Impact of Gerrymandering on 2018 Midterm Elections
Individual Report

Yashas Manjunatha
Department of Mathematics
Duke University
Durham, NC 27708
yashas.manjunatha@duke.edu

May 4, 2019

Abstract

Partisan gerrymandering is a peril to democratic elections in the United States. Given the lack of comprehensive, quantitative analysis of gerrymandering on the national level, this research aims to develop a model that assesses the impact of gerrymandering in every state and aggregates the results to determine the net effects of gerrymandering on the 2018 midterm elections. To this end, I simulated election outcomes for each state on an ensemble of congressional redistricting plans. Additionally, I developed methods to impute congressional election outcomes for uncontested races and model election results using population demographic data. Finally, I was able quantitatively identify states with potential gerrymandering in the 2018 midterm elections and evaluate the net impact of gerrymandering on the national level.

Keywords Gerrymandering · Democracy · 2018 Midterm Elections

1 Introduction

Partisan gerrymandering is a practice intended to establish a political advantage for a particular party by manipulating congressional district boundaries. Two primary tactics used to this end are cracking and packing. Cracking involves spreading a large voting bloc, a group of voters that tend to vote for a particular party, across multiple districts to prevent them from getting sufficiently large voting power in any one district. Packing aims to concentrate a voting bloc into a single electoral district, which usually has a disproportionately large margin of victory, to reduce their influence in other districts. The word "gerrymander" first appeared in 1812 to describe a salamander shaped district in Massachusetts designed to benefit the Democratic-Republican Party, and political gerrymandering has continued into the 21st century as one of the major abuses of democratic elections in the United States. [1] While endless present-day examples of partisan gerrymandering have been reported, there is a lack of comprehensive, qualitative analysis of gerrymandering on the national level. [2]

My objective is to develop a model that assesses the impact of gerrymandering in every state and aggregates the results to determine the net effects of gerrymandering on the 2018 midterm elections. To this end, I simulated election outcomes for each state on congressional maps from the 538 Atlas of Redistricting, which provided baseline gerrymandered and neutral redistricting maps. [3] Additionally, I developed methods to impute congressional election outcomes for uncontested races and model election outcomes using population demographic data.

2 Overview of Methods

The research focuses on simulating elections on an ensemble of hypothetical redistricting maps from the 538 Atlas of Redistricting: a Republican gerrymander, a Democratic gerrymander, a Proportional map that draws districts to promote proportionally partisan representation, and a Compact map which aims to draw compact districts while trying to respect county borders. [3] The latter two maps serve as a neutral baseline, while the former provide the gerrymandered baseline. Nationally, the lowest granularity of historic election voting data widely available is at the county level, which I was able to obtain from the New York Times' online election coverage (election results provided by the Associated Press). [4] However, in congressional districts with uncontested races, election data is unavailable at the county level, making it impossible to simulate elections for any district drawn over such a county. Thus, I developed a method to impute congressional election data in these counties with uncontested races using Presidential election data as a guideline. Additionally, the redistricting plans often split counties into multiple districts, making it impossible to simulate elections with only county level data. In order to reduce the granularity of the data, I developed a linear regression model to predict election outcomes at a census tract level using Census population demographic data. This is aggregated to calculate county split values, which indicate how a split county's votes should be distributed into the different districts. This is used to simulate electoral outcomes in every congressional district nationwide for the ensemble of redistricting plans.

3 Imputing House Election Data for Missing Values and Uncontested Races

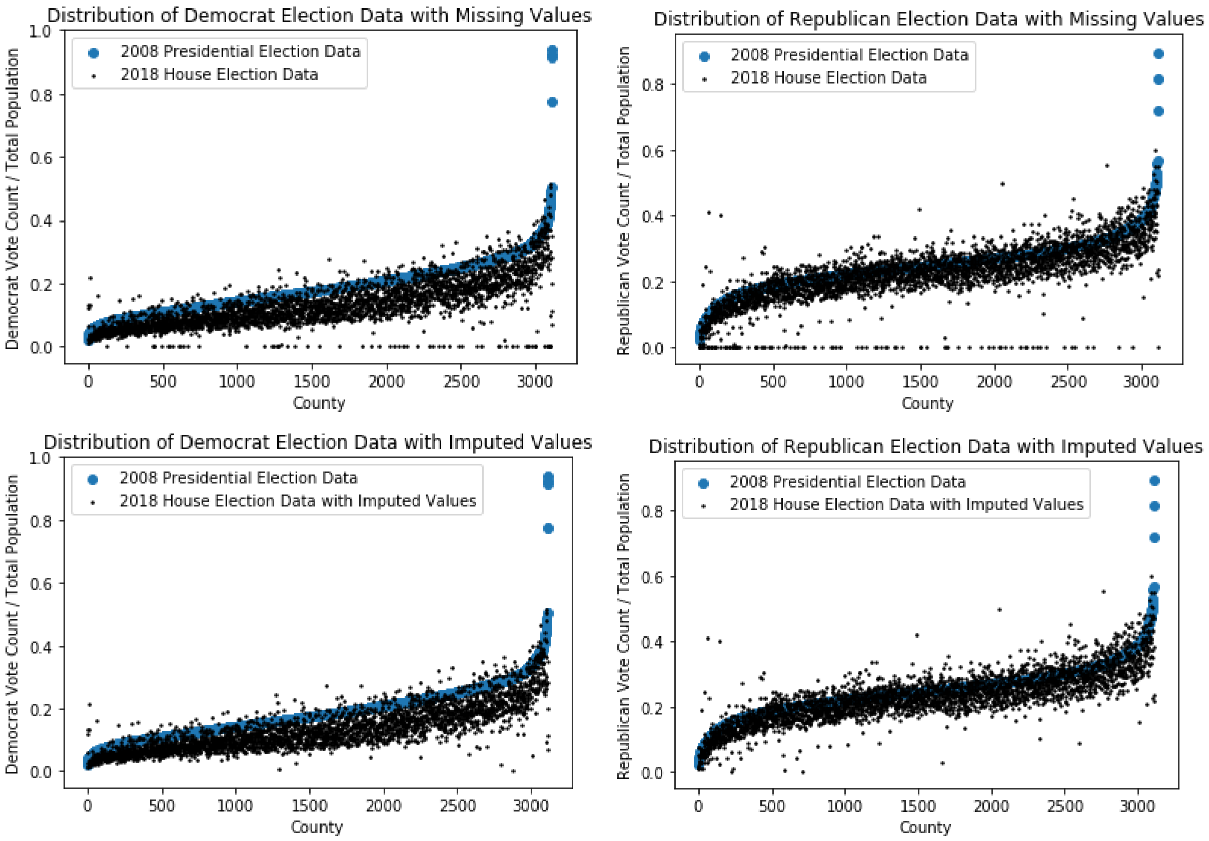


Figure 1: The top two plots show the dataset with missing values (a number of counties have a vote count of 0). The bottom two plots show the resulting dataset after the imputing process.

I web scraped county level congressional results for the 2018 midterm election from the New York Times (provided by the Associated Press). [4] However, I was unable to extract county level data for 12 out of the 435 congressional districts. Additionally, in 2018, there were a total of 41 districts that had an uncontested race (3 districts did not have a Democratic candidate and 38 districts did not have a Republican candidate). [5] Thus, the harvested dataset was missing Democrat congressional results for 61 counties and Republican congressional results for 96 counties. Since the 2008 Presidential election had a similar blue wave signature to the 2018 House midterm elections and a Presidential election grantees a competitive race in every county, I used 2008 Presidential election data as a guideline to impute missing values for the 2018 House midterm elections. Additionally, vote fraction (vote count / total population) was used as the the target of the imputing procedure to normalize the data. Missing county level congressional election data for party p was imputed with the following procedure:

1. Order all counties from lowest to highest based on the county's voting data for party p in the 2008 Presidential race over the county's total population,

$$\frac{C(P_p)}{C(TotalPopulation)}$$

   where $C(P_p)$ is the number of votes for party p's Presidential candidate in county C and $C(TotalPopulation)$ is the total population in county c according to 2010 Census data.

2. Process counties $\{C_1, C_2, C_3, ...\}$ in the order from step 1, and use a moving average of the 2018 house election data for party p for the counties to impute values for counties missing house election data for party p, $\forall \frac{C_i(H_p)}{C_i(TotalPopulation)} = 0$, where $C_i(H_p)$ is the number of votes for party p in the 2018 House election in county i,

$$\frac{C_i(H_p)}{C_i(TotalPopulation)} = \frac{\sum_x^{WindowSize} \frac{C_{i-x}(H_p)}{C_{i-x}(TotalPopulation)} + \frac{C_{i+x}(H_p)}{C_{i+x}(TotalPopulation)}}{2 \times WindowSize}$$

   .

3. And finally, calculate

$$C_i(H_p) = \frac{\sum_x^{WindowSize} \frac{C_{i-x}(H_p)}{C_{i-x}(TotalPopulation)} + \frac{C_{i+x}(H_p)}{C_{i+x}(TotalPopulation)}}{2 \times WindowSize} \times C_i(TotalPopulation)$$

This procedure was implemented with a window size of 4 (i.e. calculating a moving average of the 4 counties on either side, totalling to the 8 surrounding counties), and the results are provided in Figure 1. A drawback of the current implementation is that counties in the widow for the moving average with missing data are not ignored in the moving average calculation, and thus underestimates the value when imputing in this case. A future iteration of the procedure could expand the window until there are a window size number of counties that have existing house election data. This procedure can also be modified to use an exponential weighted moving average instead of the simple average described above. This process of imputing values for missing data in the 2018 congressional county level election dataset provides estimated data for the counties with missing values, ultimately providing a complete county level dataset to simulate elections.

4 Linear Regression Model

In order to reduce the granularity of the county level data, I developed two linear regression models to predict targets $y = \frac{\text{House Democrat Vote Count}}{\text{Total Population}}$ and $y = \frac{\text{House Republican Vote Count}}{\text{Total Population}}$ at the census tract level after training the models on county level data. I develop two separate models because only total population data was available, and voter turnout data is not widely available at such a granular scale. Thus, I cannot predict the Democrat target by taking $1-$ Republican target (or vice versa) so two separate models are required.

4.1 Training the Linear Regression Model

Table 1: Linear Regression Features

Feature	Description	Correlation with Democrat Target	Correlation with Republican Target	Linear Regression Coefficient for Democrat Target	Linear Regression Coefficient for Republican Target
HD02_S016	Percent; SEX AND AGE Total population - 70 to 74 years	-0.18972481967426633	0.48708864830170095	0.0016	0.0199
HD02_S057	Percent; SEX AND AGE Female population - 25 to 29 years	0.19697943428630005	-0.4923298225276979	0.0127	-0.0142
HD02_S078	Percent; RACE -Total population One Race - White	-0.2374571767585773	0.5411347035681564	0.0008	0.0023
HD02_S079	Percent; RACE - Total population One Race - Black or African American	0.18521744653018982	-0.39150678877192213	0.0016	0.0005
HD02_S081	Percent; RACE - Total population One Race - Asian	0.351820922382357	-0.31564055237901273	0.0114	-0.0021

Features for the linear regression model were drawn from the 2010 Census Profile of General Population and Housing Characteristics. [6] I identified features for the linear regression model by performing a visual inspection of scatter plots of the feature values against the target values and by calculating a correlation matrix of the features against the target. Features with the highest magnitudes of correlation (i.e. have the most linear relationship with the target) were chosen for the linear regression model. The selected features are detailed in Table 1.

Using statsmodels' ordinary least squares (OLS) linear regression model [7], I compute the weights/coefficients for each feature by minimizing $\sum(y_i - \vec{x}_i \times \beta)^2$, where y_i is the target value, $\vec{x}_i$ is the feature vector, β is the coefficient/weight vector, and i is the data point from the county level dataset. Additionally, a k-fold cross validation schema with an evaluation metric of average root mean square error is used to evaluate the models and prevent overfitting. A detailed summary of the final models are provided in Appendix A.

4.2 Assumptions for Multivariate Linear Regression

Next, I checked the assumptions for multivariate linear regression to better understand the performance of the models.

4.2.1 Linear Relationship

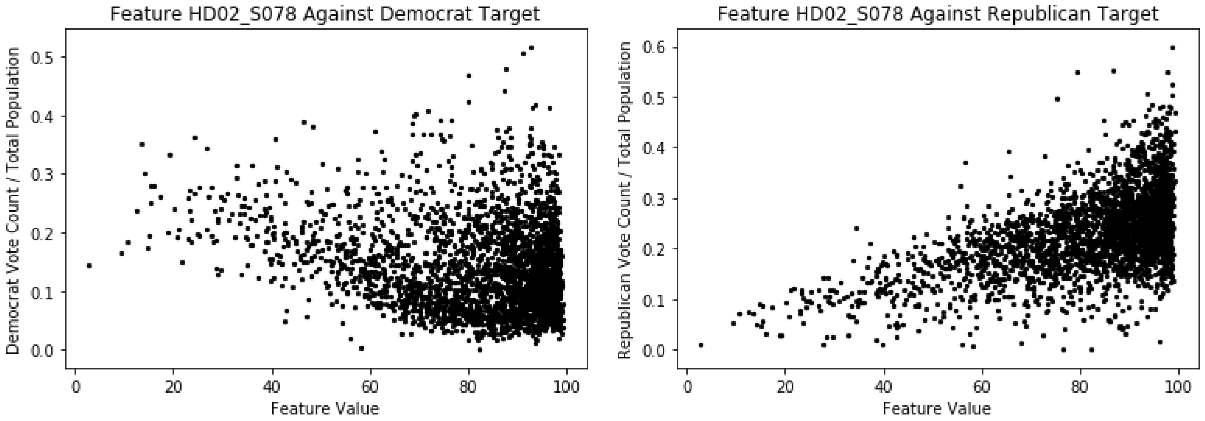


Figure 2: Scatter plot of the feature "Percent; RACE - Total population - One Race - White" against target values to demonstrate linear relationship.

First, I ensure a linear relationship between the features of the model and target values by generating scatter plots for each feature. Figure 2 shows the scatter plot for a feature with high correlation with the target values. The remaining scatter plots are provided in Appendix B. A general linear trend is observed in the scatter plots, confirming the linear approach to modeling the target values.

4.2.2 Independent Observations and Multicollinearity

Table 2: Correlation Matrix with Correlation Coefficients Between Each Feature

Feature	HD02_S016	HD02_S057	HD02_S078	HD02_S079	HD02_S081
HD02_S016	1	-0.67660641	0.32818484	-0.19659292	-0.31894467
HD02_S057	-0.67660641	1	-0.3997078	0.27169574	0.36815857
HD02_S078	0.32818484	-0.3997078	1	-0.82765461	-0.26496303
HD02_S079	-0.19659292	0.27169574	-0.82765461	1	0.03029221
HD02_S081	-0.31894467	0.36815857	-0.26496303	0.03029221	1

Further details about these features can be found in Table 1

Next, I consider the independence of observations in the dataset and multicollinearity of features. Since each county represents a single observation in the dataset, and thus each census tract is only represented once, I analytically confirm the independence of observations, ensuring that variance of each observation independently affects the overall analysis. Additionally, I generate a correlation matrix, Table 2, with correlation coefficients between each feature to understand the multicollinearity of the features, when one predictor variable can be linearly predicted from the others. I observe a high correlation between the "White Population Percentage" feature and the "Black or African American Population Percentage" feature, which can be empirically explained by inequality and the history of their population diffusion. Additionally, note that this multicollinearity is inevitable due to the features from the 2010 Census dataset limited to sex, age, and race characteristics. While multicollinearity does not reduce the predictive power or reliability of the model as a whole, it does indicate that the coefficients of regression may change erratically in response to small changes in the model or the data, and thus cannot be used to establish variable importance.

4.2.3 Normal Distribution of Residuals

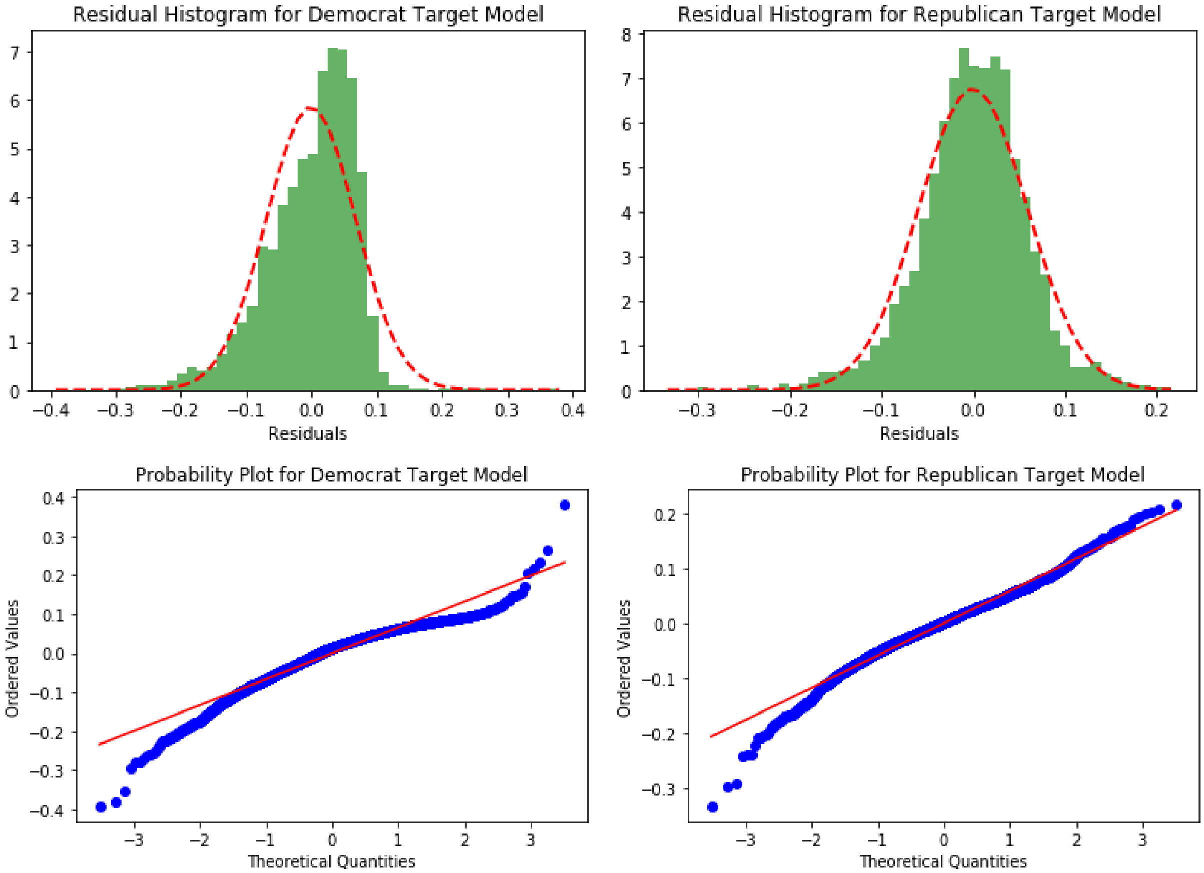


Figure 3: Plots to Visualize Distribution of Residuals: Left are for Democrat Target Model and Right are for Republican Target Model

The third condition of a multivariate linear regression model is that the error terms, or residuals, are normally distributed. To confirm this condition I generate two plots for each model, Figure 3. First, I plot the histogram of residuals with a superimposed normal curve. The model to predict the Democrat target demonstrates a skewed distribution suggesting the error terms are not normally distributed, whereas the model to predict the Republican target shows that the residuals generally follow a normal distribution. Then, I plot the theoretical percentiles of the normal distribution versus the observed sample percentiles, which should be approximately linear. This is somewhat evident in the normal probability plot for the Democrat target model, while the normal probability plot of the residuals for the Republican target model is approximately linear supporting the condition that the error terms are normally distributed.

4.2.4 Equal Distribution of Variance - Homoscedasticity

The last condition of multivariate linear regression analysis that I consider is homoscedasticity, meaning that the error term is the same across all values of the independent variables. To test this condition I plot the predicted target values against the residuals and visually inspect whether the distribution is uniform over the x-axis. While the Republican target model exhibits a relatively more uniform behavior, the Democrat target model shows an abnormal-shaped pattern of heteroscedasticity, when the size of the error term differs across values of an independent variable. OLS regression seeks to minimize residuals in order to produce the smallest possible standard errors, and by definition OLS gives equal weight to each observation, but when heteroscedasticity is observed the data points with larger disturbances have more "pull" than other observations. A weighted least squares (WLS) linear regression approach may be more

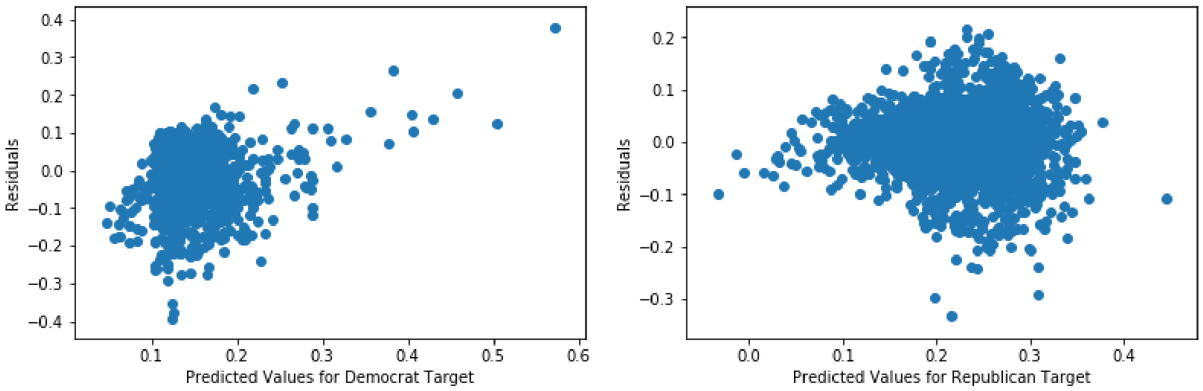


Figure 4: Plots to Test for Equal Distribution of Variance

appropriate to down-weight those data points with larger disturbances. Additionally, the bias in standard errors when heteroscedasticity is observed can lead to incorrect conclusions about the significance of the regression coefficients.

4.3 Error Analysis of Models

Table 3: Evaluation Mertics for Linear Regression Models

Metric	Democrat Target Model	Republican Target Model
Avg. RMSE	0.068572	0.059248
R-Squared Value	0.809	0.942
Standard Error of Regression	0.068289	0.059199

To evaluate the performance of the linear regression models I use a 5-fold cross validation schema with an evaluation metric of average root mean square error. Furthermore, I calculate the R-squared value to determine how well the model fit to the data. Finally, I calculate the standard error of the regression, the typical distance that the data points fall from the regression line, which provides a measure for prediction performance. The evaluation metrics for the linear regression models are provided in Table 3. The linear regression model for the Republican target values demonstrates a better performance (than the Democrat target model) with a lower error, higher fit, and better predictive performance. Additionally, since the Democrat target model showed signs of violating the assumptions for multivariate linear regression, further work must be done to improve the performance of the model, including exploring additional/different feature variables, using a variance stabilizing transformation (like a logarithmic transform) on the features, trying a WLS linear regression approach, and attempting non-linear regression methods.

4.4 Predictions from Linear Regression Model

I use the linear regression model trained on county level data to predict Republican target values, $y = \frac{\text{House Republican Vote Count}}{\text{Total Population}}$, and Democrat target values, $y = \frac{\text{House Democrat Vote Count}}{\text{Total Population}}$, on census tract level data using $y_i = \vec{x}_i \times \beta$, where where y_i is the target value, $\vec{x}_i$ is the feature vector, β is the trained coefficient/weight vector from the appropriate model, and i is the data point from the census tract level data. Furthermore, I calculate the prediction interval for the predicted target value using $y_i \pm (t_{(\alpha/2,n-p)} \times \sqrt{MSE + [\text{se}(y_i)]^2})$, where $t_{(\alpha/2,n-p)}$ is the t-multiplier and $\sqrt{MSE + [\text{se}(y_i)]^2}$ is the standard error of regression. Note that the standard error of regression term is calculated in Table 3, and thus provides a powerful metric to evaluate the predictive performance of the model. Additionally, it is important to note that the prediction interval formula strongly depends on the condition that the error terms are normally distributed.

4.5 Calculating County Split Values

Next, I aggregate the predicted values at the census tract level to calculate county split values to determine how a split county's votes should be distributed into the different districts:

$$\text{CountySplitValue}(p)_{C\cap D} = \frac{\sum_{T\in C\cap D}\left(\frac{T_p}{T_{totalpopulation}} \times T_{totalpopulation}\right)}{\sum_{T\in C}\left(\frac{T_p}{T_{totalpopulation}} \times T_{totalpopulation}\right)} = \frac{\sum_{T\in C\cap D} T_p}{\sum_{T\in C} T_p}$$

$$\text{with sanity check that } \sum_{T\in C} T_p = C_p$$

where D is district, C is county, T is census tract, p is party, T_p is house vote count for p in T, $T_{totalpopulation}$ is the total population in T according to 2010 Census data, C_p is house vote count for p in C. Note that $\frac{T_p}{T_{totalpopulation}}$ is the target value predicted from the linear regression model.

5 Simulating Elections on Redistricting Maps

Finally, I simulate electoral outcomes for each congressional district on the 538 Atlas of Redistricting ensemble of congressional plans. Results for the district are calculated as follows:

$$\text{House Result}(p)_D = \sum_{C\in D} C_p \times \text{CountySplitValue}(p)_{C\cap D}$$

where D is district, C is county, p is party, and C_p is house vote count for p in C.

Table 4: National Level Breakdown of Simulation Results

Redistricting Map	Democrat Districts	Republican Districts
2018 Midterm Results*	235	199
Current	234	201
Compact	237	198
Proportional	250	185
Democrat Gerrymander	265	170
Republican Gerrymander	213	222

*NC District 9 is Currently Undecided

Table 5: Error Bounds on Simulation Results

Redistricting Map	Democrat Districts 95% Interval	Republican Districts 95% Interval
Current	[221, 243]	[192, 214]
Compact	[226, 242]	[193, 209]
Proportional	[236, 254]	[181, 199]
Democrat Gerrymander	[262, 265]	[170, 173]
Republican Gerrymander	[180, 227]	[208, 255]

National level results from the simulation are summarized in Table 4 and Figure 5. Additionally, Table 5 provides national level error bounds on the simulation results by using the prediction interval and aggregating up the error from the census tract level prediction to give a 95% confidence interval for the national congressional results. While results from Table 4 show that the national congressional outcomes for the current map is closer to the outcomes for the compact map, one of our neutral redistricting baselines, than the results for any other map, a closer look at Figure 5 shows that the state fraction curve for the current plan (blue) follows a more similar curve to the Republican gerrymandered map (red), providing a reminder that a state level analysis aggregated up to the national level is required to accurately assess the net impact of gerrymandering on the national level.

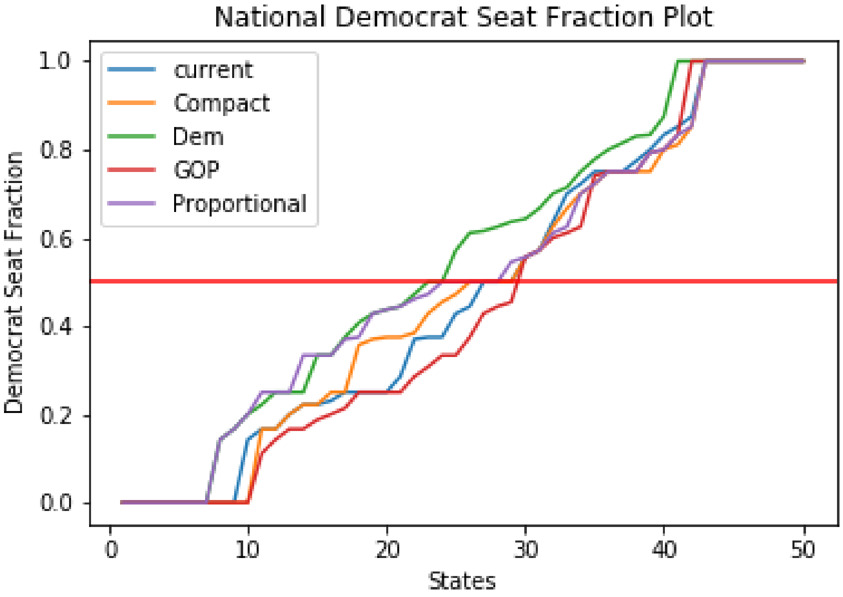


Figure 5: National plot of Democrat seat fraction in each state shows current map has a similar curve to a Republican gerrymandered map.

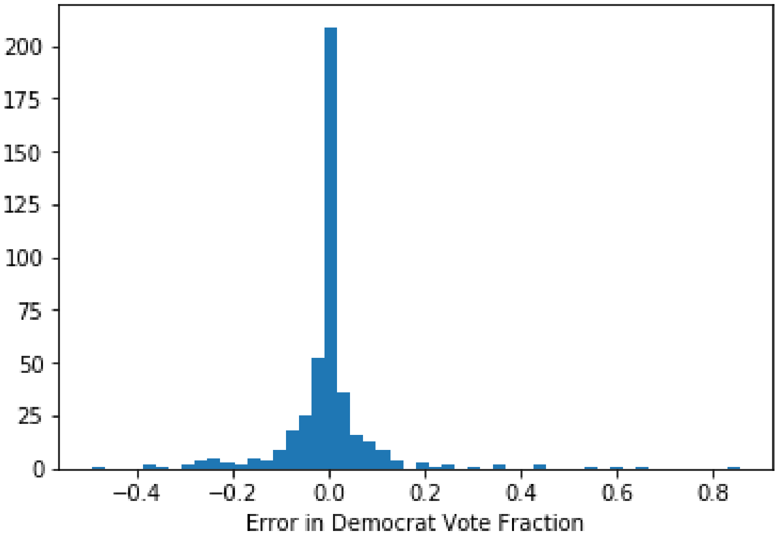


Figure 6: District level error analysis of simulated results on current map against actual election outcomes in the district.

An initial error analysis of the simulation was conducted by comparing the predicted district outcomes from the simulation on the current map to the actual district outcomes in the 2018 midterm congressional election. The simulation mispredicted the outcome in 30 districts (out of 435 total districts) with an average square error of 3.29% in the Democrat vote fraction for the district. 8 of these mispredicted districts were in Pennsylvania, and here I make an important note that the current map used by the simulation for Pennsylvania is the outdated map before the Supreme Court mandated redistricting used in the 2018 midterm elections, thus any analysis for Pennsylvania is untenable with the current results. [8] Furthermore, Figure 6 shows a district level error analysis of the simulation results with a majority of the districts showing relatively low error, with an overall average square error of 1.2% in the Democrat vote fraction for the district.

6 Results and Analysis

First I formalize the idea of similarity of two vote (or seat) fraction curves with a quantitative metric to measure the distance between redistricting maps, the Gerrymandering Index. [9] To calculate this metric the districts are ordered from lowest to highest based on the Democrat vote fraction in the district and the differences of this value is taken for each order statistic district between any two redistricting maps. Then the l^2 norm is taken on this set of differences, which results in the Gerrymandering Index.

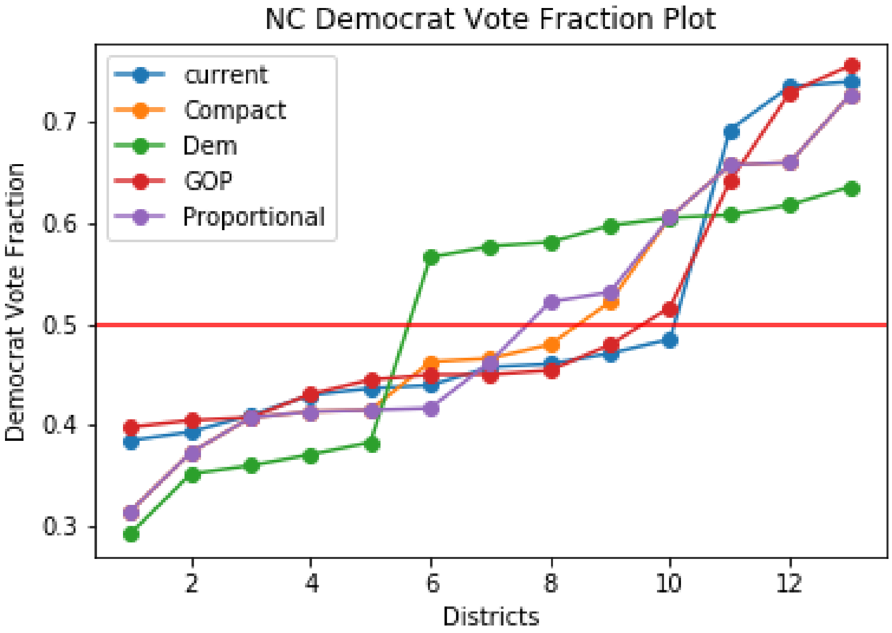


Figure 7: This Democrat vote fraction plot for North Carolina shows evidence of gerrymandering.

Figure 7 shows the state level results of the election simulation in North Carolina. For each district in the state the Democrat vote fraction, or percentage of voters in the district who voted Democrat, is calculated and districts are ordered based on this value. This figure shows evidence of Republican gerrymandering, as the vote fraction curve for the current map (blue) is very similar to the Republican gerrymandered map (red). Furthermore, evidence of packing and cracking techniques is present as a significant jump in the Democrat vote fraction is observed between the 10th and 11th districts, and the three Democrat districts present all have abnormally high Democrat vote fractions.

Figure 8 shows a state level analysis of the distance (Gerrymandering Index) from the current map to the closest gerrymandered map. The inverse of the distance to the closest gerrymandered map is calculated and used to appropriately shade the states in the figure, showing the "closeness" of a state's current map to the closest gerrymandered map. However, in the 538 Atlas of Redistricting ensemble of maps, many states have the same gerrymandered map as the current map, making it impossible to generate independent analysis and conclusions on these states with this figure.

Finally, Figure 9 shows a state level analysis of the net number of congressional seat changes from the two neutral baseline maps (compact and proportional) to the current map. This comparison of the current map to neutral baseline maps provides another avenue to observe where states lie in the neutral to gerrymandered spectrum, with states that show a large magnitude of seat change potentially indicative of gerrymandering. Looking at a national scale it is observed that from the Compact map 19 districts were flipped to Republican and 16 districts to Democrat, whereas from the Proportional map 30 districts were flipped to Republican and 14 districts to Democrat. This suggests that the national impact of gerrymandering in the 2018 midterm election benefit Republicans more than Democrats. Additionally, this analysis provides some insight on the use of compactness and proportionality as neutrality measures in redistricting policy, as well as a measure for the national net impact of gerrymandering.

This analysis is by no means perfect and further analysis on voting percentages across districts within states instead of simply looking at the winner of the districts may provide a more granular and accurate analysis. Nevertheless, using the last two figures as a cursory analysis, I can quantitatively identify states with potential gerrymandering in the 2018 midterm elections.

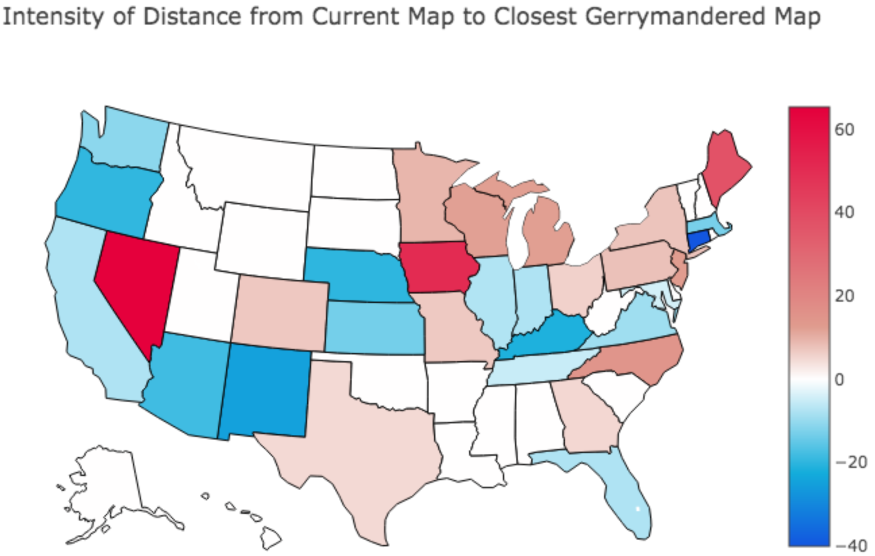

Intensity of Distance from Current Map to Closest Gerrymandered Map
60
40
20
0
−20
−40

Figure 8: Red states indicate the distance between the current map to the Republican gerrymandered map, while blue states indicate the same to the Democrat gerrymandered map. The intensity of the color represents the closeness of the current map to the gerrymandered map. States that are colored white have current maps that are the same as gerrymandered maps in the 538 Atlas of Redistricting ensemble.

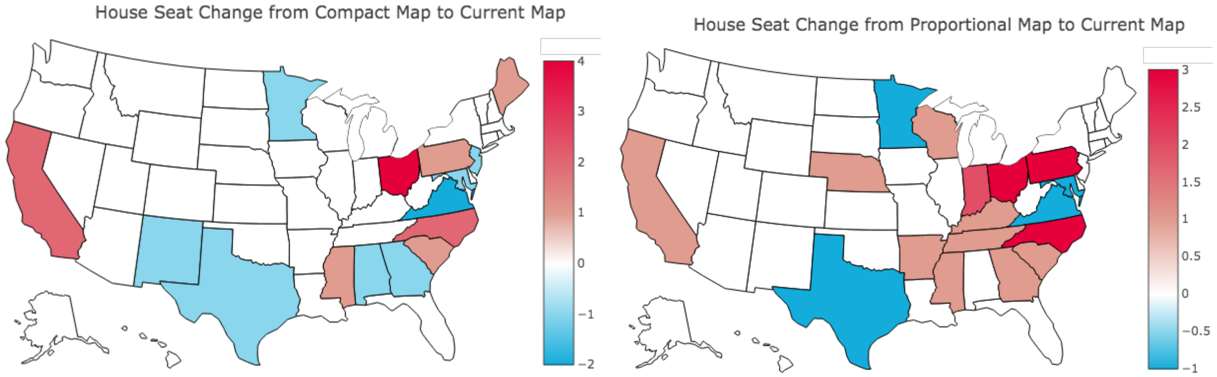

House Seat Change from Compact Map to Current Map
4
3
2
1
0
−1
−2
House Seat Change from Proportional Map to Current Map
3
2.5
2
1.5
1
0.5
0
−0.5
−1

Figure 9: Net congressional seat change from neutral maps to current map. Red indicates that the state gained more Republican seats in the current map compared to the neutral map. Blue indicates that the state gained more Democrat seats in the current map compared to the neutral map.

7 Conclusion

I aim to briefly connect my work with a comprehensive policy analysis performed by Gillian Samios and Isaac Nicchitta as part of the larger Impact of Gerrymandering research effort. States that were identified by the policy analysis as suspected gerrymanders were also quantitatively highlighted by this research (including Arkansas, Indiana, Iowa, Maine, Maryland, Missouri, Nevada, North Carolina, Oregon, Virginia, and Wisconsin). While this analysis is incomplete and limited by the fact that the 538 Atlas of Redistricting ensemble of maps has its own embedded biases by assigning gerrymandered and neutral baseline maps the same as the current map for certain states, which makes analysis of those states currently untenable, the methods used in this research set forth a framework for future analysis. Yet one fact remained quite evident throughout the project, gerrymandering continues to pose a threat to democracy in the United States and comprehensive, quantitative analysis of this phenomenon is crucial to understanding its broader effects.

8 Acknowledgements

I would like to acknowledge my team members on this project Gillian Samios, Isaac Nicchitta, Mitra Kiciman, Nima Mohammadi, and Rayan Tofique. And I would like to thank Gregory Herschlag, John Mattingly, and Frederick Mayer for their guidance with this project. I would also like to thank the Bass Connections program, Duke POLIS, and the Duke Mathematics Department.

References

[1] Martis, Kenneth C. "The Original Gerrymander." *Political Geography* 27, no. 8 (2008): 833-39. doi:10.1016/j.polgeo.2008.09.003.

[2] Lieb, David A. "Election Shows How Gerrymandering Is Difficult to Overcome." AP News. November 17, 2018. https://www.apnews.com/3b4e63717b164dc199d02bd21aa17307.

[3] Bycoffe, Aaron, Ella Koeze, David Wasserman, and Julia Wolfe. "The Atlas Of Redistricting." FiveThirtyEight. January 25, 2018. https://projects.fivethirtyeight.com/redistricting-maps/.

[4] "U.S. House Election Results 2018." The New York Times. November 06, 2018. https://www.nytimes.com/interactive/2018/11/06/us/elections/results-house-elections.html.

[5] "U.S. House Elections without a Democratic or Republican Candidate, 2018." Ballotpedia. https://ballotpedia.org/U.S._House_elections_without_a_Democratic_or_Republican_candidate,_2018.

[6] U.S. Census Bureau; 2010 Demographic Profile Data; Profile of General Population and Housing Characteristics: 2010; using American FactFinder; https://www.census.gov/prod/cen2010/doc/dpsf.pdf.

[7] Perktold, Josef, Skipper Seabold, and Jonathan Taylor. "Statsmodels.regression.linear_model.OLS." Statsmodels. Accessed May 02, 2019. https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html.

[8] "Redistricting in Pennsylvania." Ballotpedia. https://ballotpedia.org/Redistricting_in_Pennsylvania.

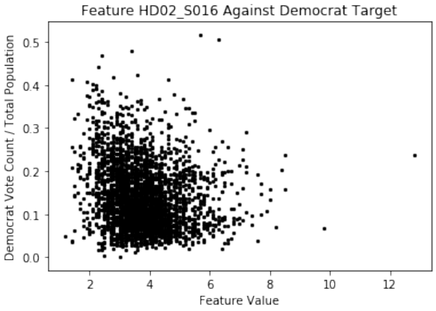
[9] Herschlag, Gregory, Han Sung Kang, Justin Luo, Christy Vaughn Graves, Sachet Bangia, Robert Ravier, and Jonathan C. Mattingly. "Quantifying Gerrymandering in North Carolina." January 10, 2018.

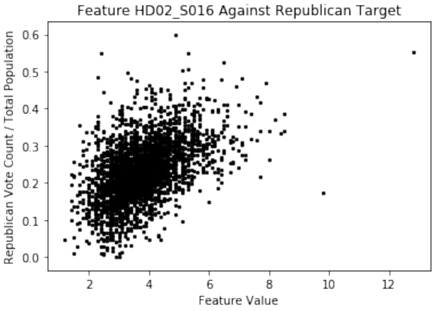
A Linear Regression Model Summaries

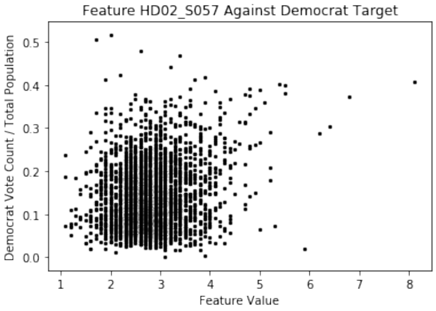
```
OLS Regression Results for Democrat Target
==============================================================================
Dep. Variable:                      y   R-squared:                       0.809
Model:                            OLS   Adj. R-squared:                  0.808
Method:                 Least Squares   F-statistic:                     2625.
Date:                Wed, 01 May 2019   Prob (F-statistic):               0.00
Time:                        12:14:57   Log-Likelihood:                 3939.4
No. Observations:                3112   AIC:                            -7869.
Df Residuals:                    3107   BIC:                            -7839.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
HD02_S016       0.0016      0.001      1.159      0.247      -0.001       0.004
HD02_S057       0.0127      0.002      5.883      0.000       0.008       0.017
HD02_S078       0.0008      0.000      8.152      0.000       0.001       0.001
HD02_S079       0.0016      0.000     12.146      0.000       0.001       0.002
HD02_S081       0.0114      0.001     19.516      0.000       0.010       0.013
==============================================================================
Omnibus:                      474.866   Durbin-Watson:                   1.220
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              937.332
Skew:                           0.932   Prob(JB):                    2.89e-204
Kurtosis:                       4.938   Cond. No.                         158.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

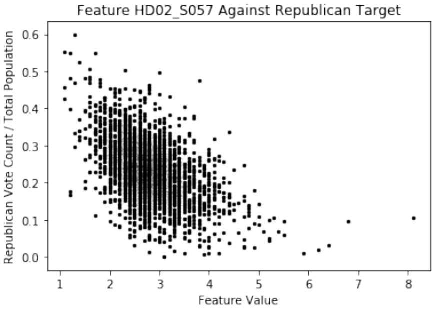
```
OLS Regression Results for Republican Target
==============================================================================
Dep. Variable:                      y   R-squared:                       0.942
Model:                            OLS   Adj. R-squared:                  0.942
Method:                 Least Squares   F-statistic:                 1.003e+04
Date:                Wed, 01 May 2019   Prob (F-statistic):               0.00
Time:                        12:14:57   Log-Likelihood:                 4383.9
No. Observations:                3112   AIC:                            -8758.
Df Residuals:                    3107   BIC:                            -8728.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
HD02_S016       0.0199      0.001     16.343      0.000       0.017       0.022
HD02_S057       -0.0142      0.002     -7.599      0.000      -0.018      -0.011
HD02_S078       0.0023   8.76e-05     26.647      0.000       0.002       0.003
HD02_S079       0.0005      0.000      4.901      0.000       0.000       0.001
HD02_S081       -0.0021      0.001     -4.095      0.000      -0.003      -0.001
==============================================================================
Omnibus:                      180.043   Durbin-Watson:                   1.360
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              406.589
Skew:                           0.365   Prob(JB):                     5.13e-89
Kurtosis:                       4.613   Cond. No.                         158.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

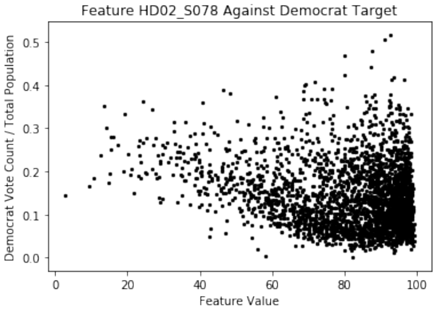
B Scatter Plots of Linear Regression Features

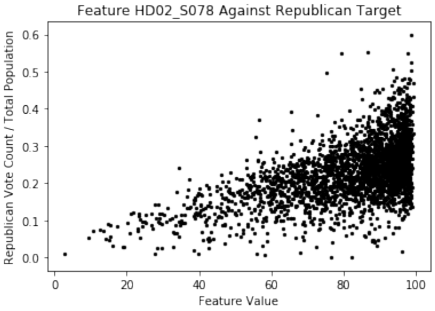

Feature HD02_S016 Against Democrat Target
Democrat Vote Count / Total Population
Feature Value

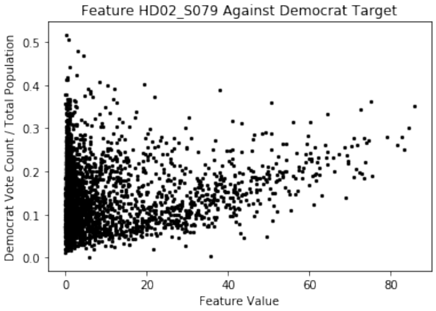

Feature HD02_S016 Against Republican Target
Republican Vote Count / Total Population
Feature Value

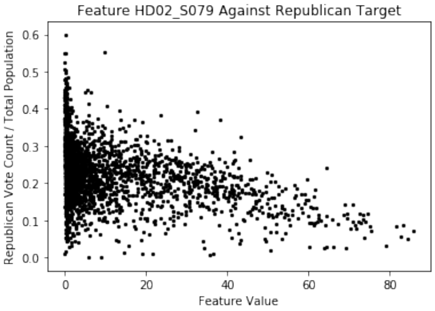

Feature HD02_S057 Against Democrat Target
Democrat Vote Count / Total Population
Feature Value

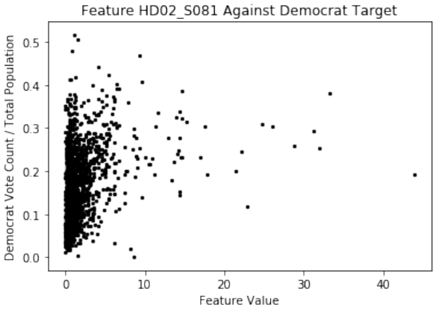

Feature HD02_S057 Against Republican Target
Republican Vote Count / Total Population
Feature Value

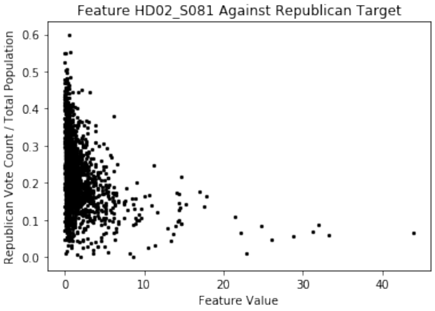

Feature HD02_S078 Against Democrat Target
Democrat Vote Count / Total Population
Feature Value


Feature HD02_S078 Against Republican Target
Republican Vote Count / Total Population
Feature Value


Feature HD02_S079 Against Democrat Target
Democrat Vote Count / Total Population
Feature Value


Feature HD02_S079 Against Republican Target
Republican Vote Count / Total Population
Feature Value


Feature HD02_S081 Against Democrat Target
Democrat Vote Count / Total Population
Feature Value


Feature HD02_S081 Against Republican Target
Republican Vote Count / Total Population
Feature Value