Attrition Assignment

Problem Statement- A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners.

A sizeable department has to be maintained, for the purposes of recruiting new talent. More often than not, the new employees have to be trained for the job and/or given time to acclimatise themselves to the company.

Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Since you are one of the star analysts at the firm, this project has been given to you.

Goal of the case study You are required to model the probability of attrition. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Step 1- Launching

```
In [1]: import pandas as pd

In [2]: import numpy as np

In [3]: import matplotlib.pyplot as plt
```

```
In [5]: dataset=pd.read_csv("general_data.csv")

In [6]: dataset.head()
Out[6]:
   Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0   51        No  ...                       0                    0
1   31       Yes  ...                       1                    4
2   32        No  ...                       0                    3
3   38        No  ...                       7                    5
4   32        No  ...                       0                    4

[5 rows x 24 columns]
```

```
      Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0      51        No  ...                       0                    0
1      31       Yes  ...                       1                    4
2      32        No  ...                       0                    3
3      38        No  ...                       7                    5
4      32        No  ...                       0                    4
...   ...       ...  ...                     ...                  ...
4405   42        No  ...                       0                    2
4406   29        No  ...                       0                    2
4407   25        No  ...                       1                    2
4408   42        No  ...                       7                    8
4409   40        No  ...                       3                    9

[4410 rows x 24 columns]
```

```
In [7]: dataset.columns
Out[7]:
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
       'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',
       'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

Step 2- Data Treatment

```
In [8]: dataset.isnull()
Out[8]:
        Age  Attrition  ...  YearsSinceLastPromotion  YearsWithCurrManager
0     False      False  ...                    False                 False
1     False      False  ...                    False                 False
2     False      False  ...                    False                 False
3     False      False  ...                    False                 False
4     False      False  ...                    False                 False
...     ...        ...  ...                      ...                   ...
4405  False      False  ...                    False                 False
4406  False      False  ...                    False                 False
4407  False      False  ...                    False                 False
4408  False      False  ...                    False                 False
4409  False      False  ...                    False                 False

[4410 rows x 24 columns]
```

Checked and Deleted Duplicate Values (here no duplicates found)

```
In [9]: dataset.duplicated()
Out[9]:
0       False
1       False
2       False
3       False
4       False
        ...
4405    False
4406    False
4407    False
4408    False
4409    False
Length: 4410, dtype: bool
```

```
In [11]: dataset=dataset.drop_duplicates()

In [12]: dataset
Out[12]:
      Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0      51        No  ...                       0                    0
1      31       Yes  ...                       1                    4
2      32        No  ...                       0                    3
3      38        No  ...                       7                    5
4      32        No  ...                       0                    4
...   ...       ...  ...                     ...                  ...
4405   42        No  ...                       0                    2
4406   29        No  ...                       0                    2
4407   25        No  ...                       1                    2
4408   42        No  ...                       7                    8
4409   40        No  ...                       3                    9

[4410 rows x 24 columns]
```

Deleted Null Values

```
In [13]: dataset=dataset.dropna()

In [14]: dataset
Out[14]:
      Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0      51        No  ...                       0                    0
1      31       Yes  ...                       1                    4
2      32        No  ...                       0                    3
3      38        No  ...                       7                    5
4      32        No  ...                       0                    4
...   ...       ...  ...                     ...                  ...
4404   29        No  ...                       1                    5
4405   42        No  ...                       0                    2
4406   29        No  ...                       0                    2
4407   25        No  ...                       1                    2
4408   42        No  ...                       7                    8

[4382 rows x 24 columns]
```

Step 3- Univariate Analysis

```
In [15]: described_data=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion',
'YearsWithCurrManager']].describe()
```

Described Data-

```
In [18]: described_data
Out[18]:
              Age  DistanceFromHome    Education  MonthlyIncome  \
count  4382.000000       4382.000000  4382.000000    4382.000000
mean     36.933364          9.198996     2.912369   65061.702419
std       9.137272          8.105396     1.024728   47142.310175
min      18.000000          1.000000     1.000000   10090.000000
25%      30.000000          2.000000     2.000000   29110.000000
50%      36.000000          7.000000     3.000000   49190.000000
75%      43.000000         14.000000     4.000000   83790.000000
max      60.000000         29.000000     5.000000  199990.000000

       NumCompaniesWorked  PercentSalaryHike  TotalWorkingYears  \
count         4382.000000        4382.000000        4382.000000
mean             2.693291          15.210634          11.290278
std              2.497832           3.663007           7.785717
min              0.000000          11.000000           0.000000
25%              1.000000          12.000000           6.000000
50%              2.000000          14.000000          10.000000
75%              4.000000          18.000000          15.000000
max              9.000000          25.000000          40.000000

       TrainingTimesLastYear  YearsAtCompany  YearsSinceLastPromotion  \
count            4382.000000     4382.000000              4382.000000
mean                2.798266        7.010497                 2.191693
std                 1.289402        6.129351                 3.224994
min                 0.000000        0.000000                 0.000000
25%                 2.000000        3.000000                 0.000000
50%                 3.000000        5.000000                 1.000000
75%                 3.000000        9.000000                 3.000000
max                 6.000000       40.000000                15.000000

       YearsWithCurrManager
count           4382.000000
mean               4.126198
std                3.569674
min                0.000000
25%                2.000000
50%                3.000000
75%                7.000000
max               17.000000
```

Mean-

```
Age                          36.933364
DistanceFromHome              9.198996
Education                     2.912369
MonthlyIncome             65061.702419
NumCompaniesWorked            2.693291
PercentSalaryHike            15.210634
TotalWorkingYears            11.290278
TrainingTimesLastYear         2.798266
YearsAtCompany                7.010497
YearsSinceLastPromotion       2.191693
YearsWithCurrManager          4.126198
dtype: float64
```

Median-

```
Age                          36.0
DistanceFromHome              7.0
Education                     3.0
MonthlyIncome             49190.0
NumCompaniesWorked            2.0
PercentSalaryHike            14.0
TotalWorkingYears            10.0
TrainingTimesLastYear         3.0
YearsAtCompany                5.0
YearsSinceLastPromotion       1.0
YearsWithCurrManager          3.0
dtype: float64
```

Mode-

```
   Age  DistanceFromHome  Education  MonthlyIncome  NumCompaniesWorked  \
0   35                 2          3          23420                 1.0

   PercentSalaryHike  TotalWorkingYears  TrainingTimesLastYear  \
0                 11               10.0                      2

   YearsAtCompany  YearsSinceLastPromotion  YearsWithCurrManager
0               5                        0                     2
```

Variance-

```
Age                        8.348974e+01
DistanceFromHome           6.569744e+01
Education                  1.050068e+00
MonthlyIncome              2.222397e+09
NumCompaniesWorked         6.239165e+00
PercentSalaryHike          1.341762e+01
TotalWorkingYears          6.061739e+01
TrainingTimesLastYear      1.662558e+00
YearsAtCompany             3.756894e+01
YearsSinceLastPromotion    1.040059e+01
YearsWithCurrManager       1.274257e+01
dtype: float64
```

Std Deviation-

```
Age                          9.137272
DistanceFromHome             8.105396
Education                    1.024728
MonthlyIncome            47142.310175
NumCompaniesWorked           2.497832
PercentSalaryHike            3.663007
TotalWorkingYears            7.785717
TrainingTimesLastYear        1.289402
YearsAtCompany               6.129351
YearsSinceLastPromotion      3.224994
YearsWithCurrManager         3.569674
dtype: float64
```

Skewness-

```
Age                        0.413048
DistanceFromHome           0.955517
Education                 -0.288977
MonthlyIncome              1.367457
NumCompaniesWorked         1.029174
PercentSalaryHike          0.819510
TotalWorkingYears          1.115419
TrainingTimesLastYear      0.551818
YearsAtCompany             1.764619
YearsSinceLastPromotion    1.980992
YearsWithCurrManager       0.834277
dtype: float64
```

Kurtosis-

```
Age                        -0.409517
DistanceFromHome           -0.230691
Education                  -0.565008
MonthlyIncome               0.990836
NumCompaniesWorked          0.014307
PercentSalaryHike          -0.306951
TotalWorkingYears           0.909316
TrainingTimesLastYear       0.494215
YearsAtCompany              3.930726
YearsSinceLastPromotion     3.592162
YearsWithCurrManager        0.170703
dtype: float64
```

IQR-

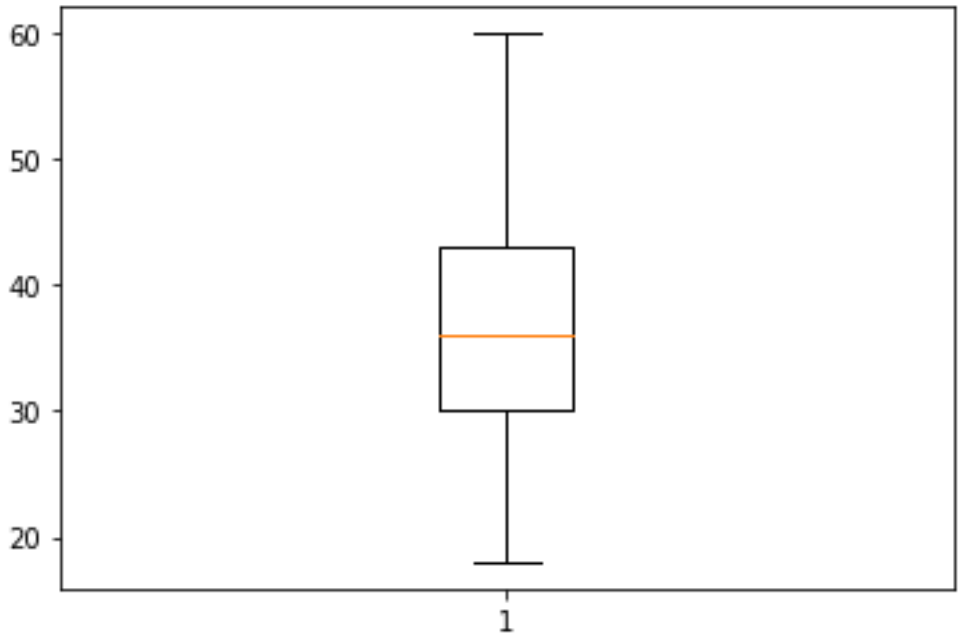
```
Age                        13.0
DistanceFromHome           12.0
Education                   2.0
MonthlyIncome           54680.0
NumCompaniesWorked          3.0
PercentSalaryHike           6.0
TotalWorkingYears           9.0
TrainingTimesLastYear       1.0
YearsAtCompany              6.0
YearsSinceLastPromotion     3.0
YearsWithCurrManager        5.0
dtype: float64
```

Step 4- Inference from the Analysis

- All the above variables are positively skewed (mean > median) except Education which is negatively skewed.
- Age, DistanceFromHome, Education and PercentSalaryHike are platykurtic in nature while all the other variables are leptokurtic.
- The MonthlyIncome's IQR is at 54K suggesting companywide attrition across all income bands.
- Age forms a near normal distribution with 13 years of IQR.

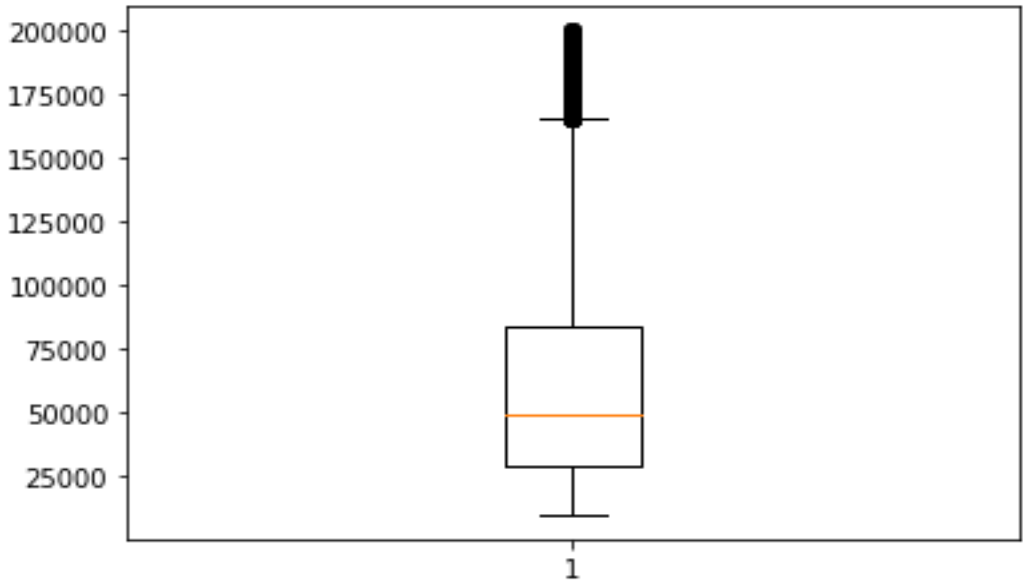
Step 5- Outliers

Age-



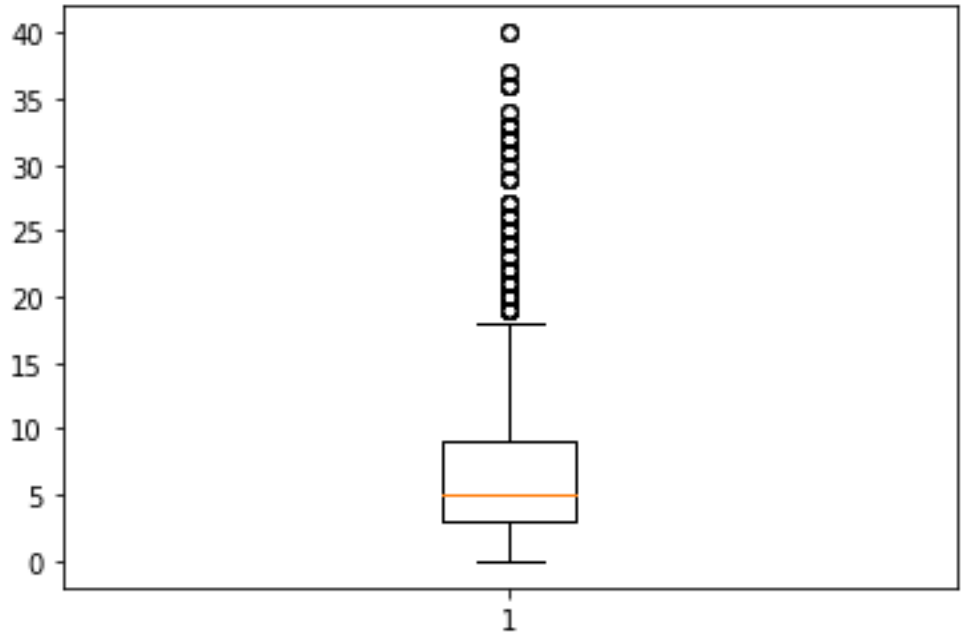
Age is normally distributed without any outliers.

Monthly Income-



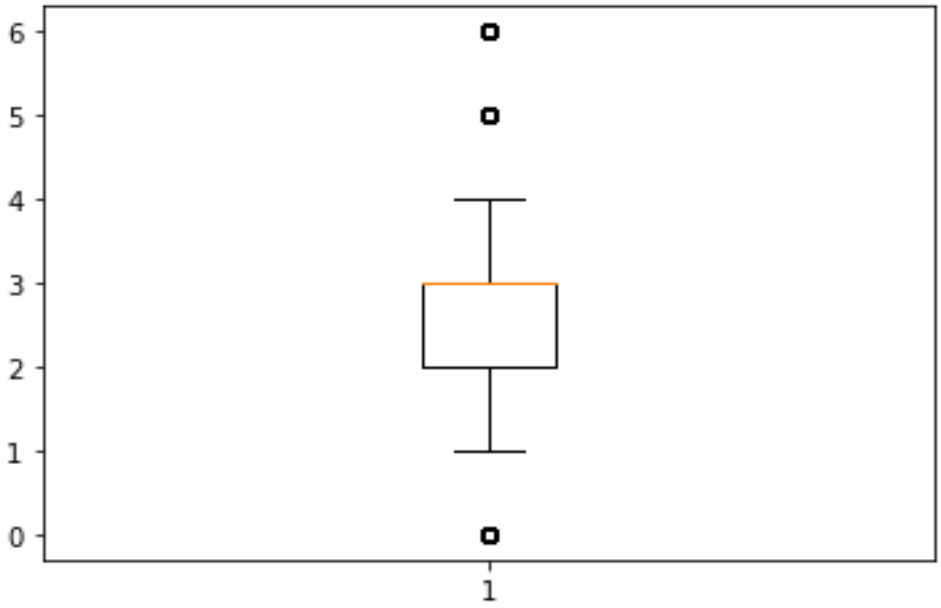
Monthly Income is positively skewed with several outliers.

Years at Company-



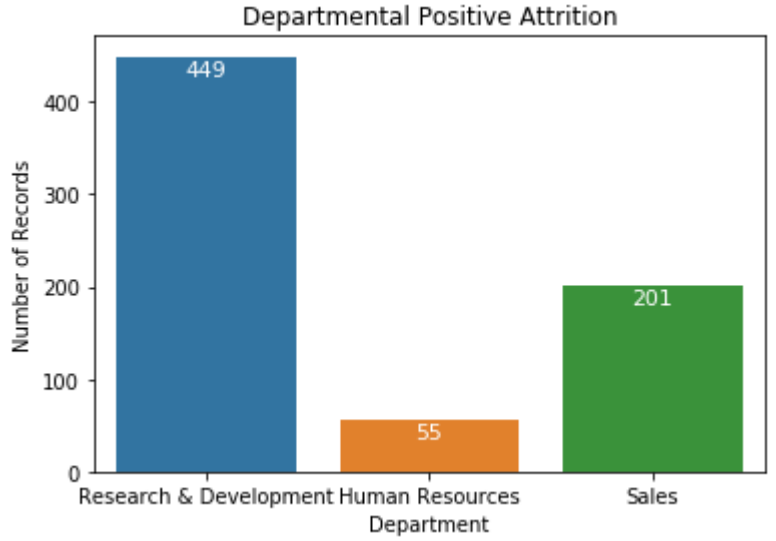
Years at Company is positively skewed with several outliers.

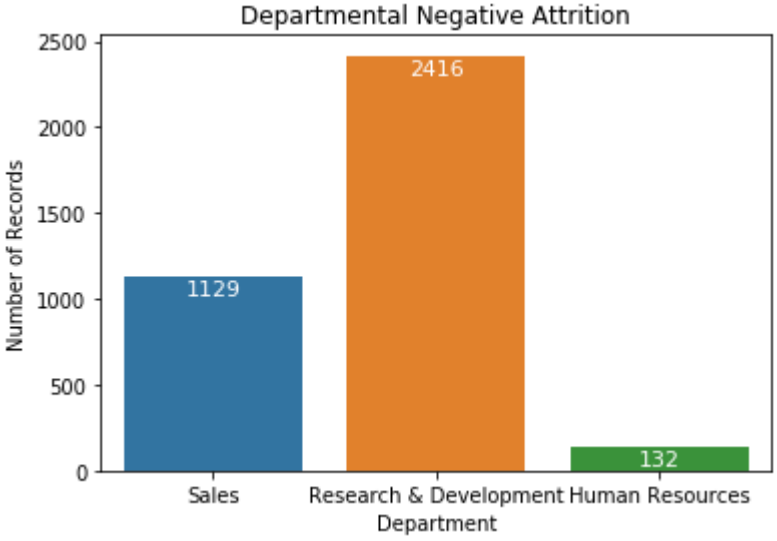
Training Times Last Year-



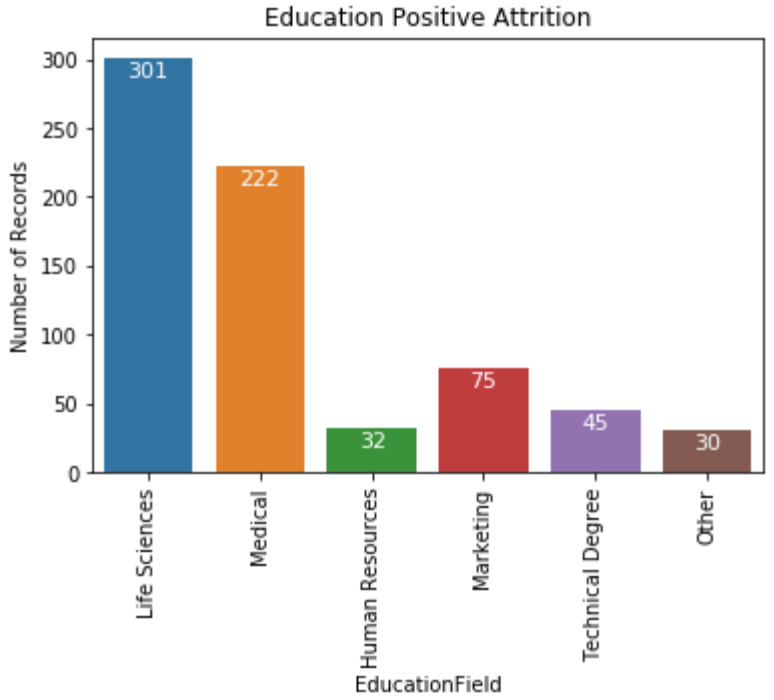
Training Times Last Year is negatively skewed with some outliers.

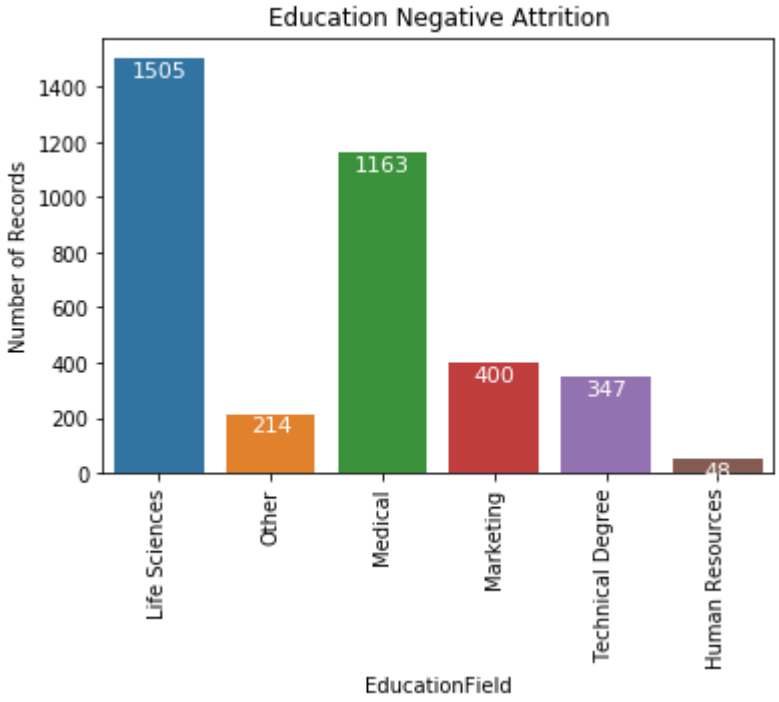
Step 6- Visualisation





Visualisation of each department with positive and negative attrition.





Visualisation of each department with positive and negative attrition.

Step 7 - Statistical Tests

Mann-Whitney Test-

Imported mannwhitneyu and defined a function for hypothesis testing.

```
In [34]: from scipy.stats import mannwhitneyu

In [35]: def manwhitney(stats, p, x):
    ...:     print('The Hypothesis statements are:')
    ...:     print('\nH0 = There is no significant difference between Attrition and ', x)
    ...:     print('\nH1 = There is significant difference between Attrition and ', x, '\n')
    ...:     print('The R value is: ',stats, '\nThe P Value is: ', p,'\n')
    ...:     if p < 0.05:
    ...:         print('The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is
accepted')
    ...:     else:
    ...:         print("The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is
accepted")
```

1. Attrition and Age-

```
In [36]: stats, p = mannwhitneyu(dataset_yes.Age, dataset_no.Age)
    ...: manwhitney(stats, p, 'Age')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  Age

H1 = There is significant difference between Attrition and  Age

The R value is:  949178.0
The P Value is:  7.98668614365882e-30

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

2. Attrition and Distance from home-

```
In [37]: stats, p = mannwhitneyu(dataset_yes.DistanceFromHome, dataset_no.DistanceFromHome)
    ...: manwhitney(stats, p, 'DistanceFromHome')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  DistanceFromHome

H1 = There is significant difference between Attrition and  DistanceFromHome

The R value is:  1295261.0
The P Value is:  0.488538986087403

The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is accepted
```

3. Attrition and Monthly Income-

```
In [38]: stats, p = mannwhitneyu(dataset_yes.MonthlyIncome, dataset_no.MonthlyIncome)
    ...: manwhitney(stats, p, 'MonthlyIncome')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  MonthlyIncome

H1 = There is significant difference between Attrition and  MonthlyIncome

The R value is:  1249573.5
The P Value is:  0.06508807631576838

The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is accepted
```

4. Attrition and Number of companies worked-

```
In [39]: stats, p = mannwhitneyu(dataset_yes.NumCompaniesWorked, dataset_no.NumCompaniesWorked)
    ...: manwhitney(stats, p, 'NumCompaniesWorked')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  NumCompaniesWorked

H1 = There is significant difference between Attrition and  NumCompaniesWorked

The R value is:  1238814.5
The P Value is:  0.02793197853866981

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

5. Attrition and Total working years-

```
In [40]: stats, p = mannwhitneyu(dataset_yes.TotalWorkingYears, dataset_no.TotalWorkingYears)
    ...: manwhitney(stats, p, 'TotalWorkingYears')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  TotalWorkingYears

H1 = There is significant difference between Attrition and  TotalWorkingYears

The R value is:  895173.5
The P Value is:  2.741211827689903e-39

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

6. Attrition and Training times last year-

```
In [41]: stats, p = mannwhitneyu(dataset_yes.TrainingTimesLastYear, dataset_no.TrainingTimesLastYear)
    ...: manwhitney(stats, p, 'TrainingTimesLastYear')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  TrainingTimesLastYear

H1 = There is significant difference between Attrition and  TrainingTimesLastYear

The R value is:  1225582.0
The P Value is:  0.008107344081224082

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

7. Attrition and Years at company-

```
In [42]: stats, p = mannwhitneyu(dataset_yes.YearsAtCompany, dataset_no.YearsAtCompany)
    ...: manwhitney(stats, p, 'YearsAtCompany')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  YearsAtCompany

H1 = There is significant difference between Attrition and  YearsAtCompany

The R value is:  912579.0
The P Value is:  3.3433144809752036e-36

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

8. Attrition and Years since last promotion-

```
In [43]: stats, p = mannwhitneyu(dataset_yes.YearsSinceLastPromotion, dataset_no.YearsSinceLastPromotion)
    ...: manwhitney(stats, p, 'YearsSinceLastPromotion')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  YearsSinceLastPromotion

H1 = There is significant difference between Attrition and  YearsSinceLastPromotion

The R value is:  1196606.0
The P Value is:  0.00037904698157957496

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

9. Attrition and Years with current manager-

```
In [44]: stats, p = mannwhitneyu(dataset_yes.YearsWithCurrManager, dataset_no.YearsWithCurrManager)
    ...: manwhitney(stats, p, 'YearsWithCurrManager')
The Hypothesis statements are:

H0 = There is no significant difference between Attrition and  YearsWithCurrManager

H1 = There is significant difference between Attrition and  YearsWithCurrManager

The R value is:  945958.5
The P Value is:  5.420302388722274e-31

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

Separate T Test-

Imported ttest_ind and defined a function for hypothesis testing.

```
In [45]: from scipy.stats import ttest_ind

In [46]: def ttest(stats, p, x):
    ...:     print('The Hypothesis statements are:')
    ...:     print('\nH0 = There is no significant difference between attrition and ', x)
    ...:     print('\nH1 = There is significant difference between attrition and ', x, '\n')
    ...:     print('The R value is: ',stats, '\nThe P Value is: ', p,'\n')
    ...:     if p < 0.05:
    ...:         print('The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is
accepted')
    ...:     else:
    ...:         print("The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is
accepted")
```

A. Attrition and Age-

```
In [47]: stats, p = ttest_ind(dataset_yes.Age, dataset_no.Age)
    ...: ttest(stats, p, 'Age')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  Age

H1 = There is significant difference between attrition and  Age

The R value is:  -10.617111568458819
The P Value is:  5.126598219406314e-26

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

B. Attrition and Distance from home-

```
In [48]: stats, p = ttest_ind(dataset_yes.DistanceFromHome, dataset_no.DistanceFromHome)
    ...: ttest(stats, p, 'DistanceFromHome')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  DistanceFromHome

H1 = There is significant difference between attrition and  DistanceFromHome

The R value is:  -0.6253536318706914
The P Value is:  0.5317715668047676

The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is accepted
```

C. Attrition and Monthly Income-

```
In [49]: stats, p = ttest_ind(dataset_yes.MonthlyIncome, dataset_no.MonthlyIncome)
    ...: ttest(stats, p, 'MonthlyIncome')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  MonthlyIncome

H1 = There is significant difference between attrition and  MonthlyIncome

The R value is:  -1.9969640177214658
The P Value is:  0.045890862744972095

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

D. Attrition and Number of companies worked-

```
In [50]: stats, p = ttest_ind(dataset_yes.NumCompaniesWorked, dataset_no.NumCompaniesWorked)
    ...: ttest(stats, p, 'NumCompaniesWorked')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  NumCompaniesWorked

H1 = There is significant difference between attrition and  NumCompaniesWorked

The R value is:  2.837197670884213
The P Value is:  0.004572057121646456

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

E. Attrition and Total working years-

```
In [51]: stats, p = ttest_ind(dataset_yes.TotalWorkingYears, dataset_no.TotalWorkingYears)
    ...: ttest(stats, p, 'TotalWorkingYears')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  TotalWorkingYears

H1 = There is significant difference between attrition and  TotalWorkingYears

The R value is:  -11.39422669317641
The P Value is:  1.1645434967153693e-29

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

F. Attrition and Training times last year-

```
In [52]: stats, p = ttest_ind(dataset_yes.TrainingTimesLastYear, dataset_no.TrainingTimesLastYear)
    ...: ttest(stats, p, 'TrainingTimesLastYear')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  TrainingTimesLastYear

H1 = There is significant difference between attrition and  TrainingTimesLastYear

The R value is:  -3.152870411721613
The P Value is:  0.00162766036355604

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

G. Attrition and Years at company-

```
In [53]: stats, p = ttest_ind(dataset_yes.YearsAtCompany, dataset_no.YearsAtCompany)
    ...: ttest(stats, p, 'YearsAtCompany')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  YearsAtCompany

H1 = There is significant difference between attrition and  YearsAtCompany

The R value is:  -8.881225486705604
The P Value is:  9.476118084889976e-19

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

H. Attrition and Years since last promotion-

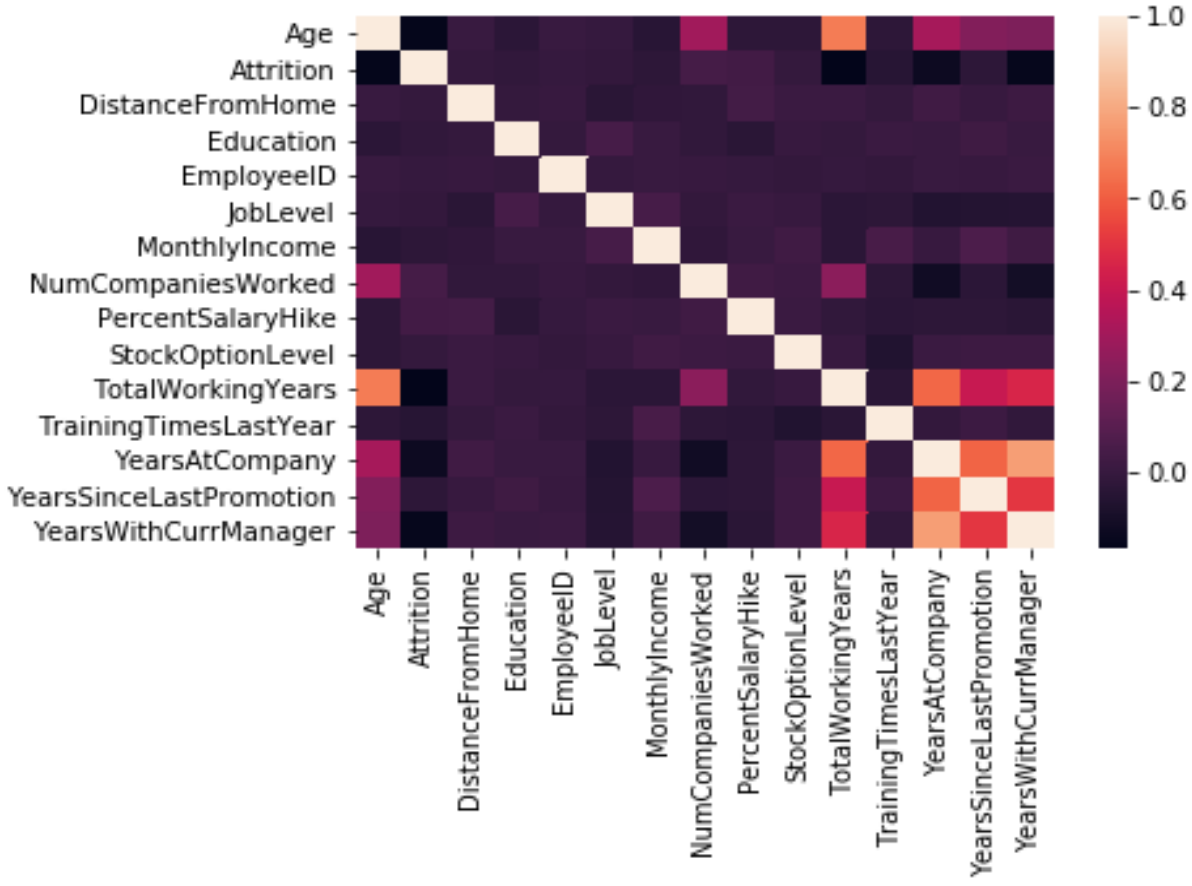
```
In [54]: stats, p = ttest_ind(dataset_yes.YearsSinceLastPromotion, dataset_no.YearsSinceLastPromotion)
    ...: ttest(stats, p, 'YearsSinceLastPromotion')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  YearsSinceLastPromotion

H1 = There is significant difference between attrition and  YearsSinceLastPromotion

The R value is:  -2.080660880277173
The P Value is:  0.03752293607413772

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

I. Attrition and Years with current manager-

```
In [55]: stats, p = ttest_ind(dataset_yes.YearsWithCurrManager, dataset_no.YearsWithCurrManager)
    ...: ttest(stats, p, 'YearsWithCurrManager')
The Hypothesis statements are:

H0 = There is no significant difference between attrition and  YearsWithCurrManager

H1 = There is significant difference between attrition and  YearsWithCurrManager

The R value is:  -10.362463400192302
The P Value is:  7.105369646808081e-25

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

Step 8 – Unsupervised Learning – Correlation Analysis

Standard Hours and Employee Count has no impact in correlation as they have same values in all records. So, we remove the columns and plot a graph to have an idea on correlation of variables.



Imported pearsonr and defined a function for calculating the correlation.

```
In [57]: from scipy.stats import pearsonr
```

```
In [62]: def corr_attrition(stats, p, x):
    ...:     print('The Hypothesis statements are:')
    ...:     print('\nH0 = There is no significant correlation between Attrition and ', x)
    ...:     print('\nH1 = There is significant correlation between Attrition and ', x, '\n')
    ...:     print('The R value is: ',stats, '\nThe P Value is: ', p,'\n')
    ...:     if p < 0.05:
    ...:         print('The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is
accepted')
    ...:     else:
    ...:         print("The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is
accepted")
```

- Attrition and Age Correlation-

```
In [67]: stats, p = pearsonr(dataset.Attrition, dataset.Age)
    ...: corr_attrition(stats, p, 'Age')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  Age

H1 = There is significant correlation between Attrition and  Age

The R value is:  -0.1583986795409671
The P Value is:  5.1265982193780794e-26

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Distance from home Correlation-

```
In [68]: stats, p = pearsonr(dataset.Attrition, dataset.DistanceFromHome)
    ...: corr_attrition(stats, p, 'DistanceFromHome')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  DistanceFromHome

H1 = There is significant correlation between Attrition and  DistanceFromHome

The R value is:  -0.009448638515156258
The P Value is:  0.5317715668019558

The Alternative Hypothesis H1 is rejected because P-Value >= 0.05, so the Null Hypothesis H0 is accepted
```

- Attrition and Monthly Income Correlation-

```
In [69]: stats, p = pearsonr(dataset.Attrition, dataset.MonthlyIncome)
    ...: corr_attrition(stats, p, 'MonthlyIncome')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  MonthlyIncome

H1 = There is significant correlation between Attrition and  MonthlyIncome

The R value is:  -0.030160293808460678
The P Value is:  0.045890862744719166

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Number of companies worked Correlation-

```
In [70]: stats, p = pearsonr(dataset.Attrition, dataset.NumCompaniesWorked)
    ...: corr_attrition(stats, p, 'NumCompaniesWorked')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  NumCompaniesWorked

H1 = There is significant correlation between Attrition and  NumCompaniesWorked

The R value is:  0.04283056724472085
The P Value is:  0.004572057121620842

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Total working years Correlation-

```
In [71]: stats, p = pearsonr(dataset.Attrition, dataset.TotalWorkingYears)
    ...: corr_attrition(stats, p, 'TotalWorkingYears')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  TotalWorkingYears

H1 = There is significant correlation between Attrition and  TotalWorkingYears

The R value is:  -0.16966991684723914
The P Value is:  1.1645434967091854e-29

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Training times last year Correlation-

```
In [72]: stats, p = pearsonr(dataset.Attrition, dataset.TrainingTimesLastYear)
    ...: corr_attrition(stats, p, 'TrainingTimesLastYear')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  TrainingTimesLastYear

H1 = There is significant correlation between Attrition and  TrainingTimesLastYear

The R value is:  -0.047585736930817205
The P Value is:  0.0016276603635474061

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Years at company Correlation-

```
In [73]: stats, p = pearsonr(dataset.Attrition, dataset.YearsAtCompany)
    ...: corr_attrition(stats, p, 'YearsAtCompany')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  YearsAtCompany

H1 = There is significant correlation between Attrition and  YearsAtCompany

The R value is:  -0.13300261842521532
The P Value is:  9.476118084840815e-19

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Years since last promotion Correlation-

```
In [74]: stats, p = pearsonr(dataset.Attrition, dataset.YearsSinceLastPromotion)
    ...: corr_attrition(stats, p, 'YearsSinceLastPromotion')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  YearsSinceLastPromotion

H1 = There is significant correlation between Attrition and  YearsSinceLastPromotion

The R value is:  -0.03142315056330984
The P Value is:  0.03752293607394267

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

- Attrition and Years with current manager Correlation-

```
In [75]: stats, p = pearsonr(dataset.Attrition, dataset.YearsWithCurrManager)
    ...: corr_attrition(stats, p, 'YearsWithCurrManager')
The Hypothesis statements are:

H0 = There is no significant correlation between Attrition and  YearsWithCurrManager

H1 = There is significant correlation between Attrition and  YearsWithCurrManager

The R value is:  -0.15469153690287285
The P Value is:  7.105369646772844e-25

The Null Hypothesis H0 is rejected because P-Value < 0.05, so the Alternative Hypothesis H1 is accepted
```

Inference from above Analysis-

i. **Attrition and Age**- As r = -0.1583, there is low negative correlation between Attrition and Age. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and Age.

ii. **Attrition and Distance from home**- As r = -0.0094, there is low negative correlation between Attrition and DistanceFromHome. As the P value is > 0.05, the null hypothesis is accepted, so there is no significant correlation between Attrition and DistanceFromHome.

iii. **Attrition and Monthly Income-** As r = -0.0301, there is low negative correlation between Attrition and Monthly Income. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and Monthly Income.

iv. **Attrition and Num of companies worked**: As r = 0.0428, there is low positive correlation between Attrition and Num of companies worked. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and Num of companies worked.

v. **Attrition and Total working years-** As r = -0.1696, there is low negative correlation between Attrition and TotalWorkingYears. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and TotalWorkingYears.

vi. **Attrition and Training times last year**- As r = -0.0475, there is low negative correlation between Attrition and TrainingTimesLastYear. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and TrainingTimesLastYear.

vii. **Attrition and Years at company**- As r = -0.1330, there is low negative correlation between Attrition and YearsAtCompany. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and YearsAtCompany.

viii. **Attrition and Years since last promotion**- As r = -0.0314, there is low negative correlation between Attrition and YearsSinceLastPromotion. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and YearsSinceLastPromotion.

ix. **Attrition and Years with current manager**- As r = -0.1546, there is low negative correlation between Attrition and YearsWithCurrManager. As the P value is < 0.05, the null hypothesis is rejected, so there is significant correlation between Attrition and YearsWithCurrManager.

Conclusion-

From the above Statistical Tests and Correlation Analysis, we can say that there are few factors which are related to Attrition and rest have no significance for Attrition. From the above Inferences, we can conclude that the Company has to make following changes to reduce the number of Attrition in the Company.

i) Hire middle aged employees having age of approximately 36 years old or above.
ii) Reduce the number of business trips of employees.
iii) Other departments should be more familiar with HR department.
iv) All departments should have more skilled and promising employees.
v) Creating familiar environment with employees so they don't leave the job.
vi) Hiring well experienced employees and train appropriate skills to the freshers.
vii) Hire more married people who are aware of the responsibilities.
viii) Don’t change the manager at a frequent interval.