Teenage Pregnancy in Bungoma County – Exploratory Data Analysis (EDA) Report

Analyst: Elizabeth Nyadimo

Organization: Elevate Analytics & Consulting

Date: June 2025

Teenage pregnancy remains a pressing public health and social issue in Bungoma County. While national and regional efforts aim to improve youth reproductive health, this analysis seeks to understand local-level dynamics. The goal is to uncover whether access to services and programs translates into reduced teenage pregnancy rates — and if not, where to look next.

STEP 1: Import Necessary Libraries

In [70]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

sns.set(style="whitegrid")

# Apply custom yellow-green color palette
custom_palette = ["#FFD700", "#32CD32"]  # Yellow & LimeGreen
sns.set_palette(custom_palette)
```

STEP 2: Read Dataset

In [58]:
```
df=pd.read_excel("Bungoma Teenage Pregnancy Raw Data.xlsx")
```

a) Display the top 5 records of the dataset

In [59]:
```
df.head()
```

Out[59]:

	SubCounty	Ward	Village	TeenPregnancyRate	SchoolDropouts	HealthCenters	C
0	Mt Elgon	Cheptais	Kamosong	16.4	20	2	
1	Mt Elgon	Cheptais	Chebwek	21.5	18	1	
2	Mt Elgon	Cheptais	Chepkube	24.1	22	0	
3	Mt Elgon	Cheptais	Kipsis	17.8	19	2	
4	Mt Elgon	Cheptais	Kapsesoi	19.8	17	0	

b) .Display the last 5 records of the dataset

In [60]:
```
df.tail()
```

Out[60]:

	SubCounty	Ward	Village	TeenPregnancyRate	SchoolDropouts	HealthC
766	Tongaren	Soysambu/Mitua	Narati	24.2	11	
767	Tongaren	Soysambu/Mitua	Maresi Maket	17.1	12	
768	Tongaren	Soysambu/Mitua	James Mwei	13.9	10	
769	Tongaren	Soysambu/Mitua	Matisi	24.0	11	
770	Tongaren	Soysambu/Mitua	Brigadia Makutano	16.7	12	

STEP 3: Data Sanity Check

a) Check Shape

In [61]:
```
df.shape
```

Out[61]:
```
(771, 8)
```

b) Check Information

In [62]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 771 entries, 0 to 770
Data columns (total 8 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   SubCounty            771 non-null    object
 1   Ward                 771 non-null    object
 2   Village              771 non-null    object
 3   TeenPregnancyRate    771 non-null    float64
 4   SchoolDropouts       771 non-null    int64
 5   HealthCenters        771 non-null    int64
 6   ContraceptiveAccess  771 non-null    object
 7   EducationPrograms    771 non-null    object
dtypes: float64(1), int64(2), object(5)
memory usage: 48.3+ KB
```

c) Finding Missing Values

In [63]:
```
df.isnull().sum()
```

Out[63]:
```
SubCounty              0
Ward                   0
Village                0
TeenPregnancyRate      0
SchoolDropouts         0
HealthCenters          0
ContraceptiveAccess    0
EducationPrograms      0
dtype: int64
```

d) Finding duplicates

In [64]:
```
df.duplicated().sum()
```

Out[64]:
```
0
```

STEP 4: Exploratory Data Analysis (EDA)

a) Descriptive Statistics

Numerical columns

In [65]:
```
df.describe().T
```

Out[65]:

	count	mean	std	min	25%	50%	75%	max
TeenPregnancyRate	771.0	19.258366	3.922952	12.5	15.9	19.5	22.7	26.0
SchoolDropouts	771.0	12.822309	2.587848	10.0	11.0	12.0	14.0	24.0
HealthCenters	771.0	1.024643	0.815859	0.0	0.0	1.0	2.0	2.0

Teenage pregnancy is widespread, averaging 19.3%, with some villages experiencing rates as high as 26%.

School dropout rates also average 12.8%, with a noticeable spread.

Health center availability is limited — nearly 25% of villages have no facility at all.

Object columns

In [66]:
```
df.describe(include="object").T
```

Out[66]:

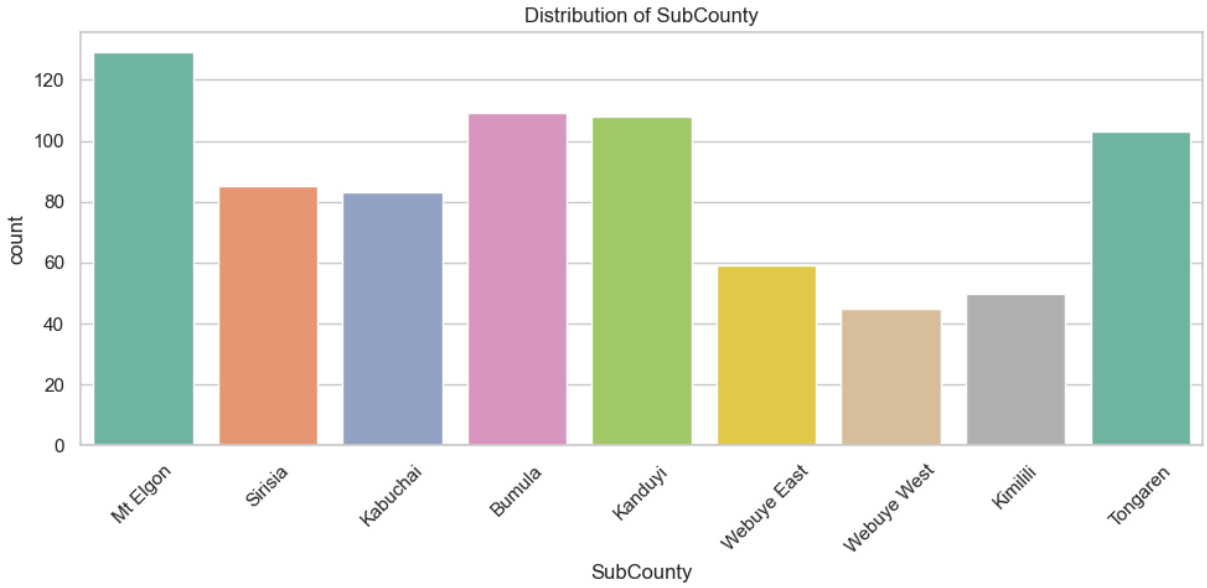
	count	unique	top	freq
SubCounty	771	9	Mt Elgon	129
Ward	771	45	Lwandanyi	34
Village	771	659	Sango	6
ContraceptiveAccess	771	2	No	486
EducationPrograms	771	2	No	501

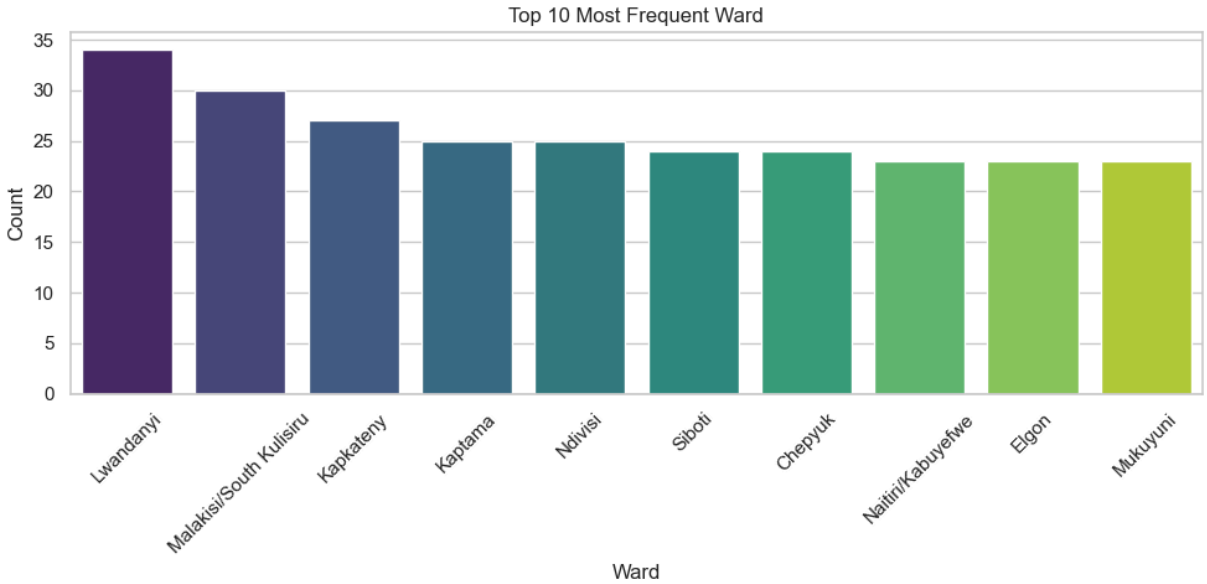
While most villages lack contraceptive access and youth-focused education programs, some with these services still show high pregnancy rates.

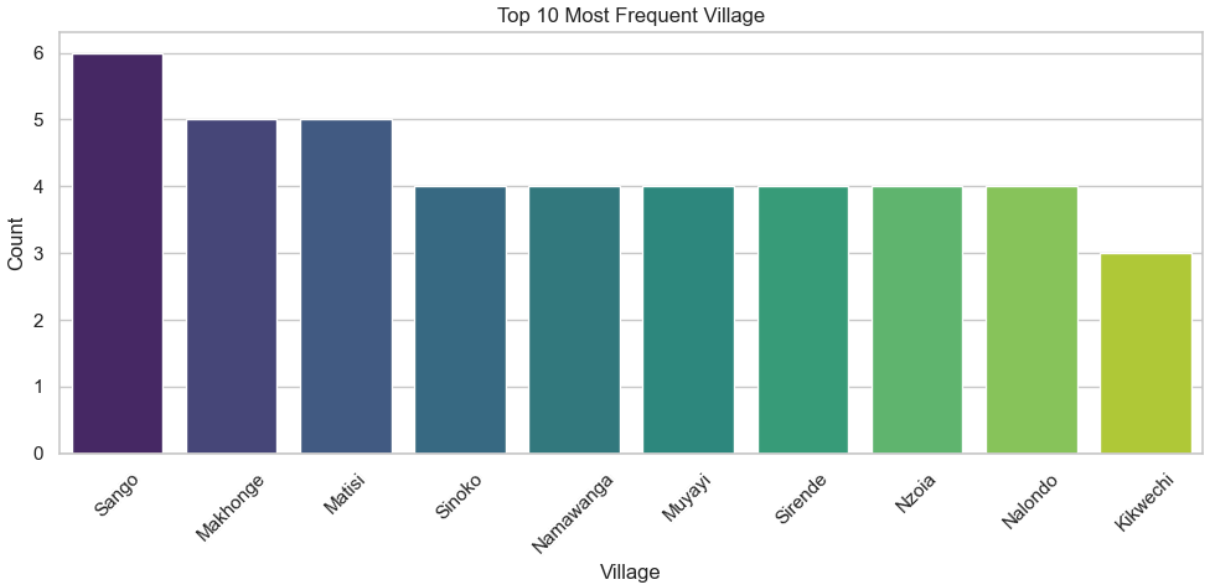
b) Univariate Analysis

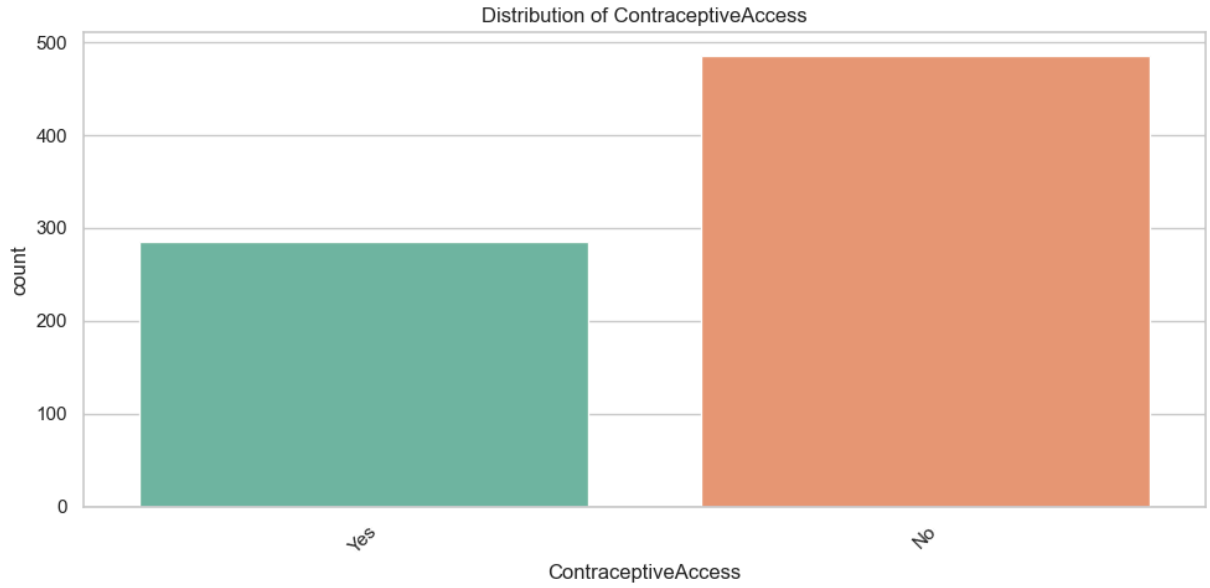
Categorical features (Object Columns)

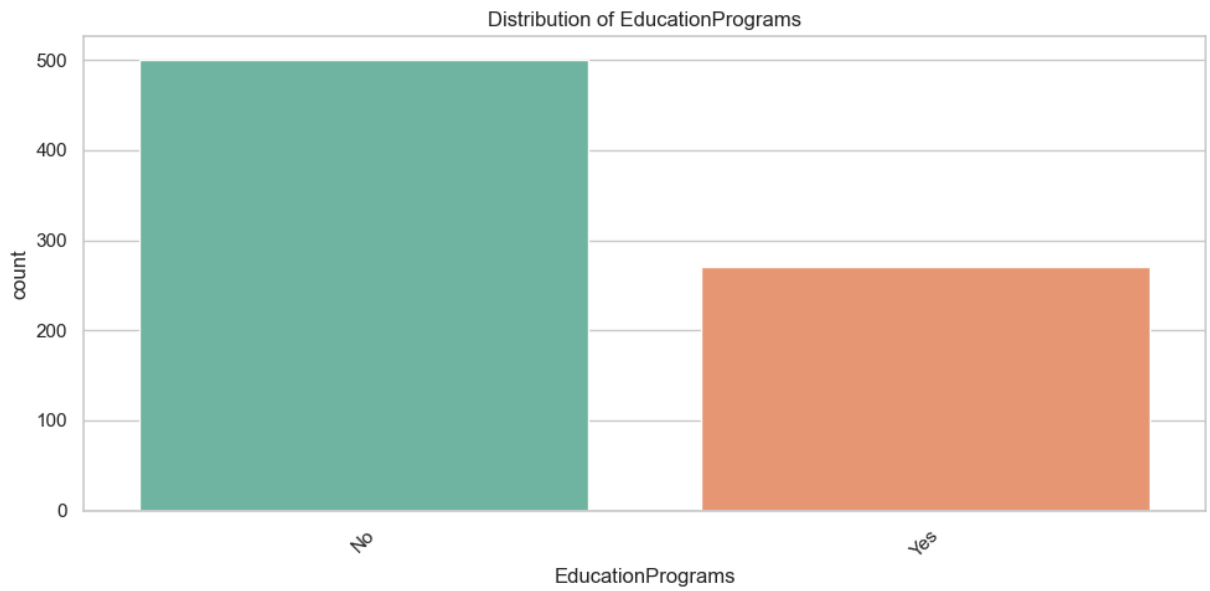
In [81]:
```
categorical_cols = ['SubCounty', 'Ward', 'Village', 'ContraceptiveAccess', 'Educati

for col in categorical_cols:
    plt.figure(figsize=(10, 5))

    # If the column has many unique categories, show top 10
    if df[col].nunique() > 15:
        top_categories = df[col].value_counts().nlargest(10)
        sns.barplot(x=top_categories.index, y=top_categories.values, palette='virid
        plt.title(f'Top 10 Most Frequent {col}')
        plt.ylabel('Count')
        plt.xlabel(col)
        plt.xticks(rotation=45)
    else:
        sns.countplot(x=col, data=df, palette='Set2')
        plt.title(f'Distribution of {col}')
        plt.xticks(rotation=45)

    plt.tight_layout()
    plt.show()
```


Distribution of SubCounty
count
0
20
40
60
80
100
120
Mt Elgon
Sirisia
Kabuchai
Bumula
Kanduyi
Webuye East
Webuye West
Kimilili
Tongaren
SubCounty


Top 10 Most Frequent Ward
Count
0
5
10
15
20
25
30
35
Lwandanyi
Malakisi/South Kulisiru
Kapkateny
Kaptama
Ndivisi
Siboti
Chepyuk
Naitiri/Kabuyefwe
Elgon
Mukuyuni
Ward

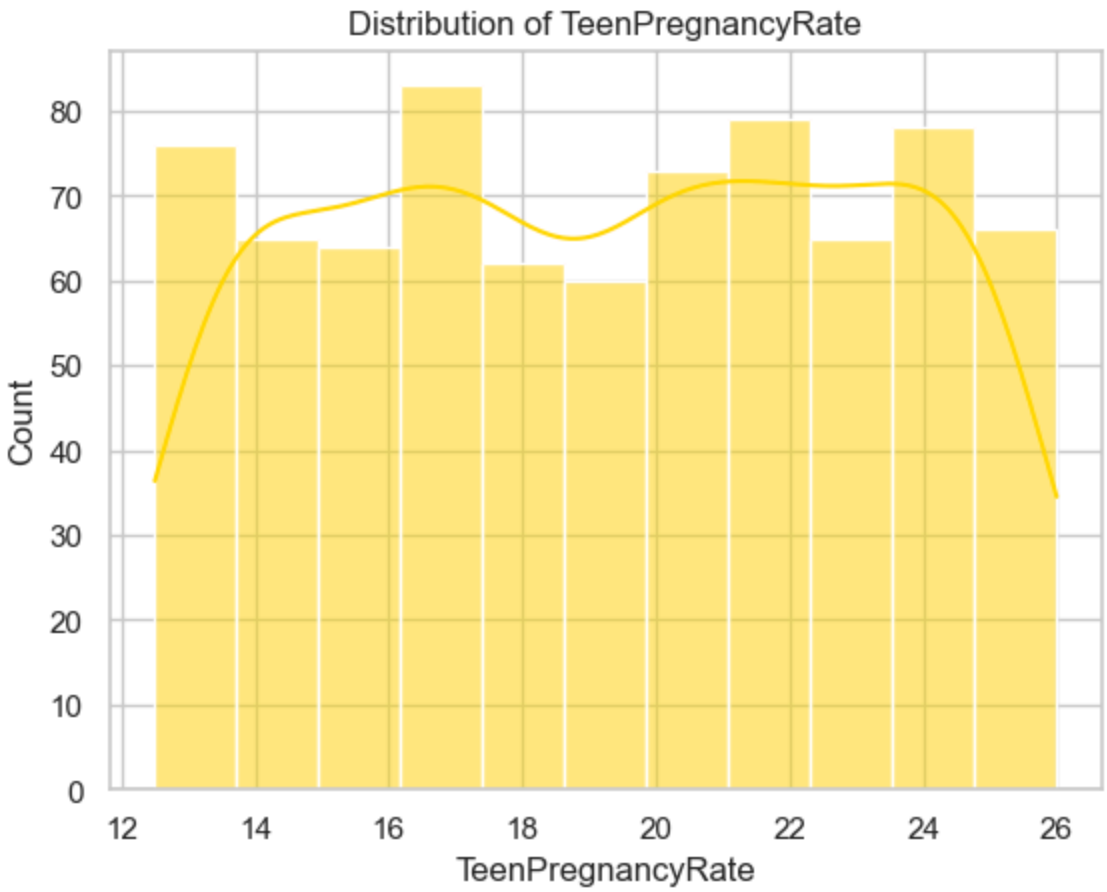

Top 10 Most Frequent Village
Count
0
1
2
3
4
5
6
Sango
Makhonge
Matisi
Sinoko
Namawanga
Muyayi
Sirende
Nzoia
Nalondo
Kikwechi
Village

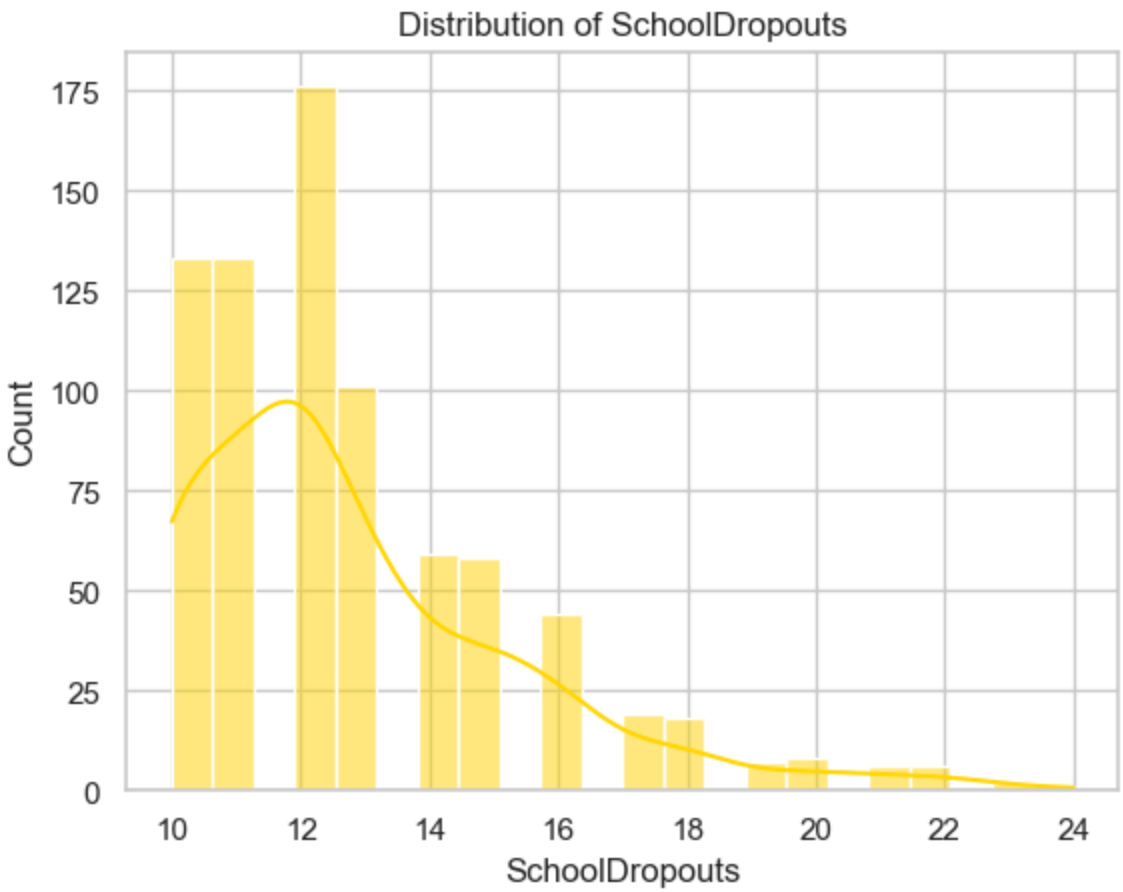


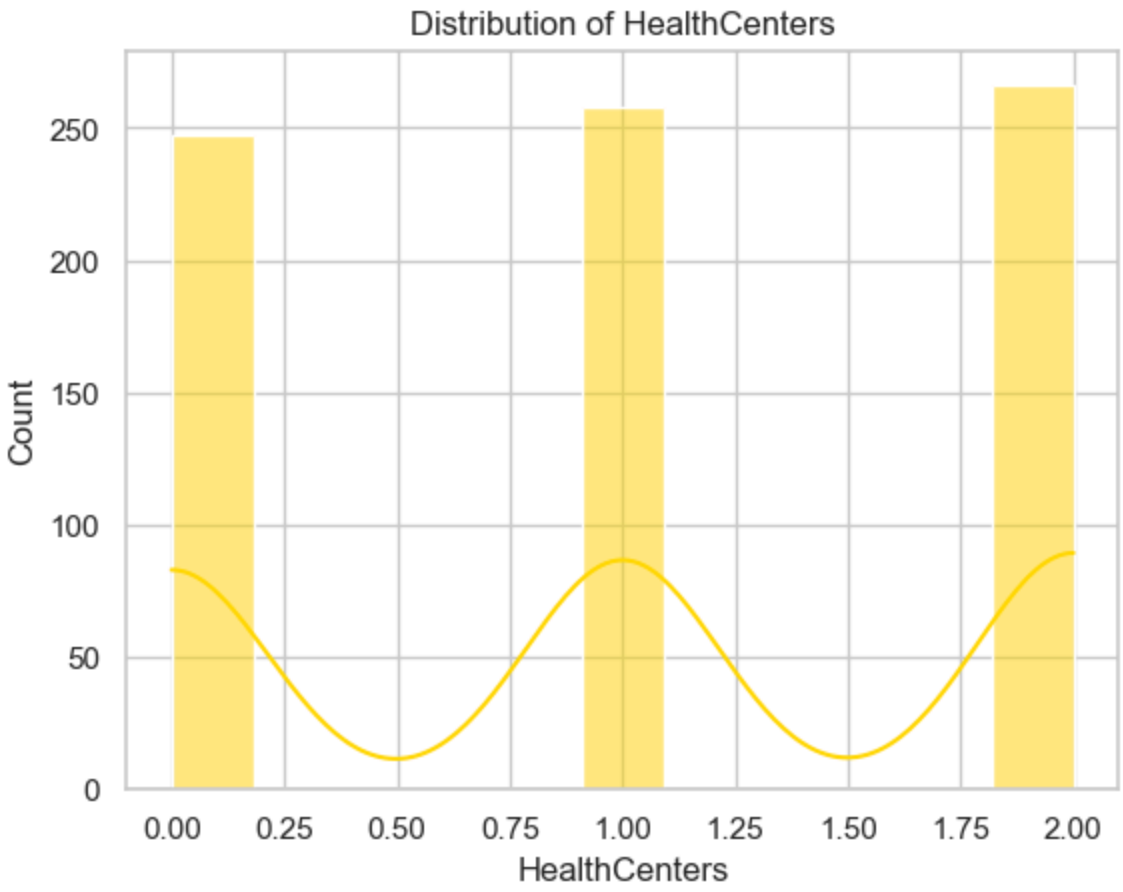


Numerical features

In [73]:
```
# List of numerical columns
numerical_cols = ['TeenPregnancyRate', 'SchoolDropouts', 'HealthCenters']

for col in numerical_cols:
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')
    plt.show()
```


Distribution of TeenPregnancyRate
Count
0
10
20
30
40
50
60
70
80
12
14
16
18
20
22
24
26
TeenPregnancyRate

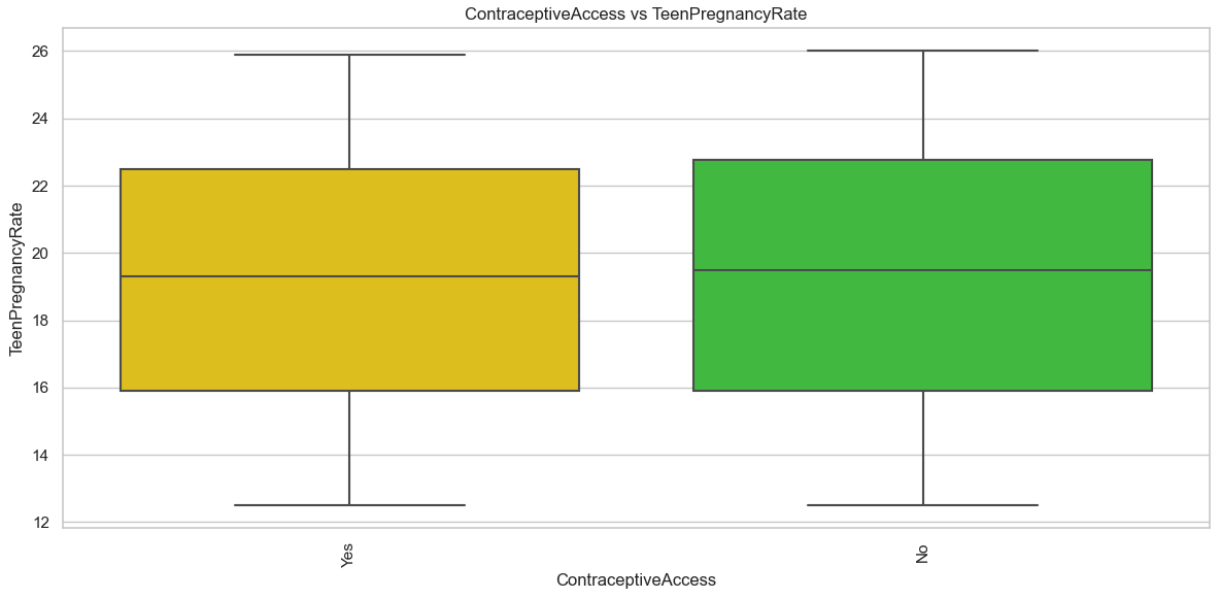

Distribution of SchoolDropouts
Count
0
25
50
75
100
125
150
175
10
12
14
16
18
20
22
24
SchoolDropouts

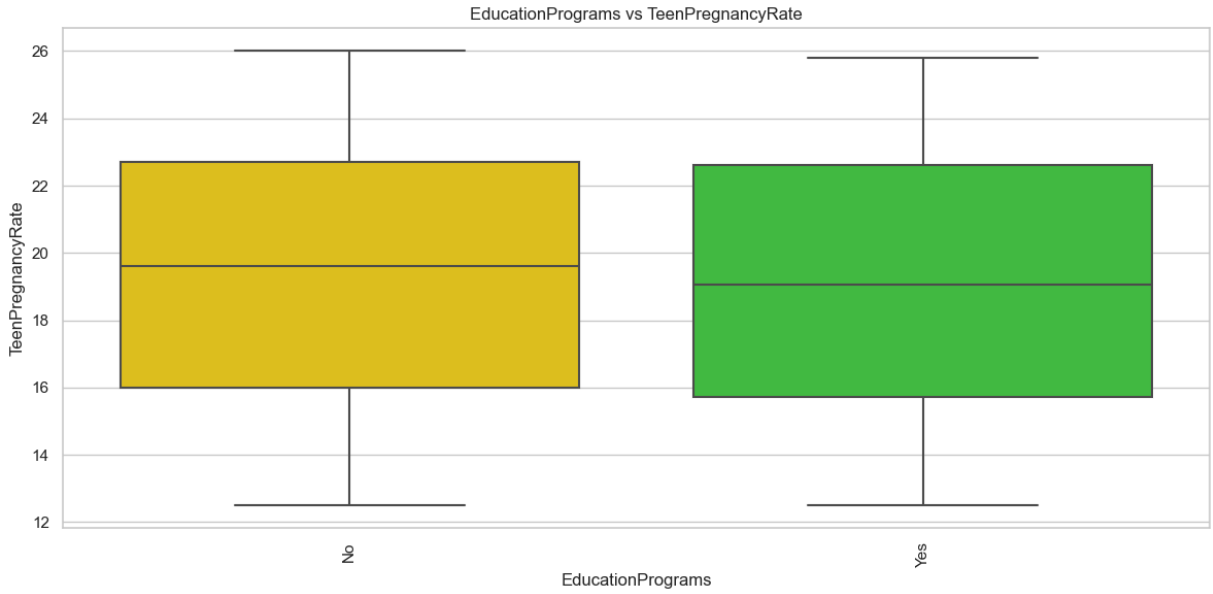


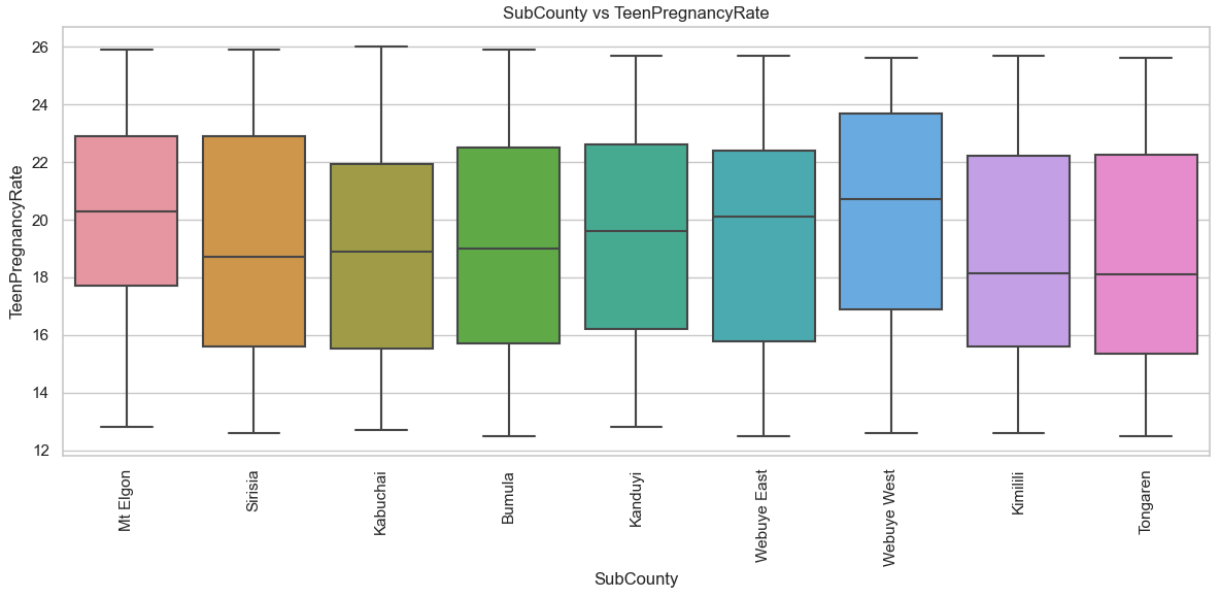
c)Bivariate Analysis

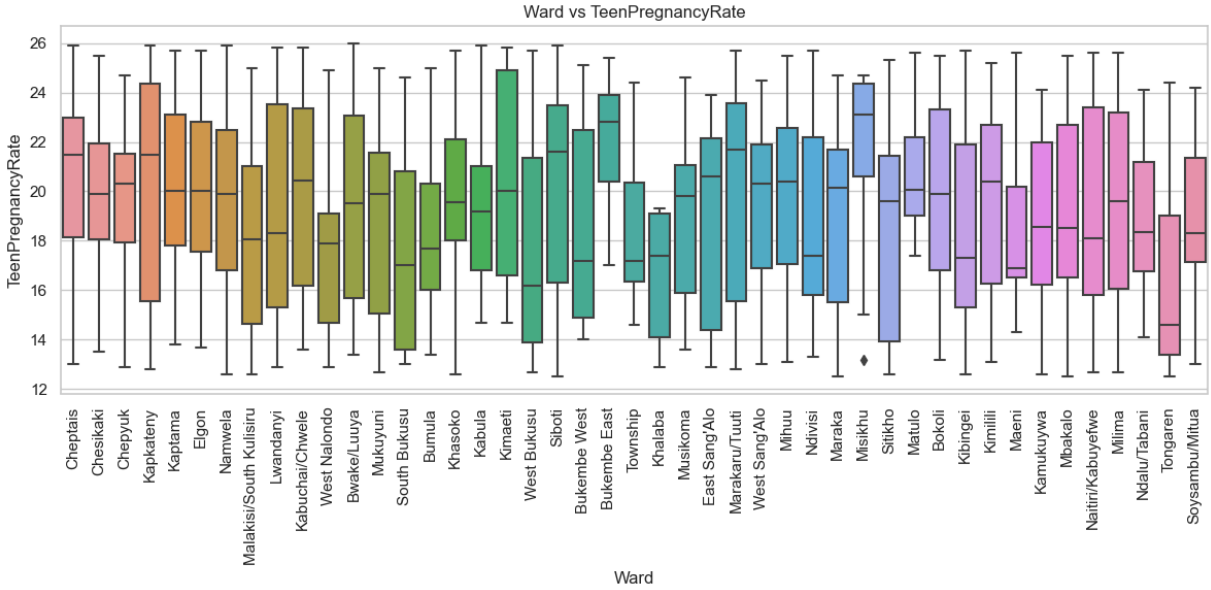
Categorical vs Target (TeenPregnancyRate)

In [84]:
```
categorical_cols = ['ContraceptiveAccess', 'EducationPrograms', 'SubCounty', 'Ward'

for col in categorical_cols:
    plt.figure(figsize=(12, 6 ))
    sns.boxplot(x=col, y='TeenPregnancyRate', data=df)
    plt.title(f'{col} vs TeenPregnancyRate')
    plt.xticks(rotation=90)
    plt.tight_layout()
    plt.show()
```

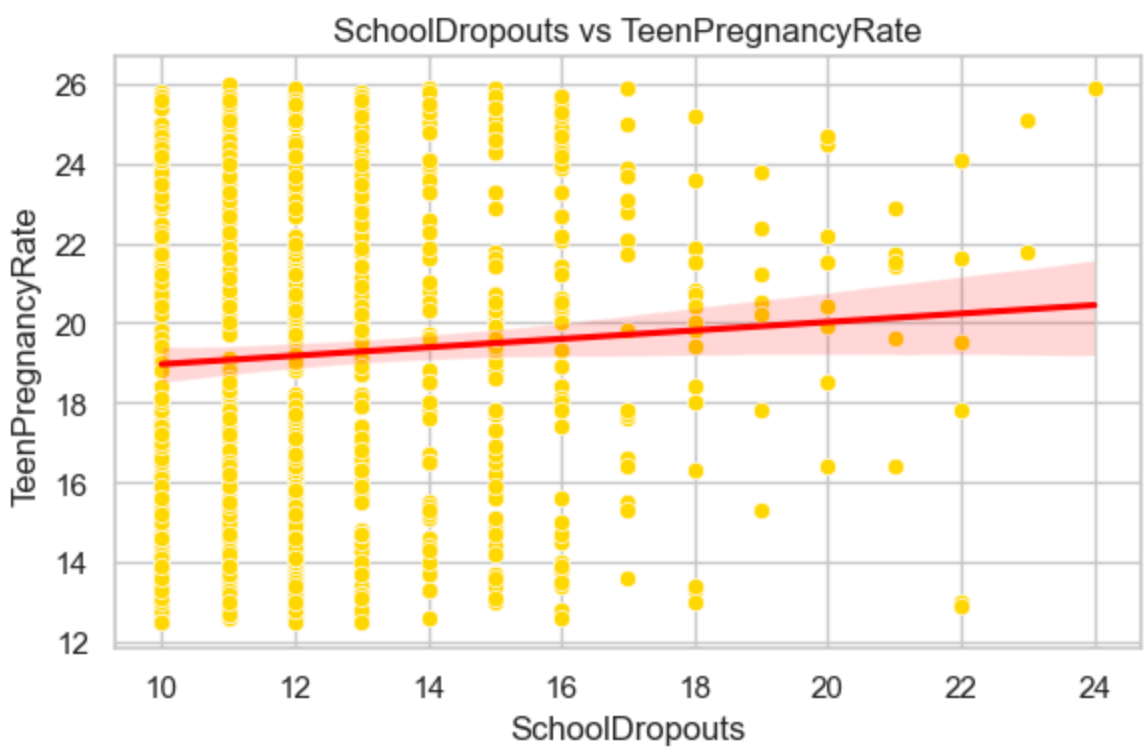

ContraceptiveAccess vs TeenPregnancyRate
TeenPregnancyRate
26
24
22
20
18
16
14
12
Yes
No
ContraceptiveAccess

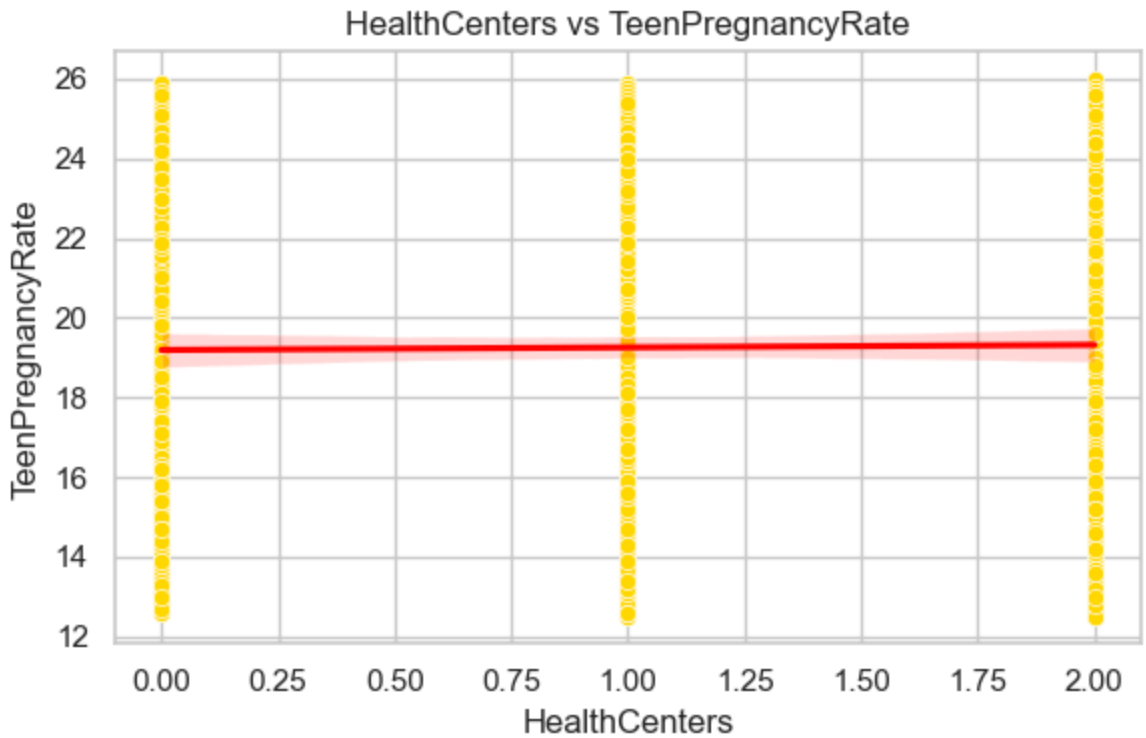

EducationPrograms vs TeenPregnancyRate
TeenPregnancyRate
26
24
22
20
18
16
14
12
No
Yes
EducationPrograms


SubCounty vs TeenPregnancyRate
TeenPregnancyRate
26
24
22
20
18
16
14
12
Mt Elgon
Sirisia
Kabuchai
Bumula
Kanduyi
Webuye East
Webuye West
Kimilili
Tongaren
SubCounty



In [86]:
```
numerical_cols = ['SchoolDropouts', 'HealthCenters']
for col in ['SchoolDropouts', 'HealthCenters']:
    plt.figure(figsize=(6, 4))
    sns.scatterplot(x=col, y='TeenPregnancyRate', data=df)
    sns.regplot(x=col, y='TeenPregnancyRate', data=df, scatter=False, color='red')
    plt.title(f'{col} vs TeenPregnancyRate')
    plt.tight_layout()
    plt.show()
```



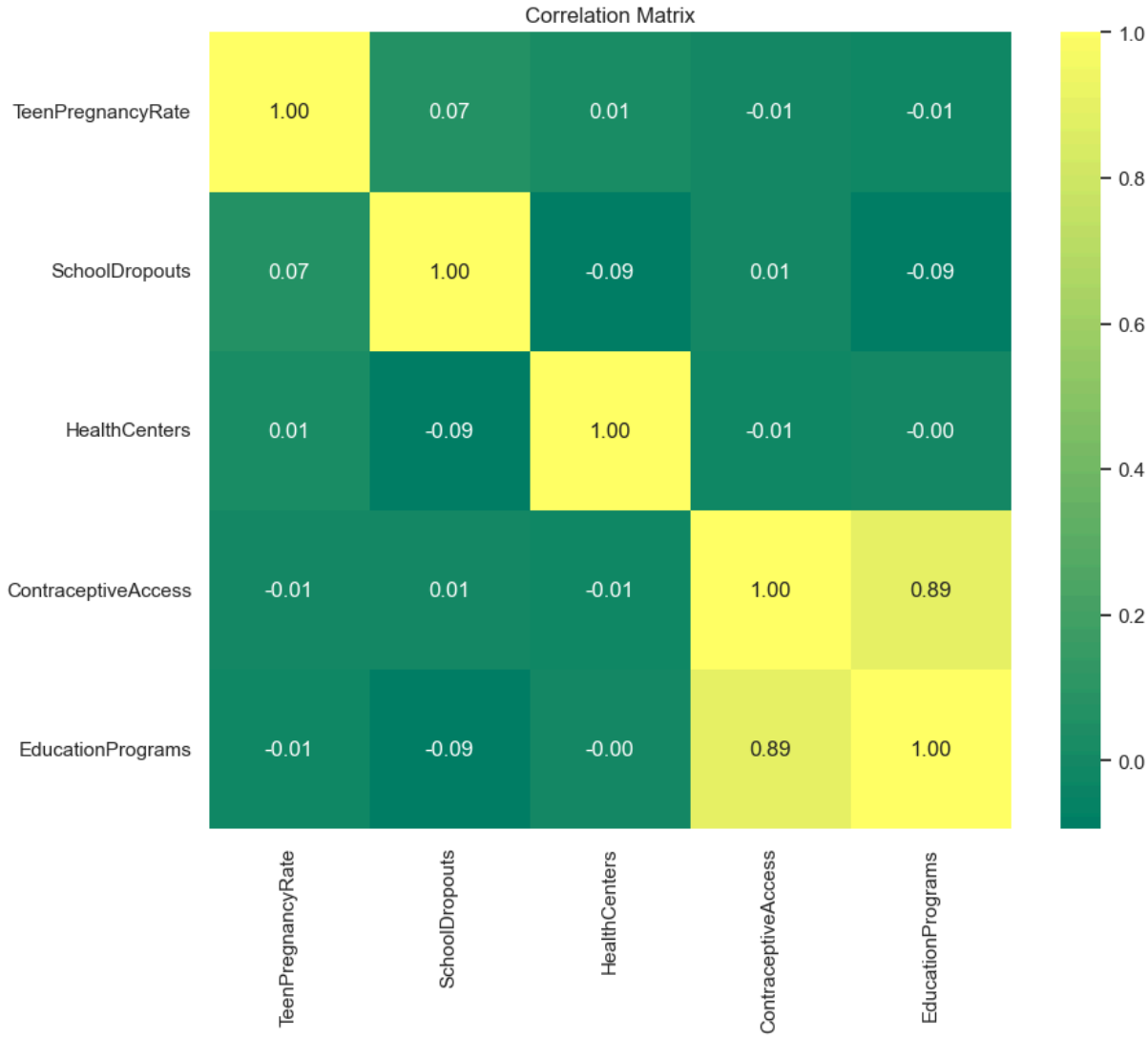


d)Correlation Analysis

In [93]:
```
#Encode Yes/No Before Correlation
df['ContraceptiveAccess'] = df['ContraceptiveAccess'].map({'Yes': 1, 'No': 0})
df['EducationPrograms'] = df['EducationPrograms'].map({'Yes': 1, 'No': 0})
```

In [94]:
```
corr_matrix = df[['TeenPregnancyRate', 'SchoolDropouts', 'HealthCenters',
                  'ContraceptiveAccess', 'EducationPrograms']].corr()
```

In [100…
```
# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap="summer", fmt=".2f")
plt.title('Correlation Matrix')
plt.show()
```



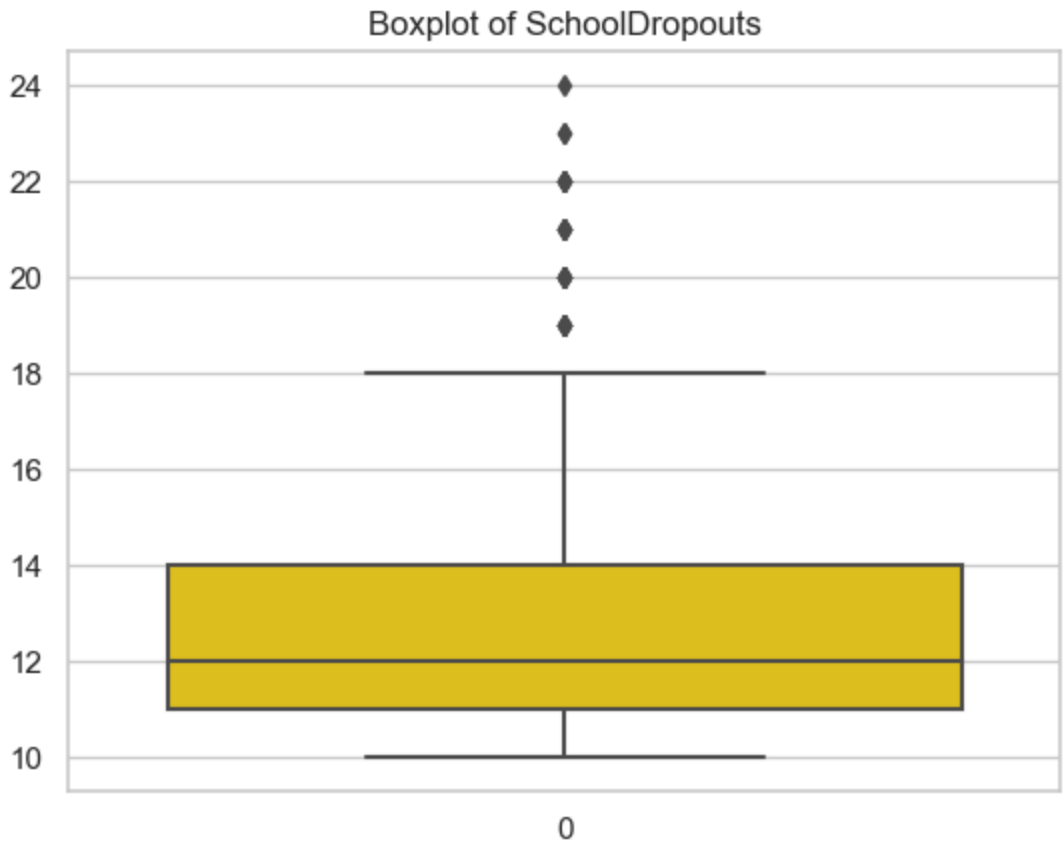
None of the structural or programmatic variables show a strong or even moderate correlation with teenage pregnancy.

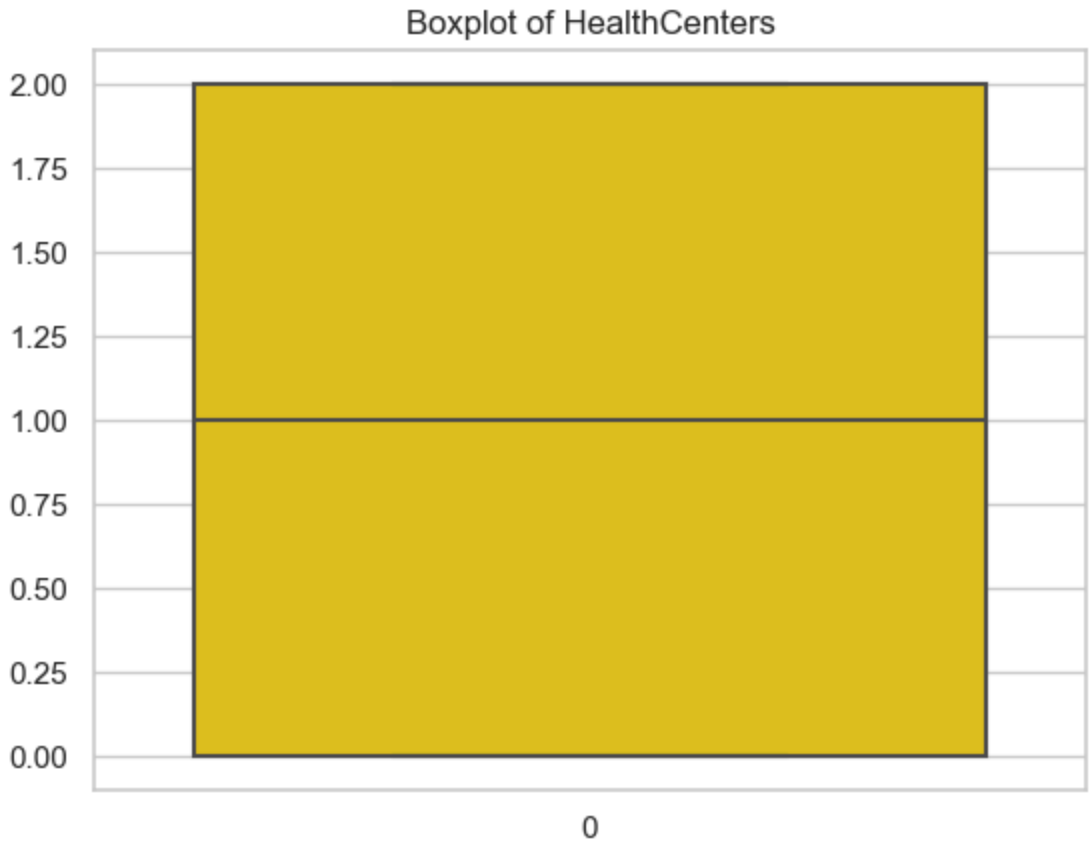
School dropouts show only a very weak positive association, suggesting potential co-occurrence rather than causation.

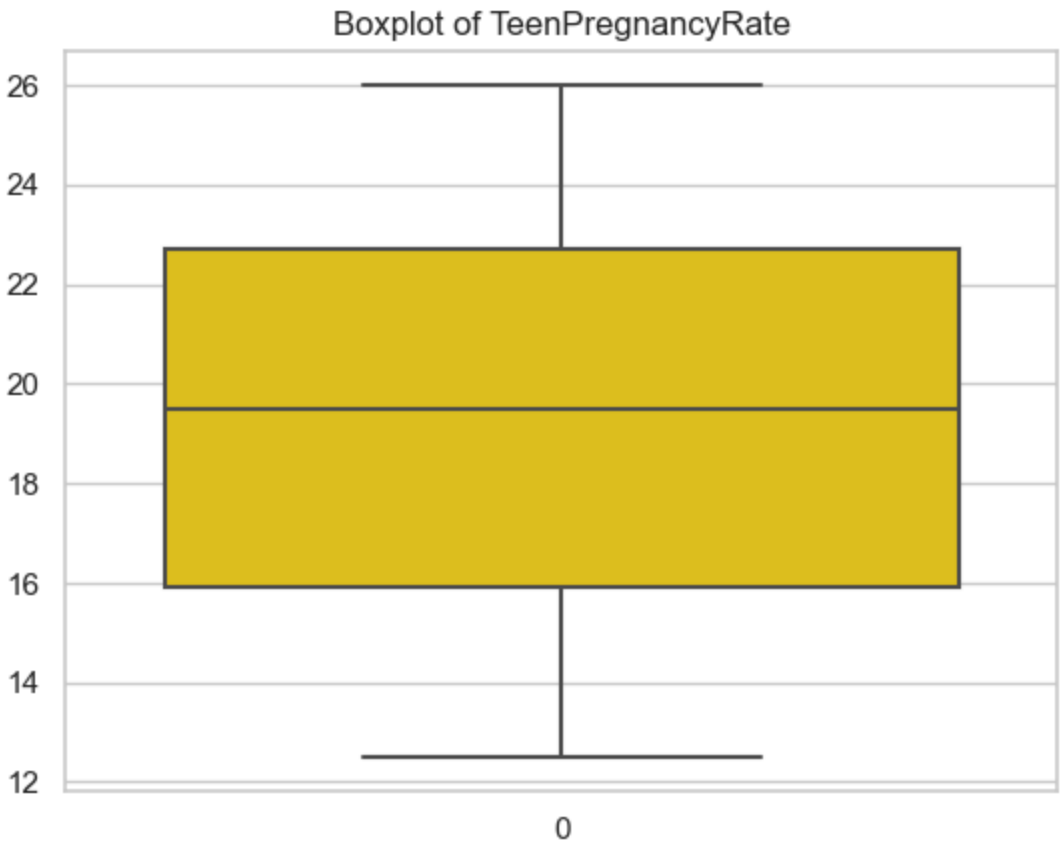
e) Outliers & Insights

In [101…

```
for col in numerical_cols:
    sns.boxplot(data=df[col])
    plt.title(f'Boxplot of {col}')
    plt.show()
```


Boxplot of SchoolDropouts
24
22
20
18
16
14
12
10
0


Boxplot of HealthCenters
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0



What the Data Is Telling Us

Widespread Issue, Poorly Explained

Even villages with health centers, contraceptives, and education programs report high teen pregnancy. This suggests:

- Structural presence ≠ Utilization or effectiveness
- The problem is deeper than service availability

Dropout Rates Are Not the Whole Story

With only **+0.07 correlation**:

- Dropouts and pregnancy may co-occur due to shared root causes like poverty, early marriage, or lack of aspiration — not necessarily causality

---

Health Centers Might Not Be Adolescent-Friendly

- A facility's existence doesn't guarantee confidentiality, non-judgmental care, or youth-specific services.

Contraceptive and Education Programs Lack Depth

Presence alone isn't enough. Are these programs:

- Age-appropriate?
- Community-accepted?
- Delivered by trained staff?
- Monitored for quality?

Recommendations

Gap Identified	Recommendation
Weak link between presence & outcomes	Conduct qualitative assessments (e.g., FGDs, interviews) on barriers to usage
Unclear program impact	Evaluate program quality, delivery, and youth engagement
Cultural/social drivers suspected	Engage parents, religious leaders, elders, and peer influencers
Lack of holistic approach	Include mental health, mentorship, life skills, and economic empowerment

---

What to Explore Next

To go beyond surface-level indicators, future data collection should include:

- Household income and poverty levels
- Parental involvement and support
- Peer pressure and relationships
- Age at first sexual encounter
- Social norms and beliefs about early marriage and gender roles

Insight: **The data shows the map — not the full terrain. It reveals where to dig deeper, not necessarily why things happen.**

Strategic Next Steps

Action Area	What to Do
Deep-Dive Assessments	Case studies, ethnographic research, community storytelling
Improve Program Quality	Evaluate content, staff capacity, delivery methods

Action Area	What to Do
Tailored Interventions	Adapt strategies to specific wards and villages
Youth Engagement	Let young people co-create solutions and be part of decision-making

Final Takeaway

> "Availability does not equal access, and access does not guarantee impact."

This analysis challenges assumptions that services alone reduce teenage pregnancy. It calls for **multi-dimensional strategies** that address the **social, behavioral, and psychological** aspects of adolescent life in Bungoma.

✅ **Done by: Elizabeth Nyadimo**

📊 **For: Elevate Analytics & Consulting**

📅 **Date: June 2025 | 📍 Bungoma County, Kenya**

"Turning Data Into Impact."

No description has been provided for this image

In [ ]: