Ben Gelman
Professor Hanke
SML 310
12/10/2021

Texas 2018 Primary Turnout Model

1. Introduction

Political parties, nonprofits, and concerned citizens all have a stake in predicting voter turnout as accurately as possible. Knowing what level of turnout to expect is also important for campaign strategy and allocation of resources as well as for election-night coverage, as news organizations can use predicted turnout levels to know how many more votes they can likely expect at a given point and when it is appropriate to call a race. There are existing models of national or state-level election turnout for general elections, though those focused on primary races are rarer. In this paper, I propose a set of statistical models that can predict turnout for the Texas state-wide primary elections (gubernatorial and senatorial races) in 2018, on the county level. This is accomplished through linear regression and ensemble models (random forest, boosted trees, and extra trees) for both the Democratic and Republican primaries.

My models utilize records of past county-level primary turnout, census statistics related to race, income, and demographics, the partisan lean of a given county, as well as other factors in order to make predictions. My training data includes statistics and election-related information from 2010-2016, with the 2018 primaries acting as my testing data (2020 census data has not been released on a granular enough level to be utilized in this research). Success is measured by the extent to which the models' can explain the county-level variation in turnout in the 2018 gubernatorial and senatorial primaries. For the Republican 2018 primaries, the best results are achieved using extra trees regression, which yields R^2 scores of 72% for the senatorial and 73%

for the gubernatorial primary. Boosted trees are the best models for both the 2018 Democratic primaries, with an R^2 of 86% for the Senate and 82% for the Governor's race.

2. Literature Review

A variety of publicly available models and academic studies have engaged with the question of voter turnout in the past. The most relevant work is a 2020 presidential election turnout model published by Bloomberg News, by Dr. Andrew Therriault. This model used data from the 2016, 2012, 2008, and 2004 presidential elections in order to predict turnout in 2020. Many of the datasets and features I used are based on what Dr. Therriault deemed important for predicting turnout, such as past turnout and racial and economic data on the county level from the American Community Survey. Therriault found that higher levels of past turnout, education, income, and homeownership predict primary turnout well. He also used random forest modeling, as well as some other techniques, to make his predictions. I took many of Dr. Therriault's basic ideas and added in variables and methods that I felt would also be relevant to turnout in a primary, single-state model.

Apart from the Bloomberg model, I have also looked toward scientific studies on the subject of voter turnout. A meta-analysis by Benny Geys, "Explaining voter turnout: A review of aggregate-level research" in *Electoral Studies*, examined 83 studies of electoral turnout in order to discover patterns in what political scientists have found to influence voter turnout and which factors have been found to be better predictors than others. Gey's findings include the fact that past turnout and concurrent elections are statistically significant predictors of present turnout (637).

There is also research on turnout in primaries, which is more pertinent to my specific project and has found results that align with many of Therriault's findings in the Bloomberg model. A 1985 study on presidential primary turnout found that, on the state level, education was a strong predictor of turnout (Kenney, Rice 106-107). Another paper on senatorial primary turnout resulted in similar findings, with education and past turnout serving as significant variables in the model (Kenney 71). A study in 2016 found that the presence of an incumbent also reduces turnout in a given election (Atkeson, Maestas 758).

Finally, the 2014 book *Who Votes Now?* also provides helpful background on what factors scholars have found to forecast higher turnout. Similar to the aforementioned studies, authors Leighley and Nagler find that higher levels of education and income lead to higher turnout (28-29).

Figure 1: Turnout by Education 1972 - 2008

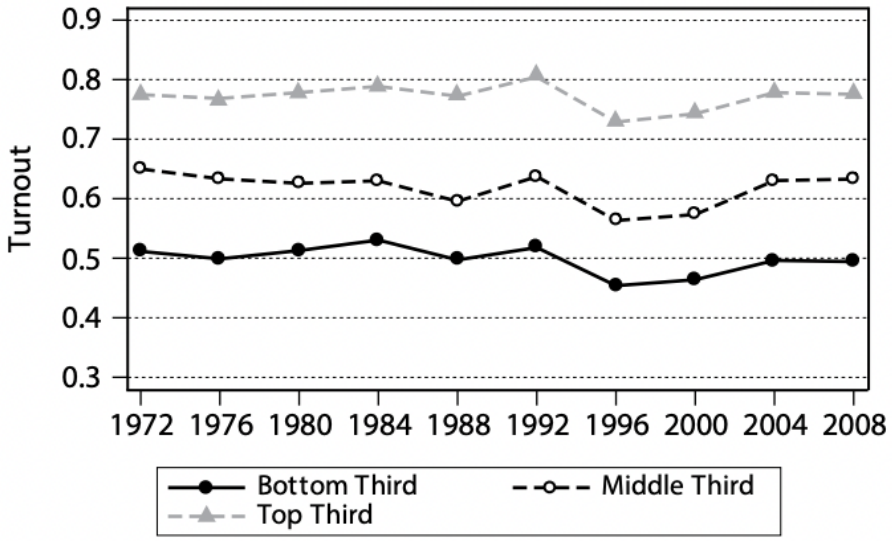
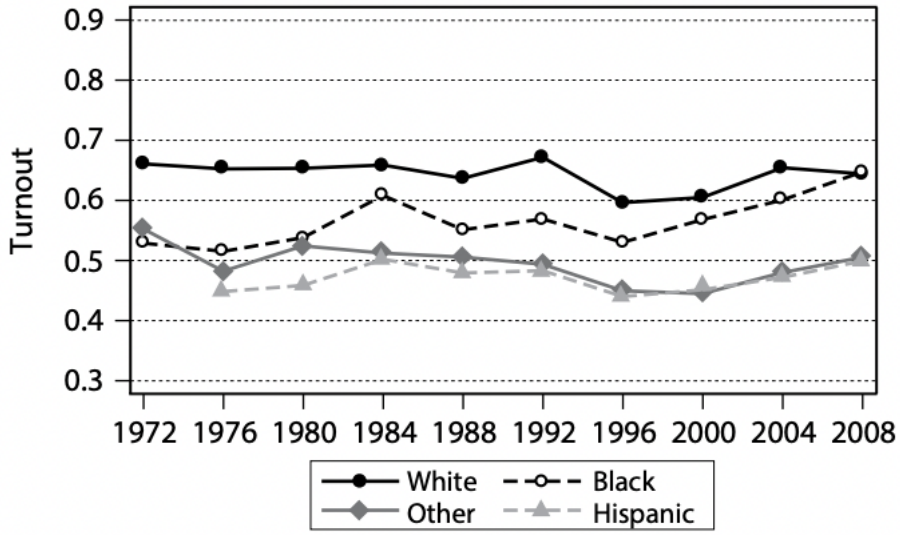


Figure 2: Turnout by Race 1972 - 2008



They also explore turnout by race, finding that while historically White people have a higher turnout than that of Black people, this discrepancy has disappeared in recent years. However, recently Black turnout has increased to match that of Whites (30). Women and married people also tend to vote more than men or singles (32-33). Therriault cites these findings as influential on his model, and I also took inspiration from this book in order to choose which variables to include in my model.

3. Data and Methods

Since 2010, Texas has run several primaries for Senator, Governor, and President, for both the Democratic and Republican parties. With the exception of 2012, these elections were all held in early March. In 2012, the primary was delayed to late May due to a redistricting dispute between the state legislature and several nonprofits (Fernandez).

Data related to the elections themselves and partisanship of each Texas county comes from the Texas Secretary of State's website. This includes records of past voter turnout (defined

as votes cast divided by total voters registered), whether or not an incumbent was running, and the most recent Republican/Democrat presidential election margin from a given county. I also included polling data in the model: National presidential approval ratings were retrieved from Gallup, state-wide presidential approval ratings specific to Texas were retrieved from the Texas Politics Project, and records of the national generic congressional ballot (defined as whether voters prefer to vote for a Democrat or Republican for the House of Representatives in the upcoming congressional election) were retrieved from RealClearPolitics. Lastly, I also included binary variables indicating the type of race (senatorial, gubernatorial, or presidential), whether an incumbent was running in the race, and whether it was a presidential election year.

Demographic/economic data was retrieved from the American Community Survey 5 year estimates, which includes annual data on race, income, housing statistics, language, employment, as well as other data points, from a sample of about 3.5 million households collected over the previous 5 years. Based on past literature, I included Texas county-level data, including gender and racial breakdowns of the counties, mean income, population by income brackets, poverty status, household size, education level, percentage of people that rent vs. own their homes, the total population, median age, population by age bracket, employment rate, and percentage of the population in the labor force. This data was acquired for the years 2010, 2012 2014, 2016, and 2018 using the CensusData Python package.

In terms of modeling, I employed a variety of different methods that differed in the party primary being modeled (Democrat vs. Republican), the type of model being used (linear regression, random forest regression, extra trees regression, and gradient boosted decision trees), and the variables included (simple vs more complicated models). As was done in the Bloomberg model, the final models for my project use all of the training years in order to make its

predictions, but for training purposes, I implemented a cross-validation method in which one year of training data was withheld at a time. I was thus able to tweak the models and study feature importances before introducing the 2018 testing data.

Another factor worth noting is that early on in the project, I made the decision to remove the 2016 data from the test set for the Republican model. During early versions of my validation tests, I saw that including this data resulted in very poor performance for my models. Upon inspection of the 2016 turnout data, I saw that the most likely reason for this was that mean county turnout was extremely high, almost double that of 2010, 2012, or 2014. This was likely due to the fact that there was no Republican incumbent president in 2016, and so the primary was more competitive and thus had higher turnout. This aligns with many of the aforementioned theories about the influence of competitiveness on elections.

I hypothesize that were I to have included a presidential primary year in which Democrats also did not have an incumbent president, such as 2008 or 2020, then these years' data would have similarly skewed my results. Evidence for this lies in the fact that turnout in those years' Democratic primaries were quite high compared to others years, with 2008 having 22.5% turnout and 2020 having 12.5%. This is much higher than 2010, 2012, or 2014, which averaged 9.07%, 4.12%, and 3.52% turnout respectively. The one exception to this would be 2016, whose Democratic primary turnout was 7.34%. In this case, I hypothesize that this primary was uniquely uncompetitive, with Hillary Clinton never really being in danger of losing to any of her rivals, and so turnout was depressed. Nonetheless, I believe it is fair to say that when it comes to modeling mid-term primary years such as 2018, one can include presidential primaries in the testing data, so long as it was not too competitive of a contest.

Figure 3: Texas Dem Turnout Over Time

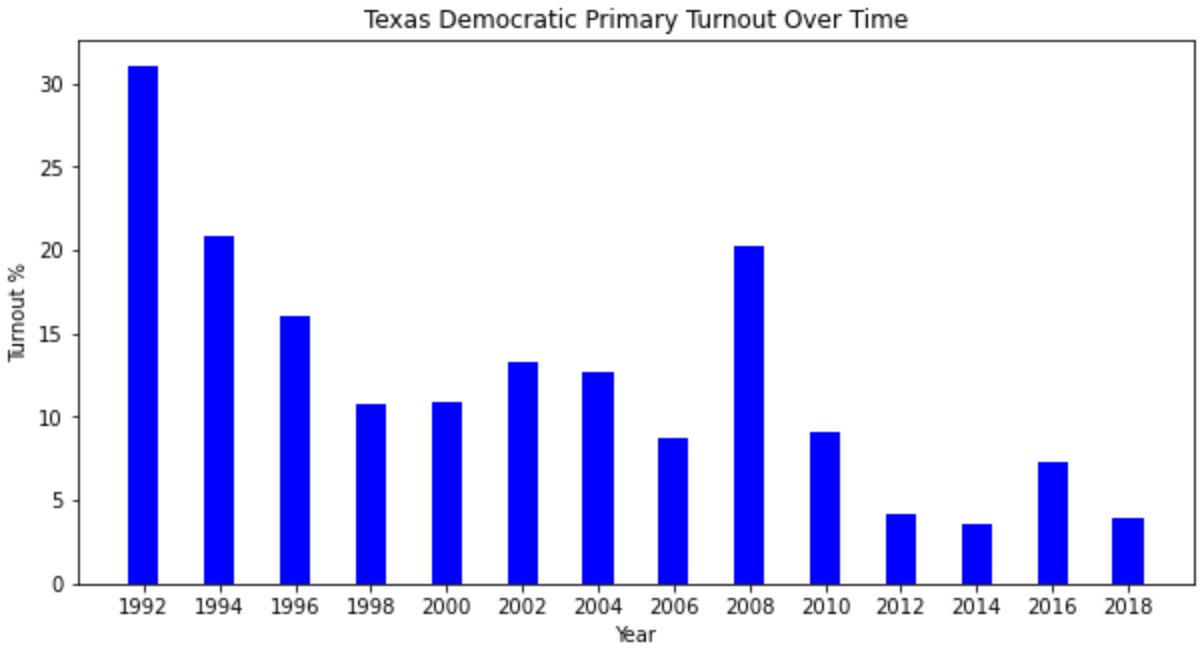
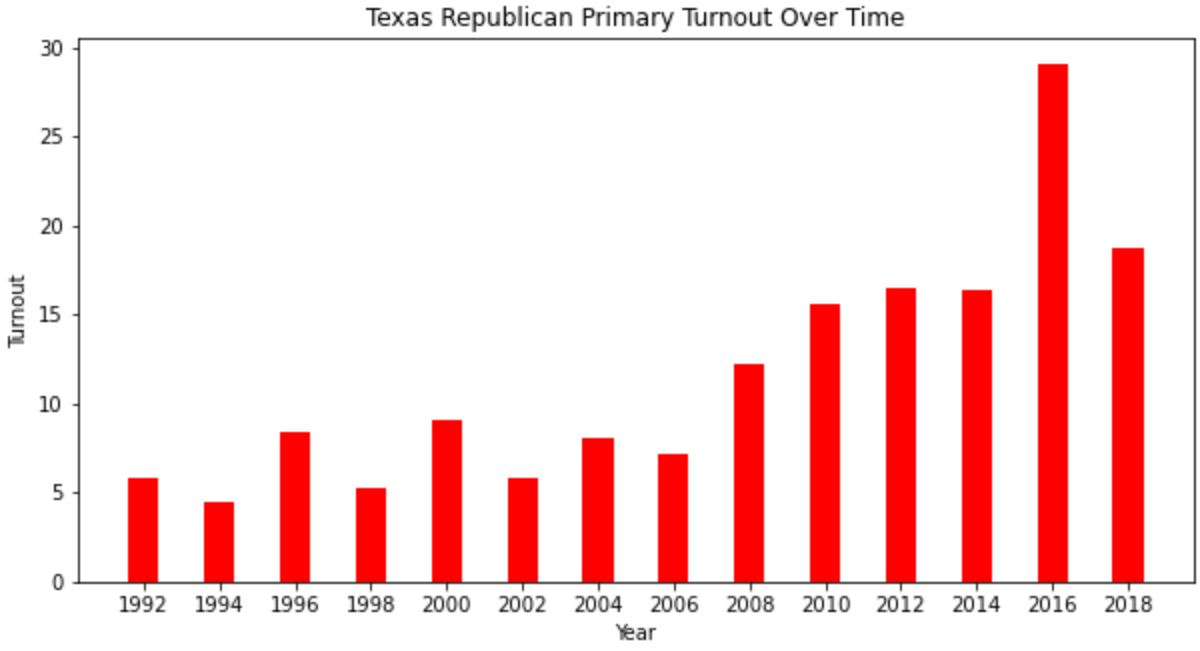


Figure 4: Texas Rep Turnout Over Time



My training data for my Democratic model contained 1524 rows, as it included records from each of Texas' 254 counties from 6 different elections. The Republican model contained

1270 rows after the removal of the 254 rows from the 2016 data. There were two sets of testing data for each party's model, one for the senatorial primary and one for the gubernatorial one, each consisting of another 254 rows, one per county.

A final note on terminology: for my project I defined "turnout" as the percentage of registered voters that cast votes in a given primary. According to Geys' study, this is the 2nd most popular way of defining turnout in the academic studies he collected, with the most frequently used definition being the percentage of the voting age population that voted (639). I use the ratio of registered voters' version of turnout as this is the statistic that the Texas Secretary of State reports. Were this research to be conducted in another state, it might be appropriate to measure turnout as the percentage of registered Republicans/Democrats that vote in their given primary. However, since Texas has open primaries, this would not make sense in this context, as anyone, not just registered Republicans/Democrats, can vote in a given primary, so long as they only vote in one.

4. EDA

Before modeling, I performed some exploratory data analysis. First, I checked the mean turnout values for each year in order to gauge what I ought to expect in terms of trends in turnout values (displayed in Figures 3 and 4 above). Then, I created histograms to show the distribution of values for all of my variables in both the Democratic and Republican datasets. My histograms (Figures 5 and 6) revealed that Republican turnout is relatively normally distributed, though it does have a leftward skew, while Democratic turnout is normally quite low, below 10%. Most racial variables skew left with the exception of White percentage (Figure 7) as well as Hispanic percentage (Figure 8), which is more normal but also skewed left. Other demographic variables are more normally distributed, including the binned income variables (10-15k per year, 15-25,

etc.), the percentage of the population with a high school education, and average household size. As would be expected from a Republican state such as Texas, the Republican presidential margin is skewed right.

Figure 5: Histogram of Rep Turnout

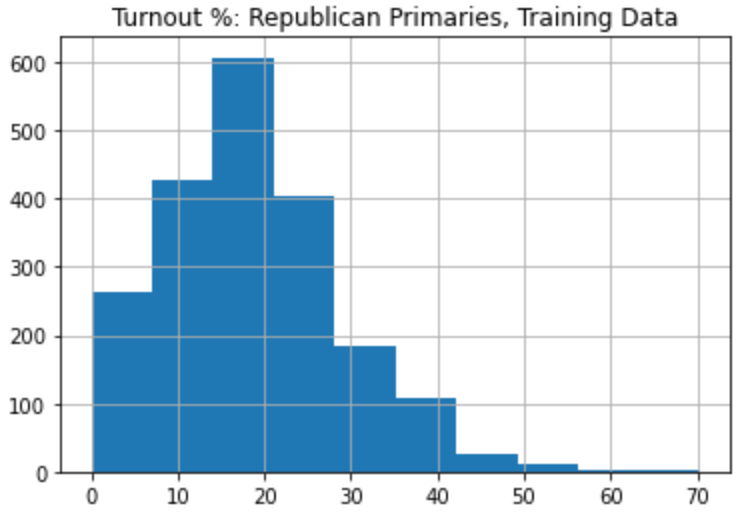


Figure 6: Histogram of Dem Turnout

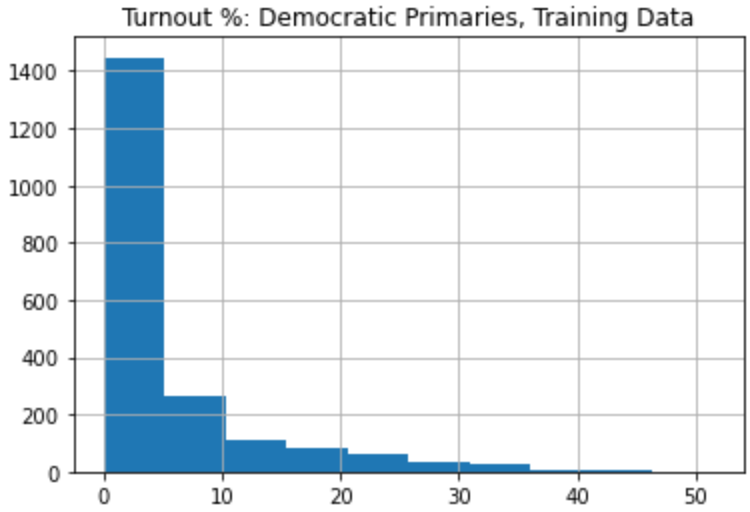


Figure 7: Non-Hispanic White % Histogram

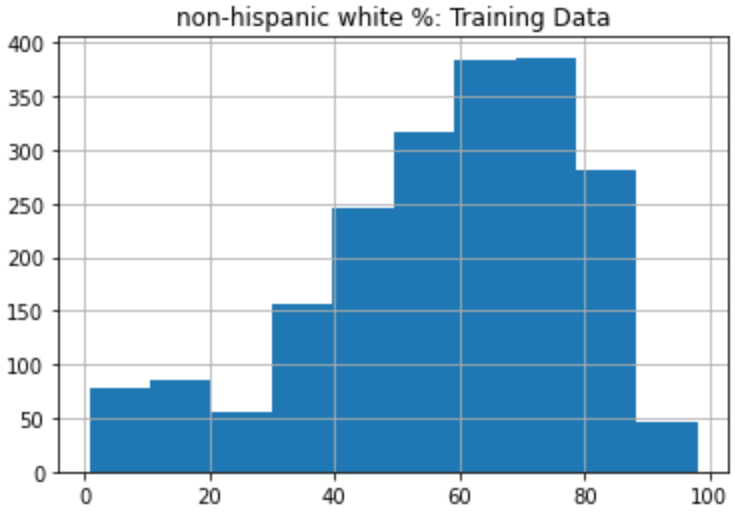
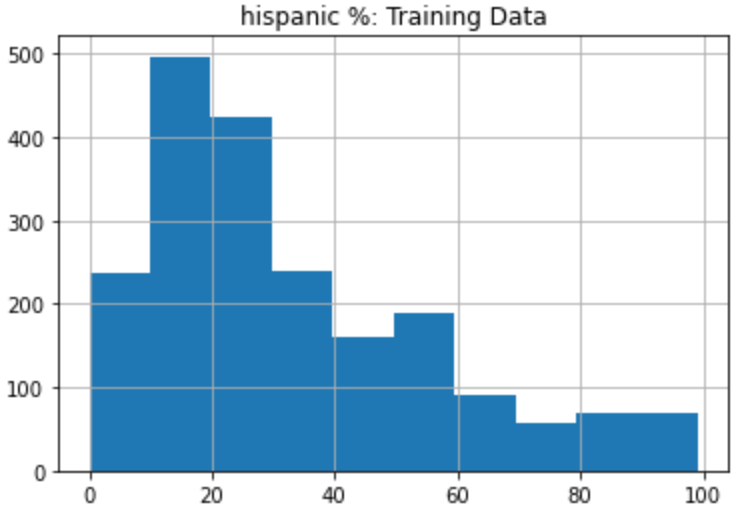


Figure 8: Hispanic % White Histogram



5. Model Tests: Training Data

The first models I created were individual linear regressions on each of my features in the training data. These were helpful in understanding whether they had positive or negative relationships with my outcome, and how important I might expect them to be in later models

based on their R^2 values. I found that of my 53 features, very few had high predictive power when tested on their own.

On the Republican side, variables that did have notable R^2 values included the most recent Republican presidential margin of the county, Hispanic percentage, non-Hispanic White (Figure 10) percentage, high school education rate, mean of past primary turnout, percentage of the population below the poverty line (Figure 9), percentage of the population making less than $10,000 per year, and household size.

Of these, Hispanic percentage, percentage making less than $10,000, percentage below poverty line, and household size, all had negative effects on turnout, while non-Hispanic White percentage, high school education rate, Republican presidential margin, and past turnout had positive effects.

Figure 9: Regression of Poverty on Republican Turnout

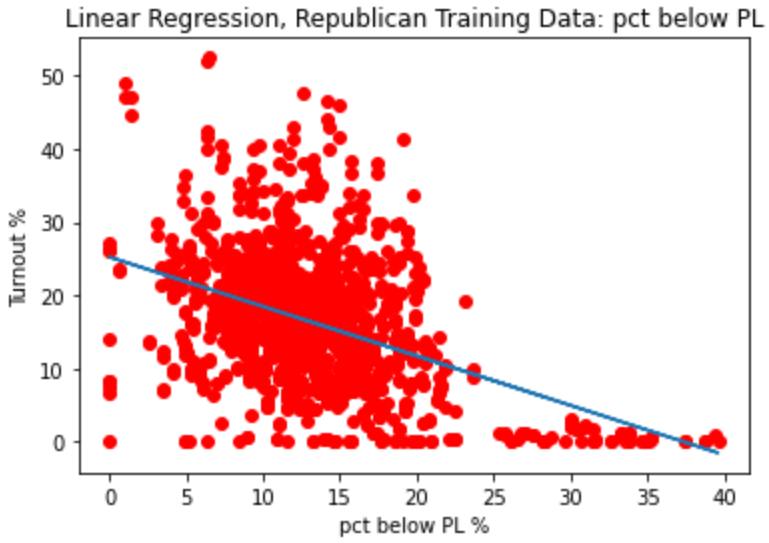
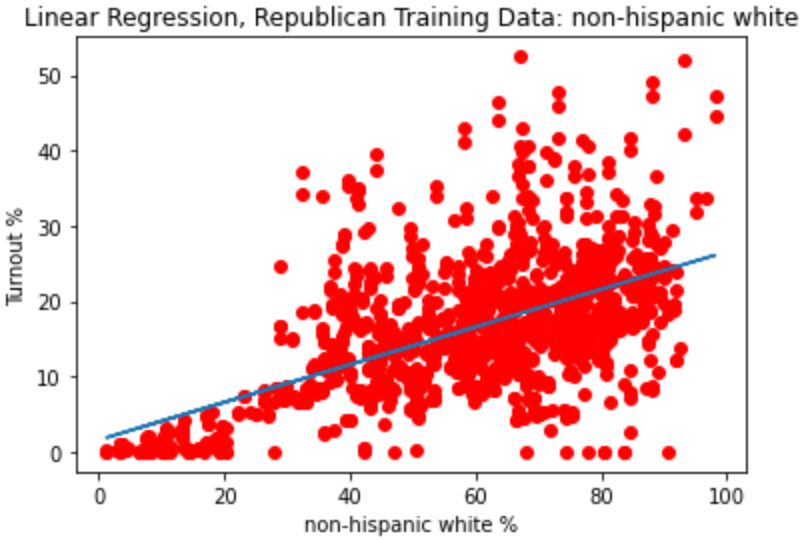


Figure 10: Regression of Non-Hispanic White % on Republican Turnout



I did the same for the Democratic training data, and found results that in many ways mirrored the ones from the Republican linear models. Hispanic percentage, percentage making below $10,000 and between $10,000 and $15,000, percentage below poverty line (Figure 11), and past turnout all had positive impacts on turnout, while non-Hispanic White percentage (Figure 12), high school education rate, and Republican presidential margin had negative effects.

Figure 11: Regression of Poverty % on Democratic Turnout

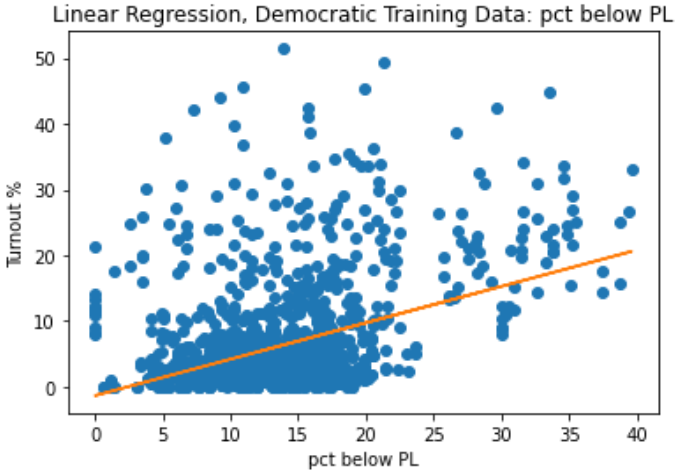
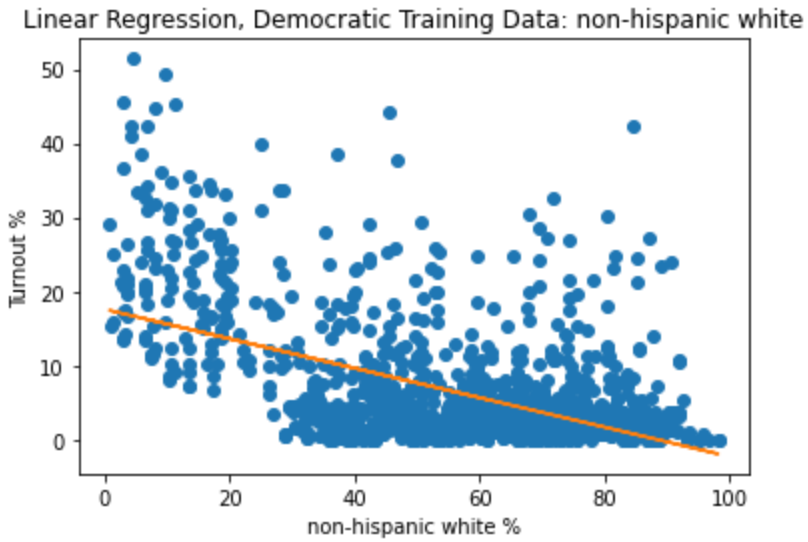


Figure 12: Regression of Non-Hispanic White % on Democratic Turnout



Apart from my full linear regression models, I created less complex linear regression models that only utilized these variables which displayed high R^2 values in their own independent linear regressions.

I examined my models' performance on my training data using validation testing, removing one year's worth of data at a time when training and then using the elections of the removed year as testing data. For example, I would remove all of the 2014 data and then test on both the 2014 senatorial and gubernatorial primary. When conducting these tests on the full Republican linear model, R^2 values ranged from 52% for the 2014 gubernatorial race to 61% for the 2012 senatorial primary (2016 was removed from the testing data). The Democratic linear models were much more varied, ranging from 5% R^2 for the 2014 Senate race to 62% for the 2012 presidential election.

When using the less complex linear regressions with only the significant variables, Republican R^2 values ranged from 57% for the 2012 Presidential race to 66% for the 2014 senatorial primary. The lowest Democratic validation score when using a smaller linear regression was 47% for the 2010 Governor's primary, and the highest, the 2012 presidential, reached all the way to 72%. Afterward, I conducted validation tests using random forest models. Performance on the Democratic side ranged from 51% for the 2010 Democratic Governor's race to 74% for the presidential race in 2012. Republican model performance ranged from 62% for the 2010 gubernatorial primary to 70% for the 2012 senatorial primary.

6. Model Tests: Testing Data

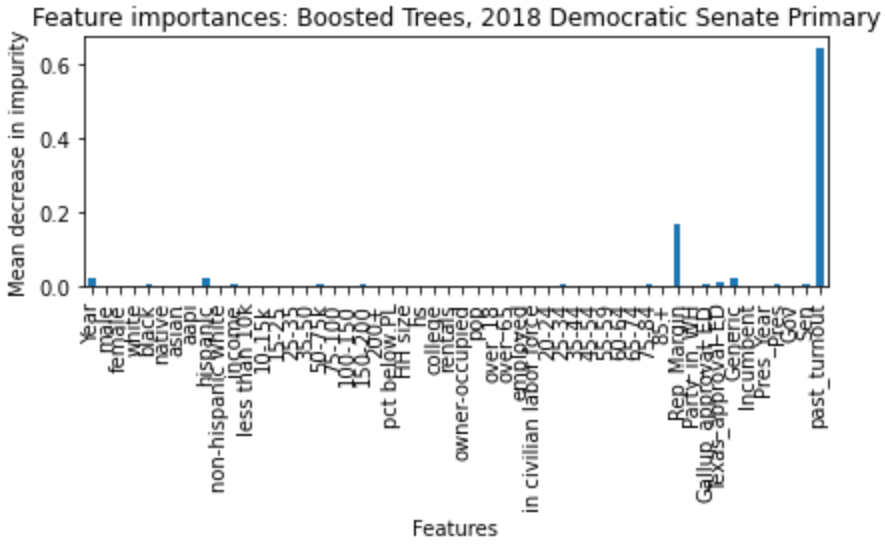
After validation testing, I moved on to testing the models, those that contained data from 2010-2016 for the Democratic model and 2010-2014 for the Republican model, on the 2018 data. I started with the full linear models, beginning with the Republican Senate primary in 2018, whose model had an R^2 of 64%, and the Republican gubernatorial primary, which also had a score of 64%. The Democratic full linear models performed extremely poorly, with R^2 scores in the negatives for both the Senate and Governor's race models. I then tried the less complex linear models. Republican model performance was 63% for the Senate and 62% for the Governor's race. The Democratic models improved significantly, achieving scores of 69% and 67% for the Senate and Governor's races, respectively.

I then tested the random forest models. Democratic model performance continued to improve, with scores of 86% for the 2018 senatorial primary and 82% for the gubernatorial primary. The Republican scores were slightly different, with scores of 70% each. Boosted trees models performed about the same for both parties, with scores of 86% and 82% for the Democratic Senate and Governor's races and 67% for both Republican races. Finally, I tested an

extra trees model for both parties. Democratic results remained similar to the other two tree based models, with scores of 79% and 78% for the Senate and Governor's races, while Republican scores were 72% and 73% for the Senate and Governor's primaries.

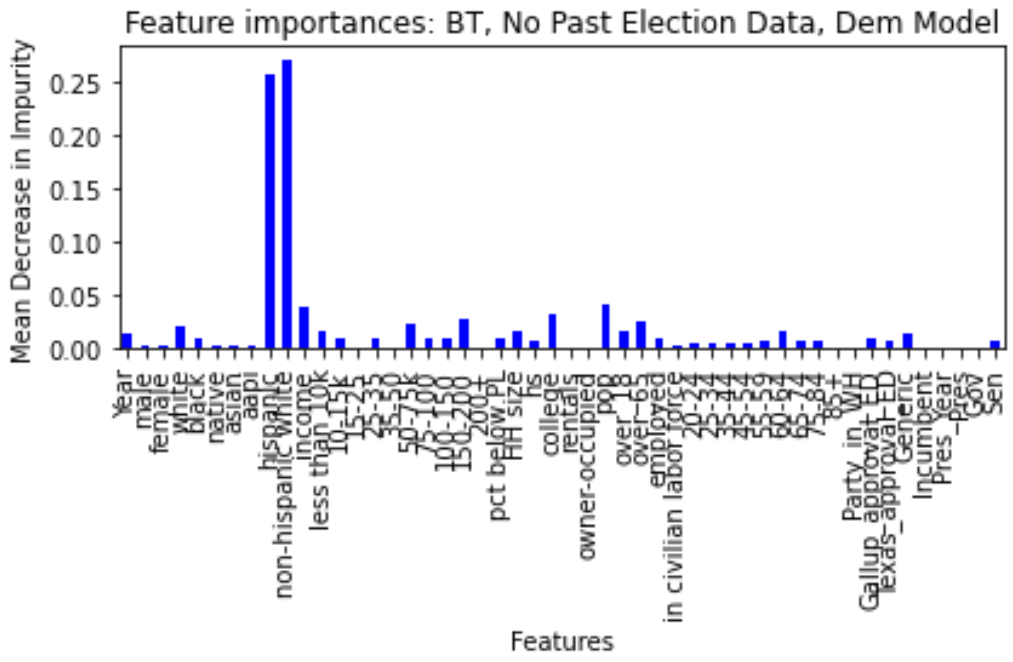
7. Feature Importances

Figure 13: Feature Importance for 2018 Democratic Primary Boosted Trees tModel



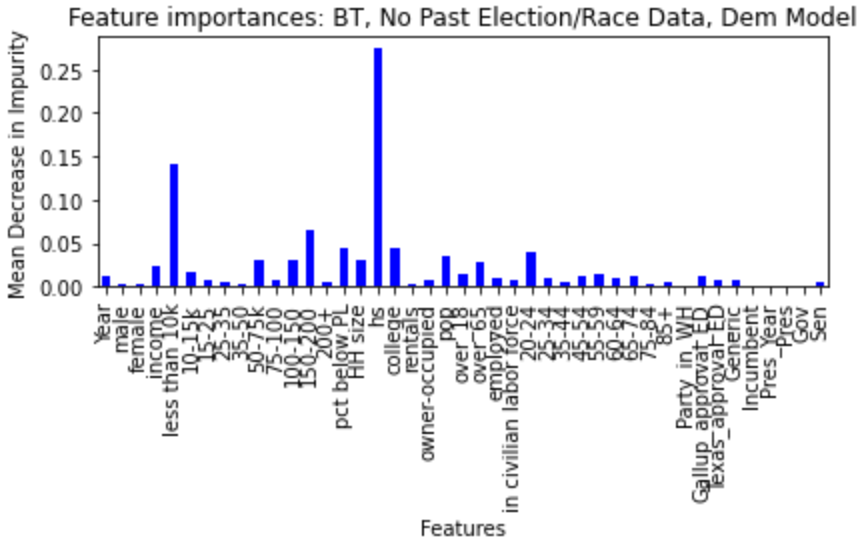
As is apparent from the above Democratic model graph for boosted trees (Figure 13), there is a high degree of asymmetry in terms of feature importance in the best performing models. This was also the case for linear regression, random forest, and extra trees. Mean of past turnout is playing a very large role in the predictions, and if it is removed then past presidential vote margin takes its place as the dominant feature without much reduction in model performance. I was thus curious to see what would happen if I removed these variables from the best performing models for each party, starting with boosted trees for the Democratic model.

Figure 14: Feature Importances in Dem Model Without Past Election Data



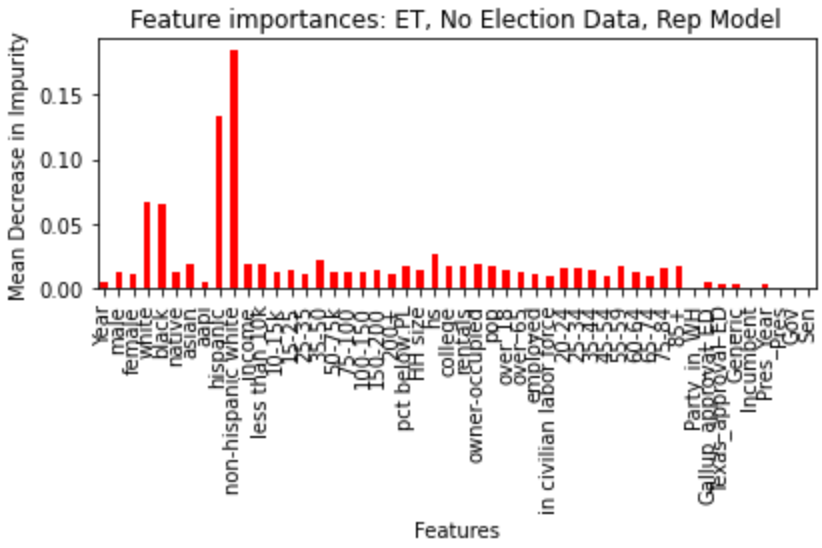
In the absence of data related to past voting behavior in the county, the model is still quite imbalanced in terms of feature importance, but shifts to racial data in order to make its predictions. The ratios of the Hispanic and non-Hispanic White population in each county become dominant. Model performance also suffers, dropping to 73% for the Senate race and 66% for the Governor race. As a final experiment, I also ran models that removed both the past election data and all data relating to race. At this point, the feature importances become more balanced, though still heavily skewed toward a few variables. High school education rates and the percentage of people making below $10,000 become the main sources of model predictions, though the percentage of people below the poverty line also plays a role. Model performance continued to decrease, reaching 53% for the Senate race and 52% for the Governor's.

Figure 15: Feature Importances in Dem Model Without Past Election or Racial Data



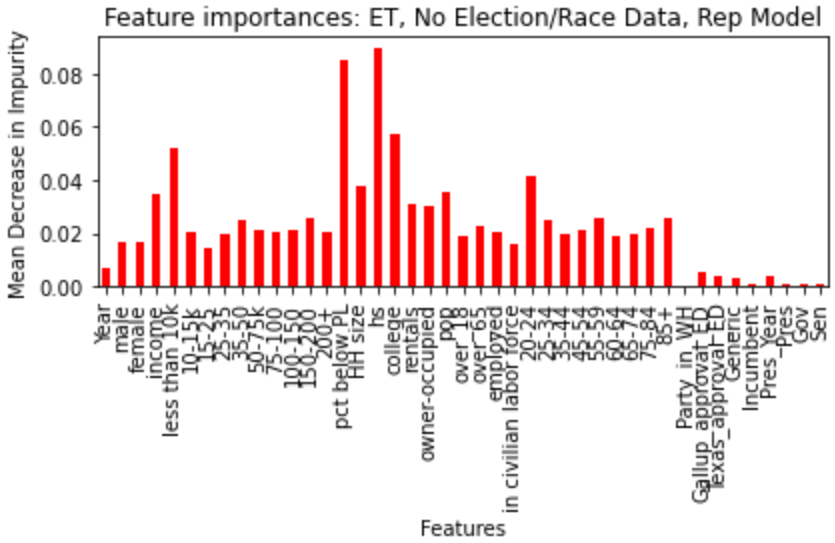
I did the same experiment with the Republican extra trees model, as it was the best performing model for that party. As was the case with the Democratic models, without past presidential margins or the means of past turnout, the percentages of Hispanic people and non-Hispanic Whites within the county play the largest roles, with the percentage of people making below $10,000 and $150,000-200,000 also dramatically increasing in importance. The models also lose much of their predictive power, dropping to 52% and 54% for the Senate and Governor's primaries.

Figure 16: Feature Importances in Rep Model Without Past Election Data



I then removed racial data, and found the following results.

Figure 17: Feature Importances in Rep Model Without Past Election or Racial Data



The feature importance in this case becomes much more evenly distributed. High school education, the poverty rate, and the less than $10,000 percentage are the most important, but college education, household size, percentage of the population between 20 and 24 years old, mean income, and the population of the county are also contributing factors. Notably, model performance does not drop after the removal of racial data, with scores 53% for both the Senate and Governor's races, indicating that in the Republican extra trees model, racial data may be acting as a proxy for information that is also contained in a combination of other demographic and economic variables. Such a proxy relationship also exists to some extent in the Democratic model, as it still performs reasonably well without the racial data. This result connects to pertinent debates about the role of race in determining political behavior, and prompts interesting questions about why this phenomenon differs between the Democratic and Republican models (McElwee et al.).

8. Limitations

My modeling approach could theoretically be used to predict other primaries beyond the 2018 races, such as 2020, using 2018 as a component of the training data and 2020 as the testing data. One potential challenge for future modeling, specifically for the Republican model, is the fact that I chose to remove 2016 from the training data due to its anomalously high turnout rate. This was done on the assumption that such a high turnout rate would not occur again in the foreseeable future, which I believe to be a fair prediction given how there is no precedent for this level of turnout in recent history and that it was likely a product of an exceptionally competitive//high profile primary race. Donald Trump, who is famous for his ability to induce high Republican turnout, as well as Texas' own senator Ted Cruz, were both running, likely contributing to higher than usual turnout (Riccardi). Nonetheless, I would even say that were

Trump and Cruz to both run again in 2024, turnout will not be as high as it is unlikely to be a competitive race, and Trump would probably have virtually secured the nomination by Texas' primary in March.

Another potential limitation is the fact that neither of the two presidential years that I included in my model had particularly competitive Democratic primaries. In 2012, the Democrats had an incumbent president running, and in 2016 the nomination was all but assured to go to Hillary Clinton. As is clear in the above turnout graph (Figure 3), turnout in the 2008 Democratic primary was quite high compared to the years that I studied, so it is possible that when the Democrats' do have a competitive primary, such as in 2020 or potentially in 2024 or 2028, my model's approach would not work as well.

Nonetheless, the model is not totally unable to handle competitive races. During validation testing, I saw that my model was able to predict results for the 2012 Republican Senate primary reasonably well, which was a competitive race between Ted Cruz and David Dewhurst. Additionally, the 2018 Democratic gubernatorial primary was a competitive race between Lupe Valdez and Andrew White. Perhaps there is something about competitive presidential races that requires a different approach, and this would be the subject of some interesting future research.

9. Conclusion

After testing on the 2018 data, my best-performing models are able to predict substantial majorities of the variation in the senatorial and gubernatorial primaries for that year, for both the Republican and Democratic races, though model performance is better for the Democratic ones. It is clear from the feature importance graphs that data related to past elections, such as past presidential margin and past turnout, play a significant role in these predictions. However, the

individual linear models proved that demographic/economic/racial features are also important. Such conclusions line up with much past research described in the literature review section, and also other known patterns such as White and/or wealthy voters' preference for Republicans and poorer and/or Hispanic voters' preference for Democrats, both in Texas and nationally, as I see in my models that counties with higher percentages of poorer/Hispanic residents have higher Democratic primary turnout, while those with higher ratios of wealthier/White residents have higher Republican primary turnout (Krogstad, Jens Manuel, et al).

My exploration of what happens when past election data and racial demographic are removed from the model could lead to some intriguing future research on the extent to which race acts as an independent factor in determining voting behavior versus a kind of shorthand for other more material data points such as education and income. Despite my models' limitations, this project demonstrates that Texas state-level primaries can be modeled relatively accurately, using both simpler linear models and more complex tree-based ones. It also illustrates how a small amount of data points related to past election history and race do most of the predictive work in such models.

Works Cited

Atkeson, Lonna Rae, and Cherie D. Maestas. "Presidential Primary Turnout 1972–2016." *PS: Political Science & Politics*, vol. 49, no. 04, Oct. 2016, pp. 755–60. *DOI.org (Crossref)*, https://doi.org/10.1017/S1049096516001608.

"Censusdata." *PyPI*, https://pypi.org/project/CensusData/.

Election Results, https://www.sos.state.tx.us/elections/historical/index.shtml.

"Explaining the Bloomberg News 2020 Election Turnout Model." *Bloomberg.com*, Bloomberg, www.bloomberg.com/graphics/2020-us-election-results/methodology.

Fernandez, Manny. "Texas: Primary Election Set for May 29." *The New York Times*, The New York Times, https://www.nytimes.com/2012/03/02/us/texas-primary-election-set-for-may-29.html.

Geys, Benny. "Explaining Voter Turnout: A Review of Aggregate-Level Research." *Electoral Studies*, vol. 25, no. 4, Dec. 2006, pp. 637–63. *DOI.org (Crossref)*, https://doi.org/10.1016/j.electstud.2005.09.002.

Kenney, Patrick J. "Explaining Primary Turnout: The Senatorial Case." *Legislative Studies Quarterly*, vol. 11, no. 1, Feb. 1986, p. 65. *DOI.org (Crossref)*, https://doi.org/10.2307/439909.

Kenney, Patrick J., and Tom W. Rice. "Voter Turnout in Presidential Primaries: A Cross-Sectional Examination." *Political Behavior*, vol. 7, no. 1, 1985, pp. 101–12. *DOI.org (Crossref)*, https://doi.org/10.1007/BF00987264.

Krogstad, Jens Manuel, et al. "Key Takeaways about Latino Voters in the 2018 Midterm Elections." *Pew Research Center*, Pew Research Center, 27 Aug. 2020, https://www.pewresearch.org/fact-tank/2018/11/09/how-latinos-voted-in-2018-midterms/.

Leighley, Jan E., and Jonathan Nagler. *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States*. Princeton University Press, 2014. *DOI.org (Crossref)*, https://doi.org/10.1515/9781400848621.

McElwee, Sean, et al. “Is America More Divided by Race or Class?” *The Washington Post*, WP Company, 7 Dec. 2021, https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/12/how-do-race-ethnicity-and-class-shape-american-political-attitudes-heres-our-data/.

Nownes, Anthony J. “Primaries, General Elections, and Voter Turnout: A Multinomial Logit Model of the Decision to Vote.” *American Politics Quarterly*, vol. 20, no. 2, Apr. 1992, pp. 205–26. *DOI.org (Crossref)*, https://doi.org/10.1177/1532673X9202000204.

Pedregosa, Fabian et al. “Scikit-learn: Machine learning in Python”. *Journal of machine learning research* 12.Oct (2011): 2825–2830. Print.

“Polling.” *The Texas Politics Project*, https://texaspolitics.utexas.edu/polling.

“The Politics of Financial Insecurity.” *Pew Research Center - U.S. Politics & Policy*, Pew Research Center, 20 Aug. 2020, https://www.pewresearch.org/politics/2015/01/08/the-politics-of-financial-insecurity-a-democratic-tilt-undercut-by-low-participation/.

“Presidential Job Approval Center.” *Gallup.com*, Gallup, 11 Nov. 2021, https://news.gallup.com/interactives/185273/presidential-job-approval-center.aspx.

RealClearPolitics - Generic Congressional Ballot. https://www.realclearpolitics.com/epolls/other/generic_congressional_vote-2170_1.html.

Riccardi, Nicholas. “Referendum on Trump Shatters Turnout Records.” *AP NEWS*, Associated Press, 9 Nov. 2020, https://apnews.com/article/referendum-on-trump-shatter-voter-record-c5c61a8d280123a1d340a3f633077800.

Texas Election Results 2016 – The New York Times. https://www.nytimes.com/elections/2016/results/texas.

U.S. Census Bureau. *2006-2010 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2010

U.S. Census Bureau. *2008-2012 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2012

U.S. Census Bureau. *2010-2014 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2014.

U.S. Census Bureau. *2012-2016 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file], 2016

U.S. Census Bureau. *2013-2018 American Community Survey 5-year Public Use Microdata Samples* [CSV Data file]. 2018

I pledge my honor that this represents my own work in accordance with University regulations.