Loading the necessary libraries, reading the data set and viewinng the data

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
df=pd.read_csv('D:/DS_Files/LetsUpgrade-AI-ML/Day-7/Assignment/general_data.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWorl
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	

5 rows × 24 columns

In [4]:

```
df
```

Out[4]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesW
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	
...	...	...	...	...	...	...	...	...	...	...	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	4406	Female	...	
4406	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	4407	Male	...	
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	4408	Male	...	
4408	42	No	Travel_Rarely	Sales	18	2	Medical	1	4409	Male	...	
4409	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	4410	Male	...	

4410 rows × 24 columns

In [5]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4410 entries, 0 to 4409
Data columns (total 24 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Age                      4410 non-null   int64  
 1   Attrition                4410 non-null   object 
 2   BusinessTravel           4410 non-null   object 
 3   Department               4410 non-null   object 
 4   DistanceFromHome         4410 non-null   int64  
 5   Education                4410 non-null   int64  
 6   EducationField           4410 non-null   object 
 7   EmployeeCount            4410 non-null   int64  
 8   EmployeeID               4410 non-null   int64  
 9   Gender                   4410 non-null   object 
 10  JobLevel                 4410 non-null   int64  
 11  JobRole                  4410 non-null   object 
 12  MaritalStatus            4410 non-null   object 
 13  MonthlyIncome            4410 non-null   int64  
 14  NumCompaniesWorked       4391 non-null   float64
 15  Over18                   4410 non-null   object 
 16  PercentSalaryHike        4410 non-null   int64  
 17  StandardHours            4410 non-null   int64  
 18  StockOptionLevel         4410 non-null   int64  
 19  TotalWorkingYears        4401 non-null   float64
 20  TrainingTimesLastYear    4410 non-null   int64  
 21  YearsAtCompany           4410 non-null   int64  
 22  YearsSinceLastPromotion  4410 non-null   int64  
 23  YearsWithCurrManager     4410 non-null   int64  
dtypes: float64(2), int64(14), object(8)
memory usage: 827.0+ KB
```

Converting the string fields to equivalent numerical labels

In [6]:

```
from sklearn import preprocessing
le=preprocessing.LabelEncoder()
df['Attrition']=le.fit_transform(df['Attrition'])
df['BusinessTravel']=le.fit_transform(df['BusinessTravel'])
df['Department']=le.fit_transform(df['Department'])
df['EducationField']=le.fit_transform(df['EducationField'])
df['Gender']=le.fit_transform(df['Gender'])
df['MaritalStatus']=le.fit_transform(df['MaritalStatus'])
df['Over18']=le.fit_transform(df['Over18'])
df['JobeRole']=le.fit_transform(df['JobRole'])
```

Dropping the null values

In [7]:

```
df.dropna()
```

Out[7]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	Over18	PercentSa
0	51	0	2	2	6	2	1	1	1	0	...	0	
1	31	1	1	1	10	1	1	1	2	0	...	0	
2	32	0	1	1	17	4	4	1	3	1	...	0	
3	38	0	0	1	2	5	1	1	4	1	...	0	
4	32	0	2	1	10	1	3	1	5	1	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
4404	29	0	2	2	4	3	4	1	4405	0	...	0	
4405	42	0	2	1	5	4	3	1	4406	0	...	0	
4406	29	0	2	1	2	4	3	1	4407	1	...	0	
4407	25	0	2	1	25	2	1	1	4408	1	...	0	
4408	42	0	2	2	18	2	3	1	4409	1	...	0	

4382 rows × 25 columns

In [8]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4410 entries, 0 to 4409
Data columns (total 25 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Age                      4410 non-null   int64  
 1   Attrition                4410 non-null   int32  
 2   BusinessTravel           4410 non-null   int32  
 3   Department               4410 non-null   int32  
 4   DistanceFromHome         4410 non-null   int64  
 5   Education                4410 non-null   int64  
 6   EducationField           4410 non-null   int32  
 7   EmployeeCount            4410 non-null   int64  
 8   EmployeeID               4410 non-null   int64  
 9   Gender                   4410 non-null   int32  
 10  JobLevel                 4410 non-null   int64  
 11  JobRole                  4410 non-null   object 
 12  MaritalStatus            4410 non-null   int32  
 13  MonthlyIncome            4410 non-null   int64  
 14  NumCompaniesWorked       4391 non-null   float64
 15  Over18                   4410 non-null   int32  
 16  PercentSalaryHike        4410 non-null   int64  
 17  StandardHours            4410 non-null   int64  
 18  StockOptionLevel         4410 non-null   int64  
 19  TotalWorkingYears        4401 non-null   float64
 20  TrainingTimesLastYear    4410 non-null   int64  
 21  YearsAtCompany           4410 non-null   int64  
 22  YearsSinceLastPromotion  4410 non-null   int64  
 23  YearsWithCurrManager     4410 non-null   int64  
 24  JobeRole                 4410 non-null   int32  
dtypes: float64(2), int32(8), int64(14), object(1)
memory usage: 723.6+ KB
```

Importing and loading the Statistical Test Package

In [11]:

```
from scipy.stats import pearsonr
```

In [15]:

```
#H0:Business Travel from haome has no effect on attrition
#H1: Business Travel from home has effect on attrition
r, p = pearsonr(df['Attrition'],df['BusinessTravel'])
print(r,p)
```

```
7.377694602220437e-05 0.9960919945440154
```

In [13]:

```
#H0:Departement has no effect on attrition
#H1: Department has effect on attrition
r, p = pearsonr(df['Attrition'],df['Department'])
print(r,p)
```

```
-0.04820581991833714 0.0013638319632111042
```

In [92]:

```
#H0:Distance from haome has no effect on attrition
#H1: Distance from home has effect on attrition
r, p = pearsonr(df['Attrition'],df['DistanceFromHome'])
print(r,p)
```

```
-0.009730141010179435 0.5182860428049617
```

In [93]:

```
#H0:Education has no effect on attrition
#H1: Education has effect on attrition
r, p = pearsonr(df['Attrition'],df['Education'])
print(r,p)
```

```
-0.015111167710968753 0.3157293177118575
```

In [94]:

```
#H0:EducationField has no effect on attrition
#H1: Education Field has effect on attrition
r, p = pearsonr(df['Attrition'],df['EducationField'])
print(r,p)
```

```
-0.05794031241568037 0.00011819790920717528
```

In [95]:

```
#H0:Gender has no effect on attrition
#H1: Gender has effect on attrition
r, p = pearsonr(df['Attrition'],df['Gender'])
print(r,p)
```

```
0.018125078877010366 0.22881970951790567
```

In [96]:

```
#H0:Joblevel has no effect on attrition
#H1: Joblevel has effect on attrition

r, p = pearsonr(df['Attrition'],df['JobLevel'])
print(r,p)
```

```
-0.010289713287495079 0.49451717271828405
```

In [97]:

```
#H0:MaritalStatus has no effect on attrition
#H1: MaritalStatus has effect on attrition
r, p = pearsonr(df['Attrition'],df['MaritalStatus'])
print(r,p)
```

```
0.025808853490974722 0.08658208267566762
```

In the above tests, p>0.05 for Business Travel,distance from home, Education, Gender, Joblevel and Marital status. so we accept the null hypothesis for these as they dont have that effect on attrition.

P<0.05 for Department, Education Field and Marital Status, which show that they have an effect on the attrition of the employees.

So the company wants to focus mainly to place resourses properly on the apt departments which sould be according to their right fied of education and matching job profiles.

In [ ]:

```
```