CS109A Project Milestone 3: EDA & Revised Project Statement

Justin Bassey, Jake Boll, Christopher Lewis, Sebastian Schwartz

PROJECT GOALS

We are looking to explain what features and information from our district level demographics data explain the 2018 midterm election results. Moreover, we want to explore how our demographic data can help explain the differences our EDA revealed between outcomes of the Democratic and Republican parties.

EXPLORATORY DATA ANALYSIS

We utilize three datasets as the basis of our analysis: 1976-2018 House Election Data from Harvard's Dataverse, Midterm Election Seat Change Data scrapped from 'www.presidency.ucsb.edu', and District Level Demographics Data from Census.gov. *We also obtained FiveThirtyEight's predictions of the 2018 midterm elections, but using their polling and model in our model felt like cheating.*

In our initial exploration of the data, we found that how states traditionally vote (Republican vs democrat) by examining the average number of votes cast for either party and examining the difference. This EDA produced Figure 1 below.

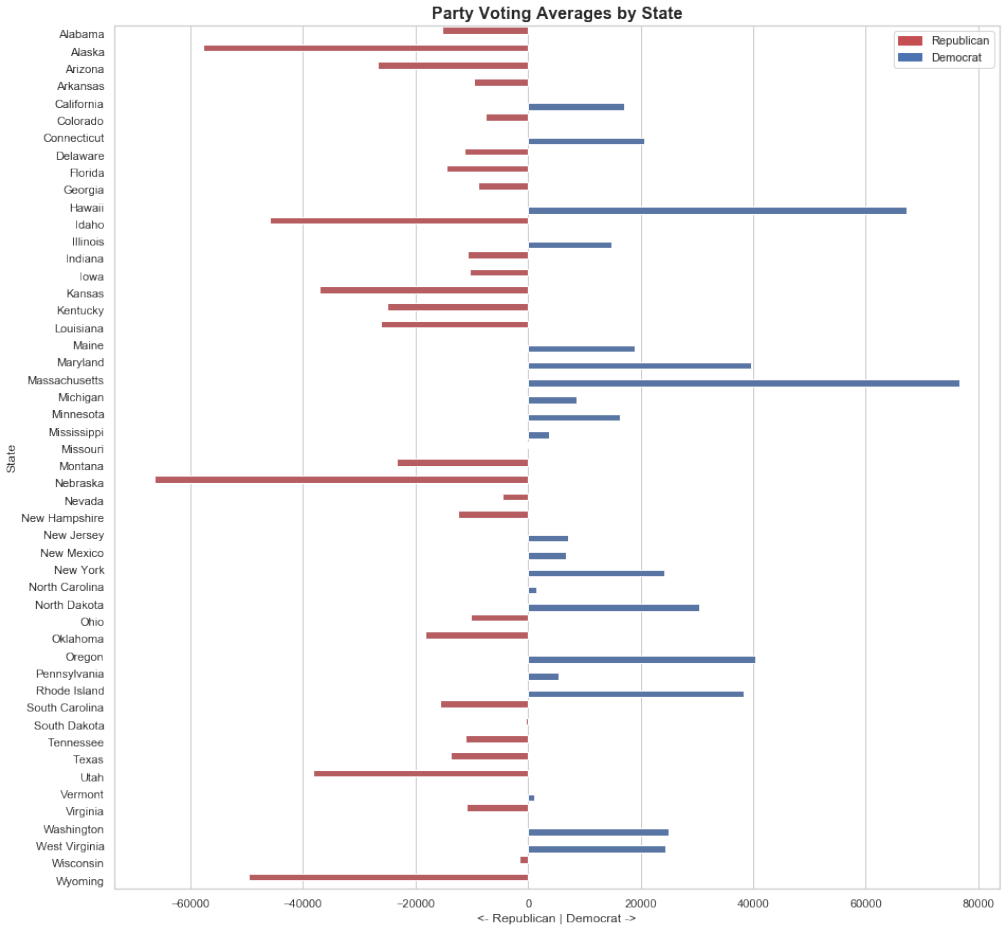

Figure 1. Traditional party lean of state voting behavior since 1976. Negative values equate to greater republican leanings. Positive values equate to greater democrat leanings.

We were also curious as to whether there are differences to how many seats were lost by the current president's party. To examine this, we pulled on our scraped seat change data, and found that on average democrats lost 31.75 seats, while republicans only lost 22.7. The distributions of these data are shown below in Figure 2.

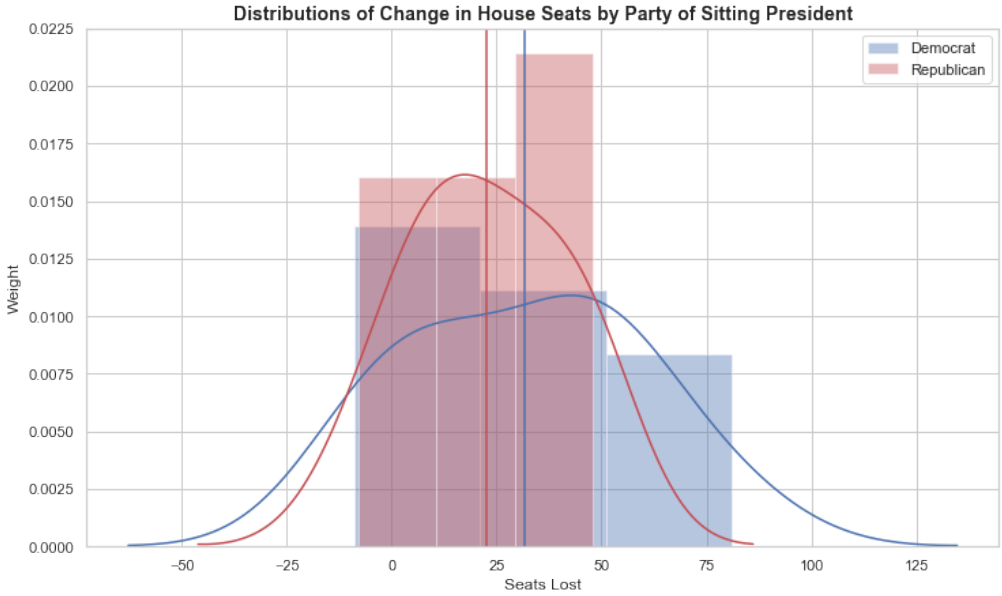

Figure 2. Differences in seats lost by the party of the sitting president.

DATA CLEANING & RECONCILIATION

In gathering our data across multiple sources, a problem we ran into was merging the datasets such that we didn't lose too much data. In particular, with our demographics data, we only have district demographics for 2018, but wanted to include it in our model. In order to prevent cutting a lot of historical data or entering a lot of Nan values, we had to perform a 1 to many merge where the 2018 demographics data is entered into rows from different years. This is currently a temporary solution.

Additionally, the data we obtained was not perfect, thus in cleaning our data we had to make decisions such as dropping a few candidates whose names and parties were unknown, and other areas where we were either missing too much data to keep the row/column, or had to impute a value for it.

We additionally created dummy variables for the categorical variable party, and created a binary variable for the presidential incumbent party.

BASELINE MODEL

To create a baseline model we ran a linear regression on the full data including all predictors after normalizing all non-binary variables using sklearn's MinMaxScaler to between 0 and 1. The linear regression was trying to predict the binary variable 'winner'.

We saw that the initial regression did not perform very well at all, achieving a poor result of $r^2 = 0.24$ for the training data, and $r^2 = 0.2$ for the test data.

Examining the p-values, the only really significant predictors were that democrat and republican candidates were much, much more likely to win than other candidates, which makes sense.

To improve this initial regression, we are aiming to experiment with adding and removing different features. In particular, interaction features between political party features and age, education and socioeconomic status features may be

promising. We will also investigate how the addition of regularization, CV, and using logistic models may improve regression accuracy.

To investigate further, we will also experiment with other models including decision tree / random forests, and also a neural net.

We will also need to look at different methods of data normalization and standardization to ensure optimum model performance.