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.

**Handling Missing Values**

A preliminary check for missing values was conducted using the isnull() method in Pandas. The sum of missing values for each column was calculated, and no missing values were detected in the dataset. This ensured the data was complete and ready for analysis without requiring imputation or removal of records.

**Exploratory Data Analysis**

**A computer screen shot of a code

Description automatically generatedFigure 1.** Count of Gender

**A graph with blue and yellow squares

Description automatically generated**

This code generated a bar chart that revealed an almost equal representation of male and female students, indicating balanced participation across genders.

**A screen shot of a computer code

Description automatically generatedFigure 2.** Count by age

A graph of different colored bars

Description automatically generated with medium confidence

The analysis demonstrated that most students fall within the typical school-age range, reflecting a standard demographic profile for the dataset.

**Figure 3.** Distribution of tutoring

**A pie chart with a number of percentages

Description automatically generatedA computer screen shot of a computer code

Description automatically generated**

A pie chart was used to examine the proportion of students receiving tutoring services. The following code generated the visualization

A screenshot of a computer program

Description automatically generated**Figure 4.** Count of student Extracurricular Participation and Gender

A graph with orange and purple bars

Description automatically generated

The results indicated significant differences in participation rates across genders, shedding light on the role gender plays in extracurricular involvement.

**Figure 5.** Ethnicity and Parental Support

**A computer screen with text

Description automatically generated**

A graph of equality and parental support

Description automatically generated

This visualization suggested that parental support levels varied across different ethnic groups, providing valuable cultural insights.

**Figure 6**. Distribution of sports

A blue and orange pie chart

Description automatically generatedA computer screen shot of a computer code

Description automatically generated

The analysis demonstrated the popularity of sports within the student population, offering a view of how extracurricular activities influence student life.

**A white screen with red text

Description automatically generatedFigure 7.** Parental Support and Parental Education

**A graph of different colored bars

Description automatically generated**

The findings showed clear trends linking higher educational attainment with increased levels of parental support.

**Figure 8.** Grade class and Tutoring

**A screenshot of a computer

Description automatically generated**

**A graph with blue and green bars

Description automatically generated**

The analysis revealed that students in higher grade classifications were more likely to receive tutoring support.

**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 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.

***Original file is located at***

***https://colab.research.google.com/drive/1RsJc0tSCn6XHxo7yL5INrls0jNTIG1rG***

***"""***

***#importing pandas library***

**import pandas as pd**

***#import matplotlib***

**import matplotlib.pyplot as plt**

***#importing seaborn library***

**import seaborn as sns**

***#loading data***

**student\_performance=pd.read\_csv('Student\_performance\_data.csv')**

**df=pd.DataFrame(student\_performance)**

***# View first five rows of the dataset***

**student\_performance.head()**

**student\_performance.isnull().sum()**

***#Total Students***

**print(df['StudentID'].count())**

***#Total Extracurricular***

**Extracurricular\_counts = df['Extracurricular'].value\_counts()**

**print(Extracurricular\_counts)**

***#Total Parental Support***

**ParentalSupport\_counts = df['ParentalSupport'].value\_counts()**

**print(ParentalSupport\_counts)**

***#Gender***

**Gender\_counts = df['Gender'].value\_counts()**

**print(Gender\_counts)**

**Gender\_counts = df.groupby(['Gender', 'Gender']).size().unstack()**

***# Plotting***

**Gender\_counts.plot(kind='bar', stacked=True,colormap='plasma',)**

**plt.ylabel('Count')**

**plt.xlabel('Gender')**

**plt.grid(True)**

**plt.title('Figure 1: Count of Gender')**

**plt.show()**

**count\_students = df.groupby(['Age', 'Age']).size().unstack()**

***# Plotting***

**count\_students.plot(kind='bar', stacked=True,colormap='plasma',)**

**plt.ylabel('Count')**

**plt.xlabel('Age')**

**plt.title('Figure 2: Count by Age')**

**plt.show()**

***# Plot Tutoring***

**Tutoring\_counts = df['Tutoring'].value\_counts()**

**plt.pie(Tutoring\_counts, labels=Tutoring\_counts.index, autopct='%1.1f%%',startangle=90)**

**plt.title('Figure 3: Distribution of Tutoring')**

**plt.axis('equal')**

**plt.show()**

**import seaborn as sns**

**import matplotlib.pyplot as plt**

***# Create a DataFrame to get counts of students with different GPA values for extracurricular participation***

**gpa\_counts = df.groupby(['Extracurricular', 'Gender']).size().reset\_index(name='Count')**

***# Bar plot for the counts***

**sns.barplot(data=gpa\_counts, x='Extracurricular', y='Count', hue='Gender', palette='inferno')**

**plt.ylabel('Count')**

**plt.xlabel('Extracurricular Participation')**

**plt.title('Figure 4: Count of Students by Extracurricular Participation and Gender')**

**plt.grid(True)**

**plt.show()**

***# Plot Ethnicity & ParentalSupport***

**sns.countplot(data=df, x='Ethnicity',hue='ParentalSupport',palette='cividis')**

**plt.xlabel('Ethnicity')**

**plt.ylabel('ParentalSupport')**

**plt.grid(True)**

**plt.title('Figure 5: Ethnicity & ParentalSupport')**

**plt.show()**

***# Plot Sports***

**Sports\_counts = df['Sports'].value\_counts()**

**plt.pie(Sports\_counts, labels=Sports\_counts.index, autopct='%1.1f%%',startangle=90)**

**plt.title('Figure 6: Distribution of Sports')**

**plt.axis('equal')**

**plt.show()**

***# Plot ParentalSupport & ParentalEducation***

**sns.countplot(data=df, x='ParentalSupport',hue='ParentalEducation',palette='viridis')**

**plt.xlabel('ParentalSupport')**

**plt.ylabel('ParentalEducation')**

**plt.grid(True)**

**plt.title('Figure 7: ParentalSupport & ParentalEducation')**

**plt.show()**

***# Plot GradeClass & Tutoring***

**sns.countplot(data=df, x='GradeClass',hue='Tutoring',palette='viridis')**

**plt.xlabel('GradeClass')**

**plt.ylabel('Tutoring')**

**plt.grid(True)**

**plt.title('Figure 8: GradeClass & Tutoring')**

**plt.show()**

**REFERENCE**

12eymanameer. (n.d.). *Prototyping data analysis task 01*. GitHub. Retrieved December 15, 2024, from <https://github.com/12eymanameer/prodigy_da_task_01>