**Data Wrangling Process: Detailed Documentation**

**1. Understanding the Dataset**

The dataset contains student performance information, focusing on demographics, academic metrics, and extracurricular participation. Column Include:

* **Demographics**: Age, Gender, Ethnicity, Parental Education, StudentID.
* **Academic Performance**: GPA, Grade Class, Study Time, and Absences.
* **Extracurriculars**: Participation in Sports, Music, Tutoring, and Volunteering.

**2. Data Cleaning Process**

**Handling Missing Values**

Missing values were identified and addressed to ensure a complete and accurate dataset:

* Categorical columns were filled with the most frequent value.
* Numerical columns that were not sensitive to outliers were filled with the mean, while data that were sensitive were filled using the median to maintain consistency and avoid skewing the results.

**Removing Duplicates**

The data did not include any duplicated data however if there were then they would be identified and removed, reducing redundancy and ensuring the reliability of the dataset.

**3. Data Transformation**

**Binning Continuous Variables**

To simplify analysis, continuous variables such as weekly study time and absences were grouped into categories:

* **Study Time** was categorized into Very Low, Low, Moderate, High, and Very High.
* **Absences** were grouped into Very Few, Few, Moderate, Many, and Very Many.

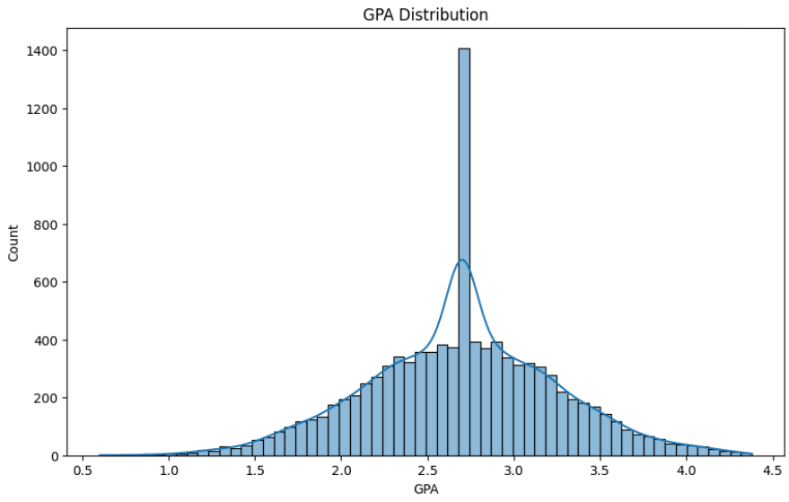
This transformation helped show patterns and trends across the dataset.

**Encoding Categorical Variables**

Categorical columns such as Ethnicity, Parental Support, Parent Education, Grade Class, and Gender were mapped to descriptive labels. For example:

* Ethnicity codes were replaced with labels like Caucasian and African American.
* Parental Support levels were converted to terms like Low and Very High, making the data more interpretable.

**4. Exploratory Data Analysis (EDA)**

**GPA Distribution**

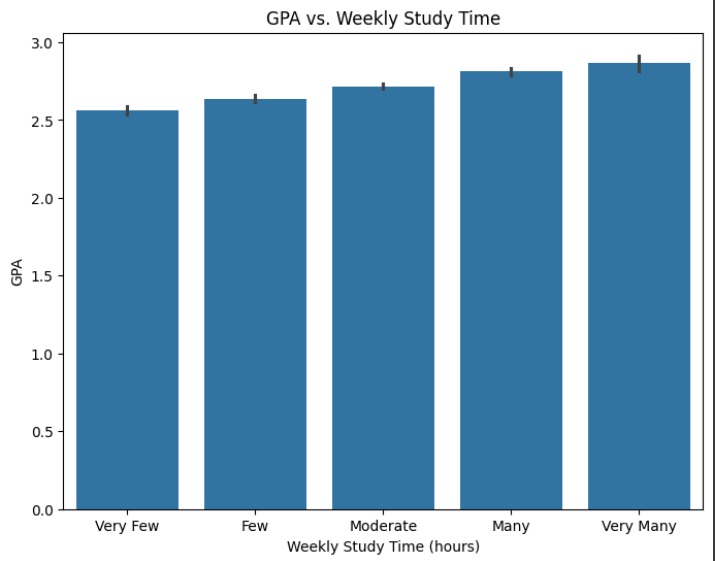
The analysis of GPA revealed a right-skewed distribution, with the majority of students achieving GPAs between **2.5 and 3.0**. This clustering suggests that most students are performing at an average academic level, with fewer excelling above 3.5 or struggling below 2.0.

**Key Insights:**

1. **Academic Clustering**:
   * Most students fall into the middle-performing range, highlighting an opportunity for targeted academic interventions.
2. **Room for Improvement**:
   * The relatively low number of students achieving GPAs above 3.5 suggests a need for enrichment programs to support high-achieving students.

**GPA vs. Study Time Weekly**

An analysis of GPA against study time showed that students who studied more tended to achieve higher GPAs. This positive correlation underscores the importance of consistent study habits for academic success.



**Absences vs. GPA**

Fewer absences strongly correlated with higher GPAs. Students with very few absences consistently outperformed those with moderate to high absences, emphasizing the importance of regular attendance in academic achievement.

A graph showing a blue line

Description automatically generated

**Extracurricular Participation**

1. **Sports and Music**:
   * Students participating in sports or music showed slightly better GPAs compared to non-participants.
2. **Volunteering and Tutoring**:
   * While tutoring had a significant positive impact on GPA, volunteering showed a smaller, yet noticeable, benefit.