Final project

PSTAT131-231

Gabrielle Smith (131) Ron Vecther (131) Sarah Haley (131)

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?

The major issue in predicting voter behavior is that there is a difference between what Silver calls the 'nowcast', a model of how people will vote if the election is held on a particular day, versus true voting intention, which often changes according to known variables such as age, race, gender, etc., as well as immeasurable factors such as effects of the economy and particularly strong campaigns. Furthermore, polls have errors, and this error can aggregate from the regional level, to the state level, and finally to the national level in a hierarchical manner, and can therefore cause large prediction errors.

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

Silver was able to achieve good predictions in 2012 by examining a full range of probabilities for each date instead of maximizing probabilities, using models from the day prior (which give reports of actual support) to measure the probability of support shifts. Using this data, Silver was able to create a model that simulated forward in time to the election day, under the assumption that the starting point of the simulation (based on most recent polling data, the 'nowcast'), to forecast the new probabilities of each level of support (state and national). Given that there was an immense amount of polling data coming out, especially closer to the end of the election campaign, the 'nowcast' could be constantly updated, and the variance of the true voting intention would decrease as the election approached and immeasurable factors such as economy, strength of campaign, etc. had less room to change.

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

In the 2016 election, individual polls were incorrect due to either statistical noise or other factors, such as nonresponse bias. Yet these errors are to be anticipated, and aggregation of individual polls throughout a state is intended to account for and reduce this error. In the case of 2016, the state polls missed in the same direction, which indicates a systematic polling error and implies error in the national polls, which are adjusted based on the results at the state levels. Many of the individual states that missed in their polls for this election were swing states, which caused the national polls to overestimate Clinton's lead over Trump. The impact of these polling errors is seen primarily in the Midwestern states (Iowa, Ohio, Pennsylvania, Wisconsin, and Minnesota), which Trump was mostly expected to lose, but mostly won. Some of the widely accepted theory for the errors made in polling relate to the percentage of Trump voters who are distrustful of poll calls, meaning they were reluctant or unwilling to disclose their voting intentions. Therefore, to improve future predictions, a strategy to implement would be a method of anonymous polling, which would enable voters to report their honest intentions and could help predictions to be more accurate.

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.

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 exclude observations where 'fips=2000' because this fips value has an associated county value of 'NA', which should only be true of nationwide and statewide tallies (where the 'fips' variable is represented by a value of either 'US' or the state's name), not countywide tallies (where the 'fips' variable is numeric).

The new dimensions of the election_raw data are 18345 x 5.

Census data

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

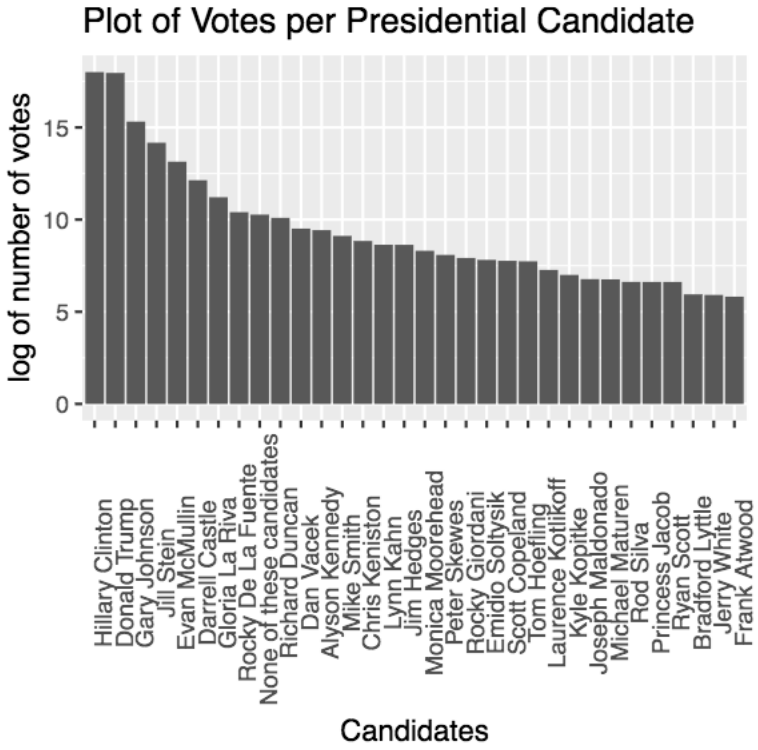
CensusTract	State	County	TotalPop	Men	Women
1001020100	Alabama	Autauga	1948	940	1008
1001020200	Alabama	Autauga	2156	1059	1097
1001020300	Alabama	Autauga	2968	1364	1604
1001020400	Alabama	Autauga	4423	2172	2251
1001020500	Alabama	Autauga	10763	4922	5841
1001020600	Alabama	Autauga	3851	1787	2064

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.)



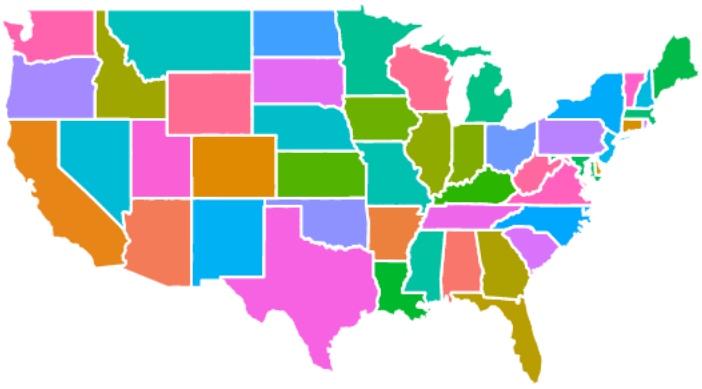
There were 32 different Presidential candidates that received votes in the 2016 US presidential election.

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).)

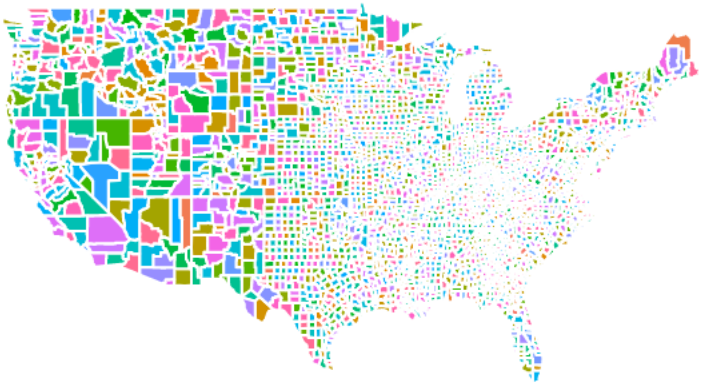
Visualization

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

```
## Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
## (as 'lib' is unspecified)
```



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`).

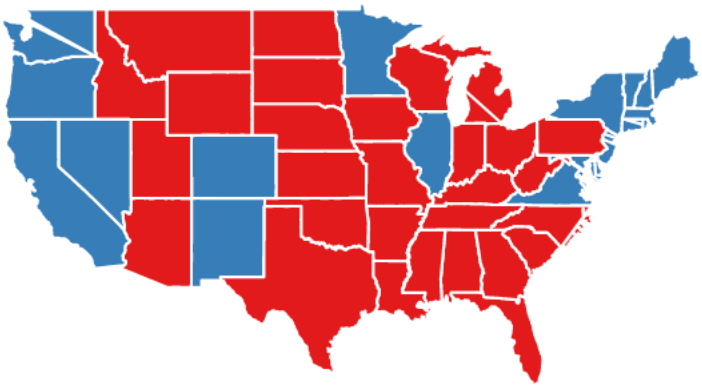
The function `left_join()` will do the trick, but needs to join the data frames on a variable with values that match. In this case, that variable is the state name, but abbreviations are used in one data frame and the full name is used in the other.

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`.

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.

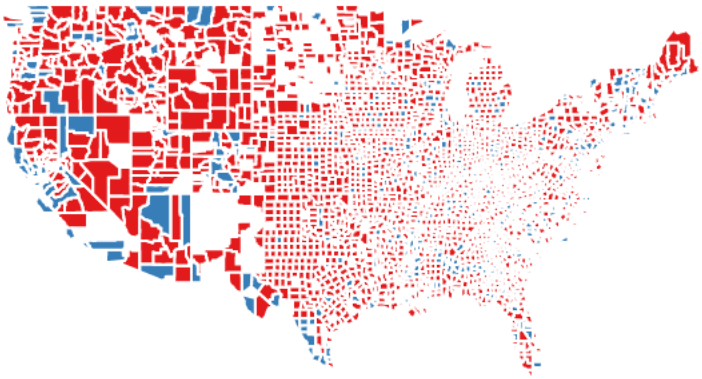
10. Use `left_join` to merge the tables and use the result to create a map of the election results by state. Your figure will look similar to this state level New York Times map. (Hint: use `scale_fill_brewer(palette="Set1")` for a red-and-blue map.)

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

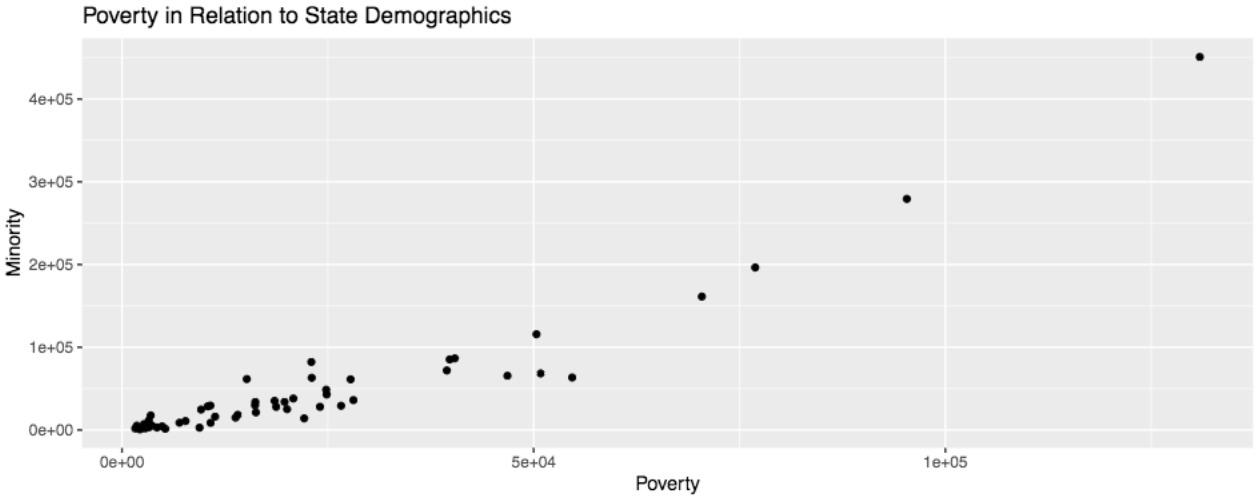


11. Now create a county-level map. The county-level map data does not have a `fips` value, so to create one, use information from `maps::county.fips`: split the `polyname` column to `region` and `subregion` using `tidyr::separate`, and use `left_join()` to combine `county.fips` with the county-level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

```
## Joining, by = "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.



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`.

State	County	CensusTract	TotalPop	Men	Women
Alabama	Autauga	1.001e+09	6486	48.43	51.57
Alabama	Baldwin	812351652	6235	39.56	41.43
Alabama	Barbour	620485497	2051	33.2	28.48
Alabama	Bibb	128297002	812.9	6.805	5.936
Alabama	Blount	318411118	2215	15.59	15.97

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.

```
## # A tibble: 1 x 7
## # Groups:   fips [1]
##   county               fips candidate       state  votes  total   pct
##   <chr>               <dbl> <chr>           <chr>  <dbl>  <dbl> <dbl>
## 1 Santa Barbara County  6083 Hillary Clinton CA    107142 174667 0.613
```

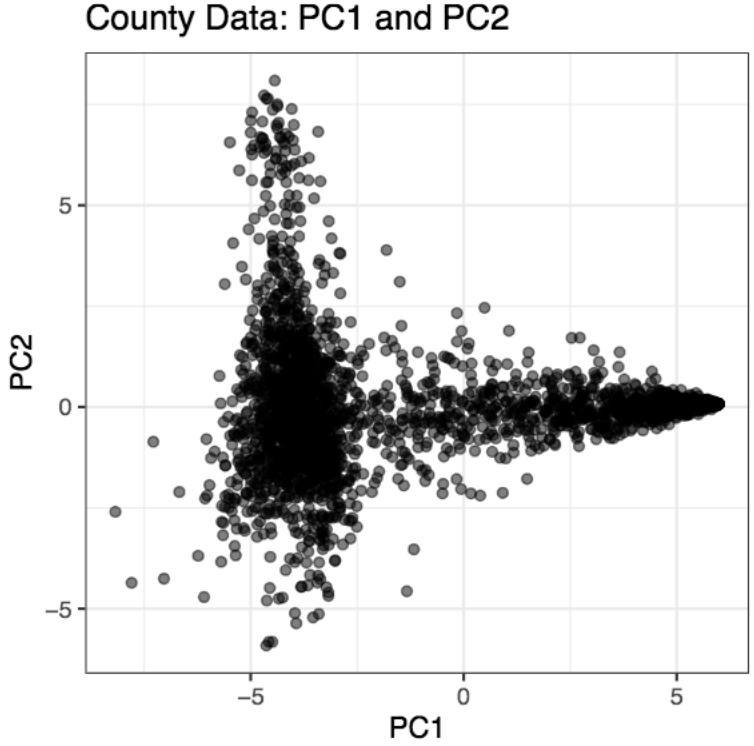
```
## # A tibble: 1 x 30
## # Groups:   State [1]
##   State County CensusTract TotalPop   Men Women White Minority Citizen Income
##   <chr> <chr>        <dbl>    <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>  <dbl>
## 1 Cali~ Santa~ 6083002003.    5886.  49.8  50.2  46.5     51.2   3598. 66498.
## # ... with 20 more variables: IncomeErr <dbl>, IncomePerCap <dbl>,
## #   IncomePerCapErr <dbl>, Poverty <dbl>, ChildPoverty <dbl>,
## #   Professional <dbl>, Service <dbl>, Office <dbl>, Production <dbl>,
## #   Drive <dbl>, Carpool <dbl>, Transit <dbl>, OtherTransp <dbl>,
## #   WorkAtHome <dbl>, MeanCommute <dbl>, Employed <dbl>, PrivateWork <dbl>,
## #   SelfEmployed <dbl>, FamilyWork <dbl>, Unemployment <dbl>
```

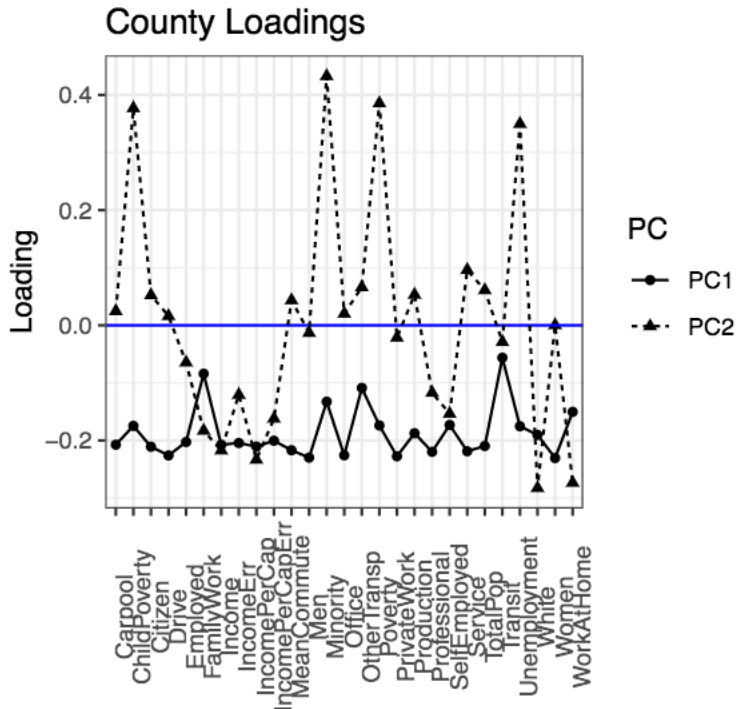
Our team was located in Santa Barbara,CA on election day in 2016. By looking at the county winner for Santa Barbara County, we see that Hillary Clinton was the winning candidate, and over 61% of eligible voters voted. Now, looking at census data for demographics in Santa Barbara, we are not surprised that Hillary was the winning representative because Santa Barbara is seen as a progressive city that would likely choose Clinton over Trump. Some key demographics of note are the diversity of race within the county, which has a greater minority population (51.18%) than white population (46.49%), as well as a large discrepancy between

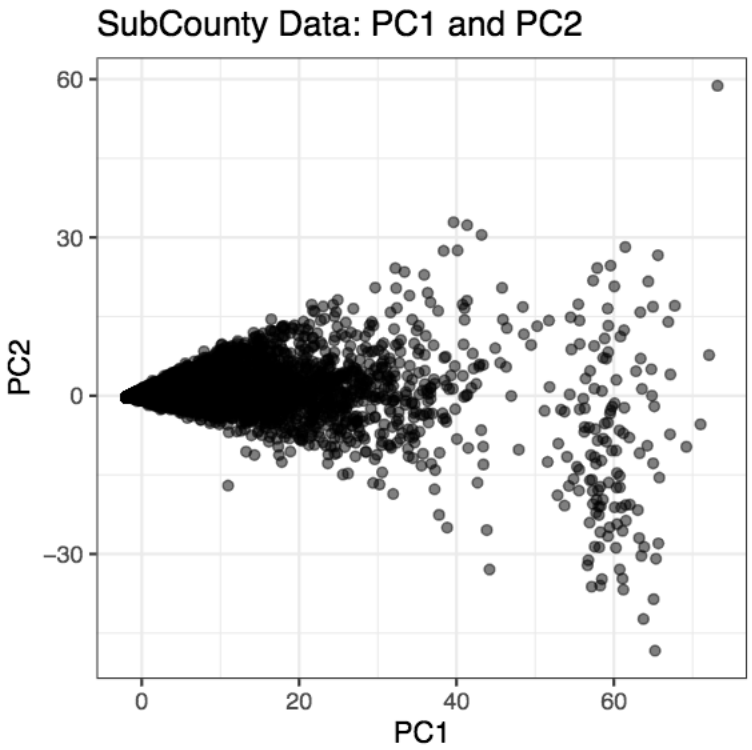
median income ($66,498.12) and income per capita ($30, 752.87). This shows that while Santa Barbara county is highly diverse, it still has a large wealth gap that could impact the way people vote within the county. This information is not surprising, and while it likely didn't greatly affect the overall California vote, it is a contributing turnout and makes us proud to be from Santa Barbara.

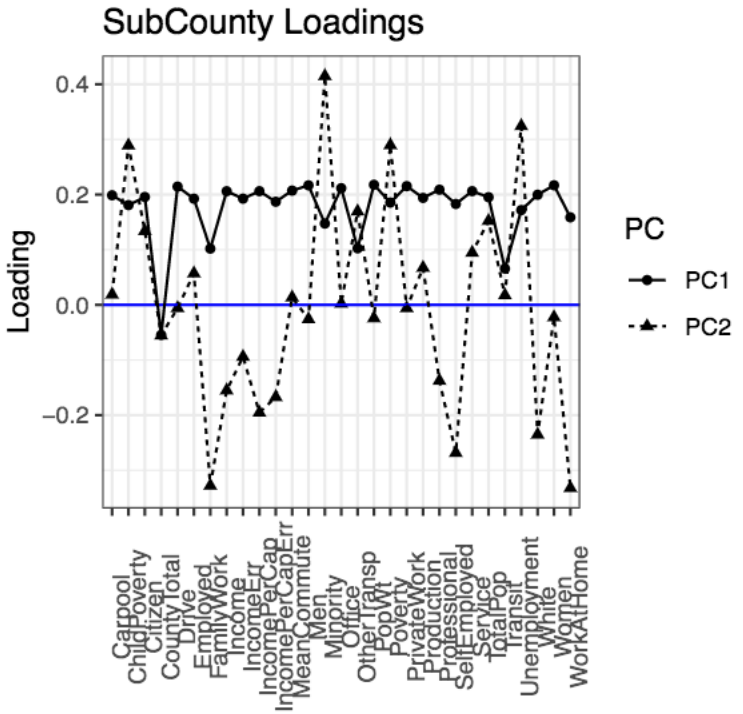
Exploratory analysis

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.




County Loadings
Loading
0.4
0.2
0.0
−0.2
PC
PC1
PC2
Carpool
ChildPoverty
Citizen
Drive
Employed
FamilyWork
Income
IncomeErr
IncomePerCap
IncomePerCapErr
MeanCommute
Men
Minority
Office
OtherTransp
Poverty
PrivateWork
Production
Professional
SelfEmployed
Service
TotalPop
Transit
Unemployment
White
Women
WorkAtHome


SubCounty Data: PC1 and PC2
PC2
60
30
0
−30
0
20
40
60
PC1

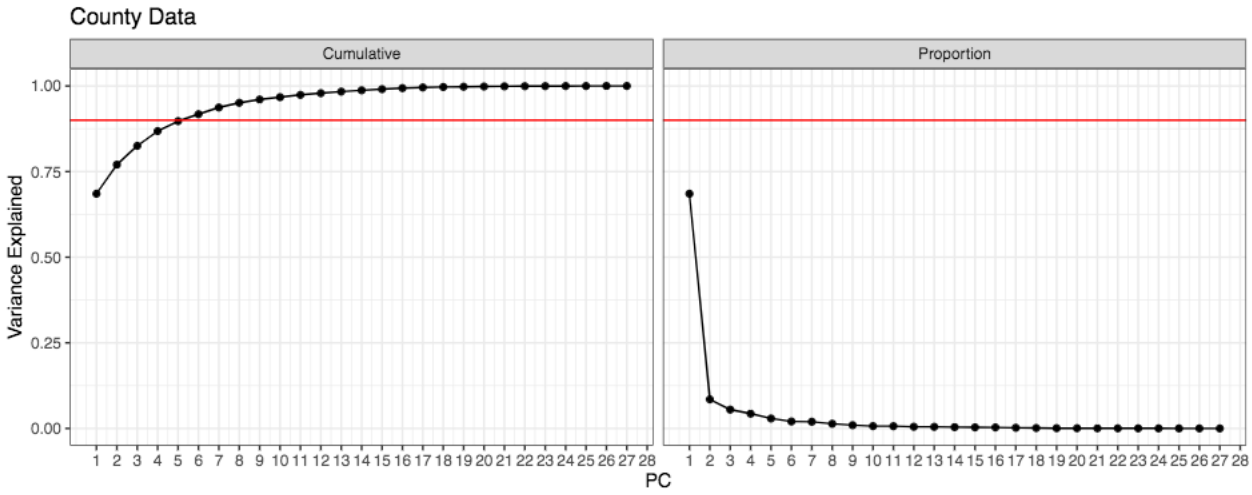


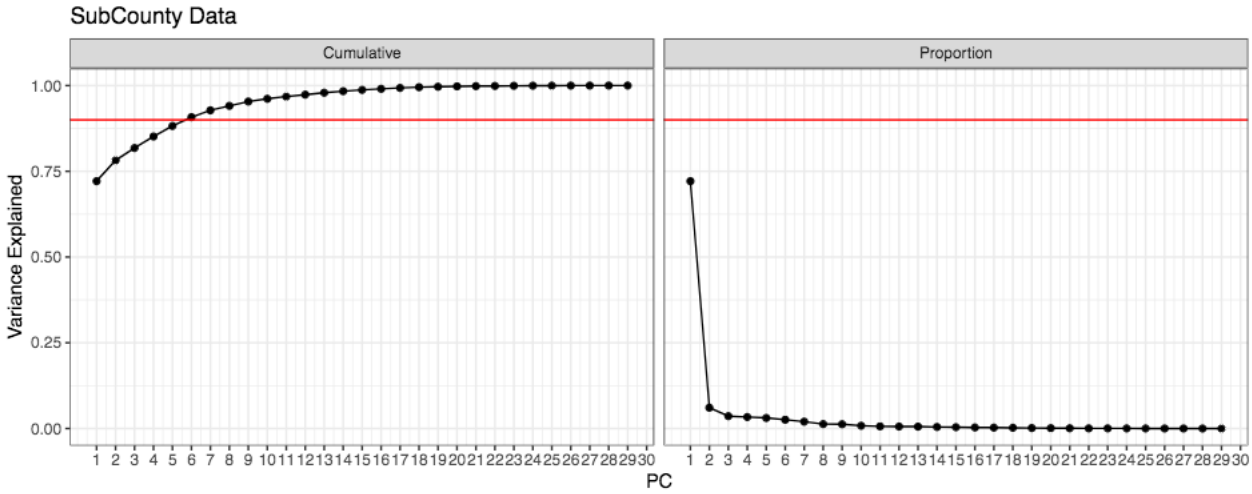
We chose to center and scale the features. The objective of centering and scaling features before performing PCA is to normalize the data so that all variables have the same standard deviation, and therefore all variables have the same weight, which helps our PCA to calculate relevant axes. In our case, since census_ct variables have a range of population-weighted averages, we opt to center and scale the features because our analysis and interpretation is sensitive to these weights, and we want to consider variables that are equally weighted.

For the sub-county loadings, we note that PC2 will be large and positive when variables including child poverty, poverty, minority, and unemployment are high, while variables like work at home, family work, self employed, and white are low. The loadings for PC1 are relatively constant across all variables, where PC1 will be large and positive when all variables are moderately high with slightly lower measures of county total, transit and other transportation.

For county loadings, we find that PC2 will be large and positive when variables including white, work at home, and income per capita are high, while variables including child poverty, poverty, minority, and unemployment are low. The loadings for PC1 are relatively constant across all variables, where PC1 will be large and positive when all variables are relatively low with slightly higher measures of family work, minority, transit and other transportation.

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.





To capture 90% of the variance in County-Level data, we need exactly 5 PCs. However, to capture 90% of the variance in SubCounty-Level data, we need a minimum of 6 PCs, where 6 principal components will capture slightly more than 90% of the variance and 5 principal components will capture slightly less than 90% of the variance.

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.

clusters	n
cluster 1	1271
cluster 2	230
cluster 3	279
cluster 4	358
cluster 5	189
cluster 6	149
cluster 7	283
cluster 8	77
cluster 9	295

clusters	n
cluster 10	87

clusters2	n
cluster 1	1308
cluster 2	1585
cluster 3	131
cluster 4	5
cluster 5	157
cluster 6	9
cluster 7	2
cluster 8	13
cluster 9	4
cluster 10	4

```
## [1] cluster 1
## 10 Levels: cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 ... cluster 10
```

```
## [1] cluster 1
## 10 Levels: cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 ... cluster 10
```

We believe that running hierarchical clustering using the first five principal components as inputs instead of the original features is more appropriate. Given that the data is highly concentrated into cluster 1 when running the hierarchical clustering using the original features, it is unlikely that further analysis can be done in regards to these classifications. Using the principal components as inputs implements data divisions based on pattern encoding the highest variance in the dataset. Using this preprocessing to reduce the dimensions of our data is an important step in the accuracy for the complete linkage model because with higher dimensions it is harder for distance models to predict accurately. Since the preprocessing of dimensionality reduction occurred and we only input the first 5 principal components, the hierarchical clustering is prone to be more accurate because of the reduced dimensionality. If we opted to use the original features in hierarchical clustering, we might consider a correlation-based similarity measure approach to be more appropriate.

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).

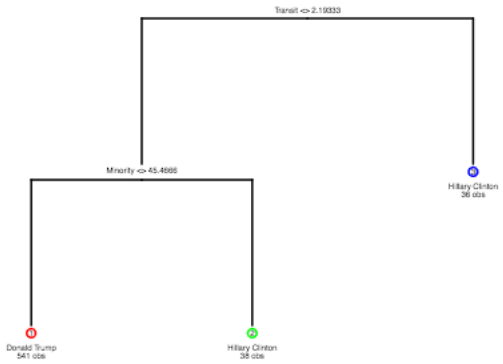
```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin

## Loaded gbm 2.1.8

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

## Loading required package: cluster

##
## Attaching package: 'cluster'

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

## Loading required package: rpart

##
## Classification tree:
## tree(formula = as.factor(candidate) ~ ., data = train, control = tree_opts,
##     split = "deviance")
## Variables actually used in tree construction:
##  [1] "Transit"         "Minority"        "SelfEmployed"    "White"
##  [5] "Men"             "CensusTract"     "FamilyWork"      "Poverty"
##  [9] "Income"          "PrivateWork"     "Unemployment"    "MeanCommute"
## [13] "Office"          "Service"         "IncomePerCapErr"
## Number of terminal nodes:  39
## Residual mean deviance:  0.04214 = 24.27 / 576
## Misclassification error rate: 0.01301 = 8 / 615
```




```
##
## Classification tree:
## snip.tree(tree = t, nodes = c(3L, 5L, 4L))
## Variables actually used in tree construction:
## [1] "Transit"  "Minority"
## Number of terminal nodes:  3
## Residual mean deviance:  0.5879 = 359.8 / 612
## Misclassification error rate: 0.09919 = 61 / 615
```

```
##                  pred
## class                      No        Yes
##   Donald Trump    0.93599615 0.06400385
##   Hillary Clinton 0.50663130 0.49336870
```

```
##                  pred
## class                      No        Yes
##   Donald Trump    0.94706449 0.05293551
##   Hillary Clinton 0.48806366 0.51193634
```

The decision tree before pruning contains 41 terminal nodes, split on 16 variables, with a total misclassification error rate of 0.01301. After cost-complexity pruning, we have a decision tree with only three terminal nodes, split on two variables, Minority and Transit, with a total misclassification error rate of 0.1171. When we examine the misclassification errors on test data, we note that the pruned tree has a higher total misclassification error rate than the initial tree. The pruned tree also has a higher false negative rate and a lower false positive rate than the initial tree; we suspect that this is the result of overfitting, where the overfit model correctly classifies the default with near perfect accuracy, but is fairly imprecise in classifying non-defaults. The first variable that the tree is split on is Transit, where we suspect that the percentage of people in a county who utilize the transit system is correlated to other demographics and wealth distribution in the county's population. Therefore, we see that counties with higher rates of public transit use, likely to be lower-income counties, are determined to be more likely to vote for Hillary. The resulting population (low-transit) is then split on a Minority variable, where counties with higher rates of Minority identification in their population are more likely to vote for Hillary, and counties with lower rates of Minority voters (and in correlation, we assume a higher rate of White voters) are more likely to elect Trump.

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?

```
##
## Call:
## glm(formula = as.factor(candidate) ~ ., family = "binomial",
##     data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1789  -0.4559  -0.1151  -0.0024   3.8204
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.700e+00  2.336e-01  -7.277 3.42e-13 ***
## CensusTract      9.854e-14  2.257e-11   0.004 0.996516
## TotalPop        -2.871e-03  1.511e-03  -1.901 0.057348 .
## Men             -1.081e-02  3.854e-01  -0.028 0.977628
## Women            2.388e-01  4.298e-01   0.556 0.578445
## White           -4.261e-01  1.868e-01  -2.281 0.022555 *
## Minority        -1.651e-01  1.683e-01  -0.981 0.326501
## Citizen          1.301e-03  1.948e-03   0.668 0.504420
## Income          -2.329e-04  1.326e-04  -1.756 0.079073 .
## IncomeErr       -1.298e-04  3.628e-04  -0.358 0.720524
## IncomePerCap     1.094e-03  3.907e-04   2.800 0.005118 **
## IncomePerCapErr -3.116e-03  1.260e-03  -2.472 0.013435 *
## Poverty          3.280e-01  2.323e-01   1.412 0.157995
## ChildPoverty    -6.167e-02  1.407e-01  -0.438 0.661262
## Professional     5.460e-01  2.291e-01   2.383 0.017162 *
## Service          1.147e+00  3.265e-01   3.514 0.000442 ***
## Office           4.369e-01  2.273e-01   1.922 0.054560 .
## Production       8.135e-01  2.472e-01   3.290 0.001001 **
## Drive           -6.854e-01  3.096e-01  -2.214 0.026837 *
## Carpool         -8.440e-01  3.792e-01  -2.226 0.026019 *
## Transit          1.358e-02  5.085e-01   0.027 0.978690
## OtherTransp     -1.316e+00  5.642e-01  -2.333 0.019659 *
## WorkAtHome      -2.450e-01  3.784e-01  -0.647 0.517335
## MeanCommute      1.312e-01  1.237e-01   1.061 0.288829
## Employed         6.498e-03  2.670e-03   2.434 0.014936 *
## PrivateWork      1.283e-01  8.816e-02   1.455 0.145587
## SelfEmployed    -2.837e-01  3.081e-01  -0.921 0.357163
## FamilyWork       1.797e+00  2.226e+00   0.807 0.419491
## Unemployment     1.851e-01  1.564e-01   1.184 0.236375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 501.36  on 614  degrees of freedom
## Residual deviance: 240.91  on 586  degrees of freedom
## AIC: 298.91
##
## Number of Fisher Scoring iterations: 9
```

```
## # A tibble: 1 x 4
##      fpr   tpr thresh youden
##    <dbl> <dbl>  <dbl>  <dbl>
## 1 0.0890 0.885  0.155  0.796
```

```
##                  y_hat_glm
##                         No         Yes
##   Donald Trump    0.990530303 0.009469697
```

```
##   Hillary Clinton 0.436781609 0.563218391
```

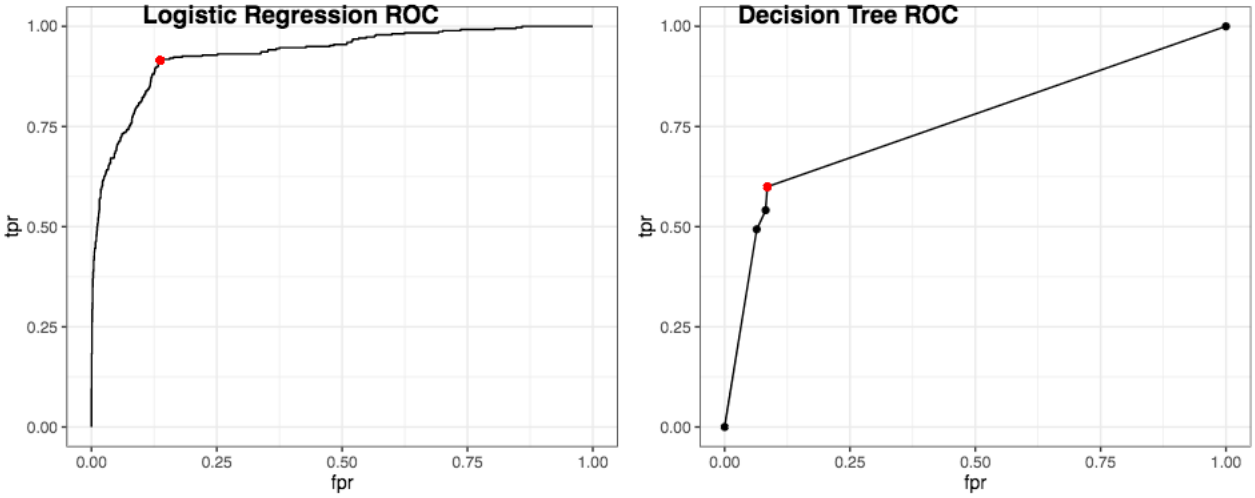
The significant variables identified in the logistic regression model (at the 0.05 significance level) are White, Income, Poverty, Professional, Service, and Drive. We find that these important variables are not consistent with our analysis done in our decision tree, which identifies two splitting variables Minority and Transit in our cost-complexity pruned model. Because logistic regression requires data to be linearly separable while decision trees capture nonlinear classification boundaries, we do not necessarily expect the identified important variables in each classification method to be consistent with one another.

We estimate that if a respondent is White, we expect the probability of Trump's election in the county to increase by 3.499e-01.

The results in our county (Santa Barbara) matched the predicted results. Hillary Clinton won the Santa Barbara County 2016 presidential election with 60.06% of the votes.

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?

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

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



From our classification results for the ROC curve, we see that the logistic regression curve has a much better optimal threshold than the decision tree, which moves in a much more step-wise pattern. Decision trees are very interpretable, but also very complex and can have large variation, which we do not want when we are trying to predict a candidate out of only 2 options. This is an easier prediction to make since there are only 2 candidates, as opposed to early in the election race. On the other hand, logistic regression maps the probabilities of each predictor being in a specific class, which could be interesting if we are trying to predict the probability of a candidate winning. However, a con to linear regression is that when it is split on nonlinear variables, the decision boundaries will be nonlinear and it will be a poor classification predictor.

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!

The main takeaway from this project is that analysis of large data sets takes diligence, but can reveal a lot about how variables relate and can lead to meaningful predictions and insightful inference. We must keep in mind that with each fitted model, there is a bias-variance tradeoff, and we must always consider the consequences of overfitting, and test our results and misclassification.

An interesting direction for this data could be to add some classification variables to the demographic based on major issues that the candidates either support or do not support. For example, when it comes to health care, Biden openly supported federally-funded health care during his campaign, while Trump claimed it was ridiculous. Citizens can eventually then fill out a survey based on current issues to them, and lead them to the candidate of their choice. This would be helpful for prediction because these major issues likely drive voting much more than demographics such as the percentage of people working from home.

An interesting question that could be asked about census data is whether economic performance affects people's vote, and subsequently the election results. This is studied greatly in political science as economic voting, which shows that Americans are sociotropic and retrospective economic voters. This means they are highly concerned with the economy at large, as well as the state of the economy in the previous term. These are other factors that could be added to census data analysis to improve predictions of election winners.

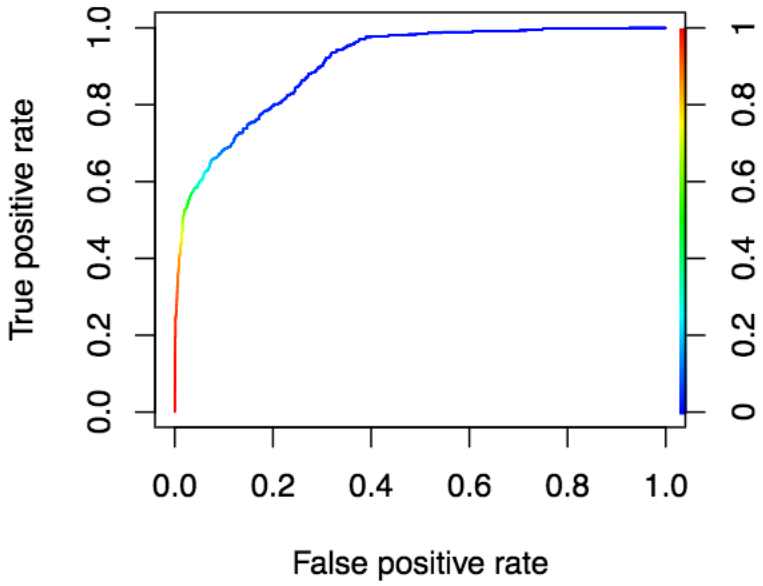
Some possibilities for further exploration are:

- 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?

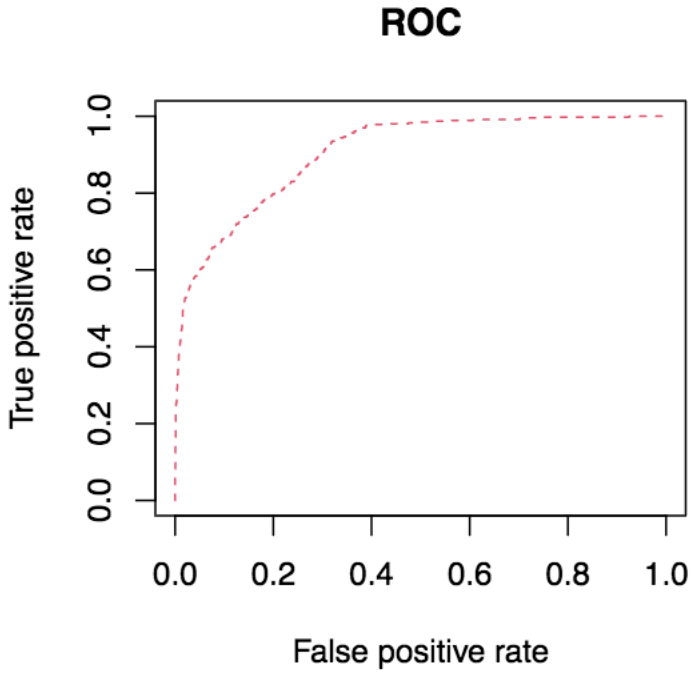
Classification at the sub-county level data before determining a winner would be much more difficult, and would likely perform worse. Depending on what variables the classes would be defined, as well as their range and density in the data, we could get very different results, and likely perform worse without knowing the winner. The main assumption we made in this project is that the census data is accurate, but another major assumption that we learned from Nate Silver's article is that people did not always vote the way they said they would in the census, such as the lot of people who claimed they would not vote for Trump, and eventually did, leading to a big surprise in the election and large errors in many prediction models.

- 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?

```
##                  pred
## class             Donald Trump Hillary Clinton
##   Donald Trump      0.97237145      0.02762855
##   Hillary Clinton   0.45474138      0.54525862
```




```
##                  pred
## class             Donald Trump Hillary Clinton
##   Donald Trump      0.94704528      0.05295472
##   Hillary Clinton   0.42672414      0.57327586
```




```
##                Donald Trump Hillary Clinton MeanDecreaseAccuracy
## Transit         0.055527238    0.0800105384         0.0589675088
## Minority        0.061537615    0.0266859699         0.0564751178
## White           0.052827827    0.0122899113         0.0471853123
## OtherTransp     0.035631390   -0.0151036038         0.0284437606
## CensusTract     0.030357903   -0.0145803354         0.0239671469
## SelfEmployed    0.027564696   -0.0094327630         0.0220181544
## Poverty         0.021739528    0.0108921009         0.0202222304
## IncomePerCap    0.023666759   -0.0063527145         0.0192396517
```

```
## Carpool          0.022554371   -0.0110570174         0.0177615626
## Service          0.021830348   -0.0081017370         0.0175946123
## MeanCommute      0.021749995   -0.0075355609         0.0175552464
## IncomePerCapErr  0.018392050   -0.0001509065         0.0157823012
## TotalPop         0.020007180   -0.0101273864         0.0156291005
## Income           0.018974493   -0.0110723640         0.0149286973
## Drive            0.017054088   -0.0030048889         0.0142448100
## Production       0.015241649   -0.0012447400         0.0131152701
## Office           0.016560726   -0.0080962448         0.0129440225
## Employed         0.016290551   -0.0074332829         0.0128406968
## Professional     0.014541712    0.0001221365         0.0124000410
## Unemployment     0.010832988    0.0175349520         0.0117940889
## PrivateWork      0.013882906   -0.0111329884         0.0102258740
## Men              0.010870141   -0.0005680353         0.0092088248
## ChildPoverty     0.009122670    0.0025634097         0.0080248483
## Citizen          0.010120770   -0.0033615123         0.0079694322
## IncomeErr        0.008388092    0.0025503961         0.0074684209
## Women            0.009072092   -0.0050248961         0.0071126498
## WorkAtHome       0.004379596   -0.0073300205         0.0026669063
## FamilyWork      -0.001513464    0.0039898403        -0.0007025043
##                 MeanDecreaseGini
## Transit                24.288495
## Minority               15.095540
## White                  10.103120
## OtherTransp             3.383905
## CensusTract             4.839292
## SelfEmployed            5.105989
## Poverty                 8.193964
## IncomePerCap            3.949300
## Carpool                 2.736859
## Service                 4.596648
## MeanCommute             3.924295
## IncomePerCapErr         3.605263
## TotalPop                2.683424
## Income                  3.870064
## Drive                   4.490469
## Production              4.591172
## Office                  3.516425
## Employed                3.337925
## Professional            4.826128
## Unemployment            5.467423
## PrivateWork             3.076971
## Men                     3.830542
## ChildPoverty            4.941327
## Citizen                 2.962310
## IncomeErr               3.215719
## Women                   2.842056
## WorkAtHome              2.841030
## FamilyWork              2.417744
```

```
##                  rtest_pred
##                   Donald Trump Hillary Clinton
##   Donald Trump      0.98556304      0.01443696
##   Hillary Clinton   0.58620690      0.41379310
```

When performing further analysis with other classification models, we note that no one model performs particularly well, with all models, including those of the decision tree and logistic regression, having a high false negative rate. This is an indicator that aligns with the widespread misprediction of the results of the election.

Similarly to the decision tree analysis, the random forest model has an extremely high true negative rate, and also a fairly high false negative rate, though the false negative rate of the random forest model is lower than that discussed in the decision tree analysis.

Linear and quadratic discriminant analysis perform fairly similarly to one another, with a more balanced true positive and true negative rate than was recorded in the decision tree and random forest models, but also had a slightly lower true negative rate than the aforementioned methods. Overall, we suspect that these models are better in context than the decision tree or random forest analysis, even if their total misclassification error is slightly higher, due to the more balanced rates that they predict.

Ultimately, we select the logistic regression as our best predictive model, which has a fairly high true negative rate while also preserving the balance between true positive and negative rates. We suspect that this model is superior because it works better in the high dimensions of our dataset and because the model only has to assign predictions to two classes: Hillary Clinton and Donald Trump.