Capstone Project 1 Data Wrangling Summary

James Olmstead

What kind of cleaning steps did you perform?

The first thing I did was categorize each survey question by its type (categorical, numerical, etc.) so I could perform cleaning operations on broad swaths of the data. I also have to change some of the questions to be categorical that began as numerical (eg. changing favorability scales from 1-100 to 1-10). The final result is that all the columns represent categorical data. For questions that tied to other questions in the survey, I combined them additivity to their corresponding question. For example, say a respondent said in Question A that they voted in the Republican primary in 2016 (a response of 2) and in Question B said they voted for Marco Rubio (a response of 4), the new aggregated response would be 2.4, which would represent a new category.

How did you deal with missing values, if any?

I first removed the survey respondents who didn't answer enough of the questions (200/1074), likely since they only participated in part of the longitudinal study. Then I had to tackle the missing values based on the type of data. For questions that had 'Don't Know' as a possible response, I clumped all the missing values into that category. For questions that didn't have such a response, I created a new category that represented a blank response. This applied to numerical data that converted to categorical. For questions that tied to other questions in the survey, I set their missing values as 0 since they would be combined additivity to their corresponding question.

Were there outliers, and how did you handle them?

Since the large majority of the data represented categorical data, I didn't need to worry much about outliers. The only possible outlier was in the question about the number of kids the respondent had of 16 and 20. After examining the raw data surrounding the information about the kids, it seemed to be an input error since the next lowest was 8. So, I decided to cap the number of kids to 8. This also allowed an easy transition to categorical data since it can simply be interpreted as 8+ as its own category.