Data Cleanup Strategy

The following cleanup and data preprocessing was done in Google Colab.

[1] Data Sourcing: Data was obtained from the link shared on the portal which contains the data of Election State Wise and Year Wise. Each link of particular year of the state included the columns as **“Unnamed: 0”, “ST_NAME”, “YEAR”, “AC”, “CANDIDATE”, “SEX”, “AGE”, “CATEGORY”, “PARTY”, “VOTES”.**

This was done using “**requests**” and “**BeautifulSoup**” Library which was later saved in .json format for further processing.

[2] Data Preprocessing for Use: The Data then obtained was used to scrape the data of the election State Wise. A separate json file was created for each year for each state. For example, Bihar had multiple json files year wise.

The Data of the required state STATE_NAME **“GUJARAT”** was then filtered from this data and stored in another folder. Similarly from the .json files of Gujarat the data of the required constituency i.e. AC_NAME **“BOTAD”** was extracted yearwise and stored in separate folder for further use.

[3] Cleaning the Data: Before extracting constituency data, state data was cleaned by:

[3.1] Removing the Duplicate Entries: The State data was checked for the duplicate entries file by file and then the entries were removed by ***‘df.drop_duplicates’,*** which kept the first entries of the duplicated data and cleaned the rest.

[3.2] Dropping the Rows with Invalid Votes (NaN, Null): The Data was then checked for and entries with invalid vote values were removed by **‘df.dropna()’** method. Also **‘pd.to_numeric()’** was used to convert values in the votes column to numeric values.

[3.3] Removing Irregularities in the Constituency Name: The Data column **“AC_NAME”** was checked and the leading and trailing whitespace and digits were removed for further use. This was implemented on State level data. For example. **“49_BOTAD”** was converted to **“BOTAD”.**

Summary

To Summarise, the data given was first scraped and stored in .json files year wise. Then the required state data was extracted and checked for duplicates and null values (Votes Column) and the rows which had these were dropped keeping the first one. The votes were converted to numeric. Finally the data was checked for irregularities and constituency names were cleaned.