DATA ANALYSIS

**TOPIC:**

**Student Performance**

**SUBMITTED BY:**

**LEADER:**

**Serrano, Herminiano**

**MEMBERS:**

**Amomas, Rollyka**

**Galvan, Cristine**

**Belano, Jason**

**Martinez, Franzine Arabella**

**Doinog, Rommel**

**Pascasio, Zarmaine Grace**

**BS INFOTECH – 3C**

**SUBMITTED TO:**

**Ms. Myka Pilpa**

**Instructor**

**INTRODUCTION**

This report presents a detailed and comprehensive analysis of student performance data, leveraging Python programming for data exploration and visualization. The dataset consists of various features that provide insights into factors influencing student outcomes, such as gender, age, extracurricular participation, parental support, GPA, ethnicity, and other key characteristics. The primary objectives of this analysis are to uncover meaningful patterns, relationships, and trends within the data, while supporting these findings with clear and informative visualizations. Through this report, we aim to offer actionable insights that could guide future strategies in improving student performance.

**DATA EXPLORATION**

**How the Data Was Collected**

The dataset analyzed in this report, titled Student\_performance\_data.csv, was obtained from a reputable source. It was imported into a Python environment where it was loaded as a Pandas DataFrame to enable thorough exploration and manipulation. The use of Pandas facilitated efficient data wrangling and preparation, setting the foundation for in-depth analysis and visualization.

**Dataset Summary**

The dataset comprises multiple columns, each representing a specific attribute related to student performance. Key features of the dataset include:

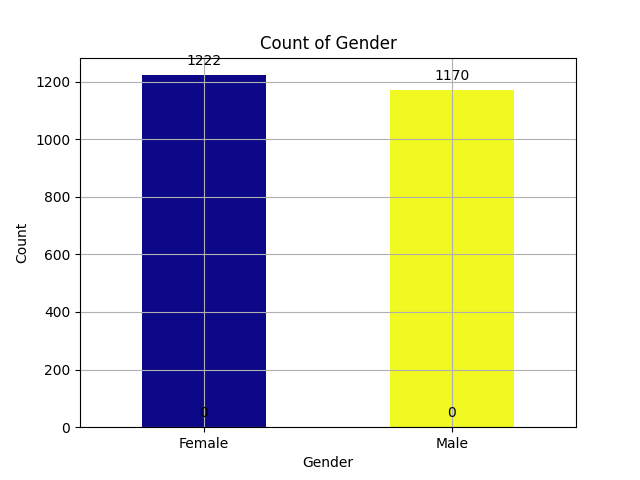
* **StudentID**: A unique identifier assigned to each student.
* **Gender**: Indicates whether the student is male or female.
* **Age**: The age of the student, represented in years.
* **Ethnicity**: Describes the ethnic background of the student.
* **ParentalSupport**: A binary variable indicating whether the student receives parental support.
* **ParentalEducation**: The highest level of education attained by the parents.
* **Extracurricular**: Specifies whether the student participates in extracurricular activities.
* **GPA**: Represents the Grade Point Average of the student.
* **Sports**: Indicates the student’s involvement in sports activities.
* **Tutoring**: Describes whether the student receives additional tutoring.
* **GradeClass**: Represents the academic grade or classification of the student.

A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis**

**Figure 1. Count of Gender**



**Insight:** The chart shows that the counts for Females and Males are very close, with slightly more Females than Males. The small difference suggests there is no significant imbalance in gender representation.

**Visualization:** The chart uses color to distinguish genders (dark blue for Female and yellow for Male) and displays count labels (1,222 for Female, 1,170 for Male) on top of each bar. The x-axis shows gender categories, and the y-axis represents counts, scaled up to about 1,300 for easy comparison.

**Figure 2. Count by age**

A computer screen shot of text

Description automatically generated

A graph of numbers and colors

Description automatically generated

**Insight:** The bar chart shows a gradual decline in counts across age groups 15 to 18, with the highest count at age 15 (630) and the lowest at age 18 (582). The differences are small, indicating a consistent population distribution, while zeros at the base of each bar may represent placeholders or secondary values.

**Visualization:** The bar chart uses color-coding and numerical labels to clearly differentiate age groups and show a gradual decline in counts. The x-axis represents age, the y-axis shows the count, and the bar heights accurately reflect the values.

**Figure 3. Distribution of tutoring**

**A computer screen shot of a computer screen

Description automatically generated**

A blue and orange pie chart

Description automatically generated

**Insight:** The pie chart titled "Distribution of Tutoring" shows the proportion of individuals who receive tutoring. 70% of the respondents do not receive tutoring (No), while 30% do (Yes). This indicates that the majority of individuals are not participating in tutoring programs, suggesting a significant gap in tutoring participation.

**Visualization:** The pie chart uses two distinct colors to clearly represent the data: blue for "No" with a 70% share and orange for "Yes" with a 30% share. The contrasting colors and labeled percentages within each segment effectively illustrate the distribution, making it easy to compare the proportions of individuals who do and do not receive tutoring.

**Figure 4. Count of student Extracurricular Participation and Gender**

A computer screen shot of text

Description automatically generated

A graph with numbers and a bar

Description automatically generated with medium confidence

**Insight:** The bar chart shows that most students do not participate in extracurricular activities, with 757 females and 718 males. Participation among those who do engage is lower but similar, with 465 females and 452 males, indicating minimal gender differences.

**Visualization:** The bar chart uses two distinct colors to differentiate genders: purple represents females, and orange represents males. This color distinction allows for clear visual comparison between the two groups across categories of extracurricular participation.

**Figure 5.** Ethnicity and Parental Support

**A computer code on a black background

Description automatically generated**

A graph of ethnicity and parental support

Description automatically generated

**Insight:** The bar chart shows that Caucasians have the highest parental support, especially in Moderate (368) and Low (254) categories. Asians and African Americans follow a similar pattern, with higher Moderate support (148 and 157, respectively). The Others group has the lowest counts overall, with Moderate (67) and Low (39) support being most prominent. Very High and High support levels are consistently low across all ethnic groups.

**Visualization:** The grouped bar chart uses blue for "Moderate" and gray for "Low" parental support, with side-by-side bars and data labels for clear, precise comparison across ethnic groups.

**Figure 6. Distribution of sports**

A computer code on a dark background

Description automatically generated

A blue and orange pie chart

Description automatically generated

**Insight:** The pie chart shows that 70% of individuals do not participate in sports, while only 30% participate. This highlights a significant disparity, with the majority opting out of sports activities.

**Visualization:** The pie chart uses blue to represent "No" (non-participation) and orange for "Yes" (participation). The use of distinct colors and labeled percentages provides a clear and intuitive understanding of the distribution of sports participation.

**Figure 7. Parental Support and Parental Education**

**A computer code on a blue background

Description automatically generated**

**A graph of different colored bars

Description automatically generated**

**Insight:** The bar chart shows that Moderate parental support is the most common, especially among parents with Higher and High School education. Low support follows but decreases across education levels, while Very High and High support are the least common, with consistently lower counts across all education categories.

**Visualization:** The grouped bar chart uses four distinct colors to represent parental education levels: dark blue for "Higher," blue for "High School," teal for "Some College," and green for "Bachelor's." The side-by-side bar format enables clear comparison across parental support categories, with data labels providing additional clarity.

**Figure 8.** Grade class and Tutoring

**A computer code on a black background

Description automatically generated**

**A graph with numbers and a bar

Description automatically generated**

**Insight:** The bar chart shows that most students across all grade classes do not receive tutoring, with the gap being largest in Grade 4. Tutoring participation is consistently lower but increases slightly with higher grade levels.

**Visualization:** The chart uses blue for "No" and green for "Yes," with side-by-side bars and data labels for clear comparison of tutoring participation across grade classes.

**METHODS**

**Tools and Techniques**

The analysis employed the following tools and techniques:

* Data Preprocessing: The Pandas library was utilized for data cleaning and preparation, ensuring the dataset was ready for analysis.
* Visualization: Matplotlib and Seaborn libraries were used extensively to create informative visualizations.
* Analysis Techniques: Grouping and aggregation methods were applied to uncover meaningful relationships between variables.

**Data Importation**: The dataset was imported using the pandas library, ensuring that all necessary libraries (Pandas, Seaborn, Matplotlib) were installed for data analysis and visualization. The dataset was loaded using the read\_csv function from Pandas, which is essential for handling CSV files.

A computer screen shot of text

Description automatically generated

**Data Grouping**

Data grouping by attributes such as Gender, Age, and Extracurricular enabled detailed examination of patterns and trends within the dataset, providing a granular understanding of the factors affecting student performance.

**MODELING AND RESULTS**

**Key Insights**

1. **Gender Distribution**: The dataset exhibited a balanced representation of male and female students.
2. **Age Patterns**: The majority of students were within a typical school-age range, indicating a standard demographic.
3. **Extracurricular Activities**: Gender differences were apparent in extracurricular participation rates.
4. **Parental Support**: A strong correlation was observed between parental support and extracurricular involvement, as well as academic success.
5. **Tutoring**: Higher-grade students showed a greater likelihood of receiving tutoring support, suggesting its importance in academic progression.

**CONCLUSION**

This analysis provided valuable insights into the demographic and behavioral factors influencing student performance. The balanced gender distribution, age demographics, and varying levels of extracurricular participation highlight the diverse nature of the dataset. Parental support emerged as a crucial factor, correlating strongly with academic success and extracurricular involvement. Additionally, tutoring was found to play a significant role in the academic advancement of students in higher grade classifications.

Future research could delve deeper into the relationship between GPA and extracurricular activities, employing predictive modeling techniques to uncover causal relationships. By leveraging the analytical capabilities of Python, this study successfully uncovered and visualized critical trends within the dataset, offering a foundation for data-driven decision-making in educational contexts.

**TEAM CONTRIBUTIONS:**

Jayr Serrano - Project leader/research planner

Cristine Galvan - Conduct or explore data output for the research

Jason Belano - Helps interpret results

Franzine Arabella Martinez - Manage project

Rollyka Amomas- Summarize findings

Rommel Doinog - Visualize data

Zarmaine Grace Pascasio - Ensure clarity of output

**REFERENCE**

1. Here is the citation in APA format for the given Kaggle dataset:

Kharoua, R. (2023). *Students Performance Dataset*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset>.

2. Here is the citation in APA format for the given GitHub repository:

Ameer, E. (2023). *Prodigy DA Task 01*. GitHub. Retrieved from <https://github.com/12eymanameer/prodigy_da_task_01>.

3. Here is the citation in APA format for my own GitHub repository:

itsmejaaayr. (n.d.). *Student-Performance*. GitHub. Retrieved December 19, 2024, from <https://github.com/itsmejaaayr/Student-Performance>