PREDICTIVE MODELING ASSIGNMENT

BY JASPER SHELDON M

Contents

List of Figures

List of Tables

1. REGRESSION TECHNIQUE FOR PREDICTING PRICE OF ZIRCONIA

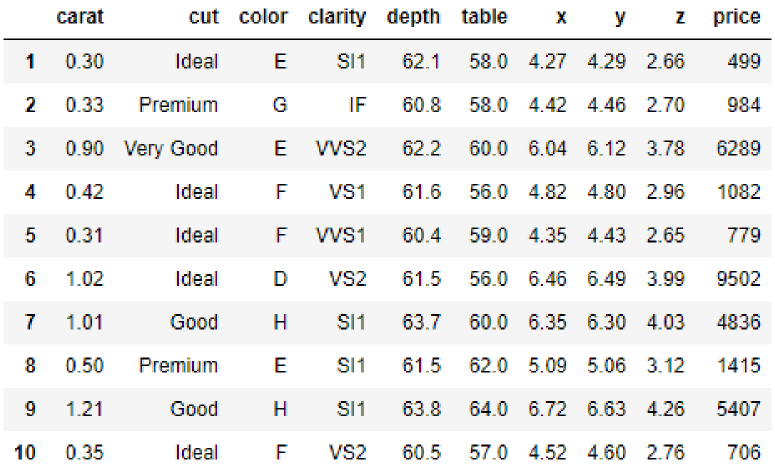
You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is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

1. DATA DESCRIPTION

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	Color of the cubic zirconia. With D being the worst and J the best.
Clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	The Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

Table 1.Data Dictionary

2. SAMPLE OF THE DATA

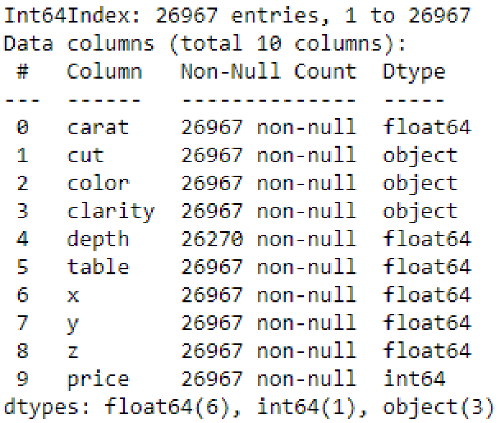


	carat	cut	color	clarity	depth	table	x	y	z	price
1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779
6	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
7	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
8	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
9	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
10	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	706

Table 2 Sample Data

3. EXPLORATORY DATA ANALYSIS

1. ATTRIBUTES OF THE DATA



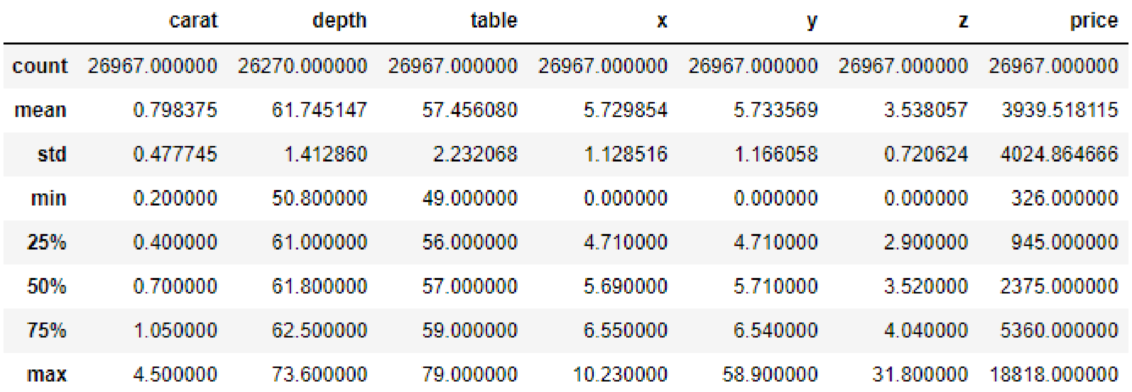
```
Int64Index: 26967 entries, 1 to 26967
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object 
 2   color    26967 non-null  object 
 3   clarity  26967 non-null  object 
 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)
```

Table 3 Information about the data

The dataset has 26967 rows and 10 columns including the dependent variable from the above table we can infer that there are 7 numerical columns and 3 categorical columns. Along with that we can see depth has few missing values in it (697 to be precise).

Numerical columns: CARAT, DEPTH, TABLE, X, Y, Z, PRICE

Categorical columns: CUT, COLOR, CLARITY



	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Table 4 Summary of the data

OBSERVATIONS:

CARAT: It has a mean of 0.79 and standard deviation of 0.47 and 50 percent of the data has value above 1.05, minimum value is 0.2 and maximum value is 4.50

DEPTH: It has a mean of 61.79 and standard deviation of 1.41 and 50 percent of the data has value above 61.8, minimum value is 50.80 and maximum value is 73.60

TABLE: It has a mean of 57.45 and standard deviation of 2.23 and 50 percent of the data has value above 57.0, minimum value is 49.0 and maximum value is 79.0

X: It has a mean of 5.79 and standard deviation of 1.12 and 50 percent of the data has value above 5.69, minimum value is 0 and maximum value is 10.23.

Y: It has a mean of 5.73 and standard deviation of 1.16 and 50 percent of the data has value above 5.71, minimum value is 0 and maximum value is 58.90

Z: It has a mean of 3.53 and standard deviation of 0.72 and 50 percent of the data has value above 3.52, minimum value is 0 and maximum value is 31.80

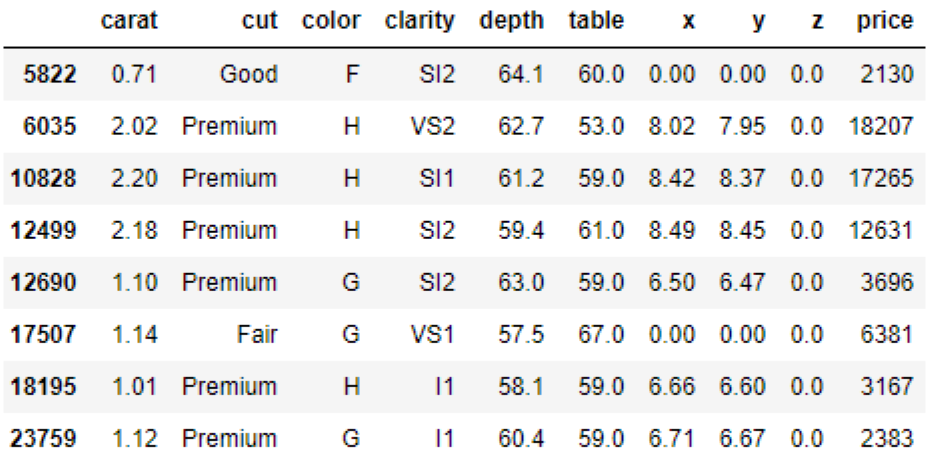
PRICE: It has a mean of 3939.51 and standard deviation of 4026 and 50 percent of the data has value above 2375, minimum value is 326 and maximum value is 18818.

Key findings:

X, Y, Z are dimensions of the zirconia stone logically it cannot be 0; we must impute the value for those particular entries. Standard deviation for price is more it denotes the spread for the price is very large.

There are 34 duplicated records in the dataset.

697 missing values for DEPTH.



	carat	cut	color	clarity	depth	table	x	y	z	price
5822	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6035	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
10828	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12499	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12690	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17507	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18195	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23759	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

Table 5 Dataset with 0s

The above table is a filtered section of the dataset which contains 0s. From business point of view these columns cannot have 0s as they describe the physical dimensions of the stone. Hence we can proceed with imputing the values for these entries are removing them. We have imputed the minimum value greater than 0 for all 3 columns. Along with the 0 imputation, we have also filled the missing values in the DEPTH column; there were 697 missing values it is imputed with the median value. Median value is chosen because outliers were present which will affect the mean. The duplicated records were removed keeping the original record.

2. UNNIVARIATE ANALYSIS

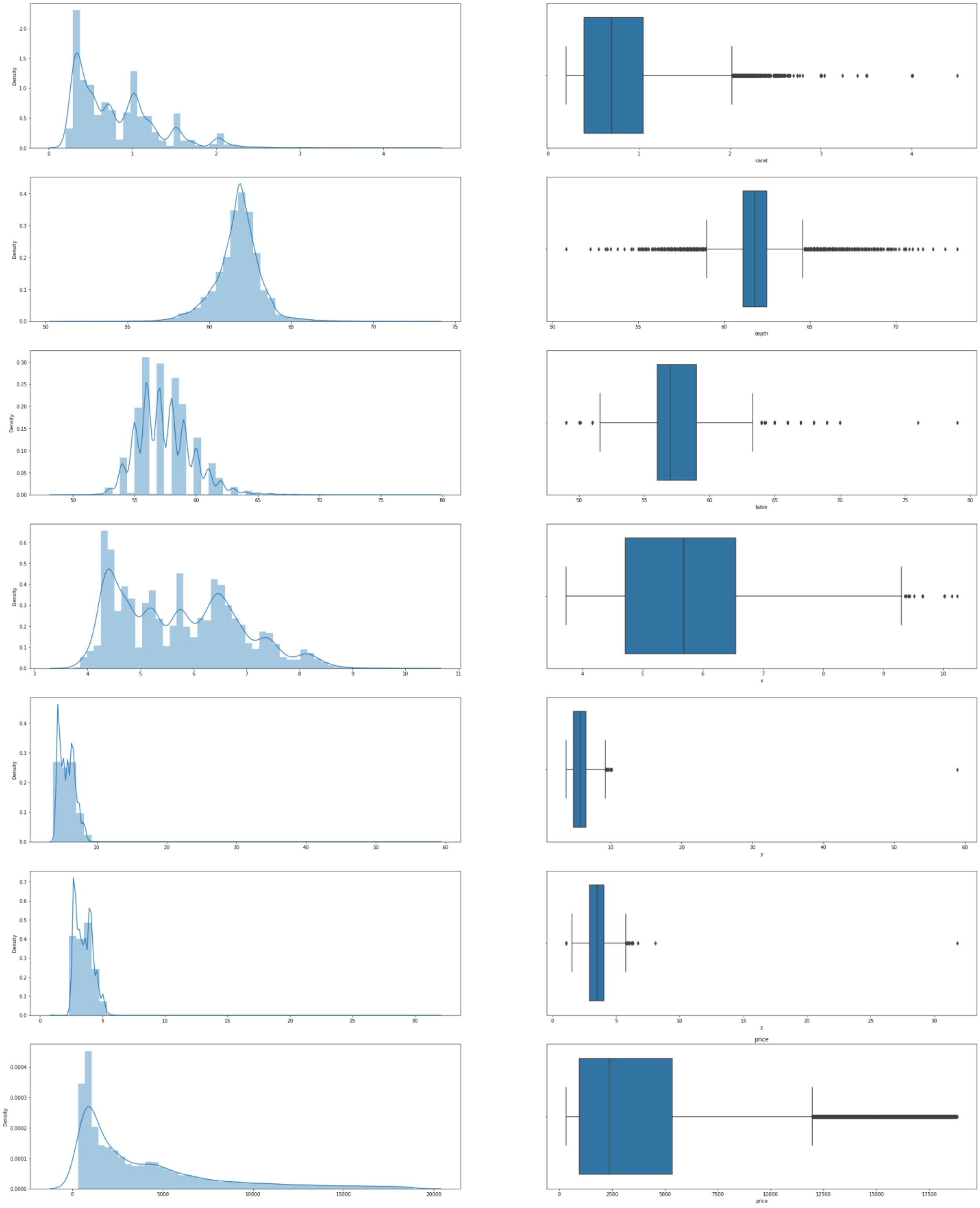
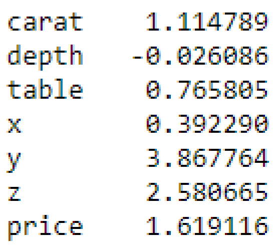
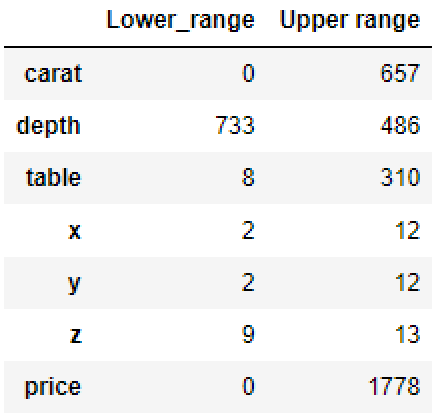


Figure 1 Histogram and boxplot of numerical cols



```
carat     1.114789
depth    -0.026086
table     0.765805
x         0.392290
y         3.867764
z         2.580665
price     1.619116
```

Table 6 Skewness values for numerical cols



	Lower_range	Upper range
carat	0	657
depth	733	486
table	8	310
x	2	12
y	2	12
z	9	13
price	0	1778

Table 7 Number of outliers

OBSERVATIONS:

CARAT: It has multiple peaks in the dataset which denotes possibility of clusters in the data it is right skewed with a value of 1.11 and it has 657 outliers only in the upper range.

DEPTH: It is almost normally distributed which has -0.02 as the skew value which is almost close to 0. It contains 733 outliers in the lower range and 486 outliers in the upper range.

TABLE: It is right skewed; its value is 0.76 which is almost close to 1. It has 8 outliers in the lower range and 310 outliers in the upper range.

X: It has right tail elongated in the box plot and the distribution is widely spread on the right side, it is right skewed and has a value of 0.39. There are 2 outliers in the lower range and 12 outliers in the upper range.

Y: It is highly right skewed that has a value of 3.86 also similar to X it has 2 outliers in the lower range and 12 in the upper range.

Z: It is also highly right skewed with a value of 2.58 and consists of 9 outliers in the lower range and 13 in the upper range.

PRICE: It is right skewed with 1.61 as skew value. It has 1778 outliers only in the upper range.

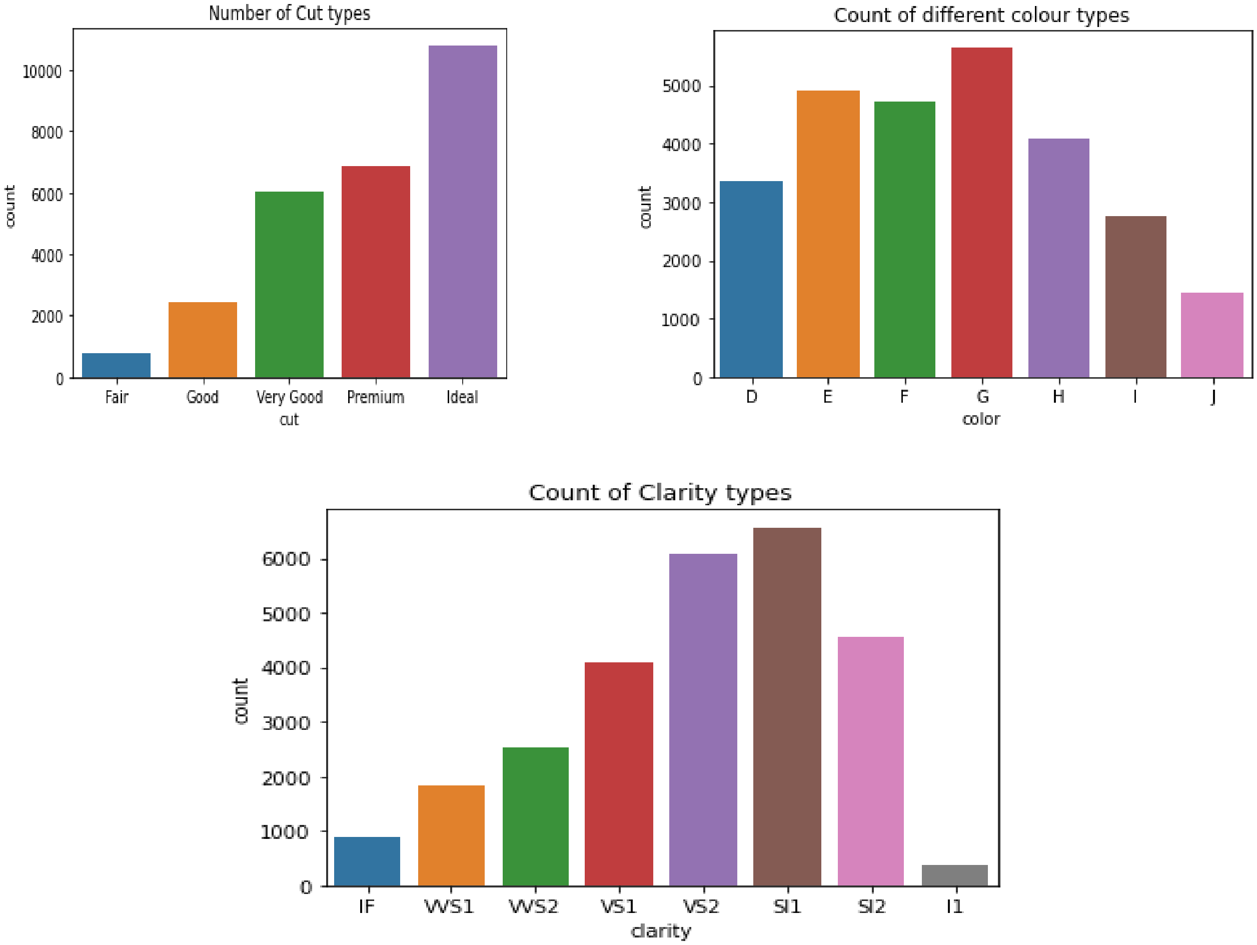


Figure 2 Count of Each levels in all categorical columns

OBSERVATIONS:

- Ideal cut type stones are sold the most and fair is the least number of cut types sold by the company , Premium and very good cut type is averagely sold
- Color G and E are predominantly sold and very less number of J color stones are sold by the company
- In clarity, S1 and VS2 are two types which are predominately sold and I1 is the least of all.

3. BIVARIATE ANALYSIS:

CATEGORICAL:

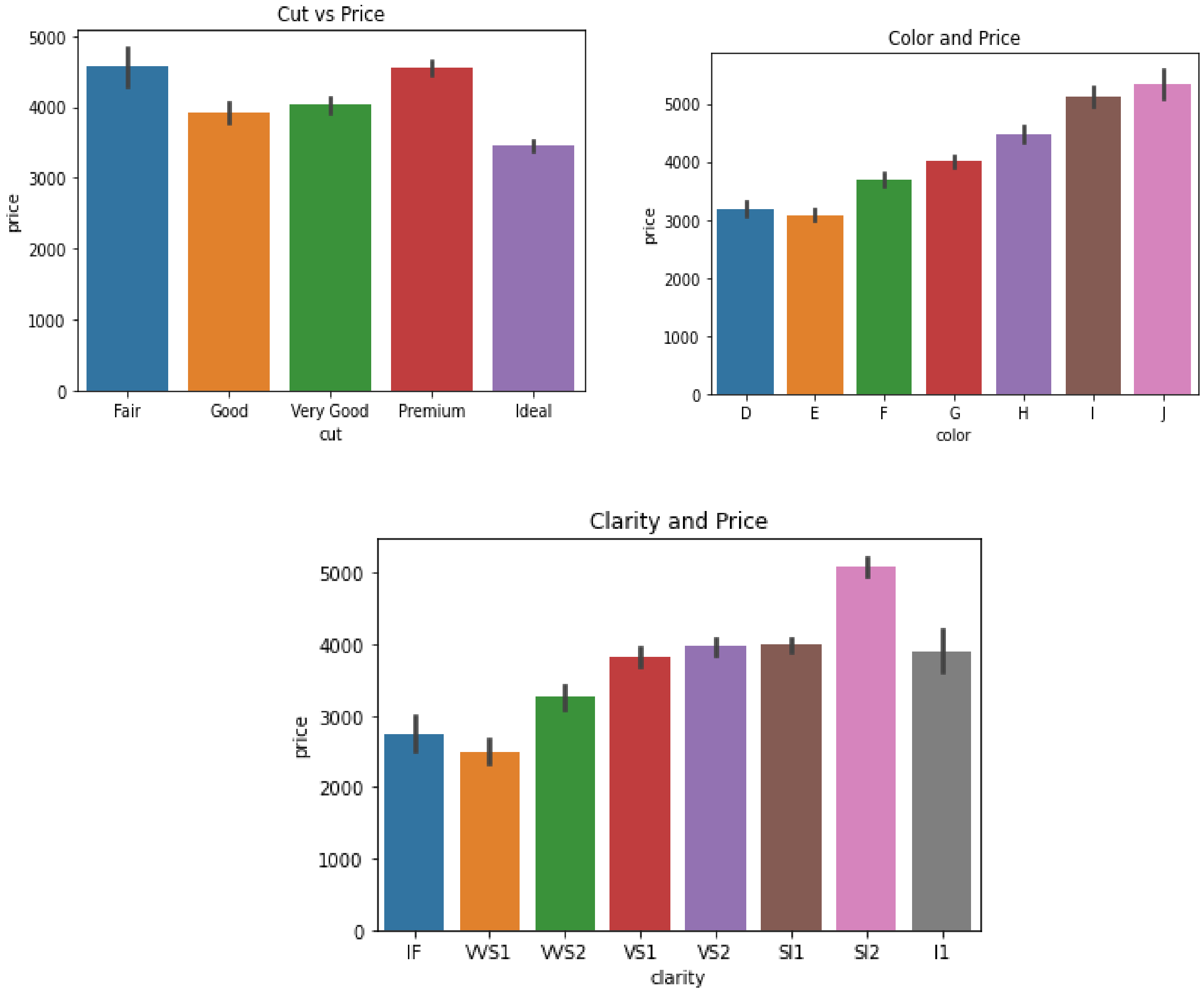


Figure 3 Avg Price vs. Categorical columns

OBSERVATIONS:

- FAIR and PREMIUM are priced very highly, GOOD and VERY GOOD CUT are priced moderately whereas IDEAL is the priced low among other cut types.
- COLOR D,E are priced low and the COLOR I and J are priced high
- In CLARITY SI2 is the costlier one and VVS1 is the lowest of all, IF and VVS2 are more or less the same value, VS1, VS2, SI1 and I1 are moderately priced and all 4 has similar means.

NUMERICAL:

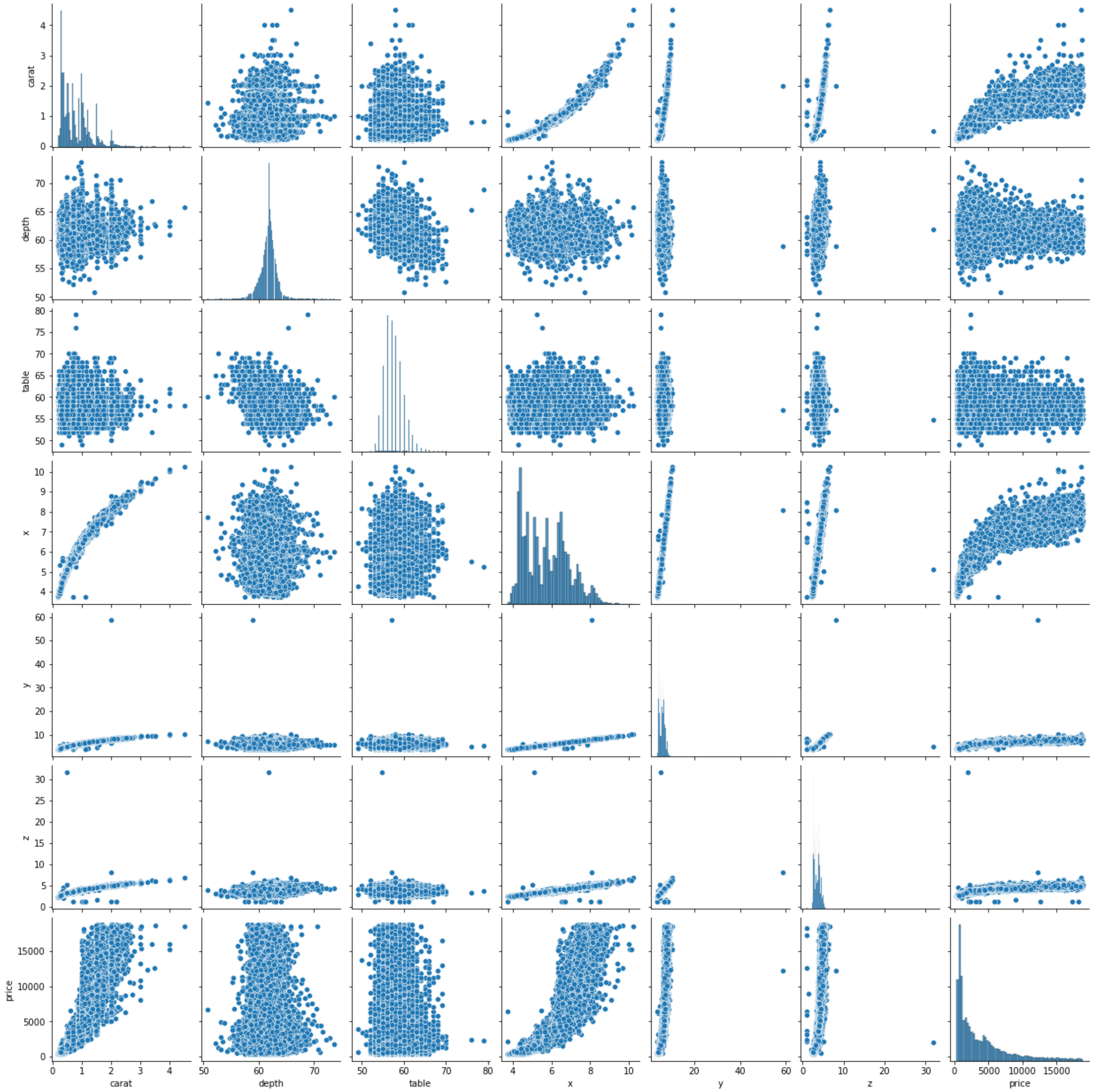


Figure 4 Pair plot of numerical columns

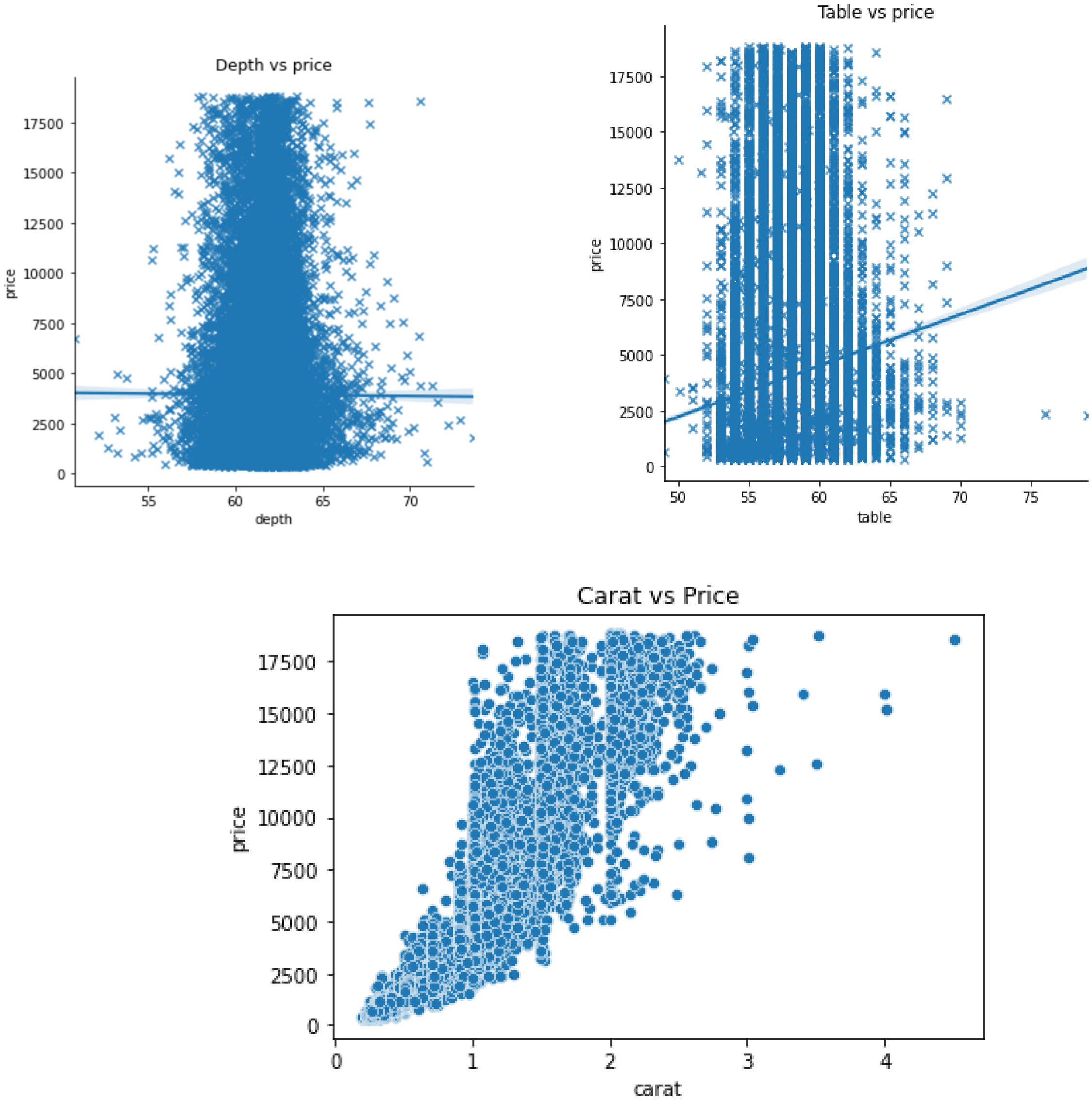


Figure 5 Regression plot for numerical columns

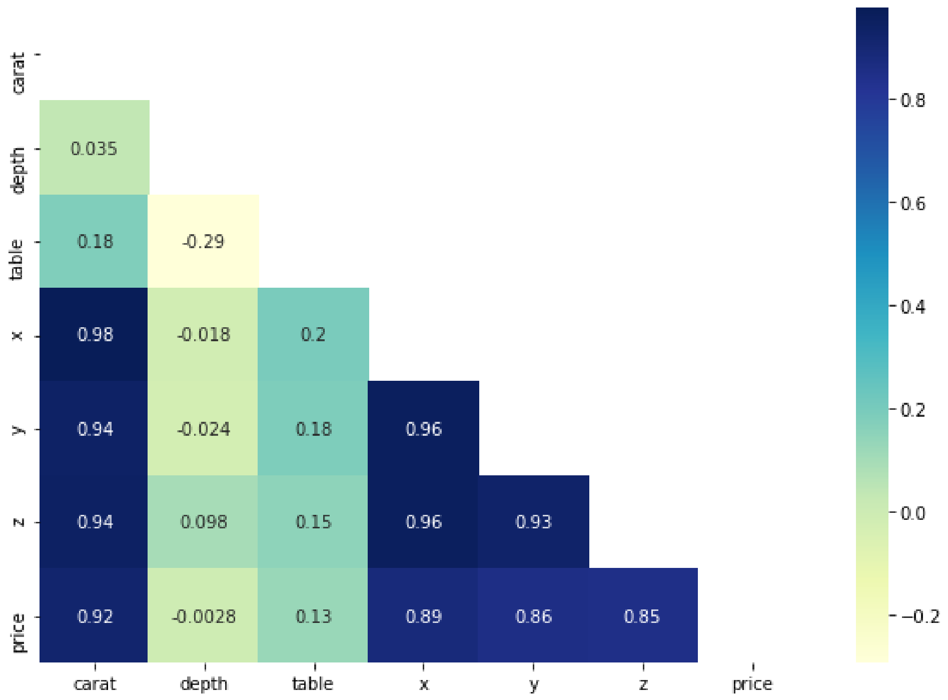


Figure 6 Correlation Heat map

OBSERVATION

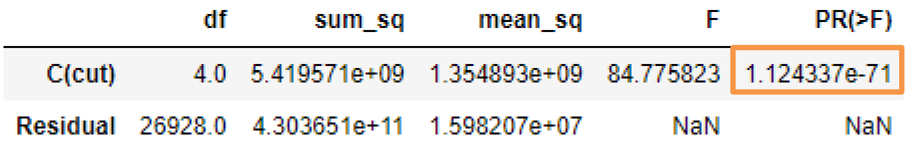
- The pair plot shows few linear relationships between the variables and few cloud like structure which denotes no correlation exists between variables.
- From the pair plot we can see X has a quadratic linear relationship on PRICE and Y, Z has a steep positive relationship on PRICE.
- In the regression plot from the figure 5 we can see DEPTH has no relationship on PRICE, the line is almost a straight line so it could not be good predictor for PRICE.
- From Fig.5 we can also see CARAT and TABLE has a high positive linear relationship on PRICE.
- In the heat map its very evident that there is a high correlation existing between the variables we can see values above 0.8 and 0.9 but DEPTH and TABLE seems to have very low correlation with the other variables since the values are low.
- PRICE is highly correlated with CARAT, X, Y, Z and there is insignificant amount of correlation with DEPTH.

4. ANOVA TEST ON CATEGORICAL COLUMNS

Null hypothesis:

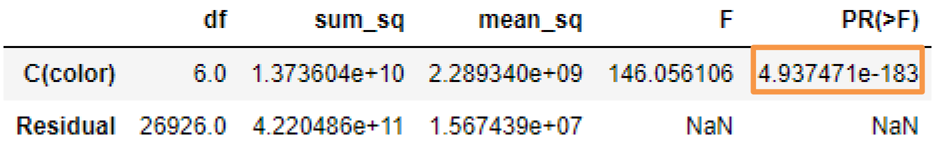
H_o: There is no significant difference in the means within the levels.

$H_{a:}$ At least one of the levels has a significant difference in the mean.



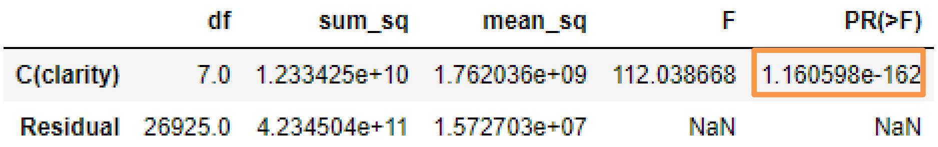
	df	sum_sq	mean_sq	F	PR(>F)
C(cut)	4.0	5.419571e+09	1.354893e+09	84.775823	1.124337e-71
Residual	26928.0	4.303651e+11	1.598207e+07	NaN	NaN

Table 8 Anova test on CUT



	df	sum_sq	mean_sq	F	PR(>F)
C(color)	6.0	1.373604e+10	2.289340e+09	146.056106	4.937471e-183
Residual	26926.0	4.220486e+11	1.567439e+07	NaN	NaN

Table 9 Anova test on COLOR



	df	sum_sq	mean_sq	F	PR(>F)
C(clarity)	7.0	1.233425e+10	1.762036e+09	112.038668	1.160598e-162
Residual	26925.0	4.234504e+11	1.572703e+07	NaN	NaN

Table 10 Anova test on CLARITY

OBSERVATIONS:

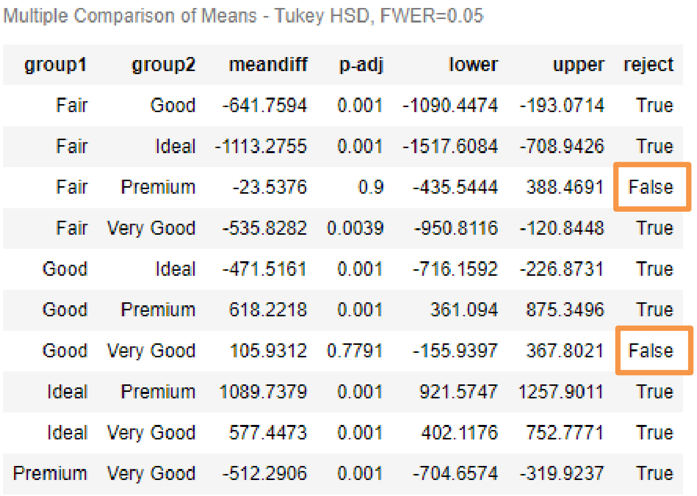
- From the table we can see all 3 p-value is insignificant so we have enough evidence to reject null hypothesis in all 3 cases. Accepting the alternate hypothesis at 95% confident interval we can say there is at least one level with significant difference in all 3 categorical variables.
- Because of the significant difference in the levels all 3 categorical variables can act as a good predictor for PRICE.

5. TUKEY HSD TEST:

Tukey's HSD is a multiple comparison technique that tests the null hypothesis that two means are equal. It tests all pairwise differences while controlling the probability of making one or more Type I errors. It can be as a follow up test done after anova to test the difference in means among the levels in the categorical column.

H_o: Two means are equal or from the same population

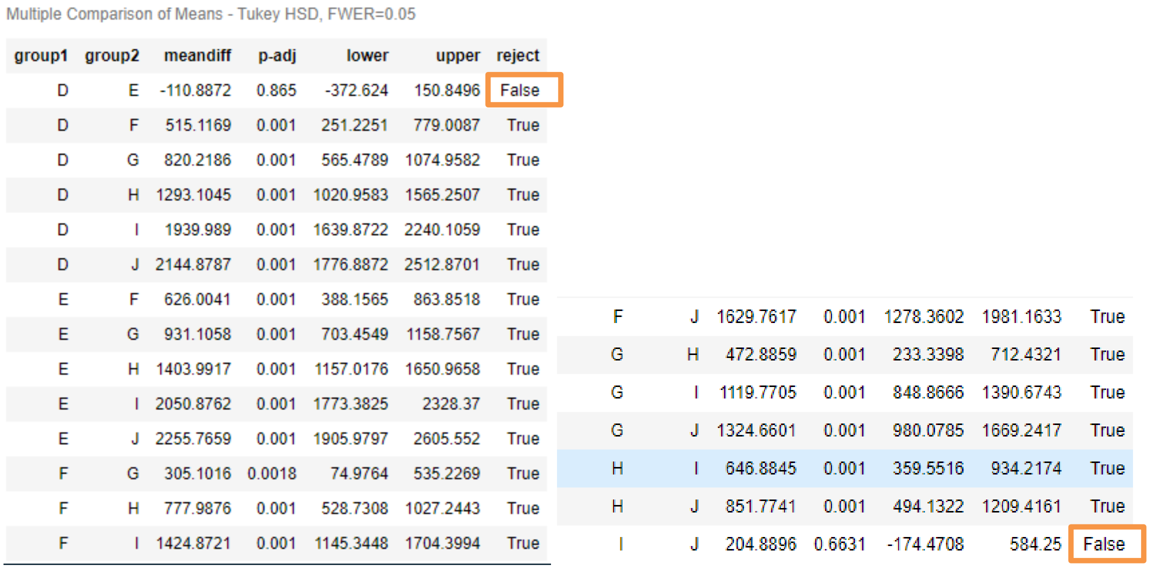
H_A: Two means are not same



Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Fair	Good	-641.7594	0.001	-1090.4474	-193.0714	True
Fair	Ideal	-1113.2755	0.001	-1517.6084	-708.9426	True
Fair	Premium	-23.5376	0.9	-435.5444	388.4691	False
Fair	Very Good	-535.8282	0.0039	-950.8116	-120.8448	True
Good	Ideal	-471.5161	0.001	-716.1592	-226.8731	True
Good	Premium	618.2218	0.001	361.094	875.3496	True
Good	Very Good	105.9312	0.7791	-155.9397	367.8021	False
Ideal	Premium	1089.7379	0.001	921.5747	1257.9011	True
Ideal	Very Good	577.4473	0.001	402.1176	752.7771	True
Premium	Very Good	-512.2906	0.001	-704.6574	-319.9237	True

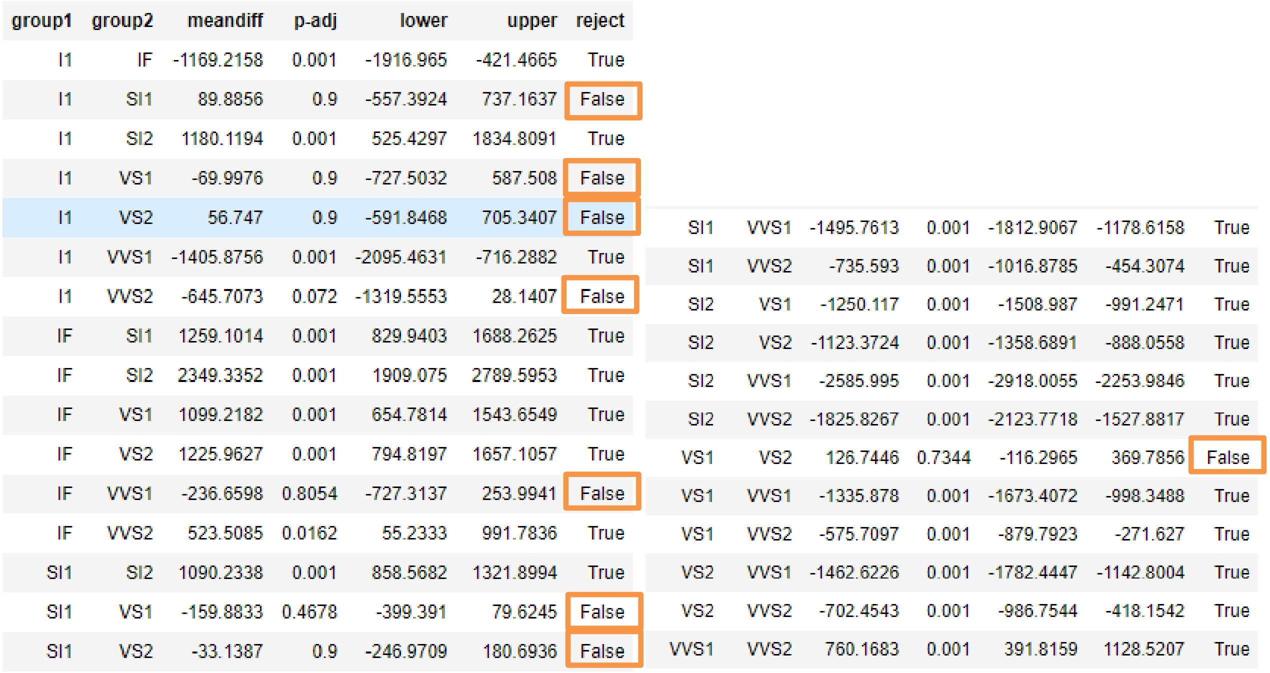
Table 11 Tukey's HSD test for CUT



Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
D	E	-110.8872	0.865	-372.624	150.8496	False
D	F	515.1169	0.001	251.2251	779.0087	True
D	G	820.2186	0.001	565.4789	1074.9582	True
D	H	1293.1045	0.001	1020.9583	1565.2507	True
D	I	1939.989	0.001	1639.8722	2240.1059	True
D	J	2144.8787	0.001	1776.8872	2512.8701	True
E	F	626.0041	0.001	388.1565	863.8518	True
E	G	931.1058	0.001	703.4549	1158.7567	True
E	H	1403.9917	0.001	1157.0176	1650.9658	True
E	I	2050.8762	0.001	1773.3825	2328.37	True
E	J	2255.7659	0.001	1905.9797	2605.552	True
F	G	305.1016	0.0018	74.9764	535.2269	True
F	H	777.9876	0.001	528.7308	1027.2443	True
F	I	1424.8721	0.001	1145.3448	1704.3994	True
F	J	1629.7617	0.001	1278.3602	1981.1633	True
G	H	472.8859	0.001	233.3398	712.4321	True
G	I	1119.7705	0.001	848.8666	1390.6743	True
G	J	1324.6601	0.001	980.0785	1669.2417	True
H	I	646.8845	0.001	359.5516	934.2174	True
H	J	851.7741	0.001	494.1322	1209.4161	True
I	J	204.8896	0.6631	-174.4708	584.25	False

Table 12 Tukey's HSD on COLOR



group1	group2	meandiff	p-adj	lower	upper	reject
I1	IF	-1169.2158	0.001	-1916.965	-421.4665	True
I1	SI1	89.8856	0.9	-557.3924	737.1637	False
I1	SI2	1180.1194	0.001	525.4297	1834.8091	True
I1	VS1	-69.9976	0.9	-727.5032	587.508	False
I1	VS2	56.747	0.9	-591.8468	705.3407	False
I1	VVS1	-1405.8756	0.001	-2095.4631	-716.2882	True
I1	VVS2	-645.7073	0.072	-1319.5553	28.1407	False
IF	SI1	1259.1014	0.001	829.9403	1688.2625	True
IF	SI2	2349.3352	0.001	1909.075	2789.5953	True
IF	VS1	1099.2182	0.001	654.7814	1543.6549	True
IF	VS2	1225.9627	0.001	794.8197	1657.1057	True
IF	VVS1	-236.6598	0.8054	-727.3137	253.9941	False
IF	VVS2	523.5085	0.0162	55.2333	991.7836	True
SI1	SI2	1090.2338	0.001	858.5682	1321.8994	True
SI1	VS1	-159.8833	0.4678	-399.391	79.6245	False
SI1	VS2	-33.1387	0.9	-246.9709	180.6936	False
SI1	VVS1	-1495.7613	0.001	-1812.9067	-1178.6158	True
SI1	VVS2	-735.593	0.001	-1016.8785	-454.3074	True
SI2	VS1	-1250.117	0.001	-1508.987	-991.2471	True
SI2	VS2	-1123.3724	0.001	-1358.6891	-888.0558	True
SI2	VVS1	-2585.995	0.001	-2918.0055	-2253.9846	True
SI2	VVS2	-1825.8267	0.001	-2123.7718	-1527.8817	True
VS1	VS2	126.7446	0.7344	-116.2965	369.7856	False
VS1	VVS1	-1335.878	0.001	-1673.4072	-998.3488	True
VS1	VVS2	-575.7097	0.001	-879.7923	-271.627	True
VS2	VVS1	-1462.6226	0.001	-1782.4447	-1142.8004	True
VS2	VVS2	-702.4543	0.001	-986.7544	-418.1542	True
VVS1	VVS2	760.1683	0.001	391.8159	1128.5207	True

Table 13 Tukey's HSD on CLARITY

The column ‘reject’ in the above tables denote whether we must reject the null hypothesis or not. Rejecting the null hypothesis means two groups have significant difference, accepting the null hypothesis will denote two groups does not have any significant difference. Using the Tukey’s HSD test results we can group the levels with no significant difference together and encode them to use it in the model this might reduce the levels in the categorical column and also increase the model’s performance.

OBSERVATIONS:

- For the column CUT , FAIR- PREMIUM and GOOD-VERY GOOD have the value false indicating these two groups has same means so these two can be encoded as one level.
- For the column COLOR , D-E and I-J have similar means so these two levels can be grouped together
- For the column CLARITY, VS1-VS2-SI1, IF-VVS1, I1-VVS1-VVS2, I1-VS1-VS2 are similar in nature.

6. DATA ENCODING:

The categorical columns are ranked in nature, that is it indicates the levels in the increasing order of quality for all three variables namely, Fair is the lowest and Ideal is the best for column CUT , D is the lowest and J is the best for column COLOR , IF is the lowest and I1 is the best for CLARITY. So we encode the above columns in ordinal manner, the encoded labels are shown below:

CUT: Fair':1, 'Good':2, 'Very Good':3, 'Premium':4, 'Ideal':5

COLOR: D':1,'E':2,'F':3,'G':4,'H':5,'I':6,'J':7

CLARITY: 'IF':1, 'VVS1':2, 'VVS2':3, 'VS1':4, 'VS2':5, 'SI1':6, 'SI2':7, 'I1':8

GROUPING LEVELS USING RESULTS OF TUKEY'S HSD TEST:

To improve the model's performance and reduce the complexity in the dataset we are grouping multiple levels together. With the help of Tukey's HSD test we found out samples with similar means so with the applying business knowledge along with that we have come up with the below strategy to combine the levels

CUT: 'Fair':3, 'Good':2, 'Very Good':2, 'Premium':3, 'Ideal':4

COLOR: 'D':2,'E':2,'F':3,'G':4,'H':5,'I':6,'J':6

CLARITY: 'IF':2, 'VVS1':2, 'VVS2':3, 'VS1':4, 'VS2':4, 'SI1':5, 'SI2':6, 'I1':7

As the test result shows fair -Premium, Good-Very Good are grouped together because of no significant difference in the mean between Fair and Premium and Good and Very Good are more or less means the same.

In the case of color, D-E and I-J are opposite extremes, D-E being the lowest quality and I-J being the best quality; both D-E and I-J are very close in terms of means so they are grouped together.

Considering the column CLARITY even though VVS1 and VVS2 seems like similar group from the business point of view VVS1 resembles to IF in most of the properties so these two are clubbed and from the tukey's test VS1-VS2 are grouped together.

By grouping the levels we have reduced the levels in CUT column from 5 to 3, COLOR column from 7 to 5 and CLARITY column from 8 to 6.

7. MODEL CREATION

Model 1: Sk-lean Liner Regression

First model was created with dataset without treating the outlier and grouping the categorical variables and with default parameters for the model.

We found the intercept to be: `9162.17`

And the coefficients to be

```
The coefficient for CARAT is 10908.09613895233
The coefficient for CUT is 118.28587226850081
The coefficient for COLOR is -330.68549397114054
The coefficient for CLARITY is -498.2101889831947
The coefficient for DEPTH is -61.73473693020163
The coefficient for TABLE is -28.690141461533173
The coefficient for X is -727.8788156065789
The coefficient for Y is 36.65699782063721
The coefficient for Z is -373.15701930485807
```

The R2 value for train and test was found to be: 0.9083, 0.9079
The MSE value for train and test was found to be: 1212.322, 1232.630

Model 2: Sk-learn model

With outliers treated in the dataset without grouping the levels one more models is created with default parameters.

We found the intercept to be: 1158.64

And the coefficients to be:

```
The coefficient for carat is 8679.681971437753
The coefficient for cut is 115.5298092012355
The coefficient for color is -276.6788906926656
The coefficient for clarity is -436.261021175506
The coefficient for depth is 10.636344960638267
The coefficient for table is -9.046832728371854
The coefficient for x is -1401.2721602058302
The coefficient for y is 1417.0494715730729
The coefficient for z is -519.1005907004283
```

The R2 value for train and test was found to be: 0.9312, 0.9310

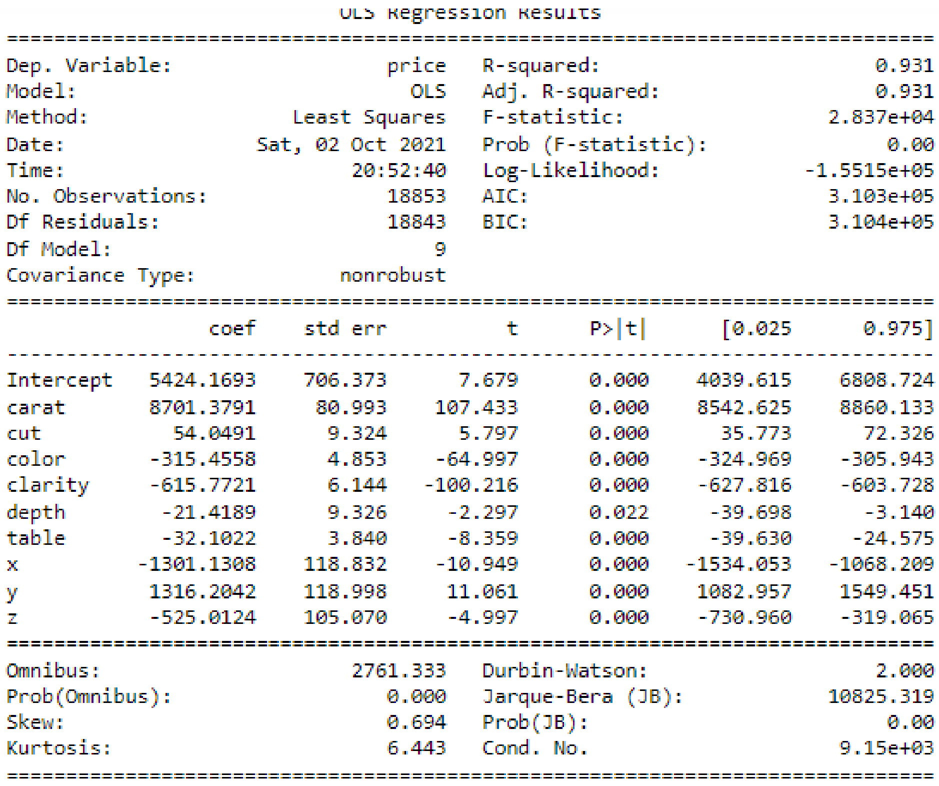
The MSE value for train and test was found to be: 907.8137, 914.6615

OBSERVATION:

We can see that there is a drop in the MSE values and the increase in R^2 value compared to the first model so outlier treatment has improved the model's performance.

Model 3: OLS model

With dataset treated with outlier and grouping the categorical column as mentioned above ols model was created.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 2.837e+04
Date:                Sat, 02 Oct 2021   Prob (F-statistic):               0.00
Time:                        20:52:40   Log-Likelihood:            -1.5515e+05
No. Observations:               18853   AIC:                         3.103e+05
Df Residuals:                   18843   BIC:                         3.104e+05
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5424.1693    706.373      7.679      0.000    4039.615    6808.724
carat       8701.3791     80.993    107.433      0.000    8542.625    8860.133
cut           54.0491      9.324      5.797      0.000      35.773      72.326
color       -315.4558      4.853    -64.997      0.000    -324.969    -305.943
clarity     -615.7721      6.144   -100.216      0.000    -627.816    -603.728
depth        -21.4189      9.326     -2.297      0.022     -39.698      -3.140
table        -32.1022      3.840     -8.359      0.000     -39.630     -24.575
x          -1301.1308    118.832    -10.949      0.000   -1534.053   -1068.209
y           1316.2042    118.998     11.061      0.000    1082.957    1549.451
z           -525.0124    105.070     -4.997      0.000    -730.960    -319.065
==============================================================================
Omnibus:                     2761.333   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10825.319
Skew:                           0.694   Prob(JB):                         0.00
Kurtosis:                       6.443   Cond. No.                     9.15e+03
==============================================================================
```

Table 14 Summary of ols model

From the summary above we can see the adjusted R^2 value to be 0.931 and the MSE for train and test was found to be 907.5213, 910.008

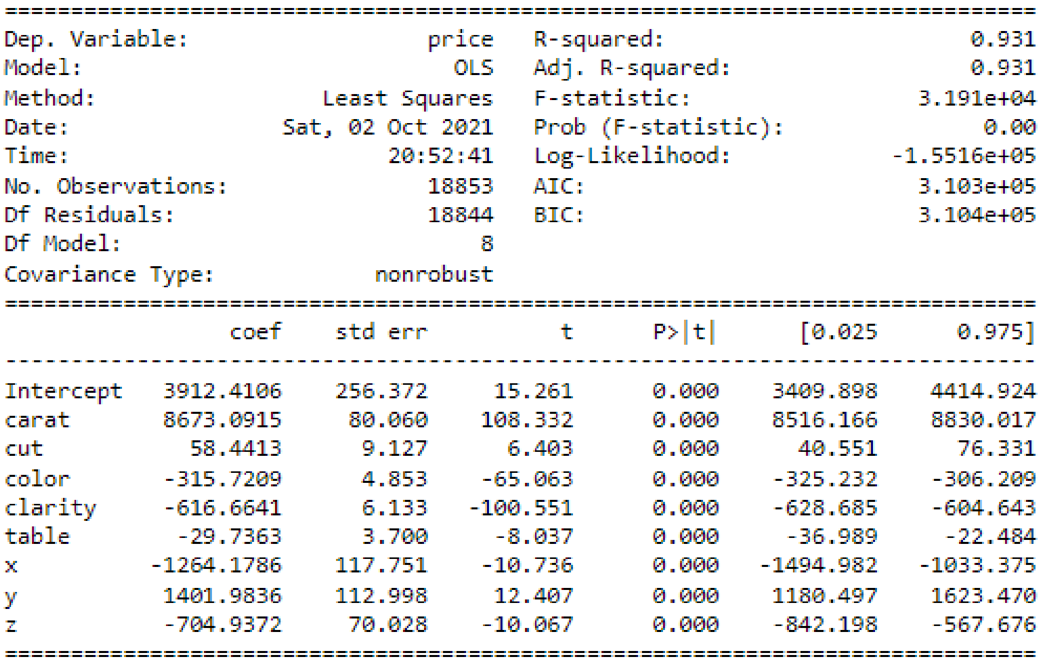
The intercept is 5424.1693 and the coefficients are to be

```
carat         8701.379059
cut             54.049103
color         -315.455754
clarity       -615.772129
depth          -21.418942
table          -32.102233
x            -1301.130848
y             1316.204192
z             -525.012425
```

We can see that the summary is showing p values for all the columns, the p value gives us the probability to test the hypothesis given the population of respective means what is the probability that the coefficients can be zero. For DEPTH variable we can see 0.022 as the p value at 98% confidence interval we can say the column DEPTH is insignificant in predicting the PRICE so we can remove it from the model.

Model 4: OLS model

Dataset without outlier and grouped categorical columns, excluding DEPTH column

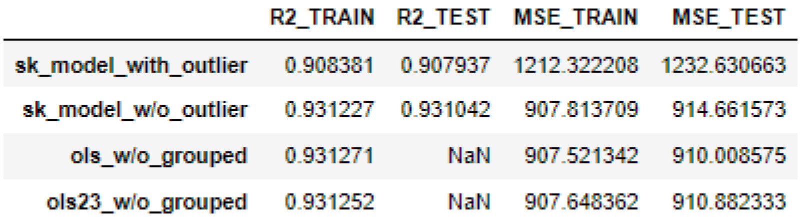


```
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 3.191e+04
Date:                Sat, 02 Oct 2021   Prob (F-statistic):               0.00
Time:                        20:52:41   Log-Likelihood:            -1.5516e+05
No. Observations:               18853   AIC:                         3.103e+05
Df Residuals:                   18844   BIC:                         3.104e+05
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3912.4106    256.372     15.261      0.000    3409.898    4414.924
carat       8673.0915     80.060    108.332      0.000    8516.166    8830.017
cut           58.4413      9.127      6.403      0.000      40.551      76.331
color       -315.7209      4.853    -65.063      0.000    -325.232    -306.209
clarity     -616.6641      6.133   -100.551      0.000    -628.685    -604.643
table        -29.7363      3.700     -8.037      0.000     -36.989     -22.484
x          -1264.1786    117.751    -10.736      0.000   -1494.982   -1033.375
y           1401.9836    112.998     12.407      0.000    1180.497    1623.470
z           -704.9372     70.028    -10.067      0.000    -842.198    -567.676
==============================================================================
```

Table 15 Summary of ols

The MSE for train and test are: 907.6483, 910.8823

```
Intercept    3912.410618
carat        8673.091483
cut            58.441270
color        -315.720880
clarity      -616.664088
table         -29.736333
x           -1264.178577
y            1401.983555
z            -704.937175
```



	R2_TRAIN	R2_TEST	MSE_TRAIN	MSE_TEST
sk_model_with_outlier	0.908381	0.907937	1212.322208	1232.630663
sk_model_w/o_outlier	0.931227	0.931042	907.813709	914.661573
ols_w/o_grouped	0.931271	NaN	907.521342	910.008575
ols23_w/o_grouped	0.931252	NaN	907.648362	910.882333

Table 16 Summary of the models

From the ols summary we can see that adjusted R^2 hasn't changed so removing DEPTH has not affected the model and only the intercept has changed. MSE value also hasn't changed a lot so we can finalize this model for our problem statement.

Table 16 shows evidently that model3 (ols23_w/o) and model 4 (ols23_w/o_grouped) works better comparing the MSE scores. Both the models have high R2 values and the least MSE value for test data but when the models is put to predict unknown data or put to production model 4 is recommended when compared to model 3. As it is free from the DEPTH variable it could generalize on the unseen data and make accurate predictions avoiding the unwanted random correlation from DEPTH variable from interfering the predictions.

The final prediction for price would be

PRICE= `3912.410 +8673.091*CARAT +58.441*CUT -315.720*COLOR -616.664*CLARITY`

`-29.736*TABLE -1264.178*X+ 1401.983*Y -704.937*Z`

CHECKING FOR ASSUMPTIONS:

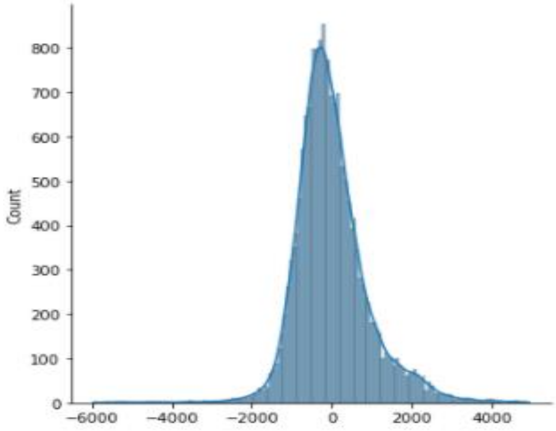


Figure 7 Histogram of Residuals

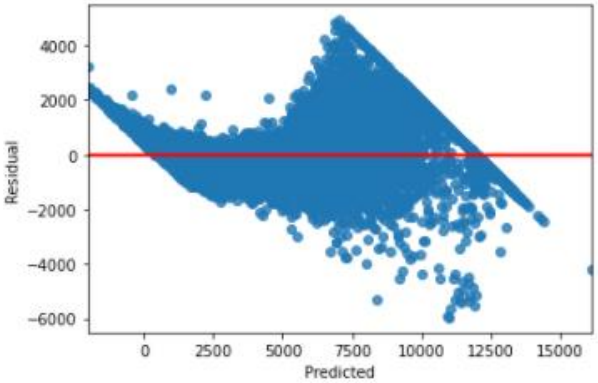


Figure 8 Residual vs. Predicted

The model passes two assumptions , the mean of residual has to be 0 ,the value we got is 9.62e-11 which is close to 0.Second assumption :the residuals must be normally distributed we can see that its normally distributed from the figure 7. The third assumption check for homoscedasticity fails as the figure 8 evidently shows a linear relationship between residual and predicted this condition is called heteroskedasticity.

SUMMARY:

Once we started to work with the data, we came to know it needed few cleaning process. There were missing values and 0's present in the dataset.0's in the dataset couldn't make any logical sense because the column referring to the physical dimension of the cubic zirconia were containing 0s.After imputing the missing values and changing all the 0's to minimum value we found some duplicated records were present in the dataset. Once all the cleaning is done outliers were detected so we imputed 5% 95% quantiles values respectively. Linear regression won't be able to handle outliers it will impair the model's performance that's why we have treated the outlier also this can be seen evidently from the initial model's performance. Once the data is cleaned Anova test was performed on the categorical variable to find out whether they have significant difference in the means so that it could act as good predictor. Performing the anova we then succeeded with Tukey's HSD test to find the groups with significant difference. After finding the groups EDA was performed

From the EDA we have the following insights on the business:

- Price of Ideal cut type stone is less compared to Fair, Good and Premium but the business is selling more number of Ideal cut types than the other categories
- Color G and E are the most sold stones and I and J are least sold stones ,where I and J are the expensive stones and G and E are the moderate ones.
- VS2 and S1 are the most sold clarity types whereas the costliest one is S2
- Price is exponentially related to carat
- Depth has no significant relation on the price.
- X, Y AND Z has a linear relationship on the price.

Once finding these insights all the categorical variables are encoded as there are ordinal data, each level is ranked as mentioned by the business requirement. After encoding the data various linear regression models were tried and finally came to conclusion that the model with following parameters is the best by using the measure of R^2 and MSE

PRICE= 3912.410 +8673.091*CARAT +58.441*CUT -315.720*COLOR -616.664*CLARITY

-29.736*TABLE -1264.178*X+ 1401.983*Y -704.937*Z

Best predictors would be CARAT, COLOR, CLARITY, X, Y, Z, as these variables with higher coefficients.

RECOMMENDATIONS:

- We should try to come with a survey to understand why Fair and Good cut types are not preferred by the customers, following the survey we could come with some price alteration in these categories to hike the sales number because there are expensive and we have a high profit margin in these types of stones.
- Convincing marketing is needed to push the sales of I and J color types which are the costlier ones.
- As we have more sales in G and E color types increasing the price by 5% will increase the profit
- Proper survey should be conducted to understand why customers prefer VS2 & S2 when I1 is available. If it's because of the pricing then I1 must be reduced or if it's because of any other reason it must be found and try to push he customers to buy I1 quality.
- Sales people must be educated enough to convince the customers to buy Fair cut type, including any offer will also do.

CLASSIFICATION FOR HOLIDAY PACKAGE

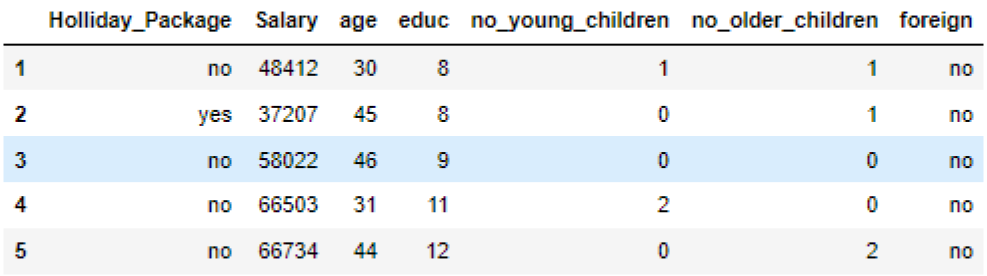
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not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

1. DATA DESCRIPTION

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

Table 17 Data dictionary

2. DATA SAMPLE:

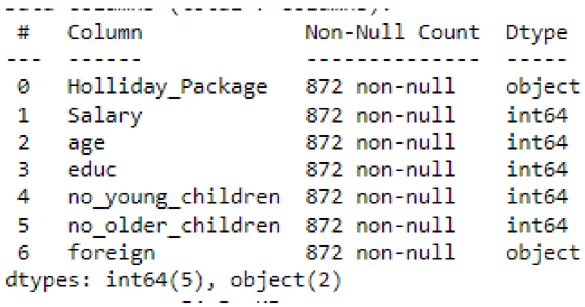


	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
1	no	48412	30	8	1	1	no
2	yes	37207	45	8	0	1	no
3	no	58022	46	9	0	0	no
4	no	66503	31	11	2	0	no
5	no	66734	44	12	0	2	no

Table 18 Sample of the data

3. EXPLORATORY DATA ANALYSIS:

1. ATTRIBUTES OF THE DATA:



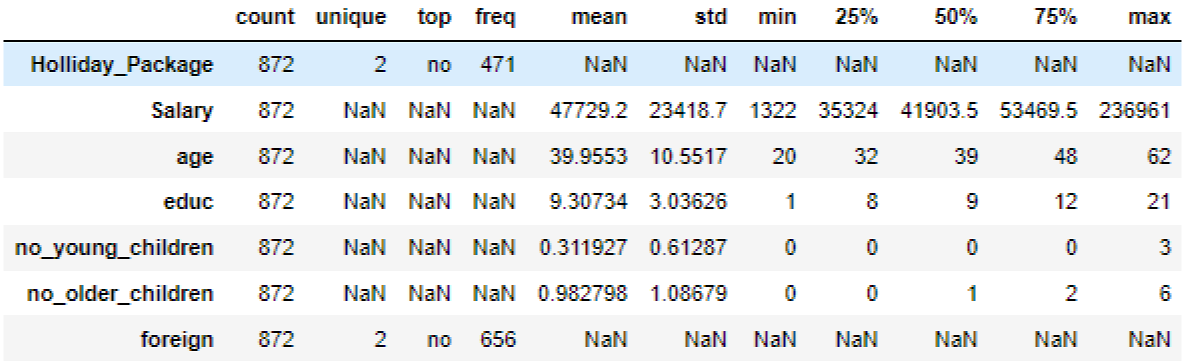
```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
```

Table 19 Information of the data

The dataset has 872 rows and 7 columns including the dependent variable from the above table we can infer that there are 3 numerical columns and 4 categorical columns. No missing values present in the dataset

Numerical columns: SALARY, AGE, EDUCATION

Categorical columns: FOREIGN, NO_YOUNG_CHILDREN, NO_OLDER_CHILDREN, HOLIDAY_PACKAGE (TARGET)



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872	NaN	NaN	NaN	47729.2	23418.7	1322	35324	41903.5	53469.5	236961
age	872	NaN	NaN	NaN	39.9553	10.5517	20	32	39	48	62
educ	872	NaN	NaN	NaN	9.30734	3.03626	1	8	9	12	21
no_young_children	872	NaN	NaN	NaN	0.311927	0.61287	0	0	0	0	3
no_older_children	872	NaN	NaN	NaN	0.982798	1.08679	0	0	1	2	6
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 20 Summary of the data

OBSERVATIONS:

SALARY: It has a mean of47729.2 and standard deviation of 23418.7.It has a huge spread which is normal in the case of salary ,50% of the data is above 41903.5 ,minimum value is 1322 and the maximum value is 236961

AGE: It is averaged at 39.95 with a standard deviation of 10.55 minimum value is 20 ,50% the data is above 39 and the maximum value is 62

EDUCATION: It has a mean value of 9.30 and standard deviation of 3.03 which means the spread is less and its closely placed to each other. 50% of the data is above 9 minimum is 1 and maximum is 12.

NO_YOUNG_CHILDREN: It has 4 unique values (0, 1, 2, and 3) and mode as 0

NO_OLD_CHILDREN: It has 7 unique values (0 to 6) and a mode of 0.

FOREIGN: It's a Boolean column with true/ false which has mode of False

Key Findings:

- It has no missing values or wrong entries in the dataset
- No duplicated records present In the dataset

2. UNIVARIATE ANALYSIS:

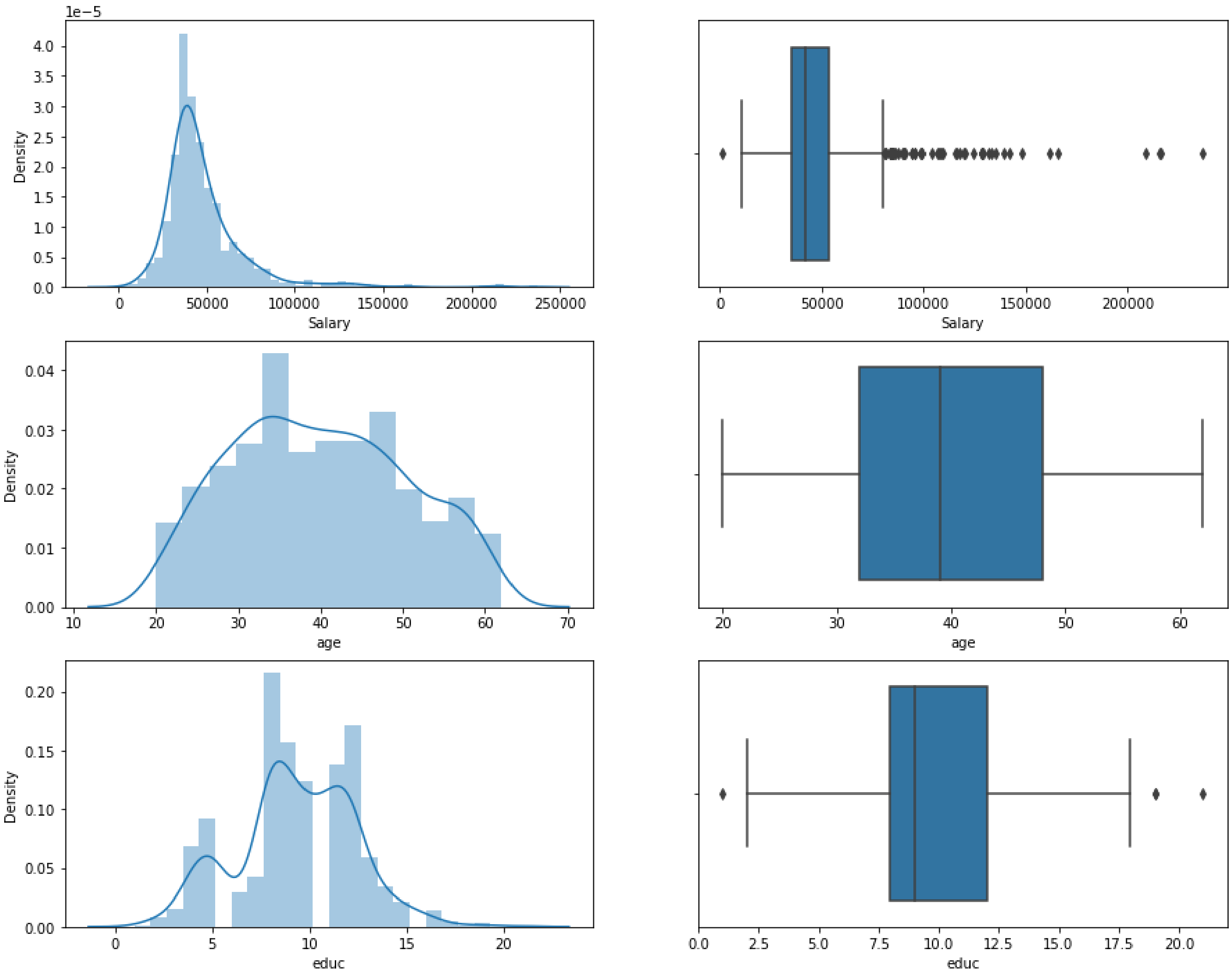
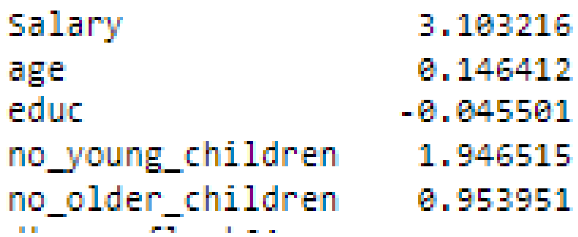
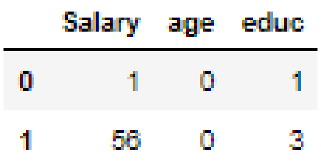


Figure 9 Histogram and Boxplot



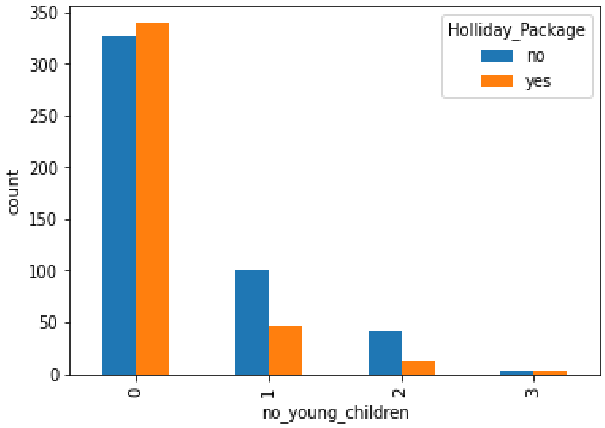
```
Salary                 3.103216
age                    0.146412
educ                  -0.045501
no_young_children      1.946515
no_older_children      0.953951
```

Figure 10 Skewness



	Salary	age	educ
0	1	0	1
1	56	0	3

Figure 11 Outliers



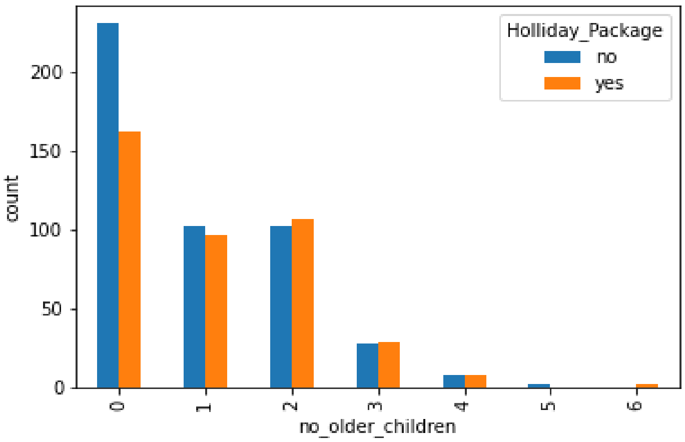


Figure 12 Count of young children

Figure 13 Count of older children

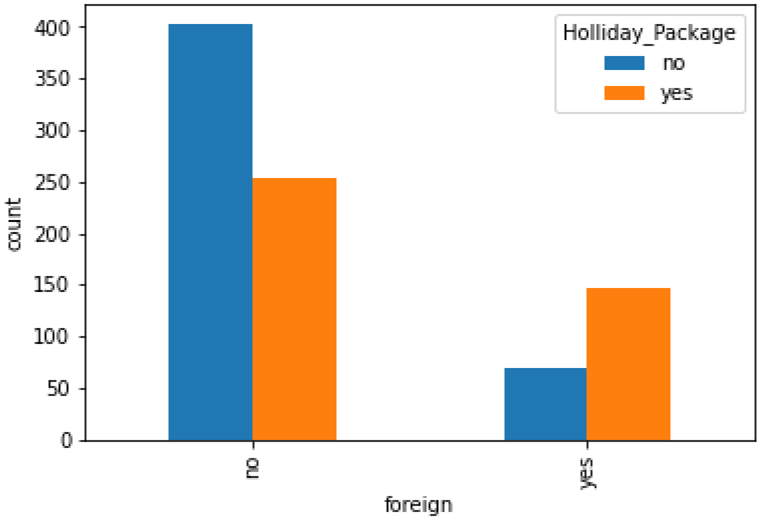


Figure 14 Count of foreign

OBSERVATIONS:

SALARY: It is normally distributed but right skewed have 56 outliers in the above range which is normal in variables like salary

AGE: It almost normally distributed with no skewness and no outliers in it.

EDUCATION: It's not normally distributed and shows partial left skewness, 4 outliers present 1 in the lower range and 3 in upper range.

The ratio of people with 0 young children taking a holiday package is more compared to other people with more children

People who are having 2 older children are more likely to take holiday package.

Foreigners are more likely to take holiday package

3. BIVARIATE ANALYISIS:

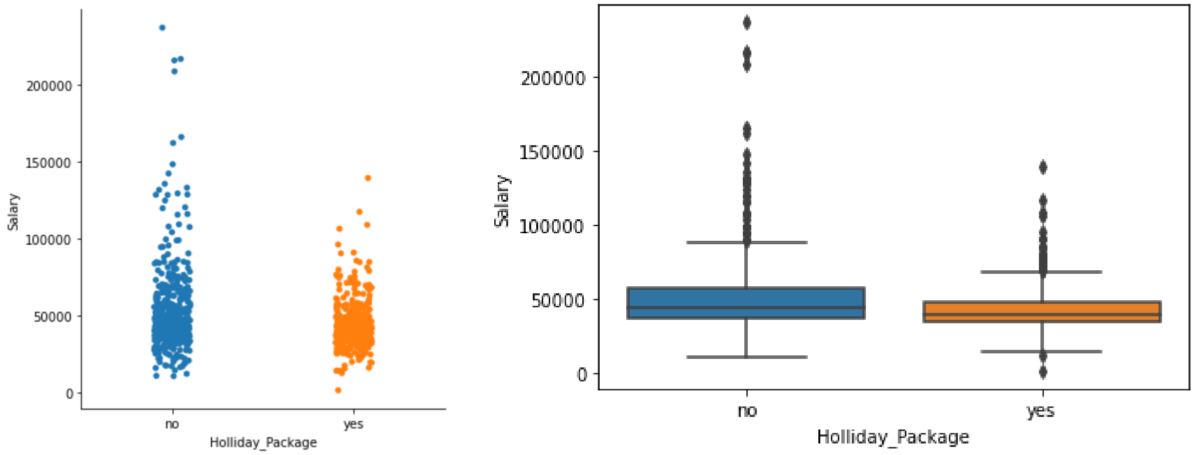


Figure 15 Salary vs. Holiday Package

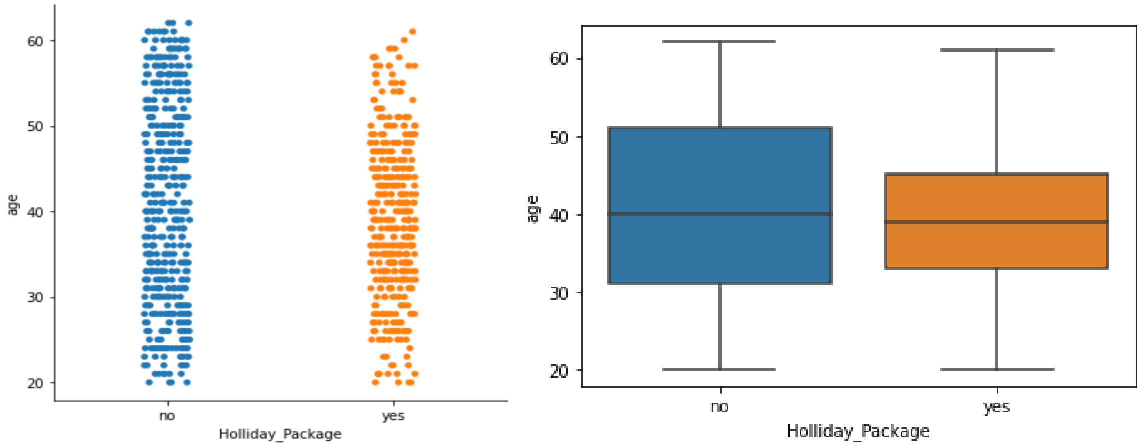


Figure 16 Age vs. Holiday Package

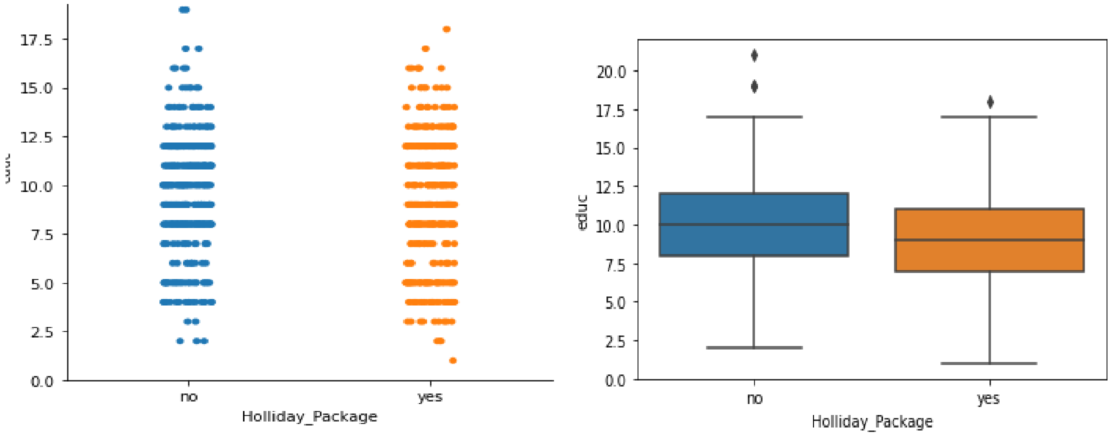


Figure 17 Education vs. Holiday Package

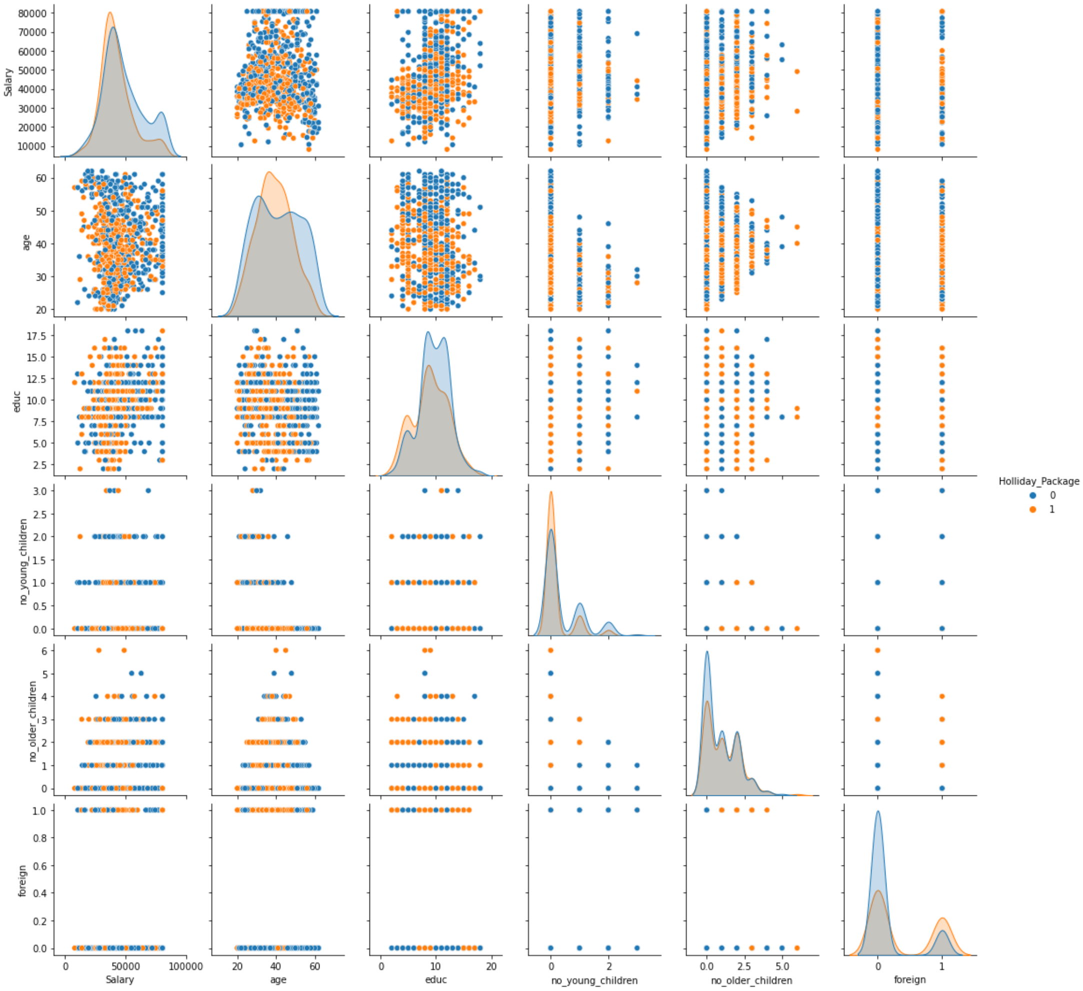


Figure 18 Pair plot for 2 dataset

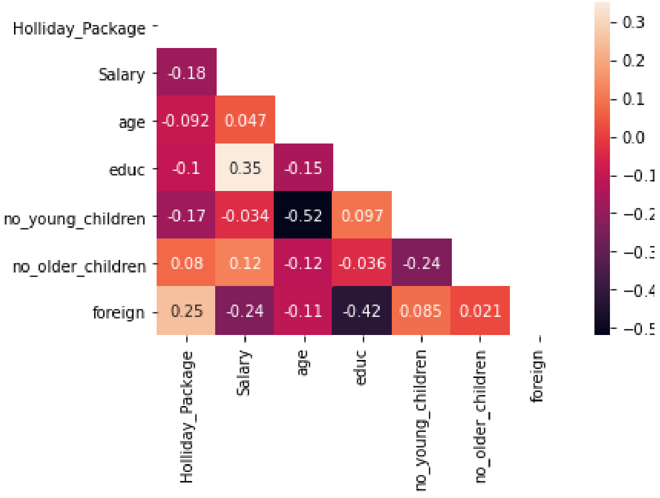


Figure 19 Heat map for Correlation

OBSERVATIONS:

- People who accept the holiday package are in the lower range than who doesn't accept the package, which is odd.
- People within the age 35 to 45 are the ones who accept the holiday package.
- Median of the people's education accepting the holiday package is less it means people with low education value are in favor of accepting the package.
- From the pair plot we can see cloud structure so there is very weak correlation present with the variables. Also, the diagonals show the distribution in which for all the predictors the classes are overlapping so all the predictors might be very weak when it's used for predictions.
- From the heat map we can conclude there is a very weak correlation most of the value ranges within 0.1 to 0.25. Only one or few moderate correlation is found i.e. Between age and young children

4. CHI-SQUARE TEST

Chi square test is done to find whether the categorical column has significant influence on the target variable so that we can choose them for model building.

H_0: Categorical column is independent of Target Variable

H_a: Categorical column is dependent on Target Variable

The P values after performing chi square test on 3 categorical columns are given below.

NO_YOUNG_CHILDREN: P_VALUE= 1.82e-06

NO_OLDER_CHILDREN: P_VALUE= 0.0933

FOREIGN: P_VALUE= 6.22e-14

In the case of NO_YOUNG_CHILDREN and FOREIGN the p value is less than 0.05 so we can reject the null hypothesis and accept the alternative hypothesis which indicates both the variables affect the target variable.

In the case of NO_OLDER_CHILDREN the p value is greater than 0.05 so we fail to reject the null hypothesis and therefore we can conclude that NO_OLDER_CHILDREN is not dependent on target variable.

5. DATA ENCODING:

Out of 3 categorical columns only 2 are significant we are using One Hot Encoding method to encode FOREIGN and NO_YOUNG_CHILDREN

This technique will check the particular level and where ever it's true will return 1 else 0. Likewise each column will be created for each level in the categorical columns.

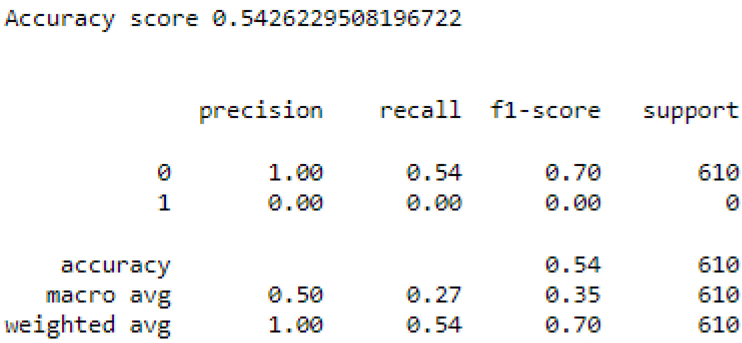
6. MODEL DEVELOPMENT:

Model 1: Logistic Classifier without any specific parameter

Coefficients are:

```
The coefficient of SALARY is: -5.632273118007217e-06
The coefficient of AGE is: 1.619535162982597e-09
The coefficient of EDUC is: 1.3387090625369774e-09
The coefficient of FOREIGN_1 is: 8.565227500234058e-10
The coefficient of NO_YOUNG_CHILDREN_1 is: -3.564461620526026e-10
The coefficient of NO_YOUNG_CHILDREN_2 is: -1.9922550002143942e-10
The coefficient of NO_YOUNG_CHILDREN_3 is: 5.070907506976367e-12
The intercept of the model is 2.20508424e
```

For Train data:



```
Accuracy score 0.5426229508196722


              precision    recall  f1-score   support

           0       1.00      0.54      0.70       610
           1       0.00      0.00      0.00         0

    accuracy                           0.54       610
   macro avg       0.50      0.27      0.35       610
weighted avg       1.00      0.54      0.70       610
```

Figure 20 Classification Report on Train data

AUC: 0.591

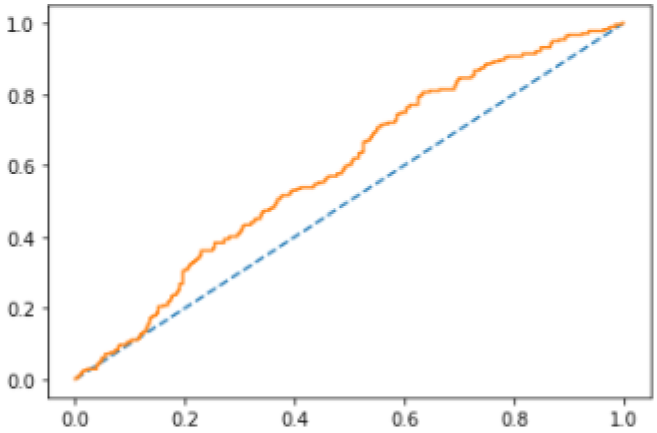
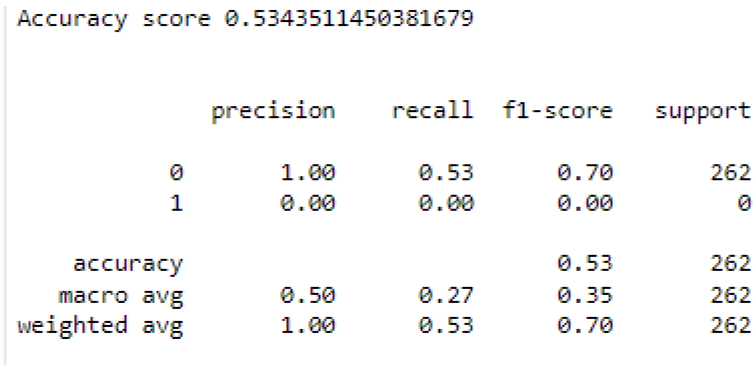


Figure 21 ROC curve

For test data:



```
Accuracy score 0.5343511450381679


              precision    recall  f1-score   support

           0       1.00      0.53      0.70       262
           1       0.00      0.00      0.00         0

    accuracy                           0.53       262
   macro avg       0.50      0.27      0.35       262
weighted avg       1.00      0.53      0.70       262
```

Figure 22 Classification report on test data

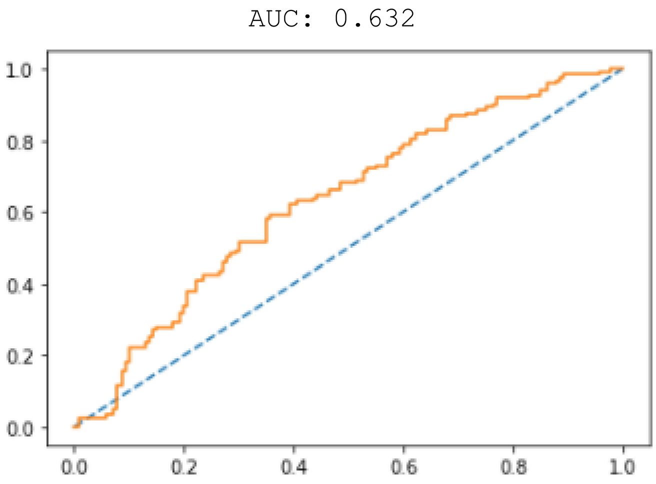


Figure 23 ROC curve for test data

OBSERVATIONS:

TRAINIG DATA

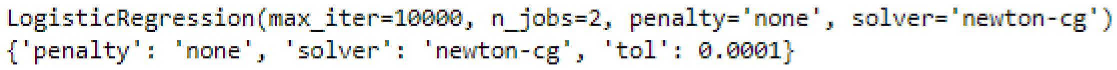
- For employee not accepting the package
  - Precision is 1.00 100% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.54 ,54% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision and Recall is 0

TEST DATA

- For employee not accepting the package
  - Precision is 1.00 100% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.53 ,53% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision and recall is 0.
- Area under the curve is 0.632 which indicates the model doesn't perform properly.
- Accuracy is also 0.53 which is not good enough.

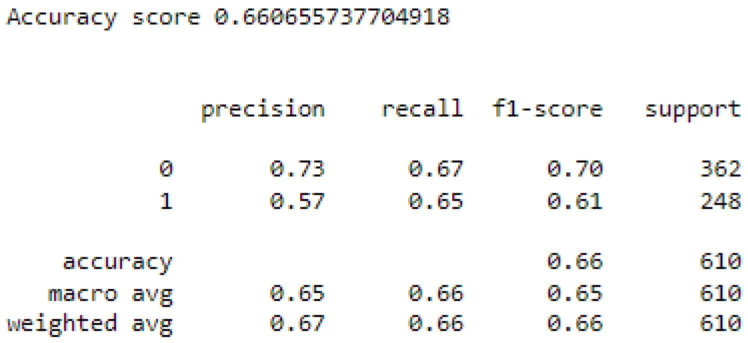
Model 2: Grid Search on Logistic classifier

Best model's parameter is given below:



```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg')
{'penalty': 'none', 'solver': 'newton-cg', 'tol': 0.0001}
```

For Training Data:



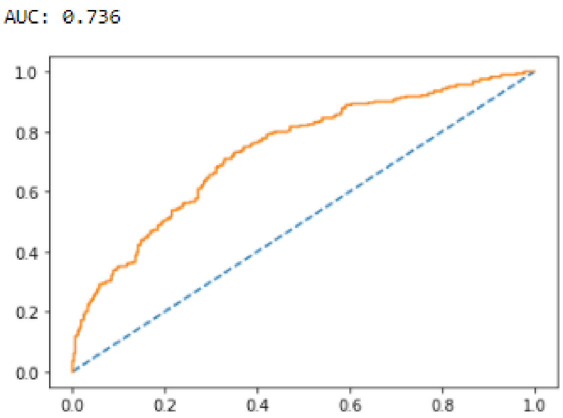
```
Accuracy score 0.660655737704918

              precision    recall  f1-score   support

           0       0.73      0.67      0.70       362
           1       0.57      0.65      0.61       248

    accuracy                           0.66       610
   macro avg       0.65      0.66      0.65       610
weighted avg       0.67      0.66      0.66       610
```

Figure 24 Classification report on Training data



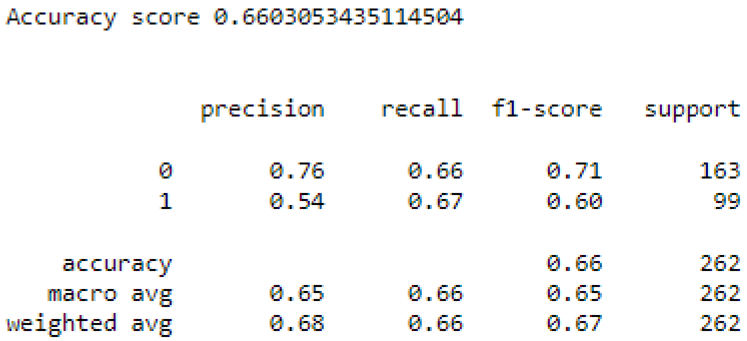
```
AUC: 0.736
```

Figure 25 ROC curve on Train data

TRAINING DATA

- For employee not accepting the package
  - Precision is 0.73 73% of the prediction of the employee who doesn’t accept the holiday package are correct.
  - Recall is 0.67 ,67% of the employees who won’t accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.57 57% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.65 ,65% of the employees who accept the package are predicted correctly
- Area under the curve is 0.736 which indicates the model is better than the previous model.
- Accuracy is also 0.66 which is better than previous mode.

TEST DATA:



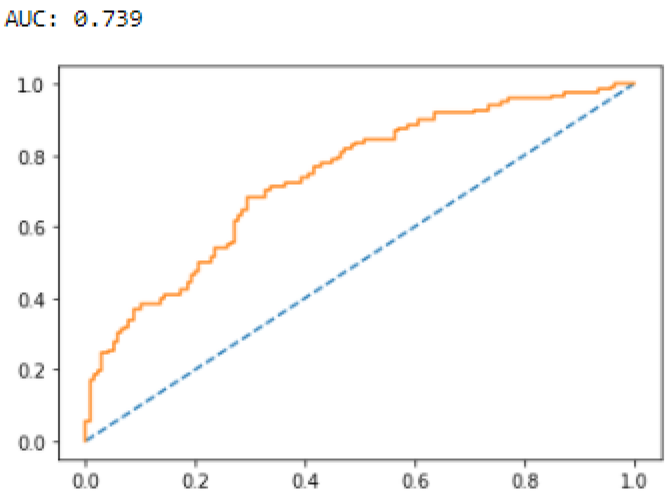
```
Accuracy score 0.6603053435114504

              precision    recall  f1-score   support

           0       0.76      0.66      0.71       163
           1       0.54      0.67      0.60        99

    accuracy                           0.66       262
   macro avg       0.65      0.66      0.65       262
weighted avg       0.68      0.66      0.67       262
```

Figure 26 Classification report on Test data



```
AUC: 0.739
```

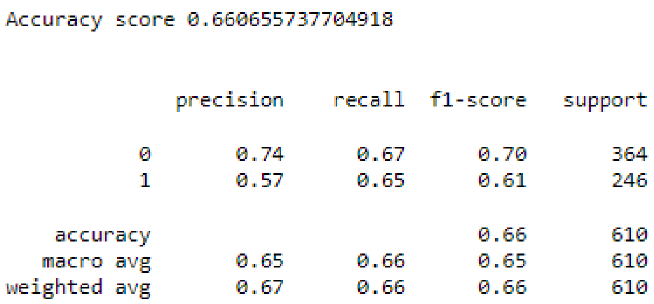
Figure 27 ROC curve for test data

TEST DATA:

- For employee not accepting the package
  - Precision is 0.76 76% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.66 ,66% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.54 54% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.67 ,67% of the employees who accept the package are predicted correctly
- Area under the curve is 0.739 which indicates the model is better than the previous model.
- Accuracy is also 0.66 which is better than previous mode.

Model 3: LDA with default parameter

FOR TRAINING DATA



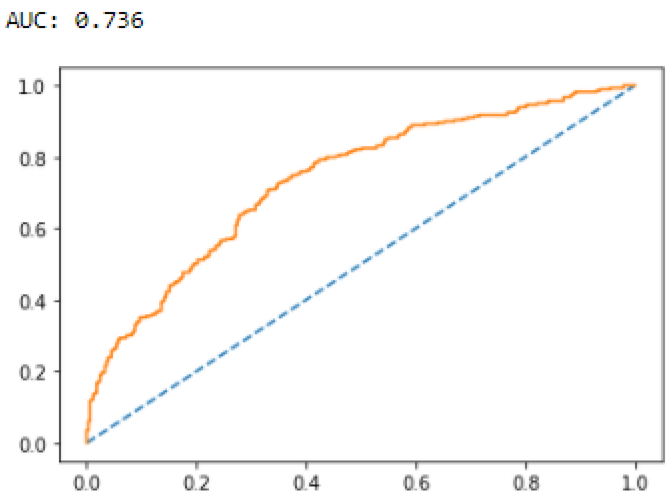
```
Accuracy score 0.660655737704918

              precision    recall  f1-score   support

           0       0.74      0.67      0.70       364
           1       0.57      0.65      0.61       246

    accuracy                           0.66       610
   macro avg       0.65      0.66      0.65       610
weighted avg       0.67      0.66      0.66       610
```

Figure 28 Classification report of LDA training



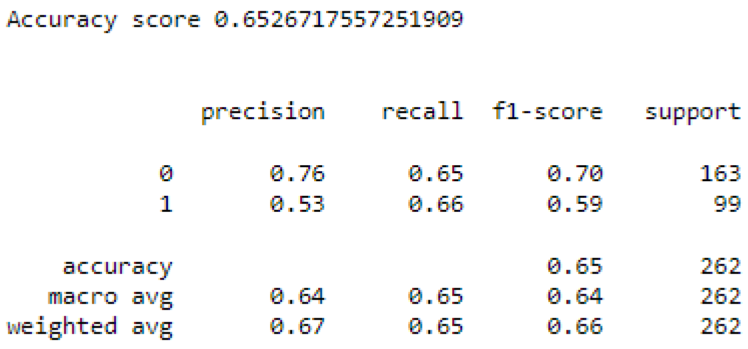
```
AUC: 0.736
```

Figure 29 ROC curve on LDA

TRAINING DATA

- For employee not accepting the package
  - Precision is 0.74 74% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.67 ,67% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.57 57% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.65 ,65% of the employees who accept the package are predicted correctly
- Area under the curve is 0.736 which indicates the model is better than the previous model.
- Accuracy is also 0.66 which is better than previous mode.

FOR TEST DATA:



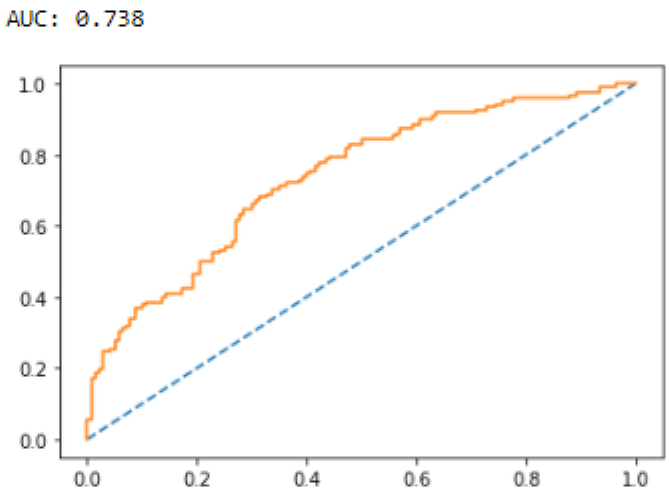
```
Accuracy score 0.6526717557251909


              precision    recall  f1-score   support

           0       0.76      0.65      0.70       163
           1       0.53      0.66      0.59        99

    accuracy                           0.65       262
   macro avg       0.64      0.65      0.64       262
weighted avg       0.67      0.65      0.66       262
```

Figure 30 Classification report on test data



```
AUC: 0.738
```

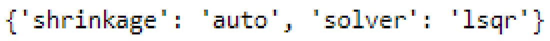
Figure 31 ROC cure on LDA test

TEST DATA:

- For employee not accepting the package
  - Precision is 0.76 76% of the prediction of the employee who doesn’t accept the holiday package are correct.
  - Recall is 0.65 ,65% of the employees who won’t accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.53 53% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.66 ,66% of the employees who accept the package are predicted correctly
- Area under the curve is 0.738 which indicates the model is better than the previous model.
- Accuracy is also 0.65 which is better than previous mode.

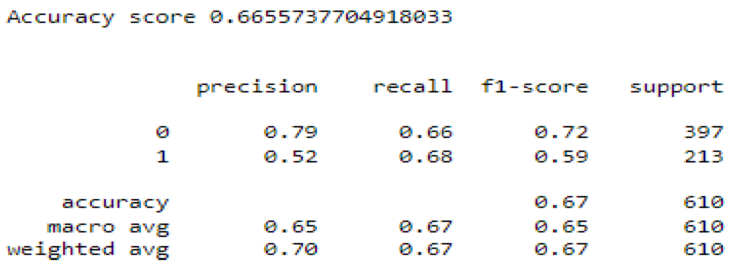
Model 4: Applying Grid Search on LDA

Best estimator’s parameters are given below:



```
{'shrinkage': 'auto', 'solver': 'lsqr'}
```

FOR TRAINING DATA:



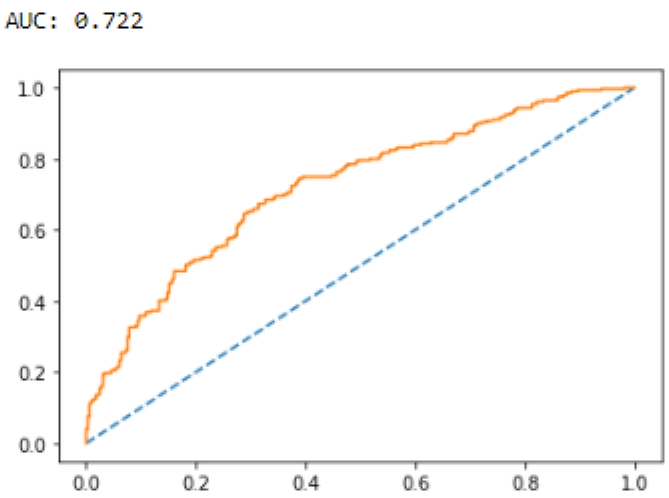
```
Accuracy score 0.6655737704918033

              precision    recall  f1-score   support

           0       0.79      0.66      0.72       397
           1       0.52      0.68      0.59       213

    accuracy                           0.67       610
   macro avg       0.65      0.67      0.65       610
weighted avg       0.70      0.67      0.67       610
```

Figure 32 Classification report on train



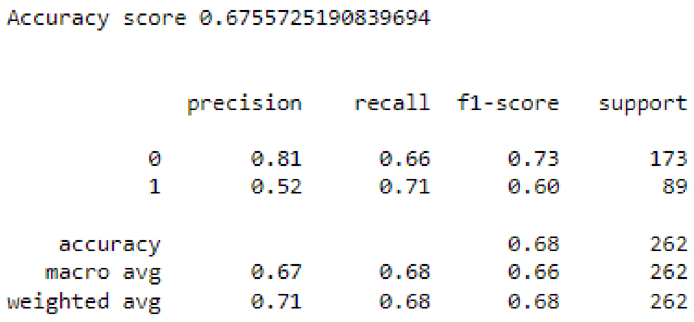
```
AUC: 0.722
```

Figure 33 ROC curve on train data

TRAINING DATA

- For employee not accepting the package
  - Precision is 0.79 79% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.66 ,66% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.52 52% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.68 ,68% of the employees who accept the package are predicted correctly
- Area under the curve is 0.722 which indicates the model is better than the previous model.
- Accuracy is also 0.66 which is better than previous mode.

FOR TEST DATA:



```
Accuracy score 0.6755725190839694

              precision    recall  f1-score   support

           0       0.81      0.66      0.73       173
           1       0.52      0.71      0.60        89

    accuracy                           0.68       262
   macro avg       0.67      0.68      0.66       262
weighted avg       0.71      0.68      0.68       262
```

Figure 34 Classification report on test

```
AUC: 0.737
```

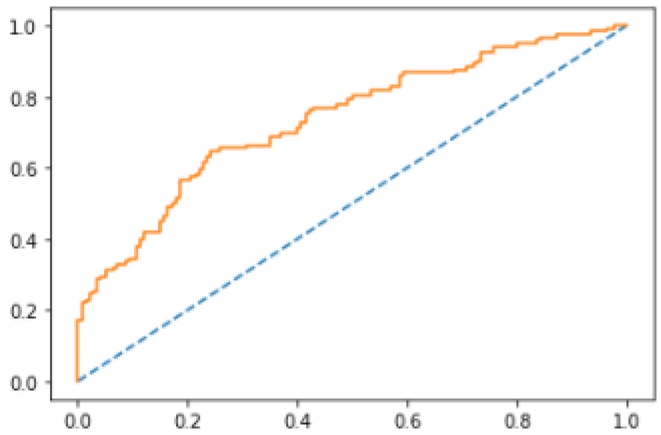


Figure 35 ROC curve test data

TEST DATA:

- For employee not accepting the package
  - Precision is 0.81 81% of the prediction of the employee who doesn't accept the holiday package are correct.
  - Recall is 0.66 ,66% of the employees who won't accept the package are predicted correctly
- For employee accepting the package
  - Precision is 0.52 52% of the prediction of the employee who does accept the holiday package are correct.
  - Recall is 0.71 ,71% of the employees who accept the package are predicted correctly
- Area under the curve is 0.737 which indicates the model is better than the previous model.
- Accuracy is also 0.67 which is better than previous mode.

7. MODEL COMPARISON

CONFUSION MATRIX ON TRAINING DATA

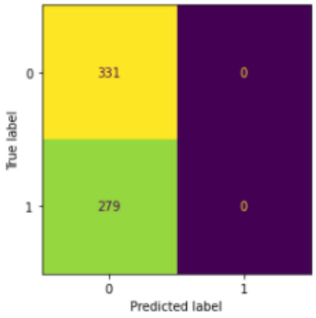


Figure 36 Logistic clf

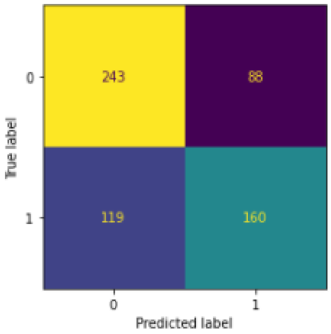


Figure 37 Grid Search on Log clf

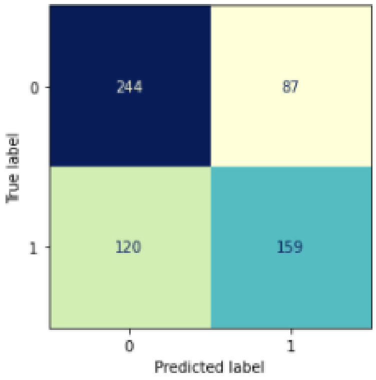


Figure 38 LDA

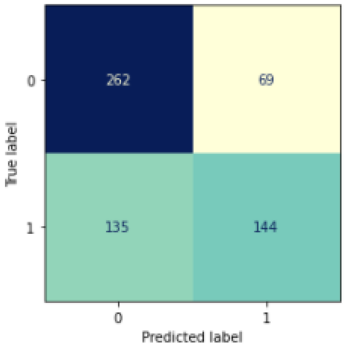


Figure 39 Grid search on LDA

CONFUSION MATRIX ON TEST DATA:

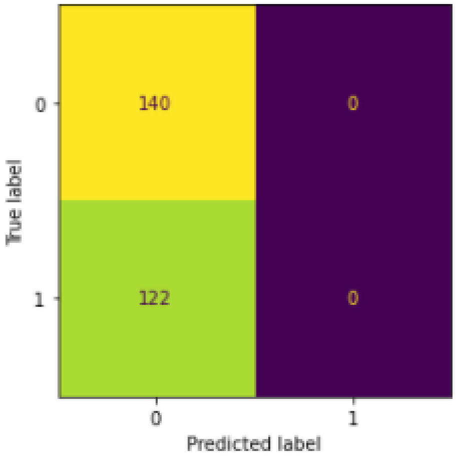


Figure 40 Logistic clf

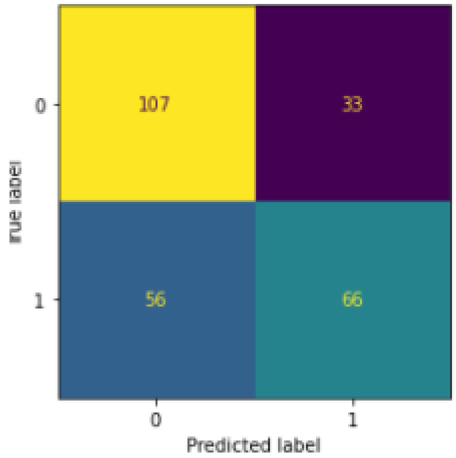


Figure 41 Grid Search on Logistic clf

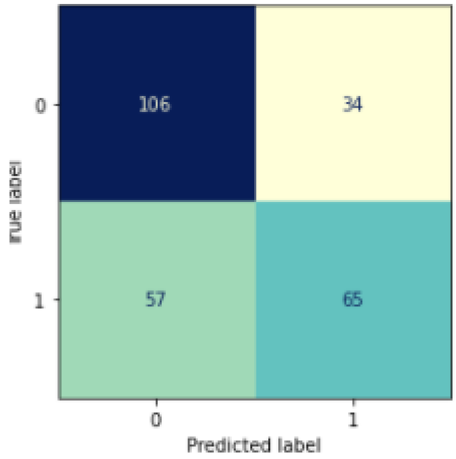


Figure 42 LDA

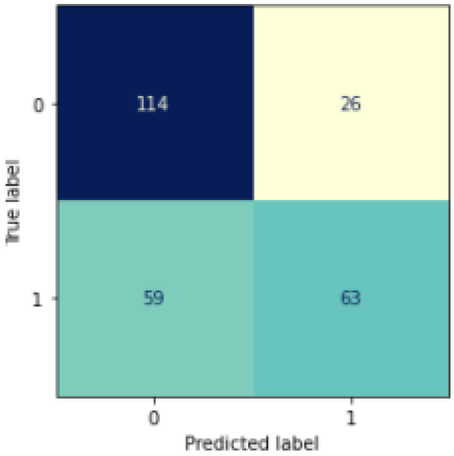
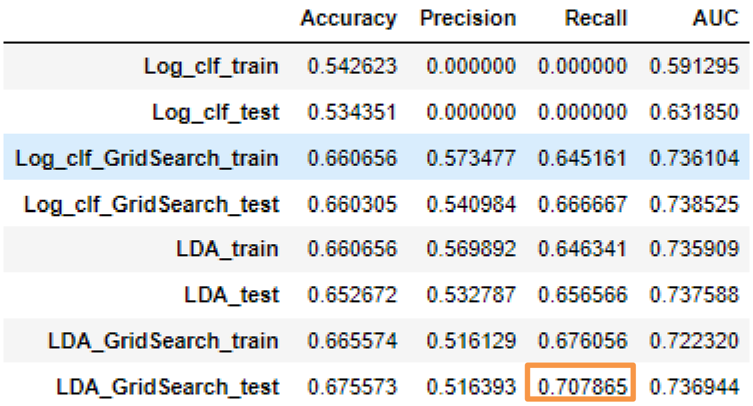


Figure 43 Grid Search on LDA



	Accuracy	Precision	Recall	AUC
Log_clf_train	0.542623	0.000000	0.000000	0.591295
Log_clf_test	0.534351	0.000000	0.000000	0.631850
Log_clf_GridSearch_train	0.660656	0.573477	0.645161	0.736104
Log_clf_GridSearch_test	0.660305	0.540984	0.666667	0.738525
LDA_train	0.660656	0.569892	0.646341	0.735909
LDA_test	0.652672	0.532787	0.656566	0.737588
LDA_GridSearch_train	0.665574	0.516129	0.676056	0.722320
LDA_GridSearch_test	0.675573	0.516393	0.707865	0.736944

8. MODEL SELECTION:

Precision = TP/ (TP+FP)

Recall= TP/ (TP+FN)

We are asked to find the employees who are most probable to take the holiday package so in that case we need to use recall as the metric to choose the best model. Precision means how many selected items are actually positive, whereas recall is how many positive items are actually selected. So by checking the recall LDA model with Grid Search has the best recall and AUC is also better than other models so for the current problem statement the above mentioned model will be accurate.

SUMMARY:

Once retrieving the data we found out the data was clean from duplicate records, missing values & wrong entries by doing the sanitary checks. Although it had outliers because we had salary variable in the dataset, it is common for salary to contain outliers.

Cleaning the data led to proceed with EDA on the dataset, the insights from the EDA are

- Salary of the employee is widely spread so the company has people earning in wide range.
- People who opt for the package earn moderately.
- Age of the people working in the company is within 35 to 45 , so most of them would be having family.
- Average education of the employee is 9
- People who have no children are most likely to accept the Holiday package
- Ratio of People with 2 older children accepting the package is more.
- Holiday Package is mostly welcomed by the foreigners.

With the help of above insights we can get a clear understanding which is a useful predictor, but to get a clear idea we used chi square test on the categorical variables and found out FOREIGN, YOUNG CHILDREN are significant predictors whereas OLDER CHILDREN is not that useful in the prediction.

Once the predictors are chosen all the categorical columns are one Hot Encoded and fed to Logistic classifier and Liner discriminant analysis classifier. By the use of recall and Area under the curve of ROC curve LDA seems to perform better compared to Logistic classifier. The best model was obtained by running Grid Search on the models.

RECOMMENDATION:

- People earning more are not really taking the package so we must find the reason why? And come with a premium package plans to attract the expensive customers.
- For those who are taking the package are from moderately earning group so to encourage more people from this class, we can introduce referral points for the employee who would recommend our program to a peer. Or increase one or two days in the package plan according to the referral.

- Mostly foreigners are choosing the package a separate strategy must be formed to reach the foreigner in the company. To attract local people we can offer some discounts on the price and alter the package plans for a short period and try to sell in those employee if they are looking for a weekend getaway.
- People with children are hesitant to take the package so more family based packages must be introduced in the plans.