PREDICTIVE MODELING PROJECT REPORT

KIRAN.N

GREAT LEARNING

Table of Contents

List Of Figures

List of Tables

Problem 1

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are 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. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

Data Dictionary

- 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 worst and J the best.
- Clarity: Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1
- Depth: The Height of cubic zirconia, measured from the Culet to the table, 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 cubic zirconia.
- X: Length of the cubic zirconia in mm.
- Y: Width of the cubic zirconia in mm.
- Z: Height of the cubic zirconia in mm.

Q 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.

Sample of the Dataset

	carat	cut	color	clarity	depth	table	x	y	z	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: Sample Dataset

Dataset has 10 columns and 26967 rows. Each row in the dataset corresponds to individual cubic zirconia's price and other attributes.

Data Types

Let us check the datatypes of variables in dataframe.

```
carat      float64
cut         object
color       object
clarity     object
depth      float64
table      float64
x          float64
y          float64
z          float64
price        int64
```

Out of 10 columns, 3 columns are of object type, one column is of integer type and remaining 6 columns are of float type.

Null Check

```
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
```

From the above results we see that there 697 are null values present in the depth column of dataset.

Duplicate Check

There are total 34 duplicate rows in the given dataset. These duplicates are removed before we proceed with the modelbuilding activity.

Check for Outliers

Let us plot the boxplot for all the numeric columns of the dataset.

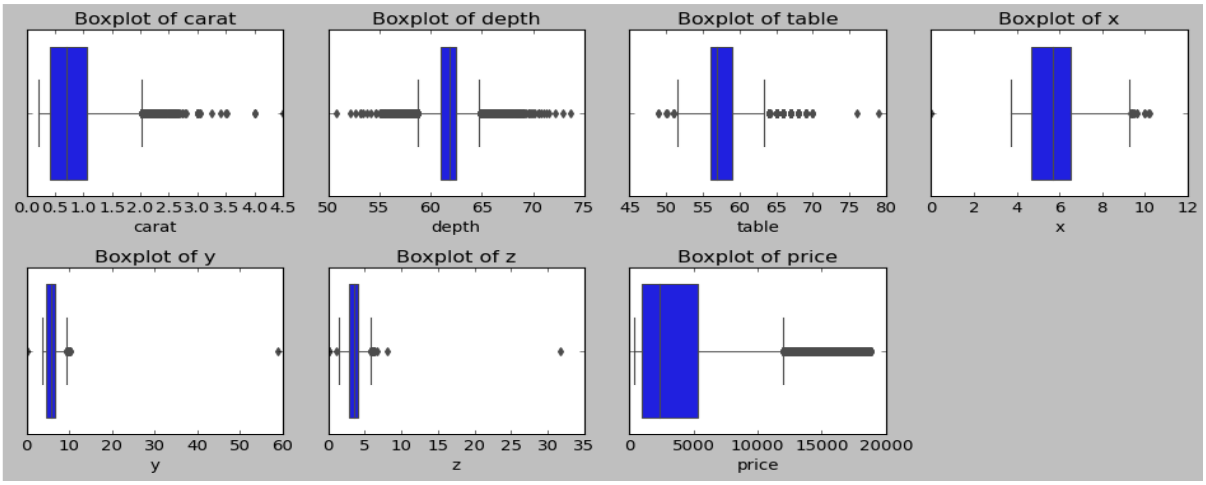

Figure 1: BoxPlot of All Numeric Columns

From the above figure we observe outliers are present in all the numeric columns

Uni-Variate Analysis

Let us plot the histogram for all the numeric columns of the dataset.

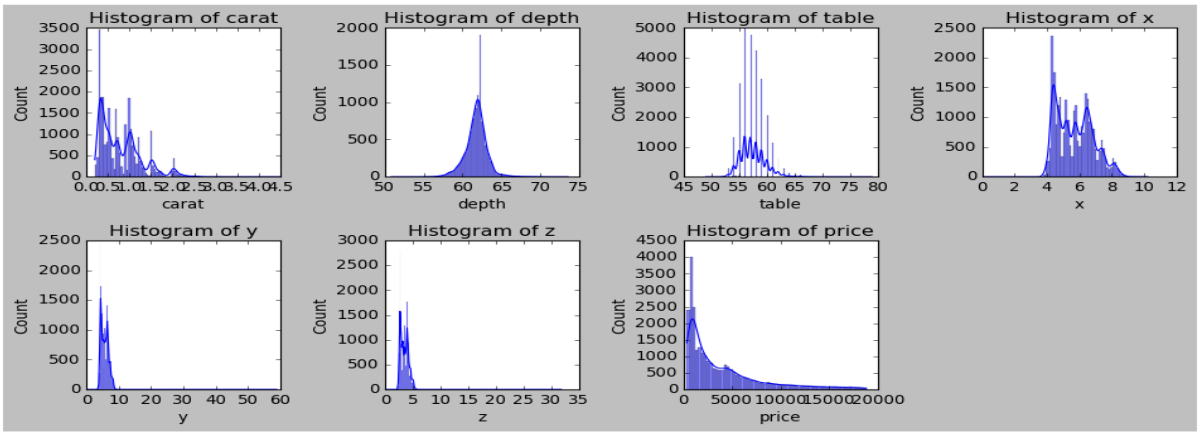

Figure 2: Histogram of All Numeric Columns

From the above plot we observe that data in carat and price columns are rightly skewed. Data in other columns are normally distributed.

Let plot the Count Plot for all object columns of the dataset.

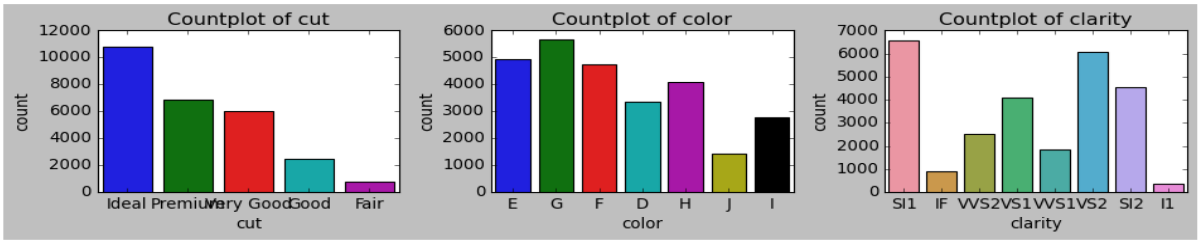

Figure 3: Count Plot of Object Columns

Bi-Variate Analysis

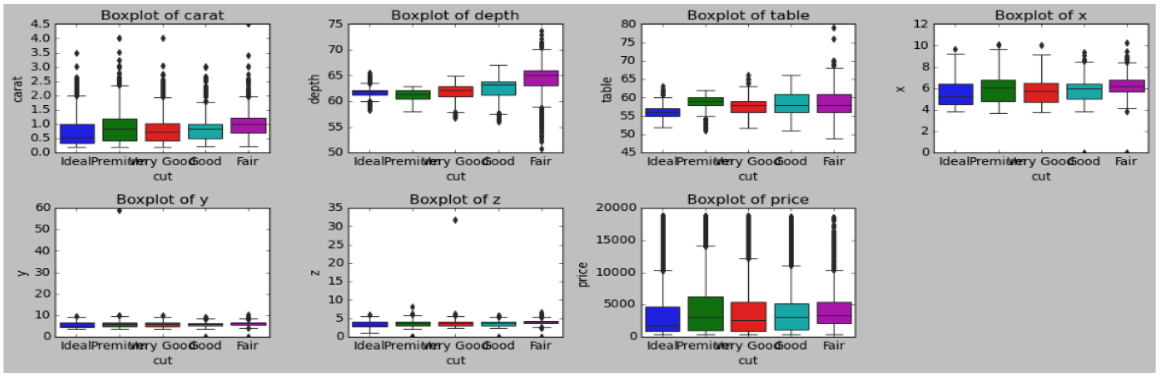

Figure 4: Bivariate Analysis Cut Verses All Numeric Columns

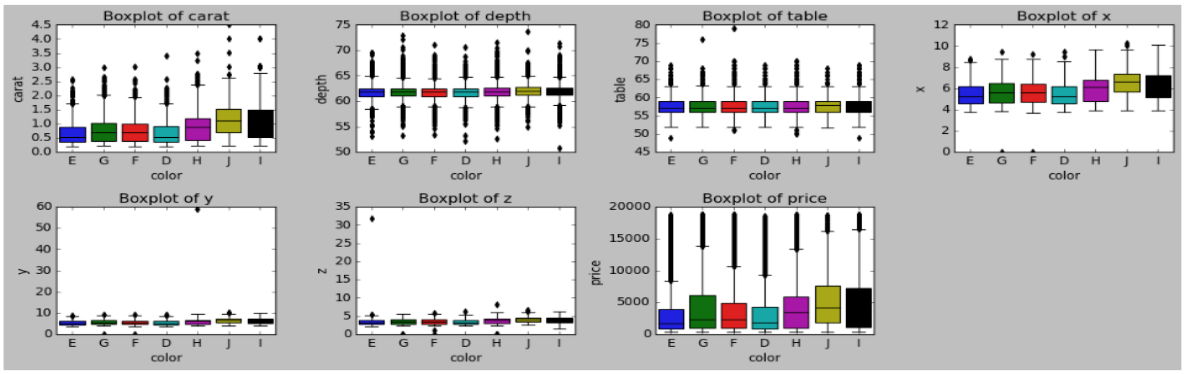

Figure 5: Bivariate Analysis Color Verses All Numeric Columns

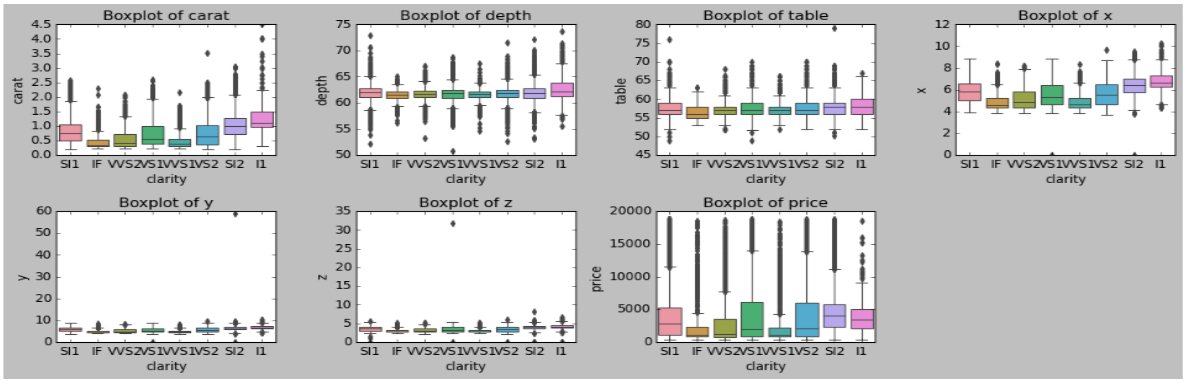

Figure 6: Bivariate Analysis Clarity Verses All Numeric Columns

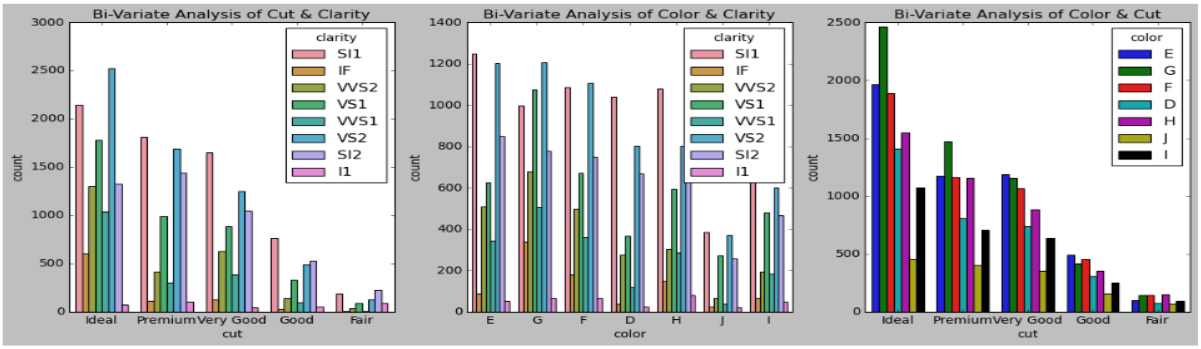

Figure 7: Bivariate Analysis of Categorical Verses Categorical Columns

Here we have plotted Categorical verses categorical and Categorical verses Continuous variables. Continuous verses continuous will be present in pair plot.

Multi-Variate Analysis

Pair-Plot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

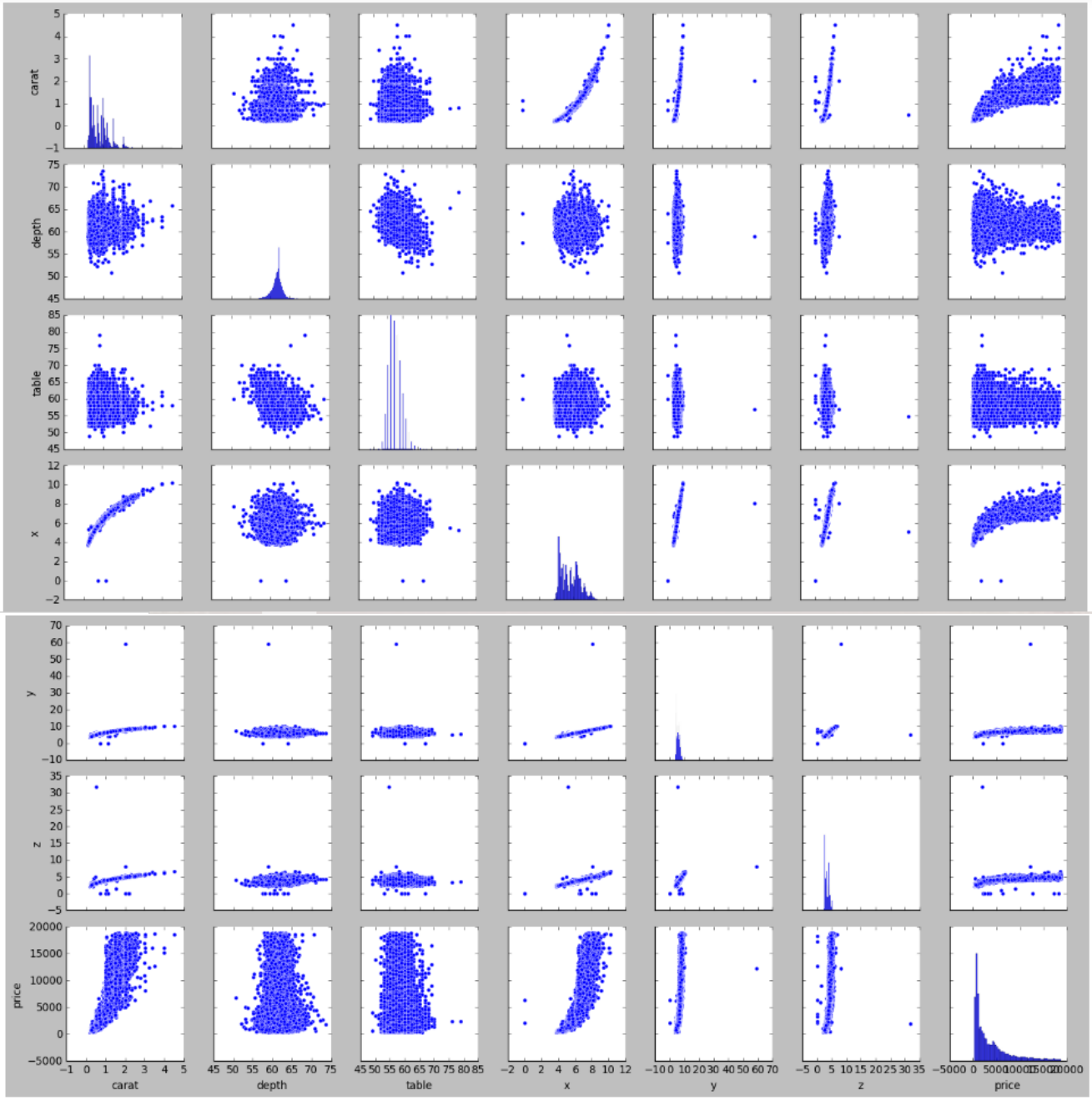

Figure 8: Pair Plot

Correlation-Plot

From the correlation plot, we can see the correlation among different variables. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

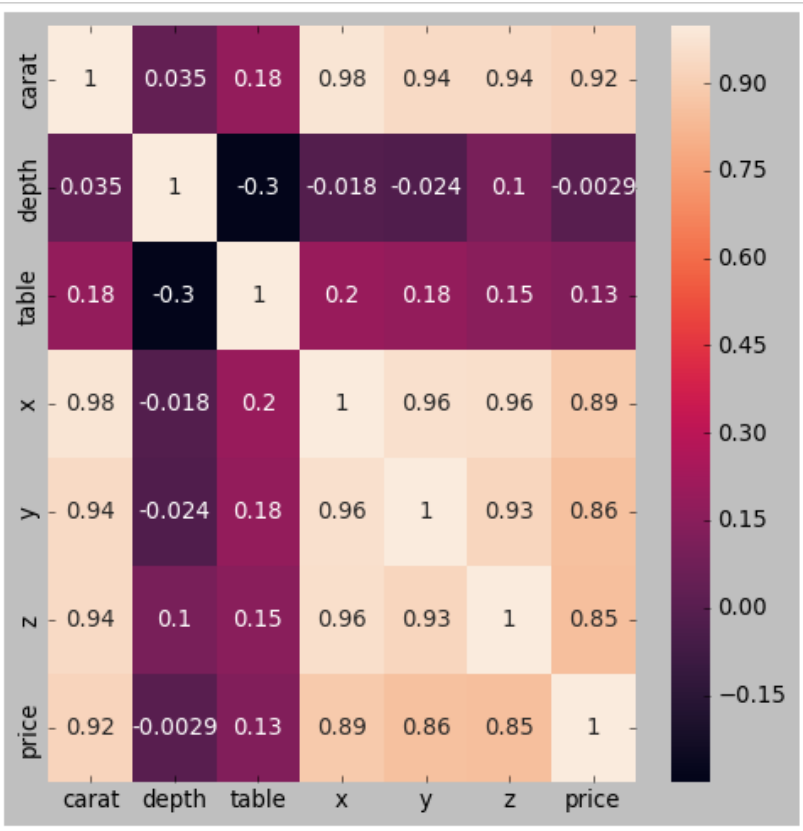

Figure 9: Correlation Plot

Descriptive Statistics

	carat	depth	table	x	y	z	price
count	26933.000000	26236.000000	26933.000000	26933.000000	26933.000000	26933.000000	26933.000000
mean	0.798010	61.745285	57.455950	5.729346	5.733102	3.537769	3937.526120
std	0.477237	1.412243	2.232156	1.127367	1.165037	0.719964	4022.551862
min	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	0.700000	61.800000	57.000000	5.690000	5.700000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5356.000000
max	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Table 2: Descriptive Statistics

Observations from above Exploratory Data Analysis:

- Price is dependent variable and other columns are Independent/Predictor Variables.
- Unnamed: 0 is a column with just serial numbers, since it doesn't have any importance, we drop it.
- Null Values are present in the depth column, need to be handled.
- Outliers were present in the dataset and are being treated.
- Carat variable is highly positively correlated with x, y, z and price variables.
- Y variable is highly positively correlated with z and price variables.
- Z variable is highly positively correlated with price variable.

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.

It is observed that we have 697 null values in depth column. We need to impute these null values before model building activity. As a standard practice null values for continuous variable are imputed with either mean or median and null values for categorical variable is imputed mode value.

Here depth column is continuous variable so we impute null values with the mean value.

Out of 26933 rows we find 2 rows have x values as 0, 2 rows have y value as zero and 8 rows have z value as 0. Since x, y and z represent length, width and height of a cubic zirconia in mm, having 0 as value for any of these columns represent a faulty value so we drop those rows. After dropping the faulty rows, we have total 26925 rows.

The categorical columns present in the dataset are cut, color and clarity, we can combine the data based on these categorical columns. Let us calculate the minimum, mean and maximum values for each this sub-category present in categorical variables.

	price		
	min	mean	max
cut			
Fair	369	4565.768935	18574
Good	335	3927.074774	18707
Ideal	326	3454.820639	18804
Premium	326	4540.186192	18795
Very Good	336	4032.267961	18818

Table 3: Descriptive stats for different Sub Categories of Cut Column

Observations:

- Cubic zirconia with
  - Very good cut has maximum price.
  - Fair cut has highest mean price
  - Ideal and premium cuts have lower prices.

	price		
	min	mean	max
color			
D	357	3184.827597	18526
E	326	3073.940399	18731
F	357	3700.277001	18791
G	361	4004.967434	18818
H	337	4469.778049	18795
I	336	5124.816637	18795
J	335	5329.706250	18701

Table 4: Descriptive stats for different Sub Categories of Color Column

Observations:

- Cubic zirconia with
  - Color G has maximum price.
  - Color J has highest mean price.
  - Color E have lowest price.

	price		
	min	mean	max
clarity			
I1	345	3915.013812	18531
IF	369	2739.534231	18552
SI1	326	3996.614564	18818
SI2	326	5088.169919	18804
VS1	338	3838.130201	18795
VS2	357	3963.159225	18791
VVS1	336	2502.874388	18445
VVS2	336	3263.042688	18718

Table 5: Descriptive stats for different Sub Categories of Clarity Column

Observations:

- Cubic zirconia with
  - Clarity Sl1 has maximum price.
  - Clarity Sl2 has highest mean price.
  - Clarity Sl1 & Sl2 have lowest prices.

While buying a Cubic zirconia few people give preference to type of cut, few give preference to color and few give preference to the clarity. So we have Sub divided the data into sub categories of each category to figure out price variation n each sub category.

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.

We have 3 object/string type columns in our dataset. For feeding the data into a model all the columns in the dataset should be of numeric type, so we encode the data before we proceed with model building activity.

We had observed outliers were present in our dataset. We will treat the outliers before we proceed with model building activity.

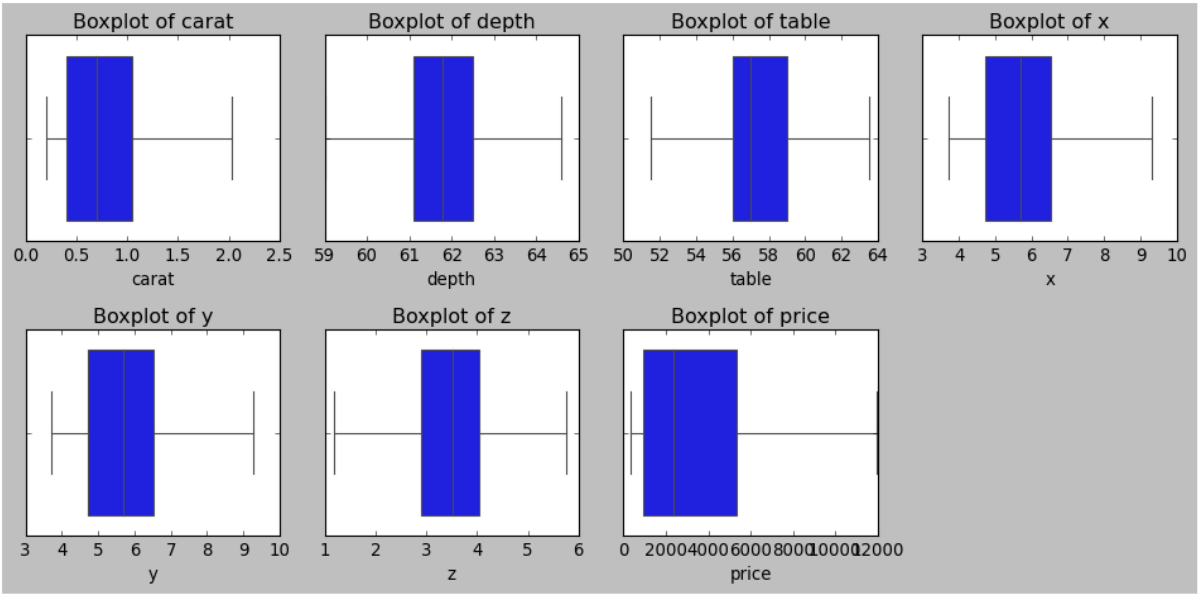

Figure 10: Box Plot of Numeric Columns after Treating Outliers:

Segregate the dependent variables and independent variable separately.

Using the train_test_split functionality present in sklearn.model_selection module we split the data into training set and testing set.

We build Linear Regression Model using LinearRegression present in sklearn.linear_model module.

After Building the Linear Regression Model following are the co-efficient values of each column.

Columns	Coefficient Value
Carat	8868.13
Cut	84.52
Color	274.06
Clarity	444.02
Depth	16.59
Table	-53.62
X	-1333.42
Y	1578.61
Z	-1012.56
Intercept	-1575.75

Table 6: Column Wise Coefficient Values

Y=mx +c (m= m1, m2, m3...m9) here 9 different co-efficients will learn align with the intercept which is "c" from the model.

From the above coefficients for each of the independent attributes we can conclude

- Every one unit increase in carat increases price by 8868.13
- Every one unit increase in cut increases price by 84.52
- Every one unit increase in color increases price by 274.06
- Every one unit increase in clarity increases price by 444.02
- Every one unit increase in y increases price by 1578.61
- Every one unit increase in depth increases price by 16.59

But

- Every one unit increase in table decreases price by 53.62
- Every one unit increase in x decreases price by 1333.42
- Every one unit increase in z decreases price by 1012.56

Model Efficiency

Model Efficiency	Training Data	Testing Data
R-squared	0.931	0.932
Adjusted R-squared	0.931	0.932
RMSE	908.9	909.4

Table 7: Model Evaluation Indexes

From the above table it is evident that model built has a better score of 0.93 for both training and testing set. Also, these is no much deviation between the values for training and testing dataset so the model is not overfit nor underfit.

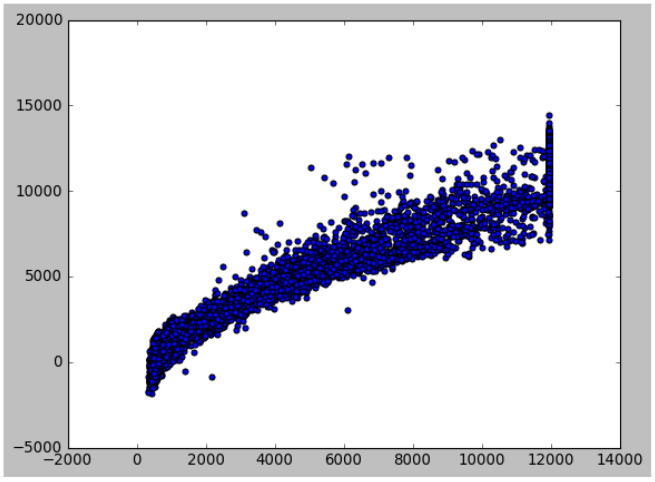

Table 8: Scatter Plot Predicted Price Verses Actual Price

When we try to build a model with scaled data, we observe that all the coefficients along with the Intercept value are also scaled and intercept value will be very less it can be neglected. Accuracy remains same.
The model built works similar for both test and train.

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

Let us use the statsmodels module to build Linear Regression model and get the Inferential Stats.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.931
Model:	OLS	Adj. R-squared:	0.931
Method:	Least Squares	F-statistic:	2.821e+04
Date:	Sat, 07 May 2022	Prob (F-statistic):	0.00
Time:	16:45:40	Log-Likelihood:	-1.5513e+05
No. Observations:	18847	AIC:	3.103e+05
Df Residuals:	18837	BIC:	3.104e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1575.7510	751.285	-2.097	0.036	-3048.338	-103.164
carat	8868.1294	82.798	107.106	0.000	8705.838	9030.421
cut	84.5206	6.780	12.467	0.000	71.232	97.810
color	274.0618	4.115	66.606	0.000	265.997	282.127
clarity	444.0157	4.463	99.480	0.000	435.267	452.764
depth	16.5922	10.951	1.515	0.130	-4.873	38.057
table	-53.6223	3.375	-15.889	0.000	-60.237	-47.008
x	-1333.4157	135.959	-9.808	0.000	-1599.907	-1066.925
y	1578.6099	133.976	11.783	0.000	1316.005	1841.215
z	-1012.5608	139.384	-7.265	0.000	-1285.766	-739.356

Omnibus:	2655.666	Durbin-Watson:	2.009
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9739.653
Skew:	0.686	Prob(JB):	0.00
Kurtosis:	6.244	Cond. No.	9.75e+03

Table 9: Inferential Statistics

From the above inferential statistics, we observe the following:

- Except depth column all other columns have p value less than 0.05, tis means except depth all other columns are strong predictors of price while depth column is a poor predictor of price.
- Price varies positively w.r.t carat, cut, color, clarity, y and depth.
- Price varies negatively w.r.t table, x and y.
- Out of all the predictors Carat has higher weightage and is most significant predictor.
- Also, color and clarity are other significant predictors of price.

Business Insights and Recommendations

- Cubic zirconia with higher carat goes for a higher price, so company must focus on them.
- Cubic zirconia with better clarity goes for a higher price, so company must focus on them.
- Cubic zirconia with better color goes for a higher price, so company must focus on them.
- Cubic zirconia with better y (width) goes for a higher price, so company must focus on them.
- Cubic zirconia with better cut goes for a higher price, so company must focus on them.

Various Steps followed

- Loaded the dataset into a pandas dataframe.
- Checked for duplicates and removed them.
- Checked for null values and imputed them with mean values.
- Look for faulty data and dropped them.
- Checked for outliers and treated them.
- Splitted the data into testing and training set.
- Built the linear Regression model using training set.
- Evaluated the model built using testing dataset.
- Evaluated the model by computing model evaluation values like R- squared, Ad-R squared and RMSEvalue.

Problem 2

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

Data Dictionary

- 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

Q 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.

Sample of the dataset

	Holliday_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 10: Sample Dataset

Dataset has 7 columns with 872 rows. Each row in the dataset corresponds to individual employee's detail.

Data Types

Let us check the datatypes of variables in dataframe.

```
Holliday_Package     object
Salary                int64
age                   int64
educ                  int64
no_young_children     int64
no_older_children     int64
foreign              object
```

There are total 7 columns, out of which 2 are of object type and remaining 5 are of integer type.

Duplicate check

There are no duplicates present in the dataset.

Null Check

```
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
```

From the above output we observe that there are no null values present in the dataset.

Check For Outliers

Let us plot the boxplot for all the numeric columns of the dataset.

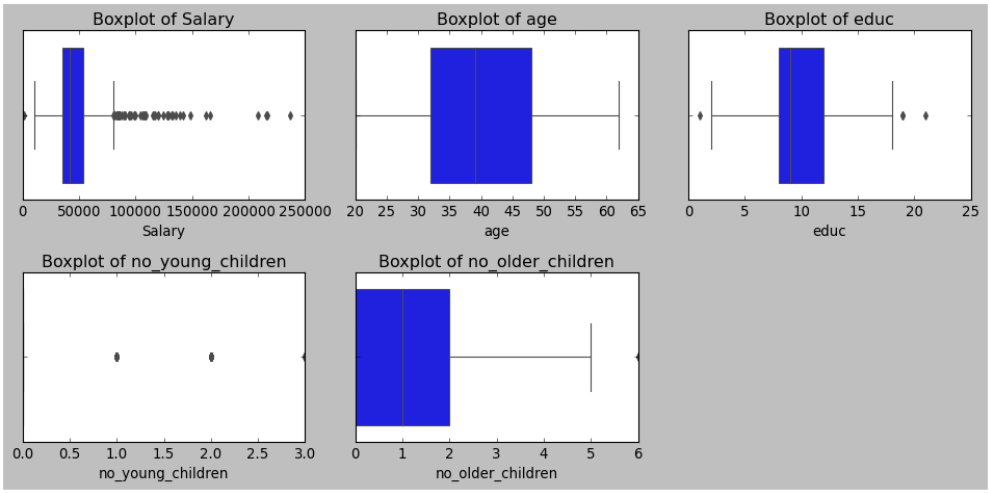

Figure 11: Boxplot of Numeric Columns

All the numeric columns present in the dataset have outliers.

Uni-Variate Analysis

Let us plot the histogram for all the numeric columns of the dataset.

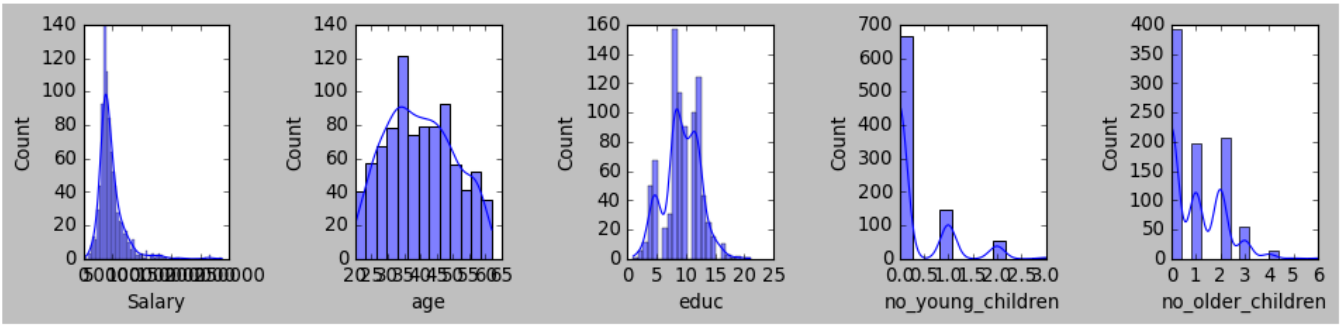

Figure 12: Histogram of Numeric Columns

From the above plot we can infer that data in Salary, Age and Educ are normally distributed. And data in number of older children is rightly skewed.

Let plot the Count Plot for all object columns of the dataset.

Figure 13: Count Plot of Object Columns

Bi- Variate Analysis

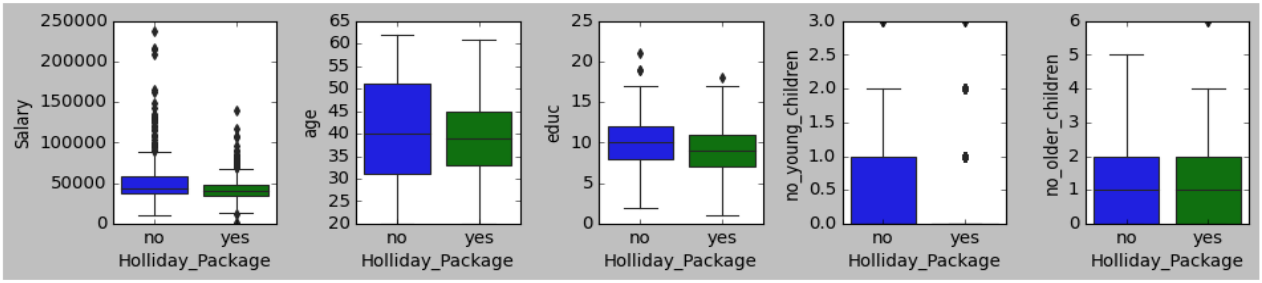

Figure 14: Bivariate Analysis Holiday Package Verses All Numeric Columns

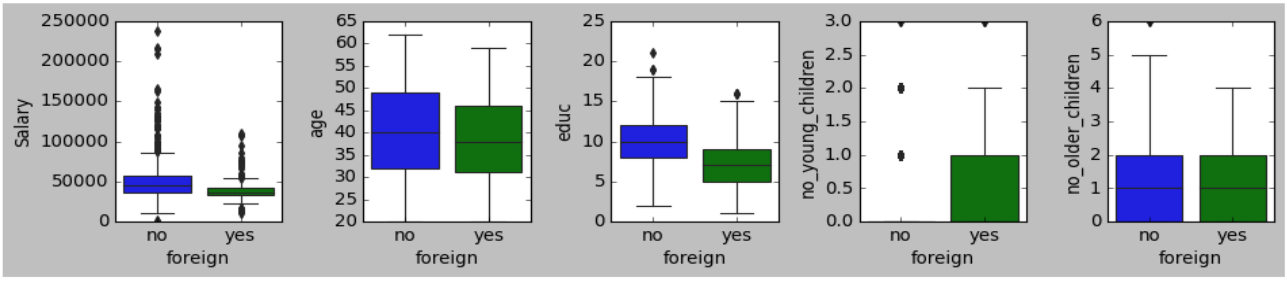

Figure 15: Bivariate Analysis Foreign Verses All Numeric Columns

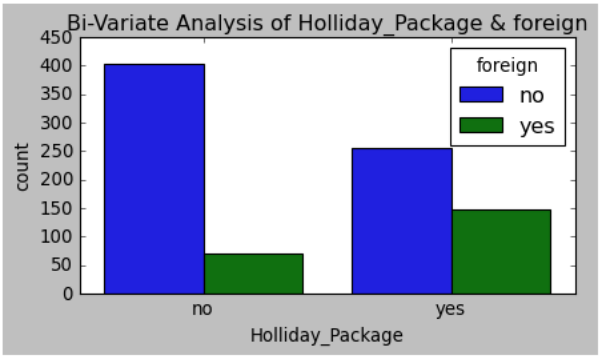

Figure 16: Bivariate Analysis of Holliday Package verses Foreign

Here we have plotted Categorical verses categorical and Categorical verses Continuous variables. Continuous verses continuous will be present in pair plot.

Multi- Variate Analysis

Pair Plot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

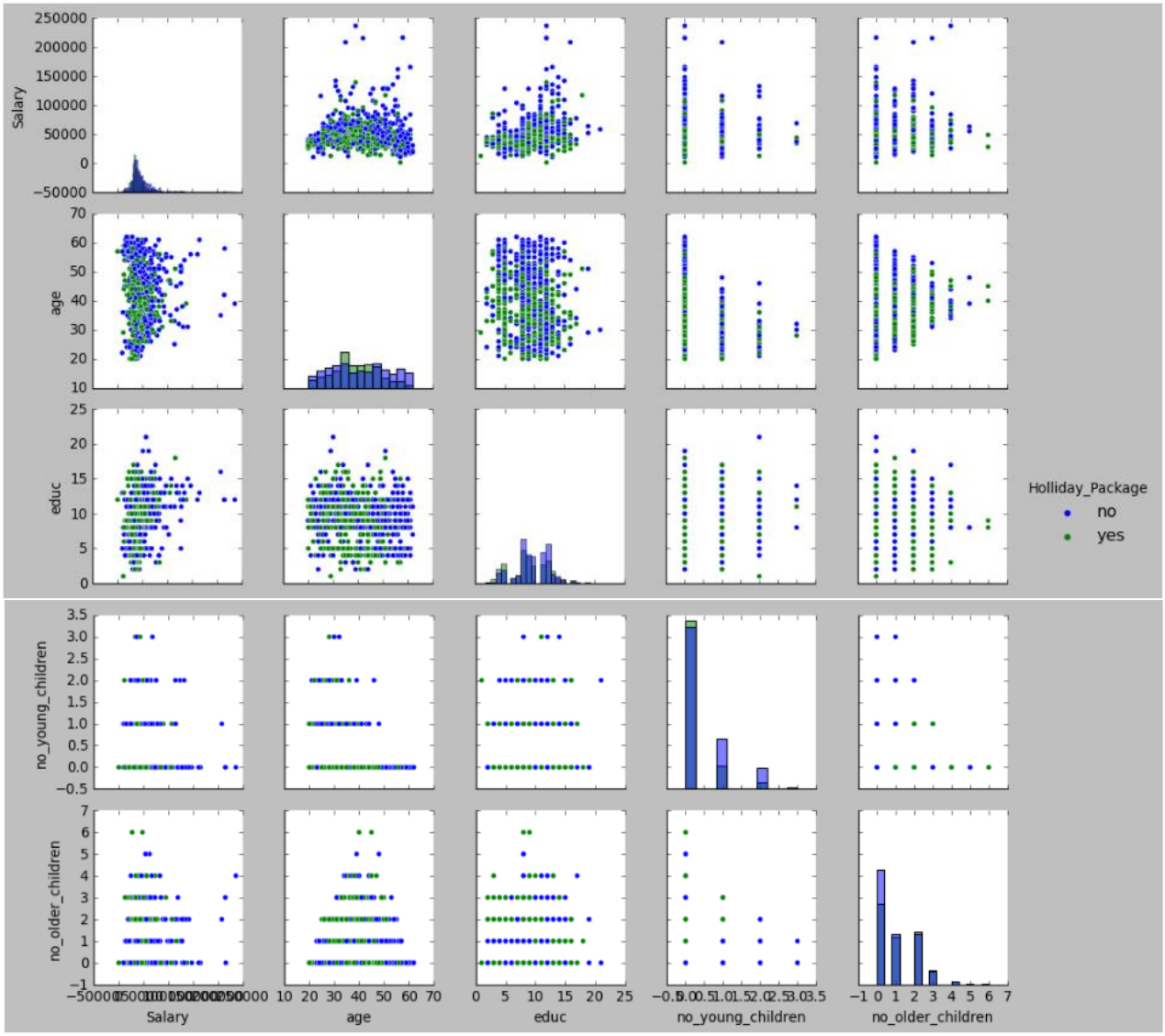

Figure 17: Pair Plot

Correlation Plot

From the correlation plot, we can see the correlation among different variables. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

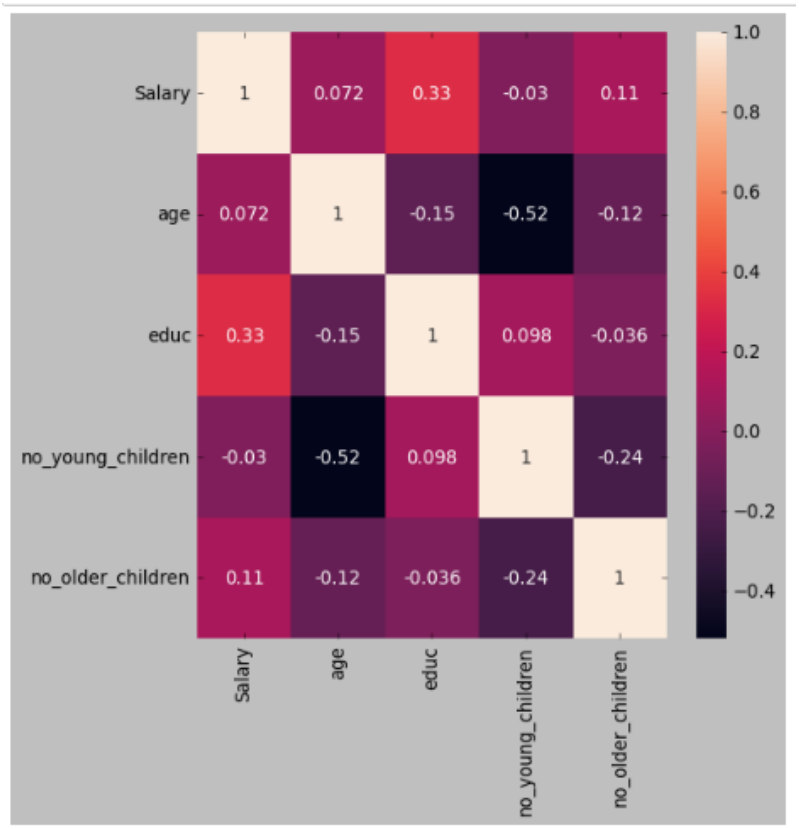

Figure 18: Correlation Plot

Observations from above Exploratory Data Analysis:

- Holliday Package is a dependent variable and other columns are Predictors.
- Unnamed: 0 is a column with just serial numbers, since it doesn't have any importance, we drop it.
- There were no Null Values and duplicates present in the dataset.
- Outliers are present in all the numeric columns.
- Correlation among the columns is very less.

Descriptive Statistics

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

Table 11: Descriptive Statistics

Q 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).

We have 2 object/string type columns in our dataset. For feeding the data into a model all the columns in the dataset should be of numeric type, so we encode the data before we proceed with model building activity.

Segregate the dependent variables and independent variable separately.

Using the train_test_split functionality present in sklearn.model_selection module we split the data into training set and testing set.

We build Logistic Regression Model using LogisticRegression present in sklearn.linear_model module.

We build the Linear Discriminant Analysis model using LinearDiscriminantAnalysis present in sklearn.discriminant_analysis module.

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.

Confusion Matrix for Logistic Regression Training Set

```
[[294  32]
 [261  23]]
```

Classification Report for Logistic Regression Training Set

	precision	recall	f1-score	support
0	0.53	0.90	0.67	326
1	0.42	0.08	0.14	284
accuracy			0.52	610
macro avg	0.47	0.49	0.40	610
weighted avg	0.48	0.52	0.42	610

Table 12: Classification Report of LR Model Training Set

Confusion Matrix for Logistic Regression Testing Set

```
[[129  16]
 [107  10]]
```

Classification Report for Logistic Regression Testing Set

	precision	recall	f1-score	support
0	0.55	0.89	0.68	145
1	0.38	0.09	0.14	117
accuracy			0.53	262
macro avg	0.47	0.49	0.41	262
weighted avg	0.47	0.53	0.44	262

Table 13: Classification Report of LR Model Testing Set

Confusion Matrix for LDA Training Set

```
[[252  74]
 [126 158]]
```

Classification Report for LDA Training Set

	precision	recall	f1-score	support
0	0.67	0.77	0.72	326
1	0.68	0.56	0.61	284
accuracy			0.67	610
macro avg	0.67	0.66	0.66	610
weighted avg	0.67	0.67	0.67	610

Table 14: Classification Report of LDA Model Training Set

Confusion Matrix for LDA Testing Set

```
[[104  41]
 [ 52  65]]
```

Classification Report for LDA Testing Set

	precision	recall	f1-score	support
0	0.67	0.72	0.69	145
1	0.61	0.56	0.58	117
accuracy			0.65	262
macro avg	0.64	0.64	0.64	262
weighted avg	0.64	0.65	0.64	262

Table 15: Classification Report of LDA Model Testing Set

Logistic Regression Model ROC Curves

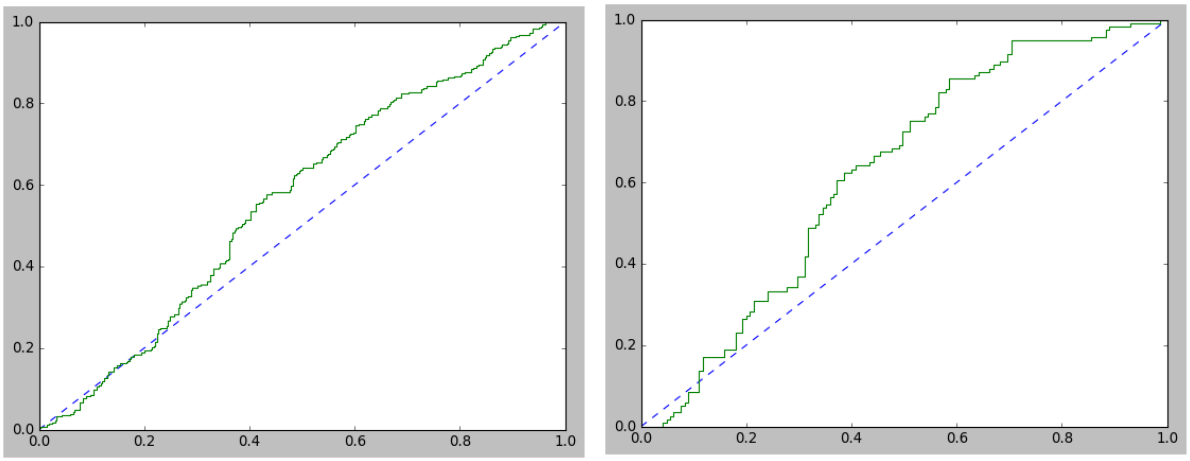

Figure 19: LR Model Training Set ROC Curve & Testing Set ROC Curve

Linear Discriminant Analysis Model ROC Curves

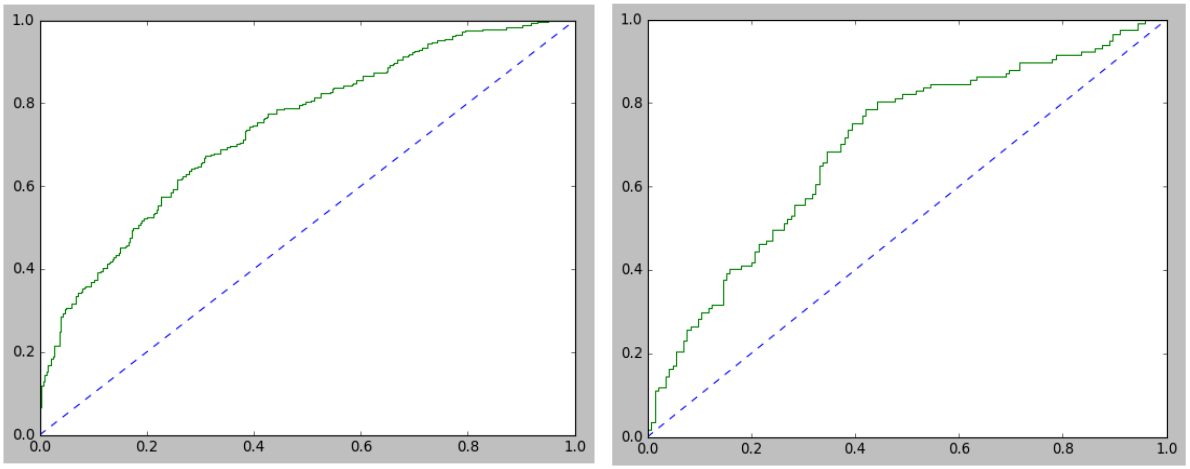

Figure 20: LDA Model Training Set ROC Curve & Testing Set ROC Curve

Summary

	Training Set	Testing Set
LR Model - AUC	0.57	0.63
LR Model - Accuracy	0.52	0.53
LDA Model - AUC	0.74	0.74
LDA Model - Accuracy	0.67	0.65

Table 16: Summary of LR and LDA Models

From the above Figures and tables, we observe:

- Area Under the curve is more for Linear Discriminant Analysis Model as compared to Linear Regression Model.
- Accuracy is more for Linear Discriminant Analysis Model as compared to Linear Regression Model.
- Also, Precision & Recall values are better for Linear Discriminant Analysis Model as compared to Linear Regression Model.

Hence, we can conclude **Linear Discriminant Analysis Model is better compared to Linear Regression Model** for given dataset.

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

Inferences from Logistic Regression:

- **Precision (59%) For {Label 0}:** 59 % of Employees who did not opt Holiday Package are correctly predicted, out of all Employees who did not opt Holiday Package that are predicted.
- **Recall (89%) For {Label 0}:** Out of all Employees who actually did not opt for holiday package, 89% of Employees who did not opt for holiday package have been predicted correctly.
- **Precision (38%) For {Label 1}:** 38 % of Employees who opted Holiday Package are correctly predicted, out of all Employees who opted Holiday Package that are predicted.
- **Recall (9%) For {Label 1}:** Out of all Employees who actually opted for holiday package, 9% of Employees who opted for holiday package have been predicted correctly.

Inferences from Linear Discriminant Analysis:

- **Precision (67%) For {Label 0}:** 67 % of Employees who did not opt Holiday Package are correctly predicted, out of all Employees who did not opt Holiday Package that are predicted.
- **Recall (72%) For {Label 0}:** Out of all Employees who actually did not opt for holiday package, 72% of Employees who did not opt for holiday package have been predicted correctly.
- **Precision (61%) For {Label 1}:** 61 % of Employees who opted Holiday Package are correctly predicted, out of all Employees who opted Holiday Package that are predicted.
- **Recall (56%) For {Label 1}:** Out of all Employees who actually opted for holiday package, 56% of Employees who opted for holiday package have been predicted correctly.

Various Steps followed

- Loaded the dataset into a pandas dataframe.
- Checked for duplicates.
- Checked for null values.
- Checked for outliers.
- Splitted the data into testing and training set.
- Built the logistic Regression model using training set.
- Built the linear discriminant analysis model using training set.
- Evaluated both the models built using testing dataset.
- Compared both the models and concluded the best model.

Thank You