DSBA APRIL 2021 BATCH

Predictive Modeling Report

- Gem Stones Co Ltd.
- Holiday Package

Prepared by Ruchita Abrange
August 2021

Table of Contents

Case Study "Gem Stones Co Ltd."

Figures & Table:

Table of Contents

Table of Contents

Case Study "Holiday Package"

Figures & Table:

About Case Study "Gem Stones Co Ltd." using Linear Regression

The data is about a company 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. The data contains a sample of 27,000 cubic zirconia that summarizes containing the prices and other attributes. The company is earning different profits on different prize slots.

Our Objective

The objective is to help the company 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 a better profit share. Also, the best 5 attributes that are most important need to be provided.
To understand the prediction let's explore the dataset.

Data Description

1. Carat - Carat weight of the cubic zirconia.
2. Cut - Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
3. Color - Colour of the cubic zirconia.With D being the worst and J the best.
4. Clarity - cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
5. Depth - The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
6. Table - The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
7. Price - the Price of the cubic zirconia.
8. X - Length of the cubic zirconia in mm.
9. Y- Width of the cubic zirconia in mm.
10. Z - Height of the cubic zirconia in mm.

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
5	6	1.02	Ideal	D	VS2	61.5	56.0	6.46	6.49	3.99	9502
6	7	1.01	Good	H	SI1	63.7	60.0	6.35	6.30	4.03	4836
7	8	0.50	Premium	E	SI1	61.5	62.0	5.09	5.06	3.12	1415
8	9	1.21	Good	H	SI1	63.8	64.0	6.72	6.63	4.26	5407
9	10	0.35	Ideal	F	VS2	60.5	57.0	4.52	4.60	2.76	705
10	11	0.32	Ideal	E	VS2	61.6	56.0	4.40	4.43	2.72	637
11	12	1.10	Premium	D	SI1	60.7	55.0	6.74	6.71	4.08	6468
12	13	0.50	Good	E	VS1	61.1	58.2	5.08	5.12	3.11	1932
13	14	0.71	Ideal	D	SI2	61.6	55.0	5.74	5.76	3.54	2767
14	15	1.50	Fair	G	VS2	66.2	53.0	7.12	7.08	4.70	10644
15	16	0.31	Ideal	G	VS2	61.6	55.0	4.37	4.39	2.70	544
16	17	0.34	Ideal	G	SI1	61.2	57.0	4.56	4.53	2.78	650
17	18	1.01	Ideal	D	VS2	59.8	56.0	6.52	6.49	3.89	7127

Table 1.

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
26947	26948	1.18	Premium	I	SI1	61.8	58.0	6.78	6.74	4.18	5617
26948	26949	0.35	Ideal	D	SI1	61.6	56.0	4.53	4.49	2.78	827
26949	26950	1.03	Ideal	G	VVS1	62.0	56.0	6.54	6.50	4.04	8398
26950	26951	1.34	Ideal	H	VS2	61.9	55.0	7.05	7.08	4.37	8771
26951	26952	1.14	Ideal	E	VVS2	61.6	57.0	6.68	6.73	4.13	11206
26952	26953	0.31	Premium	E	SI1	58.3	60.0	4.51	4.48	2.62	698
26953	26954	1.02	Premium	G	SI1	62.3	59.0	6.44	6.40	4.00	4718
26954	26955	0.50	Good	G	VVS2	63.8	56.0	5.06	5.03	3.22	1806
26955	26956	0.92	Good	E	SI1	63.3	57.0	6.17	6.22	3.92	3649
26956	26957	0.31	Ideal	E	VS2	62.3	57.0	4.32	4.35	2.70	680
26957	26958	2.09	Premium	H	SI2	60.6	59.0	8.27	8.22	5.00	17805
26958	26959	1.37	Premium	E	SI2	61.0	57.0	7.25	7.19	4.40	6751
26959	26960	1.05	Very Good	E	SI2	63.2	59.0	6.43	6.36	4.04	4281
26960	26961	1.10	Very Good	D	SI2	NaN	63.0	6.76	6.69	3.94	4361
26961	26962	0.25	Premium	F	VVS2	62.0	59.0	4.04	3.99	2.49	740
26962	26963	1.11	Premium	G	SI1	62.3	58.0	6.61	6.52	4.09	5408
26963	26964	0.33	Ideal	H	IF	61.9	55.0	4.44	4.42	2.74	1114
26964	26965	0.51	Premium	E	VS2	61.7	58.0	5.12	5.15	3.17	1656

Dataset has 11 variables with Price as a Target variable and containing other attributes of almost 27,000 cubic zirconia

Q1. 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.

Exploratory Data Analysis:

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

Table 2.

Let us check the types of variables in the data frame.

- Dataset has 6 float datatype.
- It doesn't have any null value.
- The number of rows are 26967 and the columns are 11.
- The total number of elements of the dataset are 296637.
- There are no Duplicates in the dataset.

Description:

	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	13484.000000	7784.846691	1.0	6742.50	13484.00	20225.50	26967.00
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	5.729854	1.128516	0.0	4.71	5.69	6.55	10.23
y	26967.0	5.733569	1.166058	0.0	4.71	5.71	6.54	58.90
z	26967.0	3.538057	0.720624	0.0	2.90	3.52	4.04	31.80
price	26967.0	3939.518115	4024.864666	326.0	945.00	2375.00	5360.00	18818.00

Table 3.

- We can see that the mean and the median values are almost the same. The table has high standard of deviation compare to others.

Checking for Outliers

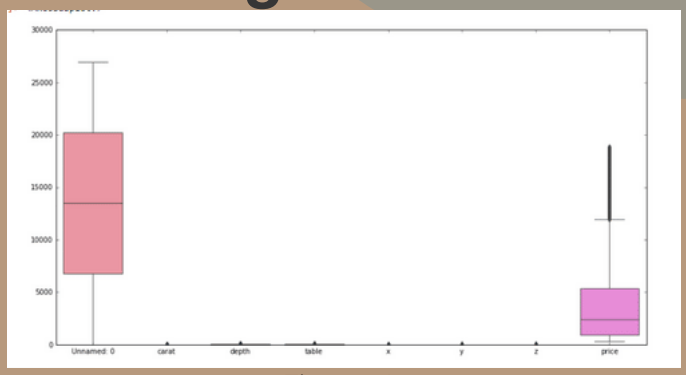


Figure 1.

We can see that The variables carat, depth, table, x, y, z and price have outliers. Treating outliers sometimes results in the models having better performance but the models lose out on generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results. On the other hand, it is perfectly fine if building models only once, i.e. either after treating or not treating the outliers.

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

Univariate Analysis:

- Univariate / Bivariate analysis helps us understand the distribution of the dataset. With the visualisation we can understand the data and solve the problem. We are going to see distribution for all the variables individually.

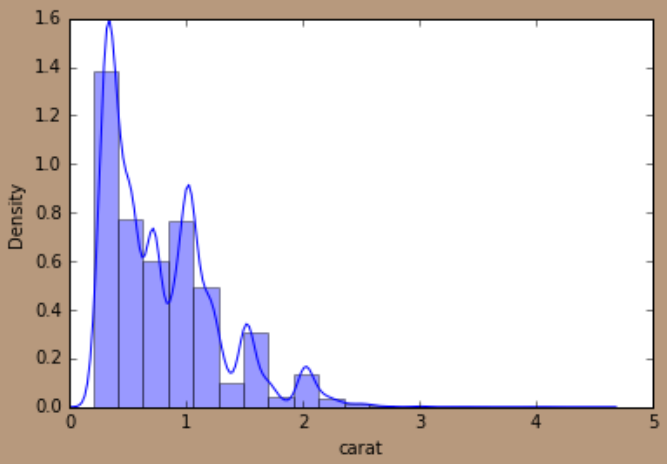


Figure 2.

Carat :

- We can see that Carat is positively skewed.
- **Positively Skewed Distribution is a type of distribution where the mean, median and mode of the distribution are positive rather than negative or zero i.e., data distribution occurs more on the one side of the scale with long tail on the right side.**

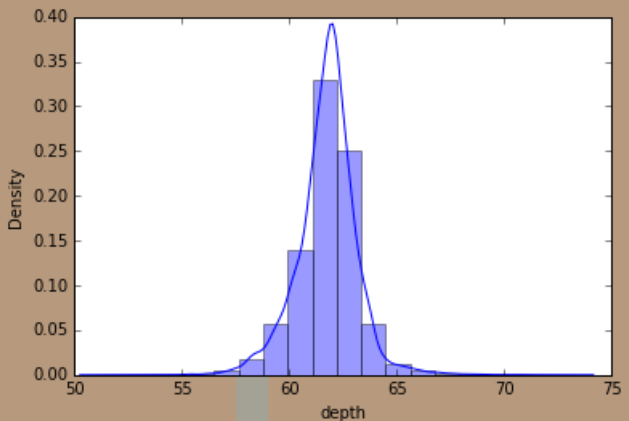


Figure 3.

Depth:

- We can see that depth is positively skewed.

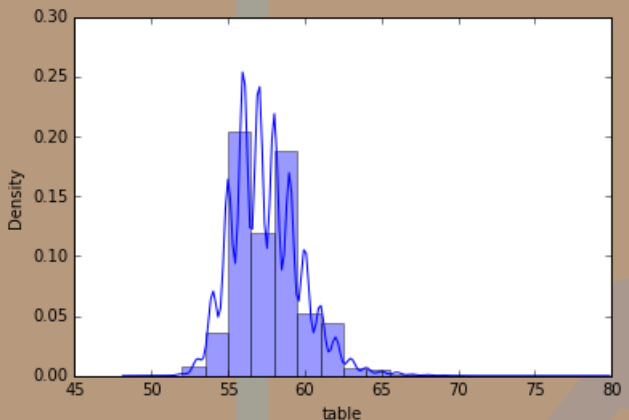


Figure 4.

Depth:

- We can see that table is positively skewed.

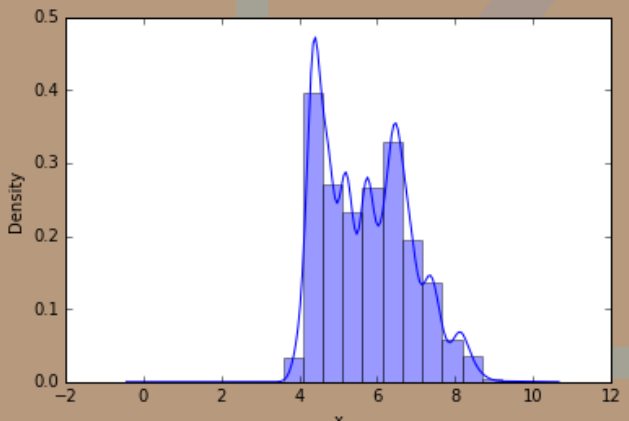


Figure 5.

X:

- We can see that table is x is positively skewed.

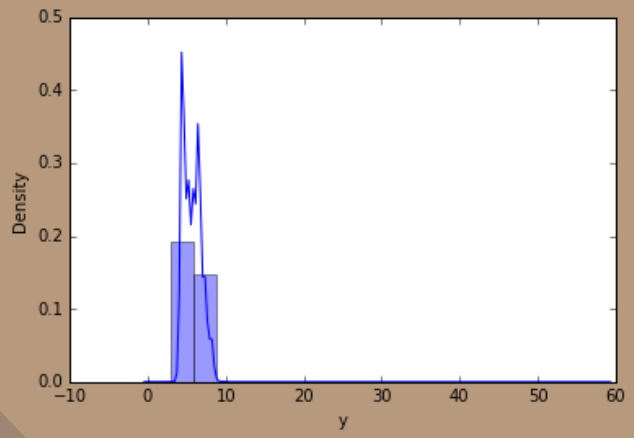


Figure 6.

Y:

- We can see that table is y is positively skewed.

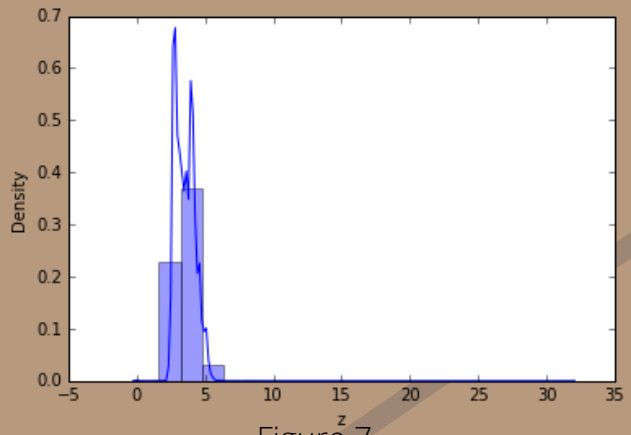


Figure 7.

Z:

- We can see that table is z is positively skewed.

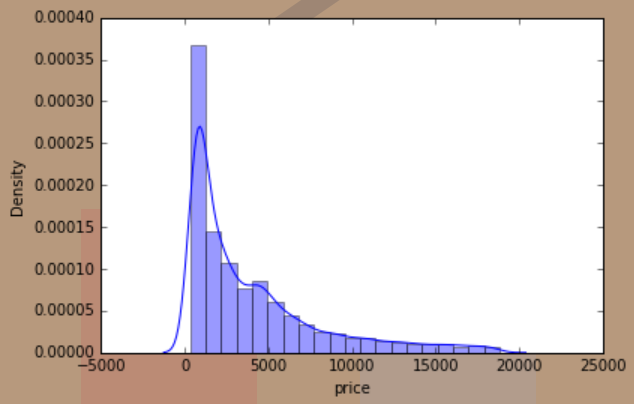


Figure 8.

price :

- We can see that price is positively skewed.

```
]: Unnamed: 0     0.000000
   carat          1.116481
   depth         -0.028618
   table          0.765758
   x              0.387986
   y              3.850189
   z              1.568157
   price          1.618550
   dtype: float64
```

Figure 9.

Skewness of data :

- We can see that y is highly rightly skewed.
- Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.
- Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.

Univariate categorical Analysis:

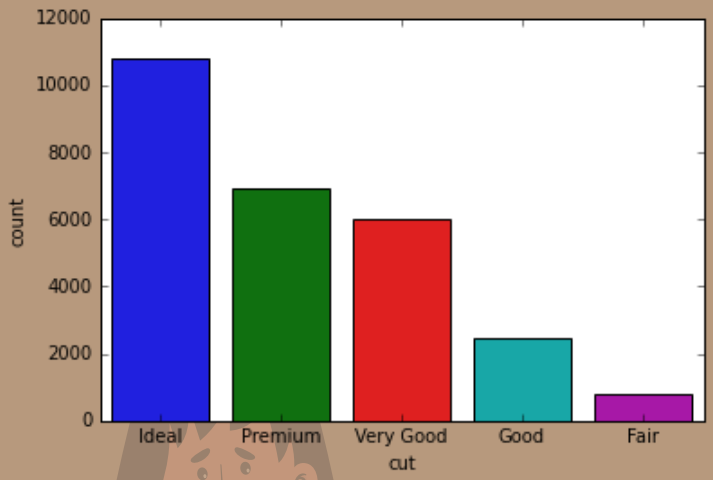


Figure 10.

cut:

- We can see that the Ideal cut is more in number 10816 to be specific and Fair is less in number ie 781.

Ideal	**10816**
Premium	**6899**
Very Good	**6030**
Good	**2441**
Fair	**781**

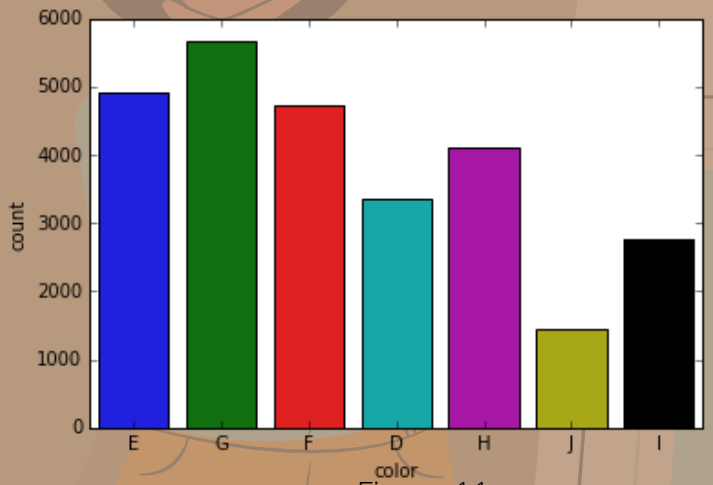


Figure 11.

color:

- We can see that the G colour is more in number ie 5661 and J which is the best one is less in number ie 1443.

G	**= 5661**
E	**= 4917**
F	**= 4729**
H	**= 4102**
D	**= 3344**
I	**= 2771**
J	**1443**

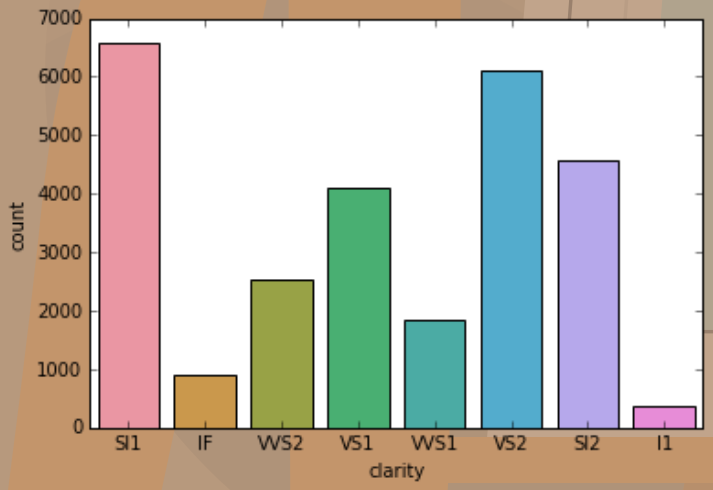


Figure 12.

clarity:

- cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
- IF is the worst which is 894 and I1 is the best which is 365.

SI1	**6571**
VS2	**6099**
SI2	**4575**
VS1	**4093**
VVS2	**2531**
VVS1	**1839**
IF	**894**
I1	**365**

Bivariate categorical Analysis:

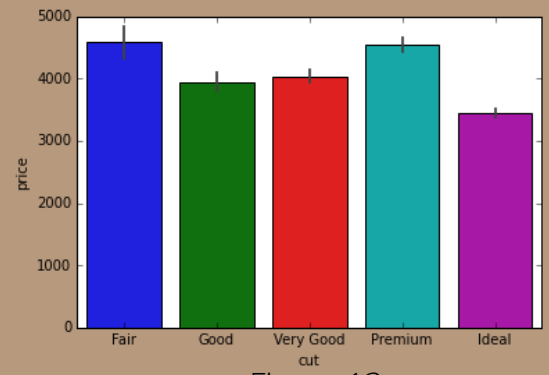


Figure 13.

price vs cut :

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

Ideal	10816
Premium	6899
Very Good	6030
Good	2441
Fair	781

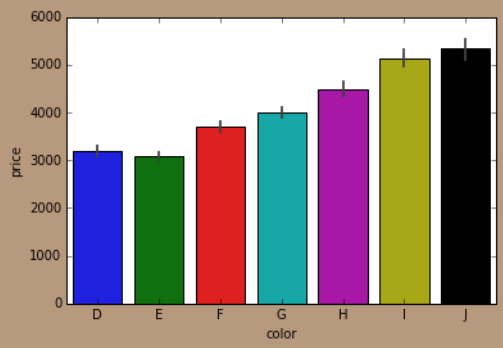


Figure 14.

price vs color :

- The price of color with J is more ie 5331.45 and the count of J is 1443..

color	price
J	5331.4
I	5129.8
H	4486.7
G	4002.6
F	3699.8

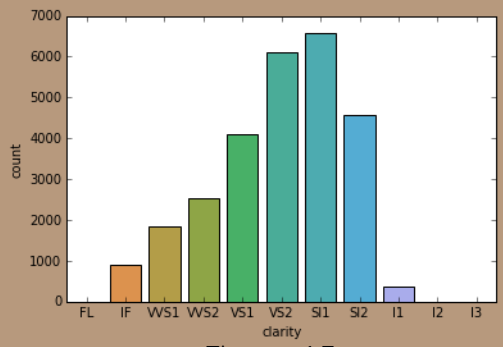


Figure 15.

price vs clarity :

- The price of clarity with SI2 is more ie 5095 and the count of SI2 is 6571.

Clarity	Price
SI2	5095.0
SI1	3998.1
VS2	3968.7
I1	3906.5
VS1	3838.8

Price vs Depth :

- We can say that zirconia with 70.6 depth has more priced compare to others.

Depth	Price
70.6	11486.0
67.7	10255.3
69.1	8736.0
69.6	7560.0
70.5	6860.0

Price vs X :

- We can say that zirconia with 9.66 Length has more priced compare to others.

x	price
9.66	18701.0
9.51	18559.0
10.23	18531.0
8.77	18485.0
9.36	18242.0

Price vs Y :

- We can say that zirconia with 9.63 Width has more priced compare to others.

Y	Price
9.63	18701.0
9.46	18559.0
10.16	18531.0
9.01	18242.0
8.94	18136.0

Price vs Z :

- We can say that zirconia with 6.72 Height has more priced compare to others.

Z	Price
6.72	18531.0
5.62	18400.5
5.90	18242.0
5.77	18242.0
5.40	17549.5

Multivariate Analysis:

Pair Plot - A pairs plot allows us to see both distributions of single variables and relationships between two variables.

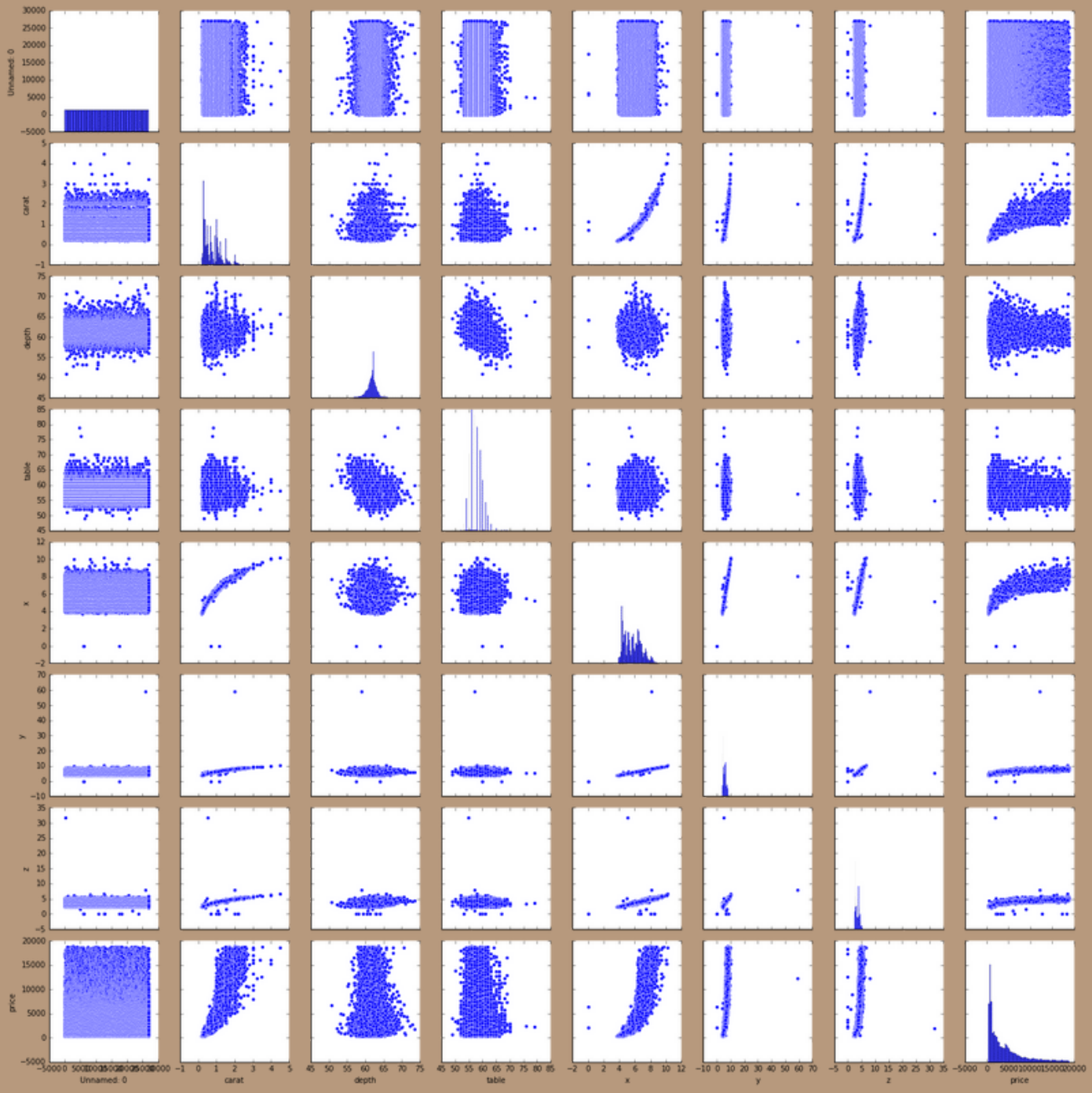


Figure 16.

	Unnamed: 0	carat	depth	table	x	y	z	price
Unnamed: 0	1.000000	0.003490	-0.001588	0.003817	0.004626	0.006844	0.001681	0.002650
carat	0.003490	1.000000	0.035364	0.181685	0.976368	0.941071	0.940640	0.922416
depth	-0.001588	0.035364	1.000000	-0.298011	-0.018715	-0.024735	0.101624	-0.002569
table	0.003817	0.181685	-0.298011	1.000000	0.196206	0.182346	0.148944	0.126942
x	0.004626	0.976368	-0.018715	0.196206	1.000000	0.962715	0.956606	0.886247
y	0.006844	0.941071	-0.024735	0.182346	0.962715	1.000000	0.928923	0.856243
z	0.001681	0.940640	0.101624	0.148944	0.956606	0.928923	1.000000	0.850536
price	0.002650	0.922416	-0.002569	0.126942	0.886247	0.856243	0.850536	1.000000

There is a strong correlation between x and carat, y and carat, z and carat, z and depth.

Heat Map:

- The heat map is one of the most useful and powerful data-analysis tools available in business intelligence. It is a visualization feature that presents multiple rows of data in a way that makes immediate sense by assigning different size and color to cells each representing a row.

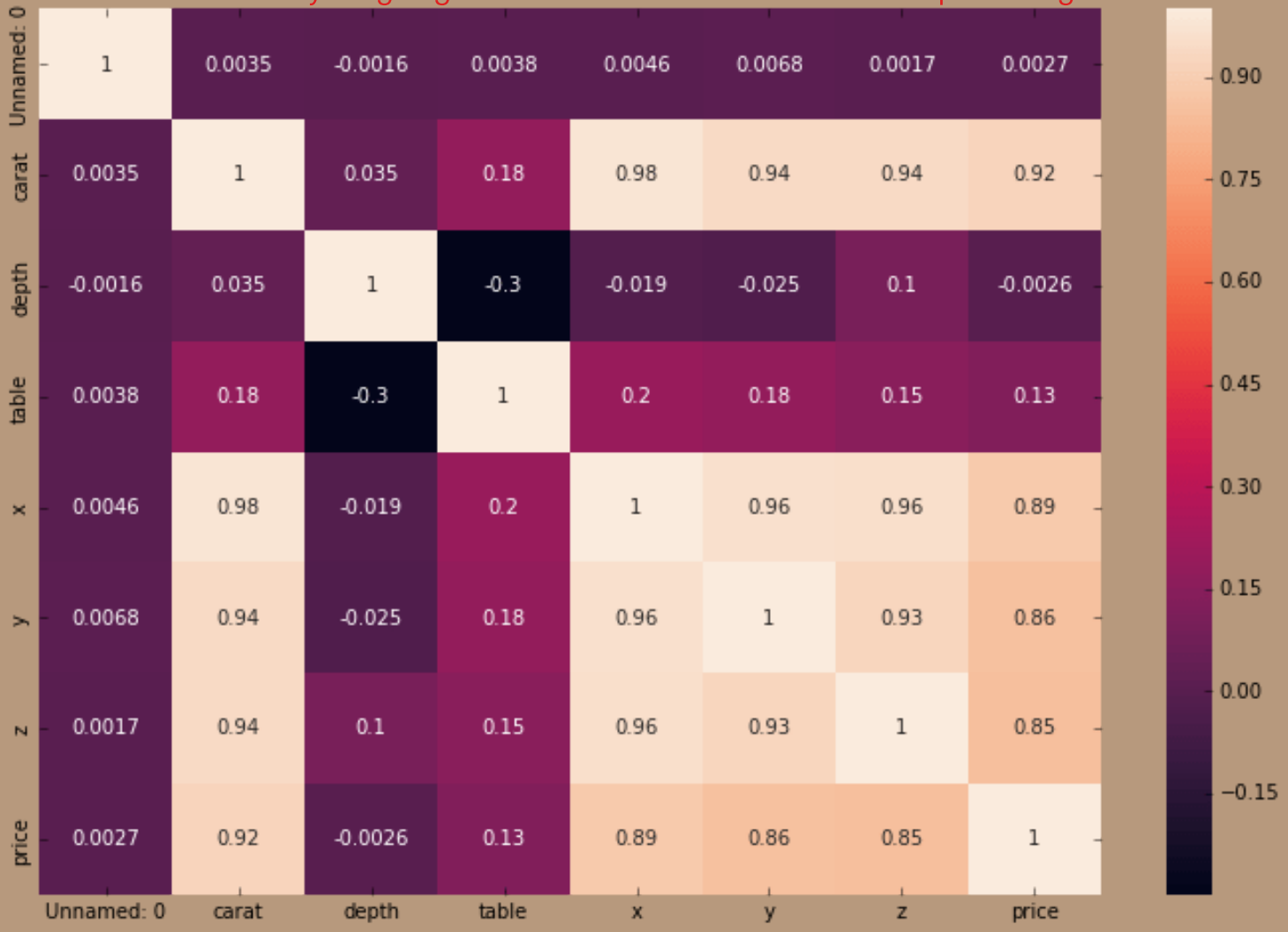


Figure 17.

From the above figure we can observe that There is a strong correlation between

1) x and carat.
2) y and carat.
3) z and carat.
4) price and carat.
5) y and x
6) z and x. This says as x,y,z increases carat also increases and as the carat increases price also increases.

This shows that there is multicollinearity in the database.

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

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

Table 2.

Data can have missing values for a number of reasons such as observations that were not recorded and data corruption.
Handling missing data is important as many machine learning algorithms do not support data with missing values.
Missing values are usually represented in the form of Nan or null or None in the dataset.

There are 697 missing values in depth variable.

<u>Imputing missing values</u>

- Deleting the columns with missing data
- Deleting the rows with missing data
- Filling the missing data with a value – Imputation
- Imputation with an additional column
- Filling with a Regression Model

Missing values can be imputed with a provided constant value, or using the statistics (mean, median or most frequent) of each column in which the missing values are located.

1. Filling the missing data with the mean or median value if it's a numerical variable.
2. Filling the missing data with mode if it's a categorical value.

After Imputing missing values
Now there are no missing values

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

Table 3.

The values which are equal to zero.

Unnamed: 0	0
carat	0
cut	0
color	0
clarity	0
depth	0
table	0
x	3
y	3
z	9
price	0
dtype: int64	

Table 4.

After imputing the values of zero :-
We are treating them by using mean value.

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

Table 5.

X Length of the cubic zirconia in mm. Y Width of the cubic zirconia in mm. Z Height of the cubic zirconia in mm. If the length, width and height of the cubic zirconia are all 0 then carat should also be zero. So it is better if we change it to the mean or median.

Checking the outliers

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population.

We can see that min_payment_amt has few outliers. Outliers Treatment is necessary as clustering is affected by them.

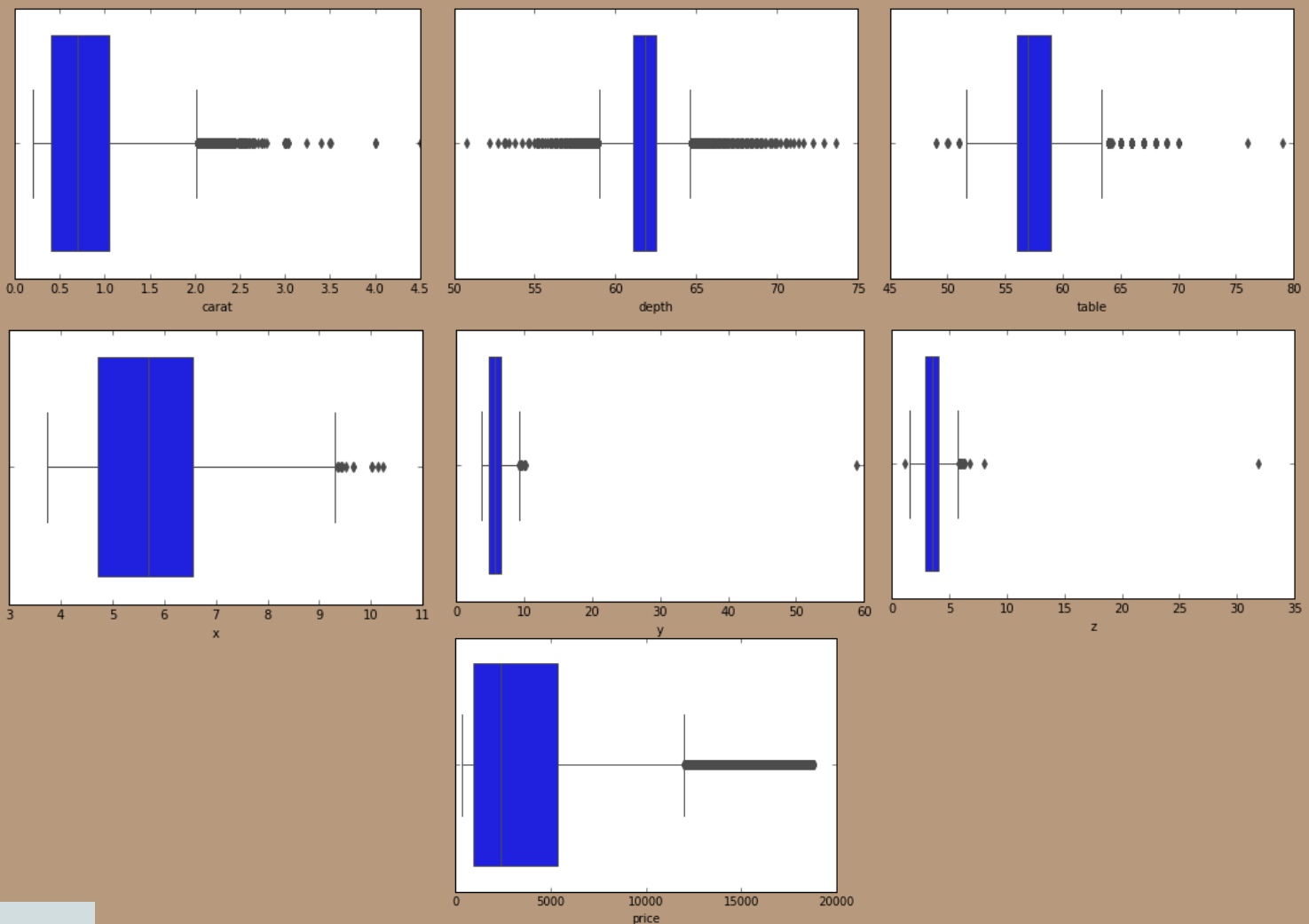


Figure. 18

After Treating the outliers

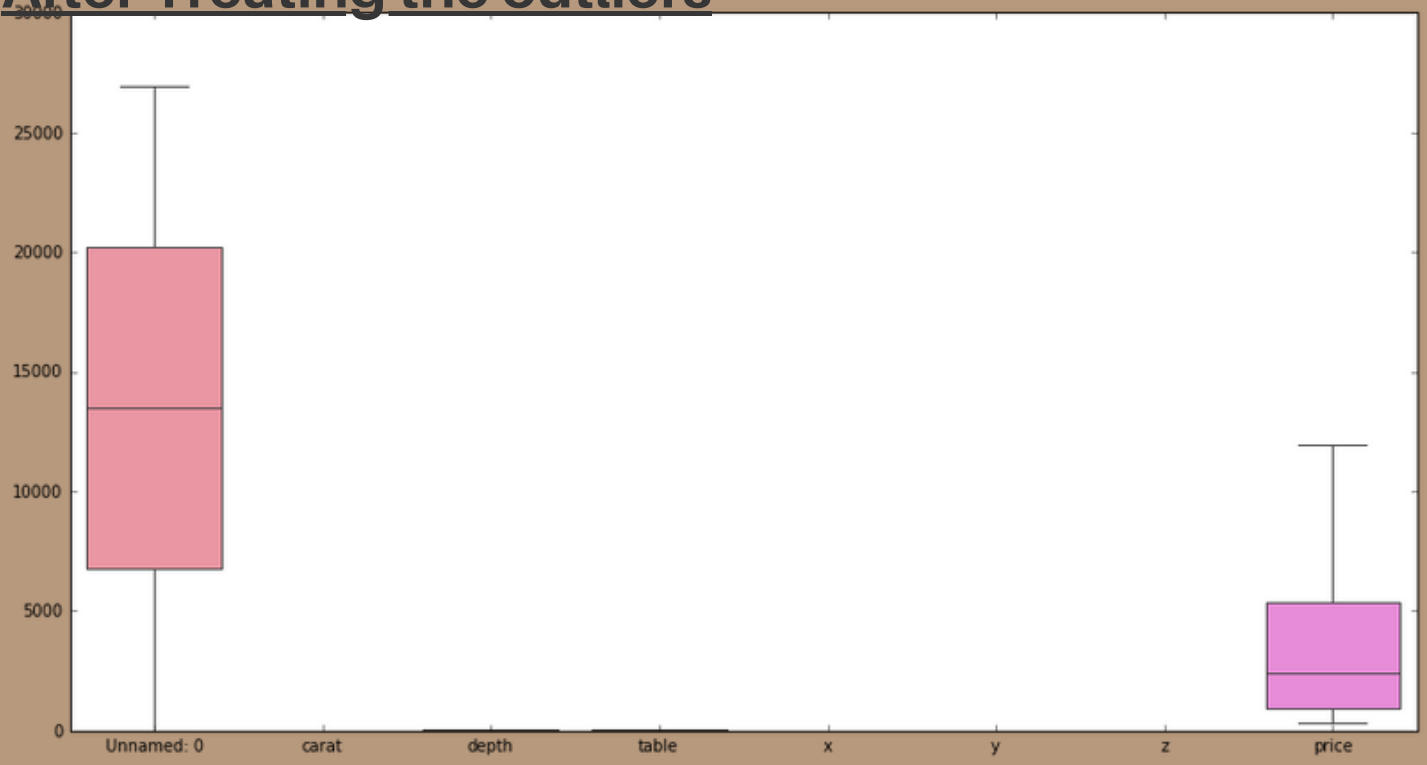


Figure. 19

Combining the sub levels of a ordinal variables

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


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


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

Table 6

An ordinal variable is the clear ordering of the categories. Ordinal variables can be considered "in between" categorical and quantitative variables.

With ordinal scales, the order of the values is what's important and significant, but the differences between each one are not really known. In this case, we know that a Ideal is better than Premium or Very good, but we don't know-and cannot quantify-how much better it is. For example, is the difference between "Ideal" and "Premium" the same as the difference between "Very Good" and "Good" We can't say.

Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

Q 1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from 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.

	Unnamed: 0	carat	depth	table	x	y	z	price	cut_Good	cut_Ideal	...	color_H	color_I	color_J	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1	clarity_VS2	clarity_VVS1	clarity_VVS2
0	1.0	0.30	62.1	58.0	4.27	4.29	2.66	499.0	0	1	...	0	0	0	0	1	0	0	0	0	0
1	2.0	0.33	60.8	58.0	4.42	4.46	2.70	984.0	0	0	...	0	0	0	1	0	0	0	0	0	0
2	3.0	0.90	62.2	60.0	6.04	6.12	3.78	6289.0	0	0	...	0	0	0	0	0	0	0	0	0	1
3	4.0	0.42	61.6	56.0	4.82	4.80	2.96	1082.0	0	1	...	0	0	0	0	0	0	1	0	0	0
4	5.0	0.31	60.4	59.0	4.35	4.43	2.65	779.0	0	1	...	0	0	0	0	0	0	0	0	1	0

5 rows × 25 columns

Table 7

Converting categorical to dummy variables

Encoding the data (having string values) for Modelling

We do Ordinal encoding to ensure the encoding of variables retains the ordinal nature of the variable. This is reasonable only for ordinal variables. This encoding looks almost similar to Label Encoding but slightly different as Label coding would not consider whether the variable is ordinal or not, and it will assign a sequence of integers

Linear regression model does not take categorical values so that we have encoded categorical values to integer for better results.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 25 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Unnamed: 0      26967 non-null  float64
 1   carat           26967 non-null  float64
 2   depth           26967 non-null  float64
 3   table           26967 non-null  float64
 4   x               26967 non-null  float64
 5   y               26967 non-null  float64
 6   z               26967 non-null  float64
 7   price           26967 non-null  float64
 8   cut_Good        26967 non-null  uint8
 9   cut_Ideal       26967 non-null  uint8
 10  cut_Premium     26967 non-null  uint8
 11  cut_Very Good   26967 non-null  uint8
 12  color_E         26967 non-null  uint8
 13  color_F         26967 non-null  uint8
 14  color_G         26967 non-null  uint8
 15  color_H         26967 non-null  uint8
 16  color_I         26967 non-null  uint8
 17  color_J         26967 non-null  uint8
 18  clarity_IF      26967 non-null  uint8
 19  clarity_SI1     26967 non-null  uint8
 20  clarity_SI2     26967 non-null  uint8
 21  clarity_VS1     26967 non-null  uint8
 22  clarity_VS2     26967 non-null  uint8
 23  clarity_VVS1    26967 non-null  uint8
 24  clarity_VVS2    26967 non-null  uint8
dtypes: float64(8), uint8(17)
memory usage: 2.1 MB
```

Table 8

Spliting the data into train and test (70:30) and applying Linear regression using scikit learn

	carat	depth	table	x	y	z	cut_Good	cut_Ideal	cut_Premium	cut_Very Good	...	color_H	color_I	color_J	clarity_IF	clarity_SI1	clarity_SI2	cl	clarity_VS1	clarity_VS2	clarity_VVS1	clarity_VVS2
0	0.30	62.1	58.0	4.27	4.29	2.66	0	1	0	0	...	0	0	0	0	1	0		0	0	0	0
1	0.33	60.8	58.0	4.42	4.46	2.70	0	0	1	0	...	0	0	0	1	0	0		0	0	0	0
2	0.90	62.2	60.0	6.04	6.12	3.78	0	0	0	1	...	0	0	0	0	0	0		0	0	0	1
3	0.42	61.6	56.0	4.82	4.80	2.96	0	1	0	0	...	0	0	0	0	0	0		1	0	0	0
4	0.31	60.4	59.0	4.35	4.43	2.65	0	1	0	0	...	0	0	0	0	0	0		0	0	1	0

5 rows × 23 columns

Table 9

Linear Regression Model

```
The coefficient for carat is 9218.950642771502
The coefficient for depth is -0.4042282058823048
The coefficient for table is -22.772438642718214
The coefficient for x is -1263.3775098104065
The coefficient for y is 1000.954703063349
The coefficient for z is -382.11834284322714
The coefficient for cut_Good is 363.53671738392933
The coefficient for cut_Ideal is 615.7097791697761
The coefficient for cut_Premium is 594.1126202366884
The coefficient for cut_Very Good is 491.1386827231898
The coefficient for color_E is -194.40003176371377
The coefficient for color_F is -269.6801171776984
The coefficient for color_G is -420.82026154507395
The coefficient for color_H is -844.187382596224
The coefficient for color_I is -1326.0306184492079
The coefficient for color_J is -1916.270585925992
The coefficient for clarity_IF is 4048.3025188780775
The coefficient for clarity_SI1 is 2604.879495464047
The coefficient for clarity_SI2 is 1791.3482767051855
The coefficient for clarity_VS1 is 3416.9830158510204
The coefficient for clarity_VS2 is 3144.437215740155
The coefficient for clarity_VVS1 is 3850.630190078691
The coefficient for clarity_VVS2 is 3817.8624984488433
```

Table 10

In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one.

Linear Regression means a kind of statistical analysis that attempts to show a relationship between two variables. Linear regression looks at various data points and plots a trend line. Linear regression can create a predictive model on apparently random data, showing trends in data, such as in cancer diagnoses or in stock prices.
Linear regression is an important tool in analytics. The technique uses statistical calculations to plot a trend line in a set of data points. The trend line could be anything from the number of people diagnosed with skin cancer to the financial performance of a company. Linear regression shows a relationship between an independent variable and a dependent variable being studied.
There are a number of ways to calculate linear regression. One of the most common is the ordinary least-squares method, which estimates unknown variables in the data, which visually turns into the sum of the vertical distances between the data points and the trend line.

The intercept for our model is -2313.055.

The intercept (often labeled the constant) is the expected mean value of Y when all X=0.

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

Predictions on Train and Test sets using Rsquare

R square on training data

0.9409

94% of the variation in the price is explained by the predictors in the model for train set

R square on testing data

0.94045

Predictions on Train and Test sets using RMSE

RMSE on Training data

845.53

RMSE on Testing data

842.65

Predictions on Train and Test sets using Adj Rsquare
Linear Regression using statsmodels

RMSE on Training data

845.53

RMSE on Testing data

842.65

```
data_train.columns
```

```
Index(['carat', 'depth', 'table', 'x', 'y', 'z', 'cut_Good', 'cut_Ideal',
       'cut_Premium', 'cut_Very_Good', 'color_E', 'color_F', 'color_G',
       'color_H', 'color_I', 'color_J', 'clarity_IF', 'clarity_SI1',
       'clarity_SI2', 'clarity_VS1', 'clarity_VS2', 'clarity_VVS1',
       'clarity_VVS2', 'price'],
      dtype='object')
```

```
data_test.columns
```

```
Index(['carat', 'depth', 'table', 'x', 'y', 'z', 'cut_Good', 'cut_Ideal',
       'cut_Premium', 'cut_Very_Good', 'color_E', 'color_F', 'color_G',
       'color_H', 'color_I', 'color_J', 'clarity_IF', 'clarity_SI1',
       'clarity_SI2', 'clarity_VS1', 'clarity_VS2', 'clarity_VVS1',
       'clarity_VVS2', 'price'],
      dtype='object')
```

Figure. 20

Predictions on Train and Test sets using Adj Rsquare

VIF values

```
carat ---> 122.78486179880461
depth ---> 1348.4342853937171
table ---> 978.9963533401854
x ---> 11968.853608180094
y ---> 11499.091021050864
z ---> 3177.9722370725503
cut_Good ---> 4.495398493285433
cut_Ideal ---> 18.01592551826808
cut_Premium ---> 10.838931616722222
cut_Very_Good ---> 10.010327160868213
color_E ---> 2.479864772837505
```

Figure. 21

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. ... This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.936
Method:	Least Squares	F-statistic:	1.458e+04
Date:	Wed, 22 Sep 2021	Prob (F-statistic):	0.00
Time:	15:37:47	Log-Likelihood:	-1.5472e+05
No. Observations:	18876	AIC:	3.095e+05
Df Residuals:	18856	BIC:	3.096e+05
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2272.1265	744.994	-3.050	0.002	-3732.381	-811.871
carat	9062.1425	79.550	113.917	0.000	8906.216	9218.069
depth	-9.4511	10.088	-0.937	0.349	-29.224	10.322
table	-19.0490	4.001	-4.886	0.000	-27.792	-12.106
x	-1313.0812	138.949	-9.450	0.000	-1585.435	-1040.728
y	1053.1879	143.046	7.363	0.000	772.805	1333.571
z	-356.7701	120.272	-2.950	0.003	-590.554	-119.026
cut_Good	375.7723	45.213	8.311	0.000	287.151	464.393
cut_Ideal	625.2058	43.878	14.249	0.000	539.201	711.210
cut_Premium	583.8845	42.186	13.867	0.000	501.354	666.416
cut_Very_Good	488.8049	43.121	11.548	0.000	404.085	573.525
color_E	-438.7383	21.477	-43.709	0.000	-480.835	-396.642
color_J	-1519.5114	29.469	-51.564	0.000	-1577.272	-1461.750
clarity_IF	3965.4229	68.148	58.195	0.000	3831.863	4098.983
clarity_SI1	2625.9952	58.700	44.736	0.000	2510.939	2741.052
clarity_SI2	1827.8538	59.040	30.960	0.000	1712.151	1943.557
clarity_VS1	3389.9502	59.849	56.641	0.000	3272.620	3507.280
clarity_VS2	3161.6488	59.012	53.576	0.000	3045.980	3277.318
clarity_VVS1	3795.2295	63.425	59.838	0.000	3670.910	3919.549
clarity_VVS2	3811.4185	61.748	61.733	0.000	3690.402	3932.435

Omnibus:	5157.897	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21899.897
Skew:	1.189	Prob(JB):	0.00
Kurtosis:	7.005	Cond. No.	9.06e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.06e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Table 11

In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable.

The overall P value is less than alpha, so rejecting H0 and accepting Ha that atleast 1 regression co-efficient is not 0. Here all regression co-efficients are not 0

Root Mean Squared Error - RMSE 878.20

Prediction on Test data

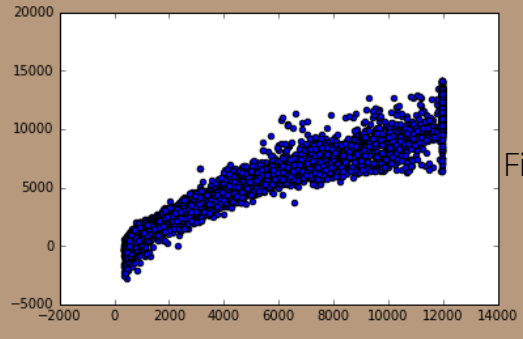


Figure. 22

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

From the EDA analysis we could understand the cut, ideal cut had number profits to the company. The colours H, I, J have bought profits for the company. In clarity if we could see there were no flawless stones and they were no profits coming from I1, I2, I3 stones. The ideal, premium and very good types of cut were bringing profits where as fair and good are not bringing profits. The predictions were able to capture 94% variations in the price and it is explained by the predictors in the training set. Using stats model if we could run the model again we can have P values and coefficients which will give us better understanding of the relationship, so that values more 0.05 we can drop those variables and re run the model again for better results. For better accuracy dropping depth column in iteration for better results. The equation,

price = (-2272.13) * Intercept + (9062.14) * carat + (-9.45) * depth + (-19.95) * table + (-1313.08) * x + (1053.19) * y + (-354.77) * z + (375.77) * cut_Good + (625.21) * cut_Ideal + (583.88) * cut_Premium + (498.8) * cut_Very_Good + (-938.74) * color_I + (-1519.51) * color_J + (3965.42) * clarity_IF + (2626.0) * clarity_SI1 + (1827.85) * clarity_SI2 + (3389.93) * clarity_VS1 + (3161.65) * clarity_VS2 + (3795.23) * clarity_VVS1 + (3811.42) * clarity_VVS2

Recommendations

1. The ideal, premium, very good cut types are the one which are bringing profits so that we could use marketing for these to bring in more profits.
2. The clarity of the diamond is the next important attributes the more the clear is the stone the profits are more

We can say that Ideal, Premium and VeryGood cut types are giving more profits. Also clarity plays an important role in profting.

The Five best attributes are :-

1) Y (Width of the cubic zirconia in mm.) of the zirconia.
2) Clarity_IF
3) clarity_VVS2
4) clarity_VVS1
5) clarity_VS1

About Logistic Regression and LDA
Case study on "Holiday Package dataset"

A tour and travel agency deals in selling holiday packages. Among these employees, some opted for the package and some didn't.

Our Objective:

The objective is to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, need to find out the important factors on the basis of which the company will focus on particular employees to sell their packages. The details of 872 employees of a company are provided.
The solution resolved for this dataset is by using Logistic Regression and LDA.
To understand the prediction let's explore the dataset.

Data Description:

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency_Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration)
7. Destination of the tour (Destination)
8. Amount of sales of tour insurance policies (Sales)
9. The commission received for tour insurance firm (Commission)
10. Age of insured (Age)

We start with loading the dataset, checking its shape and data types of variable . shape tell us how many rows and columns we have in the data and data type tell us whether the variable is object,integer or float value..

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 1.

Q2.1 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.

Exploratory Data Analysis:

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

Figure 1.

Let us check the types of variables in the data frame.

- Dataset has 2 object type, 6 integer data type.
- It doesn't have any null value.
- The number of rows are 872 and the columns are 8.
- The total number of elements of the dataset are 6976.
- There are 0 Duplicates in the dataset.
- There are 0 missing values in the dataset.

Description:

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.000000	62.000000	21.000000	3.000000	6.000000

Table 2.

Holiday package is our target variable .
The mean and the 50% of data for all variables are almost same.

no 0.540138
yes 0.459862
Name: Holliday_Package, dtype: float64

This split indicates that 45% of employees are interested in the holiday package.

Checking for Outliers:

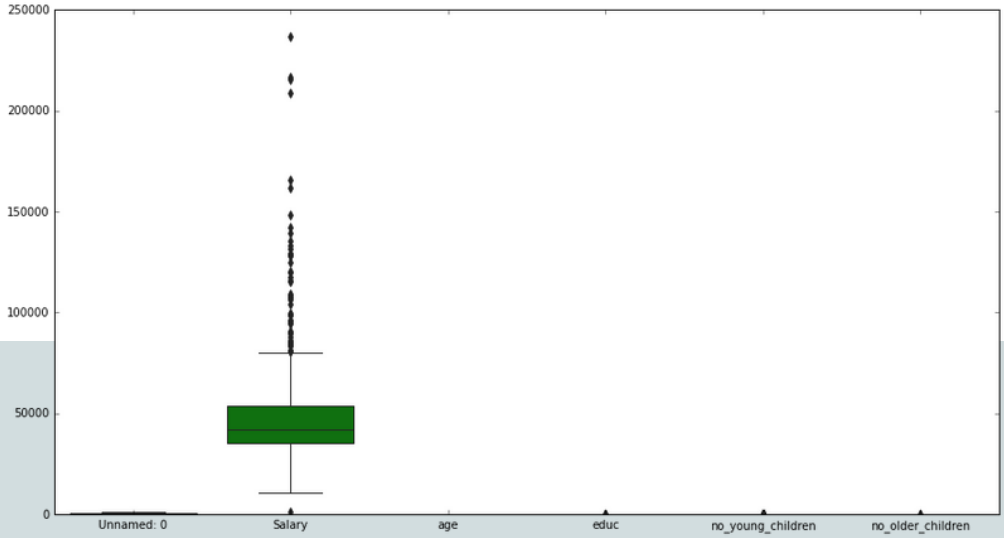


Figure 2.

All the varaibles has outliers. We are not treating the outliers as it may affect the output to more extent.

Univariate Analysis:

Univariate / Bivariate analysis helps us understand the distribution of the dataset. With the visualisation we can understand the data and solve the problem

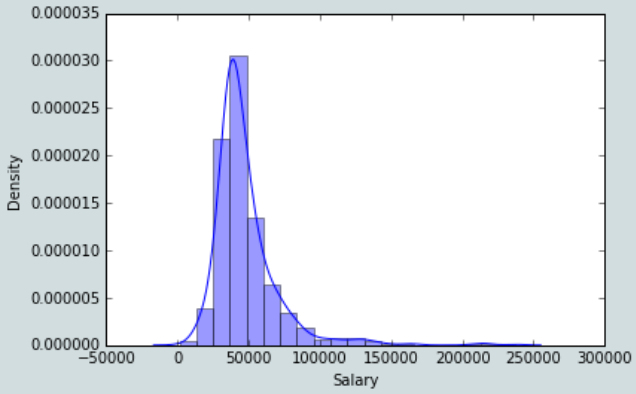


Figure 3.

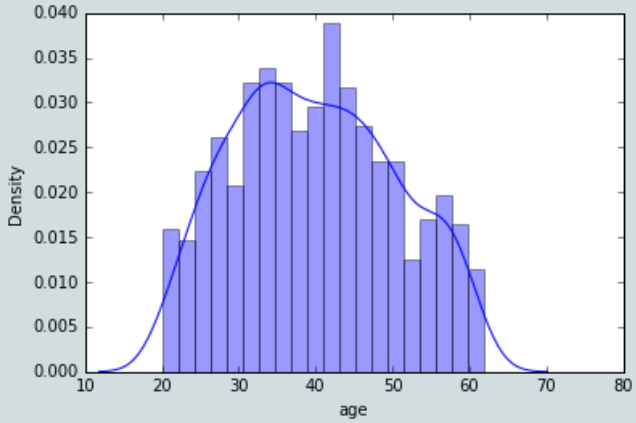


Figure 4.

Salary :

- We can see that Salary is positively skewed.
- **Positively Skewed Distribution is a type of distribution where the mean, median and mode of the distribution are positive rather than negative or zero i.e., data distribution occurs more on the one side of the scale with long tail on the right side.**

Age :

- We can see that Age is positively skewed.

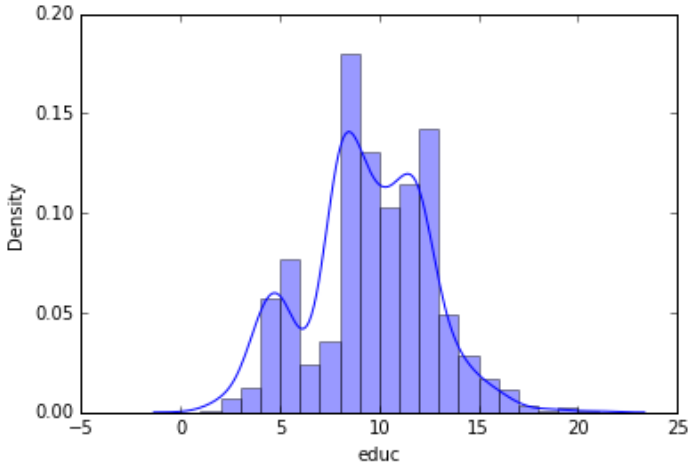


Figure 5.

Educ :

- We can see that educ is positively skewed.

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

Skewness of data :

- **We can see that Salary is highly rightly skewed.**
- **Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.**
- **Distributions can exhibit right (positive) skewness or left (negative) skewness to varying degrees. A normal distribution (bell curve) exhibits zero skewness.**

Univariate categorical Analysis:

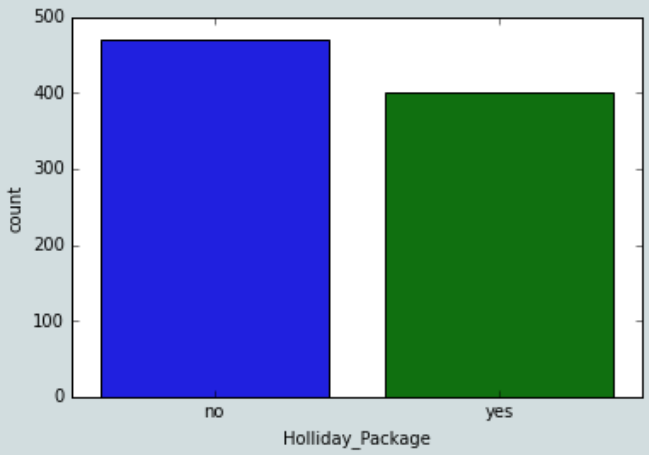


Figure 6.

Holliday_Package:
We can say that maximum employees didn't choose the holiday package.

no 471
yes 401

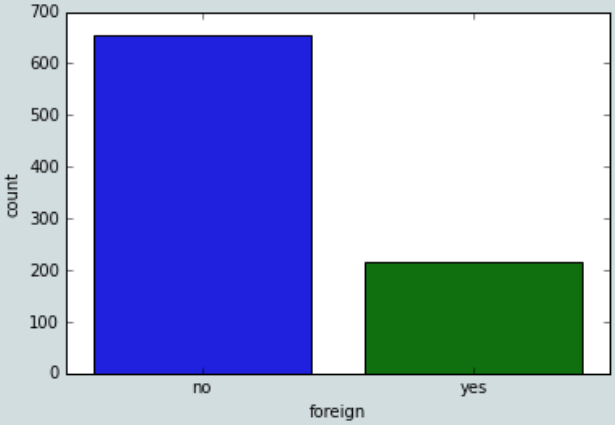


Figure 7.

foreign:
We can say that maximum foreign employees didn't choose the package.

no 656
yes 216

Bivariate categorical Analysis:

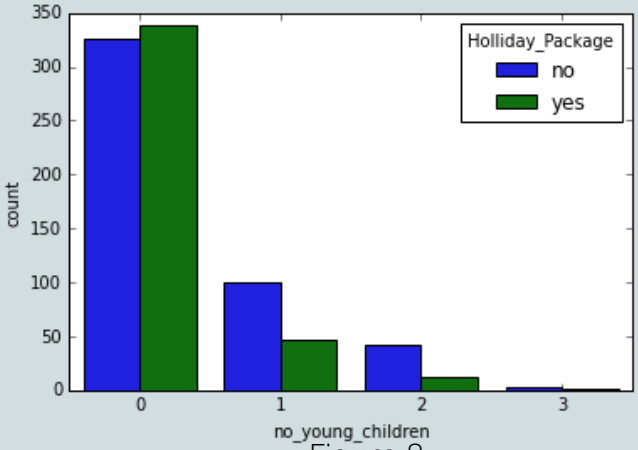


Figure 8.

no_young_children vs Holliday_Package :

- We can say that maximum employees with 0 kids choose the holiday package.

0 665
1 147
2 55
3 5

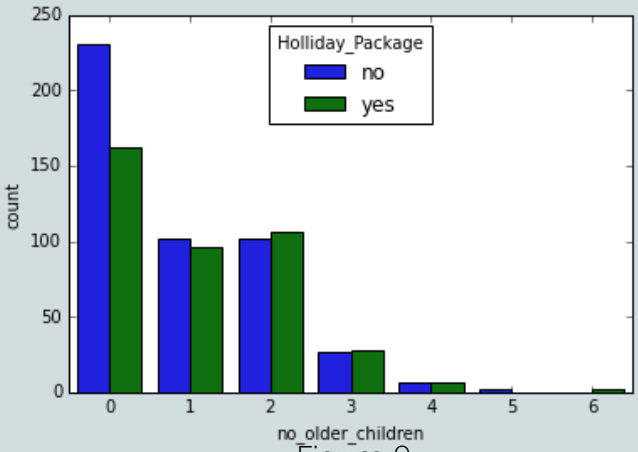


Figure 9.

no_older_children vs Holliday_Package :

- We can say that maximum employees with 2,3 and 4 older_children choose the holiday package. And employees with 0 older_children didn't choose the package

0 393
2 208
1 198
3 55
4 14
6 2
5 2

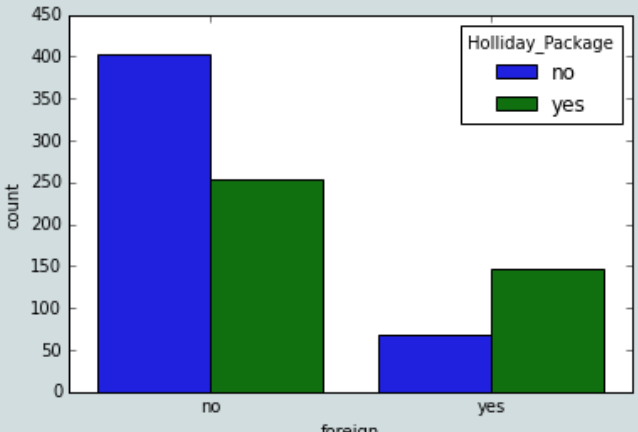


Figure 10.

foreign vs Holliday_Package :

- The maximum foreigneer employee didn't opted for Holiday package.

no 656
yes 216

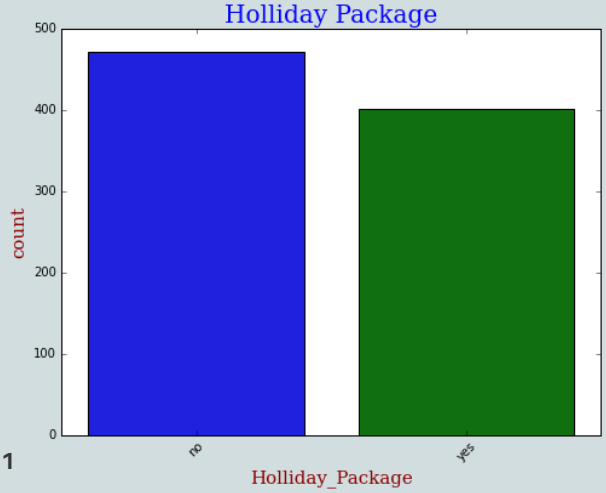


Figure 11.

Holliday_Package :

- 35 employees with age of 44 have purchased holiday package and only 3 employees with age of 62 have purchased holiday package.

Holliday_Package	age
no	40.853503
yes	38.900249

Figure 12..

Barplot of Salary vs Holliday_Package :

- Salary less than 50000 people have opted more for holiday package.

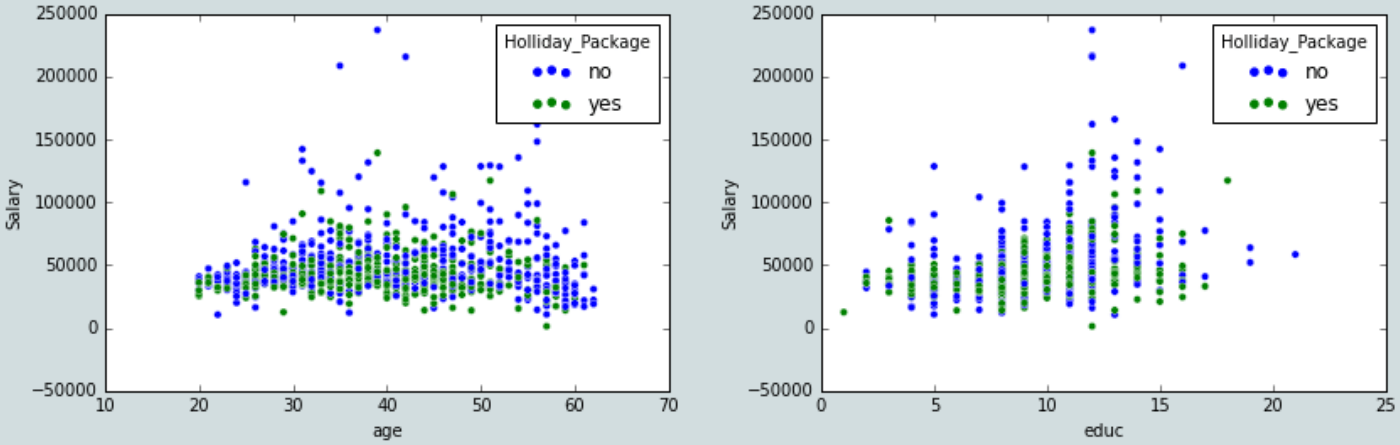


Figure 13..

Scatterplot of Salary vs age vs Holliday_Package :

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

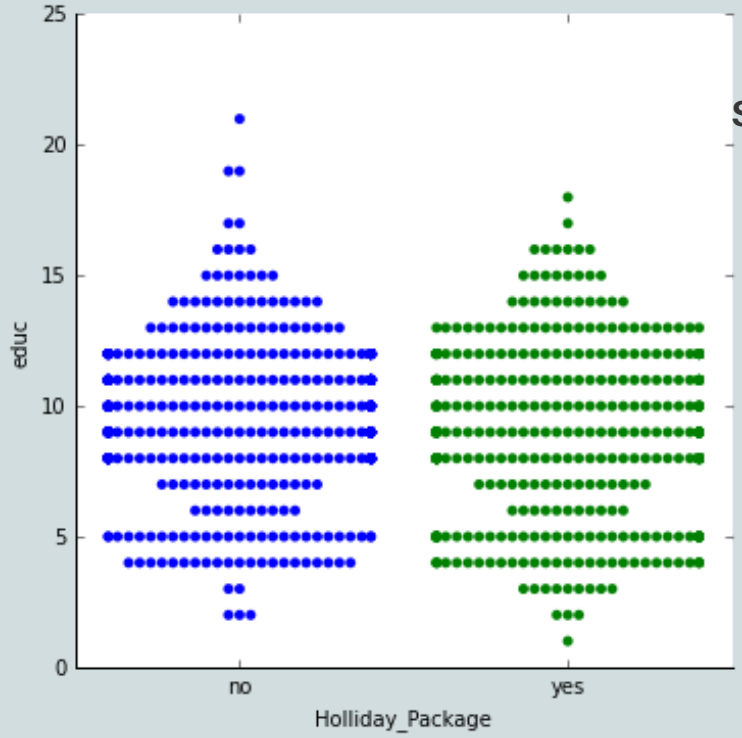


Figure 14..

Swarmplot of educ vs Holliday_Package :

- **Employee lower with 15 Years of formal education have opted more for holiday package.**

Multivariate Analysis:

Pair Plot - A pairs plot allows us to see both distributions of single variables and relationships between two variables.

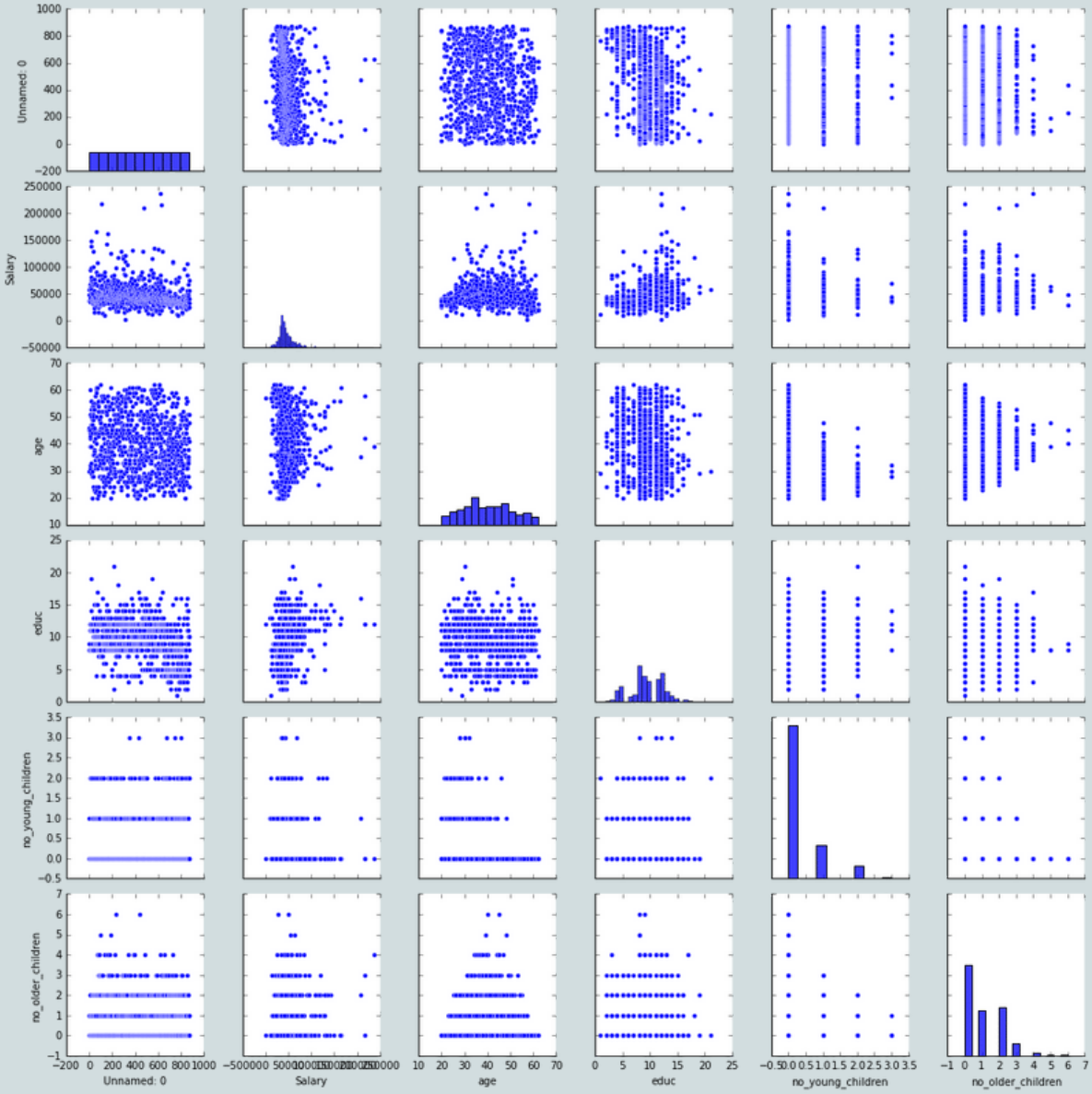


Figure 15..

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
Unnamed: 0	1.000000	-0.193249	-0.103782	-0.296015	0.052146	-0.025852
Salary	-0.193249	1.000000	0.071709	0.326540	-0.029664	0.113772
age	-0.103782	0.071709	1.000000	-0.149294	-0.519093	-0.116205
educ	-0.296015	0.326540	-0.149294	1.000000	0.098350	-0.036321
no_young_children	0.052146	-0.029664	-0.519093	0.098350	1.000000	-0.238428
no_older_children	-0.025852	0.113772	-0.116205	-0.036321	-0.238428	1.000000

We can say that there is no relation in the dataset.

Heat Map:

- The heat map is one of the most useful and powerful data-analysis tools available in business intelligence. It is a visualization feature that presents multiple rows of data in a way that makes immediate sense by assigning different size and color to cells each representing a row.

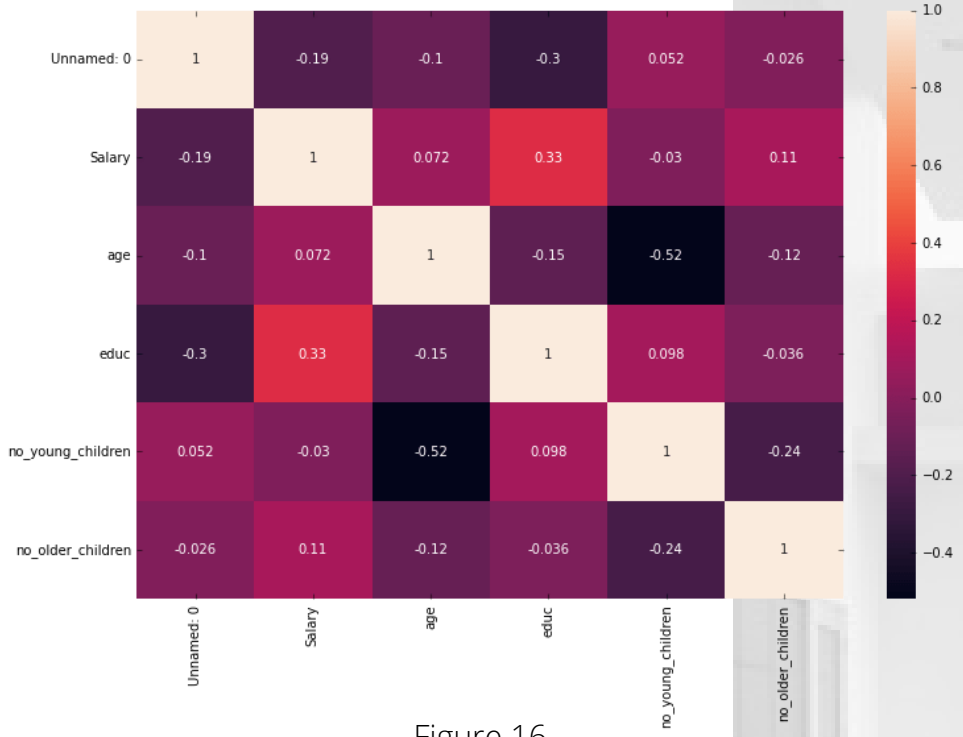


Figure 16.

From the above figure we can observe that There is a little correlation between educ and Salary

Q2.2 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).

Drop the id column as it is useless for the model. Dataset after dropping the id column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64
 2   age                872 non-null    int64
 3   educ               872 non-null    int64
 4   no_young_children  872 non-null    int64
 5   no_older_children  872 non-null    int64
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

Table 3.

Converting categorical to dummy variables in data

	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
0	48412	30	8	1	1	0	0
1	37207	45	8	0	1	1	0
2	58022	46	9	0	0	0	0
3	66503	31	11	2	0	0	0
4	66734	44	12	0	2	0	0

Table 4.

We do Ordinal encoding to ensure the encoding of variables retains the ordinal nature of the variable. This is reasonable only for ordinal variables. This encoding looks almost similar to Label Encoding but slightly different as Label coding would not consider whether the variable is ordinal or not, and it will assign a sequence of integers

The encoding helps the logistic regression model predict better results.

Spliting the data into train and test (70:30) and applying Linear regression using scikit learn

Train/ Test split

Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, you can minimize the effects of data discrepancies and better understand the characteristics of the model.

GRID SEARCH METHOD:

The grid search method is used for logistic regression to find the optimal solving and the parameters for solving

```
{'penalty': 'l1', 'solver': 'liblinear', 'tol': 0.0001}

LogisticRegression(max_iter=100000, n_jobs=2, penalty='l1', solver='liblinear')
```

Figure 17.

The grid search method gives, liblinear solver which is suitable for small datasets. Tolerance and penalty has been found using grid search method

Prediction on the training set

Getting the probabilities on the training set

]:

	0	1
0	0.675066	0.324934
1	0.574273	0.425727
2	0.686681	0.313319
3	0.525134	0.474866
4	0.544101	0.455899

Figure 18.

Confusion matrix on the training data

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset.

	precision	recall	f1-score	support
0	0.67	0.75	0.71	329
1	0.66	0.57	0.61	281
accuracy			0.67	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.67	0.66	610

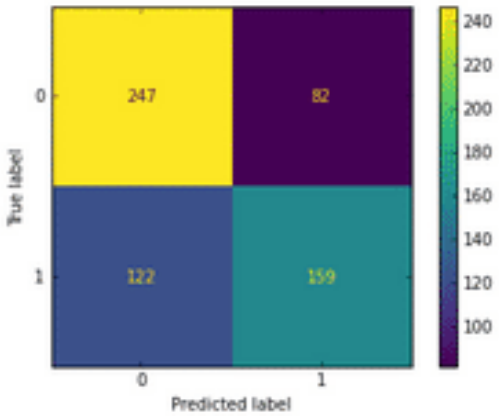


Table 6.

Confusion matrix on the test data

	precision	recall	f1-score	support
0	0.66	0.78	0.72	142
1	0.67	0.53	0.59	120
accuracy			0.66	262
macro avg	0.67	0.65	0.65	262
weighted avg	0.67	0.66	0.66	262

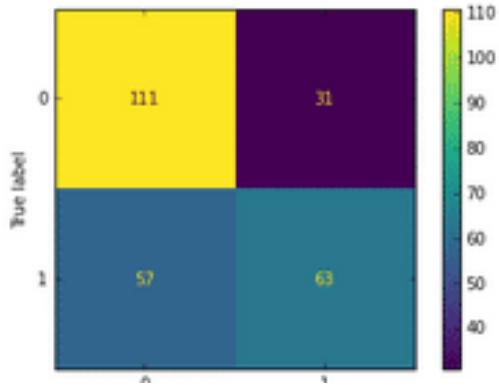


Table 7.

Accuracy - Training Data

0.6655

AUC and ROC for the training data

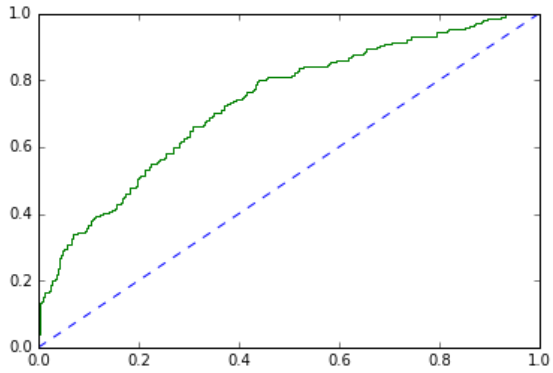


Figure 19.

AUC: 0.734

Accuracy - Test Data

0.664

AUC and ROC for the test data

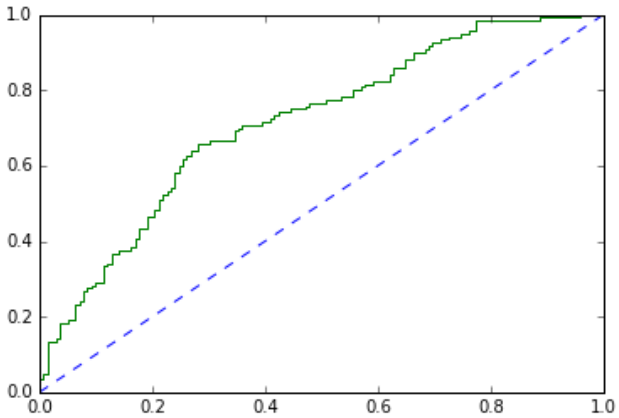


Figure 20.

AUC: 0.823

lr_train_precision 0.66

lr_train_recall 0.57

lr_train_f1 0.61

lr_test_precision 0.67

lr_test_recall 0.52

lr_test_f1 0.59

LDA MODEL

Tfeature: Holliday_Package

['no', 'yes']

Categories (2, object): ['no', 'yes']

[0 1]

feature: foreign

['no', 'yes']

Categories (2, object): ['no', 'yes']

[0 1]

Build LDA Model

lda_train_acc = 0.663

	precision	recall	f1-score	support
0	0.67	0.74	0.70	329
1	0.65	0.58	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

Figure 21.

confusion_matrix

array([[243, 86],

[119, 162]],

lda_test_acc = 0.641

	precision	recall	f1-score	support
0	0.64	0.77	0.70	142
1	0.64	0.49	0.56	120
accuracy			0.64	262
macro avg	0.64	0.63	0.63	262
weighted avg	0.64	0.64	0.63	262

Figure 22.

confusion_matrix

array([[109, 33],

[61, 59]]

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

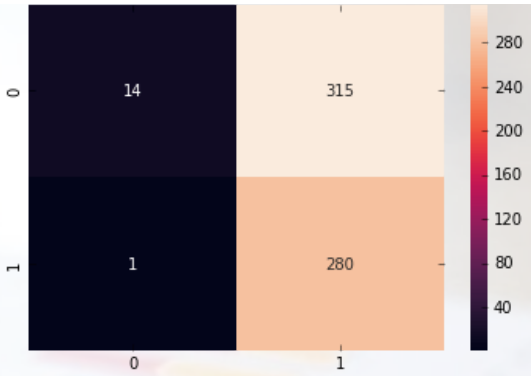
CHANGING THE CUTT OFF VALUE TO CHECK OPTIMAL VALUE THAT GIVES BETTER ACCURACY AND F1 SCORE

0.1

Accuracy Score 0.482

F1 Score 0.6393

Confusion Matrix

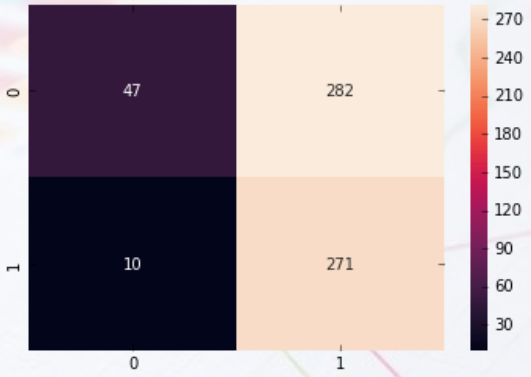


0.2

Accuracy Score 0.5213

F1 Score 0.6499

Confusion Matrix

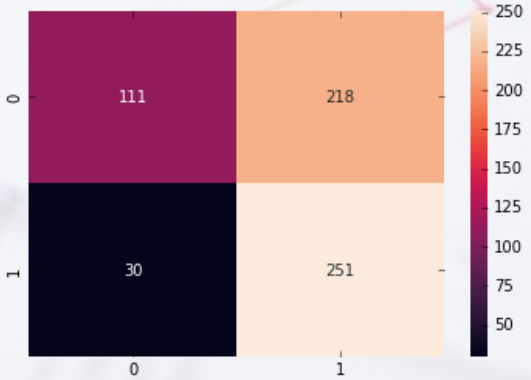


0.3

Accuracy Score 0.5934

F1 Score 0.6693

Confusion Matrix

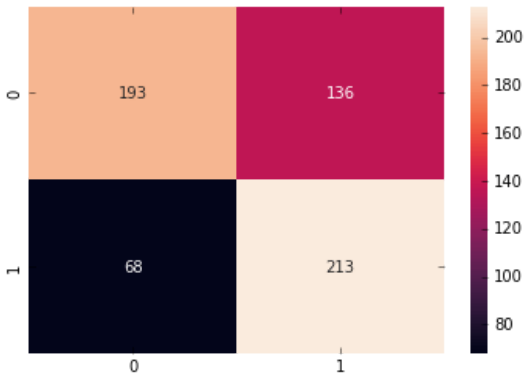


0.4

Accuracy Score 0.6656

F1 Score 0.6762

Confusion Matrix

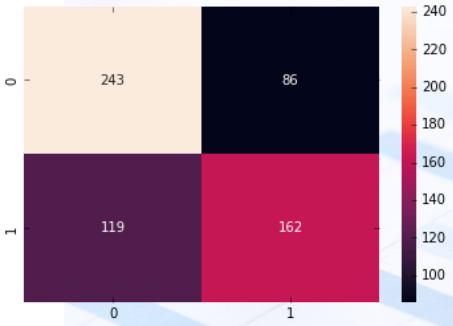


0.5

Accuracy Score 0.6639

F1 Score 0.6125

Confusion Matrix

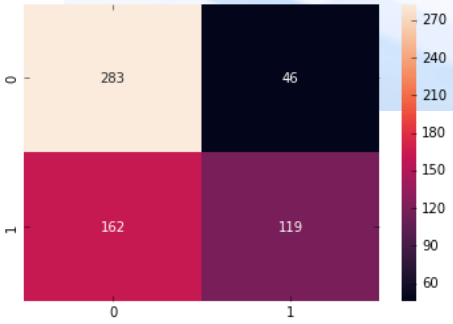


0.6

Accuracy Score 0.659

F1 Score 0.5336

Confusion Matrix

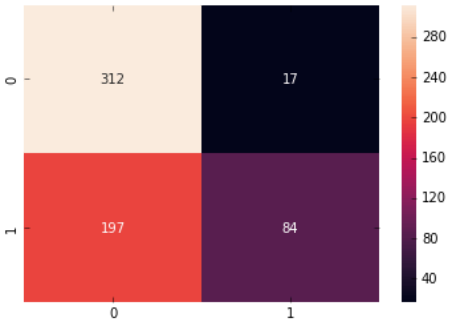


0.7

Accuracy Score 0.6492

F1 Score 0.4398

Confusion Matrix

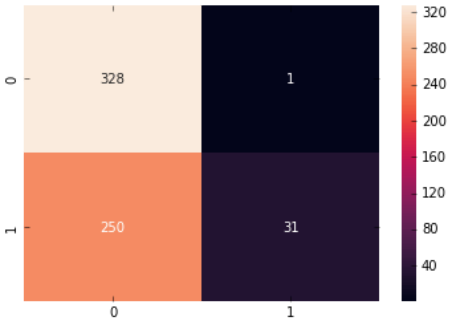


0.8

Accuracy Score 0.5885

F1 Score 0.1981

Confusion Matrix

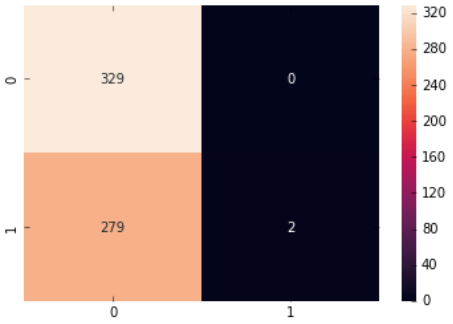


0.9

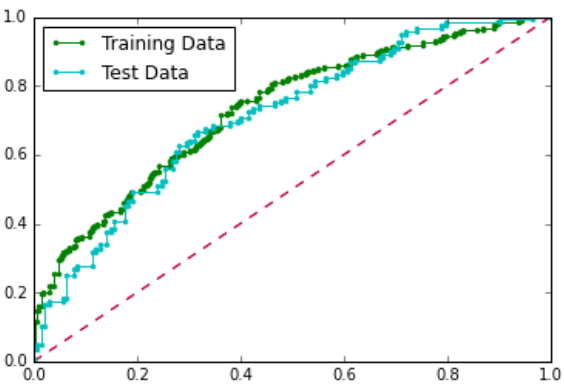
Accuracy Score 0.5426

F1 Score 0.0141

Confusion Matrix



AUC and ROC for the training and test data



AUC for the Training Data: 0.733

AUC for the Test Data: 0.714

Figure 23.

lda_train_precision 0.65

lda_train_recall 0.58

lda_train_f1 0.61

lda_test_precision 0.64

lda_test_recall 0.49

lda_test_f1 0.56

	LR Train	LR Test	LDA Train	LDA Test
Accuracy	0.67	0.66	0.66	0.64
AUC	0.73	0.72	0.73	0.71
Recall	0.57	0.52	0.58	0.49
Precision	0.66	0.67	0.65	0.64
F1 Score	0.61	0.59	0.61	0.56

Table 8.

Comparing both these models, we find both results are same, but LDA works better when there is category target variable.

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

we had done predictions on both logistic regression and linear discriminant analysis. Both results are same.

EDA Analysis

- People ranging from the age 30 to 50 generally opt for holiday packages.
- We can say that maximum foreign employees didn't choose the package.
- We can say that maximum employees with 0 kids choose the holiday package.
- We can say that maximum employees with 2,3 and 4 older_children choose the holiday package. And employees with 0 older_children didn't choose the package.
- We can say maximum employee didn't choose the package and in that maximum employee with 0 older children choose the package then employee with 2 older kids choose the package.
- 35 employees with age of 44 have purchased holiday package and only 3 employees with age of 62 have purchased holiday package.
- Salary less than 50000 people have opted more for holiday package.
- Employee age over 50 to 60 have seems to be not taking the holiday package, whereas in the age 30 to 50 and salary less than 50000 people have opted more for holiday package.

The important factors deciding the predictions are salary, age and educ.

Recommendations

- For employee having more than number of older children we can provide packages in holiday vacation places.
- For people earning more than 150000 we can provide vacation holiday packages.