Exploratory Data Analysis of PMGSY

Data Science Review Report (CA-I)

Submitted To:

Dr. Piyush Chauhan

Associate Professor

Department of Computer Science and Engineering Symbiosis Institute of Technology, Nagpur

Submitted By:
TanishqGhodpage

Semester: VII

Section: A

PRN: 22070521012

Contents

1 Dataset Overview

Metric	Value
Rows	2,292
Columns	14
Memory	0.62 MB
Missing	0.83%

2 Geographic Coverage

Level	Count
States	32
Districts	713
PMGSY Schemes	5

3 Data Content

Contains sanctioned and completed road and bridge works data:

- Work counts per region

- Infrastructure lengths

- Regional expenditures

- Missing values in `PMGSY_SCHEME` and cost columns

4 Summary

High-quality dataset (99.17% complete) covering rural road development across India with comprehensive state and district-level data.

5 Data Quality Analysis

5.1 Analysis Summary

The code loads the PMGSY dataset and performs in-depth data quality analysis. Dataset shape, memory usage, and detailed summary including column types, null percentages, and unique value counts are analyzed. Visual dashboards created using Plotly show:

[leftmargin=2em]Heatmap of missing value locations Bar charts for missing percentages and unique values Pie chart of data types distribution Gauge for overall completeness

Output shows 0.83% missing values in two columns, with rest fully complete. Data is diverse in types with rich variability across features. Overall completeness: 99.9%.

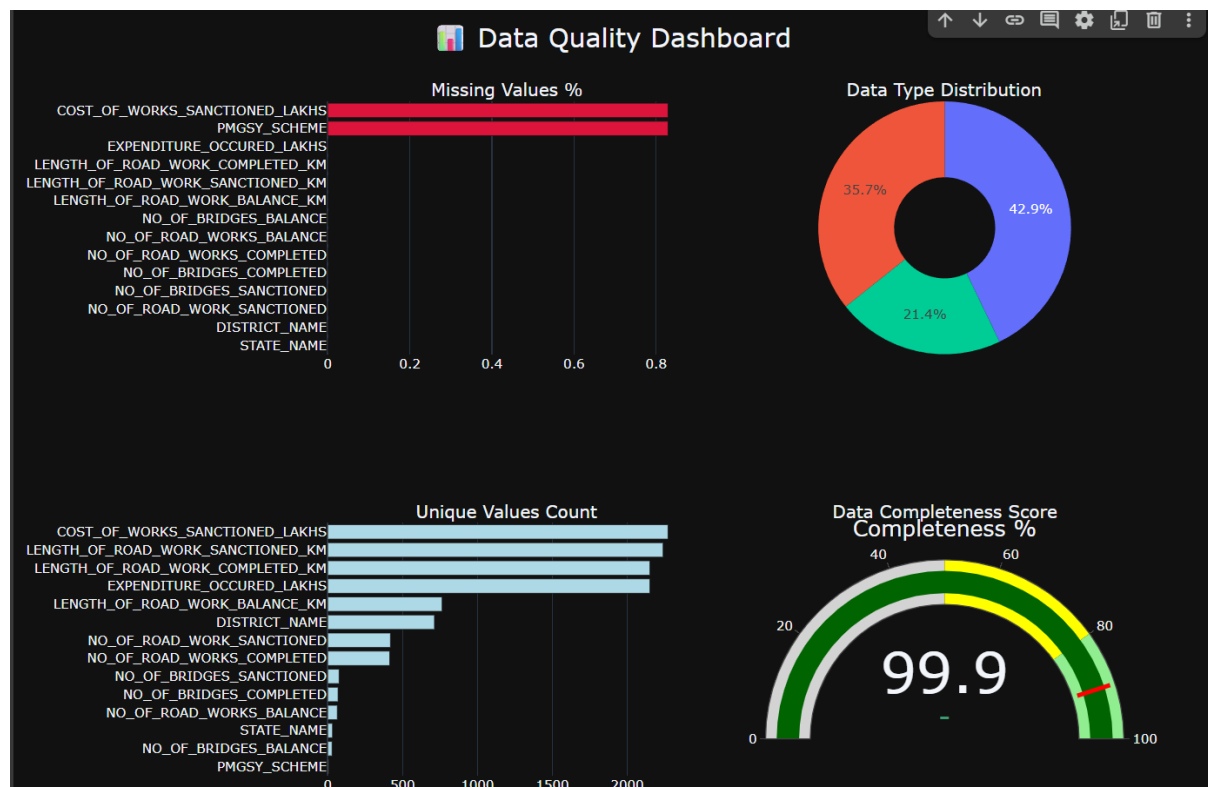## 5.2 Dashboard Visualization

Figure 1: Data Quality Dashboard - Visual confirmation of dataset completeness and structural balance

5.3 Significance

Dashboard visually confirms dataset is nearly complete and structurally well-balanced. Highlights excellent readiness for analysis with minimal preprocessing required.

6 Statistical Distribution Analysis

6.1 Analysis Summary

Statistical distribution analysis of 11 numerical columns from PMGSY dataset. Calculates descriptive statistics, identifies outliers using IQR method, and visualizes four key metrics:

[leftmargin=2em]**Skewness and Kurtosis**: Variables like `NO_OF_BRIDGES_BALANCE` are highly asymmetric and peaked **Coefficient of Variation (CV)**: Shows highest relative variability in bridge-related columns **Zero Value %**: Over 80% values in `NO_OF_BRIDGES_BALANCE` are zeros, highlighting sparsity

Outlier detection found 10-20% of entries in several columns are statistical outliers, affecting modeling or aggregation.
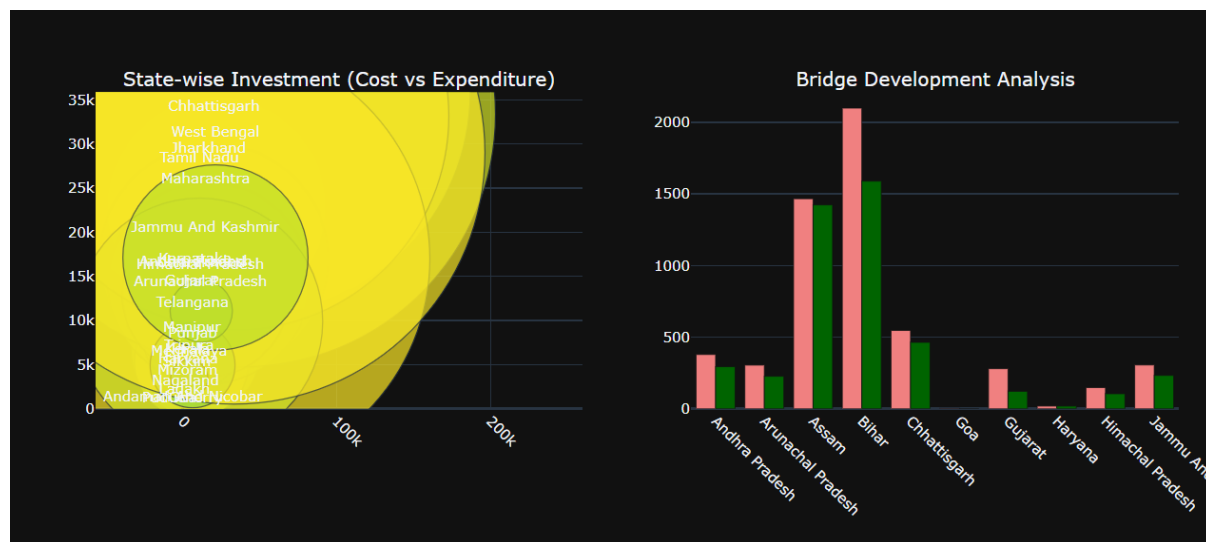
6.2 Distribution Dashboard

Figure 2: Statistical Distribution Analysis - Skewness, Kurtosis, Coefficient of Variation, and Zero Value Percentages

6.3 Significance

Dashboard reveals distribution irregularities like skewness, peakedness, and high zero-ratio features. Guides preprocessing steps such as normalization, outlier handling, or feature transformation before modeling.

7 Geographic Analysis Dashboard

7.1 Analysis Summary

Python code analyzes PMGSY implementation using state-wise data, creating comprehensive geographic dashboard. Groups dataset by `STATE_NAME` and calculates totals for road works, bridges, lengths, costs, and districts.

Key calculations include:

[leftmargin=2em]Road, Bridge, and Length Completion Rates (percentages) State-wise investment analysis (Cost vs Expenditure) District coverage metrics
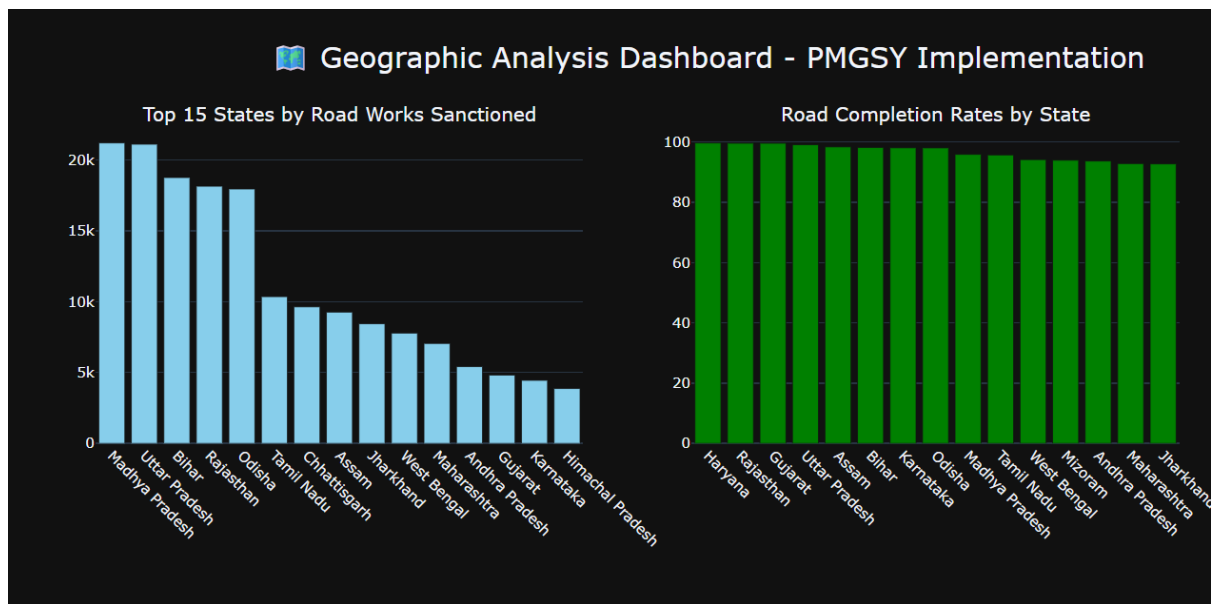
Figure 3: Top 15 States by Road Works and Completion Rates - Madhya Pradesh and Uttar Pradesh lead with 21k+ works
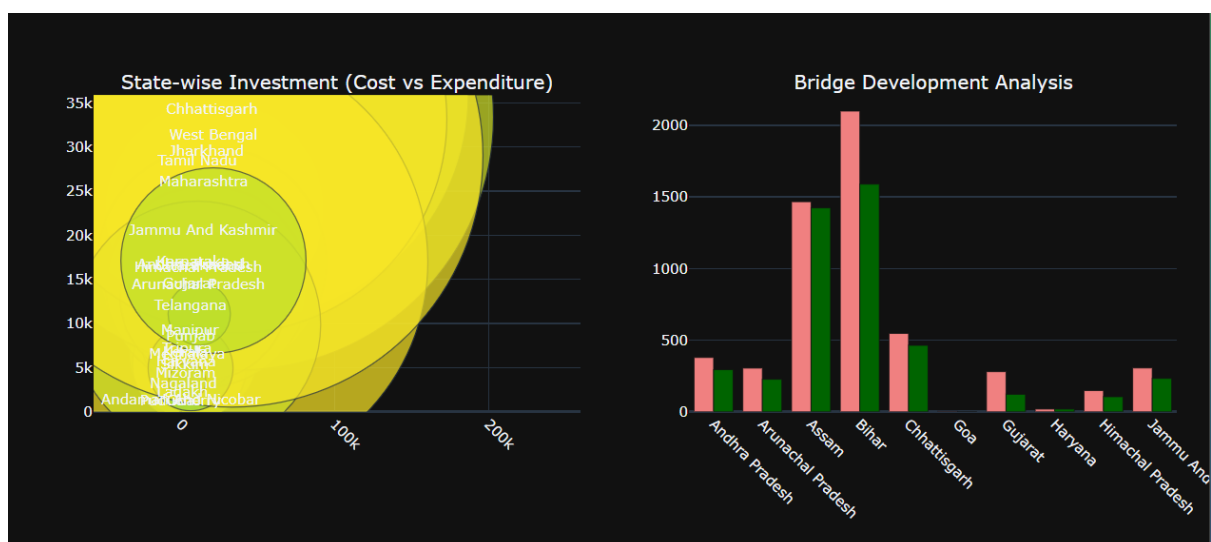

Figure 4: State-wise Investment and Bridge Development - Bihar leads in investment, significant sanctioned vs completed gaps

Category	State	Value
Highest Road Works	Madhya Pradesh	21,217 works
Best Completion Rate	Haryana	99.7%
Highest Investment	Bihar	42,010 Lakhs
Most Districts Covered	Uttar Pradesh	75 districts

Table 1: Top Performing States Across Key Metrics

7.2   Dashboard Visualizations

7.3   Top Performing States

7.4   Significance

Dashboard visually summarizes PMGSY implementation metrics across states, highlighting performance disparities. Enables policymakers to assess infrastructure progress and
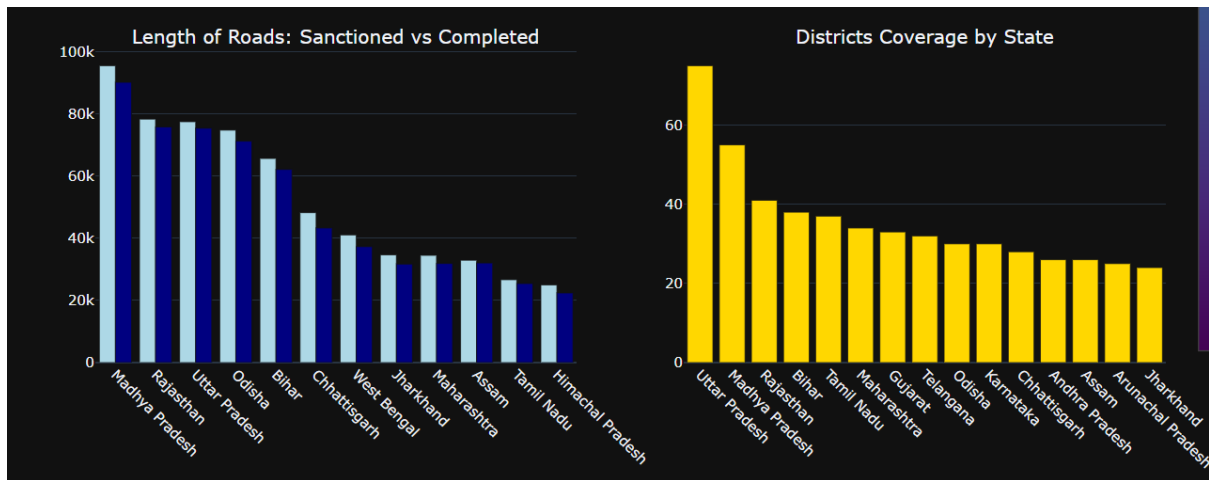
Figure 5: Road Length Analysis and District Coverage - Madhya Pradesh highest road lengths, Uttar Pradesh covers most districts

identify states needing attention or replication.

8 PMGSY Schemes Performance Analysis

8.1 Scheme-wise Analysis Summary

Comprehensive analysis of 5 PMGSY schemes covering road works, bridges, costs, and efficiency metrics. Analysis includes completion rates, investment distribution, cost efficiency, and geographic reach across states and districts.

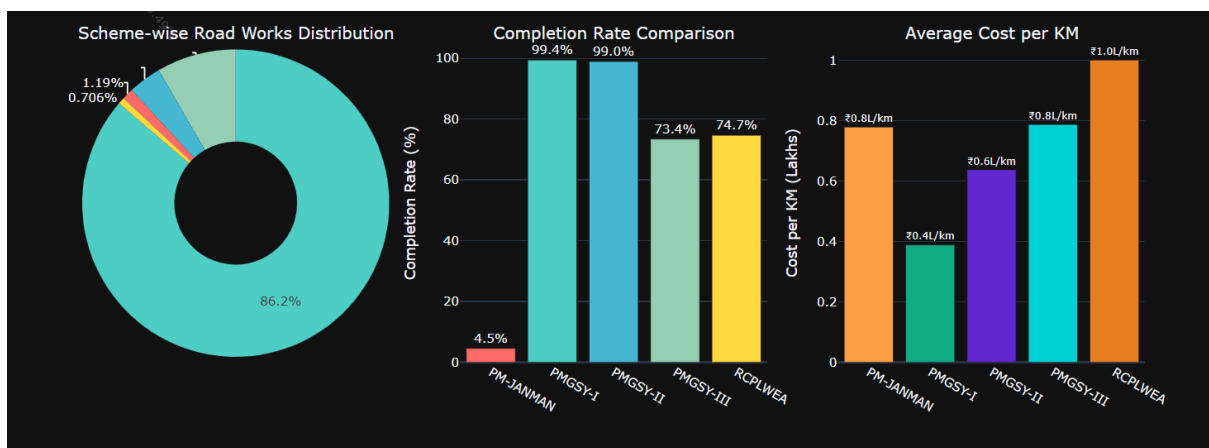## 8.2 Performance Metrics Dashboard

Figure 6: Scheme-wise Road Works Distribution, Completion Rates, and Average Cost per KM
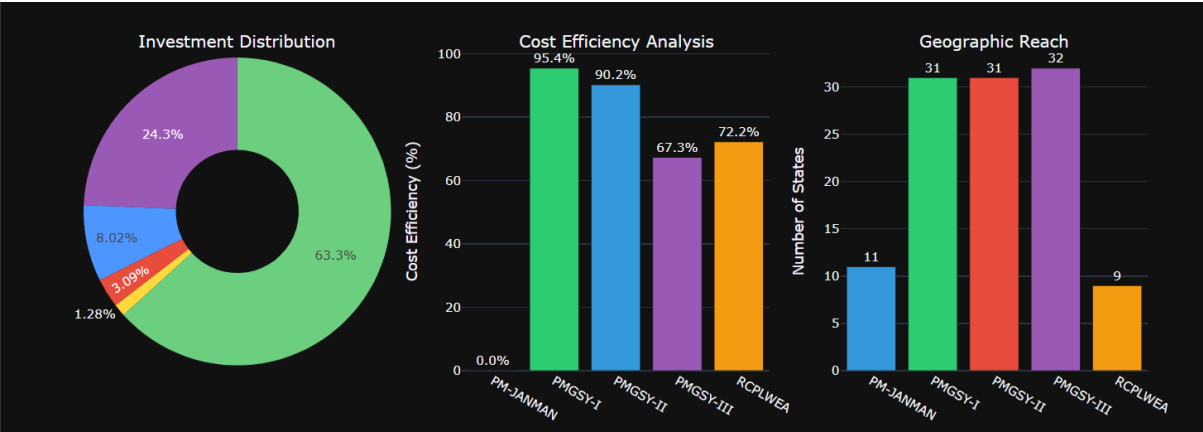
Figure 7: Investment Distribution, Cost Efficiency Analysis, and Geographic Reach by Scheme

Scheme	Roads	Completion%	Investment	States	Cost/KM
PM-JANMAN	2,269	4.5%	50.6 Cr	11	0.78L
PMGSY-I	164,600	99.4%	2,503.6 Cr	31	0.39L
PMGSY-II	6,665	99.0%	317.3 Cr	31	0.64L
PMGSY-III	15,972	73.4%	963.3 Cr	32	0.79L
RCPLWEA	1,347	74.7%	122.3 Cr	9	1.00L

Table 2: Detailed Scheme Performance Comparison

8.3 Detailed Scheme Comparison

8.4 Key Insights

Performance Metric	Leading Scheme
Best Completion Rate	PMGSY-I (99.4%)
Highest Investment	PMGSY-I (2,504 Crores)
Most Cost Efficient	PMGSY-I (95.4%)
Widest Geographic Reach	PMGSY-III (32 states)
Most Districts Covered	PMGSY-I (711 districts)

Table 3: Top Performing Schemes by Key Metrics

8.5 Significance

PMGSY-I emerges as the most successful scheme with 99.4% completion rate and highest cost efficiency (95.4%). PM-JANMAN shows lowest performance with 4.5% completion. Analysis reveals significant performance variations across schemes, highlighting need for targeted improvements in newer schemes like PM-JANMAN and PMGSY-III.

Performance Metric	Leading Scheme
Best Completion Rate	PMGSY-I (99.4%)
Highest Investment	PMGSY-I (2,504 Crores)
Most Cost Efficient	PMGSY-I (95.4%)
Widest Geographic Reach	PMGSY-III (32 states)
Most Districts Covered	PMGSY-I (711 districts)

Table 4: Top Performing Schemes by Key Metrics

8.6   Key Insights

8.7   Significance

PMGSY-I emerges as the most successful scheme with 99.4% completion rate and highest cost efficiency (95.4%). PM-JANMAN shows lowest performance with 4.5% completion. Analysis reveals significant performance variations across schemes, highlighting need for targeted improvements in newer schemes like PM-JANMAN and PMGSY-III.

8.8   Performance Analytics Dashboard

The comprehensive performance analysis reveals an overall completion rate of 95.8% with cost utilization at 86.2%. The average cost per kilometer stands at 0.63 lakhs, demonstrating efficient resource allocation across projects.
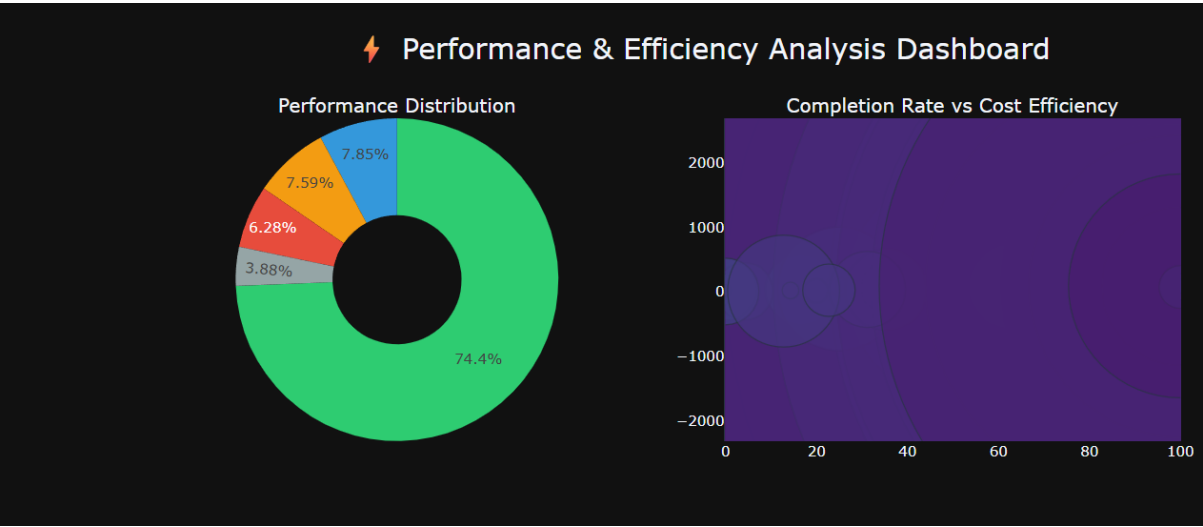

Figure 8: Performance Distribution and Completion Rate vs Cost Efficiency Analysis

The performance distribution shows 74.4% of projects achieving excellent status, while geographic variations highlight state-wise implementation challenges. The state-wise performance heatmap demonstrates strong performance in northeastern states and varying efficiency across different schemes.

Progress tracking indicates PMGSY-I's dominance with substantial investment allocation, while efficiency metrics reveal consistent performance patterns across different implementation phases. Box plot analysis shows performance variability within each scheme category.
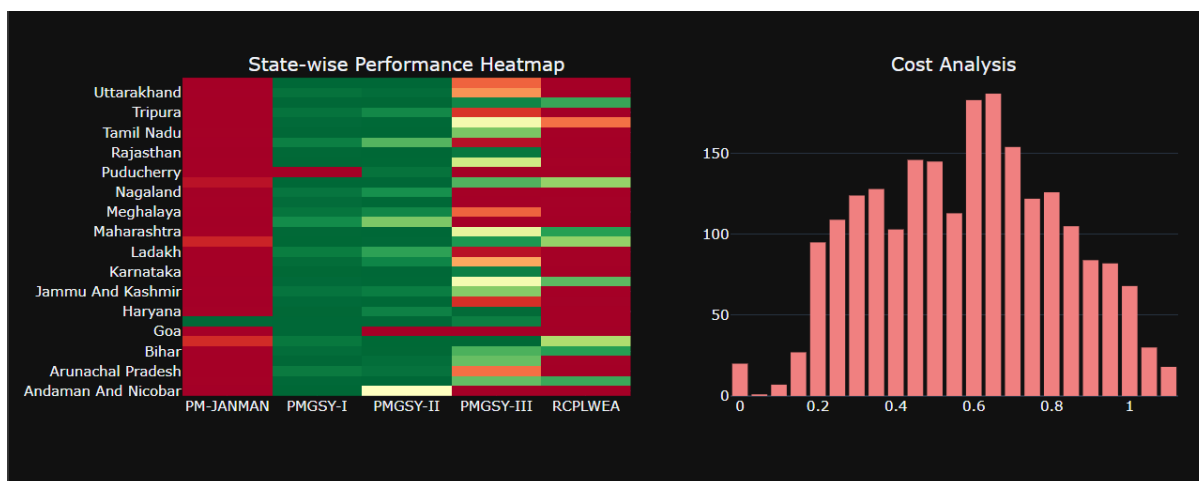
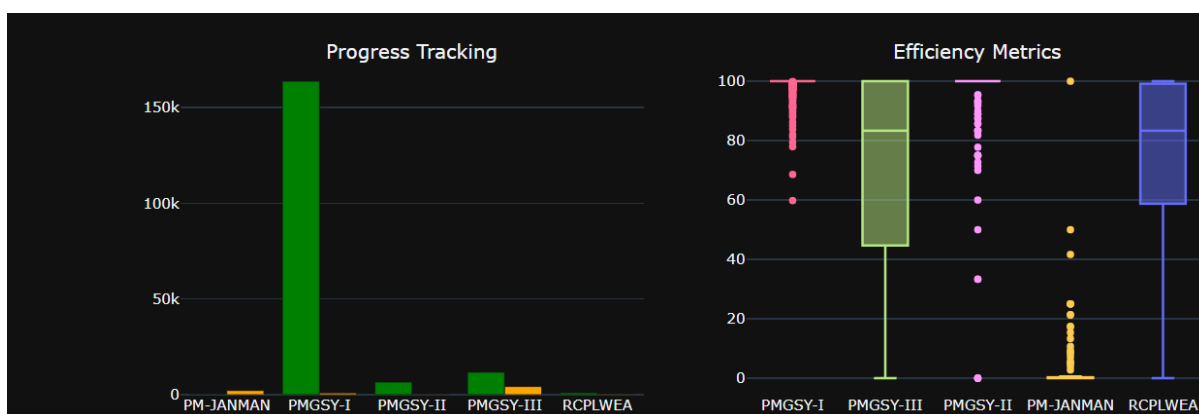Figure 9: State-wise Performance Heatmap and Cost Analysis Distribution

Figure 10: Progress Tracking and Efficiency Metrics Comparison

Top performing districts include Andaman And Nicobar and multiple Andhra Pradesh districts achieving 100% completion rates. However, attention is needed for underperforming regions, particularly in PM-JANMAN implementations where completion rates drop as low as 0.6% in certain districts.

8.9   Correlation Analysis and Cost Patterns

The correlation matrix reveals strong positive relationships between key performance indicators, with sanctioned and completed road works showing nearly perfect correlation (r=0.998). Cost-related metrics demonstrate high interdependence, indicating systematic project execution patterns across schemes.

Cost efficiency distribution analysis shows significant variation across schemes, with PMGSY-I displaying the widest range of efficiency values, while PM-JANMAN shows minimal cost utilization activity. The violin plots reveal distinct performance patterns for each implementation phase.

The expenditure versus sanctioned cost relationship demonstrates strong linear correlation (r=0.978), indicating effective budget utilization across projects. Larger projects show better completion rates (95.5% for very large projects) compared to smaller implementations, suggesting economies of scale benefits.
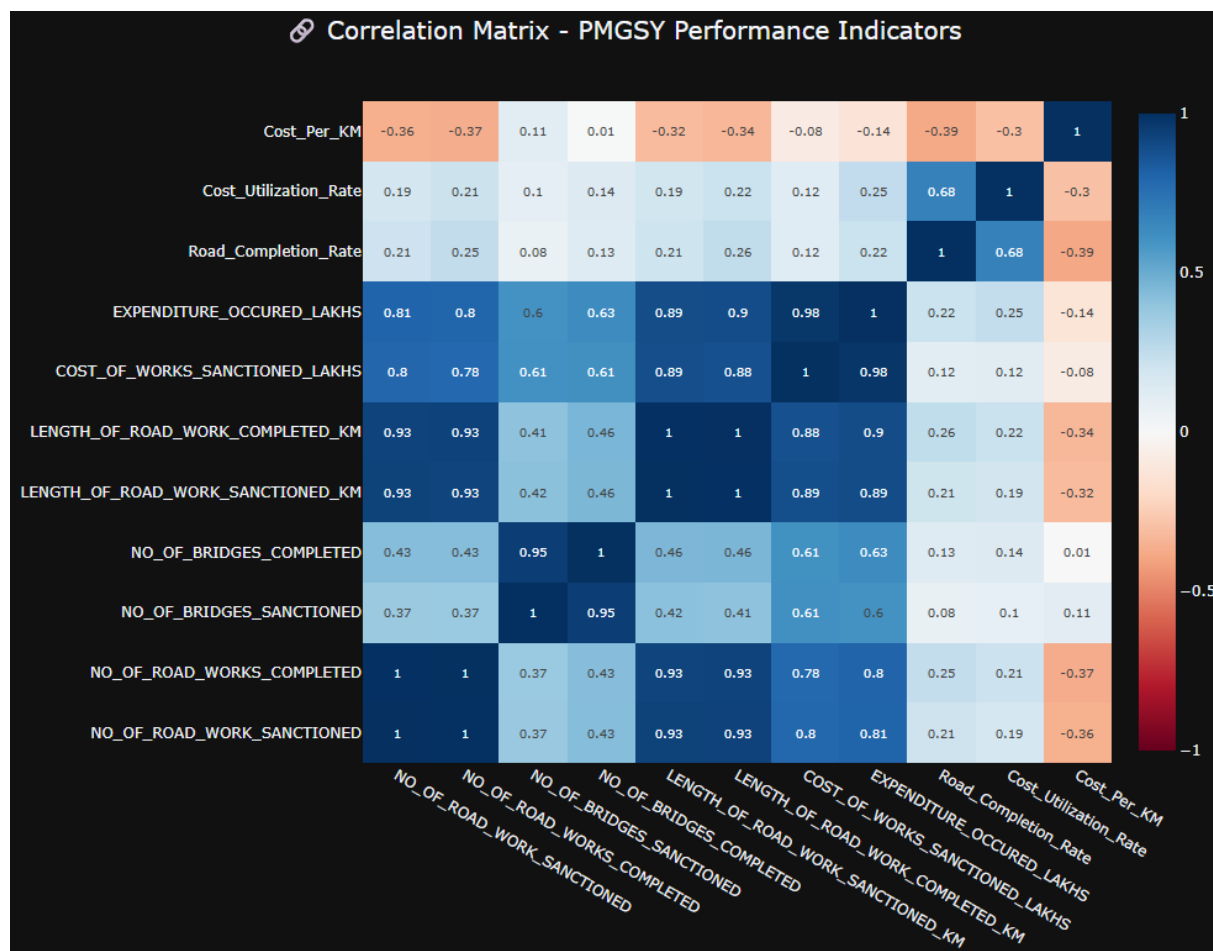
Figure 11: Correlation Matrix of PMGSY Performance Indicators
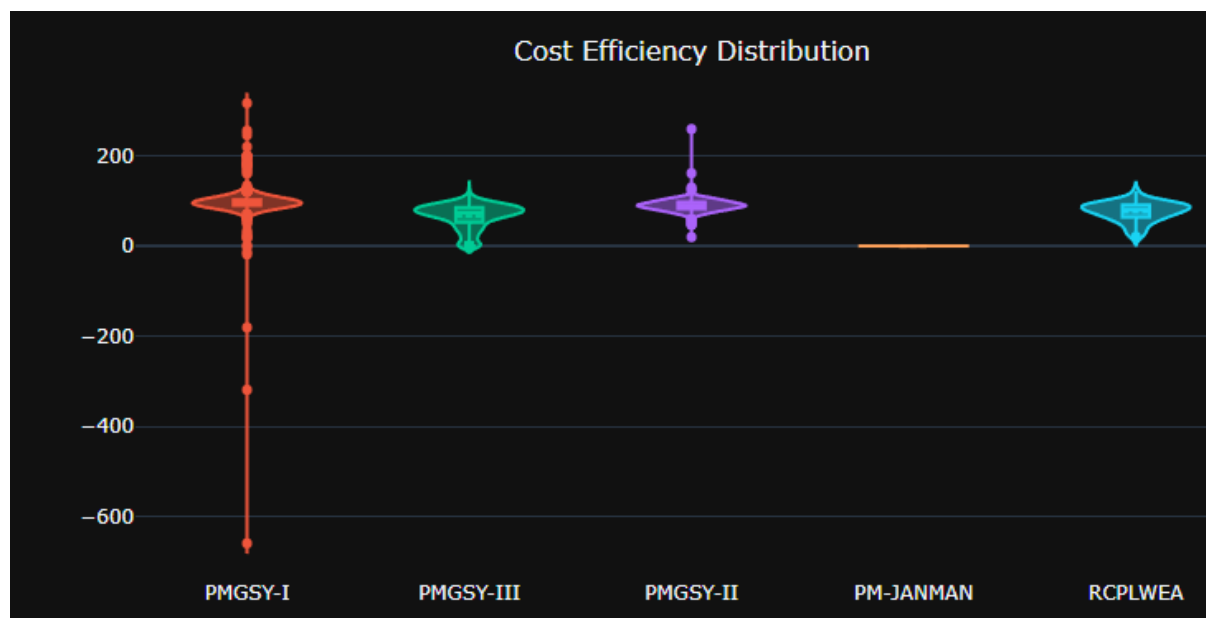

Figure 12: Cost Efficiency Distribution by Scheme

Analysis identifies 16 strong correlations with cost utilization varying significantly: PMGSY-I (95.8%), PMGSY-II (89.7%), RCPLWEA (75.8%), PMGSY-III (66.0%), while
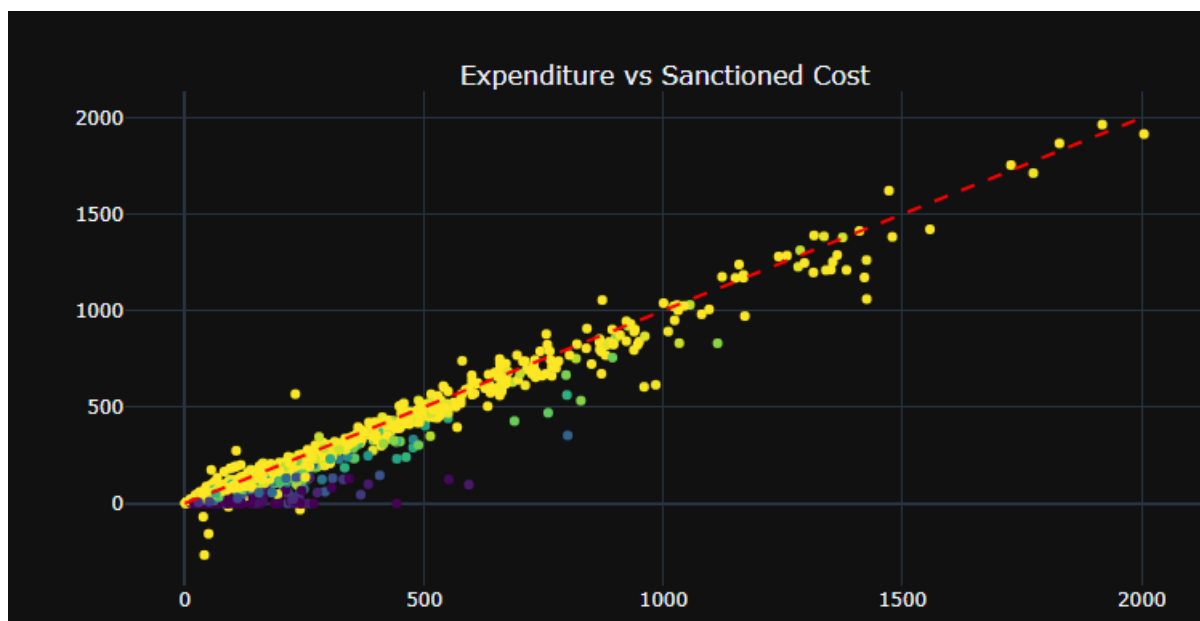
Figure 13: Expenditure vs Sanctioned Cost Analysis

PM-JANMAN shows zero cost utilization, highlighting implementation challenges in tribal area development initiatives.

8.10   Trend Analysis and Performance Evolution

Project size analysis reveals optimal performance at very large scale implementations (95.5% completion rate), demonstrating clear economies of scale benefits in rural road development programs.
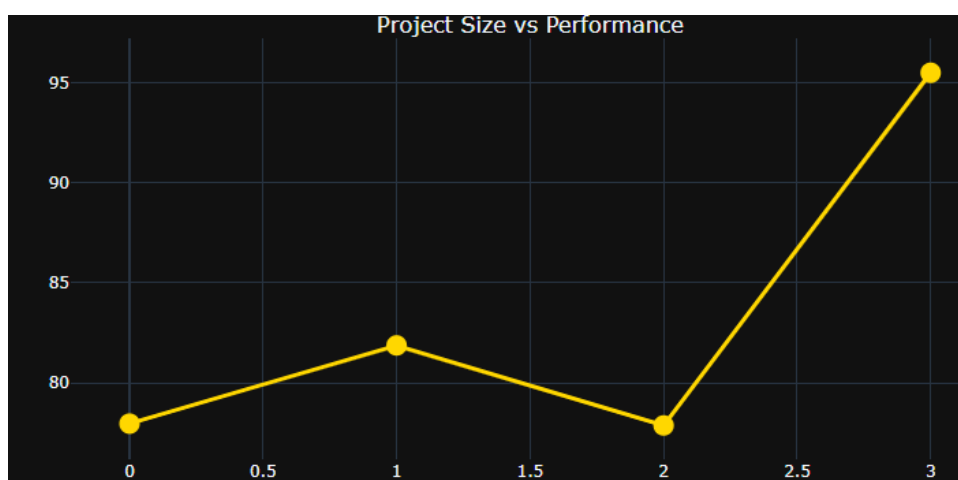

Figure 14: Project Size vs Performance Analysis

Scheme evolution demonstrates declining completion rates across phases, from PMGSY-I's 99.2% to PMGSY-III's 68.8%, while cost per kilometer increased from 0.37 to 0.81 lakhs. This trend indicates implementation challenges in newer phases despite technological advancement.
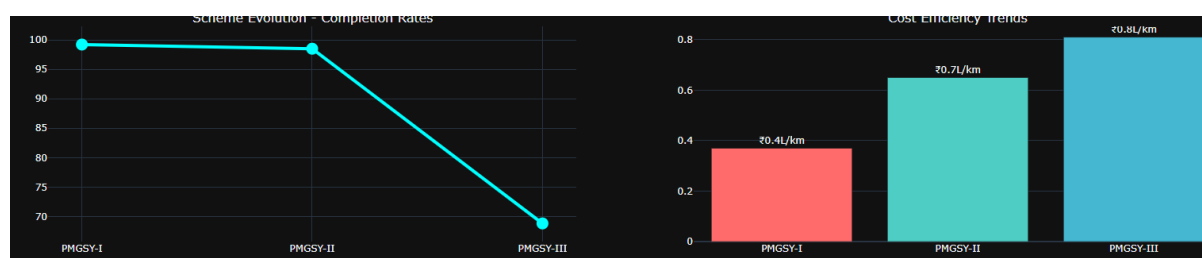
Figure 15: Scheme Evolution - Completion Rates and Cost Efficiency Trends

Project scale evolution shows PMGSY-I handling the largest volume (164,600 projects) with consistent performance, while PMGSY-III demonstrates high variability (=35.2%) compared to earlier phases. Performance consistency analysis reveals increasing unpredictability in newer implementations.

Figure 16: Project Scale Evolution and Performance Consistency

Investment patterns show declining efficiency across phases, with predictive modeling indicating potential completion rates of 58.5% for PMGSY-IV and 43.2% for PMGSY-V, suggesting need for strategic intervention to reverse the negative trajectory.
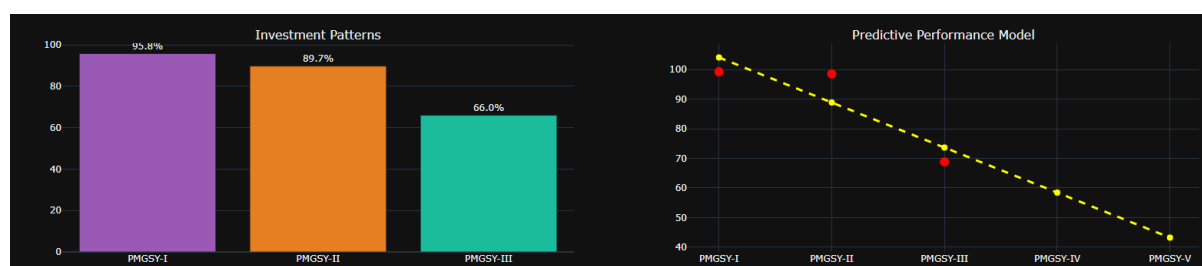

Figure 17: Investment Patterns and Predictive Performance Model

The trend analysis reveals a strong negative correlation (R^2=0.768) between phase progression and performance, with completion rates declining by 15.2% per phase while costs increase by 0.2 lakhs per kilometer, necessitating policy reforms for future implementations.

8.11 State-wise Performance Analysis and Regional Disparities

The state-wise performance analysis reveals significant regional variations in PMGSY implementation effectiveness. Top-performing states demonstrate exceptional completion rates, with Haryana and Rajasthan achieving 99.7% completion rates, followed closely

by Gujarat at 99.6%. This performance clustering among northwestern states suggests effective regional coordination and implementation strategies.
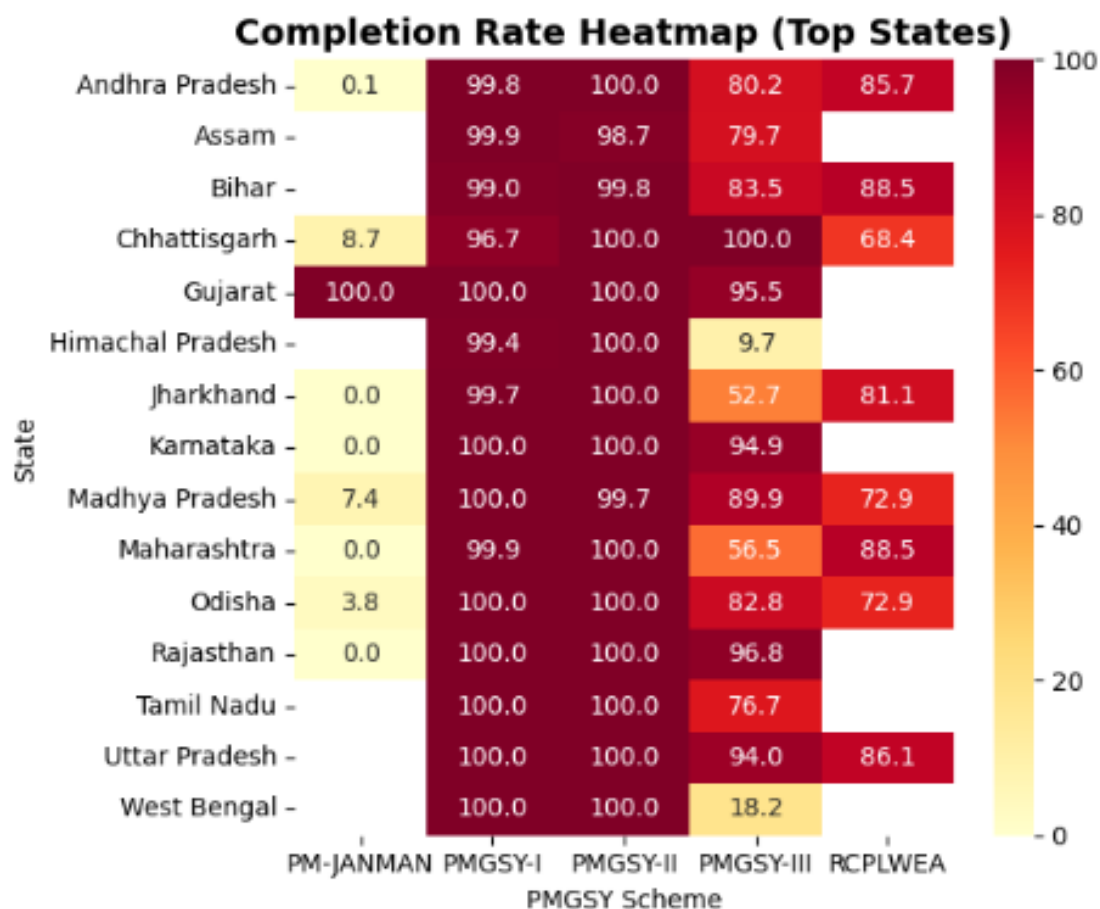

Figure 18: State-wise Performance Rankings - Completion Rates and Investment Distribution

8.12   Comprehensive Scheme Analysis and Implementation Effectiveness

Cross-scheme performance analysis reveals dramatic variations in implementation success across different PMGSY components. PMGSY-I demonstrates exceptional performance with 99.4% completion rate and substantial investment of 2,503.6 crores, establishing it as the benchmark for rural road development programs.

The scheme evolution trajectory shows concerning trends in newer implementations. While PMGSY-II maintains strong performance at 99.0% completion, PMGSY-III experiences significant decline to 73.4% despite receiving 963.3 crores investment. This pattern suggests implementation complexity increases with program evolution, requiring enhanced monitoring and adaptive management strategies. -kilometer costs.

8.13   Strategic Recommendations and Policy Implications

Based on comprehensive multivariate analysis, several critical strategic interventions emerge for optimizing PMGSY performance across future implementations. The declin-
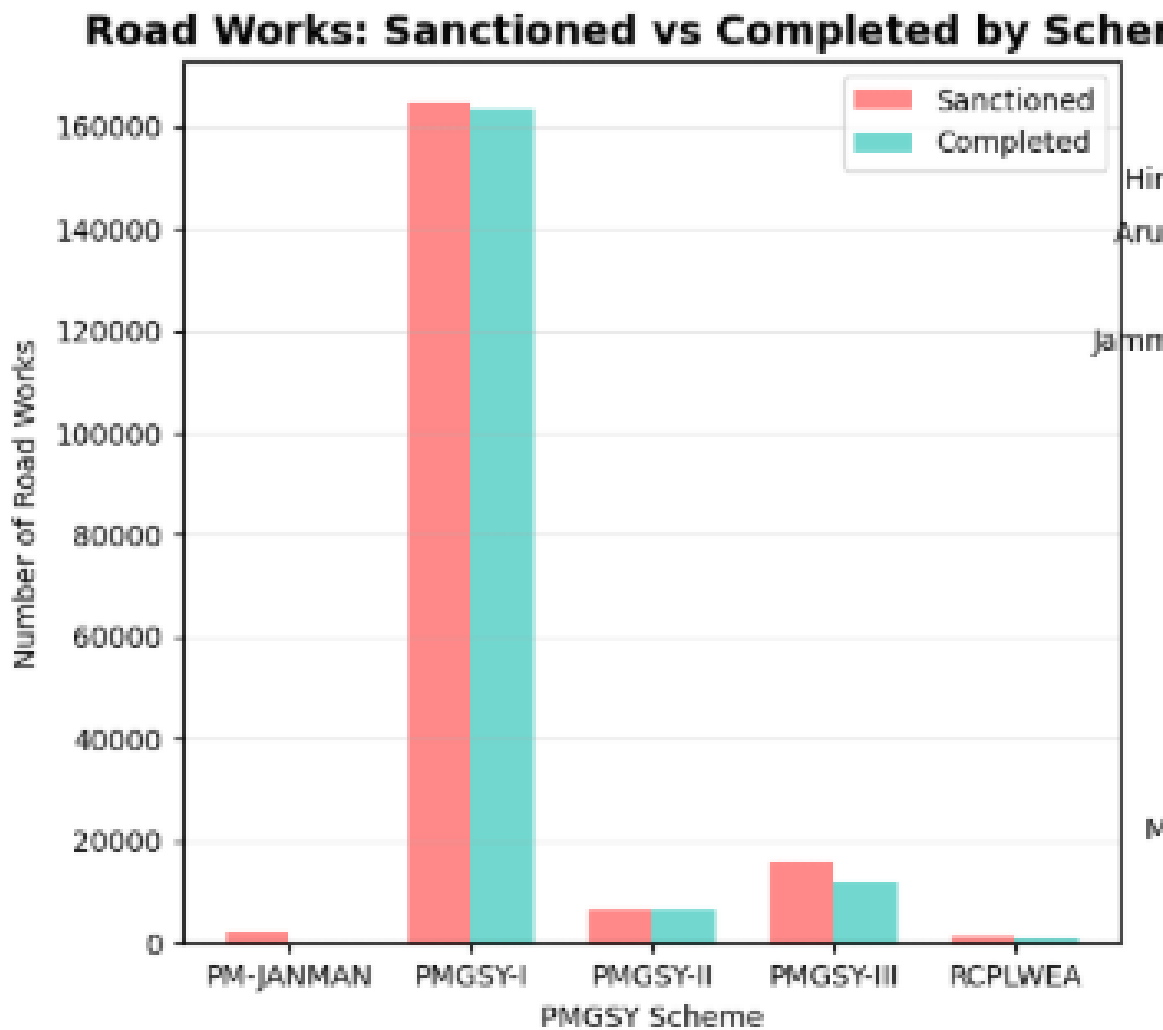
Figure 19: Comprehensive Scheme Performance Matrix - Completion Rates vs Investment Analysis

ing performance trajectory across scheme phases necessitates immediate policy attention to prevent further deterioration in program effectiveness.

Performance Optimization Strategies:

- Implementation of proven practices from high-performing states (Haryana, Rajasthan, Gujarat) across underperforming regions

- Development of specialized implementation frameworks for emerging schemes (PMGSY-III, PM-JANMAN) based on unique operational requirements

- Enhanced monitoring systems incorporating predictive analytics to identify potential project delays before completion deadlines

Investment Efficiency Improvements: Regional resource allocation optimization should prioritize states demonstrating strong administrative capabilities while providing enhanced technical support to high-investment, low-performance regions.

Future Program Design: The correlation analysis indicates that program complexity increases significantly with each successive phase, requiring simplified implementation

protocols and standardized quality assurance mechanisms. Future PMGSY phases should incorporate lessons learned from PMGSY-I's exceptional performance while addressing the systemic challenges identified in PMGSY-III implementation.

8.14 Executive Performance Dashboard and Key Metrics Analysis

The comprehensive performance dashboard reveals strong overall program effectiveness with 95.8% completion rate across 191,169 sanctioned projects, demonstrating the PMGSY program's substantial impact on rural connectivity infrastructure. The total investment of 3,957 crores with 86.2% budget utilization indicates efficient financial management within acceptable parameters.
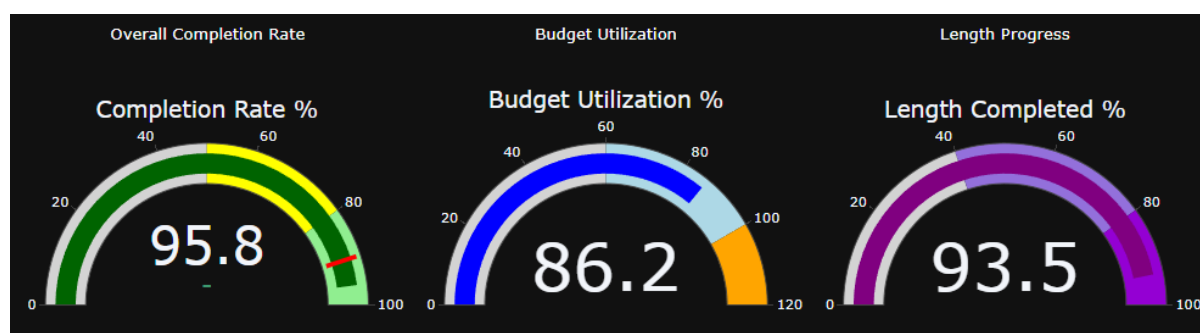

Figure 20: Executive Performance Dashboard - Key Performance Indicators Overview

Road length analysis shows impressive physical progress with 783,272 km completed out of 837,526 km sanctioned, achieving 93.5% length completion rate. The average cost per kilometer of 0.47 lakhs demonstrates cost-effective implementation across the program portfolio, though variations exist across different phases and geographical regions.

Budget utilization at 86.2% falls within optimal range, indicating balanced approach between aggressive implementation and prudent financial management.

8.15 Geographic Distribution and Risk Assessment Framework

Geographic coverage analysis reveals concentrated implementation in states with high rural connectivity needs, with Madhya Pradesh leading at 21,217 projects, followed by Rajasthan and other major states. This distribution pattern aligns with rural population density and connectivity gap priorities established in the program design phase.

Timeline progress analysis shows peak implementation during Phase 2, with systematic scaling across program duration. The bell-curve distribution indicates planned resource deployment and capacity utilization optimization, though declining activity in later phases suggests need for sustained momentum maintenance.

8.16 Strategic Implementation Framework and Future Roadmap

Based on comprehensive dashboard analysis, the strategic implementation framework emphasizes accelerated completion programs for underperforming regions while leveraging best practices from high-performing states. Gujarat's 98.5% completion rate serves as the benchmark for replication across similar geographical and administrative contexts.
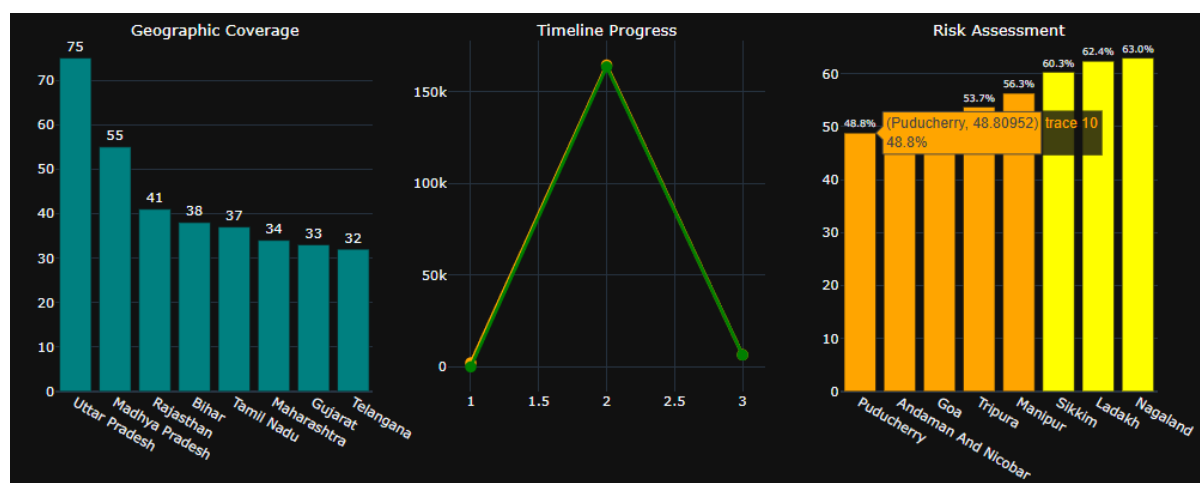
Figure 21: Geographic Distribution and Risk Assessment Matrix

Immediate Action Items:

- Deploy rapid response teams to states with completion rates below 70%

- Implement digital monitoring systems for real-time project tracking and early warning mechanisms

- Establish performance benchmarking protocols based on Gujarat and Haryana success models

- Enhance cost monitoring frameworks specifically for PMGSY-III projects showing budget variance

Medium-term Strategic Initiatives: The 86.2% budget utilization rate provides opportunity for strategic reinvestment in high-impact projects and acceleration of delayed implementations. Resource reallocation should prioritize states demonstrating administrative readiness and implementation capacity while providing technical assistance to underperforming regions.

Long-term Program Sustainability: Future PMGSY phases should incorporate lessons learned from current performance analysis, including standardized quality assurance mechanisms, technology-enabled monitoring systems, and adaptive management protocols. The average cost per kilometer of 0.47 lakhs provides baseline for future cost estimation and budget planning across diverse geographical contexts.

The strategic roadmap emphasizes maintaining the 95.8% completion rate benchmark while improving budget utilization toward 90-95% range through enhanced project management capabilities and streamlined approval processes. Success metrics indicate strong program foundation requiring targeted interventions for optimal performance across all implementation phases.