

ATTRITION PROJECT

➤ **PROBLEM STATEMENT:**

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 acclimatize 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.

➤ **SOLUTION:**

STEP1 - LAUNCHING

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

• ***To load the dataset :***

```
dataset=pd.read_csv("file:///E:/LetsUpgrade/Day_7 Assignment/general_data.csv")
```

• ***To view first few rows of data:***

```
dataset.head()
```

Out[7]:

	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 25 columns]

- **To check the column names:**

dataset.columns()

Out[9]:

Index(['Age', 'Attrition', 'Attrition1', 'BusinessTravel', 'Department', 'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'], dtype='object')

- **To view entire dataset:**

dataset

Out[10]:

	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 25 columns]



STEP2: DATA TREATMENT

- **To check if there are any null values in the dataset:**

```
dataset.isnull()
```

```
Out[11]:
```

	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 25 columns]
```

- **To check if there are any duplicate values:**

```
dataset.duplicated()
```

```
Out[13]:
```

0	False
1	False
2	False
3	False
4	False
:	:
:	:
4405	False
4406	False
4407	False
4408	False
4409	False

- **To remove duplicate values(if any):**

```
dataset.drop_duplicates()
```

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
...
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 25 columns]



STEP 3 : UNIVARIATE ANALYSIS

- Describing the entire dataset:**

```
dataset2=dataset.describe()
```

```
dataset2
```

dataset2 - DataFrame

Index	Age	Attrition	DistanceFromHome	Education	EmployeeCount	EmployeeID	JobLevel	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears
min	18	0	1	1	1	1	1	10090	0	11	8	0	0
25%	30	0	2	2	1	1103.25	1	29110	1	12	8	0	6
50%	36	0	7	3	1	2205.5	2	49190	2	14	8	1	10
std	9.1333	0.367847	8.10503	1.02393	0	1273.2	1.10669	47068.9	2.49889	3.65911	0	0.851883	7.7222
mean	36.9238	0.161298	9.19252	2.91293	1	2205.5	2.06395	65029.3	2.69483	15.2095	8	0.793878	11.2799
75%	43	0	14	4	1	3307.75	3	83800	4	18	8	1	15
max	60	1	29	5	1	4410	5	199990	9	25	8	3	40
count	4410	4408	4410	4410	4410	4410	4410	4410	4391	4410	4410	4410	440

dataset2 - DataFrame

Index	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
min	90	0	11	8	0	0	0	0	0	0
25%	10	1	12	8	0	6	2	3	0	2
50%	90	2	14	8	1	10	3	5	1	3
std	58.9	2.49889	3.65911	0	0.851883	7.78222	1.28898	6.12514	3.2217	3.56733
mean	29.3	2.69483	15.2095	8	0.793878	11.2799	2.79932	7.00816	2.18776	4.12313
75%	90	4	18	8	1	15	3	9	3	7
max	990	9	25	8	3	40	6	40	15	17
count	0	4391	4410	4410	4410	4401	4410	4410	4410	4410

- **To find the median:**

```
dataset1=dataset.median()
```

```
dataset1
```

dataset1 - Series

Index	0
Age	36
Attrition1	0
DistanceFromHome	7
Education	3
EmployeeCount	1
EmployeeID	2205.5
JobLevel	2
MonthlyIncome	49190
NumCompaniesWorked	2
PercentSalaryHike	14
StandardHours	8
StockOptionLevel	1
TotalWorkingYears	10
TrainingTimesLastYear	3
YearsAtCompany	5
YearsSinceLastPromotion	1
YearsWithCurrManager	3

- **To find the mode:**

```
dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompanies Worked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].mode()
```

```
dataset1
```

dataset1 - DataFrame

Index	Age	DistanceFromHome	Education	JobLevel	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
0	35	2	3	1	23420	1	11	10	2	5	0	2

- **To find mean:**

```
dataset1=dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].mean()
```

Index	0
Age	36.9238
DistanceFromHome	9.19252
Education	2.91293
JobLevel	2.06395
MonthlyIncome	65029.3
NumCompaniesWorked	2.69483
PercentSalaryHike	15.2095
TotalWorkingYears	11.2799
TrainingTimesLastYear	2.79932
YearsAtCompany	7.00816
YearsSinceLastPromotion	2.18776
YearsWithCurrManager	4.12313

- To find variance:**

```
dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompanies
Worked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinc
eLastPromotion','YearsWithCurrManager']].var()
```

dataset1

dataset1 - Series

Index	0
Age	83.4172
DistanceFromHome	65.6914
Education	1.04844
JobLevel	1.22476
MonthlyIncome	2.21548e+09
NumCompaniesWorked	6.24444
PercentSalaryHike	13.3891
TotalWorkingYears	60.563
TrainingTimesLastYear	1.66146
YearsAtCompany	37.5173
YearsSinceLastPromotion	10.3793
YearsWithCurrManager	12.7258

- To find the skewness:**

```
dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompanies
Worked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinc
eLastPromotion','YearsWithCurrManager']].skew()
```

dataset1

dataset1 - Series

Index	0
Age	0.413005
DistanceFromHome	0.957466
Education	-0.289484
JobLevel	1.0247
MonthlyIncome	1.36888
NumCompaniesWorked	1.02677
PercentSalaryHike	0.820569
TotalWorkingYears	1.11683
TrainingTimesLastYear	0.552748
YearsAtCompany	1.76333
YearsSinceLastPromotion	1.98294
YearsWithCurrManager	0.832884

- To find kurtosis:**

```
dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompanies
Worked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinc
eLastPromotion','YearsWithCurrManager']].kurt()
```

dataset1

dataset1 - Series

Index	0
Age	-0.405951
DistanceFromHome	-0.227045
Education	-0.560569
JobLevel	0.395525
MonthlyIncome	1.00023
NumCompaniesWorked	0.00728748
PercentSalaryHike	-0.302638
TotalWorkingYears	0.912936
TrainingTimesLastYear	0.491149
YearsAtCompany	3.92386
YearsSinceLastPromotion	3.60176
YearsWithCurrManager	0.167949

- To find standard deviation:**

```
dataset1=dataset[['Age','DistanceFromHome','Education','JobLevel','MonthlyIncome','NumCompanies
Worked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinc
eLastPromotion','YearsWithCurrManager']].std()
```

dataset1

dataset1 - Series

Index	0
Age	9.1333
DistanceFromHome	8.10503
Education	1.02393
JobLevel	1.10669
MonthlyIncome	47068.9
NumCompaniesWorked	2.49889
PercentSalaryHike	3.65911
TotalWorkingYears	7.78222
TrainingTimesLastYear	1.28898
YearsAtCompany	6.12514
YearsSinceLastPromotion	3.2217
YearsWithCurrManager	3.56733

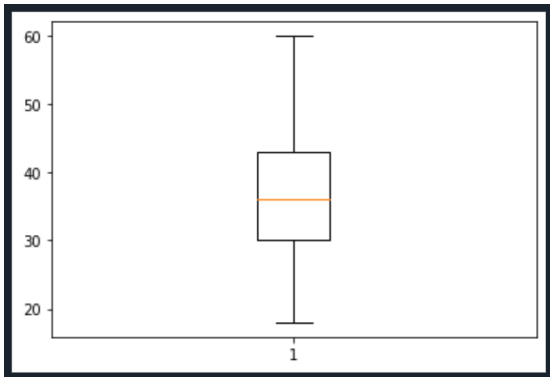
❖ INFERENCE :

	Mean	Median	Mode	Variance	Standard Deviation	IQR	Skew	Kurtosis
Age	36.9238	36	35	83.4172	9.1333	13	0.413005	-0.40595
DistanceFromHome	9.19252	7	2	65.6914	8.10503	12	0.957466	-0.22704
Education	2.91293	3	3	1.04844	1.02393	2	-0.28948	-0.56056
JobLevel	2.06395	2	1	1.22476	1.10669	2	1.0247	0.395525
MonthlyIncome	65029.3	49190	23420	2.21548e+09	47068.9	54690	1.36888	1.00023
NumCompaniesWorked	2.69483	2	1	6.24444	2.49889	3	1.02677	0.007287
PercentSalaryHike	15.2095	14	11	13.3891	3.65911	6	0.820569	-0.30263
TotalWorkingYears	11.2799	10	10	60.563	7.78222	9	1.11683	0.912936
TrainingTimesLastYear	2.79932	3	2	1.66146	1.28898	1	0.552748	0.491149
YearsAtCompany	7.00816	5	5	37.5173	6.12514	6	1.76333	3.92386
YearsSinceLastPromotion	2.18776	1	0	10.3793	3.2217	3	1.98294	3.60176
YearsWithCurrManager	4.12313	3	2	12.7258	3.56733	5	0.832884	0.167949

- Mean Age forms a near normal distribution with 13 years of IQR.
- The Mean_Monthly_Income's IQR is at 54K suggesting companywide attrition across all income bands.
- All the above variables show positive skewness except education.
- Age, Distance_from_home, Education, PercentSalaryHike are leptokurtic and all other variables are platykurtic.

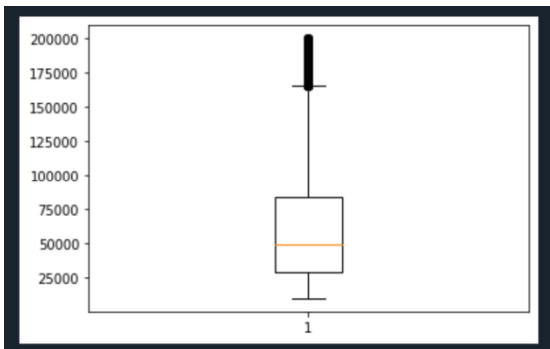
❖ OUTLIERS:

- **Checking outliers in Age:**



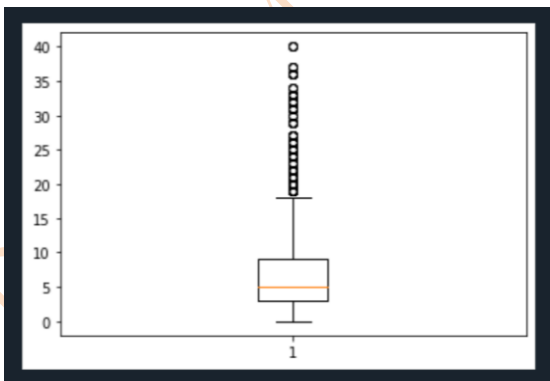
Age is normally distributed without any outliers.

- **Checking outliers in Monthly income:**



Monthly Income is Right skewed with several outliers.

- **Checking outliers in Years at Company:**



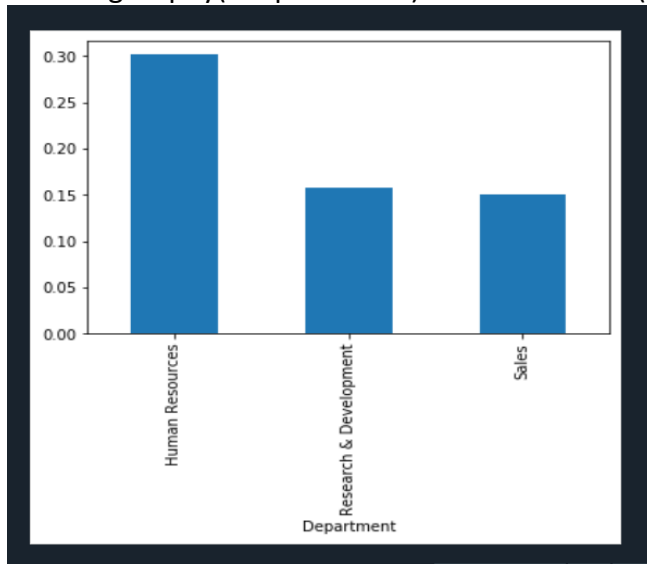
Years at company is Right Skewed with several outliers.



STEP 4 : VISUALIZATION

- **Attrition VS Department:**

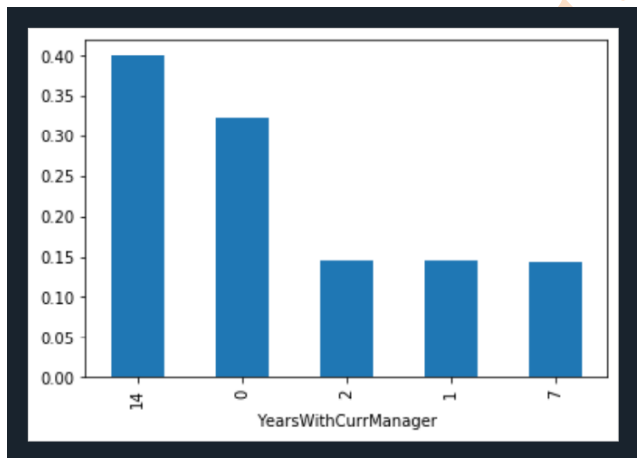
```
dataset.groupby("Department").Attrition1.mean().sort_values(ascending=False)[:5].plot.bar()
```



Attrition is maximum in HR field.

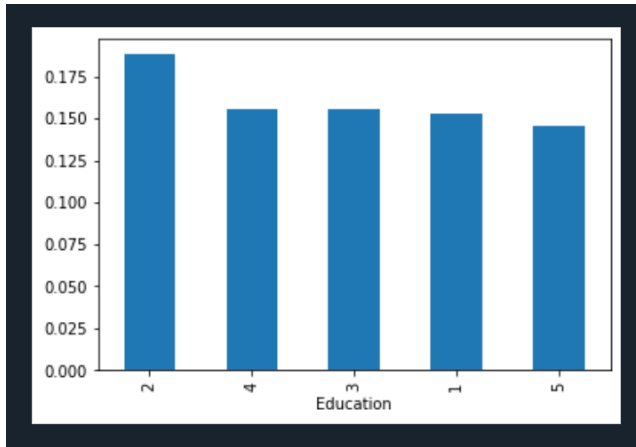
- **Attrition VS Current Manager Relationship:**

```
dataset.groupby("YearsWithCurrManager").Attrition1.mean().sort_values(ascending=False)[:5].plot.bar()
```



- **Attrition VS Education Level:**

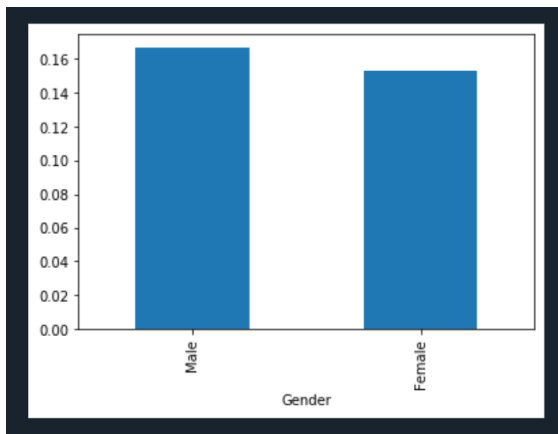
```
dataset.groupby("Education").Attrition1.mean().sort_values(ascending=False)[:5].plot.bar()
```



Attrition is major in Education Level 2 i.e, employees who have been working and also studying in college.

- **Attrition VS Gender:**

```
dataset.groupby("Gender").Attrition1.mean().sort_values(ascending=False)[:5].plot.bar()
```



Attrition is comparatively more in Male than Female.

+ **STEP 5 : STATISTICAL TESTS**

1. Mann-Whitney Test:

a) Attrition VS Distance from Home:

- H0 : There is no significance of Attrition on distance from home.
- H1 : There is significance of age on distance from home.
- `stats.p=mannwhitneyu(dataset.Attrition,dataset.DistanceFromHome)`
- `print(stats,p)-----> 230652.0 0.0`
- As $p=0.0$, hence we reject H0 and accept H1.

b) Attrition VS Age:

- H0 : There is no significance of Age on Attrition.

- H1 : There is significance of Age on Attrition.
- stats,p=mannwhitneyu(dataset.Attrition,dataset.Age)
- print(stats,p)-----> 8820.0 0.0
- As p=0.0, hence we reject H0 and accept H1.

c) Attrition VS Income:

- H0 : Monthly Income is contributing to Attrition.
- H1 : There is no affect of Monthly Income on Attrition.
- stats,p=mannwhitneyu(dataset.Attrition,dataset.MonthlyIncome)
- print(stats,p)-----> 8820.0 0.0
- As p=0.0, hence we reject H0 and accept H1.

d) Attrition VS Total Working Years:

- H0 : People having experience are not contributing to Attrition.
- H1 : There is no specific years of experience which contributes to Attrition.
- stats,p=mannwhitneyu(dataset.Attrition,dataset.TotalWorkingYears)
- print(stats,p)-----> 179296.5 0.0
- As p=0.0, hence we reject H0 and accept H1.

e) Attrition VS YearsWithCurrManager:

- H0 : The relationship with current manager doesn't cause Attrition.
- H1 : The relationship with current manager may cause Attrition.
- stats,p=mannwhitneyu(dataset.Attrition,dataset.YearsWithCuurManager)
- print(stats,p)-----> 2109319.5 0.0
- As p=0.0, hence we reject H0 and accept H1.

2. Chi-Square Test:

a) Comparison between Gender and Attrition :

- H0 : There is no significance of Gender on Attrition.
- H1 : There is significance of Gender on Attrition.
- chitable=pd.crosstab(dataset.Gender,dataset.Attrition)
- chitable

ATTRITION	Yes	No
GENDER		
Female	1494	270
Male	2205	441

- stats,p,dot,expected=chi2_contingency(chitable)
- print(stats,p)-----> 1.349904410246582 0.24529482862926827

- As $p=0.245$, hence we accept H_0 .

b) Comparison between Attrition and Department:

- H_0 : There is no significant impact of Attrition on Department.
- H_1 : There is significant impact of Attrition on Department.
- `chitable=pd.crosstab(dataset.Department,dataset.Attrition)`
- `chitable`

ATTRITION	No	Yes
DEPARTMENT		
Human Resources	132	57
Research & Development	2430	453
Sales	1137	201

- `stats,p,dot,expected=chi2_contingency(chitable)`
- `print(stats,p)`----->29.090274924488263 4.820888218170407e-07
- As $p=4.82e-07$, hence we reject H_0 and accept H_1 .

c) Comparison between Attrition and Education Field:

- H_0 : There is no significance of Education Field on Attrition.
- H_1 : There is significance of Education Field on Attrition.
- `chitable=pd.crosstab(dataset.EducationField,dataset.Attrition)`
- `chitable`

ATTRITION	No	Yes
EDUCATION FIELD		
Human Resources	48	33
Life Sciences	1515	303
Marketing	402	75
Medical	1167	225
Other	216	30
Technical Degree	351	45

- `stats,p,dot,expected=chi2_contingency(chitable)`
- `print(stats,p)`----->46.194921001730584 8.288917469574179e-09
- As $p=8.28e-09$, hence we reject H_0 and accept H_1 .

d) Comparison between Attrition and Job Role:

- H_0 : There is no significant impact of Job role on Attrition.
- H_1 : There is significant impact of Job role on Attrition.
- `chitable=pd.crosstab(dataset.JobRole,dataset.Attrition)`

ATTRITION	No	Yes
JOB ROLE		
Healthcare Representative	336	57
Human Resources	135	21

Laboratory Technician	651	126
Manager	264	42
Manufacturing Director	387	48
Research Director	183	57
Research Scientist	717	159
Sales Executive	813	165
Sales Representative	213	36

- `stats,p,dot,expected=chi2_contingency(chitable)`
- `print(stats,p)-----> 25.11631367460407` 0.0014855447448152669
- As $p=0.001$, hence we reject the H_0 and accept H_1 .

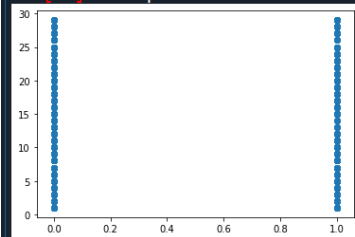


STEP 6 : UNSUPERVISED LEARNING – CORRELATION ANALYSIS

1) Correlation between Attrition and Distance From Home:

- Correlation value ' r ' = -0.00963, which means the variables are weak negatively correlated.
- Probability ' p ' value = 0.522, and $p > 0.05$, so we can accept the Null Hypothesis.
- Hence, there is no significant correlation between Attrition and Distance from Home.

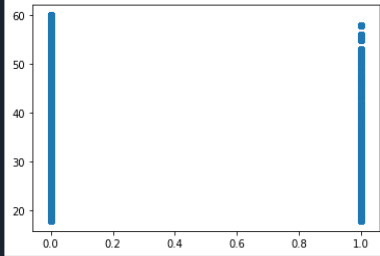
```
In [37]: stats,p=pearsonr(dataset.Attrition1,dataset.DistanceFromHome)
In [38]: print(stats,p)
-0.009638784678344565 0.5223162968450248
In [39]: plt.scatter(dataset.Attrition1,dataset.DistanceFromHome)
Out[39]: <matplotlib.collections.PathCollection at 0x1eda16ceac8>
```



2) Correlation between Attrition and Age:

- Correlation value ' r ' = -0.159, which means the variables are negatively correlated.
- Probability ' p ' value = $2.09e-26$, so we can reject the Null Hypothesis.
- Hence, there is significant correlation between Age and Attrition.

```
In [40]: plt.scatter(dataset.Attrition1,dataset.Age)
Out[40]: <matplotlib.collections.PathCollection at 0x1eda17a20c8>
```



```
In [41]: stats,p=pearsonr(dataset.Attrition1,dataset.Age)
```

```
In [42]: print(stats,p)
-0.15917551489227316 2.0935226759299425e-26
```

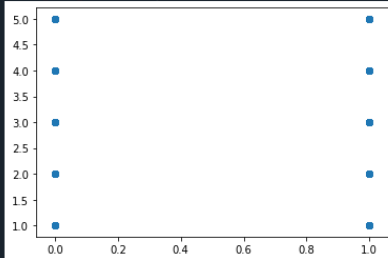
3) Correlation between Attrition and Education Level:

- Correlation value ' r ' = -0.014, which means the variables are weak negatively correlated.
- Probability ' p ' value = 0.319, so we can accept the Null Hypothesis as it affects around 32% of data.
- Hence, there is no significant correlation of Education Level and Attrition.

```
In [43]: stats,p=pearsonr(dataset.Attrition1,dataset.Education)
```

```
In [44]: print(stats,p)
-0.0149996869479451 0.31942298183450585
```

```
In [45]: plt.scatter(dataset.Attrition1,dataset.Education)
Out[45]: <matplotlib.collections.PathCollection at 0x1eda163c908>
```



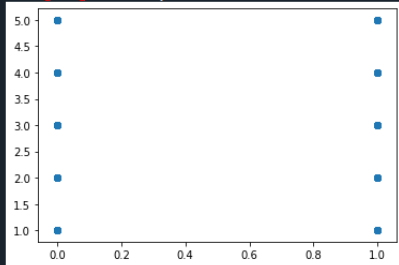
4) Correlation between Attrition and Job Level:

- Correlation value ' r ' = -0.0102, which means the variables are weakly negatively correlated.
- Probability ' p ' value = 0.497, so we can reject the Null Hypothesis.
- Hence, there is significant correlation between Attrition and Job Level.

```
In [46]: stats,p=pearsonr(dataset.Attrition1,dataset.JobLevel)

In [47]: print(stats,p)
-0.010216389132645505 0.497695798535647

In [48]: plt.scatter(dataset.Attrition1,dataset.JobLevel)
Out[48]: <matplotlib.collections.PathCollection at 0x1eda18c5488>
```



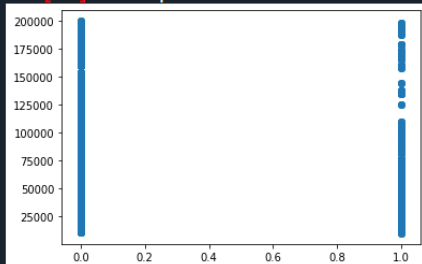
5) Correlation between Attrition and Monthly Income:

- Correlation value ' r ' = -0.0313, which means the variables are weakly negatively correlated.
- Probability ' p ' value = 0.037, so we can reject the Null Hypothesis as only 3% of the data is affected.
- Hence, there is significant correlation between Attrition and Monthly Income.

```
In [49]: stats,p=pearsonr(dataset.Attrition1,dataset.MonthlyIncome)

In [50]: print(stats,p)
-0.0313613334267405 0.03733448740624388

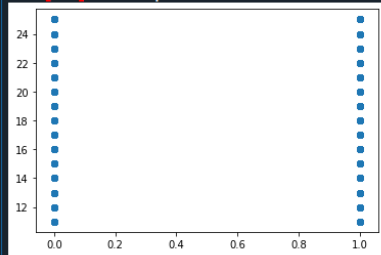
In [51]: plt.scatter(dataset.Attrition1,dataset.MonthlyIncome)
Out[51]: <matplotlib.collections.PathCollection at 0x1eda1932e48>
```



6) Correlation between Attrition and Salary Hike Percentage:

- Correlation value ' r ' = 0.0324, which means the variables are positively correlated.
- Probability ' p ' value = 0.031, so we can reject the Null Hypothesis as only 3% of data is affected.
- Hence, there is significant correlation between Attrition and Percentage of Salary hike.

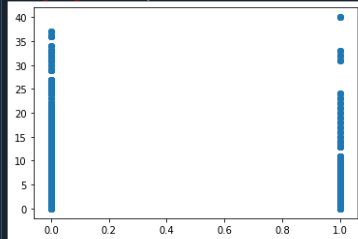

```
In [53]: stats,p=pearsonr(dataset.Attrition1,dataset.PercentSalaryHike)
In [54]: print(stats,p)
0.032469604641723576 0.03110678713888575
In [55]: plt.scatter(dataset.Attrition1,dataset.PercentSalaryHike)
Out[55]: <matplotlib.collections.PathCollection at 0x1eda2970288>
```



7) Correlation between Attrition and Number of Years at company:

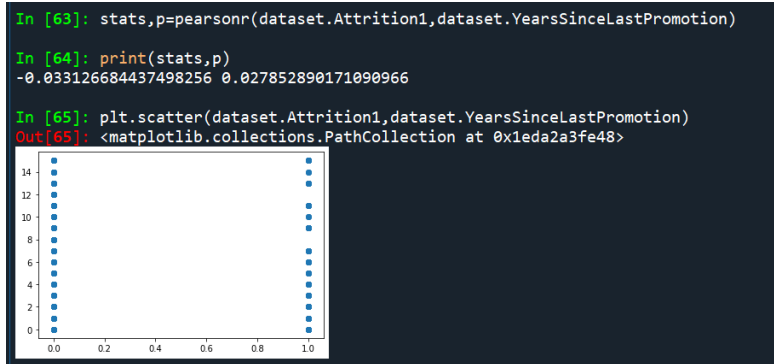
- Correlation value ' r ' = -0.134, which means the variables are negatively correlated.
- Probability ' p ' value = 3.33, so we can accept the Null Hypothesis as 33% of data is affected.
- Hence, there is no significant correlation between Attrition and Number of Years in company.

```
In [60]: stats,p=pearsonr(dataset.Attrition1,dataset.YearsAtCompany)
In [61]: print(stats,p)
-0.13433741288732426 3.33237622244819e-19
In [62]: plt.scatter(dataset.Attrition1,dataset.YearsAtCompany)
Out[62]: <matplotlib.collections.PathCollection at 0x1eda2a3ff88>
```



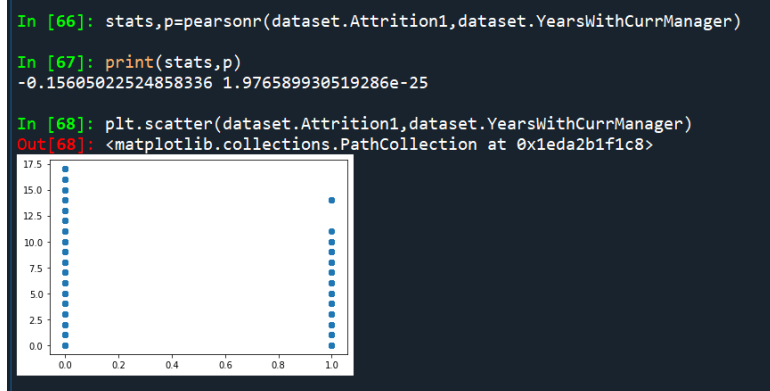
8) Correlation between Attrition and Number of years worked since last Promotion:

- Correlation value ' r ' = -0.0331, which means the variables are weakly negatively correlated.
- Probability ' p ' value = 0.027, so we can reject the Null Hypothesis as only 2% data is affected.
- Hence, there is significant correlation between Attrition and Promotion.



9) Correlation between Attrition and Number of years worked with Current Manager:

- Correlation value ' r ' = -0.156, which means the variables are negatively correlated.
- Probability ' p ' value = 1.96×10^{-25} , so we can accept the Null Hypothesis.
- Hence, there is significant correlation between Attrition and relationship with current Manager.



CONCLUSION :

From the above Statistical Tests and Correlation Analysis, we find that there are few factors that are contributing towards Attrition and rest have no significance for Attrition. From the above mentioned inferences, I conclude that the company has to make favorable changes in their rules, so that the Attrition can be reduced to maximum.

Project By-

Divya S M