Capstone Project 1: Data Wrangling

What kind of cleaning steps did you perform?

The first step in cleaning the data was removing unnecessary columns from the dataset. This will help to prevent overcomplication and potential inaccuracies and will also improve the overall processing speed of the model.

I removed the following columns from the dataset:

- **abroadPrcnt** – percent of the district that lived abroad in the census used for this term. The field was removed because Censuses are updated yearly after the 109th Congress (2005-06). Prior to the 109th Congress, censuses were generally only updated once per decade, meaning that there are a significant number of missing values.
- **prcntExAliens** – percent of the foreign-born population that became a citizen. This is intended to capture a measure of assimilation. Unfortunately, this field also has a significant number of missing values.
- **totalHouseholds** – total number of households in the district at the time of the most recent census. This feature does not add anything to the dataset that totalPopBirthPlace doesn't already add.
- **under10k** – percent of the district's households that made less than $10,000/year in 2009 dollars. The exact income changes before and after 2009 due to inflation and the census using different income buckets for reporting income prior to 1990. It is always as close as possible to the 2009 value listed. We have 10 variables that are similar in that they track some threshold of income. Our intention is to choose the one that displays the greatest difference between Democratic and Republican districts. This variable did not show the greatest variation and so it is redundant and therefore excluded from the dataset.
- **over15k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over25k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over35k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over50k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.

- **over75k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over100k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over150k** – percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **over200k** – s percent of the district's households that made more than $15,000/year in 2009 dollars. See under10k for exclusion reasoning.
- **meanIncome** – mean income of a district's households. This would be a great feature to include, however, there are a significant number of missing values.
- **medianIncome** – median income of a district's households. This would be another income feature that would great to include. Unfortunately, there a significant number of missing values. Additionally, no inflation adjustments were used on the data. The values were estimated for the 1980 census using the percent of the population in each income bracket. Thus, for the 1980s there is a heavy rounding error involved in the estimation.
- **totalEmploymentPop** - total population used for the employment variables. Only includes those over the age of around 16, depending on the cutoff for that census. The more relevant feature here is **prcntUnemp**, which tracks the percent of the district's population that is unemployed but still in the labor force.
- **prcentNotEmploy** – percent of the district's population that is not in the labor force. This is a less meaningful variables as it excludes those who are unemployed but in the labor force – including young children, stay-at-home parents, the permanently disabled, the retired, etc.
- **totalPopRaceFile** – total population for the race variables. The numbers used here are irrelevant to determining the likelihood of a district voting Democrat or Republican. However, the race variables themselves are.
- **prcntBlackNotHisp** – percent of the district that is black but not Hispanic. Unfortunately, this variable has a significant number of missing values.
- **prcntMulti** – Percent of the district that is multiracial. Unfortunately, this variable has a significant number of missing values.
- **prcntWhite** – Percent of the district that is white but not hispanic. Unfortunately, this variable has a significant number of missing values.
- **prcntNotHisp** – Percent of the district that is not hispanic. Unfortunately, this variable has a significant number of missing values.
- **prcntOld** – Percent of the district that is over 62 or 60, depending on the age buckets used by the census for that decade. Unfortunately, this variable has a significant number of missing values.

- **medianAge** – median age in the district. Unfortunately, this variable has a significant number of missing values.
- **sponID** – govtrack id number for each member of Congress. All entries have a govtrack ID number.
- **Icpsr** – icpsr number for each member of Congress. This variable is irrelevant to our analysis, plus not all entries have an ICPSR number. Additionally, there are some discrepancies in how sources use ICPSR numbers, as some sources use 9's as the first digit to indicate party switches or other irregularities
- **state** – state name, two letter abbreviation (ie AK, CA). This is a repeat of information from **stateDist**.
- **district** – congressional district number. This is a repeat of **stateDist**.
- **lastName** – Some nontraditional characters (hyphens, accents) are garbled in translation. All names are recognizable by the human eye, but id numbers should be used for matching individuals.
- **firstName** – see **lastName**
- **middleName** – see **lastName**
- **age** – age at time of being sworn into Congress for that session. We are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **gender** – gender of MC. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **numberTerms** – number of terms served in the House. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **daysServed** – number of days of the full session served. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **fracServed** – fraction of the congressional session that the MC served in. The vast majority of MCs served in the full session (2 years). Those that didn't filled vacant seats or left mid-term. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **cd** – congressional district number. This is a repeat of **stateDist**.
- **statenm** – state name. Again, a repeat of information from **stateDist**, but this is the state name used by nominate tables, all in caps, and often abbreviated (ie MINNESO)
- **dwnom1** – first dimension nominate score. This variable's significance is unknown, so it is excluded.

- **dwnom2** – second dimension nominate score. This variable's significance is unknown, so it is excluded.
- **comPower** – how powerful were the committees the MC served in. Example provided is of the 109th Congress. Committees vary by congress, so points assigned vary by Congressional session. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **chair** – Binary variable. Takes the value of 1 if the MC was the chair of a committee in that congressional session, 0 if not. No MC chaired more than one committee. Committee information is currently available for the 109th-113th Congresses. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **rank** – Binary variable. Takes the value of 1 if the MC was the ranking member of a committee in that congressional session, 0 if not. No MC was ranking member of more than one committee. Committee information is currently available for the 109th-113th Congresses. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **comName** – names of the committees served on. Committee information is currently available only for the 109th through 113th Congresses. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **numCom** – number of committees served on. Committee information is currently available for the 109th-113th Congresses. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **rankChair** – rankChair = 0.5rank + chair. This creates a single variable for power in a committee, where being a ranking member is half as valuable as being the chair of a committee. Committee information is currently available for the 109th-113th Congresses. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **black** – binary variable for being a member of the Congressional Black Caucus. 1 if a member of CBC, 0 if not. To my knowledge, all self-identifying black members of Congress are members of the CBC. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **Hispanic** – a binary variable for ebing a member of the Congressional Hispanic Caucus. 1 if a member of CHC, 0 if not. To my knowledge, all self-identifying Hispanic members of Congress are members of the CHC. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.

- **numSpon** – number of bills an MC sponsored in a term. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **numCosp** – number of bills an MC cosponsored in a term. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **numPassH** – number of bills an MC sponsored and were approved by a full House vote in a term. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **numEnact** – number of bills an MC sponsored and were signed into law by a full House vote in a term. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **passPrcnt** – percent of bills that an MC sponsored and were enacted into law. Again, we are interested in the fundamental characteristics of the districts themselves, not on who runs for Congress in them.
- **partyControl** – 'D' if Democrats controlled Congress, 'R' if Republicans controlled Congress. We are interested in district specific predictions, not party control of Congress as a whole.
- **demSeats** – number of seats held by Democrats at the beginning of the Congressional session. Again, are interested in district specific predictions, not party control of Congress as a whole.
- **repSeats** – number of seats held by Republicans at the beginning of the Congressional session. Again, we are interested in district specific predictions, not party control of Congress as a whole.
- **otherSeats** – number of seats held by third parties (independents, green party, etc.). Note that this dataset does not provide information about which third party MCs belonged to. Again, we are interested in district specific predictions, not party control of Congress as a whole.
- **ses** – socioeconomic status measure: estimated from the income and education of a district. This was estimated by finding the weight of the common factor between education and income. The method by which this variable was calculated is unknown and so is excluded from the data.
- **sesNorm** – socioeconomic status measure, normalized: estimated from the income and education of a district. It is normalized to range between 0 and 100. This allows the magnitude of this coefficient to be directly compared to the coefficients for income and education, both of which are given in percentages. The method by which this variable was calculated is unknown and so is excluded from the data.

How did you deal with missing values, if any?

The next step in cleaning the data was dealing with missing values. Once we removed the above fields there were very few rows that had any missing values. However, those rows that did have missing data were simply removed from the data. As a result, the number of rows declined from 9,312 to 9,298 – an insignificant difference.

Were there outliers, and how did you handle them?

Due to the nature of the data, there weren't any obvious outliers. However, the last step in the data cleaning process was to apply one-hot encoding to our dependent variable – **party**. The goal of our analysis is to predict, given the fundamental features of the district in question, whether that district will vote Democrat or Republican.

Therefore, our dependent variable should either be Democrat or Republican – a binary choice – which can be represented by either a 0 (for Republican) or 1 (for Democrat). We can do this easily by applying one-hot encoding, which converts text-based features into numerical, binary values, represented as "1" or "0" – "True" or "False". This will be helpful going forward as many algorithms and also scatterplots are not compatible with non-numerical data.

When I applied one-hot encoding to the parties variable, it ended up splitting into four additional features – **party_Democrat, party_Republican, party_Independent**, and **party_Republican-Conservative**. The variable **party_Indpendent** was such a small value that rows containing this value could be easily removed.

Rows containing **party_Republican** and **party_Republican-Conservative** could easily be combined, until all that was left was **party_Democrat** and **party_Republican**.

Finally, since **party_Democrat** contained information – in binary – as to the outcome of the election in each district, we could remove **party_Republican**, leaving one, final, dependent variable – **party_Democrat** – whereby a 0 signified a Republican win and a 1 signified a Democratic win.