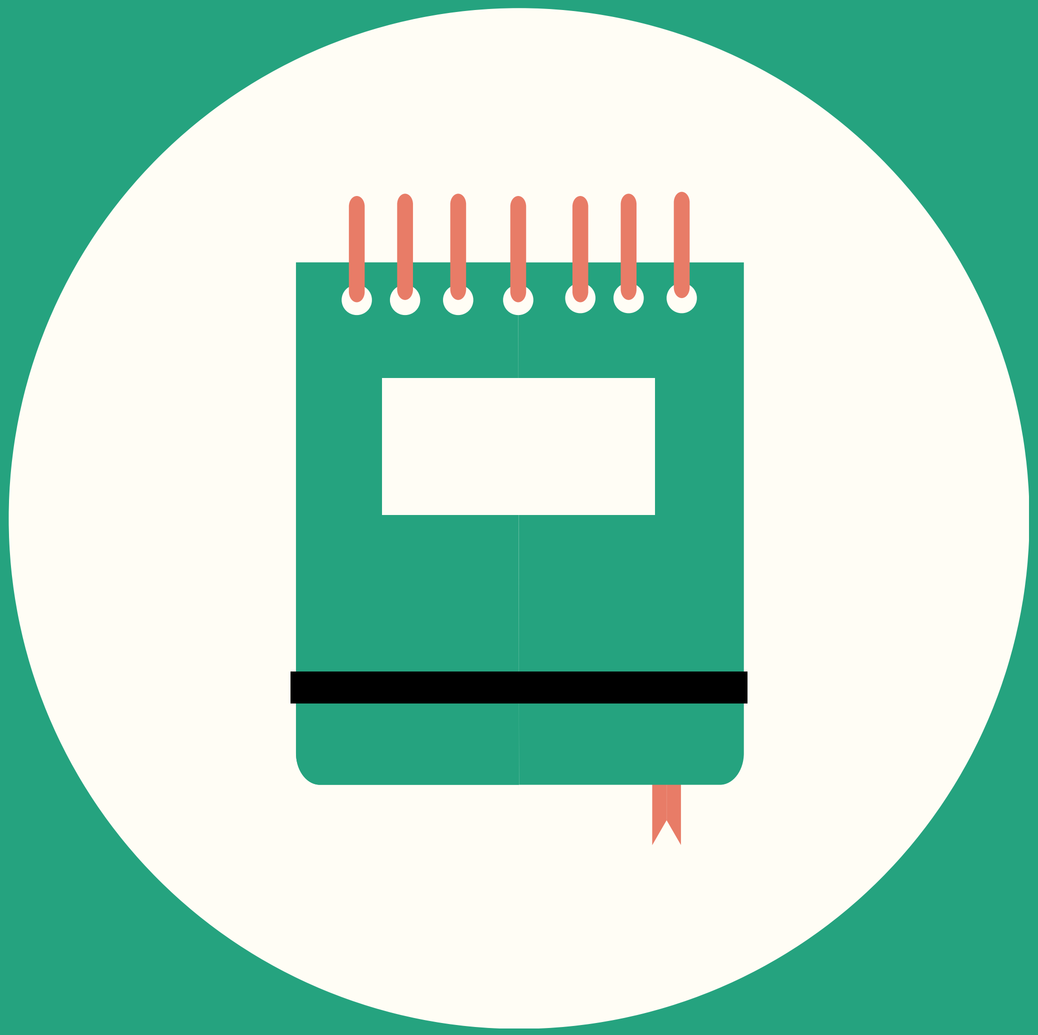
CORELATION PROJECT FINAL REPORT

Kevin(Baixuan) Tang, Nathan Zhao, Geroge Fang, Avery Cheng Mentor Bill Roth

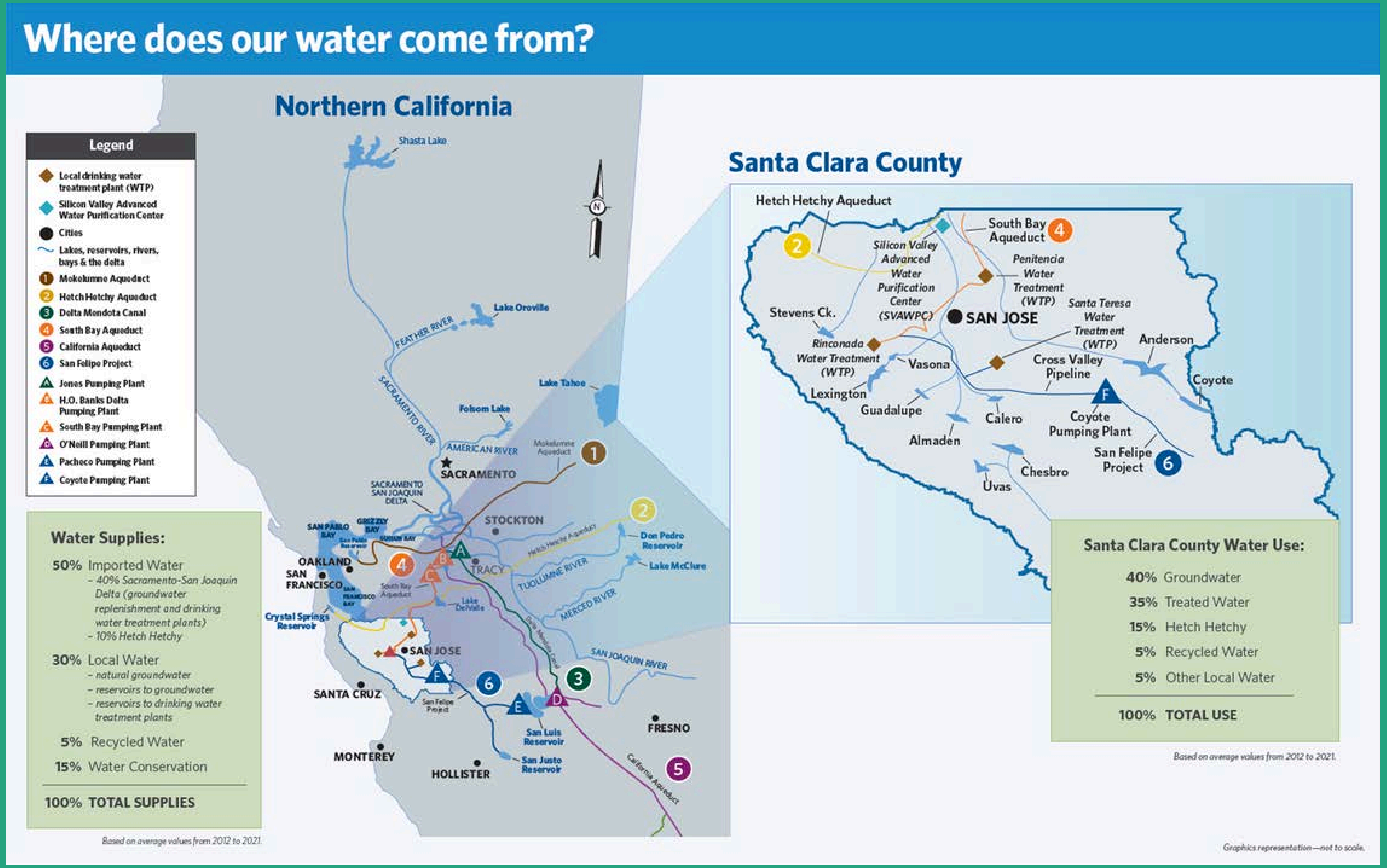
Table of Contents



Introduction

In this report, we compare two districts – one affluent and one lower-income – to explore how income levels relate to water quality. We use water quality metrics (specifically Absorbance at 499.7 nm, pH, Nitrate concentration, and Total Dissolved Solids (TDS)) collected from each district's water supply. Our focus is on identifying the correlation between income and water quality while being careful not to confuse correlation with causation. In other words, we investigate whether higher-income communities tend to have better water quality, and we discuss why that association does not necessarily mean that income causes the water to be cleaner. The goal is to present the findings clearly and accessibly for a general audience, highlighting key patterns and their implications.

The water supply comes from multiple sources across districts. For this research, it's important to note both districts draw from these shared and varied sources, so any quality differences may result from how they're treated and delivered rather than wholly different systems.



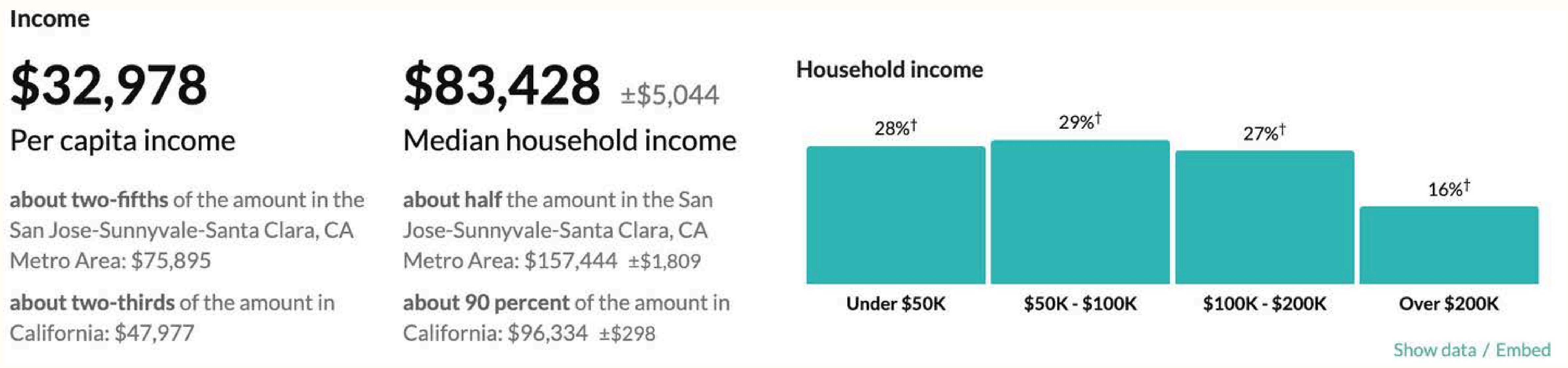
Water Quality Metrics Measured

To set the stage, we first outline the income distribution in each district and then summarize the water quality data from those areas. The two districts show a stark contrast in economic profiles, which provides the basis for examining differences in water quality. All water quality measurements were taken from local water sources in each district and include:

- Absorbance (499.7 nm): an indicator of water clarity and organic content (how much light at 499.7 nm the water absorbs – higher absorbance can indicate more color/turbidity due to dissolved organic matter or other pollutants).
- pH: the acidity/alkalinity of water (on a scale of 0-14, with ~7 being neutral; deviations can indicate contamination or treatment issues).
- Nitrates: concentration of nitrate (NO_3^-) pollutants in the water (high levels can come from agricultural runoff or poor sanitation and can be harmful to health).
- Total Dissolved Solids (TDS): the amount of dissolved substances (minerals, salts, organic matter) in the water (high TDS can affect taste and indicate overall water purity).

By comparing these metrics alongside the income data for each district, we can observe patterns that suggest a correlation between a community's economic status and the quality of its water.

District A: 95116

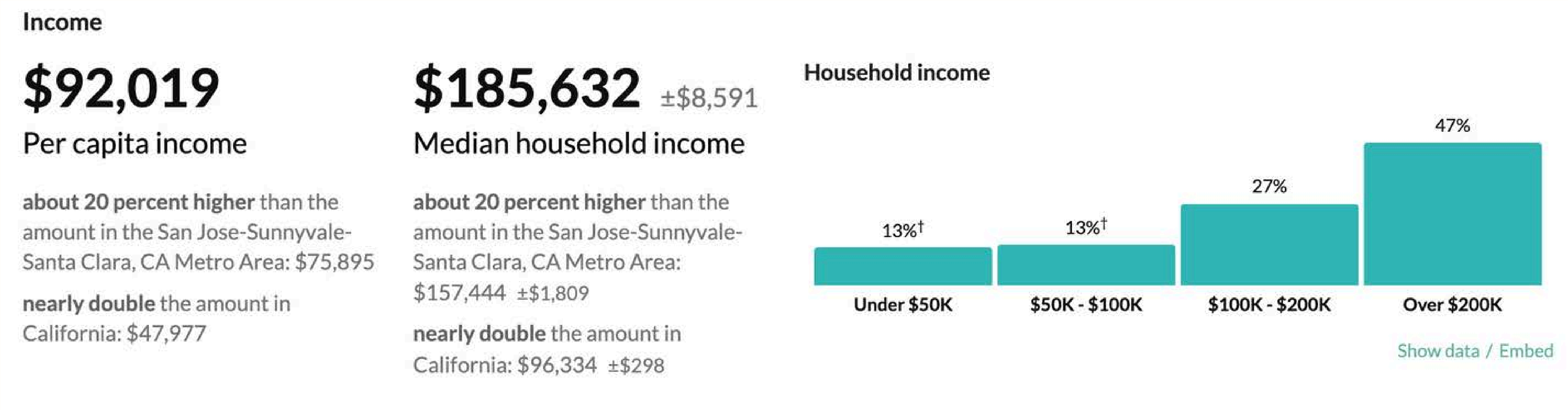


According to Census Data(Income distribution in District A), the lower-income area. District B, by contrast, is a lower-income community with a much more modest economic profile. The per capita income is roughly $33,000, and the median household income is about $83,000. These figures are less than half of those in District B, placing District A well below the metropolitan area's average income level (in fact, only about half of the regional median) and slightly below the state median. The income distribution chart for District A is far more weighted toward the lower and middle brackets: over half of households earn below $100k annually, with only about 16% of households exceeding $200k. This indicates a community that is considerably less affluent.

District A Water Analysis

Correspondingly, District A's water quality metrics were notably poorer. Water samples revealed higher absorbance at 499.7 nm, meaning the water was less clear – likely tinted or turbid due to a greater presence of organic matter or other pollutants. The pH readings in District A were somewhat less ideal; while generally not extreme, they showed more variability and tended to stray from neutral, hinting at possible issues (such as slight acidity) that aren't seen in District B's water. Nitrate levels in District A were elevated – higher than those in District B and approaching, or in some cases exceeding, recommended safe limits. This suggests that sources of nitrate pollution (perhaps agricultural runoff or aging septic systems) may be affecting District A's water supply. Additionally, TDS values were higher in District A, indicating a greater load of dissolved solids (minerals, salts, or other substances) in the water, which often correlates with lower overall water purity. (See Table 1)In short, the lower-income District A not only has fewer economic resources but also lower water quality across multiple indicators.

District B: 95051



According to Census Data, Income distribution in District B, the high-income area. District B is a wealthy district with a very high income profile. The per capita income is about $92,000, and the median household income* is around $185,000, which is significantly above regional and state averages. As the chart above shows, the majority of households in District B fall into upper income brackets – roughly three-quarters of households earn over $100k annually (with nearly a quarter of all households exceeding $200k). This skew toward higher incomes indicates a very affluent community.

District B Water Analysis

Such economic prosperity often goes hand-in-hand with well-funded infrastructure and services, and indeed, District B's water quality metrics reflect generally good water conditions. Water samples from this district showed low absorbance values (meaning the water was quite clear with little color or turbidity). The pH levels were close to neutral (around the ideal ~7), indicating balanced water with no extreme acidity or alkalinity. Nitrate concentrations in District B's water were found to be low, well within safe drinking water limits. Similarly, TDS readings were modest, suggesting relatively pure water with fewer dissolved solids or contaminants. (see table 1) In summary, the high-income District B not only has greater financial resources but also cleaner water by these measures.

Income–Water Quality Correlation Findings

Looking at the two districts side by side, a clear correlation between income levels and water quality emerges: the wealthier the community, the better the water quality metrics tend to be. District B (high-income) consistently outperforms District A (low-income) on every measured aspect of water quality. Here are the key observations that highlight this correlation(see table 4):

- Clarity (Absorbance 499.7 nm): District B's water had much lower absorbance readings, meaning the water was clearer and contained fewer light-absorbing impurities. District A's higher absorbance indicates more turbidity or contaminants in the water. In simple terms, higher income coincided with clearer water, whereas lower income coincided with murkier water.

- Nitrate Pollution: District B showed low nitrate concentrations, while District A had significantly higher nitrate levels in the water. This suggests that the lower-income area's water is more affected by pollutant sources (like fertilizer runoff or insufficient wastewater treatment). Thus, higher income community = lower nitrates; lower income community = higher nitrates

Income–Water Quality Correlation Findings

- pH Levels: The water in District B maintained a stable, neutral pH (around 7), which is ideal for drinking water. In District A, pH was observed to be less stable and occasionally outside the optimal range (though not extremely so; it hinted at minor water quality issues such as slight acidity or higher mineral content). This means the high-income area's water stayed closer to ideal conditions, whereas the lower-income area's water showed more signs of imbalance.

- Total Dissolved Solids (TDS): Measurements of TDS were lower in District B, indicating fewer dissolved substances and overall cleaner water. In District A, higher TDS pointed to more minerals and possibly contaminants dissolved in the water supply. So, wealthier District B had generally purer water, while District A's water contained more material.

All these points consistently paint the same picture: as income levels go up, water quality improves (or conversely, areas with lower income tend to have poorer water quality). The relationship was strong and consistent across multiple independent metrics. In the analysis conducted, a statistical correlation test was applied to quantify this relationship. The results showed a high correlation coefficient between the socio-economic status of an area and its water quality indicators, and this correlation was statistically significant (in fact, the analysis found the correlation to be significant at the 99% confidence level, $p < 0.01$). In plain language, this means it is very unlikely that the observed association is due to random chance – the data support a real underlying link between income and water conditions.

Interpretations of the findings&Future Action

While the correlation between higher income and better water quality in our two districts is evident, we must avoid concluding that “higher income directly causes cleaner water.” Since “Correlation ≠ Causation”, Income level itself is not likely the driving force—other extrinsic factors such as neighborhood infrastructure, housing quality, and past land use may explain why the two variables appear linked.

Examine treatment processes for each source. Document the treatment steps applied to each water source (filtration, chemical treatment, blending, etc.). Compare treatment quality and methods between sources feeding high-income vs lower-income districts.

Evaluate distribution and infrastructure differences. Investigate delivery systems: pipe age, maintenance records, length of transport, storage tanks, pumping stations.
Assess whether infrastructure in lower-income districts is older or less well-maintained, which might affect water quality.

Monitor and compare water quality by source and district. Collect water-quality data (absorbance, pH, nitrates, TDS) specific to each source and district over time.
Use this to identify which sources or delivery paths are associated with degraded water quality in lower-income areas.

Conclusion

Our comparison of the two districts demonstrates a clear relationship between socio-economic status and water quality for the two districts we observed. District B (the high-income area) enjoys clean and safe water across multiple metrics, whereas District A (the lower-income area) faces challenges with water clarity, pollutant levels, and overall quality. The correlation between higher income and better water metrics is evident and was statistically confirmed by the analysis.

For a general audience, the takeaway is straightforward: in this case, wealthier communities had cleaner water, while less affluent communities had more polluted water. This finding aligns with broader patterns that have been observed in many places – better-off neighborhoods often have superior infrastructure and environmental conditions.

Conclusion

However, it's crucial to stress that this pattern, while real, is not a justification to say the problem is income itself; rather, it suggests that inequities in resources, infrastructure, and power are likely at play. The poorer water quality in District A is a cause for concern and speaks to the need for improvements. Everyone, regardless of income, deserves access to clean and safe water. Therefore, one implication of these findings is that efforts should be made to invest in water system improvements in lower-income areas to close the gap. Policymakers and community stakeholders can look at these results and recognize that targeted action (upgrading pipes, better water treatment, stricter enforcement of standards, etc.) may be needed to ensure equitable water quality.

In closing, the analysis of income and water quality in these two districts provides a clear example of correlation: economic disparities coincide with environmental health disparities. It underscores an important point to the public: when we see such correlations, we should both acknowledge the relationship and ask deeper questions about underlying causes. By doing so, we can work toward solutions that address the root issues – ensuring that clean water isn't a luxury only for the wealthy, but a fundamental right for all communities.

Correlation Analysis Result

1. descriptive statistics

Table 1: Descriptive statistics

	N	Minimum	Maximum	Mean	Std. Deviation
pH	60	6.07	7.82	7.0827	.41436
Nitrates	60	.0	58.4	13.757	11.6817
TDS	60	120.33	367.00	247.5382	78.86349
Absorbance (499.7 nm)	60	.070	.244	.09833	.029906

This is mainly to understand the basic situation of the data. Subsequently, the data will be grouped, where 1 represents the 95116 region and 2 represents the 95051 region.

2. Analysis of variances

Table 2 Normality Test Table

	group	Kolmogorov-Smirnov			Shapiro-Wilk		
		statistic	df	sig.	statistic	df	sig.
pH	1	.174	30	.021	.925	30	.035
	2	.221	30	.001	.818	30	.000
Nitrates	1	.204	30	.003	.850	30	.001
	2	.221	30	.001	.823	30	.000
TDS	1	.145	30	.106	.847	30	.001
	2	.221	30	.001	.895	30	.006
Absorbance (499.7 nm)	1	.215	30	.001	.802	30	.000
	2	.151	30	.077	.886	30	.004

All other groups and testing methods (except for the K-S test in Group 1 of TDS and the K-S test in Group 2 of Absorbance) showed p-values below 0.05, indicating significant deviations from normal distribution. This suggests that non-parametric methods should be considered when conducting further statistical analysis.

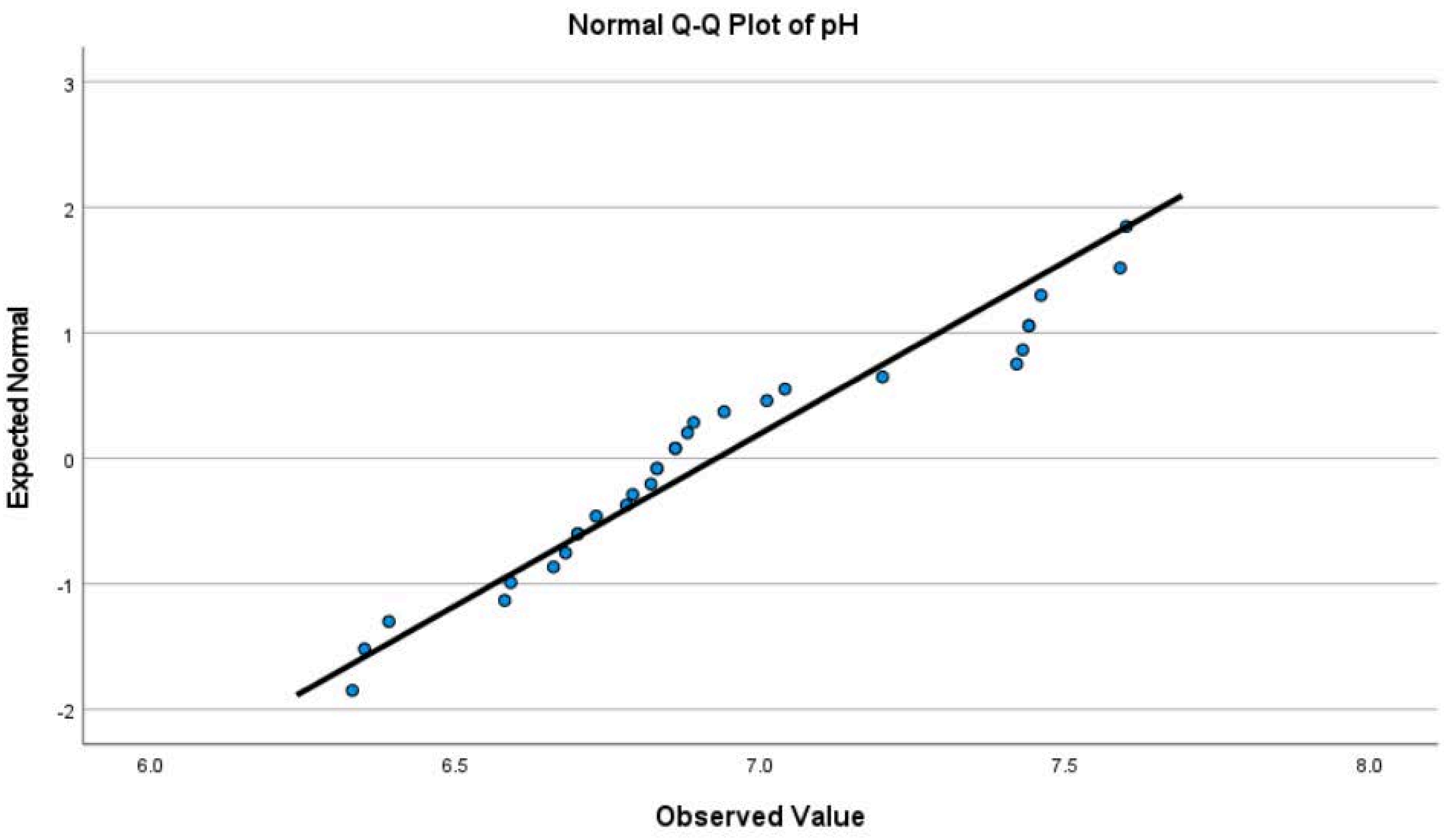

Normal Q-Q Plot of pH
Expected Normal
Observed Value
3
2
1
0
-1
-2
6.0
6.5
7.0
7.5
8.0

Figure 1 Normal Q-Q diagram of PH (other indicators are shown in the result file)

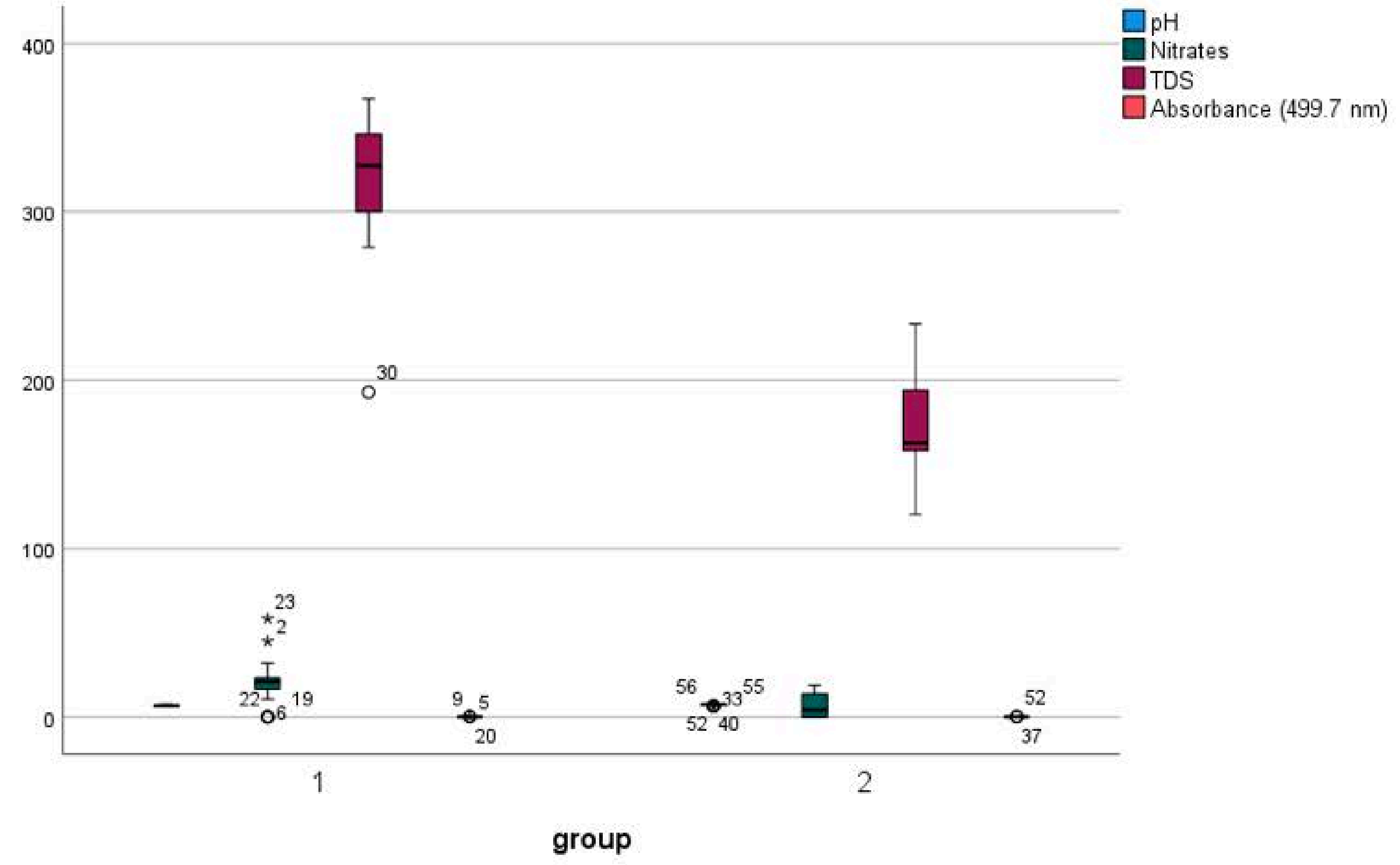


Figure 2 Box plot of each indicator

Absorbance (499.7 nm) showed a high median and wide distribution range in both groups, indicating that this variable was significantly variable in both groups.

pH, Nitrates and TDS showed lower median values and narrower distribution ranges in both groups, indicating that these variables were relatively stable in both groups with some outliers.

The difference between the two groups was more significant at Absorbance (499.7 nm), while the difference on other variables was smaller.

Table 3 Mann-Whitney U test results table

	group	Mean	Std. Deviation	Z-value	P-value
pH	1	6.9273	.36495	-2.973	0.003
	2	7.2380	.40768		
Nitrates	1	20.693	11.4227	-5.028	<0.001
	2	6.820	6.8992		
TDS	1	320.1333	33.92460	-6.514	<0.001
	2	174.9430	24.47277		
Absorbance (499.7 nm)	1	.10780	.039341	-1.739	0.082
	2	.08887	.009295		

In addition to absorbance, the other three variables (pH, nitrate content, and total dissolved solids) showed significant differences between the two groups. While pH levels demonstrated notable variations, their significance requires regional context analysis. Notably, high-income areas exhibited significantly lower nitrate and TDS concentrations compared to low-income regions, whereas absorbance differences remained insignificant. This indicates that under specific wavelengths, the spectral characteristics of the samples from both groups were comparable.

3. correlation analysis

Table 4 Results of Spearman correlation coefficient method

		group	pH	Nitrates	TDS	Absorbance (499.7 nm)
Spearman's Rho	group	1.000				
	pH	0.387**	1.000			
	Nitrates	-0.655**	-0.256*	1.000		
	TDS	-0.848**	-0.362**	0.581**	1.000	
	Absorbance (499.7 nm)	-0.226	-0.145	0.284*	0.215	1.000

**. At the 0.01 level (double tail), the correlation is significant.

*. At the 0.05 level (double tail), the correlation is significant.

In addition to absorbance (Absorbance), the other three variables (pH, nitrate Nitrates, and total dissolved solids TDS) showed significant correlations with the groupings. Specifically, pH exhibited a positive correlation with the groups, while nitrate and total dissolved solids demonstrated negative correlations with the groupings.

4. linear-regression analysis

Table 5 Summary of linear regression model table

model	R	R square	Adjusted R square	Std. Error of the	Revised statistics				
					R Square	F chan	df 1	df 2	Sig. F Change
1	.932[a]	.868	.859	.190	.868	90.67	4	55	.000

a. Predictors: (Constant), Absorbance (499.7 nm), pH, Nitrates, TDS

b. Dependent variable: group

The model's R value stands at 0.932, indicating a strong correlation between the predictors and dependent variable. The R^2 value of 0.868 demonstrates that the model accounts for 86.8% of the variance in the dependent variable. The adjusted R^2 value of 0.859, which accounts for the number of variables in the model, further enhances its validity. The standard error of estimate (SEE) is 0.190, reflecting the average difference between predicted and actual values. The model's F-change significance level reaches 0.000, confirming its overall statistical significance.

Table 6 Linear regression model ANOVA table

	model	Sum of Squares	df	Mean Square	F	Sig.
1	regression	13.025	4	3.256	90.674	.000b
	residual	1.975	55	.036		
	Total	15.000	59			

The ANOVA table presents the model's analysis of variance results. With a regression sum of squares (RSS) of 13.025, degrees of freedom (DOF) of 4, and mean square (MS) of 3.256, the F-statistic reaches 90.674 at a significance level of 0.000, demonstrating that the predictors in the model provide significant explanatory power for the dependent variable. The residual sum of squares (RSS) measures 1.975 with DOF of 55 and MS of 0.036, indicating unexplained variability within the model.

Table 7 Linear regression model coefficient table

model		Uncorrelation coefficients		standardized		
		B	Std. Error	Beta	t	Sig.
1	(constant)	2.629	.496		5.299	.000
	pH	.046	.064	.038	.722	.473
	Nitrates	-.003	.003	-.075	-1.217	.229
	TDS	-.006	.000	-.861	-13.509	.000
	Absorbance (499.7 nm)	-.506	.884	-.030	-.572	.570

The coefficient table displays the regression coefficients for each predictor variable. The constant term (intercept) stands at 2.629, indicating the predicted value of the dependent variable when all predictors are zero. The coefficient for TDS is-0.006 with a significance level of 0.000, demonstrating a significant negative impact of TDS on the dependent variable. In contrast, the coefficients for pH, Nitrate, and Absorbance all have significance levels greater than 0.05, suggesting their effects on the dependent variable are not statistically significant. Beta values represent standardized coefficients used to compare the relative importance of different predictors.

Table 8 Correlation statistics of linear regression model

model		Correlation statistics	
		tolerance	VIF
1	(constant)		
	pH	.862	1.160
	Nitrates	.628	1.592
	TDS	.589	1.697
	Absorbance (499.7 nm)	.870	1.149

The collinearity table provides the Variance Inflation Factor (VIF) and tolerance values for each predictor. A tolerance value close to 1 indicates low multicollinearity between variables. Typically, VIF values below 10 are considered acceptable for multicollinearity. In this model, all predictors show VIF values below 2, demonstrating no significant multicollinearity issues and ensuring reliable model estimation results.

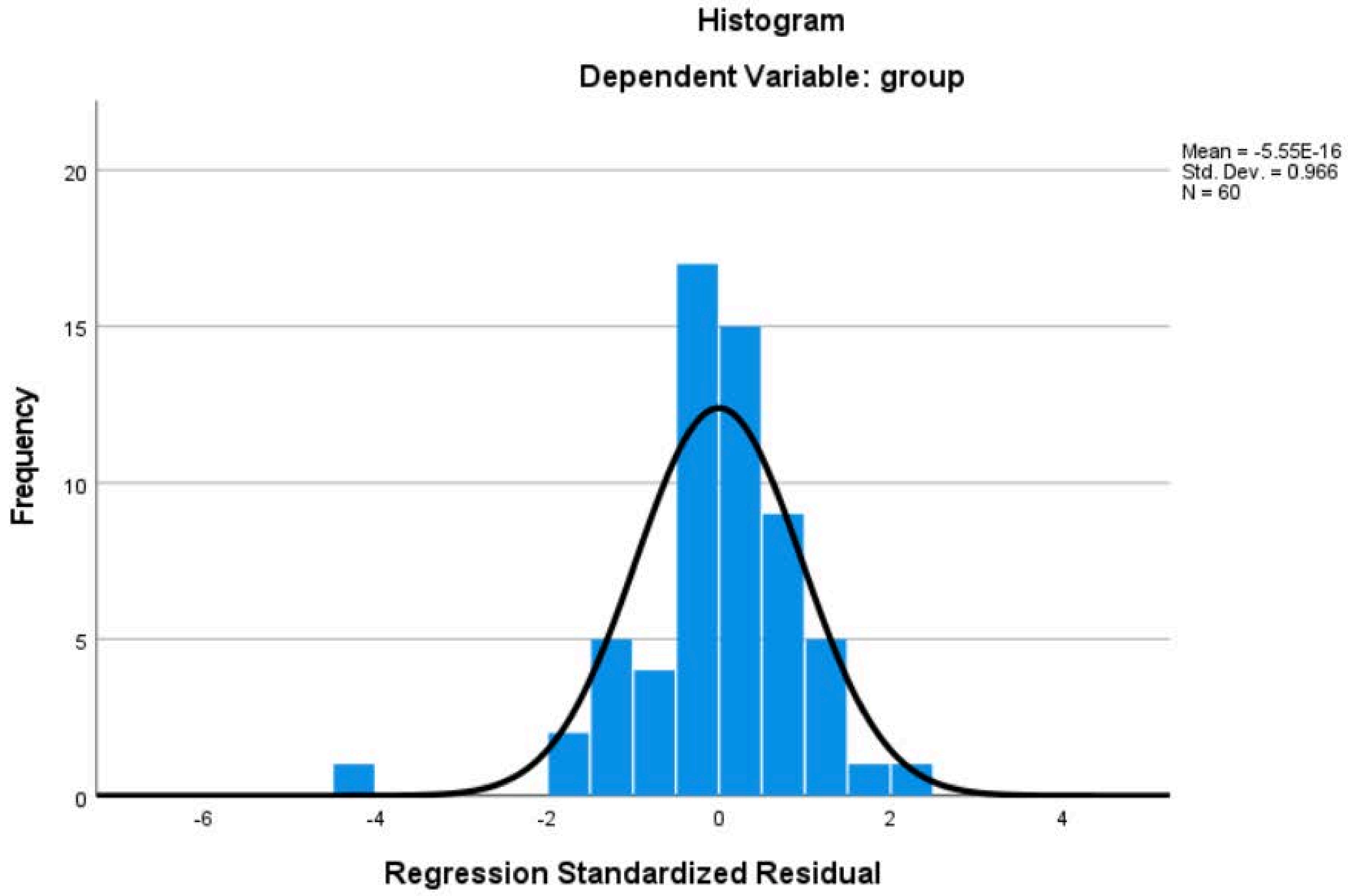


Figure 3 Linear regression residuals table

The residual distribution is normal, the normality hypothesis of the model is reasonable, and the results have reference opinions.

References

https://censusreporter.org/profiles/86000US95116-95116/

https://censusreporter.org/profiles/86000US95051-95051/

https://www.valleywater.org/your-water/where-your-water-comes