Student Performance

Exploratory Data Analysis (EDA)

Name: Mohammad Shatnawi

**Problem Statement:**

Academic success is a cornerstone of future opportunities, personal development, and societal progress. With Grade Point Average (GPA) serving as a global benchmark for assessing student performance, understanding the factors that influence it is crucial. Despite the global importance of education, many students face challenges in achieving high GPAs due to various circumstances such as socioeconomic background, parental support, individual study habits, and absenteeism. This study aims to explore and uncover the key factors that determine academic performance and provide a deeper understanding of how specific factors—such as parental education levels, extracurricular activities, weekly study time, tutoring, and support systems—influence student outcomes. By investigating these variables, we seek to identify actionable insights that can help improve academic performance across diverse demographics. Extracurricular activities can have a significant positive impact on academic success by enhancing essential skills such as time management, teamwork, and leadership. Participation in these activities often leads to increased engagement and motivation, which can translate into better academic performance. Furthermore, students who participate in extracurricular activities tend to develop a stronger sense of belonging and self-discipline, both of which contribute to their overall educational achievement.

**Business Impact:**

Understanding the factors influencing GPA is critical for improving educational outcomes and workforce readiness. Schools can design targeted programs, EdTech companies can develop tools to support learning, and policymakers can create effective reforms. This analysis empowers students, parents, and institutions to drive academic success, reduce dropout rates, and foster economic growth.

**General Dataset Information:**

**File Name:** Student\_performance.csv

**Description:** This dataset includes information on high school students, covering demographics, study habits, parental involvement, extracurricular activities, and academic performance, with Grade Class as the target variable.

**Dataset Details:** 2,393 rows & 15 columns

**Size:** The size of the dataset is 238 KB.

**Source:** [Kaggle - Students Performance Dataset](https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset/data)

**Target Features:**The problem statement outlines the essential metrics for determining the factors that influence students' academic performance. The key features in question that can be found within the dataset are:

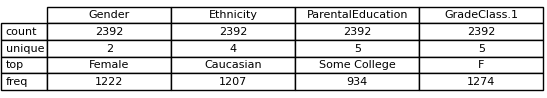
* **Gender**: Identifies the gender of the students.
* **Ethnicity**: Represents the students' racial and ethnic backgrounds.
* **Parental Level of Education**: Indicates the highest level of education attained by the parents.
* **Study Time Weekly**: Represents the time students dedicate to studying each week.
* **Parental Support**: Reflects the level of parental involvement in the student's academic activities.
* **Extracurricular Activities**: Includes involvement in activities such as sports, music, and volunteering.
* **Absences**: Represents the number of school days the student was absent, which may significantly impact performance.
* **GPA (Target Variable)**: Represents the students' overall grade point average, used as the primary measure of academic performance.
* **Tutoring**: Indicates whether the student receives additional tutoring outside regular school hours.

These metrics will be analyzed collectively to identify patterns and determine the factors contributing to higher GPA scores.

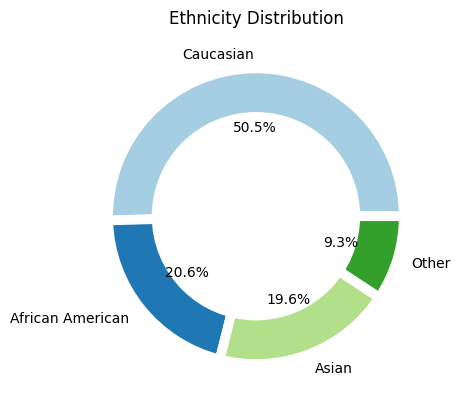
* **List Of Analysis:**
* **Analysis #1 – Statistics For Numerical Values**
* Students study an average of ~10 hours/week, but some study up to 20 hours.
* Absences vary widely, with some students missing up to 20 times/week.
* Parental support and GPA are moderate on average, but there’s a wide range.
* Most students participate in 1 extracurricular activity.

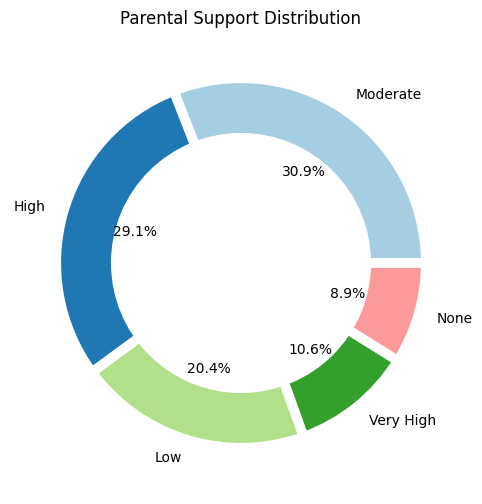
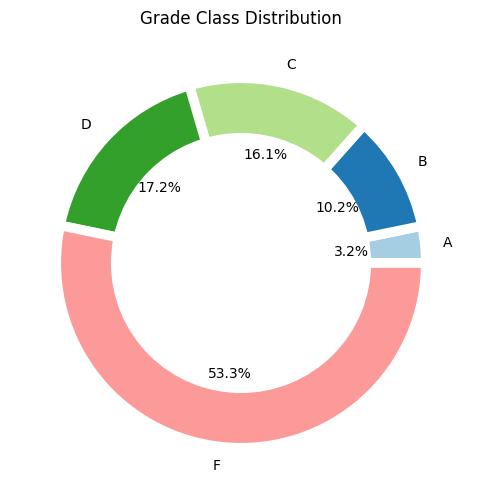
**A screenshot of a table

Description automatically generated**

* **Analysis #2 – Categorical Data Summary**
* The majority of students are Female and Caucasian.
* Most parents have Some College education.
* The most common grade class is F, indicating a potential area for improvement.
* **Analysis #3 - Number of students**

It was confirmed that all students had a unique id and the number of students in the group was studied, distributed according to gender, ethnicity, age, parental education level, the level of parental support and the grade class, to take a general view of the group used and to quote some insights. It was found that the number of female students was slightly larger than that of males, at a rate of 51 % for females and 49 % for males, and that students who came from Caucasian ethnicity were more numerous than others, at a rate of 50 % of the total number of students, and that most students received average support from their parents, as most parents had certificates for some grades and that the vast majority of students obtained an F grade, at a rate of 53 %.



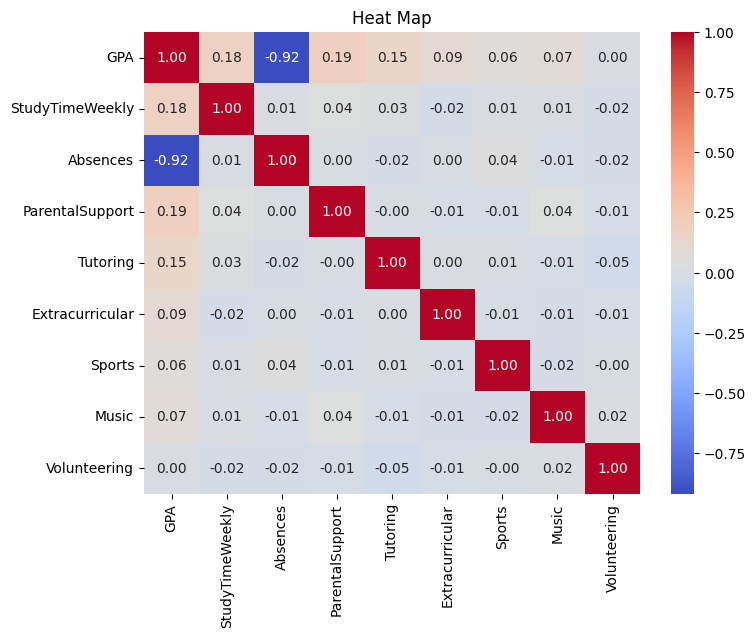


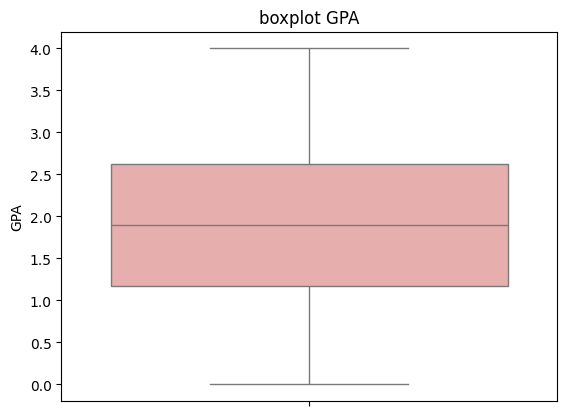
Number of students

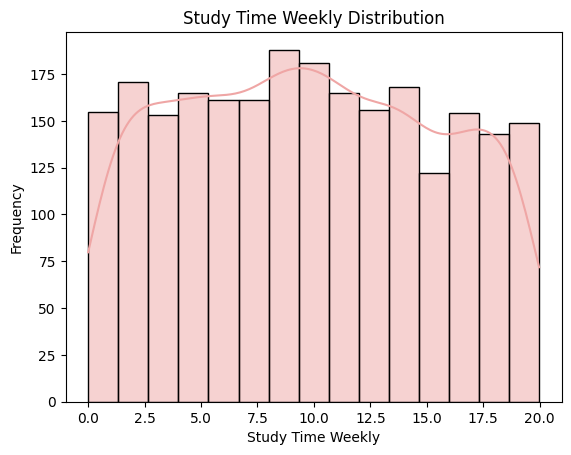
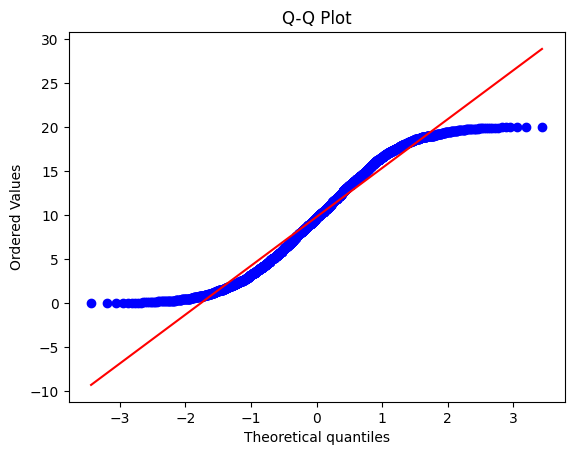
2393

* **Analysis #4 – Correlation and P\_Values**

A chart with different colored squares

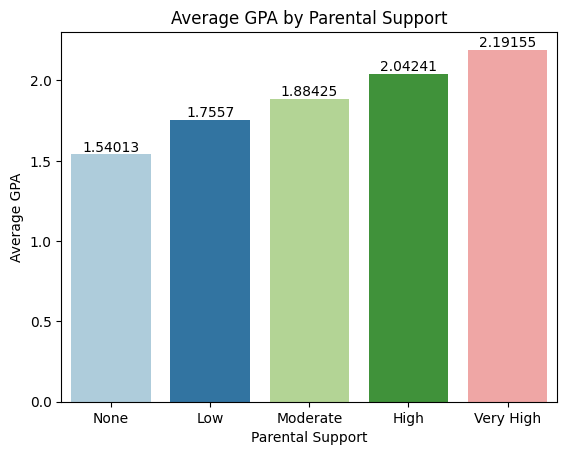
Description automatically generatedIn this part of the analysis, I decided to take a comprehensive look at the relationships in the data and the P values ​​to see between which columns the relationships were strong. It turned out that the strongest relationship was the relationship between absences and the average, as it was a negative relationship, and this is very useful. By referring to the P values, I made a heatmap that shows all the values.

* **Analysis #5 – Average and Distribution for (GPA)**
* The average GPA, initially calculated without excluding zero values, was found to be 1.906. After removing the zero values, the average increased slightly to 1.919. This minimal difference does not significantly impact the overall analysis.
* A graph of a distribution

  Description automatically generatedAs for the distribution, the GPA follows a normal distribution, which is a highly desirable characteristic in data analysis. A normal distribution indicates that the data is symmetrically distributed around the mean, allowing for reliable statistical inferences and facilitating the application of various parametric tests. This enhances the credibility and robustness of any conclusion drawn from the analysis.
* **Analysis #6 – Average and Distribution for (Study Time Weekly)**
* The **average study time per week** is **9.77 hours**. This is the central value of your data, meaning it’s the typical amount of time students spend studying weekly.
* We tested your data to see if it follows common patterns like **normal**, **exponential**, or **uniform** distribution. None of these fit your data.
* The data is **not normal**, **not exponential**, and **not uniform**.
* The data is **very slightly right-skewed** (skewness = 0.05). This means it has a tiny tail on the right side
* **Analysis #7– Average (GPA) By Gender,** **Ethnicity and Parental Education**

Based on the correlation heatmap analysis, it appears that there is no significant relationship between GPA and the columns in question. This suggests that these columns may not be useful for the analysis, as their average values are distributed very similarly and closely to one another. As a result, these variables do not provide meaningful insights or contribute to understanding GPA in this context.

* **Analysis #8 – Average (GPA) By Parental Support**

The analysis of average GPA across different levels of parental support reveals a clear positive trend: **as parental support increases, so does the average GPA**. This indicates that parental involvement plays a significant role in academic performance.

**Numerical Breakdown:**

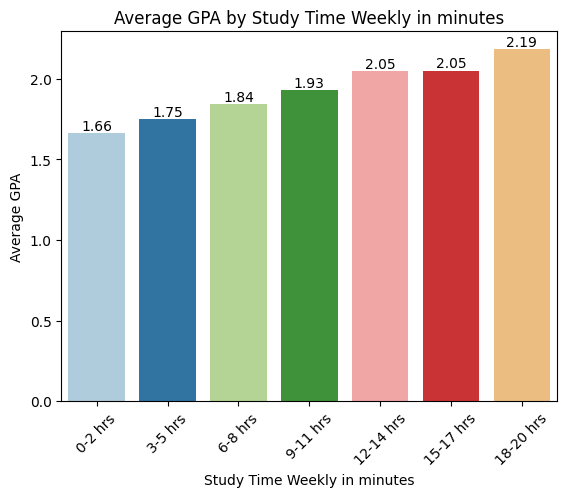
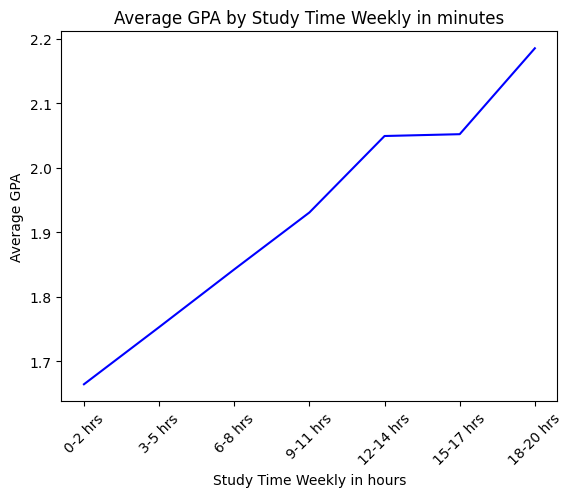
**None**: 1.54, **Low**: 1.76, **Moderate**: 1.88, **High**: 2.04, **Very High**: Highest Average GPA (2.19)

**Key Insight:**

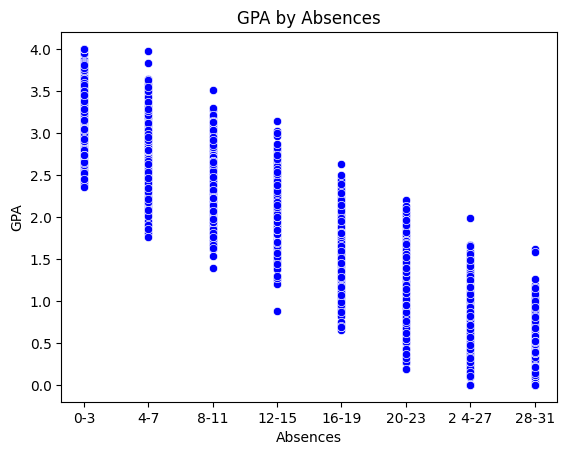
Students with **"Very High"** parental support achieve the highest average GPA, while those with **"None"** have the lowest. This numerical truth highlights the importance of parental support in fostering academic success. Interventions to enhance parental involvement could potentially improve student outcomes.

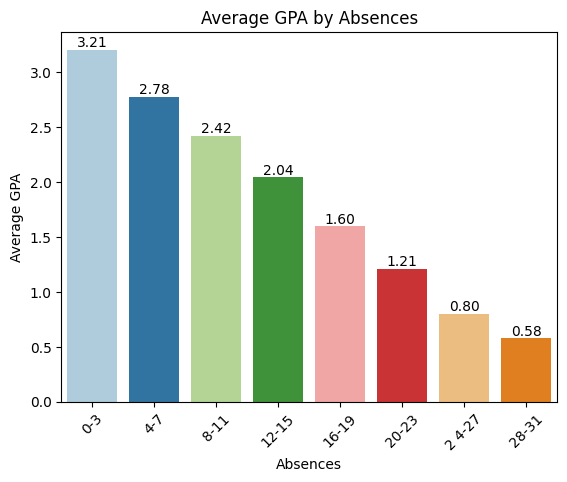
* **Analysis #9– Average (GPA) By Study Time Weekly**

The analysis shows a clear positive correlation between weekly study time and GPA. Students who study more hours tend to achieve higher GPAs. For instance, those studying 0–2 hours have the lowest GPA, while those studying 18–20 hours achieve the highest average GPA. The data suggests that consistent study time, especially above 10 hours per week, significantly contributes to academic success. However, the rate of GPA improvement starts to plateau as study time exceeds 18 hours. Visualizing this trend could help identify the optimal study time for maximum GPA improvement.



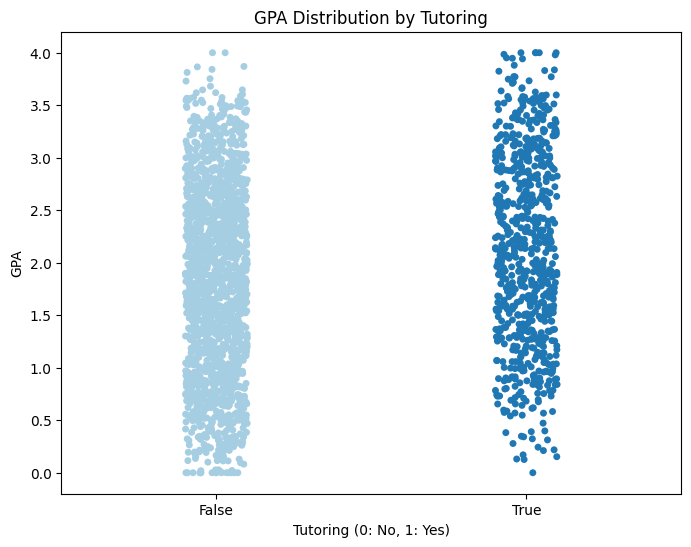
* **Analysis #10 – Average (GPA) By Absences**
* In the analysis of absences, there is a clear trend between the number of absences and the average GPA of students. As the number of absences increases, the average GPA tends to decrease. Here is a breakdown of the observed trend:
* High GPA with few absences: Students with fewer absences tend to have higher GPAs. For example, the highest GPA of 3.21 is associated with the lowest number of absences.
* Decline GPA with increased absences: As the number of absences increases, the average GPA decreases dramatically. For example, the average GPA decreases from 3.21 to 2.42, 2.04 and continues to decrease as absences increase.
* Low GPA with high absences: Students with a high number of absences have much lower GPAs, with the lowest GPA of 0.58, which corresponds to the highest number of absences.
* This analysis indicates a strong negative association between the number of absences and academic performance, as measured by GPA. It highlights the importance of regular attendance to maintain high academic achievement





* **Analysis #11 – Average And Distribution (GPA) By Tutoring**
* The average GPA for students who attended tutoring is 2.11.
* The average GPA for students who did not attend tutoring is 1.82.
* This suggests that tutoring may have a positive impact on students' GPAs, as the average GPA is higher for the tutoring group.

A graph of a bar chart

Description automatically generated with medium confidence

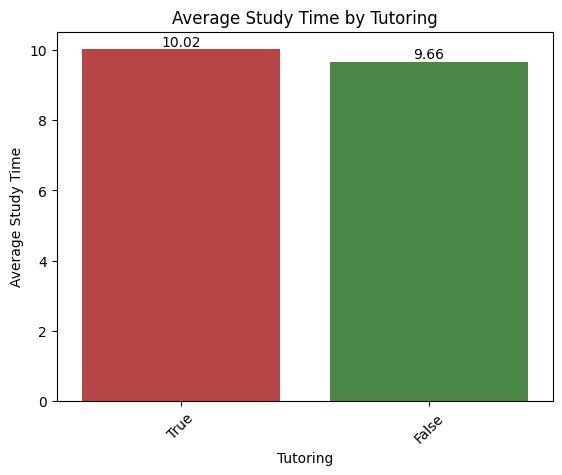
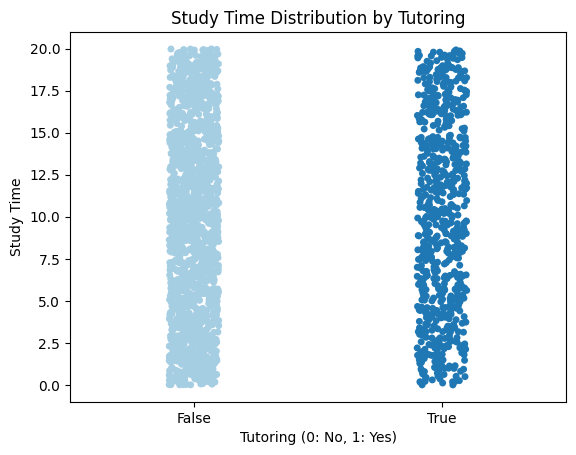
* **Analysis #12 – Average (Study Time Weekly) By Tutoring**

**Tutored students study slightly more**

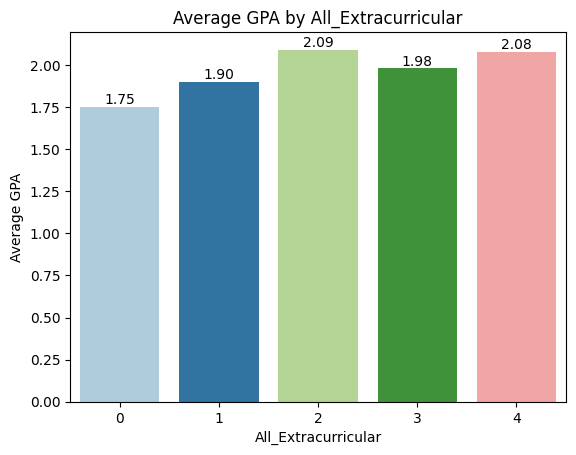
* **10.02 hours per week** for tutoring students.
* **9.66 hours per week** for students without tutoring.
* The difference is **0.36 hours (≈ 21.6 minutes per week)**.

**Very Small Difference**

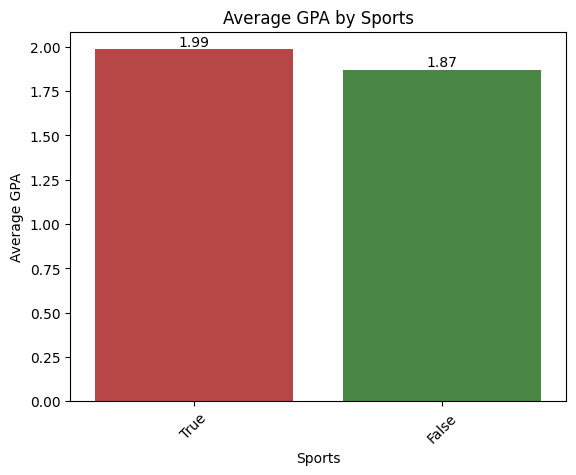
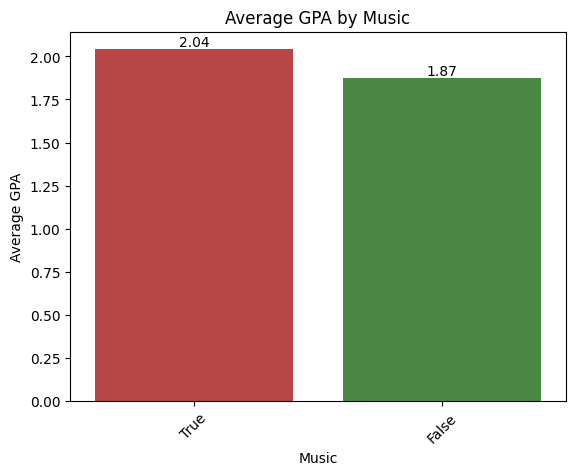
* A **21.6-minute difference per week** translates to **just over 3 minutes per day** (not a huge gap).
* It’s unclear whether tutoring has a **meaningful impact** on study time.

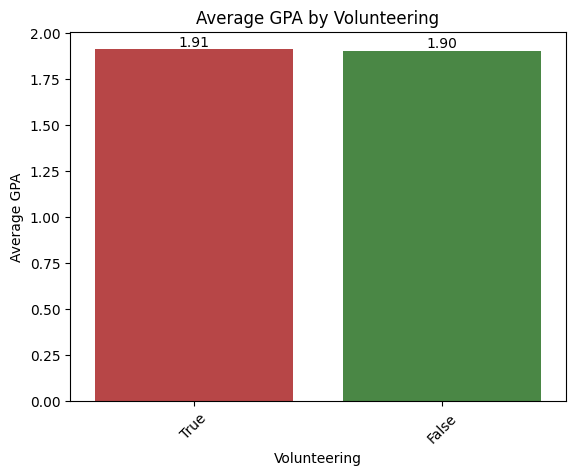


* **Analysis #13 – Average (GPA) By All Extracurricular**
* The average GPA for students with 0 extracurricular activities is 1.98.
* The average GPA for students with 1 extracurricular activity is 2.08.
* The average GPA for students with 2 extracurricular activities is 2.10.
* This suggests a slight positive trend, indicating that students who participate in extracurricular activities may have slightly higher GPAs on average.



* **Analysis #14 – Average (GPA) By Sports, Music, and Volunteering**
* **Music:** Students involved in music have a higher average GPA (2.04) than those who are not (1.87), suggesting a positive impact on academic performance.
* **Sports:** Participants in sports show a slightly higher GPA (1.99) compared to non-participants (1.87), indicating a mild positive effect.
* **Volunteering:** The GPA difference between volunteers (1.91) and non-volunteers (1.90) is minimal, implying little academic impact.





* Conclusion
* The exploratory data analysis reveals several key factors influencing academic performance, measured by GPA. Parental support and weekly study time had the most significant impacts, with GPA rising from **1.54** for students receiving no parental support to **2.19** for those with very high parental support. Similarly, students studying more than 10 hours per week achieved higher GPAs, with the highest average GPA of **3.21** observed among those studying 18–20 hours weekly.
* On the other hand, absenteeism showed a strong negative correlation with GPA. While students with the fewest absences achieved the highest GPA of **3.21**, this dropped sharply to **0.58** for those with the highest absence rates. Participation in extracurricular activities demonstrated a slight positive impact, as students involved in two activities had an average GPA of **2.10**, compared to **1.98** for those who did not participate in any.
* Tutoring also proved beneficial, with students who attended tutoring achieving an average GPA of **2.11**, compared to **1.82** for those who did not. Demographic factors, such as gender, ethnicity, and parental education, showed no statistically significant relationship with GPA, indicating that behavioral and environmental factors play a more crucial role.
* These findings highlight the importance of fostering a supportive learning environment, encouraging regular attendance, promoting consistent study habits, and providing opportunities for extracurricular involvement and tutoring. Such interventions can serve as key strategies to enhance educational outcomes and create equal opportunities for all students.
* Music shows the strongest correlation with a higher GPA, followed by sports, while volunteering has no impact.

Refer to the following notebook to check further analysis and Python code : [student performance. ipynb](https://colab.research.google.com/drive/1g6O_LUZPlXKQhWGaPlDqfmRbt-mTiVNP#scrollTo=fAjkEtrVrgHR)