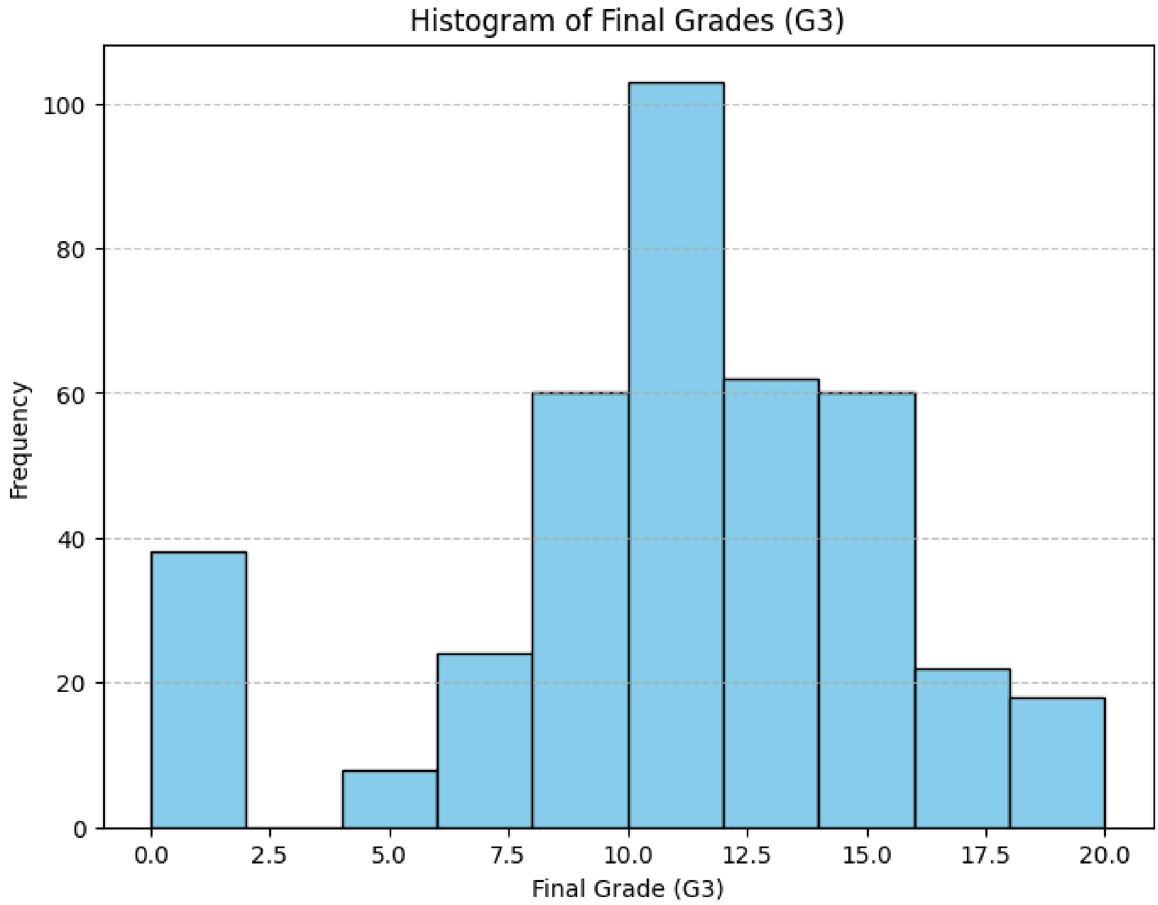
In [1]:

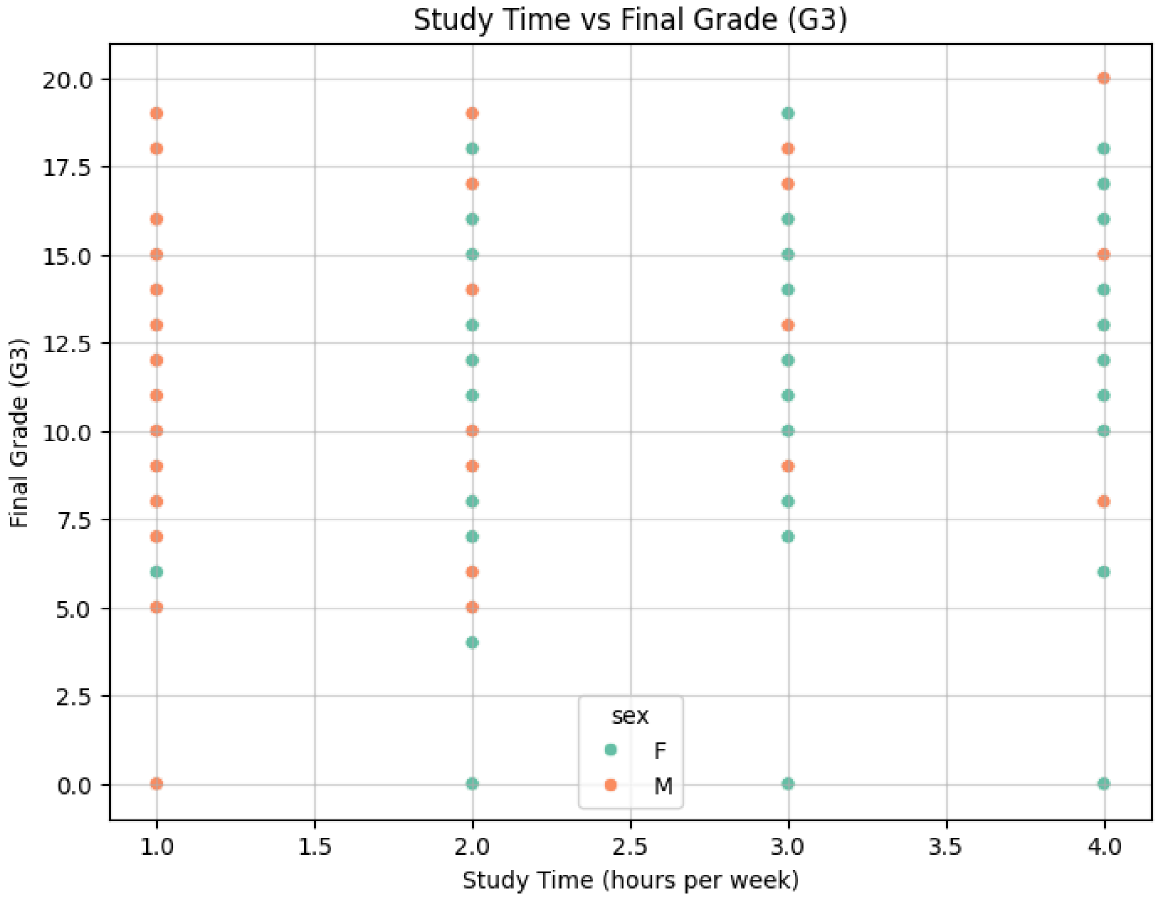
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
file_path = 'student-mat.csv'  # Update the path if necessary
data = pd.read_csv(file_path, delimiter=';')

# Display the first few rows
print("First 5 rows of the dataset:")
print(data.head())

# --- Data Exploration ---
print("\n--- Data Exploration ---")
# Check for missing values
missing_values = data.isnull().sum()
print("\nMissing Values in Each Column:")
print(missing_values)

# Check data types
print("\nData Types of Each Column:")
print(data.dtypes)

# Dataset size
print("\nDataset Size:")
print(data.shape)

# --- Data Cleaning ---
print("\n--- Data Cleaning ---")
# Remove duplicates
initial_rows = data.shape[0]
data = data.drop_duplicates()
duplicates_removed = initial_rows - data.shape[0]
print(f"Number of duplicate rows removed: {duplicates_removed}")

# --- Data Analysis ---
print("\n--- Data Analysis ---")
# 1. Average score in math (G3)
average_score = data['G3'].mean()
print(f"1. Average final grade (G3): {average_score:.2f}")

# 2. Number of students scoring above 15 in G3
students_above_15 = (data['G3'] > 15).sum()
print(f"2. Number of students scoring above 15 in G3: {students_above_15}")

# 3. Correlation between study time and G3
correlation = data['studytime'].corr(data['G3'])
print(f"3. Correlation between study time and G3: {correlation:.2f}")

# 4. Gender with a higher average G3
average_scores_by_gender = data.groupby('sex')['G3'].mean()
print("4. Average G3 by gender:")
print(average_scores_by_gender)

# --- Data Visualization ---
print("\n--- Data Visualization ---")
```

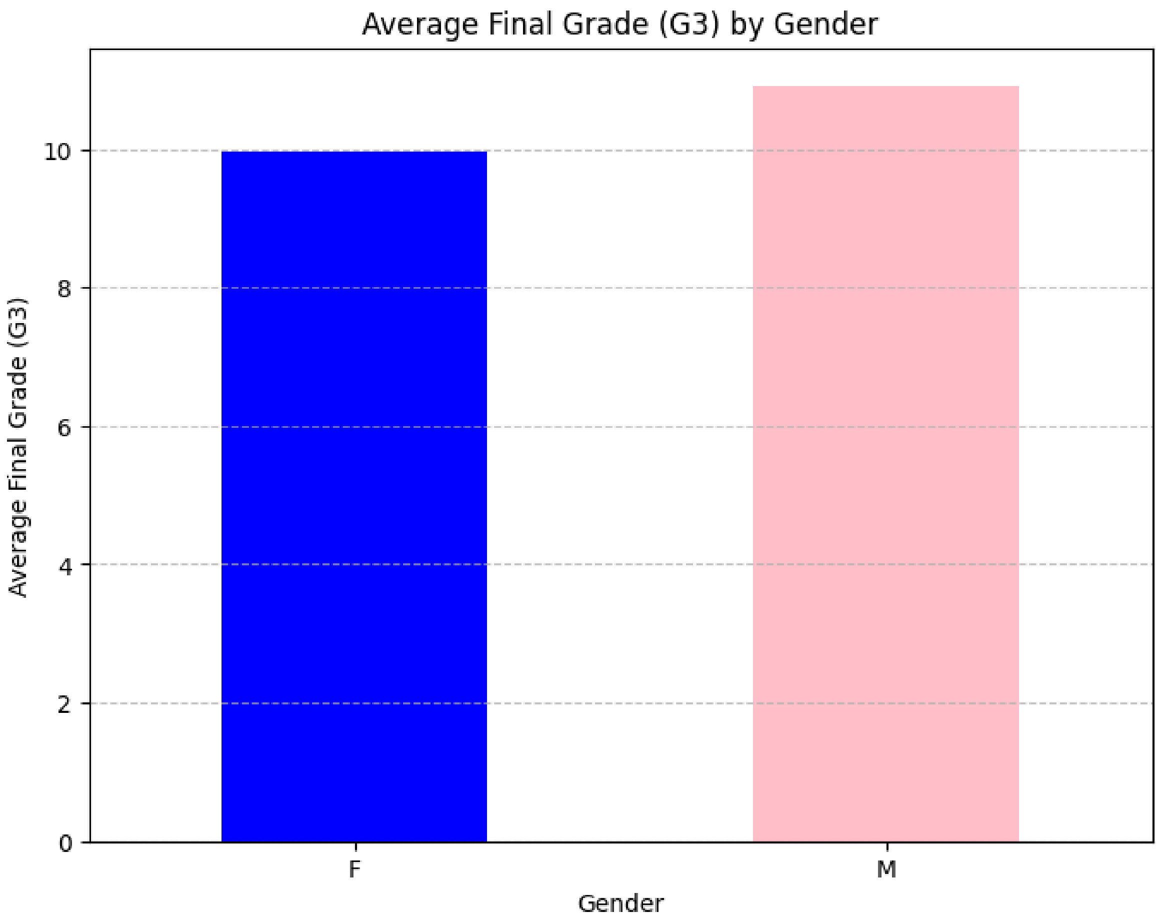
```python
# 1. Histogram of final grades (G3)
plt.figure(figsize=(8, 6))
plt.hist(data['G3'], bins=10, color='skyblue', edgecolor='black')
plt.title('Histogram of Final Grades (G3)')
plt.xlabel('Final Grade (G3)')
plt.ylabel('Frequency')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

# 2. Scatter plot between study time and final grade (G3)
plt.figure(figsize=(8, 6))
sns.scatterplot(x='studytime', y='G3', data=data, hue='sex', palette='Set2')
plt.title('Study Time vs Final Grade (G3)')
plt.xlabel('Study Time (hours per week)')
plt.ylabel('Final Grade (G3)')
plt.grid(alpha=0.5)
plt.show()

# 3. Bar chart comparing average scores of male and female students
plt.figure(figsize=(8, 6))
average_scores_by_gender.plot(kind='bar', color=['blue', 'pink'])
plt.title('Average Final Grade (G3) by Gender')
plt.xlabel('Gender')
plt.ylabel('Average Final Grade (G3)')
plt.xticks(rotation=0)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

```
First 5 rows of the dataset:
  school sex  age address famsize Pstatus  Medu  Fedu     Mjob      Fjob  ...  \
0     GP   F   18       U     GT3       A     4     4  at_home   teacher  ...
1     GP   F   17       U     GT3       T     1     1  at_home     other  ...
2     GP   F   15       U     LE3       T     1     1  at_home     other  ...
3     GP   F   15       U     GT3       T     4     2   health  services  ...
4     GP   F   16       U     GT3       T     3     3    other     other  ...

  famrel freetime  goout  Dalc  Walc health absences  G1  G2  G3
0      4        3      4     1     1      3        6   5   6   6
1      5        3      3     1     1      3        4   5   5   6
2      4        3      2     2     3      3       10   7   8  10
3      3        2      2     1     1      5        2  15  14  15
4      4        3      2     1     2      5        4   6  10  10

[5 rows x 33 columns]

--- Data Exploration ---

Missing Values in Each Column:
school        0
sex           0
age           0
address       0
famsize       0
Pstatus       0
Medu          0
Fedu          0
Mjob          0
Fjob          0
reason        0
guardian      0
traveltime    0
studytime     0
failures      0
schoolsup     0
famsup        0
paid          0
activities    0
nursery       0
higher        0
internet      0
romantic      0
famrel        0
freetime      0
goout         0
Dalc          0
Walc          0
health        0
absences      0
G1            0
G2            0
G3            0
dtype: int64

Data Types of Each Column:
school        object
sex           object
age            int64
address       object
```

```
famsize       object
Pstatus       object
Medu           int64
Fedu           int64
Mjob          object
Fjob          object
reason        object
guardian      object
traveltime     int64
studytime      int64
failures       int64
schoolsup     object
famsup        object
paid          object
activities    object
nursery       object
higher        object
internet      object
romantic      object
famrel         int64
freetime       int64
goout          int64
Dalc           int64
Walc           int64
health         int64
absences       int64
G1             int64
G2             int64
G3             int64
dtype: object

Dataset Size:
(395, 33)

--- Data Cleaning ---
Number of duplicate rows removed: 0

--- Data Analysis ---
1. Average final grade (G3): 10.42
2. Number of students scoring above 15 in G3: 40
3. Correlation between study time and G3: 0.10
4. Average G3 by gender:
sex
F     9.966346
M    10.914439
Name: G3, dtype: float64

--- Data Visualization ---
```


Histogram of Final Grades (G3)
Frequency
0
20
40
60
80
100
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Final Grade (G3)


Study Time vs Final Grade (G3)
Final Grade (G3)
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Study Time (hours per week)
sex
F
M



In [ ]:

In [2]:

```
print("Amit Kumar Jha")
```

```
Amit Kumar Jha
```

In [ ]: