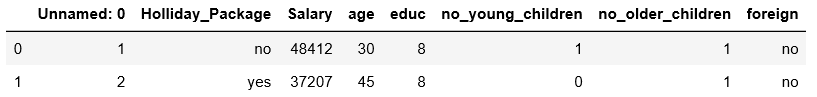
# **Problem 2: Logistic Regression and LDA**

## You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not on the basis of the information given in the data set. Also, find out the important factors on the basis of which the company will focus on particular employees to sell their packages.

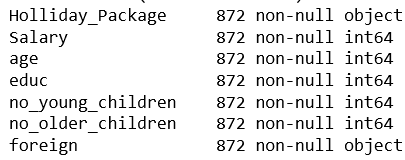
## 2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

The data is read and is glanced through to get a quick idea of what it looks like.

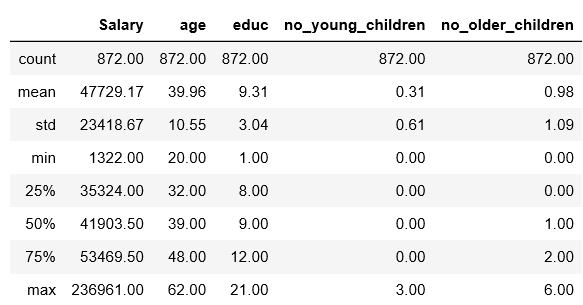


The column named “Unnamed: 0” is not useful and it is dropped.

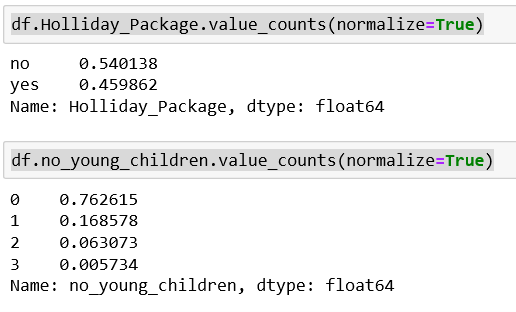
The data types of the columns are then checked for :



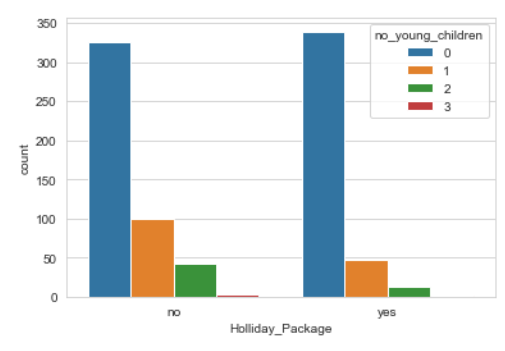
Using the Pandas describe() function, the data is viewed from statistical viewpoint.



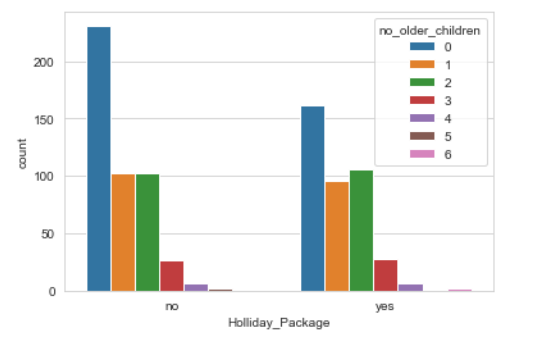
Further, checks are made on how many people amongst all in the data have taken holiday packages. It was found that 46% of the total have taken up holiday packages.



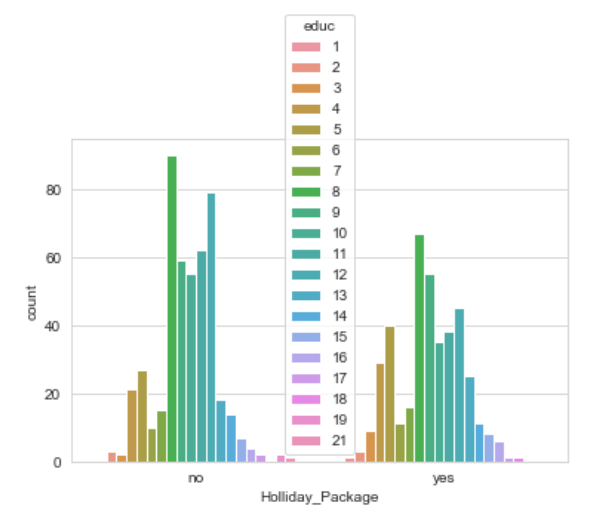
We then try to check if the number of young children a person has an effect on their decision to take a holiday by plotting a countplot such as the one shown below. It is observed that people with few or no children have a higher probability of taking up a holiday package.



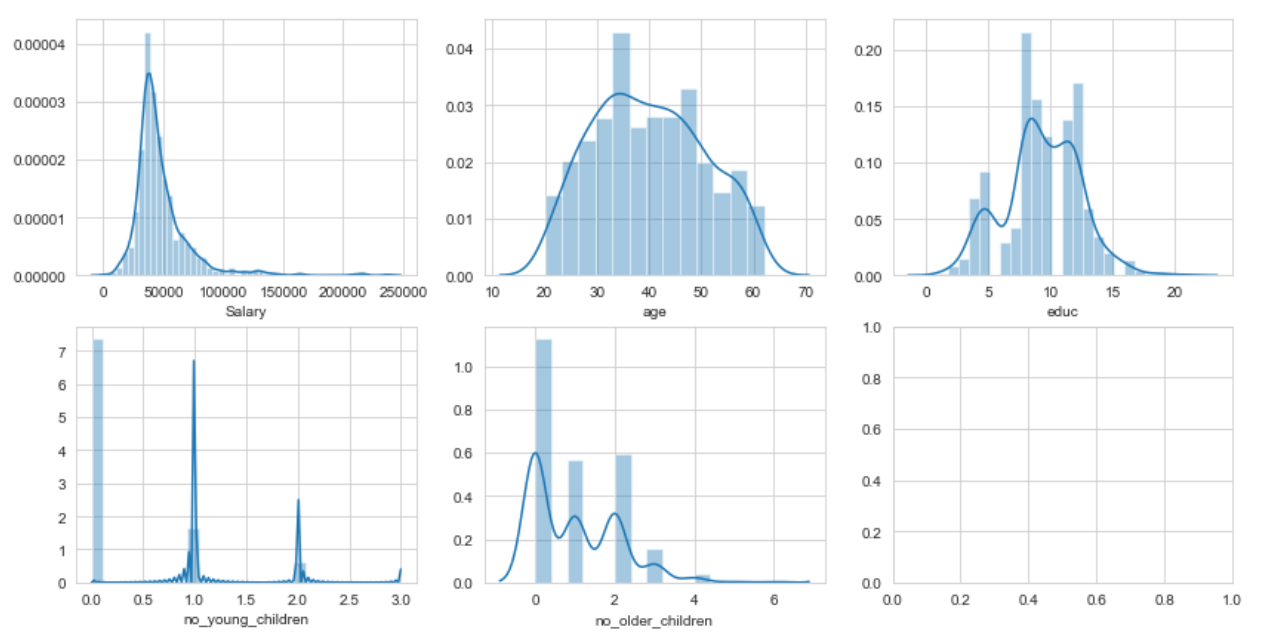
Similarly, we can check if the number of older children a person has an effect on their decision to take a holiday. Here, we observe that people with no older children tend to be the group that takes the most number of holidays.



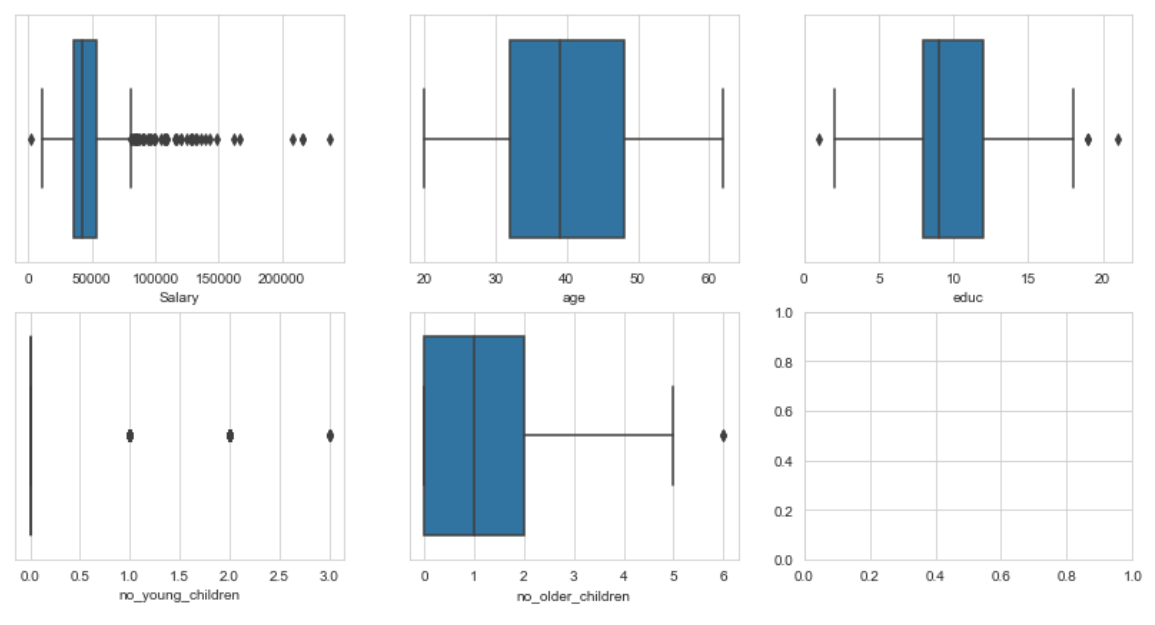
It is also observed that people within 8-12 years of education take the most number of holidays compared to the rest.



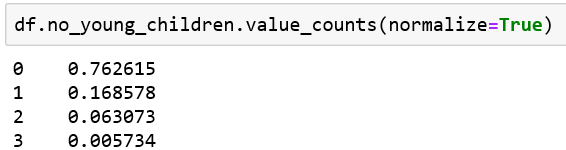
We then create distribution plots to view the distribution of data for each variable and how normal they are. Some skewness is observed in data related to the number of children and this is acceptable. It will be checked later whether grouping can be done for data with higher number of children.



**Outlier** checking is done and the continuous variables (salary, age, edu) are treated (imputed) by moving the outliers to the lower and upper limits (Q1-1.5 IQR, Q3 + 1.5 IQR) of the box plot shown. This kind of treatment is done because of lack of points with values in the outlier region. More data in this region would make the data more Gaussian/Normal.



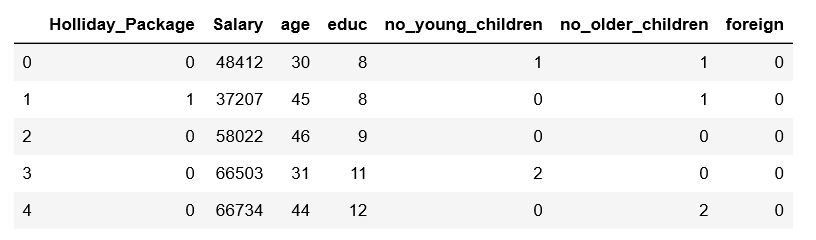
It is observed that people with younger children get flagged as outliers. This is because the number of people with zero younger children as a proportion of the set studied is large (76%).



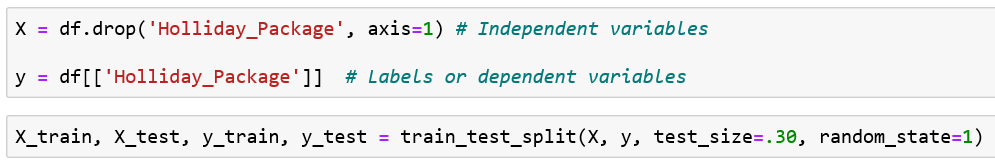
## 2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

Data encoding is done to the columns “Holliday\_Package” and “Foreign”.

Encoding is done with 0 for “no” and 1 for “yes”.

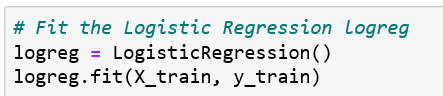


Data is then split into labels(y) and independent variables(X). The data for independent variables is further split into train and test data using a split of 70:30.



LDA and Logistic regression is performed on the data which has been split to train and test the model. The results are reviewed further.



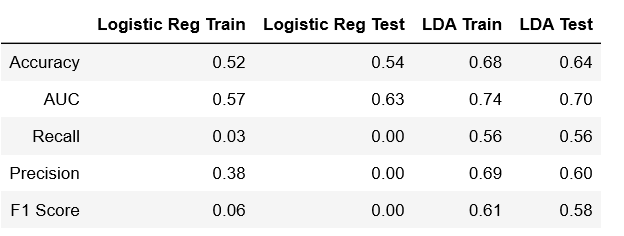


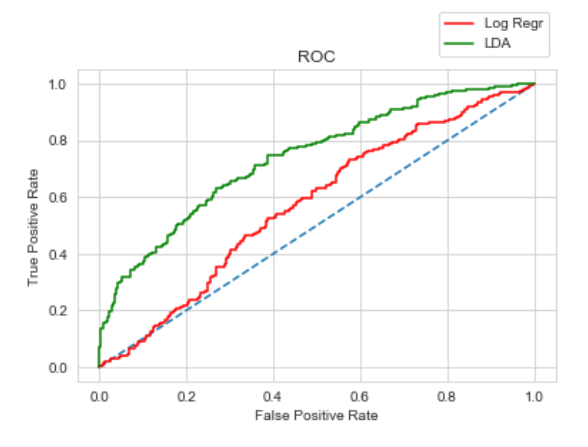
## 2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

INITIAL MODEL :

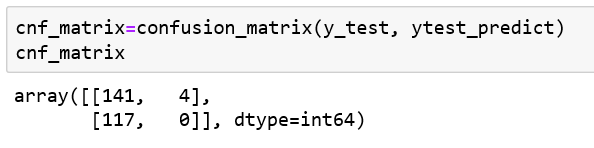
We see outliers in “no\_of\_young\_children” and “no\_of\_older\_children”. Here the number of people with zero children are higher and hence this set skews the data. We choose not to treat it as this represents a real scenario. More data that would help normalize the data could make the predictions better.

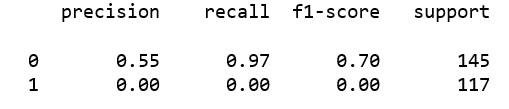
We get the following results :





When comparing these two models, we can conclude based on Accuracy, AUC and other parameters that LDA is much better compared to logistic regression in this case.

+

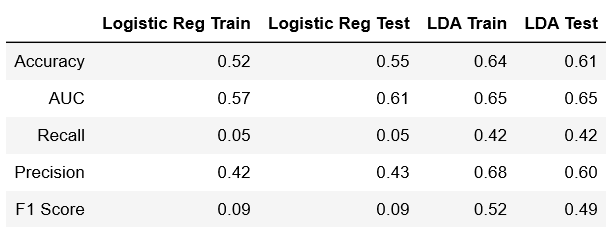


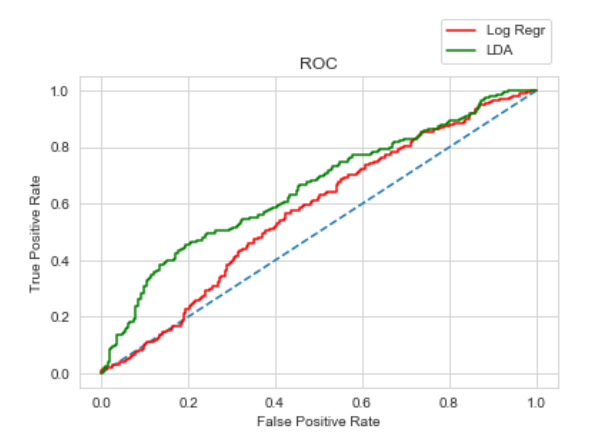
It is seen from the confusion matrix and classification report that the recall, precision and f1 scores are low in logistic regression especially for ‘1’s or positives.

This can be improved by getting more data that correspond to this case, which will help improve the prediction of ‘1’s or positives.

Further, using feature selection techniques, scaling and grouping data to make the results can be further improved.

To test this, we remove the features ‘no\_young\_children’ and ‘no\_older\_children’. We observe that using selected features improves the efficiency of the logistic regression method.





Therefore, we can try backward and forward feature elimination to improve the score. However, this is a lengthy process and beyond the scope of this exercise.

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

On the basis of these predictions, we find that LDA is more suitable for this dataset. Better accuracy can be obtained by feature reduction or having more relevant data which provides more positives (as observed from classification report).