

PROJECT ON PREDICTIVE MODELLING

**BUSINESS REPORT**

Sankesh Nagrare

Sankeshnagrare92@gmail.com

PGP – Data Science & Business Analytics

# Case Study 1 – Linear Regression

## Overview:

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.

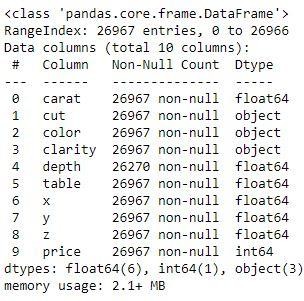
## Summary:

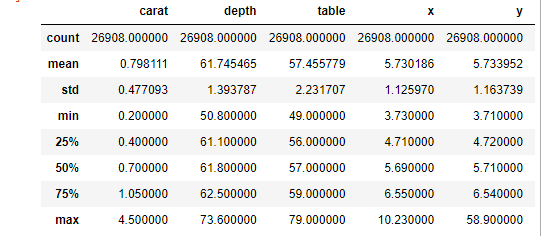
This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provide some key insights/recommendations to the business.

## Q1.1) Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).







**

### Check for Duplicates

Below is the output of python code -



Below is the number of rows and columns before and after removal of duplicate rows

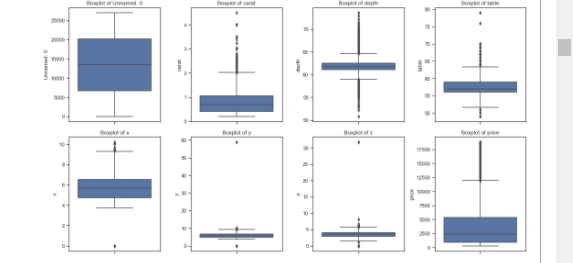


### Observations -

* The above dataset contains 26967 rows & 11 columns.
* The variables 'cut', 'color' and 'clarity' are of object datatype, whereas the rest of the variables are of numerical (integer & float) datatype.
* There are 697 null values in the dataset. Hence, we need to investigate further to treat them.
* The dataset does have a few duplicate values as well. Since, they are few in numbers compared to the entire dataset, we will remove them.
* The summary table shows mean, standard deviation, minimum & maximum values, etc. for all the variables. Variables 'x', 'y' & 'z' (namely length, width & height) have minimum values 0, which is impossible. Hence, we need to investigate it further.

### Univariate Analysis

1. Outlier Identification

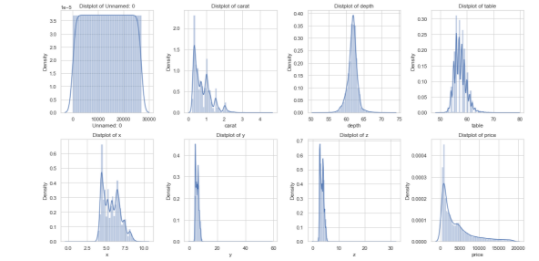


*Boxplot for Outlier Identification*

From the figure above, we can say that,

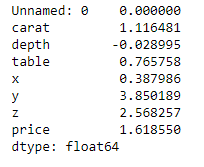
* All the variables contain outliers and since we don't know the reason behind those, we can't treat or drop them.
* Treating outliers sometimes results in the models having better performance but the models lose out on generalization. Hence, we won't be treating them in order to not lose out on generalization.

1. Distribution check



*: Distplot for distribution check for all variables*

Skew values for all the variables



From the figure and skew values above, we can say that,

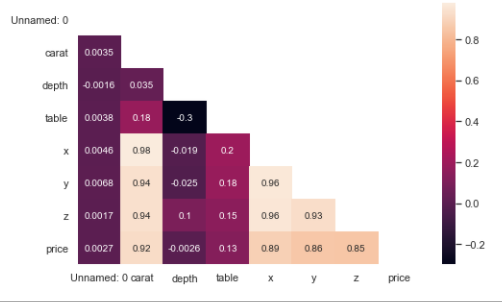
* Since, the skewness value of variables 'depth' and 'x' is between −0.5 and +0.5, they show

approximately symmetric distribution.

* Since, the skewness value of variable 'table' is between +0.5 and +1, it shows moderately skewed distribution.
* Since, the skewness value of variables 'carat', 'y', 'z' and 'price' is greater than +1, it shows highly skewed distribution.

### Multivariate Analysis

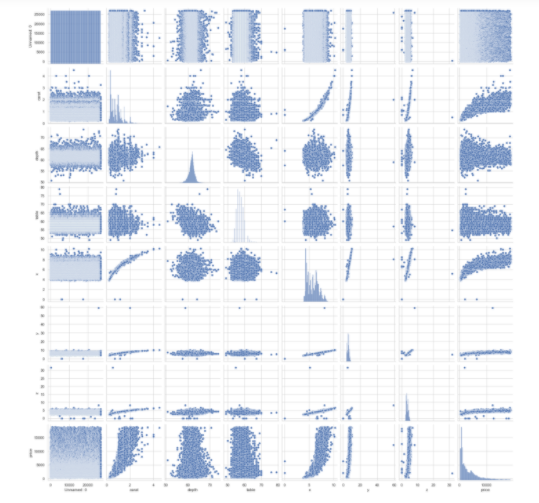
1) Heatmap to study correlation between variables



*Correlation Matrix*

From the figure above, we can say that,

* There is a strong correlation between the variables’ 'carat', 'x', 'y' 'z' & 'price' (i.e. >0.90).



*Pair plot for all variable combinations*

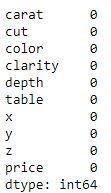
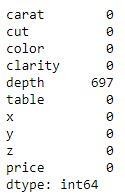
From the figure above, we can say that,

* All the correlations can be viewed as in the pairplot as the data points are closely packed to each other in the above combinations of variables.
* Some variables do have correlation, but, is very weak in strength which is evident in the graph

## 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? Check for the possibility of combining the sub levels of all the ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

### Imputing null values

Since, mean takes all the values of the dataset into consideration, we will impute the null values with the respective mean value of the variable.

Null values before imputing Null values after imputing

### Combining the sub levels of ordinal variables

* Sub-levels of 'cut' ordinal variable cannot be combined as it is well defined.
* Sub-levels of 'colour' ordinal variable cannot be combined as we don't have much information by which we can do combining.
* Sub-levels of 'clarity' ordinal variable can be combined as we can combine
  + VVS1 & VVS2 as VVS.
  + VS1 & VS2 as VS.
  + SI1 & SI2 as SI.
* This leads to ease of understanding and makes further analysis a bit easier.

Before combining After combining

CUT : 5

Fair 779

Good 2434

Very Good 6027

Premium 6880

Ideal 10805

Name: cut, dtype: int64

COLOR : 7

0 1440

1 2765

6 3341

2 4091

4 4722

5 4916

3 5650

Name: color, dtype: int64

CLARITY : 8

I1 362

IF 891

VVS1 1839

VVS2 2530

VS1 4086

SI2 4561

VS2 6092

SI1 6564

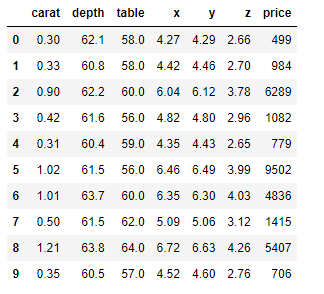
Name: clarity, dtype: int64

### Observations -

* We imputed the null values of the dataset with the mean values.
* We combined sub-levels of ordinal variable 'clarity' for ease of analysis.

## Q1.3) Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

### Converting categorical to dummy variables

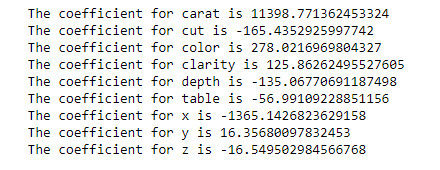
****

We can perform Linear Regression using following methods –

### Linear regression using sklearn

We know, the variable ‘price’ is the target variable and we split the data into 70:30 (Train:Test)

Once, we have built the model, we find the coefficients for all the variables & intercept for the linear equation.





R2 for training data R2 for testing data RMSE of Training data

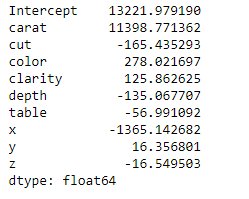
**Observations -**

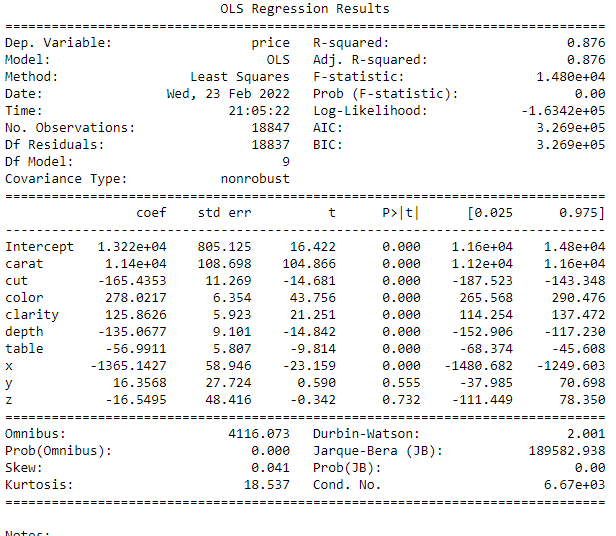
* We see that the variable 'carat' is the most important or governing parameter in order to decide the price of the cubic zirconia
* The model scores for training and testing data is almost same, hence, the model is a good fitted model.
* The accuracy of the model is approximately 67%.

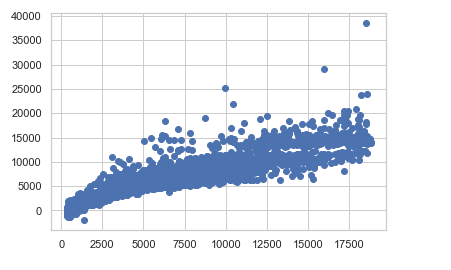
### Linear regression using stats models

First, we concatenate X and y into a single dataframe

Once, we have built the model, we find the coefficients for all the variables & intercept for the linear equation.



*OLS Test results* 

**

*Plot of predicted y vs Actual y*

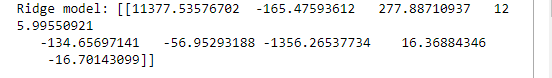
From the figure above, we can see a linear trend in the dataset.

### Observations -

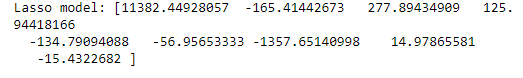
* We see that the variable 'carat' is the most important or governing parameter in order to decide the price of the cubic zirconia
* The model scores for training and testing data is almost same, hence, the model is a good fitted model.
* The accuracy of the model is 67%.

### Creating a regularized RIDGE & LASSO model and finding the coefficients for all the variables

We scale the dataset before building the model Coefficients from RIDGE model



Coefficients from LASSO model



Many of the coefficients have become 0 indicating drop of those dimensions from the model

### Model Score of Ridge for training & testing data respectively

0.9148202804295886 & 0.906787441146191

### Model Score of Lasso for training & testing data respectively



### Observations -

* From comparing the scores above for both ridge and lasso models, we can say that, ridge performs better.

### Conclusions -

From all the models we have seen above, we can say that,

* Linear Regression models using sklearn, statsmodel and ridge models behave in a similar way as they have same scores.
* Hence, we select the model of linear regression using statsmodel because it provides additional statistical information of the model and the dataset which helps us to better undestand the data and make better decisions overall.

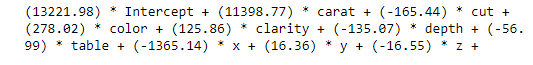
### Multi-collinearity check using Variance Inflation Factor (VIF)

### 

Since, most of the variables have VIF values greater than 1, hence, multi-collinearity exists in the dataset.

### Conclusions -

**Final Linear Regression equation for the target variable price is**



Here,

* The positive coefficients mean when the corresponding variable increases, price will also increase.
* The negative coefficients mean when the corresponding variable increases, price will decrease.

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

Now, since we have established a price predicting equation, we can find the best attributes to segregate higher and lower profitable stones

* + The most important attribute in governing the price of the cubic zirconia is 'carat' as it has highest positive coefficient. If we increase the carat, the price can go up and we can make more profit.
  + We should maintain the value of 'depth' and 'table' to optimal minimum so as to not compromise in quality and yet make more profits as they have negative coefficients.
  + The length of the cubic zirconia should be least and width & depth should be optimal minimum as all 3 have negative coefficients.
  + When the cut quality is of
    - 'Ideal' category we can make the most profitable stones of all the possible cut categories as it has the highest coefficient which leads to maximum price.
    - 'Fair' category we can make the least profitable stones of all the possible cut categories as it has the lowest coefficient which leads to minimum price.
    - The remaining 3 categories we can make the moderate profitable stones of all the possible cut categories as they have coefficient values ranging between the 'Ideal' and 'Fair' category.
  + When the colour quality is of
    - 'D' category we can make the most profitable stones of all the possible colour categories as it has highest coefficient which leads to maximum price.
    - 'J' category we can make the least profitable stones of all the possible colour categories as it has lowest coefficient which leads to minimum price.
    - The remaining 5 categories we can make the moderate profitable stones of all the possible colour categories as they have coefficient values ranging between the 'D' and 'J' category.
  + When the clarity quality is of
    - 'IF' category we can make the most profitable stones of all the possible clarity categories as it has highest coefficient which leads to maximum price.
    - 'I1' category we can make the least profitable stones of all the possible clarity categories as it has lowest coefficient which leads to minimum price.
    - The remaining 3 categories we can make the moderate profitable stones of all the possible clarity categories as they have coefficient values ranging between the 'IF' and 'I1' category.
  + We should make cubic zirconia's of all combinations of coefficients (i.e. positive & negative). By doing this, we can target all kinds of purchasing population by making gems of all possible price ranges, thereby make profits for all kinds of gems.

# Case Study 2- Logistic Regression and LDA

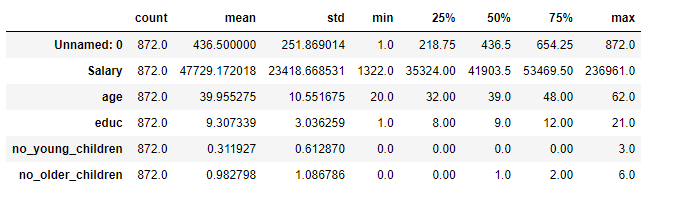
## Overview:

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.

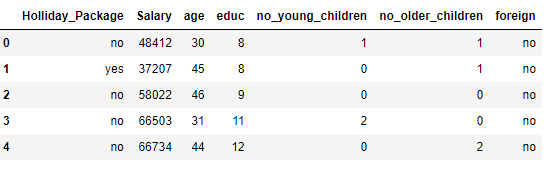
## Summary:

This business report provides detailed explanation on the approach to each problem definition, solution to those the problems provide some key insights/recommendations to the business.

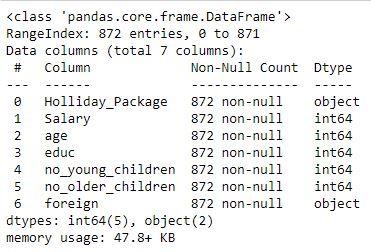
## Q2.1) Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis

****

*Original sample of the dataset*

**

*Sample of the dataset after removal of irrelevant columns*



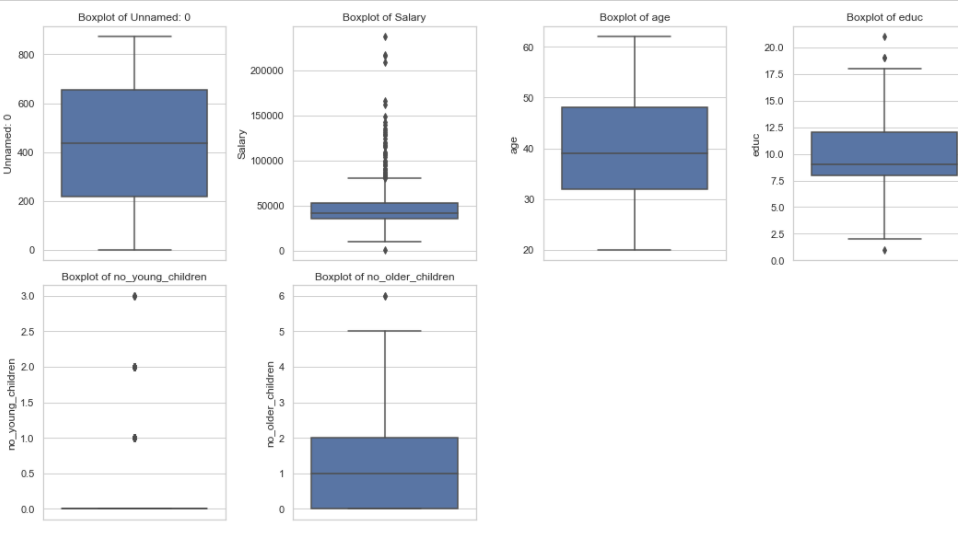
*Basic information of the dataset*

### Observations -

* The above dataset contains 872 rows & 7 columns.
* The variables 'Holiday\_Package' & 'foreign' are of object datatype, whereas the rest of the variables are of integer datatype.
* There are no null & duplicate values in the dataset.
* The summary table shows mean, standard deviation, minimum & maximum values, etc. for all the variables.
* Variables 'Holiday\_Package' & 'foreign' both have only 2 sub-levels within them.

### Univariate Analysis

1. Outlier Identification



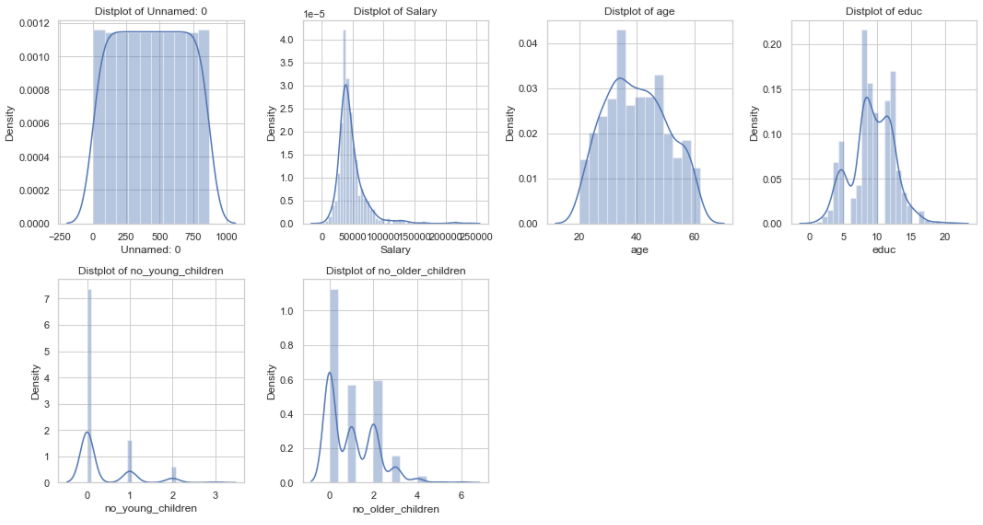
*Figure Boxplot for outlier identification*

### Observations -

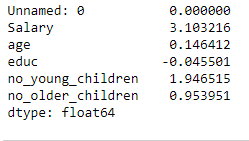
From the plot above, we can say that,

* The variable 'Salary' contains a good number of outliers.
* The variables 'education', 'no\_young\_children', 'no\_older\_children' have a very few outliers.
* The variable 'age' contains no outliers.
* Treating outliers sometimes results in the models having better performance but the models lose out on generalization. Hence, we won't be treating them in order to not lose out on generalization.

1. Distribution check



*Distplot for distribution check for all variables*

**

### Observations –

From the graph & skew values above, we can say that,

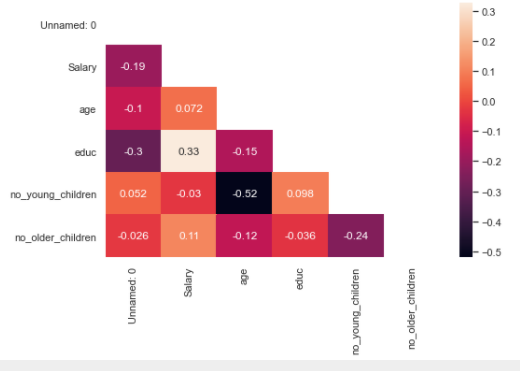
* Since, the skewness value of variables 'education' & 'age' is between −0.5 and +0.5, they show

approximately symmetric distribution.

* Since, the skewness value of variable 'no\_older\_children' is between +0.5 and +1, it shows moderately skewed distribution.
* Since, the skewness value of variables 'salary' & 'no\_young\_children' is greater than +1, it shows highly skewed distribution.

### Multivariate Analysis

1) Correlation check

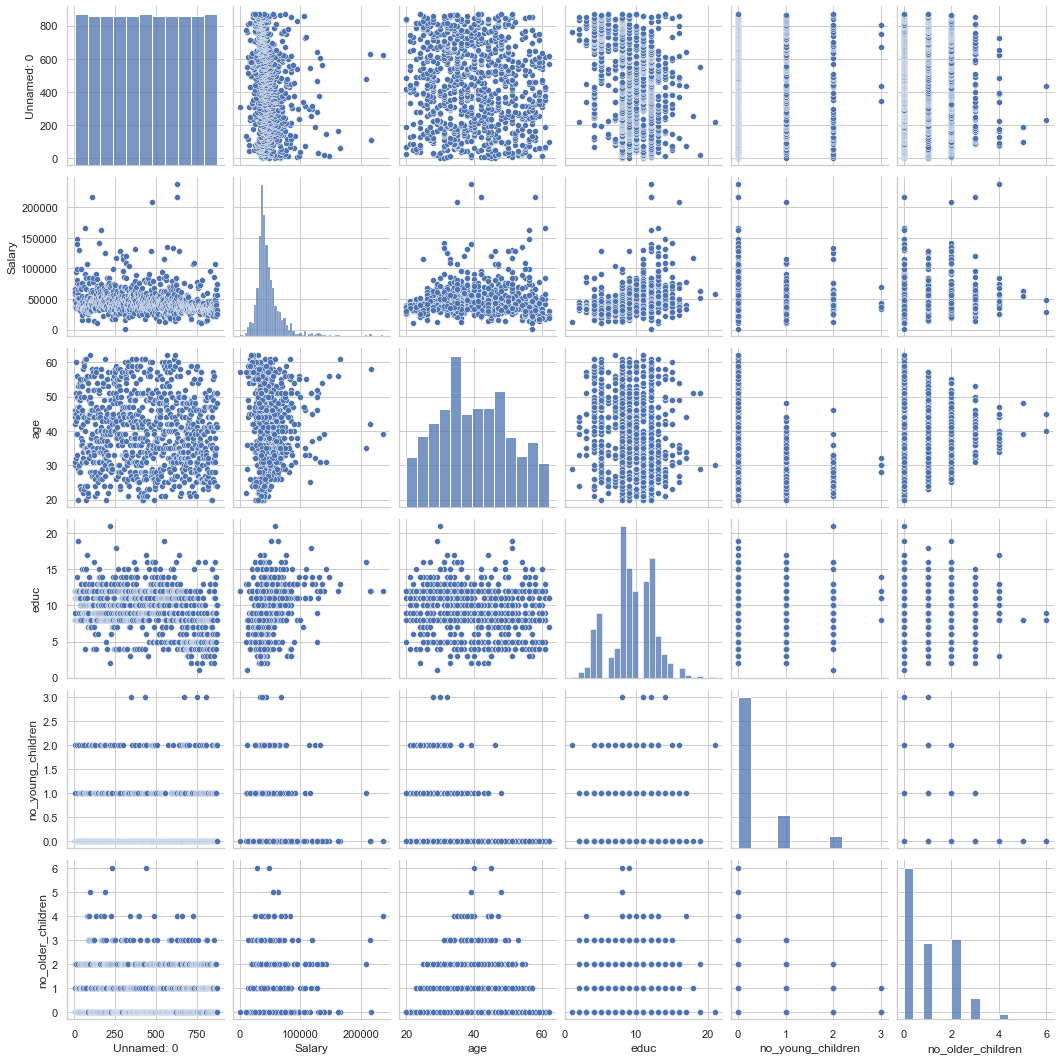


*Correlation matrix*

### Observations –

From the graphs above, we can say that,

* There is a weak correlation between the variables 'salary' & 'education' (0.33).



*Pairplot for all variable combinations*

### Observations -

* The remaining variable combinations have negligible or no correlation as seen in heatmap, which is evident in the pairplot graph of all variable combinations.

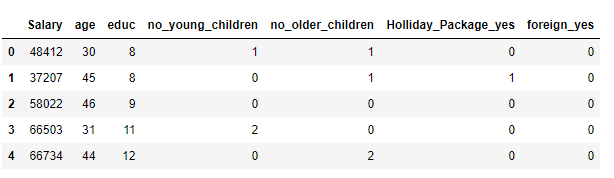
### Observations -

From the graph above, we can say that,

* + The employees who are
    - Non-foreigners, who opt out for a holiday package have higher average salary compared to those who opt in.
    - Foreigners, who opt out for a holiday package have higher average salary compared to those who opt in.
  + The employees who are
    - Non-foreigners, who opt out for a holiday package have higher average age compared to those who opt in.
    - Foreigners, who opt out for a holiday package have higher average age compared to those who opt in.
  + The employees who are
    - Non-foreigners, who opt out for a holiday package have lesser average education duration compared to those who opt in.
    - Foreigners, who opt out for a holiday package have higher average education duration compared to those who opt in.
  + The employees who are
    - Non-foreigners, who opt out for a holiday package have less number of older children on an average compared to those who opt in.
    - Foreigners, who opt out for a holiday package have less number of older children on an average compared to those who opt in.
  + The employees who are
    - Non-foreigners, who opt out for a holiday package have more number of young children on an average compared to those who opt in.
    - Foreigners, who opt out for a holiday package have more number of young children on an average compared to those who opt in.
  + The employees who are
    - Non-foreigners, are more likely to opt out for a holiday package compared to those who opt in.
    - Foreigners, are more likely to opt in for a holiday package compared to those who opt out.

## 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).

### Encoding Object datatypes -

****

*Dataset encoding*

Once, encoding is done we further

1. Copy all the predictor variables into a dataframe and then copy target into another dataframe.
2. We split the entire dataset into training and test dataset (70:30 ratio respectively). We do this to make sure model learns considering good amount of data.
3. Then, we build and fit the data into the models namely
   1. Logistic Regression
   2. Linear Discriminant Analysis
4. Once the model is built, we then predict the same for both training and test data. We do this to understand how the model behaves with new data.
5. Then, we predict the classes and probabilities which help us later to evaluate the model performance.

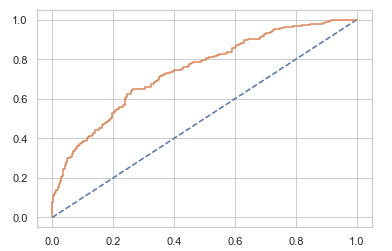
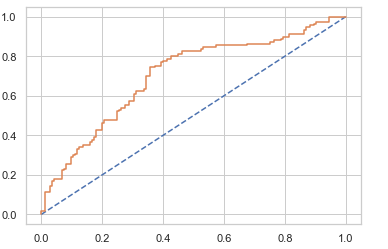
## Q2.3) Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Once, we have built the model and predicted the values, we check its performance metrics.

### Model Evaluation of Logistic Regression model

1. Model accuracy for training and test data –



1. AUC & ROC for training and test data –
2.  

: AUC & ROC curve for training data : AUC & ROC curve for test data

1. Confusion matrix for both training and test data

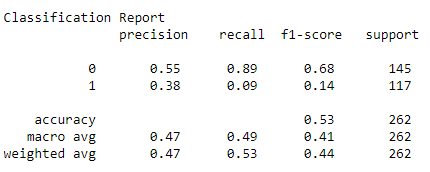
Confusion matrix train data set

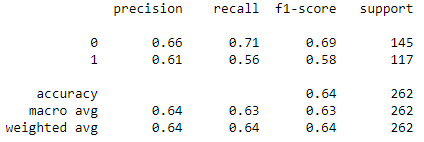


Confusion matrix test data set



1. Classification report for training and test data





### Logistic Regression Conclusion –

Train Data:

AUC: 60%

Accuracy: 53%

Precision: 53%

f1-Score: 67% Test Data:

AUC: 60%

Accuracy: 53%

Precision: 53%

f1-Score: 68%

### Model evaluation of Linear Discriminant Analysis model

1. Model accuracy for training and test data –



1. AUC & ROC for training and test data –



AUC & ROC curve for training and test data

1. Confusion matrix for both training and test data





From above, we can say that, variables number of younger children and being a foreigner or not are critical in predicting whether the employee will opt for holiday package or not.

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

From the analysis above, we came to know that the Linear Discriminant Analysis model performs much better. Also, we found that variables number of younger children & being foreigner or not plays a crucial role in determining whether the employee will opt in or out for a holiday package.

* If the employee has more number of children who are less than 7 years old, he/she is more likely to opt out for the holiday package compared to the ones who have less number of children who are less than 7 years old.
* If the employee is not a foreigner, he/she is more likely to opt out of holiday package.
* If the employee is a foreigner, he/she is more likely to opt in of holiday package.

The model can perform well if we get more number of data points, because, for now the data on which the model is built has same number of yes & no values for employees who opt in for holiday packages. However, it still performs averagely with 66% accuracy.