July 2021 | PG-DSBA-Online

PROJECT REPORT PREDICTIVE MODELLING

SUBMITTED BY
DEV TRIPATHI

Contents

List of Figures

List of Tables

Problem 1

Cubic Zirconia Dataset Analysis

Information about dataset

The dataset is provided by a company named Gem Stones co ltd, which is a cubic zirconia manufacturer. This dataset contains 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).

Data Dictionary for Cubic Zirconia Dataset

The dataset contains the following features and the information about these features is mentioned below:

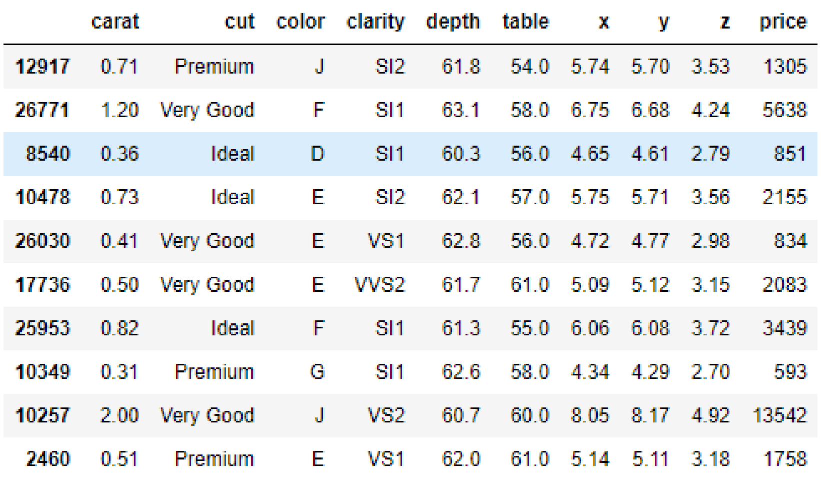
Variable Name	**Description**
Carat	Carat weight of the cubic zirconia.
Cut	Describes the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Color of the cubic zirconia with D being the best and J the worst.
Clarity	Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I1= level 1 inclusion) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, I1
Depth	The Height of cubic zirconia, measured from the Culet to the table, is divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	Price of the zirconia
X	Length of cubic Zirconia in mm.
Y	Width of cubic Zirconia in mm.
Z	Height of cubic Zirconia in mm.

Table 1: Data Dictionary (Cubic Zirconia Dataset)

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

EDA

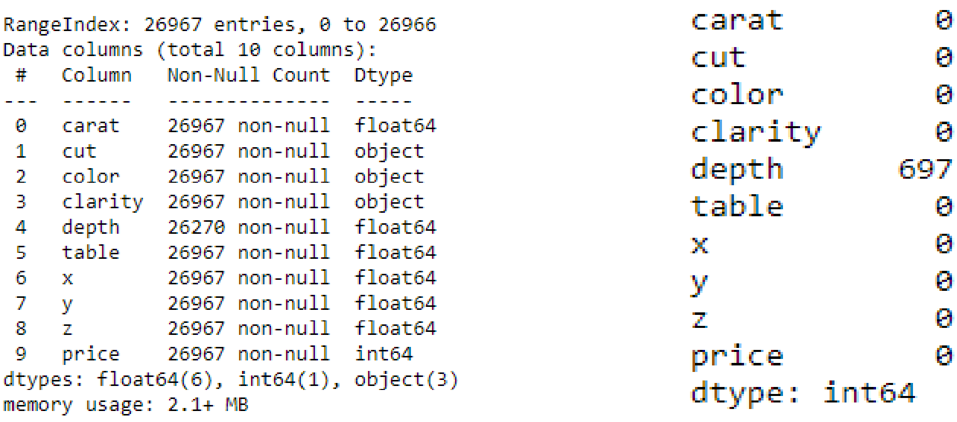
Sample of Dataset-



	carat	cut	color	clarity	depth	table	x	y	z	price
12917	0.71	Premium	J	SI2	61.8	54.0	5.74	5.70	3.53	1305
26771	1.20	Very Good	F	SI1	63.1	58.0	6.75	6.68	4.24	5638
8540	0.36	Ideal	D	SI1	60.3	56.0	4.65	4.61	2.79	851
10478	0.73	Ideal	E	SI2	62.1	57.0	5.75	5.71	3.56	2155
26030	0.41	Very Good	E	VS1	62.8	56.0	4.72	4.77	2.98	834
17736	0.50	Very Good	E	VVS2	61.7	61.0	5.09	5.12	3.15	2083
25953	0.82	Ideal	F	SI1	61.3	55.0	6.06	6.08	3.72	3439
10349	0.31	Premium	G	SI1	62.6	58.0	4.34	4.29	2.70	593
10257	2.00	Very Good	J	VS2	60.7	60.0	8.05	8.17	4.92	13542
2460	0.51	Premium	E	VS1	62.0	61.0	5.14	5.11	3.18	1758

Figure 1: Sample of Cubic Zirconia Dataset

Variable Information-



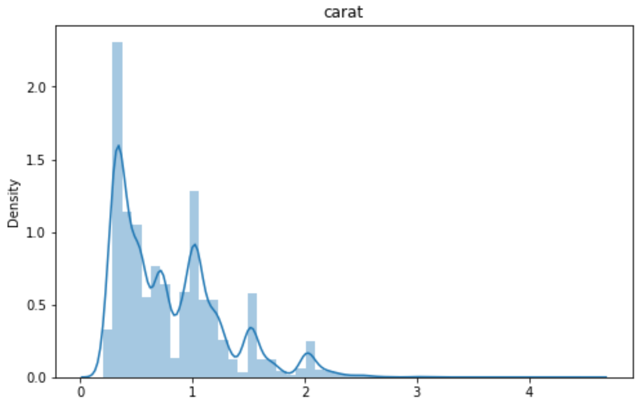
```
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object 
 2   color    26967 non-null  object 
 3   clarity  26967 non-null  object 
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64  
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

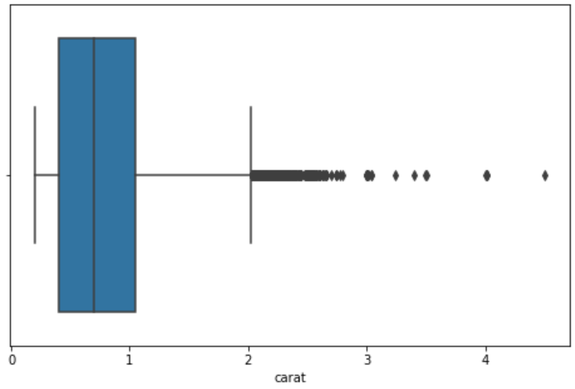
```
carat        0
cut          0
color        0
clarity      0
depth      697
table        0
x            0
y            0
z            0
price        0
dtype: int64
```

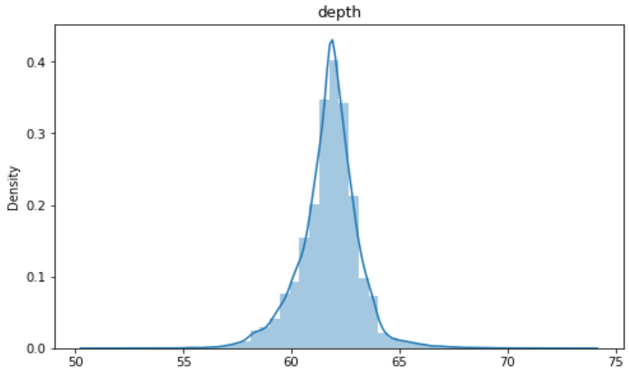
Figure 2: Feature Info and null values count of Cubic Zirconia Dataset

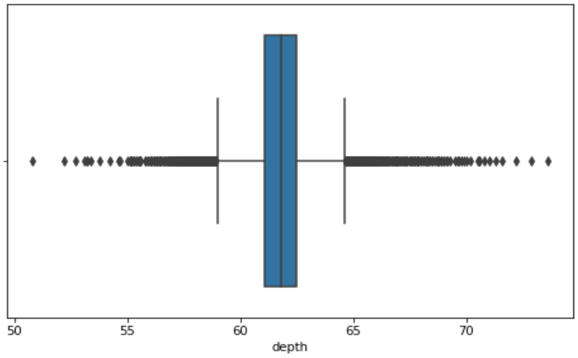
The dataset contains a total of 26967 entries and 10 features including our target variable which is **‘price’**. The **‘depth’** feature contains a total of 697 null values. None of the variables contains null values other than the **‘depth’** variable. There are three columns i.e., ’cut’, ’color’, and ‘clarity’ which are object data types. All other variables are of numeric data type.

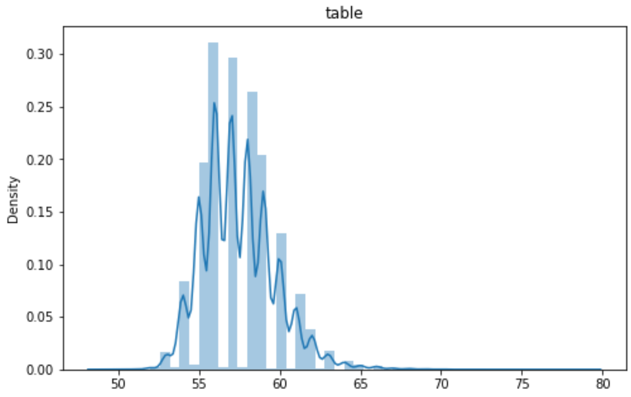
Univariate Analysis-

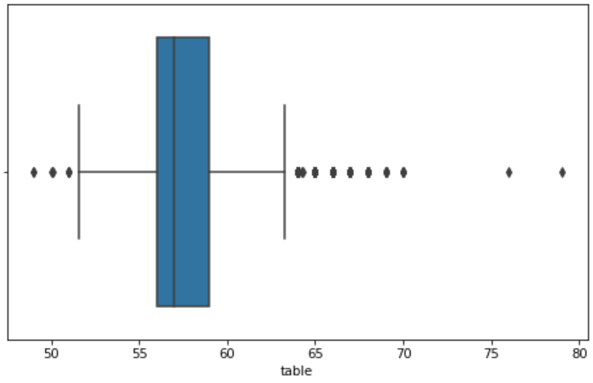


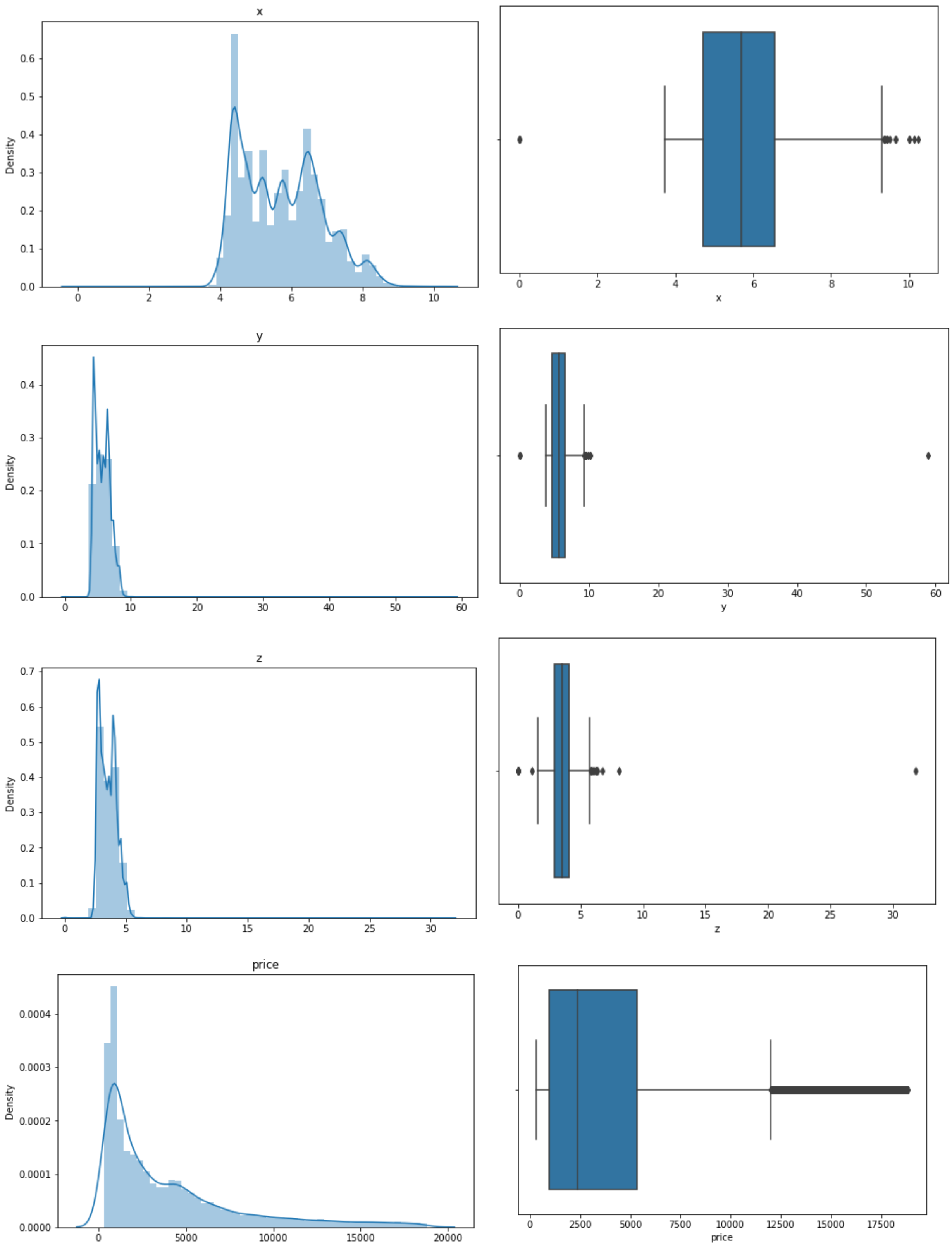











x
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
x
y
Density
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
50
60
y
z
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
5
10
15
20
25
30
z
price
Density
0.0000
0.0001
0.0002
0.0003
0.0004
0
5000
10000
15000
20000
0
2500
5000
7500
10000
12500
15000
17500
price

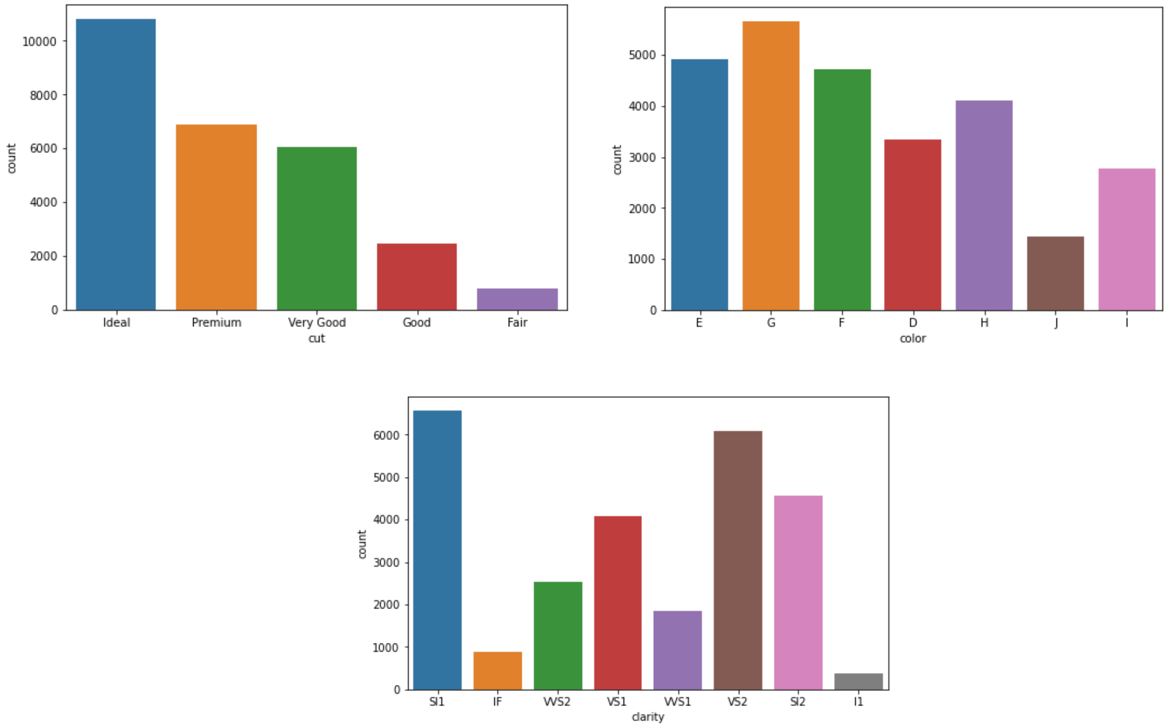


Figure 3: Univariate Curves for different features in the dataset

Now, conclusions drawn from this analysis can be briefly explained in the following points:

1. From the boxplots, it can be observed that most of the variables are having outliers
2. Though the X, Y, and Z variables have some very extreme outliers which can be problematic for further analysis.
3. All the variables are having Right skewed distributions except the ‘depth’ variable which seems to be having very little skewness.
4. Right skewed distributions show that zirconia having good values for attributes (e.g., Carat, length, width, etc.) are rare.
5. The count plot of cut shows that most of the product has Ideal or premium cut which is good.
6. The clarity of a very few zirconias lies in Fl(flawless) category. The most frequent clarity observed is the Sl1 type.

Bi-Variate Analysis-

Now for bivariate analysis, we have created a pair plot that simply shows a scatterplot for all the pairs of features present in the dataset.

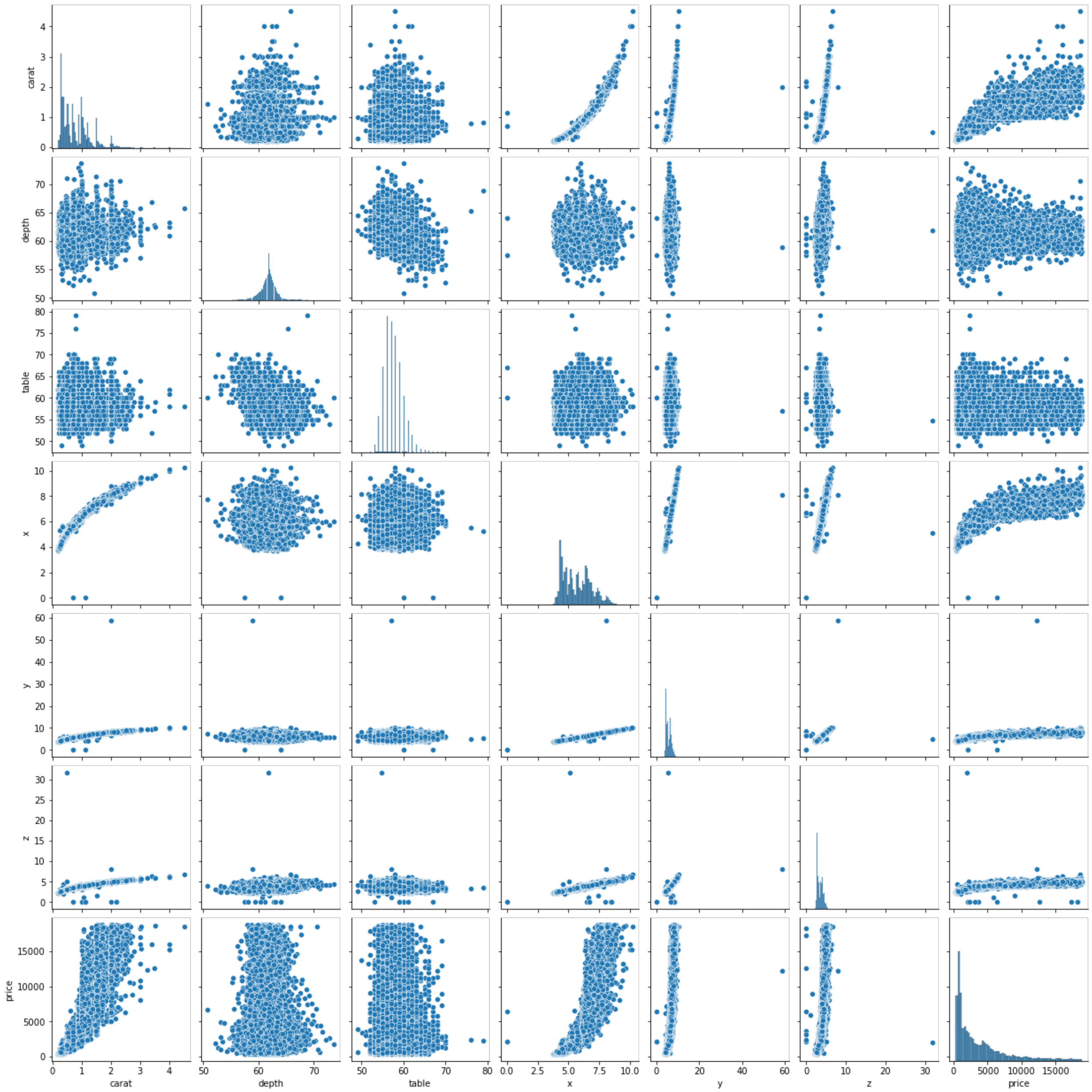


Figure 4: Pairplot for Cubic Zirconia Dataset

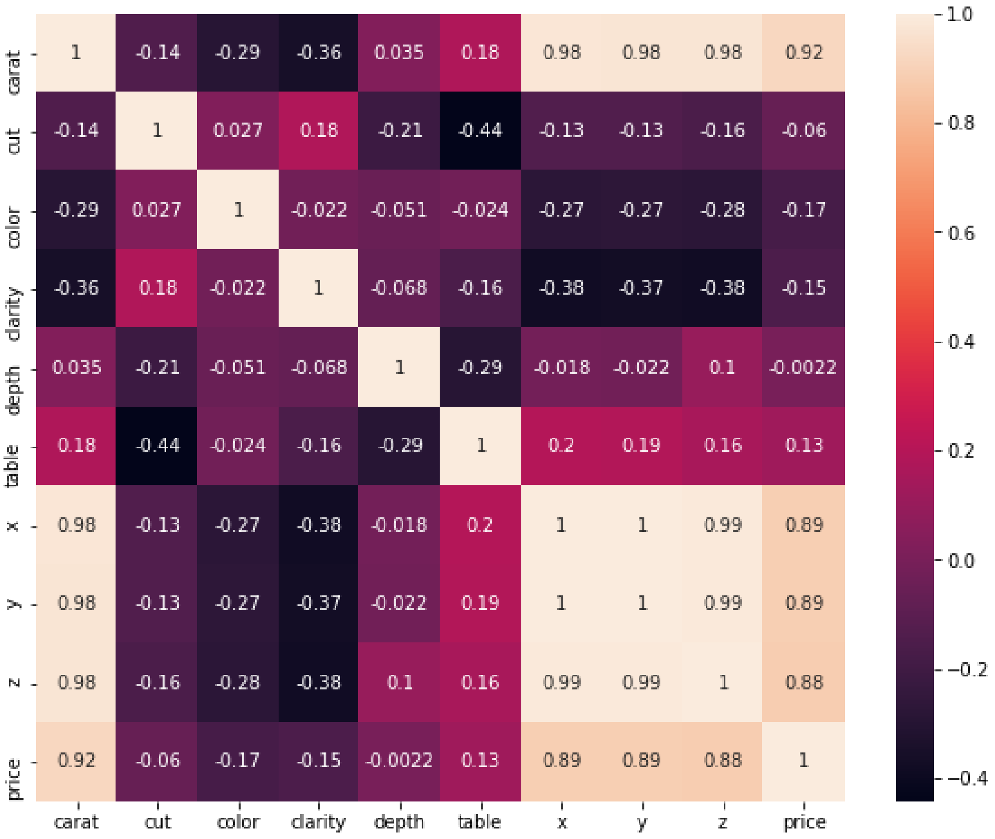


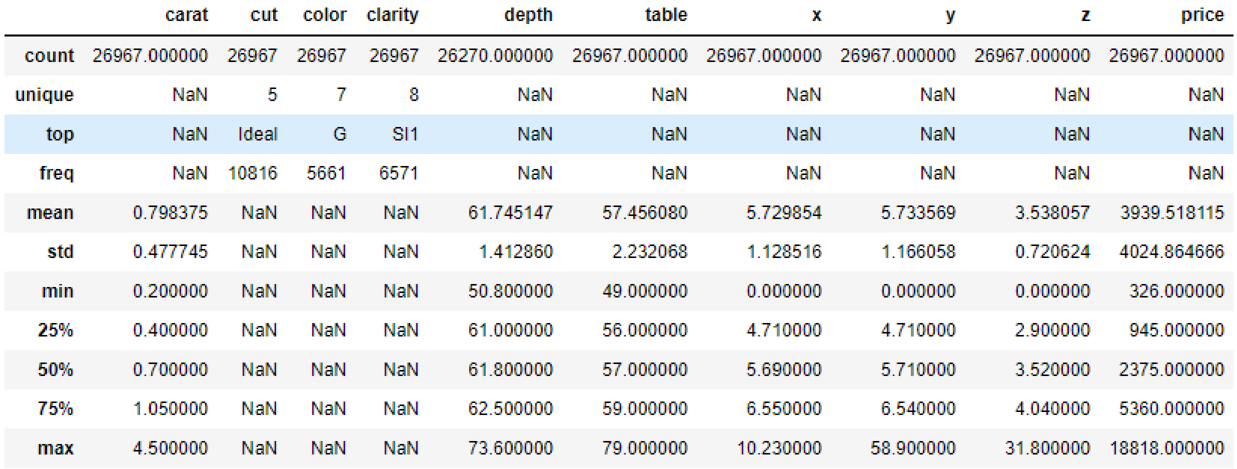
Figure 5: Correlation plot for different features

For the correlation plot encoding of categorical features is necessary. After encoding the categories in numeric form (please refer to the notebook for the labels/steps involved in encoding), the correlation plot was created. In the encoding process higher number was chosen to represent better attributes for example Fl(flawless) clarity was given 7 as the label.

Now, one can observe that the 'carat', 'x', 'y', 'z' and 'price' features have a high correlation between them (Which could be problematic for the regression model). The 'clarity' feature shows a negative correlation with x, y, z, and carat features and the 'cut' variable shows a negative correlation with the 'table' feature. The rest of the features have low to minimal correlation with any other variable. These correlations can be observed in pairplot also.

1.2 Impute null values if present, also checks for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?

Sol: As mentioned earlier, only the ‘depth’ variable has null values (null count = 697). Now, the imputation of these null values can be done in several ways like mean, median, mode, or model-based imputation. Here, **median imputation** has been done. Though the value of mean and median for the depth feature is almost the same (mean= 61.745147 and median = 61.8).



	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967.000000	26967	26967	26967	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
unique	NaN	5	7	8	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	Ideal	G	SI1	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	10816	5661	6571	NaN	NaN	NaN	NaN	NaN	NaN
mean	0.798375	NaN	NaN	NaN	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	0.477745	NaN	NaN	NaN	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	0.200000	NaN	NaN	NaN	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	NaN	NaN	NaN	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	NaN	NaN	NaN	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	1.050000	NaN	NaN	NaN	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	4.500000	NaN	NaN	NaN	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

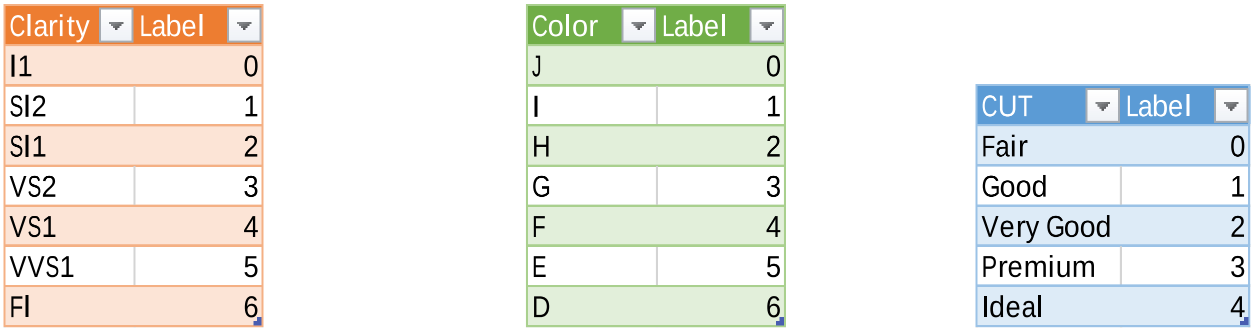
Figure 6: Description of the dataset

Now, from the description of the data, one can observe that only ‘x’, ‘y’, and ‘z’ columns have minimum values 0 (which was earlier shown in the respective boxplots). Since these are physical dimensions of zirconia the 0 value seems absurd. Though, only 11 observations were found to have the x y or z value as 0. Hence, considering the number of observations available to us which is around 27k, these samples were **dropped for further analysis**.

The scaling of the dataset becomes necessary for any model if the range of values for different features is very different. For regression models scaling is necessary for interpreting the feature importance and also preventing the model from being biased. Now, in our case, the ‘price’ ranges from around 4k to 18k whereas the values for ‘carat’ ranges from 0.2 to 4.5. Hence, due to this huge difference in ranges, scaling of the features was done before proceeding to model and prediction steps.

1.3 Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Linear regression. Performance Metrics: Check the performance of Predictions on Train and Test sets using R-square, RMSE.

Sol: As mentioned earlier the encoding was done such that the better attributes are labeled with higher values. The codes for the same are given below:



Clarity	Label
I1	0
SI2	1
SI1	2
VS2	3
VS1	4
VVS1	5
FI	6

Color	Label
J	0
I	1
H	2
G	3
F	4
E	5
D	6

CUT	Label
Fair	0
Good	1
Very Good	2
Premium	3
Ideal	4

Figure 7: Labels of different categories present in features

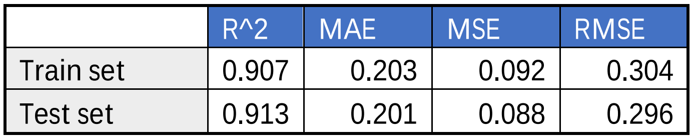
Now, after encoding and scaling, the dataset was split into train and test datasets for evaluation of the model. As per the instructions, the 70,30 split was performed (70 percent samples for train and 30 percent samples for test set).

Linear Regression Model

After the train and split, a Linear regression model was fitted to the training data The actual model obtained from the analysis can be summarized by the following equation-

(-7598.6) * Intercept + (10875.33) * carat + (140.45) * cut + (333.46) * color + (501.04) * clarity + (96.36) * depth + (-19.08) * table + (-1754.0) * x + (2594.81) * y + (-2794.93) * z = Price of zirconia

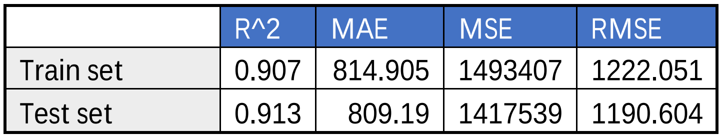
What it represents is that, if the carat value is to be increased by a single unit (1 unit) the resulting increase due to this price change is around 11k units (keeping other variables constant). Performance metrics were calculated for the train and test set which are following:



	R^2	MAE	MSE	RMSE
Train set	0.907	0.203	0.092	0.304
Test set	0.913	0.201	0.088	0.296

Table 2: Performance Metrics (Linear Regression model)

From the metrics obtained from the model, one can say that the model is working fine for both the train and test set. Also, the model performance is comparable for both sets considering the above-mentioned metrics. Though one must keep in mind that these values are for scaled data, not the actual data. For unscaled data the metrics are given below:

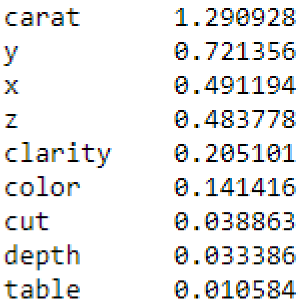


	R^2	MAE	MSE	RMSE
Train set	0.907	814.905	1493407	1222.051
Test set	0.913	809.19	1417539	1190.604

Table 3: Performance metrics for unscaled data

Hence, from the above RMSE values we can say that, if this model is used for predicting the price of zirconia, the actual and predicted value can differ by around 1200 units.

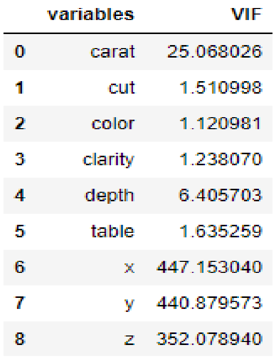
Now, if we look at the coefficients (for scaled dataset) of different features, we can say that ‘carat’, ‘x’, ‘y’, ‘z’ (means the dimensions of zirconia) and ‘clarity’ are better predictors for the price compared to other variables. But one must keep in mind that multicollinearity hasn’t been considered in the analysis.



```
carat      1.290928
y          0.721356
x          0.491194
z          0.483778
clarity    0.205101
color      0.141416
cut        0.038863
depth      0.033386
table      0.010584
```

Figure 8: Coefficient values for features present in the dataset

Though in the analysis here, we have ignored the VIF (variance inflation factor) calculation for addressing the multicollinearity problem. We can look at the VIF values to say which features are redundant for the prediction.



	variables	VIF
0	carat	25.068026
1	cut	1.510998
2	color	1.120981
3	clarity	1.238070
4	depth	6.405703
5	table	1.635259
6	x	447.153040
7	y	440.879573
8	z	352.078940

Figure 9: VIF values for features present in the dataset

As it is visible from the values that the x,y,z, and carat have very high VIF values, which means one can eliminate these features and keep only the best ones (predictors) for the model. Though in this analysis, the variables were simply assumed to be independent.

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

Sol:

Inferences

The linear regression model is one of the simplest yet powerful models available. The inferences drawn from the above analysis can be discussed in the following points:

1. The price of cubic zirconia is highly affected by its dimension and carat value.
2. The dimensions correlate with each other, we can be interpreted as most of the crystals mentioned in the dataset are not flaky.
3. For further optimization, one can remove the features with high values of VIF (>10) and try to get better results.
4. Multicollinearity is present in the dataset though for this analysis it was not considered to be problematic.
5. Error distribution obtained from the analysis was found to be slightly skewed to left.
6. Overall performance of the model was found to be good in this analysis, though this model is not recommended to be used for real-world cases because multicollinearity is present which can make the model unstable.

Problem 2

Holiday Package Dataset Analysis

Information about the Dataset

A tour and travel agency which deals in selling holiday packages. The details of 872 employees of a company are provided by the agency. Among these employees, some opted for the package and some didn't. We are to help the company in predicting whether an employee will opt for the package or not based on the information given in the data set. Also, we have to identify the important factors based on which the company will focus on particular employees to sell their packages.

Data Dictionary

Variable Name	Description
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	Foreigner yes/no?

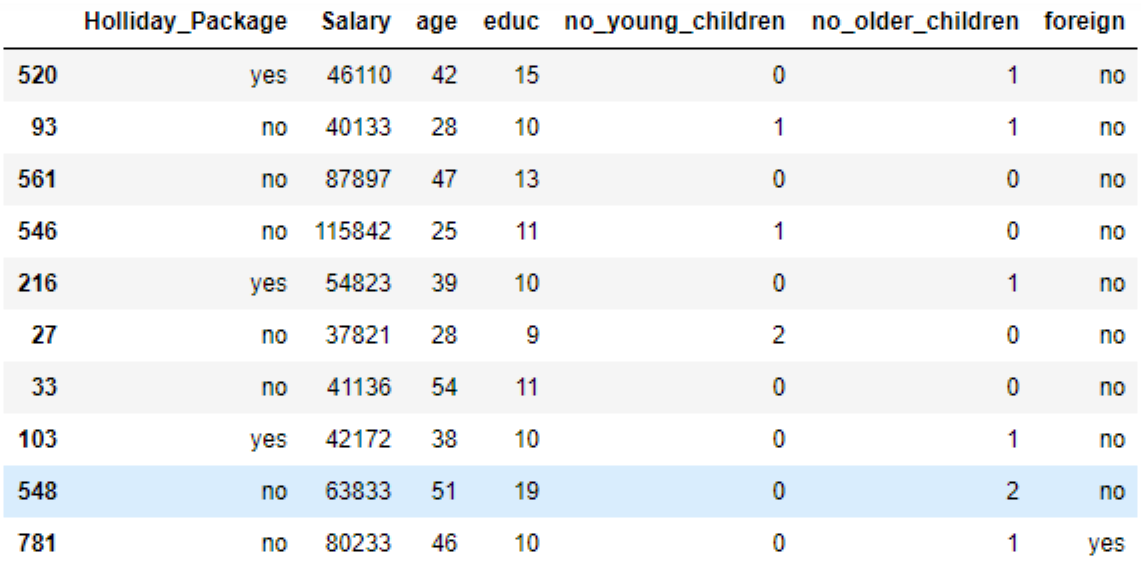
Table 4: Data Dictionary for Holiday Package Dataset

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.

Sol:

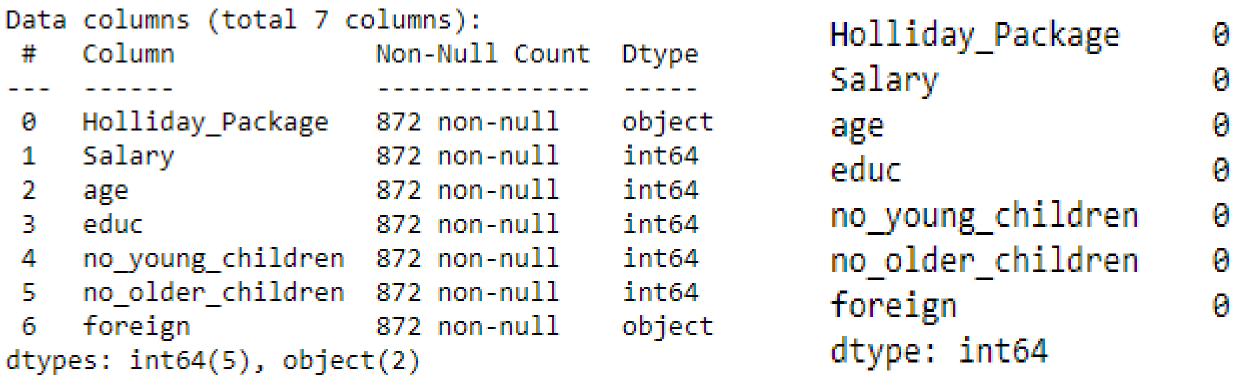
EDA

Sample of dataset



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
520	yes	46110	42	15	0	1	no
93	no	40133	28	10	1	1	no
561	no	87897	47	13	0	0	no
546	no	115842	25	11	1	0	no
216	yes	54823	39	10	0	1	no
27	no	37821	28	9	2	0	no
33	no	41136	54	11	0	0	no
103	yes	42172	38	10	0	1	no
548	no	63833	51	19	0	2	no
781	no	80233	46	10	0	1	yes

Figure 10: Sample of Holiday package dataset

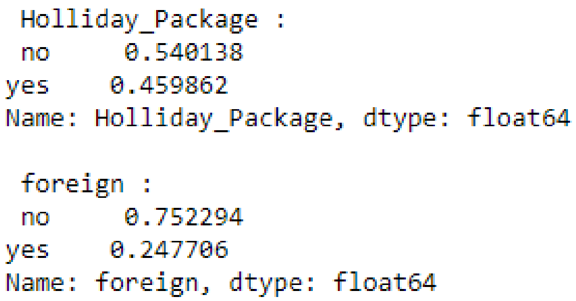


```
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2   age                872 non-null    int64 
 3   educ               872 non-null    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)
```

```
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

Figure 11: Feature info and null count of the dataset

As we can observe that the dataset has a total of 872 entries and no null values. There is a total of 6 variables including our target variable which is ‘Holiday_package’. The individual value counts for these object type variables are following:

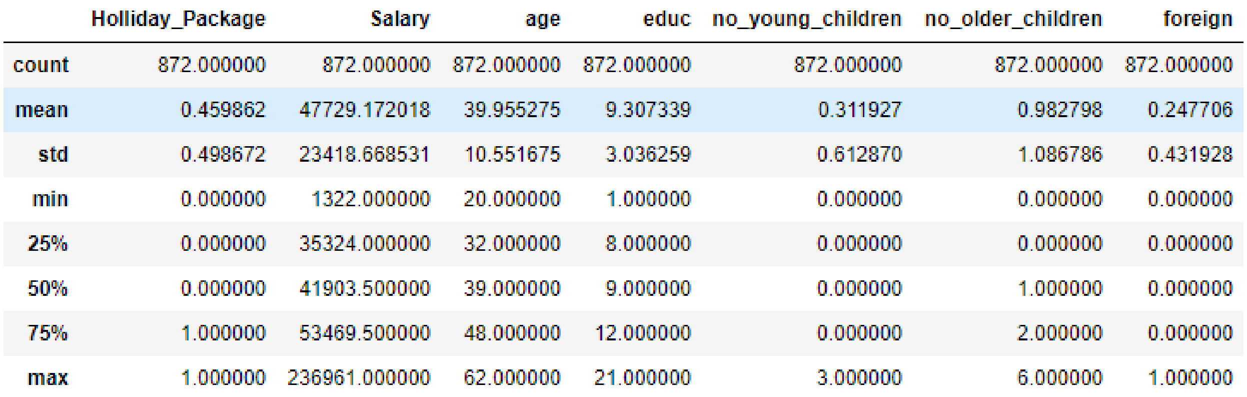


```
 Holliday_Package :
 no      0.540138
yes     0.459862
Name: Holliday_Package, dtype: float64

 foreign :
 no      0.752294
yes     0.247706
Name: foreign, dtype: float64
```

Figure 12: Value counts for each attribute in the dataset

As we can see that there are 54 percent ‘no’ values and 46 percent ‘yes’ values. So, we can say that the target variable is fairly balanced.

Descriptive Statistics



	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	0.459862	47729.172018	39.955275	9.307339	0.311927	0.982798	0.247706
std	0.498672	23418.668531	10.551675	3.036259	0.612870	1.086786	0.431928
min	0.000000	1322.000000	20.000000	1.000000	0.000000	0.000000	0.000000
25%	0.000000	35324.000000	32.000000	8.000000	0.000000	0.000000	0.000000
50%	0.000000	41903.500000	39.000000	9.000000	0.000000	1.000000	0.000000
75%	1.000000	53469.500000	48.000000	12.000000	0.000000	2.000000	0.000000
max	1.000000	236961.000000	62.000000	21.000000	3.000000	6.000000	1.000000

Figure 13: statistical Description of all the features

Univariate Analysis

For univariate analysis we can look at distribution plot and boxplots of the attributes present in the dataset:

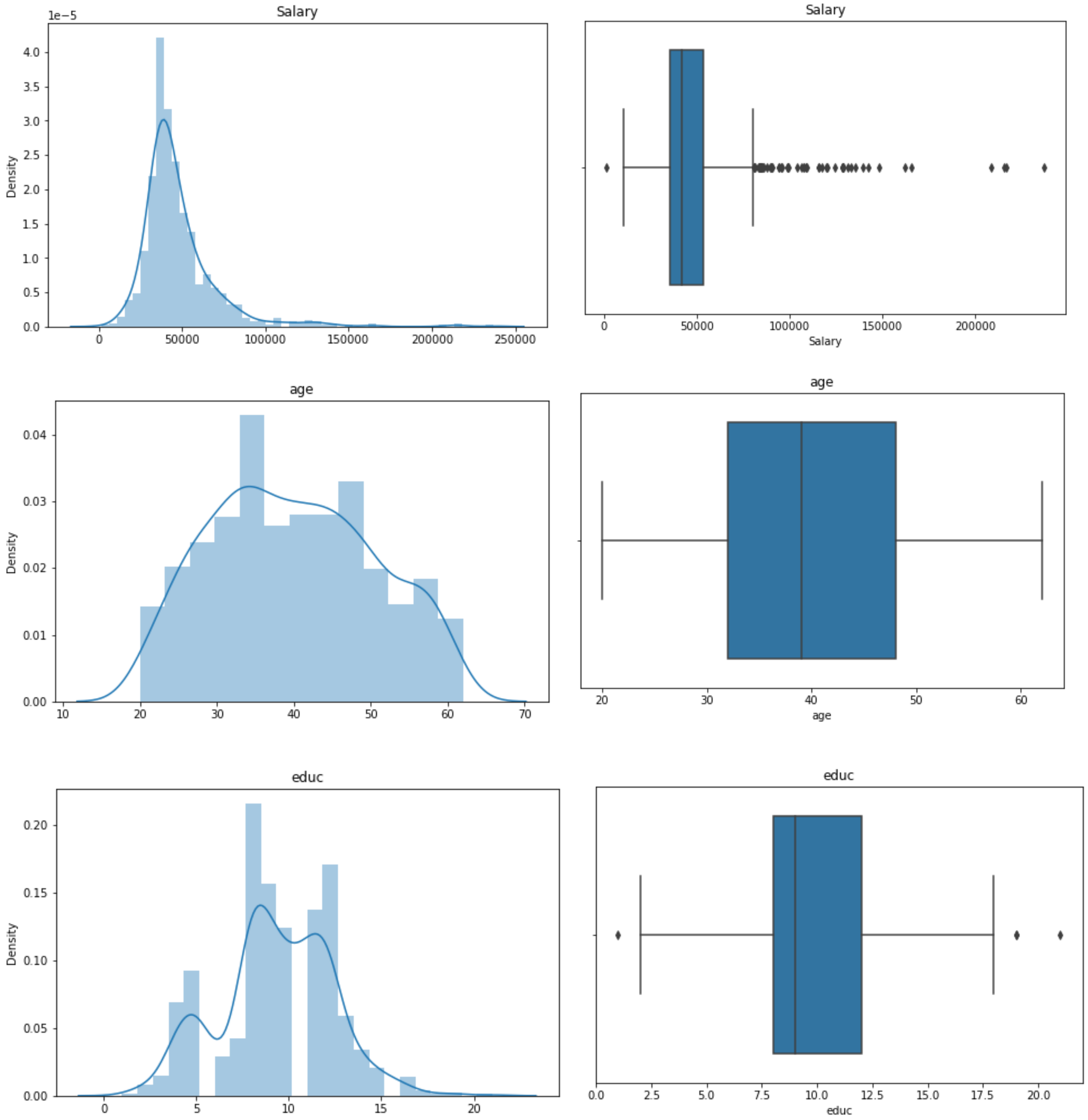


Figure 14: Univariate Analysis for continuous variables

Observations-

- The education variable has a few outliers. But these outliers are genuine outliers. Also, the frequency of the curve is highest between 8 to 12. This means a majority of the employees have 8 to 12 years of formal education.

- Age variable have no outliers. From the distribution curve, we can see that majority of the people who are an employee in the company are between the age group of 30 to 50.
- The salary variable has a lot of outliers. The distribution was found to be right-skewed.

Now, let's look at the count plot for each object type variable present in the dataset.

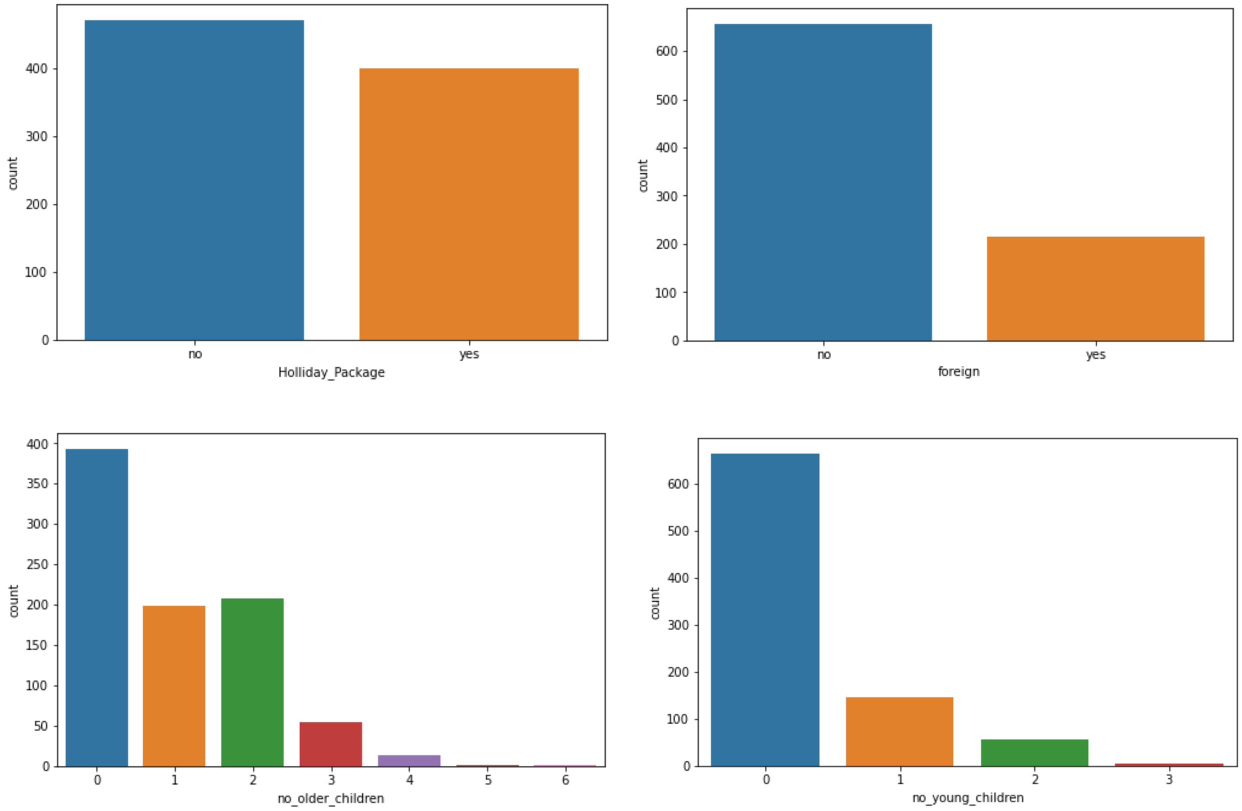


Figure 15: Count plots for object/categorical type attributes

Observations:

- The majority of the employees are natives. Only 24 percent of employees are foreigners.
- Less than 100 employees have 3 or more older children.
- Half of the number of employees have no children.
- The target variable i.e., Holiday_package is fairly balanced as found earlier from value counts.

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).

Logit model & LDA

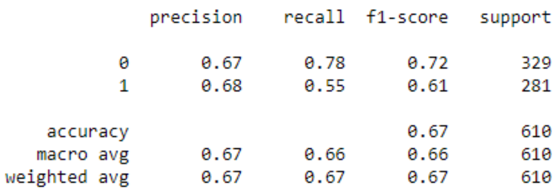
Now, before proceeding the further analysis we have to encode the string type data. Since the categories available in both the target variable (Holiday_package) and the foreigner variable is ‘yes’ and ‘no’. Hence these values were simply be replaced by 0 and 1 as, 0 refers to ‘no’ and 1 refers to the ‘yes’ category.

After doing this, we can proceed to train and test split. As per the instructions, the 70,30 split was performed (70 percent samples for train and 30 percent samples for test set). Here, one must look out for the class balance ratio in train and test. Now, the Logistic regression model was fitted to train data, and predictions were made for train and test data. Similarly, Linear Discriminant Analysis (LDA) was performed on the dataset.

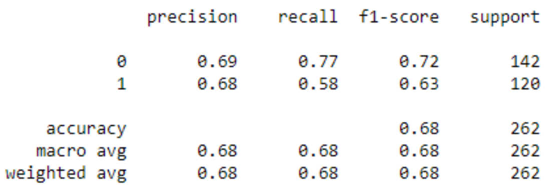
2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Performance Metrics

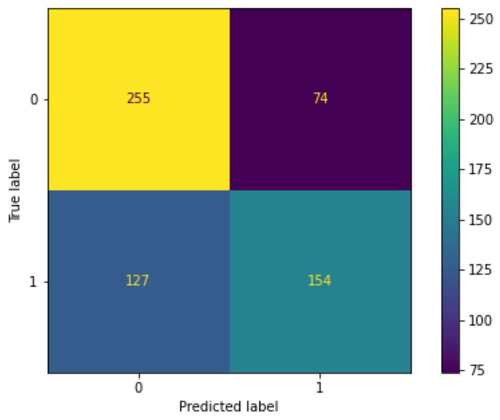
Now, if we look at the performance metrics of models, these come out to be very similar. Though the logistic regression model needed hyperparameter tuning to get the performance comparable to the LDA.

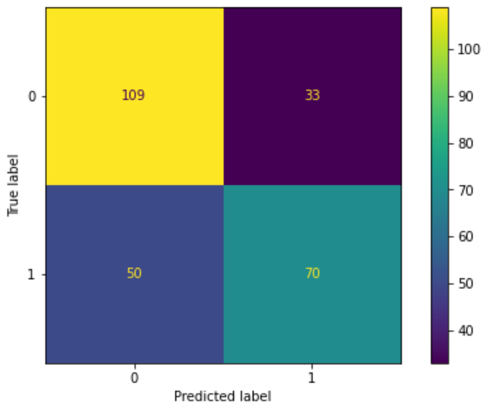


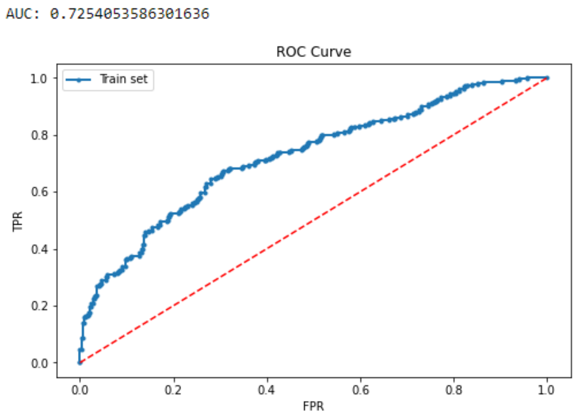
```
              precision    recall  f1-score   support

           0       0.67      0.78      0.72       329
           1       0.68      0.55      0.61       281

    accuracy                           0.67       610
   macro avg       0.67      0.66      0.66       610
weighted avg       0.67      0.67      0.67       610
```



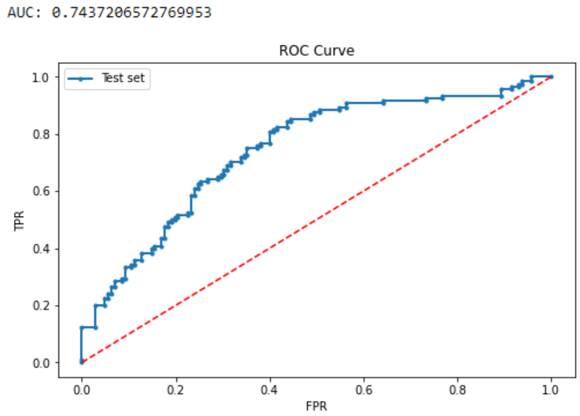
```
              precision    recall  f1-score   support

           0       0.69      0.77      0.72       142
           1       0.68      0.58      0.63       120

    accuracy                           0.68       262
   macro avg       0.68      0.68      0.68       262
weighted avg       0.68      0.68      0.68       262
```





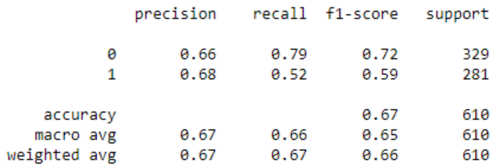


```
AUC: 0.7254053586301636
```

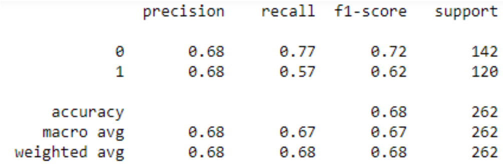


```
AUC: 0.7437206572769953
```

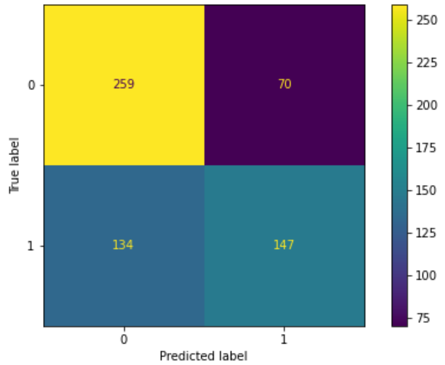
Figure 16: Performance metrics using Logit model for train set (left) and test set (Right)

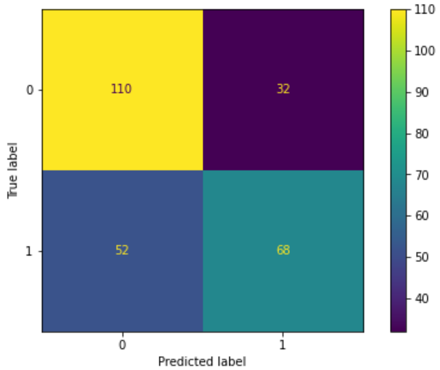


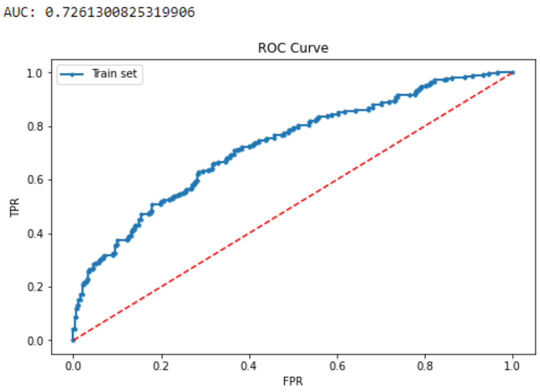
```
              precision    recall  f1-score   support

           0       0.66      0.79      0.72       329
           1       0.68      0.52      0.59       281

    accuracy                           0.67       610
   macro avg       0.67      0.66      0.65       610
weighted avg       0.67      0.67      0.66       610
```



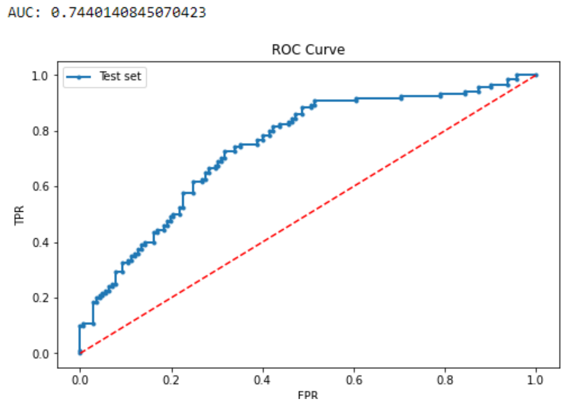
```
              precision    recall  f1-score   support

           0       0.68      0.77      0.72       142
           1       0.68      0.57      0.62       120

    accuracy                           0.68       262
   macro avg       0.68      0.67      0.67       262
weighted avg       0.68      0.68      0.68       262
```







```
AUC: 0.7261300825319906
```



```
AUC: 0.7440140845070423
```

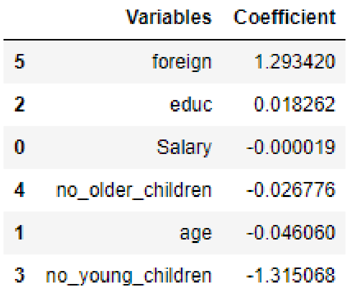
Figure 17: Performance Metrics for LDA on Train set (left) and Test set (right)

As it can be observed in the above figures, both the models are struggling to make good predictions considering the overall accuracy of the models. One of the reasons can be due to weak predictors available as we observed in Bi-variate analysis. LDA model is performing slightly better than the logit model considering the AUC scores. Hence, we can say that the LDA model is slightly more optimized for the predictions.

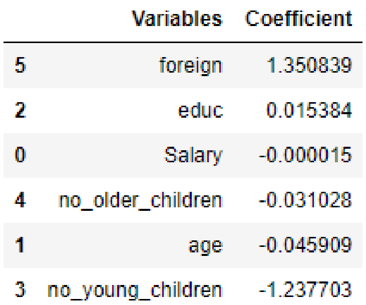
2.4 Inference: Basis on these predictions, what are the insights and recommendations.

Inferences

Now, if we look at the feature importance values calculated by the models, they come out to almost equal:



	Variables	Coefficient
5	foreign	1.293420
2	educ	0.018262
0	Salary	-0.000019
4	no_older_children	-0.026776
1	age	-0.046060
3	no_young_children	-1.315068



	Variables	Coefficient
5	foreign	1.350839
2	educ	0.015384
0	Salary	-0.000015
4	no_older_children	-0.031028
1	age	-0.045909
3	no_young_children	-1.237703

Figure 18: Coefficient for the variable present in the dataset for Logit model(left) and LDA (right)

Notice that the values for age, salary, educ, and no_older_children are almost zero which means that these are weak predictors for the target variables. On the other hand, the 'foreign' feature is more important for class 1 and 'no_young_children' is important for class 0 in both models.

Now, the following conclusions can be drawn from this analysis:

- Most of the features available in the dataset are weak predictors for the target feature.
- Model performance for both the models were coming out to be almost the same. Though for the LDA model the AUC score was slightly better on both train and test sets.
- More features and data of the employees are required for making the model perform better.
- Based on the analysis, the number of older children and foreigner status are important factors for predicting the holiday package status.