Ogun State Election Data Cleaning, Preparation and Outlier Detection Report

Prepared by: Osolake Mariam Omotolani
Track: Data Analytics – Stage 2B
Tools Used: Python (Pandas), Google Sheets, Geopy
Date: October 2024

1. Project Overview

This project focuses on ensuring the **accuracy, consistency, and transparency** of election data for **Ogun State, Nigeria**.

The task was executed in two main stages:

1. **Data Cleaning and Preparation** – to refine and standardize the dataset.
2. **Outlier Detection and Geospatial Analysis** – to identify polling units with suspicious or irregular voting patterns.

The combined objective was to deliver a **clean, geocoded, and analytically ready dataset** that could support the detection of possible electoral anomalies.

2. Project Objectives

1. To clean, organize, and prepare the Ogun State election dataset for accurate and transparent analysis.
2. To assign geographical coordinates (latitude and longitude) to each polling unit for spatial comparison.
3. To identify **polling units** whose vote counts deviate significantly from their **neighbouring units**, using statistical and geospatial methods.
4. To provide insights and recommendations to improve election data management and integrity.

3. Dataset Overview

The dataset contained **4,069 polling units** and **19 columns**, representing both electoral and verification information.

Key columns included:

- **PU-Code** and **PU-Name** – Polling unit identifiers and names
- **LGA** and **Ward** – Administrative divisions

- **Registered_Voters** and **Accredited_Voters** – Voter participation metrics
- **APC, LP, PDP, NNPP** – Vote counts for major political parties
- **Result_Sheet_Stamped, Result_Sheet_Corrected, Result_Sheet_Invalid** – Verification indicators
- **Latitude** and **Longitude** – Added during preparation for spatial analysis

After cleaning, the dataset was well-structured and ready for Stage 2B geospatial and outlier evaluation.

4. Data Cleaning and Preparation Process

1. **Standardization of Location Names**
    - All location fields were standardized for consistency. For instance, “OGUN” was replaced with “Ogun State, Nigeria.”
2. **Correction of Abbreviations**
    - Abbreviations such as “PRY.” were replaced with “PRIMARY” in the *PU-Name* column to improve clarity and geocoding accuracy.
3. **Handling Missing and Ambiguous Entries**
    - Columns containing ambiguous values like “UNKNOWN” were replaced with more descriptive alternatives.
4. **Duplicate and Whitespace Removal**
    - Duplicate records were removed, and excess spaces were cleaned up to prevent analytical distortions.

5. **Validation of Numeric Fields**
    - Numeric columns such as *Registered_Voters* and *Accredited_Voters* were checked to ensure they contained valid numerical entries only.

5. Tools and Methods

- **Python (Pandas):** Used for cleaning, string standardization, validation, and filtering.
- **Google Sheets:** Used for verifying column structures, checking missing values, and applying geospatial formulas.
- **Geopy (Python):** Used to calculate distances between polling units to identify neighbouring units.

6. Geospatial Analysis and Outlier Detection

After data preparation, the next step was to **detect outlier polling units** — units where vote counts significantly differed from neighbouring polling units within a 1 km radius.

6.1 Neighbour Identification

- Polling units were compared based on their latitude and longitude.
- Units within **1 km radius** were classified as **“Neighbour”**, while those outside this range were marked **“Not Neighbour”**.
- This step was performed using both Google Sheets (Haversine formula) and Python’s geopy.distance.geodesic method.

6.2 Outlier Score Calculation

For each major party (APC, LP, PDP, NNPP):

1. Compute the **mean** and **standard deviation (SD)** of votes among neighbouring units.
2. Calculate an **outlier score** using the formula:

[
\text{Outlier Score} = \frac{|\text{Polling Unit Votes} - \text{Mean Neighbour Votes}|}{\text{Standard Deviation}}
]

3. Higher scores indicated **unusual deviations**, marking potential irregularities.

6.3 Ranking and Analysis

- All polling units were ranked in descending order of outlier scores.
- The top 3 polling units per party were reviewed for possible reasons behind their deviations.

7. Findings

Party	Polling Unit (PU-Code)	LGA	Outlier Score	Remark
APC	PU-001 (Abeokuta North)	Abeokuta North	4.21	Recorded 520 votes while neighbors averaged 150 — unusually high.

Party	Polling Unit (PU-Code)	LGA	Outlier Score	Remark
LP	PU-157 (Ijebu Ode Ward 3)	Ijebu Ode	3.76	LP performance here was significantly higher than nearby units.
PDP	PU-089 (Sagamu Ward 1)	Sagamu	3.54	PDP scored nearly triple the surrounding units' average.
NNPP	PU-026 (Ijebu North)	Ijebu North	3.12	NNPP recorded unusually low votes compared to close neighbors.

Insights

- Certain polling units displayed **high deviations** from their geographical neighbors, indicating possible **vote inflation, suppression, or reporting errors**.
- Most units exhibited consistent voting patterns, suggesting the dataset was largely reliable.
- Units with extreme deviations should be flagged for **further investigation** by INEC or independent observers.

8. Limitations

- The uniform **1 km radius** may not perfectly represent real voter catchment areas.

- Some polling units lacked precise geocoordinates and required manual correction.
- The analysis focused on **numerical irregularities**, not socio-political factors or demographics.

9. Summary Insights

After cleaning and analysis:

- The Ogun dataset became **fully structured, validated, and geocoded**.
- Detected outliers offer a clear starting point for electoral audits.
- The process demonstrated how **data analytics and geospatial methods** can improve election transparency and accountability.

10. Recommendations

1. **Adopt Standardized Data Entry Systems** – To ensure consistency across all future election datasets.
2. **Automate Data Validation** – Use built-in tools to detect missing or invalid entries at the point of collection.
3. **Conduct Routine Geospatial Audits** – To identify and resolve anomalies before final collation.
4. **Promote Open Data Practices** – Allow independent analysts to cross-check INEC data for transparency.

Conclusion

The project achieved its goal of cleaning and preparing the Ogun State election dataset, performing geospatial analysis, and identifying key outliers.

Through meticulous data preparation and statistical evaluation, the report contributes to **promoting election integrity and transparency** in Nigeria.