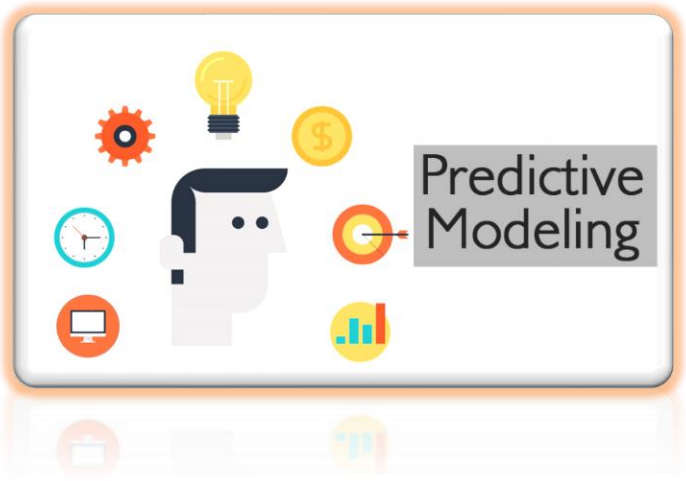


PREDICTIVE MODELING

BUSINESS REPORT

SUMMARY ABOUT TWO DIFFERENT DATAS:

Gem Stones co ltd, which is a cubic zirconia manufacturer provides a dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. The company wants to know the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share.

A tour and travel agency which deals in selling holiday packages. The agency provides details of 872 employees of a company. Among these employees, some opted for the package and some didn't. The agency wants to know whether an employee will opt for the package or not on the basis of the information given in the data set. The agency also wants to know the important factors on the basis of which the company will focus on particular employees to sell their packages.

Yashveer Kothari. A
POST GRADUATE PROGRAM IN DATA SCIENCE AND BUSINESS ANALYTICS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES		
FIGURE#	NAME	PAGE#
1	HISTOGRAM/ DISTRIBUTION PLOT	15
2	CHECKING FOR OUTLIERS	16
3	OUTLIER TREATMENT USING IQR	18
4	BAR GRAPH FOR CUT VARIABLE	19
5	BAR GRAPH FOR COLOUR VARIABLE	20
6	BAR GRAPH FOR CLARITY VARIABLE	21
7	PAIRPLOT	22
8	HEATMAP/ CORRELATION PLOT	23
9	SCATTER PLOT AFTER SCALING	32
10	SCATTER PLOT FOR UNSCALED DATA	36
11	NORMAL DISTRIBUTION	42
12	IDENTIFYING OUTLIERS USING BOXPLOT	42
13	OUTLIER TREATMENT USING IQR	43
14	PAIRPLOT	44
15	HEATMAP/ CORRELATION PLOT	45
16	BAR GRAPH FOR HOLIDAY PACKAGE VARIABLE	46
17	BAR GRAGH FOR FOREIGN VARIABLE	46
18	BOXPLOT - HOLIDAY PACKAGE AND SALARY	47
19	BOXPLOT - HOLIDAY PACKAGE AND EDUC	47
20	BOXPLOT - HOLIDAY PACKAGE AND NO YOUNG CHILDREN	48
21	BOXPLOT - HOLIDAY PACKAGE AND NO OLDER CHILDREN	48
22	BOXPLOT - HOLIDAY PACKAGE AND AGE	49
23	BAR GRAPH- HOLIDAY PACKAGE AND FOREIGN	49
24	BAR GRAPH - HOLIDAY PACKAGE AND EDUC	50
25	BAR GRAPH - HOLIDAY PACKAGE& NO OF YOUNG CHILDREN	50
26	BAR GRAPH - HOLIDAY PACKAGE& NO OF OLDER CHILDREN	51
27	BAR GRAPH - HOLIDAY PACKAGE& AGE	51
28	SPLITTING INTO TRAIN AND TEST DATA	55
29	CONFUSION MATRIX OF TRAINING DATA	57
30	CONFUSION MATRIX OF TEST DATA	58
31	AUC CURVE FOR TRAIN AND TEST DATA	60
32	CONFUSION MATRIX LDA- TRAIN AND TEST DATA	62
33	AUC CURVE LDA - TRAIN AND TEST DATA	64
34	MAXIMUM ACCURACY TEST - TRAIN DATA	65
35	MAXIMUM ACCURACY TEST - TEST DATA	69

LIST OF OUTPUTS	
NAME	PAGE#
CHECKING FOR DUPLICATES	13
CHECKING FOR DUPLICATES	28
ENCODING THE DATA	29
DIMENSIONS OF TRAIN AND TEST DATA	56
APPLYING LOGISTIC REGRESSION WITH GRID SEARCH CV	56
EXTRACTING THE BEST PARAMETERS	56
TRAIN DATA CLASS PREDICTIONS	61
TEST DATA CLASS PREDICTIONS	62

ABOUT PREDICTIVE MODELING

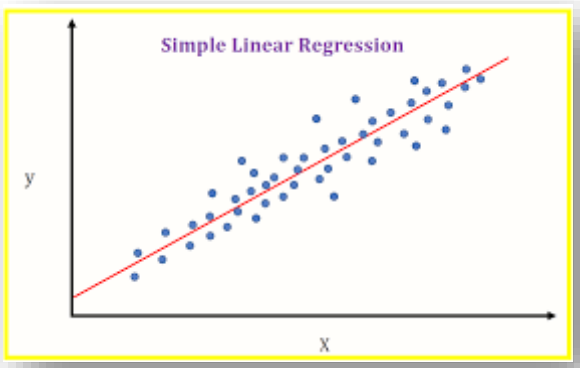
- Predictive modelling is a mathematical process used to predict future events or outcomes by analysing patterns in a given set of input data.

ABOUT LINEAR REGRESSION

- It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behaviour and trends.

- The term "regression" generally refers to predicting a real number. The term "linear" in the name "linear regression" refers to the fact that the method models data with linear combination of the explanatory variables.
- Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent and independent variable.
- A linear combination is an expression where one or more variables are scaled by a constant factor and added together.
- The linear combination used in linear regression can be expressed as:

  Dependent variable value = (weight ∗independent variable) + constant
- Simple Linear Regression figure:



- Correlation between two variables indicates how closely their relationship follow a straight line.
- Correlation of extreme possible values of -1 & +1 indicate a perfectly linear relationship between x and y. Whereas the correlation of 0 indicates absence of linear

- **Pearson's Coefficient:**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

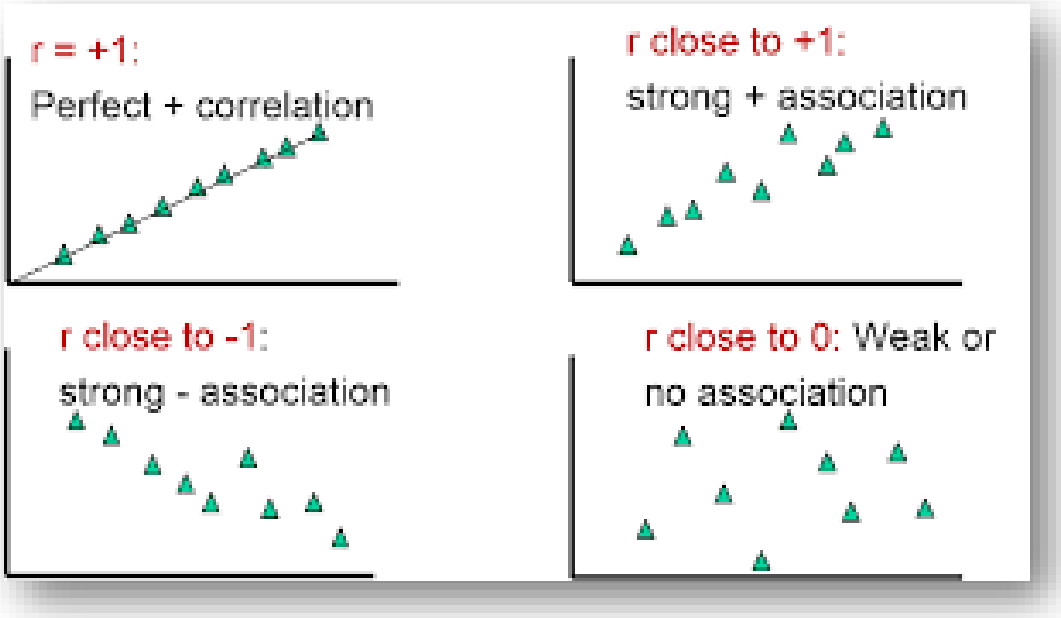
r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable



When,

1. r is close to 0: There is no correlation and no linear relation
2. r is near -1: The pattern goes near -1 in the graph, it is inversely related and has negative correlation.

3. r is near +1: It is directly related and positively correlated; the points are pointing towards +1 in the graph.

- **Best Fit Line:** The Best Fit line is the line which captures the relationship between x and y most accurately.
- **Total sum of squares:** The sum of squares total, denoted TSS, is the squared differences between the observed dependent variable and its mean.

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- **Least of sum of squared errors**: The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.
- **Coefficients of determinant:** Determines the fitness of a linear model. Represented by R^2. The closer the point gets to the line the coefficient of determinant tends to be a better model i.e., =1.
- **Assumptions of Linear Regression:**
    1. There is Low Multicollinearity.
    2. Variables should follow a normal distribution.
    3. Sensitive to outliers and extreme values.
    4. Relationships are linear.

ABOUT LOGISTIC REGRESSION

1. Also known as Logit, Maximum-Entropy classifier, is a supervised learning method for classification. It establishes relation between dependent class variable and independent variables using regression
2. The dependent variable is categorical i.e., it can take only integral values representing different classes
3. The probabilities describing the possible outcomes of a query point are modelled using a logistic function
4. Belongs to family of discriminative classifiers. They rely on attributes which discriminate the classes well.

5. There are two broad categories of Logistic Regression algorithms.

5.1. Binary Logistic Regression: When dependent variable is strictly binary

5.2. Multinomial Logistic Regression: When the dependent variable has multiple categories.

There are two types of Multinomial Logistic Regression.

I. Ordered Multinomial Logistic Regression (dependent variable has ordered values)

II. Nominal Multinomial Logistic Regression (dependent variable has unordered categories).

6. The output is the probability of belonging to a class. Probability can also be expressed in from of odds.
7. Odds have a property of ranging from 0 to infinity that makes it easy to map a regression equation to odds. That is why logistic model uses odds
8. The odds of belonging to class y = 1 is defined as the ratio of probability of belonging to class 1 to probability of belonging to class 0.

$$\text{Odds}(Y = 1) = \frac{p}{1-p}.$$

- **Assumptions of Logistic Regression:**

1. Dependent variable is categorical. Dichotomous for binary logistic regression and multi label for multi-class classification
2. Attributes and log odds i.e., log (p / 1-p) should be linearly related to the independent variables
3. Attributes are independent of each other (low or no multi-collinearity)
4. In binary logistic regression class of interest is coded with 1 and other class 0
5. In multi-class classification using Multinomial Logistic Regression or OVR scheme, class of interest is coded 1 and rest 0.

ABOUT LINEAR DISCRIMINANT ANALYSIS

1. Linear Discriminant Analysis (LDA): LDA uses linear combination of independent variables to predict the class in the response variable of a given observation.

2. LDA assumes:
    a. Independent variables are normally distributed
    b. There is equal variance /covariance for the classes
3. LDA can be used to classify and reduce the dimensionality.
4. Once the assumptions are satisfied, LDA created a Linear Decision Boundary. Though LDA performs well it violates the assumption.
5. LDA is based upon the concept of searching for a linear combination of predictor variables that best separates the classes of the target variable.
6. Quadratic Discriminant Analysis (QDA): It is a variant of LDA. It uses quadratic combinations of independent variables to predict the class in the responsible variance of a given observation.
7. LDA model:

$$DS = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \ldots + \beta kXk$$

Where:

DS = Discriminant Score

β's = Discriminant weight (coefficients)

X's = Explanatory (Predictor or independent) variables

8. LDA is used when classes are well separated, data is small and has more than two classes.

PROBLEM 1

INTRODUCTION

The dataset contains data about prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company wants us to predict the price of the stone on the bases of the details given in the dataset. The main aim of this is perform an exploratory data analysis, clean the data, do univariate, bivariate and multivariate analysis and later use linear regression models and provide insights on the best way to improvise profits.

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 worst and J the best.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best 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.

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

TABLE 1: TOP 5 SAMPLES

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

TABLE 2: DROPPING COLUMN “UNNAMED:0”

	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

TABLE 3: DATASET INFORMATION

```
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)
```

TABLE 4: CHECKING FOR MISSING VALUES

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

TABLE 5: DATASET DESCRIPTION

	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

OUTPUT: CHECKING FOR DUPLICATES

- Upon checking for duplicates we get the following output:

```
There are a total of 34 dupicate values in the dataset
```

- There are a total of 34 duplicate values in the dataset as per the output above

TABLE 6: EXTRACTING THE UNIQUE VALUES

```
CUT :  5
Fair           781
Good          2441
Very Good     6030
Premium       6899
Ideal        10816
Name: cut, dtype: int64
```

```
COLOR :  7
J     1443
I     2771
D     3344
H     4102
F     4729
E     4917
G     5661
Name: color, dtype: int64
```

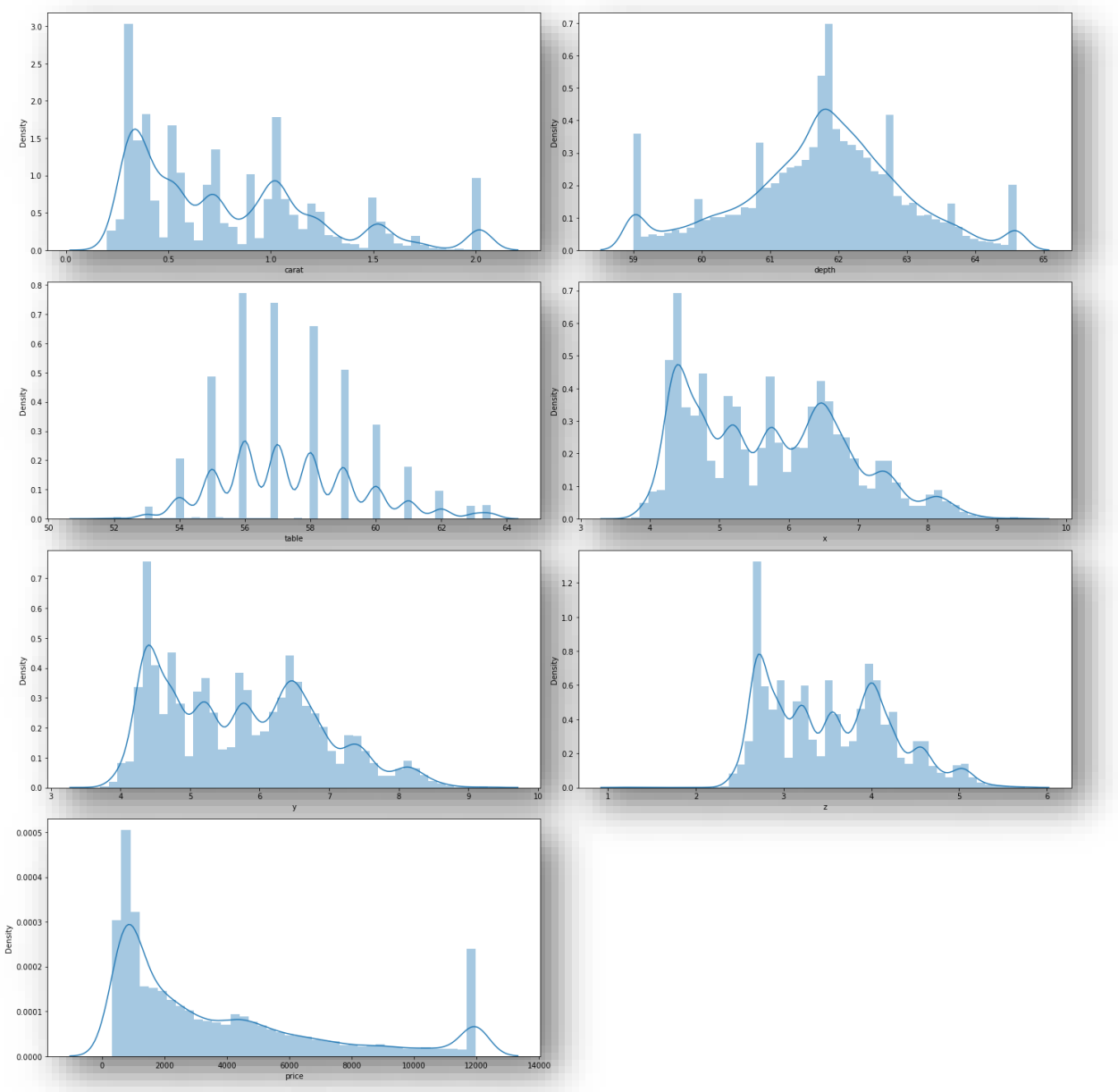
```
CLARITY :  8
I1        365
IF        894
VVS1     1839
VVS2     2531
VS1      4093
SI2      4575
VS2      6099
SI1      6571
Name: clarity, dtype: int64
```

TABLE 7: IMPUTING THE MISSING VALUES

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

We impute the missing values with the mean values of the particular columns, making the dataset free from missing values.

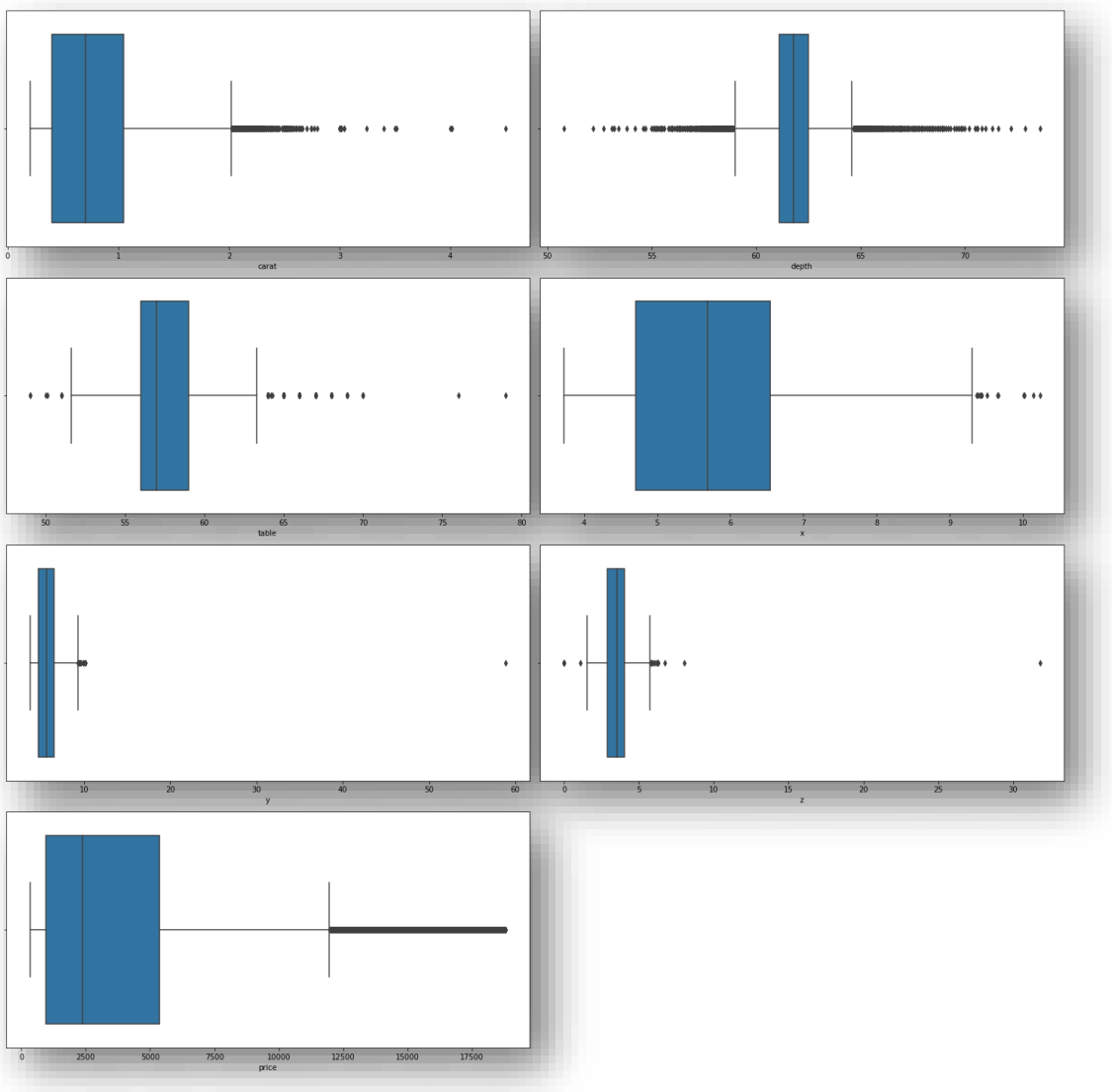
FIGURE 1: HISTOGRAM/DISTRIBUTION PLOT



- **INFERENCE FOR FIGURE 1:**

1. Depth is the only variable which can be considered as normally distributed.
2. Carat, Table, x, y, z has multiple modes as per the spread of the data.
3. Since price is being considered as the target or dependent variable, it is also right skewed

FIGURE 2: CHECKING FOR OUTLIERS



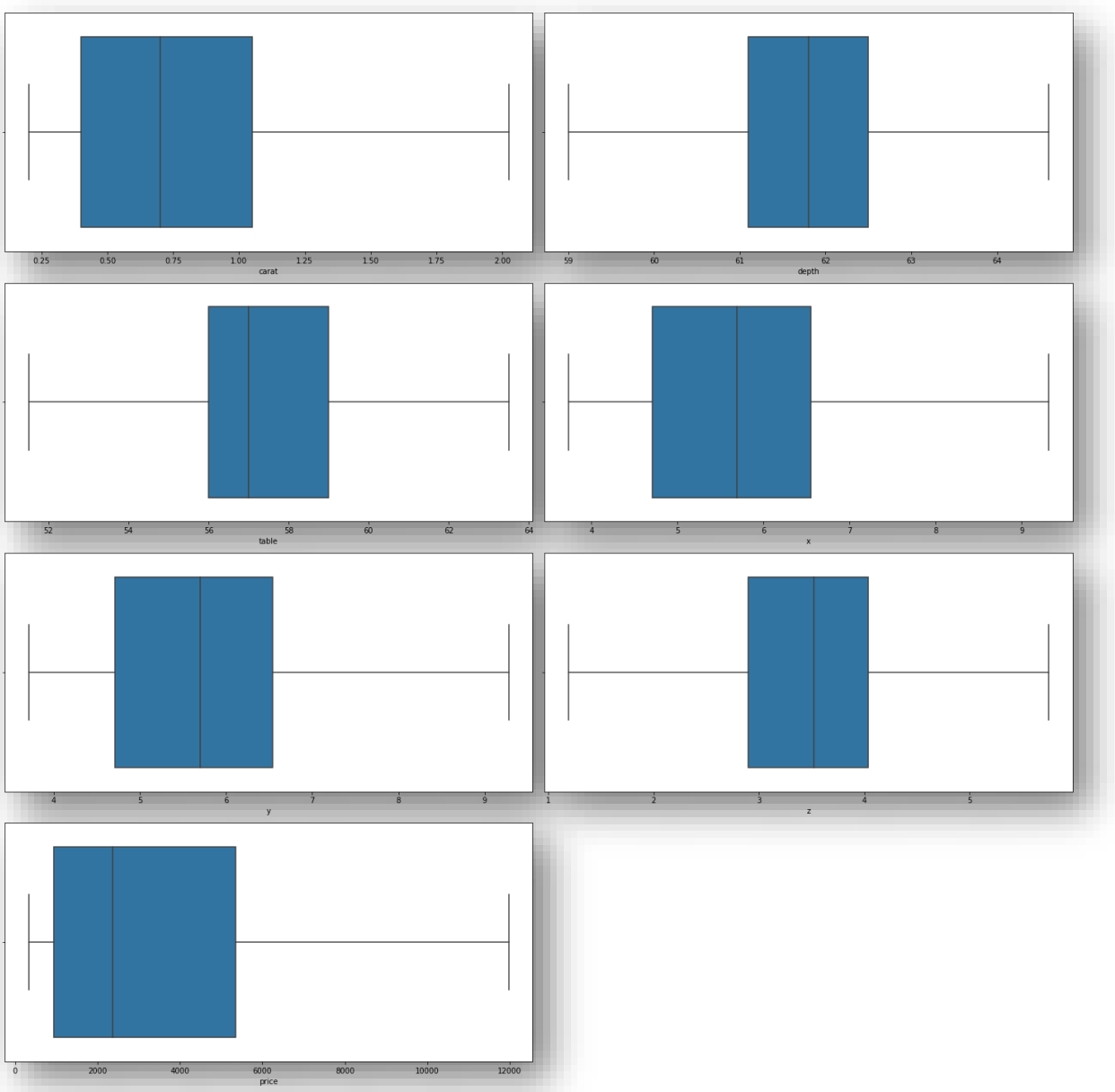

```
Shape before Outliers Treatment (26931, 10)
```

- **INFERENCE FOR FIGURE 2:**

1. As per the above figure, we can infer that there are outliers in all the variables.
2. From above data it is seen that except for carat and price variable, all other variables have mean and median values very close to each other, seems like there is no skewness in these variables.

3. Carat and price we see some difference in value of mean and median, which slightly indicates existence of some skewness in the data.

FIGURE 3: OUTLIER TREATMENT USING IQR

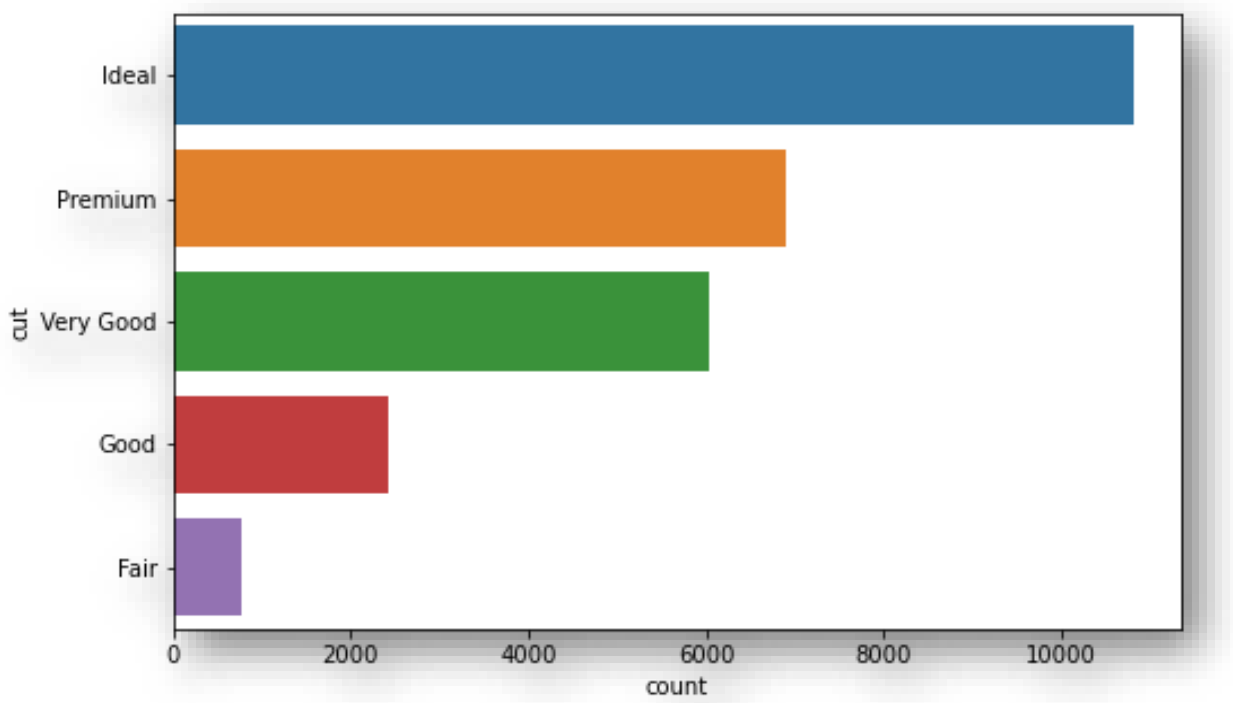



```
Shape after Outliers Treatment (26967, 10)
```

- **INFERENCE FOR FIGURE 3:**

After doing the outlier treatment using Inter quartile range, we eliminate the outliers. We can also infer that the shape of the dataset has been increased from 26931 to 26967 rows.

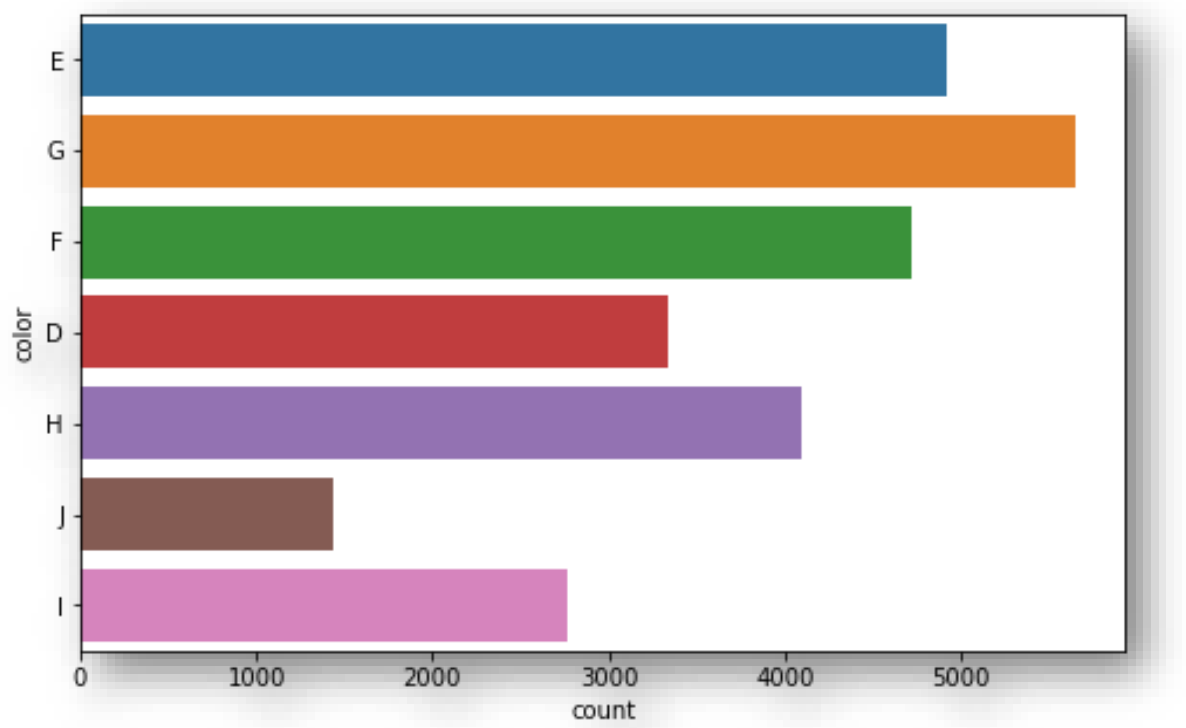
FIGURE 4: BAR GRAPH FOR CUT VARIABLE



- **INFERENCE FOR FIGURE 4:**

1. In the cut variable we see that the ideal cut is sold the highest and the fair cut is sold the least comparatively.
2. Whereas, the Premium and Very good sold at a good range and both are quiet sold at the same range.

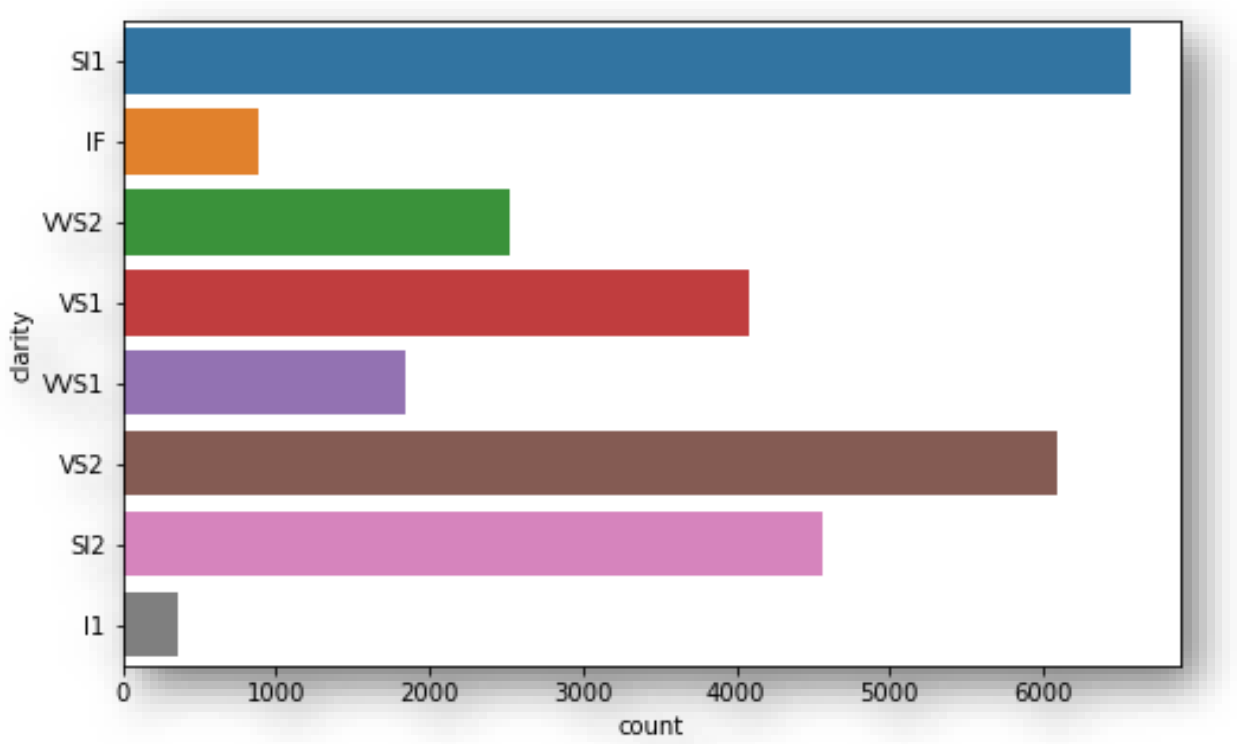
FIGURE 5: BAR GRAPH FOR COLOUR VARIABLE



- **INFERENCE FOR FIGURE 5:**

1. The colour G coloured gem is sold higher among all the other colours and the least is J coloured gem.
2. The reason for least sale of J and I can be that price of the coloured gems can be higher than the others.

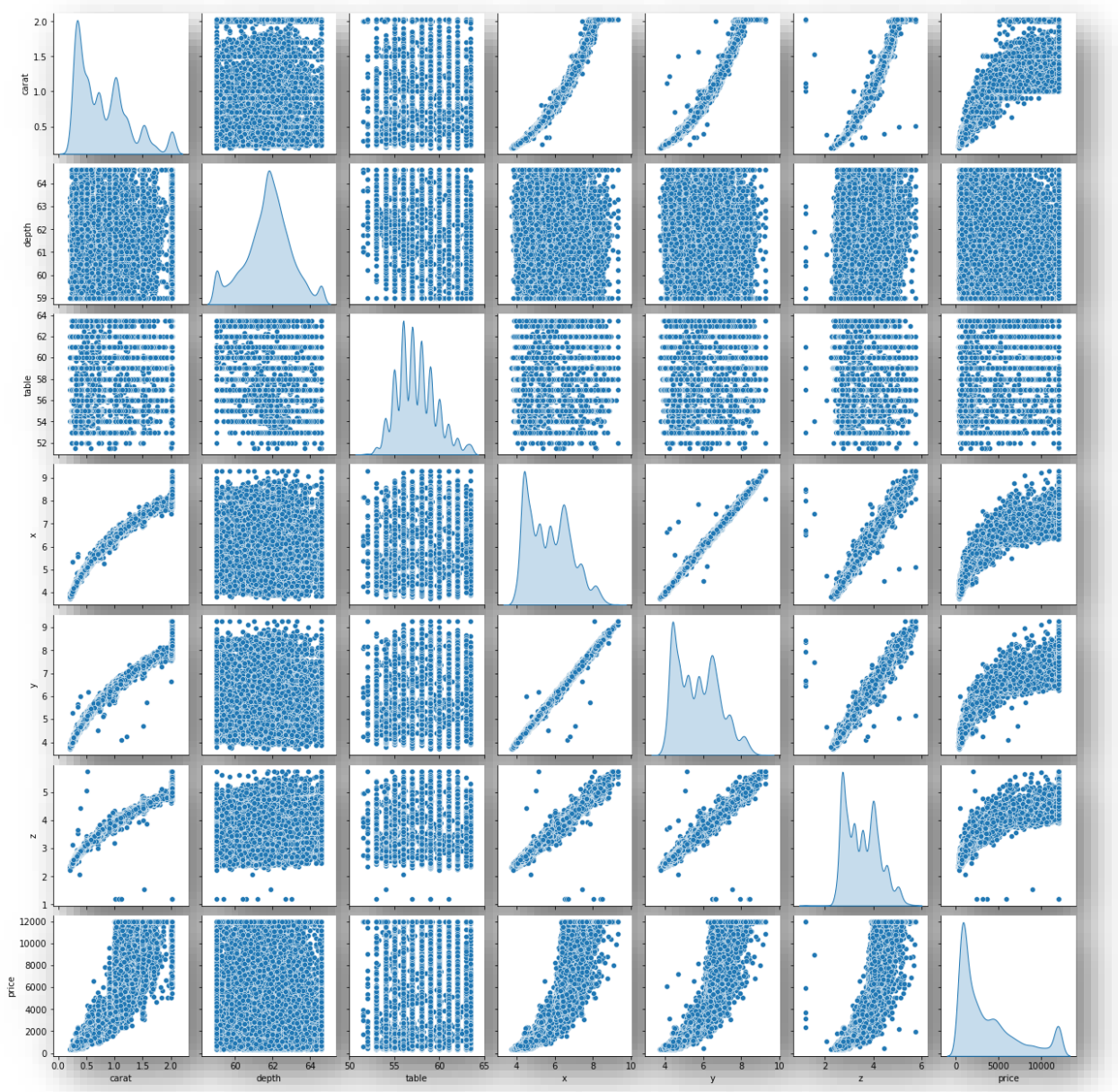
FIGURE 6: BAR GRAPH FOR CLARITY VARIABLE



- **INFERENCE FOR FIGURE 6:**

1. For the clarity variable we see the most sold is SI1 clarity gems and least is I1 clarity gems.
2. Slightly less priced seems to be SI1 type; VS2 and SI2 clarity stones seems to be more expensive.

FIGURE 7: PAIRPLOT



- From the above figure, there is no proper inference which can be made; therefore, we need to plot the correlation plot or Heatmap for clarity.

TABLE 8: CORRELATION TABLE

	carat	depth	table	x	y	z	price
carat	1.000000	0.029479	0.187107	0.982880	0.981961	0.978002	0.936763
depth	0.029479	1.000000	-0.289312	-0.019929	-0.022965	0.095242	-0.000979
table	0.187107	-0.289312	1.000000	0.199678	0.194043	0.159976	0.137849
x	0.982880	-0.019929	0.199678	1.000000	0.998490	0.988158	0.913373
y	0.981961	-0.022965	0.194043	0.998490	1.000000	0.987830	0.914805
z	0.978002	0.095242	0.159976	0.988158	0.987830	1.000000	0.906306
price	0.936763	-0.000979	0.137849	0.913373	0.914805	0.906306	1.000000

FIGURE 8: HEATMAP/CORRELATION PLOT

- **INFERENCE FOR TABLE 8 AND FIGURE 8:**

1. We see strong correlation between Carat, x, y, z and price that are demonstrating strong correlation or multicollinearity.
2. Less correlation between table with the other features.
3. Depth is negatively correlated with most the other features except for carat

CONCLUSION FOR Q.1

1. From table 3 we can infer that the dimension of the dataset is 26967 rows and 10 columns.
2. From table 4 we infer that there are 697 missing records in the depth variable.
3. From table 5 we can infer the following
    a. **Carat:** This is an independent variable, and it ranges from 0.2 to 4.5. mean value is around 0.8 and 75% of the stones are of 1.05 carat value. Standard deviation is around 0.477 which shows that the data is skewed and has a right tailed curve. Which means that majority of the stones are of lower carat. There are very few stones above 1.05 carat.
    b. **Depth:** The percentage height of cubic zirconia stones is in the range of 50.80 to 73.60. Average height of the stones is 61.80 25% of the stones are 61 and 75% of the stones are 62.5. Standard deviation of the height of the stones is 1.4. Standard deviation is indicating a normal distribution.
    c. **Table**: The percentage width of cubic Zirconia is in the range of 49 to 79. Average is around 57. 25% of stones are below 56 and 75% of the stones have a width of less than 59. Standard deviation is 2.24. Thus, the data does not show normal distribution and is similar to carat with most of the stones having less width also this shows outliers are present in the variable.
    d. **Price**: Price is the Predicted variable. Prices are in the range of 3938 to 18818. Median price of stones is 2375, while 25% of the stones are priced below 945. 75% of the stones are in the price range of 5356. Standard deviation of the price is 4022. Indicating prices of majority of the stones are in lower range as the distribution is right skewed.

4. Price variable gives continuous output with the price of the cubic zirconia stones. This is the Target Variable.
5. Carat, depth, table, x, y, z variables are numerical or continuous variables. (Table 3)
6. Cut, Clarity and colour are categorical variables. (Table 3).

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

- To find if there are any null values in the dataset we need to find if there are any missing or data which is not relevant to the dataset. Therefore, we perform value counts for all the suitable variables. (The missing value are identified and imputed in table 6 and table 7).
- There are a total of 697 missing records in depth variable and the same is being imputed using the mean value making the dataset free from missing values.

TABLE 9: CHECKING FOR VALUES EQUAL TO 0

1. **Carat:**

carat	cut	color	clarity	depth	table	x	y	z	price

- There is no value/record equal to zero in “carat” column.

2. **Depth:**

carat	cut	color	clarity	depth	table	x	y	z	price

- There is no value/record equal to zero in “Depth” column.

3. **Price:**

carat	cut	color	clarity	depth	table	x	y	z	price

- There is no value/record equal to zero in “Price” column.

4. **Column X:**

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
6215	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

- There are a total of 3 values/ records equalling to zero.

5. **Column Y:**

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
6215	0.71	Good	F	SI2	64.1	60.0	0.0	0.0	0.0	2130
17506	1.14	Fair	G	VS1	57.5	67.0	0.0	0.0	0.0	6381

6. There are a total of 3 values /records equalling to zero.

7. **Column Z:**

	carat	cut	color	clarity	depth	table	x	y	z	price
5821	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
6034	2.02	Premium	H	VS2	62.7	53.0	8.02	7.95	0.0	18207
6215	0.71	Good	F	SI2	64.1	60.0	0.00	0.00	0.0	2130
10827	2.20	Premium	H	SI1	61.2	59.0	8.42	8.37	0.0	17265
12498	2.18	Premium	H	SI2	59.4	61.0	8.49	8.45	0.0	12631
12689	1.10	Premium	G	SI2	63.0	59.0	6.50	6.47	0.0	3696
17506	1.14	Fair	G	VS1	57.5	67.0	0.00	0.00	0.0	6381
18194	1.01	Premium	H	I1	58.1	59.0	6.66	6.60	0.0	3167
23758	1.12	Premium	G	I1	60.4	59.0	6.71	6.67	0.0	2383

- There are a total of 9 values/records equalling to zero in column Z.
- Replacing these zero values as NaN values, and further dropping them. The reason to drop these 3 records is that these three values are insignificant when compared to the overall dataset thus not adding much value to the dataset. These three values also have a high multi collinearity between the columns. Hence dropping these three values of X column.

TABLE 10: MISSING VALUE AFTER TREATMENT

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

OUTPUT: CHECKING FOR DUPLICATE DATA

```
Number of duplicate rows = 0
shape after removing: (26931, 10)
```

- As per, Table 9 and the above output we can confirm that there are no missing values and duplicate values in the dataset after treatment.
- The shape of the data has been changed to 26931 rows after the treatment of duplicates.
- **SCALING:**

1. Scaling in regression is required because when one predictor variable has a very large scale. In that case, the regression coefficients may be on a very small order of magnitude which can be unclear to interpret.
2. Scaling helps to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. To supress this, we need to bring all features to the same level of magnitudes.
3. For this dataset, as per Table 5 we find out that mean and std dev numbers aren't varying significantly for numeric variables with a low std deviation and hence, even if we don't scale the numbers, our model performance will not differ much, though when scaling is done the performance is faster and conversion is also faster. The other reason for we can avoid scaling is that there is a strong correlation between independent variables – like Carat, x, y and z. All these variables are strongly correlated with the target variable price. This indicates a strong case of our dataset struggling with multicollinearity. (Table 8). Hence we shall do both scaling and without scaling to understand the difference.

TABLE 11: SCALING THE DATASET

	carat	depth	table	x	y	z	price
0	-1.067422	0.288053	0.261860	-1.296590	-1.289737	-1.259350	-0.933301
1	-1.002505	-0.778984	0.261860	-1.163343	-1.137638	-1.201918	-0.793452
2	0.230914	0.370133	1.189166	0.275730	0.347562	0.348755	0.736237
3	-0.807755	-0.122346	-0.665445	-0.808016	-0.833441	-0.828608	-0.765194
4	-1.045783	-1.107304	0.725513	-1.225525	-1.164479	-1.273708	-0.852563

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

OUTPUT: ENCODING THE DATA

```
feature: cut
['Ideal', 'Premium', 'Very Good', 'Good', 'Fair']
Categories (5, object): ['Fair', 'Good', 'Ideal', 'Premium', 'Very Good']
[2 3 4 1 0]


feature: color
['E', 'G', 'F', 'D', 'H', 'J', 'I']
Categories (7, object): ['D', 'E', 'F', 'G', 'H', 'I', 'J']
[1 3 2 0 4 6 5]


feature: clarity
['SI1', 'IF', 'VVS2', 'VS1', 'VVS1', 'VS2', 'SI2', 'I1']
Categories (8, object): ['I1', 'IF', 'SI1', 'SI2', 'VS1', 'VS2', 'VVS1', 'VVS2']
[2 1 7 4 6 5 3 0]
```

- Here in the above output, we have encoded the categorical variable into numbers as per there ranks, this will help in faster and clarity in modelling.

TABLE 12: ENCODED DATA SAMPLE

	cut	color	clarity
0	2	1	2
1	3	3	1
2	4	1	7
3	2	2	4
4	2	2	6

TABLE 13: CONCATENING THE SCALED & ENCODED DATA

	cut	color	clarity	carat	depth	table	x	y	z	price
0	2	1	2	-1.067422	0.288053	0.261860	-1.296590	-1.289737	-1.259350	-0.933301
1	3	3	1	-1.002505	-0.778984	0.261860	-1.163343	-1.137638	-1.201918	-0.793452
2	4	1	7	0.230914	0.370133	1.189166	0.275730	0.347562	0.348755	0.736237
3	2	2	4	-0.807755	-0.122346	-0.665445	-0.808016	-0.833441	-0.828608	-0.765194
4	2	2	6	-1.045783	-1.107304	0.725513	-1.225525	-1.164479	-1.273708	-0.852563

TABLE 14: TRAIN DATA SAMPLE

- After this we split the dataset into training and testing dataset for modelling purposes into 70:30 ratio

	cut	color	clarity	carat	depth	table	x	y	z	price
2275	2	1	5	-1.067422	-0.450665	-1.129098	-1.225525	-1.271843	-1.273708	-0.833821
12311	2	3	4	0.944998	-0.122346	-0.201793	0.986383	1.045428	0.994868	2.053990
5030	1	1	3	0.663692	1.273010	-0.665445	0.711005	0.759124	0.880004	0.094950
8481	3	5	5	1.529249	-0.532745	1.189166	1.421658	1.456989	1.368178	1.653763
25220	4	5	3	2.665292	0.862611	1.189166	2.007947	2.074332	2.157873	2.374706

TABLE 15: TEST DATA SAMPLE

	cut	color	clarity	carat	depth	table	x	y	z	price
19266	3	3	4	-0.829394	-1.435623	1.652819	-0.772483	-0.860282	-0.929115	-0.800084
22435	4	4	3	0.901720	-0.532745	1.189166	1.004150	1.054375	0.966152	0.361961
20645	2	5	2	-0.569727	-0.122346	-0.387254	-0.479339	-0.466614	-0.469656	-0.752218
308	1	3	4	0.101080	1.519250	-0.201793	0.222431	0.177569	0.363113	-0.121023
14666	2	0	6	-1.110700	-0.040266	-0.201793	-1.358772	-1.343419	-1.345499	-0.901871

TABLE 16: COEFFICIENTS

Intercept	-0.14
carat	1.23
cut	0.01
color	-0.07
clarity	0.07
depth	-0.03
table	-0.05
x	-0.61
y	0.44
z	-0.06

TABLE 17: MEAN SQUARED ERROR AND ROOT SQUARED ERROR FOR TRAIN AND TEST DATA

MEAN SQUARE ERROR	
TRAIN DATA	TEST DATA
0.09	0.09
ROOT MEAN SQUARE ERROR	
TABLE	TEST DATA
0.3	0.3

FIGURE 9: SCATTER PLOT AFTER SCALING

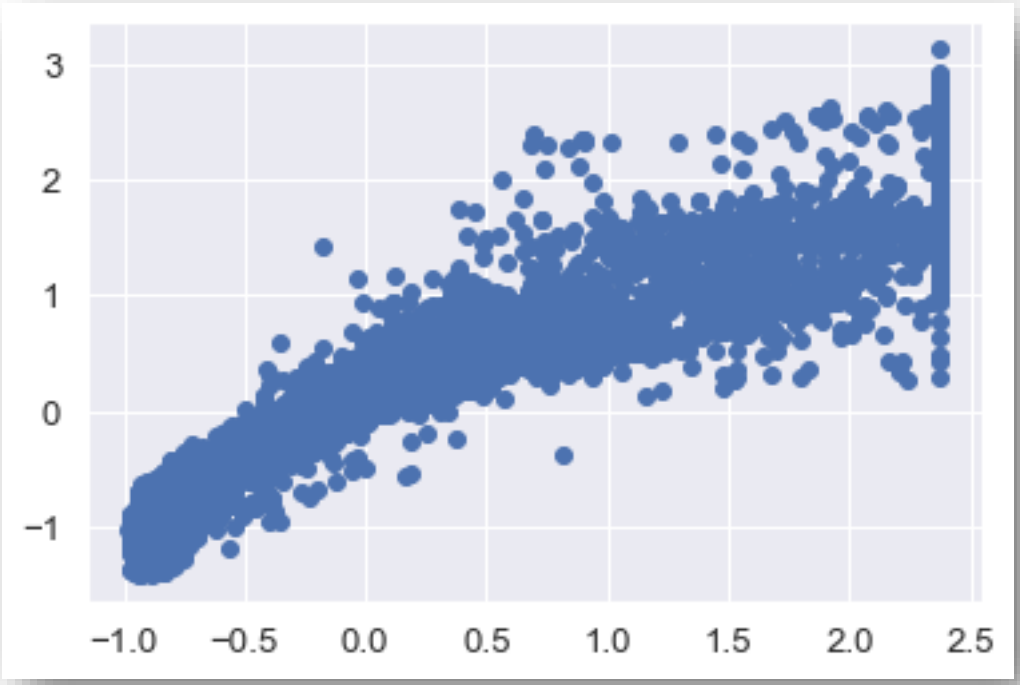


TABLE 18: VARIANCE INFLATION FACTOR

```
cut ---> 4.393244253059092
color ---> 3.1075050075605417
clarity ---> 4.295847713212699
carat ---> 32.22638853574856
depth ---> 2.5801489452711115
table ---> 1.179251706941634
x ---> 380.01456752002485
y ---> 366.81205340602287
z ---> 104.79386394021651
```

TABLE 19: SUMMARY

OLS Regression Results

Dep. Variable:	price	R-squared:	0.909
Model:	OLS	Adj. R-squared:	0.909
Method:	Least Squares	F-statistic:	2.099e+04
Date:	Mon, 08 Aug 2022	Prob (F-statistic):	0.00
Time:	11:55:31	Log-Likelihood:	-4107.3
No. Observations:	18851	AIC:	8235.
Df Residuals:	18841	BIC:	8313.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1392	0.008	-16.441	0.000	-0.156	-0.123
carat	1.2295	0.013	98.035	0.000	1.205	1.254
cut	0.0125	0.002	5.661	0.000	0.008	0.017
color	-0.0651	0.001	-48.286	0.000	-0.068	-0.062
clarity	0.0724	0.001	55.327	0.000	0.070	0.075
depth	-0.0341	0.003	-9.996	0.000	-0.041	-0.027
table	-0.0460	0.002	-19.178	0.000	-0.051	-0.041
x	-0.6100	0.042	-14.546	0.000	-0.692	-0.528
y	0.4379	0.041	10.620	0.000	0.357	0.519
z	-0.0552	0.021	-2.637	0.008	-0.096	-0.014

Omnibus:	4905.697	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25772.938
Skew:	1.156	Prob(JB):	0.00
Kurtosis:	8.241	Cond. No.	144.

INFERENCE FOR THE ABOVE:

1. Intercept of the model is 0.14 as per table 13 and table 18. Which almost equal to 0, it is because of the scaling process.
2. The Linear regression equation can be expressed as the following:

   (-0.14) * Intercept +(1.23) * carat +(0.01) * cut +(-0.07) * color +(0.07) * clarity +(-0.03) * depth +(-0.05) * table +(-0.61) * x +(0.44) * y +(-0.06) * z +

3. R Squared: R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. 100% indicates that the model explains all the variability of the response data around its mean. The value of R-squared

vary from 0 to 1. Any value inching closer to 1 can be considered a good fitted regression line. Here the R-squared is 0.90 or 90%, hence can be inferred that our model signifies a good performance.

4. Root of Mean squared error: RMSE is the standard deviation of the residuals or prediction errors. Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Here in this case the RMSE is 0.3 which is lower which means that the model is performs good.
5. Hypothesis testing for Linear Regression – The null hypothesis states that there is no relation between the dependent variable – Price and other independent variables. Looking at the summary table above, all the P values are less than 0.05 or at 95% confidence level we can say that the variables have a direct impact on the price variable. Carat and clarity variables seem to impact the price rise positively.
6. The scatter plot is near -1 which means that it is negatively correlated and has an inverse relationship.

PERFORMING LINEAR REGRESSION WITH UNSCALED DATA

TABLE 20: ENCODING THE UNSCALED DATA

	cut	color	clarity	carat	depth	table	x	y	z	price
0	2	1	2	0.30	62.1	58.0	4.27	4.29	2.66	499.0
1	3	3	1	0.33	60.8	58.0	4.42	4.46	2.70	984.0
2	4	1	7	0.90	62.2	60.0	6.04	6.12	3.78	6289.0
3	2	2	4	0.42	61.6	56.0	4.82	4.80	2.96	1082.0
4	2	2	6	0.31	60.4	59.0	4.35	4.43	2.65	779.0

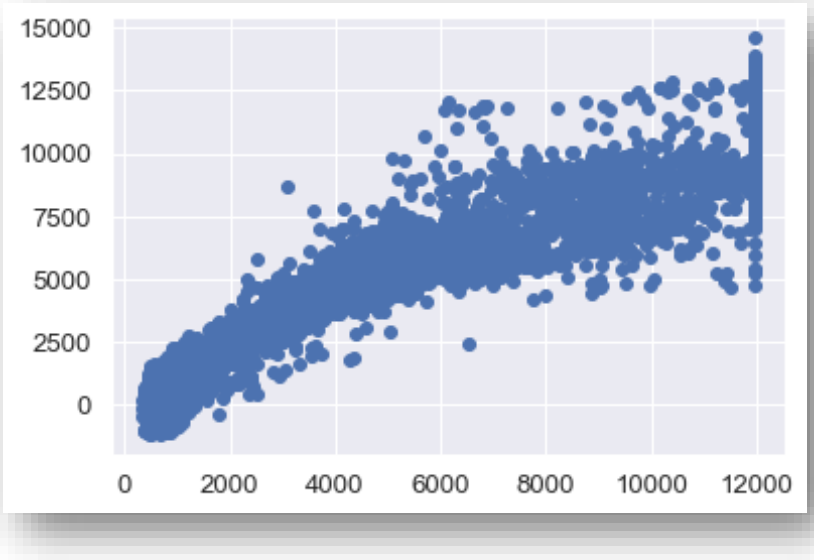
TABLE 21: COEFFICIENTS OF UNSCALED DATA

```
The coefficient for cut is 43.41885676565526
The coefficient for color is -225.7185000091394
The coefficient for clarity is 251.19760146785285
The coefficient for carat is 9227.046878634877
The coefficient for depth is -97.19552592061507
The coefficient for table is -73.92168040739321
The coefficient for x is -1879.3332650774296
The coefficient for y is 1358.7998834384387
The coefficient for z is -274.88826107349945
```

TABLE 22: MSE AND RMSE FOR UNSCALED DATA

MEAN SQUARE ERROR	
TRAIN DATA	TEST DATA
0.90	0.91
ROOT MEAN SQUARE ERROR	
TRAIN DATA	TEST DATA
1043.44	1039.48

FIGURE 10: SCATTER PLOT FOR UNSCALED DATA



- INFERENCE FOR THE ABOVE (UNSCALED DATA):

1. The intercept is now changed to 43.41 (Table 20)
2. The mean square error for the training data is unchanged to 0.9 and similarly for test data also which is 0.91.
3. The RMSE is 1043.44 for train data and 1039.48 for testing data. To understand more we do Sum of square error/ Sum of square actual error which is 0.48 for both testing and training data, which mean that the model is in the right fit zone and avoids being under or over fit.
4. The scatter plot is also similar to the scaled scatter plot.
5. This proves that scaling is not necessary in this model.

Q1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

1. The database has a strong correlation between independent variables and which means that there is high multicollinearity which can affect the results of the model. Multicollinearity makes it difficult to understand how one variable influence the target variable. However, it does not affect the accuracy of the model. As a result, while creating the model.
2. In the univariate analysis above, we can conclude that the Price variable is highly correlated with Carat variable in fact with other variables such as x, y, z and has lower

correlation with cut and table variables Hence Carat variable is strong predictor affecting the model.

3. After Linear Regression we can observe that Carat variable is affecting the target variable price majorly. We can also observe that the carat variable has the highest coefficient value comparatively. The above observation is the same when the dataset is unscaled.
4. VIF measures the intercorrelation among independent variables in a multiple regression model, here in this model we can see that there is high intercorrelation in x, y, z and should be treated for the same.
5. Carat and clarity have a higher and positive influence on the price change relatively.
6. As expected Carat is a strong predictor of the overall price of the stone. Clarity has emerged as a strong predictor of price as well. Clarity of stone types IF, VVS_1, VVS_2 and vs1 are helping the firm put an expensive price cap on the stones.
7. Colour of the stones such as H, I and J are not helping the firm put an expensive price tag on such stones. The company should rather focus colour D, E and F to have relative higher prices and support sales.
8. The company should come up with new colour stones like clear stones or a different colour/unique colour that helps impact the price positively.
9. The company should focus on the stone's carat and clarity so as to increase their prices and sales.
10. Ideal customers will anyways be contributing to more profits.
11. The company can do marketing which can be in turn be useful to educate customers about the importance of the carat of the stone and importance of clarity index.
12. Apart from this the company can make segments, and target the customer based on the customers background or capacity to afford, which can be analysed using more information.

PROBLEM 2

INTRODUCTION

A tour and travel agency which deals in selling holiday packages provides details of 872 employees of a company. Among these employees, some opted for the package and some didn't. The company wants to predict whether an employee will opt for the package or not on

the basis of the information given in the data set. Also, wants to know 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

Q.2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

TABLE 23: TOP 5 DATA SAMPLES

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

TABLE 24: DATASET INFORMATION

```
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
```

TABLE 25: EXTRACTING VALUES FROM VARIABLES

HOLLIDAY PACKAGE	
NAME	COUNT
No	54.01
Yes	45.98
FOREIGN	
NAME	COUNT
No	75.22
Yes	24.77
EDU	
NAME	COUNT
(0.979, 11.0]	74.31
(11.0, 21.0]	25.68
no_young_children	
NAME	COUNT
0	76.26
1	16.85
2	6.3
3	0.57
no_older_children	
NAME	COUNT
0	45.06
2	23.82
1	22.70
3	6.30

4	1.60
5	0.22
6	0.22
Age	
NAME	COUNT
(19.95, 30.5]	183.00
(30.5,41.0]	309.00
(41.0,51.5]	239.00
(51.5,62.0]	141.00

TABLE 26: DESCRIPTION OF THE DATASET

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

TABLE 27: MISSING VALUES

Holliday_Package	0
Salary	0
age	0
educ	0
no_young_children	0
no_older_children	0
foreign	0

TABLE 28: UNIQUE VALUE COUNTS

```
HOLLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holliday_Package, dtype: int64


EDUC :  20
1       1
21      1
18      1
19      2
17      3
2       6
16     10
3      11
15     15
6      21
14     25
7      31
13     43
4      50
5      67
10     90
11    100
9     114
12    124
8     157
Name: educ, dtype: int64
```

```
NO_YOUNG_CHILDREN :  4
3      5
2     55
1    147
0    665
Name: no_young_children, dtype: int64


NO_OLDER_CHILDREN :  7
5      2
6      2
4     14
3     55
1    198
2    208
0    393
Name: no_older_children, dtype: int64


FOREIGN :  2
yes    216
no     656
Name: foreign, dtype: int64
```

TABLE 29: CHECKING FOR DUPLICATE RECORDS

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

Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign

FIGURE 11: NORMAL DISTRIBUTION

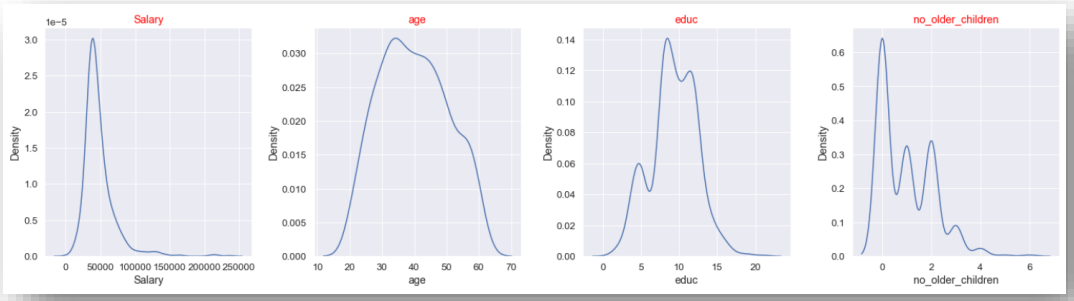


FIGURE 12: IDENTIFYING OUTLIERS USING BOXPLOT

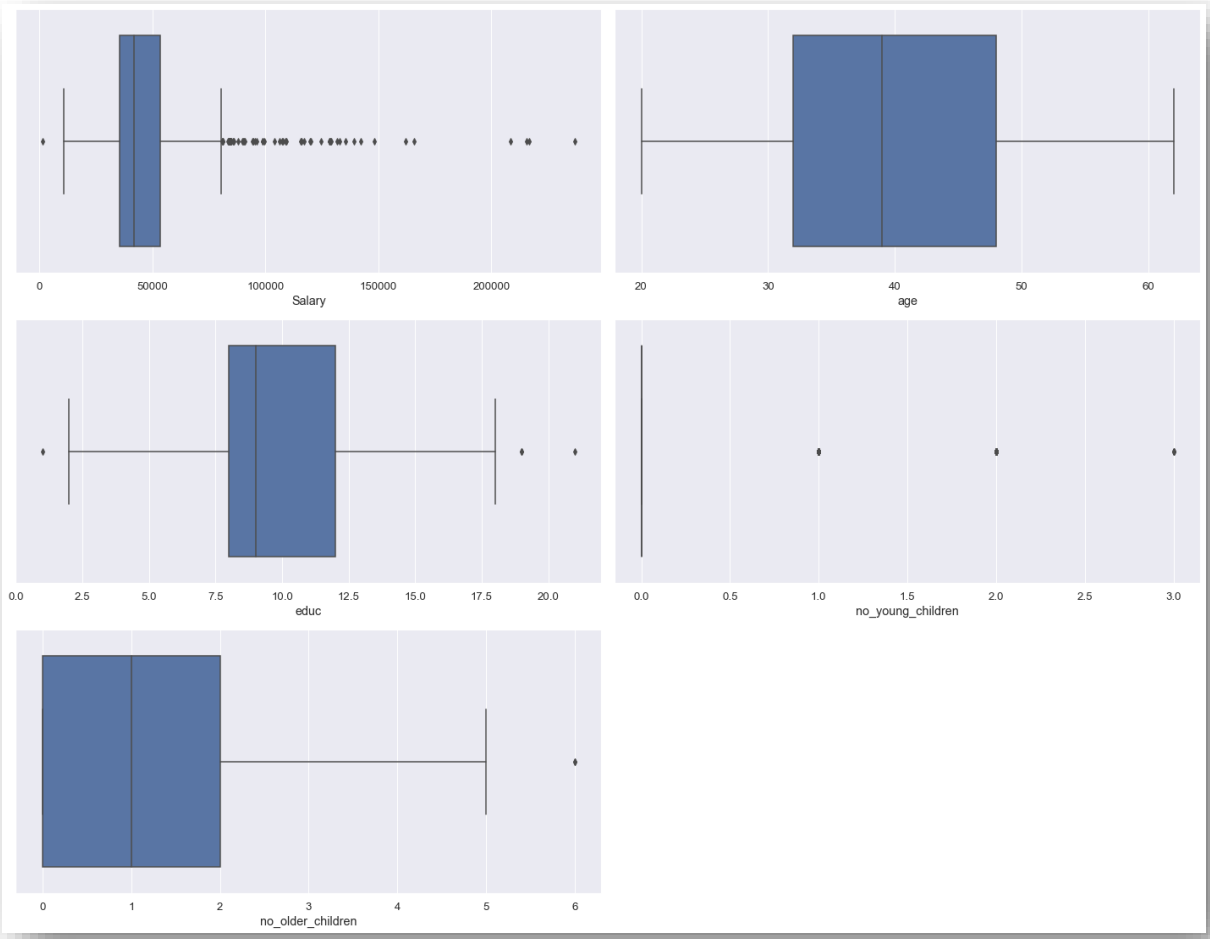
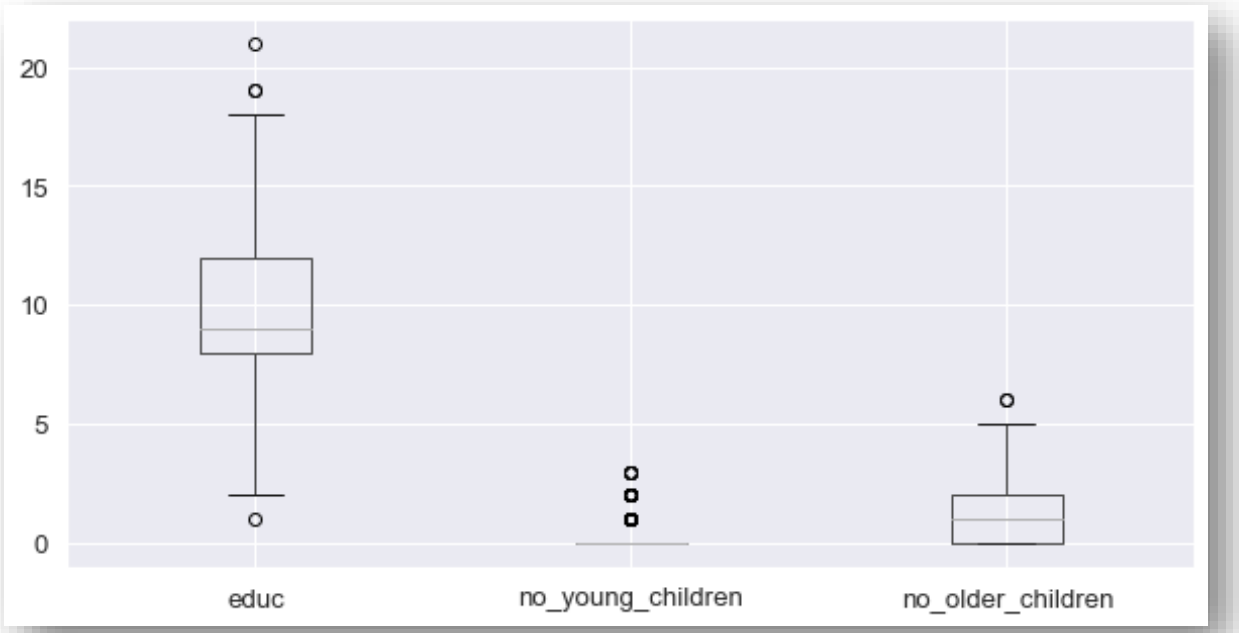


FIGURE 13: OUTLIER TREATMENT USING IQR



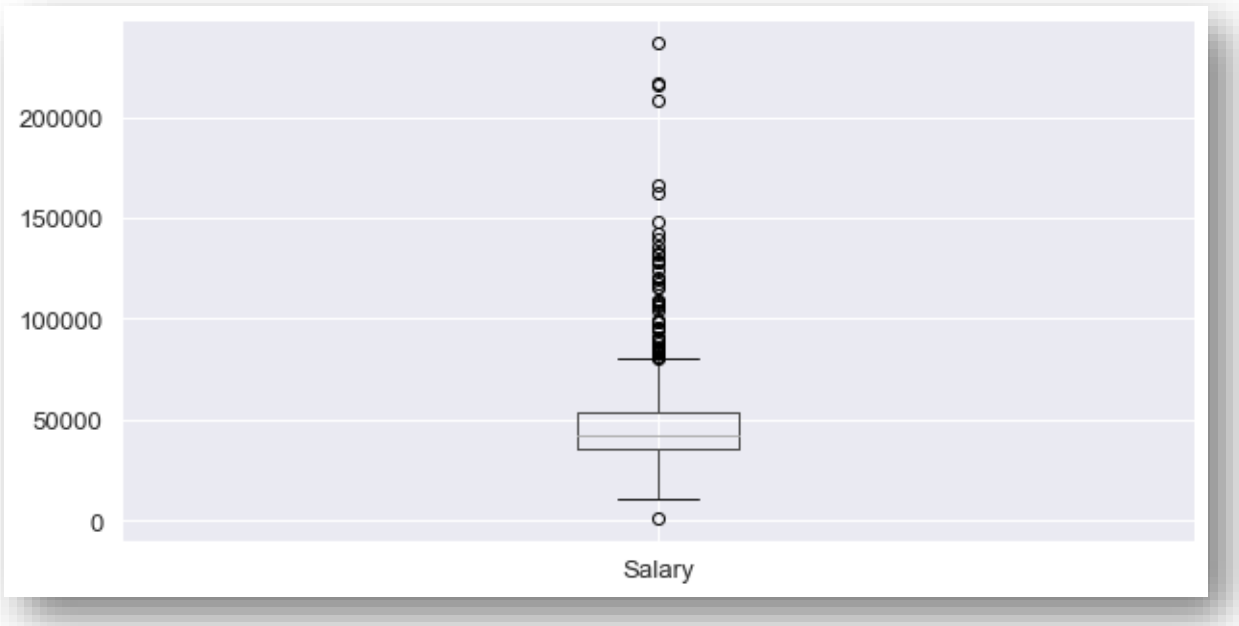


FIGURE 14: PAIRPLOT

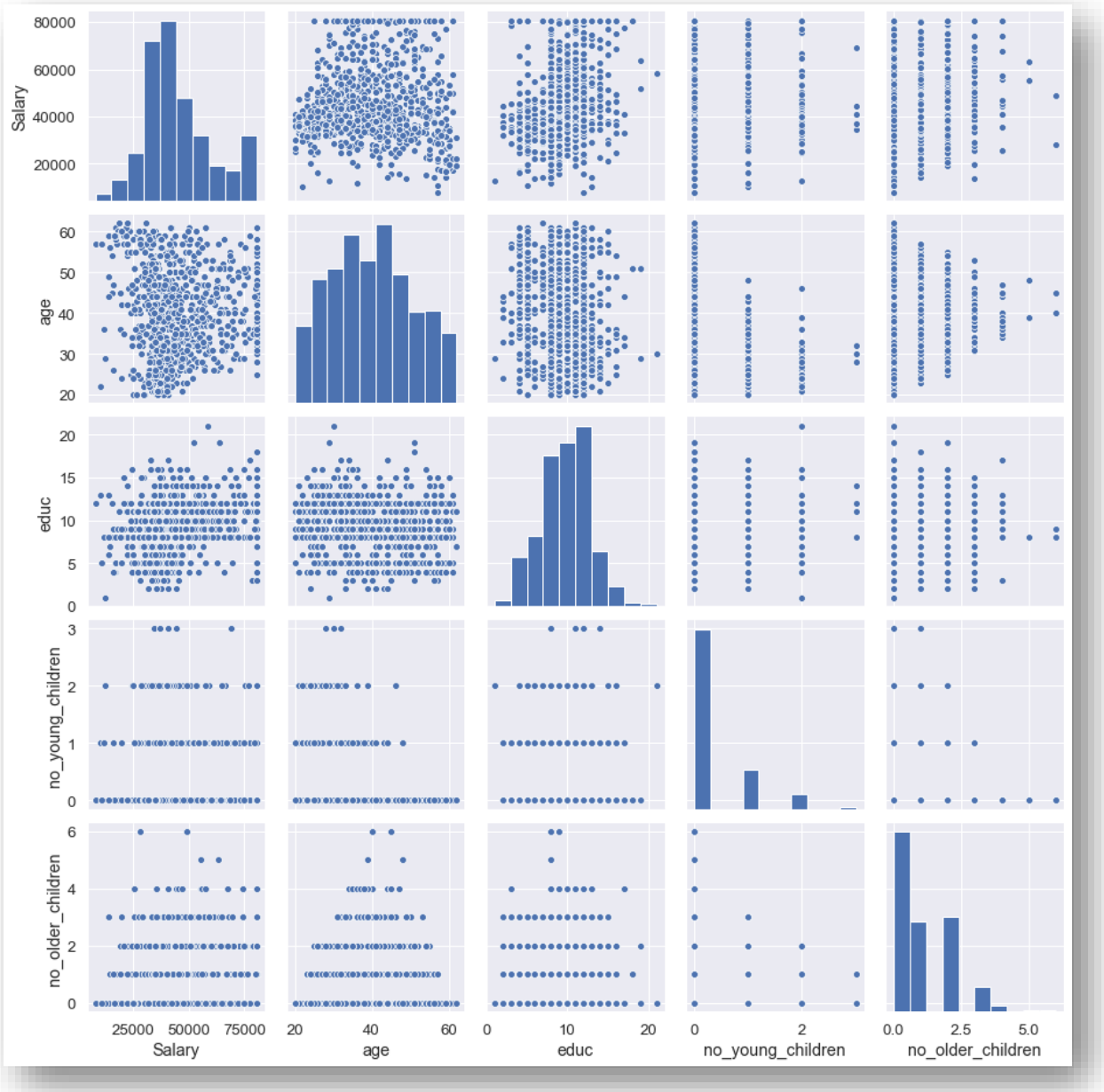


TABLE 30: CORRELATION TABLE

	Salary	age	educ	no_young_children	no_older_children
Salary	1.000000	0.047029	0.352726	-0.034360	0.121993
age	0.047029	1.000000	-0.149294	-0.519093	-0.116205
educ	0.352726	-0.149294	1.000000	0.098350	-0.036321
no_young_children	-0.034360	-0.519093	0.098350	1.000000	-0.238428
no_older_children	0.121993	-0.116205	-0.036321	-0.238428	1.000000

FIGURE 15: HEATMAP/CORRELATION PLOT

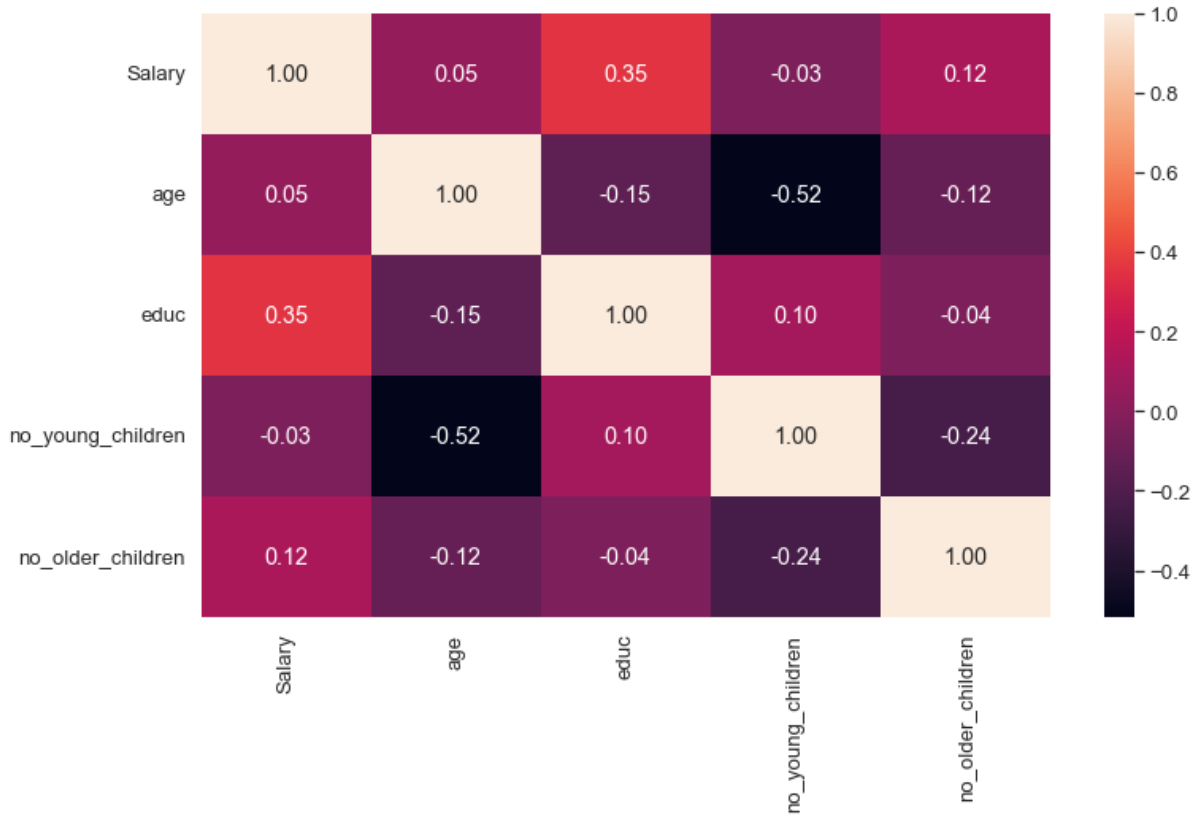


FIGURE 16: BAR GRAPH FOR HOLLIDAY PACKAGE VARIABLE

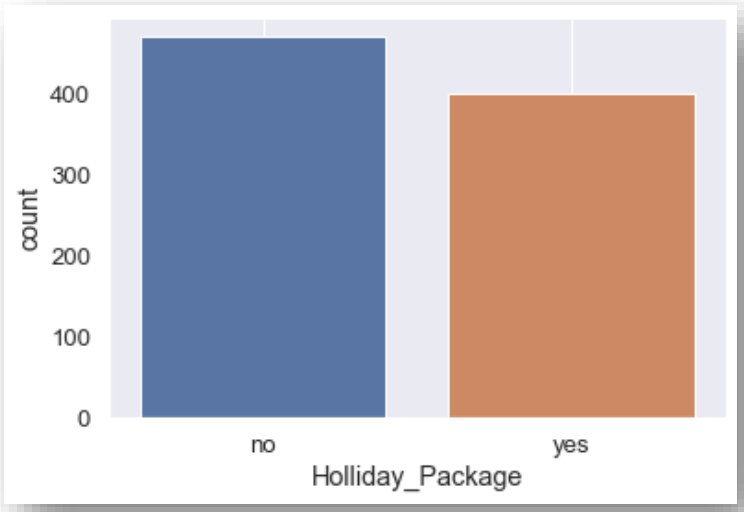


FIGURE 17: BAR GRAPH FOR FOREIGN VARIABLE

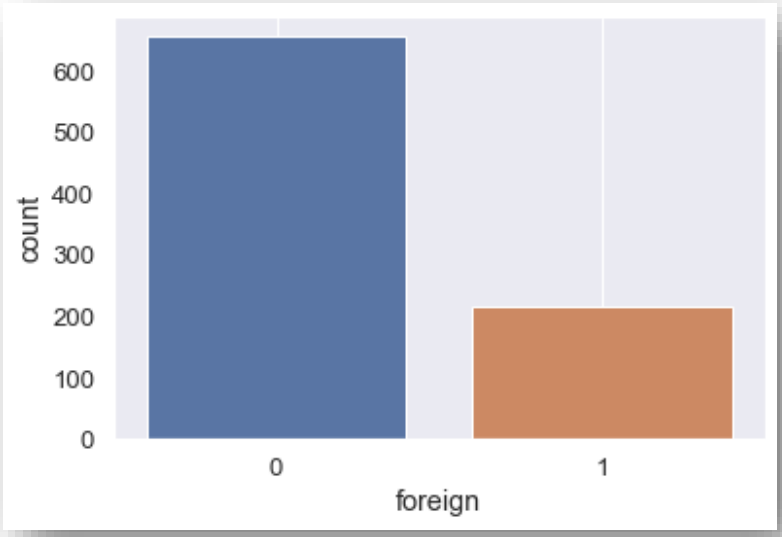


FIGURE 18: BOX PLOT - HOLIDAY PACKAGE AND SALARY

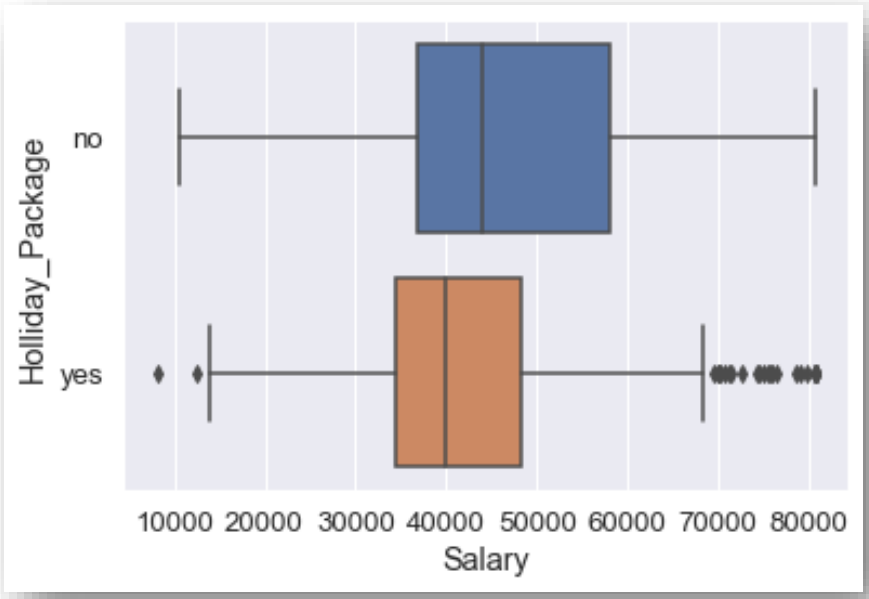


FIGURE 19: BOX PLOT - HOLIDAY PACKAGE AND EDUC

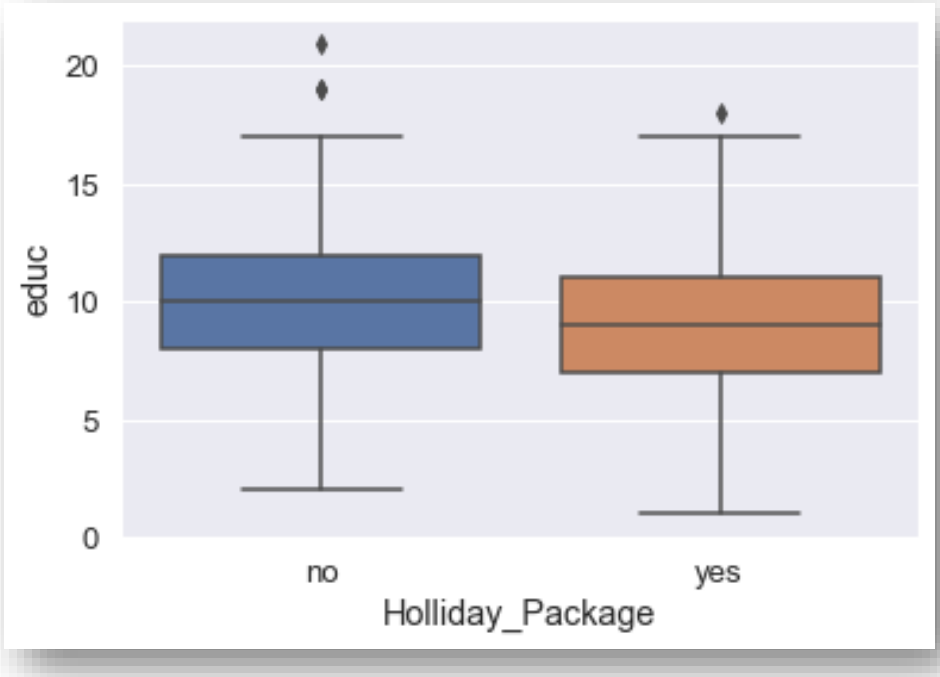


FIGURE 20: BOX PLOT - HOLIDAY PACKAGE AND NO YOUNG CHILDREN

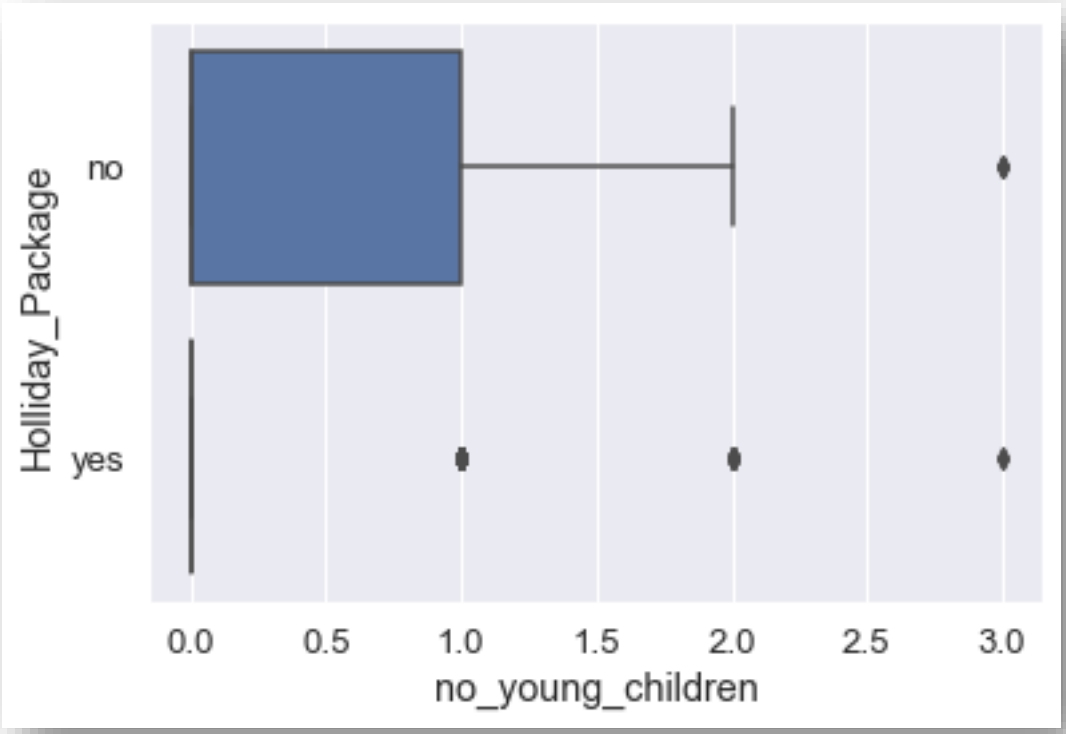


FIGURE 21: BOX PLOT - HOLIDAY PACKAGE AND NO OLDER CHILDREN

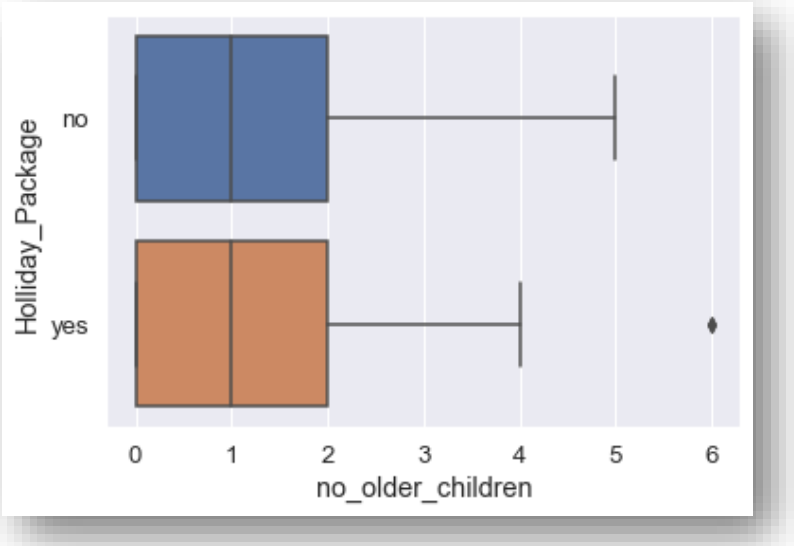


FIGURE 22: BOX PLOT - HOLIDAY PACKAGE AND AGE

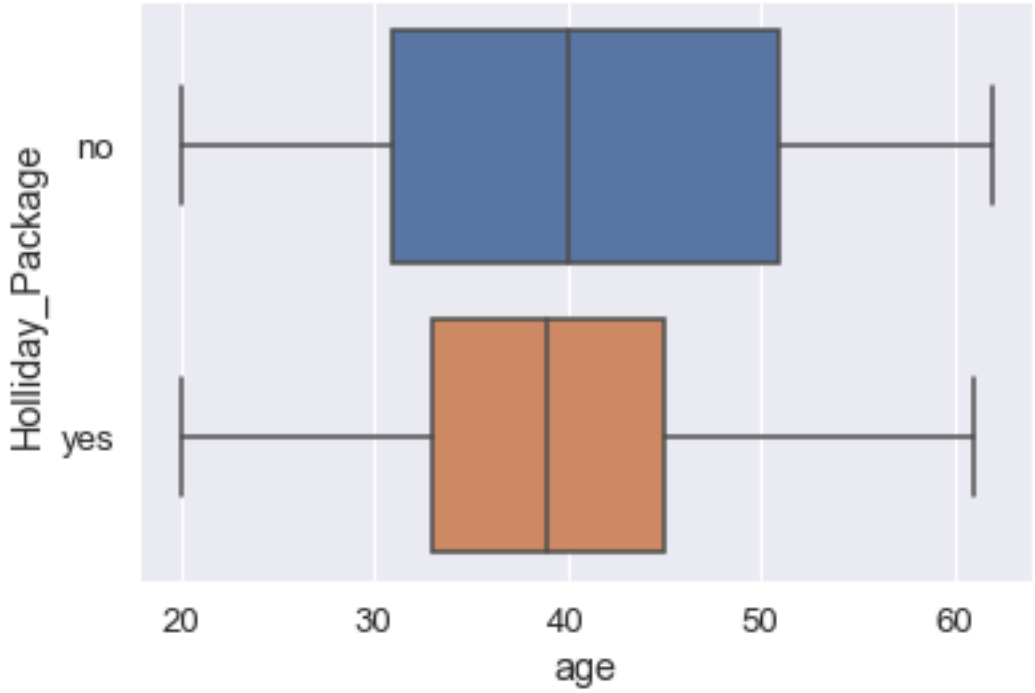


FIGURE 23: BAR GRAPH- HOLIDAY PACKAGE AND FOREIGN

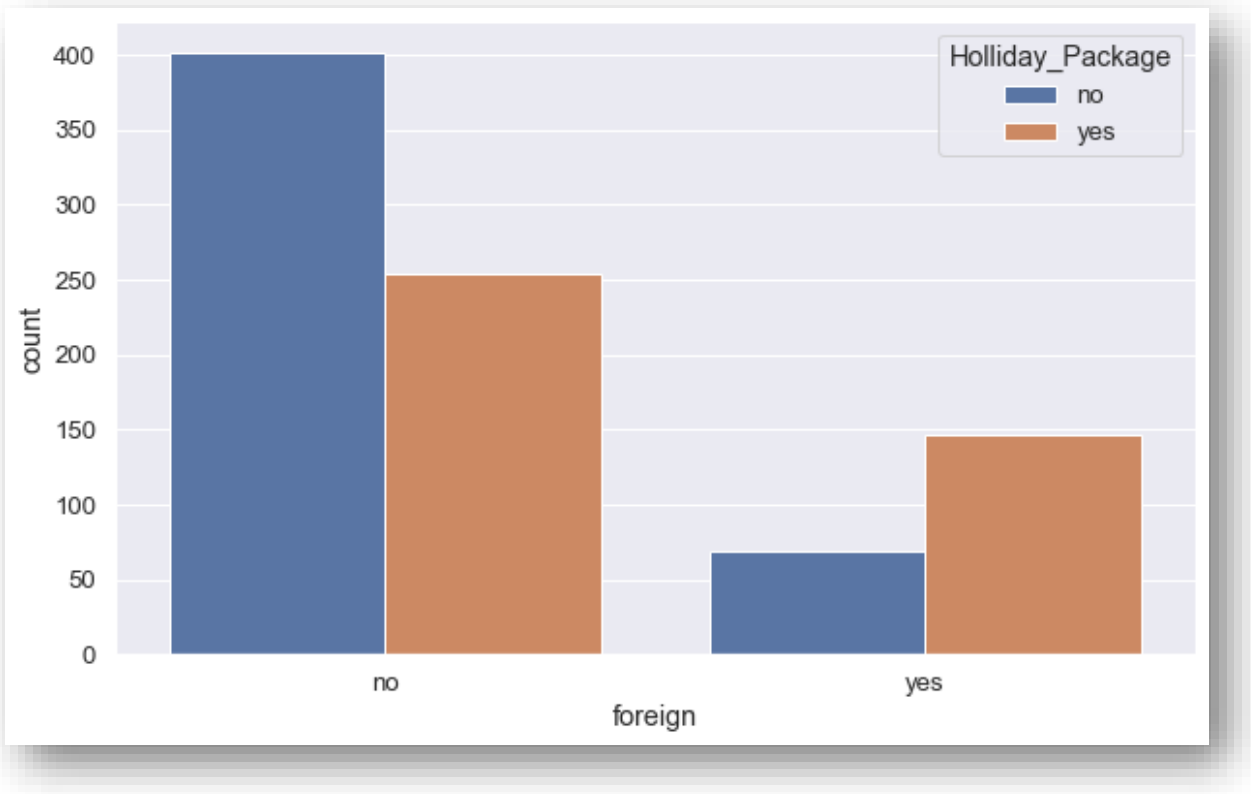


FIGURE 24: BAR GRAPH - HOLIDAY PACKAGE AND EDUC

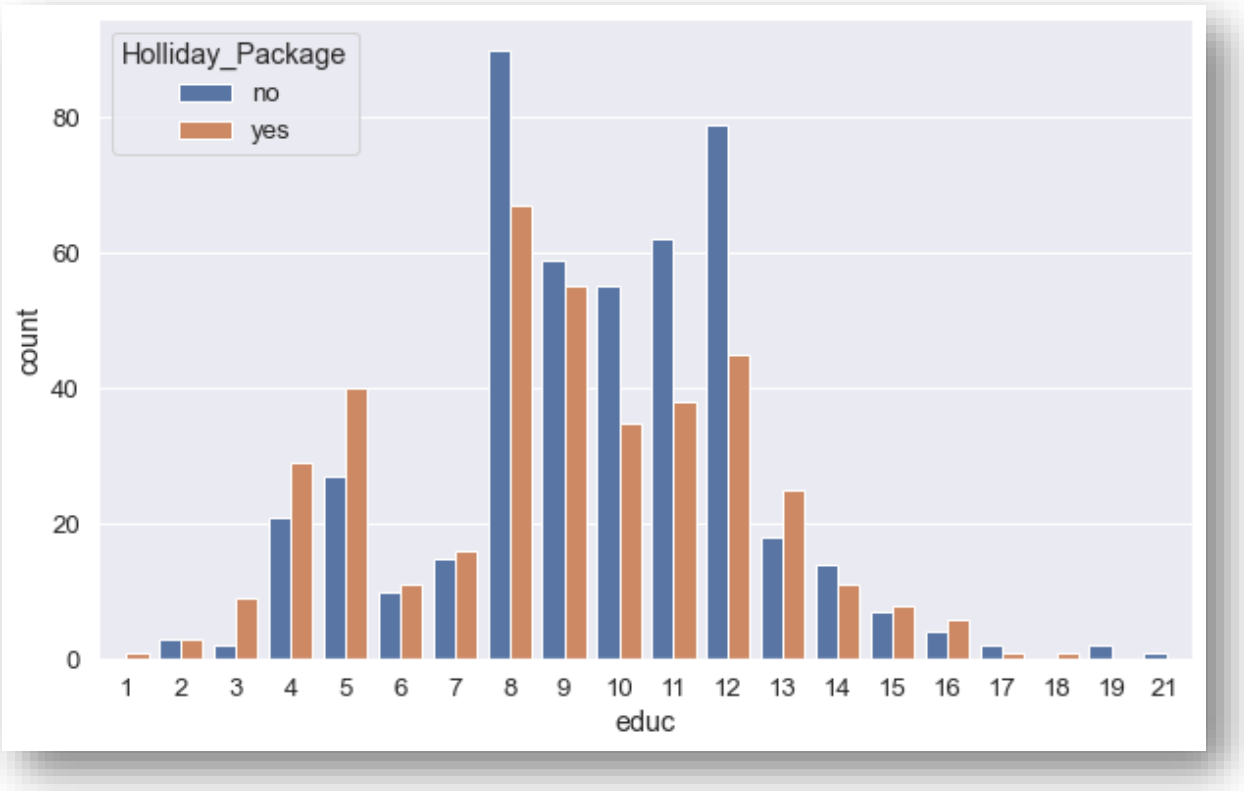


FIGURE 25: BAR GRAPH - HOLIDAY PACKAGE& NO OF YOUNG CHILDREN

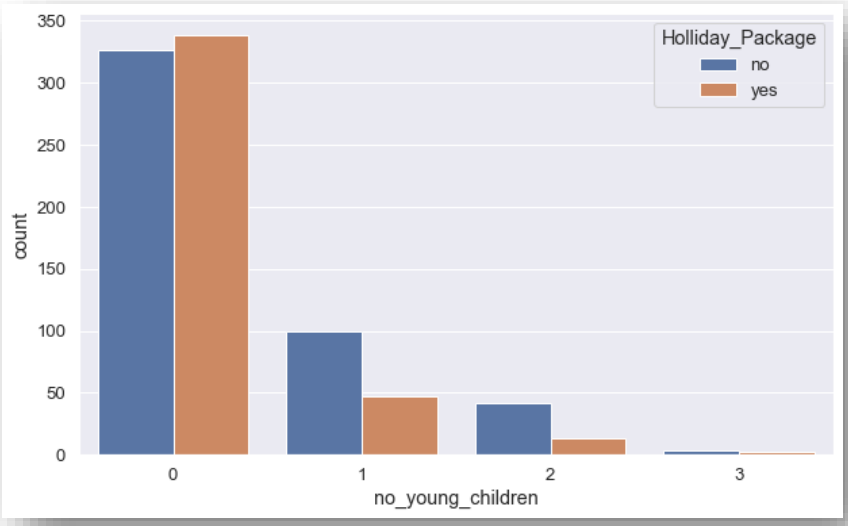


FIGURE 26: BAR GRAPH - HOLIDAY PACKAGE& NO OF OLDER CHILDREN

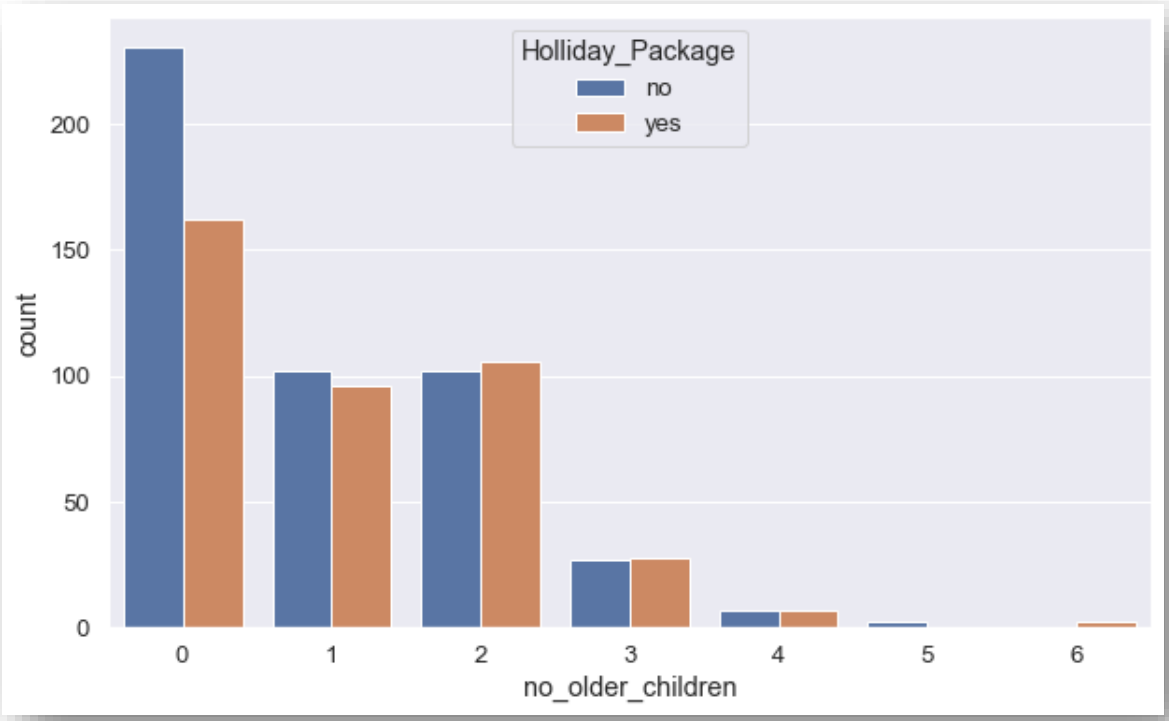
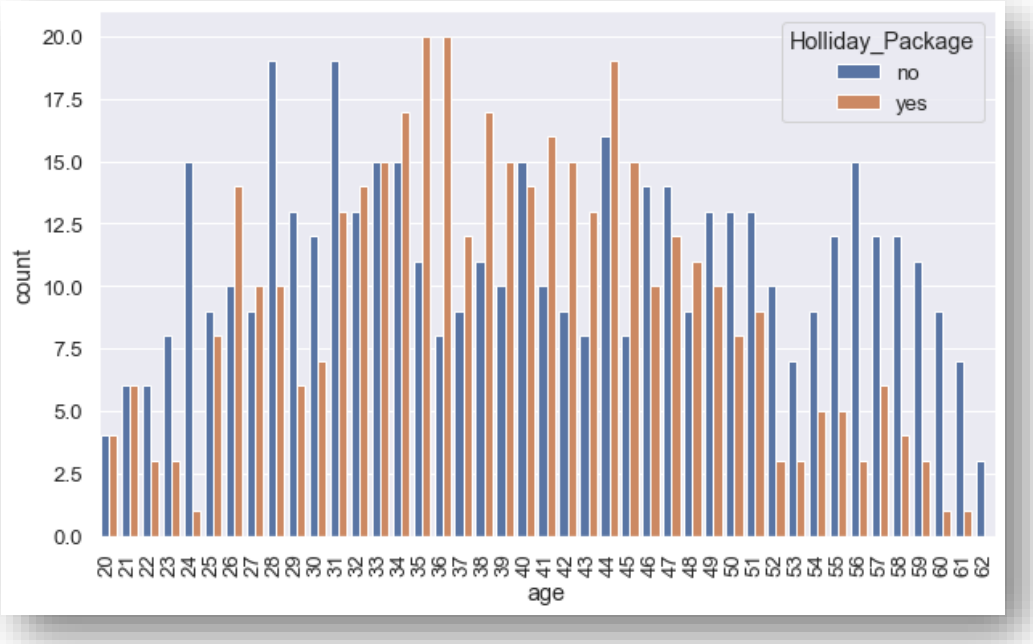


FIGURE 27: BAR GRAPH - HOLIDAY PACKAGE& AGE



INFERENCE FOR Q.2.1:

1. **Exploratory Data Analysis or (EDA)** is understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data. Plotting in EDA consists of Histograms, Box plot, pair plot and many more. It often takes much time to explore the data. Through the process of EDA, we can define the problem statement or definition on our data set which is very important.
2. The dimension of the dataset is 872 rows with 10 columns.
3. The data type in the dataset is Integer/Object type.
4. There are no missing values in any of the variables in the dataset
5. There no duplicates in the dataset as well.
6. Holiday Package – This variable is a categorical variable this is the Target Variable. Salary, age, educ, no_young_children, no_older_children, variables are numerical or continuous variables.
7. Salary ranges from 1322 to 236961. Average salary of employees is around 47729 with a standard deviation of 23418. Standard deviation indicates that the data is not normally distributed. skew of 0.71 indicates that the data is right skewed and there are few employees earning more than an average of 47729. 75% of the employees are earning below 53469 while 255 of the employees are earning 35324.
8. Age of the employee ranges from 20 to 62. Median is around 39. 25% of the employees are below 32 and 25% of the employees are above 48. Standard deviation is around 10. Standard deviation indicates almost normal distribution.
9. Years of formal education ranges from 1 to 21 years. 25% of the population has formal education for 8 years, while the median is around 9 years. 75% of the employees have formal education of 12 years. Standard deviation of the education is around 3. This variable is also indicating skewness in the data Foreign is a categorical variable
10. We can observe that 54% of the employees are not opting for the holiday package and 46% are interested in the package. This implies we have a dataset which is fairly balanced.
11. We can observe that 75% of the employees are not Foreigners and 25% are foreigners
12. From Figure 11 we can observe that there is no normal distribution in any variable, though salary and age are slightly distributed normally but because of few deviations we cannot conclude that they are normally distributed.

13. There is no strong correlation between the variables. There is minimal correlation between Salary and education and between no of older children and salary variable. [Figure 15]
14. We can observer that there are significant outliers present in variable "Salary", however there are minimal outliers in other variables like 'educ', 'no. of young children' & 'no. of older children'. There are no outliers in variable 'age'. For Interpretation purpose we would need to study the variables such as no. of young children and no. of older children before outlier treatment. For this case study we have done outlier treatment for only salary & educ.
15. Salary for employees opting for holiday package and for not opting for holiday package is similar in nature. However, the distribution is fairly spread out for people not opting for holiday packages. [Figure 18]
16. The distribution of data for age variable with holiday package is also similar in nature [Figure 22]
17. This variable is also showing a similar pattern. This means education is likely not to be a variable for influencing holiday packages for employees. [Figure 19]
18. There is a significant difference in employees with younger children who are opting for holiday package and employees who are not opting for holiday package [Figure 20]
19. The distribution for opting or not opting for holiday packages looks same for employees with older children. At this point, this might not be a good predictor while creating our logistics model. [Figure 21]
20. We observe that employees with less years of formal education (1 to 7 years) and higher education are not opting for the Holiday package as compared to employees with formal education of 8 year to 12 years. [Figure 24]
21. People with younger children are not opting for holiday packages. [Figure 25]
22. Older children also have same pattern as that of no younger children. [Figure 26]
23. We can clearly see that employees in middle range (34 to 45 years) are going for holiday package as compared to older and younger employees. [Figure 27]

Q.2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

TABLE 31: PROPORTIONS OF 1 AND 0

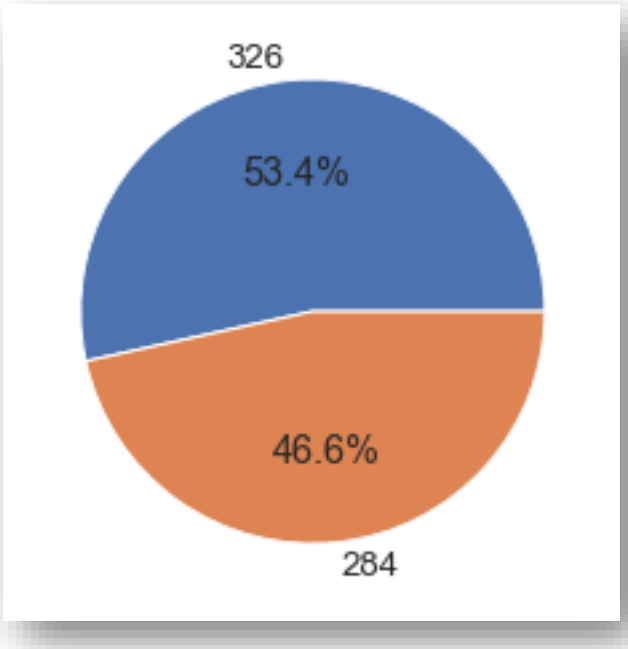
HOLIDAY PACKAGE	
0	54.01
1	45.98

FOREIGN	
0	75.22
1	24.77

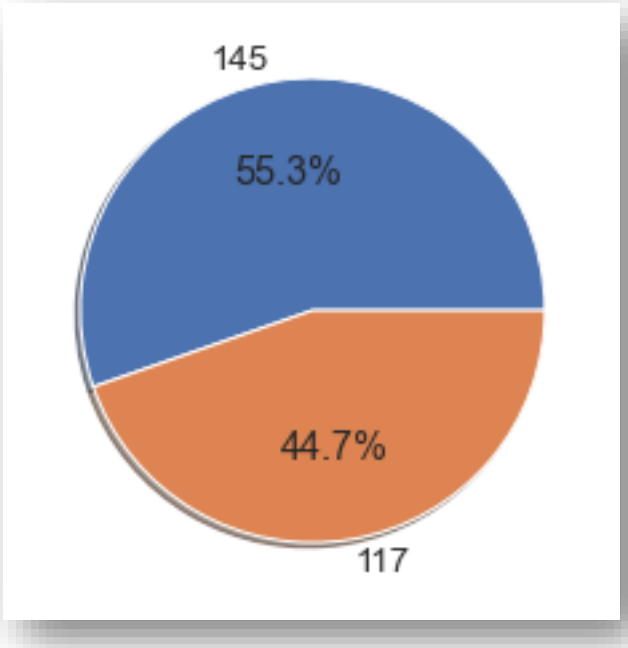
TABLE 32: DROPPING TARGET VARIABLE

	Salary	age	educ	no_young_children	no_older_children	foreign
0	48412.0	30	8	1	1	0
1	37207.0	45	8	0	1	0
2	58022.0	46	9	0	0	0
3	66503.0	31	11	2	0	0
4	66734.0	44	12	0	2	0

FIGURE 28: SPLITTING INTO TRAIN AND TEST DATA



TRAIN DATA



TEST DATA

OUTPUT: DIMENSIONS OF TRAIN AND TEST DATA

```
X_train (610, 6)
X_test (262, 6)
train_labels (610,)
test_labels (262,)
```

OUTPUT: APPLYING LOGISTIC REGRESSION WITH GRID SEARCH CV

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

OUTPUT: EXTRACTING THE BEST PARAMETERS

```
{'penalty': 'l1', 'solver': 'liblinear', 'tol': 1e-06, 'verbose': True}

LogisticRegression(max_iter=10000, n_jobs=2, penalty='l1', solver='liblinear',
                   tol=1e-06, verbose=True)
```

Q.2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

- We apply prediction and probability model for both the train and test data to produce confusion matric and other performances.
- Confusion matrix cells are populated by the terms:
  1. True Positive (TP)- The values which are predicted as True and are actually True.
  2. True Negative (TN)- The values which are predicted as False and are actually False.
  3. False Positive (FP)- The values which are predicted as True but are actually False.
  4. False Negative (FN)- The values which are predicted as False but are actually True.
- ROC CURVE: Receiver Operating Characteristic (ROC) measures the performance of models by evaluating the trade offs between sensitivity and false positive rate.
- AUC - The area under curve (AUC) is another measure for classification models is based on ROC. It is the measure of accuracy judged by the area under the curve for ROC.

FIGURE 29: CONFUSION MATRIX FOR TRAINING DATA

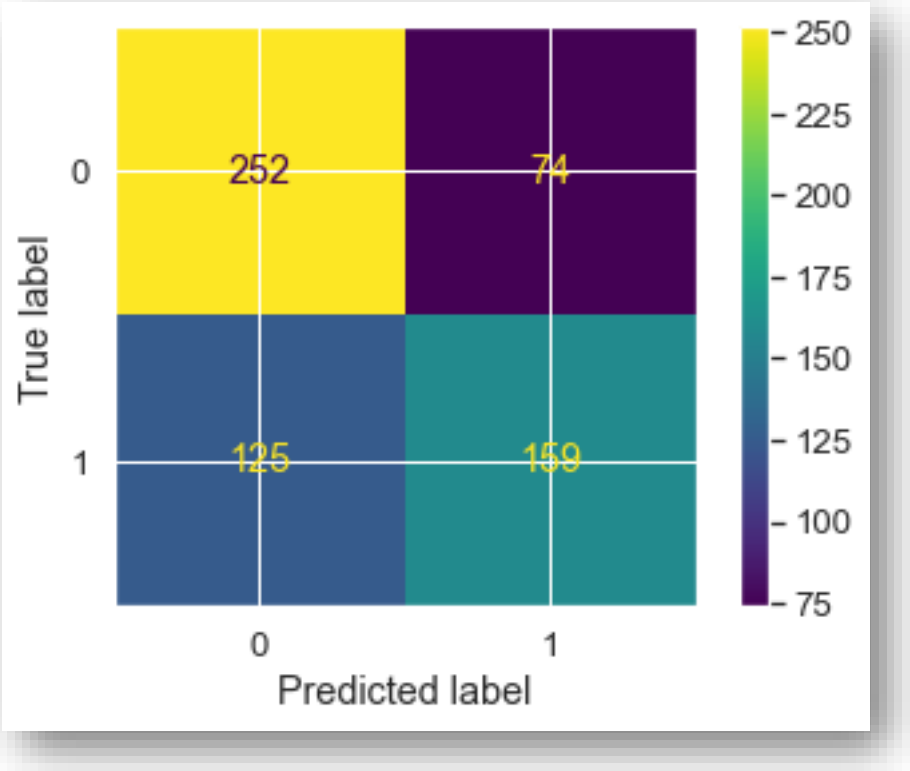


FIGURE 30: CONFUSION MATRIX FOR TEST DATA

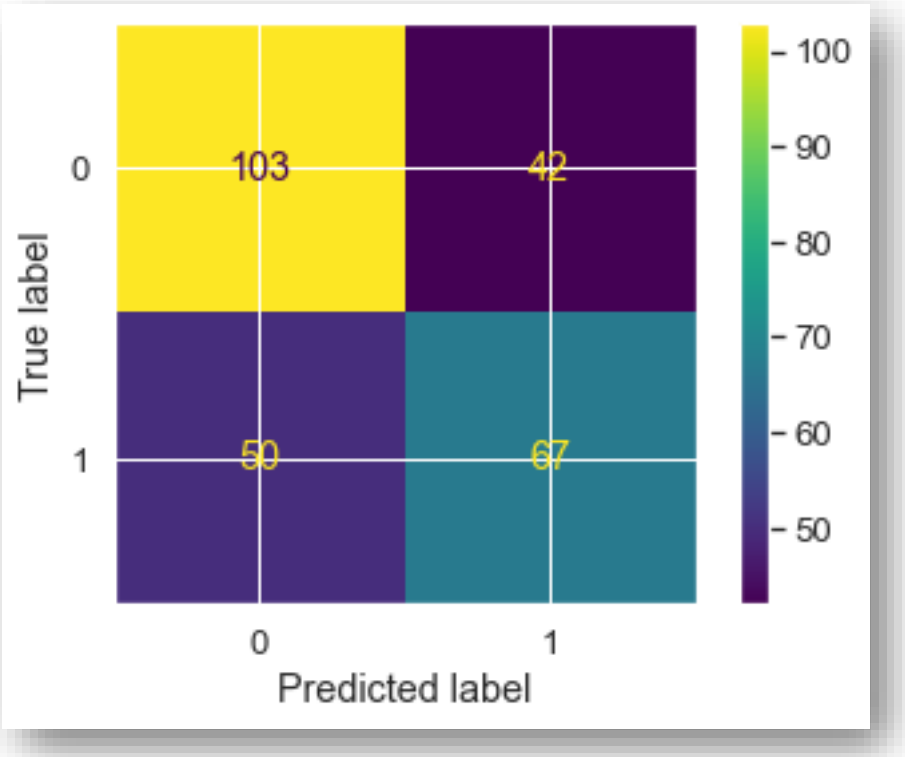


TABLE 33: CLASSIFICATION REPORT - TRAIN DATA

training data		precision	recall	f1-score	support
0	0.67	0.77	0.72	326	
1	0.68	0.56	0.62	284	
accuracy			0.68	610	
macro avg	0.68	0.67	0.67	610	
weighted avg	0.68	0.68	0.67	610	

TABLE 33: CLASSIFICATION REPORT - TEST DATA

testing data		precision	recall	f1-score	support
0	0.66	0.70	0.68	145	
1	0.60	0.56	0.58	117	
accuracy			0.64	262	
macro avg	0.63	0.63	0.63	262	
weighted avg	0.64	0.64	0.64	262	

TABLE 35:

	ACCURACY	BEST MODEL ACCURACY
TRAIN DATA	0.67	67
TEST DATA	0.63	0.64

FIGURE 31: AUC CURVE FOR TRAIN AND TEST DATA

LINEAR DISCRIMINANT ANALYSIS

- We apply Linear Discriminant analysis and fit the training and testing data.

TABLE 36: ACCURACY- TRAIN AND TEST DATA

	ACCURACY
TRAIN DATA	0.67
TEST DATA	0.64

OUTPUT: TRAIN DATA CLASS PREDICTION

```
array([0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0,
       0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1,
       0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,
       1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0,
       0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0,
       1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1,
       0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
       1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0,
       0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0,
       1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0,
       0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1,
       0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0,
       1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1,
       0, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0,
       0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0,
       1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0,
       1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0], dtype=int8)
```

OUTPUT: TEST DATA CLASS PREDICTION

```
array([0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,
       1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0,
       0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1,
       1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0,
       1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0,
       0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,
       0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0,
       0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0,
       0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1,
       0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0],
      dtype=int8)
```

FIGURE 32: CONFUSION MATRIX LDA- TRAIN AND TEST DATA

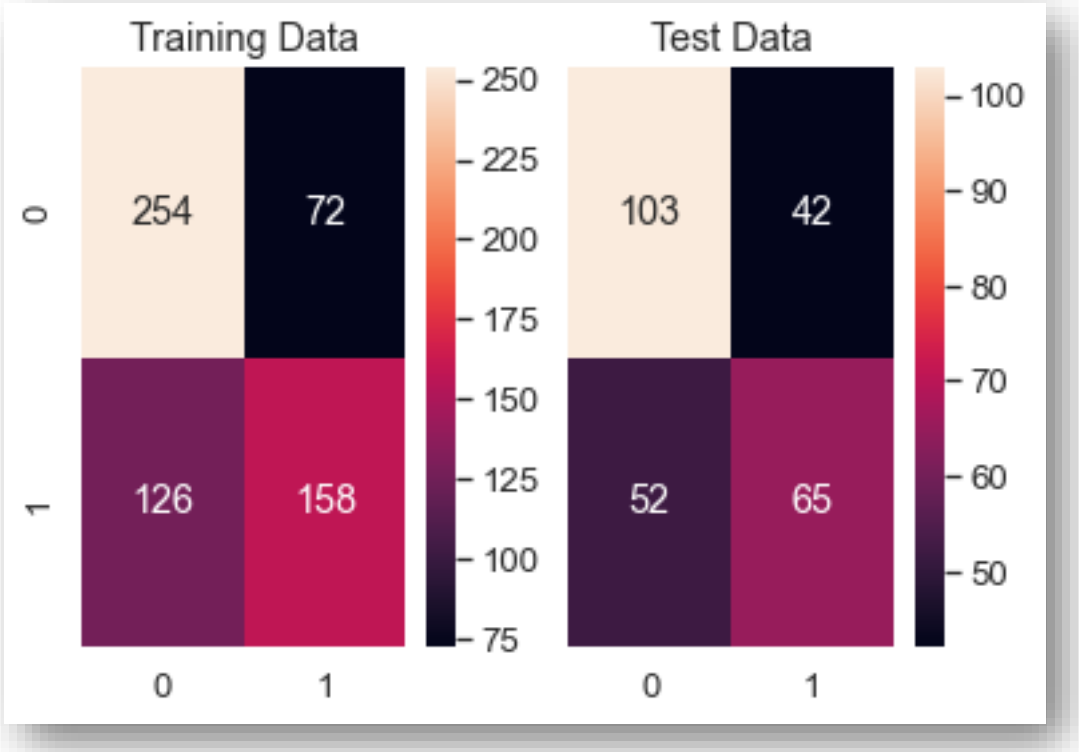


TABLE 37: CLASSIFICATION REPORT LDA- TRAIN DATA

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.67	0.78	0.72	326
1	0.69	0.56	0.61	284
accuracy			0.68	610
macro avg	0.68	0.67	0.67	610
weighted avg	0.68	0.68	0.67	610

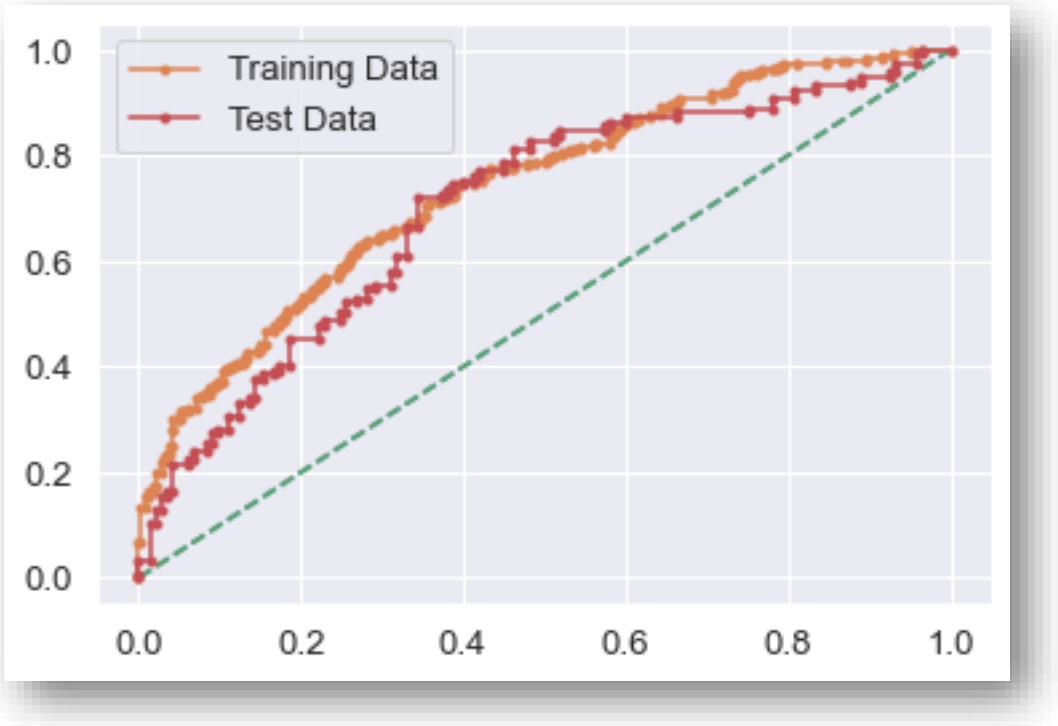
TABLE 38: CLASSIFICATION REPORT LDA- TEST DATA

Classification Report of the test data:

	precision	recall	f1-score	support
0	0.66	0.71	0.69	145
1	0.61	0.56	0.58	117
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.64	262

For the probability distribution of train and test data please refer the Jupyter notebook.

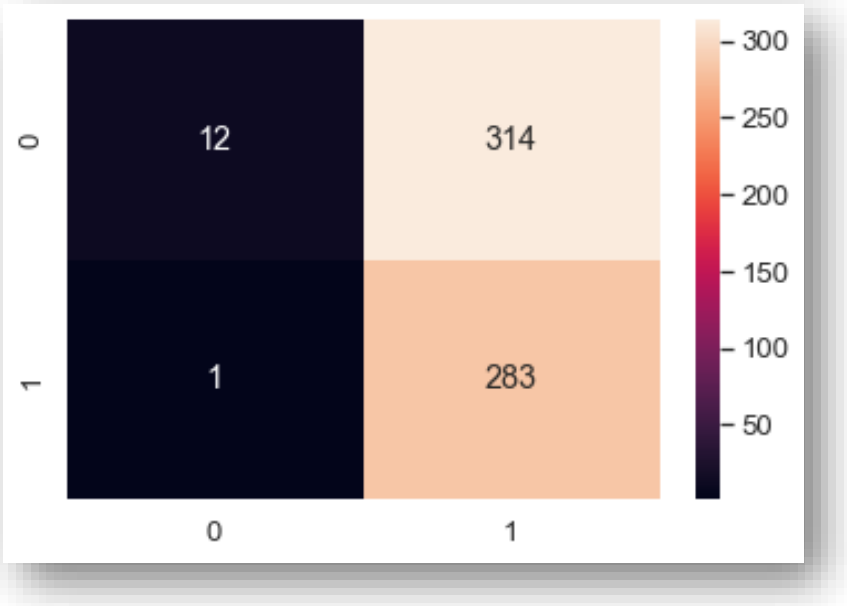
FIGURE 33: AUC CURVE LDA– TRAIN AND TEST DATA



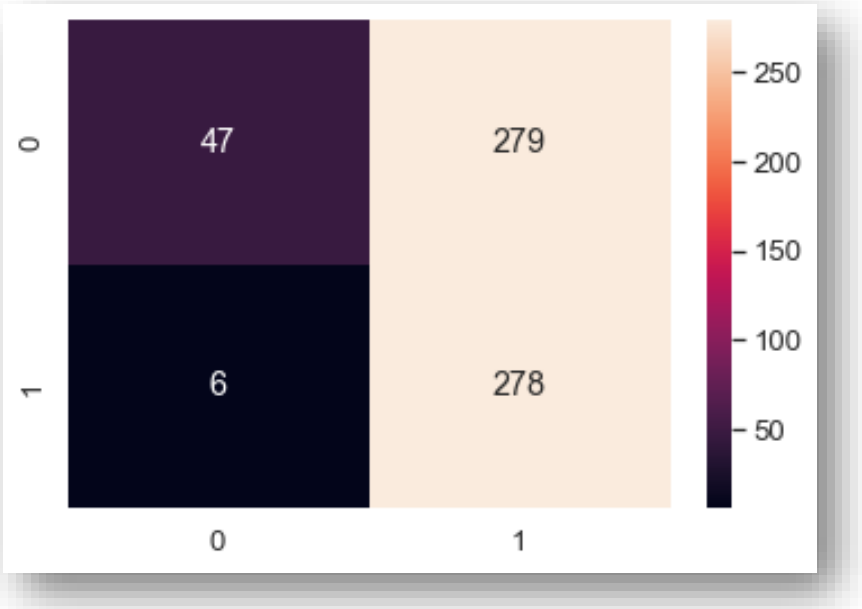
- AUC for the Training Data: 0.739
- AUC for the Test Data: 0.703

FIGURE 34: MAXIMUM ACCURACY TEST– TRAIN DATA

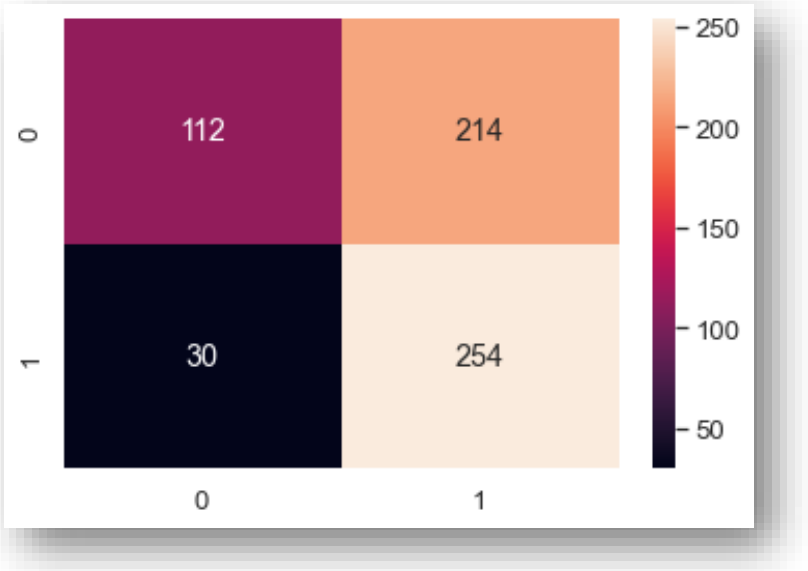
- **Accuracy Score 0.4836, F1 Score 0.6425, Confusion Matrix**



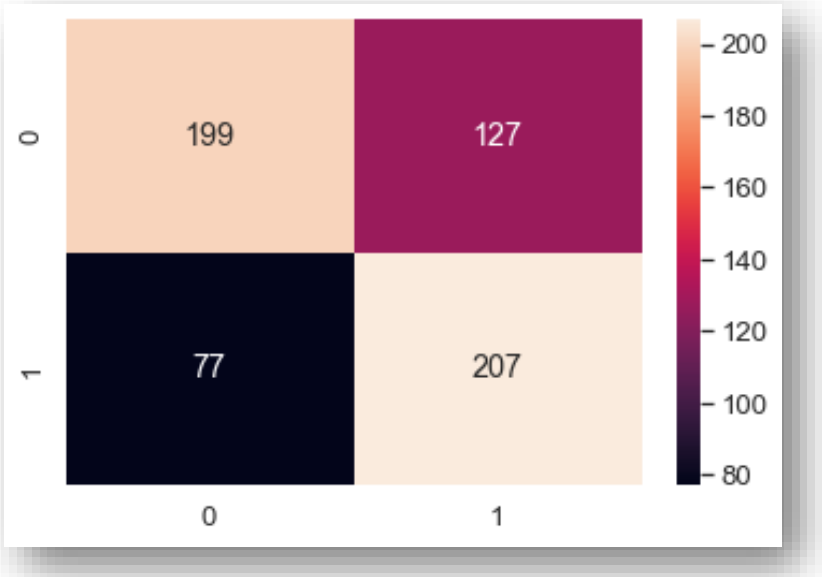
- **0.2 Accuracy Score 0.5328, F1 Score 0.6611, Confusion Matrix**



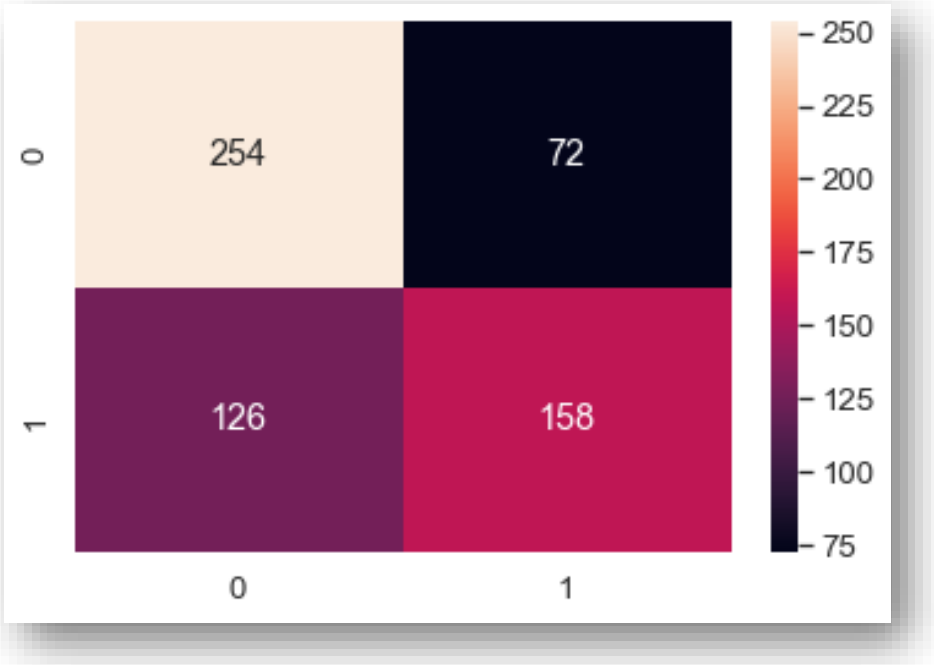
- **0.3 Accuracy Score 0.6, F1 Score 0.6755, Confusion Matrix**



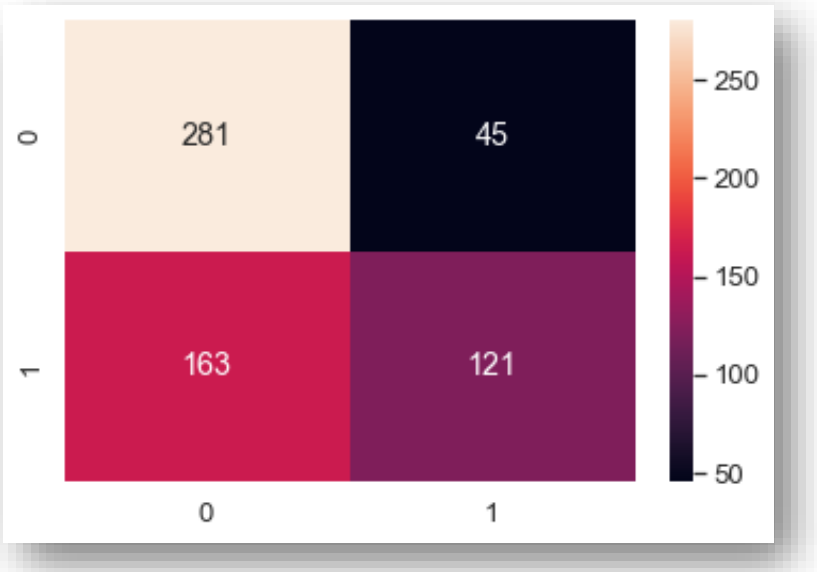
- **0.4 Accuracy Score 0.66, F1 Score 0.67, Confusion Matrix**



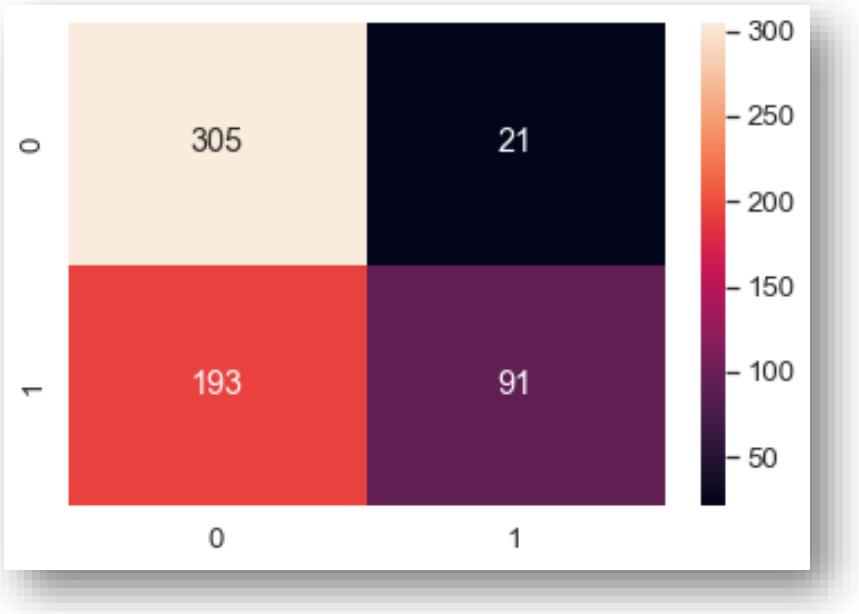
- **0.5 Accuracy Score 0.67, F1 Score 0.61, Confusion Matrix**



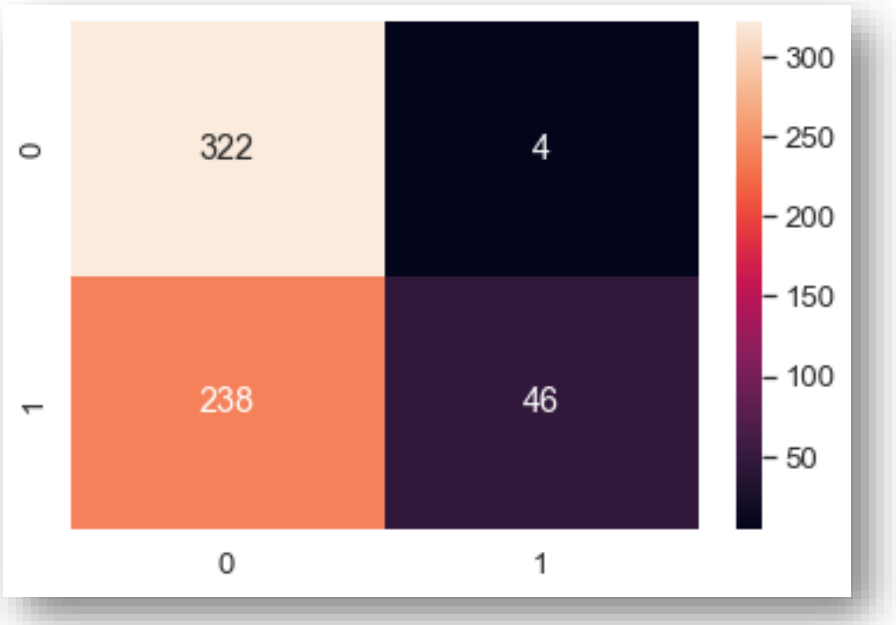
- **0.6 Accuracy Score 0.66, F1 Score 0.54, Confusion Matrix**



- **0.7 Accuracy Score 0.65, F1 Score 0.46, Confusion Matrix**



- **0.8 Accuracy Score 0.60, F1 Score 0.27, Confusion Matrix**



- **0.9 Accuracy Score 0.54, F1 Score 0.04, Confusion Matrix**

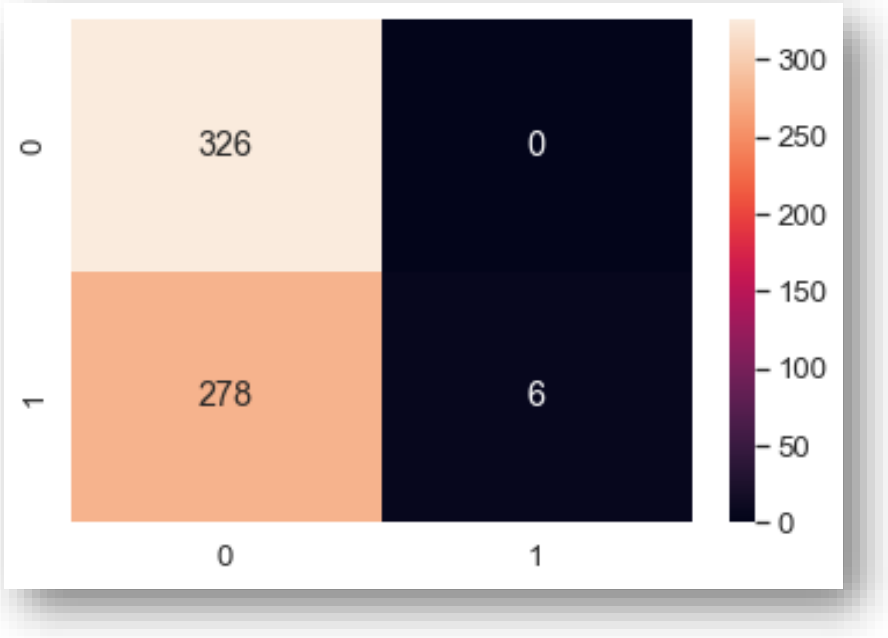
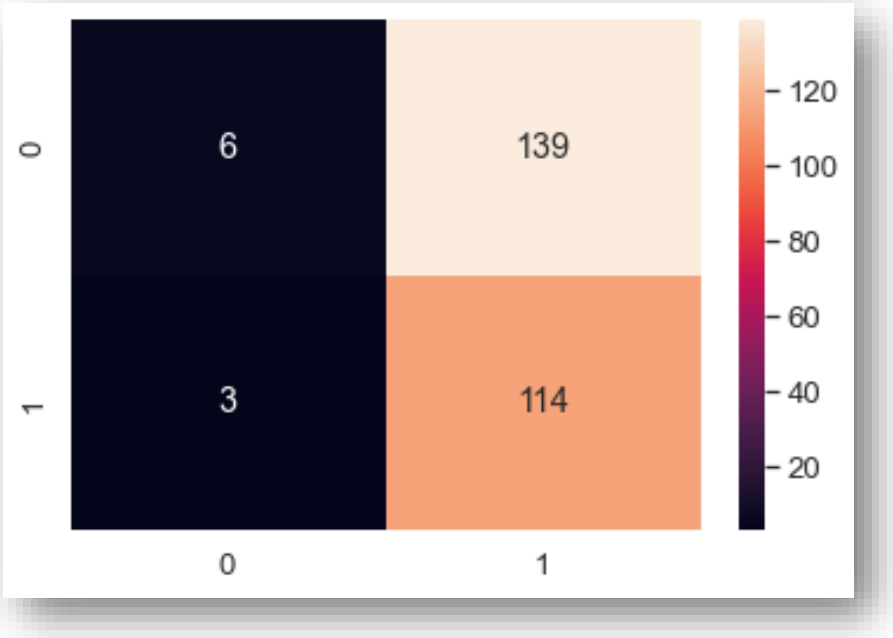
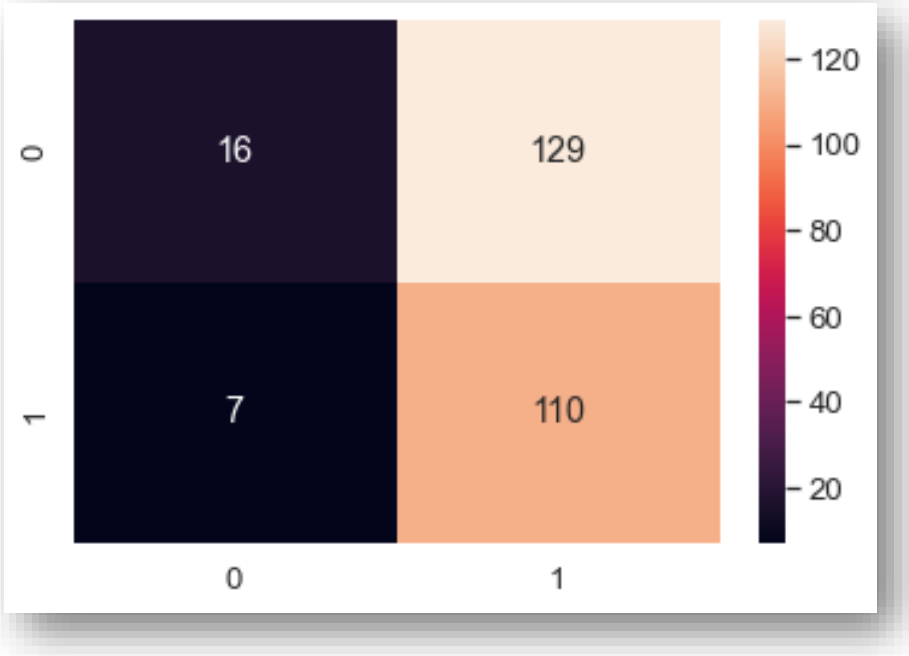


FIGURE 35: MAXIMUM ACCURACY TEST– TEST DATA

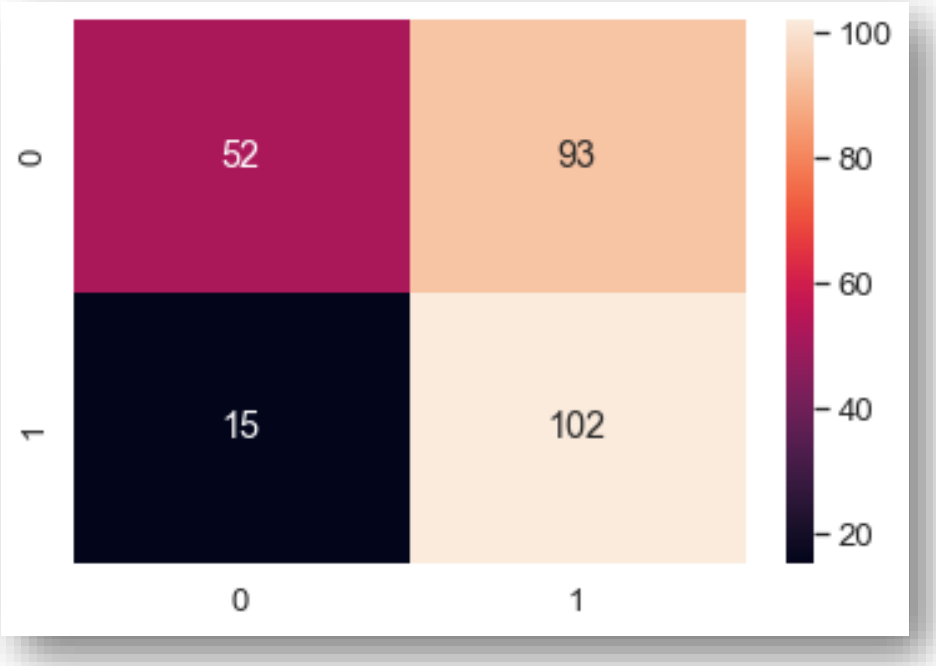
- **0.1 Accuracy Score 0.45, F1 Score 0.61, Confusion Matrix**



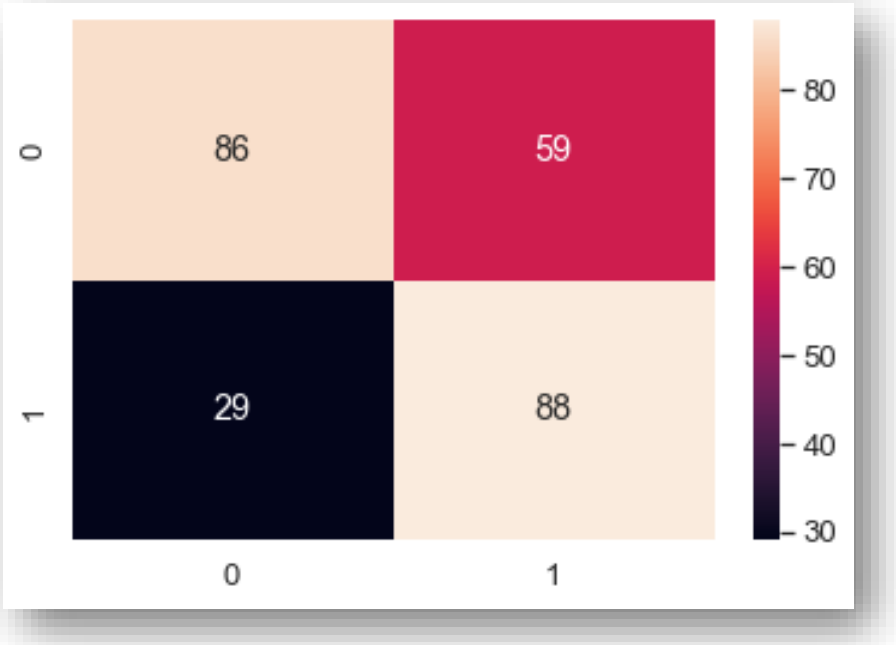
- **0.2 Accuracy Score 0.48, F1 Score 0.62, Confusion Matrix**



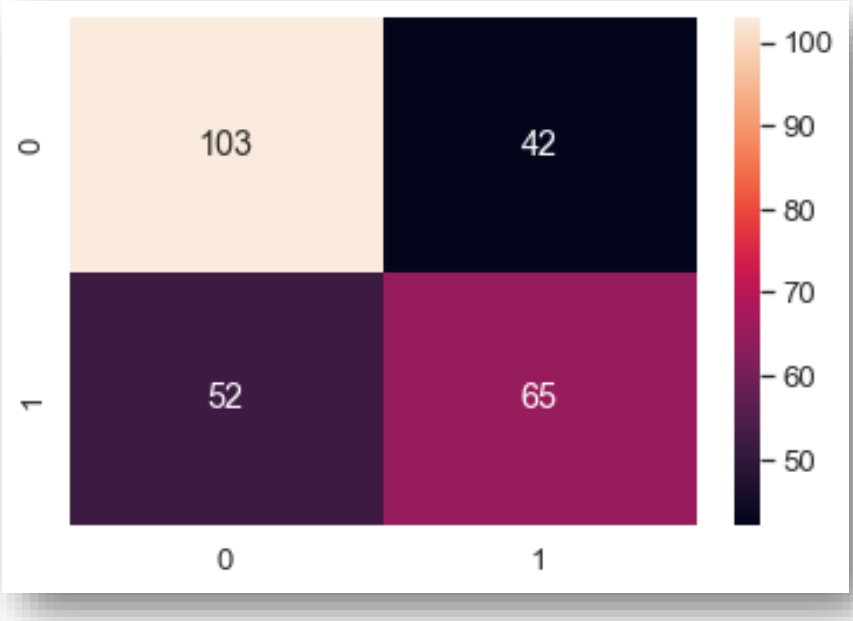
- **0.3 Accuracy Score 0.58, F1 Score 0.65, Confusion Matrix**



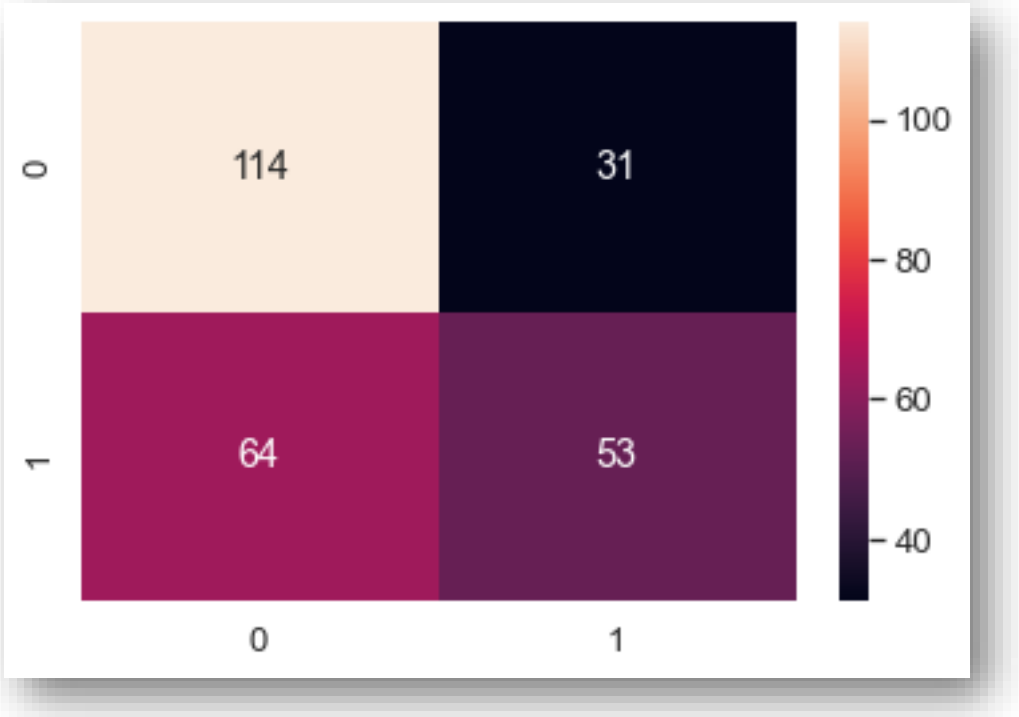
- **0.4 Accuracy Score 0.66, F1 Score 0.66, Confusion Matrix**



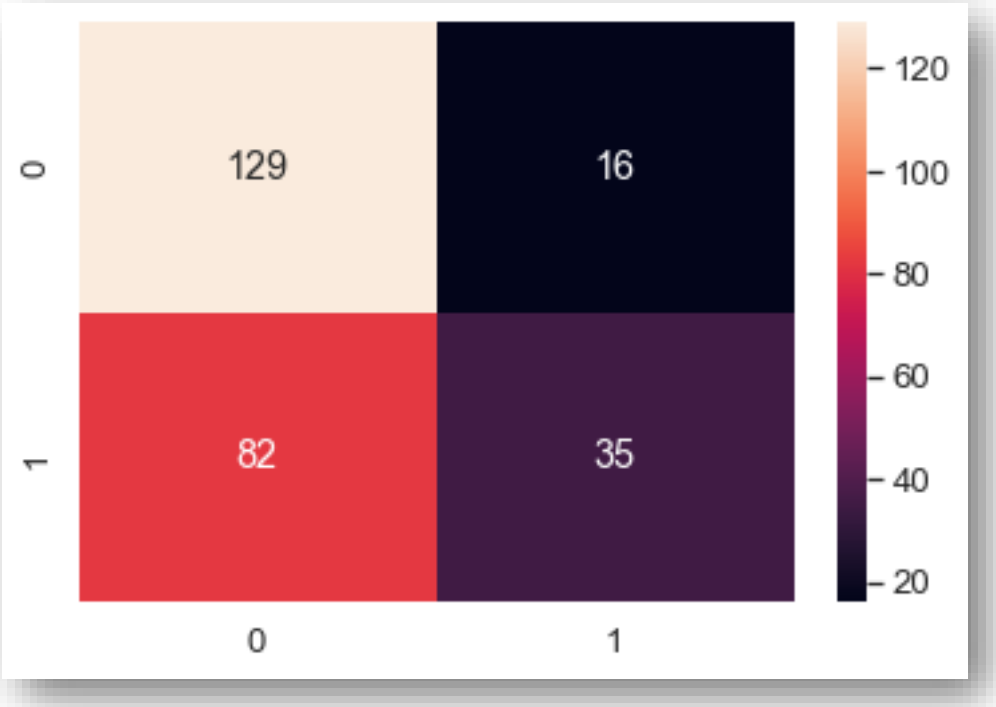
- **0.5 Accuracy Score 0.64, F1 Score 0.58, Confusion Matrix**



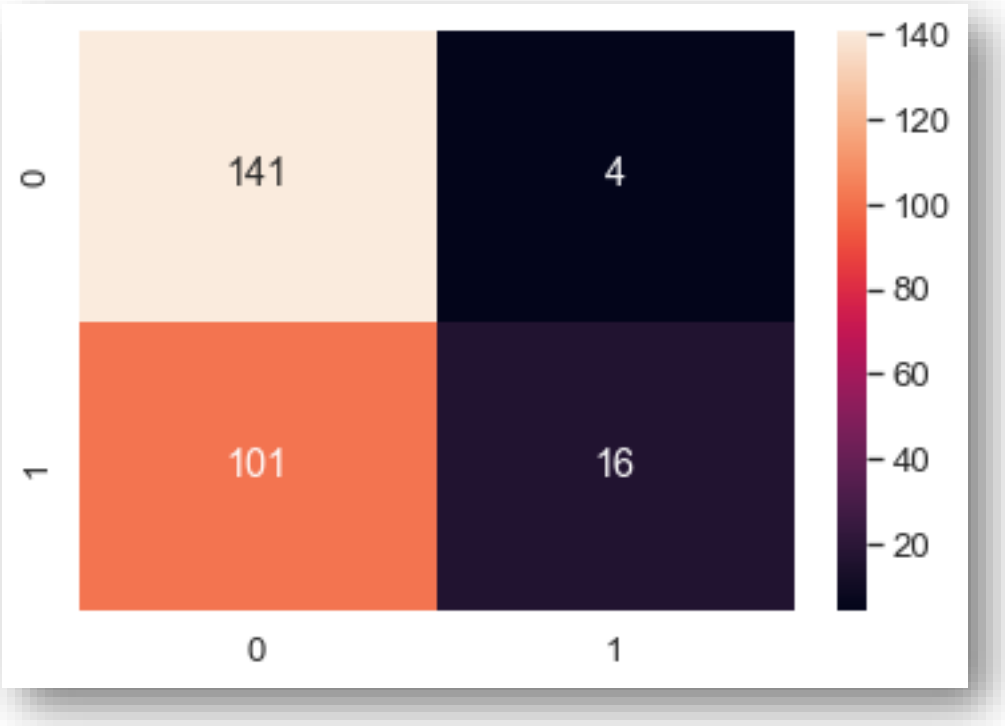
- **0.6 Accuracy Score 0.63, F1 Score 0.52, Confusion Matrix**



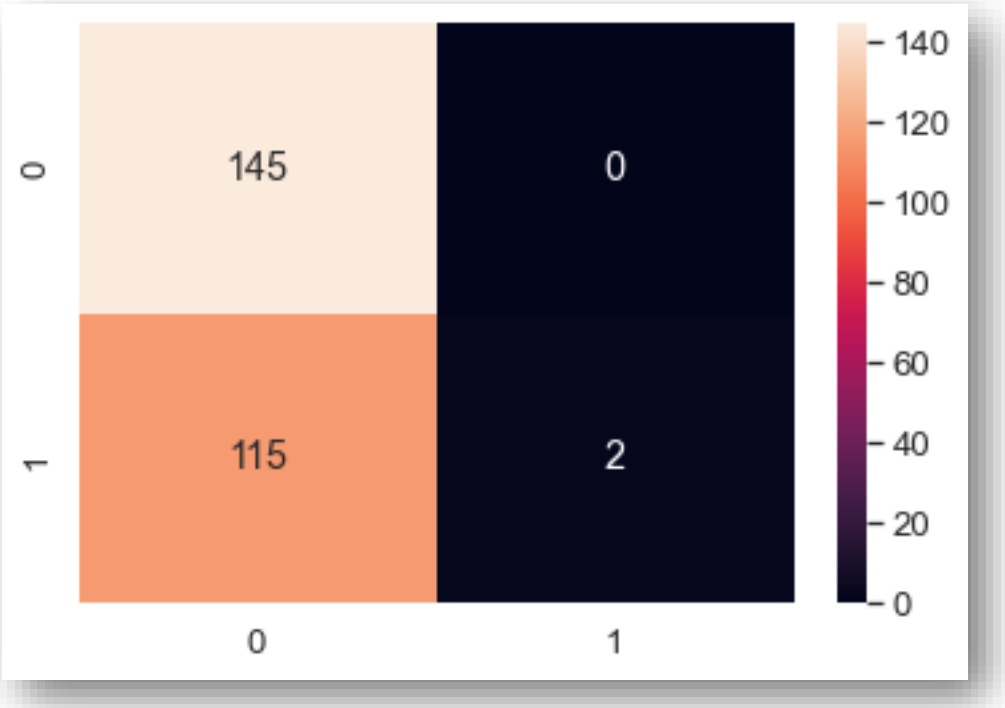
- **0.7 Accuracy Score 0.63, F1 Score 0.42, Confusion Matrix**



- **0.8 Accuracy Score 0.60, F1 Score 0.23, Confusion Matrix**



- **0.9 Accuracy Score 0.56, F1 Score 0.03, Confusion Matrix**



- **INFERENCE FOR Q.2.3.**

1. The accuracy scores aren't too different and can be considered as right fit models avoiding the scenarios of underfit and overfit models.
2. We apply Grid Search CV to see if it helps improve the results. Grid Search CV is a brute force iterative method of obtaining the best model based on a scoring metric provided by us and the parameters provided.
3. The accuracy scores after GRID SEARCH CV are 66% for both train and test data, which are similar to before applying Grid Sear CV.
4. The accuracy after applying LDA is 67% and 64% for train and test data respectively.
5. While looking the metrices for both training and the test data, it seems the accuracy scores are same on both models at 66%. Our model is close enough to be treated as a right fit model. The current model is not struggling with being a over fit model or an under fit model.
6. The AUC scores for both the training and test data are also same at 74%.
7. The model performance is good on F1 score as well with training data performing better at 61% while the test data gave a F1 score of 58%.
8. The accuracy score of the training data and test data is same at 66%. This is almost similar to the Logistic Regression model result so far. The AUC scores are marginally lower for the test data, else they are also almost similar to the Logistic Regression model. F1 scores are 61% and 58% for train and test data, respectively, which again is almost close to the logistic regression model.
9. Overall, the model seems to be a right fit model and is staying away from being referred as under fit or over fit model. Let us see if we can refine the results further and improve on the F1 score of the test data specifically.
10. Looking at the classification report for the test data, we can see that we have managed to retain our accuracy scores and have been able to improve our F1 score.
11. Logistics and LDA offers almost similar results. While LDA offers flexibility to control or change the important metrices such as precision, recall and F1 score by changing the custom cut-off. In this case study, the moment we changed the cut off, we were able to improve our precision, recall and F1 scores considerably. This is up to the business if they would allow such custom cut off values.

12. Logistics Regression is easier to implement, interpret, and very efficient to train. Also, our dependent variable is following a binary classification and hence it is ideal for us to rely on the logistic regression model to study.

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

1. There is no effect of salary, age, and education on the prediction for Holiday packages. These variables don't impact the decision to opt for holiday packages as it doesn't establish a strong relation of these variables with the target variable.
2. Foreign has emerged as a strong predictor, likelihood of a foreigner opting for a holiday package is high.
3. The probability for no young children variable to opt for holiday packages is very low. Especially at young children at 2.
4. The company should focus on foreigners to drive the sales for holiday packages as that's where the major conversion will happen.
5. The company can try to invest more on marketing efforts or offers towards foreigners for better conversions.
6. The company should also stay away from targeting parents with younger children. As the chances of selling to parents with 2 younger children is probably the lowest. This also proves the fact that parents try and avoid visiting with younger children.
7. If the firm wants to target parents with older children, that still might end up giving favourable return for their marketing efforts than spent on couples with younger children.

THE END