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.

Below provides five number summaries for all continuous variables such as carat, depth, table, x, y, z and price while providing statistics on unique value count, top value with its frequencies for each categorical variable such as cut, color and clarity.

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967	26967	26967	26967	26270	26967	26967	26967	26967	26967
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	61.74515	57.45608	5.729854	5.733569	3.538057	3939.518
std	0.477745	NaN	NaN	NaN	1.41286	2.232068	1.128516	1.166058	0.720624	4024.865
min	0.2	NaN	NaN	NaN	50.8	49	0	0	0	326
25%	0.4	NaN	NaN	NaN	61	56	4.71	4.71	2.9	945
50%	0.7	NaN	NaN	NaN	61.8	57	5.69	5.71	3.52	2375
75%	1.05	NaN	NaN	NaN	62.5	59	6.55	6.54	4.04	5360
max	4.5	NaN	NaN	NaN	73.6	79	10.23	58.9	31.8	18818

Data types of the independent variables

Independent variables	Data type
carat	float64
cut	object
color	object
clarity	object
depth	float64
table	float64
x	float64
y	float64
z	float64

The predicted variable is numeric and continuous as expected.

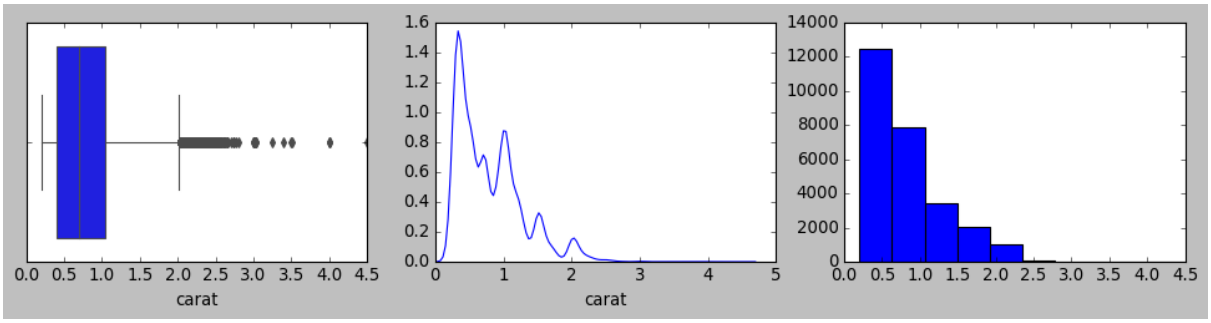
Shape:

```
26967 rows and 10 columns
```

Univariate analysis:

1. Univariate analysis for carat

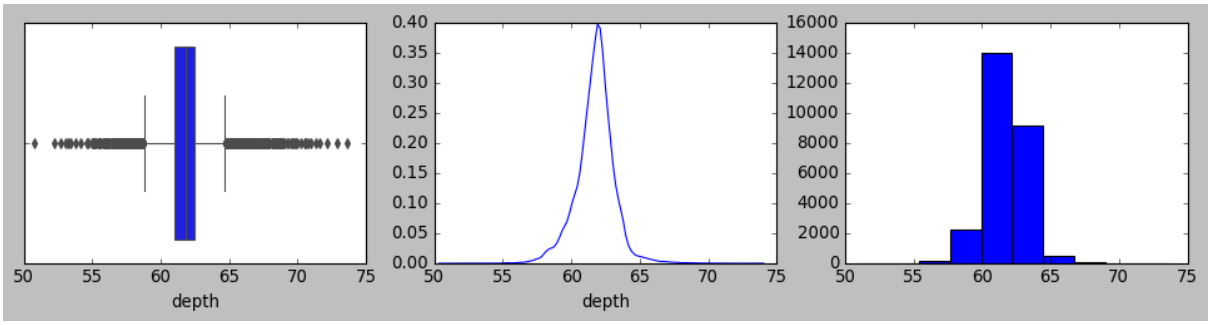
Mean is 0.798375, Median is 0.700000, Mode(s) are 0.3000
Column carat has outliers



Column carat is not normally distributed

2. Univariate analysis for depth

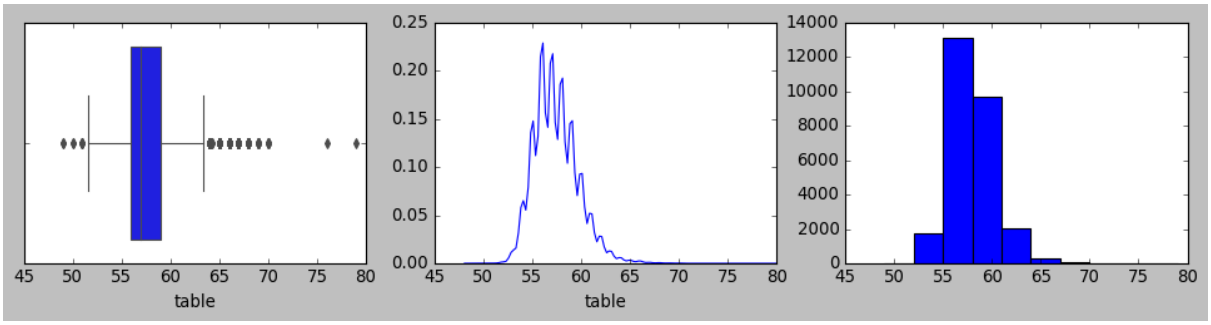
Mean is 61.745147, Median is nan, Mode(s) are 62.0000
Column depth has outliers



Column depth is normally distributed

3. Univariate analysis for table

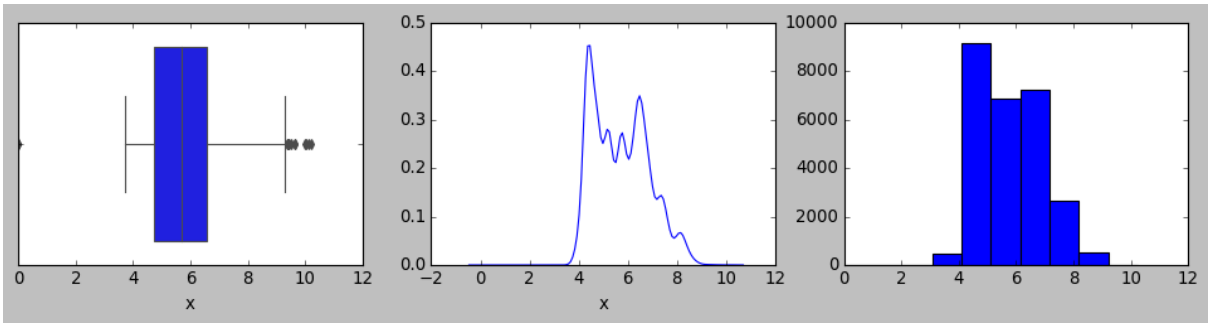
Mean is 57.456080, Median is 57.000000, Mode(s) are 56.0000
Column table has outliers



Column table is not normally distributed

4. Univariate analysis for x

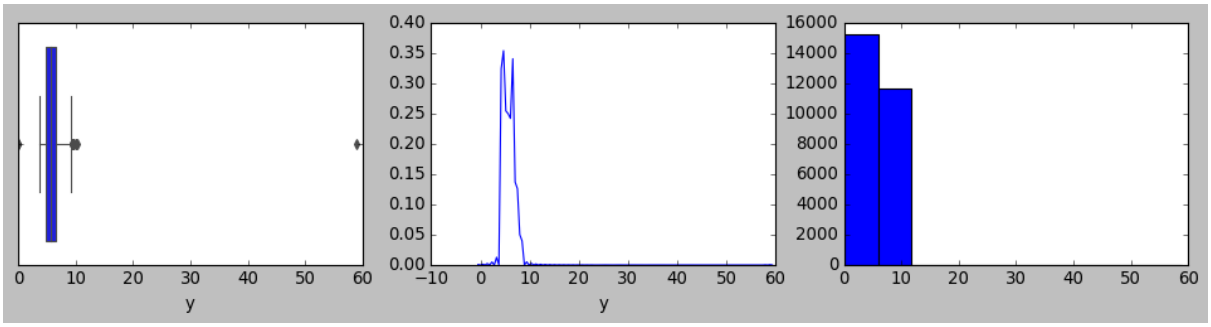
Mean is 5.729854, Median is 5.690000, Mode(s) are 4.3800
Column x has outliers



Column x is not normally distributed

5. Univariate analysis for y

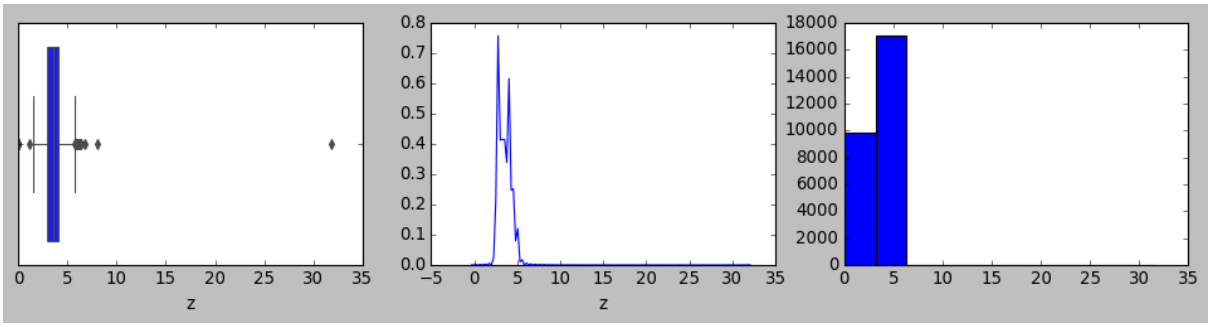
Mean is 5.733569, Median is 5.710000, Mode(s) are 4.3500
Column y has outliers



Column y is not normally distributed

6. Univariate analysis for z

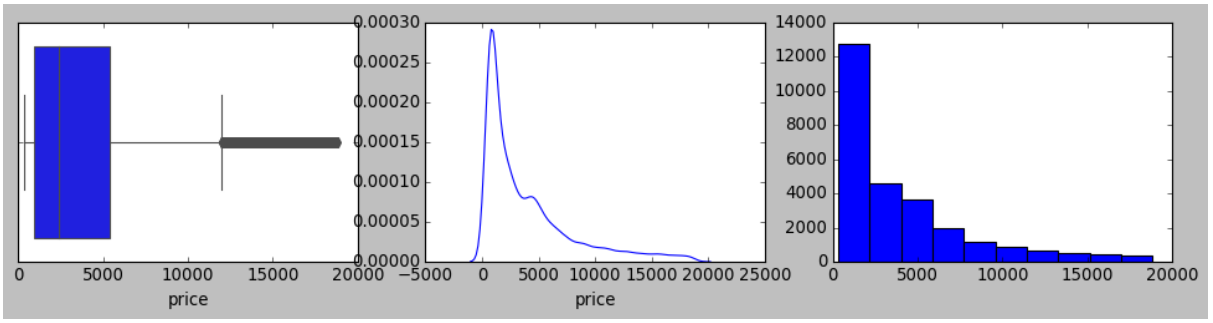
Mean is 3.538057, Median is 3.520000, Mode(s) are 2.6900
Column z has outliers



Column z is not normally distributed

7. Univariate analysis for price

Mean is 3939.518115, Median is 2375.000000, Mode(s) are 544.0000
Column price has outliers



Column price is not normally distributed

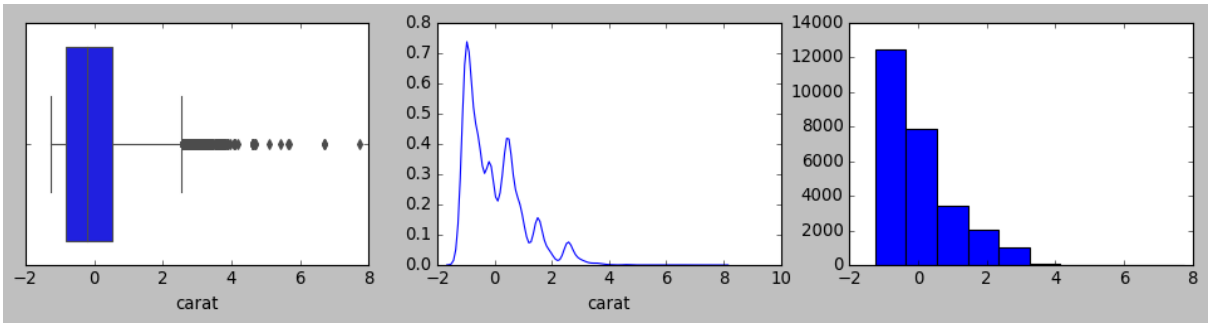
With the fact that independent variables have different units, scaling becomes necessary to remove the units variables are associated with so that the linear equation can be formed on the independent variables post standardization of them.

Z scoring based scaling of data would change the coefficient ,neutralize/remove the intercept while the accuracy score remains the same before and after. MSE would get scaled too.

Accordingly, below is the univariate analysis for the dataset that is standardized on z scores.To be noted that the below observation on the data is observed after encoding the categorical variables such as cut, color and clarity since in linear equation it demands every variable to be numeric to apply mathematical calculations on it.

1. Univariate analysis for carat

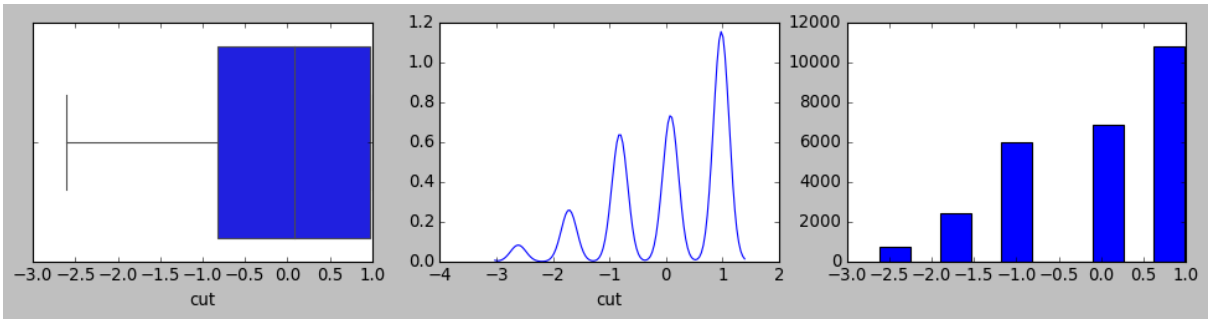
Mean is -0.000000, Median is -0.205920, Mode(s) are -1.0432
Column carat has outliers



Column carat is not normally distributed

2. Univariate analysis for cut

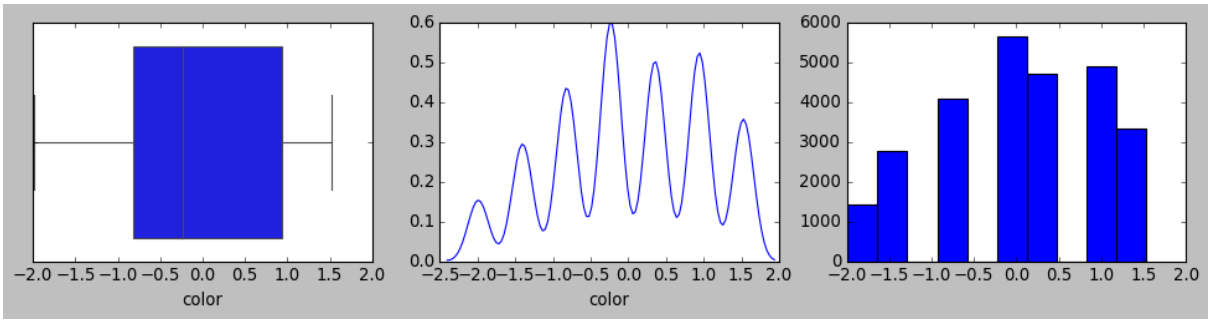
Mean is 0.000000, Median is 0.081246, Mode(s) are 0.9796
Column cut does not have outliers



Column cut is not normally distributed

3. Univariate analysis for color

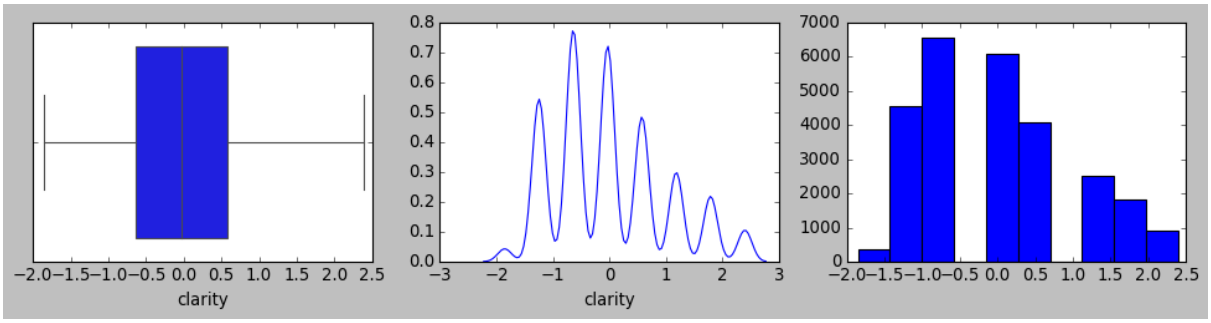
Mean is 0.000000, Median is -0.230890, Mode(s) are -0.2309
Column color does not have outliers



Column color is not normally distributed

4. Univariate analysis for clarity

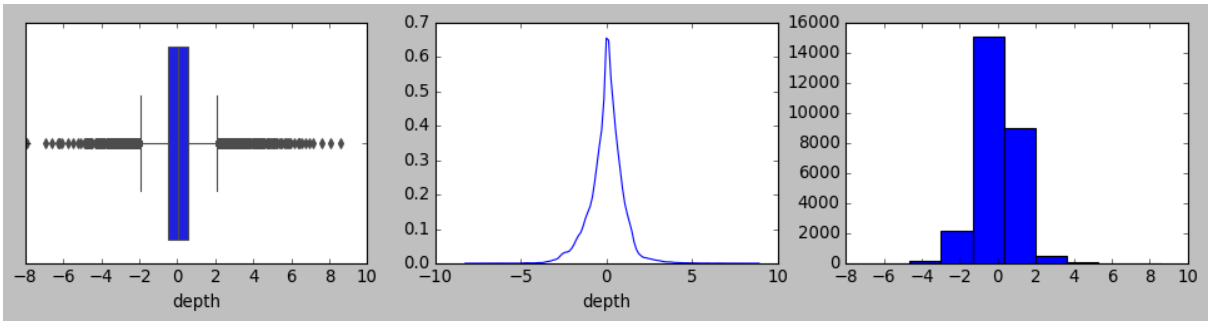
Mean is 0.000000, Median is -0.032241, Mode(s) are -0.6394
Column clarity does not have outliers



Column clarity is not normally distributed

5. Univariate analysis for depth

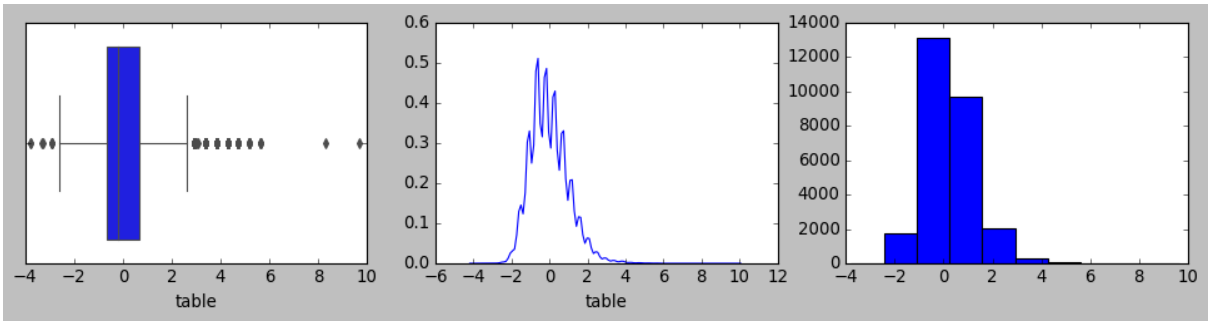
Mean is 0.000000, Median is 0.036839, Mode(s) are 0.0368
Column depth has outliers



Column depth is not normally distributed

6. Univariate analysis for table

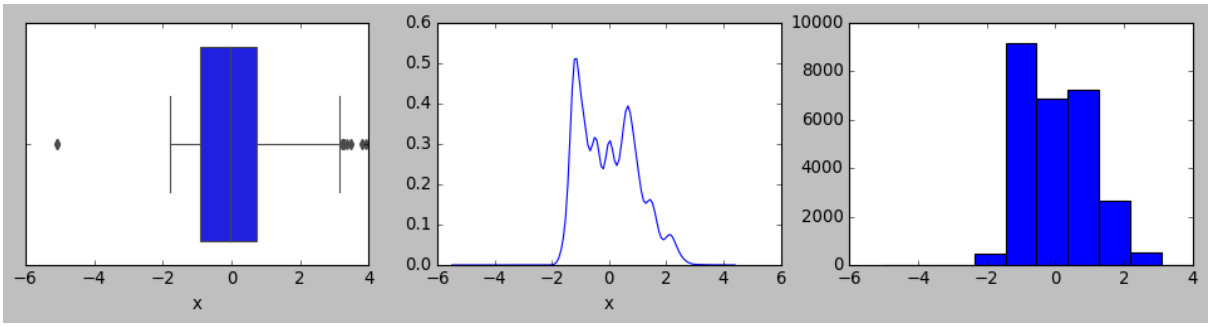
Mean is -0.000000, Median is -0.204334, Mode(s) are -0.6524
Column table has outliers



Column table is not normally distributed

7. Univariate analysis for x

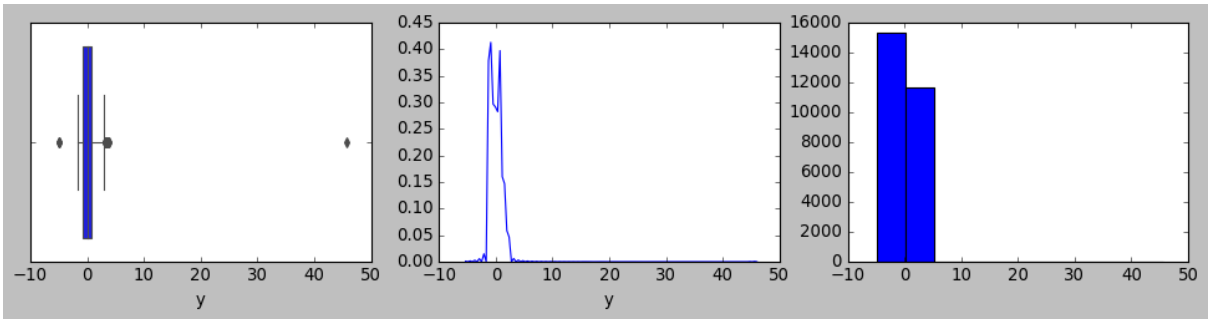
Mean is 0.000000, Median is -0.035316, Mode(s) are -1.1962
Column x has outliers



Column x is not normally distributed

8. Univariate analysis for y

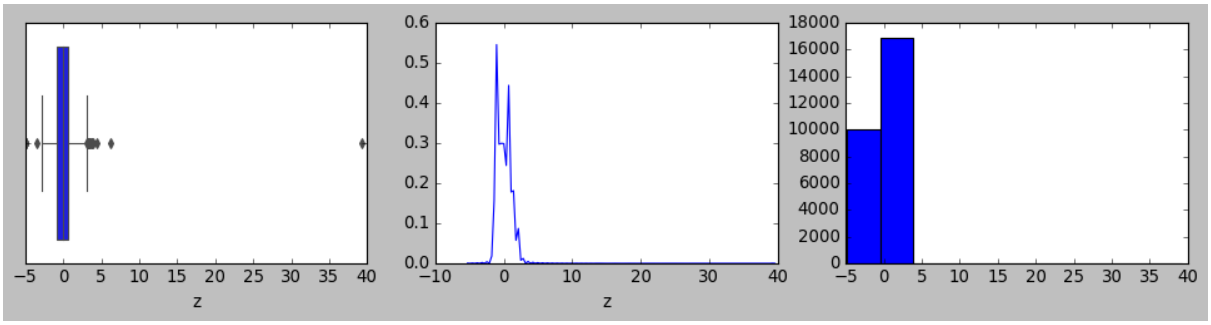
Mean is -0.000000, Median is -0.020213, Mode(s) are -1.1866
Column y has outliers



Column y is not normally distributed

9. Univariate analysis for z

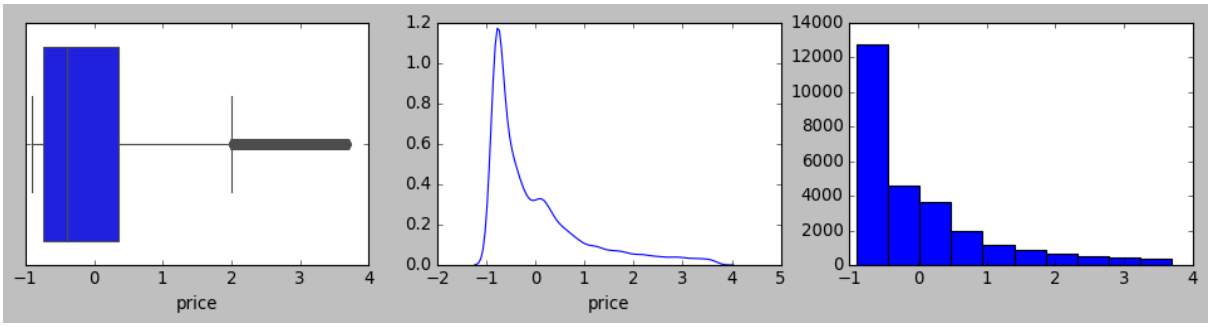
Mean is -0.000000, Median is -0.025058, Mode(s) are -1.1769
Column z has outliers



Column z is not normally distributed

10. Univariate analysis for price

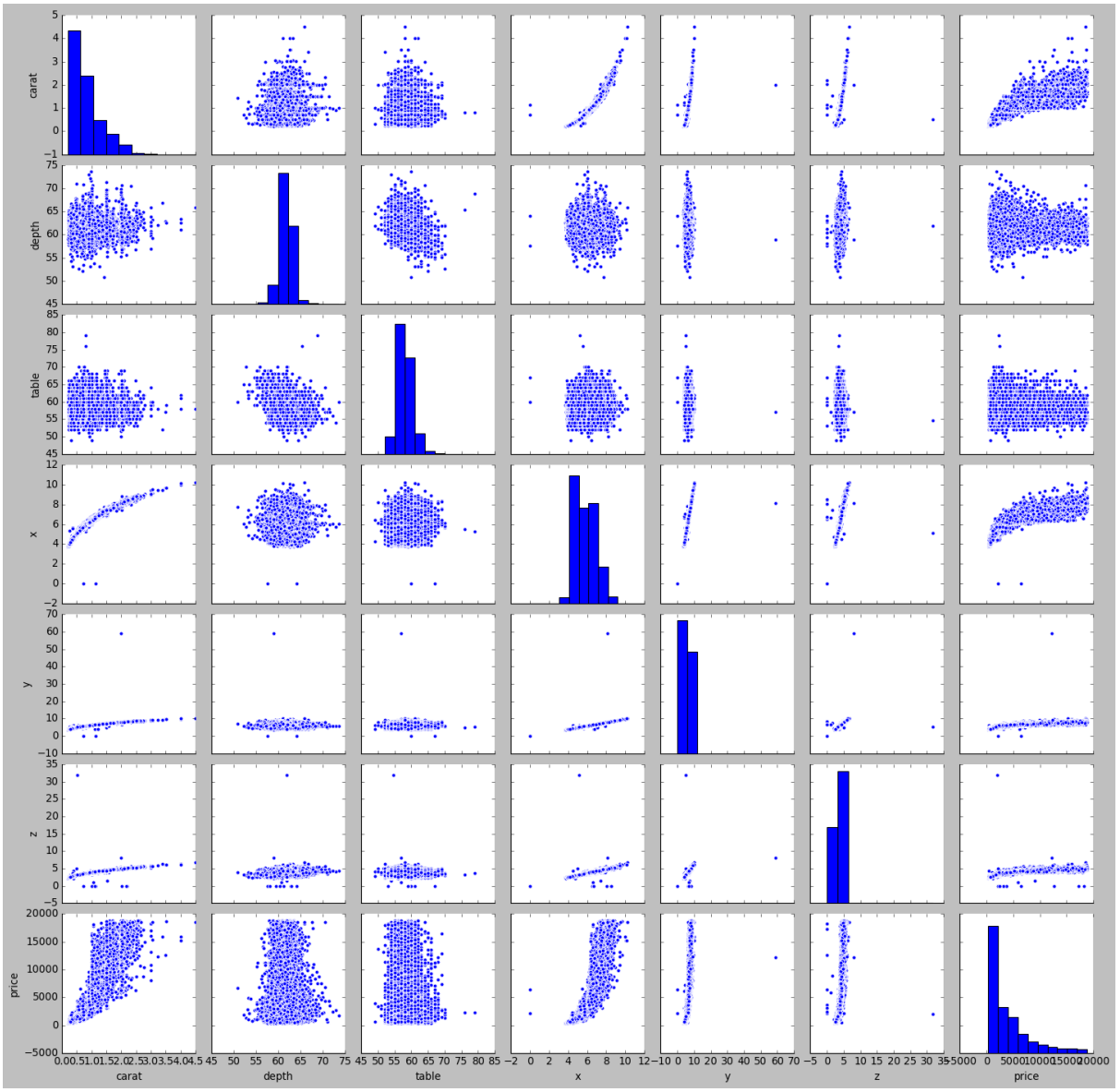
Mean is -0.000000, Median is -0.388720, Mode(s) are -0.8437
Column price has outliers



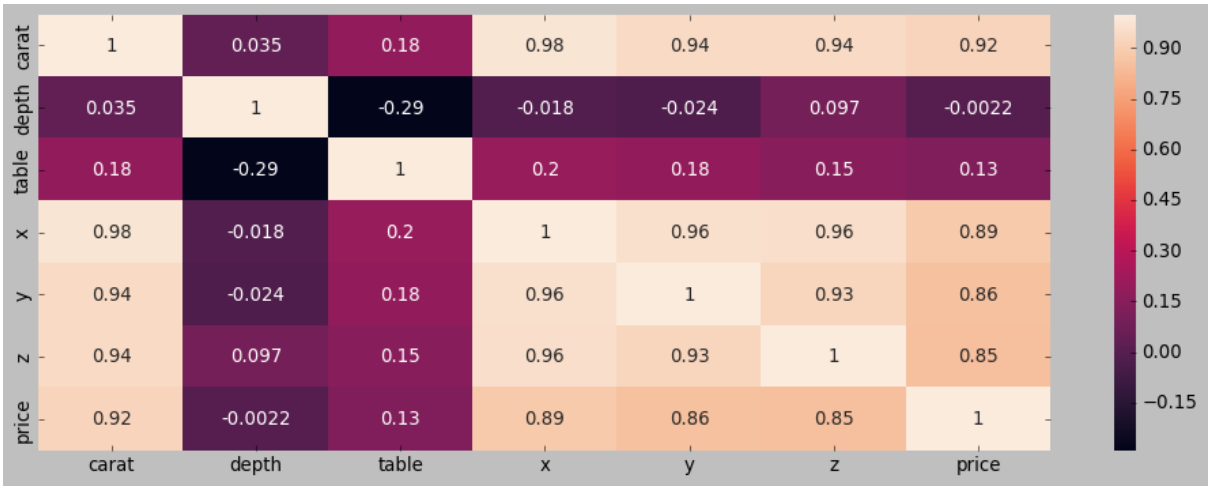
Column price is not normally distributed

Bi variate analysis:

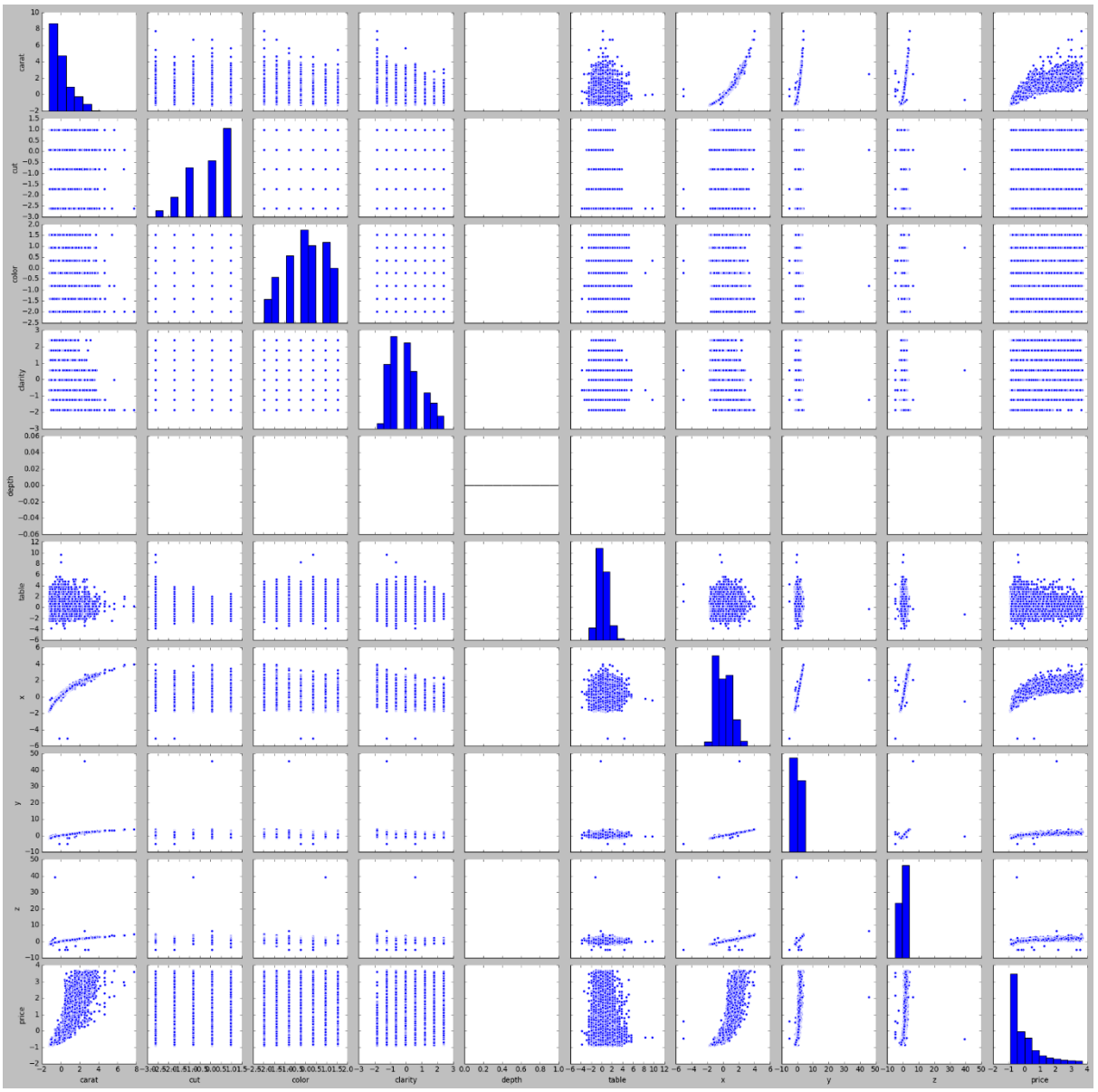
Pair plot for unscaled dataset:



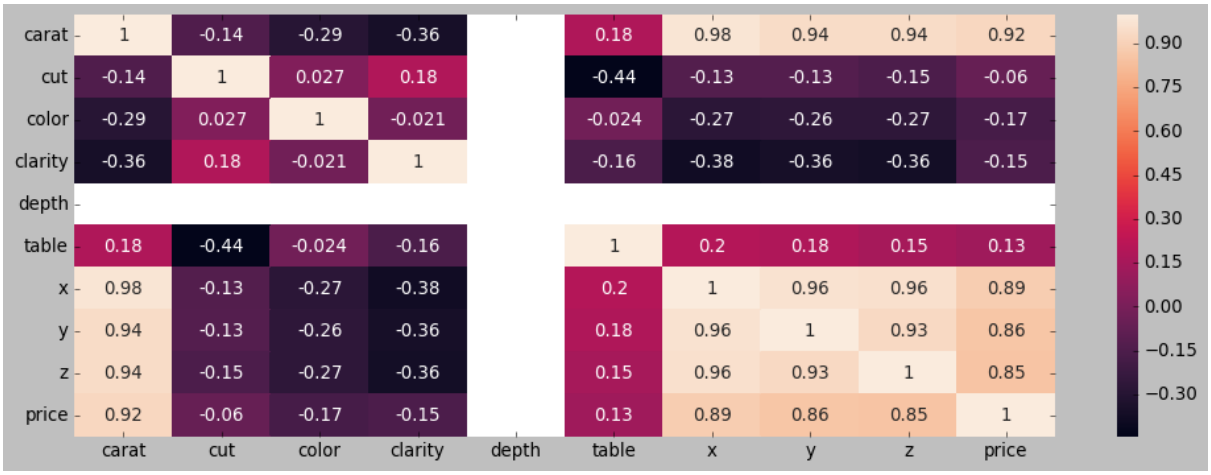
Heatmap for unscaled dataset:



Pair plot for scaled dataset:



Heatmap for scaled dataset:



Observations:

It is noticed that price (target variable) depicts great correlation among carat, X, Y and Z independent variables across all combinations of them.

We also have to further observe if these variables are potential contributors for multi collinearity which is not good for explaining the coefficients. However, from a pure prediction standpoint these multi collinearity does not matter if we don't have to explain coefficients.

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?

There are null values in depth in 697 rows. There are various ways to handle missing values. Drop the rows, replace missing values with median values etc. It is a sizable data for us to retain as they have other independent variables contributing to the linear equation algorithm. Hence depth has been imputed with the median value.

Independent variables	Count of null values
carat	0
cut	0
color	0
clarity	0
depth	697
table	0

x	0
y	0
z	0

Below table suggests variables that have zeros as its value in the given dataset. Accordingly X, Y and Z are the variables that contain zeros across one or more rows.

Independent variables	Zero values present (Yes/No)?
carat	No
cut	No
color	No
clarity	No
depth	No
table	No
x	Yes
y	Yes
z	Yes

Below are the 9 rows that have 0's either in X, Y or Z. In the linear equation these rows would not contribute to prediction and hence no impact by retaining them. However dropping the rows could have adverse effects as other variables could potentially contribute with better coefficients in predicting price (target variable) which shall not be missed.

Index	carat	cut	color	clarity	depth	table	x	y	z	price
5822	0.71	Good	F	SI2	64.1	60	0	0	0	2130
6035	2.02	Premium	H	VS2	62.7	53	8.02	7.95	0	18207
6216	0.71	Good	F	SI2	64.1	60	0	0	0	2130
10828	2.2	Premium	H	SI1	61.2	59	8.42	8.37	0	17265
12499	2.18	Premium	H	SI2	59.4	61	8.49	8.45	0	12631
12690	1.1	Premium	G	SI2	63	59	6.5	6.47	0	3696
17507	1.14	Fair	G	VS1	57.5	67	0	0	0	6381
18195	1.01	Premium	H	I1	58.1	59	6.66	6.6	0	3167
23759	1.12	Premium	G	I1	60.4	59	6.71	6.67	0	2383

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

Please check below for the unique values across categorical variables and the respective count of number of rows:

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


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


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

Since the linear algorithms require predictor variables to be numeric we shall assess if we need to convert them into codes or perform one hot encoding. Since in this case cut, color and clarity have values that indicate/depict the order we shall convert them into codes with highest value given to the variables as per below approach.

Based on the metadata clarified in the question, we wanted to assign the highest value depicting premium quality , best colour and best clarity for cut, color and clarity variables.

- Cut : Fair: 1, Good: 2, Very Good:3, Premium:4, Ideal:5
- color : D:7, E:6, F:5, G:4, H:3, I:2, J:1
- clarity: FL:11, IF:10, VVS1:9, VVS2:8, VS1:7, VS2:6, SI1:5, SI2:4, I1:3, I2:2, I3:1

Also, given that outliers can impact the model output while arriving at the linear equation and its coefficients, they shall be treated. In this case carat is treated for outliers beyond maximum whisker while depth, table, X, Y, Z and price are treated for outliers beyond both sides of the whisker. Below is the five number summary post the dataset has been treated for the outlier.

	carat	cut	color	clarity	depth	table	x	y	z	price
count	26967	2.70E+04	2.70E+04	2.70E+04	0	26967	26967	26967	26967	26967
mean	-0.01001	4.80E-16	1.49E-16	6.18E-17	NaN	-0.009131	0.000044	-0.00146	-0.00111	-0.050091
std	0.967962	1.00E+00	1.00E+00	1.00E+00	NaN	0.966443	0.998695	0.959443	0.967622	0.862377
min	-1.25252	-2.61E+00	-1.99E+00	-1.85E+00	NaN	-2.668463	-3.34946	-2.66846	-3.25843	-0.897815
25%	-0.83388	-8.17E-01	-8.17E-01	-6.39E-01	NaN	-0.652358	-0.90373	-0.87782	-0.88544	-0.744018
50%	-0.20592	8.12E-02	-2.31E-01	-3.22E-02	NaN	-0.204334	-0.03532	-0.02021	-0.02506	-0.38872
75%	0.526701	9.80E-01	9.41E-01	5.75E-01	NaN	0.691712	0.726761	0.691601	0.696552	0.352933
max	2.567575	9.80E-01	1.53E+00	2.40E+00	NaN	2.707817	3.172495	3.045731	3.069541	1.998361

Subsequently the split data into training and testing data at the ratio of 70:30 has the following five number summaries.

Train data:

	carat	cut	color	clarity	depth	table	x	y	z
count	18876	18876	18876	18876	18876	18876	18876	18876	18876
mean	-0.011	-0.00803	-0.001492	0.000503	0.004303	-0.007321	-0.00092	-0.00197	-0.00192
std	0.967735	0.999851	0.999856	0.995418	0.875865	0.964069	0.998108	0.959029	0.967608
min	-1.25252	-2.61367	-1.98943	-1.85372	-1.969594	-2.668463	-3.34946	-2.66846	-3.25843
25%	-0.83388	-0.81706	-0.81707	-0.6394	-0.463659	-0.652358	-0.91259	-0.87782	-0.88544
50%	-0.20592	0.081246	-0.23089	-0.03224	0.038319	-0.204334	-0.03532	-0.02879	-0.02506
75%	0.526701	0.97955	0.94147	0.574919	0.540298	0.691712	0.726761	0.691601	0.696552
max	2.567575	0.97955	1.52765	2.3964	2.046232	2.707817	3.172495	3.045731	3.069541

Test data:

	carat	cut	color	clarity	depth	table	x	y	z
count	8091	8091	8091	8091	8091	8091	8091	8091	8091
mean	-0.00771	0.018739	0.003481	-0.00117	-0.000625	-0.013353	0.00228	-0.00026	0.000797
std	0.968546	1.000221	1.00045	1.010731	0.87004	0.972004	1.000121	0.960467	0.967711
min	-1.25252	-2.61367	-1.98943	-1.85372	-1.969594	-2.668463	-3.34946	-2.66846	-3.25843
25%	-0.83388	-0.81706	-0.81707	-0.6394	-0.463659	-0.652358	-0.90373	-0.86924	-0.88544
50%	-0.20592	0.081246	-0.23089	-0.03224	0.038319	-0.204334	-0.02645	-0.02021	-0.02506
75%	0.526701	0.97955	0.94147	0.574919	0.540298	0.691712	0.7179	0.691601	0.696552
max	2.567575	0.97955	1.52765	2.3964	2.046232	2.707817	3.163634	2.972834	3.069541

Based on the regression model that has been built based on the above training and testing data set below are the statistical summary using the statsmodel library.

OLS Regression Results

Dep. Variable:	price	R-squared:	0.932
Model:	OLS	Adj. R-squared:	0.932
Method:	Least Squares	F-statistic:	2.857e+04
Date:	Sun, 25 Oct 2020	Prob (F-statistic):	0.00
Time:	11:00:22	Log-Likelihood:	1293.0
No. Observations:	18876	AIC:	-2566.
Df Residuals:	18866	BIC:	-2488.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0370	0.002	-22.466	0.000	-0.040	-0.034
carat	1.0464	0.010	109.827	0.000	1.028	1.065
cut	0.0311	0.002	15.385	0.000	0.027	0.035
color	0.1161	0.002	66.597	0.000	0.113	0.119
clarity	0.1772	0.002	96.577	0.000	0.174	0.181
depth	-0.0034	0.003	-1.120	0.263	-0.009	0.003
table	-0.0086	0.002	-3.946	0.000	-0.013	-0.004
x	-0.4008	0.036	-11.137	0.000	-0.471	-0.330
y	0.3530	0.037	9.426	0.000	0.280	0.426
z	-0.0427	0.016	-2.728	0.006	-0.073	-0.012

Omnibus:	2632.787	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9984.946
Skew:	0.669	Prob(JB):	0.00
Kurtosis:	6.302	Cond. No.	62.0

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Based on the null hypothesis that states that the independent variables have no co relationship with target variables from the universe, please note the coefficients of the predictor variables from the sample should be having p value lesser than the significance value of 5% . Based on the stats model based statistical observation predictor variables such as carat, cut, color, clarity, table, x, y and z shall be considered as a function of predicted variable (price) as their p value is much lesser than 5%. However for the predictor variable “depth” the p value indicates that the null hypothesis needs to be rejected and hence this will not contribute to the efficiency of the model. Hence “depth” shall be descoped from the linear equation.

Subsequently VIF has been evaluated to check multi collinearity among predictor variables. Any multi collinearity among predictor variables would mean that despite prediction itself need not be impacted with any low accuracy, the interpretation of the coefficients would be wrong without treating multi collinearity. Any VIF score between 1 to 5 is an acceptable range to retain the predictor variables in linear equation. Please find below the VIF for all the predictor variables considered so far without “depth” variable.

VIF for color :1.1199180000646427
VIF for clarity :1.239446323097618
VIF for cut :1.5097266302884316
VIF for table :1.6171628502013768
VIF for depth :2.8080406229119284
VIF for carat :31.360238644953316
VIF for z :104.8133511789732
VIF for y :366.3698715201059
VIF for x :379.79254326971653

As we can notice that variables such as x, y and z depicts much higher VIF than 5, further treatment on the linear equation to remove them one by one in the order of variable with highest VIF value taken out first, we could fine removing x, y and z results in much better VIF values removing all multi collinearity as below. Also the train and test data has been fixed to remove depth, x, y and z variables on which the model is getting executed with required predictor variables.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.931
Model:                            OLS   Adj. R-squared:                  0.931
Method:                 Least Squares   F-statistic:                 5.075e+04
Date:                Sun, 25 Oct 2020   Prob (F-statistic):               0.00
Time:                        11:08:27   Log-Likelihood:                 1174.3
No. Observations:               18876   AIC:                            -2337.
Df Residuals:                   18870   BIC:                            -2290.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0384      0.002    -23.213      0.000      -0.042      -0.035
carat          0.9438      0.002    484.268      0.000       0.940       0.948
cut            0.0311      0.002     16.823      0.000       0.028       0.035
table         -0.0089      0.002     -4.635      0.000      -0.013      -0.005
color          0.1155      0.002     66.023      0.000       0.112       0.119
clarity        0.1823      0.002    100.529      0.000       0.179       0.186
==============================================================================
Omnibus:                     2399.885   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7997.882
Skew:                           0.645   Prob(JB):                         0.00
Kurtosis:                       5.917   Cond. No.                         1.86
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Accordingly please find the revised VIF value for these remaining variables that has been considered in the revised linear equation.

VIF for color :1.1182180001886444
VIF for clarity :1.1987705393921508
VIF for cut :1.4942796531076148
VIF for table :1.580701029953695
VIF for depth :1.3231554005711066
VIF for carat :1.3031075405289636

We could notice the R squared value as well as adjusted R squared value has not changed much while we have the optimized set of predictor variables to make the linear equation to derive their coefficients.

Also the adjusted R square value (coefficient of determinant) is almost the same as R squared value indicating there is no statistical fluke indicating that there is no inconsistency on distribution between sample and the universe.

Please find below the coefficient of the predictor variables:

```
carat          0.9438
cut            0.0311
color          0.1155
clarity        0.1823
```

```
table          -0.0089

Also, the intercept is -0.0384
```

Please find below the accuracy metrics/score:

- R Squared: 0.93 (Same as adjusted R Squared)
- Root mean square error: 0.2288135274586359

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

```
(-0.04) * Intercept + (0.94) * carat + (0.03) * cut + (-0.01) * table +
(0.12) * color + (0.18) * clarity
```

Based on the coefficients making the above equation it can be understood that carat, x and y has the highest weight in that order towards predicting price with x having negative coefficient. Color and clarity also have reasonable coefficients that contribute to the predicting power for price.

So when carat increases by 1 unit price increases by 0.94 units while when cut increases by 1 unit price decreases by 0.01 unit and so on.

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.

Please find below the data types for all the columns in the given dataset.

Column	Type
Holliday_Package	object
Salary	int64
age	int64
educ	int64
no_young_children	int64
no_older_children	int64
foreign	object

Below provides five number summaries for all continuous variables such as Salary, age, educ, no_young_children and no_older_children while providing statistics on unique value count, top value with its frequencies for each categorical variable such as foreign and Holliday_Package.

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872	872	872	872	872	872
unique	2	NaN	NaN	NaN	NaN	NaN	2
top	no	NaN	NaN	NaN	NaN	NaN	no
freq	471	NaN	NaN	NaN	NaN	NaN	656
mean	NaN	47729.172	39.95528	9.30734	0.311927	0.982798	NaN
std	NaN	23418.6685	10.55168	3.03626	0.61287	1.086786	NaN
min	NaN	1322	20	1	0	0	NaN
25%	NaN	35324	32	8	0	0	NaN
50%	NaN	41903.5	39	9	0	1	NaN
75%	NaN	53469.5	48	12	0	2	NaN
max	NaN	236961	62	21	3	6	NaN

Null value check for all the columns in the dataset

Independent variables	Count of null values
Holliday_Package	0
Salary	0
age	0
educ	0
no_young_children	0
no_older_children	0
foreign	0

There are no duplicates in the dataset.

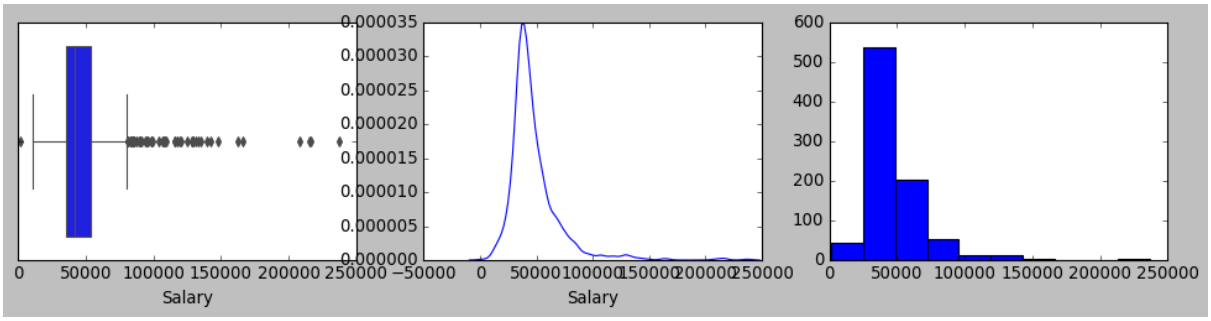
Below is the normalized distribution of the predicted variable (Holliday_Package).
No: 54.01%
Yes: 45.99%
The values across Yes and No are almost equally distributed.

Univariate analysis:
Below analysis did not count in "Holidday_Package" and "foreign" columns as they are categorical and non numeric.

1. Univariate analysis for Salary

Mean is 47729.172018, Median is 41903.500000, Mode(s) are 32197.0000

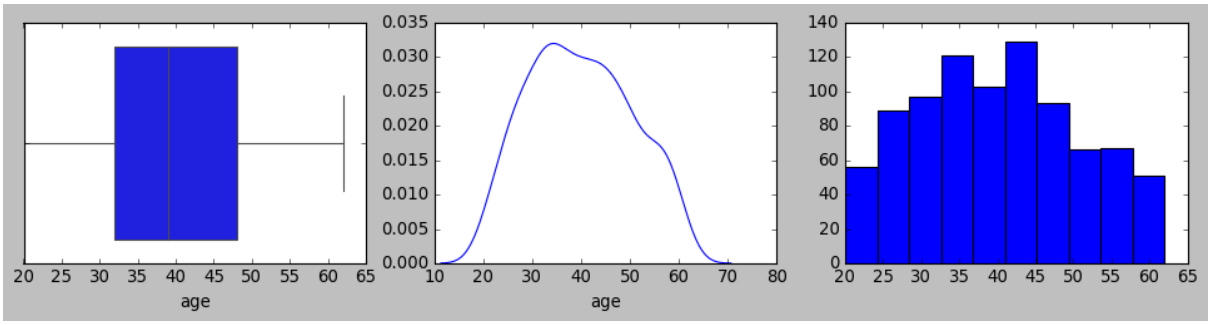
Column Salary has outliers



Column Salary is not normally distributed

2. Univariate analysis for age

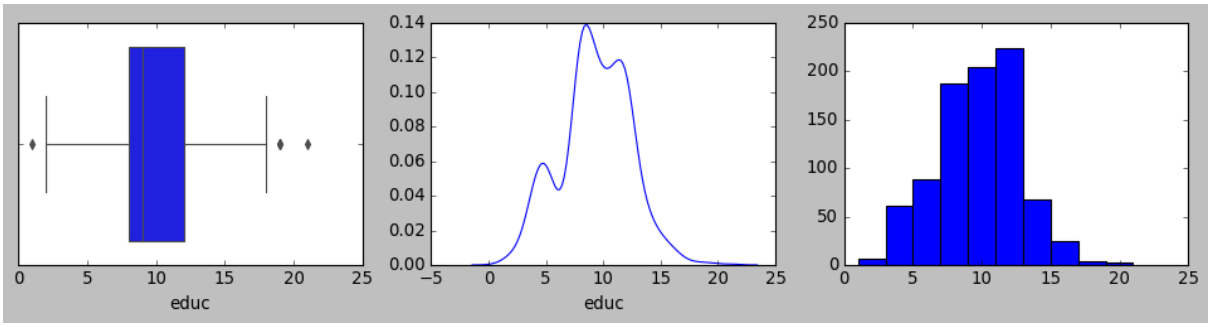
Mean is 39.955275, Median is 39.000000, Mode(s) are 44.0000
Column age does not have outliers



Column age is not normally distributed

3. Univariate analysis for educ

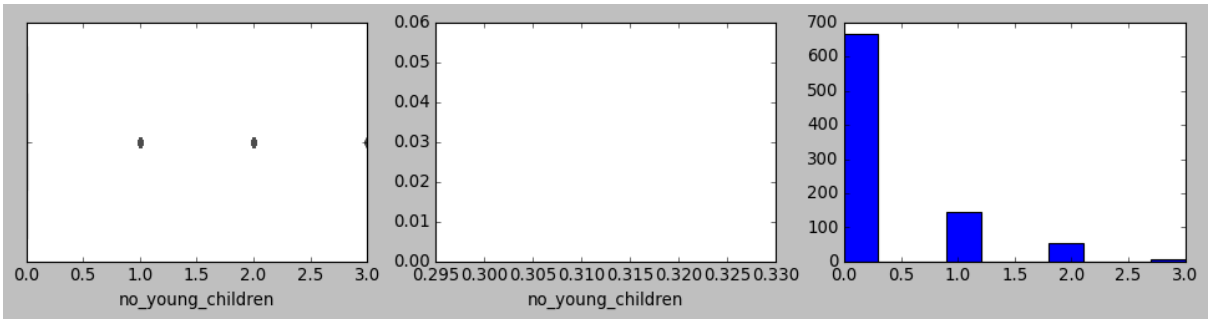
Mean is 9.307339, Median is 9.000000, Mode(s) are 8.0000
Column educ has outliers



Column educ is not normally distributed

4. Univariate analysis for no_young_children

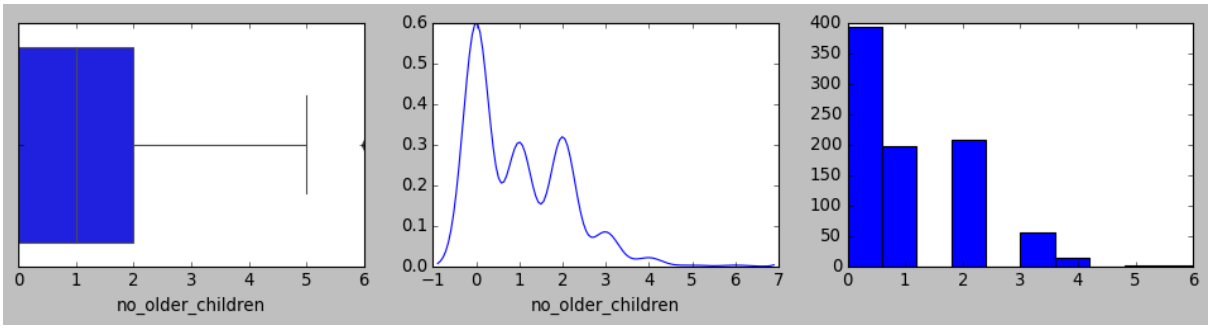
Mean is 0.311927, Median is 0.000000, Mode(s) are 0.0000
Column no_young_children has outliers



Column no_young_children is not normally distributed

5. Univariate analysis for no_older_children

Mean is 0.982798, Median is 1.000000, Mode(s) are 0.0000
Column no_older_children has outliers

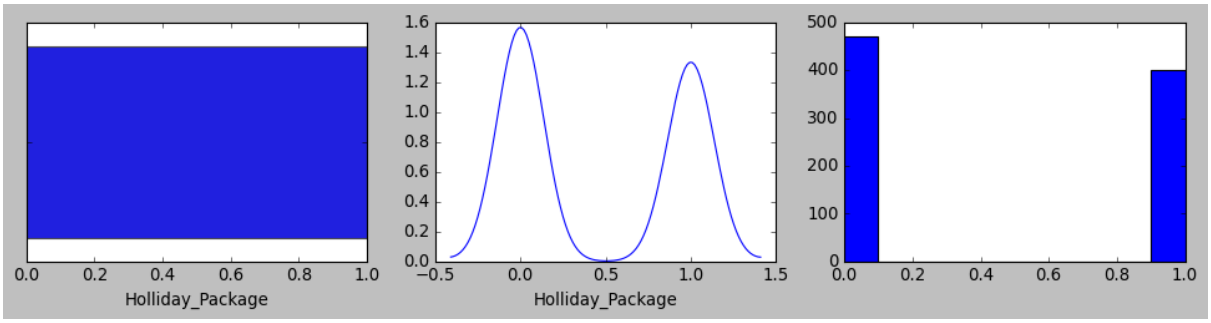


Column no_older_children is not normally distributed

Below analysis is done on the dataset after converting “Holidday_Package” into binary code (yes=1 and no=0) and “foreign” into dummies (foreign_yes and foreign_no).

1. Univariate analysis for Holliday_Package

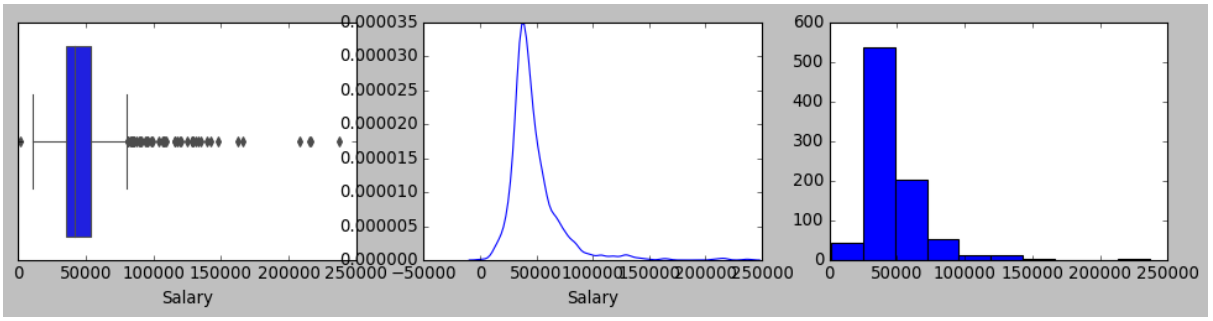
Mean is 0.459862, Median is 0.000000, Mode(s) are 0.0000
Column Holliday_Package does not have outliers



Column Holliday_Package is not normally distributed

2. Univariate analysis for Salary

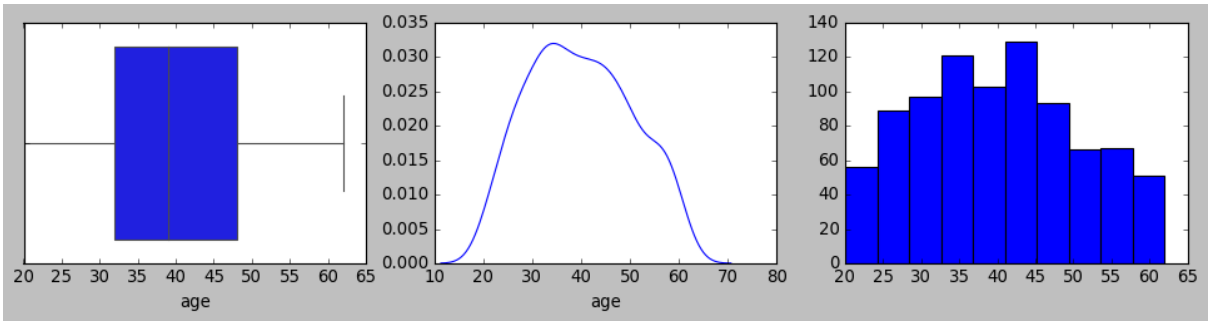
Mean is 47729.172018, Median is 41903.500000, Mode(s) are 32197.0000
Column Salary has outliers



Column Salary is not normally distributed

3. Univariate analysis for age

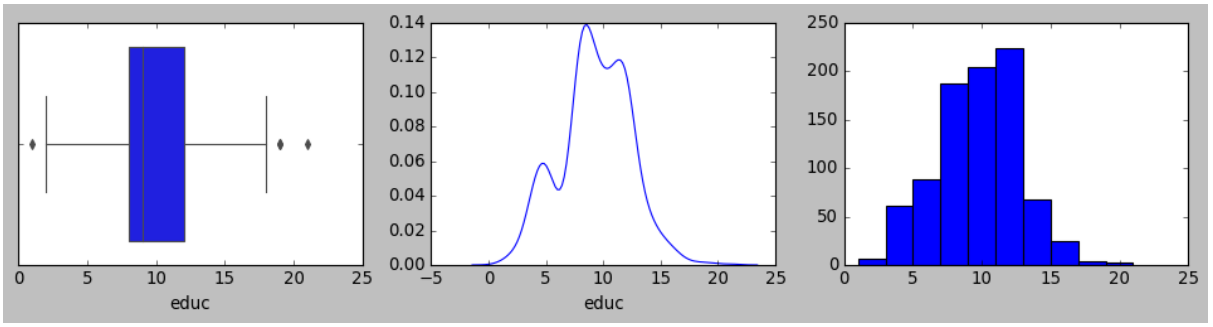
Mean is 39.955275, Median is 39.000000, Mode(s) are 44.0000
Column age does not have outliers



Column age is not normally distributed

4. Univariate analysis for educ

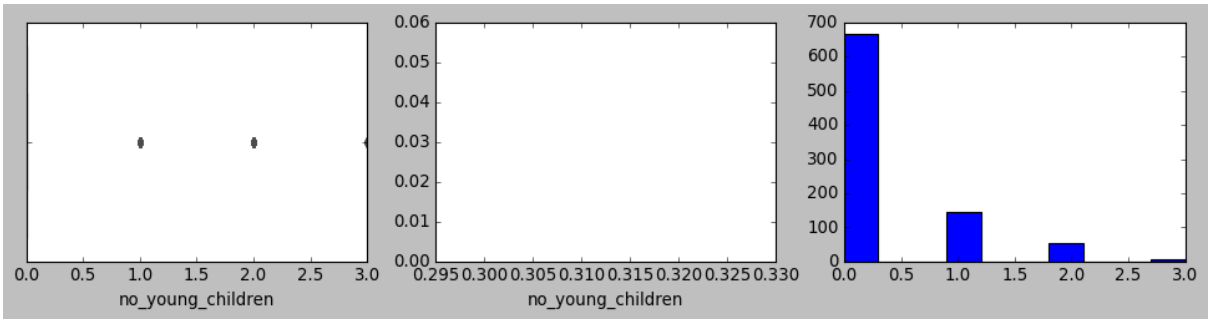
Mean is 9.307339, Median is 9.000000, Mode(s) are 8.0000
Column educ has outliers



Column educ is not normally distributed

5. Univariate analysis for no_young_children

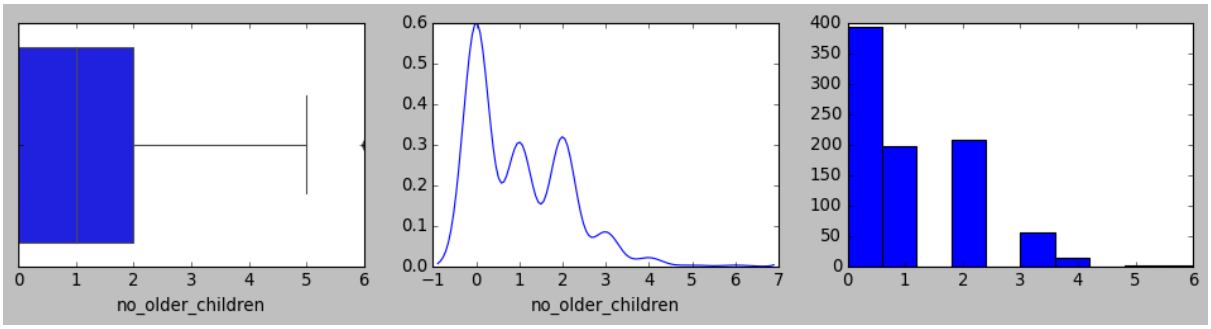
Mean is 0.311927, Median is 0.000000, Mode(s) are 0.0000
Column no_young_children has outliers



Column no_young_children is not normally distributed

6. Univariate analysis for no_older_children

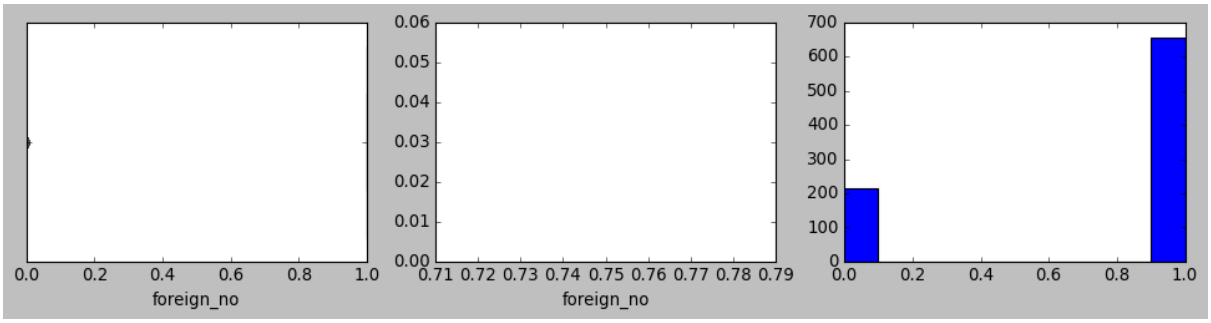
Mean is 0.982798, Median is 1.000000, Mode(s) are 0.0000
Column no_older_children has outliers



Column no_older_children is not normally distributed

7. Univariate analysis for foreign_no

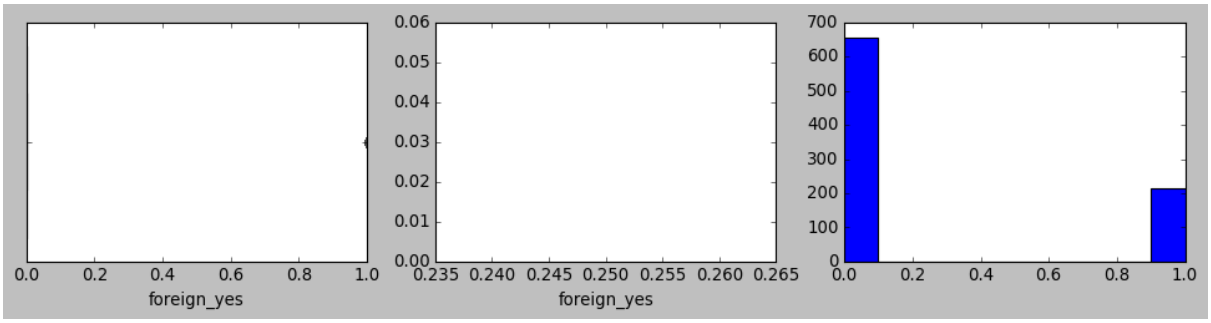
Mean is 0.752294, Median is 1.000000, Mode(s) are 1.0000



Column foreign_no is not normally distributed

8. Univariate analysis for foreign_yes

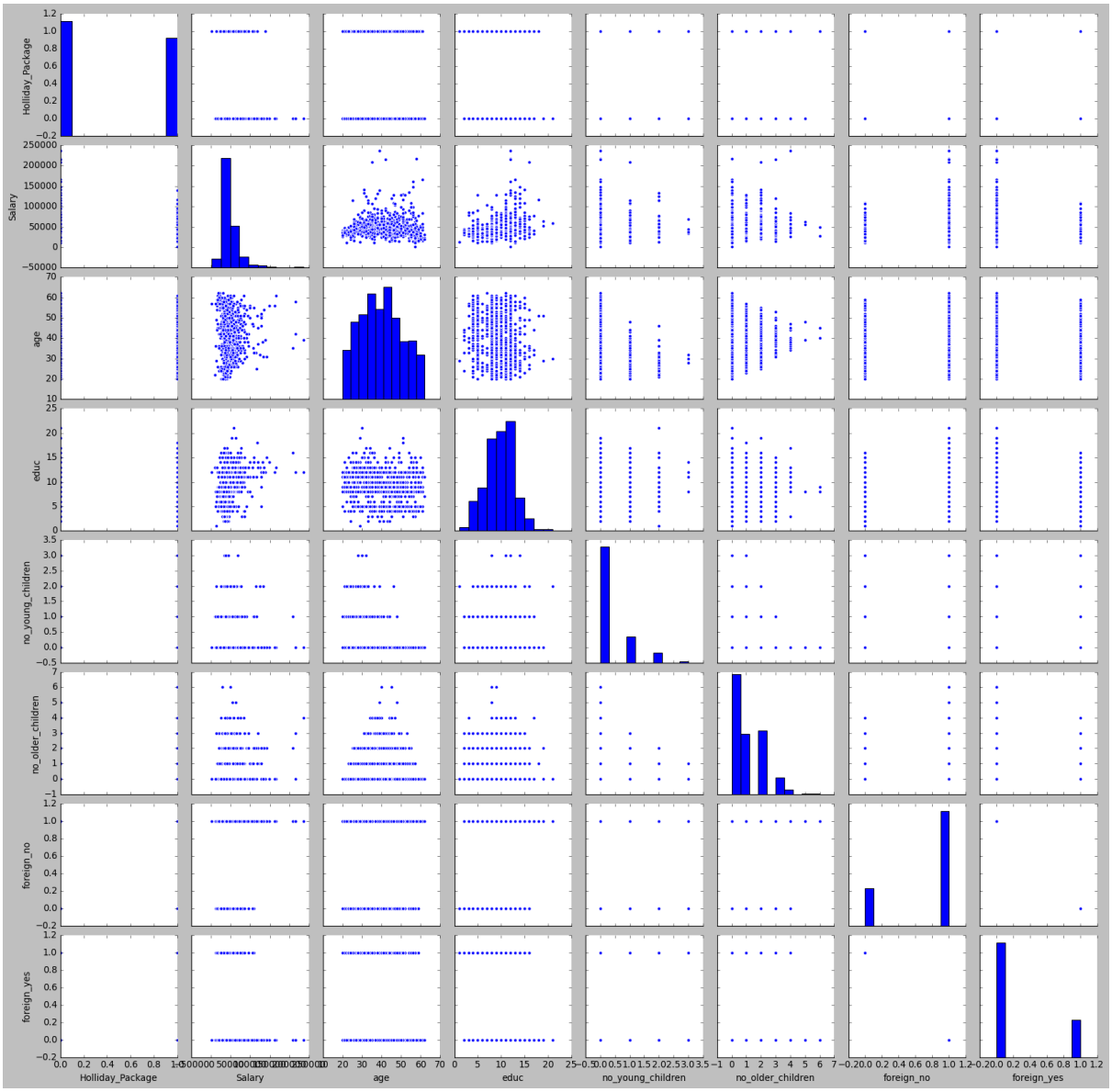
Mean is 0.247706, Median is 0.000000, Mode(s) are 0.0000
Column foreign_yes has outliers



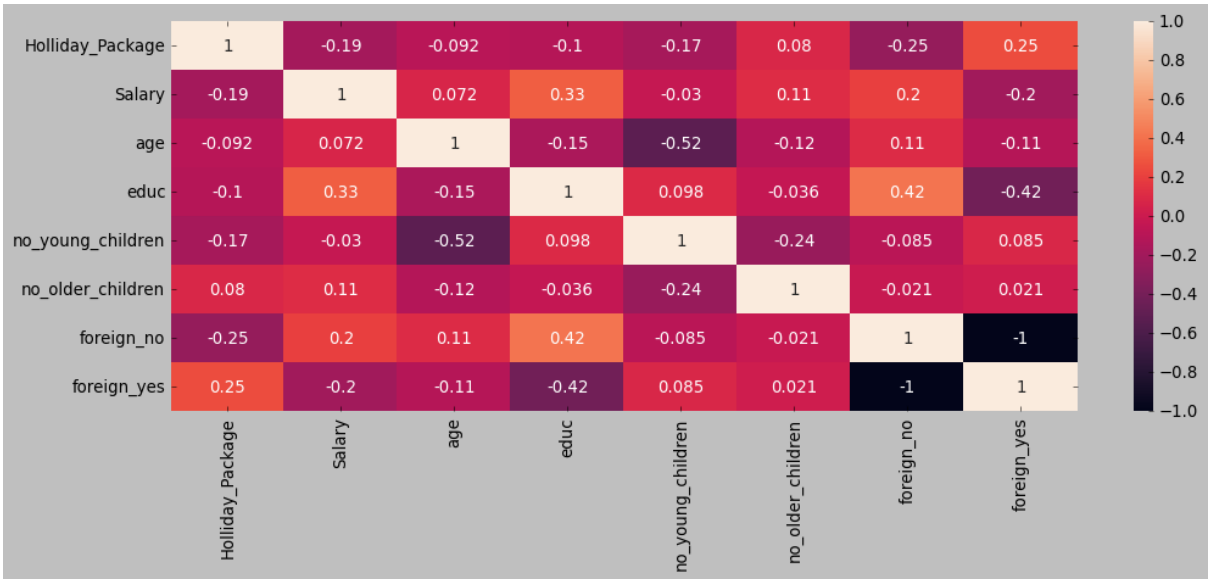
Column foreign_yes is not normally distributed

Bivariate analysis:

Please find below the pairplot for the original dataset



Please find below the heatmap for the original dataset



The above bi variate analysis clearly shows that there isn't a strong co relationship between many variables except for some traces of positive correlation observed between foreign(yes and no) and educ and some negative correlation between age and no_young_children.
Also the target variable (Holliday_Package) does not show a great deal of co relationship with other predictor variables while the predictor variables such as foreign(no and yes) followed by salary and no_young_children show minor negative co relationship with the predicted variable (target).

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).

As salary alone showed a good number of outliers beyond maximum whisker it has been treated to match the maximum whisker accordingly along with minimal outliers beyond minimum whisker also being imputed.

Please find below the unique values across categorical variables and the respective count of number of rows. Also they are converted into numeric columns by

```
HOLLIDAY_PACKAGE :  2 unique values
Yes:    401 rows
No :    471 rows

FOREIGN :  2 unique values
Yes:     216 rows
No :     656 rows
```

Also the categorical variable Holliday_Package has been type casted with binary values of 1 and 0 for 'yes' and 'no' respectively and hence making them align with reflecting on if the customer opted for vacation package(as 1) or not (as 0) respectively.

Also, the 'foreign' variable which does not have numerical significance has been converted to flag based (one hot encoding) variables with numeric binaries as their values across two

new variables namely foreign_yes and foreign_no depicting whether the customer is foreigner or not.

Please check below for the 5 number summary for the training data (70% of the sample) after the categorical object types have been converted to numeric as mentioned before.

	Salary	age	educ	no_young_children	no_older_children	foreign_no	foreign_yes
count	610	610	610	610	610	610	610
mean	46007.03	39.4557	9.372131	0.337705	0.9852	0.744262	0.255738
std	16024.82	10.4373	3.057341	0.64382	1.0542	0.436633	0.436633
min	8105.75	20	1	0	0	0	0
25%	35370.25	31	8	0	0	0	0
50%	42249.5	38	9	0	1	1	0
75%	53836.25	47	12	0	2	1	1
max	80687.75	62	21	3	5	1	1

Please check below for the 5 number summary for the testing data after the categorical object types have been converted to numeric as mentioned before.

	Salary	age	educ	no_young_children	no_older_children	foreign_no	foreign_yes
count	262	262	262	262	262	262	262
mean	44680.08	41.1183	9.156489	0.251908	0.9771	0.770992	0.229008
std	14903.91	10.7435	2.986934	0.530214	1.1611	0.420998	0.420998
min	14119	20	2	0	0	0	0
25%	35079.25	33	8	0	0	1	0
50%	40908.5	42	9	0	1	1	0
75%	52435.25	49	11	0	2	1	0
max	80687.75	62	19	3	6	1	1

Based on the regression model that has been built based on the above training and testing data set below are the statistical summary using the statsmodel library.

OLS Regression Results

Dep. Variable:	Holliday_Package	R-squared:	0.163
Model:	OLS	Adj. R-squared:	0.154
Method:	Least Squares	F-statistic:	19.54
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	6.86e-21
Time:	22:40:24	Log-Likelihood:	-386.65
No. Observations:	610	AIC:	787.3
Df Residuals:	603	BIC:	818.2
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.7917	0.084	9.404	0.000	0.626	0.957
Salary	-4.091e-06	1.28e-06	-3.198	0.001	-6.6e-06	-1.58e-06
age	-0.0123	0.002	-5.574	0.000	-0.017	-0.008
educ	0.0121	0.007	1.667	0.096	-0.002	0.026
no_older_children	-0.0096	0.019	-0.494	0.621	-0.048	0.029
no_young_children	-0.2714	0.036	-7.473	0.000	-0.343	-0.200
foreign_yes	0.5310	0.044	12.155	0.000	0.445	0.617
foreign_no	0.2607	0.053	4.953	0.000	0.157	0.364

Omnibus:	33031.174	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	51.875
Skew:	0.166	Prob(JB):	5.44e-12
Kurtosis:	1.610	Cond. No.	2.57e+20

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.19e-29. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Based on the null hypothesis that states that the independent variables in the sample have no co relationship with target variable from the universe, please note the coefficients of predictor variables such as ‘no_older_children’ and ‘educ’ from the above statistical summary with p_value above significance value of 5% states that null will fly. This means those two variables cannot be part of predictor variables. However for the rest of the predictor variables such as Salary, age, no_young_children, foreign_yes and foreign_no the p_values is 0% and hence null hypothesis is rejected, which means their samples does not reflect the universe and potentially they are co related to the target variable “Holliday_Package”.

Please find below the revised statistical summary post to the removal of “no_older_children” and “educ” as predictor variables. We are also proceeding for logistical regression and linear discriminant analysis model building by removing these predictor variables from the training and testing datasets.

OLS Regression Results

Dep. Variable:	Holliday_Package	R-squared:	0.158
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	28.45
Date:	Sat, 24 Oct 2020	Prob (F-statistic):	1.12e-21
Time:	22:35:12	Log-Likelihood:	-388.28
No. Observations:	610	AIC:	786.6
Df Residuals:	605	BIC:	808.6
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8415	0.069	12.247	0.000	0.707	0.976
Salary	-3.527e-06	1.2e-06	-2.945	0.003	-5.88e-06	-1.18e-06
age	-0.0125	0.002	-6.015	0.000	-0.017	-0.008
no_young_children	-0.2606	0.034	-7.746	0.000	-0.327	-0.194
foreign_yes	0.5401	0.040	13.526	0.000	0.462	0.619
foreign_no	0.3014	0.042	7.247	0.000	0.220	0.383

Omnibus:	15088.996	Durbin-Watson:	2.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	52.756
Skew:	0.172	Prob(JB):	3.50e-12
Kurtosis:	1.601	Cond. No.	1.08e+20

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.25e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.

Based on the revised predictor variables yielding above coefficients from statistical summary we could reject null hypothesis completely for the predictor variables Salary, age, no_young_children, foreign_yes and foreign_no predictor variables to proceed with resulting linear equation. However the coefficients of this linear equation is subject to the risk of multi collinearity and hence we check on the VIF for each of the predictor variables to check if they stay within the range of 1 to 5. Please find below the VIF accordingly:

```
VIF for Salary                : 1.061323780479254
VIF for age                   : 1.3767750106591812
VIF for no_young_children     : 1.370513314562445
VIF for foreign_yes           : 7.9420154825877445
VIF for foreign_no            : 25.500575817690464
```

Based on further analysis from the above VIF we could realize that foreign_no and Salary had to be removed in that sequence to bring the VIF within acceptable range as below:

```
VIF for age                   : 1.3291752403875914
VIF for no_young_children     : 1.1486607869453156
VIF for foreign_yes           : 1.3032977526674505
```

Accordingly the revised summary from statsmodel library for OLS regression is as below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Holliday_Package   R-squared:                       0.146
Model:                           OLS   Adj. R-squared:                  0.142
Method:                Least Squares   F-statistic:                     34.60
Date:               Sun, 25 Oct 2020   Prob (F-statistic):           1.19e-20
Time:                       12:48:18   Log-Likelihood:                -392.62
No. Observations:                610   AIC:                             793.2
Df Residuals:                    606   BIC:                             810.9
Df Model:                          3
Covariance Type:           nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept            0.9820      0.092     10.697      0.000       0.802       1.162
age                 -0.0127      0.002     -6.104      0.000      -0.017      -0.009
no_young_children   -0.2577      0.034     -7.615      0.000      -0.324      -0.191
foreign_yes          0.2685      0.043      6.246      0.000       0.184       0.353
==============================================================================
Omnibus:                     8034.241   Durbin-Watson:                   2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               54.903
Skew:                           0.173   Prob(JB):                     1.20e-12
Kurtosis:                       1.571   Cond. No.                         206.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Based on the above summary it can be observed the coefficient of determinant (R squared) is slightly lesser than the original summary built before removing the multi collinearity. Higher the R squared value better the model is.

However, we shall proceed with running logistic regression and linear determinant model towards evaluating the model scores as per the built training and test data set and compare the efficiency of these two models.

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.

Please check below for model accuracy, confusion matrix, ROC curve and AUC score for training and testing data across logistic regression and linear discriminant analysis models.

Model accuracy:

Model accuracy	Training	Testing
Logistic Regression	0.66393443	0.65267176
LDA	0.65737705	0.65267176

Confusion matrix

Logistic regression		
Training dataset	Predicted Negative	Predicted Positive
Actual Negative	240	89
Actual Negative	116	165
Testing dataset	Predicted Negative	Predicted Positive
Actual Negative	107	35
Actual Negative	56	64

LDA		
Training dataset	Predicted Negative	Predicted Positive
Actual Negative	245	84
Actual Negative	125	156
Testing dataset	Predicted Negative	Predicted Positive
Actual Negative	110	32
Actual Negative	59	61

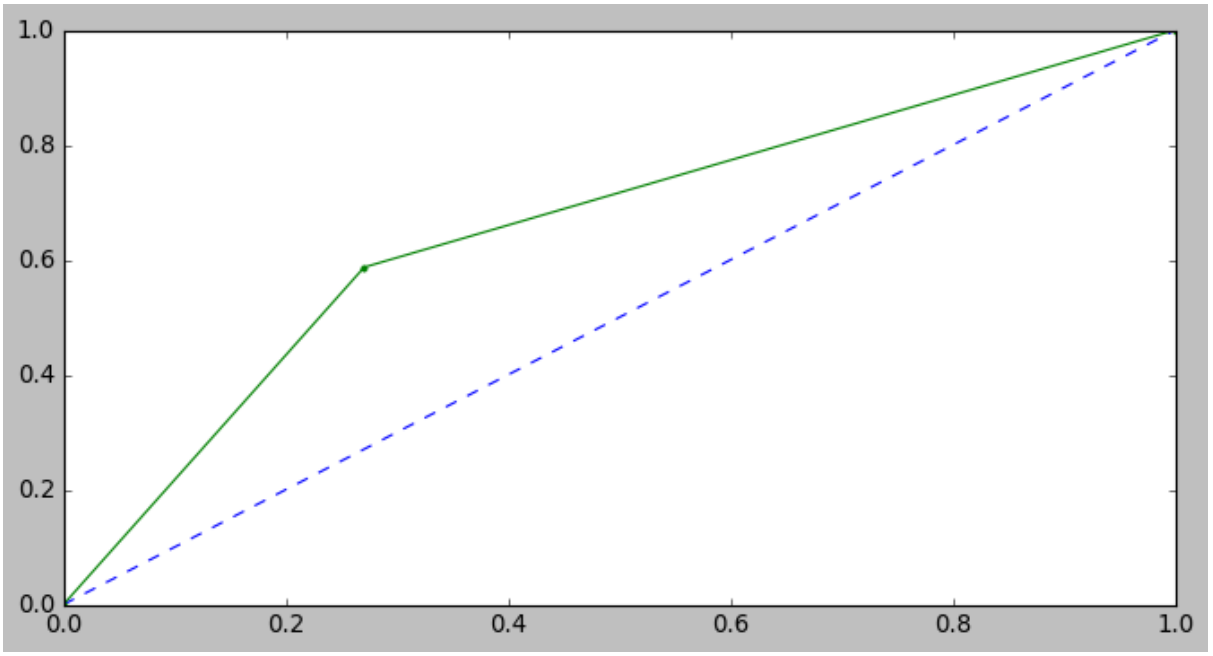
Classification report:

Logistic regression				
classification report for training data set:				
	precision	recall	f1-score	support
0	0.67	0.73	0.7	329
1	0.65	0.59	0.62	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610
classification report for testing data set:				
	precision	recall	f1-score	support
0	0.66	0.75	0.7	142
1	0.65	0.53	0.58	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.65	262

LDA				
classification report for training data set:				
	precision	recall	f1-score	support
0	0.66	0.74	0.7	329
1	0.65	0.56	0.6	281
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.65	610
classification report for testing data set:				
	precision	recall	f1-score	support
0	0.65	0.77	0.71	142
1	0.66	0.51	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.65	262

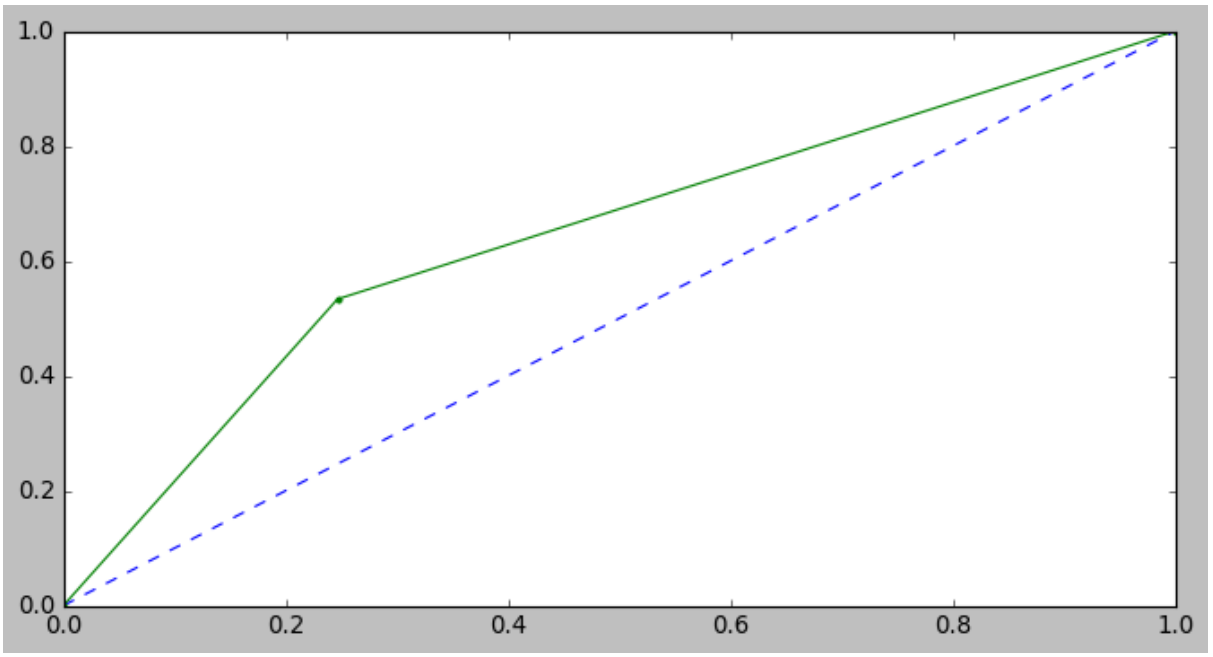
ROC/AUC from Logistic regression model:

ROC curve for training data




```
AUC for the Training Data: 0.658
```

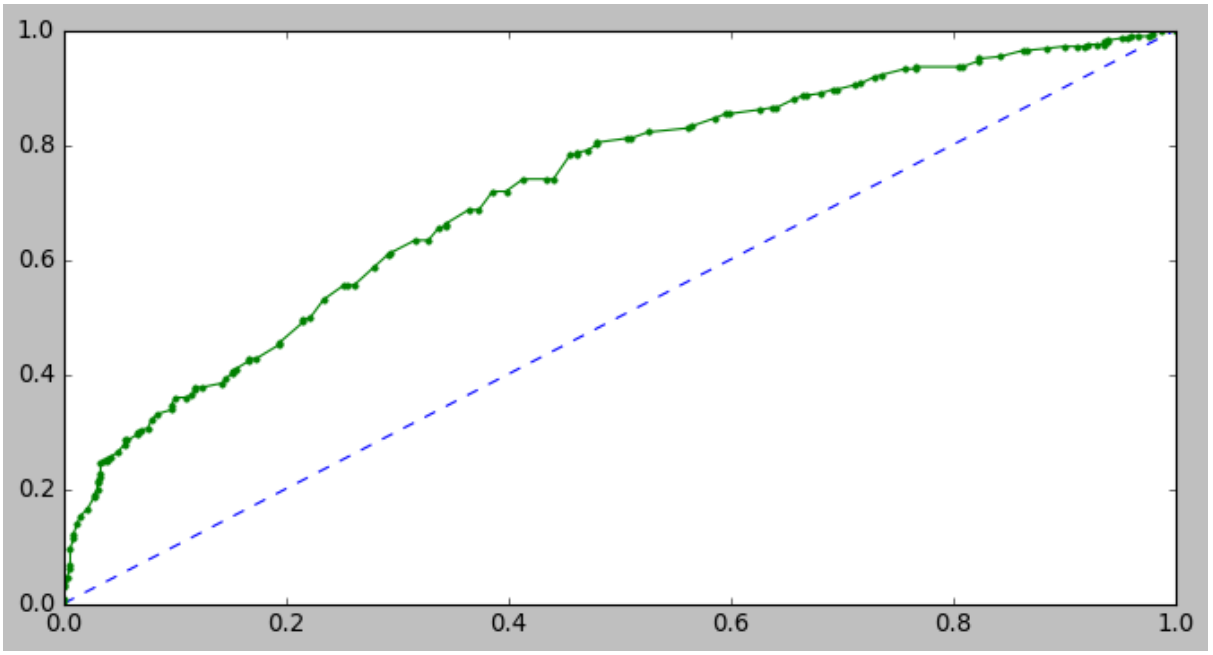
ROC curve for testing data




```
AUC for the Test Data: 0.643
```

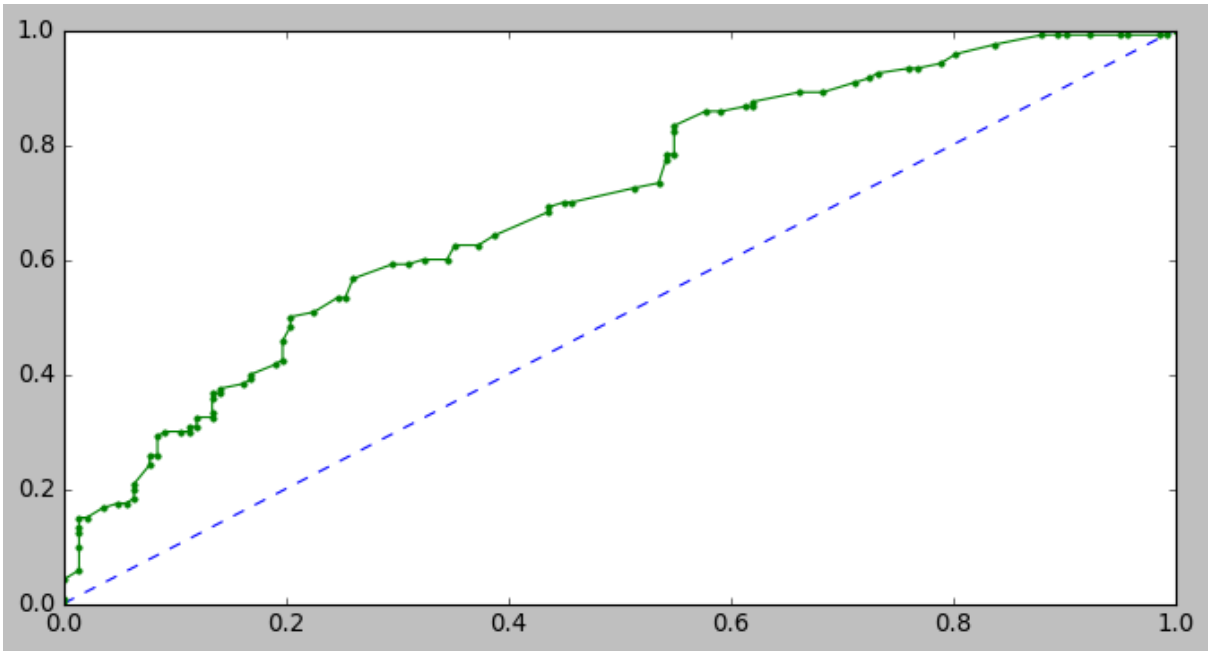
ROC/AUC from LDA:

ROC curve for training data




```
AUC for the Training Data: 0.720
```

ROC curve for testing data




```
AUC for the Test Data: 0.698
```

Accordingly, the below linear equation has been extracted upon which logistic regression model has applied sigmoid function to transform the raw numbers returned by the equation to probability to form the S curve wherein the probability ranges between 0 and 1.

```
(0.98) * Intercept + (-0.01) * age + (-0.26) * no_young_children +
(0.27) * foreign_yes
```

When we compare the performance of the model across accuracy scores and the classification report, we can notice that there isn't much difference on the performance across Logistic regression and LDA from the tuned models. There seems to be a very minor efficiency difference between the accuracy scores from the fact that logistic regression testing data outcome performed closer to the training outcome when compared to LDA.

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

Based on the analysis done on the data, it has been seen that Holliday_Package does not have a noticeable co relationship with any of the predictor variables. However, the distribution of Holliday_Package between customers who chose the vacation versus not is almost equal and hence the model has enough data to classify them well.

From an interpretation and insight perspective, there were few multicollinearity observed in the presence of Salary and non foreigners while education and number of older children data from the sample did not reflect the real data from the universe. This lead to limiting the predictive power of Holliday_Package to three aspects of the customer information on if it's a foreign customer and based on number of young children they had and the age they belong to. Out of them, foreign customers has the highest weightage in being able to sign up for the holiday package while number of young children had the next highest weightage clarifying that customer with lesser young children we able to choose the vacation package. There is also a slight tendency that customers signing up increases when the age decreases meaning mostly younger customers out of the lot has higher tendency to sign up for the vacation. This indicates potential customers could be parents of teenage students who are young adults while the parents themselves could be young enough within the age group of earning potentials coming from abroad.

Hence the recommendation would be to focus on offering family package deals with discounts during school or college holidays. This could enable upselling with that customer segment. Alternatively, for customers with younger children new vacation plans can be rolled out that are children friendly where they can spend more quality time in great resorts located near children attractions with reliable child care services that can generate interest level for that customer segment and hence provide increased market share opportunities.