PREDICTIVE MODELLING

Report by
Souravi Sinha

Table of Contents

Table of Figures

Table of Tables

PROBLEM - 1

SUMMARY

CUBIC ZIRCONIA ANALYSIS

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

Data Dictionary for cubic zirconia:

1. **Carat** : Carat weight of the cubic zirconia.
2. **Cut** : Describe the cut quality of the cubic zirconia.Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. **Colour** : Colour of the cubic zirconia.With D being the worst and J the best.
4. **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, SI1, SI2, I1.
5. **Depth** : The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6. **Table** : The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7. **Price** : the Price of the cubic zirconia.
8. **X** : Length of the cubic zirconia in mm.
9. **Y** : Width of the cubic zirconia in mm.
10. **Z** : Height of the cubic zirconia in mm.

PROBLEM 1

I. 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.Sample of the dataset

Exploratory Data Analysis

Sample Data

Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289

Table 1.1

Data types of different variable

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

Table 1.2

Observations¶

1. Dataset has a total of 26967 rows and 11 columns.
2. The fields are of following types : float64(6), int64(2), object(3)
3. There is no missing values, except for the field **depth**.

Checking for duplicates

There is No duplicates present in the dataframe.

Missing data analysis for the variables

Field	Is null present
Unnamed: 0	False
carat	False
cut	False
color	False
clarity	False
depth	True
table	False
x	False
y	False
z	False
price	False

Table 1.3

Depth has missing values present in the data frame.
Unnamed: 0 has all unique values, hence won't be contributing in the analysis, thus can be dropped.

Univariate Analysis

We need to perform univariate analysis on the fields.

Univariate Analysis Numerical fields using Histogram

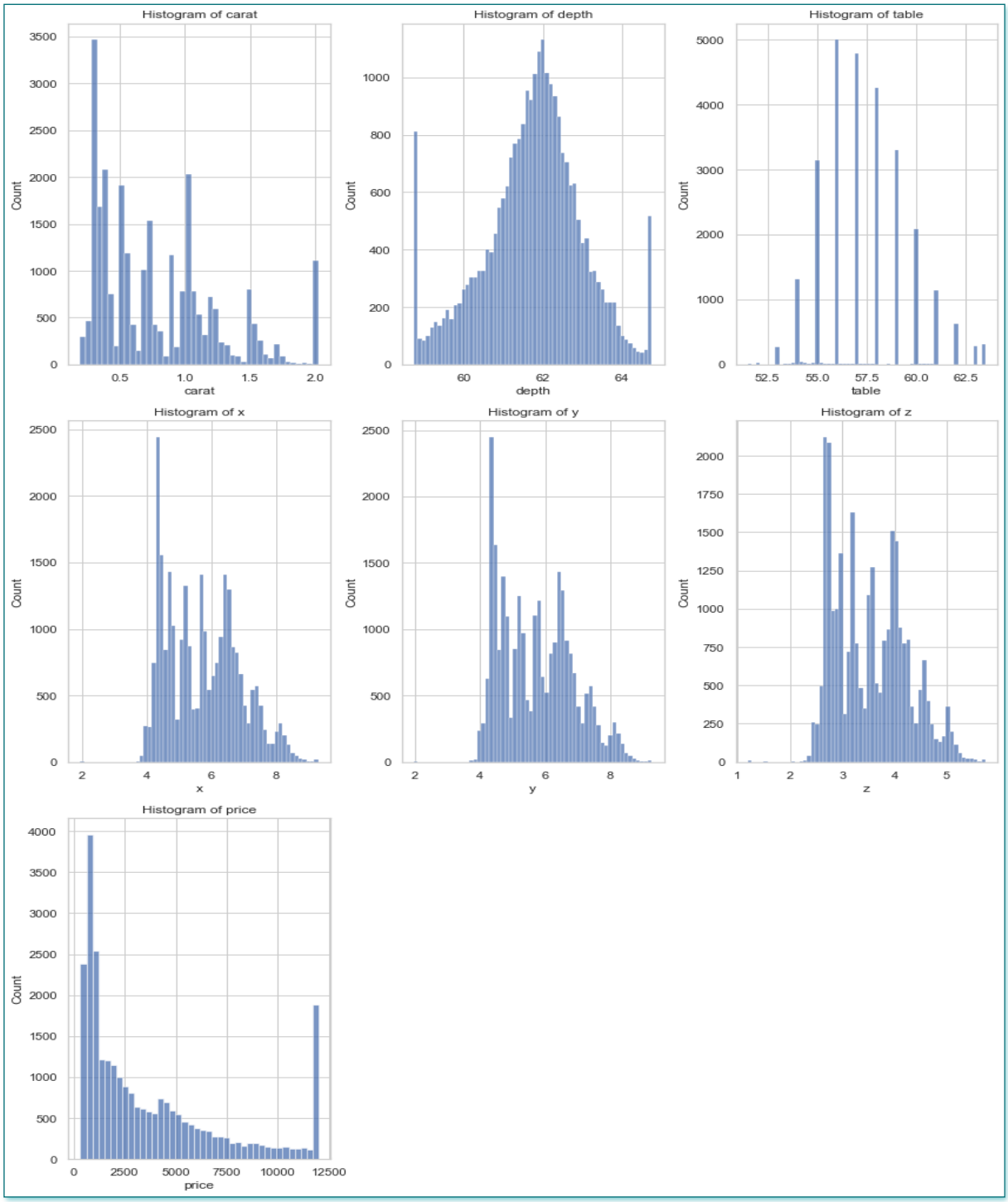

Figure 1.1 - Histogram of fields

Observations

Positively Skewed

- Skewness of carat is 1.116481
- Skewness of table is 0.765758
- Skewness of price is 1.618550

Bimodal or Multimodal Distribution

- Skewness of x is 0.387986
- Skewness of y is 3.850189
- Skewness of z is 2.568257

Approximately Normally distributed

- Skewness of depth is -0.028618

Univariate Analysis Categorical fields using Count Plot

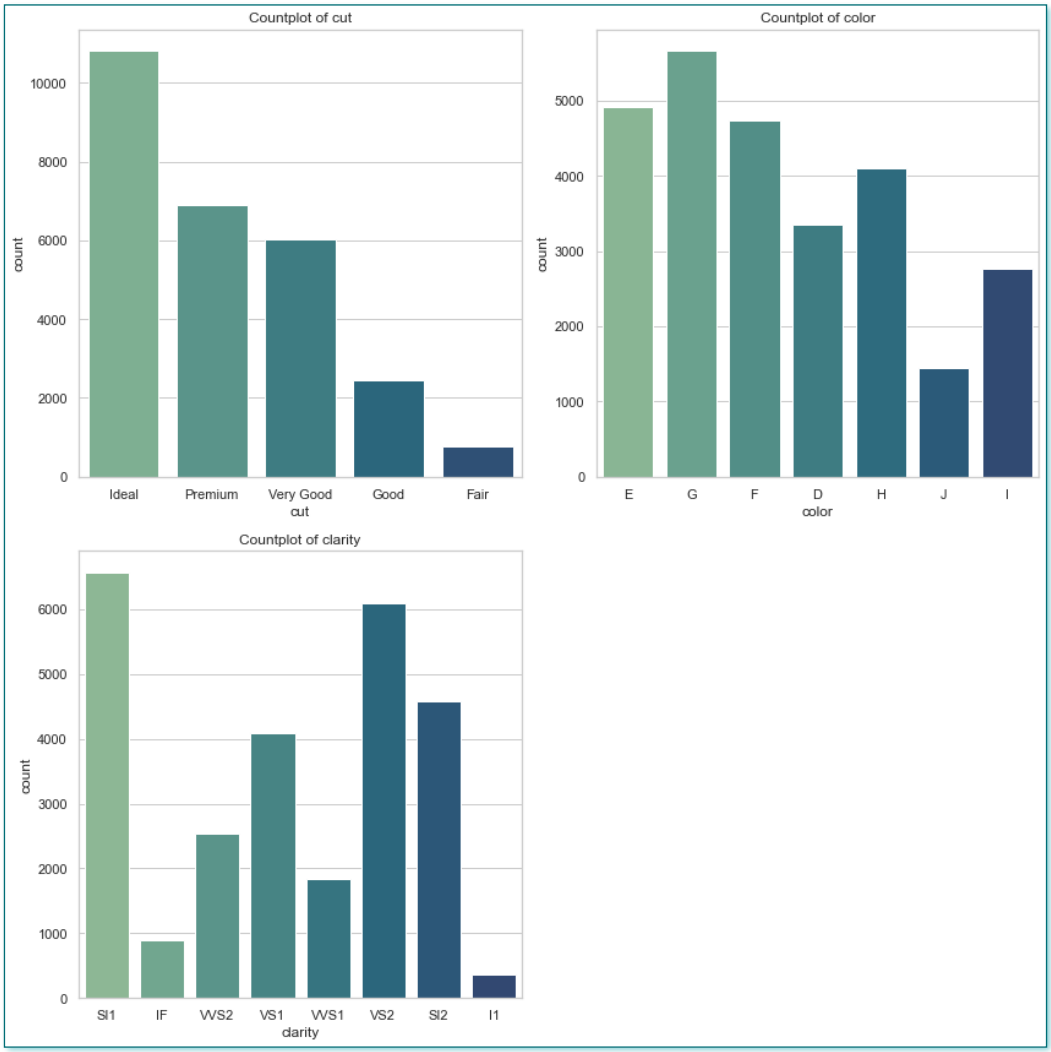

Figure 1.2 - Count plot of the fields

Observations

1. **Cut** : Maximum number of entries are present for *Ideal* and minimum number of entries are present for *Fair*.
2. **Color** : Maximum number of entries are present for *G* and minimum number of entries are present for *J*. But it seems to be a somewhat balanced category.
3. **Clarity** : Maximum number of entries are present for *SI1* and minimum number of entries are present for *I1*.

Bivariant Analysis¶

This can be done both by Correlation Heat Map and PairPlot

HeatMap

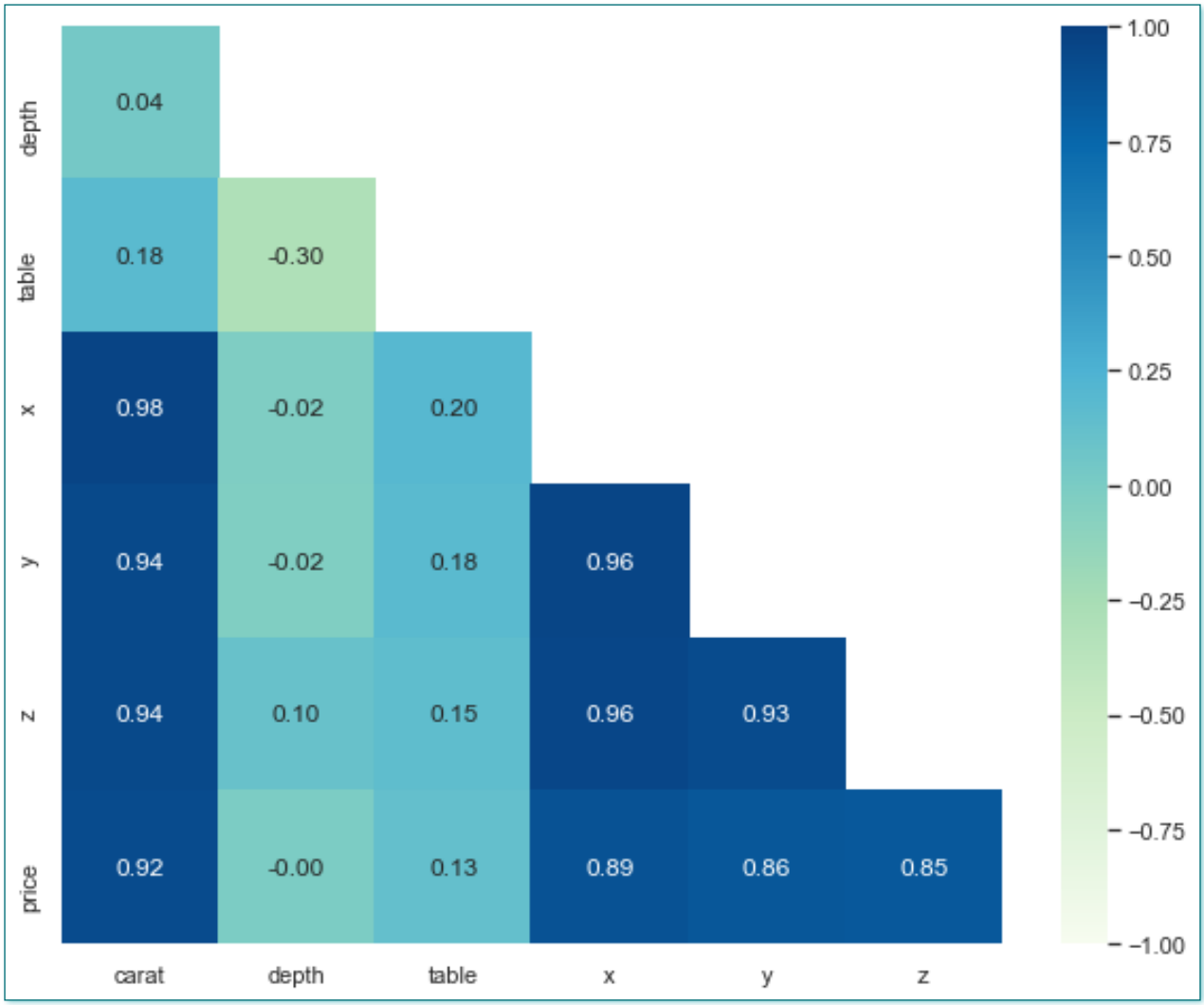

Figure 1.3 - HeatMap

Observations

1. Strong positive correlation can be observed in among many fields like-
    1. Carat to price, x, y, z and price.
    2. X to price, z and y
    3. Y to price and z
    4. Z to price

PairPlot

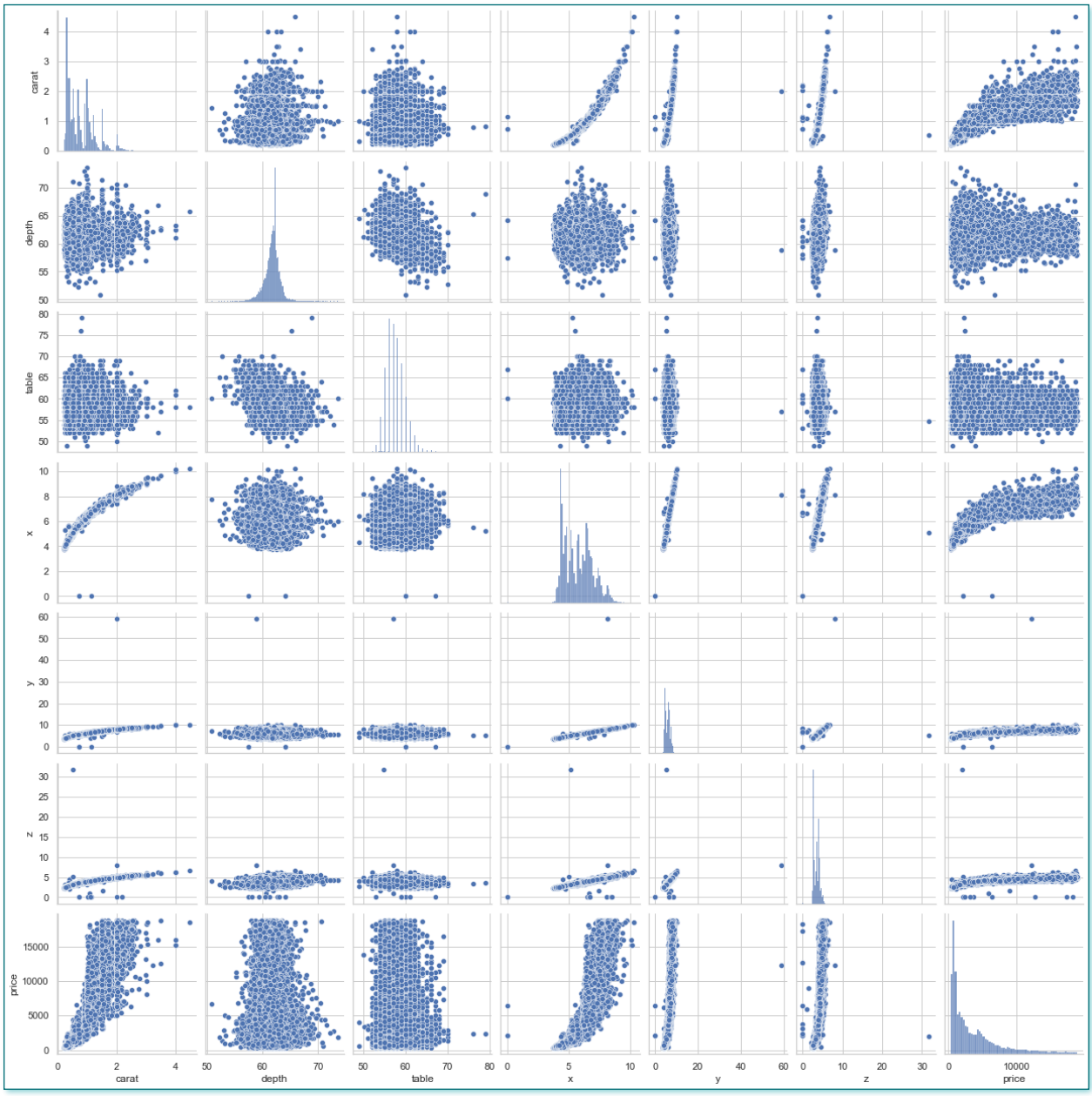

Figure 1.4 - Pairplot

II. 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.

For treating the null values, we will have to check the nulls again-

Checking for null values

Field	Number of nulls
carat	0
cut	0
color	0
clarity	0
depth	697
table	0
x	0
y	0
z	0
price	0

Table No. 2.1

Observation

1. Null values are present only in one of the fields i.e. *depth*.
2. Only *2.65%* of depth field has null values so it can be treated.
3. So, the null values in the mentioned field are treated by replacing null with median and mean in 2 different copies of dataframe, in order to check the impact of the method of imputation on the efficacy of the model.

Verifying for null values

Field	Number of nulls
carat	0
cut	0
color	0
clarity	0
depth	0
table	0
x	0
y	0
z	0
price	0

Table No. 2.2

Checking for 0's present in dataframe

1. Count of zeros in column carat is : 0
2. Count of zeros in column cut is : 0
3. Count of zeros in column color is : 0
4. Count of zeros in column clarity is : 0
5. Count of zeros in column depth is : 0
6. Count of zeros in column table is : 0
7. Count of zeros in column x is : 3
8. Count of zeros in column y is : 3
9. Count of zeros in column z is : 9
10. Count of zeros in column price is : 0

As x, y and z are dimensions of the diamond, it cannot be 0. Hence dropping records with x, z or y as 0.

Checking the Object type columns

There are 3 object type fields :-

1. **Cut :**

Cut	Count in %
Ideal	**40.108280**
Premium	25.583120
Very Good	22.360663
Good	9.051804
Fair	2.896132

Table No. 2.3

2. **Color :**

Color	Count in %
G	20.992324
E	18.233396
F	17.536248
H	15.211184
D	12.400341
I	10.275522
J	5.350985

Table No 2.4

3. Clarity :

Clarity	Count in %
SI1	24.366819
VS2	22.616531
SI2	16.965180
VS1	15.177810
VVS2	9.385545
VVS1	6.819446
IF	3.315163
I1	1.353506

Table No. 2.5

Observations

1. **Cut :** From the above counts, we can combine Good (~ 9%) and Fair (~ 2%) because fair constitute of very less percent of data and Fair doesn't mean bad hence can be combined.

2. **Color** *:* From the above counts, we can combine I (~ 10%) and J (~ 5%) as *Others* because J constitute of very less percent of data and hence can be combined.

3. **Clarity** *:* From the above counts, we can combine IF (~ 3%) and I1 (~ 1.3%) as *Others* because IF and I1 constitute of very less percent of data and hence can be combined as Others.

Verifying the Object type columns

1. **Cut :**

Cut	Count in %
Ideal	40.121671
Premium	25.569404
Very Good	22.368128
Good	11.940797

Table No. 2.6

2. **Color :**

Color	Count in %
G	20.988204
E	18.239484
F	17.534684
Others	15.631723
H	15.201424
D	12.404481

Table No. 2.7

3. **Clarity :**

Clarity	Count in %
SI1	24.371244
VS2	22.620372
SI2	16.956006
VS1	15.179168
VVS2	9.388679
VVS1	6.821723
Others	4.662809

Table No. 2.8

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

As the requirement states, multiple models with different combinations of data treatments:

Scaled dataframe with Outliers

The Model is formed in the following steps :

A. **Splitting the Data into train and test datasets for Dataframe with Outliers in 70% of train data and 30% of test data.**
B. **Scaled the data with z-score method. After Scaling the statistic summary of the same is as follows:**

	count	mean	std	min	25%	50%	75%	max
carat	18876.00	0.00	1.00	-1.29	-0.86	-0.20	0.56	2.68
depth	18876.00	0.00	1.00	-2.41	-0.52	0.05	0.61	2.42
table	18876.00	0.00	1.00	-2.75	-0.66	-0.20	0.73	2.81
x	18876.00	0.00	1.00	-3.37	-0.90	-0.03	0.73	3.19
y	18876.00	0.00	1.00	-3.39	-0.91	-0.02	0.72	3.19
z	18876.00	0.00	1.00	-3.39	-0.91	-0.03	0.72	3.19

Table No 3.1

C. **Encoded the data with OrdinalEncoder as:**
- **Cut :** Ideal - 0, Premium - 1, Very Good - 2, Good - 3
- **Color :** G - 0, E - 1, F - 2, H - 3, D - 4, Others - 5
- **Clarity** : SI1 - 0, VS2 - 1, SI2 - 2, VS1 - 3, VVS2 - 4, VVS1 - 5, Others - 6

D. **Ran Linear Regression model on the data.(Statistics are at the end along with other models)**

Normal dataframe with Outliers and Not Scaled

A. **Splitting the Data into train and test datasets for Dataframe with Outliers in 70% of train data and 30% of test data.**
B. **Encoded the data with OrdinalEncoder as:**
   - **Cut :** Ideal - 0, Premium - 1, Very Good - 2, Good - 3
   - **Color :** G - 0, E - 1, F - 2, H - 3, D - 4, Others - 5
   - **Clarity** : SI1 - 0, VS2 - 1, SI2 - 2, VS1 - 3, VVS2 - 4, VVS1 - 5, Others - 6
C. **Ran Linear Regression model on the data.(Statistics are at the end along with other models)**

Normal dataframe without Outliers

A. **Remove the Outliers with the box-plot method:**

Before Outlier Removal

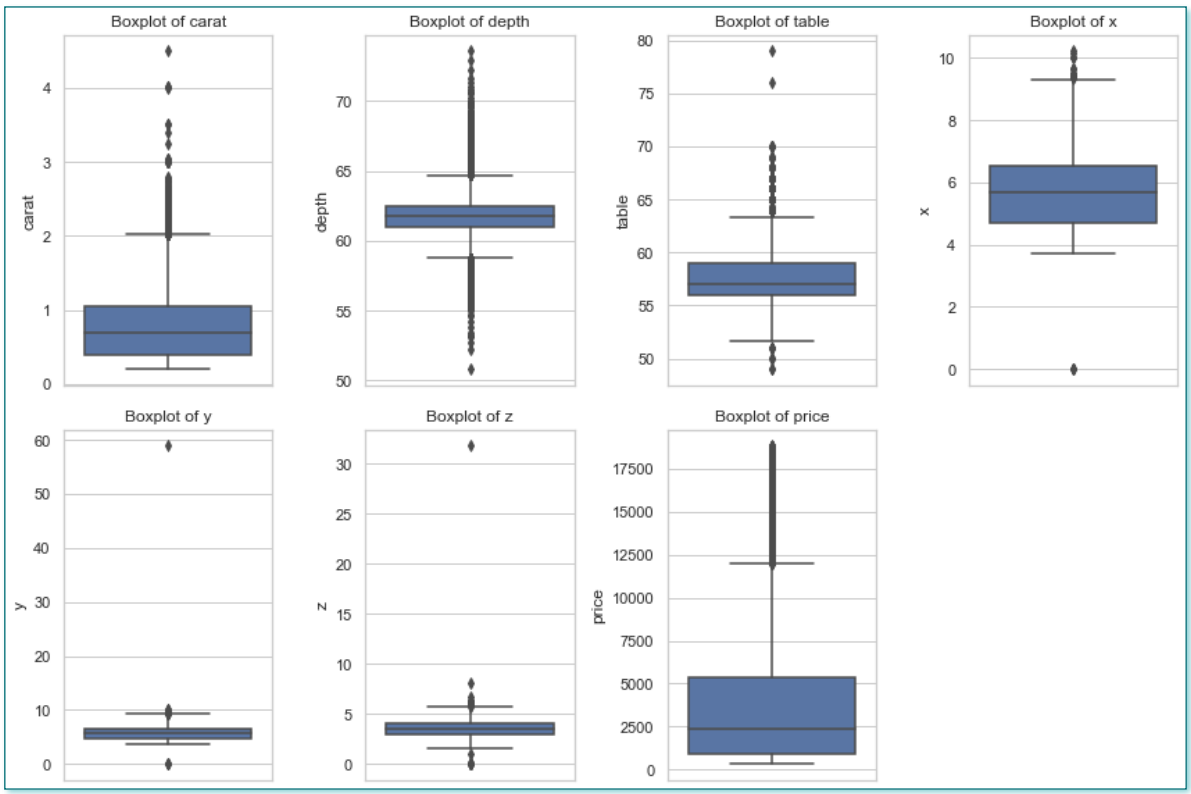

Figure No 3.1

After Outlier Removal

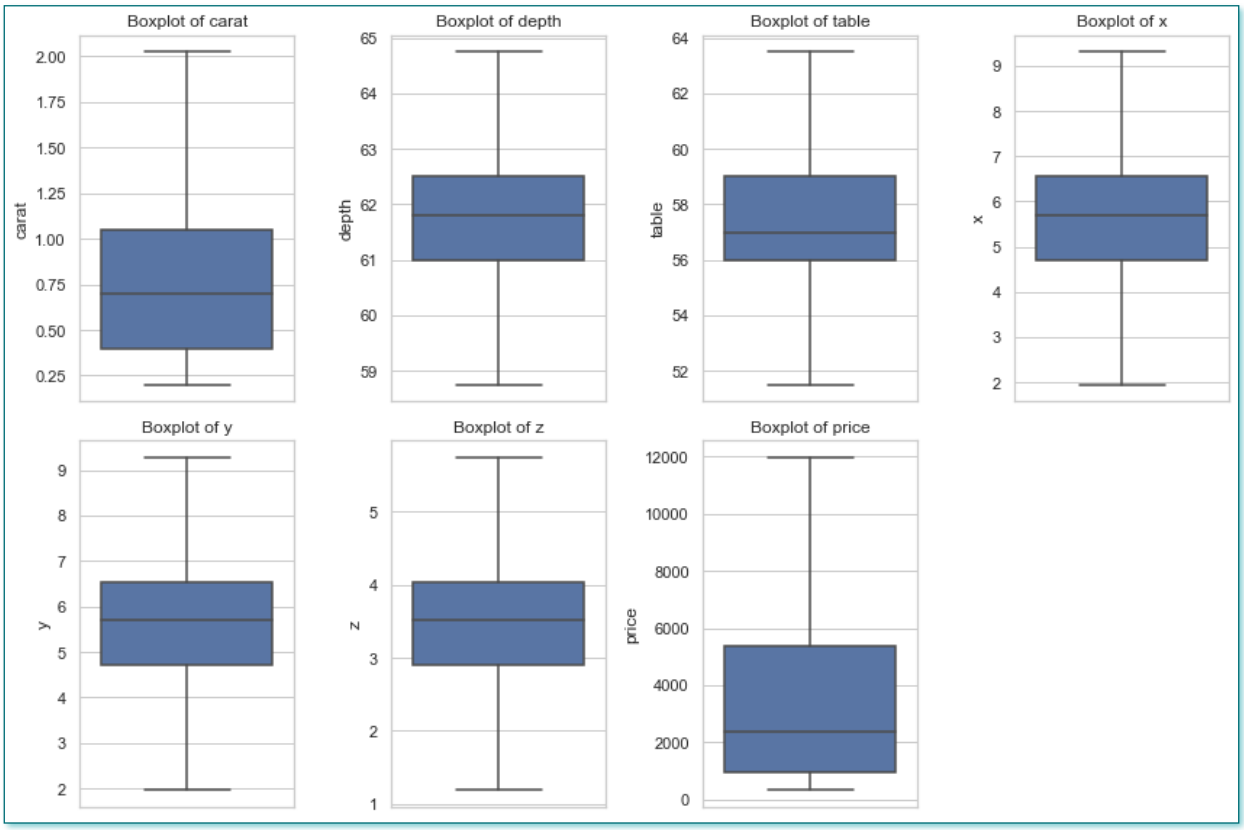

Figure No. 3.2

B. **Splitting the Data into train and test datasets for Dataframe without Outliers in 70% of train data and 30% of test data.**
C. **Encoded the data with OrdinalEncoder as:**
   - **Cut :** Ideal - 0, Premium - 1, Very Good - 2, Good - 3
   - **Color :** G - 0, E - 1, F - 2, H - 3, D - 4, Others - 5
   - **Clarity** : SI1 - 0, VS2 - 1, SI2 - 2, VS1 - 3, VVS2 - 4, VVS1 - 5, Others - 6
D. **Ran Linear Regression model on the data.(Statistics are at the end along with other models)**

Scaled dataframe without Outliers

A. **Splitting the Data into train and test datasets for Dataframe without Outliers from last step in 70% of train data and 30% of test data.**
B. **Scaled the data with z-score method. After Scaling the statistic summary of the same is as follows:**

	count	mean	std	min	25%	50%	75%	max
carat	18876.00	0.00	1.00	-1.29	-0.86	-0.20	0.56	2.68
depth	18876.00	0.00	1.00	-2.41	-0.52	0.05	0.61	2.42
table	18876.00	0.00	1.00	-2.75	-0.66	-0.20	0.73	2.81
x	18876.00	0.00	1.00	-3.37	-0.90	-0.03	0.73	3.19
y	18876.00	0.00	1.00	-3.39	-0.91	-0.02	0.72	3.19
z	18876.00	0.00	1.00	-3.39	-0.91	-0.03	0.72	3.19

Table No 3.2

C. **Encoded the data with OrdinalEncoder as:**
   - **Cut :** Ideal - 0, Premium - 1, Very Good - 2, Good - 3
   - **Color :** G - 0, E - 1, F - 2, H - 3, D - 4, Others - 5
   - **Clarity** : SI1 - 0, VS2 - 1, SI2 - 2, VS1 - 3, VVS2 - 4, VVS1 - 5, Others - 6
D. **Ran Linear Regression model on the data.(Statistics are at the end along with other models)**

Performance Statistics

After running model on the above data, the final comparison is as follows:

	Scaled dataframe with Outliers	Normal dataframe with Outliers and Not Scaled	Normal dataframe without Outliers	Scaled dataframe without Outliers
Model Score with Training data	87.02	87.02	89.34	89.34
Model Score with Test data	87.40	87.40	89.84	89.84
Rsquare	0.870	0.870	0.893	0.893
RMSE	1444.03	1444.03	3077.23	1235.98
Adj Rsquare	0.870	0.870	0.893	0.893

Table No. 3.3

P.S. In the above models, I didn't scale the price.

Scatter Plot B/W predicted values and the test price

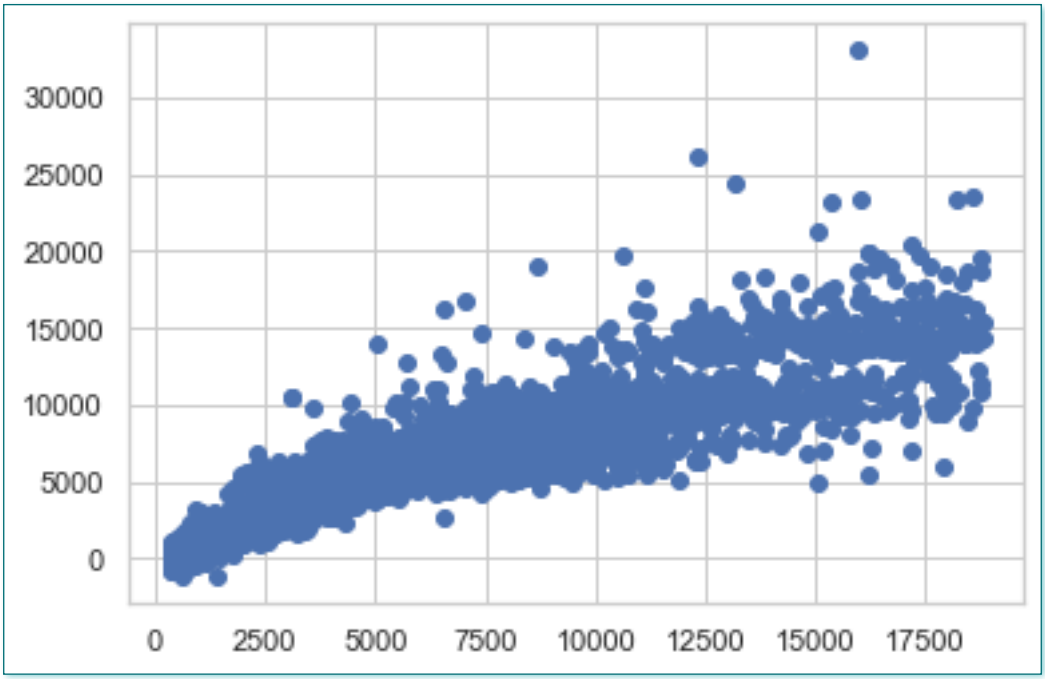

Figure No. 3.3
Scaled dataframe with Outliers

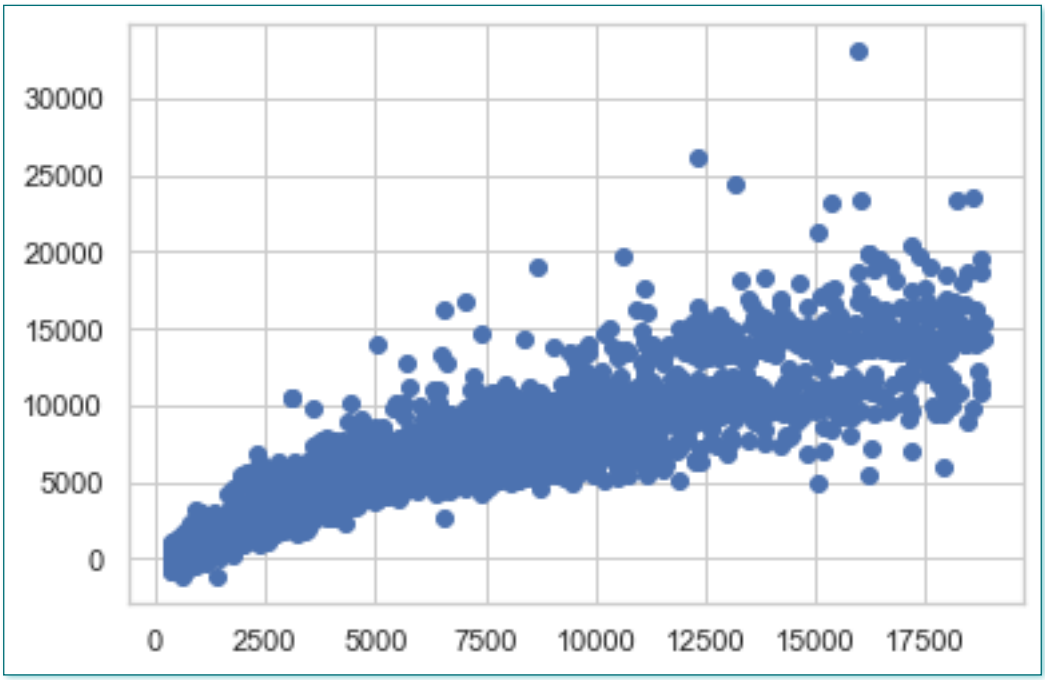

Figure No. 3.4
Normal dataframe with Outliers and Not Scaled

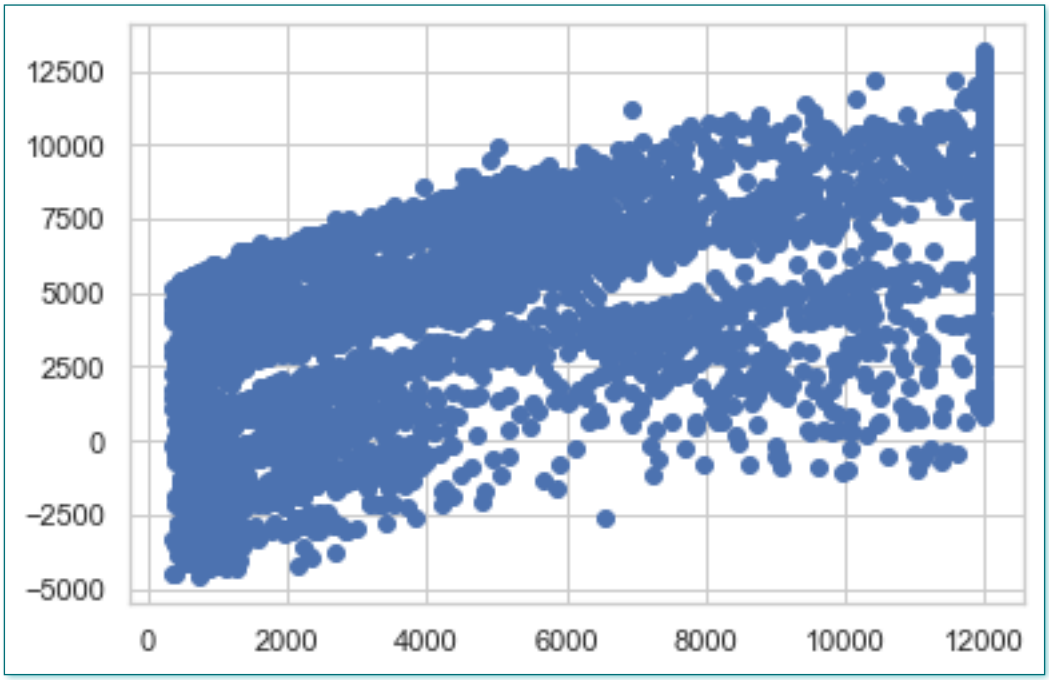

Figure No. 3.5
Normal dataframe without Outliers

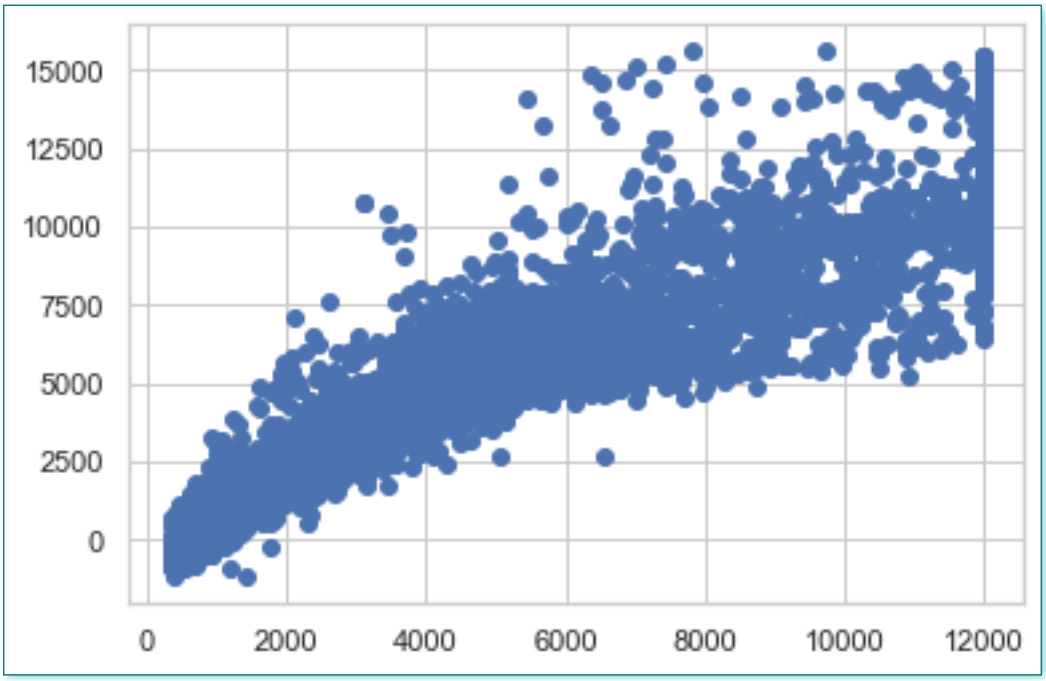

Figure No. 3.6
Scaled dataframe without Outliers

Interpretation

1. Scaling in this data frame, doesn't have any significant effect.
2. Outlier treatment has significant effect on the results.
3. **Best performance is shown by** *Scaled dataframe without Outliers*, **as the RMSE is the least with highest score. Hence I will chose the model on** *Scaled dataframe without Outliers*.

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

In the above model, following steps were performed in this model:

1. Checked for Duplicate Values
2. Dropped Unnamed: 0 field as, it had all unique values, which was unnecessary in our analysis.
3. Removed the Outliers those existed, as we could see many values out of our normal range of values.
4. Imputed the missing values.
5. Combined different sub-labels in different categories like - cut, clarity and color.
6. Scaled the data except for the Target column - "Price"
7. Encoded dataframe's object type columns.
8. Ran the model.

Interpretation of Result

1. Using the chosen model, the price of diamond with characteristics can be predicted.
2. It can be applied on other products as well.
3. Price can be set more accurately using this model.
4. Using cut, clarity, dimension, color etc, correctly price can be predicted.

PROBLEM - 2

SUMMARY

HOLIDAY PACKAGE ANALYSIS

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 for holiday package:

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

Table No. 4.1

PROBLEM 2

I. 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.

Exploratory Data Analysis

Sample Data

Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
1	no	48412	30	8	1	1	no
2	yes	37207	45	8	0	1	no
3	no	58022	46	9	0	0	no

Table No. 5.1

Data types of different variable

Column	Non-Null Count	Dtype
Unnamed: 0	872 non-null	int64
Holliday_Package	872 non-null	object
Salary	872 non-null	int64
age	872 non-null	int64
educ	872 non-null	int64
no_young_children	872 non-null	int64
no_older_children	872 non-null	int64
foreign	872 non-null	object

Table No. 5.2

Observations¶

1. Dataset has a total of 872 rows and 8 columns.
2. The fields are of following types : int64(6), object(2)
3. There is no missing values.

Checking for duplicates

There is No duplicates present in the dataframe.

Missing data analysis for the variables

Column	Number of nulls
Unnamed: 0	0
Holliday_Package	0
Salary	0
age	0
educ	0
no_young_children	0
no_older_children	0
foreign	0

Table No. 5.3

There are no null values in the Dataframe.
Unnamed: 0 has all unique values, hence won't be contributing in the analysis, thus can be dropped.

Descriptive Data Analysis

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.00	436.50	251.87	1.00	218.75	436.50	654.25	872.00
Salary	872.00	47729.17	23418.67	1322.00	35324.00	41903.50	53469.50	236961.00
age	872.00	39.96	10.55	20.00	32.00	39.00	48.00	62.00
educ	872.00	9.31	3.04	1.00	8.00	9.00	12.00	21.00
no_young_children	872.00	0.31	0.61	0.00	0.00	0.00	0.00	3.00
no_older_children	872.00	0.98	1.09	0.00	0.00	1.00	2.00	6.00

Table No 5.4

Univariate Analysis

We need to perform univariate analysis on the fields.

Univariate Analysis Numerical fields using Histogram

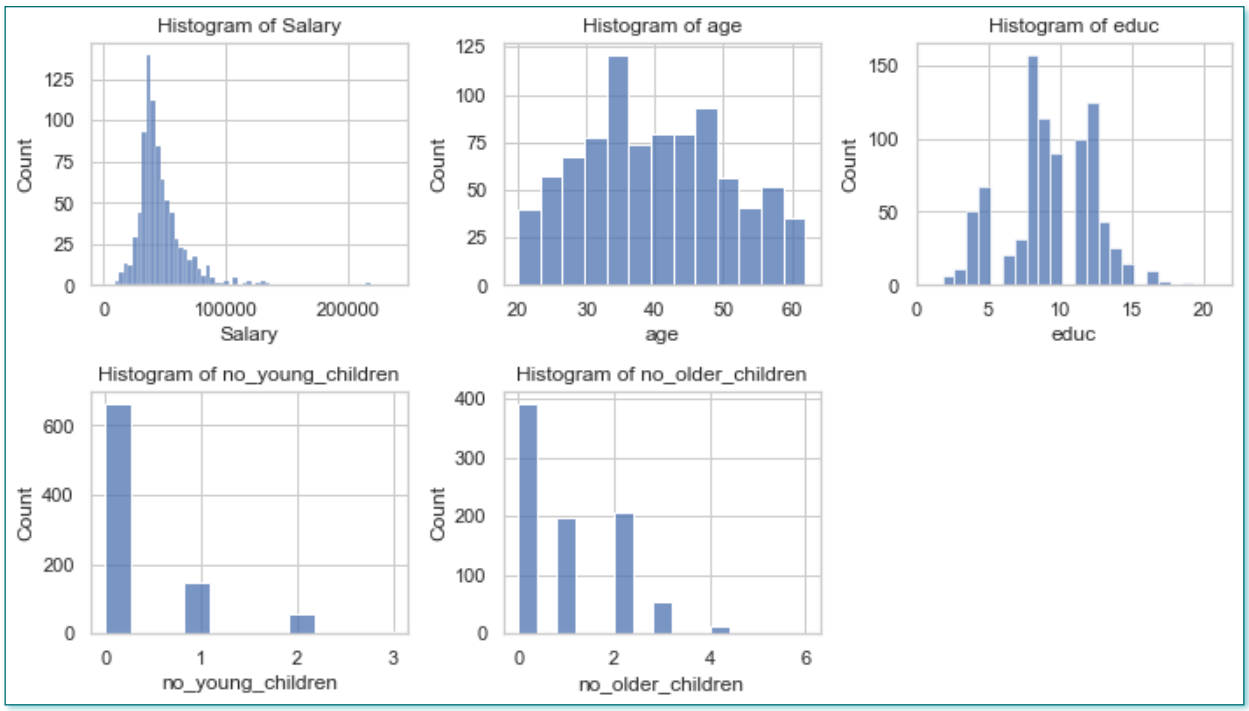

Figure No. 5.1

Observations

Positively Skewed

- Skewness of Salary is 3.10
- Skewness of no_young_children is 1.95
- Skewness of no_older_children is 0.95

Approximately Bimodal distributed

- Skewness of educ is -0.05
- Skewness of age is 0.15

Univariate Analysis Categorical fields using Count Plot

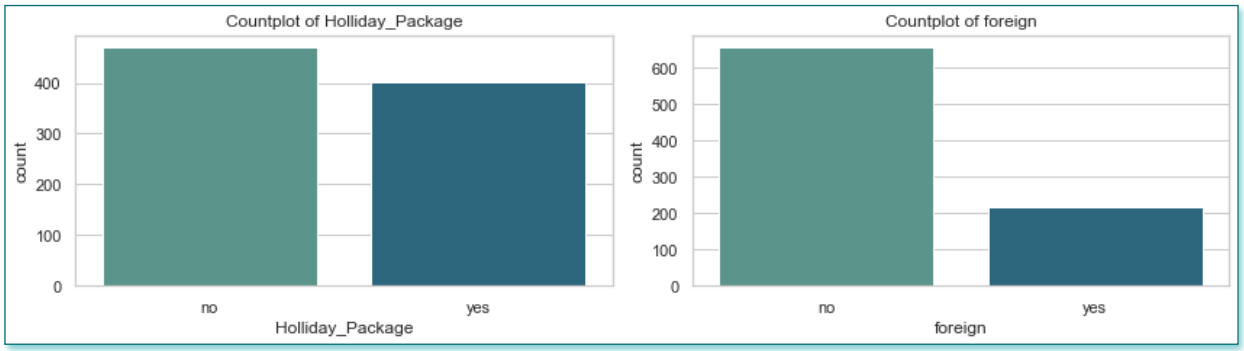

Figure No.5.2

Observations

1. **Holliday_Package** : Maximum number of entries are present for *No then Yes*.
2. F**oreign** : Maximum number of entries are present for *No then Yes*.

Bivariant Analysis¶

This can be done both by Correlation Heat Map and PairPlot

HeatMap

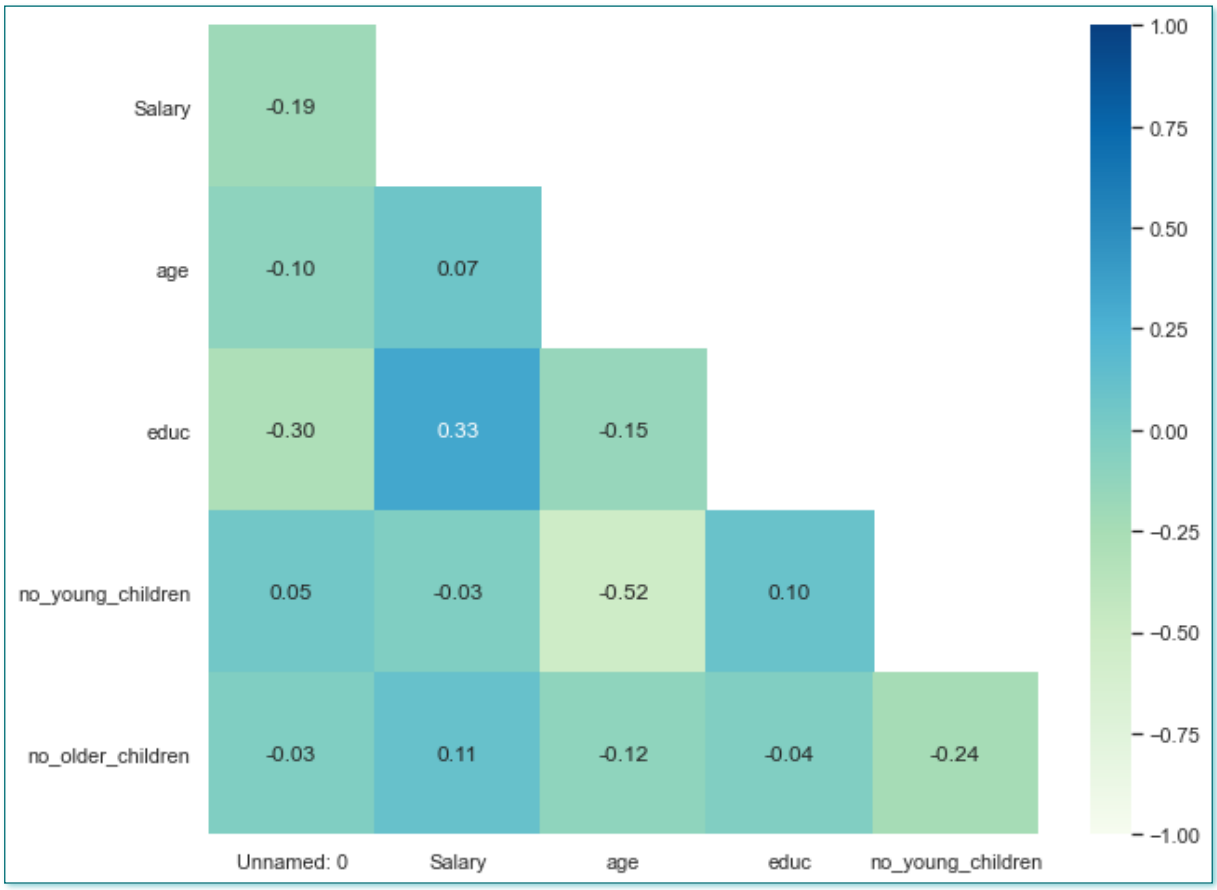

Figure No. 5.3

Observations

There is no strong correlation among the fields.

Pairplot

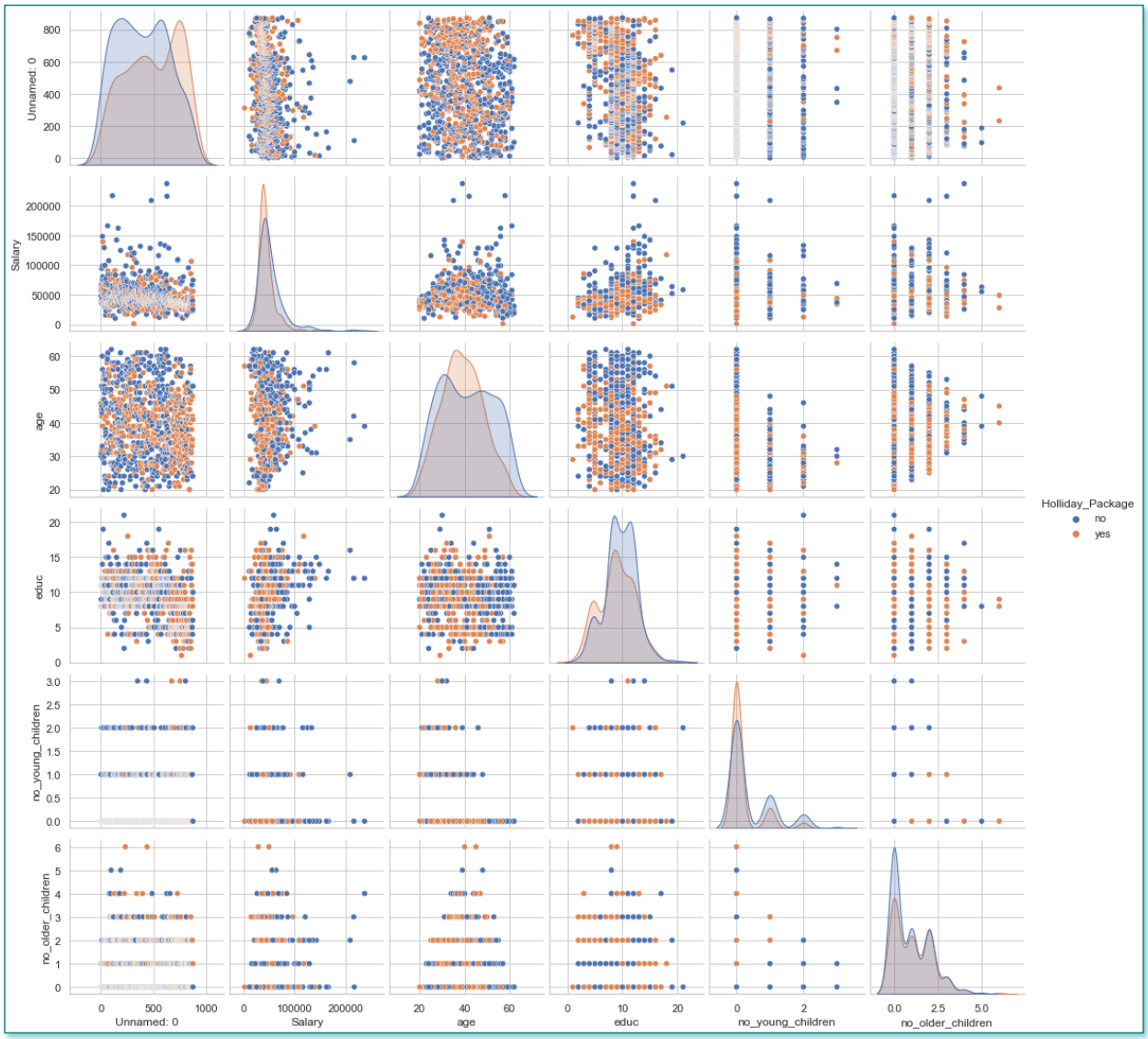

Figure No. 5.4

Observations

There is no strong covariance among the fields.

Check for Outliers using Box-plots

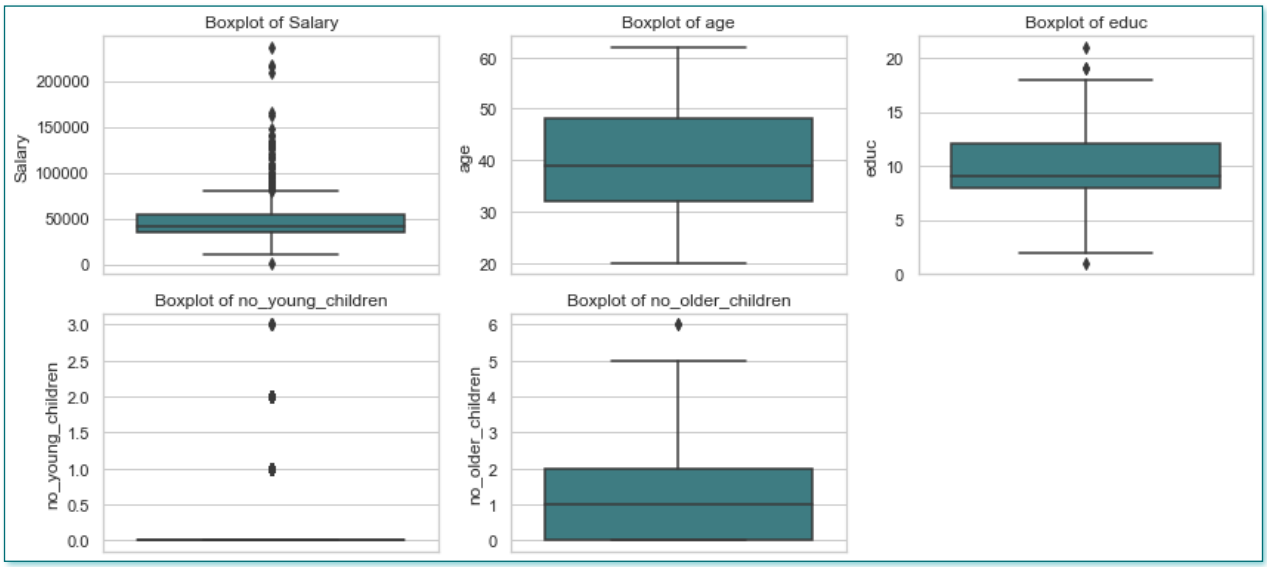

Figure No. 5.5

Observations

1. Outliers are present in Salary, educ, no_young_children and no_older_children.
2. I won't go for the outlier treatments

II. 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).

Encoding of the Data

I have encoded the data in fields Holliday_Package and foreign using OrdinalEncoder. After encoding the sample data seems like:

	Holliday_ Package	Salary	age	educ	no_young _children	no_older _children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0

Table No. 6.1

Holliday_Package	Value Counts
0	471
1	401

Table No. 6.2

foreign	Value Counts
0	656
1	216

Table No. 6.3

Data Splitting

I have split the data into 70-30%

Model Execution

1. Executed Logistic Regression model on dataframe.
2. Executed LDA model on dataframe.

III. 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/optimised.

Logistic Regression Model and LDA Stats

	Logistic Regression Model Stats	LDA Model Stats
Model Score with Training data	0.5442	0.6672
Model Score with Test data	0.5305	0.6603
AUC on Training data	0.602	0.721
AUC on Test data	0.608	0.746

Table No. 7.1

Logistic Regression Model Confusion Metrics

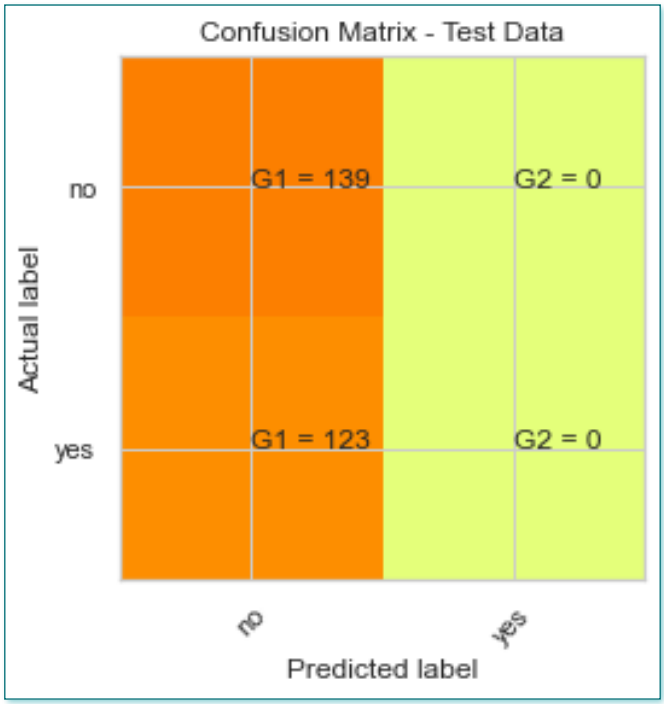

Figure No. 6.1

Logistic Regression Model Classification Report

	precision	recall	f1-score	support
0	0.53	1.00	0.69	139
1	0.00	0.00	0.00	123
accuracy			0.53	262
macro avg	0.27	0.50	0.35	262
weighted avg	0.28	0.53	0.37	262

Table No. 7.2

Logistic Regression Model ROC curve Training Data

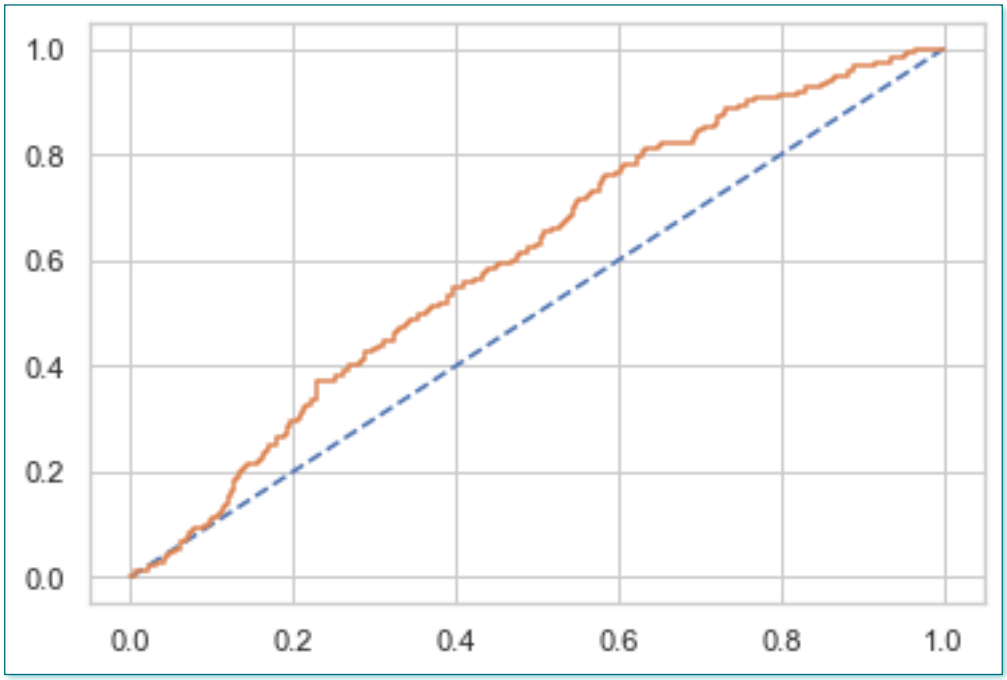

Figure No.6.2
AUC = 0.602

Logistic Regression Model ROC curve Test Data

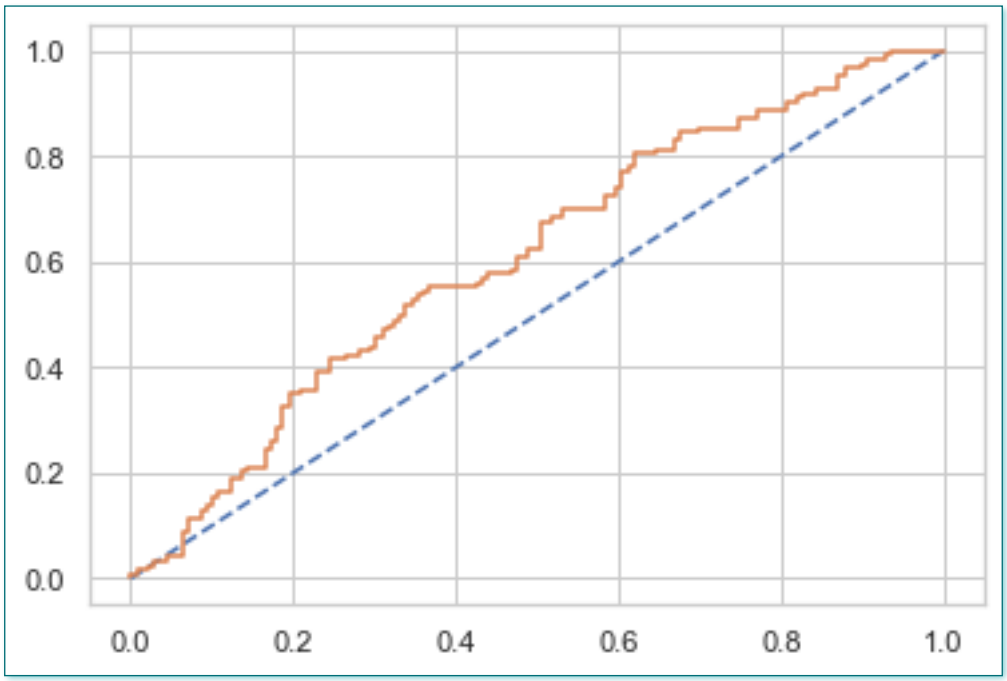

Figure No. 6.3
AUC = 0.608

LDA Model Confusion Metrics

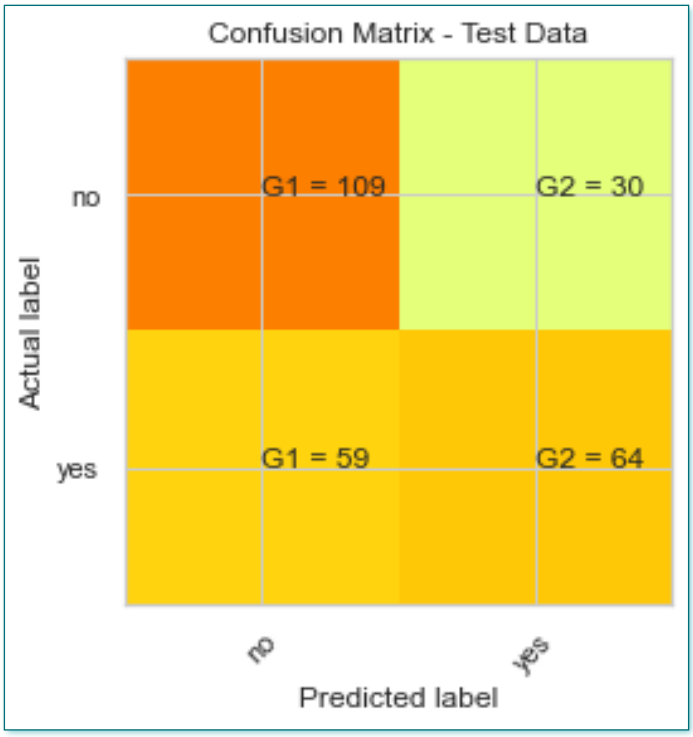

Figure No. 6.4

LDA Model Classification Report

	precision	recall	f1-score	support
0	0.65	0.78	0.71	139
1	0.68	0.52	0.59	123
accuracy			0.66	262
macro avg	0.66	0.65	0.65	262
weighted avg	0.66	0.66	0.65	262

Table No. 7.3

LDA Model ROC curve Training Data

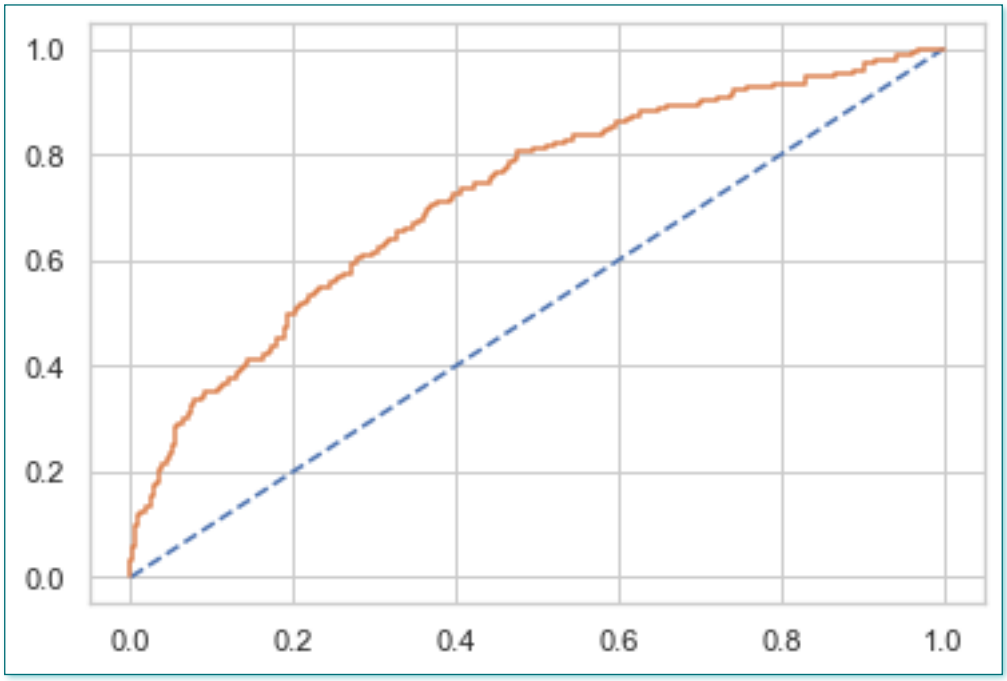

Figure No. 6.5
AUC : 0.721

LDA Model ROC curve Test Data

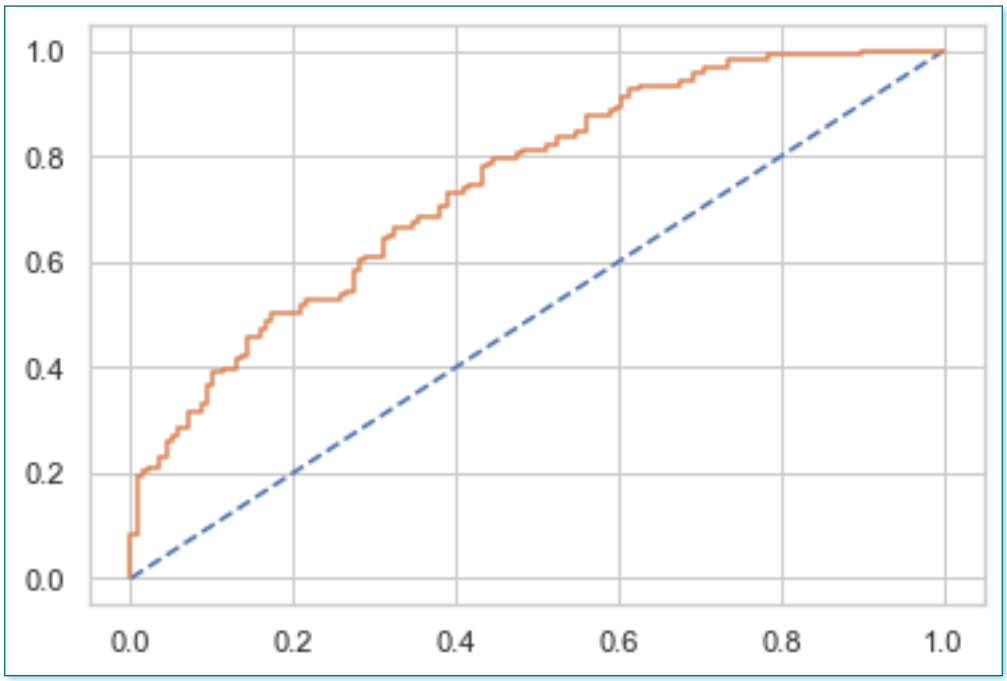

Figure No. 6.6
AUC : 0.746

Interpretation

In this, I will choose LDA model for final execution as it's performance statistics is better i.e. f1 score, precision, accuracy , AUC and ROC curve.

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

In the above model, following steps were performed in this model:

1. Checked for Duplicate Values, none present.
2. Checked for null values present in the dataset.None present.
3. Dropped Unnamed: 0 field as, it had all unique values, which was unnecessary in our analysis.
4. Checked for Outliers and but did not remove the Outliers.
5. Checked for the missing values. None present.
6. Encoded dataframe's object type columns.
7. Ran the model.
8. Executed steps to get stats on the performance of both the models.
9. Based on the performance statistics, I chose LDA model.

Interpretation of Result

Using the chosen model, we can predict whether employee will opt for the package or not