Data Analytics Track – Stage 2B Task: Outlier Detection in Election Data Using Geospatial Analysis

Case Study: Ensuring Election Integrity

Dear Intern,

Welcome aboard once again! In the recently concluded election, the Independent National Electoral Commission (INEC) has 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.

Your mission, should you choose to accept it, is to help us uncover potential voting irregularities and ensure the transparency of the election results. You will achieve this by identifying outlier polling units where the voting results deviate significantly from neighbouring units, indicating potential influences or rigging. Your keen analytical skills and attention to detail will be critical in this endeavour.

Task Overview

In this stage, you will focus on identifying outlier polling units based on the votes each party received. The analysis will involve geospatial techniques to find neighbouring polling units and calculate an outlier score for each party in each unit. The goal is to pinpoint polling units where the voting results significantly deviate from their neighbours, indicating potential irregularities or influences.

Task Objectives

1. Dataset Preparation:

- Open the drive and find your state of origin. If you are not a Nigerian, pick a random state.

- Download the `{YOUR\_SELECTED\_STATE}\_crosschecked` spreadsheet or CSV file.

- If your selected dataset does not include longitude and latitude values for each polling unit or ward, use geocoding techniques to obtain them.

2. Neighbour Identification:

- Identify neighbouring polling units based on geographical proximity. Define a radius (e.g., 1 km) to determine which units are considered neighbours.

3. Outlier Score Calculation:

- For each polling unit, compare the votes each party received with those of its neighbouring units.

- Calculate an outlier score for each party based on the deviation of votes from neighbouring units.

- Record the outlier scores along with the respective parties and neighbouring units.

4. Sorting and Reporting:

- Sort the dataset by the outlier scores for each party to identify the most significant outliers.

- Provide a detailed report explaining the methodology and findings.

- Highlight the top 3 outliers and their closest polling units, explaining why they are considered outliers.

Deliverables

1. Cleaned Dataset:

- A CSV file with added longitude and latitude values for each polling unit or ward.

- The dataset should include calculated outlier scores for each party in each polling unit and the neighboring units.

- An excel sheet showing a sorted list of polling units by outlier scores for each party.

2. Outlier Detection Report:

- A detailed report containing:

- Explanation of the methodology used for geospatial analysis and outlier detection.

- Summary of findings, including the sorted list of polling units by outlier scores for each party.

- Detailed examples of the top 3 outliers, with explanations and visualizations.

- A conclusion summarizing the findings, including key insights and visualizations.

Submission

A Google Drive Folder containing:

Your CSV file with the added longitude and latitude values should be accessible through a hosted link. You can use Google Sheets or similar.

Your EXCEL sheet showing a sorted list of polling units by outlier scores for each party should be accessible through a hosted link. You can use Google Sheets or similar.

Your report in PDF should be accessible through Google drive.