For my final project I decided to explore data from the General Social Survey. I ended up with five variables relating to the topics I was interested in exploring.

Hypotheses:

* There is a correlation between certain political views (liberal vs. conservative) and certain religious beliefs.
* There is a correlation between certain political views and some aspects of household employment (income, hours typically worked, perceived “prestige” of occupation).
* There is a correlation between certain political views and the respondent’s level of commitment to education.
* Political views can be predicted using regression analysis on these correlated variables.

Outcome of EDA:

The outcome of this project was a failure to reject any of the null hypotheses. There appeared to be some significant correlations at first, but they were extremely weak and did not hold up to further testing. Similarly, my regression analysis resulted in very weak pseudo-R-squared values that are unlikely to have any practical value.

What do you feel was missed during the analysis?

I would have liked to try some other methods for calculating correlations, such as the Kendall’s Tau. There were also many assumptions that I would have liked to check more completely.

Were there any variables you felt could have helped in the analysis?

I think there are many other variables that might have been more useful than the ones I selected. I started with many more variables of interest, and I probably should have made my selections more carefully.

Were there any assumptions made you felt were incorrect?

I tried to avoid obvious violations of assumptions, but I’m sure I did miss some. I did not the chance to perform very many assumption tests, particularly in regard to dealing with categorical variables.

What challenges did you face, what did you not fully understand?

The biggest challenge for me was that I failed to understand the variables well enough before committing to them. I did not really appreciate that the variables I selected were almost all categorical, nor did I consider how much that would change my assumptions and methods. I very possibly may have applied methods that are not suitable for categorical data, and I felt that the very small number of possible categories for each variable made the analysis less interesting.