PySpark Mini Project Report

Prepared By: K. Navadeep (2211CS010145)
Course: Big Data Analytics
Date: October 2025

Dataset Description (District_PGI_2023-24_0.csv)
Columns Identified:
- State → Name of the state.
- District → District name within the state.
- Overall → Composite PGI score indicating district performance.
- Grade → Assigned grade level (e.g., High, Moderate, Low).
- Outcome (290) → Measures educational and social outcomes.
- ECT (90) → Effective Classroom Transaction indicators.
- IF&SE (51) → Infrastructure, Facilities & Student Enrolment.
- SS&CP (35) → School Safety and Child Protection.
- DL (50) → Digital Learning initiatives.
- GP (84) → Governance Processes.

Observations from Executed Cells

1. Data Loading & Cleaning:
   Dataset loaded successfully into a PySpark DataFrame.
   Columns renamed for consistency. Missing or inconsistent data cleaned.
   Non-date fields verified. Nulls in key metrics handled.
2. State-Wise Overview:
   Dataset covers multiple states and districts across India.
   Larger states like Uttar Pradesh, Maharashtra, and Bihar have higher district counts.
   Smaller states/UTs like Goa and Sikkim have fewer districts.
3. Overall Performance Distribution:
   Wide range in Overall PGI scores — top states show strong governance and program implementation.
   Lower-performing states show scope for improvement in education and infrastructure.
4. Grade Analysis:
   Most districts fall under Moderate/Average grades.
   Fewer districts achieve High grades — those serve as benchmarks.
   Low-grade districts highlight areas requiring targeted policy action.
5. Category-Wise Performance:
   Some categories such as Outcome and DL score higher on average.
   IF&SE and GP show moderate to low scores, indicating infrastructure and governance challenges.
6. Correlation Analysis:
   Heatmap shows strong correlation between Outcome and DL categories — improvements in digital learning likely enhance outcomes.
   Weak correlations between SS&CP and others indicate independent focus areas.
7. State-Wise Grade Distribution:
   Stacked bar plots reveal that some states have clusters of high-performing districts, while others show mixed results.
   Useful for designing region-specific improvement programs.

Plots Observed
- Top 10 States by Average PGI Score (bar chart)
- Distribution of District Grades (count plot)
- Category-Wise Average Scores (multi-bar)
- Correlation Heatmap (matrix)

• State-Wise Grade Distribution (stacked bar)

Key Insights

- Overall PGI Distribution: Clear variation across states; top performers can serve as benchmarks.
- Grades: Majority of districts are moderate performers; few excel.
- Category Trends: Outcome and DL categories drive overall scores; governance (GP) needs more focus.
- Correlations: Positive links show inter-dependent improvement potential between learning and outcomes.
- State Patterns: Some states exhibit consistent high grades; others are mixed or low.

Recommendations

1. Targeted Interventions: Focus improvement programs on low-performing districts.
2. Benchmark Best Practices: Study top districts to replicate success strategies.
3. Category-Specific Focus: Strengthen weaker categories like IF&SE and GP.
4. Regular Data Audits: Maintain clean, updated datasets for accurate decision-making.
5. Cross-Category Development: Leverage correlations — e.g., improving DL can lift Outcome scores.

Conclusion

The District Performance Grading Index (PGI) 2023-24 analysis provides an extensive overview of district-level governance and educational performance.
Visual and statistical evaluation reveal strong performers, areas needing attention, and correlations across key categories.
High-performing states demonstrate effective governance, while low-performing regions highlight where policy support and capacity building are required.
Overall, the dataset reflects growing administrative transparency and provides a clear roadmap for data-driven improvement in public service delivery.