PSTAT 131 FINAL PROJECT

March 17, 2021

Grant Keith
Sampreeth Salveru
Amber Baez

BACKGROUND

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

There are many variables that affect a mathematical model of how people in the US vote in a day, such as race, wealth, etc. From there, voting intention changes with time with measurable factors such as change in employment and immeasurable factors such as a particularly successful campaign ad. Then poll result variation, day-to-day poll variation, and varying poll corrections by pollsters have to be taken into account as well. And this all comes with the assumption that every voter is truthful and samples are random, which is often not the case, making voter prediction difficult to do.

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

Nate Silver didn't use the maximum probability but rather looks at a range of probabilities. He calculates the probability of each level of support, then using the model for actual support, calculates the probability that support has shifted from one level to another.

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

There were a number of polling errors such as underestimating certain proportions of voters or failing to obtain the same level of enthusiasm from supporters of both candidates. An aggregate of individual polls that were wrong led to an overestimation of Clinton's lead, and underestimation of Donald's support. A combination of things could be done to improve future predictions like dismissing preconceived notions of how voters will vote, more accessible polling with added anonymity, and more funding for polling.

DATA

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.

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

CensusTract <dbl>	State <chr>	County <chr>	TotalPop <dbl>	Men <dbl>	Women <dbl>
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 <chr>	description <chr>	type <chr>
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

Upon observation of the election raw data set, we can see that the rows with a fips value of 2000 are duplicate values with other rows related to the state of Alaska. Therefore, we can go ahead and remove them as they are not needed and would cause confusion later on during analysis.

After removing those values, the dataframe had the following dimensions:

```
[1] 18345     5
```

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.**

In order to obtain only federal level data, we filtered out all rows that do not contain the entry "US" in the fips column and below are the first few rows of the new dataframe:

county <chr>	fips <chr>	candidate <chr>	state <chr>	votes <dbl>
NA	US	Donald Trump	US	62984825
NA	US	Hillary Clinton	US	65853516
NA	US	Gary Johnson	US	4489221
NA	US	Jill Stein	US	1429596
NA	US	Evan McMullin	US	510002
NA	US	Darrell Castle	US	186545

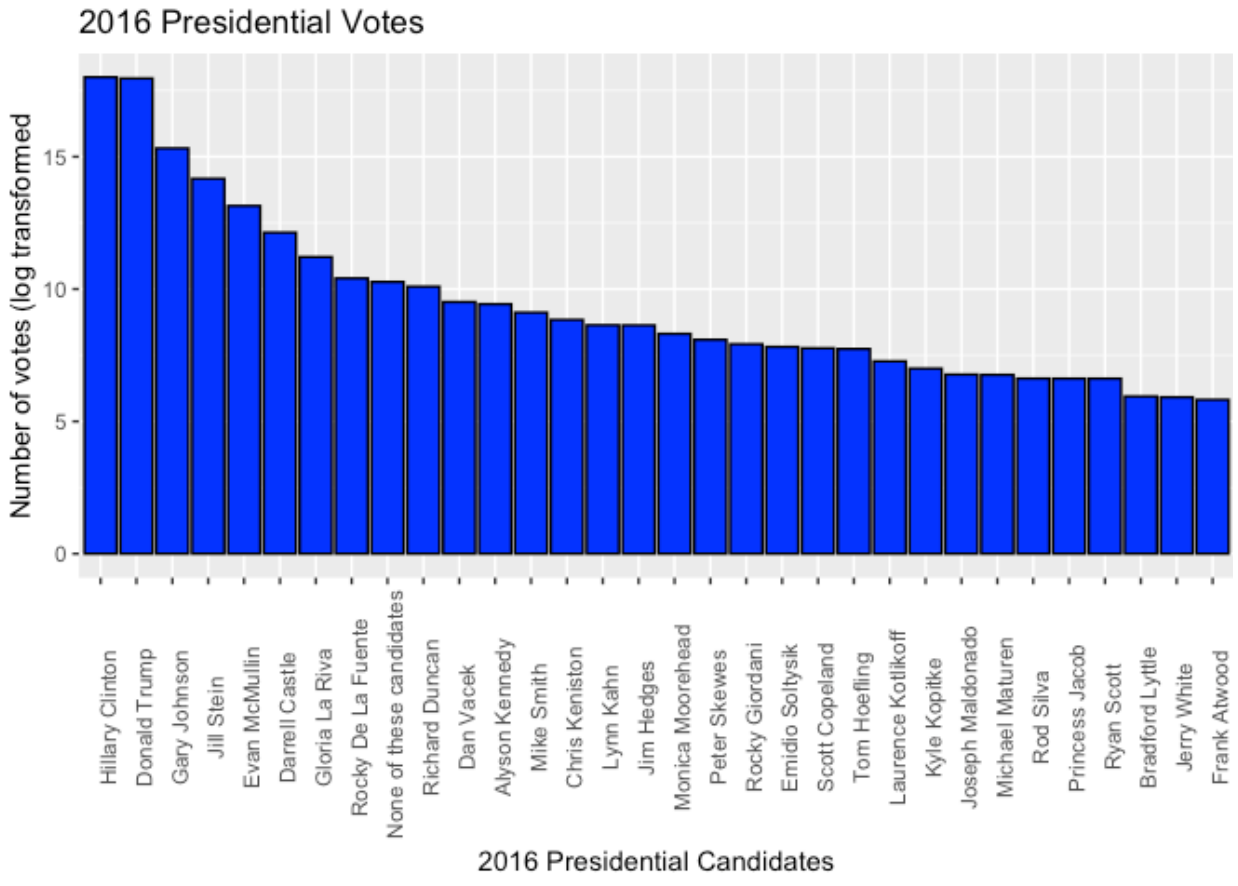
We performed a similar procedure in order to get state level tallies, which can be seen below:

county <chr>	fips <chr>	candidate <chr>	state <chr>	votes <dbl>
NA	CA	Hillary Clinton	CA	8753788
NA	CA	Donald Trump	CA	4483810
NA	CA	Gary Johnson	CA	478500
NA	CA	Jill Stein	CA	278657
NA	CA	Gloria La Riva	CA	66101
NA	FL	Donald Trump	FL	4617886

Finally, we obtained county-level results as well:

county <chr>	fips <dbl>	candidate <chr>	state <chr>	votes <dbl>
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

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.)



After log transforming the number of votes, we were able to plot a bar chart of the number of votes for each of the thirty-two candidates in descending order.

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).)

Below are the first six rows of the county_winner and state_winner data frames respectively:

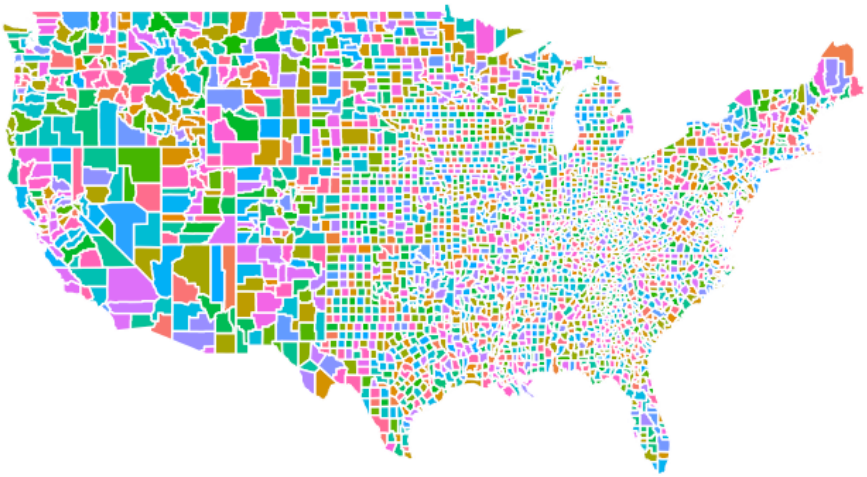
county <chr>	fips <dbl>	candidate <chr>	state <chr>	votes <dbl>	total <dbl>	pct <dbl>	County <chr>
Autauga County	1001	Donald Trump	AL	18172	24759	0.7339553	Autauga
Baldwin County	1003	Donald Trump	AL	72883	94261	0.7732042	Baldwin
Barbour County	1005	Donald Trump	AL	5454	10436	0.5226140	Barbour
Bibb County	1007	Donald Trump	AL	6738	8753	0.7697932	Bibb
Blount County	1009	Donald Trump	AL	22859	25442	0.8984750	Blount
Bullock County	1011	Hillary Clinton	AL	3530	4702	0.7507444	Bullock

county <chr>	fips <chr>	candidate <chr>	state <chr>	votes <dbl>	total <dbl>	pct <dbl>
NA	AK	Donald Trump	AK	163387	309407	0.5280650
NA	AL	Donald Trump	AL	1318255	2101660	0.6272447
NA	AR	Donald Trump	AR	684872	1130635	0.6057410
NA	AZ	Donald Trump	AZ	1252401	2554240	0.4903224
NA	CA	Hillary Clinton	CA	8753788	14060856	0.6225644
NA	CO	Hillary Clinton	CO	1338870	2780220	0.4815698

VISUALIZATION

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

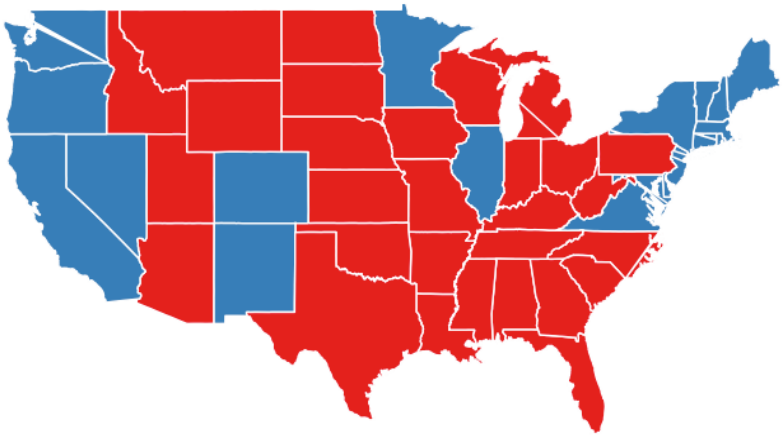
Please see the next page.



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

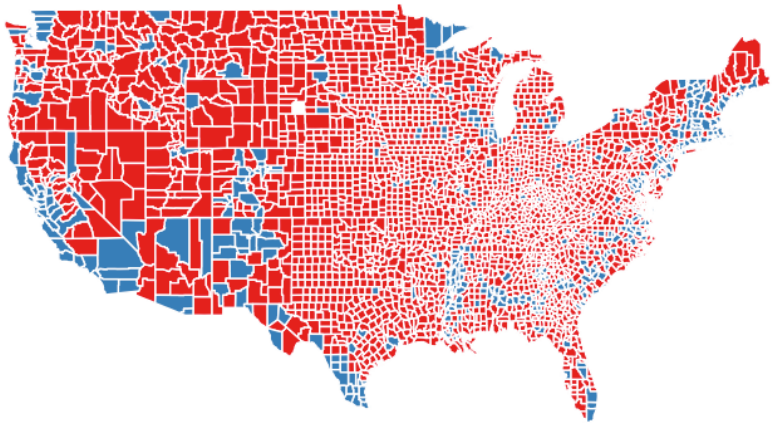
We used the provided function to create a 'fips' variable and left joined the 'states' data frame with the 'state_winner' data frame.

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](https://www.nytimes.com/elections/results/president).
(Hint: use `scale_fill_brewer(palette="Set1")` for a red-and-blue map.)



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](https://www.nytimes.com/elections/results/president).

After splitting 'polyname'in and combining it with county map and election data, we were able to plot the below map:

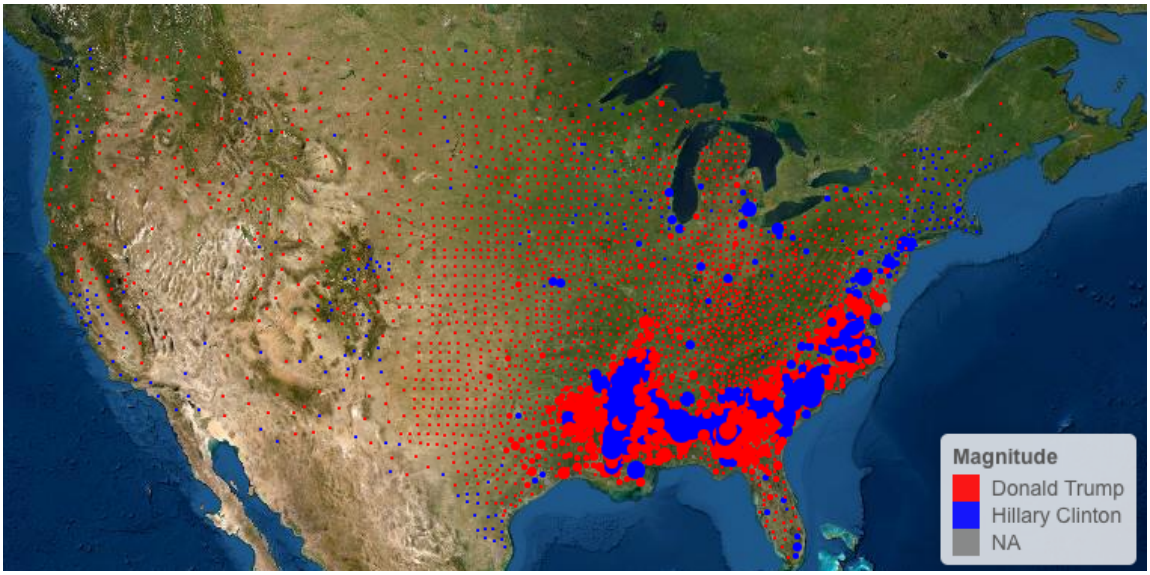


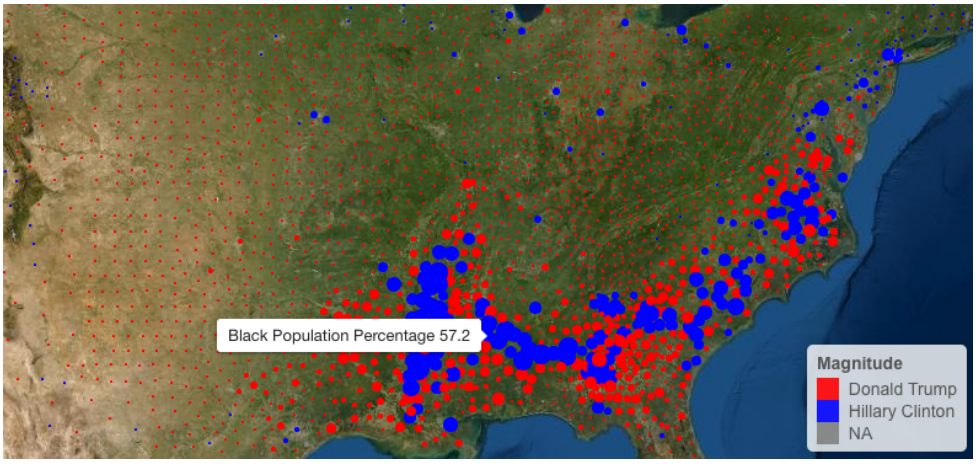
12. Create a visualization of your choice using `census` data. Many exit polls noted that [demographics played a big role in the election] (https://fivethirtyeight.com/features/demographics-not-hacking-explain-the-election-results/). If you need a starting point, use [this Washington Post article](https://www.washingtonpost.com/graphics/politics/2016-election/exit-polls/) and [this R graph gallery](https://www.r-graph-gallery.com/) for ideas and inspiration.

Here, we used county election and census data to visualize the proportions of White, Hispanic, and Black individuals in a county and the corresponding election outcome. Using the demographic information, we can see if the proportion of a certain race of Americans in a county correlated with the county aligning with either the Democratic or Republican party.

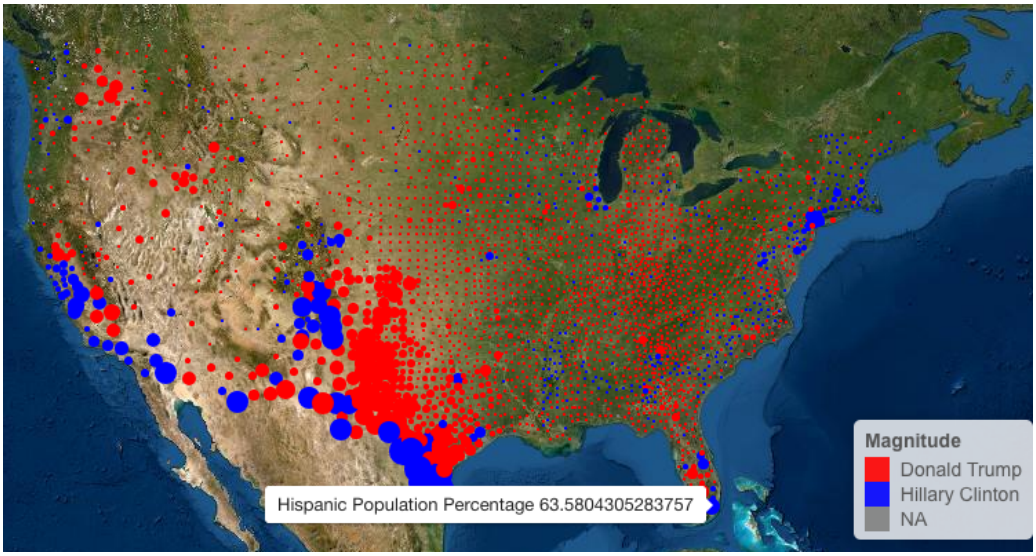
We visualized these proportions by creating a bubble graph where each point represents a county in the U.S. that we have election and census data for. The color is representative of what candidate that county voted for and the size is based on the relative proportion of citizens of a certain race in that county. This graph is also interactive and shows the numerical value for those proportions. An additional image on the next page is provided for reference to those values.

Please refer to the next page for the bubble map for the Black proportion:

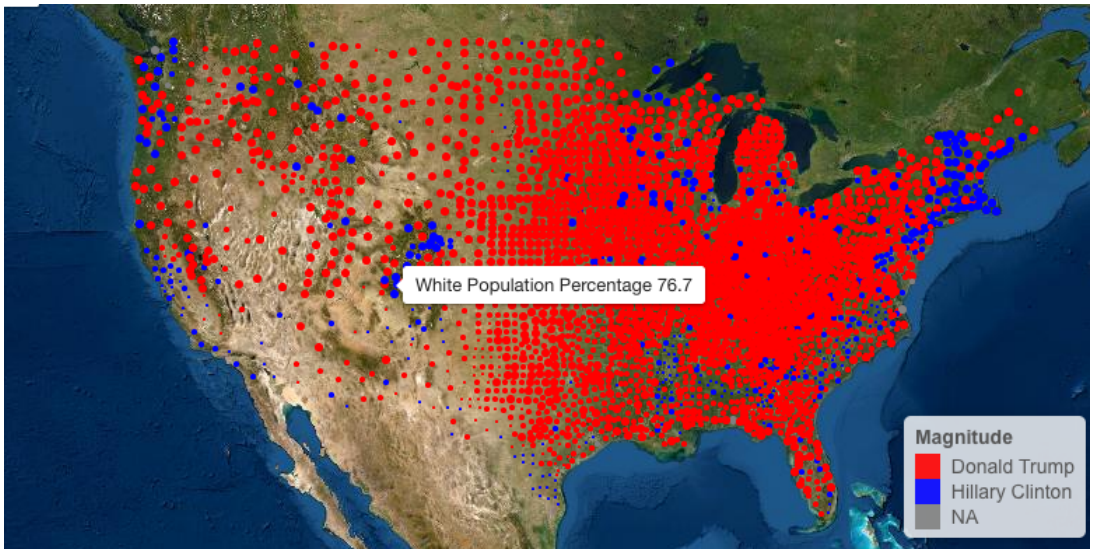




When analyzing the proportion of Black citizens within a county, we can see that, especially in the southeast portion of the country, counties who voted "blue" in the 2016 election primarily had a population that was over 50% Black. This does seem to align with the results of the election as according to the Washington Post graphic regarding exit polls, Black women and men voted blue at least 80% of the time, so it would make sense that counties with a higher Black population would be Democratic leaning.



In regards to the Hispanic percentage in a county, the same trend followed in the south western portion as well as in some parts of Florida and the north east of the United States in that it showed Democratic Party leanings in the 2016 election where higher Hispanic proportions were. There were various counties in the Texas area that had a relatively high proportion of Hispanic citizens, however voted Republican. This sort of "split" may align with exit poll information regarding Hispanic voters as they were less decisive, compared to Black voters, as at least 60% of Hispanic men and women voted for Hilary Clinton and the remaining approximate 30% voted for Donald Trump.



The White proportion in each county did not prove to be indicative of anything explanatory when it comes to voting outcome. Since white people consist of the main proportion in the majority of counties, the divide between whether a county voted "blue" or "red" was split more geographically compared to the amount of white citizens. According to exit poll information, White men and women voted for Donald Trump at least 60% of the time.

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 steps described on the handout.

After following the steps provided on the handout, the below data frame for census_ct was the output:

State <chr>	County <chr>	Men <dbl>	White <dbl>	Citizen <dbl>	Income <dbl>	IncomeErr <dbl>	IncomePerCap <dbl>	IncomePerCapErr <dbl>	Poverty <dbl>
Alabama	Autauga County	0.4843266	75.78823	0.7374912	51696.29	7771.009	24974.50	3433.674	12.91231
Alabama	Baldwin County	0.4884866	83.10262	0.7569406	51074.36	8745.050	27316.84	3803.718	13.42423
Alabama	Barbour County	0.5382816	46.23159	0.7691222	32959.30	6031.065	16824.22	2430.189	26.50563
Alabama	Bibb County	0.5341090	74.49989	0.7739781	38886.63	5662.358	18430.99	3073.599	16.60375
Alabama	Blount County	0.4940565	87.85385	0.7337550	46237.97	8695.786	20532.27	2052.055	16.72152
Alabama	Bullock County	0.5300618	22.19918	0.7545420	33292.69	9000.345	17579.57	3110.645	24.50260

ChildPoverty <dbl>	Professional <dbl>	Service <dbl>	Office <dbl>	Production <dbl>	Drive <dbl>	Carpool <dbl>	Transit <dbl>	OtherTransp <dbl>	WorkAtHome <dbl>
18.70758	32.79097	17.17044	24.28243	17.15713	87.50624	8.781235	0.09525905	1.3059687	1.8356531
19.48431	32.72994	17.95092	27.10439	11.32186	84.59861	8.959078	0.12662092	1.4438000	3.8504774
43.55962	26.12404	16.46343	23.27878	23.31741	83.33021	11.056609	0.49540324	1.6217251	1.5019456
27.19708	21.59010	17.95545	17.46731	23.74415	83.43488	13.153641	0.50313661	1.5620952	0.7314679
26.85738	28.52930	13.94252	23.83692	20.10413	84.85031	11.279222	0.36263213	0.4199411	2.2654133
37.29116	19.55253	14.92420	20.17051	25.73547	74.77277	14.839127	0.77321596	1.8238247	3.0998783

OtherTransp <dbl>	WorkAtHo... <dbl>	MeanComm... <dbl>	Employed <dbl>	PrivateWork <dbl>	SelfEmployed <dbl>	FamilyWork <dbl>	Unemployment <dbl>	Minority <dbl>	CountyPop <dbl>
1.3059687	1.8356531	26.50016	0.4343637	73.73649	5.433254	0.00000000	7.733726	22.53687	55221
1.4438000	3.8504774	26.32218	0.4405113	81.28266	5.909353	0.36332686	7.589820	15.21426	195121
1.6217251	1.5019456	24.51828	0.3192113	71.59426	7.149837	0.08977425	17.525557	51.94382	26932
1.5620952	0.7314679	28.71439	0.3669262	76.74385	6.637936	0.39415148	8.163104	24.16597	22604
0.4199411	2.2654133	34.84489	0.3844914	81.82671	4.228716	0.35649281	7.699640	10.59474	57710
1.8238247	3.0998783	28.63106	0.3619592	79.09065	5.273684	0.00000000	17.890026	76.53587	10678

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.

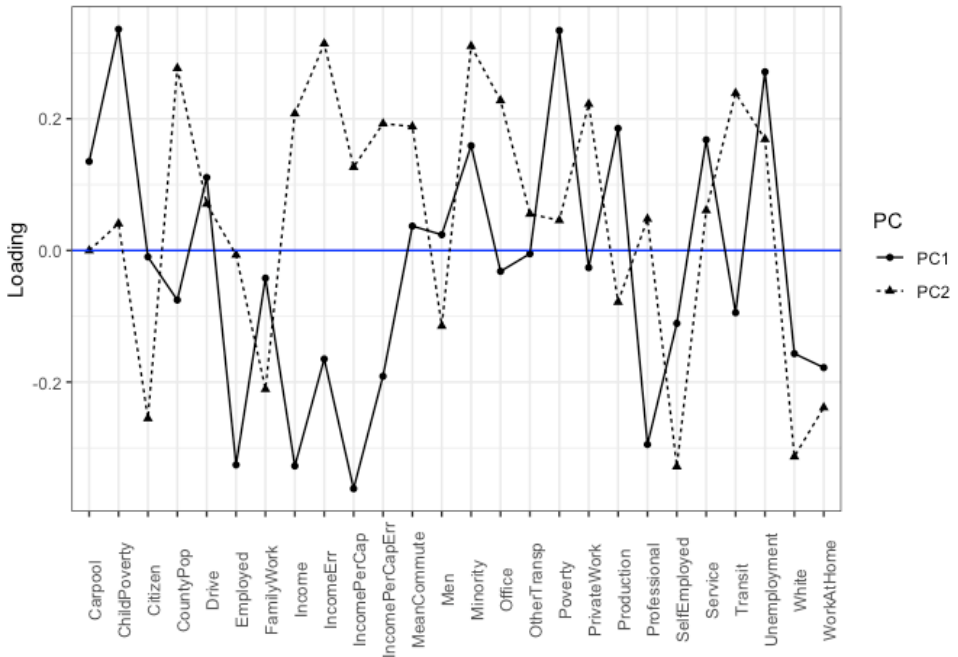
During the 2016 election, our group was in or near Santa Barbara County, which voted Democrat in the 2016 election. This aligned with the results of California and this is not surprising. Consisting of various cities with differing make ups to include Santa Barbara, Goleta, Santa Maria, and Montecito, this county encompasses an adequate sample of the California demographic. Similarly to California, Santa Barbara is essentially split in half between white and minority citizens. The average income is approximately 10,000 dollars higher than the state's average of 54,000 according to 2010 census data. The poverty rate is one percent higher than the average of California in 2010. Due to the similar make up of this county, the election results are not surprising.

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.

Due to the nature of the data, it is important to center and scale the data. As this is census information from the thousands of counties in the United States, we needed to adjust the data so that the variance is not incredibly high. This also makes it easier to work with. After conducting this important step, we created the following loading graphs for both County and Sub-County data.

County Data

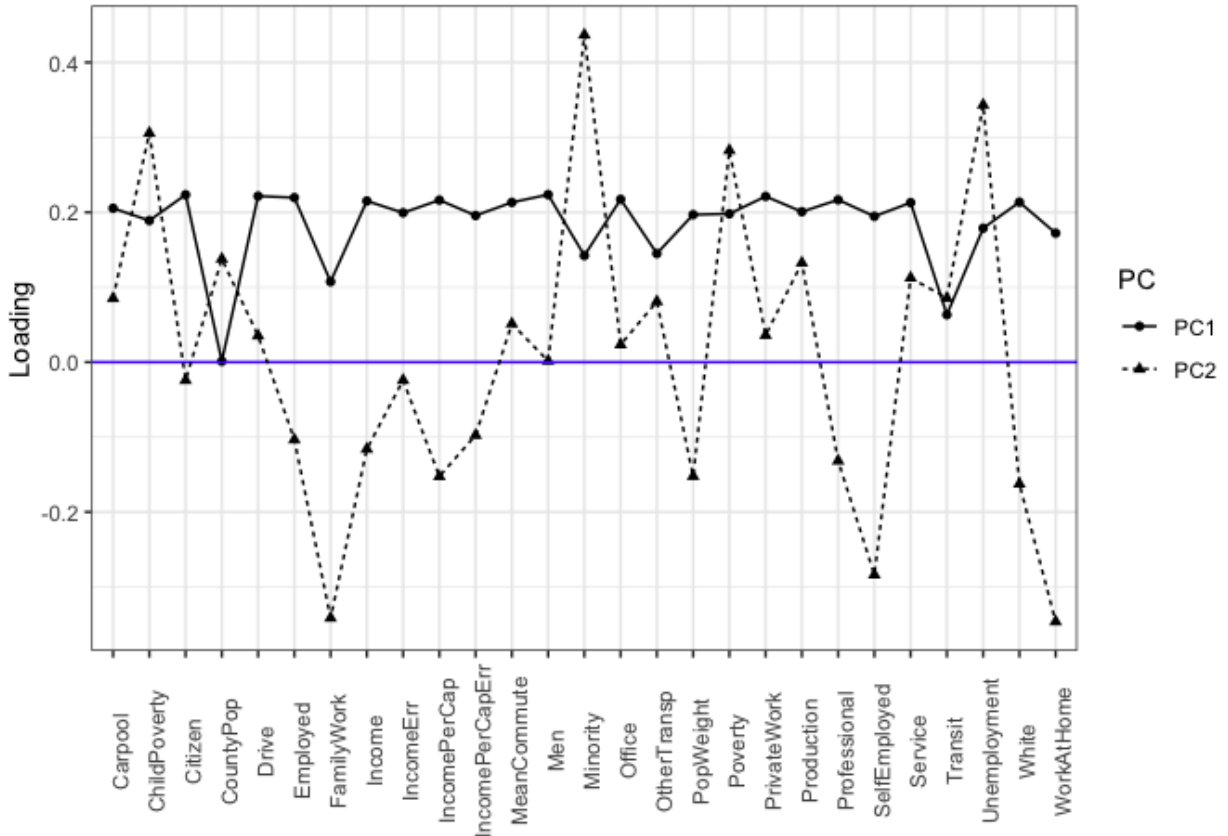


Upon observation of the principal components loading graphs, we can see that PC1 will be high when ChildPoverty, Poverty, and Unemployment are positive and when the following variables are negative: Employed, Income, IncomePerCapita, and Professional. This principal component seems to represent poorer, less economically developed areas of the country that suffer from high poverty and unemployment rates. This could be more rural and suburban areas.

Principal Component 2, on the other hand, will be high when the following variables are positive: CountyPop, Income, IncomePerCapita, Mintority, Office, Privatework while variables such as Citizen, Familywork, Selfemployed, and White are negative. This principal component seems to encapsulate high

performing economies with a predominant minority and non-citizen population such as the Los Angeles area.

Sub-County Data



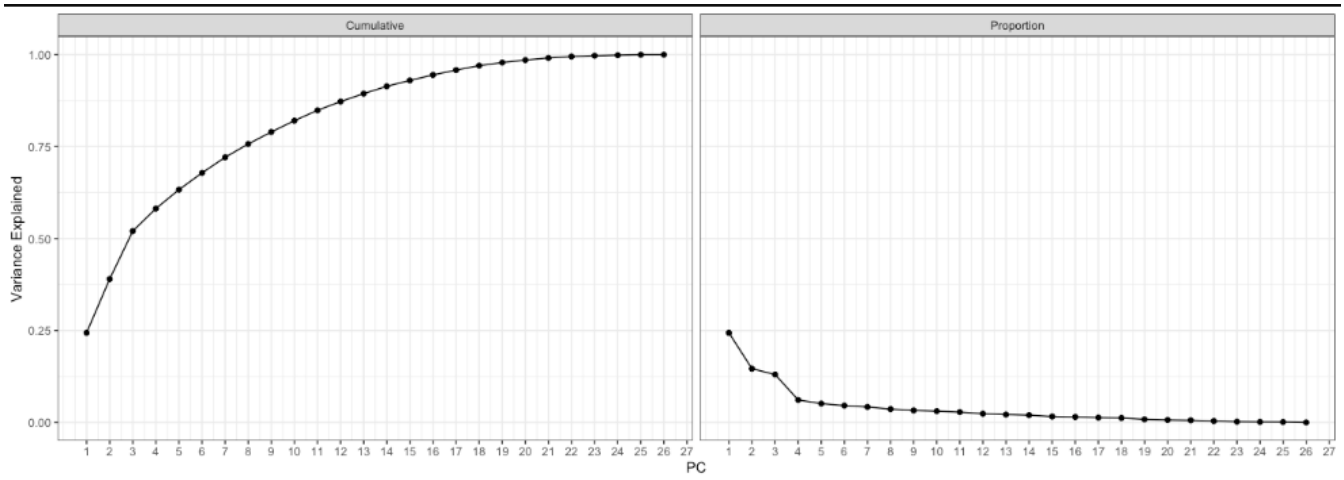
As we can see from the loadings graph above, principal component one has a very interesting shape where it does not appear that any variables are negative. The lowest values that are less positive than the others include CountyPop, and Transit. This could be representative of a less populated, less urban county.

Principal component 2 appears to have a more "dynamic" shape in that we can interpret it more specifically. This component will be high when variables to include ChildPoverty, Minority, Poverty, and Unemployment are positive and variables like FamilyWork and SelfEmployed are negative. This component appears to be representative of a poorer, possibly more rural cities in the country.

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.

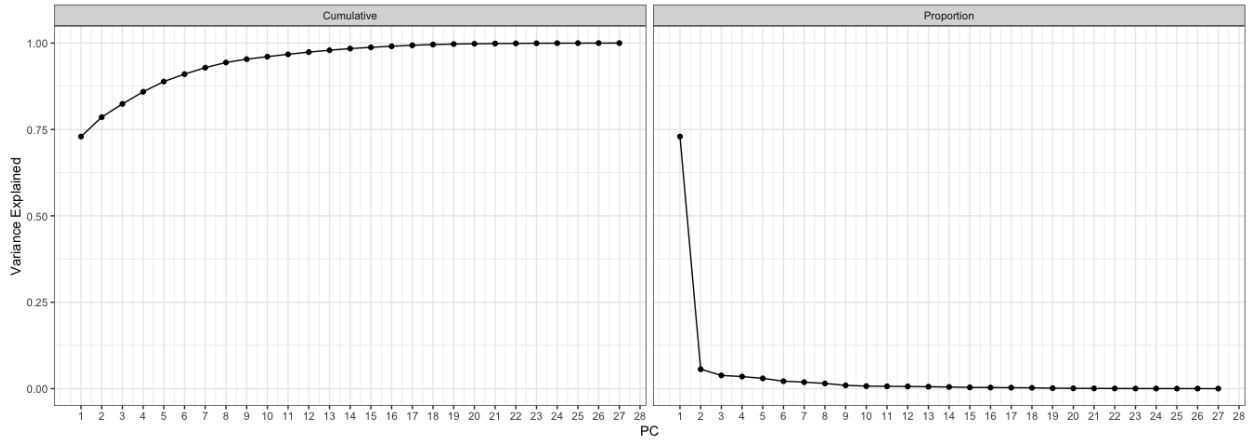
County Data:

From the below graph, we can see that we need approximately 15 principal components to explain approximately 90% of the variance.



Sub-County Data:

From the below graph, we can see that we need approximately 7 principal components to explain approximately 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.

After running hierarchical clustering with both methods as described above, we developed the following clusters. The first table lists the clusters for the conventional way of hierarchical clustering and the second lists the clusters for the clustering method performed with the first five PCAs:

	clusters <fctr>	n <int>
1	cluster 1	1098
2	cluster 2	1782
3	cluster 3	5
4	cluster 4	65
5	cluster 5	1
6	cluster 6	3
7	cluster 7	4
8	cluster 8	34
9	cluster 9	3
10	cluster 10	8

	clusters_pc <fctr>	n <int>
1	cluster 1	1703
2	cluster 2	697
3	cluster 3	126
4	cluster 4	252
5	cluster 5	150
6	cluster 6	19
7	cluster 7	8
8	cluster 8	17
9	cluster 9	30
10	cluster 10	1

We found that San Mateo County was placed into the cluster 1 and cluster 6 respectively and the first few rows of those clusters can be seen below:

county <chr>	candidate <chr>	state <chr>	total <dbl>	Men <dbl>	White <dbl>	Citizen <dbl>	Income <dbl>	IncomeErr <dbl>	IncomePerCap <dbl>
Autauga County	Donald Trump	AL	24759	0.4843266	75.78823	0.7374912	51696.29	7771.009	24974.50
Baldwin County	Donald Trump	AL	94261	0.4884866	83.10262	0.7569406	51074.36	8745.050	27316.84
Elmore County	Donald Trump	AL	36927	0.4873766	73.65151	0.7574260	54060.50	8684.428	24380.78
Jefferson County	Hillary Clinton	AL	299840	0.4727901	50.99850	0.7438371	52151.09	8348.166	27239.89
Limestone County	Donald Trump	AL	39792	0.5025168	77.41039	0.7441698	52657.32	9155.641	25568.90
Madison County	Donald Trump	AL	160217	0.4890399	65.61261	0.7470052	64645.12	9039.880	32131.20
Morgan County	Donald Trump	AL	50126	0.4909923	76.80766	0.7311622	47741.19	6711.594	24449.43
St. Clair County	Donald Trump	AL	38054	0.5024341	86.28356	0.7604002	51124.35	9512.729	24167.63
Shelby County	Donald Trump	AL	99442	0.4870732	78.81874	0.7215153	73433.95	10859.706	33493.52
Benton County	Donald Trump	AR	96824	0.4944668	75.01484	0.6577637	59216.79	8535.292	27934.22

1–10 of 1,098 rows | 1–10 of 31 columns Previous 1 2 3 4 5 6 ... 100 Next

The cluster containing San Mateo County that was produced through conventional hierarchical clustering has over one thousand counties that vary in many categories to include the candidate that they voted for, size, income, etc. This does not appear to be a very focused cluster for the given county.

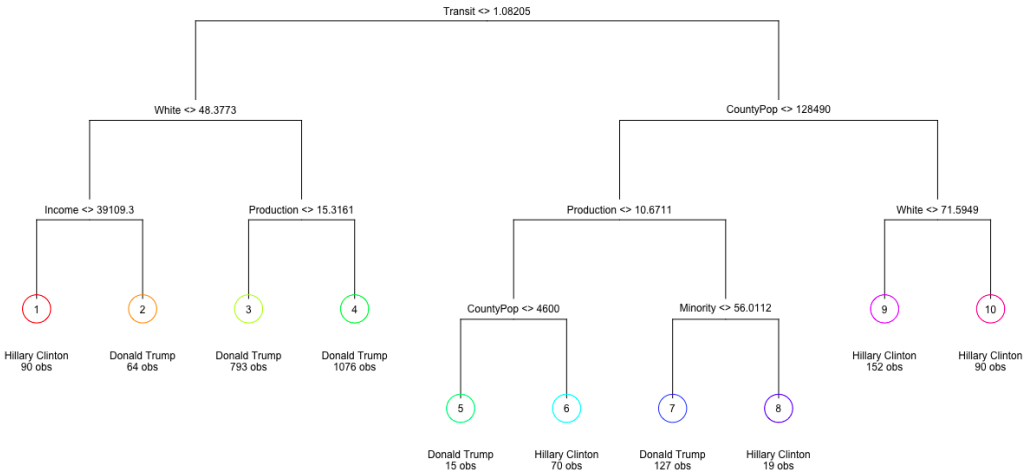
county <chr>	candidate <chr>	state <chr>	total <dbl>	Men <dbl>	White <dbl>	Citizen <dbl>	Income <dbl>	IncomeErr <dbl>	IncomePerCap <dbl>
Alameda County	Hillary Clinton	CA	648662	0.4900514	32.97244	0.6473888	83129.49	12634.54	37299.07
Contra Costa County	Hillary Clinton	CA	461569	0.4883284	45.82503	0.6564452	89623.17	13784.88	39265.13
Marin County	Hillary Clinton	CA	138118	0.4827963	72.72104	0.7002243	98924.65	17537.96	60992.69
San Francisco County	Hillary Clinton	CA	403358	0.5089265	41.25369	0.7358313	85425.17	14863.15	52230.87
San Mateo County	Hillary Clinton	CA	311424	0.4919773	40.63851	0.6420050	100369.92	16123.02	47881.29
Santa Clara County	Hillary Clinton	CA	697258	0.5026387	33.58126	0.6056144	100743.85	15214.63	43879.60
Eagle County	Hillary Clinton	CO	25223	0.5304131	67.03083	0.6283475	75660.94	19087.42	38193.28
Pitkin County	Hillary Clinton	CO	10523	0.5301378	86.29220	0.7598163	72835.73	19605.62	55518.70
Fairfield County	Hillary Clinton	CT	420486	0.4866530	63.21930	0.6552812	96820.23	16314.62	47741.97
Montgomery County	Hillary Clinton	MD	468674	0.4818740	46.15483	0.6210157	109337.05	15105.55	48416.14

Clearly, it appears that the cluster using the first 5 principal components did a much better job at grouping San Mateo County with a smaller, more focused group. All but one county voted for Hilary Clinton and they appear to have similar populations and income levels. There are also many counties included that are within a very close proximity of San Mateo County. Since the PCA method is aimed at reducing the amount of variables and maintaining as much information as possible from the dataset, it can bear a much more descriptive and appropriate cluster for the given county.

CLASSIFICATION

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]
(https://archive.nytimes.com/www.nytimes.com/imagepages/2008/04/16/us/20080416_OBAMA_GRAPHIC.html)).

```
Classification tree:
snip.tree(tree = t_0, nodes = c(8L, 14L, 26L, 11L, 25L, 15L,
9L, 10L))
Variables actually used in tree construction:
[1] "Transit"    "White"      "Income"     "Production" "CountyPop"  "Minority"
Number of terminal nodes:  10
Residual mean deviance:  0.393 = 976.9 / 2486
Misclassification error rate: 0.07131 = 178 / 2496
                preds
actual           Donald Trump Hillary Clinton
  Donald Trump            483              39
  Hillary Clinton          22              79
                preds
actual           Donald Trump Hillary Clinton
  Donald Trump      0.92528736      0.07471264
  Hillary Clinton   0.21782178      0.78217822
```



Based on the above developed pruned classification tree, we can see that the important variables in predicting voter behavior include: transit, county population, white proportion in county, minority, and production. What we can see is that voters in urban areas with a high county population and high white proportion were predicted to vote for Hilary Clinton. On the other hand, we can see that voters from less urban areas (lower public transportation rates), with a higher white proportion, and proportion of employment in production were predicted to vote for Donald 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?

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.296e+01  9.111e+00  -1.422 0.154971
Men              7.051e+00  4.771e+00   1.478 0.139494
White           -1.869e-01  6.612e-02  -2.827 0.004693 **
Citizen          1.047e+01  2.837e+00   3.690 0.000225 ***
Income          -6.576e-05  2.579e-05  -2.550 0.010769 *
IncomeErr        9.670e-06  6.243e-05   0.155 0.876908
IncomePerCap     2.870e-04  6.417e-05   4.472 7.74e-06 ***
IncomePerCapErr -3.619e-04  1.342e-04  -2.697 0.007005 **
Poverty          6.039e-02  3.601e-02   1.677 0.093521 .
ChildPoverty    -8.378e-03  2.292e-02  -0.366 0.714732
Professional     2.017e-01  3.583e-02   5.630 1.80e-08 ***
Service          2.876e-01  4.526e-02   6.354 2.10e-10 ***
Office           9.440e-02  4.502e-02   2.097 0.036014 *
Production       1.423e-01  4.066e-02   3.499 0.000468 ***
Drive           -2.115e-01  4.276e-02  -4.946 7.59e-07 ***
Carpool         -1.796e-01  5.730e-02  -3.135 0.001720 **
Transit          3.661e-02  8.846e-02   0.414 0.679016
OtherTransp     -9.802e-02  8.917e-02  -1.099 0.271642
WorkAtHome      -1.230e-01  7.001e-02  -1.756 0.079019 .
MeanCommute      3.904e-02  2.340e-02   1.668 0.095271 .
Employed         1.776e+01  3.138e+00   5.659 1.52e-08 ***
PrivateWork      5.956e-02  2.070e-02   2.878 0.004006 **
SelfEmployed    -3.689e-02  4.881e-02  -0.756 0.449822
FamilyWork      -1.180e+00  4.068e-01  -2.901 0.003722 **
Unemployment     1.818e-01  3.771e-02   4.822 1.42e-06 ***
Minority        -6.189e-02  6.387e-02  -0.969 0.332532
CountyPop        4.025e-07  3.999e-07   1.007 0.314145
```

```
                preds
actual           Donald Trump Hillary Clinton
  Donald Trump             470              52
  Hillary Clinton           14              87
                preds
actual           Donald Trump Hillary Clinton
  Donald Trump      0.90038314      0.09961686
  Hillary Clinton   0.13861386      0.86138614
```

Based on the above summary output the important variables include: White, Citizen, IncomePerCap, IncomePerCapErr, Poverty, Professional, Service, Office, Production, Drive, Carpool, Employed, PrivateWork, FamilyWork, and Unemployment. These are similar to important variables that were developed for the decision tree. However, it seemed to be more focused on commute time and other occupations besides production. The logistic regression model could tell us how more specific information relates to the election outcome.

(Intercept)	Men	White	Citizen	Income	IncomeErr	IncomePerCap	IncomePerCapErr
0.000000e+00	1.153693e+03	8.295000e-01	3.508244e+04	9.999300e-01	1.000010e+00	1.000290e+00	9.996400e-01
Poverty	ChildPoverty	Professional	Service	Office	Production	Drive	Carpool
1.062250e+00	9.916600e-01	1.223490e+00	1.333220e+00	1.099000e+00	1.152870e+00	8.093800e-01	8.355800e-01
Transit	OtherTransp	WorkAtHome	MeanCommute	Employed	PrivateWork	SelfEmployed	FamilyWork
1.037280e+00	9.066300e-01	8.843000e-01	1.039810e+00	5.145563e+07	1.061370e+00	9.637900e-01	3.072600e-01
Unemployment	Minority	CountyPop					
1.199420e+00	9.399900e-01	1.000000e+00					

For example, if we observe the exponentiated intercept value for service, 1.33, we can interpret this as meaning for every one percentage increase in proportion of people having service jobs in a county, there is a 33 percent increase in the odds of that county voting for Hilary Clinton (which is represented by the value 1).

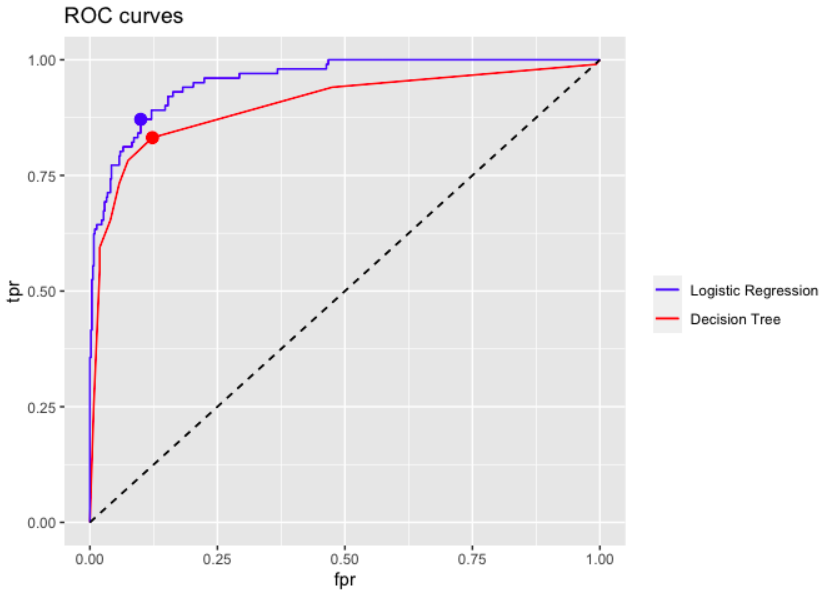
Similarly, if we observe the exponentiated intercept value for production,1.15, we can interpret this as meaning every one percentage increase in the proportion of production workers is associated with an increase in the odds of voting for Hilary Clinton by 15 percent.

```
SB_pred <- predict(lr_fit, Santa_Barbara, type="response")
SB_pred
```

```
        1
0.9054006
```

As we can see, the logistic regression model appeared to almost precisely predict the outcome of the 2016 election for Santa Barbara 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?



From the above graph, the probability threshold for both the decision tree and logistic regression model had to be adjusted to 0.709 and 0.20 respectively.

In comparing the accuracy of the two methods, we can see that both did fairly well at predicting the appropriate candidate, with the decision tree predicting a county would vote for Trump at a rate of 92.6% and Clinton at a rate of 78.2%, and the logistic regression model predicting a Trump vote at 90.0% and a Clinton vote at 86.1%. The logistic regression model did improve upon the TPR for Clinton votes by a significant amount.

Regarding interpretability, both can be utilized in different ways to gain more insight on the 2016 election outcome. The pruned decision tree allows for one to follow a path and see how multiple variables about a county lead to a either a red or blue vote Though this is a simplistic and helpful way to understand voting behavior, it does appear to be limiting in that the most important variables are not specific enough to get a solid idea on how and why certain counties voted the way they did. Visually, it is very easy to understand but lacks depth needed for further analysis.

The logistic regression model, on the other hand, offers a slightly more in-depth interpretation of how much and in what direction different independent variables affect the odds of the binary response. Another benefit to this method is that we can construct a model that can be applied in order to predict the outcome, as we did for Santa Barbara county. The drawbacks to this method include that its interpretation is not as straightforward as the decision tree and that we are not able to get a sense for how independent variables affect one another.

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!

From this in depth analysis on census and voting data, we can definitely see that predicting voting behavior and alignments is a detailed task with many different considerations. Demographics play a vital role in many people's political affiliations, however there are limitations to how much detail we can extract from census data alone. Though the decision tree and logistic regression model made for streamlined and simple ways to gain some insight on how this information could be useful in prediction, their interpretability lacks in some areas as well as their precision. Both models had noticeably higher misclassification rates when predicting a county voting for Hilary Clinton.

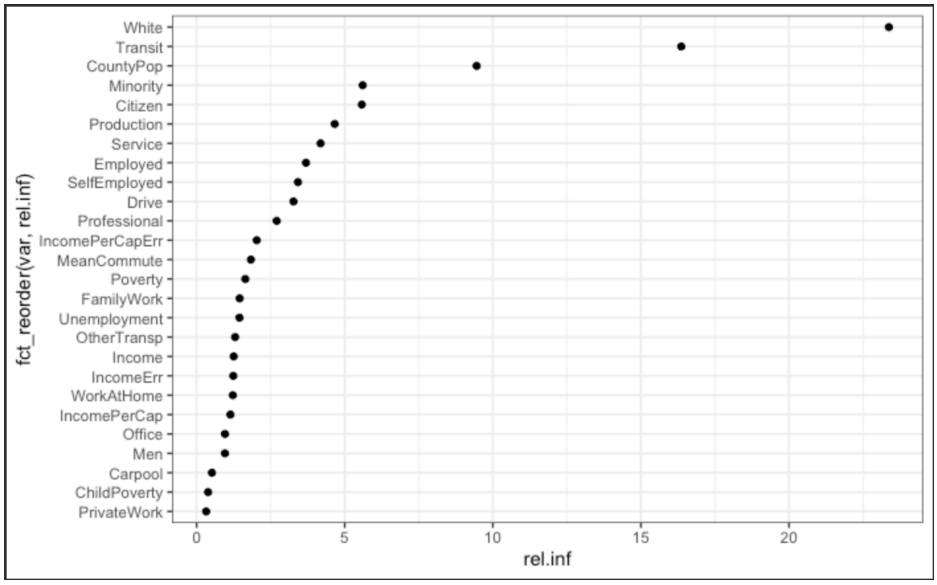
Due to these observations, we wanted to see how the results compared to other models we went over in class to include the random forest and boosted models and those outputs can be seen below.

```
Call:
 randomForest(formula = candidate ~ ., data = train, ntree = 100,      mtry = 5, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 100
No. of variables tried at each split: 5

        OOB estimate of  error rate: 6.53%
Confusion matrix:
               Donald Trump Hillary Clinton class.error
Donald Trump           2046             55  0.02617801
Hillary Clinton         108            287  0.27341772
```

For starters,a random forest model was fit and had a small misclassification error rate when predicting counties voting for Donald Trump, however had a high error rate of 27.4% in regards to Hilary Clinton. When analyzing the most important variables, we saw that many of the same are shared with the prior two methods to include White, Minority, Transit, CountyPop, and IncomePerCap.

```
gbm::gbm(formula = I(as.numeric(train$candidate) - 1) ~ ., distribution = "adaboost",
    data = train, n.trees = 100, interaction.depth = 3, shrinkage = 0.2,
    train.fraction = 0.8, cv.folds = 5, n.cores = 1)
A gradient boosted model with adaboost loss function.
100 iterations were performed.
The best cross-validation iteration was 51.
The best test-set iteration was 7.
There were 26 predictors of which 8 had non-zero influence.
[1] 51
                predicted
truee            Donald Trump Hillary Clinton
  Donald Trump      0.96917148      0.03082852
  Hillary Clinton   0.26923077      0.73076923
[1] 0.070626
```



We also fit a boosted model to the data and it resulted in similar outcomes as the other three models. This model was very precise in predicting counties for Donald Trump yet again had a high misclassification error rate when predicting a county's vote for Hilary Clinton. From the above graph, we can see that influential predictors are similar to that of the prior models and include White, Minority, Transit, and CountyPop.

These important variables that are shared throughout the models tested show that race played a role in voting behavior as well as how urban and large a county is, among other things. From the actual county voting results displayed prior in addition to the bubble maps visualizing county racial proportions , we can see that the results of our fitted models followed suit with voting behavior in 2016, to some extent. However, while these variables show that census data can serve as a strong indicator of a county's voting preference, no method we used can accurately account for all possible errors that may come as a result of the intricacies of human behaviour.

Given that four methods were conducted and each had noticeably more error in predicting a county's vote for Hilary Clinton, with logistic regression performing the best, it is difficult to concoct a solution other than possibly fitting more advanced models outside the realm of this class.

The dissonance in accurate classification can speak to the complexity of the 2016 Election and predicting results. As discussed early on in this project, erroneous polling was a major factor in the 2016 Election for numerous reasons and though these models were fit on a different assortment of data, the complex nature of predicting voting behavior is very apparent.