Flint Water Crisis Statistical Analysis

Zhishan Li, Alison Cronander, Sam Isken, Xinyi Wang
STT301 Final Project

Agenda

1. Introduction & background
2. Visual Data Analysis
3. Further Investigation of the Data
    a. Non-linear Least Squares Method
    b. Linear Model Method
    c. Statistical Results
    d. EPA guidelines (immediate actions need?)
4. Conclusion & Discussion

Introduction & Background

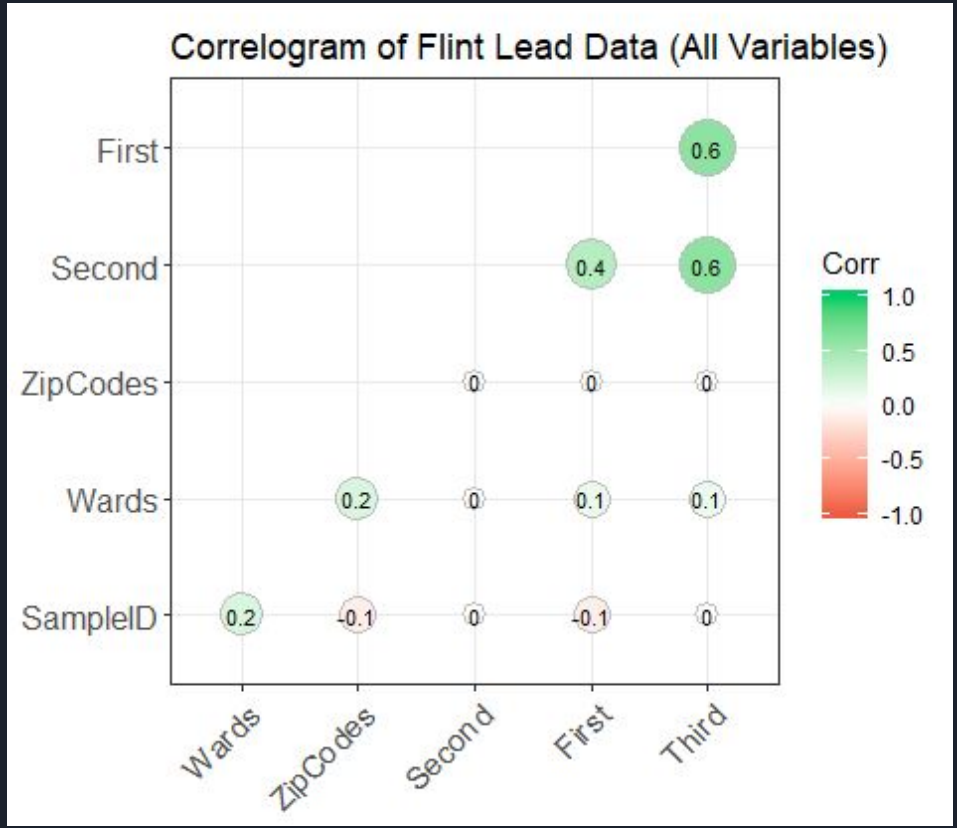
- Flint Water Crisis began in 2014
- Crisis is a result from many issues
  - Legislative issues
  - Inadequate testing
  - Misleading data
  - Aging pipes
- Learning problems and slowed growth in children
- Crisis is ongoing
- Data we were given

Goals of the Research

- Any correlations between lead levels and let water run?
- Strategy (running water) effective from a public health perspective?
- Any relationships between zip codes and the lead level?
- How does Flint's Water Crisis compare to policy standards?

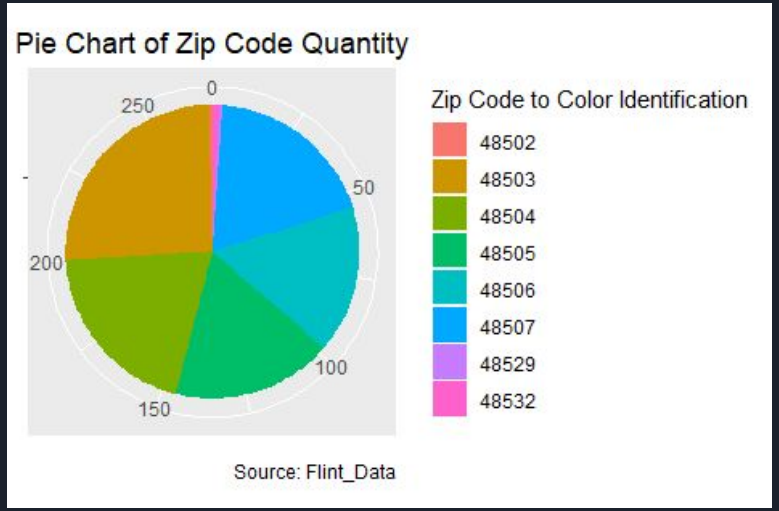
Visual Data Analysis

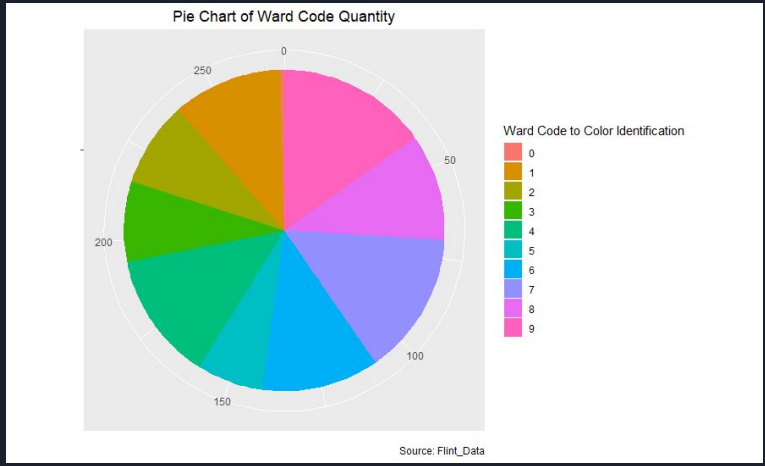
- Correlation Matrix Graph
  - Relatively high (0.6) between “First & Third” and “Second & Third”
    - Assumption: not be equivalent if time reduces the lead content
- Government supposes negative correlation



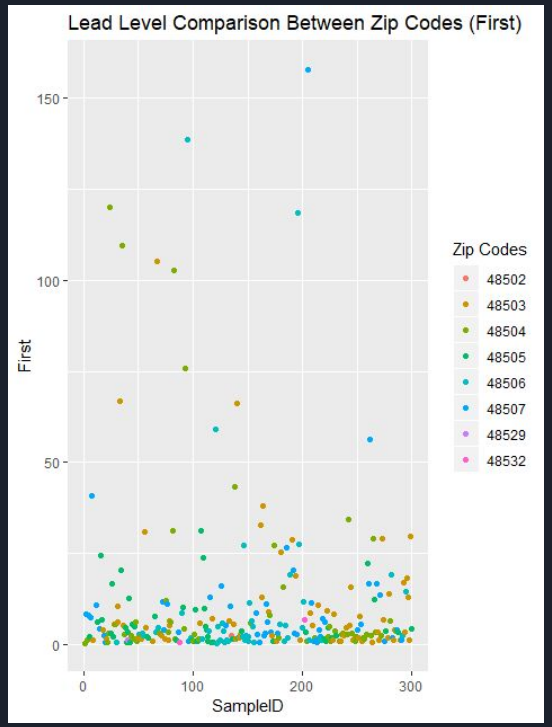
Visual Data Analysis

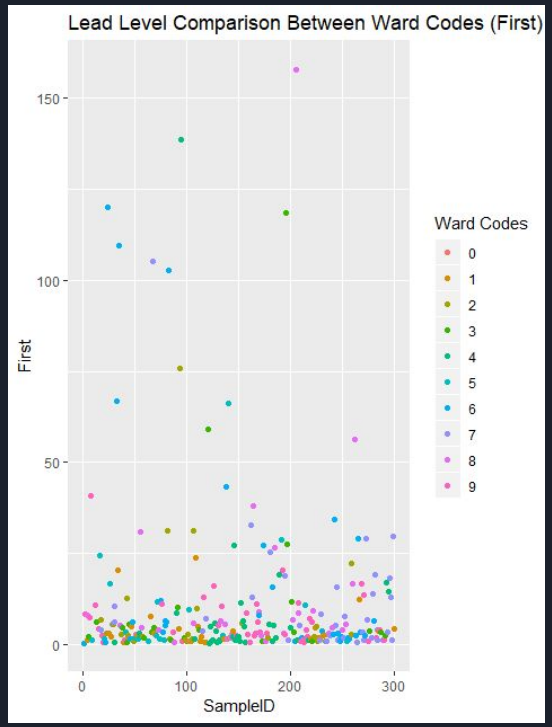
- Zip Code and Lead Content
  - Some zip codes contain 1-2 samples
  - The rest contain about the same amount of samples
- Ward Code (Neighbor) and Lead Content
  - Similar number of samples throughout wards
- Even distribution of each locations





Visual Data Analysis

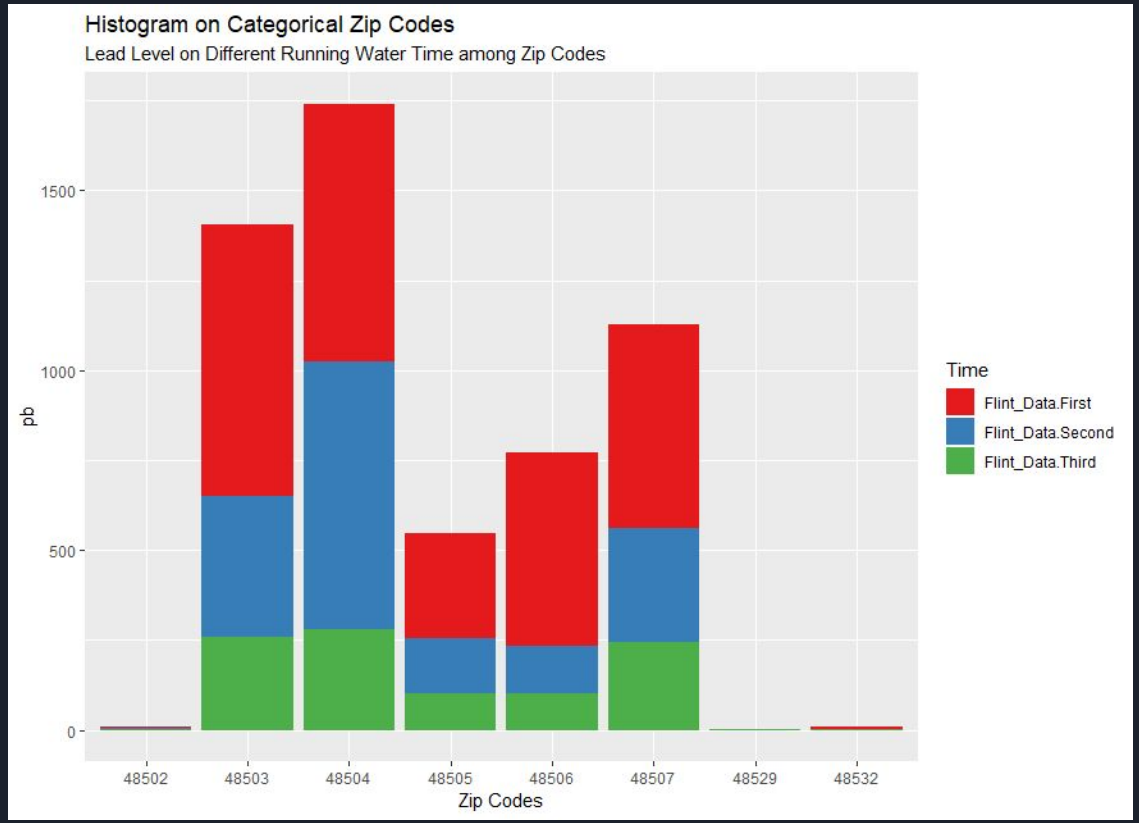




- Area and “First” Lead Content
  - No clear relationship
  - Issue does not discriminate by area

Visual Data Analysis

- Shows total lead in each zip code
- Some zip codes only have 0-5 samples, leading to misleading visualization



Further Investigation of the Data

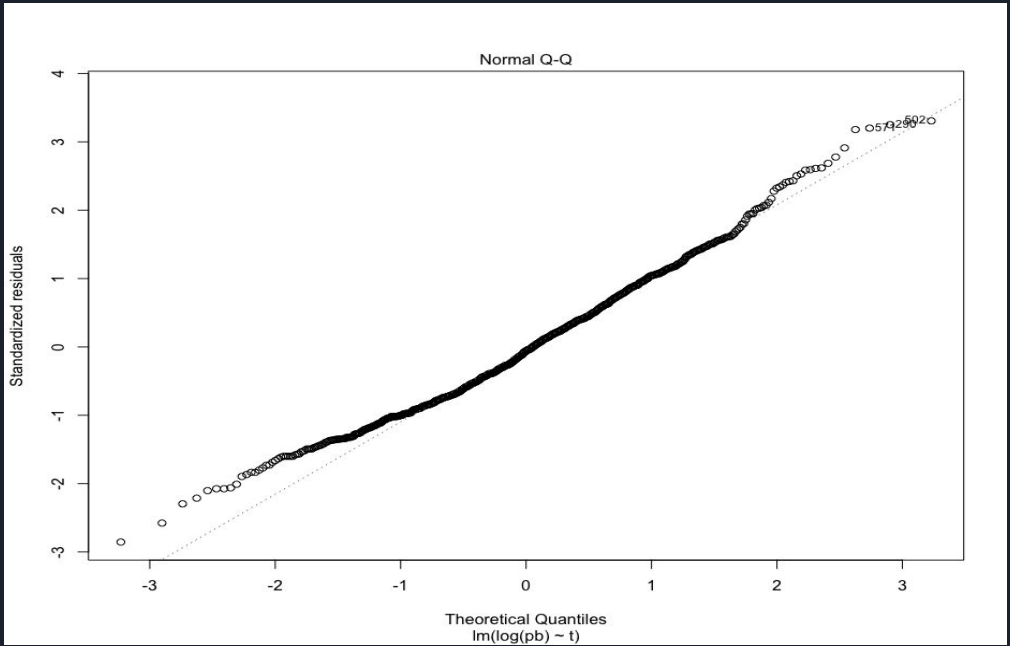
- Non-linear Least Squares Method
- Linear Model (After taking natural log)
- Exponential decay function with two parameters
- Determine the value of the parameters:
  - Running a regression on the natural log of the data

$$f(x, \theta_1, \theta_2) = \theta_1 e^{-\theta_2 x} \qquad (\theta_1 = 2,\ \theta_2 = -0.01)$$

Plots of Linear Model

Normal Q-Q:

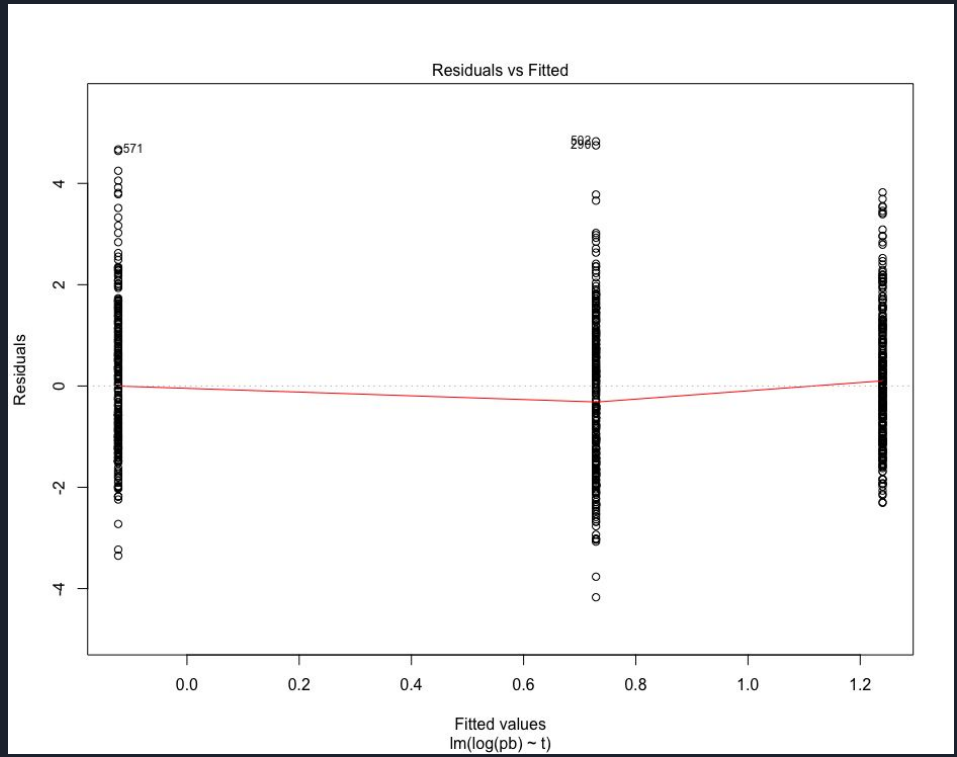
Results of our residuals are considered normal.



Plots of Linear Model

Residual vs. Fitted:

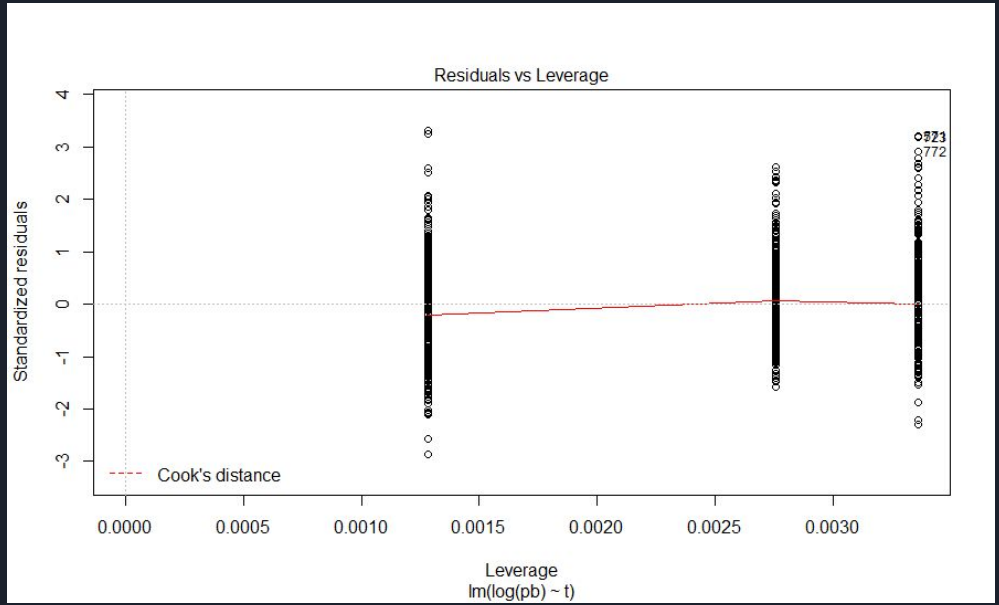
- Residuals do not possess heteroscedasticity
- Variances remain equal throughout our range of values of variable that predicts it.



Plots of Linear Model

Residuals vs. Leverage:

- Determine the X variables do not have undue influence on the model's result



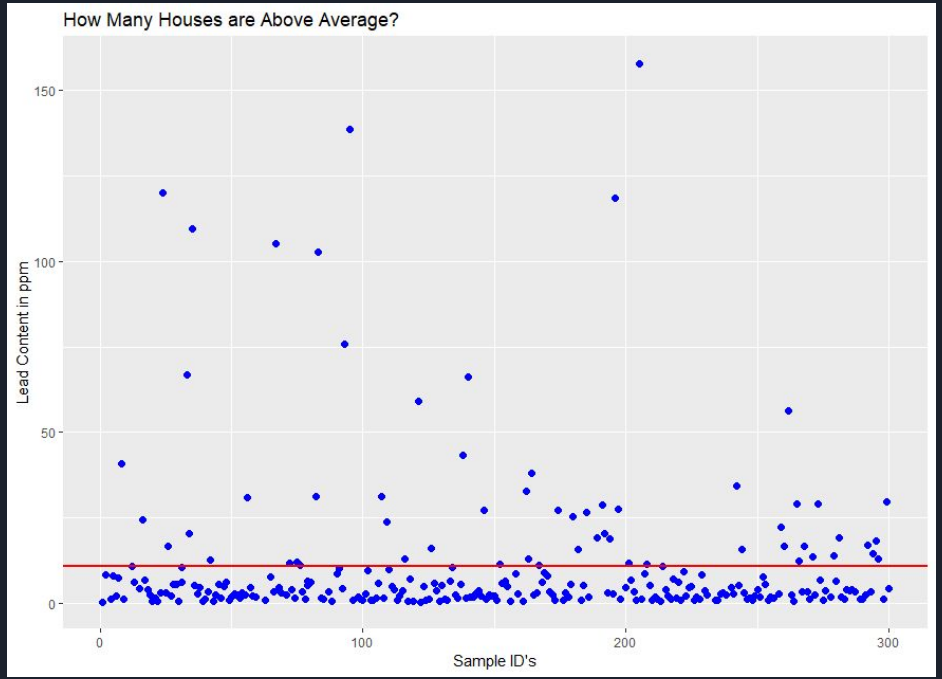
Results

- Negative relationship between time and lead content
  - Exponential decay function:
    - Negative slope = -0.01
    - Account only 12.8% of variability
      - R-Squared = .128

Environmental Protection Agency (EPA) Guidelines

- EPA guidelines state when action needs to be taken
- Percent of homes in Flint that exceed EPA standards
  - First draw | 16.67%
  - Second draw | 6.30%
  - Third draw | 4.44%

*Sources see Reference



Conclusion & Discussion

- ➔ Lead level barely decreases when letting tap run
- ➔ No relationship between zip code or ward and lead level
- ➔ Immediate actions are needed based on EPA guidelines

Reference

- EPA Guidelines: http://www.epa.gov/your-drinking-water/table-regulated-drinking-water-contaminants#seven
- Flint Water Crisis Handbook: April 2017 significancemagazine.com
- Flint Water Lead Content Data: From STT301 D2L
- Graph credits: Zhishan Li, Alison Cronander, Sam Isken, Xinyi Wang

Thank You!

Questions?