**Introduction:**

For this project we investigate the legally contested election results in Nigeria due to the questionable data integrity. Our purpose is to find out potential arears of irregularities through the available election data. This we will set out to achieve through highlighting outlier polling units whose results deviate significantly from their neighbouring units. This will serve as a pointer to possibility of rigging and undue influences.

The rest of the report will present the dataset, methods, analysis, findings and conclusions from our data cleaning, exploration and visualizations.

**Data:**

* **Data Cleaning/Preparation**:

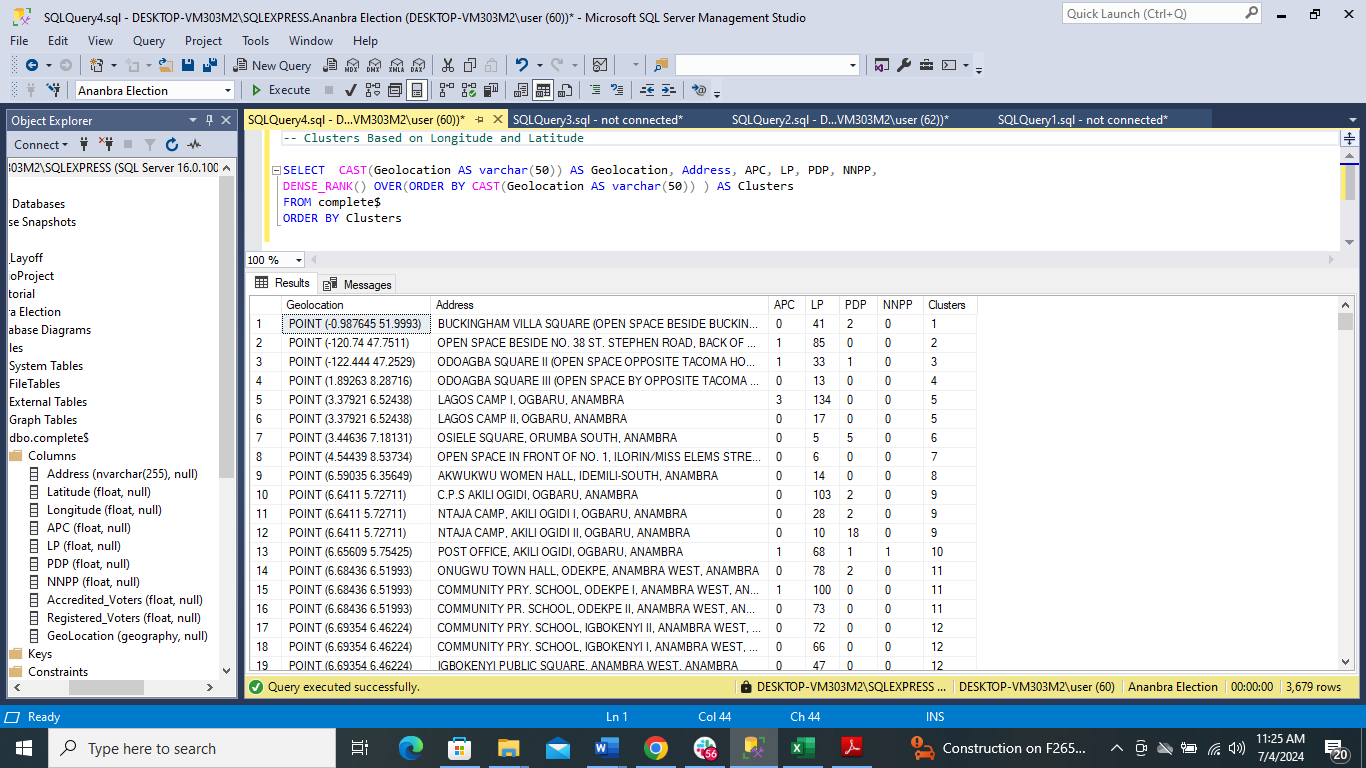
The raw data set was extracted from [drive link](https://drive.google.com/drive/folders/173oHgms6wYy5WKz_i3Lhl5mXcmobCWHz) . A Google API -Geocode Awesome Table was used to get the missing longitude and latitude of the polling units for the purpose of identifying neighbouring polling units and distances. Some columns were combined to get complete addresses in order to extract the geocodes. The dataset was cleaned for duplicates, unnecessary columns and rows, nulls and blank spaces as well as errors using Microsoft Excel. The clean dataset can be viewed in the sheet labelled CLEAN DATA in the attached spreadsheet.

**Methods:**

The dataset was then loaded into Microsoft SQL Server Management Studio (SSMS) where the longitude and latitude were converted into the geography data type (named Geolocation) to enable compute the distances. SSMS was used for the rest of the exploratory data manipulations and analysis.

At the end of the exploration, outputs were exported the Excel spreadsheet for final sorting and visualization.

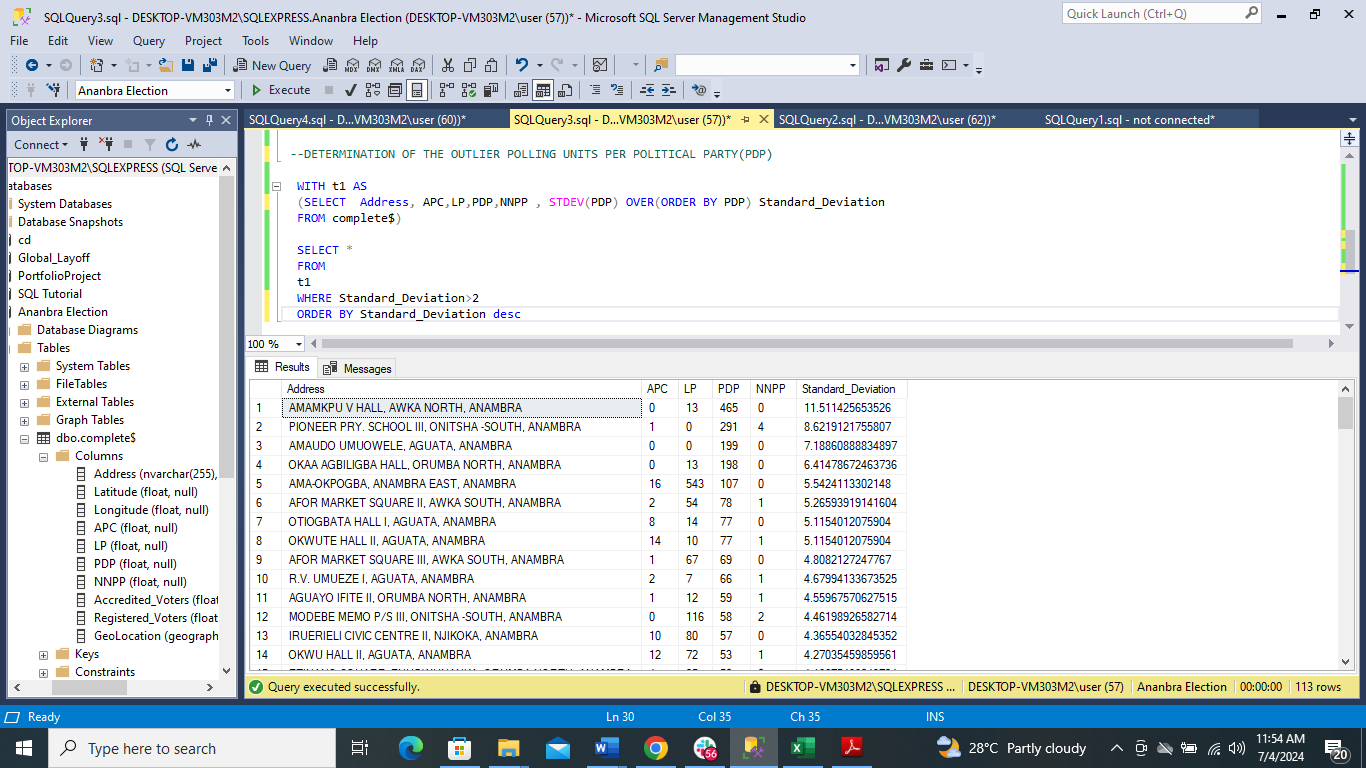
* **Clustering**: Clusters of neighbouring polling units were created in SQL based on longitudes and latitudes as shown in the sheet labelled CLUSTERS. The clusters are numbered with dense rank to make them distinct. A snippet of the procedure used can be seen below:



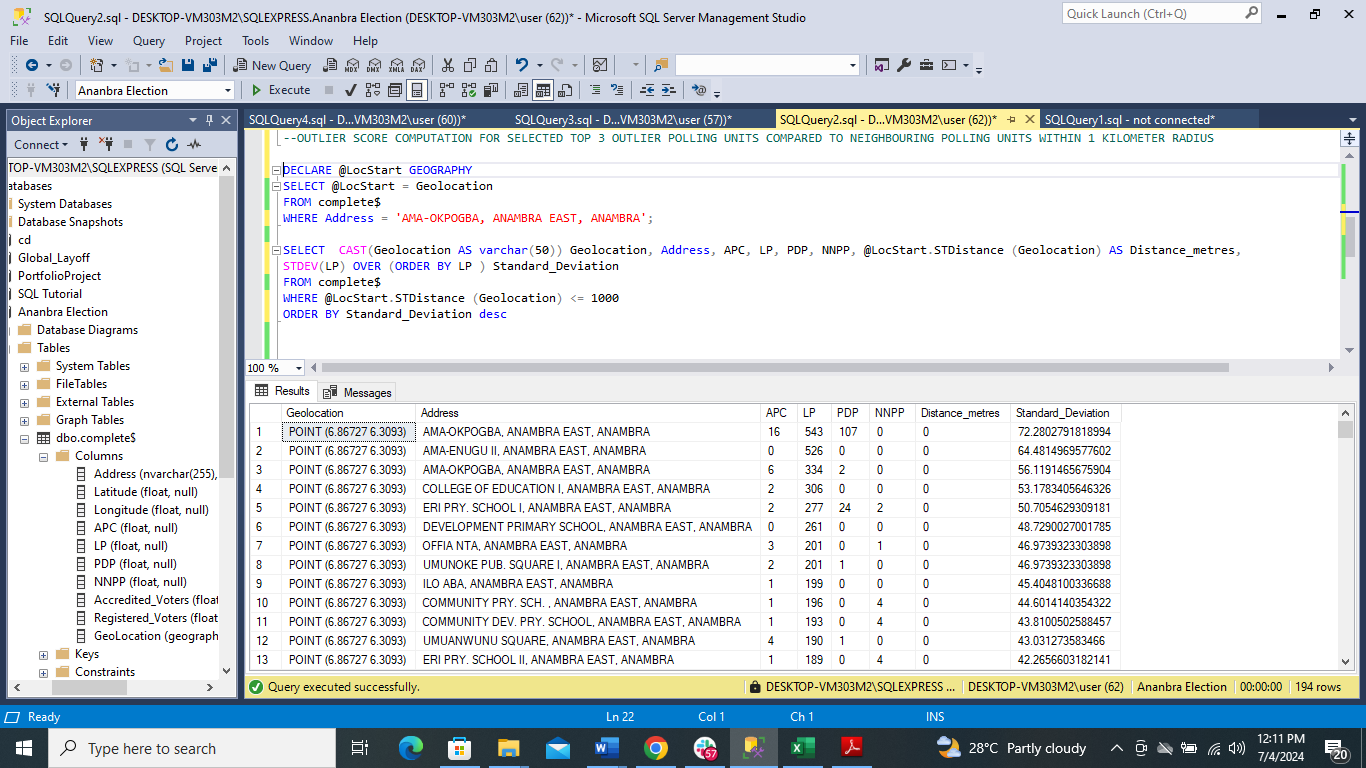
* **Outlier Score Calculation**:

Preliminary analysis of the dataset shows that LP had an overwhelming majority and this can be seen from the chart below. Hence deviations from this trend would be considered out-of-place.

Standard deviation was used as the measure of the outlier score. Outlier polling units were selected based on the polling units per political party where standard deviation is greater than 2 (STDEV >2). Below is a snippet of the procedure:



The outlier score computation was done with standard deviation to determine the extent of deviation of selected outlier polling units from neighbouring units within 1 km radius as per below procedure. The parameter for selecting the top 3 outliers was the polling units with the highest standard deviations.



**Summary of Findings**

From the overall dataset explored, it shows that LP led in majority of the polling units. The acceptable Standard deviation maximum of 2 was used to show that up to 95% of the score are close to the mean of the polling units that conform to the established trend of LP winning the election in Anambra state.

The other polling units with standard deviations greater than 2 were considered outliers. When further compared with neighbouring polling units within 1km radius, their outlier status was further established with their standard deviation being the highest among its neighbours.

The top 3 outliers for each political party top 3 highest standard deviation scores (outlier score). For example, there is clear case of over-voting among the top 3 LP outliers where the total number of votes () is higher than the accredited voters in Ama-Okpogba, Anambra East, Anambra from the dataset.

Conclusions:

According to the findings we can safely say that LP had a landslide victory. Hence polling units where they were extremely outclassed could be contested. Nevertheless, the incidence of over-voting identified in an LP outlier polling unit in Ama-Okpogba, Anambra East, Anambra, calls to question the results from the LP outlier polling units.