**Lab 8 – Flint Water**

**NAME 1 – NETID**

**NAME 2 – NETID [if applicable]**

**NAME 3 – NETID [if applicable]**

**Formatting Instructions**

* Please submit your lab report as a **pdf** to Gradescope.
* When you upload to Gradescope, please **match pages** with the **question number**.
* Be sure that all **group members** are **added** in your submission to Gradescope (click view/edit group on the top right of the page once shown your final submission after matching pages).

**Assignment Overview**

* We’ll be investigating some data on water samples taken from Flint, Michigan in 2016, during [the water crisis.](https://www.cdc.gov/nceh/casper/pdf-html/flint_water_crisis_pdf.html) We have the following analytical goals for this lab:
  + Evaluate the extent to which lead levels are unsafe across a representative sample of homes in the area
  + Investigate whether letting the faucet “flush” for a period of time lowers lead levels
  + Determine whether lead levels vary across wards (i.e., regions of the city).

**STEP 0**

* **Pre-lab work:** Complete the pre-lab tutorials (“Customizing ggplot2” and “if else statements”) for Lab 8 first: <https://stat212-learnr.stat.illinois.edu/>
* **Download** the flint.xlsx file to your computer and then **import** into your RStudio session.
  + On the pop-up screen when importing, **click the drop-down under “Ward” and change it to a character variable**. Then click “Import” on the bottom right.
  + *If using* ***RMarkdown***, put the following code after your import code to allow Ward to be read categorically. Flint$Ward = as.character(Flint$Ward)
* Create a new **R script** (or use the **RMarkdown file** if you are using that option)
* Remember to **library(tidyverse)** so that you can use the ggplot function.

**Data Description**

This data set includes lead content measurements taken from tap water across 300 homes in Flint, Michigan (of which 269 homes’ measurements are included). Researchers collected 3 water samples from each household: the water at first draw (faucet turned on), water after running the faucet for 45 seconds, and water after running the faucet for 2 minutes. Lead content is measured in parts per billion (ppb). The spreadsheet is organized such that one water sample is the unit of observation; there are 3 units of observation per household.

As a point of reference,lead measurements **above 5ppb** are considered ***somewhat unhealthy* for regular consumption,** and lead measurements **above 15ppb** are considered ***dangerous* for regular consumption.**

**Variables**

**SampleID:** Household number. There are 269 households that provided data.

**Zip\_Code:** Household’s zip code

**Ward:** The regional zone that the household was in. A ward is like a precinct. <https://www.cityofflint.com/city-of-flint-ward-map/>

**Time:** The time point at which the water sample was taken: First draw, after 45 seconds, or after 2 minutes.

**Lead\_ppb:** The lead concentration in the water sample, measured in parts per billion (ppb)

**Question 1 (5pts):** To get started, let’s visualize our response variable. Create a density curve that plots the distribution of the Lead\_ppb variable.

* Use a fill color
* Add a title, and label the x axis to state “Lead in parts per billion (ppb)”
* Scale the x axis to have more frequent tick marks than is shown by default (you be the judge!)

**Include an image of your graph in your report.** *Code is optional—partial credit for code attempt, but no graph.*

**Report the numeric summary result of this variable** (min, Q1, Q2, mean, Q3, max).

**Briefly describe the shape of this distribution**

**Question 2 (5pts):** Now, let’s createjitter plots to visually compare the lead levels at the three different time points. Have each group be represented as a column of jittered points.

* Jitter your points with a width around 0.05 or 0.1
* Set an transparency (alpha) level between around 0.2 to 0.4
* Color each column differently, and customize the colors (manual, or with a color palette)
* Add a title, and label the lead axis to “Lead in parts per billion (ppb)”
* Use a plot theme

**Report your graph in your report.** *Code is optional—partial credit for code attempt, but no graph.*

**Question 3 (5pts):** Using **a pipe**, let’s create a summary table to compare lead concentrations, grouped by the timing of the water sample. Your summaries of lead concentrations should include the following:

* The median lead level, *rounded to 3 decimal places*
* The mean lead level, *rounded to 3 decimal places*
* Proportion of water samples with a lead level above 5ppb, *rounded to 3 decimal places*
* Proportion of water samples with a lead level above 15ppb, *rounded to 3 decimal places*

Report your **code***and* report your **summary results** (you can copy/screenshot them as they are, or create a table in word if you prefer)

Notice that even though the 45 second group has a much lower median than the first draw group, it has almost the same mean. Consider your previous graph and briefly explain**: Which group do you think is producing the consistently highest lead measurements? What might be causing that spike in the mean for the other group?**

**Question 4 (5pts):** For sake of visualizing, let’s narrow in on where most of the data is.

Create side by side **boxplots** of these same two variables and…

* Color each boxplot a different color. Use custom colors (or a color palette)
* Add a title, and label the lead axis to “Lead in parts per billion (ppb)”
* Use the scale function to have the y axis go in increments of 5 and have **limits** from 0 to 50
* Add a plot theme background
* *Remove* the color legend (you can do this using the theme() function).

*Note the warning about “rows containing non-finite values” just means data points were excluded because they were outside the 0 to 50 range you limited your graph to—we did that on purpose!*

**Report your graph in your report.** *Code is optional—partial credit for code attempt, but no graph.*

**Question 5 (5pts):** Let’s now compare lead concentrations across Wards (precincts). Using **a pipe**, let’s create a summary table to compare lead concentrations across Wards. Filter the data to *only include* ***first-draw*** *observations*. Your summaries of lead concentrations should include the following:

* The median lead level, *rounded to 3 decimal places*
* The mean lead level, *rounded to 3 decimal places*
* Proportion of water samples with a lead level above 5ppb, *rounded to 3 decimal places*
* Proportion of water samples with a lead level above 15ppb, *rounded to 3 decimal places*

Report your **code***and* report your **summary results** (you can copy/screenshot them as they are, or create a table in word if you prefer)

**Question 6 (5pts):** Using a pipe, filter to only include first draw measurements. Then inside this pipe, build side-by-side jitter plots to compare lead ppb (at first draw) across each ward. *The reason we changed Ward to a factor variable was for this graph.*

* Color the points from each Ward a different color. **Use a color palette** for this one
* Add a title, and label the lead axis to “Lead in parts per billion (ppb)”
* Use the scale function to have the y axis go in increments of 5 and have **limits** from 0 to 50
* Add a plot theme background
* *Remove* the color legend (you can do this in the theme() function).

**Report your graph in your report.** *Code is optional—partial credit for code attempt, but no graph.*

Now that you have visualized and summarized the data, **do you notice much difference in lead concentration across wards?** If you had to pick, **which two or three wards seem to have the worst lead contamination?**

**Question 7 (5pts):** Now, let’s make a graph that compares lead levels across Wards and across times together. Keep Ward on the x axis, and now map Time as the fill color.

* Use the boxplot geometry for this one
* Color (or fill) each boxplot a different color. Use a color palette for this one
* Add a title, and label the lead axis to “Lead in parts per billion (ppb)”
* Use the scale function to have the y axis go in increments of 5 and have limits from 0 to 50
* Add a plot theme background

Be sure that you make the viewer window nice and big to make everything clearly visible before copying into your report.

**Report your graph in your report.** *Code is optional—partial credit for code attempt, but no graph.*

If you were advising a resident in Flint who hadn’t had their water tested for lead, **could you give them any data-based advice about how to maximize their safety if using water from their faucet?**