Predictive Modelling Project

NAME: SHOUNACK MANDAL

COURSE: PGP - DSBA Online Sep.

Date: 20/ February / 2022

Contents

LIST OF FIGURE

LIST OF TABLES

1 Problem 1: Linear Regression

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

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.

Table 1. Data Dictionary

INTRODUCTION

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), by performing EDA and then using linear regression to predict the price of cubic zirconia diamonds based on the given independent variables. Also providing the best 5 attributes that are most important.

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

Loading all necessary libraries for the model building. Reading and visualizing the data sample(head), Data type, Information of the dataset, shape, EDA- along with Description of the objective variables in the dataset separately.

Dataset sample:-

	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

Table 2. Dataset sample

Data type: -

```
Unnamed: 0      int64
carat         float64
cut            object
color          object
clarity        object
depth         float64
table         float64
x             float64
y             float64
z             float64
price           int64
dtype: object
```

Table 3 Data type

Information of the dataset:-

```
<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
```

Table 4 Information of the dataset

Descriptive statistics:-

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

Table 5 Descriptive statistics

Description of the objective variables in the dataset:-

	cut	color	clarity
count	26967	26967	26967
unique	5	7	8
top	Ideal	G	SI1
freq	10816	5661	6571

Table 6 Description of the objective variables in the dataset

Desciptive statistics help us to understand Interquartile Range like minimum, maximum, 25th, 50th and 75thpercentiles, mean or average, standard deviation and count of data observations etc. This includes the fundamental data analysis which is Measures of central tendency- Mean, Median and Mode.

Inferences of the dataset: -

1. Unnamed: 0 column has been removed as no meaning to the model making.
2. 'Carat' seems to be normally distributed.
3. Most of the 'Cut' quality is – IDEAL.
4. 'Color' of the cubic zirconia is most of the times **G.**
5. 'Clarity' of the zirconia mostly SI1
6. Price as the dependent variable has range of minimum of 326 and maximum of 18818 with a amean price of 3939.51 to a median price of 2375.
7. Most of the zirconia price is below 9920.40

Univariate Analysis:-

Distribution plot:-

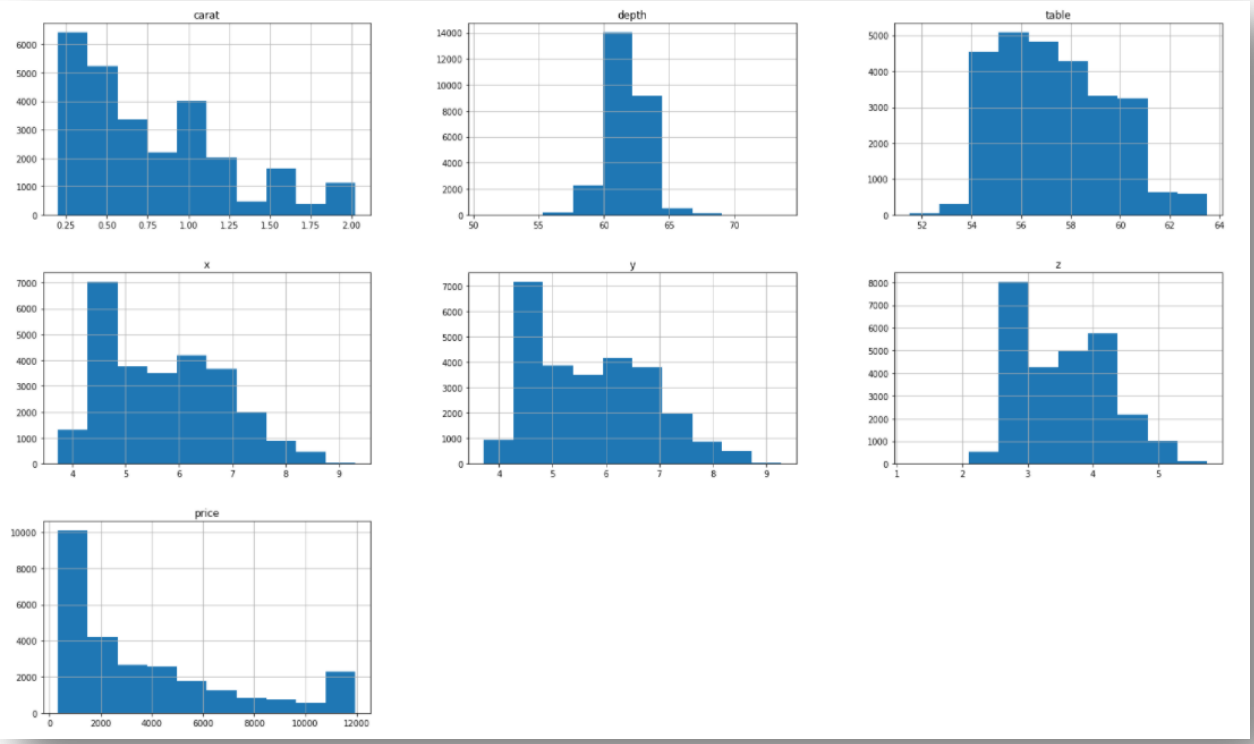

Figure 1 Distribution Plot

Boxplot of each variables:-

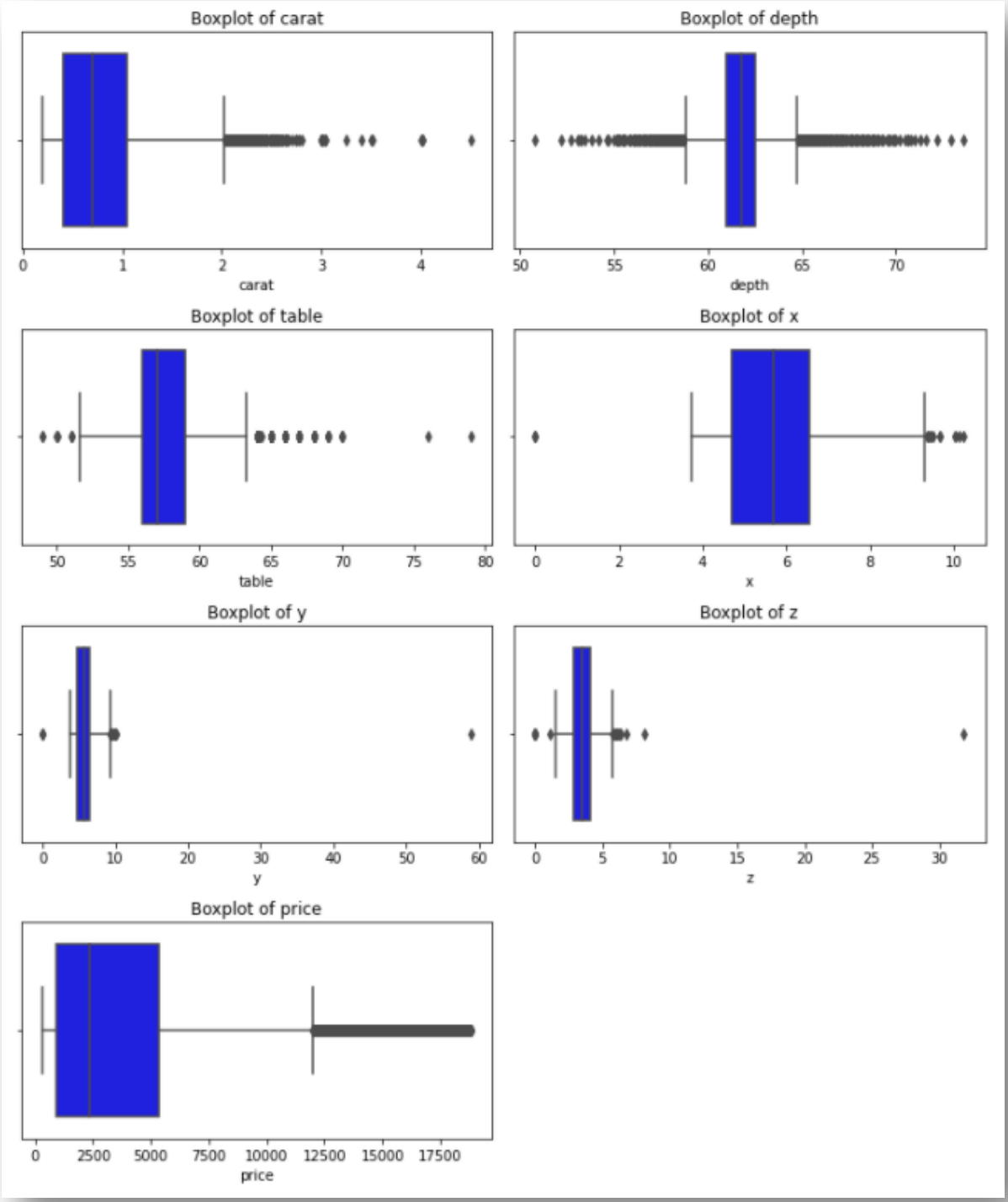

Figure 2 Box plot of each variables

We can see that all the box plot of each variable has outliers.

Count plot for the categorical variables: -

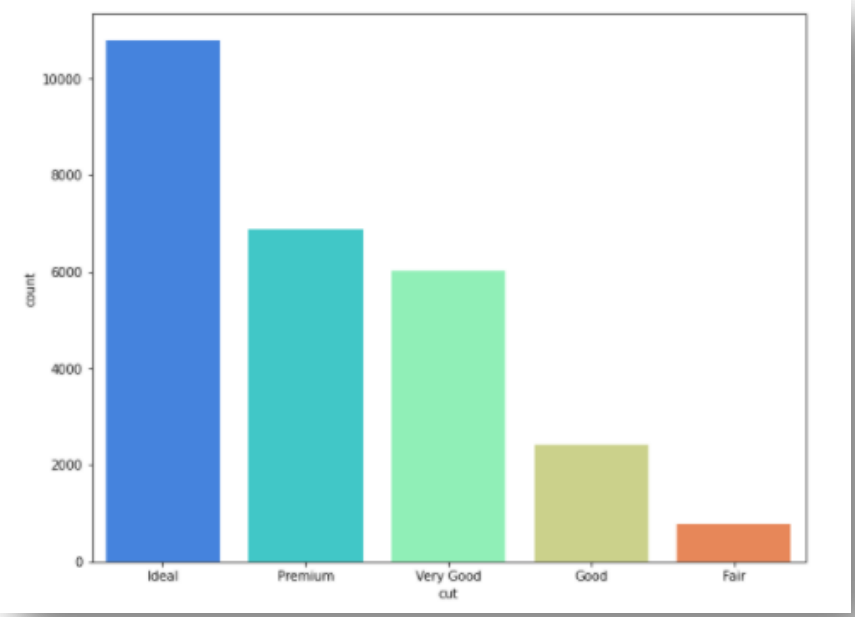

Figure 3 Count plot of CUT

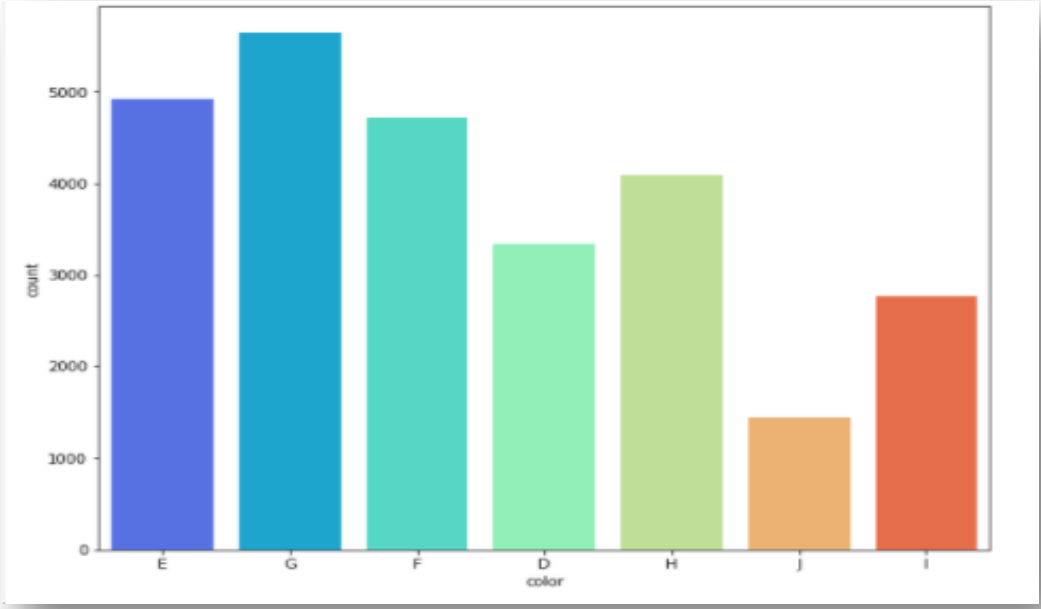

Figure 4 Count plot of Color

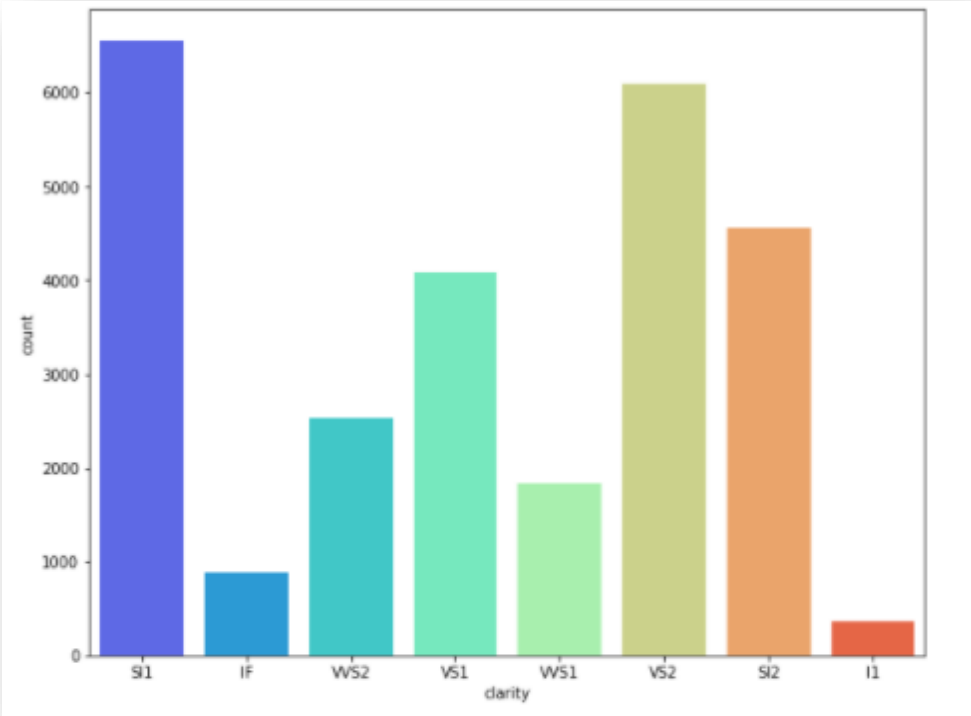

Figure 5 Count plot of Clarity

Bivariate Analysis:-

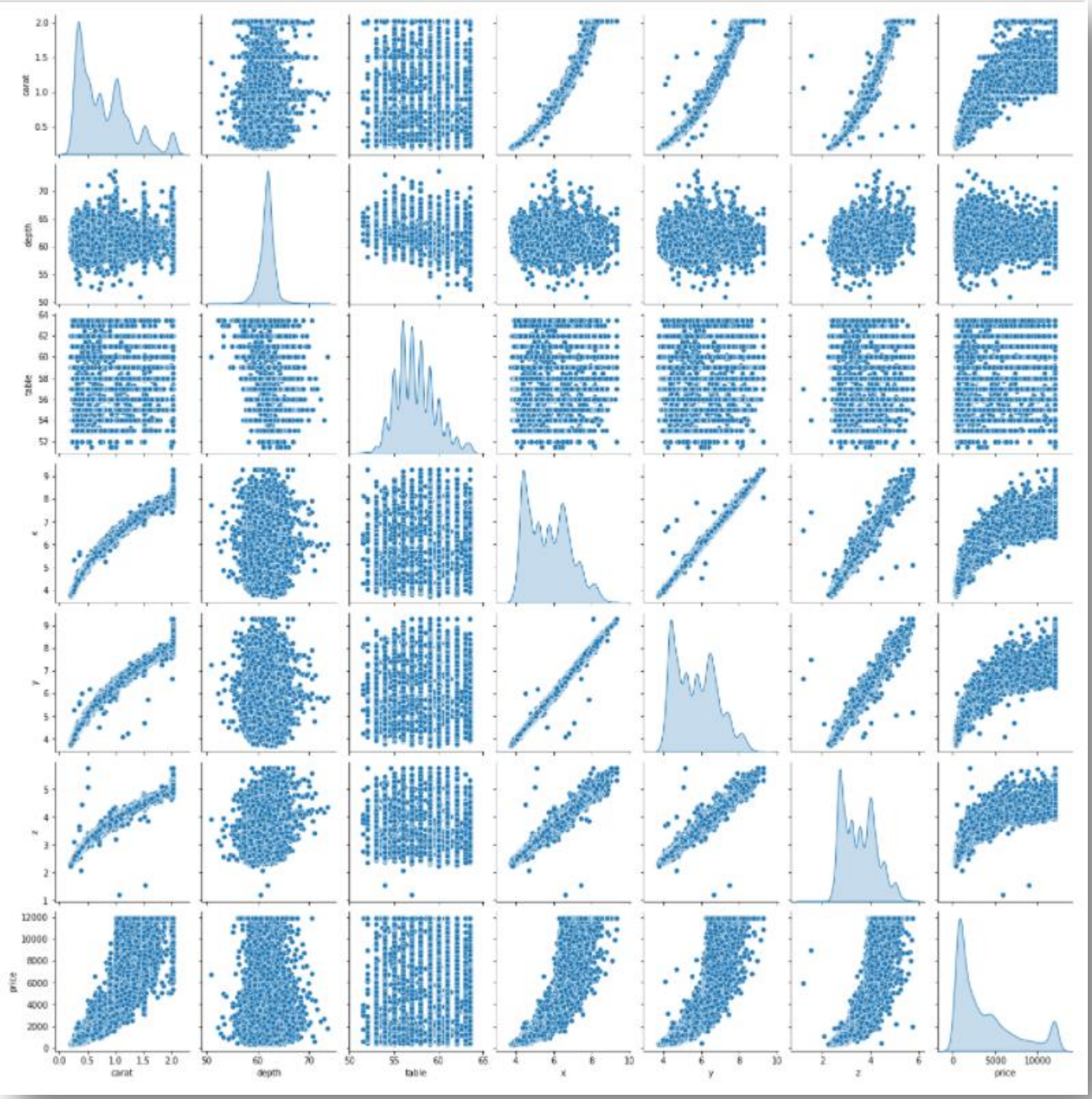

Figure 6 Pair plot

Pair plot shows the relation between each variables in the form of scatter plot and distribution plot and also shows the negative, positive or no correlation among the variables. From the above pairplot we see there is a strong positive relationship between carat and price. Similarly x, y and z have a moderately positive correlation with carat. Variables x, y and z themselves seem to have high multicollinearity. However we will also see if this is true when we check the correlation heatmap.

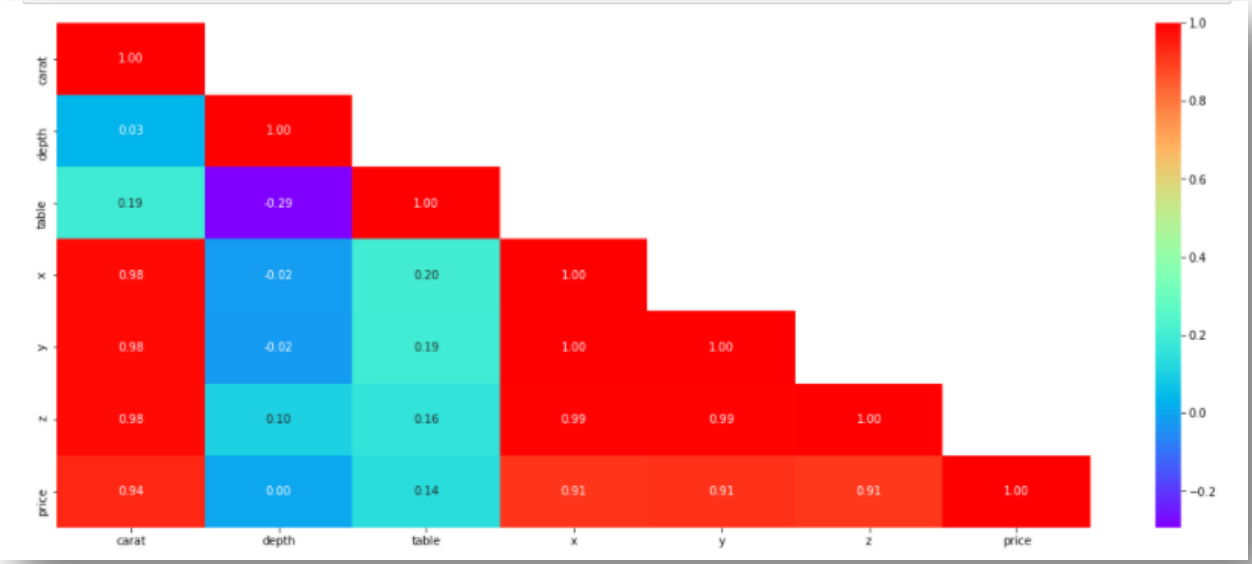

Figure 7 Correlation Heat map

From the above correlation plot we can see that x, y and z have high positive correlation with carat. There is also Multicollinearity between x, y and z variables. Price and carat have a strong positive correlation with a correlation coefficient of 0.92. Correlation between x, y and z with price is positive but they are moderate. Correlation values are always between 1 and -1. Those which are closer to 1 are positively correlated and those which near -1 are negatively correlated. Values near to 0 have no correlation.

```
price    1.000000
carat    0.922400
x        0.887467
y        0.857255
z        0.855775
table    0.126967
depth   -0.002736
Name: price, dtype: float64
```

Table 7 Price effect chart.

Relation of each feature that has price effect that differentiates the cost for diamonds. It can be inferred that most features correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (<1%).

EDA of each categorical variables:-

CUT

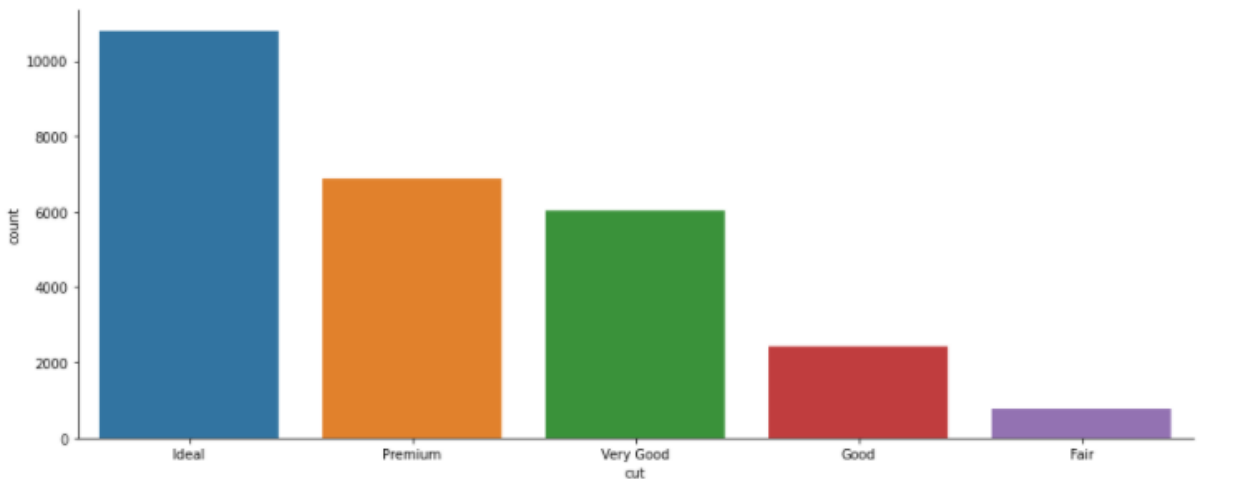

Figure 8 EDA(Catplot) of the categorical variable count with cut.

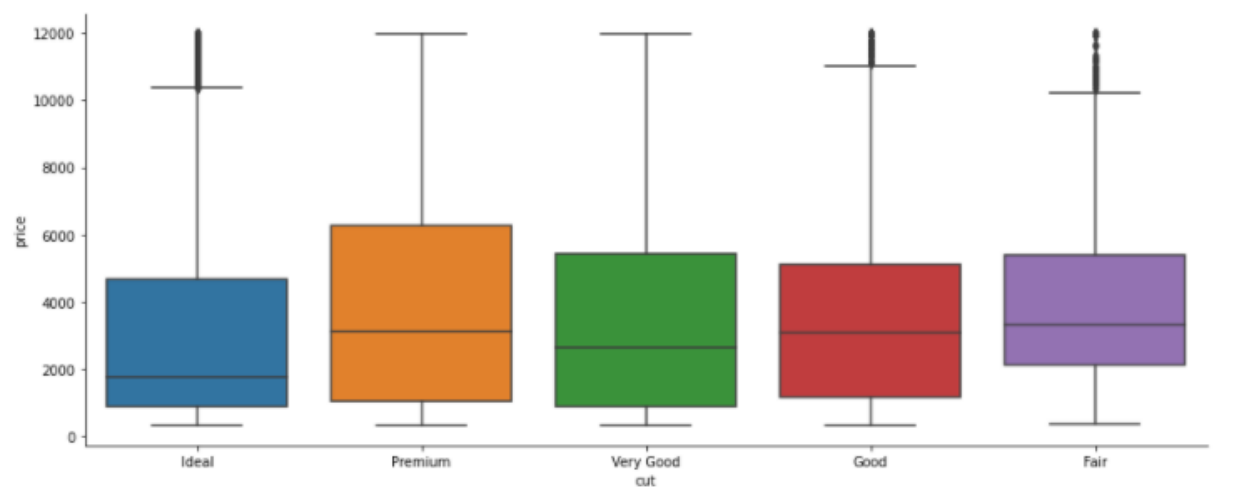

Figure 9EDA(Catplot) of the categorical variable price with cut.

COLOR

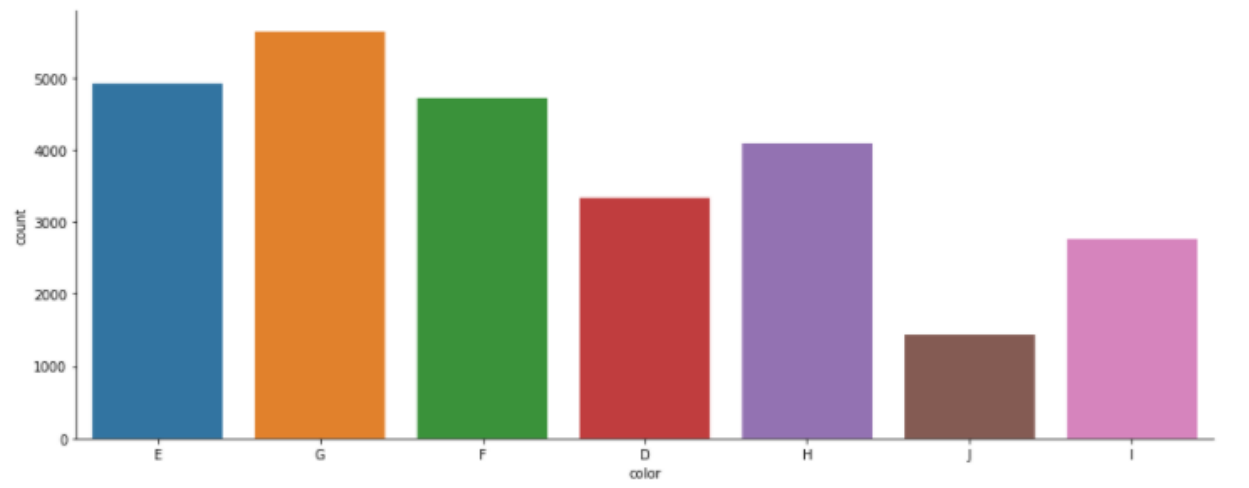

Figure 10 EDA(Catplot) of the categorical variable color with count

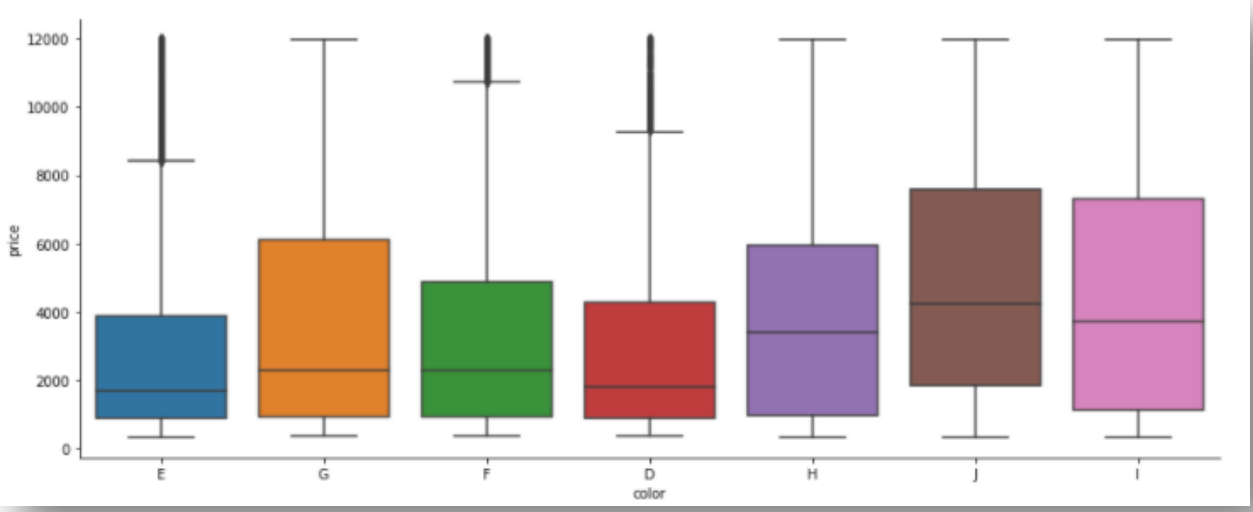

Figure 11 EDA(Catplot) of the categorical variable price with Color

CLARITY

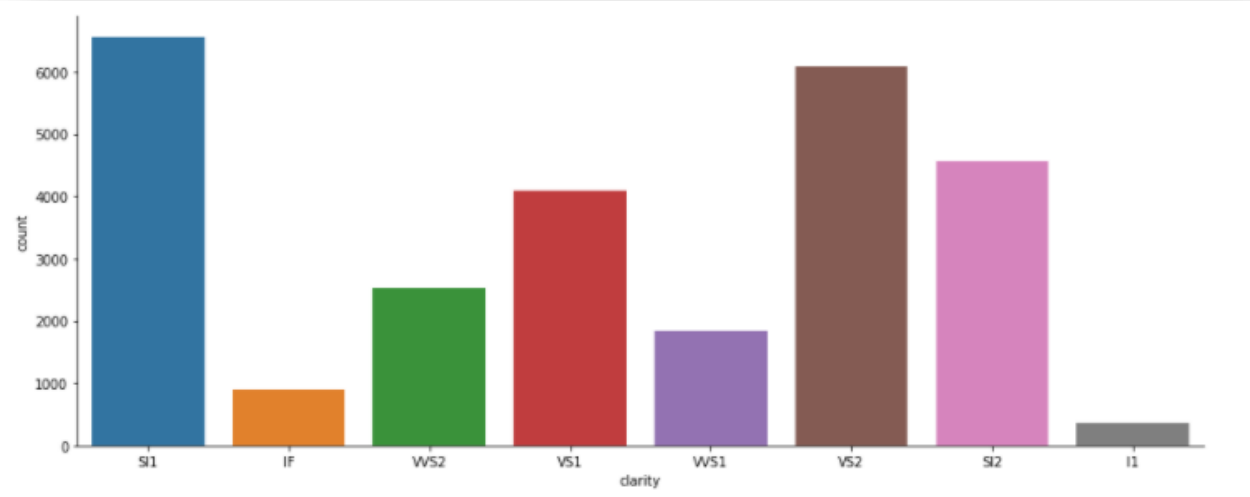

Figure 12 EDA(Catplot) of the categorical variable clarity with count

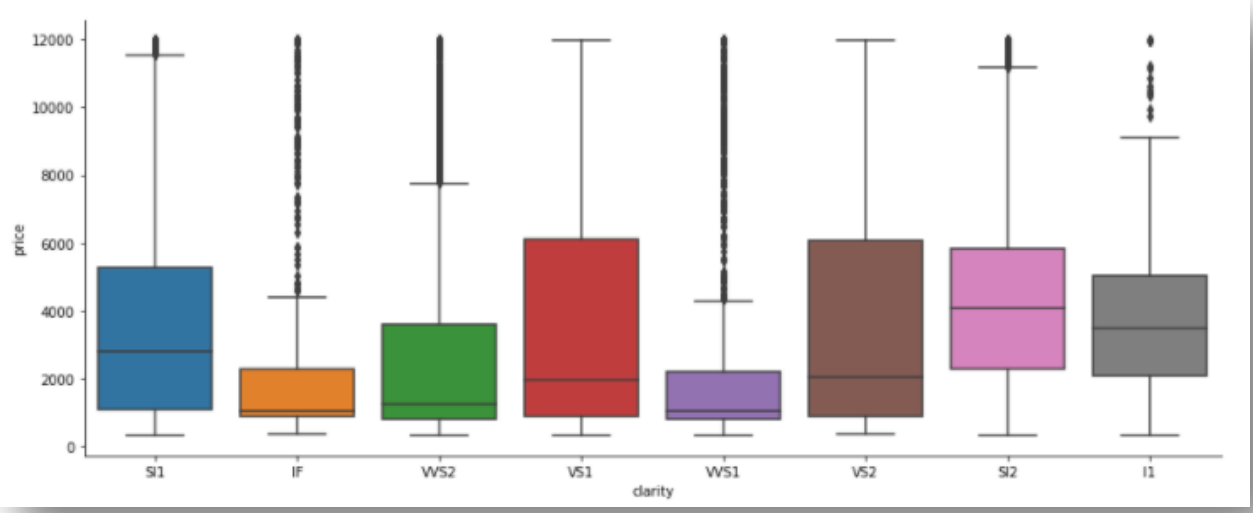

Figure 13 EDA(Catplot) of the categorical variable price with clarity

Observation on 'clarity': The Diamonds clarity with VS1 & VS2 are the most Expensive.

The inferences drawn from the above Exploratory Data analysis:

Observation-1:

(1).'Price' is the target variable while all others are the predictors.

(2).The data set contains 26967 row, 11 column.

(3).In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable and all other are predector variable.

(4)The first column is an index ("Unnamed: 0")as this only serial no, we can remove it.

Observation-2:

(1).On the given data set the the mean and median values does not have much differenc.

(2).We can observe Min value of "x", "y", "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So we have filter out those as it clearly faulty data entries.

(3).There are three object data type 'cut', 'color' and 'clarity'.

Observation-3: we can observe there are 697 missing value in the depth column. There are some duplicate row present. (33 duplicate rows out of 26958). which is nearly 0.12 % of the total data. So on this case we have dropped the duplicated row.

Observation-4: : There are significant amount of outlier present in some variable,the features with datapoint that are far from the rest of dataset which will affect the outcome of our regression model. So we have treat the outliar. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

Observation-5: It looks like most features do correlate with the price of Diamond. The notable exception is "depth" which has a negligble correlation (~1%). Observation on 'CUT': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

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

Table 8 Missing values

From the table, we saw that there are 697 Null values in depth column. We have null values in depth. We are not sure why these are kept blank and since we do not want to lose valuable information from the dataset, we will be imputing the Median value for missing values in depth column.

Dataset after imputing the median values:-

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

Table 9 Imputing missing values

Info table after missing values imputation:-

```
<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
```

Table 10 Info table with non-null value

Checking for the values which are equal to zero and dropping those values as they are fake values:-

```
Number of rows with x == 0: 3
Number of rows with y == 0: 3
Number of rows with z == 0: 9
Number of rows with depth == 0: 0
```

Table 12 Table shows number of zero values in x,y and z. (Number of dimension less cubic's.)

```
Number of rows with x == 0: 0
Number of rows with y == 0: 0
Number of rows with z == 0: 0
Number of rows with depth == 0: 0
```

Table 11 Table for checking zero values after removing the dimensionless cubic's.

From the earlier summary statistics table, the attributes x, y, z have minimum value as 0.00mm. They are clear indicators that they have been incorrectly captured as bad data. It's a data input error because there cannot be diamonds with 0.00mm length, width and height. Hence we will be dropping those rows as it is not a significant loss of information.

Before and after the outlier treatment (trimming):-

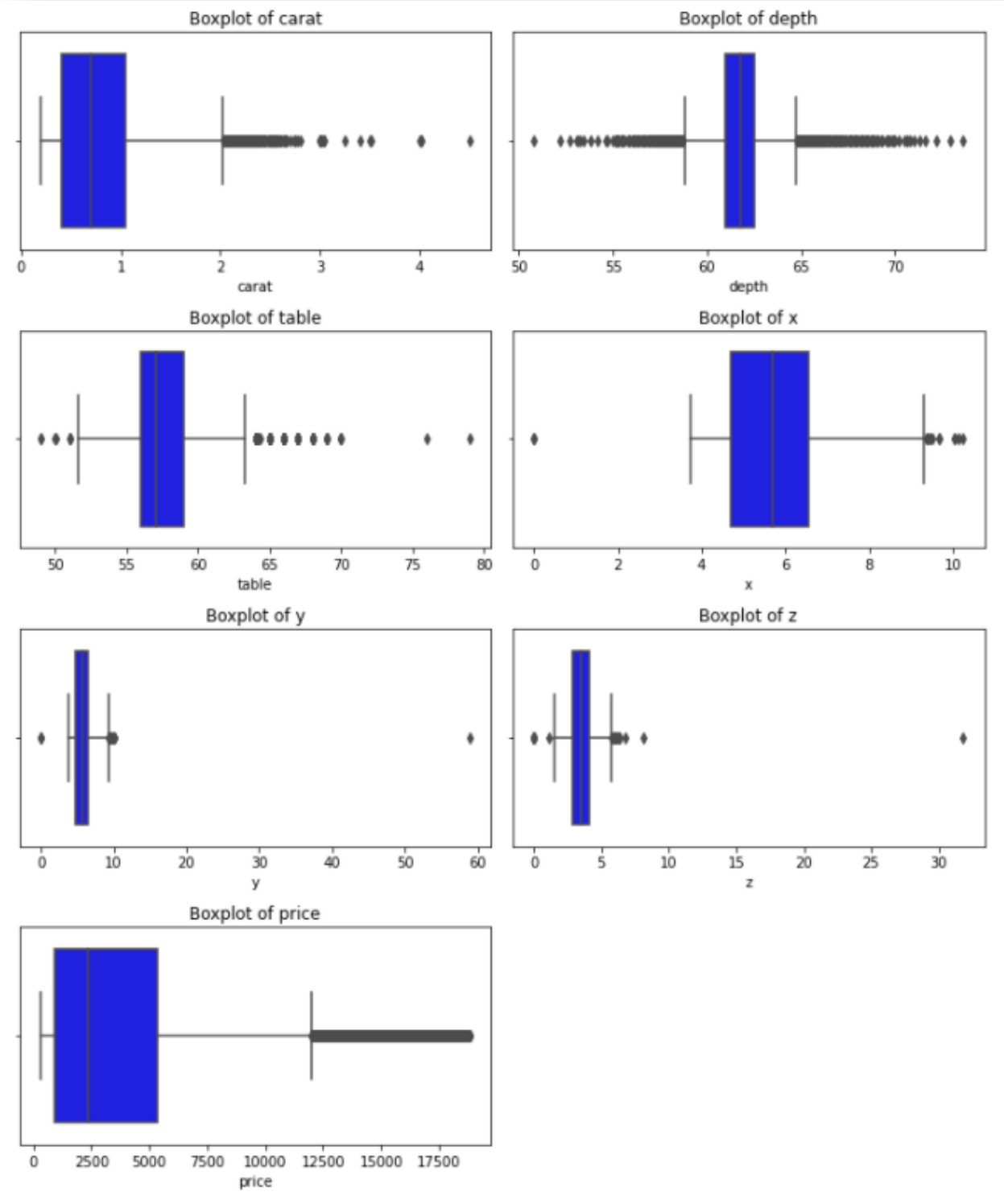

Figure 14 Boxplot of the variables with outliers

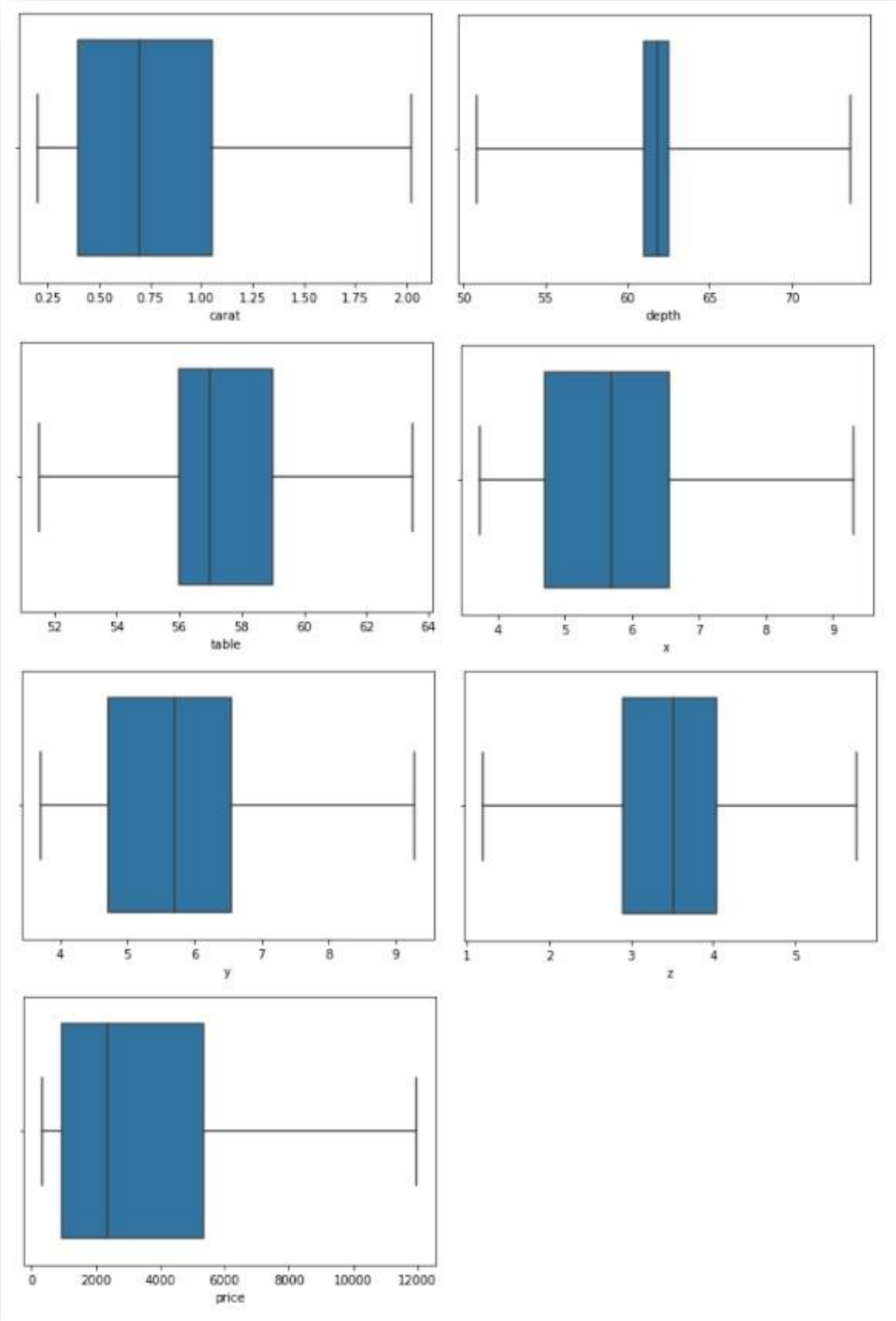

Figure 15 Boxplot of variables after trimming.

Scalling is not ecessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not. But recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in preprocessing.

For now we will process the model without scaling and later we will check the output with scaled data of regression model output.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

Linear regression

Simple definition of Linear Regression is: if there are n observations with **y** being the dependent variable and **x** being the independent variable, then variable y takes a linear combination of x to form the best fit line which predicts the y values given the x values.

Will start the liner regression after modifying the data:-

Getting unique counts for all objects:- CUT, COLOR and CLARITY.

```
cut
 Ideal        10816
Premium       6893
Very Good     6030
Good          2439
Fair           780
Name: cut, dtype: int64


color
 G    5658
E    4917
F    4727
H    4098
D    3344
I    2771
J    1443
Name: color, dtype: int64


clarity
 SI1     6570
VS2     6098
SI2     4571
VS1     4092
VVS2    2531
VVS1    1839
IF       894
I1       363
Name: clarity, dtype: int64
```

Table 13 Unique values of all the variables before encoding.

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4.0	5.0	2.0	62.1	58.0	4.27	4.29	2.66	499.0
1	0.33	3.0	3.0	7.0	60.8	58.0	4.42	4.46	2.70	984.0
2	0.90	2.0	5.0	5.0	62.2	60.0	6.04	6.12	3.78	6289.0
3	0.42	4.0	4.0	4.0	61.6	56.0	4.82	4.80	2.96	1082.0
4	0.31	4.0	4.0	6.0	60.4	59.0	4.35	4.43	2.65	779.0
5	1.02	4.0	6.0	3.0	61.5	56.0	6.46	6.49	3.99	9502.0
6	1.01	1.0	2.0	2.0	63.7	60.0	6.35	6.30	4.03	4836.0
7	0.50	3.0	5.0	2.0	61.5	62.0	5.09	5.06	3.12	1415.0
8	1.21	1.0	2.0	2.0	63.8	63.5	6.72	6.63	4.26	5407.0
9	0.35	4.0	4.0	3.0	60.5	57.0	4.52	4.60	2.76	706.0

Table 14 Data Post encoding

We have encoded the categorical variables cut, color and clarity in the ascending order from worst to best since linear regression does not take string variables as parameters into model building.
Below is the encoding for ordinal values:
CUT: Fair = 1, Good = 2, Very Good = 3, Premium = 4 and Ideal = 5
COLOR: D = 1, E = 2, F = 3, G = 4, H = 5, I = 6 and J = 7
CLARITY = IF = 1, VVS1 = 2, VVS2 = 3, VS1 = 4, VS2 = 5, SI1 = 6, SI2 = 7 and I1 = 8

Converting the categorical data type to numerical:-

```
carat      float64
cut        float64
color      float64
clarity    float64
depth      float64
table      float64
x          float64
y          float64
z          float64
price      float64
dtype: object
```

Table 15 data types from categorical to numeric/float.

Multiple Linear regression using Scikit Learn

Linear Regression Model 1 values:-

Using the sklearn.model_selection package to use train_test_split, sklearn.linear_model to use LinearRegression and sklearn.metrics to use mean_squared_error and r2_score.

```
The coefficient for carat is 8909.400896689896
The coefficient for cut is 105.5276383676879
The coefficient for color is 271.9935588326712
The coefficient for clarity is 427.61438043247085
The coefficient for depth is 53.34026921758545
The coefficient for table is -13.441905940572202
The coefficient for x is -1234.8498477426642
The coefficient for y is 1723.6544727565179
The coefficient for z is -1455.6466799234004
```

Table 16 coefficients for each of the independent attributes

The model intercept: [-6050.99917756]
Mean Squared Error (MSE): 853708.99
Root Mean Squared Error (RMSE): 923.9637393955937
Co-efficient of Determination (r-square) on the train data: 0.9312382609355447 = 93.12 %
Co-efficient of Determination (r-square) on the test data: 0.9304042555071269 = 93.04 %

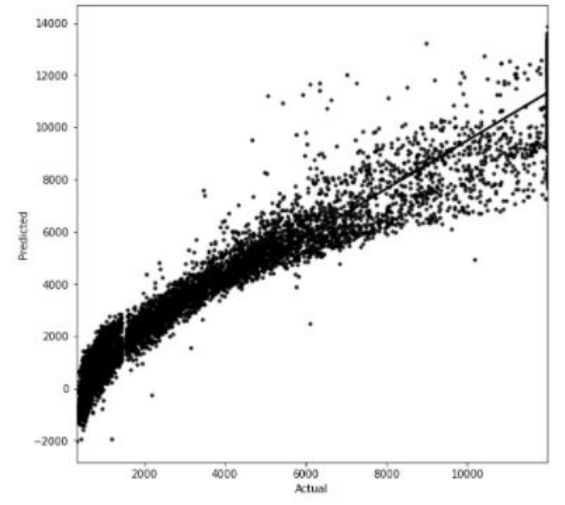

Figure 16 Scatter plot of Actual vs predicted price

Multiple Linear Regression using Stats Model –Model 2

Using the Ordinary Least Squares Method under Stats Model below are the results:

Intercept	-578.573369
carat	8856.230986
color	273.706315
clarity	441.155739
depth	-4.508633
table	-44.213788
x	-1031.575783
y	1290.323494
z	-1033.099422
dtype: float64	

Table 17 Model 2 parameters

OLS Regression Results

Dep. Variable:	price	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	4.378e+04
Date:	Sun, 20 Feb 2022	Prob (F-statistic):	0.00
Time:	02:31:00	Log-Likelihood:	-2.1635e+05
No. Observations:	26261	AIC:	4.327e+05
Df Residuals:	26252	BIC:	4.328e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-578.5734	725.517	-0.797	0.425	-2000.626	843.479
carat	8856.2310	70.306	125.967	0.000	8718.428	8994.034
color	273.7063	3.506	78.060	0.000	266.834	280.579
clarity	441.1557	3.811	115.772	0.000	433.687	448.625
depth	-4.5086	10.728	-0.420	0.674	-25.535	16.518
table	-44.2138	2.847	-15.532	0.000	-49.793	-38.634
x	-1031.5758	105.740	-9.756	0.000	-1238.831	-824.321
y	1290.3235	107.153	12.042	0.000	1080.298	1500.349
z	-1033.0994	162.982	-6.339	0.000	-1352.554	-713.645

Omnibus:	3556.422	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15146.711
Skew:	0.614	Prob(JB):	0.00
Kurtosis:	6.512	Cond. No.	1.11e+04

Table 18 Model 2 OLS regression model 2 result

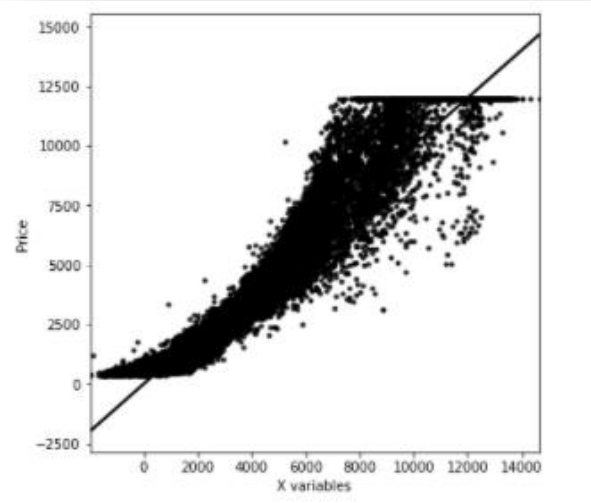

Figure 17 Scatter plot of predicted price based on X variable of model 2

1. Dependent variable is “price”.
2. Model and Method used is Ordinary Least Square (OLS) method which uses mathematical algorithm for linear regression.
3. No of observations are 26228.
4. R-squared value is 0.930 which means 93% of the outcome variability is explained by the model.
5. Adj. R-squared value is 0.930 which is the correct R-square according to the number of dependent variables.
6. F-statistic value is used for the calculation of the p-value of the model, Probability (F-statistic) which here is less than 0.05. This also tells us Python is using an ANOVA test which implies an F-distribution.
7. Coef shows the coefficients of each dependent variable.
8. Std err shows how accurate our coefficient values are. Std err is inversely related to accuracy. Lower the std err signifies higher the accuracy.
9. P>|t| is the p-value. This shows how statistically significant each independent variable is on the price (dependent variable). P-value less than 0.05 means they are statistically quite significant.

Multiple Linear Regression using Stats Model –Model 3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.930
Model:	OLS	Adj. R-squared:	0.930
Method:	Least Squares	F-statistic:	5.803e+04
Date:	Sun, 20 Feb 2022	Prob (F-statistic):	0.00
Time:	02:37:32	Log-Likelihood:	-2.1642e+05
No. Observations:	26261	AIC:	4.329e+05
Df Residuals:	26254	BIC:	4.329e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3852.7264	401.110	9.605	0.000	3066.528	4638.924
carat	8894.6958	70.228	126.654	0.000	8757.044	9032.347
color	274.0439	3.515	77.954	0.000	267.153	280.934
clarity	443.5515	3.813	116.334	0.000	436.078	451.025
depth	-71.4145	4.424	-16.141	0.000	-80.087	-62.742
table	-48.0540	2.834	-16.956	0.000	-53.609	-42.499
x	-399.4179	28.957	-13.794	0.000	-456.174	-342.661

Omnibus:	3511.519	Durbin-Watson:	2.016
Prob(Omnibus):	0.000	Jarque-Bera (JB):	15781.239
Skew:	0.589	Prob(JB):	0.00
Kurtosis:	6.610	Cond. No.	6.02e+03

Table 19 OLS regression model 3 result

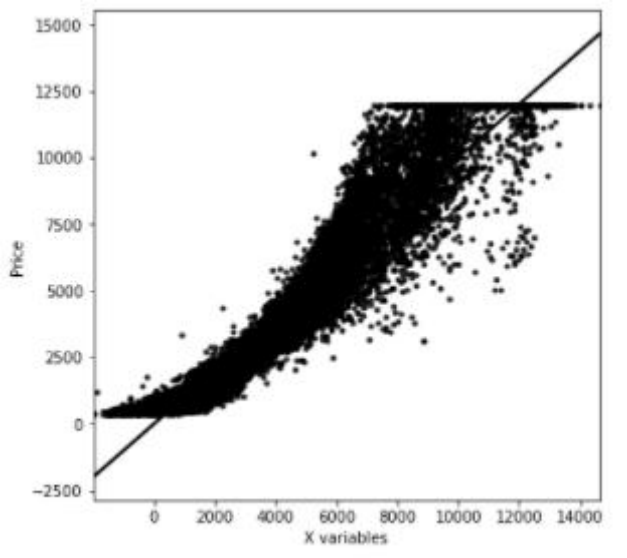

Figure 18 Scatter plot of predicted price based of X variable of model 3

Multiple Linear Regression using Stats Model 4

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	6.910e+04
Date:	Sun, 20 Feb 2022	Prob (F-statistic):	0.00
Time:	02:41:08	Log-Likelihood:	-2.1652e+05
No. Observations:	26261	AIC:	4.330e+05
Df Residuals:	26255	BIC:	4.331e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1171.3370	352.120	3.327	0.001	481.163	1861.511
carat	7945.4466	14.055	565.310	0.000	7917.898	7972.995
color	272.2507	3.526	77.219	0.000	265.340	279.161
clarity	452.7357	3.768	120.163	0.000	445.351	460.121
depth	-54.0350	4.257	-12.695	0.000	-62.378	-45.692
table	-47.1703	2.844	-16.589	0.000	-52.744	-41.597

Omnibus:	3243.589	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11961.446
Skew:	0.597	Prob(JB):	0.00
Kurtosis:	6.083	Cond. No.	5.23e+03

Table 20 OLS regression model 4 result

Multiple Linear Regression using Stats Model 5

OLS Regression Results

Dep. Variable:	price	R-squared:	0.929
Model:	OLS	Adj. R-squared:	0.929
Method:	Least Squares	F-statistic:	6.910e+04
Date:	Sun, 20 Feb 2022	Prob (F-statistic):	0.00
Time:	02:40:03	Log-Likelihood:	-2.1652e+05
No. Observations:	26261	AIC:	4.330e+05
Df Residuals:	26255	BIC:	4.331e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1171.3370	352.120	3.327	0.001	481.163	1861.511
carat	7945.4466	14.055	565.310	0.000	7917.898	7972.995
color	272.2507	3.526	77.219	0.000	265.340	279.161
clarity	452.7357	3.768	120.163	0.000	445.351	460.121
depth	-54.0350	4.257	-12.695	0.000	-62.378	-45.692
table	-47.1703	2.844	-16.589	0.000	-52.744	-41.597

Omnibus:	3243.589	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11961.446
Skew:	0.597	Prob(JB):	0.00
Kurtosis:	6.083	Cond. No.	5.23e+03

Table 21 OLS regression model 5 result

Checking Multicollinearity using Variance Inflation Factor (VIF)

Variance Inflation Factor (VIF) is one of the methods to check if independent variables have correlation between them. If they are correlated, then it is not ideal for linear regression models as they inflate the standard errors which in turn affect the regression parameters. As a result, the regression model becomes non-reliable and lacks interpretability.

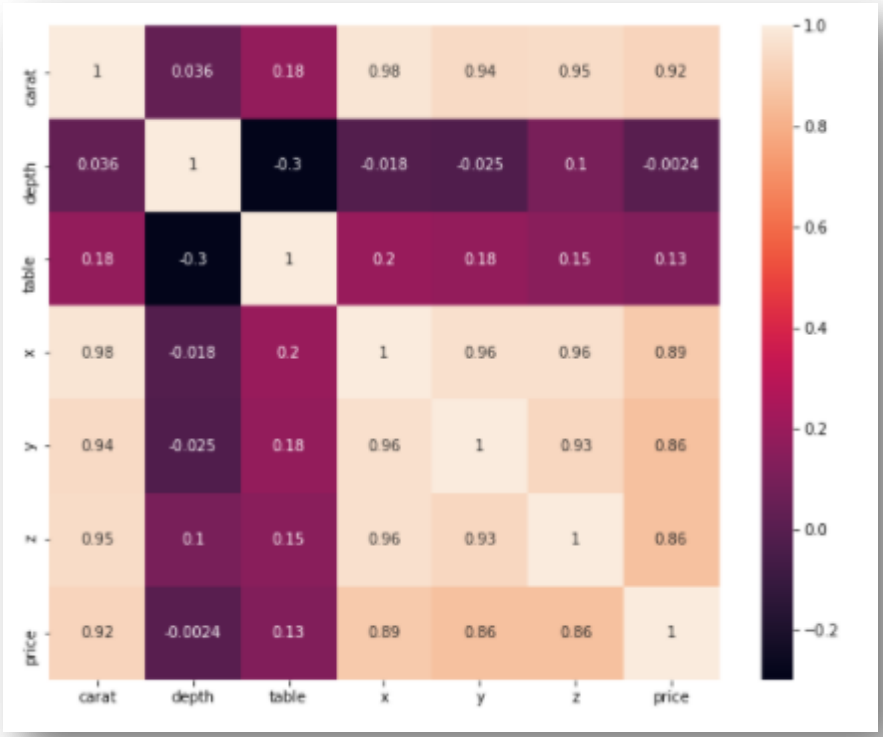

Figure 19 Heat map showing the correlation among variables.

```
carat ---> 123.4352660994068
cut ---> 10.51503773050019
color ---> 5.5596202385955875
clarity ---> 5.465442774083856
depth ---> 1272.550507399069
table ---> 902.0106597342918
x ---> 10758.780860893887
y ---> 9463.361634778345
z ---> 4026.439239231249
```

Table 22 VIF table

General rule of thumb: If VIF values are equal to 1, then that means there is no multicollinearity. If VIF values are equal to 5 or exceedingly more than 5, then there is moderate multicollinearity. If VIF is 10 or more, then that means there is high collinearity.

From the above we can conclude that variables carat, cut, clarity, depth, table, x, y and z has high multicollinearity whereas variable color has moderate correlation. However variables cut, color and clarity are categorical variables which are transformed to numerical using encoding. Hence it is difficult to say whether the VIF values indicate the right results

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Model number	Package	Outlier Treatment	Formulae	R-Squared	Adjusted R-square	Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)
Model1	SK Learn	yes	'price ~ carat+color+clarity+depth+table+x+y+z'	0.93	NA	853708.99	923.96
Model2	Stat model	yes	price ~ carat+color+clarity+depth+table+x+y+z'	0.93	0.93	NA	NA
Model3	Stat model	yes	'price ~ carat+color+clarity+depth+table+x'	0.93	0.93	NA	NA
Model4	Stat model	yes	price ~ carat+x'	0.929	0.929	NA	NA
Model5	Stat model	yes	'price ~ carat+color+clarity+depth+table'	0.929	0.929	NA	NA

Table 23 Model Performance

Price = (3852.73) * Intercept + (8894.7) * carat + (274.04) * color + (443.55) * clarity + (-71.41) * depth + (-48.05) * table + (-399.42) * x +.

Business Insights and Recommendations

1. The highest **R-squared** is **93.00%** which we achieved by using **stats_model** post outlier treatment.
2. Model 1 and 2 has better R-squared values.
3. We can see that there is **high multicollinearity** in the dataset.
4. **Intercept** of the model is **3852.73.**
5. R-squared **93.00%** shows a good accuracy which means **93%** of the price is explained by the model.
6. RMSE on testing data is **923.96.**
7. As per the above graph (Figure23) there is a **strong linear relationship** between the actual and predicted values with some noise to an extent which signifies the unexplained variance.
8. Increase in **carat weight** of the diamond will **increase the price** of the diamond considerably.

9. Width (y) of the diamond in mm also plays an important factor. As the **width** increases the **price** also increases.
10. **Brighter** the color of the diamond, the price **increases.**
11. Gem Stones Ltd should work on **carat, color** and **width** of the diamonds which are strong contributors for price.
12. Since **x, y and z had zero values** we have removed them as they are clear indicators that they have been incorrectly captured as **bad data.** It's a **data input error** because there cannot be diamonds with 0.00mm length, width and height.
13. Gem Stones Ltd can collect more data in future which helps in building more robust models for price prediction. If Gem Stones Ltd wants to use the model without outlier treatment, they can as well use the Model2 which predicts the price **93.00%** accurately.

Problem 2: Logistic Regression And LDA

You Are Hired By A Tour And Travel Agency Which Deals In Selling Holiday Packages. You Are Provided Details Of 872 Employees Of A Company. Among These Employees, Some Opted For The Package And Some Didn't. You Have To Help The Company In Predicting Whether An Employee Will Opt For The Package Or Not On The Basis Of The Information Given In The Data Set. Also, Find Out The Important Factors On The Basis Of Which The Company Will Focus On Particular Employees To Sell Their Packages.

Data Dictionary:-

Variable Name	Description
Holiday_Package	Opted For Holiday Package Yes/No?
Salary	Employee Salary
Age	Age In Years
Edu	Years Of Formal Education
No_young_children	The Number Of Young Children (Younger Than 7 Years)
No_older_children	Number Of Older Children
Foreign	Foreigner Yes/No

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of employees who either opted for holiday package or not and other attributes of almost 872 observations. We will perform exploratory data analysis to understand what the given data has to say and then use logistic regression and linear discriminant analysis (LDA) techniques to predict whether an employee will opt for the holiday package based on the independent attributes we have.

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.

Loading all necessary libraries for the model building. Reading and visualizing the data sample(head), Data type, Information of the dataset, shape, EDA.

	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

Table 24 Data info

We will be removing Unnamed: 0 column as it adds no value for our analysis.

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

Table 25 Null Values checked

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

Table 26 Info of the dataset.

The data set has 872 rows. Column indicating row number (Unnamed:0) cannot be used for analysis and needs to be deleted. Excluding row number data set has 2 object variables and 5 numerical variables. i.e. 7 variables available for analysis. 'Holliday_Package' is dependent variable and other 6 independent (predictive variables). There are no null values and duplicate values in the data set.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.000000	872.000000	872.000000	872.000000	872.000000	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

Table 27 Descriptive Statistics

Univariate Analysis

Figure 20Barplot of foreign & Salary and Holiday Package

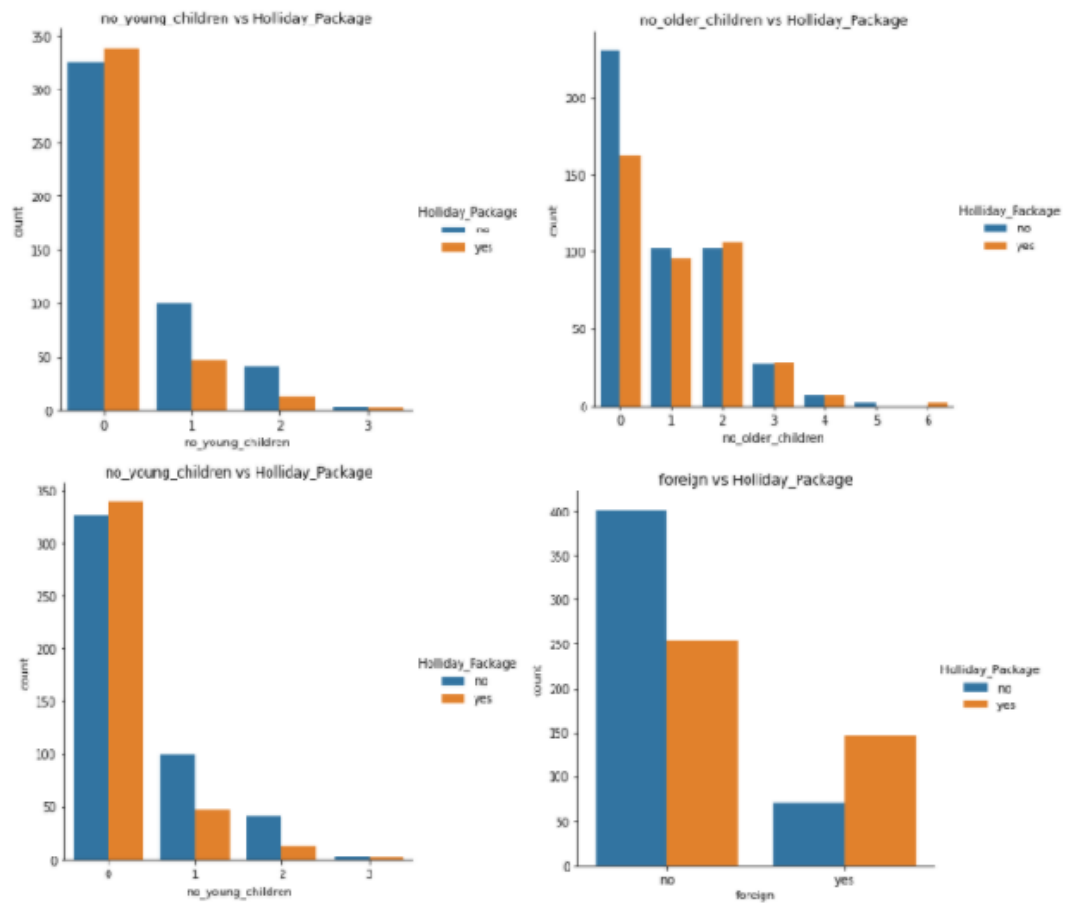

Figure 21 Count plot of different numerical variables based on holiday package

Bivariate analysis.

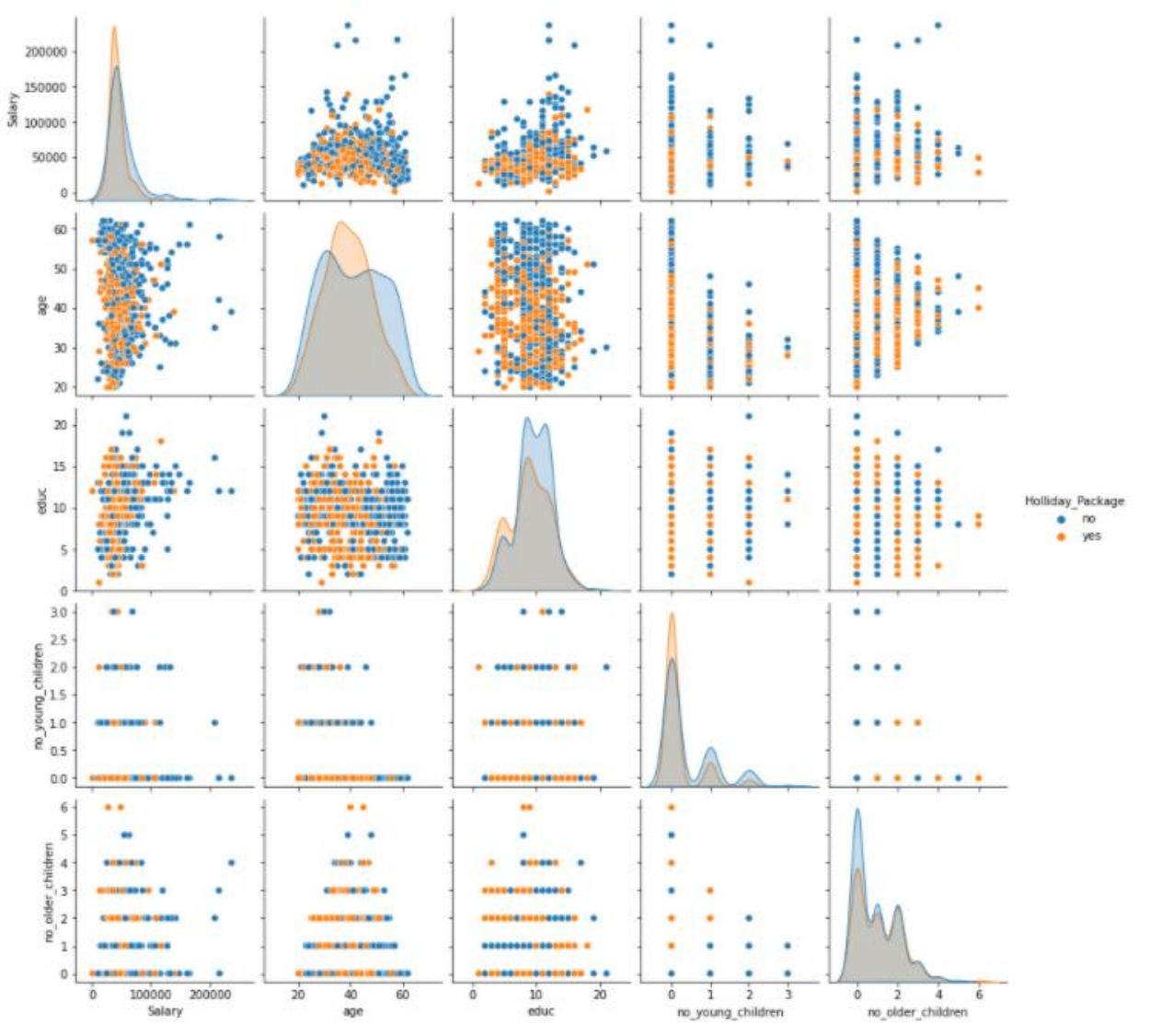

Figure 22 Pair plot

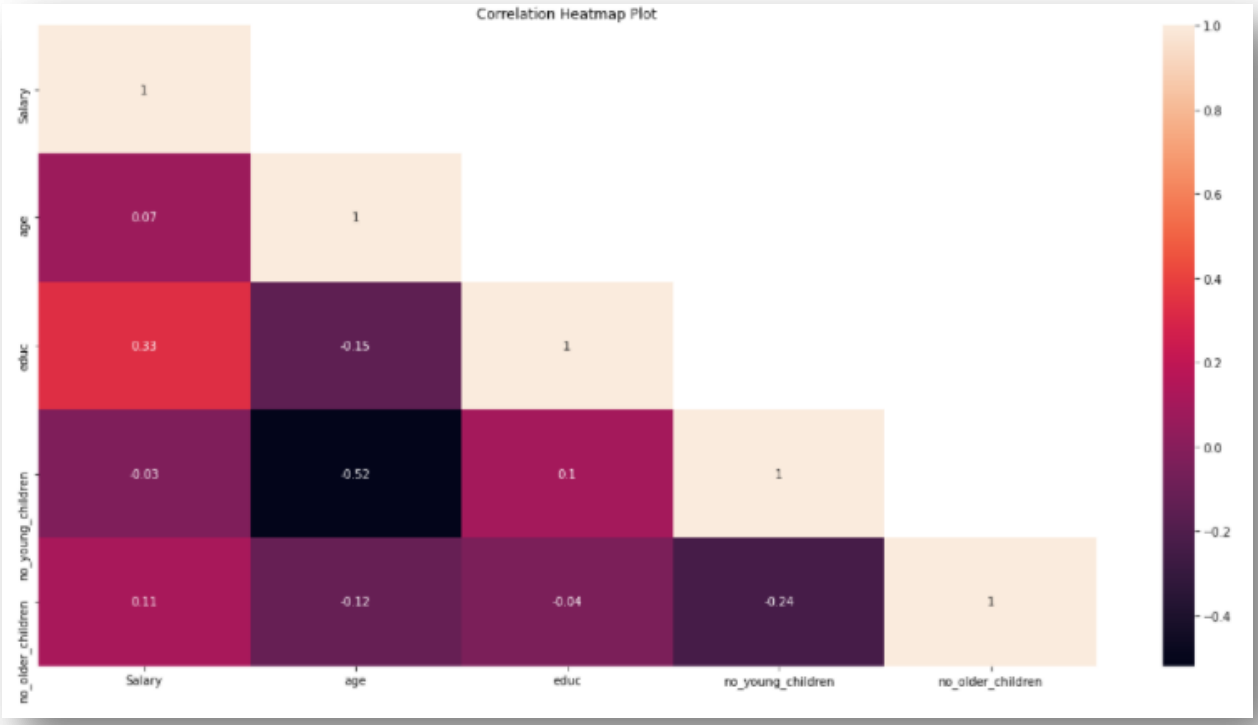

Figure 23 Correlation Heatmap

Correlation heatmap clearly evidences that there exists NO correlation between the attributes. However age and no_young_children have a less than moderate negative relationship with a correlation coefficient of -0.52 which can be considered as trivial. Correlation values are always between 1 and -1. Those which are closer to 1 are positively correlated and those which near -1 are negatively correlated. Values near to 0 have no correlation.

Boxplot of Numerical Vriables

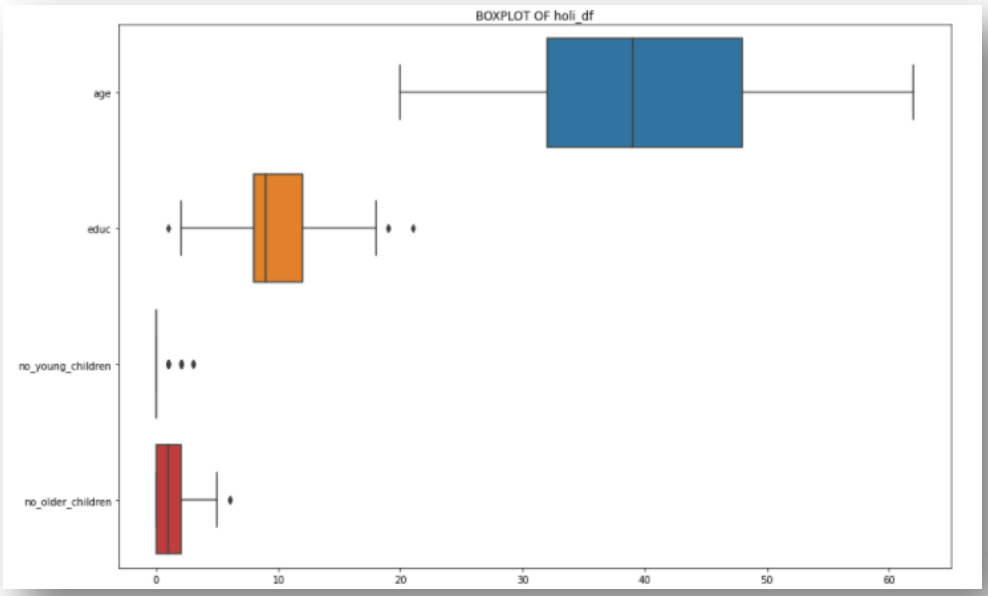

Figure 24 Before the trimming of variables

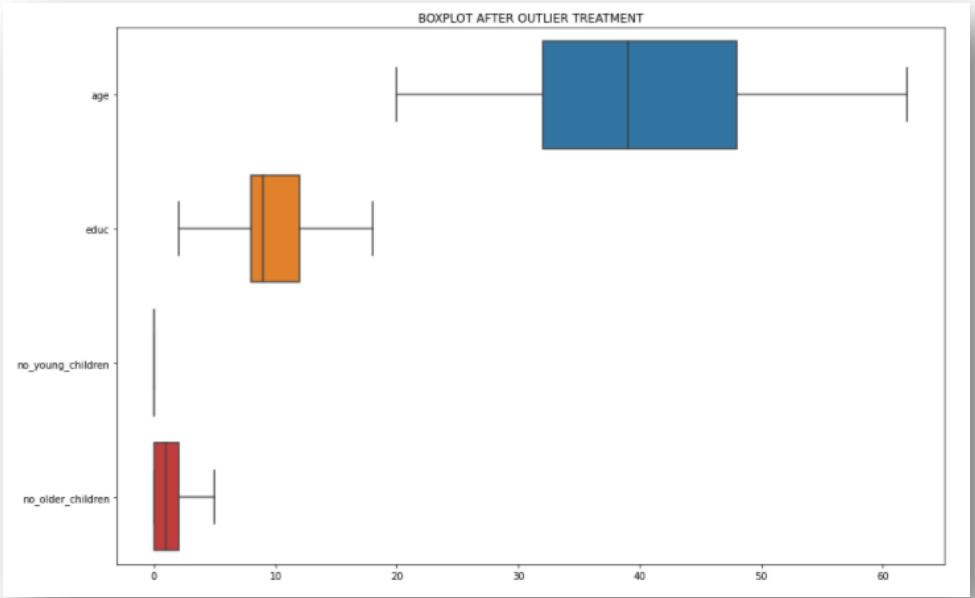

Figure 25 Box plotsthe variables after trimming

From the above plot we can see that salary is centered around the range of 30 yrs to 50 yrs of age. Employees who are more than 50 yrs have opted for holiday package which suggests that they are in the verge of retiring and hence are taking a break from their career

Similarly employees who have NO older children have a higher salary range. However the median salary range remains almost same for most of the employees who have older children.

Employees who have older children between 0 to 3 have a higher salary range whereas those who have more than 3 older children are in the moderate salary range. Again non foreign employees are higher.

Employees who have NO younger children have a higher salary range. Non foreign employees are getting more salary compared to foreign employees.

Employees with 11yrs to 17yrs of education are more in numbers which indicates that they are mostly under graduates or post graduates. Employees who are non-foreign with 12 yrs of education have the highest salary range.

Employees who have not opted for holiday package are more in numbers but we can see outliers across range of age. Non-foreign employees are higher in number.

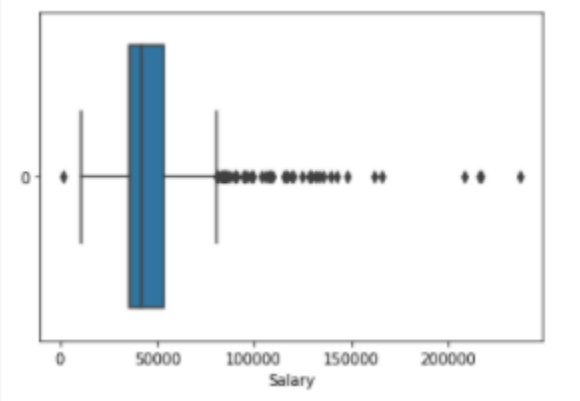

Figure 27 Salary with outliers

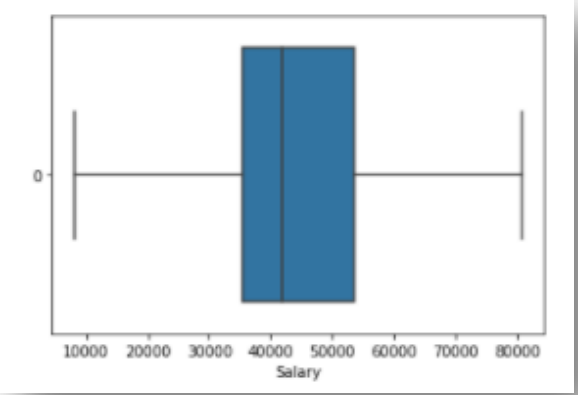

Figure 26 Salary with out outliers

If employee is foreigner and employee not having young children, chances of opting for Holiday Package is good. Independent variables are not correlated with other variables. Salary has some outliers.

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).

```
feature: Holliday_Package
['no', 'yes']
Categories (2, object): ['no', 'yes']
[0 1]


feature: foreign
['no', 'yes']
Categories (2, object): ['no', 'yes']
[0 1]
```

Table 28 Unique values of holiday package and foreign

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412.0	30.0	8.0	0.0	1.0	0
1	1	37207.0	45.0	8.0	0.0	1.0	0
2	0	58022.0	46.0	9.0	0.0	0.0	0
3	0	66503.0	31.0	11.0	0.0	0.0	0
4	0	66734.0	44.0	12.0	0.0	2.0	0

Table 30 Data after encoding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    int8
 1   Salary             872 non-null    float64
 2   age                872 non-null    float64
 3   educ               872 non-null    float64
 4   no_young_children  872 non-null    float64
 5   no_older_children  872 non-null    float64
 6   foreign            872 non-null    int8
dtypes: float64(5), int8(2)
memory usage: 35.9 KB
```

Table 29 Data info

Split the data into train and test (70:30).

```
The training set for the independent variables: (610, 6)
The training set for the dependent variable: (610, 1)
The test set for the independent variables: (262, 6)
The test set for the dependent variable: (262, 1)
```

Table 31 Split X and y into training and test set in 70:30 ratio

Logistic Regression and LDA (linear discriminant analysis).

The Accuracy LDA model is 0.53 for train and 0.54 for test data. Continuing in the later part of the answer..

Applying GridSearchCV for Logistic Regression

After the LDA model when we used GridSearchCV - The accuracy for train data increased to 0.63 whereas for test data it is 0.65. The model has performed slightly well with these parameters compared to the previous model. Model code performed in code file.

Statsmodels is a Python module which provides various functions for estimating different statistical models and performing statistical test. First, we define the set of dependent(y) and independent(X) variables. If the dependent variable is in non-numeric form, it is first converted to numeric using encoding.

```
Intercept           0.345881
Salary             -0.000019
age                -0.014465
educ                0.049285
no_older_children   0.197193
foreign             1.099271
dtype: float64
```

Table 32 Numeric statistical encoding of the model

Logit Regression Results

Dep. Variable:	Holliday_Package	No. Observations:	610
Model:	Logit	Df Residuals:	604
Method:	MLE	Df Model:	5
Date:	Sun, 20 Feb 2022	Pseudo R-squ.:	0.06391
Time:	15:53:42	Log-Likelihood:	-394.03
converged:	True	LL-Null:	-420.93
Covariance Type:	nonrobust	LLR p-value:	2.300e-10

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.3459	0.535	0.646	0.518	-0.703	1.395
Salary	-1.91e-05	6.01e-06	-3.179	0.001	-3.09e-05	-7.33e-06
age	-0.0145	0.008	-1.711	0.087	-0.031	0.002
educ	0.0493	0.034	1.462	0.144	-0.017	0.115
no_older_children	0.1972	0.082	2.392	0.017	0.036	0.359
foreign	1.0993	0.220	4.993	0.000	0.668	1.531

Table 33 Logistic regression model results

The summary table below, gives us a descriptive summary about the regression results. Foreign have coeffient of 1.099 which shows foreign is important independent variable feature. Foreign and salary having p-value <0.05 .they are statistically significant.

LDA algorithm performed with accuracy in the code file, the accuracies are: -

	precision	recall	f1-score	support
0	0.62	0.80	0.70	329
1	0.65	0.44	0.52	281
accuracy			0.63	610
macro avg	0.64	0.62	0.61	610
weighted avg	0.64	0.63	0.62	610

Table 34 LDA Classification matrix

```
array([[263,  66],
       [158, 123]], dtype=int64)
```

Table 35 LDA confusion matrix

The accuracy for train data is 0.632 whereas for test data it is 0.656.

2.3 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.

Logistic regression

Figure 28 Logistic regression Confion matrix for traing anf test data

	0	1
0	0.277645	0.722355
1	0.362880	0.637120
2	0.539021	0.460979
3	0.336028	0.663972
4	0.341973	0.658027

	0	1
0	0.637044	0.362956
1	0.564987	0.435013
2	0.651620	0.348380
3	0.557207	0.442793
4	0.537535	0.462465

Table 36 Getting probability of traing anf test data

	precision	recall	f1-score	support
0	0.63	0.78	0.70	329
1	0.65	0.46	0.54	281
accuracy			0.63	610
macro avg	0.64	0.62	0.62	610
weighted avg	0.64	0.63	0.62	610

	precision	recall	f1-score	support
0	0.64	0.82	0.72	142
1	0.69	0.46	0.55	120
accuracy			0.66	262
macro avg	0.67	0.64	0.64	262
weighted avg	0.66	0.66	0.64	262

Table 37 Classification Report of training and test data

```
The coefficient for Salary is -1.9113923814581458e-05
The coefficient for age is -0.014656494359458759
The coefficient for educ is 0.04608845740689346
The coefficient for no_young_children is 0.0
The coefficient for no_older_children is 0.19467583759199986
The coefficient for foreign is 1.048644975581004
```

Table 38 Coefficients for all variables

```
AUC for the Training Data: 0.661
AUC for the Test Data: 0.675
```

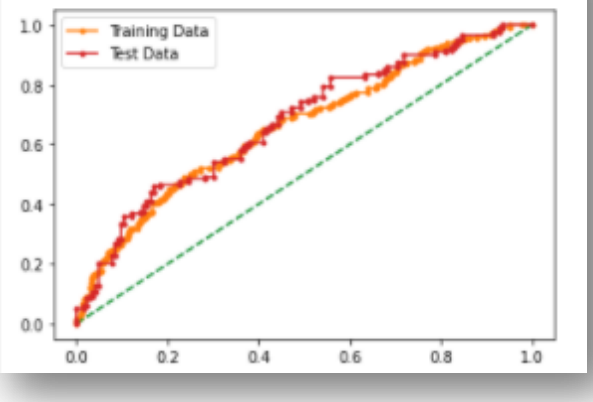

Figure 29 Auc_Roc curve of Logistic regression

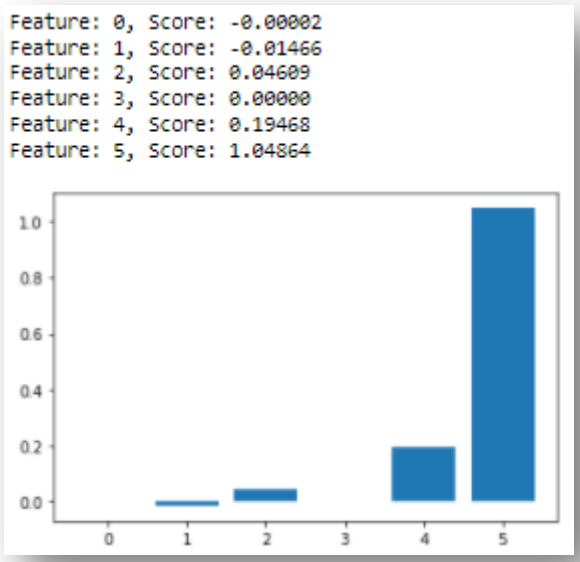

Figure 30 ploting feature importance

```
Salary ---> 10.691328770445477
age ---> 7.883717090792149
educ ---> 9.289879244969368
no_young_children ---> nan
no_older_children ---> 1.8287587912872862
foreign ---> 1.3123489821120458
```

Table 39 VIF Table

LDA

LDA is very similar to PCA (Principal Component Analysis). PCA is an unsupervised technique which is used for dimensionality reduction such that maximum variance is retained. LDA is a supervised learning technique used for classification problems. LDA uses a linear combination of data points such that the two classes are well separated i.e, the mean of class 1 is far away from the mean of class 2 however the variance of class 1 closer to mean of class 1 and similarly variance of class 2 is closer to mean of class 2. This will clearly separate the classes into two groups. LDA is not used for dimensionality reduction.

```
confusion_matrix(y_train, pred_class)
```

```
array([[263,  66],
       [158, 123]], dtype=int64)
```

```
confusion_matrix(y_test, pred_class2)
```

```
array([[118,  24],
       [ 67,  53]], dtype=int64)
```

Table 40 LDA Confusion Matrix for train and test data

	precision	recall	f1-score	support
0	0.62	0.80	0.70	329
1	0.65	0.44	0.52	281
accuracy			0.63	610
macro avg	0.64	0.62	0.61	610
weighted avg	0.64	0.63	0.62	610

	precision	recall	f1-score	support
0	0.64	0.83	0.72	142
1	0.69	0.44	0.54	120
accuracy			0.65	262
macro avg	0.66	0.64	0.63	262
weighted avg	0.66	0.65	0.64	262

Table 41 : LDA Classification Report of training and test data

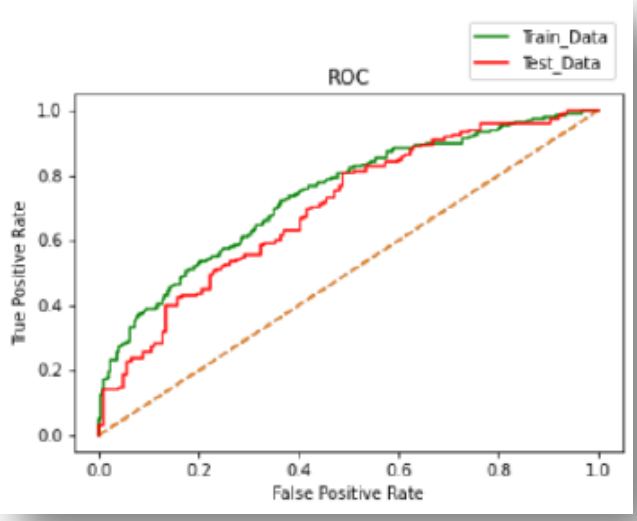

Figure 31 LDA AUC_ROC

Comparison

Accuracy score both in Training and Testing data is higher in case of Logistic Regression (LR) compared to LDA. The data set has outliers in 'salary' LR is more robust predictor in case of outliers. Therefore, it is recommended to use Logistic Regression (LR).

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

If employee is foreigner and employee not having young children, chances of opting for Holiday Package is good. Special offer can be designed to domestic employees to opt for Holiday Package.
Many high salary employees are not opting for Holiday Package, company can focus on high salary employees to sell Holiday Package. Employees having older children are not opting for Holiday Package. Age of the employee is not a material in opting for holiday package.
It can be observed from coefficient arrived from both models that opting for Holiday package has strong negative relation with number of young children. Holiday packages can be modified to make infant and young children friendly to attract more employees having young children.

We had a business problem where we need predict whether an employee would opt for a holiday package or not, for this problem we had done predictions both logistic regression and linear discriminant analysis. Since both are results are same.
The EDA analysis clearly indicates certain criteria where we could find people aged above 50 are not interested much in holiday packages.
So this is one of the we find aged people not opting for holiday packages.
People ranging from the age 30 to 50 generally opt for holiday packages.
Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package.

The important factors deciding the predictions are salary, age and educ. Recommendations:-

1. To improve holiday packages over the age above 50 we can provide religious destination places.
2. For people earning more than 150000 we can provide vacation holiday packages.
3. For employee having more than number of older children we can provide packages in holiday vacation places.

THE END…