$Hack_2O$ | Analysis of the Flint datasets

Team Number Ninjas
Angie Ngoc Dinh & Van Nguyen
February 26, 2017

Project Questions

- *How this analysis benefits both the household residents and the researchers?*
- For household residents:
  - Which households are most at risk?
  - When is the time with highest lead levels?
- For researchers:
  - Is there a significant relationship between copper/iron and lead?
  - Is there any correlation between pH and Chlorine?
  - Is there any correlation between lead level and pH or Chlorine?

Data Sources

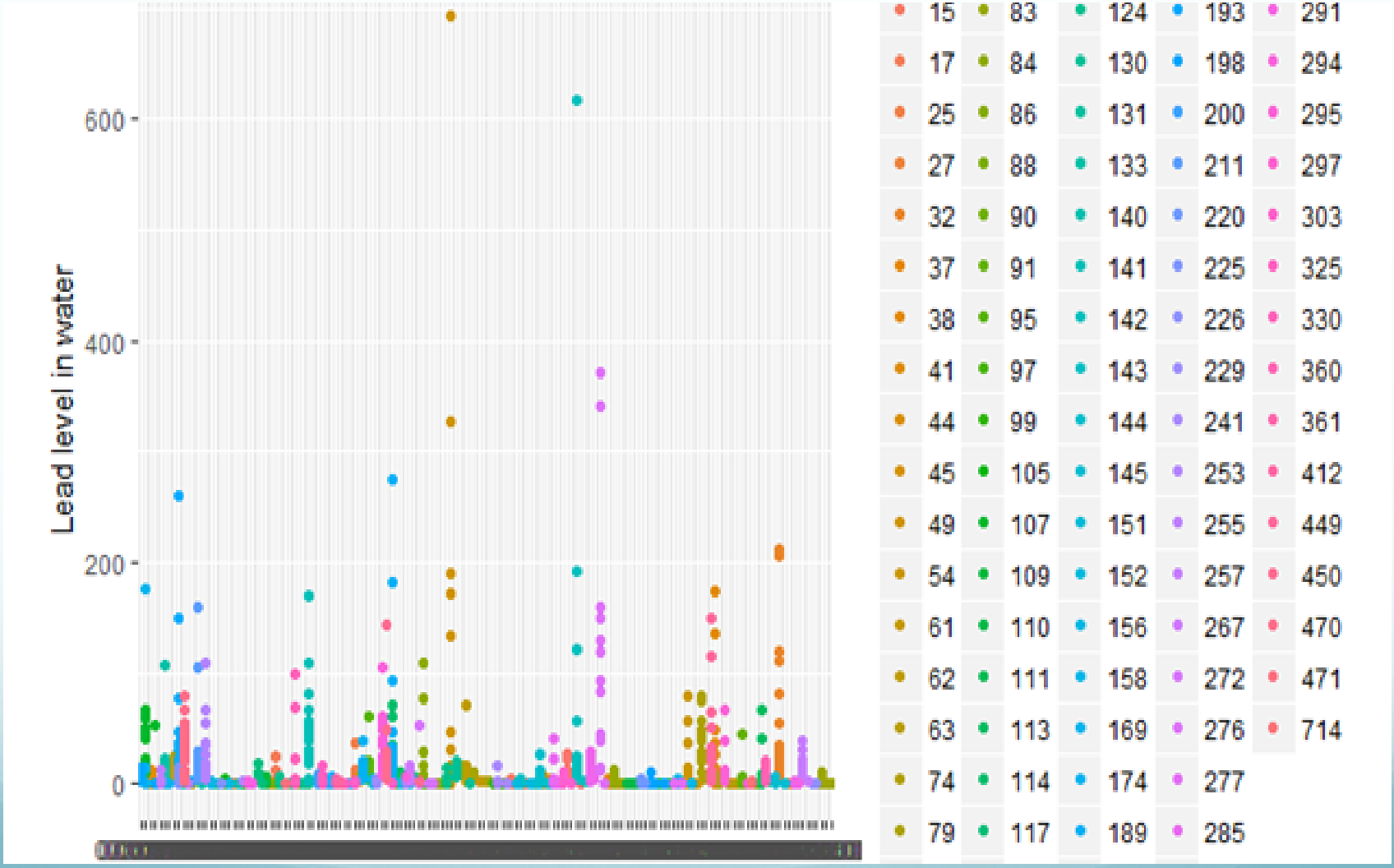
- Figure and data are based on two main datasets introduced by Joe Goodwill
  - Flint: Copper, Iron and Lead levels at different times of the day
  - Flint: Chlorine and pH levels

Our visualizations

- Which households are most at risk?
- When is the time with highest lead levels?

Lead Levels Among Households

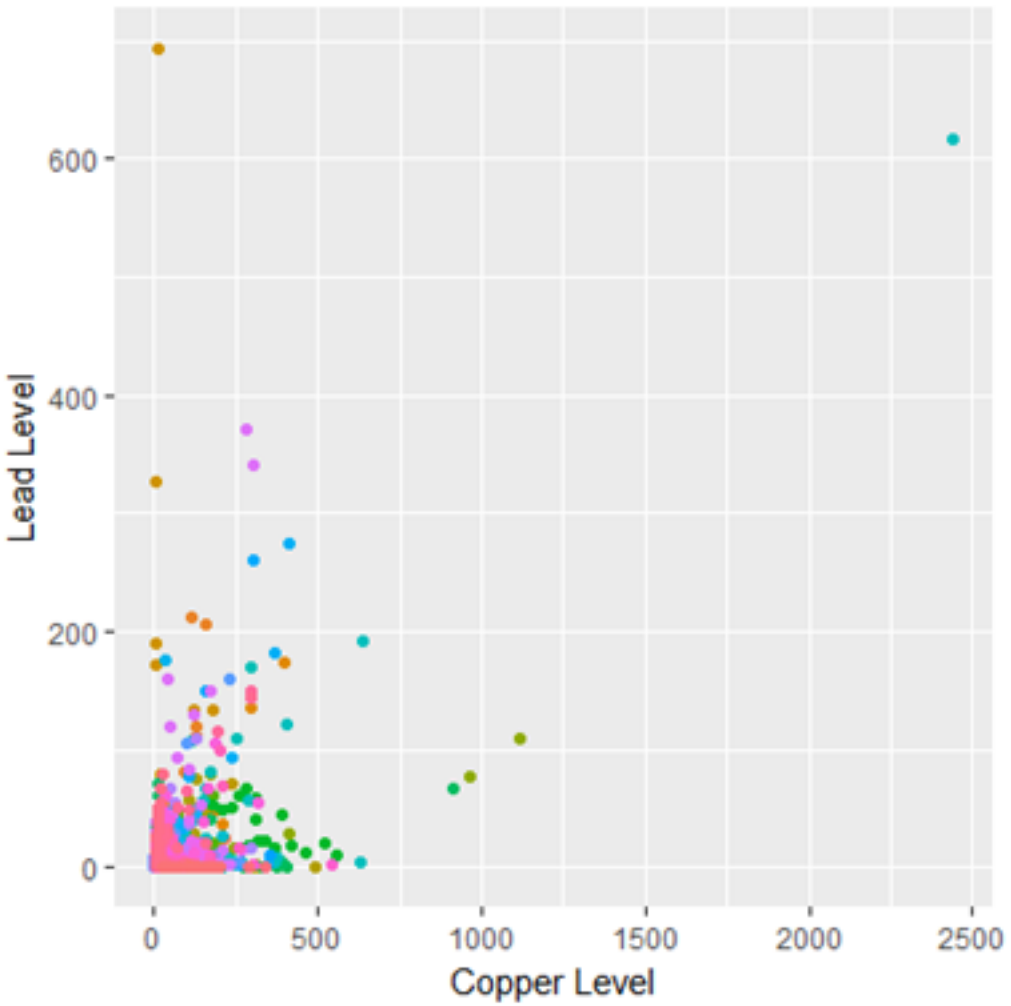
Lead level in water at different times of the day



Next questions

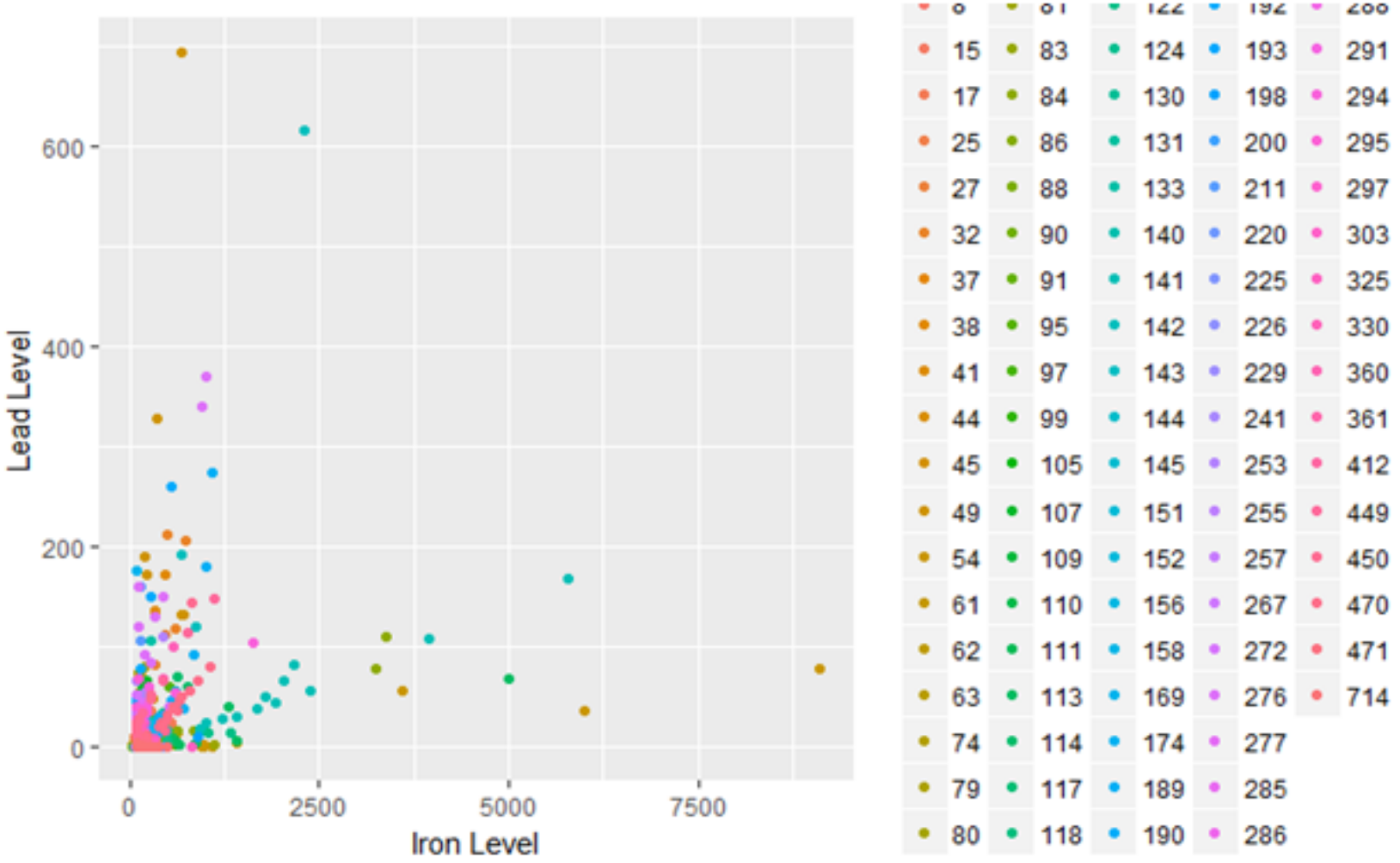
- Is there a significant relationship between copper level and lead level?
- Same question with iron level and lead level

Initial visualizations



15	83	124	193	291
17	84	130	198	294
25	86	131	200	295
27	88	133	211	297
32	90	140	220	303
37	91	141	225	325
38	95	142	226	330
41	97	143	229	360
44	99	144	241	361
45	105	145	253	412
49	107	151	255	449
54	109	152	257	450
61	110	156	267	470
62	111	158	272	471
63	113	169	276	714
74	114	174	277	
79	117	189	285	
80	118	190	286	

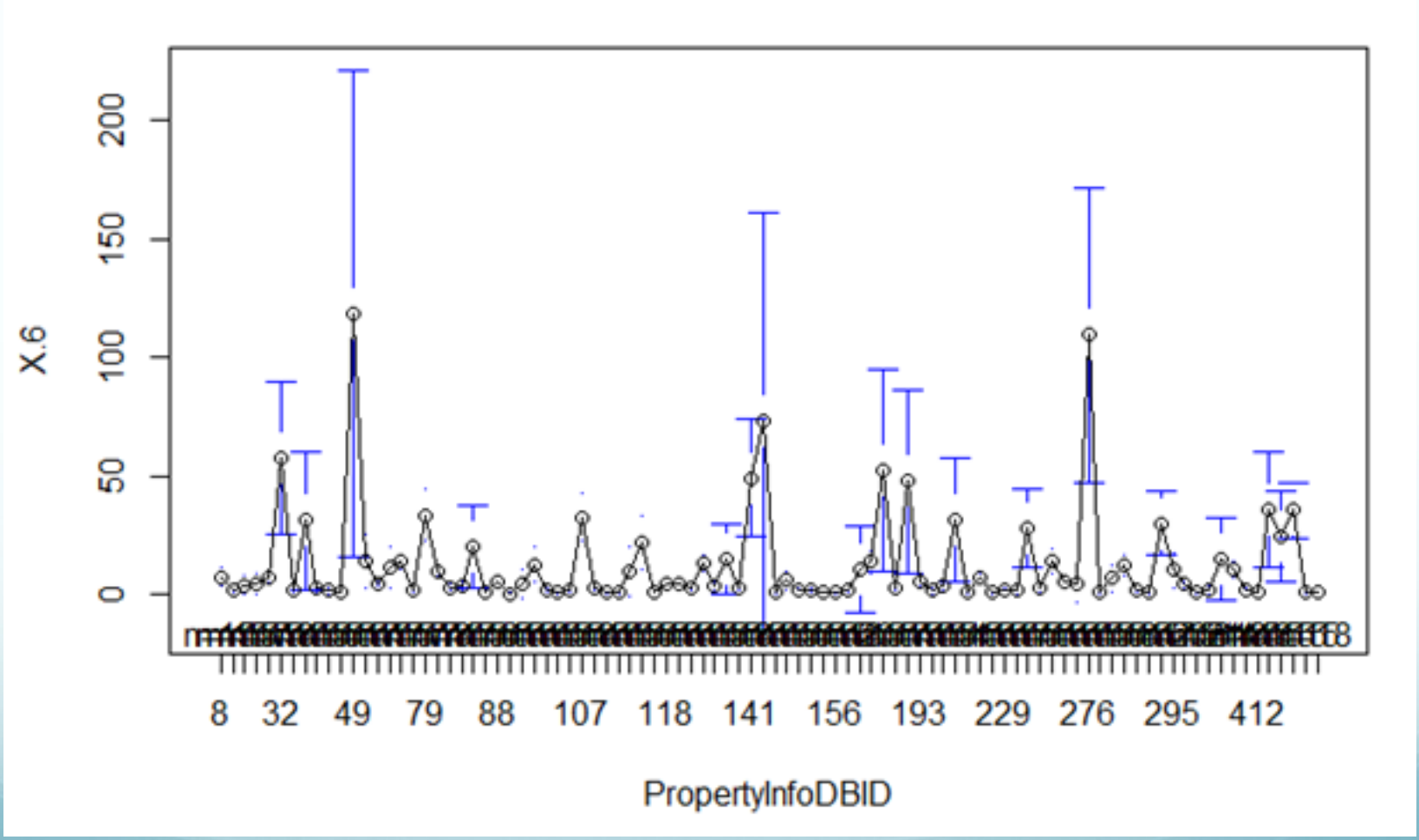
Initial visualizations (cont)

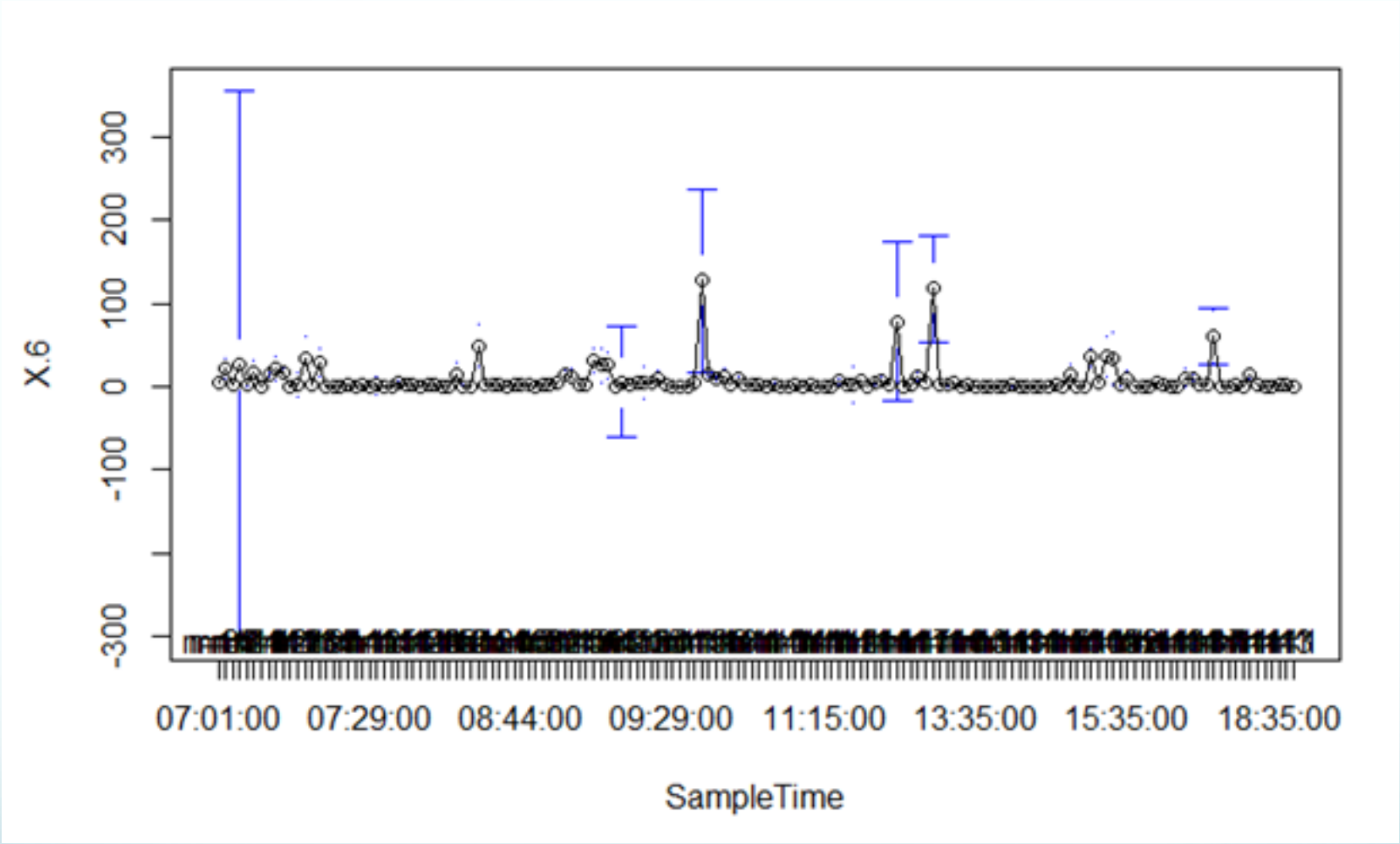


Model Selection

- Panel Data – Cross-sectional time series data
- Choices among fixed effects, random effects and regular ordinary least squares (OLS) regressions
- We used the Lagrange Multiplier Test – (Breusch-Pagan) to test for panel effects (unbalanced panels)
- We used the Hausman Test to test for fixed effects

Mean Plots


X.6
0
50
100
150
200
8
32
49
79
88
107
118
141
156
193
229
276
295
412
PropertyInfoDBID



The two mean plots show that the effects are quite random

Models & Testing

- We used the Lagrange Multiplier Test – Breusch-Pagan to test for panel effects (Lagrange Multiplier Test – Breusch-Pagan for unbalanced panels):

  data: X.6 ~ Result + Iron_Result_Ugl

  chisq = 717.17, df = 1, **p-value < 2.2e-16**

  alternative hypothesis: significant effects

- *This test shows that there is a panel effect (time effect) among the observations. Therefore, we have to use either a fixed effect or a random effect model*

Models & Testing (cont)

- We used the Hausman Test to test for fixed effects:

  data: X.6 ~ Result + Iron_Result_Ugl

  chisq = 2.9872, df = 2, **p-value = 0.2246**

  alternative hypothesis: one model is inconsistent

- This test show that we do not need to use a fixed effect model in this case.

Models & Testing (cont)

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	0.57	2.17	0.27	0.79
Copper level	0.15	0.009	17.19	**< 2.2e-16 *** **
Iron level	0.02	0.002	8.07	**1.531e-15 *** **

The model shows that there is a significant relationship between copper and lead level, as well as a significant relationship between iron level and lead level

Model Evaluation

a) Validity:

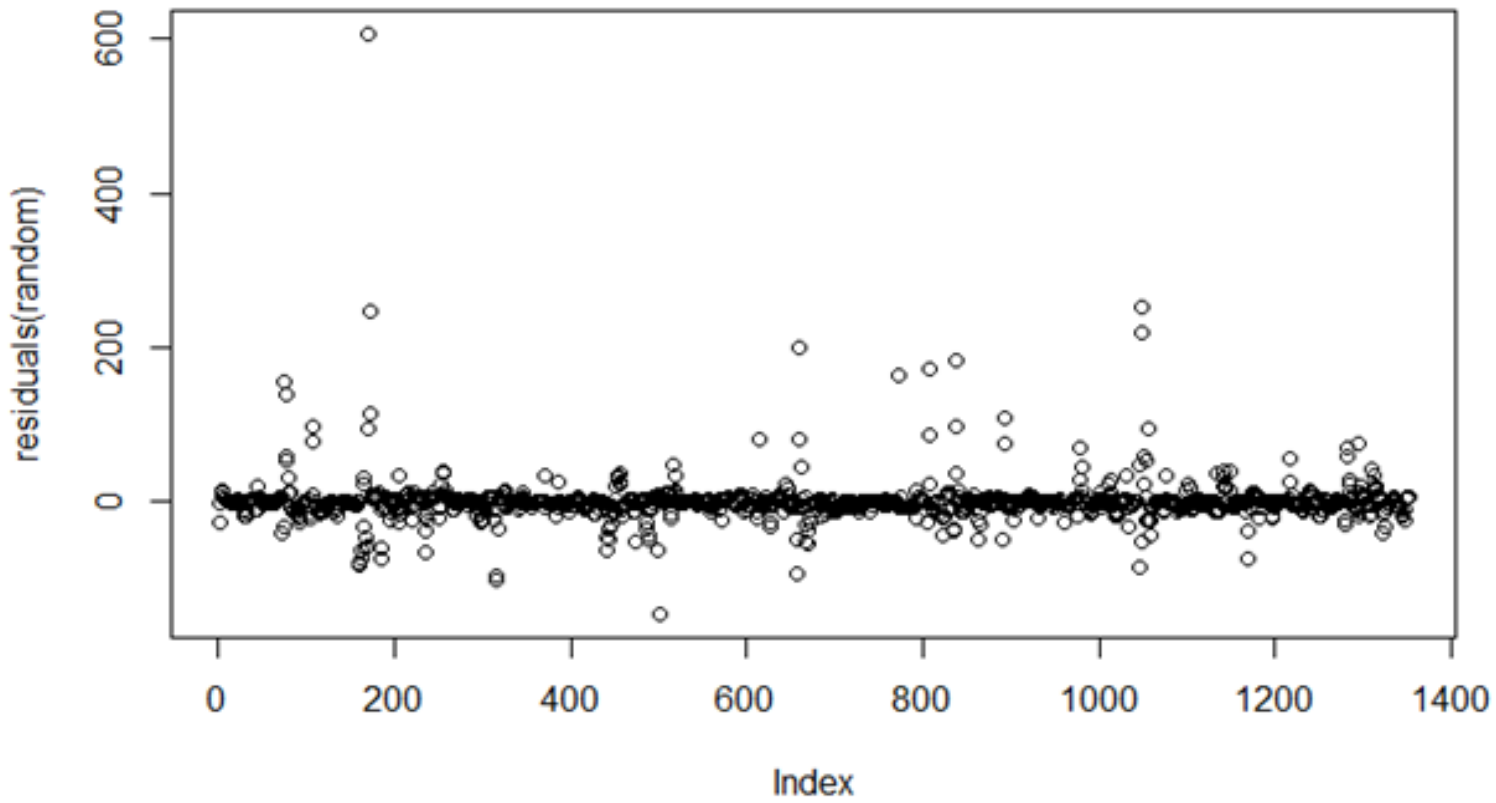
- Robust standard error:

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	0.57	2.80	0.20	0.79
Copper level	0.15	0.004	3.55	**0.004*** **
Iron level	0.02	0.007	2.40	**0.016 *** **

Controlling for unequal variance and autocorrelation, the relationship is still significant

Model Evaluation (cont)

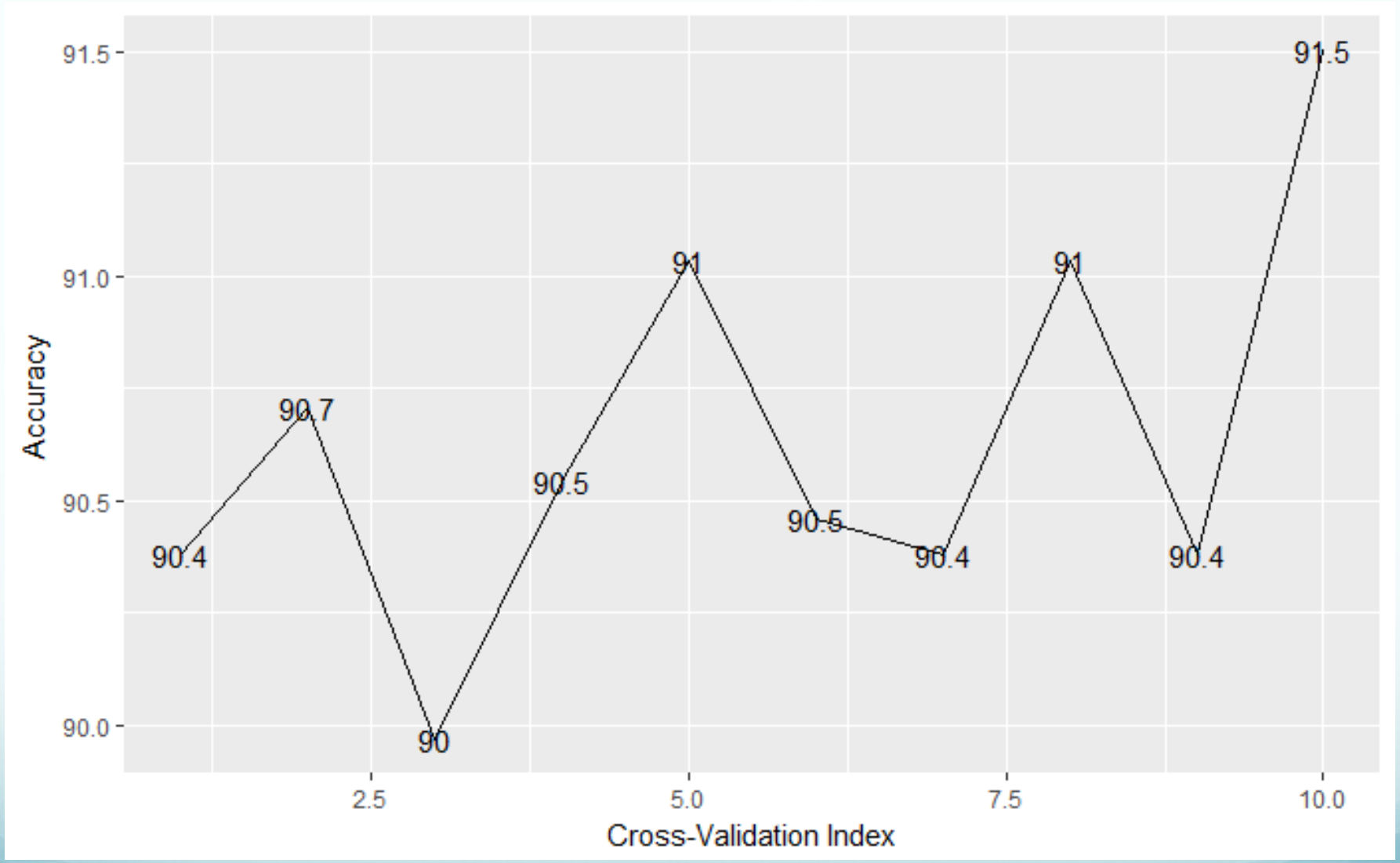
It seems that the variances are equal



Model Evaluation (cont)

b) Prediction

- Less than 15 Ug/L is safe
- More than 15 Ug/L is unsafe
- Using cross-validation, we have the Accuracy plot that shows round 90-91.5% accuracy

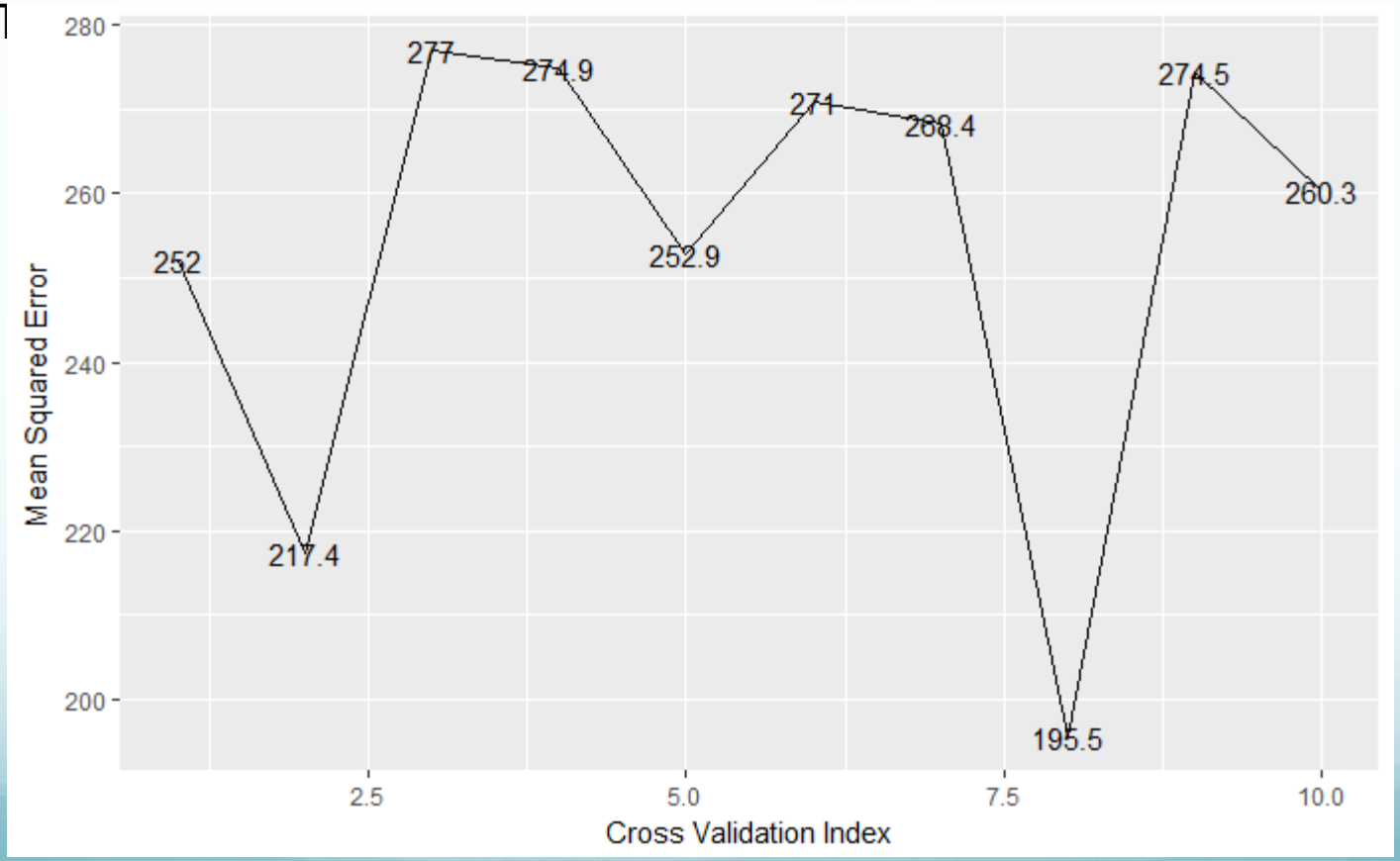

91.5
91.0
90.5
90.0
Accuracy
90.4
90.7
90
90.5
91
90.5
90.4
91
90.4
91.5
2.5
5.0
7.5
10.0
Cross-Validation Index

Model Valuation (cont)

- Base model: predicting that all water is safe, and has a 82.17 accuracy
- Cross-validation MSE – the random effect model

Mean Squared Error Plot

The MSE for the based model (the mean) is on average **1542.119**



Last Question & Answer

- Q: Is there any correlation between lead level and pH or Chlorine?
- A: We did not find any correlation between chlorine and pH, and between either of them and lead level

Questions?

Thank you for listening!

-- Number Ninjas