PREDICTIVE MODELLING PROJECT

SABITA NAIR PANCHAL
DSBA - BATCH FEBRUARY, 2021

SUBMISSION DATE - 01/08/2021

PROBLEM 1: LINEAR REGRESSION

Executive Summary:

Gem Stones co ltd, which is a cubic zirconia manufacturer, which is an inexpensive diamond alternative with many of the same qualities as a diamond. They earn different profits on different prize slots of the products. Using a dataset containing the prices and other attributes of almost 27,000 cubic zirconia, the price for the stone will be predicted, so that the company can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, the best 5 attributes that are most crucial to the price prediction process will be identified.

1.1 Exploratory data analysis:

The sample data has 26,967 entries, with values under 10 columns, the first 5 rows are given in the table below:

Table: 1.1

Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
1	0.3	Ideal	E	SI1	62.1	58	4.27	4.29	2.66	499
2	0.33	Premium	G	IF	60.8	58	4.42	4.46	2.7	984
3	0.9	Very Good	E	VVS2	62.2	60	6.04	6.12	3.78	6289
4	0.42	Ideal	F	VS1	61.6	56	4.82	4.8	2.96	1082
5	0.31	Ideal	F	VVS1	60.4	59	4.35	4.43	2.65	779

Data description:

For computational ease, the first column (Unnamed: 0) was dropped as it is merely the serial number, and will not give any insights. A brief description of the column heads:

Table: 1.2

Vaiable	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.
Color	Colour of the cubic zirconia.With D being the best and J the worst.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
Depth	The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle 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.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.

Checking datatypes:

The data types of the dataset was checked. The following data types are present in the dataset:

Table: 1.3

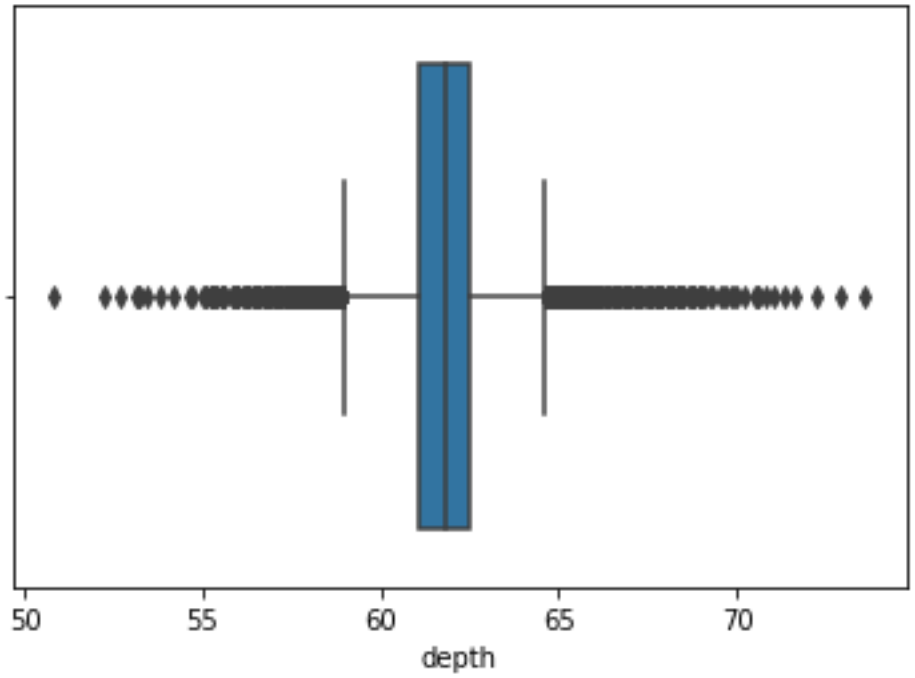
Data type	Variables
float64	carat', 'depth', 'table', 'x', 'y', 'z'
int64	price'
object	cut', 'color', clarity'

Missing value imputation:

The dataset was checked for null and missing values, using the isnull() function. It revealed that the column ‘depth’ was missing 697 values. Generally, few missing values can be deleted, but 697 is a major chunk of the data, so we imputed the missing values, to enable further data processing.

A box plot was used to check the data distribution in column ‘depth’ - it showed presence of plenty of outliers and a slight skew, as shown below:

Figure: 1.1



Since there is a substantial number of outliers and a certain skew, mean was used to impute the missing values.

Statistical summary of the data:

Post imputation of missing values, the statistical summary of the data was checked using the describe() function.

Table: 1.4

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967	26967	26967	26967	26967	26967	26967	26967	26967	26967
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.80	NaN	NaN	NaN	61.75	57.46	5.73	5.73	3.54	3939.52
std	0.48	NaN	NaN	NaN	1.39	2.23	1.13	1.17	0.72	4024.86
min	0.2	NaN	NaN	NaN	50.8	49	0	0	0	326
25%	0.4	NaN	NaN	NaN	61.1	56	4.71	4.71	2.9	945
50%	0.7	NaN	NaN	NaN	61.8	57	5.69	5.71	3.52	2375
75%	1.05	NaN	NaN	NaN	62.5	59	6.55	6.54	4.04	5360
max	4.5	NaN	NaN	NaN	73.6	79	10.23	58.9	31.8	18818

The following inferences can be made by studying the statistical summary:

- The price range of cubic zirconium is Rs. 326 - 18818, which is a very vast range, although most of them cost on an average Rs. 3940. The price also shows very high variance (approx. 4025)
- Close to half the samples are of the highest cut quality (Ideal).
- Majority of the samples belong to color G (mediocre quality).
- In terms of clarity, majority samples belong to SI1 (below average quality).

Checking for duplicate values:

The function duplicated() was used to check for duplicate values - it revealed that there was 34 duplicate rows present in the dataset. Since

the proportion of duplicate rows is very minuscule (34 / 26967 = 0.126 %), we have opted to delete the duplicate rows entirely.

Variable encoding for categorical variables (object type):

In order to convert object type variables to integer values, the columns ‘cut’, ‘color’ and ‘clarity’ were converted to numerical codes using ordinal encoding (so that the ranked order of the variables was kept intact). The following codes were used:

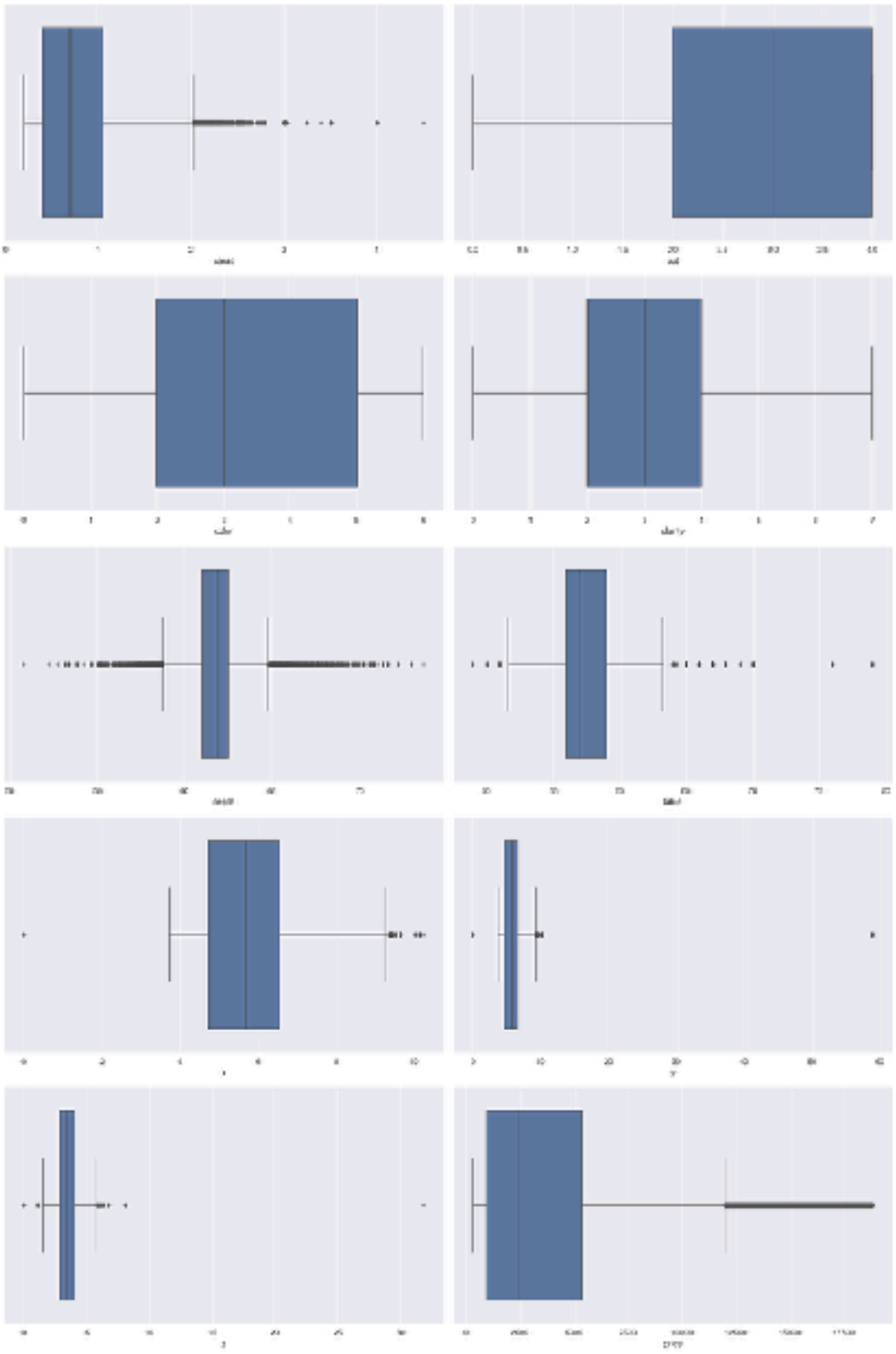
Table: 1.5

Variable	Ordinal codes
cut	Fair = 0, Good = 1, Very Good = 2, Premium = 3, Ideal = 4
color	D = 6 , E = 5, F = 4, G = 3, H = 2, I = 1, J = 0
clarity	IF = 7, VVS1 = 6 , VVS2 = 5 , VS1 = 4 , VS2 = 3, SI1 = 2 , SI2 = 1 , I1 = 0

Outlier treatment:

Outliers in the data were checked using box plot in the Seaborn package, which revealed the presence of large number of outliers in columns ‘carat’, ‘depth’ and ‘price’ (shown in figure below). These outliers were treated using Inter-quartile range (IQR) method. It was imperative to treat the outliers since Linear Regression models are very sensitive to outliers.

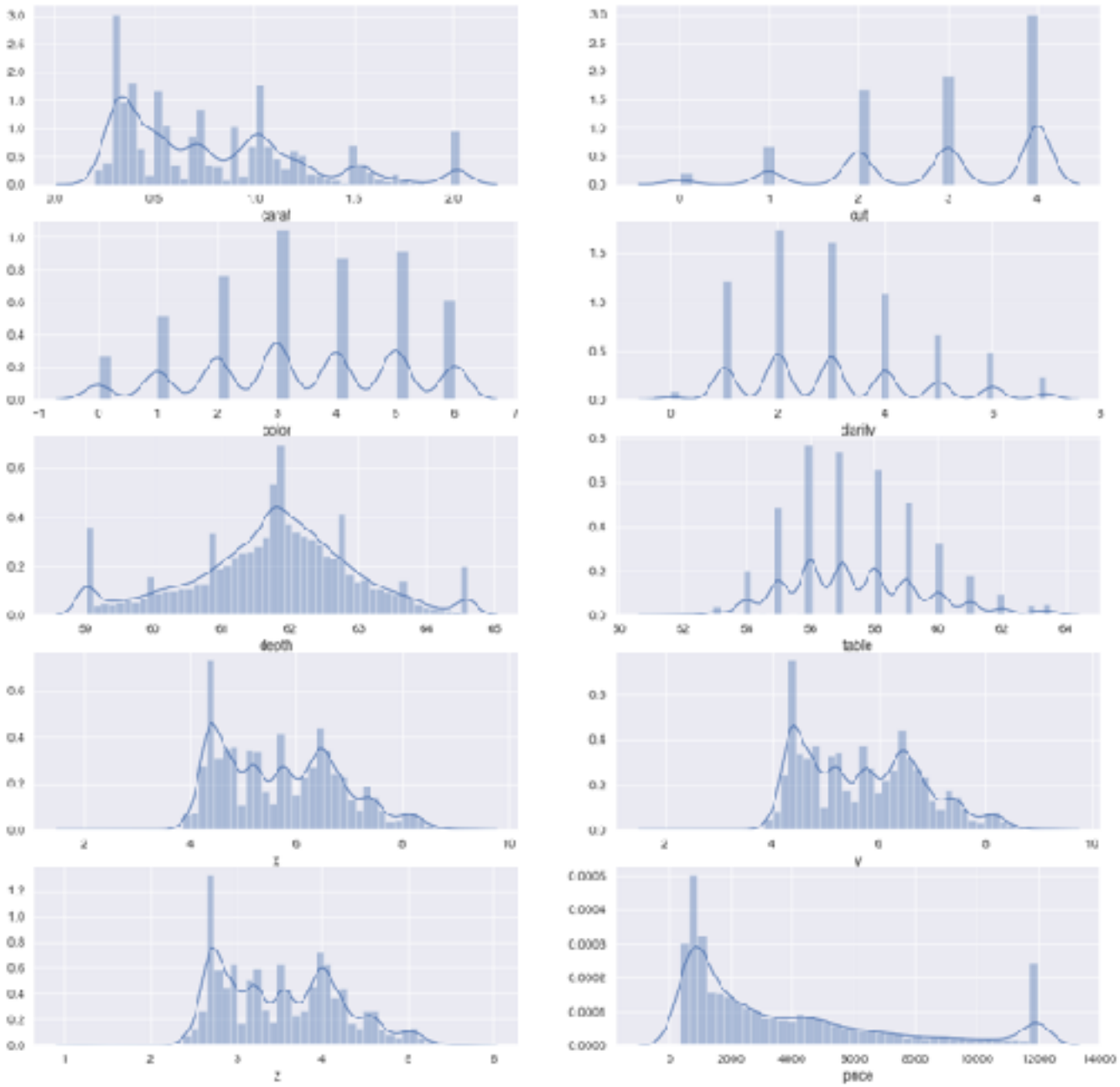
Figure: 1.2



Univariate analysis:

From the histograms, it is clear that only columns namely ‘color’, ‘depth’ and ‘table’ have a somewhat normal distribution, rest all have skewed distribution of data.

Figure: 1.3



Bivariate analysis:

A heat map was used to check correlation between variables, as shown below:

Figure: 1.4

- We can see mostly negative correlations between variables, and magnitude of correlations is also generally low.

- However, there is strong positive correlation between the target variable ‘price’ and certain independent variables like ‘carat’, ‘x’, ‘y’ and ‘z’, which means they influence the price positively - they might be good predictors of price.
- ‘cut’ and ‘depth’ have very negligible effect on the price of the product - so they might be weak predictors of price.
- We can easily infer that the price of the cubic zirconium is heavily dependent on its carat value, its length, width and height.

Further, correlation among variables was checked using both a pair plot, so that a fair idea of the interdependence of variables could be derived. The results are as seen in figure:1.5 :

From the pair plot we can see that certain variables do exhibit a linear relationship between themselves. For example, ‘carat’ shows good linear relationship with the length, width and height (‘x’, ‘y’ and ‘z’) of the product. Also, the dependent variable ‘price’ also shows a somewhat linear relationship with ‘carat’, ‘x’, ‘y’ and ‘z’.

Figure: 1.5

Scaling of data:

The given dataset has varied types of variables, wherein the input scales and the magnitude of data varies largely. Hence, scaling of data was considered. The LR model was built twice (before and after scaling the data), to check the efficacy of the model in both cases. The 'z-score' method of data scaling was used here. It was observed that the model yielded better results post scaling of data. Here, for the purpose of reporting, only post-scaling LR model is described.

Building the LR model:

The dataset was split into training and test set in the ratio 70:30 before building of the Linear Regression model, using the 'train_test_split' function from the sklearn.model_selection package. The model was built on scaled as well as unscaled data using the 'LinerRegression()' function from the sklearn.linear_model package. However, only the findings of the scaled data LR model will be elaborated here.

On running the model on the scaled data, the following coefficients were derived for the various independent variables:

Table: 1.6

The coefficient for carat is 1.172329431287857
The coefficient for cut is 0.03538735470122265
The coefficient for color is 0.1366674976657394
The coefficient for clarity is 0.2095298895319118
The coefficient for depth is -0.0025877372230944276
The coefficient for table is -0.007669299150525637
The coefficient for x is -0.45739268120861276
The coefficient for y is 0.4011897227360788
The coefficient for z is -0.06267063638166011

The coefficients determine the level of influence of each feature/variable on the dependent variable ('price'). We can observe that few features influence positively, i.e., their increase will cause a corresponding increase in the price of the cubic zirconium. While some features influence the price negatively, i.e. their increase will cause the price to decrease. The below table shows the list of feature influence on the price of the product:

Table: 1.7

Feature / variable	Change in price due to unit increase in feature
carat	increases by 1.1723 units
y (width)	increases by 0.4011 units
clarity	increases by 0.2095 units
color	increases by 0.1366 units
cut	increases by 0.0353 units
depth	decreases by 0.0025 units
table	decreases by 0.0076 units
z (height)	decreases by 0.0626 units
x (length)	decreases by 0.4573 units

The intercept for the model = **-1.365123e-16 ~ 0 (tends to 0)**

Based on these results, we can derive the linear equation for the product-price relationship, as given below:

Intercept (0.0) + (1.17) * carat + (0.4) * y + (0.21) * clarity + (0.14) * color + (0.04) * cut + (-0.0) * depth + (-0.01) * table + (-0.06) * z + (-0.46) * x

Model evaluation:

Evaluation metrics such as R2 (mean residual square) and RSME (root mean square error) were utilized to check the effectiveness of the LR model. The results are given below:

Table: 1.8

Train RSME	Test RSME	R2 (Training score)	R2 (Test score)
0.26236	0.2623	0.931167	0.931199

- Low RSME (0.2623) - for both training and test sets, signal a good indication of model robustness, since it means that the observations are closer to the best fit line derived by the LR model.
- R2 scores closer to 1 (0.9311) in both training and test sets indicate that the model is reliable.

However, often R2 can be misleading - it increases with increase in number of variables, irrespective of their contribution to prediction of dependent variable. So, 'statsmodel' library was used, which gives many more metrics of evaluation. The following additional metrics were derived by the statsmodel computation:

Table: 1.9

Metrics	Value	Implication
R2	0.931	Model is reliable
Adjusted R2	0.931	
Model P value (probability)	0.00	(p<0.05) model is reliable

To check multicollinearity, the test of VIF (variation inflation factor) was done. The following results were derived:

Table: 1.10

Variable	VIF
carat	113.0434784
cut	9.740579582
color	5.543478567
clarity	5.423435221
depth	956.9981408
table	756.8909979
x	10286.72079
y	9325.505099
z	1981.541209

Each variable has a VIF value > 5. This proves that there is a high level of multicollinearity in the data, i.e. there is a high level of correlation between the independent variables. At this point, we will not bother about this, since multicollinearity does not affect the accuracy of the LR model.

Feature importance:

From the LR model coefficients, we can say that the 5 most important features/variables that help to predict the price of the cubic zirconium are given below:

Table: 1.11

Features
carat
x (length)
y (width)
clarity
color

Business insights:

- Physical characteristics (length, width, height) of the cubic zirconium has a strong bearing on its carat value, which in turn has the highest influence on the price of the product.
- The physical features of the product (length, width, height, carat) have the most direct influence on its price, rather than the qualitative characteristics (cut, color, clarity)
- So it is imperative for the company to focus on procuring good quality crude zirconium, rather than investing heavily on its processing.
- Carat value is the biggest predictor of price - with a unit increase in carat, the price of the product increases approx. 1.8 times. The company must take care to maintain high carat value.
- Similarly, with unit increase in width, clarity and color also the price increases, so the company must pay attention to these parameters.
- Length of the cubic zirconium is another strong predictor of its price - with every unit increase in length, the price drops by 0.5 times. This is something that the company must take into account when processing the product.

PROBLEM 2: LOGISTIC REGRESSION & LDA

Executive Summary:

The client, a tour & travel agency, has supplied data about employees from one of their client companies. In this data, certain employees have opted for a particular tour package, while certain have not. Based on this data, predications need to be made whether employees in future will opt for this package or not.

Exploratory data analysis:

The sample data has 872 observations, with values under 8 columns, the first 5 rows are given in the table below:

Table: 2.1

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
4	no	66503	31	11	2	0	no
5	no	66734	44	12	0	2	no

Data description:

For computational ease, the first column (Unnamed: 0) was dropped as it is merely the serial number, and will not give any insights. A brief description of the column heads:

Table: 2.2

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

Checking datatypes:

The data types of the dataset was checked. The following data types are present in the dataset:

Table: 2.3

Data type	Variables
int64	‘Salary’, ‘age’, ‘educ’, ‘no_young_children’, ‘no_older_children’
object	‘Holliday_Package’, ‘foreign’

Checking for null and duplicate values:

- The data was checked for null values using the isnull() function. There were no null values in the dataset.
- The function duplicated() was used to check for duplicate values - there were no duplicate entries in the data set.

Statistical summary of the data:

The statistical summary of the data was checked using the describe() functions shown in table below:

Table: 2.4

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872	872	872	872	872	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.17202	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.66853	10.551675	3.036259	0.61287	1.086786	NaN
min	NaN	1322	20	1	0	0	NaN
25%	NaN	35324	32	8	0	0	NaN
50%	NaN	41903.5	39	9	0	1	NaN
75%	NaN	53469.5	48	12	0	2	NaN
max	NaN	236961	62	21	3	6	NaN

The following can be inferred from the summary:

- The variables ‘age’, ‘educ’, ‘no_older_children’ seem to be normally distributed.
- There seems to be a balanced distribution of data classes in ‘Holliday_Package’ (close to 50%) - meaning proportion of people opting for the holiday package and not opting for it are almost equal.
- However, the distribution of the categorical variable ‘foreign’ seems to be skewed in proportion.

- The dataset contains observations pertaining to working class sample (ages between 20 - 62), with the average age being approximately 40 years.
- There is a wide variation in terms of salary, with the average employee earning approx. Rs. 48,000.
- Education also shows a wide variation, but is suggestive of a normal distribution.

Checking proportion of target class variable:

The target variable in this dataset is 'Holliday_Package', having two classes - 'yes' and 'no'. The percentage of data in each class is given below:

Table: 2.5

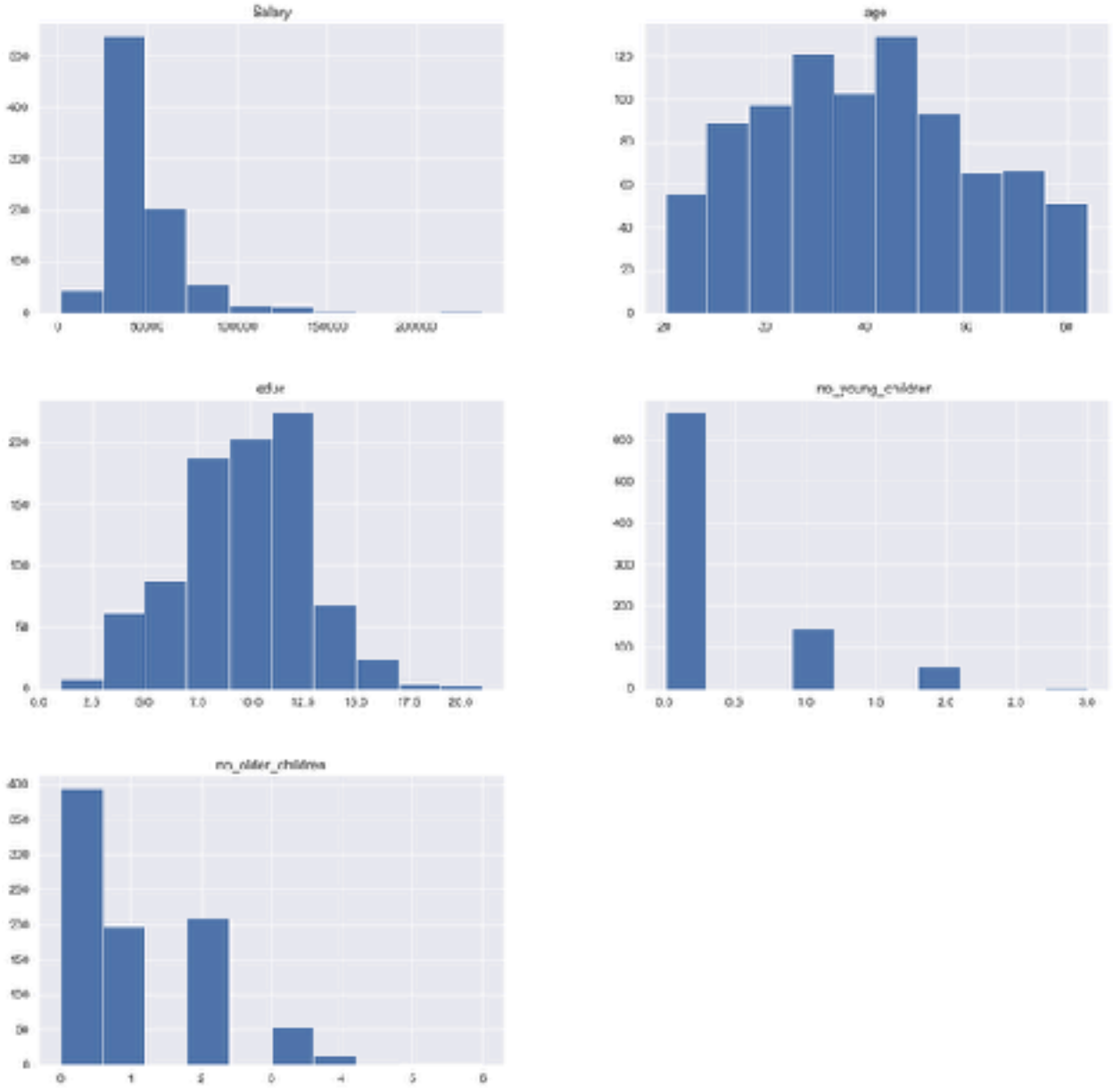
Target class (Holliday_Package)	Percentage
no	54.01%
yes	45.99%

We can see that the class distribution is not biased, rather it is very balanced, hence there is not need for balancing binary classes.

Univariate data analysis:

Barplots of the continuous variables yielded the following results.

Figure: 2.1



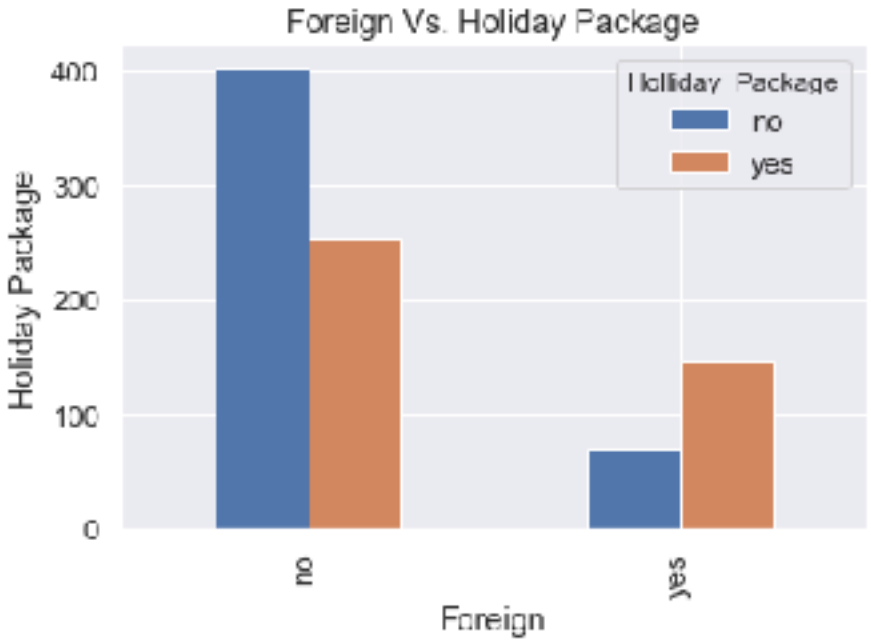
- Salary-wise, the sample is concentrated around Rs. 5000.
- 40 - 45 year olds form the largest chunk in terms of age.

- The largest concentration of sample is education at high school level (between 8th year and 12th year)
- Only a minuscule portion of the sample has young children (< 7 years).
- Any equal proportion of the sample has 1 an 2 older children. Very few have 3-4 older children.

Bivariate analysis:

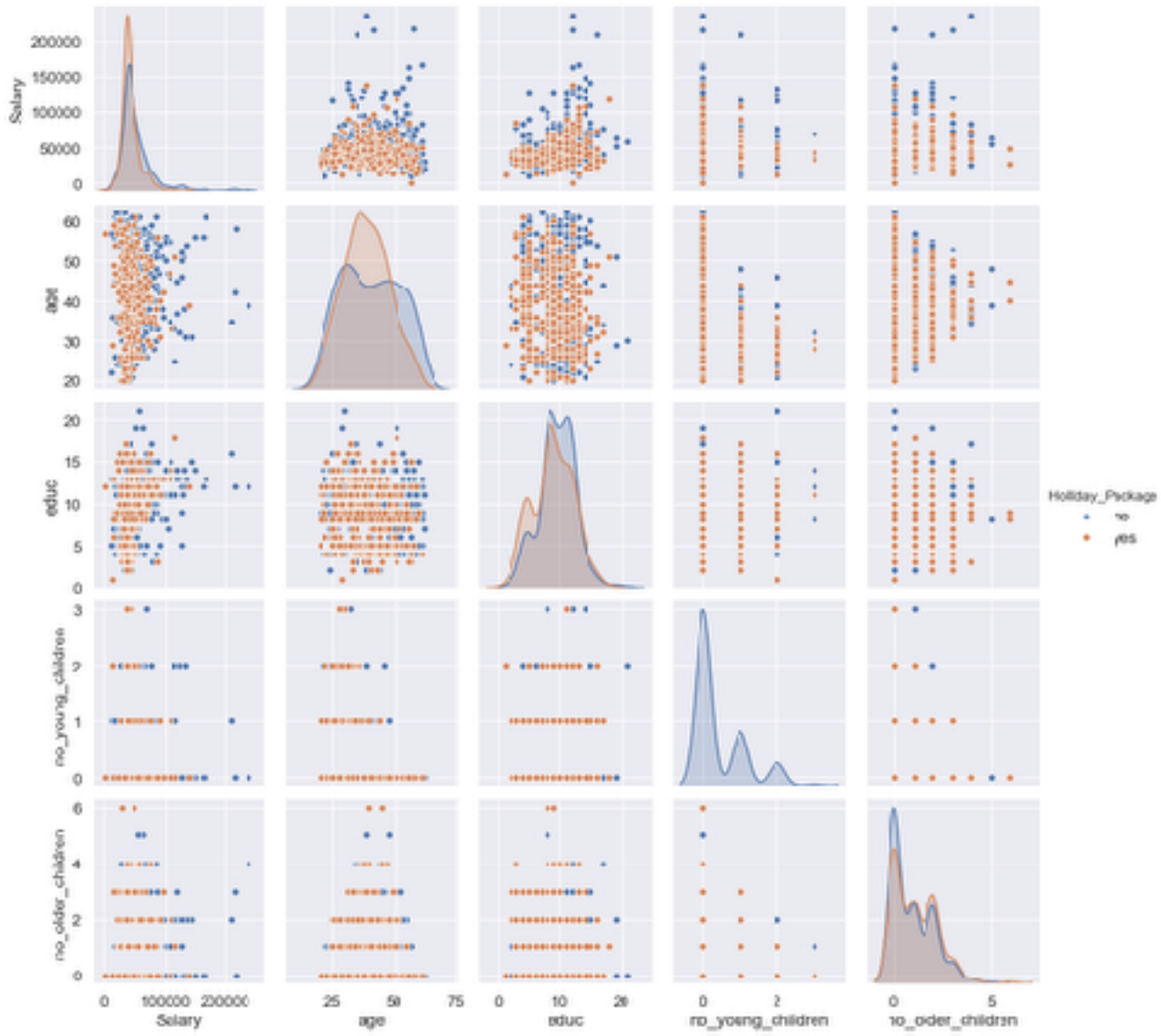
Using a barplot, the distribution of 'Holiday_Package' was checked against 'foreign', as shown in the figure below. It clearly that among the people not opting for holiday packages, a vast majority is local people. Also, it is evident that among foreigners, the proportion of people opting for the package is higher (almost double than the people not opting for a package).

Figure: 2.2



A pairplot was also used to check for the relation between the continuous variables, as shown in given figure:

Figure: 2.3



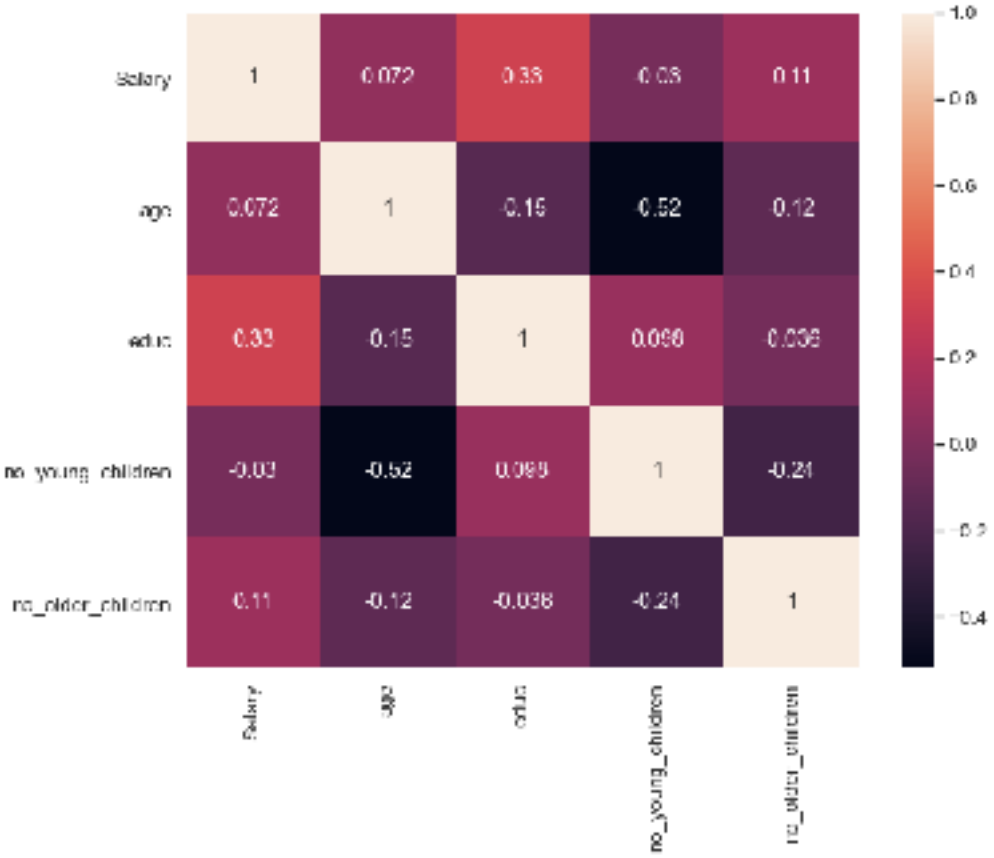
The following can be gauged from the pair plot:

- There are no evident linear relationship between any of the variables.
- In fact, education and age seem to be the only two variables with a somewhat normal distribution, rest are highly skewed.

- The plots also shows the people who have younger children (< 7 years) do not opt for holiday packages.
- Similarly, among people with more than 3 older children, very few opt for a holiday package.
- In terms of educations, it is the moderately educated who opt for packages. Neither those with very little education nor those highly educated are opting for packages.
- Surprisingly, all the 50-62 year olds have a high instance of opting for holiday packages!

A heatmap was plotted to check correlation among the continuous variables. It does not reveal any significant correlations among the variables.

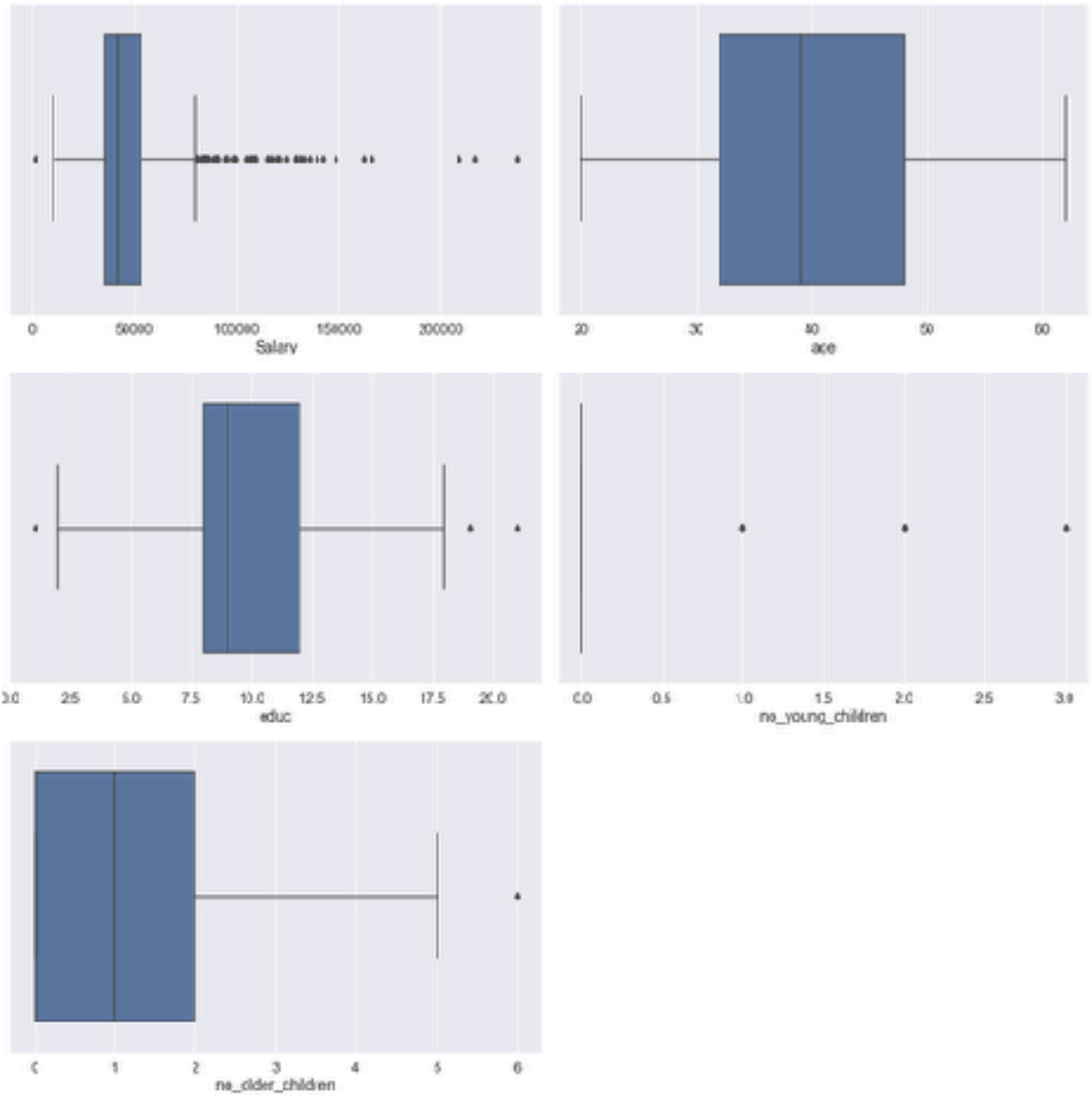
Figure: 2.4



Checking data for outliers:

The data was checked for presence of outliers using box plots. Considerable outliers were present in the variable ‘salary’, and null or very few outliers in the other variables, as shown in figure below:

Figure: 2.5



Salary is the only variable that shows a major presence of outliers.

The Inter-quartile range (IQR) method was used to treat the outliers in columns ‘Salary’ and ‘educ’. The rest were not treated as they were not many in number.

Class encoding for categorical variables (object type):

In order to convert object type variables to integer values, the columns ‘Holliday_Package’ and ‘foreign’ were converted to categorical codes using label encoding:

Table: 2.6

Variable	Codes
Holliday_Package	Yes = 1, No = 0
Foreign	Yes = 1, No = 0

Building the Logistic Regression model:

The dataset was split into training and test set in the ratio 70:30 before building of the Logistic Regression model, using the ‘train_test_split' function from the sklearn.model_selection package.

After splitting the data, the training data was fitted into the model using LogisticRegression() function.

Model evaluation:

The following metrics were used to evaluate the model:

Table 2.7

	Log-reg Train	Log-reg Test
Accuracy	0.68	0.637405
AUC	0.74	0.705158
Recall	0.56	0.56
Precision	0.68	0.6
F1 Score	0.62	0.58

We can see that the Logistical regression model is showing moderate accuracy. The recall and precision are not too high either.

The ROC curves for the Logistic Regression model are given below:

Figure: 2.6

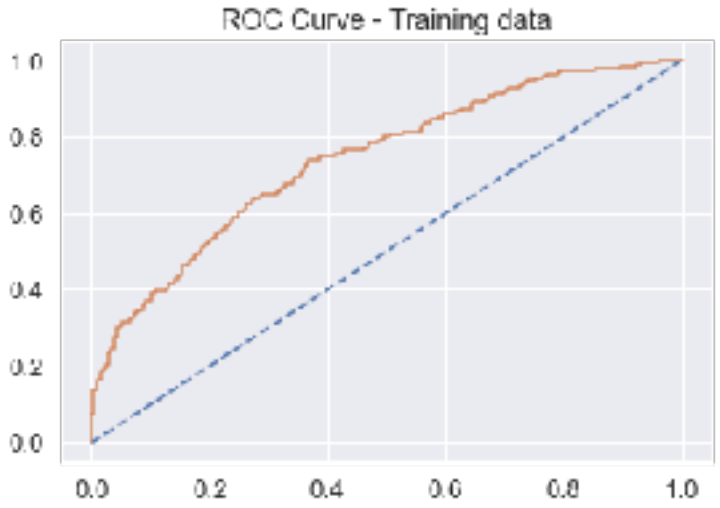
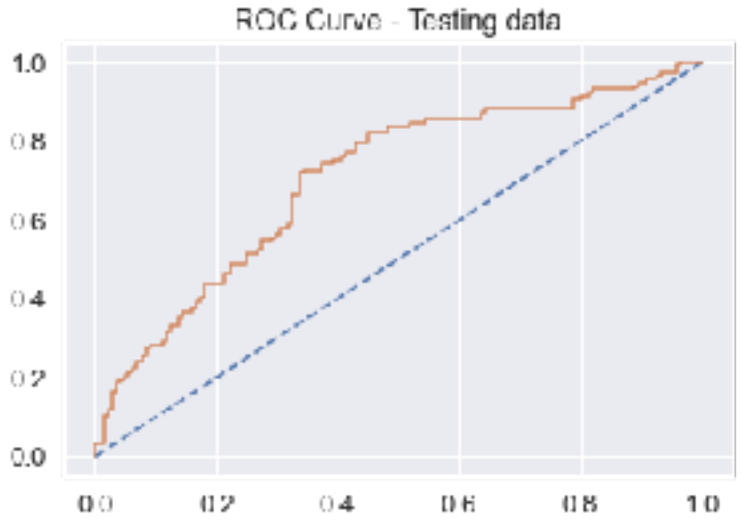


Figure: 2.7



Confusion matrix for the training and test data:

Figure: 2.8

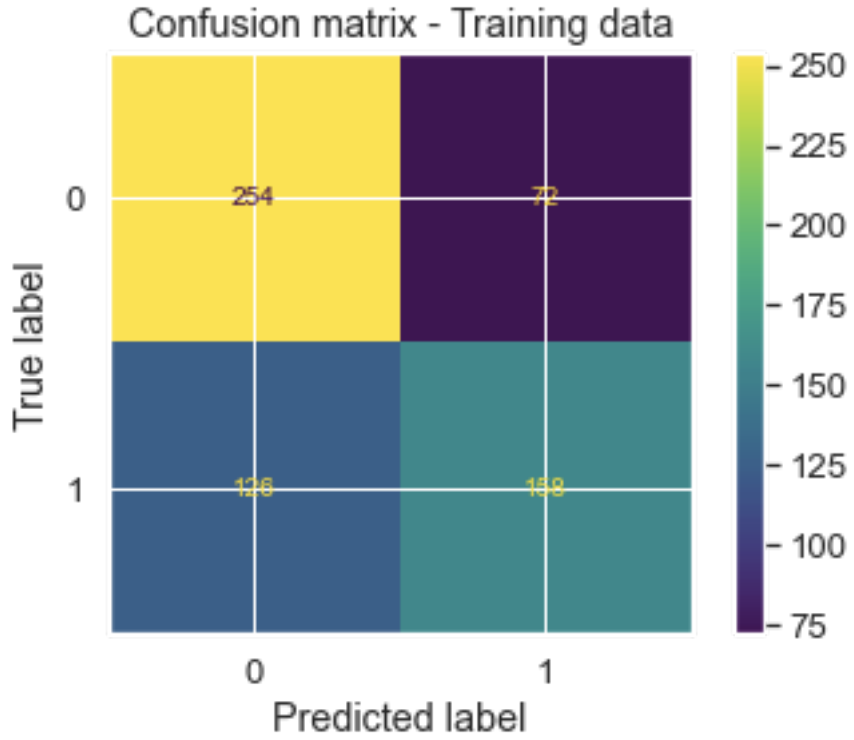
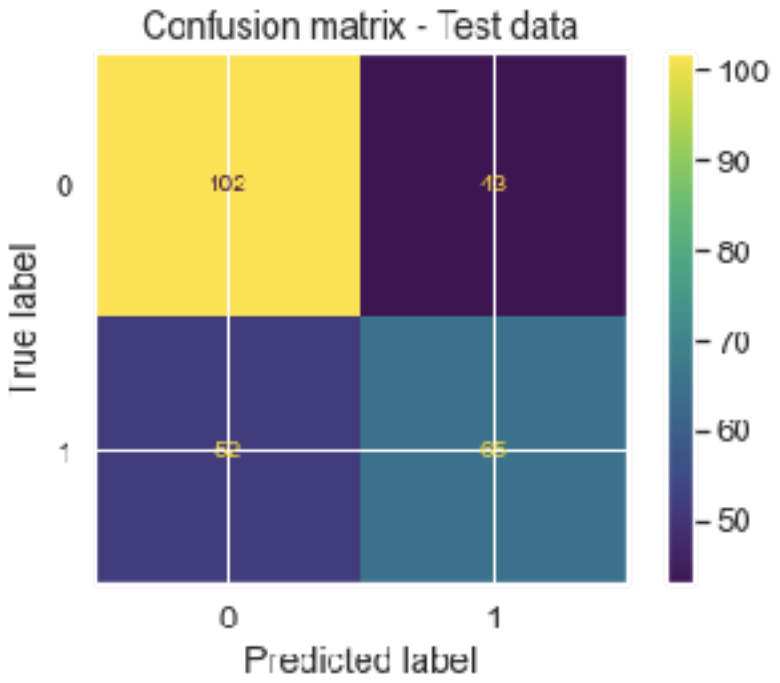


Figure: 2.9



Building the LDA (Linear Discriminant Analysis) model:

The LDA model was built using the LinearDiscriminantAnalysis function from the sklearn.discriminant_anaysis library. The following metrics were derived from the model:

Table: 2.8

	LDA Train	LDA Test
Accuracy	0.68	0.64
AUC	0.74	0.7
Recall	0.56	0.56
Precision	0.69	0.6
F1 Score	0.61	0.58

Comparison of the two models:

On comparing the metrics derived from both the Logistic Regression model and the LDA model, we can say that there exists negligible difference between them. Both exhibit moderate accuracy, recall and precision in terms of predicting the target variable.

Table: 2.9

	Log-reg Train	Log-reg Test	LDA Train	LDA Test
Accuracy	0.68	0.637405	0.68	0.64
AUC	0.74	0.705158	0.74	0.7
Recall	0.56	0.56	0.56	0.56
Precision	0.68	0.6	0.69	0.6
F1 Score	0.62	0.58	0.61	0.58

The accuracy of the models could be improved by fine-tuning the parameters of the model building.

Business insights:

Based on the dataset, certain inferences can be made regarding the sample:

- Salary seems to be a good predictor in the decision to buy holiday package. To attract the people with less incomes, the travel agency can come up with cost-effective plans to suit different budget levels, or offer installment plans.
- Also, to attract high-end customers, they could come up with niche packages targeted at the affluent market.
- The presence of young children (< 7 years) seems to be a major deterrent in opting for holiday packages. The agency could come up

with special packages targeted at people with small children, wherein the location, travel modes, activities, food choices are designed keeping in mind small children.

- Even among people with older children, the trend is to not opt for travel packages. To entice such people, the agency could tie-up with destination partners to offer youth-centric activities and itinerary.
- The foreigner clients are the cash cows here - twice the number opt for holiday packages. Hence, it is necessary for the agency to come up with differentiated products for the foreigner market, giving them a flavor and experience of the local land.