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.

Data set :

Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
no	48412	30	8	1	1	no
yes	37207	45	8	0	1	no
no	58022	46	9	0	0	no
no	66503	31	11	2	0	no
no	66734	44	12	0	2	no
yes	61590	42	12	0	1	no
no	94344	51	8	0	0	no
yes	35987	32	8	0	2	no
no	41140	39	12	0	0	no
no	35826	43	11	0	2	no
no	42643	45	11	0	2	no
no	35157	60	12	0	0	no
no	75327	33	11	2	0	no
no	148221	56	14	0	0	no
no	98870	56	11	0	0	no
no	80297	47	11	0	1	no
no	52117	50	8	0	0	no
yes	139253	39	12	0	0	no
no	62858	47	8	0	1	no

We are provided with the above data set of 872 rows and 7 columns. Of the above columns, five columns are integer data type and two columns are of object data type.

All the columns in the given dataset have no Null values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 7 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   Holliday_Package   872 non-null    object
 1   Salary             872 non-null    int64 
 2   age                872 non-null    int64 
 3   educ               872 non-null    int64 
 4   no_young_children  872 non-null    int64 
 5   no_older_children  872 non-null    int64 
 6   foreign            872 non-null    object
dtypes: int64(5), object(2)
memory usage: 47.8+ KB
```

Descriptive statistics for the dataset:

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
count	872	872.000000	872.000000	872.000000	872.000000	872.000000	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	47729.172018	39.955275	9.307339	0.311927	0.982798	NaN
std	NaN	23418.668531	10.551675	3.036259	0.612870	1.086786	NaN
min	NaN	1322.000000	20.000000	1.000000	0.000000	0.000000	NaN
25%	NaN	35324.000000	32.000000	8.000000	0.000000	0.000000	NaN
50%	NaN	41903.500000	39.000000	9.000000	0.000000	1.000000	NaN
75%	NaN	53469.500000	48.000000	12.000000	0.000000	2.000000	NaN
max	NaN	236961.000000	62.000000	21.000000	3.000000	6.000000	NaN

We have columns **'Holliday_Package'** and **'foreign'** are **categorical type** data and columns **'Salary','age',educ', 'no_young_children' and 'no_older_children' are integer type** data**.**

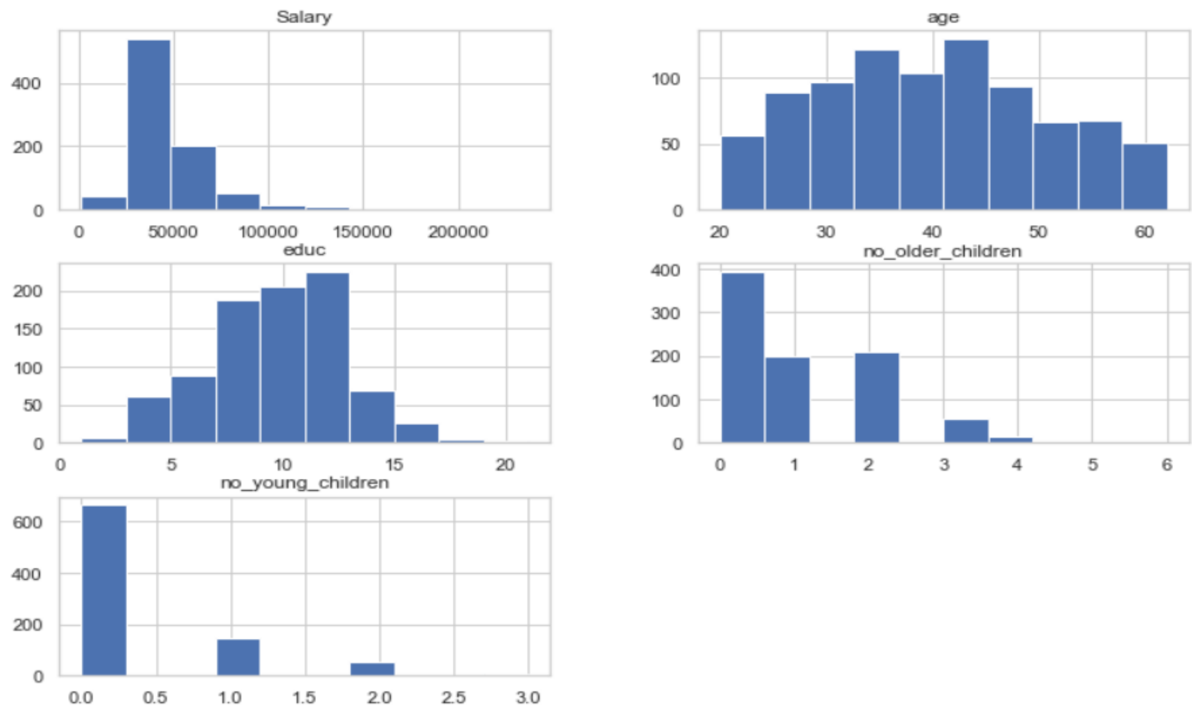
As per the details resulted from the descriptive statistics of the dataset, we can find that:

All the variables have zero null values.

Of the entire dataset, column **'Salary'** has **highest max** value of 236961 and columns **'no_young_children' and 'no_older_children'** have **least min** value of 0.00

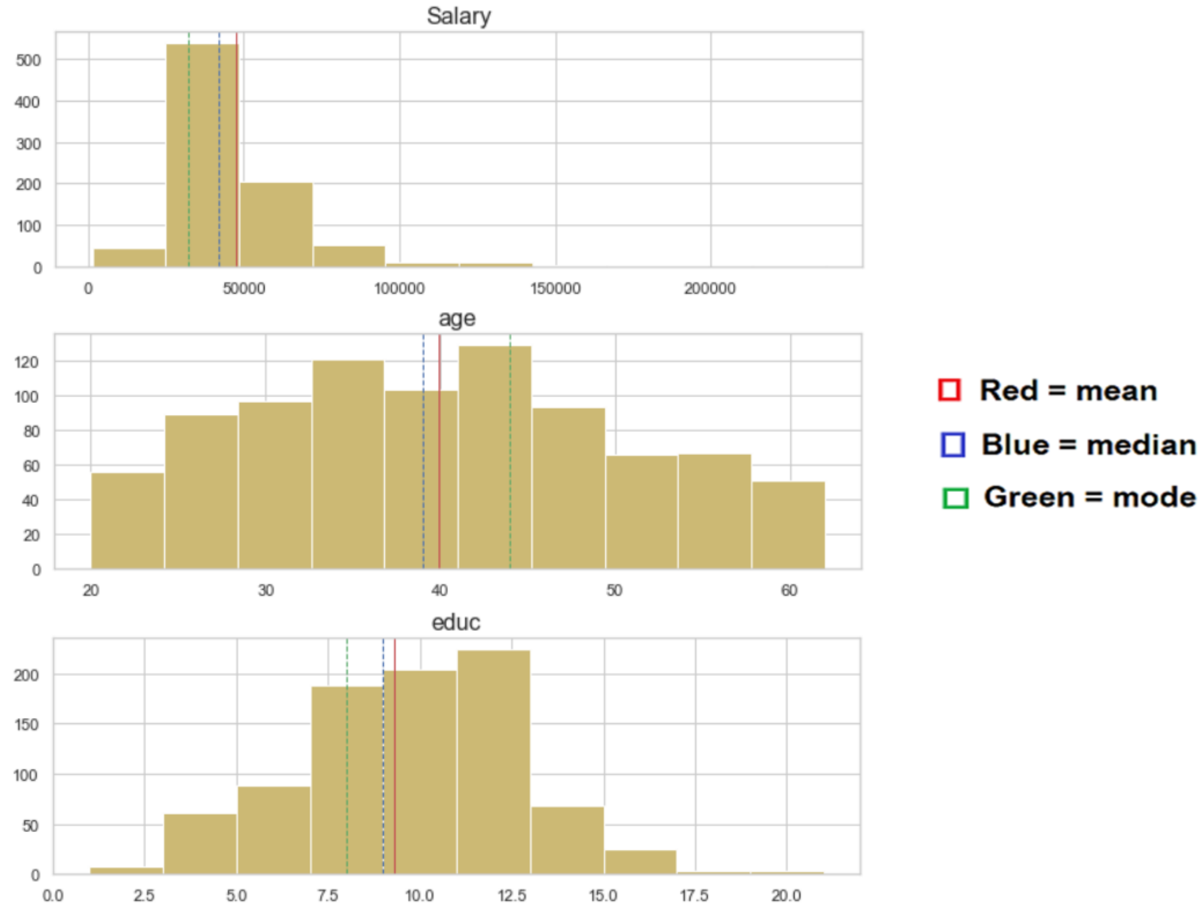
The zero value in **'no_young_children' and 'no_older_children'** means there is no younger and older children for the respective employee.

Columns **'Salary'** and **'no_young_children'** have highest mean value – 47729.172 and least mean value – 0.311927 respectively.

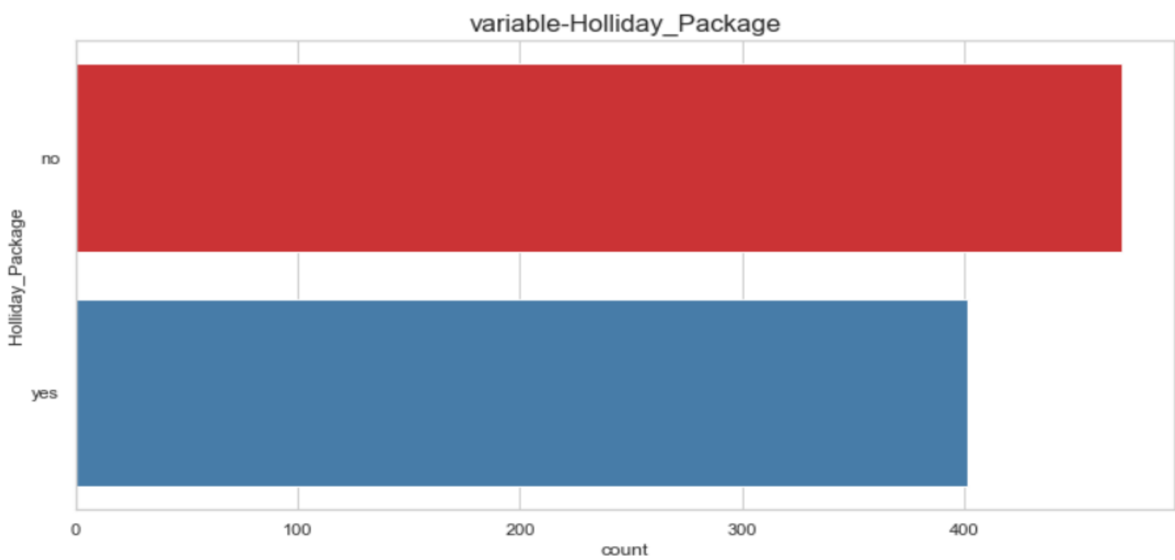


From the above histograms of the variables , we can see that majority of the variables are not symmetrical.

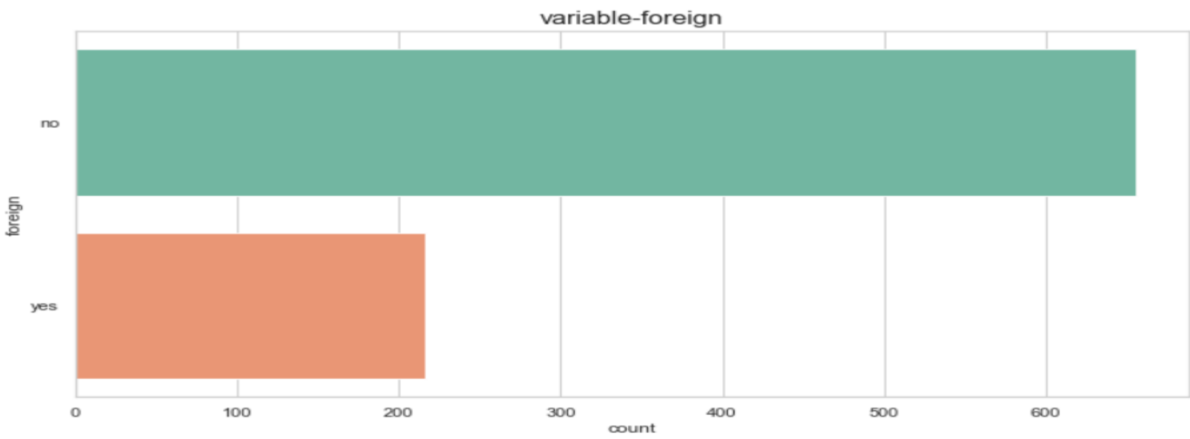
Among all the variables , Variable **'Salary'** is highly **right skewed** (skew = 3.103216) and Variable **'educ'** is highly **left ske wed** (skew = -0.045501).



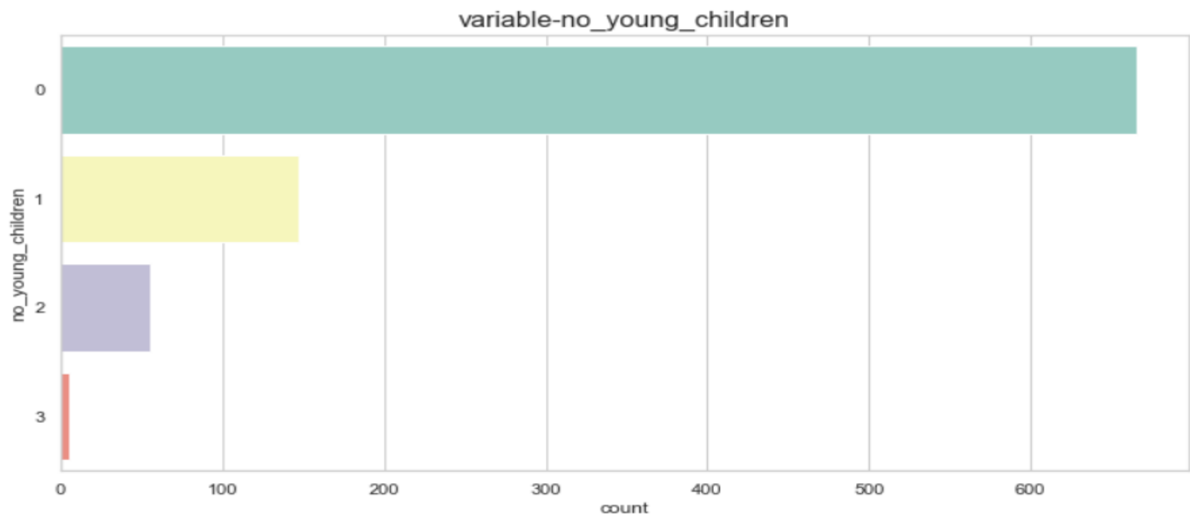
Categorical Data type:



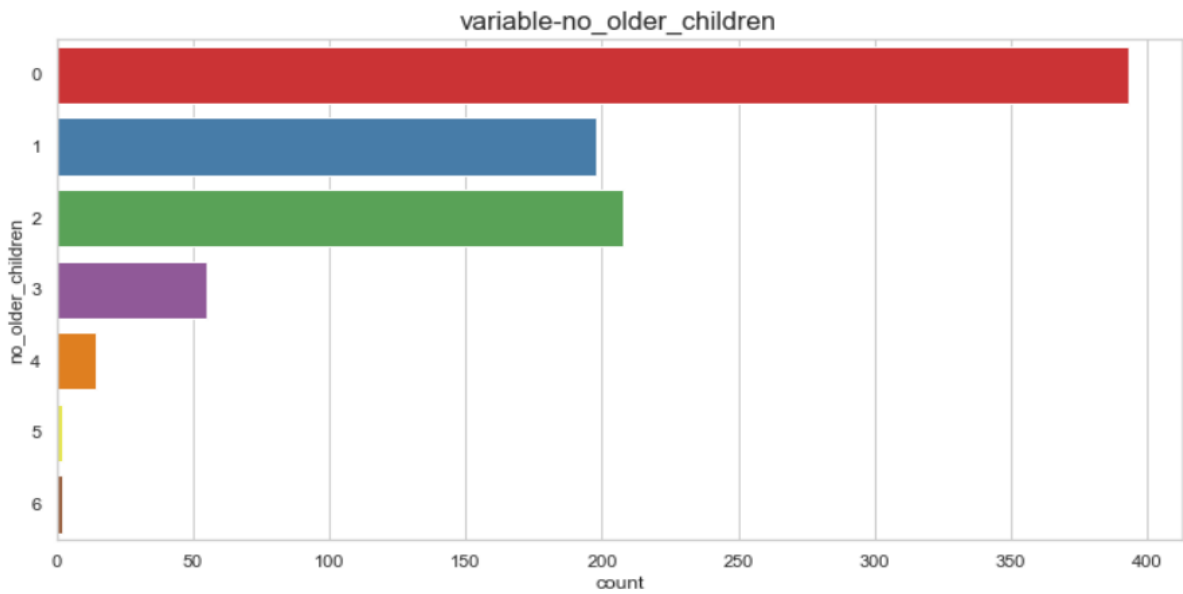
Variable **'Holliday_Packagge'** has highest presence of **'no'** and least presence of **'yes'**



Variable **‘foreign’** has highest presence of **‘no’** and least presence of **‘yes’**



Variable **‘no_young_children’** has highest presence of **‘0’** and least presence of **‘3’**

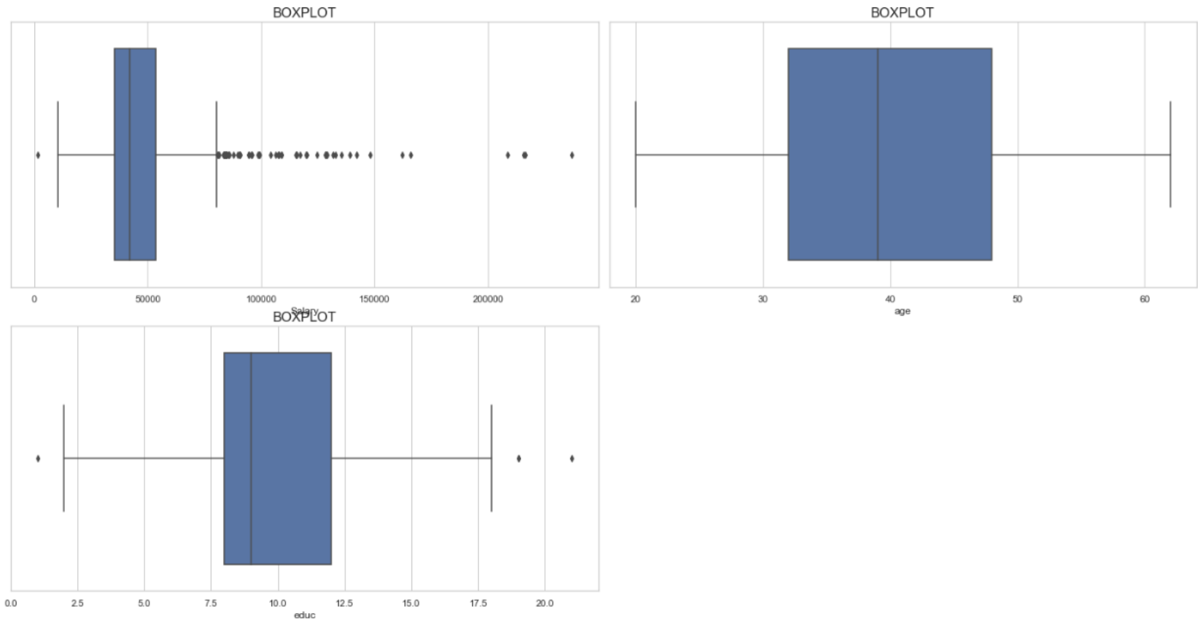


Variable **‘no_older_children’** has highest presence of **‘0’** and least presence of **‘6’**

Skewness:

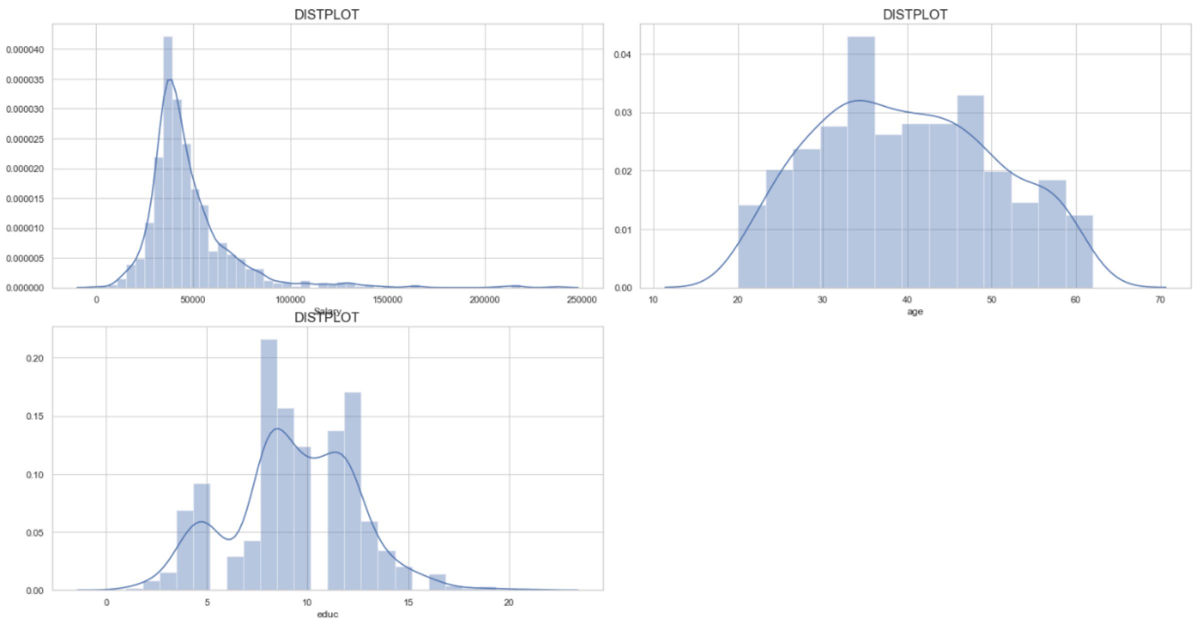
```
Salary                  3.103216
age                     0.146412
educ                   -0.045501
no_young_children       1.946515
no_older_children       0.953951
dtype: float64
```

<u>Boxplot distribution:</u>

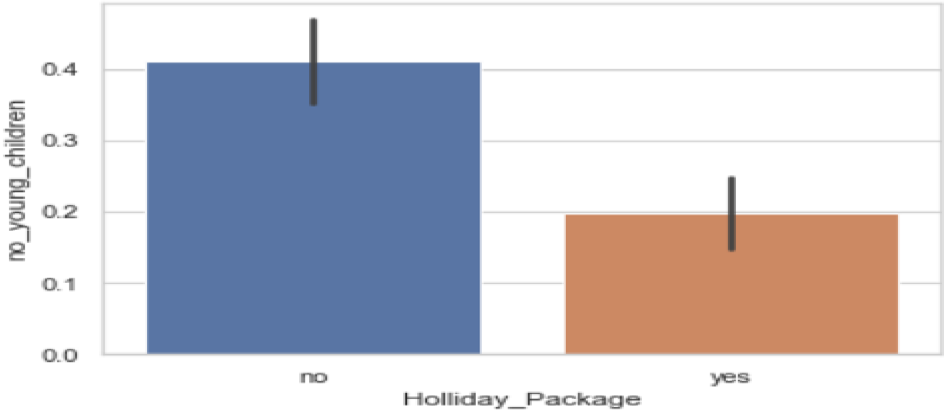


The continuous variable ‘Salary’ has outliers in the given Dataset.

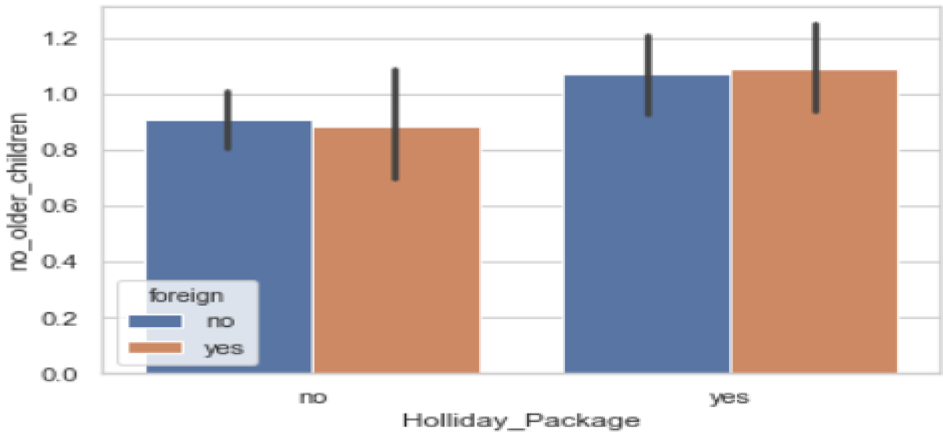
<u>DistPlot distribution:</u>



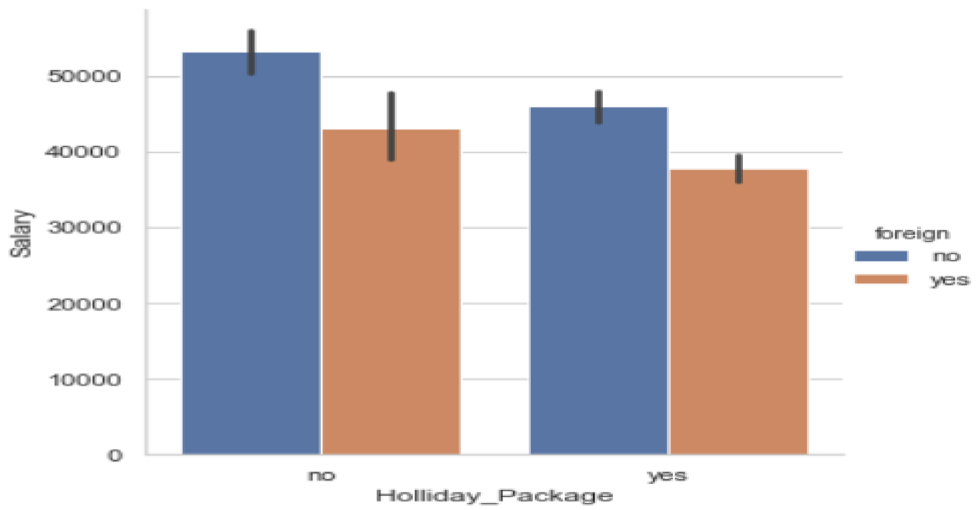
Bivariate Analysis:



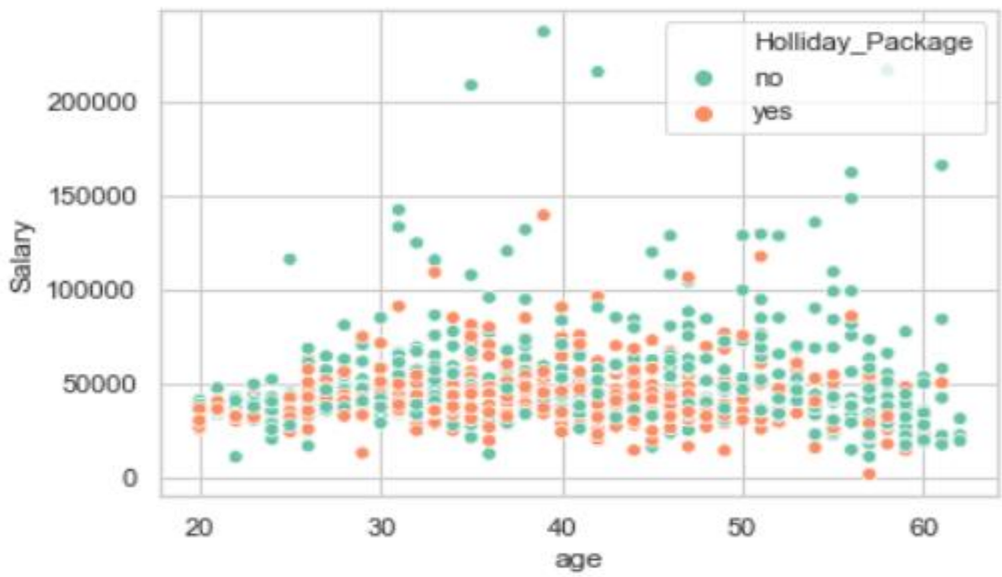
Employees with lower number of young children opted for Holiday Package more compared with higher number of younger children.



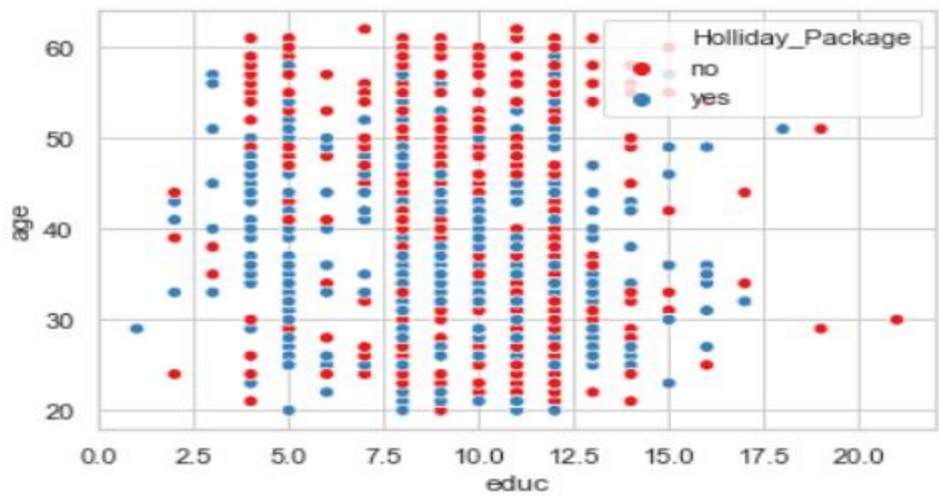
Employees with higher number of older children opted for Holiday Package more compared with lower number of older child ren.



Employees with lower salary opted for Holiday Package little more compared with higher salaried employees. Also majority of employees are not foreigners.



From above plot , we can see Holiday_Package is less preferred with higher salary.



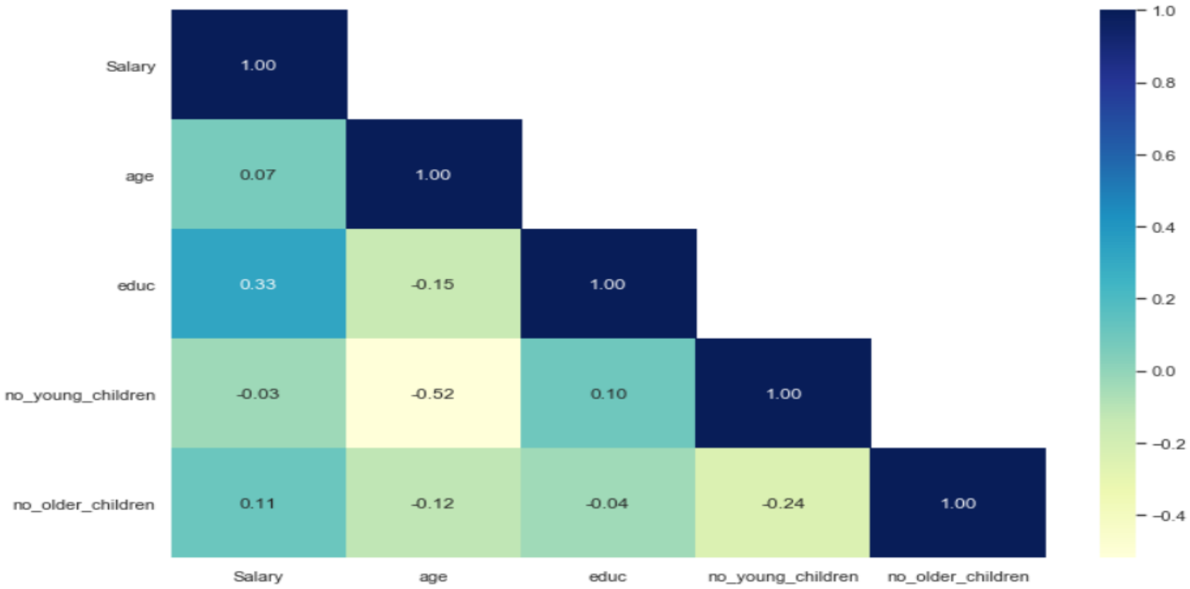
Most of the data points are in the range of 7-13 years of formal education and as age of employees increases, the employee s opted for Holiday Package decreases.

Multivariate Analysis:

We have the following correlation among the different variables given in the dataset.

	Salary	age	educ	no_young_children	no_older_children
Salary	1.000000	0.071709	0.326540	-0.029664	0.113772
age	0.071709	1.000000	-0.149294	-0.519093	-0.116205
educ	0.326540	-0.149294	1.000000	0.098350	-0.036321
no_young_children	-0.029664	-0.519093	0.098350	1.000000	-0.238428
no_older_children	0.113772	-0.116205	-0.036321	-0.238428	1.000000

HeatMap:



From the above map , we can see that many columns are co-related to each other and there is **highest positive correlation** (0.33) between variables **'educ'** and **'Salary'** . Also there is **highest negative correlation**(- 0.52) between variables **'no_young_children'** and **'age'**.

Pairplot:

In the above plot scatter diagrams are plotted for all the columns in the dataset. From the visual representation , we can understand the degree of correlation between any two columns of the given dataset.

Variables **'age' and 'educ'** shows positive linear correlation with Variable **salary'**.

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).

The given dataset contains two categorical variables 'cut','color' and 'quality'.

```
HOLLIDAY_PACKAGE :  2      FOREIGN :  2
yes     401                yes     216
no      471                no      656
```

We have to encode the data in these variables to use them in the models,

	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	0	48412	30	8	1	1	0
1	1	37207	45	8	0	1	0
2	0	58022	46	9	0	0	0
3	0	66503	31	11	2	0	0
4	0	66734	44	12	0	2	0

Splitting data into train and test (70:30)

```
x_train.head()
```

	Salary	age	educ	no_young_children	no_older_children	foreign
821	38974.0	47	12	0	2	1
805	40270.0	33	8	2	0	1
322	32573.0	30	11	1	0	0
701	43839.0	43	11	0	1	1
773	33060.0	40	5	1	1	1

```
x_test.head()
```

	Salary	age	educ	no_young_children	no_older_children	foreign
264	25118.0	58	8	0	0	0
189	40913.0	20	9	1	0	0
643	28446.0	58	8	0	0	0
65	36072.0	35	4	0	2	0
241	52736.0	40	10	0	3	0

```
y_train.head()

821    0
805    0
322    0
701    1
773    1
Name: Holliday_Package, dtype: int8
```

```
y_test.head()

264    1
189    0
643    0
65     1
241    0
Name: Holliday_Package, dtype: int8
```

Among all the variables of the given dataset, Column **'Salary'** has Outliers. So, we impute these values with lower range values (Q1-(1.5 * IQR)) and higher range values (Q1+(1.5 * IQR)) for the respective values.

Logistic Regression:

```
LogisticRegression(max_iter=10000, n_jobs=2, penalty='none', solver='newton-cg',
                   verbose=True)
```

Model score for **train data** is **0.6672131147540984**

Model score for **test data** is **0.648854961832061**

LDA (linear discriminant analysis):

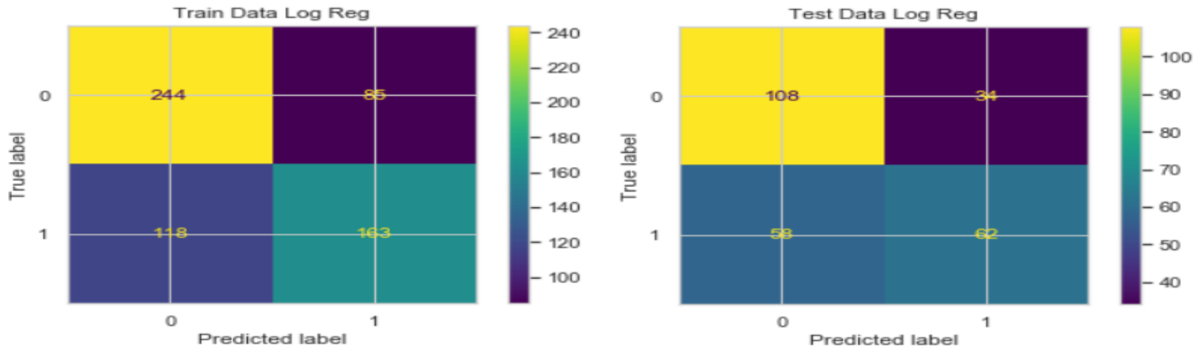
Model score of the **training data** is **0.66**

Model Score of the **testing data** is **0.65**

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 Regression model:

Confusion Matrix:



Classification Report:

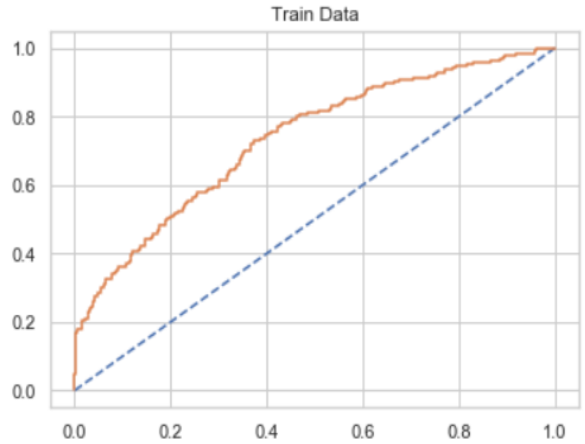
Train data :

```
              precision    recall  f1-score   support

           0       0.67      0.74      0.71       329
           1       0.66      0.58      0.62       281

    accuracy                           0.67       610
   macro avg       0.67      0.66      0.66       610
weighted avg       0.67      0.67      0.66       610
```

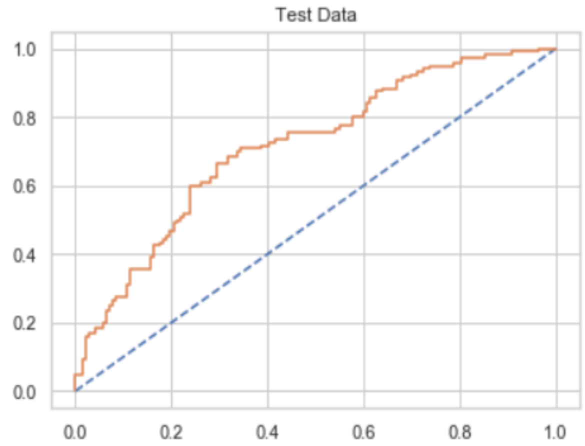
Test data :

```
              precision    recall  f1-score   support

           0       0.65      0.76      0.70       142
           1       0.65      0.52      0.57       120

    accuracy                           0.65       262
   macro avg       0.65      0.64      0.64       262
weighted avg       0.65      0.65      0.64       262
```

ROC curve :



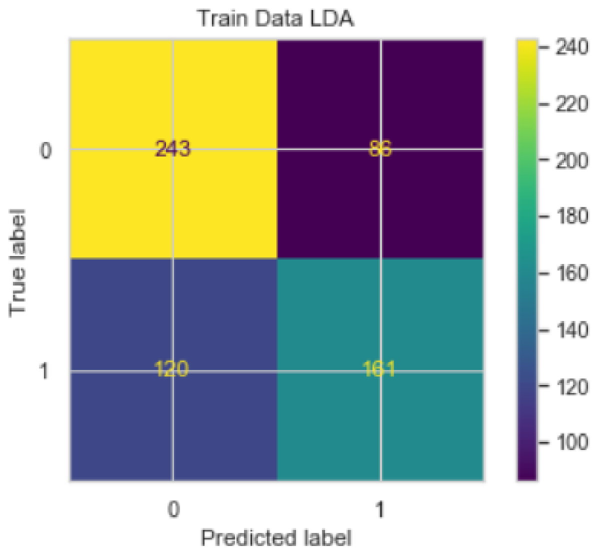
AUC - 0.733

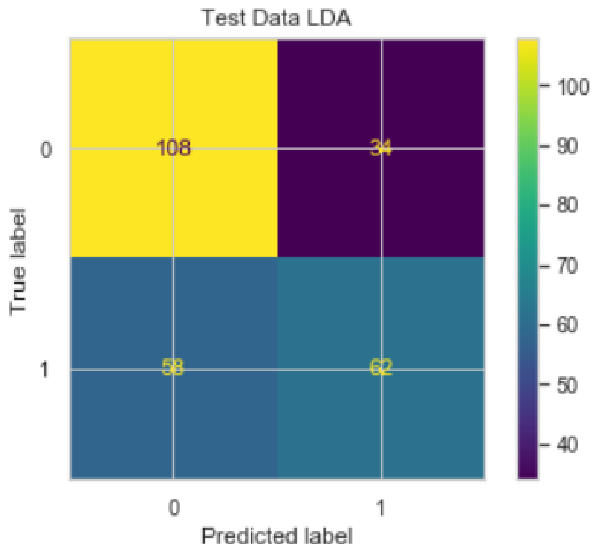


AUC - 0.715

LDA (linear discriminant analysis) model:

Confusion Matrix:





Classification Report:

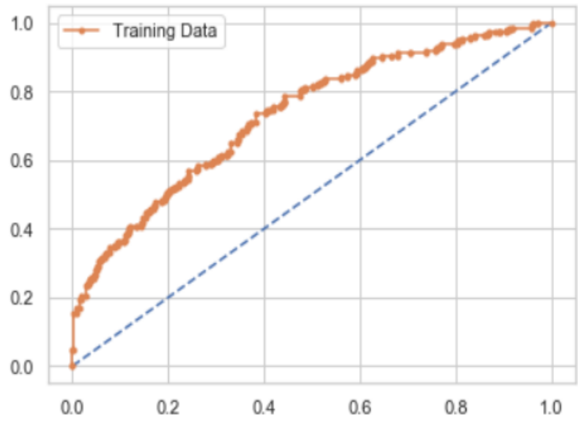
Train data :

	precision	recall	f1-score	support
0	0.67	0.74	0.70	329
1	0.65	0.57	0.61	281
accuracy			0.66	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.66	0.66	610

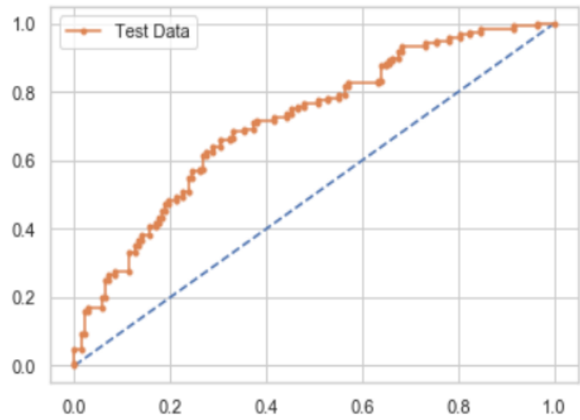
Test data:

	precision	recall	f1-score	support
0	0.65	0.76	0.70	142
1	0.65	0.52	0.57	120
accuracy			0.65	262
macro avg	0.65	0.64	0.64	262
weighted avg	0.65	0.65	0.64	262

ROC curve :



AUC - 0.731



AUC - 0.714

Comparing the performance metrics from the two models,

Looking at the details got from **test data** from the two models ,

Accuracy : Both models have equal value of 0.65

AUC : Logistic Reg model has value of 0.715 and LDA model has least value of 0.714

Recall : Both models have equal value of 0.52

Precision : Both models have equal value of 0.65

F1 Score : Both models have equal value of 0.57

We know that linear discriminate analysis and logistic regression are the most widely used statistical methods for analyzing categorical outcome variable. While both are appropriate for the development of linear classification models, linear discriminate analysis makes more assumptions about the underlying data. Hence, it is assumed that logistic regression is the more flexible and more robust method in case of violations of the assumptions also logistic regression is preferred when the dependent variable is dichotomous, while discriminant analysis is preferred when it is nominal (more than two groups).

Therefore we can use logistic regression model in predicting whether an employee will opt for the package or not.

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

Due to the importance of understanding and managing the risks in volatile business domains, it is required to find an effective aid in making decisions. The results from models show that the above algorithms are a promising opportunity in predicting whether an employee will opt for the package or not through the cause and effect relationship between the independent and dependent variables of the given dataset.

The above model will be helpful in predicting the dependent variables through the independent variables by assigning the probability of employee opting for the package to the every predictor variable to give the best predictive/dependent variable.

The proportion of the True positive(TP) to Predicted positive(TP+FP) is good for the models. So they will be useful in predicting the target variable.

As per predictions of the model, we have got the following coefficients for the independent variables of the given dataset.

The coefficient of the different attributes of the given dataset are:

The coefficient for Salary is -1.949312635383068e-05
The coefficient for age is -0.058445029274936756
The coefficient for educ is 0.055894462070796756
The coefficient for no_young_children is -1.363217023869716
The coefficient for no_older_children is -0.057106551906486225
The coefficient for foreign is 1.247624094266054

'no_young_children' is the most important feature among all the features of the dataset.

Employees with lower number of young children opted for Holiday Package more compared with higher number of younger children. Majority of employees are not foreigners.

The company must target employees based on the feature importance to get the better results in predicting whether an employee will opt for the package or not.

So, The Overall analysis of given dataset definitely helped to get insights that would help the company for the business development.