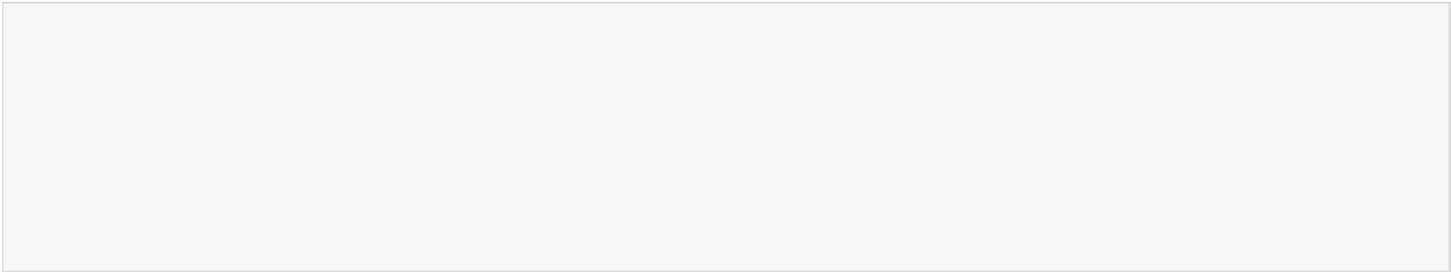
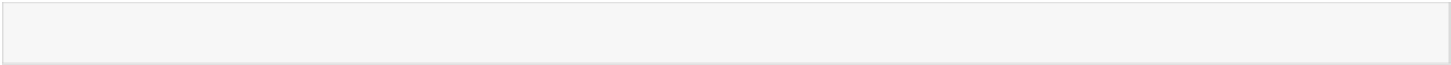
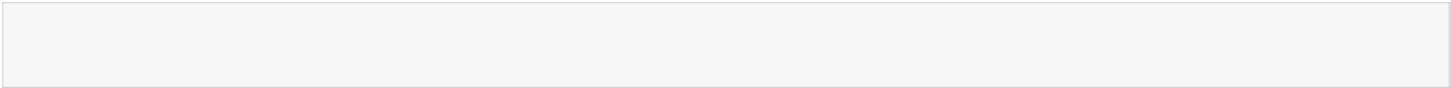
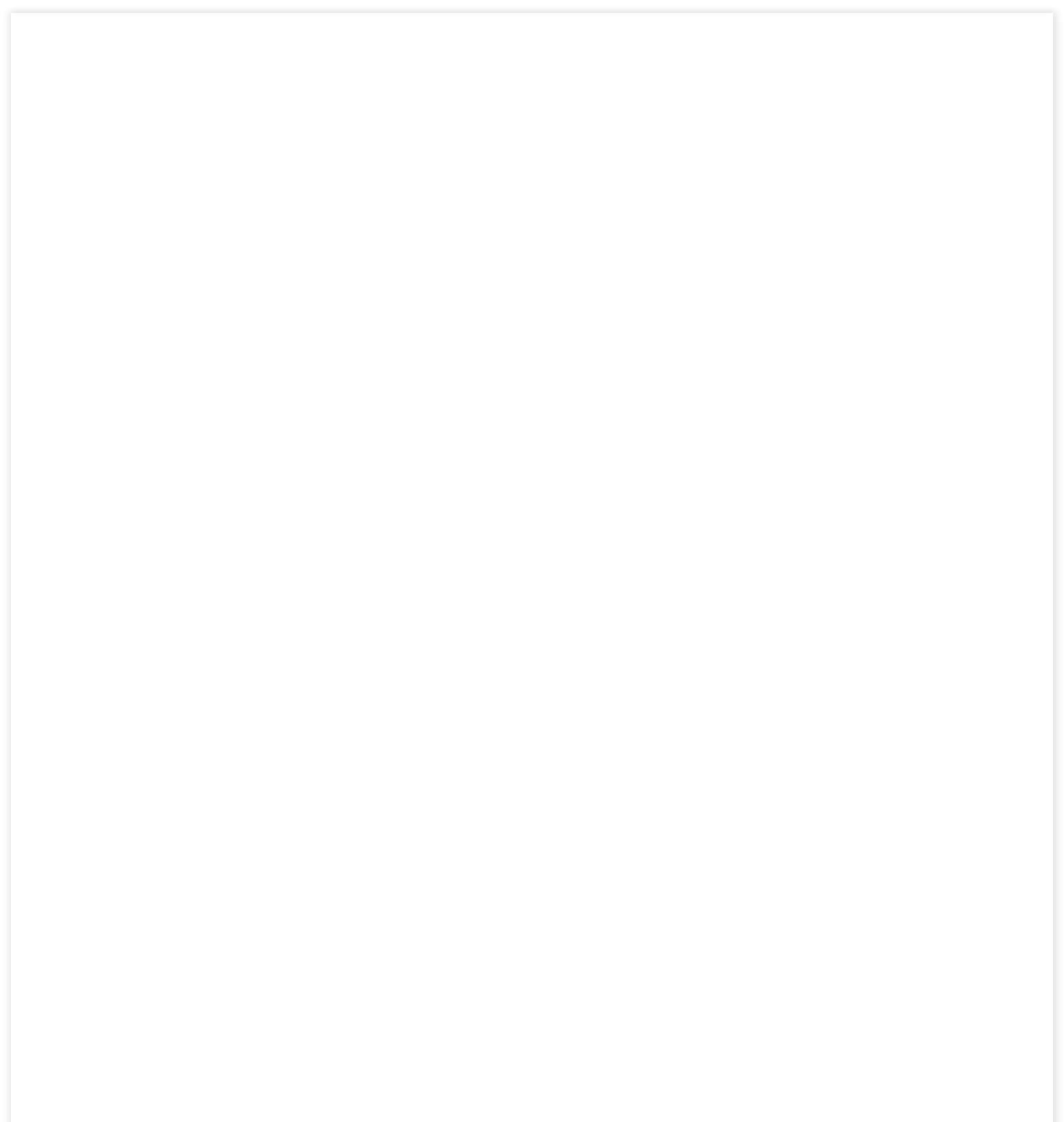
# Assignment -> Statistical Testing



In [34]:

**import pandas as pd**

In [35]:

URL = 'https://raw.githubusercontent.com/LetsUpgrade/AI-ML-July-2020/master/Day-7/Assignm ent/general\_data.csv'

data = pd.read\_csv(URL)

In [36]:

data.head(5) Out[36]:

**Age Attrition BusinessTravel Department DistanceFromHome Education EducationField EmployeeCount EmployeeI**

**0** 51 No Travel\_Rarely Sales 6 2 Life Sciences 1

1. 31 Yes Travel\_Frequently Research &

Development

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 10 | 1 | Life Sciences | 1 |  |
| 17 | 4 | Other | 1 |  |
| 2 | 5 | Life Sciences | 1 |  |
| 10 | 1 | Medical | 1 |  |

1. 32 No Travel\_Frequently Research &

Development

1. 38 No Non-Travel Research &

Development

1. 32 No Travel\_Rarely Research &

Development

In [37]:

data\_left = data[data["Attrition"]=="Yes"] data\_stay = data[data["Attrition"]=="No"]

# 1. Mann Whitney Test

## H0 : There is no significant difference in the distance from home who stay in company or who left.

**H1 : There is significant difference in the distance from home who stay in company or who left**

In [38]:

H0 = "There is no significant difference in the distance from home who stay in company or who left."

H1 = "There is significant difference in the distance from home who stay in company or wh o left."

**from scipy.stats import** mannwhitneyu

stats2,p2 = mannwhitneyu(data\_stay.DistanceFromHome,data\_left.DistanceFromHome) print("Correlation: ",stats2,"**\n** P-value: ",p2)

**if** p2<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0) Correlation: 1312110.0

P-value: 0.4629185205822659

As p>0.05, Accept null hypothesis

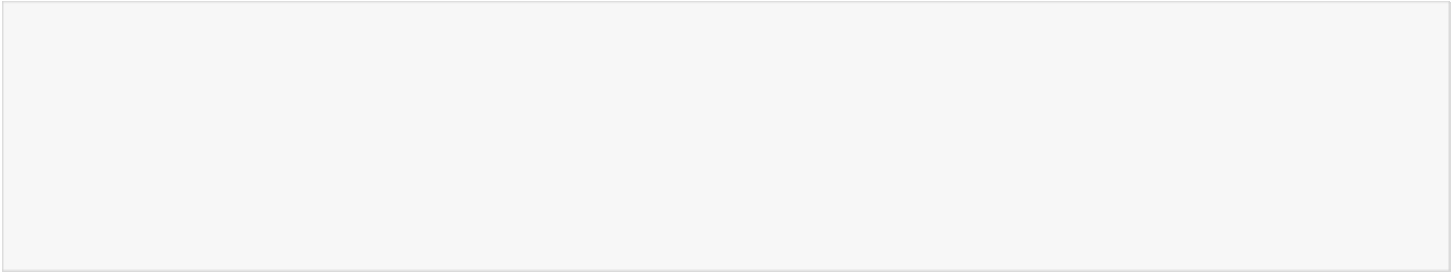
There is no significant difference in the distance from home who stay in company or who left.

# 1 sample t-test

## H0 : There is no significant difference of the sample mean of MonthlyIncome and population mean .

**H1 : There is significant difference of the sample mean of MonthlyIncome and population mean .**

In [39]:



H0 = "There is no significant difference of the sample mean of MonthlyIncome and populati on mean "

H1 = "There is significant difference of the sample mean of MonthlyIncome and population mean "

**from scipy.stats import** ttest\_1samp

stats,p = ttest\_1samp(data.MonthlyIncome , 50000) print("Correlation: ",stats,"**\n** P-value: ",p)

**if** p<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 21.204326281653234 P-value: 4.2162150272572684e-95

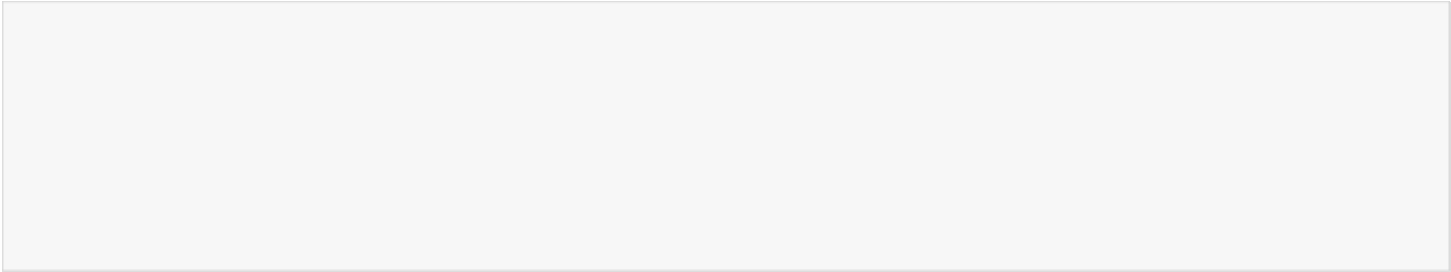
As p<0.05, Reject null hypothesis

There is significant difference of the sample mean of MonthlyIncome and population mean

## H0 : There is no significant difference of the sample mean of PercentSalaryHike and population mean .

**H1 : There is significant difference of the sample mean of PercentSalaryHike and population mean .**

In [40]:



H0 = "There is no significant difference of the sample mean of PercentSalaryHike and popu lation mean "

H1 = "There is significant difference of the sample mean of PercentSalaryHike and populat ion mean "

**from scipy.stats import** ttest\_1samp

stats,p = ttest\_1samp(data.PercentSalaryHike , 14) print("Correlation: ",stats,"**\n** P-value: ",p)

**if** p<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 21.95121411724244

P-value: 1.9094583042386026e-101

As p<0.05, Reject null hypothesis

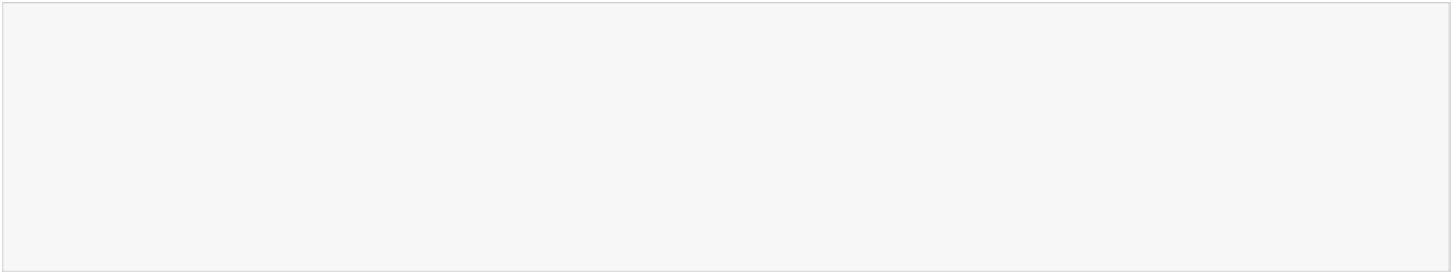
There is significant difference of the sample mean of PercentSalaryHike and population m ean

# 2 sample paired t-test

## H0 : There is no significant difference of the sample mean of YearsAtCompany and MonthlyIncome.

**H1 : There is significant difference of the sample mean of YearsAtCompany and MonthlyIncome**

In [41]:



H0 = "There is no significant difference of the sample mean of Age and NumCompaniesWorked who left the company."

H1 = "There is significant difference of the sample mean of Age and NumCompaniesWorked wh o left the company."

**from scipy.stats import** ttest\_rel

stats,p = ttest\_rel(data.YearsAtCompany,data.MonthlyIncome) print("Correlation: ",stats,"**\n** P-value: ",p)

**if** p<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: -91.73768218543985

P-value: 0.0

As p<0.05, Reject null hypothesis

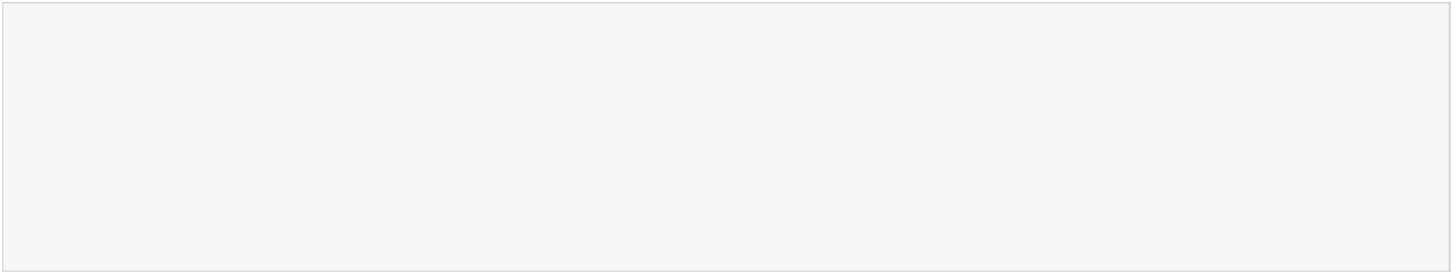
There is significant difference of the sample mean of Age and NumCompaniesWorked who lef t the company.

# 2 sample independent t-test

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

**H1: There is significant differences in the Monthly Income between attrition (Y) and attirition (N)**

In [42]:



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

H1 = "There is significant differences in the Monthly Income between attrition (Y) and at tirition (N)"

**from scipy.stats import** ttest\_ind

stats3,p3 = ttest\_ind(data\_stay.MonthlyIncome,data\_left.MonthlyIncome) print("Correlation: ",stats3,"**\n** P-value: ",p3)

**if** p3<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 2.0708863763619316

P-value: 0.03842748490605113

As p<0.05, Reject null hypothesis

There is significant differences in the Monthly Income between attrition (Y) and attirit ion (N)

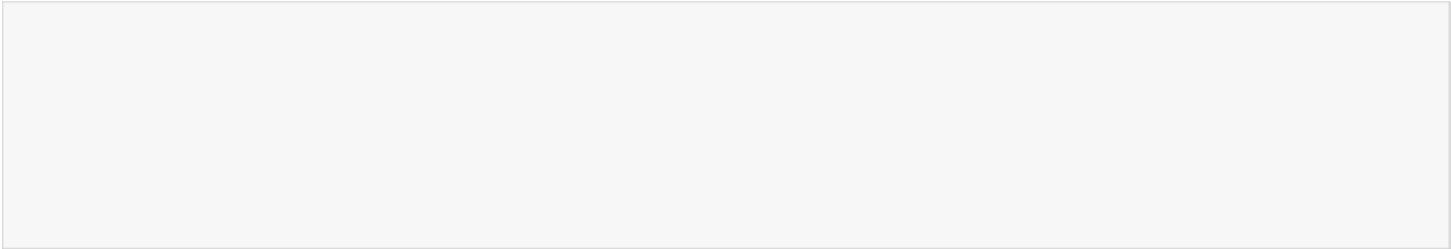
# Wilcoxon test



## H0 : There is no significant difference in the total working years and years at company.

**H1 : There is significant difference in the total working years and years at company.**

In [43]:



H0 = "There is no significant difference in the total working years and years at company. "

H1 = "There is significant difference in the total working years and years at company." **from scipy.stats import** wilcoxon

stats4,p4 = wilcoxon(data.TrainingTimesLastYear,data.DistanceFromHome) print("Correlation: ",stats4,"**\n** P-value: ",p4)

**if** p4<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 822562.5

P-value: 0.0

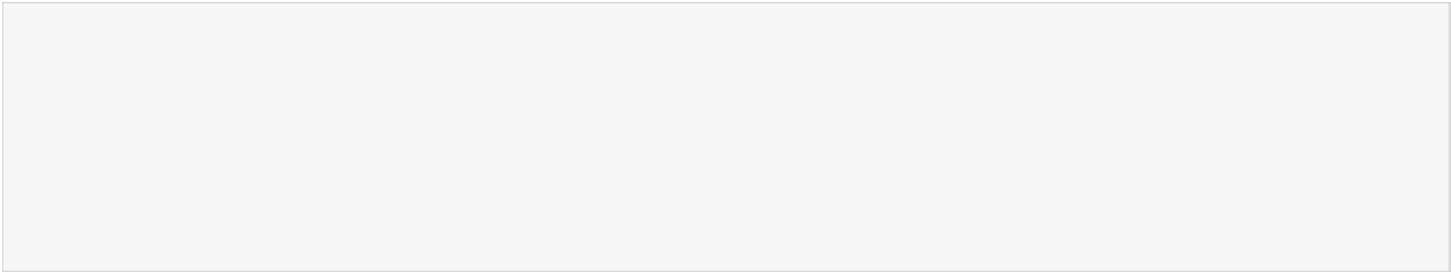
As p<0.05, Reject null hypothesis

There is significant difference in the total working years and years at company.

## H0 : There is no significant difference in the YearsSinceLastPromotion and YearsWithCurrManager.

**H1 : There is significant difference in the YearsSinceLastPromotion and YearsWithCurrManager.**

In [44]:



H0 = "There is no significant difference in the YearsSinceLastPromotion and YearsWithCurr Manager"

H1 = "There is significant difference in the YearsSinceLastPromotion and YearsWithCurrMan ager"

**from scipy.stats import** wilcoxon

stats,p = wilcoxon(data.YearsSinceLastPromotion,data.YearsWithCurrManager) print("Correlation: ",stats,"**\n** P-value: ",p)

**if** p<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 699082.5

P-value: 7.578916343856332e-279

As p<0.05, Reject null hypothesis

There is significant difference in the YearsSinceLastPromotion and YearsWithCurrManager

# 6. Friedman test

## 8. H0 : There is no significant difference in the total working years , Yerars since last promotion and years at company.

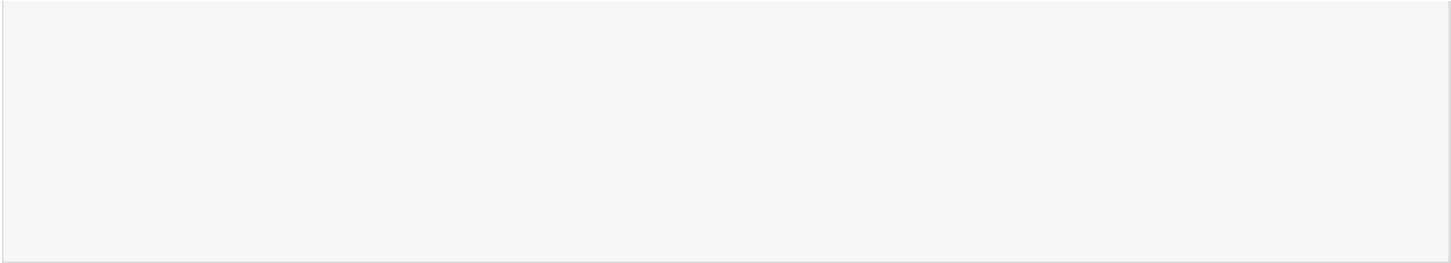
**H1 : There is significant difference in the total working years , Yerars since last promotion and years at company.**

In [45]:



H0 = "There is no significant difference in the total working years , Yerars since last p





romotion and years at company."

H1 = "There is significant difference in the total working years , Yerars since last prom otion and years at company."

**from scipy.stats import** friedmanchisquare

stats5,p5 = friedmanchisquare(data.TotalWorkingYears,data.YearsAtCompany,data.YearsSince LastPromotion)

print("Correlation: ",stats5,"**\n** P-value: ",p5) **if** p5<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 7788.320648298225

P-value: 0.0

As p<0.05, Reject null hypothesis

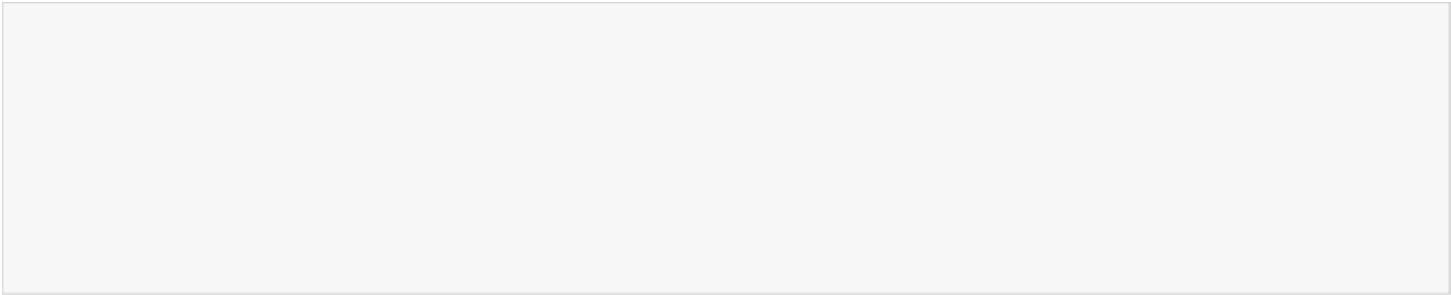
There is significant difference in the total working years , Yerars since last promotion and years at company.

# 7. Chi Square test

## H0 : There is no dependency between gender and business travel.

**H1 : There is dependency between gender and business travel.**

In [46]:



H0 = "There is no dependency between gender and business travel." H1 = "There is dependency between gender and business travel." data1 = data.dropna()

**from scipy.stats import** chi2\_contingency

chitable = pd.crosstab(data1.Gender,data1.BusinessTravel) print(chitable)

stats6,p6,dof,expected = chi2\_contingency(chitable) print("**\n**Correlation: ",stats6,"**\n** P-value: ",p6) **if** p6<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

BusinessTravel Non-Travel Travel\_Frequently Travel\_Rarely Gender

|  |  |  |  |
| --- | --- | --- | --- |
| Female | 153 | 328 | 1275 |
| Male | 295 | 497 | 1834 |

Correlation: 7.7114689255919355

P-value: 0.021158057795728463

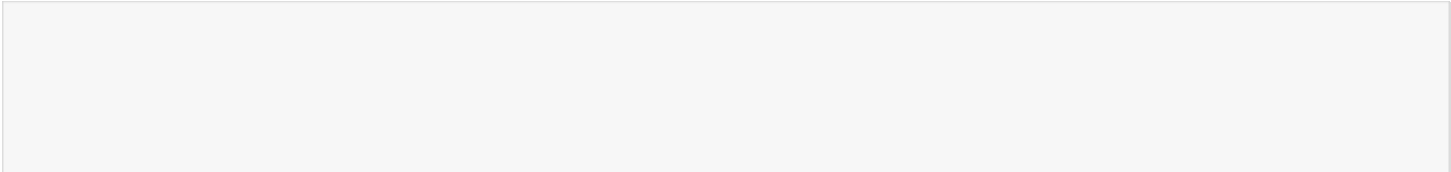
As p<0.05, Reject null hypothesis

There is dependency between gender and business travel.

## H0 : There is no dependency between maritial status and joblevel.

**H1 : There is dependency between maritial status and joblevel.**

In [47]:



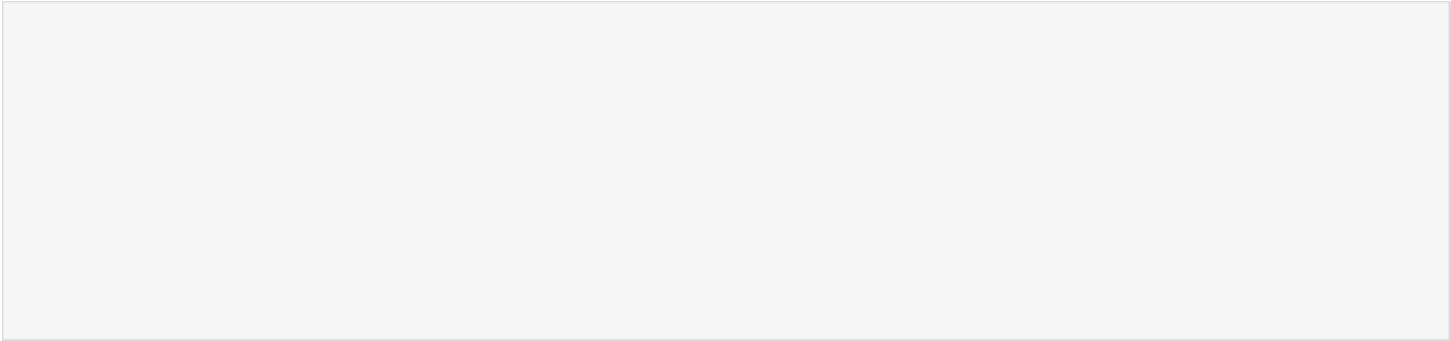
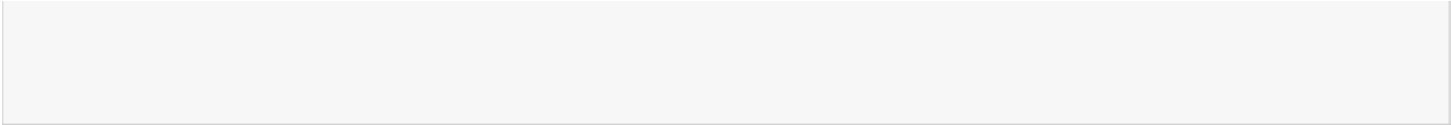
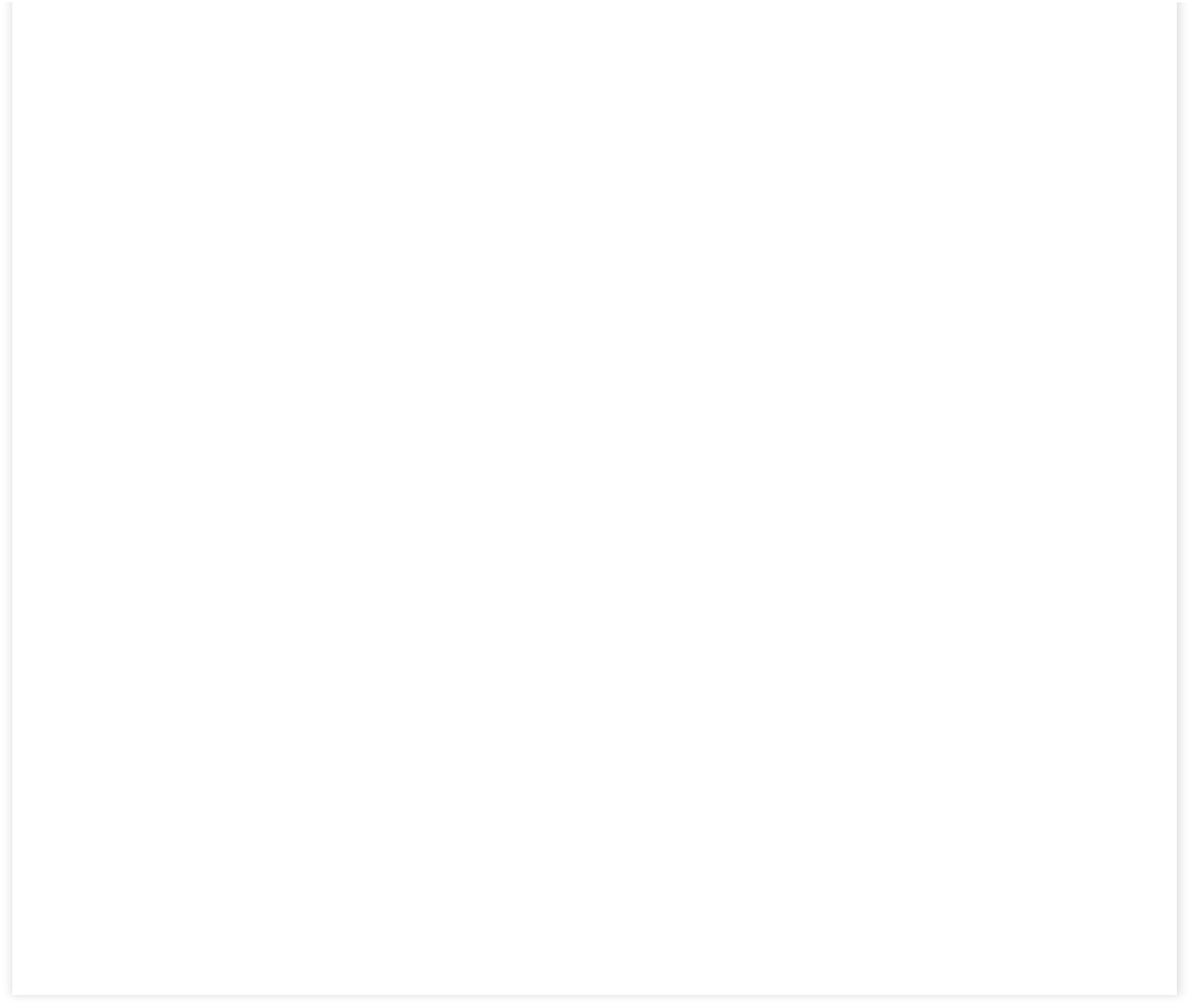
H0 = "There is no dependency between maritial status and joblevel." H1 = "There is dependency between maritial status and joblevel." data2 = data.dropna()

**from scipy.stats import** chi2\_contingency

chitable1 = pd.crosstab(data2.MaritalStatus,data2.JobLevel) print(chitable1)

stats7,p7,dof,expected = chi2\_contingency(chitable1)

print("**\n**Correlation: ",stats7,"**\n** P-value: ",p7) **if** p7<0.05:



print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| JobLevel | 1 | 2 | 3 | 4 | 5 |
| MaritalStatus |  |  |  |  |  |
| Divorced | 353 | 340 | 152 | 81 | 44 |
| Married | 734 | 713 | 320 | 156 | 84 |
| Single | 532 | 537 | 179 | 81 | 76 |

Correlation: 18.26067228007838

P-value: 0.019354877469456792

As p<0.05, Reject null hypothesis

There is dependency between maritial status and joblevel.

# 8. Krushkal wallis test

## 11. H0 : There is no significant difference years at company where business travel is travel frequently, travel rarely and non travel.

**H1 : There is significant difference years at company where business travel is travel frequently, travel rarely and non travel.**

In [48]:

H0 = "There is no significant difference years at company where business travel is travel frequently, travel rarely and non travel."

H1 = "There is significant difference years at company where business travel is travel fr equently, travel rarely and non travel."

t1 = data[data["BusinessTravel"] == 'Travel\_Rarely']

t2 = data[data["BusinessTravel"] == 'Travel\_Frequently'] t3 = data[data["BusinessTravel"] == 'Non-Travel']

**from scipy.stats import** kruskal

stats8,p8 = kruskal(t1.YearsAtCompany, t2.YearsAtCompany, t3.YearsAtCompany) print("**\n**Correlation: ",stats8,"**\n** P-value: ",p8)

**if** p8<0.05:

print("**\n\n** As p<0.05, Reject null hypothesis**\n**",H1) **else**:

print("**\n\n** As p>0.05, Accept null hypothesis**\n**",H0)

Correlation: 2.187341926985394

P-value: 0.33498451788533107

As p>0.05, Accept null hypothesis

There is no significant difference years at company where business travel is travel freq uently, travel rarely and non travel.