Module -5: Python for Data Analytics

Coding Challenge : Week 18

Real-Time Case Study

Student Performance Analysis for an Education Board

Background Story (Real-Time Scenario)

A State Education Board is concerned about the academic performance of students in secondary schools. The board has collected anonymized data of students appearing for standardized exams to understand how demographic, socio-economic, and academic preparation factors influence performance in:

- Mathematics
- Reading
- Writing

The board wants data-driven insights to:

- Improve teaching strategies
- Identify students who need academic support
- Design better preparation programs

As a Data Analyst, you are given the dataset and asked to perform exploratory data analysis and data wrangling using only Python, Pandas, and NumPy.

Data Collection

- Dataset source - https://www.kaggle.com/datasets/spscientist/students-performance-in-exams/code
- The dataset consists of 8 column and 1000 rows

Dataset Description

The dataset contains the following columns:

Column Name	**Data Description**
gender:	The gender of the student (e.g., male, female).
race_ethnicity:	The group classification of the student (e.g., group A, group B, etc.).
parental_level_of_education:	The highest level of education attained by the student's parent(s) (e.g., bachelor's degree, some college).
lunch:	The type of lunch the student receives (e.g., standard, free/reduced).
test_preparation_course:	Whether the student completed a test preparation course (e.g., none, completed).
math_score:	The student's score in mathematics.
reading_score:	The student's score in reading.

🎯 Objective

To analyze student academic performance and understand how background factors affect exam scores, using Pandas and NumPy only.

🧪 Problem-Based Question

◇ Task 1: Data Ingestion & Initial Exploration

1. Load the dataset using Pandas.
2. Display:
   - First 5 records
   - Last 5 records
3. Check:
   - Number of rows and columns
   - Column names
   - Data types of each column
4. Generate statistical summary for numerical columns.

📌 ***Objective:*** **Understand dataset structure and content.**

◇ Task 2: Data Quality & Missing Value Analysis

The education board wants to ensure data accuracy.

1. Check if the dataset contains any missing values.
2. Count missing values column-wise.
3. If missing values exist:
   - Handle them appropriately using Pandas methods.
4. Verify the dataset after cleaning.

📌 ***Objective:*** **Ensure clean data for reliable analysis**.

◇ Task 3: Overall Student Performance Analysis

The board wants an overview of student performance.

1. Calculate:
   - Average math score
   - Average reading score
   - Average writing score
2. Identify:
   - Highest score in each subject
   - Lowest score in each subject
3. Find the total score for each student (sum of all three subjects).
4. Add a new column total_score to the dataset.

📌 ***Objective:*** **Measure overall academic performance.**

◇ Task 4: Gender-Based Performance Study

The board wants to know if gender influences academic results.

1. Calculate average scores (math, reading, writing) for each gender.
2. Identify:
   - Which gender performs better in math
   - Which gender performs better in reading and writing
3. Display the performance comparison.

📌 ***Objective:*** **Identify performance patterns across genders.**

◇ Task 5: Impact of Test Preparation Course

The board invested in test preparation programs and wants to measure their impact.

1. Separate students who:
   - Completed the test preparation course
   - Did not complete the course
2. Calculate average scores for both groups.
3. Compare performance across math, reading, and writing.
4. Conclude whether test preparation improves performance.

📌 ***Objective:*** **Evaluate effectiveness of test preparation programs**.

◇ Task 6: Parental Education & Student Performance

The board believes parental education influences student outcomes.

1. Group students by parental level of education.
2. Calculate average scores for each education level.
3. Identify:
   - Highest performing parental education group
   - Lowest performing group

📌 ***Objective:*** **Understand socio-educational influence on learning**.

◇ Task 7: Lunch Program & Academic Achievement

The government provides different lunch programs to students.

1. Analyze student performance based on lunch type.
2. Calculate average scores for each lunch category.
3. Identify which lunch program group performs better overall.

📌 ***Objective:*** **Study the effect of nutrition and welfare programs.**

◇ Task 8: Performance Categorization (Data Manipulation)

To simplify reporting, the board wants students categorized.

1. Create a new column performance_level:
   - Total Score ≥ 250 → Excellent
   - Total Score 200–249 → Good
   - Total Score < 200 → Needs Improvement
2. Count number of students in each category.

📌 ***Objective:*** **Classify students based on academic achievement.**

◇ Task 9: Top & Bottom Performers Identification

The board wants to recognize excellence and support weak students.

1. Identify top 10 students based on total score.
2. Identify bottom 10 students based on total score.
3. Display relevant student details.

📌 ***Objective:*** **Detect high achievers and at-risk students.**

◇ Task 10: Insights & Conclusion (Analytical Thinking)

Based on your analysis, answer:

1. Which factor has the strongest impact on student performance?
2. Does test preparation significantly improve scores?
3. Which subject shows the highest overall performance?
4. Provide 2–3 data-driven recommendations for the education board.

📌 ***Objective:*** **Translate data into real-world decisions.**