Final Project: Analyzing the 2016 U.S. Presidential Election Results

PSTAT 131, 2021 Winter Quarter

Venk Muriki(6862379), Jessica Lopez (9886383), Karine Babadzhanian(9679804)

Abstract

In this project, we are going to be analyzing the 2016 Election Results Data along with the census data that contains features that we can use to understand more about and classify voting inclinations of certain populations. We will then analyze purple counties and determine further course of action to improve modeling election voting behavior.

Methods

After cleaning up the data and conducting pre-processing, we are going to use some R-Studio map visualization tools at the county and state level to display voting behavior. We are going to be executing Principal Component Analysis (PCA) at the county and sub-county levels and then analyze the loading plots to describe interpretable attributes of the counties. Then, we are going to be carrying out the hierarchical clustering algorithm on (1) the original features and (2) the first 5 principal components at the county-level data to determine which algorithm better encapsulates San Mateo County, CA. Then we are going to compare and contrast the decision trees versus logistic regression as classification methods in predicting candidate result based on misclassification error rates, and then we will construct ROC curves to determine the better classifier based on True Positive Rate maximization. After running linear regression to predict total vote by county, we will analyze the difficulty of predicting at the purple counties via PCA.

Introduction

Polls are a problematic source of data, they are arranged at state and national levels so both have to be taken into consideration when doing a prediction. There are many variables, such as voters behavior. What people think changes over time and also can be affected by many other factors. Changes in the economy, employment, income taxes or something as simple as a successful campaign ad causes changes in voters decision. The data collection can be bias and therefore not representative of the actual population voting on Election Day. People responding to the polls may not end up voting, or they provide false answers if they feel like their decision is being judged. To even solve the sampling error problem among other things, pollsters may correct for false or biased surveying results in different degrees in the wrong direction. Plus factors such as third party favoring votes or undecided voters changing their vote at the last minute might be hard to predict except for looking in past elections that are a lot more outdated since elections happen every 4 years, a lot of economical, health-care, globalization-related problems should be addressed periodically.

First of all, he used hierarchical clustering, which was key when one can't predict based on unpolled states but this type of modeling allows easy migration, from one state to another or from one day to another. Instead of going off the maximum likelihood of probabilities for the house effect changes in percentage points, he went

with a full range of probabilities and used that as better metric; Bayes Theorem and graph theory was used. This made him adjust for very deviant changes in the actual support. He was able to understand the possible sources of error and variation and tackle those head on to make a good model. He was also able to use his model to find out key states such as Ohio at the time based on his interpretable model and not Florida as one who would think instinctively. He also started early; he was able to build the model from this baseline in January and feed it more data to get better at prediction, especially nearing the election.

The polls were biased towards Clinton winning the election and underestimated Trump support. Compared to Clinton voters, Trump voters might have responded differently on the polls due to the group being less transparent, clear, confident in their response. James Lee of Susquehanna Polling & Research Inc. said his firm combined live-interview and automated-dialer calls, and Trump did better when voters were sharing their voting intention with a recorded voice rather than a live one. Pollsters also cited lower-than-expected turnout, particularly in the Midwest. The turnout models appear to have been badly off in many states," said Matt Towery of Opinion Savvy (fivethirtyeight.com). One solution can be for increased funding for polling places. The mentality and culture of pulling places can be transformed at the managerial level, to honestly work on trying to get the right number as opposed to the most popular or common number by other polling places. Also, exploring ways in which pollsters can gain trust of the general people can result in external validity. Also, diversifying the strategies of reaching certain votes based on the demographics, social class, jobs, etc, can result in more responses.

Results

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

The meaning of each column in `election_raw` is self-evident except `fips`. The accronym is short for Federal Information Processing Standard. In this dataset, `fips` values denote the area (nationwide, statewide, or countywide) that each row of data represent.

Let's inspect fips = 2000:

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 want to exclude fips = 2000 because it's a duplicate of the state AK (Arkansas). This fips value also has an accompanying NA value for county which doesn't get any new info to us on that aspect.

After removing fips = 2000, election_raw has 18345 observations and 5 columns.

Census data

The first few rows and columns of the `census` data are 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

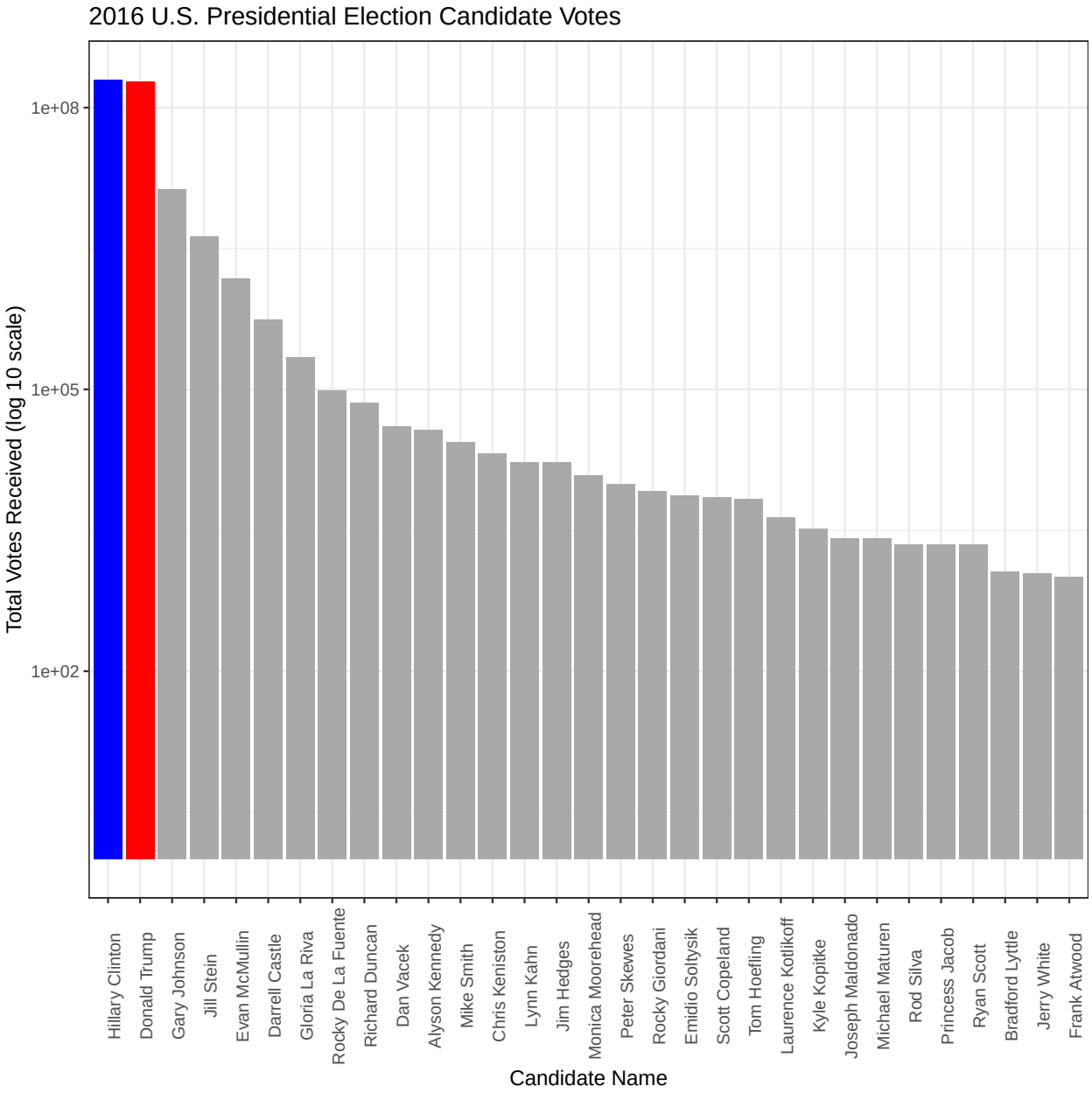
Variable descriptions are given in the `metadata` file. The variables shown above are:

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

Now we separate the rows of `election_raw` into separate federal-, state-, and county-level data frames.

There were 31 named presidential candidates in the 2016 election. Here is a bar graph of all votes received by each candidate by decreasing vote counts on a log-transformed scale.

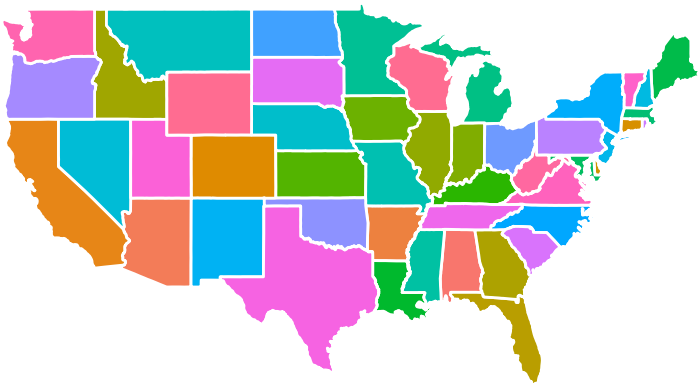
```
## `summarise()` ungrouping output (override with `.groups` argument)
```



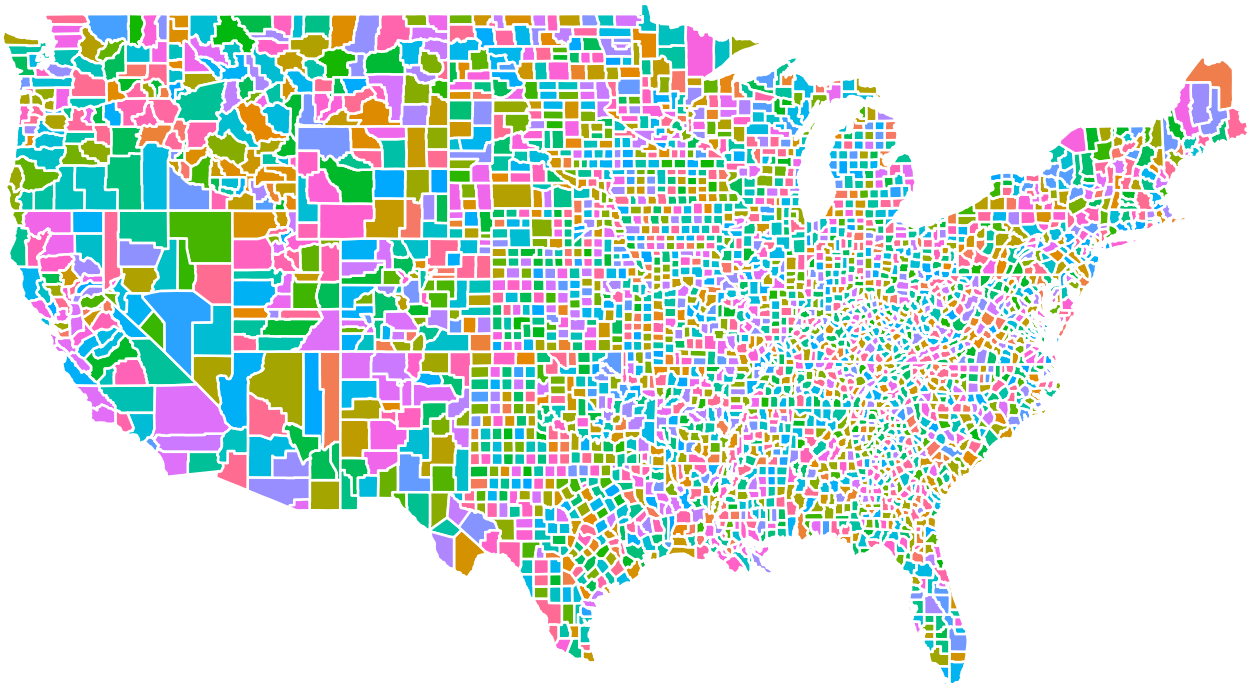
Now we create `county_winner` and `state_winner` by taking the candidate with the highest proportion of votes.

Visualization

We will conduct visualizations based on the following U.S. map:

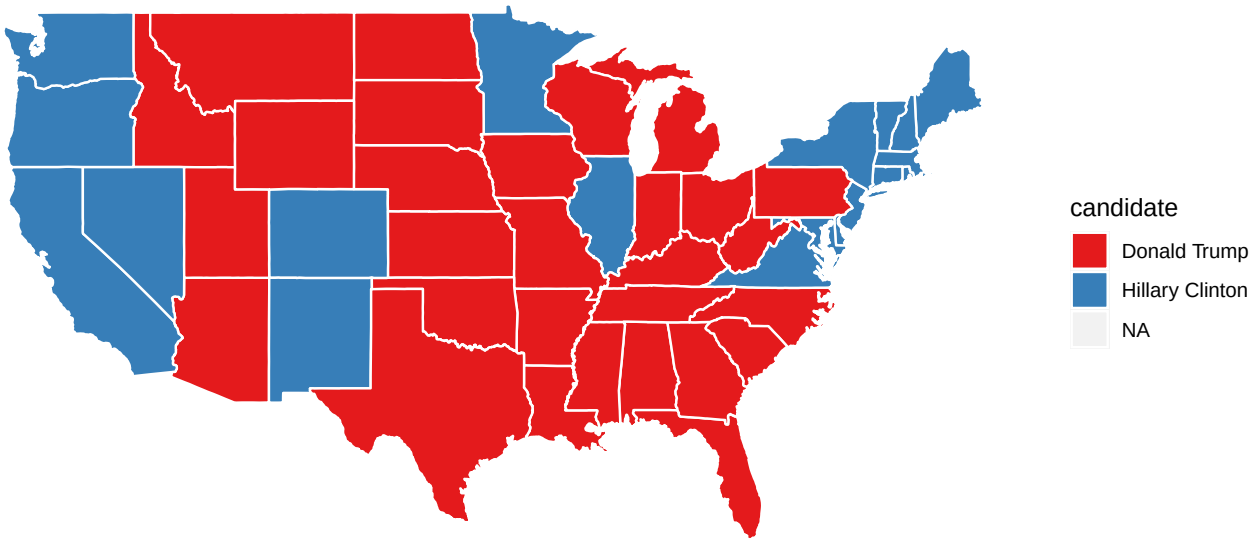


Here is a county-level map colored by county:



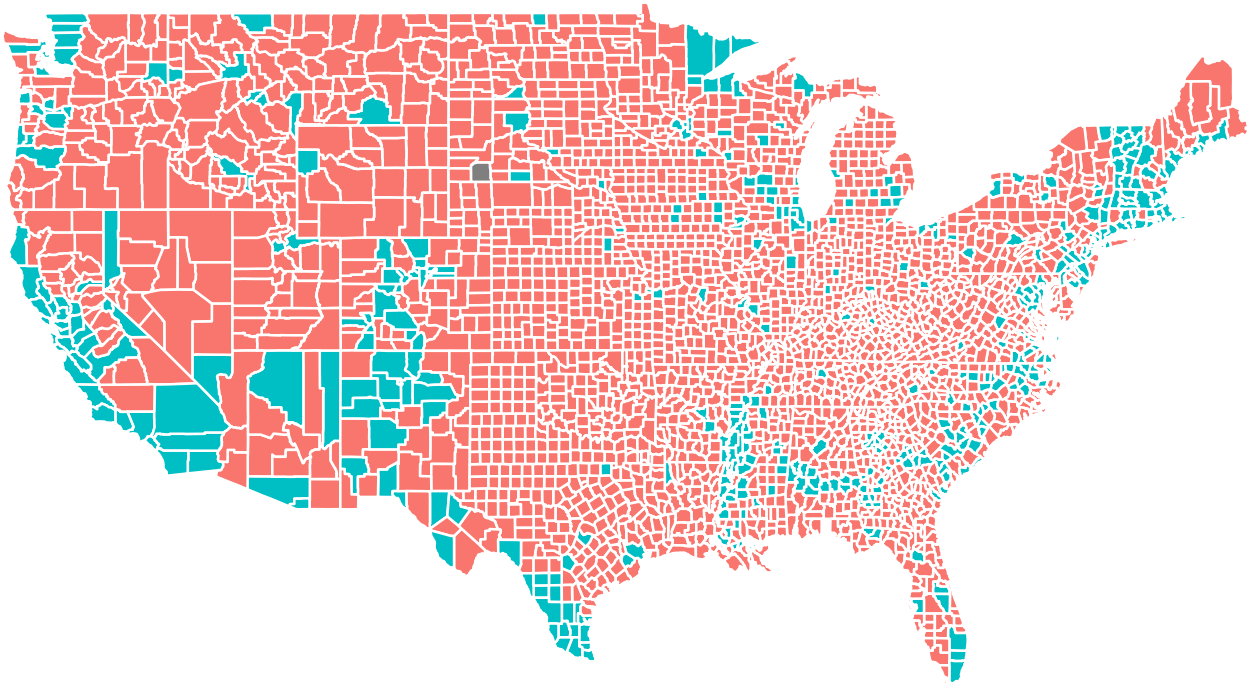
```
name2abb <- function(statename){
  ix <- match(statename, tolower(state.name))
  out <- state.abb[ix]
  return(out)
}

states$fips <- name2abb(states$region)
```

Here is the map of the election results by state:

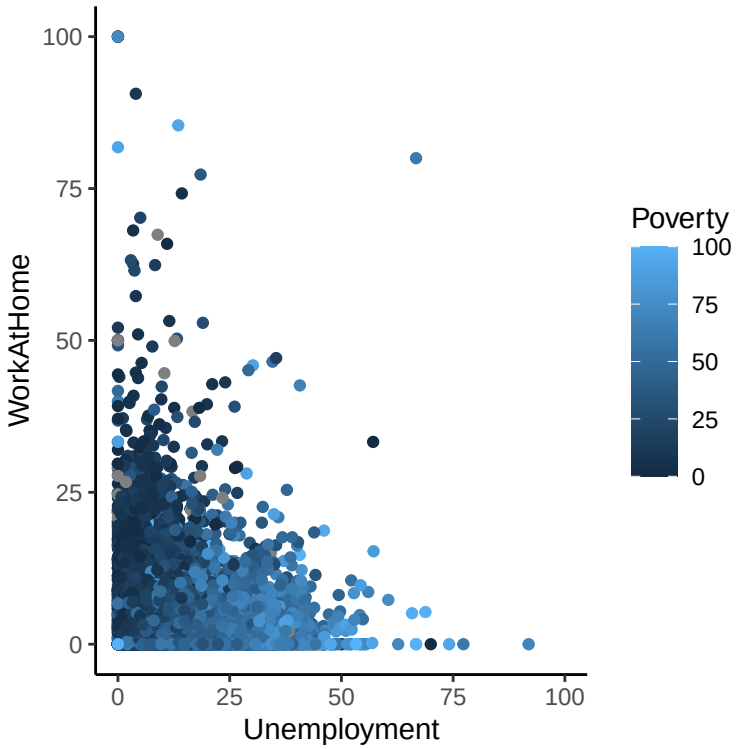


Here is the map of the election results by county:

```
## Joining, by = "fips"
```



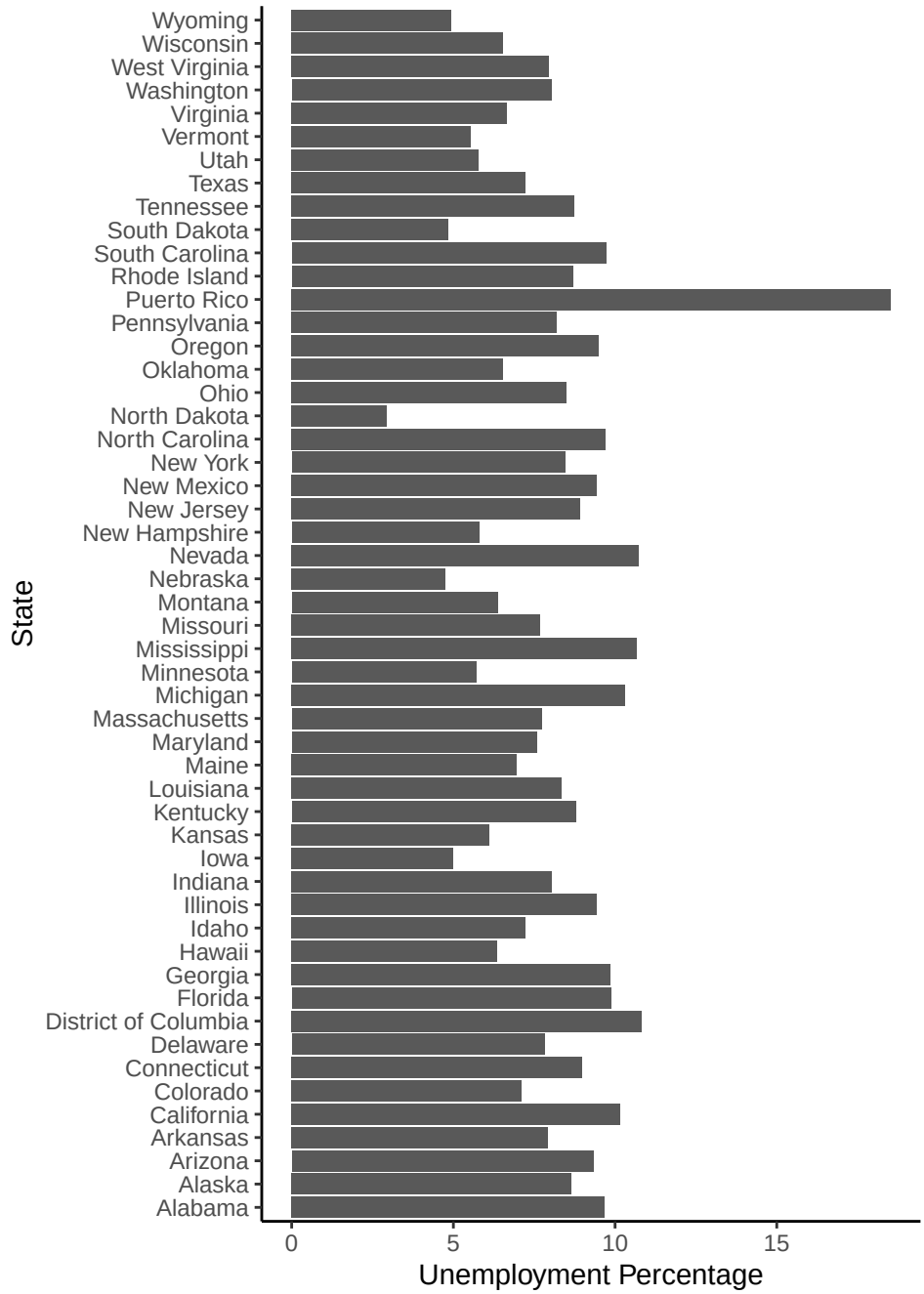
Using the `census` data, here is a scatter plot investigating the poverty percentage based on unemployment and working-at-home percentages of people. Each point is a Census Tract ID.



Based on this plot, there is a better discernability of poverty levels along the Unemployment axis and interestingly as expected, high poverty is associated with high unemployment. It seems that with low poverty comes more people working at home while high poverty comes with much less people working from home.

Now here is a bar graph for unemployment by state.

```
## Warning: `funs()` is deprecated as of dplyr 0.8.0.
## Please use a list of either functions or lambdas:
##
##   # Simple named list:
##   list(mean = mean, median = median)
##
##   # Auto named with `tibble::lst()`:
##   tibble::lst(mean, median)
##
##   # Using lambdas
##   list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```



Puerto Rico seems to be leading in unemployment while North Dakota has the lowest percentage.

13. The `census` data contains high resolution information (more fine-grained than county-level). Aggregate the information into county-level data by computing population-weighted averages of each attribute for each county by carrying out the following steps:

- Clean census data, saving the result as `census_del`:
  - filter out any rows of `census` with missing values;
  - convert `Men`, `Employed`, and `Citizen` to percentages;
  - compute a `Minority` variable by combining `Hispanic`, `Black`, `Native`, `Asian`, `Pacific`, and remove these variables after creating `Minority`; and
  - remove `Walk`, `PublicWork`, and `Construction`.

While aggregate information of cleaned county-level data based on population-weighted averages, we removed Women because it is highly correlated with the Men covariate. There is no new info from having the other if you already have one. Minority and White however do not add up to 100%, there is a mismatch due to some unaccounted factors unknown, so they are both kept.

We did the same at the sub-county level with added variable CountyPop, which is the population at the census tract at the county level.

Here are the first few rows and columns of aggregated `census_ct`.

State	County	TotalPop	Men	White	Citizen	Income	IncomeErr
Alabama	Autauga	55221	48.43	75.79	73.75	51696	7771
Alabama	Baldwin	195121	48.85	83.1	75.69	51074	8745
Alabama	Barbour	26932	53.83	46.23	76.91	32959	6031
Alabama	Bibb	22604	53.41	74.5	77.4	38887	5662
Alabama	Blount	57710	49.41	87.85	73.38	46238	8696
Alabama	Bullock	10678	53.01	22.2	75.45	33293	9000

Each of us are now going to share where we were during the 2016 election and analyze the corresponding demographic information.

Venk:

I was in Santa Clara County in NorCal at the time.

State	County	TotalPop	Men	White	Income	Poverty	Minority
California	Santa Clara	1868149	50.26	33.58	100744	9.748	63.06

Seems like there is an equal amount of men and women in Santa Clara as well as overall in California. One thing that surprised me is that the average income is almost double of a third of the other county incomes.That makes sense because my county is at the heart of Silicon Valley. I'm not that surprised that 2/3 of the population are minorities because growing up, in school or in my swim club, we would have a very diverse community. The poverty being low also makes sense because we have lots of small businesses and a very suburban area.

Jessica

State	County	TotalPop	Men	White	Income	Poverty	Minority
California	Orange	3104983	49.44	42.25	80007	12.96	55.12

I was in Orange County when the 2016 Elections. Looking at the demographics of the county, I am not very surprised. Orange County is comprised of several diverse cities such as Santa Ana, Irvine, Fullerton, Garden Grove and more so seeing a high minority percentage does not come as a surprise. On the other hand, it also as predominant white population cities such as Newport Beach, Huntington Beach, etc. When looking at the county_winner data, Hilary Clinton was the candidate with more votes in the county, this was somehow surprising. Orange County is known to be a strong republican county, so seeing the democratic candidate had the most votes was a pleasant surprise.

Karine

State	County	TotalPop	Men	White	Income	Poverty	Minority
California	Los Angeles	9995004	49.2	26.91	61385	18.23	70.66

county	candidate	state	votes
Los Angeles County	Hillary Clinton	CA	2464364

I was in Los Angeles county during the last elections. Looking at the cencus_ct data of my county I can definitely say that if among almost 10 million people, there are 71 percent of population that represents minority, no doubt that Los Angeles county contributed the vote to Hilary Clinton. In addition 18.23% of poverty is also a factor that effected the number of votes for democratic party. Los Angeles county counted in total 2,464,364 votes for Clinton which is almost 30% of a total number of votes (8,753,788) California state had for that candidate.

Discussion

Now we are going to carry out PCA for both county & sub-county level census data. The first two principal components PC1 and PC2 for both county and sub-county respectively will be computed.

We chose to center and scale the features because if we look at the following variances for county-level:

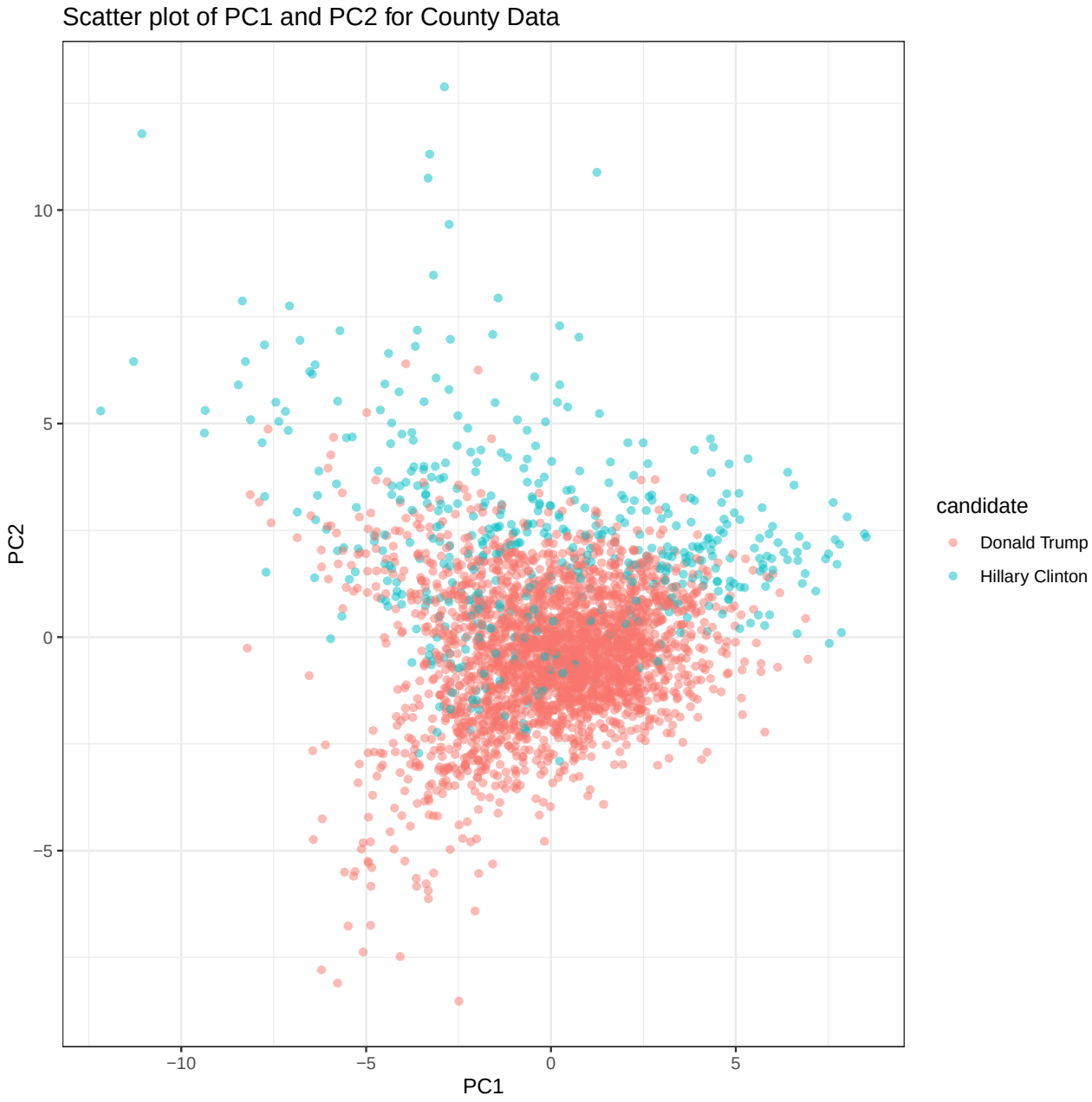
TotalPop	Income	IncomeErr	IncomePerCap	IncomePerCapErr
1.011e+11	187191546	5807218	38040721	1210144

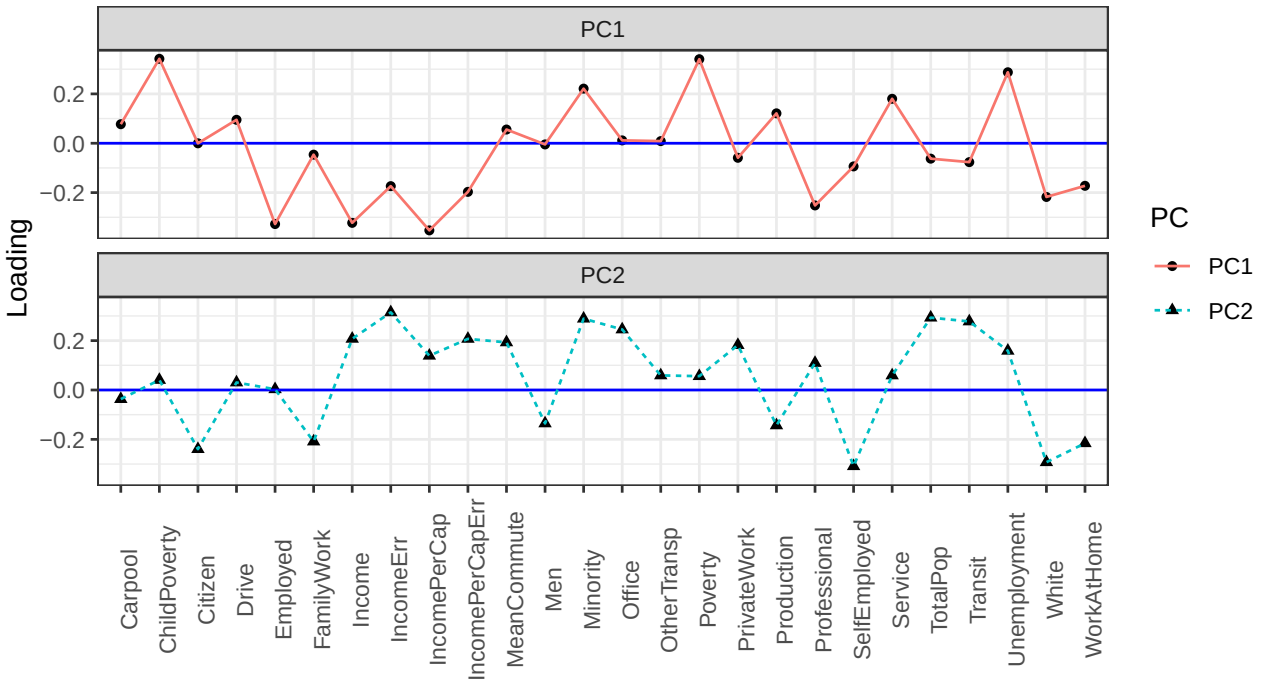
And now for sub-county level:

Income	IncomeErr	IncomePerCap	IncomePerCapErr	CountyPop
2e+07	502255	5543236	115438	3.712e+12

For both county census and subcounty census data, it looks like IncomeErr, IncomePerCap, IncomePerCapErr, and Income have very large variances, so if you do not center and scale, they will upweighted in the PCs simply because they have high variances. TotalPop for county-level and CountyPop for sub_county level are included respectively, also produce high variances. But with center and scaling, we impose unit variances for all the variables, so then we won't have PCs that are biased towards variables with high variances.

Here is a scatterplot showing the data projected onto the first two PC's by candidate at the county-level:


Scatter plot of PC1 and PC2 for County Data
PC2
10
5
0
−5
−10
−5
0
5
PC1
candidate
Donald Trump
Hillary Clinton



For PC1, based on the plot above, large negative values of PC1 are caused by high values for `Employed`, `Income`, `IncomePerCap`, `Professional` and `White`. Large positive values of PC1 are caused by high values for `ChildPoverty`, `Poverty`, and `Unemployment`.

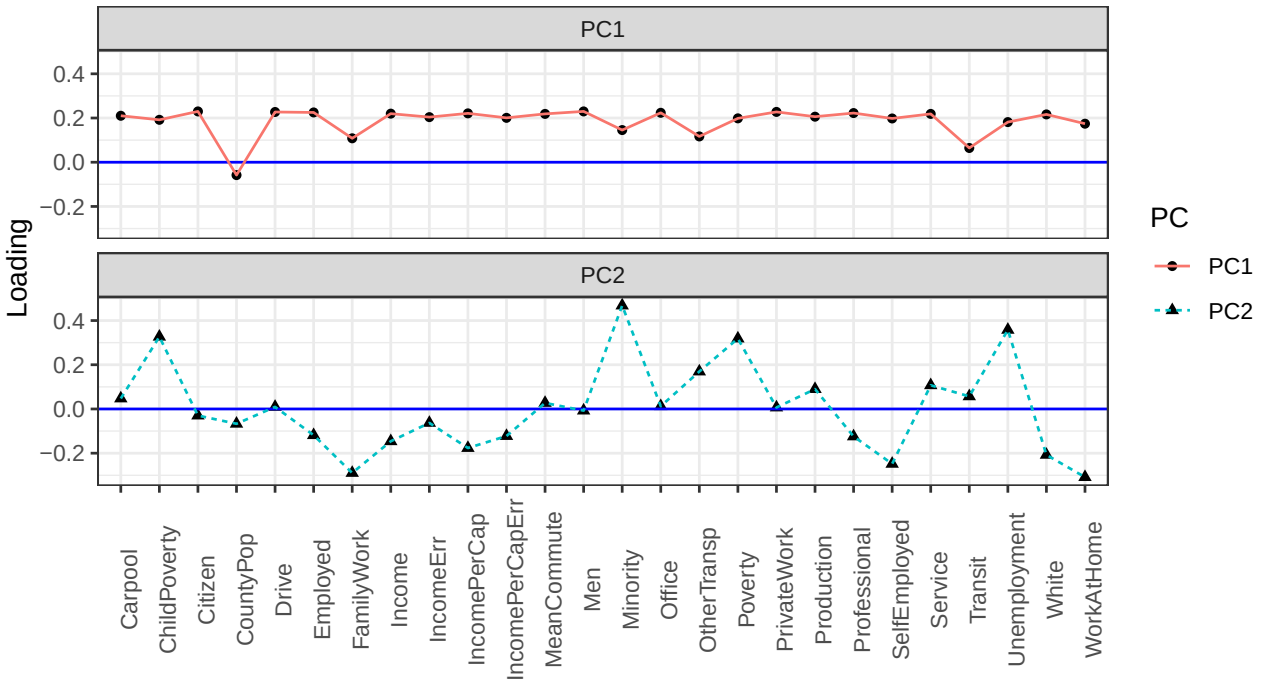
This combination of significant variables seems to indicate that PC1 is describing counties that have high levels of poverty and unemployment. These counties have thus a lower income per capita and in household, lower percentage of Whites, and lower demand for jobs in professional industry.

For PC2, based on the plot above, large negative values of PC2 are caused by high values for `SelfEmployed`, `FamilyWork`, `Citizen`, `WorkAtHome` and `White`. Large positive values of PC2 are caused by high values for `IncomeErr`,`Minority`, `Office`, `TotalPop`, and `Transit`.

This means that PC2 is describing counties with income variability, high minority presence, and a working community that is traveling a lot with a high population area, most likely urban because of high commute rates. These counties also have less people that are self-employed or work at home, less Whites, less family businesses, and low percentage of citizens.

Here is a scatterplot showing the data projected onto the first two PC's by candidate at the sub-county level:


Scatter plot of PC1 and PC2 for SubCounty Data
25
0
−25
PC2
0
20
40
PC1
candidate
Donald Trump
Hillary Clinton



For PC1, based on the plot above, there are not any covariates that associated with large negative values of PC1. Large positive values of PC1 are caused by high values for all covariates expect for `CountyPop`,`FamilyWork`, and `Transit`.

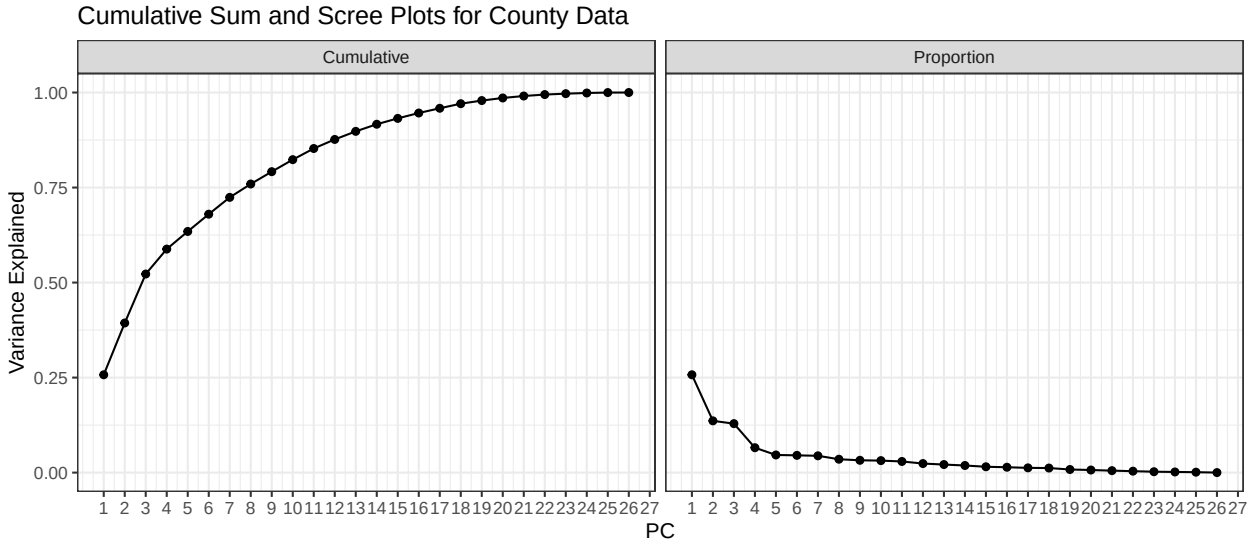
Since all of these other variables are represented, it measn that PC1 is mostly showing the average of the covariates not aforementioned. Since these are a wide range of demographics that can't be generalized, it is interpreted as the average of census data at the sub-county level.

For PC2, based on the plot above, large negative values of PC2 are caused by high values for `SelfEmployed`, `FamilyWork`, and `WorkAtHome`. Large positive values of PC2 are caused by high values for `ChildPoverty`,`Minority`, and `Unemployment`.

This means that PC2 is describing sub-counties with with high levels of poverty and unemployment, so a poor county-level economy. These sub-counties also have a high minority presence and there's a lot of reliance on other employers that are not remote via technology and less emphasis on family business.

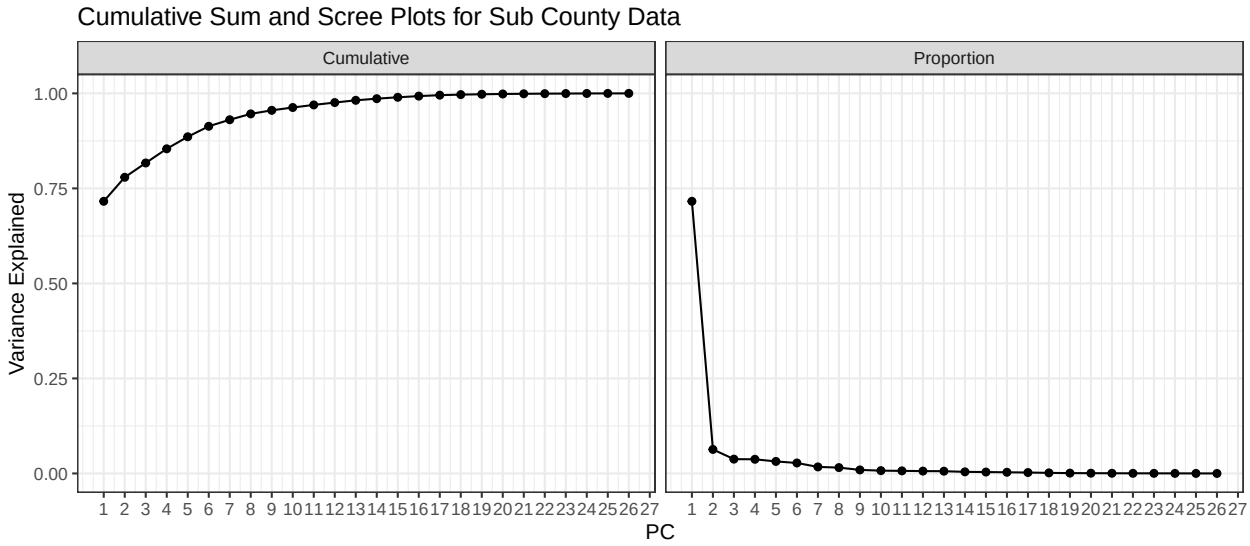
Below are the plots of the proportion of variance explained and cumulative variance explained for both county and sub-county analyses.

At the county level:



It looks like 90% of the variance here can be explained by 14-15 principal components, based on the left plot. Looking at the scree plot on the right, the elbow is right around 2 PC's.

At the sub-county level:



It looks like 90% of the variance here can be explained by 5-6 principal components, based on the left plot. Looking at the scree plot on the right, the elbow is right around 2 PC's as well.

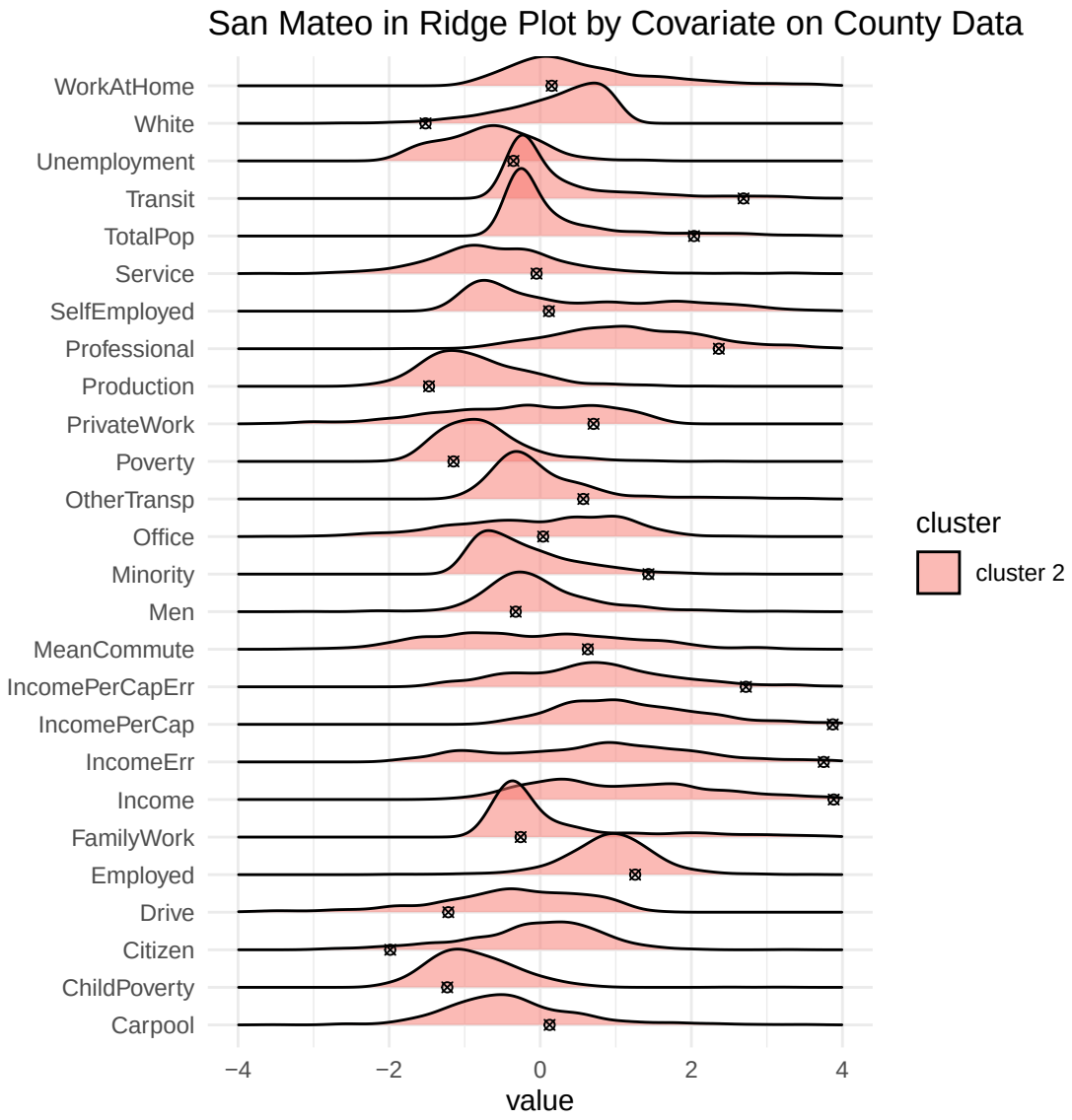
Now, we will perform hierarchical clustering with complete linkage, on all data or the first 5 principal components of the county-level data. After comparing both ways, the approach putting San Mateo County in a more appropriate cluster will be analyzed.

For the full data after standardizing the data features:

```
##
## Attaching package: 'cluster'

## The following object is masked from 'package:maps':
##
##     votes.repub
```

County	cluster
San Mateo	cluster 2

Now since it placed in cluster 2, we will now look at the ridge plots to visualize the county.

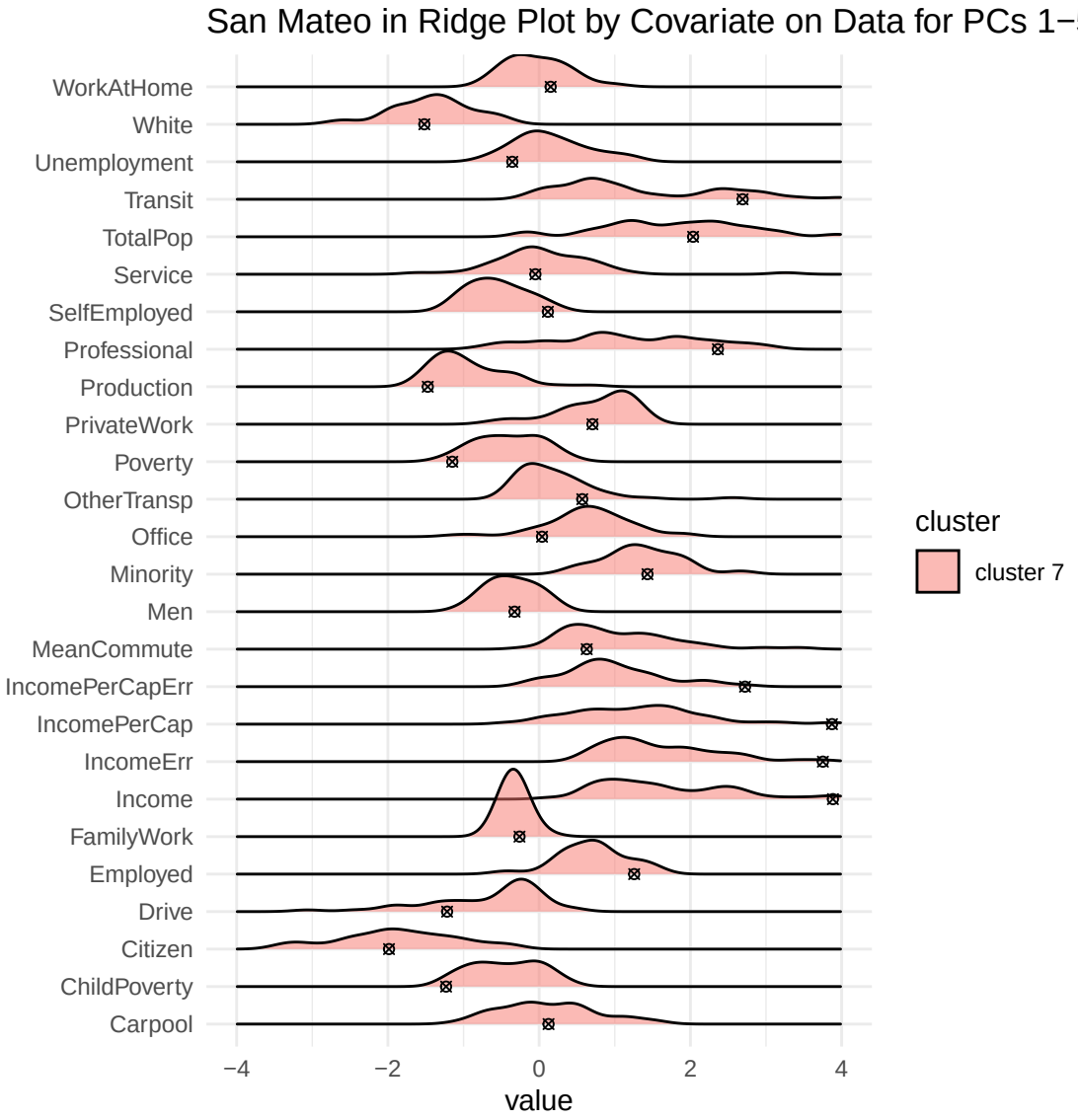


It looks the the ridges are not very concentrated, there is a lot of variance that you can see, it is not as compact or robust.

Now, consider the first 5 PC's:

County	cluster
San Mateo	cluster 7

The PC 1 through 5 place it in cluster 7, so here are the ridge plots:



These ridge plots have a a smaller and narrower, concentrated distribution, so the algorithm seems better. The trade-off of this algorithm is that we lose the valuable data provided by the other PC's.

Now, let's look at a quantitative measure of which method is better:

Data	Cluster	Mean	Variance
Original Features	2	5.734	10.14
PC Loadings 1-5	7	4.269	3.365

Since the variance is much lower with the PC's 1-5, it seems to be the better algorithm because low variance means the observations are similar, and this would be better to place San Mateo in a more appropriate cluster.

Let's look at the cluster count:

Cluster for Full Data	Count	Cluster for PCs 1-5	Count
cluster 1	2632	cluster 1	1919

Cluster for Full Data	Count	Cluster for PCs 1-5	Count
cluster 2	501	cluster 2	905
cluster 3	6	cluster 3	163
cluster 4	7	cluster 4	86
cluster 5	5	cluster 5	8
cluster 6	1	cluster 6	27
cluster 7	11	cluster 7	47
cluster 8	13	cluster 8	8
cluster 9	38	cluster 9	18
cluster 10	4	cluster 10	37

Again, since the clusters are less balanced with the original full features, the PC's 1-5 algorithm seems to work best.

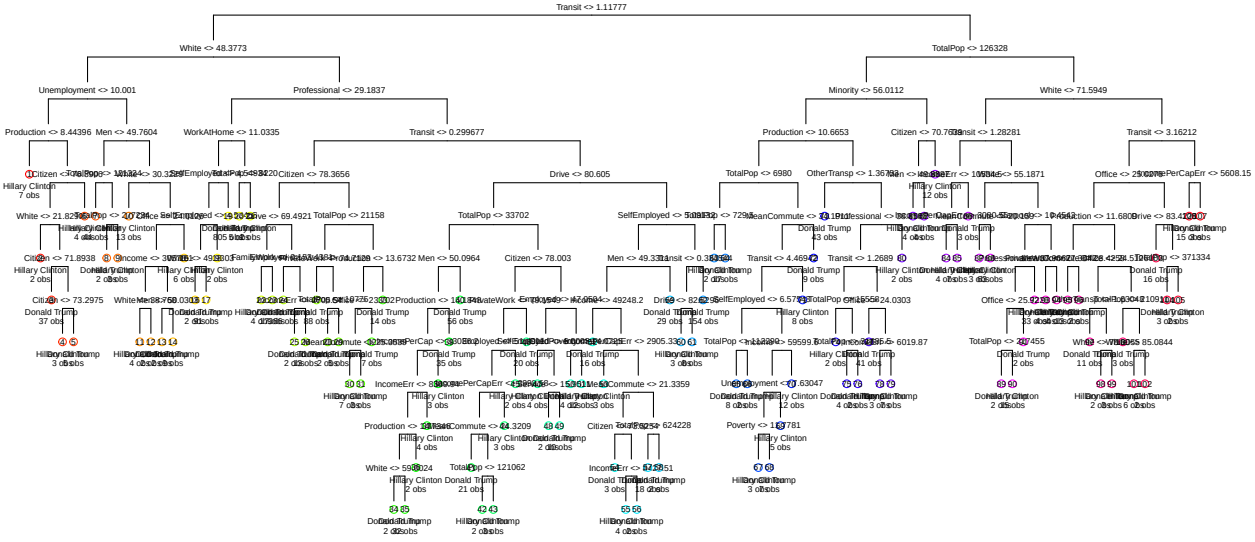
Now we will looking into classification methods.

80% training and 20% testing partitions were used.

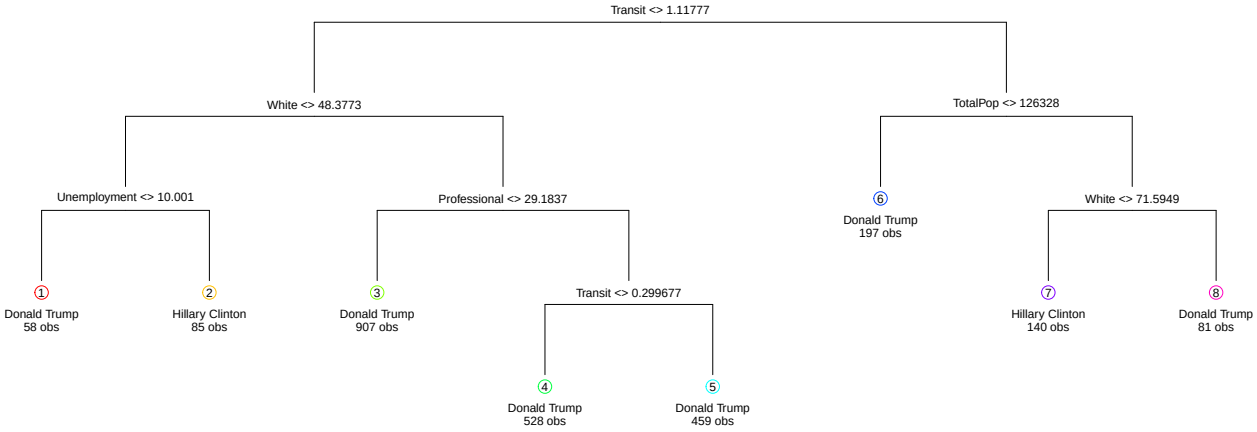
Here is the decision tree before pruning:

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
## Loading required package: rpart
```



Here is the tree after pruning:



Based on the pruned tree, people who are in a highly urbanized area because of high commute rates and large population and those that are mostly white, tend to vote for Donald Trump. But of these that are mostly minorities, seem to vote for Hillary Clinton. People with low commute rates, high percentage in the professional industry of jobs, and mostly white, tend to vote for Donald Trump solely. People in low commute rate and high unemployment areas with more minorities tend to vote for Hillary Clinton.

Let's looking at the misclassification error rates of the pruned tree:

	Donald Trump	Hillary Clinton
Donald Trump	506	10
Hillary Clinton	44	55

	Donald Trump	Hillary Clinton
Donald Trump	0.9806	0.01938
Hillary Clinton	0.4444	0.5556

Decision_Tree_Misclassification_Error
0.0878

It seems that predicting Trump winning is very good but predicting Hillary Clinton winning is very poor, so the error rates as seen in the percentage-based table, are not balanced with this method so far. We need to adjust the threshold to improve the balance with ROC curves.

Now we will looking at the logistic regression model as a classifier.

	Donald Trump	Hillary Clinton
Donald Trump	502	14
Hillary Clinton	18	81

-

Logistic_Regression_Misclassification_Error
0.05203

The misclassification error is much lower and more balanced out with logistic regression.

Here are the variables that are significant based on the logistic regression model are highlighted below:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.53	8.894	-1.859	0.06302
TotalPop	2.926e-07	3.808e-07	0.7684	0.4423
Men	0.0533	0.04931	1.081	0.2797
White	-0.1456	0.06159	-2.364	0.01806
Citizen	0.0915	0.02761	3.314	0.0009199
Income	-8.843e-05	2.631e-05	-3.361	0.000778
IncomeErr	-1.433e-05	6.393e-05	-0.2241	0.8227
IncomePerCap	0.0002617	6.468e-05	4.046	5.202e-05
IncomePerCapErr	-0.0003025	0.0001308	-2.313	0.02073
Poverty	0.01975	0.04093	0.4826	0.6294
ChildPoverty	-0.005151	0.02516	-0.2047	0.8378
Professional	0.2302	0.03607	6.382	1.75e-10
Service	0.3095	0.04408	7.021	2.205e-12
Office	0.05506	0.04236	1.3	0.1937
Production	0.1383	0.03866	3.579	0.0003456
Drive	-0.1936	0.04455	-4.346	1.386e-05
Carpool	-0.1876	0.0575	-3.262	0.001106
Transit	0.1596	0.08968	1.78	0.0751
OtherTransp	-0.09536	0.09237	-1.032	0.3019
WorkAtHome	-0.1174	0.07194	-1.632	0.1026
MeanCommute	0.04959	0.02379	2.085	0.03709
Employed	0.1721	0.0335	5.138	2.782e-07
PrivateWork	0.08214	0.02221	3.698	0.0002173
SelfEmployed	-0.007998	0.04722	-0.1694	0.8655
FamilyWork	-1.028	0.4077	-2.522	0.01168
Unemployment	0.1803	0.04129	4.366	1.263e-05
Minority	-0.01977	0.05945	-0.3326	0.7394

So to interpret the meaning of these variables, for example, if all other variables are fixed, a percentage increase of people with a job in the production industry is associated with an increase in the odds by a factor of 0.1383. If all other are variables are fixed and the a dollar increase in median household income is associated with a decrease in the odds by a factor of 8.843e-05.

Here are the variables that were used to construct the pruned tree:

var
White
Transit
TotalPop
Professional
Unemployment

There is not exactly a clear match with all the variables used to construct the tree versus the ones that are significant in the logistic regression. There are some repeating variables however, but far more variables are used for the logistic regression model.

Let's look at predicted vs actual results in our counties forementioned.

True	Prediction
Hillary Clinton	Hillary Clinton

The fit.glm model predicted the result of Hillary Clinton winning correctly in Santa Clara County. It predicted correctly for Los Angeles County as well:

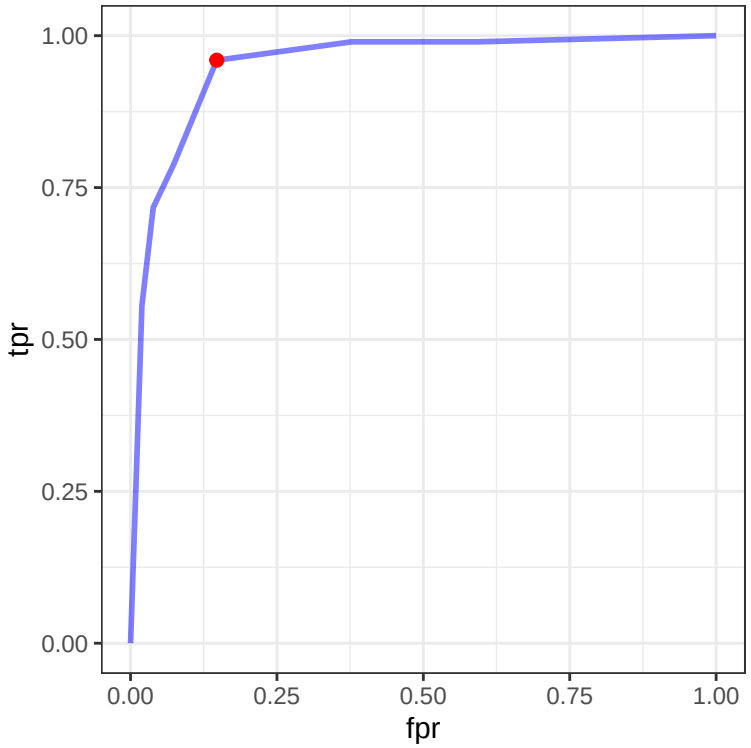
True	Prediction
Hillary Clinton	Hillary Clinton

And for orange county as well, it predicted correctly:

True	Prediction
Hillary Clinton	Hillary Clinton

Now, let's look at the ROC curves and adjusted rate for decision tree with the optimal threshold.

fpr	tpr	thresh	youden
0.1473	0.9596	0.2589	0.8123



	Donald Trump	Hillary Clinton
Donald Trump	478	38
Hillary Clinton	21	78

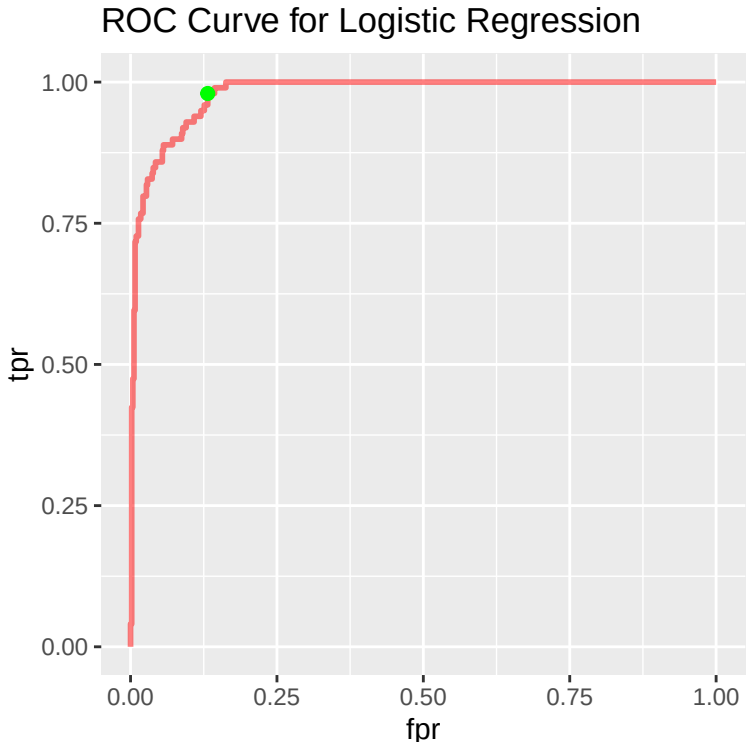
	Donald Trump	Hillary Clinton
Donald Trump	0.9264	0.07364
Hillary Clinton	0.2121	0.7879

Adjusted_Decision_Tree_Misclassification_Error
0.09593

As we learned, adjusting to the optimal threshold balances out the TPR and TNR. In other words, the benefit is that predicting Hillary Clinton results correctly improved, but predicting Donald Trump winning was poorer. As can be seen with the table of percentages, the error rates are more balanced than before. The overall misclassification error rate is higher but it is more balanced, better at predicting Clinton correctly.

Now let's look at adjusting logistic regression model.

fpr	tpr	thresh	youden
0.1318	0.9798	0.1471	0.848

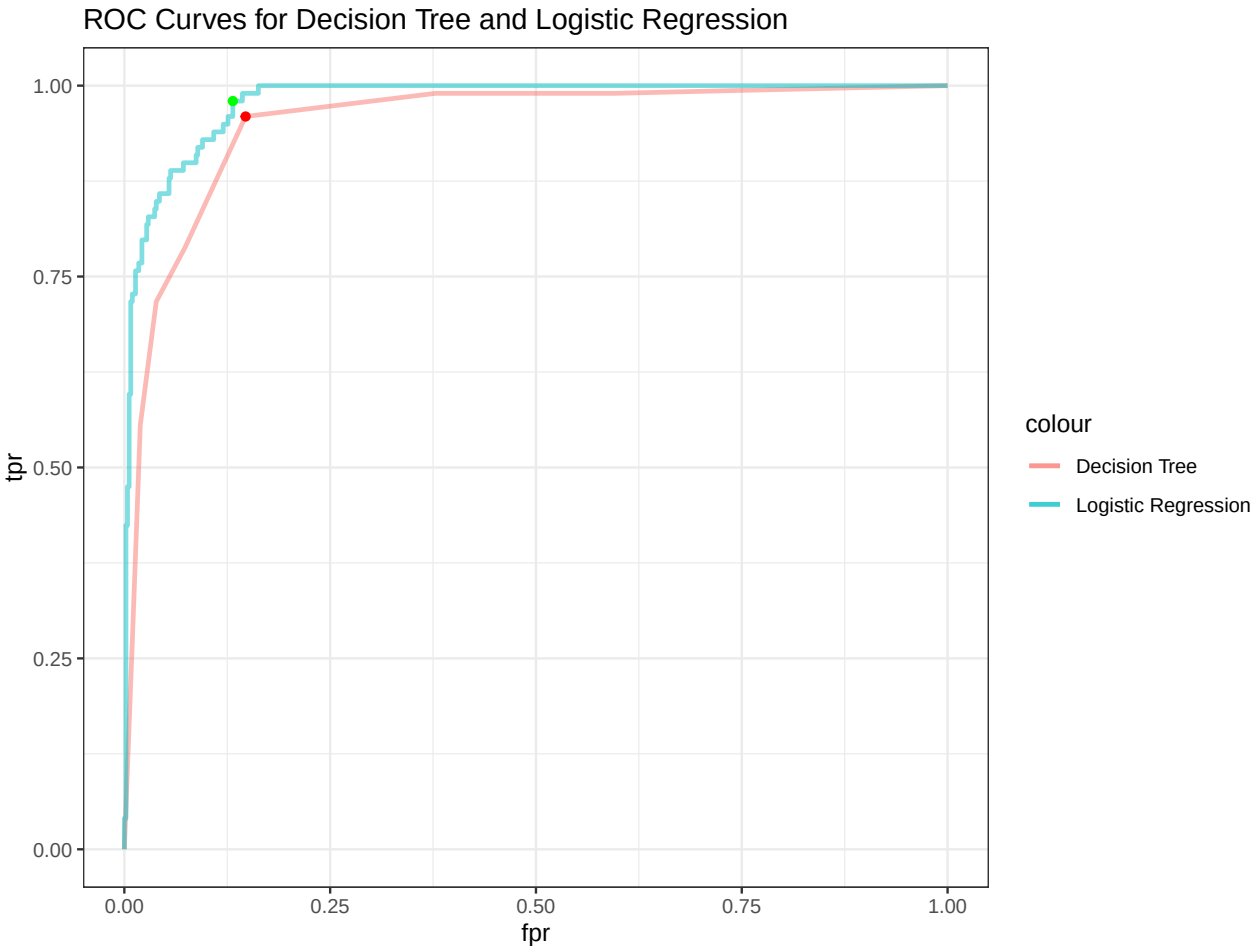


	Donald Trump	Hillary Clinton
Donald Trump	448	68
Hillary Clinton	3	96

Adjusted_Logistic_Regression_Misclassification_Error
0.1154

Again, the overall misclassification error rate is high, but predicting Hillary Clinton is migh higher, at the cost of lowering predicting ability for Donald Trump correctly.

Here are the ROC curves on the same plot:



The curve for logistic regression seems to constantly stay above the curve for the decision tree on the y-axis. This means that the TPR is maximized with logistic regression over decision tree, so logistic regression model seems to be better at discerning between Clinton and Trump for the election results.

For decision tree, the pros are that it is applicable to both regression and classification settings. (don't think is needed in our case) and no restrictions on variable types. It could fail if classification pattern of separation in the feature space is hard to approximate by rectangular regions.

Logistic regression is interpretable, works for categorical and continuous features, works in high dimensions and with noise variables but the cons are that there are big assumptions and mathematically more complicated in multiple class case.

This is why the logistic regression model is preferred with election data between Clinton and Trump and it is more interpretable and decision boundary is better capture nonlinearly instead of in rectangular regions and the more covariates used to the already large wide range of covariates.

Further Analysis

We are going to use a linear regression model to predict the `total` vote for each candidate by county.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23866	2288	10.43	4.574e-25
candidateHillary Clinton	130834	5884	22.24	1.122e-101

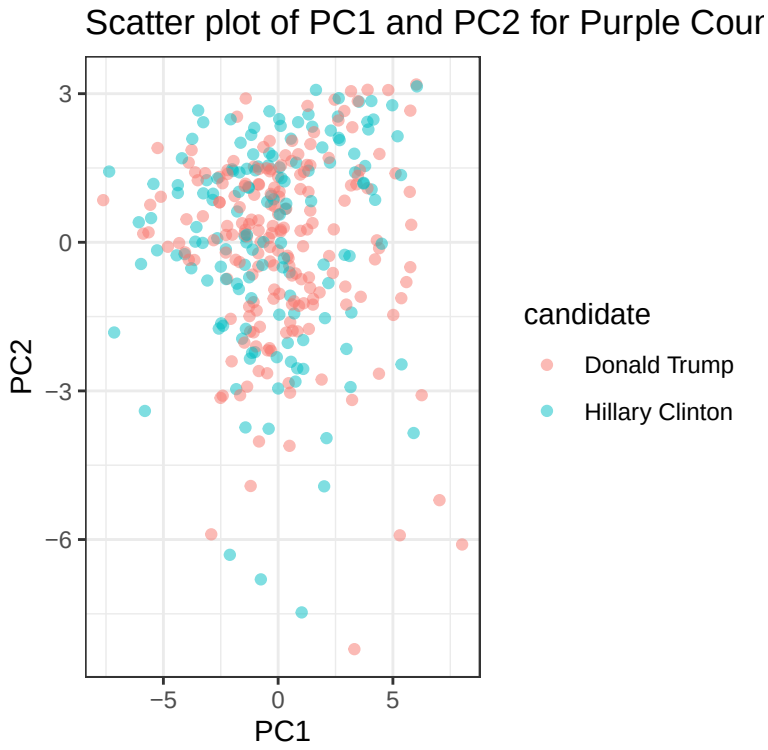
Table 34: Fitting linear model: total ~ candidate

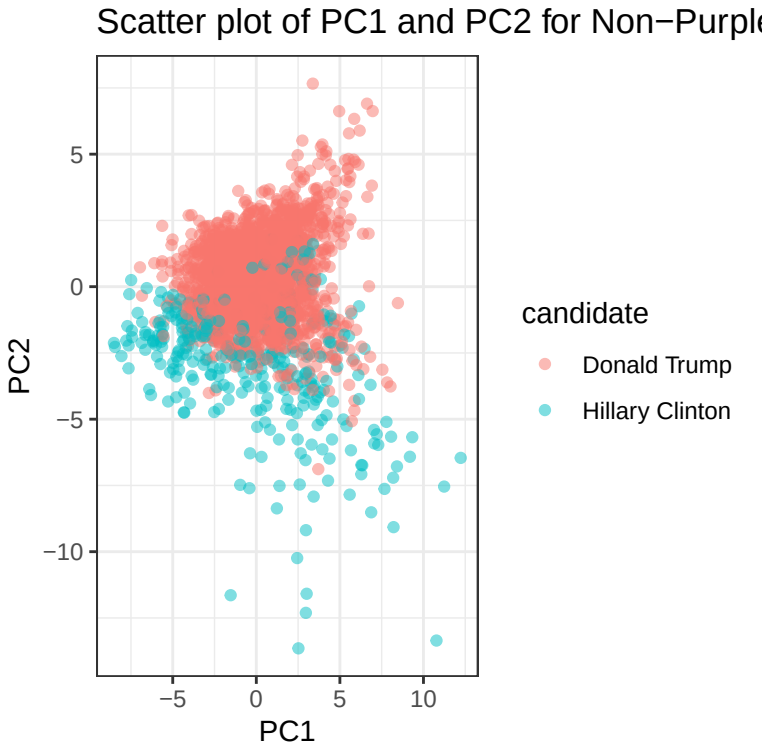
Observations	Residual Std. Error	R^2	Adjusted R^2
3070	116778	0.1388	0.1385

1.363e+10 Since the MSE is very high and the adjusted R-squared is very low, it seems the linear regression is not a very good way the other classification methods are better.

We are going to analyze purple counties. By carrying principal component analysis on this data, we are going to see what features are important to the PCs. Based on this, we are going to identify how hard these variables are to measure in the population, likely leading to tough prediction problems that is seen with purple counties.

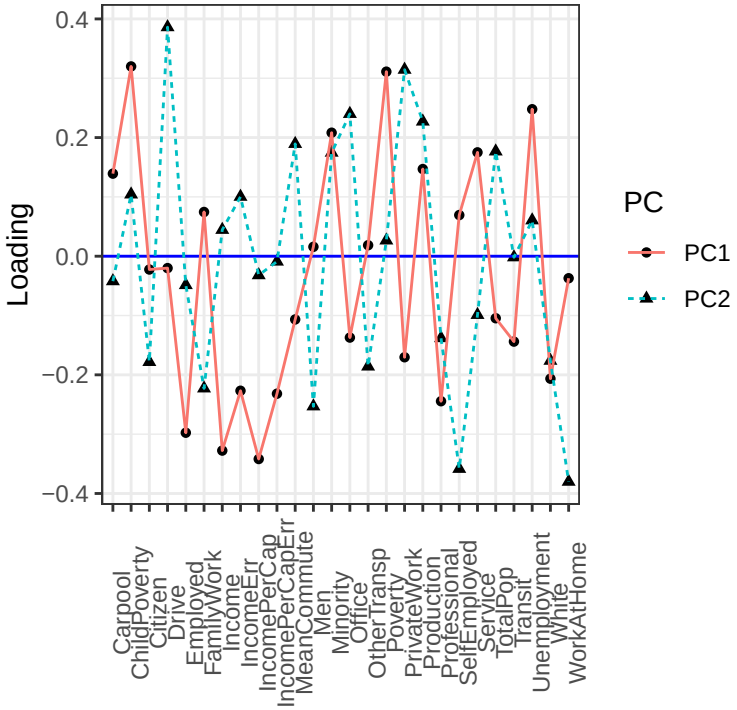
We are going to pick a +/- 3% margin as defining purple counties.





As can be seen above, the scatter plot of the non-purple counties can be discerned on the y-axis, but the purple counties scatter plot cannot be discerned on any axis, this is why it's very tough to classify these counties.

Here is the PC loading plot:



Conclusion

Some features that could be considered to improve principal component analysis on purple counties is including but not limited to age group, education level, and family status (married, number of children). We think it would also be helpful in trying to determining nonvoting vs voting behavior. We can use decision trees to see with predictors if the behavior can be discerned. Andy by using random forests, the predicted probabilities are determined without overfitting the tree applying the algorithm. We can alternatively use machine learning, that is, neural networks to be able to predict this behavior. Taking this a step further, if we have non_voting and voting in past elections as another covariate, we maybe able to discern for which candidate a person might vote for. This is yet another way of incorporating past election data and applying a more of a life-course approach to predicting election behavior that is not limited to that year only.