Day 13 Assignment

Statistical Tests (Mann-Whitney)

Attrition Vs Distance from Home

import pandas as pd

dataset=pd.read\_excel('C:/Group\_Folder/TheDataScience/Dinesh/Group 1- HR Analytics - Employee

Attrition rate analysis/Working\_sheet.xlsx', sheet\_name=1)

dataset.head()

Out[3]:

DistanceFromHome\_Yes ... YearsWithCurrManager\_No

0 0 ... 0

1 10 ... 0

2 0 ... 3

3 0 ... 5

4 0 ... 4

[5 rows x 10 columns]

dataset.columns

Out[4]:

Index(['DistanceFromHome\_Yes', 'DistanceFromHome\_No', 'MonthlyIncome\_Yes',

'MonthlyIncome\_No', 'TotalWorkingYears\_Yes', 'TotalWorkingYears\_No',

'YearsAtCompany\_Yes', 'YearsAtCompany\_No', 'YearsWithCurrManager\_Yes',

'YearsWithCurrManager\_No'],

dtype='object')

from scipy.stats import mannwhitneyu

a1=dataset.DistanceFromHome\_Yes

a2=dataset.DistanceFromHome\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3132625.5 0.0

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

H0: There is no significant differences in the Distance From Home between attrition (Y) and attirition

(N)

Ha: There is significant differences in the Distance From Home between attrition (Y) and attirition (N)

Attrition Vs Income

a1=dataset.MonthlyIncome\_Yes

a2=dataset.MonthlyIncome\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3085416.0 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the income between attrition (Y) and attirition (N)

Ha: There is significant differences in the income between attrition (Y) and attirition (N)

Attrition Vs Total Working Years

a1=dataset.TotalWorkingYears\_Yes

a2=dataset.TotalWorkingYears\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

2760982.0 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Total Working Years between attrition (Y) and attirition

(N)

Ha: There is significant differences in the Total Working Years between attrition (Y) and attirition (N)

Attrition Vs Years at company

a1=dataset.YearsAtCompany\_Yes

a2=dataset.YearsAtCompany\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

2882047.5 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Years At Company between attrition (Y) and attirition

(N)

Ha: There is significant differences in the Years At Company between attrition (Y) and attirition (N)

Attrition Vs YearsWithCurrentManager

a1=dataset.YearsWithCurrManager\_Yes

a2=dataset.YearsWithCurrManager\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3674749.5 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Years With Current Manager between attrition (Y) and

attirition (N)

Ha: There is significant differences in the Years With Current Manager between attrition (Y) and

attirition (N)

Step 6 – Statistical Tests (Separate T Test)

Attrition Vs Distance From Home

from scipy.stats import ttest\_ind

dataset.columns

Out[49]:

Index(['DistanceFromHome\_Yes', 'DistanceFromHome\_No', 'MonthlyIncome\_Yes',

'MonthlyIncome\_No', 'TotalWorkingYears\_Yes', 'TotalWorkingYears\_No',

'YearsAtCompany\_Yes', 'YearsAtCompany\_No', 'YearsWithCurrManager\_Yes',

'YearsWithCurrManager\_No'],

dtype='object')

z1=dataset.DistanceFromHome\_Yes

z2=dataset.DistanceFromHome\_No

stat, p=ttest\_ind(z2,z1)

print(stat, p)

44.45445917636664 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Distance From Home between attrition (Y) and attirition

(N)

Ha: There is significant differences in the Distance From Home between attrition (Y) and attirition (N)

Attrition Vs Income

z1=dataset.MonthlyIncome\_Yes

z2=dataset.MonthlyIncome\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

52.09279408504947 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Monthly Income between attrition (Y) and attirition (N)

Ha: There is significant differences in the Monthly Income between attrition (Y) and attirition (N)

Attrition Vs Yeats At Company

z1=dataset.YearsAtCompany\_Yes

z2=dataset.YearsAtCompany\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

51.45296941515692 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Years At Company between attrition (Y) and attirition

(N)

Ha: There is significant differences in the Years At Company between attrition (Y) and attirition (N)

Attrition Vs Years With Current Manager

z1=dataset.YearsWithCurrManager\_Yes

z2=dataset.YearsWithCurrManager\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

53.02424349024521 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted.

H0: There is no significant differences in the Years With Current Manager between attrition (Y) and

attirition (N)

Ha: There is significant differences in the Years With Current Manager between attrition (Y) and

attirition (N)

Step 8 – Unsupervised Learning - Correlation Analysis

In order to find the interdependency of the variables DistanceFromHome, MonthlyIncome,

TotalWorkingYears, YearsAtCompany, YearsWithCurrManager from that of Attrition, we executed

the Correlation Analysis as follows.

dataset=pd.read\_excel('C:/Group\_Folder/TheDataScience/Dinesh/Group 1- HR Analytics - Employee

Attrition rate analysis/Working\_sheet.xlsx', sheet\_name=0)

from scipy.stats import pearsonr

dataset['TotalWorkingYears']=dataset['TotalWorkingYears'].fillna(11.28)

dataset.columns

Out[258]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus',

'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike',

'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany',

'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

stats, p=pearsonr(dataset.Attrition, dataset.DistanceFromHome)

print(stats, p)

-0.009730141010179438 0.5182860428049617

stats, p=pearsonr(dataset.Attrition, dataset.MonthlyIncome)

print(stats, p)

-0.031176281698114025 0.0384274849060192

stats, p=pearsonr(dataset.Attrition, dataset.TotalWorkingYears)

print(stats, p)

-0.17011136355964646 5.4731597518148054e-30

stats, p=pearsonr(dataset.Attrition, dataset.YearsAtCompany)

print(stats, p)

-0.13439221398997386 3.163883122493571e-19

stats, p=pearsonr(dataset.Attrition, dataset.YearsWithCurrManager)

print(stats, p)

-0.15619931590162422 1.7339322652951965e-25

The inference of the above analysis are as follows:

Attrition & DistanceFromHome:

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 & TotalWorkingYears:

As r = -0.17, 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 & YearsAtCompany:

As r = -0.1343, 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 & YearsWithCurrManager:

As r = -0.1561, 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 & YearsWithCurrManagers