Holiday Package Analysis

Problem2: Logistic Regression and LDA

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
0	no	48412	30	8	1	1	no
1	yes	37207	45	8	0	1	no
2	no	58022	46	9	0	0	no
3	no	66503	31	11	2	0	no
4	no	66734	44	12	0	2	no
...	...	...	...	...	...	...	...
867	no	40030	24	4	2	1	yes
868	yes	32137	48	8	0	0	yes
869	no	25178	24	6	2	0	yes
870	yes	55958	41	10	0	1	yes
871	no	74659	51	10	0	0	yes

872 rows × 7 columns

- Read data from csv file and convert it into Data Frame
- For reading the data from csv file I use read_csv method of pandas

Data Dictionary:

- Data Dictionary means what all columns represent in dataset.
- Following table contain Column name and Description

Variable Name	Description
Holiday Package	Opted for Holiday Package yes/no?
Salary	Employee salary

age	Age in years
edu	Years of formal education
no_young_children	The number of young children (younger than 7 years)
no_older_children	Number of older children
foreign	foreigner Yes/No

Descriptive Statistic:

```
df.describe()
```

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

- From above figure we can say average age of people in dataset is 39.95 year
- Average education of people is 9.3 year
- Oldest person in dataset is 62 year old and youngest person is 20 year old
- Minimum education is 1 year and maximum education is 21 years.

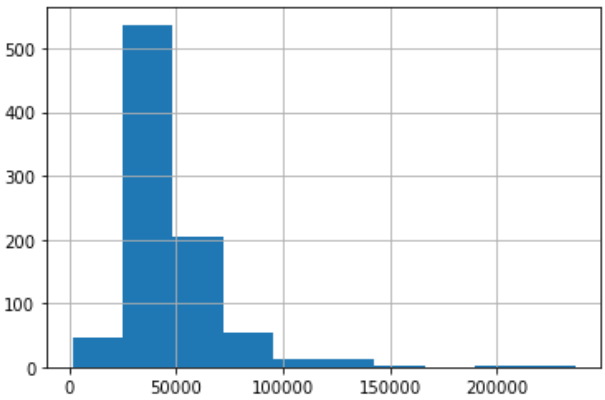
Check the null values:

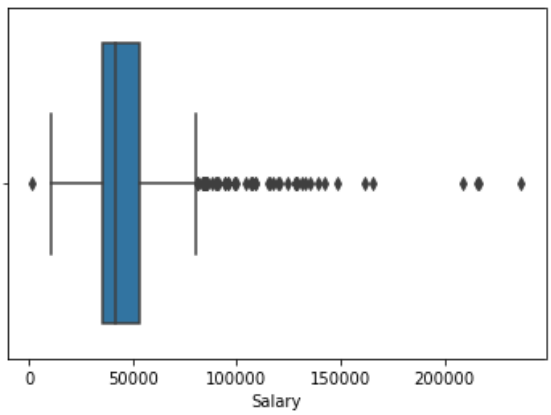
```
df.info()
```

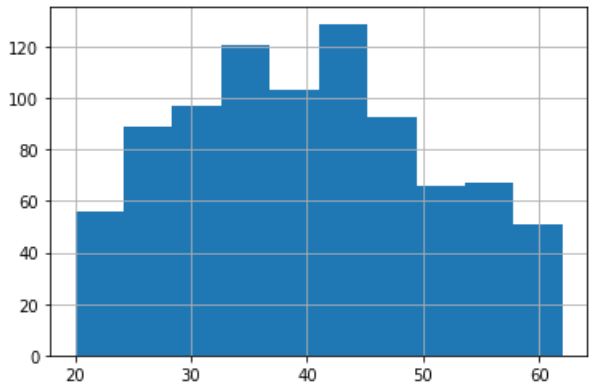
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 9 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Holliday_Package      872 non-null    object
 1   Salary                872 non-null    int64
 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
 7   foreign_num           872 non-null    int32
 8   Holliday_Package_num  872 non-null    int32
dtypes: int32(2), int64(5), object(2)
memory usage: 54.6+ KB
```

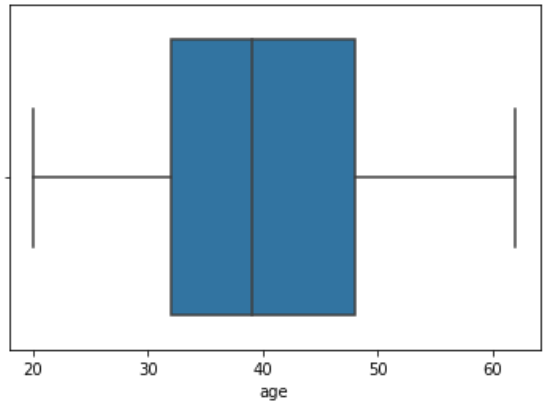
- We can see from upper image there is 0 null value in dataset.
- Holiday package and foreign columns have object data type.
- Salary, age, education, young children, older children contain integer datatype.
- There is total 872 rows and 9 columns present in data set.

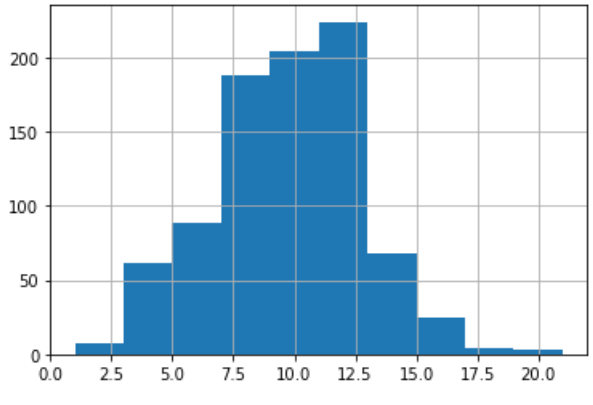
Univariate Analysis:

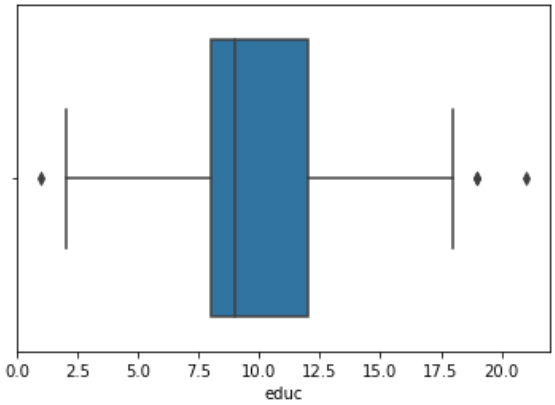


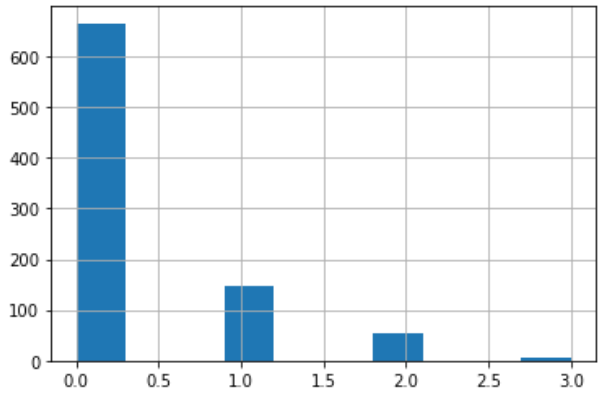


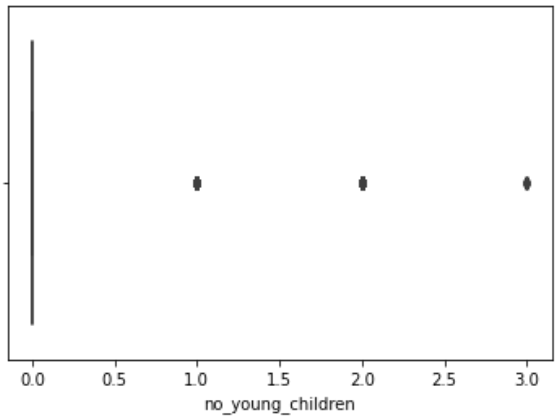


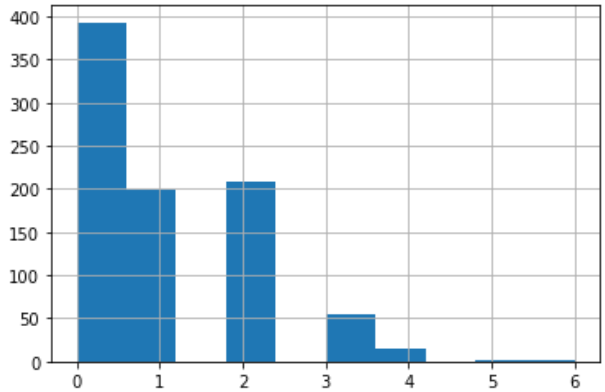


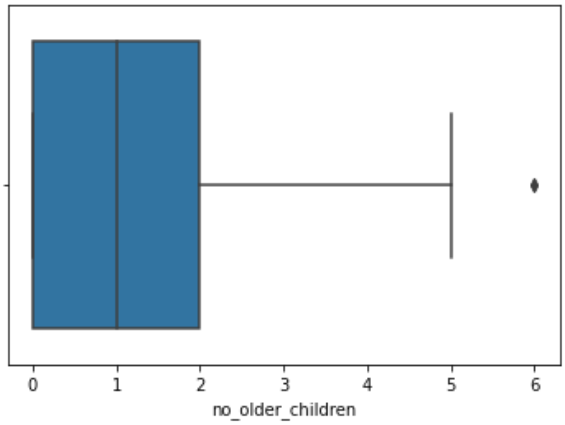








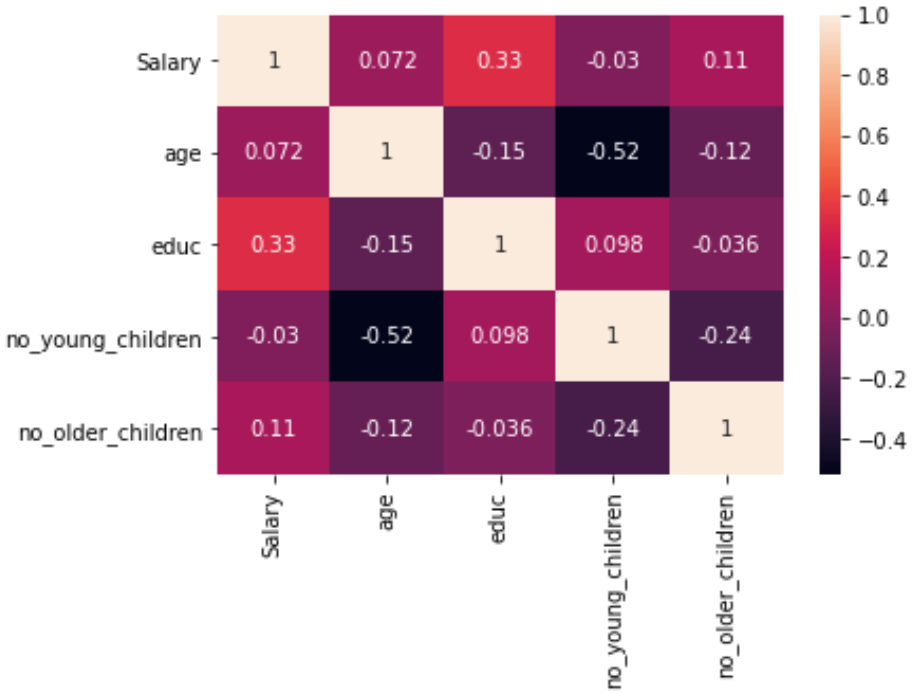




- We can say there is outlier present in salary column from boxplot.
- There are also outlier present in educ, no_young_children and no_older_children but it's not removable because the range is very short for that column.

- There is no outlier present in age column.

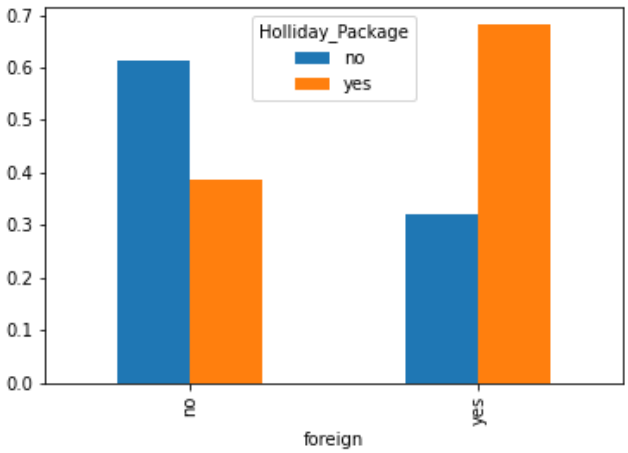
Heat Map:



- In heat map all dark shad are highly correlate each other.
- Age and number of young children are highly negative correlated.
- Number of young children and number of older children are also highly negative correlated.
- Education and salary are in positive correlation.

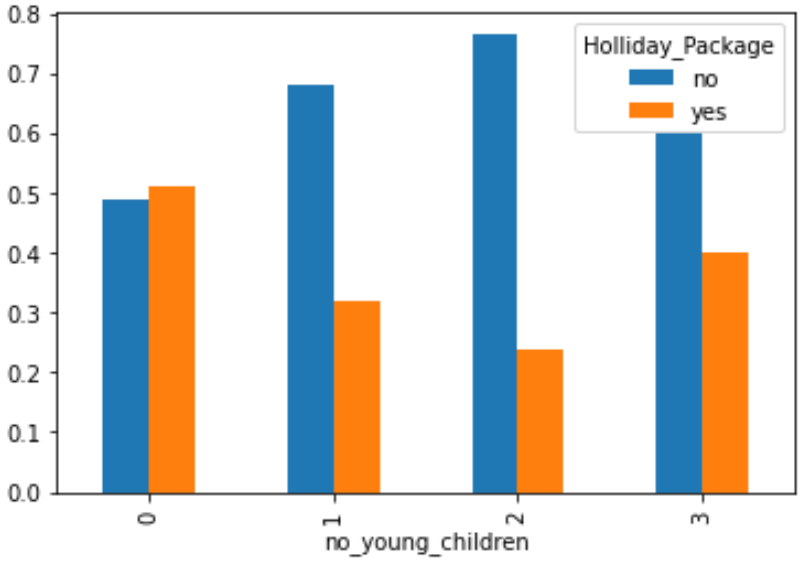
Bivariate Analysis:

- **Impact of nationality on Holiday package buying**



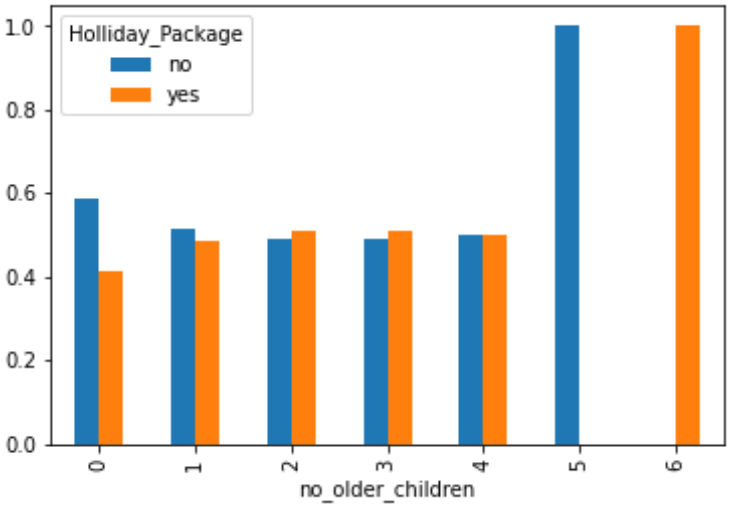
We can see from upper image person who live in foreign they are more prefer to buy holiday package.
We can see from graph around 60% people who lives in India they don't buy holiday package and only 40% people prefer to buy package.
People who live in foreign they more likely buy holiday package and ratio of holiday package buying and not buying for foreign people is 70:30 means 70% people are interested to buy holiday package.

- **Impact of no_young_children on Holiday package buying**



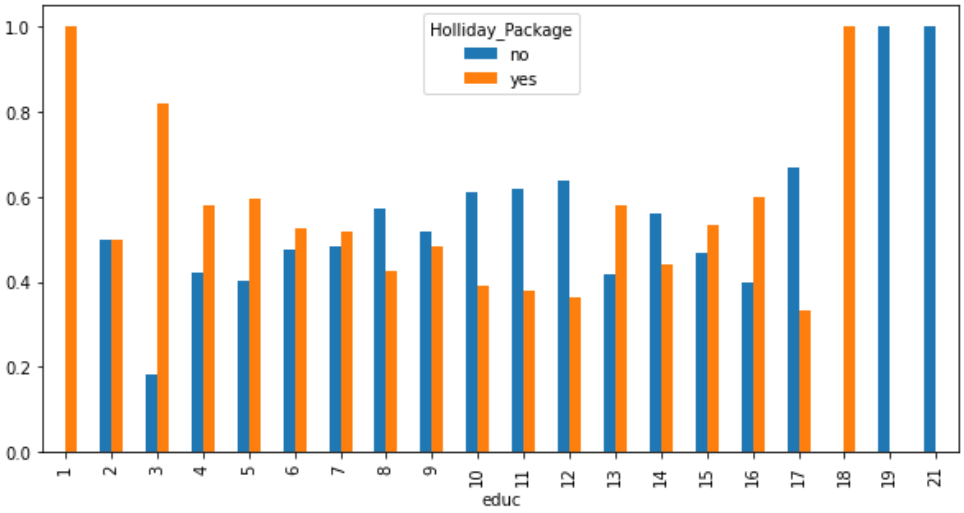
We can see from upper figure people who are not parent means which have 0 young children they are more prefer to buy holiday package.
As number of young children increase then the chance of holiday package buying is decrease so company should focus on people who don't have young children.

- **Impact of no_older_children on Holiday package buying**



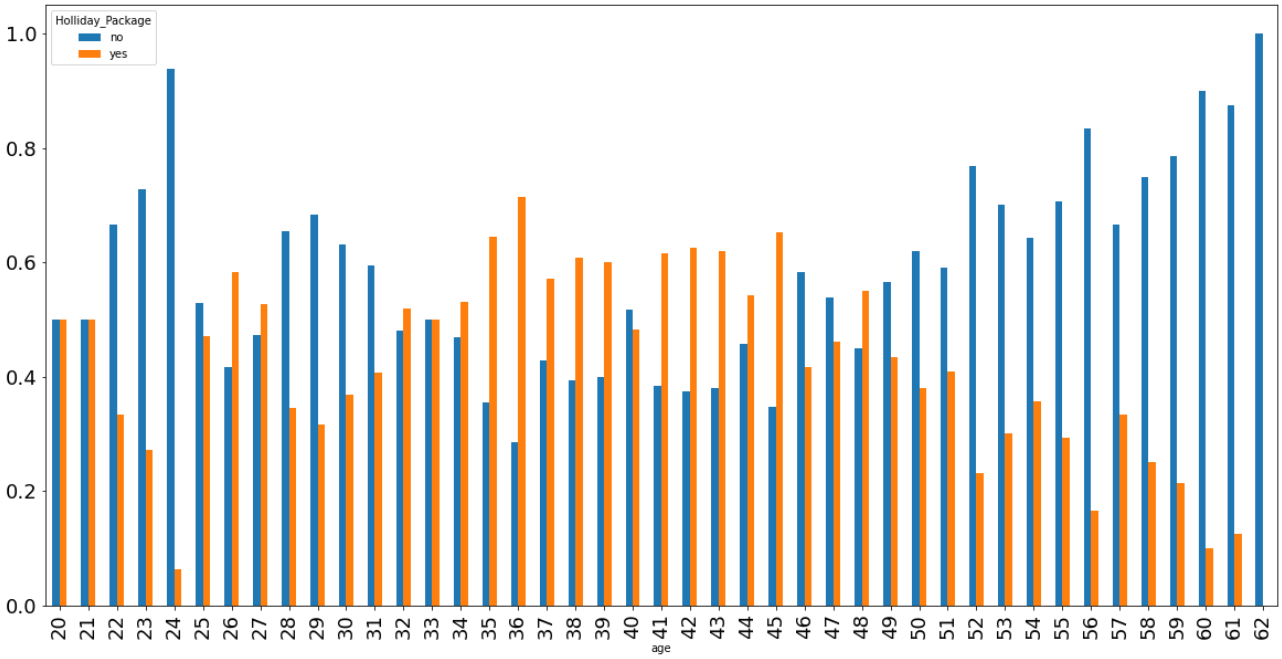
I don’t know why but according to figure we can say 100% people which have 6 older children they buy holiday package.
It is also clear from graph which people have 5 older children they don’t like to buy holiday package.
Which people have 1,2,3 or 4 older children, the chance of holiday package buying of them is around 50%.

- **Impact of education on Holiday package buying**



We can see from upper figure, people who have formal education in range 1-5 they are more prefer to buy holiday package.
As the formal education increase from 7 to 17 there is very low chance to buy holiday package.
Which people are more educated means which have around 19-20 years formal education they don’t buy the holiday package.

- **Impact of Age on Holiday package buying**



We can see from upper graph, people which age in range 20-21 they are preferring to buy holiday package and around 50% people in group range are prefer to buy package.
People which are under age group of 22-31 there is very low chance to buying holiday package.
People whose age in range 32-45 there is very high chance to buying holiday package so company should focus on that age group people for selling package.
As shown figure people whose age increase from 46 to 62, there is chance of holiday package buying is decrease and at the end of age range 60,61 and 62 there very very low chance to buying package so company should take action accordingly.

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).

- **Encode The Data:**
  Machine Learning model don't understand categorical data so first we must convert all columns which contain categorical data into numerical data.
  Here only foreign Colum contain categorical data so I convert it as following.

Foreign	Foreign_num
Yes	1
No	0

  - There are no categorical data except foreign column so there is no need do data encoding for that columns.

- **Split Data**
  - I use train_test_split method to split dataset into two parts 1) Training dataset 2) Testing dataset.
  - Using this method I devide dataset into 70:30 ratio for better model training

```
x_train
```

	Salary	age	educ	no_young_children	no_older_children	foreign_num
88	84031	44	13	0	4	0
561	46063	53	8	0	0	0
413	36409	42	8	0	2	0
58	29901	60	15	0	0	0
141	38927	31	11	0	2	0
...	...	...	...	...	...	...
728	35045	50	5	0	2	1
578	18486	60	9	0	0	0
414	49673	31	10	0	0	0
692	38874	38	3	0	3	1
697	26415	48	5	0	2	1

610 rows × 6 columns

```
y_train
```

	Holliday_Package_num
88	0
561	0
413	0
58	0
141	1
...	...
728	1
578	0
414	0
692	1
697	1

610 rows × 1 columns

- Here 6 columns and 610 rows are present in x_train dataset.
- In y_train dataset here 610 length series present.
- This both x_train and y_train are contain 70% of original dataset because we use 70:30 ratio for split dataset.

```
x_test
```

	Salary	age	educ	no_young_children	no_older_children	foreign_num
555	27598	60	8	0	0	0
813	35646	31	8	0	1	1
677	49756	32	8	0	1	1
552	42188	48	8	0	2	0
549	43940	59	10	0	0	0
...	...	...	...	...	...	...
526	62674	28	14	1	1	0
476	208561	35	16	1	2	0
719	38533	49	8	0	0	1
232	62509	51	8	0	0	0
638	33126	32	17	1	0	0

262 rows × 6 columns

```
y_test
```

	Holliday_Package_num
555	0
813	1
677	1
552	0
549	0
...	...
526	0
476	0
719	1
232	0
638	1

262 rows × 1 columns

- Here 6 columns and 262 rows are present in x_test dataset.
- In y_test dataset here 262 length series present.
- This both x_test and y_test are contain 30% of original dataset because we use 70:30 ratio for split dataset.

- **Apply Logistic Regression and LDA**

```
from sklearn.linear_model import LogisticRegression
lor = LogisticRegression()
```

```
lor.fit(x_train,y_train)
```

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
lda = LinearDiscriminantAnalysis()
```

```
lda.fit(x_train,y_train)
```

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 Compare Both the models and write inference which model is best/optimized.

- **Check the performance using Linear regression:**

```
lor.score(x_test,y_test)
```

```
0.5343511450381679
```

- **Check the performance using LDA:**

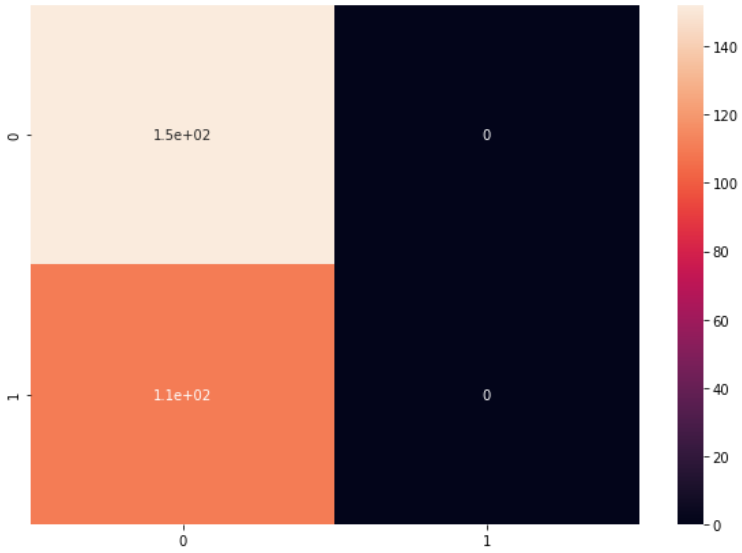
```
lda.score(x_test,y_test)
```

```
0.6755725190839694
```

  - From upper two figure we can say LDA is best algorithm for this data set and the score is very good compare to Logistic Regression.

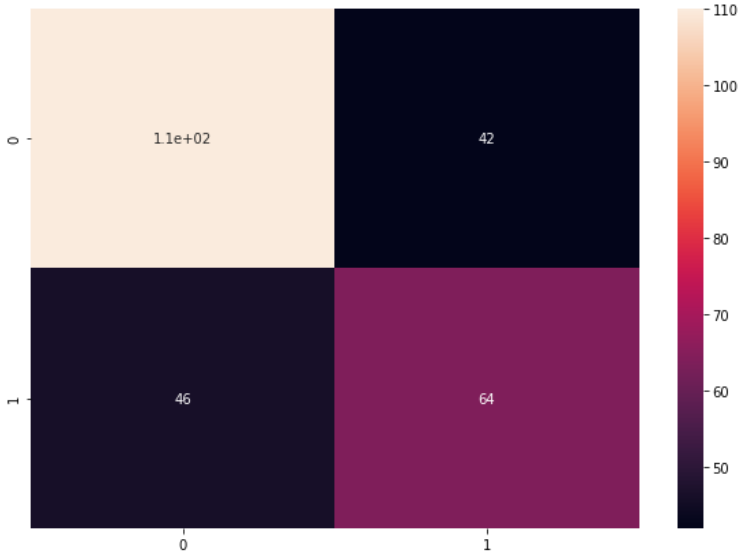
- **Confusion Matrix for Logistic Regression without dummies**

```
array([[152,   0],
       [110,   0]], dtype=int64)
```



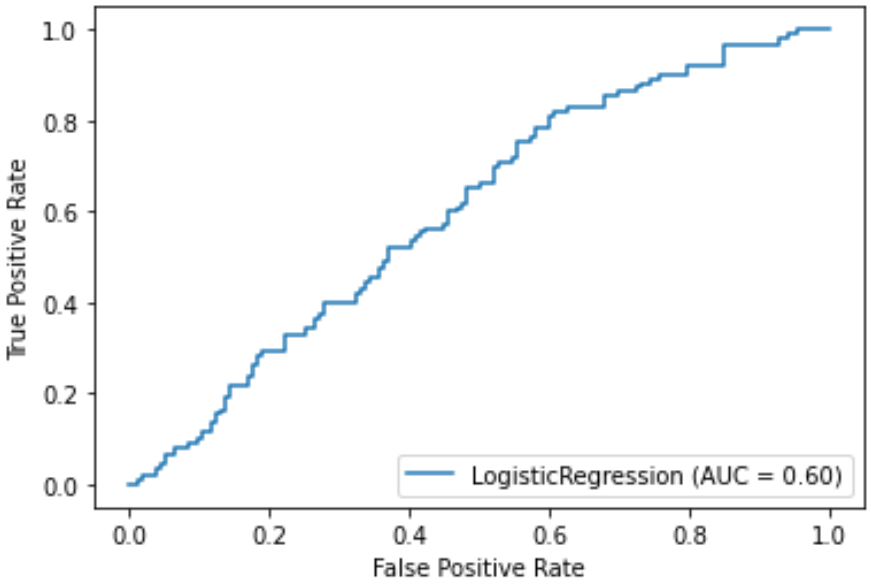
- **Confusion Matrix for LDA without dummies**

```
array([[110,  42],
       [ 46,  64]], dtype=int64)
```



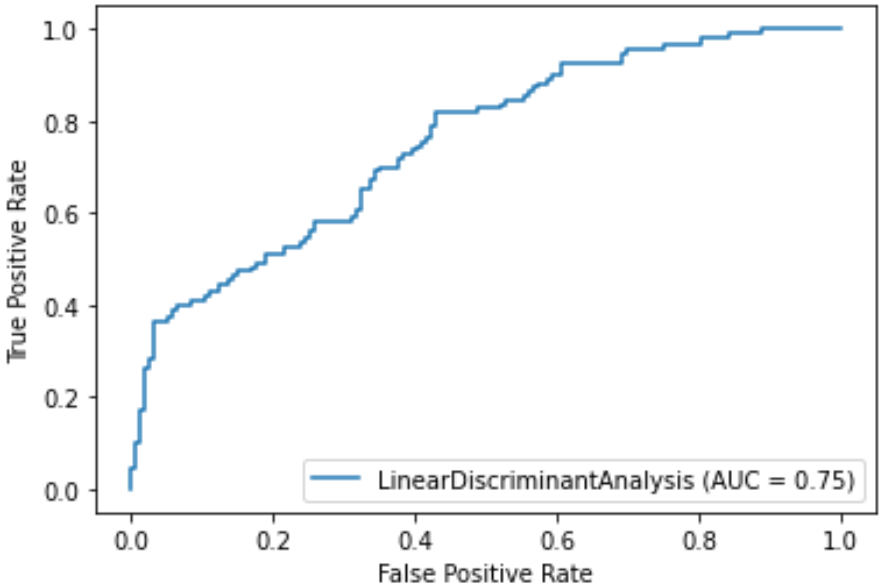
  - From above confusion metrix we can say the LDA confusion Metrix is good in compare Logistic Regression.

- **ROC Curve for Logistic Regression**



ROC_AUC Score for Linear Regression = 0.60

- **ROC Curve for LDA**



ROC_AUC Score for LDA = 0.75

- **Compare both model**

Name	Accuracy	ROC_AUC
Linear Regression	0.534	0.60
LDA	0.675	0.75

We can clearly saw from upper table the accuracy of LDA model is more than Linear Regression so LDA model is best for this particular problem.
Here ROC_AUC of LDA is also grater than Linear Regression model so we can say LDA is best algorithm for this data set.

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

- We can say around 60% people who lives in India they don't buy holiday package and only 40% people prefer to buy package but at same time people who lives in foreign they more likely buy holiday package and ratio of holiday package buying and not buying for foreign people is 70:30 means 70% people are interested to buy holiday package.
  So can say company should more promote package in foreign because there is high chance of conversion.

- We can say people which have young children they are not more interested to buy holiday package.
  So, company should advertise of package on people which have no younger children for more conversion.

- People who have formal education in range 1-5 years they are more interested to buy holiday package.

As the formal education increase from 7 to 17 years then there is very low chance to buy holiday package.
Which people are high educated means which have around 19-20 years formal education they don't prefer buy the holiday package.
So, company should promote holiday package according formal education which briefly shown in the repot.

- People which age in range 20-21 they are interested to buy holiday package and around 50% people in this age group are prefer to buy package.
  People which are in age group of 22-31 they are not interested to buy holiday package so there is very low chance to buying holiday package.
  People whose age in range 32-45 means there is very high chance to buying holiday package so company should focus on that age group people for selling package.
  People whose age increase from 46 to 62 means there is chance of buying holiday package is decrease and at the end of age range 60,61 and 62 there very very low chance to buying package so company should take action accordingly.
  So, as final conclusion company should promote holiday package according age group.