Ballot Box Proximity and Voting Likelihood in Washington State

Audrey Yu, Krishna Saxena, Noah Huck, Zachary Bi

6 December 2024

Access Analyses on Github

Abstract

Low voter turnout and overall voter likelihood has been a long-standing issue in the United States. It is clear that through higher voter representation, the democratic values of the country are better upheld. With the aim of addressing the problem of low voter turnout in Washington state and motivated by previous experimental analyzes [Col+18] [McG+20] that have discussed this issue, our project focuses on the relationship between a person's distance from the nearest ballot box and their likelihood of voting in Washington. As there are many variables in addition to the distance from a ballot box that influence someone's propensity to vote, we also analyze the effect of covariate variables in our analysis. Our analysis highlights the negative correlation between distance to ballot boxes and voter turnout, along with the relationship to other covariates.

1 Introduction

The election process in the United States has always contained many checks and balances. However, these systems often lead to close votes seemingly decided with a large divide. This is evident especially during presidential elections where votes in particularly competitive states are off by fewer than a percentage of the states votes; still leading to all of the states electoral votes going to one candidate (as by the electoral college system). Although this polarizing system is not as evident in the state of Washington, it upholds the democratic nature of our country to decrease barriers to voting and incentivize suffrage across all communities. Access to voting has long been a barrier to the election process. While mail-in ballots and ballot drop boxes have mitigated that effect in Washington, we research the question: "How much does proximity to a ballot box impact Washington residents' likelihood to vote?" We determine a quantitative measure of the percent increase in likelihood per kilometer decrease in distance for voting age individuals who are registered voters.

Past studies indicate such a relationship, such as previous work by Collingwood et al. [Col+18] and McGuire et al. [McG+20], which have experimentally found a negative relationship between ballot box distance and likelihood to vote in King and Pierce counties, respectively (although [Col+18] finds a strong correlation in non-presidential elections only). These analyses focused on the effects of additional ballot boxes added between elections. While Collingwood et al. did not have an experimental design, McGuire et al. utilized a placebo mailing box to study this relationship. We conduct an observational study on a single statewide election utilizing public data.

We expect that our approach will provide a better representation of all Washington voters' behavior and reduce possible biases that may be introduced through unanticipated changes between elections. However, we also anticipate that this design will introduce bias stemming from differences in socioeconomic background, political preference, etc. across Washington's regions. We mitigate this by weighting data points by the number of voters they represent, along with modeling covariates to get a clearer look at the effect of ballot box distance on voter turnout.

We have two research questions that we wished to address for Washington voters:

RQ1: How does the distance from a ballot box impact a voter's likelihood to vote?

RQ2: What demographic attributes most strongly correlate with overall likelihood to vote?

2 Past Work

2.1 Collingwood et al.

Collingwood et al. [Col+18] investigates the impact of ballot drop boxes on voter participation in King County. The authors analyze turnout data before and after the implementation of widespread drop box availability, assessing whether this policy enhances voter convenience and increases engagement. The study pays special attention to demographic and geographic variations in turnout, revealing how drop boxes can serve as a tool to reduce voting barriers and promote equity in electoral participation.

2.2 McGuire et al.

McGuire et al. [McG+20] examines how the geographic proximity of drop boxes influences voter turnout. Using empirical data, the authors analyze the relationship between distance to ballot drop boxes and voter participation, highlighting disparities in access and the potential for increased turnout with strategically placed drop boxes. The research is particularly focused on understanding the equity implications of ballot box placement for different demographic groups. Their study relied upon guessing different voter's races through factors such as location, name, and even their private shopping data. This may have introduced bias into this study as the race was not publicly available data and they had to generate and connect those themselves.

3 Data

3.1 U.S. Census

The U.S. Census is collected every decade for the purposes of drawing congressional districts, distributing congressional seats, determining where to distribute federal funds, etc. As it is a census performed by the government for administrative purposes, we can be assured that the dataset is high quality. For this analysis we utilized the 2020 U.S. Census [Bura] to determine voting age populations for each block group in Washington, along with several of our census block group level covariates (race [Burf], median income [Bure], poverty status [Burc], education attainment [Burb], and employment [Burd]). We chose to examine data from 2020 due to the coinciding decennial census and presidential election.

3.2 Washington Secretary of State

We also obtained voter data and ballot box locations from 2020 through the Washington Secretary of State website [Sta20b]. The purpose of this dataset was to help Washington voters find nearby

ballot boxes, increase public transparency, and archive administrative data. We additionally used the WA Secretary of State's Voter Registration Database [Sta] to obtain addresses of all Washington residents who voted in 2020. Combined, these datasets can help us estimate how far voters are to their nearest dropbox. Like the census data, we are highly confident that these datasets are unbiased and accurate, because they were made for for government administrative purposes.

3.3 Firsthand look at the data

Many variables, including voting age population, were not available at a finer granularity than by census block group, limiting our analysis from focusing on individuals. Additionally, a small proportion of voters (approximately 8%) could not be accurately geocoded to a block group, leading to a slight undercount of turnout. Some of these were due to nonresident voters, P.O. boxes, etc., while others were due to limitations of the census geocoder. Finally, we noticed that based on the available data, a few smaller block groups were estimated to have higher than 100% turnout. We believe this is due mainly to privacy protection efforts by the Census Bureau introducing uncertainty into the Census data, and we discuss efforts to mitigate this in the following section.

4 Methods

Our goal was to determine a relationship between each registered voter and their likelihood to vote given their distance to a ballot box. To accomplish this goal we aimed to create a model that would be able to find the likelihood that someone would vote given their distance to a ballot box. Letting Y_i be the event that the ith person voted and D_i be the distance of the ith person to the nearest ballot box, we were interested in finding the $\mathbb{E}[Y_i|D_i]$. Additionally, we note that there are many factors that could play a role in each voter's likelihood to vote. We aimed to utilize these covariates within our analysis as well, which we can denote to now be, letting X_i be some ith person's covariates, $\mathbb{E}[Y_i|D_i, X_i]$. To

To find this expected value, we utilized a linear regression similar to the one performed in Collingwood et al., as we believe that it would most strongly correlate to the voting behavior we aimed to analyze. To discover the relationship between these two variables, along with correlations, it is important for us to establish a regression model with a clear direction, leaving the best choice to be a linear regression. The model specifications are as follows:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \tag{1}$$

and our second model which included covariate terms followed this form:

$$Y_i = \beta_0 + \beta_1 D_i + \overrightarrow{\beta}^T \overrightarrow{X_i} + \epsilon_i, \tag{2}$$

where β_0 is a bias term (the expected turnout when all covariates and distance are 0), β_1, $\overrightarrow{\beta}$ are weights for distance and covariates, and ϵ_i is an error term.

Additionally, we utilized the Gelbach decomposition [Gel16] to determine how the linear regression coefficient for distance is influenced by other covariates.

4.1 Tools

We mainly used the **Python** and **R** programming languages along with available packages such as **Pandas** and **Geopy** to prepare our data for analysis. Pandas was utilized to read and merge CSVs along with clean them to be usable from the state and federal sources, and we used Geopy

to calculate distances from coordinate data. With additional packages, namely **Scipy** and **Scikit-learn** we performed the statistical components required of our analysis, including estimations of linear regression coefficients along with significance tests. Finally, to create the map visualization, we used **JavaScript**, along with the **Leaflet** package for rendering maps.

4.2 Data Cleaning

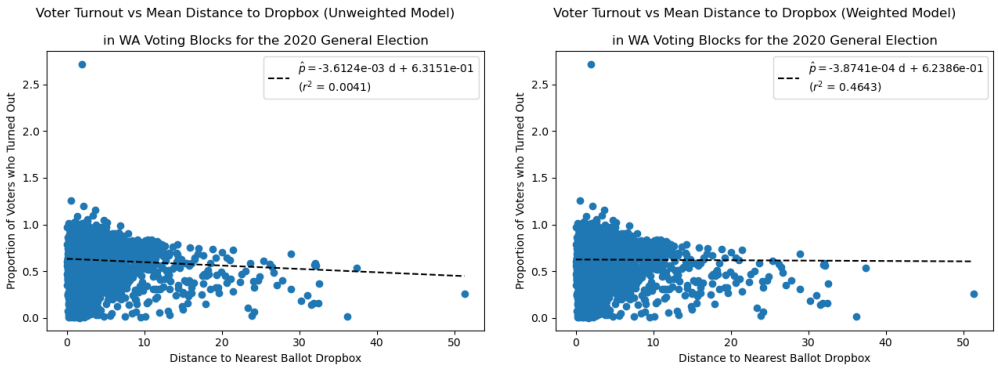
The primary data quality issue we faced was incomplete data across our data sources. The Washington Sec. of State data sources [Sta20a], [Sta] contain 5311 unique census blocks. However, we only have election results localized to 5290 of these blocks, so there are Not a Number (NaN) artifacts when we merge average distance to dropboxes for each census block to the blocks' voting rates. Thus, the regressions in 5.1 are limited to the data of these 5290 blocks. Similarly, when merging all the covariate dataframes, we found NaN artifacts for all but 3355 census blocks. Because of the uncertainty in NaN observations, the analyses in 5.2 include data from only these 3355 census blocks.

Additionally, during our analysis we discovered disparities in our data sources which caused the appearance that over 100% of voters in a census block voted. This is likely due to absentee ballots or people submitting their ballots in a different census block in their county of registration. Our later models account for this inconsistency by capping the voting rate to 100%.

5 Results

5.1 Base Regression Model

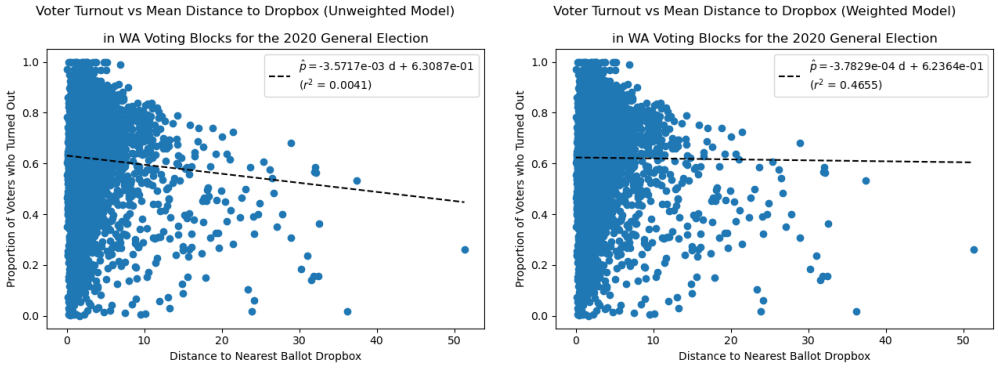
Our first analysis was a regression of the base model 1. This model showed evidence that discovered a coefficient of $-3.6124 \cdot 10^{-3}$, which was significant at the $\alpha = 0.001$ level. However, this model did not fit the data well, as shown by the r^2 of 0.0041.



To increase the analysis' robustness, we introduced weighting on block groups by their population. This produced a much better regression model with a more promising r^2 of 0.4643. The model's regression coefficient was still negative, but smaller at $-3.8741 \cdot 10^{-4}$. This coefficient was still significant at the $\alpha = 0.001$ level, so this model overall gives us confidence that increasing distance to drop boxes reduces voting likelihood.

Next, we cap voting probability to 1 for each census block. Without weighting by population, we find that every extra kilometer from a dropbox reduces the probability of voting by $\approx 0.36\%$. While this slope parameter was significant at the $\alpha = 0.001$ level, this model fit the data poorly evidenced by its r^2 value of 0.0041. After weighing observations by census block population, we

find a higher r^2 value, but the slope parameter is insignificantly different from 0 at even a $\alpha = 0.5$ level.



5.2 Covariate Analysis

Since the weighted model with capped probabilities showed an insignificant relationship between average distance to dropboxes and propensity to vote, we decided to investigate how other covariates also influence the likelihood to vote coefficient. We add additional covariates related to the racial composition [Burf], employment rates [Burd], education attainment [Burb], median income [Bure], and 2016 voting results [VT18] in each census block and run a linear regression of propensity to vote vs distance to dropbox and these covariates weighted by population. (To prevent multicolinearity, one variable from the first three of these classes is excluded: percent of block group identifying as multiracial, percent of block group employed, percent of block group with graduate degrees). The regression coefficient table is below 1:

Table 1: Full Model Regression Coefficients and their Standard Errors

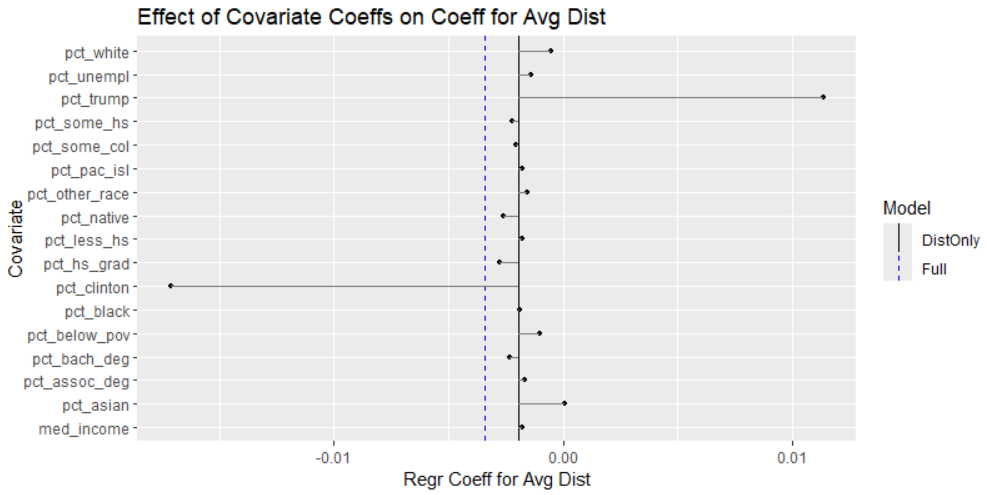
	Estimate	Std. Error	Significance
(Intercept)	-5.189e-02	1.310e-01	
Distance to Nearest Dropbox (km)	-3.436e-03	7.850e-04	***
Median Income	1.159e-06	9.370e-08	***
% white Residents	1.191e-03	4.639e-04	*
% Black Residents	-4.370e-05	6.156e-04	
% Native American Residents	-2.849e-03	6.698e-04	***
% Asian Residents	-3.095e-03	5.322e-04	***
% Pacific Islander Residents	-2.385e-03	9.862e-04	*
% Residents Other Race	-2.885e-03	5.551e-04	***
% Citizens Below Poverty Line	-4.030e-03	3.132e-04	***
% Citizens No High School	-3.383e-03	8.498e-04	***
% Citizens Some High School	-2.851e-03	6.216e-04	***
% Citizens High School Grad	-2.123e-03	4.012e-04	***
% Citizens some College	-1.363e-03	4.189e-04	**
% Citizens with Associate's Degree	2.129e-03	5.626e-04	***
% Citizens with Bachelor's Degree	9.333e-04	5.025e-04	.
% Residents Unemployed	7.491e-04	2.321e-04	**
% Trump 2016 Votes	6.226e-03	1.278e-03	***
% Clinton 2016 Votes	7.257e-03	1.250e-03	***

'.' $\alpha = 0.1$ significant, '*' $\alpha = 0.05$ significant, '**' $\alpha = 0.01$ significant, '***' $\alpha = 0.001$ significant

This new model has a coefficient between voting likelihood and mean distance to nearest dropbox (in km) of $-3.436 \cdot 10^{-3}$ and this coefficient is significant at the $\alpha = 0.001$ level. This means that, on average, every ≈ 2.91 km closer dropboxes are to voters, voters are 1% more likely to vote.

5.2.1 Gelbach Analysis

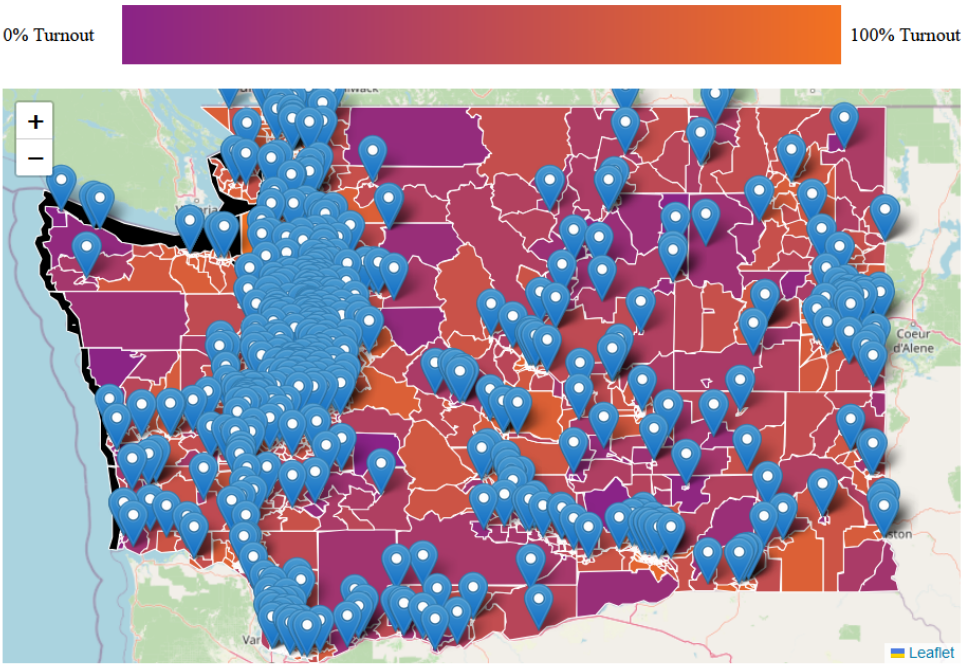
We conduct a Gelbach decomposition [Gel16] to analyze which covariates contributed to the tightening of the confidence interval of distance coefficient β_1 from the base model 1 to the full model 2. The Gelbach decomposition δ-values for each covariate added to the full model are plotted below 5.2.1:



The biggest conclusion from the Gelbach decomposition is that candidate preference in the 2016 presidential election greatly effects the relationship between voting probability and average distance to nearest dropbox in the 2020 presidential election. For example, for every percent increase in a census block's Trump support in 2016, the relationship between likelihood of voting and average distance to dropboxes in 2020 grows by 1.14%. Likewise, for every percent increase in a census block's Clinton support in 2016, the relationship between likelihood of voting and average distance to dropboxes in 2020 drops by 1.71%. This suggests that dropbox accessibility is more effective at driving turnout in Democratic voting regions. The state should study whether the sparsity of drop boxes in traditionally Republican regions (especially high turnout rural regions) is driving this phenomena or if the Secretary of State needs to do targeted outreach to increase awareness of dropboxes in/near low turnout Republican regions.

5.3 Map Visualization

We produced a visualization to better show the relationship between ballot box distance and voter turnout. The map shows turnout for each block group or census tract on a gradient, with the locations of ballot boxes marked. An interactive version can be accessed here.



6 Conclusion

We find a weighted single-variable linear regression model with $r^2 \approx 0.4643$, that provides evidence that voting rates drop 0.38741% per additional kilometer voters live from a ballot box. Meanwhile, a more sophisticated linear analysis with additional covariates revealed a more significant coefficient of $-3.436 \cdot 10^{-3}$ (0.34 % per km) , which corresponds with a 1% increase in voting likelihood for every 2.91 km closer a voter is to the nearest dropbox.

Comparatively, Collingwood et al. [Col+18] estimates a 2% decrease in voting likelihood per mile additional distance (0.8 % per km) in King County, while McGuire et al. [McG+20] experimentally found a coefficient of 0.64% per mile (0.40 % per km) in Pierce County. While our analysis differs significantly from the results from Collingwood et al., we believe this is due to their focus on King County, the most population-dense county in Washington. That is, the difference between 1 and 2 miles in a dense city may be much more impactful to a voter than the difference between 50 and 51 miles in a rural area. Meanwhile, our results were quite close to those in McGuire et al., with a difference of only 0.05 % per km, and expands the analysis to the entirety of Washington state.

7 TL;DR

We performed an analysis utilizing open source data from official sources, namely the U.S. Census and the Washington Department of State to determine the relationship between ballot box availability and voter turnout. By performing a linear regression analysis, we were able to discover a negative correlation between ballot box distance and voter likelihood, and a further covariate analysis highlighted that political preferences greatly influence how drop box accessibility correlates with voter turnout. The multivariate model gives evidence that, on average, every approximately 2.91 km closer dropboxes are to voters, voters are 1% more likely to vote. These results underscore

the importance of geographic accessibility in future policies to ensure all Washingtonians are able to exercise their constitutional right to vote.

8 Team Contributions

Our team collaborated effectively to produce a comprehensive project, with each member contributing their unique skill sets. The work was split approximately as follows: Noah and Krishna each contributed to 30% of the whole project, and Zach and Audrey each contributed 20%.

Noah played a pivotal role in data sourcing and preparation, ensuring that our dataset was clean, organized, and ready for analysis. He also contributed to data visualization, crafting clear and impactful visual representations of our findings. Additionally, Noah contributed to the writing process and conducted an in-depth literature review to ground our work in existing research.

Krishna focused on both data sourcing and performing detailed data analysis, extracting meaningful insights that informed the conclusions of our project. His analytical work on linear modeling and coefficient decomposition was essential to our interpretation of the results. Alongside this, Krishna participated in the writing process, ensuring the report was cohesive and clearly articulated.

Zach collaborated with Noah on data sourcing and preparation, assisting in preparing the dataset for analysis. He also greatly contributed to the writing process and played a key role in the literature review, working to establish a strong framework for the project.

Audrey contributed to the writing process, helping to refine and structure the report. She also worked on the preparation of the presentation, ensuring that the team's findings were communicated effectively and engagingly. Her efforts in conducting a literature review further supported the research and contextualized the results.

9 Project Reflection

This project provided valuable insights into the relationship between ballot box proximity and voter turnout in Washington State. Several aspects of the project went well, contributing to its overall success. Access to reliable, publicly available datasets from the U.S. Census and the Washington Secretary of State ensured the integrity of our analysis, allowing us to focus on meaningful patterns rather than data quality concerns. Collaboration among team members was another key strength; leveraging individual skills in data analysis, programming, and research ensured an efficient workflow, supported by tools like GitHub for version control and code sharing. Methodologically, our approach was innovative, particularly with the use of weighted regression models to account for population disparities across block groups.

Despite these achievements, the project also faced several challenges. One limitation was the granularity of the data; while block group-level analysis provided broad insights, it obscured individual voter behaviors and hyper-local socioeconomic patterns that could influence turnout. Additionally, our statistical models, though robust, captured only a portion of the variance in voter turnout, suggesting that more complex approaches, such as nonlinear models like logistic regression and neural networks, could be explored in future research. Lastly, while our covariate analysis revealed some interesting trends, further investigation into intersectional factors, such as the interplay between race and wealth, is needed for a more nuanced understanding of voter behavior.

Overall, this project demonstrated the importance of balancing statistical rigor with effective communication and highlighted the need to address potential biases proactively. It also reinforced the value of collaboration and technical skill development when it comes to complex research questions. While there is room for improvement in both methodology and interpretation, the

insights gained from this analysis contribute meaningfully to discussions on voter accessibility and democratic participation.

References

[Gel16] Jonah B. Gelbach. "When Do Covariates Matter? And Which Ones, and How Much?" In: *Journal of Labor Economics* 34.2 (2016), pp. 509–543. ISSN: 0734306X, 15375307. URL: https://www.jstor.org/stable/26553211 (visited on 12/05/2024).

[Col+18] Loren Collingwood et al. "Do Drop Boxes Improve Voter Turnout? Evidence from King County, Washington". In: *Election Law Journal: Rules, Politics, and Policy* 17 (Feb. 2018). DOI: 10.1089/elj.2017.0450.

[VT18] Voting and Election Science Team. *2016 Precinct-Level Election Results*. Version V94. Accessed at https://github.com/Davidvandijcke/Census-Block-Group-Level-2016-Presidential-Election-Results on 5 December 2024. 2018. DOI: 10.7910/DVN/NH5S2I. URL: https://doi.org/10.7910/DVN/NH5S2I.

[McG+20] William McGuire et al. "Does Distance Matter? Evaluating the Impact of Drop Boxes on Voter Turnout". In: *Social Science Quarterly* 101.5 (2020), pp. 1789–1809. DOI: https://doi.org/10.1111/ssqu.12853. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ssqu.12853. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ssqu.12853.

[Sta20a] WA Secretary of State. *2020 General Election Data*. WA Secretary of State. Accessed on 25 October 2024. 2020. URL: https://www.sos.wa.gov/elections/data-research/2020-general-election-data.

[Sta20b] WA Secretary of State. *Voting Locations and Ballot Boxes*. WA Secretary of State. Accessed on 5 December 2024. Oct. 2020. URL: https://geo.wa.gov/datasets/voting-locations-and-ballot-boxes/explore.

[Bura] U.S. Census Bureau. *CITIZEN, VOTING AGE POPULATION*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2020.B05001?q=Citizen%20Population&g=040XX00US53,53$1400000&y=2020.

[Burb] U.S. Census Bureau. *CITIZEN, VOTING-AGE POPULATION BY EDUCATIONAL ATTAINMENT*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2019.B29002?t=Educational%20Attainment&g=040XX00US53$1500000&y=2019&d=ACS%205-Year%20Estimates%20Detailed%20Tables.

[Burc] U.S. Census Bureau. *CITIZEN, VOTING-AGE POPULATION BY POVERTY STATUS*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2019.B29003?t=Poverty&g=040XX00US53$1500000&y=2019&d=ACS%205-Year%20Estimates%20Detailed%20Tables.

[Burd] U.S. Census Bureau. *EMPLOYMENT STATUS FOR THE POPULATION 16 YEARS AND OVER*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2019.B23025?t=Employment%20and%20Labor%20Force%20Status&g=040XX00US53$1500000&y=2019&d=ACS%205-Year%20Estimates%20Detailed%20Tables.

[Bure] U.S. Census Bureau. *MEDIAN HOUSEHOLD INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS)*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2019.B19013?t=Income%20(Households,%20Families,%20Individuals)&g=040XX00US53$1500000&y=2019&d=ACS%205-Year%20Estimates%20Detailed%20Tables.

[Burf] U.S. Census Bureau. *RACE*. U.S. Census Bureau. Accessed on 5 December 2024. URL: https://data.census.gov/table/ACSDT5Y2019.B02001?t=Race%20and%20Ethnicity&g=040XX00US53$1500000&y=2019&d=ACS%205-Year%20Estimates%20Detailed%20Tables.

[Sta] WA Secretary of State. *Voter Registration Database*. WA Secretary of State. Accessed on 20 October 2024. URL: https://www.sos.wa.gov/elections/data-research/election-data-and-maps/reports-data-and-statistics/washington-state-voter-registration-database-vrdb.