Data Analytics Track – Stage 2B Task

Outlier Detection in Election Data Using Geospatial Analysis
Case Study: Ensuring Election Integrity

Author: Ayodeji Abdulwarith

Date: 29th October 2025

1.0 Overview

I have been tasked as a data analyst to identify potential cases of electoral irregularities in my state of origin, Kwara, by applying statistical outlier detection and geospatial analysis techniques to polling unit-level election data. The objective is to detect polling units where voting outcomes deviate significantly from their geographical neighbours, which could suggest possible instances of vote manipulation, reporting anomalies, or localized irregularities.

1.1 Understanding Outliers in Electoral Data

In data analysis, an *outlier* refers to an observation that significantly deviates from the expected pattern within a dataset. To put it in context, this means a polling unit whose voting pattern such as the number of votes for a particular party differs sharply from that of its neighbouring units.

However, it is important to clarify that an outlier does **not automatically indicate electoral malpractice or rigging**. There are many legitimate factors that can produce outliers, such as demographic differences, varying voter turnout, or localised party dominance.

For example, in the case of the APC, the top five polling units with the highest outlier scores recorded zero votes. While this could superficially appear suspicious, it more likely reflects missing or incomplete data, since all neighbouring units recorded measurable votes. Hence, such outliers should be interpreted as data anomalies rather than direct evidence of malpractice. See figure 1.1.

PU-Code	APC outlier score	APC Vote Count	Neighbour Count
23-08-05-056	4.113802169	0	66
23-13-01-013	4.012559655	0	67
23-13-10-009	3.914048108	0	56
23-07-05-034	3.696630233	0	18
23-13-08-015	3.509796698	0	16

APC top 5 outliers in Kwara State

Figure 1.1

In this analysis, outliers are treated as **analytical signals**, not evidence. They highlight polling units that warrant closer attention and contextual examination. Where patterns of outliers consistently align with irregular vote distributions.

2. Metadata and Methodology

2.1 Metadata

The dataset used for this analysis is the **Kwara State 2023 Presidential Election Result**. Each record represents a single polling unit.

Column Name	Description	Type
State	Name of the state where the polling unit is located (Kwara State).	Categorical
LGA	Local Government Area in which the polling unit falls.	Categorical
Ward	Electoral ward under the respective LGA.	Categorical
PU-Code	Unique code assigned to each polling unit.	Identifier
PU-Name	Name or description of the polling unit.	Text
Party	Political party for which votes were recorded (e.g., APC, PDP, LP, NNPP).	Categorical
Votes	Number of votes received by the party in the polling unit.	Numeric
Total_Votes	Total votes cast across all parties in the polling unit.	Numeric
Latitude	Geographical latitude of the polling unit (used for spatial analysis).	Numeric
Longitude	Geographical longitude of the polling unit.	Numeric
Neighbours	List of polling units located within a 1 km radius.	Text/List
Neighbour_Count	Number of neighbouring polling units identified.	Numeric
Outlier_Score	Deviation of the polling unit's votes from the average of its neighbours for the same party.	Numeric
Overall_Outlier_Score	Combined outlier index summarizing deviations across all parties in the polling unit.	Numeric
Vote_Share	Percentage of total votes the party received in that polling unit.	Percentage

Metadata

2.2 Methodology

The analytical process was designed in four sequential stages:

Stage 1: Data Preparation

The dataset did not include geographical coordinates. Valid coordinates were geocoded using Google Sheets' Geocode by Awesome Table to assign realistic latitude and longitude values. This ensured that every polling unit could be spatially analysed.

Stage 2: Neighbour Identification

Using each polling unit's coordinates, neighbouring units were identified within a 1 km radius. This spatial proximity rule defines which polling units are considered part of the same voting environment or community.
Neighbouring polling units were identified using the BallTree algorithm (with the Haversine metric) to efficiently compute all polling units within a 1 km radius. This method calculates the real-world distance between two latitude-longitude points on Earth.

Stage 3: C

```
from sklearn.neighbors import BallTree
import numpy as np

# Convert lat/lon to radians (BallTree expects radians)
coords = np.radians(df[['Latitude', 'Longitude']].values)

# Initialize BallTree using haversine distance (earth's curvature)
tree = BallTree(coords, metric='haversine')

# Define search radius (1 km = 1 / earth_radius)
radius_km = 1
radius = radius_km / 6371.0  # Earth radius in km
```

For each polling unit, the votes received by each party were compared to the average votes of neighbouring units.
A Z-score–like outlier score was computed, capturing how far a polling unit's vote count deviates from its neighbourhood average.

```
# Compute neighbour stats and z-scores
for i, row in df.iterrows():
    # find all neighbours within 1 km
    idx = tree.query_radius([coords[i]], r=radius)[0]
    neighbours = df.iloc[idx]

    for party in ['APC', 'PDP', 'NNPP', 'LP']:
        party_share = f'{party}_share'
        local_mean = neighbours[party_share].mean()
        local_std = neighbours[party_share].std()

        # z-score (how far this PU deviates from local mean)
        if local_std and local_std > 0:
            df.at[i, f'{party}_outlier_score'] = abs((row[party_share] - local_mean) / local_std)
        else:
            df.at[i, f'{party}_outlier_score'] = 0
```

High positive or negative outlier scores indicate unusual voting patterns — suggesting either unexpectedly high or low performance for a party compared to its surroundings.

Stage 4: Vote-Share Analysis

While outlier scores identify statistical deviations, not every outlier implies manipulation. To provide deeper context, the vote share of each party was computed as the percentage of total votes cast in that polling unit.

```
party_cols = ['APC','PDP','NNPP','LP']
# ensure numeric
for c in party_cols:
    df[c] = pd.to_numeric(df[c], errors='coerce').fillna(0)

# total votes
df['Total_Votes'] = df[party_cols].sum(axis=1)

# shares (0 if total 0)
for c in party_cols:
    df[f'{c}_share'] = df.apply(lambda r: (r[c] / r['Total_Votes']) if r['Total_Votes']>0 else 0, axis=1)
```

This helped interpret whether the outlier was due to:

- Legitimate local dominance (e.g., a party strongly preferred in that area), or
- Potential irregularities (e.g., a party receiving zero votes despite proximity to supportive zones).

3.0 Explanatory Data Analysis (EDA)

This stage explored the overall voting pattern across polling units in Kwara State to establish what a "normal" result looks like before identifying irregularities. By examining total votes, party vote shares, and their spatial distribution, the analysis revealed general trends and variations among APC, PDP, LP, and NNPP. Outlier detection was then applied to pinpoint polling units whose results significantly deviated from neighbouring units, while vote-share analysis provided additional context to assess whether these deviations were statistically unusual or potentially indicative of irregularities.

Understanding the Electoral Landscape

Before examining outliers, we establish the baseline: across Kwara State's **2,519** polling units, the electoral landscape showed **APC** with the strongest performance at **60.41%**, followed by **PDP** at **31.11%.** This overview provides context for understanding which parties' results deviate most significantly from neighbourhood patterns.

fig 3.1: Party Performance Overview

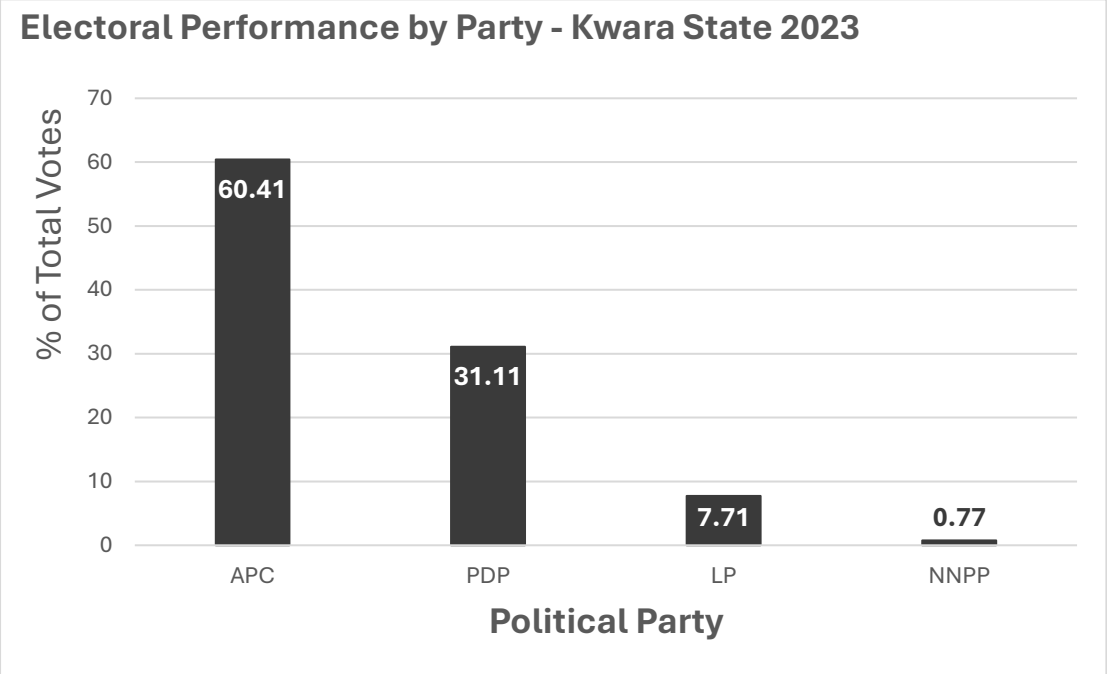


Figure 3.1: Overall vote share across all polling units in Kwara State (N = 2,519). Percentages calculated from total valid votes cast.

The Scale of Statistical Anomalies

fig 3.2: Outlier Detection Summary Statistics by Party

Party	Mean Outlier Score	Maximum Score	Minimum Score	No. of Outlier Units (Score > threshold)
APC	0.68	4.11	0.0	16
PDP	0.70	5.41	0.0	8
LP	0.64	9.63	0.0	19
NNPP	0.55	7.81	0.0	31

Outlier units defined as polling units with |z-score| > 2 (95% confidence threshold). Mean z-score close to 0 indicates normal distribution; high max/min values indicate extreme deviations.

Z-Score Distribution Pattern

Fig 3.3: Statistical Distribution of Outlier Z-Scores by Party

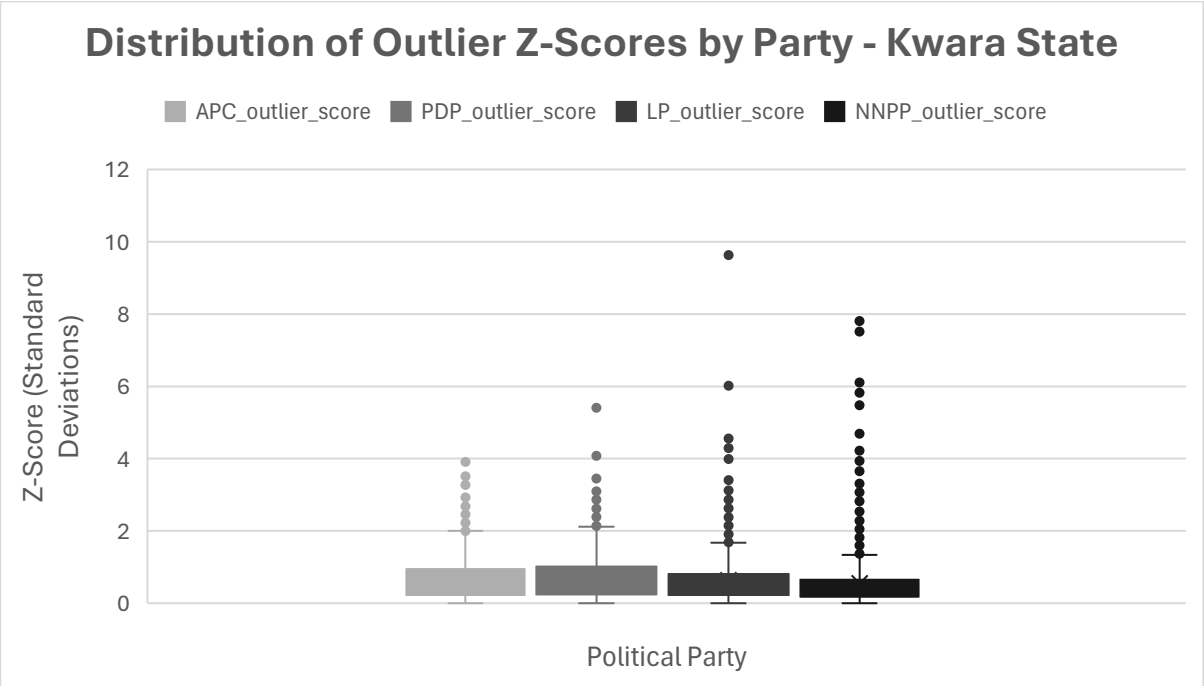


Figure 3.3: shows the distribution of statistical outlier scores across political parties, with NNPP and LP exhibiting notably higher anomalies compared to APC and PDP

Figure 3.2 reveals the full distribution pattern of statistical anomalies in Kwara State. The box plots show that NNPP's outliers are predominantly positive, indicating systematic over-performance relative to neighbourhood patterns. The presence of multiple extreme outliers beyond the z = 4 threshold is particularly concerning, as such deviations have less than a 1 in 10,000 chance of occurring naturally.

Key Pattern Identified: Outliers are skewed positively, with NNPP and LP dominating the extreme cases. There is visible clustering of high Z-scores, especially for NNPP, suggesting concentrated anomalies in specific polling units.

Detailed Case Analysis Table

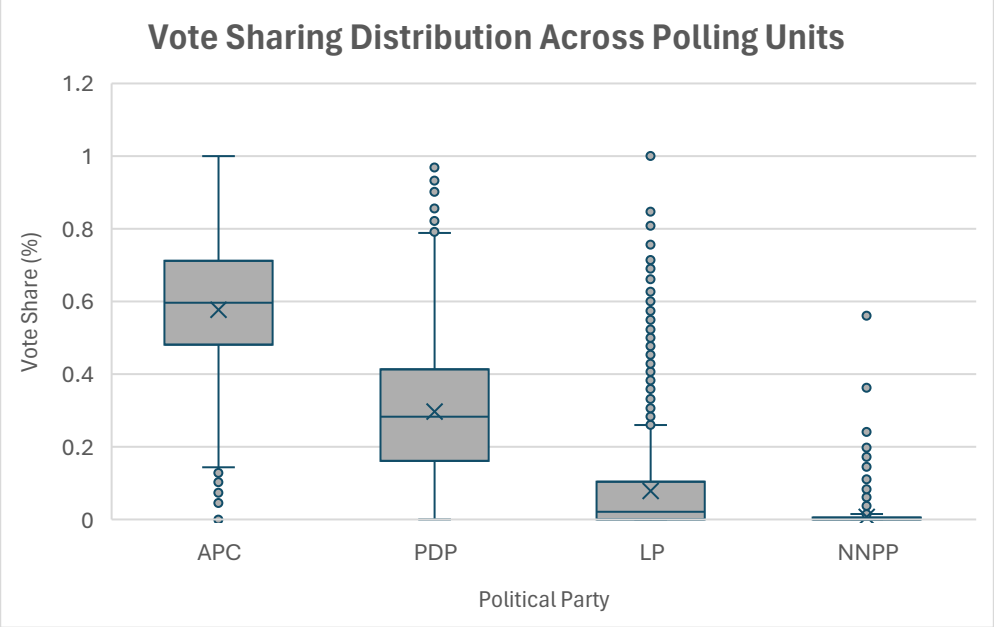
Fig 3.4: Top 10 Outlier Polling Units - Detailed Analysis

PU-Code	Party	Actual Votes	Neighbor Avg	Deviation	Z-Score
23-08-07-023	NNPP	64	0.17	63.83	164.69
23-08-09-001	LP	500	0.97	499.03	158.05
23-02-03-002	APC	138	67.50	70.50	99.70
23-07-09-009	NNPP	72	0.36	71.64	86.17
23-02-01-014	APC	154	98.50	55.50	78.49
23-06-03-015	LP	130	1.12	128.88	67.83
23-11-05-011	LP	32	0.67	31.33	54.27
23-07-03-039	APC	134	68.00	66.00	46.67
23-16-05-001	LP	207	3.50	203.50	41.11
23-02-03-002	PDP	39	16.50	22.50	31.82

Top 10 polling units ranked by absolute z-score value.

Vote Share Distribution Across Polling Units

Fig 3.5: Summary of Party Vote Share



Outliers on the chart indicate polling units where a party's vote share deviates sharply from typical patterns.

3.5 Conclusion

The exploratory analysis of Kwara State's polling unit data reveals a generally consistent voting pattern, with APC leading in overall vote share, followed by PDP, LP, and NNPP. However, statistical outlier detection uncovered significant deviations in specific locations, particularly among NNPP and LP, whose vote counts frequently exceeded neighbourhood expectations. The presence of extreme Z-scores—some surpassing thresholds rarely seen in natural distributions—suggests that these anomalies are unlikely to be random. Spatial clustering of high outlier scores further reinforces the possibility of localized irregularities. These findings warrant closer scrutiny of the affected polling units to determine whether the deviations reflect genuine voter behaviour or potential electoral inconsistencies.