United States Water Access

A Research Project

Brooke Wheeler, DATA-412 Ashley Mann, DATA-412
Natasha LaChac, DATA-412

12/15/2021

Executive Summary

The goal of this research project was to review water summary data from across the United States and analyze the relationship between states or regions and the number of violations, facilities, site visits among each group. This project was meant to show the relationship between clean water access and geographic and demographic factors. Researchers analyzed this relationship by using scatterplots and other visualizations. As of 2021, nearly 90% of the United States receive their drinking water from community water systems (CWS) (Uche et al., 2021). Examining CWS violations and their relation to demographic data will enable future researchers to predict the next water crisis, like the events in Flint, Michigan between the years 2014 and 2019. The data set used in this project was sourced directly from the Environmental Protection Agency (EPA) Drinking Water Violations Database. The project combined this data with demographic data sourced directly from the U.S. Census Bureau, which was collected in the 2019 American Community Survey (ACS). Demographic data were also collected from the state of Mississippi on a county level.

Introduction

The datasets used in this project came directly from the Environmental Protection Agency (EPA) Drinking Water Violations Database (EPA, 2015) and the U.S. Census Bureau. States under the Safe Drinking Water Act are required to report drinking water information to the EPA, which is stored in the federal data warehouse (EPA, 2015). The water violation dataset was obtained from this data warehouse for the year 2021 and has a total of 143,296 observations. There are a total of 11 variables in this dataset including PWS Id, PWS Name, EPA Region, Primacy Agency, PWS Type, Population Served, Cities Served, Counties Served, Number of Facilities, Number of Violations, and the Number of Site Visits. The code used the tidycensus package to directly access the U.S. Census data. The data were taken from the 2019 (ACS), which was the most recent available comprehensive data. The variables were taken from census data by state across the country. There were 10 variables in the data table, including GEOID, Name, Population, Median Income, and populations divided by race, including White, Black, Native, Asian, Pacific Islander, and Two or Other.

Literature Review

Previous research has shown that it may be possible to predict the next water crisis based on a number of factors. A case study of Flint, Michigan, where there was a highly publicized water crisis, showed that Flint's aging infrastructure and repeated water violations led to the water crisis (Forrer et al., 2019). It also discussed that institutional racism in a city that is over 50% non-white, as well as poverty, may have caused Flint's crisis to be ignored for approximately 18 months after citizen complaints before any action was taken

(Forrer et al., 2019). A 2019 study by the Natural Resources Defense Council (NRDC) showed that at a county level, non-white and low-income populations were more vulnerable to higher rates of water violations (Fendick et al., 2019). The NRDC also found that smaller community water systems, defined as those that served fewer than 3,3000 people, accounted for over 80% of water violations (Fendick et al., 2019). The results of this study have been corroborated by other studies. The International Journal of Environmental Research and Public Health published a study in 2021 using CWS and ACS data in California and Texas. The study found that Black and Latino communities were more likely to have more water violations and a greater cumulative risk for cancer (Uche et al., 2021). The American Water Works Association found that rural areas with smaller populations are more vulnerable to water violations because there is a greater financial burden to replace these systems (Bae, 2021).

Intial Hypotheses

This research project will investigate if for each type of water source there is an association between the size of the population served and the number of violations for each observation, and if there is a trend between state or region and the number of violations. Researchers hypothesize that smaller populations will have a greater number of water violations, and that more rural states and regions will have a greater number of water violations. Finally, researchers will investigate how population demographics affect access to clean water. The measure of access to clean water will be quantified by the number of violations each region has adjusted for population. It is hypothesized that states with lower median income will have more violations. It is hypothesized that states with higher populations of non-white residents will have more violations.

Data Preparation

First, all variables had to be renamed to remove spaces and symbols from original variable names. In preparation for data analysis, variables such as PWS Type, EPA Region, and Primacy Agency had to be changed from characters to factors. By converting these variables into factors researchers were able to more easily identify and order the variables. Researchers also found that the variable of Cities Served was missing 74,056 observations. Since a significant amount of data was missing, the researchers will not be analyzing the data by the city since a large proportion of data would be excluded. A significant finding was that no data was missing in variables pertaining to the number of facilities, violations, and site visits. The tidycensus data had to be extracted from the package and sorted. Census data was extracted from the United States by state, based on data from the 2019 ACS. The variable names were initially in the same column and during cleaning the data was pivoted.

Exploratory Data Analysis

Hypothesis One: Association between the size of the population served and the number of violations for each type of water source.

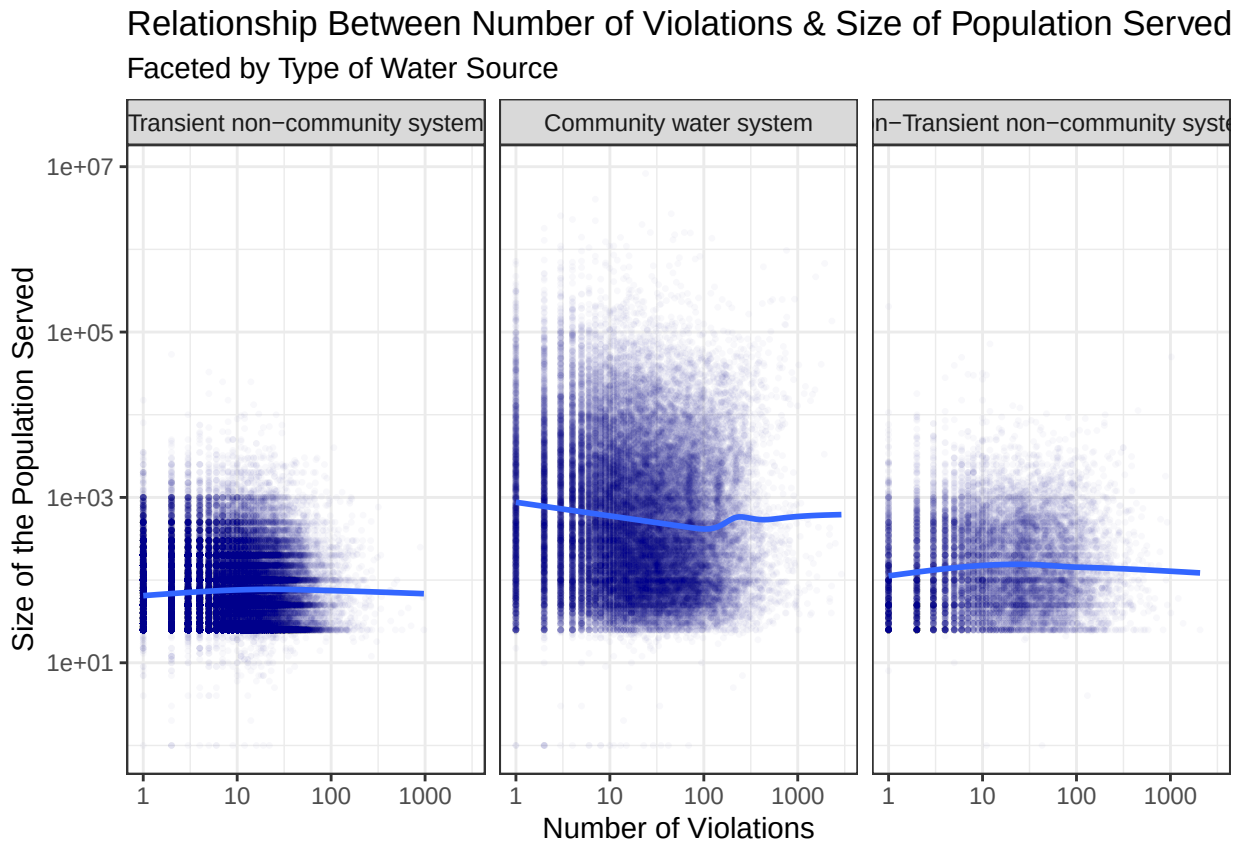
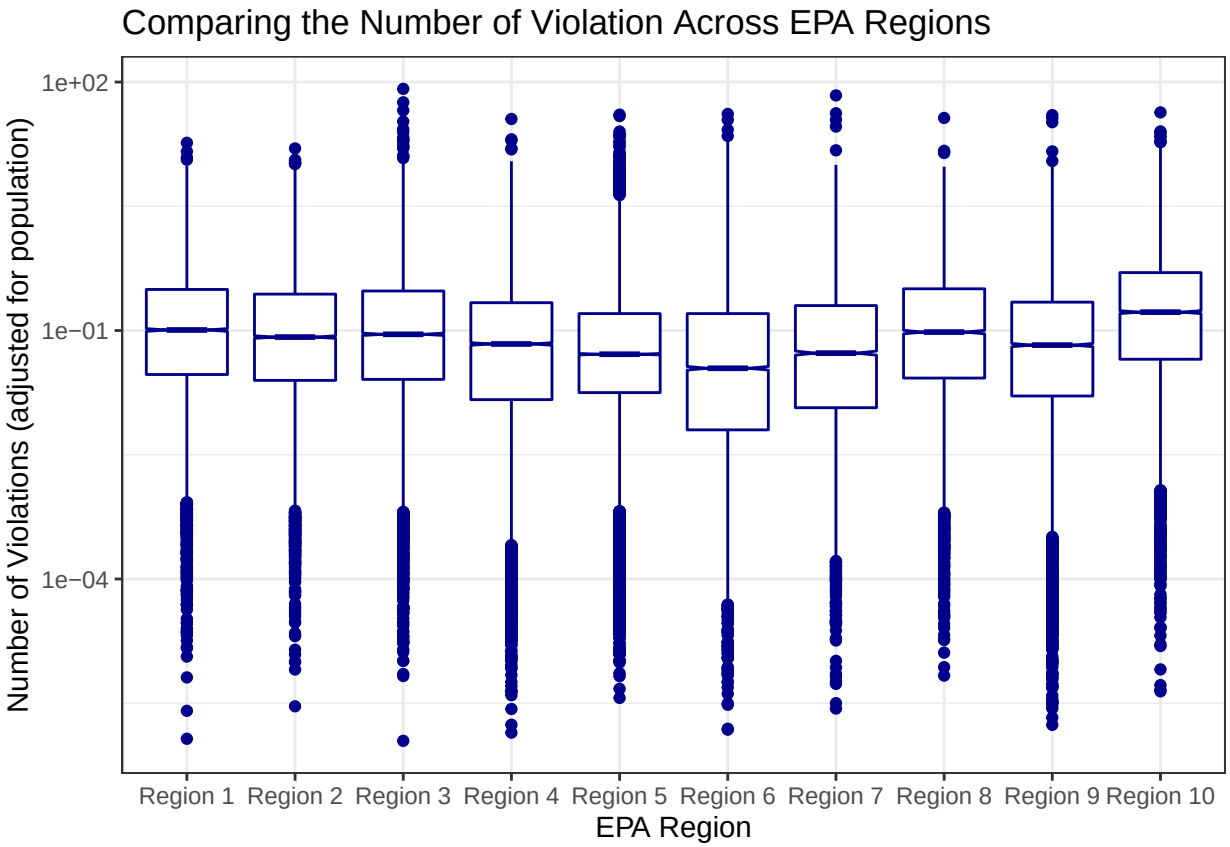


Figure 1 shows the size of the population and the number of violations for each type of water source. The blue trend line shows that there is no evidence of an association between the size of the population served and the number of violations as each line is relatively flat. The graphs show that community water systems serve larger populations while non-transient and non-transient and non-community water sources serve smaller populations. A noticeable feature in each of the three graphs is that each type of water source reported only one violation. A possible explanation could be a lack of reporting and infrastructure. Based on the graph the hypothesis that smaller populations will have more violations is not supported.

Hypothesis Two: Trend between state and region and the number of violations



As discussed above the initial hypothesis was to analyze the trend between state/region and the number of violations. The researchers used the EPA region as a categorical variable to analyze the trend of the number of violations across the 10 regions. In creating Figure 2 all regions have a large number of outliers and a huge range. By creating a log scale on the y-axis the researchers could look closely at the boxplot for each region. The researchers also adjusted the number of violations for the size of the population in each region. This was calculated by dividing the number of violations per water source by the number of people served per water source for each region. A closer analysis of states within regions five, six, and ten was conducted since region ten has the highest median number of violations and regions five and six have the lowest median number of violations. Regions five and six also have the largest population size. Based on the graphs above there was no trend found between EPA region and the number of violations.

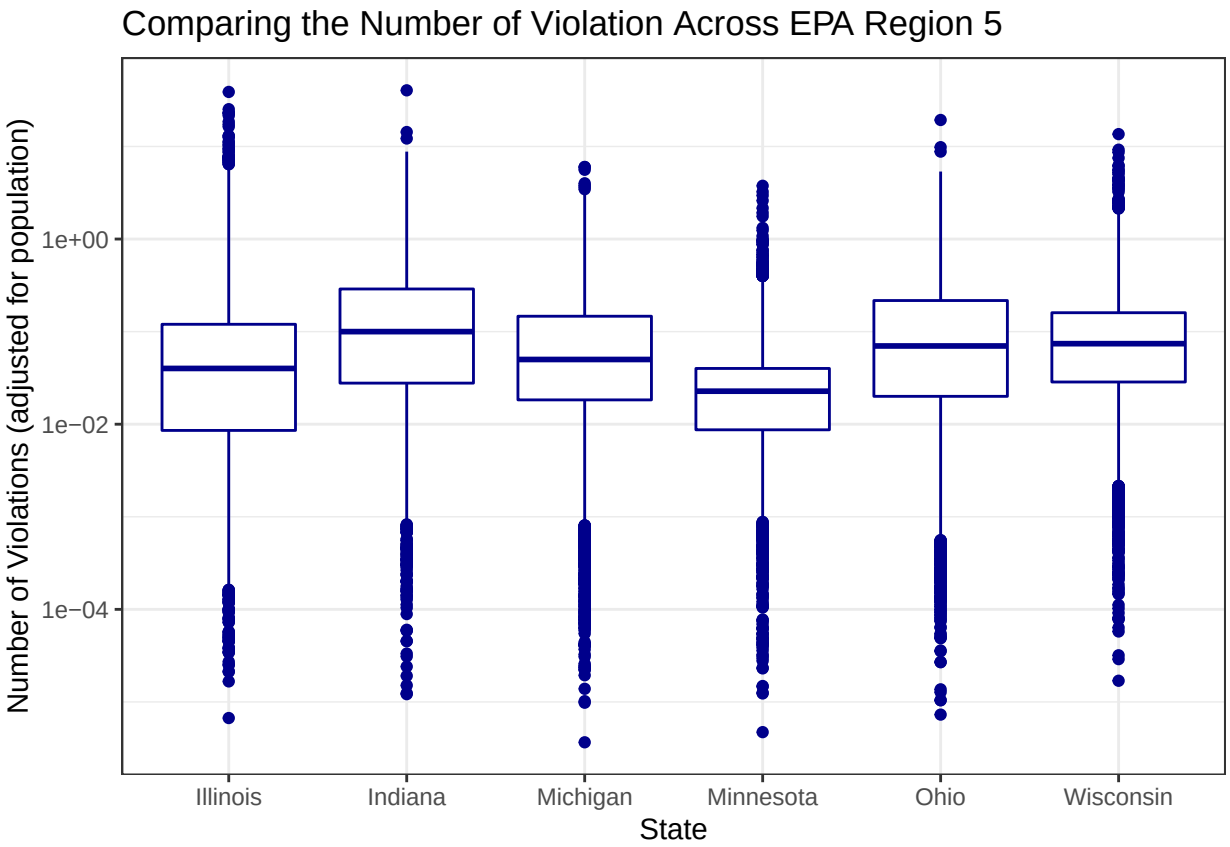


Figure 3 shows individual boxplots for each of the six states included in region five. States in region five include Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin. All six states are located in the Midwest and all have a large number of outliers. The state with the lowest median number of violations is Minnesota while Indiana has the highest.

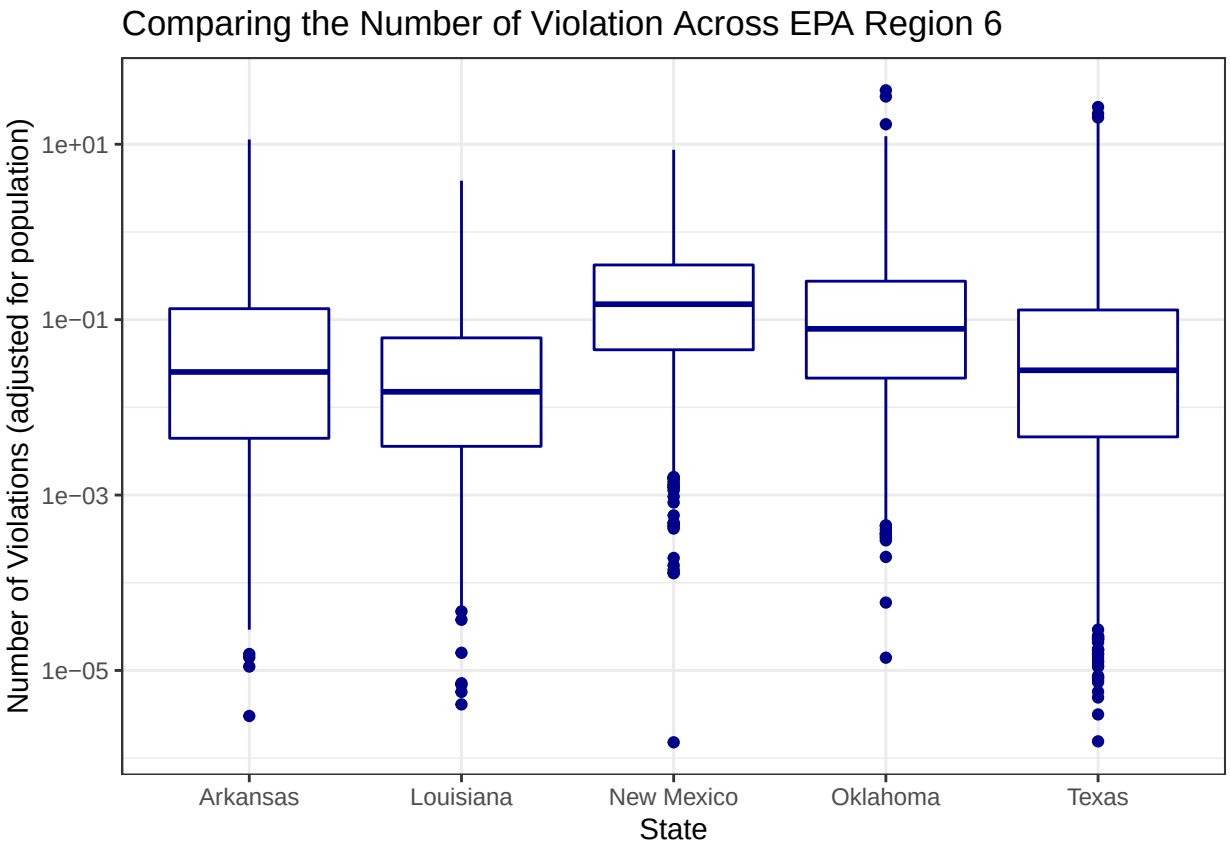


Figure 4 above displays the states within region six that have one of the highest populations sizes and one of the lowest median number of violations across all regions. States within region six include Arkansas, Louisiana, New Mexico, Oklahoma, and Texas. These southern states look relatively similar as all have multiple outliers. New Mexico has the highest median number of violations and Louisiana has the lowest.

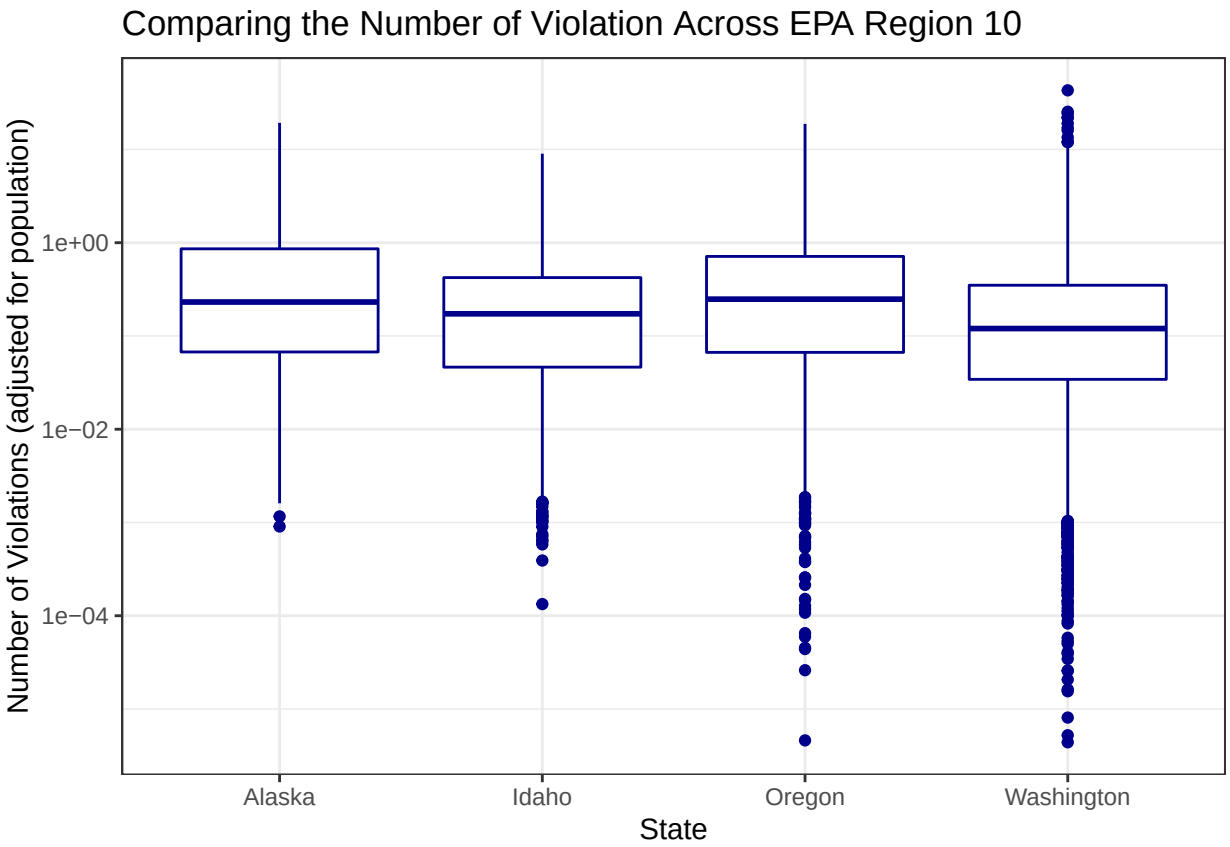


Figure 5 shows the four individual states that makeup region ten, that has the highest median number of violations across all regions. States in region 10 are Alaska, Idaho, Oregon, and Washington. Compared to states within region five, states within region ten have far fewer outliers. Washington has the most outliers while Alaska and Idaho have very few. This could be accounted for the number of water systems that operate in Alaska and Idaho compared to Washington.

Hypothesis Three: Trend between income and the number of violations

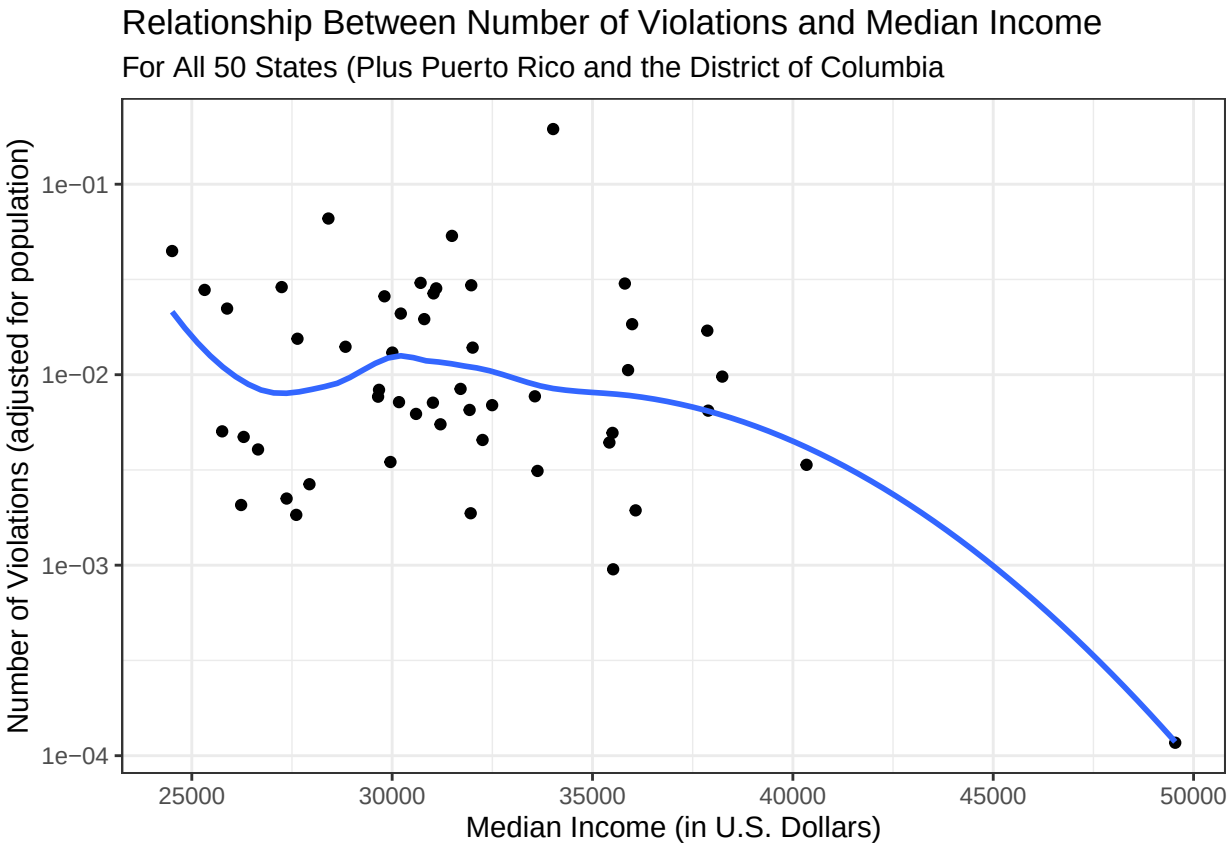


Figure 6 shows the relationship between the median number of violations adjusted for population by state and the median income for that state. There is a strong negative correlation between the variables which shows that as median income increases, the number of violations decreases. This supports the initial hypothesis that areas with lower income will have more violations.

Hypothesis Four: Trend between population of non-white residents and number of violations

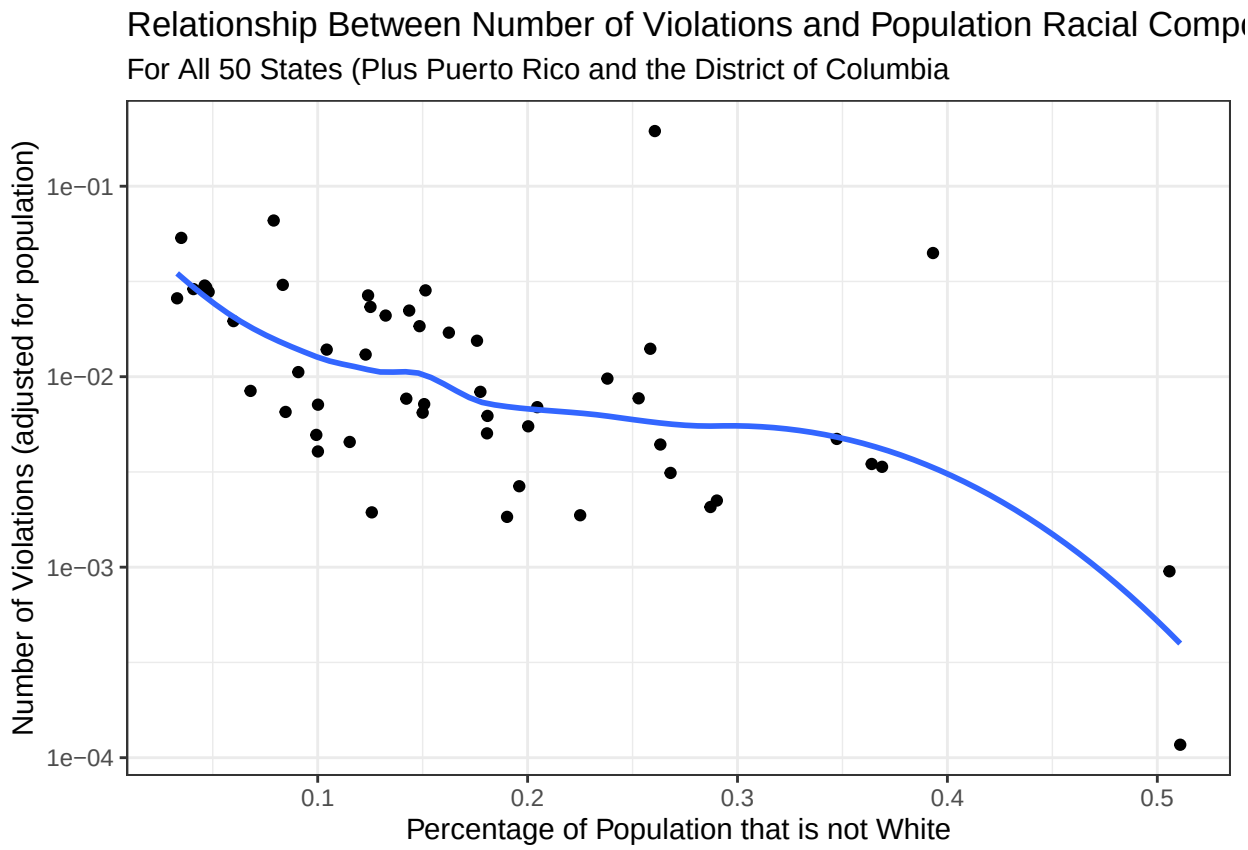


Figure 7 shows the relationship between the number of violations adjusted for population per state and the percent of the population that is not white. There is a negative correlation between the two. This does not support the initial hypothesis that areas with a higher percentage of non-white residents will have more violations. This could possibly be because the data is too broad. A state may have more non-white residents overall, but those non-white communities may experience more water violations within that state. To explore this possibility, the new hypothesis is that if the researchers look at county level in a state with a high percentage of the population that is not white, there will be a positive correlation between counties with a high percentage of a non-white population and number of water violations.

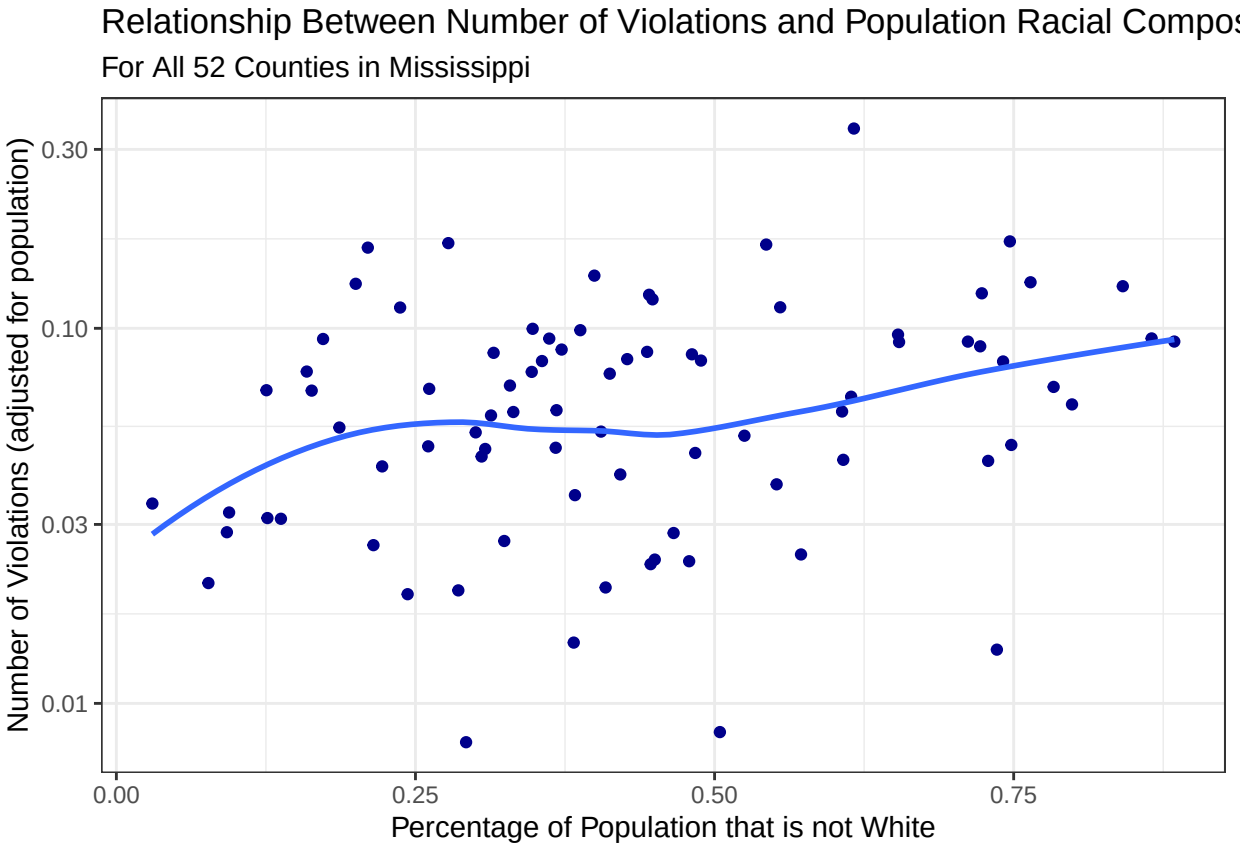


Figure 8 shows the relationship between the number of violations by county adjusted for population and the percent of the population that is not white in Mississippi. There is a positive correlation between the variables. The researchers chose Mississippi as one of states with the highest percentage of non-white residents. This supports the adjusted hypothesis.

Summary

The study supported some of the initial hypotheses. The hypothesis that states with lower median incomes had more water violations was supported by the data. The hypothesis that states with higher non-white populations had more water violations was initially not supported, but upon examining data at a county level, researchers found support for the claim that regions with greater non-white populations had less access to clean water. These hypotheses were also supported by the overall literature on clean water, which suggests that non-white and low-income communities are vulnerable populations. Future water crises may be predicted based on demographic makeup on a county level.

The researchers were unable to find support for the claim that smaller populations had more water violations. Further, the researchers were unable to determine a trend between region or state and number of water violations, and could not determine that rural regions experienced greater water violations. This could be because water violations can be under reported, as in the case of Flint (Forrer et al., 2019). This could also be because the scope of the research was too broad. Future researchers should examine data at a county level or smaller to determine whether there is a causal relationship between population and water violations in the regions with the greatest and the smallest number of water violations.

There may also have been no visible trend because the infrastructure of the United States is aging. Research suggests that the water infrastructure is failing because most of it was built in the 1890s, the 1920s, and the 1940s and it has reached the expected life span (Bae, 2021). It is possible that there was no observable

trend because there are infrastructure issues in every region. Future research should more closely examine when the infrastructure was last built by region to see if there is an association. Future research should also examine whether there is a correlation between older infrastructure and population size.

Appendix A: References

Bae, J. (2021). Clean Water for All: Examining Safe Drinking Water Act Violations of Water Systems and Community Characteristics. ProQuest Dissertations Publishing.

EPA. (2015, November 23). Safe Drinking Water Information System (SDWIS) Federal Reporting Services [Data and Tools]. United States Environmental Protection Agency. https://www.epa.gov/ground-water-and-drinking-water/safe-drinking-water-information-system-sdwis-federal-reporting

Fendick, K. P., Taylor, S., Roberts, M. (2019). Watered Down Justice. Natural Resources Defense Council. https://www.nrdc.org/sites/default/files/watered-down-justice-report.pdf

Forrer, D., McKenzie, K., Milano, T., Davada, S., McSheehy, M. G. O., Harrington, F., Breakenridge, D., Hill, S. W., & Anderson, E. D. (2019). Water Crisis In Flint Michigan – A Case Study. Journal of Business Case Studies, 15(1), 29–44. https://doi.org/10.19030/jbcs.v15i1.10282

Uche, U. I., Evans, S., Rundquist, S., Campbell, C., & Naidenko, O. V. (2021). Community-level analysis of drinking water data highlights the importance of drinking water metrics for the state, federal environmental health justice priorities in the united states. International Journal of Environmental Research and Public Health, 18(19), 10401–. https://doi.org/10.3390/ijerph181910401