Apache Airflow-Orchestrated Polling Intelligence Pipeline – Data Engineering Project

Adhish Nanda
MSc. Data Science
Univ. of Europe for Applied Sciences
Potsdam, Germany
adhish.nanda@gmail.com

***Abstract*—This project implements an automated batch data pipeline using Apache Airflow to ingest, clean, transform, and analyze election polling data. The pipeline creates analytical data marts and machine-learning-ready feature tables and executes downstream batch prediction workflows to generate candidate ranking reports and visualization artifacts. A Streamlit dashboard was developed to present the pipeline outputs interactively.**

This project demonstrates real-world data engineering practices including workflow orchestration, modular ETL design, and reproducible batch processing.

I. Introduction

The electing process in the United States, being fundamental to any democracy, can be summarized into a set of sequential events that determine public offices, including the position of President of the United States. Events include the primaries, national conventions, general elections, and the Electoral College-the building blocks critical in the conduct of selecting the nation's leadership.

Data engineering and machine learning in recent years have merged to revolutionize election forecasting. The "Polling Data Aggregator and Election Outcome Predictor" project is a perfect example of such advancement in systematically collecting, cleaning, and analyzing polling data from multiple reputable sources. It uses sophisticated algorithms to make real-time predictions about the outcome of an election, offering deep insights into the pulse of the electorate and the dynamics of electoral competition. This method will enhance forecast accuracy and add to an informed electorate.

II. Literature Review

The precision and dependability of election surveys have long been an issue of much deliberation, with regard for presidential elections in the United States. A systematic review by Panagopoulos and Tam Cho (2014) [1] examines various approaches that have been used to summarize polling data in order to project electoral outcomes; it assesses their effectiveness. The research shows pooling techniques—basically averaging or using regression models—to be extremely important in trying to improve prediction accuracy. The discussion also brings forth the trade-offs of alternative aggregation methods, methodological transparency, and possible sources of bias. While such aggregated polling data can be insightful, the pooling method applied importantly influences the accuracy of electoral prediction.

In a more recent analysis, Erikson and Wlezien (2023) [2] reexamine the performance of the polls in the 2020 U.S. presidential election and explore the implications for the next 2024 electoral cycle. The authors note that, despite numerous criticisms, polls have generally provided quite accurate predictions of election outcomes in the past. This study underlines the dynamic nature of polling accuracy, which is influenced by factors such as changes in voter behavior and the polling methodology applied. The authors call for a continuous investigation and methodological improvement to increase the reliability of pre-election polls, especially in view of the dynamic political landscapes.

These studies collectively underscore the critical role of methodological rigor in polling data aggregation and election outcome prediction. They highlight the necessity for ongoing evaluation and refinement of polling techniques to adapt to changing voter behaviors and ensure the accuracy of electoral forecasts.

III. Interesting Questions Analyzed

- Who are the top candidates based on their average polling percentages (overall and by state)? [Rank different candidates in each state]
- How do polling percentages vary between states for key candidates (e.g., Donald Trump, Kamala Harris)?
- What is the relationship between sample size and polling percentage accuracy for candidates?
- How has polling for specific candidates (e.g., top 3 candidates) evolved over time?
- Which polling organizations (pollsters) have the highest transparency scores, and how does it correlate with their poll results?

IV. Problem Statement

Accurate prediction of election outcomes is a difficult task, made more so by the fluid nature of voter sentiment, the multitude of different polling methodologies, and varying levels of reliability among data sources. Traditional methods have trouble effectively incorporating real-time data, which tends to result in forecasts that don't perfectly align with current trends. A robust set of mechanisms must be in place for consolidating polling data from various credible sources, purifying and standardizing the acquired information, and using state-of-the-art

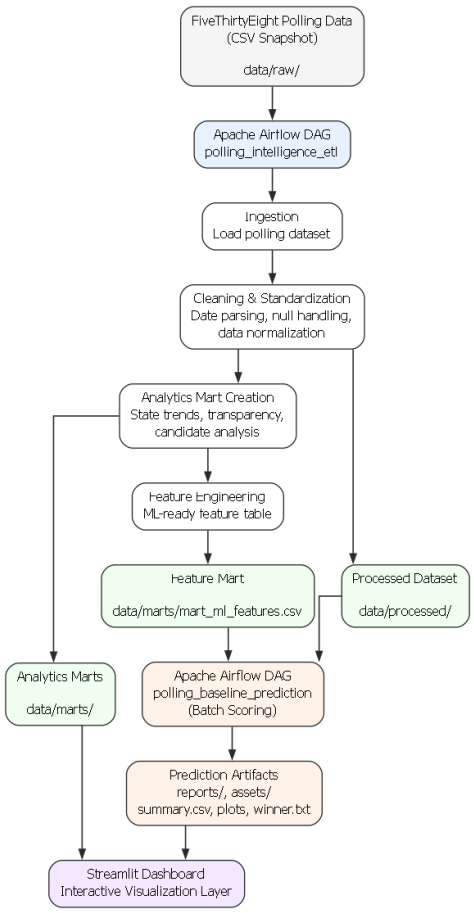

Fig. 1: Apache Airflow-orchestrated polling intelligence pipeline showing ETL workflow, batch prediction, and Streamlit visualization layer.

machine-learning methods for real-time, accurate predictions of electoral outcomes. Implementation of such a system will yield substantial insight into patterns of polling and electoral dynamics, hence empowering stakeholders to make informed, up-to-date decisions based on the latest available data.

V. Novelty of the Study

The project titled "Polling Data Aggregator and Election Outcome Predictor" is characterized by several pioneering features

- **Holistic Automation:** In contrast to conventional static polling analyses, this initiative automates the entire process—from data acquisition to prediction—thus guaranteeing ongoing updates and real-time insights
- **Incorporation of advanced technologies:** Through the integration of web scraping, data pipelines, and machine learning models, the project presents a cohesive workflow that adjusts to newly available data, thereby improving the precision of predictions.
- **Real-Time Data Processing:** The ability of the system to process and analyze data as it is generated provides stakeholders with the latest insights into the dynamics of elections.
- **Interdisciplinary Approach:** Combining political analysis with data engineering and machine-learning methodologies, this project shows a high level of technical and interdisciplinary innovation.

VI. Significance of Our Work

The project "Polling Data Aggregator and Election Outcome Predictor" is of great importance to the fields of political analysis, data engineering, and civic engagement.

- **Improving Predictive Precision:** By aggregating polling data from multiple credible sources and applying advanced machine learning methods, the project goes one step further in the search for better accuracy in electoral prediction. It handles the inherent uncertainties and possible biases of single polls, resulting in more robust predictions.
- **Real-Time Insights:** Automated data pipelines ensure that the system provides up-to-date analysis on voter sentiment and electoral dynamics. Real-time analysis capabilities like this are extremely important for stakeholders who need information quickly to make very informed decisions during election cycles. that the system provides up-to-date analysis.
- **Educational Significance:** The project serves as a practical application of data engineering and machine learning methods and, by nature, gives wide-ranging educational experiences to both students and professionals in the field. It depicts how technologies like web scraping, data pipelines, and predictive modeling are efficiently used within a real-world scenario.
- **Public Engagement:** It increases public knowledge of the electoral process by making predictive and polling trends accessible through intuitive dashboards, thus empowering the public with insights from data for informed participation in democratic processes.
- **Advancing Research:** The large dataset and analytical capabilities of the system form a foundation for further academic exploration of polling methodologies, voting behavior, and factors influencing electoral outcomes. This feeds into the wider field of political science and the study of public policy.

VII. Methodology

A. Dataset and Data Collection

The dataset used in this project consists of presidential polling data obtained from FiveThirtyEight dataset exports, consisting of presidential polling data collected across various states in the United States for the 2024 election cycle. It contains information on different candidates, their polling percentages, polling organizations, sample sizes, and transparency scores. The dataset provides insights into voter preferences and polling trends over time. Gathered Data (shown in Fig. 19) from reliable source like FiveThirtyEight. It captures both

national and state-level polls to ensure a comprehensive collection. Key attributes include:

- State – The U.S. state where the poll was conducted.
- Candidate Name – The political candidate for whom the poll results were recorded.
- Polling Percentage (pct) – The percentage of respondents supporting a particular candidate.
- Pollster – The organization conducting the poll.
- Sample Size – The number of respondents in the poll.
- Start and End Date – The time frame during which the poll was conducted.
- Transparency Score – A rating indicating the reliability of the polling data.

Apache Airflow orchestrates the data ingestion and processing pipeline, ensuring reproducible batch execution. Airflow Directed Acyclic Graphs (DAGs) manage task dependencies across ingestion, cleaning, feature engineering, and prediction stages.

B. Data Cleaning and Preprocessing

The raw data collected undergoes rigorous cleaning and treatment for inconsistent values, missing values, and biases inherent to the different methodologies used in polling. Techniques including imputation of missing data and normalization are applied in order to standardize the dataset.
Critical features are used to enhance predictive modeling, including polling averages, sample sizes, time-to-election metrics, and adjustments for known pollster biases. This ensures that the inputs to the model are relevant and informative.

C. Baseline Prediction

A statistical aggregation approach was implemented to estimate election outcomes based on average polling percentages. The prediction workflow was executed as a batch scoring task using Apache Airflow. The pipeline consumed the prepared feature dataset and generated ranked candidate summaries and visualization artifacts. This approach provides interpretable baseline predictions while demonstrating pipeline orchestration principles.

D. Visualization and Insights

The interactive dashboards on real-time polling trends, candidate trajectories, and forecasted electoral outcomes are created using Plotly or Matplotlib. The visualization of such complex data sets becomes intuitively understandable.
The system offers detailed insight into the national and state levels, which helps the user dig deep into the dynamics of the elections across different regions and demographic groups.

E. System Deployment

The whole system is developed using web frameworks like Flask or Dash, thus ensuring accessibility through web interfaces. This can be further scaled using cloud services to improve how computational needs are managed.

The overall approach goes a long way in ensuring that the project delivers real-time, accurate predictions of election outcomes through the integration of state-of-the-art data engineering techniques with intricate machine learning methodologies.

VIII. Results

A. Top Candidates by Polling Percentages

Key Insights:

- Donald Trump has strong support in many states, as indicated by the high presence of red cells in his row.
- Kamala Harris and Bernie Sanders also show some significant support in many states
- Other candidates, like Robert F. Kennedy, show a more spread-out pattern of support, which may indicate a lack of geographical consistency.

B. Top 3 Candidates by Polling Percentages Per State (Interactive Bar Chart)

Key Insights:

- Donald Trump consistently places among the top contenders in many states, reflecting his wide base of support.
- Kamala Harris and Joe Biden also frequently place in the top three, showing their competitive position.
- The differences in polling percentages at the state level show quite extensive regional differences in candidate preferences.

C. Overall Average Polling Percentages of Candidates

Key Insights:

- The top three contenders, ranked by average polling percentages, include Bernie Sanders, Kamala Harris, and Donald Trump.
- Less popular candidates, like Joseph Kishore and Joel Skousen, have a low level of overall support, evidenced by their much shorter respective bars.
- The chart really points out how there is a huge disparity in candidate popularity, shown by a large drop-off in average percentages after the top contenders.

D. State-wise Polling Percentages

Key Insights:

- The graphical form in this article focuses on how each candidate fared individually state by state.
- It eases comparing Donald Trump and Kamala Harris against each other from state to state in their poll performance.
- Some states, like Nevada, have significant differences in polling percentages between the two candidates.

E. Cumulative Polling Percentages by State

Key Insights:

- In states like Nevada, Donald Trump leads the cumulative percentage, while Kamala Harris does better in others like California.
- The visualization is effective in showing areas where the competition is tight versus states where one candidate leads by a large margin.

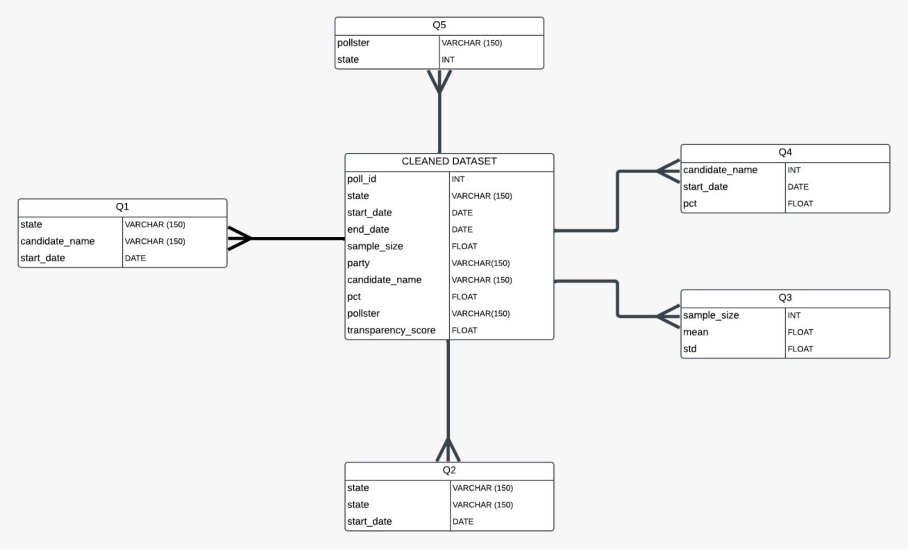

Fig. 2: ER Diagram of Polling Dataset

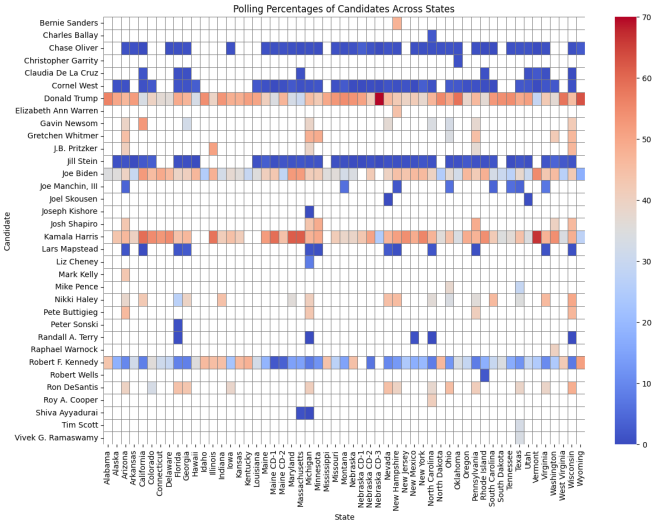

Fig. 3: Polling percentages of candidates across states.

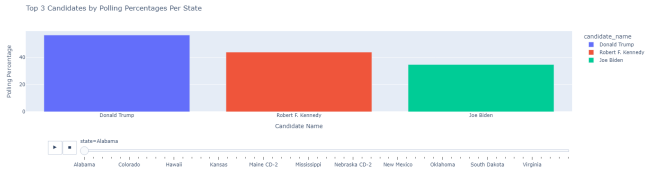

Fig. 4: Bar chart showing Top 3 Candidates Per State.

F. Trump vs. Harris Polling Percentages

Key Insights:

- Most points lie above the diagonal, indicating that in most states, Kamala Harris is leading Donald Trump.
- States closer to the diagonal represent tighter competition between the two candidates.

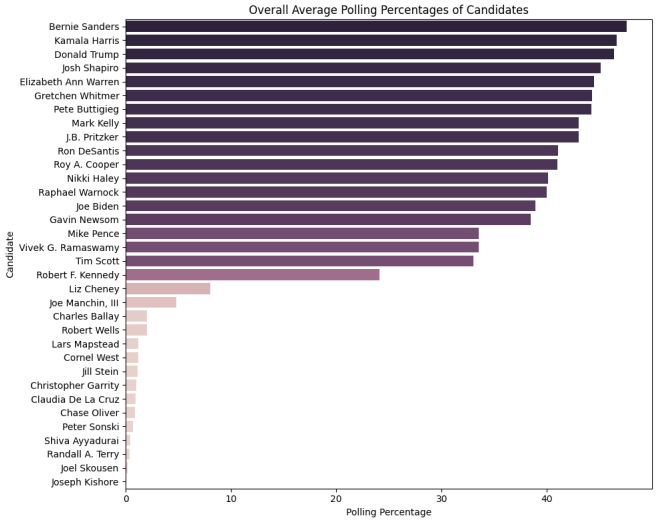

Fig. 5: Horizontal Bar Graph showing Average Polling Percentage of Candidates.

G. Sample Size vs Mean Polling Percentage

Key Insights:

- The trendline suggests a slight positive correlation, meaning that increased sample sizes are somewhat likely to result in slightly higher mean polling percentages.

H. Polling Percentage with Error Bars by Sample Size

Key Insights:

- Generally, smaller samples show more variability, as given by the taller length of error bars.
- Larger sample sizes tend to have more consistent (lower variance) polling percentages.

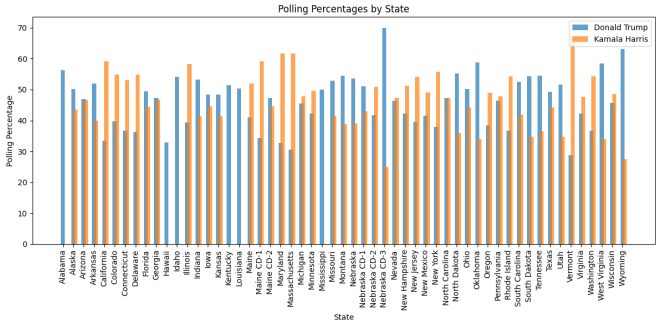

Fig. 6: State-wise distribution of polling percentages for key candidates, highlighting regional variations in support across different states.

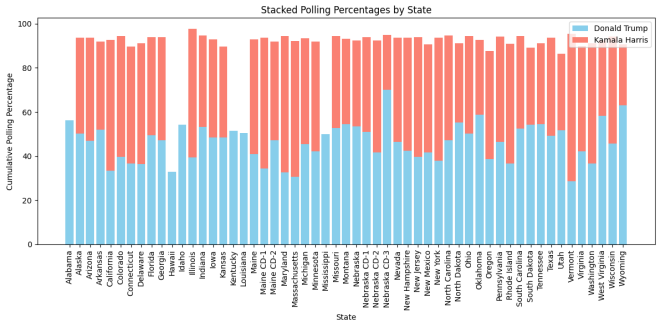

Fig. 7: Cumulative polling percentages across different states, illustrating the combined support for key candidates.

I. Distribution of Standard Deviation

Key Insights:

- The distribution is bimodal, with peaks in the lower and higher ranges of standard deviation.
- Most of the standard deviations lie between 15 and 25, indicating moderate variability in the polling data.

J. Sample Size vs Standard Deviation

Key Insights:

- There seems to be a clear trend—namely, the smaller the sample size, the higher the standard deviation, which indicates less reliable data for small datasets.
- Larger sample sizes have smaller standard deviations, indicating the polling data points to greater precision and stability.

K. Polling Percentages Over Time (Interactive Line Chart)

Key Insights:

- Kamala Harris: The initial upward curve is smooth, followed by a period of volatility as the dates of polling get closer.
- Josh Shapiro and Bernie Sanders have very consistent trends, meaning that their voter bases were stable across this period.

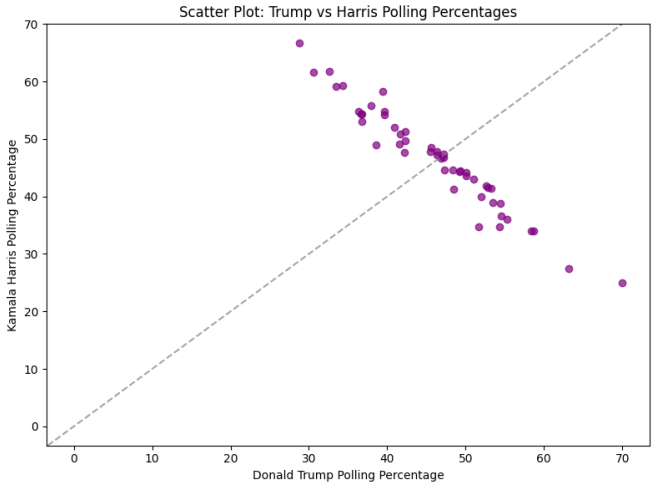

Fig. 8: Scatter plot comparing the polling percentages of Donald Trump and Kamala Harris across various states, with the diagonal line indicating equal support for both candidates.

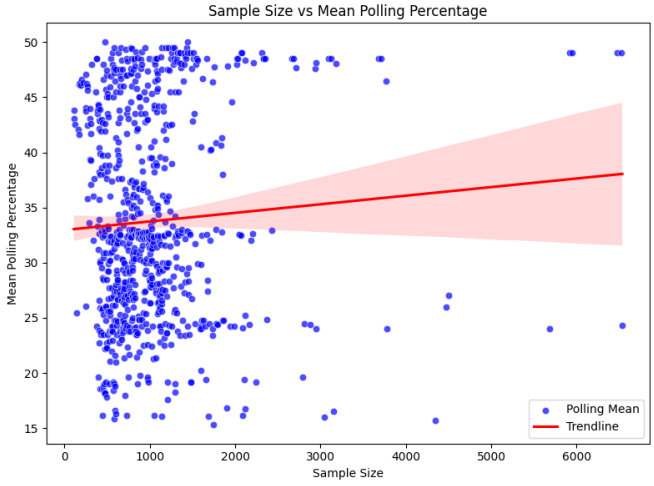

Fig. 9: Scatter plot showing the relationship between sample size and mean polling percentage, with a trend line indicating the general correlation.

L. Cumulative Polling Trends Over Time

Key Insights:

- Kamala Harris exhibits a commanding presence in the aggregate polling trends, signifying a wider resonance over various temporal spans.
- The impact of Josh Shapiro and Bernie Sanders is comparatively limited, suggesting a more pronounced leading position for Kamala Harris.

M. Polling Percentages Heatmap

Key Insights:

- Kamala Harris has high polling numbers at almost all intervals, thus leading the pack.
- Bernie Sanders shows periods of intense polling, reflecting moments of great public support.

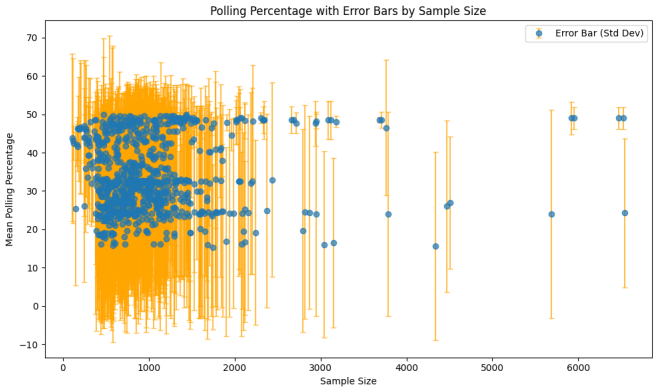

Fig. 10: Scatter plot with error bars showing average polling percentages and standard deviations for varying sample sizes.

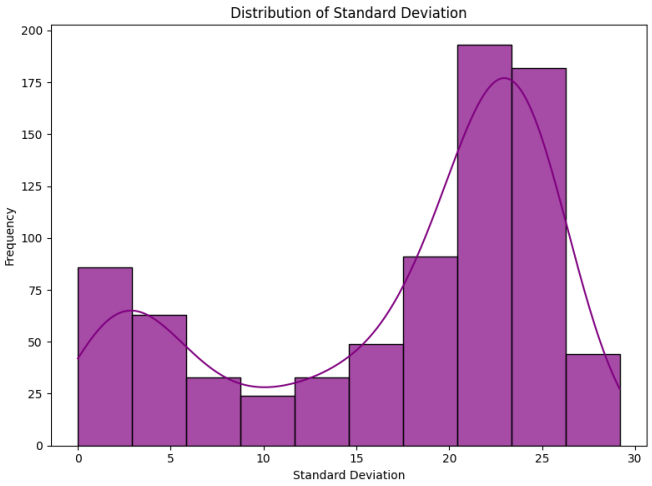

Fig. 11: Histogram displaying the distribution of standard deviations in polling percentages with a density curve.

- Josh Shapiro has only a few periods of high-intensity polling, suggesting a rather muted voter tendency.

N. Distribution of Transparency Scores Across Pollsters

Key Insights:

- Most pollsters have a transparency score of 5, 6, or 7. There is also a drop in the frequency for scores of 0 and 10, which suggests that there are fewer pollsters at either end of the transparency spectrum.

O. Transparency Scores vs. Poll Results (Scatter Plot)

Key Insights:

- There is no visible indication of a straight-line relationship between transparency scores and poll percentages, as the results of the poll are spread for all transparency scores.

P. Poll Results Distribution Across Transparency Score Ranges (Box Plot)

Key Insights:

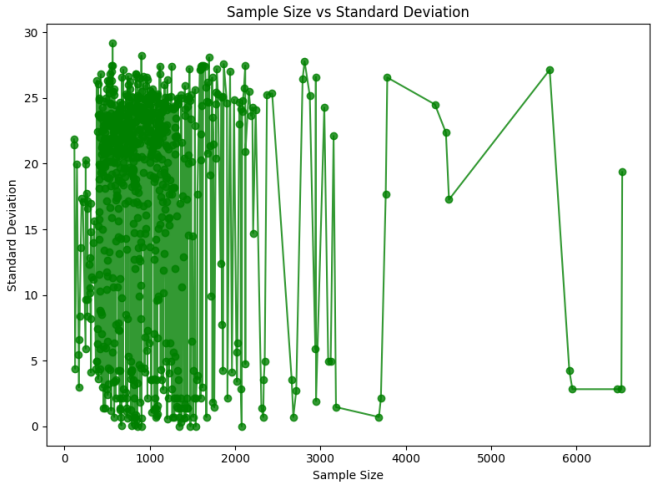

Fig. 12: Scatter plot showing the relationship between sample size and standard deviation in polling percentages, highlighting trend patterns.

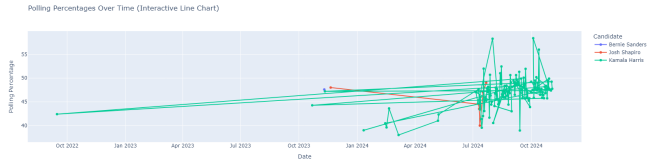

Fig. 13: Interactive line chart tracking the polling percentages of leading candidates over time.

- The results of the poll for lower transparency scores (0—2) are highly variable with a larger inter-quartile range.
- With increasing transparency scores, the poll results stabilize and the median is fairly consistent.
- Transparency scores between 2 and 4 have fewer outliers compared to other ranges.

IX. Pipeline Automation using Apache Airflow

The entire data pipeline was implemented using Apache Airflow. The ETL pipeline was structured as a Directed Acyclic Graph (DAG) with modular tasks for:

- Data ingestion
- Data cleaning and transformation
- Analytics mart generation
- Feature engineering

A downstream prediction DAG consumed the prepared features and generated prediction reports and visualization artifacts. This architecture ensured reproducibility, modularity, and clear management of task dependency.

X. Conclusion

The project provided a comprehensive analysis of the presidential poll data for the 2024 cycle, incorporating data preprocessing, data visualizations, data analysis, and statistical forecasting while automating certain areas of the project

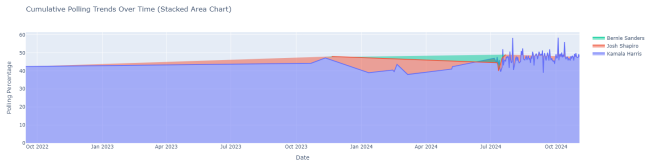

Fig. 14: The graph visually represents the cumulative polling percentages of the top three candidates over time, highlighting each candidate's proportional contribution.

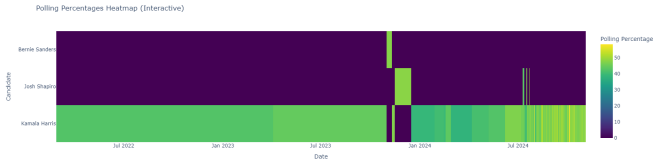

Fig. 15: This heatmap shows the polling percentages of Bernie Sanders, Josh Shapiro, and Kamala Harris over time, with darker colors representing higher percentages.

pipeline using Apache Airflow. We began by exploring the data set, identifying patterns in attributes such as polling percentages, transparency scores, sample sizes between candidates and states, and visualizing them for interpreting trends. For the predictive part, a statistical approach was adopted to estimate the likely winner of the 2024 election based on average polling percentages which was supplemented with visualizations to convey the winning probabilities for various candidates effectively.

The project workflow was automated using Apache Airflow to ensure efficient data processing, analysis, and reproduction of the results. In addition, in the prediction task, while the statistical approach provided clear and interpretable results, it helped identify the limitations and uncertainty of relying solely on polling data to predict election results. Other external factors such as voter turnout or campaign dynamics could also play a major role. However, this project has successfully demonstrated how data-driven techniques can enhance our understanding of electoral dynamics, offering valuable insights into various trends for political analysts, campaign managers, or decision makers.

References

[1] C. Panagopoulos and W. K. Tam Cho, "Polls and elections: Forecasting presidential elections in the united states," *Political Science Quarterly*, vol. 129, no. 4, pp. 703–725, 2014.

[2] R. S. Erikson and C. Wlezien, "The performance of pre-election polls in the 2020 u.s. presidential election," *Public Opinion Quarterly*, vol. 87, no. 1, pp. 15–34, 2023.

Fig. 16: The histogram displays the transparency ratings, ranging from 0 to 10, by different pollsters.

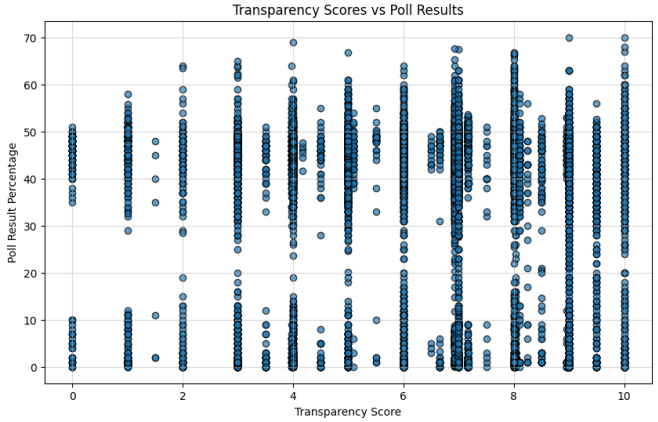

Fig. 17: Scatterplot of transparency scores versus poll result percentages.

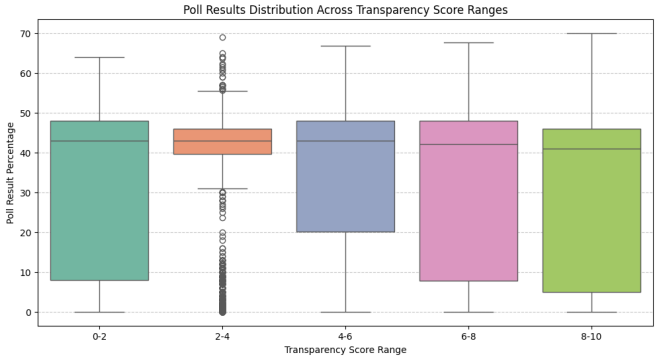

Fig. 18: A box plot for the distribution of the outcomes of polling along different ranges of transparency scores (0–2, 2–4, 4–6, 6–8, 8–10).

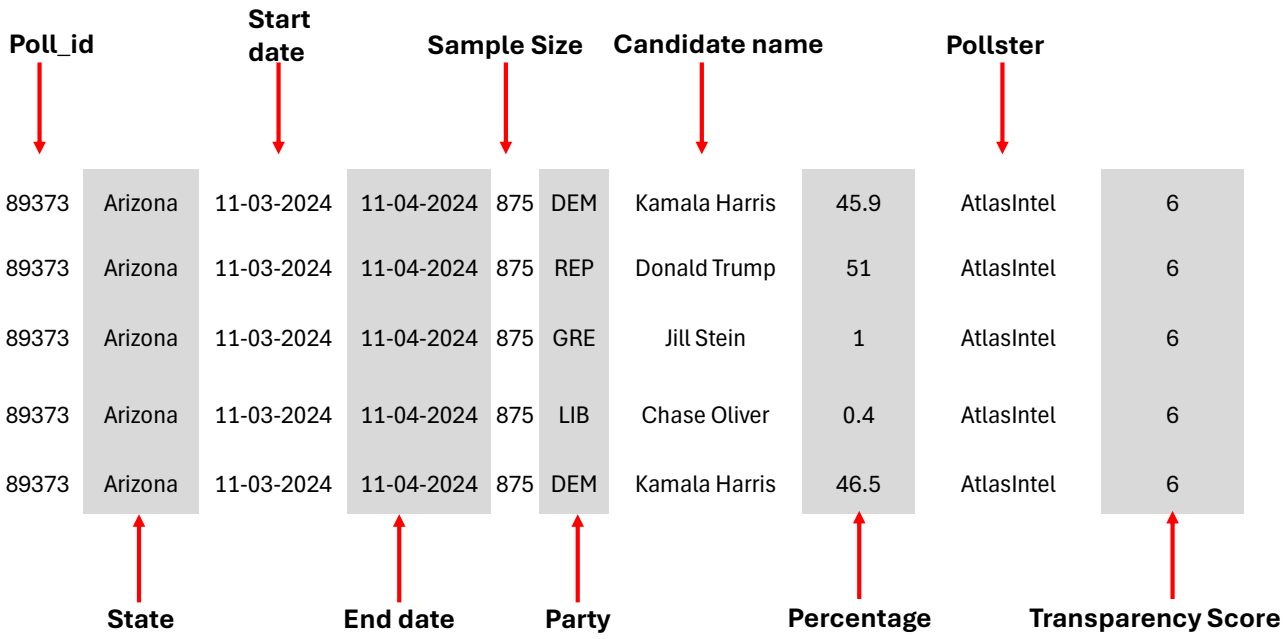

Fig. 19: Sample Dataset Image.