CH.LOKESH

2211CS010644

BIG DATA ANALYTICS

MINI PROJECT

RESEARCH PAPER

Title:

Scalable Analysis of Indian Panorama Dataset:

Government Analytics & Visualization Using PySpark

Abstract:

This paper presents a technical and analytical exploration of large-scale, tabular government data using the Indian Panorama dataset. PySpark was employed for scalable data processing, enabling the extraction of actionable insights through summary statistics, correlation analysis, and advanced visualization techniques. The study focuses on diverse use cases such as public surveys, census statistics, scheme beneficiary counts, and resource allocation analytics, highlighting strengths and under-served segments within regions and sectors.

---

1. Introduction:

- The proliferation of big data in government and creative sectors requires robust analytics pipelines.
- Datasets such as Indian Panorama (CPanorama.csv) represent multi-class information for organizational and policy analysis,

containing year, category, title, director, producer, and language.

---

2. Dataset & Domain Description:

- Data Source: Indian Panorama, loaded with Spark for parallel data workflows.
- Columns: Categorical (e.g., states, genres, schemes), Numerical (e.g., SrNo, population, budget allocation).
- Domain: Designed for analytics in government-led tasks including demographics, economic statistics, sector comparisons, and policy impact assessment.

---

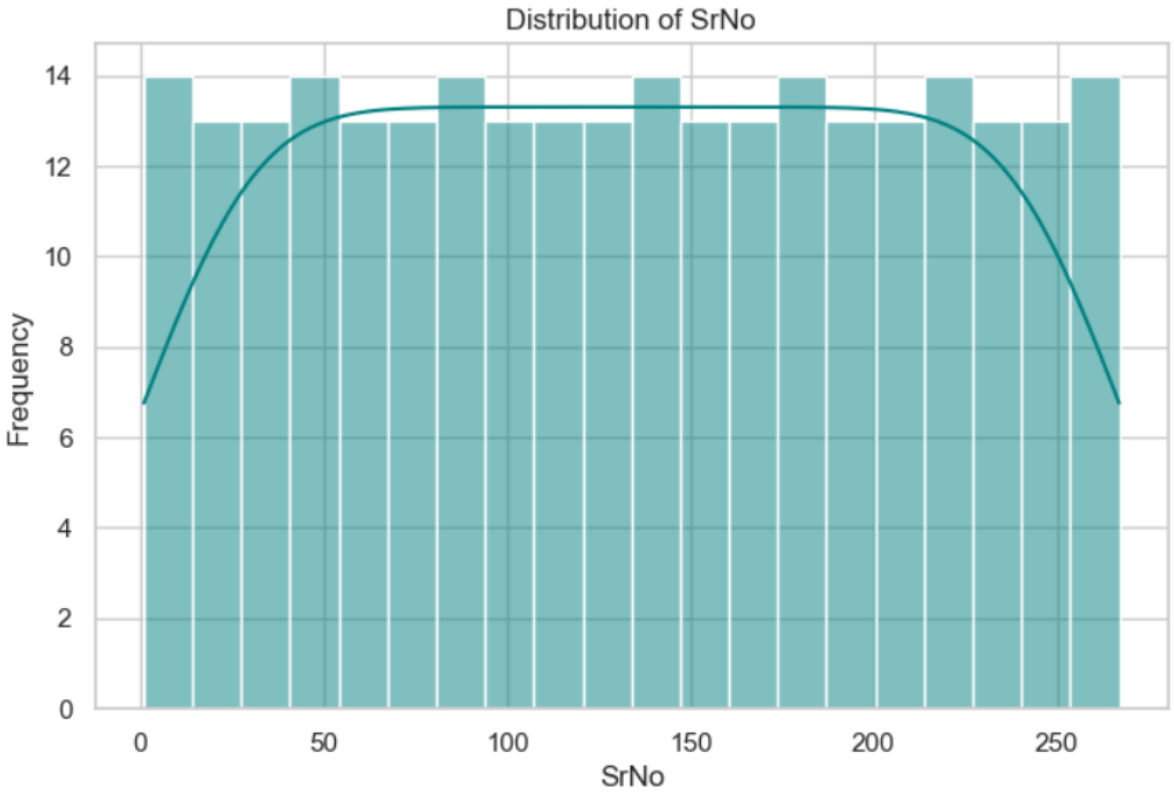
3. Data Cleaning & Preparation:

- Cleaning targeted null values and duplicates; columns renamed for consistency.
- Data types identified for effective separation of analysis strategies: categorical columns (state names/genre), numerical columns (beneficiary count/budget).

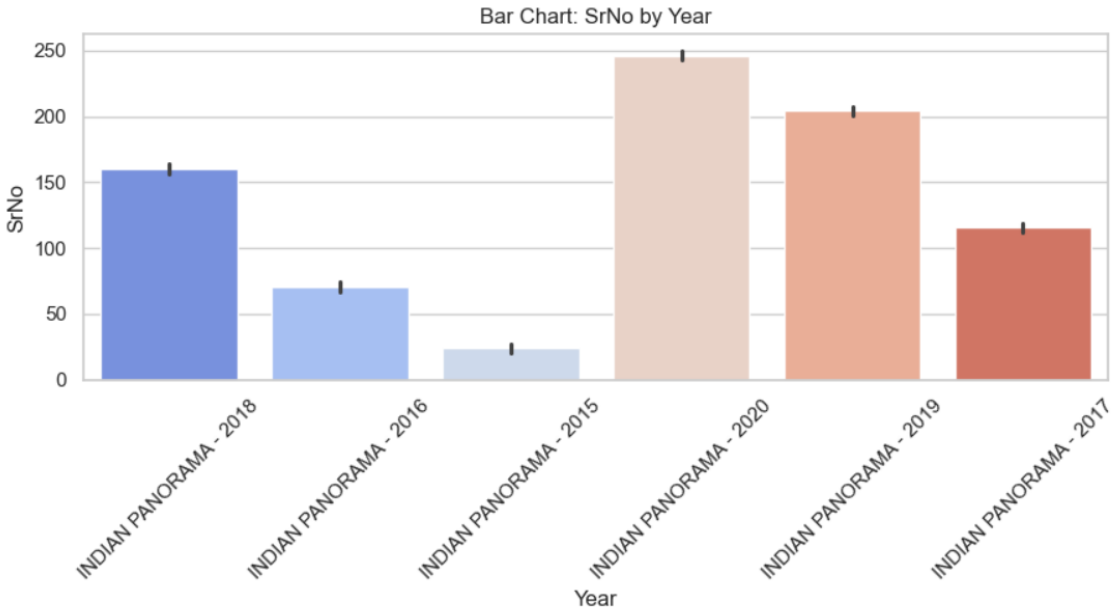
---

4. Analytical Methods:

- Summary statistics for means, minimums, maximums, standard deviations, and counts.
- Correlation analysis with heatmaps to uncover related fields (e.g., education spending vs. unemployment rates).
- Visualization gallery:

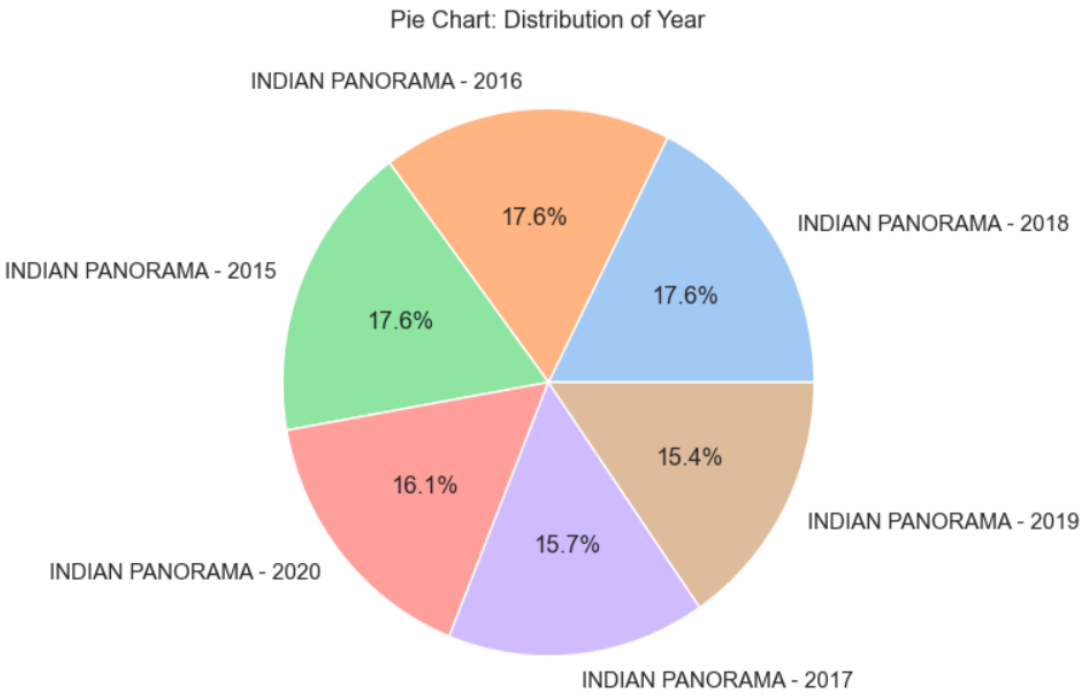


- Line Plot: Time series/growth across years or states



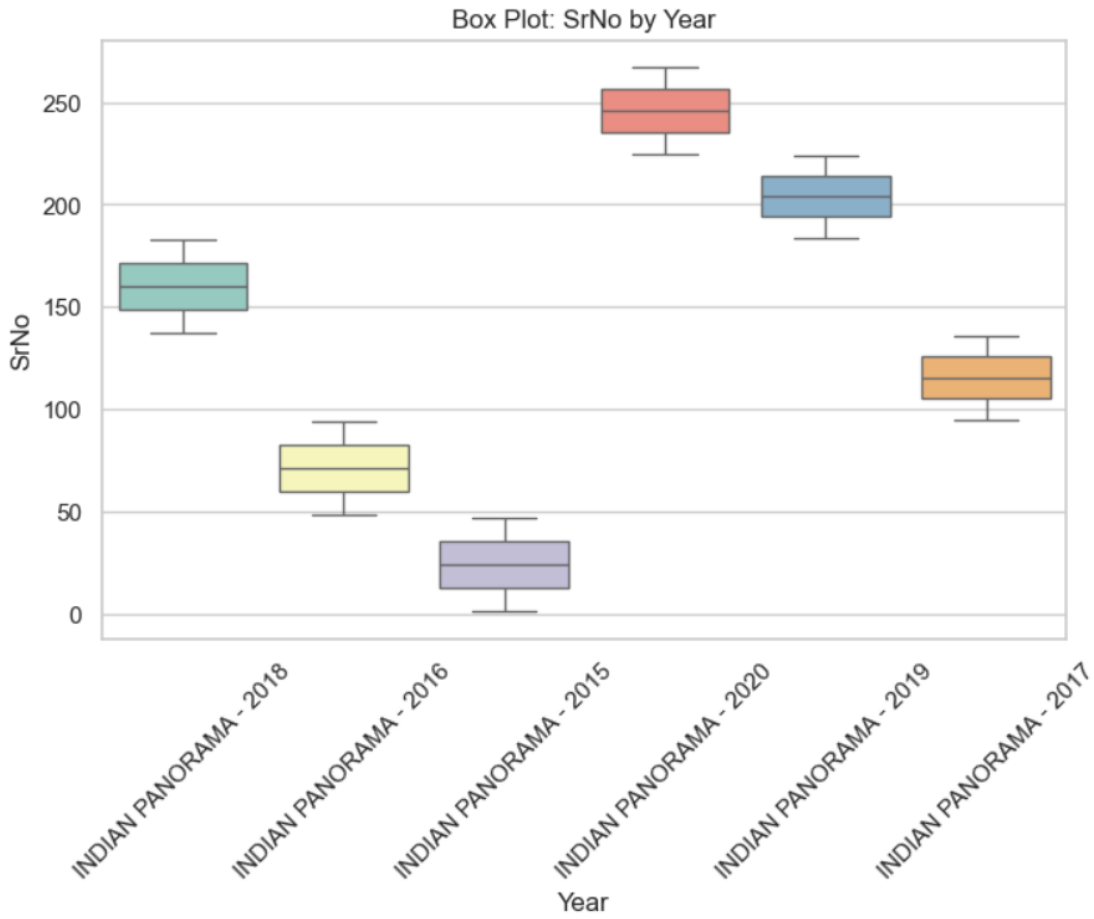
- Bar Chart: Scheme participation or genre comparison

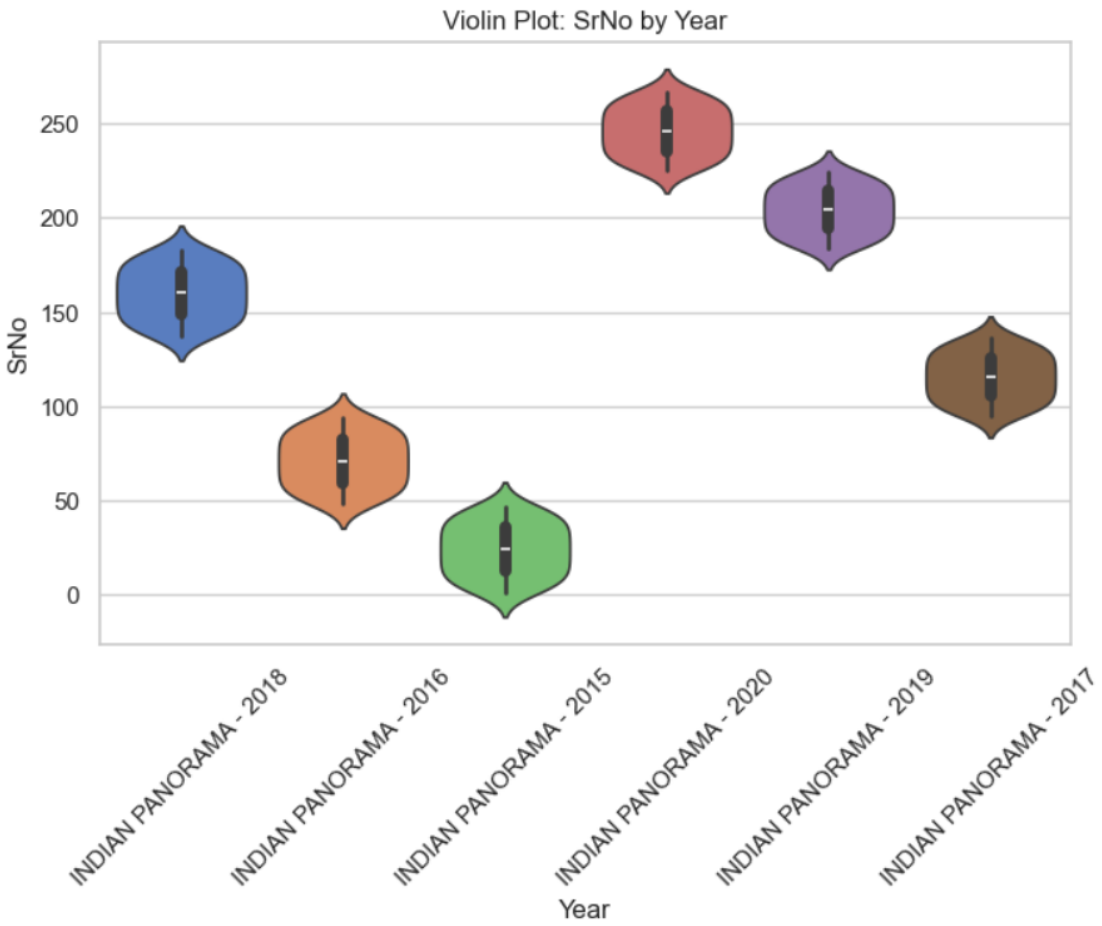
- Area Chart: Segment fill trends
    - Bubble/Scatter Plot: Outlier detection, cluster analysis
    - Pie/Violin/Box Plot: Distribution, dominance, and



categorical performance.

5. Key Findings & Hidden Insights:





- Strong correlations found—e.g., higher budget often means higher beneficiary ratios.
- Visual outlier detection highlights regions or schemes needing further intervention.
- Identification of categorical outliers, skewed distributions (under-served or over-resourced).
- Detection of outperforming states or districts irrespective of budget or size.
- Coupling of indicators (education vs. unemployment, healthcare vs. population metrics) reveals new strategic opportunities for targeted interventions.

---

6. Discussion:

- PySpark enables scalable, near-real-time analysis for large datasets, making it suitable for government use.
- Visualization tools provide intuitive dashboards for rapid stakeholder assessment and action.
- Analytical workflow can be generalized for other domains with similar multi-class, multi-column datasets.

---

7. Recommendations:

- Targeted Policy Interventions: Allocate additional resources where outliers or lagging segments are revealed.
- Resource Allocation Optimization: Utilize observed correlations to maximize sector returns.
- Strategic Monitoring: Deploy dashboards for continual scheme assessment and rapid issue detection.
- Continuous Data Quality Improvement: Standardize cleaning/integration procedures.
- Data-Driven Policy Making: Foster automated analytics across ministries, grounded in observed trends, not assumptions.

---

8. Conclusion:

By leveraging Spark-based analytics and Python visualization, this research demonstrates how complex, high-dimensional data from public sectors can be transformed into actionable intelligence. The Indian Panorama analysis provides a model for future government and large-scale organizational analytics, enabling improved governance through data-driven decision making.

---

9. References:

- Code, workflow, and examples cited from PANORAMA.ipynb & REPORT.pdf.

---

10. Appendix:

- Characteristic plots/bar charts/heatmaps illustrated in notebook, showing trends, distributions, and correlations.
- Data snippets, reporting outputs, and recommendations extracted from attached report file for presentation and strategic planning