Assignment – Attrition Analysis

Step 1: Load the sheet/Data

import pandas as pd

import matplotlib.pyplot as mplt

dataset = pd.read\_csv("D:/AI\_ML\_Course/Day 7/general\_data.csv")

dataset.columns

Out[3]:

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

dataset.isnull()

Out[4]:

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]

dataset.duplicated()

Out[6]:

0 False

1 False

2 False

3 False

4 False

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

dataset.drop\_duplicates()

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

... ... ... ... ...

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

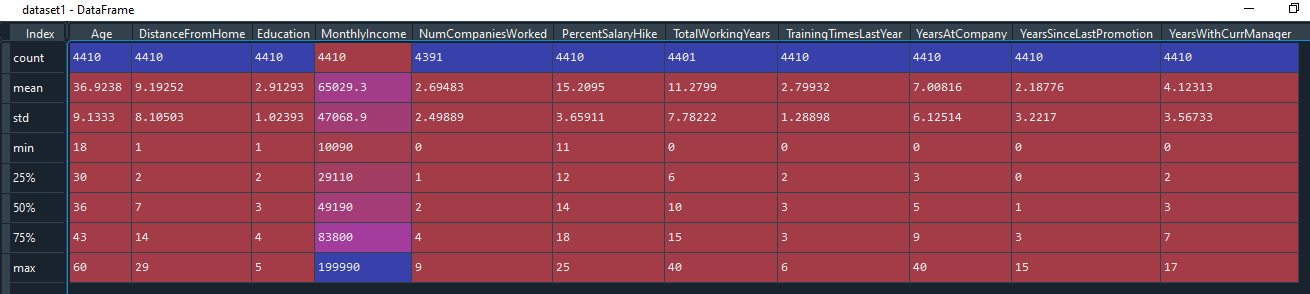
[4410 rows x 24 columns]

Step 3: Univariate Analysis:

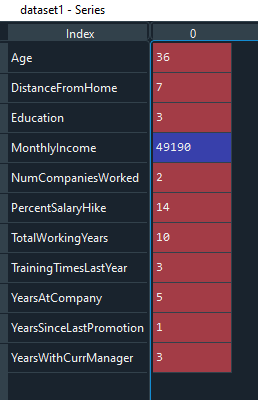
1. Analysis with Complete Dataset

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

Dataset1



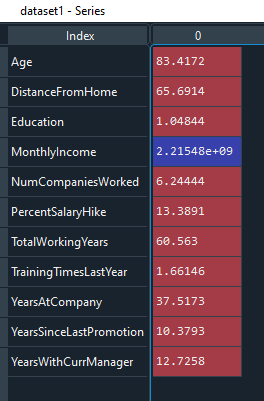
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



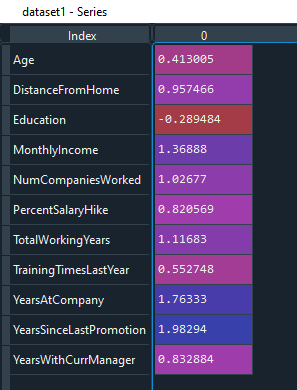
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()



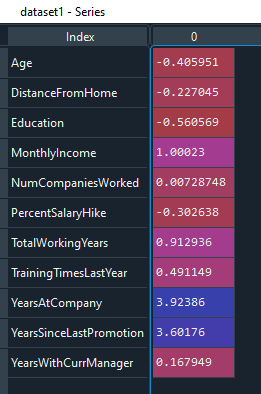
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()



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



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



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Std Deviation** | **IQR** | **Skewness** | **Kurtosis** |
| **Age (Yrs)** | 36.9 | 36 | 35 | 83.41 | 9.13 | 13 | 0.41 | -0.41 |
| **DistanceFromHome (Km)** | 9.19 | 7 | 2 | 65.69 | 8.1 | 12 | 0.96 | -0.23 |
| **Monthly Income (Rs)** | 65029 | 49190 | 23420 | 2215480270 | 47068 | 54690 | 1.37 | 1 |
| **PercentSalaryHike (%)** | 15 | 14 | 11 | 13.39 | 3.66 | 6 | 0.82 | -0.3 |
| **TotalWorkingYears (Yrs)** | 11.29 | 10 | 10 | 60.56 | 7.78 | 9 | 1.12 | 0.91 |
| **YearsAtCompany (Yrs)** | 7 | 5 | 5 | 37.52 | 6.12 | 6 | 1.76 | 3.92 |
| **YearsSinceLastPromotion (Yrs)** | 2 | 1 | 0 | 10.38 | 3.22 | 3 | 1.98 | 3.6 |
| **YearsWithCurrManager (Yrs)** | 4 | 3 | 2 | 12.73 | 3.57 | 5 | 0.83 | 0.17 |

Inference from the analysis:

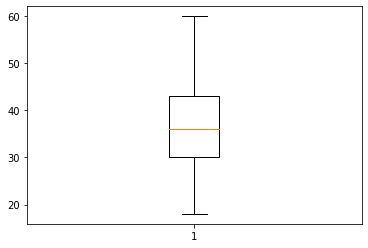
* All the above variables show positive skewness.
* Years\_At\_Company & Years\_Since\_LastPromotion are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
* The Mean\_Monthly\_Income’s IQR is at 54K suggesting companywide attrition across all income bands
* Mean age forms a near normal distribution with 13 years of IQR
* Mean Distance\_From\_Home is 12 Km of IQR which is higher.

Outliers:

There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

box\_plot=dataset.Age

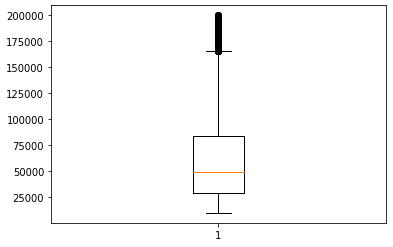
mplt. boxplot(box\_plot)



Age is normally distributed without any Outliers

box\_plot=dataset.MonthlyIncome

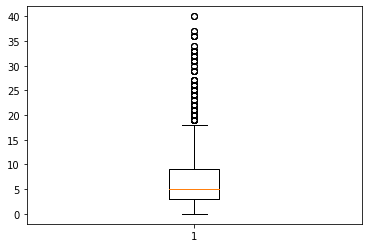
mplt.boxplot(box\_plot)



Monthly Income is right skewed with several Outliers

box\_plot=dataset.YearsAtCompany

mplt.boxplot(box\_plot)

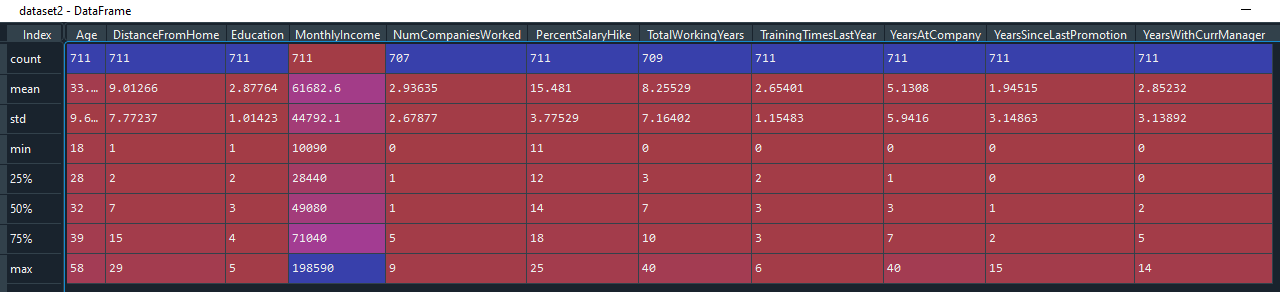


Years at company is also Right skewed with several Outliers

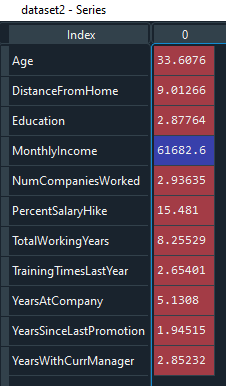
1. Analysis with dataset having Attrition as Yes

dataset1=dataset[dataset['Attrition']=='Yes']

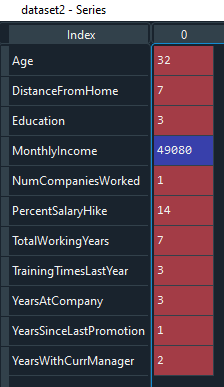
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()



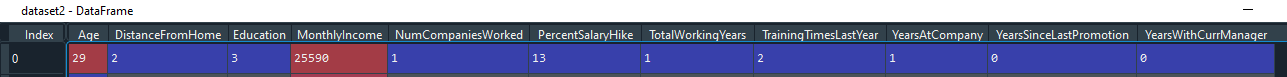
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()



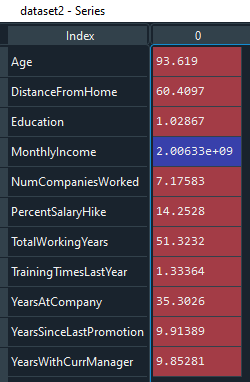
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



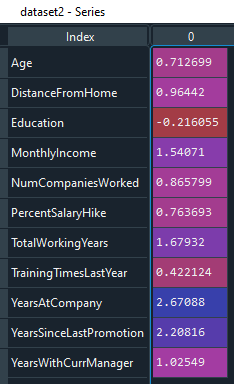
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()



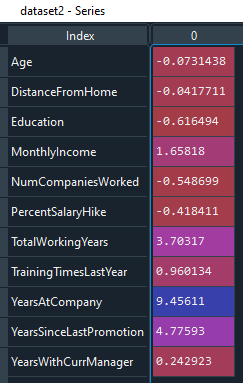
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()



dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()



dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Std Deviation** | **IQR** | **Skewness** | **Kurtosis** |
| **Age (Yrs)** | 33 | 32 | 29 | 93 | 9.67 | 11 | 0.71 | -0.073 |
| **DistanceFromHome (Km)** | 9 | 7 | 2 | 60 | 7.77 | 13 | 0.96 | -0.04 |
| **MonthlyIncome (Rs)** | 61682 | 49080 | 25590 |  | 45000 | 42600 | 1.54 | 1.65 |
| **PercentSalaryHike (%)** | 15.4 | 14 | 13 | 14.25 | 3.77 | 6 | 0.76 | -0.41 |
| **TotalWorkingYears (Yrs)** | 8 | 7 | 1 | 51 | 7.16 | 7 | 1.67 | 3.7 |
| **YearsAtCompany (Yrs)** | 5.13 | 3 | 1 | 35 | 5.94 | 6 | 2.67 | 9.45 |
| **YearsSinceLastPromotion (Yrs)** | 1.94 | 1 | 0 | 9.91 | 3.14 | 2 | 2.2 | 4.77 |
| **YearsWithCurrManager (Yrs)** | 2.85 | 2 | 0 | 9.85 | 3.13 | 5 | 1.02 | 0.24 |

Inference from the analysis:

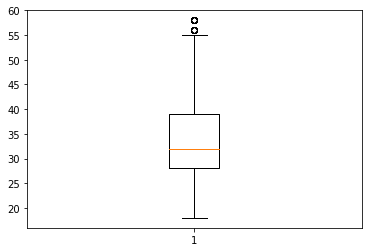
* All the above variables show positive skewness.
* Years\_At\_Company, Years\_Since\_LastPromotion & Total\_Working\_Years are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
* The Mean\_Monthly\_Income’s IQR is at 42K suggesting companywide attrition across all income bands
* Mean age forms a near normal distribution with 11 years of IQR
* Mean Distance\_From\_Home is 13 Km of IQR which is higher.

Outliers:

There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

box\_plot=dataset1.Age

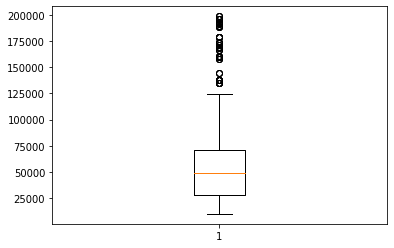
mplt.boxplot(box\_plot)



Age is normally distributed but there are few Outliers. This might be retired/voluntary retired employees.

box\_plot=dataset1.MonthlyIncome

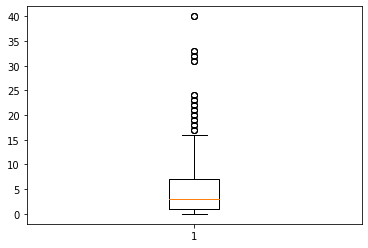
mplt.boxplot(box\_plot)



Monthly Income is Right Skewed with several Outliers

box\_plot=dataset1.YearsAtCompany

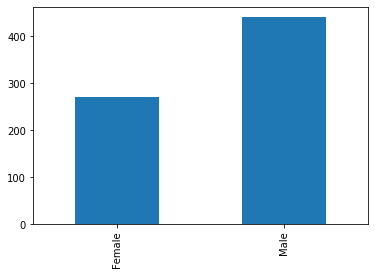
mplt.boxplot(box\_plot)



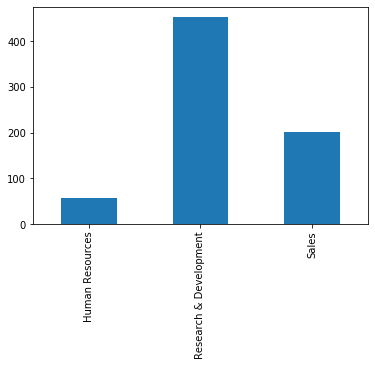
Years at company is also Right skewed with several Outliers

Step 4: Visualization

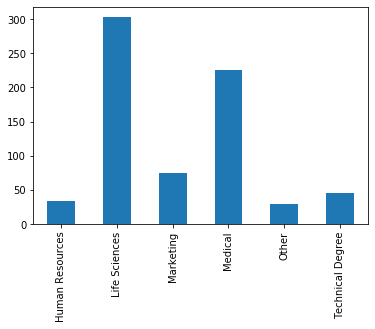
1. Attrition Vs Gender



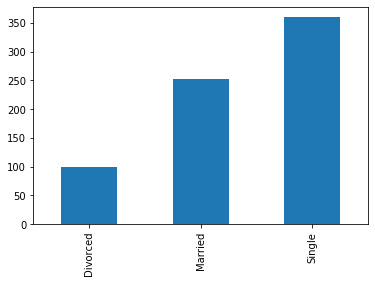
1. Attrition Vs Department



1. Attrition Vs Education Fields



1. Attrition Vs Marital Status



Step 5 – Statistical Tests (Mann-Whitney)

Attrition Vs Distance from Home

from scipy.stats import mannwhitneyu

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.DistanceFromHome

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.DistanceFromHome

stat, p = mannwhitneyu(a1,a2)

print(stat, p)

1312110.0 0.4629185205822659

As the P value of 0.46 is > 0.05, the H0 is accepted and Ha is rejected.

H0: There is no significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

Attrition Vs Monthly Income

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.MonthlyIncome

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.MonthlyIncome

stat, p = mannwhitneyu(a1,a2)

print(stat, p)

1264900.5 0.053577283839938566

As the P value of 0.053 is > 0.05, the H0 is accepted and Ha is rejected.

H0: There is no significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

Attrition Vs Years at Company

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.YearsAtCompany

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.YearsAtCompany

stat, p = mannwhitneyu(a1,a2)

print(stat, p)

923238.0 6.047598261692858e-37

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

Attrition Vs YearsWithCurrManager

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.YearsWithCurrManager

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.YearsWithCurrManager

stat, p = mannwhitneyu(a1,a2)

print(stat, p)

957253.5 1.2365483142169853e-31

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

Attrition Vs YearsSinceLastPromotion

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1. YearsSinceLastPromotion

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2. YearsSinceLastPromotion

stat, p = mannwhitneyu(a1,a2)

print(stat, p)

1209366.0 0.0002021180346719736

As the P value of 0.0002 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

Step 6: Statistical Test (Separate T Test)

Attrition Vs Distance from Home

from scipy.stats import ttest\_ind

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.DistanceFromHome

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.DistanceFromHome

stat, p = ttest\_ind(a1,a2)

print(stat, p)

-0.6460416038042738 0.518286042805572

As the P value of 0.51 is > 0.05, the H0 is accepted and Ha is rejected.

H0: There is no significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

Attrition Vs Monthly Income

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.MonthlyIncome

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.MonthlyIncome

stat, p = ttest\_ind(a1,a2)

print(stat, p)

-2.0708863763619316 0.03842748490605113

As the P value of 0.038 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

Attrition Vs Years at Company

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.YearsAtCompany

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.YearsAtCompany

stat, p = ttest\_ind(a1,a2)

print(stat, p)

-9.004357011787226 3.163883122491456e-19

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

Attrition Vs YearsWithCurrManager

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1.YearsWithCurrManager

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2.YearsWithCurrManager

stat, p = ttest\_ind(a1,a2)

print(stat, p)

-10.499379408703438 1.7339322652918153e-25

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

Attrition Vs YearsSinceLastPromotion

dataset1=dataset[dataset['Attrition']=='Yes']

a1 = dataset1. YearsSinceLastPromotion

dataset2 = dataset[dataset['Attrition']=='No']

a2 = dataset2. YearsSinceLastPromotion

stat, p = ttest\_ind(a1,a2)

print(stat, p)

-2.1934039604328843 0.028330336189428353

As the P value of 0.028 is < 0.05, the H0 is rejected and Ha is accepted.

H0: There is no significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

Ha: There is a significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

Step 7: Correlation Analysis

In order to find the interdependency of the variables DistanceFromHome, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager, YearsSinceLastPromotion from that of Attrition, we executed the Correlation Analysis as follows.

import pandas as pd

dataset1 = pd.read\_csv("D:/AI\_ML\_Course/Day 7/general\_data.csv")

from scipy.stats import pearsonr

stats, pdfh=pearsonr(dataset1.Attrition,dataset1.DistanceFromHome)

print(stats, pdfh)

-0.009730141010179674 0.5182860428050771

stats, pmi=pearsonr(dataset1.Attrition,dataset1.MonthlyIncome)

print(stats, pmi)

-0.031176281698115007 0.03842748490600132

stats, pyac=pearsonr(dataset1.Attrition,dataset1.YearsAtCompany)

print(stats, pyac)

-0.1343922139899772 3.1638831224877484e-19

stats, pywcm=pearsonr(dataset1.Attrition,dataset1.YearsWithCurrManager)

print(stats, pywcm)

-0.15619931590162847 1.7339322652896276e-25

stats, pyslp=pearsonr(dataset1.Attrition,dataset1.YearsSinceLastPromotion)

print(stats, pyslp)

-0.03301877514258434 0.028330336189396753

The Inference for the above Analysis is as follows:

Attrition & Distance from Home:

As r = -0.009, there’s low negative correlation between Attrition and DistanceFromHome

As the P value of 0.518 is > 0.05, we are accepting H0 and hence there’s no significant correlation between Attrition & DistanceFromHome

Attrition & MonthlyIncome:

As r = -0.031, there’s low negative correlation between Attrition and MonthlyIncome

As the P value of 0.038 is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition & MonthlyIncome

Attrition & YearsAtCompany:

As r = -0.1343, there’s low negative correlation between Attrition and YearsAtCompany

As the P value is 0.0 < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition & YearsAtCompany

Attrition & YearsWithCurrManager:

As r = -0.1561, there’s low negative correlation between Attrition and YearsWithCurrManager

As the P value is 0.0 > 0.05, we are accepting Ha and hence there’s significant correlation between Attrition & YearsWithCurrManager

Attrition & YearsSinceLastPromotion:

As r = -0.033, there’s low negative correlation between Attrition and YearsSinceLastPromotion

As the P value 0.028 is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition & YearsSinceLastPromotion