Capstone 1 Report

James Olmstead

Introduction

Context

With the 2020 Presidential Election underway, multiple candidates are jockeying to win the Democratic party nomination and eventually defeat incumbent Donald Trump. At the same time, the Republican party is attempting to find the best course of action to ensure the President can win a second term. Both parties rely on appealing to and understanding voters' concerns in order to win.

The Problem Statement

What are the common ideological patterns among the American electorate? At base level we can examine it from a left-right spectrum, but I hope to break it down into more detailed analysis of the ideological groupings. For example, how do voters' feelings on immigration correlate with their feelings on health care? And is it possible to categorize voters into ideological sets that can be labeled and defined?

The Dataset

I plan on using the 2018 Survey Full Data Set from the Democracy Fund Voter Study Group. This data has extensive demographic and ideological information on voters from across the political spectrum and has been utilized by high-profile media organizations such as the New York Times, Vox, and FiveThirtyEight.

Justification and Outcomes:

By categorizing voters into ideological groups, political candidates can tailor messages and platforms to appeal to certain types of voters and find a winning coalition. They can better speak the language of the average voter and appeal to whichever category(ies) to win.

Cleaning

Data Cleaning Steps

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 had 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.

Missing Values

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.

Outliers

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.

Analysis

My primary questions I wanted to answer for my initial analysis were:

1. How are the responses correlated, and what is the strength of the correlations?

2. Which political issues are priorities for the survey respondents?

3. Where do the respondents stand on the major party platforms?

Since most of the data is summarized in the documentation via frequency tables, I hoped to study correlations and visualize some key questions in the survey.

How closely correlated are the responses?

To answer this question I used the Cramer's V statistic for measuring association between nominal variables. I was able to create a matrix for each pairwise association measure and examine the results. One interesting thing I found was that the questions with the highest frequency of counts in the table had to do with watching certain media channels (e.g. Fox, MSNBC) and self-described political ideology (e.g. Conservative, Environmentalist).

A Question of Affiliation

Since I am looking to eventually categorize respondents by political labels, I decided to take a closer look at how the respondents identified themselves. This seemed especially pertinent since many of the highly correlated questions were these labels. Specifically, I decided to see what sort of overlaps there were between the affiliations based on the Cramer's V. This portion of the survey asked them to give a yes/no answer on if they identified with a certain political affiliation. Those affiliations were Libertarian, Socialist, Green, Environmentalist, Liberal, Moderate, Conservative, Radical, Progressive, Traditional, Christian, Feminist, Fundamentalist, and None.

A great way to visualize the results is a coefficient matrix. This makes it easy to see the pair-wise test results. These matrixes are on the next page below as Figures 1 and 2.

Most nearly every pair-wise test was statistically significant, with a few exceptions mostly being tests involving Libertarian, Moderate, and Fundamentalist. In addition, most of the results for the associations themselves are to be expected. The strongest associations are between Green/Environmentalist and Liberal/Progressive both having Cramer's V values above .5. On the other hand, associations between Traditional/Christian wasn't as high as I would have guessed (only .29) while Environmentalist/Feminist is higher than I thought (.33) since there is no inherit similarity besides being associated with left-wing politics. It is also important to note the relatively strong association for Conservative/Liberal (.36) is likely a negative one i.e. those who identify as Conservative don't identify as Liberal and vice versa.

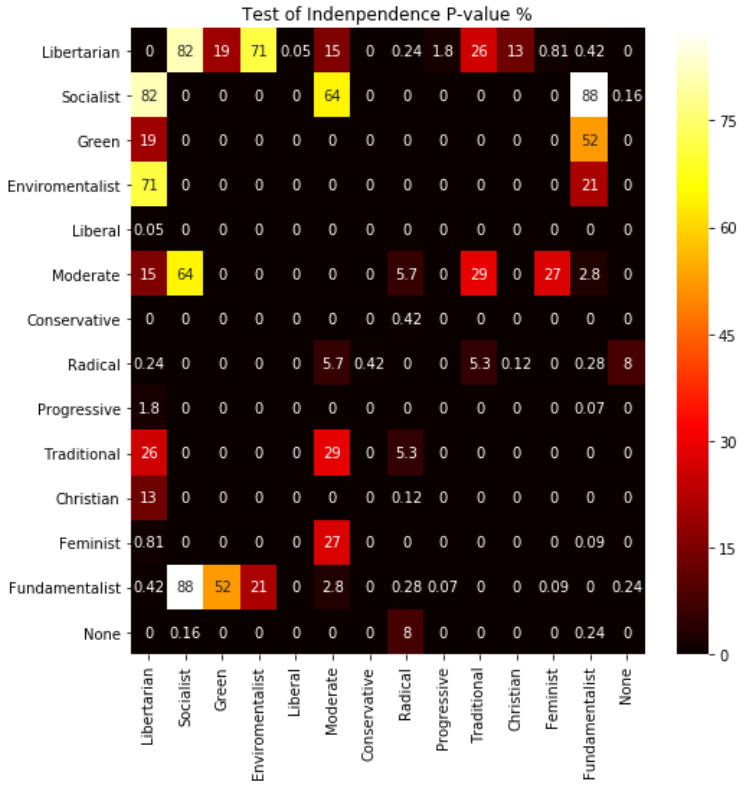

Fig. 1

Pairwise p-values for the chi-squared statistic. Any value below 5% is considered statistically significant.

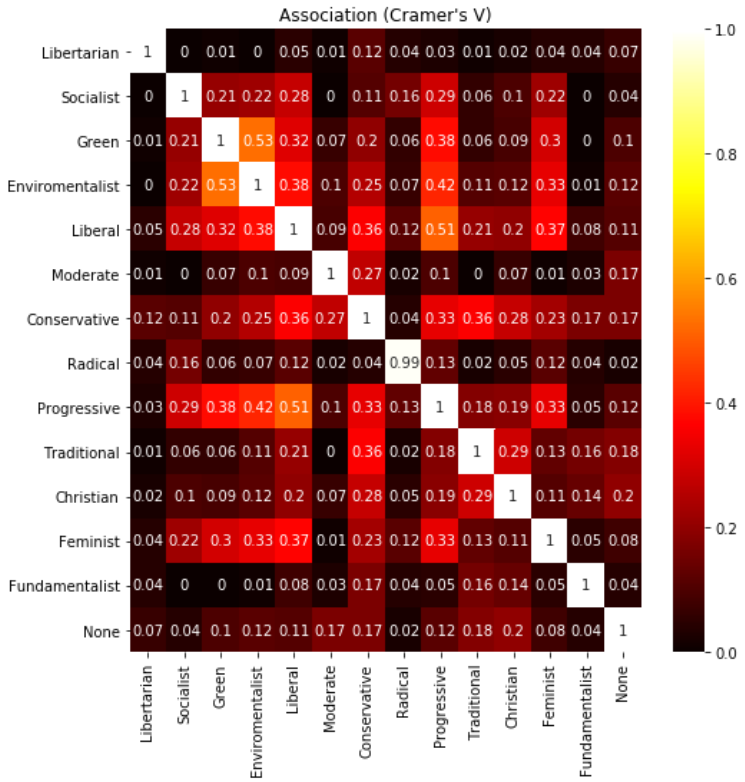

Fig. 2

Cramer's V association. Values closer to 1 signify a closer association. Keep in mind association can be positive or negative in direction. (E.g. answering 'No' to conservative may be strongly associated with 'Yes' for liberal.)

Which political issues are priorities for the survey respondents?

For this question I examined a series of questions where respondents rated how important various political issues were to them. For example, someone may rate 'Social Security' as 'Very Important' or another may rate 'Climate Change' as 'Not Very Important'. I was able to visualize these responses for each of the 23 topics and break them down by year and 2016 vote for President. One of the 23 visualizations is below as Fig 3:

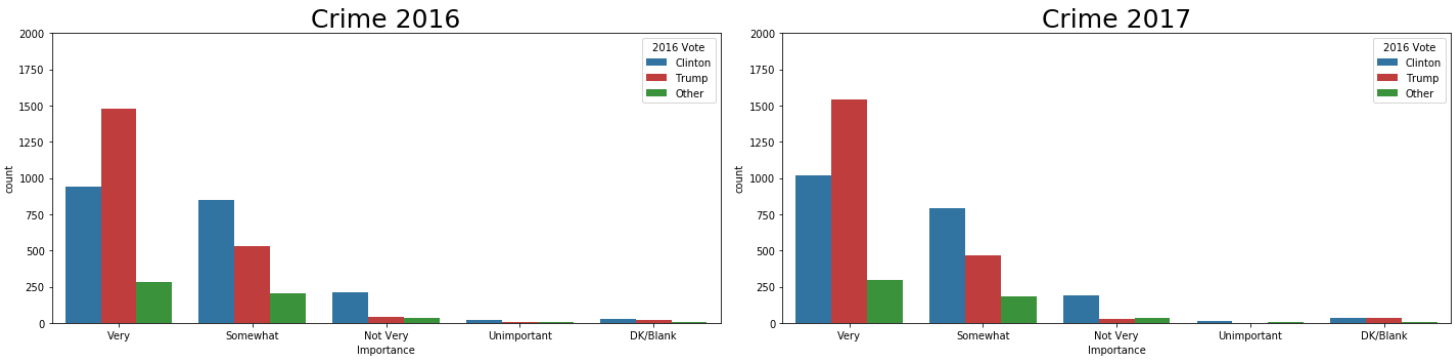

Fig. 3 Breakdown of issue importance by year and 2016 vote

As you can see, Trump voters generally put more emphasis on Crime than Clinton voters and that little changed between 2016 and 2017. The large majority of plots for the other 23 topics showed little to no change between 2016 and 2017.

Where do the respondents stand on the major party platforms?

To answer this question, I took a look at questions relating to the major party agendas (Democratic and Republican). How do respondents feel about how each party is addressing 12 major issues? And how does it differ between Trump and Clinton voters? Similarly, to the previous section, I was able to break down the results by the parties themselves and by respondents 2016 vote. Note that the party that each person was asked about was randomly assigned. One of these 12 plots is below as Fig 4:

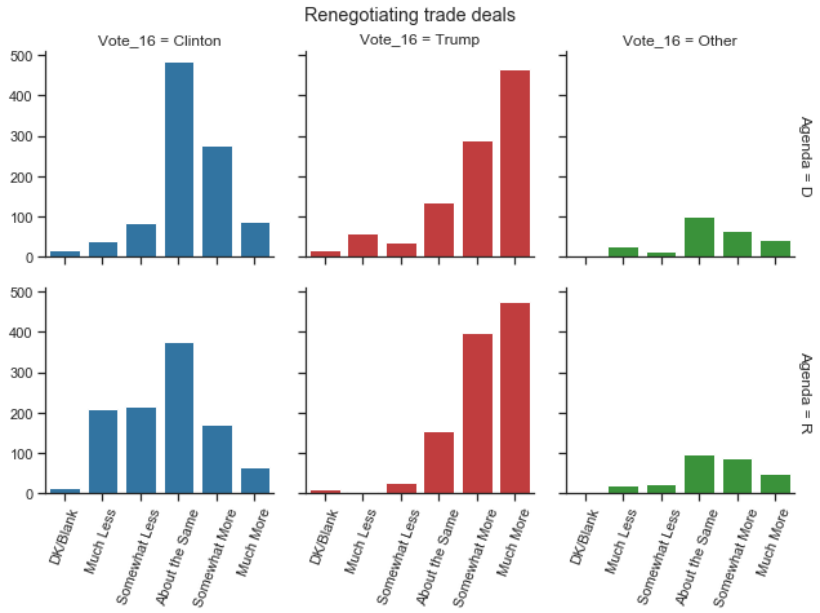

Fig. 4

Views of party platforms. 1st row is on D party, 2nd is R party. Colors/Columns are 2016 Vote

My 'Data Storytelling' and 'Data Analysis' Jupyter notebook goes into much more detail on additional observations on the results.

Checkpoint

Overall, this is an extremely deep and extensive dataset with a huge amount of insight and information to glean. While most of the in-depth analysis is in my Jupyter notebook, I'll briefly answer the questions I put forward for my analysis.

1. How are the responses correlated, and what is the strength of the correlations?

   The responses were somewhat correlated, and the strongest correlations often had to do with ideological affiliation and news consumption.
2. Which political issues are priorities for the survey respondents?

   The top issues were the Economy, Health Care, and Social Security. Some of the biggest gaps between Trump and Clinton voters were on climate change, the budget deficit, and the environment.
3. Where do the respondents stand on the major party platforms?

   As a general trend, it seemed that Clinton voters were much happier with their party's platform than Trump voters were for Republicans. Trump voters indicated with high frequency that they wanted the Republican party to focus 'much more' on a number of issues, including 'Reforming the Health Care System', 'Reducing how much Americans pay in Taxes', and 'Creating Jobs'. They also rated the Democratic party at almost the same rate, which was surprising.

In-Depth Analysis

The In-Depth Analysis of the data consists of the following:

1. Reduce the dimensionality of the data using multiple component analysis (MCA).
2. Utilize the K-means clustering algorithm to create groups for the respondents.
3. Dig further into the newly created clusters to find insights and how they differentiate from one another.

Multiple Component Analysis

As it stands now, the input data has 642 features of categorical variables, which is a lot to handle. I wanted to reduce the number of features in order to trim down the unnecessary or redundant information in the data. To this end, performed MCA on each Section of the data to reduce it to each of its core components. This allows for more interpretable components while still trimming the data significantly. It also converts the data into continuous variables instead of discrete, which is more conducive to plotting.

I needed to make sure we choose enough components to capture the essential information in the data, but not too many to add unnecessary noise. Therefore, we can look at the explained inertia of each component and see where the cumulative difference begins to shrink and choose the cutoff there. This is known colloquially as the 'elbow' method.

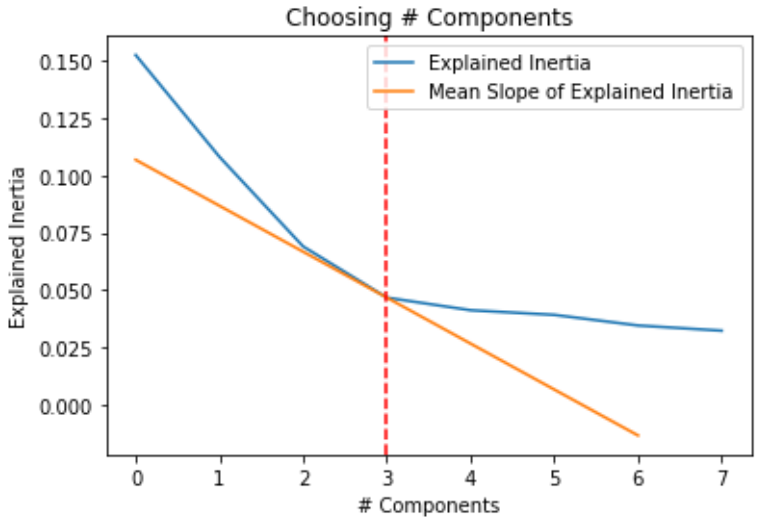

Fig. 5

An example of the 'elbow' method. In this case, we choose 3 components for this section since that is where the slope of the inertia begins to level

Once I generated the components, the next step is to perform our clustering algoritm.

K-means Clustering

One of the nessecary inputs to perform a k-means clustering is choosing the parameter k, the number of clusters. I choose k clusters based on an examination of the how the inertia changes based on different values of k, but also based on how clean the results came out with some trial and error. I end up choosing k=9 clusters.

After the algorithm finished creating clusters for the data, I needed to interpret what the clusters meant in relation to the questions. I was able to tie each cluster back to the components that generated it, and from the components to the original survey questions. This allowed me to see common threads that made up the 9 different clusters and see patterns in how they answered distinctive survey questions. After a good amount of digging, I was able to come up with my own labels for the 9 clusters.

1. Far Left: Hold views well to the left of most respondents.
2. Left: Hold views to the left of most respondents, but not as extreme as the Far Left.
3. Christian Left: Hold mostly mainstream views of political left, but are more likely to be African-American and Christian
4. Lean Left: Mainly have moderate views, but lean towards the political left in most respects.
5. Moderate/Unengaged: Don't follow political news closely and don't have strong political opinions.
6. Lean Right: Mainly have moderate views, but lean towards the political right in most respects.
7. Populist Right: Hold strict views on immigration and are distrustful towards large institutions.
8. Classic Right: Hold views to the right of most respondents, but are more open to immigration and free-markets than the Populist Right.
9. Far Right: Hold views well to the right of most respondents.

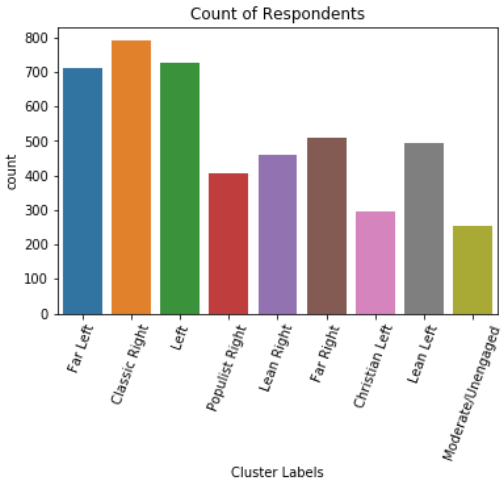

Fig. 6

Bar graph showing the counts of the different labeled clusters.

Now that I had the clusters labeled, I wanted to see how well they mapped into two dimensions for visualizations purposes and to see the performance of the algorithm. Needless to say, the plot below shows that it worked quite well. Based on the position of the clusters in 2-D, I was able to get an idea of how the axes could be interpreted as well.

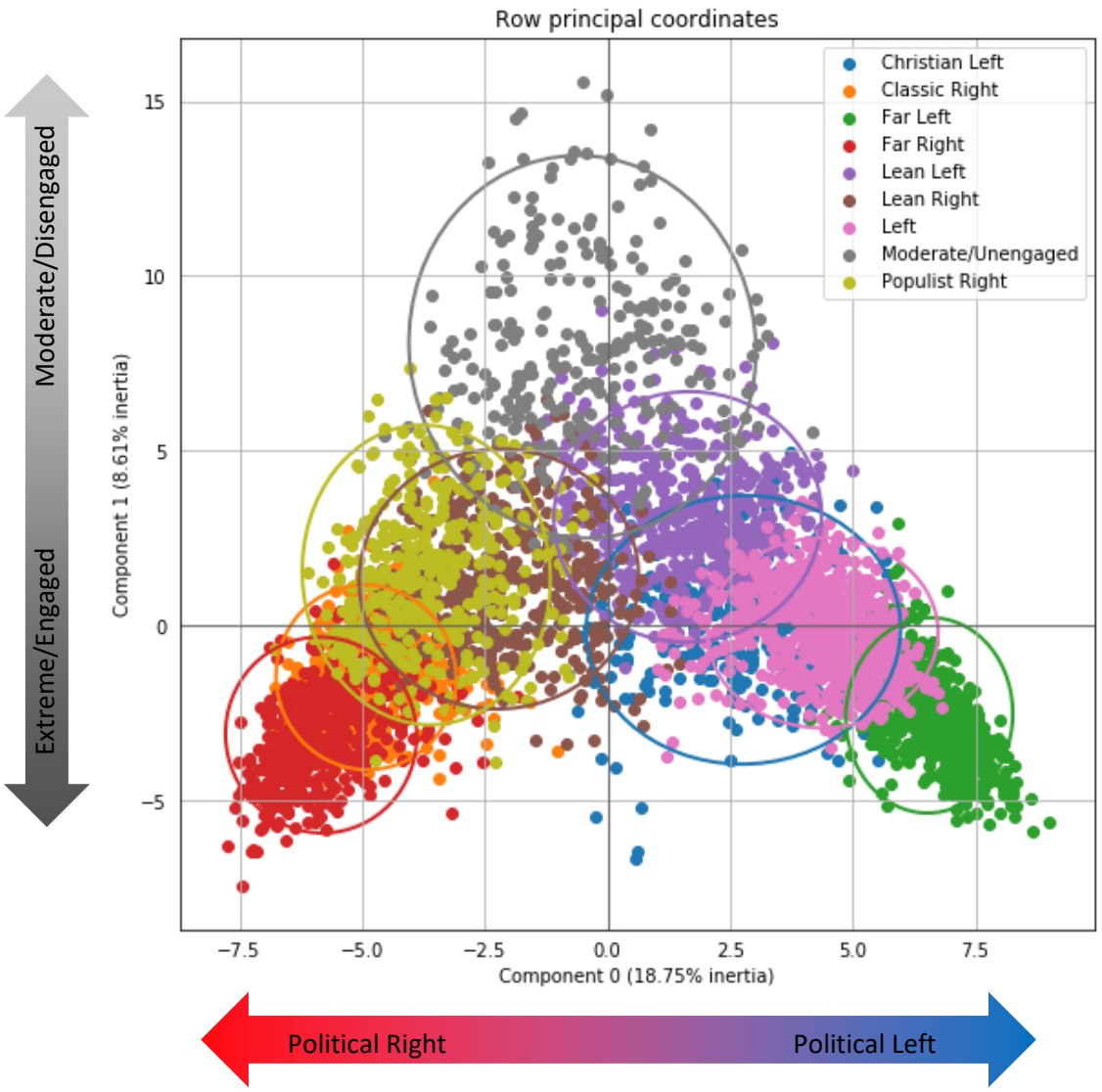

Fig. 7

2-D projection of the respondents, color-coded by their labeled cluster.

Examination of Results

Our first look at these clusters is how each of the respondents self-identified themselves. These series of questions were basically 'Do you identify as _? Yes or No?' with the blank being different political labels like 'Feminist' or 'Conservative'. The table below is what percentage of each group answered 'Yes' to identifying as the given label.

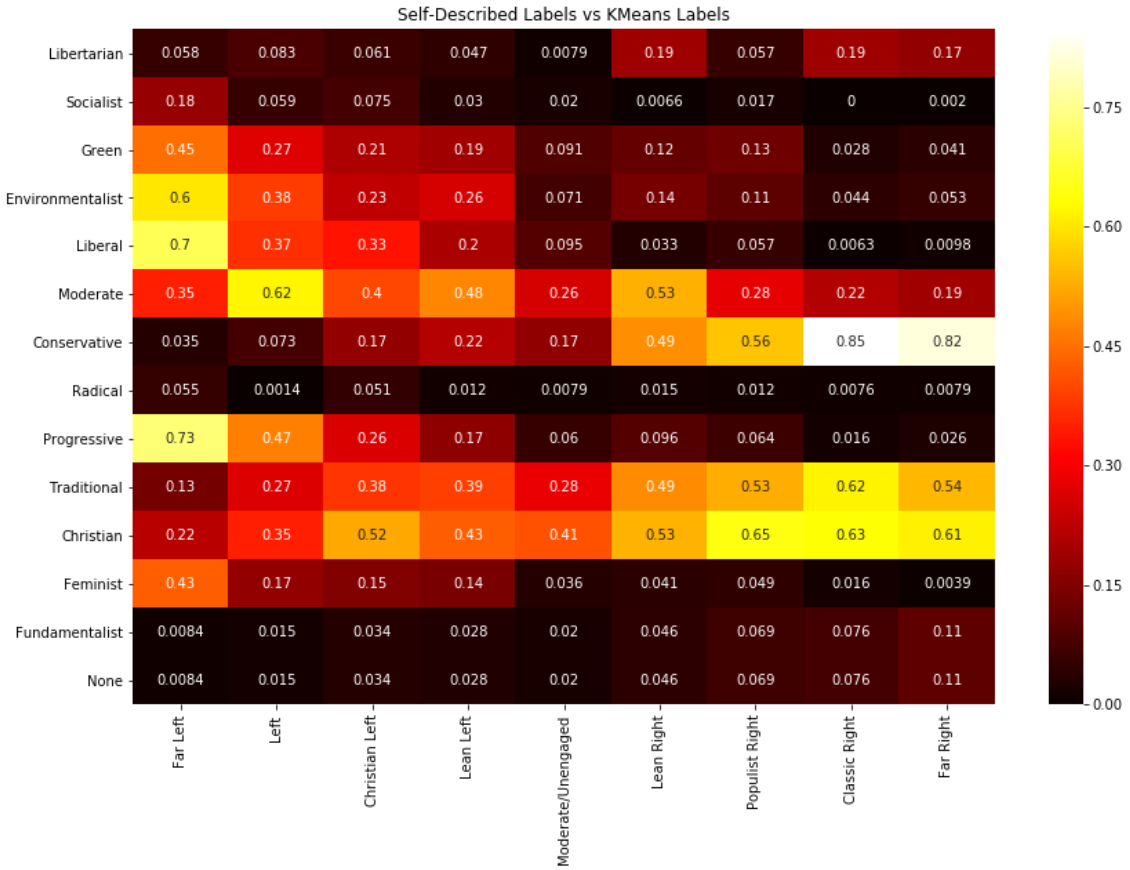

Fig. 8

Table showing how different groups identify with other political labels from the survey questions.

Next, we can look at how the different labels feel about President Trump, which is a pretty good lens into their views generally during this political climate.

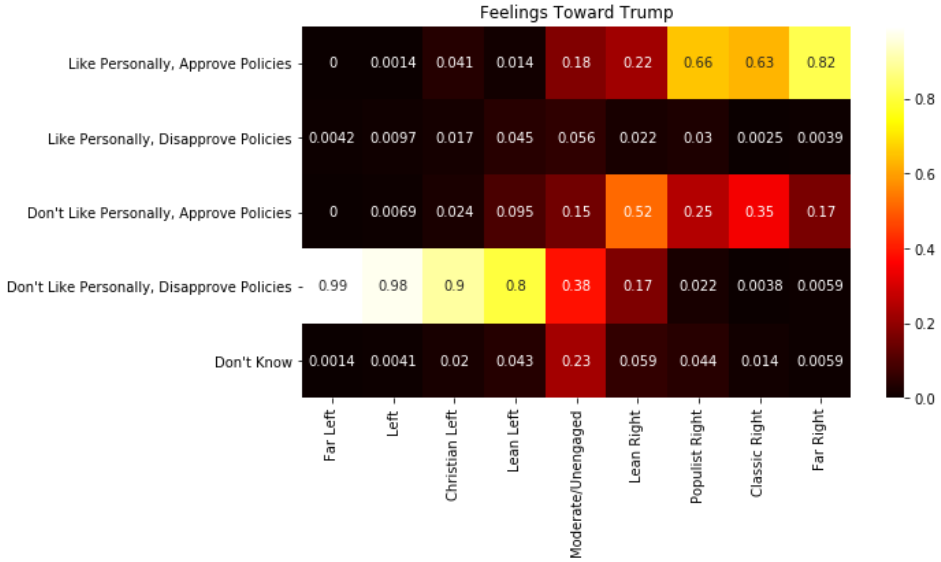

Fig. 9

Table showing how different groups feel about President Trump. Note each column adds to 1 since it is a percentage value.

Finally, we can look at how the respondents feel about immigration to the U.S.

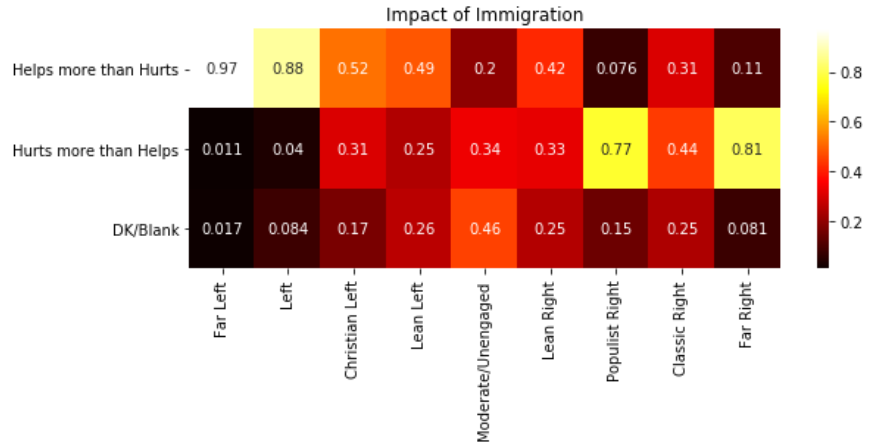

Fig. 10

Table showing how different labels feel about the impact of immigration. Note each column adds to 1 since it is a percentage value.

Note that I examined a number of additional questions of interest in my Jupyter notebook including race, big business, and religion.

Conclusion and Future

I was pleasantly surprised how defined the different clusters were and how well they map to existing political groups in the real world. There is certainly plenty of more insight to be gleaned from these clusters as well. In the future I may want to look into removing more 'Section-skipping' components to see if that allows more clusters to form without being warped. I may also want to try different clustering algorithms to see how well they perform. Finally, I could try bucketing the questions into larger Sections so that there are fewer components after the MCA is completed.