**Statistical testing**

**import packages:**

import pandas as pd

import matplotlib.pyplot as plt

**#import data:**

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

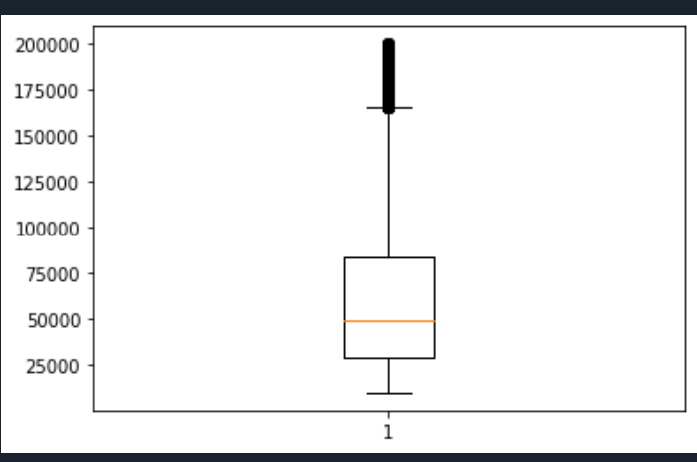
#check null and duplicated values are existing.

dataset.isnull()

da=dataset.dropna()

**Hypothesis 1:**

plt.boxplot(da.MonthlyIncome)



#Here monthlyincome is not normoly distibuted so we have to use non-parametric test.

#mannwhitneyu test

#H0=there is no significance between Attrition of yes persons monthlyincome and Attrition of no persons monthlyincome.

#H1=there is significance between Attrition of yes persons monthlyincome and Attrition of no persons monthlyincome.

from scipy.stats import mannwhitneyu

stats,p= mannwhitneyu(da[da['Attrition']=='Yes'].MonthlyIncome,da[da['Attrition']=='No'].MonthlyIncome)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

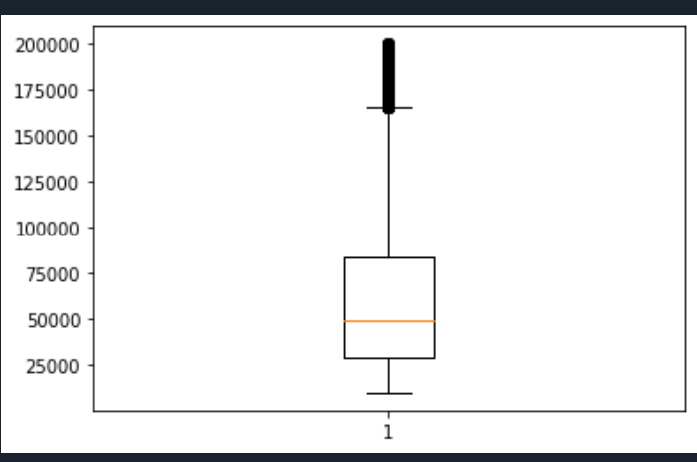
print("Reject the hypothesis")

1249573.5 0.06508807631576838

Accept the hypothesis

**Hypothesis 2:**

plt.boxplot(da.MonthlyIncome)



#here Monthlyincome is not normoly distibuted so we have to use non-parametric test.

#kruskal test

#H0=there is no significance between Department of Human Resources persons MonthlyIncome and Department of Research & Development persons MonthlyIncome

#Department of Sales persons MonthlyIncome

#H1!=there is significance between Department of Human Resources persons MonthlyIncome and Department of Research & Development persons MonthlyIncome

#Department of Sales persons MonthlyIncome.

pd.crosstab(da.Department,columns="count")

from scipy.stats import kruskal

stats,p=kruskal(da[da['Department']=='Human Resources' ].MonthlyIncome,da[da['Department']=='Research & Development'].MonthlyIncome,da[da['Department']=='Sales'].MonthlyIncome)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

print("Reject the hypothesis")

13.753271133838242 0.0010316089925355575

Reject the hypothesis

**Hypothesis 3:**

#chi square test

#H0=there is no dependency between Attrition and Department

#H1= there is dependency between Attrition and department.

from scipy.stats import chi2\_contingency

chitab=pd.crosstab(da.Attrition,da.Department)

chitab

stats,p,dof,excepted=chi2\_contingency(chitab)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

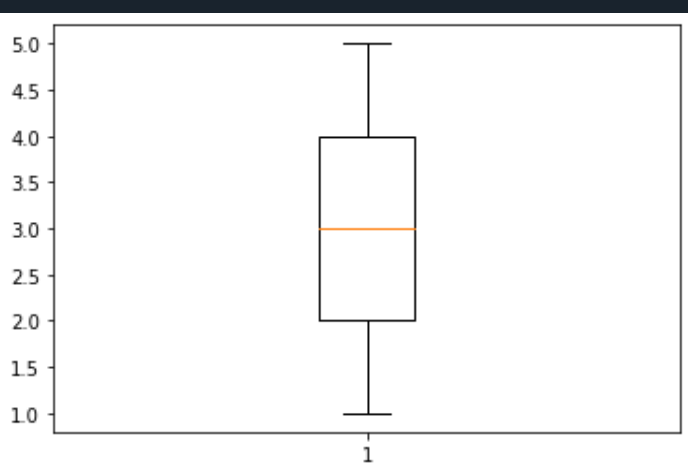
print("Reject the hypothesis")

25.89432541916022 2.382970570769315e-06

Reject the hypothesis

**Hypothesis 4:**

plt.boxplot(da.Education)



# here education is normaly distributed so we have to use parametric.

#one sample test

#H0=there is no significance between population mean of Education and sample mean of Education.

#H1=there is significance between population mean of Education and sample mean of Education.

from scipy.stats import ttest\_1samp

stats,p=ttest\_1samp(da.Education, 3 )

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

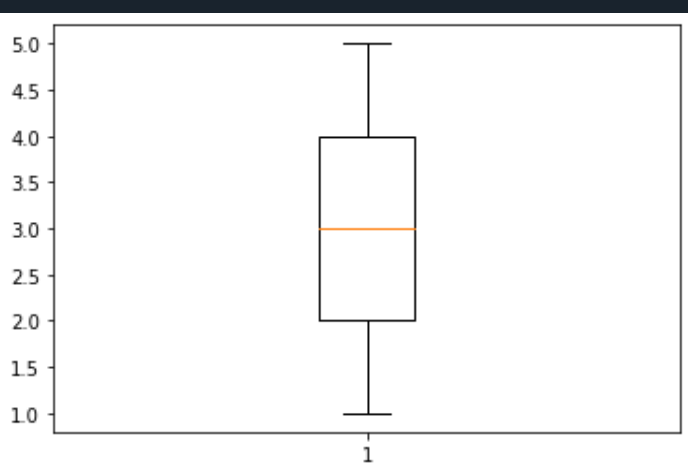
print("Reject the hypothesis")

-5.660909935667991 1.6018998035558455e-08

Reject the hypothesis

**Hypothesis 5:**

plt.boxplot(da.Education)



#Two sample independent separete t test.

#H0=there is no significance between Attrition of yes persons education mean and Attrition of no persons age mean.

#H1=there is significance between Attrition of yes persons education mean and Attrition of no persons education mean.

from scipy.stats import ttest\_ind

stats,p=ttest\_ind(da[da['Attrition']=='Yes'].Education,da[da['Attrition']=='No'].Education)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

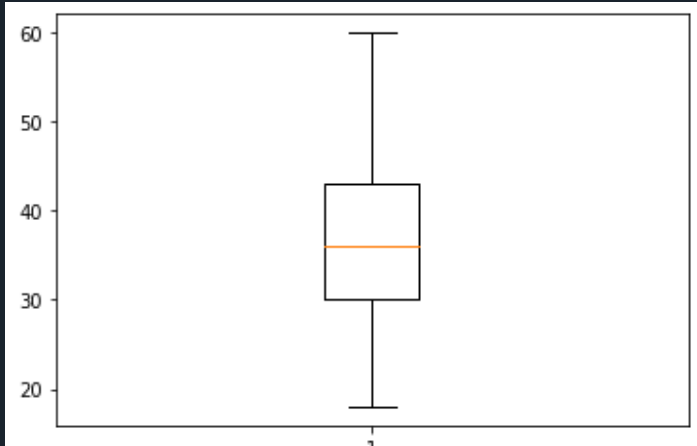
print("Reject the hypothesis")

-1.132287911067296 0.25757539308301036

Accept the hypothesis

**Hypothesis 6:**

plt.boxplot(da.Age)



# here age is normaly distributed so we have to use parametric.

#one sample test

#H0=there is no significance between population mean of age and sample mean of age.

#H1=there is significance between population mean of age and sample mean of age.

from scipy.stats import ttest\_1samp

stats,p=ttest\_1samp(da.Age, 40 )

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

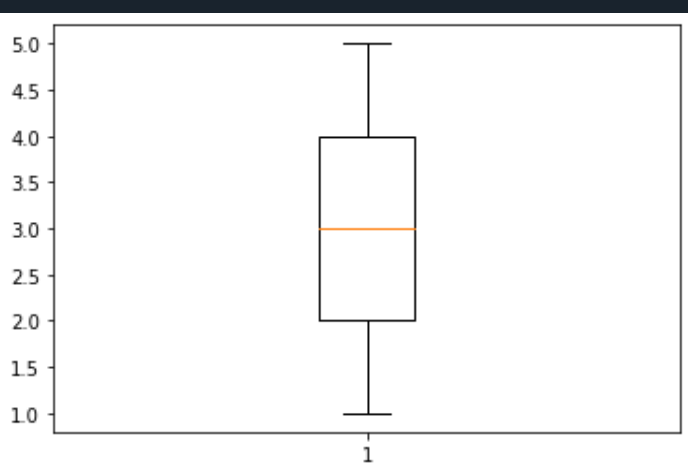
print("Reject the hypothesis")

-22.216820236363763 1.0373472954859772e-103

Reject the hypothesis

**Hypothesis 7:**

plt.boxplot(da.Education)



#here education is normoly distributed so we have to use parametric.

#Two sample independent separete t test.

#H0=there is no significance between gender "male" of education mean and gender "Female" of education .

#H1=there is significance between gender "male" of education and gender "Female" of education .

from scipy.stats import ttest\_ind

stats,p=ttest\_ind(da[da['Gender']=='Male'].Education,da[da['Gender']=='Female'].Education)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

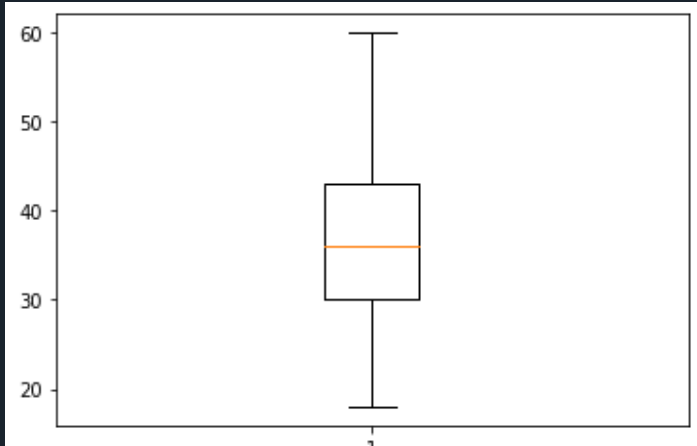
print("Reject the hypothesis")

-1.1696797492489017 0.24219351573994324

Accept the hypothesis

**Hypothesis 8:**

plt.boxplot(da.Age)



#here education is normoly distributed so we have to use parametric.

#Two sample independent separete t test.

#H0=there is no significance between gender "male" of education mean and gender "Female" of education .

#H1=there is significance between gender "male" of education and gender "Female" of education .

from scipy.stats import ttest\_ind

stats,p=ttest\_ind(da[da['Gender']=='Male'].Age,da[da['Gender']=='Female'].Age)

print(stats,p)

if p>=0.05:

print("Accept the hypothesis")

if p<0.05:

print("Reject the hypothesis")

-2.6908200170040044 0.0071547253790398445

Reject the hypothesis