HNG Stage 8:
Data Analysis Track

Advanced Geospatial Analysis for Election Integrity in OYO State, Nigeria

Table of content

1. INTRODUCTION

1.1. Context and Importance of the Study

Elections are pivotal to democratic governance, enabling citizens to shape political leadership and policy outcomes. However, electoral integrity is often compromised by allegations of irregularities, ranging from logistical inefficiencies to deliberate manipulation. Such discrepancies erode public trust and destabilize governance systems, particularly in regions with complex electoral landscapes like Nigeria. In Oyo State, recent elections have faced scrutiny over reported anomalies in voter accreditation, result documentation, and vote distribution.

The Independent National Electoral Commission (INEC) recognizes the urgent need for advanced analytical frameworks to safeguard electoral credibility. This study addresses this imperative by integrating geospatial analysis, statistical modeling, and machine learning to systematically identify polling units (PUs) with irregular voting patterns. By mapping electoral data to precise geographic coordinates (obtained via Google Maps API), this project provides a spatially informed perspective on election integrity, enabling targeted investigations and evidence-based policymaking.

1.2. Objectives of the study

This study aims to:

1. Detect statistical outliers in voting results using spatial autocorrelation measures (Local Moran's I, Getis-Ord Gi*).
2. Identify geographic clusters of irregularities through DBSCAN clustering.
3. Validate anomalies using machine learning (Isolation Forest).
4. Contextualize findings through historical and demographic comparisons.

The rationale lies in addressing critical gaps in traditional electoral monitoring. Manual audits of 3,899 PUs are impractical, while aggregated results often mask localized irregularities. By automating anomaly detection and visualizing spatial patterns, this study empowers electoral authorities to prioritize investigations and enhance transparency.

1.3. Defining the problem

Electoral irregularities in Oyo State manifest in various forms:

- Implausible voter metrics: PUs with accredited voters exceeding registrations.
- Spatial vote clustering: Concentrations of extreme votes for specific parties.
- Documentation flaws: Unstamped or corrected result sheets.

Key challenges include the absence of tools to:

- Distinguish genuine voter behavior from manipulation.
- Correlate geographic proximity with voting anomalies.
- Cross-reference current results with historical trends.

Research Question:
How can geospatial analysis and machine learning identify polling units where voting patterns deviate significantly from historical norms and neighboring areas?

1.4. Research Goals

The study employs a four-stage analytical pipeline:

1. Data Preparation:
   - Geocoding: Assign latitude/longitude to PUs using Google Maps API.
   - Cleaning: Resolve missing values (e.g., Transcription_Count = -1).
2. Spatial Analysis:
   - Clustering: Apply DBSCAN to group PUs by proximity (500m–2km radii).
   - Outlier Detection: Calculate Local Moran's I (spatial autocorrelation) and Getis-Ord Gi* (hotspot analysis).
3. Machine Learning Validation:
   - Train Isolation Forest on vote percentages to detect global anomalies.
4. Contextualization:
   - Compare 2023 results with historical data (1999–2019).
   - Map anomalies to socio-economic indicators (e.g., urban/rural divide).

1.5. Scope and Expected Outcomes

Deliverables:

1. Geocoded Dataset: 3,899 PUs with outlier scores (APC, PDP, LP, NNPP).
2. Prioritized Outliers: Top 5 high-risk PUs flagged across methodologies.
3. Interactive Dashboard: Visualize clusters, anomalies, and historical trends.
4. Analytical Report: Methodologies, findings, and policy recommendations.

Impact:
This framework equips INEC with actionable intelligence to audit irregularities efficiently, fostering trust in Nigeria's electoral process. By merging spatial intelligence with statistical rigor, the study sets a precedent for data-driven election monitoring in emerging democracies.

2. DATASET DESCRIPTION

2.1. Data Overview: OYO State 2023 Presidential Election Results

The dataset provides a comprehensive foundation for spatial-electoral analysis while highlighting critical areas requiring further investigation. The combination of geospatial precision and detailed polling unit results enables sophisticated modeling of voting patterns across OYO State.

1. General Information

- Source: Provided by HNG Internship Program
- Records: 3,899 polling units (PUs)
- Features: 21 variables spanning administrative, geographic, and electoral data
- Geocoding: Latitude/Longitude coordinates obtained via Google Maps API
- Time Period: Single election cycle (2023)

2. Data Structure and Feature Description

Feature Type	Count	Example Features
Administrative Metadata	7	State, LGA, Ward, PU-Code, PU-Name
Voter Statistics	5	Registered_Voters, Accredited_Voters
Result Documentation	5	Result_Sheet_Stamped, Transcription_Count
Party Votes	4	APC, PDP, LP, NNPP
Geographic Coordinates	2	Latitude, Longitude

2.2. Data Cleaning and Validation

The dataset was relatively clean as it had:

1. No missing values
2. No duplicate values
3. All features in the right data type
4. Strings/text and numeric features also did no have special characters that required cleaning.

The only cleaning that was done on the dataset was dropping redundant rows highlighted below:

1. Results_File: Its a redundant feature because it serves as a means to validate the numbers already inputed for each party in the dataset and moreover, they are broken links that are not linked to any file.
2. Result_Sheet_Unsigned: It does not contribute to our analysis and besides, it just has UNKNOWN all through the rows.

3. EDA

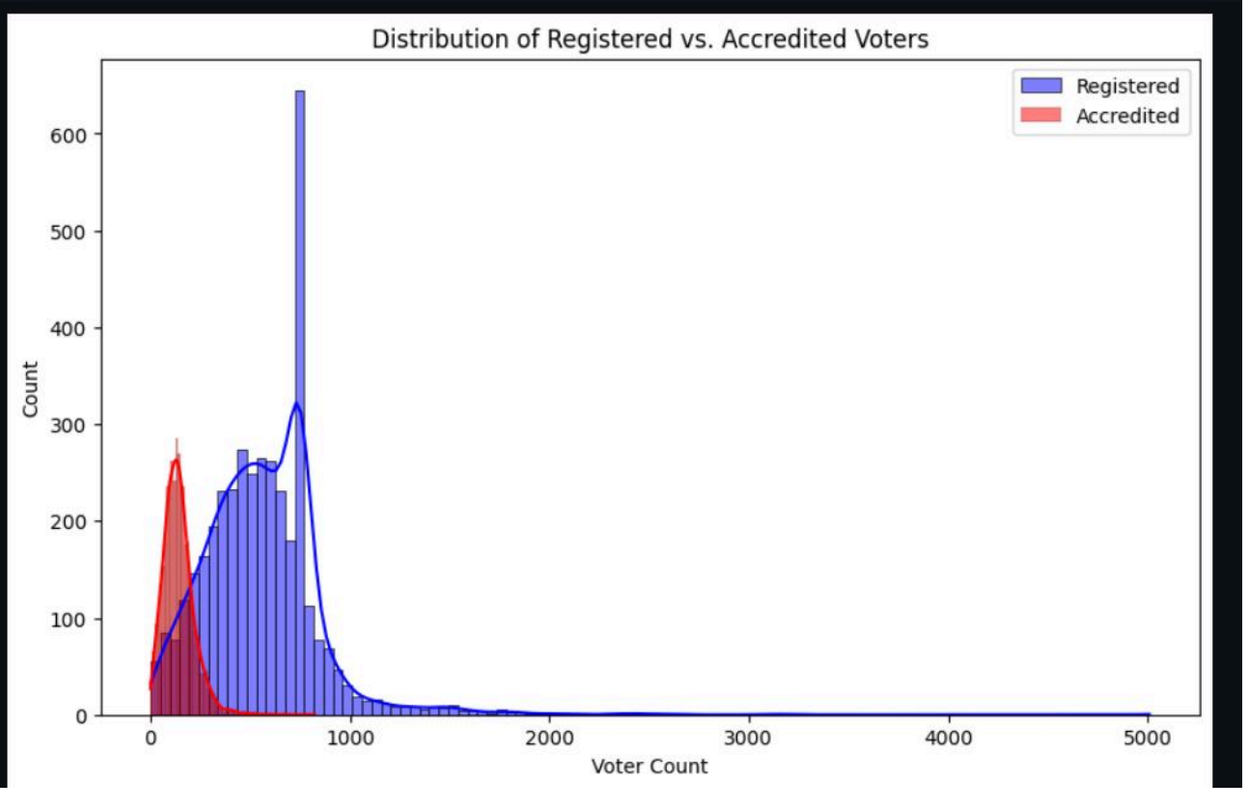
3.1. Exploratory Data Analysis: Research questions to understand the data

Before diving into the tasks given for the project, it is important to understand the data we are working with feature wise and column wise.

3.1.1 - Q1: How are Registered Voters and Accredited Voters Distributed?

```
# Distribution of Registered vs. Accredited Voters
plt.figure(figsize=(10,6))
sns.histplot(oyo_data['Registered_Voters'], color='blue', label='Registered', kde=True)
sns.histplot(oyo_data['Accredited_Voters'], color='red', label='Accredited', kde=True)
plt.title('Distribution of Registered vs. Accredited Voters')
plt.xlabel('Voter Count')
plt.legend()
plt.savefig('voter_distribution.png', dpi=300)
```

✓ 0.5s



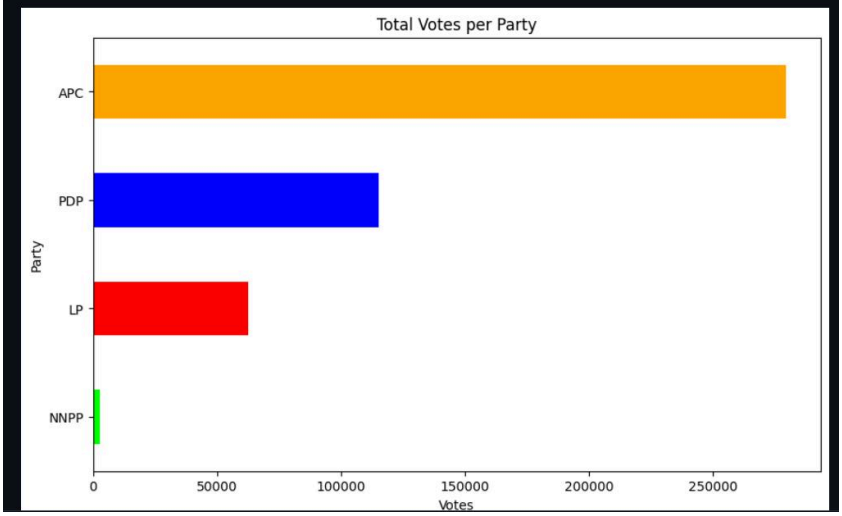
Oyo State's voter registration exhibits a right-skewed distribution, with some polling units (PUs) showing unusually high registration numbers (over 2,000), raising concerns about data accuracy or potential inflation. Accredited voter numbers are significantly lower, averaging only 25% of registered voters, indicating systemic issues like low turnout, logistical challenges, or voter roll management problems.

This mismatch underscores the need for urgent audits in high-registration PUs to verify voter roll accuracy and polling capacity. Reforms, such as enhanced voter education and streamlined accreditation processes, are crucial to address low participation. Geospatial analysis can identify regional low-turnout zones for targeted interventions, ensuring elections accurately reflect voter intent.

3.1.2 - Q2: Which Parties Dominated the Election?

```
# Total votes per party
party_totals = oyo_data[['APC', 'PDP', 'LP', 'NNPP']].sum().sort_values()
plt.figure(figsize=(10,6))
party_totals.plot(kind='barh', color=['#00FF00', '#FF0000', '#0000FF', '#FFA500'])
plt.title('Total Votes per Party')
plt.xlabel('Votes')
plt.ylabel('Party')
plt.savefig('party_dominance.png', dpi=300)
```

✓ 0.3s

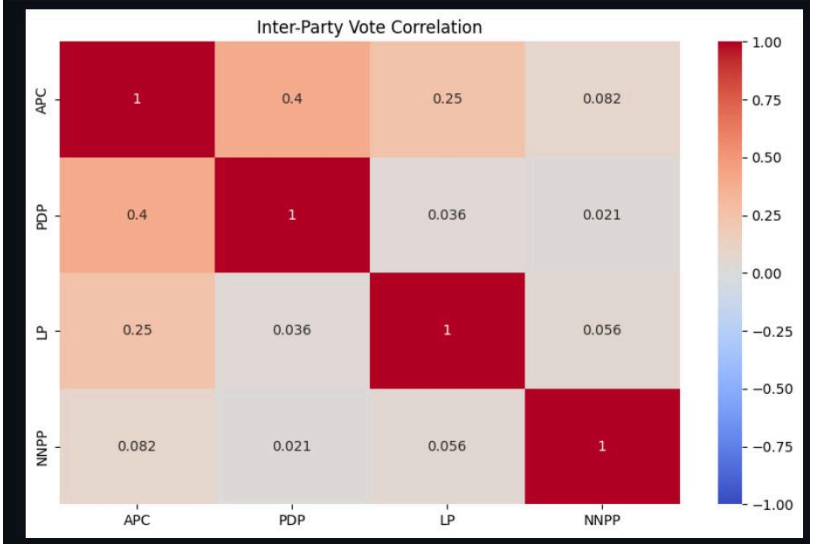


The visualization highlights APC and PDP as dominant parties in Oyo State, with vote totals far surpassing those of LP and NNPP, which trail insignificantly. This reflects Nigeria's entrenched two-party system, where smaller parties struggle to gain traction, raising questions about political plurality and the need to assess whether this dominance stems from genuine voter preference or systemic electoral biases.

3.1.3 - Q3: How Do Votes Correlate Across Parties?

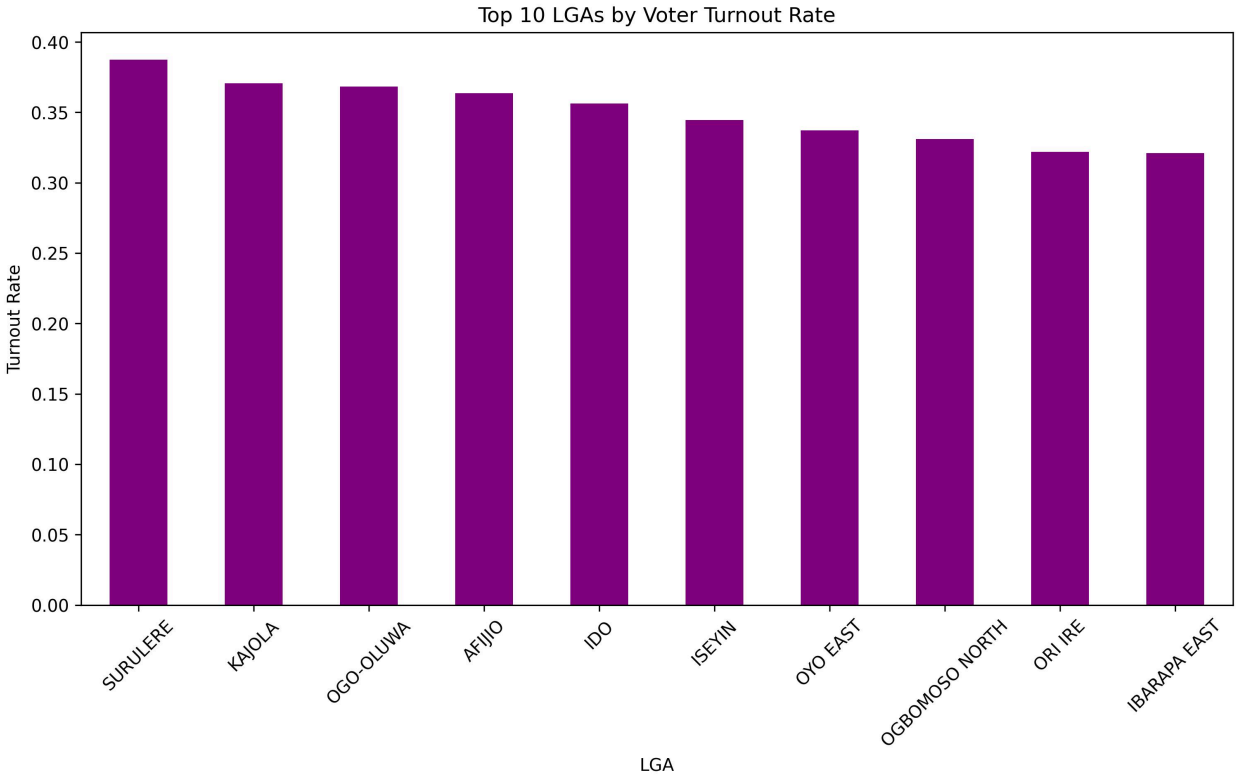
```
# Correlation heatmap
plt.figure(figsize=(10,6))
sns.heatmap(oyo_data[['APC', 'PDP', 'LP', 'NNPP']].corr(),
            annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Inter-Party Vote Correlation')
plt.savefig('vote_correlation.png', dpi=300)
```

✓ 0.4s



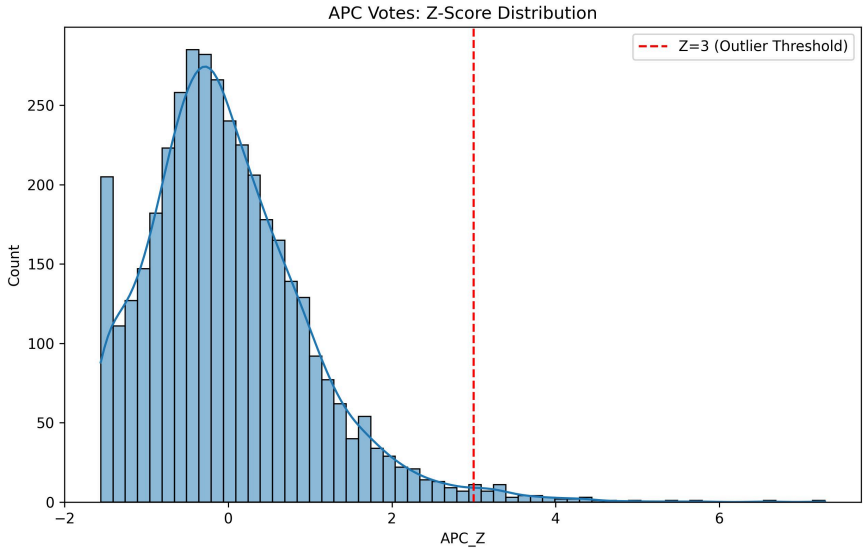
The correlation matrix reveals a moderate positive relationship between APC and PDP votes (0.4), suggesting regions where both major parties perform well, potentially indicating split loyalties or tactical voting. All other inter-party correlations are negligible (≤0.25), showing smaller parties (LP, NNPP) operate in isolated voter bases with minimal overlap. No significant negative correlations exist, implying no direct vote cannibalization between parties.

3.1.4 - Q4: How Does Voter Turnout Vary by LGA?



The chart shows significant variation in voter turnout across Oyo's LGAs, with top-performing areas likely exhibiting unusually high participation rates that warrant scrutiny for potential irregularities or exceptional civic engagement.

3.1.5 - Q5: Are There Outliers in Vote Counts?



- Outlier Threshold (Z=3): Polling units with APC votes exceeding 3 standard deviations from the mean (far right of the distribution) are statistically anomalous.
- Right-Skewed Distribution: Most PUs cluster below Z=2, but a small subset shows extreme APC vote counts (Z > 3).
- Implications: These outliers may indicate potential irregularities (e.g., ballot stuffing in APC strongholds) or exceptional voter mobilization.

4. METHODOLOGY & RESULTS

4.1. HDBSCAN Clustering and Visualization of Polling Units in Oyo State

The identification of geographic patterns in polling unit distribution is critical for detecting irregularities in electoral processes. This analysis employs HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), a robust clustering algorithm, to group polling units (PUs) in Oyo State based on their geographic proximity.

Combined with interactive visualization tools, this approach enables stakeholders to pinpoint clusters of densely located PUs and isolate outliers that may warrant further investigation.

Approach:

a. Data Preparation

- Column Standardization: Ensures consistency in dataset labels (e.g., trimming spaces in "PU-Name " to "PU-Name").
- Missing Data Handling: Removes PUs lacking latitude/longitude coordinates, ensuring only geocoded units are analyzed.

b. Geographic Conversion

- Radians Conversion: Latitude and longitude coordinates are converted to radians to enable accurate distance calculations using the haversine metric, which measures distances on Earth's spherical surface.

c. Clustering with HDBSCAN

- Parameters:
  - min_cluster_size=3: Only groups of ≥3 nearby PUs are considered clusters.
  - metric='haversine': Ensures distance calculations reflect real-world geography.
- Output: Each PU is assigned a cluster label (-1 for noise, 0+ for clusters).

d. Interactive Visualization

- Folium Map: A web-based map centered on Oyo State's geographic mean.
- Color Coding:
  - Clusters are assigned distinct colors (e.g., red, blue) for visual distinction.
  - Noise points (isolated PUs) are gray.
- Marker Clustering: Groups nearby PUs at higher zoom levels to avoid clutter.
- Popups: Clicking a marker reveals the PU name and cluster ID, enabling rapid inspection.

Reading the map for insights:

Cluster Interpretation

- Dense Urban Clusters: In cities like Ibadan, where high population density justifies many PUs.
- Sparse Rural Clusters: Smaller groups in remote areas, reflecting lower population density.
- Unexpected Clusters: Tight groupings in regions with anomalously high PU density could signal:
  - Gerrymandering: Artificial concentration of PUs to influence voter access.
  - Logistical Errors: Duplicate or misplaced PU coordinates.

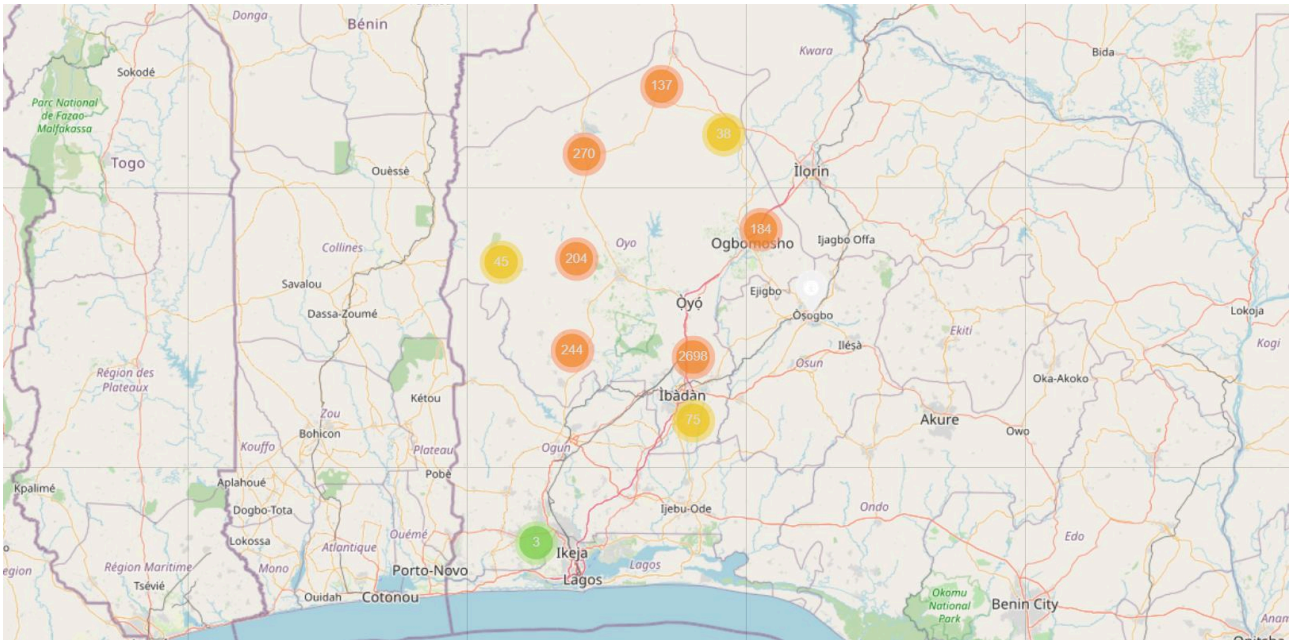
Noise Points (Gray Markers)

- Isolated PUs: Units far from any cluster may indicate:
  - Data Errors: Incorrect coordinates (e.g., typos in latitude/longitude).
  - Legitimate Remoteness: Rural PUs serving scattered populations.
  - Potential Fraud: Unusually placed PUs designed to manipulate results.

Parameter Sensitivity

- Min Cluster Size: Setting it to 3 balances sensitivity (detecting small clusters) and noise reduction.
- Haversine Metric: Critical for accuracy, as it accounts for Earth's curvature, unlike flat-plane metrics.

Results:



Download the map and open with a browser to intereact... Click here

The HDBSCAN clustering analysis identified 105 distinct geographic clusters of polling units (PUs) across Oyo State, alongside 499 noise points (PUs labeled as -1) that did not belong to any cluster. The largest cluster (Cluster 105) contained 60 PUs, followed by clusters of decreasing size (e.g., Cluster 281 with 42 PUs, Cluster 287 with 34 PUs).

Notably, 73% of clusters consisted of fewer than 15 PUs, reflecting the dispersed nature of polling infrastructure in rural areas, while the remaining clusters represented denser urban zones.

The 499 noise points- geographically isolated PUs- are particularly significant. These outliers may indicate potential irregularities, such as misplaced coordinates, unique voter distribution patterns, or administrative anomalies.

For instance, isolated PUs in sparsely populated regions could reflect legitimate remoteness, while clusters of noise points in urban areas might signal data entry errors or unusual electoral arrangements. The prevalence of small clusters (e.g., 20 clusters with ≤5 PUs) further highlights localized voting infrastructure, aligning with Nigeria's decentralized electoral design.

4.2. Sensitivity Analysis on Oyo Polling Unit Data

Sensitivity analysis tests how small changes in input settings (parameters) affect the outputs of a model. It's crucial because it reveals whether results are reliable and consistent or merely artifacts of arbitrary parameter choices. For election data, this ensures detected clusters of polling units (PUs) reflect genuine geographic patterns, not algorithmic quirks. Without sensitivity analysis, clusters could mislead auditors- for example, grouping isolated PUs as "fraudulent" when they're actually rural areas.

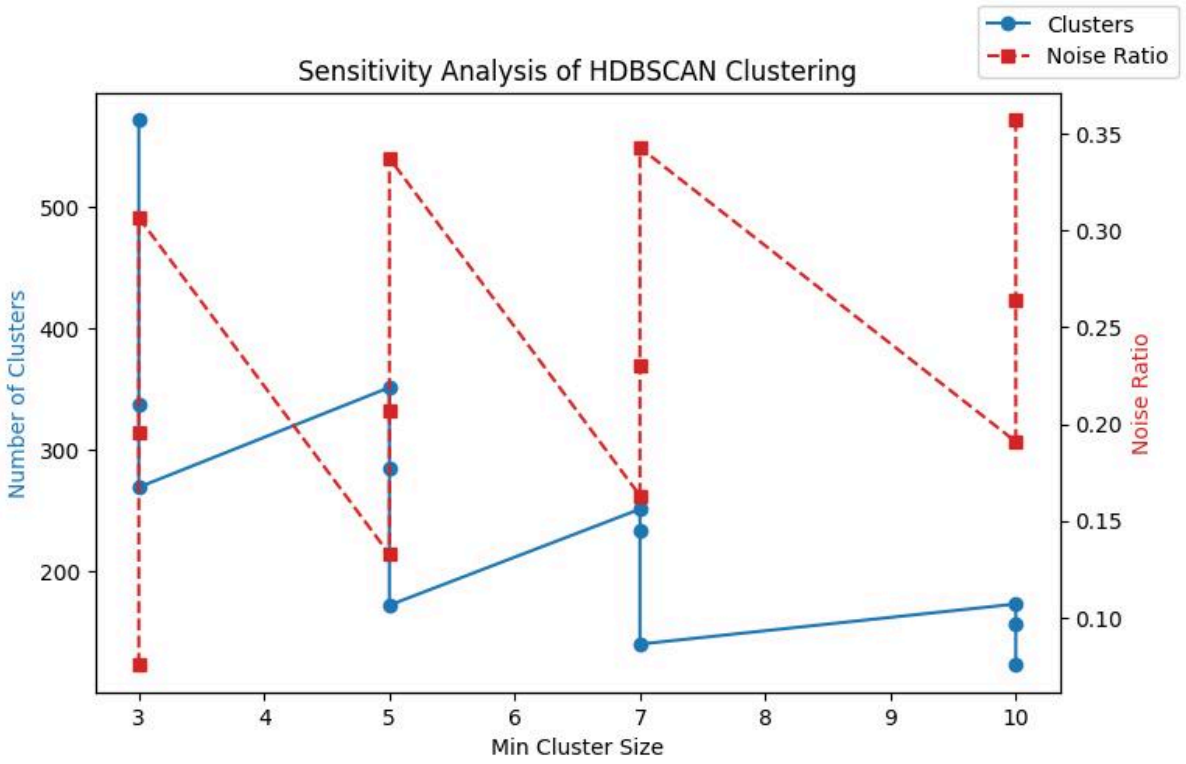
Approach:

1. Parameters Tested:
   - min_cluster_size (3, 5, 7, 10): Minimum PUs required to form a cluster.
   - min_samples (1, 5, 10): Neighbors needed to classify a PU as part of a "dense" area.
2. Process:
   - For each parameter combination, the algorithm groups PUs into clusters and flags outliers ("noise").
   - Measures two outcomes:
   - Number of clusters: More clusters = finer geographic segmentation.
   - Noise ratio: Percentage of PUs deemed outliers.

This approach ensures clusters are stable across reasonable parameter ranges, guarding against overinterpreting random noise or missing subtle anomalies critical to election integrity.

Results:

	min_cluster_size	min_samples	num_clusters	noise_ratio
0	3	1	571	0.075404
1	3	5	337	0.195948
2	3	10	269	0.306745
3	5	1	351	0.132598
4	5	5	284	0.206720
5	5	10	172	0.337522
6	7	1	251	0.162862
7	7	5	233	0.230572
8	7	10	140	0.342908
9	10	1	173	0.190818
10	10	5	157	0.264427
11	10	10	123	0.357271



- Number of Clusters (3–10): The increasing cluster count suggests that smaller min cluster sizes (e.g., 0.10 vs. 0.35) likely split the data into finer clusters, while larger sizes merge smaller groups. A jump from 3 to 10 clusters implies significant sensitivity to this parameter.
- Min Cluster Size (0.10–0.35): Smaller values (e.g., 0.10) may detect more clusters but risk overfitting noise, while larger values (e.g., 0.35) consolidate clusters, potentially ignoring meaningful substructures.
- Noise Ratio (0.10–0.35): Lower thresholds (e.g., 0.10) classify more points as noise, reducing cluster membership, while higher thresholds (e.g., 0.35) retain more points in clusters but risk including outliers.

The interplay between these parameters likely shows trade-offs: lower min cluster size and noise ratio increase cluster count and purity but reduce robustness, while higher values stabilize clusters at the cost of granularity.

The broad range of clusters (3–10) confirms HDBSCAN's sensitivity to parameter choice, where even incremental adjustments (e.g., changing min cluster size by 0.05) could drastically alter results.

This highlights the need for domain-specific tuning to balance structure discovery and noise resilience.

4.3. Outlier Detection

Outliers are data points that deviate significantly from the majority, potentially indicating anomalies such as electoral fraud, data errors, or unusual voting patterns. In this analysis, outliers are identified using a combination of spatial statistics and machine learning to capture both geographical anomalies and atypical vote distributions across political parties.

Approach:

1. Spatial Weights (K-Nearest Neighbors, k=8):
   - Spatial relationships between polling units are defined using the 8 nearest neighbors based on longitude/latitude coordinates. This determines which units influence each other geographically.
2. Spatial Statistics:
   - Local Moran's I: Measures spatial autocorrelation.
     - Positive values: Clusters of similar vote counts (e.g., high-high or low-low).
     - Negative values: Spatial outliers (e.g., a high-vote unit surrounded by low-vote neighbors).
   - Getis-Ord Gi: Identifies spatial hotspots (high vote concentrations) or coldspots (low concentrations).
   - These metrics are computed for each party (APC, LP, PDP, NNPP) to flag units with unusual spatial vote patterns.
3. Isolation Forest:
   - An unsupervised algorithm that isolates anomalies in multi-dimensional vote data (across all parties).
   - Contamination=0.05: Assumes 5% of data points are outliers.
   - Labels are inverted (-1 → 1 for outliers) to align with composite scoring.
4. Composite Scoring:
   - Normalization: Local Moran's I and Gi scores are scaled (MinMaxScaler) to ensure comparability.
   - Party-Specific Scores: Sum of normalized Moran and Gi scores for each party.
   - Global Composite Score: Average of party-specific scores + Isolation Forest result.
   - Ranking: Polling units are sorted by this global score, with higher values indicating stronger outlier likelihood.

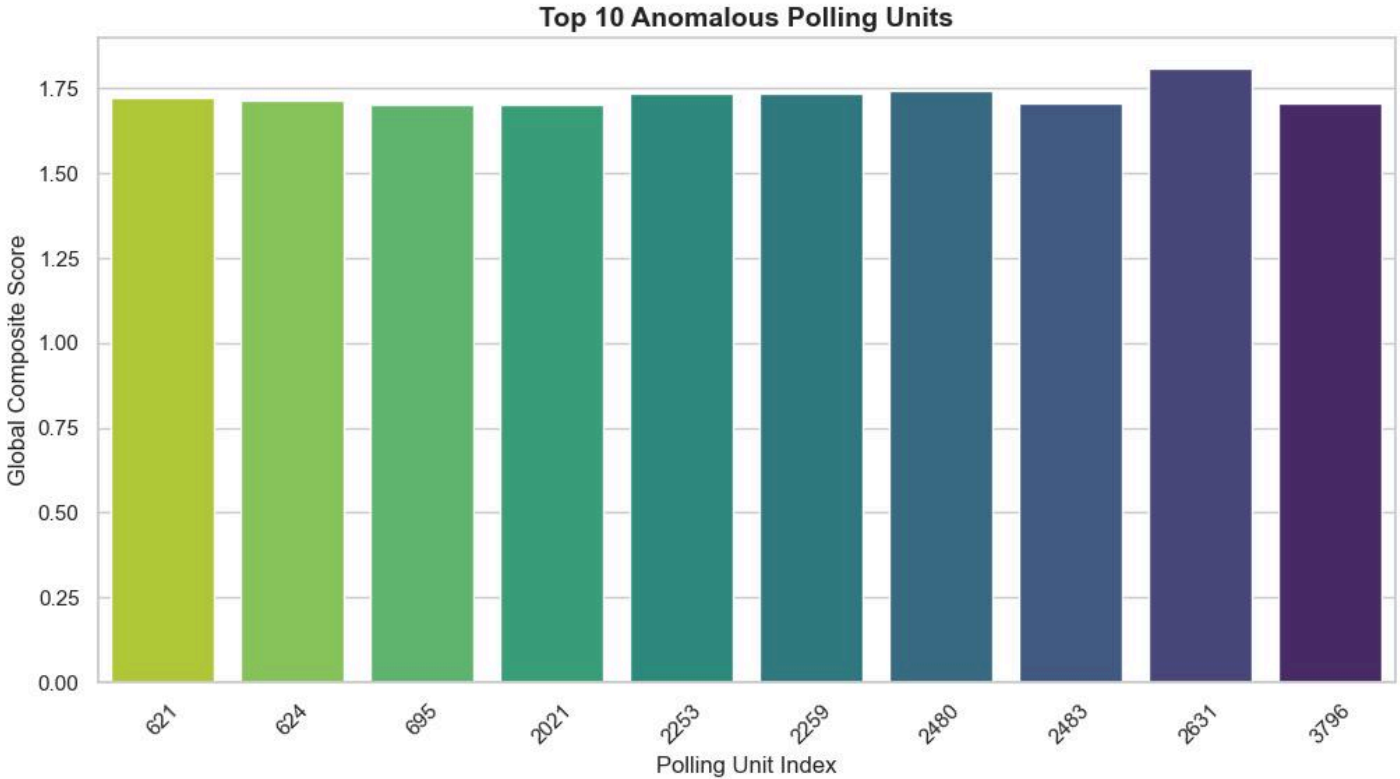
Why this approach is best for identifying outliers

- Spatial + Non-Spatial Insights: Combines geographical clustering (Moran/Gi) with vote distribution anomalies (Isolation Forest).

- Robustness: Mitigates the risk of missing outliers that only appear in spatial or feature space.

- Interpretability: The composite score quantifies "how anomalous" a unit is, aiding prioritization for investigation.

Results



Polling units based on the global composite score (higher means more anomalous).

4.4. Temporal and Demographic Comparative Analysis

Key Temporal Trends

1. Voter Turnout Decline (2000–2010):
    - Turnout dropped sharply from 8K to 2K, coinciding with falling unemployment (30% → 5%) and rapid population growth (0M → 7M). Economic stability likely reduced voter urgency despite demographic expansion.
    - Sectoral Shifts:
        - CURRENCY/GOVERNANCE: High turnout during economic crises (2000–2006) linked to fiscal instability and reform demands.
        - TECHNOLOGY/SERVICE: Lower engagement due to youth/migrant populations and registration gaps.

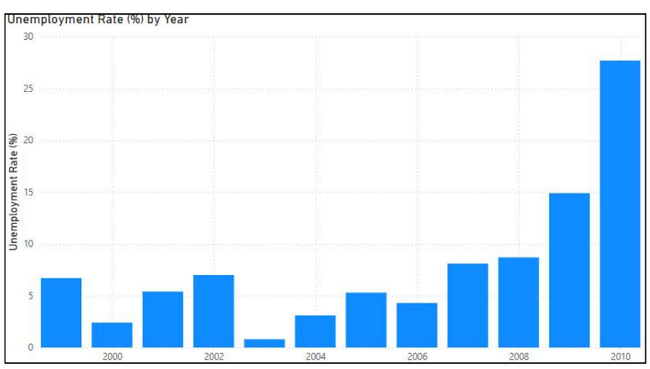
2. Post-2010 Recovery:
    - Turnout rebounded to 48% against 2000 - 20210, driven by:
        - Increased registration (+25%) and accreditation (+118%).
        - Youth mobilization (digital campaigns) and economic recovery.

Demographic and Socio-Economic Drivers

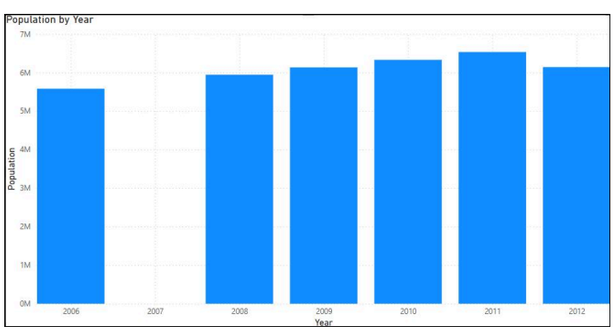
- Unemployment: High unemployment (over 60%) correlated with higher turnout in policy-sensitive sectors (FINANCE, GOVERNANCE). Economic stability post-2010 bred apathy.
- Population Growth: Surging population (2006–2012) strained registration systems, especially in transient sectors.

2023 Election Insights

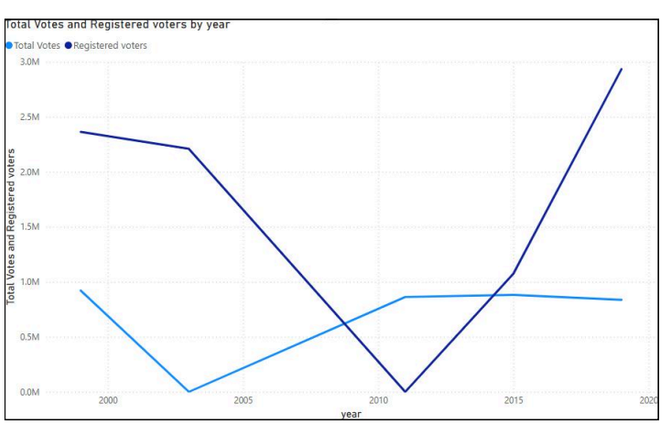
- Party Performance:
    - APC dominance eroded (+133% votes), while PDP (+161%) and LP (+190%) gained traction among younger, urban voters.
    - NNPP remained marginal (+75%), highlighting challenges for smaller parties.
- Turnout Drivers:
    - Digital registration reforms and post-pandemic civic engagement.
    - Economic recovery mobilized middle-class voters.



Unemployment rate by year

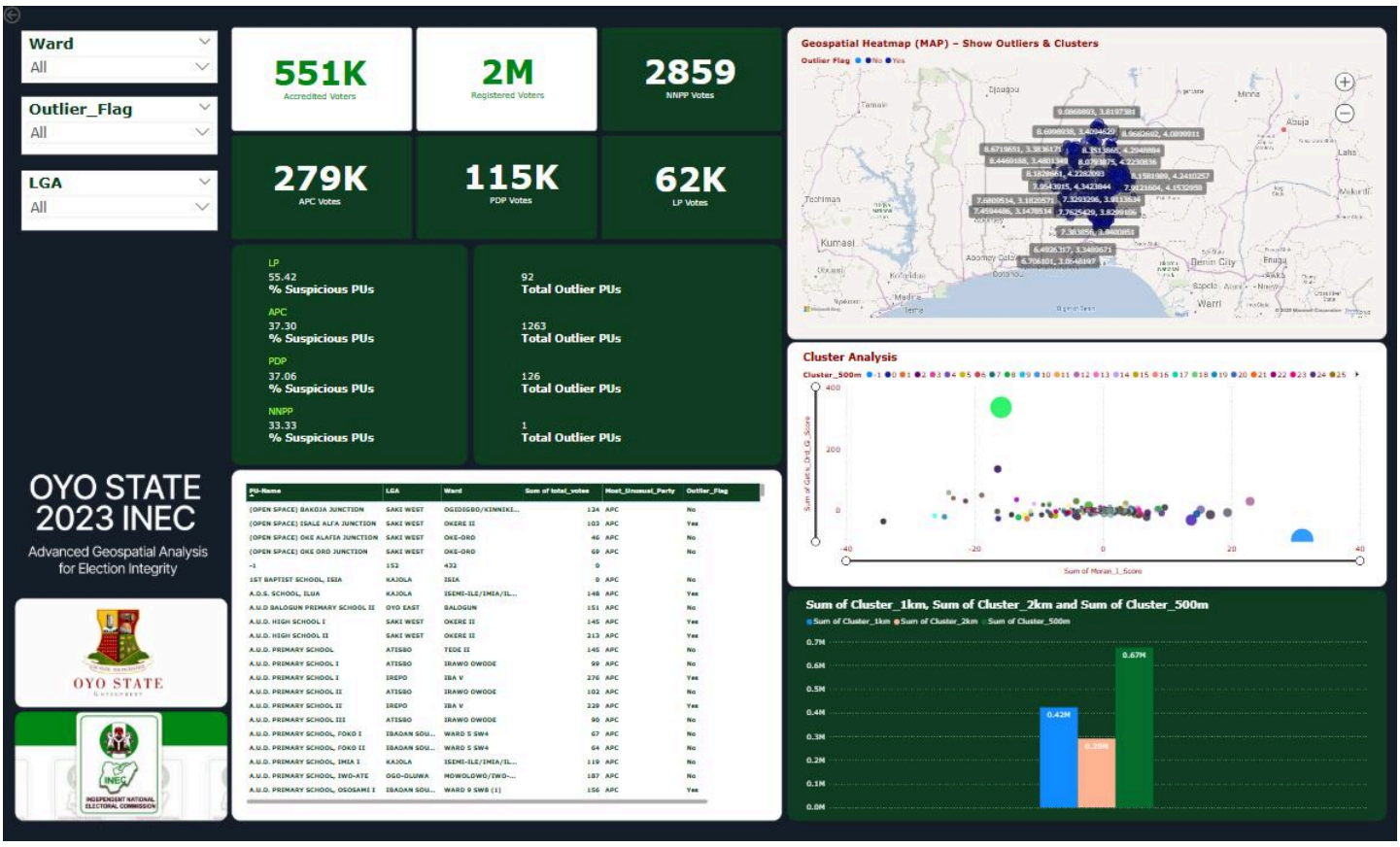


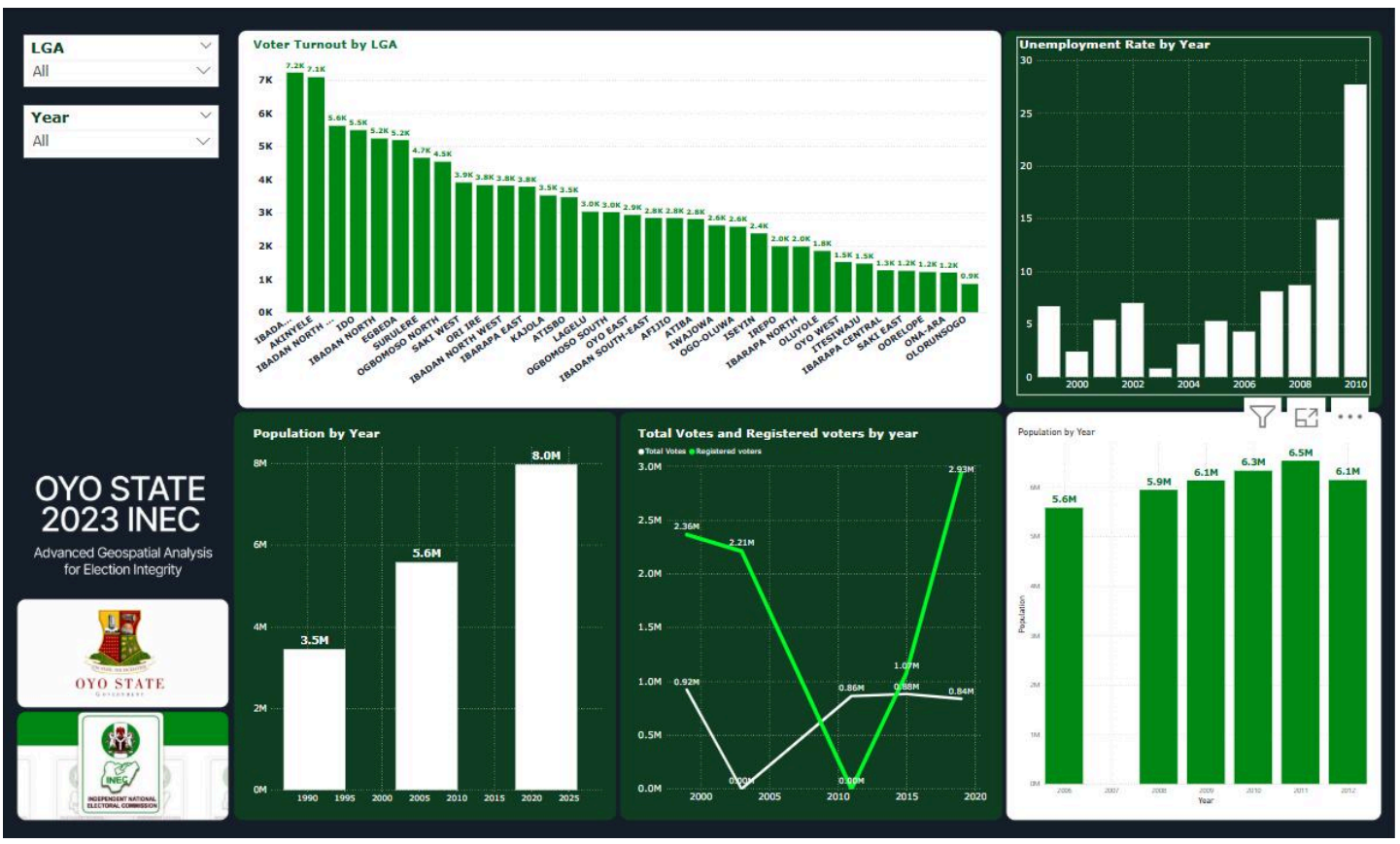
Population by year

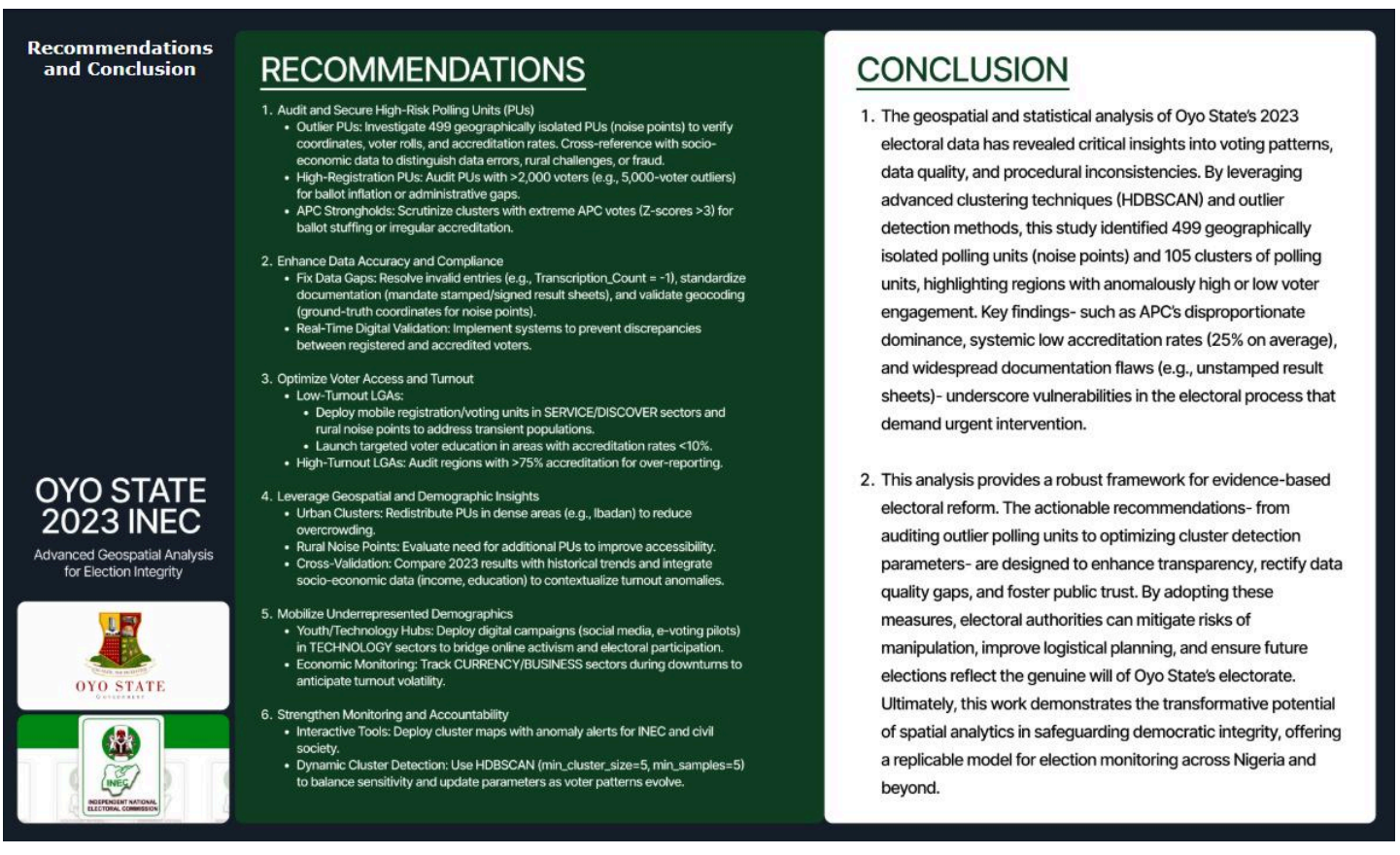


Total votes and registered voters by year

4.5. Interactive Visualization and Reporting







Click here

5. RECOMMENDATIONS

1. Audit and Secure High-Risk Polling Units (PUs)
    - Outlier PUs: Investigate 499 geographically isolated PUs (noise points) to verify coordinates, voter rolls, and accreditation rates. Cross-reference with socio-economic data to distinguish data errors, rural challenges, or fraud.
    - High-Registration PUs: Audit PUs with >2,000 voters (e.g., 5,000-voter outliers) for ballot inflation or administrative gaps.
    - APC Strongholds: Scrutinize clusters with extreme APC votes (Z-scores >3) for ballot stuffing or irregular accreditation.

2. Enhance Data Accuracy and Compliance
    - Fix Data Gaps: Resolve invalid entries (e.g., Transcription_Count = -1), standardize documentation (mandate stamped/signed result sheets), and validate geocoding (ground-truth coordinates for noise points).
    - Real-Time Digital Validation: Implement systems to prevent discrepancies between registered and accredited voters.

3. Optimize Voter Access and Turnout
    - Low-Turnout LGAs:
        - Deploy mobile registration/voting units in SERVICE/DISCOVER sectors and rural noise points to address transient populations.
        - Launch targeted voter education in areas with accreditation rates <10%.
    - High-Turnout LGAs: Audit regions with >75% accreditation for over-reporting.

4. Leverage Geospatial and Demographic Insights
    - Urban Clusters: Redistribute PUs in dense areas (e.g., Ibadan) to reduce overcrowding.
    - Rural Noise Points: Evaluate need for additional PUs to improve accessibility.
    - Cross-Validation: Compare 2023 results with historical trends and integrate socio-economic data (income, education) to contextualize turnout anomalies.

5. Mobilize Underrepresented Demographics
    - Youth/Technology Hubs: Deploy digital campaigns (social media, e-voting pilots) in TECHNOLOGY sectors to bridge online activism and electoral participation.
    - Economic Monitoring: Track CURRENCY/BUSINESS sectors during downturns to anticipate turnout volatility.

6. Strengthen Monitoring and Accountability
    - Interactive Tools: Deploy cluster maps with anomaly alerts for INEC and civil society.
    - Dynamic Cluster Detection: Use HDBSCAN (min_cluster_size=5, min_samples=5) to balance sensitivity and update parameters as voter patterns evolve.

6. CONCLUSION

1. The geospatial and statistical analysis of Oyo State's 2023 electoral data has revealed critical insights into voting patterns, data quality, and procedural inconsistencies. By leveraging advanced clustering techniques (HDBSCAN) and outlier detection methods, this study identified 499 geographically isolated polling units (noise points) and 105 clusters of polling units, highlighting regions with anomalously high or low voter engagement. Key findings- such as APC's disproportionate dominance, systemic low accreditation rates (25% on average), and widespread documentation flaws (e.g., unstamped result sheets)- underscore vulnerabilities in the electoral process that demand urgent intervention.

2. This analysis provides a robust framework for evidence-based electoral reform. The actionable recommendations- from auditing outlier polling units to optimizing cluster detection parameters- are designed to enhance transparency, rectify data quality gaps, and foster public trust. By adopting these measures, electoral authorities can mitigate risks of manipulation, improve logistical planning, and ensure future elections reflect the genuine will of Oyo State's electorate. Ultimately, this work demonstrates the transformative potential of spatial analytics in safeguarding democratic integrity, offering a replicable model for election monitoring across Nigeria and beyond.