Summary Statistics Python Notebok for Mini Project 3

Step 1: Import relevant Python packages and define functions for enerating summary statistics and producing a visualization.

In [1]:
```
import polars as pl
import matplotlib.pyplot as plt


def generate_summary_stats(file_name):
    """Using the csv file passed in as an argument, this function creates a
    dataframe from it, and then generates summary statistics (mean, median,
    mode, standard deviation, as well as percentiles) for each column of the
    using the polars describe method.
    """
    df = pl.read_csv(file_name)
    return df.describe(), df.median()


def generate_viz(file_name):
    """This function generates a scatter plot visualization of hours studied
    from the Student Performance dataset."""
    df = pl.read_csv(file_name)
    plt.scatter(df["Hours_Studied"], df["Exam_Score"], color="Green")
    plt.xlabel("Hours Studied")
    plt.ylabel("Student Exam Scores")
    plt.title("Relationship Between Hours Studied and Student Exam Scores")
    plt.savefig("performance.png")
    plt.show()
```

Step 2: Read in the StudentPerformanceFactors.csv file into a pandas dataframe.

In [2]:
```
student_df = pl.read_csv("StudentPerformanceFactors.csv")
student_df.head()
```

Out[2]: shape: (5, 20)

Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	Extracurricul
i64	i64	str	str	
23	84	"Low"	"High"	
19	64	"Low"	"Medium"	
24	98	"Medium"	"Medium"	
29	89	"Low"	"Medium"	
19	92	"Medium"	"Medium"	

Step 3: Review the summary statistics of the data set.

In [3]:
```
summary = generate_summary_stats("StudentPerformanceFactors.csv")
describe_stats = summary[0]
medians = summary[1]
print("Descriptive Statistics: \n", describe_stats, "\n")
print("Medians: \n", medians)
```

Descriptive Statistics:
shape: (9, 21)

statistic	Hours_Studied	Attendance	Parental_Involvement	…	Parental_Education_Level	Distance_from_Home	Gender	Exam_Score
str	f64	f64	str	…	str	str	str	f64
count	6607.0	6607.0	6607	…	6517	6540	6607	6607.0
null_count	0.0	0.0	0	…	90	67	0	0.0
mean	19.975329	79.977448	null	…	null	null	null	67.235659
std	5.990594	11.547475	null	…	null	null	null	3.890456
min	1.0	60.0	High	…	College	Far	Female	55.0
25%	16.0	70.0	null	…	null	null	null	65.0
50%	20.0	80.0	null	…	null	null	null	67.0
75%	24.0	90.0	null	…	null	null	null	69.0
max	44.0	100.0	Medium	…	Postgraduate	Near	Male	101.0

Medians:
shape: (1, 20)

Hours_Studied	Attendance	Parental_Involvement	Access_to_Resources	…	Parental_Education_Level	Distance_from_Home	Gender	Exam_Score
f64	f64	str	str	…	str	str	str	f64
20.0	80.0	null	null	…	null	null	null	67.0

```
| null    | 67.0       |
```

Step 4: Generate a visualization of the data set's columns (in this case, a scatterplot of hours studied vs. exam performance.)

In [4]:
```
generate_viz("StudentPerformanceFactors.csv")
```