The Influence Of Population, Voting Turnout and Ethnicity on voting outcomes

Final project STAT-312

Malachy McGovern

20/04/2021

Introduction

Politics in recent years has been turbulent to say the least. With the rise of populist politics, American society has become much more polarized where the opposing perspective has become much more unacceptable than years past. This political tension was brought to its pinnacle at the 2020 elections, which had the largest turnout in history. However, one of the main stories in that election was the potential for new swing states to come to light. The state of Georgia for example, was last blue in 1992 and had such a high percentage of democrat votes since 1980. In this case, many have claimed that the increased turnout in voters and voters from minorities was pivotal in these changes. This is at least seen by the Republican party, who since the election, have scrambled to pass stricter voting laws. This would seek to limit turnout especially from those minority groups.

This raises the theory that increased turnout favors democrats in elections. In this project I seek to test that theory and see the conditions that this may be true for counties and states across the USA. Before this, I will look to identify the political map of America and get a sense of the distribution of votes and turnout for regions of the country. I will also, implement an alternate electoral system which allocates electoral votes by percentage within the state rather than the current winner-takes all system. Finally, I will generate a model to predict the next swing counties and thus swing states using support vector machines. However, this model will rely solely on the factors of raw votes cast, turnout and the percentage distribution of votes. In doing so, I will be putting into practice the aforementioned theory by using turnout and ethnic voting as a predictor.

Hypotheses

1 - Voter turnout and party outcome

null: There will be no significant relationship between the voter turnout and difference of democrat votes to republican *Alternative:* There will be a significant relationship between voter turnout and difference of democrat votes to republican

2 - Ethnicity and party outcome

null: There will be no significant relationship between the percentage of ethnic minorities and percentage of democrat votes Alternative : There will be a significant relationship between the percentage of ethnic minorities and percentage of democrat votes

Data setup

In the chunk below I have called various packages from my library which I will use in this project. I have also set the chunks to not produce warnings or messages when knitting or running.

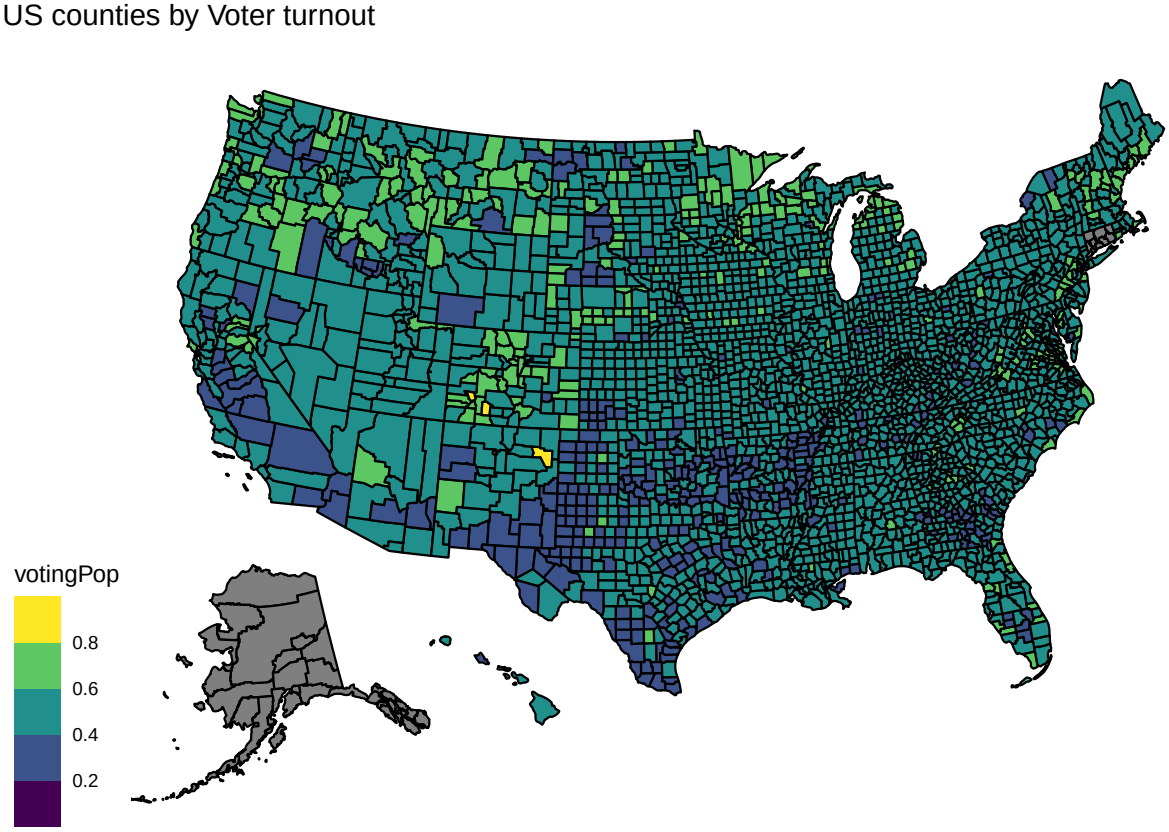
In the following chunk I have read the .csv file containing all the various statistics surrounding the counties in the USA. From there I have randomly sampled the data to produce a training, querying and testing sample. These files were then written up to form their own individual .csv file.

Exploratory analysis

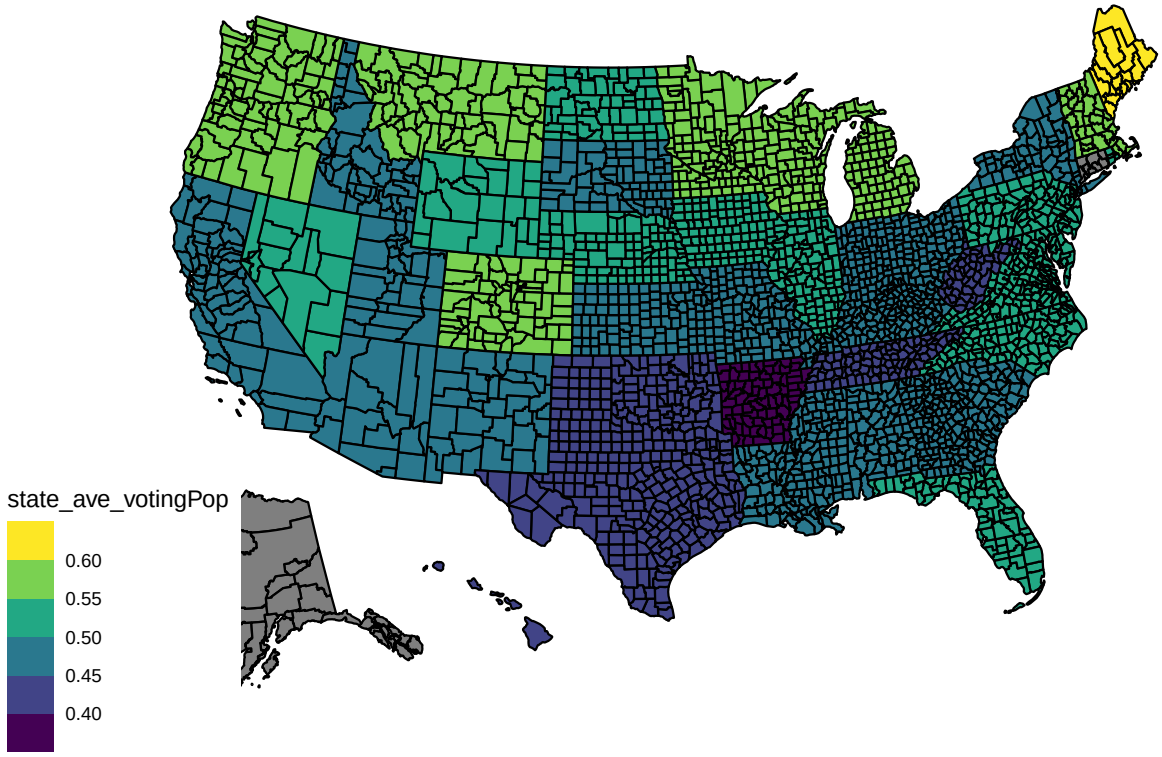
Turnout by county and state

Here marks the procedure to map the voting distributions around the country. This uses the original data as it is purely an observational stage of the project. Furthermore, by using the training data set there will be a lot of empty counties in the map which will limit the value of the visualization.

This first map gives a image of the voter turnout for each county across the US.Generally there the turnout seems to range between 40% and 60% across the country. The second plot shows the state level of this variable which verified the previous claim on turnout range.



US counties by Voter turnout

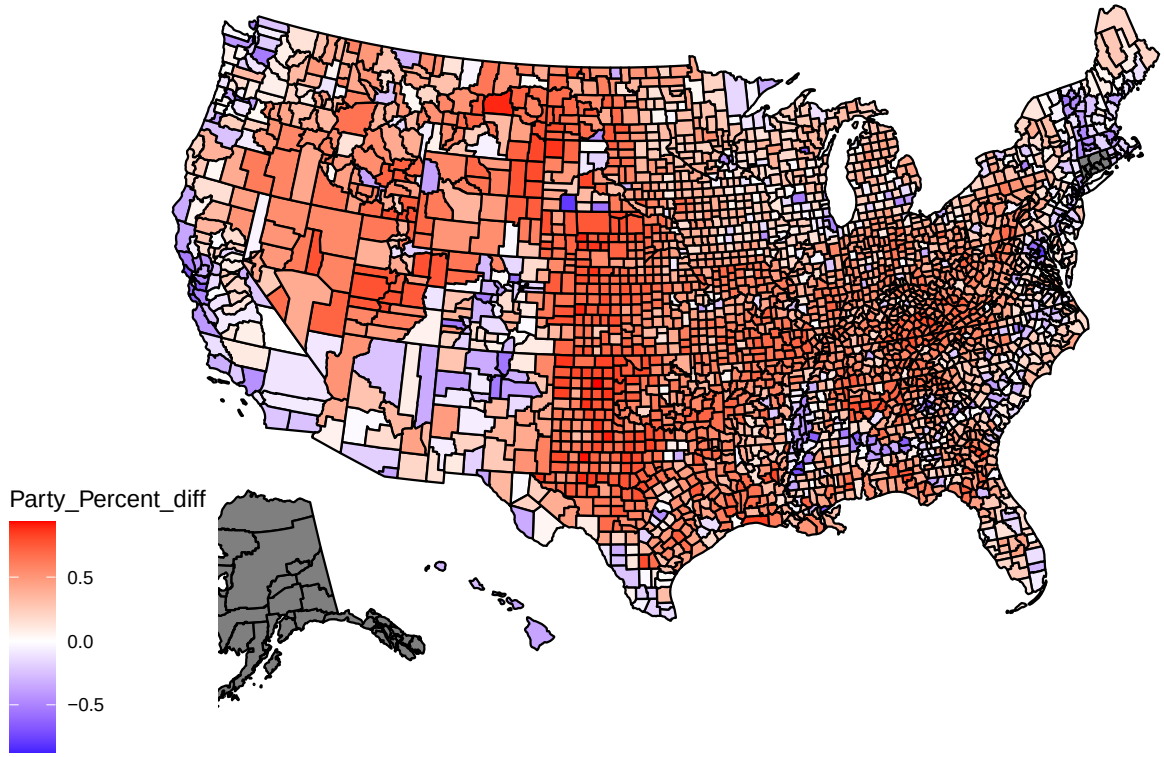


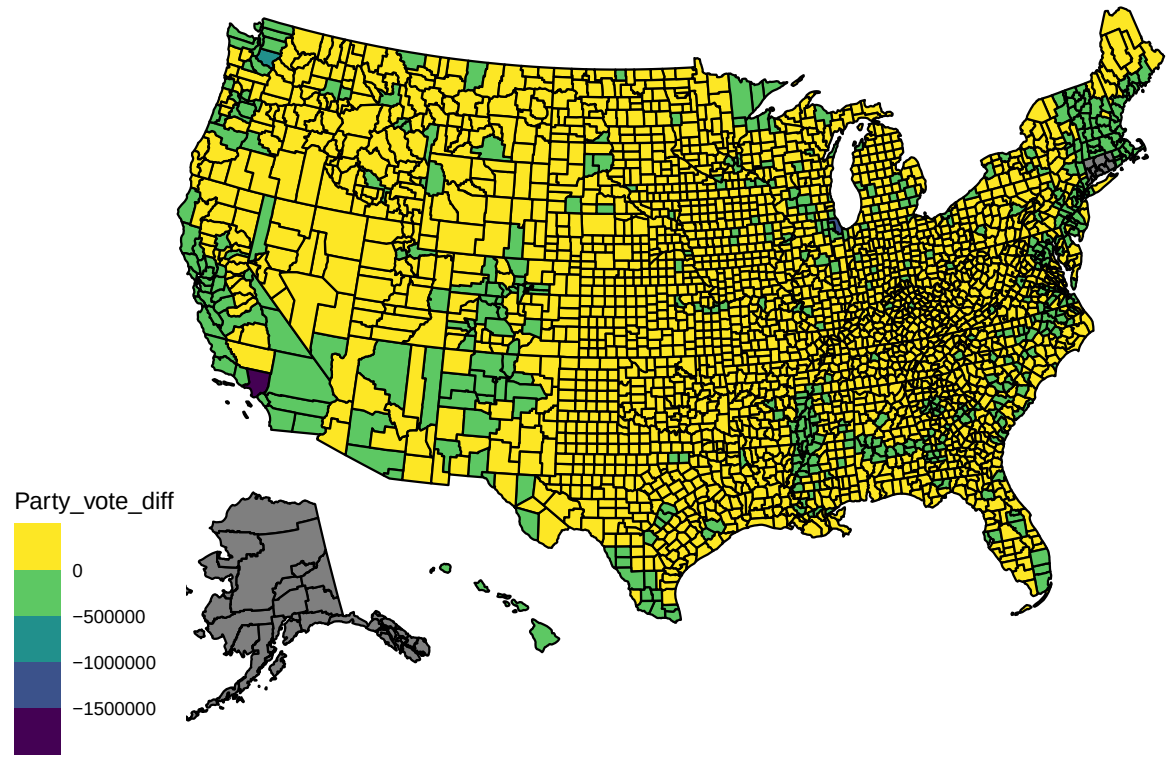
Statewise voting results

In comparison these maps below begin to describe in real and relative terms the differences in voting. The first map gives the difference in percentage votes of republicans against democrats. Here the graph is colored on a scale of red to blue with white marking the 50/50 point, which here is 0. The intensity of the color indicates the strength of win by either party.

The second graph gives the raw voting difference between the two parties. Here it is not colored by blue or red as it fails to distinguish wins as effectively. What can be observed here is that democrats have extraordinary high votes in certain counties, notably in LA, Chicago and some regions in the east coast. The likelihood here is that these are all urban counties which have the concentrations of people exponentially larger than other counties. Given that these are democrat wins, it may explain why they tend to win the popular vote.

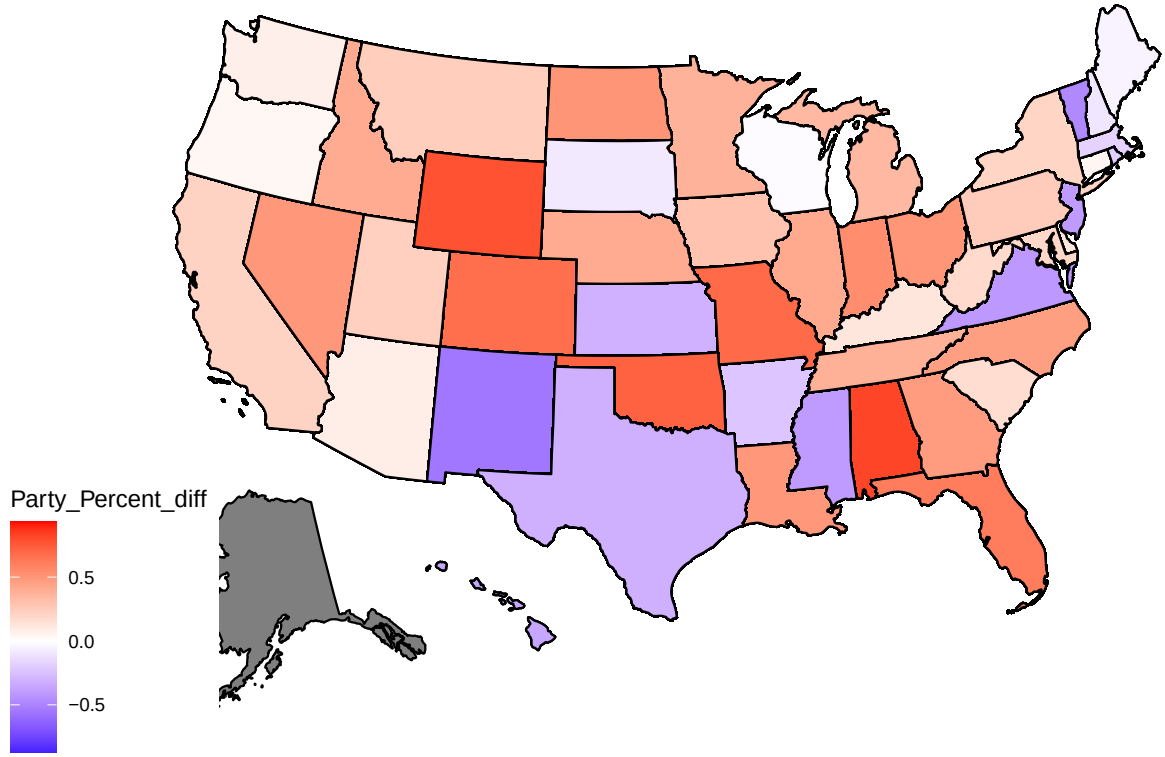
US counties by Party Percent





In this instance, the map describes these percentage and raw differences by state rather than counties. A similar trend can be observed in the raw voe difference.

US State by Party Percent



Winning elections

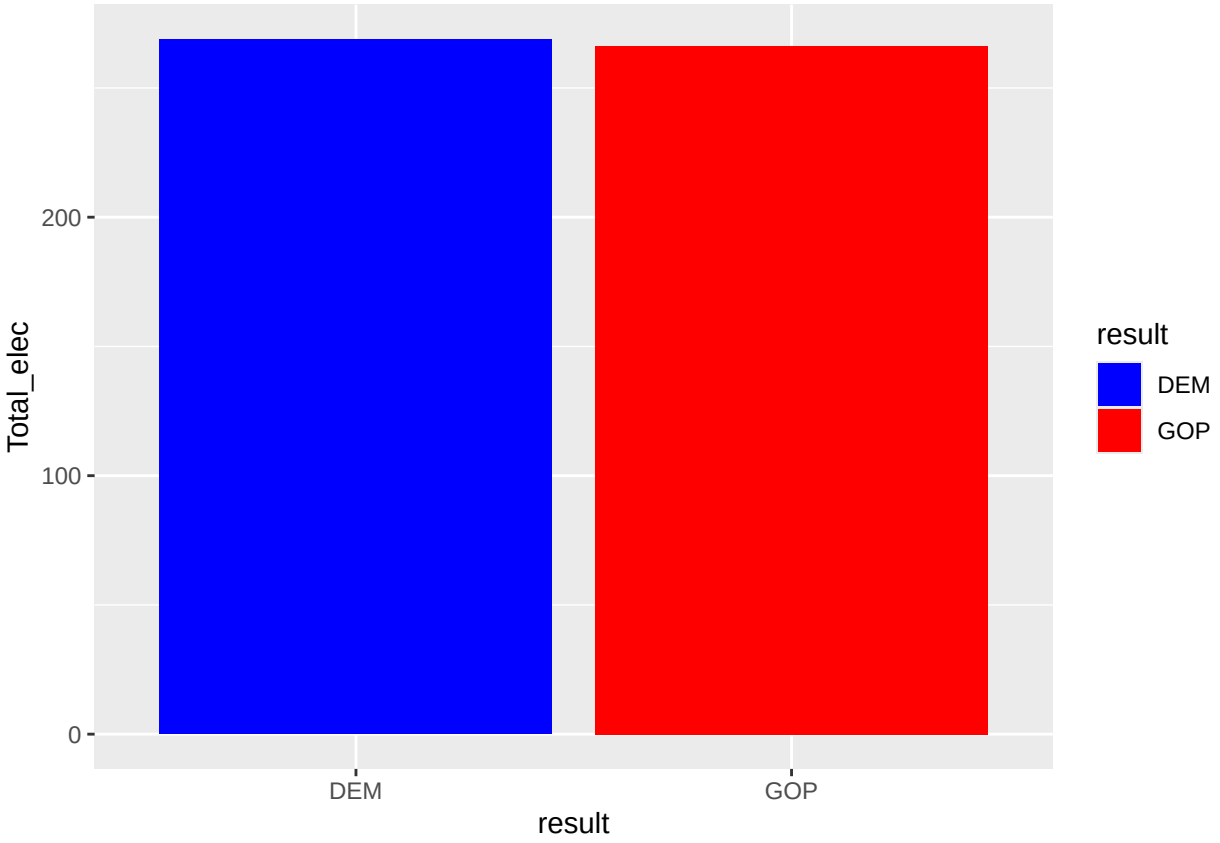
Adding the electoral college variable and a state averages

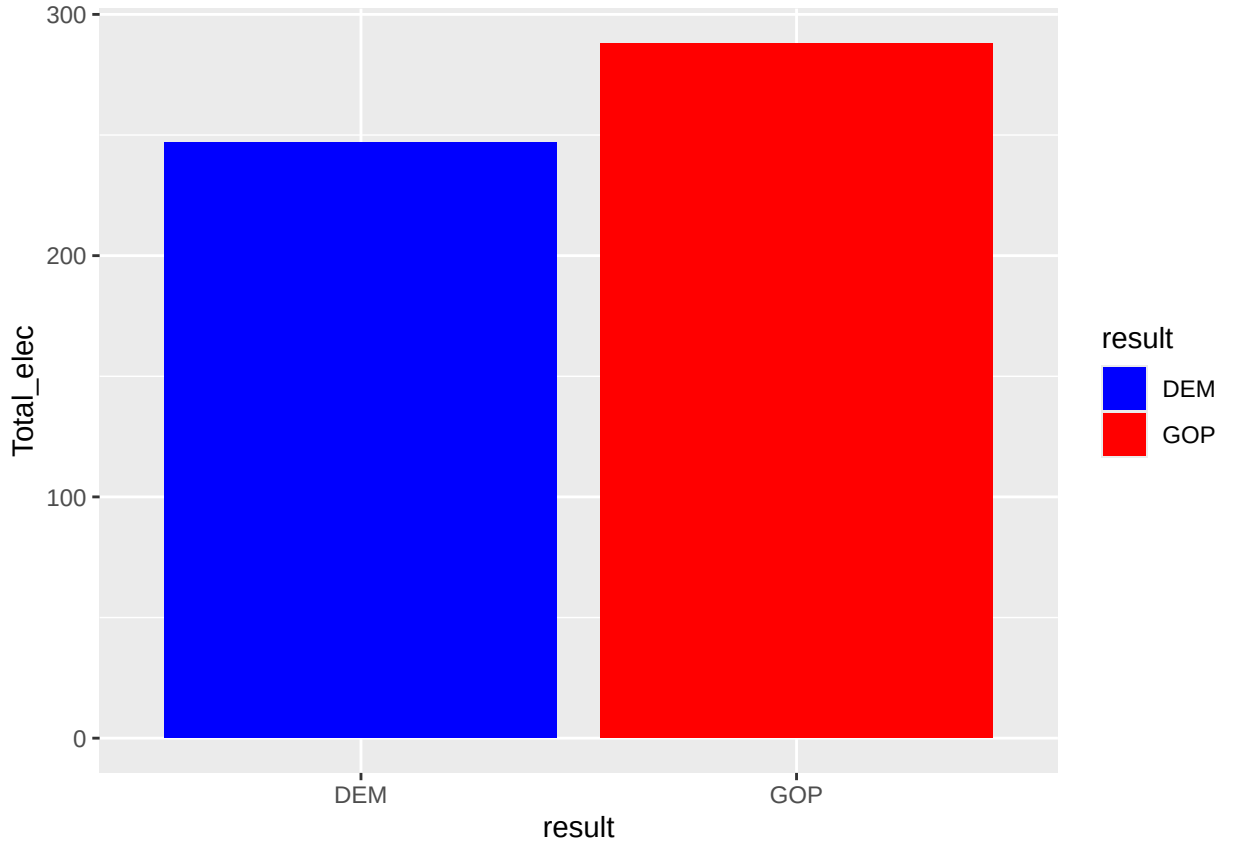
Below is the start of constructing the number of electoral votes for each party based on results. The chunk below involves the process of adding the number of electoral college votes for each state.

```
## [1] 535
```

Below I have constructed a new data frame which gives the average voting statistics for each state. This involves the total votes for each party, the percentages for each party and the difference in votes between the two. I have also added variables giving the proportion of the US population of each state. I have then added the electoral votes as a variable and then used it to construct a variable allocating electoral votes by the proportion of the US population each state holds.

I have then used this information to generate the outcomes of an election based on normal electoral votes or the adjusted electoral votes.


Total_elec
200
100
0
DEM
GOP
result
result
DEM
GOP



Applying the Proportional Electoral Vote system

Here electoral votes are allocated fractionally based on the proportion that voted in favor of the party. In this system all parties receive the number of electoral votes proportional to their votes received.

The outcome of the election by proportional electoral votes is below.

```
## # A tibble: 2 x 2
## # Groups:   Winner [2]
##   Winner     n
##   <chr>  <int>
## 1 DEM       21
## 2 GOP       29
```

```
## # A tibble: 1 x 2
##     GOP   DEM
##   <dbl> <dbl>
## 1  256.  269.
```

Regressions on election results

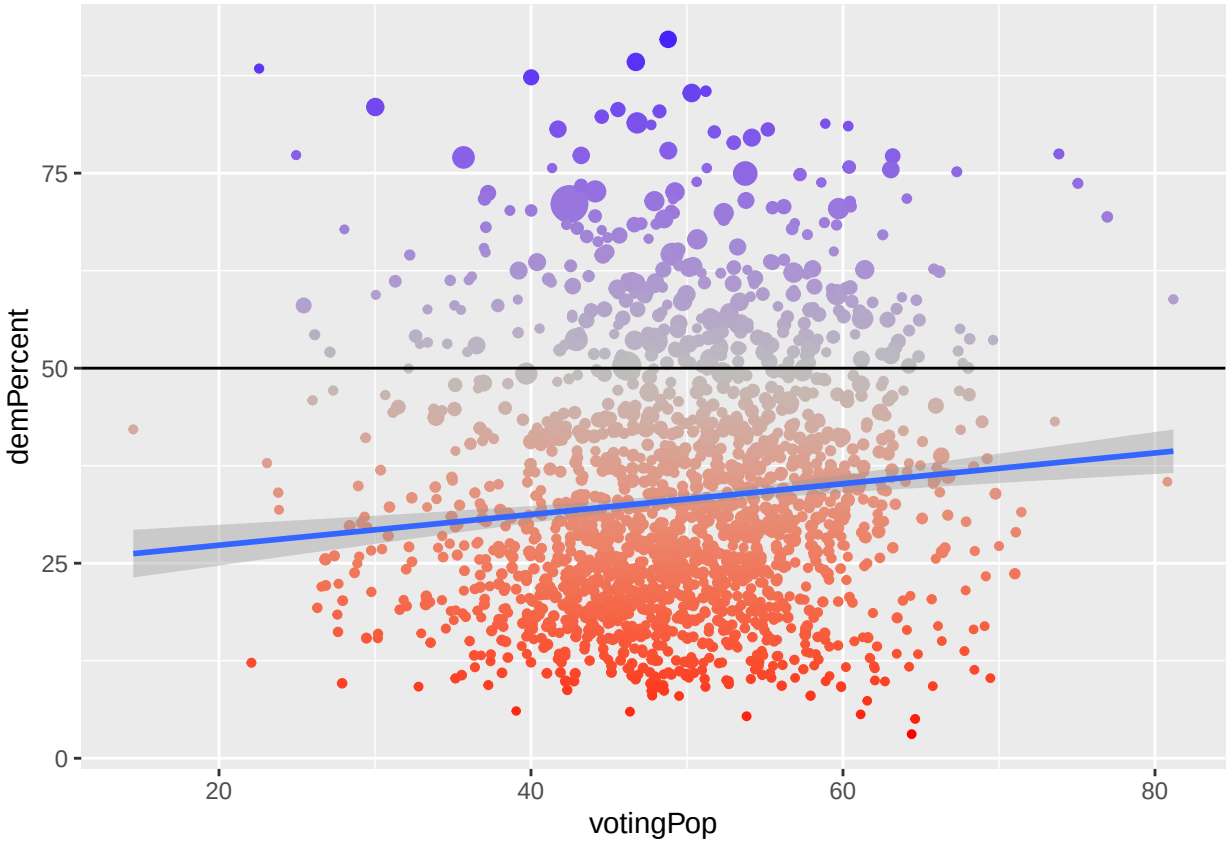
Voting turnout and democrat percent

The following section focuses on testing the hypothesis surrounding the relationship between turnout and democrat votes. This process began with the construction of the data frame 'VotingTibble. This data frame contains observations from counties with the; proportion of ethnicities; the population and voting statistics.

Using this data frame a linear model was constructed between the percentage of votes that were democratic and the voting turnout of each county. This model was found to have a significant relationship between the two variables where every increase of 10% of voter turnout led to an increase of 0.04% for democrats. This was found to be statistically significant where p = 0.000289. Whilst this relationship was found to be significant, the adjusted r-square value is quite small at 0.006489. This means that a marginal amount of variance of is explained by the voting turnout of the county.

A graphical plot of this relationship was also generated below which colors observations by the proportion of votes that were democrat. It also constructs sizes of plots based off the size of the total votes cast at the county of observation.

```
##
## Call:
## lm(formula = votingPop ~ demPercent, data = VotingTibble)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.534  -5.066   0.028   5.708  31.135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.69290    0.44920 106.173  < 2e-16 ***
## demPercent   0.05564    0.01224   4.546 5.81e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.356 on 1865 degrees of freedom
## Multiple R-squared:  0.01096,    Adjusted R-squared:  0.01043
## F-statistic: 20.67 on 1 and 1865 DF,  p-value: 5.808e-06
```



Democratic vote percent and white percentage of population

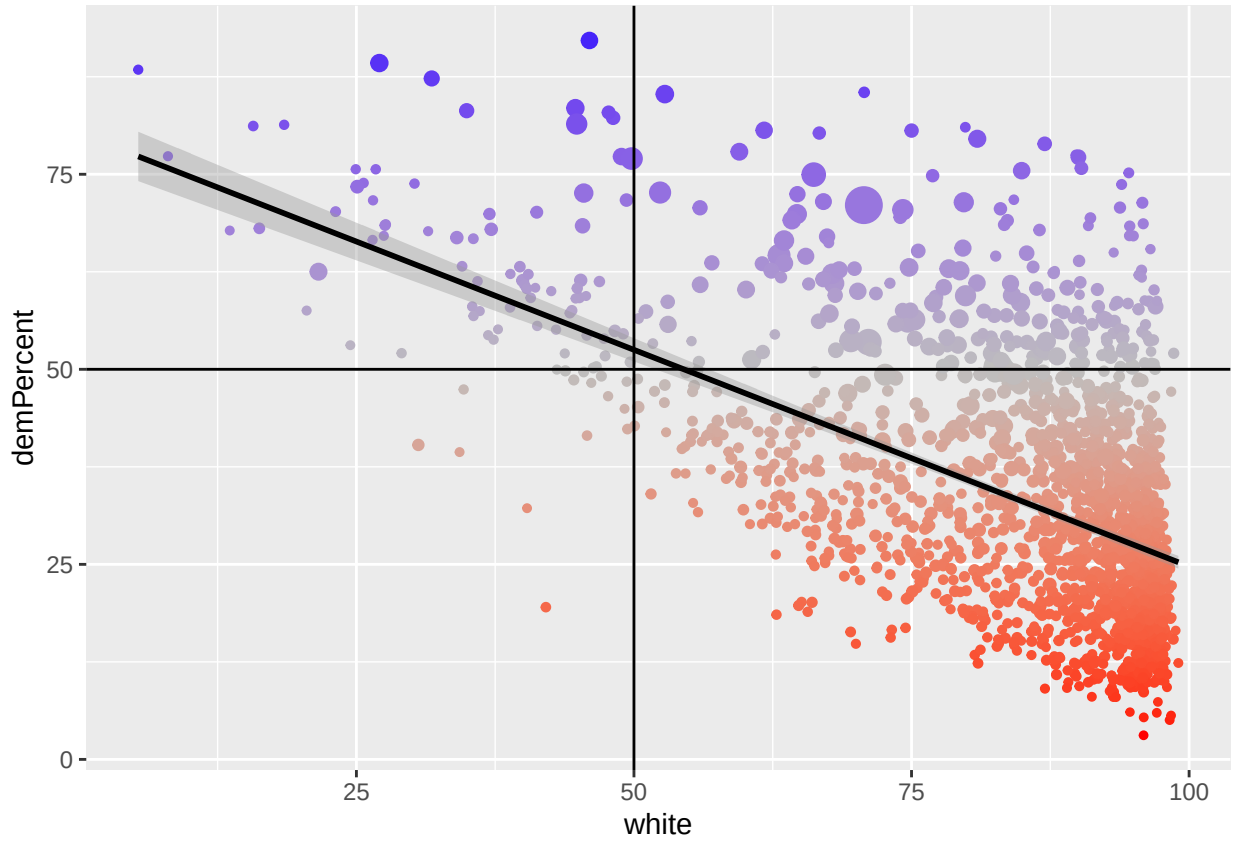
In this portion the second part of the hypothesis, surrounding the ethnic distribution of voters, is tested. Here I have opted to use the variable 'white' as an inverse measure against democratic percentage of votes. This is valid under the assumption that the US populations is at a majority white nation, meaning that remaining percentages represents the ethnic minorities.For the alternative hypothesis to be true there would have to be be a statistically significant negative relationship between white voters and the share of votes for the democratic party.

Below a linear model was constructed to do just that. As predicted the relationship was highly significant where p<2e-16. Interestingly the intercept (p<2e-16) is greater than 1 meaning that an entirely non-white county should have 103.64 % of the votes democrat. Whilst this is of course impossible, it does speak to the practical significance of the result.The coefficient itself suggests that for every 1% increase in white percentage there is a 0.56% decrease in democrat percentage. The standard error of the graph increases substantially for the region left of 50% white, but maintains the upward trajectory of democrat percentage share.This model also explains alot more variance than turnout alone, with an adjusted R-squared value of 0.3237. Meaning that the variable 'white' explains 32.37% of the total variance in democratic vote percentage.

The scatter plot shows this line and colors observation points by democrat percentage. To give greater perspective it also hols the size of points in relation to the total votes cast at the county. This larger helps visually show that democrat majorities occur in counties in larger populations.

```
##
## Call:
## lm(formula = white ~ demPercent, data = VotingTibble)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.436  -5.887   2.663   8.283  31.972
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102.4339     0.6976  146.84   <2e-16 ***
## demPercent   -0.5299     0.0190  -27.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.98 on 1865 degrees of freedom
## Multiple R-squared:  0.2942, Adjusted R-squared:  0.2939
## F-statistic: 777.5 on 1 and 1865 DF,  p-value: < 2.2e-16
```



Predicting county level election outcome by voting turnout and ethnicity

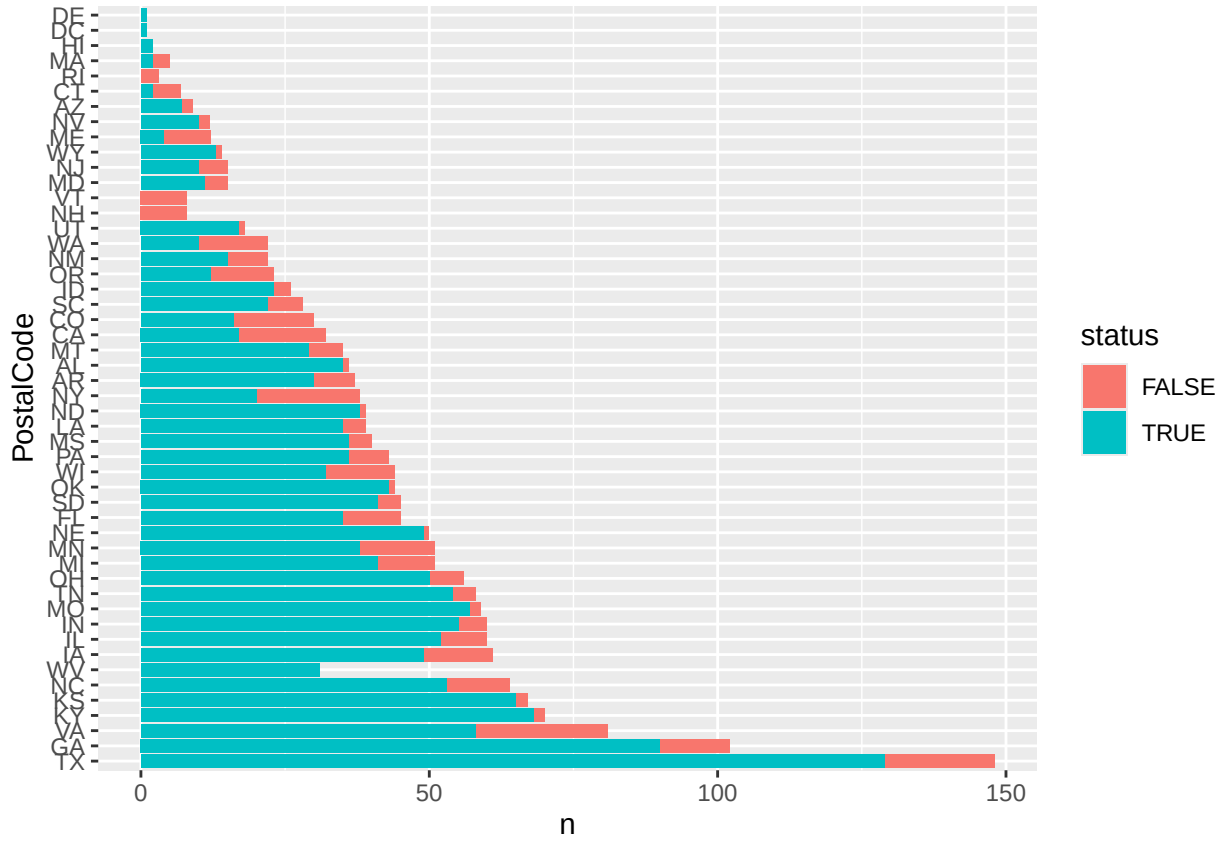
One of the important goals of this project was to use this voting information to predict the type of political battleground of each county and state. By this, I simply mean whether it is a strong win for either party or if the region is a swing county or state. This process starts with defining these categories in the training data set. Here I have considered a margin of win greater than 5% as a strong win for each party, whilst the region in between defines each observation as a swing county. This was implemented using a case_when() function.

Constructing categorical variable

After this stage we move on to utilizing support vector machines to apply different models to help predict the data. Here I have tried to use four types of models: linear, polynomial, radial and sigmoid. These will attempt to predict the win category of each county based on the following variables: white percentage, black percentage, hispanic percentage, total population, voting turnout, and total votes cast. I chose these variables as they are relevant to the theory that the outcome of a county win depends on the ethnic distribution, population and voting turnout.

In these models I used the respective kernels to model the training data and then apply predictions on the selected variables aforementioned. This was then grouped by the state of the counties and used to construct the bar graphs filled in relation to the number of sucessful (TRUE) and unsuccessful (FALSE) predictions. Finally an overall success rate was produced for each model as well as a sample of the top 6 best and worst predicted states.

Linear SVM



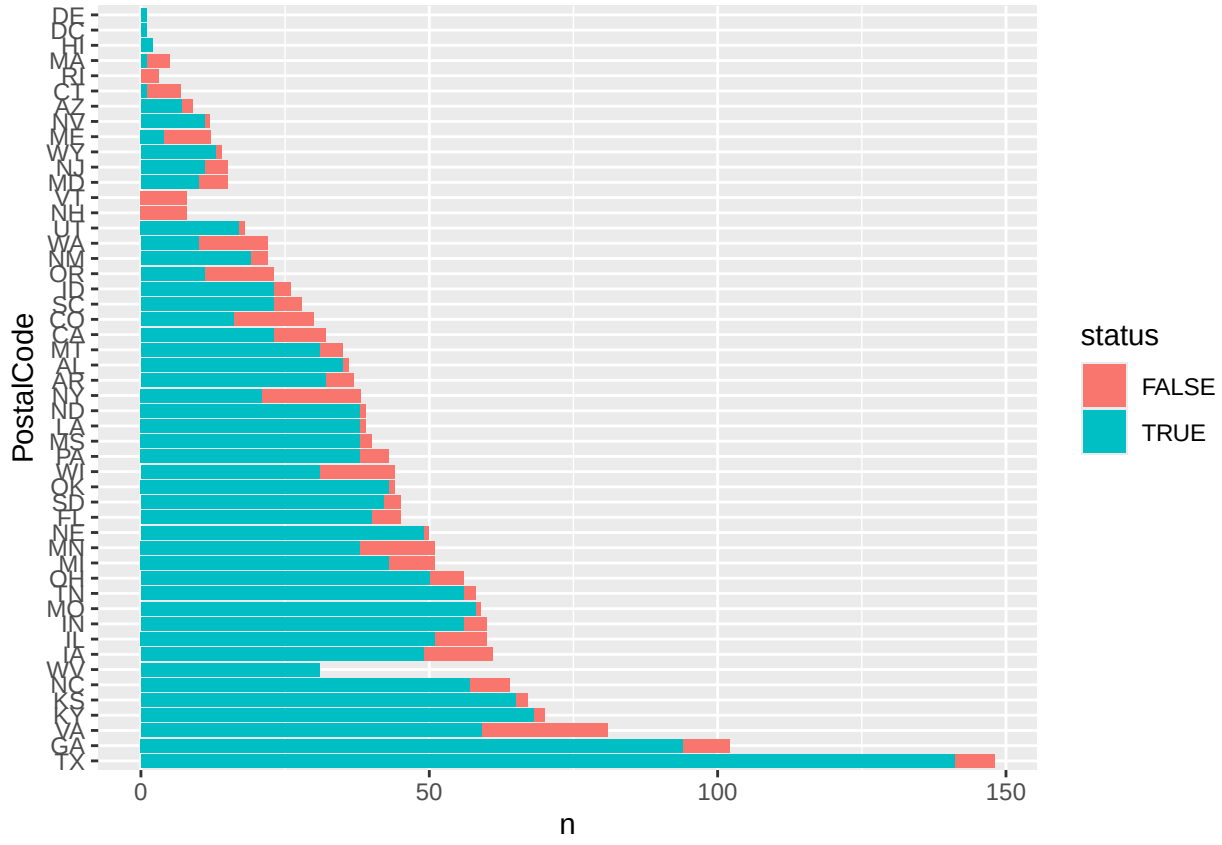

```
## # A tibble: 2 x 3
##   status success_rate    sd
##   <lgl>         <dbl> <dbl>
## 1 FALSE         0.176 0.265
## 2 TRUE          0.835 0.187
```

```
## # A tibble: 6 x 2
##   PostalCode linear
```

```
##   <fct>       <dbl>
## 1 DC          1
## 2 DE          1
## 3 HI          1
## 4 WV          1
## 5 NE          0.98
## 6 OK          0.977
```

```
## # A tibble: 6 x 2
##   PostalCode linear
##   <fct>       <dbl>
## 1 NY          0.526
## 2 OR          0.522
## 3 WA          0.455
## 4 MA          0.4
## 5 ME          0.333
## 6 CT          0.286
```

Polynomial SVM

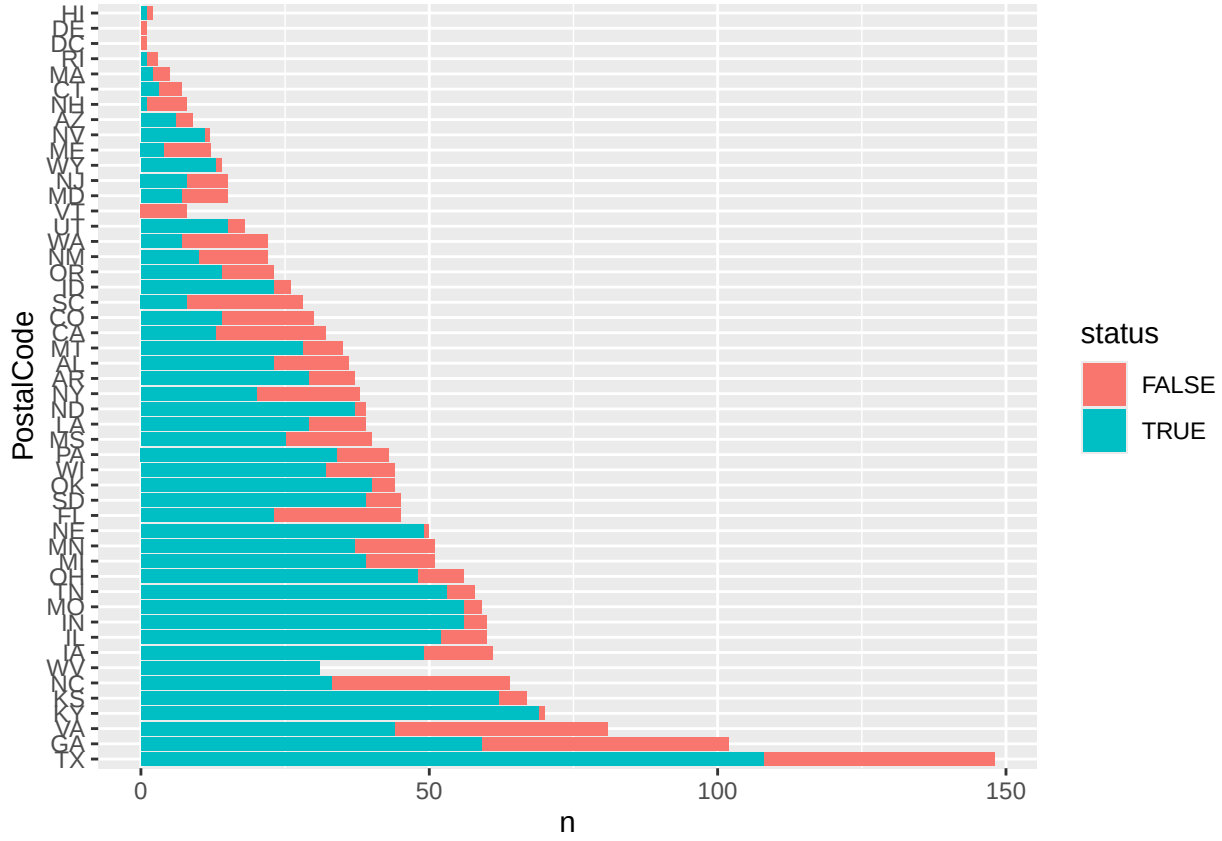



```
## # A tibble: 2 x 3
##   status success_rate    sd
##   <lgl>         <dbl> <dbl>
## 1 FALSE         0.148 0.286
## 2 TRUE          0.864 0.210
```

```
## # A tibble: 6 x 2
##   PostalCode  poly
##   <fct>      <dbl>
## 1 DC         1
## 2 DE         1
## 3 HI         1
## 4 WV         1
## 5 MO         0.983
## 6 NE         0.98
```

```
## # A tibble: 6 x 2
##   PostalCode  poly
##   <fct>      <dbl>
## 1 CO         0.533
## 2 OR         0.478
## 3 WA         0.455
## 4 ME         0.333
## 5 MA         0.2
## 6 CT         0.143
```

Sigmoid SVM



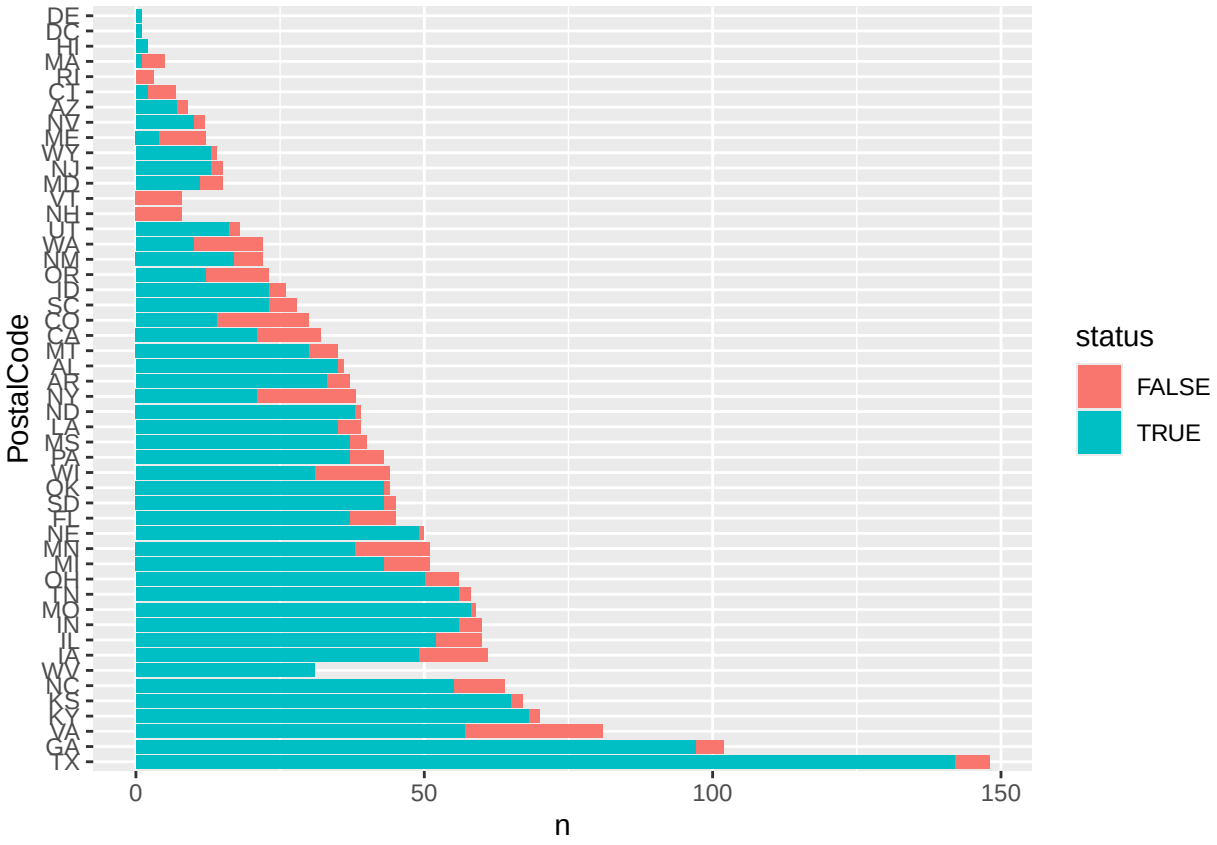

```
## # A tibble: 2 x 3
##   status success_rate    sd
```

```
##   <lgl>          <dbl> <dbl>
## 1 FALSE          0.273 0.272
## 2 TRUE           0.735 0.228
```

```
## # A tibble: 6 x 2
##   PostalCode sigmoid
##   <fct>        <dbl>
## 1 WV           1
## 2 KY           0.986
## 3 NE           0.98
## 4 MO           0.949
## 5 ND           0.949
## 6 IN           0.933
```

```
## # A tibble: 6 x 2
##   PostalCode sigmoid
##   <fct>        <dbl>
## 1 MA           0.4
## 2 ME           0.333
## 3 RI           0.333
## 4 WA           0.318
## 5 SC           0.286
## 6 NH           0.125
```

Radial SVM



```
## # A tibble: 2 x 3
##   status success_rate    sd
##   <lgl>         <dbl> <dbl>
## 1 FALSE         0.153 0.278
## 2 TRUE          0.859 0.200
```

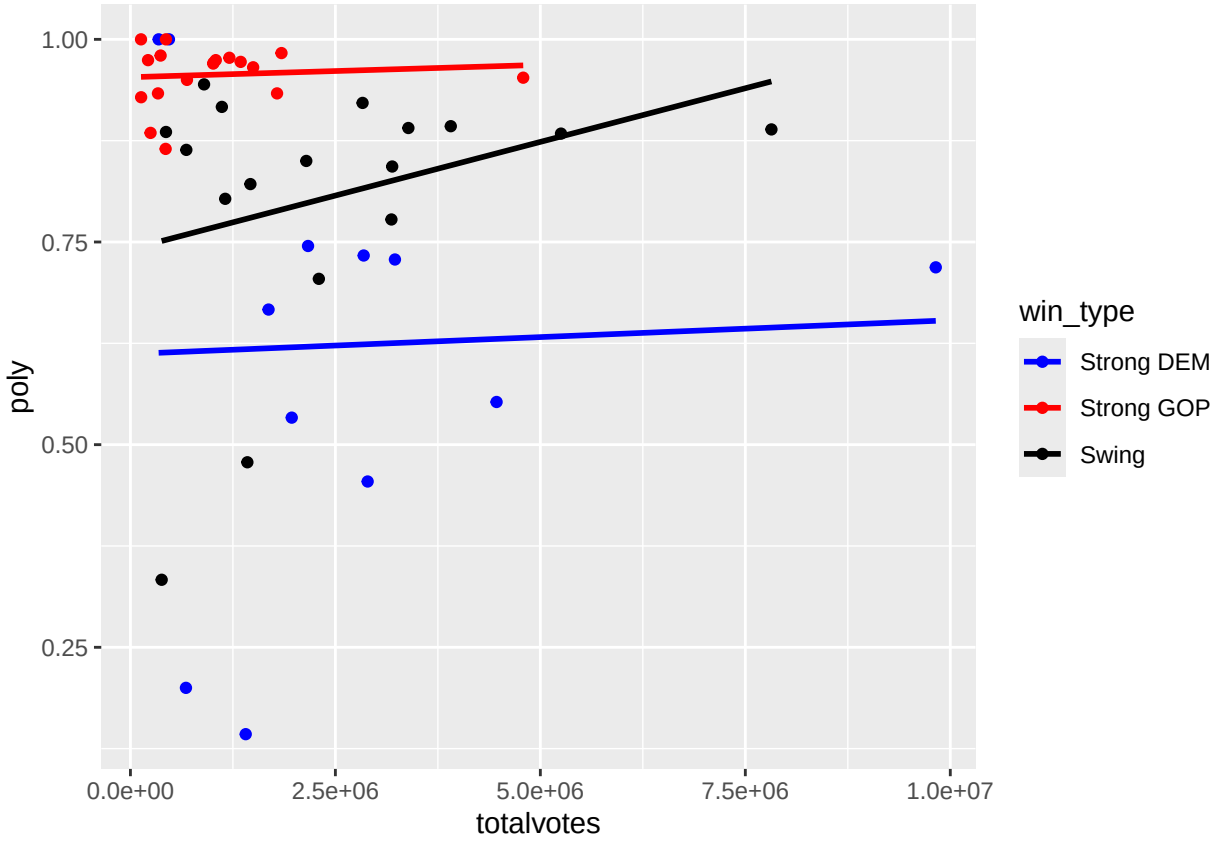
```
## # A tibble: 6 x 2
##   PostalCode radial
##   <fct>       <dbl>
## 1 DC          1
## 2 DE          1
## 3 HI          1
## 4 WV          1
## 5 MO          0.983
## 6 NE          0.98
```

```
## # A tibble: 6 x 2
##   PostalCode radial
##   <fct>       <dbl>
## 1 OR          0.522
## 2 CO          0.467
## 3 WA          0.455
## 4 ME          0.333
## 5 CT          0.286
## 6 MA          0.2
```

Of all the models the best performing kernel was in fact the polynomial model with an 84.8% accuracy. This was closely followed by the radial model at 84.6%. The standard deviation for the successes in polynomial model was also smaller indicating a lower amount of variance in predicting outcomes. The sigmoid kernel performed the worst by a large margin in making predictions at 71.3% accuracy. Unsurprising the better predicted states by proportion tended to be the smaller ones with fewer counties. Montana, DC and Delaware were all frequently on these top lists. Of states with larger numbers of counties, Tennessee and Oklahoma were predicted rather accurately. The worst predicted states were interestingly mainly in New England and north east of the US. Maine, Massachusetts and Connecticut were frequently seen on these lists.

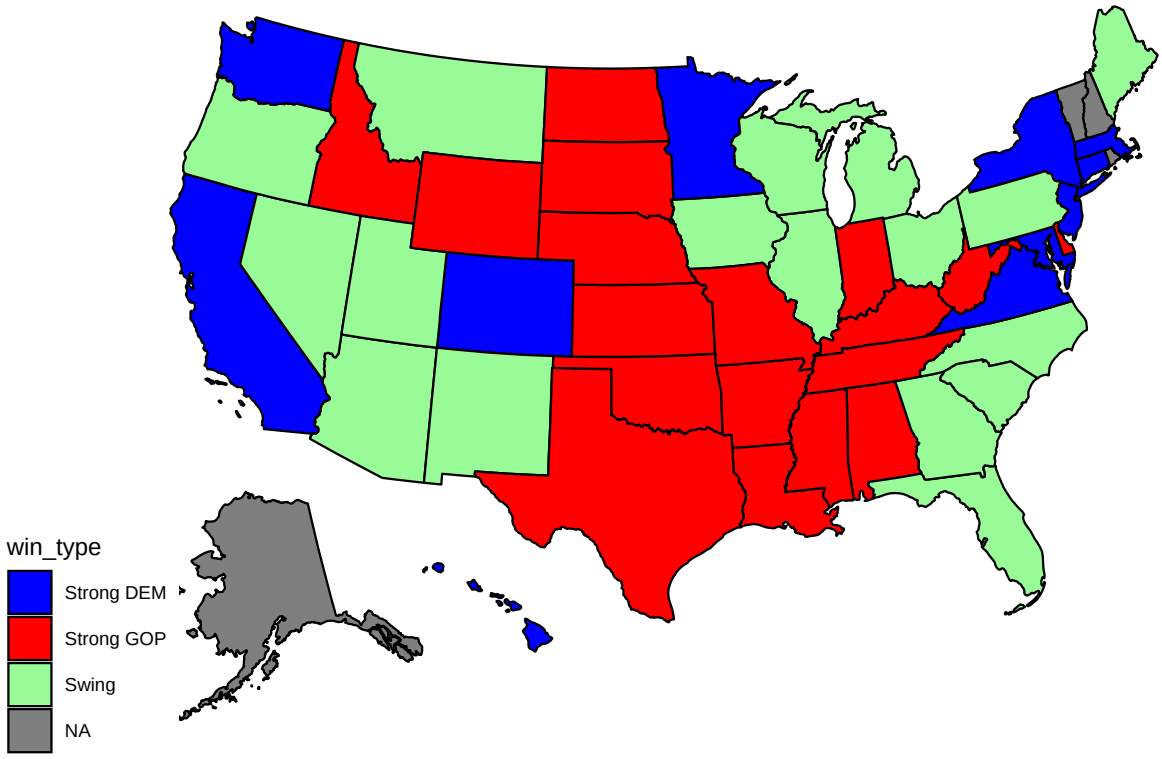
Success rate by party and total votes

Here I wanted to get a visual sense of the success of predictions when grouping by state as well as electoral votes. In the chunk below I constructed a data frame that contained the most successful model 'poly'. From there I generated a scatter plot with linear models for each win type. The data points are also colored by the win type predicted. I have also constructed a linear model to assess the total relationship between the total votes of each state and the accuracy of the polynomial SVM model.



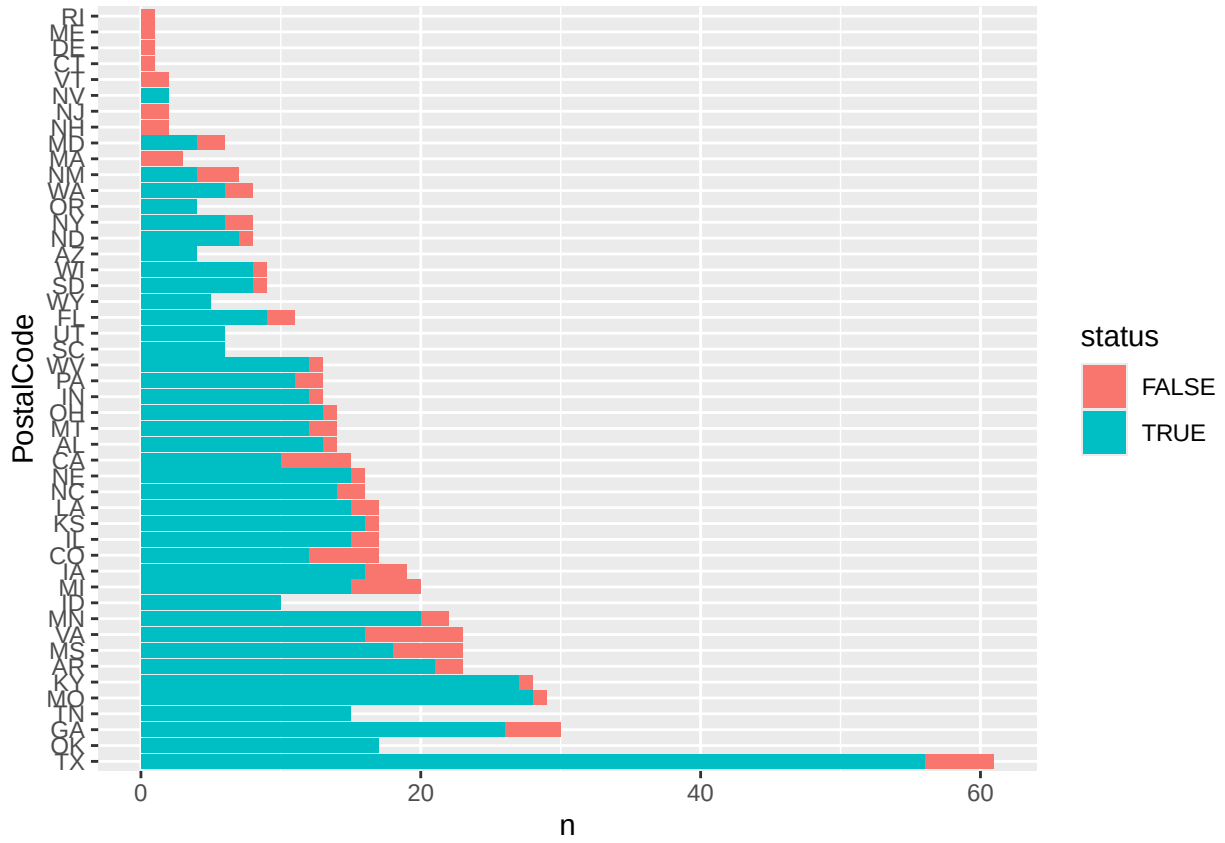

```
##
## Call:
## lm(formula = poly_statevotes$poly ~ poly_statevotes$totalvotes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68050 -0.04695  0.09016  0.14267  0.16929
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  8.399e-01  4.373e-02  19.206   <2e-16 ***
## poly_statevotes$totalvotes -1.180e-08  1.582e-08  -0.746     0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2115 on 45 degrees of freedom
## Multiple R-squared:  0.01221,    Adjusted R-squared:  -0.009745
## F-statistic: 0.5561 on 1 and 45 DF,  p-value: 0.4597
```

Quantitatively there is no overall relationship between the total votes in each state and the predictive accuracy of the polynomial model. When looking at the scatterplot it can be observed that the win types cluster together at certain predictive levels. Predicting republican wins seems to be highly accurate, whereas democrat wins are much less so. Interestingly, swing states are in between the two levels.



Applying polynomial model to query set

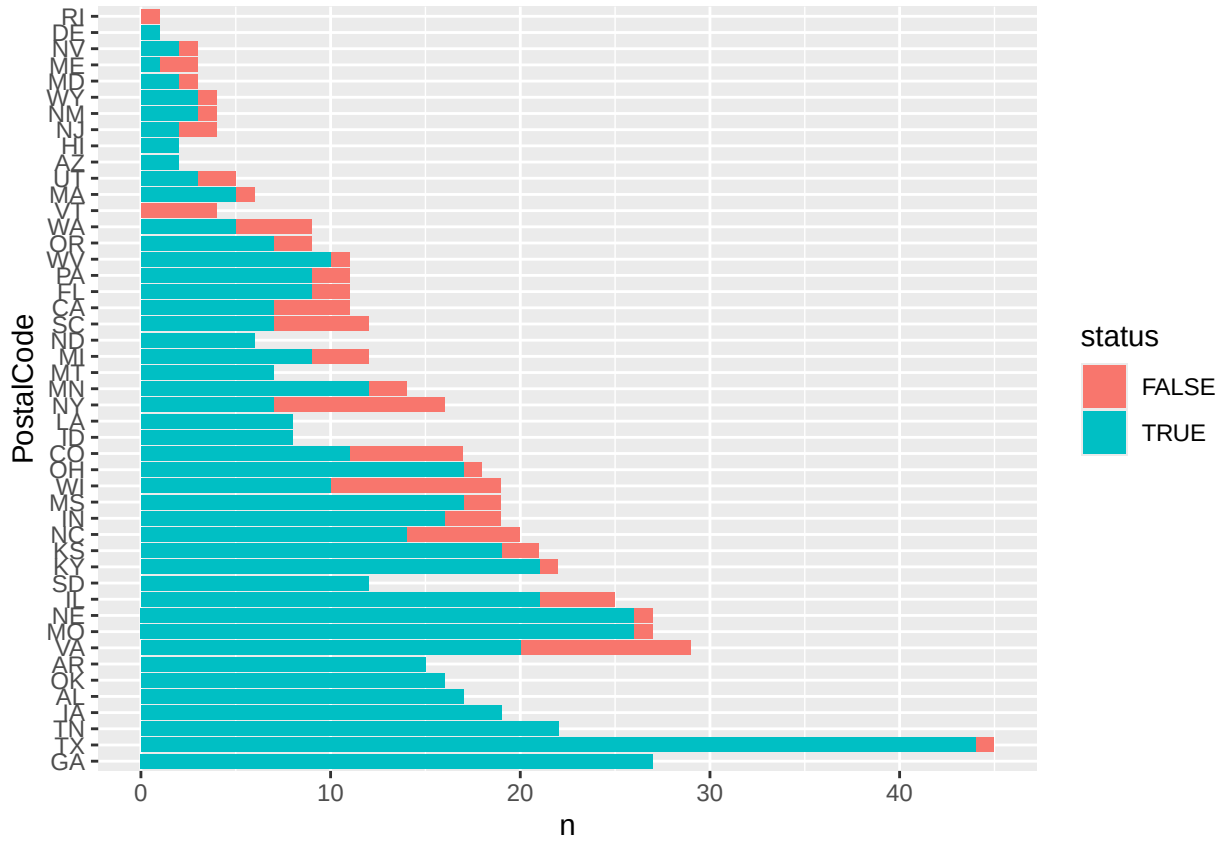
Before applying the polynomial model to the test set, I ran it with the query data set to verify it's success rate. I found that it was identical in its success rate. I constructed a US Map to display the distribution of the voting outcome categories.




```
## # A tibble: 2 x 3
##   status success_rate    sd
##   <lgl>          <dbl> <dbl>
## 1 FALSE          0.159 0.356
## 2 TRUE           0.877 0.111
```

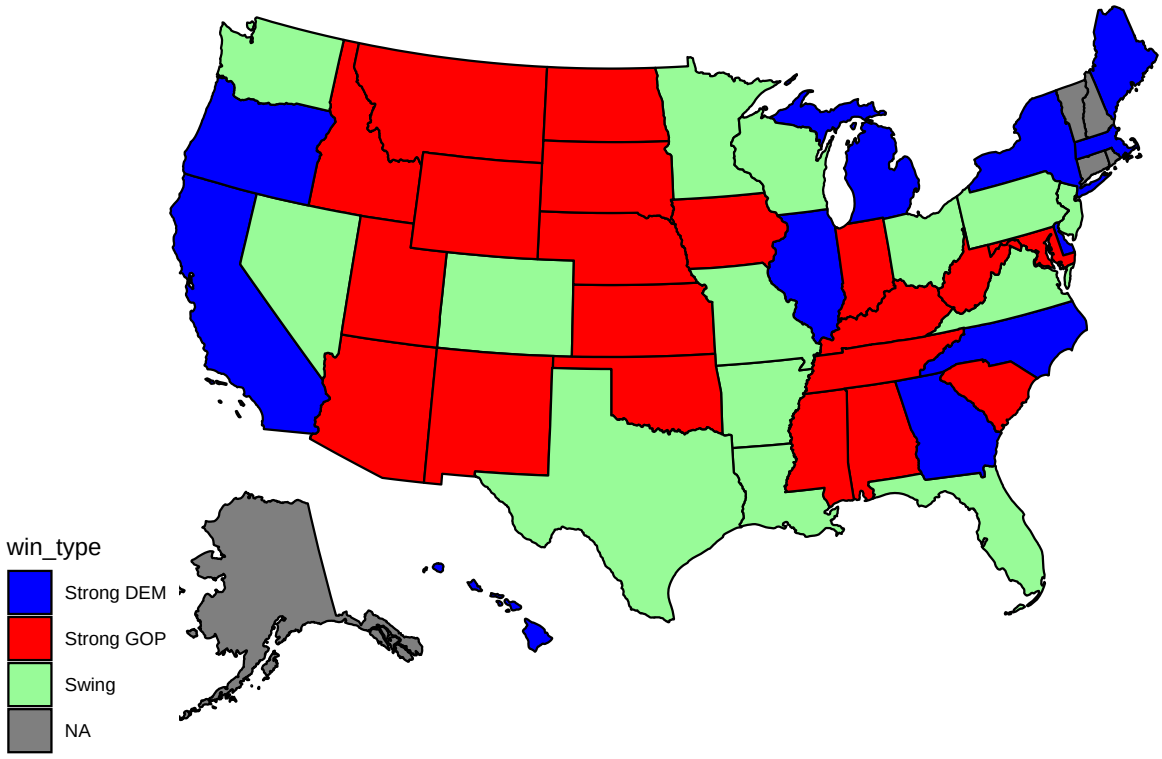
Testing polynomial model

In this portion I tested the model using the data set left aside called test. Interestingly the model here outperformed both the query set and training set. Here it had an average accuracy of 87.7%.




```
## # A tibble: 2 x 3
##   status success_rate    sd
##   <lgl>         <dbl> <dbl>
## 1 FALSE         0.208 0.243
## 2 TRUE          0.853 0.183
```

However, from the graph produced below you can see that there are considerable differences between the US Map generated in the training data versus the map generated in the test. This may be a sampling issue which arises from the fact that more counties across the US are republican than democrat. Therefore, there is greater chance of predictive error statewide and countrywide as the sampling size decreases. This being said, there was success in the consistency of certain swing states such as Texas, North Carolina and Iowa.



Conclusions

In this project there have been a number of successes. Firstly, based on the linear regressions run there is reason to believe that there is a significant relationship between voting turnout and the percentage share of democrat votes. This is a positive relationship but is ultimately of little practical significance. From the coefficient, the increase is marginal and explains a minute level of variance. Nonetheless, I can reject the null hypothesis with extremely strong statistical significance.

In contrast, the second theory surrounding the ethnic distribution in counties proved a much better predictor of the outcome of democrat votes. Here there was a large negative relationship between the proportion of white people and democrat percent vote. By deductive reasoning this would mean that the percentage of minorities would increase with democrat votes. Therefore, I can reject my second null hypothesis with incredibly strong statistical evidence.

Finally, there was some success in predicting the win types for states and counties across the US. Using support vector machine modeling I found a polynomial model to have some moderate success in predicting the outcome. In the training phase, it scored 84.8% accuracy in predicting whether a county was a strong republican, strong democrat or a swing. In the testing phase this actually increased to 87.4% which would further verify the true predictive accuracy somewhere in the mid-80% range. As a follow on, I was able to identify that the polynomial predictive model was on average more accurate for republican observations than swing states and democrat states. This model utilized statistics on population, voting turnout and ethnic proportions solely to predict these voting outcomes.In the end this puts to practice the theory introduced earlier. From this it can now be said that the proportions of minorities, voting turnouts and populations of counties are important in predicting the success of either party in elections.

Sources

https://www.pewresearch.org/fact-tank/2021/01/28/turnout-soared-in-2020-as-nearly-two-thirds-of-eligible-u-s-voters-cast-ballots-for-president/

https://www.census.gov/library/stories/2021/04/record-high-turnout-in-2020-general-election.html

https://github.com/tonmcg/US_County_Level_Election_Results_08-20 (Voting)

https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/ (Covid)

https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/ (Education, etc)

https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html (Racial/Pop.)