Day 12 Assignment - Hitik Panchal

In [23]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from scipy.stats import mannwhitneyu
from scipy.stats import ttest_ind
```

Reading the Data

In [7]:

```
gen_data=pd.read_csv('general_data.csv')
gen_data.head()
```

Out[7]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationFie
0	51	No	Travel_Rarely	Sales	6	2	Life Scienc
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Scienc
2	32	No	Travel_Frequently	Research & Development	17	4	Otl
3	38	No	Non-Travel	Research & Development	2	5	Life Scienc
4	32	No	Travel_Rarely	Research & Development	10	1	Medi

5 rows × 24 columns

In [8]:

```
gen_data.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
```

Cleaning the Data

In [9]:

```
gen_data.isnull().any()
```

Out[9]:

```
Age                       False
Attrition                 False
BusinessTravel            False
Department                False
DistanceFromHome          False
Education                 False
EducationField            False
EmployeeCount             False
EmployeeID                False
Gender                    False
JobLevel                  False
JobRole                   False
MaritalStatus             False
MonthlyIncome             False
NumCompaniesWorked         True
Over18                    False
PercentSalaryHike         False
StandardHours             False
StockOptionLevel          False
TotalWorkingYears          True
TrainingTimesLastYear     False
YearsAtCompany            False
YearsSinceLastPromotion   False
YearsWithCurrManager      False
dtype: bool
```

In [10]:

```
gen_data.fillna(0 , inplace=True)
```

In [11]:

```
gen_data.isnull().any()
```

Out[11]:

```
Age                       False
Attrition                 False
BusinessTravel            False
Department                False
DistanceFromHome          False
Education                 False
EducationField            False
EmployeeCount             False
EmployeeID                False
Gender                    False
JobLevel                  False
JobRole                   False
MaritalStatus             False
MonthlyIncome             False
NumCompaniesWorked        False
Over18                    False
PercentSalaryHike         False
StandardHours             False
StockOptionLevel          False
TotalWorkingYears         False
TrainingTimesLastYear     False
YearsAtCompany            False
YearsSinceLastPromotion   False
YearsWithCurrManager      False
dtype: bool
```

In [12]:

```
gen_data.duplicated()
```

Out[12]:

```
0       False
1       False
2       False
3       False
4       False
        ...
4405    False
4406    False
4407    False
4408    False
4409    False
Length: 4410, dtype: bool
```

In [13]:

```
gen_data.drop_duplicates()
```

Out[13]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	Educatio
0	51	No	Travel_Rarely	Sales	6	2	Life Sci
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sci
2	32	No	Travel_Frequently	Research & Development	17	4	
3	38	No	Non-Travel	Research & Development	2	5	Life Sci
4	32	No	Travel_Rarely	Research & Development	10	1	M
...	...	...	...	...	...	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	M
4406	29	No	Travel_Rarely	Research & Development	2	4	M
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sci
4408	42	No	Travel_Rarely	Sales	18	2	M
4409	40	No	Travel_Rarely	Research & Development	28	3	M

4410 rows × 24 columns

In [20]:

```
yes_data = pd.DataFrame(gen_data[gen_data['Attrition']=='Yes'])
yes_data
```

Out[20]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	Educatio
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sci
6	28	Yes	Travel_Rarely	Research & Development	11	2	M
13	47	Yes	Non-Travel	Research & Development	1	1	M
28	44	Yes	Travel_Frequently	Research & Development	1	2	M
30	26	Yes	Travel_Rarely	Research & Development	4	3	M
...	...	...	...	...	...	...	
4381	29	Yes	Travel_Rarely	Research & Development	7	1	Life Sci
4386	33	Yes	Travel_Rarely	Sales	11	4	Mar
4388	33	Yes	Travel_Rarely	Sales	1	3	Life Sci
4391	32	Yes	Travel_Rarely	Sales	23	1	Life Sci
4402	37	Yes	Travel_Frequently	Sales	2	3	Mar

711 rows × 24 columns

In [22]:

```
no_data = pd.DataFrame(gen_data[gen_data['Attrition']=='No'])
no_data
```

Out[22]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	Educatio
0	51	No	Travel_Rarely	Sales	6	2	Life Sci
2	32	No	Travel_Frequently	Research & Development	17	4	
3	38	No	Non-Travel	Research & Development	2	5	Life Sci
4	32	No	Travel_Rarely	Research & Development	10	1	M
5	46	No	Travel_Rarely	Research & Development	8	3	Life Sci
...	...	...	...	...	...	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	M
4406	29	No	Travel_Rarely	Research & Development	2	4	M
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sci
4408	42	No	Travel_Rarely	Sales	18	2	M
4409	40	No	Travel_Rarely	Research & Development	28	3	M

3699 rows × 24 columns

Statistical Tests

Seperate T-Test

Test 1 : Attrition vs Distance From Home

In [33]:

```
a1 = yes_data['DistanceFromHome']
a2 = no_data['DistanceFromHome']

tval, p = ttest_ind(a2,a1)

print("The T value is : %.3f" % tval)
print("The p value is : %.3f" % p)
```

```
The T value is : 0.646
The p value is : 0.518
```

As the P value is 0.518 , which is > than 0.05, the Ha is rejected and H0 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)

Test 2 : Attrition vs Monthly Income

In [34]:

```
a1 = yes_data['MonthlyIncome']
a2 = no_data['MonthlyIncome']

tval, p = ttest_ind(a2,a1)

print("The T value is : %.3f" % tval)
print("The p value is : %.3f" % p)
```

```
The T value is : 2.071
The p value is : 0.038
```

As the P value is 0.038 , which is < than 0.05, the Ha is accepted and H0 is rejected

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)

Test 3 : Attrition vs Years at Company

In [35]:

```
a1 = yes_data['YearsAtCompany']
a2 = no_data['YearsAtCompany']

tval, p = ttest_ind(a2,a1)

print("The T value is : %.3f" % tval)
print("The p value is : %.3f" % p)
```

```
The T value is : 9.004
The p value is : 0.000
```

As the P value is 0.0 , which is < than 0.05, the Ha is accepted and H0 is rejected

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)

Test 4 : Attrition vs Years with Current Manager

In [32]:

```
a1 = yes_data['YearsWithCurrManager']
a2 = no_data['YearsWithCurrManager']

tval, p = ttest_ind(a2,a1)

print("The T value is : %.3f" % tval)
print("The p value is : %.3f" % p)
```

```
The T value is : 10.499
The p value is : 0.000
```

As the P value is 0.0 , which is < than 0.05, the Ha is accepted and H0 is rejected

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)

Mann-Whitney Test

Test 1 : Attrition vs Distance From Home

In [37]:

```
a1 = yes_data['DistanceFromHome']
a2 = no_data['DistanceFromHome']

uval , p = mannwhitneyu(a1,a2)

print("The U value is : %.3f" % uval)
print("The p value is : %.3f" % p)
```

```
The U value is : 1312110.000
The p value is : 0.463
```

As the P value is 0.463 , which is > than 0.05, the Ha is rejected and H0 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)

Test 2 : Attrition vs Monthly Income

In [38]:

```
a1 = yes_data['MonthlyIncome']
a2 = no_data['MonthlyIncome']

uval , p = mannwhitneyu(a1,a2)

print("The U value is : %.3f" % uval)
print("The p value is : %.3f" % p)
```

```
The U value is : 1264900.500
The p value is : 0.054
```

As the P value is 0.054 , which is > than 0.05, the Ha is rejected and H0 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)

Test 3 : Attrition vs Years at Company

In [39]:

```
a1 = yes_data['YearsAtCompany']
a2 = no_data['YearsAtCompany']

uval , p = mannwhitneyu(a1,a2)

print("The U value is : %.3f" % uval)
print("The p value is : %.3f" % p)
```

```
The U value is : 923238.000
The p value is : 0.000
```

As the P value is 0.0 , which is < than 0.05, the Ha is accepted and H0 is rejected

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)

Test 4 : Attrition vs Years with Current Manager

In [40]:

```
a1 = yes_data['YearsWithCurrManager']
a2 = no_data['YearsWithCurrManager']

uval , p = mannwhitneyu(a1,a2)

print("The U value is : %.3f" % uval)
print("The p value is : %.3f" % p)
```

```
The U value is : 957253.500
The p value is : 0.000
```

As the P value is 0.0 , which is < than 0.05, the Ha is accepted and H0 is rejected

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)

In [ ]: