PGPDSBA Online FEB A 2021

Pavan Kumar R Naik

PGP-DSBA Online

Feb A 2021

31/07/2021

Table of Contents

List of Figures

List of Tables

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 best and J the worst.
Clarity	cubic zirconia 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, 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.

Q1.1. 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.

Solution:

Sample of Dataset:

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

Table 1.1. Dataset Sample (cubic zirconia)

Summary of Dataset:

	Unnamed: 0	carat	depth	table	x	y	z	price
count	26967.000000	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	13484.000000	0.798375	61.745147	57.456080	5.729854	5.733569	3.538057	3939.518115
std	7784.846691	0.477745	1.412860	2.232068	1.128516	1.166058	0.720624	4024.864666
min	1.000000	0.200000	50.800000	49.000000	0.000000	0.000000	0.000000	326.000000
25%	6742.500000	0.400000	61.000000	56.000000	4.710000	4.710000	2.900000	945.000000
50%	13484.000000	0.700000	61.800000	57.000000	5.690000	5.710000	3.520000	2375.000000
75%	20225.500000	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	26967.000000	4.500000	73.600000	79.000000	10.230000	58.900000	31.800000	18818.000000

Table 1.2. Dataset Summary (cubic zirconia)

Type of Variables:

```
<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 1.3. Type of Variables (cubic zirconia)

Check for duplicates: Number of Duplicates 0

Check for zero value in any numerical observation:

```
carat      False
cut        False
color      False
clarity    False
depth      False
table      False
x          False
y          False
z          False
price      False
dtype: bool
```

Table 1.4. Zero value in numerical observation (cubic zirconia)

Check for null values in the dataset:

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

Table 1.5. Null value in dataset (cubic zirconia)

Value count for non-numerical columns:

```
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 1.6. Count for non-numerical column (cubic zirconia)

Inference:

- Dataset has 11 columns out of which first column (Unnamed:0) is of no use for analysis and been removed
- Depth column has null values, Cut, color and clarity columns contain string values and rest of the columns are numerical
- All variables (columns) are of different scale. And there is a high chance of having outliers
- Dataset has no duplicate values and has no numerical observation with zero value
- Observed that only Depth column has 697 null values
- There are three non-numerical columns and all can be ordered

Outlier Analysis:

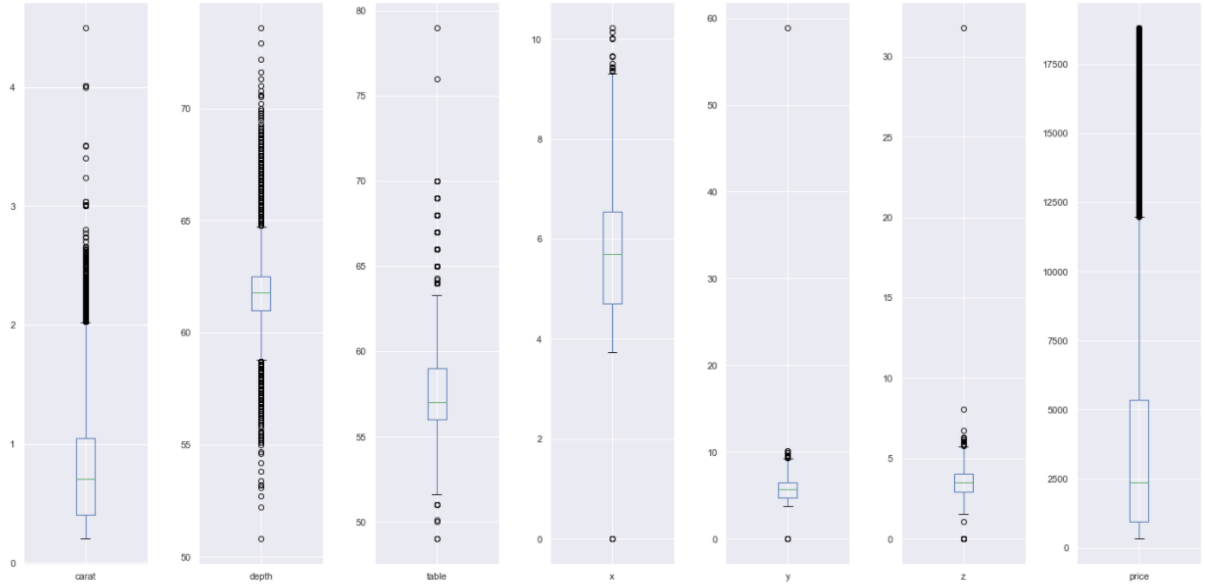

Fig 1.1. Boxplot before treating the outliers

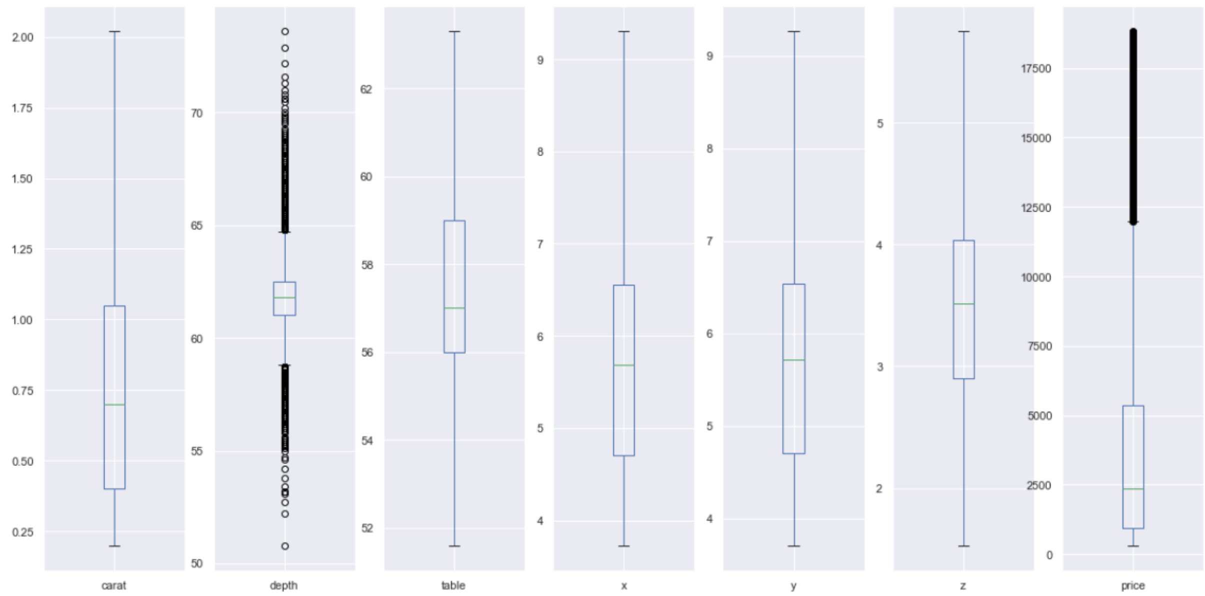

Fig 1.2. Boxplot after treating the outliers

Summary of Dataset after treating outliers:

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115
std	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	0.200000	50.800000	51.600000	3.730000	3.710000	1.530000	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	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	2.020000	73.600000	63.300000	9.300000	9.260000	5.750000	18818.000000

Table 1.7. Dataset Summary post treating outliers (cubic zirconia)

Inference:

The outliers in independent column were treated using 5-point summary, no changes were made on target column (price).

Data Visualization: EDA using sweet viz to visualize the summary for each variable as well to underrated data –

- Univariate and bivariate analysis

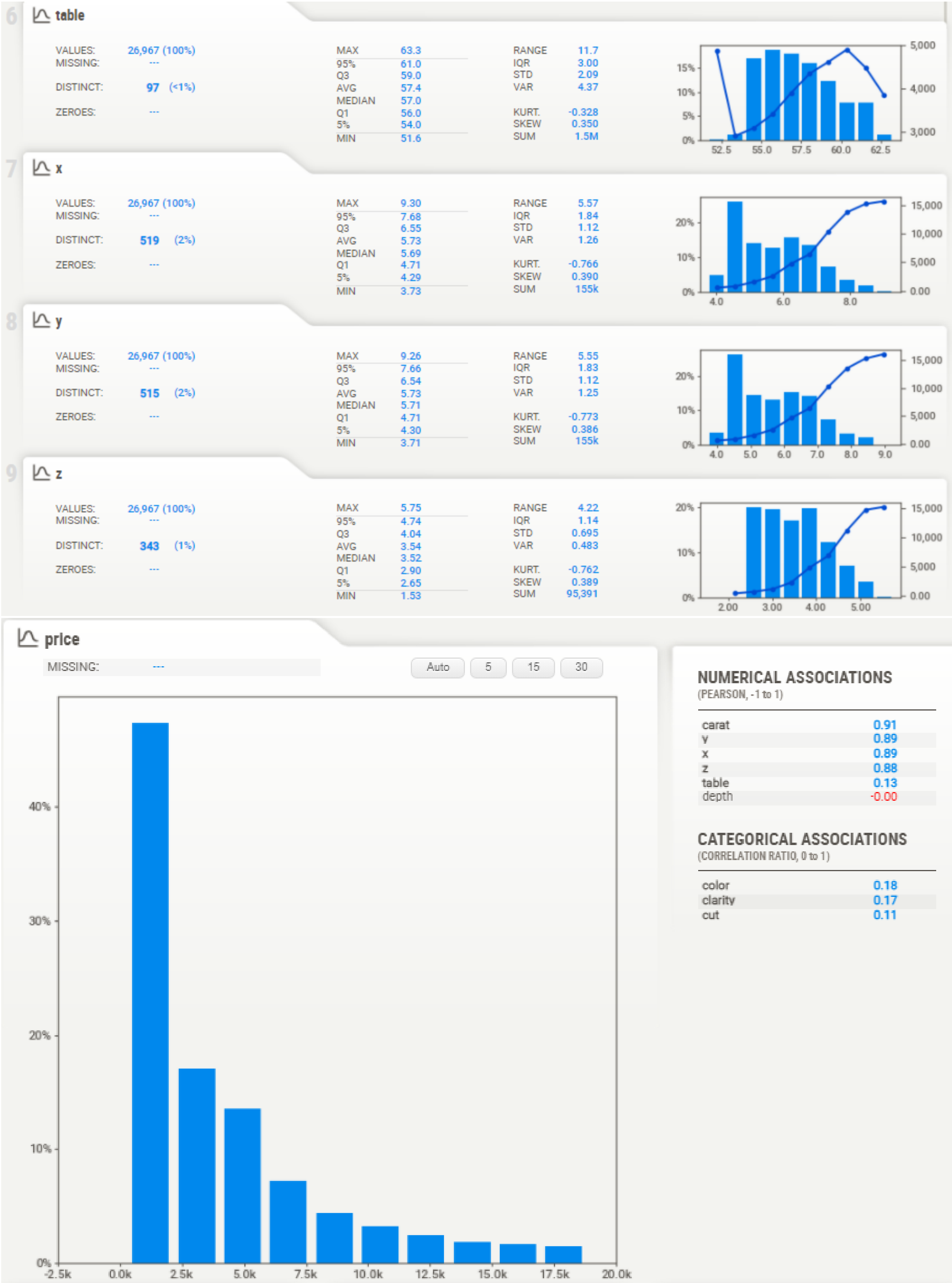

Fig 1.3. Sweet viz Univariate and bivariate analysis

- Multivariate Analysis:

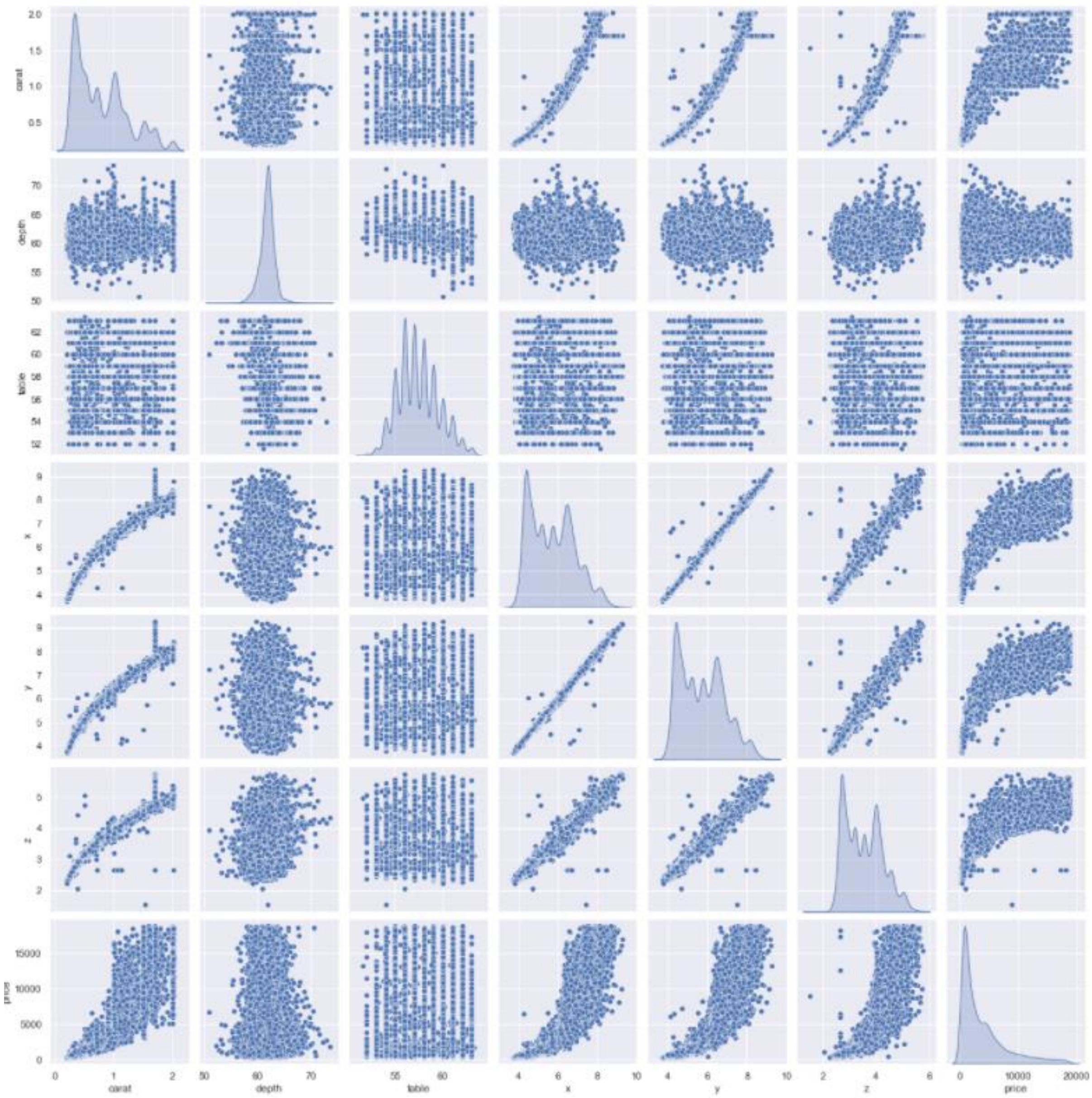

Fig 1.4. Pair plot analysis

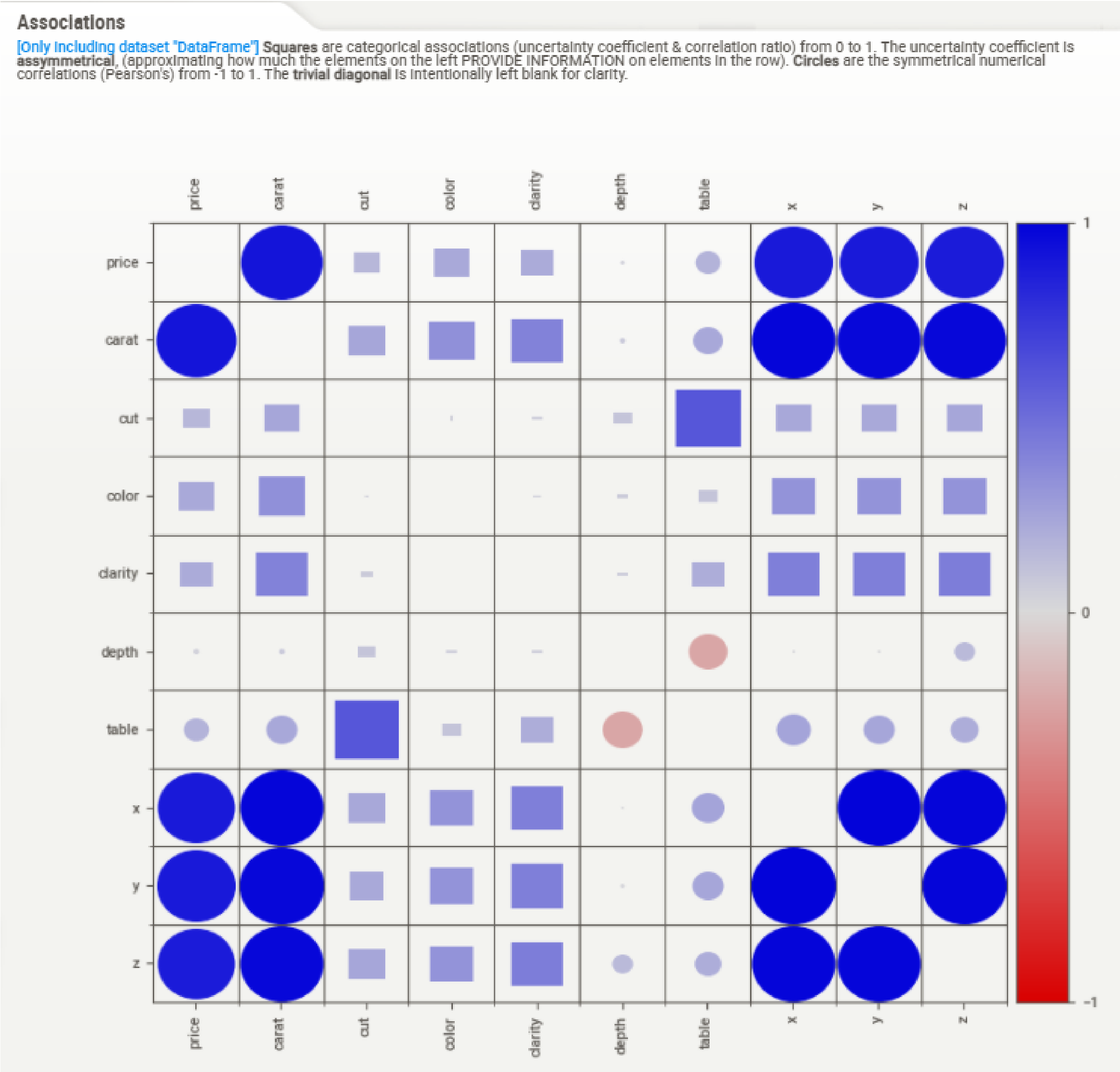

Fig 1.5. Sweet viz Multivariate analysis

Inference:

- High numerical correlation between target column price and carat, x, y, z
- High categorical correlation between price and color/clarity
- Negative correlation between price and depth
- Price and depth variables have uniform distribution. Price is right skewed

Insights:

- Information suggests that as the price increases the sales quantity reduces
- The carat and dimensions (x – length, y- width, z-height) plays an important role for pricing than other variables.

Q1.2. 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?

<u>Solution:</u>

Check for null values in the dataset:

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

Table 1.5. Null value in dataset (cubic zirconia)

Check for zero value in any numerical observation:

```
carat      False
cut        False
color      False
clarity    False
depth      False
table      False
x          False
y          False
z          False
price      False
dtype: bool
```

Table 1.4. Zero value in numerical observation (cubic zirconia)

Imputing null value using frequency:

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

Table 1.8. Null values post imputing

	carat	depth	table	x	y	z	price
count	26967.000000	26270.000000	26967.000000	26967.000000	26967.000000	26967.000000	26967.000000
mean	0.785860	61.745147	57.407702	5.729438	5.731334	3.537316	3939.518115
std	0.444042	1.412860	2.090151	1.124638	1.116593	0.694826	4024.864666
min	0.200000	50.800000	51.600000	3.730000	3.710000	1.530000	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	5.710000	3.520000	2375.000000
75%	1.050000	62.500000	59.000000	6.550000	6.540000	4.040000	5360.000000
max	2.020000	73.600000	63.300000	9.300000	9.260000	5.750000	18818.000000

Table 1.9. Dataset Summary post treating outliers and imputing null values

Inference:

- Depth column has got 697 null values and there is no zero value counts in the dataset
- As the column ‘depth’ is least correlated/ no correlation with the target variable ‘price’ imputing the null values of the depth column will not change the results on prediction much, however if we drop the column then it might affect prediction. Even though the count of null values is around 2% of the dataset, I have imputed it with most frequent value
- Yes, the scaling is required as the dataset has got different scales in the column variables. This will also help us to centre the variables and make predictors have the mean tends to zero. So, this will help in interpreting the intercept term as the expected value of the ‘price’ when the predictor value is set to their mean

Q1.3. 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.

Solution:

As the categorical columns are ordered in nature, I have encoded using ordinal encoder.

- Column cut: 'Fair':1,'Good':2,'Very Good':3,'Premium':4,'Ideal':5
- Column color: 'D':7,'E':6,'F':5,'G':4,'H':3,'I':2,'J':1 (D being the best and J being the worst
- Column clarity: 'FL':11,'IF':10,'VVS1':9,'VVS2':8,'VS1':7,'VS2':6,'SI1':5,'SI2':4,'I1':3,'I2':2,'I3':1 (from being best to worst, FL = flawless, I3 = level 3 inclusion)

Data set head after encoding:

	carat	cut	color	clarity	depth	table	x	y	z	price
0	0.3	5	6	5	62.1	58	4.27	4.29	2.66	499
1	0.33	4	4	10	60.8	58	4.42	4.46	2.7	984
2	0.9	3	6	8	62.2	60	6.04	6.12	3.78	6289
3	0.42	5	5	7	61.6	56	4.82	4.8	2.96	1082
4	0.31	5	5	9	60.4	59	4.35	4.43	2.65	779

Table 1.10. Dataset head post encoding

In the ratio of 70:30 both the target and independent variables data was split into test and train splits and used random state as 1:

	carat	cut	color	clarity	depth	table	x	y	z
11687	0.41	5	2	8	62.3	56.0	4.77	4.73	2.96
9728	1.71	5	1	5	62.8	57.0	7.58	7.55	4.75
1936	0.33	2	5	5	61.8	62.0	4.40	4.45	2.74
26220	0.70	3	3	5	62.8	57.0	5.61	5.66	3.54
18445	0.70	5	7	4	62.1	56.0	5.67	5.71	3.53

Table 1.11. X_train head

	carat	cut	color	clarity	depth	table	x	y	z
18031	2.01	1	2	4	66.5	61.0	7.81	7.75	5.17
26051	1.51	4	5	5	62.2	59.0	7.34	7.30	4.55
16279	0.50	3	3	5	60.9	61.0	5.06	5.15	3.11
16466	0.31	5	6	7	62.0	56.0	4.39	4.44	2.66
19837	1.20	3	3	7	62.0	57.0	6.77	6.81	4.21

Table 1.12. X_test head

	price
11687	1061.0
9728	6320.0
1936	536.0
26220	2214.0
18445	2575.0

Table 1.13. Y_train head

	price
18031	10671.0
26051	11607.0
16279	1133.0
16466	626.0
19837	6177.0

Table 1.14. Y_test head

Linear regression was applied and the following observations were noticed: LinearRegression()

```
The coefficient for carat is 1.2810088346596173
The coefficient for cut is 0.044086882558962044
The coefficient for color is 0.12342511078790368
The coefficient for clarity is 0.19166493838032614
The coefficient for depth is -0.003837535458018516
The coefficient for table is -0.0153963991906779
The coefficient for x is -0.53583661151529
The coefficient for y is 0.44065827297784976
The coefficient for z is -0.16422719315423415
```

Table 1.15. Coefficient of independent variable

Performance metrics:

- R-square on training data: 0.888699333687784
- R-square on testing data: 0.8836528787741355
- RMSE on training data: 0.3336175449706086
- RMSE on test data: 0.341096938165479

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

Business insights and recommendations:

- Zirconia Price = (3.363e-16) *intercept + (1.281) *carat + (0.044) *cut + (0.123) *color + (0.191) *clarity + (-0.003) *depth + (-0.015) *table + (-0.535) *x + (0.44) *y + (-0.164) *z
- The most important factors which determines the price of Zirconia are carat, length, width, clarity and height
- When carat is increased by 1unit, the price of zirconia increases by 1.28 units keeping all other predictors constant
- When length is increased by 1unit, the price of zirconia decreases by 0.535 units keeping all other predictors constant
- When width is increased by 1unit, the price of zirconia increases by 0.440 units keeping all other predictors constant
- When clarity is increased by 1unit, the price of zirconia increases by 0.191 units keeping all other predictors constant
- When height is increased by 1unit, the price of zirconia decreases by 0.164 units keeping all other predictors constant
- When color is increased by 1unit, the price of zirconia increases by 0.12 units keeping all other predictors constant

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

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.

Solution:

Sample of Dataset:

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

Table 2.1. Dataset Sample (Holiday Package)

Summary of Dataset:

	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 2.2. Dataset Summary (Holiday Package)

Type of Variables:

```
<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
```

Table 2.3. Type of Variables (Holiday Package)

Check for 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
```

Table 2.4. Null Values (Holiday Package)

Check for duplicates: Number of duplicate rows = 0

Value counts of categorical variables:

```
HOLLIDAY_PACKAGE :  2
yes    389
no     426
Name: Holliday_Package, dtype: int64


FOREIGN :  2
yes    211
no     604
Name: foreign, dtype: int64
```

```
no_young_children

0    617
1    141
2     52
3      5
Name: no_young_children, dtype: int64
```

Table 2.5. Count of Categorical variables (Holiday Package)

Check for outliers:

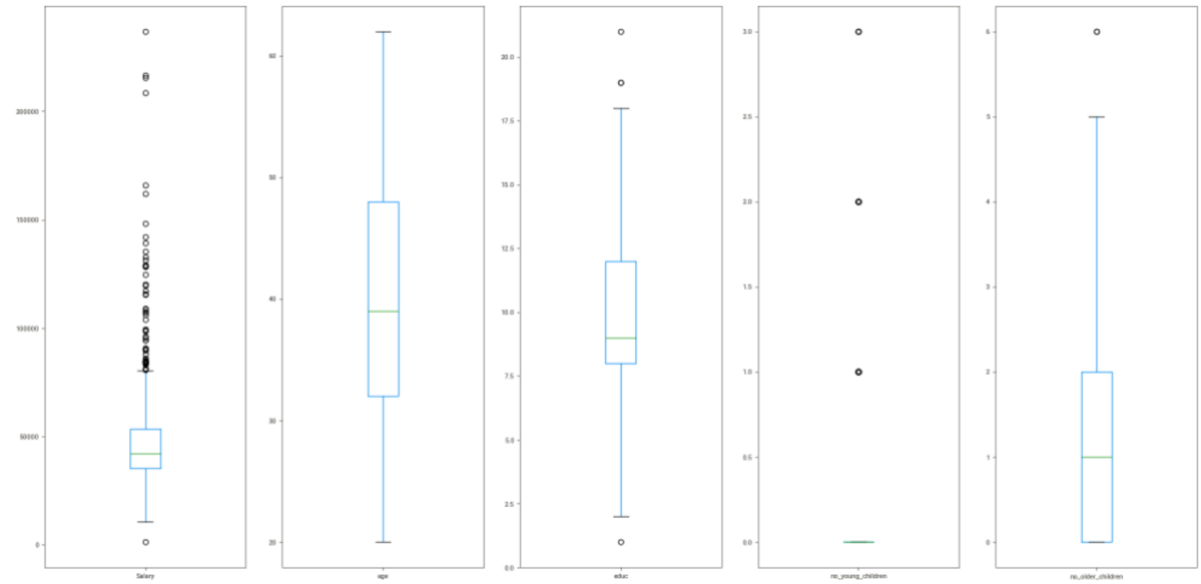

Fig 2.1. Box plot before treating outliers

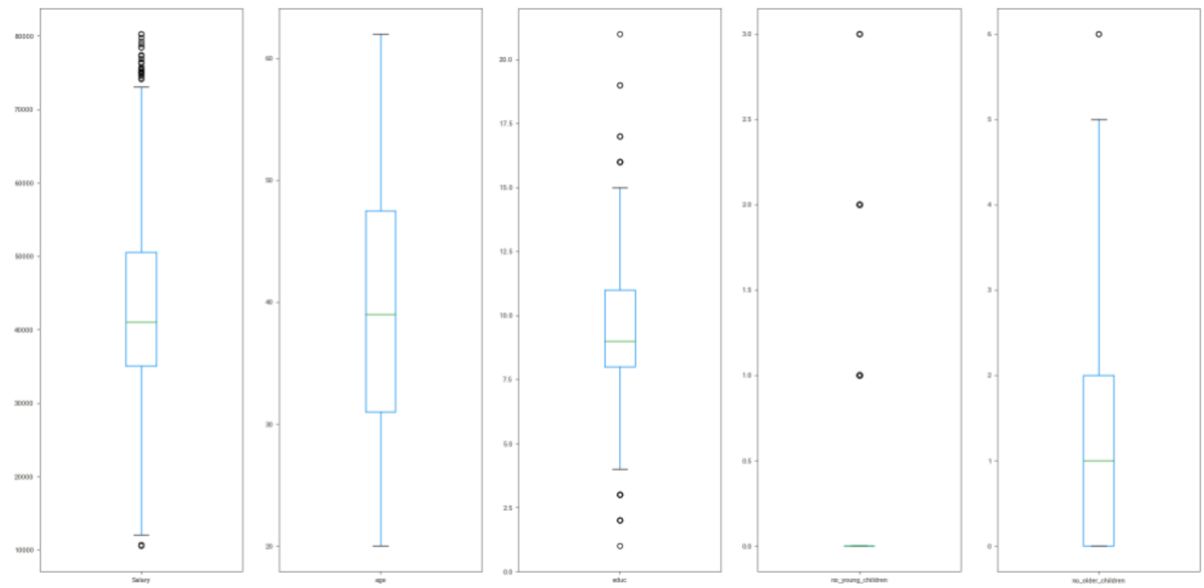

Fig 2.2. Box plot after treating outliers

Data Visualization: EDA using sweet viz to visualize the summary for each variable as well to underrated data –
Univariate and Bivariate analysis:

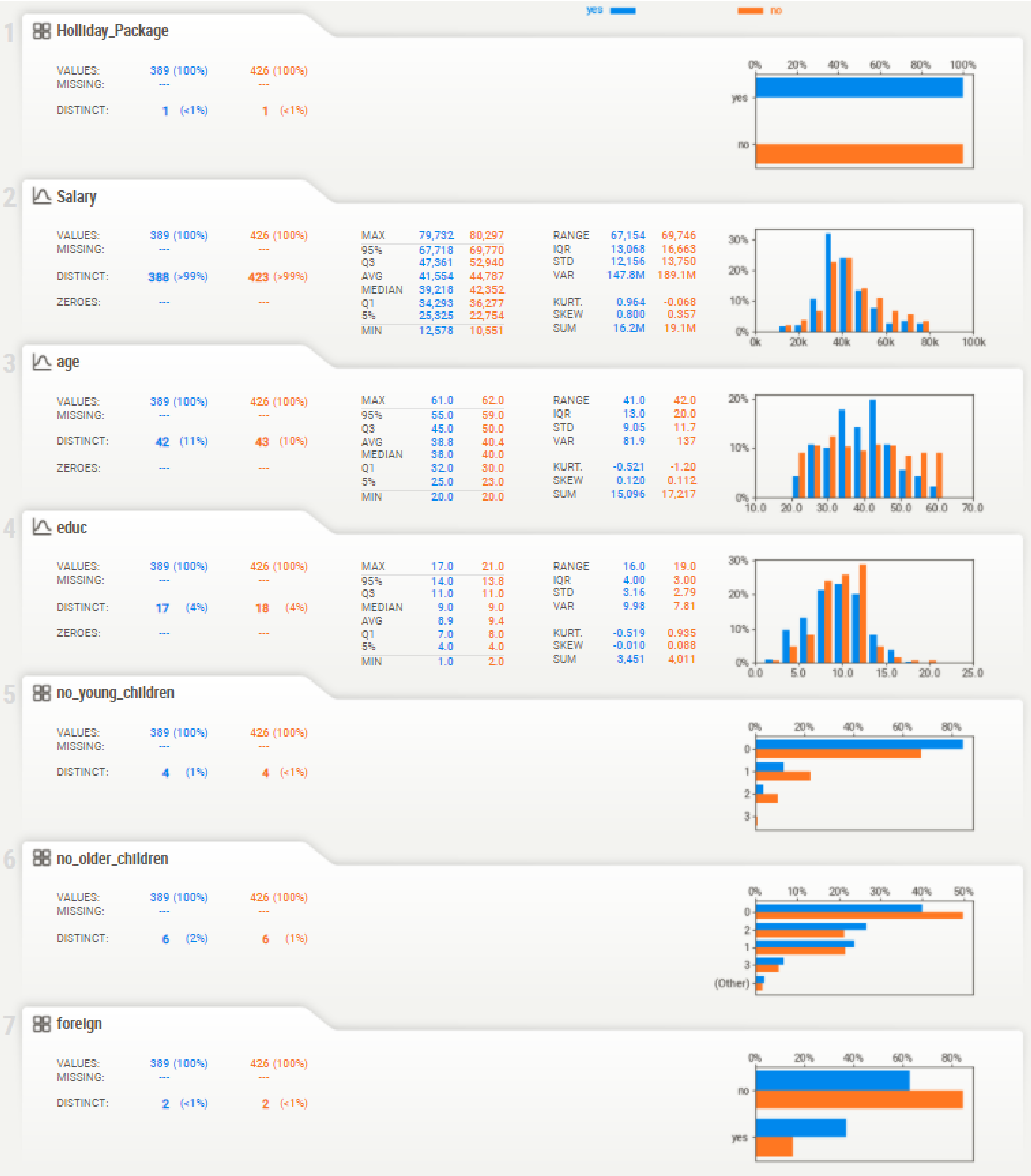

Fig 2.3. Sweet viz Univariate analysis

Multivariate analysis:

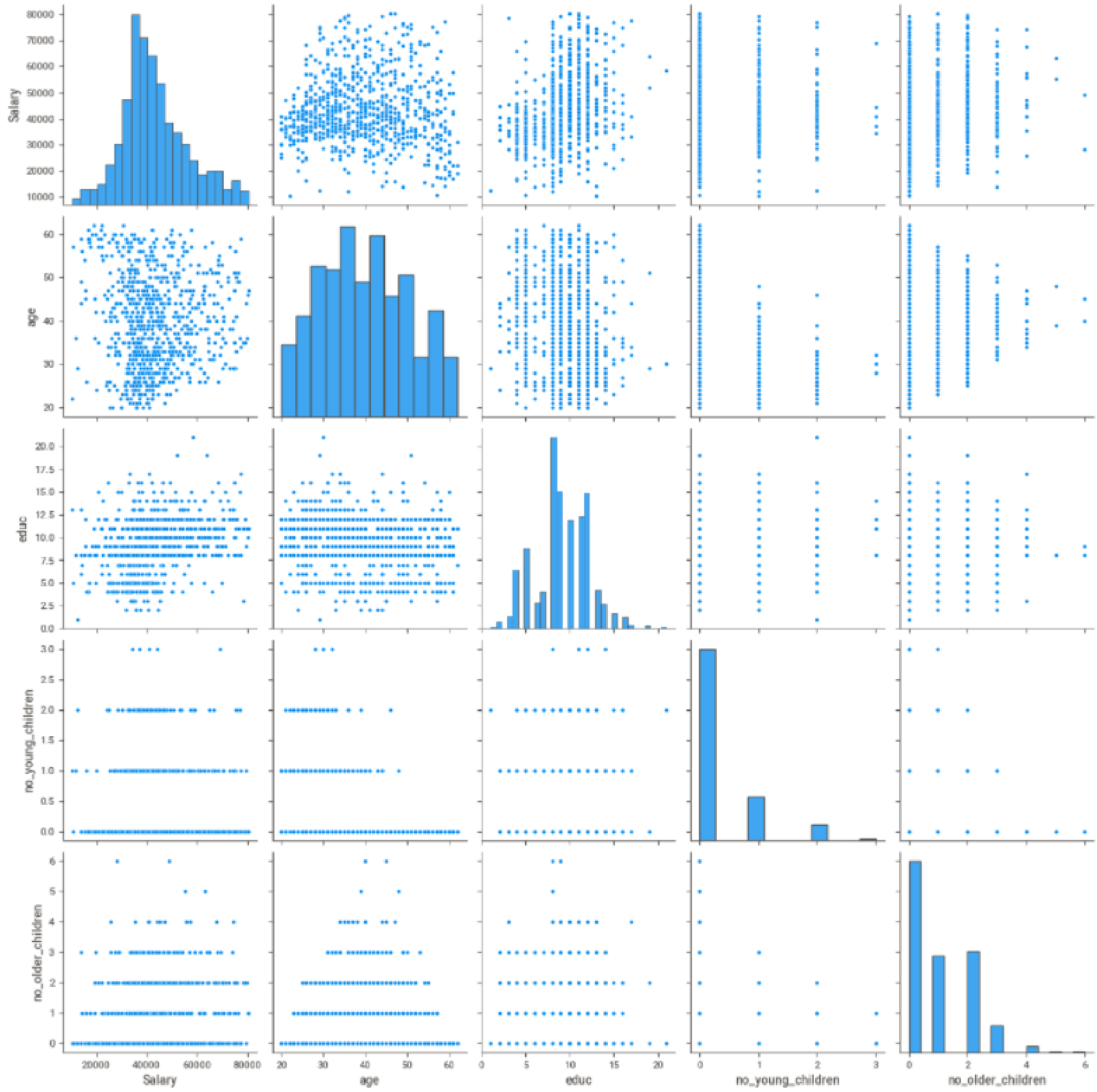

Fig 2.4. Pair plot

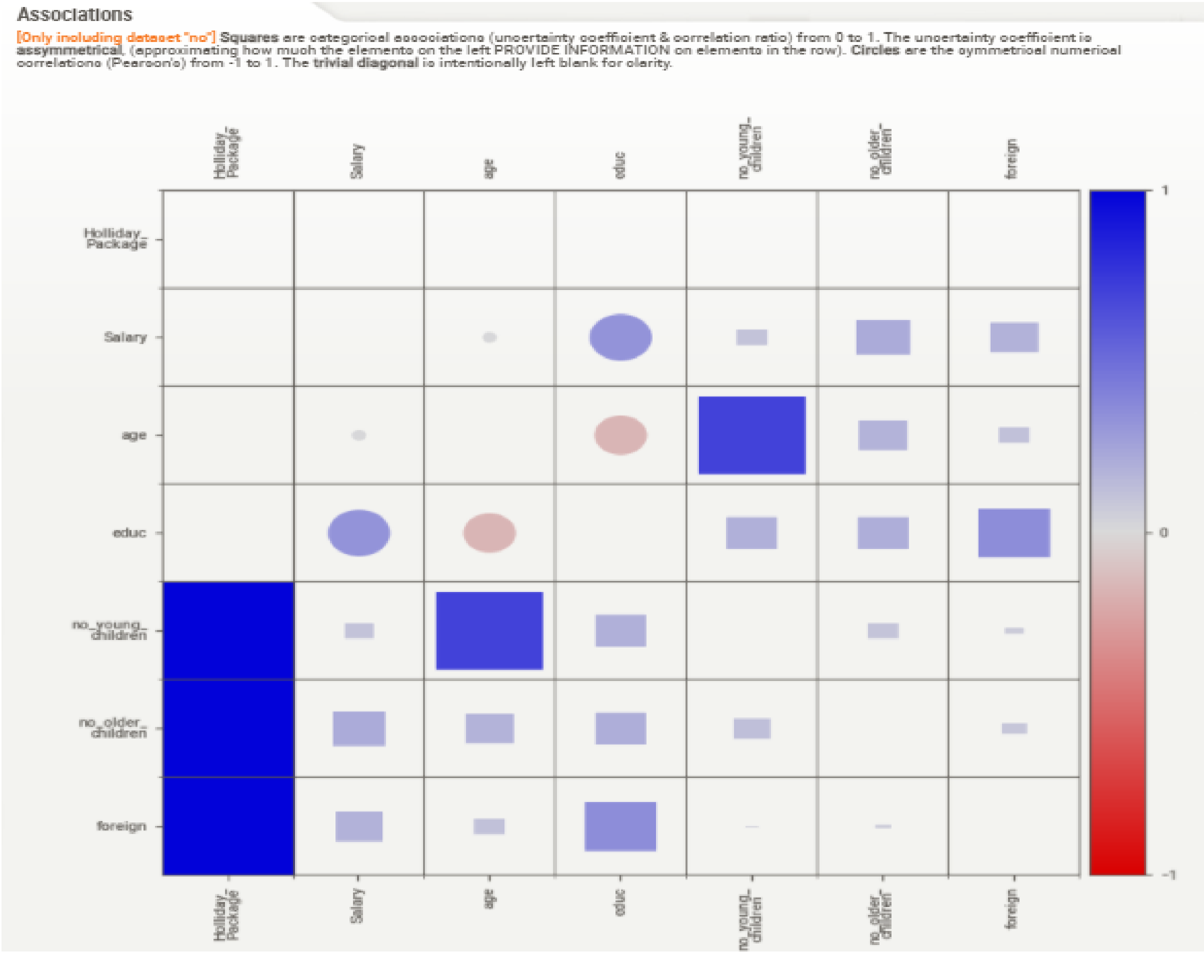

Fig 2.5. Sweet viz Multivariate analysis

Inference:

- First column (Unnamed:0) is of no use for analysis hence it's been removed
- There was a difference between 75th% and max value compared to 50th% and 75th%
- Only Salary has outliers in the continuous variable, apart from Salary and age rest all are categorical variable
- Outliers were treated using Z score
- Dataset has no null values, duplicate rows
- Noticed that there is high chance to take the package if the employee salary ranges between 30K to 40K, and if the employee age is in between 25 to 50 yrs.
- If the employee has no younger children, then there is a huge chance to tell yes
- If an employer is a foreigner, then there is chance in telling yes

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).

Solution:

Encode data: Foreign column is encoded to 0(if no) and 1(if yes)

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

Table 2.6. Encoded dataset (Holiday Package)

Data split: Data is successfully split into train and test (70:30) and random state 1 (numpy.matrix)

Model for Logistic regression and LDA is built

- LogisticRegression(solver='liblinear')
- LinearDiscriminantAnalysis()

Q2.3 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.

Solution:

Logistic Regression:

- Accuracy score for Logistic regression train variables 0.6508771929824562
- Accuracy score for Logistic regression test variables 0.6204081632653061
- Confusion Matrix

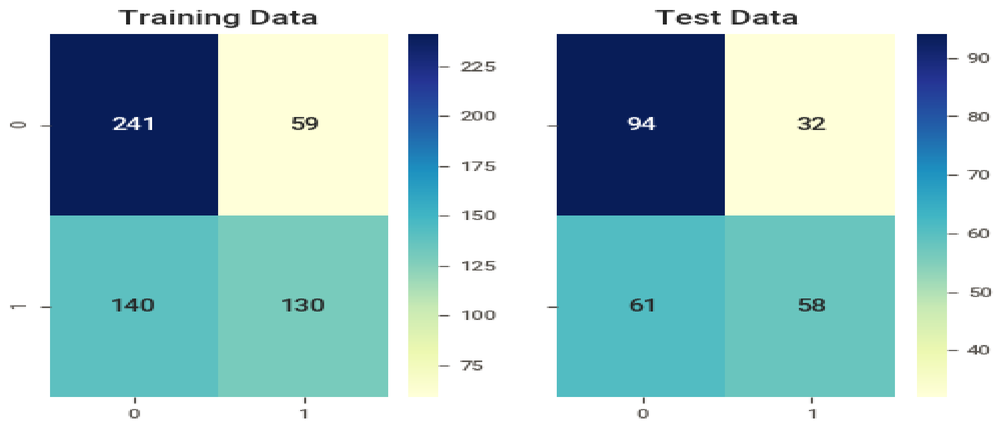

Fig 2.6. Confusion Matrix Logistic Regression

- Classification report for Logistic Regression

Classification Report of the training data:

	precision	recall	f1-score	support
no	0.63	0.80	0.71	300
yes	0.69	0.48	0.57	270
accuracy			0.65	570
macro avg	0.66	0.64	0.64	570
weighted avg	0.66	0.65	0.64	570

Classification Report of the test data:

	precision	recall	f1-score	support
no	0.61	0.75	0.67	126
yes	0.64	0.49	0.56	119
accuracy			0.62	245
macro avg	0.63	0.62	0.61	245
weighted avg	0.62	0.62	0.61	245

Table 2.7. Classification report Logistic Regression

- ROC curve and ROC_AUC score for Logistic Regression
    - AUC for the Training Data: 0.738
    - AUC for the Test Data: 0.665

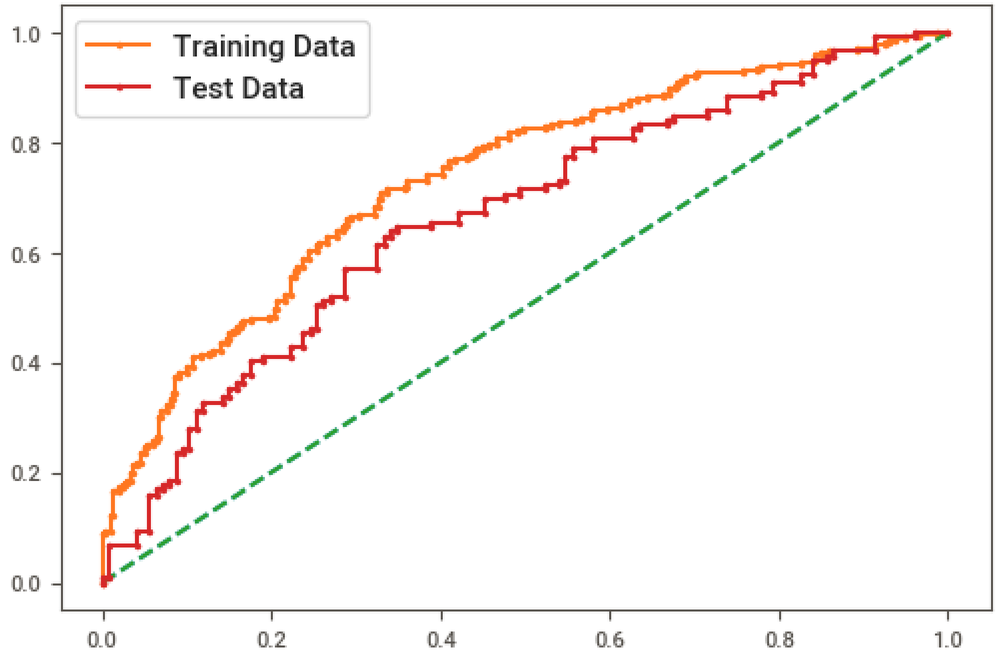

Fig 2.7. ROC Curve Logistic Regression

Linear Discriminant Analysis:

- Accuracy score for LDA train variables 0.6754385964912281
- Accuracy score for LDA test variables 0.6204081632653061
- Confusion Matrix

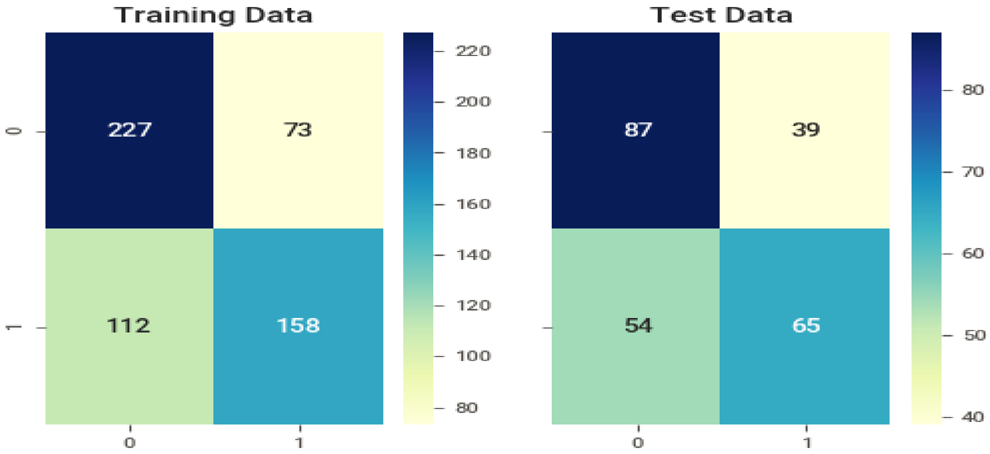

Fig 2.8. Confusion Matrix LDA

- Classification report for Linear Discriminant Analysis

Classification Report of the training data:

	precision	recall	f1-score	support
no	0.67	0.76	0.71	300
yes	0.68	0.59	0.63	270
accuracy			0.68	570
macro avg	0.68	0.67	0.67	570
weighted avg	0.68	0.68	0.67	570

Classification Report of the test data:

	precision	recall	f1-score	support
no	0.62	0.69	0.65	126
yes	0.62	0.55	0.58	119
accuracy			0.62	245
macro avg	0.62	0.62	0.62	245
weighted avg	0.62	0.62	0.62	245

Table 2.8. Classification report LDA

- ROC curve and ROC_AUC score for LDA
  - AUC for the Training Data: 0.743
  - AUC for the Test Data: 0.670

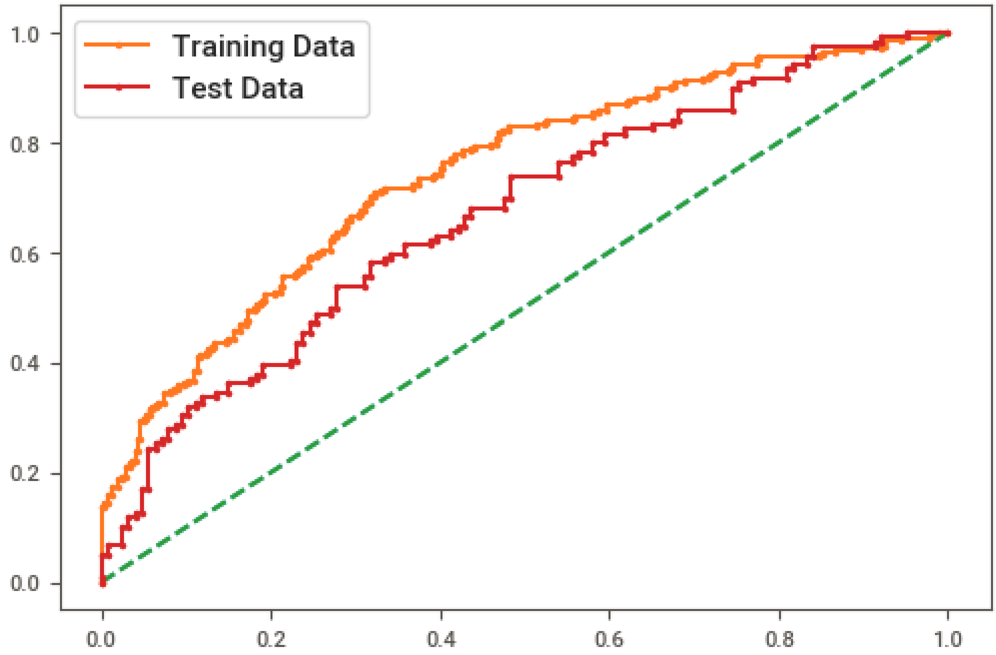

Fig 2.9. ROC Curve for LDA

Comparing both the models:

	Logistic reg Train	Logistic reg Test	LDA Train	LDA Test
Accuracy	0.65	0.62	0.68	0.62
AUC	0.74	0.67	0.74	0.67
Recall	0.48	0.49	0.59	0.55
Precision	0.69	0.64	0.68	0.62
F1 Score	0.57	0.56	0.63	0.58

Table 2.9. Comparing LR and LDA models

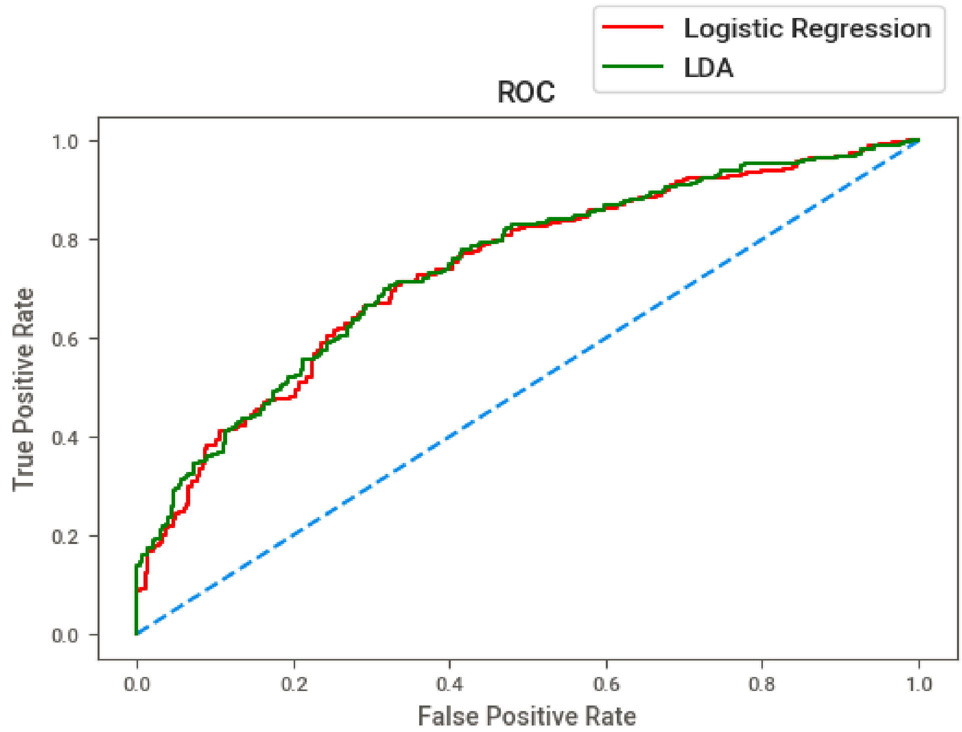

Fig 2.10. Comparing ROC Curve for Train Data

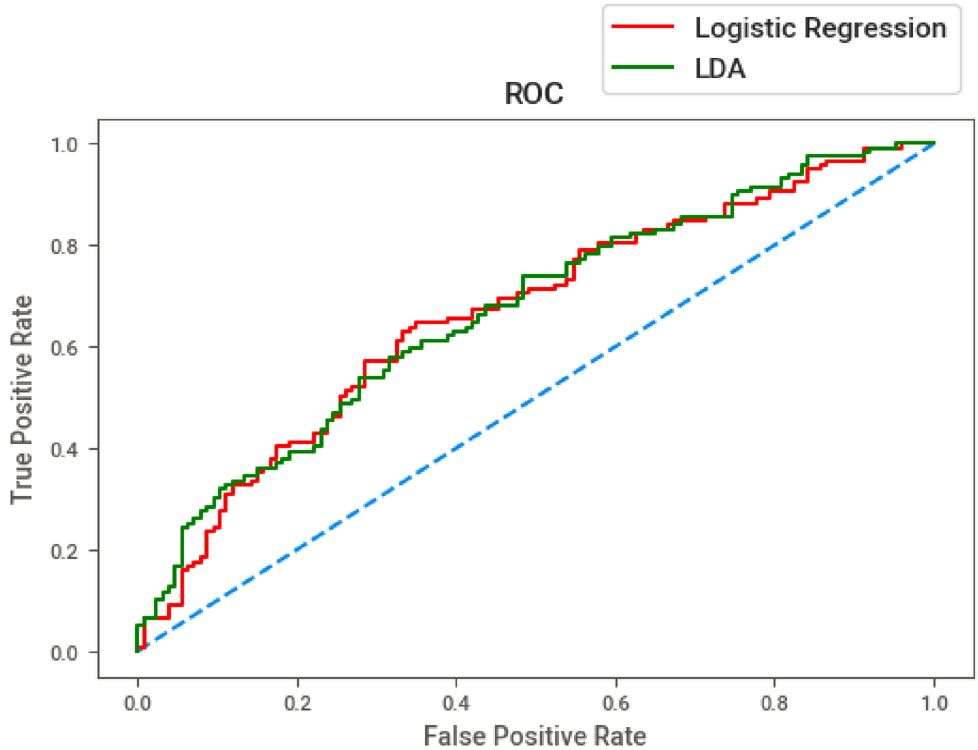

Fig 2.11. Comparing ROC Curve for Test Data

Inference:

Based on comparing the performance metrics, Linear Discriminant Analysis (LDA) performs better than the Logistic Regression because of the better recall rate and accuracy. Hence LDA is the best model.

Q2.4 Inference: Basis on these predictions, what are the insights and recommendations.

Solution:

The Linear Discriminant Analysis model will be able to predict whether an employee will opt for a package or not with around 70 percent accuracy.

Business insights:

- The important factors which determine whether an employee will opt in for package are
  - Salary
  - Age
  - No. of young children
  - Foreign
- The company must focus on the people who earns between 30k to 40k and between the age of 25 yrs. to 50 yrs. and if they have no children then there is a huge chance of opting in for a package

Business Recommendations:

- The greater number of people who are opting in for package has a salary range between 30k to 40k. It suggests that the package is of average price with medium level facilities. So, if they add some additional luxury packages with facilities like booking in star hotels, luxury cars etc. may help to increase the sales of the packages to a higher income group

- The analysis shows that a greater number of foreigners opt in for packages than non-foreigners. This along with the previous analysis which shows that most of the people are from salary group of 30k to 50k suggests that packages provided are either of local sightseeing place or of less interest to the non-foreigners
  So, suggest the company to add some more activities or places in their packages

- The analysis shows that if an employee having no young children, then there is more chance to opt in for the package. As count of children increases, the willingness to opt in for a package decrease. So, I suggest the company to provide additional discounts or children attractiveness for the employee who has young children to boost up the chance of them opting for the package

Thanks & regards,
Pavan Kumar R Naik