STAT 380 FINAL PROJECT REPORT (GROUP ONE) – Analysis on How Every Rep, Senator, and President Voted on Every Bill in History

Sid Tekumalla, Sean Dasovich, Lucas Sadoulet, Li Zhu

###INTRODUCTION:

For this project, we took a while to select a data set that would pique our interest and be insightful. We have a joint interest in politics, social behavior, and public policy. We landed with a data set from voteview.com which describes how each member of Congress has voted on bills throughout history. We are fascinated by the way congress members vote whether that's crossing party-lines or shifting positions over time. We are also interested in if there are ways one could group together members of Congress to see if there's patterns in ways members vote. Throughout discussions and initial exploratory data analysis we landed on three research questions:

Intra-Party Factions: Using the NOMINATE scores, can we identify distinct Republican factions (Freedom Caucus, moderates, Trump-aligned) and Democratic factions (progressives, moderates, Blue Dogs)? How accurate is this identification?

Age and Ideology: Using the birth year data, is there a correlation between member age and ideological positioning in 2025, controlling for party? What factors most predict a politician's age? Are younger Republicans/Democrats different from older ones, and in what ways?

Party Loyalty: Are politics getting more divided, i.e. is there less cooperation on bill / less breaking rank with each progressing congress?

Politics is tricky when it comes to data analysis because emotions and humans are involved. When these things are involved, no data will be perfect, and we are sure that there will be outliers present. Although it's not perfect data, we still would like to see if there are patterns that emerge in the way these congress members vote.

In our modern world, politics is more divided than ever. If you look at the news, there is constantly fighting between party lines. We want to see how true this is. Is America truly as divided as we see, or is that just the news? This project is giving us the opportunity to explore perception vs reality.

###DATA DESCRIPTION:

The Data set we used came from Vote View which is a public resource that has information on congress voting. We used the member ideology dataset for the 119th Congress. The Data set had one row per member of Congress whether that is House or Senate. The key variables can be broken into several categories: member identification, basic demographics, congressional information, and ideology scores.

Member Identification

bioname: Full legislator name

icpsr: Unique numeric identifier

bioguide_id: Biographical directory ID

Basic Demographics

born: Birth year

died: Year of death (if applicable)

state_abbrev: Two-letter state code

district_code: House district number (0 for Senators)

Congressional Information

congress: Session number (119th)

chamber: House or Senate

party_code:

100 = Democrat

200 = Republican

Ideology Scores

Nominate_dim1: First dimension of DW-NOMINATE ideology score

Nominate_dim2: Second dimension of DW-NOMINATE ideology score

Nominate_log_likelihood: Long-likelihood of model fit for the member's ideological estimation

Nominate_geo_mean_probability: Geometric mean of the probabilities that the model predicts each vote correctly

Nominate_number_of_votes: Total number of roll-call votes used to estimate the position

Nominate_number_of_errors: Number of incorrect predicted votes by the model

Conditional: Used internally for DW_NOMINATE conditional estimation

Nokken_poole_dim1: Alternative ideological dimension score using Nokken-poole method

Nokken_poole_dim2: Second Nokken-poole dimension score

Doing some EDA, we came to the conclusion, controlling party affiliation. There is no visible correlation between congress member age and ideological positioning in the current Congress. This is further supported if we were to try linear regression, as shown in the code appendix. We choose a range of important predictors like chamber, state, and ideological scores, and run a regression on them.

To clean the data removed rows with missing NOMINATE scores, only keeping the variables we need for the analysis. We also removed missing values from the data set. To complete the k-cluster analysis we split the dataset into Democrats and Republican.

###METHODOLOGY:

Intra-Party Factions Methodology:

We cluster members within each party using k-means on (Nominate_dim1, Nominate_dim2) and validate with elbow and silhouette diagnostics.

Nominate_dim1: liberal (−) ff conservative (+)

Nominate_dim2: often captures establishment (+) vs. populist (−) tendencies (modern era)

We choose k using elbow (WSS) and average silhouette over k_grid. Then we fit k-means and visualize a 2D cluster map.

Age and Ideology Methodology:

We built a multiple linear regression model. We had the response variable as age and the predictors were chambered (House/Senate), state, nominate_dim1, nominate_dim2, nokken_poole_dim1 and nokken_poole_dim2. This linear regression helps us to see if age is associated with ideology.

Party Loyalty Methodology:

ADD LATER

###DATA ANLYSIS RESULTS:

Intra-Party Factions Results:

Although the Democrats were split into three clusters in the k-means map (k = 3), this choice was made mainly to stay consistent with the research question. In fact, the silhouette plot shows that k = 2 has the highest average silhouette score, meaning two clusters provide the clearest separation according to the underlying NOMINATE distances. When we examine the numerical cluster centers, the three Democratic clusters are extremely close to each other. All three centers fall within a very narrow range. As a result, the labeling rules classify all three Democratic clusters as "Mainline Moderate," indicating that the algorithm's three-way split does not correspond to three meaningfully distinct ideological factions within the caucus.

With k = 3, the Republican NOMINATE map visually splits into three clusters, but the numerical results show that two of the clusters sit extremely close to each other in the ideological space and both represent the same mainstream "Traditional Conservative" group. The third cluster—located lower on the second dimension—is the only one that clearly separates, corresponding to a "Trump-aligned / Populist" faction. So even though the algorithm technically identifies three groups, the substantive interpretation is that Republicans largely divide into one major establishment bloc and a smaller Trump-aligned faction, with the two establishment clusters reflecting minor within-group variation rather than distinct ideological camps.

There is no clear point in either WSS plot where the total within-cluster sum of squares levels off; instead, both show a gradually flattening pattern as k increases. Silhouette scores above 0.7 indicate strong clustering and scores above 0.5 indicate reasonable clustering, yet for both Republicans and Democrats, the highest silhouette value occurs at k = 2 and remains below 0.5. This suggests that neither party exhibits a meaningful clustering structure. Although k = 3 was chosen to match the requirements of the research question, the more appropriate choice based on the metrics would be k = 2. Furthermore, the resulting NOMINATE dimension analysis shows that the ideological separation is not very distinct: while we expected three ideological levels for both parties, the actual clustering reveals fewer than three clearly differentiated ideological group.

Age and Ideology Results:

As the EDA shows, controlling party affiliation, there is no visible correlation between congress member age and ideological positioning in the current Congress. This is further supported if we were to try linear regression, as shown below. We choose a range of important predictors like chamber, state, and ideological scores, and run a regression on them. The multiple R-squared value is a measly 0.1811, and the adjusted R-squared is an even lower 0.05292. This means that a very small amount of the variance in the data is covered by our predictor variables, making them unfit to explain age.

It is well known that American Congressmembers skew on the old side. The median age is 59, and plenty of congressmembers are far older than this. Our analysis shows that this age is not related to location or ideology, party or chamber, suggesting that this high age is built into the institution itself. Further studies would be needed to examine the institution, probably through a study of national/state laws and party election rules. These, however, are beyond the scope of our project.

Party Loyalty Results:

ADD LATER

###CONCLUSION:

ADD LATER

###AUTHOR CONTRABUTION STATEMENT:

Sid: age-ideology analysis, built regression model, wrote methodology and results for his research question

LI: cleaned and prepared the data set, preformed the intra-party faction analysis, creating silhouette and elbow diagnostics. Wrote methodology and results for his question.

Sean: wrote introduction, data description, conclusion. Integrated everyone's work into the final report and completed this document.

Lucas: party loyalty analysis, wrote methodology and results for his research question.

###CODE APPENDEX: