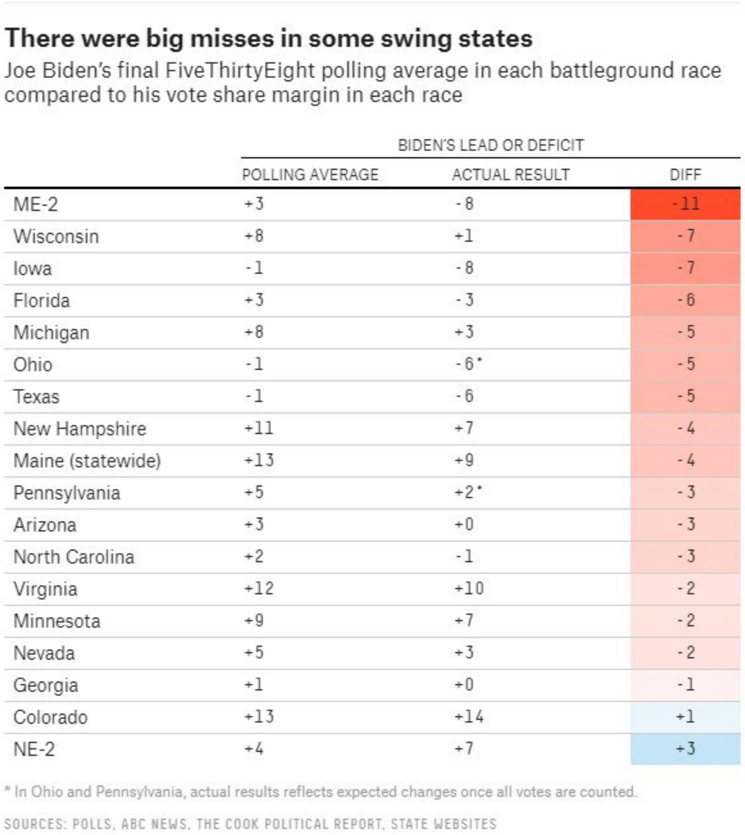
DATA621 Final Project

Mikhail Kollontai

Problem Statement

- Predictions based on Polling have an inherent margin of error
- The margins between pollster prediction and results in the 2020 (and 2016) Presidential elections were significant
- Can we find a way to use objective data to predict elections?

2020 Margins



There were big misses in some swing states

Joe Biden's final FiveThirtyEight polling average in each battleground race compared to his vote share margin in each race

	BIDEN'S LEAD OR DEFICIT		
	POLLING AVERAGE	ACTUAL RESULT	DIFF
ME-2	+3	-8	-11
Wisconsin	+8	+1	-7
Iowa	-1	-8	-7
Florida	+3	-3	-6
Michigan	+8	+3	-5
Ohio	-1	-6*	-5
Texas	-1	-6	-5
New Hampshire	+11	+7	-4
Maine (statewide)	+13	+9	-4
Pennsylvania	+5	+2*	-3
Arizona	+3	+0	-3
North Carolina	+2	-1	-3
Virginia	+12	+10	-2
Minnesota	+9	+7	-2
Nevada	+5	+3	-2
Georgia	+1	+0	-1
Colorado	+13	+14	+1
NE-2	+4	+7	+3

* In Ohio and Pennsylvania, actual results reflects expected changes once all votes are counted.

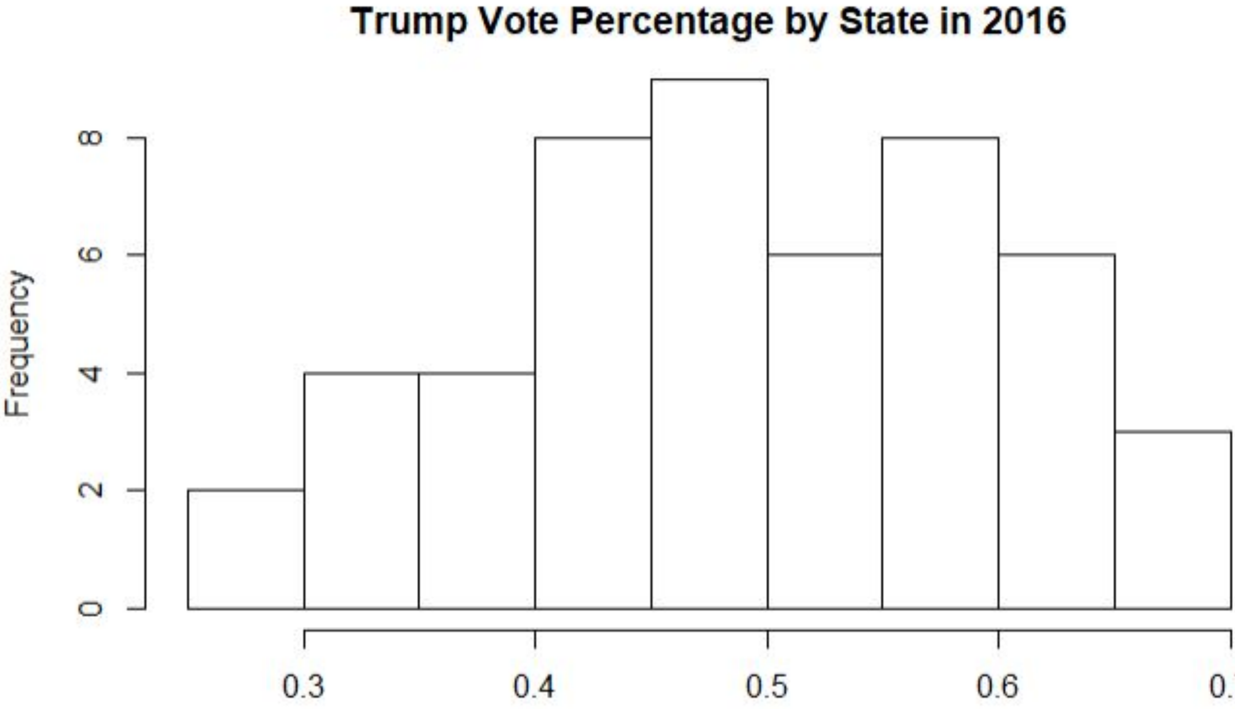
SOURCES: POLLS, ABC NEWS, THE COOK POLITICAL REPORT, STATE WEBSITES

Source: (https://fivethirtyeight.com/features/the-polls-werent-great-but-thats-pretty-normal/)

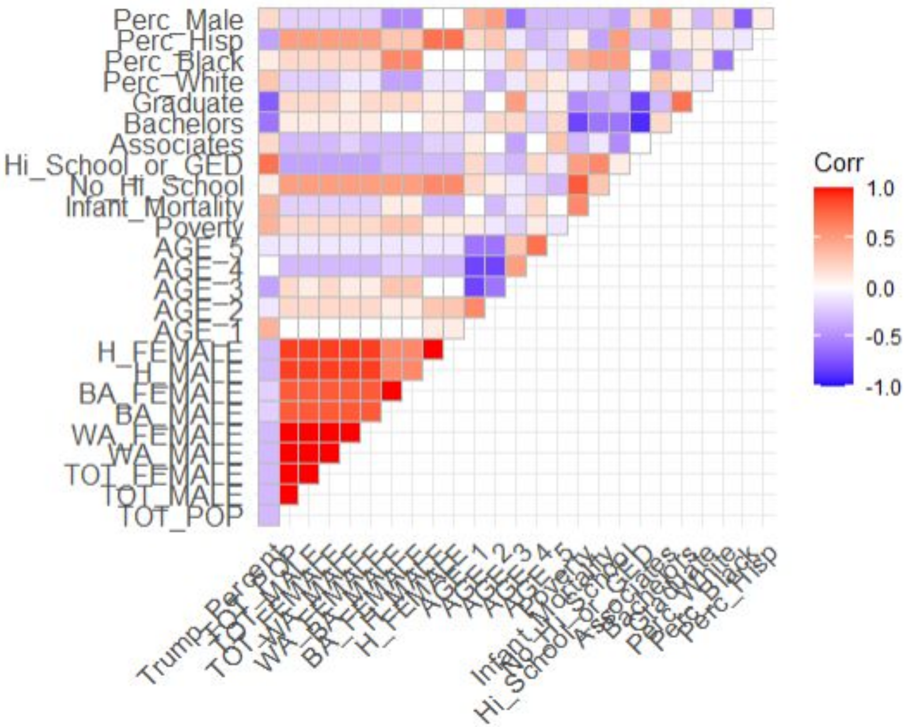
Predictors - State-Level Demographics

- Total Population
- Male/Female Breakdown
- Total Black Population
- Total Hispanic Population
- Age Breakdown
- Poverty Levels
- Education Levels
- Child Mortality Rates

Training Target Variable



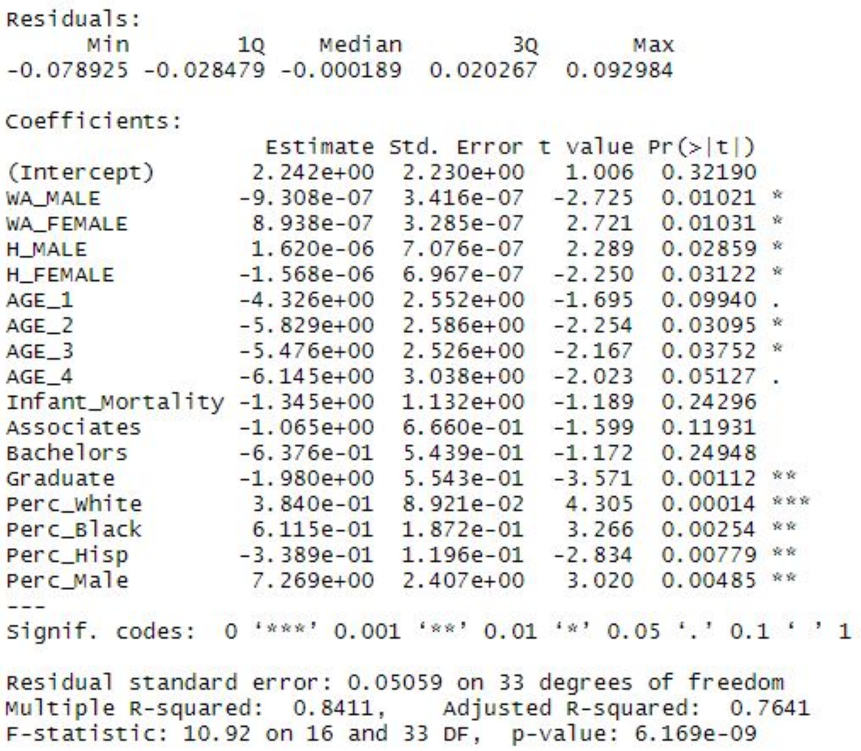
Variable Correlation Plot



- Population variables show strong interdependence - relative values are tied

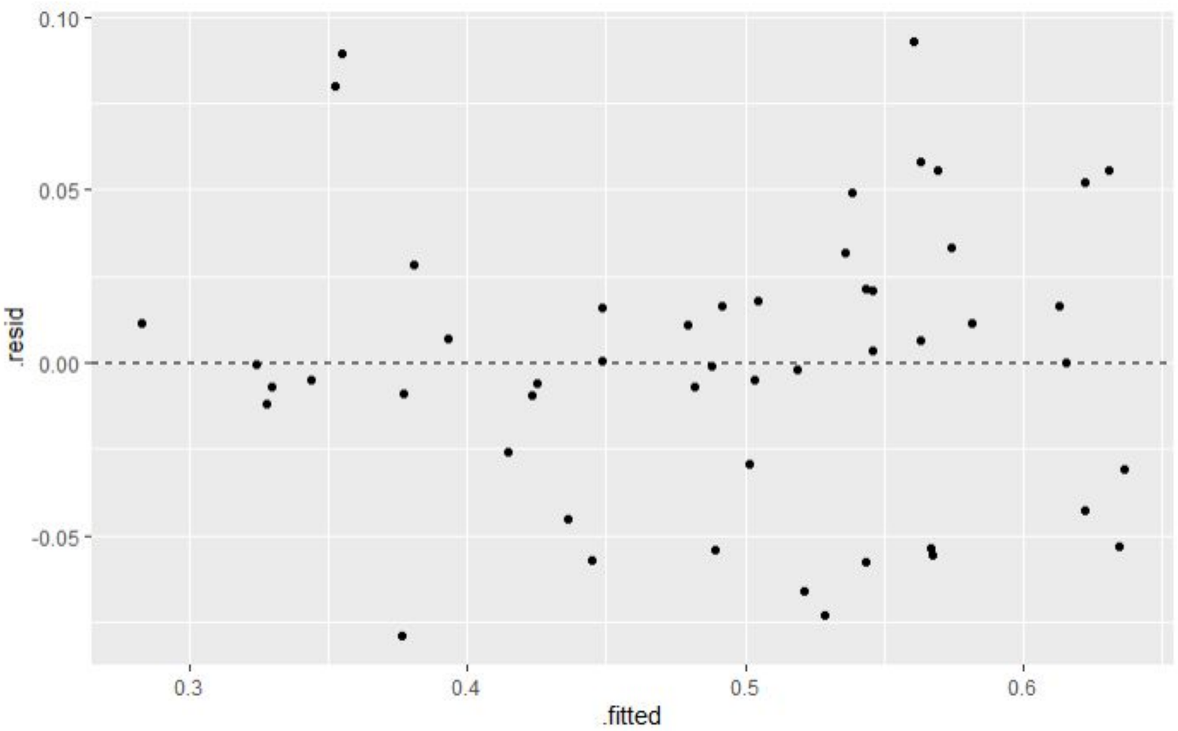
- Trump Percentage Variable
  - Strong positive relationship with Only High School Degree/GED population
  - Strongest negative relationship with Graduate degree and Bachelor’s Degree
  - Negative relationship with Hispanic Population

Final Model Coefficients

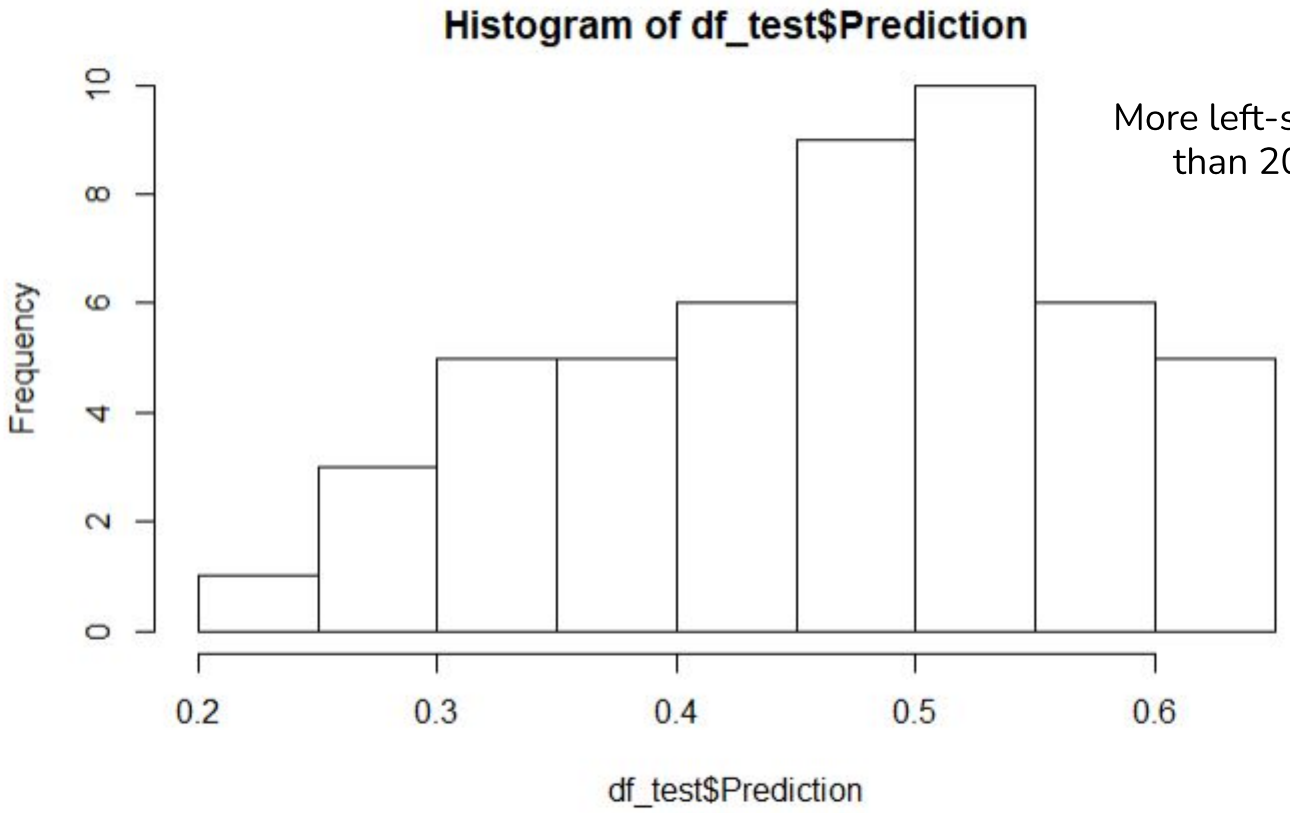


```
Residuals:
      Min        1Q    Median        3Q       Max 
-0.078925 -0.028479 -0.000189  0.020267  0.092984 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.242e+00  2.230e+00   1.006  0.32190    
WA_MALE          -9.308e-07  3.416e-07  -2.725  0.01021 *  
WA_FEMALE         8.938e-07  3.285e-07   2.721  0.01031 *  
H_MALE            1.620e-06  7.076e-07   2.289  0.02859 *  
H_FEMALE         -1.568e-06  6.967e-07  -2.250  0.03122 *  
AGE_1            -4.326e+00  2.552e+00  -1.695  0.09940 .  
AGE_2            -5.829e+00  2.586e+00  -2.254  0.03095 *  
AGE_3            -5.476e+00  2.526e+00  -2.167  0.03752 *  
AGE_4            -6.145e+00  3.038e+00  -2.023  0.05127 .  
Infant_Mortality -1.345e+00  1.132e+00  -1.189  0.24296    
Associates       -1.065e+00  6.660e-01  -1.599  0.11931    
Bachelors        -6.376e-01  5.439e-01  -1.172  0.24948    
Graduate         -1.980e+00  5.543e-01  -3.571  0.00112 ** 
Perc_White        3.840e-01  8.921e-02   4.305  0.00014 ***
Perc_Black        6.115e-01  1.872e-01   3.266  0.00254 ** 
Perc_Hisp        -3.389e-01  1.196e-01  -2.834  0.00779 ** 
Perc_Male         7.269e+00  2.407e+00   3.020  0.00485 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05059 on 33 degrees of freedom
Multiple R-squared:  0.8411,    Adjusted R-squared:  0.7641 
F-statistic: 10.92 on 16 and 33 DF,  p-value: 6.169e-09
```

Final Model Residuals

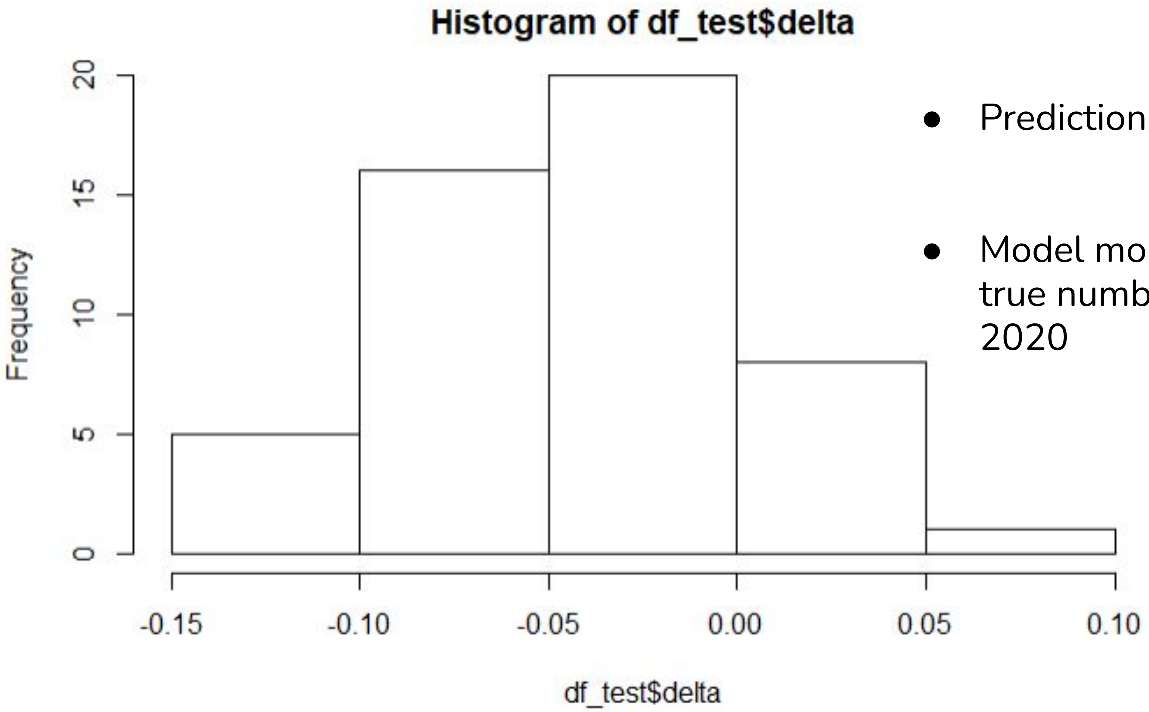


Prediction from Model



More left-skewed than 2016

Difference b/w Prediction and 2020



- Prediction margins as high as 15%
- Model more often undershot the true numbers Trump earned in 2020

Conclusions

- Predictions with our model could not improve on polling predictions

- 28 States predicted to within a 5% margin
- 17 States predicted within 5-10% margin
- 5 States predicted with margin above 10%

- This model approach does not seem sufficient for a viable prediction
- Ideally county-level data would be used to improve granularity