**Attrition Assignment**

**Step1 – Launching**

import pandas as pd

dataset=pd.read\_csv("general\_data\_org.csv")

dataset.head()

Out[8]:

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.columns

Out[14]:

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')

dataset

Out[15]:

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 2 - Data Treatment:**

dataset.isnull()

Out[16]:

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[17]:

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[18]:

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:**

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

dataset3

Out[20]:

Age ... YearsWithCurrManager

count 4410.000000 ... 4410.000000

mean 36.923810 ... 4.123129

std 9.133301 ... 3.567327

min 18.000000 ... 0.000000

25% 30.000000 ... 2.000000

50% 36.000000 ... 3.000000

75% 43.000000 ... 7.000000

max 60.000000 ... 17.000000

[8 rows x 11 columns]

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

dataset3

Out[22]:

Age 36.0

DistanceFromHome 7.0

Education 3.0

MonthlyIncome 49190.0

NumCompaniesWorked 2.0

PercentSalaryHike 14.0

TotalWorkingYears 10.0

TrainingTimesLastYear 3.0

YearsAtCompany 5.0

YearsSinceLastPromotion 1.0

YearsWithCurrManager 3.0

dtype: float64

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

dataset3

Out[24]:

Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager

0 35 2 ... 0 2

[1 rows x 11 columns]

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

dataset3

Out[26]:

Age 8.341719e+01

DistanceFromHome 6.569144e+01

Education 1.048438e+00

MonthlyIncome 2.215480e+09

NumCompaniesWorked 6.244436e+00

PercentSalaryHike 1.338907e+01

TotalWorkingYears 6.056298e+01

TrainingTimesLastYear 1.661465e+00

YearsAtCompany 3.751728e+01

YearsSinceLastPromotion 1.037935e+01

YearsWithCurrManager 1.272582e+01

dtype: float64

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()

dataset3

Out[28]:

Age 0.413005

DistanceFromHome 0.957466

Education -0.289484

MonthlyIncome 1.368884

NumCompaniesWorked 1.026767

PercentSalaryHike 0.820569

TotalWorkingYears 1.116832

TrainingTimesLastYear 0.552748

YearsAtCompany 1.763328

YearsSinceLastPromotion 1.982939

YearsWithCurrManager 0.832884

dtype: float64

dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

dataset3

Out[30]:

Age -0.405951

DistanceFromHome -0.227045

Education -0.560569

MonthlyIncome 1.000232

NumCompaniesWorked 0.007287

PercentSalaryHike -0.302638

TotalWorkingYears 0.912936

TrainingTimesLastYear 0.491149

YearsAtCompany 3.923864

YearsSinceLastPromotion 3.601761

YearsWithCurrManager 0.167949

dtype: float64

**Inference from the analysis**: All the above variables show positive skewness; while Age & Mean\_distance\_from\_home are leptokurtic and all other variables are platykurtic.

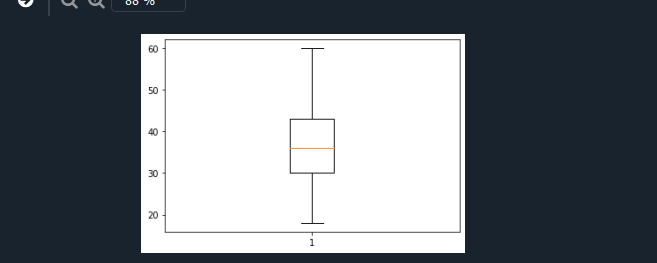
The Mean\_Monthly\_Income’s IQR is at 54K suggesting company wide attrition across all income bands Mean age forms a near normal distribution with 13 years of IQR

Outliers:

There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, DistanceFromHome , PercentSalaryHike , on a scatter plot

box\_plot=dataset.Age

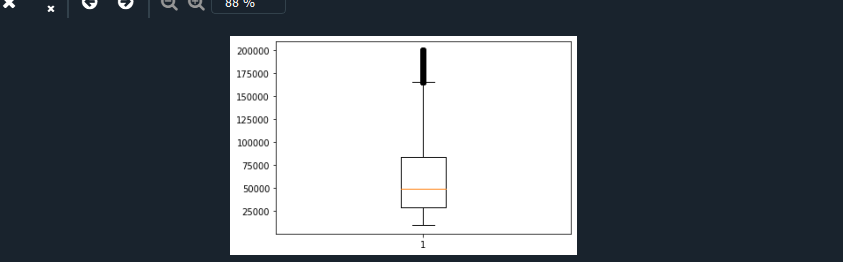
plt.boxplot(box\_plot)



Age is normally distributed without any outliers

box\_plot=dataset.MonthlyIncome

plt.boxplot(box\_plot)

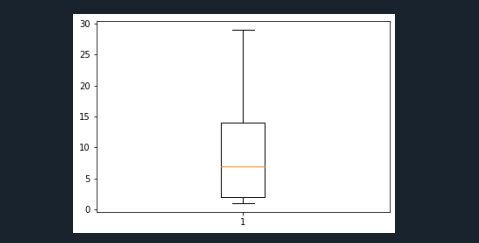


Monthly Income is Right skewed with several outliers

box\_plot=dataset.DistanceFromHome

plt.boxplot(box\_plot)

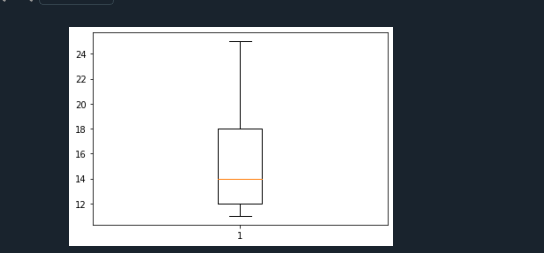
Out[56]:



DistanceFromHome is normally distributed without any outliers

box\_plot=dataset.PercentSalaryHike

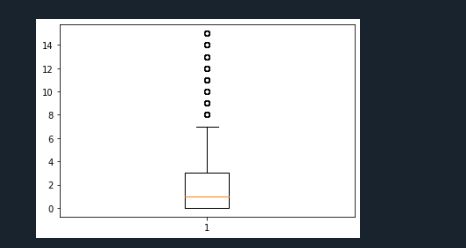
plt.boxplot(box\_plot)



PercentSalaryHike is normally distributed without any outliers

box\_plot=dataset.YearsSinceLastPromotion

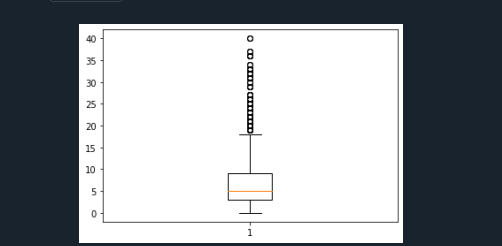
plt.boxplot(box\_plot)



YearsSinceLastPromotion is also Right Skewed with several outliers observed.

box\_plot=dataset1.YearsAtCompany

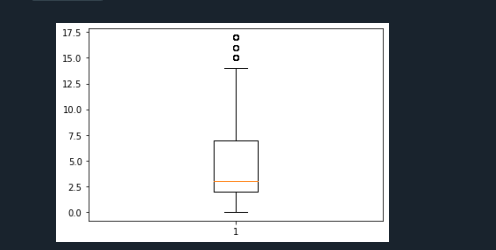
plt.boxplot(box\_plot)



Years at company is also Right Skewed with several outliers observed.

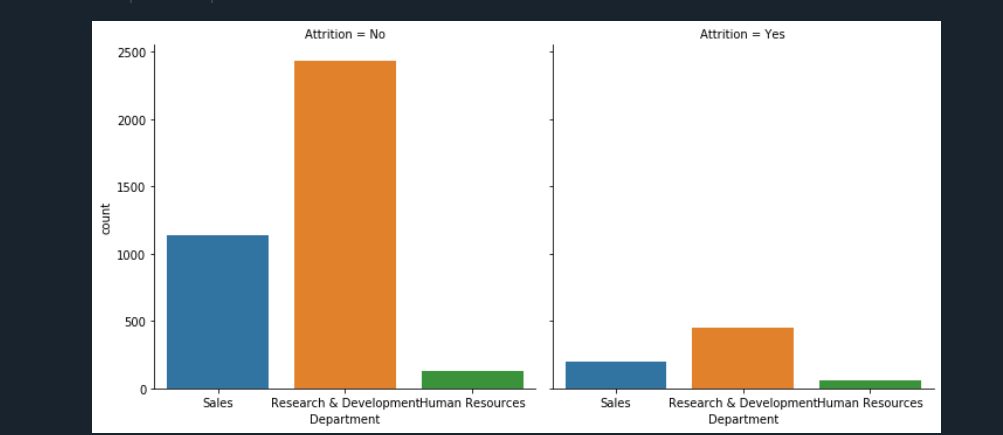
box\_plot=dataset.YearsWithCurrManager

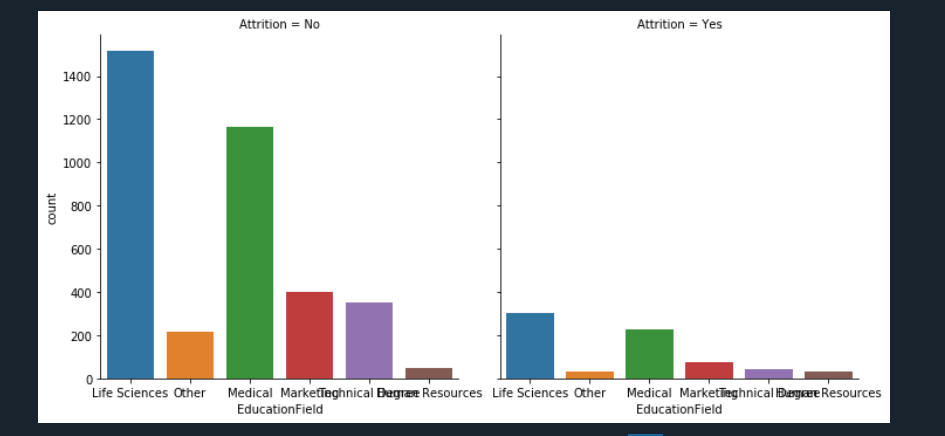
plt.boxplot(box\_plot)

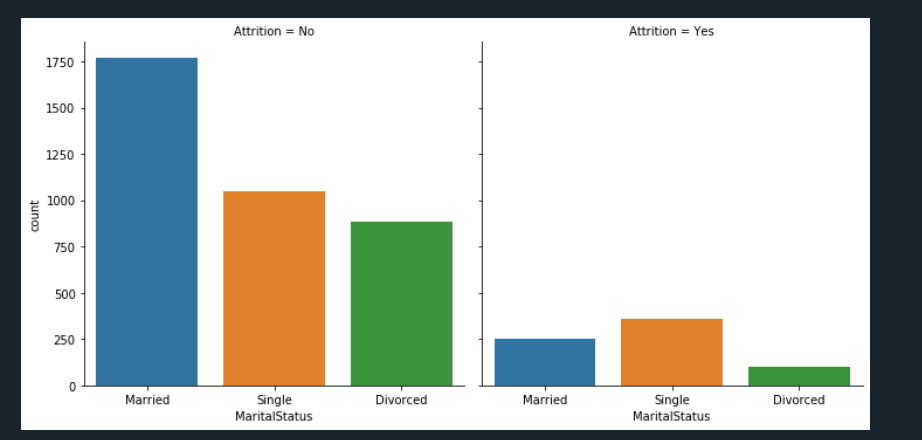


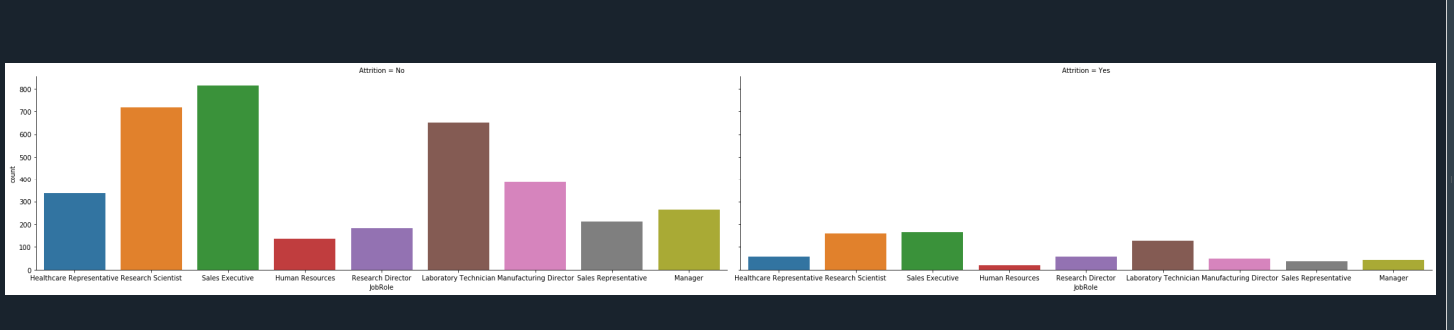
YearsWithCurrManager is also Right Skewed with several outliers observed.

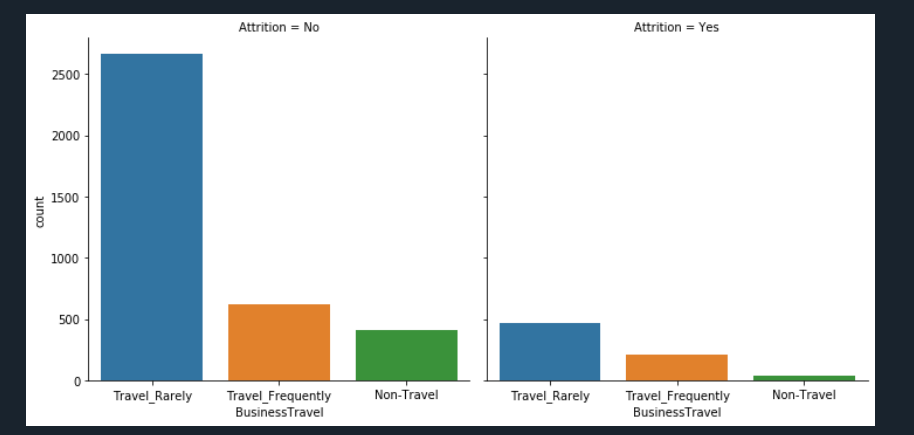
**Step 4: Visualization**



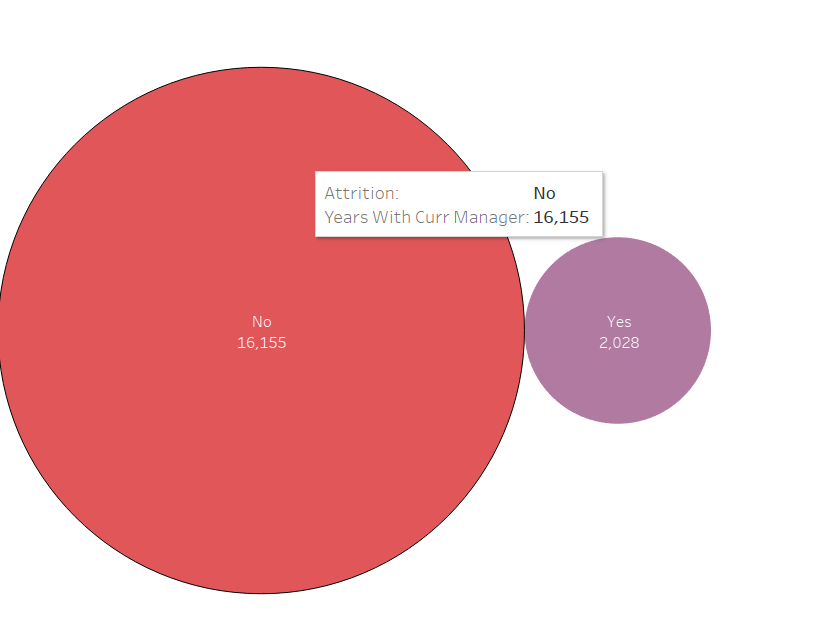




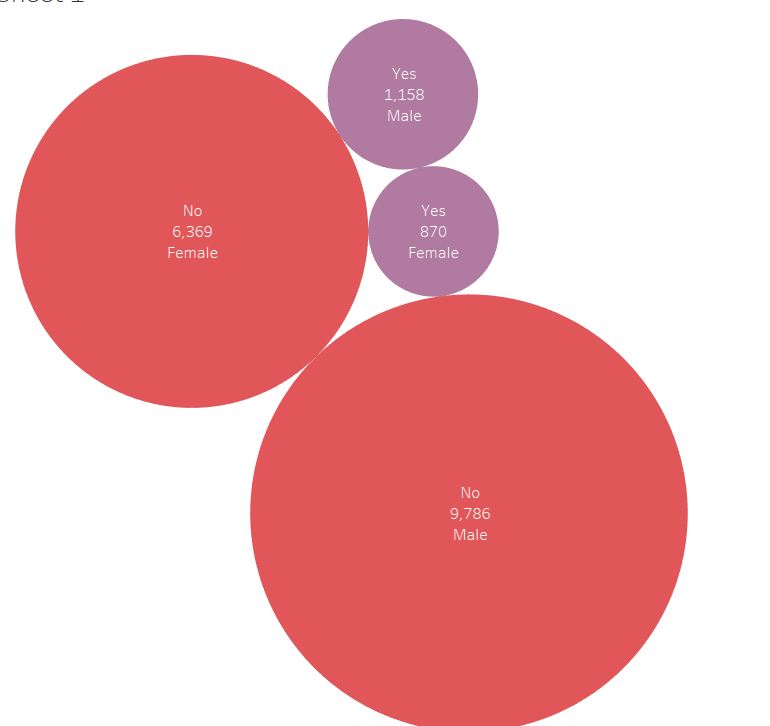




**Attrition with CurrentManager**



**Attrition with Gender**

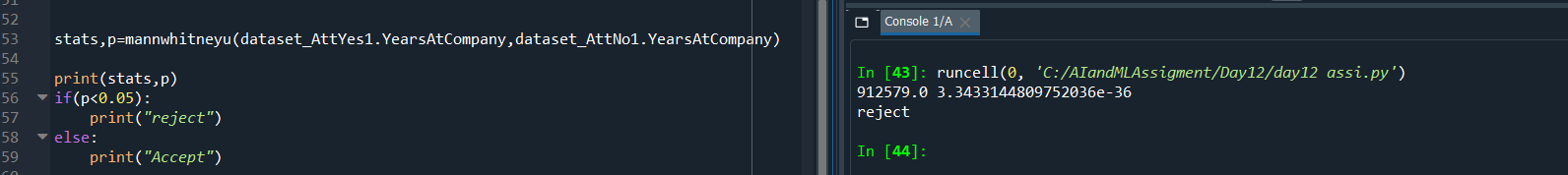


**Step 5:STATISTICAL TEST (Mann Whitney test):**

**CASE 1:**

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

**Ha:** There is significant differences in the YearsAtCompany between attrition (Y) and attrition (N)



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in YearsAtCompany** **between attrition (yes) and attrition (no)**

**CASE 2:**

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

**Ha:** There is significant differences in the Education between attrition (Y) and attrition (N)



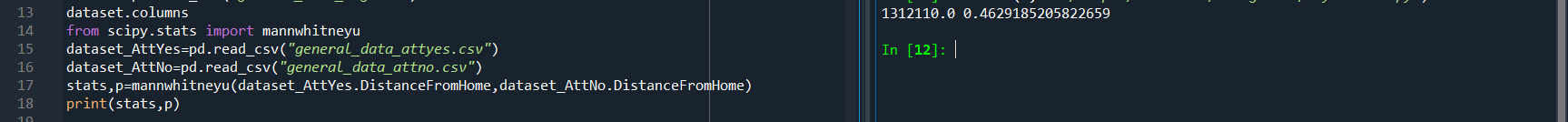
Here the **p value is greater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

**There is no significance difference in Education between attrition (yes) and attrition (no)**

**CASE 3:**

**H0:** There is no significance difference in distance from home between attrition (yes) and attrition (no)

**Ha:** There is significance difference in distance from home between attrition (yes) and attrition (no)



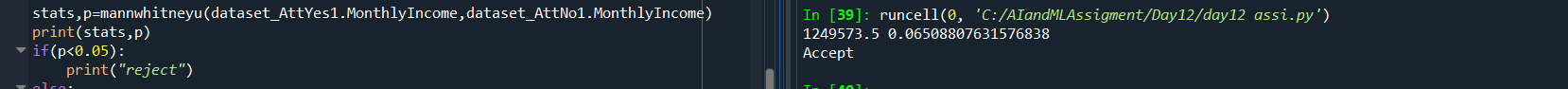
Here the **p value is greater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is**

**There is no significance difference in distance from home between attrition (yes) and attrition (no)**

**CASE 4:**

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

**Ha:** There is significant differences in the MonthlyIncome between attrition (Y) and attrition (N)



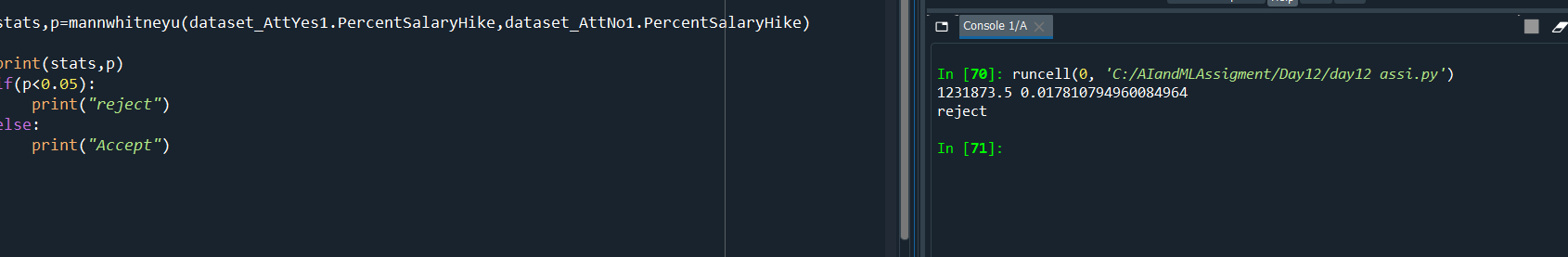
Here the **p value is greater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

**There is no significance difference in MonthlyIncome between attrition (yes) and attrition (no)**

**CASE 5:**

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

**Ha:** There is significant differences in the PercentSalaryHike between attrition (Y) and attrition (N)



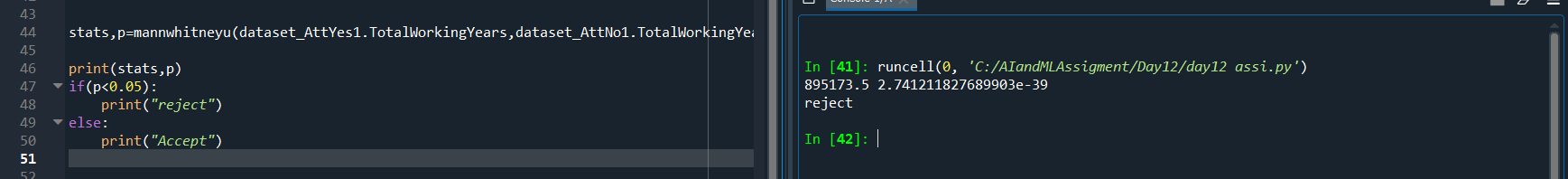
Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in PercentSalaryHike** **between attrition (yes) and attrition (no)**

**CASE 6:**

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

**Ha:** There is significant differences in the TotalWorkingYears between attrition (Y) and attrition (N)



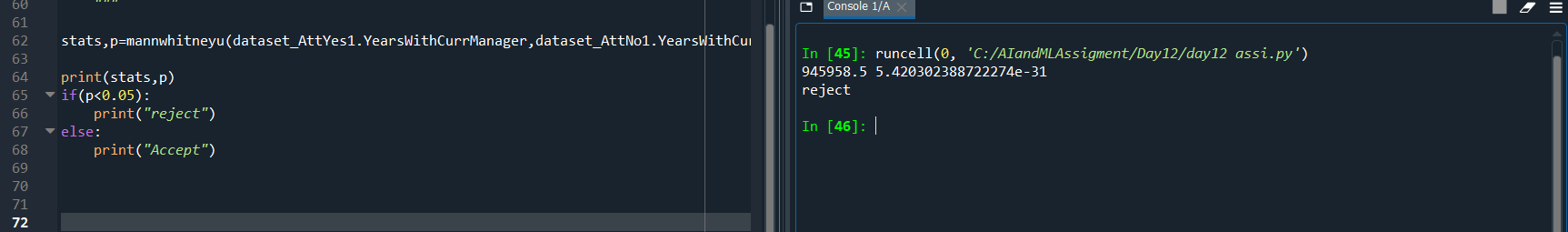
Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in TotalWorkingYears** **between attrition (yes) and attrition (no)**

**CASE 7:**

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

**Ha:** There is significant differences in the YearsWithCurrManager between attrition (Y) and attrition (N)



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

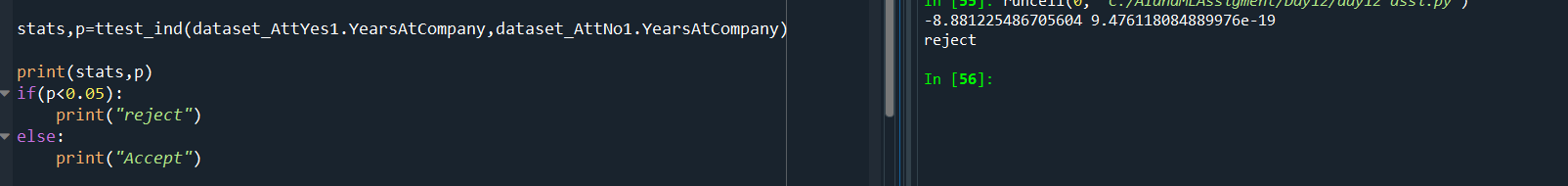
**There is significance difference in YearsWithCurrManager** **between attrition (yes) and attrition (no)**

**STEP 6: Separate T Test**

**CASE 1:**

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

**Ha:** There is significant differences in the YearsAtCompany between attrition (Y) and attrition (N)



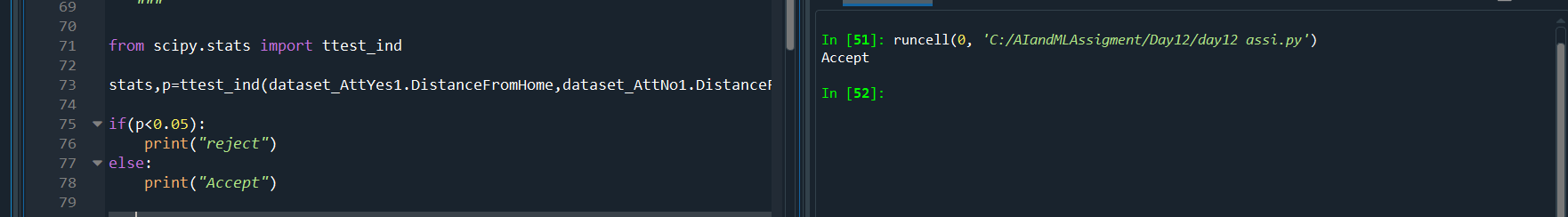
Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in YearsAtCompany** **between attrition (yes) and attrition (no)**

**CASE 2:**

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

**Ha:** There is significant differences in the DistanceFromHome between attrition (Y) and attrition (N)



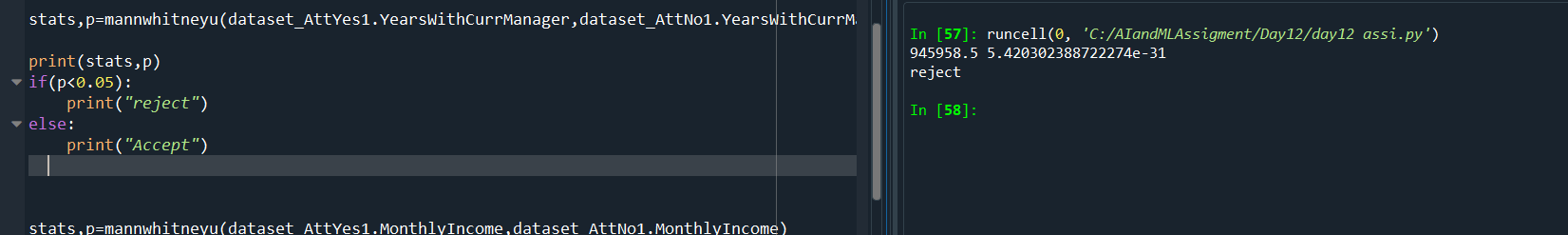
Here the **p value is greater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is**

**There is no significance difference in distance from home between attrition (yes) and attrition (no)**

**CASE 3:**

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

**Ha:** There is significant differences in the YearsWithCurrManager between attrition (Y) and attrition (N)



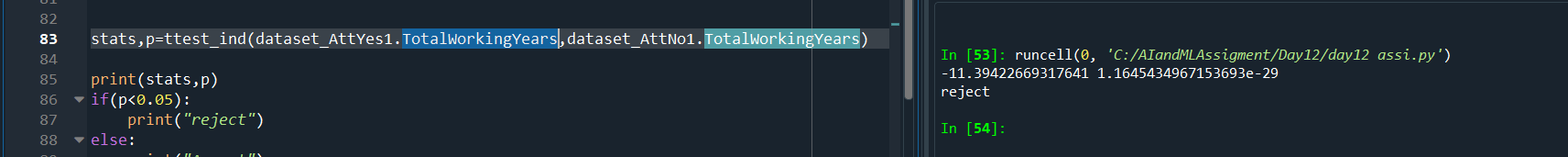
Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in YearsWithCurrManager** **between attrition (yes) and attrition (no)**

**CASE 4:**

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

**Ha:** There is significant differences in the TotalWorkingYears between attrition (Y) and attrition (N)TotalWorkingYears



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

**There is significance difference in TotalWorkingYears** **between attrition (yes) and attrition (no)**

**STEP 7: Unsupervised Learning - Correlation Analysis**

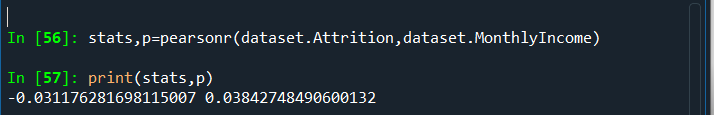
Finding Correlation

**TEST 1:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and MonthlyIncome

Ha🡪 There is significant correlation between Attrition and MonthlyIncome



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.03 there is low negative correlation between Attrition and MonthlyIncome

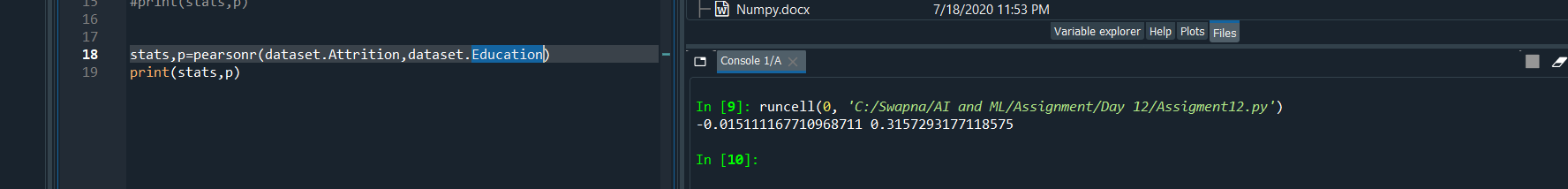
There is significant correlation between Attrition and MonthlyIncome

**TEST2:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and Education

Ha🡪 There is significant correlation between Attrition and Education



Here the **p value is grater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

R=-0.15 there is low negative correlation between Attrition and Educatio

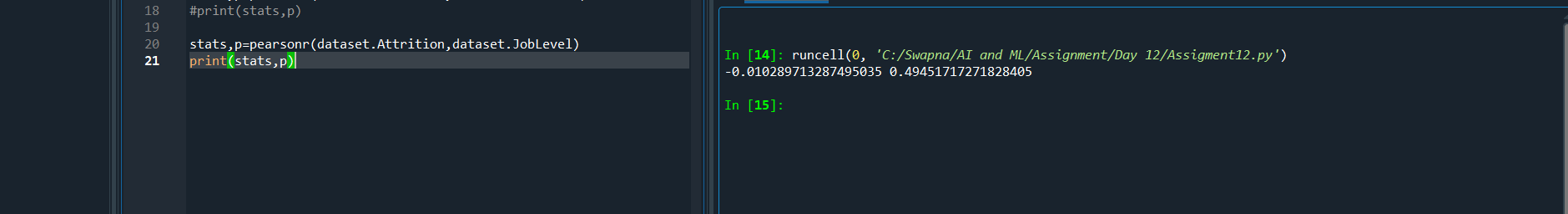
There is no significant correlation between Attrition and Education

**TEST3:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and JobLevel

Ha🡪 There is significant correlation between Attrition and JobLevel



Here the **p value is grater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

R=-0.010 there is low negative correlation between Attrition and JobLevel

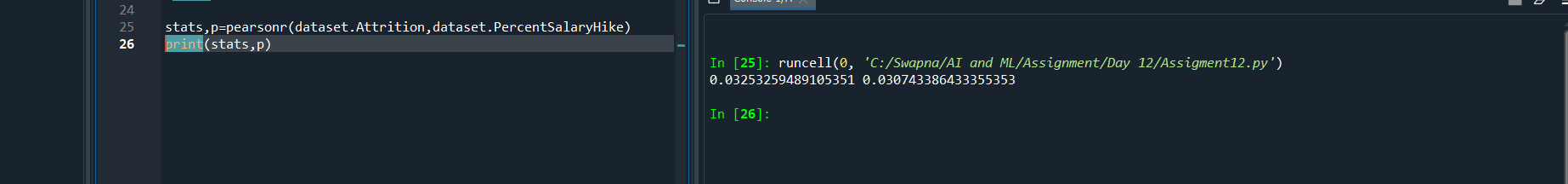
There is no significant correlation between Attrition and JobLevel

**TEST4:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and PercentSalaryHike

Ha🡪 There is significant correlation between Attrition and PercentSalaryHike



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=0.0325 there is low positive correlation between Attrition and PercentSalaryHike

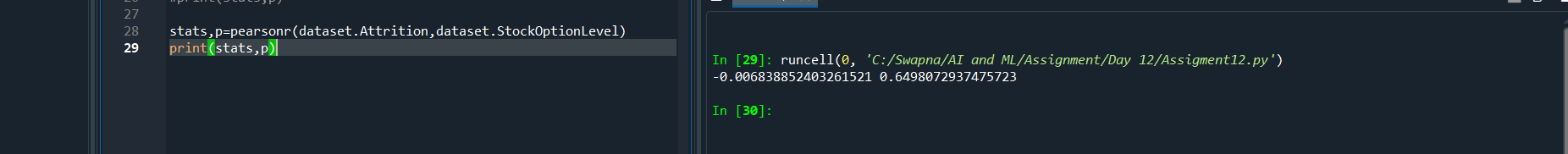
There is significant correlation between Attrition and PercentSalaryHike

**TEST5:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and StockOptionLevel

Ha🡪 There is significant correlation between Attrition and StockOptionLevel



Here the **p value is grater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

R=-0.0068 there is low negative correlation between Attrition and StockOptionLevel

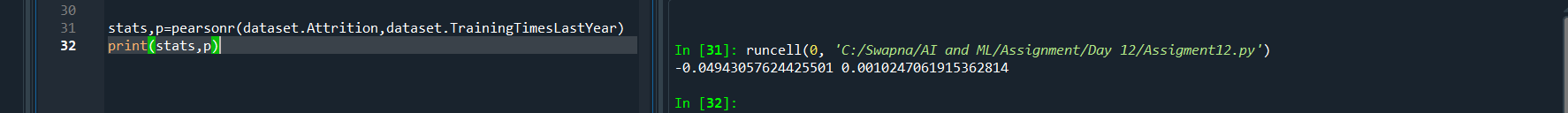
There is no significant correlation between Attrition and StockOptionLevel

**TEST6:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and TrainingTimesLastYear

Ha🡪 There is significant correlation between Attrition and TrainingTimesLastYear



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.0494 there is low negative correlation between Attrition and TrainingTimesLastYear

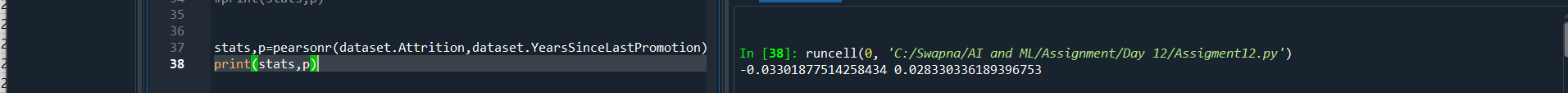
There is significant correlation between Attrition and TrainingTimesLastYear

**TEST6:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and YearsSinceLastPromotion

Ha🡪 There is significant correlation between Attrition and YearsSinceLastPromotion



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.033 there is low negative correlation between Attrition and YearsSinceLastPromotion

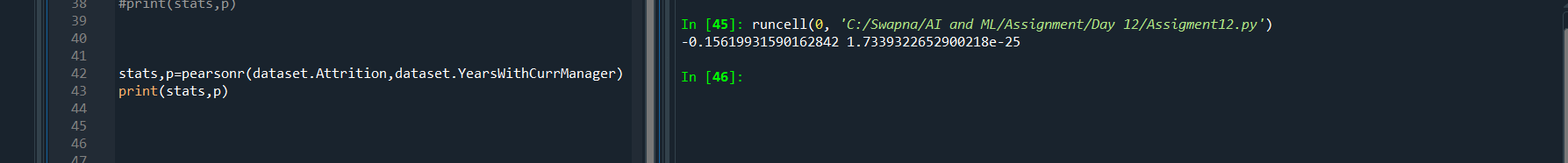
There is significant correlation between Attrition and YearsSinceLastPromotion

**TEST7:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and YearsWithCurrManager

Ha🡪 There is significant correlation between Attrition and YearsWithCurrManager



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.156 there is low negative correlation between Attrition and YearsWithCurrManager

There is significant correlation between Attrition and YearsWithCurrManager

**TEST 8:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and TotalWorkingYears

Ha🡪 There is significant correlation between Attrition and TotalWorkingYears



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.169 there is low negative correlation between Attrition and TotalWorkingYears

There is significant correlation between Attrition and TotalWorkingYears

**TEST 9:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and NumCompaniesWorked

Ha🡪 There is significant correlation between Attrition and NumCompaniesWorked



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=0.042 there is low positive correlation between Attrition and NumCompaniesWorked

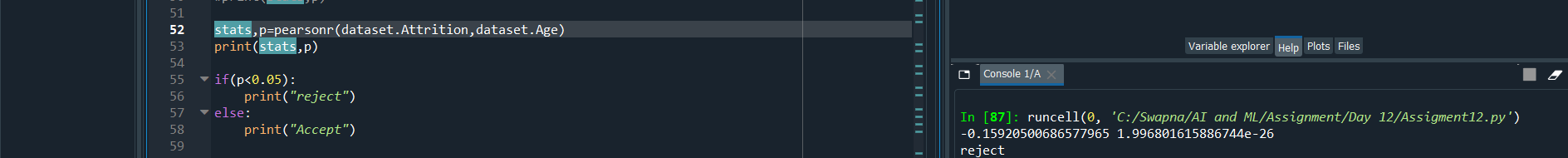
There is no significant correlation between Attrition and NumCompaniesWorked

**TEST 10:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and Age

Ha🡪 There is significant correlation between Attrition and Age



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepted**

R=-0.159 there is low -ve correlation between Attrition and Age

There is significant correlation between Attrition and Age

**TEST11:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and DistanceFromHome

Ha🡪 There is significant correlation between Attrition and DistanceFromHome



Here the **p value is greater than 0.05** so **Null Hypothesis** is **accepted** and **Alternative Hypothesis is rejected**

R=-0.009 there is low –ve correlation between Attrition and DistanceFromHome

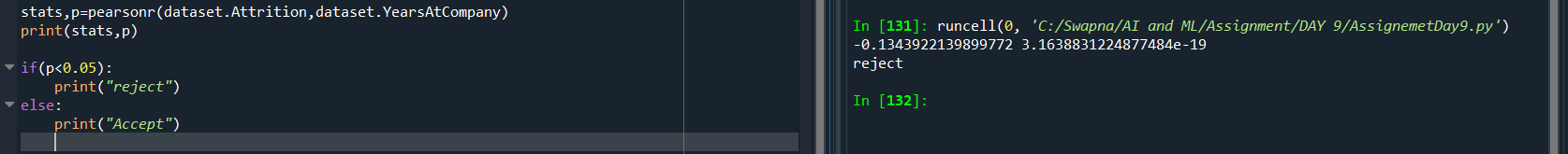
There is no significant correlation between Attrition and DistanceFromHome

**TEST12:**

Hypothesis Formulation

H0🡪There is no significant correlation between Attrition and YearsAtCompany

Ha🡪 There is significant correlation between Attrition and YearsAtCompany



Here the **p value is less than 0.05** so **Null Hypothesis** is **rejected** and **Alternative Hypothesis is accepte**

R=-0.134 there is low –ve correlation between Attrition and YearsAtCompany

There is significant correlation between Attrition and YearsAtCompany