**Problem 2:** Logistic Regression and LDA

### Problem Statement:

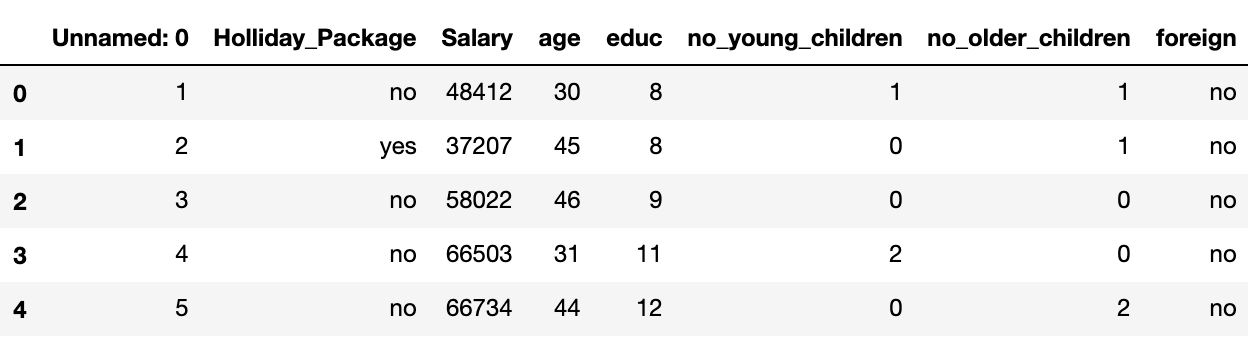
You are hired by a tour and travel agency which deals in selling holiday packages. You are provided details of 872 employees of a company. Among these employees, some opted for the package and some didn't. You have to help the company in predicting whether an employee will opt for the package or not 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.

Dataset has 872 observations and 8 attributes.

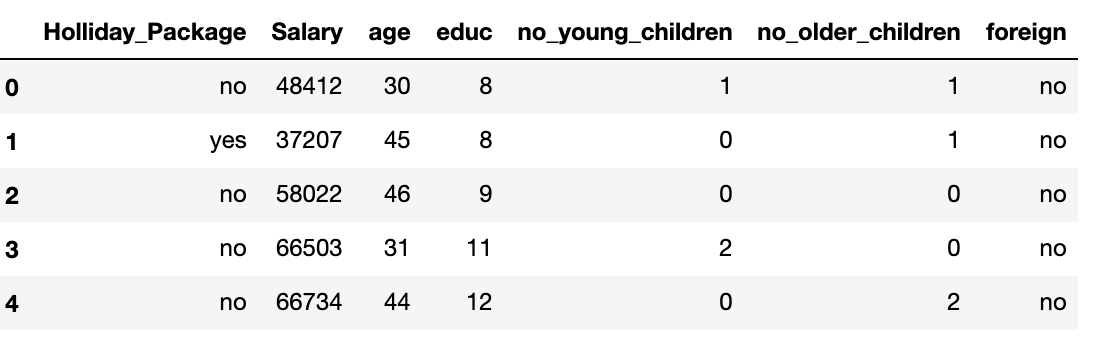
* 6 attributes named unnamed :0, salary, age, educ, no\_young\_children, no\_older\_children are of integer type
* 2 attributes named holliday\_package, foriegn are of object type.

Let’s start the data exploration step with the head function to look at the first 5 rows.



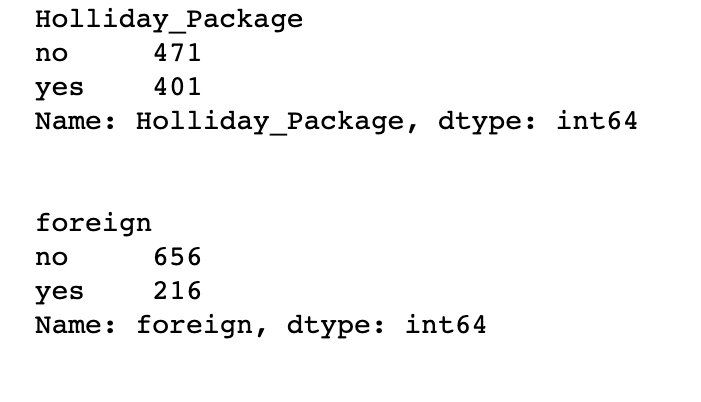
Unnamed: 0 seems to be of no use for our dataset and it’s dropped due to that.

Dataset after dropping the first column

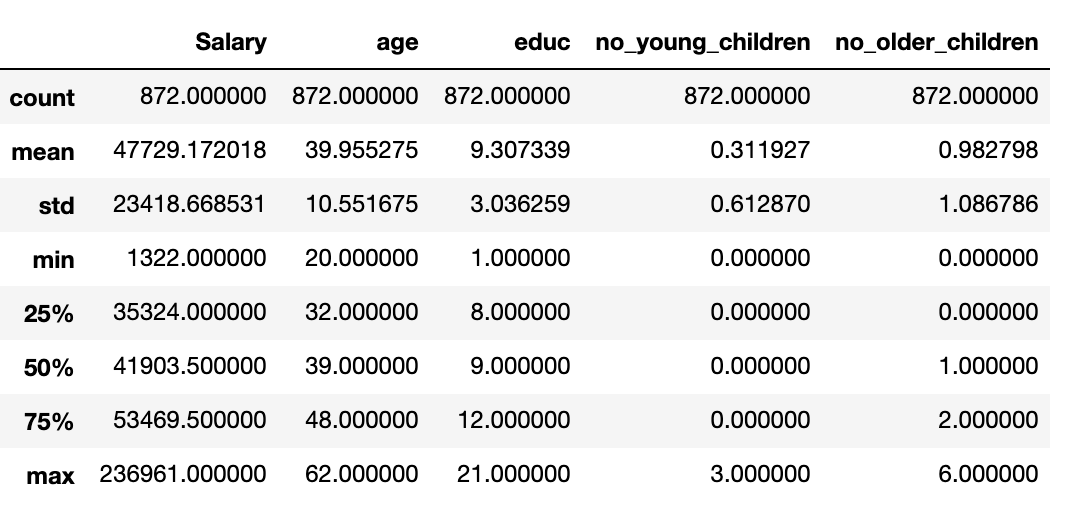


Now the shape of the dataset is (872, 7). There are 401 who opted for a Holliday\_Package and 471 who did not opt for it. The proportion of yes and no are almost the same.

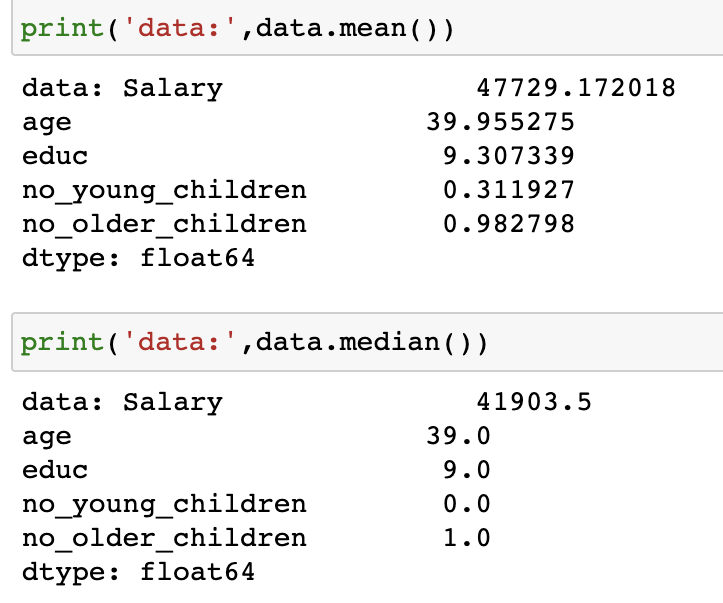
The proportion 1’s and 0’s for object data types are as shown below:



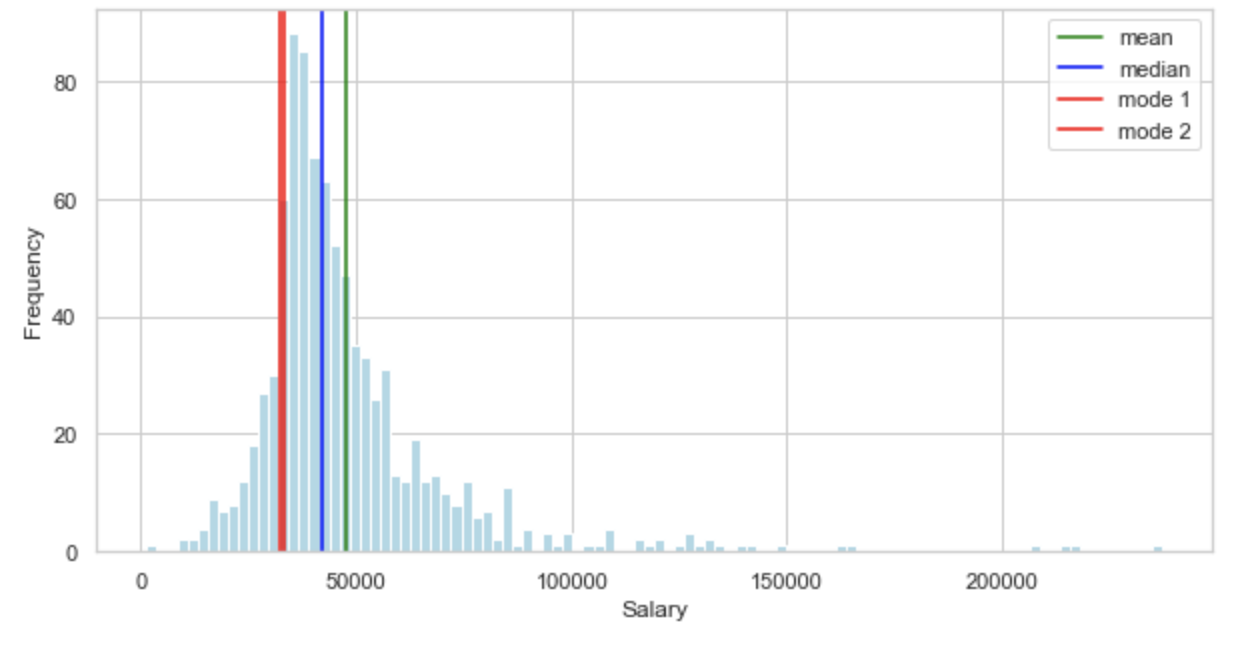
After checking the summary of the data, there seems to be some outliers in the dataset. Let’s check them by doing further exploration.



By doing descriptive statistics, we observed that mean and median are almost the same for all attributes.

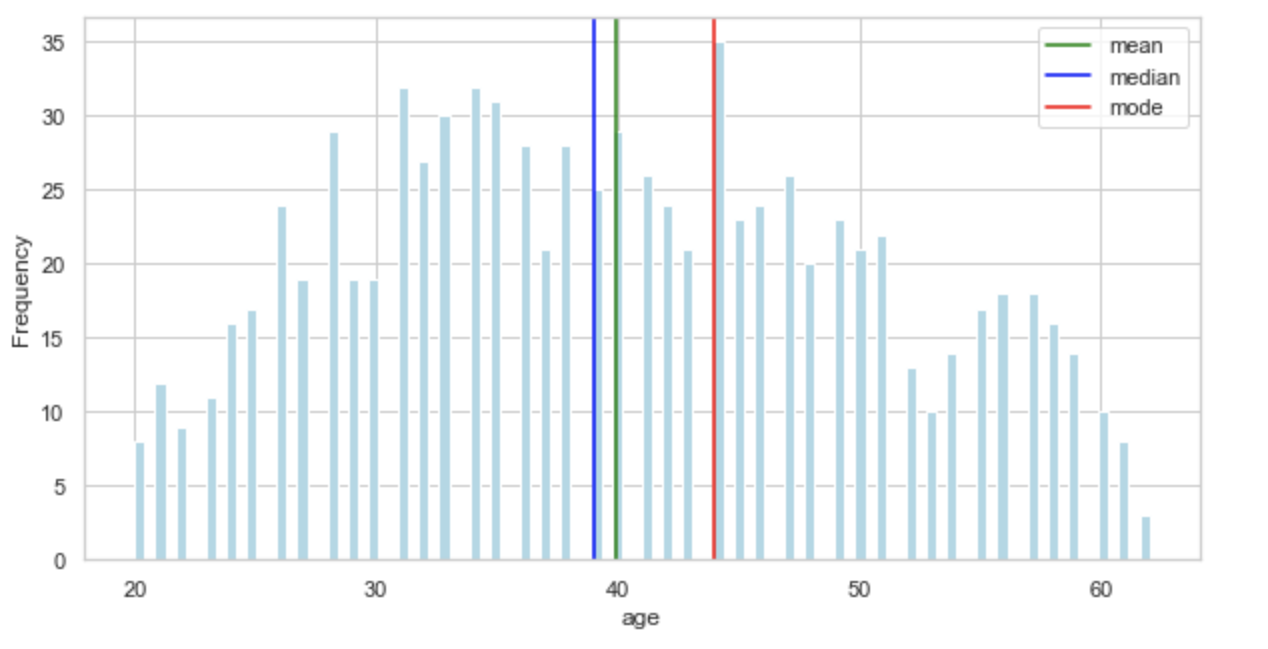


From the distribution of all attributes given in the below picture, we can see that the distribution for salary attribute seems to be positive.

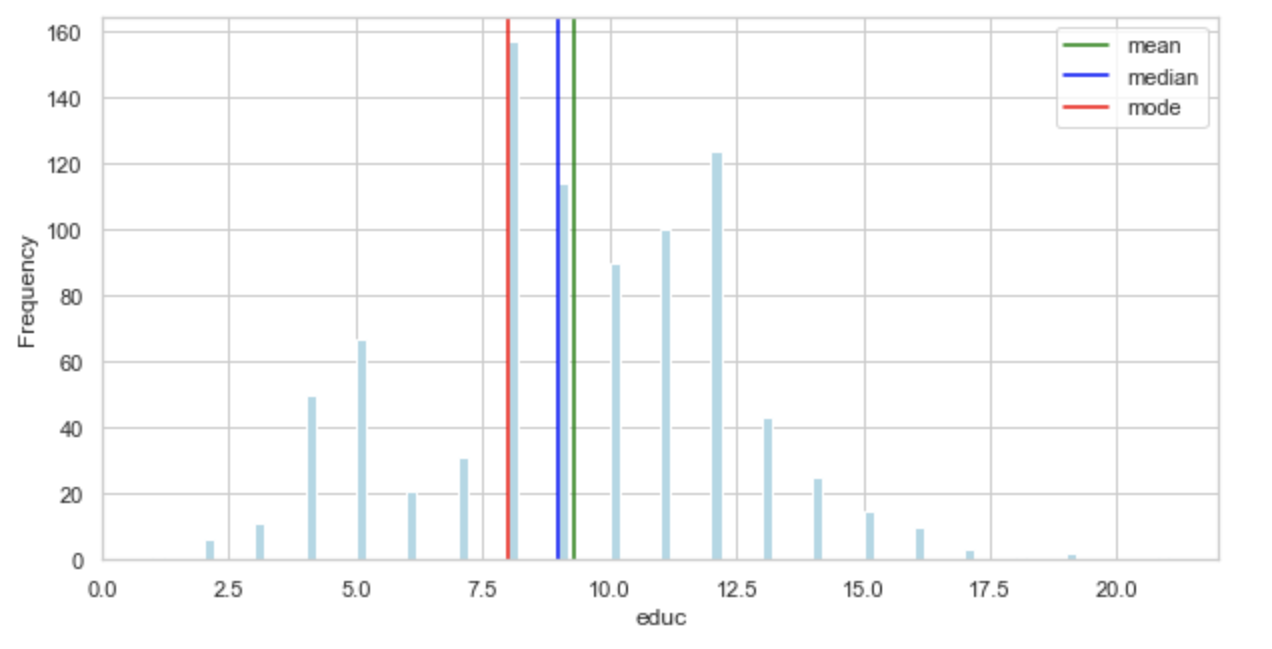


Most frequent values are low and tail is towards high values

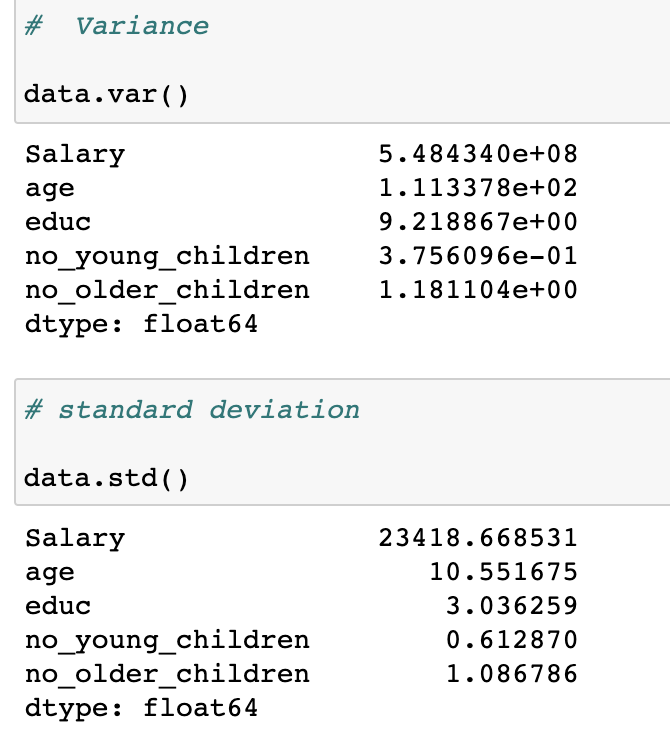
The distribution for age attribute seems to be uneven. Mean and median are almost the same.



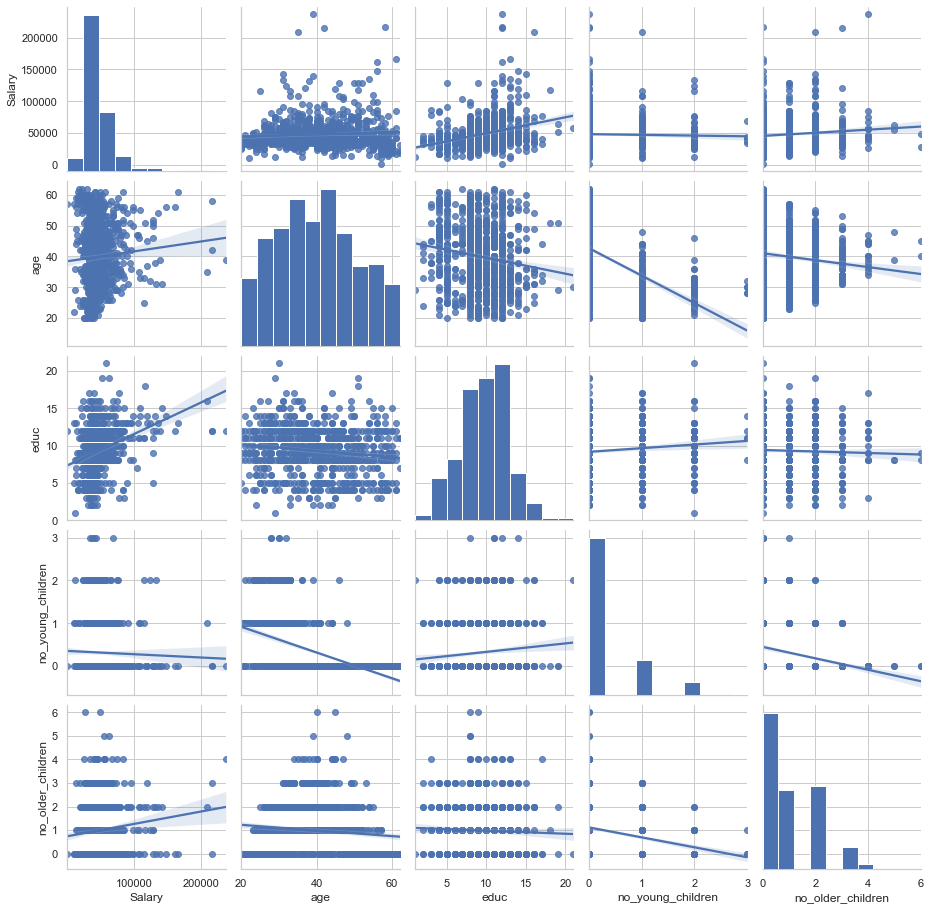
The distribution for educ attribute seems to be positive.



By checking the variance and standard deviation values, we can say that data is much deviated from the mean that means there may be some outliers for all the attributes.

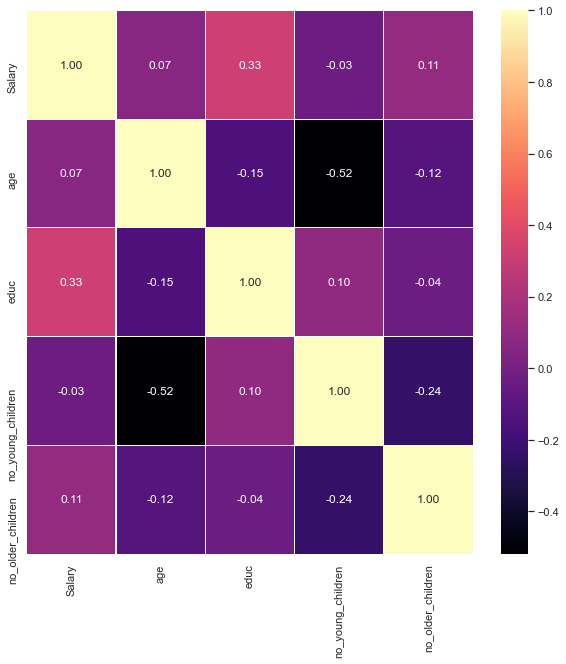


Let's check the pair plot to see the correlation between all the attributes.



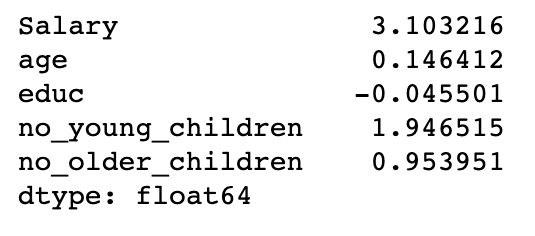
Here from the pair plot given above, we can say that the attributes are not that much correlated with each other.

Let's check heat map given below for further analysis

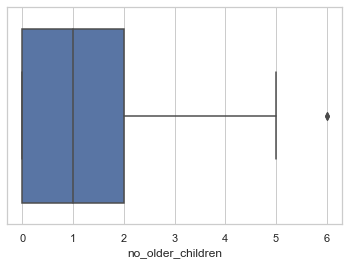
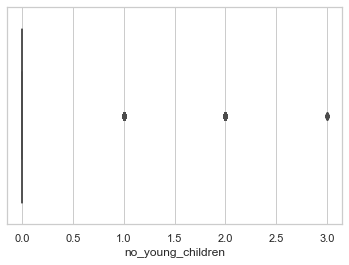
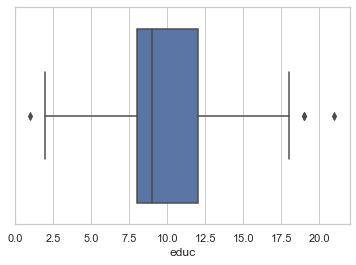
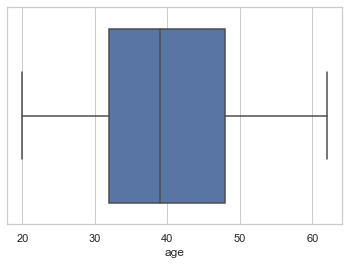
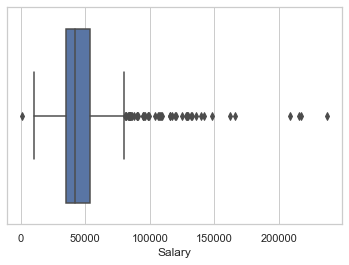


Employee salary and Years of formal education are somewhat correlated. But, there isn't much relation between other attributes.

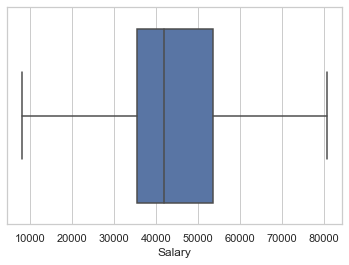
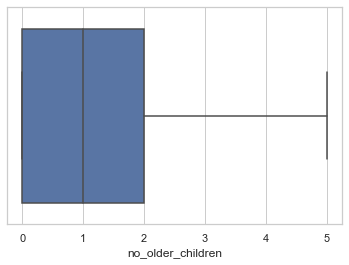
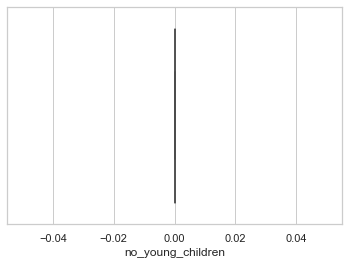
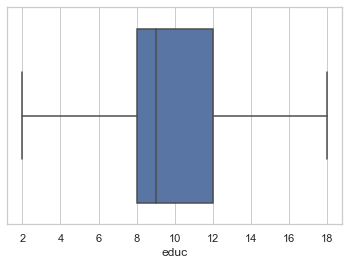
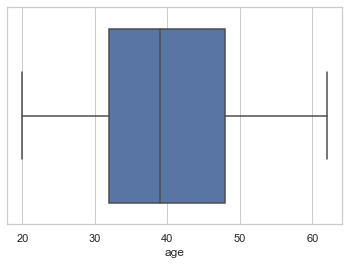
Let's check the skewness of every attribute that’s given in the picture below:



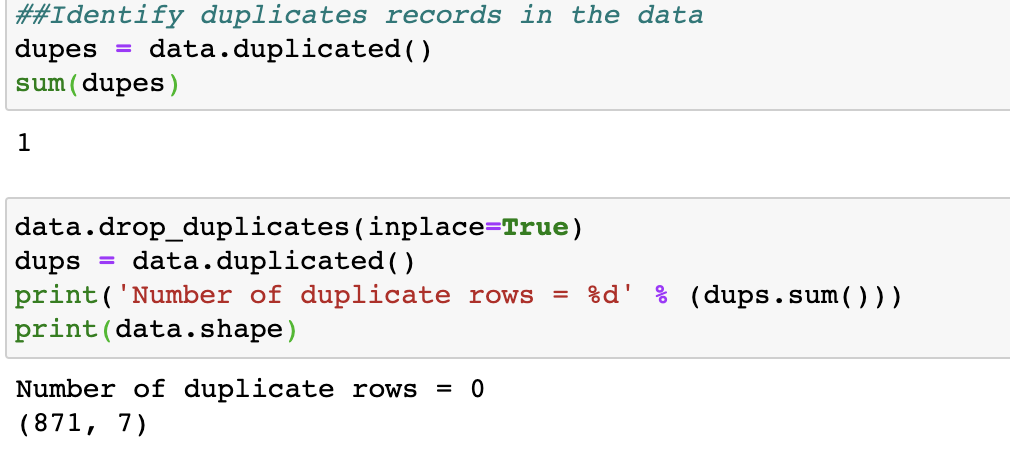
Years of formal education is negatively skewed and rest of the attributes are positively skewed

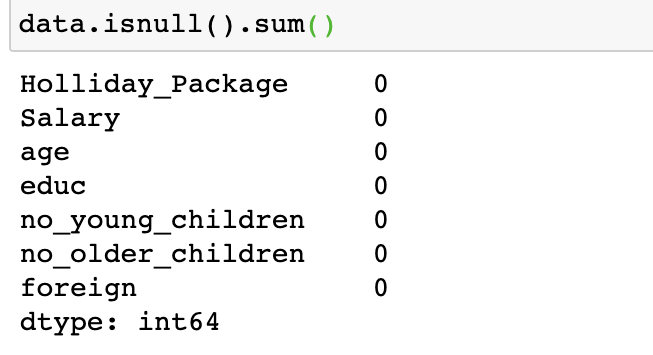


In the box-plots given above, we can see that all the attributes except age have outliers. Let’s treat them by defining custom function and check them again

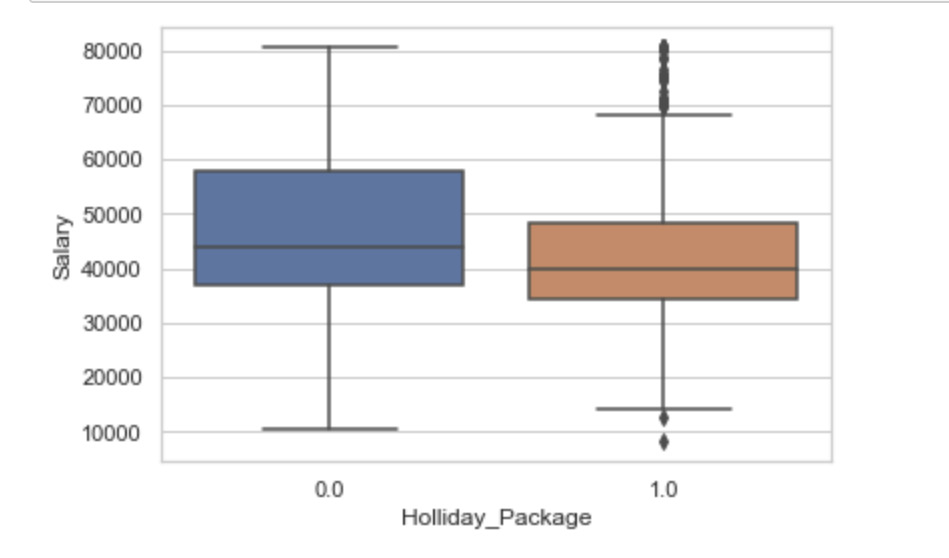


As seen in the above box-plots, there seems to be no null values. But, duplicates are in the dataset. Lets remove them.

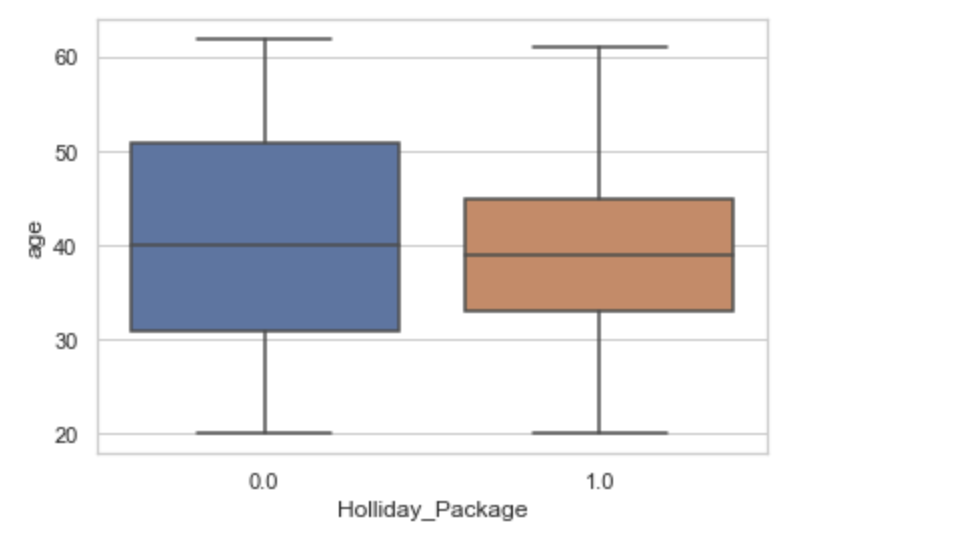




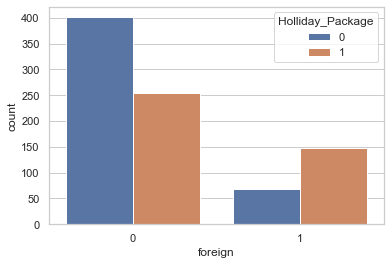
Let's check the distribution of every attribute with the target variable.

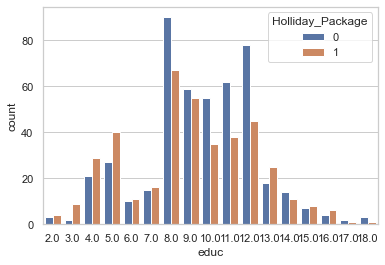
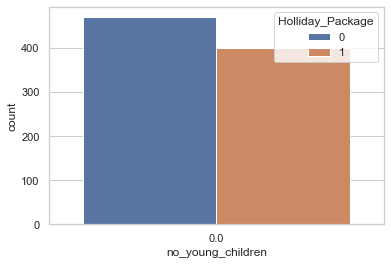


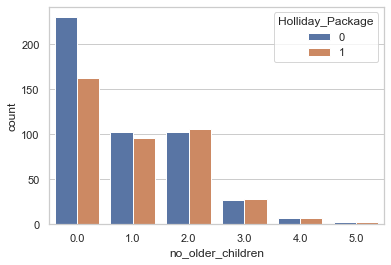
For salary attribute, it seems to be the people with high salary opted for the holiday packages most than the people with lowest salary



The age between 35 to 45 are going more for the Holliday\_Package than the rest of the people.



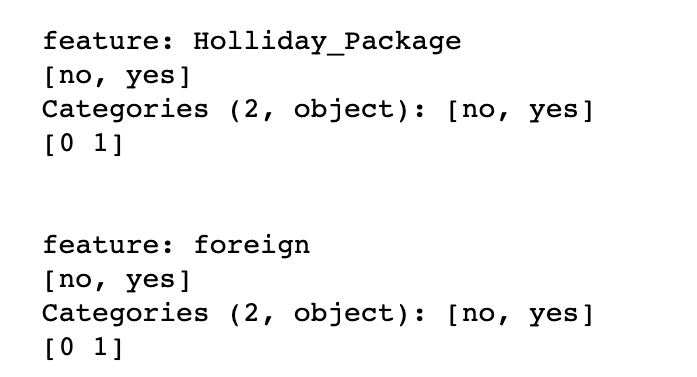




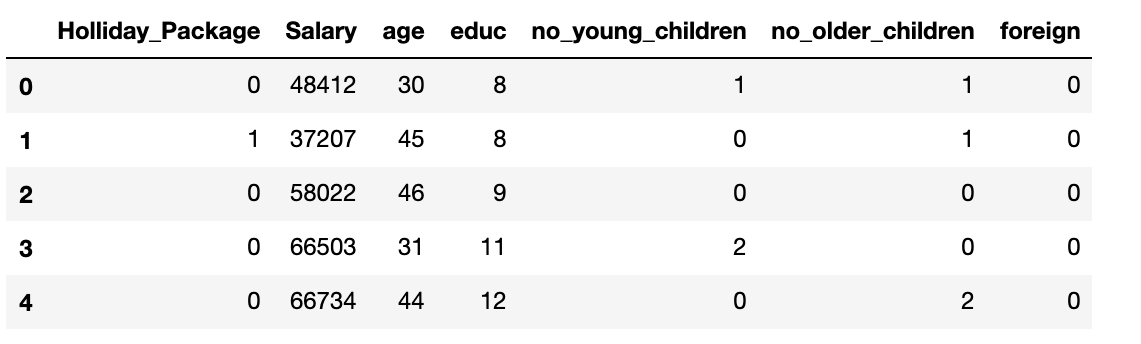
From the plots shown above, we can say that people with age ranged between 30 and 50 with a high salary and have less older children are opting for the holiday packages more. It seems that a number of years of formal education does not have that much impact on the target variable.

#### 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).

Converting object variables to categorical codes.



Let's check the first five rows in the dataset to check if the conversion is done or not.



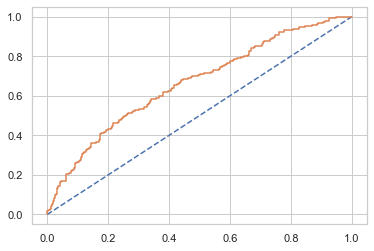
After splitting data into training and test set in 70:30 ratio and applying logistic regression and LDA

#### 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.

##### **Logistic model performance metrics**

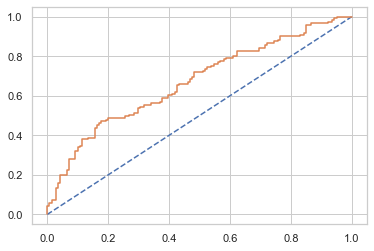
The accuracy score for the **training data** is 0.6272577996715928

* AUC = 0.663 (# AUC for the training data)



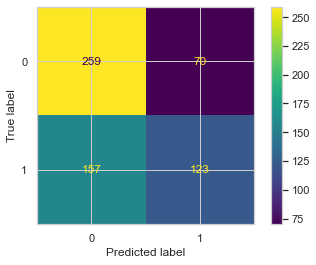
The accuracy score for the **test data** is 0.6564885496183206

* AUC = 0.663

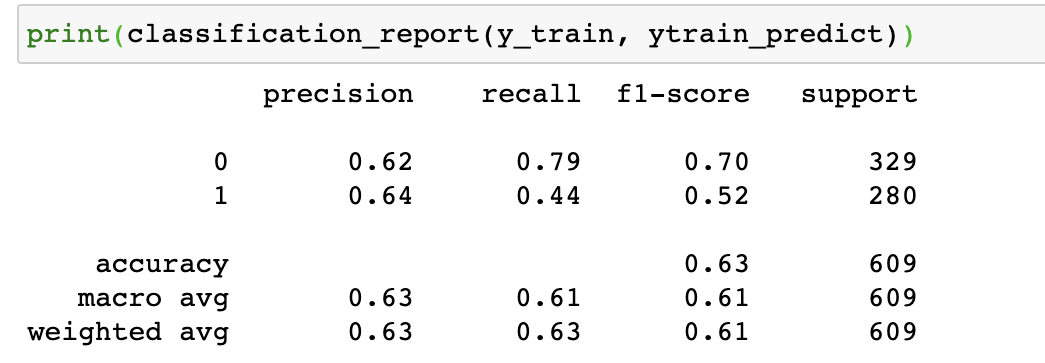


The accuracy and AUC are almost same for the **train dataset** and **test dataset**

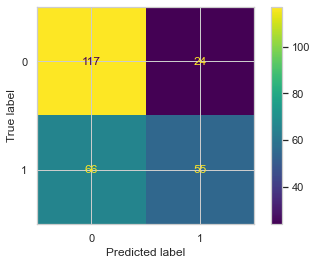
Confusion Matrix for the **training** **data**



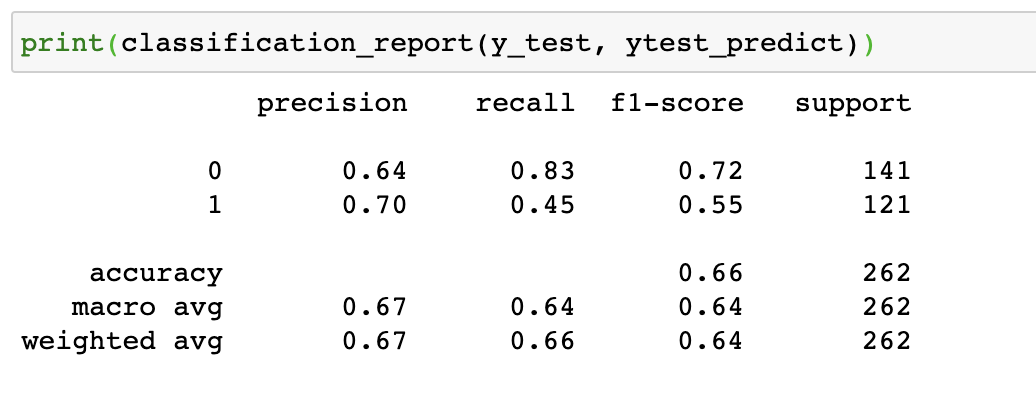
Classification report for the **training data**



Confusion Matrix for **test data**



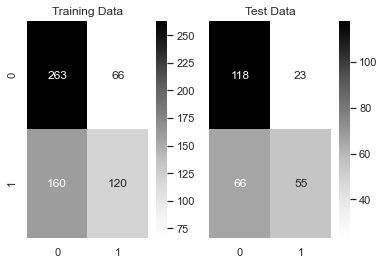
Classification report for the **test data**



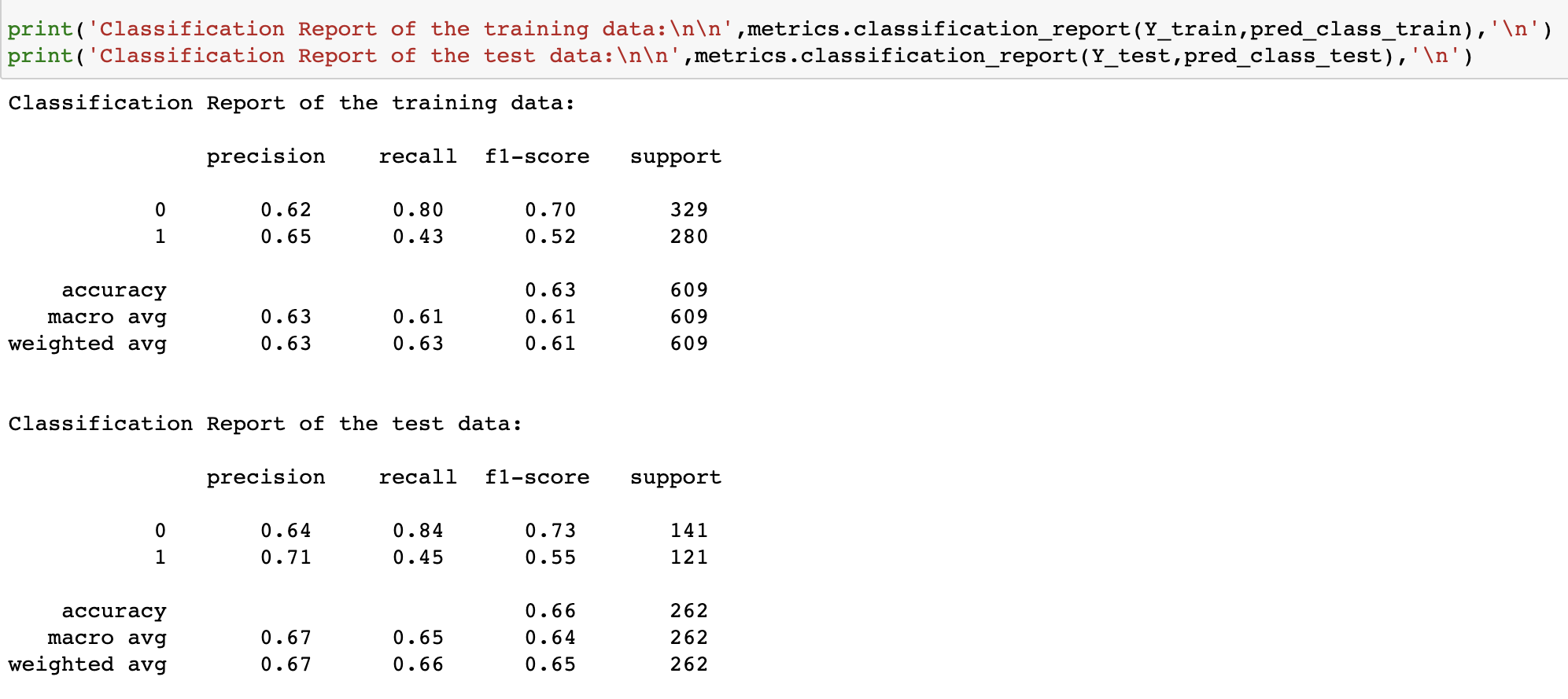
* tn - 117 (tn means “true negative”)
* tp - 55
* fp - 24
* fn - 66
* Since all the attributes are poorly correlated with each other and with the target variable, we got the low recall value(45)
* Poor predictors implies poor features

##### **LDA performance metrics**

Confusion Matrix for **Training Data** and **Test Data**



Classification Report for **Training Data** and **Test Data**



AUC for the

* Training Data = 0.663
* AUC for the Test Data = 0.670



* The recall value inferred from the above = 45, which is the same as the logistic model. Because of poor predictors, we got the low recall value.
* Test and training data are also almost the same as we got in logistic regression.
* We can see that there is a lot of difference between training and test data in the confusion matrix.

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

* The training and testing data is almost the same for both the models.
* Logistic regression gave better results than the LDA model.
* Because of the poor predictors, we didn't get a good recall value.
* We got the same recall value of 45 for the logistic model and LDA model. All the attributes aren't highly correlated with each other as well.
* The AUC score for both the training and testing dataset is almost the same for both the models.
* We can say that based on the predictions,
  + Employees with medium to high level salary and with age range between 30 and 45 are choosing holliday\_package.
  + Employees with very young children are not interested in choosing holliday\_package.
* It is recommended for the company to focus on employees with medium level salaries and who are in middle ages (30-45).
* They can neglect the factor years of formal education, since it is not showing much relation with the target variable or other attributes.
* Foreigners are also not the main factor but they can give the second importance to that factor.
* Employees with children who are not very young and very old are opting for holiday packages
* The important factors that company can focus on to sell their packages are salary and age