Data Analysis and Visualization

Analysing Sleep Patterns of Students

By M.L.Prasanna
Electronics and
Communication Engineering

Introduction

Data visualization is a crucial step in any data analysis pipeline, helping to uncover patterns, trends, and insights. However, raw datasets are often messy, inconsistent, and incomplete. To ensure accurate and meaningful visualizations, thorough data preparation is essential.

Data Set Description

The dataset comprises **500** records capturing the sleep patterns and associated lifestyle factors of students. It aims to uncover how demographic and behavioral elements influence sleep quality and duration. The data includes a blend of numeric and categorical variables, providing a comprehensive view of student habits and demographics.

Key attributes of the dataset include student demographics such as age, gender (categorized as Male, Female, or Other), and the year of study in the university. These variables help us understand how sleep patterns vary across different groups.

Preparing and Handling the Data

M

Data Munging

Is the data structured properly? Are the column names clear and consistent?

C

Data Cleaning

Are there missing values or duplicates? Is the data consistent and error-free?

F

Filtering

Which data is relevant? What conditions should I use to filter?

A

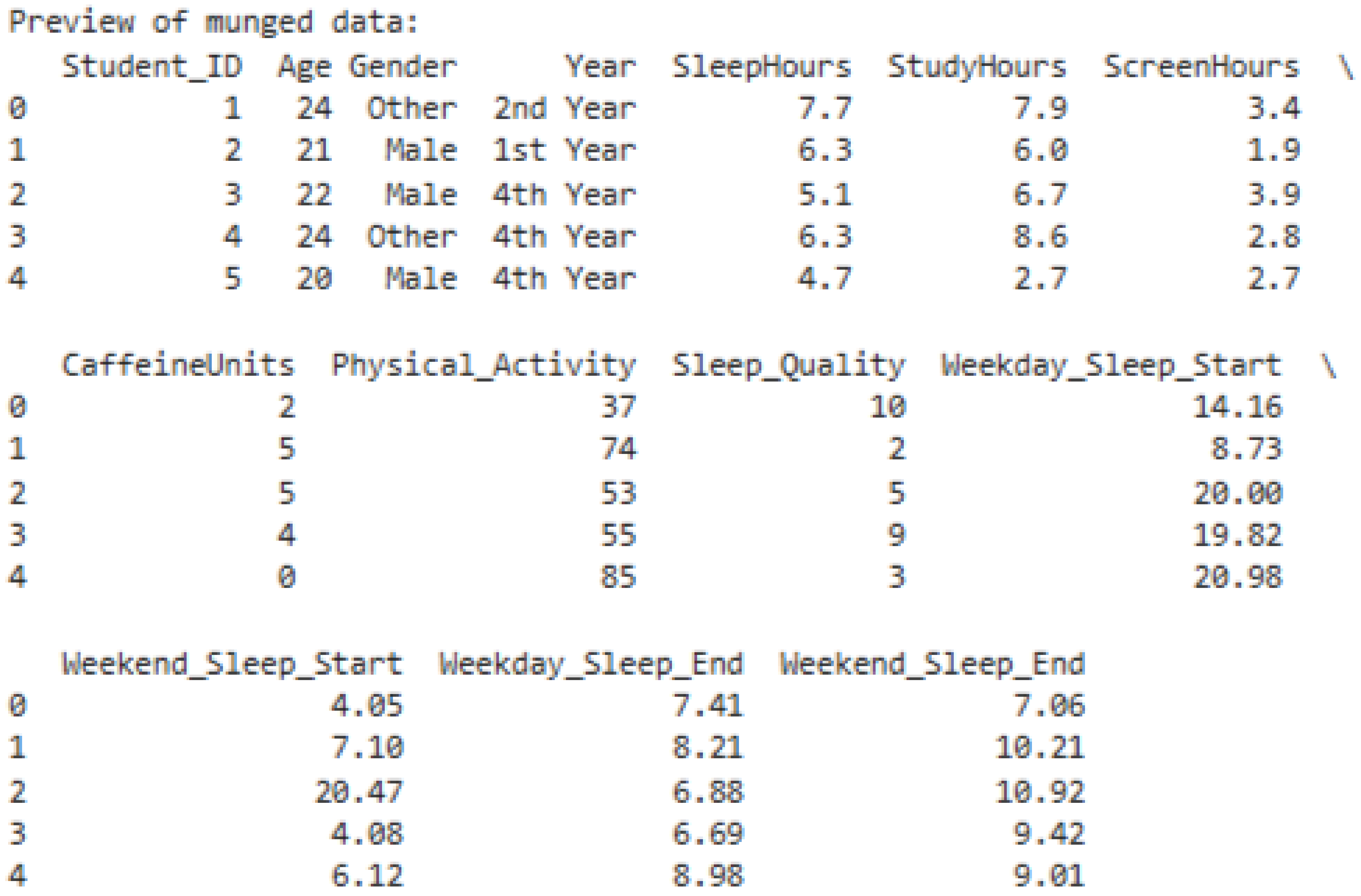
Aggregation

Do I need to combine datasets? How can I summarize the data effectively?

/05

Data Munging

Data munging transforms raw, unstructured data into a more usable format. This process involves renaming columns, converting data types, parsing dates, and reshaping the dataset to ensure compatibility with analytical and visualization tools. It’s the first step in making raw data ready for exploration and analysis.

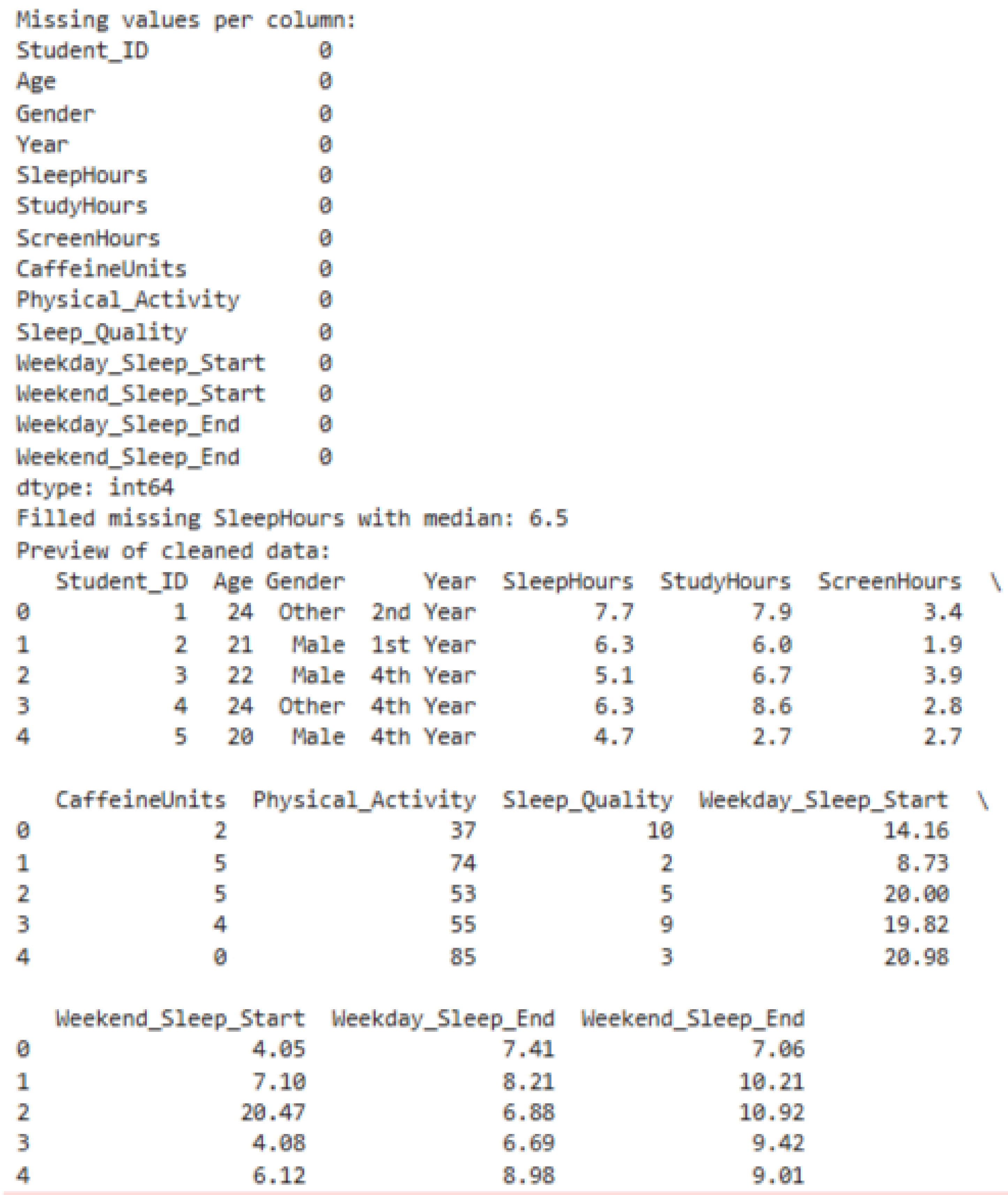


```
Preview of munged data:
   Student_ID  Age Gender      Year  SleepHours  StudyHours  ScreenHours  \
0           1   24  Other  2nd Year         7.7         7.9          3.4   
1           2   21   Male  1st Year         6.3         6.0          1.9   
2           3   22   Male  4th Year         5.1         6.7          3.9   
3           4   24  Other  4th Year         6.3         8.6          2.8   
4           5   20   Male  4th Year         4.7         2.7          2.7   

   CaffeineUnits  Physical_Activity  Sleep_Quality  Weekday_Sleep_Start  \
0              2                 37             10                14.16   
1              5                 74              2                 8.73   
2              5                 53              5                20.00   
3              4                 55              9                19.82   
4              0                 85              3                20.98   

   Weekend_Sleep_Start  Weekday_Sleep_End  Weekend_Sleep_End  
0                 4.05               7.41               7.06  
1                 7.10               8.21              10.21  
2                20.47               6.88              10.92  
3                 4.08               6.69               9.42  
4                 6.12               8.98               9.01  
```

/06

Data Cleaning

Data cleaning ensures the dataset is accurate and consistent. This step addresses missing values, removes duplicates, and standardizes entries. Clean data reduces the risk of misleading analyses and ensures reliability.

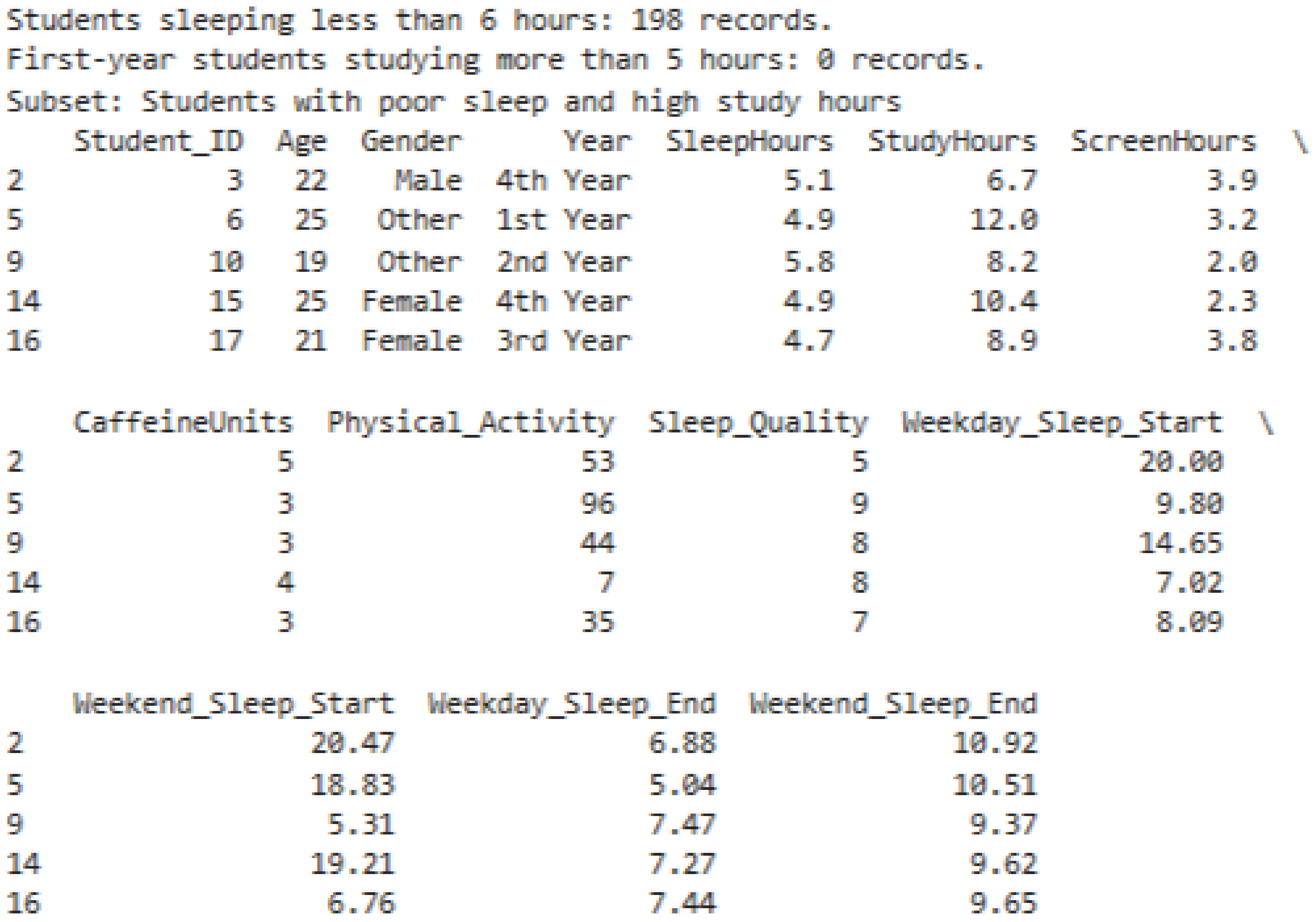


```
Missing values per column:
Student_ID           0
Age                  0
Gender               0
Year                 0
SleepHours           0
StudyHours           0
ScreenHours          0
CaffeineUnits        0
Physical_Activity    0
Sleep_Quality        0
Weekday_Sleep_Start  0
Weekend_Sleep_Start  0
Weekday_Sleep_End    0
Weekend_Sleep_End    0
dtype: int64
Filled missing SleepHours with median: 6.5
Preview of cleaned data:
   Student_ID  Age Gender      Year  SleepHours  StudyHours  ScreenHours  \
0           1   24  Other  2nd Year         7.7         7.9          3.4
1           2   21   Male  1st Year         6.3         6.0          1.9
2           3   22   Male  4th Year         5.1         6.7          3.9
3           4   24  Other  4th Year         6.3         8.6          2.8
4           5   20   Male  4th Year         4.7         2.7          2.7

   CaffeineUnits  Physical_Activity  Sleep_Quality  Weekday_Sleep_Start  \
0              2                 37             10                14.16
1              5                 74              2                 8.73
2              5                 53              5                20.00
3              4                 55              9                19.82
4              0                 85              3                20.98

   Weekend_Sleep_Start  Weekday_Sleep_End  Weekend_Sleep_End
0                 4.05               7.41               7.06
1                 7.10               8.21              10.21
2                20.47               6.88              10.92
3                 4.08               6.69               9.42
4                 6.12               8.98               9.01
```

/07

Filtering

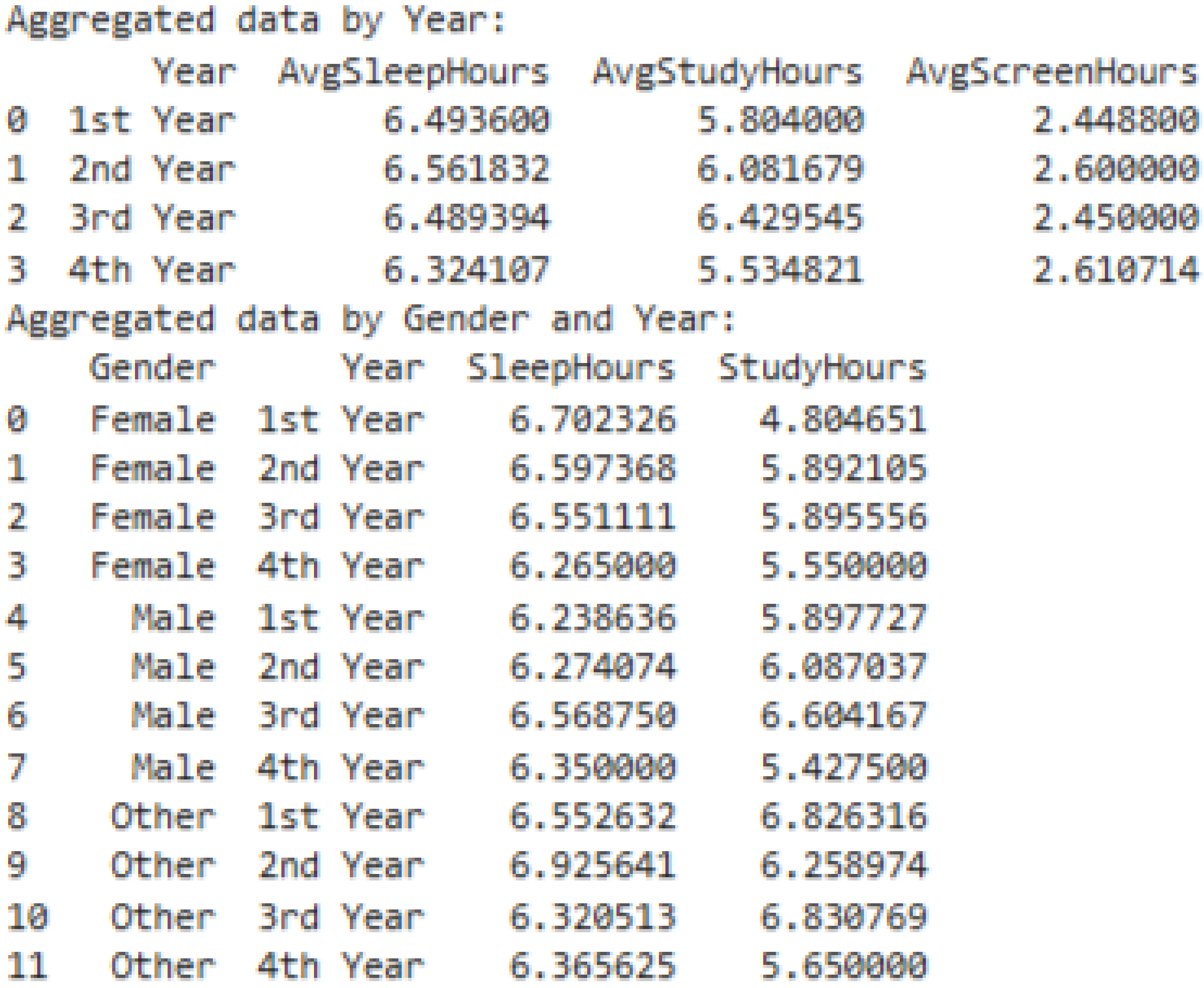
Filtering isolates relevant subsets of data based on specific conditions. This step focuses on answering questions or highlighting patterns. Filters can be numerical thresholds, categorical matches, or combinations of conditions



```
Students sleeping less than 6 hours: 198 records.
First-year students studying more than 5 hours: 0 records.
Subset: Students with poor sleep and high study hours
    Student_ID  Age  Gender      Year  SleepHours  StudyHours  ScreenHours  \
2            3   22    Male  4th Year         5.1         6.7          3.9
5            6   25   Other  1st Year         4.9        12.0          3.2
9           10   19   Other  2nd Year         5.8         8.2          2.0
14          15   25  Female  4th Year         4.9        10.4          2.3
16          17   21  Female  3rd Year         4.7         8.9          3.8

    CaffeineUnits  Physical_Activity  Sleep_Quality  Weekday_Sleep_Start  \
2               5                 53              5                20.00
5               3                 96              9                 9.80
9               3                 44              8                14.65
14              4                  7              8                 7.02
16              3                 35              7                 8.09

    Weekend_Sleep_Start  Weekday_Sleep_End  Weekend_Sleep_End
2                 20.47               6.88              10.92
5                 18.83               5.04              10.51
9                  5.31               7.47               9.37
14                19.21               7.27               9.62
16                 6.76               7.44               9.65
```

Aggregation

Data aggregation summarizes data to reveal trends and insights. By grouping data based on specific criteria (e.g., year or gender), you can compute metrics like averages, sums, and counts to analyze patterns across categories.

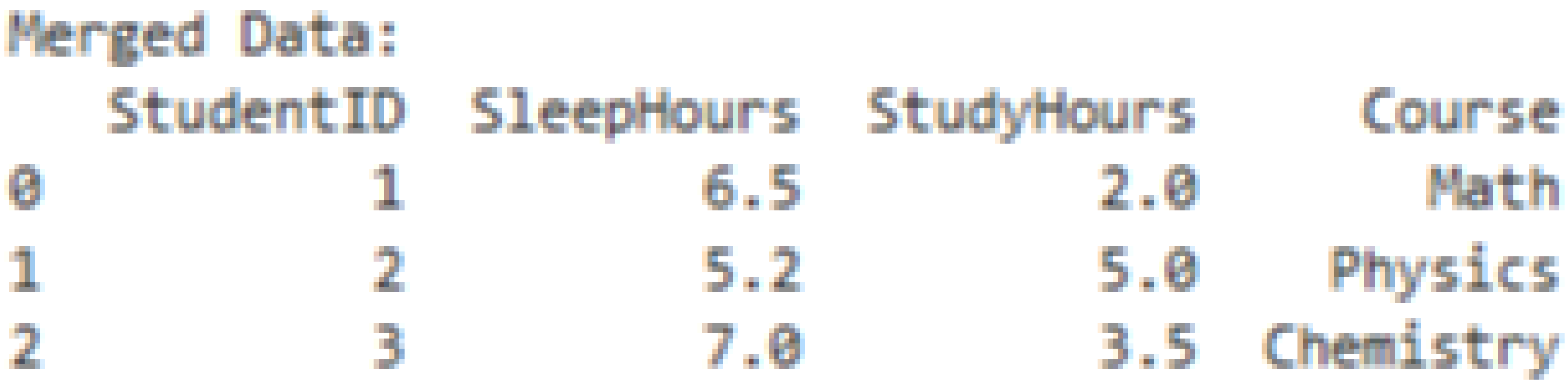


```
Aggregated data by Year:
       Year  AvgSleepHours  AvgStudyHours  AvgScreenHours
0  1st Year       6.493600       5.804000        2.448800
1  2nd Year       6.561832       6.081679        2.600000
2  3rd Year       6.489394       6.429545        2.450000
3  4th Year       6.324107       5.534821        2.610714
Aggregated data by Gender and Year:
    Gender      Year  SleepHours  StudyHours
0   Female  1st Year    6.702326    4.804651
1   Female  2nd Year    6.597368    5.892105
2   Female  3rd Year    6.551111    5.895556
3   Female  4th Year    6.265000    5.550000
4     Male  1st Year    6.238636    5.897727
5     Male  2nd Year    6.274074    6.087037
6     Male  3rd Year    6.568750    6.604167
7     Male  4th Year    6.350000    5.427500
8    Other  1st Year    6.552632    6.826316
9    Other  2nd Year    6.925641    6.258974
10   Other  3rd Year    6.320513    6.830769
11   Other  4th Year    6.365625    5.650000
```

/09

Data Merging

Merging combines two or more datasets based on a common column or index. It’s essential for integrating different sources of information, enabling a comprehensive analysis. Merge types include inner, outer, left, and right, each serving different purposes depending on the need.

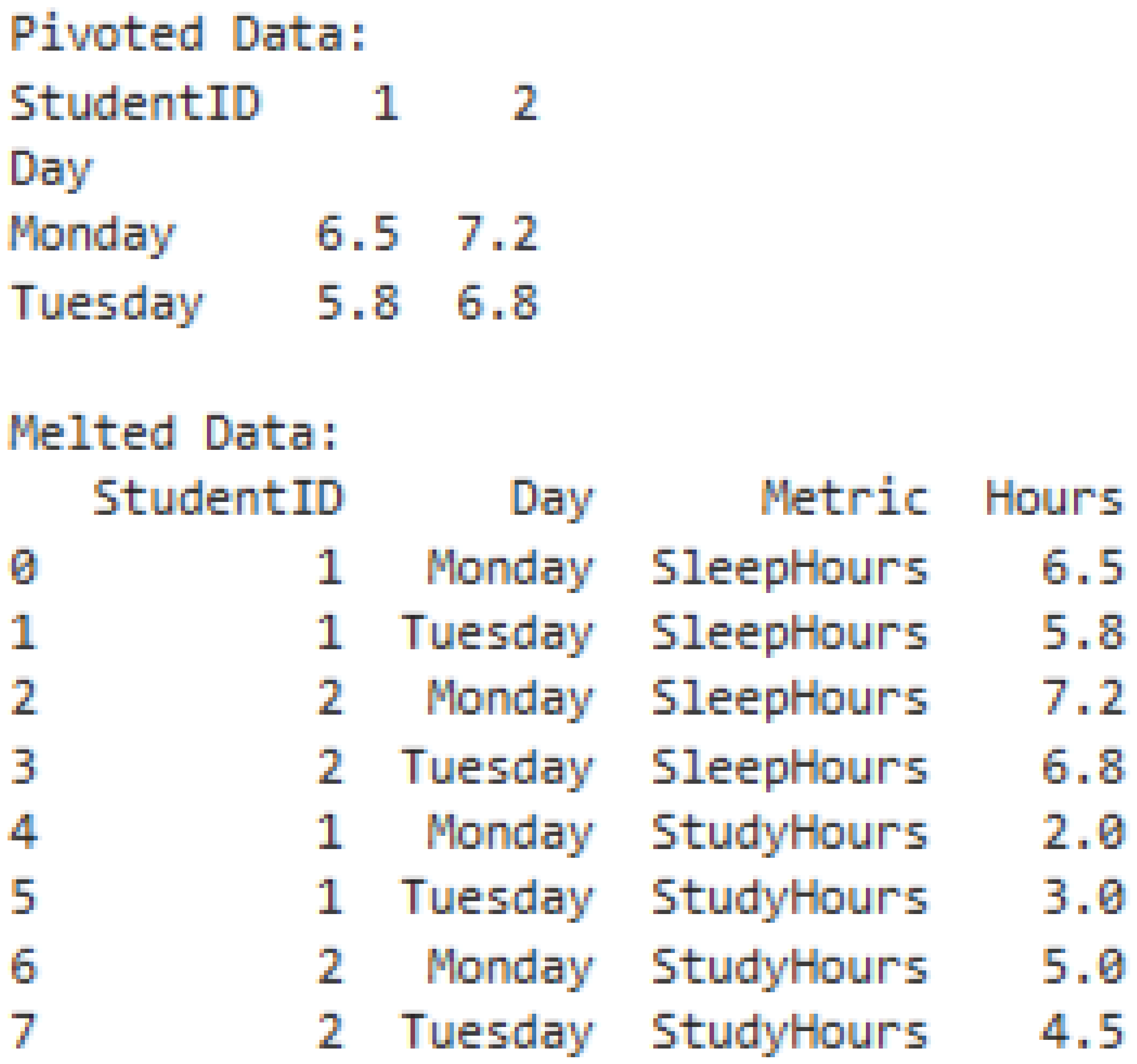


```
Merged Data:
   StudentID  SleepHours  StudyHours     Course
0          1         6.5         2.0       Math
1          2         5.2         5.0    Physics
2          3         7.0         3.5  Chemistry
```

Data Reshaping

Reshaping changes the structure of the dataset to suit different types of analysis. The common operations are:

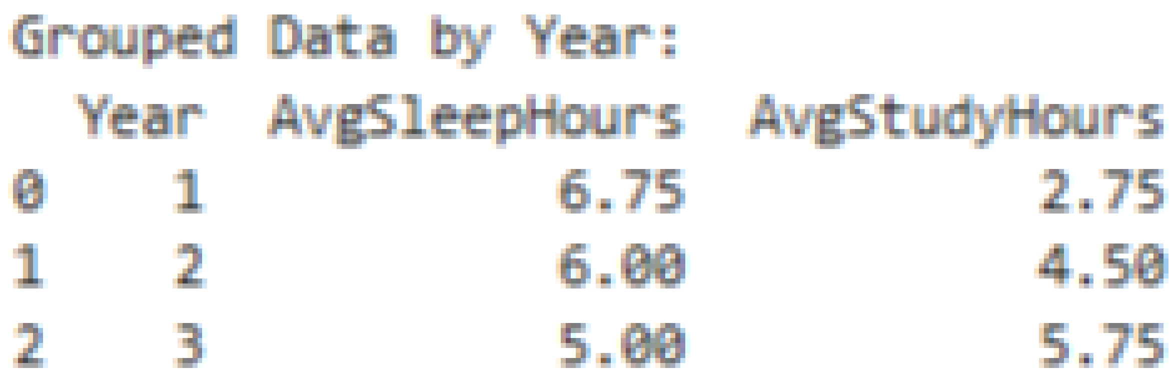
- Pivot: Rearranges data to summarize information.
- Melt: Converts wide-format data into a long format.
- Stack/Unstack: Modifies multi-level indexed data for flexibility.



```
Pivoted Data:
StudentID    1    2
Day
Monday     6.5  7.2
Tuesday    5.8  6.8

Melted Data:
   StudentID      Day       Metric  Hours
0          1   Monday   SleepHours    6.5
1          1  Tuesday   SleepHours    5.8
2          2   Monday   SleepHours    7.2
3          2  Tuesday   SleepHours    6.8
4          1   Monday   StudyHours    2.0
5          1  Tuesday   StudyHours    3.0
6          2   Monday   StudyHours    5.0
7          2  Tuesday   StudyHours    4.5
```

Data Grouping

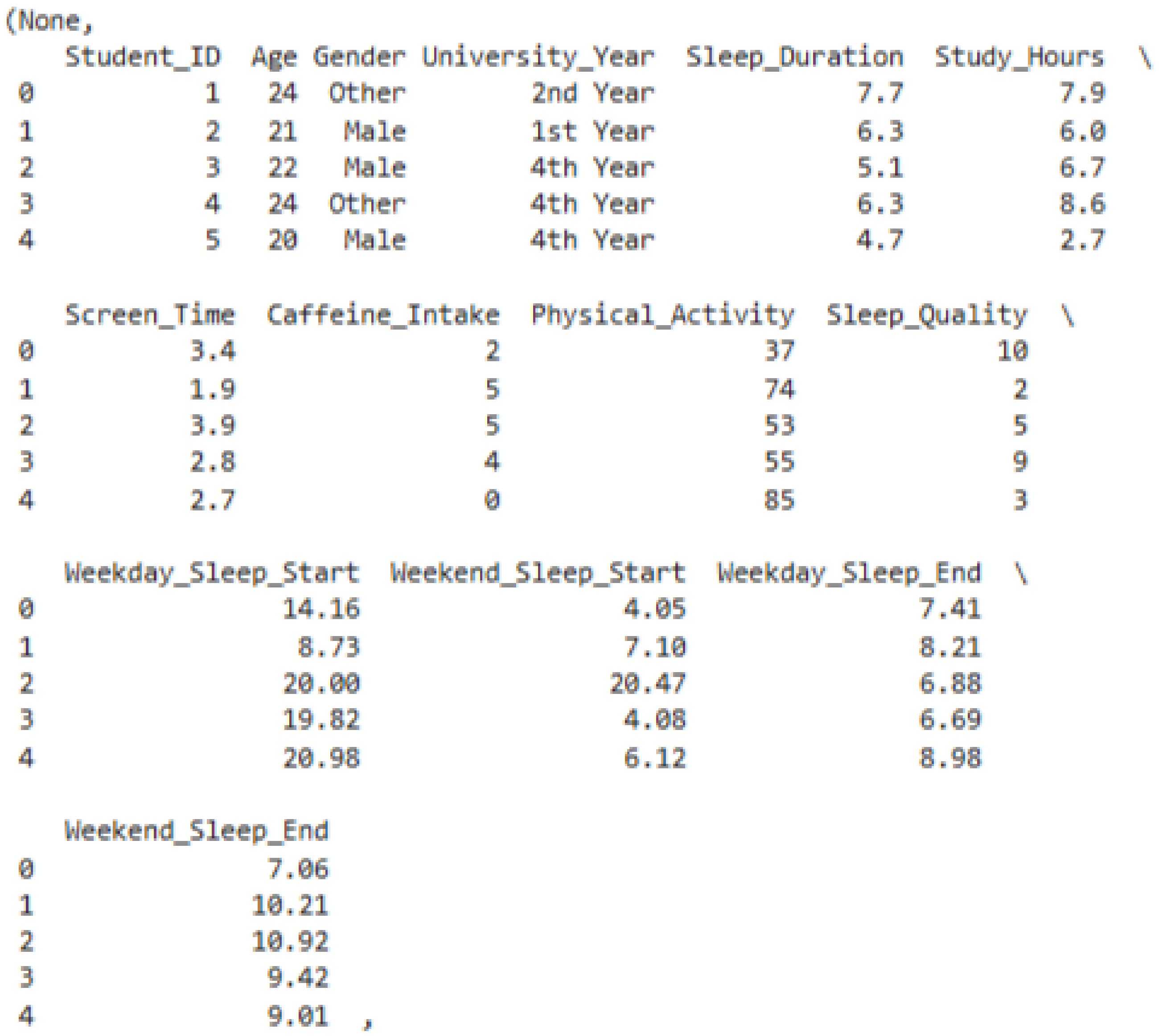
Grouping organizes data into categories and performs aggregate operations like sum, mean, count, etc. It’s ideal for identifying trends or patterns in categorical data.



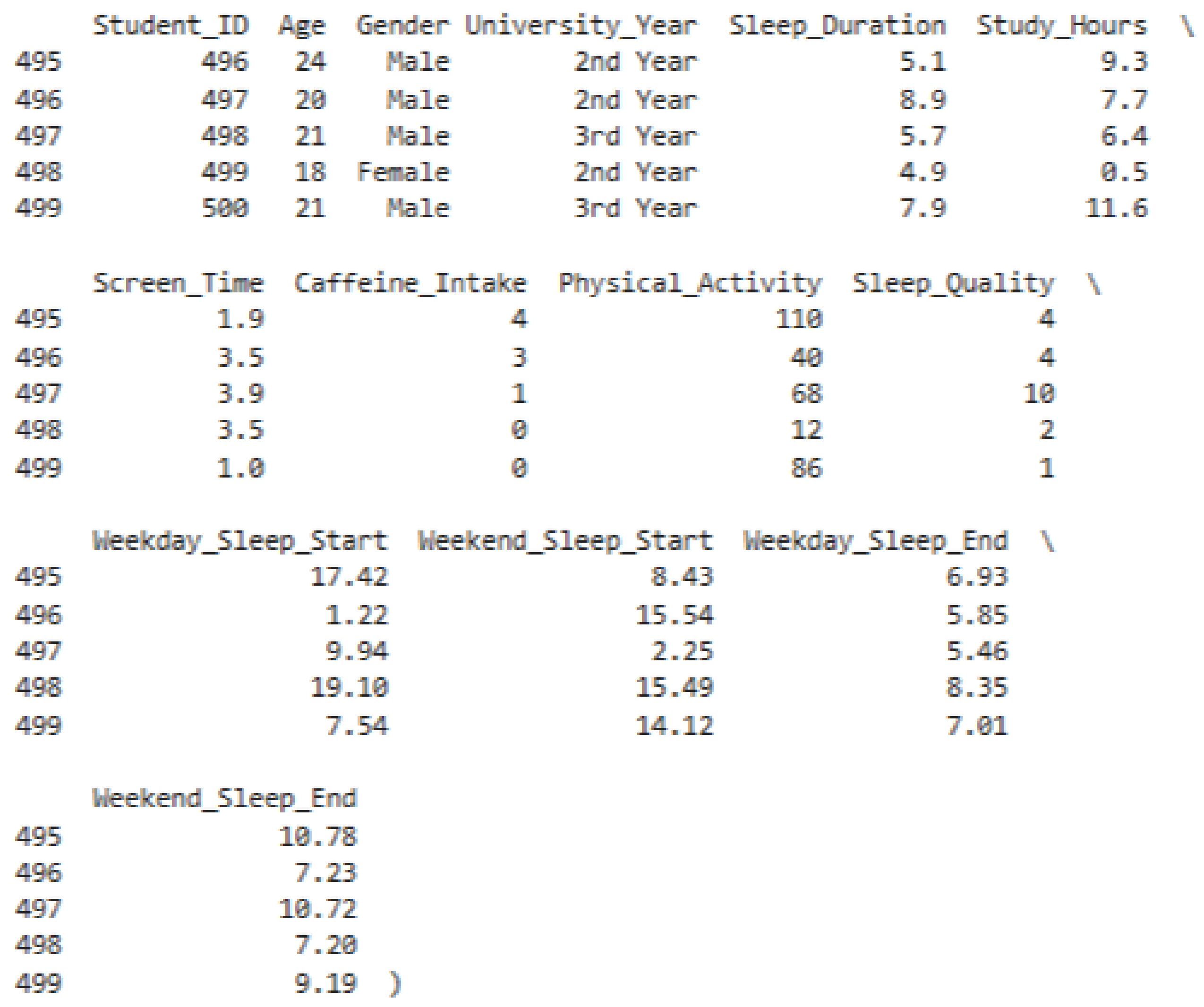
```
Grouped Data by Year:
   Year  AvgSleepHours  AvgStudyHours
0     1           6.75           2.75
1     2           6.00           4.50
2     3           5.00           5.75
```

HEAD AND TAIL

In data analysis, head displays the first few rows (default is 5) of a dataset, providing a quick overview of its structure, columns, and initial values. Tail shows the last few rows, useful for checking the dataset's conclusion, identifying missing data, or understanding patterns at the end. Both are essential for data inspection and exploration.



```
(None,
   Student_ID  Age Gender University_Year  Sleep_Duration  Study_Hours  \
0           1   24  Other        2nd Year             7.7          7.9
1           2   21   Male        1st Year             6.3          6.0
2           3   22   Male        4th Year             5.1          6.7
3           4   24  Other        4th Year             6.3          8.6
4           5   20   Male        4th Year             4.7          2.7

   Screen_Time  Caffeine_Intake  Physical_Activity  Sleep_Quality  \
0          3.4                2                 37             10
1          1.9                5                 74              2
2          3.9                5                 53              5
3          2.8                4                 55              9
4          2.7                0                 85              3

   Weekday_Sleep_Start  Weekend_Sleep_Start  Weekday_Sleep_End  \
0                14.16                 4.05               7.41
1                 8.73                 7.10               8.21
2                20.00                20.47               6.88
3                19.82                 4.08               6.69
4                20.98                 6.12               8.98

   Weekend_Sleep_End
0               7.06
1              10.21
2              10.92
3               9.42
4               9.01  ,
```



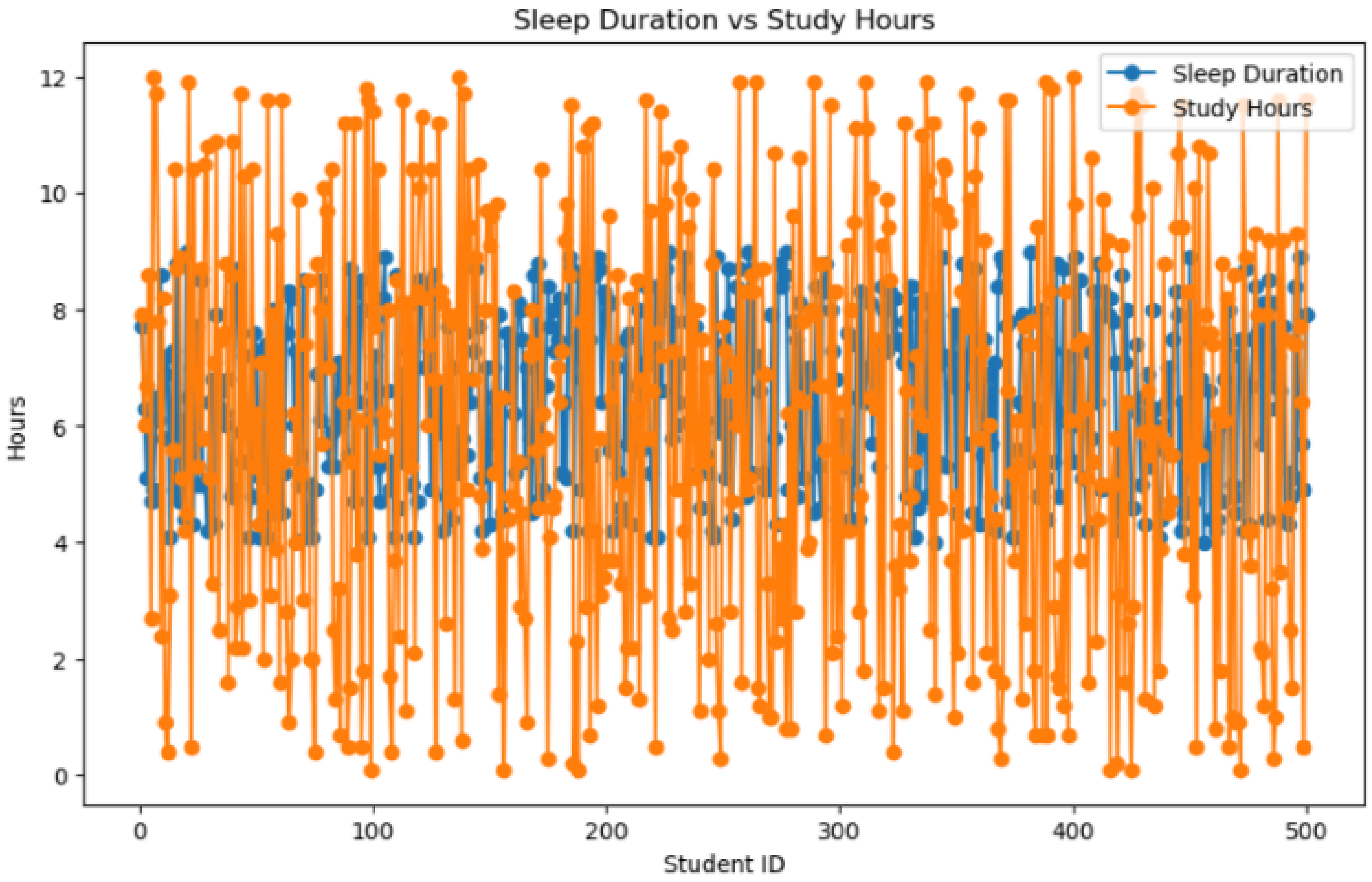
```
     Student_ID  Age  Gender University_Year  Sleep_Duration  Study_Hours  \
495         496   24    Male        2nd Year             5.1          9.3
496         497   20    Male        2nd Year             8.9          7.7
497         498   21    Male        3rd Year             5.7          6.4
498         499   18  Female        2nd Year             4.9          0.5
499         500   21    Male        3rd Year             7.9         11.6

     Screen_Time  Caffeine_Intake  Physical_Activity  Sleep_Quality  \
495          1.9                4                110              4
496          3.5                3                 40              4
497          3.9                1                 68             10
498          3.5                0                 12              2
499          1.0                0                 86              1

     Weekday_Sleep_Start  Weekend_Sleep_Start  Weekday_Sleep_End  \
495                17.42                 8.43               6.93
496                 1.22                15.54               5.85
497                 9.94                 2.25               5.46
498                19.10                15.49               8.35
499                 7.54                14.12               7.01

     Weekend_Sleep_End
495              10.78
496               7.23
497              10.72
498               7.20
499               9.19  )
```

Data Visualization

Data visualization transforms complex datasets into graphical formats, aiding better comprehension. It includes line plots for trends, scatter plots for correlations, bar graphs for comparisons, pie charts for proportions, and histograms for distributions. Techniques like subplots, legends, and annotations enhance clarity. Tools like Matplotlib, Pandas, and NumPy make creating insightful visual representations straightforward and efficient.

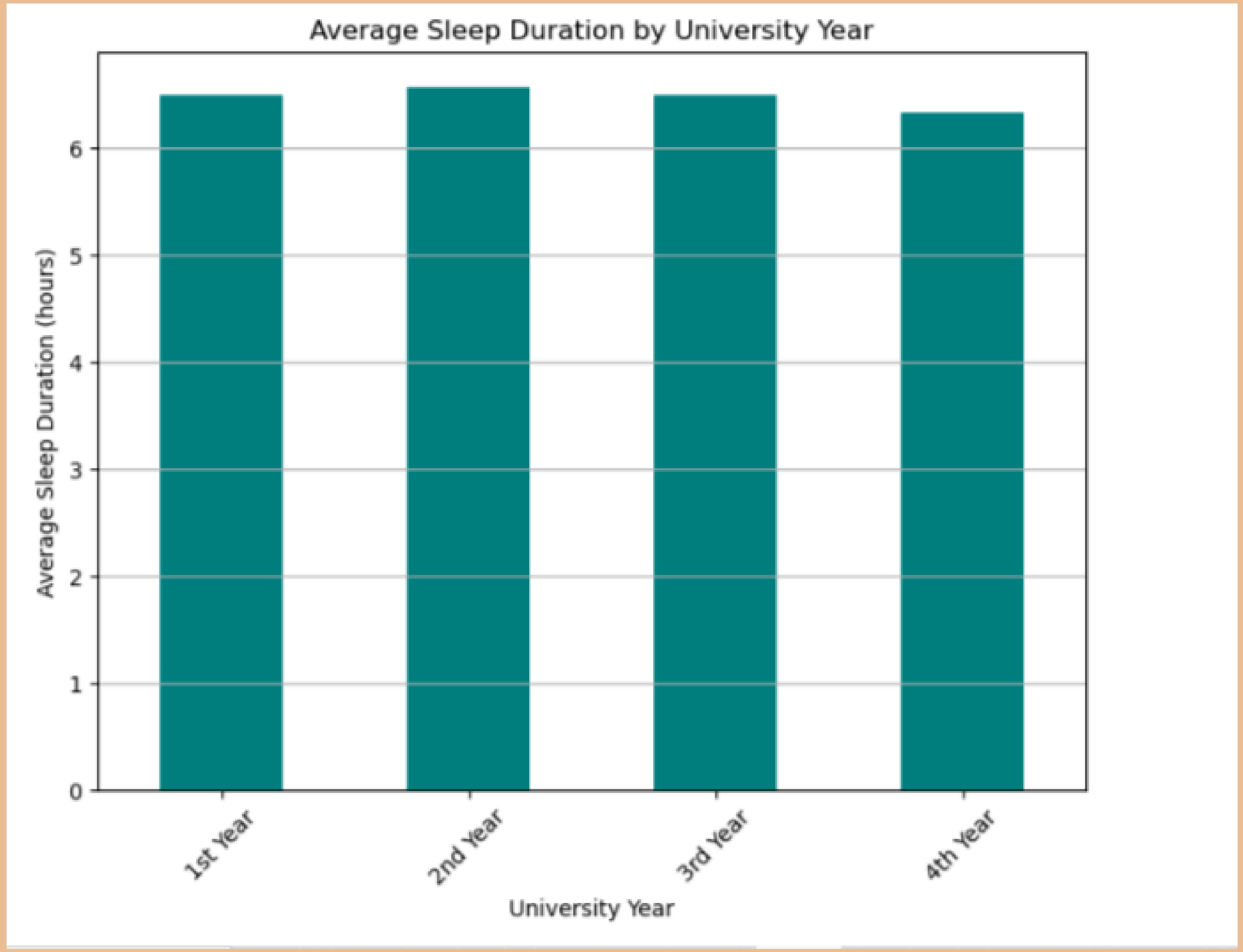
Line Plots

Line plots are ideal for visualizing trends in data over time or across ordered categories. They are particularly useful for showing changes in variables like temperature, stock prices, or, in this case, the relationship between sleep duration and study hours. By connecting data points with a line, line plots allow for easy identification of patterns and fluctuations. They are commonly used when both the x-axis and y-axis represent continuous or sequential data points.



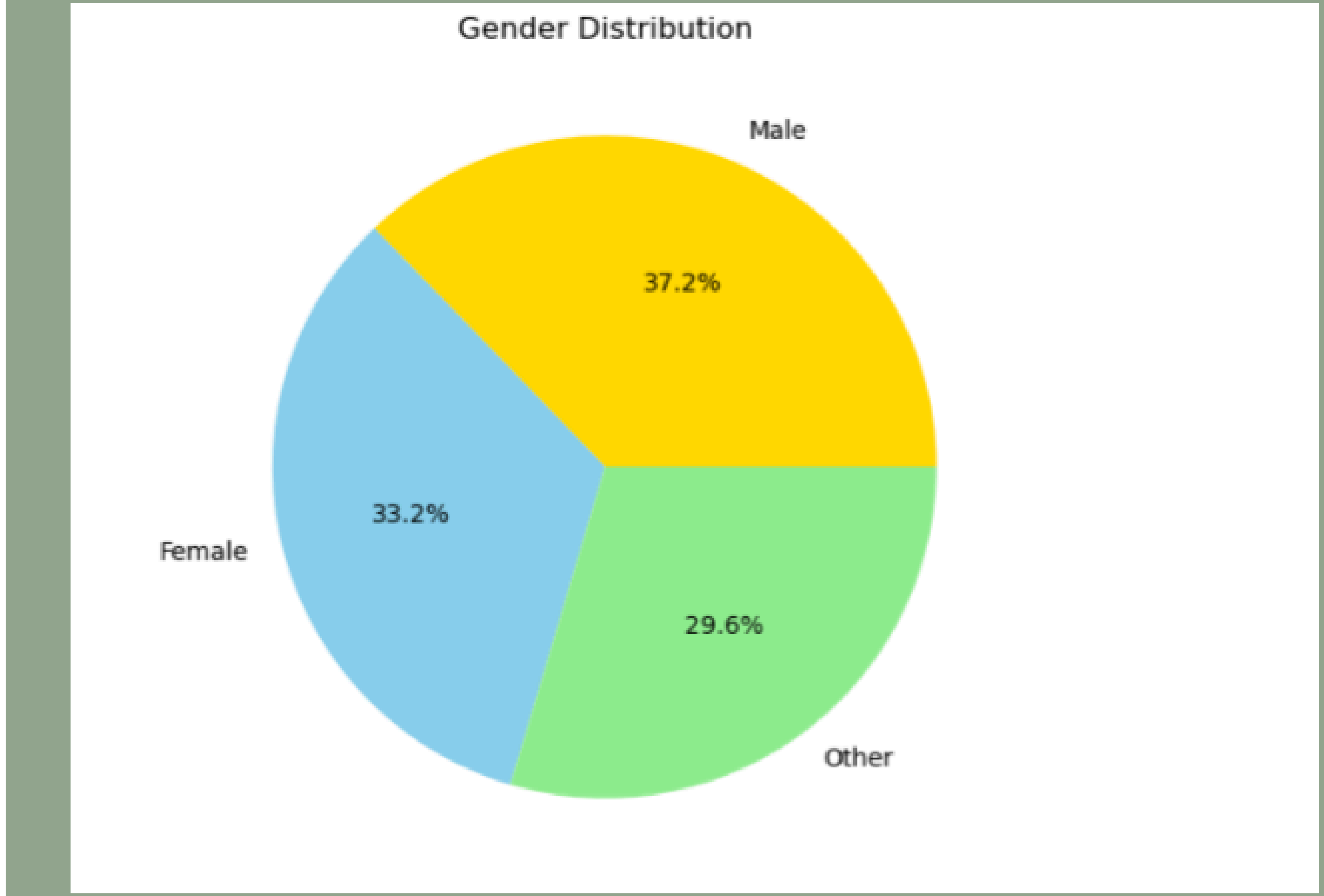
Bar Graphs

.Bar graphs are effective for comparing data across distinct categories, making them ideal for visualizing differences in groups like academic years, product sales, or survey results. Each bar represents a category, with the height reflecting its value. For instance, a bar graph showcasing average sleep duration by university year provides a clear comparison of how sleep patterns vary among students. Bar graphs are versatile, easy to interpret, and widely used in categorical data analysis.



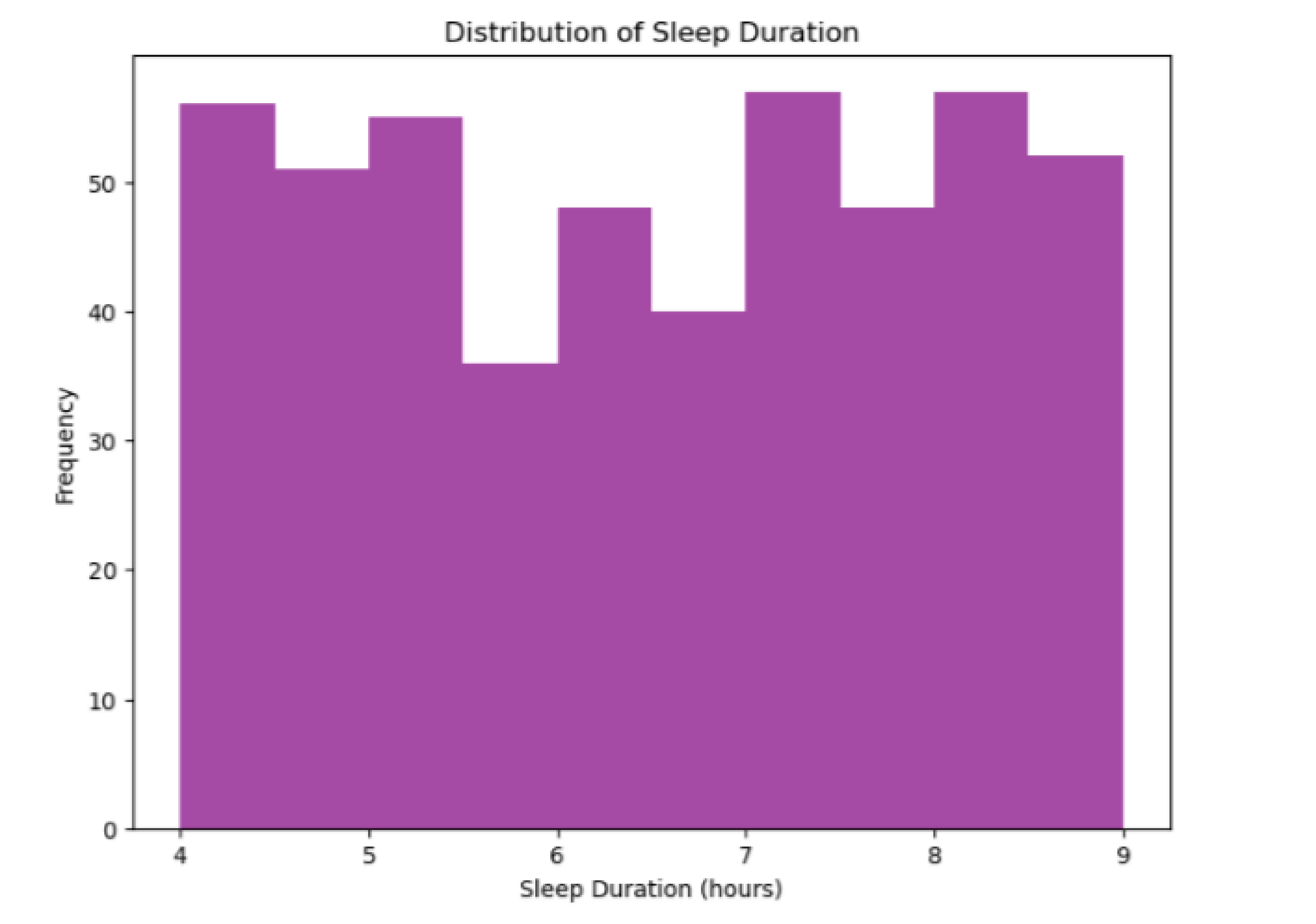
Pie Chart

Pie charts are perfect for visualizing proportions and distributions within a dataset, displaying how parts contribute to a whole. Each slice represents a category, with its size corresponding to its percentage of the total. For example, a pie chart showing gender distribution in a dataset effectively illustrates the share of male and female participants. Pie charts are simple and visually appealing, making them ideal for representing percentage-based data.



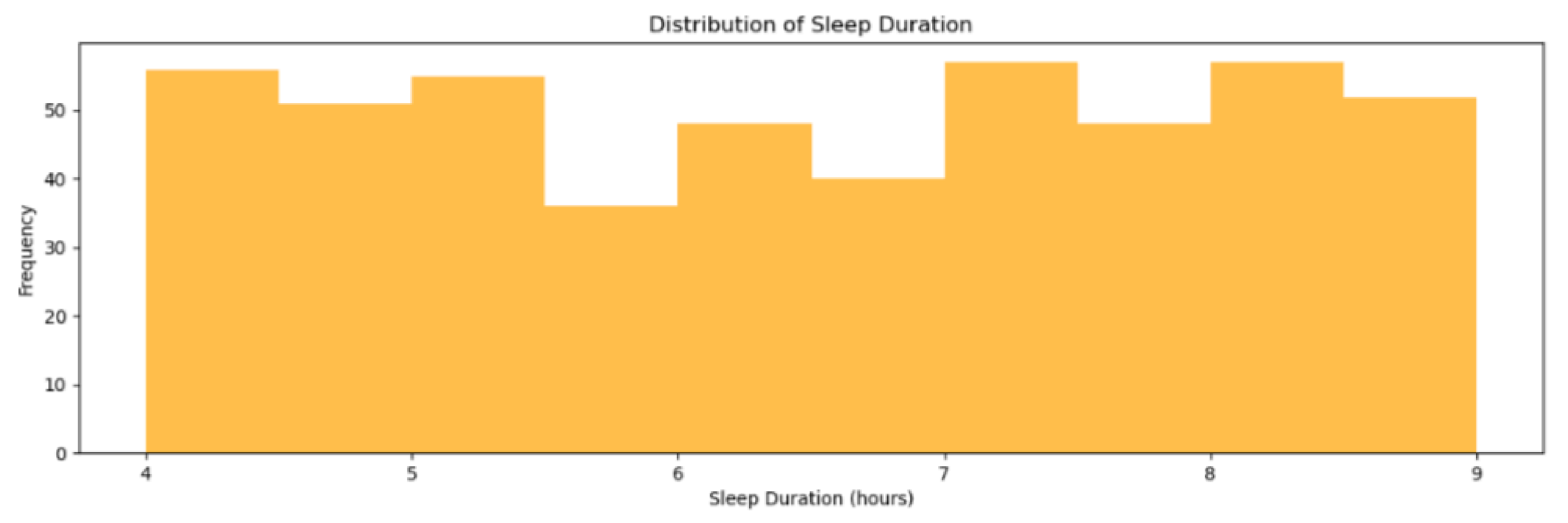
Histograms

Histograms are useful for understanding the frequency distribution of numerical data. They divide data into intervals, or bins, and display the number of occurrences within each bin. This helps identify patterns such as skewness, peaks, and gaps. For instance, a histogram showing the frequency of sleep duration highlights how often certain sleep intervals occur among participants, providing insights into common and outlier patterns in the dataset.

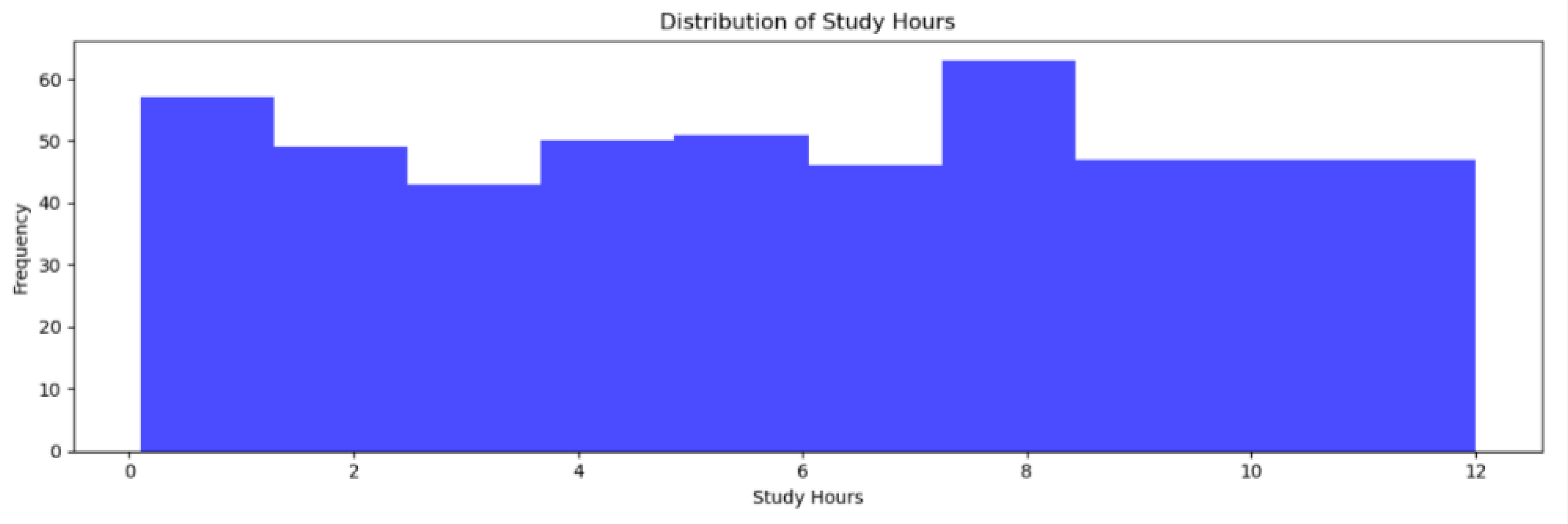


Subplots

Subplots allow multiple plots to be displayed within a single figure, making it easy to compare various data attributes side-by-side. They are particularly useful for identifying relationships and differences across datasets. For example, subplots showing the distributions of sleep and study hours enable a simultaneous view of how these variables vary. This approach simplifies analysis by providing a cohesive and organized presentation of multiple visualizations in one layout.



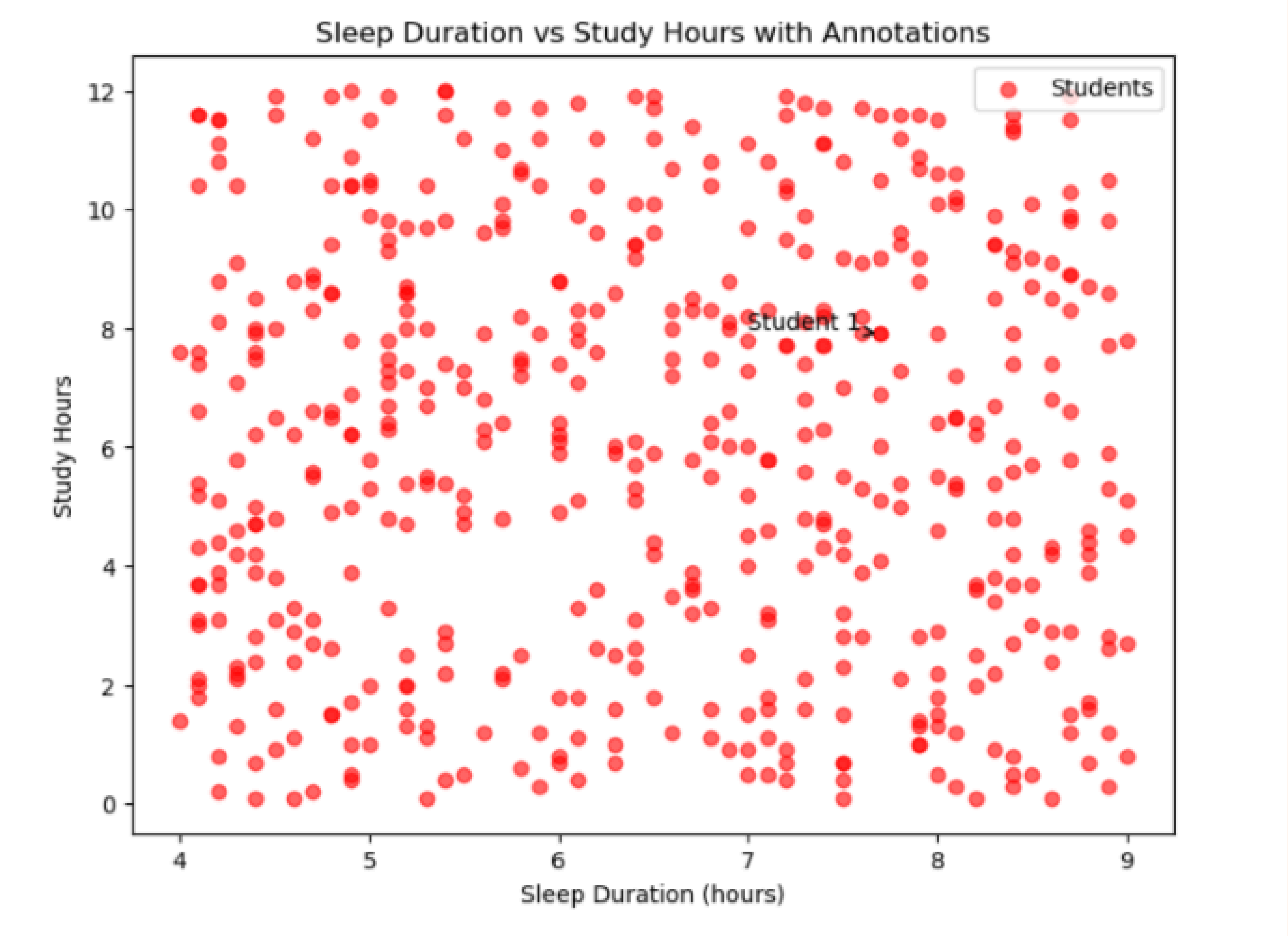
Distribution of Sleep Duration



Distribution of Study Hours

Annotations

Legends and annotations enhance data visualizations by providing context and emphasizing important details. Legends identify different elements in the plot, such as lines or categories, while annotations highlight specific data points or trends with text or arrows. For example, an annotated scatter plot can pinpoint key data points, offering insights into anomalies or significant trends. These tools make visualizations more informative and easier to interpret for the audience.



Thank You

for more details : thehoarseandhumble@gmail.com

Scan for code snippets