```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
```

```
# Creating a synthetic Academic Performance dataset
data = {
    'Student_ID': range(1, 21),  # Unique Student IDs
    'Math_Score': [80, 85, 78, 90, 95, 60, 55, 45, 88, 92, 79, 83, 50, 97, 65, 75, 81, 89, 72, 90],
    'Science_Score': [85, 88, 75, 92, 94, 58, 54, 43, 86, 90, 77, 81, 48, 95, 62, 73, 79, 85, 70, 89],
    'English_Score': [78, 82, 80, 85, 89, 55, 50, 40, 83, 87, 76, 80, 45, 90, 60, 70, 78, 85, 68, 88],
    'Attendance': [90, 85, 80, 95, 98, 60, 50, 40, 88, 92, 77, 84, 55, 96, 65, 75, 82, 90, 70, 91]
}
df = pd.DataFrame(data)
```

```
df.to_csv("/kaggle/working/academic_performance.csv", index=False)
```

```
df = pd.read_csv("/kaggle/working/academic_performance.csv")
print(df.head())  # Display first few rows
```

```
   Student_ID  Math_Score  Science_Score  English_Score  Attendance
0           1          80             85             78          90
1           2          85             88             82          85
2           3          78             75             80          80
3           4          90             92             85          95
4           5          95             94             89          98
```

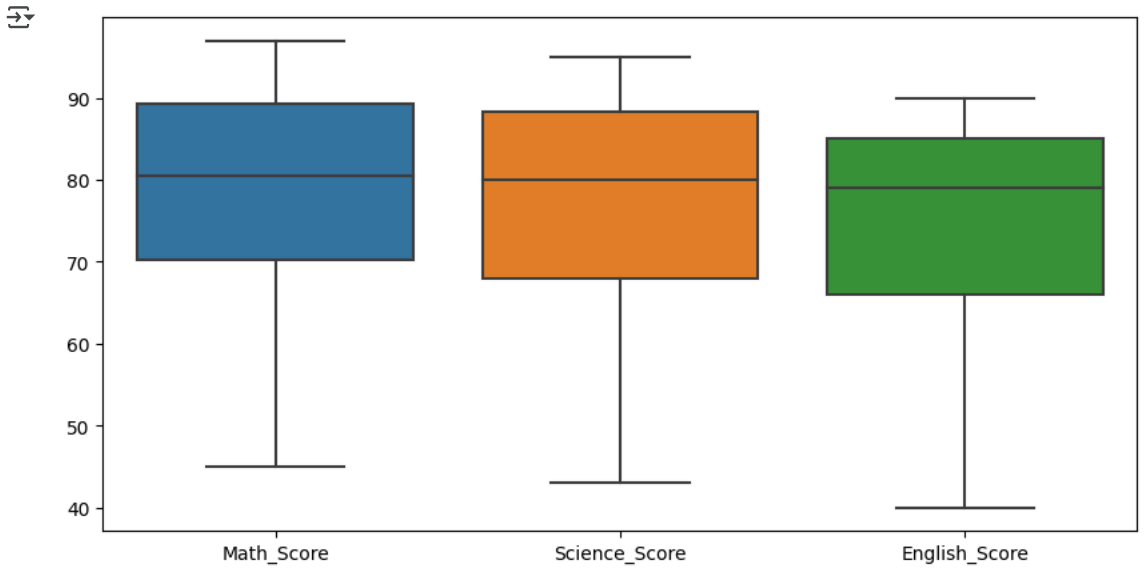
```
print(df.isnull().sum())  # Check missing values
print(df.info())  # Check data types
```

```
Student_ID       0
Math_Score       0
Science_Score    0
English_Score    0
Attendance       0
dtype: int64
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Student_ID     20 non-null     int64
 1   Math_Score     20 non-null     int64
 2   Science_Score  20 non-null     int64
 3   English_Score  20 non-null     int64
 4   Attendance     20 non-null     int64
dtypes: int64(5)
memory usage: 928.0 bytes
```

```
None
```

```
df.fillna(df.median(), inplace=True)
```

```
plt.figure(figsize=(10,5))
sns.boxplot(data=df[['Math_Score', 'Science_Score', 'English_Score']])
plt.show()
```

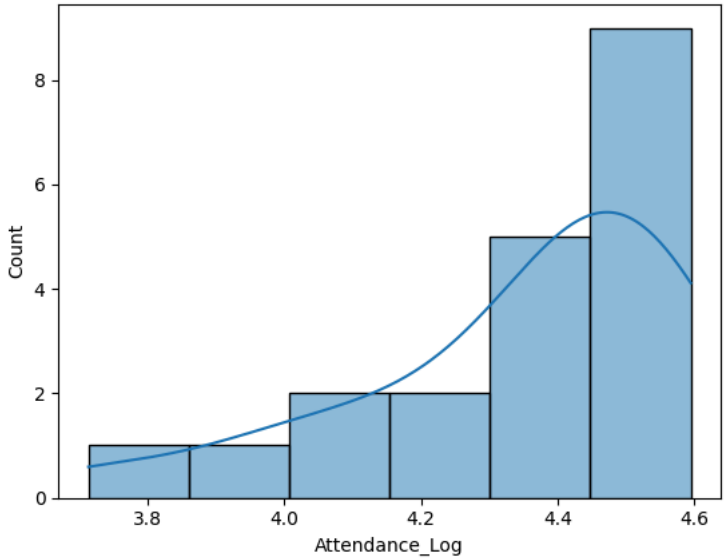



```
z_scores = np.abs(stats.zscore(df[['Math_Score', 'Science_Score', 'English_Score']]))
df = df[(z_scores < 3).all(axis=1)]  # Keep values within 3 standard deviations
```

```
df['Attendance_Log'] = np.log1p(df['Attendance']).replace([np.inf, -np.inf], np.nan)
df.dropna(subset=['Attendance_Log'], inplace=True)

# Ensure Seaborn gets a clean dataset
sns.histplot(df.loc[df['Attendance_Log'].notna(), 'Attendance_Log'], kde=True)
plt.show()
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed in a future version. Convert inf values to NaN before operat
  with pd.option_context('mode.use_inf_as_na', True):
```




```
df.to_csv("/kaggle/working/cleaned_academic_performance.csv", index=False)
```