**Assignment No 2**

**Code:**

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read\_csv('Students data.csv')

print("===========IsNa=============")

print(df.isna().sum())

print("===========IsNull===========")

print(df.isnull().sum())

print("===========NotNull===========")

print(df.notnull().sum())

print("============FillNa============")

df['class'].fillna('A',inplace=True)

print(df)

print("===========Replace============")

df['class'].replace(['A','B'],[0,1],inplace=True)

print(df)

print("==========Rename==================")

df.rename(columns={'class':'Div'},inplace=True)

print(df)

print("===========DropNa============")

df.dropna(how='all',inplace=True)

print(df)

print("===========Z-score===========")

race=df['race']

#print(race)

mean=np.mean(race)

std=np.std(race)

print("Mean:",mean)

print("Standard deviation:",std)

threshold=3

outlier=[]

for i in race:

    z=(i-mean)/std

    if z>threshold:

        outlier.append(i)

print("Outlier :",outlier)

print("==========IQR==========")

#print(race)

algebra=df['Statistics']

#Nrace=sorted(race)

#print(Nrace)

q1,q3=np.percentile(algebra,[25,75])

print("Q1,Q3:",q1,q3)

iqr=q3-q1

print("IQR:",iqr)

lower\_fence=q1-(1.5\*iqr)

upper\_fence=q3+(1.5\*iqr)

print("Lower fence,upper\_fence:",lower\_fence,upper\_fence)

outlier=[]

for x in algebra:

    if((x>upper\_fence)or(x<lower\_fence)):

        outlier.append(x)

print('Outliner in the dataset is',outlier)

fig=plt.figure(figsize=(10,7))

plt.boxplot(df['Statistics'])

plt.show()

ua=np.where(df['Statistics']>=upper\_fence)[0]

la=np.where(df['Statistics']<=lower\_fence)[0]

df.drop(index=ua,inplace=True)

df.drop(index=la,inplace=True)

print("\*\*\*\*\*\*\*\*\*\*\*After removing outliner\*\*\*\*\*\*\*\*\*\*")

print(df['Statistics'])

fig=plt.figure(figsize=(10,7))

plt.boxplot(df['Statistics'])

plt.show()

print("\*\*\*\*\*\*\*\*\*\*Data Transformation\*\*\*\*\*\*\*\*\*\*")

df['Log\_attendance']=np.log(df['Statistics'])

print('\*\*Display dataset after data transformation\*\*')

print(df)

**Output:**

===========IsNa=============

ID                     0

class                  0

gender                 0

race                   0

GPA                    0

Algebra                0

Calculus1              0

Calculus2              0

Statistics             0

Probability            0

Measure                0

Functional\_analysis    0

from1                  0

from2                  0

from3                  0

from4                  0

y                      0

dtype: int64

===========IsNull===========

ID                     0

class                  0

gender                 0

race                   0

GPA                    0

Algebra                0

Calculus1              0

Calculus2              0

Statistics             0

Probability            0

Measure                0

Functional\_analysis    0

from1                  0

from2                  0

from3                  0

from4                  0

y                      0

dtype: int64

===========NotNull===========

ID                     105

class                  105

gender                 105

race                   105

GPA                    105

Algebra                105

Calculus1              105

Calculus2              105

Statistics             105

Probability            105

Measure                105

Functional\_analysis    105

from1                  105

from2                  105

from3                  105

from4                  105

y                      105

dtype: int64

============FillNa============

       ID class  gender  race    GPA  ...  from1  from2  from3  from4  y

0    1141     A    male     1  73.47  ...      A      A      A      3  0

1    1142     A  female     1  71.22  ...      B      A      A      2  0

2    1143     A  female     2  74.56  ...      C      A      A      0  1

3    1144     A  female     1  72.89  ...      D      A      A      0  0

4    1145     A  female     1  70.11  ...      E      A      A      0  0

..    ...   ...     ...   ...    ...  ...    ...    ...    ...    ... ..

100  1241     A  female     1  88.34  ...      M      B      A      0  1

101  1242     B    male     1  89.84  ...      A      B      A      0  1

102  1243     B    male     1  88.82  ...      T      B      A      0  2

103  1244     A    male     1  86.60  ...      S      B      A      0  2

104  1245     A    male     1  93.71  ...      K      B      A      0  2

[105 rows x 17 columns]

===========Replace============

c:\Users\Himanshu Lonkar\OneDrive\Documents\Academics\TE\DSBDA\A2\A2.py:22: FutureWarning: Downcasting behavior in `replace` is deprecated and will be removed in a future version. To retain the old behavior, explicitly call `result.infer\_objects(copy=False)`. To opt-in to the future behavior, set `pd.set\_option('future.no\_silent\_downcasting', True)`

  df['class'].replace(['A','B'],[0,1])

       ID class  gender  race    GPA  ...  from1  from2  from3  from4  y

0    1141     A    male     1  73.47  ...      A      A      A      3  0

1    1142     A  female     1  71.22  ...      B      A      A      2  0

2    1143     A  female     2  74.56  ...      C      A      A      0  1

3    1144     A  female     1  72.89  ...      D      A      A      0  0

4    1145     A  female     1  70.11  ...      E      A      A      0  0

..    ...   ...     ...   ...    ...  ...    ...    ...    ...    ... ..

100  1241     A  female     1  88.34  ...      M      B      A      0  1

101  1242     B    male     1  89.84  ...      A      B      A      0  1

102  1243     B    male     1  88.82  ...      T      B      A      0  2

103  1244     A    male     1  86.60  ...      S      B      A      0  2

104  1245     A    male     1  93.71  ...      K      B      A      0  2

[105 rows x 17 columns]

==========Rename==================

       ID class  gender  race    GPA  ...  from1  from2  from3  from4  y

0    1141     A    male     1  73.47  ...      A      A      A      3  0

1    1142     A  female     1  71.22  ...      B      A      A      2  0

2    1143     A  female     2  74.56  ...      C      A      A      0  1

3    1144     A  female     1  72.89  ...      D      A      A      0  0

4    1145     A  female     1  70.11  ...      E      A      A      0  0

..    ...   ...     ...   ...    ...  ...    ...    ...    ...    ... ..

100  1241     A  female     1  88.34  ...      M      B      A      0  1

101  1242     B    male     1  89.84  ...      A      B      A      0  1

102  1243     B    male     1  88.82  ...      T      B      A      0  2

103  1244     A    male     1  86.60  ...      S      B      A      0  2

104  1245     A    male     1  93.71  ...      K      B      A      0  2

[105 rows x 17 columns]

===========DropNa============

       ID class  gender  race    GPA  ...  from1  from2  from3  from4  y

0    1141     A    male     1  73.47  ...      A      A      A      3  0

1    1142     A  female     1  71.22  ...      B      A      A      2  0

2    1143     A  female     2  74.56  ...      C      A      A      0  1

3    1144     A  female     1  72.89  ...      D      A      A      0  0

4    1145     A  female     1  70.11  ...      E      A      A      0  0

..    ...   ...     ...   ...    ...  ...    ...    ...    ...    ... ..

100  1241     A  female     1  88.34  ...      M      B      A      0  1

101  1242     B    male     1  89.84  ...      A      B      A      0  1

102  1243     B    male     1  88.82  ...      T      B      A      0  2

103  1244     A    male     1  86.60  ...      S      B      A      0  2

104  1245     A    male     1  93.71  ...      K      B      A      0  2

[105 rows x 17 columns]

===========Z-score===========

Mean: 1.7904761904761906

Standard deviation: 1.6658773641208555

Outlier : [7, 7, 7, 7, 7]

==========IQR==========

Q1,Q3: 80.0 92.0

IQR: 12.0

Lower fence,upper\_fence: 62.0 110.0

Outliner in the dataset is [60, 51, 60, 60, 60, 61]

\*\*\*\*\*\*\*\*\*\*\*After removing outliner\*\*\*\*\*\*\*\*\*\*

0       60

1       51

2       60

3       60

4       60

      ...

100     98

101     98

102     98

103     99

104    100

Name: Statistics, Length: 105, dtype: int64

\*\*\*\*\*\*\*\*\*\*Data Transformation\*\*\*\*\*\*\*\*\*\*

\*\*Display dataset after data transformation\*\*

       ID class  gender  race    GPA  ...  from2  from3  from4  y  Log\_attendance

0    1141     A    male     1  73.47  ...      A      A      3  0        4.094345

1    1142     A  female     1  71.22  ...      A      A      2  0        3.931826

2    1143     A  female     2  74.56  ...      A      A      0  1        4.094345

3    1144     A  female     1  72.89  ...      A      A      0  0        4.094345

4    1145     A  female     1  70.11  ...      A      A      0  0        4.094345

..    ...   ...     ...   ...    ...  ...    ...    ...    ... ..             ...

100  1241     A  female     1  88.34  ...      B      A      0  1        4.584967

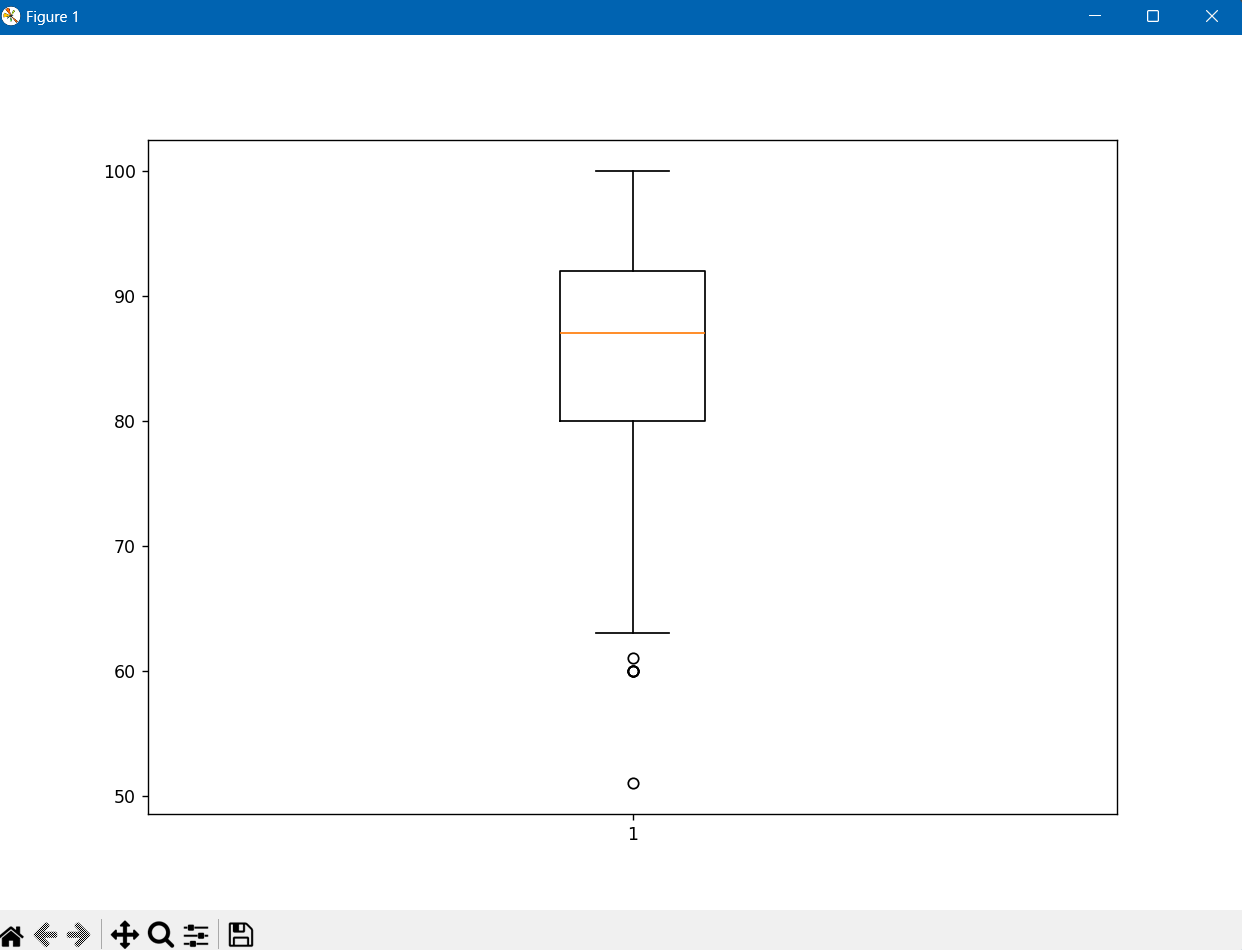
101  1242     B    male     1  89.84  ...      B      A      0  1        4.584967

102  1243     B    male     1  88.82  ...      B      A      0  2        4.584967

103  1244     A    male     1  86.60  ...      B      A      0  2        4.595120

104  1245     A    male     1  93.71  ...      B      A      0  2        4.605170

[105 rows x 18 columns]

****