**HNG DATA ANALYST INTERNSHIP STAGE 1 TASK**

**Report on Outlier Detection in Election Data Using Geospatial Analysis**

**Selected state:** Anambra

**Tools:** Pandas, Geocode by Awesome Table, Math, Matplotlib, Seaborn, and KDTree

**1.0 Introduction**

In the 2023 presidential election, the Independent National Electoral Commission (INEC) faced multiple legal challenges concerning the integrity and accuracy of the election results. Allegations of vote manipulation and irregularities have been widespread, prompting a thorough investigation into the matter. This report aims to uncover potential voting irregularities by identifying outlier polling units where the voting results deviate significantly from neighboring units. The analysis uses geospatial techniques to find neighboring polling units and calculate an outlier score for each party in each unit.

**2.0 Methodology**

**2.1 Dataset Preparation**

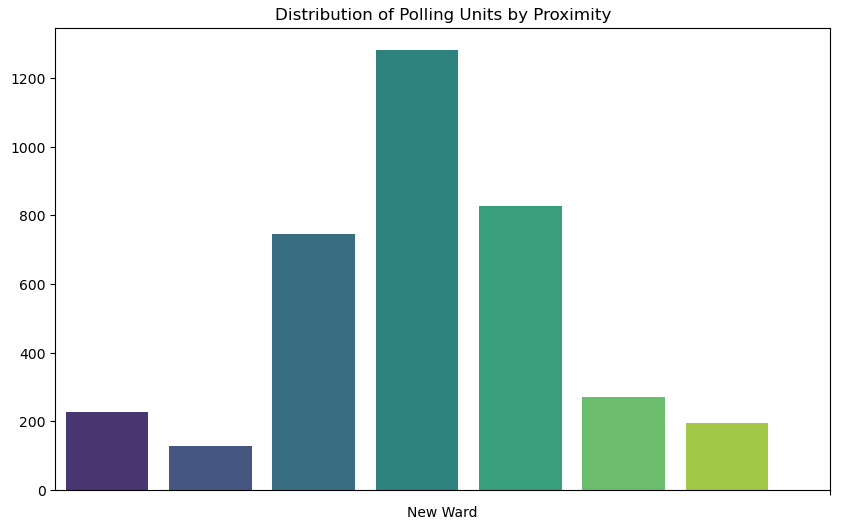
**1. Data Inspection:** The dataset had 19 rows and 3600 columns. Columns not needed for the analysis were dropped while columns containing location details were concatenated in the appropriate order to form a new column `address`, This new column is needed for determining the longitude and latitude of the polling units.

**2. Geocoding:** Using a Google Sheets add-on, Geocode by Awesome Table, the geo-locations of the polling units were obtained were obtained by passing their addresses to Geocode in 4 batches, 1000 addresses per session.

**3. Distance Calculation**

Polling units were grouped into clusters based on the distance bins. The distance between polling units was calculated using the Haversine formula, which considers the curvature of the Earth.

**4. Clustering:** Proximity radius of 10km was defined. Polling units were clustered based on this geographical proximity using predefined distance bins (0-10 km, 10-20 km, etc.).

**2.2 Nearest Polling Unit Identification**

The nearest polling unit was retrieved by using a KDTree data structure to efficiently find the nearest polling unit for each polling unit based on their geographical coordinates. I obtained the indices of the nearest neighbors which are then used to add a new column to the DataFrame.

**2.3 Outlier Score Calculation**

For each polling unit, the votes received by each party were compared with those of neighboring units within the same cluster. The outlier score for each new ward (cluster) was calculated using the Interquartile (IQR) method, which measures how far a party's votes deviates from the cluster’s median.

The Interquartile Range was computed as IQR = Q3 – Q1.

A lower bound was established as lower bound was established as:

Lower Bound = Q1 – 1.5 x IQR

Upper bound:

Upper Bound = Q3 + 1.5 x IQR

The Votes falling below the lower bound or above the upper bound were considered outliers. The outlier score for each party becomes how much each party’s vote count deviates from the clusters mean.

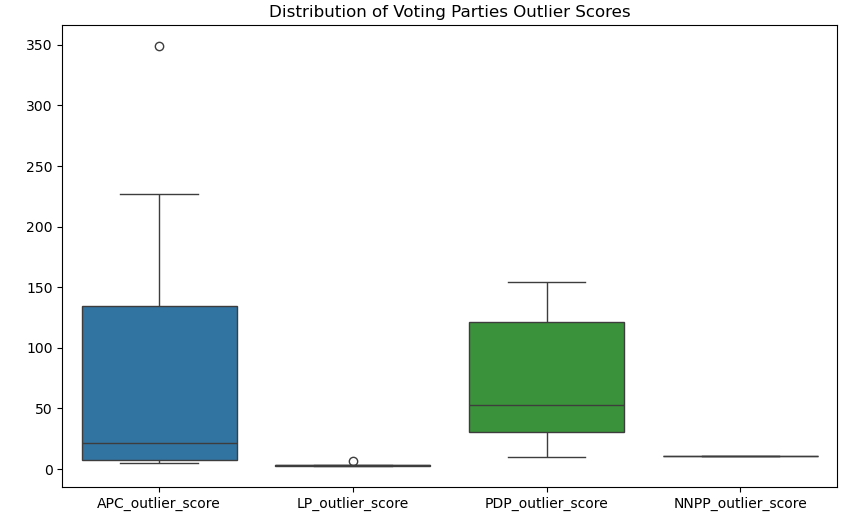
**2.4 Sorting and Reporting**

The dataset was sorted in descending order by the outlier scores for each party to identify the most significant outliers. The top 3 outliers and their closest units were highlighted, and explanations were provided for why these were considered outliers.

**2.4.1 Top 3 Voting Outlier Scores**

| **Outlier Polling Unit** | **Closest Polling Unit** | **Outlier Score** | **Party** |
| --- | --- | --- | --- |
| AKABO HALL I, NNEWI SOUTH, ANAMBRA | ISIGWU HALL I, NNEWI SOUTH, ANAMBRA | 348.7604 | APC |
| NKWO AMENYI SQUARE II, AWKA SOUTH, ANAMBRA | EBE DUNU SQUARE, AWKA SOUTH, ANAMBRA | 226.456 | APC |
| AKWA VILLAGE HALL III, DUNUKOFIA, ANAMBRA | ILOUGWU VILLAGE HALL I, DUNUKOFIA, ANAMBRA | 146.456 | APC |

**2.4.2 Visualizing Outlier Scores Distribution**



**2.4.3 Explanation of Outliers**

These votes are regarded as outliers because the polling unit had significantly higher votes compared to its neighboring units. The standard deviations of the party votes are outrageously above the mean of the cluster, raising suspicions of vote inflation for the political party.

**2.5 Conclusion**

The analysis identified significant outliers in the election data, highlighting potential voting irregularities. The identified outliers showed substantial deviations from the average voting patterns within their clusters, indicating potential vote concentration, suppression, or inflation. These findings provide valuable insights into the integrity of the election results and can aid in further investigations to ensure transparency and accuracy.