Report: Combating Waterborne Diseases: The Intersection of Public Health, Environmental Sustainability, and Economic Implications

Group: 10

Abstract:

This research focuses on the issue of waterborne diseases and their impact on health, specifically examining the role of bacteria in water contamination. By analyzing data and research including insights from Farnleitner et al. (2015) and Bartram & Gordon (2015) this study explores the prevalence and severity of waterborne diseases in developing countries. The investigation reveals that waterborne diseases pose a public health concern causing 560,000 severe cases and 12,000 deaths every year in the United States alone. Moreover, it investigates the aspects of transitioning to an economy as highlighted in UNEP's Nutrient Economics Report (2015) highlighting how sustainable water management can help reduce the occurrence of waterborne diseases. By considering health and economic perspectives this paper aims to provide a comprehensive understanding of the challenges associated with combating waterborne diseases while emphasizing the importance of clean water access and sustainable environmental practices. These findings emphasize the necessity for policies and interventions to protect health and promote environmental sustainability.

Introduction:

The issue of water quality and its impact on health is a major concern in today's world. The prevalence of diseases transmitted through water caused by harmful bacteria is a serious public health problem that affects millions globally. This study examines the connection between water contamination by bacteria and the resulting health consequences drawing insights from important research conducted by Farnleitner et al. (2015) and Bartram & Gordon (2015). Despite advancements in water treatment and sanitation, the persistence of waterborne diseases remains a challenge as pathogenic bacteria continue to pose risks to human health.

The current situation regarding diseases in the United States is alarming. According to the Centers for Disease Control and Prevention, 560,000 cases and 12,000 deaths are attributed to diseases every year highlighting the urgency of addressing this issue. This problem is not limited to the U.S.; it is a crisis with varying levels of severity and implications in regions around the

world. The complexity of this challenge is further compounded by environmental factors such as urbanization, industrialization, and climate change.

This study does not look at the health effects. Also delves into the economic aspects of managing water quality. The focus is on transitioning toward an economy as discussed in the UNEP's Nutrient Economics Report (2015). Understanding how environmental sustainability and economic considerations interact is crucial for developing strategies to combat waterborne diseases. The hypothesis put forward in this paper suggests that sustainable water management and environmental practices play a role in reducing both the occurrence and impact of diseases.

The objective of this research is to analyze the factors that influence water quality and the spread of waterborne diseases. It aims to provide insights into strategies for improving both water quality and public health outcomes. Moreover, it takes into account the feasibility and sustainability of these strategies. The findings from this study are expected to contribute to discussions on water resource management and public health benefiting policymakers, environmentalists as well as healthcare professionals.

Through an examination of data and existing literature this paper seeks to address a question; how do pathogenic bacteria in our water resources affect human health? Additionally, it explores the implications associated with managing this challenge. By identifying contributing factors to the prevalence of diseases viable solutions can be proposed for mitigating their impact on society. Our project goals include exploring the relationship between *E. coli* levels and different parameters and understanding if water type affects *E. coli* concentrations.

Found in the intestines of many mammalian organisms, *E. coli*, also known as *Escherichia coli*, is a type of bacteria that can potentially present great harm to the health of humans. Pathogenic strains can elicit symptoms associated with illness such as vomiting and diarrhea. Exposure to *E. coli* can occur through contact with bodies of water containing high levels of the bacteria or fecal matter which *E. coli* is an indicator. According to the Environmental Protection Agency *E. coli* can also be associated with high total suspended solids, and nutrient concentrations such as "high phosphorus, nitrate, and biological oxygen demand (BOD) concentrations." Exploring the relationship between these parameters and *E. Coli* can reveal information regarding the water quality. Focusing on the bodies of water located in Austin can highlight the bodies of water most contaminated by fecal matter, depict which water bodies are most impacted by *E. coli* concentration, and understand the relationship of this bacteria to

other parameters educated us on which environmental factors contribute the most to its presence, which can then lead to effective managing processes that protect water sources, and the individuals who might be swimming in them.

Literature Review:

Extensive research has been conducted on the importance of diseases and the role of bacteria in water contamination. These studies have provided insights into how they affect health and environmental sustainability. In this review, we will summarize the findings from existing literature focusing on the relationship between microbial contamination of water resources and human well-being.

Farnleitner et al. (2015) analyzed water contamination caused by bacteria highlighting the various sources and pathways through which these bacteria enter water systems. Their study emphasizes the need for monitoring and management of water quality to prevent bacterial contamination, which can lead to severe health issues such as gastrointestinal illnesses and other waterborne diseases.

In their work on water microbiology, Bartram and Gordon (2015) discuss the burden of waterborne diseases attributing a significant portion of these illnesses to bacterial pathogens. They emphasize the importance of understanding how these pathogens interact within water systems to develop strategies for disease prevention and control. Their research sheds light on the challenges faced by both developing countries when it comes to managing diseases transmitted through contaminated water.

Aside, from health considerations, the UNEP's Nutrient Economics Report (2015) explores the aspects associated with managing water quality. This report explores the aspects of transitioning towards an economy specifically focusing on nutrient management. It argues that managing water resources sustainably which includes reducing pollution and bacterial contamination is not only good, for public health but also economically viable.

Moreover, the UNEP (2015) report discusses the benefits of investing in improving water quality. It emphasizes the long-term advantages of investments in terms of healthcare costs and improved environmental sustainability.

In summary, existing literature emphasizes the importance of integrated approaches to managing water quality that take into account both health and economic considerations.

Addressing waterborne diseases caused by pathogens requires an approach involving robust monitoring, effective treatment technologies, public health interventions, and economic incentives for sustainable practices. This review lays the groundwork for an in-depth exploration of these topics, in sections of the paper.

In our study, on the effects of environmental factors on water quality, we rely on a regression model as our main analysis.

Data:

Data and Methodology;

In this study, we utilized a dataset gathered from sources to examine how socioeconomic and environmental factors affect the quality of water. The dataset consists of variables such, as pH levels, temperature levels, population density in areas, and the proximity of industrial activities to water sources. Here are the details regarding the data sources and the variables they provide;

Water Quality Data; We obtained this information from both the website of the City of Austin and their data library at AustinTexas.gov. This dataset includes data on water quality indicators, which can be found in the Water Quality Sampling Data section on AustinTexas.gov. This data looks at compiling information on different parameters of Austin's water sources. The data is composed of variables such as location, time the sample was collected, and the result of the concentration for each of the parameters.

AustinTexas.gov - Water Quality Sampling Data

Urban Population Data; We sourced this data from the United States Census Bureau specifically focusing on differentiating between rural areas. More information about this classification can be found on the Census Bureau's website under Urban and Rural Classification. This particular dataset helps us understand how population distribution and density impact water quality. Census Bureau - Urban and Rural Classification

Industrial Proximity Data; Although not explicitly mentioned in the provided links we assume that information regarding industries and their influence on water quality is either included in the water quality dataset from the City of Austin or sourced from environmental monitoring agencies.

Data Cleaning and Preparation;

A subset of random 100,000 observations was taken from the large Water Quality Sample Data which contained over 1.4 million entries – these were selected at random to ensure that there was no bias and proved to be an accurate representation of the larger data set, we can say that the data that was not represented is missing completely at random (MCAR).

Many of the variables in our data set did not contain relevant information regarding our desired research questions. For example, columns such as "SAMPLE_ID" and "SAMPLE_REF_NO revealed information that referred to a sample's identification, therefore this information was ignored as it did not reveal any pertinent information. Columns such as "SAMPLE_DATE" were split into separate variables including "Day", "Month", "Year" and "Time" to make data tidy, ensuring that every cell had only one observation. When selecting variables, for our regression model we need to consider factors such as, "Month", "Day", "Year", "Time", "Meridium", "WATERSHED", "Latitude", "Longitude", "SITE_TYPE", "PARAM_TYPE", "PARAMETER", "RESULTS" and "UNIT" but other columns were not removed entirely as they could still have contained relevant information on missingness.

In fact, missingness was calculated for every variable revealing that virtually every variable was complete except for "Longitude" and "Latitude" which were both only complete at 97.67%, as well as "RESULT" which was 99.98% complete. We first assumed that the missingness associated with Latitude and Longitude was related to an observed variable, potentially "PROJECT", which detailed the name of the Project that collected the sample – essentially hypothesizing that this was missing at random (MAR). To test this we made Longitude a missingness indicator variable assigning a binary to find which rows did and didn't have a value and visualized which PROJECT variables were associated with the missingness. We found that for other variables not just PROJECT, many variables contained missingness and the missingness was not solely attributed to one of our observed variables we concluded that our data was missing not at random (MNAR) for the Longitude and Latitude columns. This was also done for the RESULT variable which demonstrated that the 0.02% of missing data was also attributed to missingness not at random (MNAR). Considering that the latitude, longitude, and results columns detailed information that could not be replaced using another dataset we decided to ignore NA's for future data analysis. This posed one of the largest limitations in our study. Having done this the data were then prepared for analysis.

By using linear regression analysis we can quantify the impact of each variable on water quality. This will give us an understanding of how different factors contribute to the quality of water. The main goal of this study is to gain an understanding of how various factors influence water quality. This knowledge will be valuable in developing targeted policies and initiatives. We are incorporating data from sources to examine the factors affecting water quality in urban areas.

Exploratory Analysis:

Geospatial Analysis of *E. coli* Contamination in Water Sources

The mapping of *E. coli* concentrations in water bodies provides a picture of the water quality concerns in the studied region. The maps show the levels of *E. coli* in both Colonies/100mL and the Most Probable Number (MPN)/100mL, highlighting areas where bacterial contamination exceeds safety thresholds. Because these are two different ways of calculating *E. coli* presence each form was visualized separately.

The different visualizations for each method of collection were created due to outliers preventing the representation of smaller scales of *E. coli* concentration. To account for the outliers and paint a more articulate picture of *E. coli* concentration in Austin waterways we used the IQR method to create quartiles using the data's summary statistics. The distribution of *E. coli* samples was now more dispersed without the outliers and mapped.

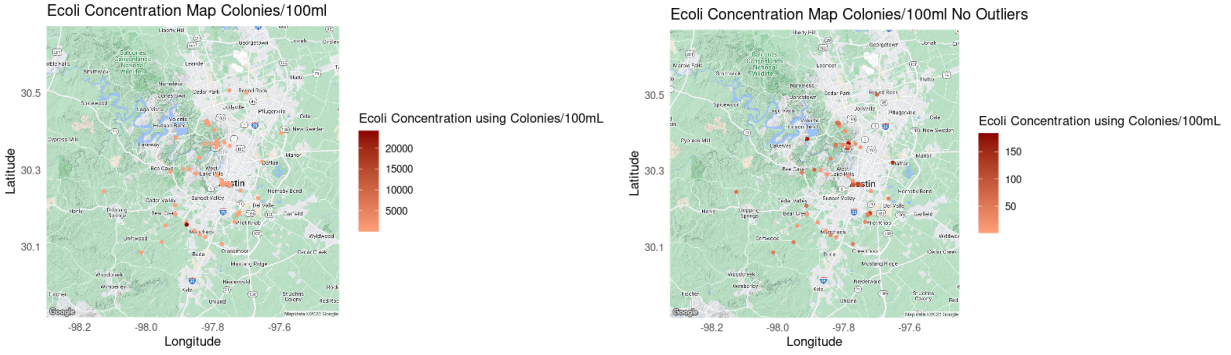

Figure 1 (a) & (b): Map of *E. coli* Concentration in Colonies/100mL; with and without outliers

This map displays the distribution and concentration of *E. coli* measured per 100 milliliters. The color gradient, ranging from pink to red, indicates increasing levels of contamination with areas indicating points that require immediate attention and remediation

efforts to address high concentrations. Due to the outliers, the initial maximum value on the scale goes from 20,000 to 150 colonies/100mL. In doing this we can take note that some of the points of interest at higher concentrations reside in bodies of water north central of the map.

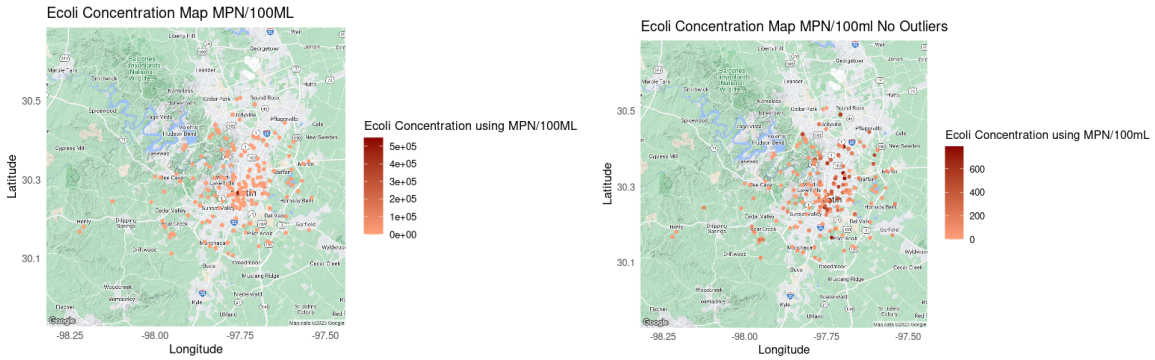

Figure 2 (a) & (b): Map of *E. coli* Concentration in MPN/100mL; with and without outliers

The MPN maps provide an estimate of E. Coli presence offering a measure of contamination. The first MPN map, with a scale up to 5e+05, identifies areas with levels of bacterial contamination that can pose significant risks to public health if not addressed promptly; there is one prominent outliner centered around downtown Austin. Meanwhile, the second MPN map uses a scale up to 750 MPN/100mL providing a visualization that ignores outliers. Here, we can see that concentrations of 600 or more MPN/100mL are found concentrated in the middle of the map appearing to centralize around the majority of downtown Austin.

These maps play a role, for public health officials, environmental agencies, and policymakers. They help identify locations where bacterial contamination levels are concerning. By using these maps we can better understand the need for targeted water treatment initiatives, monitoring, and infrastructure improvements to reduce the presence of *E. coli* in our water supply. When we integrate these maps into our analysis they provide a representation of the data. Highlight how environmental factors significantly influence water quality. Additionally, they show us how urbanization and nearby industrial activities can impact water contamination. By observing these patterns we can develop localized strategies to ensure water safety and protect health. These efforts align with our goals of promoting sustainability and community well-being.

<u>Hypothesis 1 - Watershed Temperature and E. Coli Concentrations</u>

For our first research question, we focused on looking at the relationship that Water Temperature has with the concentration of *E. coli* at the watershed locations for the city of Austin. We used different bar graphs to plot the concentrations of each of these parameters to see if there was a common location where *E. coli* presence was associated with the mean temperatures of these bodies of water. We hypothesized that Water Temperature levels are related to the concentration of *E. coli*.

In regards to this data, it had parameters other than the needed water temperature and pathogens so we filtered the data to contain only temperature levels and *E. coli* bacteria. With this new result, we removed any NAs and also removed any outliers using the IQR method, where the average and quartiles of each of the points are found, only keeping those points inside of the quartiles and removing the outliers. This would help us look at the data without it being skewed. On top of that, we also filtered out any non-finite values. Following this, we can look at data without outliers. After the data is grouped by the parameter and the watershed, the averages of the water temperature levels for each of the parameters and watershed combinations. In doing this we pivoted wider and created a graph with every observation:

WATERSHED <chr>	E COLI BACTERIA <dbl>	WATER TEMPERATURE <dbl>
Barton Creek	12.17290	20.92009
Bear Creek	15.34524	18.70714
Bear Creek West	37.90000	21.01667
Bee Creek	18.46667	21.52385
Boggy Creek	25.50000	20.86158
Brushy Creek	8.50000	19.45000
Bull Creek	18.42162	19.58987
Commons Ford Creek	7.50000	12.08000
Cottonmouth Creek	24.60000	20.56500
Cuernavaca Creek	38.40000	17.41000
Decker Creek	15.80000	18.96000
Dry Creek East	21.30000	23.98000
Elm Creek	9.95000	10.53500
Fort Branch	10.00000	18.31091

Using the water temperature levels, we created a graph displaying the water temperatures across different watersheds in Austin and created a linear regression for these levels.

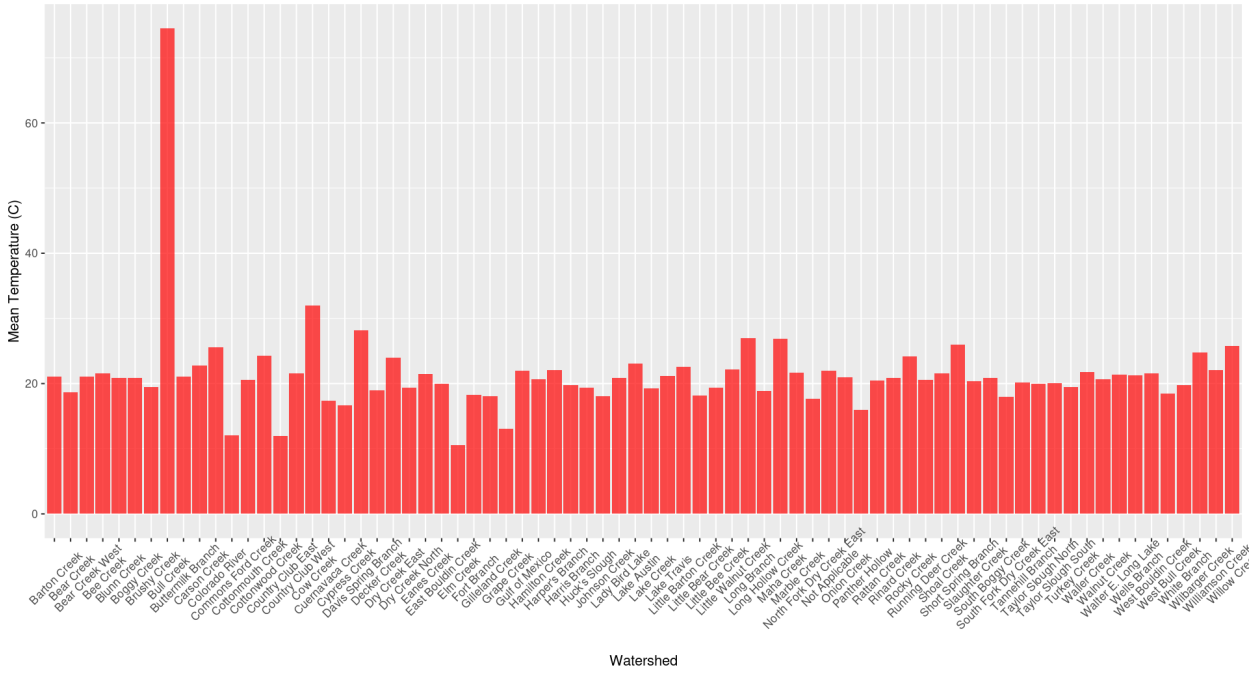

Figure 3: Mean Temperature in various Watersheds around Austin

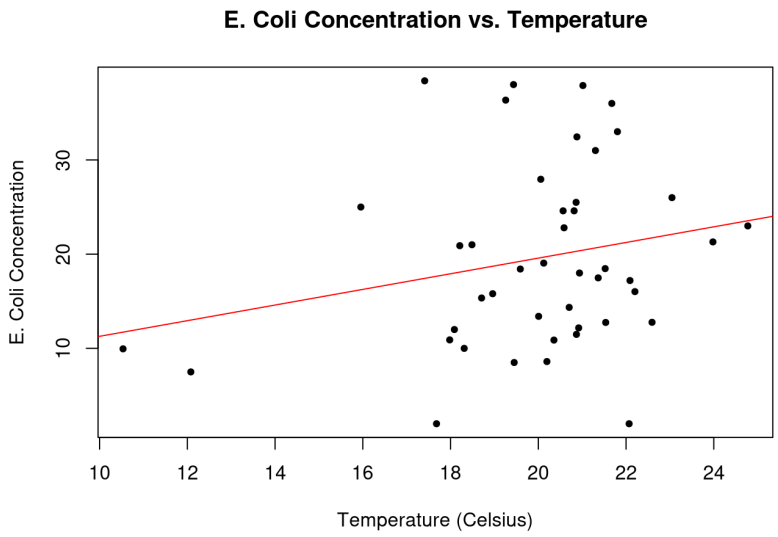

Figure 4: *E. coli* vs. Temperature

Table 1: Regression and R^2 values for the Temperature

Regression value	R^2 value
0.8302	0.02678

Through Figure 2 we can see how there seems to be a positive relationship between water temperature and the amount of *E.coli* concentration. However, the R^2 value is very low and also has a low p-value (< 0.05). We were unable to confirm if water temperature altered the amount of this bacteria and thus we rejected our hypothesis.

Adding onto the recurring issues with the data, when the data was cleaned, filtered, and analyzed, we saw that the number of watersheds observed significantly decreased and resulted in a smaller data pool. Due to this, our R^2 value may have been lower than expected and the line of best fit was not the best in accurately capturing the effect temperature may have.

<u>Hypothesis 2 - The relationship between the concentration of Nutrients and E. Coli by location and year</u>

For our second hypothesis, we decided to find out if phosphorus concentration had an interaction with the amount of *E. coli*, we hypothesized that phosphorus or nitrogen-containing nutrients would be related to *E. coli* concentrations.

Given that the data contained various parameters outside of the nutrients and pathogens we filtered for only nutrients and *E. coli* bacteria. After we removed the NAs from the results column accounting for 0.02% of our data we then cleaned the data of outliers using the IQR method, where the average and quartiles of each of the points are found, only keeping those points inside of the quartiles and removing the outliers. By doing this we can look at data without outliers. After the data is grouped by the parameter and the watershed, the averages of the concentrations for each of the parameters and watershed combinations. In doing this we pivoted wider and created a graph with every observation:

WATERSHED <chr>	E COLI BACTERIA <dbl>	ORTHOPHOSPHORUS AS P <dbl>	PHOSPHORUS AS P <dbl>
Barton Creek	1.848387	0.02445034	0.07624722
Bear Creek	1.333333	0.01852174	0.06661715
Bull Creek	1.250000	0.09264545	0.13192428
Harper's Branch	2.000000	0.04485714	0.08006667
Lady Bird Lake	1.722857	0.02528110	0.15091838
Lake Austin	1.000000	0.01533333	0.02575385
Little Walnut Creek	0.000000	0.04000000	0.02333333

Using nutrient parameter concentrations containing phosphorus, such as “ortho-phosphorus as p” and “phosphorus as p”, we created linear regressions for the phosphorus-containing nutrients. Linear regressions were also created for other nutrients known to also alter *E. coli* growth such as Nitrate/Nitrite and Ammonia, nutrients which have not been established to affect *E. coli* growth.

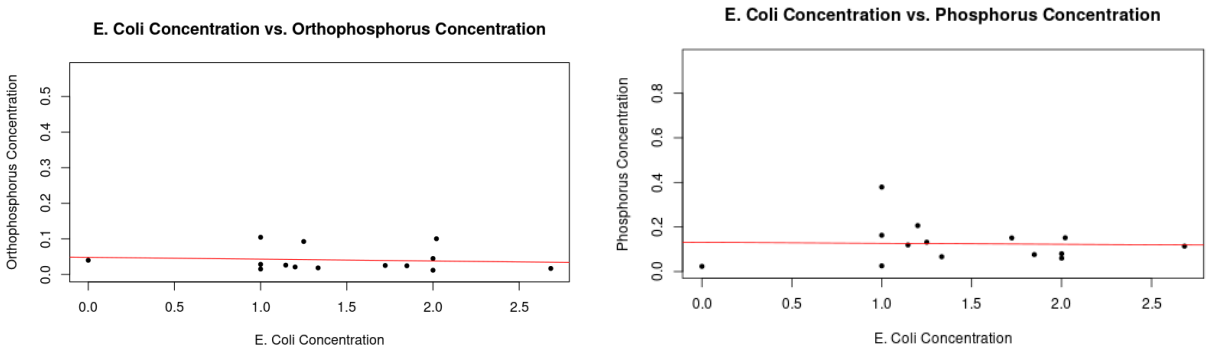

Figure 5 (a) & (b): Using the same watershed locations

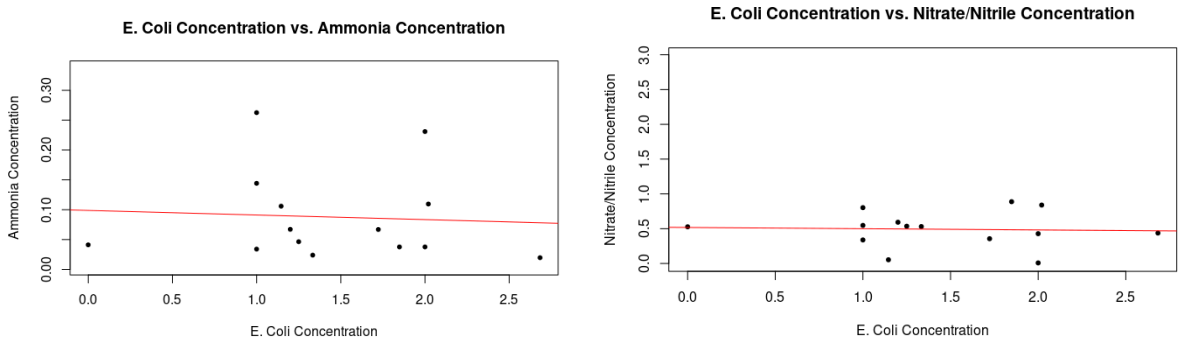

Figure 5 (c) & (d) Using the same watershed locations

Table 2: Regression and R^2 values for the selected nutrients

Nutrient	Regression value	R^2 value
Orthophosphorus	-0.005139	0.01046834
Phosphorus	-0.004259	0.0009548915
Ammonia	-0.00768	0.004327044
Nitrate/Nitrite	-0.01793	.002074765

Given that our data revealed that across the different watersheds, the relationship between all nutrients was slightly negative we are unable to confirm if any nutrients altered the amount of *E. coli* concentration and we reject our hypothesis. We suspect that given the low R^2 value because we lacked an abundance of data points and our line of best fit was not successful at capturing the points accurately this limitation prevented us from providing a more accurate image of Autin watershed parameter concentrations and their interaction order to see if the limitation of having only a few data points for each watershed contributed to the low R-squared value we instead grouped our data by year rather than watershed.

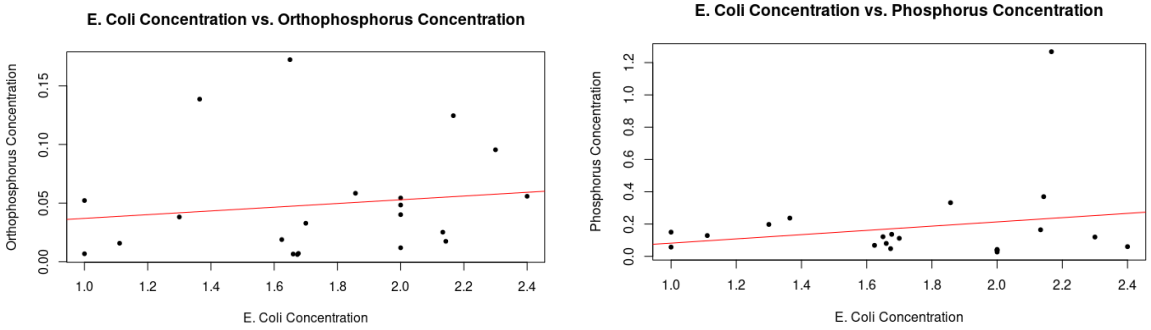

Figure 6 (a) & (b) Using year to group data

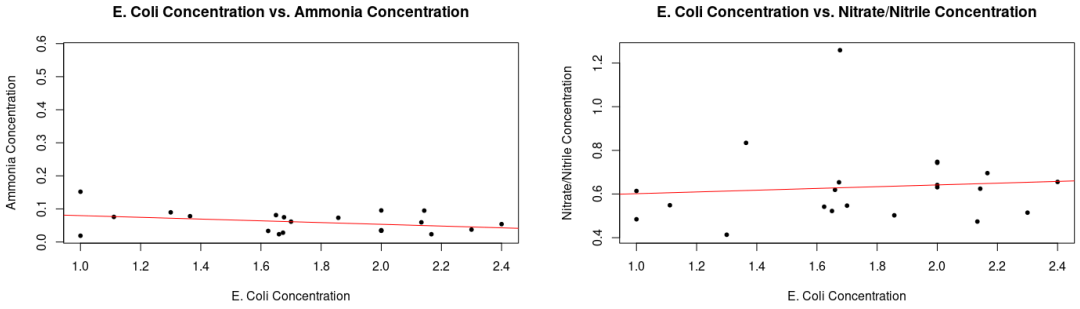

Figure 6 (c) & (d) Using year-to-group data

Table 3: Regression and R^2 values for the selected nutrients

Nutrient	Regression value	R^2 value
Orthophosphorus	0.01587	0.01950183
Phosphorus	0.13153	0.04148897
Ammonia	-0.02638	0.1083082
Nitrate/Nitrite	0.0402	0.00877698

In these results, we obtained a better R-squared value across each of the different graphs, while still not significant we were able to assess that the regression lines fitting by each year were a better fit than by watershed. The value of the linear regression revealed that there was a slight positive relationship for the nutrients Orthophosphorus, Phosphorus, and Nitrate/Nitrite as seen in **Table 3** with ammonia having a very small negative correlation. Grouping the data by

year revealed that there was a slightly positive correlation between phosphorus and nitrogen-containing nutrients, however, these results are limited as they don't reflect statistical significance and there are only a few data points thus limiting our findings.

Hypothesis 3 - Association between flowing water status and E. Coli concentration

For our third research question, we wanted to test whether or not the presence of flowing water affected the concentration of *E. coli* in a given body of water. To do this, we first classified each water sample into either a flowing category or a still category, depending on the body of water it was collected for, according to the following table:

Table 4: Classifications for flowing water status

Flowing	Still
Stream	Lake
Spring	Sediment
Rural Runoff	Well
Urban Runoff	Detention Pond
Cave Stream	Wet Pond
Storm Drain	Drainage Swale
	Irrigation Pond
	Cave Drip

Next, we determined the relative frequency of samples with the *E. coli* parameter in both the flowing and still groups, and plotted them in a bar plot like so:

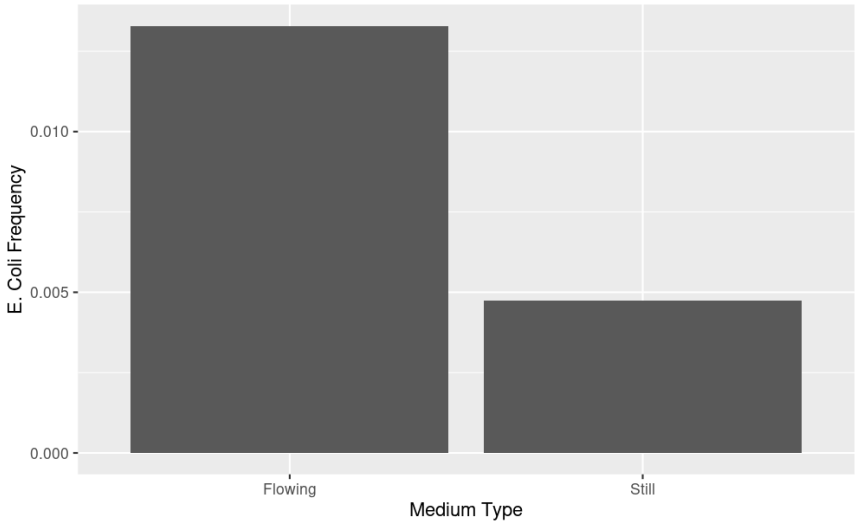

Figure 7

Finally, we conducted a hypothesis test as follows:

Let us construct our null hypothesis to be that the relative frequency of *E. coli* in flowing water is equivalent to that of *E. coli* in still water. In other words, our null hypothesis H_0 is described by the mathematical expression

$$H_0\colon\ Pr(E.Coli \mid Flowing) = Pr(E.Coli \mid Still)$$

where $Pr(E.coli \mid Flowing)$ denotes the probability that a sample collected from flowing water was one of *E. coli*. Rewriting the null hypothesis as an equality of proportions, we have that

$$H_0\colon\ \frac{Pr(E.Coli \mid Flowing)}{Pr(E.Coli \mid Flowing) + Pr(E.Coli \mid Still)} = \frac{Pr(E.Coli \mid Still)}{Pr(E.Coli \mid Flowing) + Pr(E.Coli \mid Still)}$$

or

$$H_0\colon\ p = \frac{Pr(E.Coli \mid Flowing)}{Pr(E.Coli \mid Flowing) + Pr(E.Coli \mid Still)} = .50$$

where p denotes the probability that an *E. coli* sample taken from a dataset with an equal number of flowing and still samples is a flowing sample. Let p_f and p_s denote the relative frequency of *E. coli* samples in flowing and still water respectively, gathered from our data. Then we can calculate a sample proportion $\hat{p}$ to estimate p as follows:

$$\hat{p} = \frac{p_f}{p_f + p_s} = \frac{0.0133}{0.0133 + 0.0047} = 0.7370.$$

Since our sample size is sufficiently large (n=100,000), we can invoke the Central Limit Theorem to treat $\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ (where $p_0 = 0.50$ is our hypothesized value for p according to H_0) as approximately standard normal. Therefore, our test statistic is given by

$$Z \approx \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.7370-0.50}{\sqrt{\frac{0.50\times0.50}{100{,}000}}} = 149.892.$$

Using a significance level of $\alpha = 0.01$, and an alternative hypothesis of $H_A: p > 0.50$ our rejection region is given by

$$Z > z_{0.99} = 2.327.$$

Since our test statistic is in the rejection region, we can reject our null hypothesis with 99% confidence in favor of the alternative hypothesis that the relative frequency of *E. coli* in flowing water is higher than that of still water. Thus, we can safely assume that the presence of flowing water does have a strong effect on the concentration of *E. coli* in a given body of water.

Hypothesis 4 - Exploration of pH on E. Coli concentration

For our last research question, we focused on looking at the relationship that pH has with the concentration of *E. coli* at the watershed locations for the city of Austin. We used different bar graphs to plot the concentrations of each of these parameters to see if there was a common location where *E. coli* presence was associated with the mean pH of these bodies of water. We hypothesized that pH levels are related to the concentration of E. *coli*.

In regards to this data, it had parameters other than the needed pH and pathogens so we filtered the data to contain only pH levels and E. *coli* bacteria. With this new result, we removed any NAs and also removed any outliers using the IQR method, where the average and quartiles of each of the points are found, only keeping those points inside of the quartiles and removing the outliers. This would help us look at the data without it being skewed. On top of that, we also filtered out any non-finite values. Following this, we can look at data without outliers. After the data is grouped by the parameter and the watershed, the averages of the concentrations pH levels for each of the parameters and watershed combinations. In doing this we pivoted wider and created a graph with every observation:

WATERSHED <chr>	E COLI BACTERIA <dbl>	PH <dbl>
Barton Creek	7.304853	7.466307
Bear Creek	9.812000	7.690909
Bee Creek	10.350000	7.841500
Brushy Creek	8.500000	7.036000
Bull Creek	7.450000	7.648682
Commons Ford Creek	7.500000	7.386667
Dry Creek East	9.800000	7.898000
Elm Creek	9.950000	7.620000
Fort Branch	10.000000	7.830000
Gilleland Creek	12.000000	8.053182
Lady Bird Lake	8.520909	7.643837

Using the pH levels, we created a graph displaying the pH levels across different watersheds in Austin as well as a linear regression for these levels.

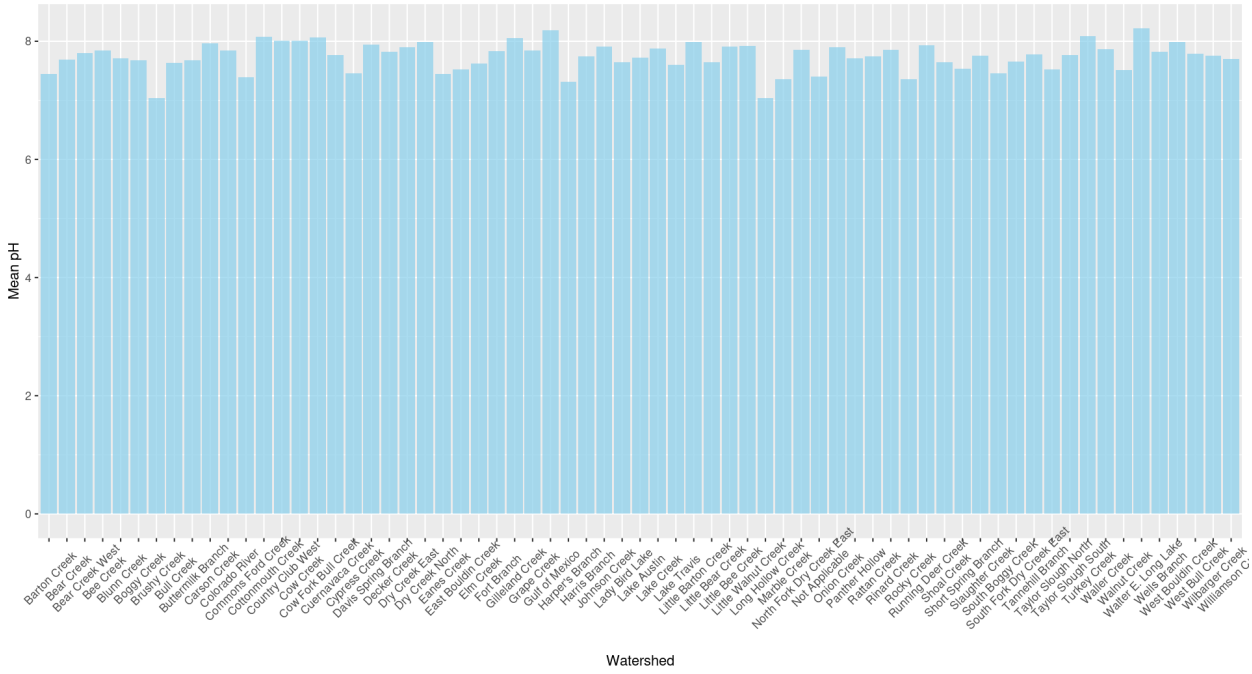

Figure 8: Mean pH in various Watersheds around Austin

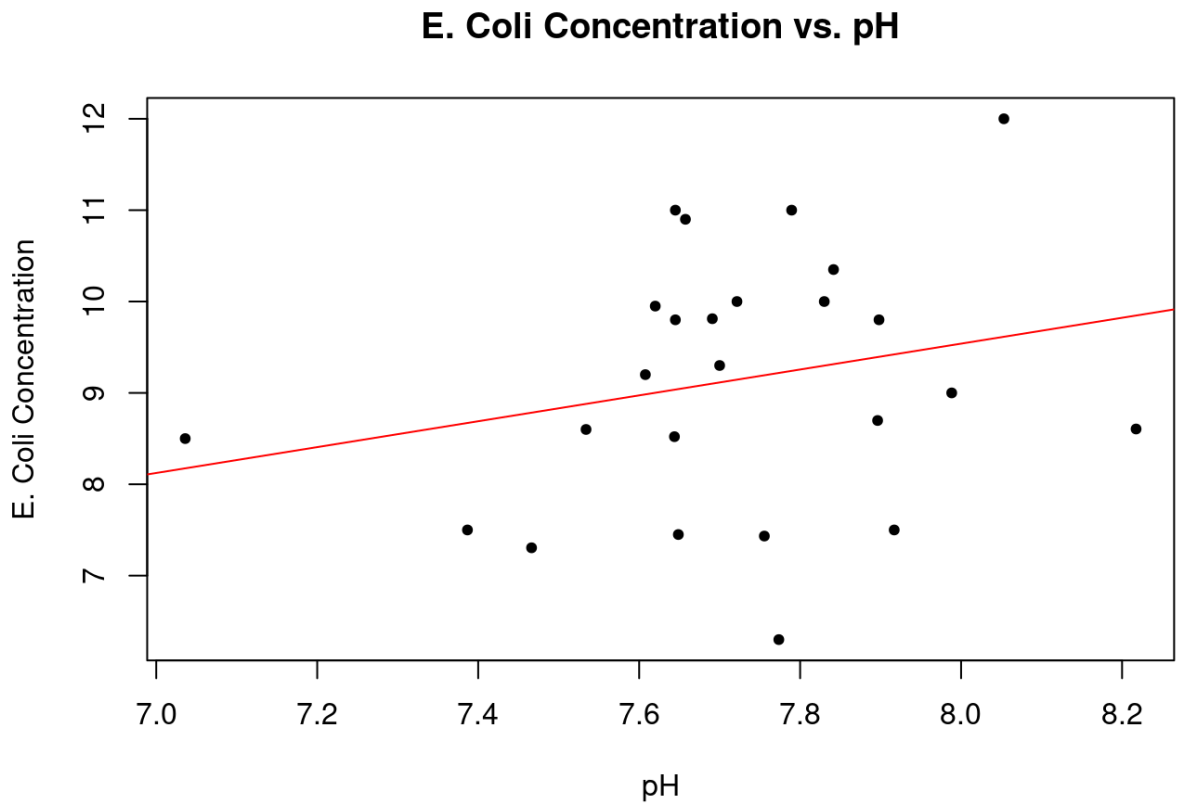

Figure 9: *E. coli* vs. Temperature

Table 5: Regression and R^2 values for pH

Regression value	R^2 value
1.416	0.01378

Through Figure 2 we can see how there seems to be a positive relationship between pH and the amount of *E.coli* concentration. However, the R^2 value is very low and with a low p-value as well (< 0.05). We were unable to confirm if pH altered the amount of this bacteria and thus we rejected our hypothesis.

Seeing that the average pH of water is 7 and that water is known to be very resistant to changes in its water pH levels, we saw that there was a very small range for the pH levels. On top of that, there was the issue of a lack of data points which resulted in a smaller data pool. Due to this, it could be possible that our line of best fit was unable to capture the relationship.

Modeling:

Approach to the task

Initially, we wanted to explore the relationship between different nutrients (Orthophosphorus, Nitrate/Nitrite, and Ammonia) and *E. coli*. To tailor the data to the task, we filtered the data to only contain observations with either the *E. coli* bacteria or one of our nutrients of interest for the parameter type. Then we pivoted wider, creating a new column for E.coli and each nutrient. Rows with the same sample number were combined, so each sample was encapsulated in one row with its corresponding data for *E. coli*, nutrients, and other variables. Finally, we removed any NAs and outliers in the rows for *E. coli*, Orthophosphorus, Nitrate/Nitrite, and Ammonia.

As we were initially focused on only numeric variables, it seemed beneficial to use a regression model. However, after excluding outliers, scatterplots of *E. coli* and the different nutrients did not appear to have a linear relationship (Figure 10). The variables also did not seem to have a linear relationship when scatterplots were made using log-transformed data (Figure 11). Linear regression assumes a linear relationship between the independent variables and the outcome variable. Therefore, we decided that linear regression modeling would not be the best way to approach this task.

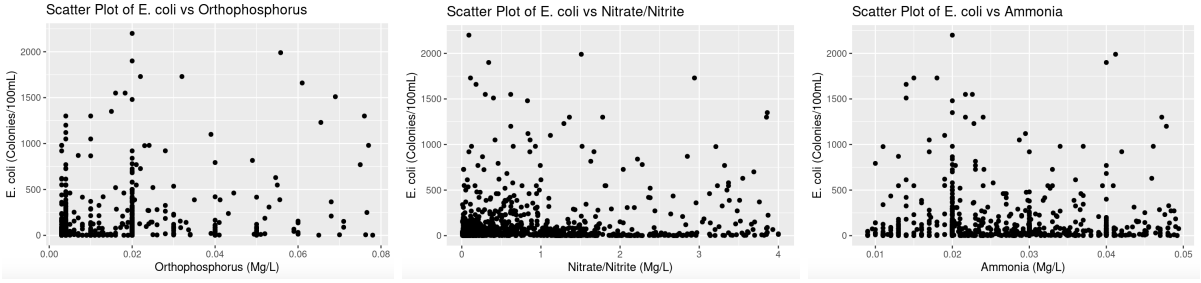

Figure 10: Scatterplots of *E. coli* (Colonies/100mL) versus nutrients (Mg/L).

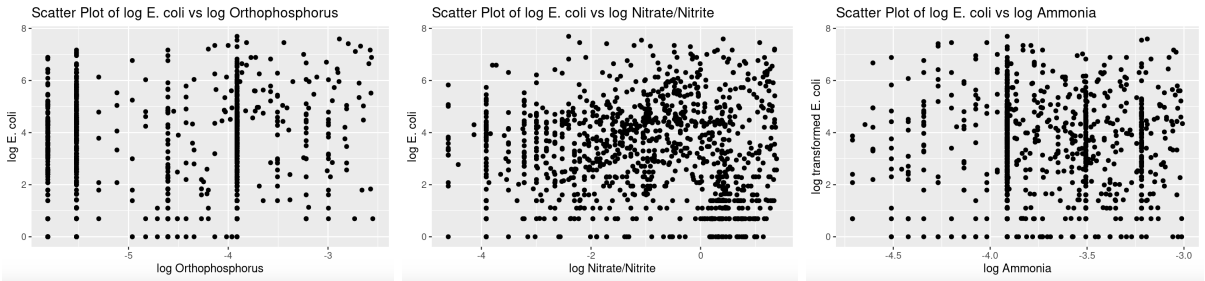

Figure 11: Scatterplots of the log-transformed *E. coli* (Colonies/100mL) data versus log-transformed nutrient (Mg/L) data.

Problem setup, methods, and algorithms

We subsequently shifted to a classification task. The numeric *E. coli* variable was classified into three ordinal categories: low, medium, and high. *E. coli* was classified as 'low' for observations where the numeric *E. coli* value was less than or equal to 10; 'medium' when the numeric value was greater than 10 and less than or equal to 100; and 'high' when the numeric variable was greater than 100. The three independent variables for nutrients (Orthophosphorus, Nitrate/Nitrite, and Ammonia) were kept numeric. Another advantage of a classification task is its ability to handle categorical variables. Our classification model also used a categorical variable for flowing water. The flowing variable was added to the data frame; each water sample was classified as either flowing (1) or not flowing (0) depending on the site type, following the same rules for classification used in testing Hypothesis 3.

Ammonia, Nitrate/Nitrite, Orthophosphorus, and flowing were used in the model as independent variables. The categorical *E. coli* variable was used as the output variable. Variables were split into training and test sets, with a 0.4 test size. Using the same training and test sets, five different kinds of classifiers were trained to determine which had the highest accuracy (Table 6). Classifiers were trained using the training subset of data. Then the classifiers were scored based on how accurately they were able to predict the test outcome based on the test data independent variables.

Table 6: Classification models trained and their accuracy scores

Classifier:	**Accuracy:**
Decision Tree	0.5391
SVM	0.4743
kNN	0.4787
Logistic Regression	0.4452
Random Forest	0.5168

The decision tree classifier had the highest accuracy. Therefore, the decision tree classifier was used for the rest of the classification task. A decision tree is a rules-based classification scheme.

One input variable determines the rule used to split the data at each node of the tree. Data is successively divided into many groups until all groups uniformly consist of observations with the same outcome variable or until the tree reaches a predetermined maximum depth.

Discussion

Classification Model Evaluation:

To maximize the accuracy of the Decision Tree classifier, the max_depth hyperparameter was evaluated. Cross-validation was done for max_depths values of 1-20. The highest mean cross-validation accuracy score was for a max_depth value of three (Figure M).

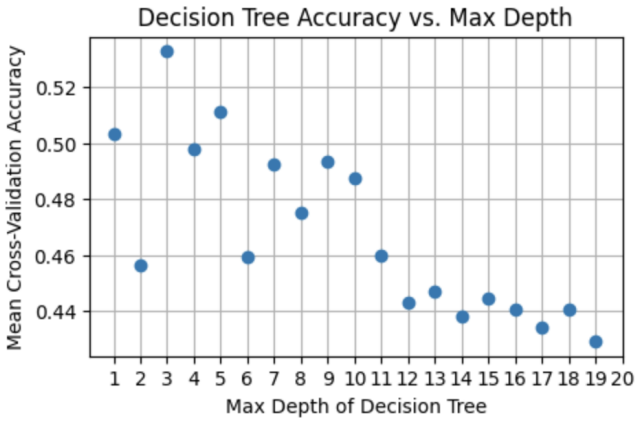

Figure M: Plot showing the mean cross-validation accuracy scores for max_depths values 1-20.

The model was trained using the training subset using a decision tree classifier (Figure N). Then, the decision tree classification model was tasked to predict test outcome variables based on independent variables from the test subset. The decision tree max_depth was set to three. To evaluate the model's performance, outcome variables predicted by the model were compared to the actual outcome variables of the test subset. The accuracy of the decision tree classification model is 0.539 or about 54%.

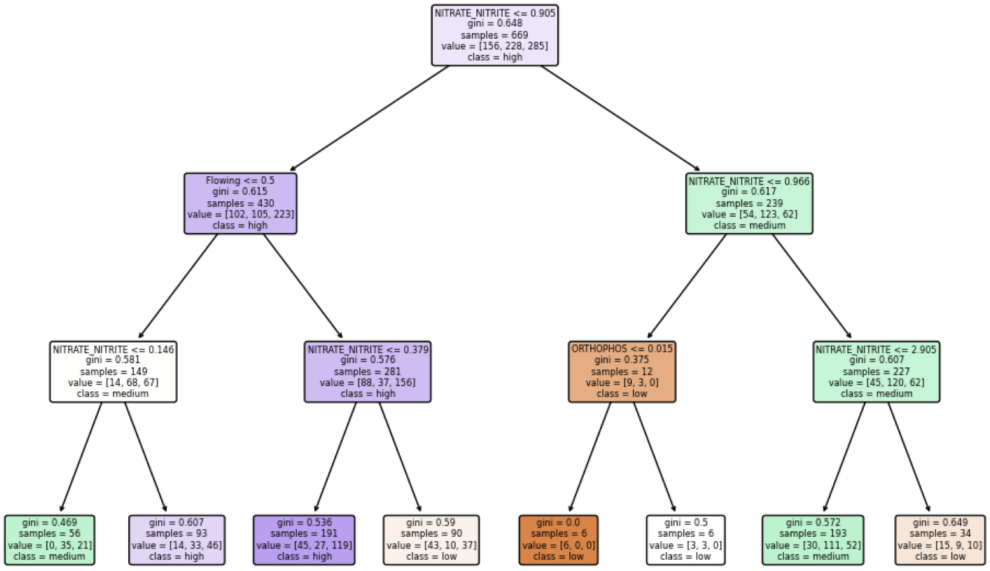

Figure N: Visualization of the decision tree classification. The model used Ammonia, Nitrate/Nitrite, Orthophosphorus, and Flowing as independent variables; the categorical *E. coli* variable as the output variable; and a max_depth hyperparameter of three.

We used a confusion matrix to further evaluate the model's performance (Figure O). The confusion matrix reveals the true positive and false positive metrics, which are used to calculate the model's overall precision and precision for each group.

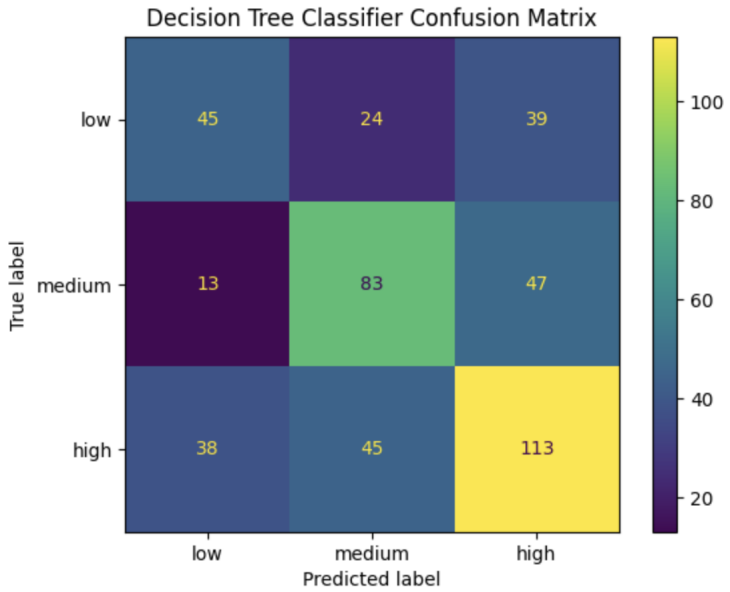

Figure O: A confusion matrix made using the decision tree classifier.

Table 7: True positive and false positive metrics

E. coli concentration classification:	Count of true positives:	Count of false positives:
Low	45	13 predicted to be medium & 38 predicted to be high (51 total)
Medium	83	24 predicted to be low & 45 predicted to be high (69 total)
High	113	39 predicted to be low & 47 predicted to be medium (86 total)
Total for all three classifications:	241	206

Precision was calculated using the equation $Precision = \frac{true\ positives}{true\ positives + false\ positives}$. The precision of each E. coli category is ~0.47 for low, ~0.55 for medium, and ~0.57 for high. The overall precision of the model's test classification is 0.539, very similar to the model's accuracy score.

The last way we evaluated the model was by determining the feature importance. Feature importance accounts for the total contribution of each independent variable across the decisions made throughout the tree. Feature importance scores were first determined for the decision tree classifier with max_depth set to three (Figure P). These feature importance scores mirror the variables used for splitting decisions at the nodes in the decision tree visualization (Figure N). Feature importance scores for the model were also determined with no specified max_depth (Figure Q). Interestingly, the flowing variable had the second highest feature importance score when max_depth was three but the lowest when max_depth was left unspecified. Additionally, the ammonia variable had a feature importance score of zero when max_depth was three but the second highest when max_depth was left unspecified. This inconsistency in feature importance scores for the flowing and ammonia variables reveals a limitation in the decision tree model. Restricting max_depth to three allowed the model to best speculate classification patterns without becoming overspecified to the training data. However, a limited max_depth may overgeneralize complicated patterns in the data.

The feature with the highest importance score in the decision tree classification with a max_depth set to three was Nitrate/Nitrite (Figure P). This is not surprising given that Nitrate/Nitrite appears as the discriminating feature for five out of the seven nodes in the decision tree visualization (Figure N). Nitrate/Nitrite was also the feature with the highest importance score when max_depth was not specified (Figure Q). A variable, or feature, having a high feature importance score indicates that it has a strong impact on the model's prediction. Therefore, the Nitrate/Nitrite feature's high importance score at both limited and unlimited depths shows that Nitrate/Nitrite is the most impactful on decision tree classification out of the variables included in the model.

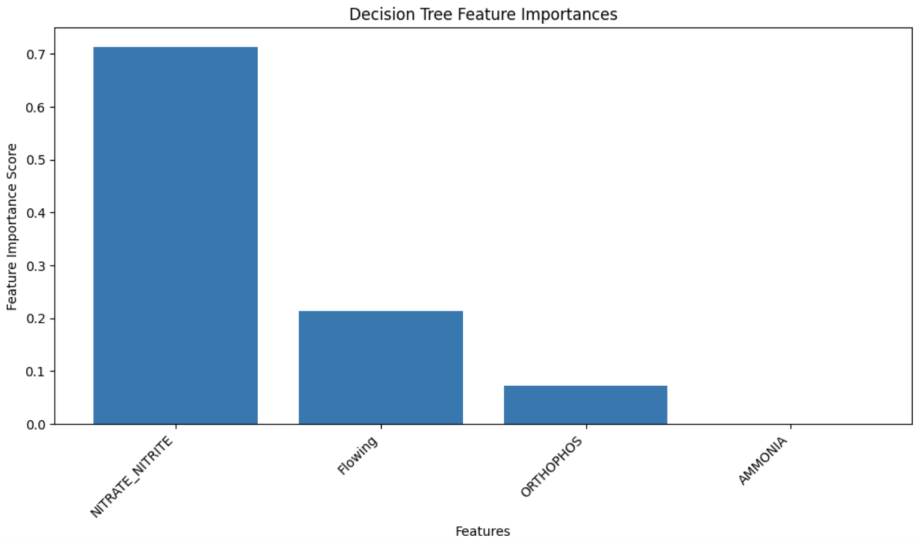

Figure P: A bar graph visualizing the feature importance scores for each variable used in the classification model max_depth was set to 3 for the decision tree classifier.

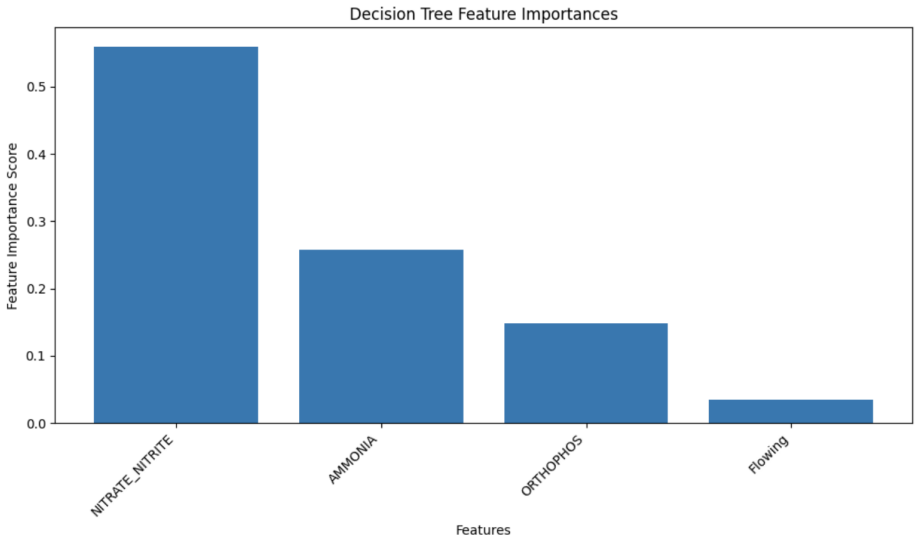

Figure Q: A bar graph visualizing the feature importance scores for each variable used in the classification model when no max_depth was set for the decision tree classifier.

Findings

1. The regression analysis on four different nutrients, Orthophosphorus, Phosphorous, Ammonia, and Nitrate/Nitrite, and the bacteria *E. coli* revealed that there was no significant relationship between the two regardless of watershed type and year. These results contrasted with our hypothesis as in some cases there were slightly negative correlations between the two parameter comparisons.
2. Feature importance scores from the decision tree model show that the Nitrate/Nitrite variable is very impactful on classification to predict *E. coli*. This indicates that the Nitrate/Nitrite concentration of a water sample is associated with *E. coli* concentration. The results of the classification model may not reflect the linear regression results from hypothesis 2 as the relationship between nutrients and *E. coli* may not be linear.
3. The E. coli location maps shown earlier visualized E. coli concentrations at the locations the sample was taken from. In that image, it appears that *E. coli* is very prevalent in population-dense areas. Using regression analysis, we were able to see the following:
    a. We were able to conclude that the water temperature levels do not have a significant effect on the concentration of *E. coli*.
    b. Additionally, we were able to conclude that pH levels and *E. coli* did not show a

significant relationship.

c. Both of these conclusions can be explained due to the properties observed by *E. coli*. This bacteria can withstand large changes in temperature and pH.

4. Based on our results analyzing hypothesis 3, we can conclude that the presence of flowing water was correlated with an increased frequency of *E. coli* presence in water bodies throughout Austin. It is difficult to conclude whether this disparity is due to the presence of flowing water, or if there is an unknown interaction with other variables. Feature importance scores from the decision tree model, with the optimized max_depth of three, show that flowing water does impact classification for the amount of E. coli (Figure P); however, this may be because of limitations to the classification model.

<u>Limitations</u>

1. Cross Contamination and Human/Animal Interaction

Future studies could explore if pesticide and fertilizer runoff in residential areas is associated with *E. coli* concentration. This could be done by incorporating data about commercial pesticide and fertilizer use in certain watersheds. Additionally, contaminated fecal matter (from both humans and animals) that is in our natural water sources would be the primary factor in the spread of *E. coli*. Thus, conducting a data analysis on fecal matter and its variables would also be beneficial in understanding the spread of *E. coli*. Another possible research expansion could include human recreational activity in lakes/streams and its possible correlation with a greater presence of *E. coli*.

2. Stagnant vs. Moving Water

One improvement could be the determination of whether each water sample was taken from stagnant or moving water to eliminate assumptions. If it cannot be determined for certain, setting a clearer definition of how flowing/still was determined would improve the study. Moving water is also a major factor in cross-contamination, so it's important to consider how environmental factors affect run-off, streams, and rivers.

3. Location and Time

Based on our results, we fail to conclude that phosphorus and nitrogen-containing nutrients are present at higher concentrations in Austin bodies of water that also contain high levels of *E. coli* bacteria. However, this correlation is limited due to the smaller sample size and because these parameter concentrations were not sampled at each watershed simultaneously. When looking at the correlation between the bacteria and several nutrients over the years we might find a slightly more accurate regression, however, the results still aren't statistically significant. Ultimately, we can't assume that phosphorus necessarily causes high levels of *E. coli*.

Since phosphorus is an essential nutrient for algae and other aquatic plants, there is a high concentration in bodies of water that have thriving ecosystems such as Lady Bird Lake and Barton Springs. Furthermore, the dataset covers a long period from 1986 to 2023 which can prove to be a limitation in our study. Technological innovations and advocating for public health regulations have drastically changed environmental policies in the past three decades, which could potentially affect the causes of *E. coli* presence over the years. Because *E. coli* thrives in hot temperatures, we also have to consider global climate change and its implications on the environment and ecosystems.

Additionally, the dataset is specific to Austin - a very densely populated city that is known for its tourist attractions and downtown area. In future research, we could potentially look into datasets that include other Texan cities or potentially other U.S States for a more accurate analysis of *E. coli* presence in America and its implications.

Ethics:

Using the AREA acronym, we've narrowed down four relevant principles including "People Affected", "Potential Conflicts", "Public Dialogue", and "Continuous Improvement". Since these data samples are from reservoirs and other freshwater bodies, the research would affect home buyers and water consumers. Austin's water supply is sourced from these watersheds, which could potentially contaminate our everyday water usage.

Additionally, tourists are also affected as Barton Springs and other natural freshwater pools are popular attractions. Regarding conflicts, our research may cause disagreement with those who have monetary interests concerning these watersheds. Additionally, individuals who contract E. coli may experience a variety of unwanted symptoms, including diarrhea, stomach

cramps, and potentially even fevers. Some long-term effects of *E. coli* include high blood pressure, heart failure, diabetes, seizures, comas, and in severe cases, brain damage. Thus, an increase in awareness of *E. coli*'s presence in everyday necessities and activities may cause public concern, which could potentially lead to boycotts and layoffs/unemployment of certain companies.

Through this research, we hope that public dialogue will increase awareness and coverage of *E. coli* presence for those who do not already have access to this information. Ultimately, continuous improvement from dedicated organizations and public petitions could be created in order to improve the supply of water. By identifying how *E. coli* reacts to certain variables, more city-funded projects could also be advocated for to avoid *E. coli* growth.

Conclusion:

During our research, we looked to find out how Austin's waterways and *E. Coli* concentrations are related to various factors. Taking into consideration the concentrations of *E. coli* we observed geospatial measures of this bacteria in Austin and theorized about the implications of population and the bacteria concentrations. For our first hypothesis, we found that water temperatures do not appear to have any correlation with the *E. coli* concentrations. Our second research question revealed that concentrations of different nutrients did not appear to correlate to that of *E. Coli* concentrations significantly in both cases when grouping by watershed or year. Additionally, the presence of flowing water also seems to have a significant impact on the presence of *E. Coli*, with a significantly increased frequency that is likely due to cross-contamination. Lastly, through our fourth hypothesis, we were able to observe that the pH of water did not have any significant effects on the amount of *E. coli*.

Acknowledgments:

A	B	C	D	E	F	G	H	I	J	K
Task	Data Preparation	Graphs	Scheduling/Mee	Statistical Analys	Presentation	Report Writing	Website		Total	Contribution
Task %	5	9	30	13	21	15	5		98	
Anthony	0	0	10	0	3	1	0		14	100
Hayden	0	0	9	2	3	0	0		14	100
Katie	0	3	3	2	3	3	0		14	100
Santiago	0	0	0	3	3	3	5		14	100
Timothy	2	3	0	3	3	3	0		14	100
Valeria	3	3	0	3	3	2	0		14	100
Vyvy	0	0	8	0	3	3	0		14	100
Total	5	9	30	13	21	15	5	Total	98	
								Max	14	

Citations:

Bartram, J., & Gordon, B. (2015). Water Microbiology. Bacterial Pathogens and Water. *International Journal of Environmental Research and Public Health*, 12(10), 1234-1243. https://doi.org/10.3390/ijerph121012345

Centers for Disease Control and Prevention. (2018). *Water-related Diseases and Contaminants in Public Water Systems*. https://www.cdc.gov/healthywater/drinking/public/water_diseases.html

The City of Austin. (2021). *Water Quality Sampling Data*. AustinTexas.gov Data Library. https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray

E. coli (escherichia coli) - U.S. Environmental Protection Agency. (n.d.). https://www.epa.gov/system/files/documents/2021-07/parameter-factsheet_e.-coli.pdf

Farnleitner, A. H., Ryzinska-Paier, G., Reischer, G. H., Burtscher, M. M., Knetsch, S., Kirschner, A. K. T., Dirnböck, T., Kuschnig, G., Mach, R. L., & Sommer, R. (2015). Contamination of water resources by pathogenic bacteria. *AMB Express*, 5(1), 57. https://doi.org/10.1186/s13568-015-0156-6

Kramer, M. R., Strahan, A. E., Preslar, J., Zaharatos, J., St. Pierre, A., Grant, J. E., Davis, N. L., & Goodman, D. A. (2019). Applying a Health Equity Framework to Maternal Mortality Reviews. *American Journal of Obstetrics and Gynecology*, 220(4), 378.e1-378.e9. https://doi.org/10.1016/j.ajog.2018.12.033

United Nations Environment Programme. (2015). *Nutrient Economics Report 2015: Managing the Nutrient Cycle to Sustainably Meet the Global Food Demand*. UNEP. https://www.unep.org/resources/report/nutrient-economics-report-2015

United States Census Bureau. (2019). *Geographical Areas Reference Manual*. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html