Water quality and socioeconomic status in California

Anastasiya Markova
Halıcıoğlu Data Science Institute
University of California, San Diego
anmarkova@ucsd.edu
LinkedIn
GitHub

Zoe Ludena
Halıcıoğlu Data Science Institute
University of California, San Diego
zludena@ucsd.edu
LinkedIn
GitHub

Steven Luong
Halıcıoğlu Data Science Institute
University of California, San Diego
sxluong@ucsd.edu
LinkedIn
GitHub

ABSTRACT

This study investigates the relationship between socioeconomic factors, race, and water quality in California, a state known for its water scarcity and droughts. Based on previous studies, there seems to be conflicting conclusions regarding how influential socioeconomic status and race affects water quality. Our investigation aims to clarify this relationship by analyzing the recent datasets encompassing water quality metrics, income levels, county demographics, and education levels. We hypothesize water quality will be affected by income levels, education levels. We are also curious to find out if there is a correlation between water quality and race.

We utilized data from various sources, which included sources such as the Drinking Water - SAFER Dashboard, ACS DEMOGRAPHIC AND HOUSING ESTIMATES - U.S. Census Bureau, Income Table For California - U.S. Census Bureau, Education Table For California - U.S. Census Bureau, and Zip to ZCTA - Github users. Through various statistical methods, including linear regression, logistic regression, ANOVA and Kruskal-Wallis Test, we tested our hypothesis.

We found that columns attributed to education, income, and race can act as determining factors of water quality. This means we have concluded that socioeconomic factors affect water quality, but we are unsure of which groups affect water quality in a positive or negative way.

1. INTRODUCTION

1.1 Background Research

A sustainable and clean source of water is not only a necessity, but a fundamental human right for all citizens. Water is vital in a health and environment sense, but becomes a matter of social justice when there is not an equitable distribution of clean water for those of different socioeconomic status, ethnic background, and geographical areas.

In the research article, "Socioeconomic factors and water quality in California,"[1] Y. H. Farzin and Kelly Grogan explored the key factors affecting California's water quality. They worked with data

from 1993-2006 that includes water quality and socioeconomic data. They used three classes of models, the Environmental Kuznets Curve (EKC), a more inclusive model containing socioeconomic variables, and a model that included socioeconomic variables and spatial correlation. Note EKC states that as income increases then environmental quality declines, but after a certain per capita income level, quality begins to and continues to improve as income increases. They investigated whether purely economic factors affect water quality or California agriculture, race, and education. They found the per capita income was not a significant factor in explaining variability in water quality, but agriculture and industrial activities did.

In the research article, "Disparities in drinking water quality evidence,"[2] Sarah Acquah and Maura Allaire address disparities between water qualities based on income and race. Their goal was to create more empirical evidence for California's government in an effort to improve the drinking water quality for those in disadvantaged communities. They worked with data from 2000 to 2018 that included Community Water Systems (CWSs) in California, EPA's Safe Drinking Water Information Systems, and the United States census. They used Probit regression models to examine the likelihood of violations as a function of the demographics of the CWS service area. They found low-income communities and minority groups (like hispanics) are more likely to face health-related water quality violations.

Our study aims to clear up these contradicting findings by finding if income, education, and race affect water quality in different parts of California. The tests we chose to use do not tell us if specific income levels, education levels, or races affect the water quality in a positive or negative way, but we encourage others to explore our research to find out.

1.2 Hypothesis

In this paper we want to see if socioeconomic factors affect water quality in California. In research articles we found, like the ones provided above, we found contradicting results. These differences probably come from the differences in models and data. We noticed the times looked at also differed between the two research papers. We want to conduct hypothesis tests to illuminate the connections between socioeconomic factors, demographic attributes, and water quality today. We hope by analyzing datasets detailing water quality metrics across California, demographic insights from the U.S. Census and mapping resources can uncover statistically significant correlations between economic status, racial demographics, education levels, and various parameters of water quality. We hypothesize water quality will be affected by income levels and education levels. We are also curious to discover if there is a correlation between water quality and race.

1.3 Our data

Our first dataset is called Drinking Water - SAFER Dashboard Failing and At-Risk Drinking Water Systems.[3] The number of observations inside of this dataset is 3232 and the number of variables is 138. This dataset contains information on the water system, the location, and water information (E-coli, accessibility, violations, concerns, drought, groundwater, median household income, socioeconomic burden, and deficiencies). Our second dataset is called DP05|ACS DEMOGRAPHIC AND HOUSING ESTIMATES.[4] The number of observations inside of this dataset is 33120 and the number of variables is 138. This dataset contains information on sex, age, and race with total populations, margin of error, and estimates. Our third dataset is called zip_to_zcta.[5] The number of observations inside of this dataset is 41131 and the number of variables is 3. This dataset contains a mapping from zip codes to ZCTAs. Our

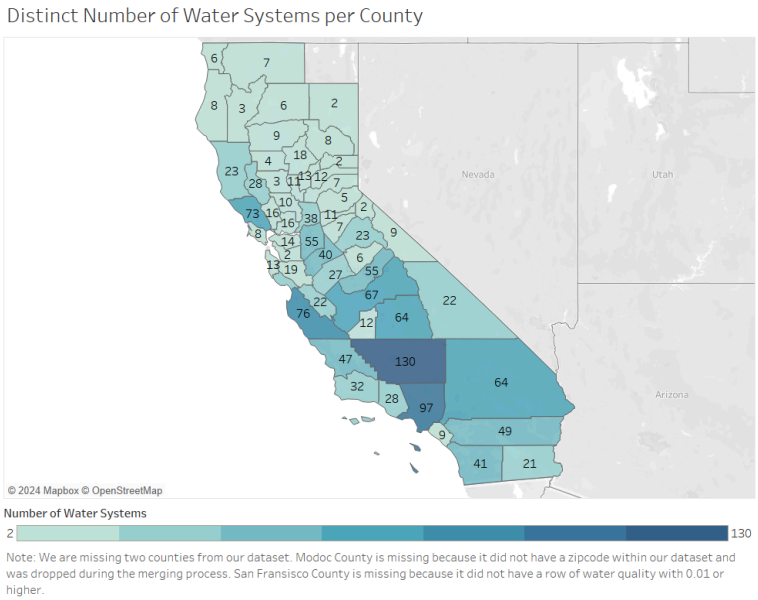
fourth dataset is called California Income Table.[6] The number of observations inside of this dataset is 60 and the number of variables is 4. This dataset contains county, the Federal Information Processing Standards (FIPS), percentages, number of people with at least a Bachelor's Degree, and rank within the United States. Our fifth dataset is called California Education Table.[7] The number of observations inside of this dataset is 60 and the number of variables is 5. This dataset contains county, the Federal Information Processing Standards (FIPS), median household income in US dollars, and rank within the United States.

2. MATERIALS AND METHODS

2.1 Exploratory Data Analysis

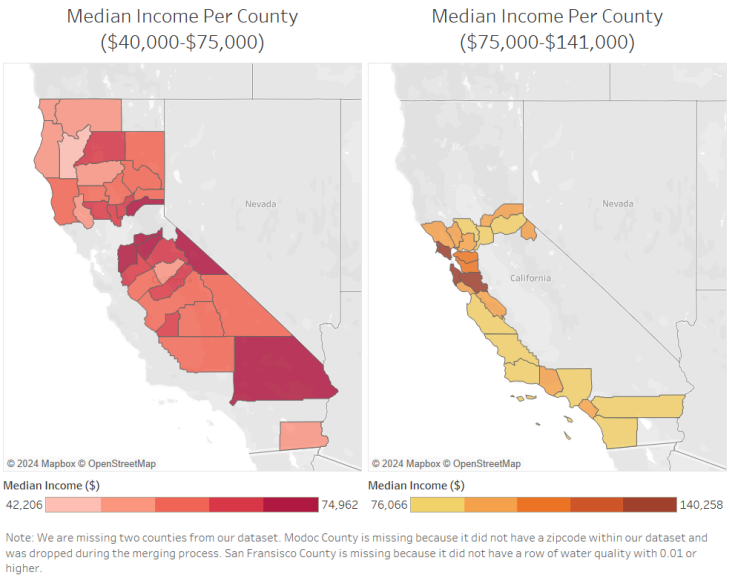
To better understand our data we performed exploratory data analysis on different aspects of our dataset. We focused on the number of water systems, median income per county, education level per county, water quality scores, and groups of races.

Fig. 1 The number of water systems per county in California.



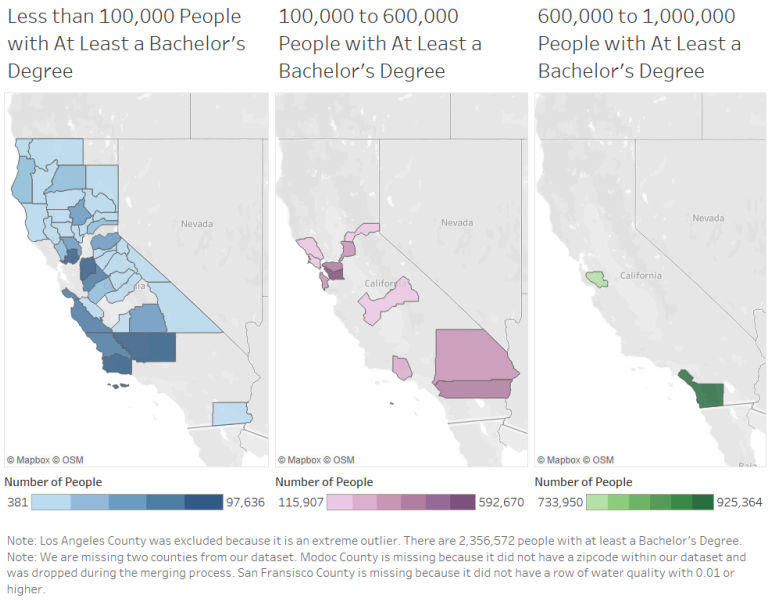
Pictured in Figure 1 is a choropleth map that illustrates the number of water systems per county inside of California. Inside of Tableau we were able to use the geographical location of the county then grouped the distinct number of water systems. From Figure 1 we found that some counties have many more water systems than others. This caused us to become more cautious when making assumptions about a single county based on the number of water systems.

Fig. 2 The median income in dollars per county in California.



Pictured in Figure 2 is a choropleth map that illustrates the median income in dollars per county inside of California. Inside of Tableau we were able to use the geographical location of the county then display the median income in dollars. We made two separate maps to better display the data. We learned there is a large variety of incomes inside of California.

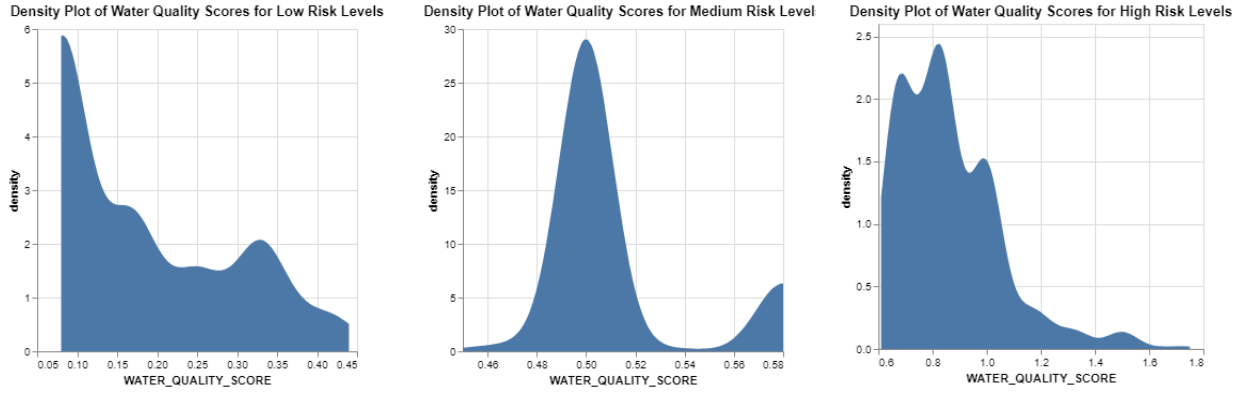
Fig. 3 The number of people with a minimum of bachelor's degrees per county in California.



Pictured in Figure 3 is a choropleth map that illustrates the number of people with a minimum of a bachelor's degree per county inside of California. Inside of Tableau we were able to use the geographical

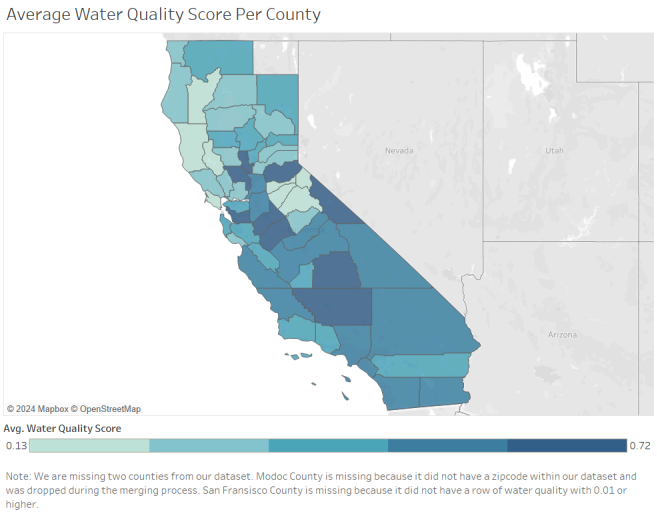
location of the county then display the number of people. We made three separate maps to better display the data. We omitted Los Angeles County which had an outlier of 2,356,572 people with a minimum of a bachelor's degree. We learned there is a wide range of people with at least a bachelor's degree. It appears there are more counties with 100,000 people with bachelor's degrees or lower.

Fig 4. Density plots for water quality scores low, medium, and high.



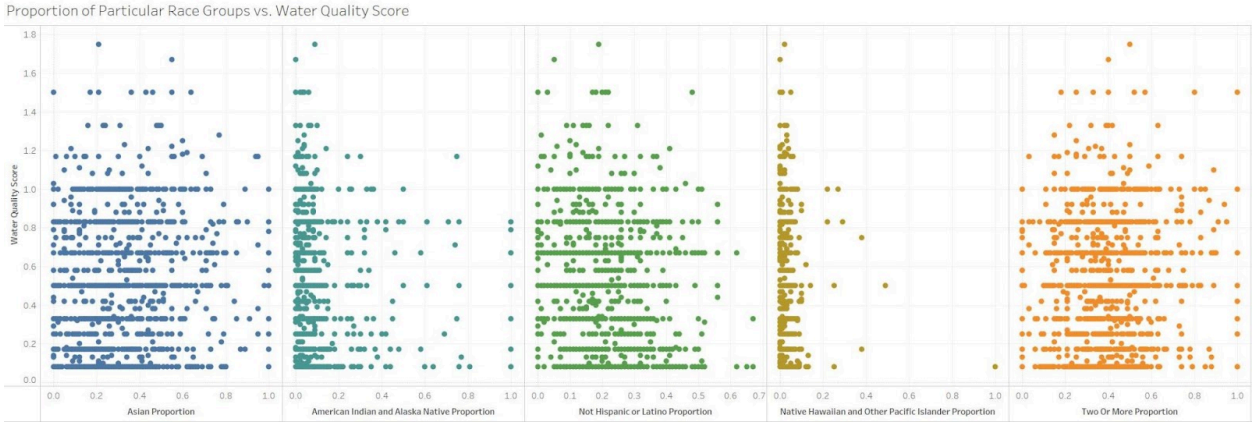
Pictured in Figure 4 are three density plots. The risk level is determined by the unweighted water quality category score. There is one for low risk levels, which is 0.01 to 0.44, medium risk levels, which is 0.45 to 0.59, and high risk levels, which is 0.6 or above. The density plot shows us the overall shape of the distribution of water quality scores at different risk levels. We can see the shape for low risk levels and high risk levels are skewed to the right. We see the shape for medium risk levels appears almost normal, but has another, smaller peak around 0.58. This tells us that if there is a low or high risk level it is more likely to be on the lower side of its range. For medium risk level it is more likely to be in the middle of its range.

Fig 5. The average water quality score by county.



Pictured in Figure 5, this choropleth map illustrates the average water quality score in each county starting from the lowest echelon of 0.13 to 0.72. Inside Tableau we were able to use the geographical location of the county then display then the average water quality score. The southern counties of california seem to have much better water quality on average than that of the counties of central/northern California.

Fig 6. Scatter plots of race group proportions and their associated water quality scores.



X-axis Respective to the Scatterplots: Asian, American Indian and Alaska Native, Not Hispanic or Latino Group, Native Hawaiian and Other Pacific Islander, and Two or more Races.

Pictured in Figure 6, are scatterplots that showcase the correlations between the proportion of race groups within a county and the water quality score of that county. From these visuals generated by Tableau, it is very hard to make out an obvious trend. After calculating Pearson's correlation score for all groupings, we got the following: 0.1, -0.03, -0.18, -0.01, and 0.02. These correlation scores, all nearing the neutral score of 0, all indicate a very weak relationship between the race groups and the water quality.

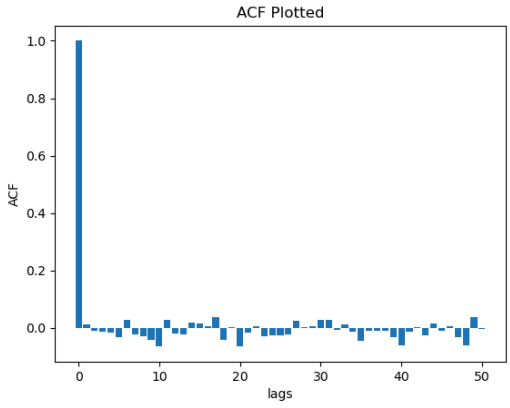
2.2 Linear Regression

We wanted to use linear regression with forward selection, which iteratively adds the predictors that improve the model based on Bayesian Information Criterion (BIC) and backward selection, which iteratively removes the predictors that contribute least to the model according to BIC. We performing linear regression on all of our variables, county (for geographical location), education value percent, the number of people who at least have a bachelor's degree in the county, and the United States rank for the education for the county (for education), median income dollars and the United States rank for the income for the county (for income), and total populations for different races per county (for races) to predict water quality scores.

To perform linear regression the relationship between our independent variables and our dependent variables should be linear. The observations should be independent from each other. The residuals should have constant variance (homoscedasticity). The residuals should be approximately normally distributed. There should not be multicollinearity (independent variables should not be highly correlated with each other).

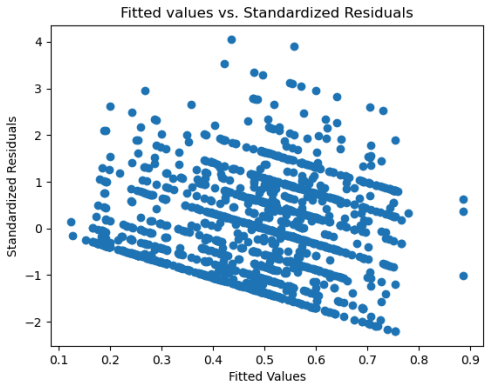
To test linearity we plotted each of our variables against water quality score in a scatter plot. We found none of the variables were linear with water quality score. Despite our best judgment we continue to test our assumptions. We graphed an autocorrelation function (ACF) plot to test for independence.

Fig. 7 Autocorrelation function plot to test for independent variables.



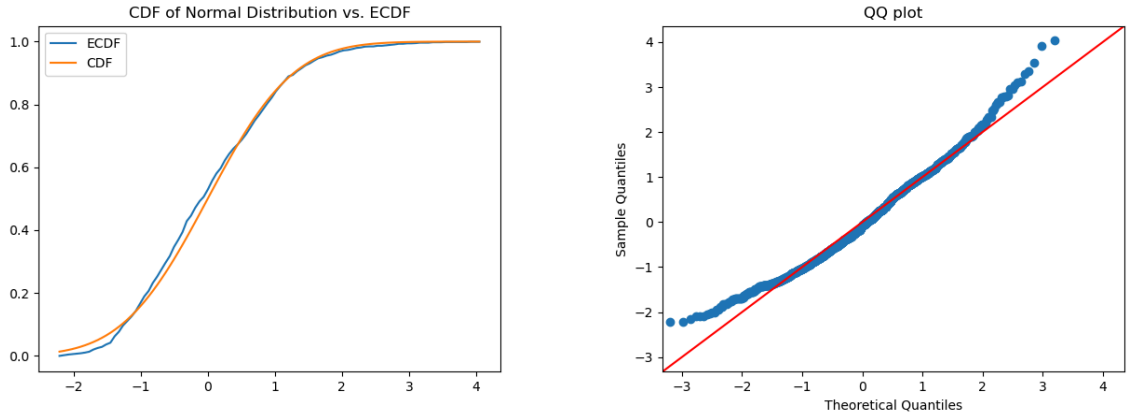
We plotted Figure 7 by plotting the residuals of our linear regression model as a barplot. We learned that the data looks fairly independent. You can see most of the lags are zero. When lags are near zero it means there is little to no correlation between the values of the time series at different lags. This implies the time series are random and independent of past observations. We looked at a scatter plot of the residuals and found they did not have a constant variance. See this below in Figure 8.

Fig 8. Our fitted values plotted against the standardized residuals.



We plotted Figure 8 by plotting the model's fitted values against the standardized residuals of our model. We would hope for this plot to be completely random. As you can see above there is a sharp edge with lines inside of it. This means we have some kind of constant variance. We can also see from Figure 9 that our residuals are not normally distributed. To be confident in this observation we plotted the cumulative distribution function (CDF) and empirical distribution function (ECDF) and a QQ-plot.

Fig. 9 CDF of Normal Distribution vs. ECDF and QQ-plot to test for normality.



We plotted Figure 9 by plotting the standardized residuals in the QQ-plot and by comparing the true normal distribution's CDF and plotting it against the ECDF of our data. If the data were normal then the scatterplot points in the QQ-plot would line up with the red line and the CDF and ECDF would look the same. We confirmed that our data does not meet the assumption for normality. Before scratching the idea for linear regression we decided to create a heatmap to test for multicollinearity.

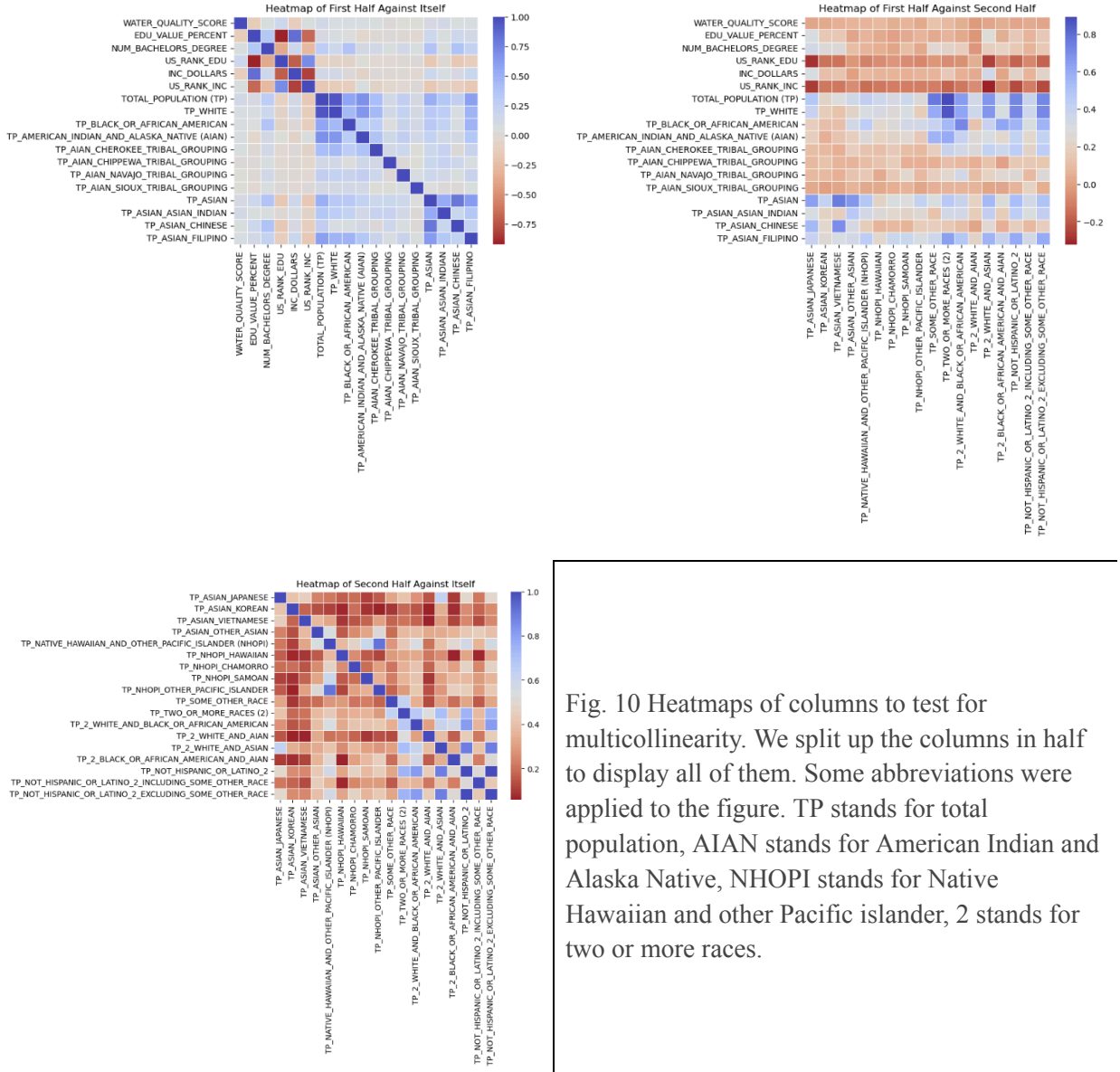


Fig. 10 Heatmaps of columns to test for multicollinearity. We split up the columns in half to display all of them. Some abbreviations were applied to the figure. TP stands for total population, AIAN stands for American Indian and Alaska Native, NHOPI stands for Native Hawaiian and other Pacific islander, 2 stands for two or more races.

We plotted Figure 10 by using seaborn's heatmap method on our dataframe's correlation. Our first heatmap has a scale of what appears to be about -1 to 1. We can see in our heatmap of the first half against itself that there is a noticeably positive correlation between median income and education value percentages and true population and true population white. We also found a negative correlation between the United States rank of education and education value percentages, the United states rank of income and education value percentages, the United states rank of income and the United states rank of education, the United states rank of income and median income. This makes sense since a rank of 1, would correspond to highest statistics, so a region ranked as number one for income would have the highest median income.

In our second heatmap notice the change in scale from above 0.8 to a little below -0.2.We can see in our heatmap of the first half against the second half there is a positive correlation between the total population and total population of two or more races, the total population of white and total population of two or more races, the total population of asian and total population of asian chinese, total population of

white and total population of white and asian, total population and total population of non hispanic or latino two or more races. We see a negative correlation between the United States rank of education and total population of asian, Japanese, the United States rank of Education and the total population of white and asian, the United States rank of income and asian, Japanese, and the United States rank of income the total population of white and asian.

We learned in our heatmap of the second half against the second half there are many negative and positive correlations, but the lower correlations do not have the same scale as the first heatmap. The range appears to be around 0 to 1. There appears to be many with low correlation rather than high correlation. Positive correlations found are total population of not hispanic or latino two or more races and total population of not hispanic or latino two or more races excluding some other race and true population of native Hawaiian and other pacific islander and the true population of native Hawaiian and other pacific islander other pacific islander.

These heatmaps tell us there is some correlation between some of our variables. A few of the ones we found have a very high correlation, which tells us there is multicollinearity of our data. This means we fail this assumption. With this we can see we failed every assumption for linear regression. However, this did not stop us from doing the linear regression. We decided to carry through and found the following statistics:

Fig. 11 Linear Regression OLS results.

OLS Regression Results

R-squared:	0.182
Adj. R-Squared:	0.133
F-statistic:	3.698
Log-Likelihood:	-33.18
AIC:	824.4
BIC:	1256

To interpret our results we will first explain what each value in the table means. R-squared represents the coefficient of determination, which means it measures how well the regression prediction approximates the real data points. An R-squared of 0 means the model does not explain any of the variability of the response data around its mean, which means it fails to fit the data. R-squared represents the model fitting the data perfectly and indicates the model explains all the variability of the response data around the mean. Our R-squared is 0.182, which is close to zero, which means our model does not fit the data well.

Adjusted R-squared (Adj. R-squared) penalizes R-squared values for including additional predictors that do not improve the model's performance sufficiently. A higher adjusted R-squared means it is a better fit of the model for the data and a lower adjusted R-squared means the additional predictors do not contribute significantly to explain the variability in the dependent variable. Our adjusted R-squared is 0.133 which means we have predictors that do not improve the model's performance and are not a good fit for our data.

An F-statistic tests the null hypothesis that all of the regression coefficients are equal to zero. It evaluates if the independent variables collectively have a significant effect on the dependent variable. A larger F-statistic indicates the model is more likely to be statistically significant than a model with no independent variables. Our p-value of the F-statistic is less than our significance level of 0.05, meaning that our model is statistically significant compared to the model with no independent variables.

Log likelihood measures how well the mode's predicted probabilities match the observed outcomes. Log likelihood sums the contributions of each observation to the overall likelihood. We want to find parameter values that maximize the log likelihood. We would need to compare to other models to say if -33.18 is a good or bad likelihood. The one that is higher would generally be considered a better fit for the data.

Akaike Information Criterion (AIC) is used to compare the goodness of fit of statistical models. It balances the model's goodness of fit with its complexity, penalizing models that are too complex. A lower AIC indicates a better balance between model fit and complexity. Our AIC is high at 824.4. However, we cannot say if it is high or low for this model without creating different models and comparing our value with the new AICs.

Bayesian Information Criterion (BIC) is used for model selection in statistics, particularly for linear regression. It is another measure used to balance the goodness of fit with complexity. BIC penalizes models more heavily for complexity than AIC. It prefers simpler models. A lower BIC indicates a better balance between model fit and complexity. Our BIC is high at 1256. However, we cannot say if it is high or low for this model without creating different models and comparing our value with the new BICs.

To select our variables we used mixed selection with the BIC criterion. Backward selection is usually determined by education value percent, median income ($), U.S. rank education, U.S. rank income, and some other race total populations were the best variables we should use. Note that the list of variables changes as backward selection minimizes to a different set of variables each time. Forward selection told us the usual total population of some races, U.S. rank education, and number of bachelor degrees would be most helpful. Just like backward selection, these columns vary depending on which minimum the selection process finds. This indicates that race, income, and education seem to play some role with water quality.

That being said, we cannot trust these results because we failed every assumption for linear regression. Using linear regression is not a good fit for our data. This was reiterated by our low R-squared. We can see our AIC and BIC values are pretty large numbers in general, so it makes us also think that we do not have a good balance of goodness of fit with complexity. We want to say that income, education, and race affect water quality score, but more testing is needed.

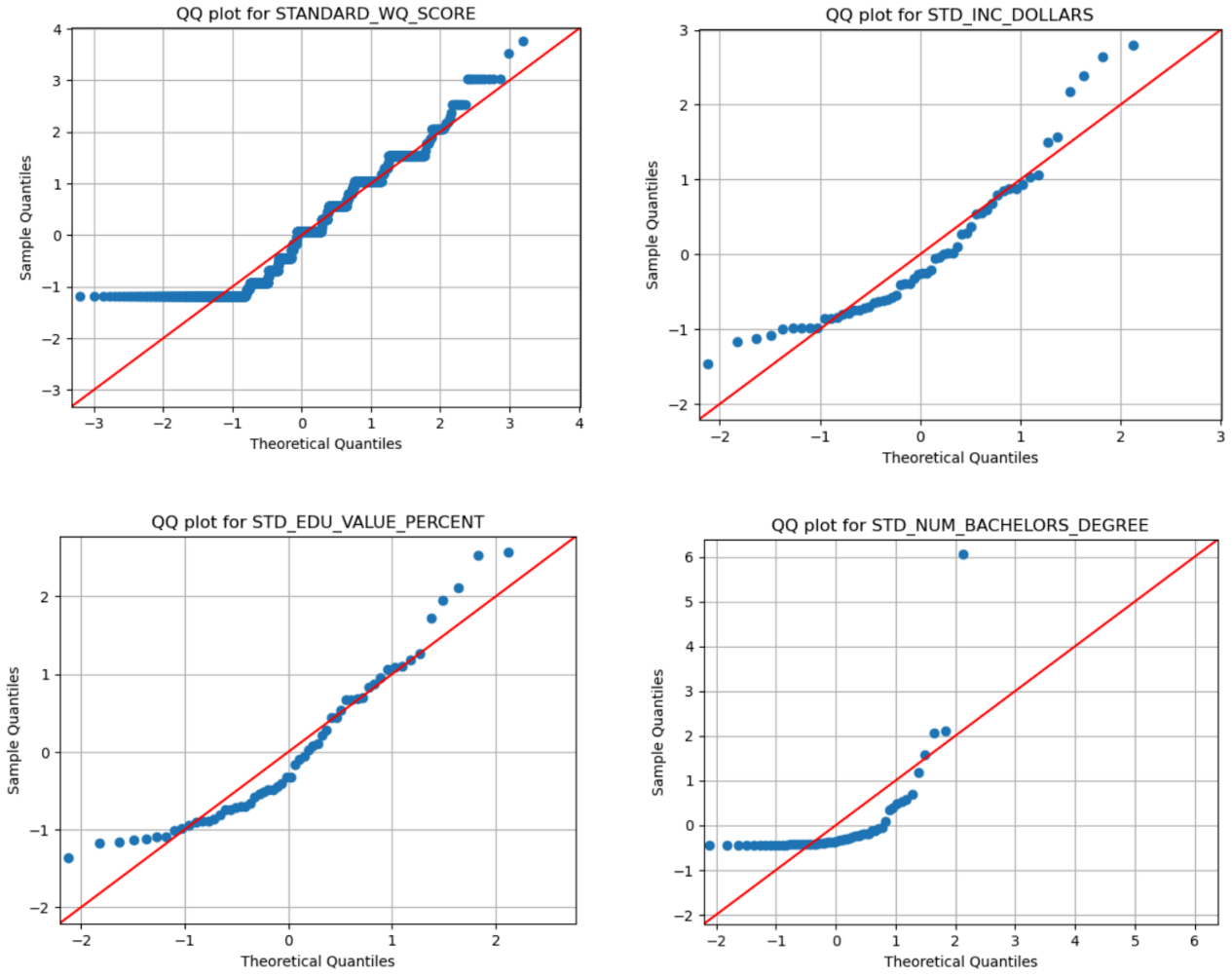
2.3 ANOVA

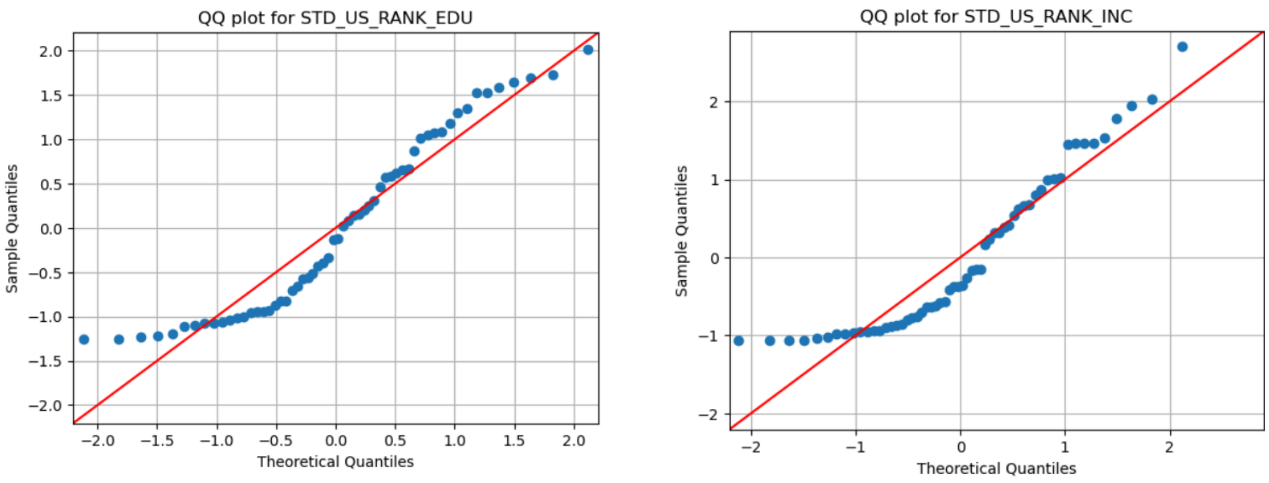
ANOVA is used to determine if there is a statistically significant difference in means between independent groups. We want to use this test to see if there is a statistically significant difference in means between different levels of the same variable. The assumptions to perform ANOVA hypothesis testing is normality, independence, and homogeneity of variances.

To test our assumption of normality we used the Kolmogorov Smirnov test. We found none of the races came from normal distributions. This makes sense because the United States is a melting pot of different ethnicities. We found the standardized data inside of the water quality score did not come from a normal distribution either. This did not surprise us because of figure 4. We saw how the density plots did not look normal and were instead right skewed. We found none of our education or income data was

normal either. This makes sense because some places in California are known to be more wealthy than others and we associate wealthier locations with high paying jobs that might require more education. We were curious to see how close to normal the water quality score, income, and education columns were, so we plotted QQ-plots.

Fig. 12 QQ-plots to test ANOVA normality.





We decided in Figure 12 to plot the QQ-Plots of our standardized water quality score, income columns, and education columns to visualize how close to normal the distributions are. We can see the tails of all of the graphs come off of the reference line for a normal distribution. We can see the standardized United States rank of education and income, the standardized education value percentage, and the standardized median income all take on a little of an s-curve shape. Our standardized number of bachelor's degrees shoots up exponentially. The standardized water quality score looks the most similar to the line, but flattens out occasionally.

Since our data as a whole does not come from a normal distribution, we were not surprised to find that the groups do not come from normal distribution either. We first performed the Kolmogorov Smirnov test to check the normality of each group for our ANOVA test. We are doing an ANOVA for each column of interest (education statistics, income statistics, race statistics columns) so we performed the normality test for each group per column of interest. What we found is that the majority of the groups do not come from the normal distribution for any of the columns. This means that we do not meet the normality assumption for ANOVA.

To test our assumption of independence we used the chi-square test of independence. We chose this test because our observed data (water quality score) of focus is not time-series based, precluding the use of the Autocorrelation Function. We separated the data into groups by quartile per column of interest and then assessed the independence of the water quality levels between the groups using the chi-square test.

Fig. 13 The chi-square test results showed the following for each risk level category:

Columns	Is it independent?
Percentage of Total White Population	No
Percentage of Total Population Black or African American	No
Percentage of Total Population American Indian and Alaska Native	Yes

Percentage of Total Population Asian	No
Percentage of Total Population Asian Filipino	No
Percentage of Total Population some other race	No
Percentage of Total Population two or more races	No
Percentage of Total Population two or more races white and american indian and Alaska Native	No
Percentage of Total Population two or more races white and asian	Yes
Percentage of Total Population not hispanic or latino two or more races	Yes
Percentage of Total Population not hispanic or latino two or more races two races excluding some other race	No
Education Value Percentage	No
Num Bachelor's Degrees	No
U.S. Rank Education	No
Median Income Dollars	No
U.S. Rank Income	No

Of our sixteen groups only three of them were found to be independent because they rejected the null hypothesis, which means that the observed values of the specified groups are not independent. This means we fail this assumption for ANOVA.

Taking into account that the data within each group is not normal, we chose to use Levene Test to check for homogeneity of variances. The Levene test has two assumptions: the data comes from a random sample, and the data of the samples is independent of each other. Unfortunately, we do not meet these assumptions, so we proceed with caution. Under the null hypothesis Levene test states that the variances of all the groups are the same. We meet the criteria of these assumptions. After performing the Levene test we found that the majority of the time we reject the null hypothesis, and almost all of the groups seem to have inconsistent variances between groups.

After we tested our assumptions we decided to bin our data from the minimum to the first quartile, from the first quartile to the second quartile, from the second quartile to the third quartile, and from the third quartile to the maximum value. From here we performed a one way ANOVA test on the bins. Our null hypothesis for these tests say the means of each group is the same. This means if we reject the null hypothesis the column the test was performed on impacts water quality. If we fail to reject the null hypothesis then the column the test was performed on does not impact water quality. For our ANOVA testing we chose a p-value of 0.05. We created a function in python that used scipy.stats'

f_oneway function to help us determine if we fail to reject the null hypothesis or reject the null hypothesis. We performed ANOVA on all of our columns that fit within our designated bins.

Fig. 14 ANOVA test results for those with unique bins. Recall bins were minimum to first quartile, first quartile to second quartile, second quartile to third quartile, and third quartile to maximum. The other columns in our dataset did not have enough data and in the binning process were skipped.

Column:	**P-value:**	**Conclusion:**
Median Income ($)	0.00000001	Reject the Null Hypothesis
U.S. Rank of Income	0.00000001	Reject the Null Hypothesis
Number of Bachelor Degrees	0.00000012	Reject the Null Hypothesis
Education Value Percentage	0.00000001	Reject the Null Hypothesis
U.S. Rank of Education	0.00000867	Reject the Null Hypothesis
Percentage of Total White Population	0.00000008	Reject the Null Hypothesis
Percentage of Total Population Black or African American	0.00444551	Reject the Null Hypothesis
Percentage of Total Population American Indian and Alaska Native	0.0445517	Reject the Null Hypothesis
Percentage of Total Population Asian	0.04815210	Reject the Null Hypothesis
Percentage of Total Population Asian Filipino	0.00528847	Reject the Null Hypothesis
Percentage of Total Population some other race	0.00000008	Reject the Null Hypothesis
Percentage of Total Population two or more races	0.00109753	Reject the Null Hypothesis
Percentage of Total Population two or more races white and american indian and Alaska Native	0.00180463	Reject the Null Hypothesis
Percentage of Total Population two or more races white and asian	0.00088532	Reject the Null Hypothesis
Percentage of Total Population not hispanic or latino two or more races	0.00000004	Reject the Null Hypothesis
Percentage of Total Population not hispanic or latino two or more races two races excluding some other race	0.00000277	Reject the Null Hypothesis

Every instance of rejecting the null hypothesis means the column affects water quality. We can safely conclude that median income and education impact water quality. We discovered some races seem to affect water quality as well. However, since we were unable to determine for all of our race columns due to binning issues we decided to use percentages of the race within the county.

We suspected that there may be interaction between race variables and education and income variables when determining water quality. To test this theory we decided to run a Two-Way ANOVA between education value percentage, Number of Bachelor Degrees, U.S. Rank of Education, Median Income ($), U.S. Rank of Income, and every race column. Our results were as following: (a checkmark indicates that we found statistically significant evidence for interaction, nothing indicates we didn't find statistically significant evidence for interaction)

Fig 15. Two-way ANOVA test results for those with unique bins. Vertical and horizontal index indicates the combinations of independent variables whose interaction we were assessing. Checked values indicate that the p-value was less than 0.05, while missing values indicate that p-value is greater than 0.05.

	Education Value Percentage	Number of Bachelor Degrees	U.S. Rank of Education	Median Income ($)	U.S. Rank of Income
% White Population	✓				
% Black or African American		✓	✓		
% American Indian and Alaska Native	✓	✓			
% Asian		✓	✓	✓	
% Asian Filipino			✓		
% Some Other Race		✓	✓	✓	
% Two+ races	✓		✓	✓	✓
% Two+ races white, Native American/Alaskan	✓		✓		
% Two+n races white and asian			✓		✓
% not hispanic or latino two+ races		✓	✓		

Based on the table we can see that the U.S. Rank of Education tends to have the most interaction with race columns. Additionally, Percentage of Total Population two or more races and Percentage of Total Population some other race tend to have the most interaction with education and income demographics columns. However, there is a consistent interaction pattern between race demographics and the income and education demographics. Thus, it seems that race has its own impact on water quality separate of the education or income data. However, we must take these results with a grain of salt, since we did not meet any of the assumptions for ANOVA.

2.4 Kruskal-Wallis Test

As an alternative to ANOVA, we decided to use Kruskal-Wallis Test, because it is a non-parametric method, and makes minimal assumptions about the distribution, to test if the samples are coming from the same distribution. The assumptions for Kruskal-Wallis Test are that the observations are independent of each other, population is not necessarily normal and the variances are not necessarily equal, and the observations must be drawn from the population through random sampling. All of these assumptions except for independence are met in our case, so we chose to proceed with a lot of caution.

Under the null hypothesis Kruskal-Wallis Test states that the median of all groups are the same indicating that the data of each group is coming from the same distribution. Our results showed that for every column with enough data we reject the null hypothesis. This was true for the following columns: Education Value Percentage, Number of Bachelor Degrees, U.S. Rank of Education, Median Income ($), U.S. Rank of Income, Percentage of Total White Population, Percentage of Total Population Black or African American, Percentage of Total Population American Indian and Alaska Native, Percentage of Total Population Asian, Percentage of Total Population Asian Filipino, Percentage of Total Population some other race, Percentage of Total Population two or more races, Percentage of Total Population two or more races white and american indian and Alaska Native, Percentage of Total Population two or more races white and asian, Percentage of Total Population not hispanic or latino two or more races, Percentage of Total Population not hispanic or latino two or more races two races excluding some other race. This means that at least one group within these columns has a statistically significantly different median at the alpha level of 0.05. Thus, it is probable that all different groups from all columns mentioned above can act as determining factors of water quality.

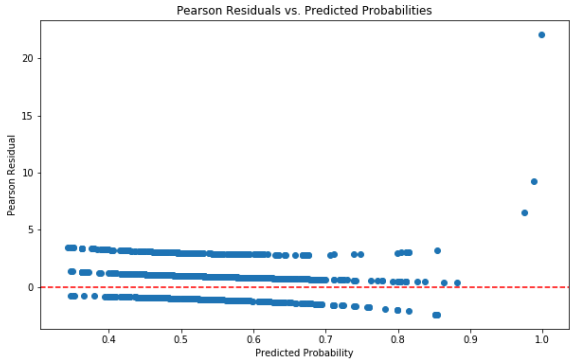
Interestingly, our Kruskal-Wallis Test and ANOVA came to the same conclusions. These conclusions also contradict our EDA finding that there is a weak correlation between proportions of race and water quality (Fig 6.). The results of our tests may not match the EDA findings because of the way we binned the scores or because of the lack of independence. We may need to further investigate this issue.

2.5 Logistic Regression

This section has been added after the rest of our models because we wanted to try one last model. We found logistic regression did not work.

The assumptions to perform logistic regression are independence, linearity of independent variables, no perfect multicollinearity, and homoscedasticity. We know that our data is not independent because of Figure 13. After running the chi-square test of independence within each water quality risk level group, categorized by quartiles, we found that the samples in each group were not independent. We know that our data does not have perfect multicollinearity because of Figure 10. We have to test for linearity of independent variables and homoscedasticity.

Fig. 16 Pearson residual plot to test for linearity of independent variables and homoscedasticity.



We plotted Figure 16 by calculating the Pearson residual, which is $r_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$, where y is our actual value, $\hat{p}$ is our predicted probability value. We can see that our Pearson residual plot has a clear pattern, which means that we fail the assumptions of linearity of independent variables and homoscedasticity. Regardless of failing some of the assumptions needed for logistic regression we continued.

Fig. 17 Multinomial Logistic Regression Results.

MLN Logistic Regression Results

Pseudo R-squared:	0.07283
Log-Likelihood:	-1360.9
AIC:	2849.79
BIC:	3186.8

Unlike the traditional R-squared in linear regression, the pseudo R-squared is used when the outcome variable is categorical. However, just like the R-squared coefficient scores, they are interpreted in the same manner as in Figure 11 (it measures how well the regression prediction approximates the real data points). The model's pseudo R-squared coefficient is 0.07283 indicating that the model doesn't effectively fit the data well. The log-likelihood value is at -1360.9 highly suggests that the model has much room for improvement in fitting the data's underlying patterns. The scores for both AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion), metrics that measure model fit while also penalizing for complexity, are the following: 2849.79 and 3186.8. This means that the model has much room for improvement in terms of complexity and effectively fitting the data.

When training the full logistic regression model, it was expected to get results that were as lackluster as the OLS regression model. This was due to the fact that we did not employ any variable selection processes to eliminate model complexity, hence the expectation for metrics worse than that of the OLS regression model.

RESULTS

Linear Regression: When looking at the linear regression, our data also failed the majority of assumptions. Primarily, the homoscedasticity and linearity tests were the ones that deviated the most from the desired result. Thus, we can not interpret the results with confidence. Backward and forward selection also yielded extremely different covariate recommendations, which is indicative of non-linearity. The AIC and BIC of our model is also quite large even after minimization techniques indicating that the model is a poor fit. Our R-squared is very low which indicates that the model doesn't explain the variability around the mean well. Overall, we couldn't really conclude much from the linear regression since the assumptions were not met, and the statistics for the quality of our model were quite poor.

ANOVA: It is important to note that our data did not meet any assumptions for ANOVA, so we take those results with a grain of salt. Our ANOVA found that all of the education, income, and race demographics which had enough data were significant in determining water quality. We considered the possibility that the reason race is significant in predicting water quality is because of its interaction with other demographic variables in education and income. The two-way ANOVA proved this hypothesis wrong, since there was no evident pattern in interaction between race and other demographic factors. While some interaction existed, because we don't meet the assumptions for ANOVA the randomness of interactions could perhaps be attributed to that.

Kruskal-Wallis: To gain more statistically significant insights into the relationship between demographics and water quality score, we decided to try a non-parametric test which would work despite the fact that our data doesn't meet normality nor variance homogeneity requirements. This is the Kruskal-Wallis test. The conclusions from this test were the same as ANOVA. It found statistically significant differences in medians between groups of each column. Meaning depending on the group the median water quality changes, thus demonstrating that there is a relationship between water quality and level of the column. Confirming ANOVA results, we can now be more sure that education, income, and race demographics were statistically significant in determining water quality score.

Logistic Regression: Our data did not meet all the required assumptions when using the logistic regression model to predict the water quality risk level. Furthermore, we failed four of four assumptions: independence, linearity of independent variables, no perfect multicollinearity, and homoscedasticity. This lack of confidence in our assumptions are met by the metrics that we found in the model summary. The pseudo R-squared coefficient and the log-likelihood score are really low, indicating the model poorly fit the data. In terms of gauging the model complexity and model fit together, the AIC and BIC scores are extremely high. Overall, we can't conclude that a multinomial logistic regression model would be an accurate model to predict the water quality risk level. Note that logistic regression is in a separate jupyter notebook because it was done after our initial models.

ETHICS

An important consideration we wanted to address were the ethics of our datasets. We recognize that the data we are using and our results can affect people's lives. Our datasets do not collect personal information from specific individuals that could be traced back to them. All of our data was collected from public data sources in an effort to avoid using data that could compromise an individual's privacy.

Another aspect we would like to address is the biases we found within our datasets. Inside of our datasets we found a geographical bias. We found that for some counties there was no data provided and we discovered the possibility of more reported cases for certain counties and/or zipcodes. Therefore, we

mapped out the distribution of reports inside of our exploratory data analysis. Furthermore we found potential bias in how our data was composed. For our first dataset it contains self reported data, which can introduce bias that comes with human interaction. More severe and noticeable water issues were more likely to be reported than minor ones. The Census does a good job of collecting data (our second, fourth, and fifth datasets use data from the Census), but may not capture the whole population because some might not participate or be recorded properly. The census might not capture those "Hard-to-Reach Populations" that might have poor water quality. Furthermore, those without stable housing are less likely to participate in the survey. For our third dataset not all zip codes have a corresponding ZCTA, so we only work with the data that has both.

LITERATURE CITED

1. ^ Farzin, Y.H., Grogan, K.A. Socioeconomic factors and water quality in California. Environ Econ Policy Stud 15, 1–37 (2013). https://doi.org/10.1007/s10018-012-0040-8
2. ^ Sarah Acquah and Maura Allaire. "Disparities in Drinking Water Quality: Evidence from California." Water Policy, IWA Publishing, 1 Feb. 2023, iwaponline.com/wp/article/25/2/69/93385/Disparities-in-drinking-water-quality-evidence
3. ^ California State Water Resources Control Board - Division of Drinking Water. "Drinking Water - Safer Dashboard Failing and at-Risk Drinking Water Systems - Dataset - California Open Data." *CA.GOV OPEN DATA PORTAL*, 6 June 2024, data.ca.gov/dataset/safer-failing-and-at-risk-drinking-water-systems.
4. ^ Bureau, United States Census. "DP05 | ACS DEMOGRAPHIC AND HOUSING ESTIMATES." *United States Census Bureau*, 2020, data.census.gov/table/ACSDP5Y2020.DP05.
5. ^ Censusreporter. "ZIP Code to ZCTA Crosswalk." *GitHub*, 2022, github.com/censusreporter/acs-aggregate/blob/master/crosswalks/zip_to_zcta.
6. ^ HDPulse. "California Income - Table." *An Ecosystem of Health Disparities and Minority Health Resources*, 2021, https://hdpulse.nimhd.nih.gov.
7. ^ HDPulse. "California Education - Table." *An Ecosystem of Health Disparities and Minority Health Resources*, 2021, https://hdpulse.nimhd.nih.gov.