Beta 538

•••

Bisher, Caio, Max, Surabhi, Vivek, and Will

Overview

Motivation and Questions

Data Exploration

Feature Selection

Models and Specifications

Conclusion

Motivation and Questions

Elections are multi-faceted

- Where to allocate funding
- How to strategize as parties
- Multiple levels of government

Question 1: What level of turnout for an election?

Question 2: Which counties and seats will change party hands?

Question 3: Can we predict the winner of a specific election?

Data Exploration

Data

-From Kaggle

-Had extensive amount of features, all very descriptive

-Good size of data

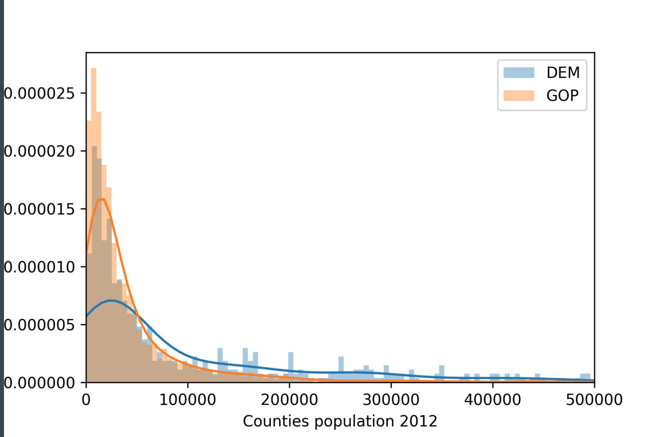
-Allowed us to work on different questions/parts of the data science life cycle

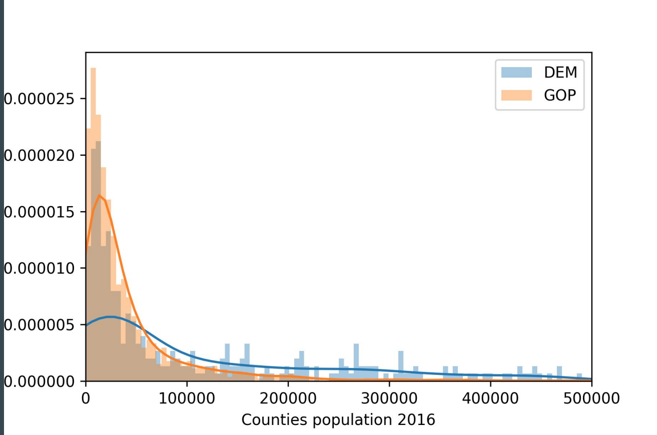
Data Exploration

We got multiple datasets which were pretty similar to each other. Each was a list of election results based on counties paired with in total over 80 features.

Features included demographic information, wealth and poverty reports and even economic performance indicators.

Dataset correlations

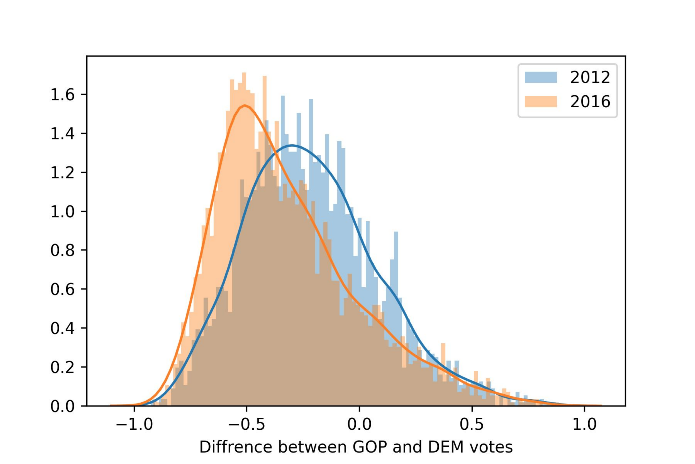




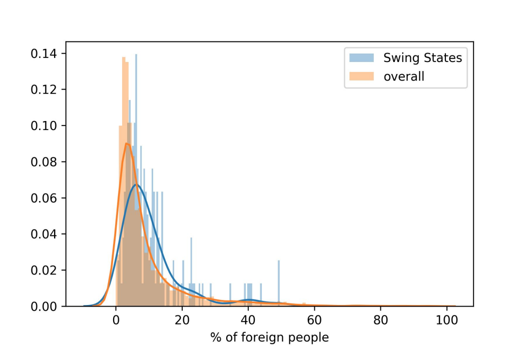
First we see that county wins are not uniformly distributed. Republicans tend to win counties with less population more often.

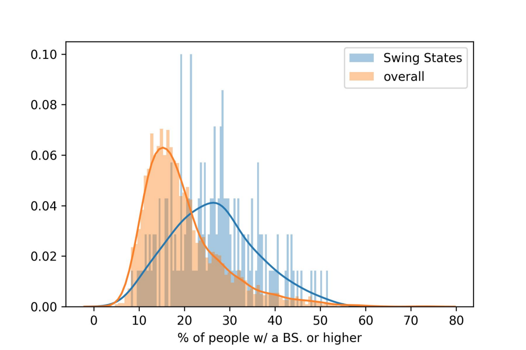
Dataset correlations

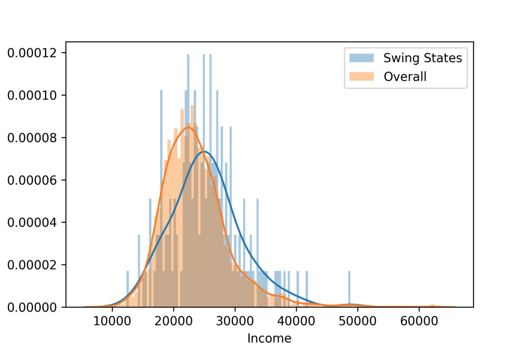
But not only do they win those counties more often, their winning margin is also significantly greater. In 2016 there was a strong shift emphasizing this behaviour.

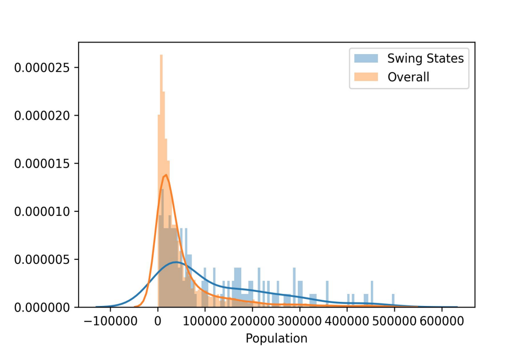


Swing states









Feature Selection

Feature Selection

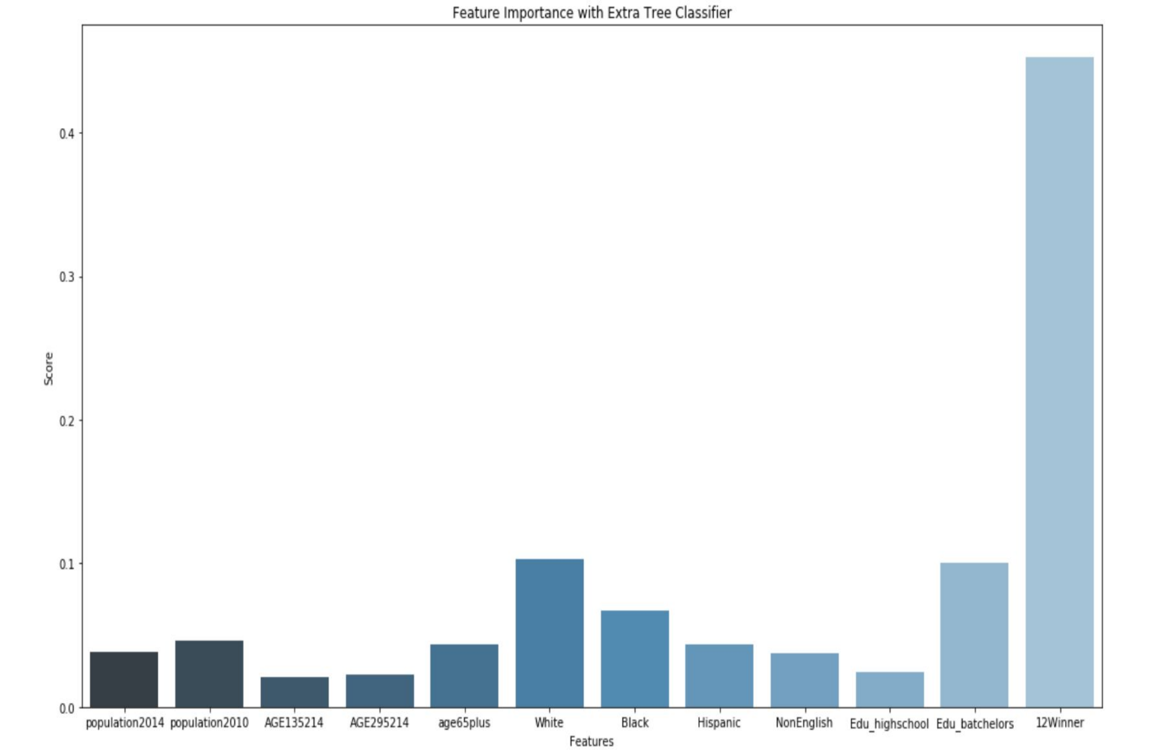
What?

- Automatically select features most useful to problem at hand

Why?

- “If you put garbage in, you will only get garbage to come out”
- Enables machine learning algorithm to train faster
- Improves accuracy of a model
- Reduces the complexity of a model
- Reduces overfitting

Feature Selection: Extra Tree Classifier



Feature Importance

1: 2012 Election (.461)

2: Bachelors Education (.119)

3: % Population White (.072)

4: % Population Black (.059)

5: Veterans (.035)

Feature Selection: Recursive Feature Elimination

1) 2012 Election

2) % Population Hispanic

3) % Population White

4) % Population Black

5) Bachelors Education

6) % People Under 5

7) % People Under 18

8) Non-English

9) Age 65 +

10) Poverty

11) % Population Female

12) Highschool Education

13) Population 2010

14) Population 2014

15) Income

16) Veterans

Exploratory Data Analysis and Feature Engineering

•••

Question 1

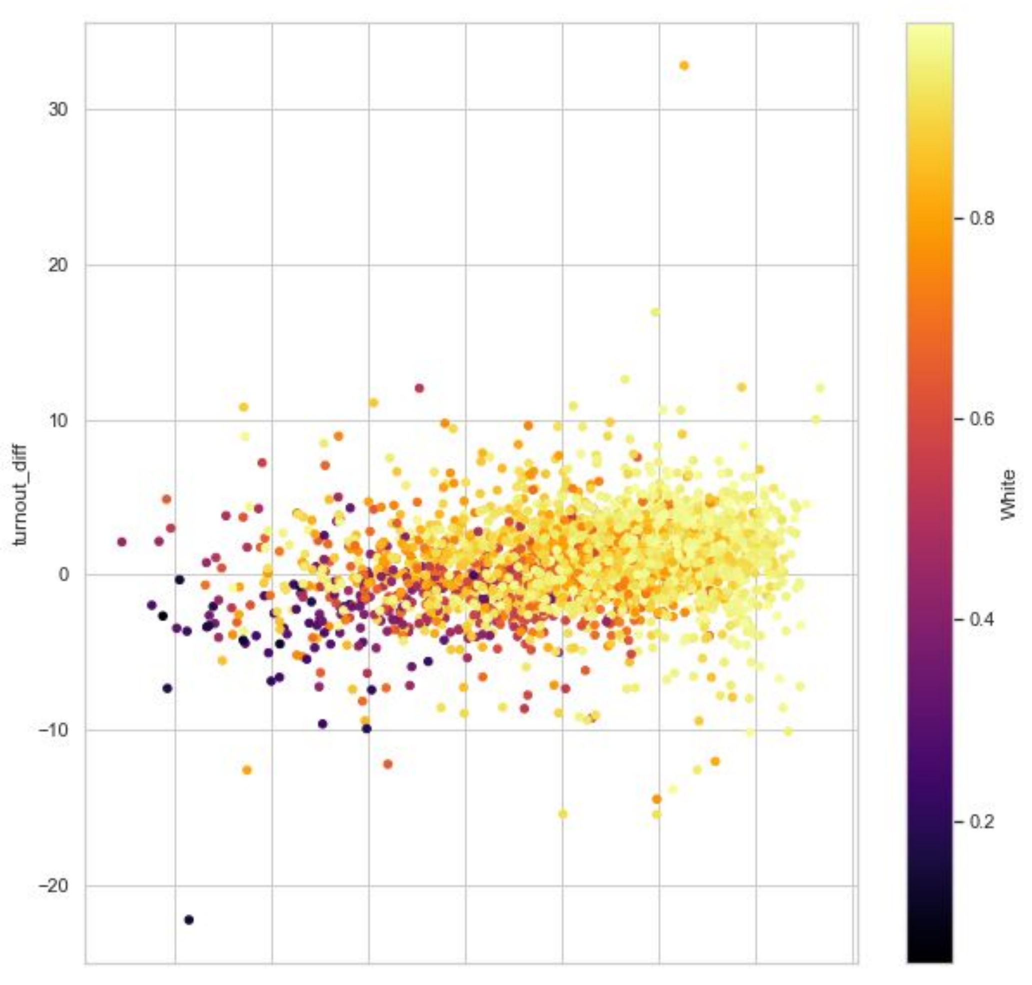
Election Turnout

Feature engineering:

1) 2016 Turnout: Using county population and total number of votes for 2016
2) 2012 Turnout: Using county population and total number of votes for 2012
3) Turnout difference: 2016 Turnout - 2012 Turnout
4) Turnout difference discretized (for classifier algorithms)
5) Candidate popularity (percentage difference): difference between the percentage of the population that voted for Trump and Hillary

Election Turnout, Plot 1

- Let’s try to visualize the data
- Scatter plot where x = percentage difference in votes (+ for Trump, - for Hillary), y = turnout difference between 2016 and 2012, and the color gradient is percentage of people who are white



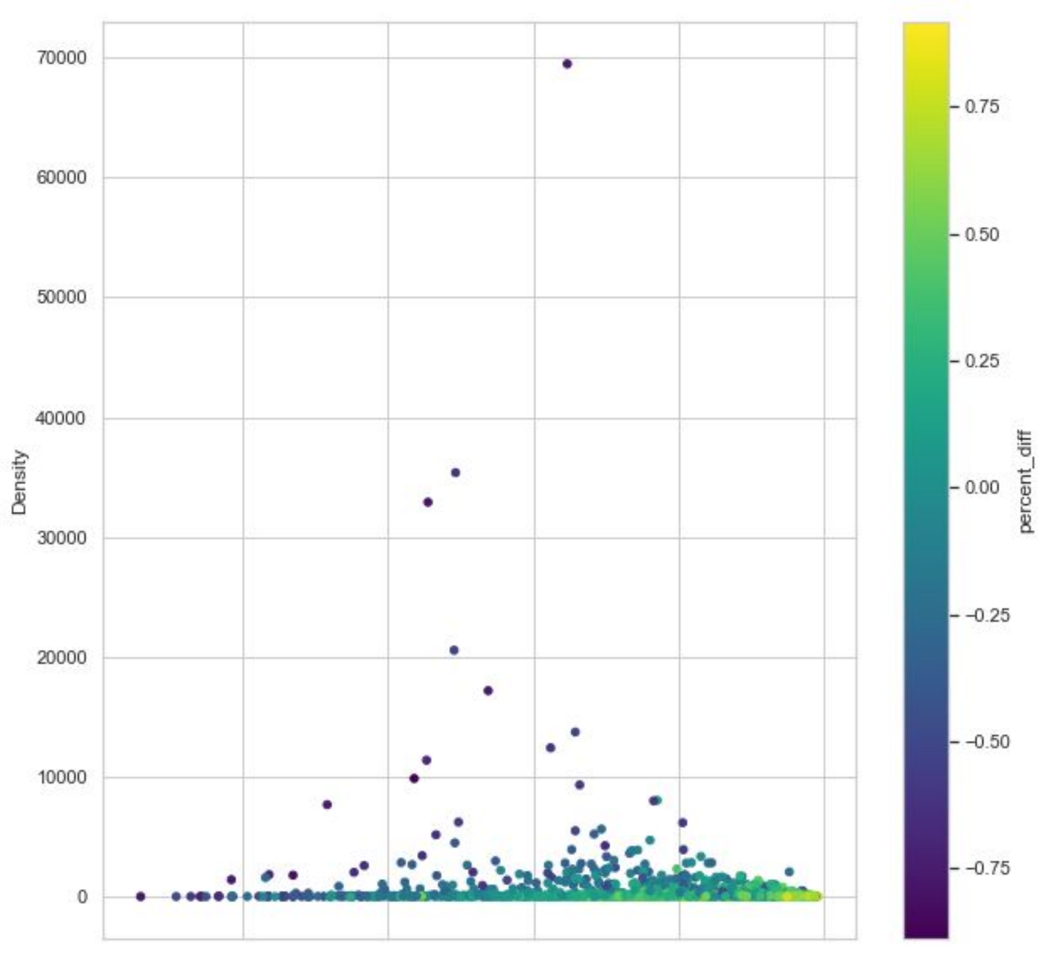
Percentage difference in votes

Election Turnout, Plot 1

- Trends:
  - Slight upward trend with percentage difference between the candidates; i.e. counties that went to Trump had the tendency to turn out at a higher rate
  - Larger White population in a county meant more votes for Trump and slightly higher turnout

Election Turnout, Plot 2

- Scatter plot where x = percentage of white population, y = population density (people per square mile), and the color gradient is percentage turnout difference



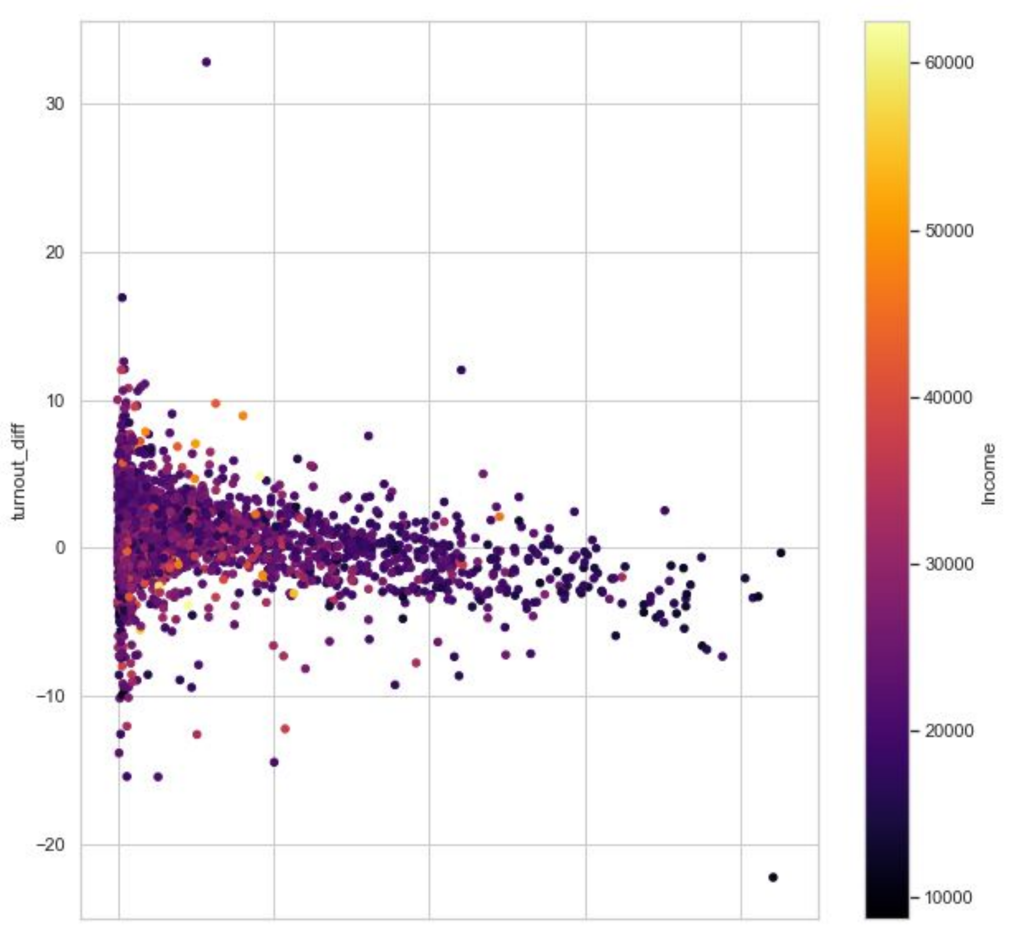
Percentage of white population

Election Turnout, Plot 2

- Trends:
  - Low density, high white population counties had high turnout:
    - Most likely rural areas
  - Very high density counties had negative turnout difference
    - Trend concurs with recent election
    - High density counties usually also have high minority/immigrant populations

Election Turnout, Plot 3

- Scatter plot where x = percentage of black population, y = turnout difference, and the color gradient is income



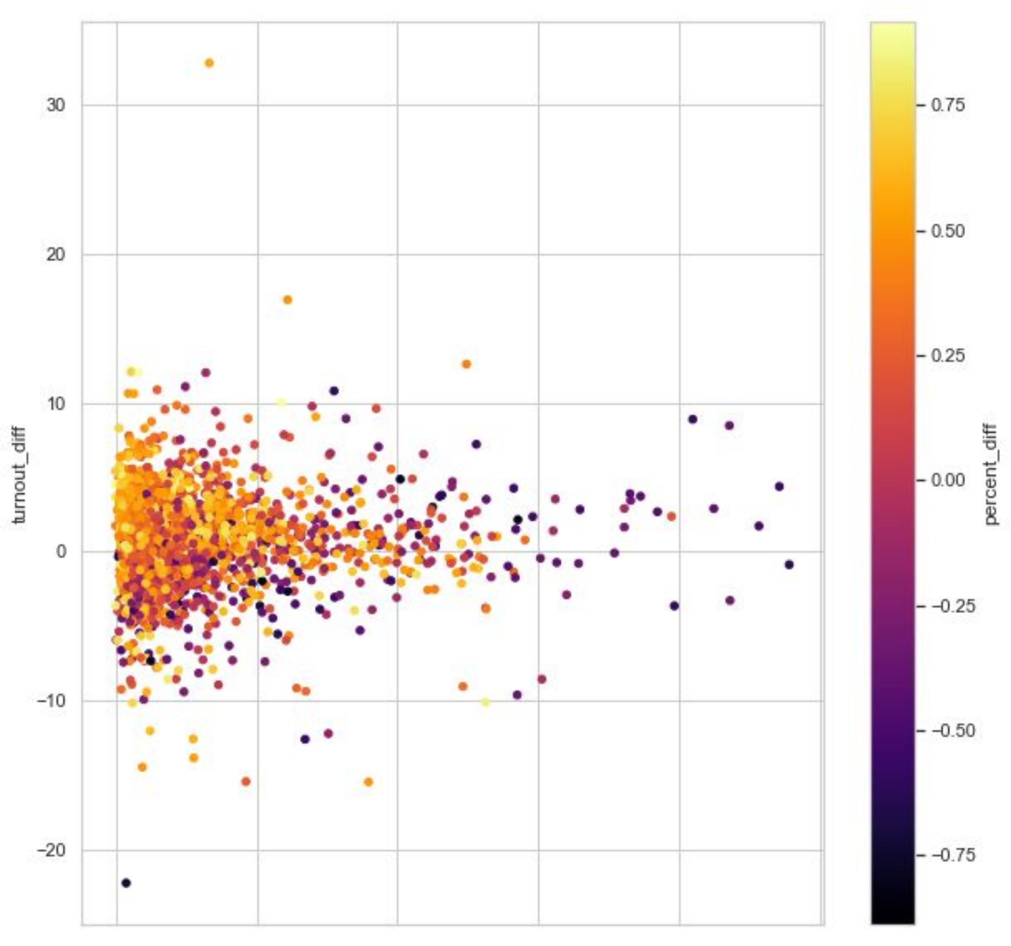
Percent of black population

Election Turnout, Plot 3

- Trends:
  - Downward trend: higher Black population counties had lower turnout
    - High density urban areas
    - Consistent with political pundits discussing lower Black turnout in post-Obama era
  - Very few counties with large black populations are high income
    - Income inequality data could be vital
    - High density counties usually also have high minority/immigrant populations

Election Turnout, Plot 3.5

- Scatter plot where x = percentage of population that speak a non-English language, y = turnout difference, and the color gradient is percentage difference between candidates
- We looked at data from Black and White voters. But who do voters with immigrant backgrounds more closely align with?



Percentage of “immigrant-background”

Election Turnout, Plot 3.5

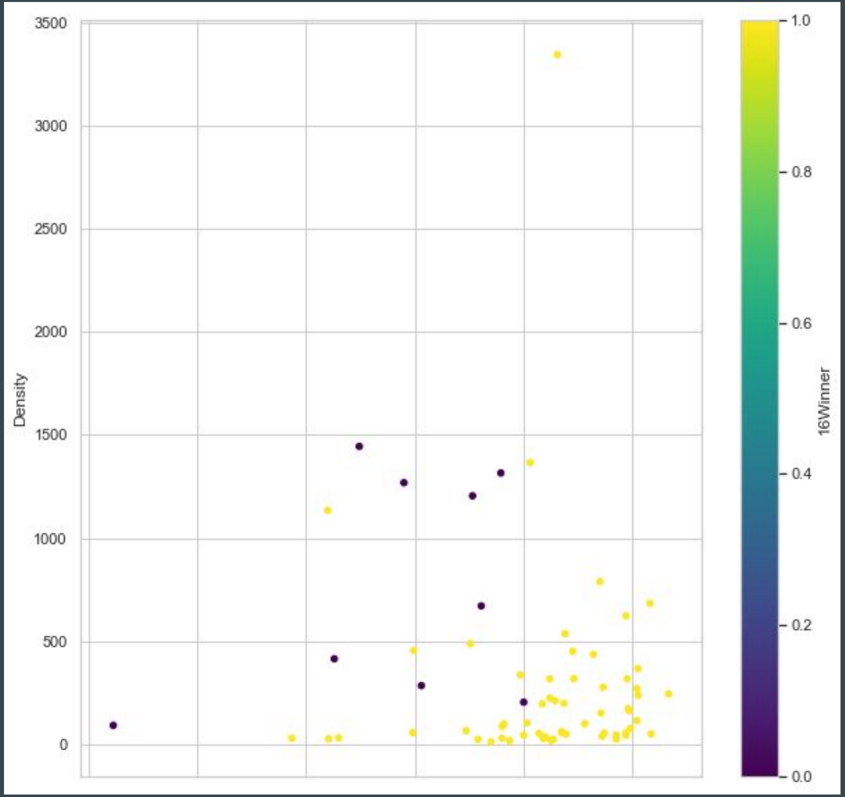
- Caveats:
  - Data set lacked detail information about voters with immigrant background
  - Used percentage of population that a speak a second language as closest indicator of new immigrant population
- Trends:
  - Consistent with Black population in candidate of choice
    - Non-White populations tend to consistently vote Democrat, regardless of
    - Consistent with political pundits discussing lower Black turnout in post-Obama era
  - In contrast with Black population, immigrant turnout had a slight upward trend
    - More immigrants turned out to the election in 2016 than in 2012

Election Turnout, Plot 4

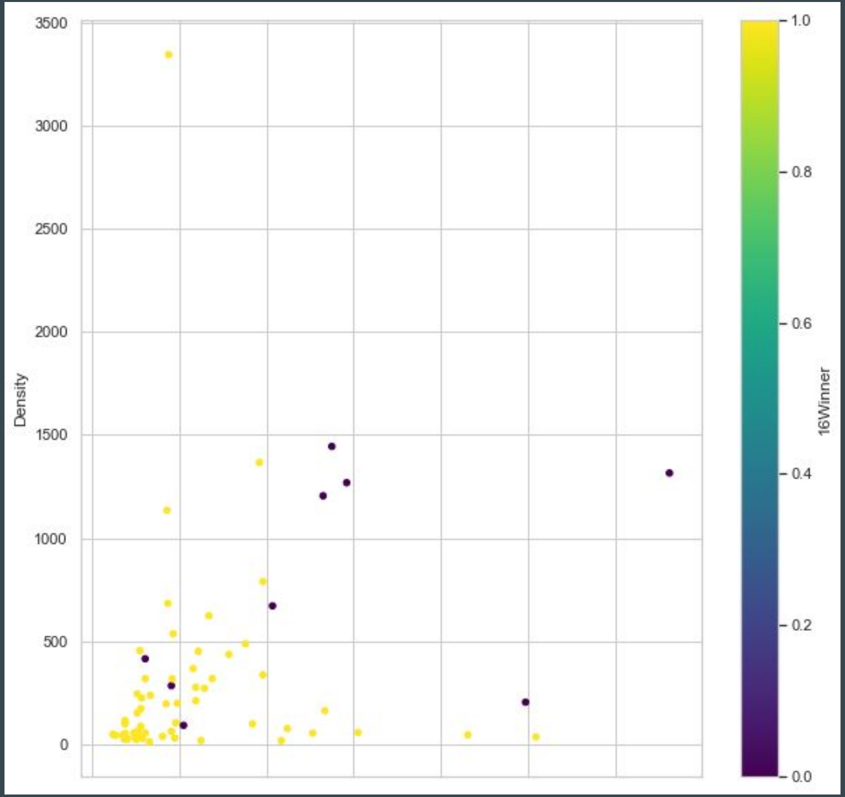
- Let’s take a deeper dive look at an important swing state, Florida
- Was initially projected to go Democrat but ended up voting for Trump
- Florida demographics are pretty diverse and actually closely reflection nationwide demographics in ethnic background and age
- Big Hispanic and retiree populations

Election Turnout, Plot 4

- The next slide has two scatter plot that compare the percentage of Hispanic and White populations vs Density, and colored differently based on which candidate won (Yellow = Trump, Purple = Hillary)



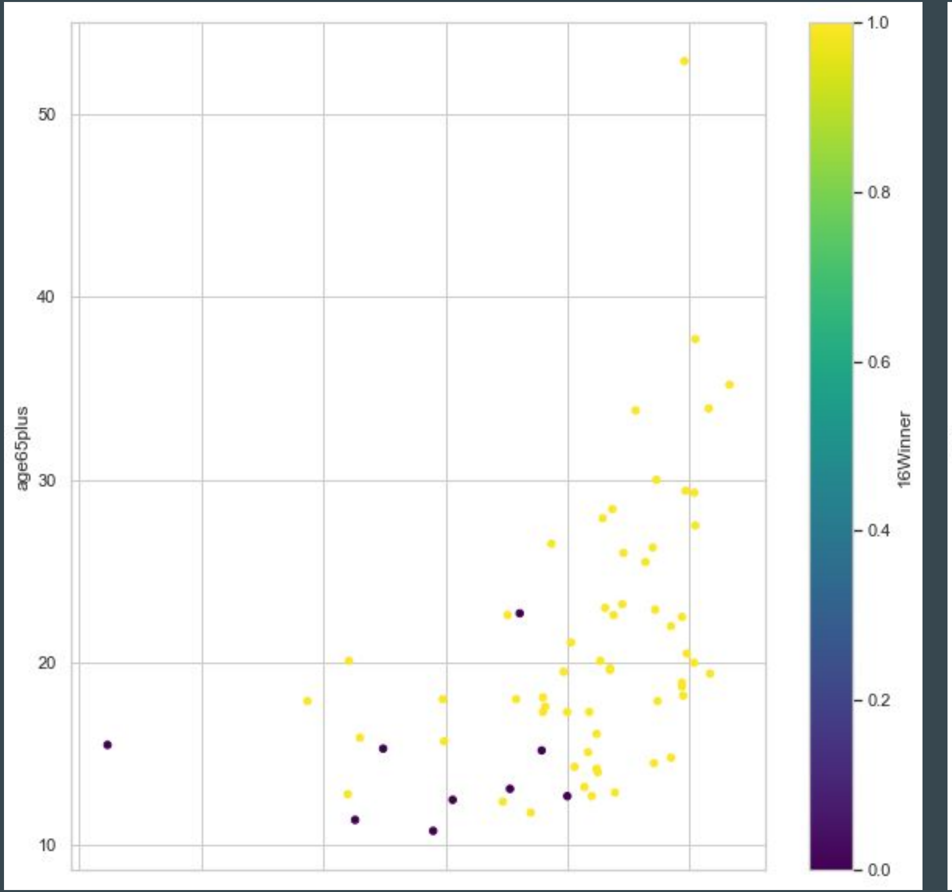
White population



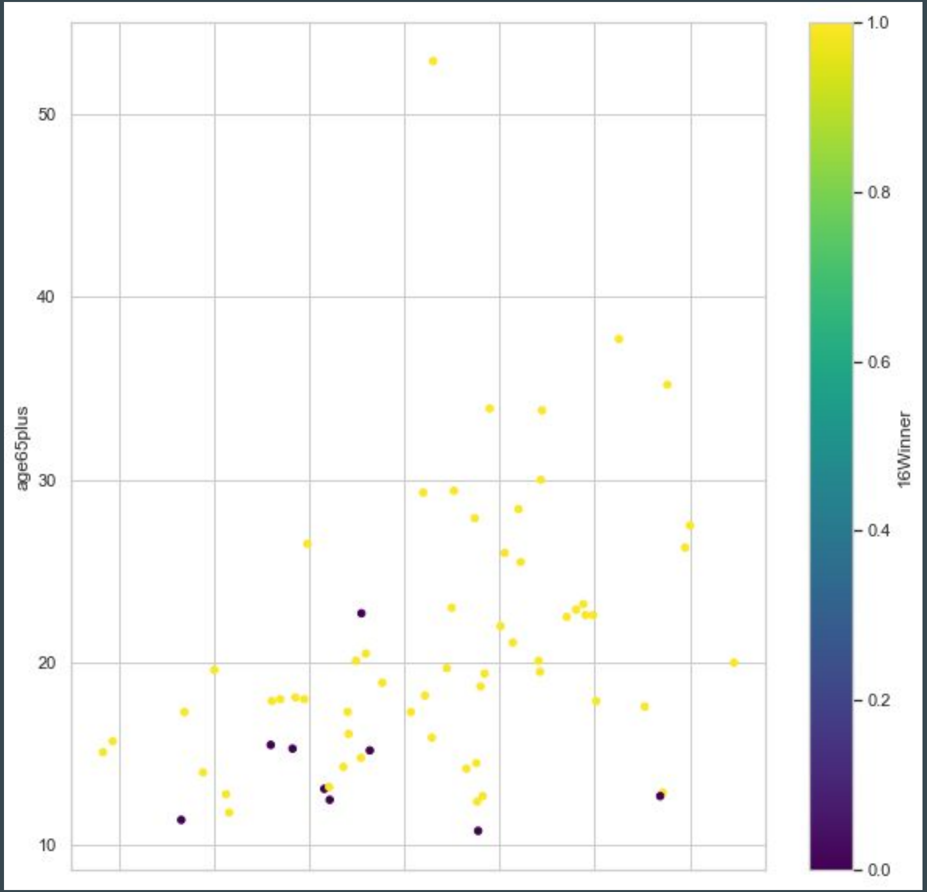
Hispanic population

Election Turnout, Plot 4.5

- The next slide has two scatter plot has x = percentage of white population/turn out, y = percentage of voters age 65+, and color for voted for Trump or Hillary (Yellow = Trump, Purple = Hillary)



White population



Turnout diff

Election Turnout, Plot 4 & 4.5

- Trends
  - Retirees, especially white, voted overwhelmingly for Trump
  - Hispanics voted overwhelmingly for Hillary
  - Florida is indicative of the situation across the country: dense, more diverse, urban areas tend to vote democrat whereas rural, sparsely populated areas tend to vote Republican

Election Turnout, Predictions

- What features can be used to predict future turnout?
- Can turnout be accurately predicted?

Election Turnout, Prediction

- First attempt: Decision Tree Classifier
  - Classify counties into two categories:
    - Higher turnout in 2016 than 2012, or lower
- Feature Engineering:
  - Already extracted useful fields
- Important features:
  - Ran classifier on a group of features to find out what features are the most important:
  - Features in order of importance:
    - Percentage difference between candidates, White, Density, Voters age 65+, Black
  - Most important features are consistent with findings

Election Turnout, Prediction Results

- Prediction accuracy:

```
[[111  93]
 [119 300]]
             precision    recall  f1-score   support


          0       0.48      0.54      0.51       204
          1       0.76      0.72      0.74       419


avg / total       0.67      0.66      0.66       623
```

Election Turnout, Prediction Results

- Analysis:
  - More accurately predicted higher turnout that lower turnout
  - The overall accuracy is still fairly low which can be due to many reasons:
    - Black voters turned out at a much lower rate than expected
    - Rural white voters turned out at a much higher rate than expected
    - Rallying for Trump amongst his supporters was much stronger than Hillary
  - Inaccurate turnout predictions could have been reason behind 2016 Dem loss as Hillary counted on "guaranteed" Black votes
- Deductions:
  - Percentage of population age 65+ is still an accurate predictor of turnout
  - The Black vote might be aligned with other minority populations but their turnout is significantly lower than immigrants or Hispanics
  - Income in a county in general does not seem to be a major indicator of where the vote is heading
  - More income analysis like income inequality between races can provide more accurate predictions

Election Turnout, Regression

- Using the same features and Decision Tree Regressor, predict turnout rate
- Results:

  Mean Absolute Error: 2.6056625996341314
  Mean Squared Error: 13.689617109159897
  Root Mean Squared Error: 3.699948257632787

- RMSE > 1, too high
  - Can be due to too many outliers
  - Model just not a great fit

Models and Specifications

•••

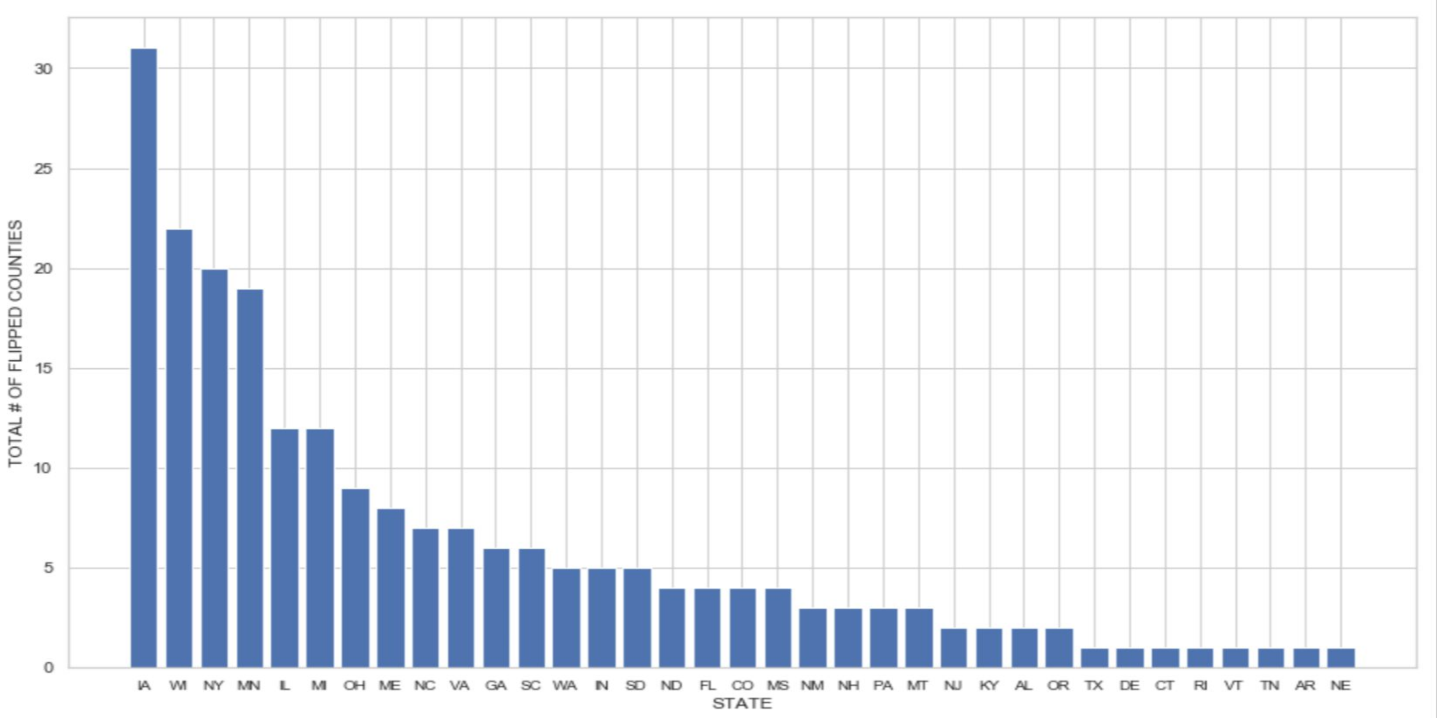
Question 2

Flipped Counties

- A huge question in the 2016 election: Were Democrats voting for Trump?
- Random forest classifier gave great results, wanted to see where the most common occurrences of flipped counties was
- From the votes dataset, best approach was to see how many flipped counties there were per state
- Looked at both:
  - Republican in 2012 (Romney) => Democratic in 2016 (Clinton)
  - Democratic in 2012 (Obama) => Republican in 2016 (Trump)

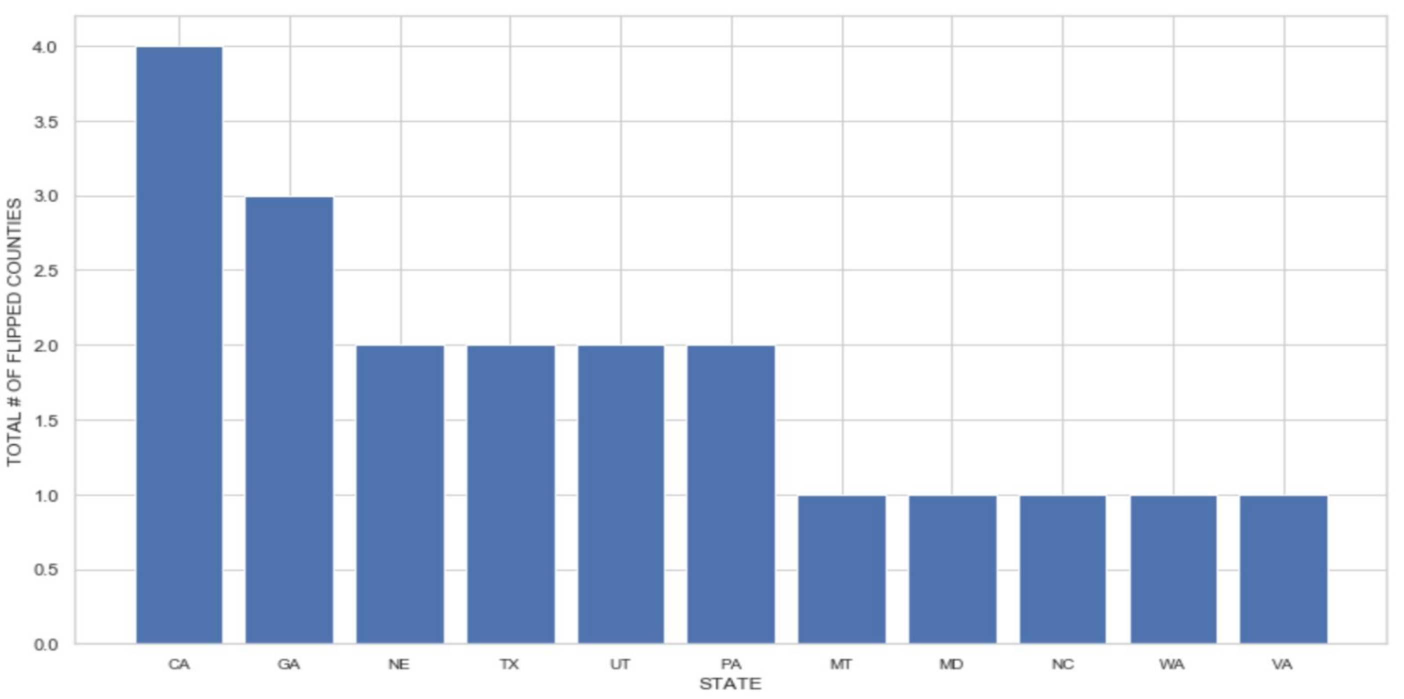
Flipped Counties

- Democratic in 2012 (Obama) => Republican in 2016 (Trump)



Flipped Counties

- Republican in 2012 (Romney) => Democratic in 2016 (Clinton)



Specifications: which seats flip

Probit

Cox Proportional Hazard Model

Instrumental Variable Regression

---Data: Historical House Election--

Feature Engineering

Presidential Election Year (mod[Year,4])

Hazard Data (time until fail, number previous fails, first failure, spell)

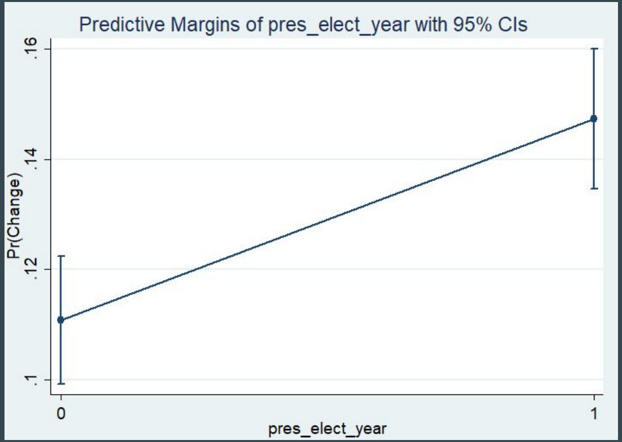
Flip or not Flip

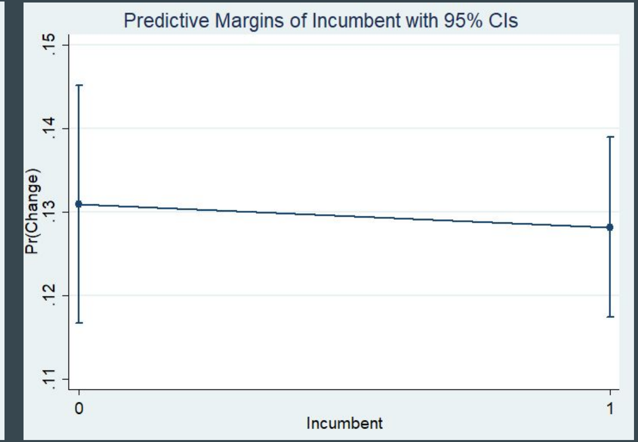
Probit Model

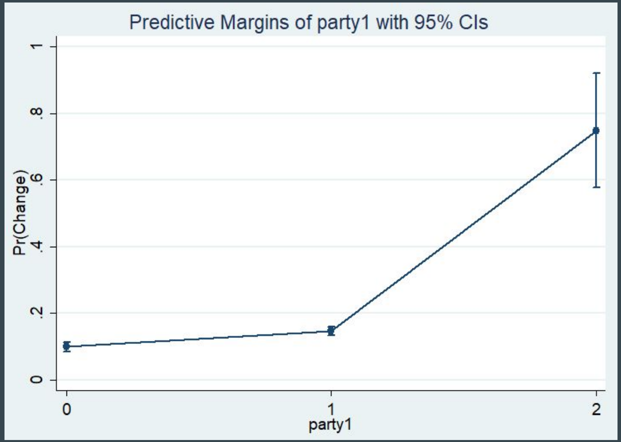
$$p = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \Phi(z)dz$$

72,319 Observations

9,309 are flips	Predicted: 3,965
63,010 are not flips	Predicted: 68,354







Hazard Model

$$\lambda_t(t_j|X_{it}, u) = \lambda_o(t)\exp(X_{it}'\beta + u)$$

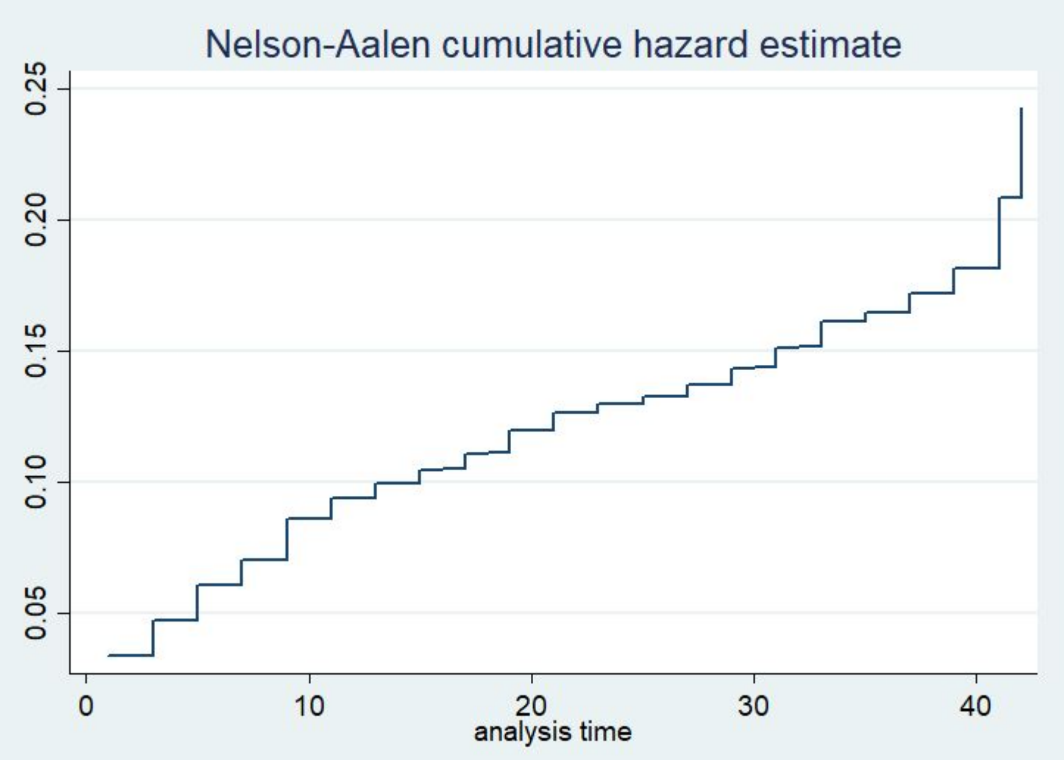
Recall:

72,319 Observations

- 9,309 are flips
- 63,010 are not flips

Success rate:

- 4,289 are flips
- 68,030 are not flips


Nelson-Aalen cumulative hazard estimate

Instrumental Variables

$$X_{it} = \beta^0 IV_{it} + \varepsilon_{it}$$

$$Y_{it} = \beta_{iv}\widehat{X_{\imath t}} + covariates + u_{it}$$

Recall:

72,319 Observations

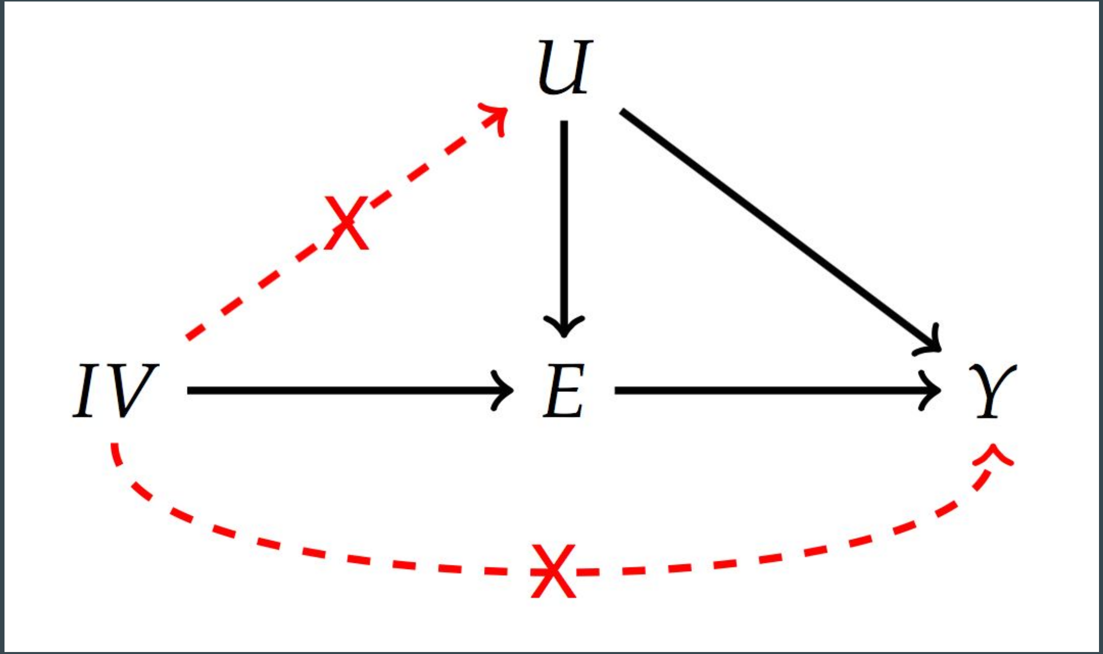
9,309 are flips

63,010 are not flips

Success rate:

6,437 are flips

65,882 are not flips



Models and Specifications

•••

Question 3

Specifications: predicting outcomes

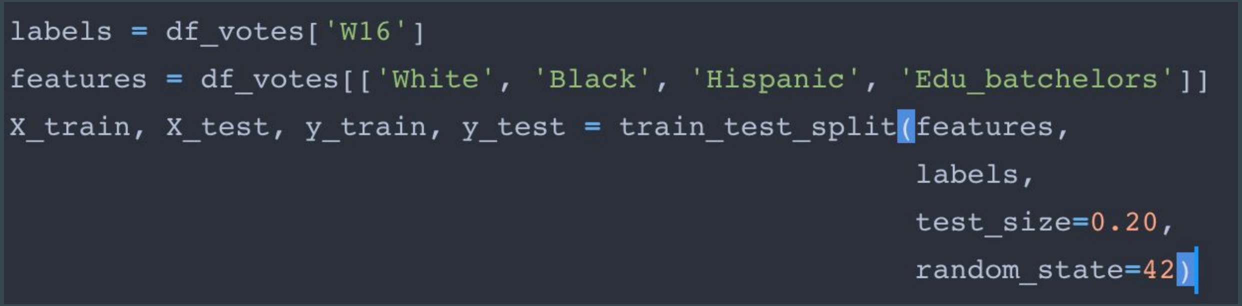
- Random Forest
- Decision Tree
- Support Vector Machine
- Markov Chain

Random Forest Classifier…Preparation

1. Data exploration
    a. Looking at data types of features
    b. Finding which features to use
    c. Feature importance
2. Feature engineering
    a. Extracting county winners
    b. Label encoding
    c. Finding combinations of features that were better than others
3. Reason for Random Forest
    a. Have used it in many projects before
    b. Generally produces good results/accurate predictions

Random Forest Results

1. Using just the previous winners as input feature (feature: ‘W12’):
   a. Accuracy rate: 93%
   b. Makes a lot of sense, a majority of counties remain unflipped every election.
2. Using % population White, % population Black, % population Hispanic, % of Edu_bachelors
   a. Accuracy rate: 90.4%
   b. Still incredibly high accuracy, even with no knowledge of past winner



```
labels = df_votes['W16']
features = df_votes[['White', 'Black', 'Hispanic', 'Edu_batchelors']]
X_train, X_test, y_train, y_test = train_test_split(features,
                                                    labels,
                                                    test_size=0.20,
                                                    random_state=42)
```

Random Forest Results

1. Overall, Random Forest worked really well for this problem
2. There is a lot of correlation between all features and the county outcomes.
3. Future work:
    a. Feature importance to see which have the most impact in this particular model
    b. Feature correlation

Support Vector Machine

- Run SVM classifier on data using features we have already determined to have high importance to predict which candidate the county will go to
- Only using two features at a time:
  - SVM can handle a lot more but disadvantage of SVM is that it’s time consuming
  - More difficult to visualize (very cool) decision maps
- Using only Linear, Gaussian, or low degree polynomial kernels because SVM takes a very long time to compute

SVM, Prediction

- Results using White and voters with immigrant background, linear kernel

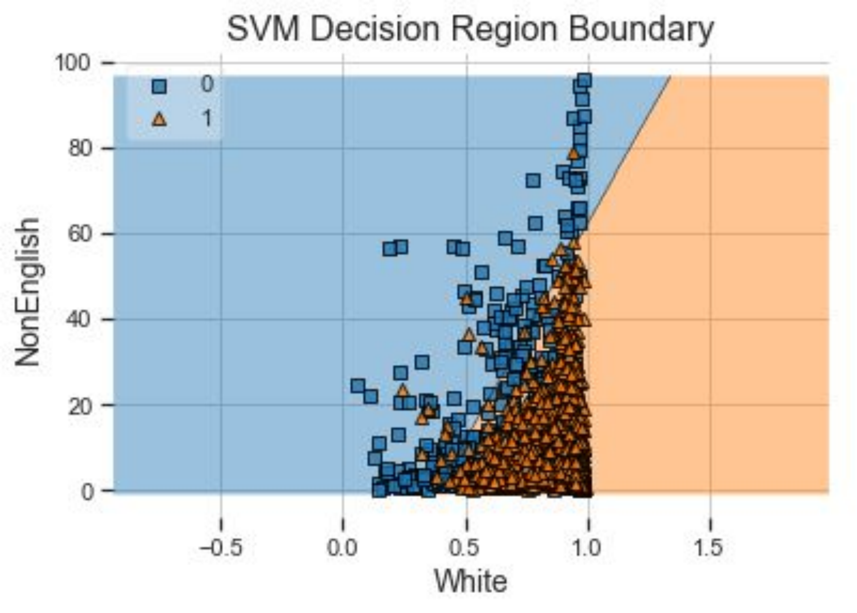
```
[[ 47  84]
 [  6 486]]
             precision    recall  f1-score   support


          0       0.89      0.36      0.51       131
          1       0.85      0.99      0.92       492

avg / total       0.86      0.86      0.83       623
```

SVM, Prediction

- Analysis:
  - 86% accuracy
  - Linear Kernel was successful because of almost clear distinction between immigrant-background voters and White voters



SVM, Prediction

- Results using percentage of voters with Bachelor’s degree and home ownership rate, polynomial kernel 3rd degree

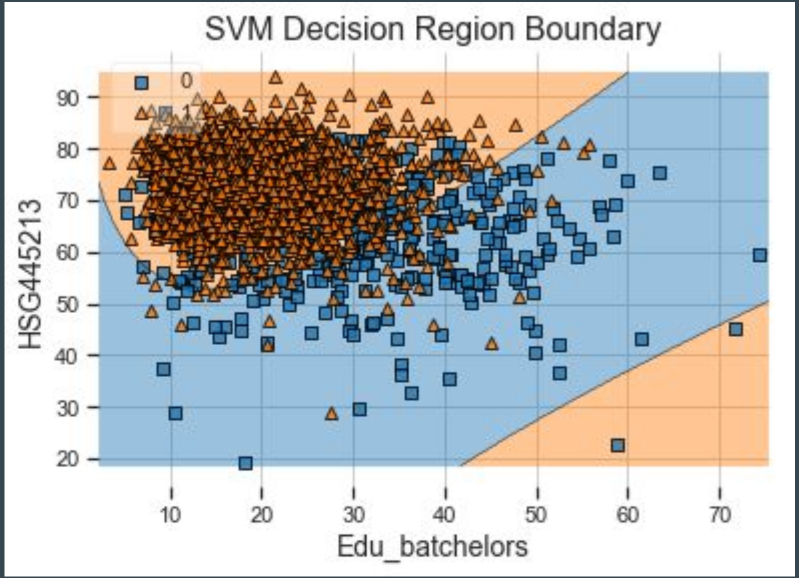
```
[[ 47  84]
 [ 16 476]]
             precision    recall  f1-score   support

          0       0.75      0.36      0.48       131
          1       0.85      0.97      0.90       492

avg / total       0.83      0.84      0.82       623
```

SVM, Prediction

- Analysis:
  - 83% accuracy
  - Homeownership here can be seen as just another measure of wealth
  - A higher education means more likely to vote Democrat, regardless of homeownership status



Decision Tree: 2016 Election Predictions

Single Feature						
Feature	Bachelors Education	%White	%Black	%Hispanic	2010 Population	2012 Winner
Score	.8122	.8138	.8218	.772	.7544	.9229

Multiple Features			
Features	Race Demographic	Race Demo + Bach Ed	Race Demographic + Bachelors Education + 2012 Winner
Score	.835	.878	.947

Election 2016

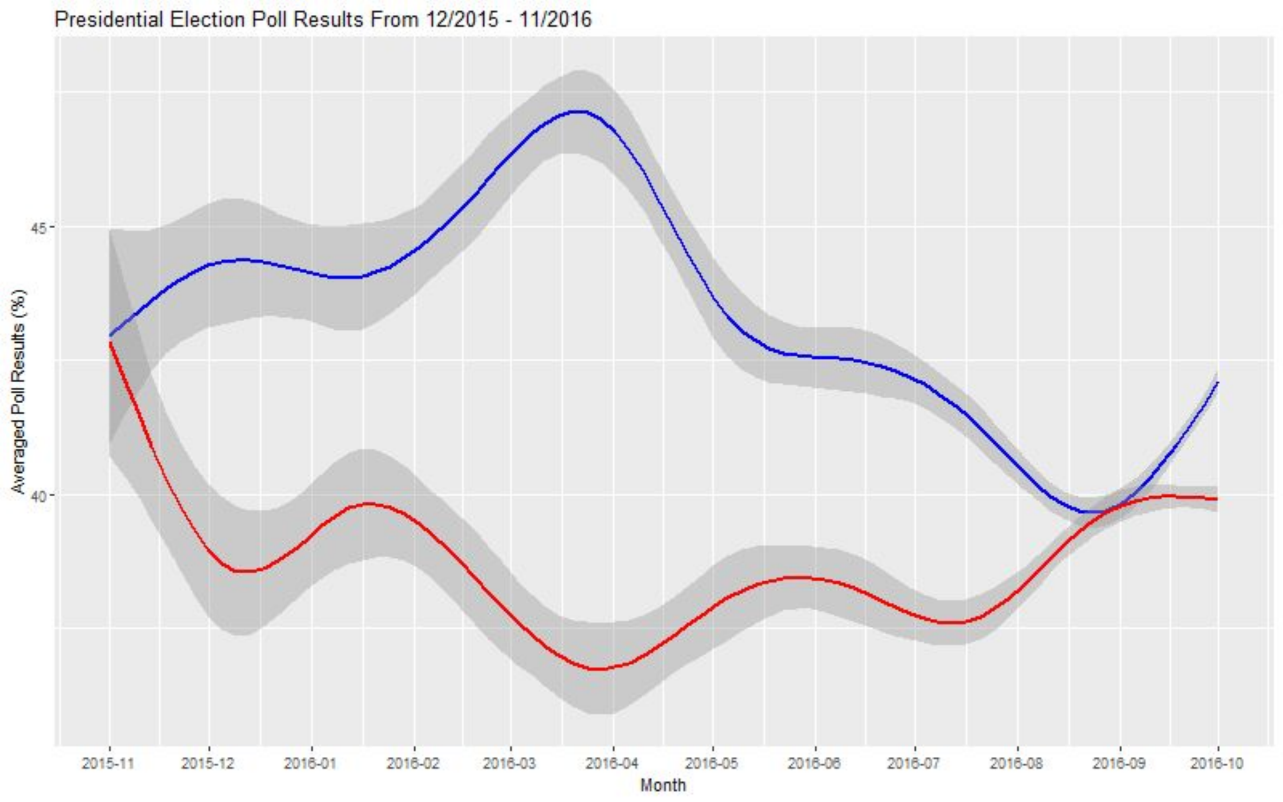
- Another question that may pop up:

Can we predict the results of the 2016 presidential election using data from polls?

- Usually polls become more frequently around one prior to the election.
- The frequency of released polls increase as the election day approaches.
- For U.S. elections, a large number of companies/institutes release polls.

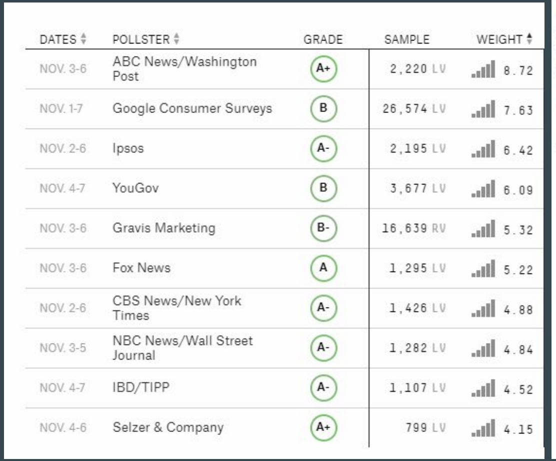
2016 Presidential Election Polls (Raw data)

- Data from **fivethirtyeight**
- Contains polls from the beginning of Oct/ 2015
- More than 10,200 polls

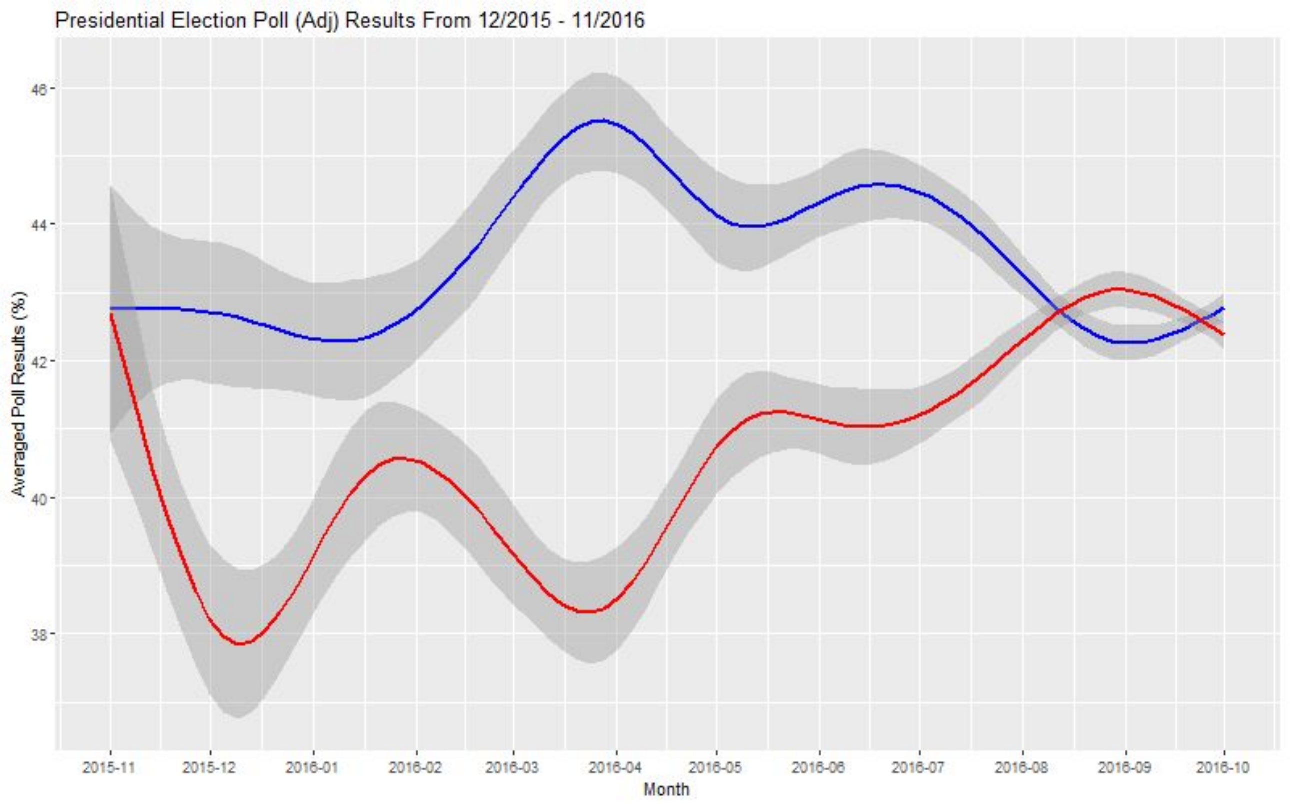


2016 Presidential Election Polls (Adjusted data)

Fivethirtyeight assigns to different companies/institute different grades for the quality of polls

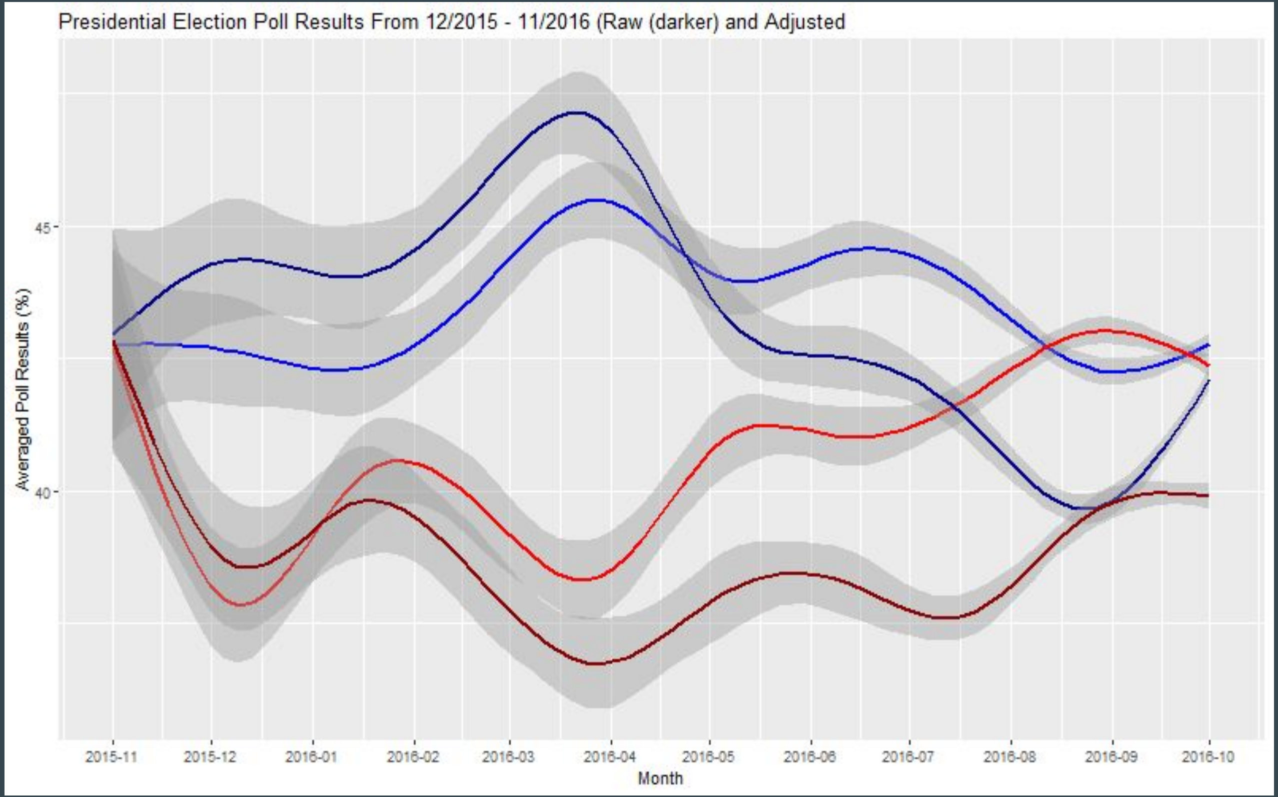


DATES	POLLSTER	GRADE	SAMPLE	WEIGHT
NOV. 3-6	ABC News/Washington Post	A+	2,220 LV	8.72
NOV. 1-7	Google Consumer Surveys	B	26,574 LV	7.63
NOV. 2-6	Ipsos	A-	2,195 LV	6.42
NOV. 4-7	YouGov	B	3,677 LV	6.09
NOV. 3-6	Gravis Marketing	B-	16,639 RV	5.32
NOV. 3-6	Fox News	A	1,295 LV	5.22
NOV. 2-6	CBS News/New York Times	A-	1,426 LV	4.88
NOV. 3-5	NBC News/Wall Street Journal	A-	1,282 LV	4.84
NOV. 4-7	IBD/TIPP	A-	1,107 LV	4.52
NOV. 4-6	Selzer & Company	A+	799 LV	4.15

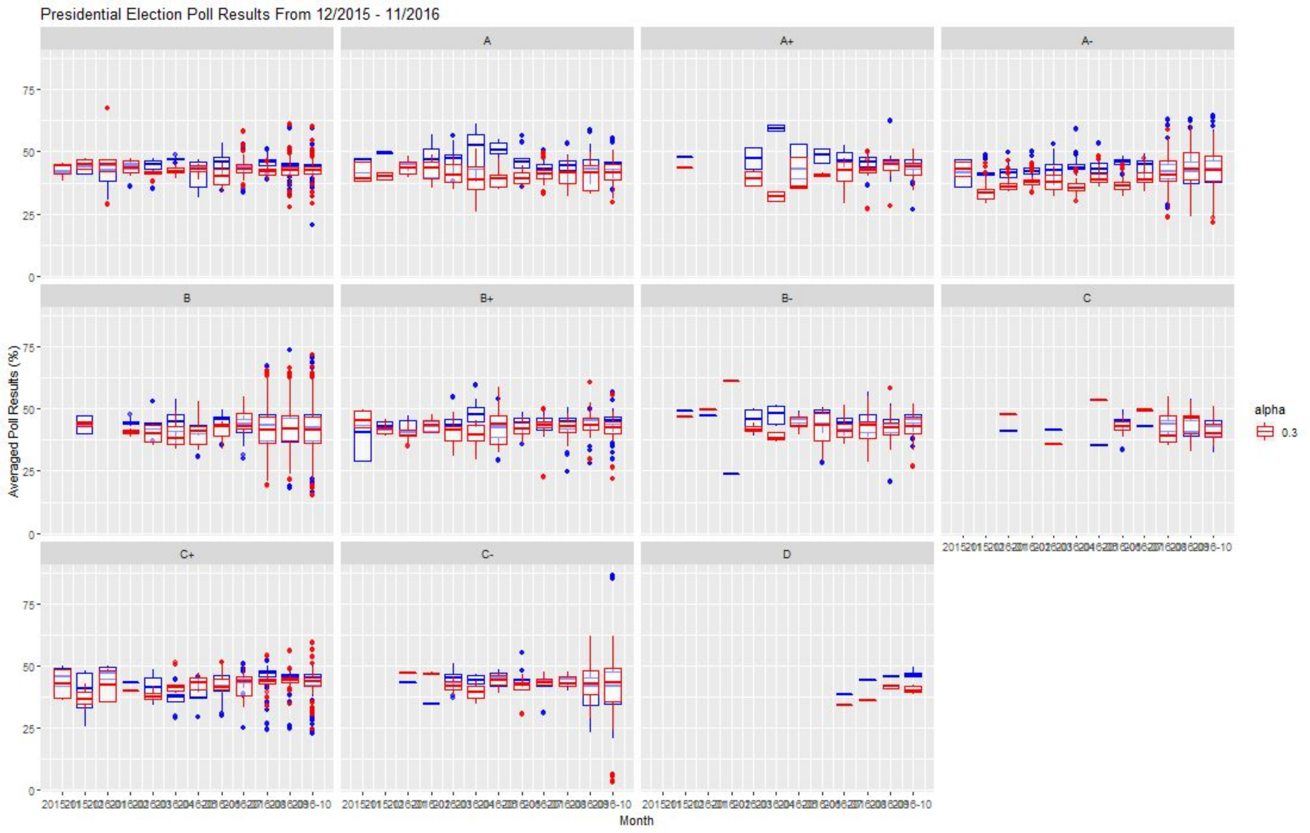


Presidential Election Poll (Adj) Results From 12/2015 - 11/2016

2016 Presidential Election Polls (Raw and Adjusted data)



2016 Presidential Election Polls (Adjusted data)



ARIMA Models

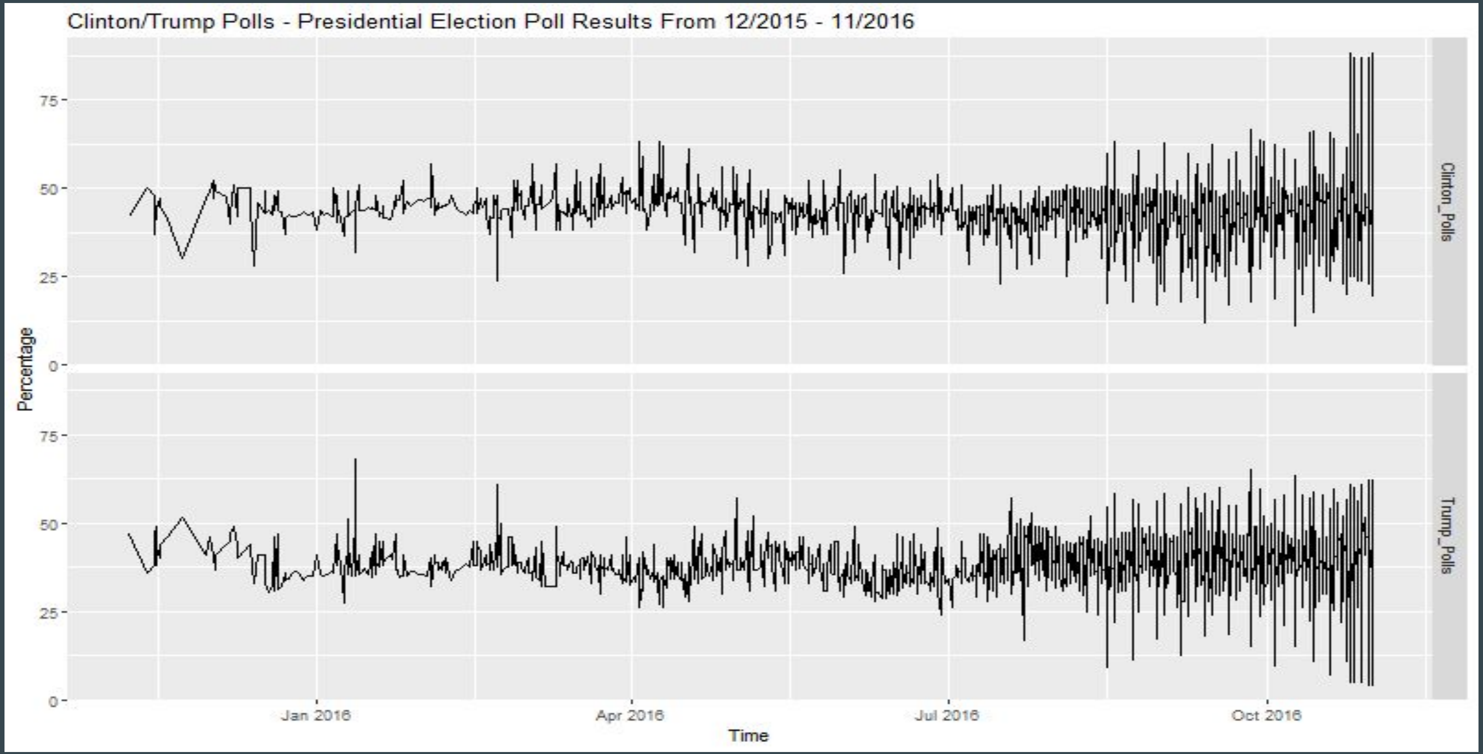
- Among the standard statistical methods, this is the most **suitable**

- We can’t decompose using seasonality models

- Our data is not evenly spaced

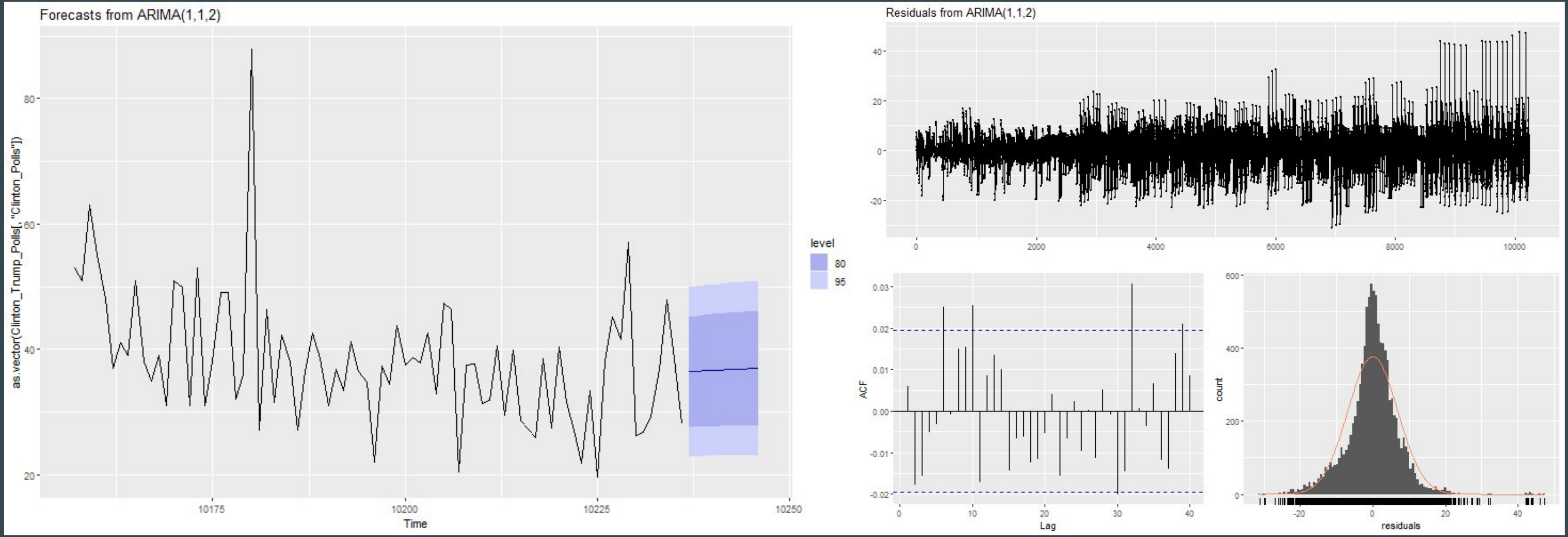
- ARIMA models aim to describe the autocorrelations in the data.

Clinton/Trump Polls Time Series (2015 - 2016)



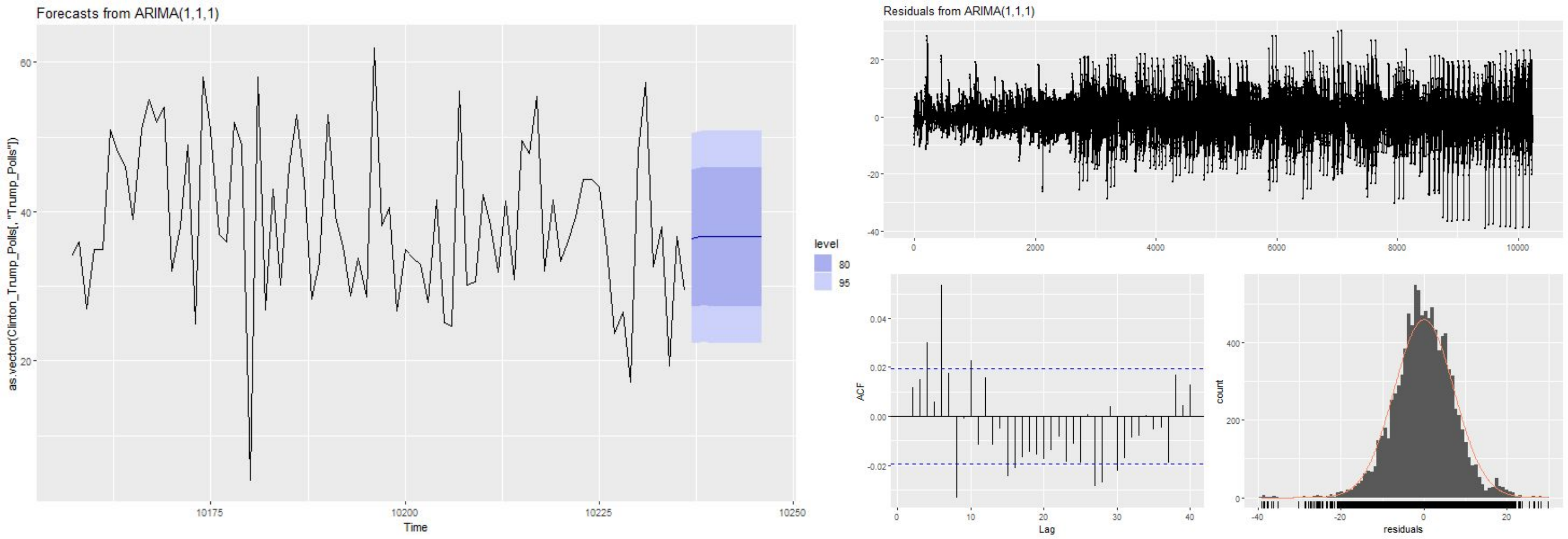
Fitted ARIMA for Clinton Polls

- Best fit: ARIMA (1,1,2)



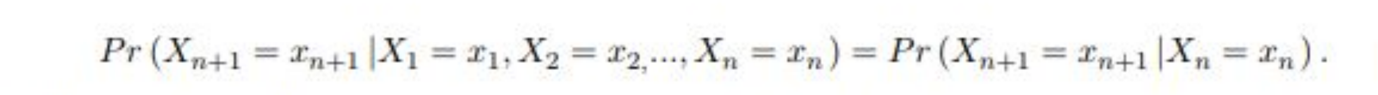
Fitted ARIMA for Trump Polls

- Best fit: ARIMA (1,1,1)



Markov Chains

- **Markov Property:** the distribution of the forthcoming state X_{n+1} depends only on the current state X_n and doesn't depend on the previous ones



$$Pr\left(X_{n+1}=x_{n+1}\,|X_1=x_1,X_2=x_2,\ldots,X_n=x_n\right)=Pr\left(X_{n+1}=x_{n+1}\,|X_n=x_n\right).$$

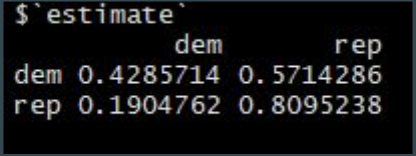
- For our problem we have 2 states (perhaps 3?):
  - S={s1, s2}={democratic, republican}

Markov Chains

- The transition probability (p_{ij}) is given by: $$p_{ij} = Pr\left(X_1 = s_j \middle| X_0 = s_i\right).$$
- For instance p_{ij} could represent the probability of transition from democratic to republican.
- In this case we want $p_{\text{dem,rep}}$
- Having the transition probability, we can build the transition matrix:

$$P = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}.$$

- For example, for Wyoming next election we estimate the transition matrix to be:

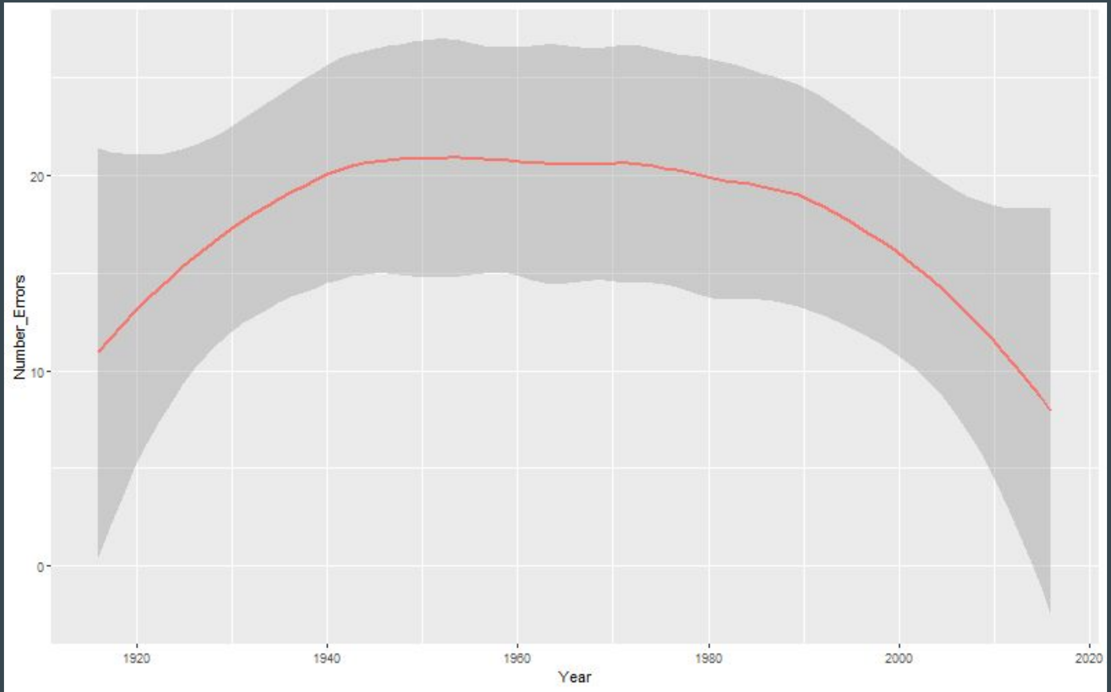


```
$`estimate`
          dem       rep
dem 0.4285714 0.5714286
rep 0.1904762 0.8095238
```

Markov Chain model - Number of Errors per Election
(from 1912 to 2016)

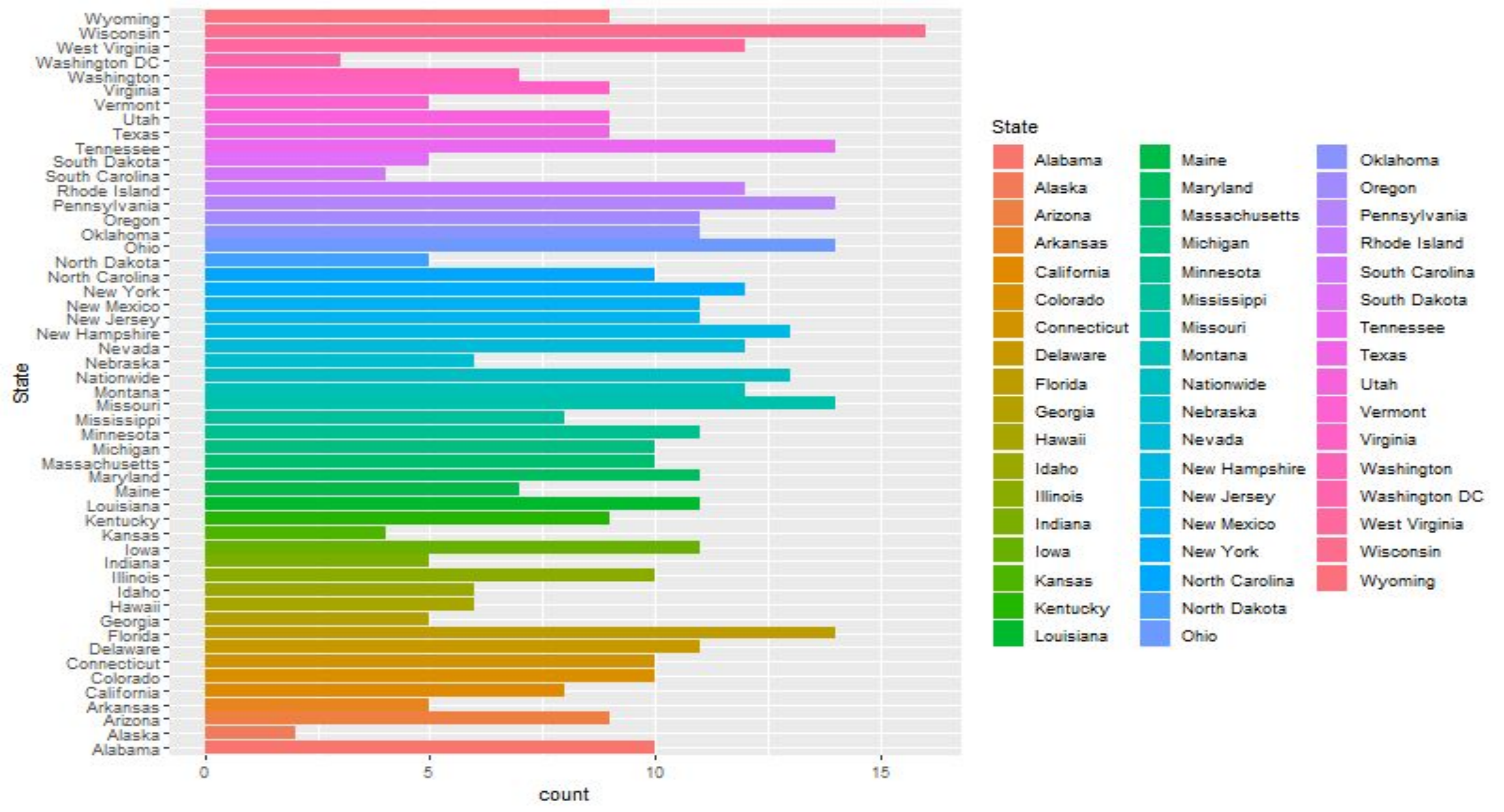
- We created a loop that updates the **Transition Matrix** after each election;

- Start forecasting from the 3rd election in our sample: **1912**

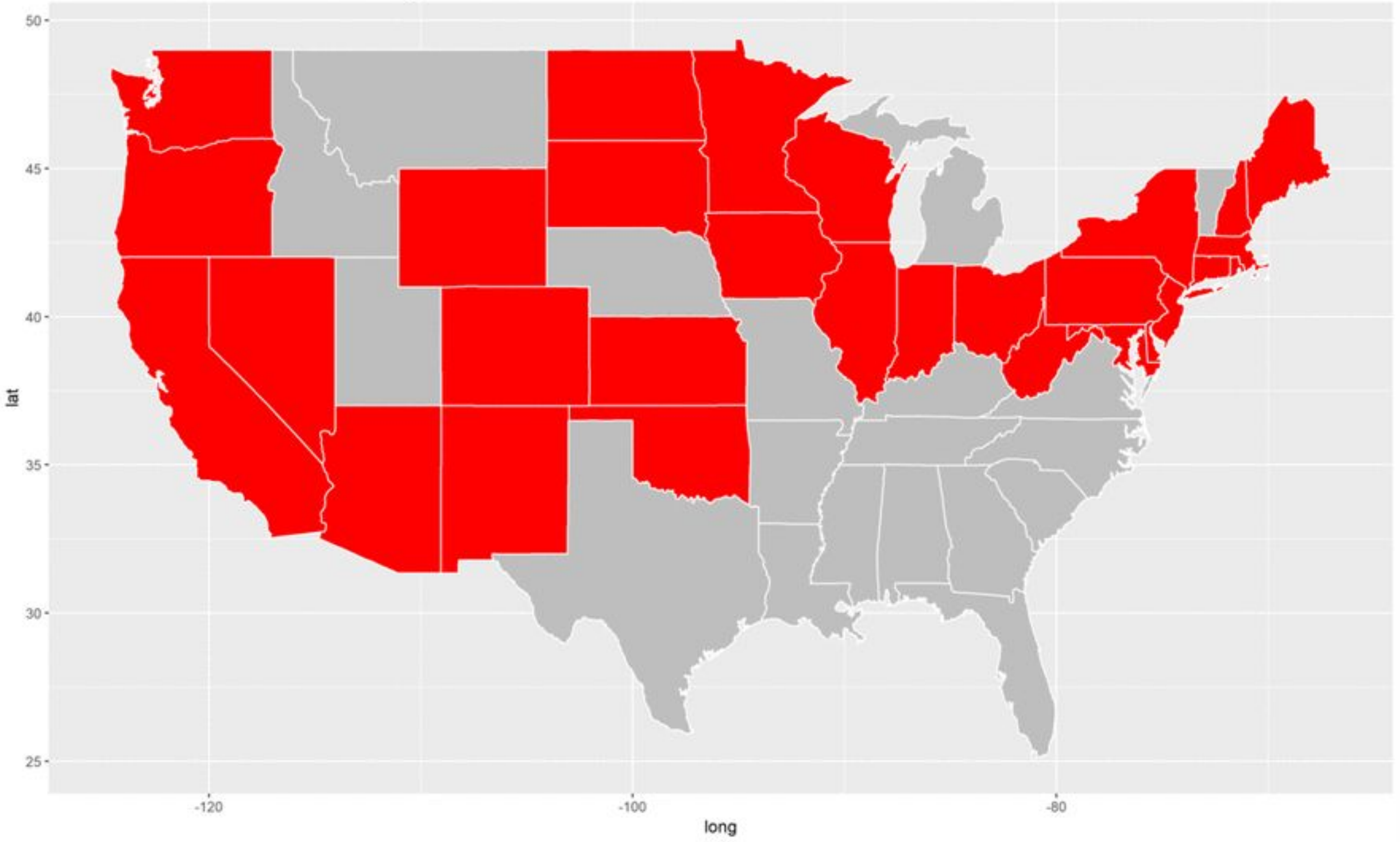


Number of Errors (Markov Chain)

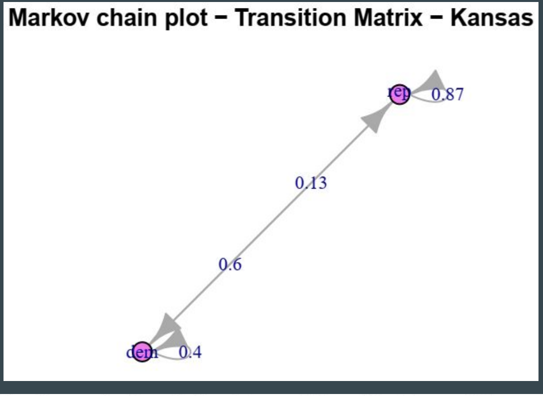
(Elections from 1912 to 2016)



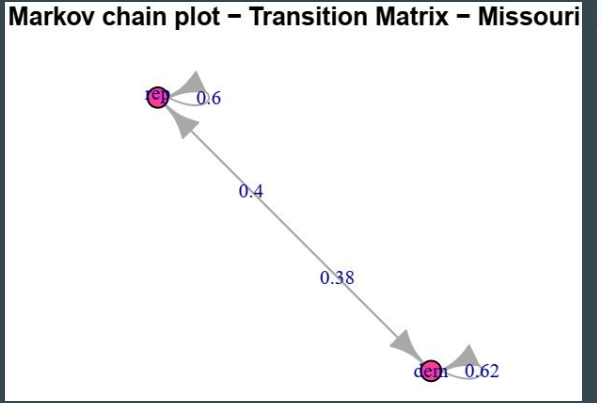
1908 - Presidential Election Errors using Markov Chain



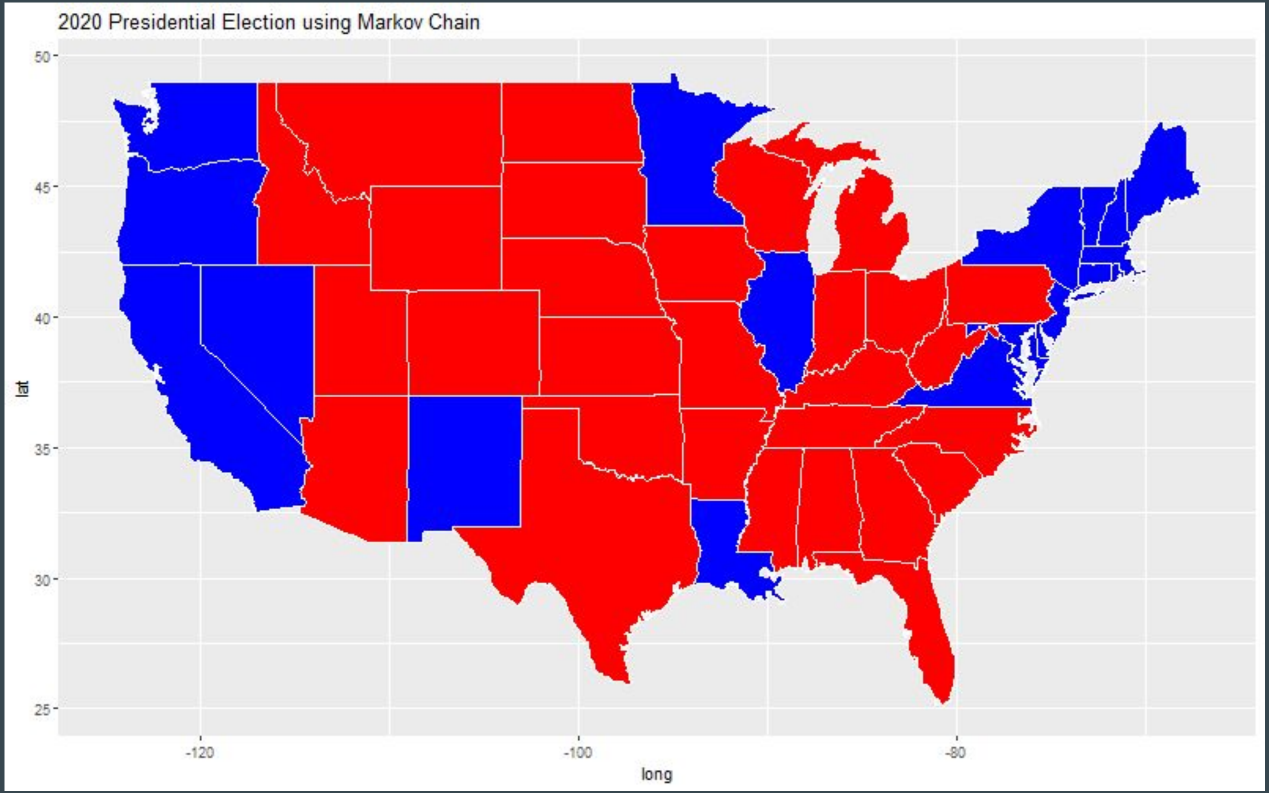
2020 Presidential Election Prediction



Markov chain plot − Transition Matrix − Kansas
rep
0.87
0.13
0.6
dem
0.4



Markov chain plot − Transition Matrix − Missouri
rep
0.6
0.4
0.38
dem
0.62



2020 Presidential Election using Markov Chain
50
45
40
35
30
25
lat
-120
-100
-80
long