GREAT LEARNING

POST GRADUATE PROGRAM IN DATA SCIENCE & BUSINESS ANALYTICS

BUSINESS REPORT

CASE STUDIES ON:

- Linear Regression – Gem Stones Co Ltd
- Logistic Regression - Tour and Travel Agency

Submitted By:

STEFFIN JOHN

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CASE STUDY

1. **Linear Regression – Gem Stones Co Ltd**
2. **Logistic Regression - Tour And Travel Agency**

PROBLEM 1 - LINEAR REGRESSION

Overview:

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).

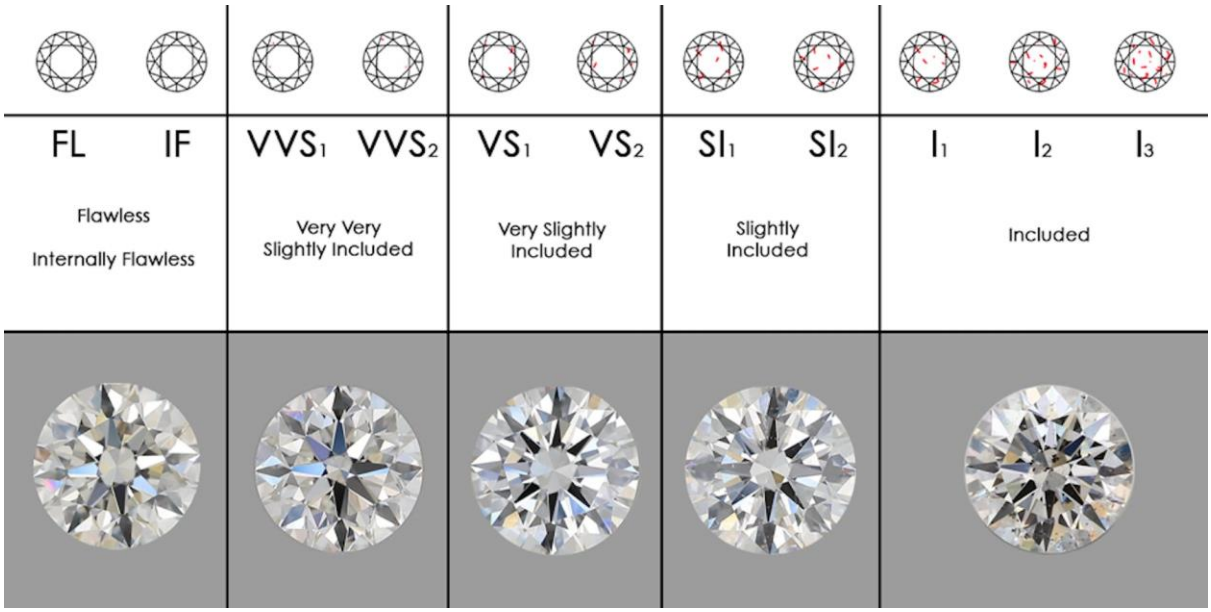
Your objective is to accurately predict prices of the zircon pieces. Since the company profits at a different rate at different price levels, for revenue management, it is important that prices are predicted as accurately as possible. At the same time, it is important to understand which of the predictors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

The data dictionary is given below.

Data Dictionary:

Variable Name	Description
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Colour	Colour of the cubic zirconia.With D being the best and J the worst.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
Depth	The Height of cubic zirconia, measured from the Culet to the

	table, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.



Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

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.**

DATASET HEAD AND TAIL

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

Table 1: Dataset Head

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
26962	26963	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	26964	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	26965	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656
26965	26966	0.27	Very Good	F	VVS2	61.8	56.0	4.19	4.20	2.60	682
26966	26967	1.25	Premium	J	SI1	62.0	58.0	6.90	6.88	4.27	5166

Table 2: Dataset Tail

Observations:

1. Dataset has 11 columns.
2. The first column (Unnamed column :0) is of no use for analysis and can be removed.

DATASET DESCRIBE

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	NaN	NaN	NaN	13484.0	7784.846691	1.0	6742.5	13484.0	20225.5	26967.0
carat	26967.0	NaN	NaN	NaN	0.798375	0.477745	0.2	0.4	0.7	1.05	4.5
cut	26967	5	Ideal	10816	NaN	NaN	NaN	NaN	NaN	NaN	NaN
color	26967	7	G	5661	NaN	NaN	NaN	NaN	NaN	NaN	NaN
clarity	26967	8	SI1	6571	NaN	NaN	NaN	NaN	NaN	NaN	NaN
depth	26270.0	NaN	NaN	NaN	61.745147	1.41286	50.8	61.0	61.8	62.5	73.6
table	26967.0	NaN	NaN	NaN	57.45608	2.232068	49.0	56.0	57.0	59.0	79.0
x	26967.0	NaN	NaN	NaN	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	NaN	NaN	NaN	5.733569	1.166058	0.0	4.71	5.71	6.54	58.9
z	26967.0	NaN	NaN	NaN	3.538057	0.720624	0.0	2.9	3.52	4.04	31.8
price	26967.0	NaN	NaN	NaN	3939.518115	4024.864666	326.0	945.0	2375.0	5360.0	18818.0

Table 3: Dataset Describe

Observation:

1. The column 'Price' is target variable and others are predictor variables
2. From the data sample, the column 'Unnamed:0' is not relevant for analysis
3. All the columns except depth contains 26,967 rows of data but depth column has 26,270 rows of data
4. x, y, z columns having values '0'
5. In the given data set there are 2 Integer type features,6 Float type features. 3 Object type features. Where 'price' is the target variable and all other are predictor variable.

DATASET INFORMATION

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

Table 4: Dataset Information

Observation:

1. Depth has missing values
2. Total 3 columns i.e. "cut"," colour", "clarity" are of object type.

DATASET SHAPE

```
(26967, 11)
```

Observation:

1. Total no. of Rows = 26967
2. Total no. of Columns = 11

EXPLORATORY DATA ANALYSIS

STEP 1: CHECK AND REMOVE ANY DUPLICATES IN THE DATASET

a) Checking for values which are equal to zero

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

Observation:

1. Successfully dropped the rows having values zero as don't have any meaning in model building.
2. On the given data set the mean and median values does not have much difference. We can observe Min value of "x", "y", & "z" are zero this indicates that they are faulty values. As we know dimensionless or 2-dimensional diamonds are not possible. So, we need to filter out those as it clearly faulty data entries. There are three object data type 'cut', 'colour' and 'clarity'.

b) Checking for duplicate records in the data.

```
Number of duplicate rows = 33
```

Observation:

1. Duplicate records are been treated thus no duplicates in the dataset.

STEP 2: CHECK AND TREAT ANY MISSING VALUES IN THE DATASET

a) Treatment of Missing Values

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

(Before)

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

(After)

Observation:

1. Total 297 values we missing in the Depth row

b) Shape of the Dataset after dropping Unnamed: 0, Zero's and Duplicates

```
Before (26958, 10)
After (26925, 10)
```

Observation:

1. Total no. of after dropping Rows = 26968
2. Total no. of after dropping Columns = 10

c) Dropping of Unnamed Column:

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

Observation:

1. Successfully dropped Unnamed Column : 0

d) Summary of Dataset before after treatments

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.000000	13484.000000	7784.846691	1.000000	6742.500000	13484.000000	20225.500000	26967.000000
carat	26967.000000	0.798375	0.477745	0.200000	0.400000	0.700000	1.050000	4.500000
depth	26270.000000	61.745147	1.412860	50.800000	61.000000	61.800000	62.500000	73.600000
table	26967.000000	57.456080	2.232068	49.000000	56.000000	57.000000	59.000000	79.000000
x	26967.000000	5.729854	1.128516	0.000000	4.710000	5.690000	6.550000	10.230000
y	26967.000000	5.733569	1.166058	0.000000	4.710000	5.710000	6.540000	58.900000
z	26967.000000	3.538057	0.720624	0.000000	2.900000	3.520000	4.040000	31.800000
price	26967.000000	3939.518115	4024.864666	326.000000	945.000000	2375.000000	5360.000000	18818.000000

(Summary Before Treatments)

	count	mean	std	min	25%	50%	75%	max
carat	26925.000000	0.797821	0.477085	0.200000	0.400000	0.700000	1.050000	4.500000
depth	26925.000000	61.746982	1.393457	50.800000	61.100000	61.800000	62.500000	73.600000
table	26925.000000	57.455305	2.231327	49.000000	56.000000	57.000000	59.000000	79.000000
x	26925.000000	5.729385	1.126081	3.730000	4.710000	5.690000	6.550000	10.230000
y	26925.000000	5.733152	1.163820	3.710000	4.710000	5.700000	6.540000	58.900000
z	26925.000000	3.538820	0.717483	1.070000	2.900000	3.520000	4.040000	31.800000
price	26925.000000	3936.249991	4020.983187	326.000000	945.000000	2373.000000	5353.000000	18818.000000

(Summary After Treatments)

Observation:

1. We could easily clearly see so many imbalances in the dataset (Before Treatment).
2. After the proper treatments we can see proper balance in the dataset (After Treatment).

STEP 3: OUTLIER TREATMENT

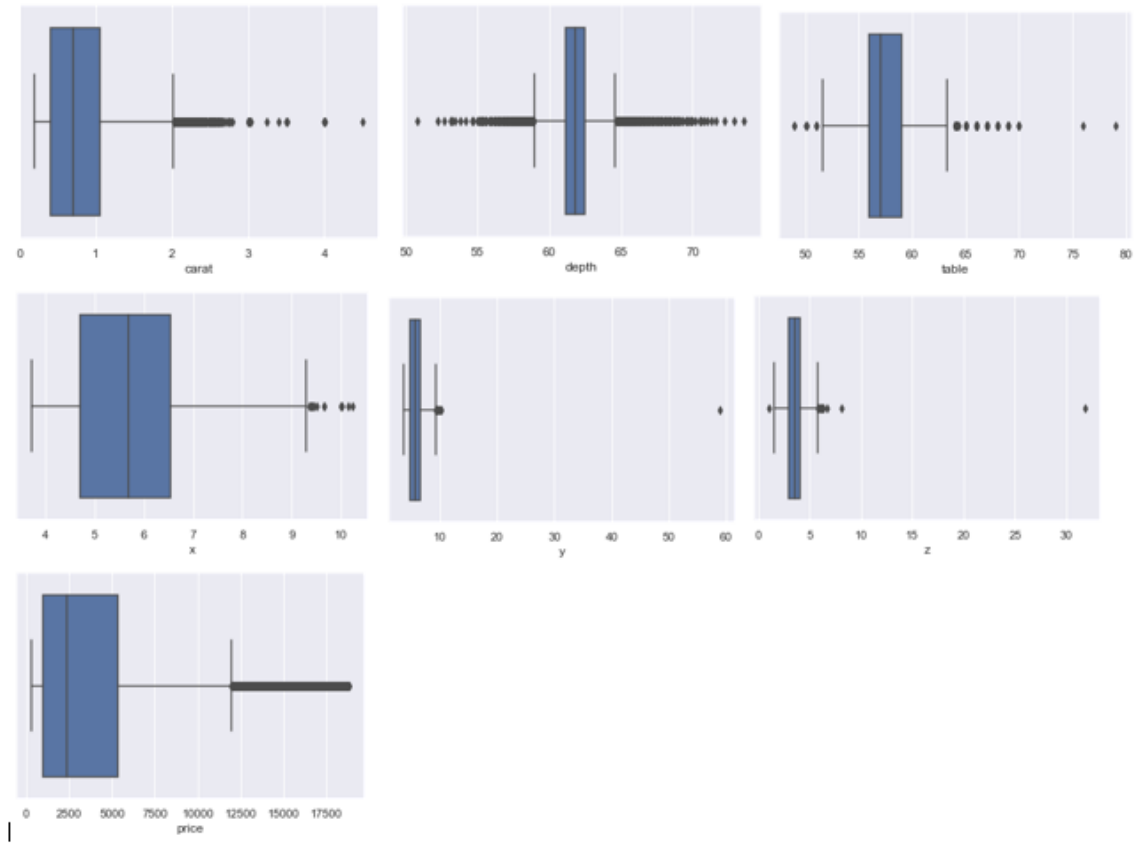


Figure 1: Outliers (Before Treatment)

Observation:

- **carat:** Lower Range: -0.5750000000000001; Upper Range: 2.0250000000000004
- **depth:** Lower Range: 59.0; Upper Range: 64.6
- **table:** Lower Range: 51.5; Upper Range: 63.5
- **x:** Lower Range: 1.9500000000000002; Upper Range: 9.309999999999999
- **y:** Lower Range: 1.9649999999999999; Upper Range: 9.285
- **z:** Lower Range: 1.1899999999999997; Upper Range: 5.75
- **price:** Lower Range: -5667.0; Upper Range: 11965.0

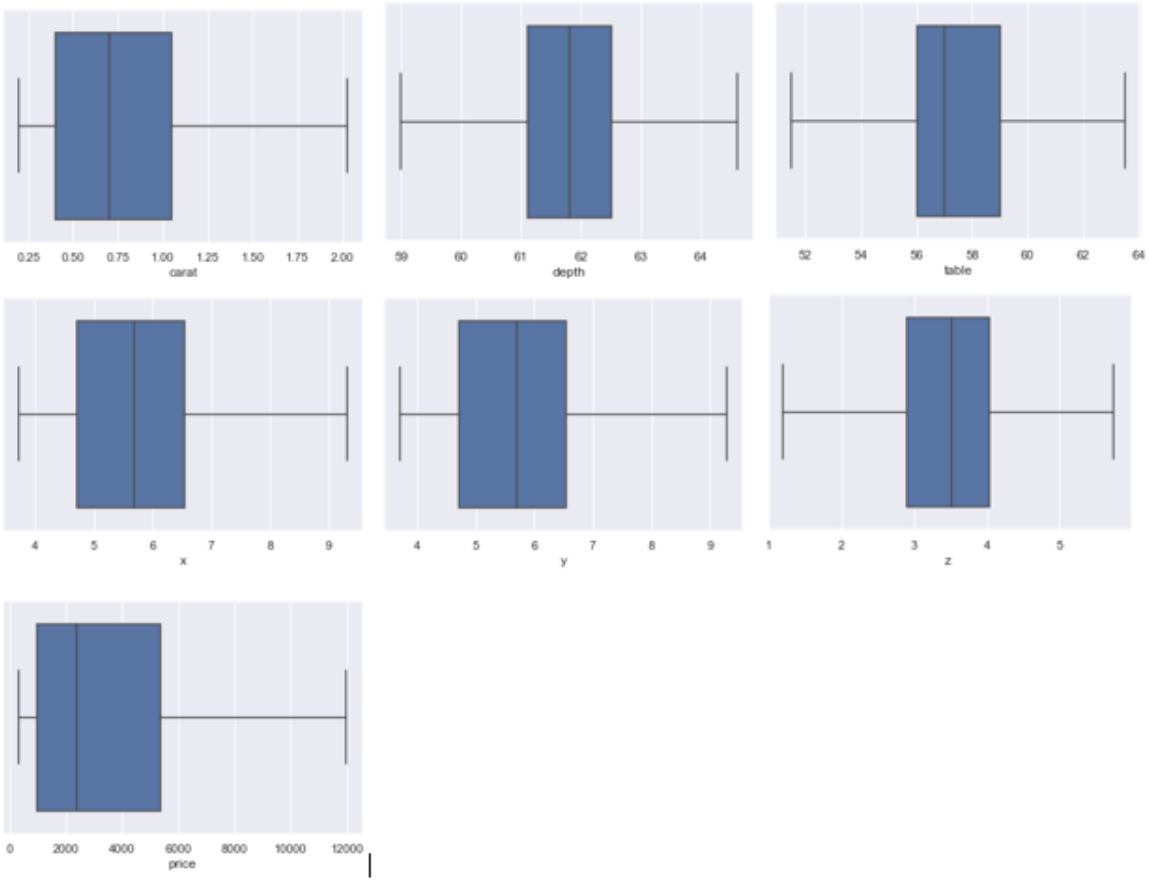


Figure 2: Outliers (After Treatment)

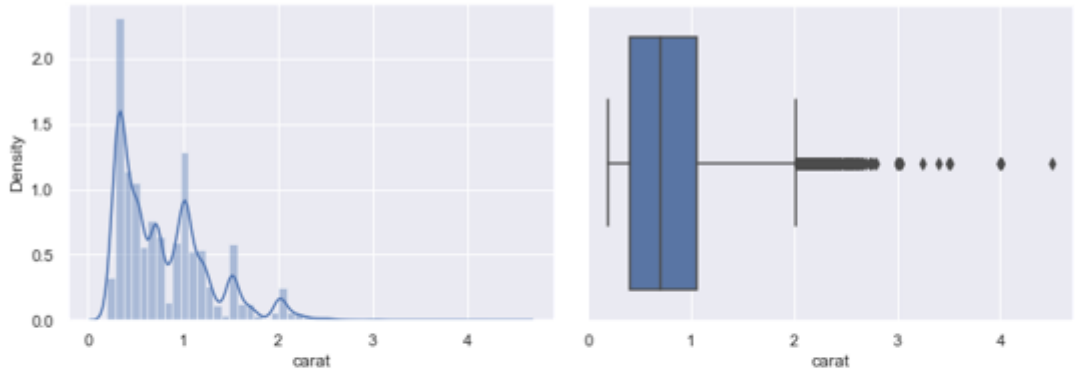
Observation:

We can clearly see that after doing the treatment of outliers there are no outliers in the data set all the outliers are imputed using median.

STEP 4: UNIVARIATE ANALYSIS

Numerical Variables:

a) Carat:



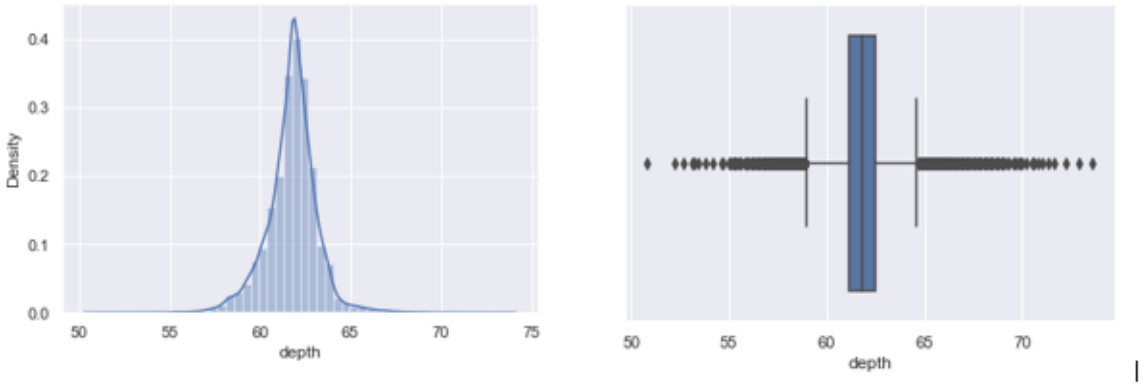

```
Description of carat
.............................
count    26925.000000
mean         0.797821
std          0.477085
min          0.200000
25%          0.400000
50%          0.700000
75%          1.050000
max          4.500000
Name: carat, dtype: float64
Description of carat
```

Figure 3: Distribution of carat

Observation:

The distribution of data in carat seems to positively skewed, as there are multiple peaks points in the distribution there could multimode and the box plot of carat seems to have large number of outliers. In the range of 0 to 1 where majority of data lies.

b) Depth:



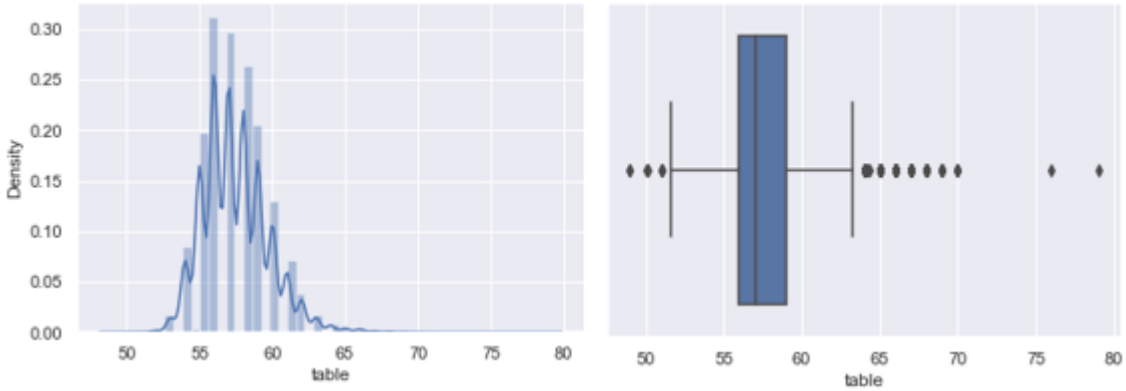

```
Description of depth
.............................
count    26925.000000
mean        61.746982
std          1.393457
min         50.800000
25%         61.100000
50%         61.800000
75%         62.500000
max         73.600000
Name: depth, dtype: float64
Description of depth
```

Figure 4: Distribution of depth

Observation:

1. The distribution of depth seems to be normal distribution,
2. The depth ranges from 55 to 65.
3. The box plot of the depth distribution holds many outliers.

c) Table:



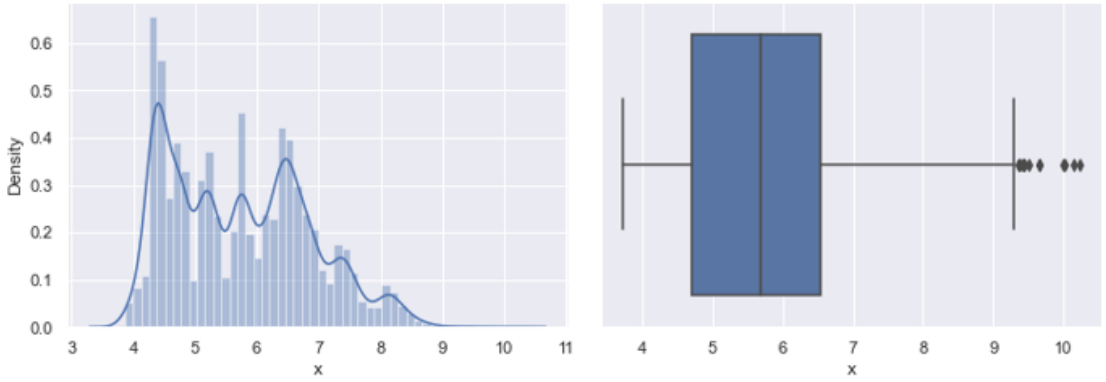

```
Description of table
.............................
count    26925.000000
mean        57.455305
std          2.231327
min         49.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         79.000000
Name: table, dtype: float64
Description of table
```

Figure 5: Distribution of Table

Observation:

1. The distribution of table also seems to be positively skewed.
2. The box plot of table has outliers.
3. The data distribution where there is maximum distribution is between 55 to 65.

d) X (Length of the cubic zirconia)



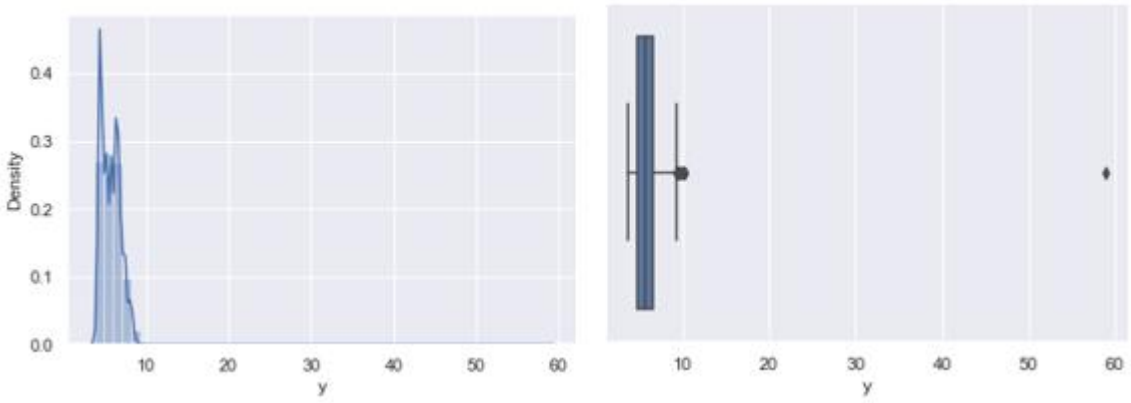

```
Description of x
.........................
count    26925.000000
mean         5.729385
std          1.126081
min          3.730000
25%          4.710000
50%          5.690000
75%          6.550000
max         10.230000
Name: x, dtype: float64
Description of x
```

Figure 6: Distribution of x

Observation:

1. The distribution of x (Length of the cubic zirconia in mm.) is positively skewed.
2. The box plot of the data consists of many outliers.
3. The distribution rages from 4 to 8.

e) Y (Width of the cubic zirconia)




```
Description of y
.........................
count    26925.000000
mean         5.733152
std          1.163820
min          3.710000
25%          4.710000
50%          5.700000
75%          6.540000
max         58.900000
Name: y, dtype: float64
Description of y
```

Figure 7: Distribution of y

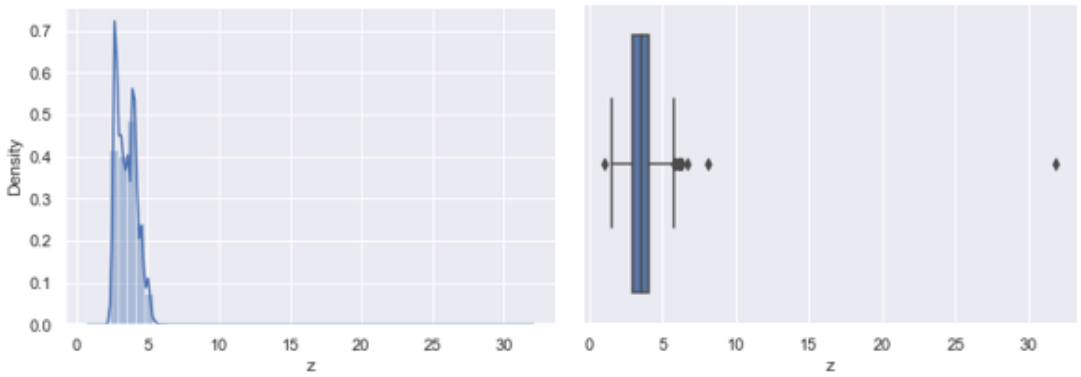
Observation:

1. The distribution of Y (Width of the cubic zirconia in mm.) is positively skewed.

2. The box plot also consists of outliers.

3. The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market.

f) Z (Height of the cubic zirconia)



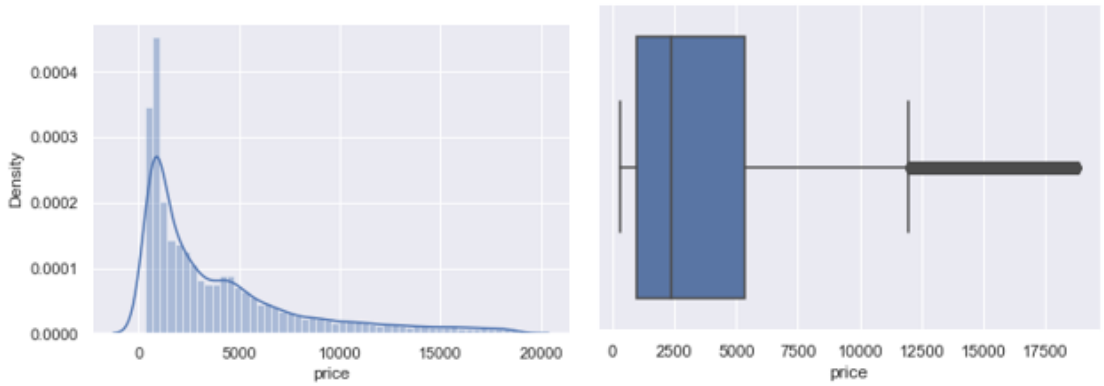

```
Description of z
.........................
count    26925.000000
mean         3.538820
std          0.717483
min          1.070000
25%          2.900000
50%          3.520000
75%          4.040000
max         31.800000
Name: z, dtype: float64
Description of z
```

Figure 8: Distribution of z

Observation:

1. The distribution of z (Height of the cubic zirconia in mm.) is positively skewed.
2. The box plot also consists of outliers.
3. The distribution too much positively skewed. The skewness may be due to the diamonds are always made in specific shape. There might not be too much sizes in the market.

g) Price




```
Description of price
.............................
count    26925.000000
mean      3936.249991
std       4020.983187
min        326.000000
25%        945.000000
50%       2373.000000
75%       5353.000000
max      18818.000000
Name: price, dtype: float64
Description of price
```

Figure 9: Distribution of price

Observation:

1. The price has seemed to be positively skewed. The skew is positive.
2. The price has outliers in the data.
3. The price distribution is from rs 326 to 18818.

Categorical Variables

a) CUT:

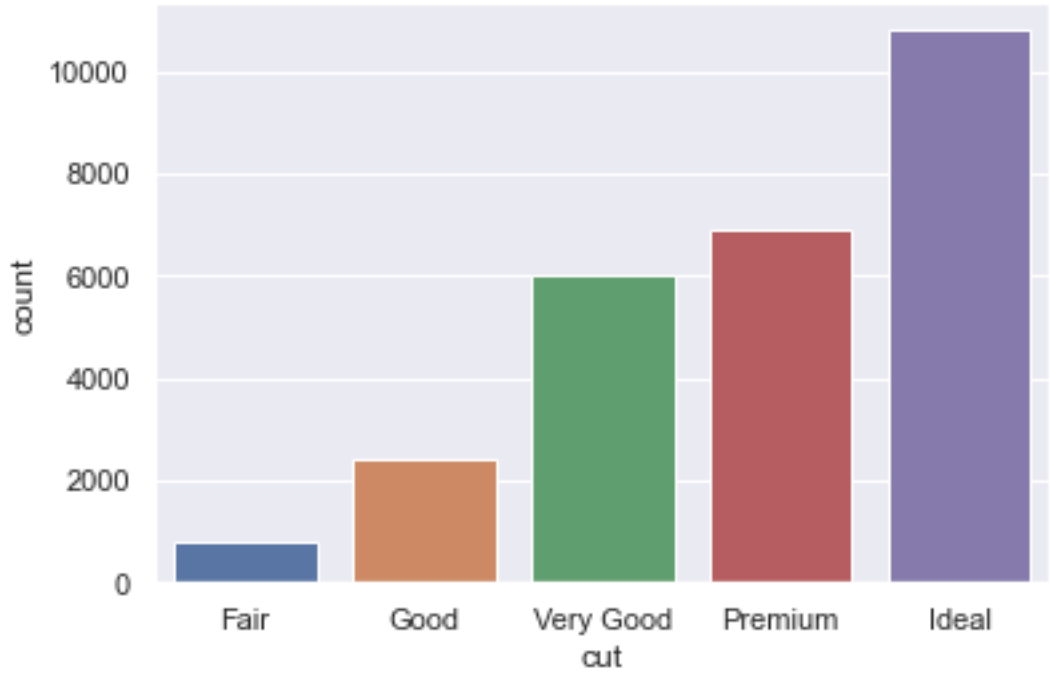


Figure 10: Distribution of cut

Observation:

The most preferred cut seems to be ideal cut for diamonds.

b) COLOUR

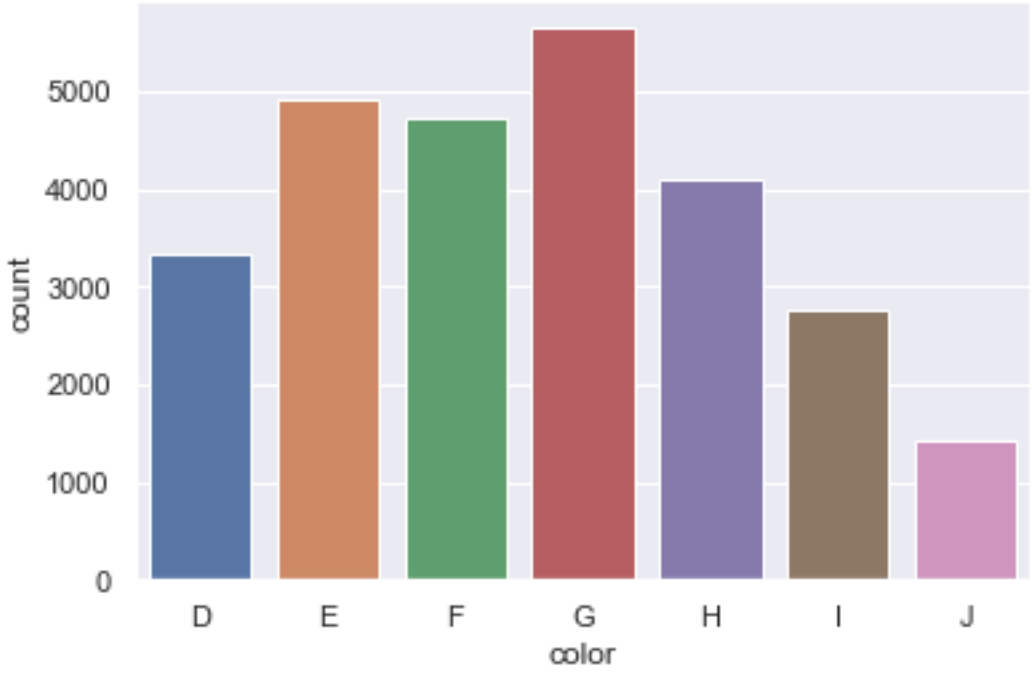


Figure 11: Distribution of colour

Observation:

We have 7 colours in the data, The G seems to be the preferred colour.

c) CLARITY

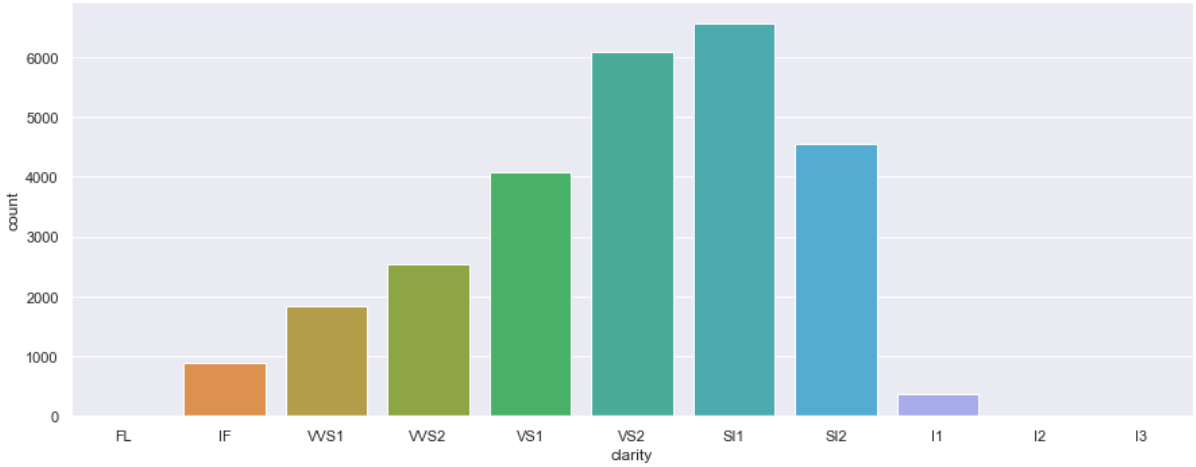


Figure 12: Distribution of clarity

Observation:

We can see here that SL1 has highest no. of diamonds

STEP 5: BI-VARIATE ANALYSIS

a) CUT vs PRICE

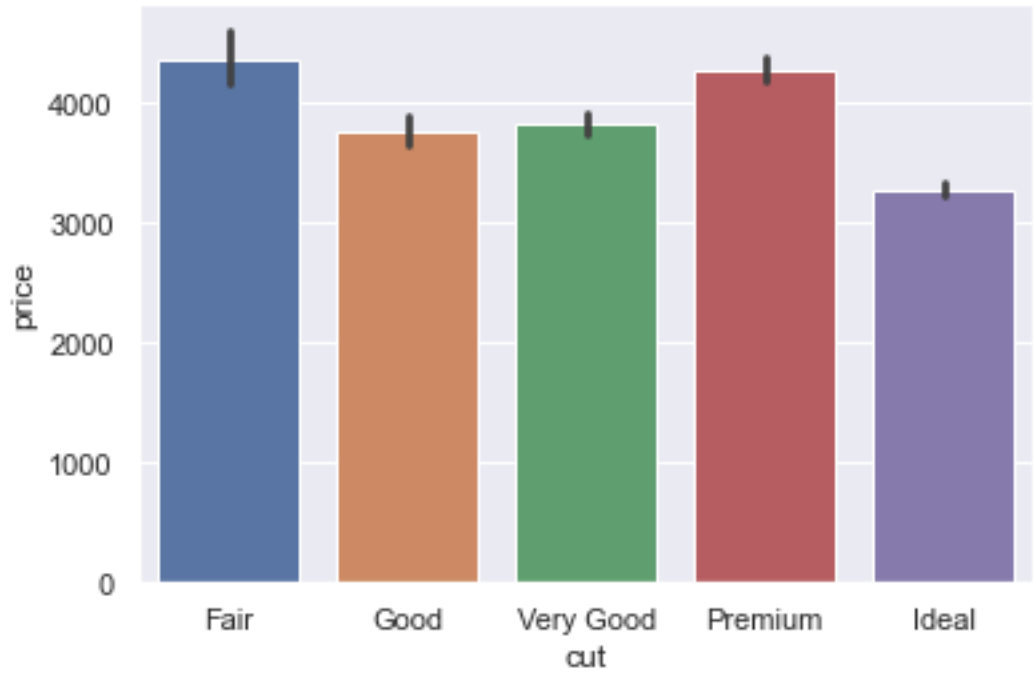


Figure 13: Distribution of cut vs price

Observation:

The reason for the **most preferred cut Ideal** is because those diamonds are **priced lower than other cuts.**

b) COLOUR vs PRICE

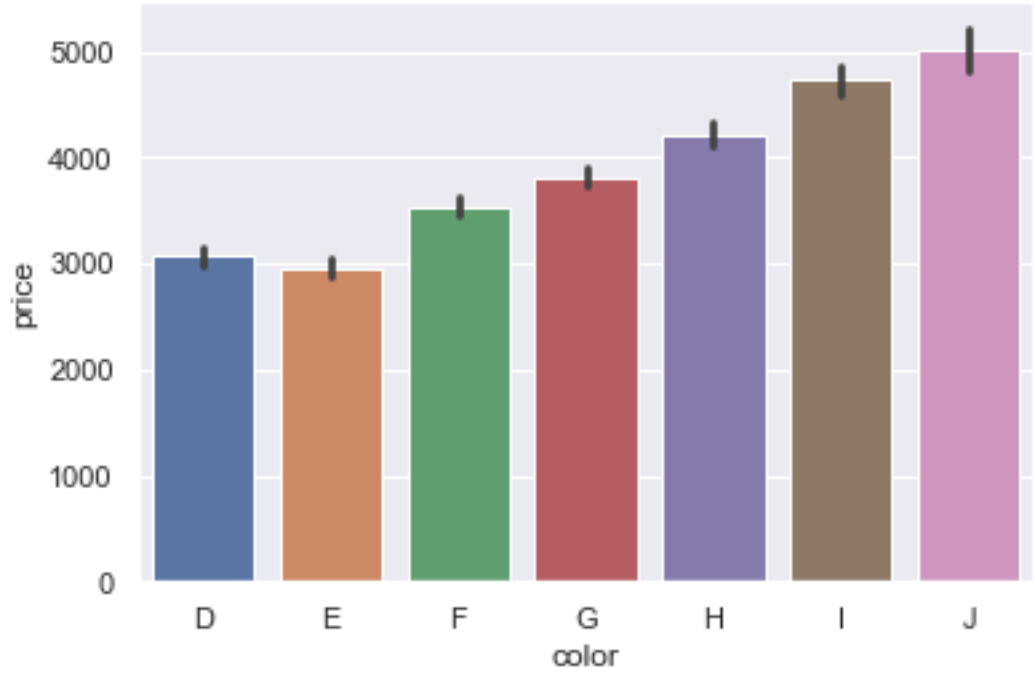


Figure 14: Distribution of colour vs price

Observation:

We see the **G** is priced in the **middle** of the seven colours, whereas **J** being the **worst colour price seems too high.**

c) CLARITY vs PRICE

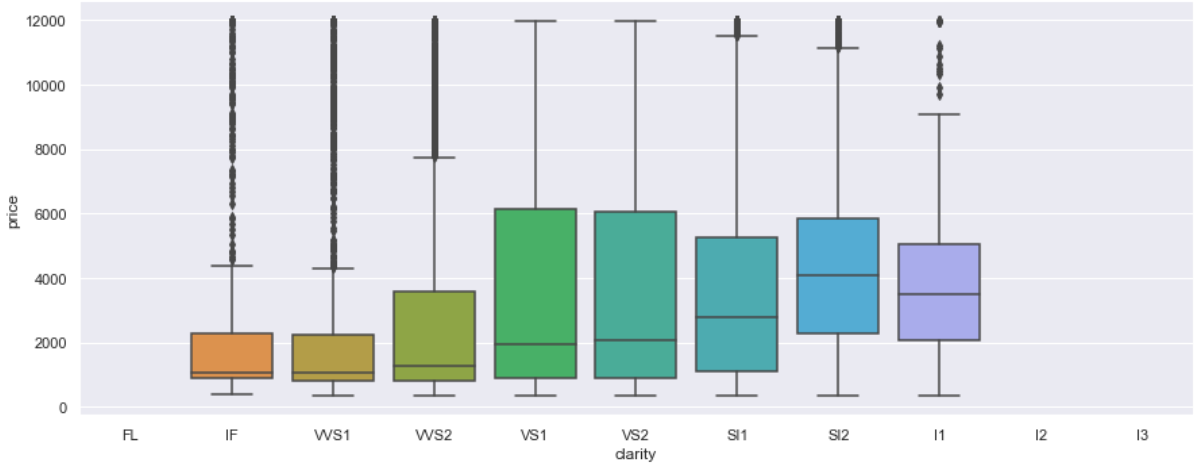


Figure 15: Distribution of clarity vs price

Observation:

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

d) X,Y,Z vs PRICE

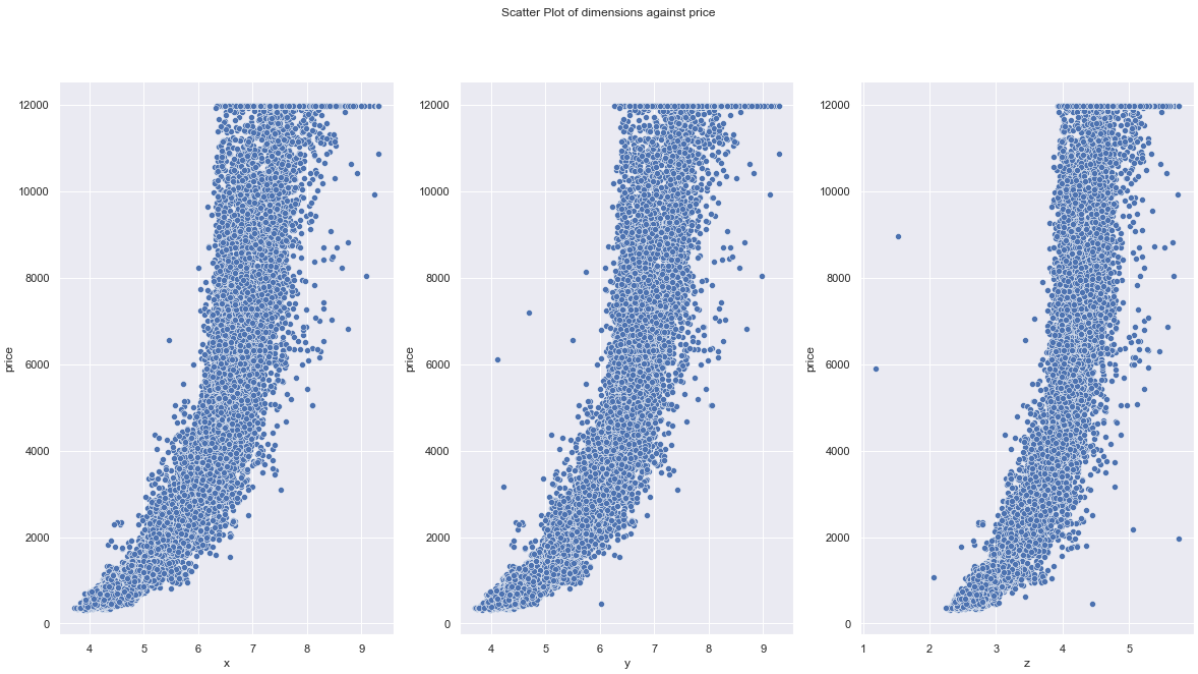


Figure 16: Distribution of X,Y,Z vs price

Observation:

We can see here as the prices increases according to the increase in size of x, y and z

STEP 6: MULTIVARIATE ANALYSIS

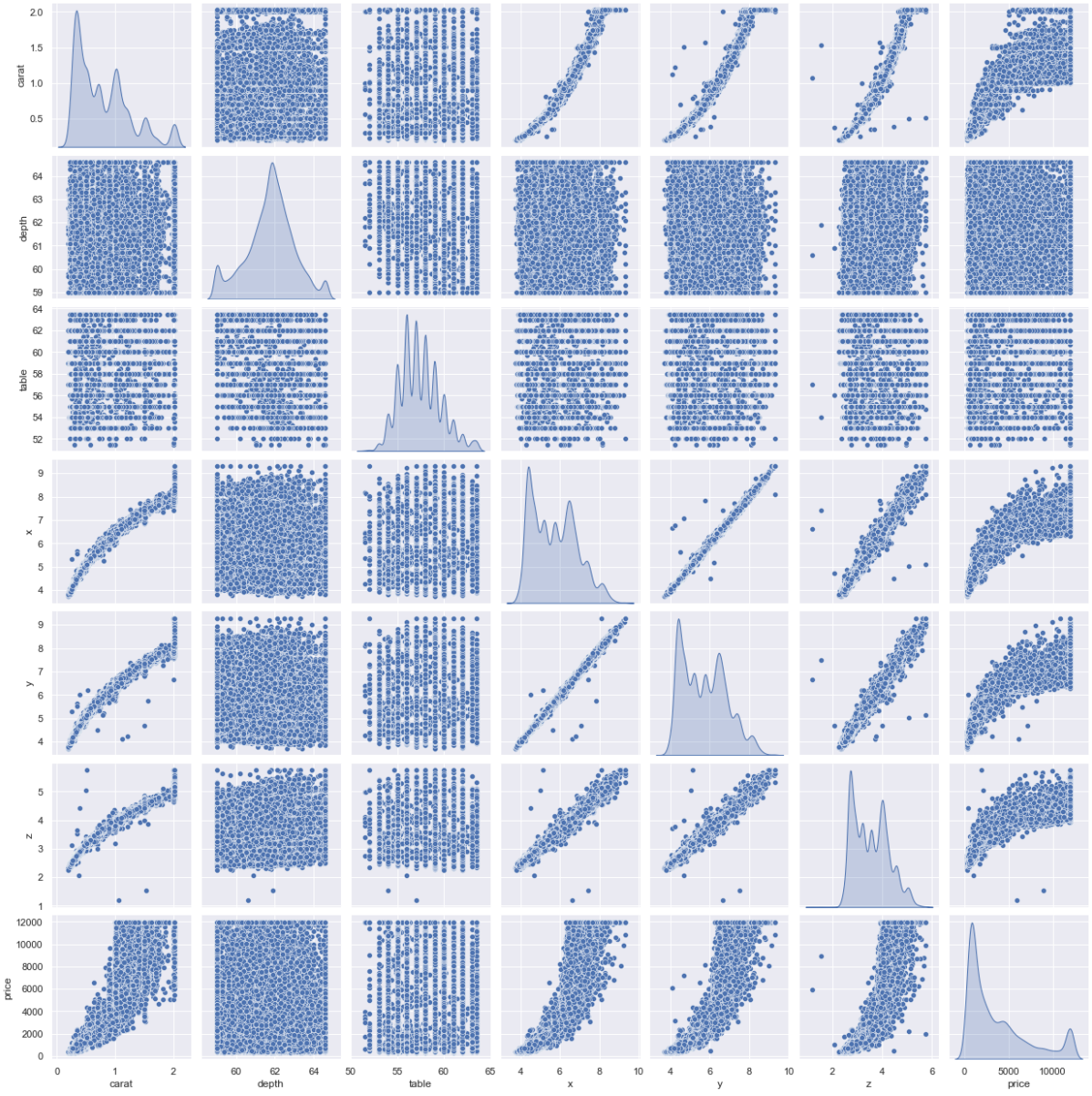


Figure 17: Multivariate Analysis of Dataset

Figure 18: Multivariate Analysis of carat vs price

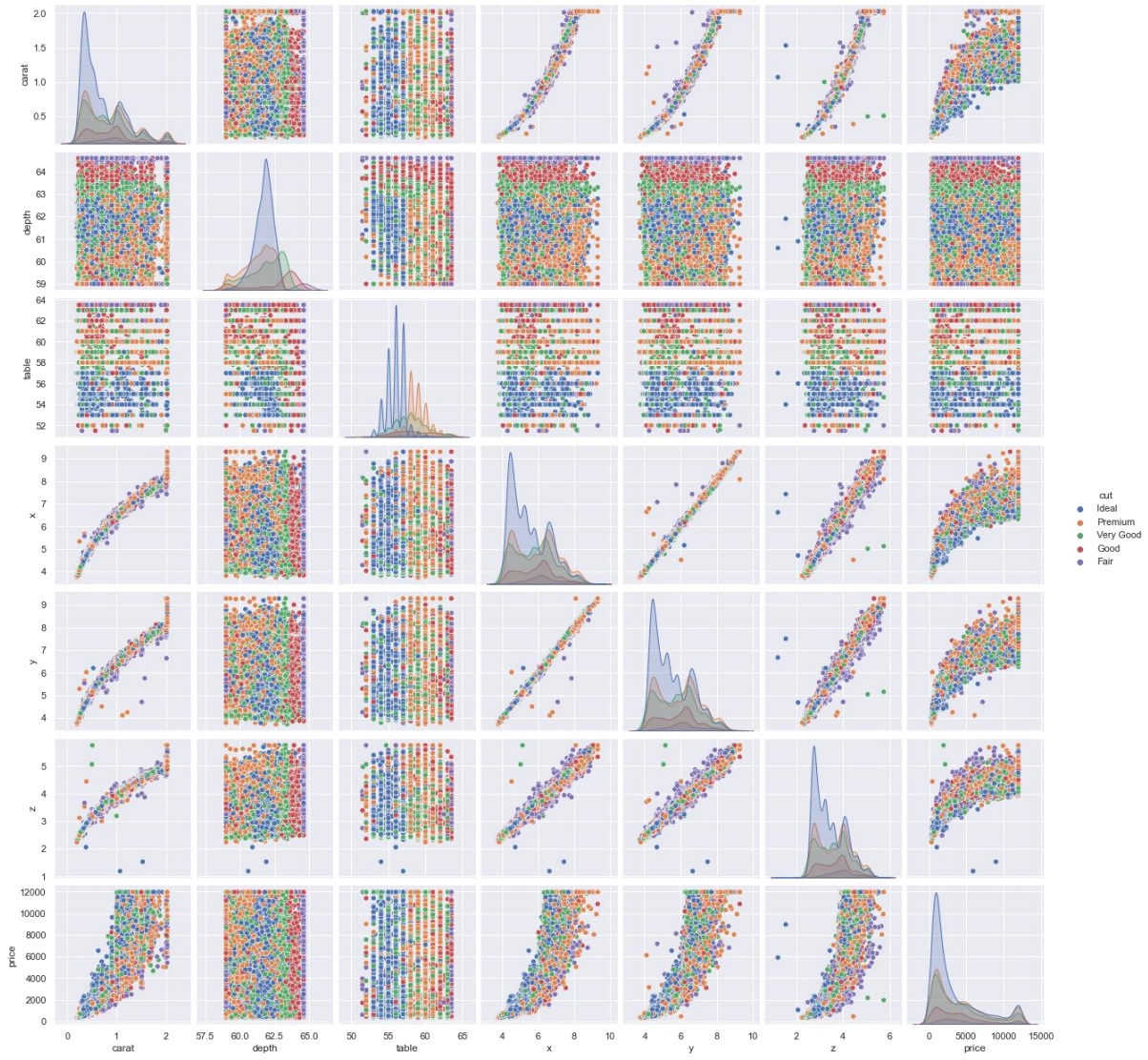


Figure 19: Multivariate Analysis of Dataset vs cut

Observation:

- Ideal cut gems are clustered at low priced region compared to high priced region
- x, y, z and price is a non-linear relation
- Price is highly corelated with carat, dimensions (x, y, z)

CORRELATION MATRIX

	carat	depth	table	x	y	z	price
carat	1.000000	0.029735	0.187134	0.982880	0.981960	0.980882	0.936765
depth	0.029735	1.000000	-0.289163	-0.019676	-0.022720	0.094916	-0.000845
table	0.187134	-0.289163	1.000000	0.199653	0.194015	0.160519	0.137915
x	0.982880	-0.019676	0.199653	1.000000	0.998489	0.990898	0.913409
y	0.981960	-0.022720	0.194015	0.998489	1.000000	0.990533	0.914838
z	0.980882	0.094916	0.160519	0.990898	0.990533	1.000000	0.908599
price	0.936765	-0.000845	0.137915	0.913409	0.914838	0.908599	1.000000

Table 5: Correlation Matrix of Dataset

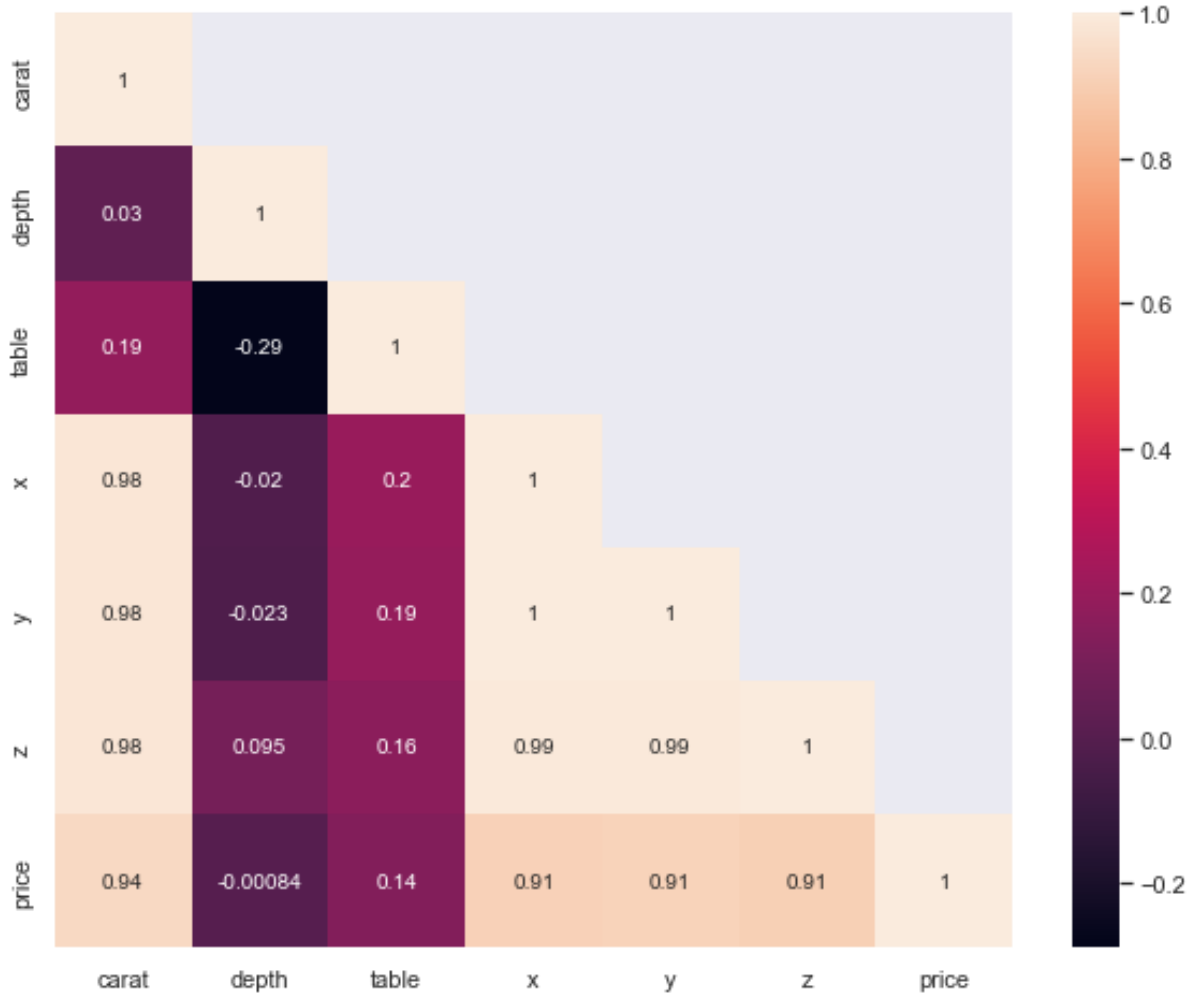


Figure 20: Heat Map of Dataset

Observation:

1. This matrix clearly shows the presence of multi collinearity in the dataset.
2. There is high correlation between carat and x,y,z,price and price with x,y,z and z with x and y

SKWEWESS TABLE

```
price    1.000000
carat    0.936765
y        0.914838
x        0.913409
z        0.908599
table    0.137915
depth   -0.000845
Name: price, dtype: float64
```

Table 6: Skweness of Dataset

Observation:

- It can be inferred that most features correlate with the price of Diamond. The notable exception is "depth" which has a negligible correlation (<1%).

Exploratory Data analysis Summary

The inferences drawn from the above **Exploratory Data analysis:**

Observation-1:

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

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

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

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

Observation-2:

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

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

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

Observation-3:

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

Observation-4:

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

Observation-5:

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

Do you think scaling is necessary in this case?

Scaling or Standardizing the features around the centre and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bias.

For example, A variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization will give the variable with the larger range weight of 1000 in the analysis. Transforming the data to comparable scales can prevent this problem.

In this data set we can see the all the variable are in different scale i.e price are in 1000s unit and depth and table are in 100s unit, and carat is in 10s. So its

necessary to scale or standardise the data to allow each variable to be compared on a common scale. With data measured in different "units" or on different scales (as here with different means and variances) this is an important data processing step if the results are to be meaningful or not dominated by the variables that have large variances.

But is scaling necessary in this case?

No, it is not necessary, we'll get an equivalent solution whether we apply some kind of linear scaling or not. But recommended for regression techniques as well because it would help gradient descent to converge fast and reach the global minima. When number of features becomes large, it helps is running model quickly else the starting point would be very far from minima, if the scaling is not done in pre-processing.

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

2. Use the Pre-processed <u>Full Data</u> 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?

<u>a) Check Correlation Matrix in form of HeatMap</u>

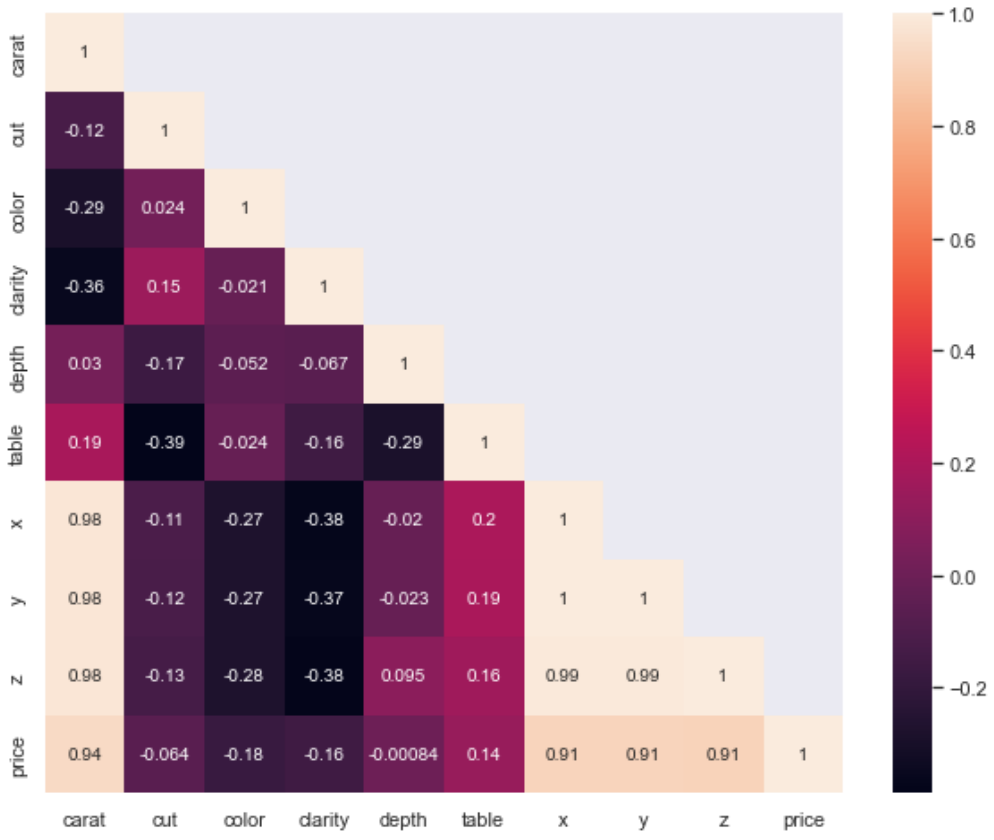


Figure 21: Correlation Matrix in form of Heat Map of Dataset

<u>b) Calculate the Variance Inflation Factor (VIF).</u>

```
carat ---> 122.72929963878924
cut ---> 7.246960772034961
color ---> 5.545479730513101
clarity ---> 5.426739259682176
depth ---> 1173.4714394827827
table ---> 841.1483504379042
x ---> 10845.48353444513
y ---> 9411.890074682238
z ---> 3180.4894711998472
```

Observation:

We can observe there are very strong multi collinearity present in the data set. Ideally it should be within 1 to 5.

We can consider a rule of thumb that if vif is greater than 5, we can choose to drop the variable as there can be a problem of multicollinearity. This essentially means that we can choose to drop a predictor variable whose 80% variation is being explained by the other predictor variables

C) Linear Regression using statsmodels

OLS Regression Results

Dep. Variable:	price	R-squared:	0.931
Model:	OLS	Adj. R-squared:	0.931
Method:	Least Squares	F-statistic:	4.037e+04
Date:	Sun, 25 Sep 2022	Prob (F-statistic):	0.00
Time:	08:42:31	Log-Likelihood:	-2.2167e+05
No. Observations:	26925	AIC:	4.434e+05
Df Residuals:	26915	BIC:	4.434e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-4005.9987	667.037	-6.006	0.000	-5313.426	-2698.571
carat	8840.4426	68.873	128.358	0.000	8705.448	8975.438
cut	68.9302	5.095	13.528	0.000	58.943	78.918
color	273.9660	3.442	79.598	0.000	267.220	280.712
clarity	437.5862	3.750	116.694	0.000	430.236	444.936
depth	28.9087	9.491	3.046	0.002	10.306	47.511
table	-24.0383	3.136	-7.665	0.000	-30.185	-17.891
x	-1190.6108	101.434	-11.738	0.000	-1389.427	-991.795
y	1520.8029	99.591	15.270	0.000	1325.599	1716.007
z	-1138.6637	122.273	-9.312	0.000	-1378.326	-899.002

Omnibus:	3675.822	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	14831.316
Skew:	0.636	Prob(JB):	0.00
Kurtosis:	6.406	Cond. No.	1.03e+04

Table 7: Linear Regression Model 1

Observation:

R2R2 value and the Adjusted R2R2 value says **Model 1 can easily explain 93% of the data**

For the $t-statistic$ t−statistic_ for every co-efficient of the Linear Regression the null and alternate Hypothesis is as follows:

H0 *: The variable is significant.*

H1*: The variable is not significant.*

Lower the p-value for the t-statistic more significant are the variables.
Assuming null hypothesis is true, i.e there is no relationship between this variable with price. from that universe we have drawn the sample and on this sample we have found this co-efficient for the variable shown above.

Now we can ask what is the probability of finding this co-efficient in this drawn sample if in the real world the co-efficient is zero. As we see here the overall P value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression co-efficient is not '0'. Here all regression co-efficients are not '0'.

So we can say that the attribute which are having p value greater than 0.05 are poor predictor for price.

d) Check The Scatter Plot And Displot Of Residuals

Figure 22: Scatter Plot of Actual vs Precited residuals

Figure 23: Displot of Actual vs Precited residuals

e) *Linear Relationship between Dependent and Independent Variables*

Figure 24: Scatter Plot of Linear Relationship between Independent and Dependent

f) Calculate Mean Squared Error – MSE and Root Mean Squared Error - RMSE

Mean Squared Error – MSE: `828690.7545368613`

Root Mean Squared Error – RMSE: `910.3245325359859`

g) Best Parameters

```
Intercept   -4005.998683
carat        8840.442626
cut            68.930222
color         273.966039
clarity       437.586239
depth          28.908675
table         -24.038261
x           -1190.610819
y            1520.802939
z           -1138.663749
dtype: float64
```

The final Linear Regression equation is:

(-4006.0) * Intercept + (8840.44) * carat + (68.93) * cut + (273.97) * color + (437.59) * clarity + (28.91) * depth + (-24.04) * table + (-1190.61) * x + (1520.8) * y + (-1138.66) * z

price = b0 + b1 *carat + b2 * cut + b3 * color + b4 * clarity+ b5 * depth + b6 * table + b7 * x + b8 * y + b9 *z

1) When **carat** increases by 1 unit, diamond price increases by 8840.442626 units, keeping all other predictors constant.

2) When **cut** increases by 1 unit, diamond price increases by 68.930222 units, keeping all other predictors constant.

3) When **color** increases by 1 unit, diamond price increases by 273.966039 units, keeping all other predictors constant.

4) When **clarity** increases by 1 unit, diamond price increases by 437.586239 units, keeping all other predictors constant.

5) When **y** increases by 1 unit, diamond price increases by 1520.802939 units, keeping all other predictors constant.

6) When **depth** increases by 1 unit, diamond price increases by 28.908675 units, keeping all other predictors constant.

Below are the Six attributes that are most important attributes for predicting the price:

(1) 'Carat',

(2) 'Cut',

(3) 'color',

(4) 'clarity',

(5) depth,

(6) width i.e 'y'.

There are also some negative co-efficient values, for instance, corresponding co-efficient (-1190.610819) for 'x',(-1138.663749) for z and (-24.038261) for table This implies, these are inversely proportional with diamond price.

Conclusion:

- On the given data set we can see the 'X' i.e Length of the cubic zirconia in mm. having negative co-efficient i.e. -1190.610819. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stone.
- Similarly, for the 'z' variable having negative co-efficient i.e -1138.663749. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stone.
- Also we can see the 'y' width in mm having positive co-efficient i.e 1520.802939 . And the p value is less than 0.05, so we can conclude that higher the width of the stone is a higher profitable stone.
- Finally, we can conclude that best 6 attributes that are most important are **'Carat', 'Cut', 'color','clarity', depth and width i.e 'y'** for predicting the price.

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.

When prediction accuracy of the price is the only objective, then we have divided the data into a training and a test set, chosen randomly in 70:30 ratio.

Exploring the coefficients for each of the independent attributes

- The coefficient for carat is 8910.226902696053
- The coefficient for cut is 74.0661724839455
- The coefficient for color is 273.06848774088866
- The coefficient for clarity is 439.1103520124743
- The coefficient for depth is 17.390942555597725
- The coefficient for table is -25.941584193425747
- The coefficient for x is -1408.720639183714
- The coefficient for y is 1594.9332611533919
- The coefficient for z is -942.1707088511994

Observation:

Y=mx +c (m= m1,m2,m3...m9) here 9 different co-efficient will learn align with the intercept which is "c" from the model.

From the above coefficients for each of the independent attributes we can conclude:

- The one unit increase in carat increases price by 8910.226902696053.
- The one unit increase in cut increases price by 74.0661724839455.
- The one unit increase in colour increases price by 273.06848774088866.
- The one unit increase in clarity increases price by 439.1103520124743.
- The one unit increase in y increases price by 1594.9332611533919.
- The one unit increase in depth increases price by 17.390942555597725,
- But the one unit increase in table decreases price by -25.941584193425747,
- The one unit increase in x decreases price by -1408.720639183714,
- The one unit increase in z decreases price by -942.1707088511994.

Checking the intercept for the model

The intercept for our model is -3128.1000410056035

Observation:

The intercept (often labelled the constant) is the expected mean value of Y when all X=0. If X never equals 0, then the intercept has no intrinsic meaning.

The intercept for our model is -3128.1000410056035. In present case when the other predictor variable are zero i.e like carat, cut, color, clarity all are zero then the C=-3128. (Y = m1X1 + m2X2+ ….. + mnXn + C + e) that means price is -3128. which is meaningless. We can do Z score or scaling the data and make it nearly zero.

R square on training and testing data

R square on training data: 0.9308919278860922
R square on testing data: 0.9312649702204575

Observation:

R-square is the percentage of the response variable variation that is explained by a linear model. Or:

R-square = Explained variation / Total variation

R-squared is always between 0 and 100%: 0% indicates that the model explains none of the variability of the response data around its mean.100% indicates that the model explains all the variability of the response data around its mean. In this regression model we can see the R-square value on Training and Test data respectively 0.9308919278860922 and 0.9312649702204575.

RMSE on Training and Testing data

Training data RSME : 909.1175874626282
Testing data RSME : 913.6992869600517

Plot the predicted y value vs actual y values for the test data

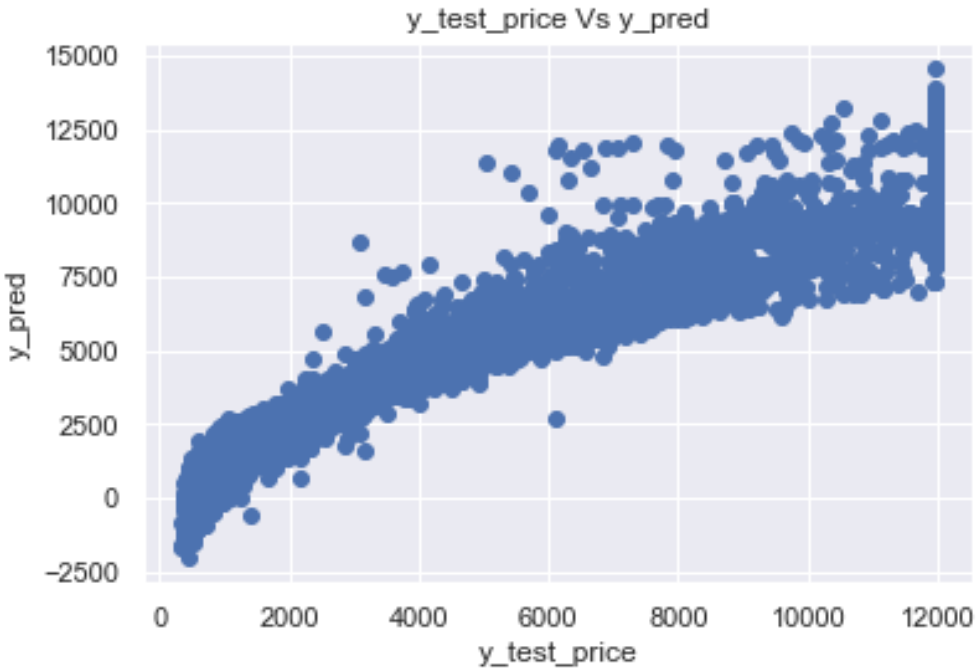


Figure 25: Scatter Plot of Predicted y and Actual y

Applying zscore statsmodels

a) The coefficients for each of the independent attribute:

```
The coefficient for carat is 1.1848755295962892
The coefficient for cut is 0.027059658339948636
The coefficient for color is 0.13456521080175787
The coefficient for clarity is 0.20937050696453188
The coefficient for depth is 0.00613077231923572
The coefficient for table is -0.016176279595997163
The coefficient for x is -0.4567109739737822
The coefficient for y is 0.5135558015931092
The coefficient for z is -0.1887684644215627
```

b) Intercept for our model

```
The intercept for our model is -6.862327432518792e-16
```

c) R2 or coeff of determinant

R2 : `0.931226208782979`

Observation:

Now we can observe by applying z score the intercept became -6.862327432518792e-16. Earlier it was -3128.1000410056035. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

Summary

Inference:

We can see that the is a linear plot, very strong corelation between the predicted y and actual y. But there are lots of spread. That indicated some kind noise present on the data set i.e Unexplained variances on the output.

Linear regression Performance Metrics:

- intercept for the model: -3128.1000410056035
- R square on training data: 0.9308919278860922
- R square on testing data: 0.9312649702204575
- RMSE on Training data: 909.1175874626282
- RMSE on Testing data: 913.6992869600517

As the training data & testing data score are almost inline, we can conclude this model is a Right-Fit Model.

Impact of scaling:

Now we can observe by applying z score the intercept became -6.862327432518792e-16. Earlier it was -3128.1000410056035. the co-efficient has changed, the bias became nearly zero but the overall accuracy still same.

Multi collinearity:

We can observe there are very strong multi collinearity present in the data set.

From statsmodels:

we can see R-squared:0.931 and Adj. R-squared: 0.931 are same. The overall P value is less than alpha.

Below are the Six attributes that are most important attributes for predicting the price:

1) When **carat** increases by 1 unit, diamond price increases by 8840.442626 units, keeping all other predictors constant.

2) When **cut** increases by 1 unit, diamond price increases by 68.930222 units, keeping all other predictors constant.

3) When **color** increases by 1 unit, diamond price increases by 273.966039 units, keeping all other predictors constant.

4) When **clarity** increases by 1 unit, diamond price increases by 437.586239 units, keeping all other predictors constant.

5) When **y** increases by 1 unit, diamond price increases by 1520.802939 units, keeping all other predictors constant.

6) When **depth** increases by 1 unit, diamond price increases by 28.908675 units, keeping all other predictors constant.

Similarly,

- On the given data set we can see the **'X' i.e Length** of the cubic zirconia in mm. having **negative co-efficient** i.e. -1190.610819. And the p value is less than 0.05, so can conclude that as higher the length of the stone is a lower profitable stone.
- Similarly for the **'z' variable having negative co-efficient** i.e -1138.663749. And the p value is less than 0.05, so we can conclude that as higher the 'z' of the stone is a lower profitable stone.
- Also we can see the **'y' width in mm having positive co-efficient** i.e 1520.802939 . And the p value is less than 0.05, so we can conclude that higher the width of the stone is a higher profitable stone.
- Finally we can conclude that best 6 attributes that are most important are **'Carat', 'Cut', 'color','clarity',depth and width i.e 'y'** for predicting the price.

4. Basis on these predictions, what are the insights and recommendations.

Insights and Recommendations:

1. The Gem Stones company should consider the features ‘Carat', 'Cut', 'color','clarity' and width i.e 'y' as most important for predicting the price.
2. To distinguish between higher profitable stones and lower profitable stones so as to have better profit share.
3. As we can see from the model Higher the width('y') of the stone is higher the price.
4. So the stones having higher width('y') should consider in higher profitable stones. The 'Premium Cut' on Diamonds are the most Expensive, followed by 'Very Good' Cut, these should consider in higher profitable stones.
5. The Diamonds clarity with 'VS1' &'VS2' are the most Expensive. So these two category also consider in higher profitable stones.
6. As we see for 'X' i.e Length of the stone, higher the length of the stone is lower the price.
7. So higher the Length('x') of the stone are lower is the profitability. higher the 'z' i.e Height of the stone is, lower the price. This is because if a Diamond's Height is too large Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light. That is why
8. Stones with higher 'z' is also are lower in profitability.

PROBLEM 2 - LOGISTIC REGRESSION

Overview:

You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

Data Dictionary:

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

Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provides some key insights/recommendations to the business.

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.**

DATASET HEAD AND TAIL

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

Table 8: Dataset Head

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
867	868	no	40030	24	4	2	1	yes
868	869	yes	32137	48	8	0	0	yes
869	870	no	25178	24	6	2	0	yes
870	871	yes	55958	41	10	0	1	yes
871	872	no	74659	51	10	0	0	yes

Table 9: Dataset Tail

Observation:

- Dataset has 8 columns.
- The first column (Unnamed column :0) is of no use for analysis and can be removed.

DATASET SHAPE

```
(872, 8)
```

Observation:

- Total no. of Rows = 26967
- Total no. of Columns = 11

DATASET SUMMARY AND SKEWNESS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.0	NaN	NaN	NaN	436.5	251.869014	1.0	218.75	436.5	654.25	872.0
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

Table 10: Dataset Summary

```
Unnamed: 0           0.000000
Salary               3.103216
age                  0.146412
educ                -0.045501
no_young_children    1.946515
no_older_children    0.953951
dtype: float64
```

Table 11: Dataset Skweness

Observation :

From summary, we can see that :-

- max salary(236K) is very high as compared to mean(47K) and median(42K). Hence it contains outlier
- Mean and median of age are approximately similar 39-40. It doesn't contains outlier.
- Education middle 50% of data lies in between 8 to 12 range with few outliers.
- Most employees have no of young children as 0.
- Most of the employees have 1 child who is older than 7 years
- All the columns are positively skewed except education

DATASET INFORMATION

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Unnamed: 0         872 non-null    int64
 1   Holliday_Package   872 non-null    object
 2   Salary             872 non-null    int64
 3   age                872 non-null    int64
 4   educ               872 non-null    int64
 5   no_young_children  872 non-null    int64
 6   no_older_children  872 non-null    int64
 7   foreign            872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

Table 12: Dataset Information

Observation:

- The data set contains 872 observations of data and 7 features. Since non null count is same in every column variable there appears **no null data.**
- Two object variables and six numeric variables.
- Variable "Unnamed: 0" seems useless variable.

EXPLORATORY DATA ANALYSIS

STEP 1 - CHECK AND REMOVE ANY DUPLICATES IN THE DATASET

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

Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign

Observation

No duplicated data is present.

STEP 2 - CHECKING MISSING VALUE

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

Observation

We can confirm that there are no NULL values in the data

STEP 3 - OUTLIER CHECKS AND TREATMENT

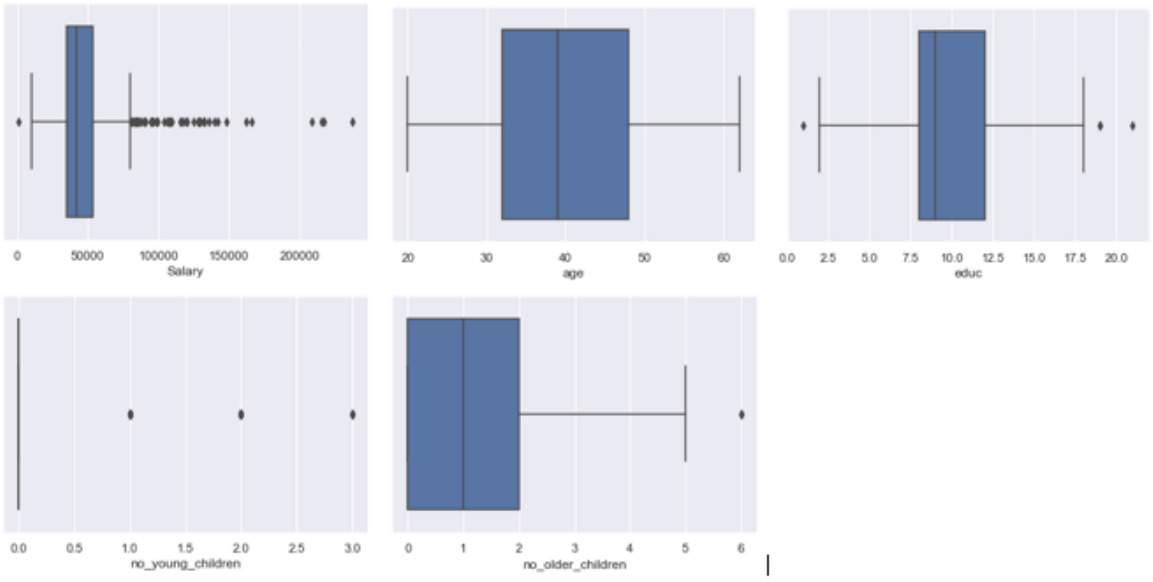


Figure 26: Outliers (Before Treatment)

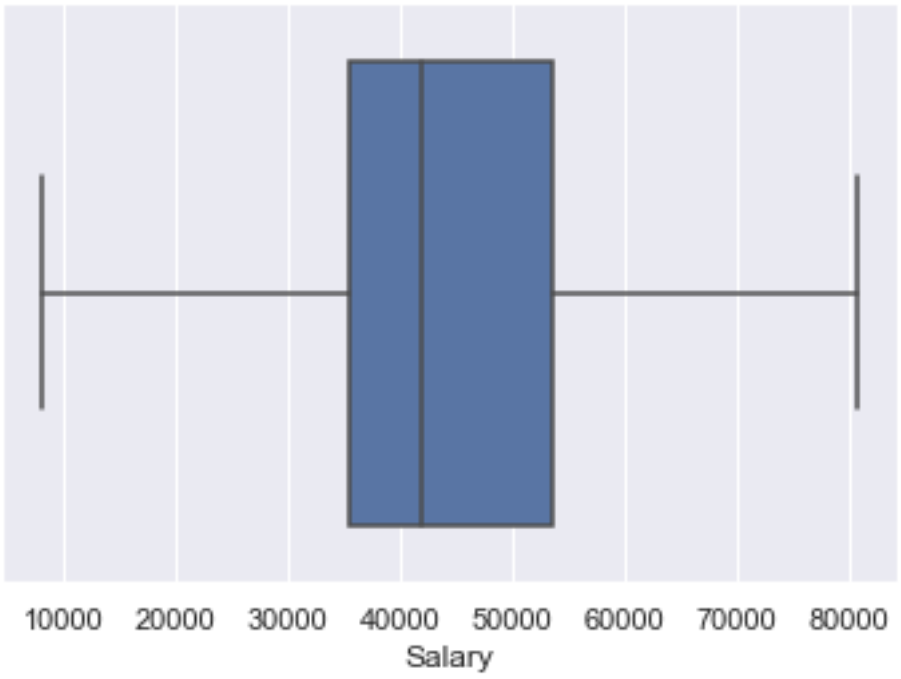


Figure 27: Outliers (After Treatment only Salary)

Observation:

We are only doing outlier treatment for Salary attribute as other columns have very less outliers and that are near lower and upper ranges

STEP 4 - UNIVARIATE ANALYSIS

a) Salary

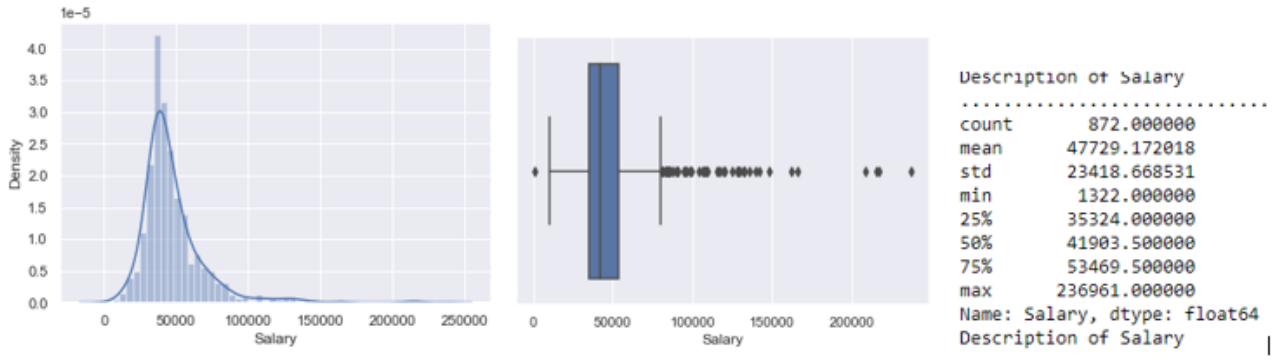


Figure 28: Distribution of Salary

b) age

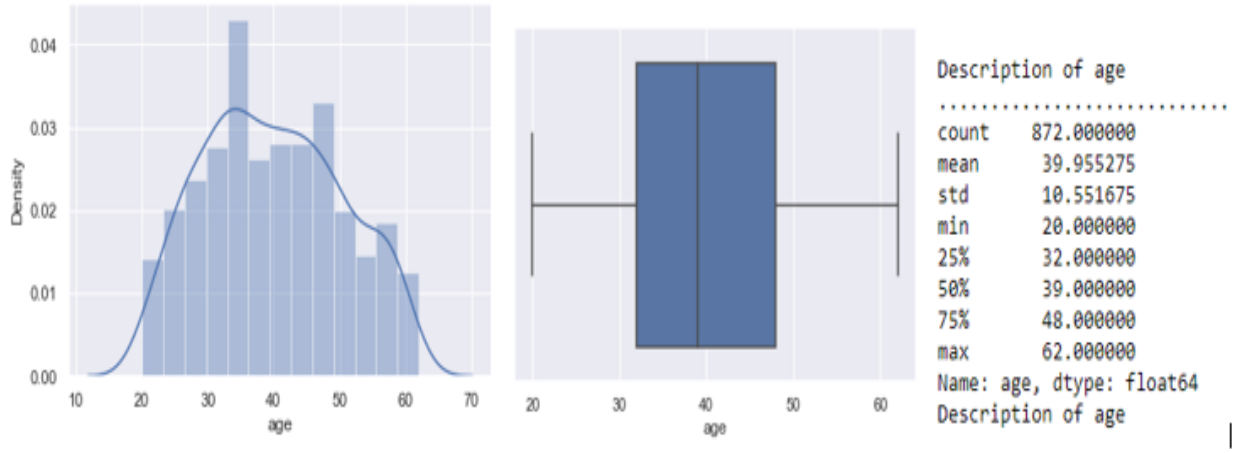


Figure 29: Distribution of Age

c) educ

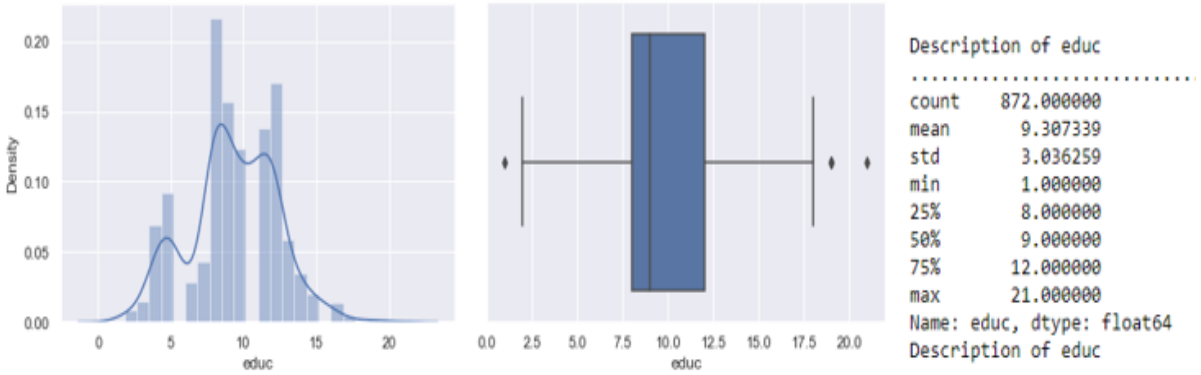


Figure 30: Distribution of Edu

d) no_young_children

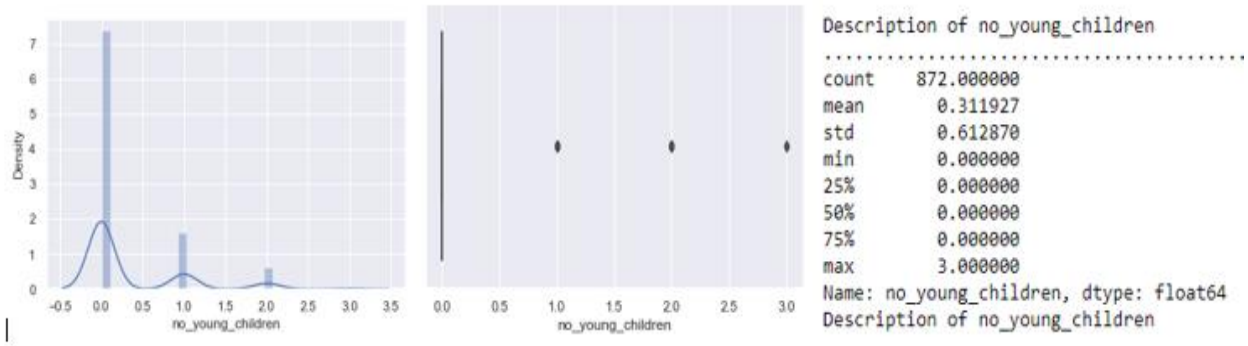


Figure 31: Distribution of No of young children

e) no_older_children

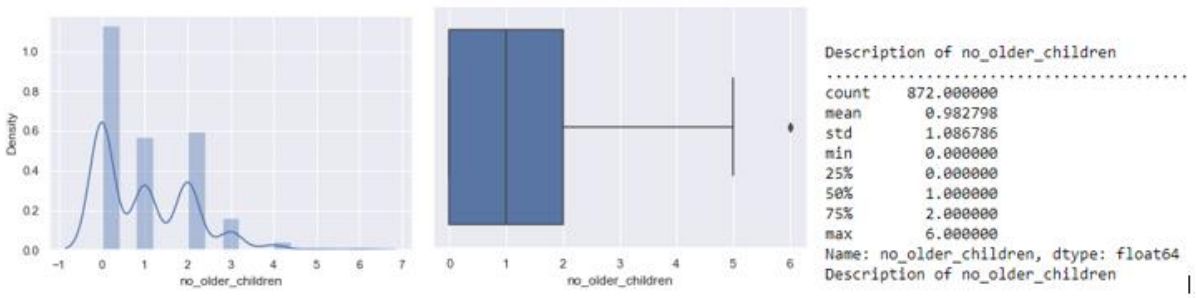


Figure 32: Distribution of No of Older children

Observation

From Distplot and Boxplot we can see that:

- Salary range is 0-100000 for most of the employees. However few employees are getting more salary causing skewness
- Age appears to be normally distributed
- Around 650 employees out of 872 have their young children as 0.
- Around 380 out of 872 employees have no of older children as 0
- Education middle 50% of data lies in between 8 to 12 range with few outliers.
- As evident from above box plot, there are many outliers in salary column.
- education, no of young children and old children's columns have very few outliers.

f) Holiday Package

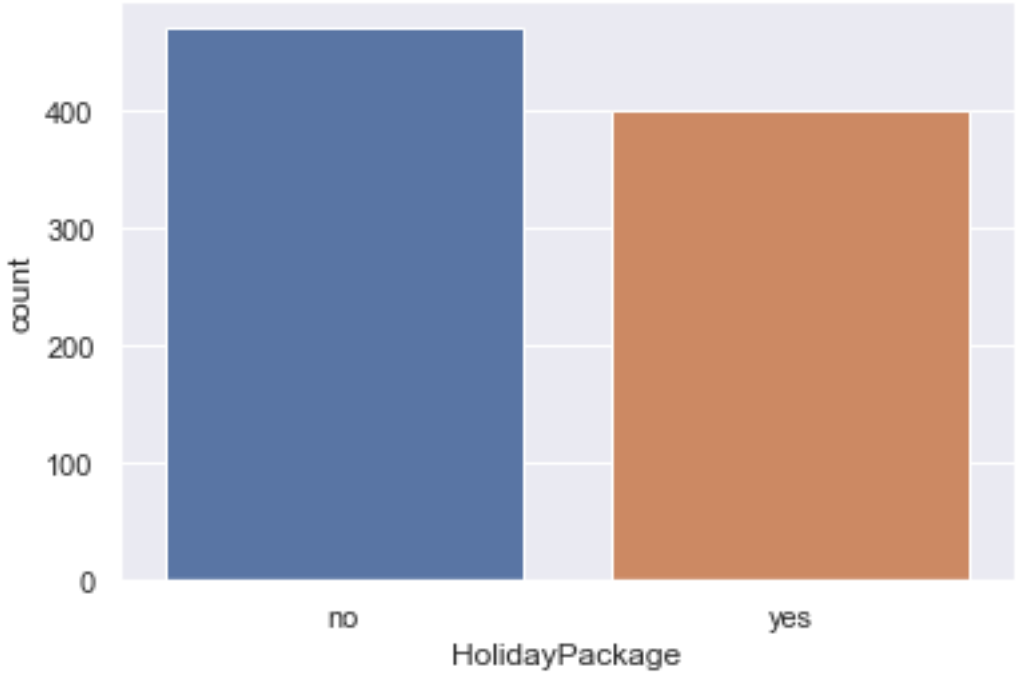


Figure 33: Distribution of Holiday Package

Observation:

We can see count of Employee not using Holiday Packages are more

g) Foreign

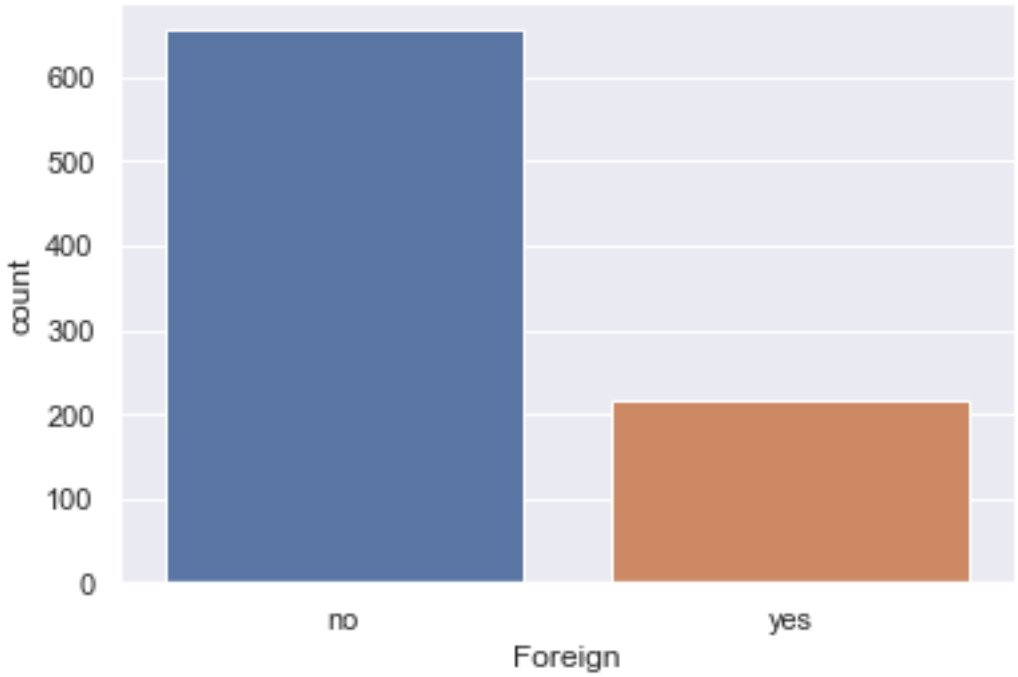


Figure 34: Distribution of Foreign Employees

Observation:

We can see count of Foreign Employee not using Holiday Packages are more

STEP 5 - BIVARIATE ANALYSIS

a) Salary and Age vs Holiday Package

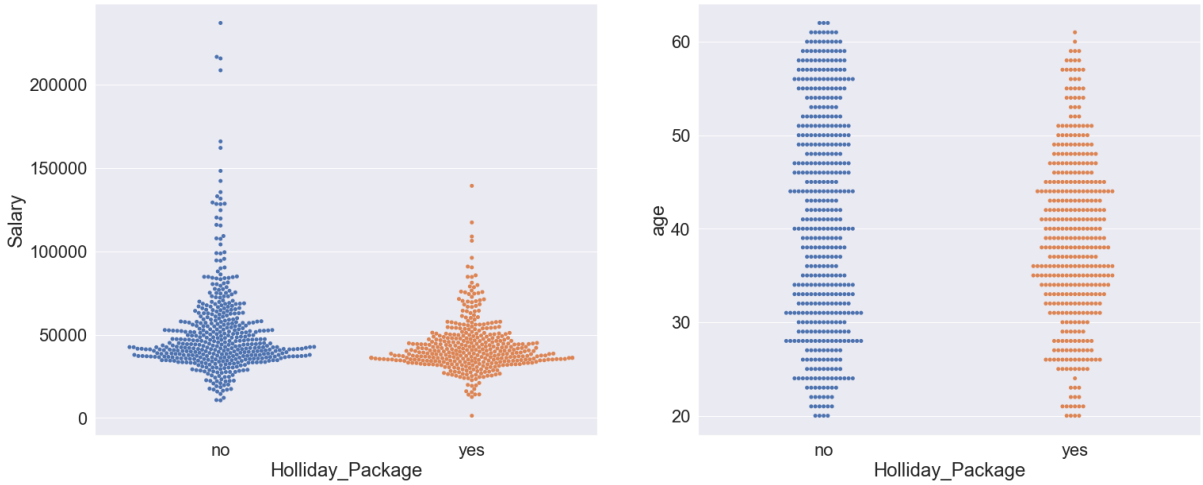


Figure 35: Distribution of Salary and Age vs Holiday Package

Observation:

- As Salary increases to the max value, employees count increases for the not opting for the holiday package. As Age increases beyond 50 level, less employees opt for the holiday package
- We can see employees below salary 150000 have always opted for holiday package.

b) Edc, no. of young children, no. of older children, foreign Vs Holiday Package

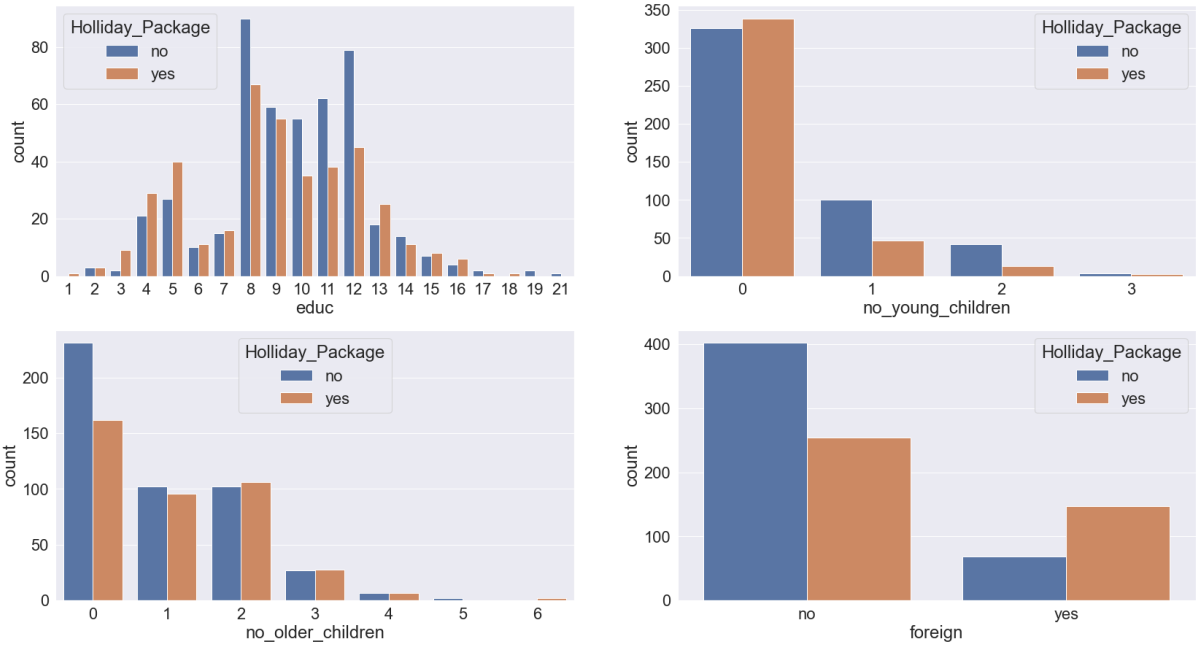


Figure 36: Distribution of Edc, no. of young children, no. of older children, foreign Vs Holiday Package

Observations :

- More Employees opt for Tours if their education level is 3,4,5,6,7,13,14,15,16
- Employees don't opt for tours if they have young child
- Older children count doesn't appears to have much impact on tour opted by employees or not
- Foreigner employees tends to opt more for the tour

c) No. of young Children Vs Holiday Package

Holliday_Package	no	yes	All
no_young_children			
0	326	339	665
1	100	47	147
2	42	13	55
3	3	2	5
All	471	401	872

Table 13: Distribution of No. of young Children Vs Holiday Package

Observation :

We can see that around 24% of employees have one or more young child. Out of these employees, 70% ((100+42+3)/(147+55+5)) are not opting for tours.

d) Foreign Vs Holiday Package

Holliday_Package	no	yes	All
foreign			
no	402	254	656
yes	69	147	216
All	471	401	872

Table 14: Distribution of Foreign Vs Holiday Package

Observation :

As per the data, we can say that 68% of foreign employees are opting for the tour packages.

e) Salary Vs age Vs Holiday Package

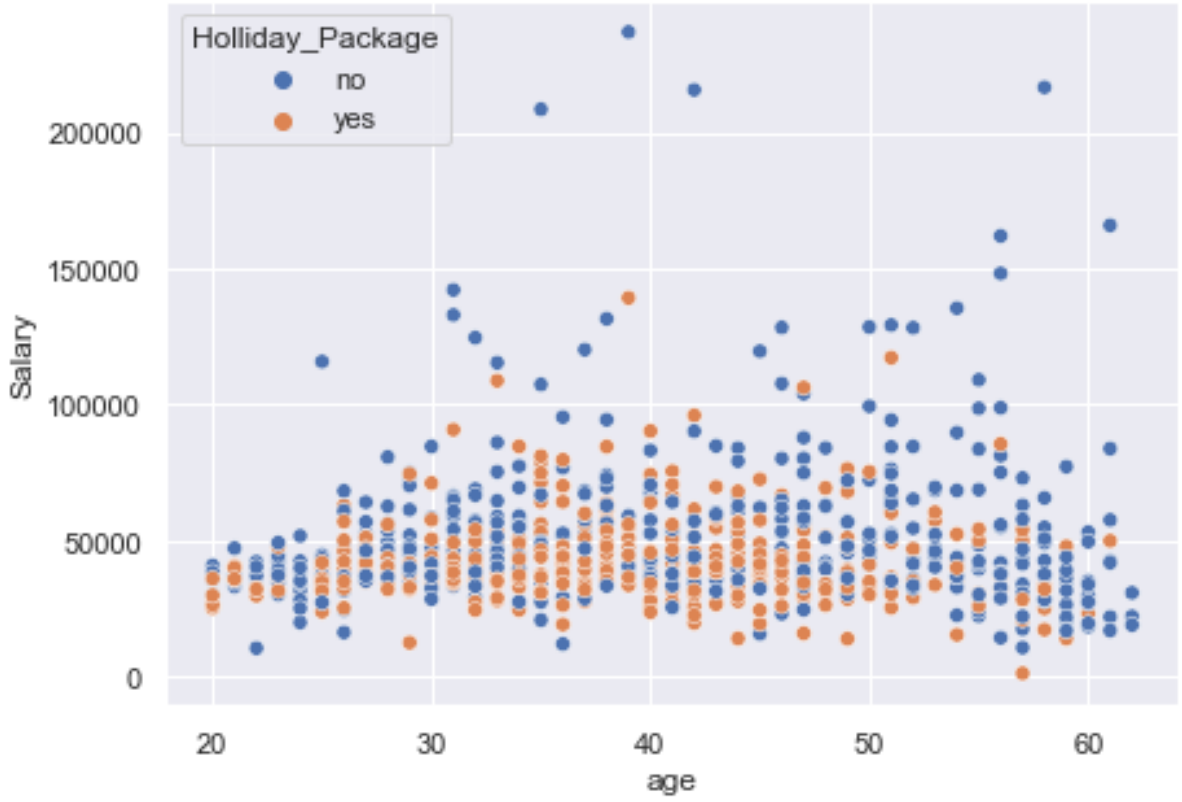


Figure 37: Distribution of Salary Vs age Vs Holiday Package

Observation :

Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package

STEP 6 - MULTIVARIATE ANALYSIS

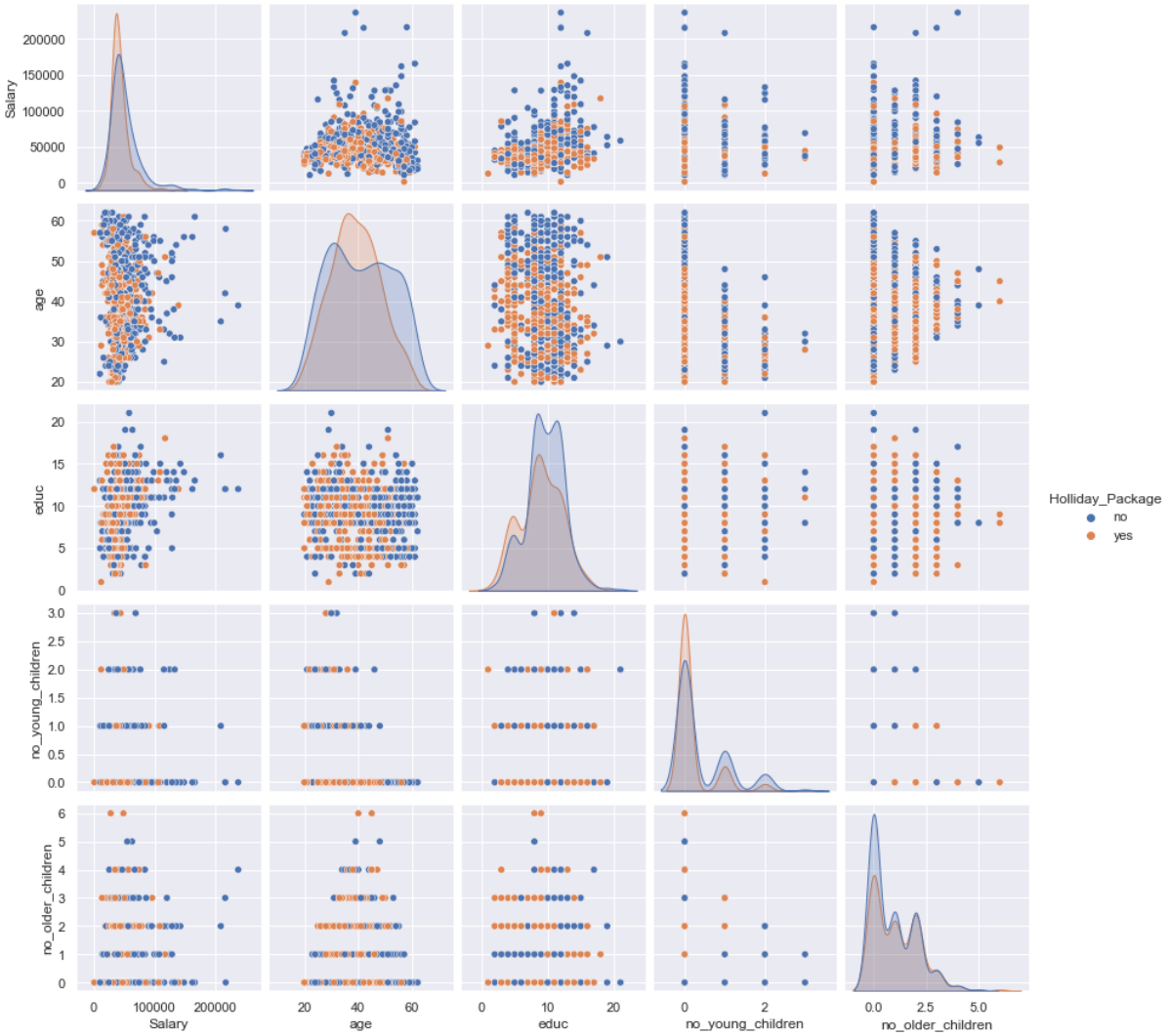


Figure 38: Multi-Variate analysis of Travel Data

Observation :

- There is not much correlation between the data, the data seems to be normal.
- There is no huge difference in the data distribution among the holiday package, I don't see any clear two different distribution in the data

STEP 7 – CORRELATION MATRIX TABLE AND HEAT MAP

	Salary	age	educ	no_young_children	no_older_children
Salary	1.000000	0.071709	0.326540	-0.029664	0.113772
age	0.071709	1.000000	-0.149294	-0.519093	-0.116205
educ	0.326540	-0.149294	1.000000	0.098350	-0.036321
no_young_children	-0.029664	-0.519093	0.098350	1.000000	-0.238428
no_older_children	0.113772	-0.116205	-0.036321	-0.238428	1.000000

Table 15: Correlation Table

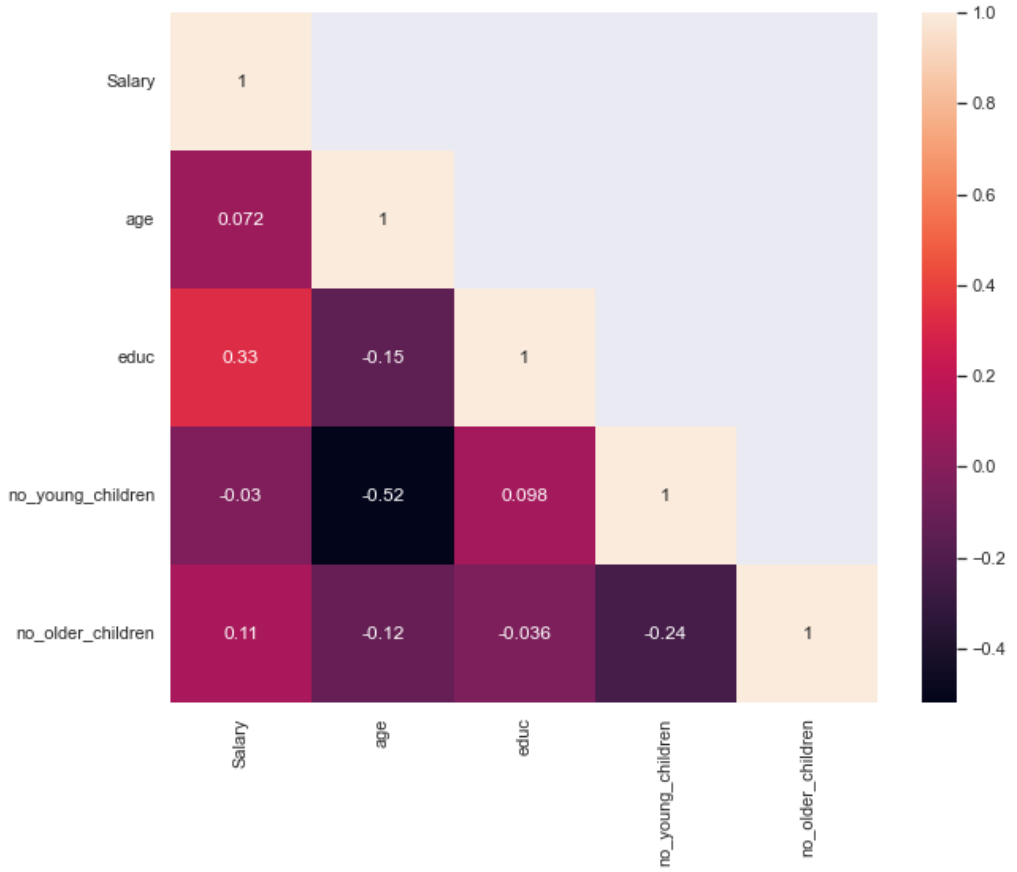


Figure 39: Correlation Heat Map

Observation :

We can see in heatmap & correlation matrix that

- Salary has correlation with educ.
- Age is negatively correlated with No_young_children

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.**

Variance Inflation Factor (VIF) for checking Multicollinearity

Salary VIF = 1.2
Age VIF = 1.56
Educ VIF = 1.41
No_young_children VIF = 1.57
No_older_children VIF = 1.19
Foreign_yes VIF = 1.27

Observation

If VIF is greater than 5, we can choose to drop the variable as there can be a problem of multicollinearity.

Since all the VIF less than 5 thus **No Multicollinearity**

Building Logistical Regression Model

a) Model 1

Logit Regression Results

Dep. Variable:	HolidayPackage_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	865
Method:	MLE	**Df Model:**	6
Date:	Sun, 25 Sep 2022	**Pseudo R-squ.:**	0.1244
Time:	08:42:55	**Log-Likelihood:**	-526.78
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	9.138e-30

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.5432	0.559	4.550	0.000	1.448	3.639
Salary	-2.088e-05	5.26e-06	-3.970	0.000	-3.12e-05	-1.06e-05
Age	-0.0496	0.009	-5.491	0.000	-0.067	-0.032
Educ	0.0342	0.029	1.172	0.241	-0.023	0.091
No_young_children	-1.3287	0.180	-7.386	0.000	-1.681	-0.976
No_older_children	-0.0251	0.074	-0.341	0.733	-0.169	0.119
Foreign_yes	1.3037	0.200	6.519	0.000	0.912	1.696

Observation :

- We can see that the p value of No_older_children is the highest (.733) and it is greator than 0.05.
- Hence it confirms that No_older_children attribute has no impact on dependent variable HolidayPackage

- The adjusted pseudo R-square value is 0.11440780487420366

a) Model 2 (Droping 'No_older_children ')

Logit Regression Results

Dep. Variable:	HolidayPackage_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	866
Method:	MLE	**Df Model:**	5
Date:	Sun, 25 Sep 2022	**Pseudo R-squ.:**	0.1243
Time:	08:42:55	**Log-Likelihood:**	-526.84
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	1.671e-30

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.4783	0.525	4.724	0.000	1.450	3.506
Salary	-2.117e-05	5.19e-06	-4.079	0.000	-3.13e-05	-1.1e-05
Age	-0.0487	0.009	-5.677	0.000	-0.065	-0.032
Educ	0.0351	0.029	1.209	0.227	-0.022	0.092
No_young_children	-1.3080	0.169	-7.747	0.000	-1.639	-0.977
Foreign_yes	1.3028	0.200	6.517	0.000	0.911	1.695

Observation :

Salary VIF = 1.17
Age VIF = 1.42
Educ VIF = 1.4
No_young_children VIF = 1.37
Foreign_yes VIF = 1.27

- VIF indicates there is no Multicollinearity problem
- Based on 'p value' lets drop Educ as its highest (0.227) and it is greator than 0.05. Thus its not significant
- The adjusted pseudo R-square value is 0.11597348232009375

b) Model 3 (Droping 'Educ ')

Logit Regression Results

Dep. Variable:	HolidayPackage_yes	**No. Observations:**	872
Model:	Logit	**Df Residuals:**	867
Method:	MLE	**Df Model:**	4
Date:	Sun, 25 Sep 2022	**Pseudo R-squ.:**	0.1231
Time:	08:42:55	**Log-Likelihood:**	-527.58
converged:	True	**LL-Null:**	-601.61
Covariance Type:	nonrobust	**LLR p-value:**	5.267e-31

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.8128	0.448	6.282	0.000	1.935	3.690
Salary	-1.932e-05	4.94e-06	-3.911	0.000	-2.9e-05	-9.64e-06
Age	-0.0504	0.008	-5.962	0.000	-0.067	-0.034
No_young_children	-1.3023	0.169	-7.707	0.000	-1.633	-0.971
Foreign_yes	1.2092	0.183	6.592	0.000	0.850	1.569

Observation:

- Now all p values are less than 0.05. Hence all these attributes and their coeffficients have importance in deciding the target variable HolidayPackage.
- Also we can see that coef value is highest for No_young_children followed by foreign, Age and salary
- Salary coefficient value is very low i.e -00001932. So its impact is almost 0 on dependent variable
- 'p values' indicate that all the variable are significant at 95% confidence level
- The adjusted pseudo R-square value is 0.11641507873864654 . We notice that the Adjusted pseudo R-square value have increased

Conclusion:

Logistic regression equation is as shown below:-

(2.81) * Intercept + (-0.0) * Salary + (-0.05) * Age + (-1.3) * No_young_childre n + (1.21) * Foreign_yes +

Log (odd) = (2.81) + (-0.0) Salary + (-0.05) Age + (-1.3) No_young_children + (1.21) Foreign_yes

We can see that salary coefficient is very small, this it can be removed. So our equation would become:-

Log (odd) = (2.81) + (-0.05) Age + (-1.3) No_young_children + (1.21) * Foreign_yes

Most important attribute here is No of young children followed by Foreign and age

\

Prediction on Data

a) Boxplot of Holiday Package Actual vs Predicted

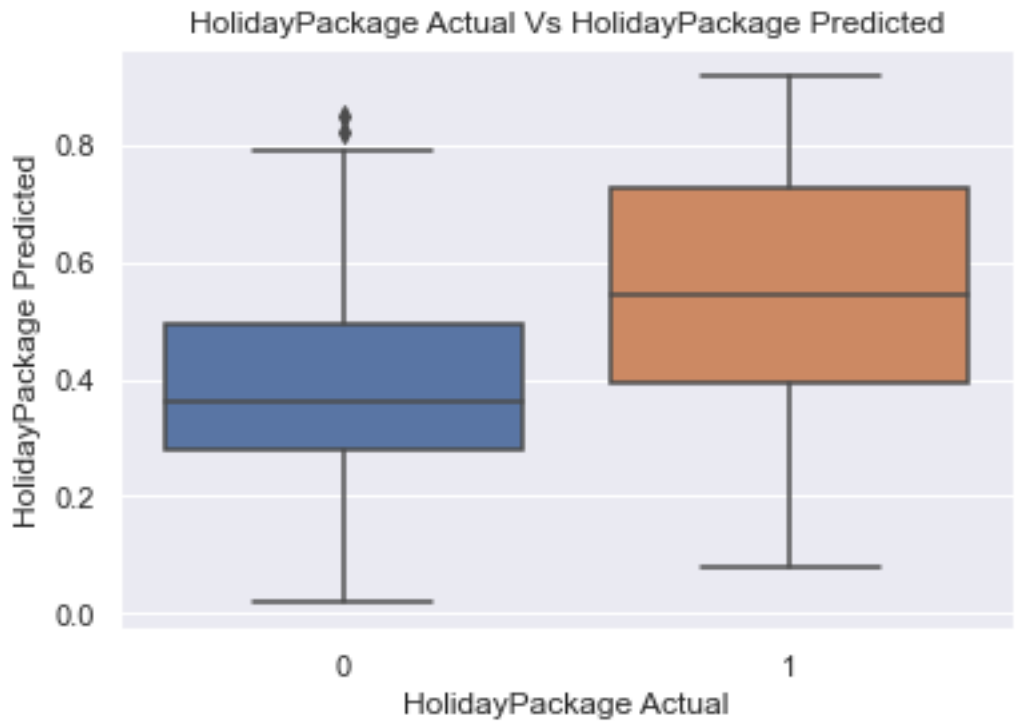


Figure 40: Boxplot of Holiday Package Actual vs Predicted

Observation :

From the above boxplot, we need to decide on one such value of a cut-off which gives most reasonable power of the model

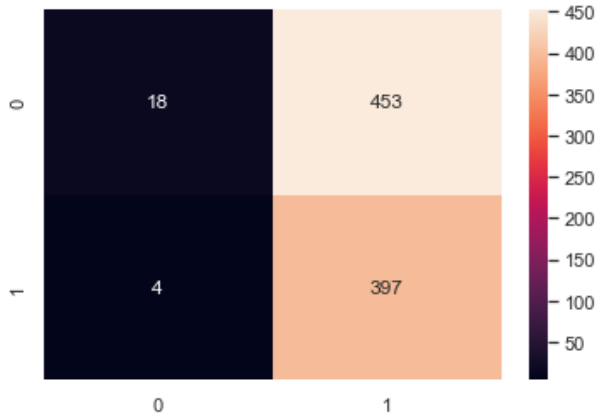
b) Different cut-off method for the predictions on the Probability Predictions Data

- **Cut – off at 0.1**

```
0.1

Accuracy Score 0.4759
F1 Score 0.6347

Confusion Matrix
```



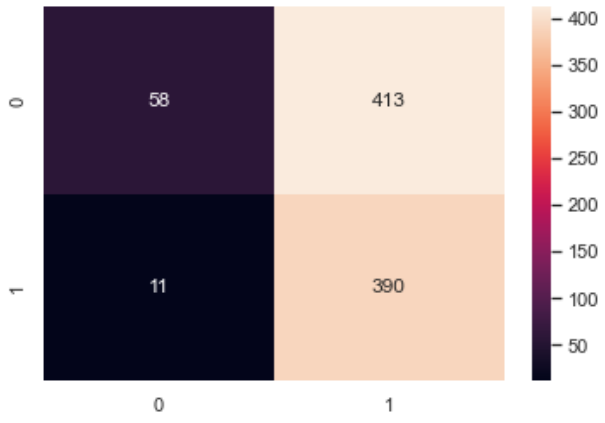

```
True Negative: 18
False Positives: 453
False Negatives: 4
True Positives: 397
```

- **Cut – off at 0.2**

```
0.2

Accuracy Score 0.5138
F1 Score 0.6478

Confusion Matrix
```



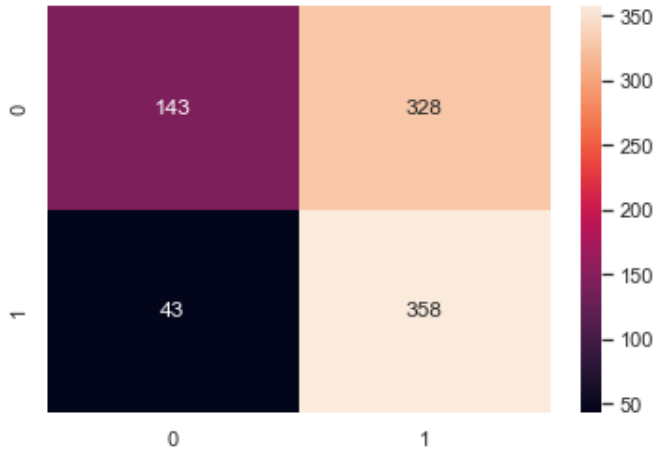

```
True Negative: 58
False Positives: 413
False Negatives: 11
True Positives: 390
```

- **Cut – off at 0.3**

```
0.3

Accuracy Score 0.5745
F1 Score 0.6587

Confusion Matrix
```



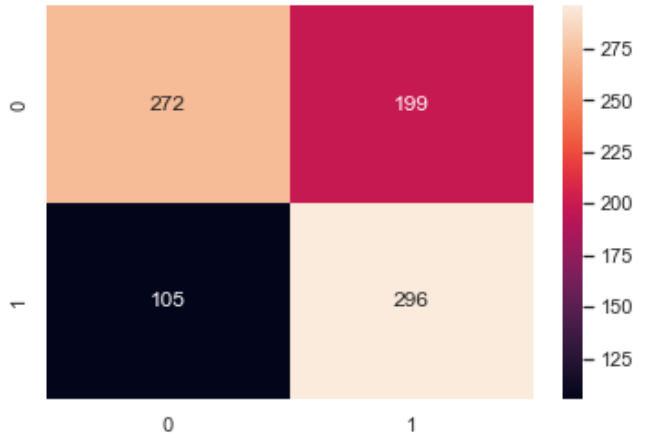

```
True Negative: 143
False Positives: 328
False Negatives: 43
True Positives: 358
```

- **Cut – off at 0.4**

```
0.4

Accuracy Score 0.6514
F1 Score 0.6607

Confusion Matrix
```



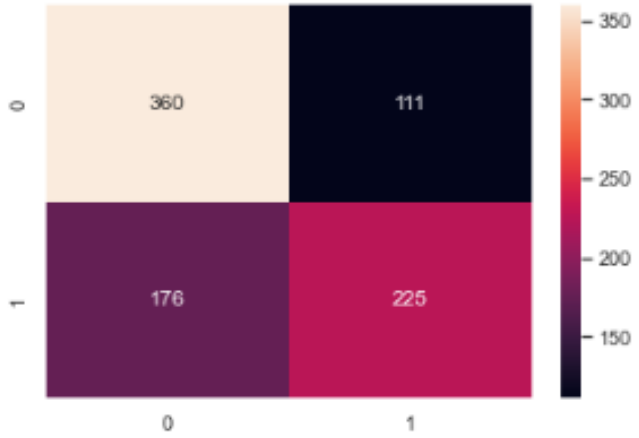

```
True Negative: 272
False Positives: 199
False Negatives: 105
True Positives: 296
```

- **Cut – off at 0.5**

```
0.5

Accuracy Score 0.6709
F1 Score 0.6106

Confusion Matrix
```



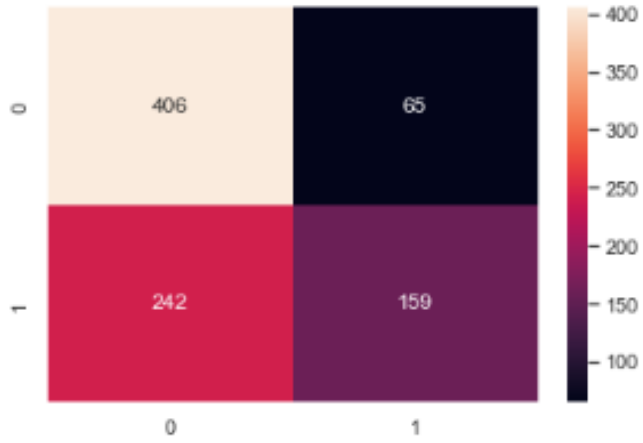

```
True Negative: 360
False Positives: 111
False Negatives: 176
True Positives: 225
```

- **Cut – off at 0.6**

```
0.6

Accuracy Score 0.6479
F1 Score 0.5088

Confusion Matrix
```



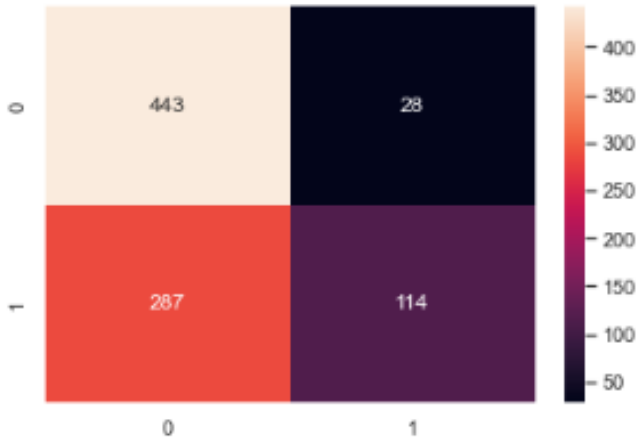

```
True Negative: 406
False Positives: 65
False Negatives: 242
True Positives: 159
```

- **Cut – off at 0.7**

```
0.7

Accuracy Score 0.6388
F1 Score 0.4199

Confusion Matrix
```



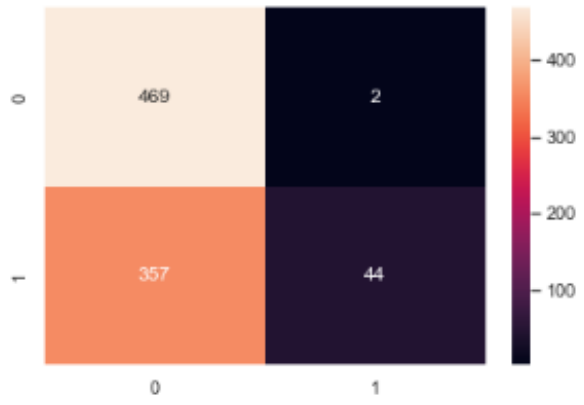

```
True Negative: 443
False Positives: 28
False Negatives: 287
True Positives: 114
```

- **Cut – off at 0.8**

```
0.8

Accuracy Score 0.5883
F1 Score 0.1969

Confusion Matrix
```



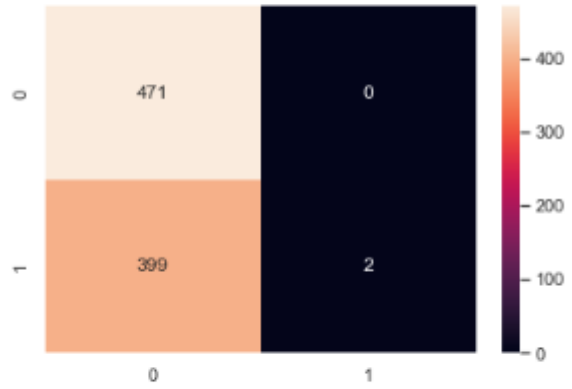

```
True Negative: 469
False Positives: 2
False Negatives: 357
True Positives: 44
```

- **Cut – off at 0.9**

```
0.9

Accuracy Score 0.5424
F1 Score 0.0099

Confusion Matrix
```




```
True Negative: 471
False Positives: 0
False Negatives: 399
True Positives: 2
```

Figure 41: Confusion Matrix Figures at different Cut-Off

Observation:

Cut off 0.05 looks the most promising as its accuracy and F1 score is most balanced

Accuracy Score 0.6709
F1 Score 0.6106

c) Classification Report on Data at cut-off 0.5

	precision	recall	f1-score	support
0	0.672	0.764	0.715	471
1	0.670	0.561	0.611	401
accuracy			0.671	872
macro avg	0.671	0.663	0.663	872
weighted avg	0.671	0.671	0.667	872

Table 16: Classification Report on Data at cut-off 0.5

a) Calculate the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve

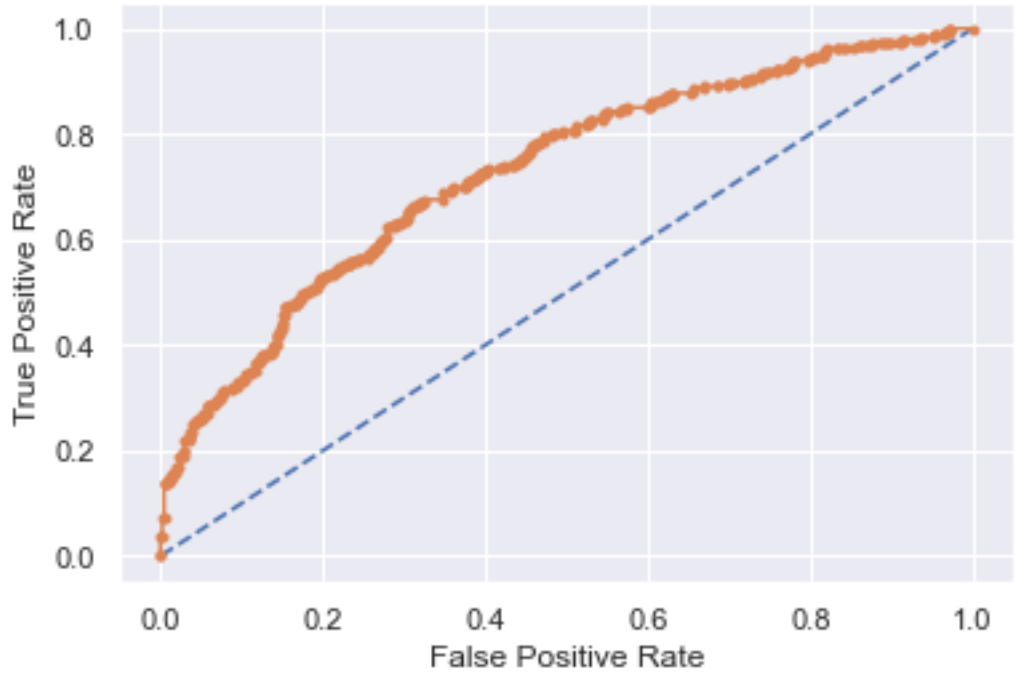


Figure 42: Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve

Observation :

- With accuracy of 67% and recall rate of 56%, model is only able to predict 56% of total tours which were actually claimed as claimed.
- Precision is 67% of data which means, out of total employees predicted by model as opt for tour, 67% employees actually opted for the tour
- F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 values is low.
- Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).
- If a employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if an employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. Its a loss of potential lead for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.
- As Recall rate of test dataset is very poor around 56% thus this doesn’t looks good enough for classification

3. **Alternatively, if prediction accuracy of employee opting for holiday package or not 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.**

If we only wanted to predict using Logistic Regression and was not looking for the model building aspect of it, we can do that as well. For this, we will use the same variables as of Model 2 and Model 3.

First, we will split the data into train and test at 70:30 ratio. We will build the model on the training data and check the model performance metrics on the test data.

We will take the default cut-off of 0.5 to get the class predictions.

a) Confusion Matrix check for the Models 2 and model 3 built Train and Test Data

Figure 43: Confusion Matrix check for the Models 2 and model 3 built Train and Test Data

```
Accuracy Score - Model 2
Training: 0.6688524590163935
Testing:  0.648854961832061

Accuracy Score - Model 3
Training: 0.6688524590163935
Testing:  0.6603053435114504
```

a) AUC and ROC for the Training and Testing data Model 2

Figure 44: AUC and ROC for the Training and Testing data Model 2

Observation:

Model 2 ROC curve have AUC train (0.660) and test (0.647) score.

b) Classification_report Model 2

	precision	recall	f1-score	support
0	0.667	0.772	0.715	329
1	0.672	0.548	0.604	281
accuracy			0.669	610
macro avg	0.670	0.660	0.660	610
weighted avg	0.669	0.669	0.664	610

Table 17 :Classification_report Model 2

Observation Model 2:

- With accuracy of 67% and recall rate of 54%, model is only able to predict 54% of total tours which were actually claimed as claimed.
- Precision is 67% of data which means, out of total employees predicted by model as opt for tour, 67% employees actually opted for the tour
- F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 values is low.
- Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).
- If a employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if am employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. Its a loss of potential lead for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.
- As Recall rate of test dataset is very poor around 54% thus this doesn’t look good enough for classification

c) AUC and ROC for the Training and Testing data Model 3

Figure 45: AUC and ROC for the Training and Testing data Model 3

Observation :

Model 3 ROC curve have better AUC train(0.660) and test(0.658) score compared to Model 2 ROC curve

d) Classification_report Model 3

	precision	recall	f1-score	support
0	0.667	0.772	0.715	329
1	0.672	0.548	0.604	281
accuracy			0.669	610
macro avg	0.670	0.660	0.660	610
weighted avg	0.669	0.669	0.664	610

Table 18: Classification_report Model 3

Observation Model 3:

- With accuracy of 67% and recall rate of 54%, model is only able to predict 54% of total tours which were actually claimed as claimed.
- Precision is 67% of data which means, out of total employees predicted by model as opt for tour, 67% employees actually opted for the tour
- F1-score is the harmonic mean of precision and recall, it takes into the effect of both the scores and this value is low if any of these 2 values is low.
- Since we are building a model to predict if whether employee will opt for tour or not, for practical purposes, we will be more interested in correctly classifying 1 (employees opting for tour) than 0(employees not opting for tour).
- If a employee not opting for tour is incorrectly predicted to be "opted for tour" by the model, then the impact on cost for the travel company would be bare minimum. But if an employee opted for tour is incorrectly predicted to be not opted by the model, then the cost impact would be very high for the tour and travel company. It's a loss of potential lead for the company. Hence recall rate (actual data point identified as True by model) is very important in this scenario.
- As Recall rate of test dataset is very poor around 54% thus this doesn't looks good enough for classification

Running other Classification models

Making 5 Classification models using:

1. Artificial Neural Network,
2. Decision Tree,
3. Random Forest,
4. Logistic Regression,
5. Linear Regression (Linear Discriminant Analysis)

a) Summary Of Train and Test Accuracy on the 5 Models (Before Pruning)

	Logistic Regression	LDA	Decision Tree	Random Forest	ANN
Train Accuracy	0.67	0.66	1.00	1.00	0.72
Test Accuracy	0.66	0.65	0.58	0.64	0.67
Train AUC	0.73	0.73	1.00	1.00	0.80
Test AUC	0.72	0.71	0.58	0.68	0.73
Train Recall	0.57	0.57	1.00	1.00	0.69
Test Recall	0.52	0.52	0.51	0.59	0.63
Train precision	0.66	0.65	1.00	1.00	0.70
Test precision	0.67	0.65	0.55	0.61	0.64
Train f1	0.61	0.61	1.00	1.00	0.69
Test f1	0.58	0.57	0.53	0.60	0.64

Table 19: Summary Of Train and Test Accuracy on the 5 Models (Before Pruning)

b) GridSearchCV for Models

- **Logistic Regression**

```
 Showing best parameters for the grid search

{'penalty': 'l1', 'solver': 'liblinear', 'tol': 1e-05}

 Classification Report for Train dataset

              precision    recall  f1-score   support

           0       0.67      0.74      0.71       329
           1       0.66      0.57      0.61       281

    accuracy                           0.66       610
   macro avg       0.66      0.66      0.66       610
weighted avg       0.66      0.66      0.66       610


 Classification Report for Test dataset

              precision    recall  f1-score   support

           0       0.65      0.77      0.71       142
           1       0.66      0.52      0.58       120

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

- **Decision Tree**

Showing best parameters for the grid search

{'max_depth': 4, 'min_samples_leaf': 4, 'min_samples_split': 67}

Classification Report for Train dataset

	precision	recall	f1-score	support
0	0.70	0.82	0.76	329
1	0.74	0.59	0.66	281
accuracy			0.72	610
macro avg	0.72	0.71	0.71	610
weighted avg	0.72	0.72	0.71	610

Classification Report for Test dataset

	precision	recall	f1-score	support
0	0.64	0.75	0.69	142
1	0.63	0.51	0.56	120
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.63	262

➢ Random Forest

```
Running grid search

Showing best parameters for the grid search

{'max_depth': 4, 'max_features': 3, 'min_samples_leaf': 8, 'min_samples_split': 56, 'n_estimators': 490}

Classification Report for Train dataset

              precision    recall  f1-score   support

           0       0.72      0.81      0.76       329
           1       0.74      0.63      0.68       281

    accuracy                           0.73       610
   macro avg       0.73      0.72      0.72       610
weighted avg       0.73      0.73      0.73       610


Classification Report for Test dataset

              precision    recall  f1-score   support

           0       0.68      0.73      0.70       142
           1       0.65      0.58      0.61       120

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

➢ ANN (Artificial Neural Network)

```
Showing best parameters for the grid search

{'activation': 'relu', 'hidden_layer_sizes': (100, 100), 'max_iter': 10000, 'solver': 'adam', 'tol': 0.001, 'verbose': True}

Classification Report for Train dataset

              precision    recall  f1-score   support

           0       0.74      0.81      0.77       329
           1       0.75      0.67      0.71       281

    accuracy                           0.74       610
   macro avg       0.74      0.74      0.74       610
weighted avg       0.74      0.74      0.74       610


Classification Report for Test dataset

              precision    recall  f1-score   support

           0       0.69      0.73      0.71       142
           1       0.65      0.62      0.64       120

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

c) ROC and AUC curve of all the models Training Data and Testing Data

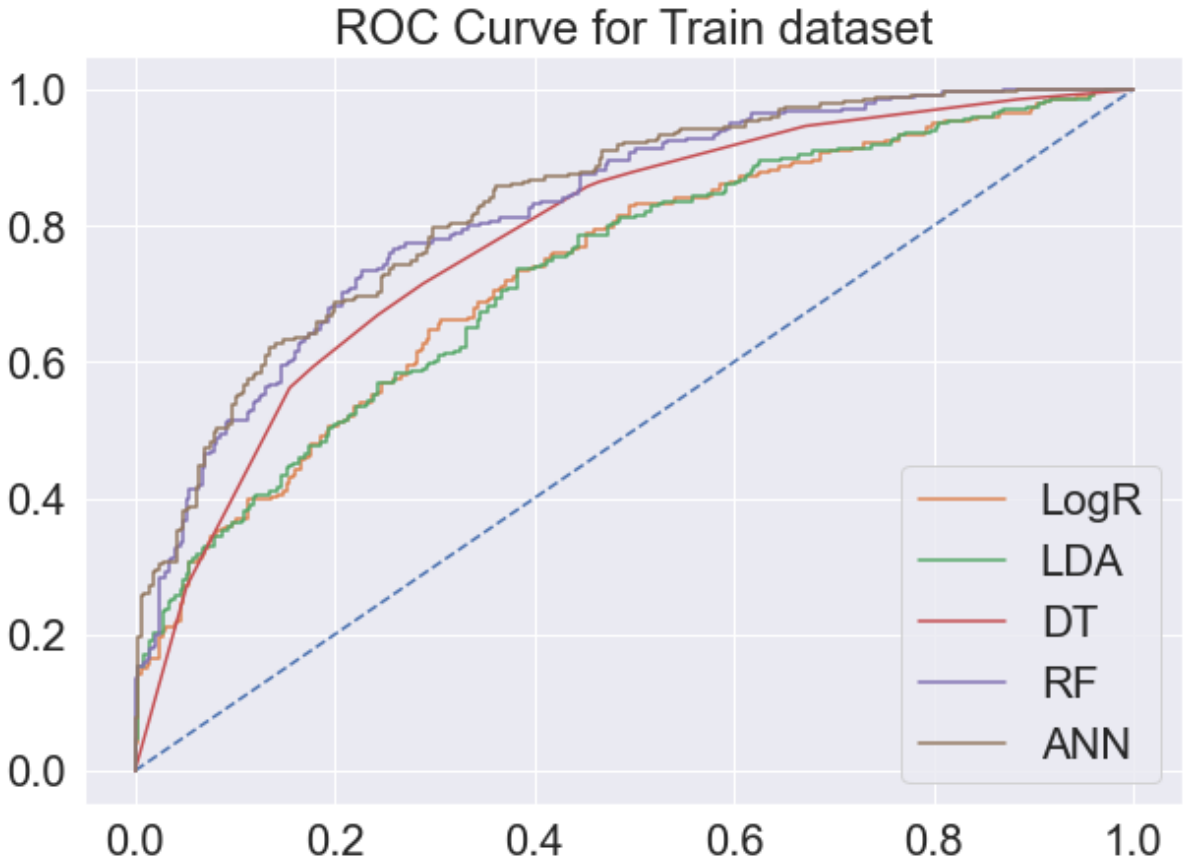


Figure 46: ROC and AUC curve of all the models Training Data

Observations:

AUC for LogR is: 0.73
AUC for LDA is: 0.73
AUC for DT is: 0.78
AUC for RF is: 0.82
AUC for ANN is: 0.83

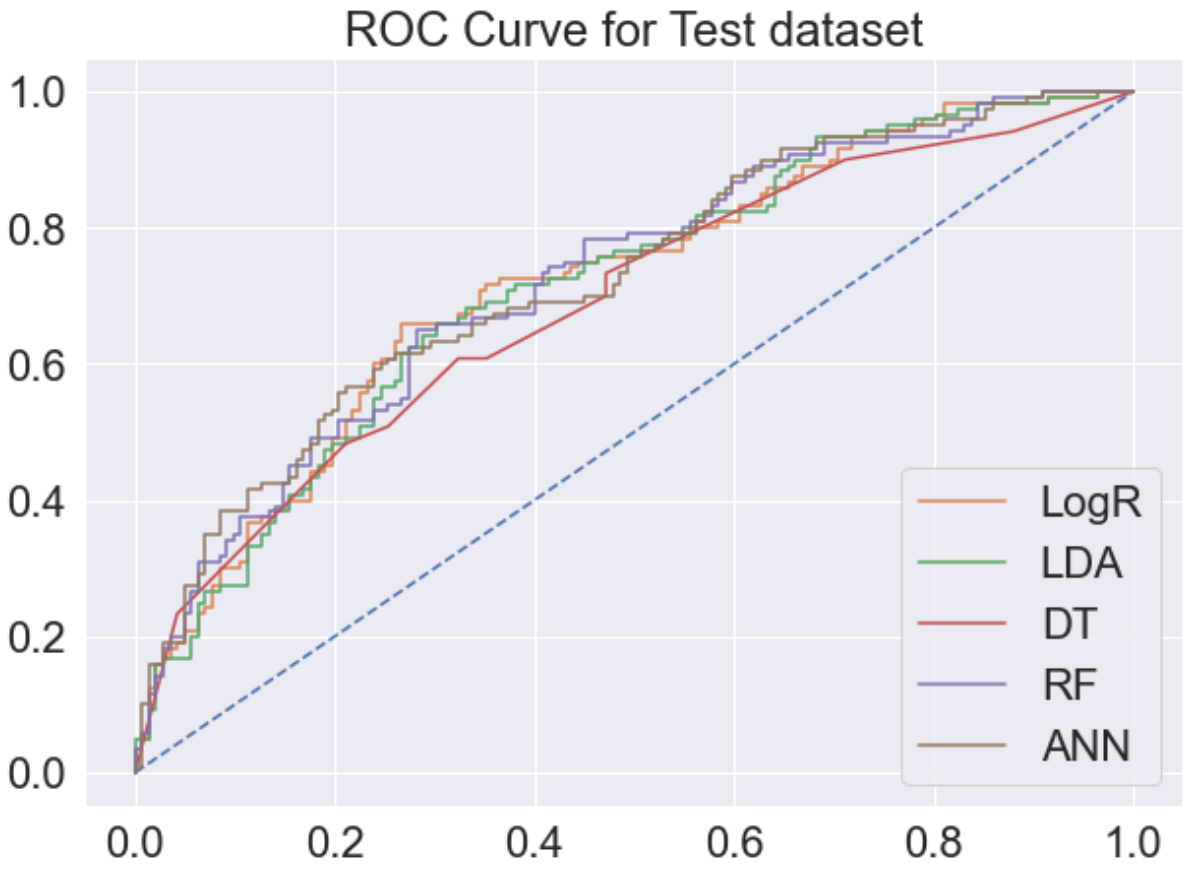


Figure 47: ROC and AUC curve of all the models Testing Data

Observations:

AUC for LogR is: 0.72
AUC for LDA is: 0.71
AUC for DT is: 0.69
AUC for RF is: 0.72
AUC for ANN is: 0.73

d) Summary Of Train and Test Accuracy on the 5 Models (After Pruning)

	Logistic Regression	LDA	Decision Tree	Random Forest	ANN
Train Accuracy	0.67	0.66	0.72	0.74	0.74
Test Accuracy	0.66	0.65	0.64	0.67	0.68
Train AUC	0.73	0.73	0.78	0.82	0.83
Test AUC	0.72	0.71	0.69	0.72	0.73
Train Recall	0.57	0.57	0.59	0.66	0.67
Test Recall	0.52	0.52	0.51	0.60	0.62
Train precision	0.66	0.65	0.74	0.75	0.75
Test precision	0.67	0.65	0.63	0.65	0.65
Train f1	0.61	0.61	0.66	0.70	0.71
Test f1	0.58	0.57	0.56	0.62	0.64

Table 20: Summary Of Train and Test Accuracy on the 5 Models (After Pruning)

Observation:

- On comparing all the models, it looks like that no model is over-fitting/under fitting.
- All models test and train score are comparable and within 5-6% range.
- We can see that all models are giving similar results with not much of difference in accuracy.
- Random Forest and Artifical Neural Network gives better f1 score and better recall rate as compared to the logistics/LDA
- Among all these models we will go for Artificial Neuro Network MLP classifier as its test f1 score and test accuracy is the highest.

4. Basis on these predictions, what are the insights and recommendations.

We have run five different models (Logistic Regression/ Linear Discriminant Analysis/ CART/ Random Forest/ Artificial Neuro Network) for predicting whether an employee is opting for holiday package or not

Based on the reports and analysis done it was found that all models were not good enough for classification as accuracy coming out is 66%.

So our recommendation to the business is as shown below:-

- In order to further improve the predictive model results for finding the employees which will opt for tour in future more accurately, more data sample is required for analysis. The greatest number of people who are opting in for the package has a salary of range between 30 to 40 k. It suggests that the package is of average price with medium level facilities

- Current model is useful to predict when tours are not getting claimed with more than 66% accuracy.

- Most important attribute here is No of young children of an employee followed by Foreigner column and lastly the age.

- As seen in EDA, 68% of foreign employees are opting for the tour packages. So the travel company should make dedicated tours for these foreigners keeping in mind which places/areas that these foreigners would like to travel. If these customers are satisfied then when they will travel back to their country they will refer more of their friends/family members for tours. This way company can retain and increase its customer base.

- Currently 24% of employees have 1 or more young child. It was found during EDA that out of these employees 70% are not opting for the tours. So the travel company should make dedicated tour for the employees who have young child and provide them with some extra child care benefits (like play area for child, child food, medical facilities etc.) so as to lure these employees. So, if they add some additional luxury packages with facilities like booking in star hotels, luxury cars etc. it may help to increase the sales of packages to a higher income group.

- The analysis shows that a greater number of foreigners opt in for packages than the non-foreigners. This along with the previous analysis which shows that most of the people are from salary group of 30 to 50k (so it is not expensive package) suggest that packages provided are either of local sightseeing place or of less interest to the non-foreigners. So, suggest the company to add some more activities or places in their packages

- Old age employees (age greater than 50) opt less for the tours. So company can provide dedicated tour plans for old aged senior employees. To improve holiday packages over the age above 50 we can provide religious destination places.

- As per the analysis, Salary of the employees is not an important attribute in deciding that whether employee will opt for tour or not. So the travel company should not focus on salary of the employee. May be salary can decide which tour (economical or lavish) that particular employee will be interested in but he/she will opt for some tour irrespective of his/her salary

- The analysis shows that data if the employee as no young children, there is more Chance of them taking up the package. As count of children increases, the willingness to opt in for a holiday package decreases. So, I suggest the company to provide additional discounts or children attractiveness for the employee who has young children to boost up the chance of them opting in for the package.