In [26]:
```
import pandas as pd
import numpy as np
df = pd.read_csv('A1.csv')
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [27]:
```
# Missing values in data

df.isnull().sum()
```

Out[27]:
```
Roll No       0
Name          0
Subject 1     0
Subject 2     0
Subject 3     1
Subject 4     1
Attendance    0
dtype: int64
```

In [28]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1001 entries, 0 to 1000
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Roll No     1001 non-null   int64
 1   Name        1001 non-null   object
 2   Subject 1   1001 non-null   int64
 3   Subject 2   1001 non-null   int64
 4   Subject 3   1000 non-null   float64
 5   Subject 4   1000 non-null   float64
 6   Attendance  1001 non-null   int64
dtypes: float64(2), int64(4), object(1)
memory usage: 54.9+ KB
```

In [29]:
```
bs = pd.notnull(df)
print(bs)
```

```
      Roll No  Name  Subject 1  Subject 2  Subject 3  Subject 4  Attendance
0        True  True       True       True       True       True        True
1        True  True       True       True       True      False        True
2        True  True       True       True       True       True        True
3        True  True       True       True      False       True        True
4        True  True       True       True       True       True        True
...       ...   ...        ...        ...        ...        ...         ...
996      True  True       True       True       True       True        True
997      True  True       True       True       True       True        True
998      True  True       True       True       True       True        True
999      True  True       True       True       True       True        True
1000     True  True       True       True       True       True        True

[1001 rows x 7 columns]
```

In [30]:
```
df.fillna(0, inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        0.0
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        0.0       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [31]:
```
df = pd.read_csv('A1.csv')
df.fillna(method='pad')
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [32]:
```
df = pd.read_csv('A1.csv')
df.fillna(method='bfill')
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [33]:
```
df = pd.read_csv('A1.csv')
df.interpolate(method='linear', limit_direction='forward')
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [34]:
```
df = pd.read_csv('A1.csv')
nc = ['Subject 1', 'Subject 2', 'Subject 3', 'Subject 4', 'Attendance']
for col in nc:
    df[col].fillna(df[col].mean(), inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62      73.00     92.000
1           2     Student_2         72         97      82.00     80.545
2           3     Student_3        100         88      71.00     99.000
3           4     Student_4         72         99      79.89     84.000
4           5     Student_5         97         70      84.00     70.000
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68      84.00     66.000
997       998   Student_998         61         96      62.00     84.000
998       999   Student_999         72         76      90.00     72.000
999      1000  Student_1000         68         87     100.00     76.000
1000        1     Student_1        100         62      73.00     92.000

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [35]:
```
df = pd.read_csv('A1.csv')
df.replace(to_replace=np.nan, value=85)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [36]:
```
df = pd.read_csv('A1.csv')
df.dropna()
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [37]:
```
df = pd.read_csv('A1.csv')
df.dropna(how='all', inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
1           2     Student_2         72         97       82.0        NaN
2           3     Student_3        100         88       71.0       99.0
3           4     Student_4         72         99        NaN       84.0
4           5     Student_5         97         70       84.0       70.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
1             78
2            -94
3             86
4             86
...          ...
996           98
997           83
998           90
999           79
1000          96

[1001 rows x 7 columns]
```

In [38]:
```
df = pd.read_csv('A1.csv')
df.dropna(how='any', inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
2           3     Student_3        100         88       71.0       99.0
4           5     Student_5         97         70       84.0       70.0
5           6     Student_6         98         76       89.0       92.0
6           7     Student_7         61         64       97.0       98.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
2            -94
4             86
5             82
6             83
...          ...
996           98
997           83
998           90
999           79
1000          96

[999 rows x 7 columns]
```

In [39]:
```
df = pd.read_csv('A1.csv')
df.dropna(how='any', axis=1, inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Attendance
0           1     Student_1        100         62          96
1           2     Student_2         72         97          78
2           3     Student_3        100         88         -94
3           4     Student_4         72         99          86
4           5     Student_5         97         70          86
...       ...           ...        ...        ...         ...
996       997   Student_997         88         68          98
997       998   Student_998         61         96          83
998       999   Student_999         72         76          90
999      1000  Student_1000         68         87          79
1000        1     Student_1        100         62          96

[1001 rows x 5 columns]
```

In [40]:
```
df = pd.read_csv('A1.csv')
df.dropna(axis=0, inplace=True)
print(df)
```

```
      Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0           1     Student_1        100         62       73.0       92.0
2           3     Student_3        100         88       71.0       99.0
4           5     Student_5         97         70       84.0       70.0
5           6     Student_6         98         76       89.0       92.0
6           7     Student_7         61         64       97.0       98.0
...       ...           ...        ...        ...        ...        ...
996       997   Student_997         88         68       84.0       66.0
997       998   Student_998         61         96       62.0       84.0
998       999   Student_999         72         76       90.0       72.0
999      1000  Student_1000         68         87      100.0       76.0
1000        1     Student_1        100         62       73.0       92.0

      Attendance
0             96
2            -94
4             86
5             82
6             83
...          ...
996           98
997           83
998           90
999           79
1000          96

[999 rows x 7 columns]
```

In [41]:
```
df.drop_duplicates(inplace=True)
print(df)
```

```
     Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0          1     Student_1        100         62       73.0       92.0
2          3     Student_3        100         88       71.0       99.0
4          5     Student_5         97         70       84.0       70.0
5          6     Student_6         98         76       89.0       92.0
6          7     Student_7         61         64       97.0       98.0
..       ...           ...        ...        ...        ...        ...
995      996   Student_996         74         89       85.0       71.0
996      997   Student_997         88         68       84.0       66.0
997      998   Student_998         61         96       62.0       84.0
998      999   Student_999         72         76       90.0       72.0
999     1000  Student_1000         68         87      100.0       76.0

     Attendance
0            96
2           -94
4            86
5            82
6            83
..          ...
995          87
996          98
997          83
998          90
999          79

[998 rows x 7 columns]
```

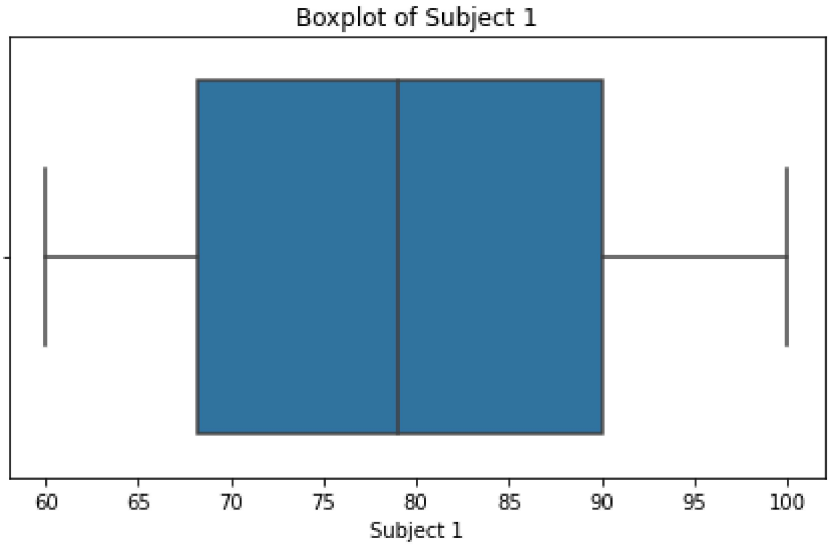
In [42]:
```
x = df[['Subject 1', 'Subject 2', 'Subject 3', 'Subject 4', 'Attendance']]
print(df[x<0])
```

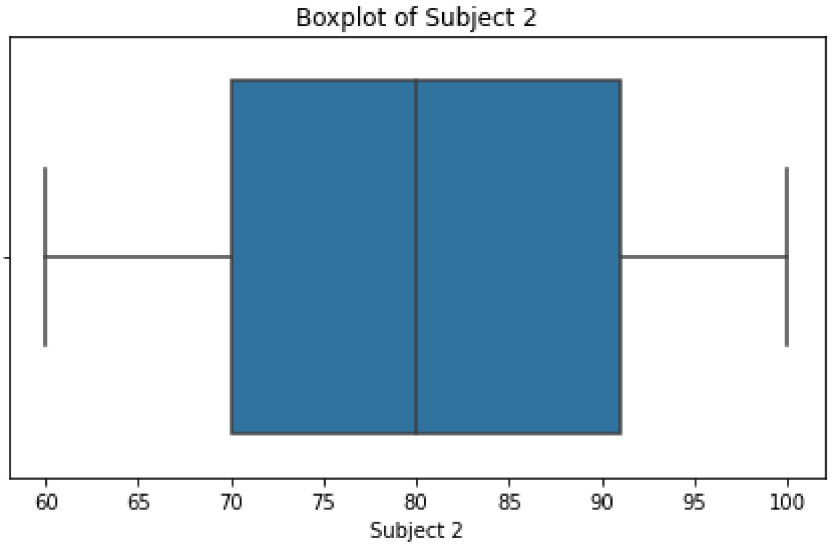
```
     Roll No Name  Subject 1  Subject 2  Subject 3  Subject 4  Attendance
0        NaN  NaN        NaN        NaN        NaN        NaN         NaN
2        NaN  NaN        NaN        NaN        NaN        NaN       -94.0
4        NaN  NaN        NaN        NaN        NaN        NaN         NaN
5        NaN  NaN        NaN        NaN        NaN        NaN         NaN
6        NaN  NaN        NaN        NaN        NaN        NaN         NaN
..       ...  ...        ...        ...        ...        ...         ...
995      NaN  NaN        NaN        NaN        NaN        NaN         NaN
996      NaN  NaN        NaN        NaN        NaN        NaN         NaN
997      NaN  NaN        NaN        NaN        NaN        NaN         NaN
998      NaN  NaN        NaN        NaN        NaN        NaN         NaN
999      NaN  NaN        NaN        NaN        NaN        NaN         NaN

[998 rows x 7 columns]
```

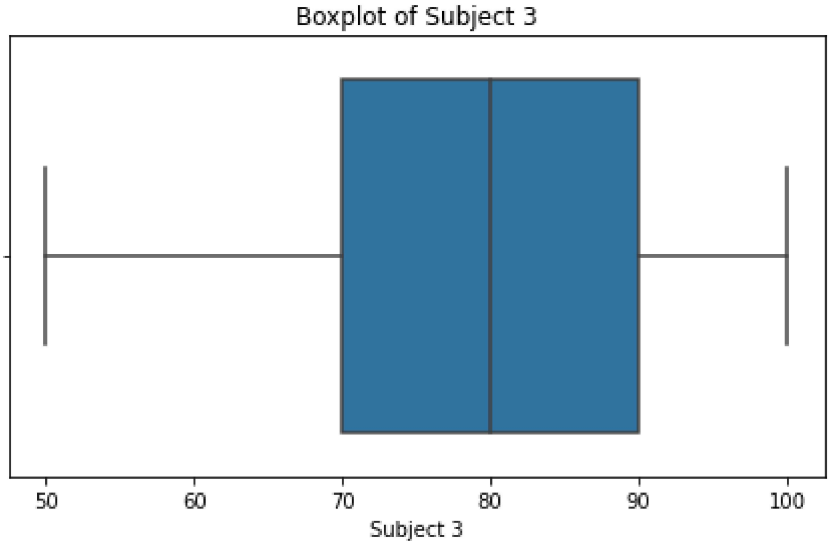
In [43]:
```
# Inconsistencies in the data

x = x.clip(lower=0)
print(x)
```

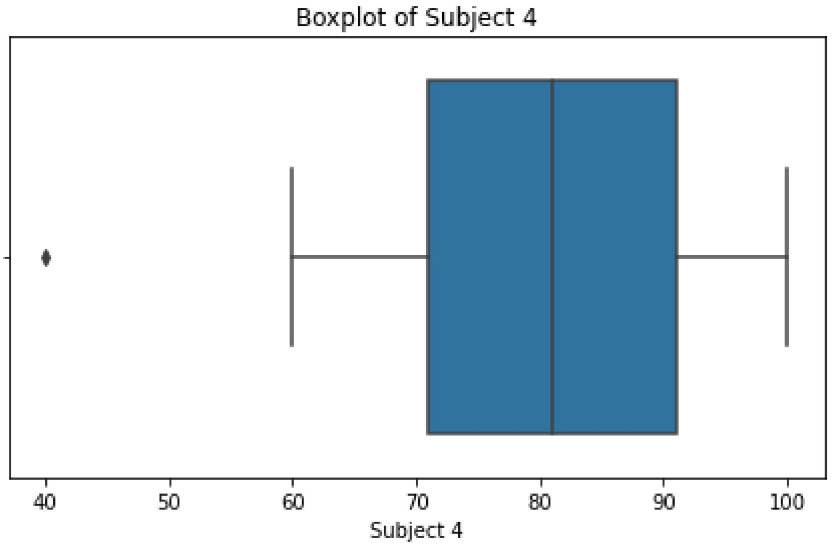
```
     Subject 1  Subject 2  Subject 3  Subject 4  Attendance
0          100         62       73.0       92.0          96
2          100         88       71.0       99.0           0
4           97         70       84.0       70.0          86
5           98         76       89.0       92.0          82
6           61         64       97.0       98.0          83
..         ...        ...        ...        ...         ...
995         74         89       85.0       71.0          87
996         88         68       84.0       66.0          98
997         61         96       62.0       84.0          83
998         72         76       90.0       72.0          90
999         68         87      100.0       76.0          79

[998 rows x 5 columns]
```

In [44]:
```
# Handling Outliers

import seaborn as sns
import matplotlib.pyplot as plt

nc = ['Subject 1', 'Subject 2', 'Subject 3', 'Subject 4']
for col in nc:
    sns.boxplot(x=df[col])
    plt.title(f'Boxplot of {col}')
    plt.tight_layout()
    plt.show()
```









In [48]:
```
for col in nc:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lb = Q1 - 1.5*IQR
```

```
    ub = Q3 + 1.5*IQR
    df[col] = np.where((df[col]<lb) | (df[col]>ub), df[col].median(), df[col])

print(df)
```

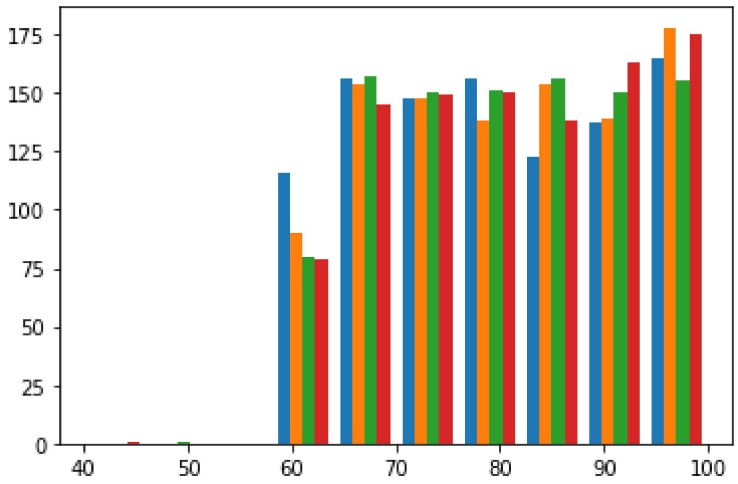
```
     Roll No          Name  Subject 1  Subject 2  Subject 3  Subject 4  \
0          1     Student_1      100.0       62.0       73.0       92.0
2          3     Student_3      100.0       88.0       71.0       99.0
4          5     Student_5       97.0       70.0       84.0       70.0
5          6     Student_6       98.0       76.0       89.0       92.0
6          7     Student_7       61.0       64.0       97.0       98.0
..       ...           ...        ...        ...        ...        ...
995      996   Student_996       74.0       89.0       85.0       71.0
996      997   Student_997       88.0       68.0       84.0       66.0
997      998   Student_998       61.0       96.0       62.0       84.0
998      999   Student_999       72.0       76.0       90.0       72.0
999     1000  Student_1000       68.0       87.0      100.0       76.0

     Attendance
0            96
2           -94
4            86
5            82
6            83
..          ...
995          87
996          98
997          83
998          90
999          79

[998 rows x 7 columns]
```

In [57]:
```
df = pd.read_csv('A1.csv')

z = (df - df.mean())/df.std()
print(z)

plt.hist(df[nc])
plt.show()

print(df.skew())
```

```
/tmp/ipykernel_19954/784410992.py:3: FutureWarning: Dropping of nuisance columns in
DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version t
his will raise TypeError.  Select only valid columns before calling the reduction.
  z = (df - df.mean())/df.std()
```

	Attendance	Name	Roll No	Subject 1	Subject 2	Subject 3	Subject 4
0	1.029784	NaN	-1.726012	1.718218	-1.517511	-0.597082	0.971282
1	-0.721646	NaN	-1.722553	-0.598529	1.422599	0.182851	NaN
2	-17.457528	NaN	-1.719094	1.718218	0.666570	-0.770400	1.564820
3	0.056767	NaN	-1.715635	-0.598529	1.590605	NaN	0.292953
4	0.056767	NaN	-1.712176	1.469995	-0.845486	0.356169	-0.894122
...	...	...	...	...	...	...	...
996	1.224387	NaN	1.719087	0.725327	-1.013492	0.356169	-1.233287
997	-0.235138	NaN	1.722546	-1.508679	1.338595	-1.550333	0.292953
998	0.445974	NaN	1.726005	-0.598529	-0.341467	0.876124	-0.724540
999	-0.624344	NaN	1.729464	-0.929493	0.582567	1.742716	-0.385376
1000	1.029784	NaN	-1.726012	1.718218	-1.517511	-0.597082	0.971282

[1001 rows x 7 columns]




```
Roll No       0.000021
Subject 1     0.091410
Subject 2    -0.010696
Subject 3     0.006017
Subject 4    -0.068298
Attendance   -5.348146
dtype: float64
```

```
/tmp/ipykernel_19954/784410992.py:9: FutureWarning: Dropping of nuisance columns in
DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version t
his will raise TypeError.  Select only valid columns before calling the reduction.
  print(df.skew())
```

In [ ]: