PREDICTIVE MODELING
PROJECT REPORT

Table of Contents

Problem 1: Linear Regression

You are hired by a company named Gem Stones Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of a diamond).

Your objective is to accurately predict prices of the zircon pieces. Since the company profits at a different rate at different price levels, for revenue management, it is important that prices are predicted as accurately as possible. At the same time, it is important to understand which of the predictors are more important in determining the price.

Following is a guideline for developing a solution:

1. The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. Since this is a regression problem, the dependence of the response on the predictors needs to be thoroughly investigated.

2. Use the Pre-processed Full Data to develop a model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors?

3. Alternatively, if prediction accuracy of the price is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.
Write a project report on the solution method, clearly highlighting the benefits of all the approaches. Your report must indicate the logic of model selection and business interpretation.

Data Dictionary for Cubic Zirconia:

1. Carat: Carat weight of the cubic zirconia
2. Cut: Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. Colour: Colour of the cubic zirconia. D being the best and J the worst.
4. Clarity: Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg price) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
5. Depth: The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6. Table: The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7. Price: The Price of the cubic zirconia.

8. X: Length of the cubic zirconia in mm.
9. Y: Width of the cubic zirconia in mm.
10. Z: Height of the cubic zirconia in mm.

Solution:

Checking the first 5 rows of the data:

To start with the analysis, let's look at the sample data and perform basic checks.

	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

Dropping Unnamed: 0 column as it is index and we don't have any use in building models:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Checking the dimensions of data:

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

Inferences:

1. The dataset has a total of six independent variables - All are continuous, one target variable and three categorical variables
2. Shape (dimension) of the Dataset is (26967, 10)
3. There are 697 NULL values present in the dataset
4. 34 duplicate values are present in the dataset
5. There are a total of 10 variables and 26967 records
6. We can choose to drop the rows with missing values
7. However, here we are not dropping the rows
8. We are replacing the missing values with median value and dropping the duplicate rows

Exploratory Data Analysis:

Univariate Analysis:

To perform Univariate analysis on continuous variables, let us start with looking at the summary statistics of the dataset.

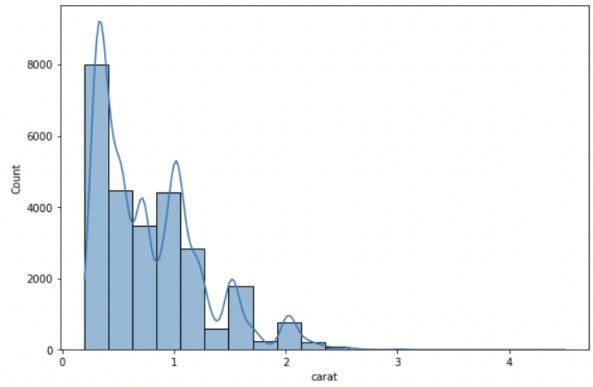
	carat	depth	table	x	y	z	price
count	26933.000000	26933.000000	26933.000000	26933.000000	26933.000000	26933.000000	26933.000000
mean	0.798010	61.746701	57.455950	5.729346	5.733102	3.537769	3937.526120
std	0.477237	1.393875	2.232156	1.127367	1.165037	0.719964	4022.551862
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.100000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.700000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5356.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

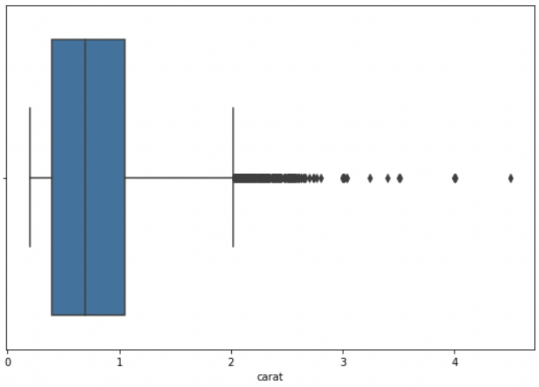
	count	mean	std	min	25%	50%	75%	max
carat	26933.0	0.798010	0.477237	0.2	0.40	0.70	1.05	4.50
depth	26933.0	61.746701	1.393875	50.8	61.10	61.80	62.50	73.60
table	26933.0	57.455950	2.232156	49.0	56.00	57.00	59.00	79.00
x	26933.0	5.729346	1.127367	0.0	4.71	5.69	6.55	10.23
y	26933.0	5.733102	1.165037	0.0	4.71	5.70	6.54	58.90
z	26933.0	3.537769	0.719964	0.0	2.90	3.52	4.04	31.80
price	26933.0	3937.526120	4022.551862	326.0	945.00	2375.00	5356.00	18818.00

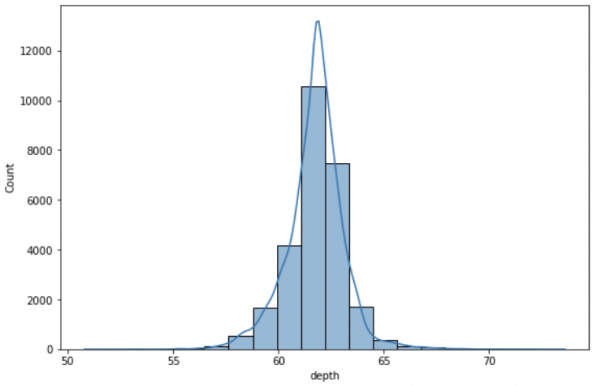
From above table:

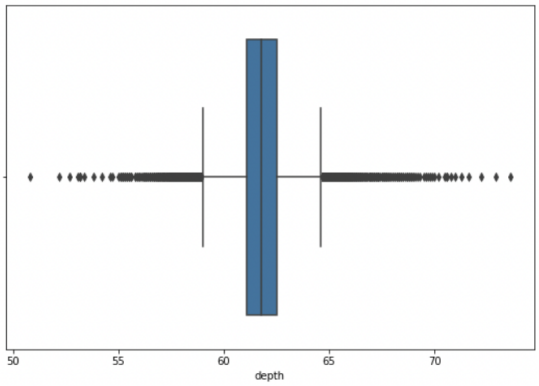
- x,y,z are length, width and height respectively
- length, width and height cannot be '0' practically as it becomes dimensionless
- From the above table, we can see that the minimum value of x,y,z seems to be '0'
- We removed the rows with x=0,y=0,z=0 and continue with plotting

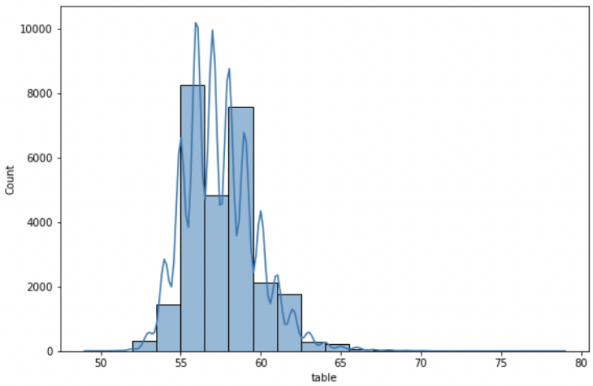
Plotting Box plots and hist plots:

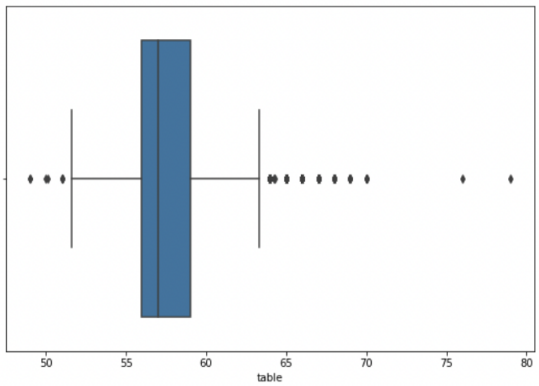


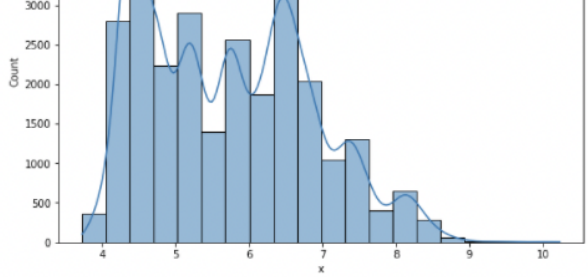


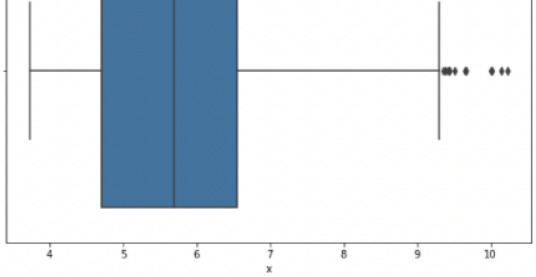


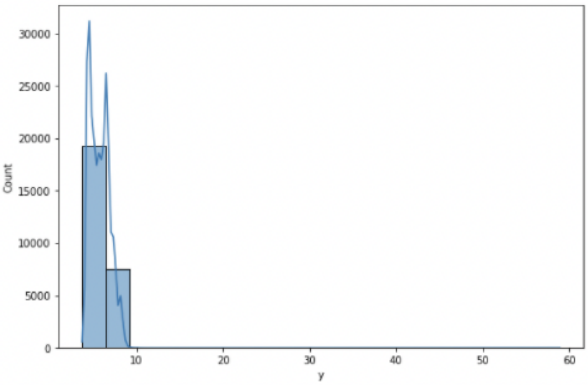


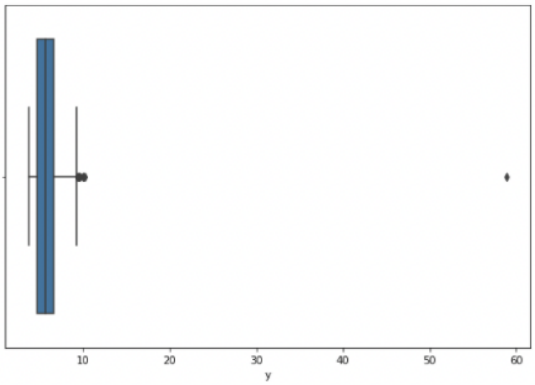


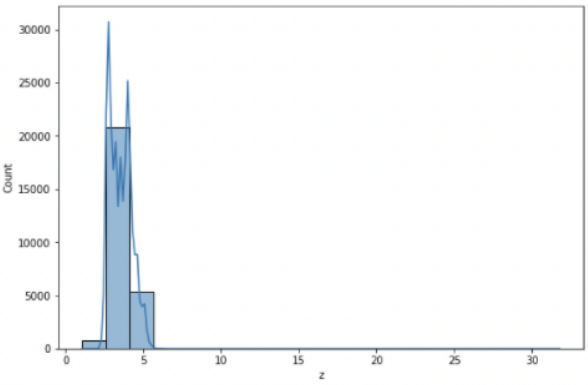


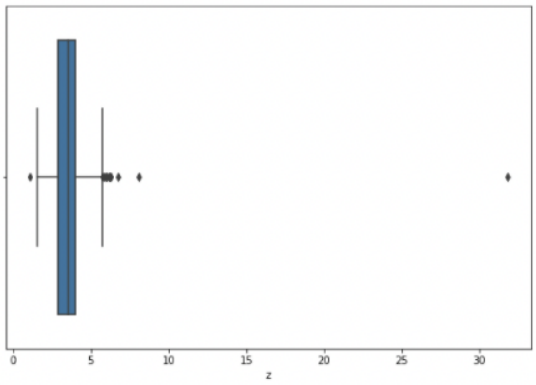

Count
x

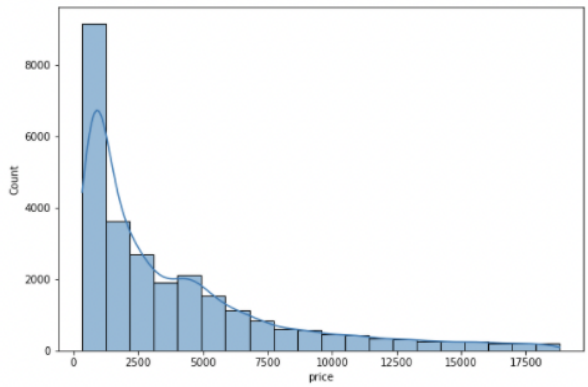

x

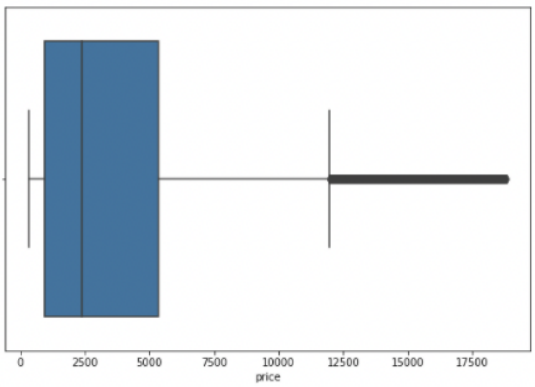

Count
y


y


Count
z


z


Count
price


price

Percentage of Outliers:

	Outlier%
carat	2.432683
clarity	0.000000
color	0.000000
cut	0.000000
depth	5.240483
price	0.000000
table	1.177344
x	0.044568
y	0.044568
z	0.051996

- Three columns have outlier values with a Percentage of 1.18,2.43,5.24 which is not very high as 20%
- We can choose to treat outliers, however here we are not treating outliers as there might be an exceptional case where these values are possible for the business in certain cases.
- For example, carat variable gives the carat weight of the cubic zirconia. From the dataset, we can see that the weight of 3 and more than 3 are very less in number which are shown as outliers and there is no harm using them as the business can have a cubic zirconia with large carat weight.
- As per the dataset, none of the values are so unrealistic. Hence, I am considering not to treat outliers for my analysis
- However, at the end we will build models after treating outliers just to check if that is making any difference in the performance

Kurtosis and Skewness in the Dataset:

	Skewness
carat	1.114871
depth	-0.028403
table	0.764890
x	0.402010
y	3.888607
z	2.639529
price	1.619055

	Kurtosis
carat	1.212235
depth	3.863895
table	1.579418
x	-0.720965
y	160.727513
z	88.516471
price	2.152993

Inferences on Univariate Analysis:

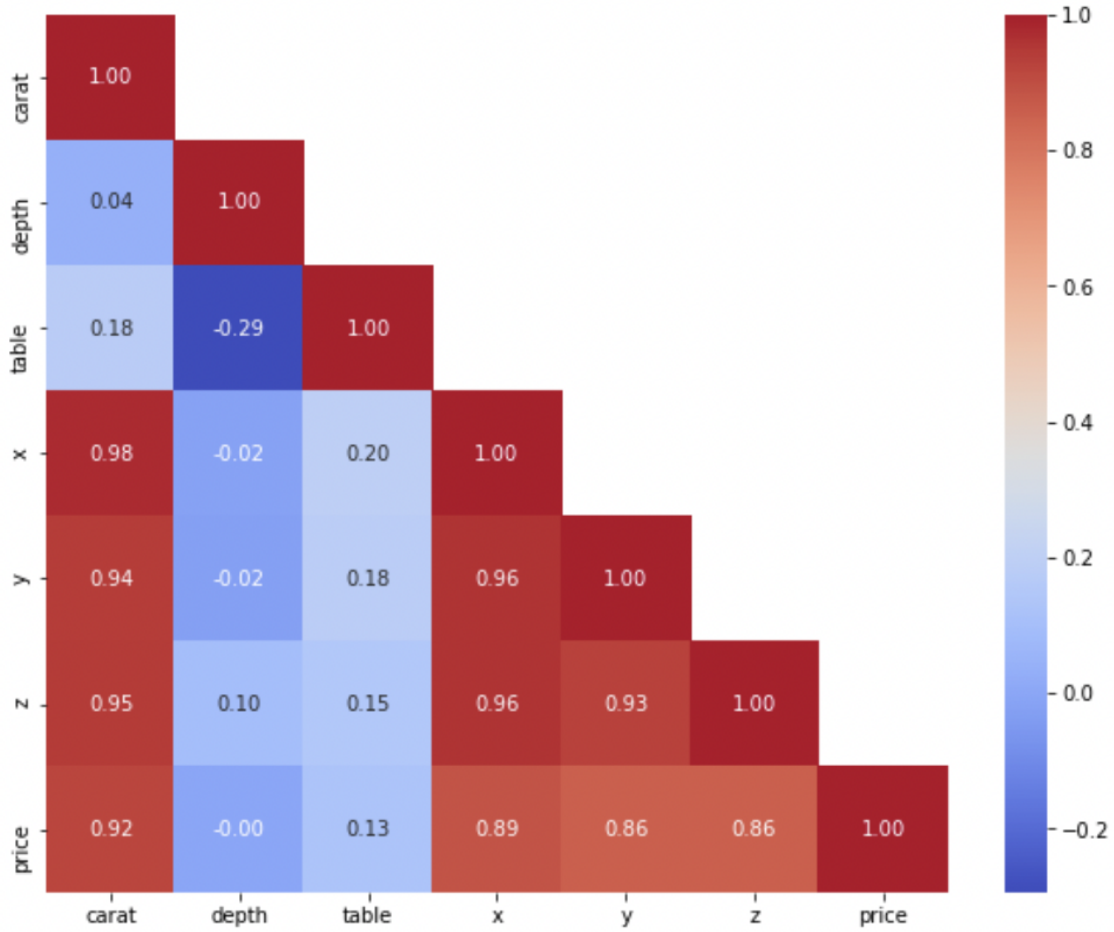
1. From the Box plots, we can conclude that there are few outliers present in x,y,z,table variables.
2. Variable depth is normally distributed and variable carat is right skewed.
3. Variables Carat, Table, x, y and z have multi-modal distributions.
4. Kurtosis values for all variables are very high, not near to normal distribution
5. There is extreme positive kurtosis seen in few variables which is > 3 which means more of the values are located in the tails of the distribution rather than around the mean.
6. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".

Multivariate analysis:

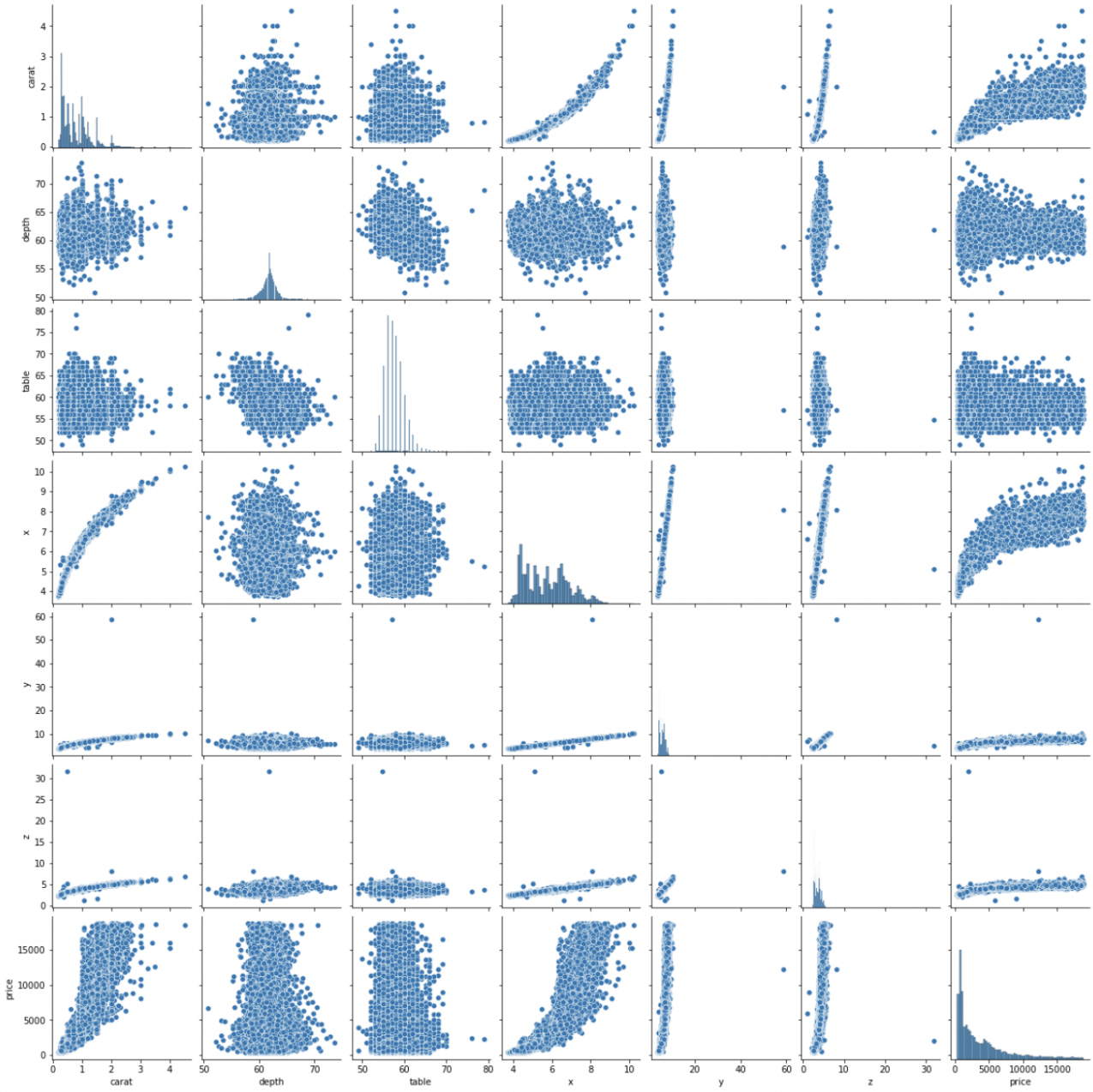
Analyzing the relationship among continuous variables by using Pair plot and Correlation Heatmap.

Checking for multicollinearity:

Heatmap:



Pair Plot:

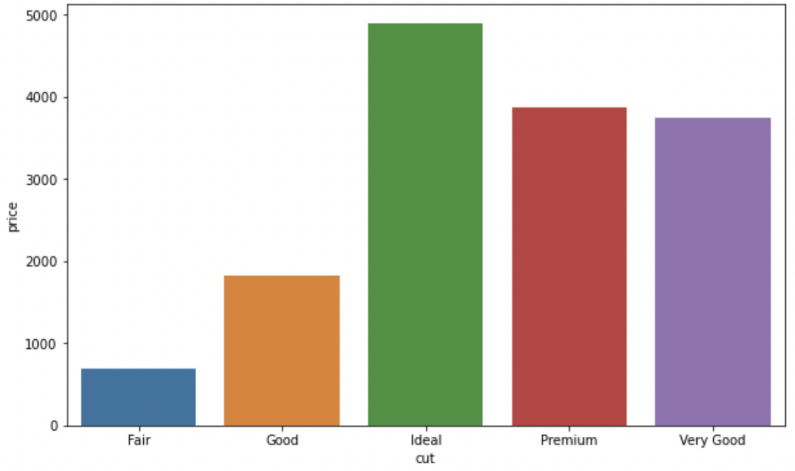


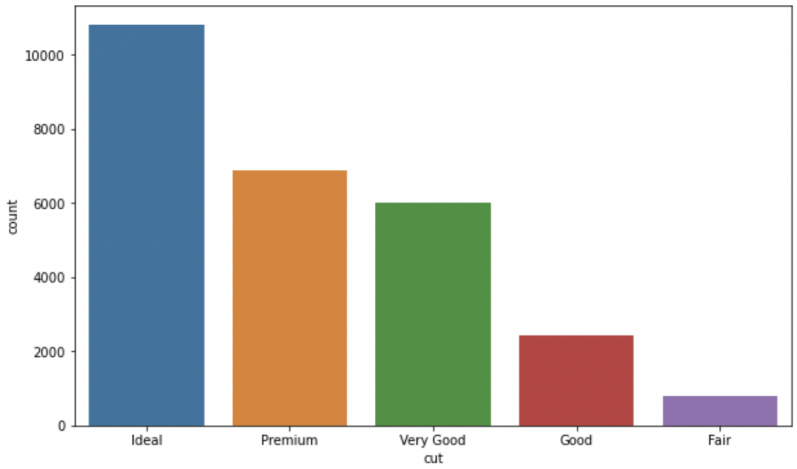
Inferences on Multivariate Analysis:

1. x,y,z variables are highly positively correlated with carat variable
2. z is highly correlated with x and y
3. y is highly correlated with x
4. Top Negative correlation pair is table and depth (-0.29)
5. From pairplot, we see that as the carat weight increases the length(x) is also increasing showing a positive relationship
6. Few of the variables have no correlation with each other, i.e., datapoints are spread wideacross(ex: carat and depth, carat and table). There might be small amount of correlation but there is no particular trend

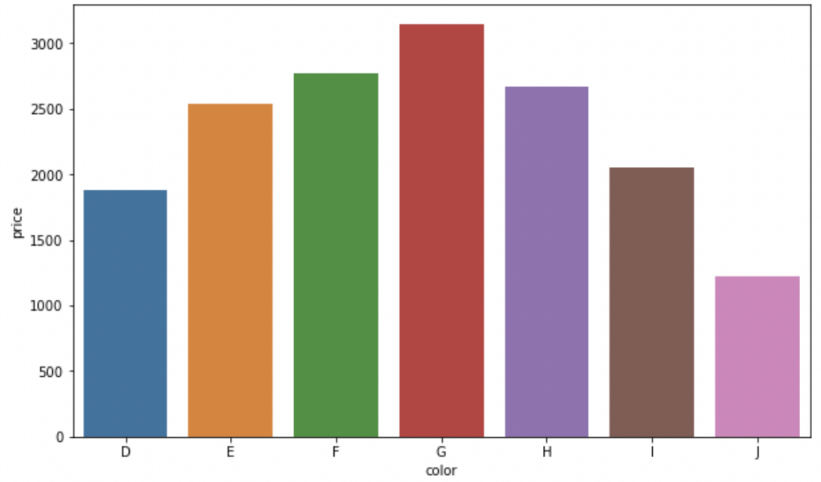
EDA for Categorical Variables: We have 3 categorical variables - cut, color, clarity

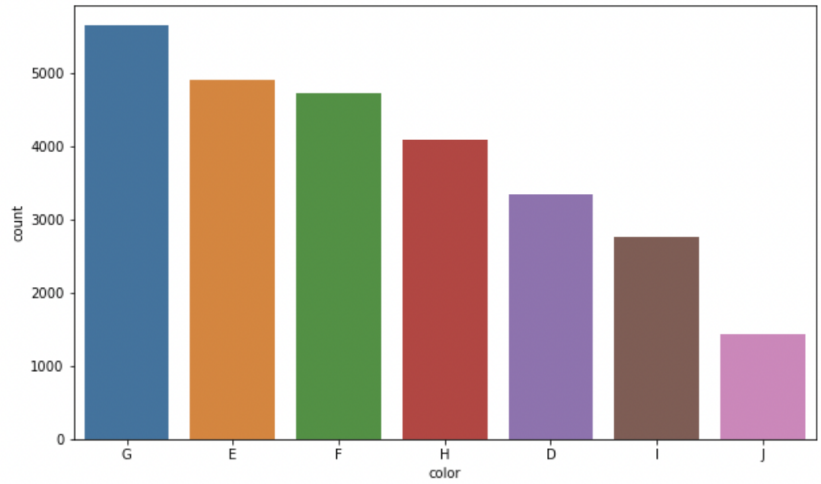
Variable Cut:



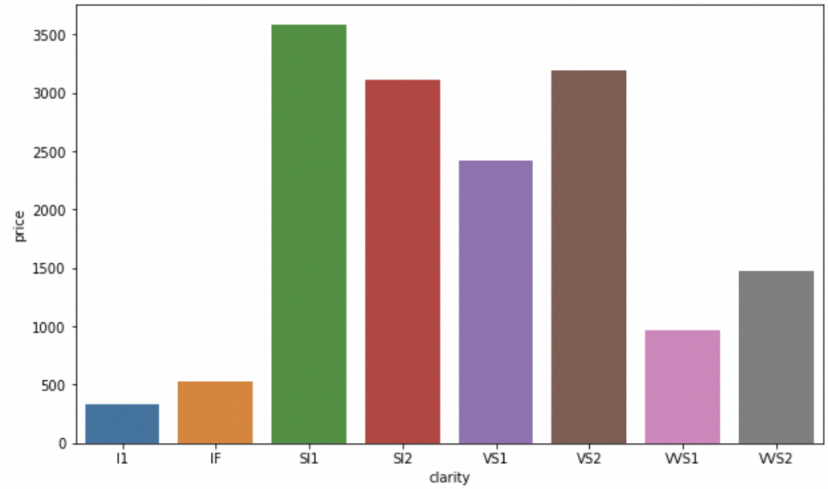


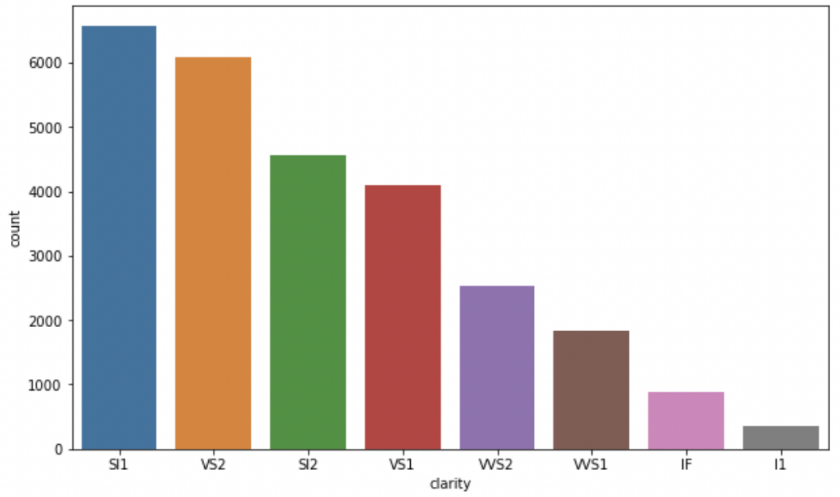
Variable Color:





Variable Clarity:





The inferences drawn from above EDA:

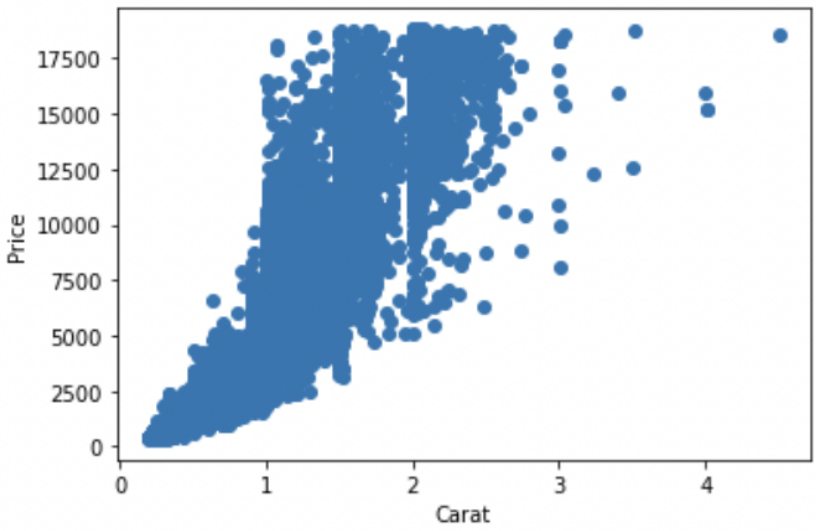
1. 'Price' is the target variable while all others are the predictors. The data set contains 26967 rows, 11 columns.
2. In the given data set there are 2 Integer type features, 6 Float type features, 3 Object type features.
3. The first column is an index ("Unnamed: 0"). As this is only serial no, we can remove it.
4. In the given data set, the mean and median values does not have much difference.
5. 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, we have filtered those as it clearly shows faulty data entries.
6. There are three object data types 'cut', 'color' and 'clarity'.
7. We can observe there are 697 missing values in the depth column. There are some duplicate rows present, nearly 0.12 % of the total data. So in this case we have dropped the duplicated rows.
8. There are significant amount of outliers present in some variables, the features with datapoints that are far from the rest of the dataset which will affect the outcome of our regression model. However, they are not unrealistic values.
9. So we can choose to treat or not treat the outliers. We can see that the distribution of some quantitative features like "carat" and the target feature "price" are heavily "right-skewed".
10. It looks like most features do correlate with the price of cubic zirconia.Observation on 'cut': The Premium Cut on Diamonds are the most Expensive, followed by Very Good Cut.
11. Scaling is not necessary, 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 in running model quickly else the starting point would be very far from minima if the scaling is not done in preprocessing.
12. For now we will process the model without scaling and we will also check the output with scaled data of regression model output to check if the performance is changing

Simple Linear Regression:

Correlation of independent variables with the target variable:

```
carat     0.922400
depth    -0.002683
table     0.126967
x         0.887467
y         0.857255
z         0.855775
price     1.000000
dtype: float64
```

- Carat variable has high correlation of 0.92
- Hence, using carat variable as independent variable and price as target variable
- Checking the distribution using scatter plot and building a simple regression model with one variable



- There is a positive correlation, as carat weight increases, price increases

Model before scaling:

OLS Regression Results

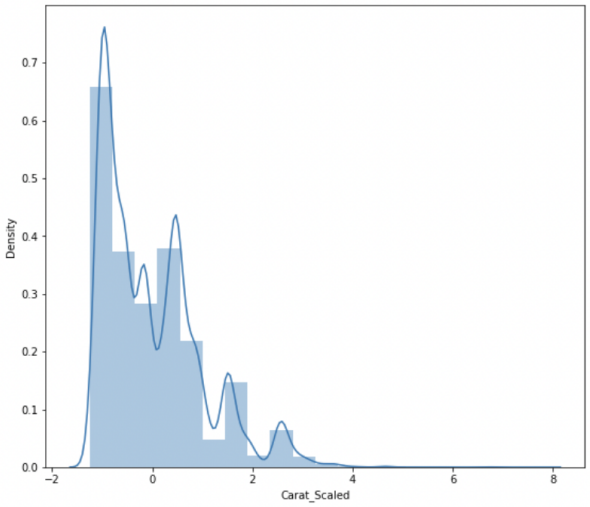
Dep. Variable:	price	**R-squared:**	0.851
Model:	OLS	**Adj. R-squared:**	0.851
Method:	Least Squares	**F-statistic:**	1.536e+05
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	21:42:48	**Log-Likelihood:**	-2.3605e+05
No. Observations:	26925	**AIC:**	4.721e+05
Df Residuals:	26923	**BIC:**	4.721e+05
Df Model:	1		
Covariance Type:	nonrobust		

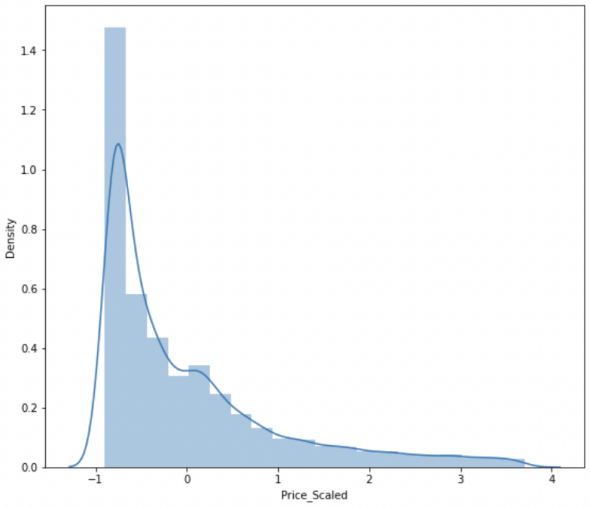
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2266.1749	18.442	-122.880	0.000	-2302.323	-2230.027
carat	7774.2063	19.839	391.859	0.000	7735.320	7813.092

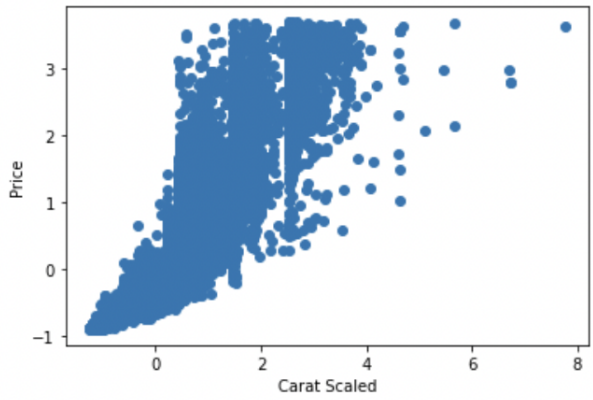
Omnibus:	6768.730	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	63464.268
Skew:	0.942	**Prob(JB):**	0.00
Kurtosis:	10.282	**Cond. No.**	3.63

- Around 85% variability in the dependent variable is being explained by the 'Carat' variable.

Simple Linear Regression on scaled data:







- The distributions are similar for both scaled and unscaled data

Model after scaling:

OLS Regression Results

Dep. Variable:	df_price_scaled	**R-squared:**	0.851
Model:	OLS	**Adj. R-squared:**	0.851
Method:	Least Squares	**F-statistic:**	1.536e+05
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	21:50:32	**Log-Likelihood:**	-12591.
No. Observations:	26925	**AIC:**	2.519e+04
Df Residuals:	26923	**BIC:**	2.520e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-3.903e-18	0.002	-1.66e-15	1.000	-0.005	0.005
df_carat_scaled	0.9224	0.002	391.859	0.000	0.918	0.927

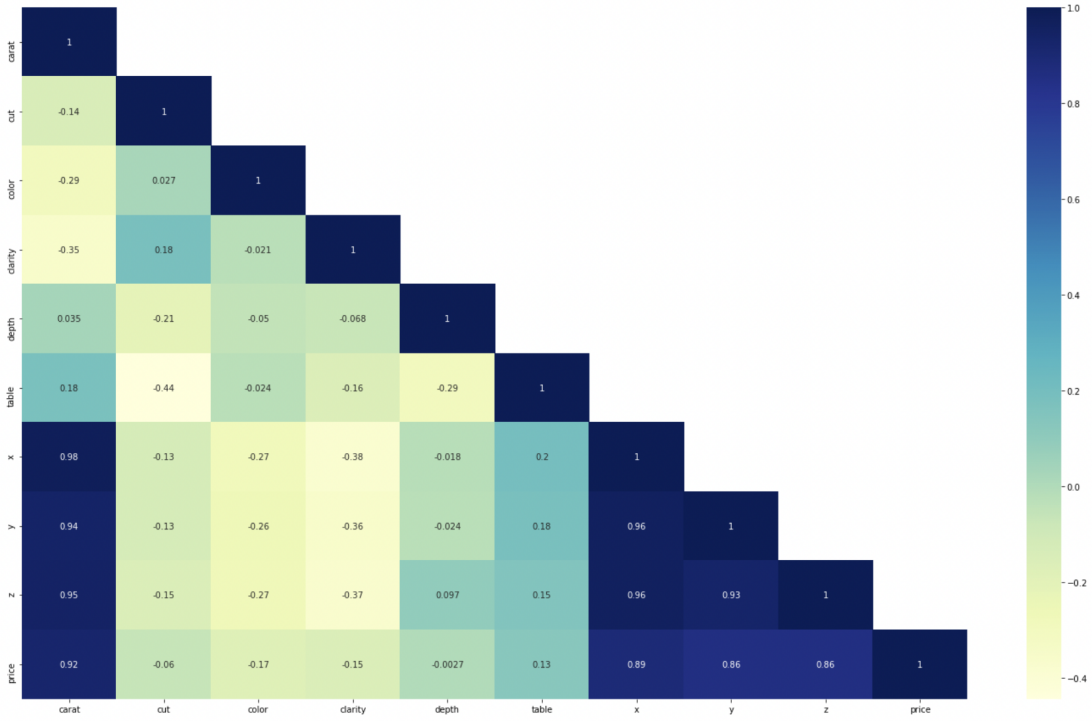
Omnibus:	6768.730	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	63464.268
Skew:	0.942	**Prob(JB):**	0.00
Kurtosis:	10.282	**Cond. No.**	1.00

- Around 85% variability in the dependent variable is being explained by the 'Carat' variable after scaling the data.
- Performance of the model is same for both scaled and unscaled data.
- We see that the R^2 value has remained the same after this transformation. We can say that scaling a variable for Linear Regression will give us the same values as compared to the unscaled variables.

Data Preprocessing: Converting all the categorical variables into numerical variables

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.30	4	5	2	62.1	58.0	4.27	4.29	2.66	499
1	0.33	3	3	7	60.8	58.0	4.42	4.46	2.70	984
2	0.90	2	5	5	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	4	4	4	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	4	4	6	60.4	59.0	4.35	4.43	2.65	779

Let us now check the correlation amongst the predictor variables just to make sure that the predictor variables are not highly correlated amongst themselves:



- There is high correlation between few variables.
- However, we will build a model with all levels of categorical variables to check the performance

Model 1: Model with all variables included

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.909
Model:	OLS	**Adj. R-squared:**	0.908
Method:	Least Squares	**F-statistic:**	2.969e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:03:50	**Log-Likelihood:**	-2.2947e+05
No. Observations:	26925	**AIC:**	4.590e+05
Df Residuals:	26915	**BIC:**	4.590e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4604.1277	605.265	7.607	0.000	3417.777	5790.479
carat	1.103e+04	77.677	141.963	0.000	1.09e+04	1.12e+04
cut	117.0032	8.141	14.372	0.000	101.046	132.960
color	328.1337	4.601	71.325	0.000	319.116	337.151
clarity	499.0037	4.994	99.921	0.000	489.215	508.792
depth	-84.5133	6.716	-12.585	0.000	-97.676	-71.350
table	-29.6312	4.195	-7.063	0.000	-37.854	-21.409
x	-958.7198	45.667	-20.994	0.000	-1048.229	-869.211
y	29.6693	23.714	1.251	0.211	-16.812	76.150
z	-56.6500	41.432	-1.367	0.172	-137.858	24.558

Omnibus:	6033.171	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	290706.193
Skew:	-0.127	**Prob(JB):**	0.00
Kurtosis:	19.095	**Cond. No.**	6.95e+03

Model 1 inferences:

- Independent variables explain around 90.8% of variation in the target variable
- The p-values for the variable 'y' and variable 'z' are high, i.e., greater than 0.05. These variables are statistically not important. But we need to understand these variables from a business point of view and then only drop the variables if required.
- We see that the conditional number is high, i.e., 6.06e+03 which clearly states that there is a problem of multicollinearity.
- We will treat multicollinearity by dropping the variables having high VIF value than threshold(2)
- If p-value is still greater than 0.05 after treating multicollinearity, we will drop features based on high p-values
- Because if p-value is higher, there is no correlation between x(independent) variable and y(target) variable
- It can be a poor predictor of target variable y
- It doesn't make sense to take that column even if it has a value for its coefficient. The coefficient value obtained might be by chance
- We will check the problem of multi-collinearity and treat it. we will check multicollinearity using variance inflation factor

VIF Values after Model 1:

carat VIF = 24.99
cut VIF = 1.49
color VIF = 1.12
clarity VIF = 1.23
depth VIF = 1.59
table VIF = 1.59
x VIF = 48.11
y VIF = 13.86
z VIF = 16.08

- Above, we calculated VIF values for all the variables.
- Since there was a presence of multicollinearity, VIF values are high for few variables
- Predictor variable variation on target variable is explained by other predictor variables
- Hence, we can choose to drop one variable with high vif and build new model
- Let us consider the vif threshold as 2
- From above result, we see that VIF of 'x' variable is 48.11 which is high
- In model 2, dropping 'x' as it has high vif value

Model 2 - Model after dropping variable 'x'

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.907
Model:	OLS	**Adj. R-squared:**	0.907
Method:	Least Squares	**F-statistic:**	3.281e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:13:13	**Log-Likelihood:**	-2.2969e+05
No. Observations:	26925	**AIC:**	4.594e+05
Df Residuals:	26916	**BIC:**	4.595e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	146.4043	571.409	0.256	0.798	-973.587	1266.396
carat	9913.4303	57.195	173.328	0.000	9801.326	1e+04
cut	114.7475	8.207	13.982	0.000	98.662	130.833
color	326.2168	4.637	70.350	0.000	317.128	335.306
clarity	512.1482	4.995	102.534	0.000	502.358	521.939
depth	-41.0639	6.441	-6.376	0.000	-53.688	-28.440
table	-32.2119	4.227	-7.620	0.000	-40.498	-23.926
y	-184.1468	21.590	-8.529	0.000	-226.464	-141.829
z	-475.5191	36.606	-12.990	0.000	-547.270	-403.769

Omnibus:	5564.823	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	183028.939
Skew:	0.229	**Prob(JB):**	0.00
Kurtosis:	15.765	**Cond. No.**	6.49e+03

Inferences from Model 2:

- Model performance, i.e., adjusted R-square is 0.907 which is same as model 1.

- Hence, dropping 'x' variable didn't change the performance. We can say that the variation of x in target variable is explained by other independent variables
- p-values for y and z dropped and are less than 0.05
- The conditional value is also high(6.49e+03), hence there is a problem of multicollinearity
- We will check the vif values again and treat multicollinearity

VIF Values after Model 2:

carat VIF = 13.33
cut VIF = 1.49
color VIF = 1.12
clarity VIF = 1.21
depth VIF = 1.44
table VIF = 1.59
y VIF = 11.3
z VIF = 12.35

- Here, VIF is calculated after dropping x variable and we see that vif values are high for carat,y and z
- We removed carat as it has high VIF of 13.33 and built a model to check if performance is dropped and we see that adj R2 value dropped by 10%. Hence, we can say that dropping 'carat' is not a good idea as it explains certain amount of variation in target variable.
- 'carat' is an important variable whose variance is not explained by other variables in its absence.
- In model 3, dropping 'z' variable which has vif of 12.35 and keeping the 'carat' variable as it explains variance in the target variable

Model 3 - Model after dropping variable 'z' :

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.906
Model:	OLS	**Adj. R-squared:**	0.906
Method:	Least Squares	**F-statistic:**	3.725e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:20:04	**Log-Likelihood:**	-2.2977e+05
No. Observations:	26925	**AIC:**	4.596e+05
Df Residuals:	26917	**BIC:**	4.596e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	658.0749	571.823	1.151	0.250	-462.729	1778.878
carat	9513.4608	48.349	196.765	0.000	9418.694	9608.228
cut	116.6628	8.231	14.174	0.000	100.530	132.796
color	325.0685	4.651	69.897	0.000	315.953	334.184
clarity	515.9703	5.002	103.158	0.000	506.167	525.774
depth	-61.4680	6.266	-9.810	0.000	-73.749	-49.187
table	-31.3660	4.240	-7.398	0.000	-39.677	-23.055
y	-302.3014	19.641	-15.391	0.000	-340.799	-263.803

Omnibus:	5521.555	**Durbin-Watson:**	2.013
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	153907.431
Skew:	0.314	**Prob(JB):**	0.00
Kurtosis:	14.696	**Cond. No.**	6.47e+03

Inferences from Model 3:

- About 90.6% of variance is explained by predictor variables on the target variable
- Adj R-square value is similar to model 2 and dropping variable 'z' does not have any significant effect on the target variable
- Now, lets check the vif values to see if there is any change

VIF Values after Model 3:

carat VIF = 9.47
cut VIF = 1.49
color VIF = 1.12
clarity VIF = 1.21
depth VIF = 1.36
table VIF = 1.59
y VIF = 9.3

- After dropping variable 'z', the performance is almost same as model 2
- Hence, the variance of 'z' is explained by others. Also vif values decreased.
- 90.6% of variance is observed on target variable by x attributes without the column named 'z'
- As we consider carat as important variable, we will drop 'y' variable in model 4 which also has high vif.

Model 4 - Model after dropping variable 'y' :

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.906
Model:	OLS	**Adj. R-squared:**	0.906
Method:	Least Squares	**F-statistic:**	4.304e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:33:38	**Log-Likelihood:**	-2.2989e+05
No. Observations:	26925	**AIC:**	4.598e+05
Df Residuals:	26918	**BIC:**	4.598e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1722.4394	552.915	-3.115	0.002	-2806.182	-638.697
carat	8822.3988	18.012	489.800	0.000	8787.094	8857.704
cut	120.4538	8.263	14.577	0.000	104.257	136.650
color	323.6415	4.670	69.301	0.000	314.488	332.795
clarity	522.4880	5.006	104.381	0.000	512.677	532.299
depth	-44.8006	6.198	-7.228	0.000	-56.950	-32.651
table	-28.8690	4.255	-6.784	0.000	-37.210	-20.528

Omnibus:	5228.267	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	98462.112
Skew:	0.420	**Prob(JB):**	0.00
Kurtosis:	12.331	**Cond. No.**	6.21e+03

VIF Values after Model 4:

carat VIF = 1.3
cut VIF = 1.49
color VIF = 1.12
clarity VIF = 1.2
depth VIF = 1.32
table VIF = 1.59

Inferences from Model 4:

- Adjusted R-square value of model 4 is 0.906 which is same as model 3
- 90.6% of variance is explained by the factors considered above on the y-predicted
- We can say that y has no significant effect on the variance observed in target variable
- Here, the vif of carat is also reduced. Hence, we can say that carat is a significant variable when compared to z and y.
- Multicollinearity problem has been reduced.
- Multicollinearity is reduced, however we will drop table and depth in next models as its coefficient is less
- This is just to check the performance, if it is reduced or same
- If there is no change in performance after dropping these variables, we will have a model with less features

Model 5 - Model after dropping variable 'table' :

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.905
Method:	Least Squares	**F-statistic:**	5.155e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:38:29	**Log-Likelihood:**	-2.2991e+05
No. Observations:	26925	**AIC:**	4.598e+05
Df Residuals:	26919	**BIC:**	4.599e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4626.0738	350.321	-13.205	0.000	-5312.720	-3939.427
carat	8808.1504	17.904	491.957	0.000	8773.057	8843.244
cut	149.6720	7.058	21.206	0.000	135.838	163.506
color	323.6897	4.674	69.254	0.000	314.529	332.851
clarity	524.6151	5.000	104.925	0.000	514.815	534.415
depth	-25.9390	5.544	-4.678	0.000	-36.807	-15.072

Omnibus:	5242.250	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	97818.072
Skew:	0.427	**Prob(JB):**	0.00
Kurtosis:	12.299	**Cond. No.**	2.88e+03

VIF Values after Model 5:

carat VIF = 1.28
cut VIF = 1.09
color VIF = 1.12
clarity VIF = 1.19
depth VIF = 1.05

Inferences from Model 5:

- Adjusted R-square value = 0.905 which is similar to model 4
- Hence, there is no change in performance

Model 6 - Model after dropping variable 'depth' :

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.905
Model:	OLS	**Adj. R-squared:**	0.905
Method:	Least Squares	**F-statistic:**	6.438e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	22:43:22	**Log-Likelihood:**	-2.2992e+05
No. Observations:	26925	**AIC:**	4.599e+05
Df Residuals:	26920	**BIC:**	4.599e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-6255.1366	38.376	-162.994	0.000	-6330.356	-6179.917
carat	8809.9221	17.907	491.975	0.000	8774.823	8845.021
cut	156.3891	6.913	22.622	0.000	142.839	169.939
color	324.8040	4.670	69.556	0.000	315.651	333.957
clarity	525.4872	4.998	105.132	0.000	515.690	535.284

Omnibus:	5231.586	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	98842.344
Skew:	0.419	**Prob(JB):**	0.00
Kurtosis:	12.349	**Cond. No.**	31.2

Inferences from Model 6:

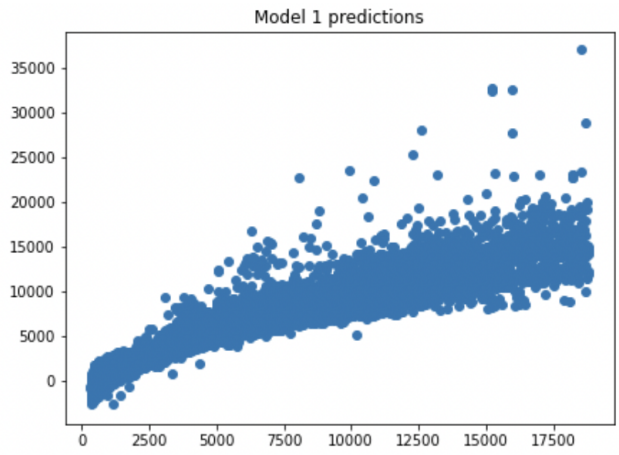
- After dropping table and depth, we see that 90.5% of variance is explained which is same as model 4
- Hence, we can say that table and depth has no significant effect on the variance observed in target variable though the vif values are low for both of them

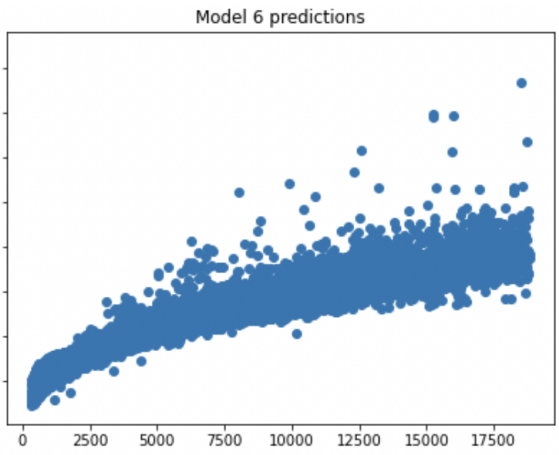
Model Evaluation:

	model_name	adj_R_squared	no_of_x_variables	x_variables_included
0	simple linear regression	0.850817	1	carat
1	simple linear regression scaled	0.850817	1	carat(scaled)
2	model_1_all	0.908470	9	carat,cut,color,clarity,depth,table,x,y,z
3	model_2_drop_x	0.906974	8	carat,cut,color,clarity,depth,table,y,z
4	model_3_drop_z	0.906395	7	carat,cut,color,clarity,depth,table,y
5	model_4_drop_y	0.905574	6	carat,cut,color,clarity,depth,table
6	model_5_drop_table	0.905416	5	carat,cut,color,clarity,depth
7	model_6_drop_depth	0.905343	4	carat,cut,color,clarity

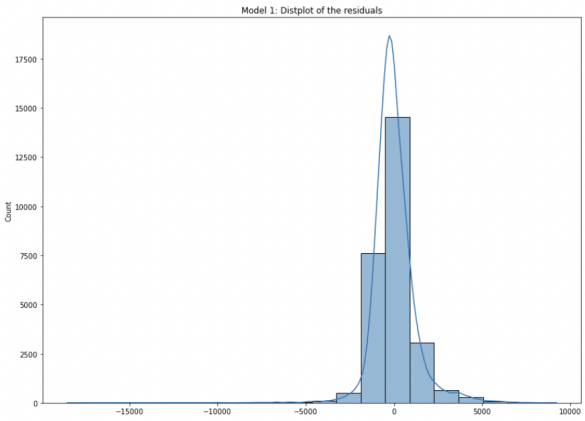
- We will use Model 6 to predict and check the model evaluation.
- Model 6 because, it has a high Adjusted R Square, with least number of features
- It is evident that model 6 is better model compared to all other models

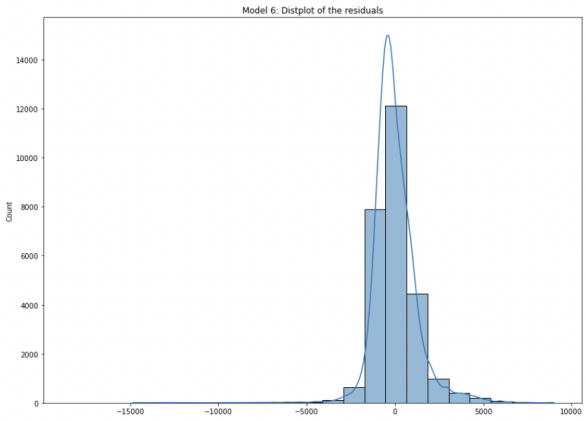
Model 1 & 6 - Prediction and Scatterplot:

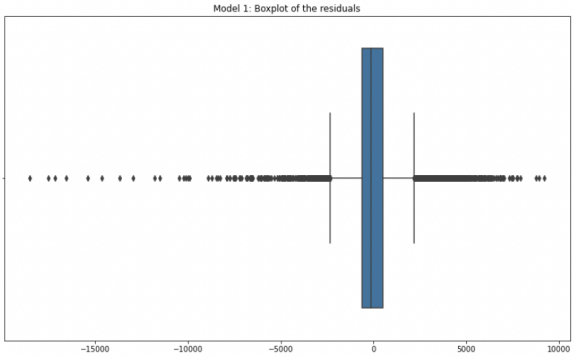


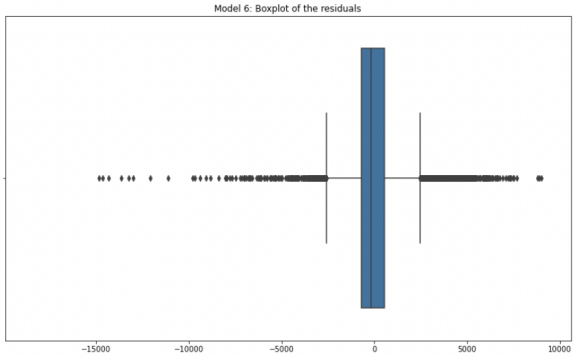


Checking the boxplot and the distplot of the residuals:









Calculating RMSE (Root mean square error):

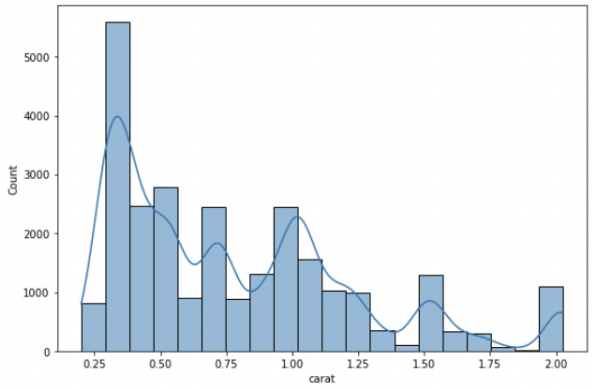
Model 1 RMSE: 1216.280875799657
Model 6 RMSE: 1236.9954355893326

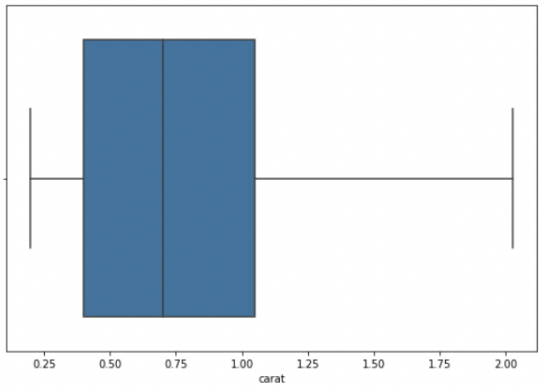
Inferences:

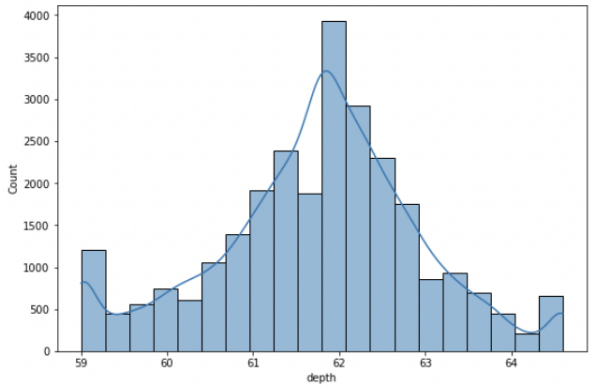
- We can conclude that model 6 is the better model with a performance of 90.5% with least number of features(4)
- Out of all the models, linear regression performs the best predicting price of cubic zirconia with an adjusted $R2$ of 90.5% and an RMSE of about 1236.99. This indicates that our model can explain about 90.5% of the variation in price
- The test error of the linear model is low for model 1. However, in model 6, multicollinearity is reduced by using less number of features(4) compared to model 1 which used 9 features to explain the variance in predicted variable.

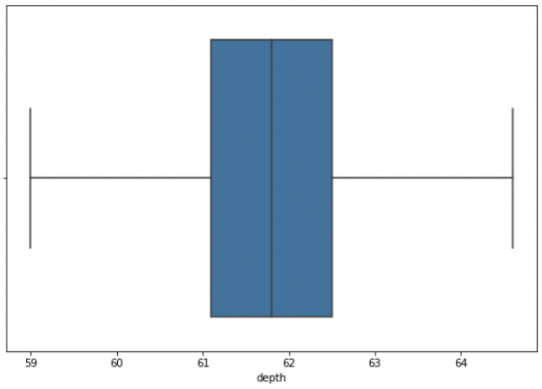
Treating Outliers and building model:

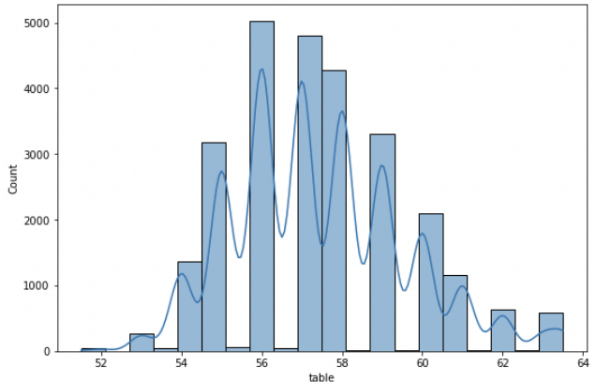
Box Plots and Distribution Plots after treating outliers:

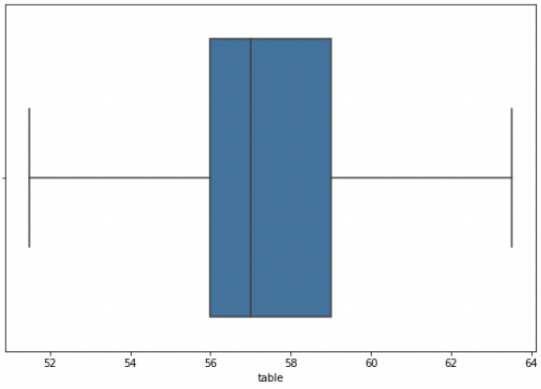


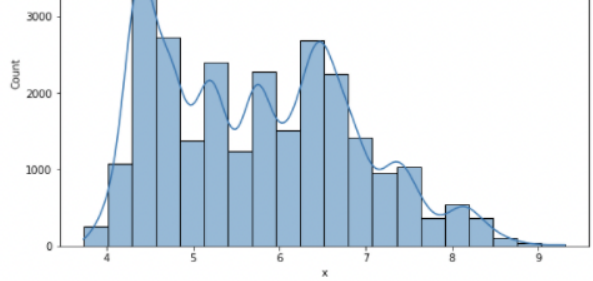


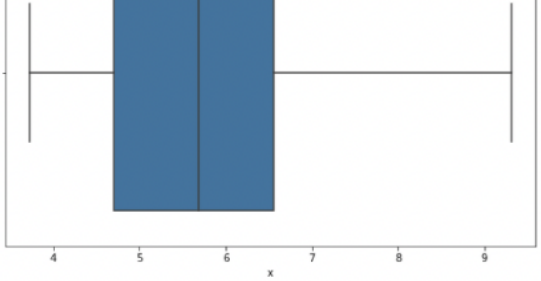


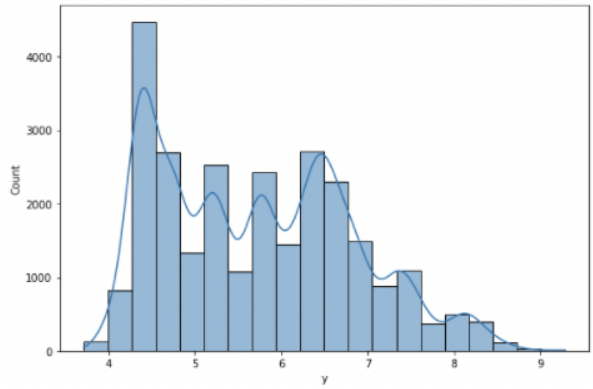


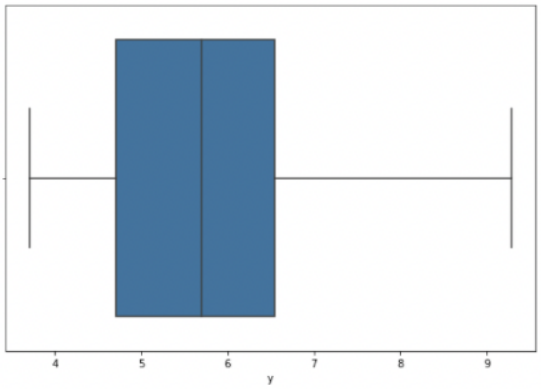


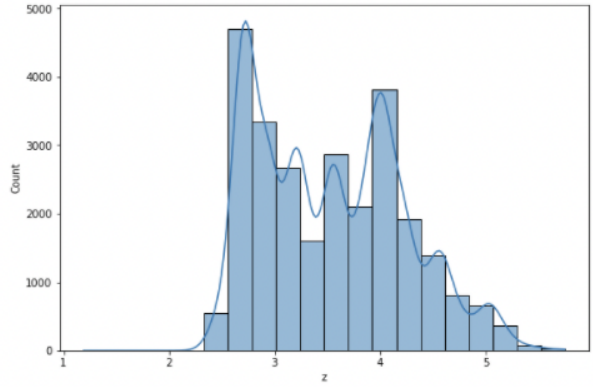


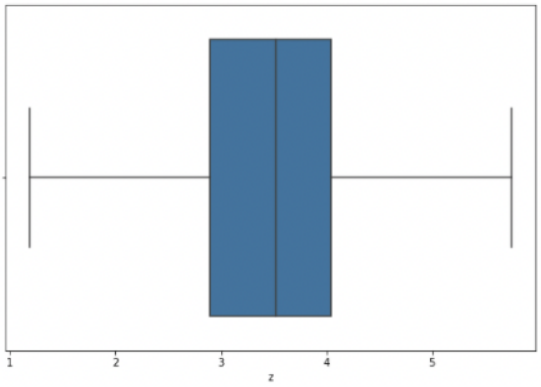

Count
x

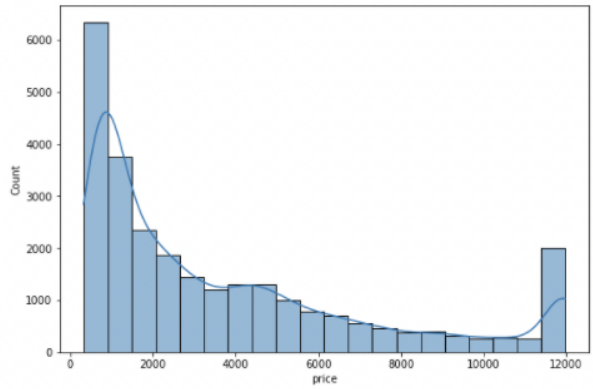

x

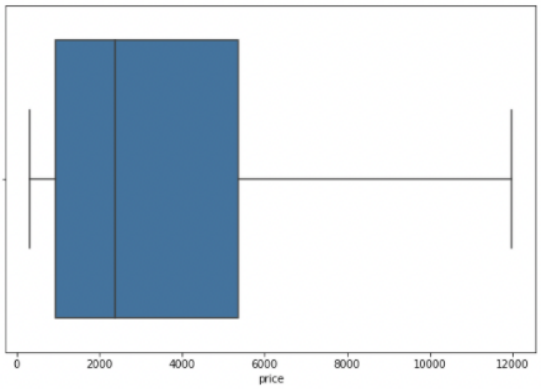

Count
y


y


Count
z


z


Count
price


price

Final model after treating outliers: (Model 6)

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.930
Model:	OLS	**Adj. R-squared:**	0.930
Method:	Least Squares	**F-statistic:**	8.971e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	23:27:15	**Log-Likelihood:**	-2.2183e+05
No. Observations:	26925	**AIC:**	4.437e+05
Df Residuals:	26920	**BIC:**	4.437e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5209.8081	28.603	-182.140	0.000	-5265.872	-5153.744
carat	7933.3529	13.684	579.749	0.000	7906.531	7960.174
cut	121.2053	5.118	23.681	0.000	111.173	131.237
color	273.1859	3.454	79.092	0.000	266.416	279.956
clarity	449.2426	3.700	121.406	0.000	441.990	456.495

Omnibus:	3370.319	**Durbin-Watson:**	2.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11672.915
Skew:	0.624	**Prob(JB):**	0.00
Kurtosis:	5.974	**Cond. No.**	31.6

Model Evaluation after treating outliers:

	model_name	adj_R_squared	no_of_x_variables	x_variables_included
0	model 1	0.931350	9	carat,cut,color,clarity,depth,table,x,y,z
1	model 2	0.931020	8	carat,cut,color,clarity,depth,table,y,z
2	model 3	0.930763	7	carat,cut,color,clarity,depth,table,z
3	model 4	0.930278	6	carat,cut,color,clarity,depth,table
4	model 5	0.930204	4	carat,cut,color,clarity

Inferences on models built after treating outliers:

- Model 5 has an adjusted R-square value of 93% which is similar to all the models from 1 to 4
- There is no change in performance even after treating multicollinearity
- 93% of variance is explained by the factors considered above(carat,cut,color and clarity) on the target variable, whereas before treating outliers, 90.5% of variance is explained
- The outcome of the regression model is better after treating outliers with a variance of 93%, however the values are not unrealistic. Hence, we can choose the model we built without treating outliers with a variance of 90.5%

Predictive Approach of Linear Regression:

- Splitting the data into dependent and independent variables
- Splitting the data into train(70%) and test(30%)

Let us explore the coefficients for each of the independent attributes:

The coefficient for carat is 11054.457219364136
The coefficient for cut is 107.3887926665351
The coefficient for color is 329.6026406281934
The coefficient for clarity is 502.9591100351053
The coefficient for depth is -84.34477277815833
The coefficient for table is -35.57555822491017
The coefficient for x is -951.975419286857
The coefficient for y is 6.672990548429425
The coefficient for z is -42.13557708600473

The intercept for our model is : 4966.082681057811

Regression results on Training data:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.908
Model:	OLS	**Adj. R-squared:**	0.908
Method:	Least Squares	**F-statistic:**	2.065e+04
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	23:53:05	**Log-Likelihood:**	-1.6062e+05
No. Observations:	18847	**AIC:**	3.213e+05
Df Residuals:	18837	**BIC:**	3.213e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4966.0827	717.308	6.923	0.000	3560.094	6372.072
carat	1.105e+04	93.361	118.406	0.000	1.09e+04	1.12e+04
cut	107.3888	9.738	11.027	0.000	88.301	126.477
color	329.6026	5.507	59.853	0.000	318.809	340.397
clarity	502.9591	5.957	84.431	0.000	491.283	514.635
depth	-84.3448	7.863	-10.727	0.000	-99.757	-68.933
table	-35.5756	5.011	-7.100	0.000	-45.397	-25.754
x	-951.9754	50.868	-18.715	0.000	-1051.681	-852.270
y	6.6730	23.895	0.279	0.780	-40.164	53.510
z	-42.1356	41.730	-1.010	0.313	-123.929	39.658

Omnibus:	4196.806	**Durbin-Watson:**	1.993
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	205178.221
Skew:	-0.059	**Prob(JB):**	0.00
Kurtosis:	19.164	**Cond. No.**	6.89e+03

Regression results on Test data:

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.911
Model:	OLS	**Adj. R-squared:**	0.911
Method:	Least Squares	**F-statistic:**	9172.
Date:	Thu, 03 Feb 2022	**Prob (F-statistic):**	0.00
Time:	23:53:18	**Log-Likelihood:**	-68794.
No. Observations:	8078	**AIC:**	1.376e+05
Df Residuals:	8068	**BIC:**	1.377e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7961.8831	1779.493	-4.474	0.000	-1.15e+04	-4473.617
carat	1.102e+04	139.740	78.866	0.000	1.07e+04	1.13e+04
cut	145.0145	14.791	9.804	0.000	116.020	174.009
color	325.2370	8.319	39.094	0.000	308.929	341.545
clarity	486.2829	9.112	53.368	0.000	468.421	504.145
depth	99.9556	26.035	3.839	0.000	48.920	150.991
table	-9.9008	7.687	-1.288	0.198	-24.970	5.168
x	-1351.1237	211.755	-6.381	0.000	-1766.218	-936.029
y	2268.3609	221.033	10.263	0.000	1835.080	2701.642
z	-3042.5716	392.800	-7.746	0.000	-3812.561	-2272.582

Omnibus:	1779.903	**Durbin-Watson:**	2.014
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74432.773
Skew:	-0.184	**Prob(JB):**	0.00
Kurtosis:	17.866	**Cond. No.**	1.15e+04

- Adjusted R square value for both train and test dataset are similar i.e 90.8% and 91.1% of variance is explained in training and test data respectively
- As the adj R square value is high for test data, predictions can be accurate in production
- Next, we build models on training data and predict them on both training and test data

RMSE check for the models:

Training Data RMSE of model_1: 1215.8035529418505
Test Data RMSE of model_1: 1217.778984191432

Training Data RMSE of model_6: 1236.8326134303509
Test Data RMSE of model_6: 1237.4497650968

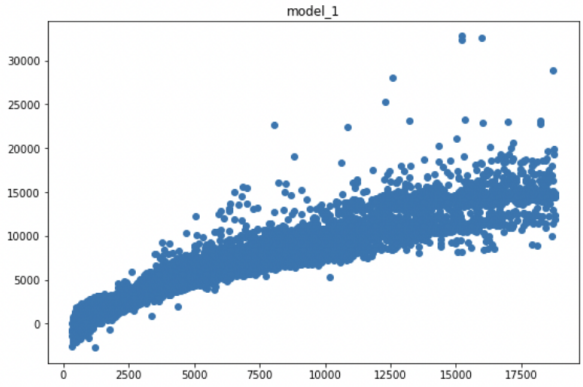
Inferences:

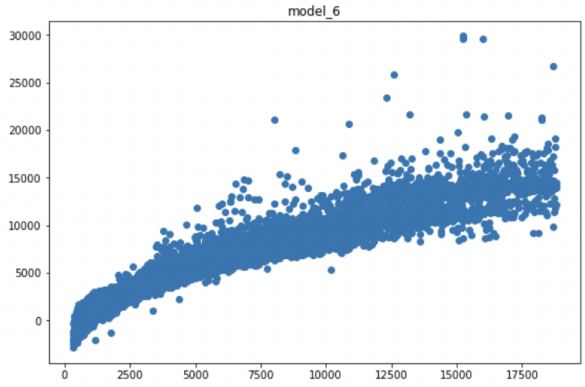
- RMSE value is higher for test data in both model 1 and 6
- If rmse is high on test data, we can say that model is performing well on training data but not so good on unseen data
- In our case, rmse for test data is only little bit high than training data. Hence, we cannot state model will not perform well in production
- The error between yactual and ypredict must be as low as possible even on test data
- RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

- RMSE for model 1(1217) is lower compared to model 6(1237). We predicted model 6 as the better model in descriptive approach. However, as accuracy is important in predictive approach not the number of variables used, we can consider model 1 as the better possible model for predictions

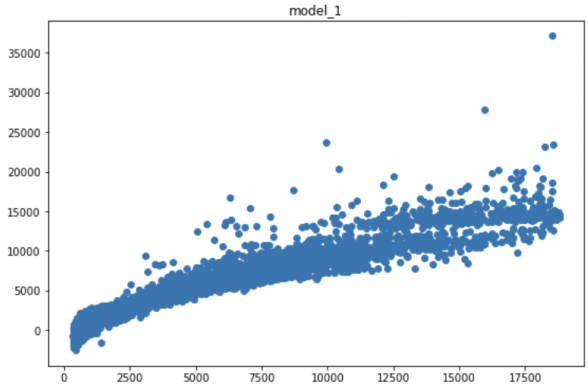
Scatterplot for the predictions:

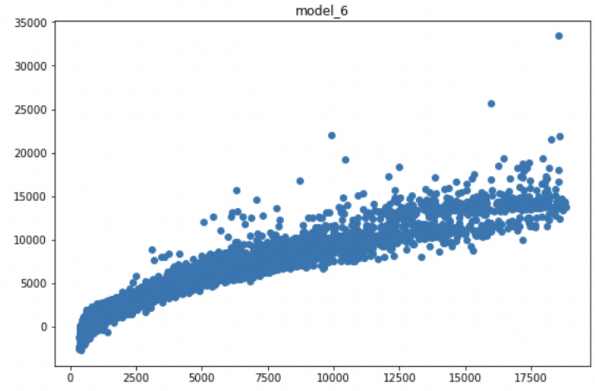
Training Data:





Test Data:





- From the scatterplot, we can observe that there is more linear relationship between actual y and predicted y when compared to the model 1
- Visually both looks like same pattern, but the datapoints are a bit closely arranged in model 6 when compared to model 1

Problem 2: Logistic Regression

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.

Following is a guideline for developing a solution:

1. The very first step of any data analysis assignment is to do the exploratory data analysis (EDA). Once you have understood the nature of all the variables, especially identified the response and the predictors, apply appropriate methods to determine whether there is any duplicate observation or missing data and whether the variables have a symmetric or skewed distribution. Note that data may contain various types of attributes and numerical and/or visual data summarization techniques need to be appropriately decided. Both univariate and bivariate analyses and pre-processing of data are important. Check for outliers and comment on removing or keeping them while model building. For this is a classification problem, the dependence of the response on the predictors needs to be investigated.

2. Use the Pre-processed Full Data to develop a logistic regression model to identify significant predictors. Check whether the proposed model is free of multicollinearity. Apply variable selection method as required. Show all intermediate models leading to the final model. Justify your choice of the final model. Which are the significant predictors? Compare values of model selection criteria for proposed models. Compare as many criteria as you feel are suitable.

3. Alternatively, if prediction accuracy of the full scholarship is the only objective, then you may want to divide the data into a training and a test set, chosen randomly, and use the training set to develop a model and test set to validate your model. Use the models developed in Part (2) to compare accuracy in training and test sets. Compare the final model of Part (2) and the proposed one in Part (3). Which model provides the most accurate prediction? If the model found in Part (2) is different from the proposed model in Part (3), give an explanation.

Data Dictionary:

1. Holiday_Package: Opted for Holiday Package yes/no?
2. Salary: Employee salary
3. Age: Age in years
4. Edu: Years of formal education
5. no_young_children: The number of young children (younger than 7 years)
6. no_older_children: Number of older children
7. foreign: foreigner Yes/No

Solution:

Checking first five rows of the dataset:

	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

Dropping Unnamed: 0 Column:

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

Checking dimensions of the dataset:

```
<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
```

Inferences:

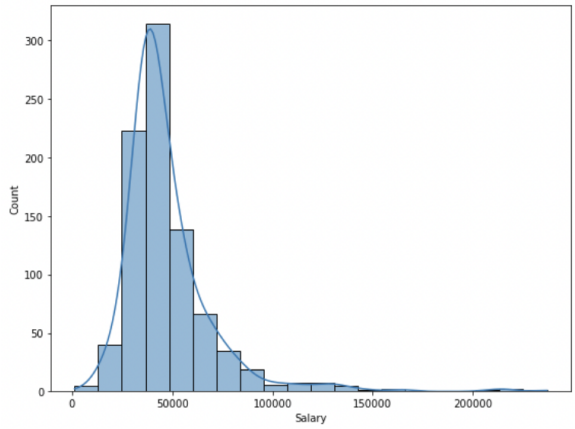
- The dataset has a total of five independent variables - All are continuous, two categorical variables one of which is target variable
- Shape (dimension) of the Dataset is (872, 7)
- There are zero NULL values present in the dataset
- No duplicate values present in the dataset
- There are a total of 7 variables and 872 records

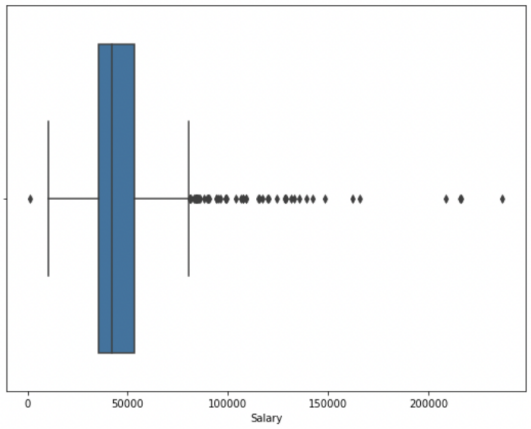
Univariate Analysis:

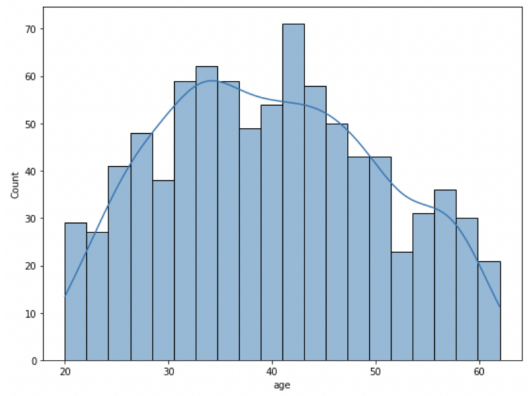
Summary statistics of the dataset: To perform Univariate analysis on continuous variables, let us start with looking at the summary statistics of the dataset.

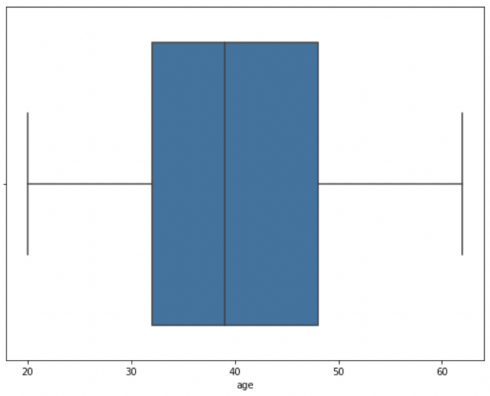
	count	mean	std	min	25%	50%	75%	max
Salary	872.0	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0

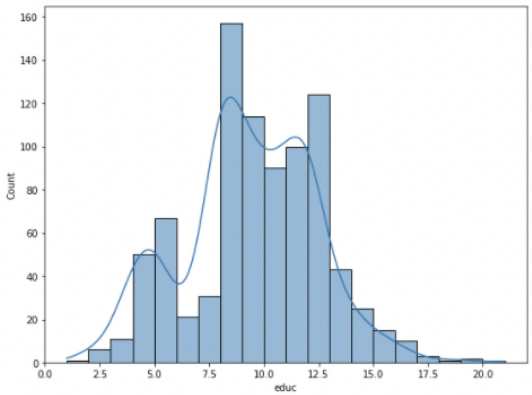
Plotting boxplots and distribution plots:

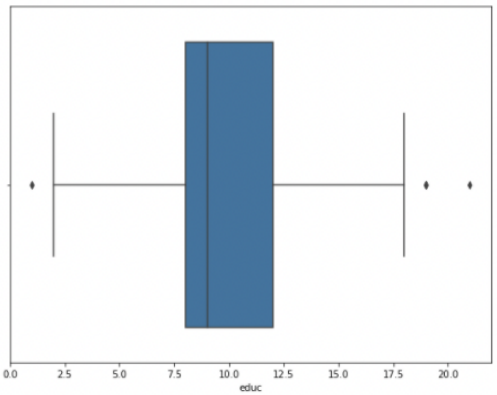


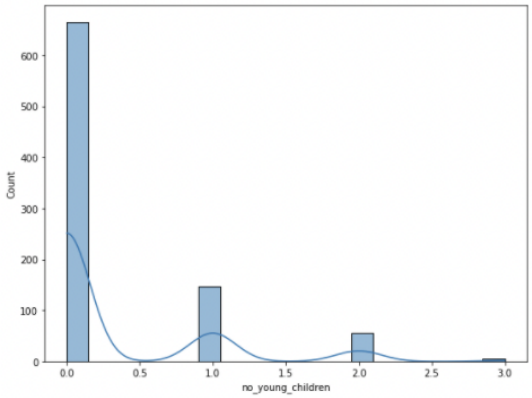


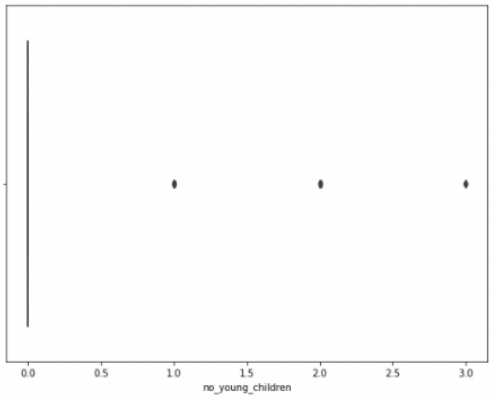

Count
age

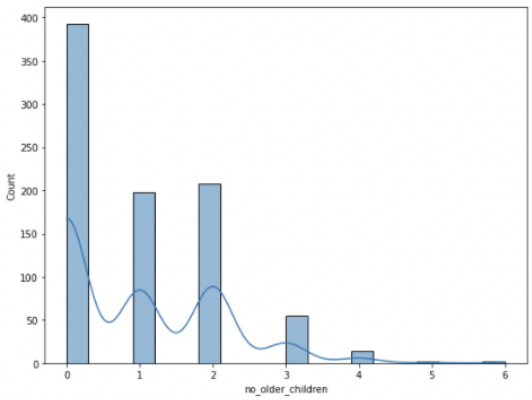

age

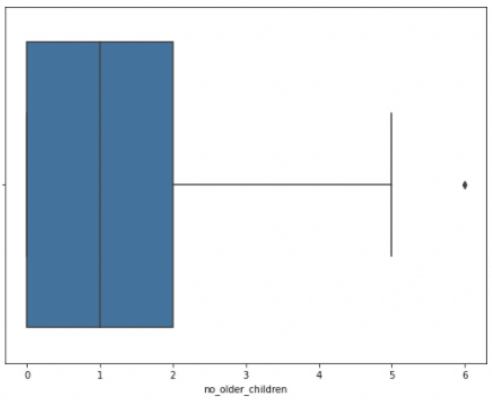

Count
educ


educ


Count
no_young_children


no_young_children


Count
no_older_children


no_older_children

Checking Outlier% of the dataset:

	Outlier%
Holliday_Package	0.000000
Salary	6.536697
age	0.000000
educ	0.458716
foreign	0.000000
no_older_children	0.229358
no_young_children	23.738532

- Three columns have outlier values with a Percentage of 6.53,0.45,0.22 which is not very high as 20%
- One column named 'no_young_children' has a very high outlier percentage of 23.73%
- Logistic Regression models are not much impacted due to the presence of outliers because the sigmoid function tapers the outliers. But the presence of extreme outliers may somehow affect the performance of the model and lowering the performance
- Hence, we are not considering treating outliers. If we treat outliers for variable no_young_children, we will loose the only data we have as it is clear from the above box plot

Checking skewness in the dataset:

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

Checking kurtosis values:

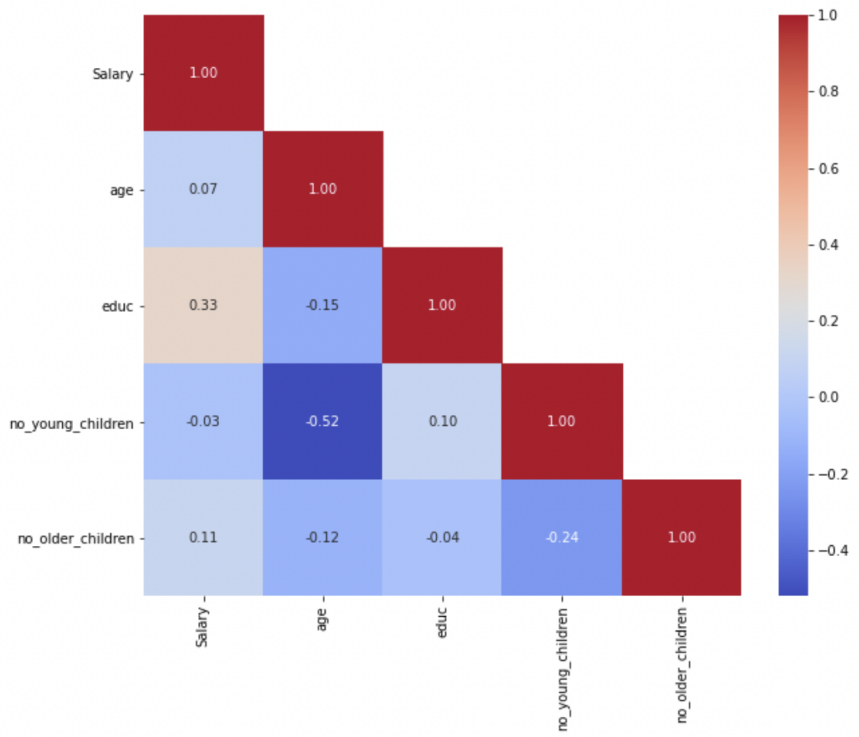
	Kurtosis
Salary	15.852557
age	-0.909962
educ	0.005558
no_young_children	3.109892
no_older_children	0.676017

Inferences on Univariate Analysis:

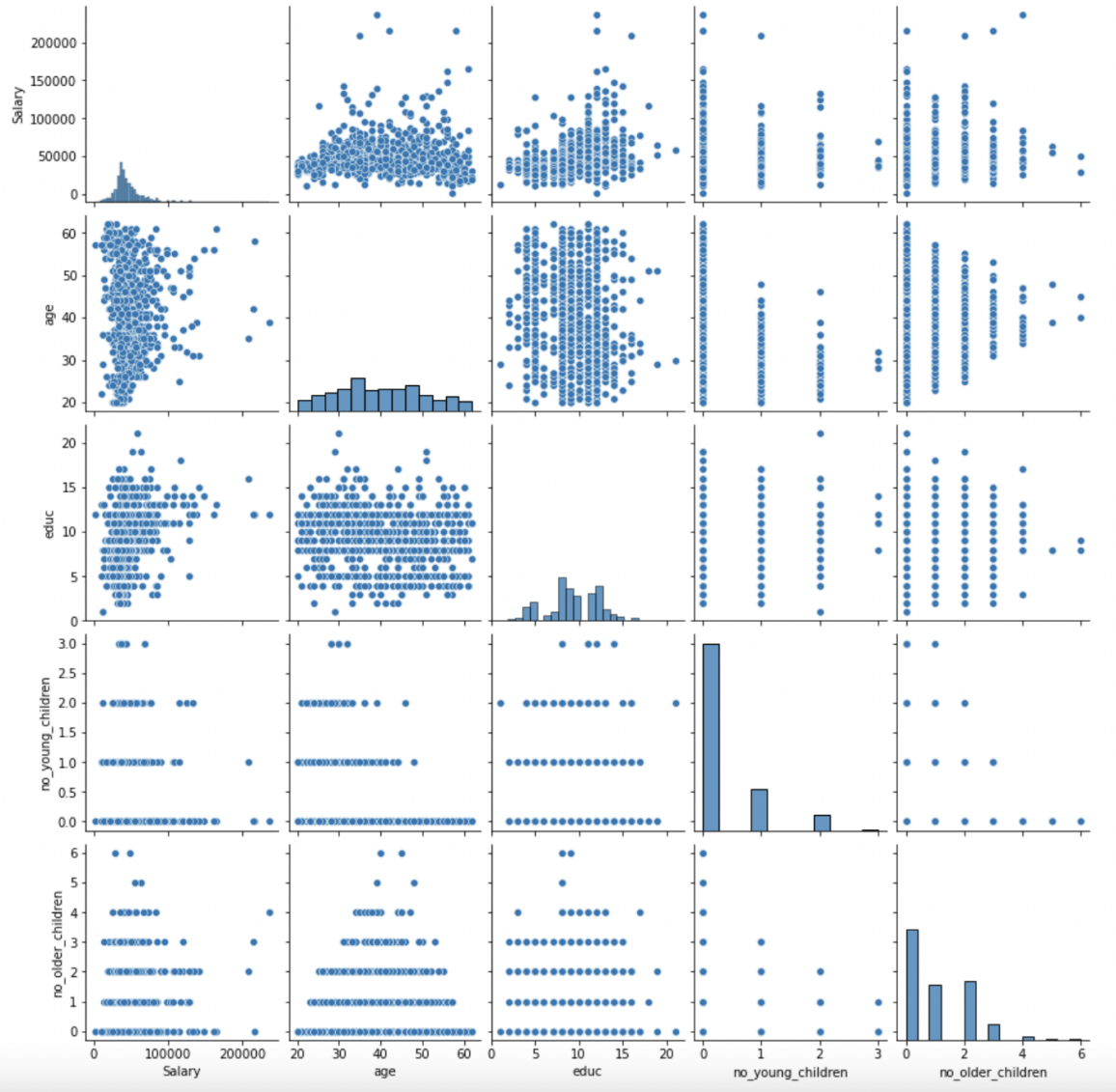
1. From the Box plots, we can conclude that there are few outliers present in educ,no_young_children and no_older_children variables.
2. There are huge outliers for salary variables but they are realistic values
3. Kurtosis value for 2 variables is high, not near to normal distribution
4. There is extreme positive kurtosis seen in few variables which is > 3 which means more of the values are located in the tails of the distribution rather than around the mean i.e., salary variable

Multi variate Analysis: Analyzing the relationship among continuous variables by using Pair plot and Correlation Heatmap.

Heatmap:



Pairplot:



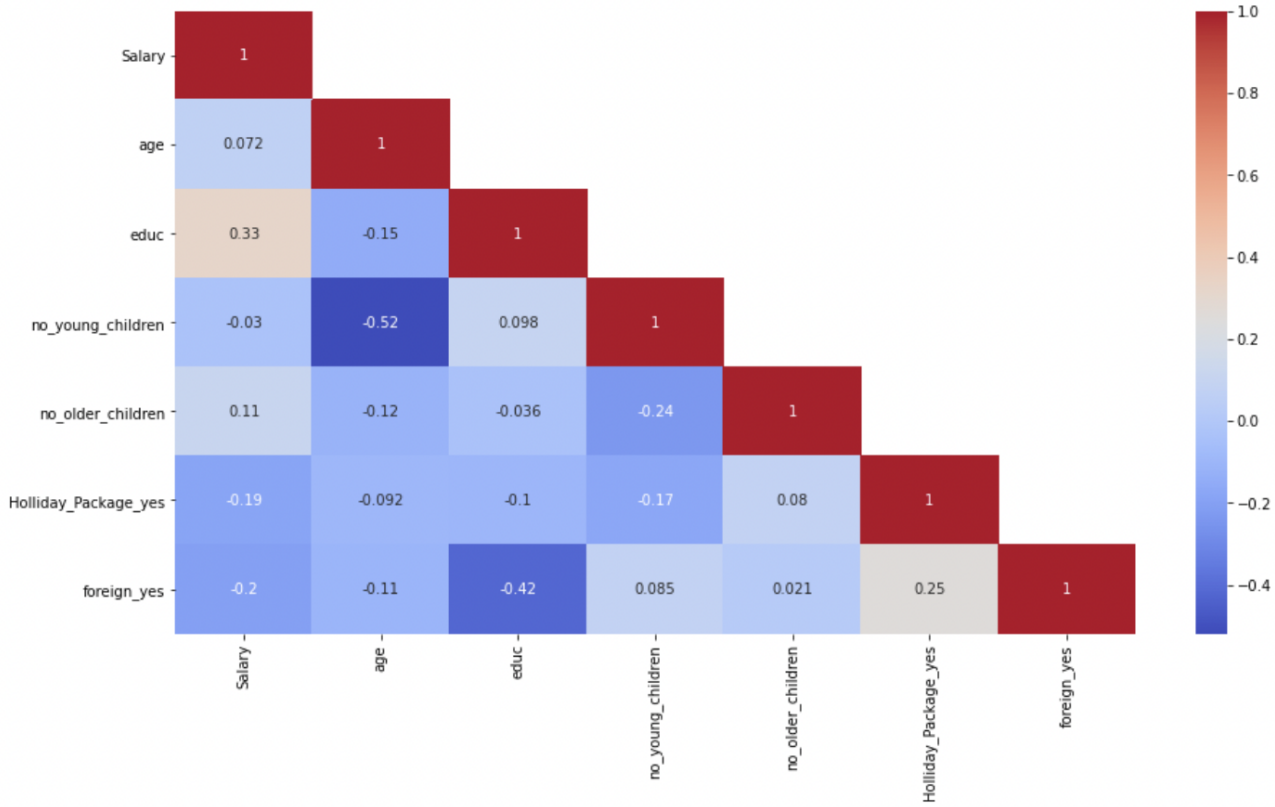
Inferences on Multi variate Analysis:

1. - educ variable is a bit positively correlated with salary variable
2. - Top Negative correlation pair is no_young_children and age (-0.52)
3. - From pairplot, we see that there is no normal distribution for all the variables
4. - Few of the variables have no correlation with each other, i.e., datapoints are spread wideacross(ex: educ and age). - There might be small amount of correlation for salary and educ but there is no particular trend

Handling categorical test variables:

	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
0	48412	30	8	1	1	0	0
1	37207	45	8	0	1	1	0
2	58022	46	9	0	0	0	0
3	66503	31	11	2	0	0	0
4	66734	44	12	0	2	0	0

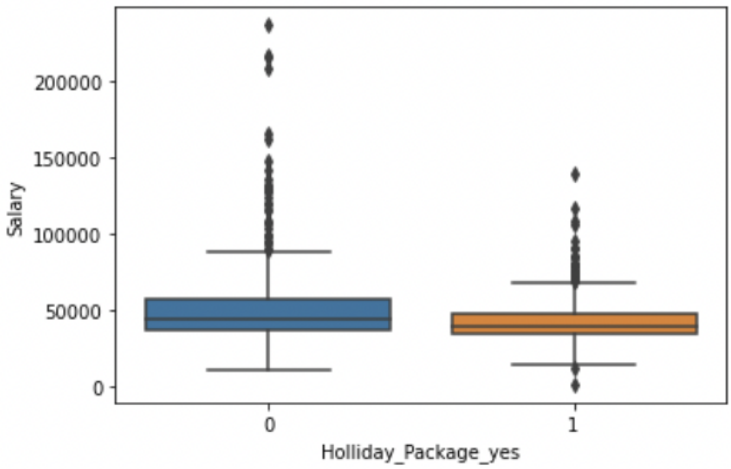
Heatmap:

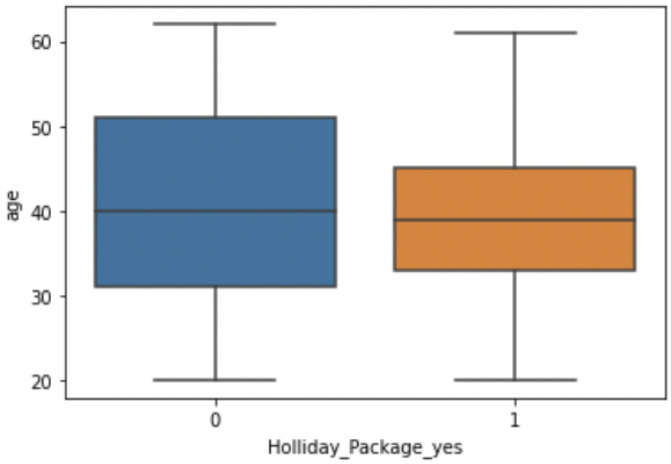


Building the Logistic Regression Models - Descriptive approach:

We will build the Logistic Regression model with only the variable 'Salary' and see how that affects the probability.

But first, let us plot a boxplot to understand the variability in the 'Holiday_Package_yes' with respect to the 'Salary' variable.





- We can see that people with high salary tend not to opt for holiday packages and people with less salary are equal in opting and not opting a holiday package
- We can see that people with age higher than median are not choosing holiday package

Model 1:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	870
Method:	MLE	**Df Model:**	1
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.02863
Time:	01:34:33	**Log-Likelihood:**	-584.39
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	4.378e-09

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.7875	0.190	4.151	0.000	0.416	1.159
Salary	-2.035e-05	3.9e-06	-5.223	0.000	-2.8e-05	-1.27e-05

- - Adj pseudo r-square value is very less in model 1
- - We will add 'age' variable in model 2 and check the performance

Model 2:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	869
Method:	MLE	**Df Model:**	2
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.03346
Time:	01:34:39	**Log-Likelihood:**	-581.48
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	1.811e-09

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.4070	0.322	4.364	0.000	0.775	2.039
Salary	-2.009e-05	3.93e-06	-5.112	0.000	-2.78e-05	-1.24e-05
age	-0.0159	0.007	-2.402	0.016	-0.029	-0.003

- Adj pseudo r-square value is similar for model 2 and 1
- We will add 'foreign_yes' variable in model 2 and check the performance

Model 3:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	868
Method:	MLE	**Df Model:**	3
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.06632
Time:	01:34:42	**Log-Likelihood:**	-561.72
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	3.401e-17

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.7918	0.336	2.359	0.018	0.134	1.450
Salary	-1.544e-05	3.87e-06	-3.987	0.000	-2.3e-05	-7.85e-06
age	-0.0123	0.007	-1.813	0.070	-0.026	0.001
foreign_yes	1.0496	0.171	6.145	0.000	0.715	1.384

VIF values:

Salary VIF = 1.04
age VIF = 1.01
foreign_yes VIF = 1.05

- Adj pseudo r-square value is not much increased in model 3
- VIF values are below threshold and are not high
- We will build model 4 by adding educ variable

Model 4:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	867
Method:	MLE	**Df Model:**	4
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.06722
Time:	01:34:48	**Log-Likelihood:**	-561.17
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	1.132e-16

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.4894	0.444	1.103	0.270	-0.380	1.359
Salary	-1.665e-05	4.07e-06	-4.087	0.000	-2.46e-05	-8.66e-06
age	-0.0105	0.007	-1.514	0.130	-0.024	0.003
foreign_yes	1.1282	0.187	6.025	0.000	0.761	1.495
educ	0.0290	0.028	1.042	0.297	-0.026	0.084

VIF values:

Salary VIF = 1.14
age VIF = 1.08
foreign_yes VIF = 1.26
educ VIF = 1.39

- Adj pseudo r-square value is not increased in model 4
- VIF values are below threshold(2) and are not high
- We will build model 5 by adding no_young_children variable

Model 5:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	867
Method:	MLE	**Df Model:**	4
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.1265
Time:	01:34:55	**Log-Likelihood:**	-525.51
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	6.885e-32

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.6725	0.426	6.278	0.000	1.838	3.507
Salary	-1.664e-05	4.08e-06	-4.075	0.000	-2.46e-05	-8.64e-06
age	-0.0495	0.008	-5.843	0.000	-0.066	-0.033
foreign_yes	1.2124	0.183	6.634	0.000	0.854	1.571
no_young_children	-1.2946	0.169	-7.669	0.000	-1.625	-0.964

VIF values:

Salary VIF = 1.05
age VIF = 1.38
foreign_yes VIF = 1.05
no_young_children VIF = 1.37

- Adj pseudo r-square value is 0.119 for model 5 and 11.9% of variance is explained in the target variable
- VIF values are below threshold(2) and are not high
- We will build model 6 by adding no_older_children variable

Model 6:

Logit Regression Results

Dep. Variable:	Holliday_Package_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	866
Method:	MLE	**Df Model:**	5
Date:	Fri, 04 Feb 2022	**Pseudo R-squ.:**	0.1266
Time:	01:35:05	**Log-Likelihood:**	-525.43
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	4.204e-31

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.7419	0.461	5.953	0.000	1.839	3.645
Salary	-1.644e-05	4.11e-06	-3.999	0.000	-2.45e-05	-8.38e-06
age	-0.0505	0.009	-5.681	0.000	-0.068	-0.033
foreign_yes	1.2161	0.183	6.644	0.000	0.857	1.575
no_young_children	-1.3191	0.180	-7.323	0.000	-1.672	-0.966
no_older_children	-0.0294	0.073	-0.400	0.689	-0.173	0.115

VIF values:

Salary VIF = 1.07
age VIF = 1.51
foreign_yes VIF = 1.06
no_young_children VIF = 1.57
no_older_children VIF = 1.18

- Adj pseudo r-square value is 0.118 for model 6 which is similar to model 5

Model Evaluation:

	model_name	model_perf	variables
0	model 1	0.026967	salary
1	model 2	0.030135	salary + age
2	model 3	0.061329	salary + age + foreign_yes
3	model 4	0.060573	salary + age + foreign_yes + educ
4	model 5	0.119842	salary + age + foreign_yes + no_young
5	model 6	0.118313	salary + age + foreign_yes + no_young + no_old

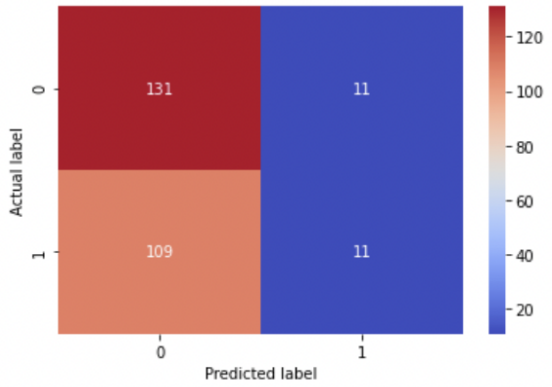
- Model 5 is better compared to all other models because 11.9% of variance is explained by 4 variables
- In all other models, less variance is explained compared to model 5

Predictive Approach:

- Using 4,5,6 model variables to build the models on training data and we will predict for both train and test data
- We will evaluate models based on classification metrics
- Accuracy Score of Model 4: 0.5229508196721312
- Accuracy Score of Model 5: 0.6377049180327868
- Accuracy Score of Model 6: 0.5704918032786885

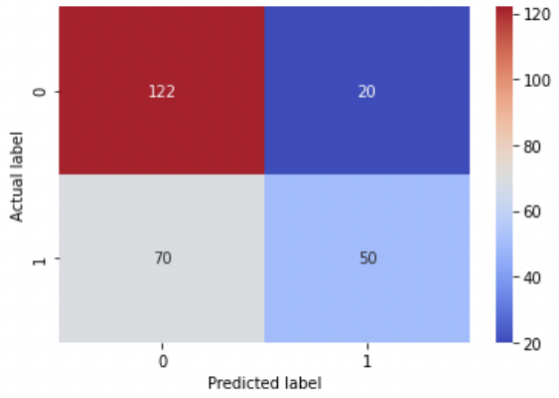
Confusion matrix and classification report:

Model 4:



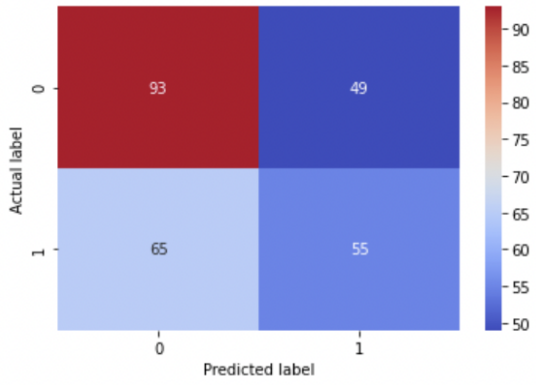
	precision	recall	f1-score	support
0	0.55	0.92	0.69	142
1	0.50	0.09	0.15	120
accuracy			0.54	262
macro avg	0.52	0.51	0.42	262
weighted avg	0.52	0.54	0.44	262

Model 5:



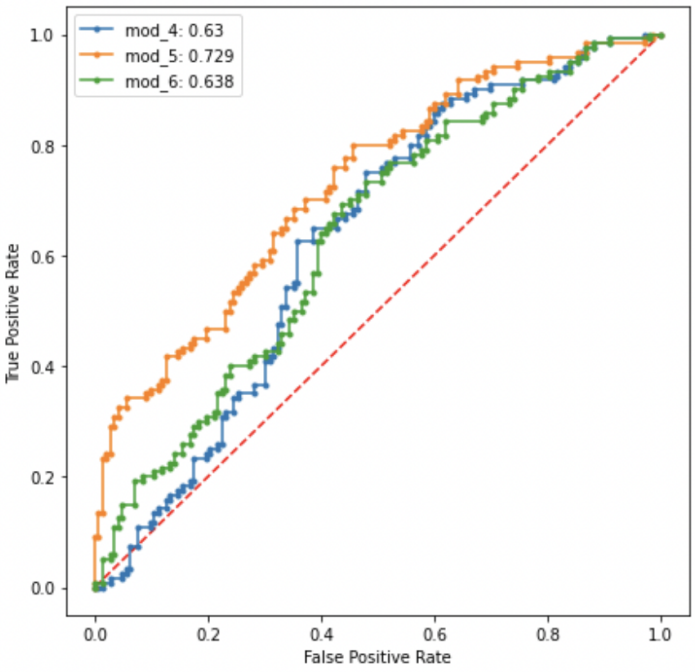
	precision	recall	f1-score	support
0	0.64	0.86	0.73	142
1	0.71	0.42	0.53	120
accuracy			0.66	262
macro avg	0.67	0.64	0.63	262
weighted avg	0.67	0.66	0.64	262

Model 6:



	precision	recall	f1-score	support
0	0.59	0.65	0.62	142
1	0.53	0.46	0.49	120
accuracy			0.56	262
macro avg	0.56	0.56	0.56	262
weighted avg	0.56	0.56	0.56	262

ROC - AUC:



Inferences:

- Model 5 is better model to suggest because, accuracy is important in predictions
- In model 4, recall and f1-score are very low
- Though recall is a bit low in model 5 when compared to model 6, we can propose model 5 as it has good f1-score and accuracy along with recall
- Model 5 is better in the predictive approach as the area under the curve is high
- Also, accuracy for model 5 is high(63.7) compared to other models