Project Report on Midterm Marks Analysis using PySpark

Dataset Description

The dataset contains midterm marks of students across multiple sections. Each subject is scored out of **20 marks**.

During the initial inspection, the following issues were observed:

- All first 20 displayed records belonged to **ALPHA section only**, suggesting possible data imbalance in sample display.
- **Data types**: All marks columns were stored as strings, making them unsuitable for direct numerical analysis.
- **Invalid entries**: Some non-numeric values were found, such as `"A"`, `"MP"`, `"○"`, `"2○"`, and `"I9"`.
- **Typographical errors** in section names were identified (e.g., `"GAMA"` → `"GAMMA"`, `"SGMA"` → `"SIGMA"`).

This indicated the need for thorough **data cleaning and preprocessing** before analysis.

Data Cleaning Process

The preprocessing steps carried out were as follows:

1. **Section corrections**
   - Missing section values were replaced with `"ZETA"`.
   - Typographical corrections were made:
     - `"GAMA"` → `"GAMMA"`
     - `"SGMA"` → `"SIGMA"`
2. **Subject-wise corrections**
   - **DV (Design & Verification)**:
     - `"A"`, `"MP"` → `0`
     - `"2○"` → `20`
     - `"I9"` → `19`
   - **M-II (Mathematics-II)**:
     - `"A"`, `"○"`, `"AB"` → `0`
     - `"I2"` → `12`
     - `"II"` → `11`
     - `"I"` → `1`
   - **PP (Programming Principles)**:

        - `"A"`, `"AB"`, `"MP"` → `0`
3. **Final conversion**
    - All subject marks were cast from **string** to **integer type** for valid statistical analysis.

Analysis Observations

1. Subject-wise Trends

- **Strong subjects**: Fundamentals of Logic (FL) and BEEE showed consistently higher marks, with many students scoring **15 or above**.
- **Weak subjects**: Mathematics-II (M-II) and Programming Principles (PP) displayed significant weaknesses. Some students scored **0** in these subjects, even if they performed well in others.

2. Section-wise Performance

- After cleaning, student distribution across sections was **balanced**.
- Section averages were largely uniform.
- However, **failure counts in M-II** were slightly higher in some sections, indicating a broader struggle with Mathematics.

3. Student Performance Groups

- **High Achievers**: Consistently scored between **18–20** across subjects.
- **Average Performers**: Scored in the **15–18** range, showing balanced performance with room for improvement.
- **Low Performers**: Scored **below 15**, often with weaknesses in M-II and PP.

4. Hidden Performance Gap – Programming Weakness

- Some **high scorers** (overall achievers) still showed **poor results in PP (Programming Principles)**.
- This suggests a **skills gap**: students may excel in theoretical or non-programming subjects but struggle with practical programming concepts.

Visualizations Observed

The notebook generated several useful plots:

- **Bar charts** for subject-wise mark distributions (highlighting stronger vs weaker subjects).
- **Histograms** showing performance clustering of students.
- **Section-wise average marks** plots, which confirmed near-uniform performance.
- **Grade distribution charts**, clearly segmenting students into High, Average, and Low achievers.

Recommendations

1. **For Low Performers (<15 marks)**
    - Conduct **remedial classes** in weak subjects, particularly **M-II and PP**.
    - Provide **extra practice sessions**, doubt-clearing classes, and mentoring.
2. **For Average Performers (15–18 marks)**
    - Encourage with **continuous evaluation tasks** and **weekly assignments**.
    - Motivate them with **small improvement goals** to push into the high achiever category.
3. **For High Achievers (≥18 marks)**
    - Challenge them with **advanced problem-solving tasks** and competitions.
    - Engage them in **peer mentoring** to help weaker classmates.
4. **For High Achievers with Poor Programming Skills (Strong overall, Weak in PP)**
    - Introduce **specialized programming classes** (extra labs, coding practice sessions).
    - Focus on **hands-on coding exercises** rather than theory.
    - Pair them with programming mentors or encourage participation in **hackathons, coding clubs, and project-based learning**.

Conclusion

The PySpark-based analysis of midterm marks shows:

- **Strengths** in subjects like FL and BEEE.
- **Weaknesses** in M-II and PP that need urgent attention.
- **Sectional performance** remains fairly balanced across groups.
- Students can be clearly classified into **High, Average, and Low performers**.