PSTAT 131: Final project

Emanuel Medina: 5095906, Razeen Ahmed: 6941736, Cameron Joe: 6114367

Instructions and Expectations

- You are allowed and encouraged to work with two partners on this project. Include your names, perm numbers, and whether you are taking the class for 131 or 231 credit.
- You are welcome to write up a project report in a research paper format – abstract, introduction, methods, results, discussion – as long as you address each of the prompts below. Alternatively, you can use the assignment handout as a template and address each prompt in sequence, much as you would for a homework assignment.
- There should be no raw R *output* in the body of your report! All of your results should be formatted in a professional and visually appealing manner. That means that visualizations should be polished – aesthetically clean, labeled clearly, and sized appropriately within the document you submit, tables should be nicely formatted (see `pander`, `xtable`, and `kable` packages). If you feel you must include raw R output, this should be included in an appendix, not the main body of the document you submit.
- There should be no R *codes* in the body of your report! Use the global chunk option `echo=FALSE` to exclude code from appearing in your document. If you feel it is important to include your codes, they can be put in an appendix.

Background

The U.S. presidential election in 2012 did not come as a surprise. Some correctly predicted the outcome of the election correctly including Nate Silver, and many speculated about his approach.

Despite the success in 2012, the 2016 presidential election came as a big surprise to many, and it underscored that predicting voter behavior is complicated for many reasons despite the tremendous effort in collecting, analyzing, and understanding many available datasets.

Your final project will be to merge census data with 2016 voting data to analyze the election outcome.

To familiarize yourself with the general problem of predicting election outcomes, read the articles linked above and answer the following questions. Limit your responses to one paragraph for each.

1. What makes voter behavior prediction (and thus election forecasting) a hard problem?

Voter behavior prediction (and election forecasting) is a difficult problem because voting intention changes over time. Furthermore, polls tend to vary as a result of multiple factors such as sampling error and bias. As a result, it is extremely difficult to accurately predict voter behavior and election results as a whole.

2. What was unique to Nate Silver's approach in 2012 that allowed him to achieve good predictions?

Silver used hierarchical modeling that accounts for voting behavior on both the national and the state level to achieve highly accurate predictions. In addition, instead of using the most likely voter support percentage as an estimate of actual vote percentages, Silver calculated the probabilities for a range of voter support. Then, using these values, Baye's Theorem, and graph theory, he was able to calculate new likelihoods for a range of support levels.

3. What went wrong in 2016? What do you think should be done to make future predictions better?

Polls often try to bolster sample size and reduce possible bias and variance by looking to other polls for information. In the case of the 2016 election, small (potentially unavoidable) systematic biases became amplified by this inter-poll reliance which likely caused a much larger trend where polls were systematically off in the same direction - overestimating Hillary's lead over Trump. This could be due to multiple key oversights; it is possible that Trump supporters were reluctant to express their true opinions or distrusted the polls themselves. The unexpected results may also be attributed to lower-than-expected voter turn out or last minute shifts in voter opinions.

One possible way to make polling predictions better may be to reduce the amount of inter-poll reliance. The final paragraph of Bialik and Enten's article expressed a sentiment that polls seem to be incentivized to converge on one opinion or definitive value. Relying less on other polls may introduce bias and variance, but it avoids the risk repeating the mistakes of past polls only to come to a similar conclusion (which isn't that informative at all).

Data

The `project_data.RData` binary file contains three datasets: tract-level 2010 census data, stored as `census`; metadata `census_meta` with variable descriptions and types; and county-level vote tallies from the 2016 election, stored as `election_raw`.

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.

Nationwide and statewide tallies are included as rows in `election_raw` with `county` values of `NA`. There are two kinds of these summary rows:

- Federal-level summary rows have a `fips` value of `US`.
- State-level summary rows have the state name as the `fips` value.

4. Inspect rows with `fips=2000`. Provide a reason for excluding them. Drop these observations – please write over `election_raw` – and report the data dimensions after removal.

Table 2: Observation with 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

```
## [1] 18345     5
```

Observations with a fips of 2000 have no county information on them. In addition, they do not appear to be federal-level data (as their fips value is not "US") and they do not appear to be state-level data (as their fips value is not the state name "AK"). So, observations with a fips value of 2000 are anomylous as they are neither county, state, or federal data. Removing these observations results in an object of dimensions 18345 x 5.

Census data

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

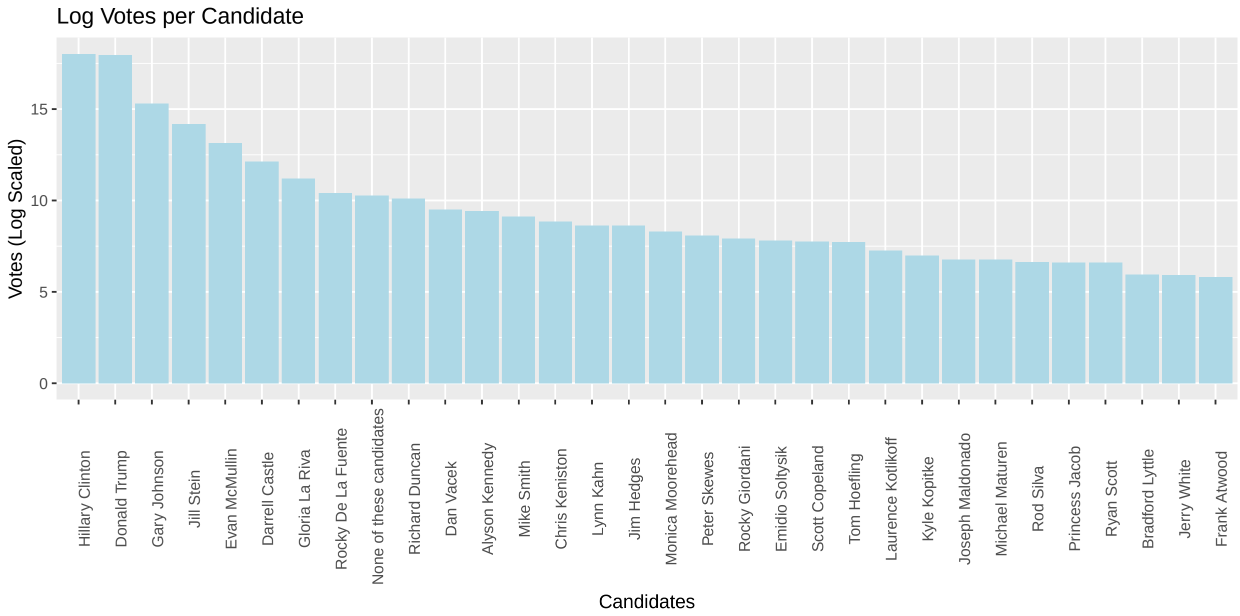
CensusTract	State	County	TotalPop	Men	Women
1001020100	Alabama	Autauga	1948	940	1008
1001020200	Alabama	Autauga	2156	1059	1097
1001020300	Alabama	Autauga	2968	1364	1604
1001020400	Alabama	Autauga	4423	2172	2251
1001020500	Alabama	Autauga	10763	4922	5841
1001020600	Alabama	Autauga	3851	1787	2064

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

Data preprocessing

5. Separate the rows of `election_raw` into separate federal-, state-, and county-level data frames:
   - Store federal-level tallies as `election_federal`.
   - Store state-level tallies as `election_state`.
   - Store county-level tallies as `election`. Coerce the `fips` variable to numeric.
6. How many named presidential candidates were there in the 2016 election? Draw a bar graph of all votes received by each candidate, and order the candidate names by decreasing vote counts. (You may need to log-transform the vote axis.)

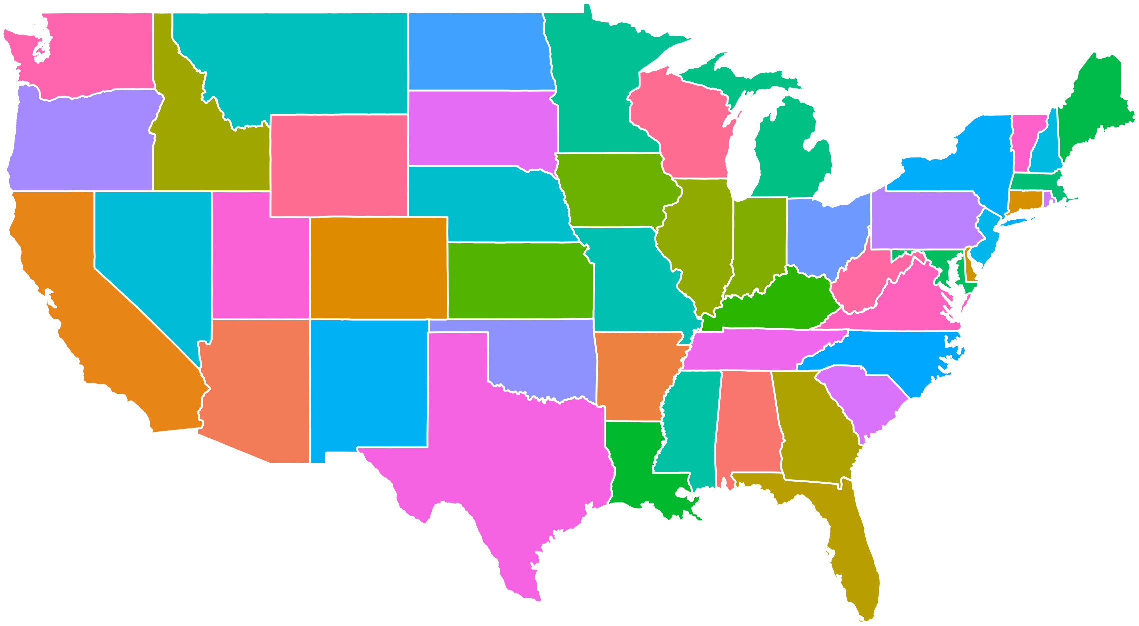


In total, there were 32 named presidential election candidates in the 2016 race.

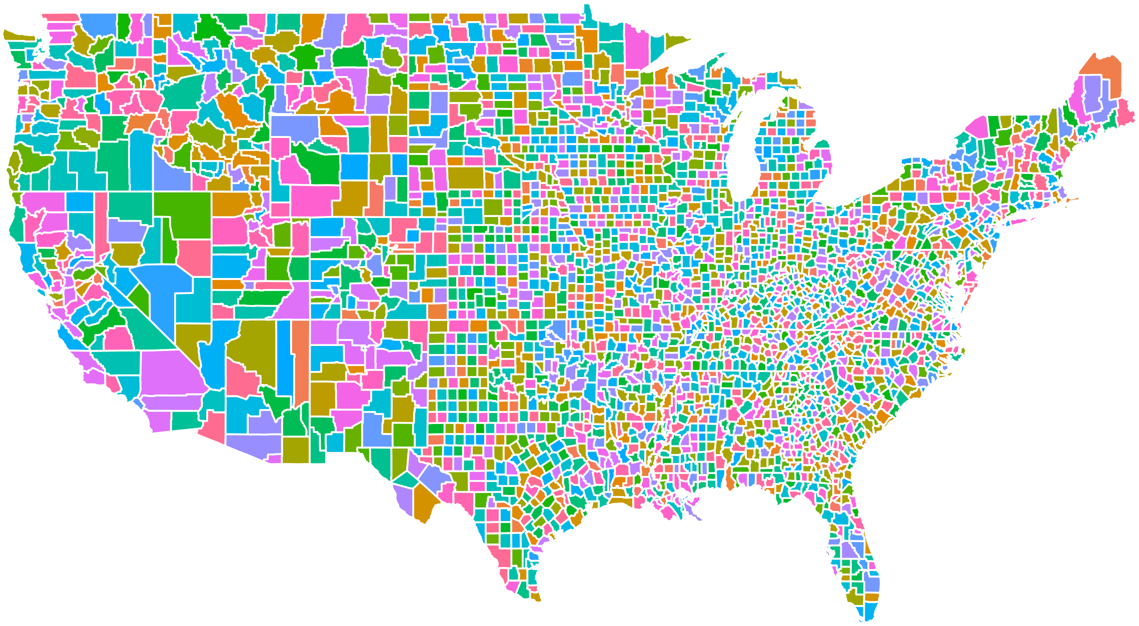
7. Create `county_winner` and `state_winner` by taking the candidate with the highest proportion of votes. (Hint: to create `county_winner`, start with `election`, group by `fips`, compute `total` votes, and `pct = votes/total`. Then choose the highest row using `slice_max` (variable `state_winner` is similar).)

Visualization

Here you'll generate maps of the election data using `ggmap`. The .Rmd file for this document contains codes to generate the following map.



8. Draw a county-level map with `map_data("county")` and color by county.



In order to map the winning candidate for each state, the map data (`states`) must be merged with with the election data (`state_winner`).

The function `left_join()` will do the trick, but needs to join the data frames on a variable with values that match. In this case, that variable is the state name, but abbreviations are used in one data frame and the full name is used in the other.

9. Use the following function to create a `fips` variable in the `states` data frame with values that match the `fips` variable in `election_federal`.

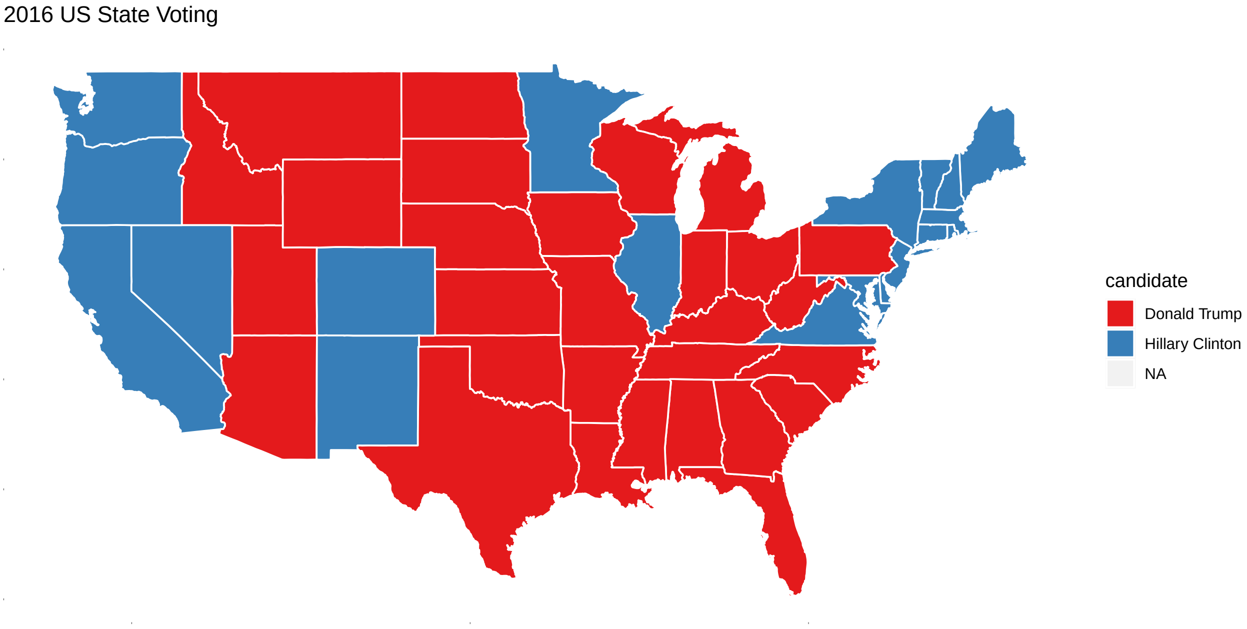
```
name2abb <- function(statename){
  ix <- match(statename, tolower(state.name))
  out <- state.abb[ix]
  return(out)
}

states <- states %>%
  mutate(fips = name2abb(region))
```

Now the data frames can be merged. `left_join(df1, df2)` takes all the rows from `df1` and looks for matches in `df2`. For each match, `left_join()` appends the data from the second table to the matching row in the first; if no matching value is found, it adds missing values.

10. Use `left_join` to merge the tables and use the result to create a map of the election results by state. Your figure will look similar to this state level New York Times map. (Hint: use `scale_fill_brewer(palette="Set1")` for a red-and-blue map.)

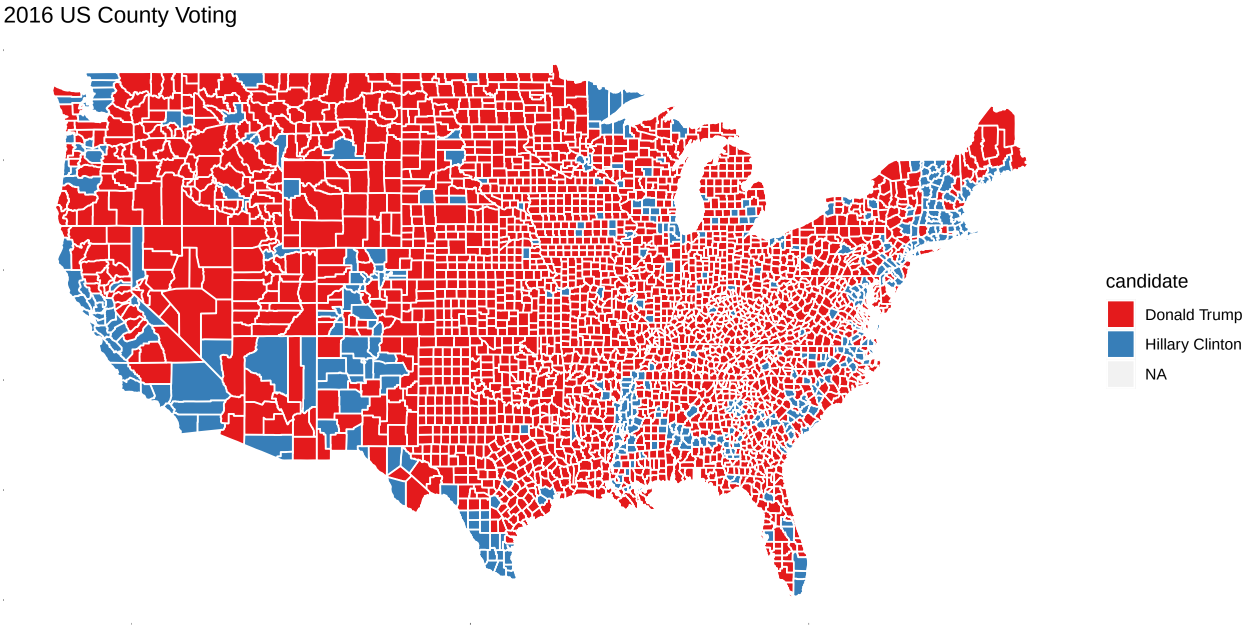
```
## Joining, by = "fips"
```



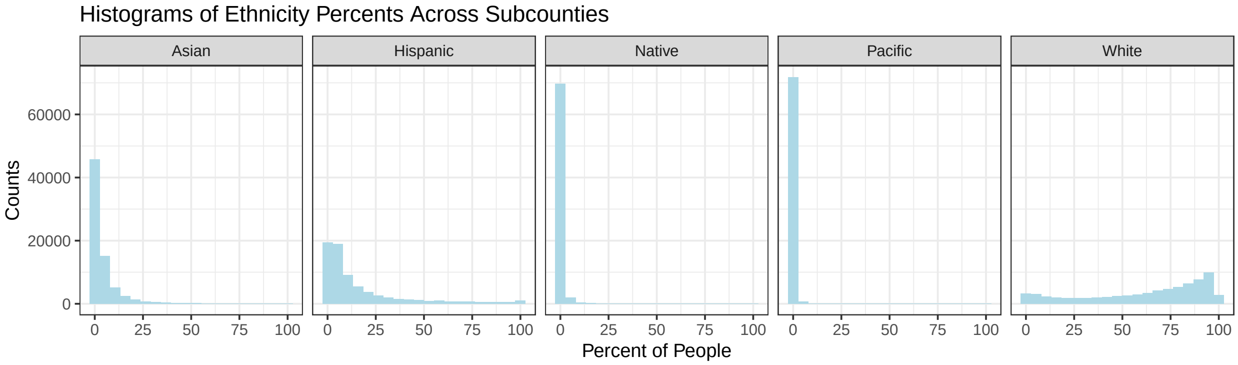
11. Now create a county-level map. The county-level map data does not have a `fips` value, so to create one, use information from `maps::county.fips`: split the `polyname` column to `region` and `subregion` using `tidyr::separate`, and use `left_join()` to combine `county.fips` with the county-level map data. Then construct the map. Your figure will look similar to county-level New York Times map.

```
## Joining, by = c("region", "subregion")
```

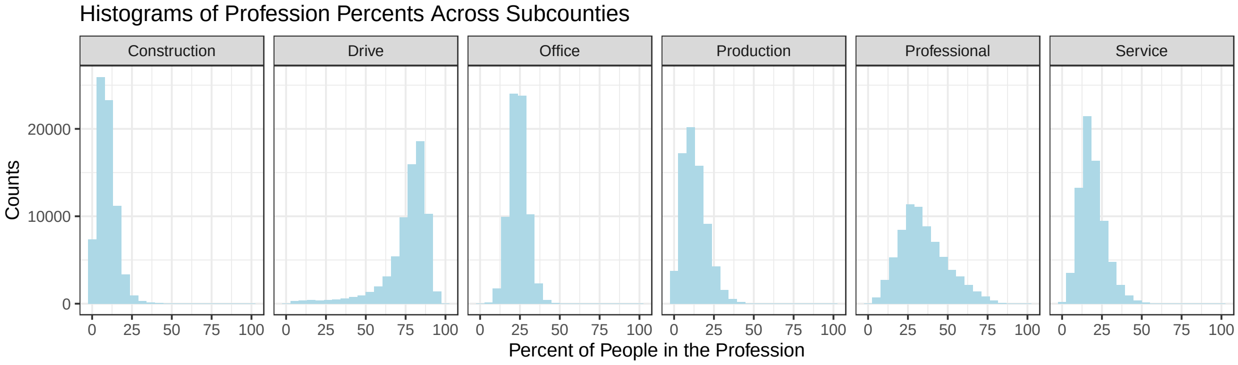
```
## Joining, by = "fips"
```



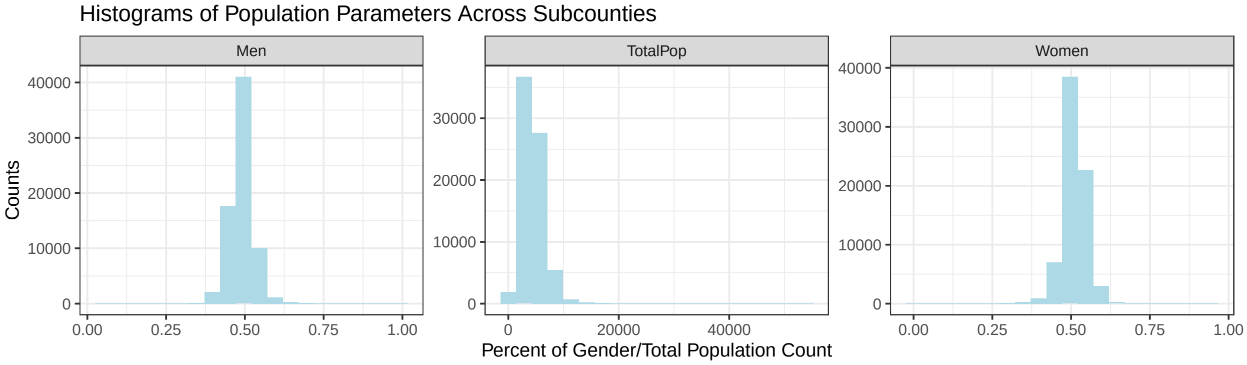
12. Create a visualization of your choice using `census` data. Many exit polls noted that demographics played a big role in the election. If you need a starting point, use this Washington Post article and this R graph gallery for ideas and inspiration.



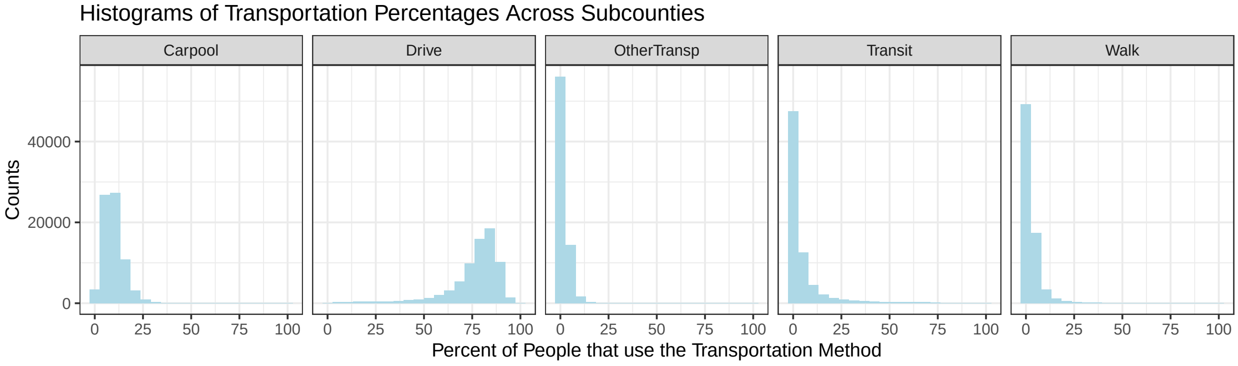
Looking at the histograms, it can be seen that those of Pacific islander and native american ethnicity often make up less than 10% of the population. Asian and Hispanic ethnicities make up a larger percent of the population more often - usually within the range of 0 - 25%. White people consist of a majority most often. Frequently exceeding 75% of the population and sometimes even close to 100% of the population.



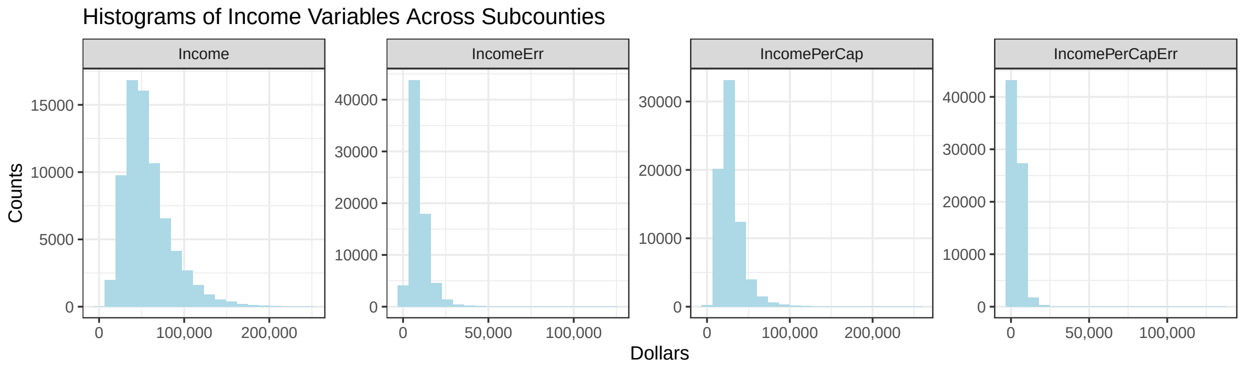
Construction, service, and production related jobs seem to consist of around 10% of the population jobs. Professional and office jobs are most commonly observed to consist of around 25% of the population jobs. Surprisingly, driving jobs are most commonly observed at consisting of 80% of population jobs.



Both men and women are most commonly seen to consist of 50% of the population (with men more commonly observed as being under 50% and women more commonly observed as being over 50%).



Walking, other transportation, and transit are most commonly observed at consisting of around 0% of the population transportation. Carpool appears to be seen at higher percentages more frequently, with its mode being at around 10%. Driving is observed at higher frequencies most often with its mode being at around 80%.



The most frequently observed average income is seen to be around 50,000.

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`.
- Create population weights for sub-county census data, saving the result as `census_subct`:
  - group `census_del` by `State` and `County`;
  - use `add_tally()` to compute `CountyPop`;
  - compute the population weight as `TotalPop/CountyTotal`;
  - adjust all quantitative variables by multiplying by the population weights.
- Aggregate census data to county level, `census_ct`: group the sub-county data `census_subct` by state and county and compute popluation-weighted averages of each variable by taking the sum (since the variables were already transformed by the population weights)
- Print the first few rows and columns of `census_ct`.

```
## `summarise()` regrouping output by 'State' (override with `.groups` argument)
```

Table 5: First Few Rows and Columns of Census_ct

State	County	Men	Women	White
Alabama	Autauga	48.4327	3348.81	75.7882
Alabama	Baldwin	48.8487	3934.17	83.1026
Alabama	Barbour	53.8282	1491.94	46.2316
Alabama	Bibb	53.4109	2930.11	74.4999
Alabama	Blount	49.4056	3562.08	87.8539
Alabama	Bullock	53.0062	1968.03	22.1992

14. If you were physically located in the United States on election day for the 2016 presidential election, what state and county were you in? Compare and contrast the results and demographic information for this county with the state it is located in. If you were not in the United States on election day, select any county. Do you find anything unusual or surprising? If so, explain; if not, explain why not.

```
## `summarise()` regrouping output by 'State' (override with `.groups` argument)
## `summarise()` ungrouping output (override with `.groups` argument)
```

Table 6: California Census Information (continued below)

State	Men	Women	White	Citizen	Income	IncomeErr
California	49.54	2875	38.72	63.07	67908	10908

Table 7: Table continues below

IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty	Professional
30391	4110	16.36	20.58	35.65

Table 8: Table continues below

Service	Office	Production	Drive	Carpool	Transit	OtherTransp
19.19	23.73	11.63	73.64	11.17	4.949	2.465

Table 9: Table continues below

WorkAtHome	MeanCommute	Employed	PrivateWork	SelfEmployed	FamilyWork
5.172	28.03	45.07	77.58	8.199	0.1773

Unemployment	Minority	StatePop
10.16	58.28	38221472

Table 11: San Diego Census Information (continued below)

State	County	Men	Women	White	Citizen	Income	IncomeErr
California	San Diego	50.12	3221	46.96	66.32	69943	10850

Table 12: Table continues below

IncomePerCap	IncomePerCapErr	Poverty	ChildPoverty	Professional
31282	3983	14.48	17.13	39.26

Table 13: Table continues below

Service	Office	Production	Drive	Carpool	Transit	OtherTransp
20.13	23.69	8.689	76.55	9.594	3.108	1.923

Table 14: Table continues below

WorkAtHome	MeanCommute	Employed	PrivateWork	SelfEmployed	FamilyWork
6.315	25.29	45.52	76.86	7.666	0.1605

Unemployment	Minority
8.948	49.75

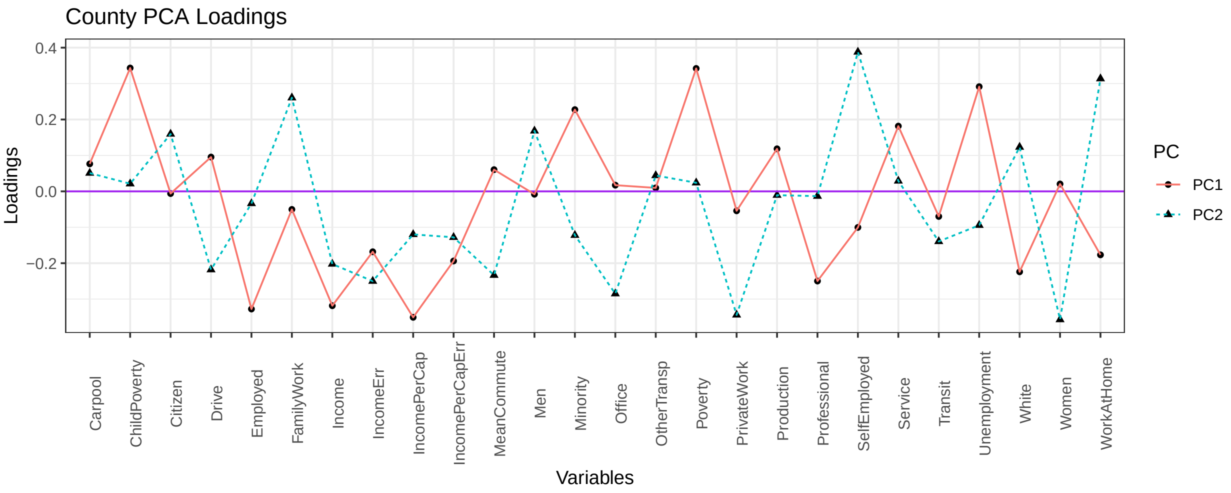
For the 2016 election, one of the group members was in California, specifically the county of San Diego. It is surprising to see that only 66% percent of the population were citizens at the time, but that statistic is not odd considering California has a 63.0% citizen percent. In addition, the average for San Diego income was $69,943.33 which intuitively feels high, but again is not different from the Californian average of $67,908.21. The average percent of people that worked at home was a measly 6.32% for the county and 5.17% for the state (this number has probably drastically increased in the current times). In addition, I was very surprised to see that 47% of the population was white in San Diego while only 38.72% of the population was white for California as a whole. This surprised me cause I had assumed that San Diego was more eclectic than California. Looking at information from the census bureau, it appears that the class of White sometimes includes those of Hispanic and Latino ethnicities (which explains why some sources report percents of up to 72% white population for California). This dataset appears to separate the two ethnicities definitively.

Exploratory analysis

15. Carry out PCA for both county & sub-county level census data. Compute the first two principal components PC1 and PC2 for both county and sub-county respectively. Discuss whether you chose to center and scale the features and the reasons for your choice. Examine and interpret the loadings.

Table 16: Head of County PCA Loadings

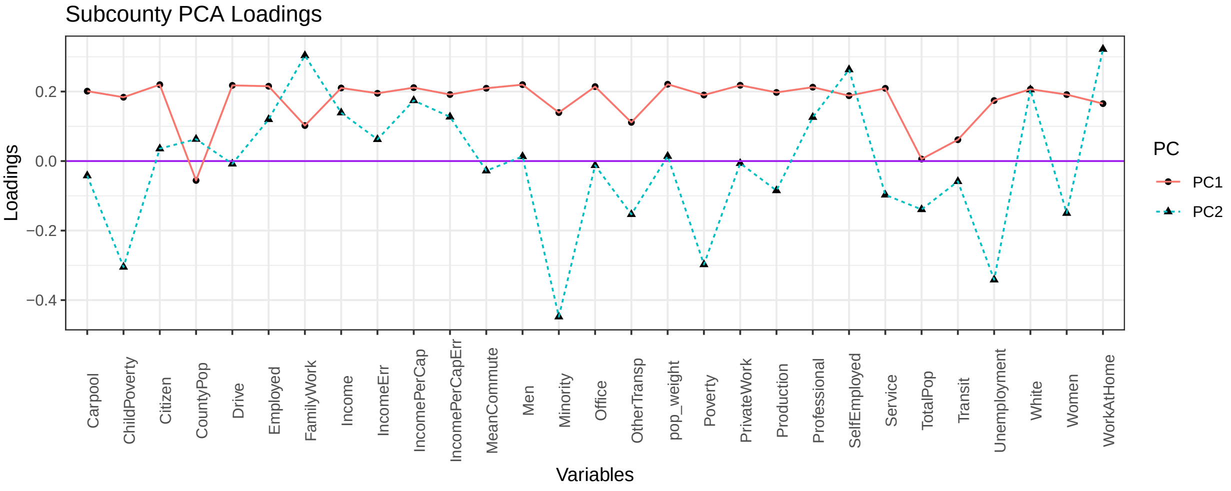
PC1	PC2
-0.3539	-2.116
-1.05	-2.404
4.197	0.1772
1.904	-0.0162
0.695	-2.211
4.032	-0.8589



Single value decomposition is dependent on squared error loss and as a result, is dependent upon variable scale. So, we chose to scale variables before conducting principle component analysis. PC1 up-weights Child Poverty, Minority, Poverty, Service and Unemployment the most, and down weights Employed, Income, IncomePerCap, Professional and White. Therefore, PC1 mostly represents the employment/income/minority/or not components, and a high PC1 corresponds to high poverty and likelihood of being a minority while a low PC1 corresponds to higher level employment/income/likelihood to be white. PC2 up-weights FamilyWork, Men, SelfEmployed, and Work-At-Home the most, and down weights Women, PrivateWork, Drive, Office, and IncomeErr. So, PC2 represents sex and what someone's job is like. High PC2 means high likelihood of being a man, self employed, and working from home, and vice versa.

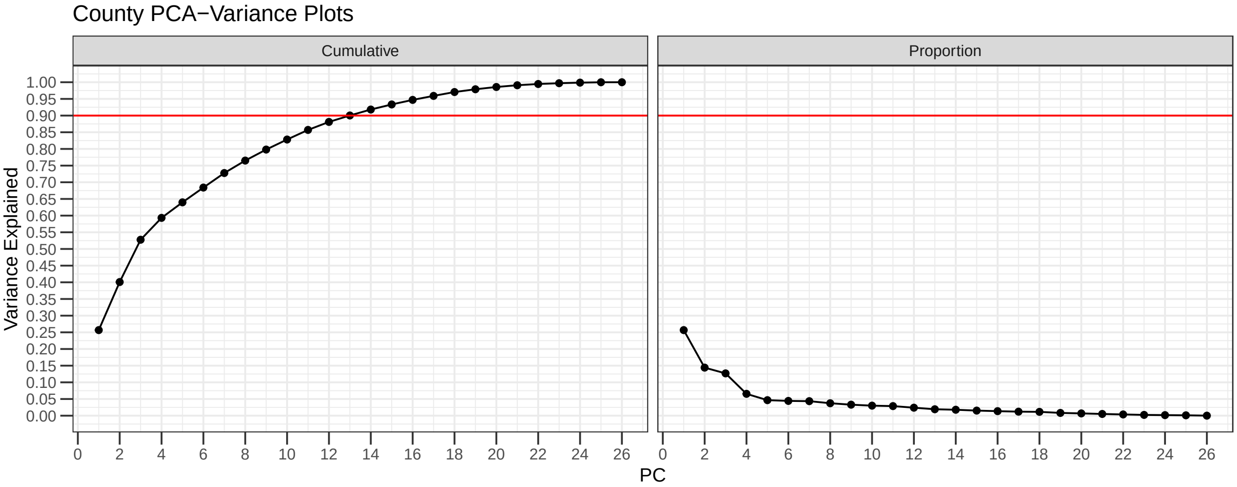
Table 17: Head of Subcounty PCA Loadings

PC1	PC2
-0.4911	0.3867
-0.1619	-0.4171
0.2176	0.1807
1.454	-0.007746
7.22	-1.044
1.547	-0.124

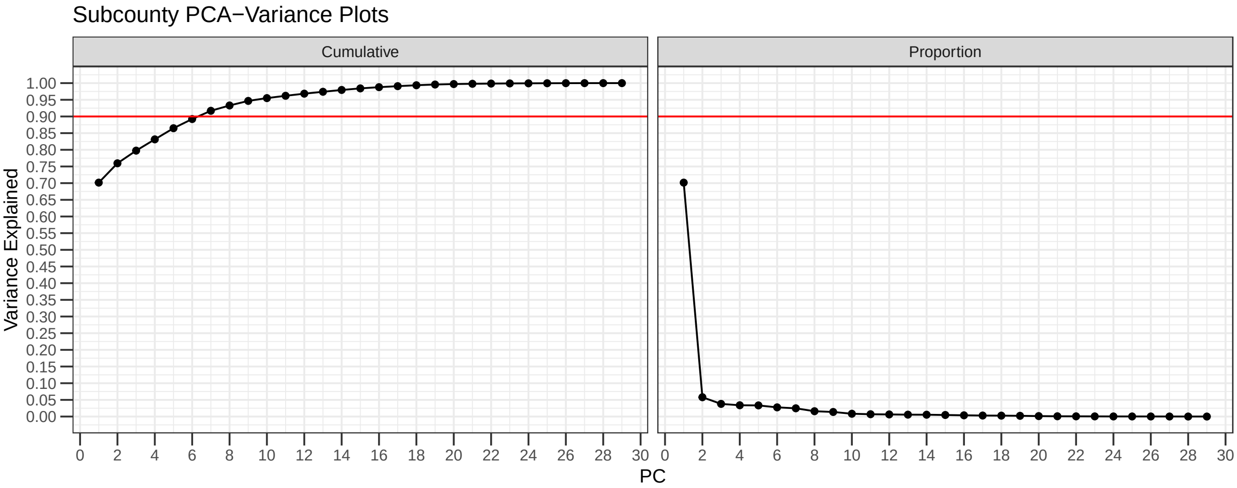


PC1 seems to upweight all variables except for TotalPop and CountyPop, which are weighted nearly neutrally. Of the up-weighted variables, Family Work, Minority, OfficeTransport and Transit are weighted the least. PC2 mostly up-weights FamilyWork, SelfEmployed, White, and WorkAtHome, and mostly down weights ChildPoverty, Minority, Poverty and Unemployment. So a high PC2 represents high rates of SelfEmployment, Whiteness, FamilyWork and WorkAtHome and low rates of Minority-hood, Unemployment and Poverty, and the opposite is true for low PC2.

16. Determine the minimum number of PCs needed to capture 90% of the variance for both the county and sub-county analyses. Plot the proportion of variance explained and cumulative variance explained for both county and sub-county analyses.



Looking at the cumulative variance plot, it can be seen that 13 PCs are required to capture 90% of the variance of the county data.



Looking at the cumulative variance plot, it can be seen that 6 PCs are required to capture 90% of the variance of the subcounty data.

17. With `census_ct`, perform hierarchical clustering with complete linkage. Cut the tree to partition the observations into 10 clusters. Re-run the hierarchical clustering algorithm using the first 5 principal components the county-level data as inputs instead of the original features. Compare and contrast the results. For both approaches investigate the cluster that contains San Mateo County. Which approach

seemed to put San Mateo County in a more appropriate cluster? Comment on what you observe and discuss possible explanations for these observations.

```
## Warning in dist(census_ct, method = "euclidean"): NAs introduced by coercion
```

Table 18: Hierarchical Clustering Counts on Census Variables

clusters	n
cluster 1	1505
cluster 2	243
cluster 3	794
cluster 4	409
cluster 5	145
cluster 6	30
cluster 7	19
cluster 8	8
cluster 9	5
cluster 10	60

Table 19: Hierarchical Clustering Counts on first 5 PCs

clusters	n
cluster 1	1950
cluster 2	631
cluster 3	312
cluster 4	57
cluster 5	8
cluster 6	167
cluster 7	43
cluster 8	7
cluster 9	40
cluster 10	3

Table 20: Comparing San Mateo Clustering

Data	Cluster	Mean	San_Mateo_Distance	Variance
Original Features	7	24489	16801	388620521
First 5 PCs	3	3.894	5.681	3.432

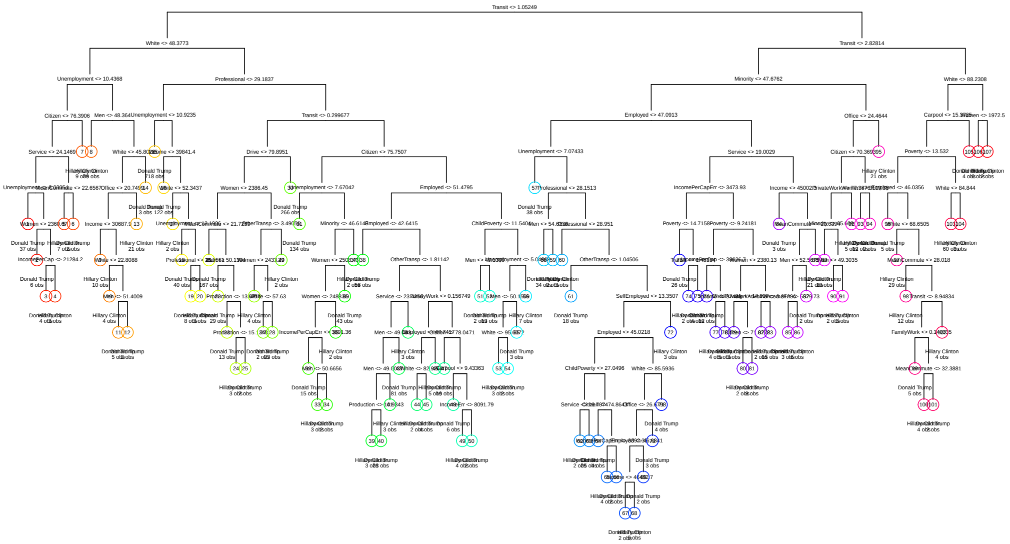
The original features place San Mateo County in cluster 7. In constrast, the first 5 PCs place San Mateo County in cluster 3. The first 5 PCs approach appears to put San Mateo County in a more appropriate cluster.

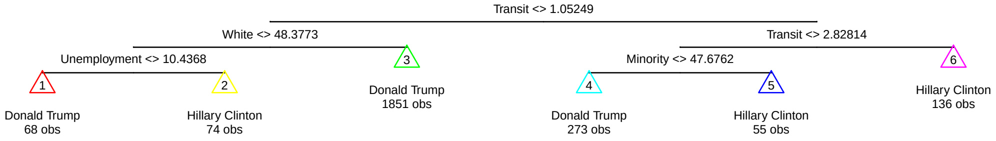
Classification

In order to train classification models, we need to combine `county_winner` and `census_ct` data. This seemingly straightforward task is harder than it sounds. Codes are provided in the .Rmd file that make the necessary changes to merge them into `election_cl` for classification.

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

18. Decision tree: train a decision tree on the training partition, and apply cost-complexity pruning. Visualize the tree before and after pruning. Estimate the misclassification errors on the test partition, and intepret and discuss the results of the decision tree analysis. Use your plot to tell a story about voting behavior in the US (see this NYT infographic).





Looking at the pruned tree, it appears that the percent of the population that utilizes public transportation is one of the most important variables as it is split upon the earliest and twice. It appears that increased percent of public transportation use favors Hillary Clinton. In addition, increased percent minority and increased population unemployment favors Hillary Clinton. In contrast, an increased white population percent favors Donald Trump.

Table 21: Pruned Tree Misclassification Error

	Donald Trump	Hillary Clinton
Donald Trump	0.9785	0.0215

	Donald Trump	Hillary Clinton
Hillary Clinton	0.3956	0.6044

19. Train a logistic regression model on the training partition to predict the winning candidate in each county and estimate errors on the test partition. What are the significant variables? Are these consistent with what you observed in the decision tree analysis? Interpret the meaning of one or two significant coefficients of your choice in terms of a unit change in the variables. Did the results in your particular county (from question 14) match the predicted results?

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Table 22: Logist Model Misclassification Error

	Donald Trump	Hillary Clinton
Donald Trump	0.9752	0.02484
Hillary Clinton	0.3049	0.6951

Table 23: Logistic Model Significant variables

Variables	Estimate	Exponential_coef	Pr(>|z|)
Service	0.3373	1.401	7.618e-12
Professional	0.2479	1.281	6.396e-11
Employed	0.1948	1.215	2.983e-09
Unemployment	0.1986	1.22	6.789e-07
PrivateWork	0.1025	1.108	1.349e-06
Drive	-0.2303	0.7943	1.894e-06
Production	0.1667	1.181	8.047e-05
Citizen	0.1055	1.111	0.0001528
IncomePerCap	0.0002387	1	0.000295
Carpool	-0.2118	0.8092	0.0007599
Office	0.1272	1.136	0.005071
Men	0.1405	1.151	0.005857
Intercept	-27.39	1.275e-12	0.006213
IncomePerCapErr	-0.0003048	0.9997	0.0234
WorkAtHome	-0.1664	0.8467	0.02533
FamilyWork	-0.9826	0.3743	0.02714
Income	-5.699e-05	0.9999	0.03578
MeanCommute	0.05013	1.051	0.04073

In general, the significant variables in our logistic model do not match those found in the pruned tree model from the prior question. The variables white, transit, and minority were all significant in the pruned tree, but are absent from the logistic regression. Unemployment was the only variable shared between the two. It is very possible that variables found significant in the tree and unsignificant in the logistic model covaried with other variables found significant in the logistic model and so their variance was explained indirectly through other variables. A one percent increase in people in the service job field is associated with an increase in the odds of voting for Hillary Clinton by a factor of 1.401.

Table 24: Predicted vs Actual Candidate for San Diego 2016 (continued below)

county	candidate	predicted_candidate	Men	Women	White
san diego	Hillary Clinton	Hillary Clinton	50.12	3221	46.96

Table 25: Table continues below

Citizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr	Poverty
66.32	69943	10850	31282	3983	14.48

Table 26: Table continues below

ChildPoverty	Professional	Service	Office	Production	Drive	Carpool
17.13	39.26	20.13	23.69	8.689	76.55	9.594

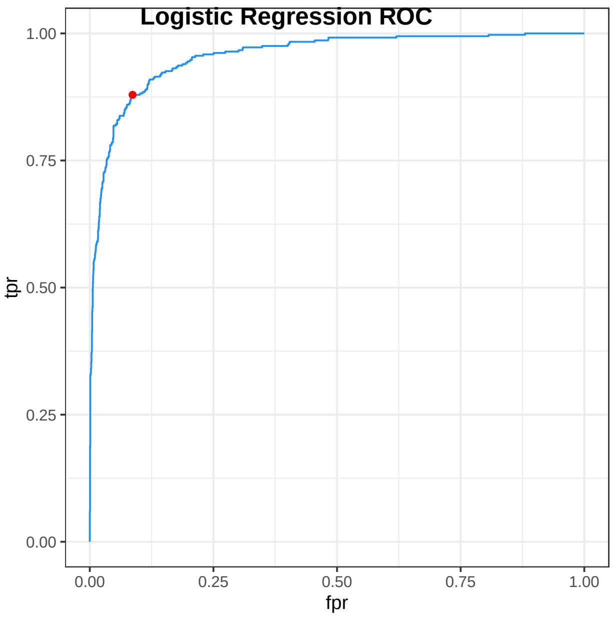
Table 27: Table continues below

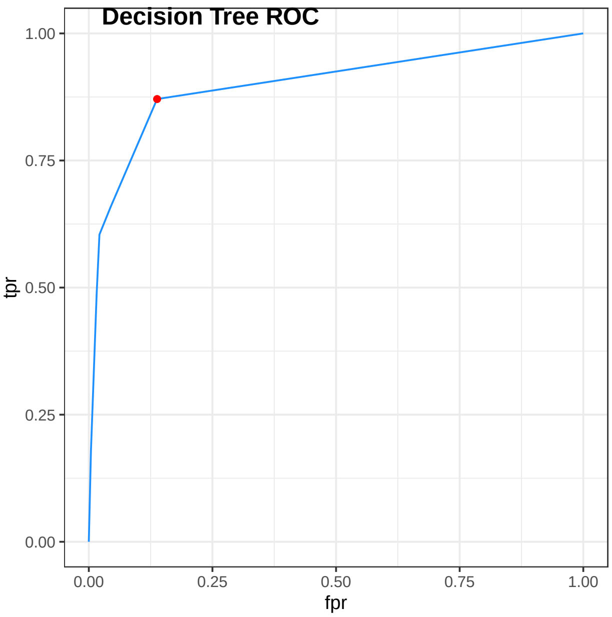
Transit	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork
3.108	1.923	6.315	25.29	45.52	76.86

SelfEmployed	FamilyWork	Unemployment	Minority
7.666	0.1605	8.948	49.75

Our model correctly predicted that Hillary Clinton would be the winning candidate for San Diego county.

20. Compute ROC curves for the decision tree and logistic regression using predictions on the test data, and display them on the same plot. Based on your classification results, discuss the pros and cons of each method. Are the different classifiers more appropriate for answering different kinds of questions about the election?





One of the biggest advantages of logistic regression is interpretability. With logistic regression coefficients, you can quantitavely explain the association between predictors and class label probabilities. Logistic modeling can also help identify important predictors, as seen in question 19. However, any logistic regression model is limited as it has assumed functional form and is increasingly more difficult to extend beyond two categories. Decision trees offer clear visualization and are very intuitive as seen with the prune tree in question 18. Another advantage is that decision trees do not require normalization or scaling of data. However, a small change in the data can cause a large change in the structure of the decision tree causing instability or overfitting. In addition, the logic behind trees does not often fit the way things happen in the real world - people or events do not behave according to hard boundary lines.

Taking it further

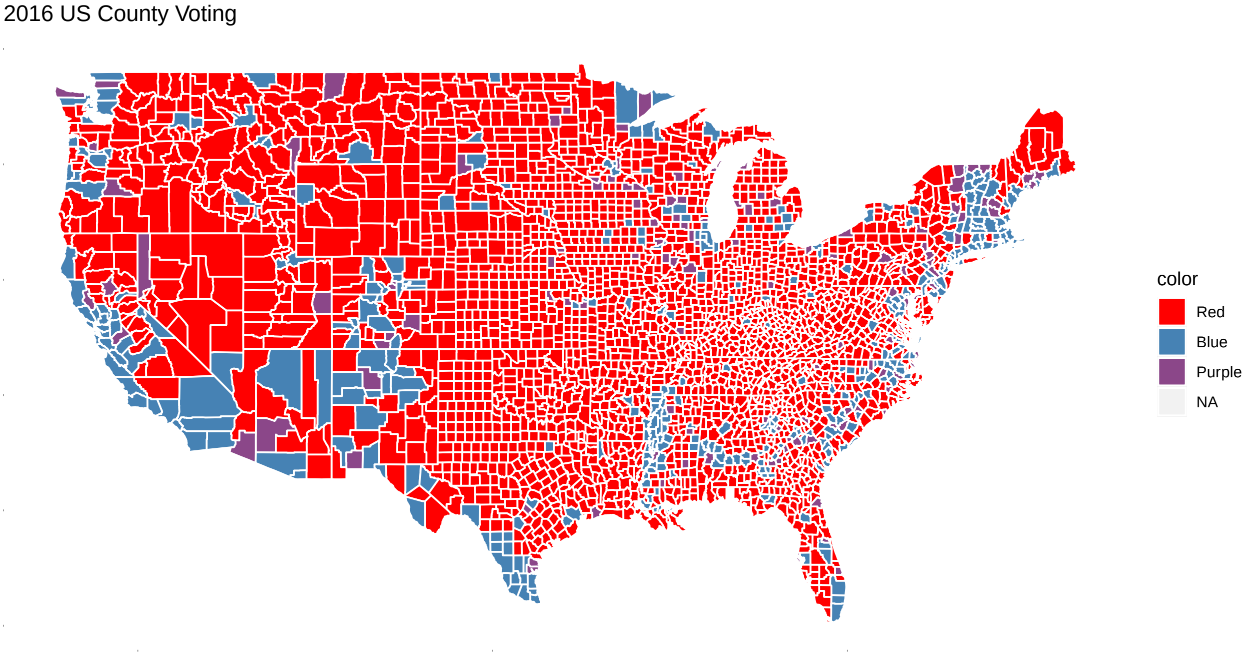
21. This is an open question. Interpret and discuss any overall insights gained in this analysis and possible explanations. Use any tools at your disposal to make your case: visualize errors on the map, discuss what does or doesn't seem reasonable based on your understanding of these methods, propose possible directions (for example, collecting additional data or domain knowledge). In addition, propose and tackle *at least* one more interesting question. Creative and thoughtful analyses will be rewarded!

Some possibilities for further exploration are:

- Data preprocessing: we aggregated sub-county level data before performing classification. Would classification at the sub-county level before determining the winner perform better? What implicit assumptions are we making?
- Exploring one or more additional classification methods: KNN, LDA, QDA, random forest, boosting, neural networks. (You may research and use methods beyond those covered in this course). How do these compare to logistic regression and the tree method?
- Use linear regression models to predict the `total` vote for each candidate by county. Compare and contrast these results with the classification models. Which do you prefer and why? How might they complement one another?
- Conduct an exploratory analysis of the "purple" counties– the counties which the models predict Clinton and Trump were roughly equally likely to win. What is it about these counties that make them hard to predict?

- Instead of using the native attributes (the original features), we can use principal components to create new (and lower dimensional) sets of features with which to train a classification model. This sometimes improves classification performance. Compare classifiers trained on the original features with those trained on PCA features.

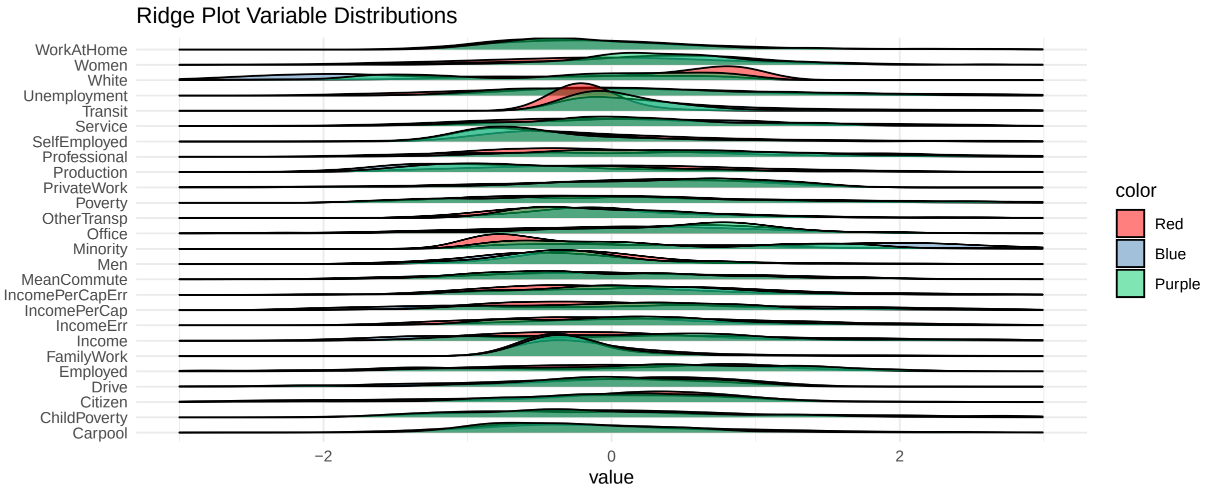
```
## Joining, by = "fips"
```



For the above map, purple states were defined as having less than a 4% total vote difference between the top two candidates for the county.

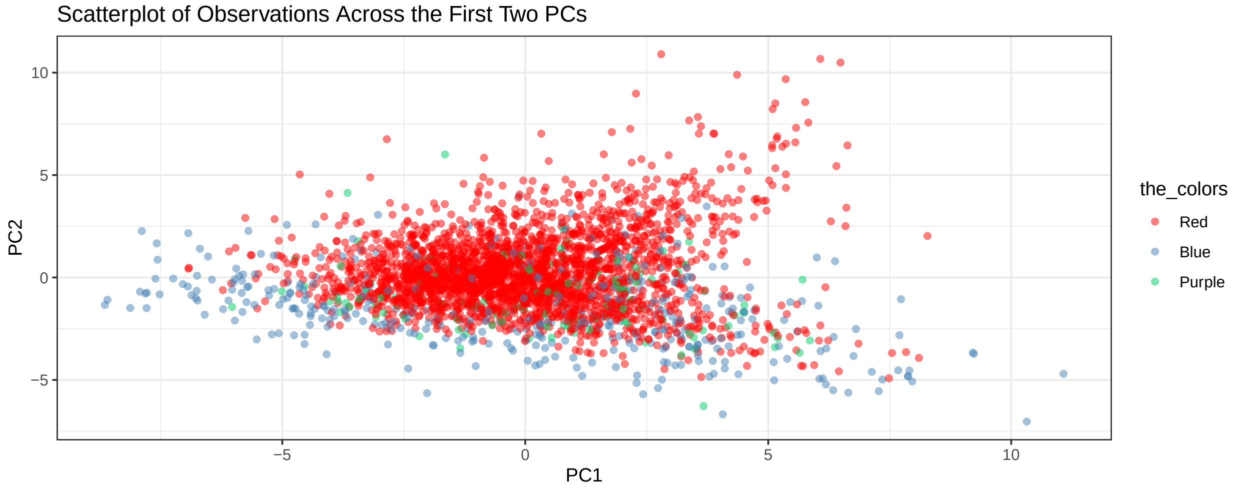
```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## Warning: Removed 1018 rows containing non-finite values (stat_density_ridges).
```



Purple counties were chosen to be represented as green for the ridge plot as it would be difficult to distinguish purple from red and blue when they overlap. In general, we see that purple counties tend to have distributions between red and blue counties. This can be exemplified when looking at the transit variable, purple counties have a taller distribution than blue counties, but a shorted distribution than red counties. In general, there are very few variables were the distribution of purple counties stand out from the other two categories. The distribution for purple counties looks especially left skewed for the office variable.

```
## New names:
## * NA -> ...3
```



Looking at the observations in respect to the first two principle components we see a similar story. We can somewhat distringuish red counties from blue counties - red counties are clustered in the middle while blue counties are more towards the periphery. But the purple counties - represnted in green are very spread out and difficult to group together.

Table 29: Table continues below

k1_test	k5_test	k10_test	k15_test	k20_test	k25_test	k1_train
0.1319	0.09967	0.09837	0.09707	0.1003	0.1026	0.0005212

k5_train	k10_train	k15_train	k20_train	k25_train
0.07375	0.08423	0.08801	0.09218	0.09635

Training the data using k nearest neighbors, we see that the smallest test error occurs at a k of 15 and the smallest training error occurs at a k of 1. A larger k implies that states with similar features tend to vote in a similar fashion.

Table 31: True vs Predicted Error Rates: KNN LOOCV

	Red	Blue	Purple
Red	0.9323	0.03777	0.0299
Blue	0.2183	0.6878	0.09391
Purple	0.5896	0.2985	0.1119

Applying LOOCV K nearest neighbors, we see that the correct prediction rate for red counties is relatively high 0.9323. While the correct prediction rate for blue counties is much lower at 0.6878 and the prediction for purple counties is almost useless 0.1119.

In summary, purple counties were defined as those that had a voting percent differential of less than or equal to 4%. Looking at the variable distribution of the three categories, at appeared that purple counties tended to have features inbetween blue and purple counties. Visualing across the first two principle components, it could be observed that purple counties were very difficult to extricate from blue and especially red counties. Attempting to apply k nearest neighbors to the data confirmed this hunch as purple county accuracy was especially poor.