Slack ID: CoBird

Track: Data Analysis

Stage: Stage 8 task

Election Outlier Detection Report

1. Introduction

The Independent National Electoral Commission (INEC) has requested an in-depth geospatial and statistical analysis to identify potential anomalies in voting patterns for the 2023 presidential election that took place in Osun state. This report outlines the methodologies employed, findings, and recommendations to enhance election integrity.

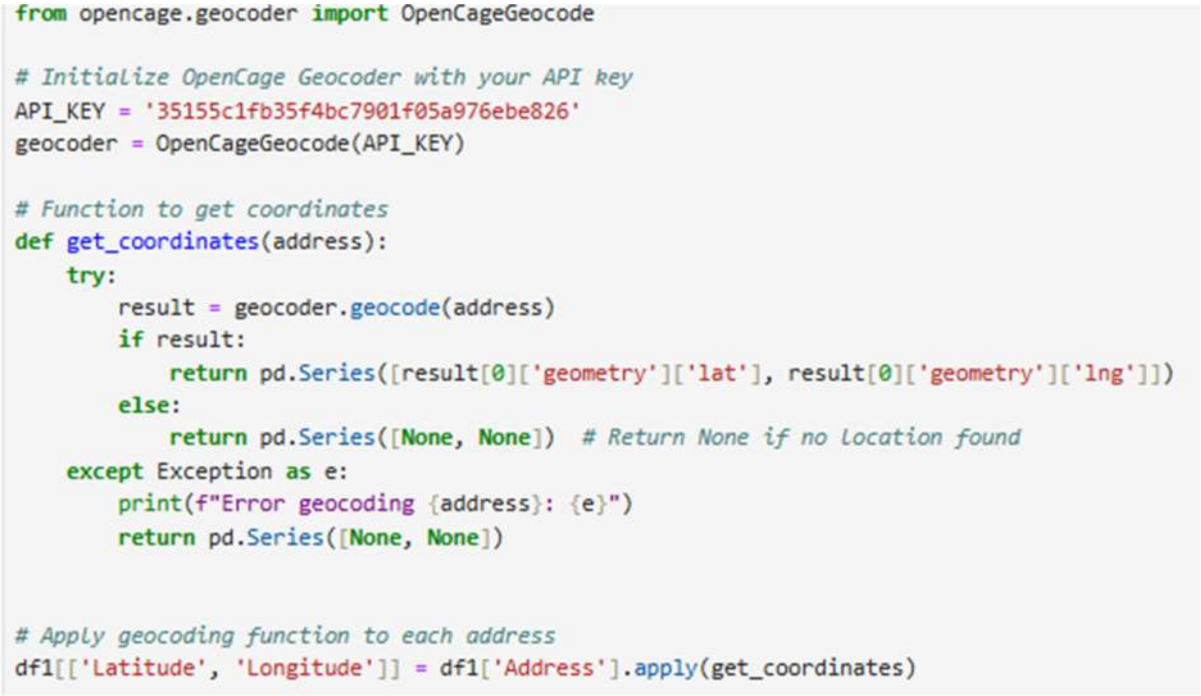
2. Methodology

2.1 Dataset Preparation

- The dataset for the selected state was downloaded and processed.
- The data was observed to be clean and devoid of any duplicates.
- The original data further contained 21 columns and 2248 rows.
- From the preliminary data analysis, Osun state had 2,247 polling units across 324 wards and 30 local government areas.
- Geospatial coordinates (latitude and longitude) were verified using both the Hdbscan and dbscan algorithms, while the most appropriate algorithm was incorporated into the dataset and used for further analysis.

2.2 Geocoding with Longitudinal and Latitudinal Data

- The dataset used for this exercise did not contain location coordinates and, therefore, provided a setback to the analysis.
- As a result, each polling unit was geolocated using appropriate python codes
- Open Cage Geocoder was used for this task due to its high accuracy and precision.



```python
from opencage.geocoder import OpenCageGeocode

# Initialize OpenCage Geocoder with your API key
API_KEY = '35155c1fb35f4bc7901f05a976ebe826'
geocoder = OpenCageGeocode(API_KEY)

# Function to get coordinates
def get_coordinates(address):
    try:
        result = geocoder.geocode(address)
        if result:
            return pd.Series([result[0]['geometry']['lat'], result[0]['geometry']['lng']])
        else:
            return pd.Series([None, None])  # Return None if no location found
    except Exception as e:
        print(f"Error geocoding {address}: {e}")
        return pd.Series([None, None])


# Apply geocoding function to each address
df1[['Latitude', 'Longitude']] = df1['Address'].apply(get_coordinates)
```

- Of the total 2,248 locations contained in the dataset, the Geocoder could not get coordinates for 131 polling units.
- This was handled by forward-filling the null coordinates since that was the common pattern observed in the coordinate column of the dataset.

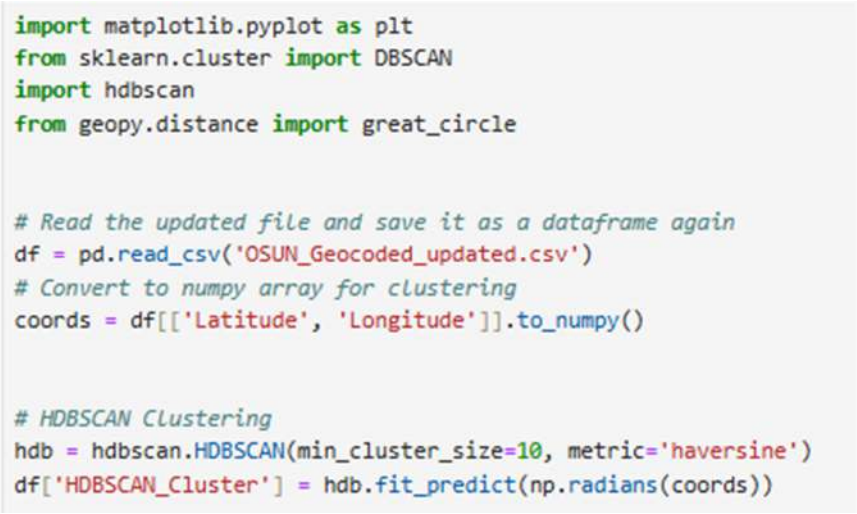
2.3 Neighbor Identification

- The DBSCAN algorithm was initially used to group the polling units according to geographical spread.

- Despite varying the neighborhood radius used in the algorithm, only two polling clusters could be generated.

- Consequently, the DBSCAN operation was discarded for the more automatic HDBSCAN clustering algorithm.

- The HDBSCAN clustering algorithm was used to dynamically group polling units based on their geographic proximity using appropriate Python code.



```
import matplotlib.pyplot as plt
from sklearn.cluster import DBSCAN
import hdbscan
from geopy.distance import great_circle


# Read the updated file and save it as a dataframe again
df = pd.read_csv('OSUN_Geocoded_updated.csv')
# Convert to numpy array for clustering
coords = df[['Latitude', 'Longitude']].to_numpy()


# HDBSCAN Clustering
hdb = hdbscan.HDBSCAN(min_cluster_size=10, metric='haversine')
df['HDBSCAN_Cluster'] = hdb.fit_predict(np.radians(coords))
```

- A total of 30 clusters were generated covering the entire polling units of Osun state.

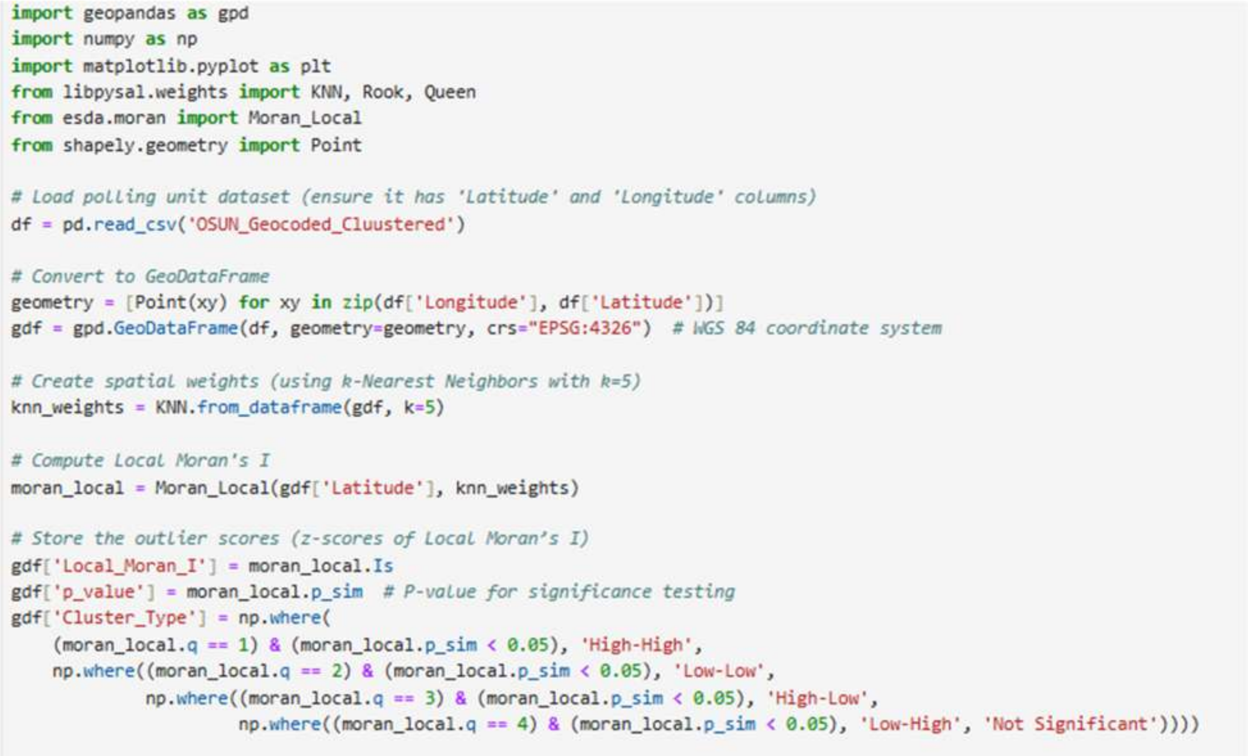
- The data on the generated clusters were updated to the dataset with the appropriate column name.
- Sensitivity analysis was conducted by varying neighborhood radii (500m, 1km, 2km) to assess the robustness of clusters.

2.4 Outlier Score Calculation

- **Local Moran's I**: This spatial autocorrelation method was employed to detect polling units that are sited in uncharacteristic locations.

- Via this algorithm, polling locations that are spatial outliers can be identified and detected.

- Using a combination of the Morans score and p-value, polling units and voting clusters that represent spatial outliers are identified.



```
import geopandas as gpd
import numpy as np
import matplotlib.pyplot as plt
from libpysal.weights import KNN, Rook, Queen
from esda.moran import Moran_Local
from shapely.geometry import Point

# Load polling unit dataset (ensure it has 'Latitude' and 'Longitude' columns)
df = pd.read_csv('OSUN_Geocoded_Cluustered')

# Convert to GeoDataFrame
geometry = [Point(xy) for xy in zip(df['Longitude'], df['Latitude'])]
gdf = gpd.GeoDataFrame(df, geometry=geometry, crs="EPSG:4326")  # WGS 84 coordinate system

# Create spatial weights (using k-Nearest Neighbors with k=5)
knn_weights = KNN.from_dataframe(gdf, k=5)

# Compute Local Moran's I
moran_local = Moran_Local(gdf['Latitude'], knn_weights)

# Store the outlier scores (z-scores of Local Moran's I)
gdf['Local_Moran_I'] = moran_local.Is
gdf['p_value'] = moran_local.p_sim  # P-value for significance testing
gdf['Cluster_Type'] = np.where(
    (moran_local.q == 1) & (moran_local.p_sim < 0.05), 'High-High',
    np.where((moran_local.q == 2) & (moran_local.p_sim < 0.05), 'Low-Low',
             np.where((moran_local.q == 3) & (moran_local.p_sim < 0.05), 'High-Low',
                      np.where((moran_local.q == 4) & (moran_local.p_sim < 0.05), 'Low-High', 'Not Significant'))))
```

- The result of the analysis revealed that polling units in 612 locations and covering 7 clusters (clusters 0,1,2,3,4,5, and 7) are spatial outliers. See figure below and the visualization dashboard for a more detailed map view of spatial outliers

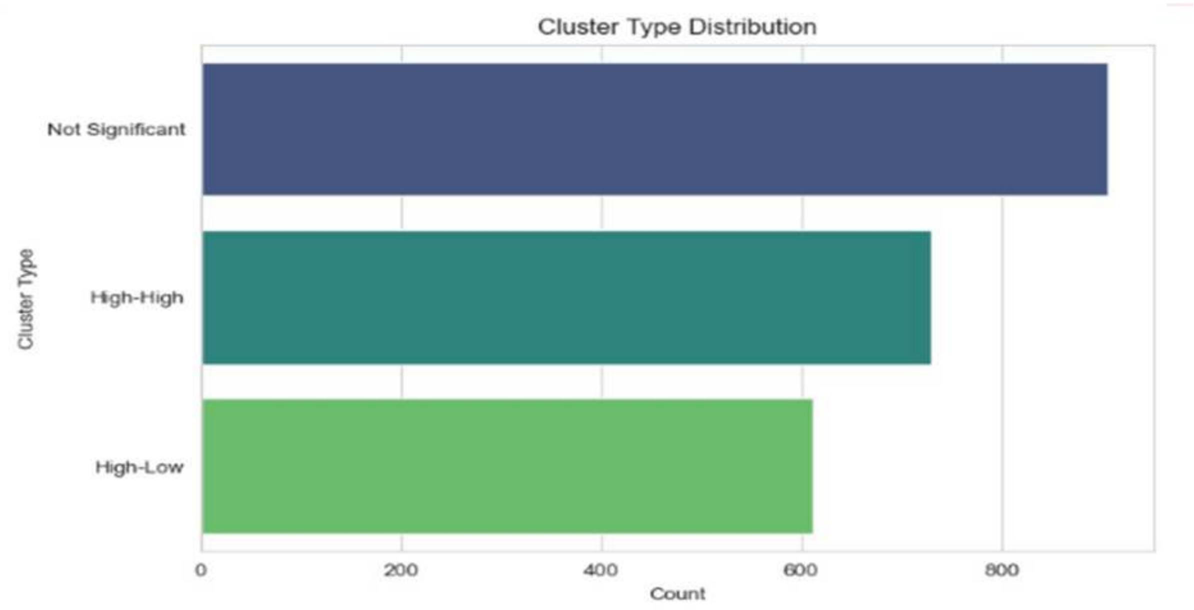


Figure 1:Bar Chart of Cluster Type Spatial Distribution

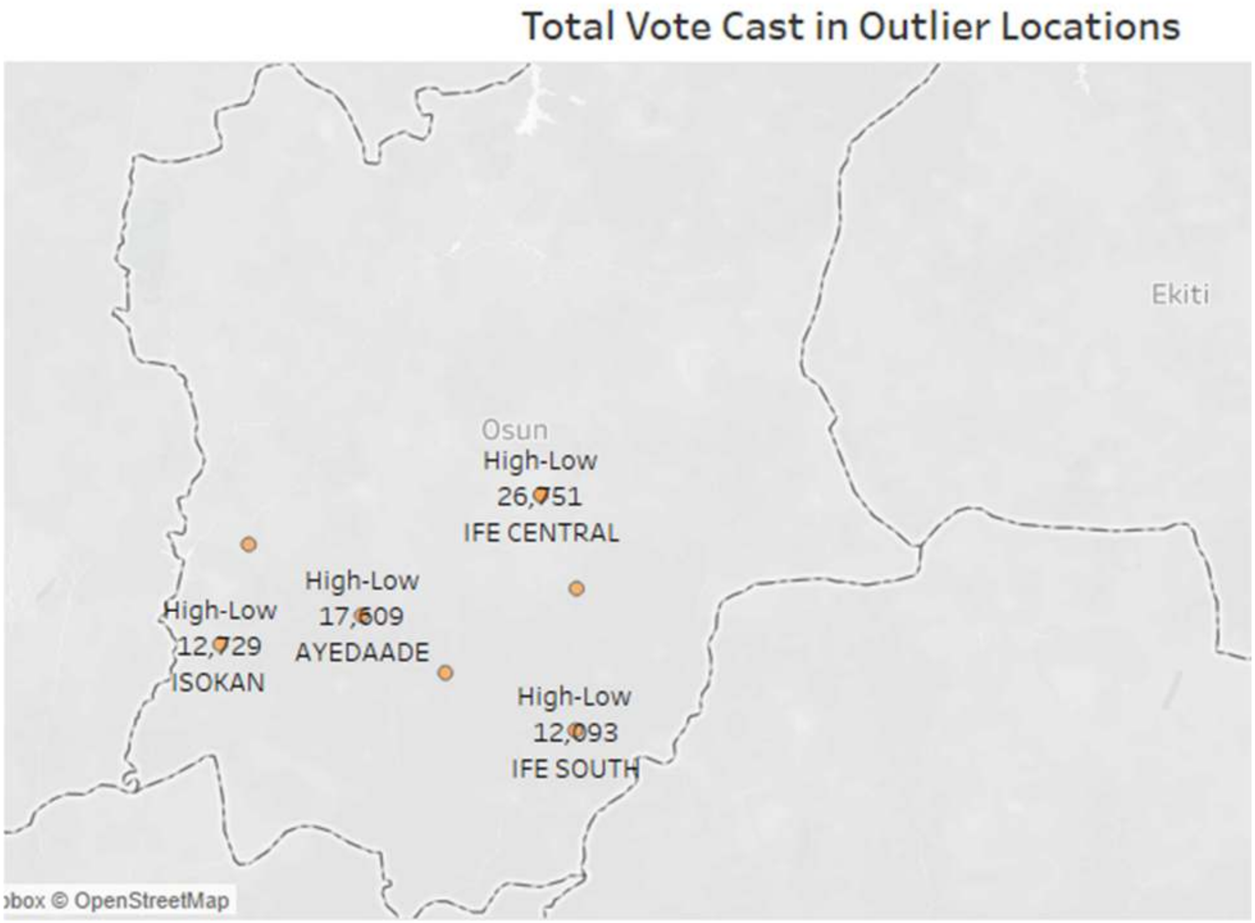


Figure 2:Map Distribution of Cluster Type Spatial Distribution

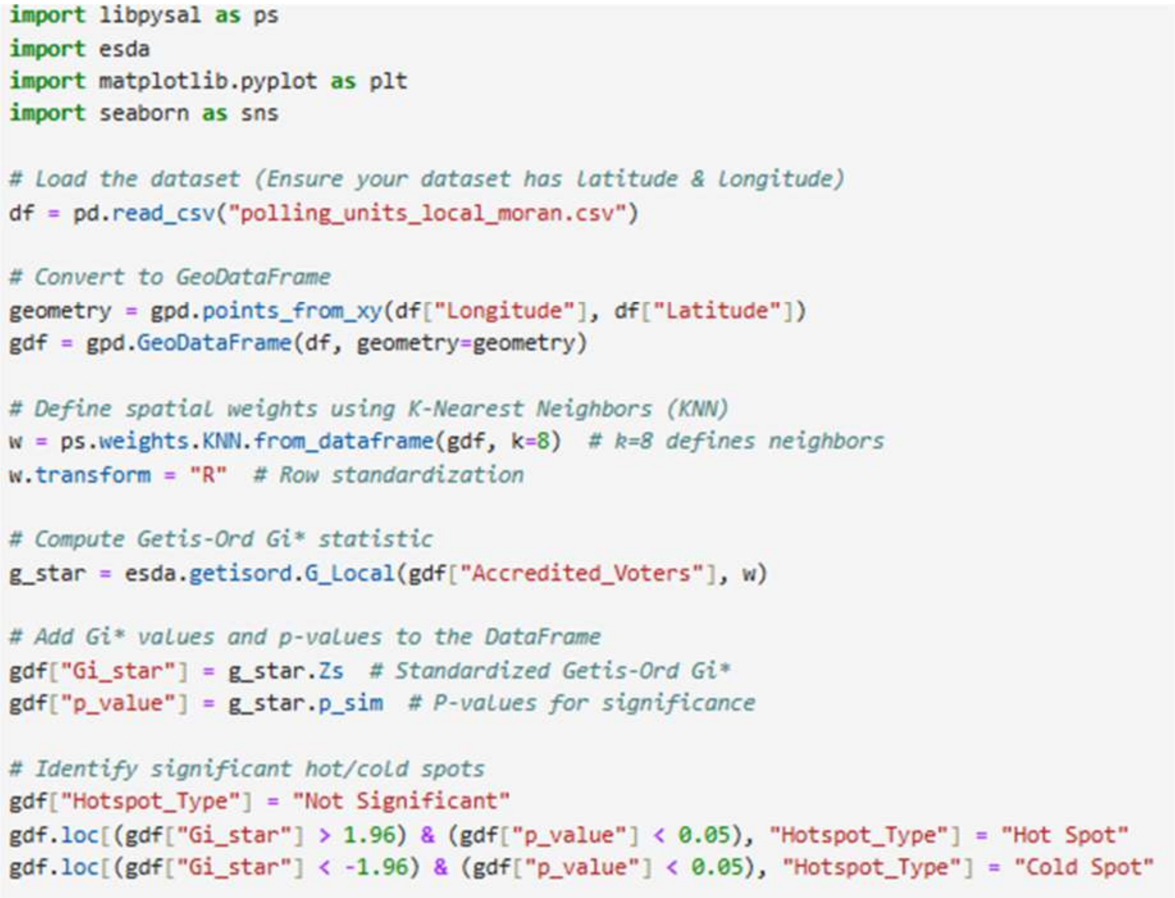
- Further analysis also revealed that the average total votes scored by each party at the outlier locations were not significantly different from the average votes scored by each party at non-outlier polling clusters. **(See table below).** This helps to disprove the hypothesis that the location of clusters at outlier locations favored one party over the other, thereby improving the credibility of the poll results.

Table 1: Average Party Votes in Outlier and Non-Outlier Clusters

PARTY	Non-Spatial Outliers Average Vote	Spatial Outliers Average Vote
APC	101.8	159.5
LP	80.7	48.1
PDP	190.8	208.3
NNPP	4.4	2.3

OVERALL	190.8	207.5

- **Getis-Ord Gi***: A hotspot analysis was performed to identify significant vote concentrations using the appropriate Python code.
- This analysis helps to identify areas with voting patterns that differ significantly from other areas or their adjoining polling units/ clusters.



```
import libpysal as ps
import esda
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset (Ensure your dataset has latitude & longitude)
df = pd.read_csv("polling_units_local_moran.csv")

# Convert to GeoDataFrame
geometry = gpd.points_from_xy(df["Longitude"], df["Latitude"])
gdf = gpd.GeoDataFrame(df, geometry=geometry)

# Define spatial weights using K-Nearest Neighbors (KNN)
w = ps.weights.KNN.from_dataframe(gdf, k=8)  # k=8 defines neighbors
w.transform = "R"  # Row standardization

# Compute Getis-Ord Gi* statistic
g_star = esda.getisord.G_Local(gdf["Accredited_Voters"], w)

# Add Gi* values and p-values to the DataFrame
gdf["Gi_star"] = g_star.Zs  # Standardized Getis-Ord Gi*
gdf["p_value"] = g_star.p_sim  # P-values for significance

# Identify significant hot/cold spots
gdf["Hotspot_Type"] = "Not Significant"
gdf.loc[(gdf["Gi_star"] > 1.96) & (gdf["p_value"] < 0.05), "Hotspot_Type"] = "Hot Spot"
gdf.loc[(gdf["Gi_star"] < -1.96) & (gdf["p_value"] < 0.05), "Hotspot_Type"] = "Cold Spot"
```

- The result of the analysis revealed that no polling units or cluster had a significant indication of heavy voting or voter suppression with respect to adjacent clusters/polling units.
- This result further improved the credibility of the election results being analyzed.

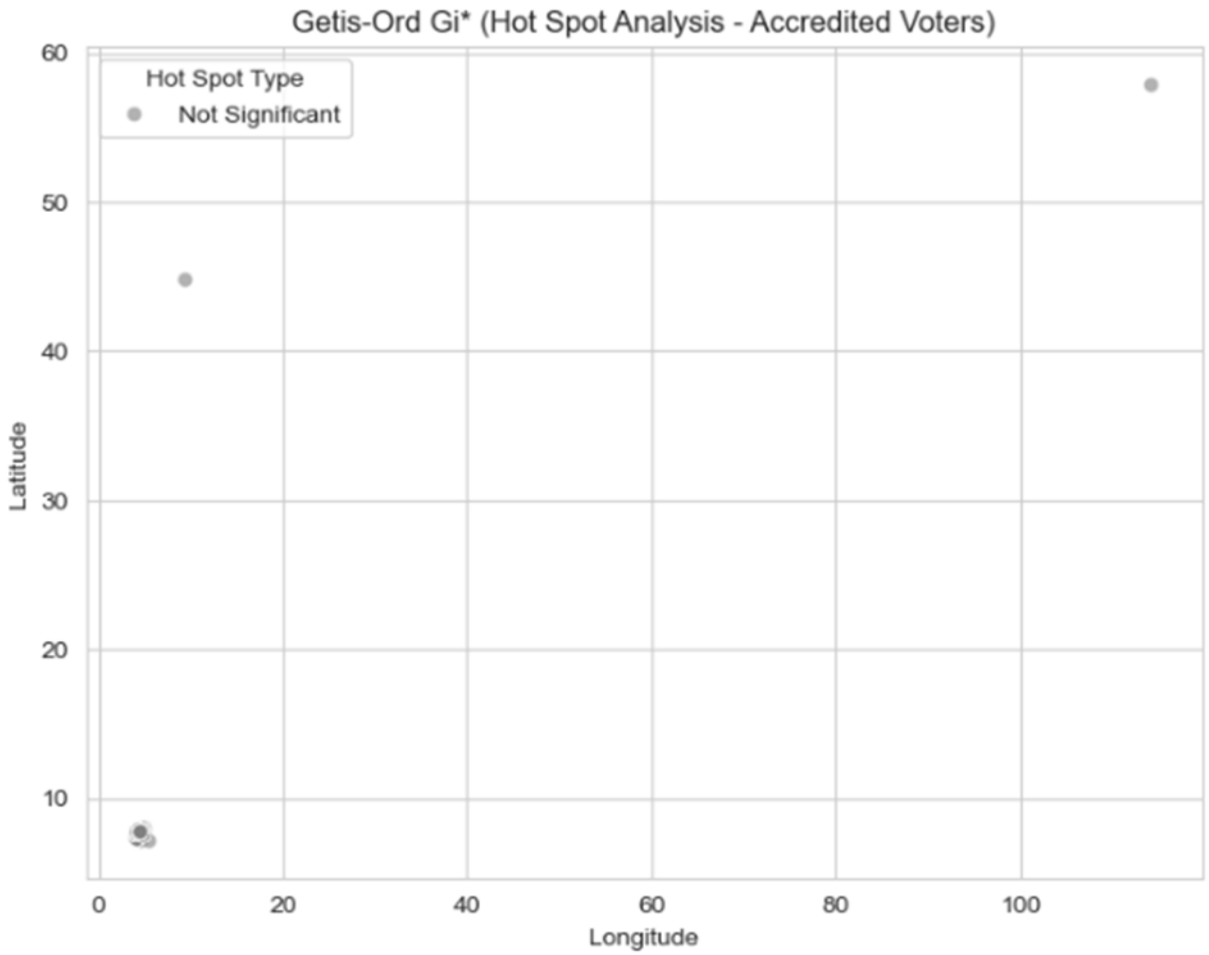


Figure 3: Chart of Hotspot Analysis Result

- **Machine Learning Validation**: Isolation Forest was used to cross-validate statistical findings and detect additional anomalies.

- The sum of votes cast, registered voters, and accredited voters alongside location coordinates were used to perform the analysis.

- The result of the analysis identified 113 polling units as anomalous and having significant voting results deviation from the normal. However, voting results from 27 of the identified location shows that no voting occurred there.

- This essentially implies that anomalous result was suspected in 86 polling units spanning 23 local governments or 21 clusters.

- Integration of the results of the Getis-Ord hotspot analysis and the Isolation Forest algorithm was then used to extract information on the top 5 Polling Units with significant deviation in their vote result.

- Table 2 below shows the top 5 outliers alongside their coordinates, Getis-Ord score, and other relevant parameters.

Table 2: Top 5 Outliers and Relevant Parameters

LGA	Ward	PU-Name	Latitude	Longitude	HDBSCAN_ Cluster	Gi_star	Anomaly_ Score	Anomaly_Label
EJIGBO	IFEODAN 'B'/MASIFA	OGURO/IFE ODAN /MASIFA JUNCTION	7.90292	4.31419	9	1.22432	-1	Anomalous
EJIGBO	ELEJIGBO 'B'/OSOLO	MOYOFADE COM. BANK	7.90292	4.31419	9	1.22383	-1	Anomalous
EJIGBO	ILAWO/ISOKO/ ISUNDUNRIN	ISUNDUNRIN BAPT. DAY SCHOOL,	7.90292	4.31419	9	1.22371	-1	Anomalous
EJIGBO	ELEJIGBO 'D'/EJEMU	OLOWOLA/OLOGBIN	7.90292	4.31419	9	1.22252	-1	Anomalous
EJIGBO	ELEJIGBO 'C'/MAPO	BEULAH BAPT. SCHOOL, EJIGBO	7.90292	4.31419	9	1.22195	-1	Anomalous

2.5 Temporal and Demographic Analysis

- Historical election data was compared to detect sudden deviations in voting trends.

- Due to the unavailability of relevant data, temporal analysis was performed using only the immediate past election result.

- The result of the historical evaluation showed that while the two most dominant parties (APC and PDP) have lost supporters and or voters between 2019 to 2023, the other two parties (LP and NNPP) have increased their support base.

- LP has improved its voting strength in the state by over 4500%, while the NNPP has increased theirs by about 750%.

Bar Chart showing Historical Trend in Osun State Presidential Election Result

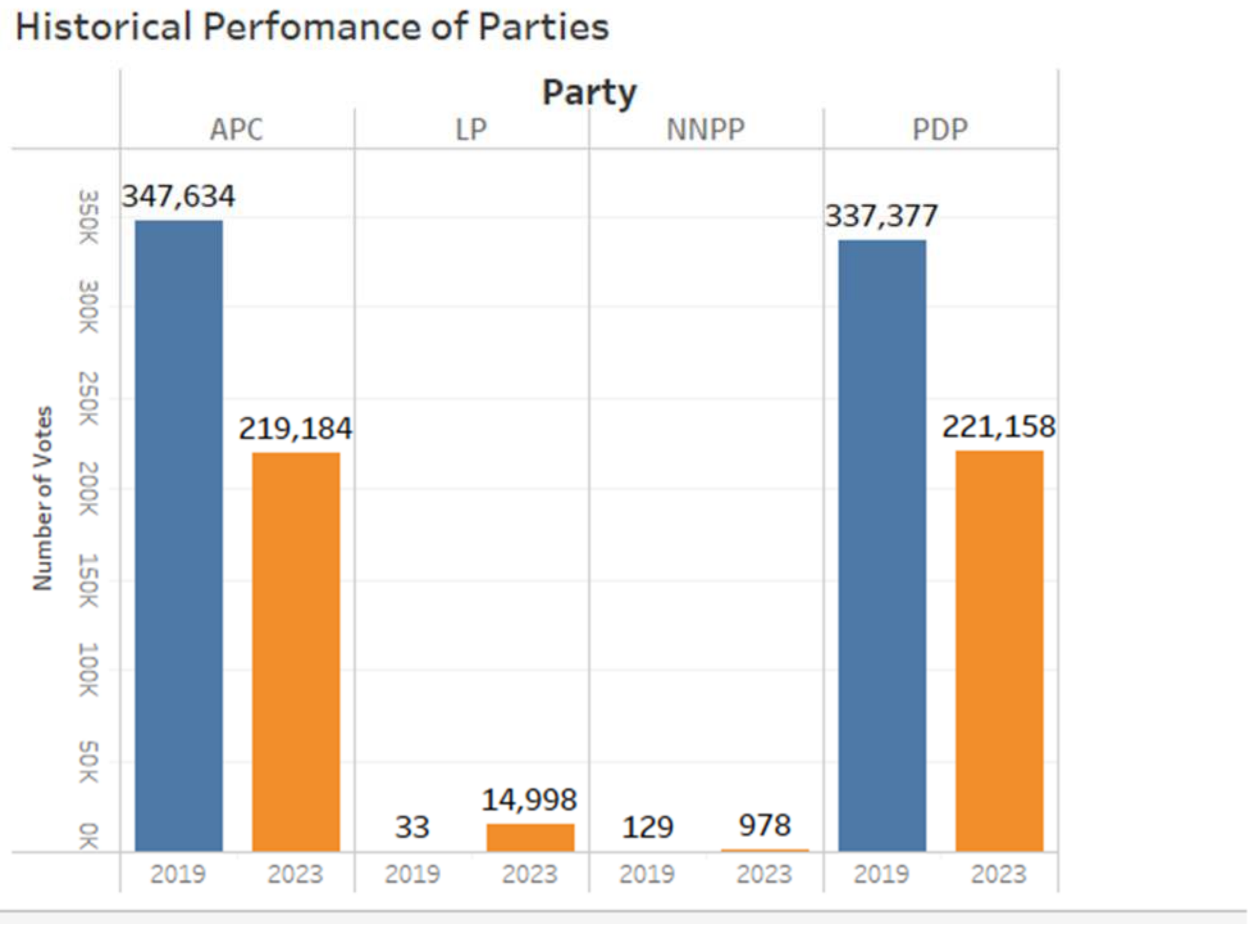


Figure 4

3. Recommendations

- **Voter Turnout/Apathy**: There appears to be a general level of malaise or apathy among the electorates. This is evident in the wide disparity of accredited voters and the sum of votes cast in all of the local governments. Public institutions and democratic organizations are enjoined to seek out public opinion to investigate the cause of voter apathy. Such root causes should be tackled to encourage public participation in the process and improve the robustness of the elections.

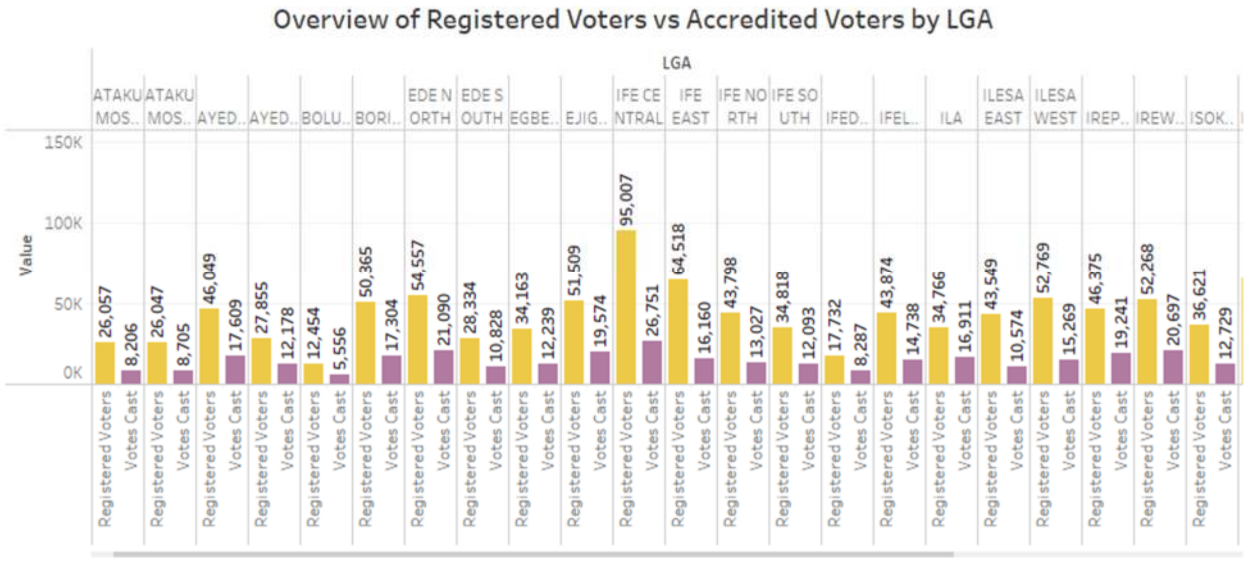


Figure 6: Chart Comparing the Number of Registered Voters to Sum of Votes Cast

- **Enhanced Monitoring**: Deploy independent observers to flagged polling units in future elections, especially areas where anomalies are suspected
- **Data Transparency**: Implement open-access election data dashboards for public scrutiny.
- **Refined Voter Registration**: Conduct audits to ensure accuracy in voter registries.

4. Conclusion This report presents a comprehensive spatial and statistical approach to detecting electoral anomalies. The findings should be used to improve electoral transparency and safeguard democracy. Further validation and continuous monitoring are recommended.