```
# Group A-2

# Data Wrangling II
# Create an “Academic performance” dataset of students and perform the
following operations using
# Python.

# 1. Scan all variables for missing values and inconsistencies. If there are
missing values and/or
# inconsistencies, use any of the suitable techniques to deal with them.
# 2. Scan all numeric variables for outliers. If there are outliers, use any
of the suitable
# techniques to deal with them.
# 3. Apply data transformations on at least one of the variables. The purpose
of this
# transformation should be one of the following reasons: to change the scale
for better
# understanding of the variable, to convert a non-linear relation into a
linear one, or to
# decrease the skewness and convert the distribution into a normal
distribution.

# Reason and document your approach properly.

!pip install pandas numpy matplotlib seaborn
```

```
Requirement already satisfied: pandas in d:\study
material\dsbd\venv\lib\site-packages (2.2.3)
Requirement already satisfied: numpy in d:\study material\dsbd\venv\lib\site-
packages (2.2.2)
Requirement already satisfied: matplotlib in d:\study
material\dsbd\venv\lib\site-packages (3.10.0)
Requirement already satisfied: seaborn in d:\study
material\dsbd\venv\lib\site-packages (0.13.2)
Requirement already satisfied: python-dateutil>=2.8.2 in d:\study
material\dsbd\venv\lib\site-packages (from pandas) (2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in d:\study
material\dsbd\venv\lib\site-packages (from pandas) (2025.1)
Requirement already satisfied: tzdata>=2022.7 in d:\study
material\dsbd\venv\lib\site-packages (from pandas) (2025.1)
Requirement already satisfied: contourpy>=1.0.1 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (1.3.1)
Requirement already satisfied: cycler>=0.10 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (4.56.0)
Requirement already satisfied: kiwisolver>=1.3.1 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (1.4.8)
Requirement already satisfied: packaging>=20.0 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (24.2)
```

```
Requirement already satisfied: pillow>=8 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (11.1.0)
Requirement already satisfied: pyparsing>=2.3.1 in d:\study
material\dsbd\venv\lib\site-packages (from matplotlib) (3.2.1)
Requirement already satisfied: six>=1.5 in d:\study
material\dsbd\venv\lib\site-packages (from python-dateutil>=2.8.2->pandas)
(1.17.0)


[notice] A new release of pip is available: 24.3.1 -> 25.0.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
 # Importing libraries
 import pandas as pd
 import numpy as np
 import matplotlib.pyplot as plt
 import seaborn as sns

def DetectOutlier(df,var):
 # IQR method is used to deal with outliers
  Q1 = df[var].quantile(0.25)
  Q3 = df[var].quantile(0.75)
  IQR = Q3 - Q1
  high, low = Q3+1.5*IQR, Q1-1.5*IQR
  print("Highest allowed in variable:", var, high)
  print("lowest allowed in variable:", var, low)
  count = df[(df[var] > high) | (df[var] < low)][var].count()
  print('Total outliers in:',var,':',count)
 # new dataframe is created which contains outliers
  df1 = df[((df[var] < low) | (df[var] > high))] #these are outliers
  print('Outliers : \n', len(df1))
  print(df1.T)
  df = df[((df[var] >= low) & (df[var] <= high))] #now filter out data which
is not outlier
  return(df)

# Reading dataset
df = pd.read_csv('academic.csv')

# Display basic information
print('Information of Dataset:\n', df.info)
```

```
Information of Dataset:
 <bound method DataFrame.info of     gender NationalITy PlaceofBirth
StageID GradeID SectionID  \
0        M          KW       KuwaIT    lowerlevel    G-04         A
1        M          KW       KuwaIT    lowerlevel    G-04         A
2        M          KW       KuwaIT    lowerlevel    G-04         A
3        M          KW       KuwaIT    lowerlevel    G-04         A
4        M          KW       KuwaIT    lowerlevel    G-04         A
..     ...         ...          ...           ...     ...       ...
```

```
475      F      Jordan       Jordan  MiddleSchool    G-08         A
476      F      Jordan       Jordan  MiddleSchool    G-08         A
477      F      Jordan       Jordan  MiddleSchool    G-08         A
478      F      Jordan       Jordan  MiddleSchool    G-08         A
479      F      Jordan       Jordan  MiddleSchool    G-08         A

         Topic Semester Relation  raisedhands  VisITedResources  \
0           IT        F   Father         15.0                16
1           IT        F   Father          NaN                20
2           IT        F   Father         10.0                 7
3           IT        F   Father         30.0                25
4           IT        F   Father          0.0                50
..         ...      ...      ...          ...               ...
475  Chemistry        S   Father          5.0                 4
476    Geology        F   Father         50.0                77
477    Geology        S   Father         55.0                74
478    History        F   Father         30.0                17
479    History        S   Father         35.0                14

     AnnouncementsView  Discussion ParentAnsweringSurvey  \
0                    2          20                   Yes
1                    3          25                   Yes
2                    0          30                    No
3                    5          35                    No
4                   12          50                    No
..                 ...         ...                   ...
475                  5           8                    No
476                 14          28                    No
477                 25          29                    No
478                 14          57                    No
479                 23          62                    No

    ParentschoolSatisfaction StudentAbsenceDays Class
0                       Good            Under-7     M
1                       Good            Under-7     M
2                        Bad            Above-7     L
3                        Bad            Above-7     L
4                        Bad            Above-7     M
..                       ...                ...   ...
475                      Bad            Above-7     L
476                      Bad            Under-7     M
477                      Bad            Under-7     M
478                      Bad            Above-7     L
479                      Bad            Above-7     L

[480 rows x 17 columns]>
```

```
print('Shape of Dataset (row x column): ', df.shape)
```

```
Shape of Dataset (row x column):  (480, 17)
```

```
print('Columns Name: ', df.columns)
```

```
Columns Name:  Index(['gender', 'NationalITy', 'PlaceofBirth', 'StageID',
'GradeID',
       'SectionID', 'Topic', 'Semester', 'Relation', 'raisedhands',
       'VisITedResources', 'AnnouncementsView', 'Discussion',
       'ParentAnsweringSurvey', 'ParentschoolSatisfaction',
       'StudentAbsenceDays', 'Class'],
      dtype='object')
```

```
print('Total elements in dataset:', df.size)
```

```
Total elements in dataset: 8160
```

```
print('Datatype of attributes (columns):', df.dtypes)
```

```
Datatype of attributes (columns): gender                      object
NationalITy                  object
PlaceofBirth                 object
StageID                      object
GradeID                      object
SectionID                    object
Topic                        object
Semester                     object
Relation                     object
raisedhands                 float64
VisITedResources              int64
AnnouncementsView             int64
Discussion                    int64
ParentAnsweringSurvey        object
ParentschoolSatisfaction     object
StudentAbsenceDays           object
Class                        object
dtype: object
```

```
print('First 5 rows:\n', df.head().T)
```

```
First 5 rows:
                                  0           1           2           3  \
gender                            M           M           M           M
NationalITy                      KW          KW          KW          KW
PlaceofBirth                 KuwaIT      KuwaIT      KuwaIT      KuwaIT
StageID                  lowerlevel  lowerlevel  lowerlevel  lowerlevel
GradeID                        G-04        G-04        G-04        G-04
SectionID                         A           A           A           A
Topic                            IT          IT          IT          IT
Semester                          F           F           F           F
Relation                     Father      Father      Father      Father
raisedhands                    15.0         NaN        10.0        30.0
VisITedResources                 16          20           7          25
AnnouncementsView                 2           3           0           5
```

```
Discussion                    20     25     30     35
ParentAnsweringSurvey        Yes    Yes     No     No
ParentschoolSatisfaction    Good   Good    Bad    Bad
StudentAbsenceDays         Under-7 Under-7 Above-7 Above-7
Class                         M      M      L      L

                              4
gender                        M
NationalITy                  KW
PlaceofBirth              KuwaIT
StageID                lowerlevel
GradeID                    G-04
SectionID                     A
Topic                        IT
Semester                      F
Relation                 Father
raisedhands                0.0
VisITedResources            50
AnnouncementsView           12
Discussion                  50
ParentAnsweringSurvey       No
ParentschoolSatisfaction   Bad
StudentAbsenceDays       Above-7
Class                         M
```

```
print('Any 5 rows:\n',df.sample(5).T)
```

```
Any 5 rows:
                              44          385         170         65   \
gender                         F           F           M           M
NationalITy                   KW        Iraq          KW          KW
PlaceofBirth              KuwaIT        Iraq      KuwaIT      KuwaIT
StageID                HighSchool  lowerlevel  lowerlevel  HighSchool
GradeID                    G-09        G-02        G-02        G-12
SectionID                     A           B           B           A
Topic                        IT      Arabic      French     English
Semester                      F           S           S           F
Relation                 Father         Mum      Father      Father
raisedhands                33.0        79.0        40.0        13.0
VisITedResources             33          93          62           5
AnnouncementsView            30          49          83          18
Discussion                   90          23          33          19
ParentAnsweringSurvey        No         Yes         Yes          No
ParentschoolSatisfaction    Bad        Good        Good         Bad
StudentAbsenceDays       Under-7     Under-7     Under-7     Above-7
Class                         M           H           H           L

                             345
gender                         F
NationalITy                Jordan
```

```
PlaceofBirth                 Jordan
StageID                   lowerlevel
GradeID                       G-02
SectionID                        B
Topic                       French
Semester                         F
Relation                       Mum
raisedhands                   13.0
VisITedResources                82
AnnouncementsView               20
Discussion                      30
ParentAnsweringSurvey           No
ParentschoolSatisfaction      Good
StudentAbsenceDays          Under-7
Class                            H
```

```
# Display Statistical information
print('Statistical information of Numerical Columns: \n',df.describe())
```

```
Statistical information of Numerical Columns:
        raisedhands  VisITedResources  AnnouncementsView  Discussion
count   478.000000        480.000000         480.000000  480.000000
mean     46.939331         54.797917          37.918750   43.283333
std      31.375699         33.080007          26.611244   27.637735
min       0.000000          0.000000           0.000000    1.000000
25%      15.000000         20.000000          14.000000   20.000000
50%      50.000000         65.000000          33.000000   39.000000
75%      75.000000         84.000000          58.000000   70.000000
max     170.000000         99.000000          98.000000   99.000000
```

```
 # Display Null values
print('Total Number of Null Values in Dataset: \n', df.isna().sum())
```

```
Total Number of Null Values in Dataset:
 gender                     2
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 2
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
```

```
Class                       0
dtype: int64
```

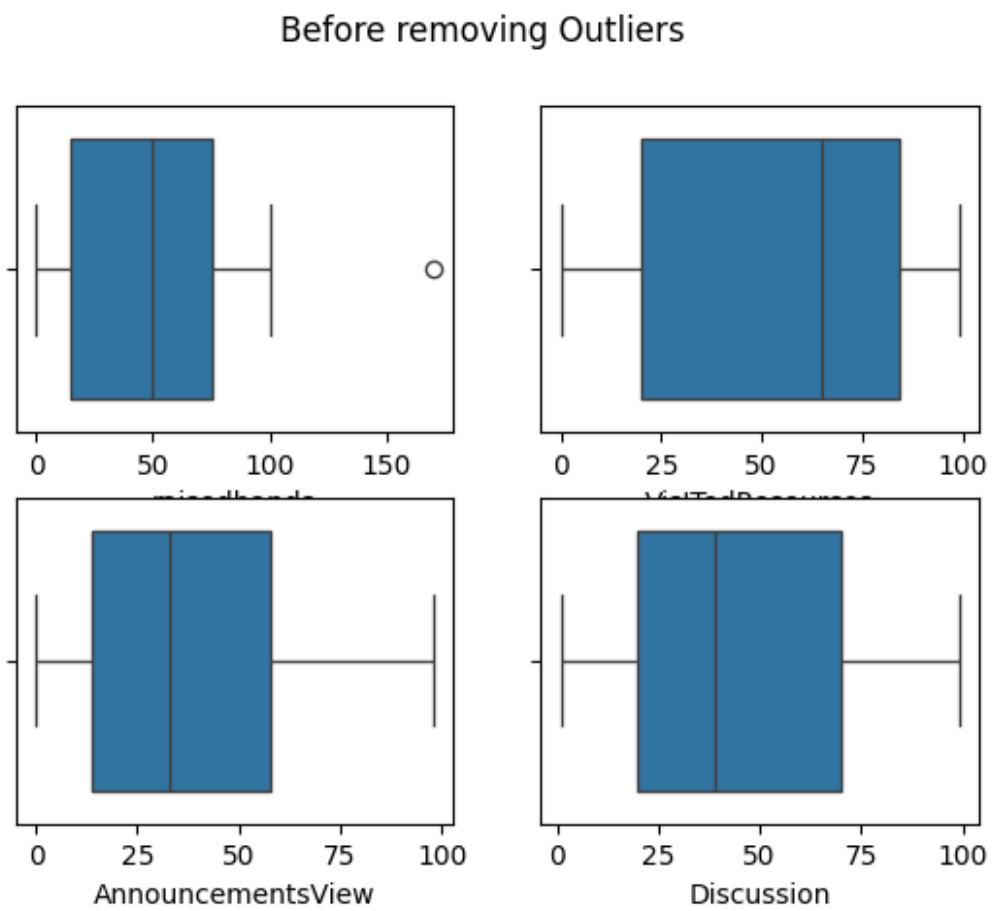
```
 # Fill the missing values
df['gender'].fillna(df['gender'].mode()[0], inplace=True)
df['raisedhands'].fillna(df['raisedhands'].mean(), inplace=True)
print('Total Number of Null Values in Dataset: \n', df.isna().sum())
```

```
Total Number of Null Values in Dataset:
 gender                     0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 0
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

```
C:\Users\Aishwarya
Bhansali\AppData\Local\Temp\ipykernel_20256\3014334111.py:2: FutureWarning: A
value is trying to be set on a copy of a DataFrame or Series through chained
assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves
as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  df['gender'].fillna(df['gender'].mode()[0], inplace=True)
C:\Users\Aishwarya
Bhansali\AppData\Local\Temp\ipykernel_20256\3014334111.py:3: FutureWarning: A
value is trying to be set on a copy of a DataFrame or Series through chained
assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves
as a copy.
```

```
For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.
```

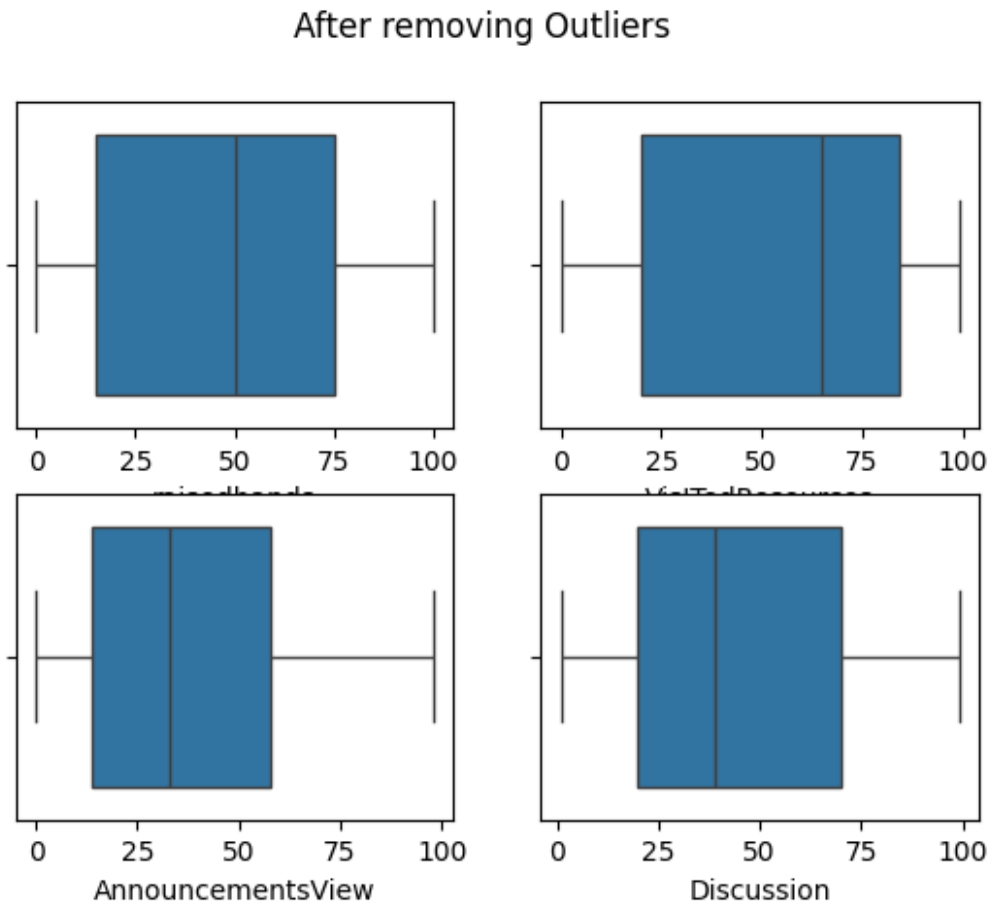
```
  df['raisedhands'].fillna(df['raisedhands'].mean(), inplace=True)

 # Converting categorical to numeric using Find and replace method
df['Relation']=df['Relation'].astype('category')
df['Relation']=df['Relation'].cat.codes

 # Outliers can be visualized using boxplot
 # using seaborn library we can plot the boxplot
fig, axes = plt.subplots(2,2)
fig.suptitle('Before removing Outliers')
sns.boxplot(data = df, x ='raisedhands', ax=axes[0,0])
sns.boxplot(data = df, x ='VisITedResources', ax=axes[0,1])
sns.boxplot(data = df, x ='AnnouncementsView', ax=axes[1,0])
sns.boxplot(data = df, x ='Discussion', ax=axes[1,1])
plt.show()
```



```
#Display and remove outliers
df = DetectOutlier(df, 'raisedhands')
fig, axes = plt.subplots(2,2)
fig.suptitle('After removing Outliers')
sns.boxplot(data = df, x ='raisedhands', ax=axes[0,0])

sns.boxplot(data = df, x ='VisITedResources', ax= axes[0,1])
sns.boxplot(data = df, x ='AnnouncementsView', ax= axes[1,0])
sns.boxplot(data = df, x ='Discussion', ax= axes[1,1])
plt.show()
```

```
Highest allowed in variable: raisedhands 165.0
lowest allowed in variable: raisedhands -75.0
Total outliers in: raisedhands : 1
Outliers :
 1
```

	28
gender	M
NationalITy	KW
PlaceofBirth	KuwaIT
StageID	MiddleSchool
GradeID	G-08
SectionID	A
Topic	Science
Semester	F
Relation	0
raisedhands	170.0
VisITedResources	85
AnnouncementsView	52
Discussion	43
ParentAnsweringSurvey	Yes
ParentschoolSatisfaction	Good
StudentAbsenceDays	Under-7
Class	M




```
 print('---------------- Data Skew Values before Yeo John Transformation ----
------------------')
 # There are two types
 # 1. Left skew
 # 2. Right skew
 # Formula to find out data skewness = 3*(mean-median)/std
 # = 0 (no skew) print
 # = negative (Negative skew) left skewed data
 # = positve (Positive skew) Right skewed data
 # = -0.5 to 0 to 0.5 (acceptable skew)
 # = -0.5> <-1 moderate negative skew
 # = 0.5> <1 moderate positive skew
 # = > -1 high negative
 # = > 1 high positive

# Checking skewness for 'raisedhands' column
print('raisedhands: ', df['raisedhands'].skew())

# Checking skewness for 'VisITedResources' column
print('VisITedResources: ', df['VisITedResources'].skew())

# Checking skewness for 'AnnouncementsView' column
```

```
print('AnnouncementsView: ', df['AnnouncementsView'].skew())

# Checking skewness for 'Discussion' column
print('Discussion: ', df['Discussion'].skew())

# Create subplots to visualize data distribution before and after
transformation
fig, axes = plt.subplots(2,2)
fig.suptitle('Handling Data Skewness')

# Plot histogram for 'AnnouncementsView' before transformation
sns.histplot(ax = axes[0,0], data = df['AnnouncementsView'], kde=True)

# Plot histogram for 'Discussion' before transformation
sns.histplot(ax = axes[0,1], data = df['Discussion'], kde=True)

# Apply Yeo-Johnson Power Transformation to handle skewness
from sklearn.preprocessing import PowerTransformer
yeojohnTr = PowerTransformer(standardize=True)

# Transform 'AnnouncementsView' column
df['AnnouncementsView'] =
yeojohnTr.fit_transform(df['AnnouncementsView'].values.reshape(-1,1))

# Transform 'Discussion' column
df['Discussion'] = yeojohnTr.fit_transform(df['Discussion'].values.reshape(-
1,1))

# Print skewness values after transformation
print('---------------- Data Skew Values after Yeo John Transformation ------
----------------')
print('AnnouncementsView: ', df['AnnouncementsView'].skew())
print('Discussion: ', df['Discussion'].skew())

# Plot histogram for 'AnnouncementsView' after transformation
sns.histplot(ax = axes[1,0], data = df['AnnouncementsView'], kde=True)

# Plot histogram for 'Discussion' after transformation
sns.histplot(ax = axes[1,1], data =df['Discussion'], kde=True)

# Display the plots
plt.show()


---------------- Data Skew Values before Yeo John Transformation ------------
----------
raisedhands:  0.028374079559687623
VisITedResources:  -0.3388404568312024
AnnouncementsView:  0.4021955128761278
```

```
Discussion:  0.3621541732143617
---------------- Data Skew Values after Yeo John Transformation -------------
---------
AnnouncementsView:  -0.1800377395845211
Discussion:  -0.13328782723929383
```