W25 CASE STUDY 01
DUE DATES:
Lectures 100/200: Saturday, February 15, at 8:00PM Ann Arbor Time via Gradescope
Lectures 300/400: Sunday, February 16, at 8:00PM Ann Arbor Time via Gradescope

Late submission policy: We offer a 1-hour grace period without any penalty. Submissions uploaded after the 1-hour grace period but within 24 hours of the deadlines listed above will be accepted but are subject to a 10 percent late penalty. No submissions will be accepted thereafter.

Purpose:
The overall purpose of this case study is to build your skills in conducting an applied statistical analysis in a real-world context. Specifically, this assignment is designed to (1) improve your statistical writing skills; (2) assess your proficiency with conducting and interpreting one-sample z-tests and t-tests; and (3) challenge you to make data-based arguments in a setting where decisions based on statistical analyses have substantial real-world implications.

The context - Revisiting the Flint Water Crisis:
Lead poisoning can be a serious problem associated with drinking tap water. Many older water pipes are made of lead. Over time, the pipes corrode, releasing lead into the drinking water. This can be most harmful to children, resulting in behavior and learning problems, slowed growth, iron deficiency, hair loss, and rashes, among other consequences. These effects can become irreversible. The Environmental Protection Agency (EPA)'s Lead and Copper Rule (LCR) of 1991 indicates that lead levels exceeding 5 parts per billion (ppb) can cause health problems and requires that if **more than 10 percent** of tested homes in a jurisdiction have lead levels greater than 15 parts per billion (ppb), action must be taken to mitigate the risks of lead poisoning.

In April 2014, the city of Flint, Michigan changed their water supply from Lake Huron to Flint River in order to save money. The Michigan Department of Environmental Quality (MDEQ) collected water samples from a random sample of households across the city. They measured lead levels (in ppb) for each sample and then created an indicator of whether the lead level was unsafe. After analyzing the sample of indicators, they declared Flint's water safe to drink.

However, within only a few weeks, citizens were concerned as they noticed early warning signs of lead poisoning. Upon retesting by a third party at the University of Virginia Tech, Flint's water was declared unsafe for consumption because of high-lead levels. Water from the Flint River was much more corrosive than Lake Huron water and it was apparent that the new chemical composition of the water was allowing lead to leach into the city water supply. In October of 2015, the city's water supply was switched back to Lake Huron. A state of emergency was declared, and state and federal funds were used to improve the water quality. By late 2016, the water was safe to drink again without a filter.

After the high-lead levels were discovered, the state of Michigan brought criminal charges against 13 states and local officials, including various accounts of misconduct, tampering with evidence, and conspiracy. Residents themselves issued class action lawsuits against 14 people, including then-Governor Rick Snyder.

Your task:
Your professors are particularly interested in the data collected as part of the water tests that led the MDEQ to incorrectly believe the Flint River water was safe to drink. Your task is to use the available data to first replicate the analysis MDEQ conducted when declaring the water safe to drink. Specifically, answer the research question:

> *Research Question A: Did more than **10 percent** of homes in Flint exceed the EPA action level of **15 ppb**?*

Then, conduct a separate analysis that answers a similar research question MDEQ did not answer:

> *Research Question B: Was the mean lead level in Flint water (in ppb) above the EPA threshold of **5 ppb**?*

Finally, compare and contrast your analyses and state whether your analysis suggests MDEQ is liable for providing false assurances of public safety.

The data:

When the source of Flint's water was first moved to Flint River, the [MDEQ took 71 samples from local homes](#). The `FlintWater` data set contains variables recorded for each of these 71 homes. The variables are:

- **house.id**: an ID number representing the house sampled
- **lead**: lead level (measured in parts per billion; ppb) in the water of the house that was sampled
- **warning**: an indicator of whether the lead level exceeded the EPA action threshold of 15 ppb ("Above EPA Threshold" or "Below EPA Threshold") Note: this is a binary representation of the quantitative variable `lead`

The assignment:

This case study is broken up into two parts: Part 1 (40 points), analyzing the MDEQ data; and Part 2 (60 points), writing up your results in a brief report.

1. **Data Analysis Task**: (Completed during lab - Worth 40 points, see below for rubric)
   a. **Descriptive analysis**: Perform a detailed exploratory data analysis (EDA) on the variables *lead* and *warning*. Create graphical representations of both variables. Calculate appropriate numerical summaries to describe the observed data distributions in terms of shape, center, and variability (as applicable).
   b. **Inferential analysis**: Conduct a z-test using the categorical coding of the MDEQ results and a t-test using the quantitative coding.
   c. NOTE: We recommend checking in with your GSI to confirm your results, such that you can correct any errors / improve your analyses before proceeding to your written report.

2. **Writing Task** (Completed outside of lab - Worth 60 points, see below for rubric)
   Write a report summarizing your findings from Part 1. Your report should provide a brief summary of the goals of your analysis, demonstrate a clear understanding of its results, and provide a statement that summarizes your opinion regarding whether your analysis suggests MDEQ officials were guilty of criminal negligence. Successful statistical reports should follow the structure recommended in the outline below and range between approximately 700 and 1000 words. **Submissions that are between 1100 and 1500 words will receive a 5-point penalty, and those that exceed 1500 words will receive a 10-point penalty.**

   i. **Introduction**: Describe the purpose of the analysis and justify its importance. Summarize the data that have been collected. Successful summaries will…
      a. state the observational units of the study, sample size, and describe all variables analyzed
      b. include complete descriptions/interpretations of the graphical and numerical displays (reference plots by their title)

   ii. **Analysis overview**:
   Summarize the methods and findings of both the z- and t-tests. Highlight and interpret the key results. A successful analysis overview will…
      a. define in context the parameters tested by each testing procedure
      b. report the results (test statistic & *p*-value) of each testing procedure
      c. *interpret* the test statistic *and p*-value of at least one of the testing procedures (note: no discussion of effect sizes is necessary)
      d. include a conclusion for each testing procedure with respect to the original research question (your conclusions should explicitly identify the level of evidence provided by the *p*-value)

   iii. **Discussion**:
   Provide a discussion of whether the analyses you run support MDEQ's initial conclusion that the water was safe to drink. A successful discussion will…
      a. evaluate whether the two tests provide concurring or contradictory results with respect to water safety
      b. if the tests lead to contradictory results, explain why that might be
      c. state whether your analysis suggests MDEQ officials were guilty of criminal negligence and justify your opinion
      d. Summarize the key takeaways about the importance of appropriate statistical test selection

Criteria for Success:

Rubric for Data Analysis Task (40 points)

		EDA - Categorical variable (10 points)	EDA - Quantitative variable (10 points)	z-test results (10 points)	t-test results (10 points)
Statistical Analysis	**Exemplary** Correctly addresses all elements of the prompt, demonstrating expert understanding and interpretations of required concepts / skills.	10	10	10	10
	Proficient Correctly addresses all elements of the prompt, demonstrating understanding and interpretations of required concepts / skills with just one minor error.	8	8	8	8
	Emerging Correctly addresses most elements of the prompt, demonstrating solid understanding and interpretations of required concepts / skills. May include a few minor computational/conceptual mistakes.	6	6	6	6
	Needs Improvements Demonstrates limited understanding of statistical concepts / skills. Large errors or conceptual mistakes are present.	2	2	2	2
	Missing Completely missing any attempted analysis.	0	0	0	0

Criteria for Success:

Rubric for the Written Report (60 points)

		Overall Report **(5 points)**		
Presentation	**Proficient** Report flows in a logical order, containing minimal grammatical or spelling errors. Writing is clear, creative, and informative.	5		
	Emerging Report includes passages that are vague, unclear, or include substantial grammatical or spelling errors.	3		
	Needs Improvements Report displays little logical order, suffering from many grammatical or spelling errors.	1		
		Introduction **(10 points)**	Analysis Overview **(20 points)**	Discussion **(25 points)**
Statistical Reasoning	**Exemplary** Correctly addresses all elements of the prompt, demonstrating expert understanding and interpretations of required concepts / skills.	10	20	25
	Proficient Correctly addresses all elements of the prompt, demonstrating solid understanding and interpretations of required concepts / skills. May include one or two minor conceptual mistakes.	9	18	22
	Fair Correctly addresses most elements of the prompt, demonstrating an understanding and interpretations of required concepts / skills. Includes several mistakes, one of which may be major.	7	15	18
	Emerging Correctly addresses some elements of the prompt, demonstrating an understanding and interpretations of required concepts / skills. May include more than two minor conceptual mistakes OR demonstrates major misunderstandings.	5	12	14
	Needs Improvements Demonstrates very limited understanding of statistical concepts / skills. Includes more than one major misunderstanding in core concepts.	3	8	10