PSTAT 131 Final Project

Mateo Vasquez and Elijah Castro

Background

In 2012, the outcome of the U.S. presidential election did not come as a surprise to many. Some statisticians correctly predicted the outcome of the election, but none as precisely as Nate Silver. Yet, despite the success in 2012, the 2016 presidential election came as a enormous surprise, and it underscored that predicting voter behavior is complicated for many reasons despite the tremendous effort in collecting, analyzing, and understanding many available datasets. Our goal, therefore, is to merge census data with 2016 voting data to analyze this election outcome.

To preface, it is important to first comment on the difficulties of voter behavior prediction (and thus election forecasting). Voter behavior prediction is especially challenging because although polling error is an inevitability that statisticians always seek to mitigate, the accuracy of their forecasts are ultimately determined by the the quality of the polling data used. Statisticians can only use information about how people "think" they will vote on election day, and the temporality of people's responses gathered from polling needs to be taken into consideration since changes in a voter's allegiance can and will happen. The quality of polls are also limited to only take their respondents' answers on a surface level, which itself brings a plethora of potential issues that could detract from their validity. For example, respondents could either provide a non-response or lie about their chosen candidates out of embarrassment. Also, the polling process itself can introduce plenty of (un-)intentional bias to the data caused by a variety of factors, such as the use of a not-so-popular surveying method (e.g. phone surveys) which makes the random samples non-random or pollsters themselves being biased towards certain candidates.

Despite the challenges of accurately predicting voter behavior, in 2012, statistician Nate Silver was perfect in his predictions of the winners for every one of the 50 states during the presidential election. This is attributed to his unique approach of observing the full range of probabilities based on an agglomeration of collected polls rather than maximizing specific probabilities of certain variables like his contemporaries. Applying Bayes' Theorem, Silver would incorporate the full range of probabilities by calculating new probabilities of each level of percentage support for Obama in each state and use the polling data to determine how much of those were above 50% (i.e., the probability that Obama wins each state if the election were called on that day). Silver's model would be simulated forward in time to the election day for both the state and national level of support, and then, it would weigh each forwarded simulation by the probability that the starting point is correct in order to predict the probability that Obama would win the election. The 2012 presidential elections also featured a huge influx of polling data, especially nearing its end, that could have been fed into Silver's model and lent it more confident estimates (while also accounting for potential biases by fitting the previous 2008 presidential election's data into the model and determining how much its approximated support deviated from the actual results).

Unfortunately, in 2016, Nate Silver and many other statisticians failed in predicting Trump's victory. This was rooted in certainty bias fostered by both the polling data and their subsequent analyses. As aforementioned before, every poll will inevitably have error and statisticians always aim to mitigate any of it, but the quality of their analyses will ultimately be determined by the data gathered. In this instance, the polls conducted at every state were vulnerable to systematic errors such as underestimating the proportion of voters who were unemployed whites, or not accurately measuring the difference in enthusiasm between supporters of Trump relative to another candidate like Clinton. The unevenness of these nationwide errors further compounded to their inaccuracy as underestimates of certain Trump-supporting demographics, such as white men and women

without degrees, made it harder for statisticians to decipher an interpretable pattern and thus, de-emphasized their significance in their analyses. As a result of polls being skewed towards Clinton due to these systematic errors, forecasts based on them would start missing in the same direction and cause a snowball effect to occur. Subsequent polls would then lean further towards Clinton while their analyses by statisticians would follow suit, fostering more certainty bias by both pollsters and analysts. To improve future predictions, we believe more lengths should be taken to identify patterns (or lack thereof) in the omission or insignificance of certain variables in order to interpret any possible ways a statistician's model could go wrong. Perhaps when many different statisticians collaborate and identify a trend on what variables are considered unimportant in their models, they could bring their insights to the pollsters' attention and have them focus more data gathering on those underestimated variables to improve future forecastings.

Election Data

To begin analyzing the outcome of the 2016 presidential election, we first look at the election data.

Some example rows of the election data are shown below:

county	fips	candidate	state	votes
Los Angeles County	6037	Hillary Clinton	CA	2464364
Los Angeles County	6037	Donald Trump	CA	769743
Los Angeles County	6037	Gary Johnson	CA	88968
Los Angeles County	6037	Jill Stein	CA	76465
Los Angeles County	6037	Gloria La Riva	CA	21993
Cook County	17031	Hillary Clinton	IL	1611946

If we inspect the rows with `fips = 2000`, we obtain the following results.

county	fips	candidate	state	votes
NA	2000	Donald Trump	AK	163387
NA	2000	Hillary Clinton	AK	116454
NA	2000	Gary Johnson	AK	18725
NA	2000	Jill Stein	AK	5735
NA	2000	Darrell Castle	AK	3866
NA	2000	Rocky De La Fuente	AK	1240

We notice that the `county` column are all missing values, which implies that the votes tallied are either statewide or nationwide. These are clearly statewide votes because if we also observe when `fips = 'AK'`:

county	fips	candidate	state	votes
NA	AK	Donald Trump	AK	163387
NA	AK	Hillary Clinton	AK	116454
NA	AK	Gary Johnson	AK	18725
NA	AK	Jill Stein	AK	5735
NA	AK	Darrell Castle	AK	3866
NA	AK	Rocky De La Fuente	AK	1240

We see that the votes are the exact same as before, which means that `fips=2000` is also representing the statewide votes for Alaska instead of a county. Therefore, we exclude its rows from our analysis to remove

any redundancies in our data set.

Our new data set after the removal of the `fips=2000` observations then gives us dimensions of 5 variables (columns) and 18,345 observations (rows).

Census Data

The first few rows and columns of the `census` data are also shown below.

CensusTract	State	County	TotalPop	Men	Women
1001020100	Alabama	Autauga	1948	940	1008
1001020200	Alabama	Autauga	2156	1059	1097
1001020300	Alabama	Autauga	2968	1364	1604
1001020400	Alabama	Autauga	4423	2172	2251
1001020500	Alabama	Autauga	10763	4922	5841
1001020600	Alabama	Autauga	3851	1787	2064

The variables shown above are:

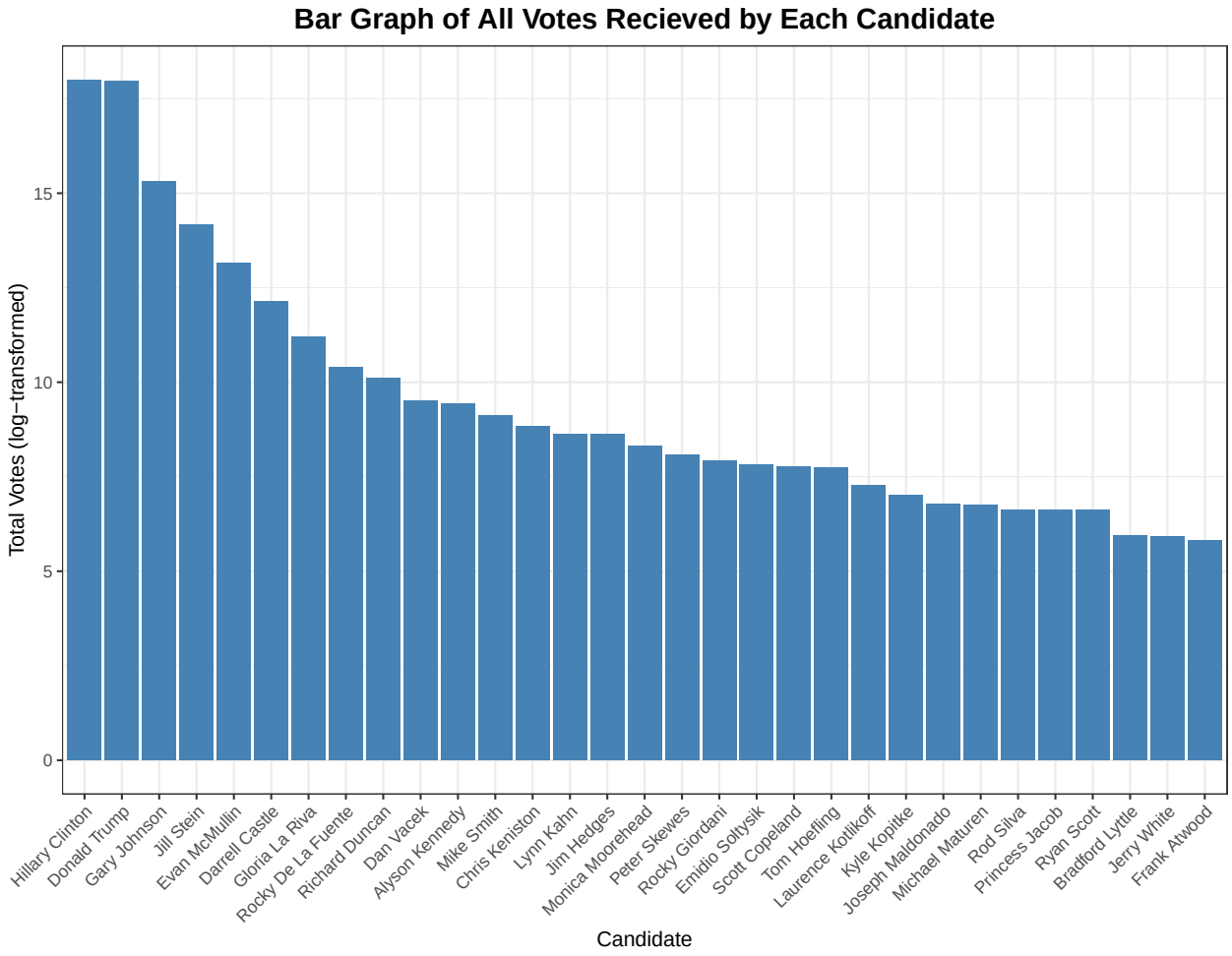
variable	description	type
CensusTract	Census tract ID	numeric
State	State, DC, or Puerto Rico	string
County	County or county equivalent	string
TotalPop	Total population	numeric
Men	Number of men	numeric
Women	Number of women	numeric

Data Preprocessing

To preprocess our data, we first separate the rows of our election data into three data frames: 1.) federal-level vote tallies 2.) state-level vote valles 3.) county-level vote tallies

We then draw a bar graph of all votes received by each candidate (candidate names ordered by decreasing vote counts).

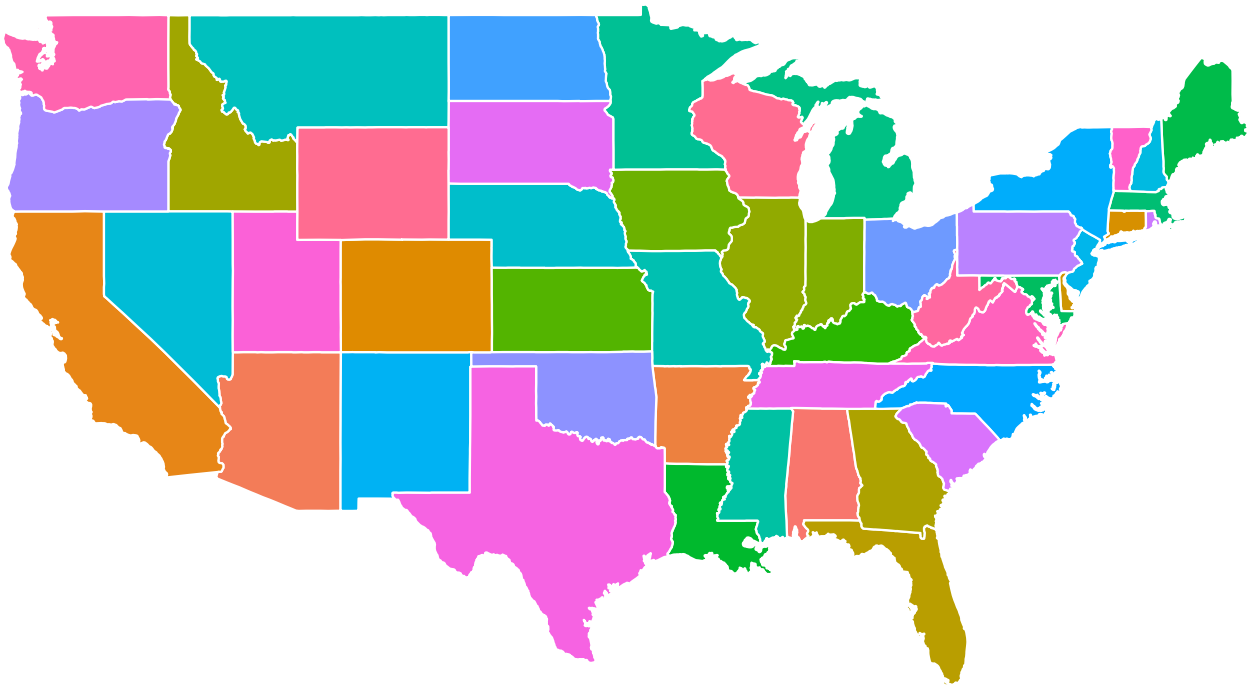
There were 31 named presidential candidates during the 2016 election.



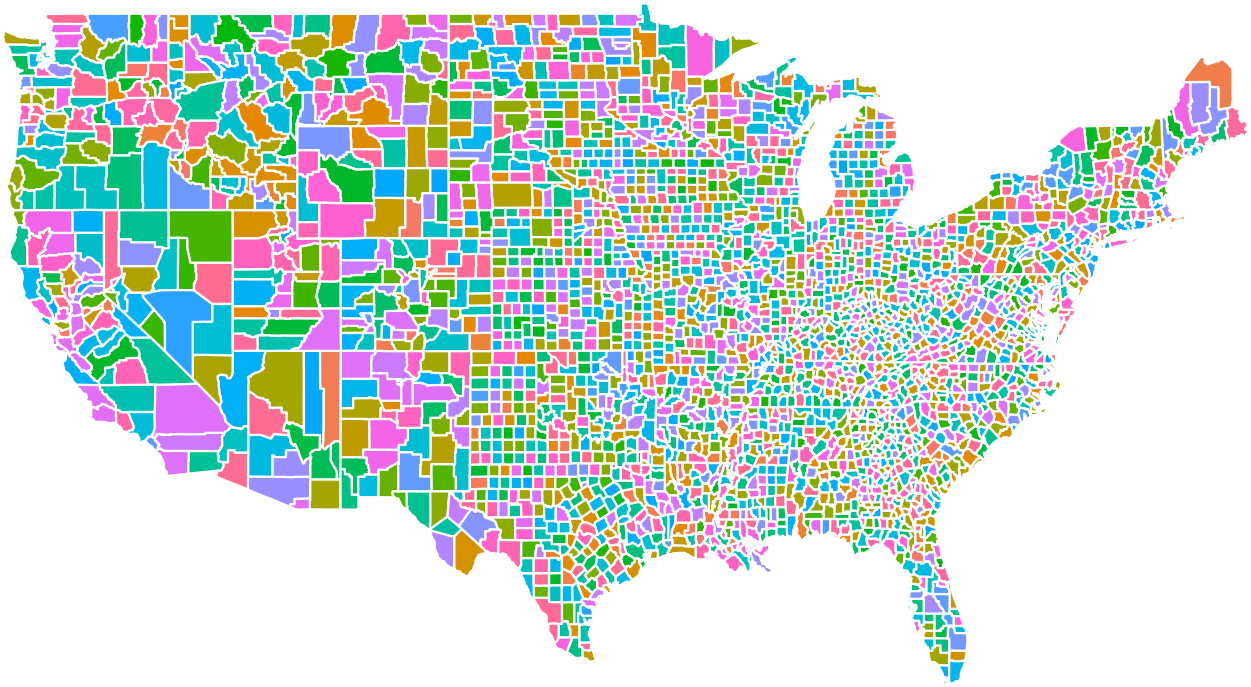
Next, we create two more data frames that show the winning candidate (with the highest proportion of votes) by county and state.

Visualization

Below, we can see a state-level map of the election data colored by state.

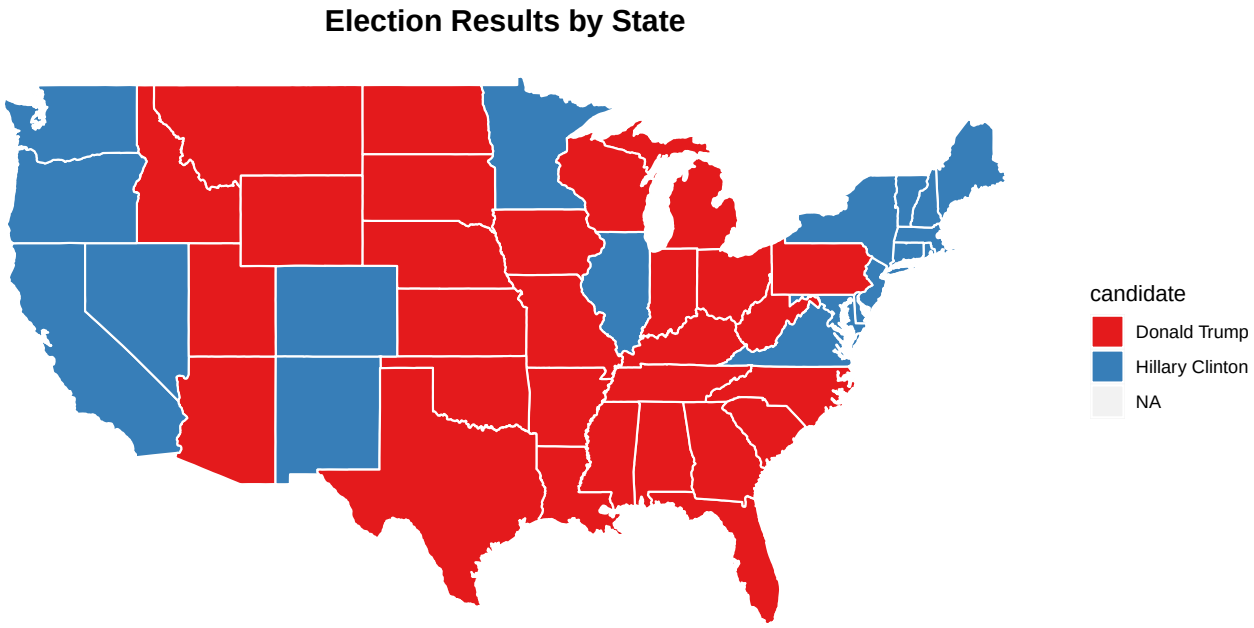


We can also draw a county-level map colored by county.

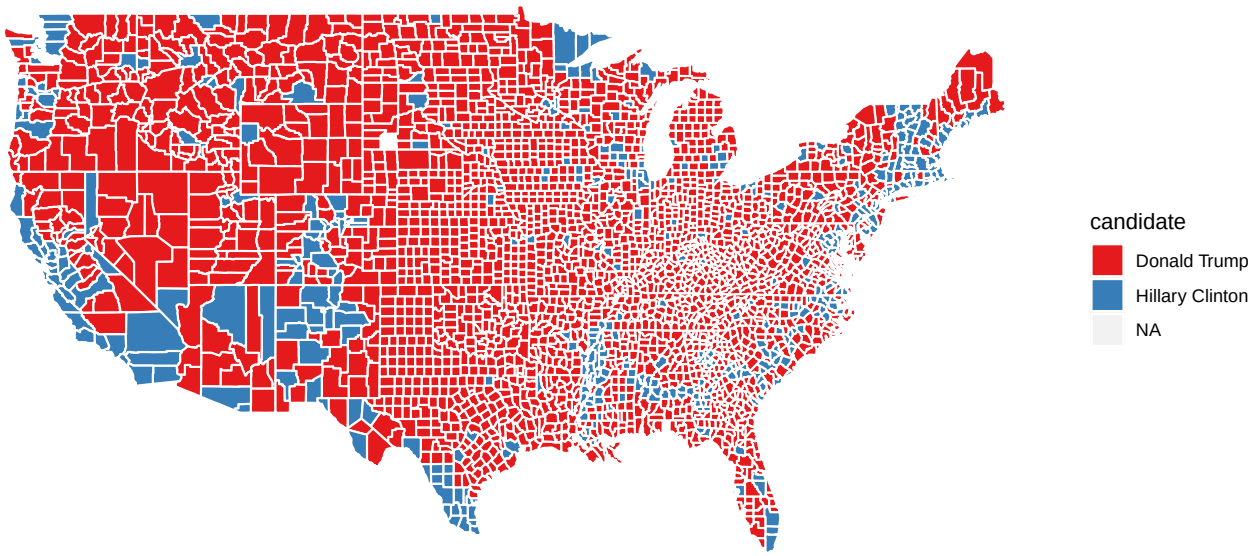


In order to create a map with the winning candidate for each state, we need to merge the map data with the data frame we previously created of the candidate winner by state.

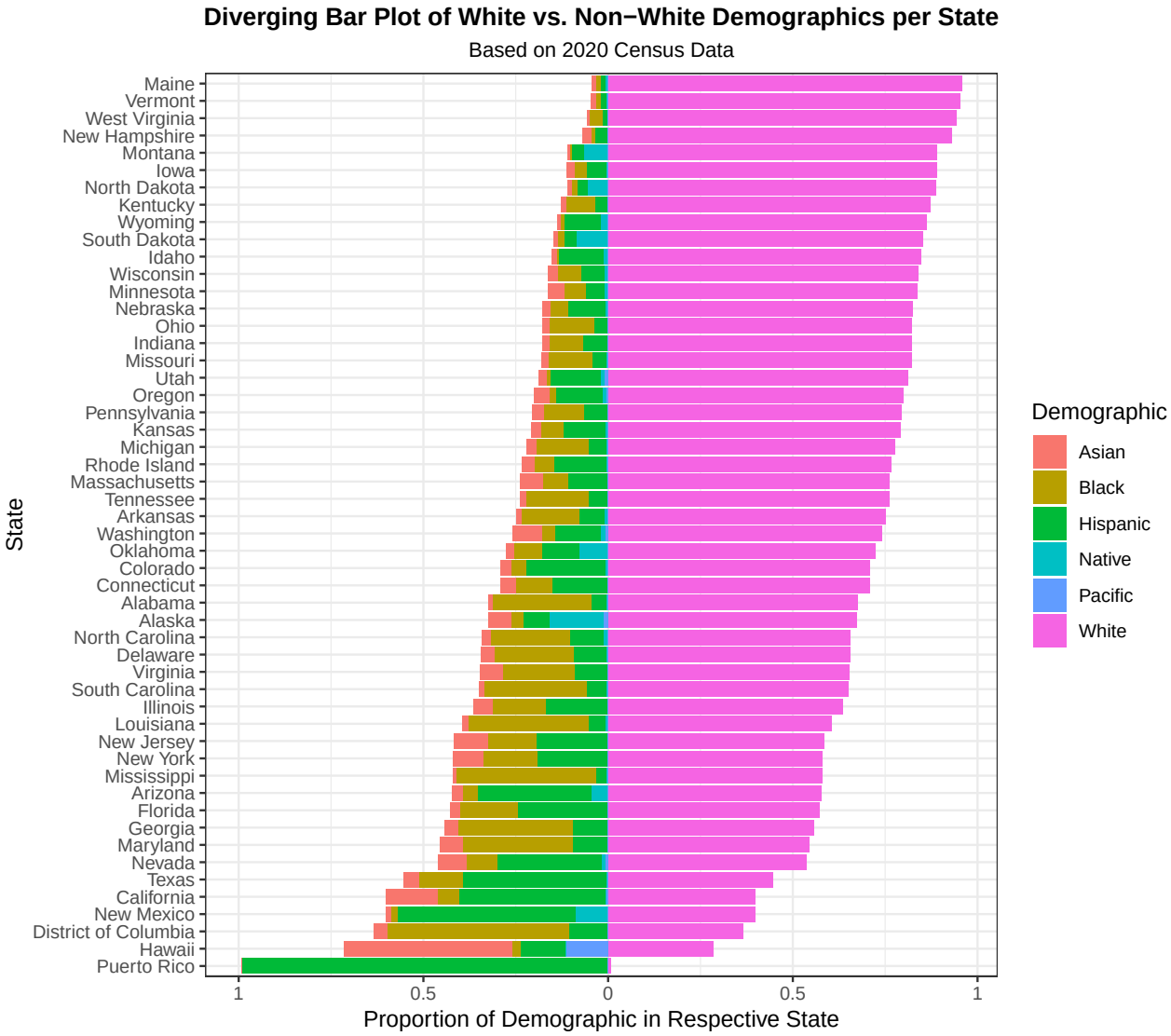
Once the data is merged, we can create a map of the election results by state:

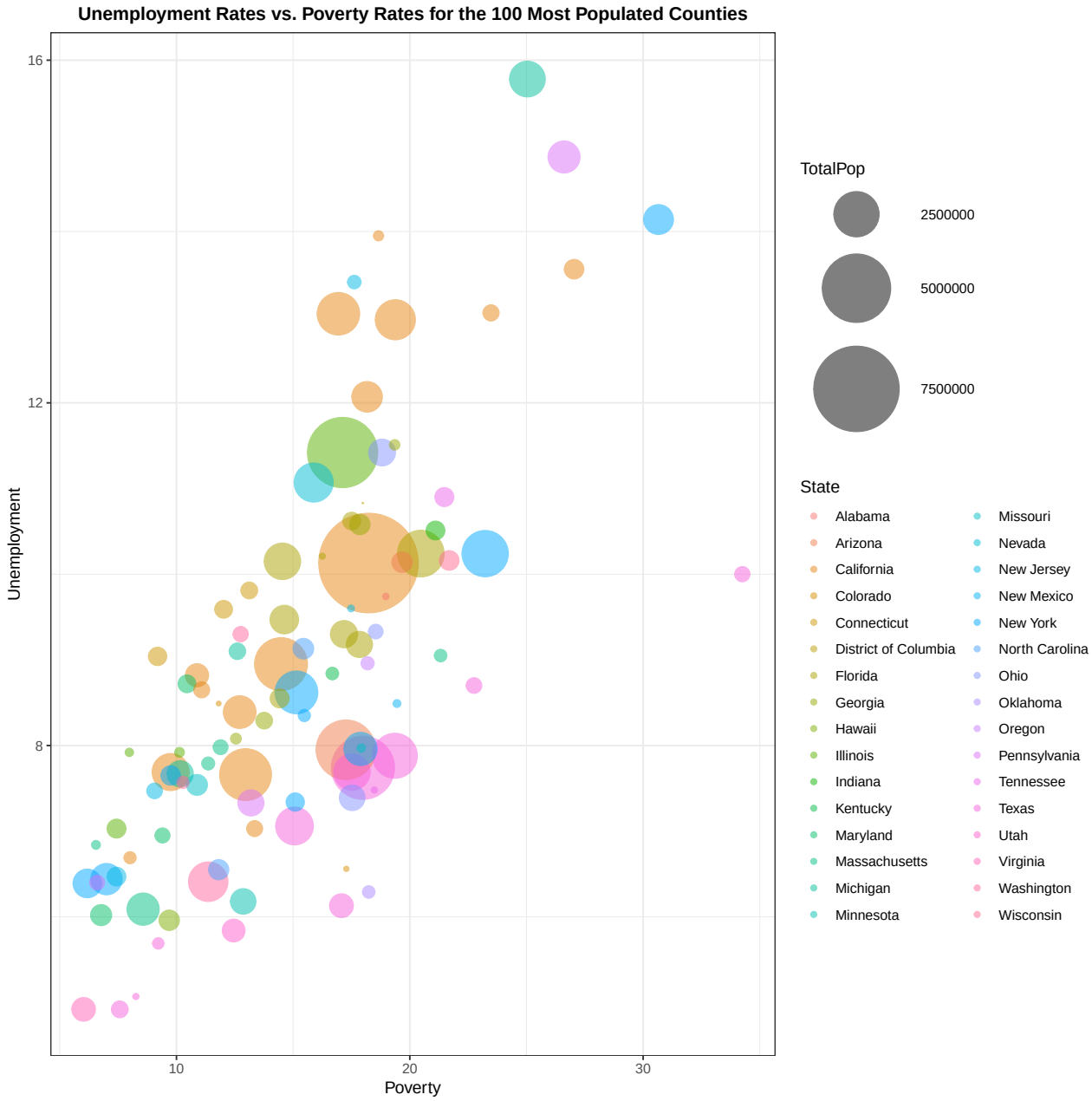


A similar map can be created of the election results by county:



Using the census data, we can also create different visualizations. For example, the first visualization is a diverging bar plot representing White vs Non-White populations for each state, and the second is a bubble plot representing Unemployment vs. Poverty for the top 100 most populated counties.





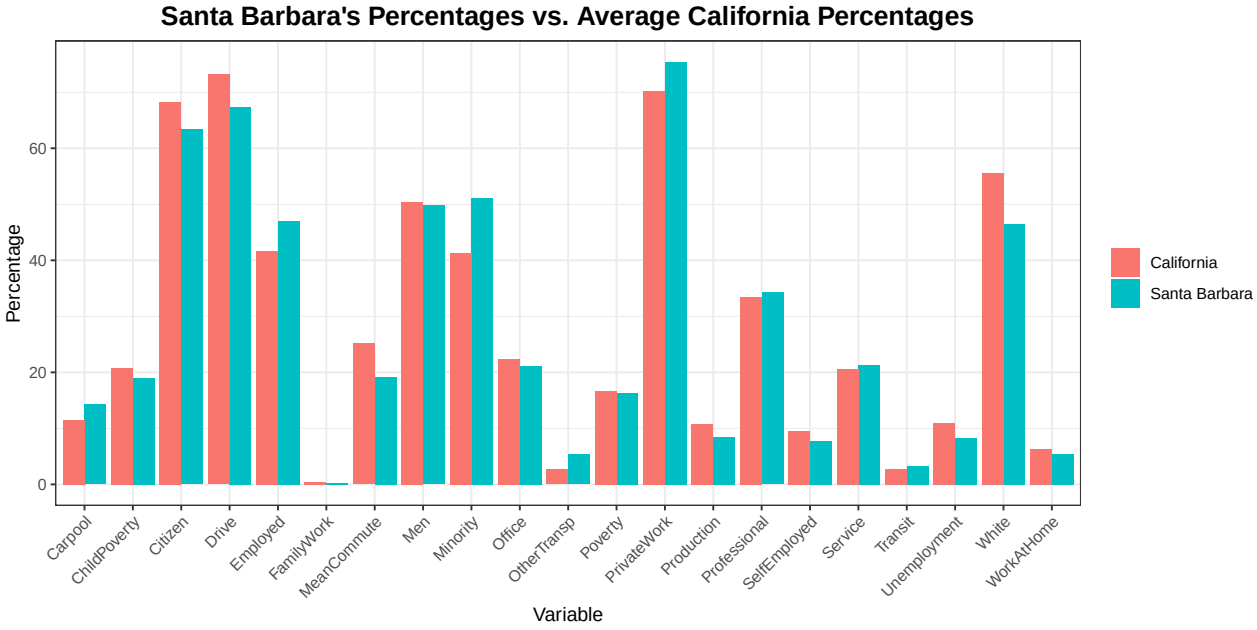
Next, since the census data is more fine-grained than needed for our use, we can aggregate the information into county-level data. Below is the first 6 rows and columns after aggregating:

State	County	TotalPop	Men	White	Citizen
Alabama	Autauga	55221	48.43	75.79	73.75
Alabama	Baldwin	195121	48.85	83.1	75.69
Alabama	Barbour	26932	53.83	46.23	76.91
Alabama	Bibb	22604	53.41	74.5	77.4
Alabama	Blount	57710	49.41	87.85	73.38
Alabama	Bullock	10678	53.01	22.2	75.45

To delve deeper into the relationship between census data and election data, we think as a preliminary, it

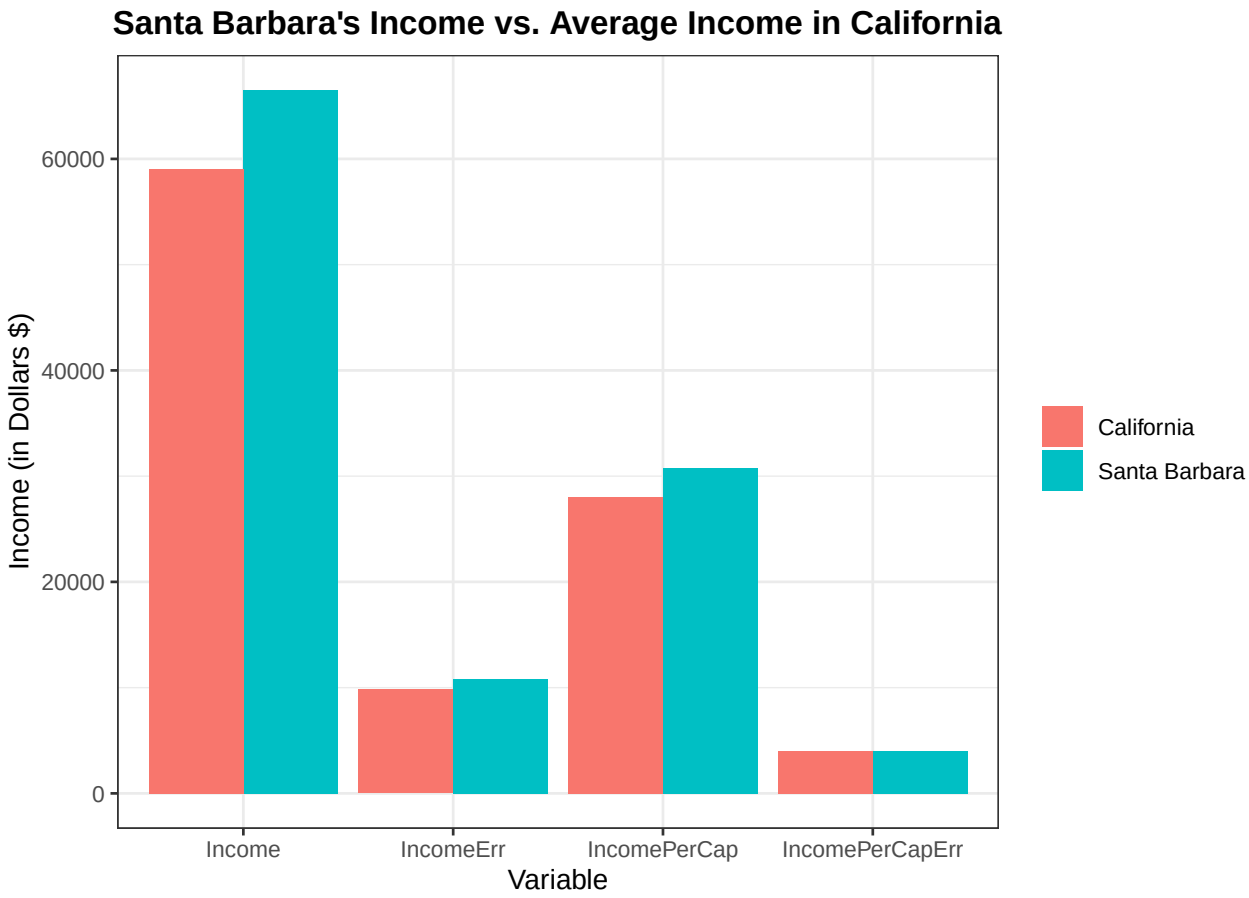
is important to compare and contrast the results and demographic information for the county we were in during the 2016 presidential election with the state it is located in.

During election day of the 2016 presidential election, we were in Santa Barbara County, California, where Hillary Clinton won the majority vote. We can visually compare and contrast results and demographic information from Santa Barbara County to California's averages using dodge bar-plots. We begin by observing any significant differences between demographic information.



Based on the bar-plot, there does not seem to be many discernible differences in percentages between Santa Barbara and California's averages. However, the minority and white percentages are fairly notable since the percentage difference appears larger than most of the bars. Santa Barbara features a higher percentage of minority groups in its population than California on average as well as a smaller white population. This difference in demographics is a bit surprising considering that Santa Barbara County has a higher cost of living than most counties in California; so, one would expect the difference in white and minority percentages to either be closer or skewed more in favor for the former. Furthermore, the percentage difference of the population who drives alone to work (`Drive`), who are employed (`Employed`), or employed in a private industry (`PrivateWork`) are worth noting as well, even if it is to a lesser degree.

We then observe any significant differences in terms of the income variables between Santa Barbara and California's averages:

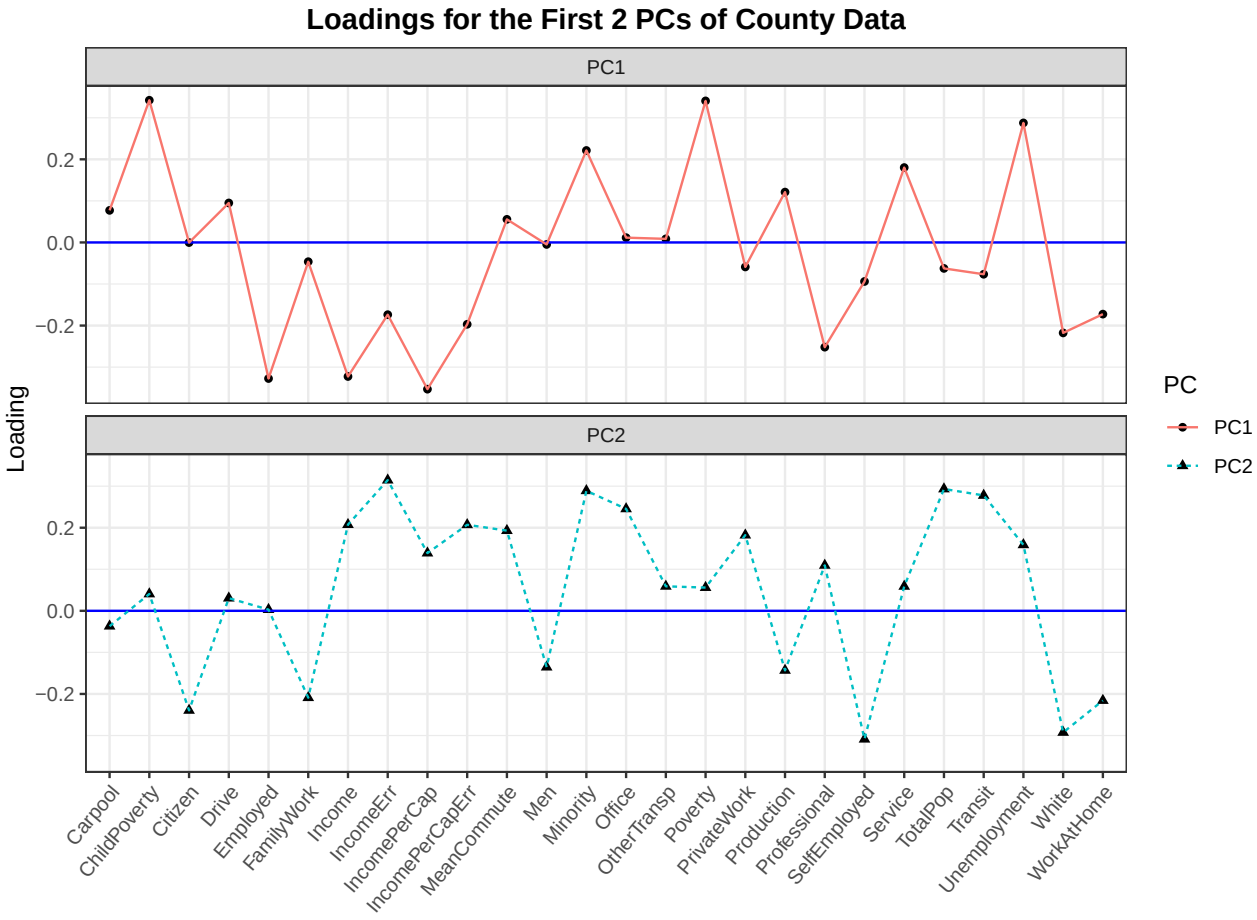


The most significant difference in this bar-plot is the median household income (`Income`), as Santa Barbara's is higher than the average in California. This aligns well with the prior statement that Santa Barbara County has a higher cost of living, which makes the previous difference in demographic percentages more surprising and worth further analysis.

Exploratory Analysis

We will perform Principal Component Analysis (PCA) on both county and sub-county level census data by computing the first two principal components (PC1 and PC2) for each and seeing if we can infer their latent structure. For our purposes, we will center and scale the features for both data sets because the income, population and percentage-related variables are not under the same scale, which is evident from the prior bar-graphs where we had to plot the income and percentage variables separately.

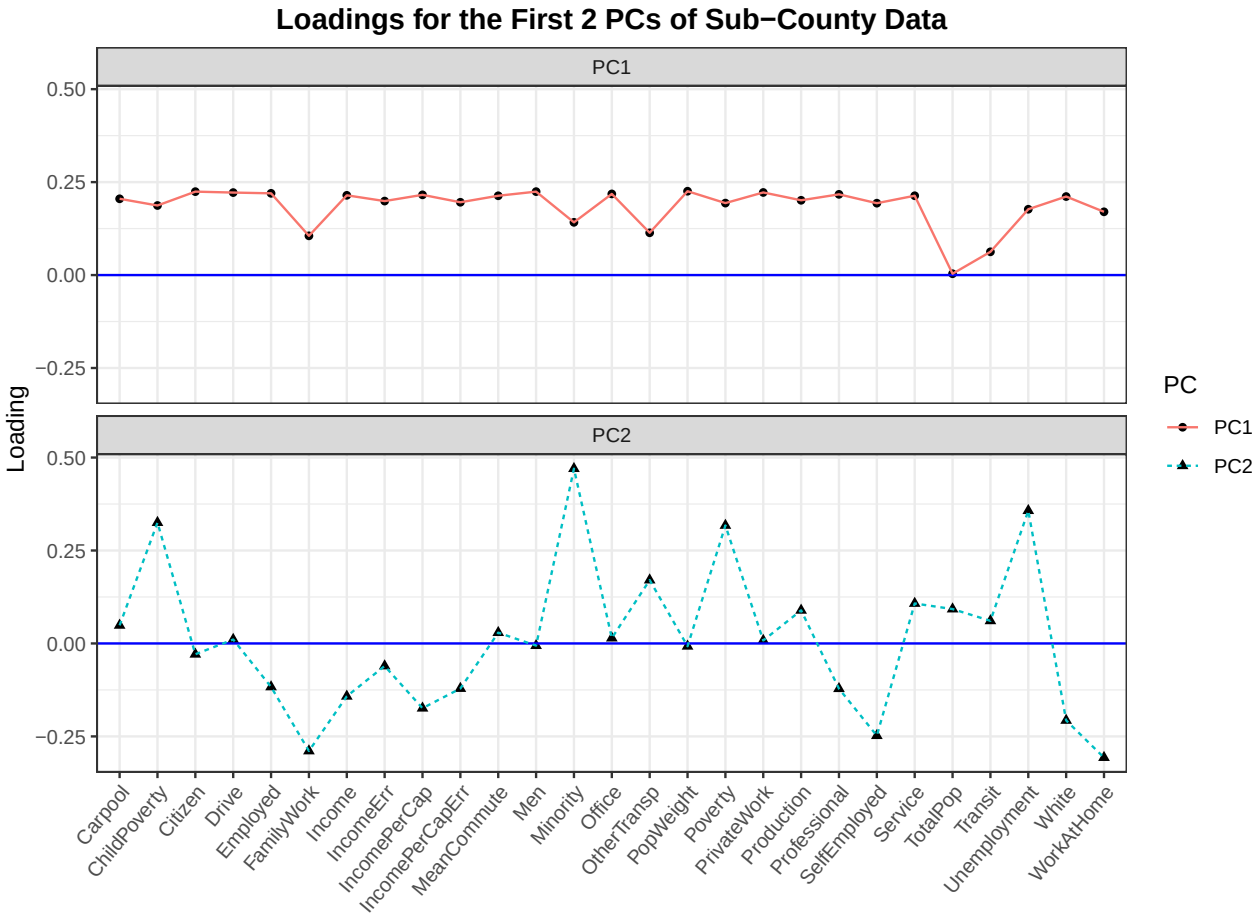
We observe the loadings plot from the county level census data first:



For PC1, it will have a large value whenever a county's poverty rates, the percentage of its population who are a minority, the unemployment rate, and the proportion of the workforce who are employed in service jobs are high while the employment rate, the total income and income per capita (plus their respective errors), the percentage of the workforce who are either employed in a professional industry or work at home, and the population proportion who are white are low. A large PC1 value seems to describe a rural county whose populations mostly comprises of minority groups and high unemployment rates due to a lack of job opportunities.

For PC2, a large value occurs whenever the income levels, the mean commute time in minutes, and the overall total population of the county (combined with the minority population) who are employed in office jobs, or unemployed, and commute via public transportation are higher than average while the number of citizens, the percentage of the county's population who are in unpaid family work, self-employed, white, or who work at home are lower than average. A high PC2 value seems to characterize a US county with a minority-majority population and has a higher-than-average income level thanks to its workforce mostly comprising of office workers who frequently commute via public transportation.

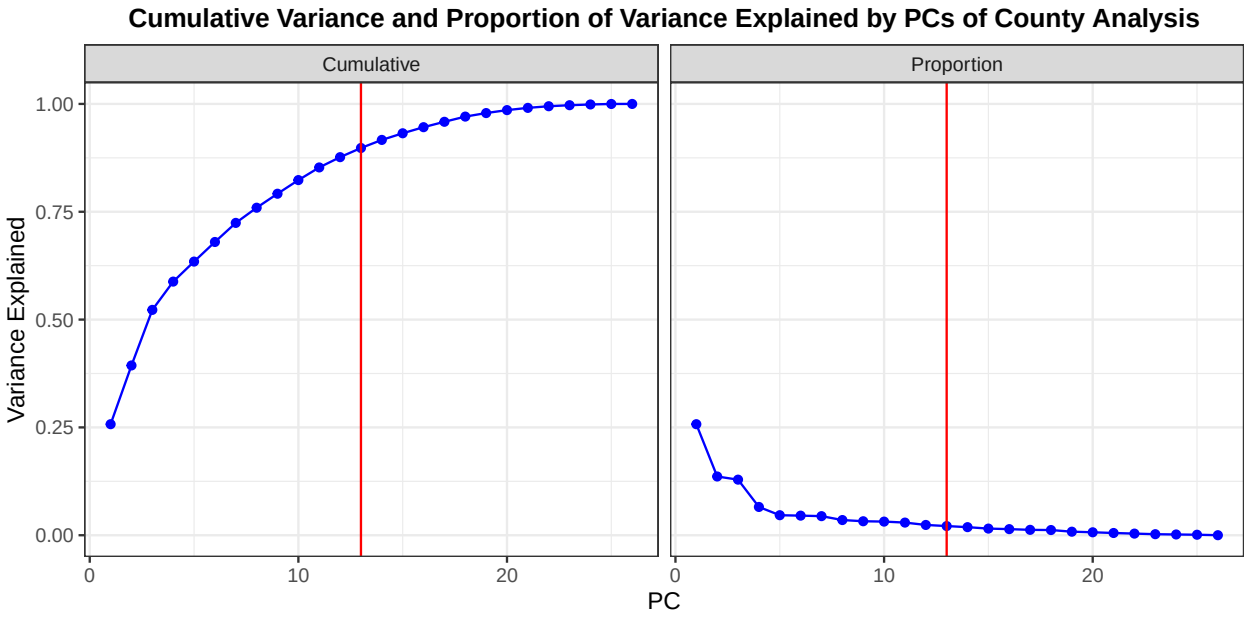
Next, we analyze the loadings plot from the sub-county data:



For PC1, the only variables that are not above average by a significant margin are the sub-county's total population and the percentage that commute on public transportation. Every other variable is above the average, which implies that a high PC1 value describes a high-income level sub-county with low unemployment rates and a diverse population with diverse job opportunities, whether working from home or at an office.

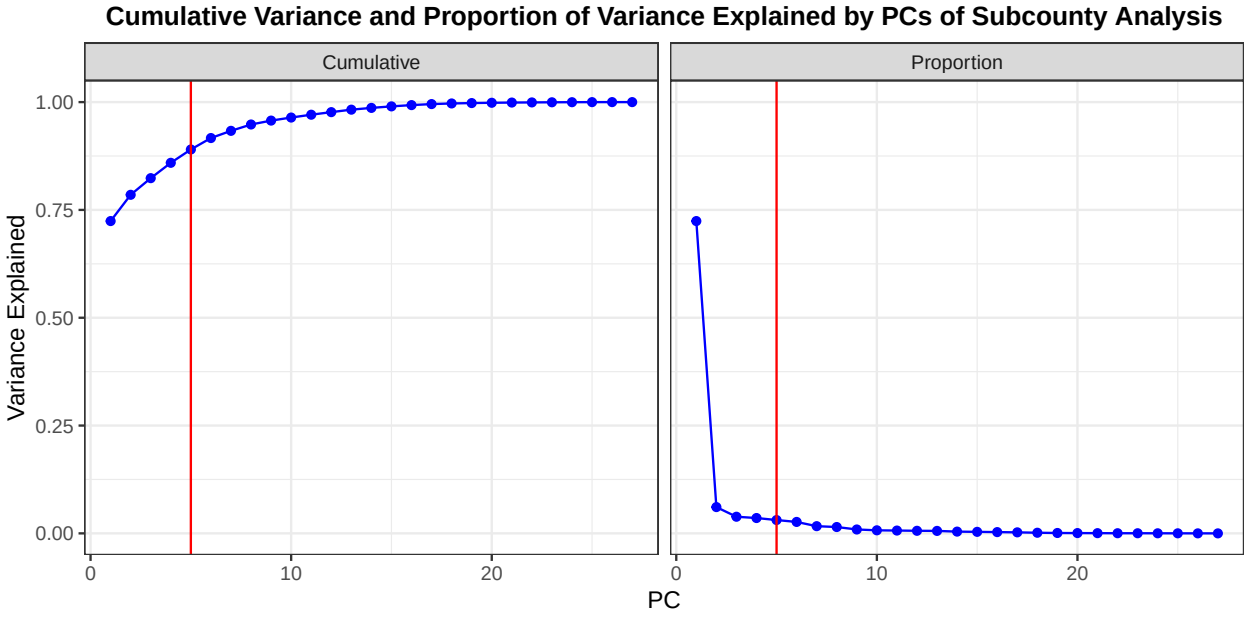
PC2 will have a high value when the percentages of the sub-county's population that are under the poverty level, who are in a minority group, and the unemployment rate are significantly high, whereas the percentage of the sub-county's population who are white and either self-employed, doing unpaid family work and/or working at home are low. High PC2 values indicates a sub-county with a minority-majority populations with a rampant poverty problem due to its high unemployment rate.

Next, we determine the minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. We begin by plotting both the proportion of variance explained and cumulative variance explained for the county data:



Based on our plots, we have determined that 13 is the minimum number of PCs needed to capture 90% of the variance in the county data.

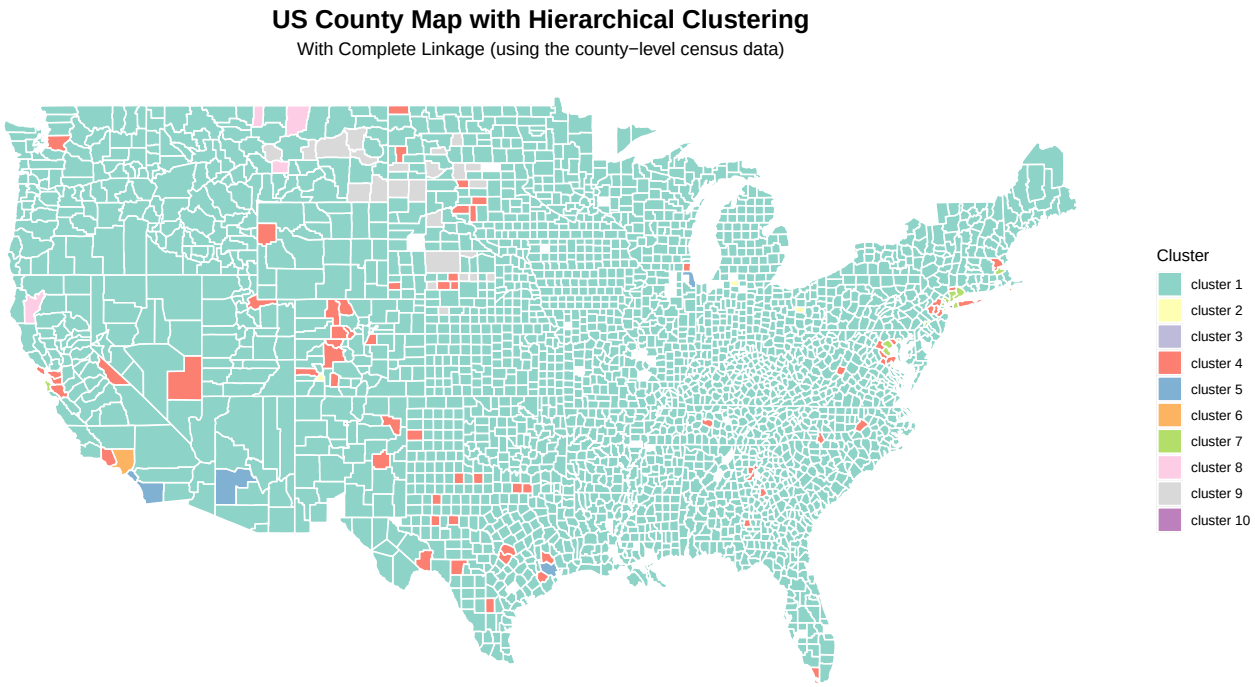
Next, we observe the plots depicting the proportion of variance and cumulative variance explained for the sub-county data:



Based on the cumulative plot, we see that we need 5 PCs at minimum to capture 90% of the variance in the sub-county data, with the first PC being the most significant as it is capturing nearly 75% of the data.

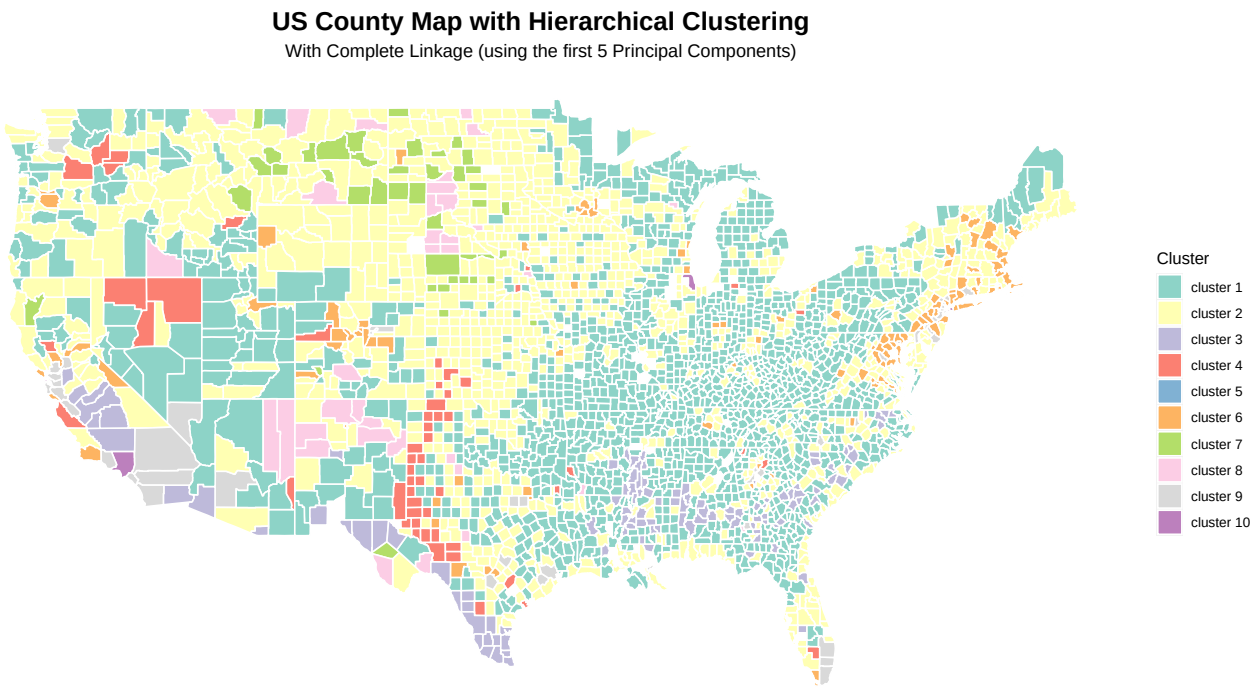
To examine if we can allocate labels to the county-level census data, we perform hierarchical clustering with complete linkage on both the original features and its first 5 principal components. We are aiming to cut the tree created by the clustering to partition the observations into 10 clusters.

We begin with clustering the original county-level census data and plotting the clusters on the US county map as follows:



We see that using the original features leads to a poorly-balanced clustering since a vast majority of the counties are in cluster 1.

Next, we cluster the first 5 PCs of the county-level data to see if it features a better balance:



Compared to the prior plot, this visualization looks much more promising since cluster 1 is nowhere near as dominant across the US as it was before, even if it looks as if it is still the most popular one. Every other cluster received a significant bump in popularity, particularly clusters 2 and 7, which were near nonexistent in the prior clustering.

To determine which data set gives us a more appropriate clustering, let us observe which cluster San Mateo County was assigned to for both approaches:

State	County	Assigned Cluster	Data Used
California	San Mateo	cluster 9	First 5 principal components of county-level data
California	San Mateo	cluster 7	Original county-level data

The clusters that the county was assigned to are both rare occurrences based on the prior visualizations. However, the cluster that San Mateo was assigned to based on the first 5 PCs (cluster 9) shows fairly frequently in California as well as in bits and pieces across the West Coast and East Coast, while the originial features' clusters were much rarer and harder to decipher a pattern.

In order to find a more measurable way in determining which data set clusters San Mateo County better, we take the average of all the numerical variables of the counties contained in each cluster of their respective data sets, and then find the absolute difference between the averages and the data from San Mateo County. Unfortunately when we do this, we are at risk of producing biased differences because cluster 7 from the original county-level data has undoubtedly way less observations than cluster 9 from the first 5 PCs. Regardless, it is worth performing such a calculation to see if we can determine in what ways does each clustering approach succeed in where the other one falters. When performing the calculation, we get the following table:

Table 8: Table continues below

Cluster & Data Used	TotalPop	Men	White	Citizen	Income
Cluster 9; First 5 PCs	414671	0.2207	1.737	1.932	33212
Cluster 7; Original Data	10820	0.7279	9.699	3.779	12632

Table 9: Table continues below

IncomeErr	IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty
5561	16599	1699	7.647	11.1
2310	3649	165	4.119	5.233

Table 10: Table continues below

Professional	Service	Office	Production	Drive	Carpool	Transit
8.214	1.213	1.742	3.271	1.723	0.05907	1.074
2.723	0.7826	0.9595	0.2795	12.21	2.551	12.93

Table 11: Table continues below

OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork	SelfEmployed
0.5517	0.6889	1.857	4.324	0.2067	2.781
0.3684	0.5053	5.351	0.3232	2.908	2.747

FamilyWork	Unemployment	Minority
0.03426	2.615	0.7807
0.05632	1.327	8.627

Based on the table, we can see that the first 5 PCs performs surprisingly worse in many of the numerical variables compared to the original features. Although we could attribute this discrepancy to the imbalance of observations for each clustering, the interesting exceptions where the first 5 PCs' clustering performs better are in the demographic variables, particularly the `White` and `Minority` columns which represent the percentage of the county's population who are white and of a minority group respectively. Other notable exceptions like all of the transportation-based variables (i.e. the `Drive`, `Carpool`, and `Transit` variables which correspond to commuting via driving alone, carpooling, and public transportation respectively) are worth noting too, since there is a clear definite pattern as to where the PCs' cluster outperforms the original features'.

Although the hierarchical clustering based on the first 5 PCs does perform measurably worse than the original features' in many of the variables, we affirm that it puts San Mateo County in a more appropriate cluster. This is because the few variables it does perform better at are worth further research, i.e. the demographic and transportation-based variables as well as their more diverse clustering based on their visualizations lending more promise to their validity.

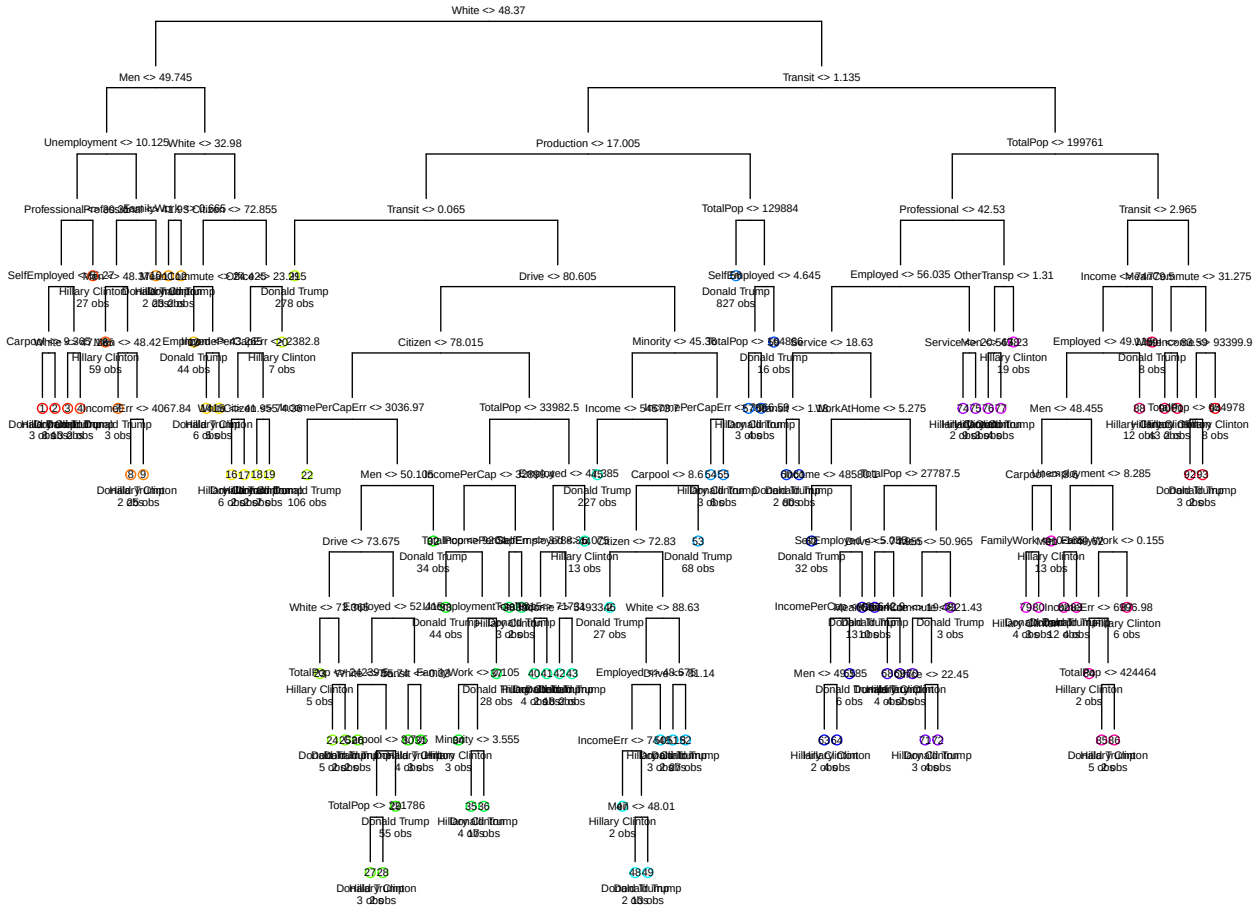
Classification

In order to train classification models, we need to combine the data frame of the winning candidate by county and the aggregated county census data.

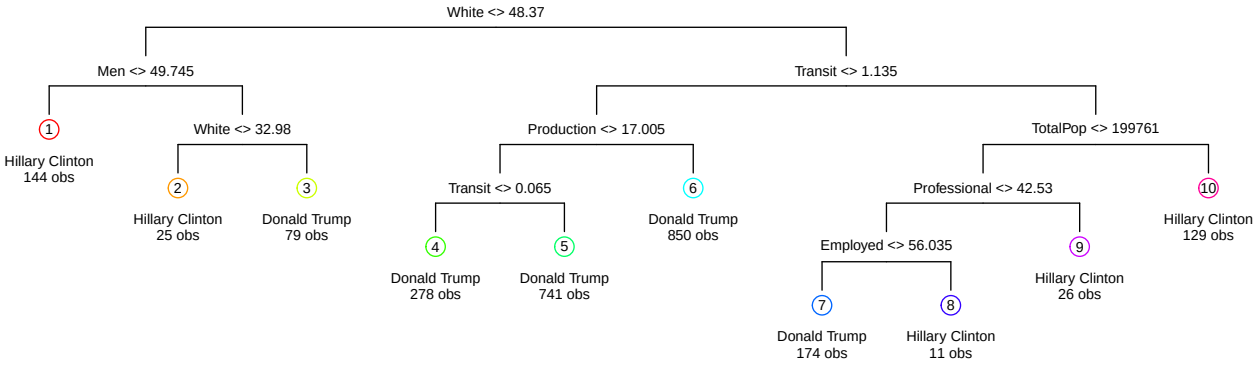
After merging the data, we partition the result into 80% training and 20% testing partitions.

Decision Tree

Let us train a decision tree on our training partition and apply cost-complexity pruning. Before we prune, let's visualize how our decision tree looks without any pruning:



As we can see, the tree is way too busy and too uninterpretable, which highlights the importance of pruning. We then apply cost-complexity pruning to the initial tree to get a much cleaner visualization:



We get a much more interpretable visualization of the optimized decision tree. When we compute the misclassification rates of our optimized tree:

	Donald Trump	Hillary Clinton
Donald Trump	0.9363	0.06367
Hillary Clinton	0.1266	0.8734

We see that it is correctly predicting the candidates pretty well, although it is not perfectly balanced since

the misclassification error rate for Hillary Clinton is two times the misclassification rate for Trump. The overall misclassification rate is *0.07178* which is still pretty low, but could definitely be reduced. Regardless, our optimized decision tree has done an admirable job predicting the voting behavior of US counties based on certain factors.

Using our more interpretable decision tree visualization, we can construct a narrative about which kinds of counties will vote for who based on this decision tree:

If a county featured a minority-majority population (i.e. if the percentage of white citizens is less than 48.37%), then Hillary Clinton would win its majority vote if the majority of its population were women or if less than 33% of the men were white. However, if a county did feature a white-dominated population, then Clinton could still win its majority vote as long as the percentage of the population that commuted via public transition was greater than 1.14% and the total population is greater than 199,761. But if its total population was less, then Clinton could still win the majority vote as long as over 42.5% percent of the workforce was employed in a professional industry or, if less, the employment rate was above 56%. Based on this decision tree diagram, the narrative we can construct for Hillary Clinton is that she would win counties with female-minority majority populations or with an urbanized population with a high employment rate and a significant proportion of the workforce employed in a professional industry like management, business, science or arts.

Meanwhile, for Donald Trump to win the majority vote of a county with a minority-majority population, then it must also be a male-majority population where over 33% are white. If a county did feature a white-majority population, then Trump could win its majority vote as long as the percentage of the population that commutes via public transportation is less than 1.14% or if greater, the total population would have to be less than 199,761 with less than 42.5% of the workforce employed in a professional industry and with a sub-56% employment rate. Counties that featured populations with whites comprising a majority of the male demographic or a white-majority population with a significant unemployment rate were more likely going vote for Donald Trump than any other candidate.

Logistic Regression

Next, we will train a logistic regression model on the training partition of the county-level census data to predict the winning candidate in each county. Let's observe the following summary report from the model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-19.62	9.501	-2.065	0.03895
TotalPop	2.189e-07	3.735e-07	0.5862	0.5577
Men	0.05216	0.05044	1.034	0.3011
White	-0.1756	0.06269	-2.802	0.00508
Citizen	0.1005	0.02859	3.516	0.0004383
Income	-7.398e-05	2.669e-05	-2.772	0.00557
IncomeErr	-1.585e-05	6.64e-05	-0.2387	0.8114
IncomePerCap	0.0002222	6.834e-05	3.251	0.001149
IncomePerCapErr	-0.0003853	0.0001879	-2.05	0.04034
Poverty	0.02159	0.04077	0.5296	0.5964
ChildPoverty	0.002692	0.0256	0.1052	0.9162
Professional	0.2559	0.0391	6.545	5.954e-11
Service	0.3315	0.04945	6.704	2.026e-11
Office	0.0808	0.04564	1.77	0.07667
Production	0.1366	0.04181	3.268	0.001081
Drive	-0.1728	0.0455	-3.798	0.0001459
Carpool	-0.1643	0.06037	-2.722	0.006489
Transit	0.09621	0.09716	0.9902	0.3221
OtherTransp	-0.06553	0.09462	-0.6926	0.4886
WorkAtHome	-0.0998	0.07702	-1.296	0.195

	Estimate	Std. Error	z value	Pr(>|z|)
MeanCommute	0.04437	0.02428	1.828	0.06758
Employed	0.1955	0.03249	6.019	1.755e-09
PrivateWork	0.09299	0.02177	4.272	1.941e-05
SelfEmployed	-0.01521	0.0508	-0.2994	0.7646
FamilyWork	-0.8365	0.3921	-2.133	0.03291
Unemployment	0.2028	0.04026	5.037	4.735e-07
Minority	-0.05053	0.06042	-0.8364	0.4029

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	2051 on 2456 degrees of freedom
Residual deviance:	857 on 2430 degrees of freedom

The most significant variables (i.e. the ones with a p-value way below the 0.05 threshold based on the `Pr(>|z|)` column) are the proportion of a county's population who are white, the number of citizens, the median household income, the income per capita, the percentage of the workforce employed in a professional industry, production, private industry, in service jobs and unpaid family work, the percentage of the population who commute via driving alone and carpooling, and both the employment and unemployment rate.

The variables deemed significant by the logistic regression model is not consistent with those of the pruned decision tree, since there are more significant variables and some glaring omissions as well. The percentage of the workforce employed in the professional or production industry as well as the employment rate are still deemed significant in both models, but there are a lot of new variables considered just as important too such as the number of citizens or the percentage of employees working at service jobs. Furthermore, there are some variables from the decision tree cut from this model such as the male demographic or the percentage who commute via public transportation.

Even with the variables that both methods consider important, the degree of importance to the overall classification is not always equal. For instance, although the white demographic of a county is considered significant for both methods, the logistic regression model downplays its significance more as its p-value is only marginally below 0.05. This means that it has a 1% probability of being insignificant, whereas in the decision tree, it is the very root of the classification, highlighting its central importance to its classification.

Furthermore, we can express a few of the estimated coefficients in terms of a unit change in the variable. For example, a 10% increase in employment in service jobs in a county would correspond to an effect of 10 * 0.33 = 3.3 unit increase in the candidate variable, bumping the chances that Hillary Clinton will win its majority vote. Meanwhile, a 15% bump in the percentage of a county's population who drive to work will result in an effect of 15 * -0.1728 = -2.592 unit decrease in the candidate variable, meaning that the county would edge closer to voting for Donald Trump as its majority vote candidate.

Now let us test our logistic regression model on the test partition of our county-level data and observe the error rates:

	Donald Trump	Hillary Clinton
Donald Trump	0.9403	0.0597
Hillary Clinton	0.07792	0.9221

We observe that the error rates look a lot more balanced here than they did with the decision tree error rates. Our overall misclassification rate is

0.06199 which is smaller than what we got before. It looks like our logistic regression model features better

predictive performance compared our decision tree model.

We put our model to the test by using it on the entirety of the 2016 presidential election results. We get the following error rates:

	Donald Trump	Hillary Clinton
Donald Trump	0.945	0.055
Hillary Clinton	0.1662	0.8338

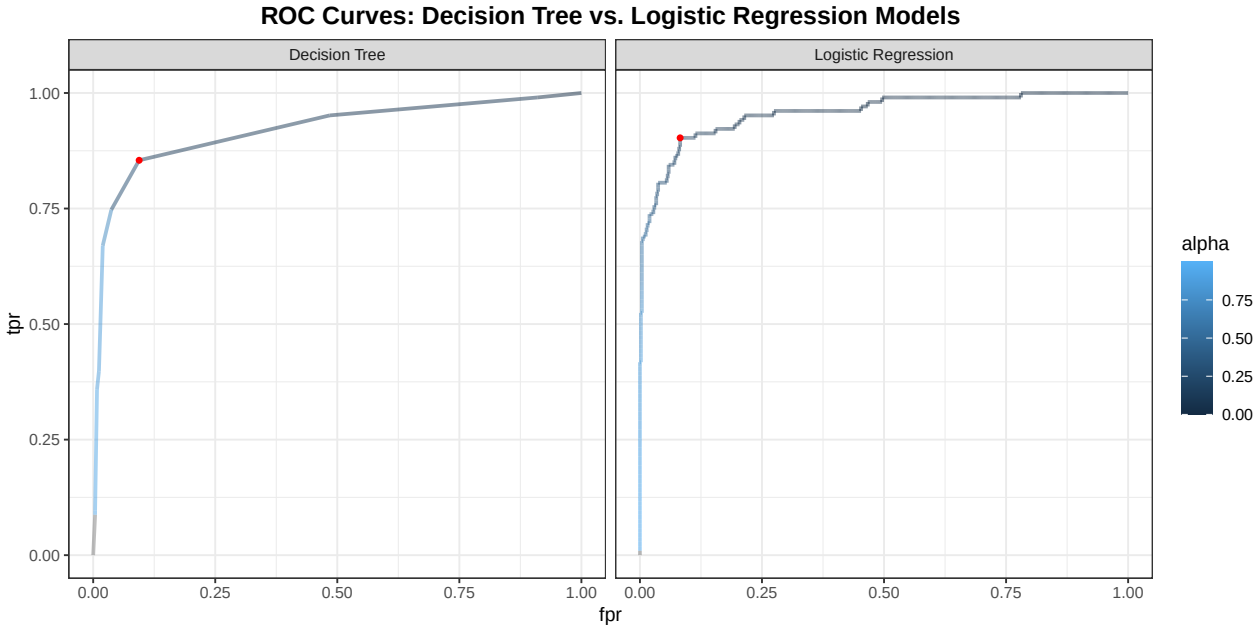
Although the model continues to do a solid job predicting which counties would vote for Donald Trump, it certainly took a hit when it came to correctly predicting counties that voted for Hillary Clinton, as the misclassification rate for predicting Clinton jumped from 0.07 to a worrying 0.1662. Let us see if the model correctly predicted the winning candidate of Santa Barbara County:

candidate	predicted	State	County
Hillary Clinton	Hillary Clinton	california	santa barbara

It looks like the results of this particular county match the predicted results, though the jump in misclassification rates for Hillary Clinton specifically is a cause for concern.

Decision Tree vs. Logistic Regression

Next, let us compute ROC curves for the decision tree and logistic regression model using predictions on the test data. We want to see if we can visualize the true positive rates and false negative rates for the sake of comparing the two classification methods:



An immediate pro of using a decision tree is that they are intuitive and easy to visualize and interpret. Although logistic regression is interpretable in the sense of being able to determine which variables are important for predicting class labels (such as which presidential candidate a county will vote for), a decision tree is immediately interpretable even to an outsider of the statistic field as its visualization will neatly lay out a roadmap of how a label will be determined according to a chosen set of variables deemed important.

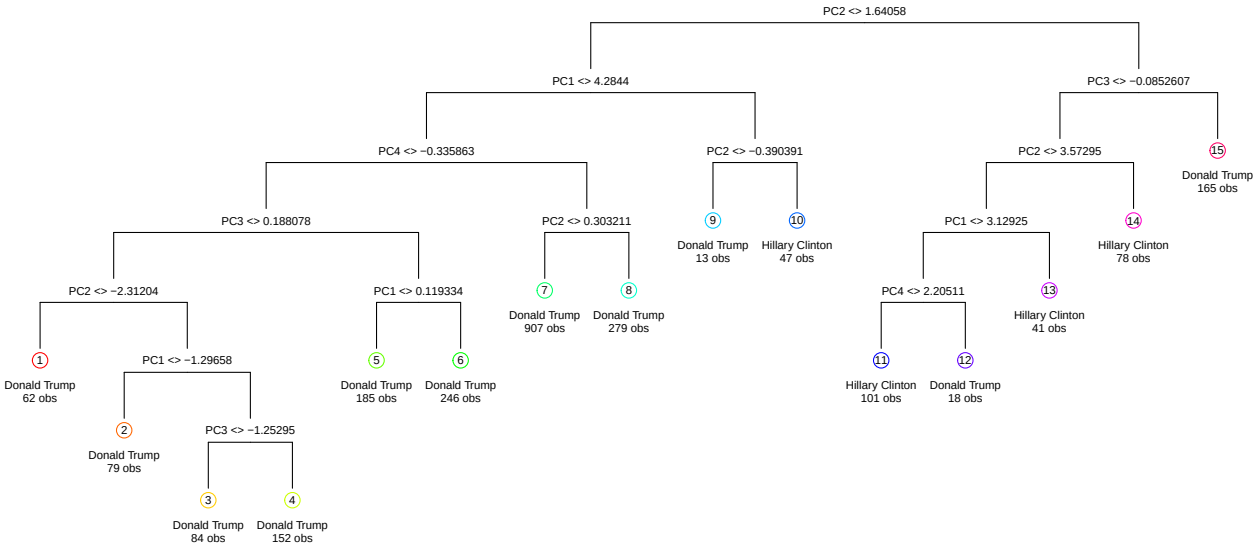
Decision trees also do a seamless and tidy job portraying relationships between certain variables. Although logistic regression could do the same using interaction terms, the amount of relationships between variables could increase to a ridiculous number to the point where the number of predictors would surpass the number of observations and cause trouble for the model's prediction accuracy. Finally, decision trees can handle more than three variables unlike logistic regression. This would be crucial if there was a situation where there was a third candidate that actually won some counties; the tree could easily handle an occurrence such as that.

With as many benefits as decision trees have along with their ease of use, it seems obvious to use one over a logistic regression model. However, due to how highly interpretable the trees are, there is definitely a big trade-off. The decision tree's predictive capabilities are compromised as demonstrated earlier where the misclassification error rates of our decision tree on the test partition of the county-level data were more imbalanced than the logistic regression's. Furthermore, logistic regression features a more detailed estimation regarding which predictors are important unlike decision trees, where most of those calculations are hidden away for the sake of interpretability. Trees are also very sensitive to small changes in the data, whereas logistic regression can handle lots of noise variables while still being able to accurately decipher which variables are significant. If we were to have kept the `Women` variable in our census data sets, then that would have impacted the tree's structure significantly, whereas the logistic regression model would have immediately identified it as a redundant variable for being highly correlated with the `Men` variable and render it insignificant immediately.

There is really no clear answer for which classification method to use. It is a matter of what contexts call for their uses. In the case of the 2016 presidential election, using a decision tree $after$ the election would be effective in describing what factors and relationships would cause a U.S. county to vote for a specific candidate. However, a logistic regression model's effectiveness would be employed during an election, so it can continually be trained by new polling data to determine which counties will vote for who and predict the overall winner in each of the 50 states.

Taking It Further

Let's say that we want to train a classifier model that sacrifices all interpretability for the sake of obtaining more accurate predictions. How can we accomplish this? Looking back at the hierarchical clustering, we observed that PCA gave us more promising clusters than using the original features. So let's use them again for training a decision tree and logistic regression model. Perhaps we will increase our chances of getting better prediction results than the original models. Let us begin with finding the first 5 PCs of the county-level election results and split it into a training and testing partition like we did before. Then we will draw out the tree diagram constructed by a new decision tree model that utilizes the first 5 PCs training data:



Immediately, this tree is nowhere near as interpretable as the original tree because we are now using the PCs for determining splits. However, as we stated before, that was to be expected if it meant we were getting better prediction results. Let us observe the misclassification rates and see if the predictions were actually better:

	Donald Trump	Hillary Clinton
Donald Trump	0.9133	0.08672
Hillary Clinton	0.2113	0.7887

Our overall predictions using the first 5 PCs actually gave us worse misclassification rates than the original features! It does not look like reducing our county-level data set to the first 5 PCs actually improved anything in terms of predictions. Perhaps that was to be expected, since decision trees do lean more towards interpretability rather than prediction accuracy and using the 5 PCs to determine our tree's splits would have defeated its whole purpose.

Let us see if logistic regression based on the first 5 PCs performs any better:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.738	0.1091	-25.1	4.683e-139
PC1	0.03722	0.02663	1.398	0.1621
PC2	0.9335	0.05463	17.09	1.876e-65
PC3	-0.6133	0.04718	-13	1.264e-38
PC4	-0.3973	0.06205	-6.403	1.528e-10
PC5	-0.2838	0.07459	-3.805	0.0001417

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	2051 on 2456 degrees of freedom
Residual deviance:	1225 on 2451 degrees of freedom

Immediately the first thing to note is that almost every PC is considered significant to the regression *except* PC1. One would hope that every PC would help with the classification given that we are using only five, but apparently not in our case. Another interesting observation to note is that PC3 and onwards feature negative

estimates for their coefficients, which imply that a unit increase in those PCs would be in favor of Donald Trump winning a county. Once again, it is hard to decipher any true meaning from the PCs compared to the prior logistic regression where it was immediately obvious which factors contributed the most to the winning candidate of a county's majority vote.

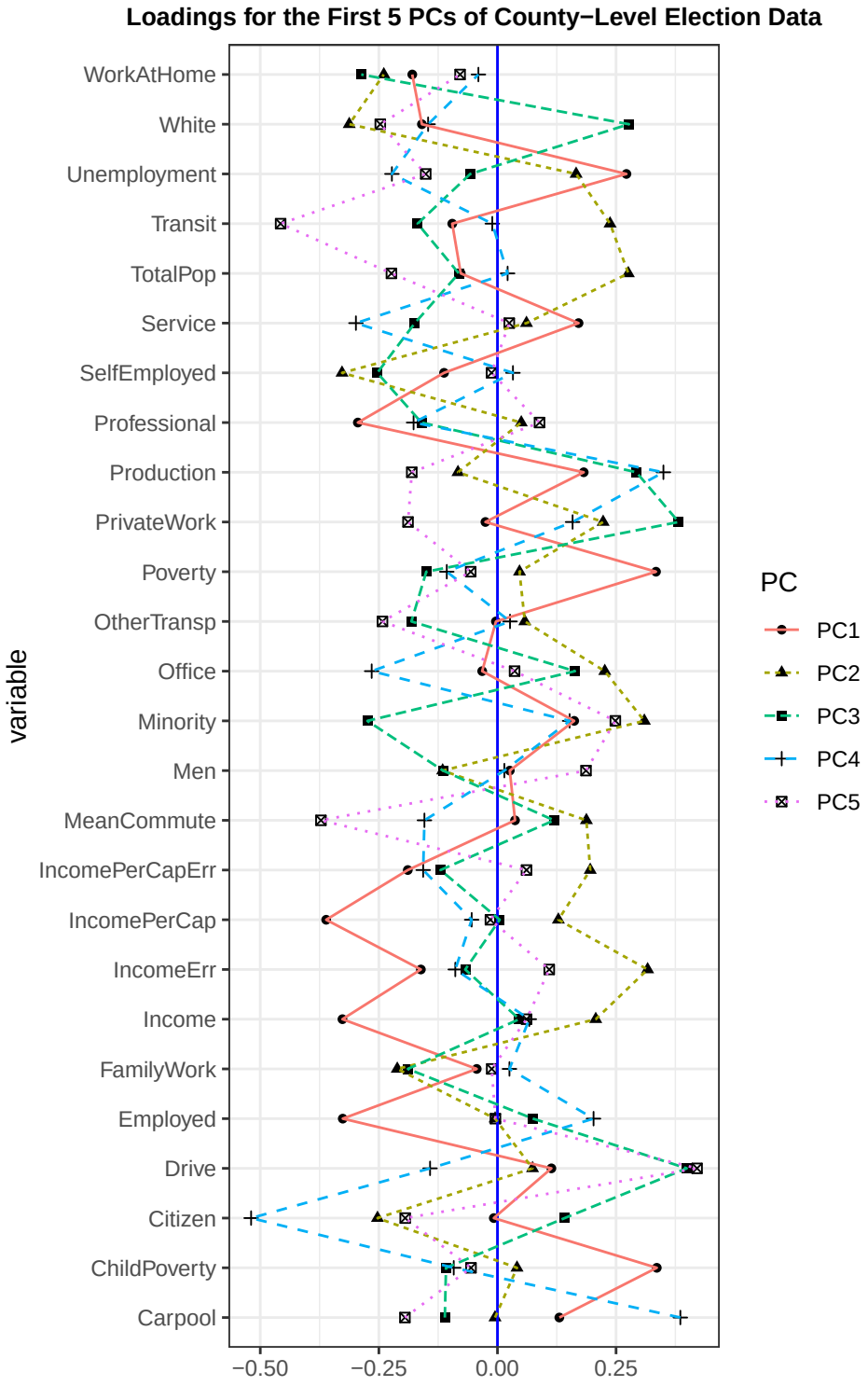
Next, let us see if the misclassification rates are an improvement from the original data sets:

	Donald Trump	Hillary Clinton
Donald Trump	0.8991	0.1009
Hillary Clinton	0.1897	0.8103

Nope! It actually performed worse than the original logistic regression model as it features significantly higher misclassification rates on the test partition.

Even with sacrificing more interpretability for the sake of better predictions, we actually got worse results using PCA than we did with the original model. Was this whole detour completely useless then? Probably, but PCA's failure here does demonstrate that the relationship between prediction results and predictors was perhaps never linear from the start. The whole purpose of analysis was to find a lower-dimension *linear* approximation of the predictors. This failure does also illustrate the critical importance of balancing interpretability and predictive accuracy in the context of election results. In an attempt to sacrifice all interpretability for the sake of better predictions, we actually managed to get worse results than we did using the original features! Perhaps it is because of that failure that we realize that predicting election results are ultimately dependent on interpretability to obtain more meaningful information on which predictors are the most important or even insignificant. Maybe that is a factor as to why a vast majority of the election forecasting during the 2016 presidential election failed to correctly predict Trump's victory, including Silver's; they traded off too much interpretability for the sake of presumably more accurate predictions. Through that sacrifice, less interpretability lead to a lack of understanding in where their models could potentially be performing poorly and, thus, foster certainty bias.

Let us see if we can somehow salvage our PCA analysis through interpreting the loading plots for all 5 of our PCs:



Based on our logistic regression analysis, PC1 was considered insignificant, so let us choose to ignore that PC for our plot analysis here. We can see that there are strong gaps between loadings at the demographic variables (`White` and `Minority`), the percentage of a county's population that commute via public transportation, by themselves, or by carpooling as well as the average commute time, a county's total population, its number of citizens, and the proportion of the workforce employed in a production industry, private industry or in office jobs. One particularly interesting aspect about this plot that is consistent with our logistic regression model summary is that PC2 is often the PC that is driving a majority of the separation as evident in certain

variables like `Transit` or the demographic variables. It is also worth noting that in the logistic regression summary, a unit increase in PC2 would lead a county to lean more towards voting for Hillary Clinton as its majority vote candidate, so its divergence from the other PCs makes sense.

We can interpret a lot more from this plot when utilizing what we computed prior in our logistic regression analysis, such as the kind of county archetypes the PCs represent with respect to which presidential candidate they align more towards. Therefore, our PCA analysis was not a complete waste since we could utilize our failed classification models from before for even more in-depth interpretations!

Thus, looping back to the central question of this whole tangent: is sacrificing interpretability in our models worth it for better predictions? Based on our implementation of PCA in our classification models, we have to argue against making that drastic of a trade-off. While prediction accuracy is critical for our model's effectiveness, interpretability is just as integral as it can lead into even more interesting angles that could warrant further research (such as identifying what general conditions a county would have to meet to be more aligned towards a specific candidate). Furthermore, it can help us identify problems in our models such as which variables are suspiciously considered less important than others. For instance, we chose to ignore PC1 in our loading plot analysis because the logistic regression model's summary deemed it insignificant, but doing so means we also would choose to ignore the clear separation of income-based and poverty-based variables driven by PC1.

Perhaps the temptation of winning the glory of one's model best predicting the outcome of a presidential election out of every other competing model is considered worthy enough to sacrifice all interpretability. But doing so means we cannot interpret what could go wrong with the model or any possible alternative routes worth exploring. For those reasons, the interpretability-prediction trade-off is an especially important topic in predicting election outcomes.