

```

In [4]: # 1. Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style("whitegrid")

# 2. Load the Dataset
# قراءة ملف بيانات الطلاب الحقيقي
import tkinter as pd_file
from tkinter import filedialog
root = pd_file.Tk()
root.withdraw()
root.attributes('-topmost', True)
file_path = filedialog.askopenfilename(title="Select your Kaggle CSV file", filetype='csv')
df = pd.read_csv(file_path)
print("Data Shape:", df.shape)
print("\n--- First 5 Rows of Student Data ---")
print(df.head())

# -----
# Step 1: Handle Missing Values
# -----
print("\n--- Missing values before handling ---")
print(df.isnull().sum())

# تقسيم الأعمدة تلقائياً حسب نوعها
num_cols = df.select_dtypes(include=[np.number]).columns
cat_cols = df.select_dtypes(exclude=[np.number]).columns

# تعبئة القيم المفقودة بالأرقام بالمتوسط والنصوص بالقيمة الأكثر تكراراً
for col in num_cols:
    df[col] = df[col].fillna(df[col].median())
for col in cat_cols:
    df[col] = df[col].fillna(df[col].mode()[0])

print("\n--- Missing values after handling ---")
print(df.isnull().sum())

# -----
# Step 2: Handle Outliers (IQR Method)
# -----
print("\nRows before removing outliers:", df.shape[0])

for col in num_cols:
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower = Q1 - 1.5 * IQR
    upper = Q3 + 1.5 * IQR
    df = df[(df[col] >= lower) & (df[col] <= upper)]

print("Rows after removing outliers:", df.shape[0])

```

```
# للتأكد من معالجة القيم الشاذة لبيانات الطلاب Boxplots رسم
fig, axes = plt.subplots(1, len(num_cols), figsize=(4 * len(num_cols), 4))
for ax, col in zip(axes, num_cols):
    sns.boxplot(y=df[col], ax=ax, color="skyblue")
    ax.set_title(f"Boxplot: {col}")
plt.tight_layout()
plt.show()

# -----
# Step 3: Perform Data Visualization
# -----
# 1. رسم التوزيع التكراري للمتغيرات الرقمية (مثل raisedhands و VisITEDResources)
fig, axes = plt.subplots(1, len(num_cols), figsize=(4 * len(num_cols), 4))
for ax, col in zip(axes, num_cols):
    sns.histplot(df[col], kde=True, ax=ax, color="teal")
    ax.set_title(f"Distribution: {col}")
plt.tight_layout()
plt.show()

# 2. بين سلوكيات الطلاب الرقمية (Correlation Heatmap) مصفوفة الارتباط
plt.figure(figsize=(6, 5))
sns.heatmap(df[num_cols].corr(), annot=True, cmap="coolwarm", fmt=".2f")
plt.title("Correlation Heatmap (Student Behavior)")
plt.show()

# 3. رسم بياني يوضح توزيع الطلاب حسب المادة الدراسية (Topic)
plt.figure(figsize=(8, 4))
df['Topic'].value_counts().plot(kind="bar", color="coral")
plt.title("Count of Students by Topic")
plt.ylabel("Count")
plt.xlabel("Topic")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

# 4. حفظ البيانات النظيفة
df.to_csv("Student_Data_Cleaned.csv", index=False)
print("\nSaved cleaned dataset successfully.")
```

Data Shape: (480, 17)

--- First 5 Rows of Student Data ---

	gender	NationalITY	PlaceofBirth	StageID	GradeID	SectionID	Topic \
0	M	SaudiArabia	SaudiArabia	HighSchool	G-11	B	IT
1	F	lebanon	lebanon	MiddleSchool	G-06	B	Science
2	M	KW	KW	MiddleSchool	G-02	B	Science
3	F	KW	KW	lowerlevel	G-11	A	Arabic
4	F	SaudiArabia	SaudiArabia	lowerlevel	G-04	C	Arabic

	Semester	Relation	raisedhands	VisITedResources	AnnouncementsView \
0	F	Father	52.0	100.0	NaN
1	S	Mum	96.0	16.0	62.0
2	S	Father	29.0	0.0	52.0
3	S	Father	60.0	0.0	50.0
4	F	Father	40.0	85.0	41.0

	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction \
0	74.0	Yes	Bad
1	10.0	Yes	Bad
2	76.0	No	Bad
3	41.0	Yes	Bad
4	67.0	Yes	Good

	StudentAbsenceDays	Class
0	Above-7	H
1	Under-7	H
2	Under-7	L
3	Under-7	L
4	Above-7	M

--- Missing values before handling ---

gender	0
NationalITY	12
PlaceofBirth	0
StageID	0
GradeID	0
SectionID	0
Topic	0
Semester	0
Relation	0
raisedhands	12
VisITedResources	12
AnnouncementsView	12
Discussion	12
ParentAnsweringSurvey	0
ParentschoolSatisfaction	0
StudentAbsenceDays	0
Class	0
dtype: int64	

--- Missing values after handling ---

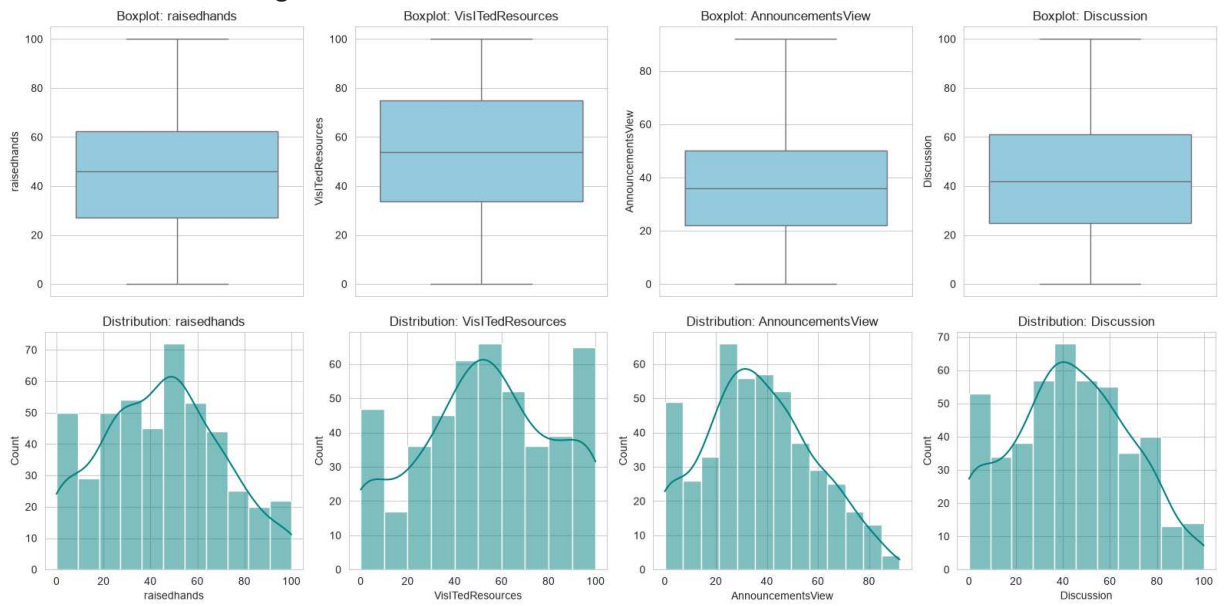
gender	0
NationalITY	0
PlaceofBirth	0
StageID	0

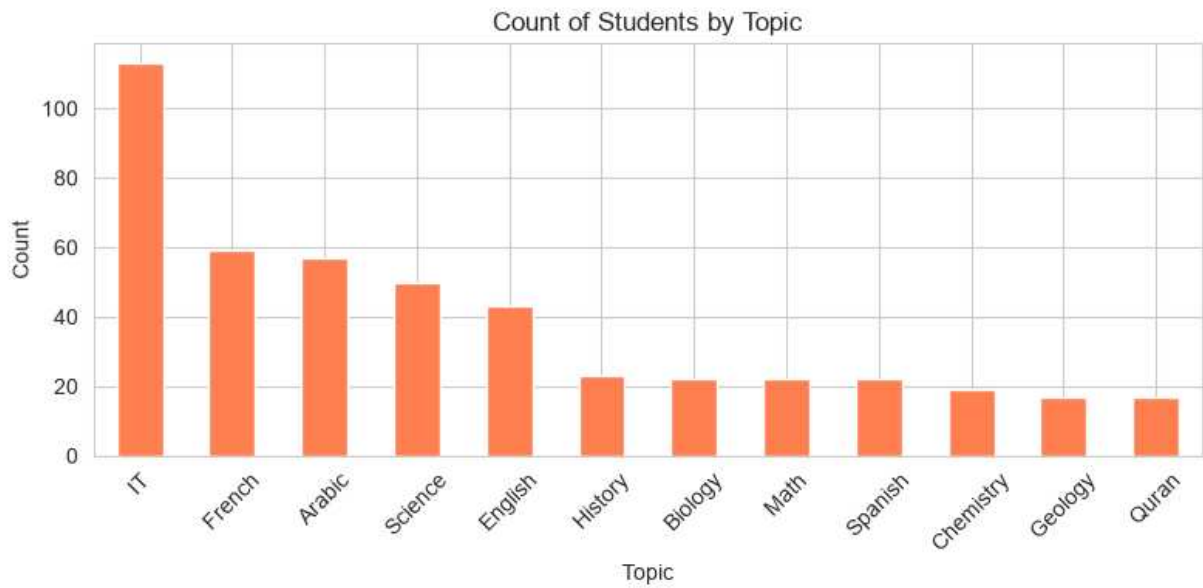
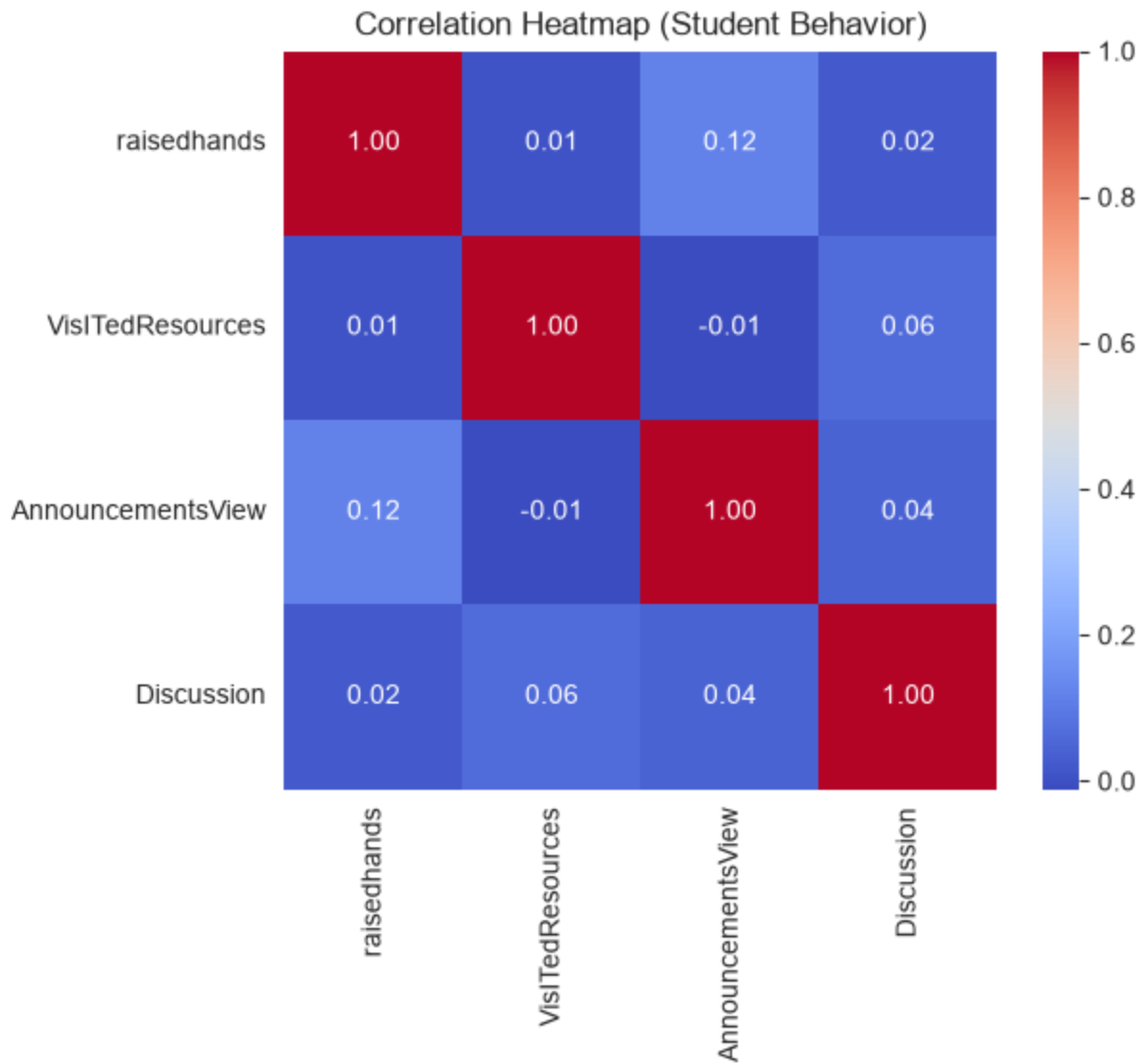
```

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
    
```

Rows before removing outliers: 480

Rows after removing outliers: 464





Saved cleaned dataset successfully.

In []: