**Analyzing the American Electorate**

Springboard Data Science

Capstone Project 2

Josh You

**Introduction**

Politics is a major part of American life. Many Americans vote and participate in politics, and the stakes in elections are often high. In this project, I investigated the nature of the American population and electorate using clustering analysis to determine how many different political groups Americans are divided into, and how these groups differ in their political and demographic characteristics.

**Client**

My client for this project is politicians and political campaigns interested in learning more about the American electorate. Knowing more about different groups of voters may provide useful insights. For example, if a large bloc of voters is separated out by their views on a given issue, that provides information for how to message to that group of voters. This can help campaigns understand the voters that tend to already support their party, and learn which groups of voters may be open to persuasion and outreach.

**Dataset**

The data for this project comes from the Cooperative Congressional Election survey, which is an annual survey that is carried out as a cooperative effort between dozens of universities across the US. In 2016, the survey had 64,600 respondents, and was conducted by 60 different teams. The dataset and guide are publically available [here](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GDF6Z0&version=4.0). The survey consisted of over five hundred different questions, covering demographic and socioeconomic characteristics such as race, ethnicity, religion, income, and education level, as well as a variety of questions about the respondents’ political opinions, party identification, and how they voted in the 2016 election for President, Congress, and governor races. The survey was originally conducted before the election, and the teams followed up with the same respondents after the election in a post-survey.

**Data Cleaning**

Because of the large number of columns, perhaps the most important cleaning step was choosing which columns I would use in my project. I first uploaded the dataset to a Pandas dataframe, interpreting blank values and cells containing “\_\_NA\_\_” as missing values, and looked at how many missing values there were in each column in order to get a list of candidate columns to be included. I then compiled a spreadsheet where I reviewed each column’s description in the dataset guide and made a decision whether to include that column or not.

*Missing values*

I then dropped rows that did not answer the post-election survey. I decided to do this because presidential vote is one of the columns I am most interested in, and I wanted to look at how people actually voted, not just their plans before their election. There were 52899 rows remaining after this.

After these steps, there were still some columns with a fair number (several thousand or more) missing values. I first checked whether I could infer the missing values from other columns. I could do this because some survey questions were not asked to some respondents based on their responses to other questions. For example, a respondent who already said they did not vote were not then asked whether they voted for president. I was able to fill in values this way for the presidential vote column (based on whether the answer to whether the respondent voted), and the hispanic column (based on the respondent’s answer to the race question).

I then had to deal with missing values that were not encoded with blank or NA values. For example, many questions had a “Don’t know” or “Prefer not to say” response, which is implicitly missing information. I then replaced these values with NaN in the Pandas dataframe. This left several columns with over 1000 missing values, including the “sexuality” and “trans” columns. For these columns, the large majority of responses were “Heterosexual” and “No”, respectively, so I imputed the missing values using those values. This still left a few columns with a small number of missing values, and I dropped the rows that had those missing values. This resulted in 48130 rows remaining, out of the 52899 who took the post-survey.

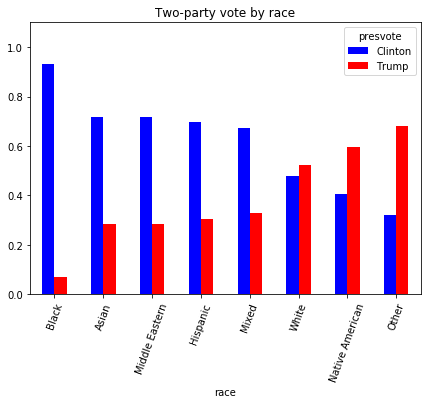
*Column and value transformation*

After dealing with missing values, the next step was transforming values and columns into a format that is amenable to machine learning analysis. Essentially, I had to make sure all columns were either (categorical) binary variables or numerical variables. For binary text responses (Yes/no, support/oppose, etc.), I converted those values to 1 and 0. For categorical text responses with multiple options, I had to create dummy variables (a binary variable for each category). To reduce the number of columns, I collapsed some of the categories. For example, I converted five-option questions (strongly support, support, neutral …) into three-option questions (support, neutral, oppose), before turning those options into dummies.

I also had to convert the family income variable, which previously contained information about which income range the respondent’s family belonged to (e.g. $10,000 to $19,999, $20,000 to $29,999, etc.), into single numerical values. I did this by taking the midpoint of each range. For the response “$500,000 or more”, I converted this to the bottom of the range (500,000).

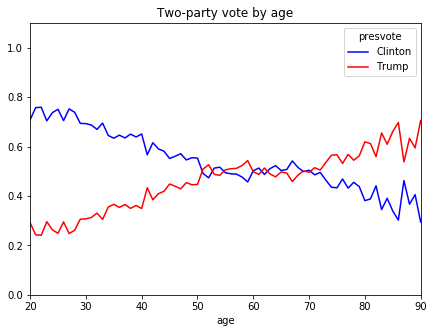
**Exploratory Data Analysis**

For exploratory data analysis, I created plots breaking down respondents’ presidential vote by the respondent’s characteristics. For example, here is presidential vote by race, just including Clinton and Trump:

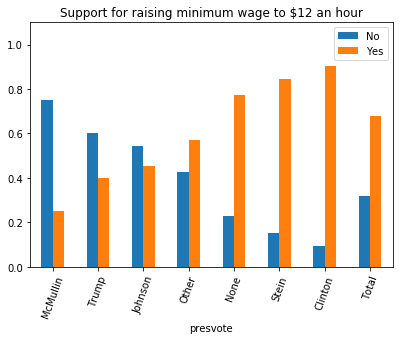


Black and Asian respondents largely preferred Clinton over Trump, while white respondents were more likely to prefer Trump, and Native American and "Other" respondents were most likely to support Trump. The Native American result surprised me given that other ethnic minorities mostly support Democrats, though a cursory Google search did not turn up information on how Native Americans usually vote. Note that there were only 401 respondents who identified as Native Americans in this dataframe.

Now here is presidential vote by age. The relationship is quite striking; support for Trump rises fairly consistently with age.

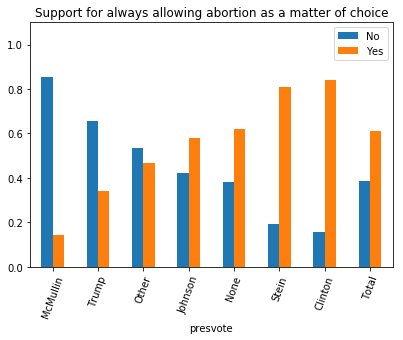


I also created plots showing respondents’ responses to questions about their political views, broken down by how they voted for president. For example, here is support for raising the minimum wage to $12/hour, by presidential vote:



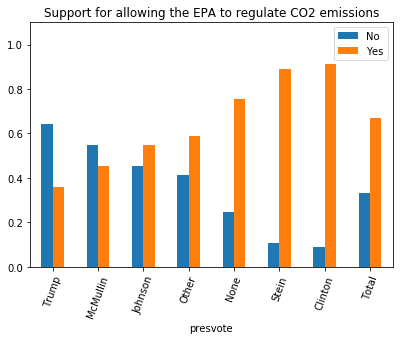
On raising the minimum wage, Clinton voters are the most supportive, while McMullin voters the least supportive, followed by Trump voters.

This graph shows respondents’ support for abortion:



On abortion, McMullin voters are by far the least supportive, followed by Trump voters, and Clinton voters are most supportive.

Here is support for allowing the EPA to regulate CO2 emissions:



On CO2 regulations, Clinton and Stein voters are most supportive, and Trump voters are the least supportive.

Overall, these responses are consistent with the positions of the presidential candidates and their parties, showing that voters mostly agree with the views of the candidate they support.

The positions of Johnson/Stein voters may be surprising. Stein voters were less "liberal" on minimum wage and CO2 emissions, despite my impression that Jill Stein is very liberal and the Green Party is focused on environmental issues. In addition, Johnson voters were more likely to support raising the minimum wage and allowing the EPA to regulate CO2 emissions than Trump voters, suggesting that Johnson voters are not all "hard-core" libertarians.

In addition, while non-voters are have opinions that are in between those of the conservative candidates and the liberal candidates, they more often than not take the “liberal” side of issues such as raising the minimum wage.

**Clustering Analysis**

The next phase in my project was using unsupervised learning to separate the data into clusters. I used an implementation of the k-means algorithm from Python’s scikit-learn. K-means clustering randomly picks *k* points as cluster centers, assigns each example to a cluster based on which point is closest to it, and then iteratively recalculates the cluster centers as the average of the clusters that were created and reassigns each datapoint to the closest cluster center.

*Removing correlated features*

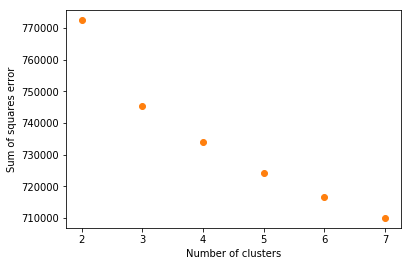
Since distances are computed across every column in unsupervised learning, if some features are very similar to others, this introduces a potential source of bias, as the same underlying feature gets counted multiple times. Accordingly, I computed the Pearson correlation coefficient between pairs of features to determine which features were most correlated with each other.

Unsurprisingly, responses many questions about the same political topic were highly correlated. For example, answers to whether they support allowing employers to decline abortion coverage in their healthcare plans and whether they support banning federal spending on abortion are correlated with a coefficient of 0.65. Support for strengthening the Clean Air act is correlated with allowing the Environmental Protection Agency to regular CO2 emissions (0.58) and with creating renewable energy mandates (0.58). I chose to drop the federal abortion question because it was more correlated with the other abortion questions than the abortion coverage question. I dropped the Clear Air and renewable energy mandate questions because the EPA question felt more important - climate change is perhaps the most important and salient environmental issue, and the question is closely linked to the Obama administration’s Clear Power Plan, a major policy to regulate carbon emissions through the EPA.

In addition to this, there were five different questions about state legislature spending (do you support increasing/maintaining/decreasing spending on X?). I noticed, after looking at the clusters when choosing k = 4, there appeared to be a pattern between two clusters in terms of whether they chose "increase" or "maintain" on the state spending questions. This seemed to be a major factor that divided these two clusters, and this felt fairly spurious and exaggerated by the large number of questions about state legislature spending.  
  
I excluded three of these questions (about spending on healthcare, infrastructure, and law enforcement) on the grounds that other questions already cover healthcare policy (repeal ACA), infrastructure policy, and police spending (a question which asks whether they support increasing the number of police by 10 percent).

*Choosing k*

Obviously, choosing the number of clusters was an important decision for this project. One measure of how good a clustering is is the sum of squared errors (SSE), or the sum of squared distances between every datapoint and its cluster center. Increasing the number of clusters will always reduce the SSE, but that does not mean more clusters is always good. One method of choosing which cluster based on SSE is the “elbow” method, which means looking for a value of k after which the SSE declines more slowly as k increases. The graph below shows SSE plotted against number of clusters.



3 is perhaps a good candidate for an “elbow”, but that is pretty unclear.

Another method for evaluating a cluster is the silhouette score. The silhouette score measures how similar a point is to other points in its own cluster compared to how similar points are to points in other clusters. Thus, it measures both how well points match their own clusters, and how well the clusters are separated. The average silhouette scores for each point is shown for several choices of k:

For k = 2: 0.108305054698  
For k = 3: 0.101409875779  
For k = 4: 0.093843294352  
For k = 5: 0.0884931801442  
For k = 6: 0.0544745113861  
For k = 7: 0.0490876142195

K = 2 had the highest silhouette score. This is an interesting result because we generally think of Americans as being divided among two major parties, and this provides evidence in favor of that. The silhouette score goes down somewhat as k increases to 5, and sharply decreases after that.

However, another consideration is which number of clusters will produce the most informative results when we analyze how the clusters differ, when thinking about what questions people actually have about the American public. Two clusters may lead to the best separation between clusters, but we are still interested in how those two clusters break down into subgroups. This ultimately led me to choose to focus my analysis on five clusters, which ended up breaking down into two clusters that were mostly liberal/Clinton voters, two that were mostly conservative Trump voters, and one that was more moderate and had a high proportion of non-voters.

*Feature Analysis*

Now that I had my clusters, I need to find out how they differ. My basic strategy was to train supervised learning models to predict each point’s cluster label, and then see which features played the biggest roles in these models.

I first trained a random forests model. After optimizing hyperparameters, the model achieved an average accuracy of 84.4% after five-fold cross-validation. This was a model predicting five different classes - the recall (the percentage of points in a category that are successfully labeled as that category) is was 0.83 and 0.95 for the five clusters, and precision (the number of points labeled as a given category that are actually in that category) was between 0.78 and 0.95. Thus, the model was fairly good at predicting all five classes.

Below, I list the top ten most important features of this model. Feature importance is a measure of how much influence a given feature had on the model, compared to if that feature was excluded:

pres\_Trump: 0.0804831520081  
repealACA: 0.0629656018686  
Obama: 0.0599725219875  
EPACO2: 0.0590967565517  
Party\_Republican: 0.0345146793422  
religimp\_Very important: 0.0344066894898  
whiteadvantage\_Disagree: 0.0317513998146  
minwage12: 0.0297117108141  
primary: 0.0293260029879  
pew\_bornagain: 0.0285907038281

Presidential vote (pres\_Trump is a dummy variable created from presidential vote), party identification, and a few key policy views (regarding the Affordable Care Act, EPA, and raising the minimum wage $12/hour) were among the most important features.

I also trained a logistic regression model, which achieved 90% accuracy. Logistic regression produces coefficients for each feature indicating whether that coefficient positively or negatively predicts a given class. I print the five coefficients with the highest absolute values for each cluster below. A positive coefficient means that the feature predicts membership in the class, and a negative coefficient means that the feature is negatively associated with that class.

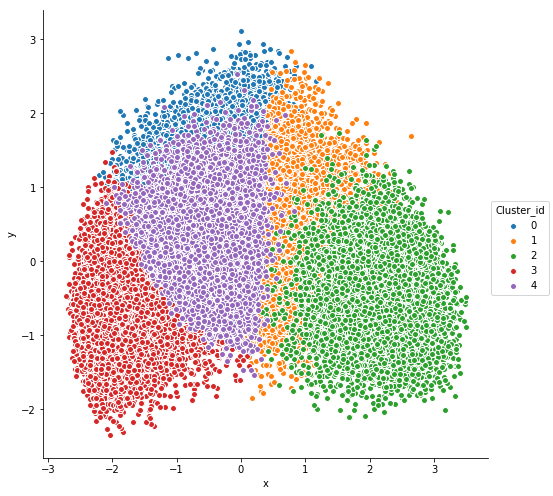
Cluster 0   
  
pew\_bornagain: -1.21241895109  
gaymarriage: 1.1819284388  
Romney: -1.16311661427  
ideo\_Very liberal: -1.12564206944  
religimp\_Very important: -1.10598932116  
  
Cluster 1   
  
pres\_McMullin: 2.01096436001  
pres\_Trump: 1.96535232927  
pres\_Other: 1.74433986275  
pres\_Johnson: 1.71886246057  
Party\_Republican: 1.03887752788  
  
Cluster 2   
  
EPACO2: -6.07877399917  
banassault: -5.0758811388  
fuelefficiency: -4.7141164115  
minwage12: -4.5890247662  
TPP: -3.94274754674  
  
Cluster 3   
  
repealACA: -4.01345745144  
ideo\_Moderate or Not sure: -3.7566010487  
abortion20wks: -3.69284571163  
pres\_None: -3.3756046346  
campaigndonate: 3.36551036545  
  
Cluster 4   
  
religimp\_Very important: 1.62500079969  
pres\_Trump: -1.39192731233  
Romney: -1.38143020145  
Obama: 1.24849158841  
campaigndonate: -1.22098746112

There is significant overlap between these features and the most important features for the random forest model. For example, repealACA is the highest coefficient for cluster 3 and is one of the most important features for random forests, and religious importance is in the top five for clusters 0 and 4 (with opposite signs) while also being in the top ten for random forests. However, each cluster has some unique features that distinguish them, which I will explore below.

**Exploratory Data Analysis with Clusters**

*Principal Component Visualization*

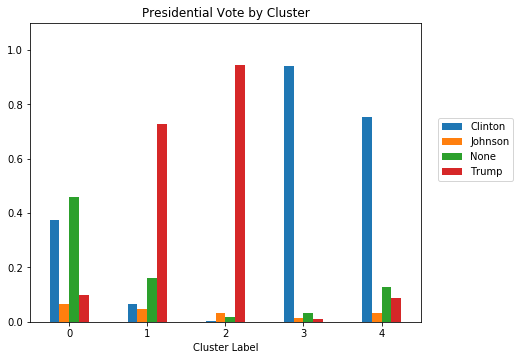
There were over 100 columns in the data used to create the clusters, so the data cannot be visualized directly. Principal Component Analysis (PCA) provides a means to reduce the dimensionality of the data. PCA identifies the vectors that account for the highest variance possible, or best describe the dimensions along which the data differs. Using scikit-learn to generate two principal components and graph the points in two dimensions using these components as the axes. This visualization is shown below:



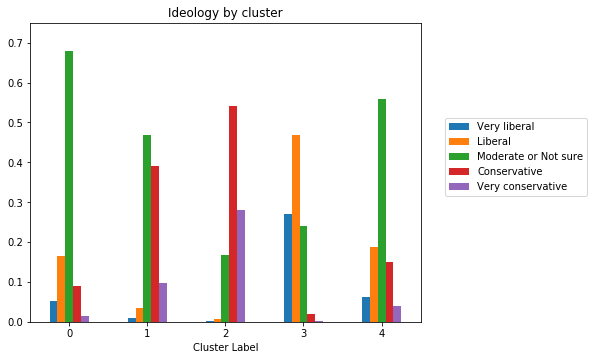
The size of the clusters may be misleading - cluster 0 actually had 8776 points, out of 48144 in total. The cluster sizes ranged from 6561 to 13343.

*Visualizing differences between clusters*

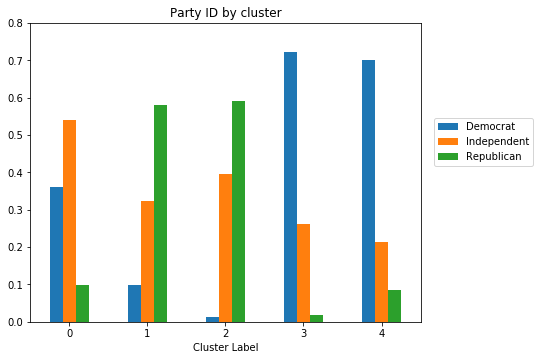
I will now show a few graphs illustrating the differences between the five clusters, using some of the key features identified. Here is presidential vote by cluster:



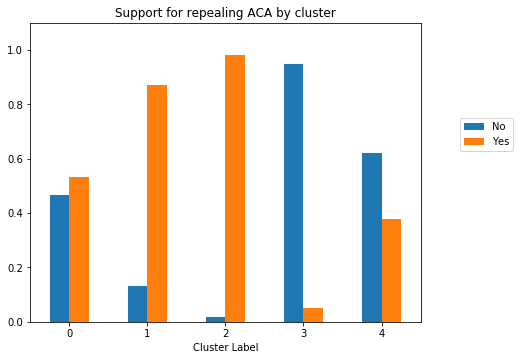
Cluster 0 stands out for having by far the highest proportion of non-voters; those that did pick a candidate tended to vote for Clinton. 1 and 2 are largely Trump voters, but 2 supported Trump to an overwhelming margin. 3 and 4 are largely Clinton voters, but 3 is overwhelming Clinton voters.



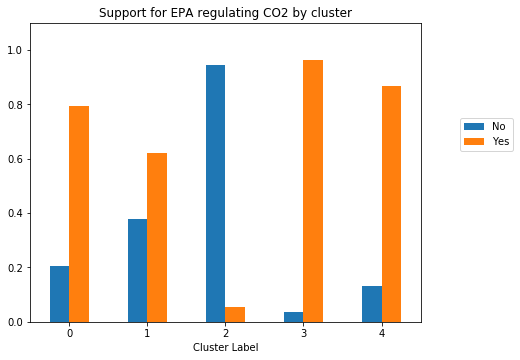
0 is largely moderate, with a small number of liberals and a smaller number of conservatives. 1 has a plurality of moderates, with the rest mostly conservative, while 2 is mostly conservative or very conservative. 3 is mostly very liberal or liberal, while 4 is mostly moderate.



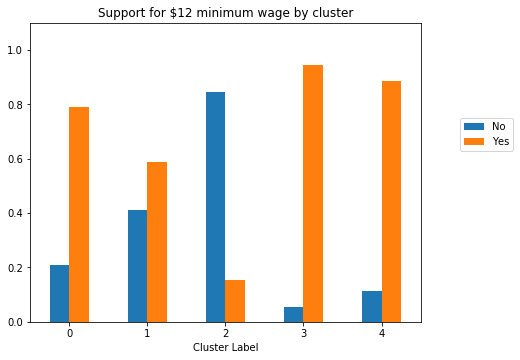
In terms of party identification, 0 is mostly independent, not surprisingly. Among the other clusters, the more ideological clusters (2 and 3) actually have a higher proportion of independents than 1 and 4, but a lower proportion of people identifying with the minority party within their cluster.



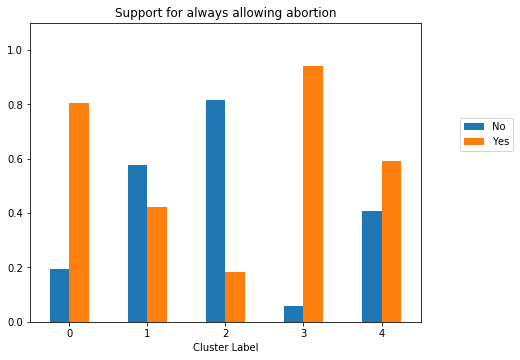
For repealing the Affordable Care Act, the two conservative clusters (1 and 2) are fairly similar, but far more people in 4 and 0 supporting repealing the Affordable Care Act than in cluster 3.



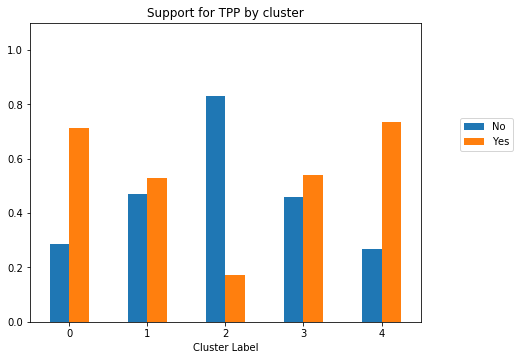
This plot shows that there is a major difference in environmental policy between the conservative clusters, with 2 overwhelmingly opposing the EPA regulating CO2, and cluster 1 supporting it.



Minimum wage is another issue that divides the clusters 1 and 2, with cluster 1 supporting raising the minimum wage and cluster 2 opposing it. Every other cluster overwhelming supports raising the minimum wage.



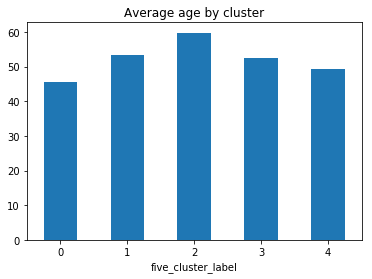
Support for abortion is another issue in which there is a noticeable difference between the two Trump-heavy and the two Clinton-heavy clusters. Cluster 4 is far more likely to oppose always allowing abortion.



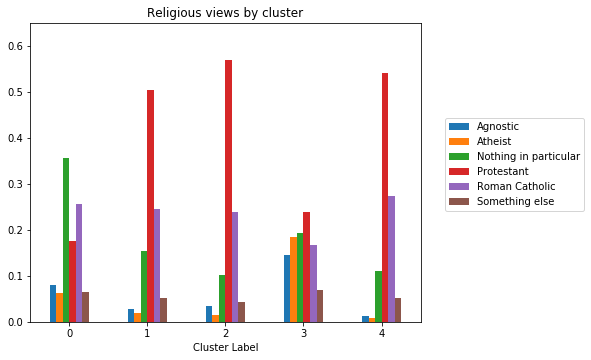
Every cluster supports the TPP (Trans-Pacific partnership, a trade agreement negotiated between the US and several Asian countries which Trump withdrew from in 2017) except for cluster 2. Interestingly, the most liberal cluster (3) is less likely to support the TPP than the less liberal Clinton cluster (4). This likely reflects a split that occurred during the Democratic primary, wherein Bernie Sanders was vocally opposed to the TPP while Clinton was more ambivalent. Unfortunately, this dataset does not contain information on whom the respondents voted for in the primary.

Now I will turn to looking at the clusters’ demographic features:

Here is average age by cluster:

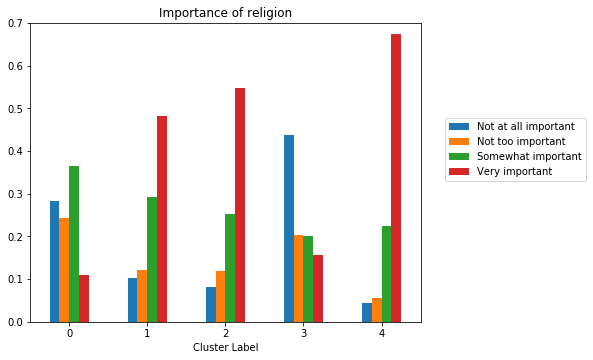


0 is youngest of all, 2 is older than 1, and 3 is older than 4. Thus, the more ideologically extreme clusters have older members.

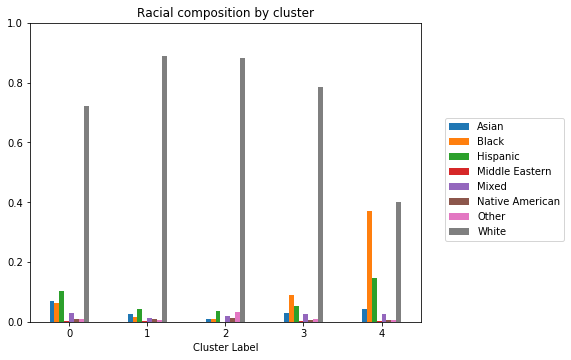


The most common religious view is Protestant for all clusters except 0, which has a plurality saying Nothing in Particular. 4 is much more Protestant, and more Catholic than 3. 3 has the most atheist and agnostic respondents.

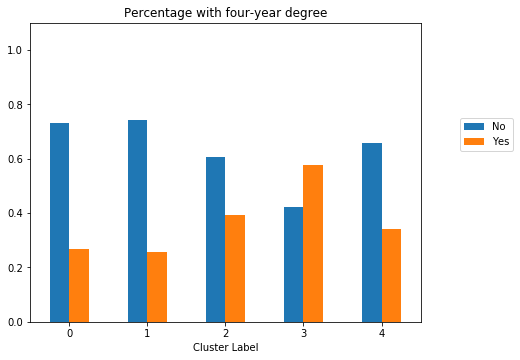
Next, here is how important religion is in the respondent’s life, grouped by cluster:



There is a striking gap between clusters 3 and 4, with cluster 4 being far more likely to say that religion is very important in their life. Clusters 1 and 2 both tend to view religion as important compared to cluster 3 and cluster 0.

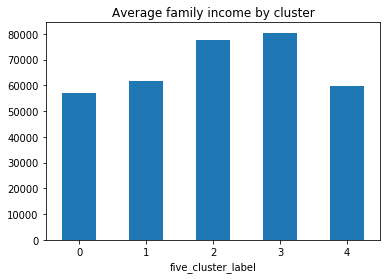


In terms of race, cluster 4 stands out by having by far the highest proportion of non-whites. Clusters 1 and 2 are more white than clusters 0 and 3.



In terms of college education, cluster 2 is more educated than cluster 1, which is the least educated of all the clusters. Cluster 3 is the only cluster where the majority of people have a four-year degree.

Finally, here is average family income by cluster:



Clusters 0 and 4 have the lowest incomes, followed by cluster 1. Clusters 2 and 3 (the highly conservative/highly liberal clusters) have the highest incomes.

**Discussion**

The first takeaway is that the United States really is a partisan nation that is roughly divided into two camps: Democrats and Republicans, or alternatively liberals and conservatives. There are multiple lines of evidence. First, there are high correlations between party identification, how people vote, and how people think about political issues, with a sharp divide in opinion between Republicans/Trump voters and Democrats/Clinton voters. Additionally, the number of clusters with the highest silhouette score was 2, indicating that dividing the American population in half leads to the best separation into groups of people that are similar to others in their group, while being different from the other groups. Finally, when moving to higher numbers of clusters, most people belonged to the four clusters that overwhelmingly voted for either Clinton or Trump.

That being said, there are meaningful differences between voters for the two major candidates. On the Clinton side, there is a bloc (cluster 4) that is less educated, has more racial minorities, is much more religious, and is somewhat more likely to disagree with the “liberal” position on some issues such as repealing the ACA and allowing abortion, while the other group is more ideological and much more educated than the rest of the population.

On the Trump side, voters are divided into two groups, one of which is older, wealthier, more educated, and more ideologically conservative, and the other is younger, much less educated, and less ideological to the point that they tend to agree with the “left” position on political issues such as economic issues.

Finally, the cluster containing the most non-voters is young, is relatively uneducated, and has mostly liberal-learning political opinions.

*Practical implications:*

This analysis has several potential implications for political campaigns. Knowing how political opinions are associated with each other, as well as with demographic features, may influence how parties tailor their strategy by community and location. Richer and more educated liberals and conservatives are more ideological. Ideological conservatism (e.g. economic conservatism) will play better among richer and more educated communities than in poorer and less educated communities, even if both communities are nominally Republican.

On the other side, Democratic voters are split into along education, religion, and race. While non-white Americans largely vote for Democrats, they may have different views on religion and social issues than white Democrats/liberals, which should impact how Democratic campaigns discuss these issues.

Persuading moderate voters, and people who tend not to vote, is a major goal for any politician. These people are grouped together by my clustering analysis, and have tend to have political opinions that are closer to the Democrats than to the Republicans. The marginal non-voter may be more easily persuaded to turn out to vote for a politician that promises to raise the minimum wage than one that does not promise to do so.