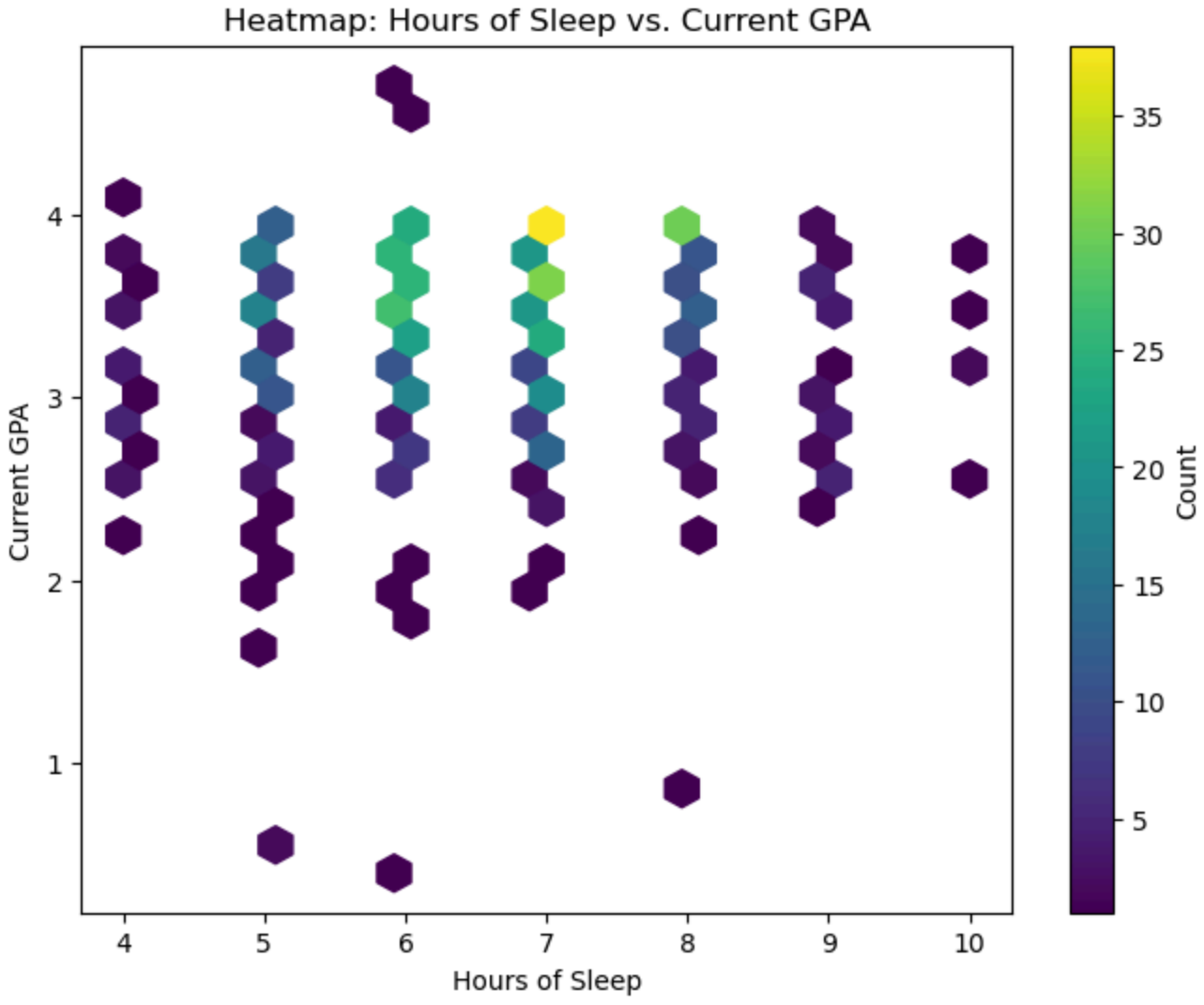
Introduction

https://github.com/uic-cs418/group-project-data-minds/blob/main/ProgressReport.ipynb

Our project, StudySync, aims to identify and analyze behavioral and lifestyle factors that significantly impact students' academic performance. We are using real-world datasets that include features like sleep patterns, study habits, attendance, drug use, and GPA. The goal is to visualize correlations (like attendance vs. GPA) and apply binary classification models (like predicting poor academic performance based on drug use) to help students understand what behaviors positively or negatively influence their grades.

Relationship between Sleep and GPA

In [1]:
```
from main import create_sleep_dataframe, create_sleep_visualization
df_sleep = create_sleep_dataframe()
create_sleep_visualization(df_sleep)
```



1. Hypothesis Testing:

The hexbin heatmap of **Hours of Sleep vs. Current GPA** tests the hypothesis that there is a relationship between the amount of sleep a student gets and their academic performance. This hypothesis is interesting because:

- Sleep is a modifiable behavior, so if an optimal sleep range correlates with higher GPA, students can adjust their habits.
- The heatmap reveals clusters of data, showing where most students fall and highlighting a potential "sweet spot" for sleep.
- It helps determine whether the relationship is linear or if there is a plateau beyond a certain sleep duration.

2. Data Cleaning:

For the **Sleep** sheet:

- The original **"Hours of sleep"** column contained mixed text entries (e.g., "10 or more hours").
- A regular expression was used to extract the first numeric value from each entry, converting "10 or more hours" to **10**.
- The extracted values were then converted to a numeric type. Non-convertible values become NaN (which are ignored by the plotting functions).
- The dataset is at the individual student level, allowing us to analyze each student's sleep habits relative to their GPA.

3. Exploratory Data Analysis:

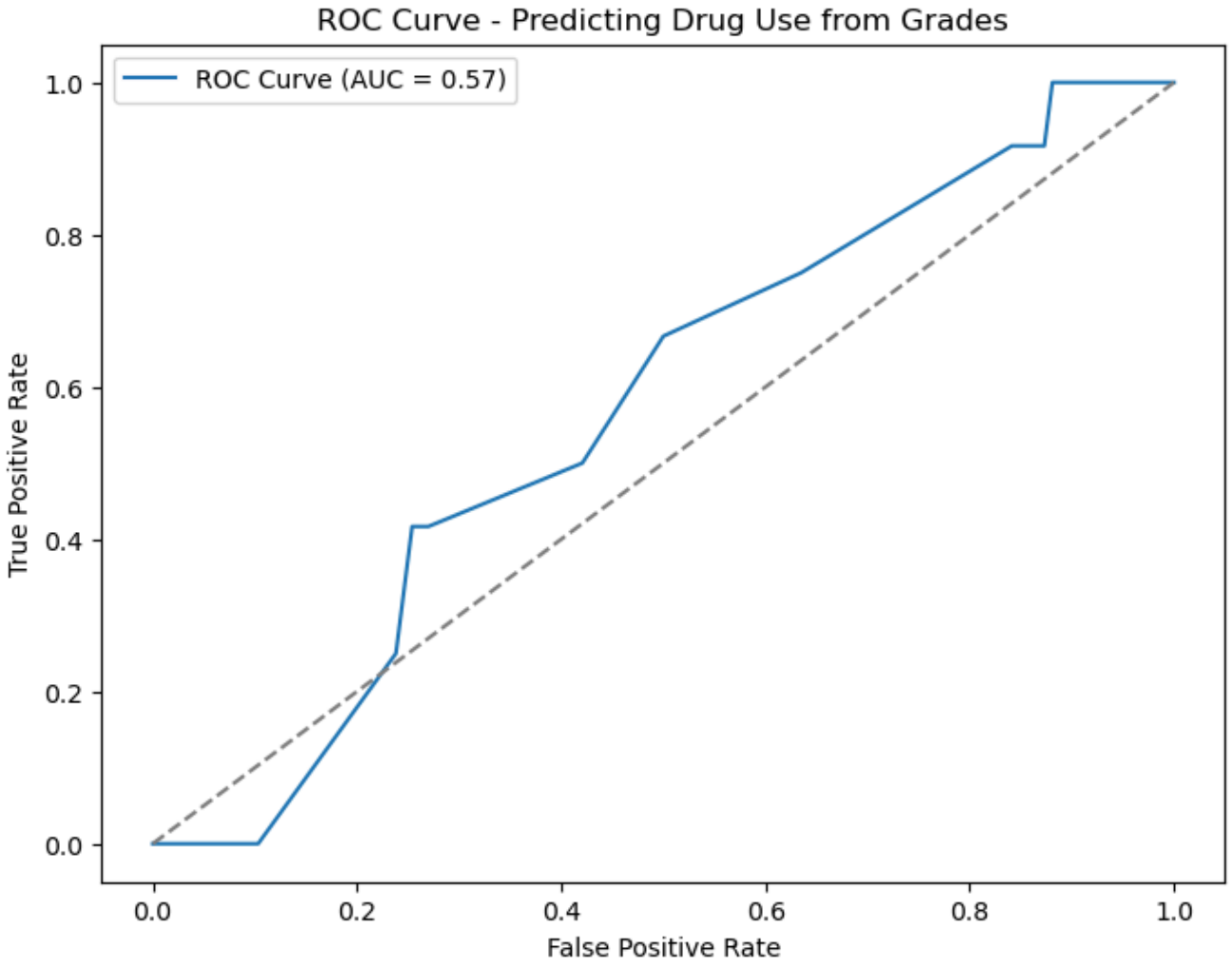
- **Data Distribution:** The heatmap shows that many students cluster around specific sleep durations, with a notable density between 6–8 hours.
- **Preliminary Insight:** Higher GPA values appear to be associated with certain sleep ranges, suggesting a potential optimal range of sleep for academic success.
- **Variability:** There are areas with sparse data, which might indicate outliers or less common sleep patterns.

Predicting Drug Use from Grades

In [1]:
```
from main import (
    load_grade_data,
    prepare_grade_data,
    plot_roc_curve,
    plot_grade_distributions
)
```

```
df = load_grade_data()
prepared_df = prepare_grade_data(df)
plot_roc_curve(prepared_df)
```

```
Correlation between Drug Use and Average Grade:
                ADrugs_binary  AvgGrade
ADrugs_binary        1.000000 -0.066803
AvgGrade            -0.066803  1.000000
```



1. Data Cleaning:

After performing data cleaning, we conducted a logistic regression analysis to evaluate the predictive relationship between students' academic performance and reported drug use. The F1Grade, F2Grade, and F3Grade values were first mapped from letter grades to numeric scores and averaged to create a new `AvgGrade` feature. The `ADrugs` column was cleaned and standardized into a binary format (yes → 1, no → 0).

To assess how well academic performance predicts drug use, we trained a logistic regression model and plotted the Receiver Operating Characteristic (ROC) curve. The ROC curve illustrates the model's ability to distinguish between students who reported drug use and those who did not, across different classification thresholds.

2. Exploratory Data Analysis:

- **ROC Curve: Predicting Drug Use From Grades**

The Area Under the Curve (AUC) value was calculated to quantify the model's performance. A higher AUC (closer to 1) indicates a better-performing model. The ROC curve helps visualize the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate, offering insights into how effectively grades alone can act as predictors for potential drug use among students.

Reflection

What is the hardest part of the project that you've encountered so far?

- Trying to find good datasets to use for our project. It was hard finding datasets that would help find correlations and solve the problem that our project is trying to help with.
- Handling inconsistencies different scales for attendance, missing or malformed GPA entries, “Yes/No” vs. “yes/no” drug use.
- Deciding which rows to drop vs. how to impute missing data without biasing the results.

What are your initial insights?

- We observed a strong correlation between student attendance and GPA, supporting our hypothesis that students who attend more classes tend to perform better academically. Additionally, early visualizations and exploratory data analysis suggest that behavioral factors such as study time, sleep habits, and drug use may also play a role in academic outcomes. These findings indicate that lifestyle factors are measurable predictors of GPA.

Are there any concrete results you can show at this point? If not, why not?

- We created a bar chart showing the positive relationship between attendance and GPA, trained a logistic regression model to classify students based on GPA and drug use, and performed EDA to clean, prepare datasets, and visualize distributions of key features like study time and health behaviors. These steps gave us good starting evidence for our hypothesis, and the machine learning model shows some promise in predicting low GPA.

Going Forward, what are the current biggest problems you're facing?

- Experimenting with our machine learning models to improve accuracy and enhance generalizability.

Do you think you are on track with your project? If not, what parts do you need to dedicate more time to?

- Yes, we believe we are on track with our project, though we plan to dedicate more time to refining our machine learning models, interpreting the results in more detail, and finalizing our visualizations.

Given your initial exploration of the data, is it worth proceeding with your project, why? If not, how are you going to change your project and why do you think it's better than your current results?

- Yes, based on our initial exploration, it is worth proceeding with the project because we've already identified meaningful patterns—such as a clear relationship between attendance and GPA—and the data is rich enough to support further analysis and modeling.

Any changes: A discussion of whether your scope has changed since the check-in proposal slides. What did you aim to do that you will not do, and what have you added to the project?

- Since our proposal, we've narrowed our focus to the FGCU, Harvard, and Mendeley datasets. These were the most usable in terms of structure and relevance to our research questions. Although we originally planned to use four datasets, including Dartmouth, we've prioritized the ones that allowed us to make quicker progress. We're still open to using the Dartmouth dataset later if time permits. We also decided to concentrate on three main behavioral factors: sleep, drug use, and attendance.

Next steps

What you plan to accomplish in the next month and how you plan to evaluate whether your project achieved the goals you set for it.

- In the next few weeks, we'll refine our current visuals, explore regression models for GPA prediction, and decide whether to include the Dartmouth dataset. We'll also begin writing the final report, ensuring that our findings are clearly explained and supported by the data.