17 July 2022

Predictive Modelling Project

Submitted by – Akashatra Sharma

Contents

Table Of Figures

Executive Summary

There were basically two types of Dataset provided which gives us a lot of information. The first data set was named as “cubic zirconia” and it consists of the attributes of approximately 27000 pieces of cubic zirconia that shows the measurements, prices etc. Next data we were provided was named as “Holiday_Package” which contains information on various employees showing their salary, education etc. In both of the datasets, we performed different analytical techniques and build different models in order to get better understanding and give business implications regarding each case study of data set.

Introduction

The purpose of this assignment is to explore the data sets. For that, we’ll do different inferential & statistical operations in order to get the most of out the data and help in building a very good model for the company.

Starting with the data sets, we had gone through the both the data sets and the briefing of the data sets are as follows :

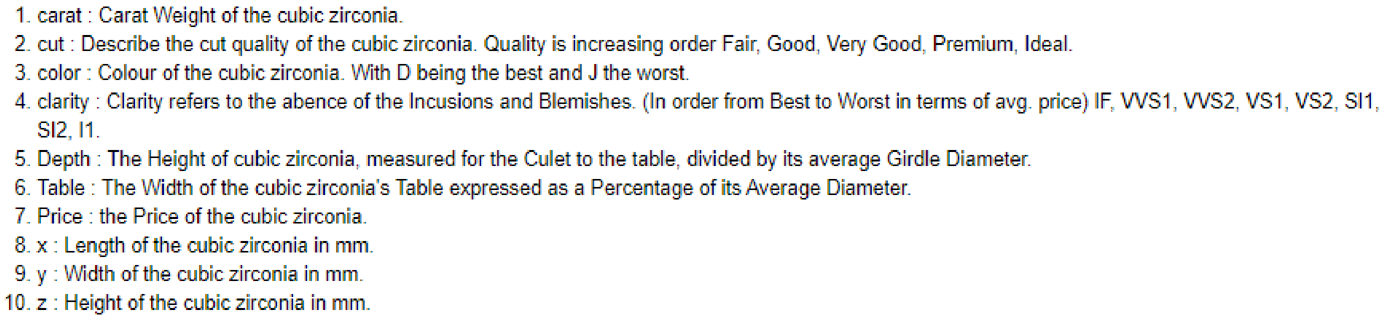
- First Data set that was ‘cubic zirconia’ consists of prices and others features of 26966 pieces of cubic zirconia and using that data to find out the best effective predictors to determine the price of cubic zirconia on the basis of these features.
- Second data set that was ‘Holiday_Package’ consists variety of information of 872 employees who had opted for the Holiday Package or not.

Data Dictionary

The data dictionary is mainly for the understanding of meaning of columns provided in the data set .

1. Cubic zirconia

It is as follows :

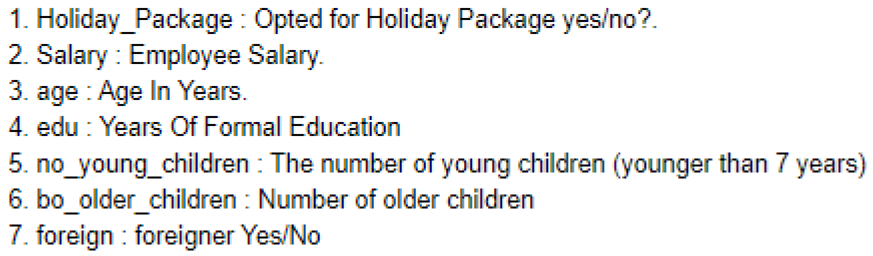


1. carat : Carat Weight of the cubic zirconia.
2. cut : Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. color : Colour of the cubic zirconia. With D being the best and J the worst.
4. clarity : Clarity refers to the abence of the Incusions 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 for 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.

Figure 1 : Data Dictionary (Cubic Zirconia)

2. Holiday Package

It is as follows :



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. bo_older_children : Number of older children
7. foreign : foreigner Yes/No

Figure 2 : Data Dictionary (Holiday Package)

Data Description

Description of both data sets are as follows :

1. Cubic Zirconia

- carat : Continuous Data from 0.2 to 4.50
- cut : Categorical Data from Ideal to Premium
- color : Categorical Data from E to J
- clarity : Categorical Data from SI1 to SI1
- depth : Continuous Data from 50.8 to 73.60

- table : Continuous Data from 49.0 to 79.00
- x : Continuous Data from 0.0 to 10.23
- y : Continuous Data from 0.0 to 58.90
- z : Continuous Data from 0.0 to 31.80
- price : Continuous Data from 326.0 to 18818.00

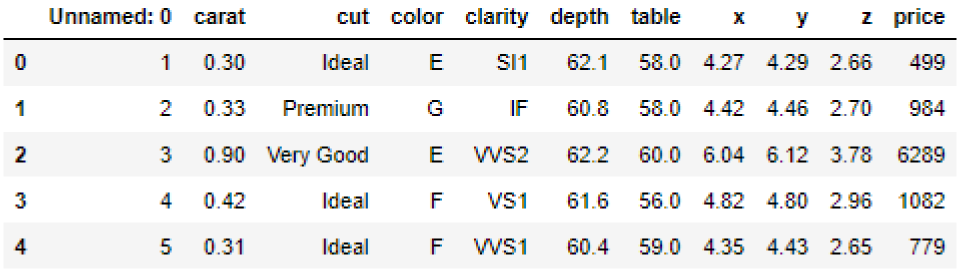
2. Holiday Package

- Holliday_Package : Categorical Data from no to no
- Salary : Continuous Data from 1322.0 to 236961.0
- age : Continuous Data from 20 to 62
- educ : Continuous Data from 1.0 to 21.0
- no_young_children : Continuous Data from 0.0 to 3.0
- no_older_children : Continuous Data from 0.0 to 6.0
- foreign : Categorical Data from no to yes

Datasets

1. Cubic Zirconia

Here were the first five observations of this data set :

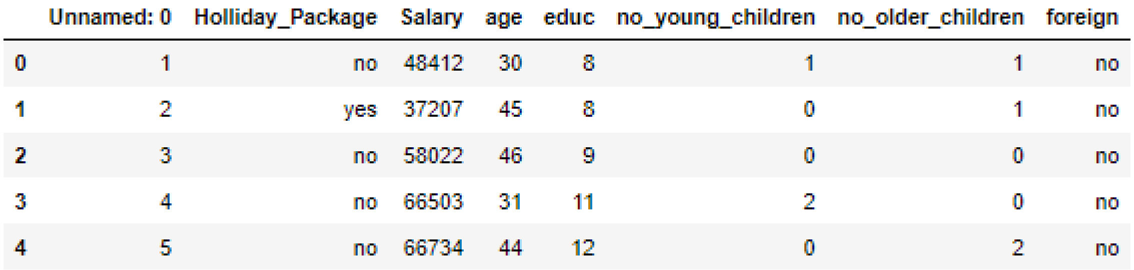


	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Figure 3 : First Five Observations(Cubic Zirconia)

2. Holiday Package

Here were the first five observations of this data set :



	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

Figure 4 : First Five Observations (Holiday Package)

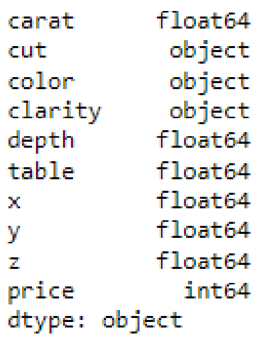
Note : We observed that the very first column 'Unnamed: 0' was not of any use as the default index was already provided so, we must drop that column from both the datasets.

Data Analysis

- Data Types

○ Cubic Zirconia

The data types of variables in the data set were :



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

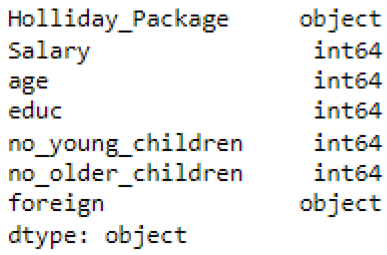
Figure 5 : Data Type (Cubic Zirconia)

Interpretations :

As per the insights, 1 column consists of integer values , while on the other hand 6 columns had float values and 3 columns had object data types values.

○ Holiday Package

The data types of variables in the data set were :



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

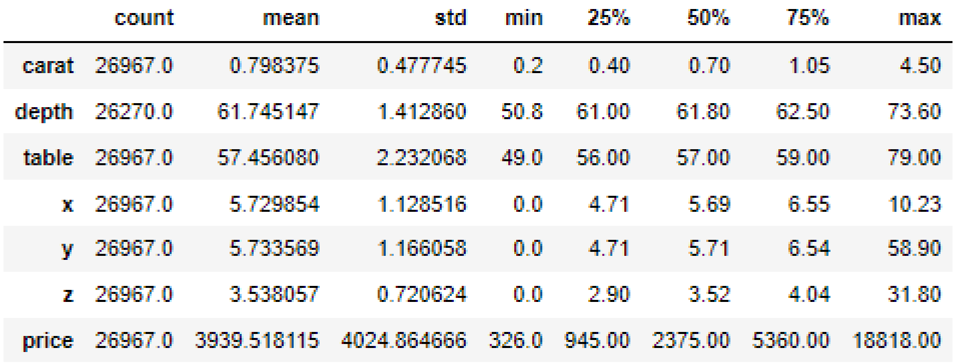
Figure 6 : Data Type (Holiday Package)

Interpretations :

There were 5 integer data types while there were only 2 object data types.

- Descriptive Statistics

1. Cubic Zirconia



	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

Figure 7 : Descriptive Statistics (Cubic Zirconia)

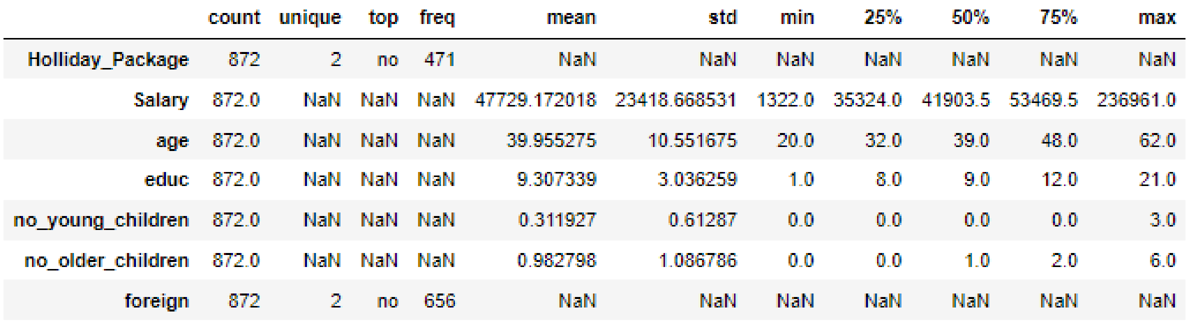
Interpretations :

- Almost each and every column except for 'price', the mean and median between them were almost equal to each other, which indicates the very less skewness exists in their distribution.

- 'Price' ranges from minimum of 326 to 18818. Average price is 3939.518115 and median sale is 75 dollars indicating that the distribution is right skewed.

- Here the minimum value for 'x', 'y', 'z' variables were shown 0 which can be considered as absurd because all these three variables represent length, width and height of the cubic zirconia and these measurements cannot be zero because it would mean that cubic zirconia doesn't exist.

2. Holiday Package



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Holliday_Package	872	2	no	471	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Salary	872.0	NaN	NaN	NaN	47729.172018	23418.668531	1322.0	35324.0	41903.5	53469.5	236961.0
age	872.0	NaN	NaN	NaN	39.955275	10.551675	20.0	32.0	39.0	48.0	62.0
educ	872.0	NaN	NaN	NaN	9.307339	3.036259	1.0	8.0	9.0	12.0	21.0
no_young_children	872.0	NaN	NaN	NaN	0.311927	0.61287	0.0	0.0	0.0	0.0	3.0
no_older_children	872.0	NaN	NaN	NaN	0.982798	1.086786	0.0	0.0	1.0	2.0	6.0
foreign	872	2	no	656	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure 8 : Descriptive Statistics (Holiday Package)

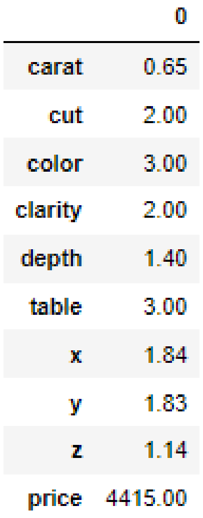
Interpretations :

- For ‘age’, ‘no_older_children’ & ‘educ’ the mean and median were almost equal to each other which indicates the very less skewness exists in their distribution.
- 'Salary' ranges from minimum of 1322 to 23661.0 . Average salary was 47729.172 and median salary was 41903.5 indicating that the distribution is right skewed.

- IQR

1. Cubic Zirconia

The Inter Quartile Range for the data set was as follows :



	0
carat	0.65
cut	2.00
color	3.00
clarity	2.00
depth	1.40
table	3.00
x	1.84
y	1.83
z	1.14
price	4415.00

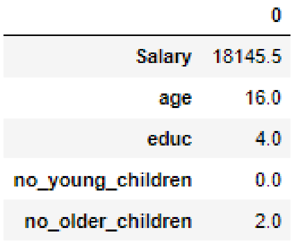
Figure 9 : IQR (Cubic Zirconia)

Interpretations :

- We inferred that ‘cut’ & ‘table’ column had the highest IQR value meaning the range between quantile 1 & quantile 3 were very high as compared to others in the dataset.

2. Holiday Package

The Inter Quartile Range for this data set was as follows :



	0
Salary	18145.5
age	16.0
educ	4.0
no_young_children	0.0
no_older_children	2.0

Figure 10 : IQR (Holiday Package)

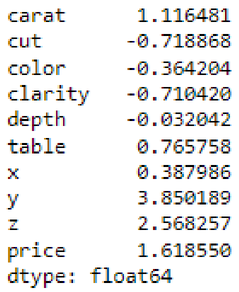
Interpretations :

- We inferred that 'Salary' variable had very high IQR value meaning the range between quantile 1 & quantile 3 were very high as compared to others in the dataset.

- Skewness

1. Cubic Zirconia

The skewness of the data set was as follows :



```
carat       1.116481
cut        -0.718868
color      -0.364204
clarity    -0.710420
depth      -0.032042
table       0.765758
x           0.387986
y           3.850189
z           2.568257
price       1.618550
dtype: float64
```

Figure 11 : Skewness (Cubic Zirconia)

Interpretations :

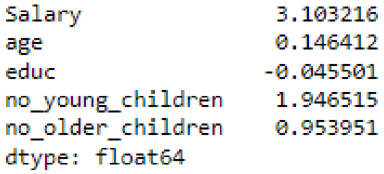
-We observed that "y" & "z" column had the maximum skewness among all.

- Among every feature in the dataset, "cut", "color", "clarity" & "depth" had negative value for skewness which indicated that these 4 columns are negatively left skewed.

- "x" column had skewness value of 0.387 which was close to 0 that means this column was close to normal distribution.

2. Holiday Package

The skewness of this data set was as follows :



```
Salary               3.103216
age                  0.146412
educ                -0.045501
no_young_children    1.946515
no_older_children    0.953951
dtype: float64
```

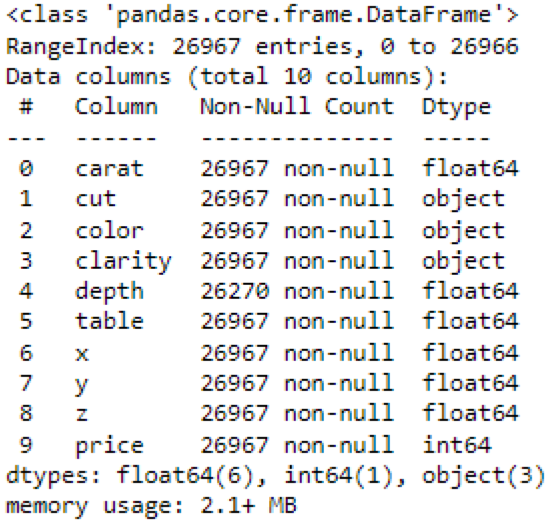
Figure 12 : Skewness (Holiday Package)

Interpretations :

-We observed that "Salary" & "no_young_children" variable had the maximum skewness among all.
- Among all the features, 'educ' was the only variable which had negative value for skewness which confirmed that it was negatively left skewed.
- 'age' attribute had very less skewness value which indicated that it was very close to normal distribution.

Checking For Null Values

1. Cubic Zirconia



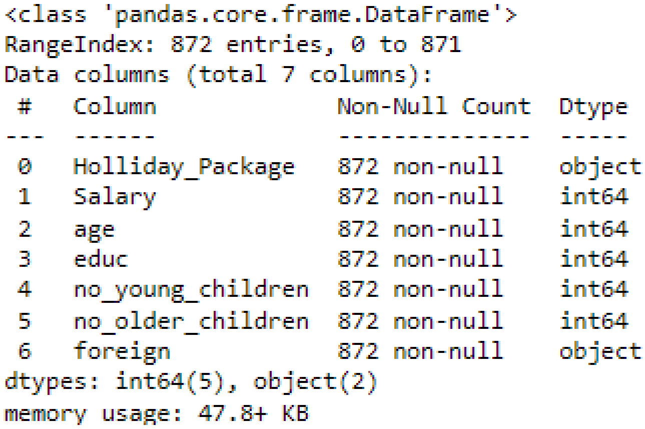
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object 
 2   color    26967 non-null  object 
 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
```

Figure 13 : Data Info & Null Values Check (Cubic Zirconia)

Interpretations :

- There were zero null values present in the data set.
- Also, we noted that the shape/ dimensions of data set is (26967, 10) which means that there were 26967 entries and 10 columns in the data set.

2. Holiday Package



```
<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
```

Figure 14 : Data Info & Check Null Values(Holiday Package)

Interpretations :

- As per the code, there were zero null values present in the dataset.
- Also, we noted that the shape/ dimensions of data set is (872,7) which means that there were 872 entries and 7 columns in the data set.

Problem 1 – Linear Regression

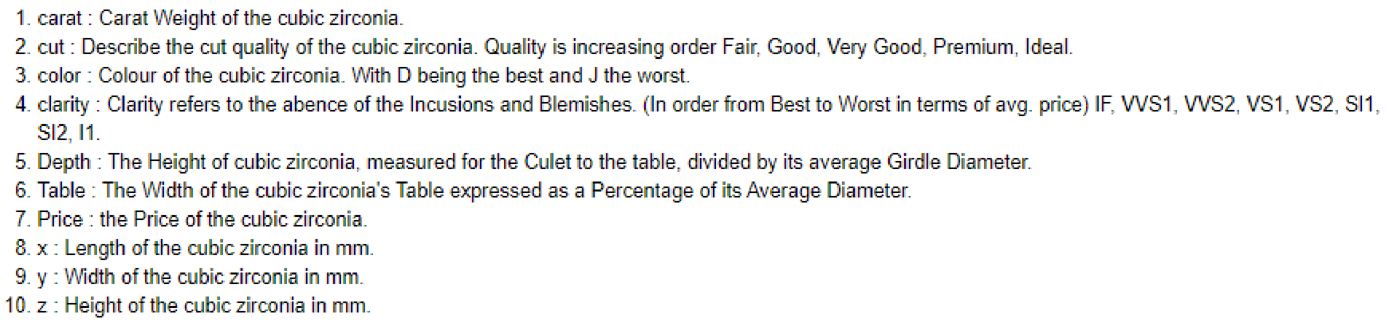
You are hired by a company named Gem Stones Co Ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of approximately 27,000 pieces of cubic zirconia (which is an inexpensive synthesized diamond alternative with similar qualities of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

Problem 1.1 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.

At first we loaded the Data Dictionary for understanding of column name for data set "Cubic Zirconia".

The Data Dictionary is as follows :



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. color : Colour of the cubic zirconia. With D being the best and J the worst.
4. clarity : Clarity refers to the abence of the Incusions 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 for 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.

Figure 15 : Data Dictionary (Problem 1)

EDA

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is of various types such as univariate, bi-variate and multi-variate.
After getting a brief understanding of what is EDA, we did analyze the data and here it is what we had found :

Univariate Analysis

For Univariate analysis, we plotted a Distribution plot and a Boxplot for each column provided in the data set .

The Distribution plot was used for univariant set of observations and visualizes it through a histogram i.e. only one observation and hence we choose one particular column of the dataset.

And, Boxplot was used as a measure of how well the data is distributed in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data and also shows us whether there are outliers or not.

Here were the Boxplot for **Cubic Zirconia** data set :

1. **Carat**

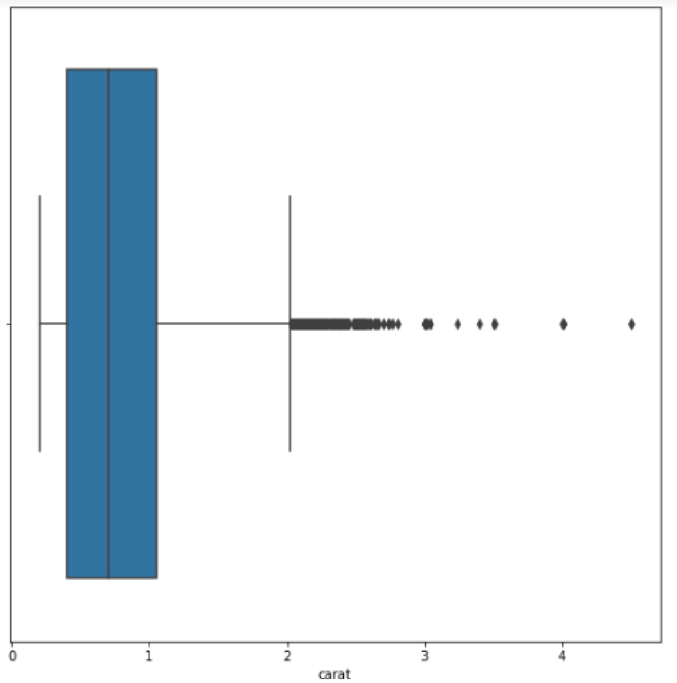


Figure 16 : Boxplot (Carat)

- The boxplot told us that there were many extreme values which exceed the upper limit but these values were not outliers.
- It was because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there. This boxplot indicated that it was positively right skewed.

2. **Cut**

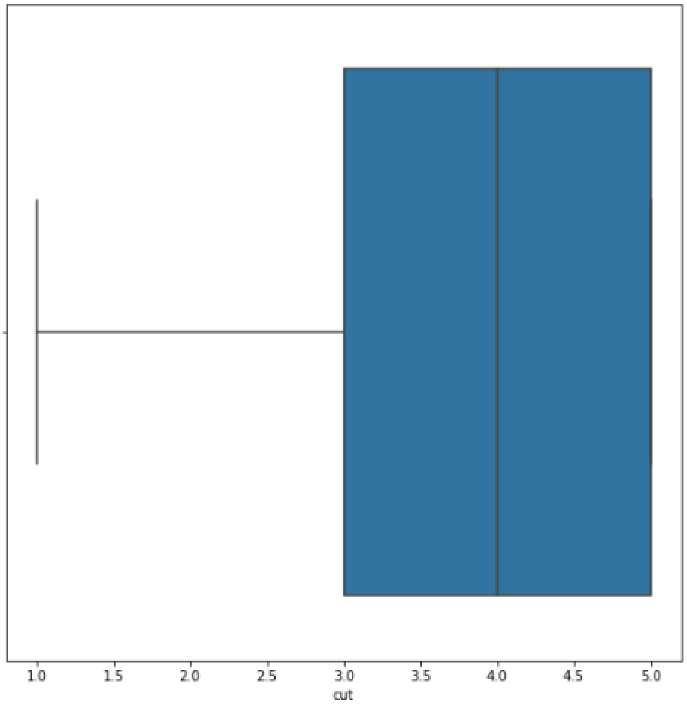


Figure 17 : Boxplot (Cut)

- The boxplot of 'cut' told us that there were no extreme values at either of the ends. The median tends to be at the right side of the boxplot indicating that 'color' variable was negatively left skewed.
- The 'cut' variable showed no outliers as per the Boxplot which means all values were laying inside the range from minimum to maximum.

3. Color

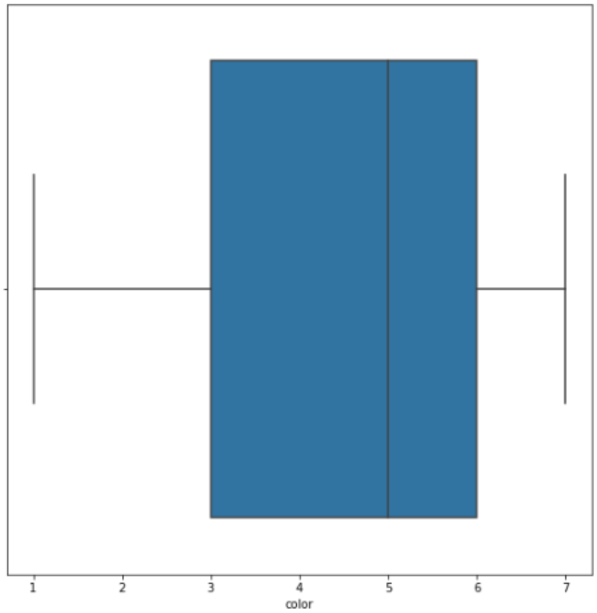


Figure 18 : Boxplot (color)

- The ‘color’ variable showed no outliers as per the Boxplot which means all values were laying inside the range from minimum to maximum.
- The boxplot tends to be at the right side of the boxplot indicating that ‘color’ variable was negatively left skewed.

4. Clarity

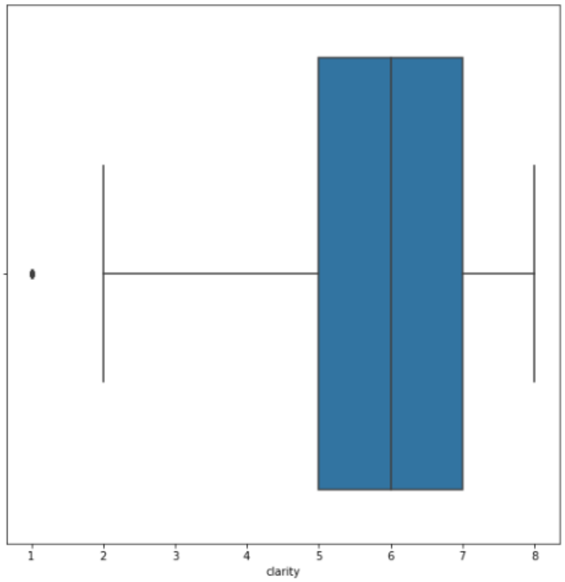


Figure 19 : Boxplot(Clarity)

- The ‘clarity’ variable showed some extreme values on the left of minimum value as per the Boxplot which means most values were laying inside the range from minimum to maximum except for some values
- The boxplot tends to be at the right side of the boxplot indicating that ‘depth’ variable was negatively left skewed.

5. Depth

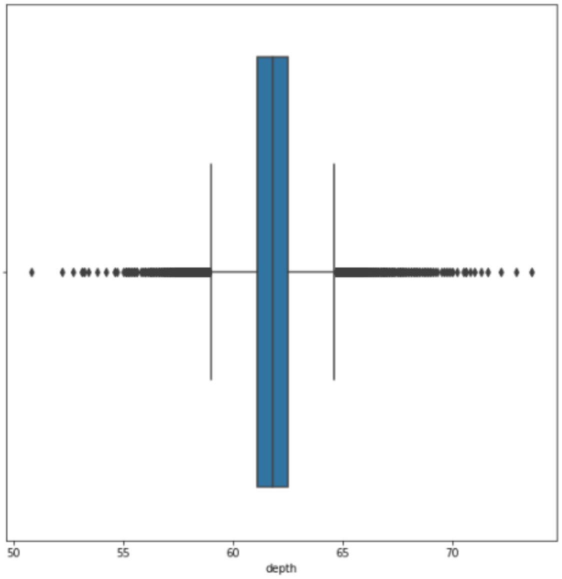


Figure 20 : Boxplot (Depth)

- As per the Boxplot of ‘Depth’ , there were many extreme values present at both left and right side of the boxplot. These extreme values were not outliers because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there.
- The boxplot indicated that ‘depth’ variable was negatively left skewed.

6. Table

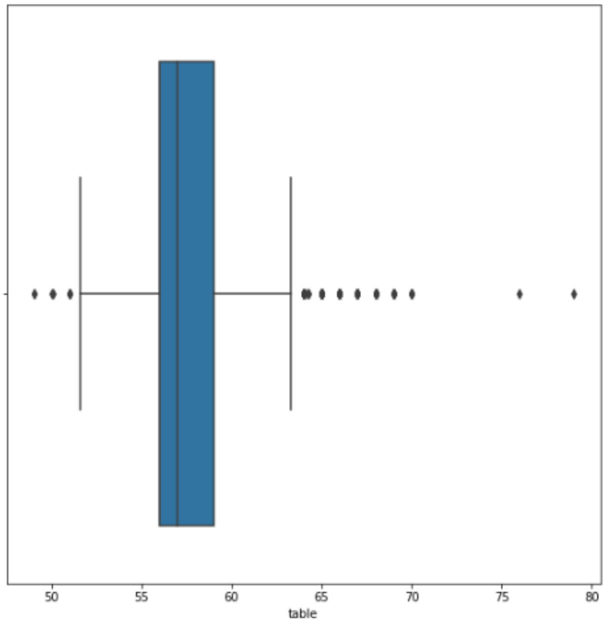


Figure 21 : Boxplot (Table)

- As per the Boxplot of ‘table’, there were many extreme values present at both left and right side of the boxplot. These extreme values were not outliers because these extreme values were checked thoroughly and no absurd or irrelevant values were not present there.
- The boxplot tends to be at the left side of the boxplot indicating that ‘table’ variable was positively right skewed.

7. x

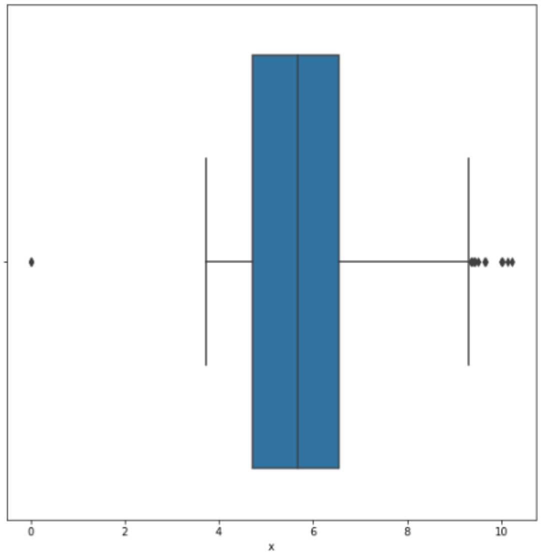


Figure 22 : Boxplot (x)

- The ‘x’ variable showed some extreme values on the left & right side of minimum value as per the Boxplot which means most values were laying inside the range from minimum to maximum except for some values. These extreme values were absurd & meaningless as ‘x’ variable represents the length of the cubic zirconia in mm and length cannot be zero that why these values were considered as outliers.
- The boxplot tends to be at the right side of the boxplot indicating that ‘x’ variable was negatively left skewed.

8. y

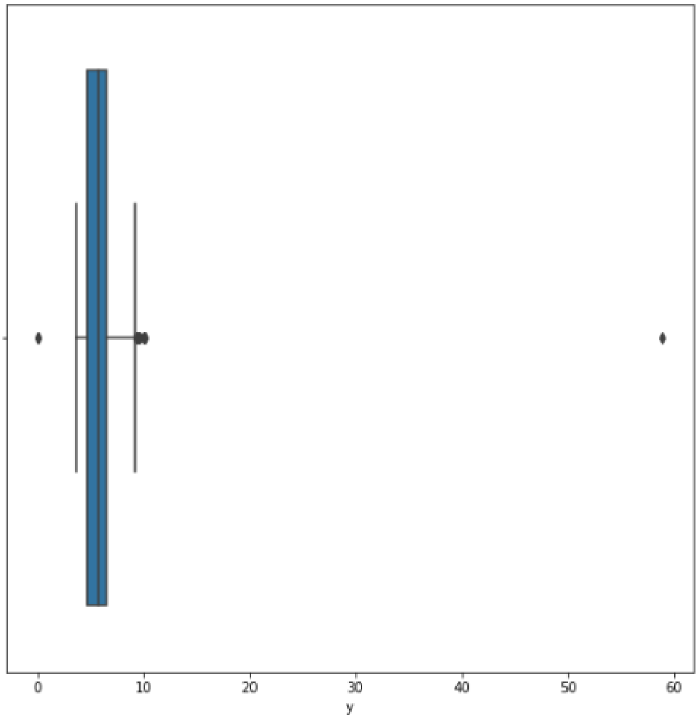


Figure 23 : Boxplot (y)

- The ‘y’ variable showed some extreme values on the left & right side of minimum value as per the Boxplot which means most values were laying inside the range from minimum to maximum except for some values. These extreme values were absurd & meaningless as ‘y’ variable represents the width of the cubic zirconia in mm and width cannot be zero that why these values were considered as outliers.
- The boxplot tends to be at the right side of the boxplot indicating that ‘y’ variable was positively right skewed.

9. z

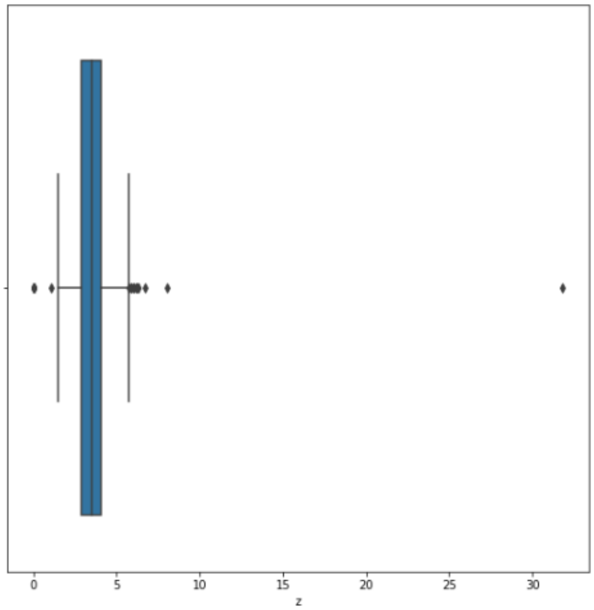


Figure 24 : Boxplot (z)

- The ‘z’ variable showed some extreme values on the left & right side of minimum value as per the Boxplot which means most values were laying inside the range from minimum to maximum except for some values. These extreme values were absurd & meaningless as ‘z’ variable represents the height of the cubic zirconia in mm and height cannot be zero that why these values were considered as outliers.
- The boxplot tends to be at the right side of the boxplot indicating that ‘y’ variable was positively right skewed.

10. Price

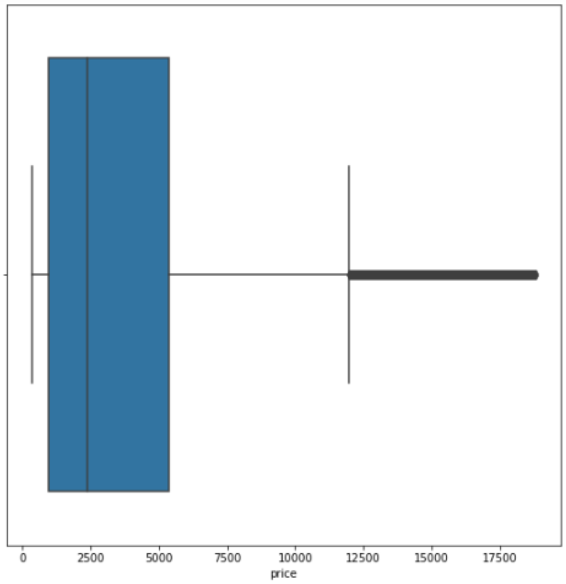


Figure 25 : Boxplot (price)

- The ‘price’ variable showed some extreme values on the right side of minimum value as per the Boxplot which means many values were laying inside the range from minimum to maximum except for some values which lie outside the boxplot. These extreme values were not absurd & meaningless as ‘price’ variable represents the price of the cubic zirconia and ‘price’ can differ from high to low as per the features of the cubic zirconia.
- The boxplot tends to be at the right side of the boxplot indicating that ‘price’ variable was positively right skewed.

Conclusion :

- We noticed that some other variables also contain extreme low and high values such as 'carat', 'clarity', 'depth', 'table' & 'price'. But we cannot consider them as outliers as those values were looked thoroughly and concluded that those values were logical and appropriate.
- Hence, they were not been considered as outliers and therefore shall not be removed/treated and we used those values in the data set while building the linear regression model.
- However, ‘x’ , ‘y’ & ‘z’ variables showed outliers which needed to treated/removed before building the linear regression model.

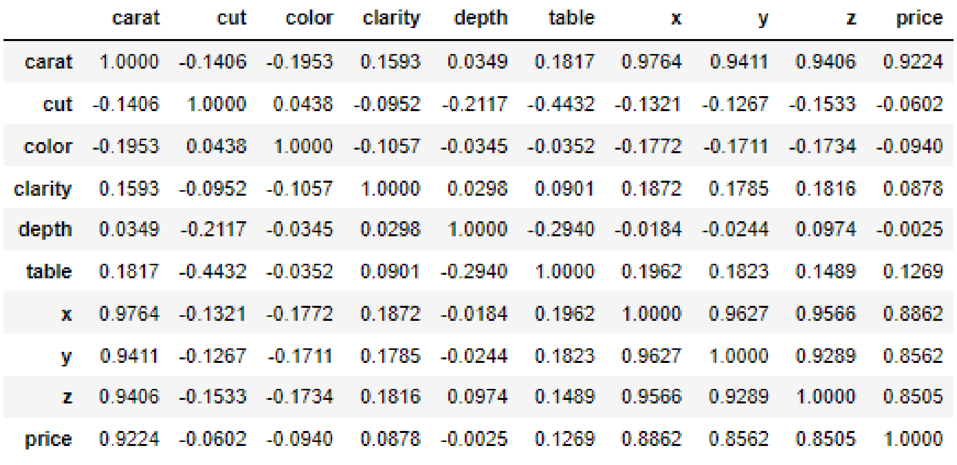
Bi-Variate Analysis :

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

There are many types of bivariate analysis such as scatter plot**,** regression analysis, correlation matrix analysis and much more.

1. **Correlation Matrix :**

For this data, we did the correlation matrix and find out many insights from it. It is as follows :



	carat	cut	color	clarity	depth	table	x	y	z	price
carat	1.0000	-0.1406	-0.1953	0.1593	0.0349	0.1817	0.9764	0.9411	0.9406	0.9224
cut	-0.1406	1.0000	0.0438	-0.0952	-0.2117	-0.4432	-0.1321	-0.1267	-0.1533	-0.0602
color	-0.1953	0.0438	1.0000	-0.1057	-0.0345	-0.0352	-0.1772	-0.1711	-0.1734	-0.0940
clarity	0.1593	-0.0952	-0.1057	1.0000	0.0298	0.0901	0.1872	0.1785	0.1816	0.0878
depth	0.0349	-0.2117	-0.0345	0.0298	1.0000	-0.2940	-0.0184	-0.0244	0.0974	-0.0025
table	0.1817	-0.4432	-0.0352	0.0901	-0.2940	1.0000	0.1962	0.1823	0.1489	0.1269
x	0.9764	-0.1321	-0.1772	0.1872	-0.0184	0.1962	1.0000	0.9627	0.9566	0.8862
y	0.9411	-0.1267	-0.1711	0.1785	-0.0244	0.1823	0.9627	1.0000	0.9289	0.8562
z	0.9406	-0.1533	-0.1734	0.1816	0.0974	0.1489	0.9566	0.9289	1.0000	0.8505
price	0.9224	-0.0602	-0.0940	0.0878	-0.0025	0.1269	0.8862	0.8562	0.8505	1.0000

Figure 26 : Correlation Matrix (Cubic Zirconia)

Interpretations :

- It shows the Correlation Matrix of the original dataset using Pearson method.
- After this, a heatmap was created w.r.t to correlation matrix for visualization.

2. Heatmap :

A **Correlation Heatmap** is a rectangular representation of data and it repeats the same data description twice because the categories are repeated on both axis for computing analysis. Hence, the same result is obtained twice. A correlation heatmap that presents data only once without repetition that is categories are correlated only once is known as a **Triangle Correlation Heatmap**.

Since data is symmetric across the diagonal from left-top to right bottom the idea of obtaining a triangle correlation heatmap is to remove data above it so that it is depicted only once. The elements on the diagonal are the parts where categories of the same type correlate.

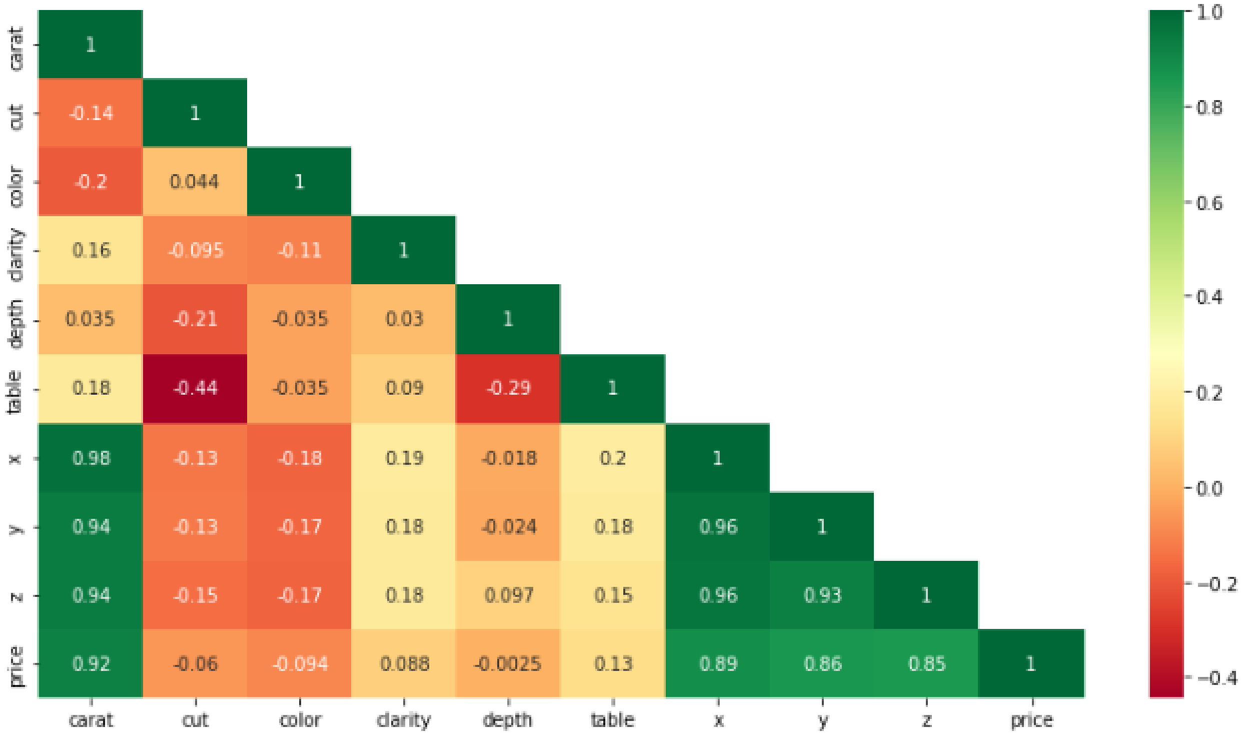


Figure 27 : Heatmap (Cubic Zirconia)

Interpretations :

- We noticed that there was a very high positive correlation value between 'carat' & other variables ('x', 'y', 'z') where 'x' has highest value of 98% and next to it were 'y' & 'z' having 94% value of correlation. This indicated that as the length, width, and height of the zirconia (in mm) , the price of the carat also increases.

- In addition to it, after these length, width & height of zirconia('x', 'y', 'z'), the next variable with a high positive correlation with 'carat' variable was the 'price' variable with 92% correlation value. It meant that as the weight of the carat increases, the price of the cubic zirconia also increases.
- Moving on to 'cut' variable, it had a medium strong negative correlation relationship with 'table' variable with 44% of correlation value. This meant that, the width of the cubic zirconia expressed in percentage of its average diameter in indirectly proportional to the cut quality of cubic zirconia. In other words, as the 'table' variable value increases, the cut quality decreases resulting in only fair of good quality of cubic zirconia.
- We inferred that, 'color' variable had very low correlation value with all the other variables. Only 'x', 'y', 'z' had some of weak relationship with 'color' variable.
- Similarly, 'clarity' variable also had a mild correlation with 'x', 'y', 'z' with value of 18-19%.

Multivariate Analysis :

- **Multivariate analysis** (**MVA**) is a Statistical procedure for analysis of data involving more than one type of measurement or observation. It may also mean solving problems where more than one dependent variable is analyzed simultaneously with other variables . Multivariate analysis is one of the most useful methods to determine relationships and analyze patterns among large sets of data. It is particularly effective in minimizing bias if a structured study design is employed. However, the complexity of the technique makes it a less sought-out model for novice research enthusiasts. Therefore, although the process of designing the study and interpretation of results is a tedious one, the techniques stand out in finding the relationships in complex.

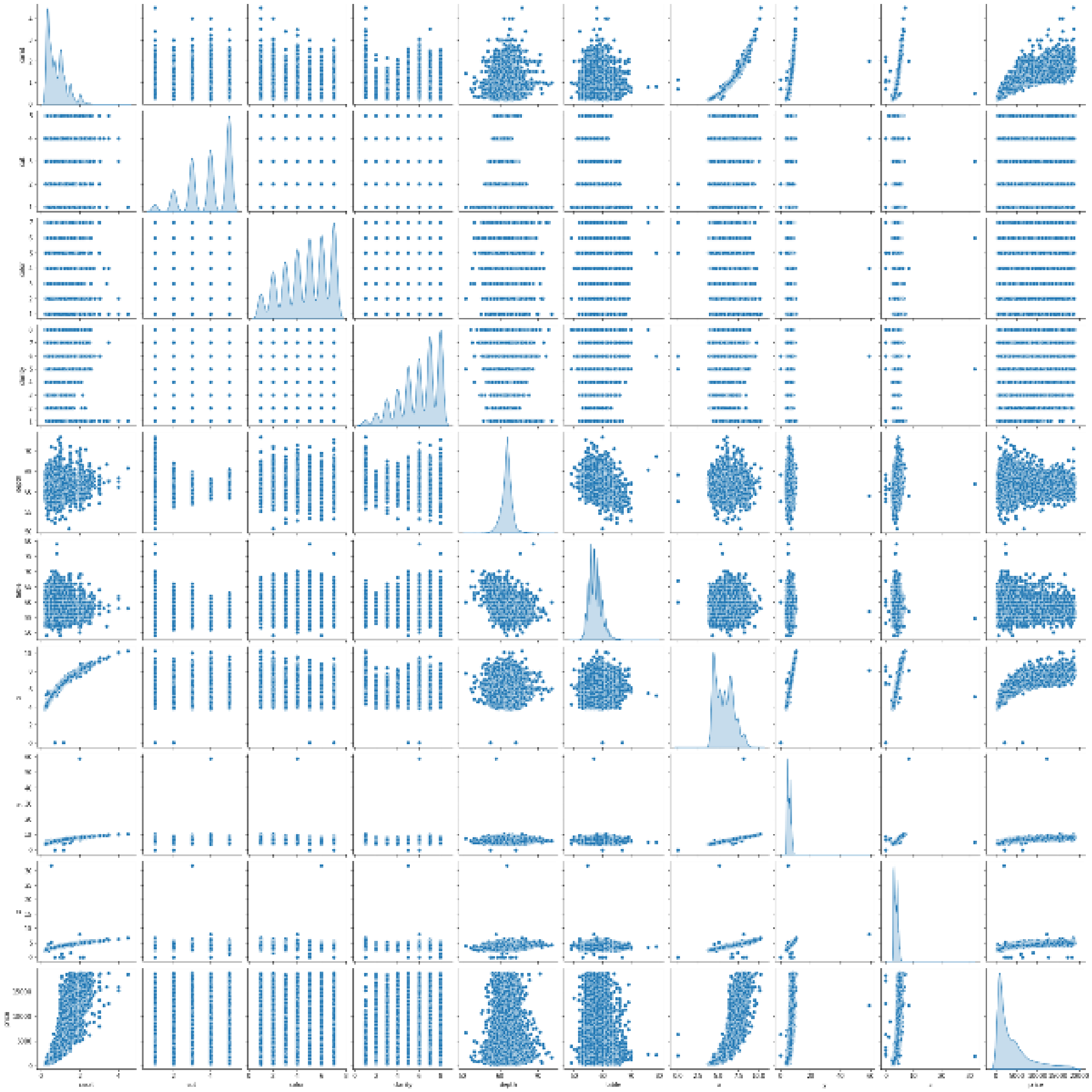
Pairplot :

- In this we found out the **Pairplot** of the original dataset .

- **Pairplot** function allows the users to create an axis grid via which each numerical variable stored in data is shared across the X- and Y-axis in the structure of columns and rows. We can create the Scatter plots in order to display the pairwise relationships in addition to the distribution plot displaying the data distribution in the column diagonally.

- The **pairplot()** function can also be used to showcase the subset of variables, or we can plot different types of variables on rows and columns.

- The Pairplot of the following data set was as follows :



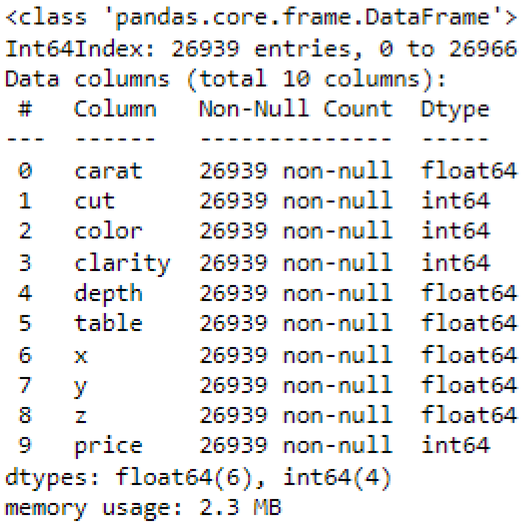
Interpretations :

- The data distribution across various dimensions except 'depth' do not look normal. Only 'depth' looked close to a normal distribution.
- It has been observed that heteroscedasticity exists between 'carat' & 'depth' as well as between 'carat' & 'table' which will impact the model accuracy.
- The 'x' variable had a strong positive curvilinear relationship with 'carat' variable in addition with outliers being present at the lower left of the scatter plot existing between them as shown in the pairplot.

- Furthermore, 'y' variable also had a strong relation with 'carat' variable. This relation also comprises of outliers at extreme right side & extreme bottom left as being presented in the scatter plot shown inside the pairplot. This relation tells us that a little increase in 'y' variable (width of cubic zirconia in mm) can lead to very large increment in the 'carat' weight
- In addition, 'z' variable also had a strong relation with 'carat' variable in addition to outliers being present at the extreme left and on the extreme bottom right as well. This relation tells us that a little increase in 'z' variable (height of cubic zirconia in mm) can lead to very large increment in the 'carat' weight.

Outlier Treatment

- The presence of outliers and influential cases can dramatically change the magnitude of regression coefficients and even the direction of coefficient signs (i.e., from positive to negative or vice versa).
- So, these outliers must be find out and shall be treated in order to perform linear regression.
- Using the user defined function and IQR outliers values from ‘x’, ‘y’, ‘z’ were detected and hence after that those values were being removed from the data set.



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26939 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26939 non-null  float64
 1   cut      26939 non-null  int64  
 2   color    26939 non-null  int64  
 3   clarity  26939 non-null  int64  
 4   depth    26939 non-null  float64
 5   table    26939 non-null  float64
 6   x        26939 non-null  float64
 7   y        26939 non-null  float64
 8   z        26939 non-null  float64
 9   price    26939 non-null  int64  
dtypes: float64(6), int64(4)
memory usage: 2.3 MB
```

Figure 28 : Data Info After Outlier Treatment (Cubic Zirconia)

Problem 1.2 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?

Linear Regression Model Using Statsmodels Library

The data set had been pre processed and now the we further proceed to build the linear regression model. We builded the Linear Regression model using statsmodels library.

Checking For Multicollinearity

The multicollinearity was being for the base model using Variance Inflation Factor(VIF).

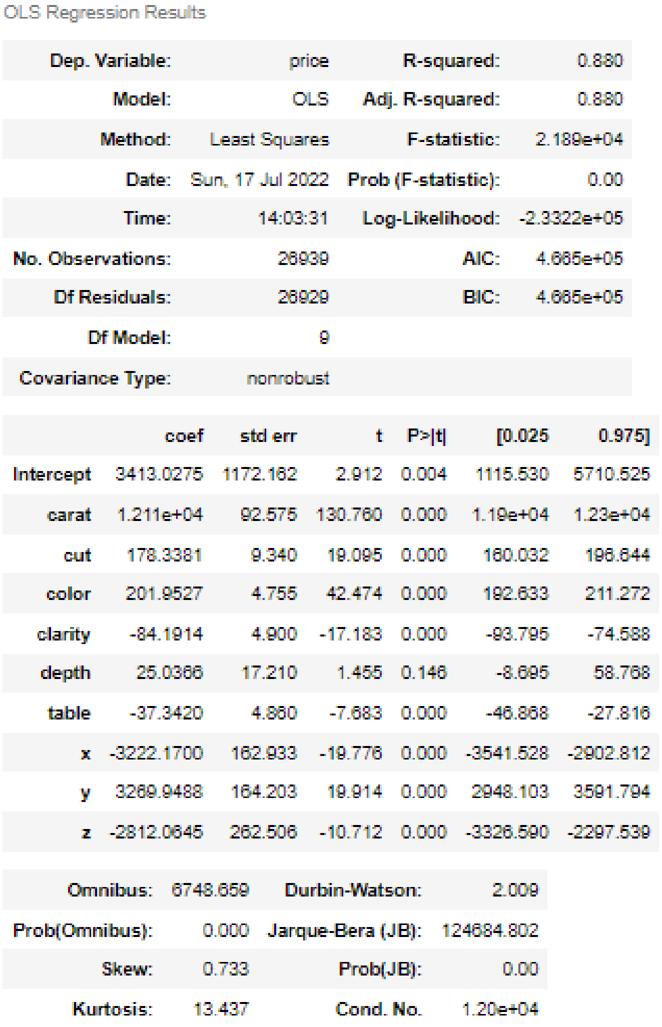
VIF is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable Is highly collinear with the other variables in the model.

Base Model / Model 1

We build a model without scaling the data.

The linear regression model was build using the ols function and then fit that model and finally print the summary of the **base model**.



OLS Regression Results

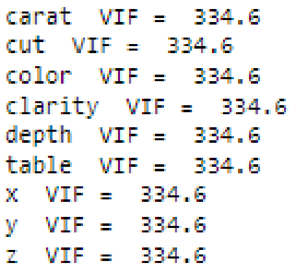
Dep. Variable:	price	R-squared:	0.880
Model:	OLS	Adj. R-squared:	0.880
Method:	Least Squares	F-statistic:	2.189e+04
Date:	Sun, 17 Jul 2022	Prob (F-statistic):	0.00
Time:	14:03:31	Log-Likelihood:	-2.3322e+05
No. Observations:	26939	AIC:	4.665e+05
Df Residuals:	26929	BIC:	4.665e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3413.0275	1172.162	2.912	0.004	1115.530	5710.525
carat	1.211e+04	92.575	130.760	0.000	1.19e+04	1.23e+04
cut	178.3381	9.340	19.095	0.000	160.032	196.644
color	201.9527	4.755	42.474	0.000	192.633	211.272
clarity	-84.1914	4.900	-17.183	0.000	-93.795	-74.588
depth	25.0366	17.210	1.455	0.146	-8.695	58.768
table	-37.3420	4.860	-7.683	0.000	-46.868	-27.816
x	-3222.1700	162.933	-19.776	0.000	-3541.528	-2902.812
y	3269.9488	164.203	19.914	0.000	2948.103	3591.794
z	-2812.0645	262.506	-10.712	0.000	-3326.590	-2297.539

Omnibus:	6748.659	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	124684.802
Skew:	0.733	Prob(JB):	0.00
Kurtosis:	13.437	Cond. No.	1.20e+04

Figure 29 : Base Model Summary (Linear Regression)

The coefficients for carat , cut , color, y were very large while for depth, table, clarity, x & z the coefficient value were very low but that doesn't really mean they were more influential compared to price. It was simply because out data set has not been normalized and the data range for each of these columns vary widely. VIF for base model was as follows :



```
carat  VIF =  334.6
cut  VIF =  334.6
color  VIF =  334.6
clarity  VIF =  334.6
depth  VIF =  334.6
table  VIF =  334.6
x  VIF =  334.6
y  VIF =  334.6
z  VIF =  334.6
```

Figure 30 : VIF for Base Model (Linear Regression)

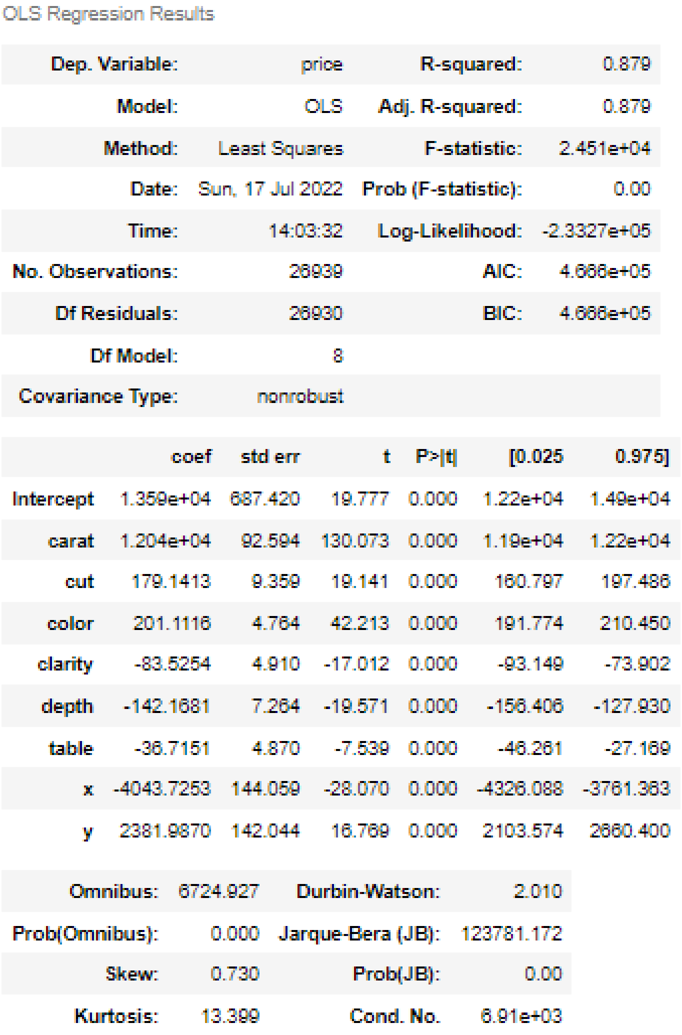
Interpretations :

- The R2 value and Adjusted R2 value are almost equal.
- All variables had same VIF . This indicates that there was very high multicollinearity between the independent variables and the dependent variable.
- Thus, we need to remove those variables with high correlation and considering VIF values in order to remove multicollinearity for a better model.

2nd Iteration Model

As our priority was to remove the multicollinearity first, we referred to the pairplot of the dataset. From there we noticed that, there was very high correlation between 'z' & 'price which could lead to high multicollinearity in the linear regression model as shown in the pairplot(scatter plot of z and price). So, this variable was removed/not included in the model building.

The 2nd Iteration model was as follows :



OLS Regression Results

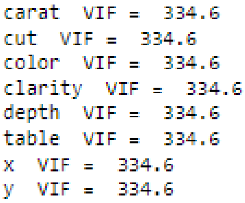
Dep. Variable:	price	**R-squared:**	0.879
Model:	OLS	**Adj. R-squared:**	0.879
Method:	Least Squares	**F-statistic:**	2.451e+04
Date:	Sun, 17 Jul 2022	**Prob (F-statistic):**	0.00
Time:	14:03:32	**Log-Likelihood:**	-2.3327e+05
No. Observations:	26939	**AIC:**	4.666e+05
Df Residuals:	26930	**BIC:**	4.666e+05
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.359e+04	687.420	19.777	0.000	1.22e+04	1.49e+04
carat	1.204e+04	92.594	130.073	0.000	1.19e+04	1.22e+04
cut	179.1413	9.359	19.141	0.000	160.797	197.486
color	201.1116	4.764	42.213	0.000	191.774	210.450
clarity	-83.5254	4.910	-17.012	0.000	-93.149	-73.902
depth	-142.1681	7.264	-19.571	0.000	-156.406	-127.930
table	-36.7151	4.870	-7.539	0.000	-46.261	-27.169
x	-4043.7253	144.059	-28.070	0.000	-4326.088	-3761.363
y	2381.9870	142.044	16.769	0.000	2103.574	2660.400

Omnibus:	6724.927	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	123781.172
Skew:	0.730	**Prob(JB):**	0.00
Kurtosis:	13.399	**Cond. No.**	6.91e+03

Figure 31 : 2nd Iteration Model (Linear Regression)

VIF for 2nd Iteration Model was as follows :



```
carat  VIF =  334.6
cut  VIF =  334.6
color  VIF =  334.6
clarity  VIF =  334.6
depth  VIF =  334.6
table  VIF =  334.6
x  VIF =  334.6
y  VIF =  334.6
```

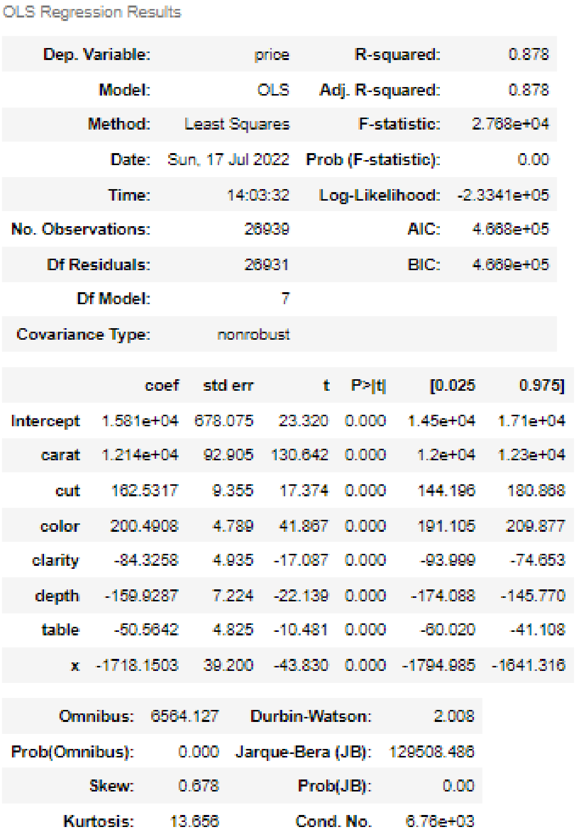
Figure 32 : VIF for 2nd Iteration Model (Linear Regression)

Interpretations :

- As we removed 'z' variable , we got to notice that the p value for depth was reduced and was below 0.05 as compared to the previous base model where the p value for 'depth' was greater than 0.05 .
- Now, all the attributes had p value lower than 0.05 but the VIF value for all the variables are very high and same also. Thus we need to remove more variables in order to build a more efficient model.

3rd Iteration Model

Again we referred to the pairplot of the data to find out which independent variables had very high correlation and we find out that ‘y’ had a high correlation with ‘price’. Thus, we need to remove it and then build another model.



OLS Regression Results

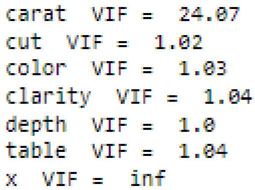
Dep. Variable:	price	R-squared:	0.878
Model:	OLS	Adj. R-squared:	0.878
Method:	Least Squares	F-statistic:	2.768e+04
Date:	Sun, 17 Jul 2022	Prob (F-statistic):	0.00
Time:	14:03:32	Log-Likelihood:	-2.3341e+05
No. Observations:	26939	AIC:	4.668e+05
Df Residuals:	26931	BIC:	4.669e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.581e+04	678.075	23.320	0.000	1.45e+04	1.71e+04
carat	1.214e+04	92.905	130.642	0.000	1.2e+04	1.23e+04
cut	162.5317	9.355	17.374	0.000	144.196	180.868
color	200.4908	4.789	41.867	0.000	191.105	209.877
clarity	-84.3258	4.935	-17.087	0.000	-93.999	-74.653
depth	-159.9287	7.224	-22.139	0.000	-174.088	-145.770
table	-50.5642	4.825	-10.481	0.000	-60.020	-41.108
x	-1718.1503	39.200	-43.830	0.000	-1794.985	-1641.316

Omnibus:	6564.127	Durbin-Watson:	2.008
Prob(Omnibus):	0.000	Jarque-Bera (JB):	129508.486
Skew:	0.678	Prob(JB):	0.00
Kurtosis:	13.656	Cond. No.	6.76e+03

Figure 33 : 3rd Iteration Model (Linear Regression)

VIF value for 3rd Iteration Value was as follows :



```
carat  VIF =  24.07
cut  VIF =  1.02
color  VIF =  1.03
clarity  VIF =  1.04
depth  VIF =  1.0
table  VIF =  1.04
x  VIF =  inf
```

Figure 34 : VIF for 3rd Iteration Model (Linear Regression)

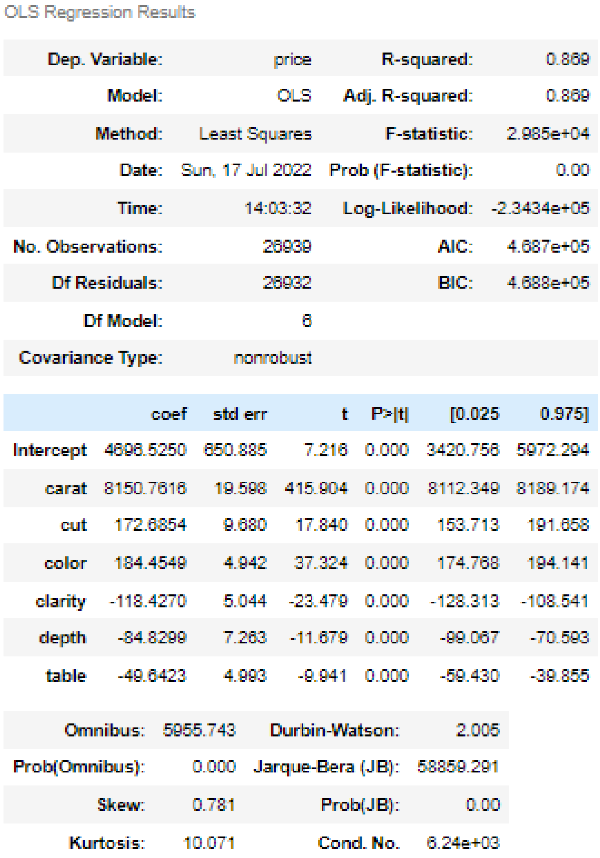
Interpretations :

- As we can see that the VIF had been improved as we removed the 'y' variable.
- But 'x' variable is showing infinite VIF value which is absurd and thus is should removed.

4th Iteration Model

As per the 3rd Iteration Model, the VIF value for ‘x’ variable was infinite which clearly indicated the presence of multicollinearity and this variable should be removed.

Using the ols function, we build the 4th Iteration Model . It was as follows :



OLS Regression Results

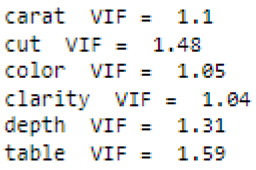
Dep. Variable:	price	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	2.985e+04
Date:	Sun, 17 Jul 2022	Prob (F-statistic):	0.00
Time:	14:03:32	Log-Likelihood:	-2.3434e+05
No. Observations:	26939	AIC:	4.687e+05
Df Residuals:	26932	BIC:	4.688e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4696.5250	650.885	7.216	0.000	3420.756	5972.294
carat	8150.7616	19.598	415.904	0.000	8112.349	8189.174
cut	172.6854	9.680	17.840	0.000	153.713	191.658
color	184.4549	4.942	37.324	0.000	174.768	194.141
clarity	-118.4270	5.044	-23.479	0.000	-128.313	-108.541
depth	-84.8299	7.263	-11.679	0.000	-99.067	-70.593
table	-49.6423	4.993	-9.941	0.000	-59.430	-39.855

Omnibus:	5955.743	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58859.291
Skew:	0.781	Prob(JB):	0.00
Kurtosis:	10.071	Cond. No.	6.24e+03

Figure 35 : 4th Iteration Model (Linear Regression)

VIF value for 4TH Iteration Model was as follows :



```
carat  VIF =  1.1
cut  VIF =  1.48
color  VIF =  1.05
clarity  VIF =  1.04
depth  VIF =  1.31
table  VIF =  1.59
```

Figure 36 : VIF value for 4th Iteration Model (Linear Regression)

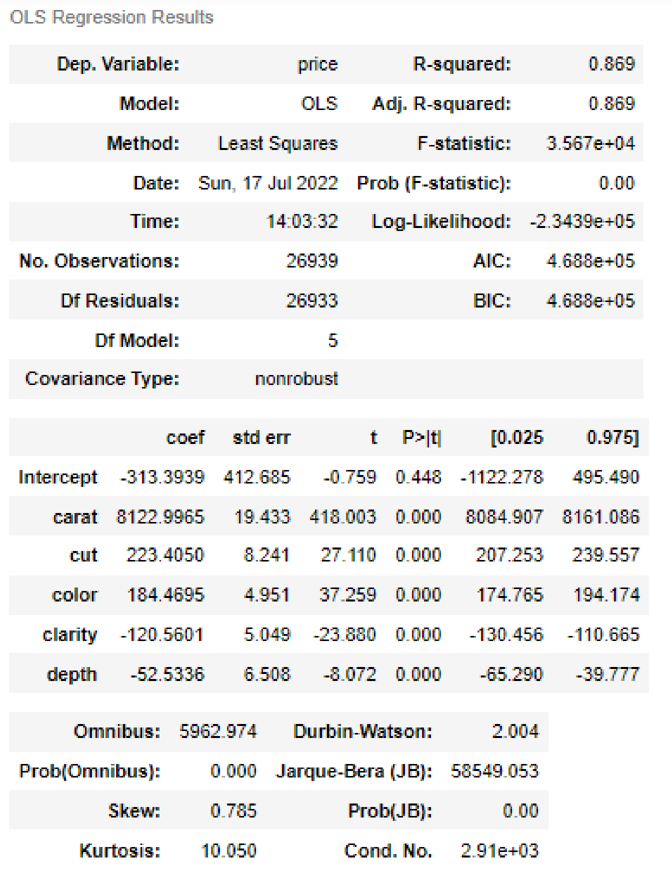
Interpretations :

- The R2 and Adj R2 was also looking good for the model.
- As inferred , we can say that VIF value had been improved a lot as compared to previous models but still 'table' variable had some high VIF value so it should be treated/removed.

5th Iteration Model

Now, we checked the 4th Iteration model and noticed that 'table' column had maximum VIF value among others. Thus, we removed that variable and build another model for more tuning.

Hence the 5th Iteration Model was as follows :



OLS Regression Results

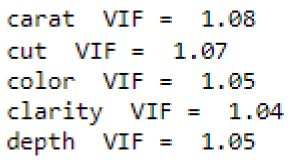
Dep. Variable:	price	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	3.567e+04
Date:	Sun, 17 Jul 2022	Prob (F-statistic):	0.00
Time:	14:03:32	Log-Likelihood:	-2.3439e+05
No. Observations:	26939	AIC:	4.688e+05
Df Residuals:	26933	BIC:	4.688e+05
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-313.3939	412.685	-0.759	0.448	-1122.278	495.490
carat	8122.9965	19.433	418.003	0.000	8084.907	8161.086
cut	223.4050	8.241	27.110	0.000	207.253	239.557
color	184.4695	4.951	37.259	0.000	174.765	194.174
clarity	-120.5601	5.049	-23.880	0.000	-130.456	-110.665
depth	-52.5336	6.508	-8.072	0.000	-65.290	-39.777

Omnibus:	5962.974	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	58549.053
Skew:	0.785	Prob(JB):	0.00
Kurtosis:	10.050	Cond. No.	2.91e+03

Figure 37 : 5th Iteration Model (Linear Regression)

VIF value for 5th Iteration Model was as follows :



```
carat  VIF =  1.08
cut  VIF =  1.07
color  VIF =  1.05
clarity  VIF =  1.04
depth  VIF =  1.05
```

Figure 38 : VIF value for 5th Iteration Model (Linear Regression)

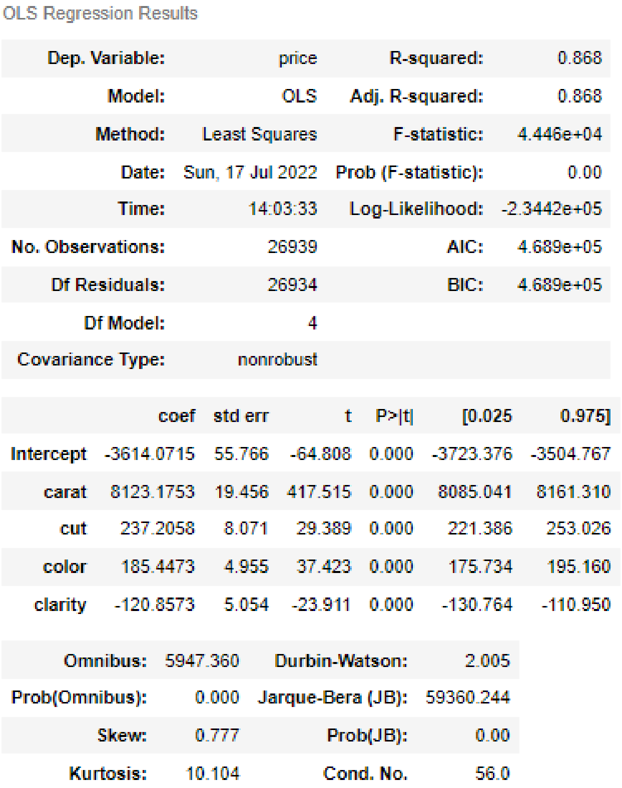
Interpretations :

- The P value had increased significantly for intercept with value greater than 0.05. So, it shall be corrected by making some changes in the model.
- The VIF values are now significant but P value is not good so that shall be treated.

- We removed 'depth' column as in the scatter plot(pairplot) , it had almost no relation with price but as we can see the coefficient value in model 5, it had value of - 52.53. This means that for every change of 1 in depth variable, the price of cubic zirconia gets decreased by 52.53. Thus 'depth' column should be removed.

6th Iteration Model

We tuned the linear regression model by removing ‘depth’ column and builded a better linear regression model . It was as follows :



OLS Regression Results

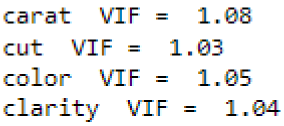
Dep. Variable:	price	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.868
Method:	Least Squares	F-statistic:	4.446e+04
Date:	Sun, 17 Jul 2022	Prob (F-statistic):	0.00
Time:	14:03:33	Log-Likelihood:	-2.3442e+05
No. Observations:	26939	AIC:	4.689e+05
Df Residuals:	26934	BIC:	4.689e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3614.0715	55.766	-64.808	0.000	-3723.376	-3504.767
carat	8123.1753	19.456	417.515	0.000	8085.041	8161.310
cut	237.2058	8.071	29.389	0.000	221.386	253.026
color	185.4473	4.955	37.423	0.000	175.734	195.160
clarity	-120.8573	5.054	-23.911	0.000	-130.764	-110.950

Omnibus:	5947.360	Durbin-Watson:	2.005
Prob(Omnibus):	0.000	Jarque-Bera (JB):	59360.244
Skew:	0.777	Prob(JB):	0.00
Kurtosis:	10.104	Cond. No.	56.0

Figure 39 : 6th Iteration Model (Linear Regression)

VIF value for 6th Iteration Model was as follows :



```
carat  VIF =  1.08
cut  VIF =  1.03
color  VIF =  1.05
clarity  VIF =  1.04
```

Figure 40 : VIF Value for 6th Iteration Model (Linear Regression)

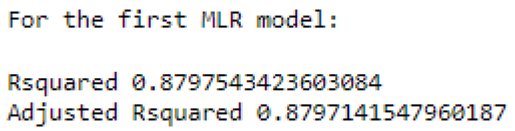
Interpretations :

- Now, all statistical parameters were being checked such as p value, VIF value etc , this model had all the appropriate things which must be required in our linear regression model.
- It had all the qualities of an ideal best regression model.

Comparing the R2 & Adj R2 values for Base Model and Model 6th Iteration Model

Comparing R2 & Adj R2 Value for Base and 6th Iteration Model

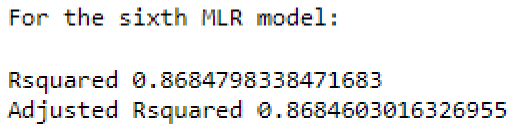
1. **Base Model :**



```
For the first MLR model:

Rsquared 0.8797543423603084
Adjusted Rsquared 0.8797141547960187
```

Figure 41 : R2 & Adj R2 Values for Base Model

2. **6th Iteration Model :**



```
For the sixth MLR model:

Rsquared 0.8684798338471683
Adjusted Rsquared 0.8684603016326955
```

Figure 42 : R2 & Adj R2 Values for 6th Iteration Model

Interpretations :

- This shows that, after removing all the insignificant variables and tuning the model, the 6th Iteration Model showed a very good R2 & Adj R2 value.

Before concluded the best model, we checked the prediction for model 6.

For that, we plotted the destiny distribution plot. It was as follows :

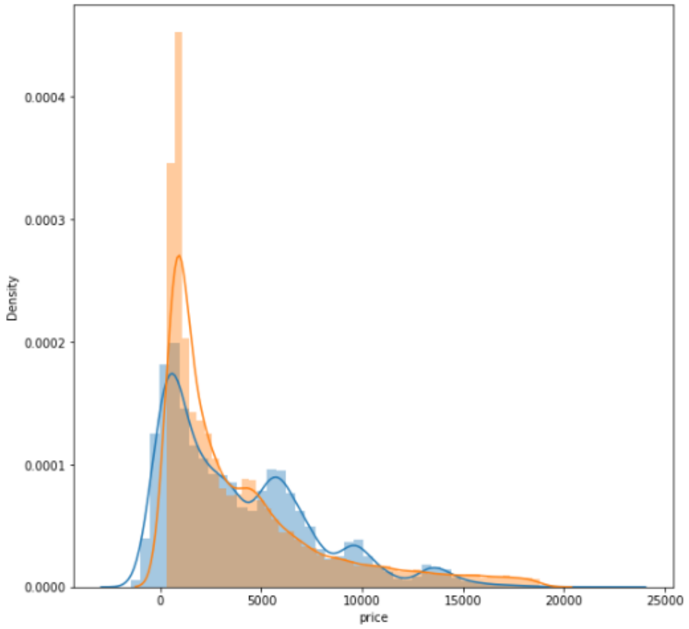


Figure 43 : Distribution Plot (6th Iteration Model)

Interpretations :

- Blue is the fitted values(predicted) and orange is the actual values.
- We observed that the blue curve almost overlaps the orange curve which meant that many values were correctly being predicted . So, this model was considered a very good model.

Linear Relationship Check Between Variables

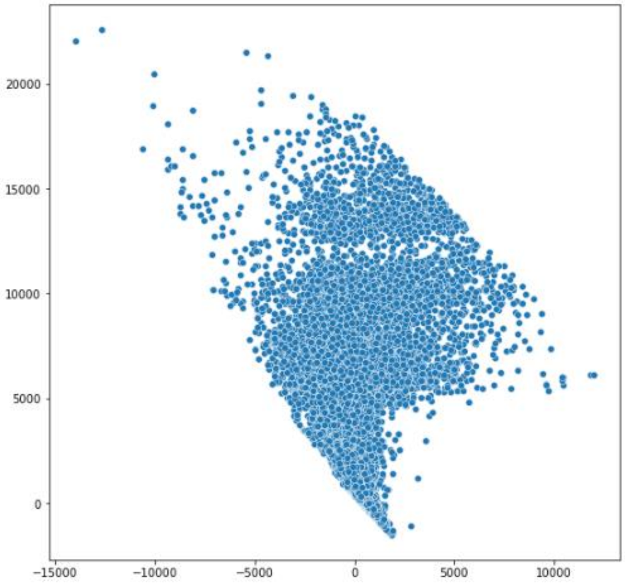


Figure 44 : Linear Relationship between Independent & Dependent Variables

This represents the linear negative relationship between Independent & Dependent Variables of the data residuals.

Afterwards, we plotted the distribution plot and boxplot for the residual data. It was as follows :

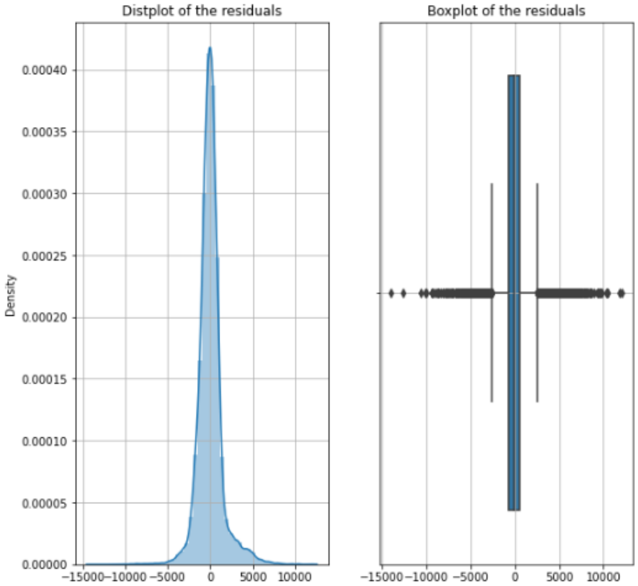


Figure 45 : Distplot & Boxplot Of Data Residuals

Accuracy Assessment For Model

A scatter plot was plotted showing the linear relationship between predicted and actual values for the 6th Iteration Model.

The plot was shown below as follows :

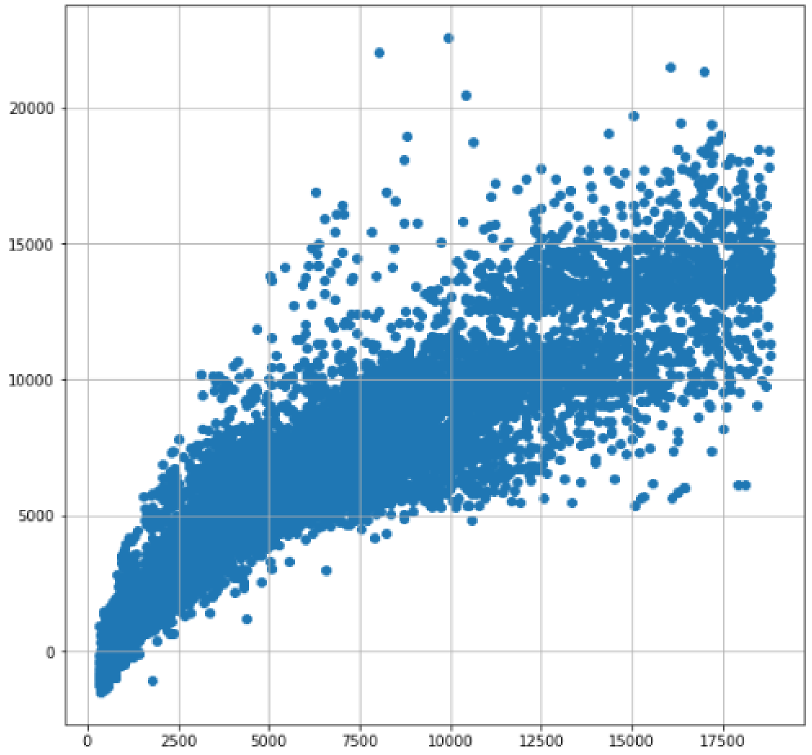


Figure 46 : Scatter Plot (Model 6) - Linear Relationship

This shows a linear relationship as the **predicted** and **actual** values were very close to each other. Hence the R2 is also high

Hence **Model 6** is the **Best Model** said to be a very good linear regression model as its **prediction value** is very close to the **actual value**.

Conclusion :

Price = (carat x 8123.1753) + (cut x 237.2058) + (color x 185.4473) + (clarity x (-120.8573))

- The above equation is the best equation for predicting the price of cubic zirconia.
- Hence, **6th Iteration Model** was the **Best Model** as per the statsmodels library.

Problem 1.3 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.¶

After building a Linear Regression Model using Statsmodels Library, we builded another Linear Regression Model using the Sklearn Library.

Sklearn library works on predictors while statsmodels work on statistical approach.

Linear Regression Model Using Sklearn Library

If we only wanted to predict using Linear Regression and was not looking for the model building aspect of it, we can do that as well.

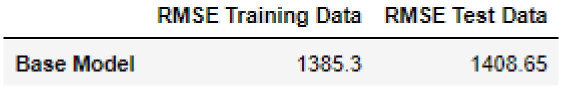
First we will split the data into train and test. We will build the model on the training data and check the RMSE on the test data.

Note: We are going to build all the models and use those to predict first and then go on to evaluate those models.

- Firstly, we copied all the predictor variables into X dataframe. Since 'price' is dependent variable we dropped it. Next, we copied the target variable i.e. ‘price’ into Y dataframe.
- Then we split the data using train_test_split function.
- Afterwards, we invoked the Linear Regression function and find the best fit model on training data.

Base Model building using Sklearn Linear Regression

- We fit the train variables into the base model and run it.
- After fitting the model, we used predict function in order to predict for train & test set data.
- In Sklearn library, for linear regression we check the Root Mean Square Error (RMSE) Value to determine the model fitness.
- The output of the RMSE for base model was as follows :



	RMSE Training Data	RMSE Test Data
Base Model	1385.3	1408.65

Figure 47 : RMSE Value For Base Model

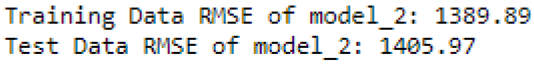
Interpretations :

- As we can see that RMSE(Root Mean Square Error) for train and test data were not similar (close to each other) and there was some difference between them.
- Hence, we shall remove some predictor variables and then build another model.

2nd Iteration Model

As 'z' had very high correlation with price as shown in the scatter plot(pairplot), thus 'z' variable must be removed from the model.

- We removed the ’z’ variable from the data and rebuild the model.
- After fitting the model, we again used the predict function in order to predict for train & test data set.
- Also, we determined the RMSE value for this 2nd Iteration Model. It was as follows :



```
Training Data RMSE of model_2: 1389.89
Test Data RMSE of model_2: 1405.97
```

Figure 48 : RMSE Value For 2nd Iteration Model

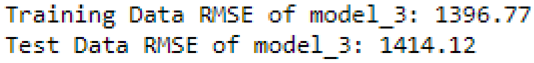
Interpretations :

- Some improvement had been done since 'z' variable had been removed but still the model was not fit significantly.
- Thus, we need to improve the model by removing 'y' variable as it had a very high strong correlation with 'price' and thus this was leading to multicollinearity which should be fixed.

3rd Iteration Model

As 'y' had very high correlation with price as shown in the scatter plot(pairplot), thus 'z' variable must be removed from the model.

- We removed the ’y’ variable from the data and rebuild the model.
- After fitting the model, we again used the predict function in order to predict for train & test data set.
- Also, we determined the RMSE value for this 3rd Iteration Model. It was as follows :



```
Training Data RMSE of model_3: 1396.77
Test Data RMSE of model_3: 1414.12
```

Figure 49 : RMSE Value For 3rd Iteration Model

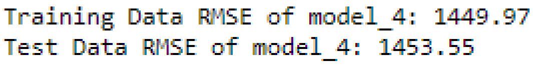
Interpretations :

- Instead of improvement, the difference between test and train data had increased which should be fixed.
- For that, other high correlation variables with price must be removed .

4th Iteration Model

As 'x' had a high correlation value with price as shown in the scatter plot, thus it should be removed because it is impacting the model.

- Following the same above steps after removal of ‘x’ variable , a new 4th Iteration build and prediction function was run for it.
- The RMSE value for this model was as follows :



```
Training Data RMSE of model_4: 1449.97
Test Data RMSE of model_4: 1453.55
```

Figure 50 : RMSE Value for 4th Iteration Model

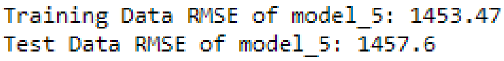
Interpretations :

- Improvement has been there now but still the RMSE values between test and train were different to some levels and that must be fixed by appropriate measures.
- For that, other variables must be removed .

5th Iteration Model

Removing the ‘depth’ column because of the scatter plot(pairplot) , it had almost no relation with price.

- Following the same above steps after removal of ‘depth’ variable , a new 5th Iteration build and prediction function was run for it.
- The RMSE value for this model was as follows :



```
Training Data RMSE of model_5: 1453.47
Test Data RMSE of model_5: 1457.6
```

Figure 51 : RMSE Value for 5th Iteration Model

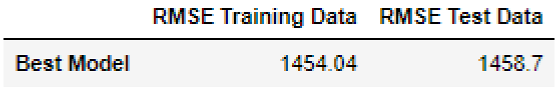
Interpretations :

- Improvement has been there now but still the RMSE values between test ans train were different to some levels and that must be fixed by appropriate measures.
- For that, other variables with price must be removed.

6th Iteration Model

In this model, we removed the ‘table’ attribute.

- After removal of ‘table’ attribute, all the same code functions were run and a better tuned model was build.
- The RMSE Value for 6th Iteration Model was as follows :



	RMSE Training Data	RMSE Test Data
Best Model	1454.04	1458.7

Figure 52 : RMSE Value For 6th Iteration Value

Interpretations :

- We noticed that, the RMSE(Root Mean Square Error) for both train and test were very close to each other and similar.
- Hence, we can state that 'model 6' was the best model that was build.
- Thus we can say that the best model had a very good prediction and can be considered as a very good fit model.

Comparing the final model(best model) from both libraries (statsmodels & Sklearn)

We shall compare it by comparing the scatter plots for both the best models and the comparison can be made out of it.

Scatter Plot For Best Model (statsmodel library)

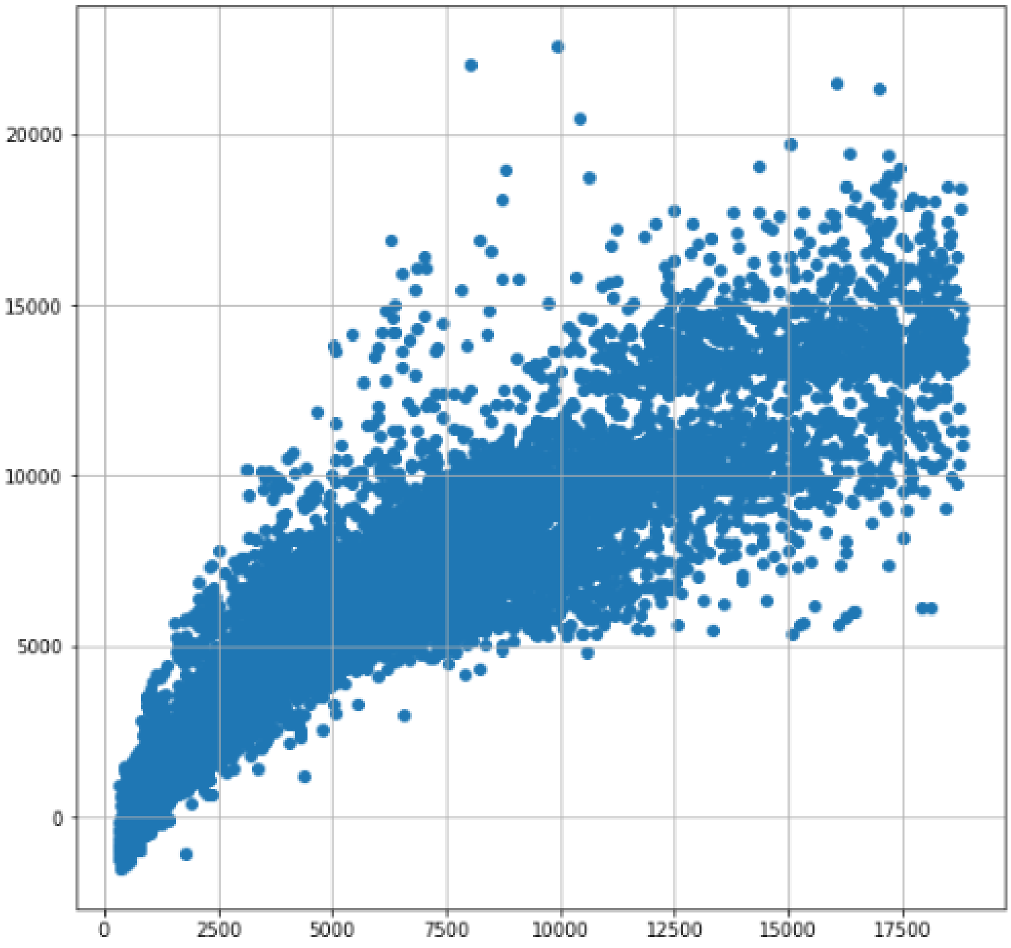


Figure 53 : Scatter Plot For Best Model (Statsmodels Library)

Scatter Plot For Best Model (Sklearn Library)

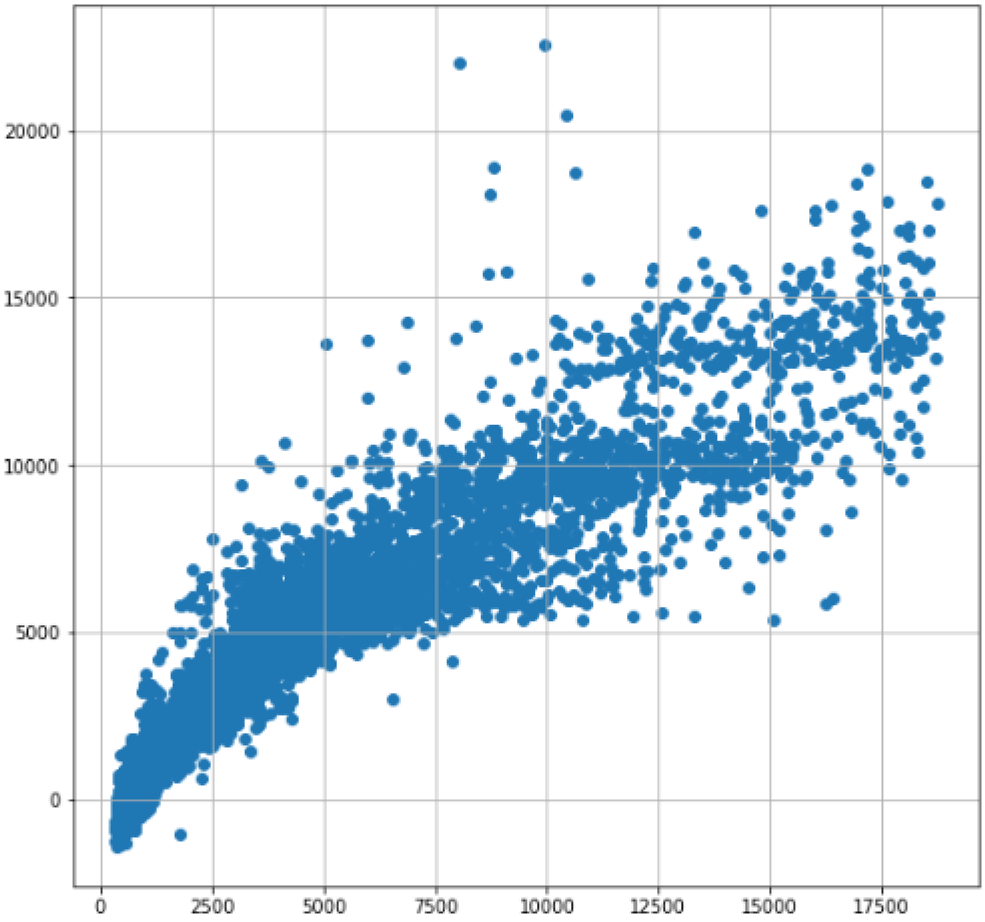


Figure 54 : Scatter Plot For Best Model (Sklearn Library)

Interpretations :

- From **Best Model From Statsmodel Library**, the accuracy of the predicted values to actual values were quite good as most of the points lie very close to each other. While looking at the **Best Model From Sklearn Library**, we can say that the accuracy for the model was not that good with to **Best Model From Statsmodel Library**.
- Some values were found missing in the **Best Model From Sklearn Library** at the top right side of the scatter plot, while as we look at the scatter plot of **Best Model From Statsmodel Library** , many of the values were present and providing a high accuracy for the model.

Conclusion :

- Both of the model gave strong high positive linear relationship but still if looking the gap values and other parameters such as accuracy etc, the best model among both of them was considered **Best Model From Statsmodel Library** as this model provided a very strong accuracy prediction which is the objective for this linear regression question.

Highlighting the benefits of the model approach and business interpretations

- The model approach was very simple from the beginning, we just observed the question what has been asked and what kind of model we need to build in order to achieve the target.
- Since, the benefits of the model approach were to be highlighted, the predictor variables with high correlation was the main thing that made out model approach easier.
- As we removed those high correlation variables, the multicollinearity from the model disappeared/ got reduced which were the main headache for building a linear regression model.
- As for the business interpretations, all we need to note was that the quality of cut & weight of carat used in cubic zirconia should be main approach in order to predict the price of cubic zirconia. As a business strategist, we must think that whenever a customer comes to buy a cubic zirconia, he would check the quality , weight of the carat (how much weight should it be) . These were the key factors that must be taken into account for the pricing.
- Apart from 'carat weight' & 'cut quality' , incusions and blemishes were something to be taken care of while making a cubic zirconia. The less the blemishes , the greater the price of cubic zirconia.
- Color can be considered as a very good predictor of determining the price, as color represents the beauty, the attractiveness of the cubic zirconia. People often love to have a attractive thing around themselves and thus they buy such kind of products. So, color was also a very good predictor variable for determining the price of cubic zirconia.

Price = (carat x 8123.1753) + (cut x 237.2058) + (color x 185.4473) + (clarity x (-120.8573))

The above equation is the correct and best equation for predicting the price of cubic zirconia.

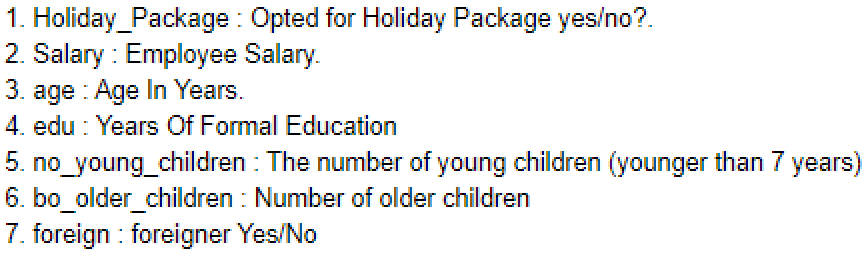
Problem 2 – Logistic Regression

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.

Problem 2.1 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.

At first we loaded the Data Dictionary for understanding of column name for data set “Holiday Package”.

The Data Dictionary was as follows :

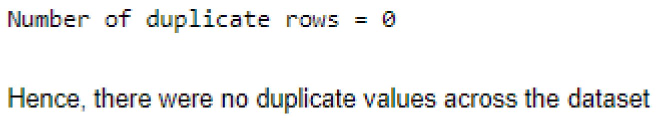


1. Holiday_Package : Opted for Holiday Package yes/no?.
2. Salary : Employee Salary.
3. age : Age In Years.
4. edu : Years Of Formal Education
5. no_young_children : The number of young children (younger than 7 years)
6. bo_older_children : Number of older children
7. foreign : foreigner Yes/No

Figure 55 : Data Dictionary (Problem 2)

- After loading the data set and reading the data dictionary and data, we got to know that 471 employees didn't opt for the holiday package while only 401 opted for the same.

- We also inferred that, out of 872 employees, only 216 employees were foreigners while rest of the employees were of the same nation.
- There were duplicate rows present in the data.



```
Number of duplicate rows = 0
```

Hence, there were no duplicate values across the dataset

Figure 56 : Duplicate Row Check (Problem 2)

EDA

Univariate Analysis

Outliers in logistic regression models did not impact the model if they were in small quantity. But it can impact the output of the model , if they were in large amount. So, removing them or not must be totally dependent on this factors.

Here were the Boxplot for **Holiday Package** data set :

1. **Salary**

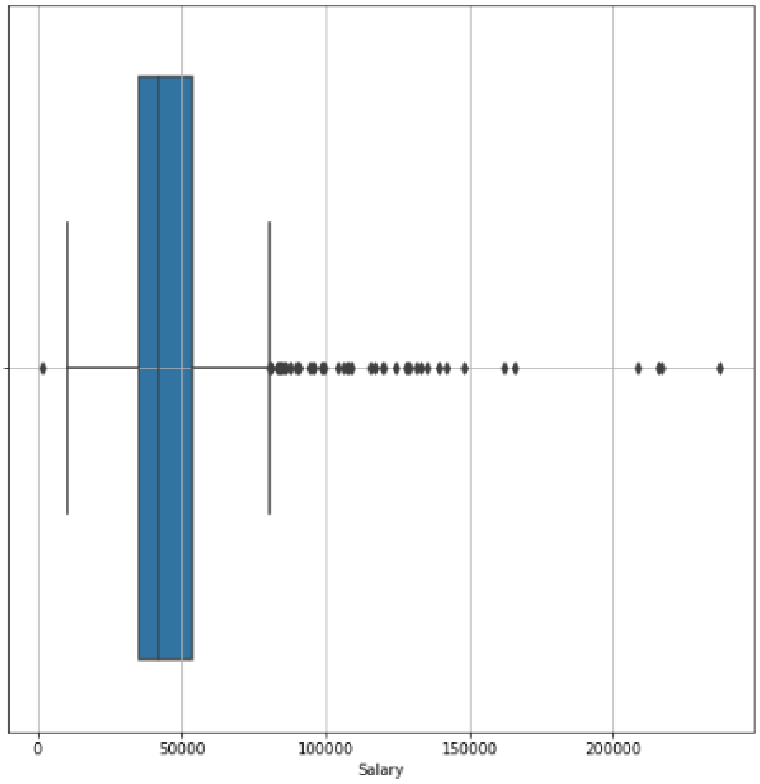


Figure 57 : Boxplot (Salary)

- Many extreme values were present in the ‘Salary’ variable but no of them were outliers as those values were logical so they were not required to be removed.
- The boxplot tends to be at the right side of the boxplot indicating that ‘Salary’ variable was positively right skewed.

2. Age

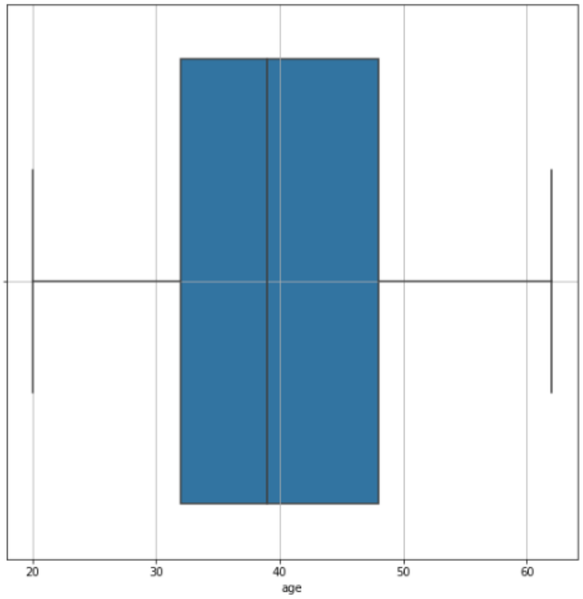


Figure 58 : Boxplot (age)

- The boxplot of ‘age’ does not show any extreme values and hence no outliers were present in it.

3. Educ

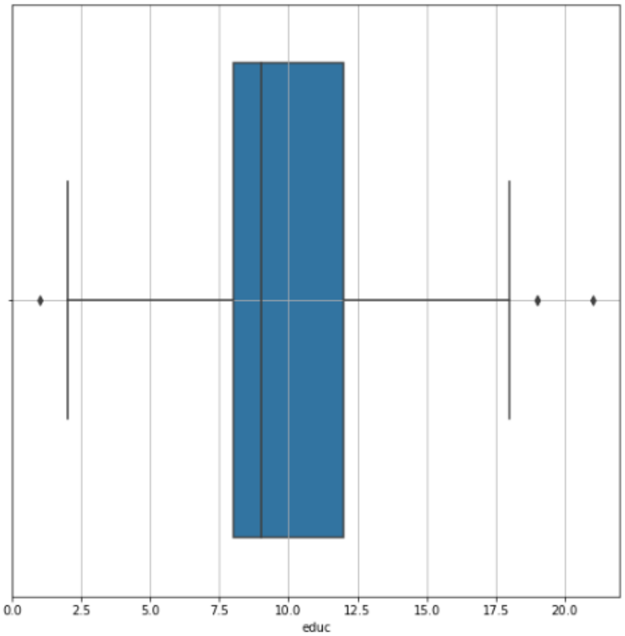


Figure 59 : Boxplot (educ)

- Some extreme values were present at both extreme left & right side of the boxplot but they were considered as outliers.
- The boxplot tends to be at the right side of the boxplot indicating that ‘educ’ variable was positively right skewed.

4. no_young_children

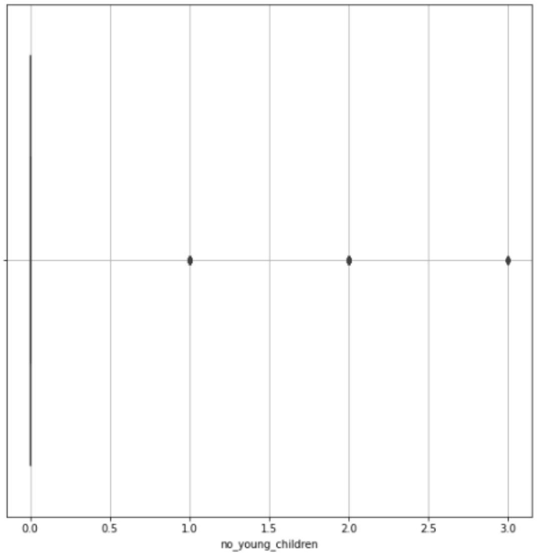


Figure 60 : Boxplot (no_young_children)

- The boxplot of ‘no_young_children’ showed many values at the 0 point but that doesn’t mean those were outliers. Those values represents the total number of young children of the employee working.
- This boxplot interpret different outcomes such as , the employee doesn’t have kids which is the reason the number of young children mentioned were zero. The other reason could be that the young children had already grown up and come into the category of older children which was the reason number of young children were zero. And many more reasons could be there.

5. no_older_children

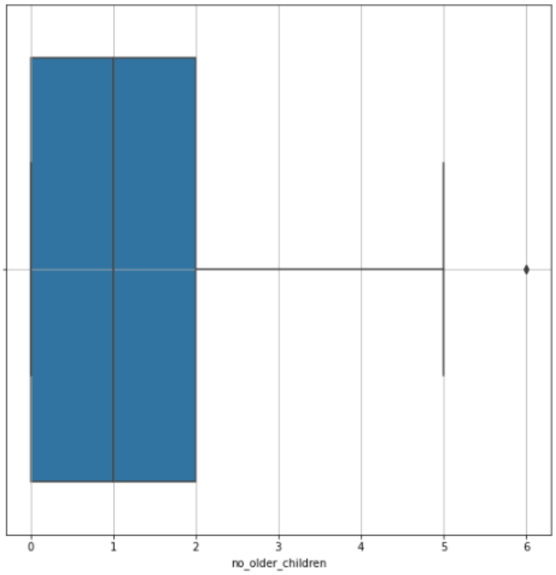


Figure 61 : Boxplot (no_older_children)

- The boxplot tends to be at the right side of the boxplot indicating that ‘no_older_children’ variable was positively right skewed.
- There were some extreme values for this boxplot but they were not outliers thus we do not need to remove them.

Note :

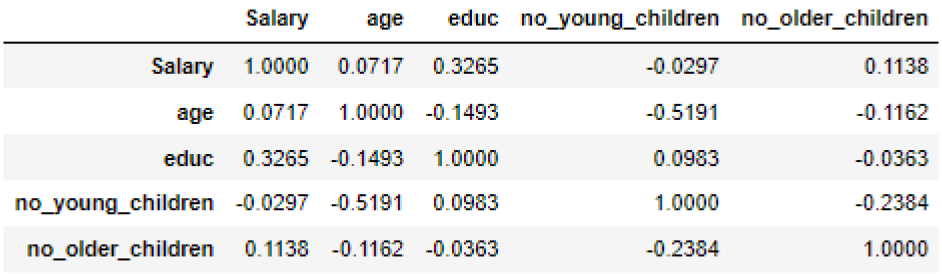
- As per the boxplot, **'no_young_children'** had maximum no. of zero values. The value count function also tells us the exact count of no young children in the data set.
- We should not remove this variable as it had a logical reason that some employees had children who were older or some employees had no children .

Should we treat outliers or not?

Logistic Regression models are not much impacted due to the presence of outliers because the sigmoid function tapers the outliers. But due to some parameters we can make the model predictions much reliable and effective.

Bi Variate Analysis

- **Correlation Matrix**



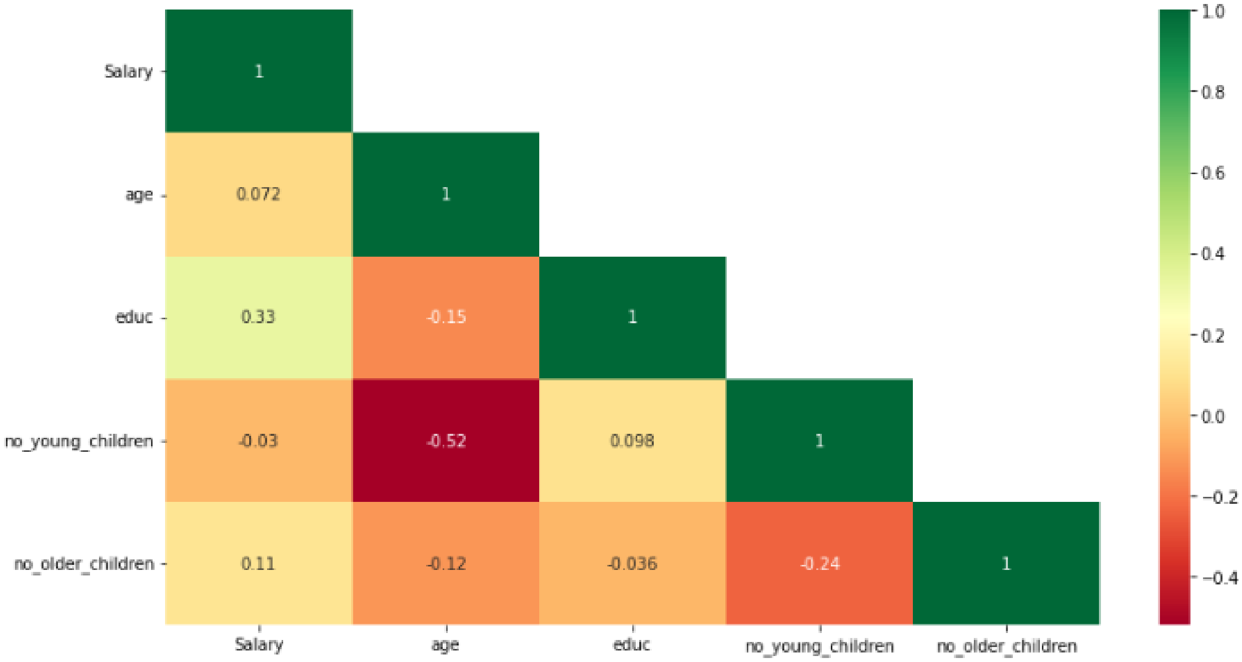
	Salary	age	educ	no_young_children	no_older_children
Salary	1.0000	0.0717	0.3265	-0.0297	0.1138
age	0.0717	1.0000	-0.1493	-0.5191	-0.1162
educ	0.3265	-0.1493	1.0000	0.0983	-0.0363
no_young_children	-0.0297	-0.5191	0.0983	1.0000	-0.2384
no_older_children	0.1138	-0.1162	-0.0363	-0.2384	1.0000

Figure 62 :Correlation Matrix (Problem 2)

Interpretations :

- A Correlation Matrix was created above using the 'Pearson' method.
- For better understanding, a graphical representation in the form of heatmap was also created with respect to Correlation Matrix.

- **Heat Map**



Interpretations :

- We inferred from the correlation matrix that, 'age' & 'no_young_children' variable had lowest relation with a correlation value of -52 %. It means that, as the age of employee increases , the no of young children gets decreased. There can be reasons for that such as ,as the age of employees increases, the age of their respective children also gets increased logically.
- We also noticed that 'educ' & 'Salary' had the maximum correlation among all the variables. This relationship was a positive correlation with a correlation value of 33% . It means that years of formal education plays a role in getting a higher Salary among all the other variables.
- We observed that 'no_young_children' & 'no_older_children' also had a negative correlation with a correlation value of -24 %. This meant that, they both were indirectly proportional to each other which is technically logical. As the age of young children increases, they would become older and hence the no. of young children would decrease and the no. of older children increases as per the universal fact.
- Moreover, 'age' & 'Salary' didn't had a very good correlation. They had very low relation between each other meaning they did not had a impact on each other considerably.
- Similarly, 'no_young_children' & 'Salary' and 'no_older_children' & 'Salary' also didn't had a very good correlation with each other. It means that no of young or older children didn't affect/impact the Salary of the employee.

Multi-Variate Analysis

We created a pairplot for analysis of the dataset and finding insights from it.

- **Pairplot**

  The pairplot for this dataset was as follows :

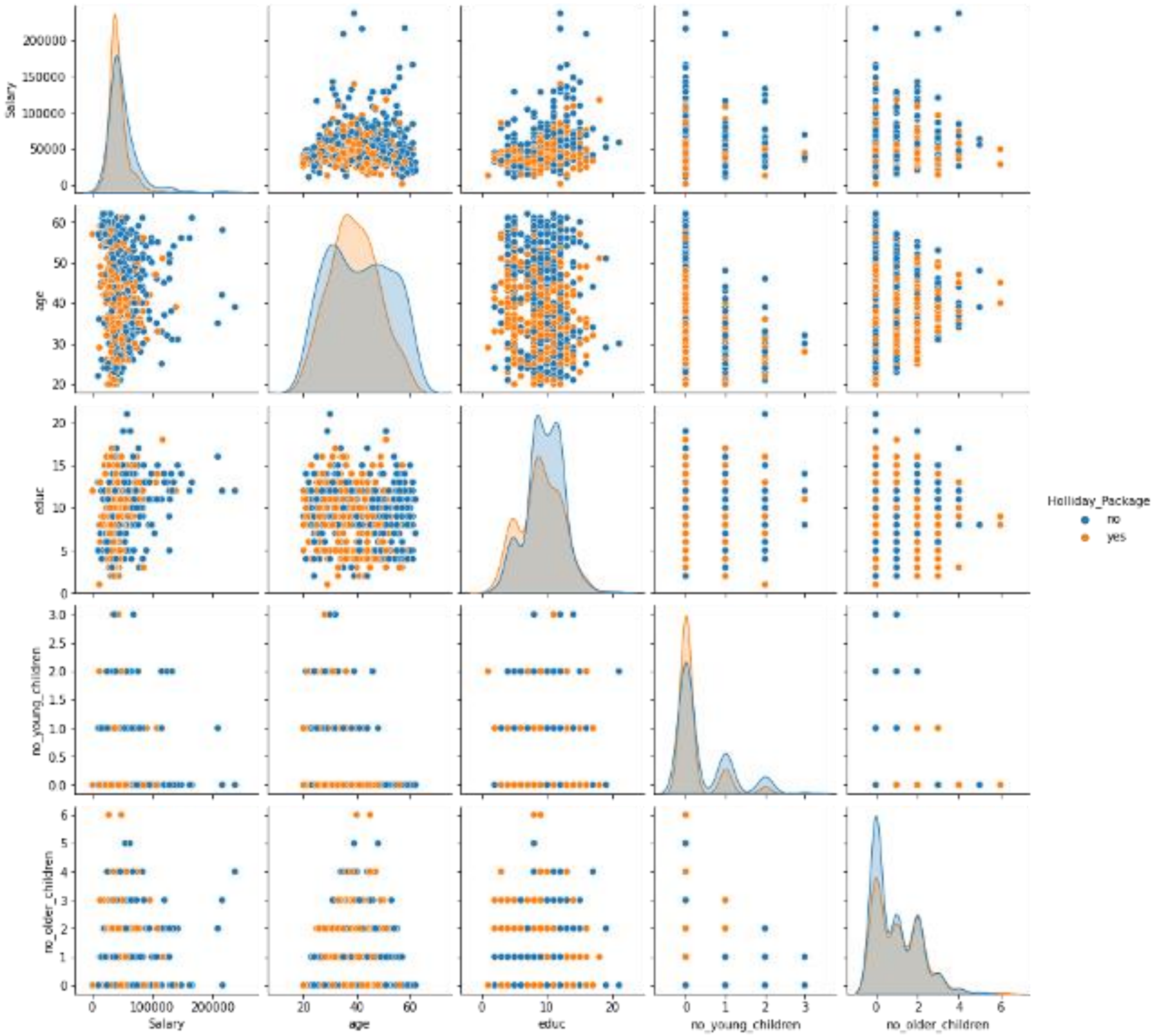


Figure 63 : Pairplot (Problem 2)

Interpretations :

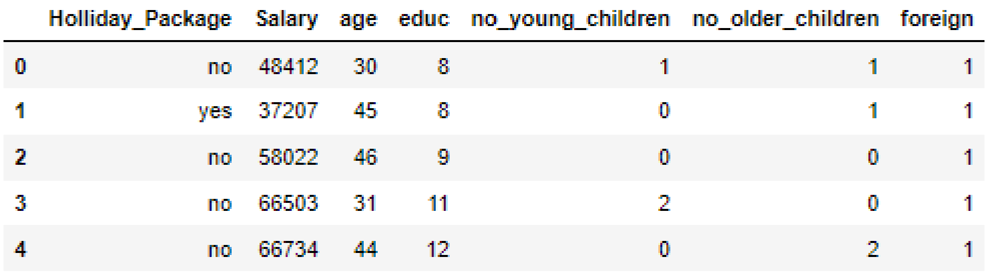
- As we know logistic regression is a classification model, so we looked for the insights of the diagonal from the pairplot.
- We inferred that **'Salary'** variable had distribution plot with **'holiday package'** as hue in the diagonal was overlapping each other which means that this variable cannot distinguish whether the employee had opted/ not opted for the holiday package. Furthermore, the **'Salary'** distribution plot had highest skewness and it was positively right skewed.
- Similarly, in the distribution plot with **'holiday package'** as hue for **'no_young_children'** and **'no_older_children'** , the plots overlap each other indicating that

these variables were also not able to predict whether an employee opted the holiday package or not. Such attributes were not considered as good attributes for classification model. Hence , they can be considered as poor predictors. These two variables were also positively right skewed while looking at the distribution plot.

- Moreover, as we looked at the **'age'** attribute, the distribution plot were not completely overlapping but some values were surely overlapping. Also, the distribution plot with hue 'yes' had higher value than the distribution plot with hue 'no'. These attributes were going to be weak predictors.
- Looking at the scatter plots, we inferred that, all the values are mostly overlapping each other with some exceptions.
- Only **'educ'** variable turns out to be negatively left skewed as we checked the mean and median value for it.

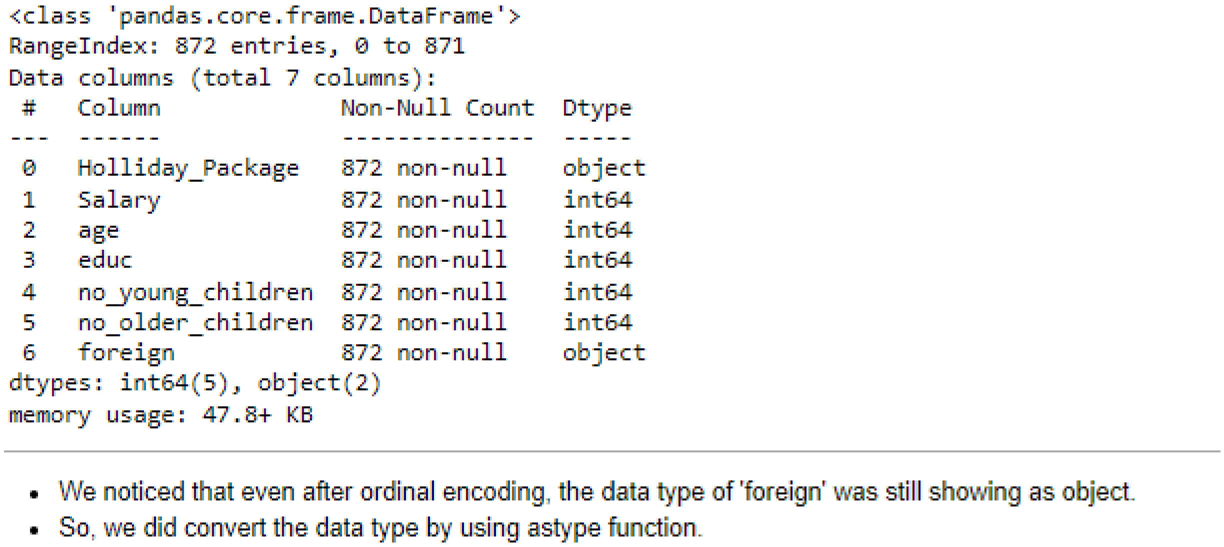
Transforming the object data types

- We can treat the object variables appropriately by either creating dummy variables (One-Hot Encoding) or coding it up in an ordinal manner.
- We choose ordinal manner in this case to treat 'foreign' attribute.



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

Figure 64 : Data Set After Ordinal Coding

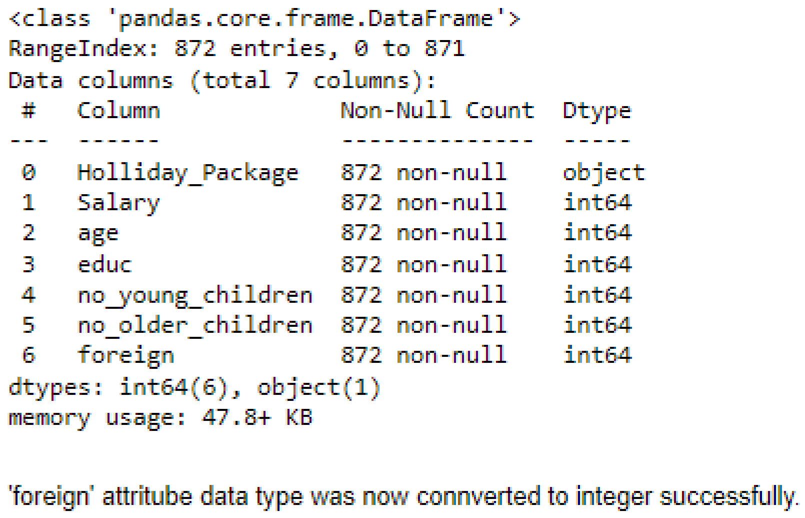


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2   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
```

- We noticed that even after ordinal encoding, the data type of 'foreign' was still showing as object.
- So, we did convert the data type by using astype function.

Figure 65 : Insights

We converted the data type by astype function

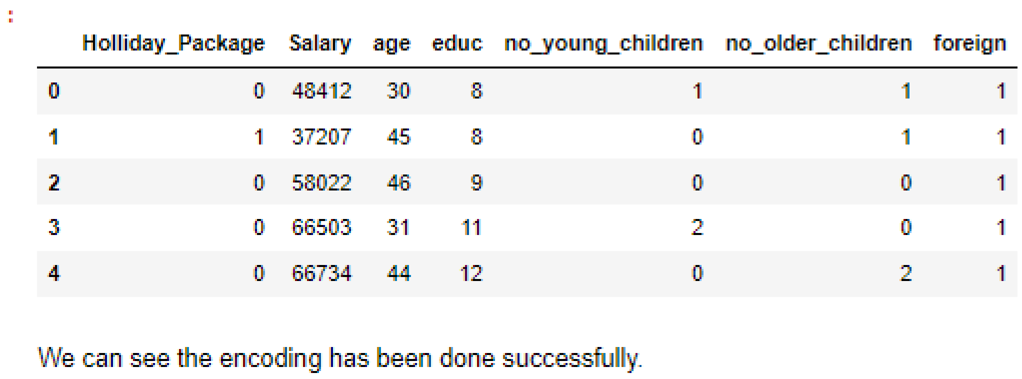


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2  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

'foreign' attritube data type was now connverted to integer successfully.

Figure 66 : Converting data type

Then by label encoding, I converted the target variable i.e. ‘Holliday_Package ‘.



:

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

We can see the encoding has been done successfully.

Figure 67 : Label Encoding Successful

Now, the data has been pre-processed and was now ready to build the logistic regression model.

Problem 2.2 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.

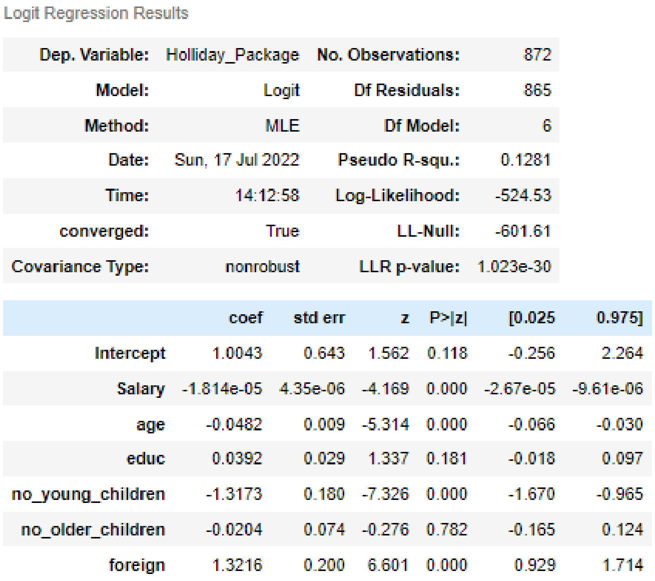
Logistic Regression Model

In the case of linear regression, the target variable ‘y’ is a continuous variable but let us assume that the ‘y’ is a categorical variable which has two classes then linear regression should not be used to predict the value of target variable. The output of the Linear Regression is not bound within [0,1] as it can take any real value from $(-\infty ,\infty)$. Logistic regression is used to solve such problem which gives us the corresponding probability outputs and then we can decide the appropriate cut-off points to get the target class outputs.

- We are now building the Logistic Regression Model using all the variables on the full data and check the summary statistics of the model. Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF).

Base Model

We had builded the base model as follows using the necessary libraries and function.



Logit Regression Results

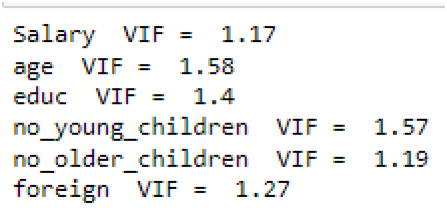
Dep. Variable:	Holliday_Package	No. Observations:	872
Model:	Logit	Df Residuals:	865
Method:	MLE	Df Model:	6
Date:	Sun, 17 Jul 2022	Pseudo R-squ.:	0.1281
Time:	14:12:58	Log-Likelihood:	-524.53
converged:	True	LL-Null:	-601.61
Covariance Type:	nonrobust	LLR p-value:	1.023e-30

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.0043	0.643	1.562	0.118	-0.256	2.264
Salary	-1.814e-05	4.35e-06	-4.169	0.000	-2.67e-05	-9.61e-06
age	-0.0482	0.009	-5.314	0.000	-0.066	-0.030
educ	0.0392	0.029	1.337	0.181	-0.018	0.097
no_young_children	-1.3173	0.180	-7.326	0.000	-1.670	-0.965
no_older_children	-0.0204	0.074	-0.276	0.782	-0.165	0.124
foreign	1.3216	0.200	6.601	0.000	0.929	1.714

Figure 68 : Base Model (Logistic Regression)

Check for multicollinearity in the predictor variables using Variance Inflation Factor (VIF)

We made a user defined function for checking the VIF value .



```
Salary  VIF =  1.17
age  VIF =  1.58
educ  VIF =  1.4
no_young_children  VIF =  1.57
no_older_children  VIF =  1.19
foreign  VIF =  1.27
```

Figure 69 : VIF Value (Problem 2)

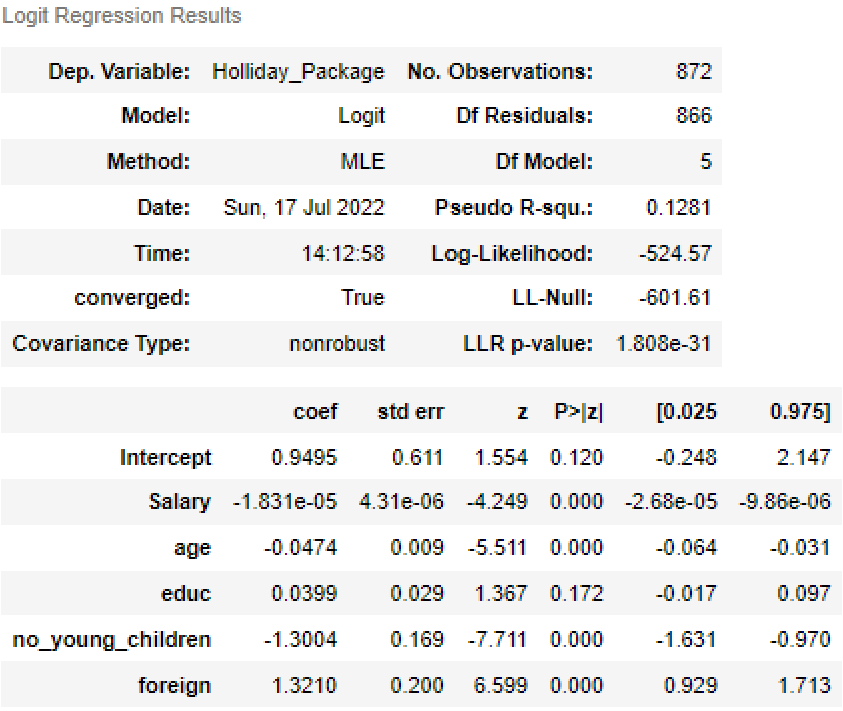
Interpretations :

- From inferential statistics, we got to know that p value for 'no_older_children', 'educ' was greater than 0.05 which indicated that these attributes were not significant for the logistic regression model. But only 'no_older_children' had p value = 0.782 which was very high so, this attribute should be removed and a new model shall be build.
- As per the VIF values, 'age' had the maximum VIF value with a value of 1.58 followed by 'no_young_children' with a VIF value of 1.57

2nd Iteration Model

We removed ‘no_older_children’ as said above and builded another model and checked its fitness.

The 2nd Iteration Model summary was as follows :



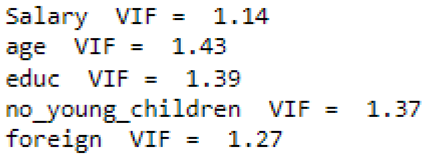
Logit Regression Results

Dep. Variable:	Holliday_Package	No. Observations:	872
Model:	Logit	Df Residuals:	866
Method:	MLE	Df Model:	5
Date:	Sun, 17 Jul 2022	Pseudo R-squ.:	0.1281
Time:	14:12:58	Log-Likelihood:	-524.57
converged:	True	LL-Null:	-601.61
Covariance Type:	nonrobust	LLR p-value:	1.808e-31

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.9495	0.611	1.554	0.120	-0.248	2.147
Salary	-1.831e-05	4.31e-06	-4.249	0.000	-2.68e-05	-9.86e-06
age	-0.0474	0.009	-5.511	0.000	-0.064	-0.031
educ	0.0399	0.029	1.367	0.172	-0.017	0.097
no_young_children	-1.3004	0.169	-7.711	0.000	-1.631	-0.970
foreign	1.3210	0.200	6.599	0.000	0.929	1.713

Figure 70 : 2nd Iteration Model (Problem 2)

Afterwards , we checked the VIF value for the 2nd Iteration Model as follows :



```
Salary  VIF =  1.14
age  VIF =  1.43
educ  VIF =  1.39
no_young_children  VIF =  1.37
foreign  VIF =  1.27
```

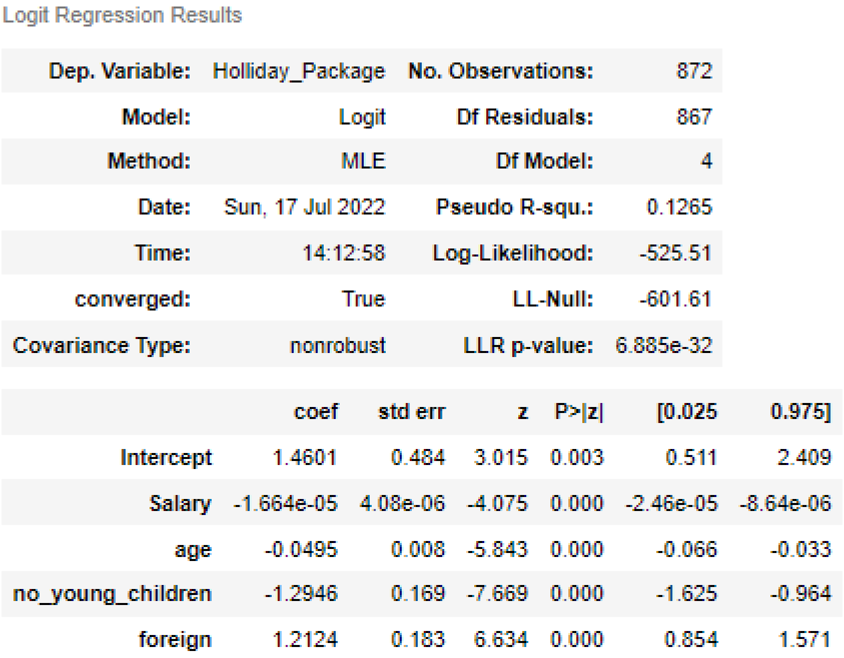
Figure 71 : VIF Value for 2nd Iteration Model

Interpretations :

- As we dropped the 'no_older_children' variable, the p value for 'educ' has reduced but it was still greater than 0.05. Therefore, we should 'educ' attribute and another model and check its performance.

3^{RD} Iteration Model

We removed ‘educ’ attribute and rebuild another model to check its fitness.



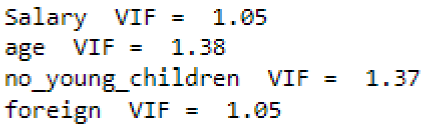
Logit Regression Results

Dep. Variable:	Holliday_Package	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	867
Method:	MLE	**Df Model:**	4
Date:	Sun, 17 Jul 2022	**Pseudo R-squ.:**	0.1265
Time:	14:12:58	**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	1.4601	0.484	3.015	0.003	0.511	2.409
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
no_young_children	-1.2946	0.169	-7.669	0.000	-1.625	-0.964
foreign	1.2124	0.183	6.634	0.000	0.854	1.571

Figure 72 : 3rd Iteration Model

VIF Value for 3^{RD} Iteration Model was as follows :



```
Salary  VIF =  1.05
age  VIF =  1.38
no_young_children  VIF =  1.37
foreign  VIF =  1.05
```

Figure 73 : VIF Value (3RD Iteration Model)

Interpretations :

- Now, every variable were having p value less than 0.05 and the VIF value for every variable were good enough and in the range also.

Conclusion :

- As per the insights and making appropriate changes into the model, the **3rd Iteration model** was considered to be the **Best Model** from all the models.

- The p value, VIF value and factors of all the attributes were appropriate and this model looks pretty good model.

As we believe that, 3rd **Iteration Model** was the **Best Model** so, we checked the prediction for the same and confirmed that.

We used the fitted values function and then plotted the distribution plot for the **Best Model.**

The distribution plot was as follows :

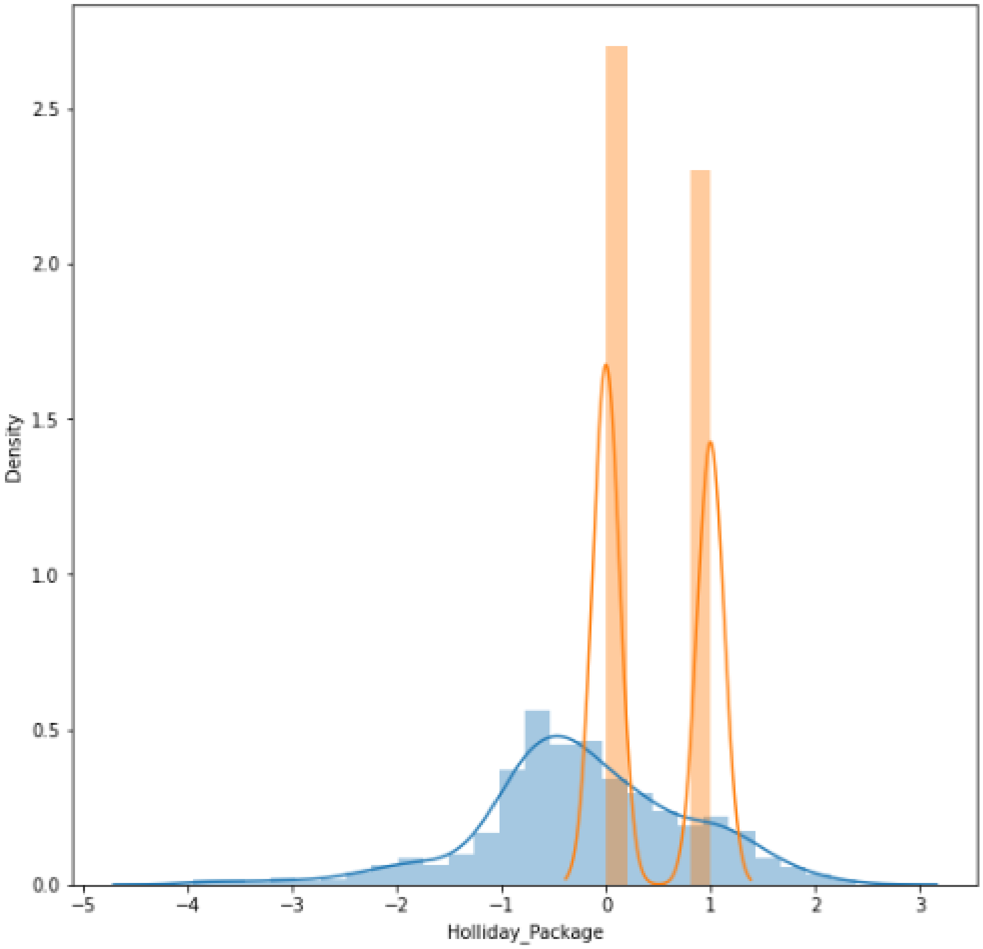


Figure 74: Distplot For Best Model

- Blue is the fitted values (predicted) and orange is the actual values.

Conclusions From 3th Iteration that is Best Model

Holliday_Package = (Salary x -1.664e-05) + (age x -0.0495) + (no_young_children x -1.2946) + (foreign x 1.2124)

- The above equation represents the best predictors used in finding out the target variable which is 'Holliday_Package' in order to determine whether an employee opted for the Holiday Package or not.
- Hence, **Model 3** was the best model.

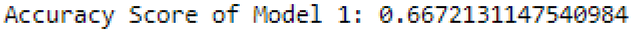
Logistic Regression Model using Sklearn Library

We builded the Logistic Regression Model using Sklearn Library .

- At first, we loaded the Logistic Regression with parameter solver = ‘newton-cg’ and other parameters for a better model.
- Then , we spilt the data into Train & Test and fit that model.

Base Model Logistic Regression

- We builded the base model and checked the accuracy of the train model.
- After that we trained the model checking the prediction for the test data.
- The results were as follows :

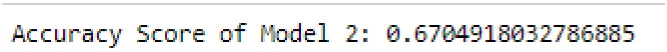


```
Accuracy Score of Model 1: 0.6672131147540984
```

Figure 75 : Accuracy Score Of Base Model

2nd Iteration Model

- We removed ‘no_older_children’ for the same reason as we did in the Logistic Regression Model from statsmodels library.
- We used the same function for building the model while doing some tuning.
- The results were as follows :

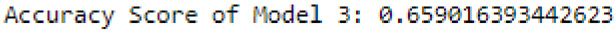


```
Accuracy Score of Model 2: 0.6704918032786885
```

Figure 76 : Accuracy Score Of Model 2

3rd Iteration Model

- We removed ‘educ’ for the same reason as we did in the Logistic Regression Model from statsmodels library.
- Then, we fit the model and checked the accuracy .
- Afterwards, we used the predict function for predicting the test data.
- The results were as follows :



```
Accuracy Score of Model 3: 0.659016393442623
```

Figure 77 : Accuracy Score of 3rd Iteration Model

Evaluating the three models on the test data using the various statistics of the confusion matrix.

Confusion Matrix summary statistics Evaluation on the Test Data

We loaded the necessary libraries and plotted the heatmap for the confusion matrix for the test set for all the models .

The output was as follows :

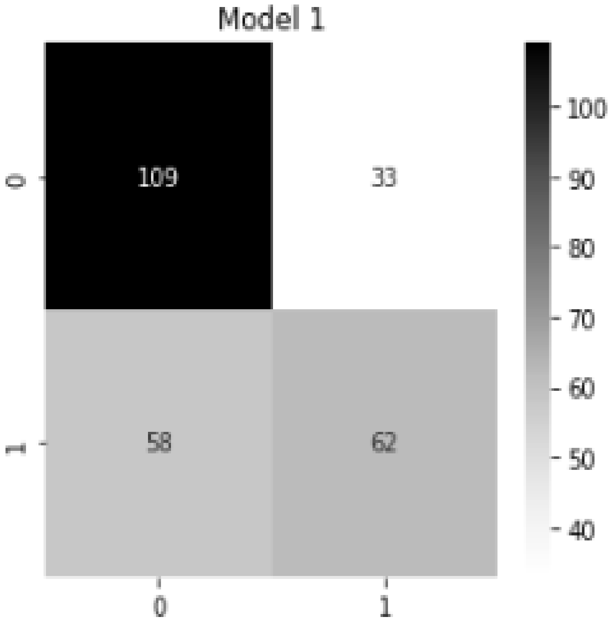


Figure 78 : Confusion Matrix (Base Model)

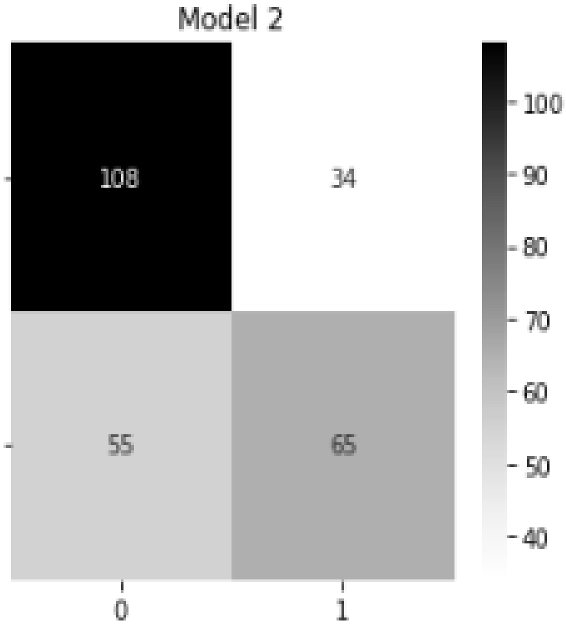


Figure 79 : Confusion Matrix (2nd Iteration Model)

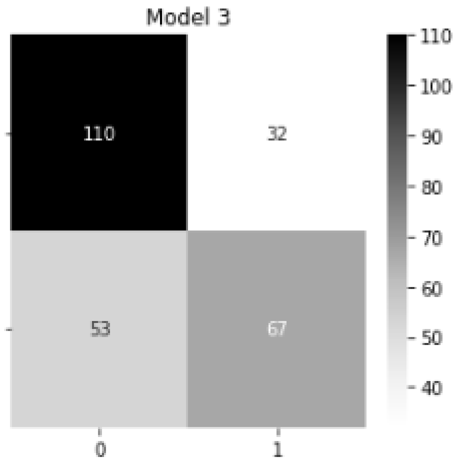
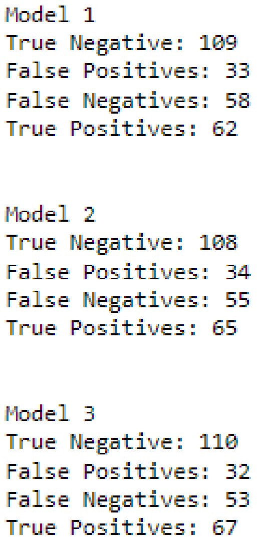


Figure 80 : Confusion Matrix (3rd Iteration Model)

By comparing all the confusion matrix of all the models, we drawed some insights from them . They were as follows :

- There were 4 quadrants in the each model i.e. true negative, false negative, false positive & true positive.
- From all these 4 quadrants, the values were as follows :



```
Model 1
True Negative: 109
False Positives: 33
False Negatives: 58
True Positives: 62


Model 2
True Negative: 108
False Positives: 34
False Negatives: 55
True Positives: 65


Model 3
True Negative: 110
False Positives: 32
False Negatives: 53
True Positives: 67
```

Figure 81 : Model Evaluation

Interpretations :

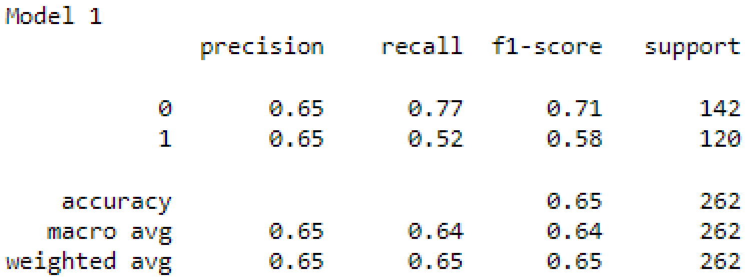
- The **Base Model** predicts that True Positives value = 62 which means that 62 employees were correctly predicted that they will opt the Holiday Package. While True Negative value = 109 which means that 109 employees were correctly predicted that they would not opt for the Holiday Package.
- The base model predicts that True Positives value = 62 which means that 62 employees were correctly predicted that they will opt the Holiday Package.

Classification Report Of All the Models

Note : - In all models, **'0'** represented **'No'** to Holiday Package Opted & **'1'** represented **'Yes'** to Holiday Package.

In logistic regression, **accuracy** is not considered as a good criteria for determining a good model. Instead , **f1-score** is used to determine the fitness(how good the model is built) of the logistic regression model.

1. Model 1



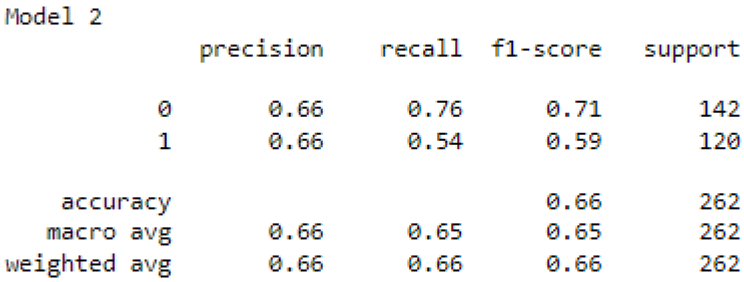
```
Model 1
              precision    recall  f1-score   support

           0       0.65      0.77      0.71       142
           1       0.65      0.52      0.58       120

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

Figure 82 : Classification Report - Model 1

Interpretations :

- From **Model 1** , for **'0'** ,the **precision** and **recall** were 0.65 & 0.77 and there **f1-score** was 0.71 which indicates that the model is fair . While on the other hand, for **'1'** in **Model 1**, the **precision** and **recall** were 0.65 & 0.52 and their respective **f1-score** was 0.58 which indicated that the prediction for employees opting for Holiday Package was not very good. Therefore, the model can be improved by removal of poor predicators.

2. Model 2



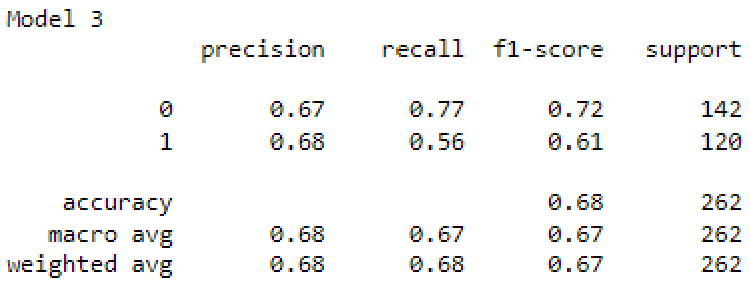
```
Model 2
              precision    recall  f1-score   support

           0       0.66      0.76      0.71       142
           1       0.66      0.54      0.59       120

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

Figure 83 : Classification Report - Model 2

Interpretations :

- From **Model 2**, for **'0'** ,the **precision** and **recall** were 0.66 & 0.76 and there **f1-score** was 0.71 which indicates that the model is fair . While on the other hand, for **'1'** in **Model 2**, the **precision** and **recall** were 0.66 & 0.54 and their respective **f1-score** was 0.59 which indicated that the prediction for employees opting for Holiday Package was not that good. Therefore, the model can be improved by removal of poor predicators.

Model 3



```
Model 3
              precision    recall  f1-score   support

           0       0.67      0.77      0.72       142
           1       0.68      0.56      0.61       120

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

Figure 84 : Classification Report - Model 3

Interpretations :

- From **Model 3**, for **'0'** ,the **precision** and **recall** were 0.67 & 0.77 and there **f1-score** was 0.72 which indicates that the model is fair . While on the other hand, for **'1'** in **Model 3**, the **precision** and **recall** were 0.68 & 0.56 and their respective **f1-score** was 0.61 which indicated that the prediction for employees opting for Holiday Package was good.

Conclusion :

- **Model 3** was now free of multicollinearity by removing the poor predictors.
- Therefore, the **Model 3** built can be considered as the **best model**.

Check the summary statistics of the AUC-ROC curve for all the three Logistic Regression Models built.

- **AUC-ROC curve** is a performance measurement for the classification problems at various threshold settings.
- **ROC** is a probability curve and **AUC** represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.
- Higher the AUC, the better the model is at predicting 0 classes as 0 & 1 classes at 1. This was done only for the test data.
- We used the probability function for all the models and plotted the ROC curve for all the curve.

- The results were as follows :

```
Model 1 AUC: 0.71684
```

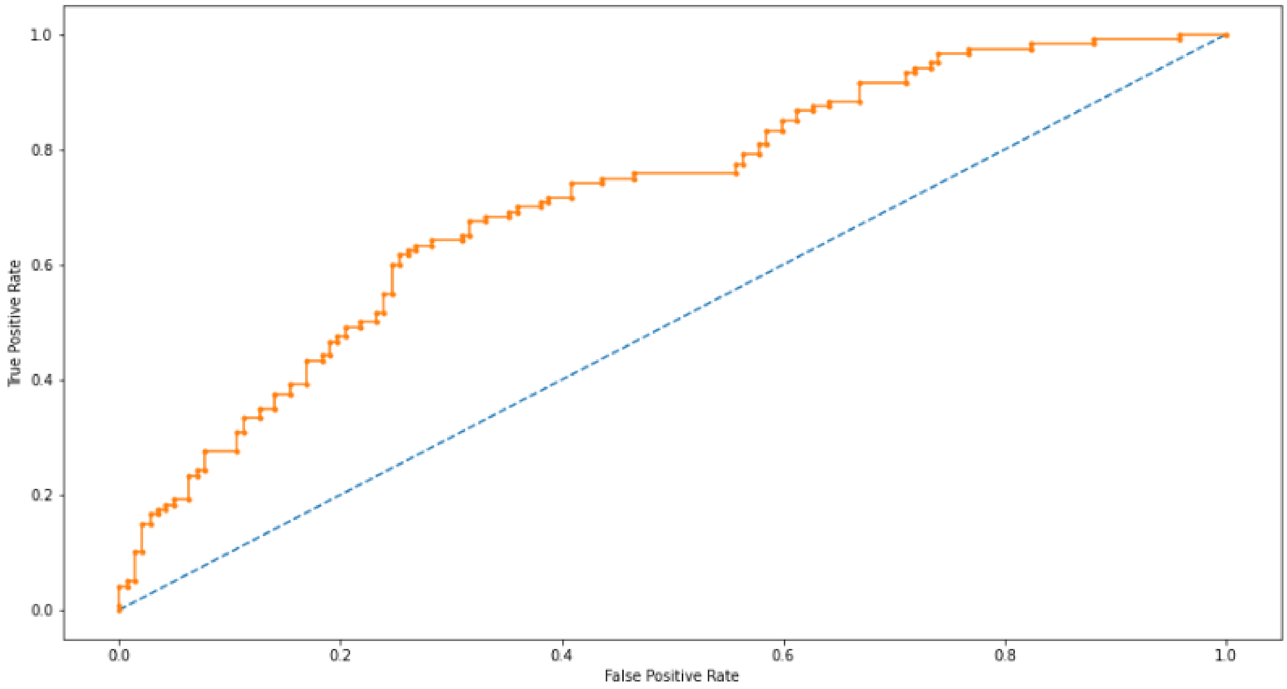


Figure 85 : ROC Curve - Model 1

```
Model 2 AUC: 0.71719
```

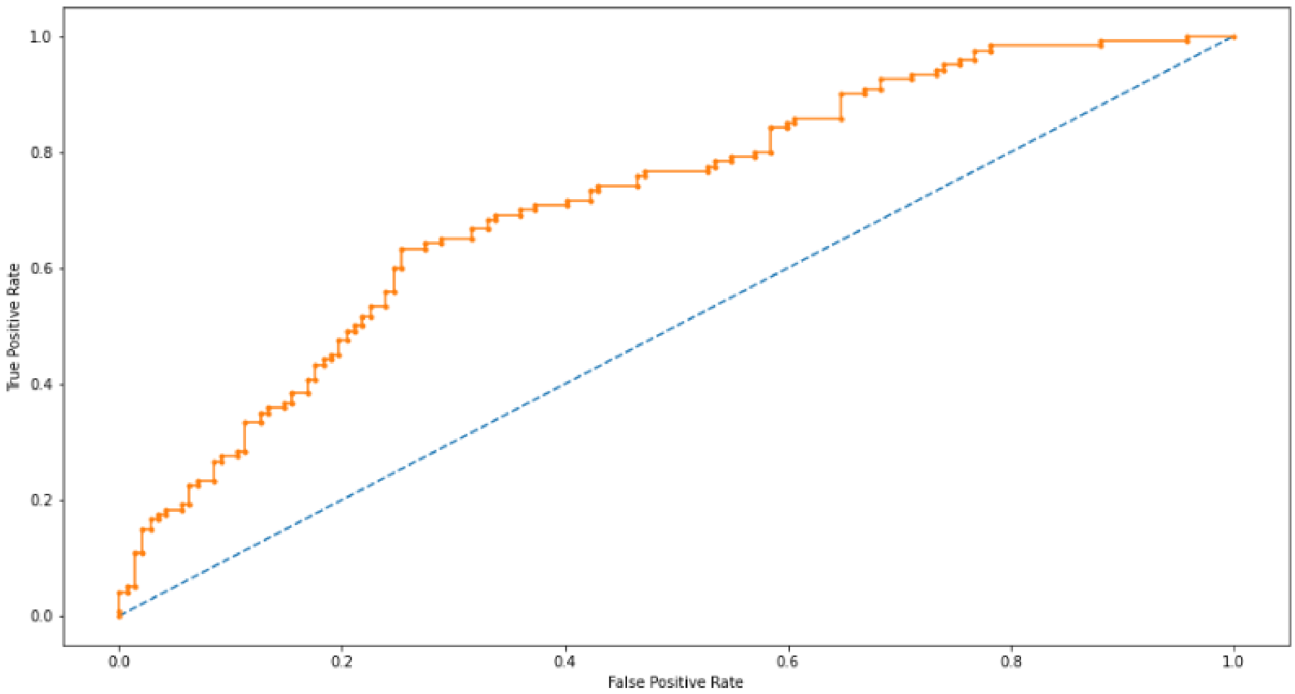
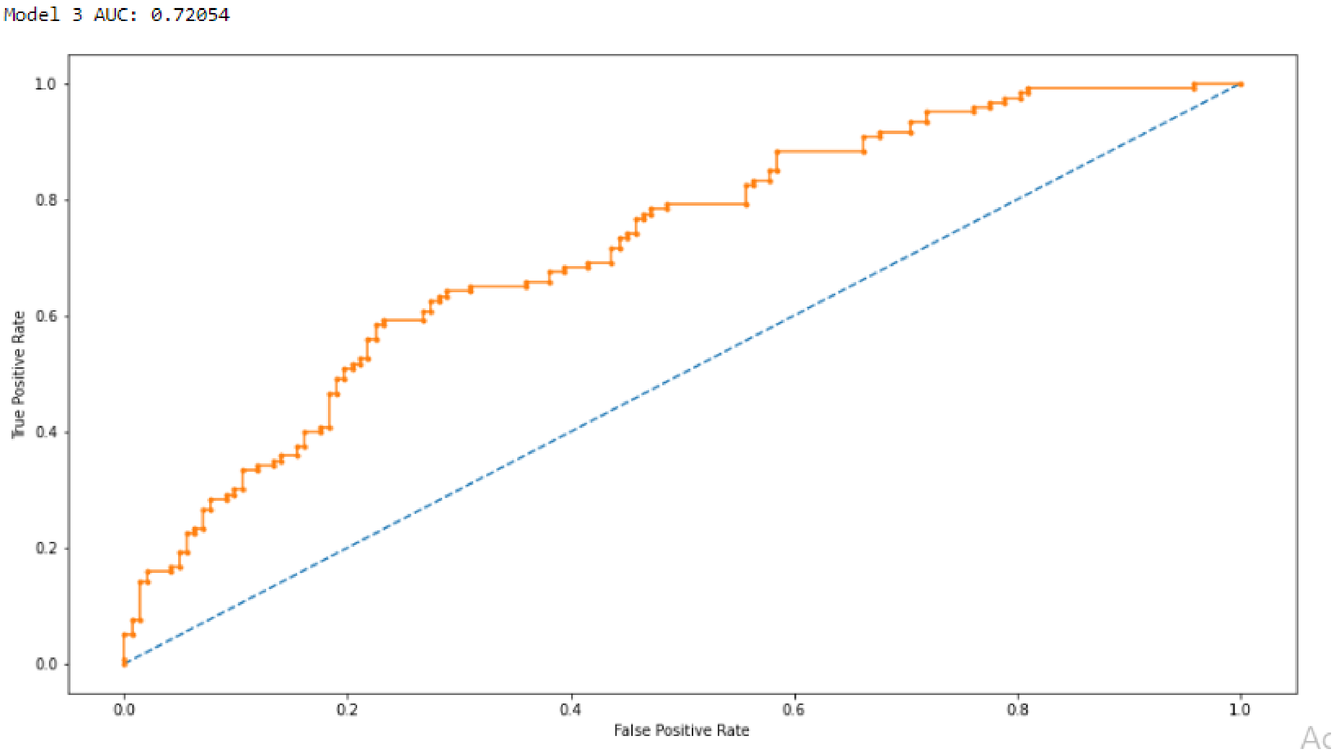


Figure 86 : ROC Curve - Model 2



```
Model 3 AUC: 0.72054
```

Figure 87 : ROC Curve - Model 3

Conclusion :

- After comparing all the **ROC curve** , **Model 3** has the best ROC curve.
- As per the **Model 1** & **Model 2**, the **AUC score(Area Under Curve)** were found out to be 0.71684 & 0.71719 which were comparatively low to **Model 3** . This indicated that the **Model 3** was a very good model.
- As per the **Model 3** , the **AUC score(Area Under Curve)** was found out to be 0.72054 which was actually the highest among all the other models. It's true positive rate was greater than the other models. This indicated that the **Model 3** was a very good model which means it has a good measure of separability.

Compare the final best model of Part (II) and the proposed one in Part (III)

We shall use the scatter plot to compare both the best models build using statsmodel and Sklearn library.

Scatter Plot For Best Model (statsmodel library)

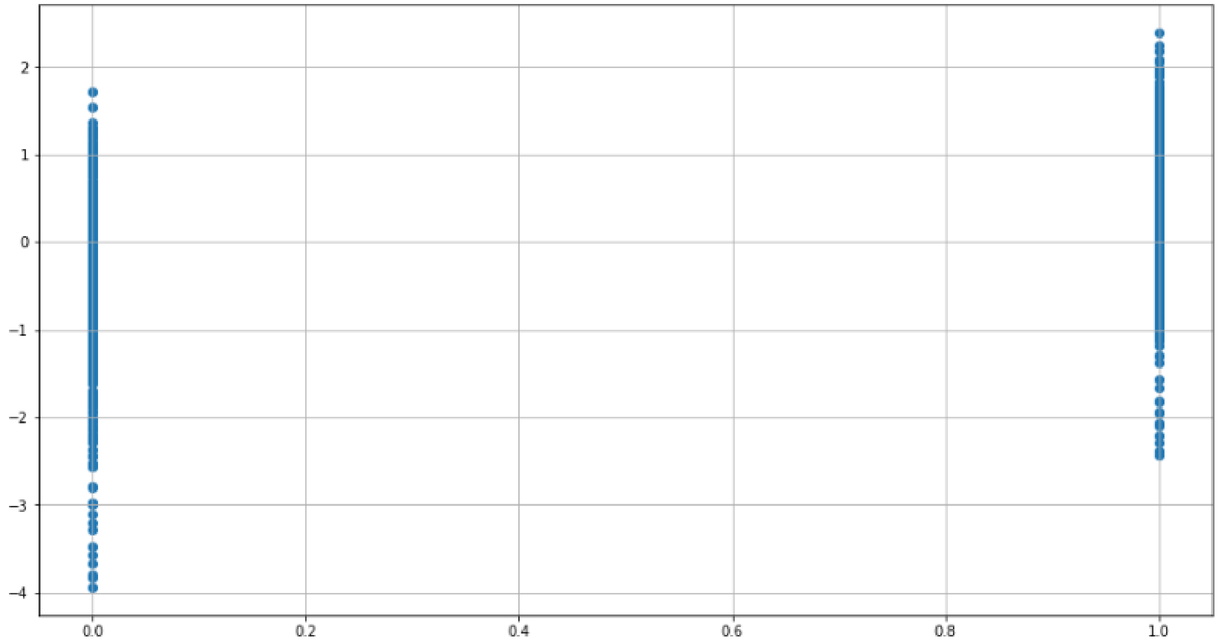


Figure 88 : Scatter Plot For Best Model (Statsmodels Library)

Scatter Plot For Best Model (Sklearn library)

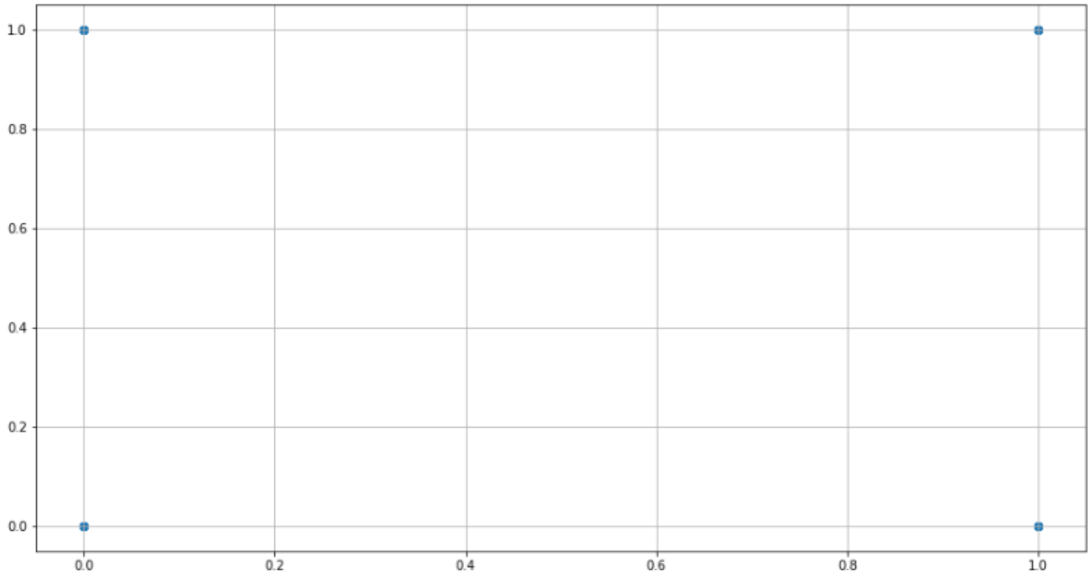


Figure 89 : Scatter Plot For Best Model (Sklearn Library)

Conclusion :

- By comparing the scatter plot for both test set of the best model (statsmodel & Sklearn) , we clearly observed that the **Best Model** from **statsmodel library** had a good plot as compared to that of the **Best Model** from **Sklearn Library**.
- **Best Model** from **Statsmodels Library** was a positive inclined plot while the plot of **Best Model** from **Sklearn Library** remained the same, as the value of x increases , the y value remained the same. This means that the value of y depends on x as per the scatter plot of **Best Model** from **Statsmodels Library** which indicated a very good model and fitness of that model was very good.
- Hence, by looking at both the models, we concluded that **Best Model** from **Statsmodel Library** was the **Best Model** out of both the models. Therefore, it should be used for model prediction by the tour and travel agency.