PGP - DSBA

Predictive Modeling

Project Report – October 2022

Shruti Jha
10-19-2022

Contents

List of Tables

List of Figures

Problem 1 – Linear Regression

Introduction

Gem Stones co ltd, which is a cubic zirconia manufacturer, has 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. The company needs help in predicting the price for the stone on the basis 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

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	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst in terms of avg price) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
Depth	Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
Table	Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.
Price	Price of the cubic zirconia.

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.

1.1.1 Sample of dataset

Here are the top 5 rows (sample) of the dataset:

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

Table 1. 1: Dataset Sample

- Dataset has 10 valid variables and as we can see the first column (*Unnamed: 0*) only contains serial numbers which are not relevant, we can remove it from our dataset.
- Also, changed the names of columns 'x', 'y' and 'z' to make them self-explanatory.

	carat	cut	color	clarity	depth	table	length	width	height	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 1. 2: Modified Dataset Sample

1.1.2 Check for Duplicate Records

```
Number of duplicate records: 34
```

We have 34 duplicate entries in our database. As per the problem statement, company needs help in predicting the price for the stone on the bases of unique attributes provided for different types of stones.

Keeping this in mind, we have removed the duplicates as the duplicate values of various attributes can manipulate the end result / prediction.

1.1.3 Types of variables in the dataset

```
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    26933 non-null  float64
 1   cut      26933 non-null  object
 2   color    26933 non-null  object
 3   clarity  26933 non-null  object
 4   depth    26236 non-null  float64
 5   table    26933 non-null  float64
 6   length   26933 non-null  float64
 7   width    26933 non-null  float64
 8   height   26933 non-null  float64
 9   price    26933 non-null  int64
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

- There are 7 variables in numeric (float64, int64) and 3 variables in categorical (object) format.
- There are a total of 10 columns and 26933 rows in the dataset, after removing duplicates.
- We can see there are missing values in 'depth' column, which needs to be taken care of.

1.1.4 Missing values in the dataset

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

There are 697 missing values in 'depth' column, which are fixed in further steps of data preprocessing.

1.1.5 Descriptive Statistics

Describe function provides a table indicating the count of variables, mean, standard deviation and other values for the 5-point summary that includes (min, 25%, 50%, 75% and max) for numeric variables. 50% in the table is also known as median.

	count	mean	std	min	25%	50%	75%	max
carat	26933.0	0.798010	0.477237	0.2	0.40	0.70	1.05	4.50
depth	26236.0	61.745285	1.412243	50.8	61.00	61.80	62.50	73.60
table	26933.0	57.455950	2.232156	49.0	56.00	57.00	59.00	79.00
length	26933.0	5.729346	1.127367	0.0	4.71	5.69	6.55	10.23
width	26933.0	5.733102	1.165037	0.0	4.71	5.70	6.54	58.90
height	26933.0	3.537769	0.719964	0.0	2.90	3.52	4.04	31.80
price	26933.0	3937.526120	4022.551862	326.0	945.00	2375.00	5356.00	18818.00

Table 1. 3: Data Description for integer variables

For object/categorical columns, describe function shows the observation count, unique values in each column, most frequent value and value frequency in each column.

	count	unique	top	freq
cut	26933	5	Ideal	10805
color	26933	7	G	5653
clarity	26933	8	SI1	6565

Table 1. 4: Data Description for object variables

CUT : 5		COLOR : 7		CLARITY : 8	
Fair	780	J	1440	I1	364
Good	2435	I	2765	IF	891
Very Good	6027	D	3341	VVS1	1839
Premium	6886	H	4095	VVS2	2530
Ideal	10805	F	4723	VS1	4087
		E	4916	SI2	4564
		G	5653	VS2	6093
				SI1	6565

Table 1.4: Categories under object variables

From the above descriptive statistics, we can infer:

- The maximum carat weight is 4.5 and the average is 0.8. From the above table, 75% (Q3) indicates that majority of the stones in the data weigh below 1.05 carat weight.
- As we had already observed, 'depth' column has missing values, which needs to be taken care of. Out of the available values, depth ranges between 50.8% and 73.6%.
    - The **ideal depth of a good stone ranges between 59% and 74%** [1]. Hence, we can infer that there are some shallow stones present in our data with depth below 59%.

[1] *International Gem Society*

- 'table' is denoted in percentage (%) form, with respect to the average diameter of individual stones. The **ideal table of a good quality stone ranges between 54% till 75%** [2]. The 'table' of stones present in our data ranges between 49% till 79%. Hence, we can say there are some stones in our data which are dull and lacklustre.
- We can observe some 0.0 values in length, width and hight columns, which are anomalies and need to be treated.
- In width and height columns, maximum values are 58.9 and 31.8mm, respectively, which don't seem be valid either.
- Price of the stones ranges from 326.0 till 18,818.0 INR, with the average of 3,937.53 INR.
- 'carat' and 'price' variables seem to be right skewed with most of values being on the left side of the curve.
- For rest of the variables – mean and median are very close to each other, indicating that those variables are somewhat normally distributed.
- The highest costing cubic zirconia weighs 2 carat, with depth of 63% and table of 56%, which is well within the ideal limits.

price	carat	depth	table	cut	color	clarity
18818	2.0	63.5	56.0	Very Good	G	SI1

Table 1. 5: Maximum priced cubic zirconia

- The lowest costing cubic zirconia stones have depth and table falling well within the ideal limit, as well as their cut, color and clarity are also reasonably fine. This indicates that even the lowest costing stones are of good quality, and the price is less only because of the less carat weight of the stone.

price	carat	depth	table	cut	color	clarity
326	0.23	61.5	55.0	Ideal	E	SI2
326	0.21	59.8	61.0	Premium	E	SI1

Table 1. 6: Minimum priced cubic zirconia

- There are 5 categories of 'cut' of the stones, with highest no. of stones (10805) of 'ideal' cut.
- There are 7 categories of 'color' of the stones, with highest no. of stones (5653) of 'G' grade of color.
- There are 8 categories of 'clarity' of the stones, with highest no. of stones (6565) of 'SI1' clarity.

[2] *International Gem Society*

1.1.6 Check for outliers

To check for outliers in numeric variables, box plots have been plotted:

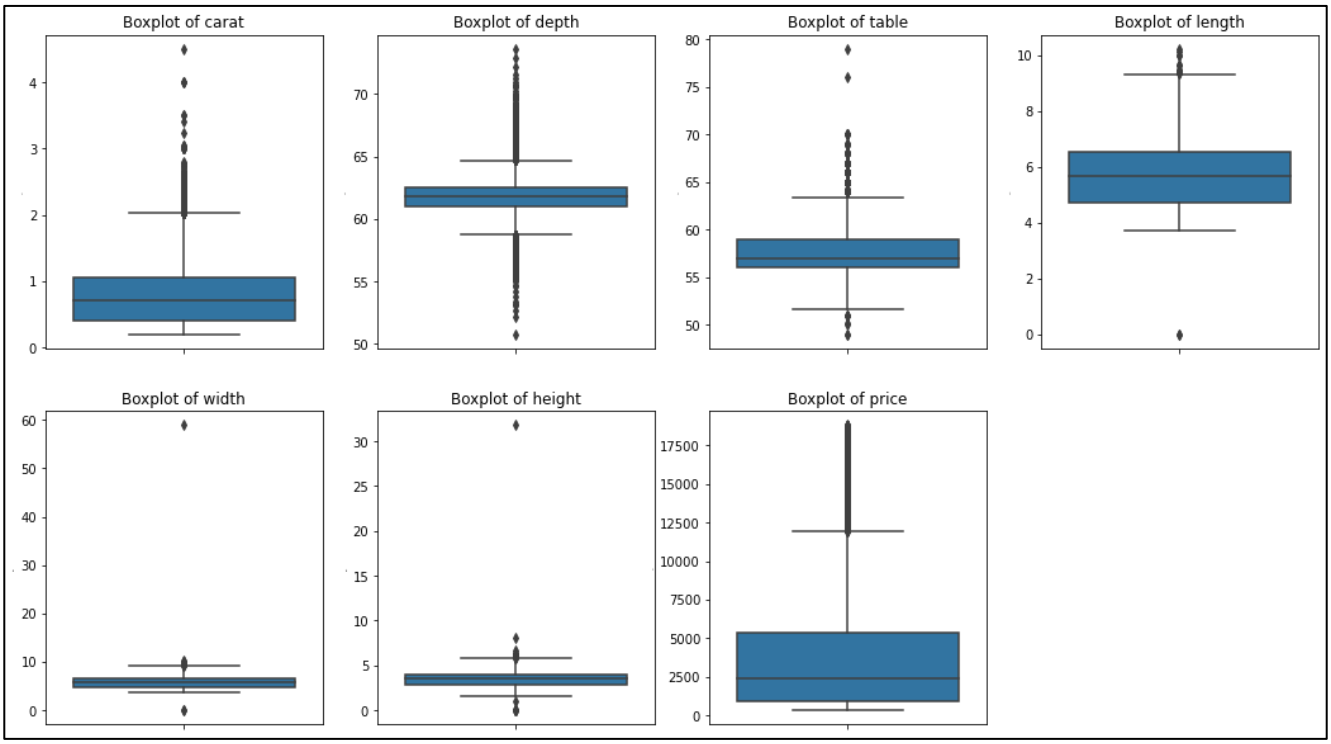


Figure 1. 1: Boxplot for Outliers

- The small dots outside the whiskers of boxplots denote outliers. All the numeric columns in our data have outliers present.
- The 0 data values present in length, width and height are anomalies and need to be treated.
- The maximum values in width and height seem to be invalid as well, as the width and height of a stone cannot be that large as compared to other dimensions. These columns need to be treated as well.

width	length	height	carat	depth	table	price	cut	color	clarity
58.9	8.09	8.06	2.0	58.9	57.0	12210	Premium	H	SI2
5.15	5.12	31.8	0.51	NaN	54.7	1970	Very Good	E	VS1

Table 1. 7: Minimum priced cubic zirconia

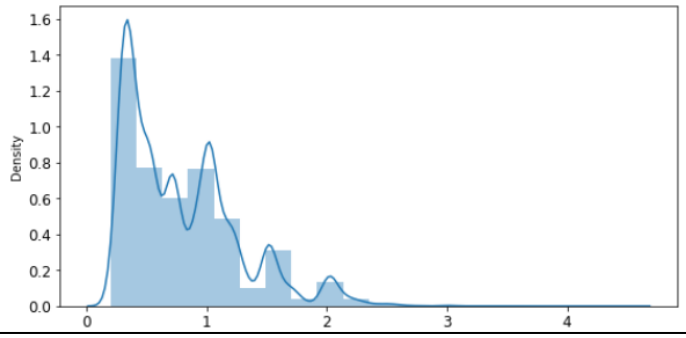
- Other than above mentioned invalid values, rest all outliers appear to be valid and significant.
- We have imputed 0 data values present in length, width and height with the lower limit (Q1 – 1.5*IQR) of the respective variables.
- Invalid values in width and height columns are treated using upper limit (Q3 + 1.5*IQR).

1.1.7 Univariate analysis

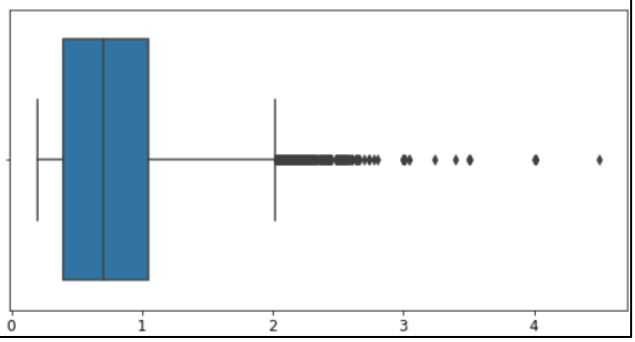
Univariate analysis is performed for all the numeric variables individually to display their statistical description. Visualized the variables using distplot to view the distribution and the box plot to view 5-point summary and outliers if any.

```
Description of carat
.....................................................................
count    26933.000000
mean         0.798010
std          0.477237
min          0.200000
25%          0.400000
50%          0.700000
75%          1.050000
max          4.500000
Name: carat, dtype: float64
```

Distribution of carat

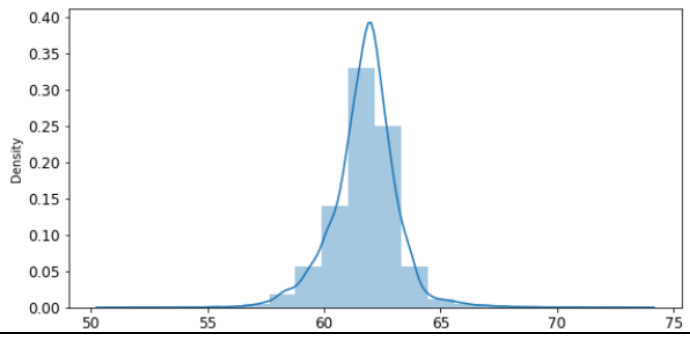


boxplot of carat

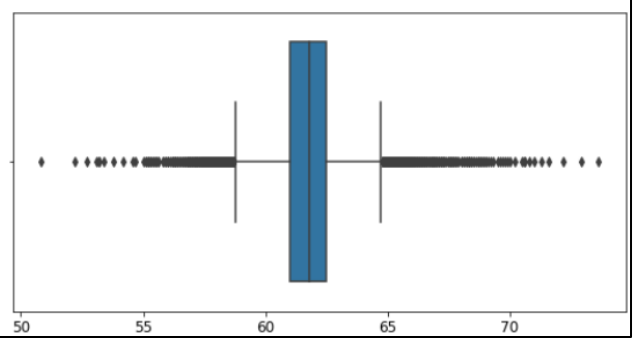



```
Description of depth
.....................................................................
count    26236.000000
mean        61.745285
std          1.412243
min         50.800000
25%         61.000000
50%         61.800000
75%         62.500000
max         73.600000
Name: depth, dtype: float64
```

Distribution of depth

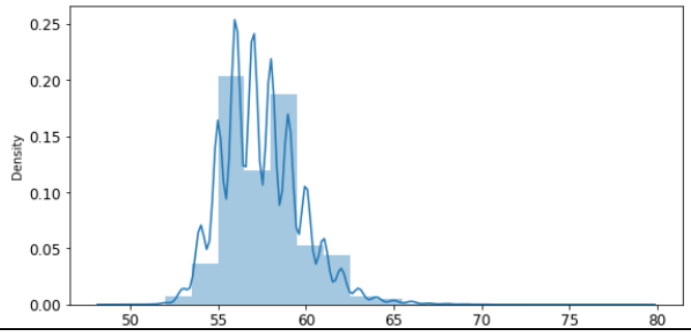


boxplot of depth

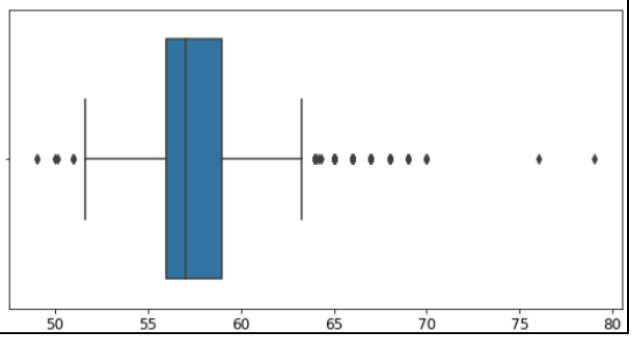


```
Description of table
.....................................................................
count    26933.000000
mean        57.455950
std          2.232156
min         49.000000
25%         56.000000
50%         57.000000
75%         59.000000
max         79.000000
Name: table, dtype: float64
```

Distribution of table

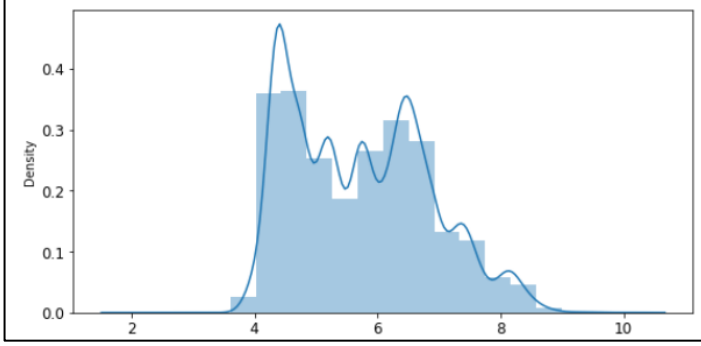


boxplot of table

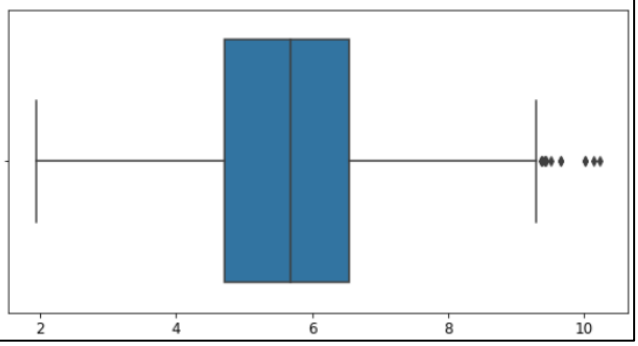



```
Description of length
.....................................................................
count    26933.000000
mean         5.729491
std          1.126756
min          1.950000
25%          4.710000
50%          5.690000
75%          6.550000
max         10.230000
Name: length, dtype: float64
```

Distribution of length

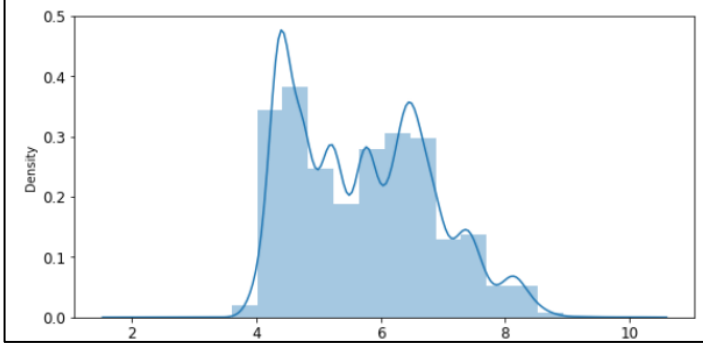


boxplot of length

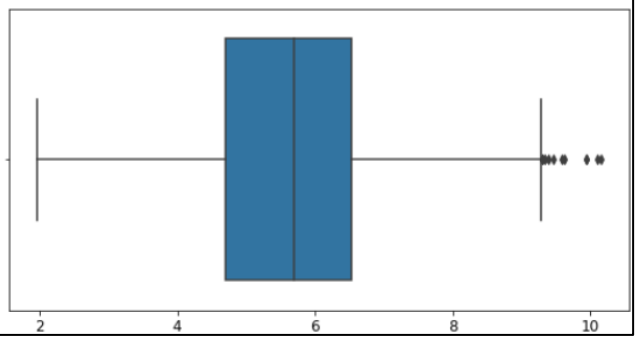


```
Description of width
.....................................................................
count    26933.000000
mean         5.731406
std          1.118675
min          1.965000
25%          4.710000
50%          5.700000
75%          6.540000
max         10.160000
Name: width, dtype: float64
```

Distribution of width

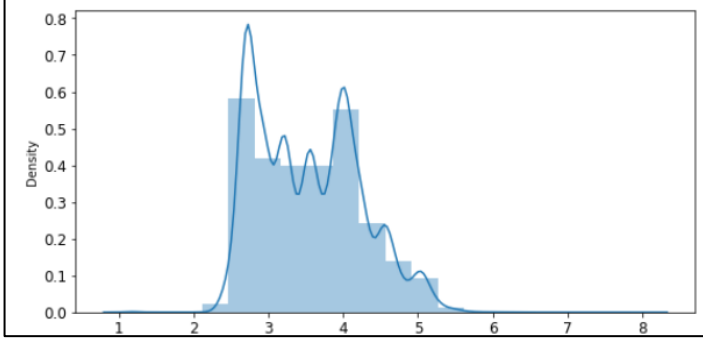


boxplot of width

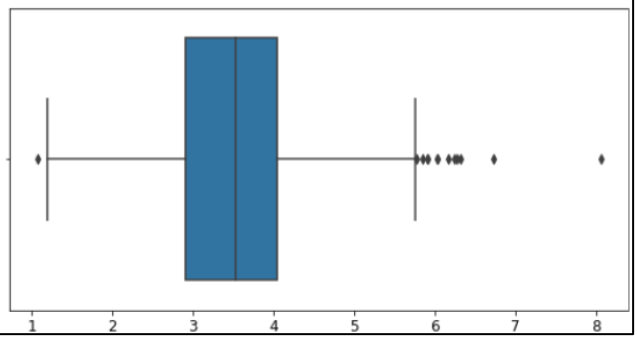



```
Description of height
.....................................................................
count    26933.000000
mean         3.537156
std          0.697704
min          1.070000
25%          2.900000
50%          3.520000
75%          4.040000
max          8.060000
Name: height, dtype: float64
```

Distribution of height



boxplot of height



```
Description of price
.....................................................................
count    26933.000000
mean      3937.526120
std       4022.551862
min        326.000000
25%        945.000000
50%       2375.000000
75%       5356.000000
max      18818.000000
Name: price, dtype: float64
```

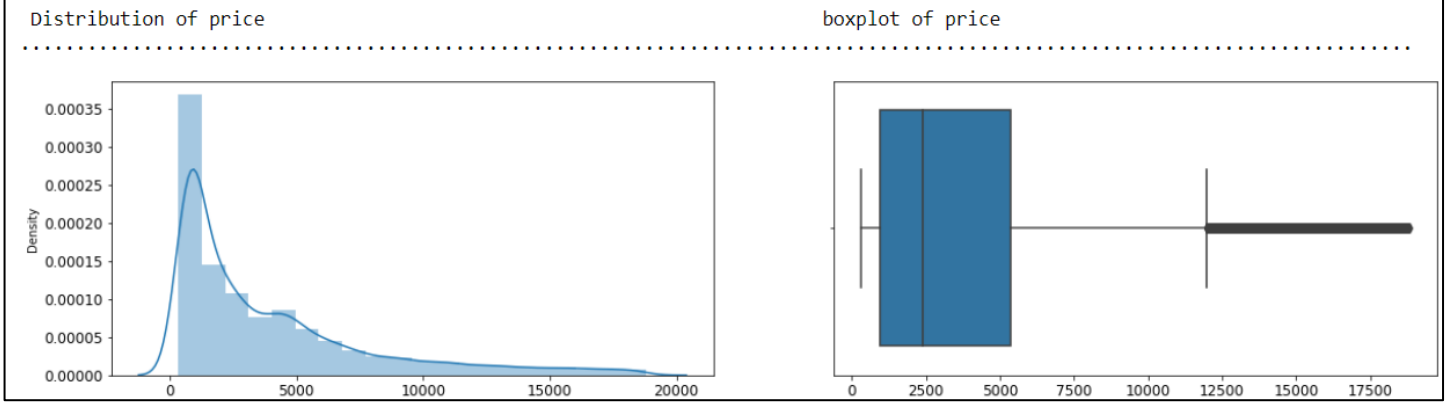


Figure 1. 2: Univariate Analysis

	Kurtosis	Skewness
carat	1.21	1.11
depth	3.68	-0.03
table	1.58	0.77
length	-0.71	0.40
width	-0.72	0.40
height	-0.62	0.40
price	2.15	1.62

Table 1. 8: Kurtosis & Skewness

Observations

- There are 7 numeric fields in the dataset.
- From the boxplots we can see that there are no invalid values present in the data anymore. The outliers present in the data is valid.
- The distributions for 'carat', 'table', 'length', 'width', 'height and 'price' are multimodal.
- 'depth' seems to have data that is normally distributed, as the skewness is very close to zero and creates a perfect bell curve.
- The distribution is right/positive skewed for 'carat' and 'price'.

1.1.8 Multivariate analysis

Pair plot:

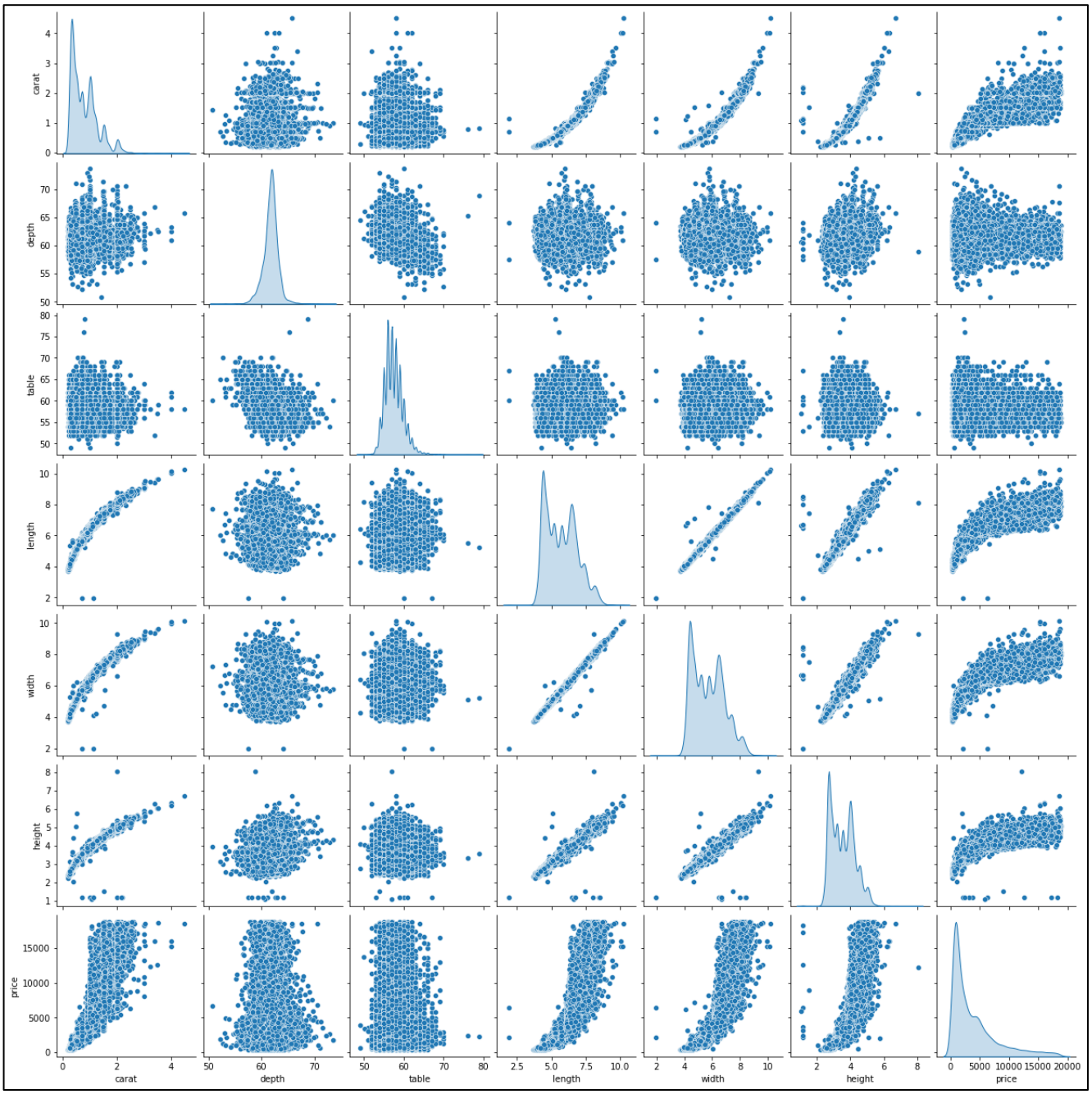


Figure 1. 3: Pairplot

- As the length, width and height of stones increases the carat weight also increasing. As the carat weight increasing the price is also going up.
- When the depth and table is extremely low or high, the carat weight and price remain less. Carat weigh and price are higher for the stones with depth and table almost between 60% to 65%. Stones with an extremely low & high depth and table % often have a dull appearance, and deemed to be of low quality.
- Length, weight and height are highly positively correlated with each other and carat and price.

Correlation plot (Heatmap):

Figure 1. 4: Correlation Plot

- As we have observed in the pairplot as well – carat, length, weight, height and price are highly positively correlated with each other, with correlation pretty much close to 1.
- Correlation for carat and price with depth and table is close to 0 (*no correlation*).
- Table and depth are negatively correlated but it is not significant enough to derive any inferences.

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 a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

1.2.1 Imputing null values

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

- There are 697 missing / null values in 'depth' column.

- As mentioned in the data dictionary, the depth (in %) of cubic zirconia is measured by height of the stone divided by its average diameter. So, rather than employing traditional methods (SiimpleImputer / mean / median), we will use the below formula to impute missing values in depth column to get true values:

$$\text{Depth} = \frac{\text{Height}}{(\text{length} + \text{width}) / 2} * 100$$

- In previous steps we had tackled the issue of 0 value in length, width and height column. With the help of these values in above formula, we have imputed missing values in depth column.

1.2.2 Zero value in columns

Some of the values in length, width and height columns were zero, which were invalid, as these attributes cannot contain a 0 value, and were imputed with the lower limit (Q1 – 1.5*IQR) of the respective variables, before performing descriptive statistics.

1.2.3 Combining sub levels of variables

We have ordinal variables in the dataset, namely – cut, color and clarity, denoting order of the quality of the stones.

CUT : 5		COLOR : 7		CLARITY : 8	
Fair	780	J	1440	I1	364
Good	2435	I	2765	IF	891
Very Good	6027	D	3341	VVS1	1839
Premium	6886	H	4095	VVS2	2530
Ideal	10805	F	4723	VS1	4087
		E	4916	SI2	4564
		G	5653	VS2	6093
				SI1	6565

Table 1. 9: Categorical Sublevels

CUT	
Ideal	**Premium**
Premium	
Very Good	**V_Good**
Good	**Good**
Fair	

COLOR	
D	**Colorless**
E	
F	
G	**Near_colorless**
H	
I	
J	

CLARITY	
IF	**IF**
VVS1	**VVS**
VVS2	
VS1	**VS**
VS2	
SI1	**SI**
SI2	
I1	**I**

Table 1. 10: Combined Sublevels

- We have combined the sublevels to reduce the categories in object columns. Since object datatype cannot be fit into the linear regression line, we will be converting these sublevels into ordinal codes. Having fewer ordinal codes will make model interpretation easier.
- The sublevels are combined in a such a way that the values don't lose their grade. For example, 'Good' and 'Fair cut stones would have attribute values very close to each other; same with 'Premium' and 'Ideal'.
    - 5 sub-levels of 'cut' have been combined into 3 sub-levels – **Premium**, **V_Good**, **Good**.
- For 'color' column - D, E, F fall under colorless grade and G, H, I, J fall under near colorless grade.[3]
    - 7 sub-levels of 'color' have been combined into 2 sub-levels – **Colorless**, **Near_colorless**.

[3] *Color-scale - http://aginewyork.com/diamond-colors-education.php*

- For 'clarity' column, according to the below clarity chart, combined into 5 sub-levels – **IF**, **VVS**, **VS**, **SI**, **I**.

Figure 1. 5: Color scale of precious stones[4]

- All the above-mentioned sub-categories marked in yellow are in decreasing (highest to lowest) ranking.

Updated sample of the dataset:

	carat	cut	color	clarity	depth	table	length	width	height	price
0	0.30	Premium	Colorless	SI	62.1	58.0	4.27	4.29	2.66	499
1	0.33	Premium	Near_colorless	IF	60.8	58.0	4.42	4.46	2.70	984
2	0.90	V_Good	Colorless	VVS	62.2	60.0	6.04	6.12	3.78	6289
3	0.42	Premium	Colorless	VS	61.6	56.0	4.82	4.80	2.96	1082
4	0.31	Premium	Colorless	VVS	60.4	59.0	4.35	4.43	2.65	779

Table 1. 11: Modified Dataset Sample

[4] *Source - https://www.u7jewelry.com/blogs/jewelry-guide/czcubic-zirconia-vs-diamond*

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 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.

1.3.1 Data Encoding

We have 3 variables of object datatype, of which we combined the sublevels. These variables cannot be read in the linear regression equation while running the model. To convert the object values into ordinal numeric values, we performed 2 techniques:

1. Label Encoding – We got the sublevels converted into integers, but the order is not highlighted that is mentioned in the problem statement. See the comparison of sublevels before and after the label encoding:

BEFORE	AFTER
Premium 17691 V_Good 6027 Good 3215 Name: cut, dtype: int64	1 17691 2 6027 0 3215 Name: cut, dtype: int64
Near_colorless 13953 Colorless 12980 Name: color, dtype: int64	1 13953 0 12980 Name: color, dtype: int64
SI 11129 VS 10180 VVS 4369 IF 891 I 364 Name: clarity, dtype: int64	2 11129 3 10180 4 4369 1 891 0 364 Name: clarity, dtype: int64

Table 1. 12: Numeric Sublevels 1

2. We then tried manually assigning ordinal numbers to the subcategories (5 being the highest level; 1 being the lowest), and this is how before and after look:

BEFORE	AFTER
Premium 17691 V_Good 6027 Good 3215 Name: cut, dtype: int64	3 17691 2 6027 1 3215 Name: cut, dtype: int64
Near_colorless 13953 Colorless 12980 Name: color, dtype: int64	1 13953 2 12980 Name: color, dtype: int64
SI 11129 VS 10180 VVS 4369 IF 891 I 364 Name: clarity, dtype: int64	2 11129 3 10180 4 4369 5 891 1 364 Name: clarity, dtype: int64

Table 1. 13: Numeric Sublevels 2

Here are how the datatypes turned out after encoding the data:

```
RangeIndex: 26933 entries, 0 to 26932
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26933 non-null  float64
 1   cut      26933 non-null  int32  
 2   color    26933 non-null  int32  
 3   clarity  26933 non-null  int32  
 4   depth    26933 non-null  float64
 5   table    26933 non-null  float64
 6   length   26933 non-null  float64
 7   width    26933 non-null  float64
 8   height   26933 non-null  float64
 9   price    26933 non-null  int64  
dtypes: float64(6), int32(3), int64(1)
```

All the features are in integer format.

1.3.2 Data Split

- The target variable in our encoded dataset is 'price'.
- The data has been first divided in to independent and dependent (target) variables, x and y respectively.
    - Where x contains all the predictor variables
    - y contains the target variable.
- After that the data is split into training and testing set with both sets having 70% and 30% of the data, respectively. Here is how the x and y containing training dataset look like:

	carat	cut	color	clarity	depth	table	length	width	height
25343	1.00	3	1	2	61.8	58.0	6.44	6.40	3.97
17	1.01	3	2	3	59.8	56.0	6.52	6.49	3.89
12415	0.53	2	2	2	59.3	61.0	5.26	5.29	3.13
18263	0.33	3	2	3	61.1	56.0	4.45	4.48	2.73
26526	1.01	1	2	2	58.3	59.0	6.59	6.61	3.85

Table 1. 14: Independent Variables

	price
25343	3808
17	7127
12415	1226
18263	723
26526	5747

Table 1. 15: Dependent Variable

1.3.3 Linear Regression Model

Using scikit learn Library:

The coefficients for each of the independent attributes:

```
The coefficient for carat is 10547.38
The coefficient for cut is 184.96
The coefficient for color is 751.33
The coefficient for clarity is 881.28
The coefficient for depth is -25.04
The coefficient for table is -30.67
The coefficient for length is -2478.34
The coefficient for width is 2106.02
The coefficient for height is -809.93
```

- We can see that 'carat' has the highest coefficient. It indicates that with 1 unit increase in carat, the price of stone will increase by 10,547.38 INR and the price change will be the highest when the value of carat increases among other variables, given that all other variables remain constant.
- If the value increases for any of the attributes with positive coefficients, from the above list, the price will also increase, given that all other variables remain constant.
- If the value increase for any of the attributes with negative coefficients, from the above list, the price will decrease, given that all other variables remain constant.
- Intercept for the model is = -224
- From the below given R-square values, we can infer that almost 90% of the variation in the cubic zirconia price is explained by the predictors in the model for both training and testing set.

```
R-square training data: 0.8992
R-square testing data: 0.896
```

- The Root Mean Square Error (RMSE) tells the average distance between the predicted values and the actual values in the dataset. The lower the RMSE, the better a given model is able to "fit" a dataset. As we can see below, the RMSE of training data is the lowest, but the difference is not huge between training and testing dataset RMSE.

```
Root Mean Square Error for training data: 1272.34
Root Mean Square Error for testing data: 1307.89
```

Using statsmodels (ols):

The difference between linear regression models created using scikit learn library and statsmodels is that statsmodels generates the regression summary automatically which makes it easy to display all the parameters at once.

We passed x & y of train datasets. Here is the regression summary for training set:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	1.868e+04
Date:	Wed, 19 Oct 2022	Prob (F-statistic):	0.00
Time:	13:50:44	Log-Likelihood:	-1.6152e+05
No. Observations:	18853	AIC:	3.231e+05
Df Residuals:	18843	BIC:	3.231e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-224.0045	881.393	-0.254	0.799	-1951.614	1503.605
carat	1.055e+04	95.306	110.668	0.000	1.04e+04	1.07e+04
cut	184.9610	15.256	12.124	0.000	155.057	214.865
color	751.3252	19.271	38.987	0.000	713.552	789.098
clarity	881.2788	12.083	72.935	0.000	857.595	904.963
depth	-25.0365	11.298	-2.216	0.027	-47.182	-2.891
table	-30.6660	4.934	-6.215	0.000	-40.337	-20.995
length	-2478.3361	166.147	-14.917	0.000	-2803.999	-2152.673
width	2106.0216	171.844	12.255	0.000	1769.193	2442.851
height	-809.9327	140.343	-5.771	0.000	-1085.017	-534.848

Omnibus:	3786.673	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	124080.897
Skew:	0.157	Prob(JB):	0.00
Kurtosis:	15.564	Cond. No.	8.14e+03

Table 1. 16: Regression Summary 1

- We can observe that the intercept (-224) and coefficients for the train dataset are same in both the models.
-

Interpretation of R-squared

- R-squared value tells us that our model can explain almost 90% of the variance in training set.

Interpretation of Coefficients (coef)

- The coefficients indicate how one unit change in x can affect y.
- The sign of coefficient indicates if the relationship is positive or negative.
- For instance, as per the above summary, 1 unit increase in depth will decrease 25.04 units in price, as the coefficient value given is negative for depth (-25.04).
- However, when we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

Interpretation of P-value (P>|t|)

- For each predictor variable there is a null (Ho) and alternate (Ha) hypothesis:
  - Ho: Predictor variable is not significant
  - Ha: Predictor variable is significant
- If the level of significance is set to 5% (0.05), the p-value greater than 0.05 indicates that the corresponding predictor variables are not significant.
- However, if multicollinearity is present in our data, the p-values will also change.

Multicollinearity check

- Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, p-values cannot be trusted to identify independent variables which are statistically significant.
- We have used Variation Inflation Factor (VIF) method to detect multicollinearity.
- VIF measures the inflation of the variances of the regression coefficients estimates due to collinearities that exist among the predictors. It is a measure of how much the variance of the estimated regression coefficient is inflated by the existence of correlation among the predictor variables in the model.

```
VIF Values:
 const      9042.41
carat        24.13
cut           1.33
color         1.08
clarity       1.21
depth         3.18
table         1.40
length      409.16
width       431.31
height      112.11
dtype: float64
```

- The VIF values indicate that the features carat, length, width and height are correlated with one or more independent features, as their VIFs are more than 5. The ideal value of VIF should range between 1 to 5.
- We now need to understand by dropping which of these features will not affect our regression model. If the regression model is completely fine after dropping these features, maybe we don't need them because the information they bring to the table is already been brought by other features because they are linearly related.
- Ultimately, we will drop one or more features that have the least impact on the R-squared of the model.
- We removed multicollinear columns one by one and observed the effect on our predictive model.
- We already know that depth and table features are derivatives of length, width and height. In other words, length, width and height are already explained by depth and table. And since the VIFs of

length, width and height are extremely high, we started by removing these 3 variables one after the other.

- After dropping 'width' column:

```
After dropping width:
R-squared: 0.898    |    Adjusted R-squared: 0.898


VIF Values:
 const      7788.54
carat        24.08
cut           1.27
color         1.08
clarity       1.20
depth         2.75
table         1.34
length      119.51
height       99.20
dtype: float64
```

- There is only a drop of 0.001 in R-square after dropping 'width'. We dropped 'length' column next:

```
After dropping length:
R-squared: 0.898    |    Adjusted R-squared: 0.898


VIF Values:
 const      4720.12
carat        19.86
cut           1.27
color         1.08
clarity       1.19
depth         1.38
table         1.34
height       20.21
dtype: float64
```

- Dropping 'length' did not affect the R-square value at all. We dropped 'height' column next:

```
After dropping height:
R-squared: 0.896    |    Adjusted R-squared: 0.896


VIF Values:
 const      4701.28
carat         1.23
cut           1.27
color         1.08
clarity       1.17
depth         1.29
table         1.34
dtype: float64
```

- Dropping 'height' column decreased R-squared value by only 0.002. The overall drop in R-square value is on 0.003, which is very less. We can conclude that 'length', 'width' and 'height' columns are not contributing significantly to the model.
- The variance inflation factors are now less than 2 and very close to 1. We can conclude that the variables are not correlated now and multicollinearity doesn't exist at this stage.

- This is the final model we have:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.896
Method:	Least Squares	F-statistic:	2.699e+04
Date:	Wed, 19 Oct 2022	Prob (F-statistic):	0.00
Time:	14:19:21	Log-Likelihood:	-1.6184e+05
No. Observations:	18853	AIC:	3.237e+05
Df Residuals:	18846	BIC:	3.238e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1407.0301	646.329	-2.177	0.029	-2673.892	-140.168
carat	8533.6941	21.889	389.858	0.000	8490.789	8576.599
cut	150.7543	15.165	9.941	0.000	121.030	180.479
color	745.1896	19.593	38.034	0.000	706.786	783.593
clarity	936.4666	12.080	77.522	0.000	912.789	960.145
depth	-51.2966	7.335	-6.993	0.000	-65.675	-36.919
table	-41.7989	4.912	-8.509	0.000	-51.427	-32.171

Omnibus:	3613.569	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	49695.533
Skew:	0.524	Prob(JB):	0.00
Kurtosis:	10.884	Cond. No.	5.79e+03

Table 1. 17: Regression Summary 2

- Now that we do not have multicollinearity in our data, the coefficients and p-value of the coefficients have become reliable for interpretation.
- If the level of significance is set to 5% (0.05), the p-value greater than 0.05 indicates that the corresponding predictor variables are not significant. But for our data, all the p-values are less than 0.05, indicating that all the predictor variables remaining in the training data set are significant to the dependent variable - price.

Model parameters:

```
const     -1407.0301
carat      8533.6941
cut         150.7543
color       745.1896
clarity     936.4666
depth       -51.2966
table       -41.7989
dtype: float64
```

Linear Regression Equation:

price = -1407.03 + (8533.69 * carat) + (150.75 * cut) + (745.19 * color) + (936.47 * clarity) + (-51.3 * depth) + (-41.8 * table)

1.3.4 Assumptions of Linear Regression

The assumptions are essential conditions that should be met before we draw inferences regarding the model estimates or use the model to make prediction.

For linear regression, we need to check if the following assumptions hold:

1. **Linearity**: As we can observe in the below plot, there is no pattern in the residual vs fitted values, as the data points seem to be randomly distributed. We can say that the model is linear.

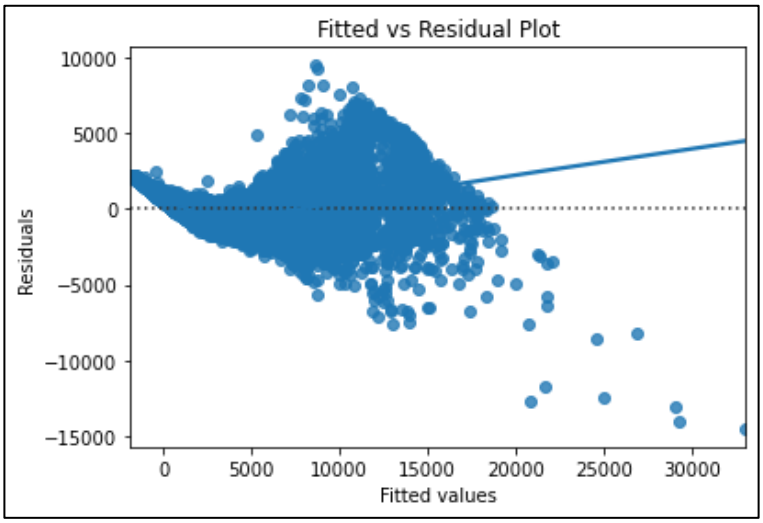


Figure 1. 6: Linearity – Fitted vs Residuals

2. **Independence**: Since there is a lack of pattern, we can conclude that the residuals are independent of one another.
3. **Homoscedasticity**: If the variances of the residuals are symmetrically distributed across the regression line, then data is said to be homoscedastic. Using goltfeldquandt, which is a hypothesis test:

   Ho: Residuals are homoscedastic

   Ha: Residuals are hetroscedastic

   As the p-value (0.01) is less than the significance level of 0.05, this assumption doesn't hold true.
4. **Normality of residuals**: Since the p-value (0.0) received from Shapiro Wilk test is less than significance level of 0.05, strictly speaking – the residuals are not normally distributed.

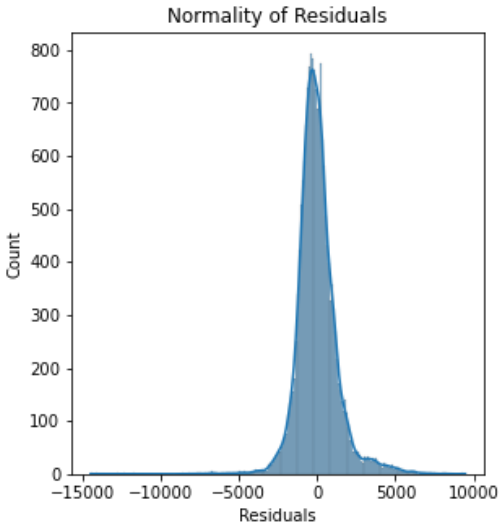


Figure 1. 7: Normality of Residuals

5. **No or low multicollinearity**: We had already checked for multicollinearity and adjusted our data accordingly. Now the data has no or very less correlation.

As we can see that most of the assumptions are satisfied, we can move ahead with the interpretation and derive inferences from the model.

1.3.5 Predictions on the test data

We dropped the columns from the testing dataset which we removed from the training dataset (length, width, height). Here is the sample of test dataset:

	const	carat	cut	color	clarity	depth	table
25823	1.0	0.51	3	1	3	61.90	56.0
17845	1.0	0.38	3	1	2	60.49	58.0
8441	1.0	0.90	2	2	2	61.50	58.0
2075	1.0	2.24	3	1	2	62.10	53.6
3998	1.0	0.40	3	1	3	62.10	58.0

Table 1. 18: Sample of Testing Dataset

Here is the regression summary for test set:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.893
Model:	OLS	Adj. R-squared:	0.893
Method:	Least Squares	F-statistic:	1.125e+04
Date:	Wed, 19 Oct 2022	Prob (F-statistic):	0.00
Time:	14:48:14	Log-Likelihood:	-69557.
No. Observations:	8080	AIC:	1.391e+05
Df Residuals:	8073	BIC:	1.392e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-744.5372	1032.710	-0.721	0.471	-2768.916	1279.841
carat	8685.4720	34.373	252.683	0.000	8618.092	8752.852
cut	192.2117	23.944	8.028	0.000	145.276	239.148
color	852.6235	30.599	27.865	0.000	792.642	912.605
clarity	963.5256	18.636	51.703	0.000	926.995	1000.056
depth	-58.8698	11.866	-4.961	0.000	-82.131	-35.609
table	-52.8833	7.669	-6.896	0.000	-67.916	-37.850

Omnibus:	1642.832	Durbin-Watson:	2.021
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20093.531
Skew:	0.620	Prob(JB):	0.00
Kurtosis:	10.625	Cond. No.	5.91e+03

Table 1. 19: Regression Summary 3

Comparison of train vs test values:

	R-sqr	Adj.R-sqr	RMSE	MAE
Train	0.896	0.896	1294.062	29.905
Test	0.893	0.893	1328.344	30.029

Table 1. 20: Train vs Test

- We can see that R-square and Adjusted R-square values for both train and test dataset is very clos to equal, there is difference is only of 0.003.
- The root mean square error (RMSE) score is also very close for both the sets.
- MAE indicates that our current model is able to predict price within a mean error of 30.029 units on the test data.
- Hence, we can conclude that the train model is good for prediction as well as inference purposes.

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

As per the below linear equation from training set, we can make following business insights:

price = -1407.03 + (8533.69 * carat) + (150.75 * cut) + (745.19 * color) + (936.47 * clarity) + (-51.3 * depth) + (-41.8 * table)

Carat	for 1 unit increase in carat, the price will increase by 8,533.69 INR.
Cut	for 1 unit increase in cut of cubic zirconia, the price will increase by 150.75 INR.
Color	for 1 unit increase in color, the price will increase by 745.19 INR.
Clarity	for 1 unit increase in clarity, the price will increase by 936.47 INR.
Depth	for 1 unit increase in depth, the price will decrease by 51.3 INR.
Table	for 1 unit increase in table, the price will decrease by 41.8 INR.

Table 1. 21: Linear Equation Interpretation

- The above-mentioned changes in price will occur for each feature when other features remain constant.
- In case of ordinal features – cut, color and clarity, the interpretation will be a little different. If the quality of cut, color or clarity improves, the price will go up according to their respective coefficients.

The best 5 attributes that are most important are:

carat	8533.6941
clarity	936.4666
color	745.1896
cut	150.7543
table	-41.7989

- Attributes with positive values are directly linked to the price of cubic zirconia, as discussed above. More the value, higher the price.
- 'table' value is negative, which means the cubic zirconia stone with less value of table will be more profitable.

Recommendation:

- The price is tremendously high for the cubic zirconia stone when it is of a higher carat weight.
- The price is higher for the cubic zirconia stone of "Internally Flawless" (IF) clarity. The price is getting lesser as the clarity is going down in this order: Internally Flawless > Very Very Small Inclusions > Very Small Inclusions > Small Inclusions > Inclusions.
- The price is higher for the cubic zirconia stone of "Colorless" color grade. The price is lesser for "Near_colorless" stones.
- The price is higher for the cubic zirconia stone of "Premium" cut. The price is less for "Very Good" then "Good" cut stones.

- There is a decrement in price of cubic zirconia stone by a larger factor if the depth of the stone is high.
- There is a decrement in price of cubic zirconia stone by a larger factor if the size of table of the stone is high.
- The company need to manufacture more and more cubic zirconia stones with higher carat weight and lesser depth and table size, also the stones will be more profitable when their cut, clarity and color is of higher grade as we discussed above. The stones of these standards will be of highest quality and very profitable.
- The average quality of the stone attributes will result in medium profitability.
- Qualities lesser than that will be less profitable, as the market for low grade cubic zirconia is not wide. People will go with other type of stones of lower grade if they are ready to compromise on the quality. Business can invest less in the manufacturing of less quality cubic zirconia stones.

Problem 2 – Logistic Regression & Linear Discriminant Analysis

Introduction

A tour and travel agency which deals in selling holiday packages, has provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. 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

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

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.

2.1.1 Sample of dataset

Here are the top 5 rows (sample) of the dataset:

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

Table 2. 1: Data Sample

- Dataset has 7 valid variables and as we can see the first column (*Unnamed: 0*) only contains serial numbers which are not relevant, we can drop it from the dataset.
- 'Holliday_Package' column name has spelling error and hence been fixed.

	Holiday_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 2. 2: Modified Data Sample

2.1.2 Check for Duplicate Records

```
Number of duplicate records: 0
```

As we can see there are no duplicate records present in the data.

2.1.3 Types of variables in the dataset

```
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holiday_Package    872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
```

- There are a total of 872 observations (rows) under 7 features (columns) in the dataset.

- There are 5 variables of int64 and 2 of object datatype.
- 'Holiday_Package' is the dependent / target variable

2.1.4 Missing / null values in the dataset

```
Holiday_Package      0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
```

There are no missing values present in the dataset. There are some 0 (zero) values in 'no_young_children' and 'no_older_children' columns, but those values are valid, as some people may have either young or older children or no children at all, which is perfectly normal.

2.1.5 Descriptive Statistics

Describe function provides a table indicating the count of variables, mean, standard deviation and other values for the 5-point summary that includes (min, 25%, 50%, 75% and max) for numeric variables. 50% in the table is also known as median.

	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 2. 3: Data Description for Continuous Columns

For object/categorical columns, describe function shows the total count, unique values in each column, most frequent value and value frequency in each column.

	count	unique	top	freq
Holiday_Package	872	2	no	471
foreign	872	2	no	656

Table 2. 4: Data Description for Categorical Columns

- Salary of employees ranges from 1,322 till 236,961.
- The data captures employees with age starting from 20 till 62. The average age of employees is 39.
- 75% of employees have completed only 12 years of formal education.
- There are very less employees with children younger than 7 years. Employees with no young children fall under 75% of the data. The maximum number of young children employees have is 3.
- Employees falling under 25th quartile of the data have no older children.

- Employee who is earning highest salary is aged 39, has 4 older children, but has not opted for the holiday package.

	Holiday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
623	no	236961	39	12	0	4	no

Table 2. 5: Emp with highest salary

- Employee who is earning the lowest salary, is aged 57 and has no children, but has opted for the holiday package.

	Holiday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
310	yes	1322	57	12	0	0	no

Table 2. 6: Emp with lowest salary

- Data is focused on 2 preferences of opted for Holiday_Package, with 'no' (not opted) being the most popular preference among employees.

```
HOLIDAY_PACKAGE :  2
yes    401
no     471
Name: Holiday_Package, dtype: int64


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

- Majority of employees – 656 are not foreigner.

2.1.6 Check for outliers

Boxplots have been plotted for numerical variables to check for outliers:

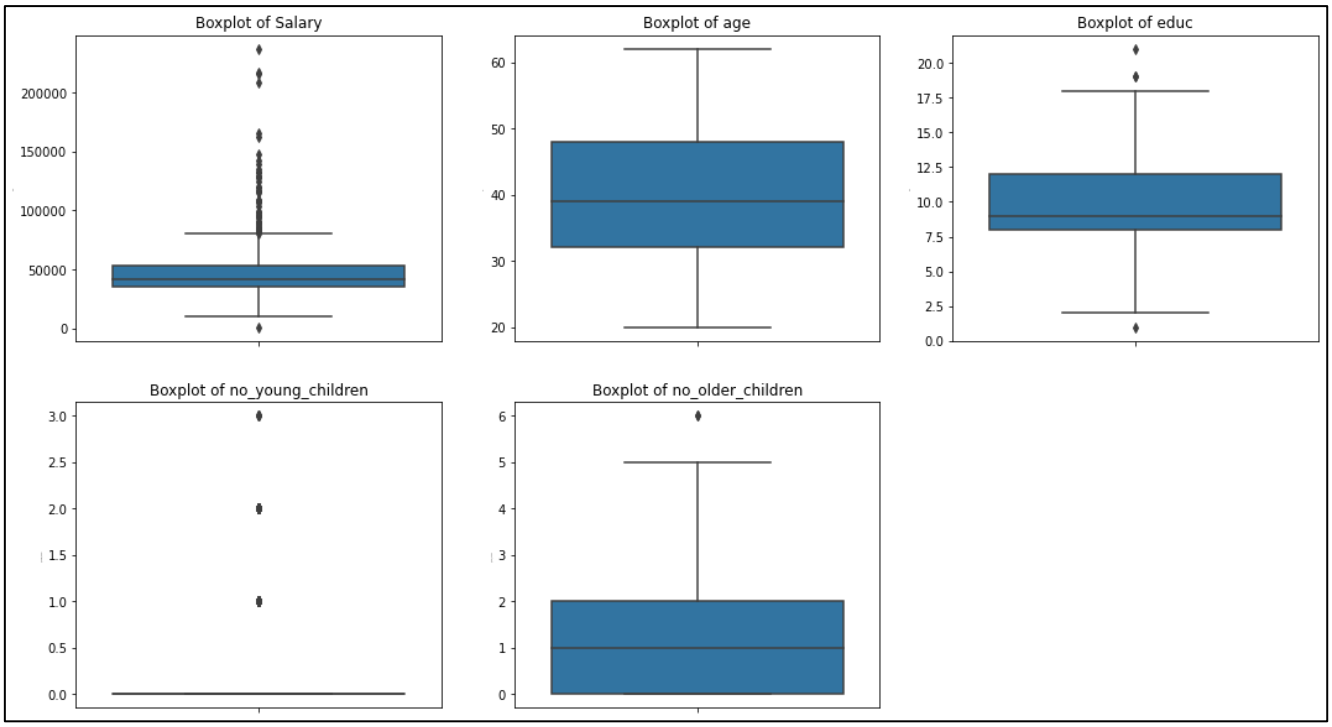


Figure 2. 1 Boxplot for Outliers

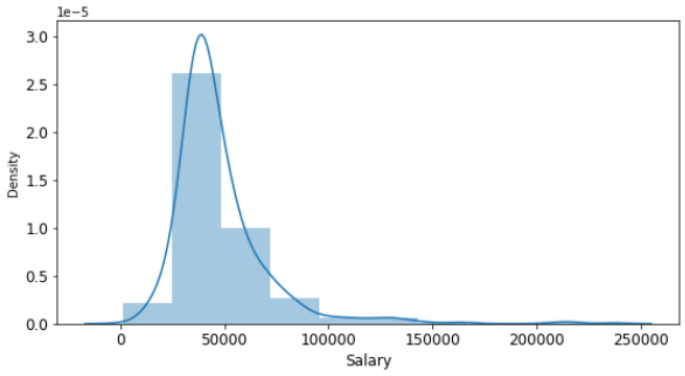
- The small dots outside the whiskers of boxplots denote outliers. All the numeric columns in our data have outliers present.
- There are no outliers in 'age' column.
- The outliers present in the data appear to be valid. Hence, outlier treatment is not necessary.

2.1.7 Univariate analysis

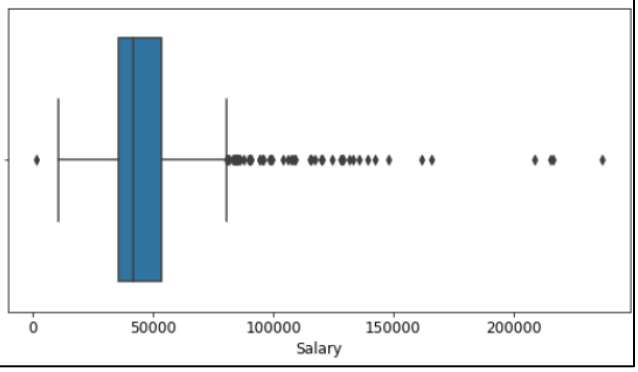
Univariate analysis is performed for all the numeric variables individually to display their statistical description. Visualized the variables using distplot to view the distribution and the box plot to view 5-point summary and outliers if any.

```
Description of Salary
.....................................................................
count       872.000000
mean      47729.172018
std       23418.668531
min        1322.000000
25%       35324.000000
50%       41903.500000
75%       53469.500000
max      236961.000000
Name: Salary, dtype: float64
```

Distribution of Salary

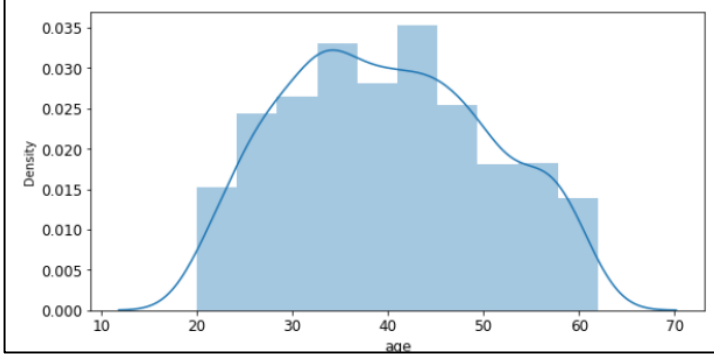


boxplot of Salary

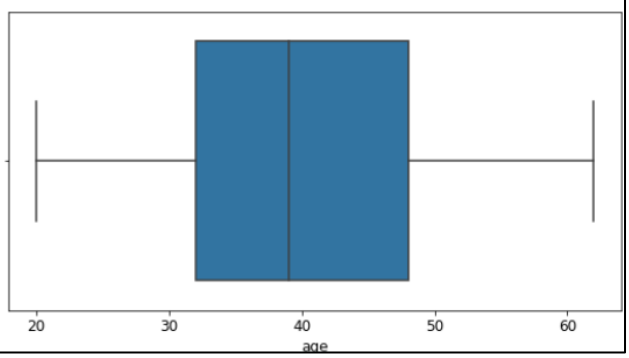



```
Description of age
.....................................................................
count    872.000000
mean      39.955275
std       10.551675
min       20.000000
25%       32.000000
50%       39.000000
75%       48.000000
max       62.000000
Name: age, dtype: float64
```

Distribution of age

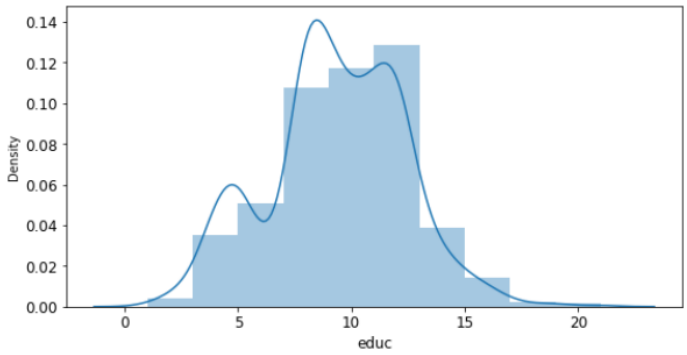


boxplot of age

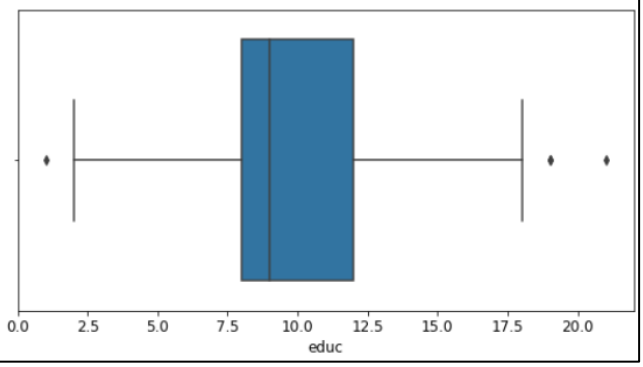


```
Description of educ
.....................................................................
count    872.000000
mean       9.307339
std        3.036259
min        1.000000
25%        8.000000
50%        9.000000
75%       12.000000
max       21.000000
Name: educ, dtype: float64
```

Distribution of educ

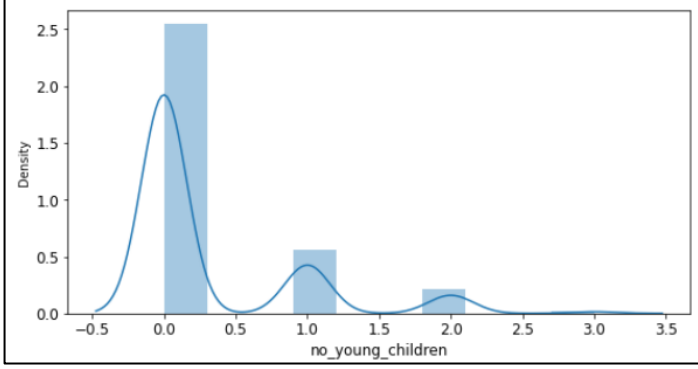


boxplot of educ

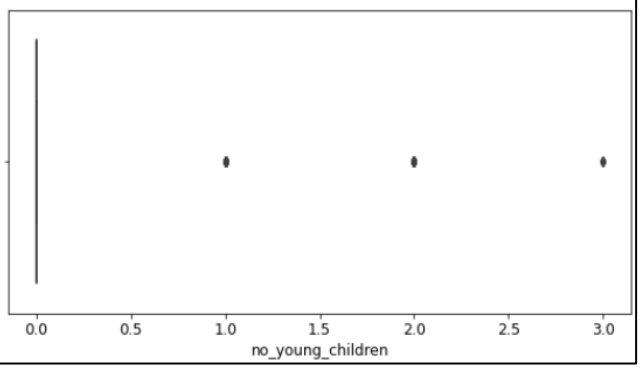



```
Description of no_young_children
.....................................................................
count    872.000000
mean       0.311927
std        0.612870
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        3.000000
Name: no_young_children, dtype: float64
```

Distribution of no_young_children



boxplot of no_young_children



```
Description of no_older_children
.....................................................................
count    872.000000
mean       0.982798
std        1.086786
min        0.000000
25%        0.000000
50%        1.000000
75%        2.000000
max        6.000000
Name: no_older_children, dtype: float64
```

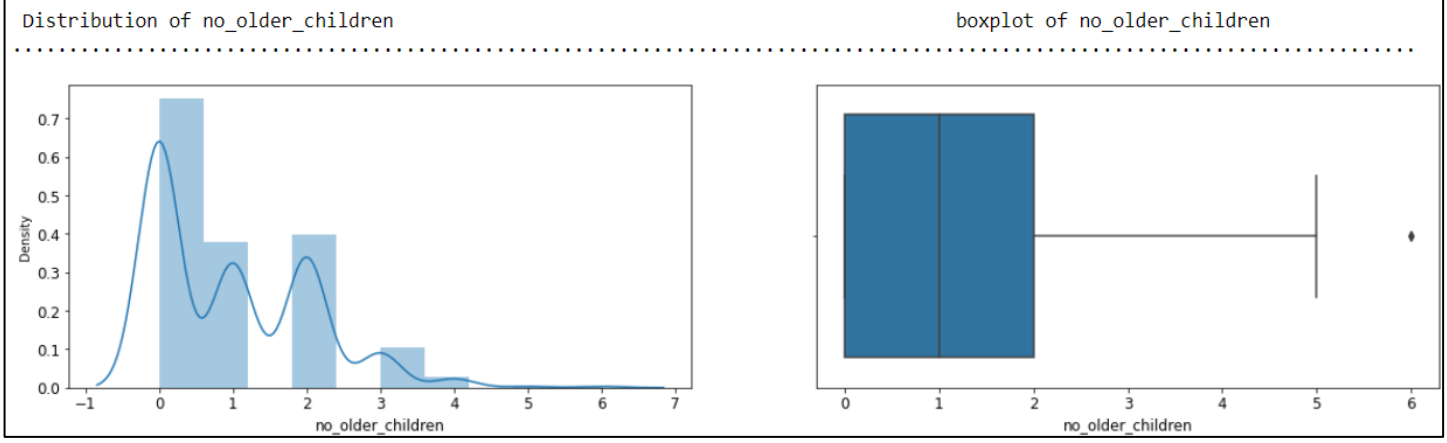


Figure 2. 2: Univariate Analysis

	Kurtosis	Skewness
Salary	15.85	3.10
age	-0.91	0.15
educ	0.01	-0.05
no_young_children	3.11	1.95
no_older_children	0.68	0.95

Table 2. 7: Kurtosis & Skewness

- There are 5 numeric fields in the dataset.
- From the boxplots we can observe the presence of outliers, but there is no need to treat them since they are valid.
- The variables show multimodal pattern.
- Distribution for 'Salary' is highly positively skewed, and has the highest kurtosis/peak.
- Around 50% of the employees fall under the age bracket of 30 to 50.
- Employees who have completed formal education more than 12.5 year fall beyond 75% of total employees. Which is very less.

2.1.8 Bivariate analysis

Bivariate analysis of continuous columns with target variable:

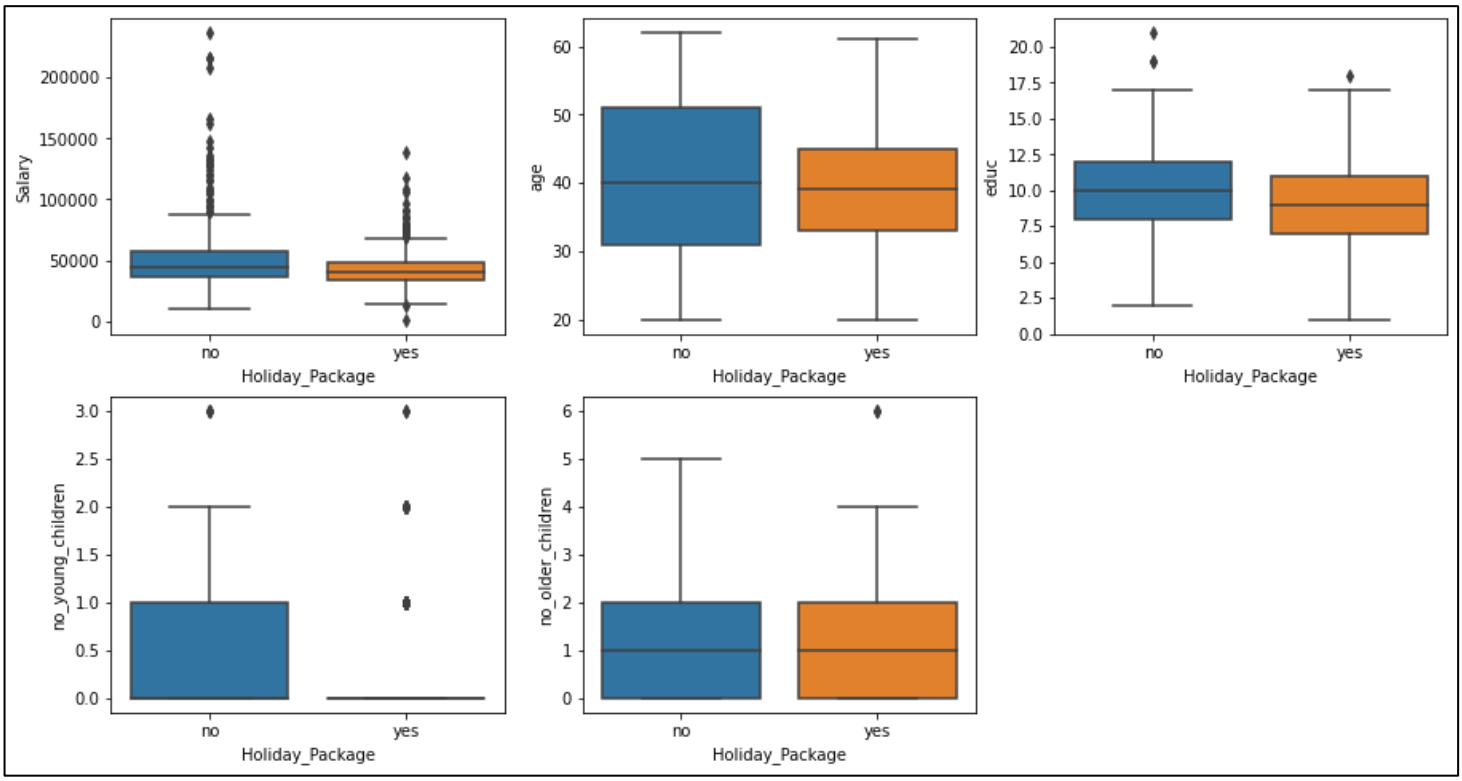


Figure 2. 3: Bivariate Analysis - Continuous

- The median is more or less the same for employees opting and not option for holiday package, for all the variables. Slight patterns can be observed, showcasing majority of the employees have not opted for the holiday package.
- We can observe that the people earning more than 50,000 have not opted for the package.
- Age shows normal distribution for opted and not opted, but the distribution is wider for opted. We can also infer that, employees with more than around 45 years of age are not opting for the packages.
- Educ shows some skewness in the distribution between opted and not opted. Median of not opted is slightly higher than opted. Employees who have completed 11 year of formal education have not opted for the package.
- Number of young children column shows clear distinction between employees who opted and not opted for package. Employees who have opted for holiday package have fewer young children. Whereas employees who have not opted for package shows a wider distribution indicating more number of young children.
- Number of older children column shows similar distribution between opted and not opted.

Bivariate analysis of categorical columns with target variable:

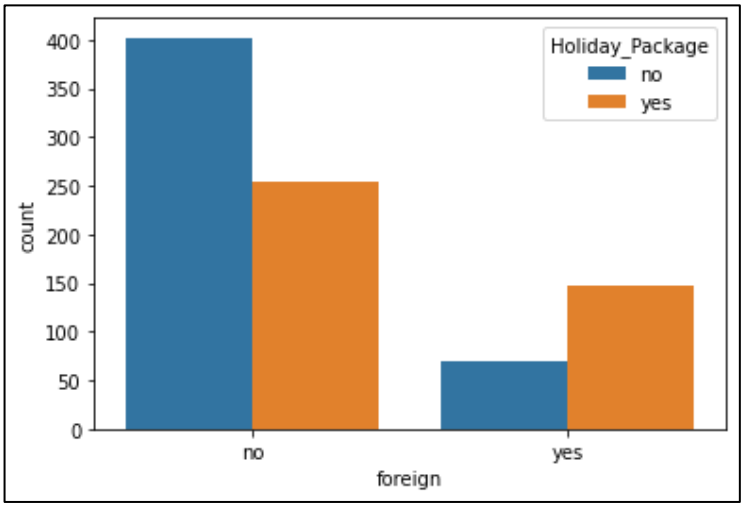


Figure 2. 4: Bivariate Analysis - Categorical

- Employees who are not foreigners have the highest number of employees who have not opted for the holiday package.
- More employees who are foreigners have opted for the holiday package.

2.1.9 Multivariate analysis

Pair plot (numeric variables):

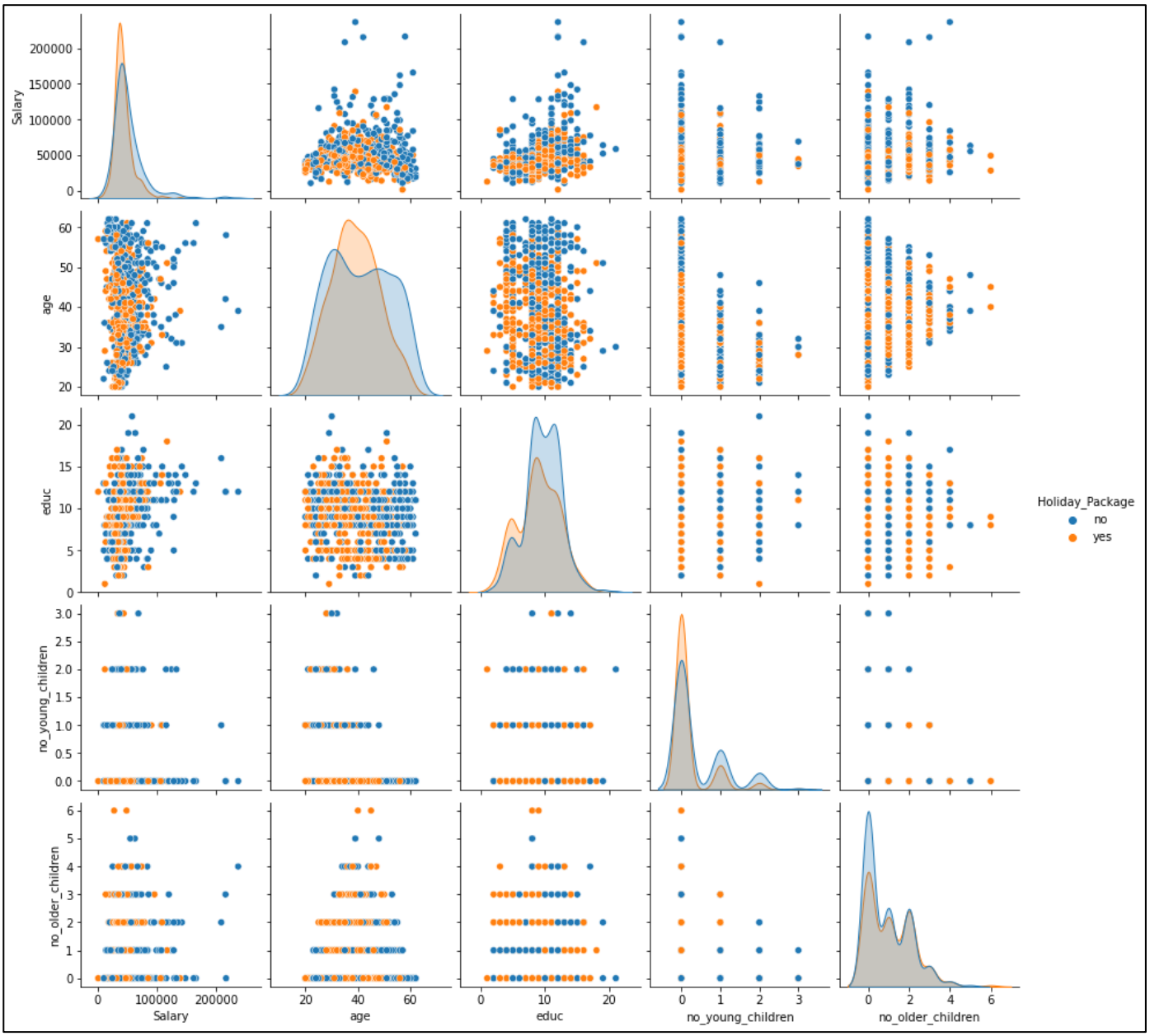


Figure 2. 5: Pairplot

- As we can observe from distribution plots on the diagonal, that the values in target variable (yes & no) is not well separated for any of the independent variables.
- From the scatter plots we can infer that the distribution of opted and not opted are more or less the same for all the variables.
- Most of the attributes are right skewed and the distribution is multimodal.
- As the age and education is increasing, salary is not increasing at the same pace. We can say that only for a few employees who are older and have higher education are earning more.

Correlation plot (Heatmap) of numeric variables:

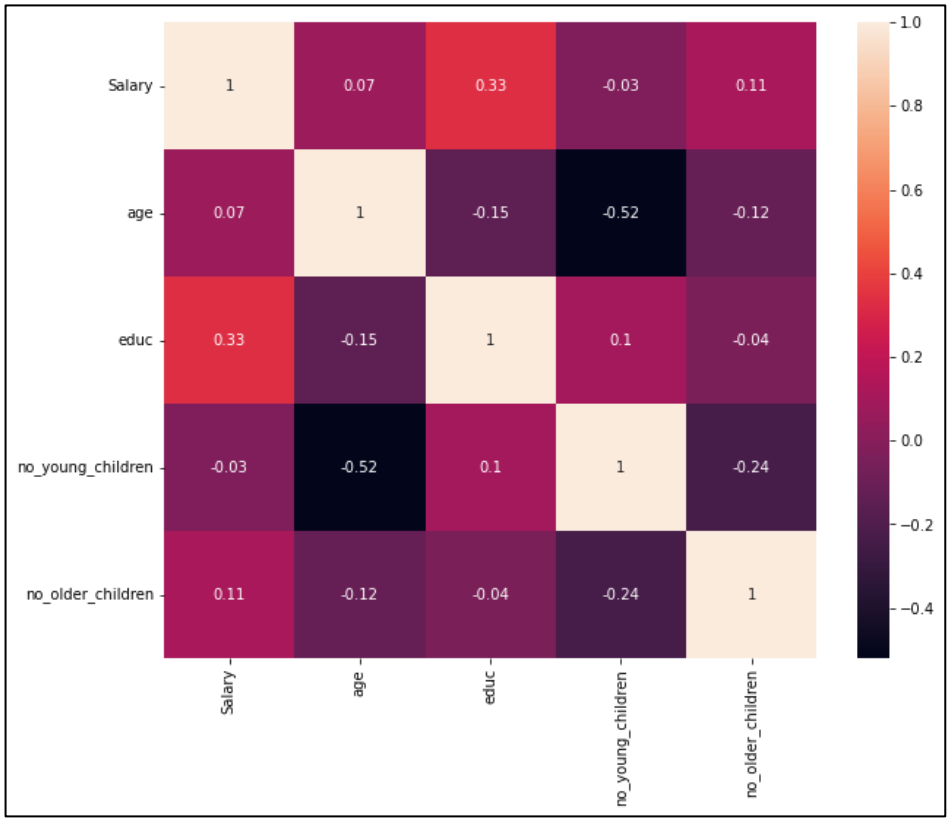


Figure 2. 6: Correlation Plot

- We can see moderate negative correlation between no_young_children and age. It can be said that the employees who are older, have less children below the age of 7.
- Other than this there is no significant correlation among the variables.
- The only assumption for logistic regression model of low or no multicollinearity holds true in this case.

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).

2.2.1 Data Encoding

For prediction models the data to pass should be in numeric/categorical codes format only. The variables with string values in our dataset need to be converted to integer format.

- Target variable 'Holiday_Package' has been converted into integer by replacing 'no' with 1, because 'no' is the class of interest in this case, we need to focus on the employees who have not opted for holiday packages. Value 'yes' has been replaced by class 0.

BEFORE	AFTER
no 471 yes 401 Name: Holiday_Package	1 471 0 401 Name: Holiday_Package

Table 2. 8: Data Encoding 1

- 'educ' column is in integer format but we can get better interpretation out of it if the education data can be categorized as 'Primary' – for 1 to 8 years, 'Secondary' – for 9 to 12 years and 'Tertiary' – for 13 years of education and above. Then manually encoded 'educ' column and assigned an order to the category, where Tertiary = 3 being the heights and Primary = being the lowest.

BEFORE	AFTER
Secondary 428 Primary 344 Tertiary 100 Name: educ, dtype: int64	2 428 1 344 3 100 Name: educ, dtype: int64

Table 2. 9: Data Encoding 2

- Furthermore, we have performed dummy variable encoding on 'foreign' column, using drop_first to get n-1 dummies out of n categorical levels by removing the first level.

After encoding, this is how the dataset appear:

	Holiday_Package	Salary	age	educ	no_young_children	no_older_children	foreign_yes
0	1	48412	30	1	1	1	0
1	0	37207	45	1	0	1	0
2	1	58022	46	2	0	0	0
3	1	66503	31	2	2	0	0
4	1	66734	44	2	0	2	0

Table 2. 10: Modified Data Sample

Types of variables:

```
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holiday_Package    872 non-null    int32
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int32
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign_yes        872 non-null    uint8
dtypes: int32(2), int64(4), uint8(1)
```

- All the variables are in numeric format.
- We have a total of 7 variables in our encoded dataset, out of which 6 are independent variables and 'Holiday_Package' is the target variable.

2.2.2 Data Split

The target variable in our encoded dataset is 'Holiday_Package', where 1 = not opted; 0 = opted. Here is the proportion of values in the target variable:

```
Percentage of "Not Opted" in target variable: 54.01 %
Percentage of "Opted" in target variable: 45.99 %
```

The proportion seems to be balanced enough to move forward with models building.

The data has been first divided in to independent and dependent (target) variables, x and y respectively.

The data is now split into training and testing set with both sets having 70% and 30% of the data, respectively. Here is the proportion of target variable in both the sets:

```
Percentage of "Not Opted" in target variable in Training set: 53.93 %
Percentage of "Opted" in target variable in Training set: 46.07 %


Percentage of "Not Opted" in target variable in Testing set: 54.2 %
Percentage of "Opted" in target variable in Testing set: 45.8 %
```

2.2.3 Logistic Regression Model

In the first instance, we built the model without changing default values of parameters. After observing performance of the model, we will decide the best parameters to better fit the model.

Confusion matrix and classification report:

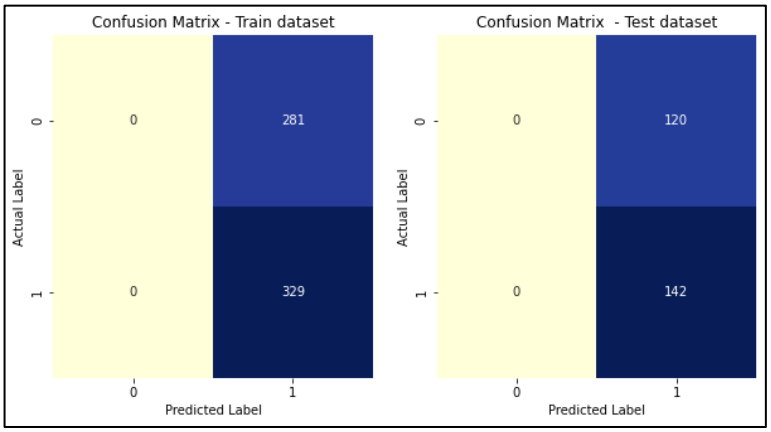


Figure 2. 7: LR Confusion Matrix 1

Classification Report - Train dataset

	precision	recall	f1-score	support
0	0.00	0.00	0.00	281
1	0.54	1.00	0.70	329
accuracy			0.54	610
macro avg	0.27	0.50	0.35	610
weighted avg	0.29	0.54	0.38	610

Classification Report - Test dataset

	precision	recall	f1-score	support
0	0.00	0.00	0.00	120
1	0.54	1.00	0.70	142
accuracy			0.54	262
macro avg	0.27	0.50	0.35	262
weighted avg	0.29	0.54	0.38	262

Figure 2. 8: LR Classification Report 1

As we can see that the logistic regression model did not account for 0 class at all. The Recall is 1 for both train and test set, which appears to be over fitted.

We performed GridSearch crossvalidation for this model, by passing multiple combination for hyper-parameters, to find out the best parameters to build a model that performs well. After running GridSearch cross validation, here are the observations:

- Best parameters: 'max_iter': 450, 'penalty': 'none', 'solver': 'newton-cg', 'verbose': True.

We again built the model using best parameters, obtained from GridSearch crossvalidation:

Confusion matrix and classification report:

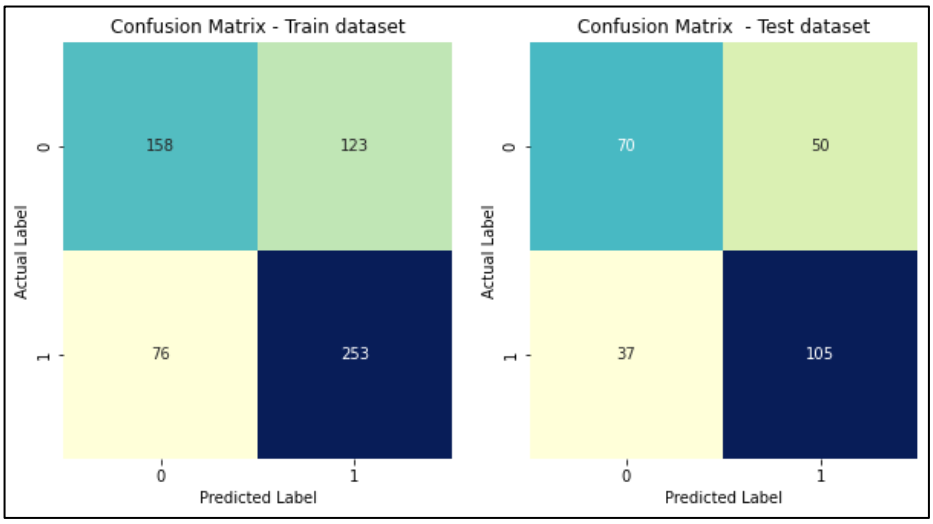


Figure 2. 9: LR Confusion Matrix 2

Classification Report - Train dataset

	precision	recall	f1-score	support
0	0.68	0.56	0.61	281
1	0.67	0.77	0.72	329
accuracy			0.67	610
macro avg	0.67	0.67	0.67	610
weighted avg	0.67	0.67	0.67	610

Classification Report - Test dataset

	precision	recall	f1-score	support
0	0.65	0.58	0.62	120
1	0.68	0.74	0.71	142
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.67	262

Figure 2. 10: LR Classification Report 2

- Above results indicate that we have reduced the overfitting of the logistic regression model, and now the Recall and F1-score of train and test has less difference.
- The ROC-AUC score for both the sets are equal. Based on this observation, we can say that the testing sample is performing exactly as good as the training sample.

```
ROC - AUC score for training set is 0.73
ROC - AUC score for testing set is 0.74
```

- ROC Curve:

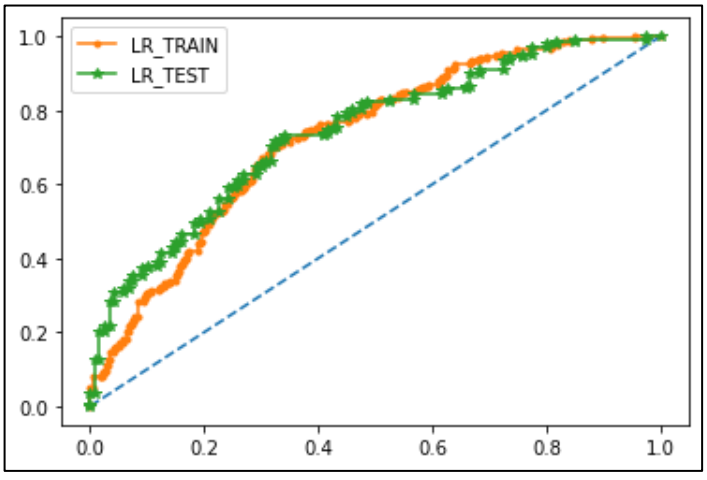


Figure 2. 11: LR ROC Curve

2.2.4 Linear Discriminant Analysis

In the first instance, we built the model without changing default values of parameters. After observing performance of the model, we will decide the best parameters to better fit the model.

Confusion matrix and classification report:

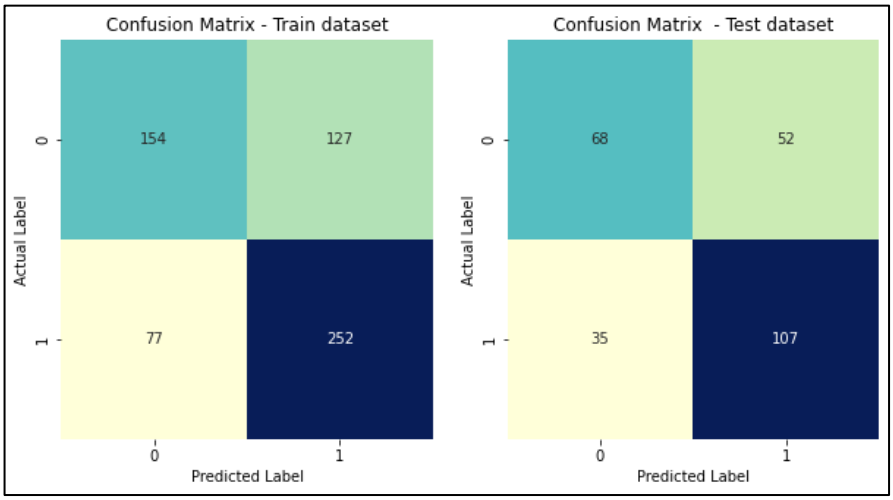


Figure 2. 12: LDA Confusion Matrix 1

Classification Report - Train dataset

	precision	recall	f1-score	support
0	0.67	0.55	0.60	281
1	0.66	0.77	0.71	329
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Classification Report - Test dataset

	precision	recall	f1-score	support
0	0.66	0.57	0.61	120
1	0.67	0.75	0.71	142
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.66	262

Figure 2. 13: LDA Classification Report 1

As we can observe that the default parameters have performed considerably well. Let's try with different parameters to see if the results can be improved.

We performed GridSearch crossvalidation for this model, by passing multiple combination of values for the parameters, to find out the best parameters to build a model that performs well. After running GridSearch cross validation, here are the observations:

- Best parameters: 'solver': 'svd', 'tol': 0.01.

We again built the model using best parameters, obtained from GridSearch crossvalidation:

Confusion matrix and classification report:

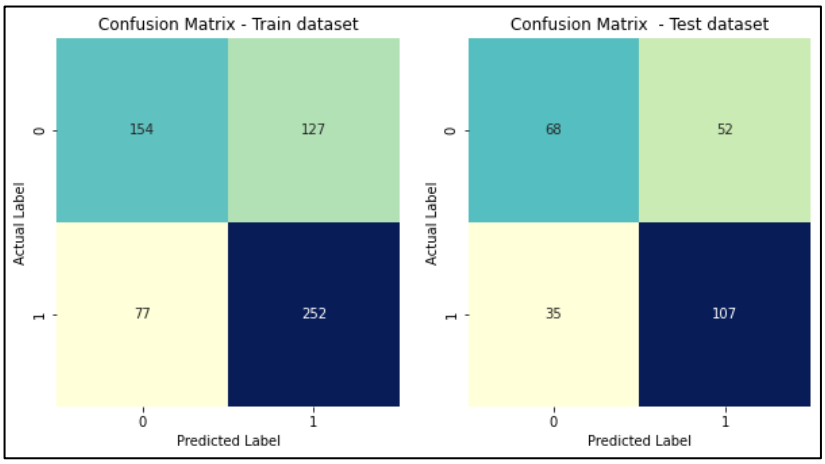


Figure 2. 14: LDA Confusion Matrix 2

Classification Report - Train dataset

	precision	recall	f1-score	support
0	0.67	0.55	0.60	281
1	0.66	0.77	0.71	329
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.66	610

Classification Report - Test dataset

	precision	recall	f1-score	support
0	0.66	0.57	0.61	120
1	0.67	0.75	0.71	142
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.66	262

Figure 2. 15: LDA Classification Report 2

- The results remained same after performing gridsaerch crossvalidation.
- The ROC-AUC score for test set is slightly higher than that of a train set. Based on this observation, we can say that the testing sample is performing a little better than the training sample.

```
ROC - AUC score for training set is 0.73
ROC - AUC score for testing set is 0.75
```

- ROC Curve:

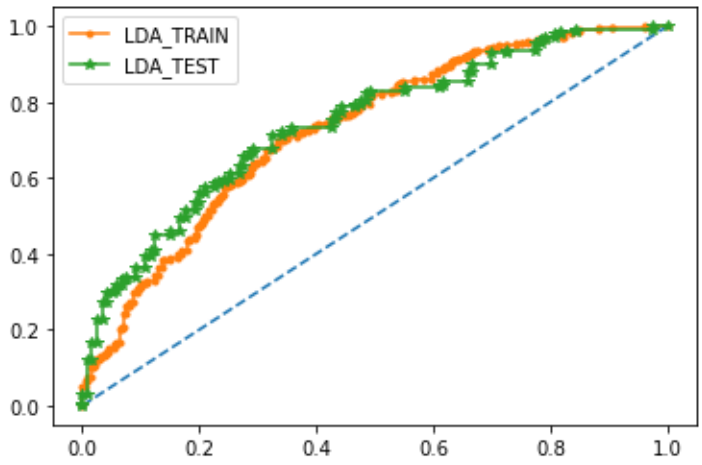


Figure 2. 16: LDA ROC Curve

Custom cut-off values to check for better performance:

We checked for possible improvement in the performance of Linear Discriminant model, by changing the custom cut-off values (default is 0.5). Below are the recall score, F1 score and confusion matrix at different cut-off values:

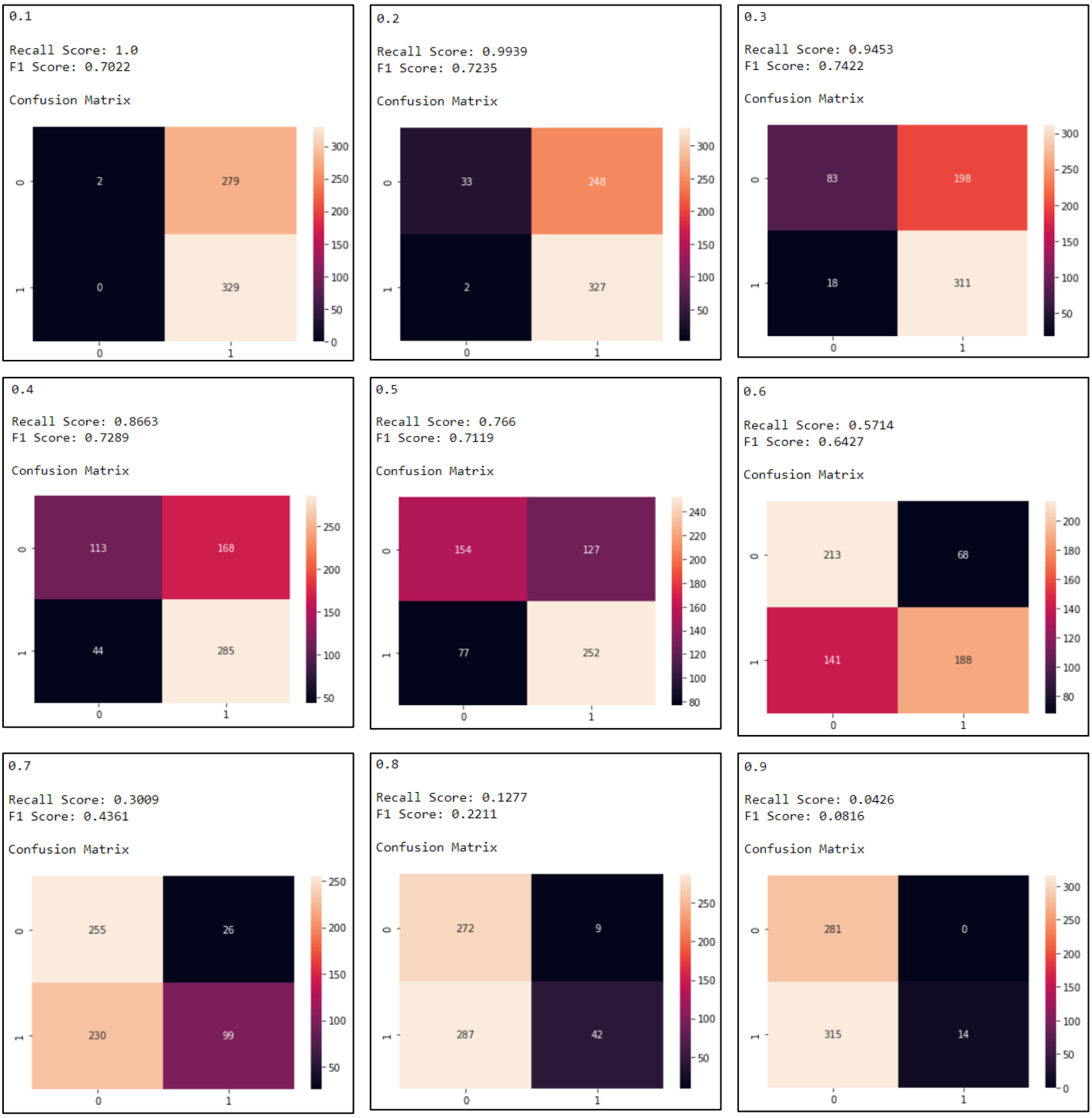


Figure 2. 17: Scores of Different Cut-off Values

- The Recall value is highest at 0.1 and then 0.2 cut-off level, this could be because of the over fitting.
- At 0.3 cut-off level the recall value is the significantly higher than what we had at 0.5 (original) cut-off. The F1 score is also the highest at 0.3 cut-off. And at 0.4 and beyond, both values are dropping largely.
- As such, we can take 30% threshold to get optimum Recall and F1 score, as it gives the best performance without manipulations in the data.

- In other words, from the model generated using 0.3 as cut-off value we can infer that the employees who have probability of not opting more than 30%, they will not opt for holiday packages for sure.
- The 30% threshold needs to be confirmed with the business if they are okay with it or the threshold is too low for them. We are basing our predictions on the default threshold of 50% (0.5 cut off).

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.

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.67	0.67	0.67	0.67
Recall	0.77	0.74	0.77	0.75
AUC	0.73	0.74	0.73	0.75
Precision	0.67	0.68	0.66	0.67
F1 score	0.72	0.71	0.71	0.71

Table 2. 11: Models Comparision

ROC Curve Comparision:

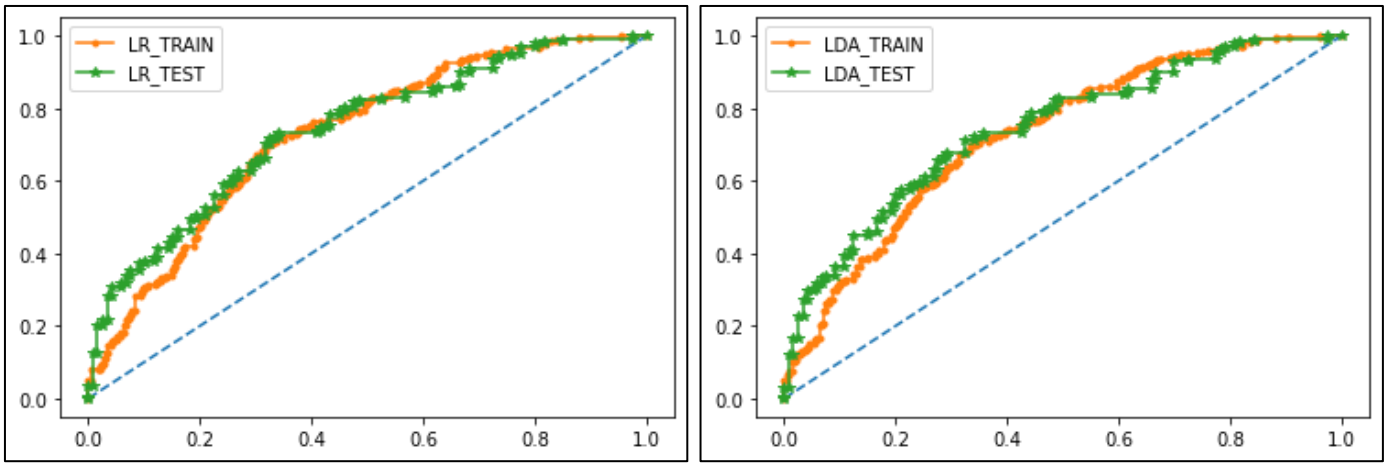


Figure 2. 18: ROC Curve Comparision

- Looking at the above table values, both Logistic Regression and Linear Discriminant models have performed very well, and the scores obtained are also very close.
- The scenario of **False Negative, where "prediction is that holiday package was opted but actually package was not opted"** as well as **True Positive, where "prediction is that holiday package was not opted and actually package was not opted"** need to be of main focus for the business. As such, the Recall score becomes of utmost importance for this case study.
- The Recall score of on training set are same for both models, but the score of testing set is slightly higher for Linear Discriminant Analysis model.
- AUC score is also slightly higher for test dataset for Linear Discriminant Model.
- The ROC curve seems to be best fit for the Linear Discriminant Analysis model, where testing set performing slightly better than training set.
- After evaluating all above factors, we can conclude that Linear Discriminant Analysis model is the best optimized for this business problem.

2.4 Inference: Based on the whole Analysis, what are the business insights and recommendations.

- We observed that, employees who are earning higher have not opted for the holiday packages. It could be because of the facilities or other components of the packages are not as per their liking or comfort level. Another reason for that could be higher earning people are not getting enough time to spare to go on holidays.
    - Agency can create some exclusive packages for high earning individuals, with 5-star hotel stay with all-inclusive package, and offer them on discount initially to attract them.
    - Similar packages designed for the span of 2-3 days holiday of lesser cost, so that this class of employees can enjoy their holidays over the weekend.
    - Workcation packages can be designed, so that employees can work at the comfort of their stay at an exotic location.
- More number of individuals of more than 45 years of age are not opting for the packages. At that age some individuals may fall under the class of high earning individuals as well. They can be attracted towards packages providing more comfort and less adventures.
- Employees with more number of young children are not opting for the package as it may be challenging for them to commute to the holiday destination with young kids onboard. Packages with commutation to nearby holiday destinations may attract this class of individuals.
- If the agency focuses on the above-mentioned factors and customize holiday packages according to the needs of specific individual class, they may be able to get more customers onboard.
- A small survey of sorts can also be done to identify the common needs of individuals in order to cater them well.

End of Report