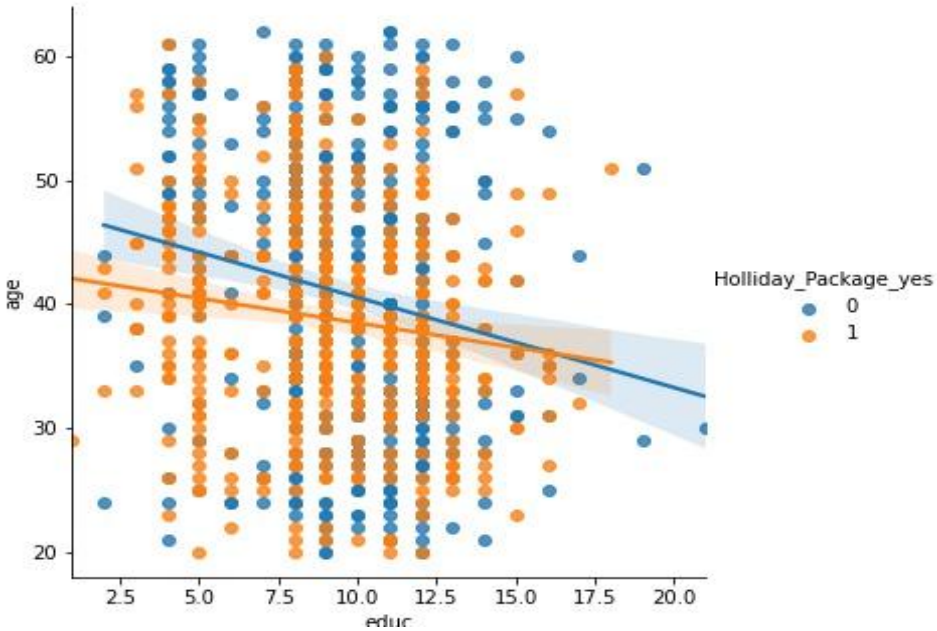


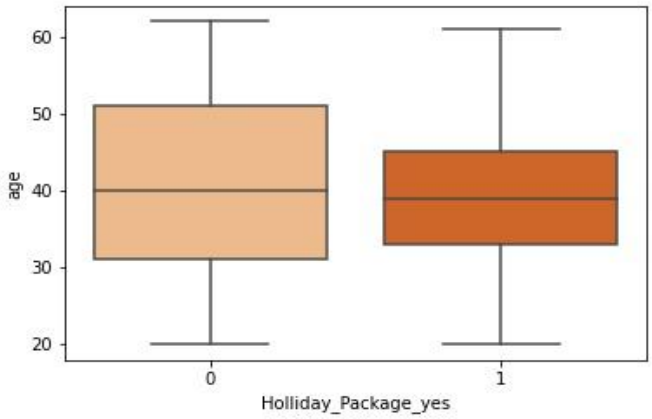
INFERENCE:Here we have done some univariate analysis . We have one one hot encoding to the data and kept the holiday_package_yes column and foregin_yes column. It seems like people opting for a holiday package is less than not opting for one and choosing the foreign package is much less than not choosing one. Other variables are normally distributed but the salary column is a little bit negatively skewed. There are some outliers present in the salary column and the education column. Though I don't have an idea about the outliers of the education column,I won't be treating the outliers here without consulting with the client. And the outliers in the salary columns are seem to be valid salary amount.

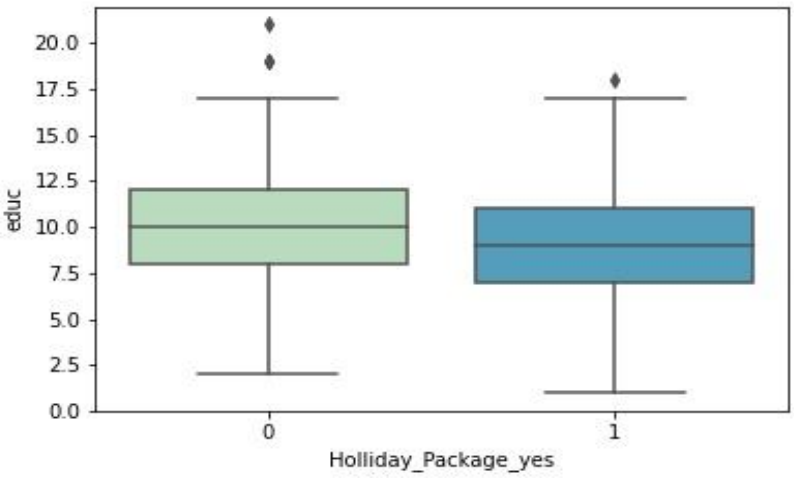
BIVARIATE ANALYSIS:

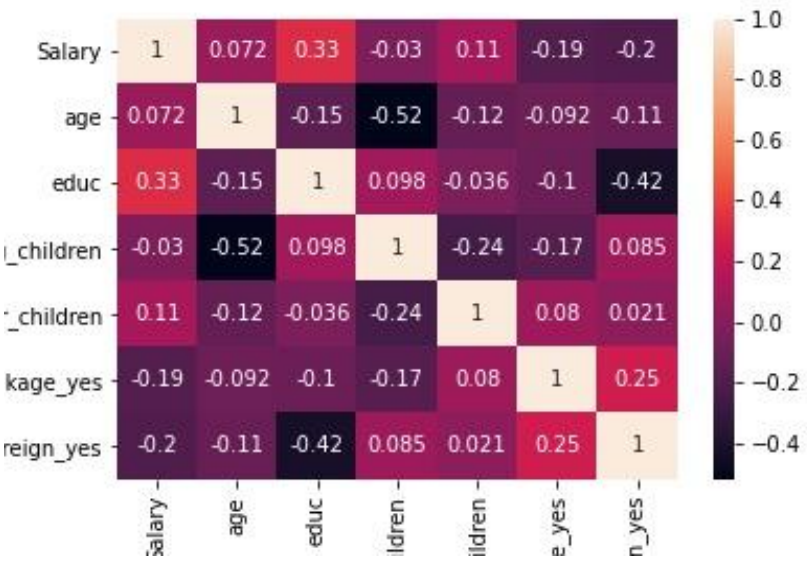



Salary
200000
150000
100000
50000
0
20
25
30
35
40
45
50
55
60
age
Holliday_Package_yes
0
1


age
60
50
40
30
20
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
educ
Holliday_Package_yes
0
1


age
60
50
40
30
20
0
1
Holliday_Package_yes





INFERENCE: From the analysis we can infer that there is a relatively fine correlation between the no_older_chilren and holiday_package_yes and foreign_yes and holiday_package_yes. There seems to be a relation between age and holiday package. More the age less there is a slightly less chance that they will choose the holiday package.We can also see that higher the salary more people are avoiding the holiday package.

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)

Have done one hot encoding for the two columns foreign and holiday_package and split the data into train set and test set in 70:30 proportion and random state is 1 an stratify is 'y', build the model .

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Linear discriminant analysis(LDA) confusion matrix

```
Confusion Matrix
 [[243  86]
 [119 162]]

Classification Report
              precision    recall  f1-score   support

           0       0.67      0.74      0.70       329
           1       0.65      0.58      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

Confusion Matrix
 [[109  33]
 [ 61  59]]

Classification Report
              precision    recall  f1-score   support

           0       0.64      0.77      0.70       142
           1       0.64      0.49      0.56       120

    accuracy                           0.64       262
   macro avg       0.64      0.63      0.63       262
weighted avg       0.64      0.64      0.63       262
```

Logistic Regression Confusion Matrix:

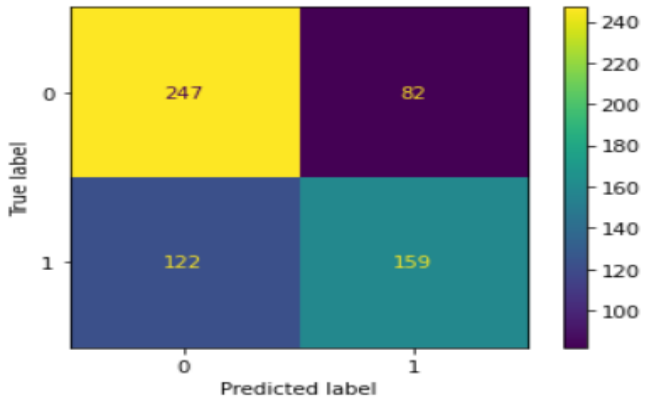
Classification Report

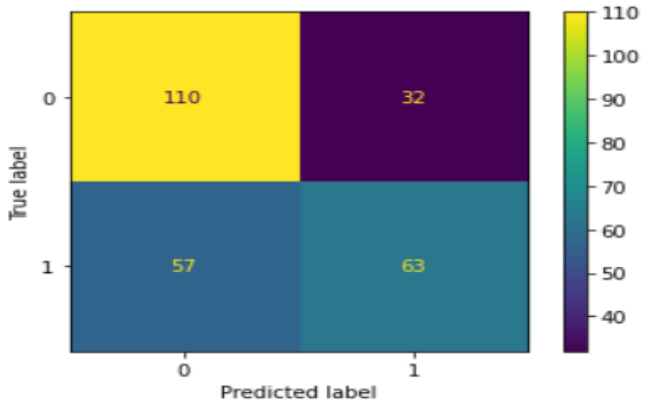
	precision	recall	f1-score	support
0	0.67	0.75	0.71	329
1	0.66	0.57	0.61	281
accuracy			0.67	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.67	0.66	610

Classification Report

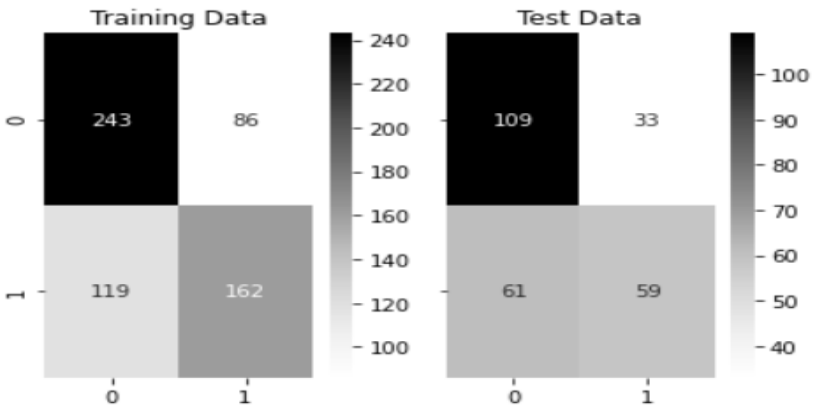
	precision	recall	f1-score	support
0	0.66	0.77	0.71	142
1	0.66	0.53	0.59	120
accuracy			0.66	262
macro avg	0.66	0.65	0.65	262
weighted avg	0.66	0.66	0.65	262

CONFUSION MATRIX OF LOGISTIC REGRESSION MODEL:



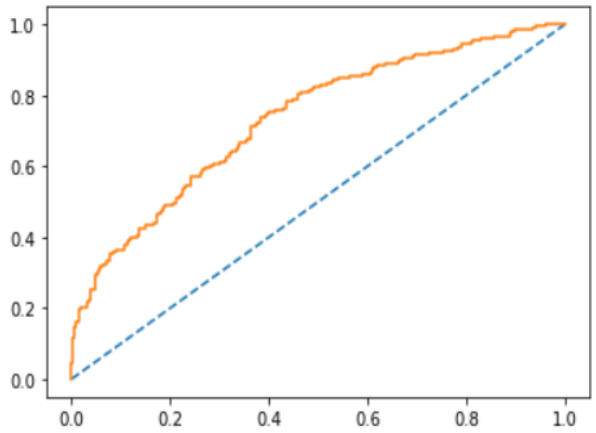


CONFUSION MATRIX OF LDA MODEL;

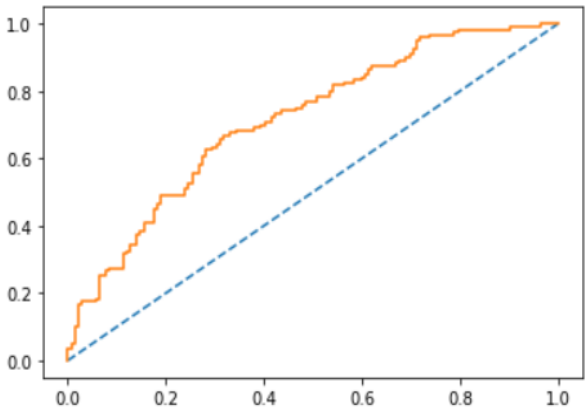


AUC AND ROC CURVE OF TRAIN AND TEST DATA IN LDA MODEL:

AUC: 0.733



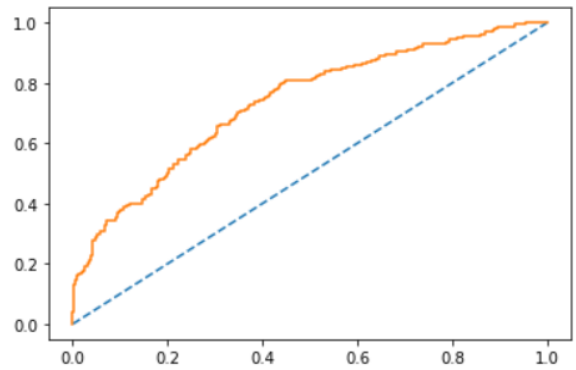
AUC: 0.733



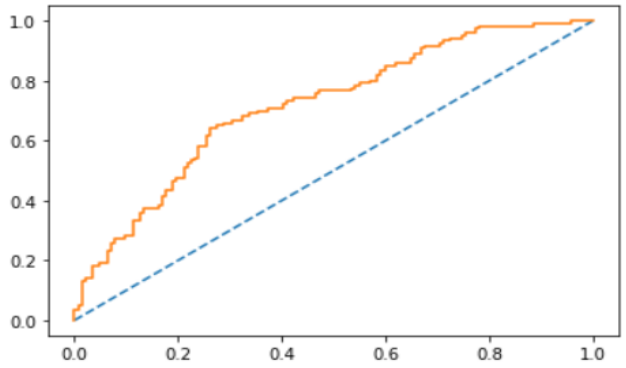
-

AUC AND ROC CURVE OF THE TRAIN AND TEST DATA OF LOGISTIC REGRESSION MODEL:

AUC: 0.735



AUC: 0.735



INFERENCE: Here we can see that the problem needs to predict the employees who are surely thinking of opting the holiday package option which is the true positive value is what the agency is searching for and so the recall is the most important aspect for judging the model . We are here interested in the '1' value not the zero because it won't affect the agency much.

Recall of the lda model train data is 58% and 49% for the test data whereas the recall value for the logistic regression model for the train dataset is 57% and for the test dataset is 53%.

The area under the curve is little bit high for the logistic regression model which is 73.5% and for the LDA model is 73.3% which is 0.2 % less so the logistic regression model is performing little bit better than the LDA model.The accuracy score of the logistic regression for the train is 67% an for test set is 66% where as the accuracy score for the LDA model is

66% for the training dataset and 64% for test dataset.Through the accuracy doesn't play an important role in this problem predicting still if we consider the accuracy then logistic regression is performing much better than LDA.

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

INFERENCE ACCORDING TO THE PREDICTION AND INSIGHTS AND RECOMMENDATION:

- As we have discussed earlier that the recall is very important in this problem and another thing to notice that in case of the recall of the lda model train data is 58% and 49% for the test data whereas the recall value for the logistic regression model for the train dataset is 57% and for the train dataset is 53% thus the difference between the training and test data is high in case of LDA model thus it may happen that in the realtime the accuracy there may be some cases where the data is predicted as true where the actual result is false which is if an employee will be opting for the package it won't be known by the agency and the agency may face some loss on sudden change in the scenario. Predicting a false negative isn't a problem here but the true negative may cause trouble for the agency. Though according to the logistic regression model the model is predicting with 66% accuracy that the recall of 53% of the test data as true and positive .
- Inorder to predict more accurately we will be needing more data or the prediction purpose from the agency.
- In the EDA we could notice that the aged persons are not opting for the holiday package above 50 . More the middle aged employees are going for the package with a age range of 30-40 .So the agency could create age specific holiday package like for old age people could plan for a religious place for the middle age employees could select for somewhere hill station cool and calm away from the city crowd for relaxation from work life.
- With higher age and higher salary people are less opting for the holiday package.Though salary is not that of an important attribute for selecting the holiday package.

THANK YOU.