"Who votes for Trump ?"

Pramod Padmanabhan

21st July, 2016

Introduction

The year 2016 will be remembered down the ages for the 58th quadrennial Presidential election of the United States as it gets highlighted by the candidates running for the most powerful position on this planet. The Democrats being, most-likely, represented by Hillary Clinton and the Republicans by the exuberant billionaire, Donald Trump. Though these are the two main contenders in these elections, Senator Bernie Sanders will live in the minds of the younger generation, that witnessed these elections, for his feasible socialism and populist ideas.

Despite Sanders wanting to rewrite the election process in the United States through his revamped funding campaigns by seeking donations from individuals and his fresh breath of what leadership should look like, he seems unlikely to achieve the endorsement of the Democratic party. Nevertheless history will still be made by the end of this year with the United States either electing its first ever Woman President or giving the job to a fascist leader in Donald Trump.

In this project we will look at how on earth did Donald Trump get nominated by the Republicans by studying the kind of Americans who favored him. The characteristics we will look at include the race, the occupation, their income levels, the amount of education, their cosmopolitan outlook etc.

Data and the Variables

To study the above mentioned characteristics of Trump's voters we need data sets from quite disparate sources. So our final data set with all the relevant variables is one obtained as a result of cleaning up and standardising all these different data sets. We briefly look at the election process in the United States and then describe the dependent quantity and the variables that we would use to predict it.

The United States is one of the largest democracies on this planet and they have a long and tedious election process which stretches for an entire year. Though the Constitution of the United States allows any number of parties to field candidates for the election, there are two major political forces that have determined the politics in this nation since its birth. They are the Republican Party or the Grand Old Party (GOP) and the Democratic Party. Before every general election they choose the candidate to run for the President through a process known as the primaries. The primaries is a democratic way of choosing the candidate from each party to represent them in the general elections for the President's office. All registered voters who have pledged support to either the Republican or the Democratic party take part in this process. This process varies by state and takes place over the course of a few months before the general elections. The fraction of votes obtained by the various candidates in a primary in a particular state, is converted to a number of delegates from the state according to a rule set by the party. The candidate who reaches the cutoff, of a minimum number of delegates needed to run for the Presidency, first represents the respective party in the general election.

Thus our dependent quantity is the number of votes obtained by Trump in the primaries held in the 46 states in 2016. The primaries are held in all the counties of each state and as the sizes of the states vary, the number of counties also vary. To nullify the effect of this we normalize the dependent variable by computing the percentage of votes obtained by the candidate, Trump in this case, from each state. Thus we end up with 46 observations. The results of the primaries from all the counties in the 46 states was obtained from - https://www.kaggle.com/benhamner/2016-us-election.

We would like to study the dependence of the percentage of Republican votes obtained by Trump on a number of variables. The variables chosen are based on Trump's ideologies and how they may have influenced Republican voters. These include :

Hate group Variable :

Variable 1 - The **Klans** variable measures the number of active Ku Klux Klans (KKK) in each state from the 1920's and 30's. Though this data is from the past we suspect that they still have an effect on the present day voters. These numbers are also divided county wise in each state and we work with the percentage of Klans from each state as our normalized variable.

Education/Employment/Poverty Variables :

Variable 2 - The **Education** variable indicates the percentage of counties of a state that was considered poorly educated. A county is considered poorly educated if at least 20 % of its voters, between the ages of 25 - 64 years, did not have a high school diploma or its equivalent.

Variable 3 - The **Employment** variable indicates the percentage of counties from a state that was considered to have low employment. A county is considered to have low employment if less than 65 % of its residents between the age of 25-64 years were employed in the years 2008 - 2012.

Variable 4 - The **Poverty** variable measures the percentage of persistent poverty of counties of a given state. A county was classified as having persistent poverty if 20 percent or more of its residents were poor as measured by the 1980, 1990, and 2000 decennial censuses and the American Community Survey 5-year estimates for 2007-11.

Economic Variables :

Variable 5 - The **Farming** variable indicates the percentage of counties that have farming as a major source of revenue. A county is considered as farming-dependent if it accounted for 25% or more of the county's earnings or 16% or more of the employment averaged over 2010-2012.

Variable 6 - The **Mining** variable shows the percentage of counties that have mining as a major profession. This is so if mining accounted for 13% or more of the county's earnings or 8% of the employment averaged over 2010-12.

Variable 7 - The **Manufacturing** variable shows the percentage of counties that have manufacturing as a major revenue source. This is the case when manufacturing accounted for 23% or more of the county's earnings or 16% of the employment averaged over 2010-12.

Variable 8 - The **Government** variable shows the percentage of counties where a lot of people were employed in Federal/State government jobs. This is consdiered to be true when Federal and State government accounted for 14% or more of the county's earnings or 9% or more of the employment averaged over 2010-2012.

Variable 9 - The **Recreation** variable compiles the percentage of counties that depend on recreation as a source of income. See documentation page: http://www.ers.usda.gov/data-products/county-typology-codes.aspx

Variable 10 - The **Nonspecialized** variable shows percentages of counties that were not involved in any of the above economic variables.

Development Variable

Variable 11 - The **Urbanization** variable indicates the percentage of Metros in a state. Metro is considered Urban whereas non-metro is considered rural. Metro areas include all counties containing one or more urbanized areas: high-density urban areas containing 50,000 people or more; metro areas also include outlying counties that are economically tied to the central counties, as measured by the share of workers commuting on a daily basis to the central counties. Nonmetro counties are outside the boundaries of metro areas and have no cities with 50,000 residents or more.

Racial Indicators

Variable 11 - **White** shows the percentage of Whites in the state.

Variable 12 - **Black** shows the percentage of Blacks in the state.

Variable 13 - **Native** shows the percentage of natives in the state.

Variable 14 - **Asian** shows the percentage of Asians in the state.

Variable 15 - **Islander** shows the percentage of Pacific Islanders in the state.

Variable 16 - **Mixed** shows the percentage of people belonging to two or more races in the state.

Variable 17 - **Hispanic** shows the percentage of Hispanics in the state.

Income Variable

Variable 18 - The **Income** variable describes the Median household income during 2009-2013 for each state.

Visualizing the Characteristics of Donald Trump's Voters

In this section we will study the characteristics of Trump's supporters. The regression tree plots that follow have the percentage of votes from the Republican voters of each state on the nodes. These votes were taken from the results of the Republican primaries held in 46 states as of July 13th, 2016. The states not included are Colorado, District of Columbia (DC), Maine, Minnesota and North Dakota making the number of observations 46.

We also make scatter plots using the important variables. The points here vary in size according to the percentage of votes obtained in that particular state.

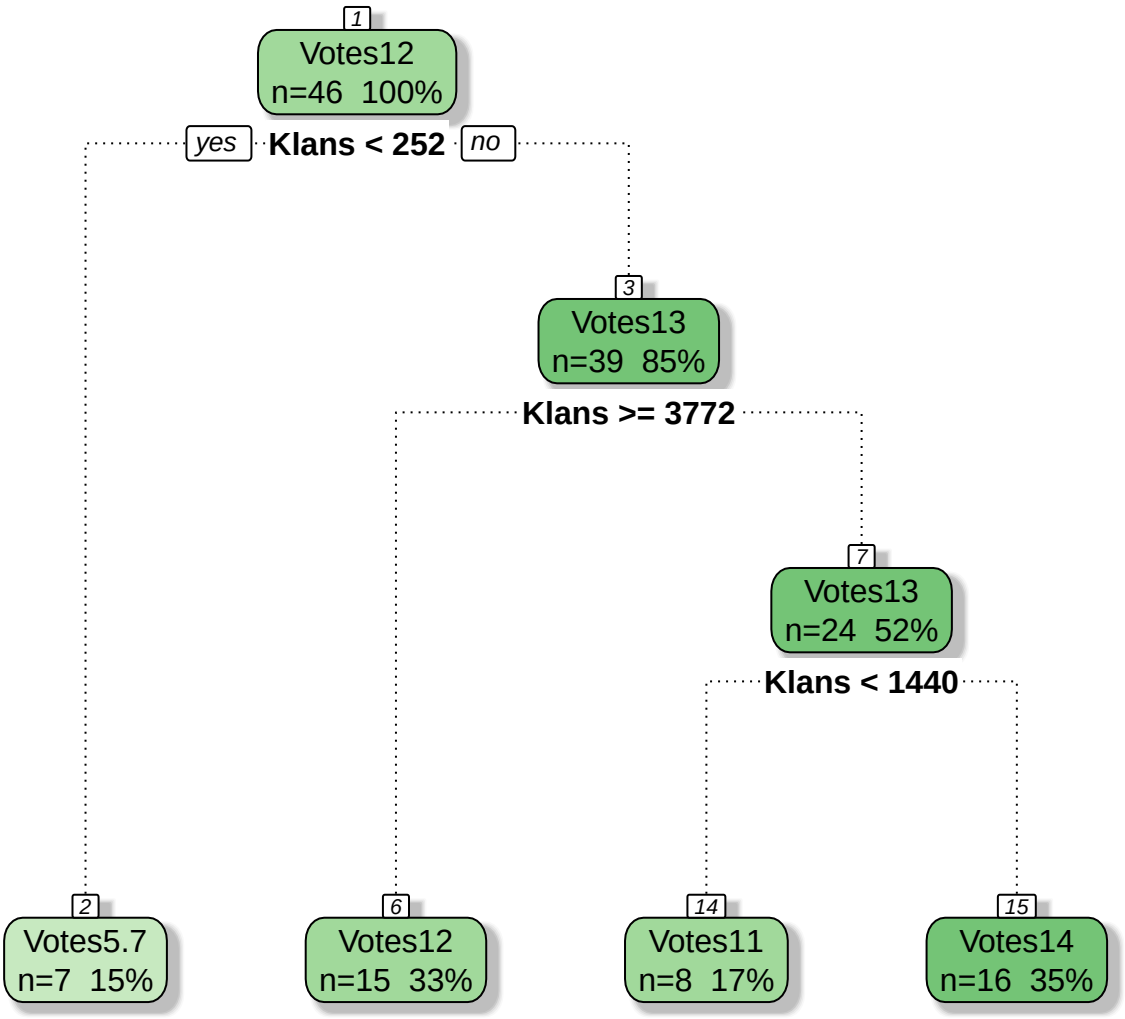
The variables or the characteristics that we wish to study are the following : the racial composition of the state, the level of urbanization of the state, the type of economy of the state (that is if the state is predominantly a farming, mining, manufacturing state), the level of education of the voting poulation, the level of employment and the poverty levels of the various states, the mean household income in the state and finally the number of Ku Klux Klans that were operating in the state in the past !

Several of these variables can have correlations with each other and most of them are completely independent. Taking this into account we make regression plots based on groups of variables that can possibly interact with each other.

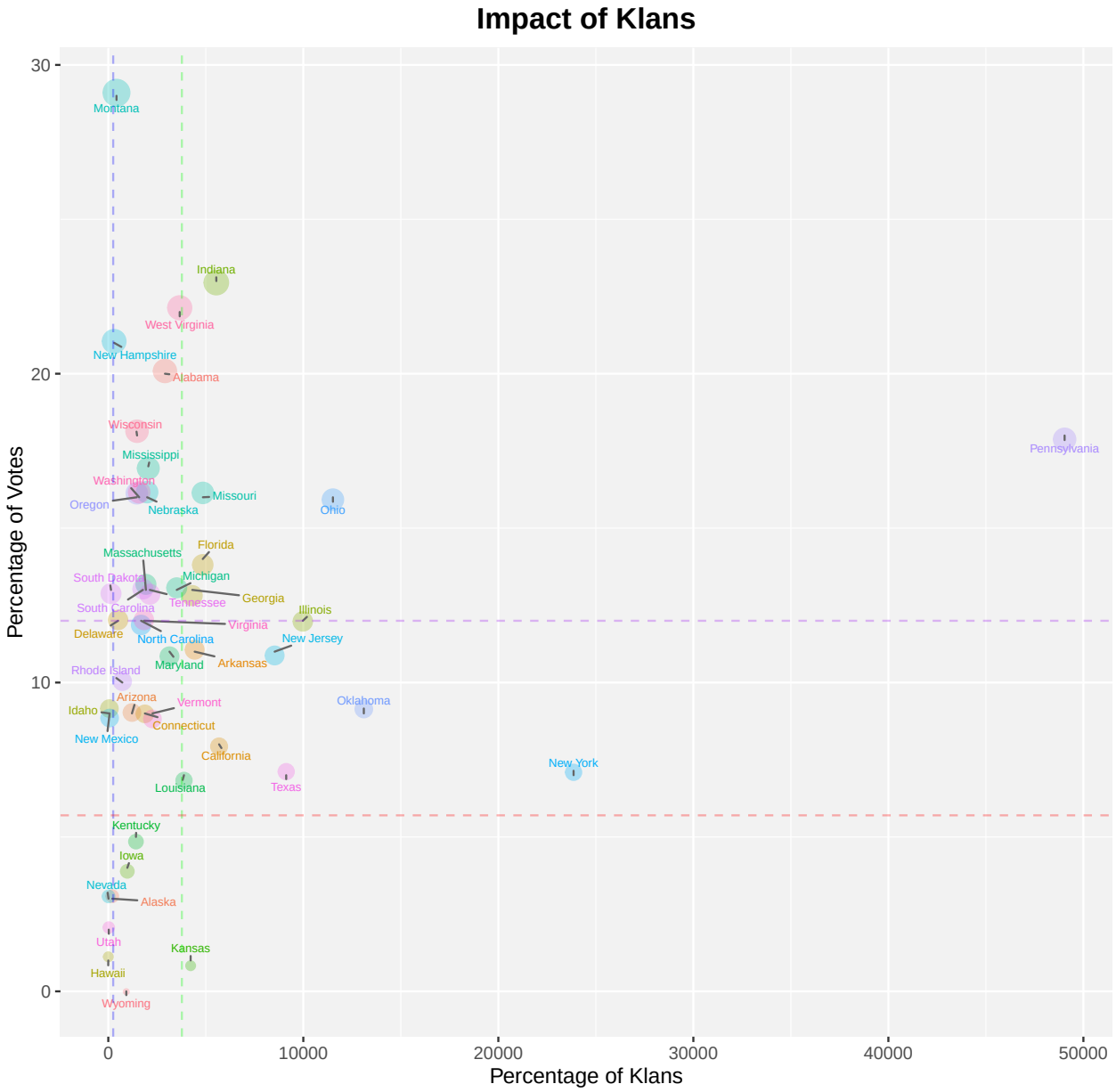
Ku Klux Klans

With its long history of violence, the Klan is the most infamous and oldest of American hate groups. When the Klan was formed in 1865, it was single, unitary organization. Today, there are dozens of competing Klan groups. Although black Americans have typically been the Klan's primary target, it has also attacked Jews, immigrants, homosexuals, and Catholics.

The data we used was from the 1920's and 30's and gave us information about the number of Klans in each state. They were separated county-wise which we aggregated state-wise and computed the percentages from them for each state.



Regression tree for Votes vs Klan Percentage, n denotes the number of states.



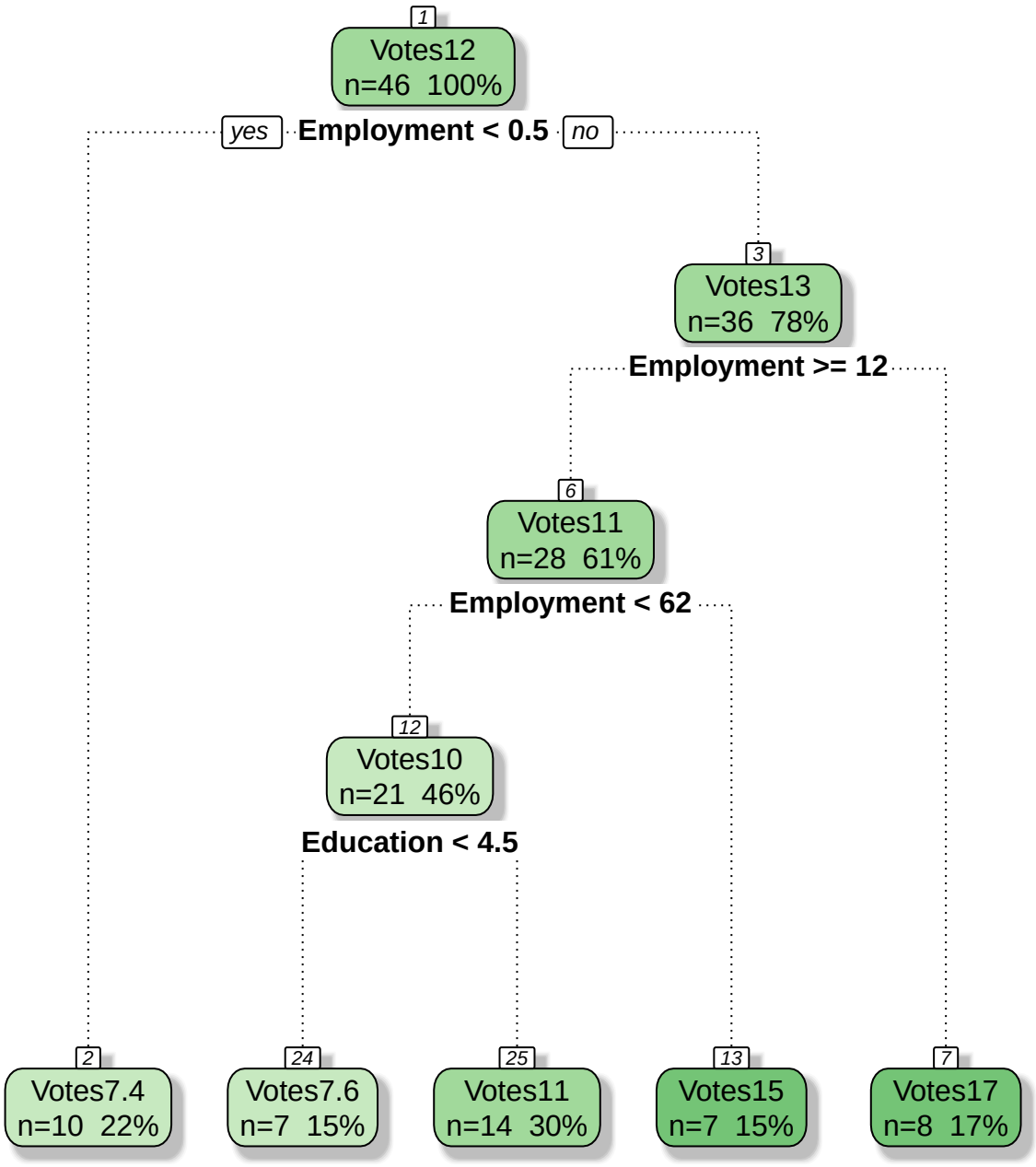
The plot shows the percentage of Klans in each state and how it affected voters of the present. The average percentage of votes garnered by Mr. Trump in the primaries is 12 % of the Republican voters. The mean number of Klan percentage is 4459.
Clearly we see from the tree plot that nearly 31 states, with a large concentration of Klans, account for above average votes for Trump and out of these 31 states, 15 of them have a Klan percentage above the average Klan percentage.

The division in the scatter plot is fixed by the splits in the regression tree for the same. We have used the extreme splits of less than 252 % Klans and more than 3772 % Klans and their corresponding voting percentages as the divison of the scatter plot.

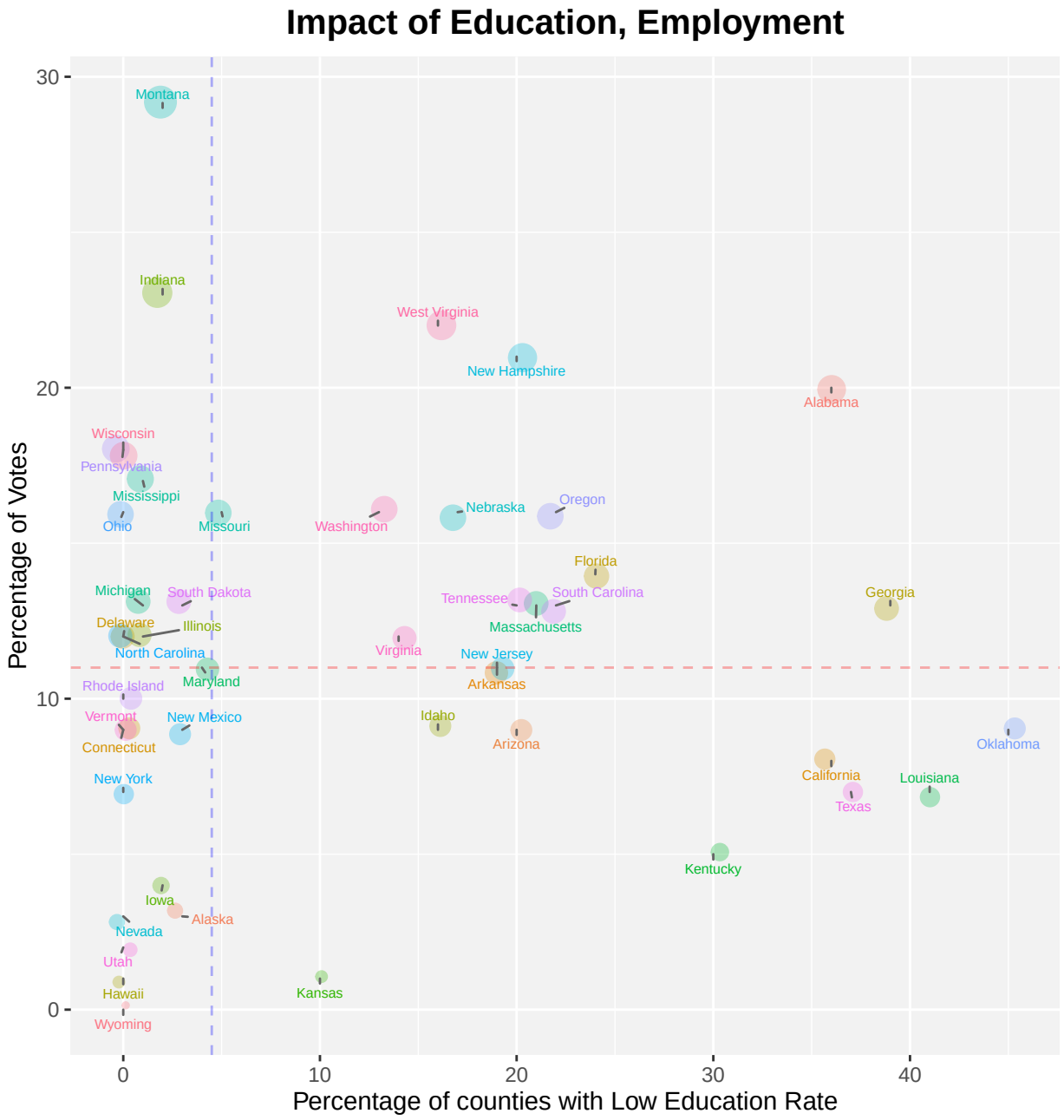
The data used was from the early part of the century and so is not as reliable. However it still gives us an indication of the impact of Klan population on voter sentiment. Hate groups are still existent in the United States with the KKK being just one example. For more info : https://www.splcenter.org/fighting-hate/intelligence-report/2016/active-hate-groups-united-states-2015.

Impact of Education, Poverty and Low Employment on the Votes

Ignorance due to low education and frustration of losing out jobs to skilled immigrants and the poverty that follows can be crucial factors for the empowered voter who can make a difference by electing a favorable representative. Thus we expect these three interdependent factors to have a significant impact on Trump's candidacy as can be seen in the tree plot below.



Regression tree for Votes vs Employment/Education/Poverty.



Low employment above 12 % accounts for 28 states with an average of 11 % of the votes in favor of Trump, making it dangerously close to his national average of 12 %. Out of these 28 states 21 of them have a low employment of more than 62 %. This number is twice the national mean for low employment rate of 31 %.

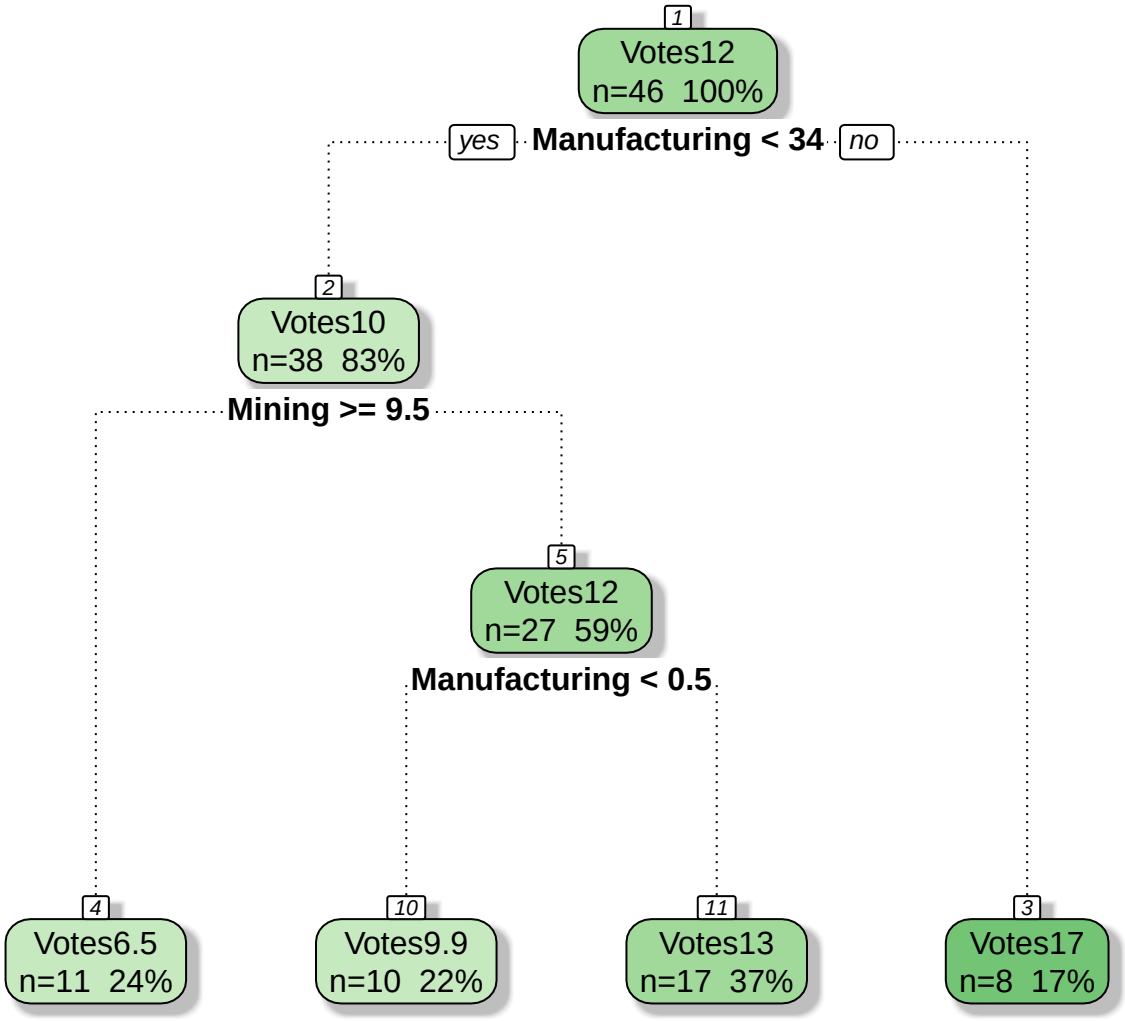
The split among these 21 states get determined by low education, where 7 states which have a high education rate vote in low numbers for Trump, an average of 7.6 %. However 14 states with a low education rate of more than 4.5 % are again precariously close to Trump's national average of 12 %. This number is well below the national mean of low education 12 %.

Furthermore the scatter plot shows the division with regards to low education at 4.5 % and the corresponding voting % of 11.

The impact of both these factors on Trump's votes are in line with our expectations of him being popular among areas with poorly educated people and low rate of employment.

Dependence on the Type of Economy

States with very little farming support the candidacy more but overall this is not a significant factor as farming as an occupation is much less compared to other vocations. In the same manner, Recreation, Government jobs and other nonspecialized counties do not have a significant impact on Trump's votes. This trend can be seen in the tree below.



Regression tree for Votes vs Type of Economy.

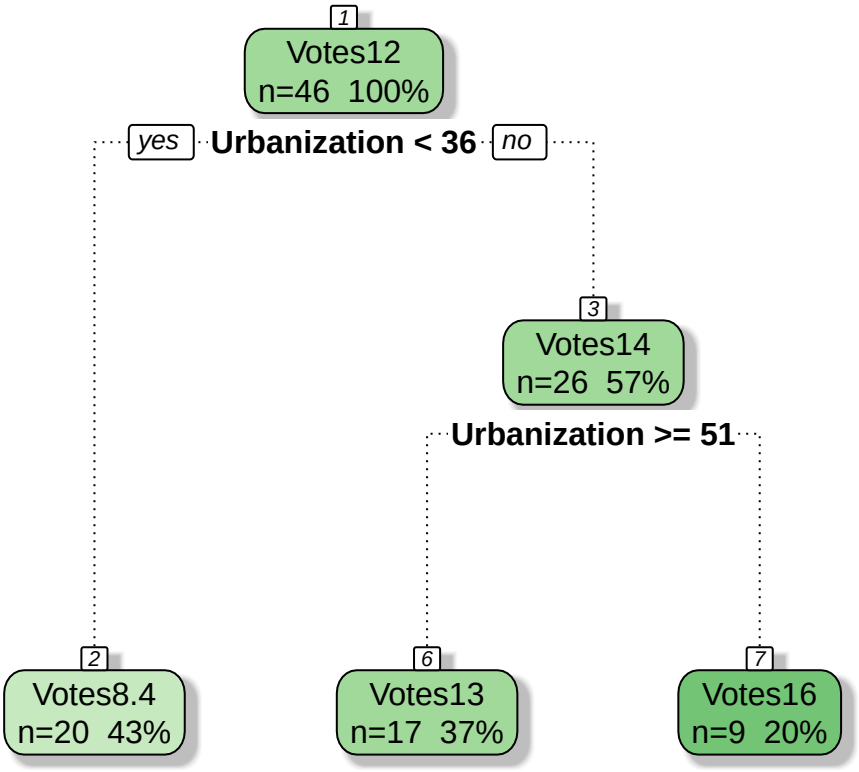
The less industrialized states tend to vote for Trump as they make up 83 %, or 38 states, of the favorable vote. These states have less than 34 % of the counties having manufacturing as a major source of revenue which is much more than the national mean of 22 %. On the other hand highly industrialized states with more than 34 % of the counties having manufacturing as a major profession also voted in large numbers with an average of 17 %, more than Trump's national average of 12 %.

Out of these 38 states 27 of them are Mining states, with more than 9.5 % of the counties engaged in the activity. This is comparable with the national mean of 11 %. They have voted with average of 12 % which mimicks Trump's national trend.

Thus states with a lot of mining and highly industrialized states tend to vote for Trump more.

Urbanized Voters

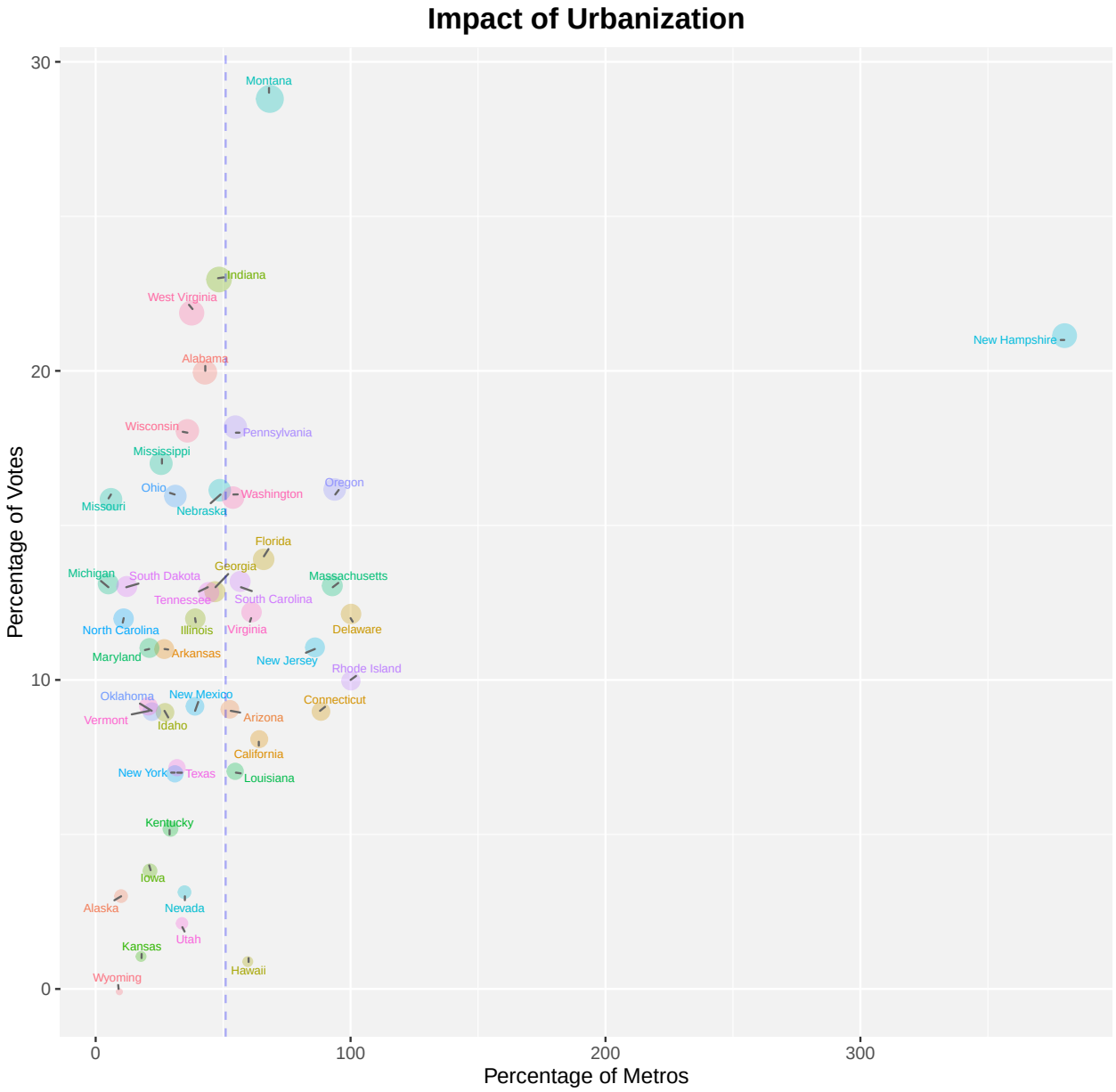
A measure of urbanization of a state is obtained by counting the number of metros in a given state. The regression tree for this dependence is shown below.



Regression tree for Votes vs Urbanization Percentage.

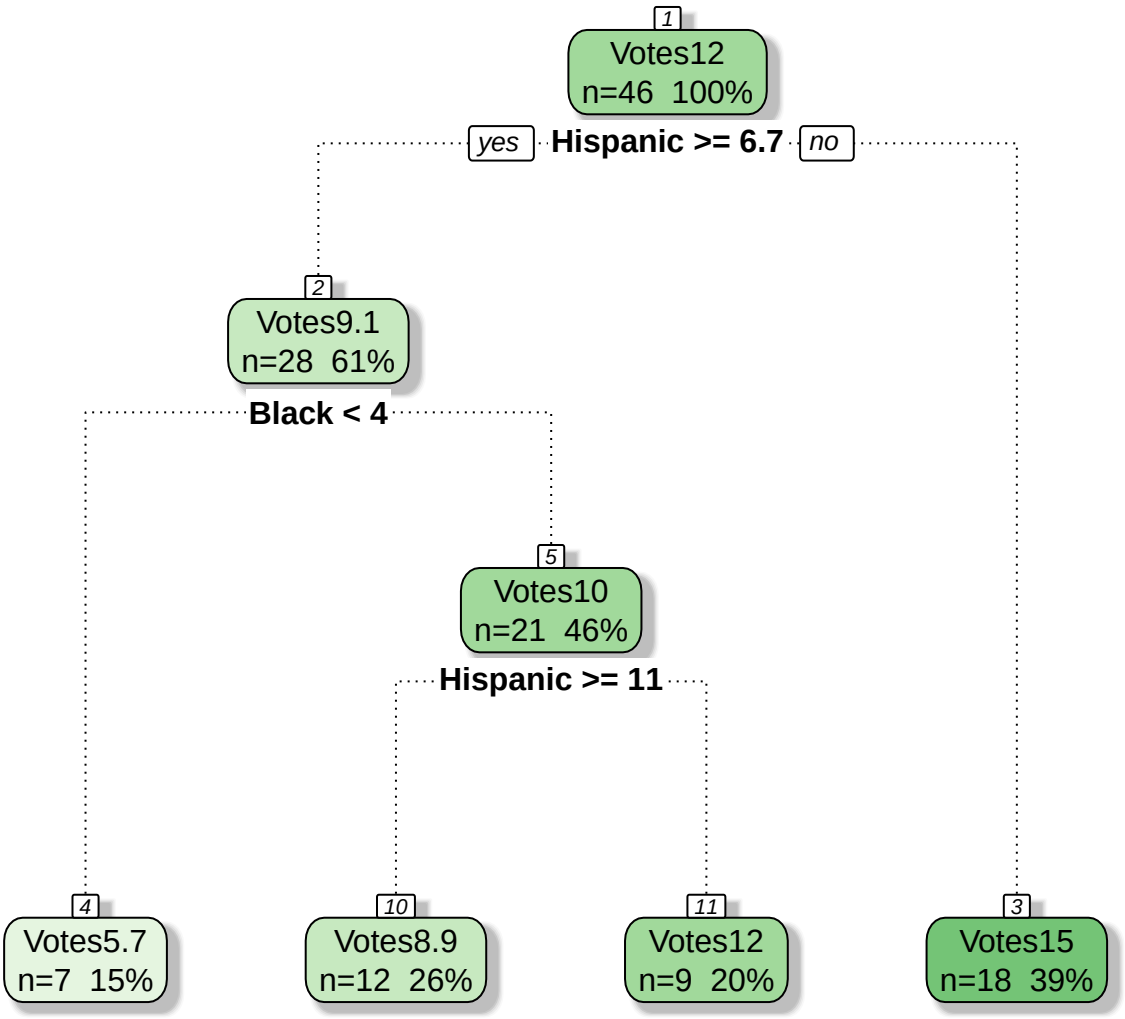
The mean percentage of urbanization for the United States is 51 %. We find that 4 % or 20 states have a high concentration of rural areas. These have voted less in favor or Trump showing that the farming community is clearly against his views. Highly urban regions of more than 51 % metros have voted in large numbers for Trump with an average of 13 % votes, above his national average of 12 %. Semi urban regions, between 36 and 51 % have also voted in his favor.

The scatter plot below shows how urbanization affected Trump's votes with the plot divided by the national mean of 51 %.



Preferences of different Racial Groups

With Trump's scathing remarks on the minorities it is imperative to check the inclination of the different races. As expected the Blacks and Hispanics tend to skew the voting percentages for Mr. Trump as seen in the regression tree below.

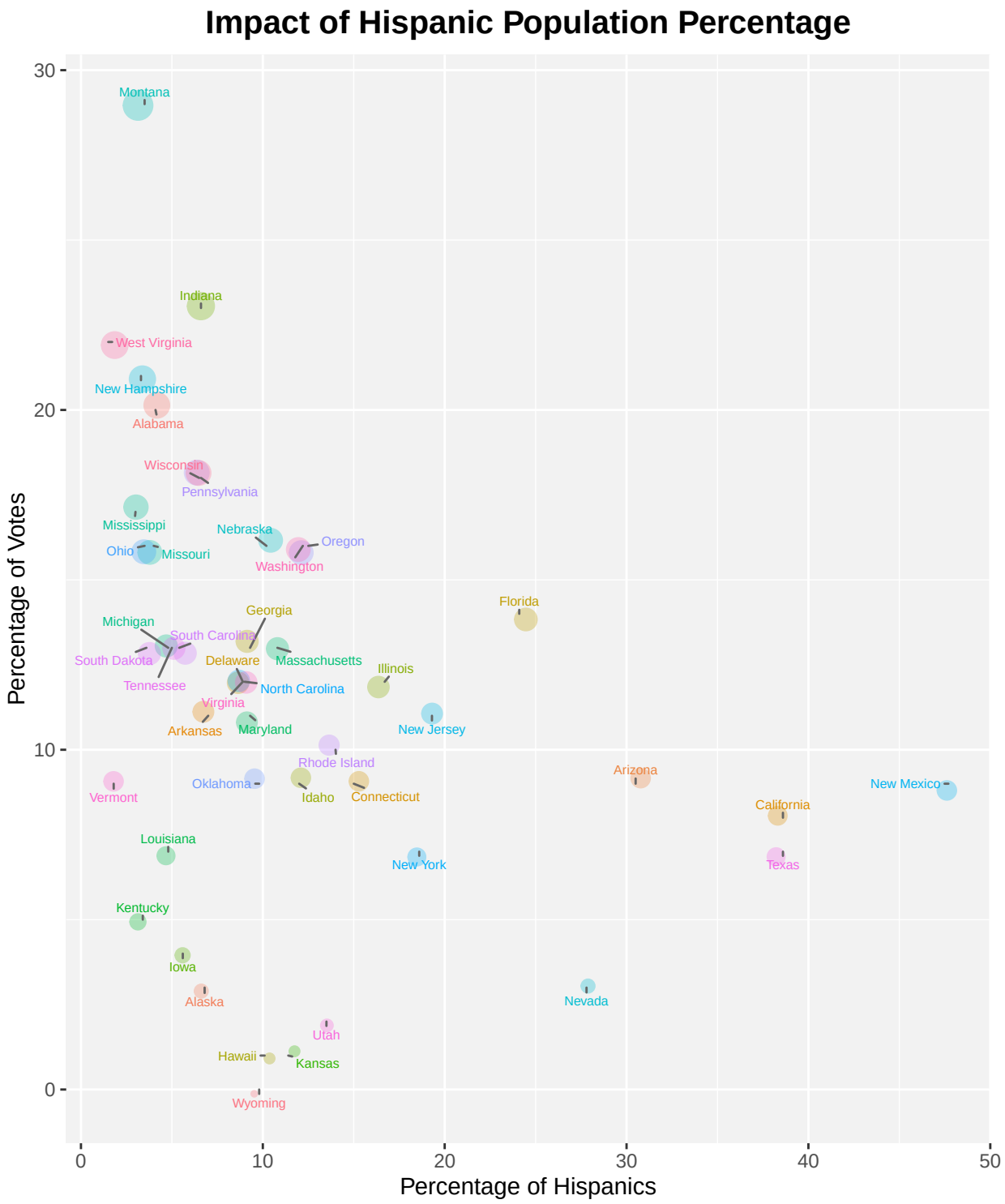


Regression tree for Votes vs Race.

States with less than 6.5 % Hispanic population voted overwhelmingly in favor of Trump, with an average of 15 % votes, well above Trump's national average of 12 %. This trend is seen in 18 of the 46 states.

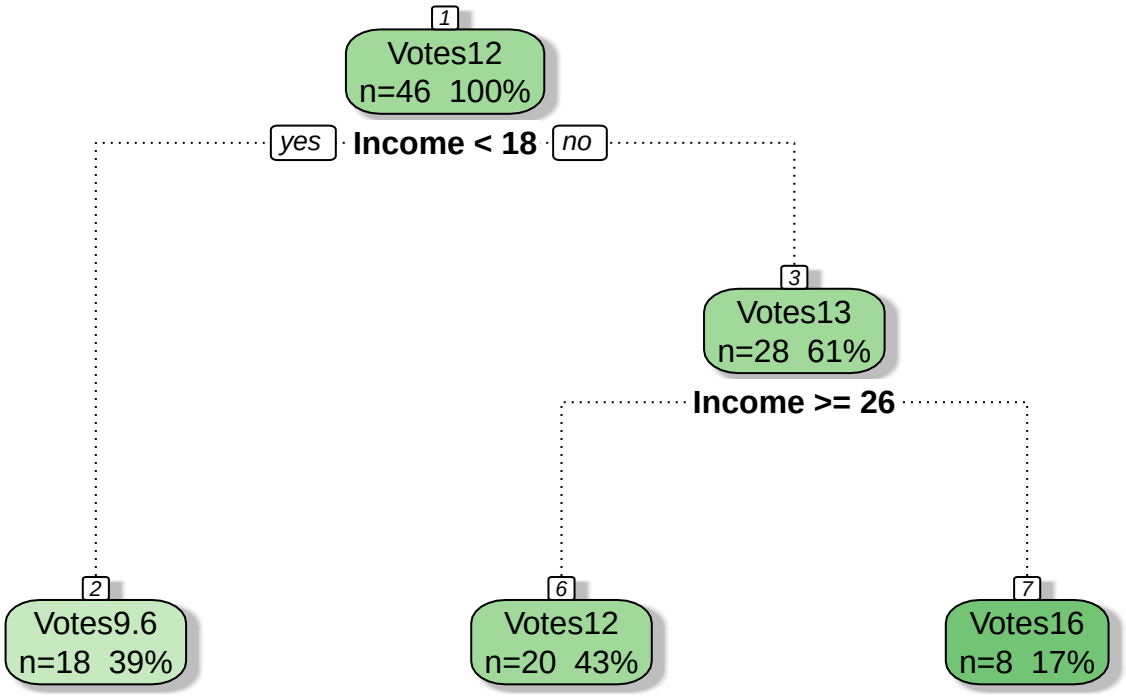
The remaining 28 states are those where the Hispanic population and Black poulation determine Trump's fate. The national means for Hispanics is 12 % and for Blacks is 12 %. States with a low Black population of less than 4 % have voted in very low numbers for Trump with an average of 5.7 %. States with both a high Black and hispanic population have voted with just about the average votes for Trump of 12 %.

In general states with higher Black and Hispanic population haveThis is fu voted less favorably for Trump. This is further confirmed by the scatter plot showing the voting trend of the Hispanics.



Income Dependence

The lower income groups with mean household income of less than 18,000 $ per annum clearly do not favor Trump when compared to the higher income groups as can be seen in the regression tree below. These include 18 of the 46 states.



Regression tree for Votes vs Mean Household Income.

The national mean for the mean household income is 24,000 $. The upper middle class and the rich with an annual mean household income of more than 26,000 $ are clearly in favor of the capitalist Trump. This accounts for 20 among the 46 states. The remaining 8 states have an average household income between 18,000 $ and 26,000 $ comprising the middle class and they have voted in large numbers for Trump.

Inferences

Having analysed these factors affecting the voting pattern for Trump, let us put together our inferences based on this non-exhaustive list of factors.

Our data is based on the primaries held in 46 states and so the voting percentage comprises of only Republican voters from each state. As can be seen from the root of each of the tree plots, an average of 12 % of the Republican voters from each state opted for Trump. Using this as a baseline let us summarize how each factor has built this average.

States with a siginficant Hispanic and Black population have voted much below the average making Trump's supporters predominantly White. This is consistent with the fact that these are the three biggest racial groups in the US and Trump's rhetorics against non-White population.

Mean household income data show that the upper middle class and the rich Republicans are the ones who endorse Trump while the lower income groups have voted in much lower numbers.

Regions where more than half the state is urbanized have also gone in favor of Trump when compared to more rural states that have been clearly repelled by Trump's agenda.

States with a heavy presence of manufacturing and mining have preferred Trump, with the non-industrialized states voting in much lower numbers for a production friendly Trump. This is especially seen with states having farming as its economic backbone.

Unemployment seems to be a huge factor in deciding votes for Trump, in line with his anti-immigrant agenda. States with a high percentage of unemployment seem to favor Trump. As expected states with low education

percentage, that is those states where 20 % or more of the voting population do not have a high school diploma or equivalent, have supported Trump.

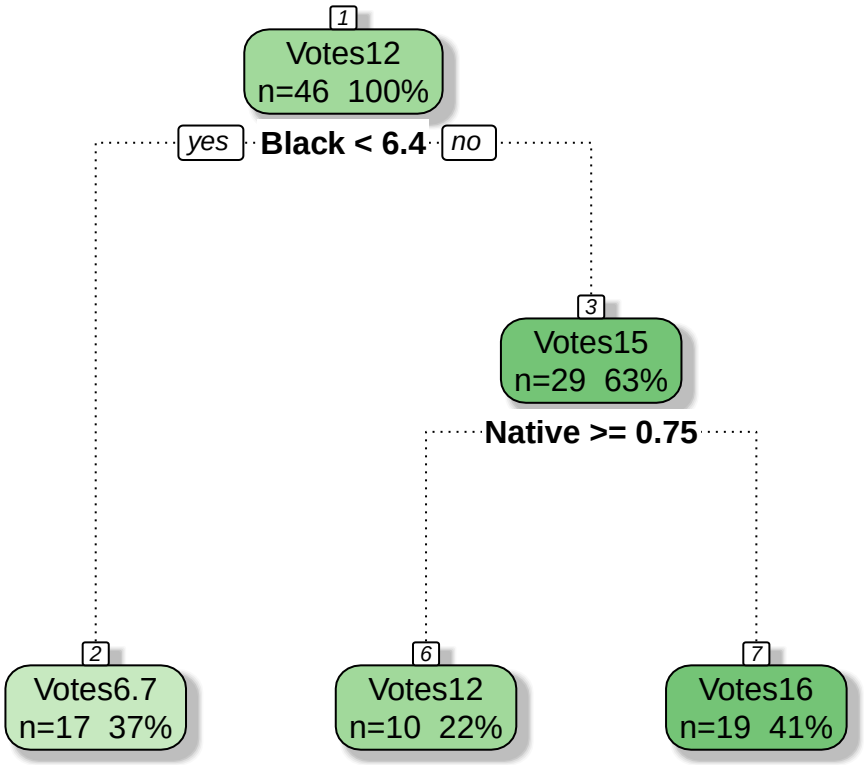
Finally we saw the impact of hate groups such as the Ku Klux Klan in deciding the votes. Though this data is from the past presence of the KKK in the various states, it still seemed to have an effect on the voters with states which had a large presence of the KKK voting above the average number of votes obtained by Mr. Trump from all the states.

Thus far the inferences obtained seems to be in line with the ideologies of Donald Trump. It would also be interesting to study the effect of the religion of the voters in their choice of Trump given his anti-semitic and anti-Muslim speeches. We were unable to do this due to lack of sufficient data.

Future Directions

We could perform a similar analysis for the Democrats Hillary Clinton and Bernie Sanders and study the characteristics of their voters. The data we collected allows us to make these filters and complete the analysis. A comparative study could lead to an estimate of the number of voters who could change sides and ultimately lead to a prediction of the presidential election later this year. However this is much a much more ambitious goal than our exploratory analysis done up to this point.

Nevertheless to give a flavor of Hillary Clinton's voters let us look at the racial characteristic of her voters and compare them to that of Trump's voters.

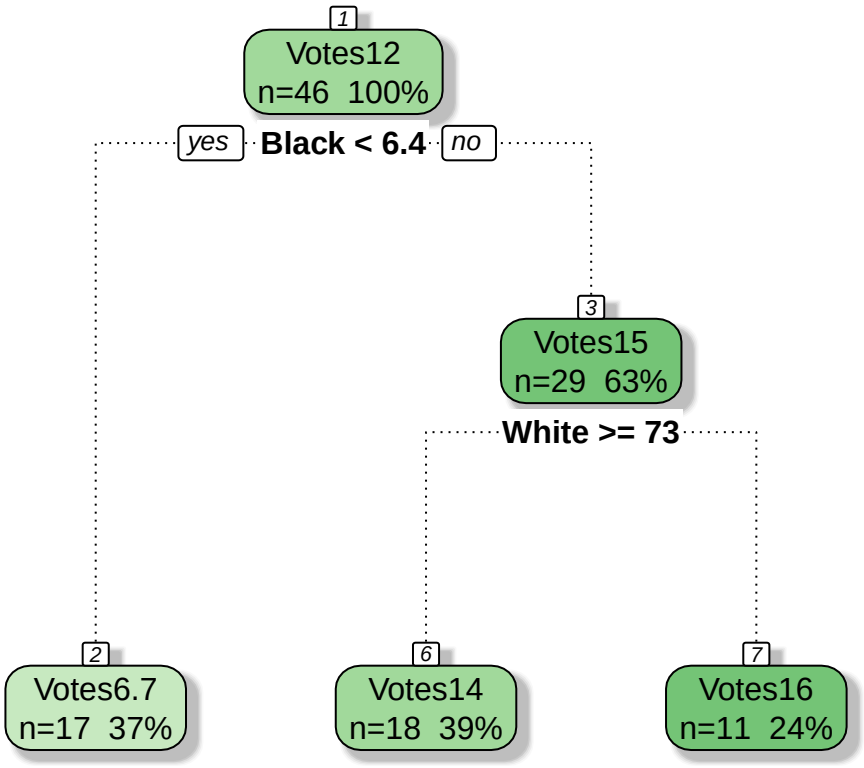


Regression tree for Votes vs Race for Hillary Clinton.

Hillary Clinton obtanied a national mean of 12 % of the Democrat voters, the same as Donald Trump ! We should keep in mind that the number of registered voters pledged to the Democrats and the Republicans are

not the same. States with a low Black population have voted in fewer numbers for her while those with a Black population of higher than 6.7 % have voted in larger numbers. Her voting trends seem to affected more by minority groups as well.

As the Hispanic and Black population determined the voting trend for Trump let us make another regression tree for Hillary Clinton with just the Whites, Blacks and the Hispanics, the three biggest racial groups in the United States to make a better comparison between the two candidates.



Regression tree for Votes vs Major Races for Hillary Clinton.

This shows that states with high percentage of Black and White population contributed in a big way to Clinton's voting numbers. Clearly Hillary has an healthy advantge among the Black population when compared to Trump.

We can perform similar comparisons for the other characteristics of the two voters which will definitely help the campaign managers of the two candidates to study whose votes they should usurp to obtain the majority. The candidates can appropriately change their policies to achieve this.

Acknowledgements

I thank my mentor Srdjan Santic for helping me out throughout the workshop and the Springboard team for their resources and guidance.