

✓ Basic Statical Data Visualization

Load the pandas, seaborn, and matplotlib libraries.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df = pd.read_csv('student_depression_dataset.csv')
```

The first 5 rows of the data frame for visualization

```
df.head()
```



	id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Stuc Satisfactic
0	2	Male	33.0	Visakhapatnam	Student	5.0	0.0	8.97	2
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5
3	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3

✓ Data Preparation

```
df["Dietary Habits"].value_counts()
```



	count
Dietary Habits	
0	10317
1	9921
2	7651
-1	12

dtype: int64

```
df["Financial Stress"].value_counts()
```



	count
Financial Stress	
5.0	6715
4.0	5775
3.0	5226
1.0	5121
2.0	5061

dtype: int64

To improve the visualization and analysis of non-numeric variables, they will be converted into numerical values. For example, in the case of **Dietary Habits**, the categories will be mapped as follows: **Unhealthy** = 0, **Moderate** = 1, **Healthy** = 2, and **Other** = 3 .

```
import numpy as np
```

```
df = pd.read_csv('student_depression_dataset.csv')
```

```
gender_mapping = {'Male': 1, 'Female': 0}
df['Gender'] = df['Gender'].map(gender_mapping).fillna(-1).astype(int)
```

```
dietary_mapping = {'Unhealthy': 0, 'Moderate': 1, 'Healthy': 2, 'Other': 3}
df['Dietary Habits'] = df['Dietary Habits'].map(dietary_mapping).fillna(-1).astype(int)
```

```
suicide_mapping = {'Yes': 1, 'No': 0}
df['Have you ever had suicidal thoughts ?'] = df['Have you ever had suicidal thought
```

```
family_history = {'Yes': 1, 'No': 0}
```

```
df['Family History of Mental Illness'] = df['Family History of Mental Illness'].map(

df["Financial Stress"] = df["Financial Stress"].replace("?", np.nan) # Now np is def

df["Financial Stress"] = df["Financial Stress"].astype(float)
```

✓ General Information of The Dataset

```
df.shape
```

```
➦ (27901, 18)
```

The shape function indicates that the dataset contains information on 27,901 students and includes 18 variables describing various aspects of each individual.

```
df.info()
```

```
➦ <class 'pandas.core.frame.DataFrame'>
RangeIndex: 27901 entries, 0 to 27900
Data columns (total 18 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   id                                         27901 non-null  int64
1   Gender                                    27901 non-null  int64
2   Age                                       27901 non-null  float64
3   City                                      27901 non-null  object
4   Profession                               27901 non-null  object
5   Academic Pressure                        27901 non-null  float64
6   Work Pressure                           27901 non-null  float64
7   CGPA                                     27901 non-null  float64
8   Study Satisfaction                      27901 non-null  float64
9   Job Satisfaction                        27901 non-null  float64
10  Sleep Duration                          27901 non-null  object
11  Dietary Habits                          27901 non-null  int64
12  Degree                                  27901 non-null  object
13  Have you ever had suicidal thoughts ?  27901 non-null  int64
14  Work/Study Hours                       27901 non-null  float64
15  Financial Stress                        27898 non-null  float64
16  Family History of Mental Illness       27901 non-null  int64
17  Depression                             27901 non-null  int64
dtypes: float64(8), int64(6), object(4)
memory usage: 3.8+ MB
```

✓ Variables

The variables analyzed in relation to depression in students.

```
df.columns
```

```
Index(['id', 'Gender', 'Age', 'City', 'Profession', 'Academic Pressure',  
      'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction',  
      'Sleep Duration', 'Dietary Habits', 'Degree',  
      'Have you ever had suicidal thoughts ?', 'Work/Study Hours',  
      'Financial Stress', 'Family History of Mental Illness', 'Depression'],  
      dtype='object')
```

Identification of their type.

```
df.dtypes
```

```
0
```

id	int64
Gender	int64
Age	float64
City	object
Profession	object
Academic Pressure	float64
Work Pressure	float64
CGPA	float64
Study Satisfaction	float64
Job Satisfaction	float64
Sleep Duration	object
Dietary Habits	int64
Degree	object
Have you ever had suicidal thoughts ?	int64
Work/Study Hours	float64
Financial Stress	float64
Family History of Mental Illness	int64
Depression	int64

dtype: object

✓ Analysis of the Variables

Age:

Mean: 25.82

Median (50%): 25.00

Standard Deviation: 4.91

Range: 18 to 59

Conclusion: Most students are in their mid-20s, with a relatively narrow age distribution, likely indicating that the majority are undergraduate or graduate students.

Academic Pressure:

Mean: 3.14

Median: 3.00

Standard Deviation: 1.38

Range: 0 to 5

Conclusion: Academic pressure is moderate on average, with a significant portion experiencing high pressure, as indicated by the 75th percentile being close to the maximum.

Work Pressure:

Mean: 0.00

Median: 0.00

Standard Deviation: 0.04

Range: 0 to 5

Conclusion: Almost no students report work pressure, likely because most are not working while studying.

CGPA:

Mean: 7.66

Median: 7.77

Standard Deviation: 1.47

Range: 0 to 10

Conclusion: The average CGPA is high, but the wide range suggests some students are struggling academically.

Study Satisfaction:

Mean: 2.94

Median: 3.00

Standard Deviation: 1.36

Range: 0 to 5

Conclusion: On average, students have moderate satisfaction with their studies, but there is a wide range in responses.

Job Satisfaction:

Mean: 0.00

Median: 0.00

Standard Deviation: 0.04

Range: 0 to 4

Conclusion: Most students do not work, as indicated by the low mean and median.

Work/Study Hours:

Mean: 7.16

Median: 8.00

Standard Deviation: 3.71

Range: 0 to 12

Conclusion: Students tend to study or work around 7-8 hours a day, with some students potentially overworking.

Depression:

Mean: 0.59


Median: 0.00

Standard Deviation: 0.49

Range: 0 to 1

Conclusion: Over half of the students report symptoms of depression, suggesting mental health concerns are significant in this group.

```
summary = pd.DataFrame({  
    'Minimum': df.min(numeric_only=True),  
    'Maximum': df.max(numeric_only=True)  
})  
print(summary)
```



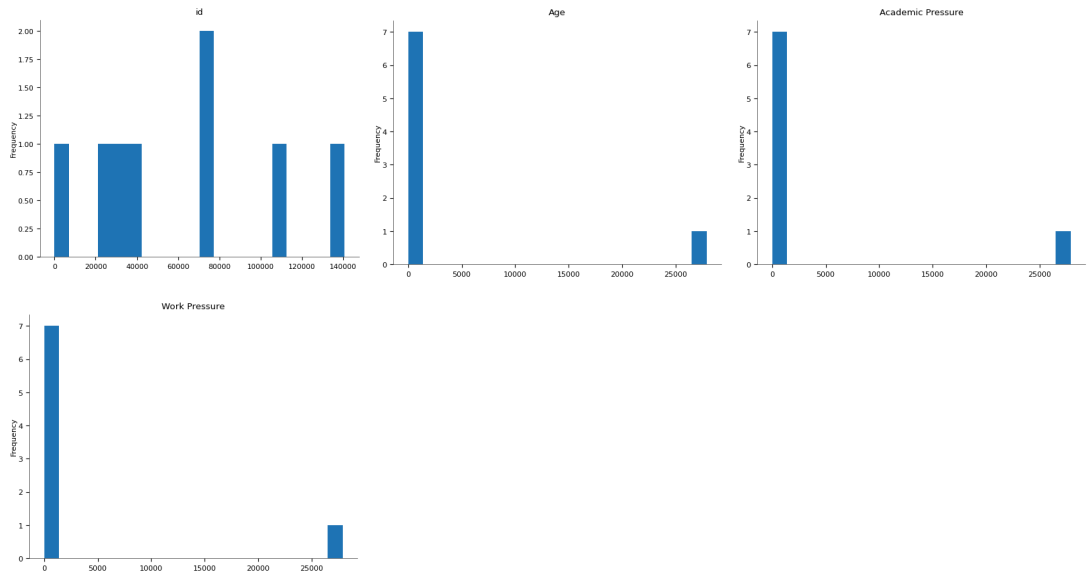
	Minimum	Maximum
id	2.0	140699.0
Age	18.0	59.0
Academic Pressure	0.0	5.0
Work Pressure	0.0	5.0
CGPA	0.0	10.0
Study Satisfaction	0.0	5.0
Job Satisfaction	0.0	4.0
Work/Study Hours	0.0	12.0
Depression	0.0	1.0

```
df.describe()
```

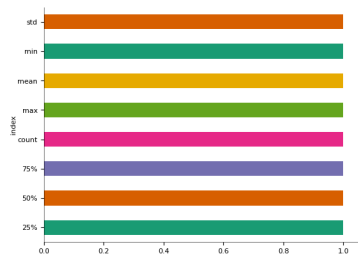


	id	Age	Academic Pressure	Work Pressure	CGPA	Student Satisfaction
count	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000	27901.000000
mean	70442.149421	25.822300	3.141214	0.000430	7.656104	2.943830
std	40641.175216	4.905687	1.381465	0.043992	1.470707	1.361140
min	2.000000	18.000000	0.000000	0.000000	0.000000	0.000000
25%	35039.000000	21.000000	2.000000	0.000000	6.290000	2.000000
50%	70684.000000	25.000000	3.000000	0.000000	7.770000	3.000000
75%	105818.000000	30.000000	4.000000	0.000000	8.920000	4.000000
max	140699.000000	59.000000	5.000000	5.000000	10.000000	5.000000

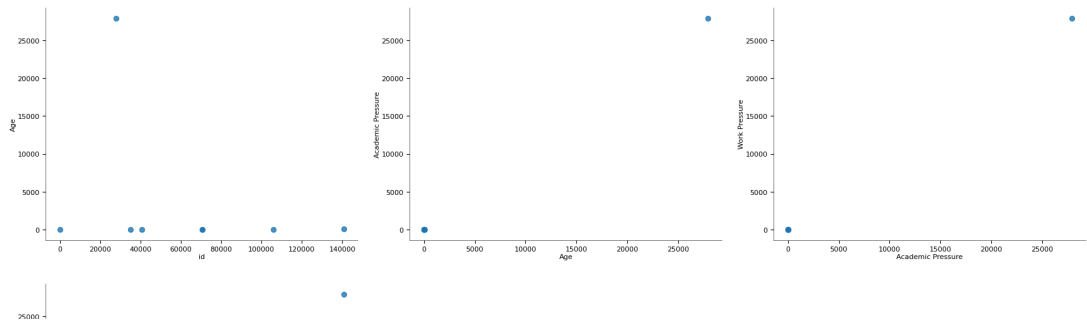
Distributions

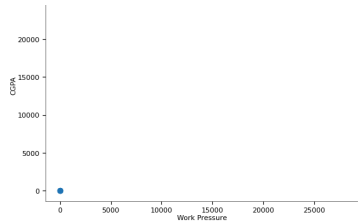


Categorical distributions

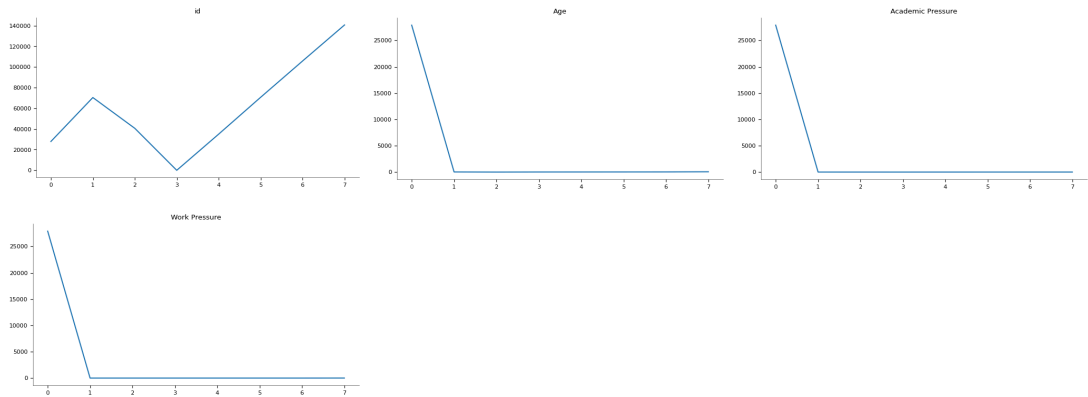


2-d distributions





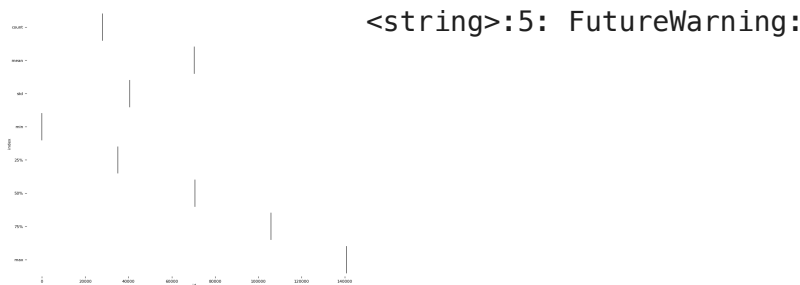
Values



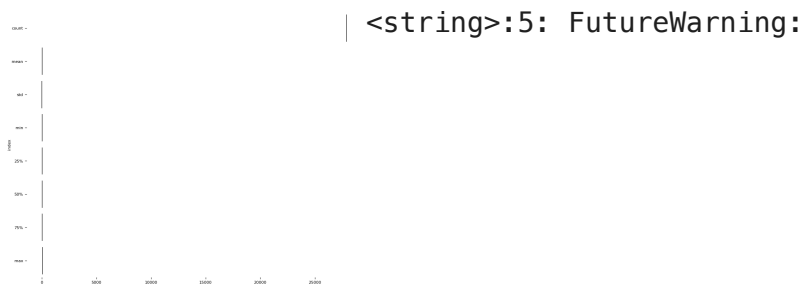
Faceted distributions

<string>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v



Passing `palette` without assigning `hue` is deprecated and will be removed in v



✓ Boxplots and Histograms for Variables

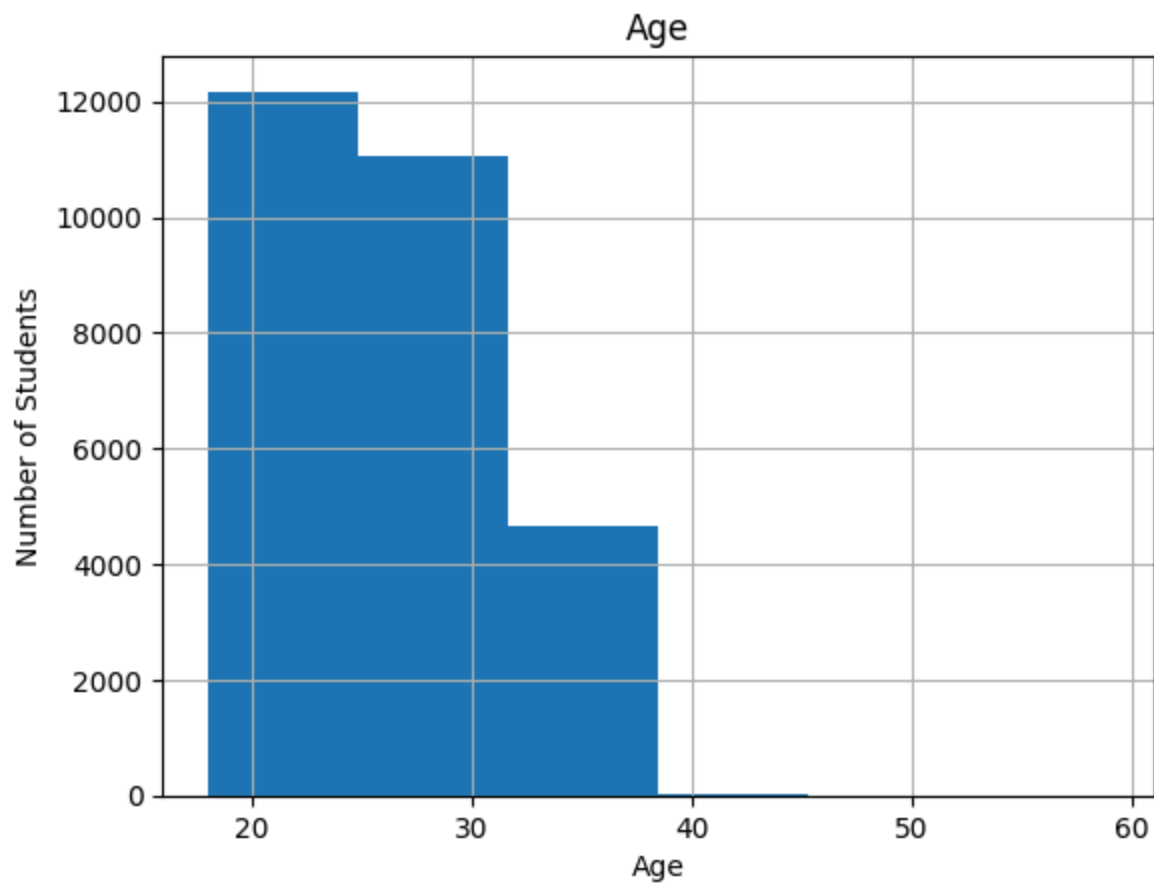
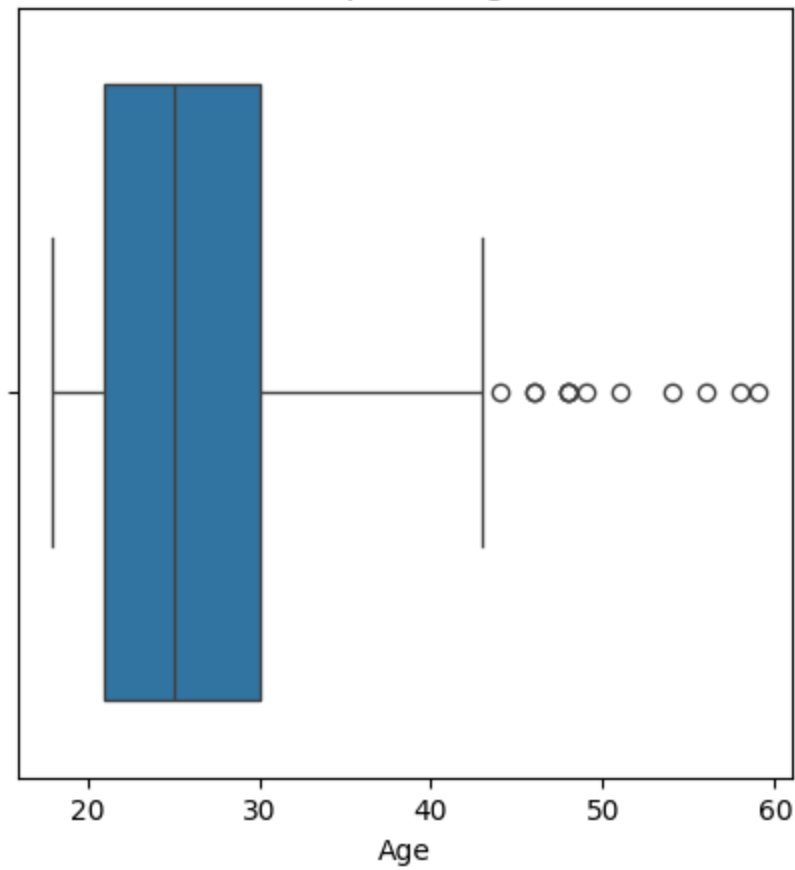
Boxplot and Histogram for Age

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='Age', data=df)
plt.title('Boxplot of Age')
plt.show()

df['Age'].hist(bins=6)
plt.title('Age')
plt.xlabel('Age')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of Age



Start coding or [generate](#) with AI.

- Most students are between 22 and 30 years old.
- There are some students over 40, who are outliers in this context.
- The boxplot shows that the data is slightly skewed upwards, as there are more outliers above than below.

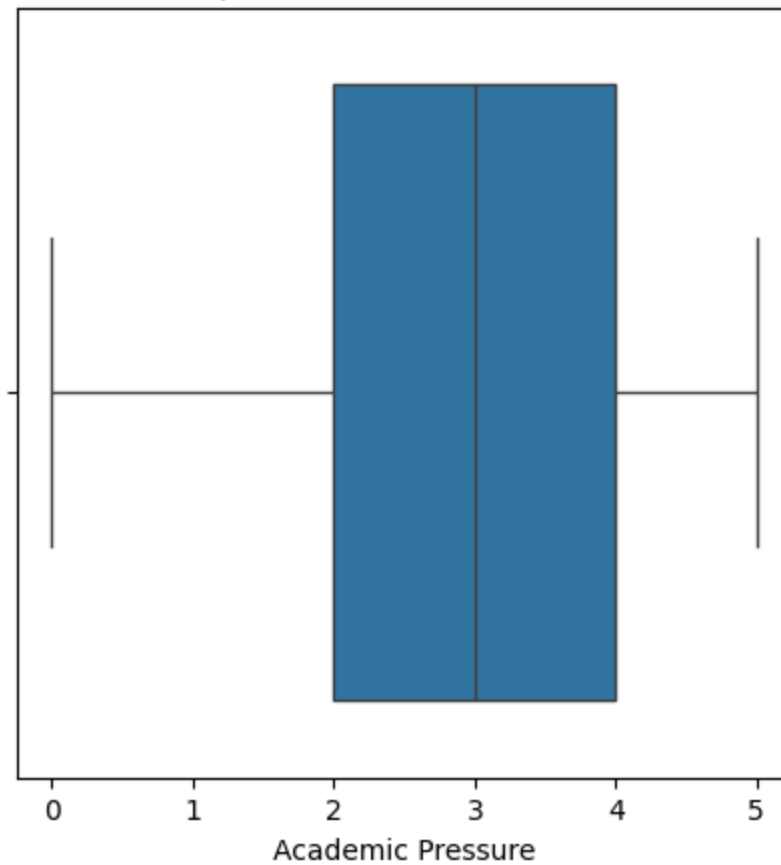
Boxplot and Histogram for Academic Pressure

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='Academic Pressure', data=df)
plt.title('Boxplot of Academic Pressure')
plt.show()
```

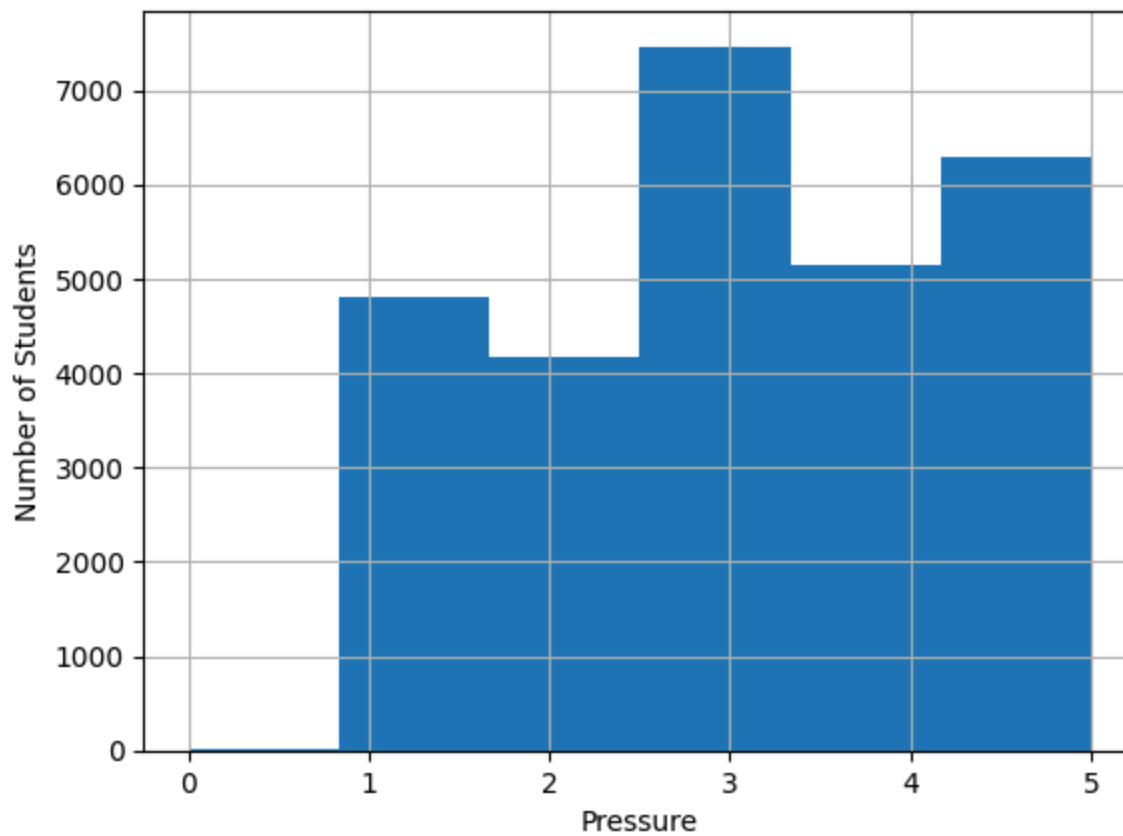
```
df['Academic Pressure'].hist(bins=6)
plt.title('CGPA')
plt.xlabel('Pressure')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of Academic Pressure



CGPA



- The majority of students report an academic pressure level between 2 and 4.
- The median academic pressure is approximately 3, indicating a moderate level.
- There are no visible outliers, suggesting that all responses fall within the expected range.

Boxplot for Work Pressure and Job Satisfaction

```
sns.boxplot(data=df[['Work Pressure', 'Job Satisfaction']])  
plt.title('Boxplot of Work Pressure and Job Satisfaction')  
plt.ylabel('Value')  
plt.show()
```



```
df["Work Pressure"].value_counts()
```



	count
Work Pressure	
0.0	27898
5.0	2
2.0	1

dtype: int64

```
df["Job Satisfaction"].value_counts()
```



	count
Job Satisfaction	
0.0	27893
2.0	3
4.0	2
1.0	2
3.0	1

dtype: int64

Most students do not work, therefore:

- Work Pressure has values that are mostly zero.
- Job Satisfaction shows very little variability (or may not even apply to many students).

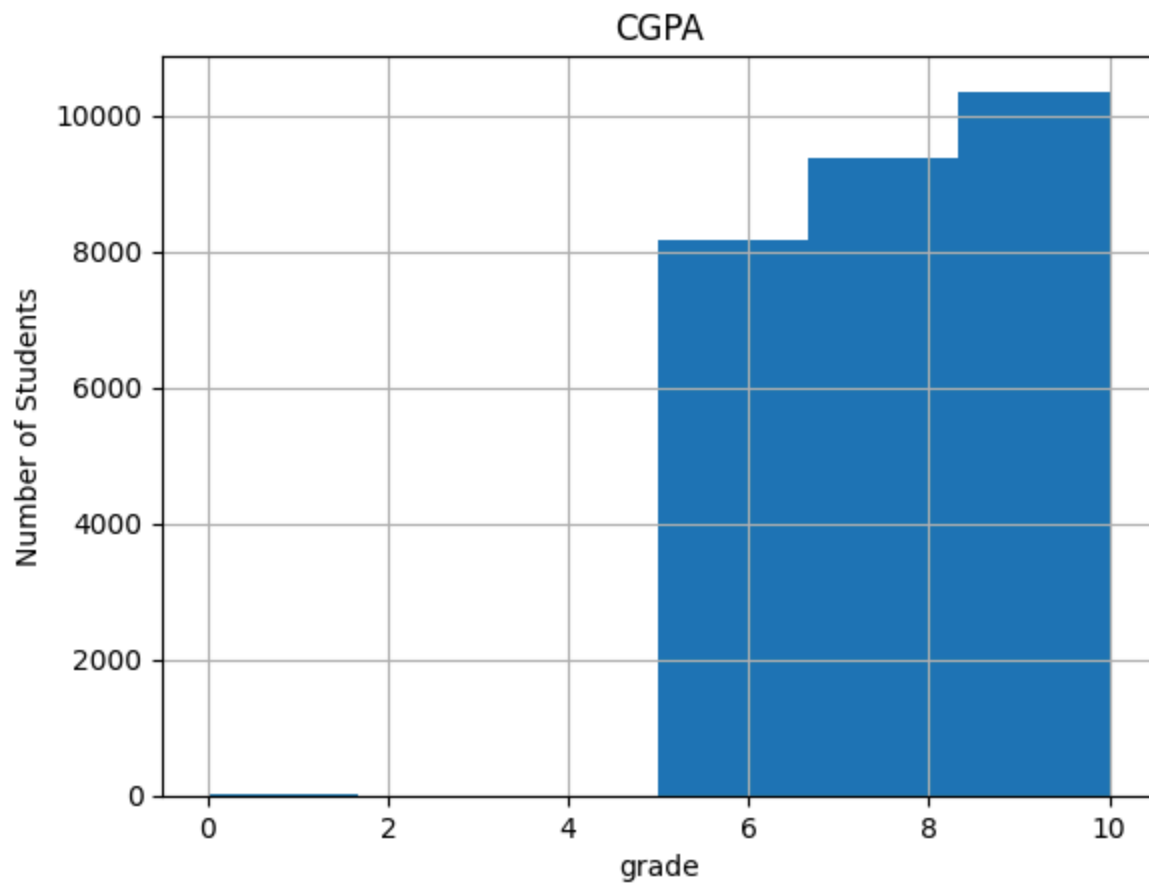
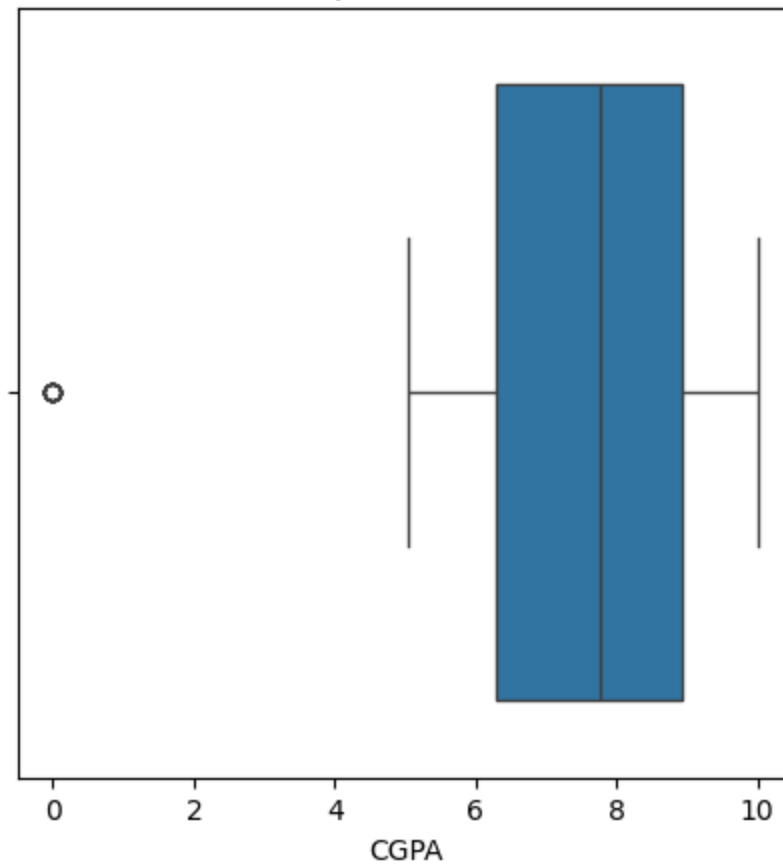
Boxplot for CGPA

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='CGPA', data=df)
plt.title('Boxplot of CGPA')
plt.show()
```

```
df['CGPA'].hist(bins=6)
plt.title('CGPA')
plt.xlabel('grade')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of CGPA



- The majority of students have a CGPA between 6 and 10, indicating generally good academic performance.
- The median CGPA is around 8, showing that 50% of students score above and below this value.
- The interquartile range (IQR), represented by the box, spans from approximately 6 to 9, covering the middle 50% of the data.
- There is at least one outlier below 1.0, which suggests a student with an unusually low CGPA compared to the rest.

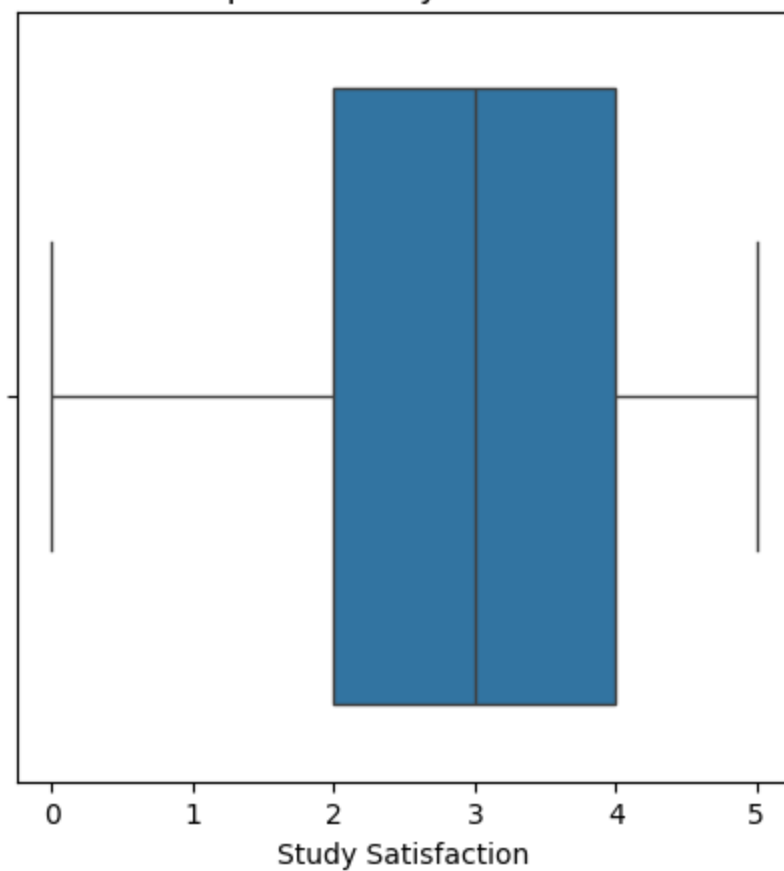
Boxplot and Histogram for Study Satisfaction

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='Study Satisfaction', data=df)
plt.title('Boxplot of Study Satisfaction')
plt.show()

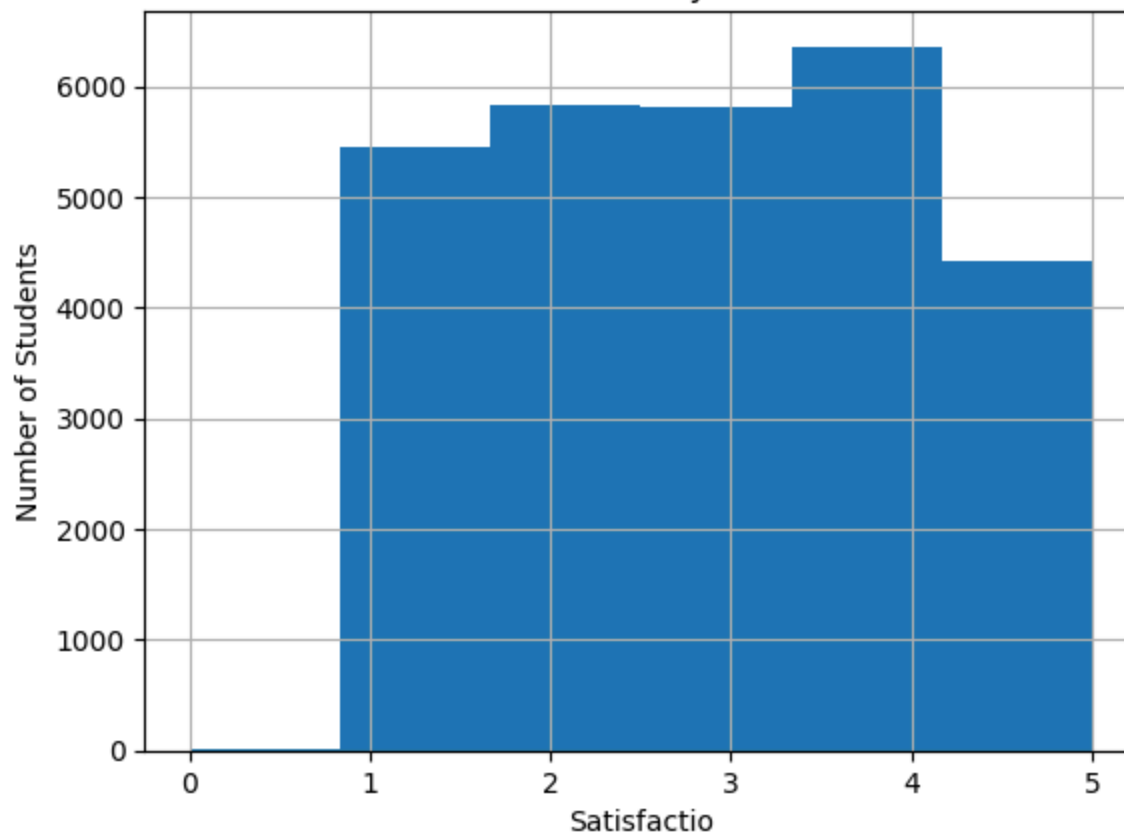
df['Study Satisfaction'].hist(bins=6)
plt.title('Distribution of Study Satisfaction')
plt.xlabel('Satisfactio')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of Study Satisfaction



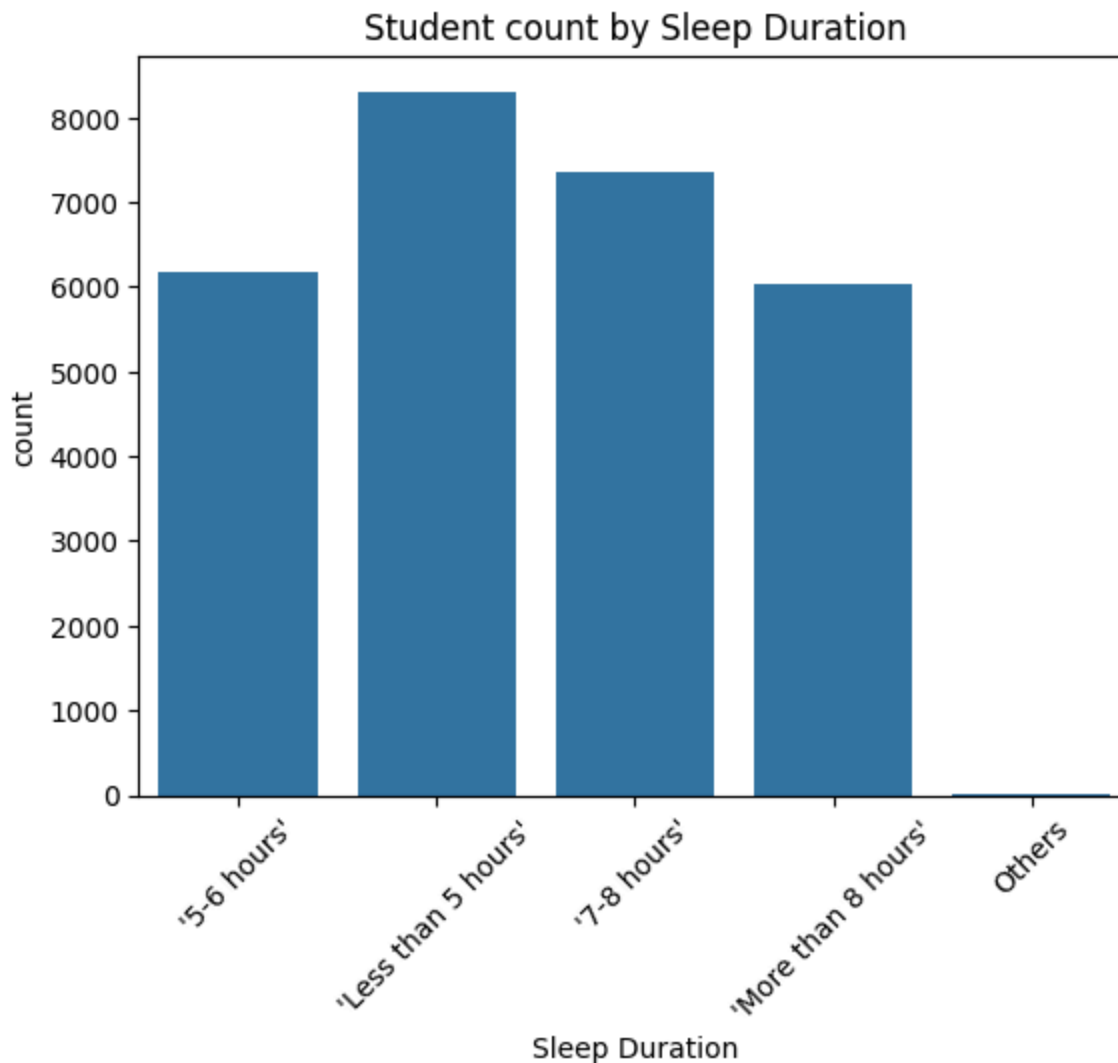
Distribution of Study Satisfaction



- Most students have a study satisfaction level between 2 and 4, indicating a generally moderate to high satisfaction with their studies.
- The median is close to 3, showing that half of the students rate their satisfaction at or above this level.
- The whiskers extend from about 0 to 5, meaning there are students who are both completely dissatisfied (0) and fully satisfied (5).
- There are no extreme outliers, suggesting that the responses are relatively well distributed within the scale.

Bar Chart for Sleep Duration

```
sns.countplot(x='Sleep Duration', data=df)  
plt.title('Student count by Sleep Duration')  
plt.xticks(rotation=45)  
plt.show()
```



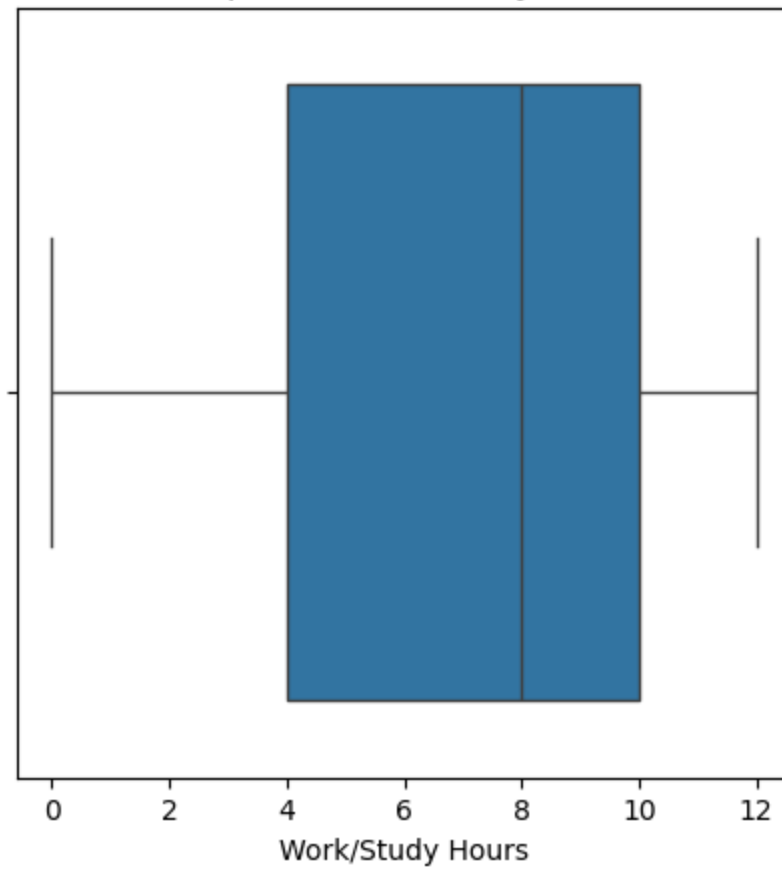
Boxplot and Histogram Work/Study Hours

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='Work/Study Hours', data=df)
plt.title('Boxplot of Work/Study Hours')
plt.show()

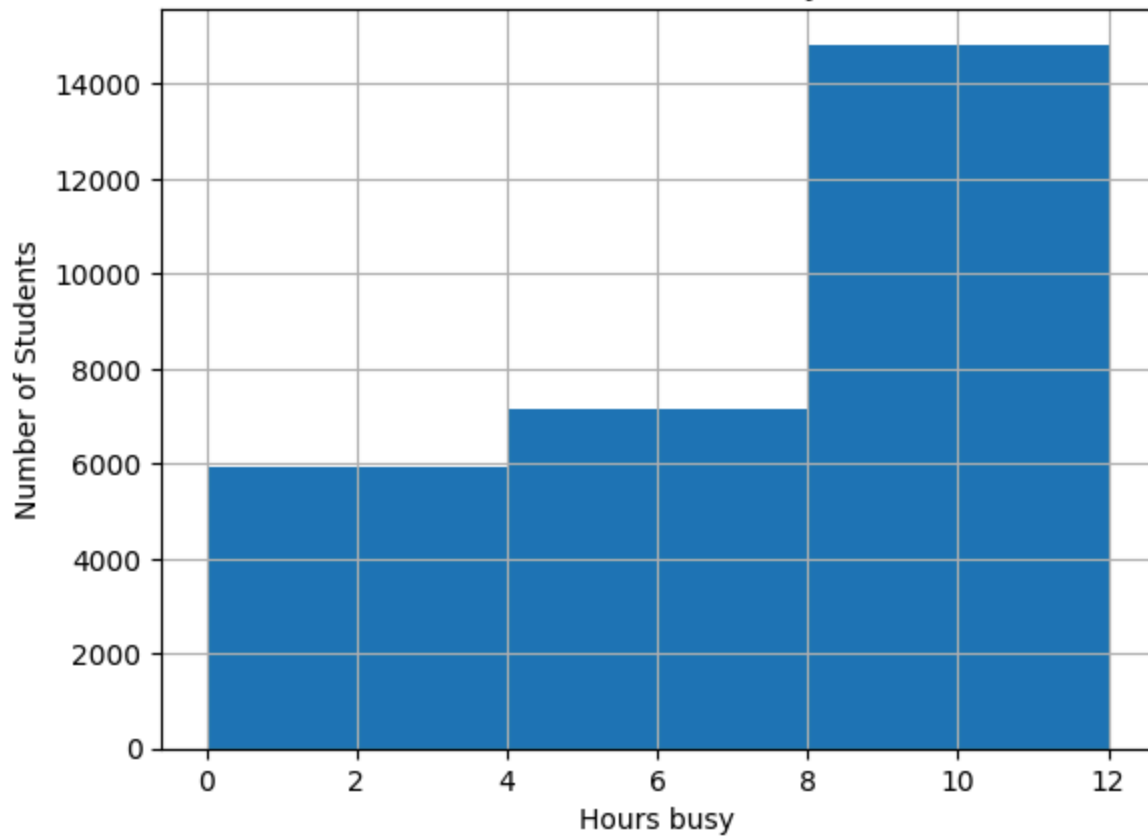
df['Work/Study Hours'].hist(bins=3)
plt.title('Distribution of Work/Study Hours')
plt.xlabel('Hours busy')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of Work/Study Hours



Distribution of Work/Study Hours



- Most students spend between 4 and 10 hours per day working or studying.
- The median is around 8 hours, suggesting that's a typical workload for many.
- The interquartile range (IQR) spans from about 4 to 10 hours, showing moderate variation.
- The minimum and maximum values range from 0 to around 12 hours.
- No extreme outliers are visible, indicating a consistent distribution.

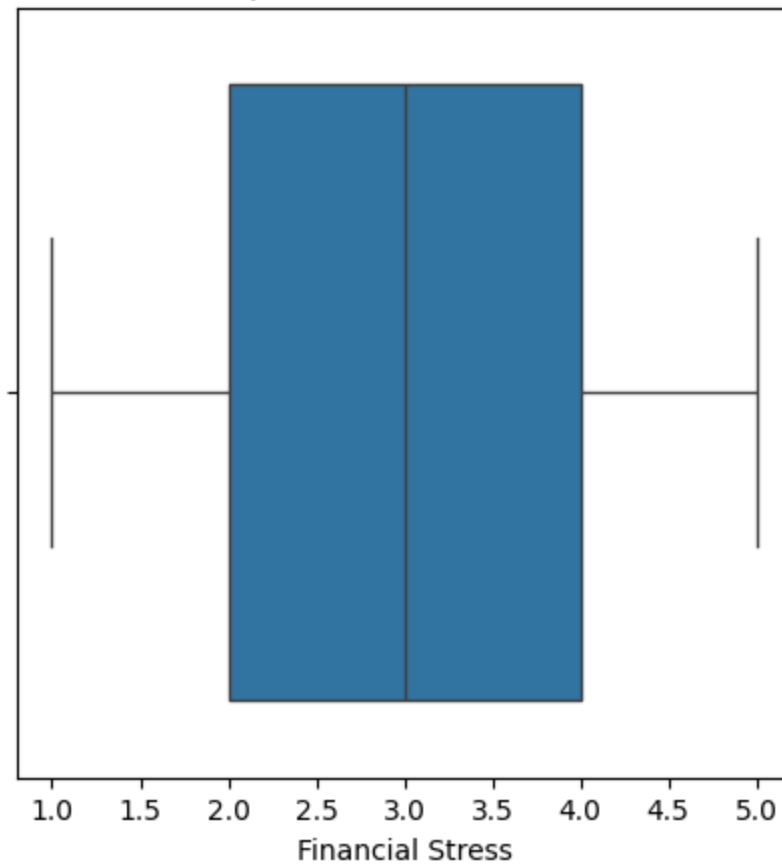
Boxplot and Histogram for Financial Stress

```
plt.figure(figsize=(5, 5))
sns.boxplot(x='Financial Stress', data=df)
plt.title('Boxplot of Financial Stress')
plt.show()

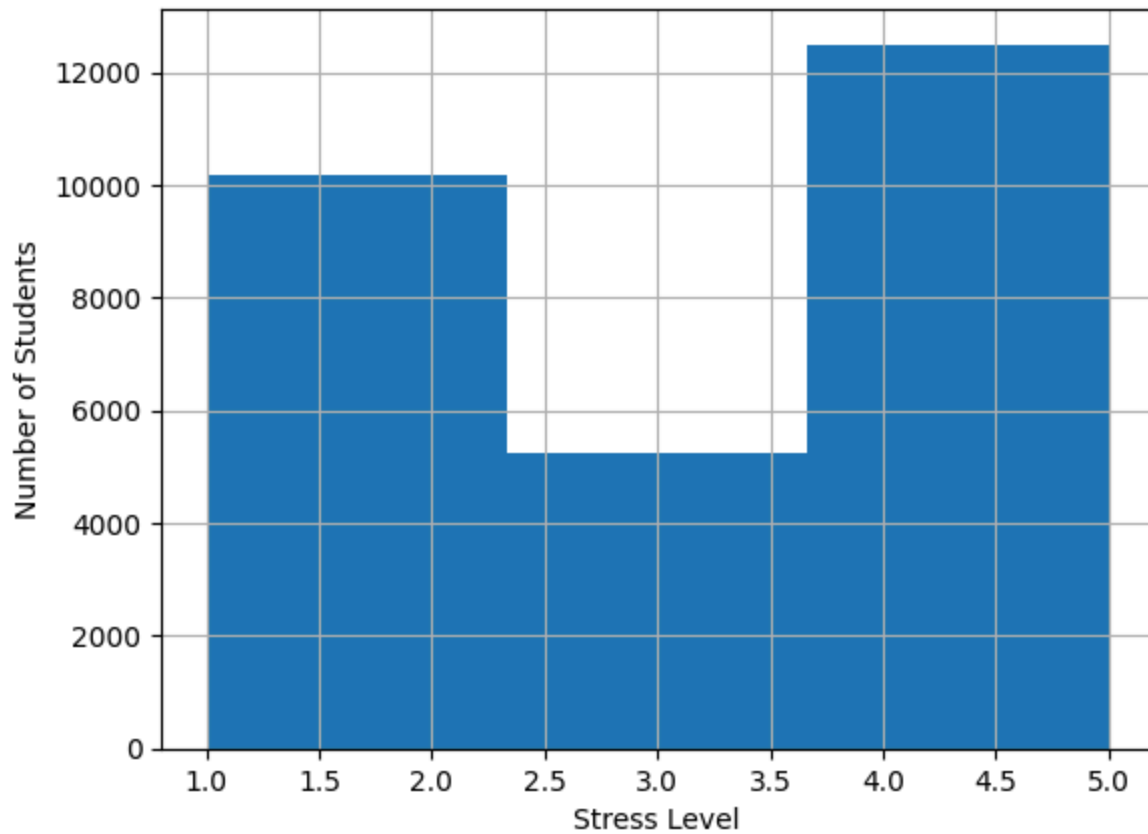
df['Financial Stress'].hist(bins=3)
plt.title('Distribution of Financial Stress')
plt.xlabel('Stress Level')
plt.ylabel('Number of Students')
plt.show()
```



Boxplot of Financial Stress



Distribution of Financial Stress



- Most students report a financial stress level between 2 and 5.
- The median is approximately 3, indicating moderate stress on average.
- The interquartile range (IQR) spans from around 2 to 5, showing variability in financial concerns.
- The minimum and maximum values extend from 1 to possibly 6 or 7 (the last tick is unclear in the plot).
- There are no visible outliers, which means students' responses are relatively consistent.

✓ Pearson Correlation

```
numeric_df = df.select_dtypes(include=['int64', 'float64'])  
correlation_pearson = numeric_df.corr(method='pearson')  
print("Pearson Correlation:\n", correlation_pearson)
```