Final Project Part I
Math-014-01 Intro to Data Science
Camille Settles
Dr. Meenakshi Nerolu
April 24, 2025
Project Title: U.S. Senate Elections Analysis (1976-2020)

Objective:

The objective of this project is to analyze over four decades of U.S. Senate election data to uncover trends in voter turnout, party dominance, and candidate performance. Using Python-based tools, the goal was to identify patterns across different states, track changes in total votes over time, and understand how political engagement and party strength have evolved. This project aims to equip students with hands-on data science skills by applying cleaning techniques, exploratory data analysis (EDA), and visualization to derive meaningful insights.

Introduction:

This dataset contains historical U.S. Senate election results from 1976 to 2020. It consists of more than 18,000 rows, each representing a candidate's performance in a specific election year and state. The data includes key attributes such as year, state, party affiliation, total votes, candidate names, write-in status, and coded state identifiers (state_fips, state_cen, state_ic). Through this project, I sought to gain insight into voter behavior and the dynamics of Senate races at both the state and national levels. The scale and timespan of this dataset offer a unique opportunity to observe macro-level patterns and micro-level anomalies.

The significance of this project lies in understanding how population growth, political polarization, and regional demographics influence Senate outcomes. Data analysis can not only reveal cyclical voting behaviors but also highlight emerging shifts in party influence, voter turnout, and candidate popularity. In addition, this project demonstrates the power of data science to turn raw historical data into a compelling narrative that informs both political science and civic understanding.

Method: Data Wrangling and Cleaning

The original dataset contained inconsistencies and missing values that needed resolution before meaningful analysis could begin. The first step was displaying the structure of the data: column types, unique value counts, and descriptive statistics. The dataset had over 18,000 rows and 28 columns, with numeric fields like totalvotes and candidatevotes as floats or integers, and categorical fields such as state, party_detailed, and candidate as strings.

I began by identifying columns with missing data. Fields like party_detailed and candidate contained null values. These were imputed with placeholder values such as "Unknown Party" or "Unknown Candidate" to preserve the integrity of the data. Missing values in non-critical

columns, like write-in, were filled with Boolean defaults (False). Since there were no missing values in numeric fields, there was no need to apply median or mean imputation. I avoided dropping rows or columns to prevent loss of information unless the row was entirely invalid.

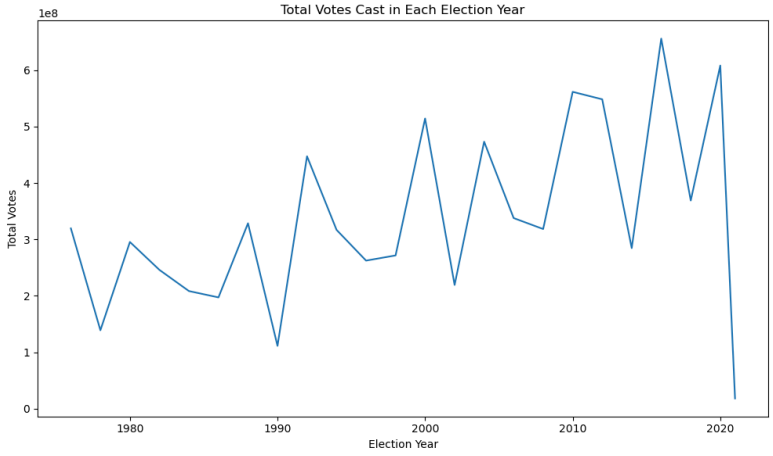
Using the Interquartile Range (IQR) method, I identified 196 outliers in totalvotes and 398 in candidatevotes. Initially, I removed all outliers, but this eliminated high-population states like California, Texas, and New York contributed significantly to national vote counts. To address this, I filtered the data to retain rows that either fell within the IQR or belonged to a list of high-population states. This approach preserved the accuracy and representativeness of the dataset, ensuring these influential states were not excluded from the analysis.

Before filtering, there were 196 total vote outliers; after filtering with population preservation, 196 remained. For candidate votes, 398 outliers were reduced to 226. This strategy maintained data integrity while allowing outlier logic to inform analysis.

To address text inconsistencies, all string values in columns like state, candidate, and party_detailed were converted to title case and had leading/trailing white spaces removed. These standardizations avoided discrepancies like "democrat" vs "Democrat" and ensured accurate groupings. Additionally, a new column, vote_percentage, was created to represent the percentage of votes received by a candidate relative to the total votes in their election. This derived feature allowed for clearer comparison across states and years, beyond raw vote totals.

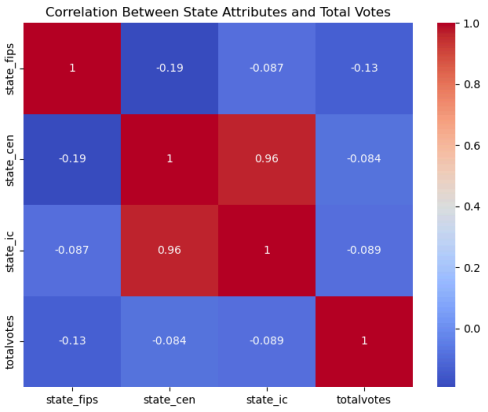
Storytelling: Data Visualization & Interpretation

Total Votes Over Time:



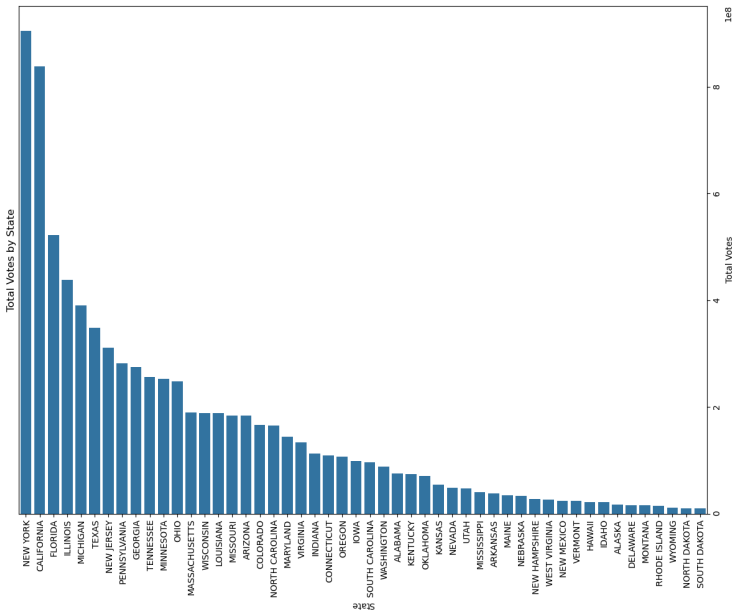
A line plot of total votes cast from 1976 to 2020 shows an upward trend in voter participation. Peaks occurred in 2008, 2016, and 2020, which were nationally significant election years. The rise in voter turnout reflects increasing political engagement and population growth. Notably, around 2020 recorded the highest number of votes cast.

Correlation and Pair Plot:



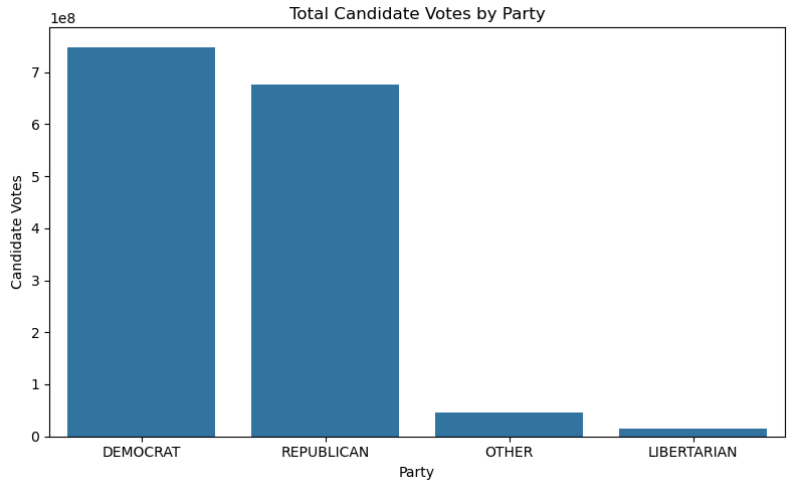
Correlation matrices and pair plots between totalvotes, state_fips, state_cen, and state_ic show negligible correlation. This demonstrates that numerical state identifiers do not predict voting behavior. These codes serve primarily as metadata.

State-Wide Total Votes:



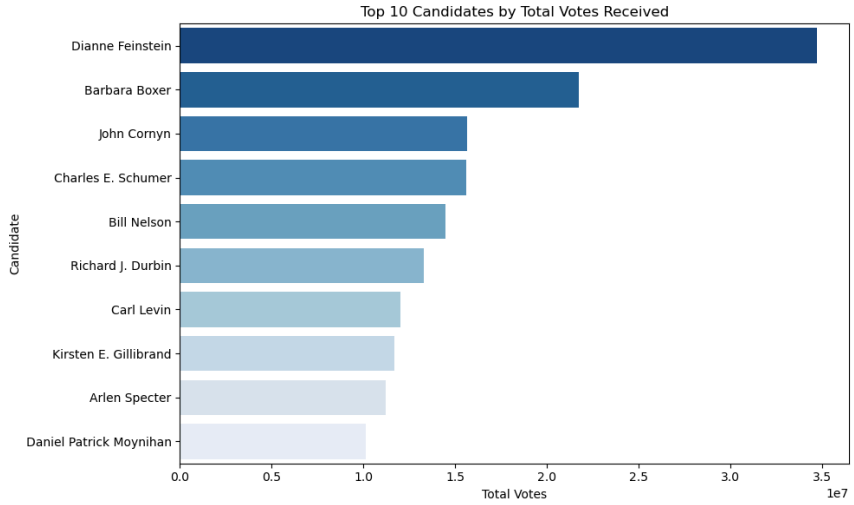
Horizontal bar plots revealed that states like California, Texas, New York, and Florida consistently report the highest vote counts due to their higher population. Their presence validates the decision to retain them during outlier filtering. Small states like Wyoming and Vermont appear at the bottom, which aligns with expectations based on population size.

Party Performance Comparison:



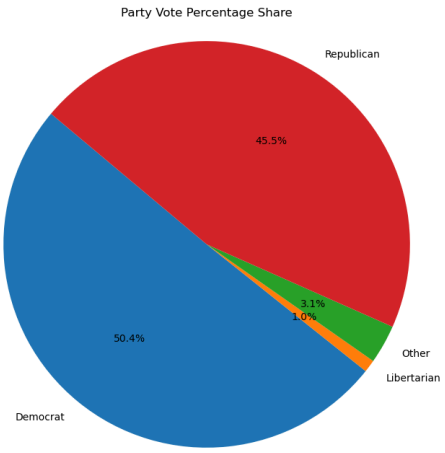
This visualization compares the total number of votes received by each political party across all Senate elections from 1976 to 2020. The graph clearly shows a two-party dominance, with the Democratic and Republican parties receiving the overwhelming majority of votes. While Democrats slightly surpass Republicans in total votes, the margin is relatively narrow, underscoring the competitive nature of Senate races. Third parties, such as the Libertarian and Green parties, garnered significantly fewer votes, rarely exceeding 2% in any given election cycle. This trend highlights the structural barriers and limited influence faced by smaller parties in the American political system. The party performance chart provides a broader understanding of how consistently major parties maintain electoral control, reflecting both institutional entrenchment and voter alignment.

Top Candidates by Vote Totals:



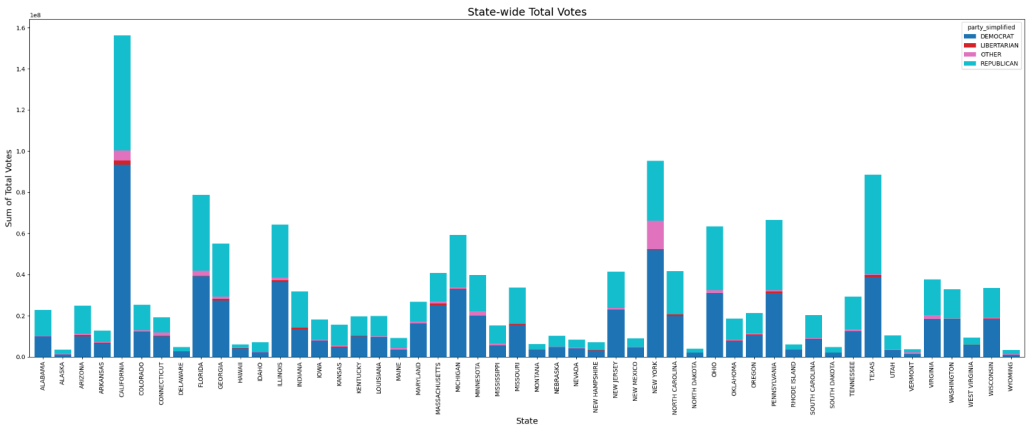
The top vote-getters are highlighted as Dianne Feinstein, Barbara Boxer, and John Cornyn as the top three candidates. Feinstein received over 34 million votes, which reflects not only incumbency but also California's large electorate. These numbers illustrate how vote totals can be driven by both political popularity and state demographics.

Party Vote Share and Dominance:



A pie chart shows Democrats receiving 50% of the total votes, Republicans 46%, and other parties under 2%. This supports the two-party dominance in U.S. politics, especially in Senate elections.

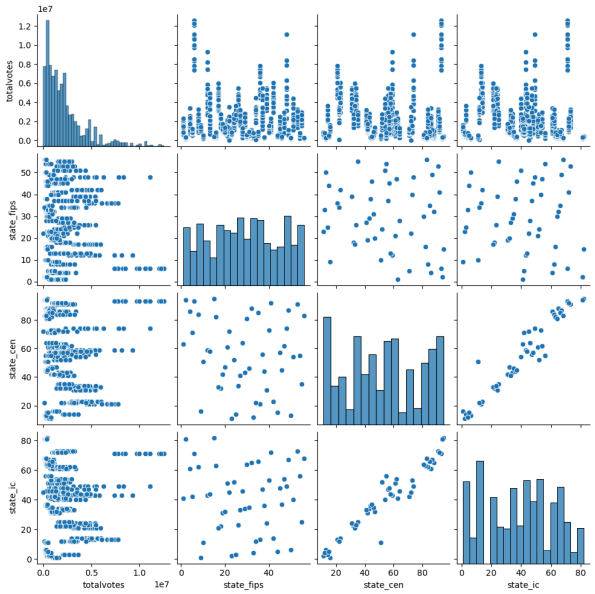
Plotly Bar Chart:



The visualization highlights how vote counts are distributed not just in total, but across Democratic, Republican, Libertarian, and Other party categories within each state. States like California, Texas, New York, and Florida once again emerge as dominant contributors to national vote totals, with both Democratic and Republican parties sharing substantial segments of the vote. The inclusion of minor parties like the Libertarian and "Other" categories also provides insight into third-party traction, which, while small, is more noticeable in select states. This chart

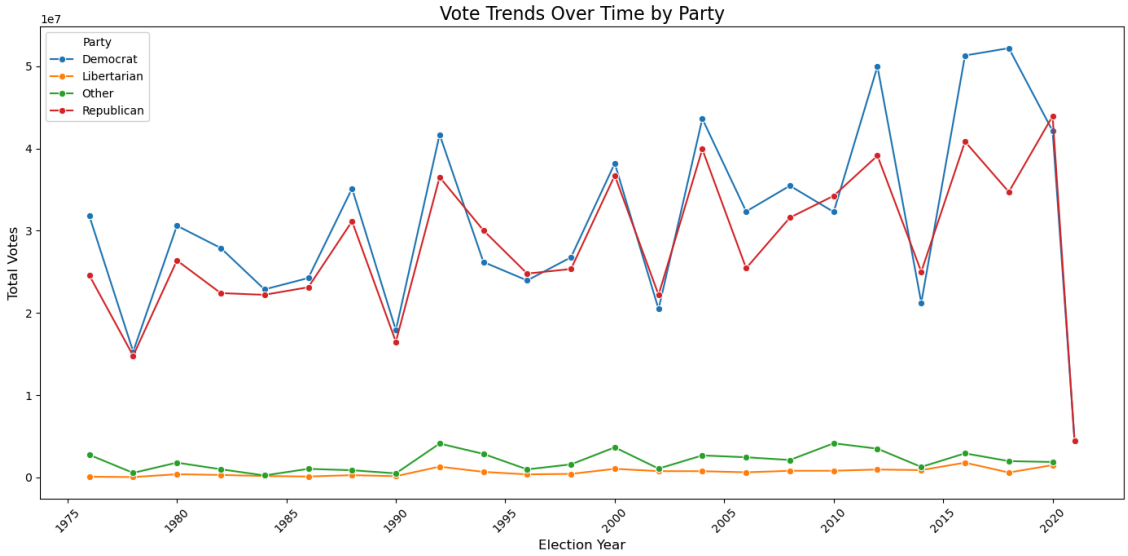
reinforces the critical role of high-population states in shaping federal election outcomes and reveals the persistent two-party concentration in Senate contests.

Pair Plot:



The pair plot offers a visual cross-check of the relationships between numerical attributes in the dataset. It further validates that state-level codes (state_fips, state_cen, state_ic) do not demonstrate strong patterns when compared to total vote counts. The lack of clustering or directionality reinforces earlier findings from the correlation matrix.

Party Vote Trends Over Time:



This line graph shows how vote totals for the Democratic and Republican parties have evolved across time. While both parties have maintained strong participation, Democrats slightly edge out Republicans overall. This temporal view captures fluctuations in party strength and provides a longitudinal context for interpreting vote share dominance.

Conclusion:

This analysis of U.S. Senate elections from 1976 to 2020 reveals several key insights. Voter turnout has steadily increased, with especially large spikes in recent decades reflecting both political engagement and population growth. The vote count distribution shows that states like California, Texas, New York, and Florida play an outsized role in Senate outcomes due to their population size. The analysis confirmed the dominance of the two major parties, with Democrats slightly ahead in overall vote share. Although write-in voting is rare, it provides subtle indications of voter dissatisfaction in specific years. In terms of data features, the newly engineered vote percentage column allowed for clearer interpretation of candidate performance. Lastly, numeric state codes such as state_fips, state_cen, and state_ic serve primarily as identifiers and do not meaningfully explain variance in voting outcomes. These attributes help distinguish states, but they do not significantly predict voting volume, which is instead more driven by population and state size.

References:

Kaggle. *U.S. Senate Elections (1976–2020)* Dataset. Retrieved from https://www.kaggle.com/datasets/unanimad/us-election-2020

Acknowledgments:

I would like to thank Dr. Meenakshi Nerolu for her consistent support, guidance, and thoughtful feedback throughout the semester. Her lectures and project guidance provided the framework that allowed me to apply technical concepts to real-world political data.