Mercury Levels in Freshwater Lakes in Maine: Risk Assessment and Environmental Factors Identification

The Government of Maine

Miles Xi

Executive Summary

Mercury contamination in fish has become a concern for public health. Fish, living in water, can accumulate chemicals in their bodies (USEPA, 2025). This study analyzes a sample of freshwater lakes in Maine, U.S. It addresses three questions: Do the mercury levels at lakes in Maine exceed the 0.43 ppm warning threshold? Are high mercury levels related to the presence of dams? Can the lake type explain the variation in mercury levels of lakes?

Using data collected from a sample of lakes, we found that 45% of lakes have mercury levels above 0.43 ppm. Statistical modeling showed no significant relationship between mercury levels and the presence of dams or lake type.

The study is useful for policy decision-making, environmental monitoring, and public health assessment.

Introduction

Waters where people fish usually contain chemical contamination such as mercury, plastics, and pesticides. These contaminants can cause serious harm to human health through the food chain (USEPA, 2025). As a result, states in the U.S. monitor mercury levels in local waters to issue fish consumption advisories. At the request of the Maine State government, this report analyzes freshwater lakes in Maine. It investigates three problems:

(a) Do the mercury levels at lakes in Maine exceed the 0.43 ppm warning threshold?
(b) Are high mercury levels related to the presence of dams?
(c) Can the lake type explain the variation in mercury levels of lakes?

Answering questions (b) and (c) requires significance tests using the sample data, a statistical approach to determine if the hypothesized relationships are statistically significant or happen by chance.

Data and Methods

The data consists of 120 observations on 18 variables. Table 1 shows the variable names and the explanations.

Table 1. Variables names and explanations. The original variable names were changed for readability.

Variable Name	Explanation
LakeName	Lake or pond name
MercuryLevel	Mercury level, response variable
NumFishSampled	Number of fish in the composite (i.e., sampled)
Elevation	Elevation (feet)
SurfaceArea	Surface area (acres)
MaxDepth	Maximum depth (feet)
LakeType	Lake type: 1 = oligotrophic, 2 = mesotrophic, 3 = eutrophic
TempStratification	Temperature stratification: 1 = stratified, 0 = non-stratified
DrainageArea	Drainage area (square miles)
RunoffFactor	Runoff Factor
FlushingRate	Flushing rate
DamPresence	Dam presence: 0 = no functional dam, 1 = some man-made flowage
LatitudeDegrees	Latitude degrees
LatitudeMinutes	Latitude minutes
LatitudeSeconds	Latitude seconds
LongitudeDegrees	Longitude degrees
LongitudeMinutes	Longitude minutes
LongitudeSeconds	Longitude seconds

During data pre-processing, we first identified five duplicate values for the *LakeName* and appended suffixes to differentiate them. Ten of 120 observations contained missing values, which were not discarded until the statistical modeling step. For categorical variables, i.e., *LakeType*, *TempStratification*, and *DamPresence*, missing values were labeled as 'Unknown.' Finally, we set *LakeName* as the index; the dataset effectively contained 17 variables.

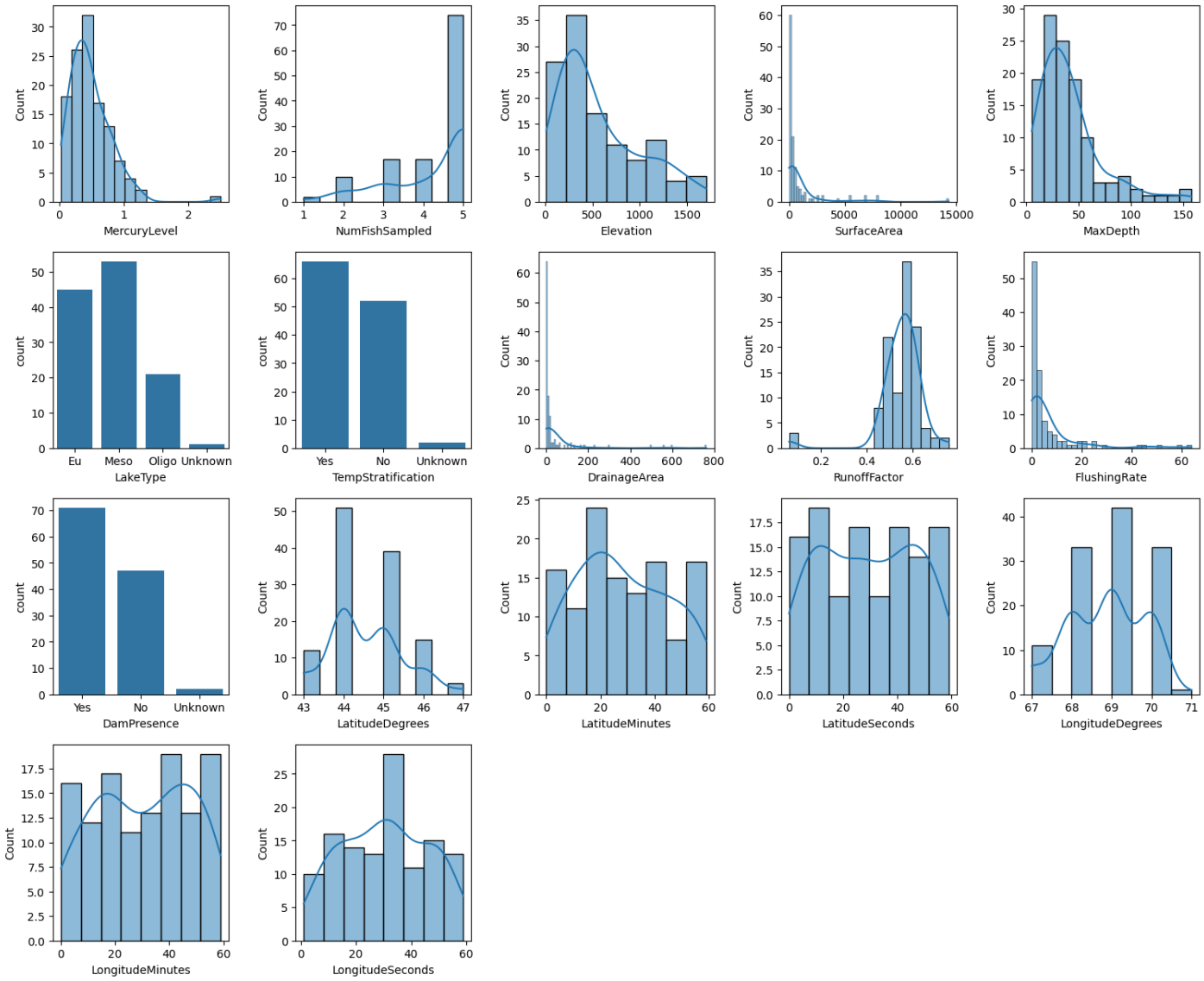

Figure 1. Histograms and bar charts for numerical variables and categorical variables. Kernel density estimation (blue curves) was used to generate a smooth distribution for numerical variables.

To answer the questions, we first explored the data through data visualization, to get an intuitive understanding of the data and preliminary answers. Then, we used multiple linear regression to formally test the hypotheses (b) and (c).

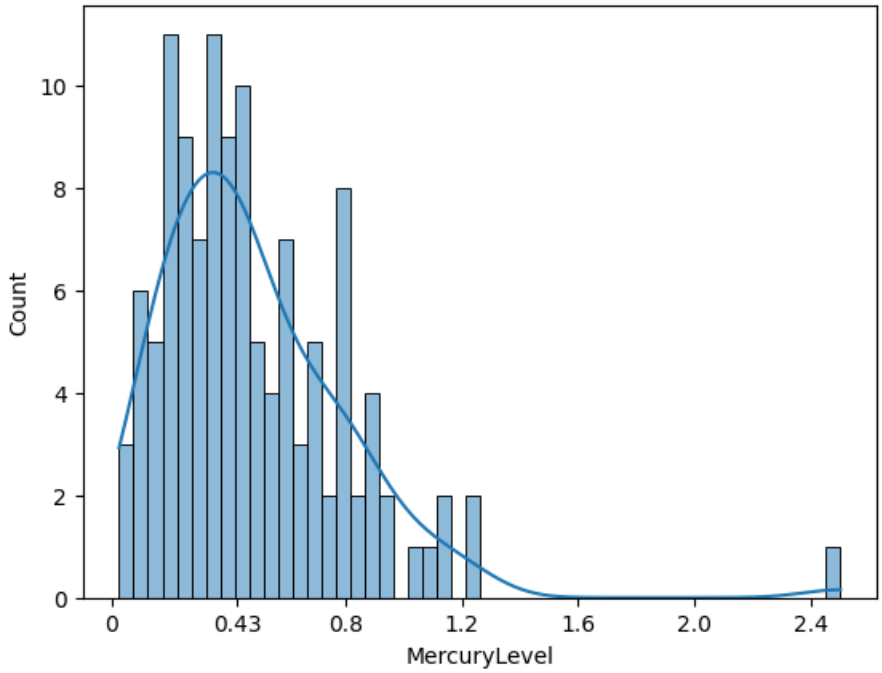

Figure 2. A detailed distribution of *MercuryLevel* of the lakes in Maine. Forty-five percent of them (54 out of 120) exceed 0.43 ppm warning level.

Exploratory Data Analysis

To explore the data, we created histograms and bar charts for continuous variables and categorical variables to visualize their distributions, as shown in Figure 1. Figure 2 is a more detailed histogram of *MercuryLevel*.

From Figure 2, we can see that the distribution of *MercuryLevel* is right-skewed. Forty-five percent (45%) of the sampled lakes and ponds exceed the 0.43 ppm threshold, which is considered unsafe in Maine. This finding indicates a concern.

Figure 3 gives an intuitive examination of questions (b) and (c). Boxplots display the distribution of data given the level of a variable. Non-overlapping boxes across different levels indicate that the levels influence the data, because the data is distributed differently under different conditions. From Figure 3, neither *DamPresence* nor *LakeType* appears to be related to the mercury level, as the boxes overlap substantially.

Statistical Modeling and Results

Linear regression is a statistical method to test the linear relationship between a response variable and a set of predictors. In this case, we are interested in the effect of *DamPresence* and *LakeType* on *MercuryLevel*, controlling for other predictors or influences. Therefore the multiple linear regression model is specified as follows:

$$MercuryLevel = \beta_0 + \beta_1 DamPresence + \beta_2 LakeType + \sum \beta_i Control_i + \varepsilon$$

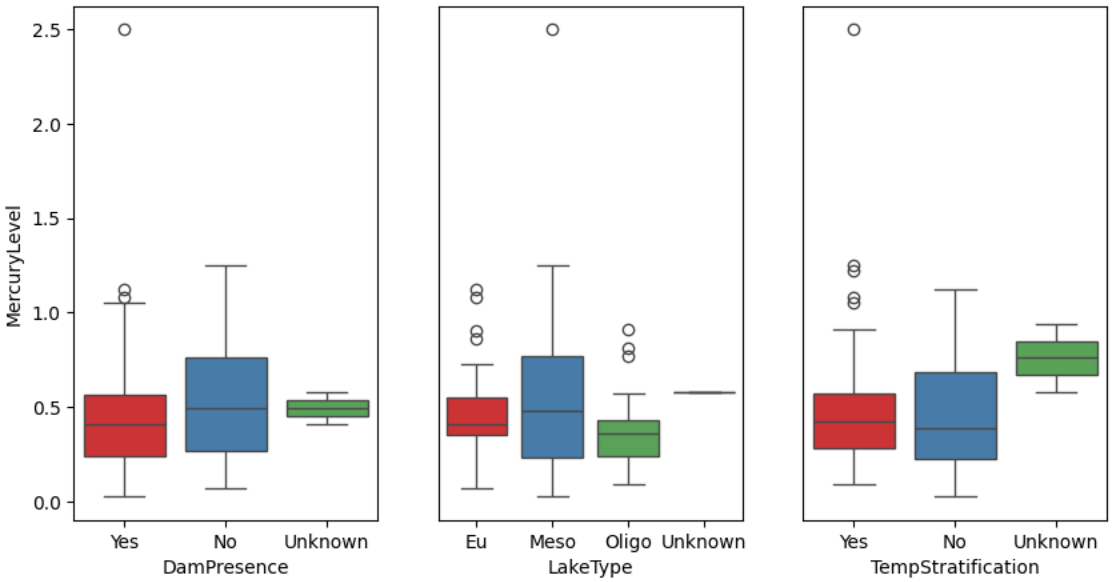

Figure 3. Box plots of *DamPresence*, *LakeType*, and *TempStratification* against *MercuryLevel*. Significant overlaps indicate no relationship.

where β's are coefficients to be estimated. They measure the effect of a predictor on the response variable, controlling for other predictors. To answer the questions, we (1) tested the model's significance, and (2) tested the significance of the coefficients of *DamPresence* and *LakeType* (β_1 and β_2).

We standardized each of the continuous predictors, created dummy variables for categorical variables, and removed the observations with missing values. We then estimated the coefficients in the model using a statistical package in Python. Results are presented in the Table 2.

The proposed linear relationship did not hold for either model, as F test statistics were not significant at 5% level. In addition, individual coefficients β_1 and β_2 were not significant. This means, based on the current sample, we do not have enough evidence to prove that there is a linear relationship between *MercuryLevel* and *DamPresence* or *LakeType*. Therefore the answers to the question (b) and (c) are negative.

Discussion and Conclusion

Using statistical analysis, this report answers three questions. The mercury levels in lakes and ponds in Maine are concerning, with 45% of the sampled waters exceeding the 0.43 ppm threshold. However, there is no evidence to support the claims that mercury levels are associated with the lake type or the presence of dams.

The limitations in data and the methodology may risk the robustness of the conclusion. Future research could benefit from a larger sample size and better handling of missing values.

Table 2. Multiple linear regression results for a model and a smaller model excluding Latitude and Longitude variables. Estimated coefficients (β's) are shown with standard errors in parentheses. Significance levels are indicated using ***p-value < 0.01, ** p-value < 0.05, * p-value < 0.10.

Variable	Model 1	Model 2
const	0.4432***	0.4080***
	(0.130)	(0.123)
LakeType_2.0	0.0983	0.1102
	(0.133)	(0.123)
LakeType_3.0	-0.0497	-0.0140
	(0.127)	(0.118)
TempStratification_1.0	0.0783	0.0926
	(0.088)	(0.083)
DamPresence_1.0	-0.0452	-0.0314
	(0.077)	(0.071)
NumFishSampled	0.0169	0.0004
	(0.034)	(0.033)
Elevation	-0.0178	-0.1118**
	(0.063)	(0.035)
SurfaceArea	-0.0700	-0.0499
	(0.052)	(0.047)
MaxDepth	-0.0068	-0.0269
	(0.056)	(0.055)
DrainageArea	0.0382	0.0389
	(0.049)	(0.044)
RunoffFactor	-0.0284	-0.0009
	(0.038)	(0.033)
FlushingRate	-0.0178	-0.0285
	(0.036)	(0.035)
LatitudeDegrees	-0.1129	
	(0.074)	
LatitudeMinutes	-0.0537	
	(0.042)	
LatitudeSeconds	-0.0220	
	(0.035)	
LongitudeDegrees	-0.1043	
	(0.063)	
LongitudeMinutes	-0.0694*	
	(0.036)	
LongitudeSeconds	-0.0293	
	(0.035)	
F statistic	1.501	1.792*
Adjusted R^2	0.073	0.074

References

United States Environmental Protection Agency. (2025, January 31). How Do I Know if a Fish I Caught is Contaminated? https://www.epa.gov/choose-fish-and-shellfish-wisely/how-do-i-know-if-fish-i-caught-contaminated

Appendix

For code and data, see Github repository: https://github.com/miles-xi/Data-analytics-consulting-case-study