Water Quality Disparity in California

Storytelling with data: Identifying the socioeconomic status of the 1 million Californian lacking clean water

Victor Ding, Shin Kim, Claire Zhang, Zehui (Joyce) Zhou

Advised by Professor David Culler (EECS, UC Berkeley)

1 Executive Summary

As of early 2019, about one million people in California are being served water by public water systems with active water quality violations. The violations are not uniformly distributed, with a notable geographic concentration in the San Joaquin Valley, suggesting the existence of a disparity in the access to clean water. With the backdrop of the recent Open and Transparent Water Data Act (AB 1755) in California to make better use of water data and the 2018 Safe Drinking Water Data Challenge, our team analyzes the water compliance and quality datasets published by the California Water Boards in relation to socioeconomic datasets. At the individual public water system (PWS) level we find a concerning correlation between the low-income of a community and the number of water quality violations incurred by the PWS serving that community. We also find a correlation with longer total violation duration as well as a correlation with a higher repeat rate of the same violation in the next year.

2 Introduction and Motivation

2.1 Why water in California? Why data science?

Water is an important issue in California. Due to limited rainfall during the dry season, water is a naturally limited resource. Ongoing climate change has intensified this problem, with the state experiencing a severe drought in 2011-2017 and the period between 2011-2014 being the driest period in California's history since record-keeping.

The importance of safe drinking water is also underlined by government actions and legislation. In 2012, California Governor Jerry Brown signed Assembly Bill (AB) 685 to make California the first state in the United States that recognized water as a human right [1]. Despite this emphasis on water, water quality problems have persisted. As recently as February 2019, during the State of the State Address, current California Governor Gavin Newsom called it a "moral disgrace and a medical emergency" that over one million Californians lack access to clean water in California today [2].

In addressing these myriad water-related issues, California has turned to open data as a strategy to improve the efficiency of water operators and raise public awareness of water issues. In 2016, California passed the Open and Transparent Water Data Act (AB 1755) to create a statewide integrated water data platform and develop protocols for management and utilization of water data [3]. The state is currently in the process of implementing this bill.

With this context, the 2018 Safe Drinking Water Data Challenge was held with support from numerous state agencies including the California Water Boards. The challenge encouraged community members and data science practitioners to explore the potential of numerous water-related datasets covering broad areas such as safety, reliability, and affordability of drinking water. Despite the challenge having concluded in October 2018, our capstone project team decided to take on the challenge in spirit. Our aim was to utilize

the provided datasets to help envision what a data-driven insight extraction and storytelling process may look like.

2.2 Prior studies have looked at water quality disparity

There are many related studies that examined the factors related to water quality problems. Allaire, Wu, and Lall (2018) studied the temporal and spatial patterns of the health-related water quality violations of 17,900 community water systems and found that more violations happen in rural areas than urban areas and in publicly-owned community water systems than privately-owned ones [4]. They also suggested underperforming water systems could potentially improve their performance with assistance. In addition, VanDerslice (2011) emphasized the lack of data characterizing the water infrastructure to systematically examine disparities in drinking water infrastructure, which plays an essential role of serving 99% of the United States population [5]. Such relevant literature encouraged us to explore the potential of data science in helping to address the disparity between different social groups and identify possible suggestions and solutions.

2.3 Our datasets are from the Census Bureau and the State Water Resources Control Board

We used the American Community Survey (ACS) as our main data source for the socioeconomic data in California. Every year, the Census Bureau collects information on jobs and occupations, educational attainment, veterans, whether people own or rent their homes, and other topics among over 3.5 million households in the US and present the information in the ACS 1-year estimate dataset. The Census Bureau will release the ACS 5-year estimate dataset based on the data in the past five years. In both datasets, the census tract is the smallest geographical unit for the census. Because ACS 1-year estimate only includes data for census tract with population more than 65,000, for census tract with population more than 65,000, we use the latest ACS 1-year estimate in 2017 for the sake of currency and for census tract with population less than 65,000, we use the latest ACS 5-year estimate in 2012-2017 for the completion of the data [6].

We also used the dataset from the State Water Resources Control Board. According to AB 1775, the state aims to maintain an integrated data platform, where all water-related data will be collected. Every year, water utility units need to submit annual reports on the water usage data, and quality monitoring data will also be published. The water usage data includes the amount of water delivered for different end users and the quality monitoring data includes the measurement of different substances in water.

3 Key Background Concepts

3.1 Data storytelling: a visual approach to a data-based story

We present our project using data storytelling. Data storytelling is a visual approach to telling a data-based story, which pinpoints the message via data plots and tables. Data storytelling uses many kinds

of plots and narrative tools to guide the audience towards understanding the material presented based on strong data and relevant analysis [7].

3.2 Public water system (PWS): the fundamental units for water supply

Public Water Systems are the fundamental units for water supply. In general, it works in 3 steps: the water system gets water from upstream sources where the water may have been contaminated by pollutants such as agricultural fertilizer. Then, the water will be processed via filtration, purification, and disinfection. Finally, the clean water is served to the consumer through pipes or other constructed conveyances. There are about 3000 PWS's in California. PWS's vary in sizes, and a large PWS can serve millions of people as the Los Angeles Department of Water and Power does while small ones only serve tens to hundreds of people. The largest 10% of the PWS's serve 90% of the population in California.

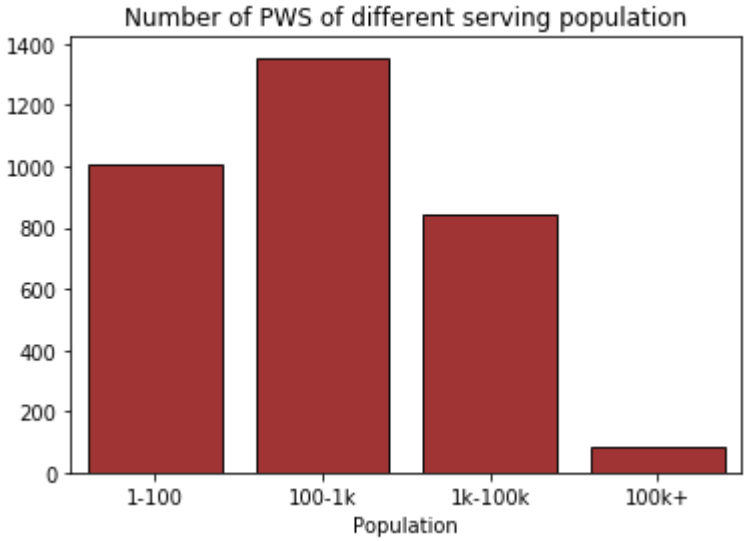


Fig 3.2. PWS's vary in sizes and many PWS's are small.

PWS's can either be publicly or privately owned. In California, 82% of the population is served by PWS's that are publicly owned. PWS's can also use surface water (e.g. from lakes, rivers, and reservoirs) or groundwater (e.g. wells) as their primary source of water. In California, 82% of the population is served by PWS's that use surface water as their primary source of water.

3.3 Maximum contaminant level (MCL): the regulatory threshold of substances

Maximum contaminant level represents the regulatory threshold that the concentration of any specific chemical substance should not exceed. This maximum value is usually established nationally by the United States Environmental Protection Agency (EPA) with the effect on public health in mind [8]. In California, the State Water Resources Control Board further set maximum contaminant levels for different contaminants, which are usually the same or more strict than the national maximum contaminant levels.

The MCL can change over time as more evidence on the harmful health impact of a contaminant is established. For example, the MCL for arsenic in California was reduced fivefold from 0.050 mg/L to 0.010mg/L in 2008 [9].

Contaminant Name	**Maximum Contaminant Level**
Arsenic	0.010 mg/L
Uranium	20 pCi/L
Nitrate	10 mg/L
TTHM	0.080 mg/L
HAA5	0.060 mg/L

Table 3.3. Five Common Contaminants in Water and their Maximum Contaminant Level

3.4 Water quality violations ("Violations"): violation issued to the PWS

There are two major types of water quality violations: exceeding the maximum contaminant level, and the flawed water treatment. The former violation means the concentration of a contaminant at a water system has exceeded the MCL. More than 91% of all the contamination events in California from 2012 to 2018 involve the exceeding maximum contaminant level. To give a clear example of the contaminants that frequently triggers contamination events, nitrate levels frequently exceed the MCL in the Central Valley region, which is an agricultural region with frequent usage of nitrogen-based fertilizers. The second type of violation occurs when water is contaminated during the process of water treatment itself. In fact, the most common cause for the second type of violation is the by-product generated during the treatment process. These by-products, if not carefully removed from the body of water, might have complex chemical reactions with other contaminants, which produce harmful materials and pollute the water.

4 The Story: Identifying the Socioeconomic Status of the 1 million Californians Lacking Clean Water

4.1 18% of Californians have been exposed to drinking water violations in the past 7 years

Water System Name	Population	County	City
Los Angeles-City, Dept. Of Water & Power	4,072,307	Los Angeles	Los Angeles
San Gabriel Valley Water Co.-El Monte	257,500	Los Angeles	El Monte
Modesto, City of	214,181	Stanislaus	Modesto
City Of Stockton	175,530	San Joaquin	Stockton
City Of Sunnyvale	147,055	Santa Clara	Sunnyvale

Table 4.1. (1). Top 5 PWS's with Water Quality Violation ranked by Population, 2012-2018

Water System Name	Population	County	City	Number of Violations
Pappas & Co (Coalinga)	25	Fresno	Fresno	105
Edgewater Mobile Home Park	40	Sacramento	Rio Vista	69
Keeler Community Service District	50	Inyo	Keeler	69
Encinal Rd Ws #01	41	Monterey	Salinas	67
Hat Creek Water Company, LLC	50	Shasta	Old Station	65

Table 4.1. (2). Small PWS's with a large number of Water Quality Violations, 2012-2018

In 2012-2018, 18% of the population in California has been exposed to drinking water violation at least once. The water quality problem is universal among big and small PWS's. Table 4.1.(1) shows the PWS's with problems in 2012-2018, ranked by the population the PWS is serving. It is noticeable that the PWS that serves over four million population in Los Angeles City also was also issued water quality violation in the past seven years. On the other hand, PWS's that serve small population could possibly be issued multiple water quality violations in the past seven years, as shown in Table 4.1.(2).

Although the water quality problem is universal among different sizes of PWS's, the type of contaminants that cause the problems is highly concentrated. The most frequent chemicals that have triggered violations are arsenic, nitrate, and total trihalomethanes (TTHM). Table 4.1.(3) shows the number of violations associated with each contaminant and the population affected by that contaminant in California during 2012-2018. The high number of violations suggests that the top three chemicals are either very common in California or hard to treat for the PWS's.

Contaminant	Total Population Affected	Number of Violations
Arsenic	417,484	172
Combined Nitrate	395,873	88
TTHM	970,172	87
1,2,3-Trichloropropane	276,047	57
SWTR	555,615	43
Combined Uranium	62,811	39
Total Haloacetic Acids (HAA5)	134,108	33
Fluoride	15,406	12
Turbidity	5,417	12

Table 4.1. (3). Type of contaminant, the total population affected and number of violation in California, 2012-2018

4.2 53% of the water quality violations happen in San Joaquin Valley in the past 7 years

Water quality violations in California are geographically concentrated in San Joaquin Valley, a region within the Central Valley of California. In 2018, around 100,000 people in the Central Valley did not have access to clean water [10], and both arsenic and nitrates are known issues in the Central Valley [11]. Furthermore, San Joaquin Valleys have the many PWS's with violations [12].

In Figure 4.2, a plot of water quality violations in California, we observe the concentration of violations in the San Joaquin Valley. In fact, 53% of the water quality violations in the last 7 years in California have occurred in San Joaquin Valley. In contrast, San Joaquin Valley only represents 17.2% of the land mass and 10.6% of the population in California.

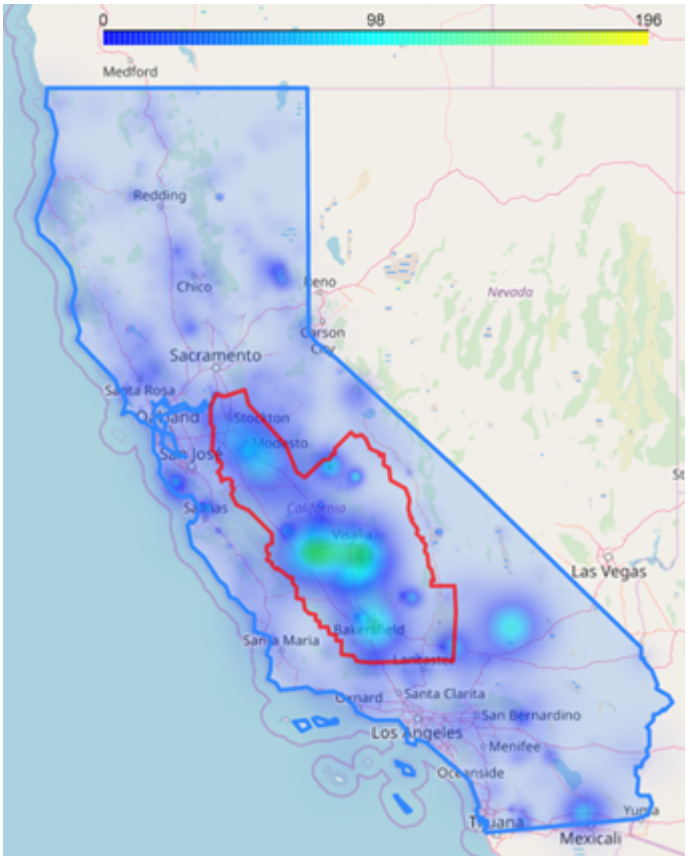


Figure 4.2. Water quality violations in California. Brighter colors indicate more violations happened in that region. The red box denotes the San Joaquin Valley.

On average, each PWS in San Joaquin Valley has experienced 5.15 violations in the last 7 years.

Table 4.2 shows the eight counties in San Joaquin Valley, the average of San Joaquin Valley and California along with the information on the population and the PWS violations summed or averaged over all the PWS's within the regions. From the average number of violations, we observe that the two counties in the San Joaquin Valley with most population – Kern County and Fresno County – suffered 2.5 to 3.5 times as many water quality violations compared to the rest of California.

	Number of PWS with Violation	Total Number of Violations	Avg. Number of Violations per PWS	Avg. Violation Duration (Month)	Population (2017)
Kern	77	1,481	7.96	16.21	878,744
Fresno	58	1,547	10.89	13.88	971,616
Tulare	44	832	6.71	10.89	458,809
Madera	36	751	8.63	13.67	154,440
Stanislaus	27	412	4.96	11.57	535,684
San Joaquin	17	195	1.67	3.18	724,153
Merced	13	116	2.97	4.49	267,390
Kings	9	228	11.40	19.80	150,183
SJ Valley	281	5,562	5.15	8.74	4,141,019
California	3295	10,226	3.10	4.78	38,982,847

Table 4.2. The violation records based on Public Water Systems in counties in San Joaquin Valley, 2012-2018

As Kern County is a county with a large population as well as both a high number of violations and long violation duration in San Joaquin Valley, we further conducted our study into the water condition at Kern County.

4.3 Kern County has suffered from frequent and persistent water quality violations

Kern County is located in the south of the Central Valley. It has four relatively distinctive seasons and its economy is heavily dependent on agriculture and petroleum extraction. Kern County has constantly suffered from water quality issues and undesirable air quality in recent years.

Figure 4.3 shows the average number of violations per PWS and the percentage of incompliant PWS in 2012-2018 in Kern County. Every year, there are around 200 violations in Kern County. There are more than 20% of the PWS's in Kern County being out of compliance each year, which is much higher than the average 8% of all PWS's in California being out of compliance each year. From this we observe that Kern County has suffered from frequent and persistent water quality violations.

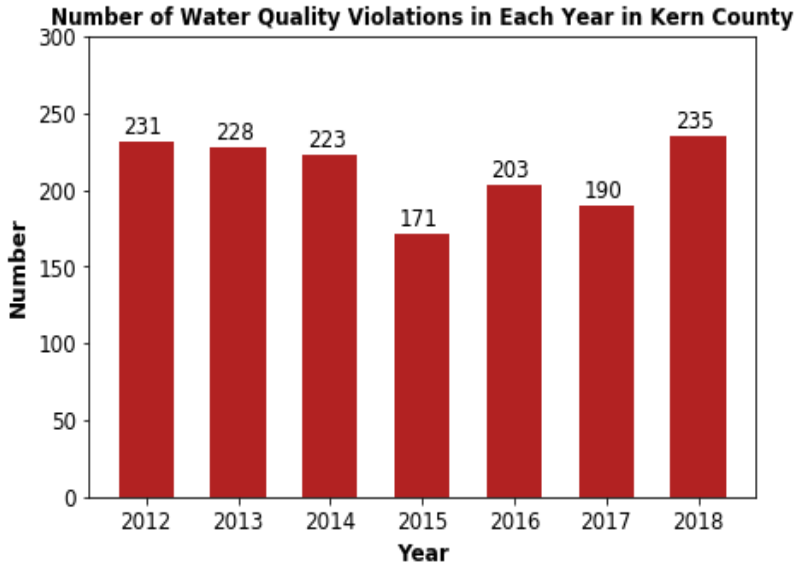


Figure 4.3. (1). Large number of violations happen in Kern County each year

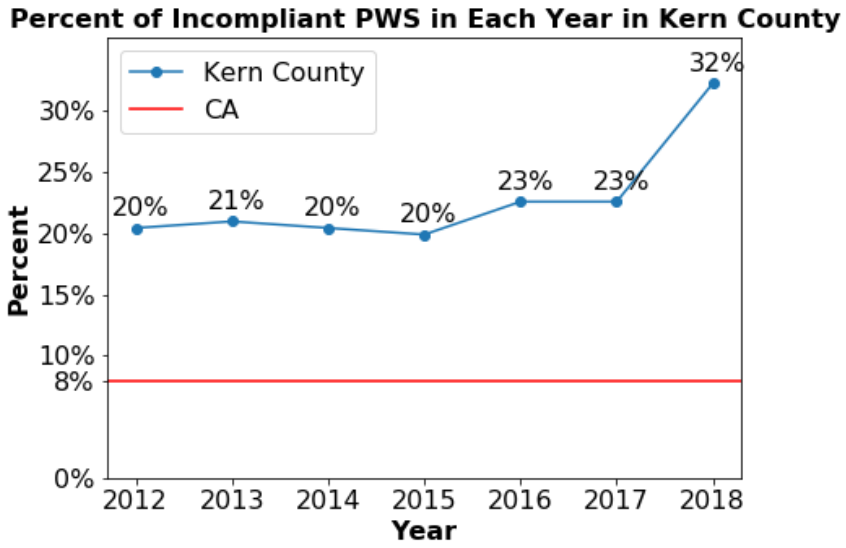


Figure 4.3. (2). Kern County always have a higher percentage of incompliant PWS

4.4 The most common contaminants in Kern County are arsenic and nitrate

The contaminants that cause the water quality problems in Kern County is highly concentrated among top two contaminants, arsenic and nitrate, which made up about 80% (1,197 out of 1,481) of the total number of water quality violations in Kern County in 2012-2018.

Ranked the first among the contamination triggering chemicals, arsenic could occur naturally from rocks and sediments or resulted from human activities, such as mining and use of arsenic-related wood preservative and pesticide. Exposure to a high level of arsenic in drinking water could potentially increase the risk of cancer in the bladder, kidney, lung, and skin [13].

Nitrate in the drinking water mainly comes from fertilizers, septic systems, and manure storage or spreading operations. Exposure to nitrate related contaminants could cause methemoglobinemia, or "blue baby" disease and also negatively affect one's digestion system and the ability of red blood cells to carry oxygen [14].

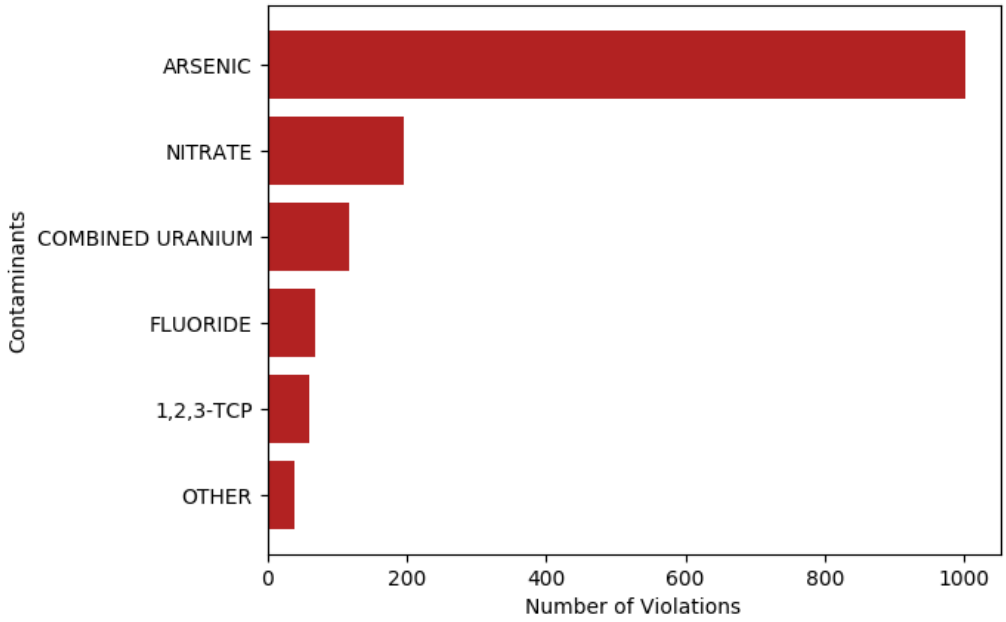


Figure 4.4. Number of violations related to different contaminants in Kern County, 2012-2018

4.5 Kern County has a struggling economy

As the paper "Socioeconomic Disparities and Air Pollution Exposure: A Global Review" points out [15], there are some potential relationships between low socioeconomic status and air quality issues, so we also want to understand the environmental issues with socioeconomic inequality and disparity. Thus, we decide to analyze the economic status of Kern County.

According to Table 4.5, Kern County lags behind when compared to California in the following three metrics. Kern County has a higher poverty rate and lower household median income, and a higher unemployment rate than those of California.

	Poverty Rate	**Household Median Income**	**Unemployment Rate**
Kern County	22.6%	$50,826	9.1%
California	15.1%	$67,169	4.8%

Table 4.5. Poverty rate, household median income and Unemployment Rate in Kern County vs. in California, 2017

The observation of the poor economic shape and water quality in Kern County leads us to ask the following question: is the co-occurrence of low-income and high water contaminations purely a coincidence? Or is this a prevalent trend in California?

4.6 A correlation exists between low-income and poor water quality across California

For the analyses in this subsection, socioeconomic features (e.g. median income) were assigned to PWS's based on those of the cities that they serve. We devised a "mapping algorithm" to handle cases where there isn't a one-to-one relationship between the cities and the PWS's. This algorithm is explained in further detail in the "Methodology" section of this paper.

Many small PWS's did not have boundary shapefiles available and thus did not get socioeconomic features assigned to them. Given their smaller size and significance, we dropped these PWS's from the dataset. We also dropped all PWS's serving less than 1,000 population given the inherent noise in their features due to the small population they cover. This preprocessing yielded 880 PWS's that we used for the analysis below.

4.6.1 Low-income PWS's have a higher chance of being exposed to a large number of violations

Figure 4.6.1 shows the percentage of PWS's with more than 10 violations from 2012 to 2018 serving different income groups. We found that among low-income PWS's ($0-50K median income), 11.2% had more than 10 water quality violations during the last 7 years. This is substantially higher than for middle-income PWS's ($50-100K median income) or high-income PWS's ($100K+ median income) among which only 4.2% and 1.8% had more than 10 violations. Our findings show that low-income PWS's have 6.3x the likelihood of high-income PWS's of having more than 10 violations.

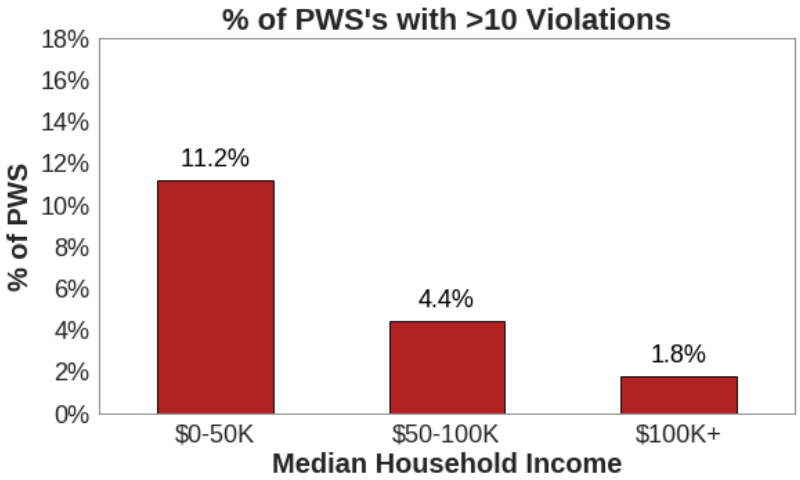


Figure 4.6.1. Percentage of PWS's with >10 violations for each income group, 2012-2018.

Median Income	Number of PWS's with...			Total
	0 Violation	1-10 Violations	11+ Violations	
$0-50K	137	30	21	188
$50-100K	442	54	23	519
$100K+	147	20	3	170

Table 4.6.1. Number of PWS's with 0, 1-10, or 11+ violations for each income group, 2012-2018.

4.6.2 Low-income PWS's have a higher chance of being exposed to long violation duration

Figure 4.6.2 shows the percentage of PWS's with violations duration longer than 12 months serving different income groups from 2012 to 2018. Similarly, we found that among low-income PWS's, 13.8% had a total violation duration exceeding 12 months during the last 7 years. This is also substantially higher than 4.8% and 2.4% for medium-income and high-income PWS's. Our findings show that low-income PWS's have 5.9x the likelihood of high-income PWS's of the total violation duration exceeding 12 months.

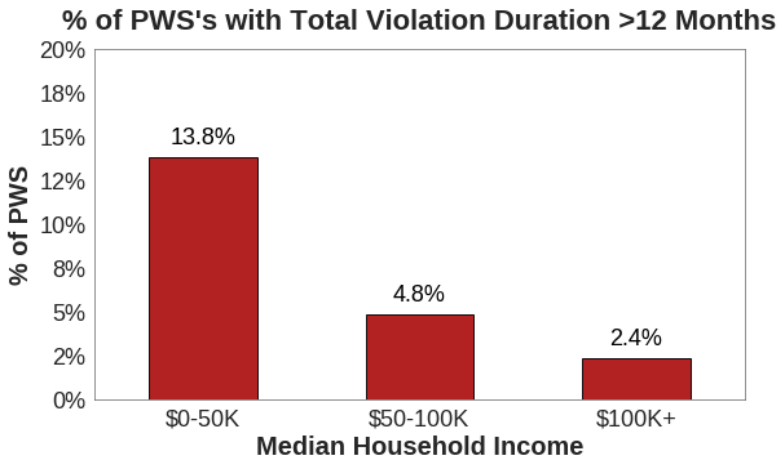


Figure 4.6.2. Percentage of PWS's with long violations duration serving different income groups, 2012-2018.

Median Income	Number of PWS's with violation duration of...			Total
	0 Months	1-12 Months	12+ Months	
$0-50K	137	25	26	188
$50-100K	442	52	25	519
$100K+	147	19	4	170

Table 4.6.2. Number of PWS's with 0, 1-12, or 12+ months of total violation duration for each income group, 2012-2018.

4.6.3 Low-income PWS's have a higher chance of the same violation being repeated next year

Lastly, we looked at the violation repeat rate for PWS's at each median household income level. We defined the "repeat rate" as the proportion of violations for which the same violation is repeated in the next calendar year. Figure 4.6.3 shows the repeat rate of the PWS's serving different income groups from 2012 to 2017. We interpret the repeat rate as the likelihood a water quality problem is not fixed during that year and is carried over to the next year.

We found that repeat rates were significantly higher for low-income PWS's, which were 68.9%, than for high-income PWS's whose repeat rates were 22.9%. Water quality violations being carried over into the next year for low-income PWS's was 3.0x that of high-income PWS's.

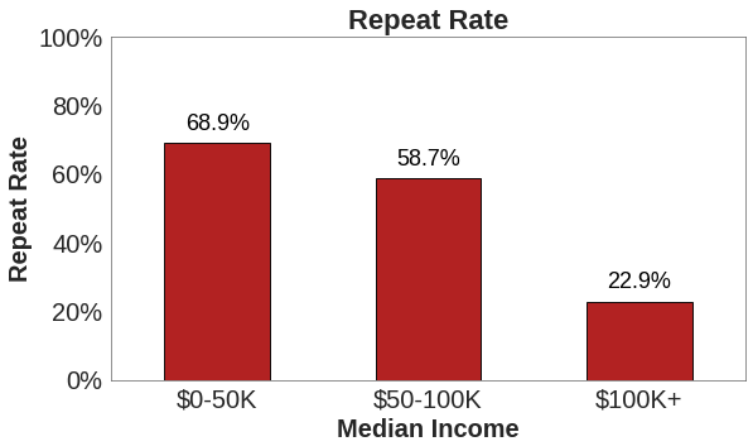


Figure 4.6.3. Repeat rate for each income group, 2012-2018.

Median Income	All Violations*	Repeated Violations	Repeat Rate
$0-50K	148	102	68.9%
$50-100K	172	101	58.7%
$100K+	35	8	22.9%

* Counts unique violation for each PWS in each year (e.g. repeated violations in same year not counted)

Table 4.6.3. Repeated violations and repeat rate for each income group, 2012-2017.

4.6.4 Hypothesis testing for our findings show statistical significance

We conducted hypothesis testing using permutation tests to ensure the statistical significance of our findings above. In each iteration of a permutation test, low-income and high-income labels of PWS's are randomly shuffled (permuted) and the test statistic is recalculated. After iterating a large number of times, the resulting distribution of the generated test statistic is compared against the observed test statistic to obtain a p-value.

The p-value represents the likelihood of observing an outcome more extreme than our observed outcome. We were able to reject all three null hypotheses each with p-values < 0.01.

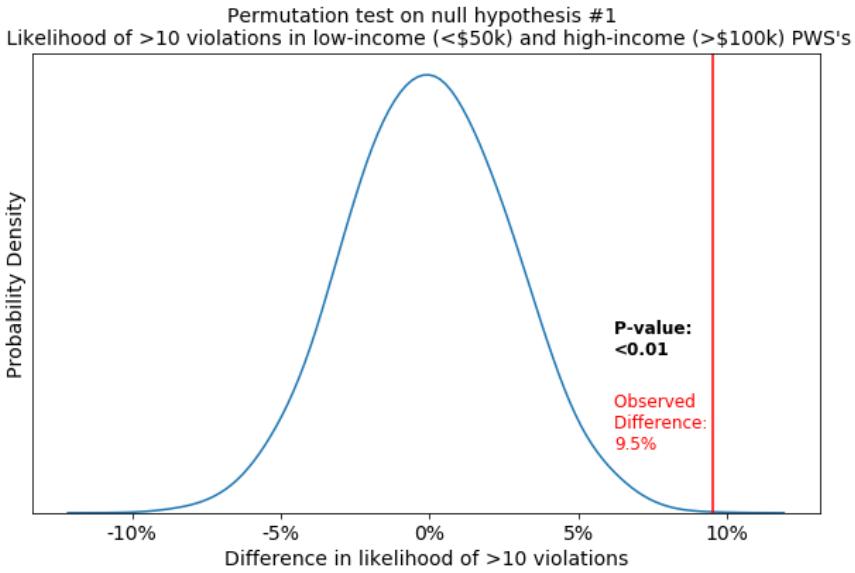


Figure 4.6.4. (1)

*Null hypothesis 1: The chance of a PWS being issued >10 violations in the last 7 years is **equal** for low-income PWS's ($0-50K median income) and high-income PWS's ($100K+ median income)*

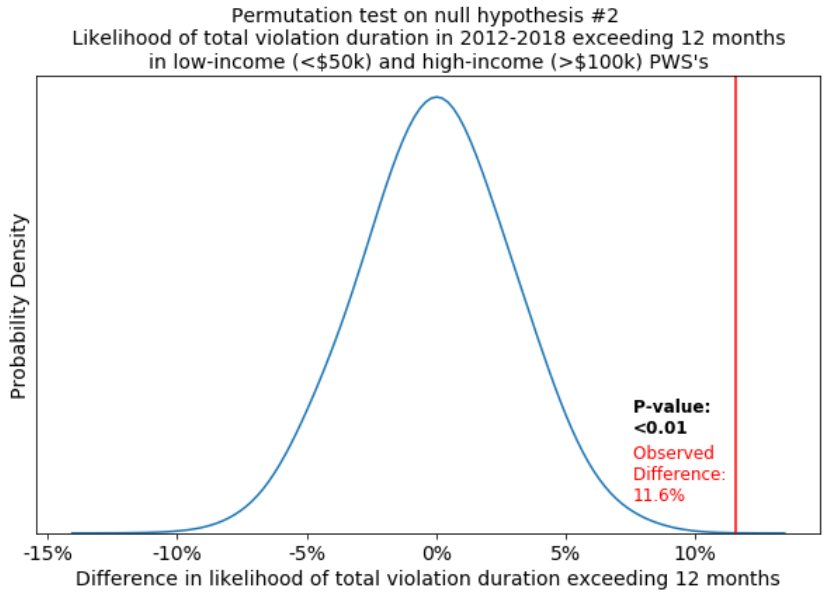


Figure 4.6.4. (1)

*Null hypothesis 2: The chance of a PWS's total violation duration in the last 7 years exceeding 12 months is **equal** for low-income PWS's ($0-50K median income) and high-income PWS's ($100K+ median income)*

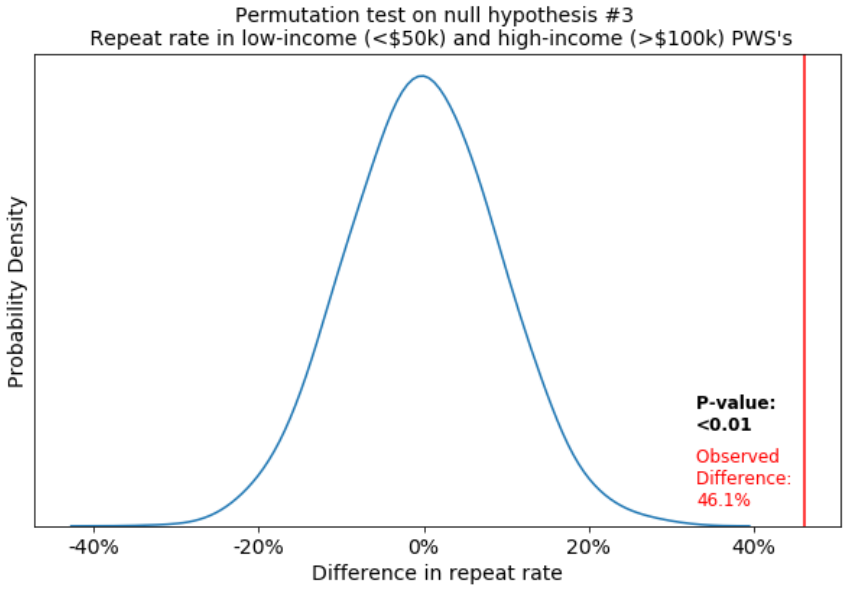


Figure 4.6.4. (1)

*Null hypothesis 3: The chance of a PWS's violation re-occurring in next year is **equal** for low-income PWS's ($0-50K median income) and high-income PWS's ($100K+ median income)*

4.7 A case study on a low-income PWS with effective water treatment

In order to study the water condition faced by those PWS's which serve a socioeconomically vulnerable population, we conducted a case study of 4 PWS's serving 4 communities in San Joaquin Valley: Delano, Allensworth, McFarland, and Kern Valley State Prison (KVSP).

4.7.1 Delano, Allensworth, McFarland, and KVSP share similar natural environments, socio-economic status, and water quality.

Delano, Allensworth, and McFarland are three cities at the border of Kern County and Tulare County. KVSP is located at Delano. As shown in Figure 4.7.1, these 4 communities are within close proximity so they are exposed to similar natural environments.

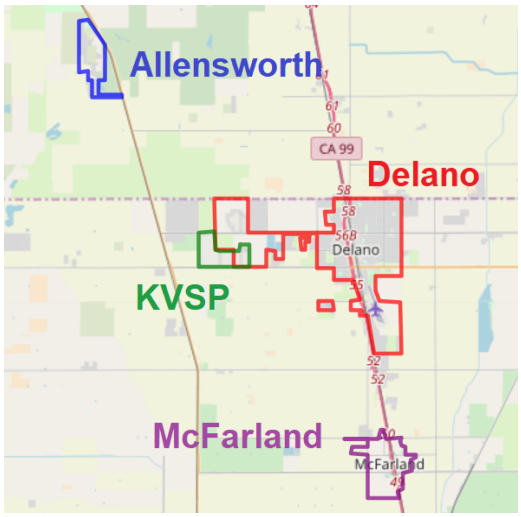


Figure 4.7.1. The map of the 4 communities showing their close proximity.

Furthermore, Delano, Allensworth, and McFarland are all cities with economic statuses below average as we can tell by their low income, high unemployment rate and high poverty rate in Table 4.7.1 (1). Additionally, KVSP serves a population mainly constituted with prisoners, who have even more limited choice on their drinking water.

	Population	Household Median Income	Unemployment Rate	Poverty Rate
Delano	56,632	$38,708	24.7%	24.4%
McFarland	15,105	$35,069	11.9%	35.9%
Allensworth*	512	$31,042	N/A	44.6%
KVSP	5,300	N/A	N/A	N/A
California	38,000,000	$67,169	4.8%	15.1%

* Unemployment Rate in Allensworth unreported from CA EDD due to small size

Table 4.7.1.(1). The economic status of the 4 communities.

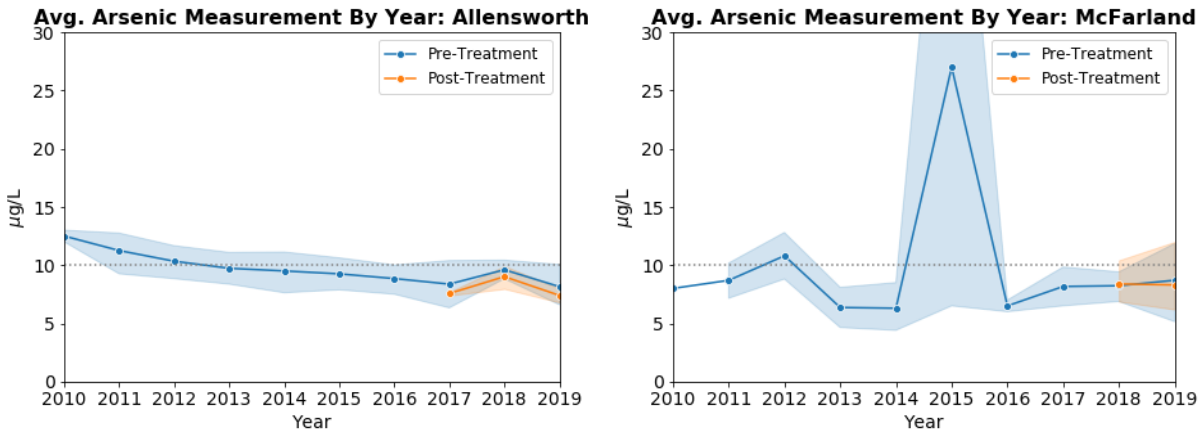
San Joaquin Valley is known for its arsenic contamination, in the past 7 years, all of these 4 PWS's have cumulative violation duration of one year or more.

	Violation Duration (Month)
Delano	18
McFarland	15
Allensworth	24
KVSP	12

Table 4.7.1.(2). The economic status of the 4 communities.

4.7.2 Delano shows effective arsenic treatment while KVSP is mediocre.

All of these 4 PWS's we chose have implemented some arsenic treatments. We studied the arsenic measurements both pre and post treatments for these 4 PWS's.



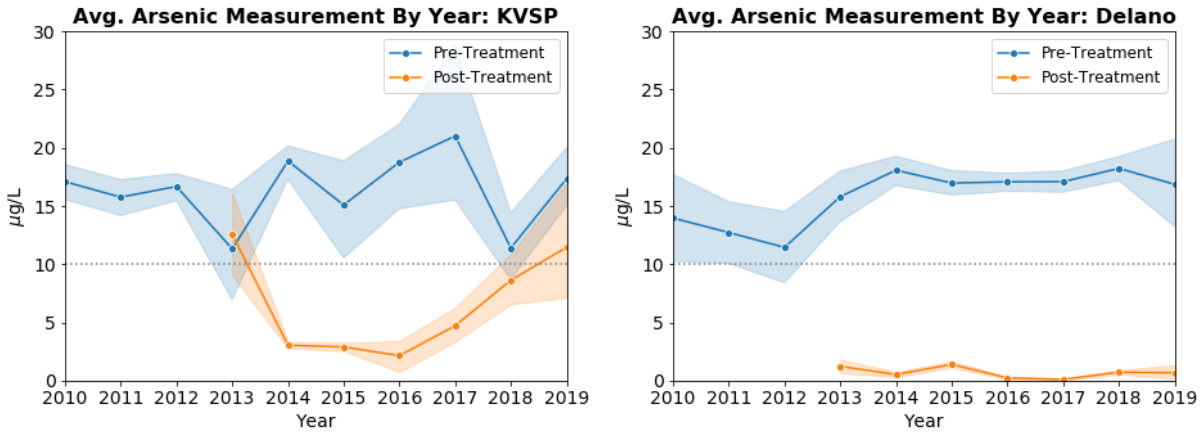


Figure 4.7.2. Arsenic measurement of pre and post treatment for the 4 PWS's. The dotted lines in the graph indicate the MCL for arsenic.

For Allensworth and McFarland, according to the data (Figure 4.7.2), these two PWS's have only started to implement their arsenic treatment recently.

Delano and KVSP have been implementing arsenic treatment for a while. The installation of the arsenic removal plant at KVSP finished between December 2012 and January 2013 [16]. As we can observe in Figure 4.7.2 above, Delano has always been successful in terms of arsenic treatment, as the post-treatment arsenic measure is almost zero after the implementation of treatment. As a comparison, KVSP shows the effective treatment of arsenic in 2014-2016; however, since 2017, KVSP shows an increasing trend in the post-treatment arsenic measurement.

In order to understand the reason behind the difference in the treatment outcome, we suggest further study to be conducted. Maybe Delano was using different treating methods. Maybe the fact that KVSP is a prison lessens the effort put into water treatment. Delano is a good example of a PWS serving socioeconomically vulnerable communities with effective water treatment, and understanding the water condition in Delano helps us alter the harmful water condition faced by other PWS's serving low-income communities.

5 Our Procedures and Methods

5.1 Exploratory data analysis (EDA)

We conducted EDA on our datasets to understand the structure of our datasets, identify needs for data preprocessing (e.g. outliers), performing sanity checks on the consistency of data (e.g. incorrect units), spotting interesting trends, and form preliminary research questions. We mainly relied on data visualization techniques for EDA. Key types of visualizations include geographic maps, scatterplots, distribution plots, temporal line plots, and categorical bar plots.

5.2 Data cleaning and preprocessing

Incorrect and inconsistent units were a persistent problem throughout the dataset as both mg/L and µg/L were used for contaminant measurements. For example, an arsenic measurement of 0.01mg/L and 10µg/L are identical measurements but apparent data entry errors were present in the form of 0.01µg/L and 10mg/L. We corrected for these unit mistakes globally.

The datasets include incorrect MCLs as well. For example, the MCL of arsenic in California was reduced from 50µg/L to 10µg/L in 2008, but many data entries in our dataset still listed arsenic's MCL as 50µg/L even though our dataset starts in 2012. We corrected for incorrect MCLs globally.

When appropriate to do so, we combined contaminant labels into a common contaminant group. For example, nitrate was measured in the dataset under three different contaminant labels – nitrate measured as nitrogen, total Nitrate/nitrite measured as nitrogen, and nitrate measured as NO3. Since the three labels aim to measure a common contaminant which is nitrate, we combined these contaminant labels as one.

Lastly, the overwhelming majority of the violation records listed the start date and end date of the violation as the first or last day of the month. Given that fact, we designated months (as opposed to days) as the basic counting unit for violation duration. There were a small minority of records where violations did not start or end on the first or last day of the month, where violation durations would not immediately round to whole numbers. In those cases, we matched the dates to the nearest first or last day of the months.

5.3 Water quality features

In order to compare the water quality in different public water systems, we create quantitative features. These are the dependent variables that we will want to regress in regression models later on. The four features are shown in the table below.

Feature Name	**Calculation**[1] $(\forall i \in \{Violations\ of\ a\ specific\ PWS\})$	**Description**
Number of Violations	$\left\lvert\{Violation_i\}\right\rvert$	Count of all violations issued to a PWS since 2012
Number of Violation Types	$\left\lvert\{Type(Violation_i)\}\right\rvert$	Count of all unique violation types issued to a PWS since 2012

[1] | | operator denotes size of set. A set does not allow duplicate elements.

Violation Duration	$\bigcup_{i}^{n} Duration(Violation_i)$	Total period a PWS has been out of compliance since 2012
MCL Multiple	$\frac{1}{n}\sum_{i}^{n} \frac{Measurement_i}{MCL_i}$	Average multiple of MCL. A 2x MCL multiple indicates that a PWS measurement

Table 5.3. 4 water quality features generated through our study

Number of Violations simply counts the total number of violations, which is the most naive feature. Number of Violation Types captures the diversity of the violations issued. One public water system may have repeated violations for a single contaminant, while others may have violations for multiple contaminants. Violation Duration captures temporality by looking at the total period during which a public water system was out of compliance. MCL Multiple captures the "seriousness" of the violations by calculating the average multiple of MCL of the violations.

5.4 Socioeconomic features

We also gathered 36 socioeconomics features from the Census American Community Survey (ACS) and the Employment Development Department (EDD) datasets. The first dataset contains the most detailed information of the US, from which we get data on race and origins, political affiliation, unemployment, median income, and household income quintile distribution of about 1500 census designated places in California. Because completing ACS is not mandatory, we have some missing/incorrect values. In order to preserve data integrity, we performed the following data transformation:

- Leave the missing values to be not a number (NaN) instead of 0 to eliminate the bias the missing could potentially have on our result.
- Change the number that doesn't make realistic sense (such as negative income) to be NaN.
- Merge features such as households with Hispanic Asian and Not Hispanic Asian into households with Asian to make each feature more significant.

We obtained the 3 leading industries that recruit the most people in around 1,500 cities in California and vectorize the data into one hot encoding, which aids the machine to read the data and our later regression.

5.5 Mapping socioeconomic features to PWS's

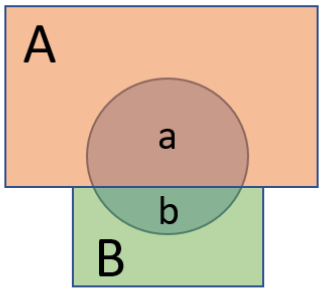
Our raw data has varying granularities. Our water quality data is based on public water systems, which contains the water quality features described above for each public water system. Our socioeconomic status data comes with different units. Some of the data are based on city boundaries and others are based on census regions.

In order to perform any correlation analysis, we need to unify the granularity of our datasets. Since our project is water-focused rather than socioeconomics focused, we created an algorithm to transform the unit socioeconomic data into the unit for our water quality data. For example, originally, we have a total

income for each city, our algorithm will transform this data into total income for each public water system.

In order to map socioeconomic data onto water quality data, we assumed that the population is distributed evenly within each region of socioeconomic data. The algorithm is illustrated below.

Circle: the target PWS
Orange box: Region A
Green box: Region B



Socio-Economical Data:

	Area	Total Income
Region A	A	t_a
Region B	B	t_b

$$total\ income\ for\ this\ PWS = t_a * \frac{a}{A} + t_b * \frac{b}{B}$$

Figure 5.5. The mapping algorithm

5.6 Regression methods

We attempted to use various regression-based methods to establish correlations between water quality features and socioeconomic features. They were generally unsuccessful due to the strong class imbalance present in our dataset. 83% of the PWS's in our dataset did not have any water quality violations in the last 7 years. This made it very difficult for regression-based methods to identify strong trends in the dataset. Here is a summary of the regression-based methods that we tried to use.

Method	**Description**
Linear Regression	Inputs: Industry, racial demographics, median income Target: Numeric value (# of violations, total violation duration) Results: Low R^2 and flat trendline due to class imbalance
Logistic Regression	Inputs: Industry, racial demographics, median income Target: Classification label (has violation record vs. not) Results: Classifier predicts only the dominance class (i.e. no violation)
Quantile Regression	Inputs: Industry, racial demographics, median income Outputs: Water quality violation data (numeric) Results: 95% quantile regression successfully generates trendlines with non-trivial slopes but difficult to extract meaningful interpretations

Table 5.6. Regression Methods used for our study

5.7 Tools and Software Packages

Our analyses were conducted and documented in the Jupyter Notebook and Google Colab environments. Our code was written using Python.

Specific Python packages that we used are as follows. For data manipulation, we used Pandas and Numpy. For visualization, we used Seaborn and Matplotlib packages. For statistical analysis, we used Scikit-learn and StatsModels packages. For working with GIS data and generating maps we used Geopandas and Folium.

6 Conclusions and Future Work

6.1 Findings Summary

Even in California, there are 1 million people currently exposed to contaminated drinking water. 53% of the water quality violations in California happens in the San Joaquin Valley, and Kern County in the San Joaquin Valley is exposed to frequent and persistent water quality violations.

Furthermore, our study shows a correlation between low-income and poor water quality. PWS's serving communities with low median household income are more likely to be exposed to a large number of water contamination violations for a longer duration. Moreover, those PWS's have a higher chance of experiencing the repetition of the same violation.

For a resident living in a place with median income lower than $50K, s/he is more likely to be exposed to more water contamination for a longer duration compared to those who live in a place with a median income over $100K.

Additionally, our case study indicates that effective water treatment can be implemented in low-income communities. With a median income of $38,708, Delano shows successful arsenic treatment as the post-treatment arsenic measurements were consistently close to zero.

6.2 Future Work

Our study shows the existence of water quality disparity among communities with different economic status. Future studies can be conducted on the cause of the difference between Delano and KVSP. Besides, the future study can look into how fast are violations get fixed in communities of different incomes.

6.2.1. Advanced methods and modeling

As mentioned in Our Procedures and Methods, we attempted to go beyond just establishing correlation to perform more advanced analyses such as various regression analyses and building classification models. We discovered that the plain-vanilla models were not meaningfully fitting to our dataset and did not pursue these modeling methods much further. However, we believe that these methods may have benefitted from some additional preprocessing of the data to reduce class imbalance and multicollinearity. Our compliance dataset is heavily imbalanced given that 83% of the PWS's do not have any violation records. It may be desirable to use apply upsampling or downsampling to address this class imbalance issue. Also, in our regression analysis, we used as many as 20+ socioeconomic features as independent variables to regress against water quality features used as target variables. It is likely that multicollinearity among the independent variables was present and selecting features to minimize multicollinearity might have had meaningful improvement on regression performance.

In addition, there were additional models that we considered but did not implement due to time constraints that may be worthwhile. These include support vector machines, decision tree based methods such as random forest, and principal component analysis.

6.2.2. Other directions

Given the important role that the government plays in regulating and operating public water systems, an impactful direction of future work would be investigating the public policy implications. For example, do low-income and high-income PWS's respond to different government measures differently? What type of enforcement actions issued by the Water Boards results in a swifter and permanent resolution to water quality violations? One might try to create a model that policymakers could use measure and estimate the differential impact of their actions on PWS's serving communities of varying socioeconomic status.

Our research highlighted the contrast in arsenic treatment outcomes between Delano and KVSP, but further work will be warranted to fully understand the root causes of those differences. Different treatment methods, equipment, or processes may be contributing to the difference. Understanding the resurgence in KVSP's post-treatment arsenic measurements in 2018-2019 may provide an explanation. Ultimately one would hope to replicate the learnings from Delano's effective arsenic treatment that reduce post-treatment measurements to near-zero levels.

Our research focused on establishing a correlation between low-income and water quality and did not actively explore potential causal links. Perhaps underfunded budgets or less vocal customer bases are mechanisms that contribute to poor water quality at low-income PWS's. Or perhaps there are confounding factors such as the local industry composition that drive the correlation. These are meaningful avenues of further research with key implications on determining the correct interventions to mitigate the problem but were beyond the scope of our project.

In our regression analysis, an early result that we were not able to expand much further on was the apparent correlation between agriculture as an industry and poor water quality. This is in itself is no big

surprise given agriculture is a key industry in San Joaquin Valley, where a majority of the water quality violations have been concentrated. Whether this correlation between the agriculture industry and poor water quality generalizes to the entire state and if so what the causal links are would be interesting questions to ask.

6.3 Team Reflection

6.3.1. Our learnings

Working with the water quality datasets and census datasets altered our view of data science life cycle and how we perceive the water scarcity problem in California. First, when working with real-world datasets, we need to perform data cleaning thoroughly to identify and fix the incomplete, incorrect, and inaccurate parts in the datasets. Second, because we are free to explore any direction that we think is meaningful and insightful using the datasets, we found the ability the relevant questions very crucial to our project as well. Specifically, we tried brainstorming to come up with new ideas and examining the feasibilities meaningfulness of these ideas by consulting domain experts and group decision making.

Besides, although we always knew the severity of drought in California, it's not until we really worked with the dataset that we truly understood the complexity and urgency of the water problems in California. While access to clean drinking water is an essential human right, water contamination is very widespread and frequent in certain regions. Touched by the gravity of the water problems, we are urged to devote our data science skills to exploring useful information that could help to alleviate the problems.

Working with the datasets also allowed us to trial different statistical tools on the datasets and gained a deeper insight into these tools. For example, we tried performing linear regression between the ratio of population recruited by different industries and the number of water contamination events in each region. The regression ended up giving us a high miss rate (false negative rate) due to the imbalanced dataset. Experimenting with different tools on real-world datasets helped us understand what assumptions are made for different classifiers and models and which one should we choose in different situations.

6.3.2. Suggestions for the California Water Data Challenge

The website of the California Water Data Challenge provides a good starting point for our project, with the recommended dataset covering a wide scope of water-related topics.

With the massive number of datasets, we find it a little hard to navigate at the beginning. When we first started our project, our team performed an EDA on every dataset without knowing which ones are more used. If the Challenge can give participants some metric of how often is each dataset used in the past, participants would be able to prioritize.

There are some data we hope to have access to while working on this project. When studying the supply of water to different usages, we realize that the water supply towards agriculture was partially missing. As a huge portion of water usage goes to agriculture in California, it would be great if the Challenge could

provide participant access to agriculture water usage data. We would also like to know about the historical MCL of different contaminants. For example, we know that the MCL for arsenic has been changed from 0.05mg/L to 0.010 mg/L, but when we tried to perform on the study of the historical change of MCL for different chemicals, we found it very hard to find the suitable data.

We have explored many possibilities with the datasets. As we are interested in making contributions to society via data science, we really hope that we could know whether there are any questions that state agencies would find helpful when answered. As a participant, some suggestions on the direction of exploration would be appreciated.

It would be helpful if the Challenge could provide participants with some domain specific background knowledge. During initial EDA phase of our project, we spent substantial amount of time trying to understand some basic concepts. What is a PWS and what is its basic functionality? How frequent are lab test required for each PWS? What are the different regulatory bodies in California and what are their responsibilities? These questions are heavily domain specific and it would be nice if the Challenge can provide participants a quick overview of these background concepts.

6.4 Acknowledgements

First of all, we would like to express the most sincere gratitude to our advisor Prof. David Culler for his help and guidance for our capstone project. We obtained his valuable suggestions not only for the research and data analysis but also for the paper writing, slides, and presentations. We all appreciate much that we could have such a patient and motivating professor with huge knowledge as our capstone Advisor Prof.

In the meantime, we also thank the Berkeley Water Center and Dr. Meredith Lee, which both provided useful resources and insightful comments for our case study and some relevant researches. Without them, it would be hard to further our analysis and get more non-obvious insights from our research.

Last but not least, we would like to give our thank to Ms. Mayasari Lim, who is our 270K advisor. She gave lots of helpful suggestions and inspiring motivation for us to solve all potential conflicts that would happen in our team. Her continuous supports helped us go more smoothly in our capstone project.

7 References

[1] California State Water Resources Control Board. "Human Right to Water | California State Water Resources Control Board." SWRCB.gov. Accessed April 11, 2019. https://www.waterboards.ca.gov/water_issues/programs/hr2w/.
[2] Gavin Newsom, “State of the State Address.” Office of the Governor. February 12, 2019. Accessed April 11, 2019. https://www.gov.ca.gov/2019/02/12/state-of-the-state-address/.
[3] California State Water Resources Control Board. "2018 Safe Drinking Water Data Challenge." The White House CEQ and The State of California Water Data Challenge. Accessed April 11, 2019. http://waterchallenge.data.ca.gov/.

[4] Allaire, Maura, Haowei Wu, and Upmanu Lall. "National Trends in Drinking Water Quality Violations." PNAS. February 27, 2018. Accessed April 11, 2019. https://www.pnas.org/content/115/9/2078.
[5] VanDerslice, James. "Drinking Water Infrastructure and Environmental Disparities: Evidence and Methodological Considerations." American Journal of Public Health. December 2011. Accessed April 11, 2019.
[6] U.S. Census Bureau, American Community Survey 1 Year, 2017, accessed February 20, 2018. https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2017/1-year.html
[7] Genard, Gary. "How to Be Persuasive If You're a Data Scientist." The Genard Method. Accessed April 11, 2019. https://www.genardmethod.com/blog/how-to-be-persuasive-if-youre-a-data-scientist.
[8] Trenton, NJ. "Maximum Contaminant Levels (MCLs) for Perfluorononanoic Acid and 1,2,3-Trichloropropane; Private Well Testing for Arsenic, Gross Alpha Particle Activity, and Certain Synthetic Organic Compounds". New Jersey Department of Environmental Protection. 2018-09-04.
[9] Carton, Robert J. "Review of the 2006 United States National Research Council report: fluoride in drinking water." Fluoride39, no. 3 (2006): 163-172.
[10] Bland, Alastair. "100,000 Residents In Bountiful Central Valley Still Lack Access to Clean Water." KCET. May 18, 2018. Accessed April 11, 2019. https://www.kcet.org/shows/earth-focus/100000-residents-in-bountiful-central-valley-still-lack-access-to-clean-water.
[11] Greenberg, Alissa. "Sinking Land, Poisoned Water: The Dark Side of California's Mega Farms." The Guardian. July 18, 2018. Accessed April 11, 2019. https://www.theguardian.com/environment/2018/jul/18/california-central-valley-sinking-arsenic-water-farming-agriculture.
[12] "Water Quality." Community Water Center. Accessed April 22, 2019. https://www.communitywatercenter.org/contamination.
[13] "Arsenic." American Cancer Society. Accessed April 10, 2019. https://www.cancer.org/cancer/cancer-causes/arsenic.html.
[14] "Fact Sheets: Nitrate: Health Effects in Drinking Water." PSEP. Accessed April 10, 2019.
[15] Hajat, Anjum, Charlene Hsia, and Marie S. O'Neill. "Socioeconomic disparities and air pollution exposure: a global review." Current environmental health reports 2, no. 4 (2015): 440-450.
[16] "United States District Court: Eastern District of California" Case 1:11-cv-00809-AWI-SKO. https://www.govinfo.gov/content/pkg/USCOURTS-caed-1_11-cv-00809/pdf/USCOURTS-caed-1_11-cv-00809-38.pdf
[17] Gavin Newsom, "State of the State Address." Office of the Governor. February 12, 2019. Accessed April 11, 2019. https://www.gov.ca.gov/2019/02/12/state-of-the-state-address/.
[18] "Drinking Water - Water System Service Area Boundaries", Distributed by California Open Data Portal. https://data.ca.gov/dataset/drinking-water-water-system-service-area-boundaries.
[19] "Drinking Water - Public Water System Information", Distributed by California Open Data Portal. https://data.ca.gov/dataset/drinking-water-public-water-system-information.

[20] “Drinking Water - Human Right to Water Regulatory (including Enforcement Actions) Information”, Distributed by California Open Data Portal.
https://data.ca.gov/dataset/drinking-water-human-right-water-regulatory-including-enforcement-actions-information.
[21] ”Drinking Water - Laboratory Water Quality Results”, California Water Boards.
https://www.waterboards.ca.gov/drinking_water/certlic/drinkingwater/EDTlibrary.html.
[22] U.S. Census Bureau, American Community Survey 5 Year, 2013 to 2017, accessed February 20, 2018.
https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2017/5-year.html
[23] U.S. Census Bureau, American Community Survey 1 Year, 2017, accessed February 20, 2018.
https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2017/1-year.html