Yash Joshi

29-08-2021

Predictive Modeling - **Project Report**

Problem 1: Linear Regression:

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.

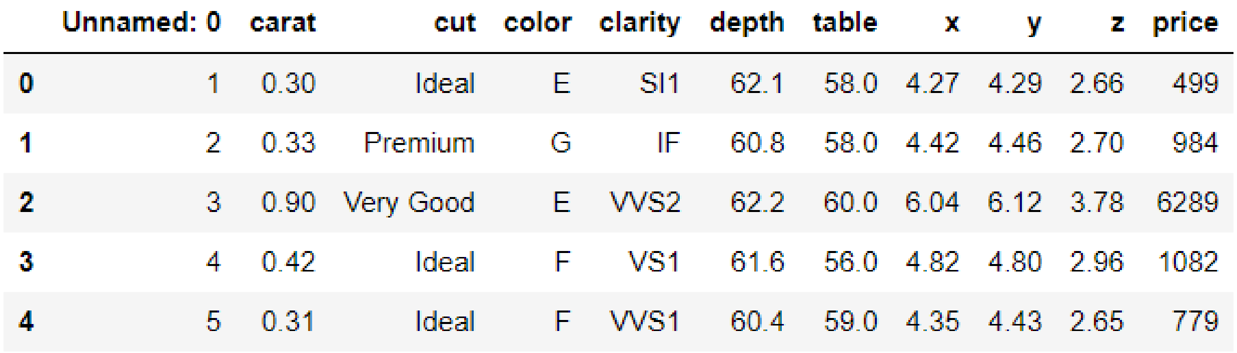
Data Dictionary:

Variable Name	Description
Carat	Carat weight of the cubic zirconia.
Cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
Color	Colour of the cubic zirconia.With D being the worst and J the best.
Clarity	cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, IF = 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, divided by its average Girdle Diameter.
Table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
Price	the Price of the cubic zirconia.
X	Length of the cubic zirconia in mm.
Y	Width of the cubic zirconia in mm.
Z	Height of the cubic zirconia in mm.

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.

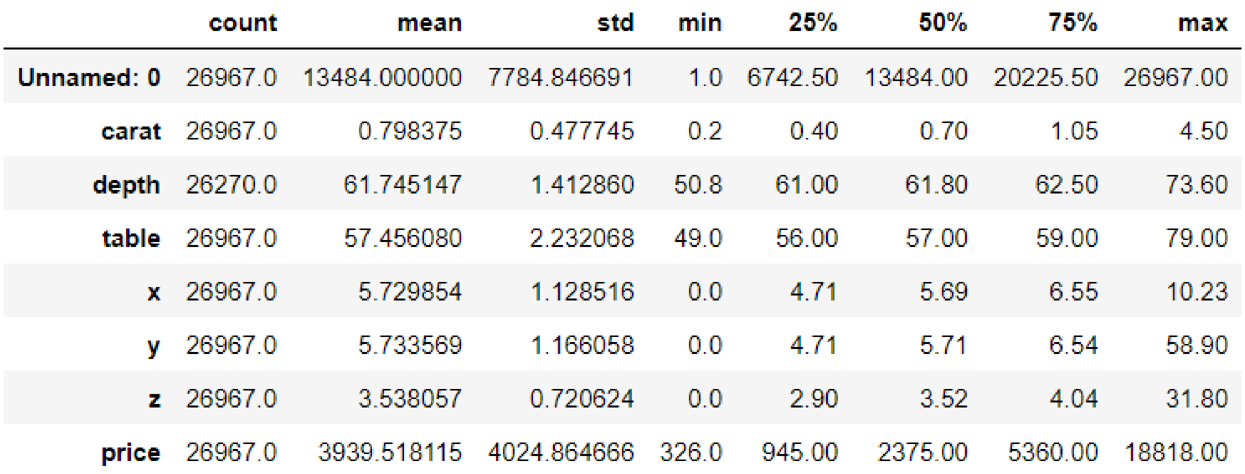
Ans 1.1

Looking at the top 5 records:



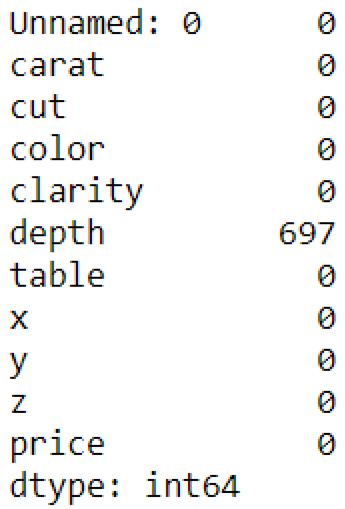
	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35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

Describing the data:



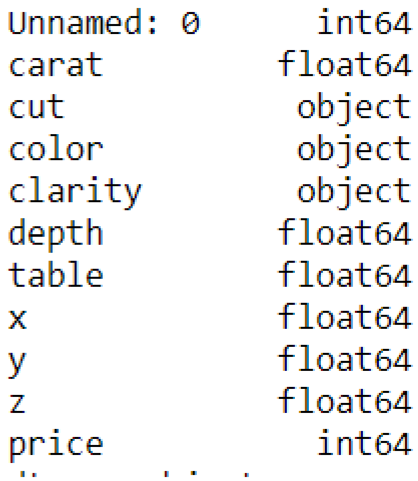
	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	26967.0	13484.000000	7784.846691	1.0	6742.50	13484.00	20225.50	26967.00
carat	26967.0	0.798375	0.477745	0.2	0.40	0.70	1.05	4.50
depth	26270.0	61.745147	1.412860	50.8	61.00	61.80	62.50	73.60
table	26967.0	57.456080	2.232068	49.0	56.00	57.00	59.00	79.00
x	26967.0	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

Checking Null Values:



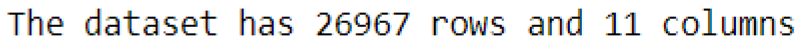
```
Unnamed: 0      0
carat           0
cut             0
color           0
clarity         0
depth         697
table           0
x               0
y               0
z               0
price           0
dtype: int64
```

Checking Data Types:



```
Unnamed: 0      int64
carat         float64
cut            object
color          object
clarity        object
depth         float64
table         float64
x             float64
y             float64
z             float64
price           int64
```

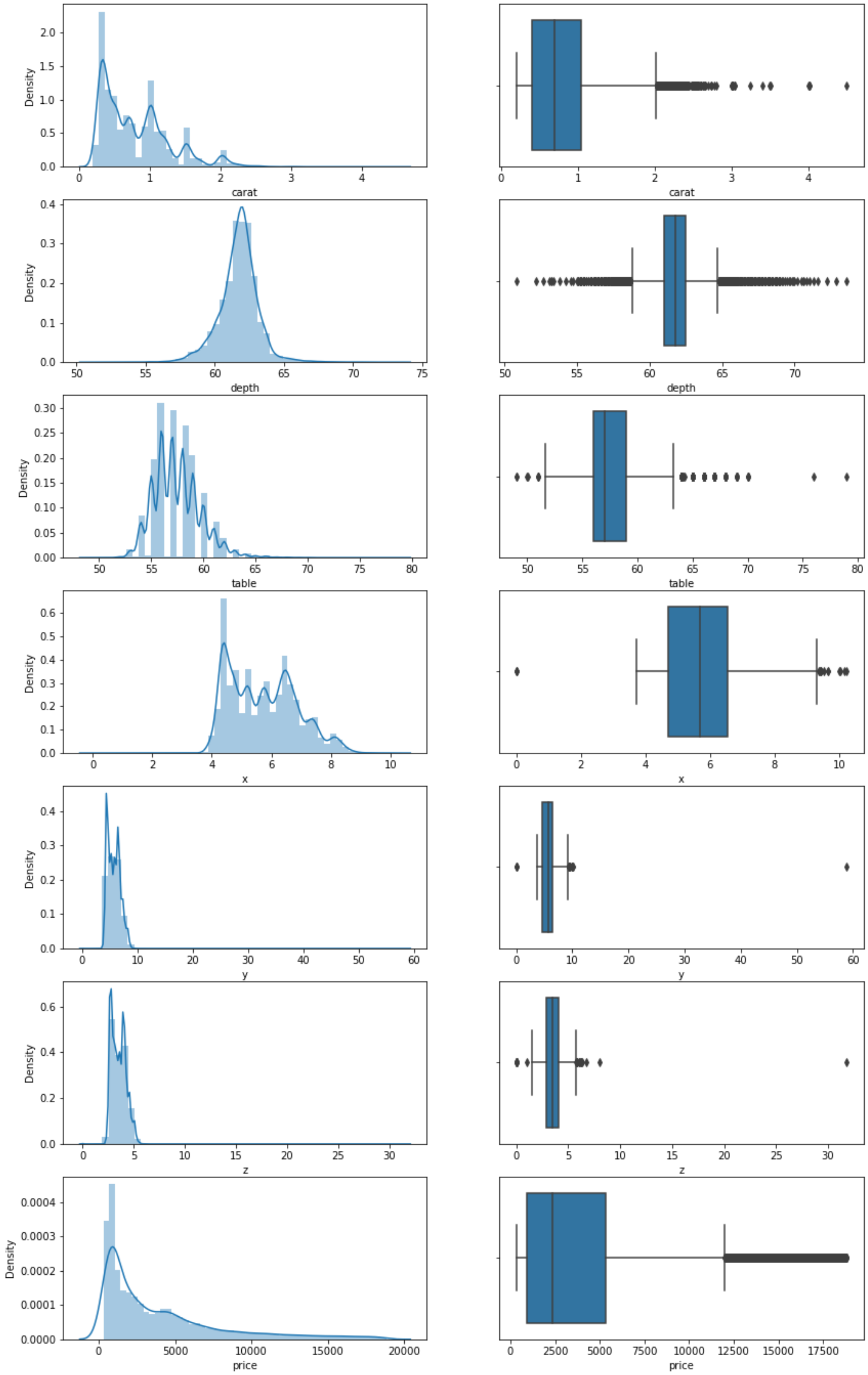
Checking the Dataset shape:



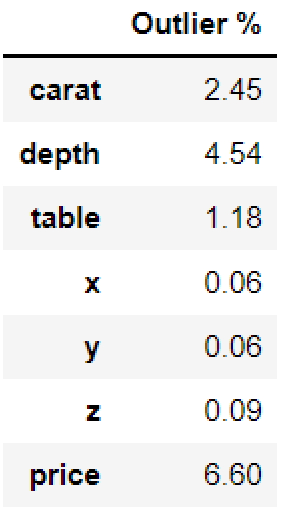
```
The dataset has 26967 rows and 11 columns
```

Univariate Analysis:

Box Plot & Distribution Plot:



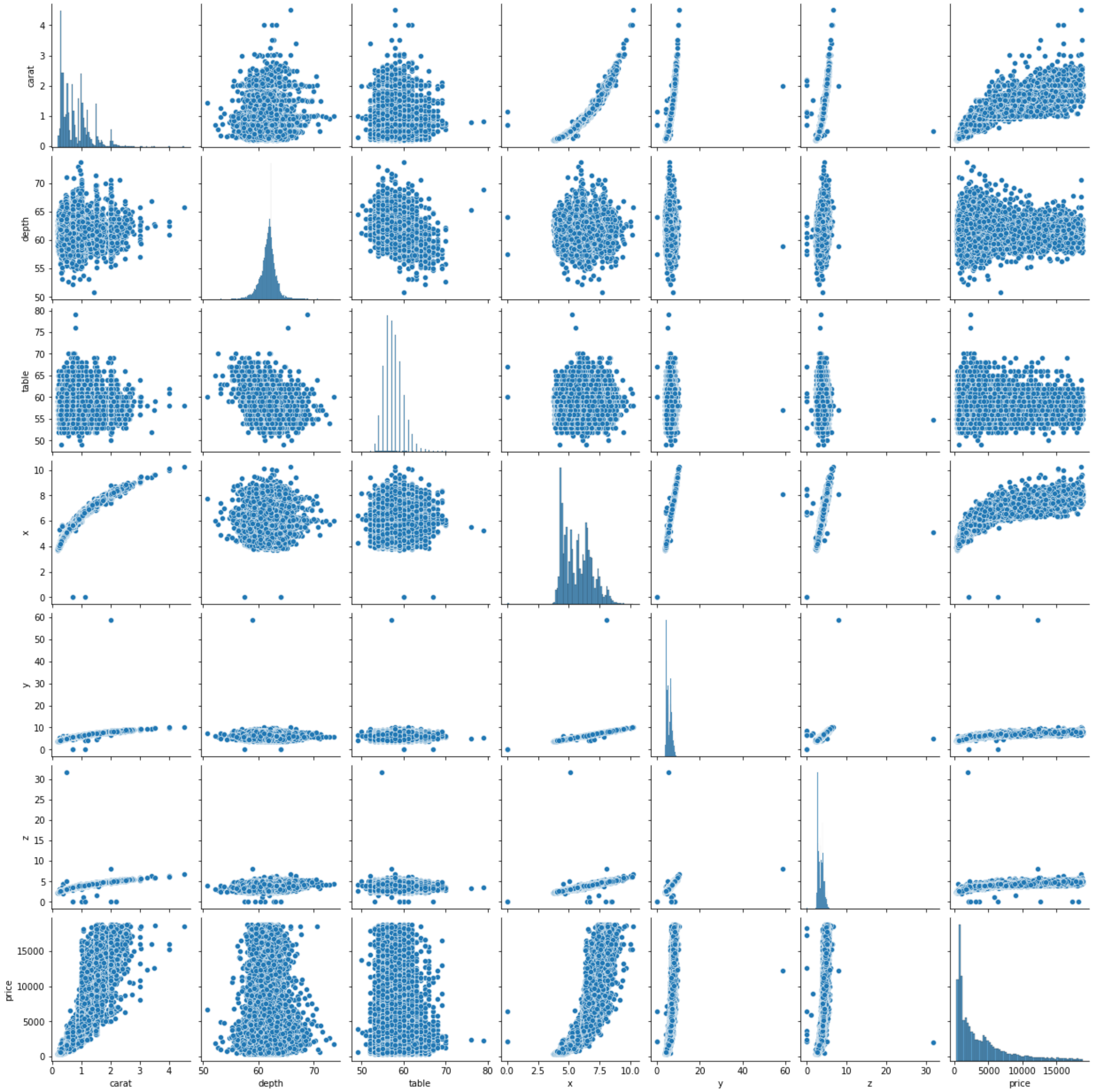
Percentage of Outliers:



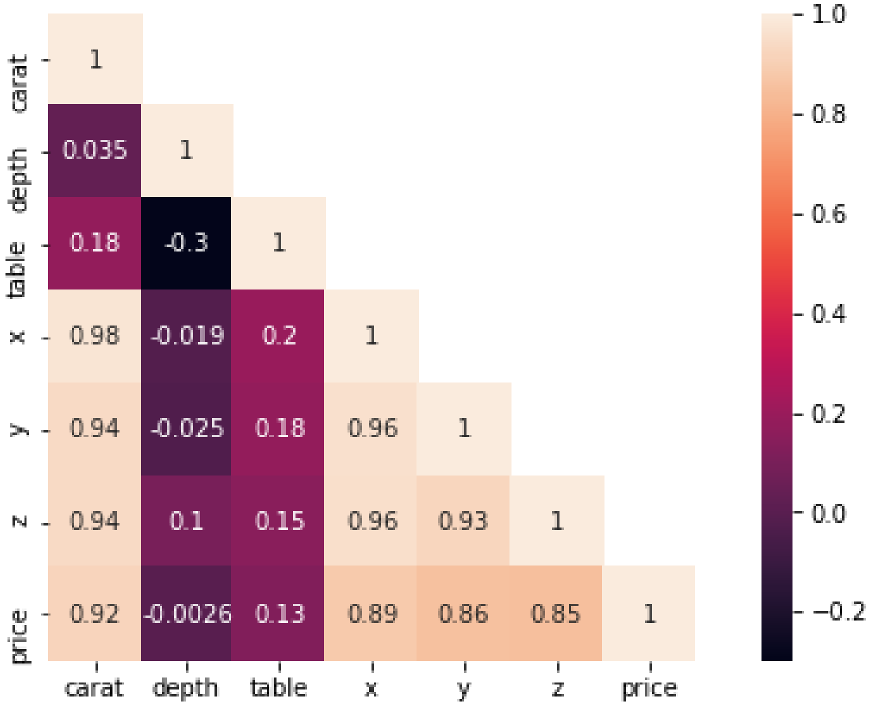
	Outlier %
carat	2.45
depth	4.54
table	1.18
x	0.06
y	0.06
z	0.09
price	6.60

Bi-Variate Analysis:

Pairplot:



Correlation Plot:



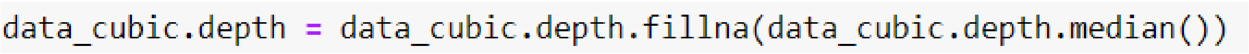
1.1 Inferences:

- The dataset has 9 independent variables and 1 dependent variable.
- Out of the 9 independent variables - 2 are categorical while the rest 7 are numerical.
- The columns - Depth and Table have much higher median while compared to other columns.
- The column - Depth has a few Null Values.
- The Dataset has around 27000 rows and 11 columns.
- 4 columns have % outliers less than 1.5%.
- The variable Carat is highly correlated with X, Y and Z.
- X is highly correlated with Y and Z.

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? Do you think scaling is necessary in this case?

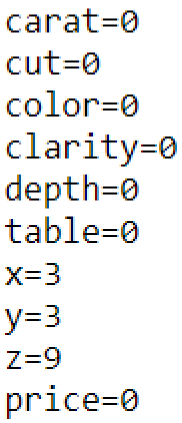
Ans 1.2

Replacing the Null values in the Depth Column:



```
data_cubic.depth = data_cubic.depth.fillna(data_cubic.depth.median())
```

Checking for 0 values:



```
carat=0
cut=0
color=0
clarity=0
depth=0
table=0
x=3
y=3
z=9
price=0
```

1.2 Inferences:

Yes, the Zero values **should be dropped**, the reason being that the dimensions X, Y and Z can't be zero.

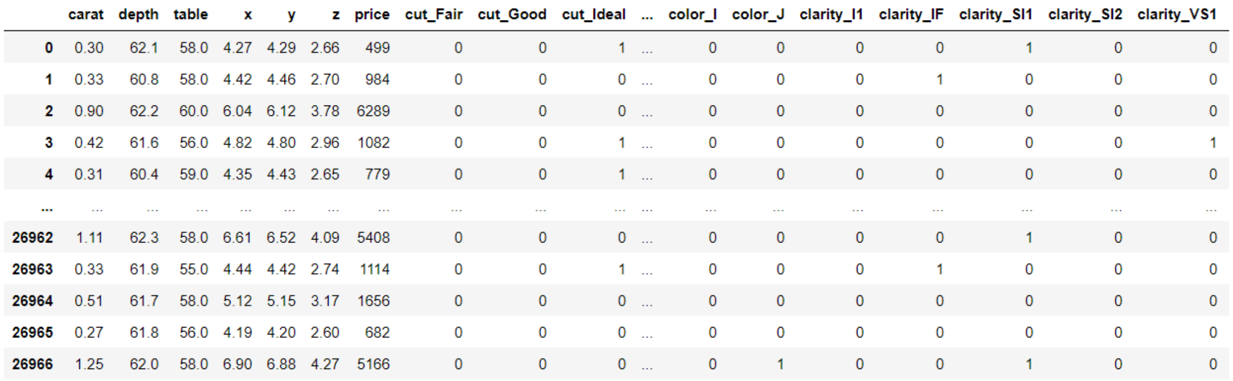
Yes, Scaling **is necessary** in this case as the values vary a lot in magnitude. The columns - Depth and Table have much higher median while compared to other columns.

Scaling of the data is done at a later stage in the model building phase using z score.

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 RSquare, RMSE.

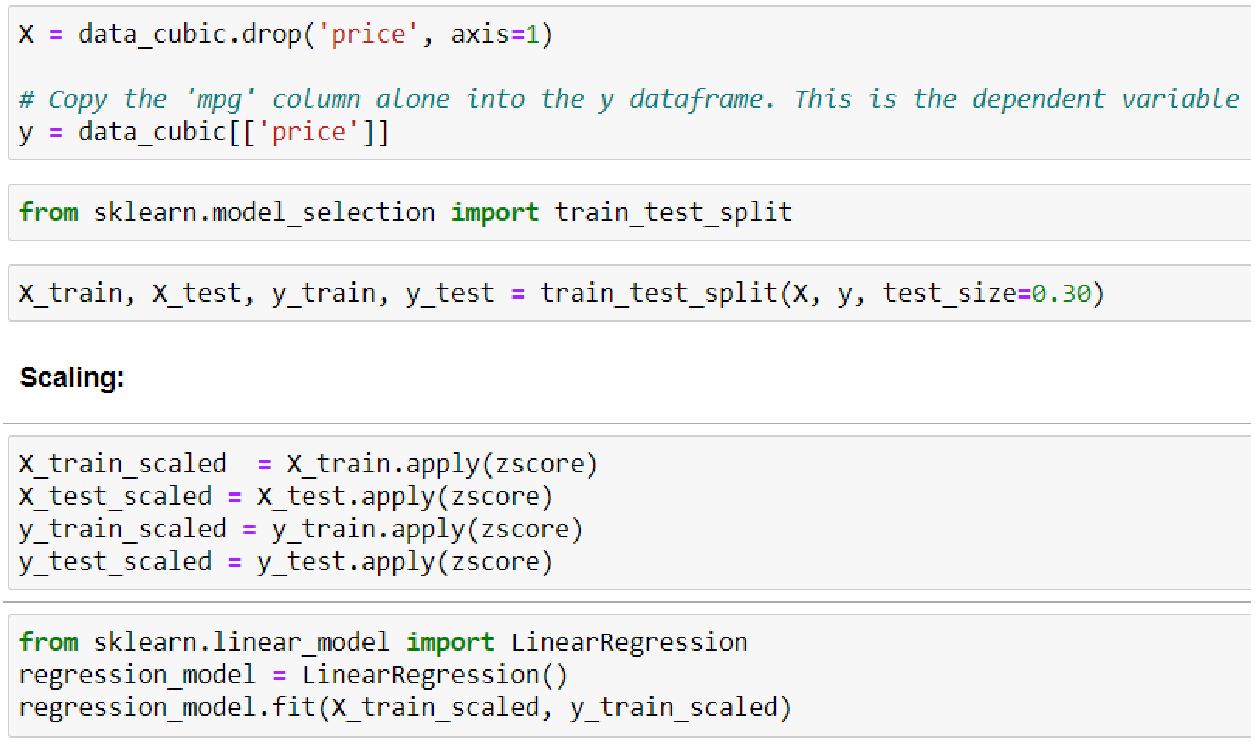
Ans 1.3

Encoding of the 3 categorical columns - cut, color and clarity is done using pd.get_dummies function.



	carat	depth	table	x	y	z	price	cut_Fair	cut_Good	cut_Ideal	...	color_I	color_J	clarity_I1	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1
0	0.30	62.1	58.0	4.27	4.29	2.66	499	0	0	1	...	0	0	0	0	1	0	0
1	0.33	60.8	58.0	4.42	4.46	2.70	984	0	0	0	...	0	0	0	1	0	0	0
2	0.90	62.2	60.0	6.04	6.12	3.78	6289	0	0	0	...	0	0	0	0	0	0	0
3	0.42	61.6	56.0	4.82	4.80	2.96	1082	0	0	1	...	0	0	0	0	0	0	1
4	0.31	60.4	59.0	4.35	4.43	2.65	779	0	0	1	...	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26962	1.11	62.3	58.0	6.61	6.52	4.09	5408	0	0	0	...	0	0	0	0	1	0	0
26963	0.33	61.9	55.0	4.44	4.42	2.74	1114	0	0	1	...	0	0	0	1	0	0	0
26964	0.51	61.7	58.0	5.12	5.15	3.17	1656	0	0	0	...	0	0	0	0	0	0	0
26965	0.27	61.8	56.0	4.19	4.20	2.60	682	0	0	0	...	0	0	0	0	0	0	0
26966	1.25	62.0	58.0	6.90	6.88	4.27	5166	0	0	0	...	0	1	0	0	1	0	0

The data is split using train_test_split function, the values are scaled, and the model is fit:



```
X = data_cubic.drop('price', axis=1)

# Copy the 'mpg' column alone into the y dataframe. This is the dependent variable
y = data_cubic[['price']]
```

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30)
```

Scaling:

```
X_train_scaled  = X_train.apply(zscore)
X_test_scaled = X_test.apply(zscore)
y_train_scaled = y_train.apply(zscore)
y_test_scaled = y_test.apply(zscore)
```

```
from sklearn.linear_model import LinearRegression
regression_model = LinearRegression()
regression_model.fit(X_train_scaled, y_train_scaled)
```

Model Performance:

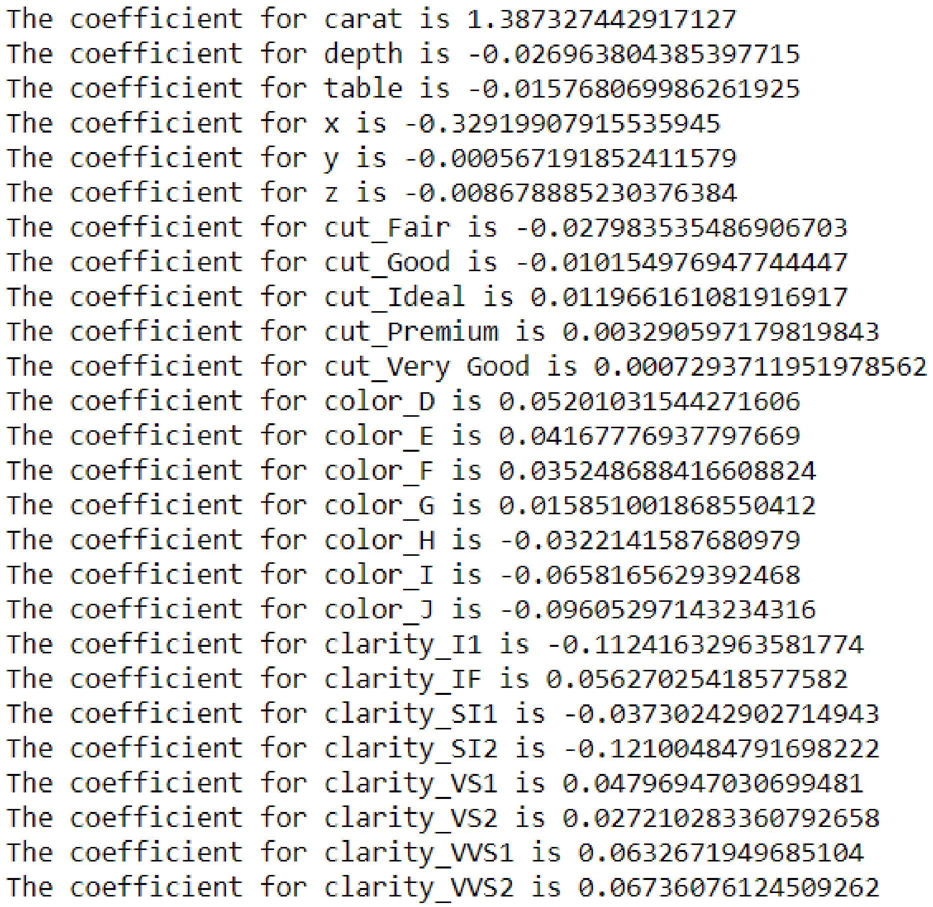
Training Data:

- R Squared: 0.922
- RMSE: 0.277

Test Data:

- R Squared: 0.917
- RMSE: 0.287

Feature Importance:



```
The coefficient for carat is 1.387327442917127
The coefficient for depth is -0.026963804385397715
The coefficient for table is -0.015768069986261925
The coefficient for x is -0.32919907915535945
The coefficient for y is -0.000567191852411579
The coefficient for z is -0.008678885230376384
The coefficient for cut_Fair is -0.027983535486906703
The coefficient for cut_Good is -0.010154976947744447
The coefficient for cut_Ideal is 0.011966161081916917
The coefficient for cut_Premium is 0.003290597179819843
The coefficient for cut_Very Good is 0.0007293711951978562
The coefficient for color_D is 0.05201031544271606
The coefficient for color_E is 0.04167776937797669
The coefficient for color_F is 0.035248688416608824
The coefficient for color_G is 0.015851001868550412
The coefficient for color_H is -0.0322141587680979
The coefficient for color_I is -0.0658165629392468
The coefficient for color_J is -0.09605297143234316
The coefficient for clarity_I1 is -0.11241632963581774
The coefficient for clarity_IF is 0.05627025418577582
The coefficient for clarity_SI1 is -0.03730242902714943
The coefficient for clarity_SI2 is -0.12100484791698222
The coefficient for clarity_VS1 is 0.04796947030699481
The coefficient for clarity_VS2 is 0.027210283360792658
The coefficient for clarity_VVS1 is 0.0632671949685104
The coefficient for clarity_VVS2 is 0.06736076124509262
```

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

- The features - **Carat, X, Clarity being - I1 or SI2 and the Color being J** are the most important attributes and they have the **most impact on the price**.
- Certain columns such as **X, Y and Z** have **zero values** which should be avoided in the future, these are **incorrect entries by the inventory manager**.
- 3 columns have **% outliers** more than 1.5%, for these columns - **Carat, Depth and Price**. It **could** be the case that a few values entered here are incorrect, a **sanity check** must be done to ensure that these columns have the right values.
- The Linear Regression Model fits the dataset well, Gem Stones co ltd can use this model to **maximize their profit share**.
- The various steps performed include - Reading Data, doing EDA, imputing Null and Zero values, encoding data, splitting data, applying model and drawing inference.

Problem 2: Logistic Regression and LDA:

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:

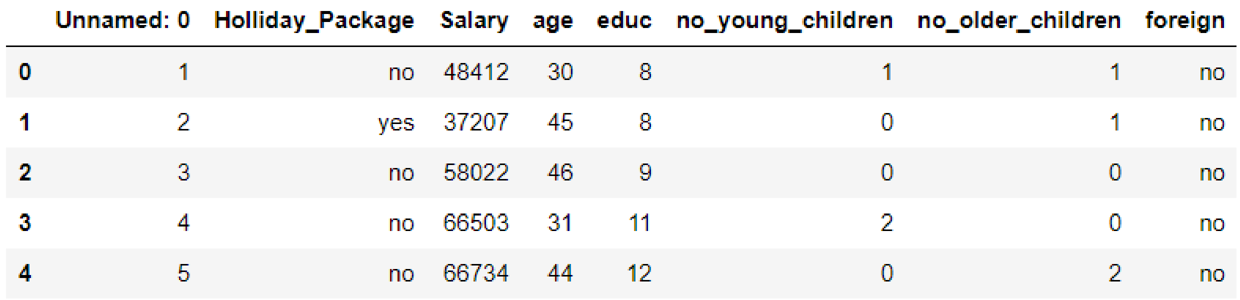
Variable Name	Description
Holiday_Package	Opted for Holiday Package yes/no?
Salary	Employee salary
age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children

foreign	foreigner Yes/No

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.

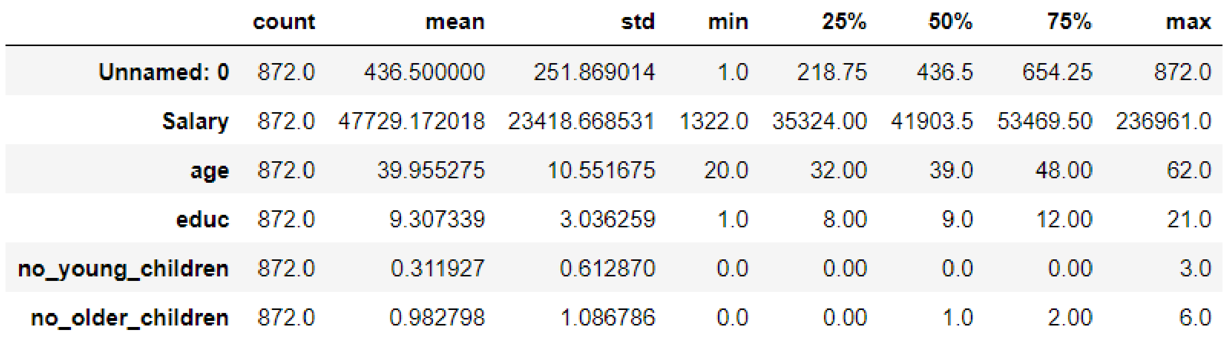
Ans 2.1

Displaying the first few rows:



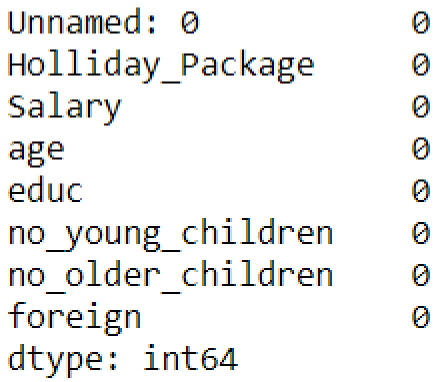
	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

Describing the data:



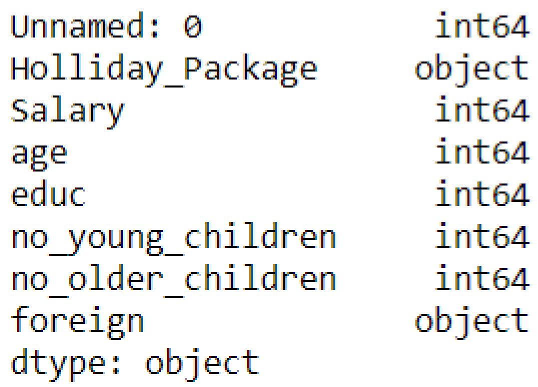
	count	mean	std	min	25%	50%	75%	max
Unnamed: 0	872.0	436.500000	251.869014	1.0	218.75	436.5	654.25	872.0
Salary	872.0	47729.172018	23418.668531	1322.0	35324.00	41903.5	53469.50	236961.0
age	872.0	39.955275	10.551675	20.0	32.00	39.0	48.00	62.0
educ	872.0	9.307339	3.036259	1.0	8.00	9.0	12.00	21.0
no_young_children	872.0	0.311927	0.612870	0.0	0.00	0.0	0.00	3.0
no_older_children	872.0	0.982798	1.086786	0.0	0.00	1.0	2.00	6.0

Checking Null Values:



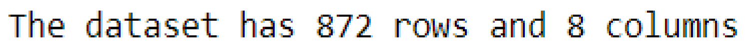
```
Unnamed: 0           0
Holliday_Package     0
Salary               0
age                  0
educ                 0
no_young_children    0
no_older_children    0
foreign              0
dtype: int64
```

Checking Data Types:



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

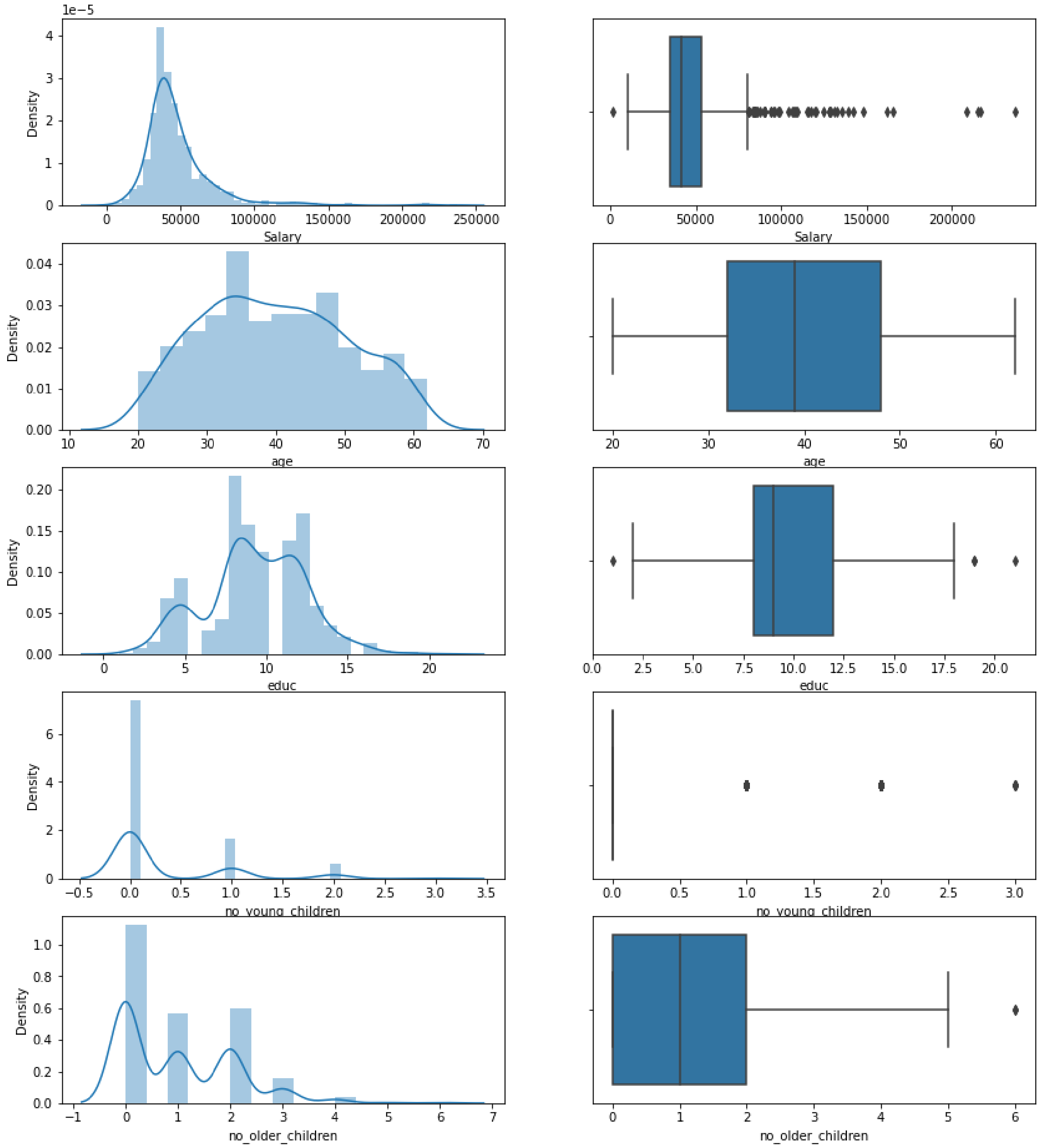
Checking the Dataset shape:



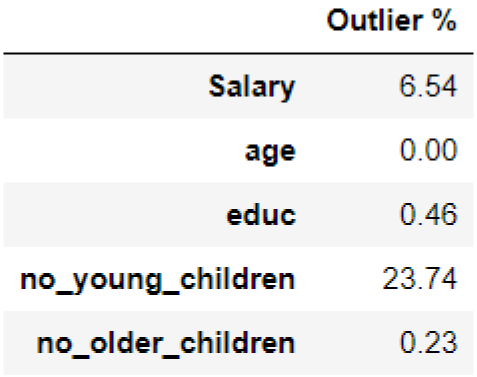
```
The dataset has 872 rows and 8 columns
```

Univariate Analysis:

Box Plot & Distribution Plot:



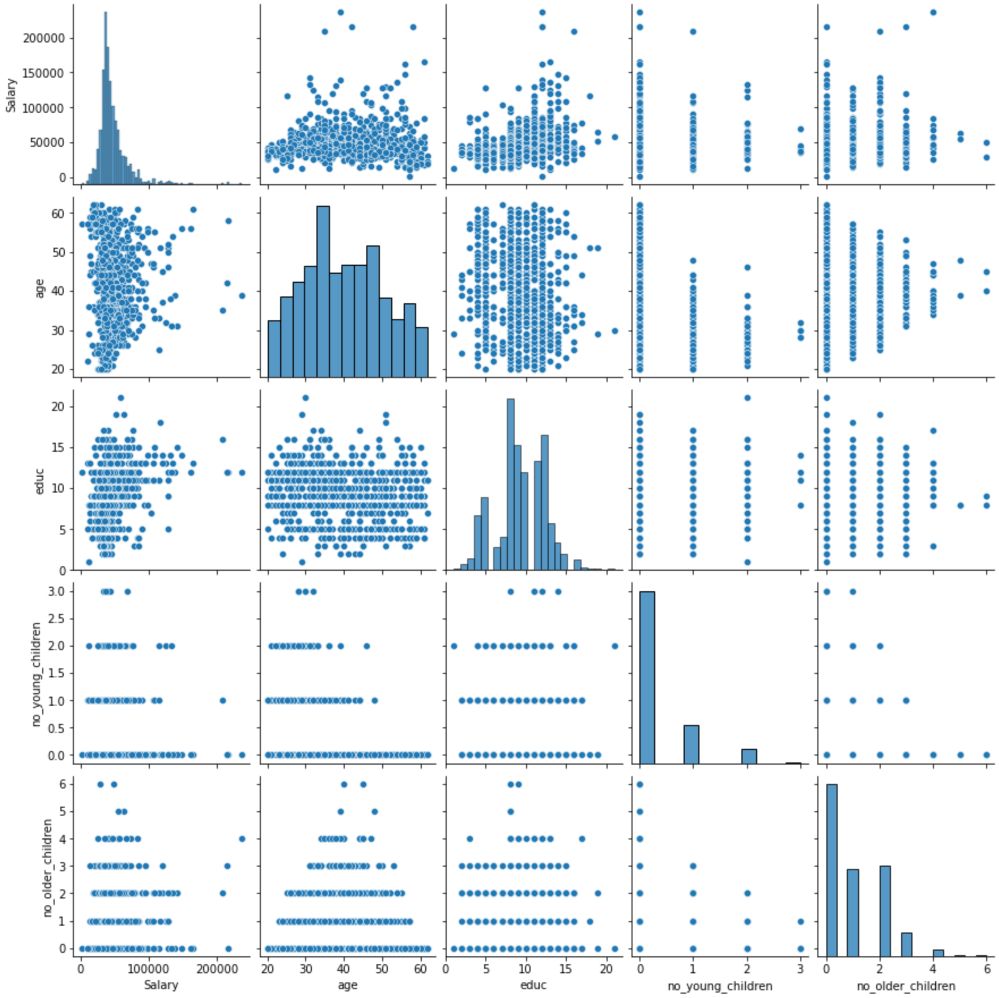
Percentage of Outliers:



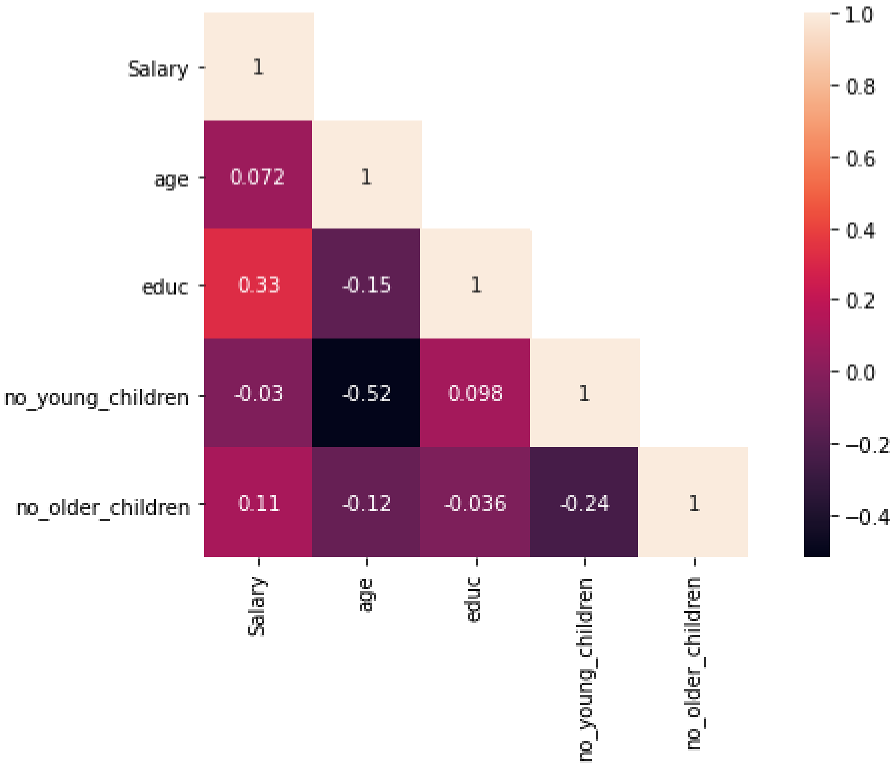
	Outlier %
Salary	6.54
age	0.00
educ	0.46
no_young_children	23.74
no_older_children	0.23

Bi-Variate Analysis:

Pairplot:



Correlation Plot:



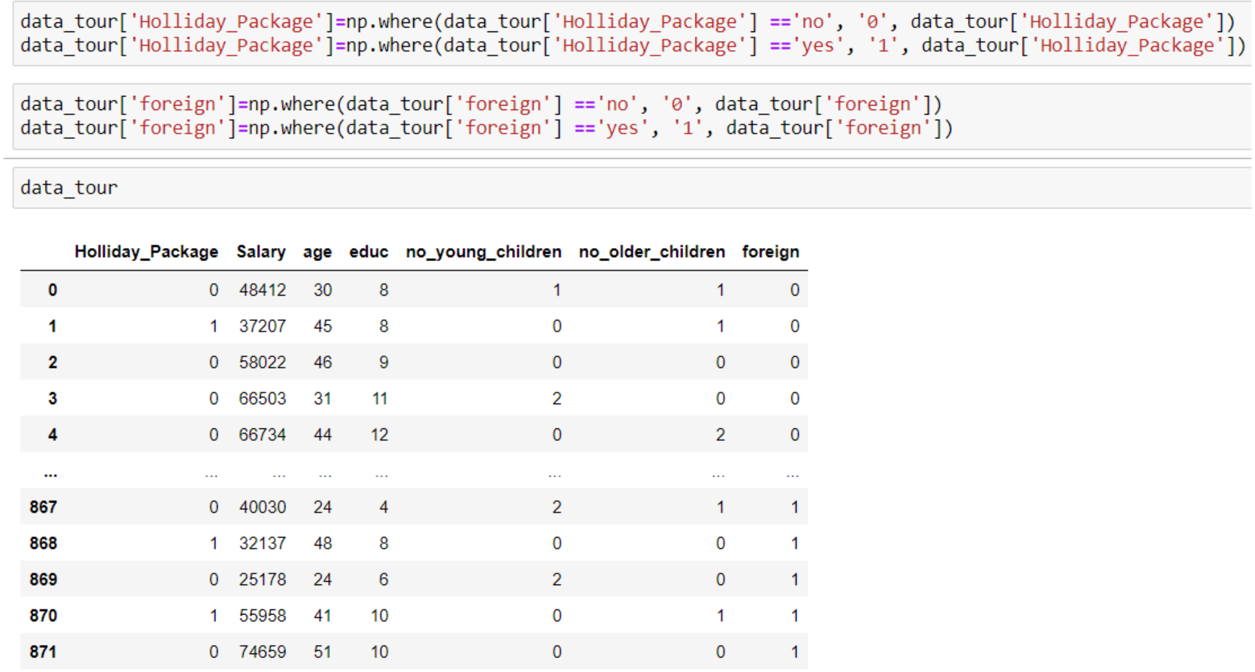
2.1 Inferences:

- The dataset has 6 independent variables and 1 dependent variable.
- Out of the 6 dependent variables - 1 is categorical, while the other 5 are numeric.
- The columns - Salary has a much higher median when compared to the rest of the columns.
- No column has Null values.
- The Dataset has 872 rows and 8 columns.
- The columns - no_young_children, Salary have % outliers more than 1.5%.
- Age has a slight correlation to no_young_children. Most columns don’t have any strong positive or negative correlation with the rest of the columns.

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).

Ans 2.2

Encoding the categorical columns - Holliday_Package, and foreign:



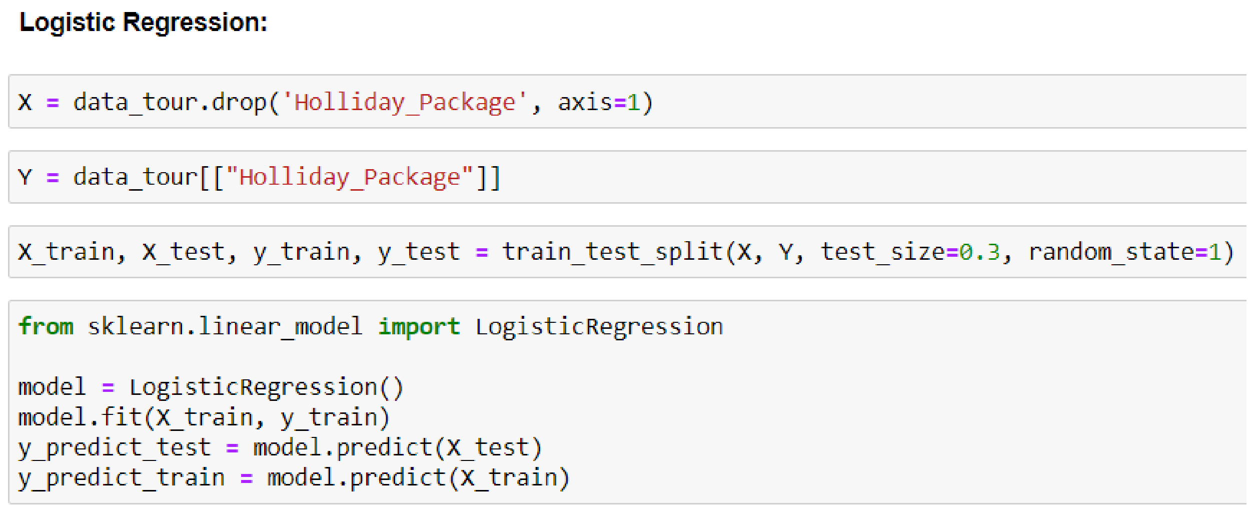
```
data_tour['Holliday_Package']=np.where(data_tour['Holliday_Package'] =='no', '0', data_tour['Holliday_Package'])
data_tour['Holliday_Package']=np.where(data_tour['Holliday_Package'] =='yes', '1', data_tour['Holliday_Package'])
```

```
data_tour['foreign']=np.where(data_tour['foreign'] =='no', '0', data_tour['foreign'])
data_tour['foreign']=np.where(data_tour['foreign'] =='yes', '1', data_tour['foreign'])
```

```
data_tour
```

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0
...	...	...	...	...	...	...	...
867	0	40030	24	4	2	1	1
868	1	32137	48	8	0	0	1
869	0	25178	24	6	2	0	1
870	1	55958	41	10	0	1	1
871	0	74659	51	10	0	0	1

Applying Logistic Regression:



Logistic Regression:

```
X = data_tour.drop('Holliday_Package', axis=1)
```

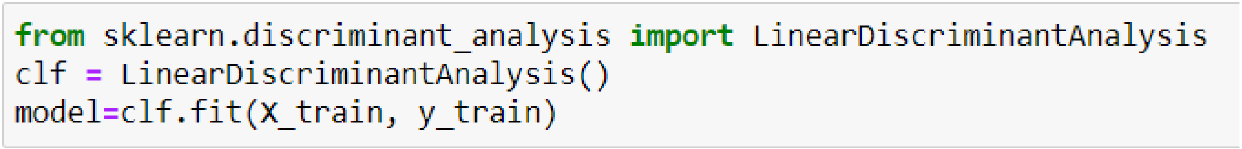
```
Y = data_tour[["Holliday_Package"]]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=1)
```

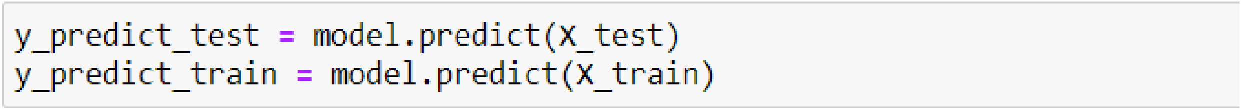
```
from sklearn.linear_model import LogisticRegression

model = LogisticRegression()
model.fit(X_train, y_train)
y_predict_test = model.predict(X_test)
y_predict_train = model.predict(X_train)
```

Applying Linear Discriminant Analysis:

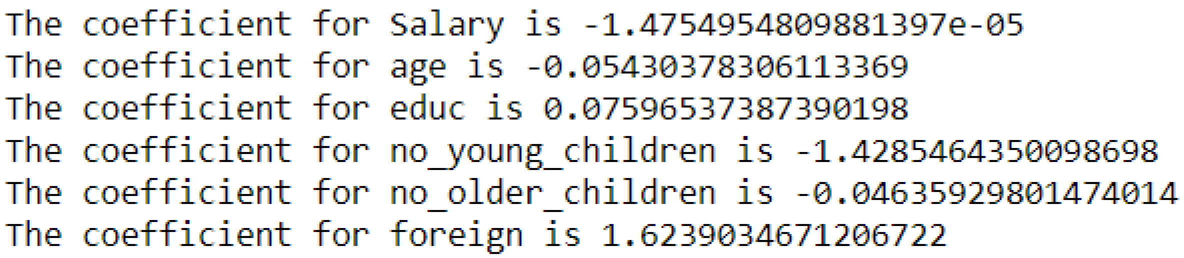


```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
clf = LinearDiscriminantAnalysis()
model=clf.fit(X_train, y_train)
```



```
y_predict_test = model.predict(X_test)
y_predict_train = model.predict(X_train)
```

Finding the most important features:

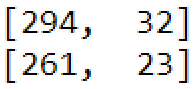
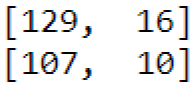
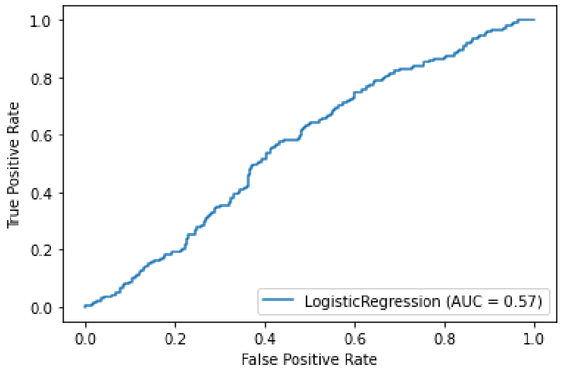
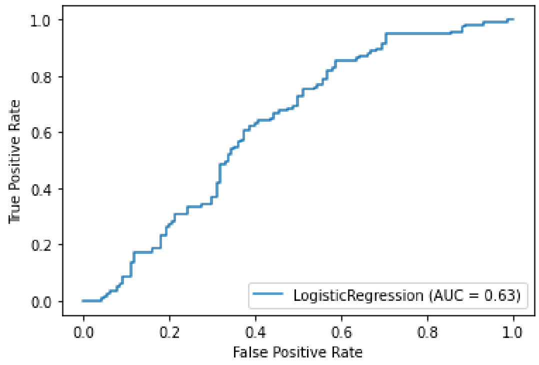


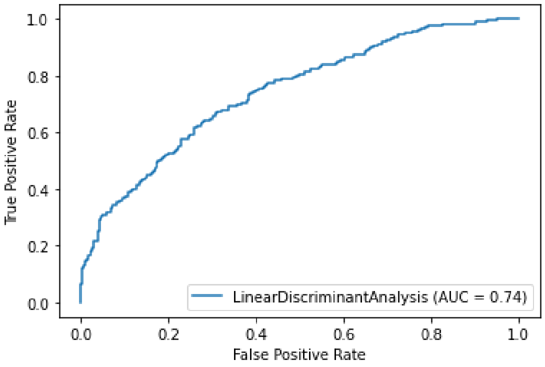
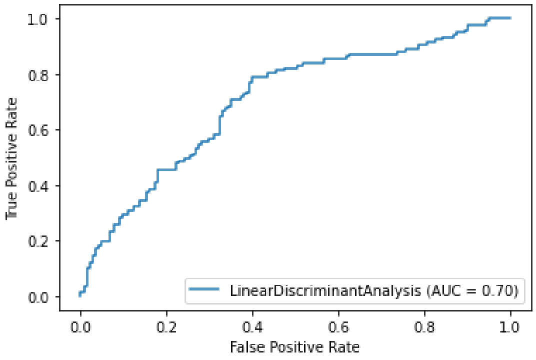
```
The coefficient for Salary is -1.4754954809881397e-05
The coefficient for age is -0.05430378306113369
The coefficient for educ is 0.07596537387390198
The coefficient for no_young_children is -1.4285464350098698
The coefficient for no_older_children is -0.04635929801474014
The coefficient for foreign is 1.6239034671206722
```

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.

Ans 2.3

Model Comparison:

	Logistic Regression	
	Train Dataset	**Test Dataset**
Accuracy	0.519	0.530
Confusion Matrix	 [294, 32] [261, 23]	 [129, 16] [107, 10]
ROC Curve:		
ROC_AUC score:	0.566	0.626

	Linear Discriminant Analysis	
	Train Dataset	**Test Dataset**
Accuracy	0.672	0.641
Confusion Matrix	[252, 74] [126, 158]	[103, 42] [52, 65]
ROC Curve:		
ROC_AUC score:	0.742	0.702

2.3 Inferences:

- Looking at the training and test performance metrics - It can be concluded that the **Linear Discriminant Analysis** model performs the best.
- The LDA model is better as it has higher accuracy and AUC values, which indicate better performance.

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

- The various steps performed include - Reading Data, doing EDA, imputing Null and Zero values, encoding data, splitting data, applying model and drawing inference.
- The features - **Salary, no_young_children, and Foreign** are the most important attributes and they have the **most impact on the price**.
- The columns - **no_young_children, Salary** have **% outliers more** than 1.5%. A sanity check must be done to ensure that these columns have the right values.
- The **Linear Discriminant Analysis** fits the dataset well, the tour and travel agency can use this model to predict if an employee will opt for a package or not.