**Analyzing Factors Influencing Student Performance Using Python**

**Project Description:**

This project aims to explore and analyze various factors that impact the academic performance of students. By leveraging Python libraries like **pandas**, **matplotlib**, and **seaborn**, we performed an in-depth analysis of a dataset containing demographic and academic information of over **30,000 students**. The primary objective is to identify trends and correlations that may influence students' scores in subjects such as Math, Reading, and Writing.

**Objective:**

The key objectives of this project are:

1. **Data Cleaning & Preparation**: Handle missing values and drop irrelevant columns.
2. **Data Analysis**: Analyze the impact of various factors like gender, parental education, marital status, ethnic group, transportation means, and study hours on student performance.
3. **Visualization**: Use visual tools such as count plots, heatmaps, box plots, and pie charts to derive insights from the data.
4. **Outlier Detection**: Identify outliers in the student scores to understand score distribution.

**Data Overview:**

* **Dataset**: The dataset consists of **30,641 entries** and **15 columns**, including information like Gender, Ethnic Group, Parent Education, Marital Status, Transportation Means, Weekly Study Hours, and scores in Math, Reading, and Writing.
* **Data Fields**:
  + Demographic details (Gender, Ethnic Group, Parent Education)
  + Academic performance (MathScore, ReadingScore, WritingScore)
  + Socioeconomic factors (LunchType, ParentMaritalStatus, TransportationMeans)
  + Other factors (Test Preparation, PracticeSport, Weekly Study Hours)

**Key Analyses and Insights:**

1. **Gender Distribution**:
   * The analysis revealed that there are more female students than male students in the dataset.
   * Visualized using a count plot.
2. **Parental Education vs. Student Scores**:
   * A heatmap analysis showed that students whose parents have a higher level of education tend to perform better in Math, Reading, and Writing.
   * This highlights the positive correlation between parental education and academic success.
3. **Parental Marital Status**:
   * Grouping by marital status showed minimal to no impact on students' academic performance.
   * A heatmap confirmed that there’s negligible correlation between marital status and student scores.
4. **Outlier Detection**:
   * Box plots were used to detect outliers in Math, Reading, and Writing scores.
   * Identifying these outliers can help understand exceptional cases or errors in data entry.
5. **Ethnic Group Distribution**:
   * The distribution of students across various ethnic groups was analyzed using pie charts and count plots.
   * This provided insights into the demographic composition of the dataset.
6. **Impact of Transportation Means on Scores**:
   * Students who use a school bus as a means of transport tend to have better scores compared to those using private transportation.
   * This insight can guide schools in improving transportation policies to support student performance.
7. **Weekly Study Hours**:
   * A count plot showed that students who study between **5-10 hours per week** achieve better scores.
   * This suggests that moderate but consistent study hours are more effective for better performance.

**Future Scope:**

* Applying **machine learning algorithms** to predict student performance based on demographic and behavioral factors.
* Conducting further analysis on other aspects such as test preparation, sports involvement, and study environments.
* Extending the analysis to a broader dataset to validate the insights across different regions or educational systems.

**Analysis:**

**Step 1: Importing Libraries**

The code starts by importing the necessary Python libraries for data analysis and visualization:

* **pandas**: For data manipulation and analysis.
* **matplotlib.pyplot**: For plotting graphs.
* **numpy**: For numerical operations.
* **seaborn**: For creating statistical data visualizations.

**Step 2: Reading and Inspecting the Dataset**

* The dataset is loaded using pd.read\_csv().
* The initial dataset is displayed to get an overview of the data.

**Data Overview:**

* The dataset contains **30,641 entries** and **15 columns** with various data types, such as integers, floats, and objects (strings).

**Data Summary:**

* df.describe() gives a statistical summary of the numerical columns.
* df.info() reveals information about the dataset, including non-null counts and data types for each column.
* df.isnull().sum() shows that several columns have missing values.

**Step 3: Dropping Unnecessary Columns**

The column "Unnamed: 0" is dropped as it doesn't provide useful information for analysis.

**Step 4: Gender Distribution Visualization**

A **count plot** is created to visualize the distribution of genders among students:

plt.figure(figsize=(5, 5))

ax = sns.countplot(data=df, x="Gender")

ax.bar\_label(ax.containers[0])

plt.title("Gender Distribution")

plt.show()

* **Insight**: The number of females is higher than the number of males in the dataset.

**Step 5: Parent Education vs. Student Scores**

The code groups the data by "ParentEduc" (parent's education level) and calculates the mean scores for Math, Reading, and Writing:

gb = df.groupby("ParentEduc").agg({"MathScore": "mean", "ReadingScore": "mean", "WritingScore": "mean"})

* A **heatmap** is used to visualize the relationship between the education level of parents and the students' performance.
* **Insight**: Higher parental education seems to correlate with better student scores.

**Step 6: Parent Marital Status vs. Student Scores**

Similarly, the dataset is grouped by "ParentMaritalStatus":

gb1 = df.groupby("ParentMaritalStatus").agg({"MathScore": "mean", "ReadingScore": "mean", "WritingScore": "mean"})

* A heatmap is created to analyze the impact of parents' marital status on student scores.
* **Insight**: The marital status of parents has a negligible or no impact on students' performance.

**Step 7: Identifying Outliers in Scores**

The code uses **box plots** to detect outliers in the Math, Reading, and Writing scores:

sns.boxplot(data=df, x="MathScore")

sns.boxplot(data=df, x="ReadingScore")

sns.boxplot(data=df, x="WritingScore")

* **Insight**: The box plots help identify any outliers in student scores across subjects.

**Step 8: Distribution of Ethnic Groups**

* The unique ethnic groups are identified:

df["EthnicGroup"].unique()

* The distribution of students across ethnic groups is visualized using a **pie chart** and a **count plot**.

**Insights**:

* The pie chart shows the proportion of each ethnic group.
* The count plot displays the count of students in each ethnic group.

**Step 9: Transportation Method vs. Student Scores**

* The dataset is grouped by "TransportMeans" to see how transportation impacts student scores:

trans = df.groupby("TransportMeans").agg({"MathScore": "mean", "ReadingScore": "mean", "WritingScore": "mean"})

* A count plot is used to visualize the distribution of transportation means.
* **Insight**: Students using school buses tend to have better performance compared to those using private transport.

**Step 10: Weekly Study Hours vs. Student Performance**

* A count plot is created for the "WklyStudyHours" column:

ax = sns.countplot(data=df, x="WklyStudyHours")

plt.title("Impact on Study Hours")

ax.bar\_label(ax.containers[0])

* **Insight**: Students who study between **5-10 hours per week** tend to have higher scores.

**Overall Analysis Summary**

1. **Gender Distribution**: There are more female students than male students.
2. **Parental Education Impact**: Higher parental education correlates positively with better student scores.
3. **Parental Marital Status**: Little to no effect on students' academic performance.
4. **Outliers**: Outliers were detected in Math, Reading, and Writing scores.
5. **Ethnic Group Distribution**: Students are distributed across different ethnic groups with varying performance levels.
6. **Transportation Method**: Using a school bus appears to positively influence academic performance.
7. **Study Hours**: The optimal weekly study hours for better performance are between 5-10 hours.