



ALLIANCE
UNIVERSITY

*Private University established in Karnataka State by Act No.34 of year 2010
Recognized by the University Grants Commission (UGC), New Delhi*

Project Report

Statistic For Data Science

Semester – 2

“Youth smoking and drug”

By

Laishram Ritikumar Singh

Reg no: 2411021240040

GitHub link: <https://github.com/Ritikumar2007/SDS-project>

Department of Computer Application

Alliance University Chandapura — Anekal Main Road,

Anekal Bengaluru — 562 106

April 2025

Project Overview

This project investigates youth behavior in relation to smoking and drug usage by analyzing a dataset comprising responses from teenagers aged 12 to 17. The dataset, which includes variables such as age, gender, smoking habits, and drug usage patterns, aims to uncover correlations and trends that can inform preventative strategies and public health policies. Through exploratory data analysis and visualizations, the project seeks to understand how various demographic and behavioral factors are linked to substance use among adolescents.

Challenges

1. Data Cleaning & Preprocessing:

- a. The dataset required handling missing values and converting categorical data into a suitable format for analysis.
- b. Some variables needed to be restructured for consistency and clarity.

2. Data Imbalance:

- a. Certain behaviors (e.g., hard drug use) were underrepresented, making trend detection and inference more challenging.

3. Interpretability:

- a. Understanding the nuanced relationships between multiple behavioral and demographic variables involved subjective interpretation.

4. Limited Contextual Data:

- a. The dataset did not include contextual or environmental factors (e.g., socioeconomic status, peer pressure) that could influence youth behavior.

Introduction

Adolescence is a critical developmental period, often marked by experimentation and identity formation. Unfortunately, it is also a stage where many young individuals are exposed to smoking and drug use. Understanding the patterns and factors contributing to

substance use among youth is crucial for designing targeted intervention programs. This project utilizes a dataset containing survey responses from teenagers aged 12 to 17, aiming to explore patterns in smoking and drug behavior and provide actionable insights into the factors influencing these behaviors.

Project Goals

- To analyze the prevalence of smoking and drug use among adolescents.
- To identify age and gender trends in substance use.
- To explore the correlation between smoking and the use of other substances.
- To provide visualizations that aid in understanding youth behavior patterns.
- To derive insights that can help inform policymakers, educators, and health professionals.

Conclusion

The analysis revealed notable trends in youth smoking and drug use, with older teenagers showing a higher propensity for both behaviors. Gender-based differences were also observed, along with a strong correlation between smoking and the likelihood of using other substances. These findings highlight the importance of early prevention strategies and the need for gender-sensitive approaches in intervention programs. While the dataset provides valuable insights, integrating additional contextual factors could enhance the depth and applicability of future research.

```
#data set of youth smoking and drug
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv(r"C:\Users\abung\Downloads\Youth smoking SDS.csv")
df
```

	Year	Age_Group	Gender	Smoking_Prevalence	Drug_Experimentation
\					
0	2024	15-19	Both	18.85	32.40
1	2024	Oct-14	Female	34.88	41.57
2	2023	Oct-14	Both	42.00	56.80
3	2024	40-49	Both	33.75	42.90
4	2023	15-19	Male	47.90	39.62
...
9995	2023	15-19	Male	49.17	10.21
9996	2020	80+	Female	48.00	30.85
9997	2021	25-29	Both	47.62	39.54
9998	2022	40-49	Male	9.37	11.64
9999	2023	Oct-14	Male	43.77	21.95

	Socioeconomic_Status	Peer_Influence	School_Programs
Family_Background \			
0	High	5	Yes
1			
1	High	6	Yes
10			
2	High	6	Yes
2			
3	Middle	10	No
9			
4	High	1	No
2			
...
...			
9995	Low	7	Yes
4			
9996	Middle	8	Yes
8			
9997	High	1	No

```

7
9998          Low          7          No
10
9999          High          4          Yes
3

      Mental_Health Access_to_Counseling Parental_Supervision \
0          5          No          4
1          5          No          9
2          7          Yes          2
3          7          Yes          2
4          4          Yes          4
...          ...          ...          ...
9995          5          No          7
9996          8          No          4
9997          2          Yes          1
9998          1          No          2
9999          4          Yes          1

```

```

      Substance_Education Community_Support Media_Influence
0          No          3          1
1          Yes          9          3
2          No          5          1
3          No          10         9
4          No          10         3
...          ...          ...          ...
9995          Yes          2          9
9996          Yes          8          9
9997          No          5          10
9998          Yes          10         4
9999          No          6          3

```

```
[10000 rows x 15 columns]
```

```
df.tail()
```

```

      Year Age_Group Gender Smoking_Prevalence Drug_Experimentation
\
9995  2023   15-19   Male          49.17          10.21
9996  2020    80+  Female          48.00          30.85
9997  2021   25-29   Both          47.62          39.54
9998  2022   40-49   Male           9.37          11.64
9999  2023   Oct-14   Male          43.77          21.95

```

```

      Socioeconomic_Status Peer_Influence School_Programs
Family_Background \

```

9995	Low	7	Yes
4			
9996	Middle	8	Yes
8			
9997	High	1	No
7			
9998	Low	7	No
10			
9999	High	4	Yes
3			

	Mental_Health	Access_to_Counseling	Parental_Supervision	\
9995	5	No	7	
9996	8	No	4	
9997	2	Yes	1	
9998	1	No	2	
9999	4	Yes	1	

	Substance_Education	Community_Support	Media_Influence
9995	Yes	2	9
9996	Yes	8	9
9997	No	5	10
9998	Yes	10	4
9999	No	6	3

df.head()

	Year	Age_Group	Gender	Smoking_Prevalence	Drug_Experimentation	\
0	2024	15-19	Both	18.85	32.40	
1	2024	Oct-14	Female	34.88	41.57	
2	2023	Oct-14	Both	42.00	56.80	
3	2024	40-49	Both	33.75	42.90	
4	2023	15-19	Male	47.90	39.62	

	Socioeconomic_Status	Peer_Influence	School_Programs
Family_Background			\
0	High	5	Yes
1			
1	High	6	Yes
10			
2	High	6	Yes
2			
3	Middle	10	No
9			
4	High	1	No
2			

	Mental_Health	Access_to_Counseling	Parental_Supervision	\
0	5	No	4	
1	5	No	9	

2	7	Yes	2
3	7	Yes	2
4	4	Yes	4

	Substance_Education	Community_Support	Media_Influence
0	No	3	1
1	Yes	9	3
2	No	5	1
3	No	10	9
4	No	10	3

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 10000 entries, 0 to 9999
```

```
Data columns (total 15 columns):
```

#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	Year	10000 non-null	int64
1	Age_Group	10000 non-null	object
2	Gender	10000 non-null	object
3	Smoking_Prevalence	10000 non-null	float64
4	Drug_Experimentation	10000 non-null	float64
5	Socioeconomic_Status	10000 non-null	object
6	Peer_Influence	10000 non-null	int64
7	School_Programs	10000 non-null	object
8	Family_Background	10000 non-null	int64
9	Mental_Health	10000 non-null	int64
10	Access_to_Counseling	10000 non-null	object
11	Parental_Supervision	10000 non-null	int64
12	Substance_Education	10000 non-null	object
13	Community_Support	10000 non-null	int64
14	Media_Influence	10000 non-null	int64

```
dtypes: float64(2), int64(7), object(6)
```

```
memory usage: 1.1+ MB
```

```
df.describe()
```

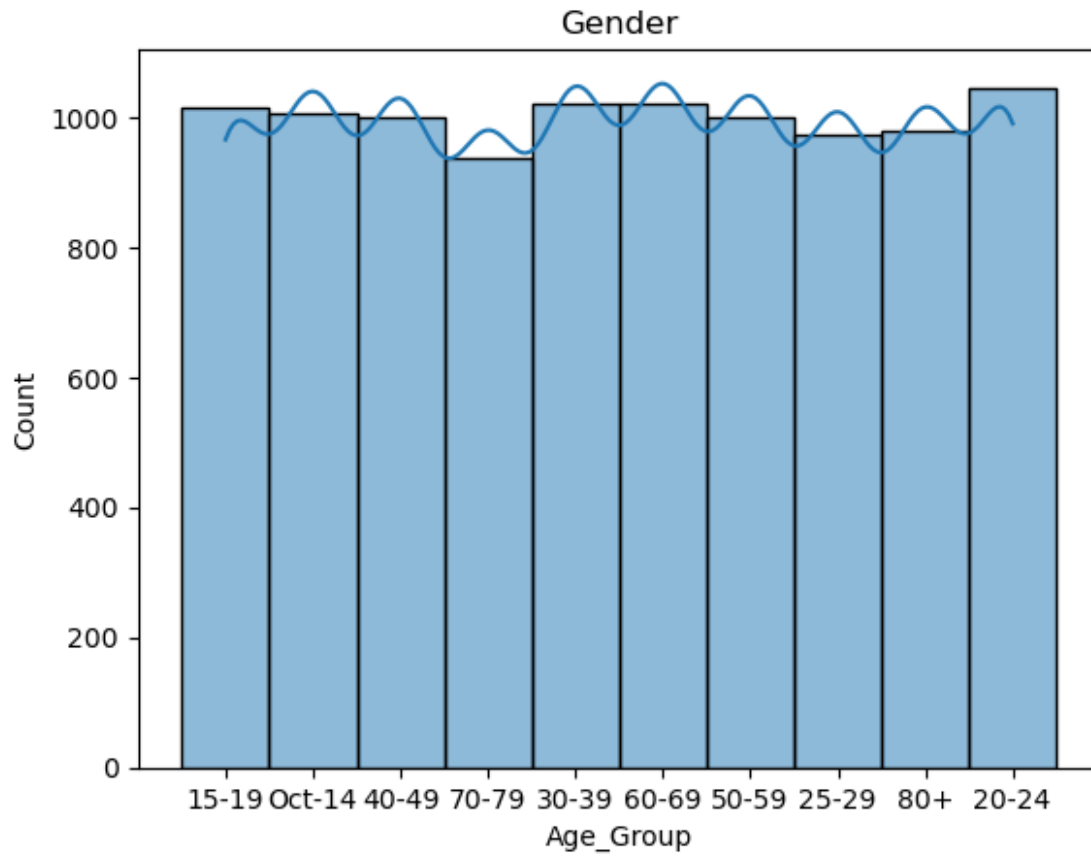
	Year	Smoking_Prevalence	Drug_Experimentation
Peer_Influence \			
count	10000.000000	10000.000000	10000.000000
mean	2022.000500	27.439257	40.150182
std	1.425027	12.975528	17.515917
min	2020.000000	5.000000	10.000000
25%	2021.000000	16.160000	24.920000
3.000000			

50%	2022.000000	27.355000	40.100000
5.000000			
75%	2023.000000	38.672500	55.462500
8.000000			
max	2024.000000	50.000000	69.990000
10.000000			

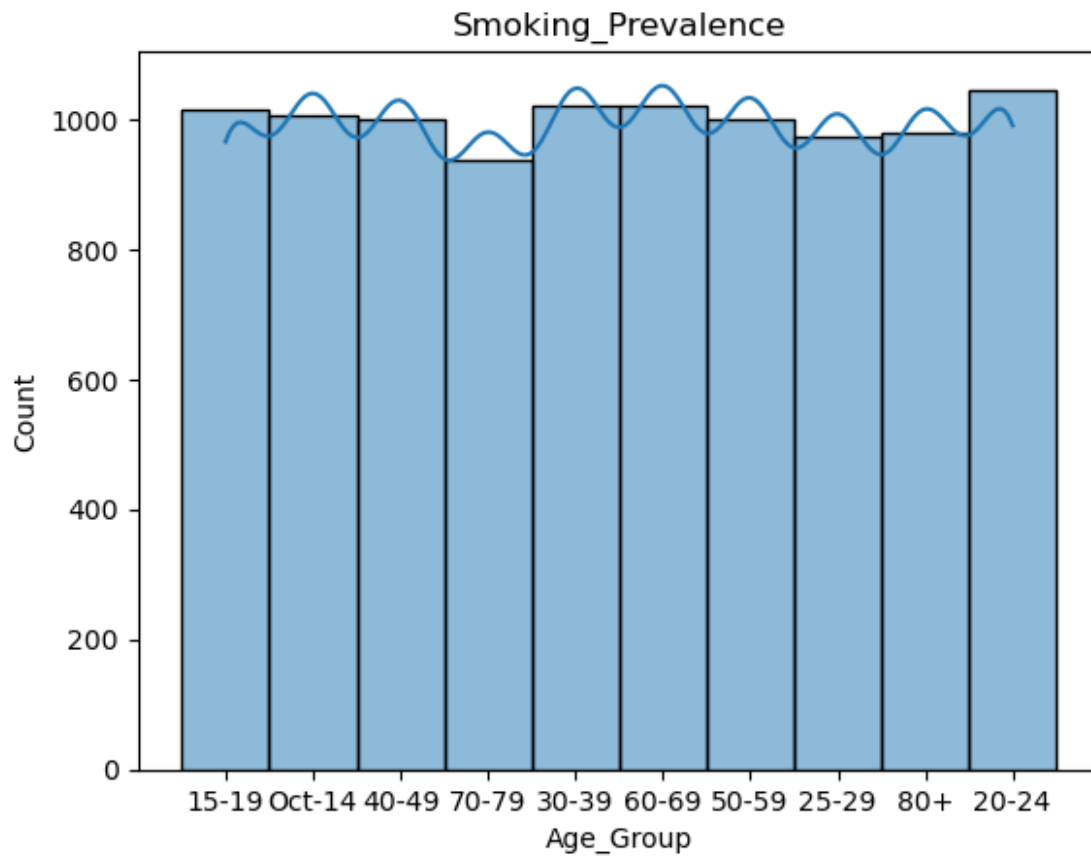
	Family_Background	Mental_Health	Parental_Supervision \
count	10000.000000	10000.000000	10000.000000
mean	5.513300	5.469800	5.528000
std	2.865038	2.879326	2.891514
min	1.000000	1.000000	1.000000
25%	3.000000	3.000000	3.000000
50%	6.000000	5.000000	6.000000
75%	8.000000	8.000000	8.000000
max	10.000000	10.000000	10.000000

	Community_Support	Media_Influence
count	10000.000000	10000.000000
mean	5.544600	5.506200
std	2.870302	2.872836
min	1.000000	1.000000
25%	3.000000	3.000000
50%	6.000000	6.000000
75%	8.000000	8.000000
max	10.000000	10.000000

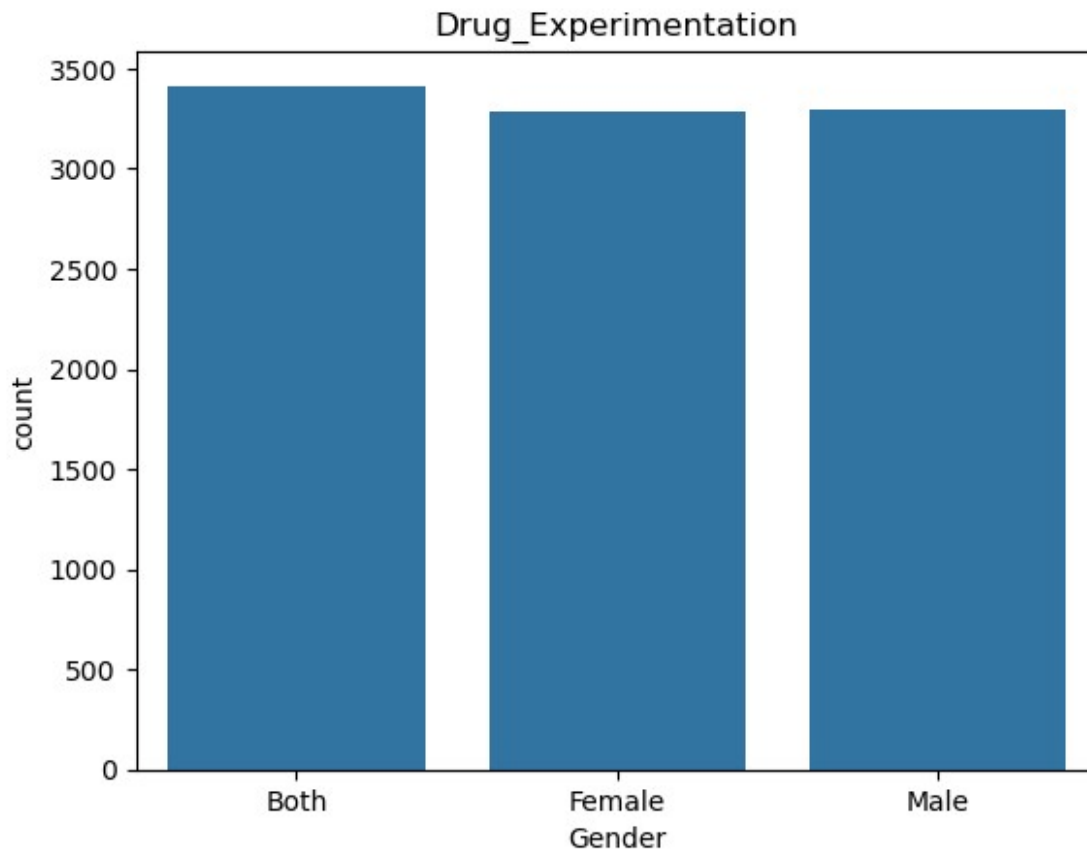
```
# Univariate Analysis: Numerical
sns.histplot(df['Age_Group'], kde=True).set_title('Gender')
plt.show()
```

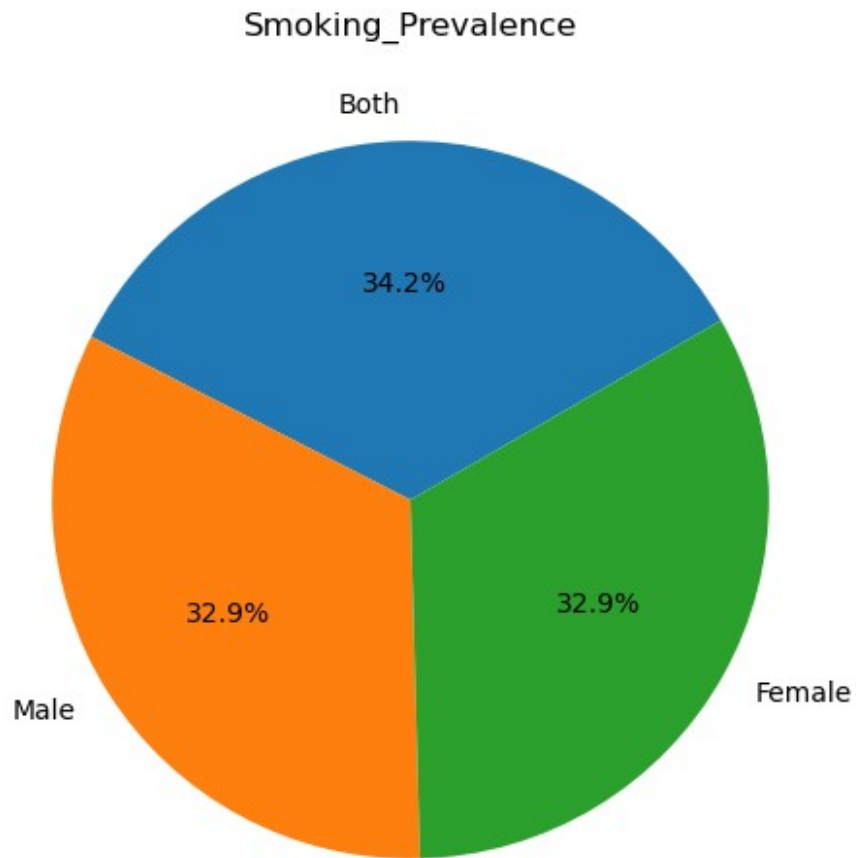
```
# Univariate Analysis: Numerical
sns.histplot(df['Age_Group'],
kde=True).set_title('Smoking_Prevalence')
plt.show()
```



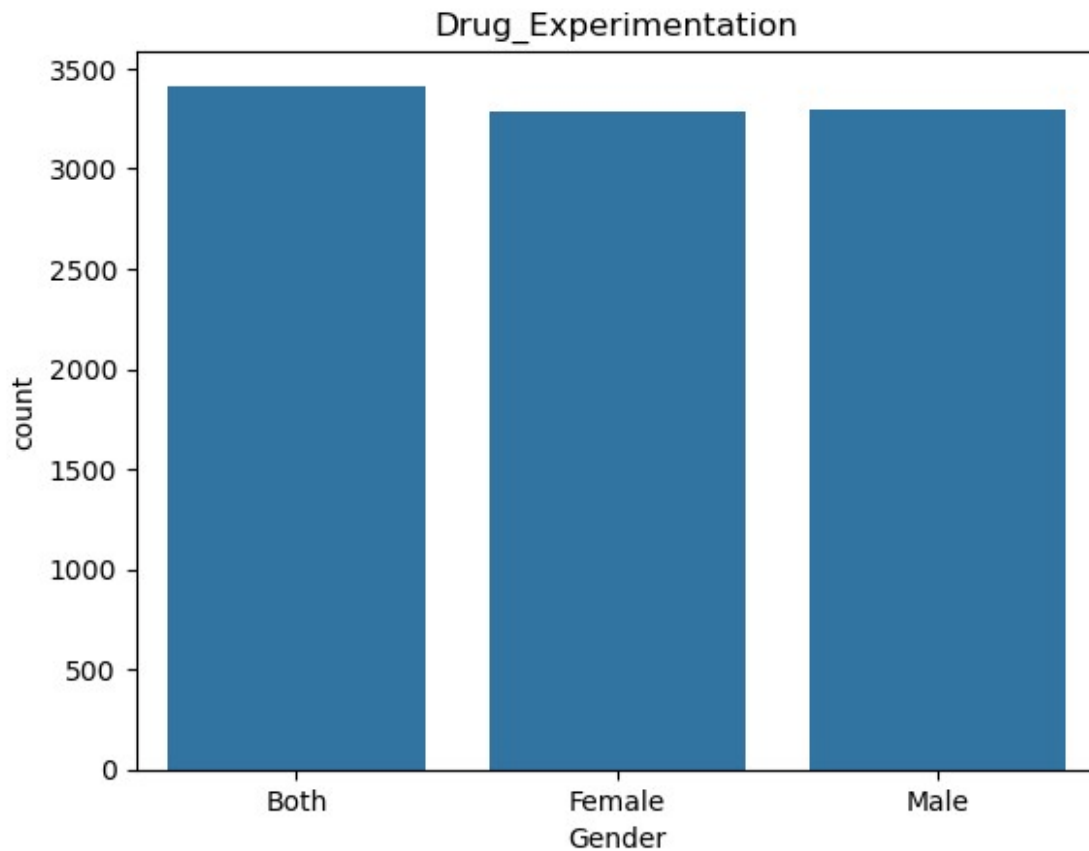
```
# Univariate Analysis: Categorical  
sns.countplot(x='Gender', data=df).set_title('Drug_Experimentation')  
plt.show()
```



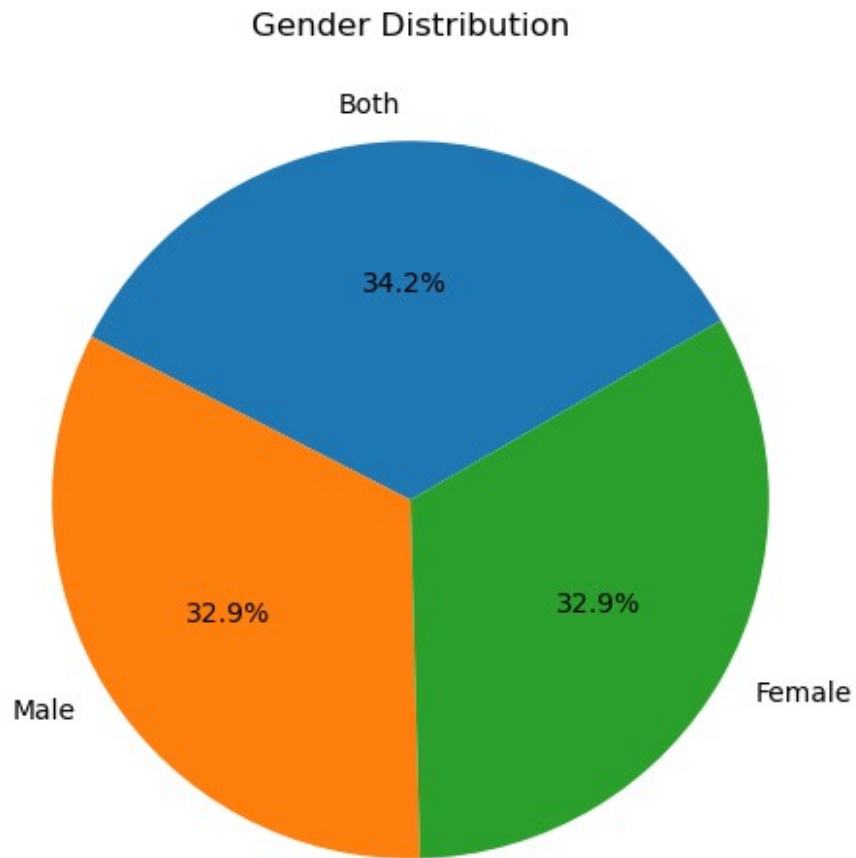
```
# Univariate Pie Chart:Smoking_Prevalence
Smoking_Prevalence= df['Gender'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(Smoking_Prevalence, labels=Smoking_Prevalence.index,
autopct='%1.1f%%', startangle=30)
plt.title('Smoking_Prevalence')
plt.show()
```



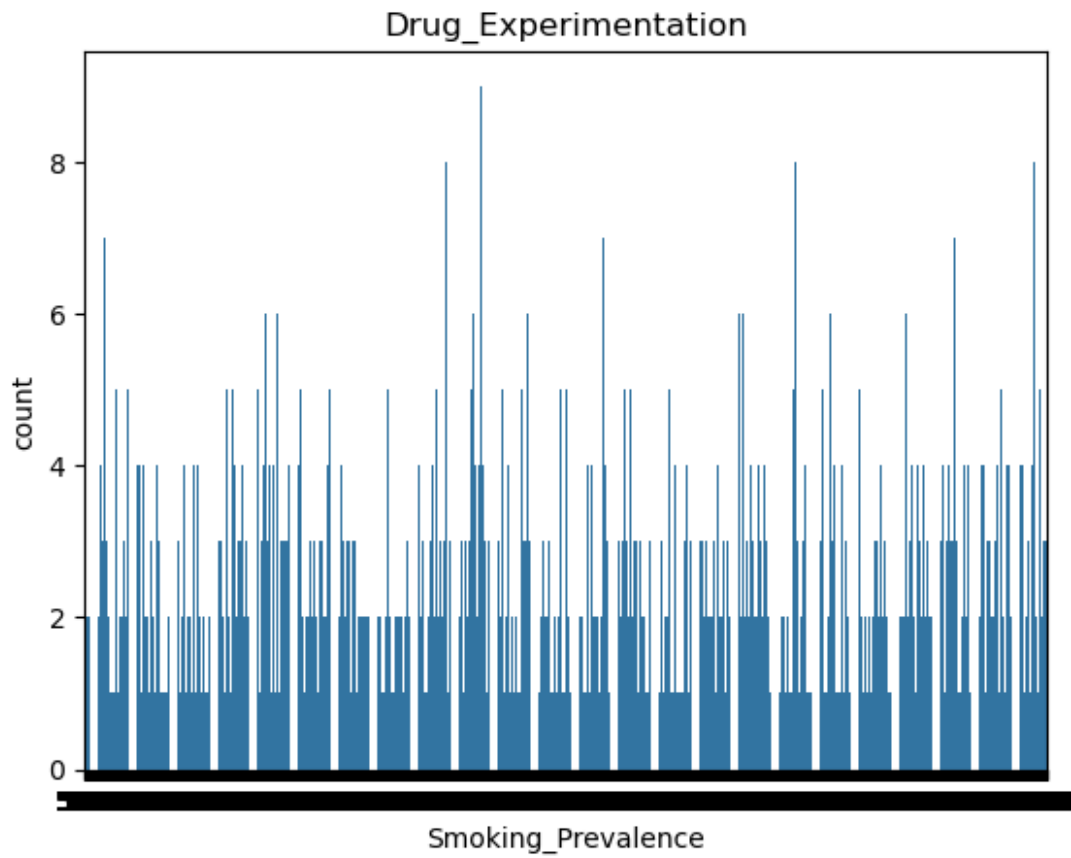
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv(r"C:\Users\abung\Downloads\Youth smoking SDS.csv")
sns.countplot(x='Gender', data=df).set_title('Drug_Experimentation')
plt.show()
```



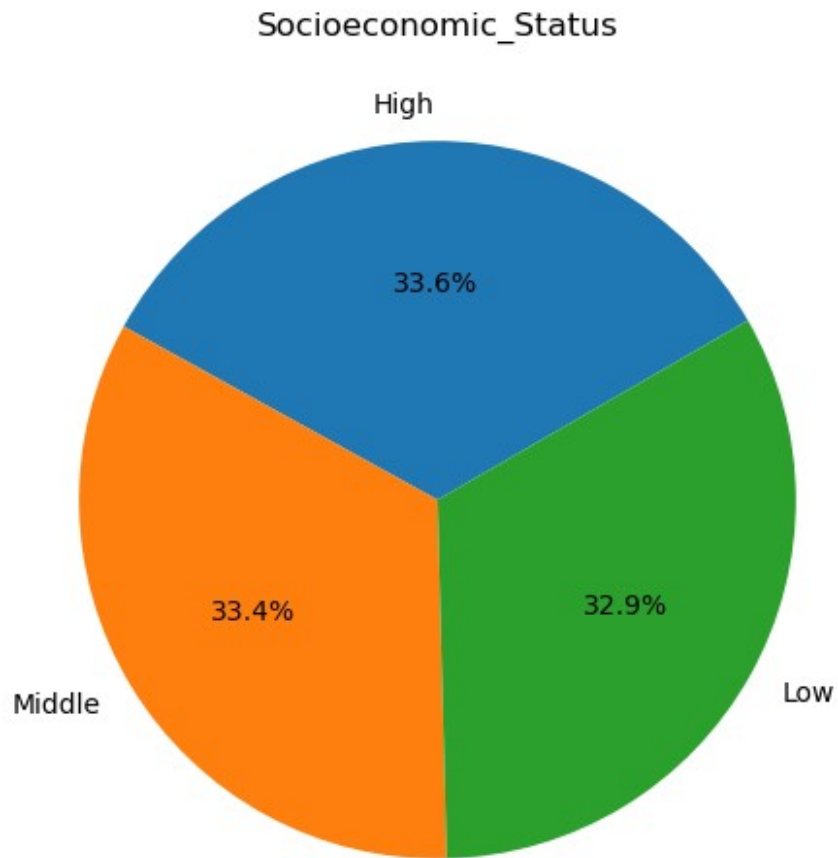
```
gender_counts = df['Gender'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%',
startangle=30)
plt.title('Gender Distribution')
plt.show()
```



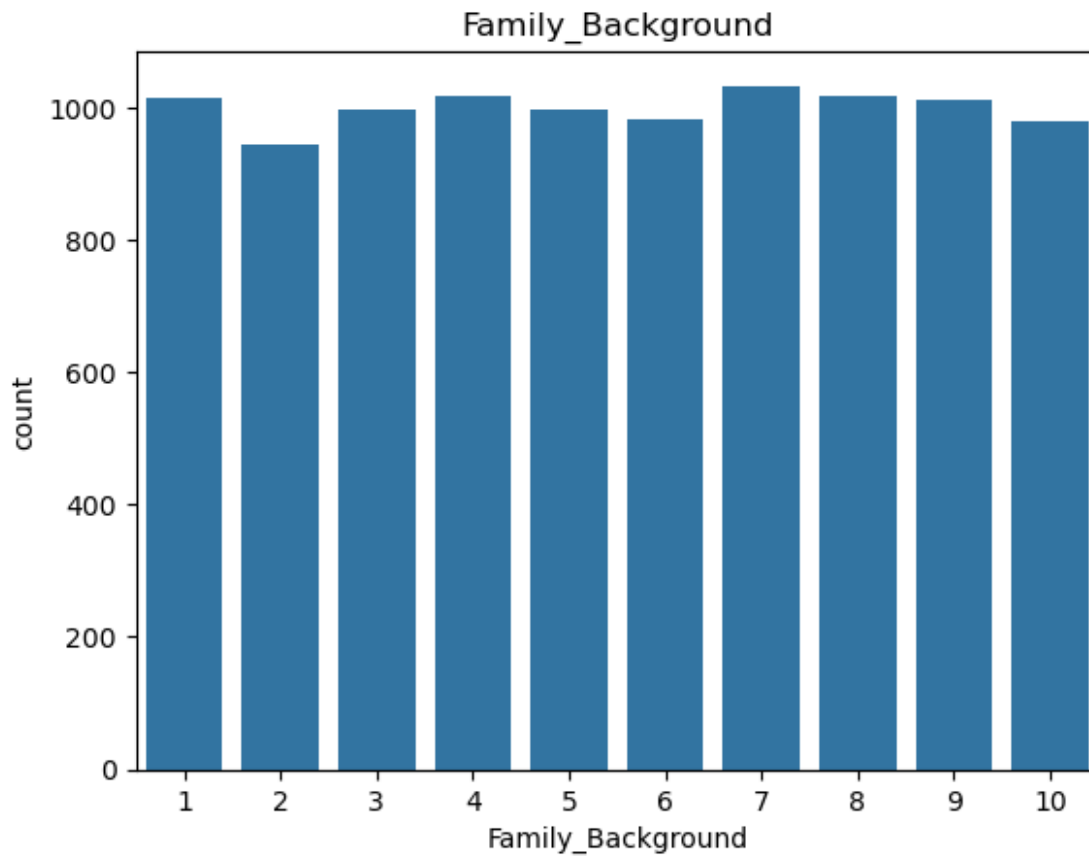
```
sns.countplot(x='Smoking_Prevalence',data=df).set_title('Drug_Experimentation')  
plt.show()
```



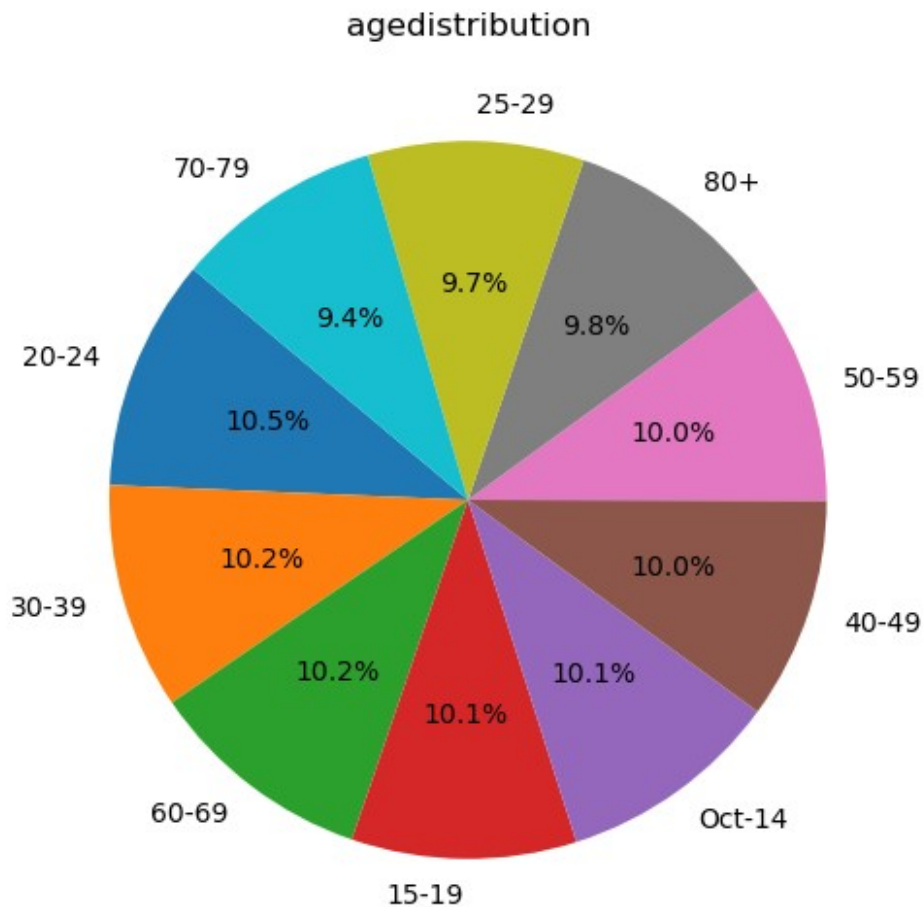
```
import matplotlib.pyplot as plt # Add this if not already done
Smoking_Prevalence_counts = df['Socioeconomic_Status'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(Smoking_Prevalence_counts, labels=Smoking_Prevalence_counts.index, autopct='%1.1f%%', startangle=30)
plt.title('Socioeconomic_Status')
plt.show()
```



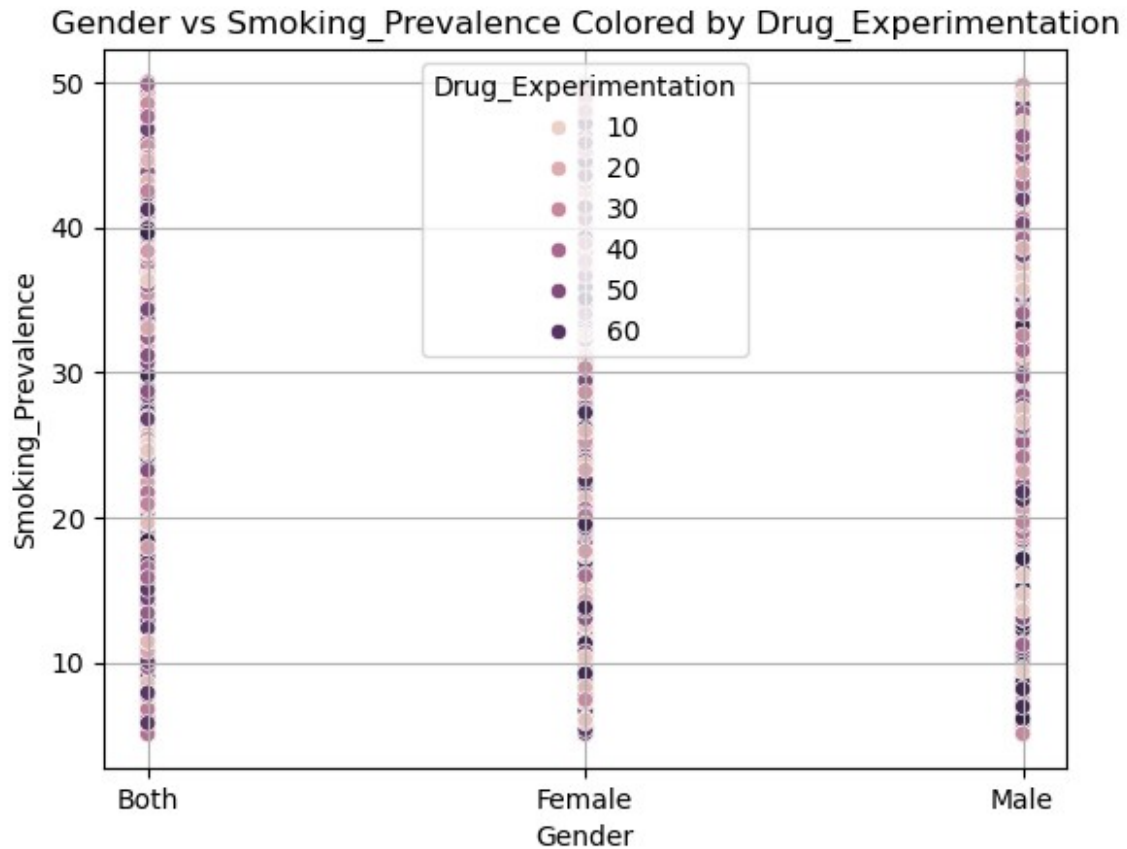
```
# Univariate Analysis: Categorical  
sns.countplot(x='Family_Background',  
data=df).set_title('Family_Background')  
plt.show()
```

```
# Univariate Pie Chart:Age_Group
age_count = df['Age_Group'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(age_count, labels=age_count.index, autopct='%1.1f%%',
startangle=140)
plt.title('agedistribution')
plt.show()
```



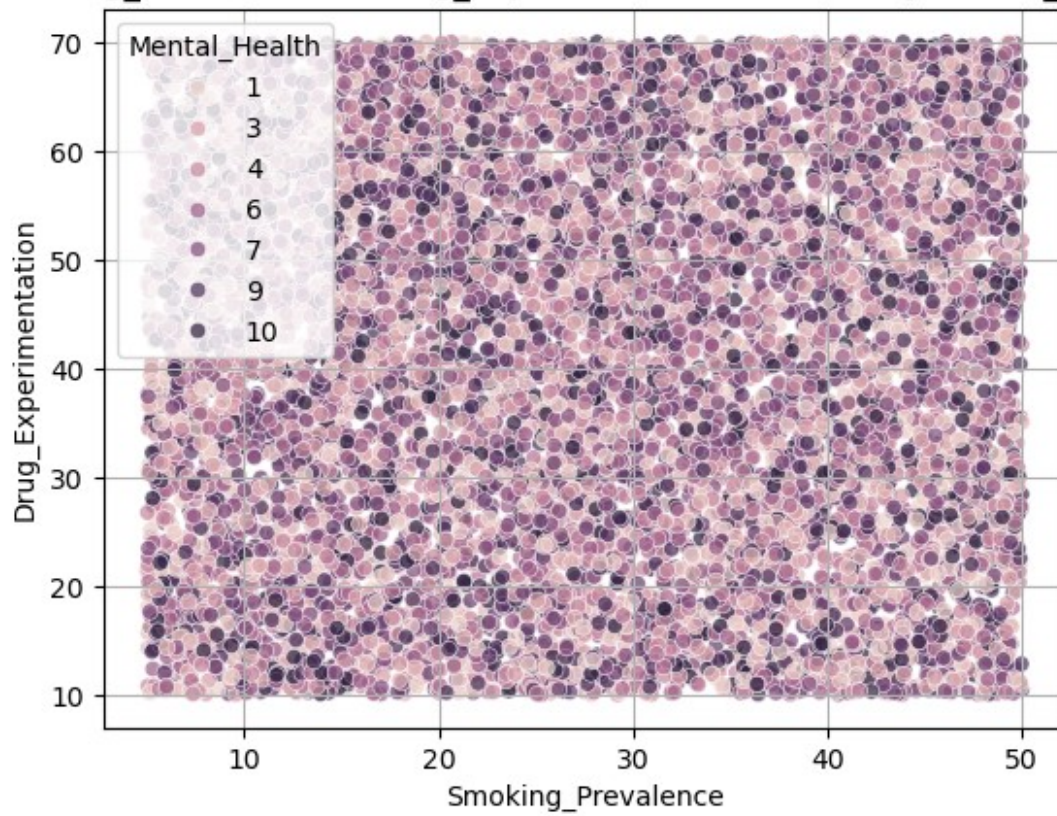
```
# Gender vs Smoking_Prevalence Colored by Drug_Experimentation  
scaatter plot  
sns.scatterplot(x='Gender',  
y='Smoking_Prevalence',hue='Drug_Experimentation',data=df)  
plt.title('Gender vs Smoking_Prevalence Colored by  
Drug_Experimentation')  
plt.xlabel('Gender')  
plt.ylabel('Smoking_Prevalence')  
plt.grid(True)  
plt.show()
```



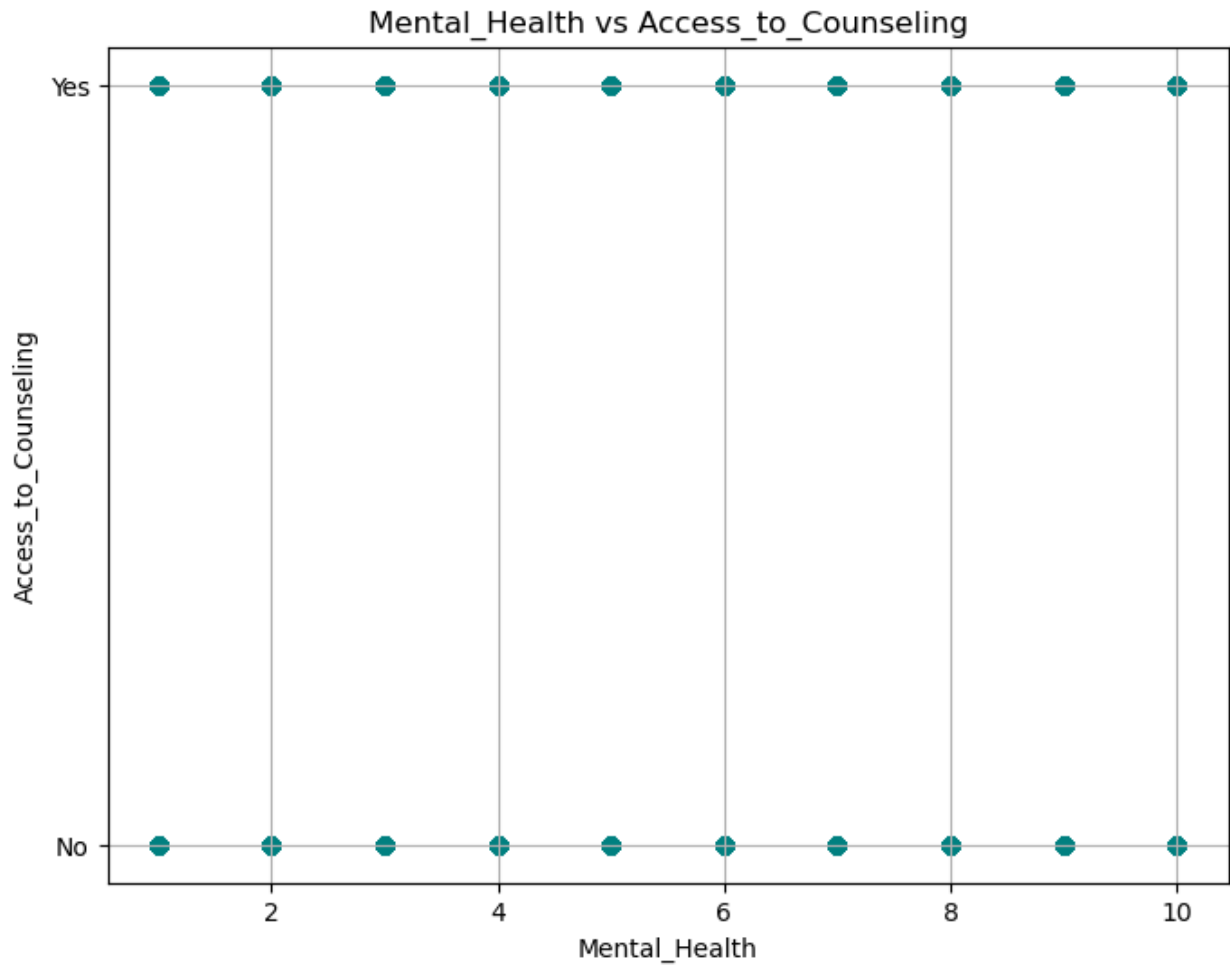
```
# Scatter plot colored by Mental_Health
sns.scatterplot(x='Smoking_Prevalence', y='Drug_Experimentation',
hue='Mental_Health',data=df, alpha=0.7)
plt.title('Smoking_Prevalence vs Drug_Experimentation Colored by
Mental_Health')
plt.xlabel('Smoking_Prevalence')
plt.ylabel('Drug_Experimentation')
plt.grid(True)
plt.show()
```

```
C:\Users\abung\AppData\Roaming\Python\Python312\site-packages\IPython\
core\pylabtools.py:170: UserWarning: Glyph 9 ( ) missing from
current font.
fig.canvas.print_figure(bytes_io, **kw)
```

Smoking_Prevalence vs Drug_Experimentation Colored by Mental_Health



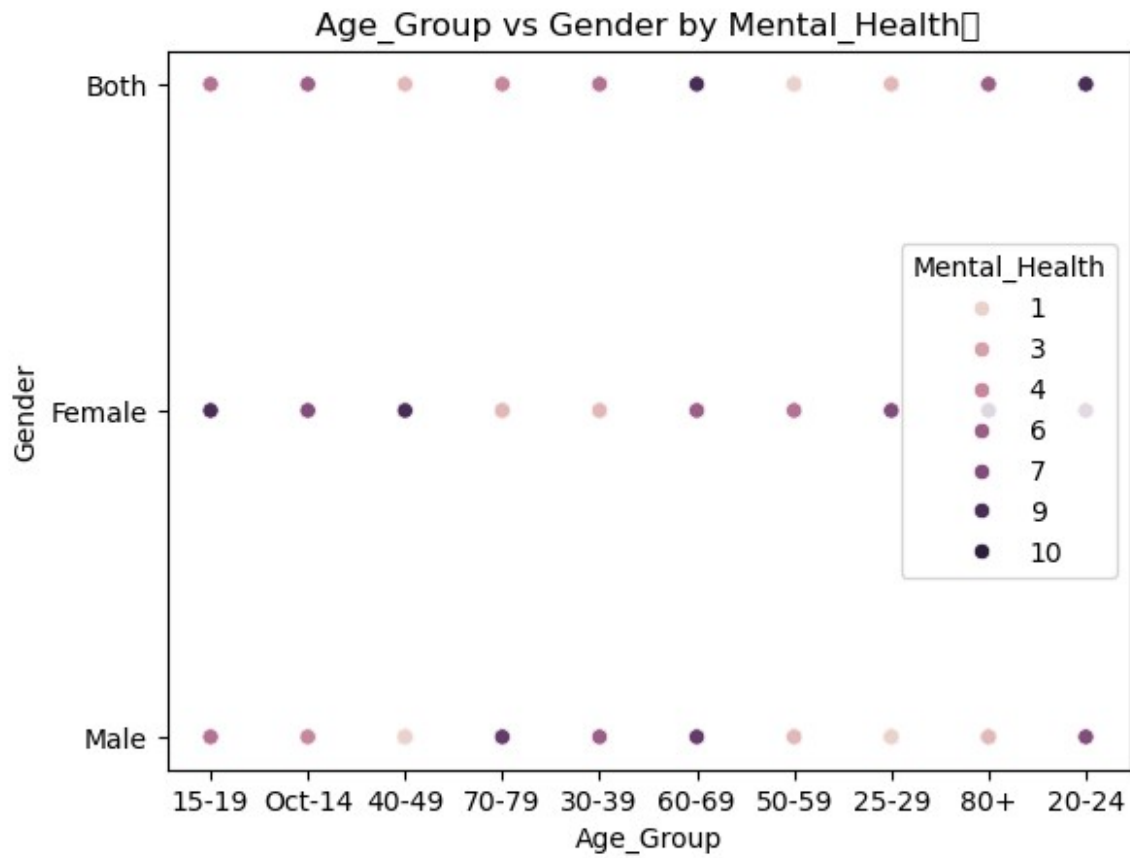
```
# Mental_Health vs Access_to_Counseling scatter plot
plt.figure(figsize=(8, 6))
plt.scatter(df['Mental_Health'], df['Access_to_Counseling'],
            color='teal', alpha=0.5)
plt.title('Mental_Health vs Access_to_Counseling')
plt.xlabel('Mental_Health')
plt.ylabel('Access_to_Counseling')
plt.grid(True)
plt.show()
```



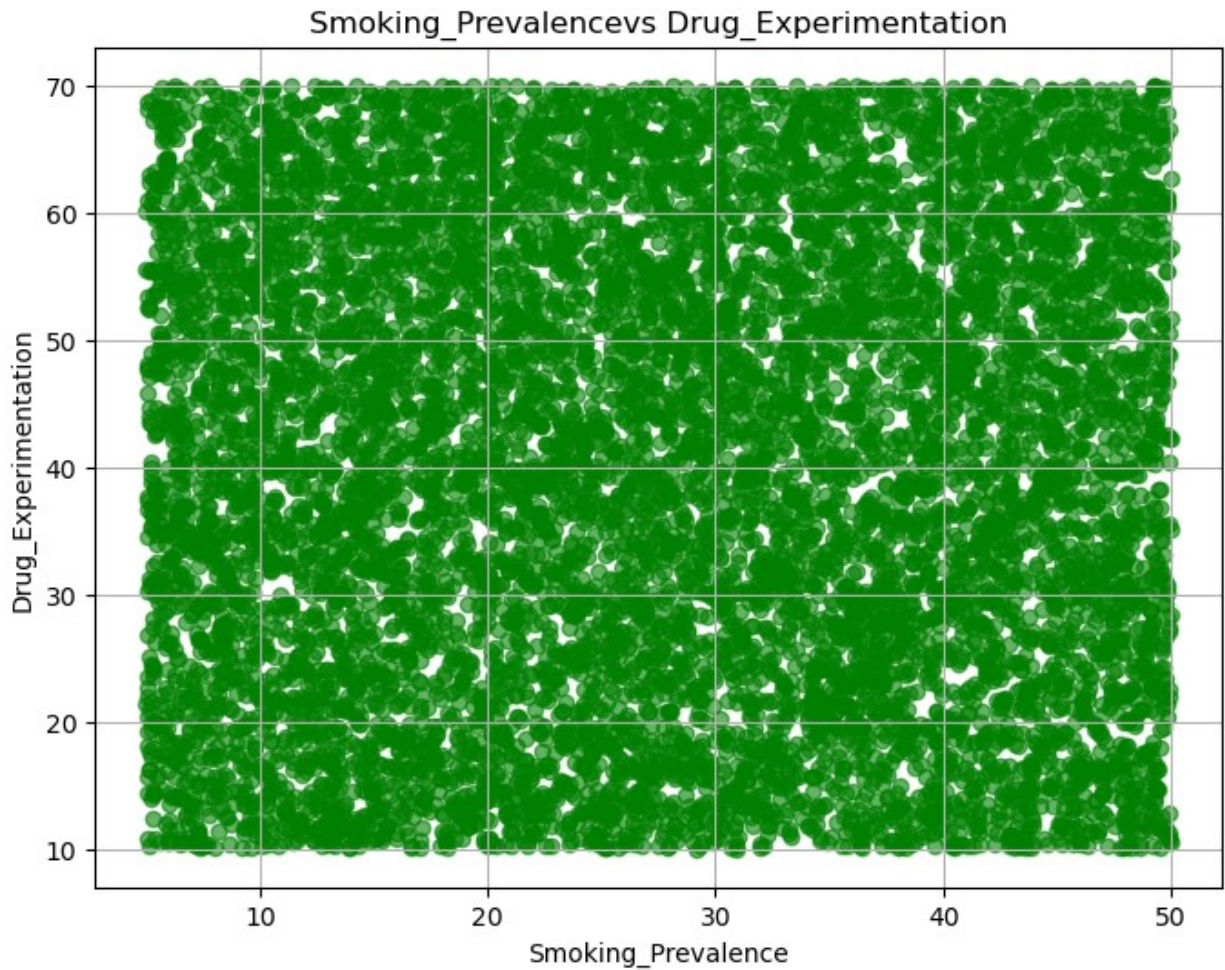
```
# Scatter plot with hue based on Mental_Health
sns.scatterplot(x='Age_Group', y='Gender',
hue='Mental_Health',data=df)
plt.title('Age_Group vs Gender by Mental_Health ')
plt.xlabel('Age_Group')
plt.ylabel('Gender')
plt.show()
```

C:\Users\abung\AppData\Roaming\Python\Python312\site-packages\IPython\
core\pylabtools.py:170: UserWarning: Glyph 9 () missing from
current font.

```
fig.canvas.print_figure(bytes_io, **kw)
```

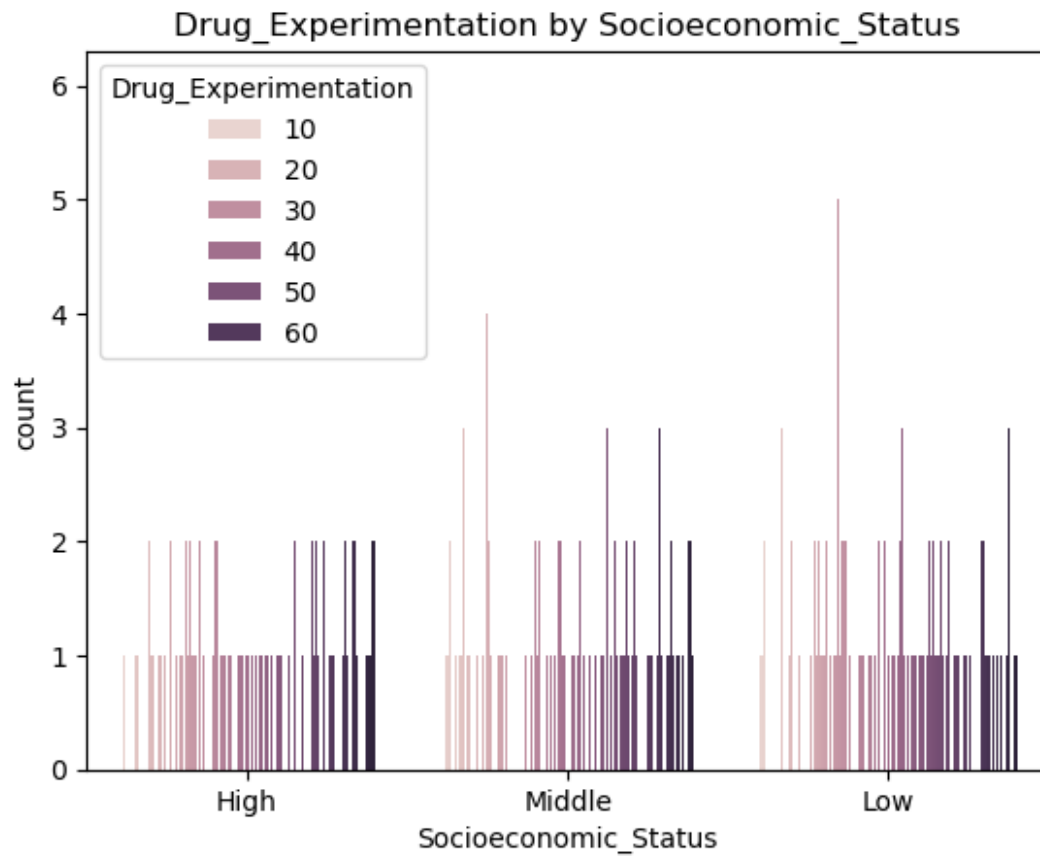


```
# Scatter plot(Smoking_Prevalence vs Drug_Experimentation)
plt.figure(figsize=(8, 6))
plt.scatter(df['Smoking_Prevalence'],df['Drug_Experimentation'],
alpha=0.6, c='green')
plt.title('Smoking_Prevalencevs Drug_Experimentation')
plt.xlabel('Smoking_Prevalence')
plt.ylabel('Drug_Experimentation')
plt.grid(True)
plt.show()
```

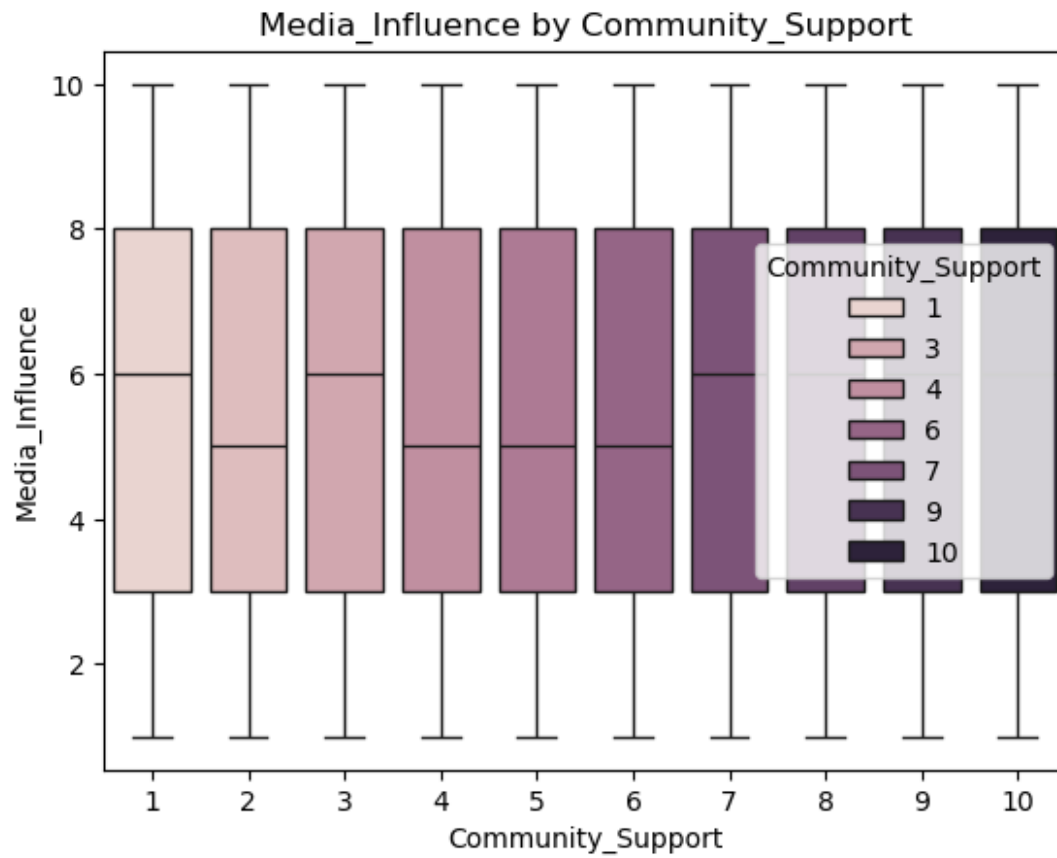



```
# Bivariate Socioeconomic_Status vs Drug_Experimentation
sns.countplot(x='Socioeconomic_Status',
hue='Drug_Experimentation',data=df)
plt.title('Drug_Experimentation by Socioeconomic_Status')
plt.show()
```

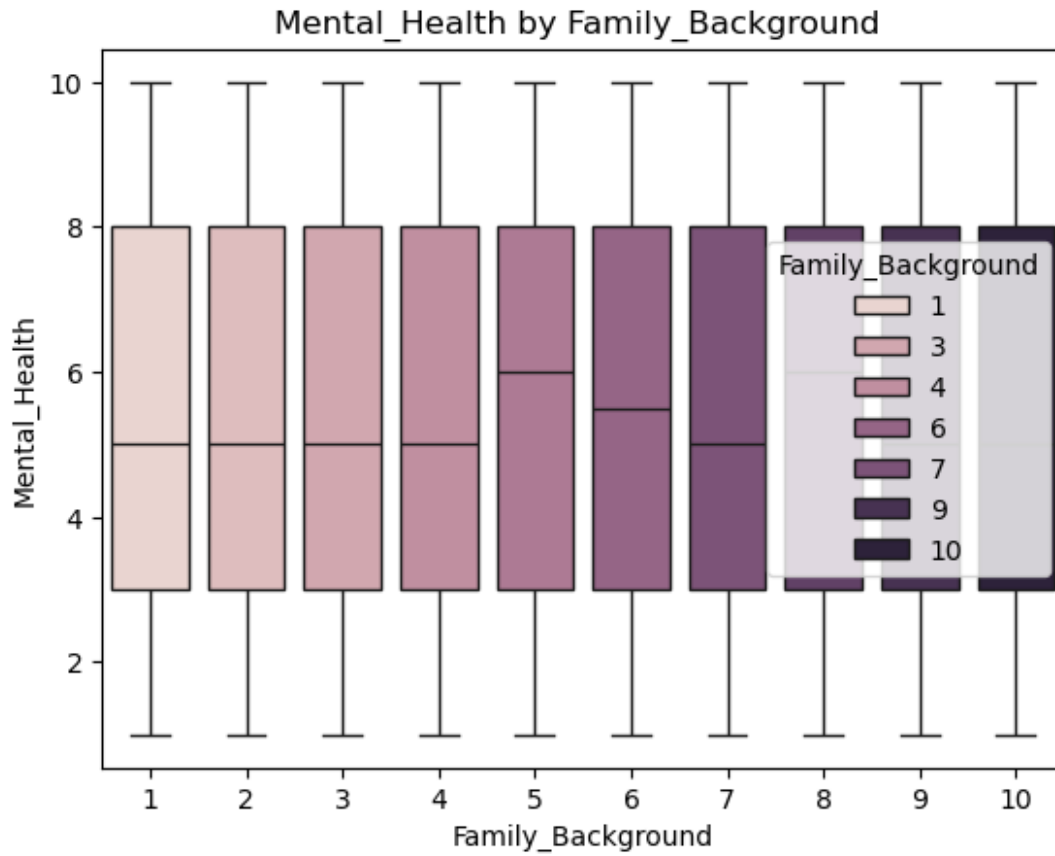
```
C:\Users\abung\AppData\Roaming\Python\Python312\site-packages\IPython\
core\pylabtools.py:170: UserWarning: Creating legend with loc="best"
can be slow with large amounts of data.
  fig.canvas.print_figure(bytes_io, **kw)
```



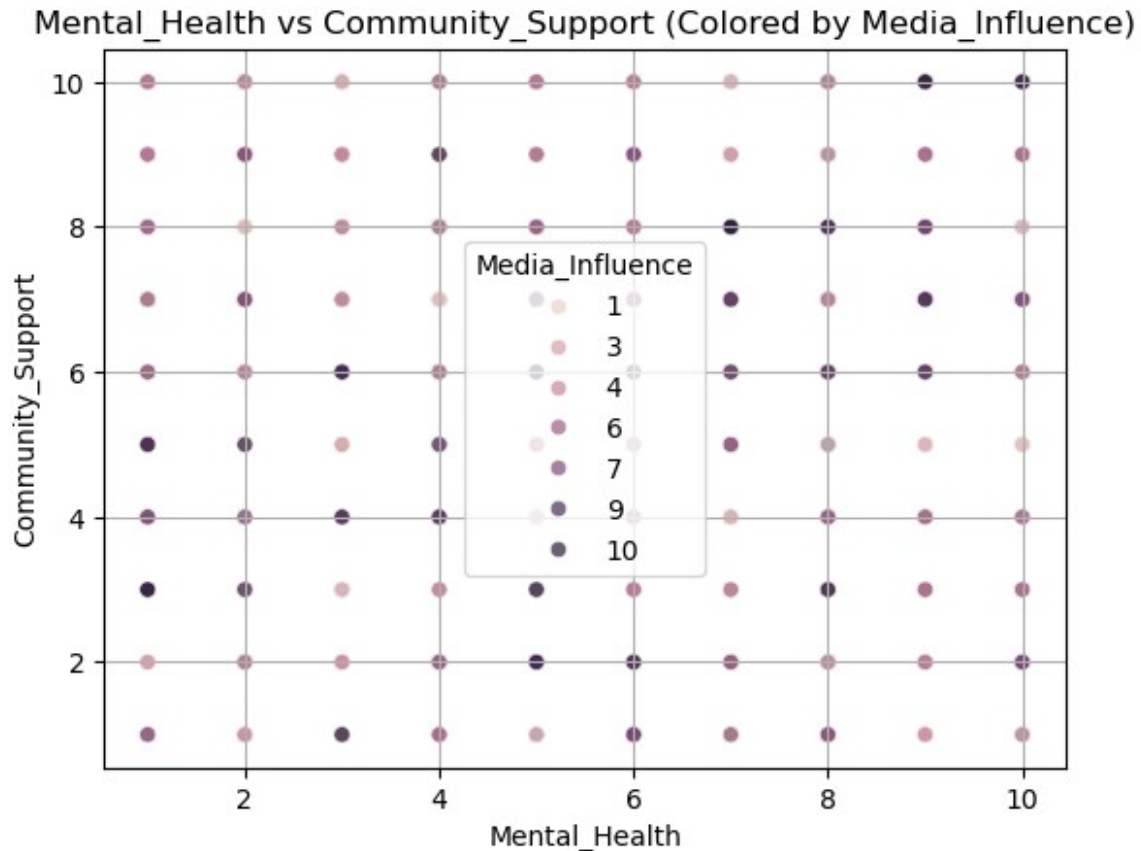
```
# Bivariate Analysis Community_Support vs Media_Influence
sns.boxplot(x='Community_Support',
y='Media_Influence',hue='Community_Support', data=df)
plt.title('Media_Influence by Community_Support')
plt.show()
```

```
# Bivariate Analysis Family_Background vs Mental_Health
sns.boxplot(x='Family_Background',
y='Mental_Health',hue='Family_Background',data=df)
plt.title('Mental_Health by Family_Background ')
plt.show()
```



```
# Scatter Plot (Bivariate): Mental_Health vs Community_Support,  
# colored by Media_Influence  
sns.scatterplot(x='Mental_Health',y='Community_Support',hue='Media_Influence',data=df,alpha=0.7)  
plt.title('Mental_Health vs Community_Support (Colored by Media_Influence)')  
plt.xlabel('Mental_Health')  
plt.ylabel('Community_Support')  
plt.grid(True)  
plt.show()
```



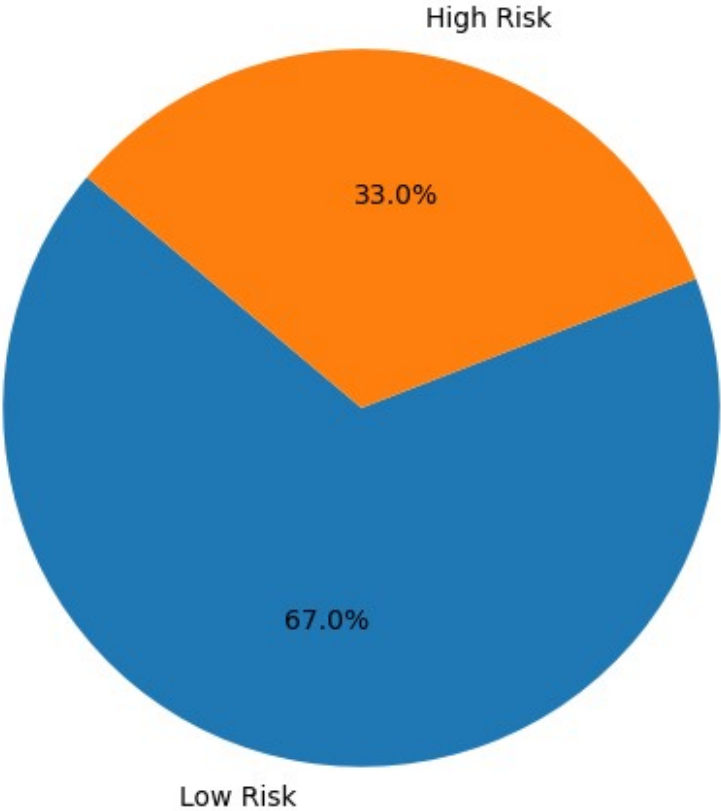
```
male_df = df[df['Gender'] == 'Male']
female_df = df[df['Gender'] == 'Female']

# Use 'Drug_Experimentation' as a risk indicator
# Define: >= 50 as High Risk (1), < 50 as Low Risk (0)
male_risk = (male_df['Drug_Experimentation'] >=
50).astype(int).value_counts()
female_risk = (female_df['Drug_Experimentation'] >=
50).astype(int).value_counts()

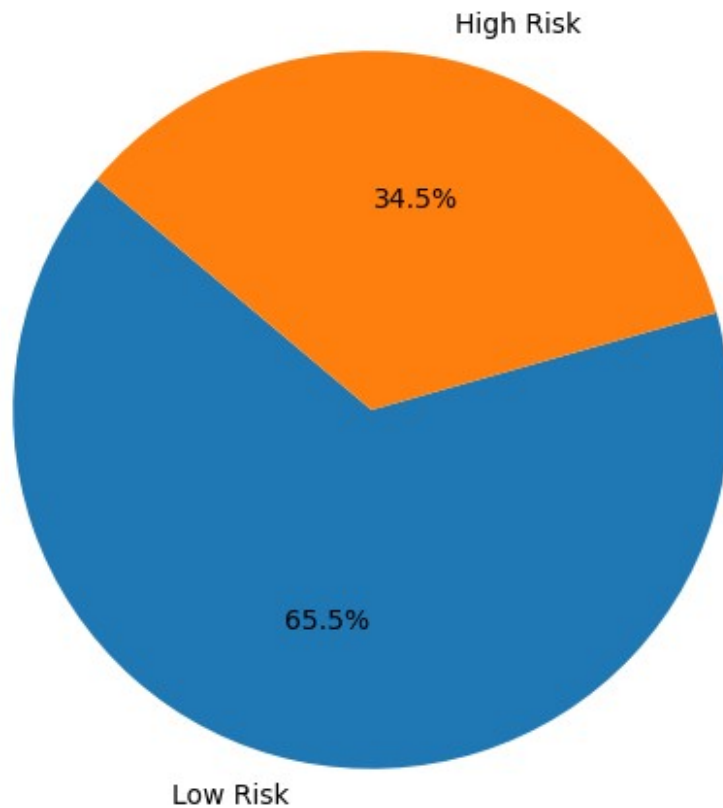
# Plot Male Risk Pie Chart
plt.figure(figsize=(6, 6))
plt.pie(male_risk, labels=['Low Risk', 'High Risk'], autopct='%1.1f%%',
startangle=140)
plt.title('Drug Experimentation Risk in Male Youth')
plt.show()

# Plot Female Risk Pie Chart
plt.figure(figsize=(6, 6))
plt.pie(female_risk, labels=['Low Risk', 'High Risk'], autopct='%1.1f%%',
startangle=140)
plt.title('Drug Experimentation Risk in Female Youth')
plt.show()
```

Drug Experimentation Risk in Male Youth



Drug Experimentation Risk in Female Youth



```
# Define selected numeric columns for multivariate analysis
selected_columns = [
    'Smoking_Prevalence',
    'Drug_Experimentation',
    'Peer_Influence',
    'Family_Background',
    'Mental_Health',
    'Parental_Supervision'
]

# Rename the dataset variable
df = df.rename(columns=str.strip) # Optional cleanup, if needed

# Create a proxy binary label for Drug Risk
df['High_Risk'] = (df['Drug_Experimentation'] >= 50).astype(int)

# Drop missing values from selected columns and 'High_Risk'
plot_data = df[selected_columns + ['High_Risk']].dropna()

# Pairplot with hue based on drug risk
```

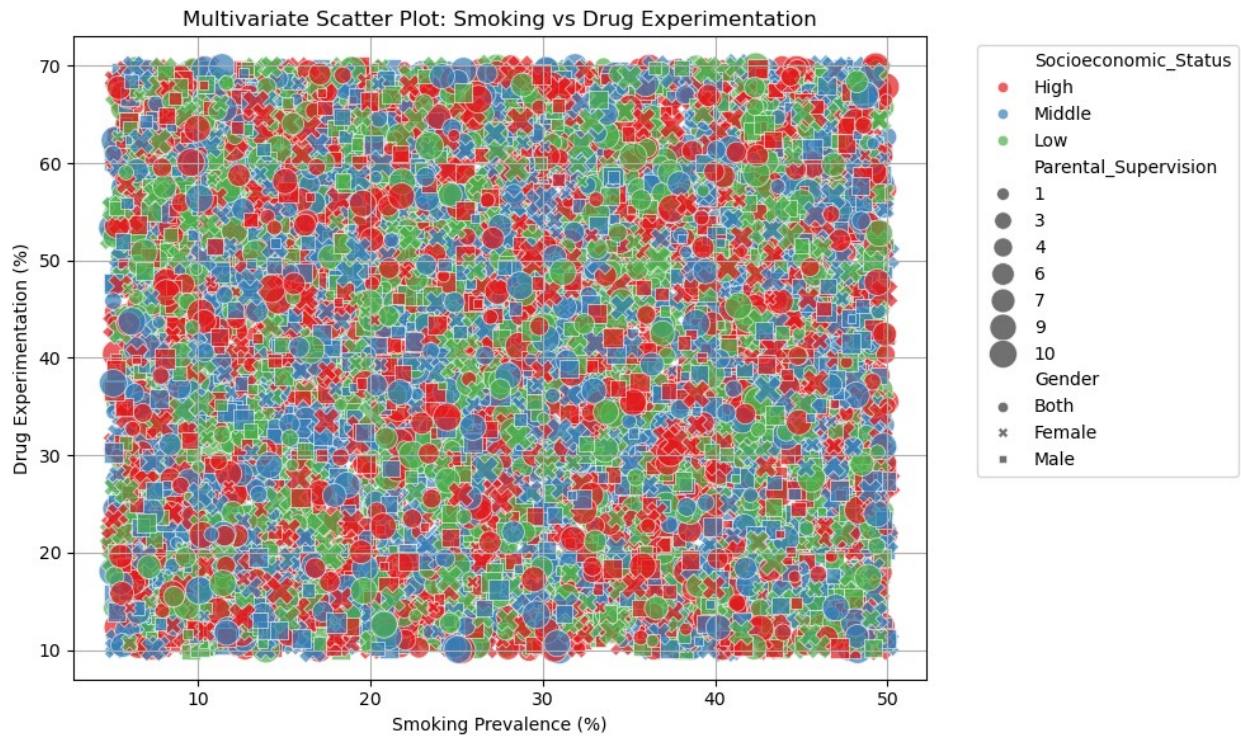
```
sns.pairplot(plot_data, hue='High_Risk', diag_kind='hist',
palette='Set2')
plt.suptitle("Pairplot of Selected Features Colored by Drug Risk",
y=1.02)
plt.show()
```



```
plt.figure(figsize=(10, 6))
sns.scatterplot(x='Smoking_Prevalence', y='Drug_Experimentation',
hue='Socioeconomic_Status', size='Parental_Supervision',
style='Gender', data=df, palette='Set1', sizes=(50, 250), alpha=0.7)
plt.title('Multivariate Scatter Plot: Smoking vs Drug
Experimentation')
```



```
plt.xlabel('Smoking Prevalence (%)')
plt.ylabel('Drug Experimentation (%)')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.tight_layout()
plt.show()
```



```
corr_matrix = df.corr(numeric_only=True)

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f",
            linewidths=0.5)
plt.title('Correlation Heatmap (Youth Smoking & Drug Use)')
plt.show()
```

