PREDICTIVE MODELING

HOLIDAY PACKGES

Microsoft Corporation

CONTENT

HOLIDAY PACKAGES

PROJECT 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.

Data Dictionary:

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

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.

Loading all the Data from the dataset .

HEAD :

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
0	1	no	48412	30	8	1	1	no
1	2	yes	37207	45	8	0	1	no
2	3	no	58022	46	9	0	0	no
3	4	no	66503	31	11	2	0	no
4	5	no	66734	44	12	0	2	no

TAIL :

	Unnamed: 0	Holliday_Package	Salary	age	educ	no_young_children	no_older_children	foreign
867	868	no	40030	24	4	2	1	yes
868	869	yes	32137	48	8	0	0	yes
869	870	no	25178	24	6	2	0	yes
870	871	yes	55958	41	10	0	1	yes
871	872	no	74659	51	10	0	0	yes

DATA DESCRIPTION :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 872 entries, 0 to 871
Data columns (total 8 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Unnamed: 0          872 non-null    int64
 1   Holliday_Package    872 non-null    object
 2   Salary              872 non-null    int64
 3   age                 872 non-null    int64
 4   educ                872 non-null    int64
 5   no_young_children   872 non-null    int64
 6   no_older_children   872 non-null    int64
 7   foreign             872 non-null    object
dtypes: int64(6), object(2)
memory usage: 54.6+ KB
```

- There are 872 Rows and 8 Columns present in the dataset.
- Among them , 6 Numerical variables and 2 Categoical variables.
- Unnamed:0,Salary,educ,no_young_children,no_olderchildren are Numerical variables.
- Holliday_package,foreign are Categorical variables.
- Target variables : Holliday_package

Checking NULL VALUES :

```
Unnamed: 0           0
Holliday_Package      0
Salary                0
age                   0
educ                  0
no_young_children     0
no_older_children     0
foreign               0
dtype: int64
```

No null values present .

Checking DUPLICATES :

No duplicates present.

```
dups=holiday.duplicated().sum()
dups
```

```
0
```

DATA SUMMARY :

	Unnamed: 0	Salary	age	educ	no_young_children	no_older_children
count	872.000000	872.000000	872.000000	872.000000	872.000000	872.000000
mean	436.500000	47729.172018	39.955275	9.307339	0.311927	0.982798
std	251.869014	23418.668531	10.551675	3.036259	0.612870	1.086786
min	1.000000	1322.000000	20.000000	1.000000	0.000000	0.000000
25%	218.750000	35324.000000	32.000000	8.000000	0.000000	0.000000
50%	436.500000	41903.500000	39.000000	9.000000	0.000000	1.000000
75%	654.250000	53469.500000	48.000000	12.000000	0.000000	2.000000
max	872.000000	236961.000000	62.000000	21.000000	3.000000	6.000000

UNIQUE VALUES :

```
holiday['Holliday_Package'].value_counts()
```

```
no      471
yes     401
Name: Holliday_Package, dtype: int64
```

```
holiday['foreign'].value_counts()
```

```
no      656
yes     216
Name: foreign, dtype: int64
```

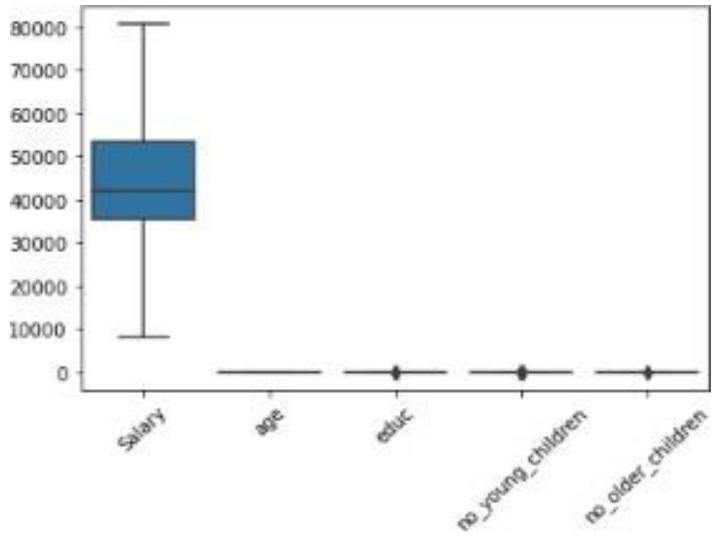
```
holiday['no_young_children'].value_counts()
```

```
0    665
1    147
2     55
3      5
Name: no_young_children, dtype: int64
```

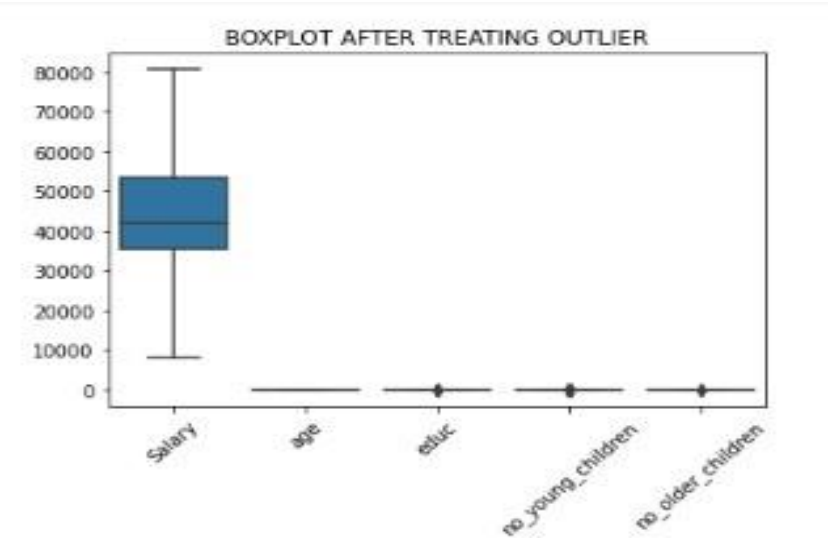
```
holiday['no_older_children'].value_counts()
```

```
0    393
2    208
1    198
3     55
4     14
5      2
6      2
Name: no_older_children, dtype: int64
```

CHECKING OUTLIERS :

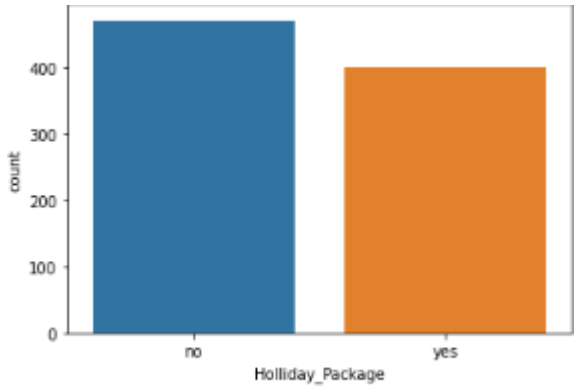


Presence of Outliers in all the columns except Age.Now its enough to treat outliers in salary column for the given problem.

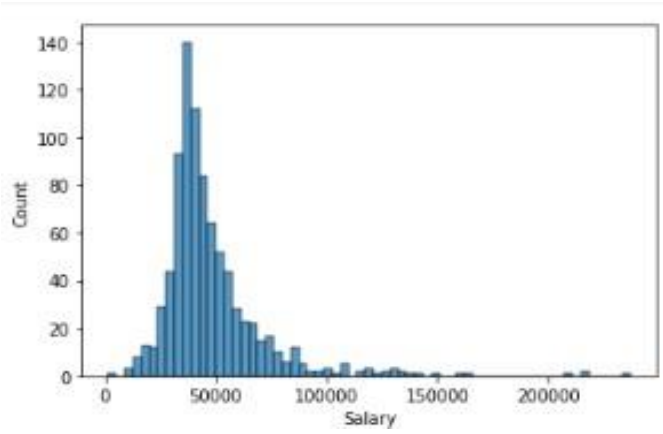


UNIVARIATE ANALYSIS :

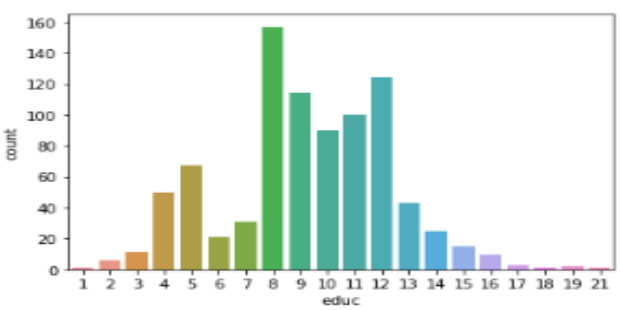
Holliday_package :



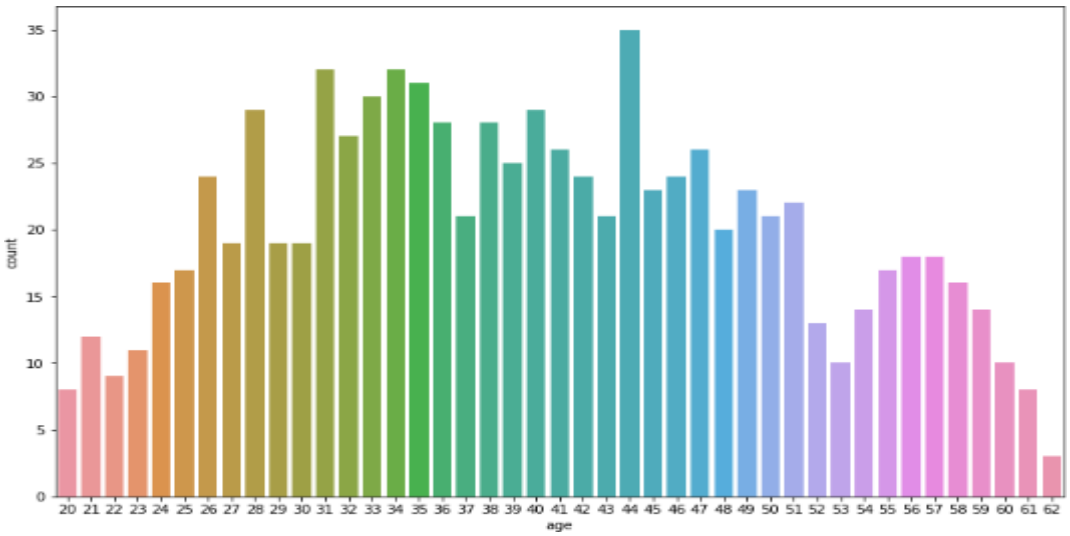
Salary :



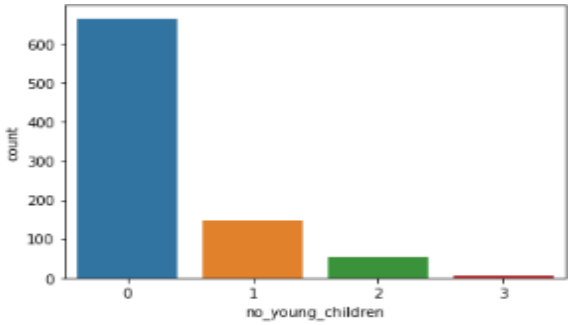
Educ (no. of years of Formal Education) :



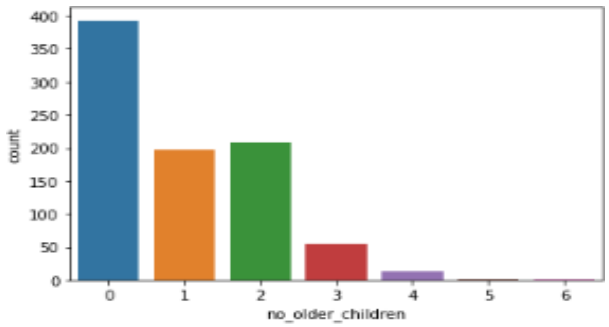
Age :



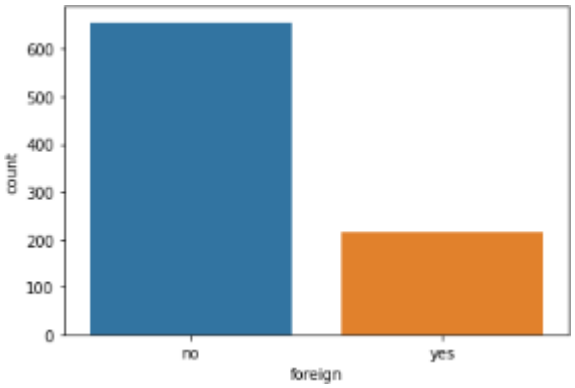
No_young _children :



No_older_children :

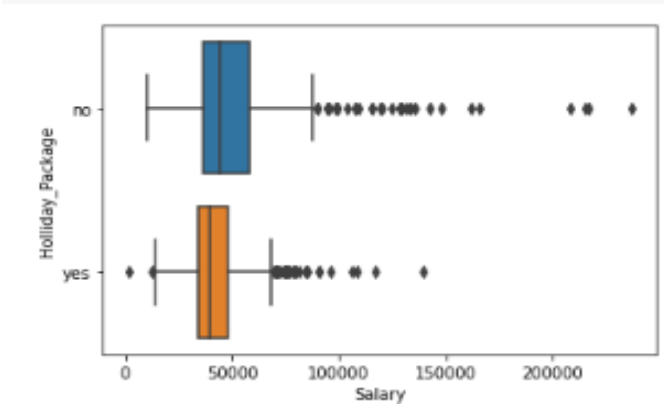


FOREIGN :

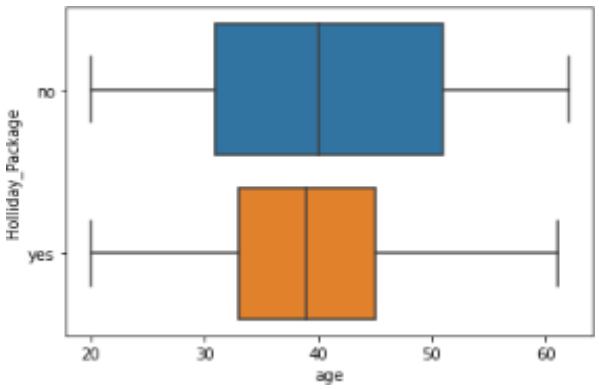


BIVARIATE ANALYSIS :

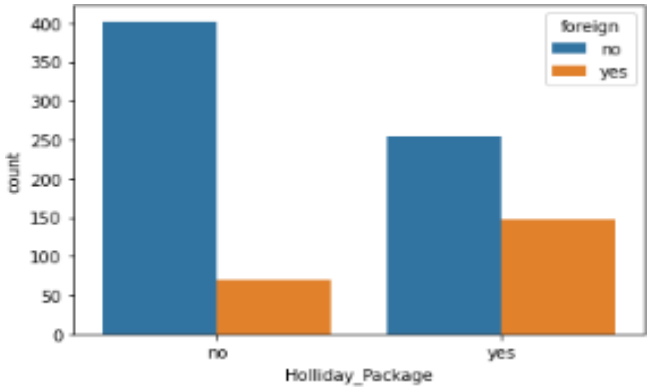
SALARY VS HOLIDAY_PACKAGE :



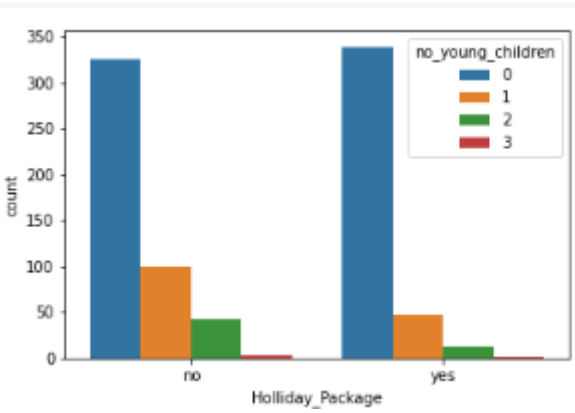
HOLLIDAY_PACKAGE VS AGE :



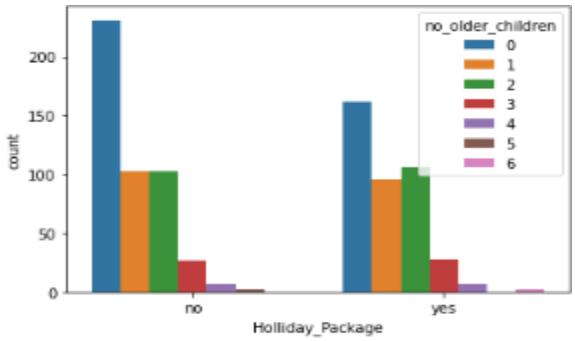
HOLIDAY_PACKAGE VS FOREIGN :



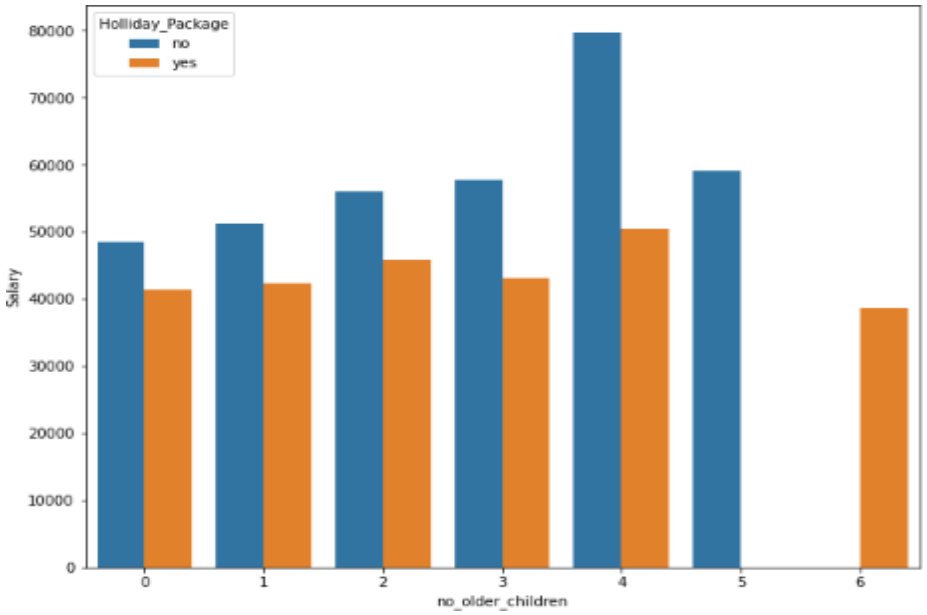
HOLIDAY_PACKAGE VS NO_YOUNG_CHILDREN :



HOLIDAY_PACKAGE VS NO_OLDER_CHILDREN :

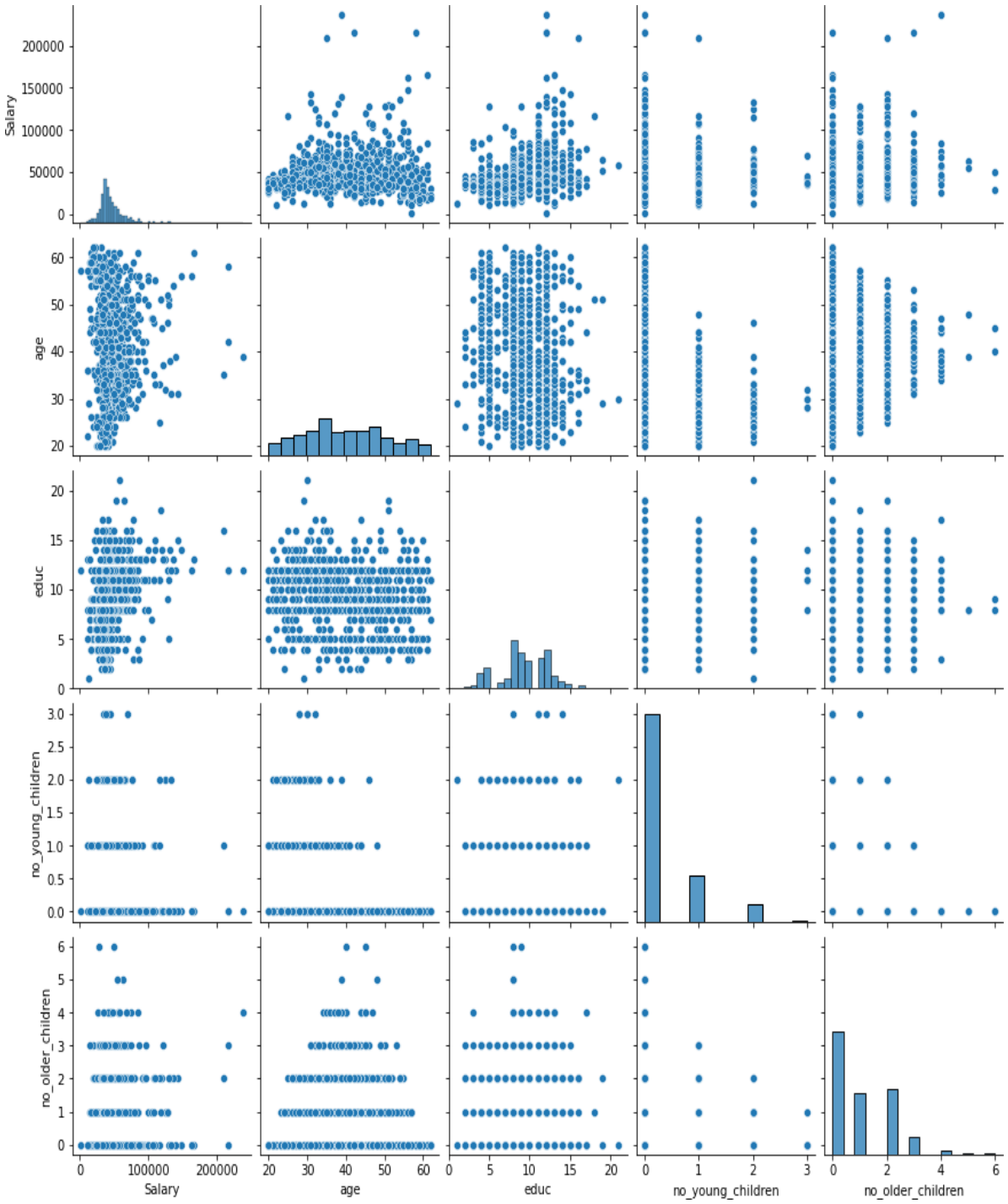


HOLLIDAY_PACKAGE with respect to SALARY & NO_OLDER_CHILDERN :



CORRELATION MAP :

DATA DISTRIBUTION :



OBSERVATIONS :

- There is high chance for them to take the package if the employee salary is between 35k to 50 k.
- There Is a higher chance of taking up the package if the employee age is between age of 35 to 45 years. After 50 years there is less chance for accepting the package.
- If the employee has no young children then there is huge chance to accept.
- If employee is a foreigner there is a huge chance to accept the package

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 : GET DUMMIES

	Salary	age	educ	no_young_children	no_older_children	Holliday_Package_yes	foreign_yes
0	48412.0	30	8	1	1	0	0
1	37207.0	45	8	0	1	1	0
2	58022.0	46	9	0	0	0	0
3	66503.0	31	11	2	0	0	0
4	66734.0	44	12	0	2	0	0
...	...	...	...	...	...	...	...
867	40030.0	24	4	2	1	0	1
868	32137.0	48	8	0	0	1	1
869	25178.0	24	6	2	0	0	1
870	55958.0	41	10	0	1	1	1
871	74659.0	51	10	0	0	0	1

872 rows × 7 columns

SPLIT THE DATA INTO TRAINING & TEST DATA :

```
X= hp.drop('Holliday_Package_yes', axis=1)

# Copy target into the y dataframe.
Y = hp[['Holliday_Package_yes']]
```

```
X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.3,random_state=123)
type(X_train)
```

```
pandas.core.frame.DataFrame
```

LOGISTIC REGRESSION :

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

LINEAR DISCRIMINATARY ANALYSIS :

```
clf=LinearDiscriminantAnalysis()
model2=clf.fit(X,Y.values.ravel())
model2
```

```
LinearDiscriminantAnalysis()
```

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.

❖ Performance Metrics for Logistic Regression :

✓ Accuracy score :

- Train data :

```
# Accuracy - Training Data
print('Accuracy score for Logistic Regression Train variables', model1.score(X_train, Y_train))
```

```
Accuracy score for Logistic Regression Train variables 0.6655737704918033
```

- Test data :

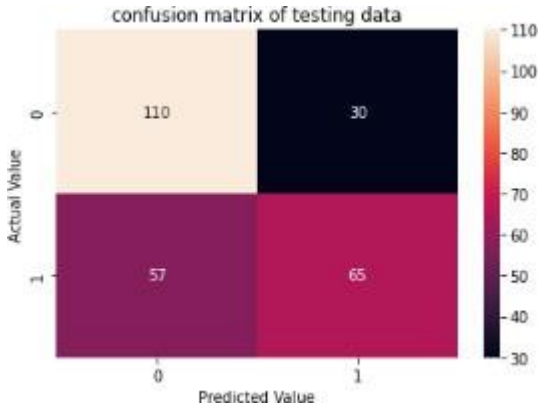
```
# Accuracy - Test Data
print('Accuracy score for Logistic Regression Test variables', model1.score(X_test, Y_test))
```

```
Accuracy score for Logistic Regression Test variables 0.6679389312977099
```

✓ Confusion Matrix :

- Training data :

- Testing data :



✓ Classification Report :

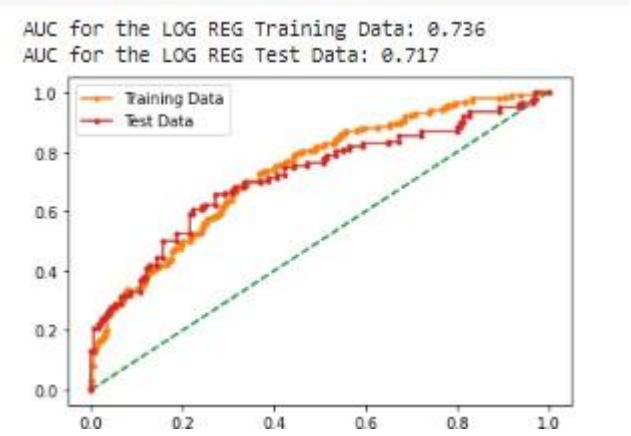
- Training data :

	precision	recall	f1-score	support
0	0.67	0.74	0.71	331
1	0.65	0.58	0.61	279
accuracy			0.67	610
macro avg	0.66	0.66	0.66	610
weighted avg	0.66	0.67	0.66	610

- Testing data

	precision	recall	f1-score	support
0	0.66	0.79	0.72	140
1	0.68	0.53	0.60	122
accuracy			0.67	262
macro avg	0.67	0.66	0.66	262
weighted avg	0.67	0.67	0.66	262

ROC & AUC :



❖ **Performance metrics for LDA :**

✓ **Confusion matrix :**

✓ Classification Report :

Classification Report of the LDA training data:

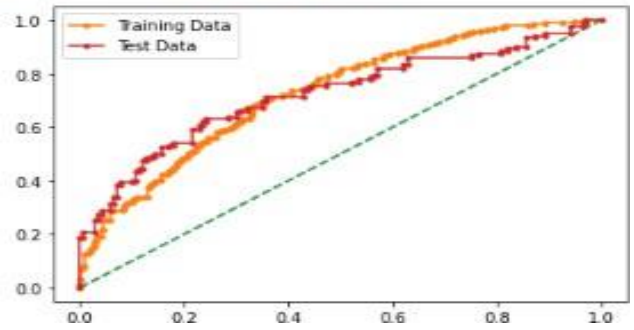
	precision	recall	f1-score	support
0	0.67	0.76	0.71	331
1	0.66	0.55	0.60	279
accuracy			0.66	610
macro avg	0.66	0.65	0.65	610
weighted avg	0.66	0.66	0.66	610

Classification Report of the LDA test data:

	precision	recall	f1-score	support
0	0.67	0.81	0.74	140
1	0.72	0.54	0.62	122
accuracy			0.69	262
macro avg	0.69	0.68	0.68	262
weighted avg	0.69	0.69	0.68	262

✓ ROC & AUC CURVE :

AUC for the LDA Training Data: 0.728
AUC for the LDA Test Data: 0.722



✓ LDA EQUATION :

LDA = 2.45 -0.0000208(Salary) - 0.05 (age) + 0.04 (educ) - 1.24 (no_young_children) - 0.02 (no_older_children) + 1.35 (foreign_yes)

COMPARISON BETWEEN LOGISTIC REG & LDA :

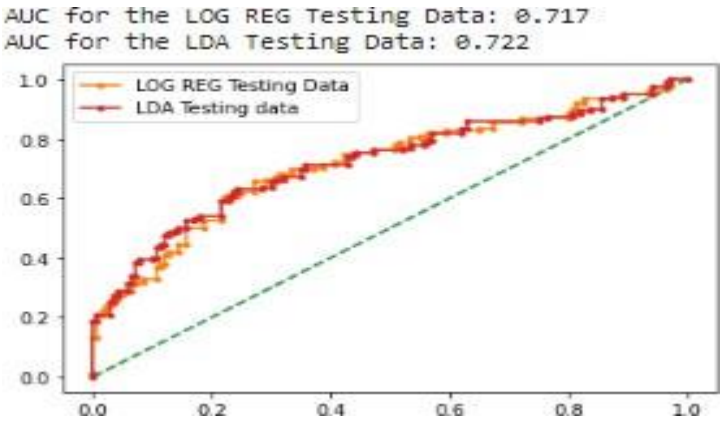
- **TABULATION :**

	LOGISTIC REG TRAINING DATA	LOGISTIC REG TESTING DATA	LDA TRAINING DATA	LDA TESTING DATA
ACCURACY	0.665	0.667	0.66	0.69
RECALL	0.54	0.53	0.55	0.54
PRECISION	0.65	0.68	0.66	0.72
F1 SCORE	0.61	0.60	0.60	0.62
AUC	0.74	0.72	0.73	0.72

❖ **ROC & AUC CURVE** :

- Training data :

- Testing data :



INFERENCES :

Based on comparison of Performance Metrics , LDA looks better because it has better RECALL rate and ACCURACY when compared with Logistic Regression.So LDA is the best model.

INFERENCE: BASIS ON THESE PREDICTIONS, WHAT ARE THE INSIGHTS AND RECOMMENDATIONS.

Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.

INSIGHTS :

- With 70 percent (approx.) accuracy , LDA model can predict whether an employee will opt for the package or not.

- Important factors which determine whether an employee opt for the package are Salary , Age , No_young_children, Foreign.

- Employee who

    - earn between 35k to 50k and
    - Age between 35 to 45 and
    - having no children have higher chance to accept the Holiday Package.
    - Foreigner have huge chance to accept the package.

- With 72 percent precision that the employees accept the Holiday Package.

- Accuracy, AUC, Precision and Recall for test data is almost inline with training data. This proves no overfitting or underfitting has happened, and overall the LDA model is a good model for classification.

RECOMMENDATIONS :

- The analysis shows that a greater number of foreigners opt in for packages than the non-foreigners. This along with the previous analysis which shows that most of the people are from salary group of 35 to 45k (so it is not expensive package) suggest that packages provided are either of local sightseeing place or of less interest to the non-foreigners. So, suggest the company to add some more activities or places in their packages.

- The greatest number of people who are opting in for the package has a salary of range between 35 to 45 k. It suggests that the package is of average price with medium level facilities So, if they add some additional luxury packages with facilities like booking in star hotels, luxury cars etc. it may help to increase the sales of packages to a higher income group.

- The analysis shows that data if the employee as no young children, there is more chance of them taking up the package. As count of children increases, the willingness to opt in for a holiday package decreases. So, I suggest that the company has to provide additional discounts or attract the children of employees who has young children to boost up the chance of opting for the package .

THANK YOU