STAT6021 Project 2 Proposal

Diana Morris - dmd8a

Grace Lyons - kat3ac

Alex Bass - ujb3bu

Andy Ortiz - eao7r

### Objectives/Goals

The data set we chose for our project consists of survey responses from people who were surveyed about political beliefs and lifestyle choices. The study focused on identifying political and demographic divides between urban, suburban, and rural communities. The data set consists of many categorical variables with two or more response options that give the respondents options to choose from and are then coded with a corresponding number. There are three quantitative variables that measure the respondents' level of confidence for Donald Trump, Barack Obama, and Mike Pence.

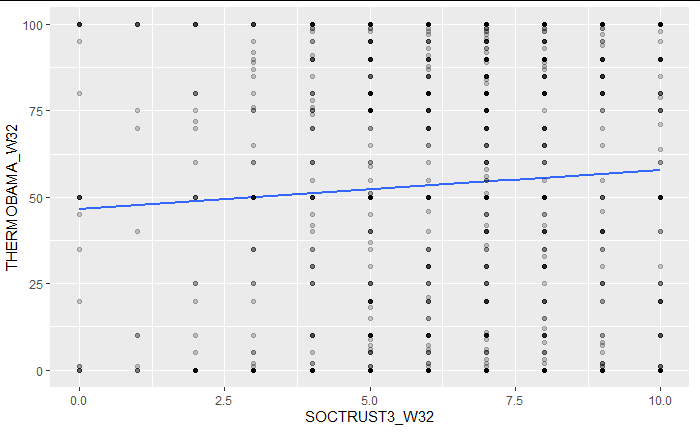
Since this data set is based on survey data, we expect there to be some data cleaning that will need to be done. The survey did not have forced responses, so if a question was skipped by the respondent, it was coded with a ‘99’ or ‘999’. There are also many NAs in the data set that will need to be removed. The number of NA responses differs by variable.

We selected this dataset because we are interested in observing how political beliefs and lifestyle choices differ in the three different communities. We plan to look at how the individual presidential rating is predicted by different variables such as Urban/Suburban/Rural status, abortion stance, social trust level, etc. We are interested in looking at how these answers may vary between respondents who rated Trump and Pence similarly or differently. We are also interested in exploring how fear, or lack of social trust, predicts a presidential candidate’s rating. Finally, we plan to perform logistic regressions on categorical variables we find interesting, such as if the respondent currently lives near where they grew up; if the respondent currently has enough income to lead the life they want; or if the respondent is satisfied with the role of government. We are interested in using presidential rating, demographic variables, and age as predictors in the logistic regressions.

### Visuals

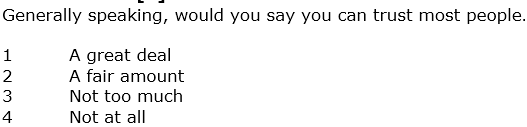
**Linear regression hypothesis and related exploratory figures:**

With linear regression models in mind to test our question about fear/distrust in relation to presidential candidates, we created a scatter plot and regression line of the trust rating (on a scale of 1 to 10) vs. feelings (on a scale of 1 to 100) about former President Obama.



We see a positive correlation between the two, but further analysis will be necessary to see if this is a meaningful result. Ultimately, we will involve several categorical variables in our linear regression, potentially including the respondent’s economic optimism, rootedness in the community, and feelings towards newcomers.

To obtain a clearer view of this particular question (how trust affects the rating of various politicians), the following plots were created:

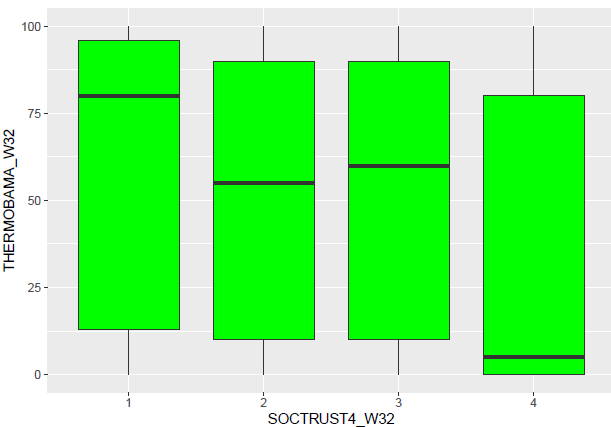
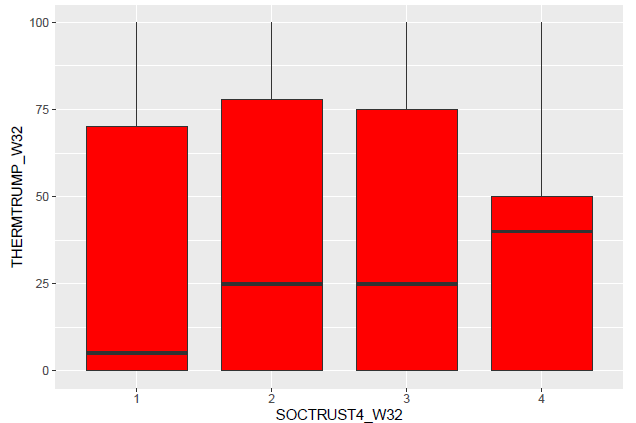
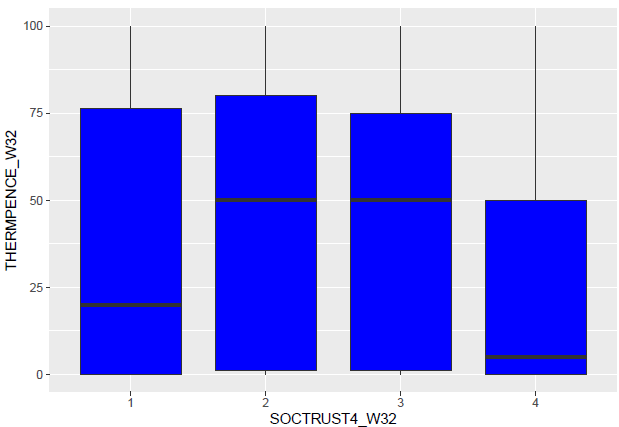
X axis : categorical variable “SOCTRUST4\_W32”

Y axes: quantitative variables (0-100)

**THERMTRUMP.** How do you feel toward Donald Trump?

**THERMBOBAMA.** How do you feel toward Barack Obama?

**THERMPENCE.** How do you feel toward Mike Pence?

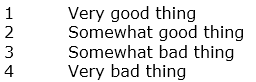


Although the medians diverge a bit, there is large overlap. Our hypothesis about using the trust variable may not be enough by itself, or there may be other trust related variables we could use.

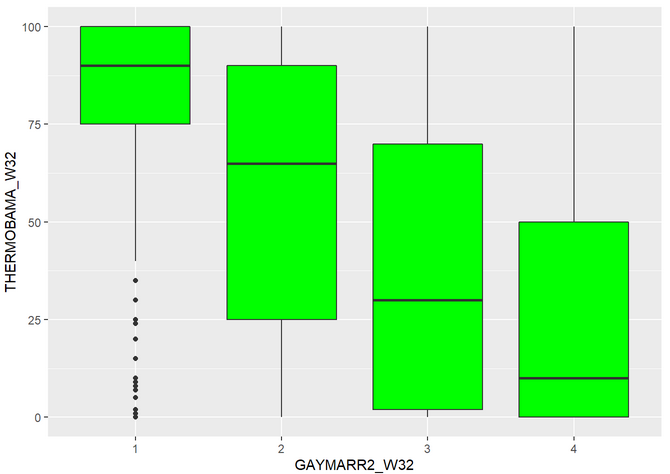
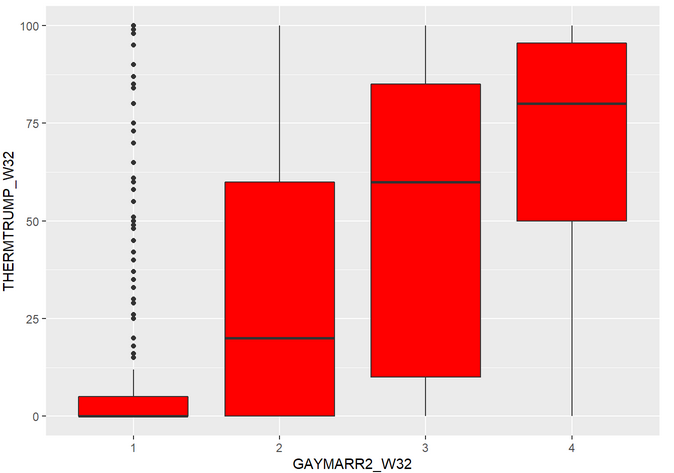
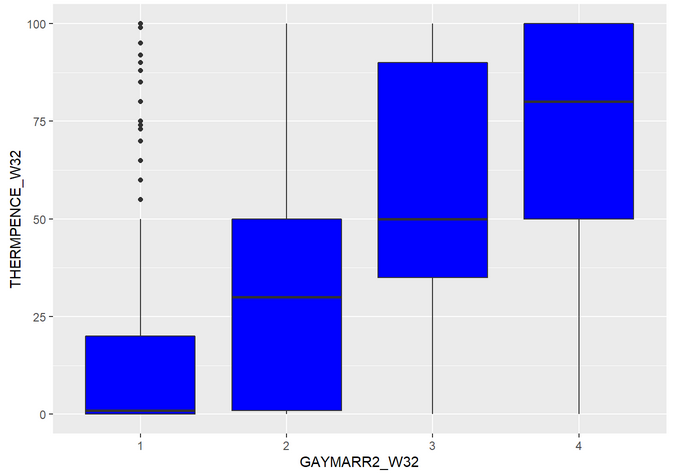
Using a different categorical variable,

X axis : categorical variable “GAYMARR2\_W32”

“As you may know, same-sex marriage is now legal in the U.S. Do you think this isa good thing or a bad thing for our society?”



Y axes: quantitative variables (as above)



There is still overlap between the categories, but now the medians diverge more. Since this is a first step, we will need to further explore the data to determine if a combination of variables can be found to better predict the presidential candidate rating.

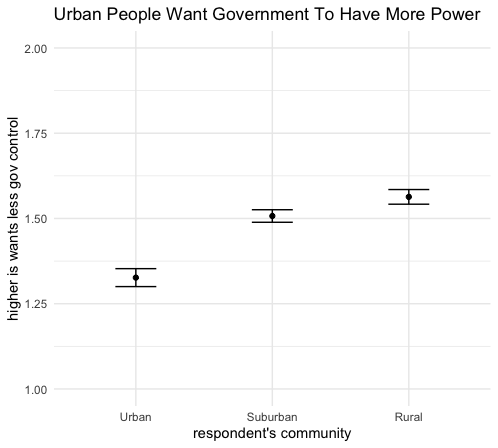
**Logistic regression hypothesis and two related exploratory figures:**

Respondents were asked, “Which statement comes closest to your view even if neither is exactly right?” and were given the following two options…

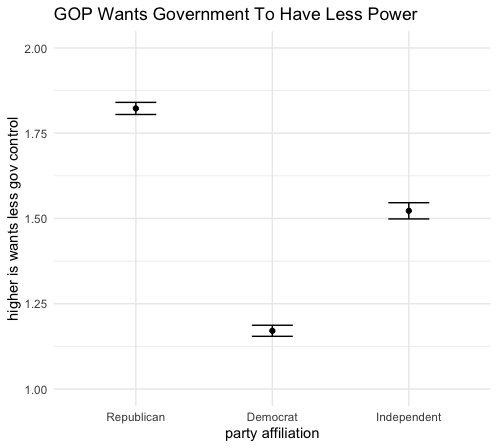
1. Government should do more to solve problems
2. Government is doing too many things better left to businesses and individuals

Hypothesis: Controlling for other variables such as a person’s race, age, and party, **rural US residents** are more likely to say the government is doing too many things, and **urban US residents** are more likely to say the government should do more to solve problems. This hypothesis is founded on our theory that living in a rural area perhaps is one way to express skepticism of government generally as government presence is more visible in a city environment.

In the figures below, ***the mean of each group is displayed*** where 1 = Government should do more to solve problems, and 2 = Government is doing too many things better left to businesses and individuals.



One of the primary challenges of our hypothesis is trying to measure the independent effect of a respondent’s community on government preference. Because this is a binary tradeoff question, we can use logistic regression to estimate the independent effect of a respondent’s community of government preference. One important variable we will control for in our analysis is the respondent’s party affiliation. As seen in the figure below, Republicans largely prefer statement 2 and Democrats prefer statement 1.



For our predictors, we plan to include demographic information such as age, gender, race, and party. We will also evaluate other demographic variables asked in the survey and include them in the model based on feature selection techniques.