**Correlation Analysis on Attrition Dataset**

**Step 1: Launching**

import pandas as pd

dataset = pd.read\_excel("Attrition Analysis Data.xlsx", sheet\_name = 0)

dataset.head()

Out[3]:

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]

dataset.tail()

Out[4]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

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

[5 rows x 24 columns]

dataset.columns

Out[5]:

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.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

dataset1 = dataset.drop\_duplicates()

dataset1.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]

working\_dataset = dataset1.dropna()

working\_dataset.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 4382 entries, 0 to 4408

Data columns (total 24 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 4382 non-null int64

1 Attrition 4382 non-null object

2 BusinessTravel 4382 non-null object

3 Department 4382 non-null object

4 DistanceFromHome 4382 non-null int64

5 Education 4382 non-null int64

6 EducationField 4382 non-null object

7 EmployeeCount 4382 non-null int64

8 EmployeeID 4382 non-null int64

9 Gender 4382 non-null object

10 JobLevel 4382 non-null int64

11 JobRole 4382 non-null object

12 MaritalStatus 4382 non-null object

13 MonthlyIncome 4382 non-null int64

14 NumCompaniesWorked 4382 non-null float64

15 Over18 4382 non-null object

16 PercentSalaryHike 4382 non-null int64

17 StandardHours 4382 non-null int64

18 StockOptionLevel 4382 non-null int64

19 TotalWorkingYears 4382 non-null float64

20 TrainingTimesLastYear 4382 non-null int64

21 YearsAtCompany 4382 non-null int64

22 YearsSinceLastPromotion 4382 non-null int64

23 YearsWithCurrManager 4382 non-null int64

dtypes: float64(2), int64(14), object(8)

memory usage: 855.9+ KB

**Step 3: Correlation Analysis**

from scipy.stats import pearsonr

working\_dataset.info()

<class 'pandas.core.frame.DataFrame'>

Int64Index: 4382 entries, 0 to 4408

Data columns (total 24 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 Age 4382 non-null int64

1 Attrition 4382 non-null object

2 BusinessTravel 4382 non-null object

3 Department 4382 non-null object

4 DistanceFromHome 4382 non-null int64

5 Education 4382 non-null int64

6 EducationField 4382 non-null object

7 EmployeeCount 4382 non-null int64

8 EmployeeID 4382 non-null int64

9 Gender 4382 non-null object

10 JobLevel 4382 non-null int64

11 JobRole 4382 non-null object

12 MaritalStatus 4382 non-null object

13 MonthlyIncome 4382 non-null int64

14 NumCompaniesWorked 4382 non-null float64

15 Over18 4382 non-null object

16 PercentSalaryHike 4382 non-null int64

17 StandardHours 4382 non-null int64

18 StockOptionLevel 4382 non-null int64

19 TotalWorkingYears 4382 non-null float64

20 TrainingTimesLastYear 4382 non-null int64

21 YearsAtCompany 4382 non-null int64

22 YearsSinceLastPromotion 4382 non-null int64

23 YearsWithCurrManager 4382 non-null int64

dtypes: float64(2), int64(14), object(8)

memory usage: 855.9+ KB

from sklearn import preprocessing

test = preprocessing.LabelBinarizer()

working\_dataset['Attrition']= test.fit\_transform(working\_dataset['Attrition'])

\_\_main\_\_:1: SettingWithCopyWarning:

A value is trying to be set on a copy of a slice from a DataFrame.

Try using .loc[row\_indexer,col\_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.Age)

print(stats,p)

-0.15839867954096706 5.1265982193780794e-26

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.DistanceFromHome)

print(stats,p)

-0.009448638515156243 0.5317715668019558

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.JobLevel)

print(stats,p)

-0.012381569720790865 0.4125489150380087

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.MonthlyIncome)

print(stats,p)

-0.030160293808460668 0.045890862744719166

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.NumCompaniesWorked)

print(stats,p)

0.042830567244720875 0.004572057121620842

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.PercentSalaryHike)

print(stats,p)

0.03315303713546665 0.028192446935106235

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.TotalWorkingYears)

print(stats,p)

-0.16966991684723917 1.1645434967091854e-29

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.YearsAtCompany)

print(stats,p)

-0.13300261842521538 9.476118084840815e-19

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.YearsSinceLastPromotion)

print(stats,p)

-0.03142315056330995 0.03752293607395154

stats,p = pearsonr(working\_dataset.Attrition,working\_dataset.YearsWithCurrManager)

print(stats,p)

-0.15469153690287274 7.105369646771178e-25

**Inference:**

**Attrition and Age**

As r = -0.15, there’s low negative correlation between Attrition and Age

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

**Attrition and DistanceFromHome**

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

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

**Attrition and JobLevel**

As r = -0.012, there’s low negative correlation between Attrition and JobLevel

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

**Attrition and MonthlyIncome**

As r = -0.03, there’s low negative correlation between Attrition and MonthlyIncome

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

**Attrition and NumCompaniesWorked**

As r = 0.04, there’s low positive correlation between Attrition and NumCompaniesWorked

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

**Attrition and PercentSalaryHike**

As r = 0.033, there’s low positive correlation between Attrition and PercentSalaryHike

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

**Attrition and TotalWorkingYears**

As r = -0.016, there’s low negative correlation between Attrition and TotalWorkingYears

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

**Attrition and YearsAtCompany**

As r = -0.13, there’s low negative correlation between Attrition and YearsAtCompany

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

**Attrition and YearsSinceLastPromotion**

As r = -0.03, there’s low negative correlation between Attrition and YearsSinceLastPromotion

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

**Attrition and YearsWithCurrManager**

As r = -0.15, there’s low negative correlation between Attrition and YearsWithCurrManager

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