* 1. **Question: Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA). Perform Univariate and Bivariate Analysis.**

**EDA:**

Below provides five number summary 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.



Data types of the independent variables



Please note that the predicted variable is numeric and continuous as expected.

Shape:

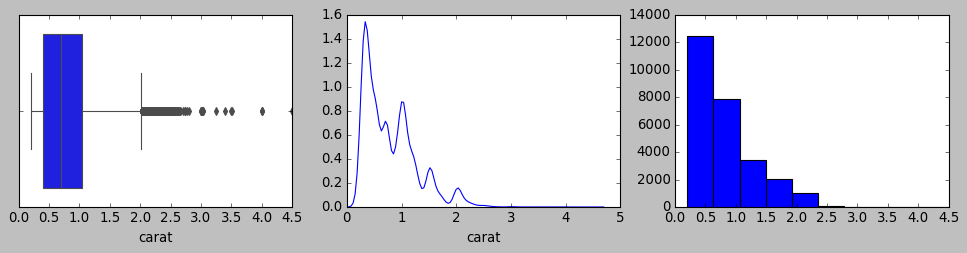
26967 rows and 10 columns

Univariate analysis:

1. Univariate analysis for carat

Mean is 0.798375, Median is 0.700000, Mode(s) are 0.3000

Column carat has outliers

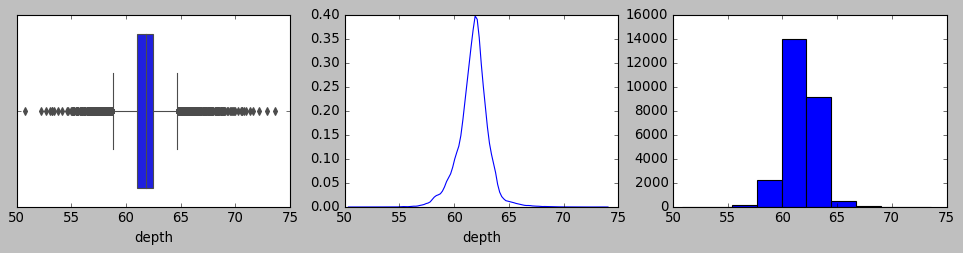


Column carat is not normally distributed

2. Univariate analysis for depth

Mean is 61.745147, Median is nan, Mode(s) are 62.0000

Column depth has outliers

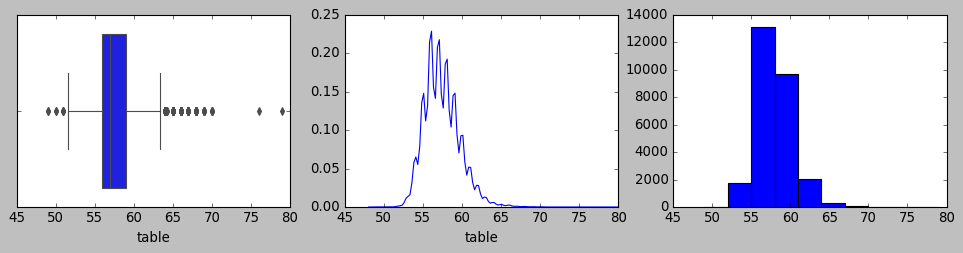


Column depth is normally distributed

3. Univariate analysis for table

Mean is 57.456080, Median is 57.000000, Mode(s) are 56.0000

Column table has outliers

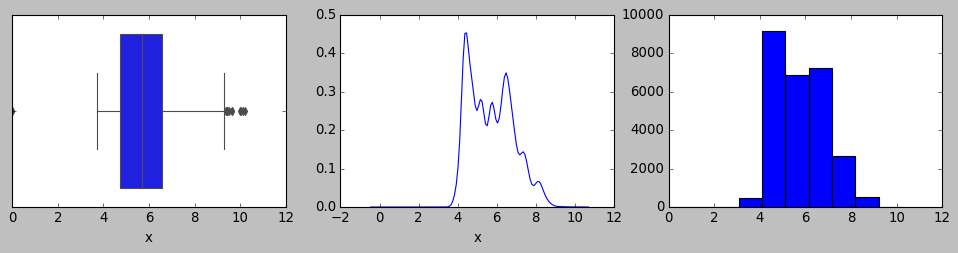


Column table is not normally distributed

4. Univariate analysis for x

Mean is 5.729854, Median is 5.690000, Mode(s) are 4.3800

Column x has outliers

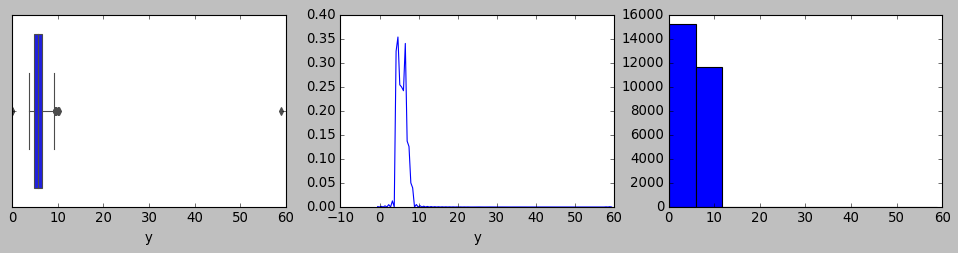


Column x is not normally distributed

5. Univariate analysis for y

Mean is 5.733569, Median is 5.710000, Mode(s) are 4.3500

Column y has outliers

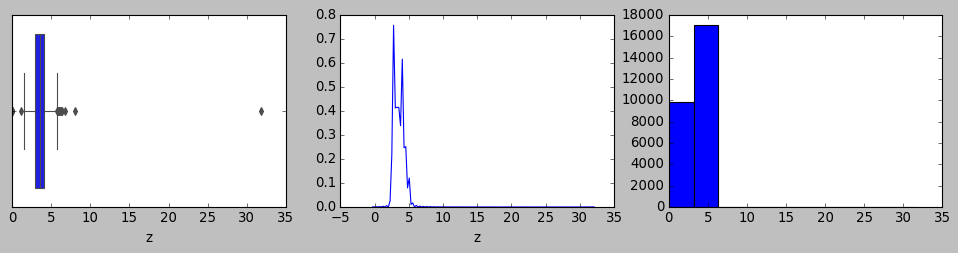


Column y is not normally distributed

6. Univariate analysis for z

Mean is 3.538057, Median is 3.520000, Mode(s) are 2.6900

Column z has outliers

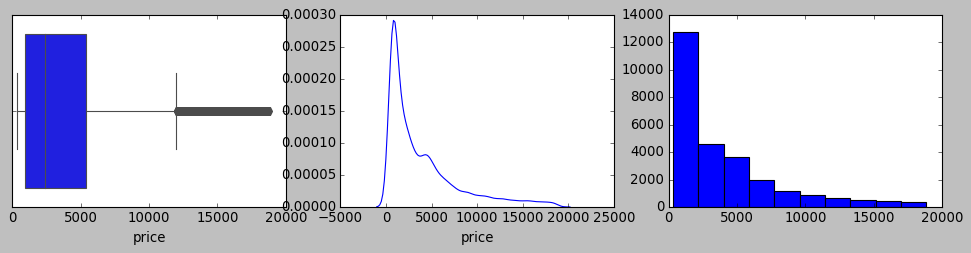


Column z is not normally distributed

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

Going by 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.

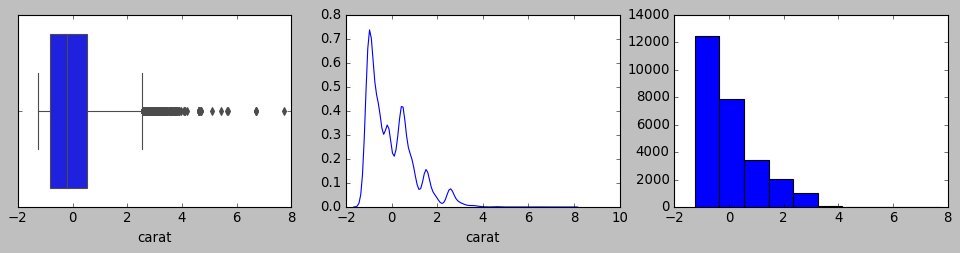
Z scoring based scaling of data would change the co efficient ,neutralize/remove the intercept while the accuracy score remains the same before and after. MSE would get scaled too.

Accordingly, below is the univariate analysis for the dataset that is standardized on z scores.Please note 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. Further explanation on this will be shared in Question 1.3.

1. Univariate analysis for carat

Mean is -0.000000, Median is -0.205920, Mode(s) are -1.0432

Column carat has outliers

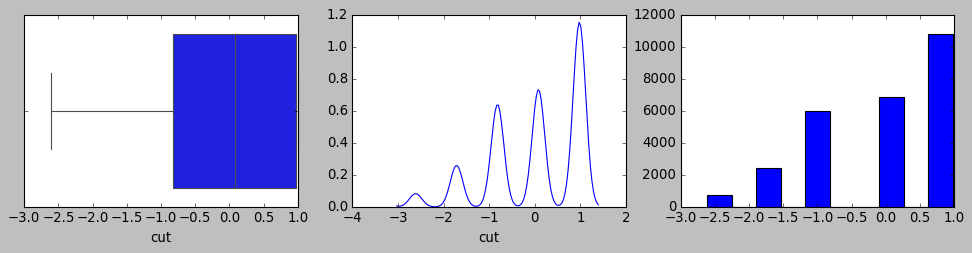


Column carat is not normally distributed

2. Univariate analysis for cut

Mean is 0.000000, Median is 0.081246, Mode(s) are 0.9796

Column cut does not have outliers

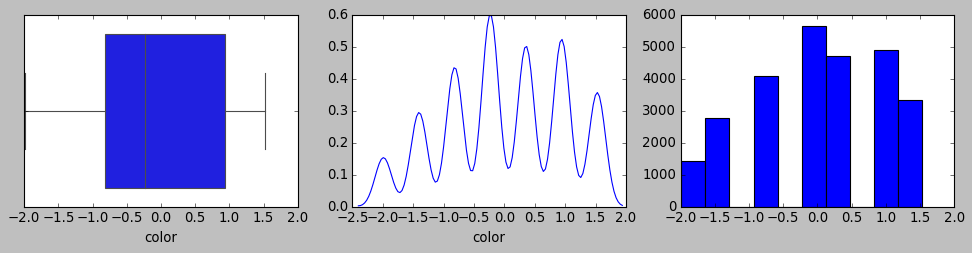


Column cut is not normally distributed

3. Univariate analysis for color

Mean is 0.000000, Median is -0.230890, Mode(s) are -0.2309

Column color does not have outliers

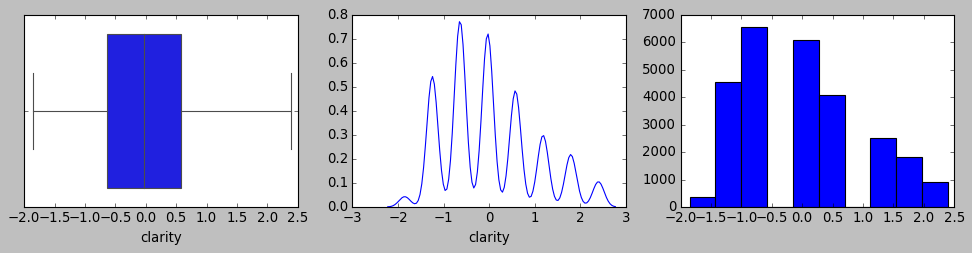


Column color is not normally distributed

4. Univariate analysis for clarity

Mean is 0.000000, Median is -0.032241, Mode(s) are -0.6394

Column clarity does not have outliers

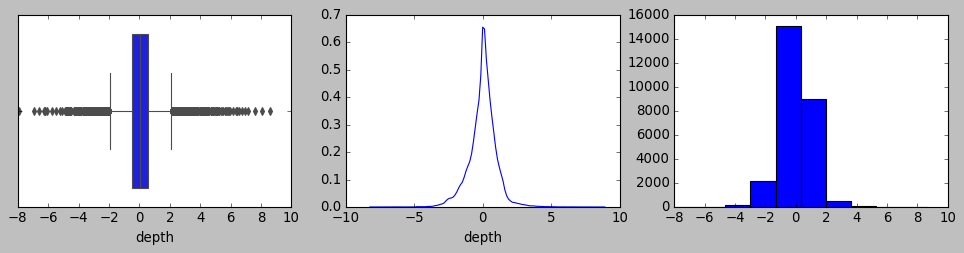


Column clarity is not normally distributed

5. Univariate analysis for depth

Mean is 0.000000, Median is 0.036839, Mode(s) are 0.0368

Column depth has outliers

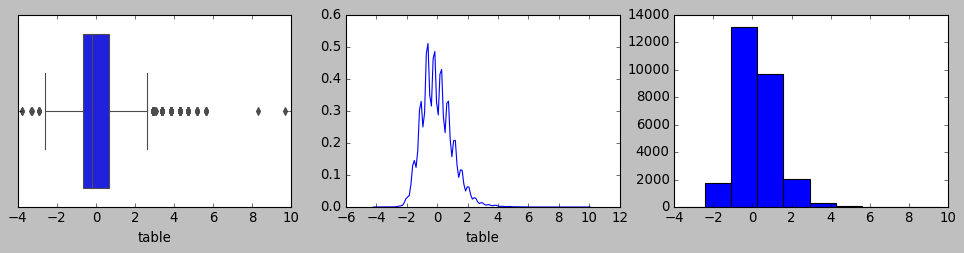


Column depth is not normally distributed

6. Univariate analysis for table

Mean is -0.000000, Median is -0.204334, Mode(s) are -0.6524

Column table has outliers

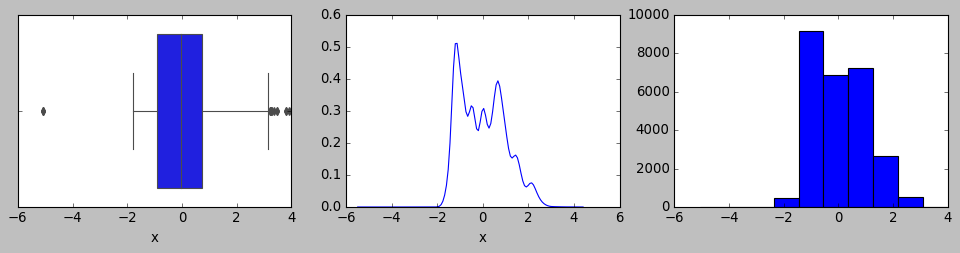


Column table is not normally distributed

7. Univariate analysis for x

Mean is 0.000000, Median is -0.035316, Mode(s) are -1.1962

Column x has outliers

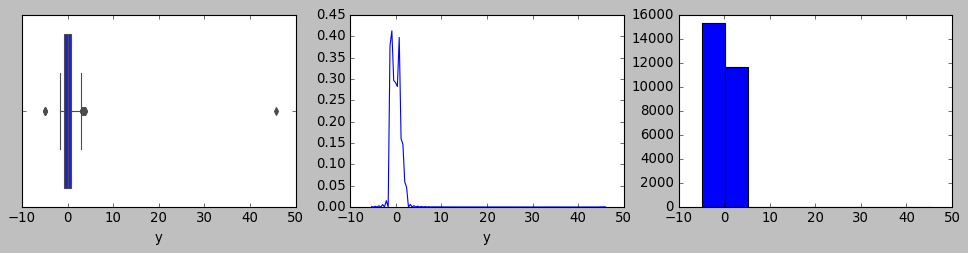


Column x is not normally distributed

8. Univariate analysis for y

Mean is -0.000000, Median is -0.020213, Mode(s) are -1.1866

Column y has outliers

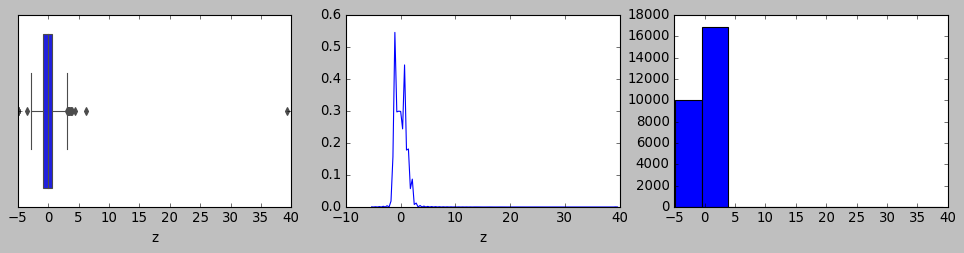


Column y is not normally distributed

9. Univariate analysis for z

Mean is -0.000000, Median is -0.025058, Mode(s) are -1.1769

Column z has outliers

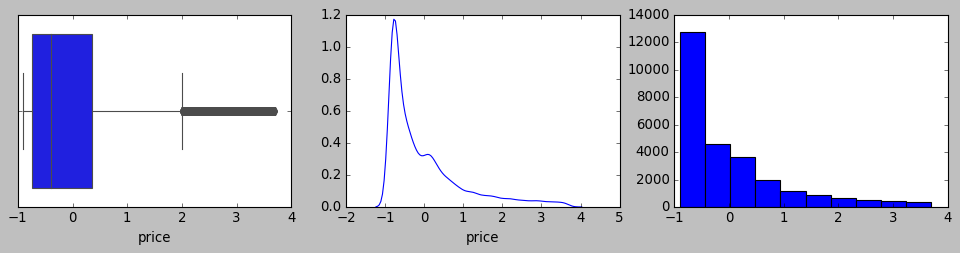


Column z is not normally distributed

10. Univariate analysis for price

Mean is -0.000000, Median is -0.388720, Mode(s) are -0.8437

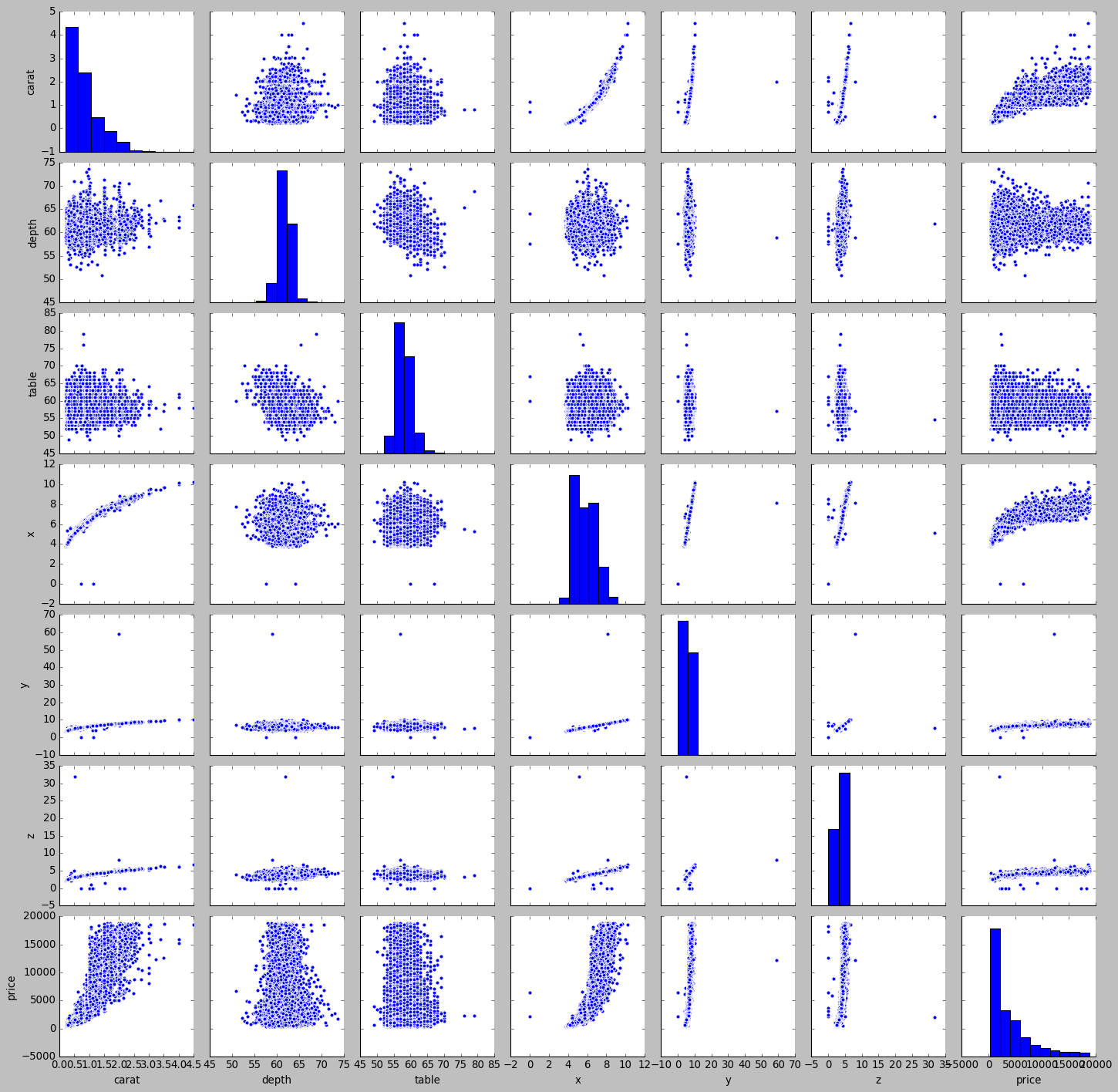
Column price has outliers



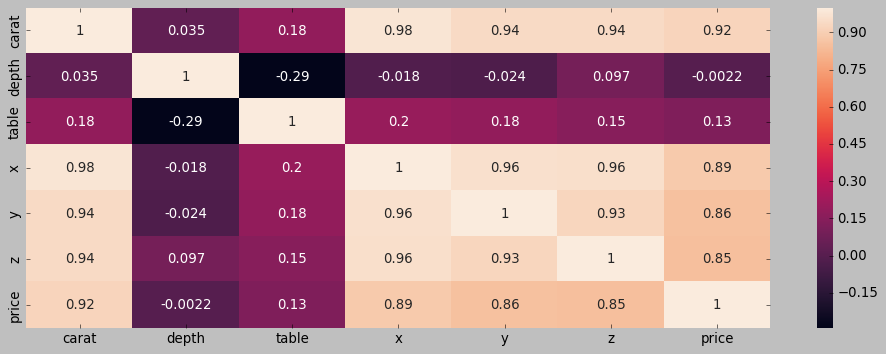
Column price is not normally distributed

Bi variate analysis:

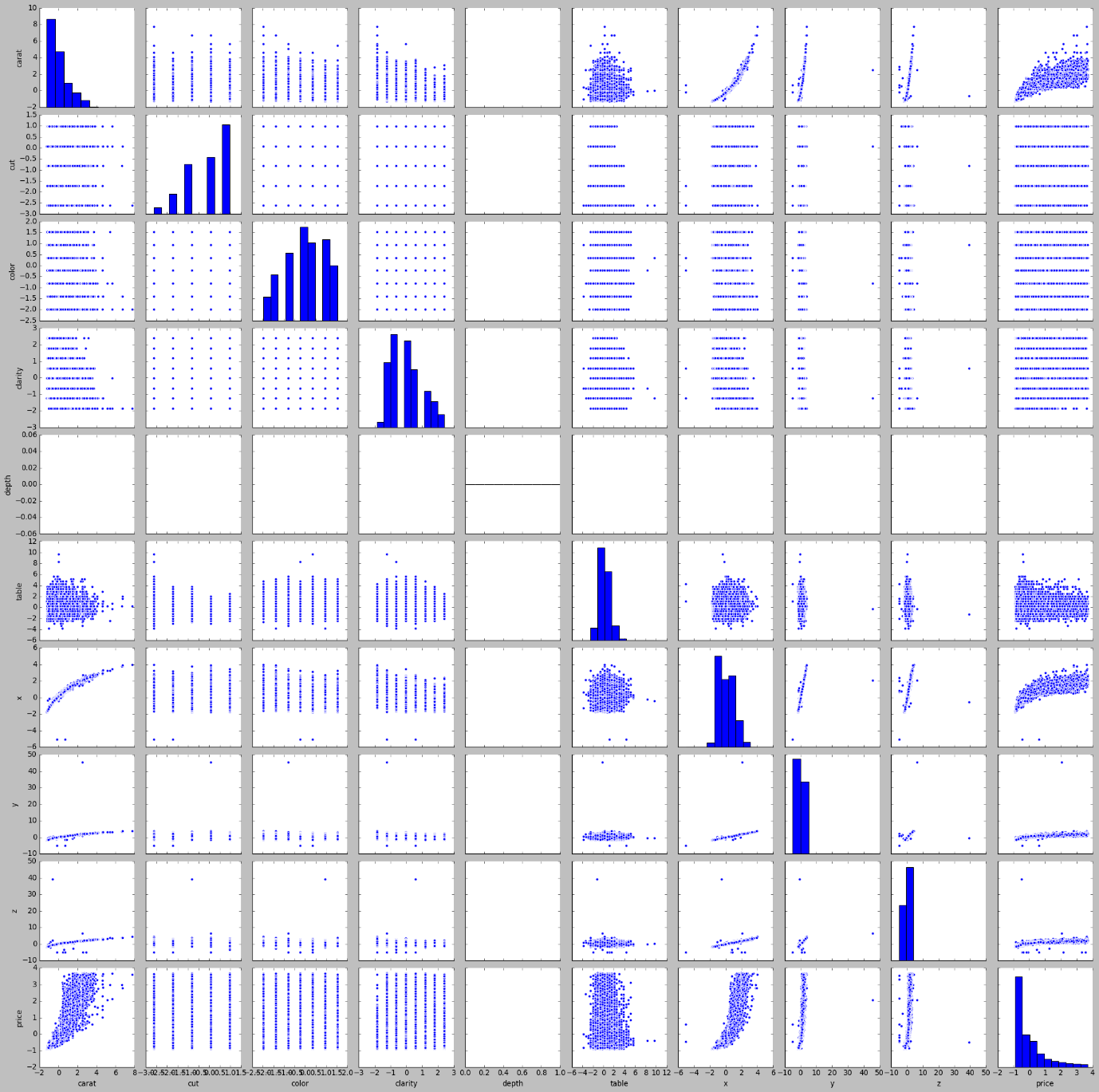
Pair plot for unscaled dataset:



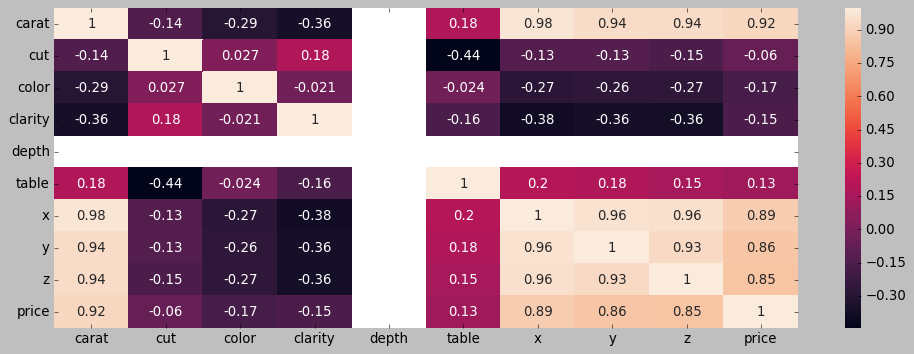
Heatmap for unscaled dataset:



Pair plot for scaled dataset:



Heatmap for scaled dataset:



Observations:

We could notice that price (target variable) depicts great corelation among carat, X, Y and Z independent variables across all combinations of them.

We also have to further observe if these variables are potential contributor for multi collinearity which is not good for explaining the coefficients. However, from 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 have 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 has zeros as its value in the given dataset. Accordingly X, Y and Z are the variables that contain zeros across one or more rows.



Below are the 9 rows that has 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 effect as other variables could potentially contribute with better coefficients in predicting price (target variable) which shall not be missed.



**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 find below 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 requires 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 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 outlier beyond maximum whisker while depth, table, X, Y, Z and price are treated for outliers beyond both side of the whisker. Below is the five number summary post the dataset has been treated for the outlier.



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

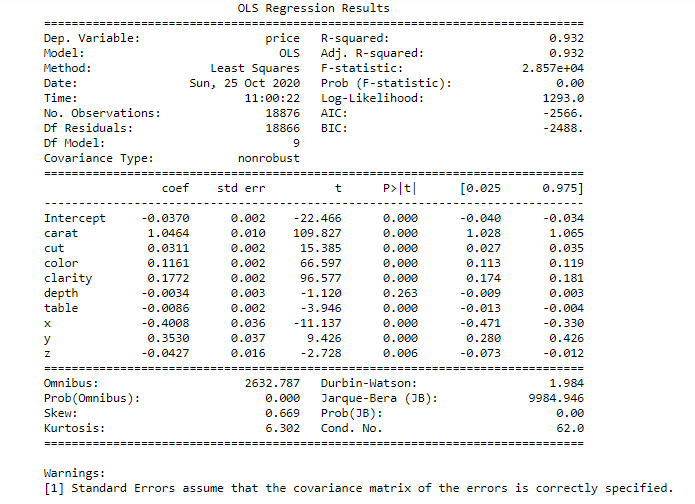
Train data:



Test data:



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.



Based on the null hypothesis that states that the independent variables have no co relationship with target variable 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 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

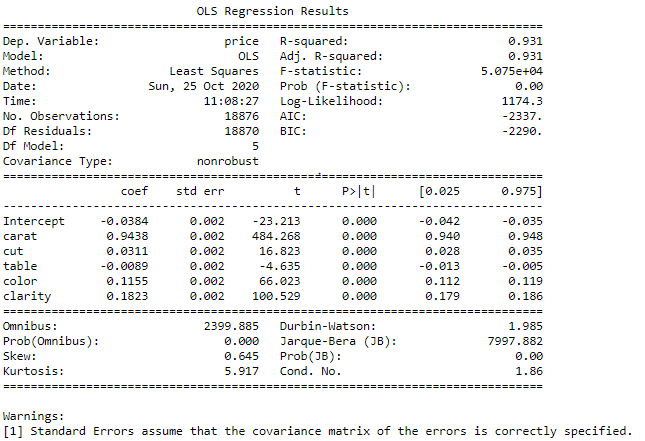
VIF for carat :31.360238644953316

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.



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 at the optimized set of predictor variables to make the linear equation to derive their coefficients.

Also the adjusted R square value (co efficient 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 co efficient 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 co efficients making the above equation it can be understand that carat, x and y has the highest weight in that order towards predicting price with x having negative co efficient. Color and clarity also have reasonable co efficients that contributes to the predicting power for price.

So when carat increases by 1 unit price increases by 0.94 unit while when cut increase 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.**

**EDA:**

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 summary 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.

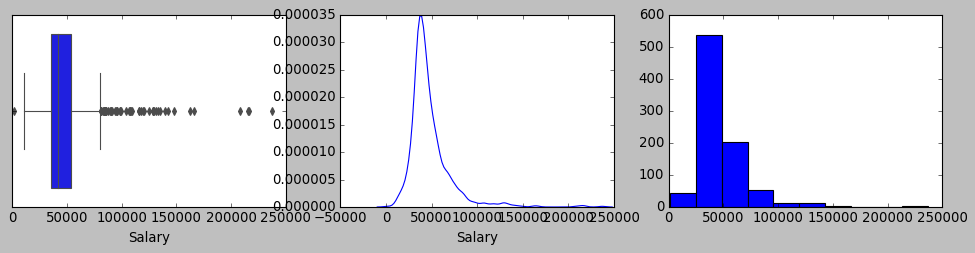
Univariate analysis:

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

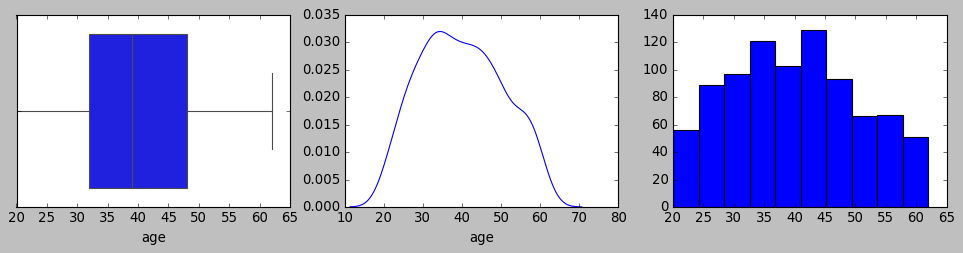


Column Salary is not normally distributed

2. Univariate analysis for age

Mean is 39.955275, Median is 39.000000, Mode(s) are 44.0000

Column age does not have outliers

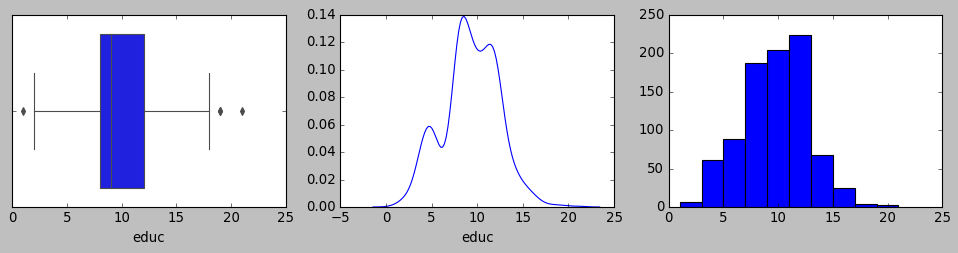


Column age is not normally distributed

3. Univariate analysis for educ

Mean is 9.307339, Median is 9.000000, Mode(s) are 8.0000

Column educ has outliers

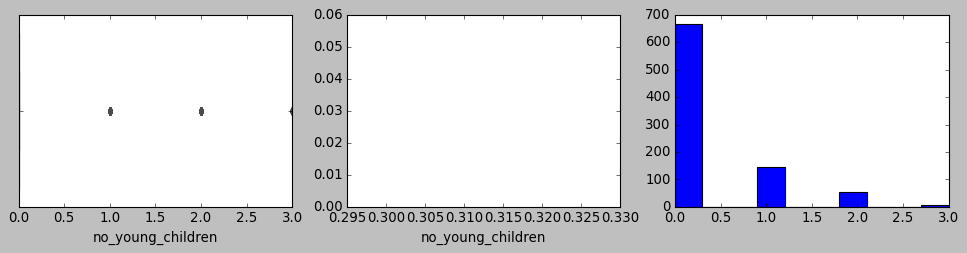


Column educ is not normally distributed

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

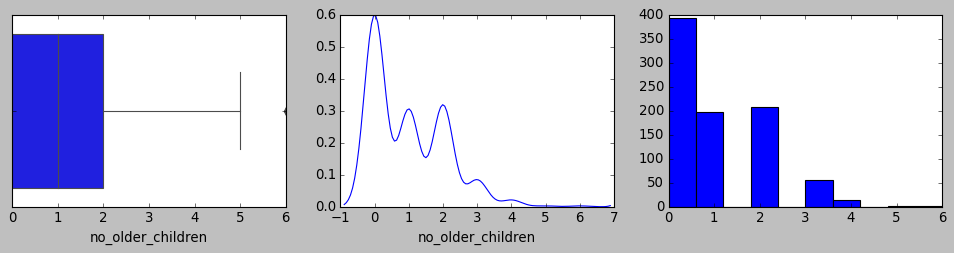


Column no\_young\_children is not normally distributed

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



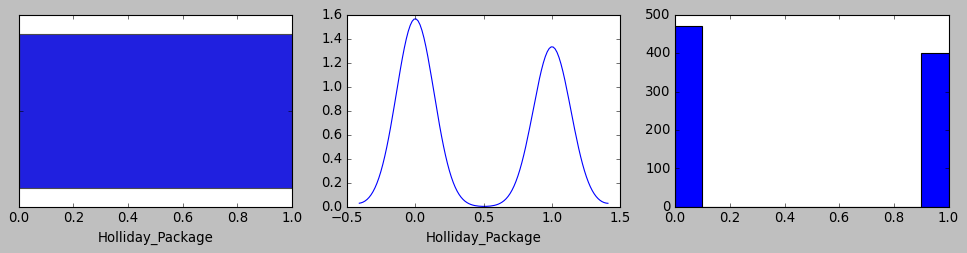
Column no\_older\_children is not normally distributed

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

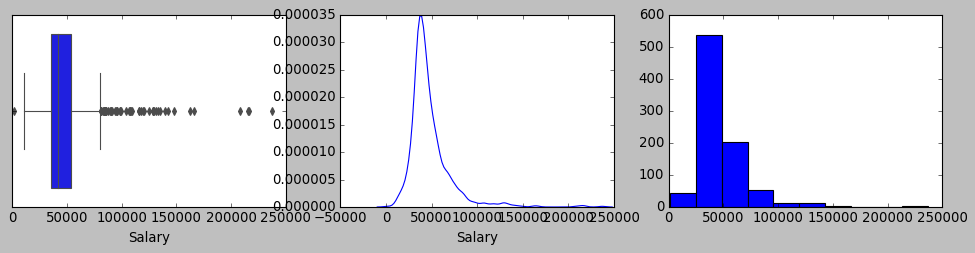


Column Holliday\_Package is not normally distributed

2. Univariate analysis for Salary

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

Column Salary has outliers

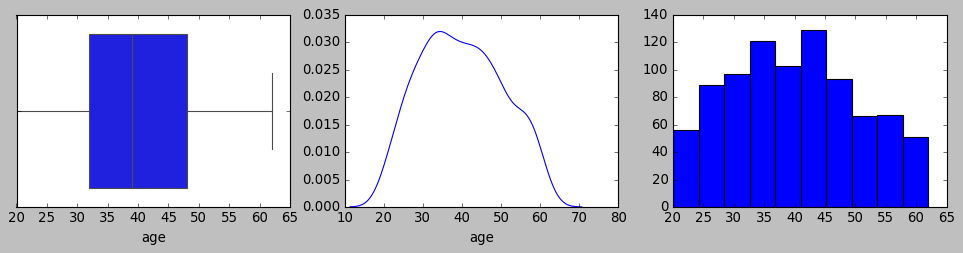


Column Salary is not normally distributed

3. Univariate analysis for age

Mean is 39.955275, Median is 39.000000, Mode(s) are 44.0000

Column age does not have outliers

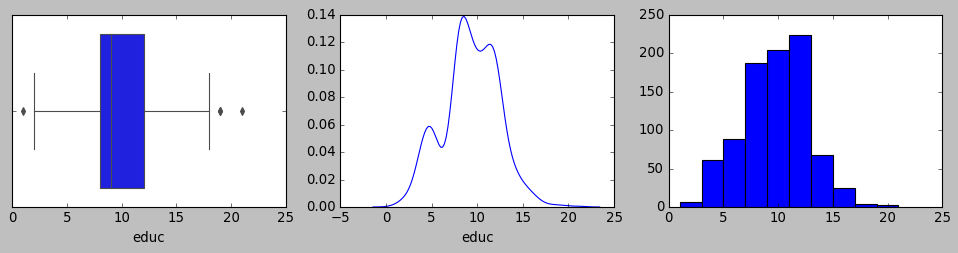


Column age is not normally distributed

4. Univariate analysis for educ

Mean is 9.307339, Median is 9.000000, Mode(s) are 8.0000

Column educ has outliers

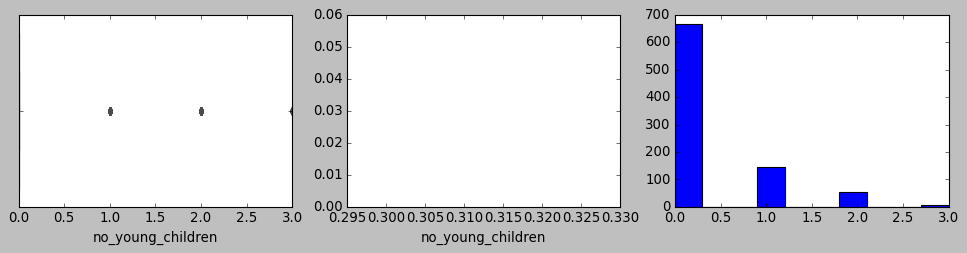


Column educ is not normally distributed

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

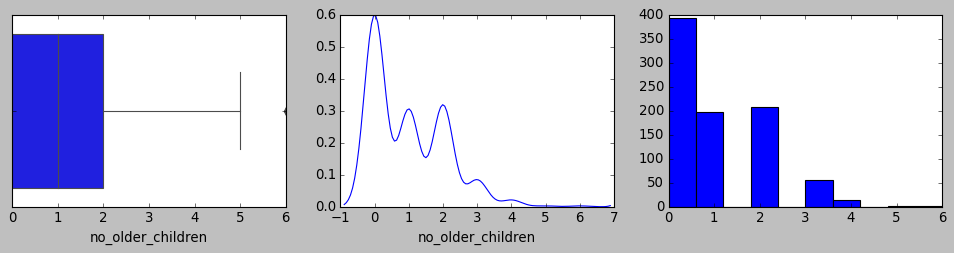


Column no\_young\_children is not normally distributed

6. Univariate analysis for no\_older\_children

Mean is 0.982798, Median is 1.000000, Mode(s) are 0.0000

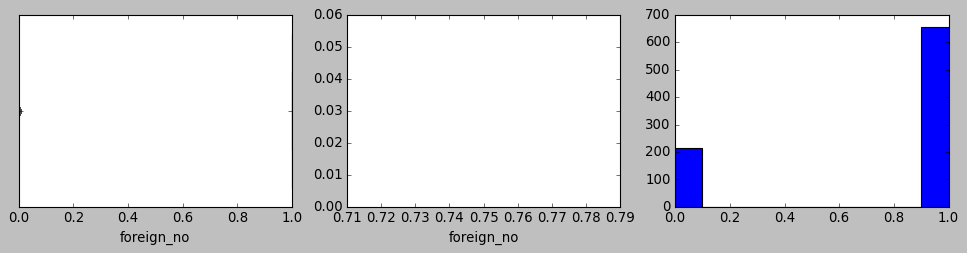
Column no\_older\_children has outliers



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

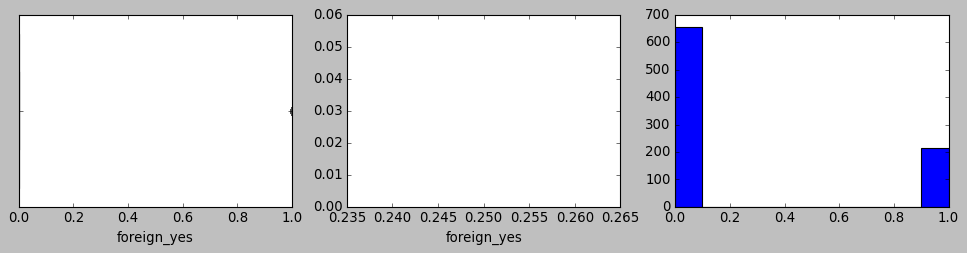


Column foreign\_no is not normally distributed

8. Univariate analysis for foreign\_yes

Mean is 0.247706, Median is 0.000000, Mode(s) are 0.0000

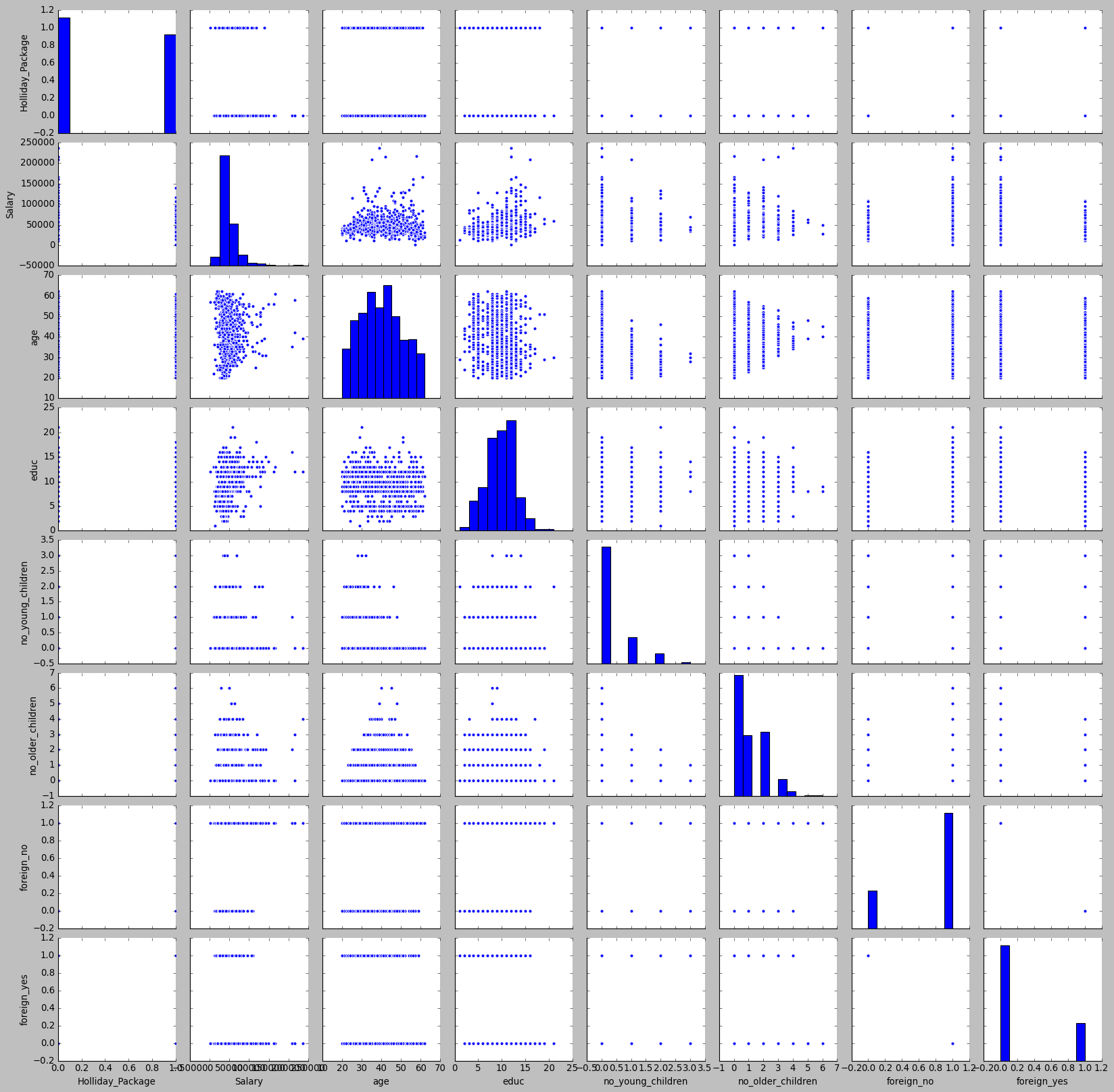
Column foreign\_yes has outliers



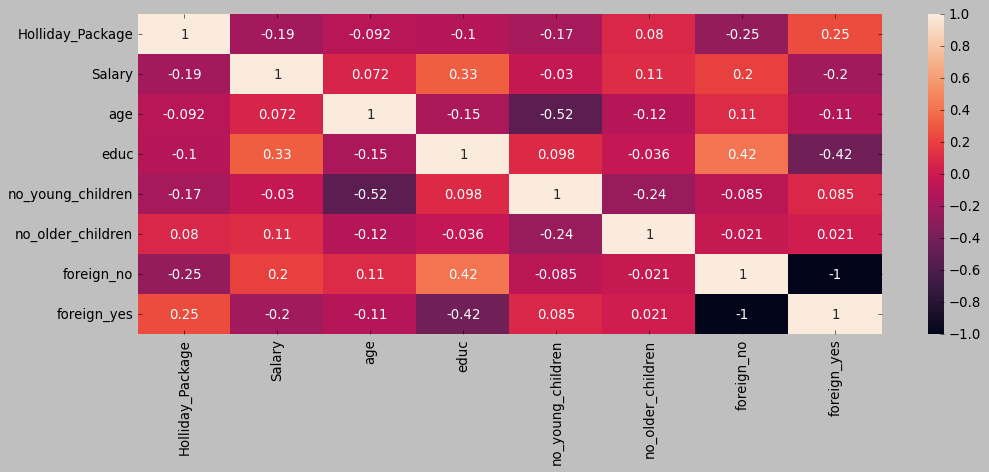
Column foreign\_yes is not normally distributed

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 co relation observed between foreign(yes and no) and educ and some negative corelation between age and no\_young\_children.

Also the target variable (Holliday\_Package) does not show 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).**

Please note that since 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.

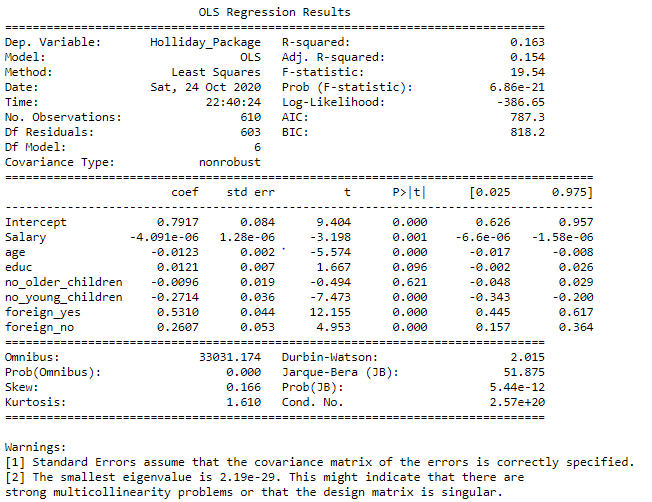
Please find below the 5 number summary for the training data (70% of the sample) after the categorical object types has been converted to numeric as mentioned before.



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

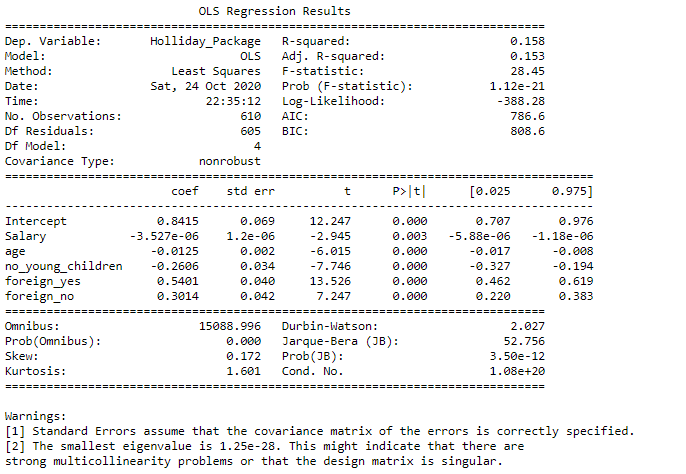


Based on the regression model that has been built based on the above training and testing data set below are the statistical summary using the statsmodel library.



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.



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

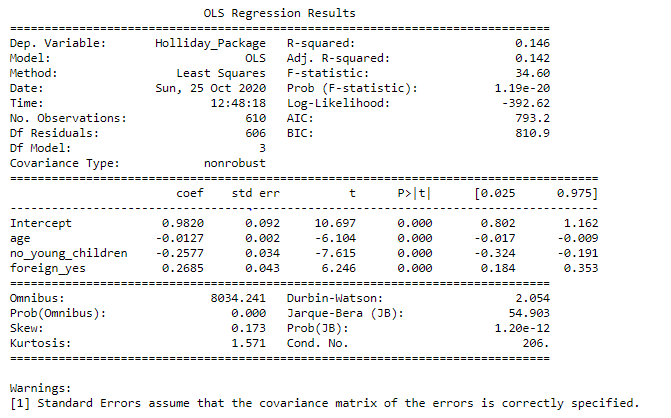
Based on further analysis from the above VIF we could realize that foreign\_no and Salary had to be removed in that sequence to bring the VIF within acceptable range as below:

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:



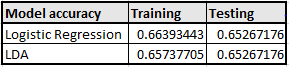
Based on the above summary it can be observed the coefficient of determinant (R squared) is slightly lesser than the original summary built before removing the multi collinearity. Higher the R squared value better the model is.

However, we shall proceed with running logistical 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.

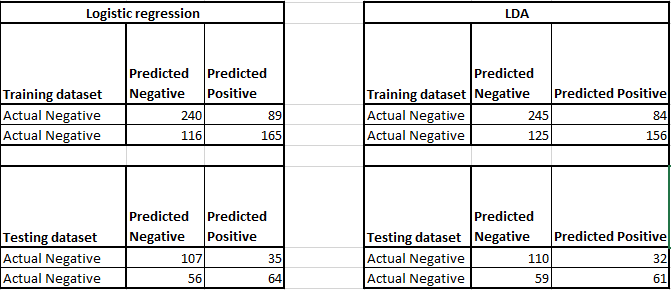
**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

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

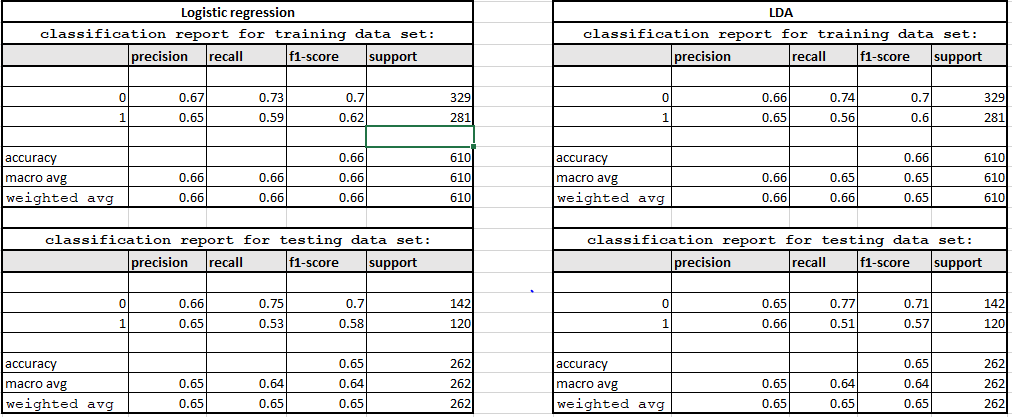
**Model accuracy:**



**Confusion matrix**

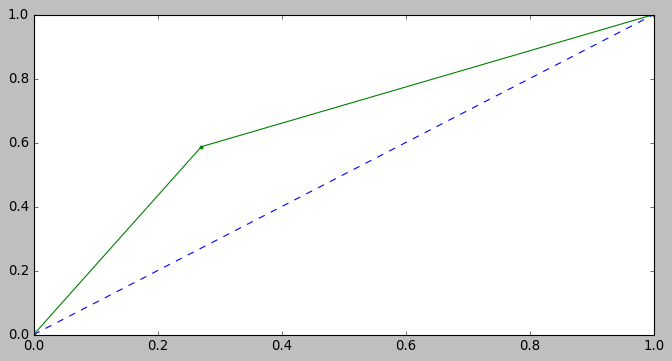


**Classification report:**



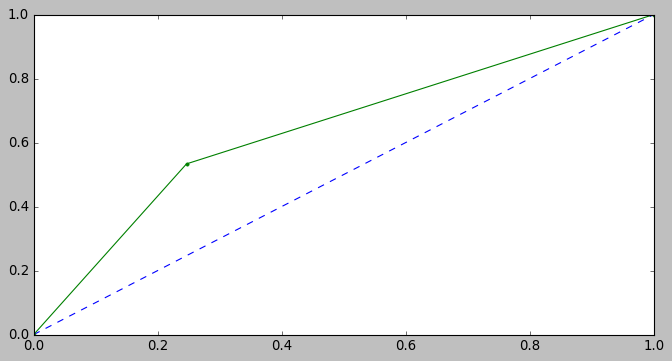
ROC/AUC from Logistical regression model:

ROC curve for training data



AUC for the Training Data: 0.658

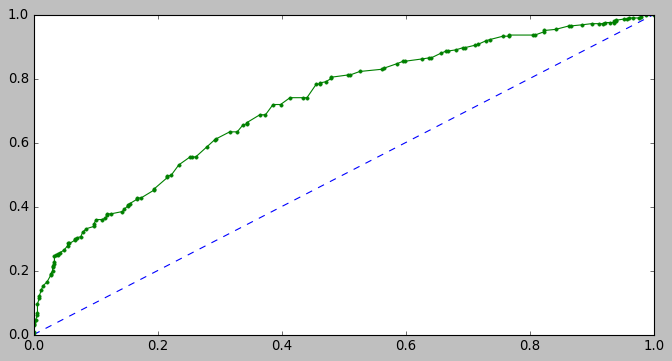
ROC curve for testing data



AUC for the Test Data: 0.643

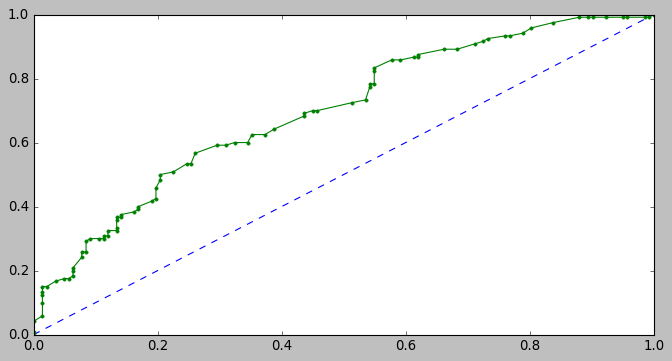
ROC/AUC from LDA:

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 logistical 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 revealed that Holliday\_Package does not have a noticeable co relationship with any of the predictor variable. However, the distribution of Holliday\_Package between customer who chose the vacation versus not is almost equal and hence model has enough data to classify them well.

From an interpretation and insight perspective, there were few multi collinearity 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 resort located near children attractions with reliable child care services can generate interest level for that customer segment and hence provide increased market share opportunities.