**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.

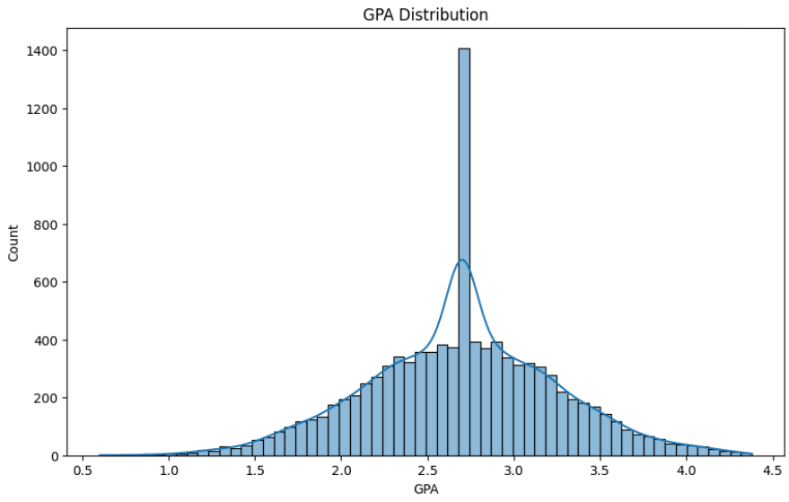
This transformation helped show patterns and trends across the dataset.

**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**

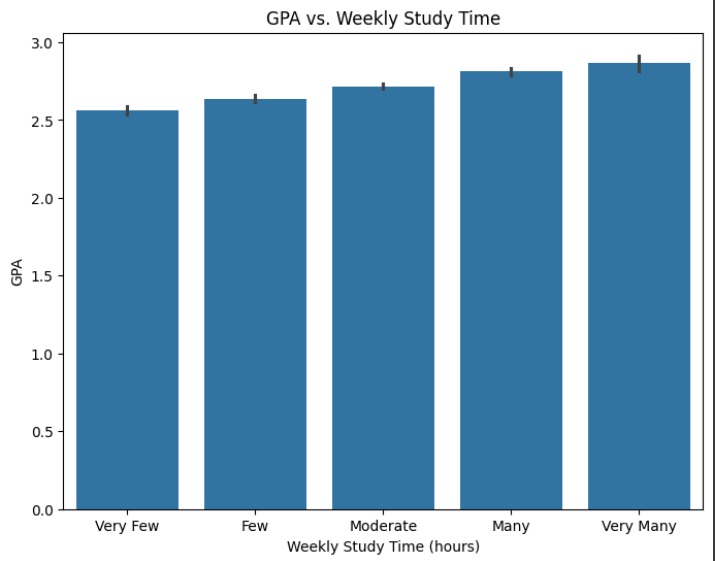
The analysis of GPA revealed a normalised distribution, with the majority of students achieving GPAs between **2.5 and 3.0**. This clustering suggests that most students are performing at an average academic level, with fewer excelling above 3.5 or struggling below 2.0.

**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.

**5. SQL Table Creation Process**

To implement the star schema based on the ERD diagram:

1. **Fact and Dimension Tables:**
   * The star schema revolves around a central fact table containing measures and keys linking to surrounding dimension tables.
   * In this dataset:
     + The **fact table** is StudentDetails, as it contains measurable metrics like Age, Gender, Ethnicity, Parental Education.
     + The **dimension tables** include:

* StudentPerformance: Contains academic performance data such as GPA, Study Time Weekly, and Grade Class.
* StudentEngagement: Tracks engagement metrics like Attendance and Extracurricular Participation.
* Gender, Ethnicity, ParentalEducation, ParentalSupport, and GradeClass for categorical attributes.

1. **SQL Table Creation:**
   * Each table was created based on the schema and relationships defined in the ERD diagram
2. **Relationships and Data Integration:**

* **Fact Table Relationships:**
  + StudentDetails acts as the central table, storing demographic data (Age, Gender, Ethnicity, and ParentalEducation) and linking to:
    - StudentPerformance for academic metrics like GPA and StudyTimeWeekly.
    - StudentEngagement for engagement-related metrics.
    - Gender, Ethnicity, ParentalEducation, ParentalSupport, and GradeClass for categorical dimensions.
* **Dimension Table Relationships:**
  + Each dimension table provides descriptive data for querying and categorization:
    - Gender, Ethnicity, and ParentalEducation define demographic attributes.
    - GradeClass categorizes academic performance.

1. **Star Schema Design Benefits:**

* **Simplicity:** The star schema design clearly organizes relationships, with StudentDetails at the center and dimension tables surrounding it.
* **Efficiency:** Indexed foreign keys enable fast queries, such as joining StudentDetails with StudentPerformance or StudentEngagement.
* **Scalability:** Dimensions can be updated or expanded independently of the fact table, reducing maintenance complexity.

1. **ERD Alignment:**

* The ERD diagram visually aligns with the adjusted schema:
  + Dimensions like Gender, Ethnicity, and ParentalEducation are directly linked to StudentDetails.
  + Additional tables like StudentPerformance and StudentEngagement provide further details about the students, emphasizing StudentDetails as the core fact table.

Education Performance Analysis

**Counts:**

* There are 2392 total students
* The grade “F” has the greatest number of students with 1211 students meanwhile the grade “A” has the least.
* The Caucasian ethnicity has the greatest number of students with 1207 students, and the average being around 50.46% of all students, which could be due to the fact of the school being located in Europe.

**GPA Analysis**

* **Parental Education:**
  + Students whose parents completed high school or college had the highest average GPAs of 1.95 and 1.30, respectively.
  + Interestingly, students whose parents held bachelor’s degrees or higher had slightly lower average GPAs of 1.81 and 1.82, suggesting a potential inverse correlation between higher parental education and student GPA.
* **Parental Support:**
  + Very high parental support was associated with the highest average GPA of 2.19.
  + Students with no parental support had the lowest average GPA of 1.54.
* **Gender:**
  + Male students slightly outperformed females, with an average GPA of 1.91 compared to 1.89 for females.
* **Absences:**
  + **Attendance strongly influenced performance:**
    - Students with very few absences had the highest average GPA of 3.16.
    - Students with very many absences had the lowest average GPA of 0.93.
  + **Key Insight:** Attendance is a critical factor in student success.

**Extracurricular Activities**

* **Tutoring:**
  + Students who received tutoring achieved a higher average GPA of 2.10, compared to 1.81 for those who didn’t.
* **Participation in Activities:**
  + Students involved in extracurricular activities had an average GPA of 2.01, outperforming those who didn’t (1.83).
* **Sports:**
  + Students who played sports had a slightly higher average GPA of 1.89, compared to 1.87 for those who didn’t.
* **Music:**
  + Students who played musical instruments achieved an average GPA of 2.04, outperforming those who didn’t (1.87).
* **Volunteering:**
  + The difference here was minimal: students who volunteered had an average GPA of 1.91, compared to 1.90 for those who didn’t.

**Key Takeaways:**

1. Absences have the most significant impact on GPA, with very few absences linked to much higher performance.
2. Parental support plays a crucial role, with very high support resulting in better GPAs.
3. Extracurricular activities, especially tutoring and music, show a positive impact on academic performance.