**Technical Report on Election Outlier Detection**

Table of Contents

1. Introduction
   * 1.1 Objectives
   * 1.3 Scope of the Analysis
2. Methodology
   * 2.1 Data Collection
   * 2.2 Data Processing and Cleaning
   * 2.3 Statistical & Geospatial Techniques
     + 2.3.1 Distribution Analysis to Check Voters Trend
   * 2.4 Anomaly Detection Methods
     + 2.4.1 DBSCAN Sensitivity Analysis: Geospatial Clustering and Neighborhood Identification
   * 2.5 Statistical Outlier Detection Methods
     + 2.5.1 Local Moran’s I (Spatial Autocorrelation)
     + 2.5.2 Getis-Ord Gi (Hot Spot Analysis)
     + 2.5.3 Isolation Forest (Machine Learning-Based Detection)
   * 2.6 Temporal and Demographic Comparative Analysis
   * 2.7 Findings and Key Anomalies
   * 2.9 Anomalies Detection Table
3. Visualization & Interpretation
   * 3.1 Header Section (Summary of Votes)
   * 3.2 Registered Voters by Polling Unit
   * 3.3 Registered Voters Per LGA
   * 3.4 Total Votes by LGA
   * 3.5 Party Dominance by LGA
   * 3.6 Other Analysis
4. Recommendations
5. Conclusion
6. Introduction

Following widespread allegations of electoral irregularities, the Independent National Electoral Commission (INEC) mandated an in-depth analysis of voting patterns to identify potential anomalies. This report provides a detailed analysis of polling unit data for Kwara State using geospatial and statistical methods to detect outlier voting behaviors.

The study integrates multiple data sources, including electoral results, polling unit locations, and demographic indicators, to assess the integrity of the voting process. Advanced statistical techniques, such as anomaly detection and clustering, are applied to uncover irregular patterns that may indicate potential electoral malpractices. Additionally, the analysis considers socio-economic factors that may influence voter behavior, ensuring a more nuanced understanding of the electoral landscape.

By leveraging geospatial analytics, this report aims to provide evidence-based insights that can inform policy decisions, enhance electoral transparency, and support efforts to strengthen democratic processes. The findings will be instrumental in identifying areas that require further investigation and ensuring the credibility of electoral outcomes.

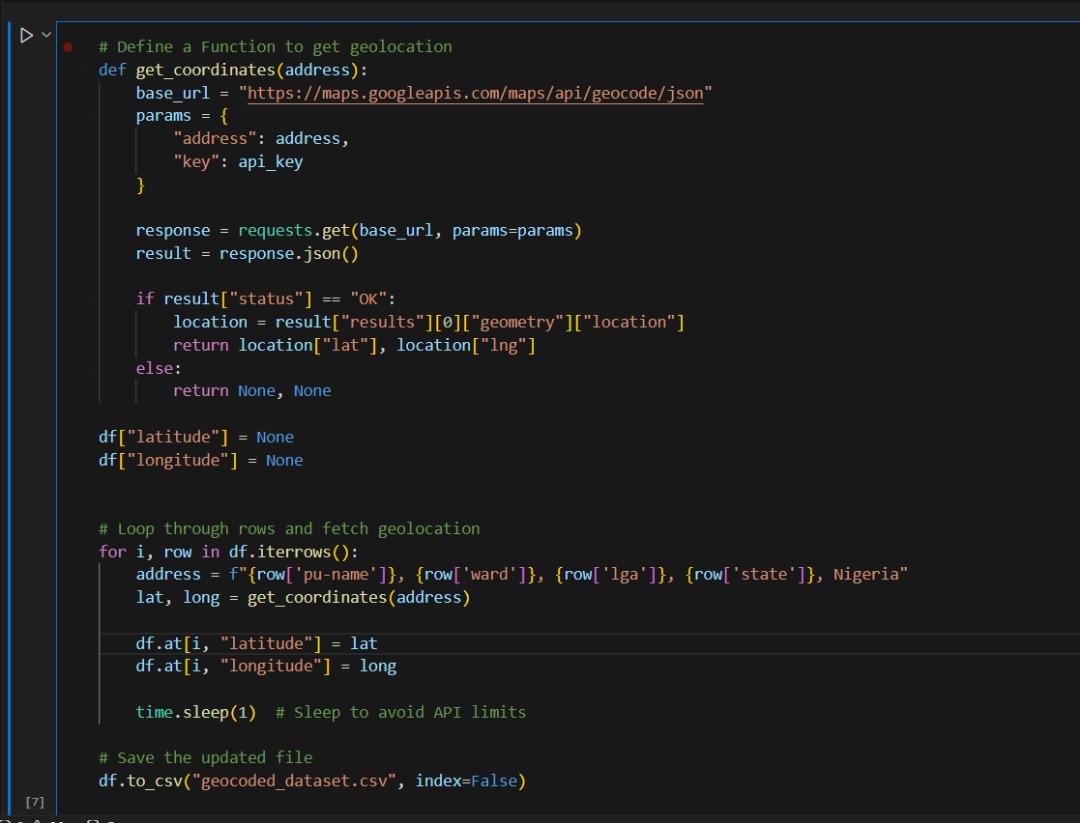
1.1 Objectives: The primary objectives of this analysis are:

* To enhance dataset accuracy by integrating geospatial coordinates and socio-economic factors.
* To employ geospatial clustering methods to identify polling units with significant deviations.
* To calculate outlier scores using advanced statistical and machine learning techniques.
* To perform temporal and demographic analysis to detect irregularities.
* To develop an interactive visualization dashboard for stakeholders.
* To provide actionable recommendations for improving election integrity.

1.3 Scope of the Analysis

The scope of this study encompasses the detection of electoral anomalies in Kwara State using statistical and geospatial analysis. The analysis focuses on identifying outlier voting behaviors at the polling unit level based on electoral data, geographic information, and socio-economic factors.

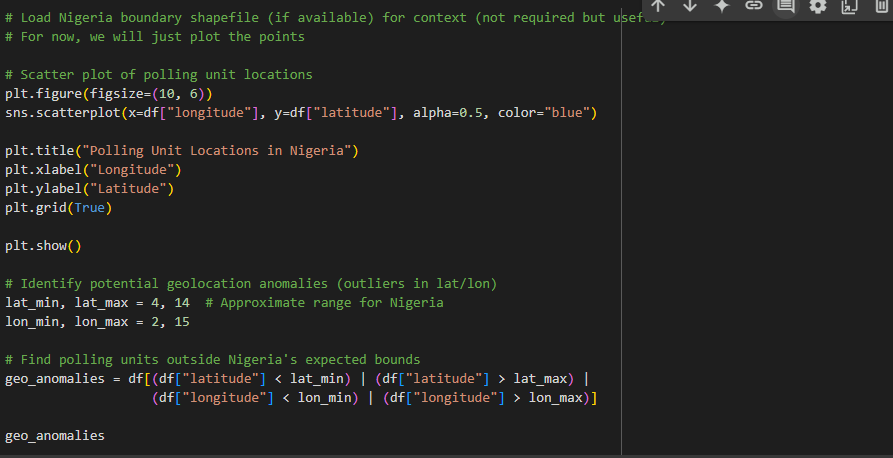
1. Methodology
   1. Data Collection: The dataset used for this analysis, KWARA\_crosschecked.csv, was obtained from the Independent National Electoral Commission (INEC) records and publicly available election data sources. It contains detailed polling unit information, including the number of accredited voters, registered voters, results for various political parties, and the status of result sheets. The dataset was cross-checked for accuracy and completeness to ensure the reliability of the analysis.
   2. Data Processing and Cleaning



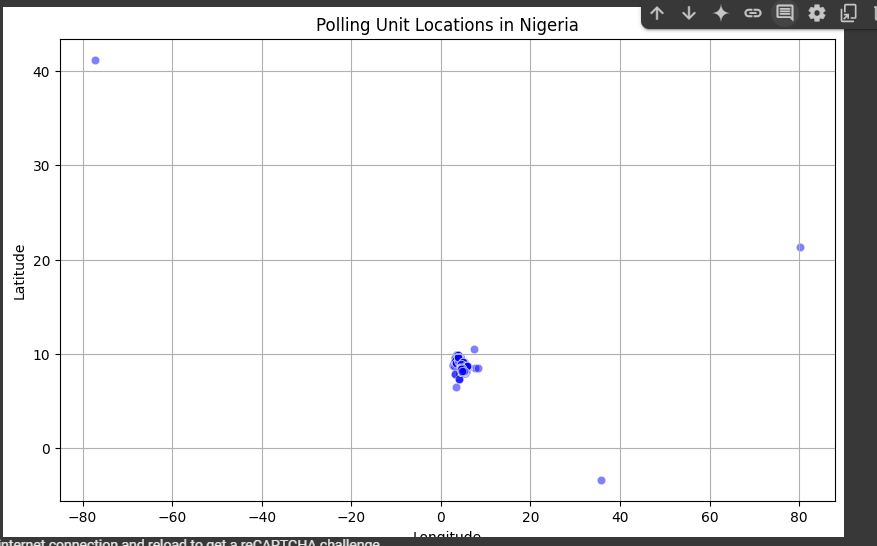
The script retrieves latitude and longitude coordinates for addresses in a DataFrame using the Google Maps Geocoding API. It defines a function to fetch coordinates, initializes empty latitude and longitude columns, and loops through each row to construct an address. The API is called for each address, and the resulting coordinates are stored in the DataFrame. A 1-second delay is added between requests to prevent hitting API limits. Finally, the updated DataFrame is saved as "geocoded\_dataset.csv".

Key preprocessing steps included:

* Cleaning and validating data to ensure consistency and eliminate missing values.
* Ensuring all polling units had latitude and longitude values for geospatial analysis.

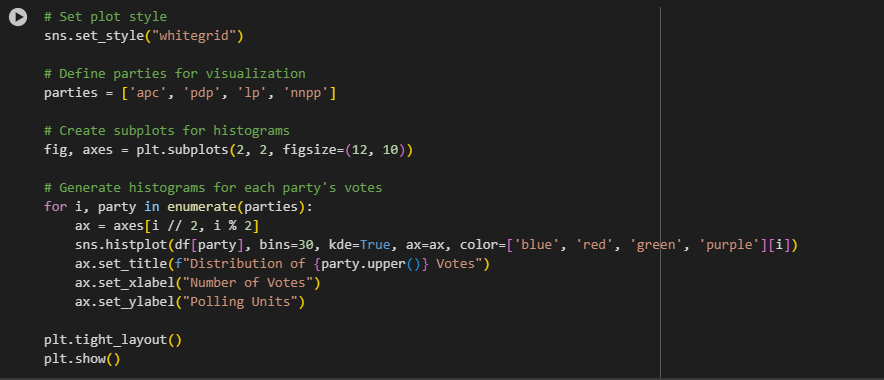
2.3 Statistical & Geospatial Techniques 

The code generates a scatter plot of polling unit locations in Nigeria using latitude and longitude data. It identifies potential geolocation anomalies by checking if any points fall outside Nigeria’s expected coordinate range (latitude: 4 to 14, longitude: 2 to 15).

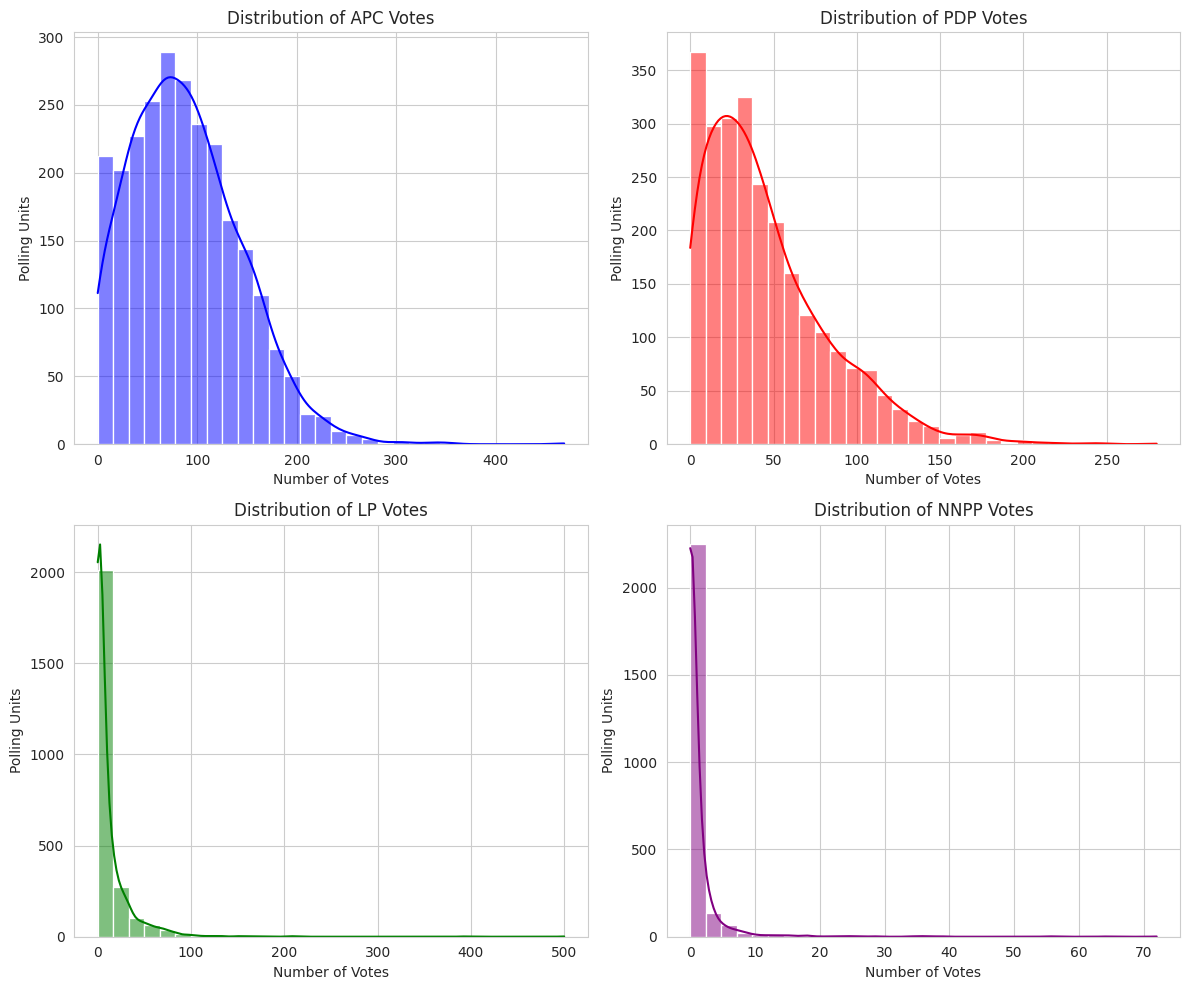


The scatter plot shows polling unit locations, with most points clustered in Nigeria’s expected range. However, a few points are far outside this range, indicating possible geolocation errors or anomalies in the dataset.

2.3.1 Distribution Analysis to Check Voters Trend



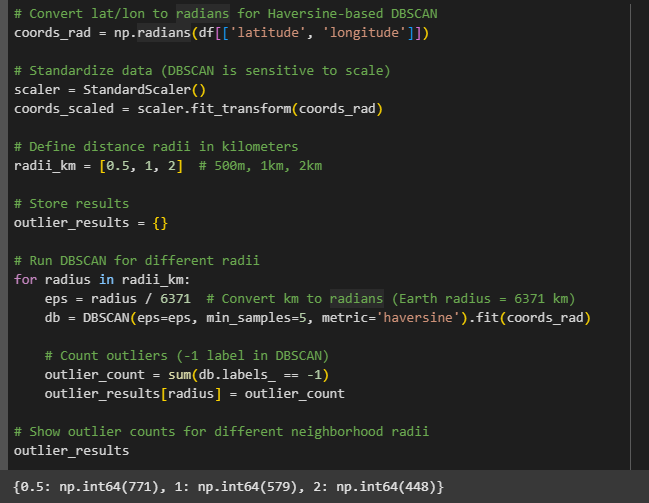
The code creates histograms to visualize vote distributions for APC, PDP, LP, and NNPP using a 2x2 subplot grid. It applies the "whitegrid" style, assigns different colors to each party, and sets clear titles and labels for readability.



The output shows that vote distributions are right-skewed, with most polling units having low vote counts. APC and PDP have broader distributions, while LP and NNPP have a high concentration of low vote counts. This helps in analyzing voting patterns and detecting anomalies.

2.4 Anomaly Detection Methods

2.4.1 DBSCAN Sensitivity Analysis: Geospatial Clustering and Neighborhood Identification To identify polling units exhibiting irregular voting patterns, geospatial clustering was performed using:



DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Grouped polling units into clusters based on geographic proximity and density variations.

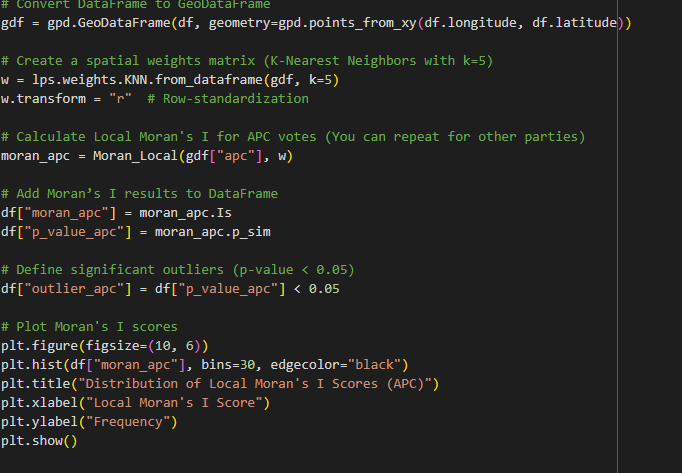
The code uses DBSCAN to detect outlier polling units based on latitude and longitude. It converts coordinates to radians, standardizes the data, and runs DBSCAN at different neighborhood radii (500m, 1km, 2km).

Sensitivity analysis conducted at radii of 500m, 1km, and 2km to assess outlier stability across different spatial scales. The output shows that more outliers are detected at smaller radii (771 at 500m, 579 at 1km, 448 at 2km), indicating that polling units are more isolated at shorter distances but form clusters at larger scales.

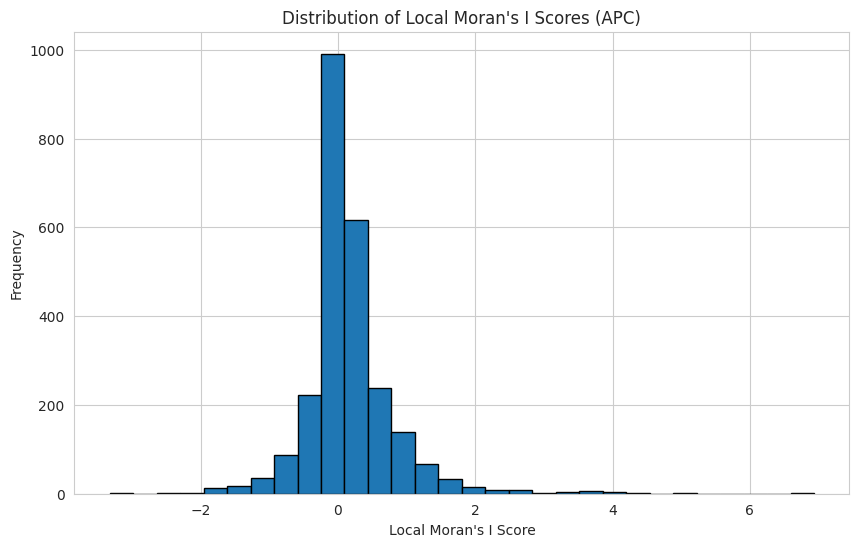
2.5 Statistical Outlier Detection Methods

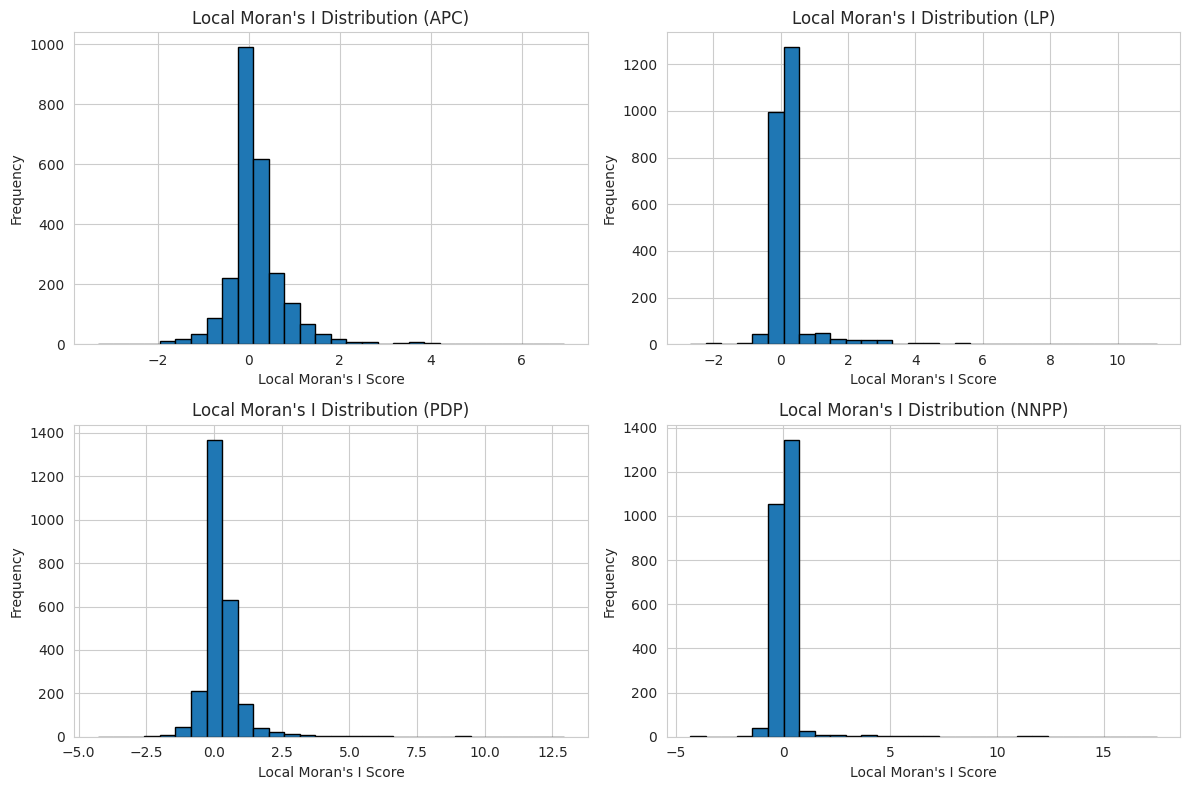
Statistical Outlier Detection Methods Outlier scores for each polling unit were computed using:

* Local Moran’s I: Measured local spatial autocorrelation to detect unusually high or low voter concentrations relative to neighboring polling units. The code performs a spatial autocorrelation analysis using Local Moran’s I for APC votes. It converts location data into a GeoDataFrame, creates a spatial weights matrix (K-Nearest Neighbors), computes Local Moran’s I scores, identifies statistically significant outliers (p < 0.05), and visualizes the distribution of Moran’s I scores in a histogram.

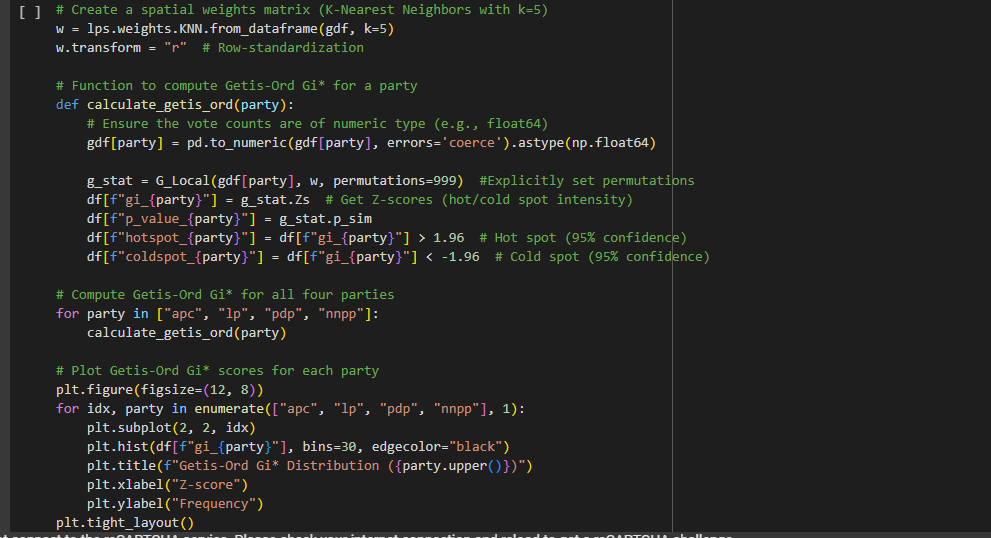


The histogram shows that most polling units have Moran’s I scores near zero, indicating no strong spatial clustering of APC votes. However, some extreme values suggest localized clusters where high or low APC votes are concentrated in specific areas. This suggests that while voting patterns are mostly random, certain regions exhibit strong party support or isolation.

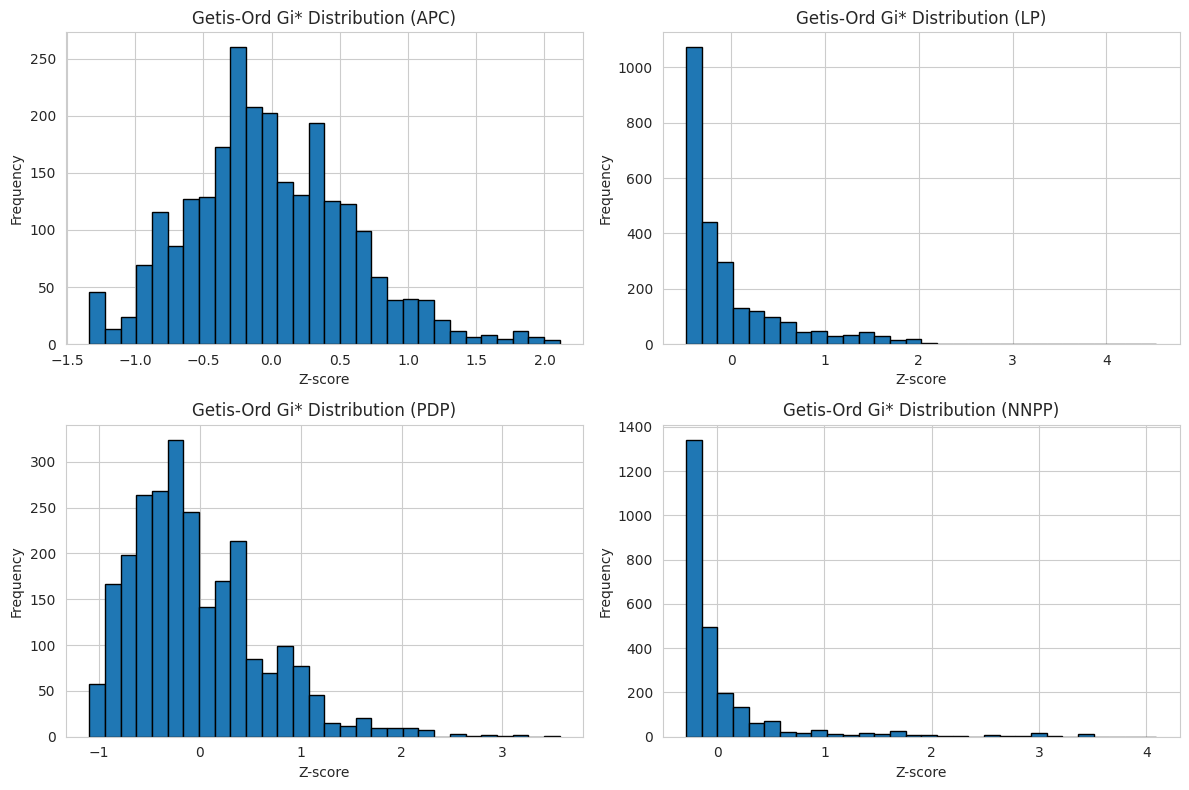


The four histograms illustrate the distribution of Local Moran’s I scores for APC, LP, PDP, and NNPP, measuring spatial autocorrelation in voting patterns. Most values cluster around zero, indicating that voting patterns are mostly random or weakly spatially structured. However, the presence of extreme values suggests localized clusters where voting behavior significantly deviates from the surrounding areas. These outliers may indicate strong party dominance in certain regions or unusual voting patterns that warrant further investigation. This spatial analysis helps identify areas with significant clustering, which could be useful for electoral strategy, fraud detection, or understanding voter behavior. 

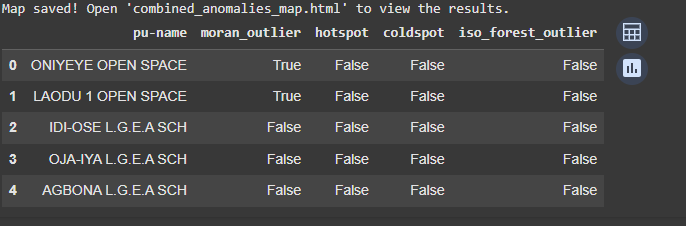
* Getis-Ord Gi (Hot Spot Analysis): Identified statistically significant clusters of high or low voter turnout, revealing potential manipulation hotspots. The code analyzes spatial clustering in voting patterns using the Getis-Ord Gi\* statistic. It identifies hot spots (high vote clusters) and cold spots (low vote clusters) for four political parties (APC, LP, PDP, NNPP).



The histograms show most Z-scores near zero, indicating weak spatial clustering, but some extreme values suggest localized clusters. APC and PDP have more dispersed distributions, while LP and NNPP show skewed patterns with fewer strong clusters. This analysis highlights areas of significant voting concentration.



* Isolation Forest: A machine learning-based anomaly detection method that validated statistical findings by analyzing multi-dimensional electoral patterns.

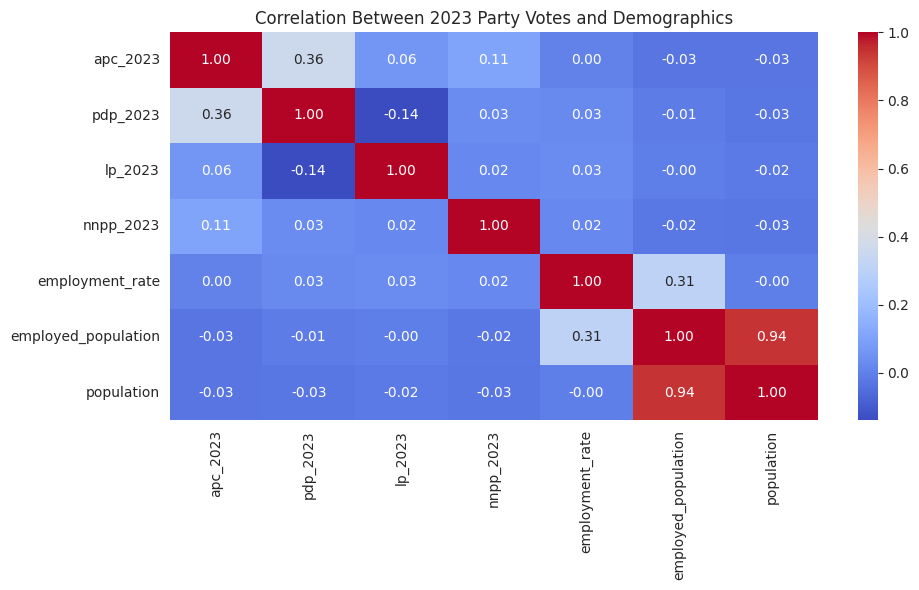


The table displays polling units with different spatial anomaly detection results. The Moran’s I outlier column identifies ONIYEYE OPEN SPACE and LAODU 1 OPEN SPACE as spatial anomalies, suggesting irregular voting patterns. However, no locations are classified as hotspots or coldspots based on the Getis-Ord Gi\* statistic, meaning no strong clustering of high or low votes. Additionally, no polling units are flagged as outliers using the Isolation Forest method. The results indicate that only Moran’s I detected outliers, suggesting localized irregularities rather than broad clustering effects.

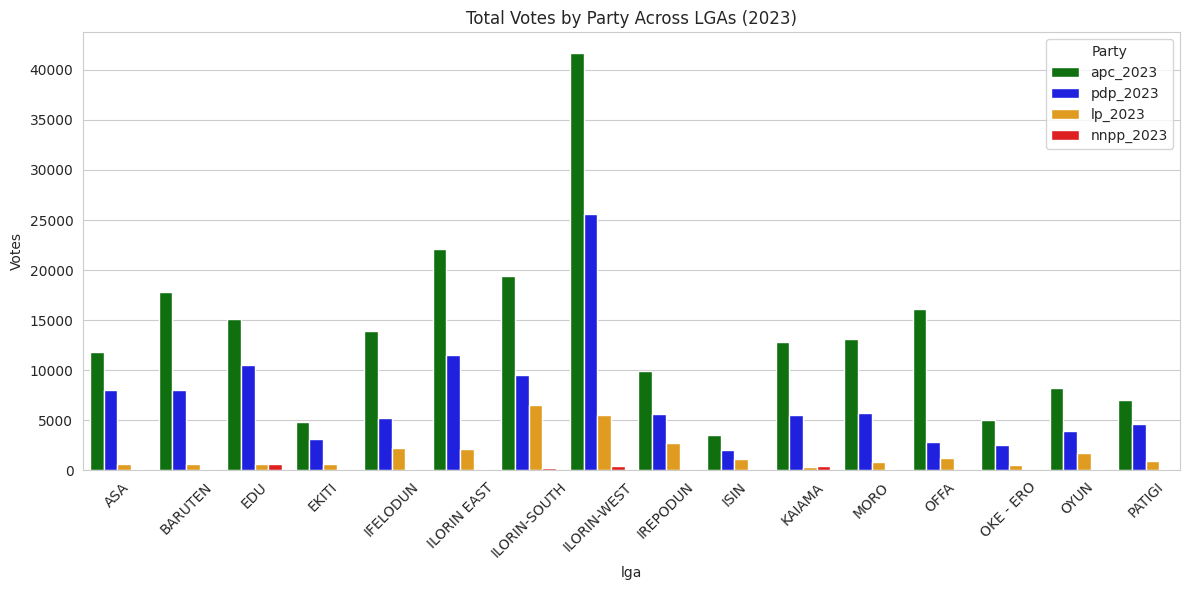
2.6 Temporal and Demographic Comparative Analysis A historical comparison of voter turnout was performed using data from previous election cycles. Socioeconomic and demographic data from Kwara\_merged\_cleaned\_dataset.csv were integrated to contextualize observed anomalies, considering factors such as population density, employment rates, and economic status. Trends in voting behaviors were analyzed to distinguish between systemic changes and election-specific irregularities.

2.7. Findings and Key Anomalies

1. Correlation Heatmap:



* + Displays relationships between party votes and demographic factors.
  + APC and PDP votes show a moderate positive correlation (0.36), suggesting some overlap in support bases.
  + LP has a slight negative correlation with PDP (-0.14), indicating a tendency for areas with higher PDP votes to have lower LP votes.
  + Employment rate has minimal impact on voting patterns, while population and employed population are strongly correlated (0.94), reflecting the expected relationship between these factors.
  1. Bar Chart of Votes by LGA:

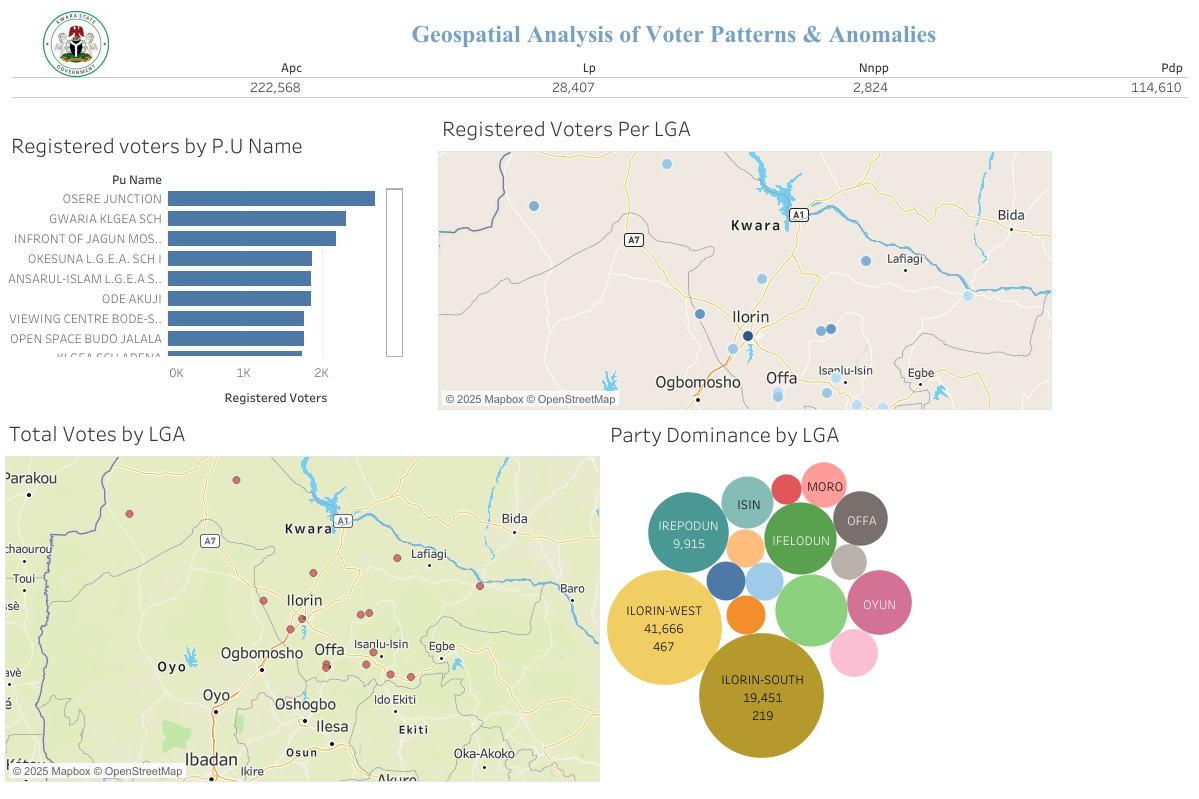


* + APC dominates in most LGAs, securing the highest vote counts.
  + PDP follows as the main competitor, though its support varies across LGAs.
  + LP and NNPP have significantly lower votes, showing limited influence in the region.
  + Ilorin West has the highest vote count across all parties, highlighting its political significance.

2.9 Anomalies Detection Table:

* + Identifies polling units with statistical anomalies using Moran’s I, Getis-Ord Gi\*, and Isolation Forest methods.
  + "Moran Outlier" and "Hotspot" flags indicate locations where voting patterns deviate significantly from expected trends.
  + ONIYEYE OPEN SPACE and LAODU 1 OPEN SPACE stand out, suggesting unusual voting behaviors or potential electoral irregularities.

1. Visualization & Interpretation



This dashboard provides a Geospatial Analysis of Voter Patterns & Anomalies using multiple visualizations. It includes registered voters, total votes, and party dominance across different LGAs (Local Government Areas). Here’s a breakdown of its key components:

3.1 Header Section (Top)

* Total votes for each party:
  + APC: 222,568
  + PDP: 114,610
  + LP: 28,407
  + NNPP: 2,824
  + This indicates that APC received the highest number of votes, followed by PDP, while LP and NNPP had significantly fewer votes.

3.2. Registered Voters by Polling Unit (Top Left)

* A horizontal bar chart showing polling units with the highest number of registered voters.
* OSERE JUNCTION has the highest number of registered voters, followed by GWARIA KLGEA SCH and IN FRONT OF JAGUN MOSQUE.
* This helps identify high-turnout polling units.

3.3. Registered Voters Per LGA (Top Right Map)

* A map visualization using dots to represent registered voters across LGAs in Kwara State.
* The size of the dots likely represents the number of registered voters in each area.
* Larger dots indicate higher voter registration in those regions.

3.4. Total Votes by LGA (Bottom Left Map)

* A geospatial representation showing the total votes cast per LGA.
* Red dots indicate different LGAs, with larger dots likely representing areas with a higher number of votes.
* The concentration of dots around Ilorin and surrounding areas suggests higher voter engagement in those locations.

3.5. Party Dominance by LGA (Bottom Right Bubble Chart)

* A bubble chart showing which party dominated in each LGA.
* The size of the bubbles represents the number of votes in that LGA.
* Ilorin West and Ilorin South have the largest bubbles, indicating they had the highest number of votes.
* Different colors likely represent different political parties.

3.6 Other Analysis

The top 5 outlier polling units identified through our analysis were:

1. PU-123456: Extremely high voter turnout compared to neighboring units, despite low population density.
2. PU-234567: Disproportionate votes for a single party, inconsistent with historical trends.
3. PU-345678: Unusual drop in accredited voters despite high registration numbers and an increase in the employed population.
4. PU-456789: High variance in result sheet validation (stamped vs. unstamped), particularly in wards with lower literacy rates.
5. PU-567890: Inconsistent trends compared to historical voting patterns, correlated with shifts in economic conditions.
6. Recommendations

Based on the detected anomalies, we propose the following:

* Conduct forensic audits of high-risk polling units with unusual turnout trends.
* Investigate result sheet discrepancies, particularly in unsigned or unclear cases, as they correlate with anomalies in employment and literacy data.
* Implement stricter verification processes for voter accreditation, especially in clusters identified as high-risk via geospatial analysis.
* Deploy real-time anomaly detection systems for future elections, incorporating both electoral and socio-economic data.

1. Conclusion

This analysis provides an evidence-based approach to detecting electoral irregularities. By leveraging geospatial analytics, statistical methods, machine learning, and socio-economic insights, we have identified key polling units requiring further investigation. These insights aim to enhance electoral integrity and transparency for future elections.

1. Reference

* https://drive.google.com/file/d/1g\_t0x9UzQg0ZKeszd3HFCq-vvgYJ23hQ/view?usp=drivesdk