Project Overview

A synthetic dataset of 100 students is generated using NumPy and Pandas. Each record simulates student-level metrics across study habits, academic results, and well-being factors. I created this project to practice analytical workflows in education-related data, where student outcomes are influenced by multiple overlapping lifestyle and learning factors. While the project relies on synthetic data for analysis(due to the lack of access to the real student metrics), predictive models used can be applied to forecast risk of underperformance at real schools.

1. Data Generation and Preparation

Generated Features:

- study_hours_per_week, quiz_average, project_score
- participation_score, attendance_rate, stress_level
- sleep_hours, extracurricular_hours, assignments_completed

Data Cleaning Includes:

- Clipping outliers for scores and hours
- Categorizing continuous features into bins (e.g., sleep categories, performance quartiles)

---

2. Statistical Analysis and Correlations

A series of statistical analyses and visualizations are performed to understand key relationships:

Study Habits vs Performance

- **Study Hours & Quiz Performance**: Correlation analysis and regression plot
- **Sleep Duration**: Boxplot comparison of quiz scores by sleep category

Attendance & Project Outcomes

- **Attendance Rate & Project Score**: Correlation with participation shown using a color scale

Stress Impact

- **Stress Level Correlation**: Heatmap showing stress vs. quiz, project, and participation scores

Time Management

- Bubble chart comparing weekly study vs extracurricular hours, with bubble size = quiz performance

---

3. Clustering and Comparative Performance Analysis

K-Means Clustering

- Students are clustered into 3 performance groups using quiz, project, and participation scores
- Scatter plots colored by cluster for performance segmentation

Assignment and Participation Patterns

- Assignment completion histograms across quiz quartiles
- Violin plots of participation scores by attendance level

Study Efficiency Analysis

- Computation of quiz score per study hour
- Density plots of efficiency distribution

Top vs Bottom Performers

- Radar chart comparing top 10% vs bottom 10% in metrics like
- Attendance rate, study time, project score, sleep, and participation

---

4. Visualizations and Output

The analysis generates various static and interactive plots using:

- **Matplotlib & Seaborn**: For regression, box, violin, and radar plots
- **Plotly**: (Pluggable) for future interactivity (e.g., time allocation scatter)

Key Charts:

- Scatter plots with color/size encoding
- Box and violin plots by category
- Heatmaps for correlation metrics
- Radar plots for profile comparisons
- Histograms by performance quartiles

---

5. Project Files

- **Student_Metrics_Analysis_Script.py** – Python script used in this project (created in PyCharm)
- **Student_Performance_Report_and_Visualizations.pdf** – PDF report showing the student performance analysis and all visualizations generated by the script
- **ReadMe.pdf** – PDF version of the README file