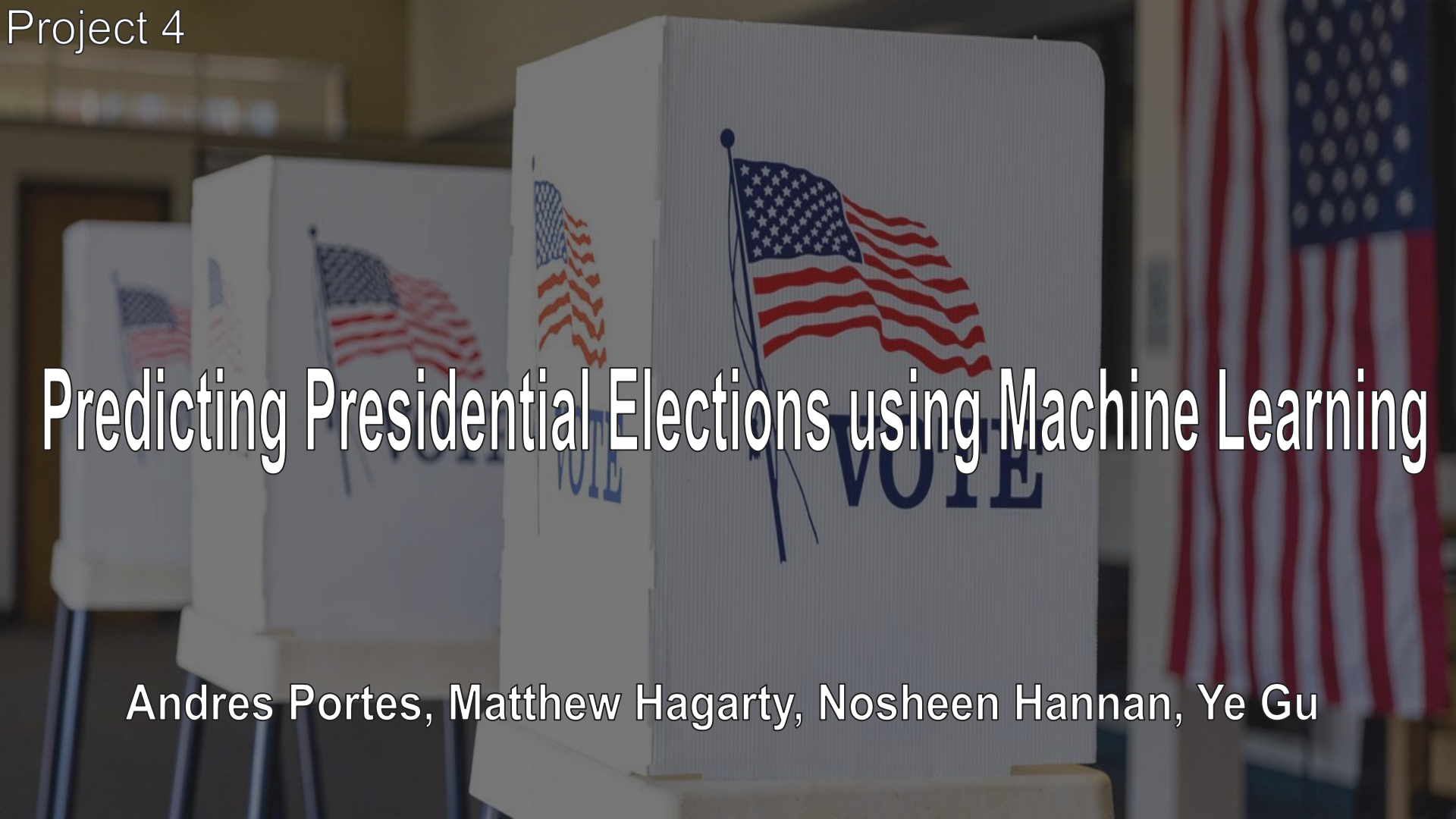
Project 4



Predicting Presidential Elections using Machine Learning

Andres Portes, Matthew Hagarty, Nosheen Hannan, Ye Gu

Overview

- Objective
- Data Sources
- Data Preprocessing Steps
- Machine Learning Methods: Logistic vs. Linear Regression.
- Model Evaluation
- Regression Results
- Key Takeaways & Future Improvements
- Q&A session.

Objective

Our objective is to understand and predict the percentage of votes earned for each party by state in a presidential election by leveraging both census data and historical election results. Accurate predictions of voting behavior can provide valuable insights into the political landscape and help political parties refine their campaign strategies. By analyzing the demographic factors from census data and historical voting patterns, we aim to uncover the underlying drivers of political outcomes and make informed predictions.

Data Sources

- **Census Data**: Collected data from the Census Bureau
- **Election Results Data**: Obtained 2000-2020 past presidential elections results by county (dataverse.harvard.edu)

Data Preprocessing

- Scale the data by numerical values converting to percentage values
- Census Data:
  - Grouped into four categories:
    - Gender
    - Race (White, Non-White)
    - Hispanic/Non-Hispanic
    - Age (Under 50 but at least voting age)
  - These categories were chosen to capture demographic factors that influence voting behavior
- Election Result: percentage voted Republican vs. Democrat

Methods Explored

We explored various data details and machine learning methods, including:

- Considered County/State data
    i. Granularity
    ii. Focused on State level data
- Logistic Regression (Categorical): Binary analysis. Winner: 1=Democrat, 0 = Republican
- Linear Regression (Numerical): predicting numerical vote percentages.
    i. Multiple Linear Regression: multiple variables
    ii. Polynomial Regression: Relationship between variables is nonlinear
    iii. Assessed model performance using Mean Squared Error (MSE)

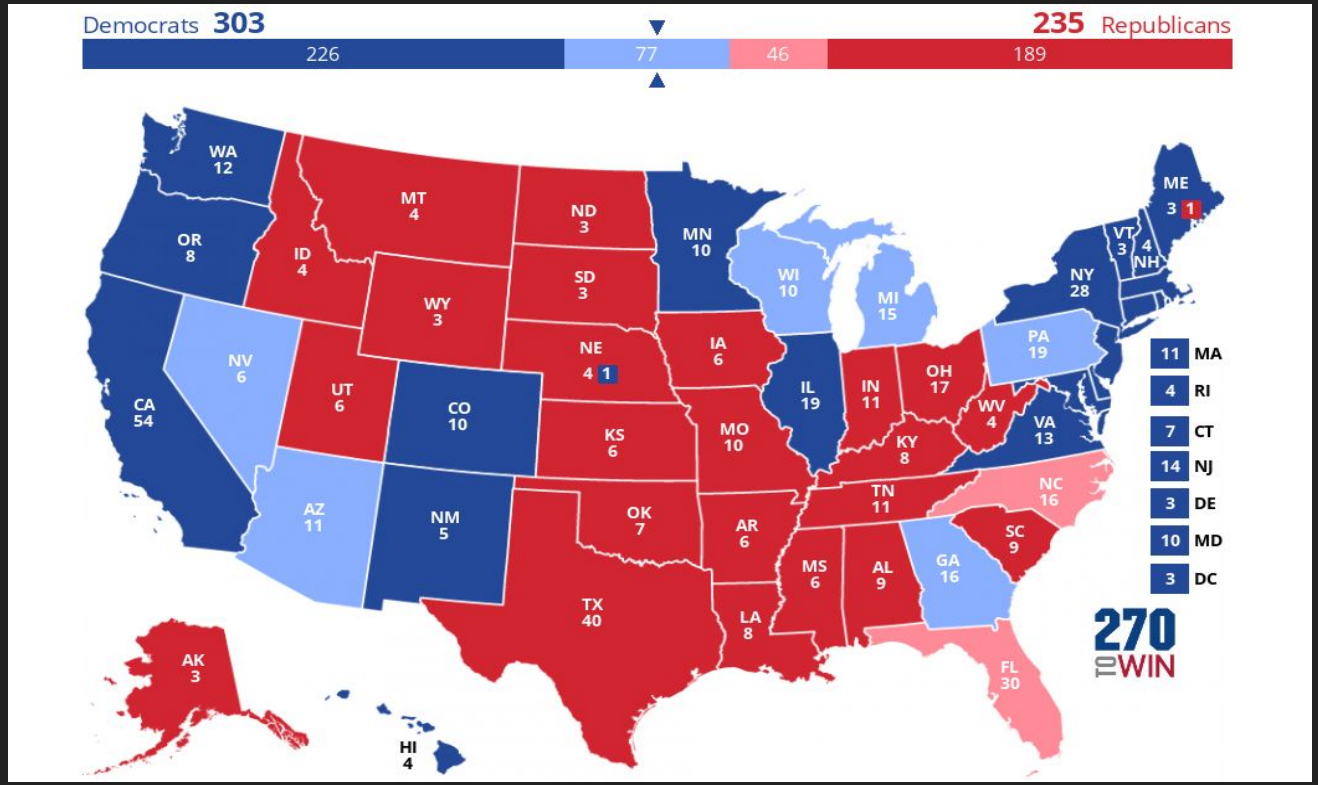
Historically Democrat vs. Republican

- Include as feature
  - Model accuracy was higher
- Not included as feature, model accuracy was lower
- Considerations of including this feature
  - Historical vote is strong predictor of the election outcome
  - Relevance to our goal: goal is not to understand the relationship between voting patterns and election outcome
- Overall we chose not to include this data for our project

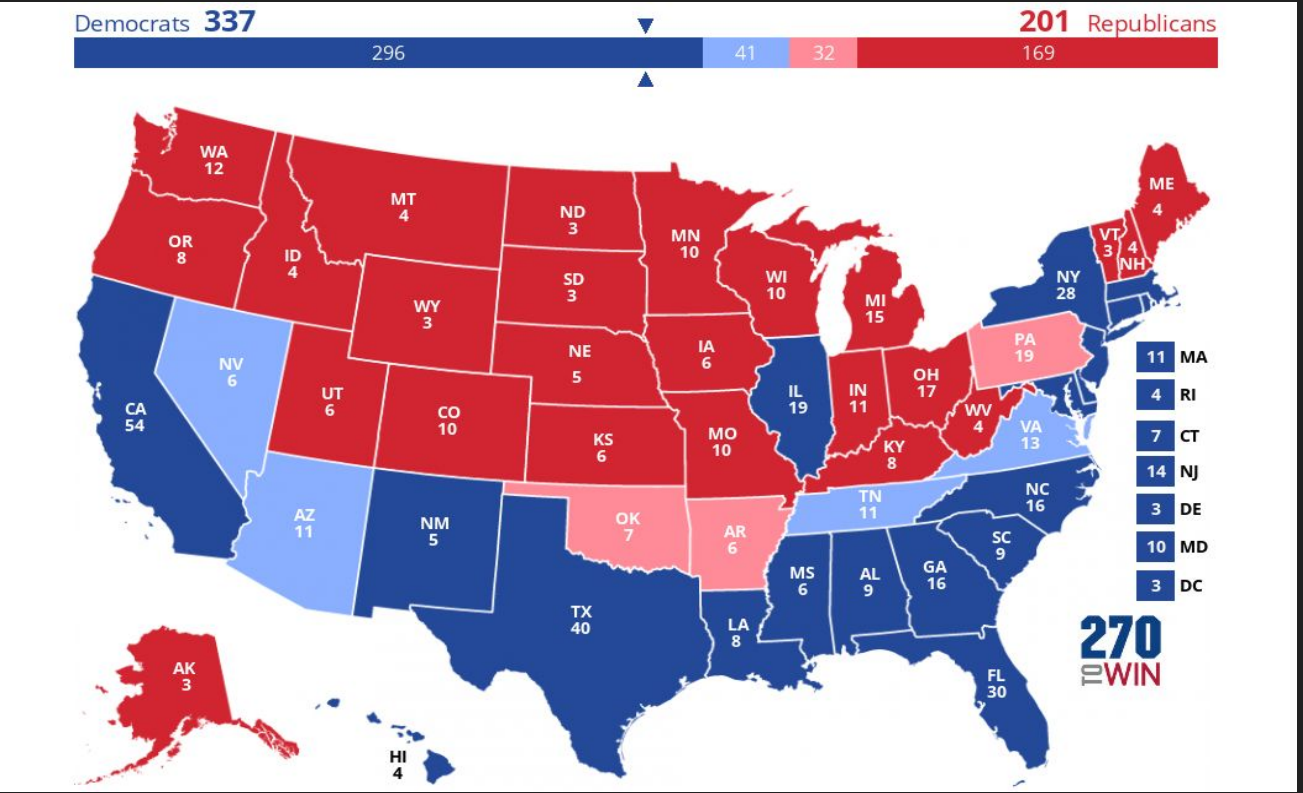
Linear Regression

- Multiple Linear Regression:
  - Mean Squared Error: 0.009046195543439153
- Polynomial Regression:
  - Mean Squared Error of Degree 2 Polynomial: 0.0067626505678602 4
  - Mean Squared Error of Degree 3 Polynomial: 0.0056900610182295 6
  - Mean Squared Error of Degree 4 Polynomial: 0.0054117217248442 6

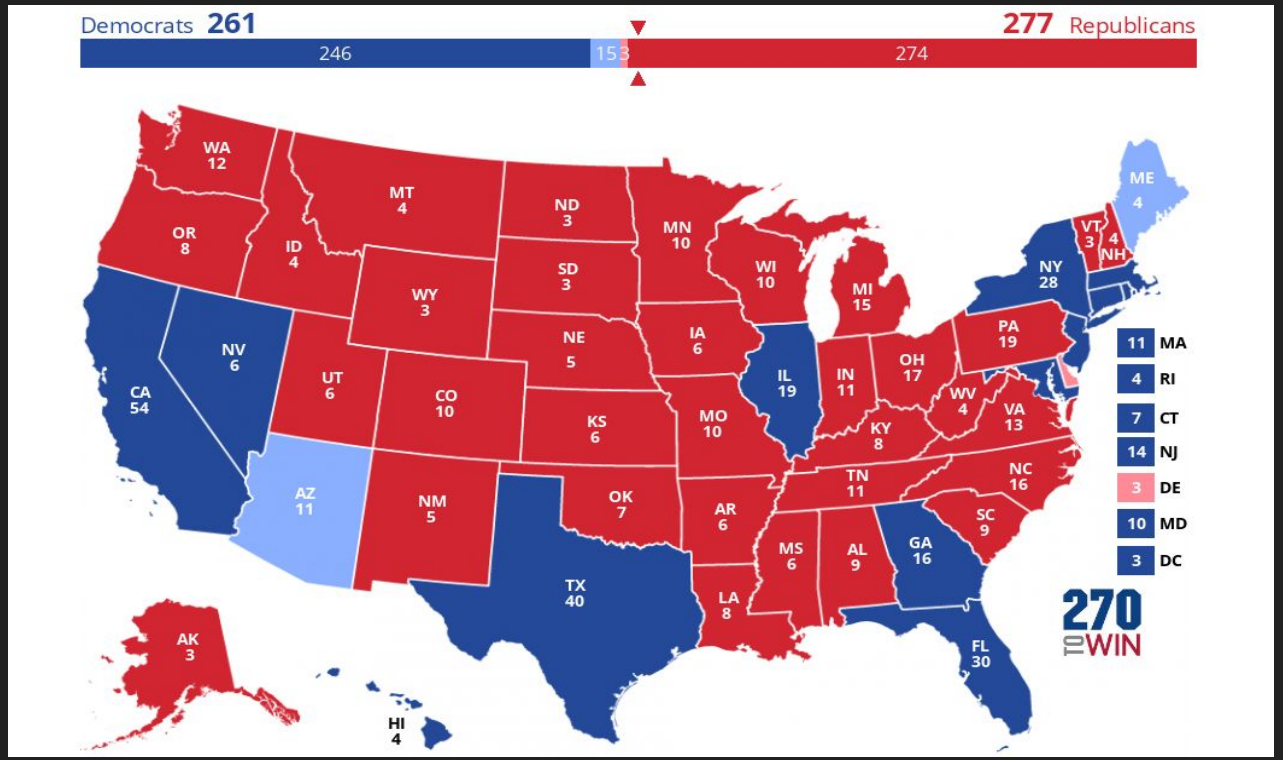
Actual 2020 Election Results



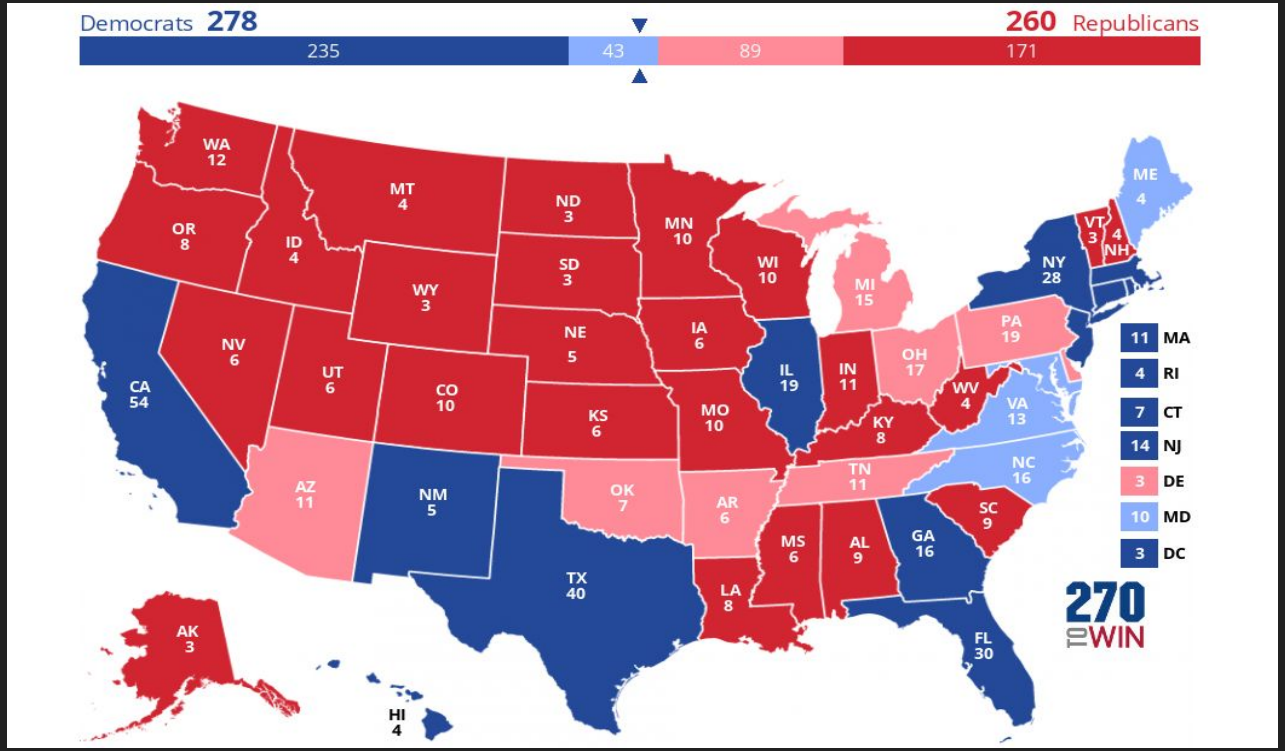
2020 Linear Predictions



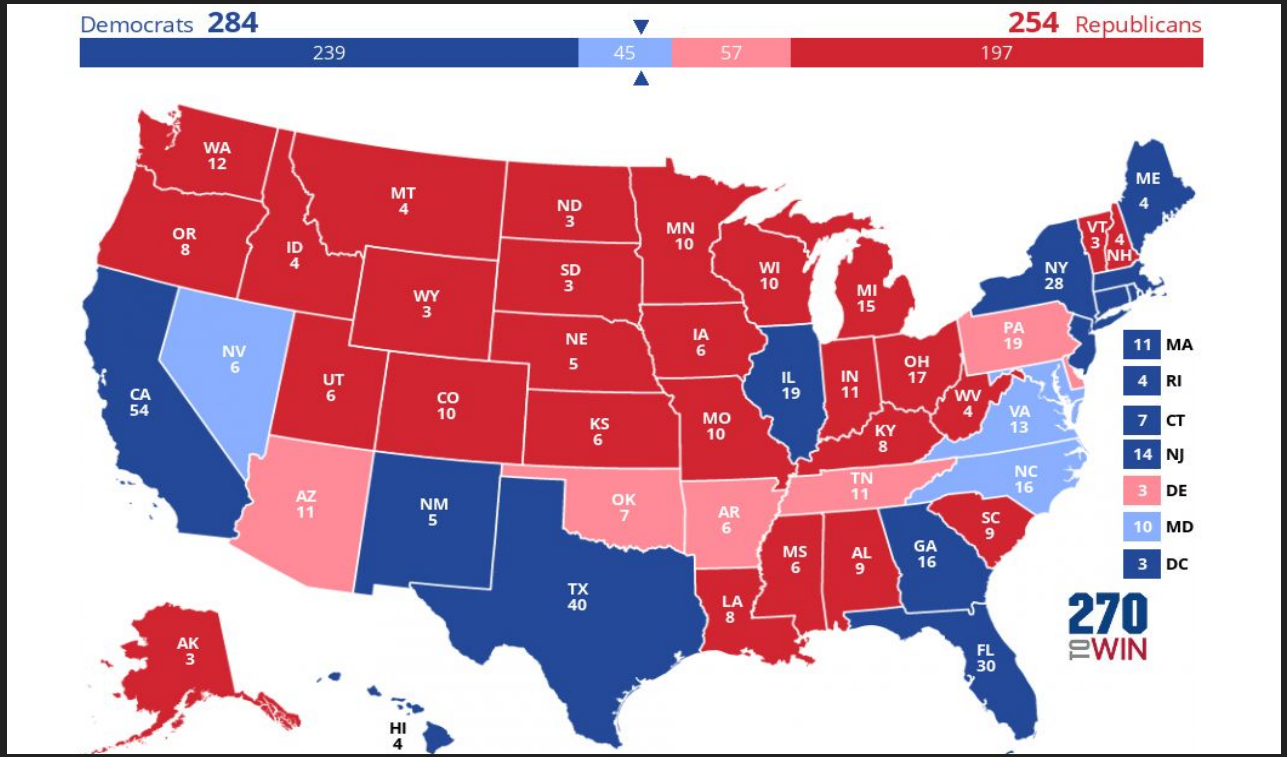
2020 Polynomial Degree 2 Predictions



2020 Polynomial Degree 3 Predictions



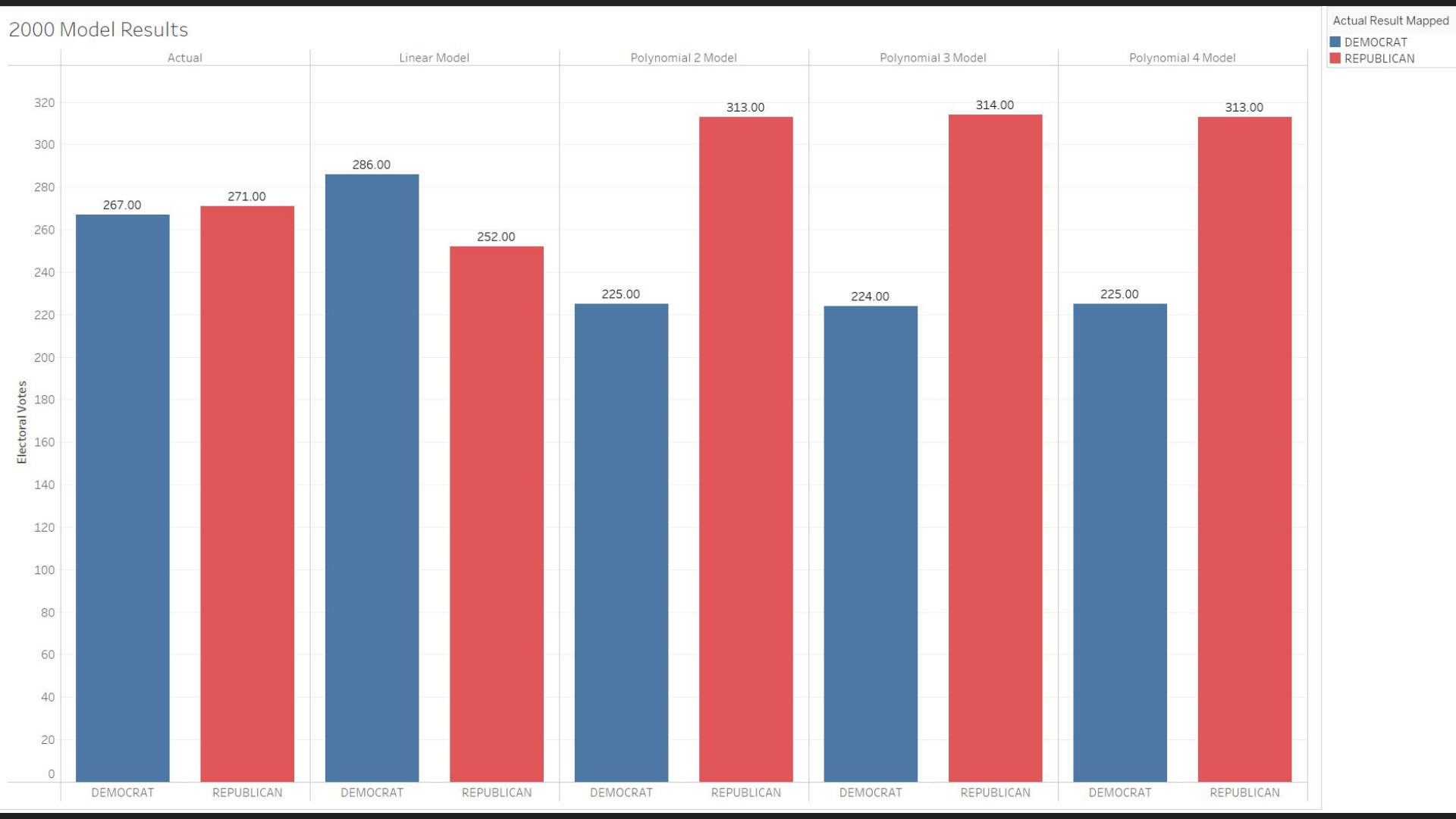
2020 Polynomial Degree 4 Predictions



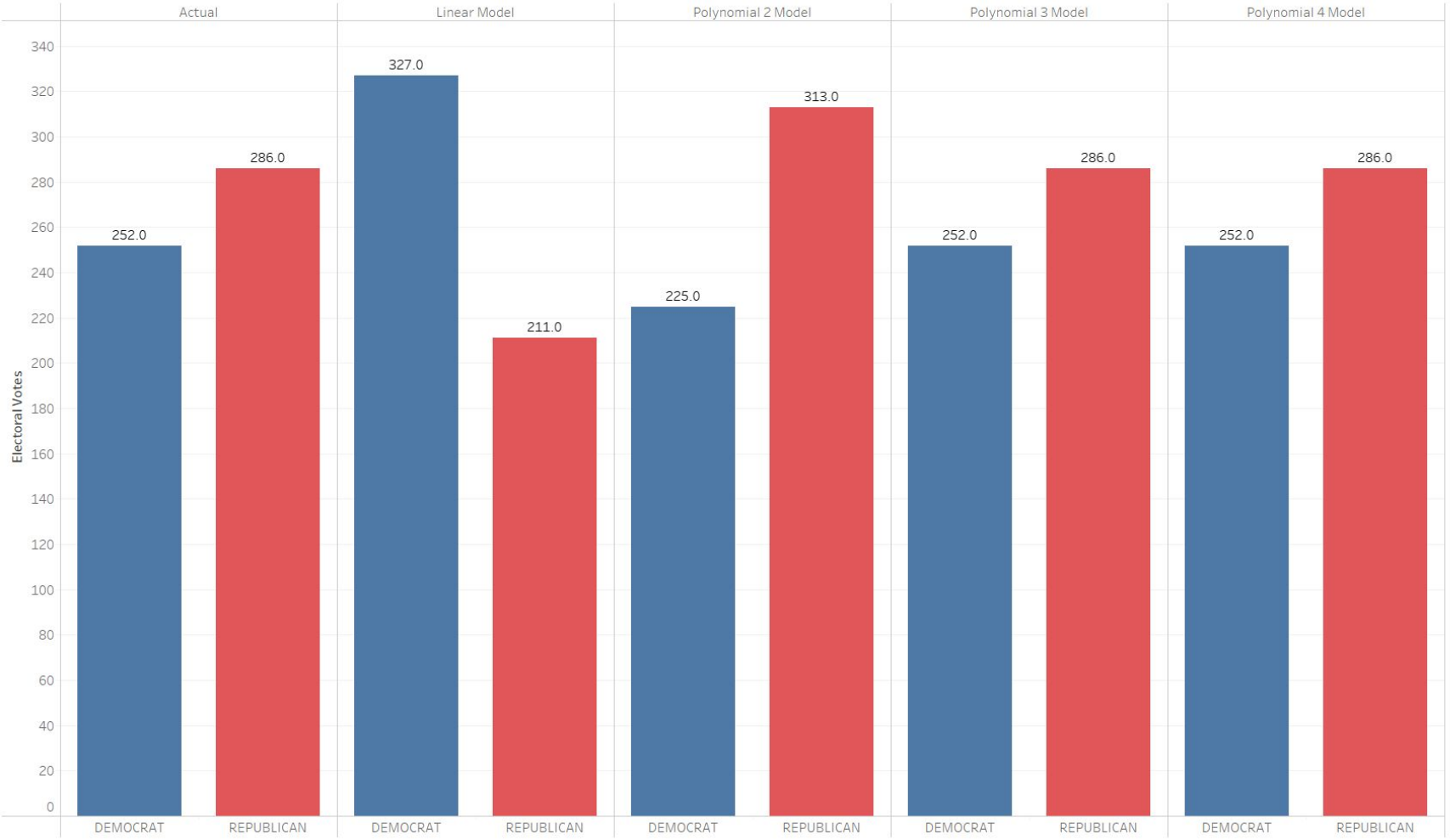
Thoughts on 2020 Prediction Maps

- Linear Model was very off in the Southern US
- All models predicted Texas and Florida as Democrat, and Michigan and Wisconsin as Republican, when the opposite was true
  - Note that those were all considered competitive swing states in the 2020 election
- Midwest was very accurate
- Northeast was mostly accurate as well
- All in all the Degree 3 and 4 models did fairly well predicting the 2020 election

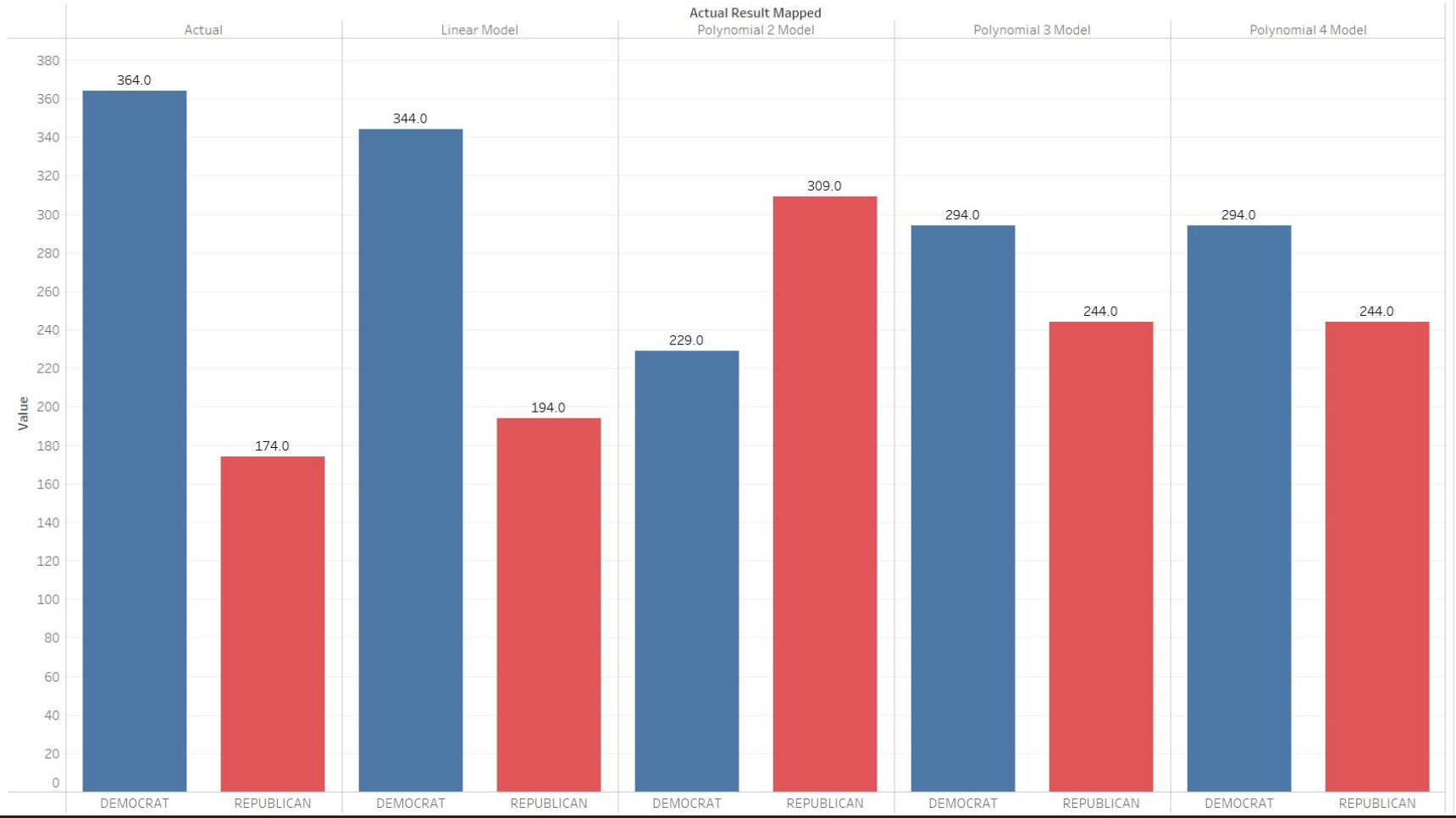
2000 Model Results



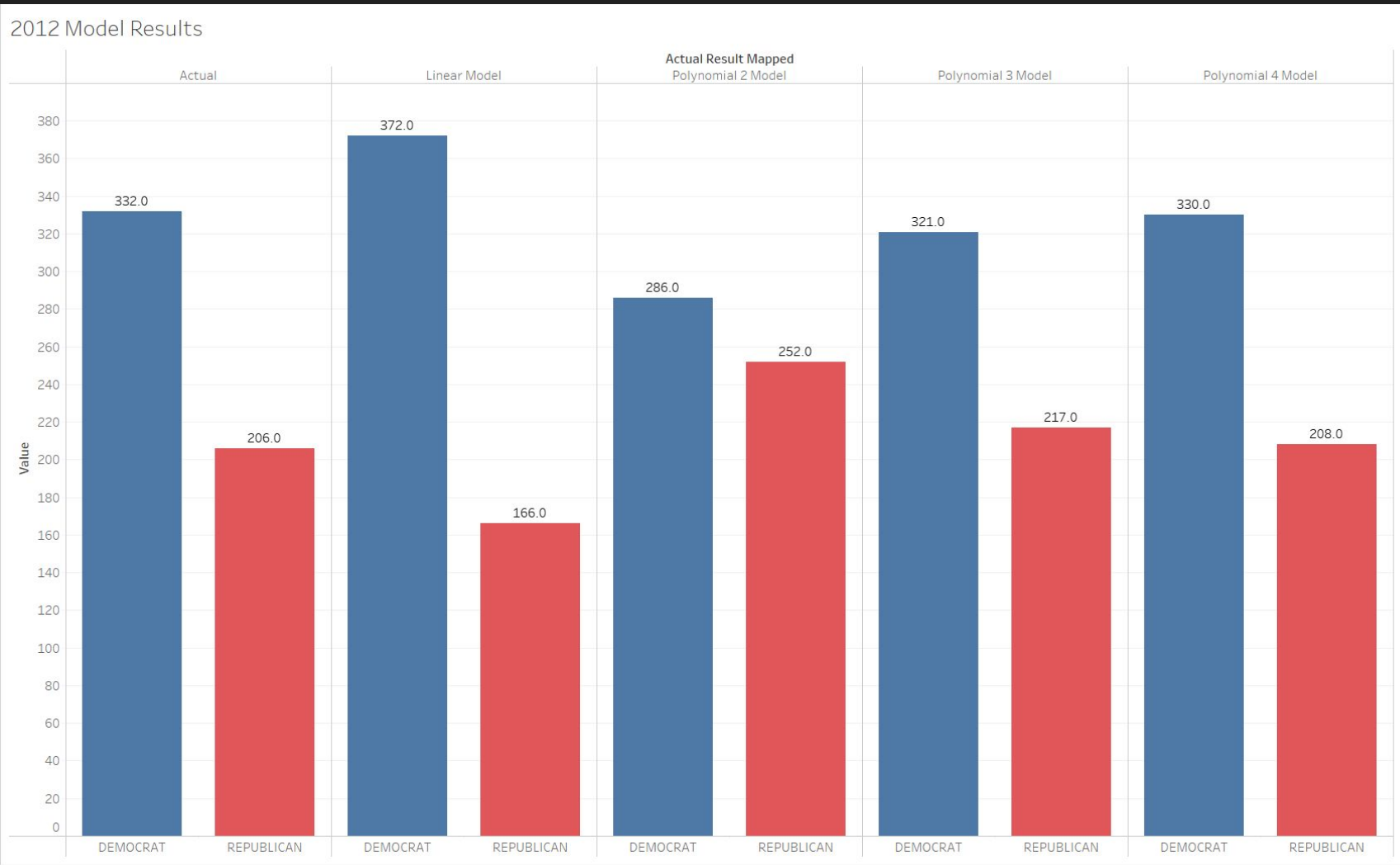
2004 Model Results



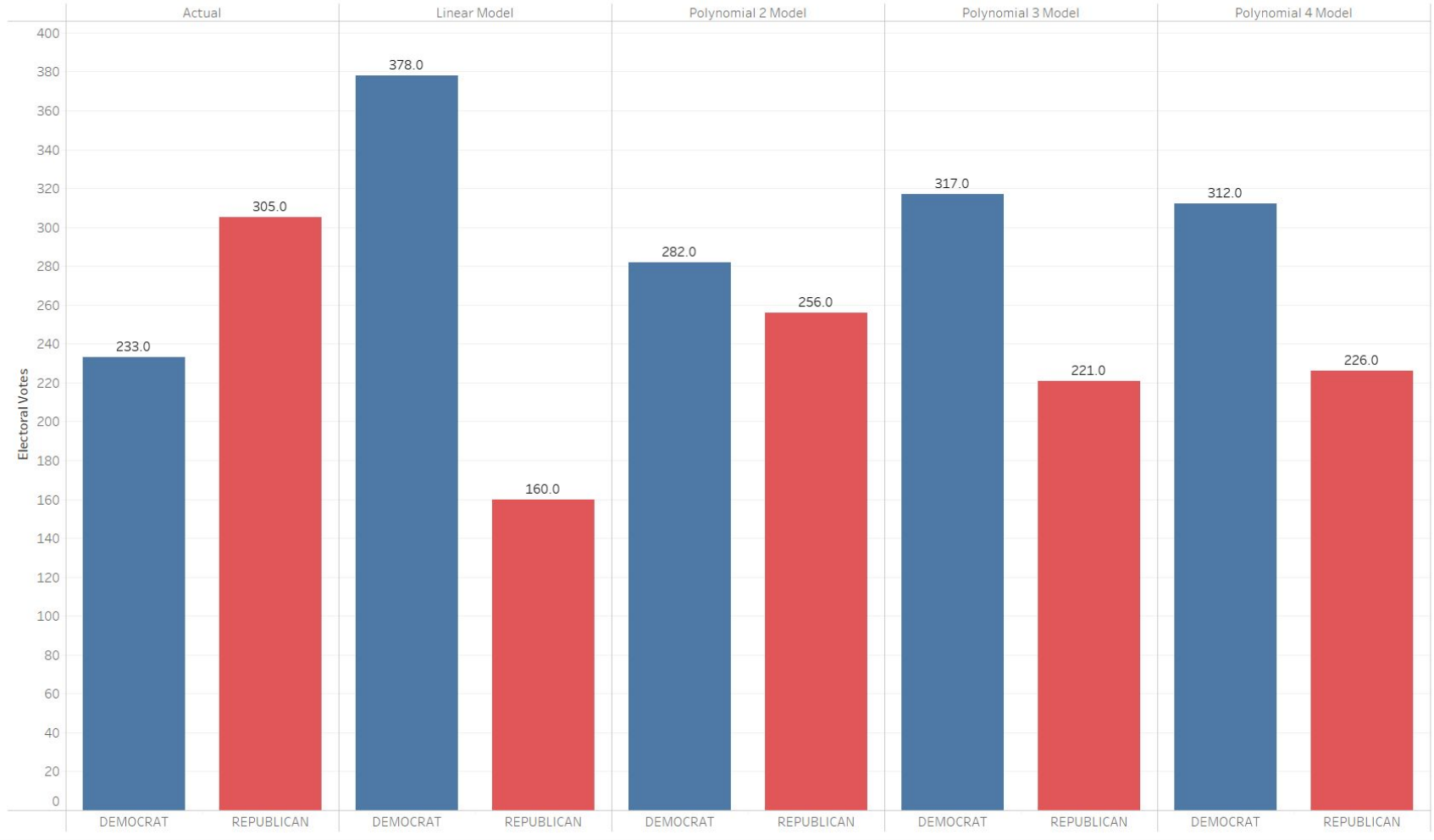
2008 Model Results



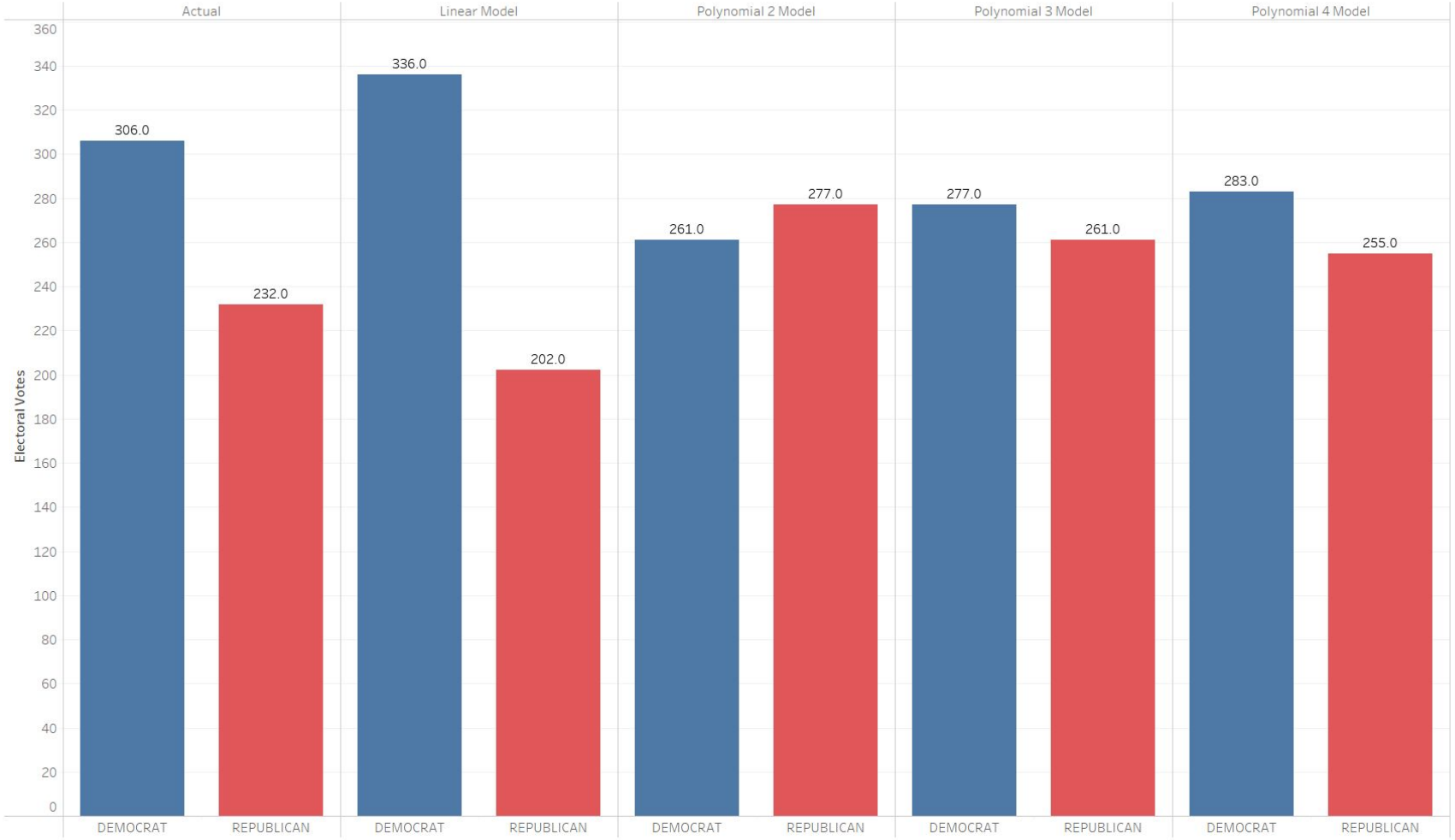
2012 Model Results



2016 Model Results



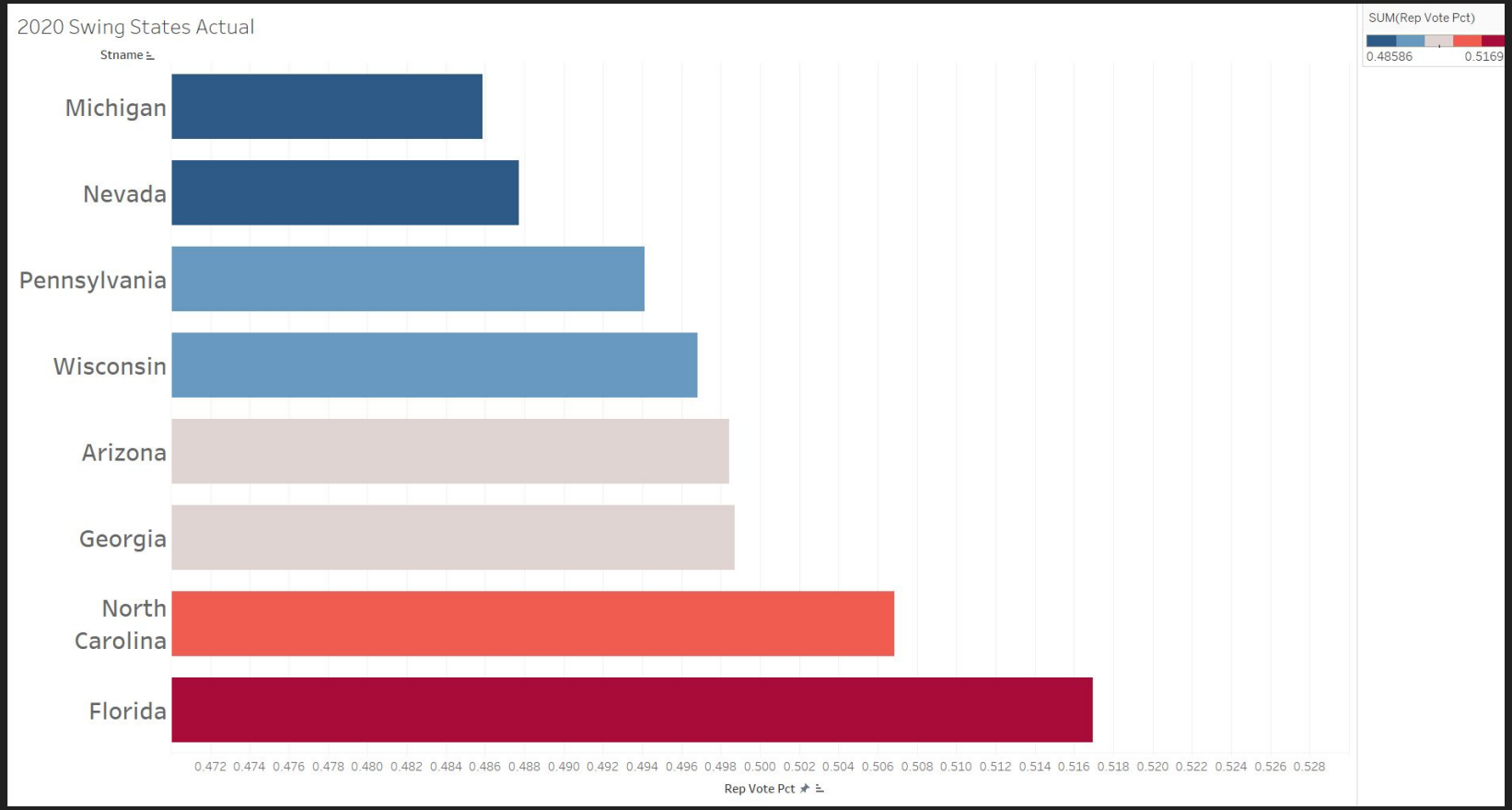
2020 Model Results



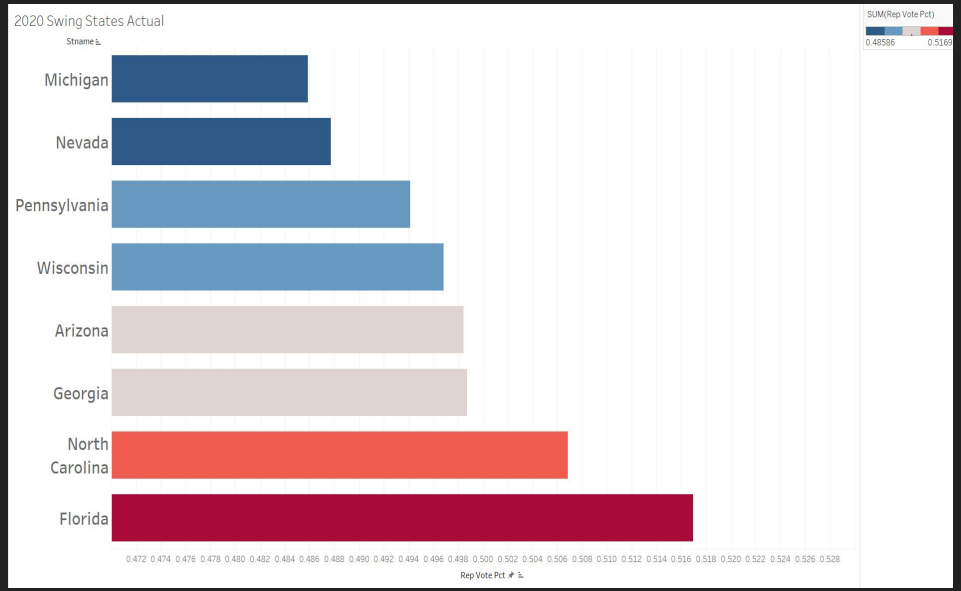
Thoughts on Electoral Vote Predictions

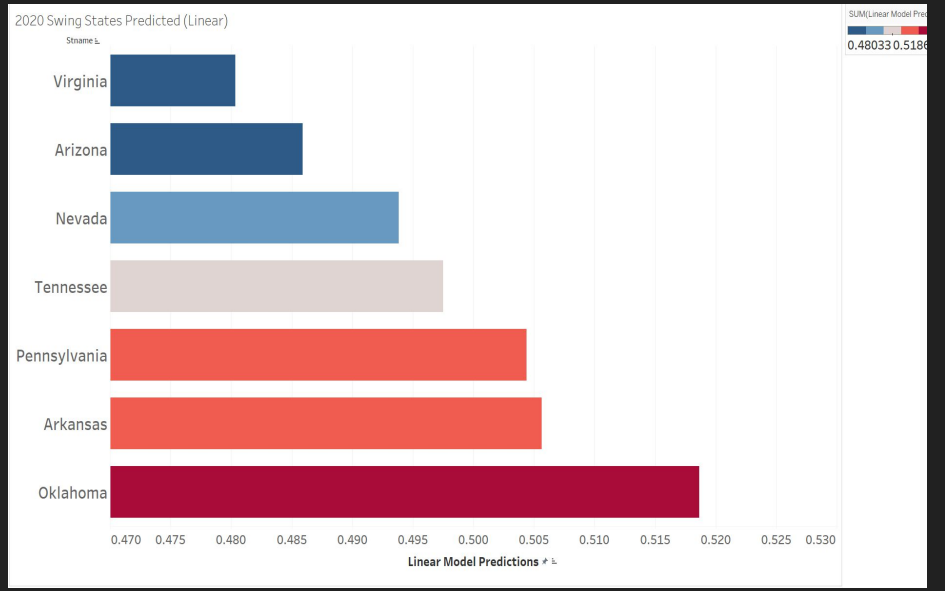
- Again, the linear model didn’t perform nearly as well, predicting Democrat victories every year
- The Degree 3 and Degree 4 electoral vote predictions were most accurate
  - Based off of Winner (only missed one election winner each)
  - Based off of difference in actual and predicted electoral votes won by each side
- 2016 was a very wild year for our models
  - All models missed the Republican winner that year
    - As did most pollsters and political experts
  - It’s possible this threw off our results for other years when using the machine learning

Actual Swing States in 2020 Election

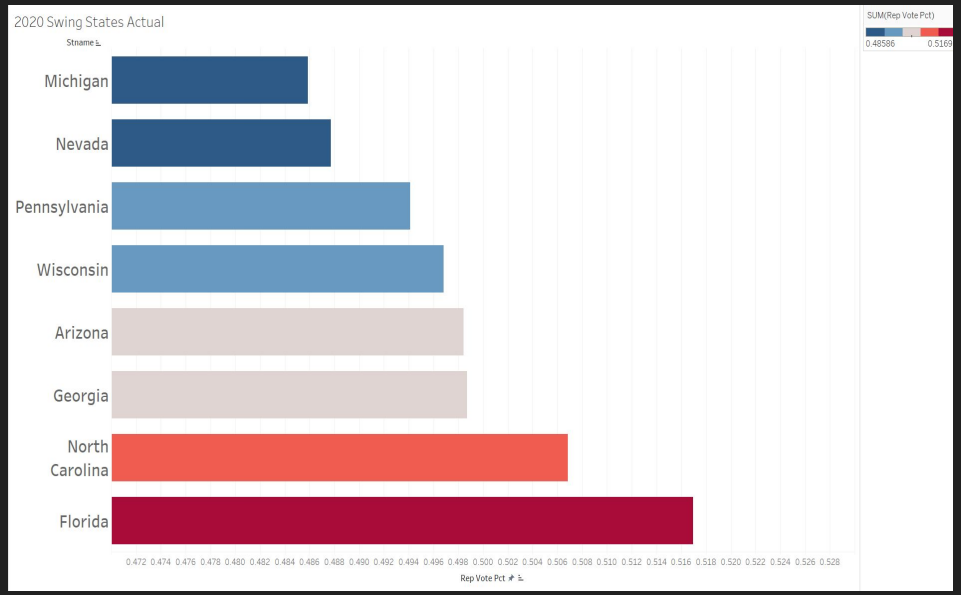


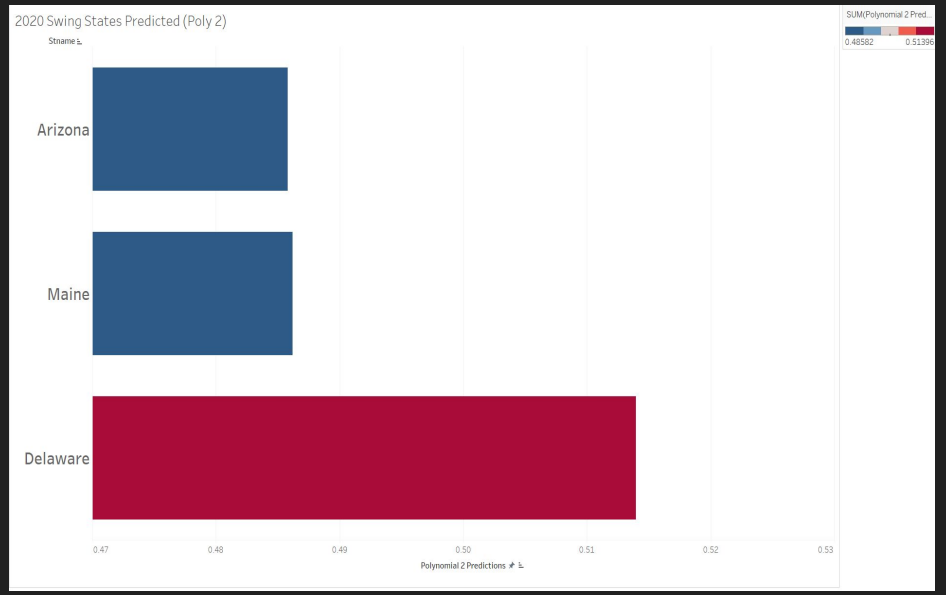
Linear Model



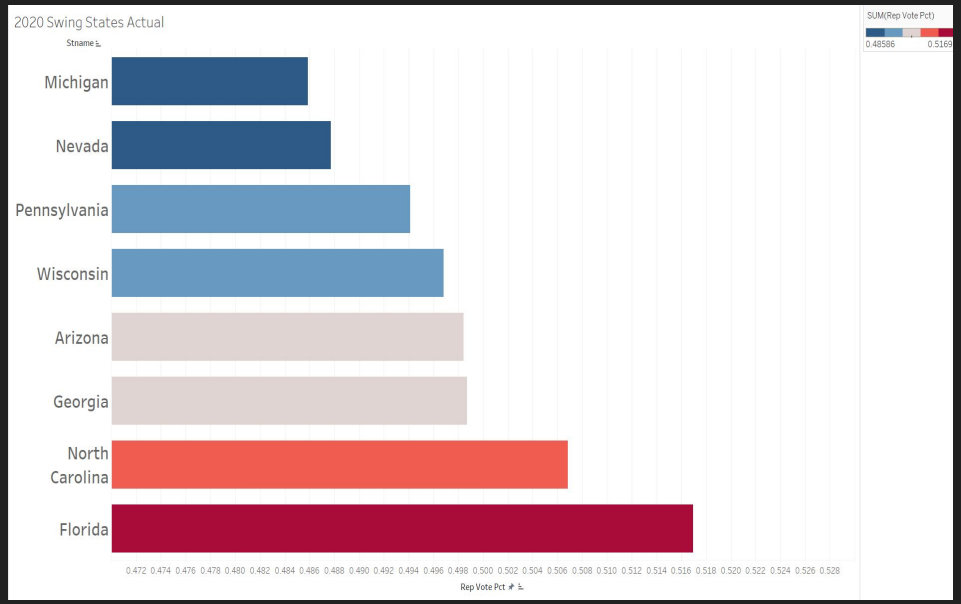


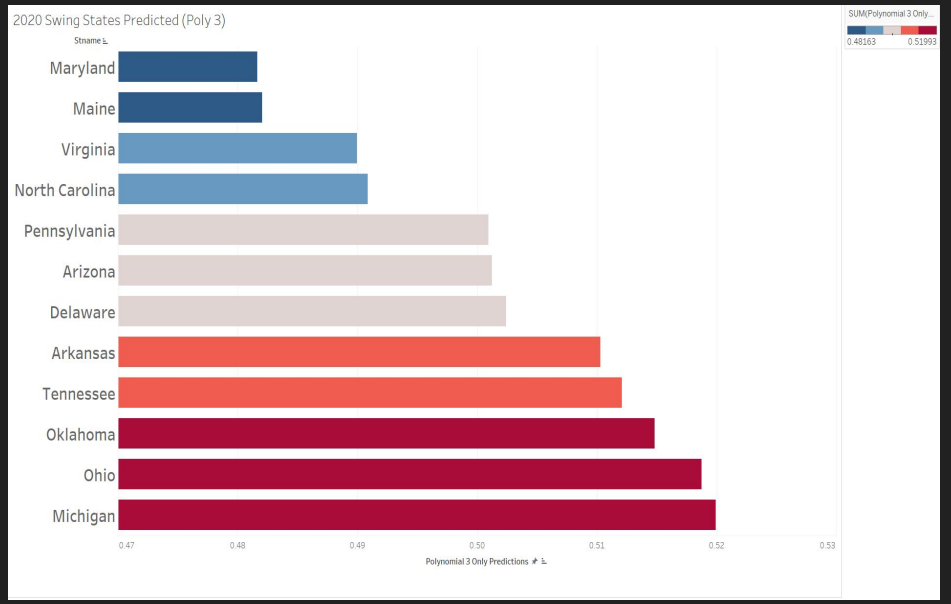
Degree 2 Polynomial Model



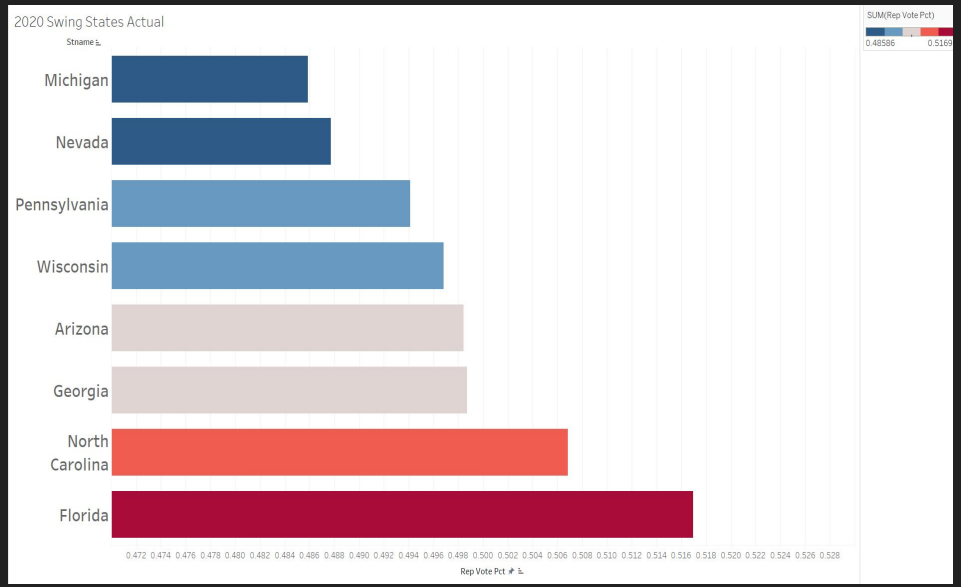


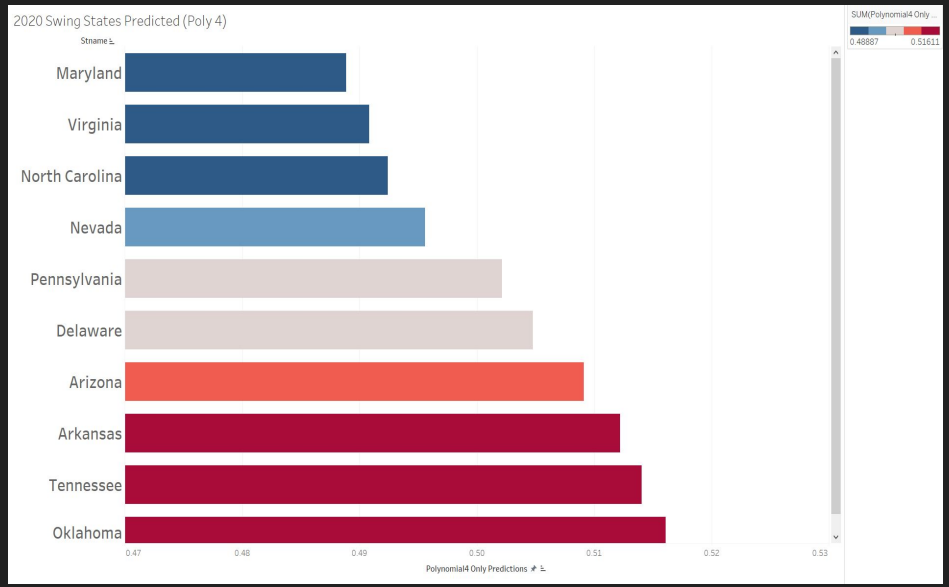
Degree 3 Polynomial Model





Degree 4 Polynomial Model





Thoughts on Swing State Predictions

- Linear regression and Degree 2 is not accurate at predicting swing states
  - Linear: 3
  - Degree 2: 1
- Degree 3 and Degree 4 is more accurate: both predicted 4 out of 8 of the swing states

Future Improvements

- County Data
  - Working up from smaller data can give us more insight into each state’s specific demographics
- More and Different Demographic Data
  - Socioeconomic Data
  - Population Density
  - Education Level
  - Multiple categories for race, age, gender
  - Eligibility to Vote
  - Religious affiliations
- Use of non-Demographic Data
  - People don’t vote along strict demographic lines
  - Options include: polling data, prior vote pattern by state
- Different Machine Learning Techniques
  - Deep Learning: neural networks

Q & A