Lab 8 – Flint Water

Name: Yiran Hu, netid: yiranhu3

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/ (https://stat212-learnr.stat.illinois.edu/)

- Load `tidyverse` package.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────────────── tidy
verse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────────────────────
───── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
me errors
```

- Download the Flint.xlsx file to your computer and then load it into your RMarkdown file. You might also need to library `readxl` to do that.

```
library(readxl)
Flint = read_excel("Flint.xlsx")
# Note, your data file should be saved in the same folder as your RMarkdown.
# Also match the name of the data reference above with the name of the data file. For example,
it might save as Flint.xlsx, or Flint_1_.xlsx or something else.
```

- Use the following code after your import code to allow `Ward` to be read categorically.
  `Flint$Ward = as.character(Flint$Ward)`

```
Flint$Ward = as.character(Flint$Ward)

#View(Flint)
```

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**.

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/ (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)

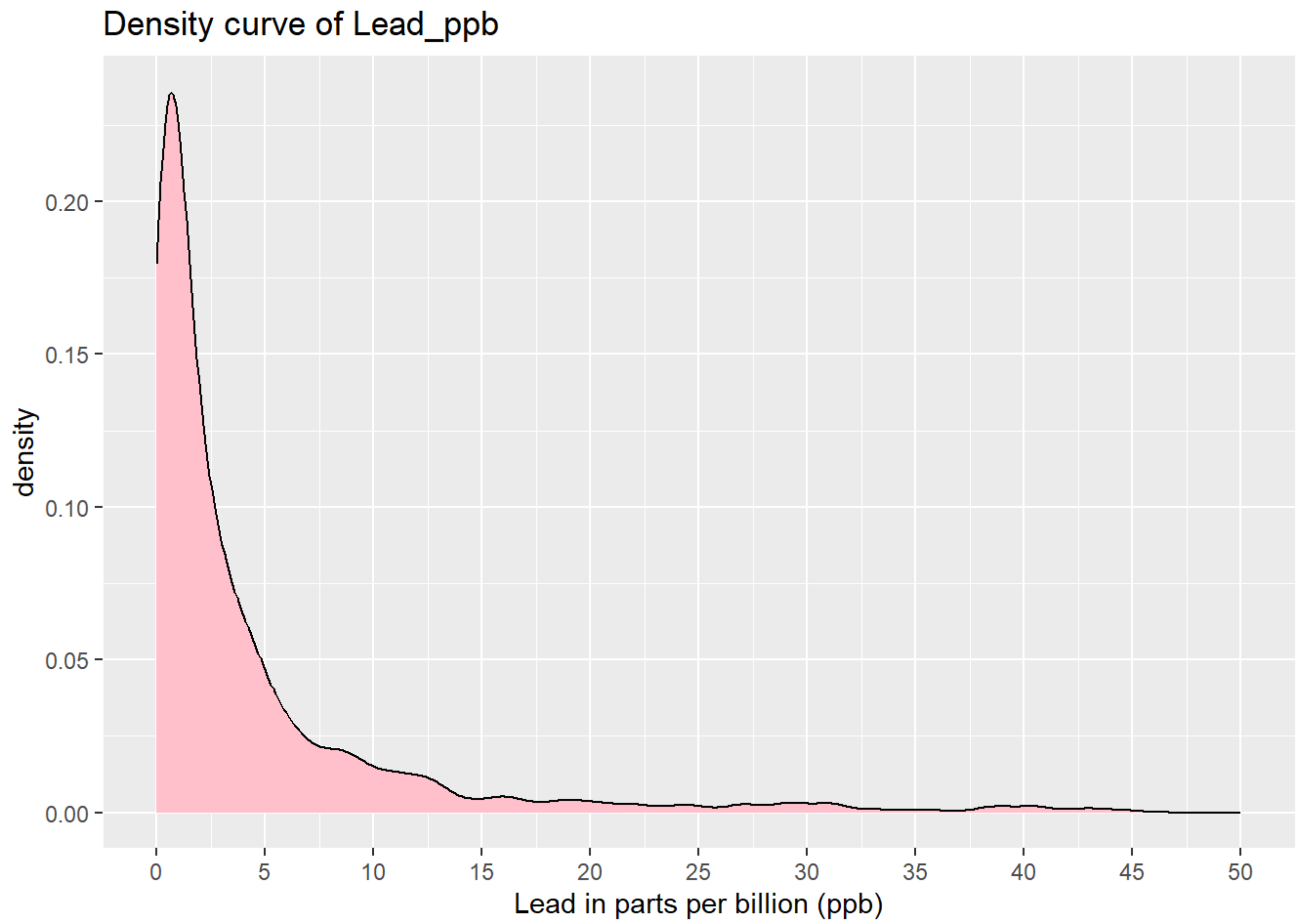
Question 1

To get started, let's visualize what our variable of interest looks like. 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)

```
ggplot(data=Flint, aes(x=Lead_ppb))+
  geom_density(fill="pink")+
  labs(title="Density curve of Lead_ppb", x="Lead in parts per billion (ppb)")+
  scale_x_continuous(breaks = seq(0,50,5), limits = c(0,50))
```

```
## Warning: Removed 21 rows containing non-finite values (`stat_density()`).
```



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

```
summary(Flint$Lead_ppb)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.031    0.509    1.843    8.208    5.155 1051.000
```

min: 0.031, Q1: 0.509, Q2: 1.843, mean: 8.208, Q3: 5.155, max: 1051.000

Describe the shape of this distribution.

The distribution of the variable “Lead_ppb” is highly right skewed.

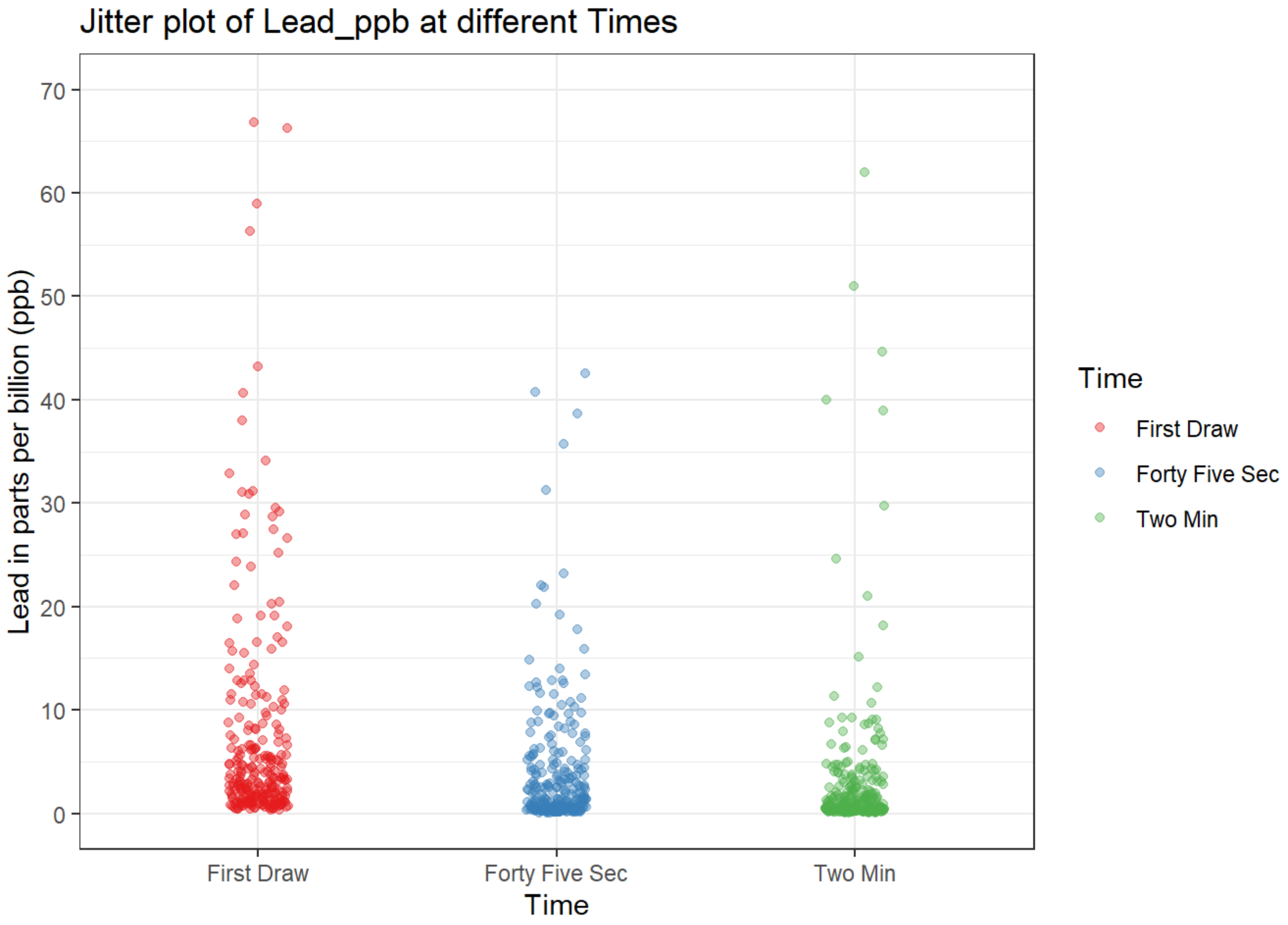
Question 2

Now, let’s create jitter 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

```
ggplot(data=Flint, aes(x=Time, y=Lead_ppb, color=Time))+
  geom_jitter(width=0.1, alpha=0.4)+
  scale_color_brewer(palette="Set1")+
  scale_y_continuous(breaks = seq(0,70,10), limits = c(0,70))+
  theme_bw()+
  labs(title="Jitter plot of Lead_ppb at different Times", y="Lead in parts per billion (ppb)")
```

```
## Warning: Removed 15 rows containing missing values (`geom_point()`).
```



Question 3

Using a pipe, create a summary table to compare lead levels, separated in rows by the three time points of data (first draw, 45 seconds, and 2 minutes). Your summaries of lead levels 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

```
Flint |>
  group_by(Time) |>
  summarise(Median_Lead_Level = round(median(Lead_ppb), digits = 3),
            Mean_Lead_Level = round(mean(Lead_ppb), digits = 3),
            Lead_above5 = round(mean(Lead_ppb>5), digits = 3),
            Lead_above15=round(mean(Lead_ppb>15),digits = 3))
```

```
## # A tibble: 3 × 5
##   Time          Median_Lead_Level Mean_Lead_Level Lead_above5 Lead_above15
##   <chr>                     <dbl>           <dbl>       <dbl>        <dbl>
## 1 First Draw                3.48            10.7        0.401        0.167
## 2 Forty Five Sec            1.4             10.3        0.245        0.063
## 3 Two Min                   0.831            3.68       0.123        0.045
```

Notice that the first draw and 45-second groups have nearly identical means. 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?**

I would say the “First Draw” group is producing the consistently highest lead measurements. The high mean lead level in “Forty Five Sec” group may be caused by several very high measurements. For example, one point is above 1000 ppb.

Question 4

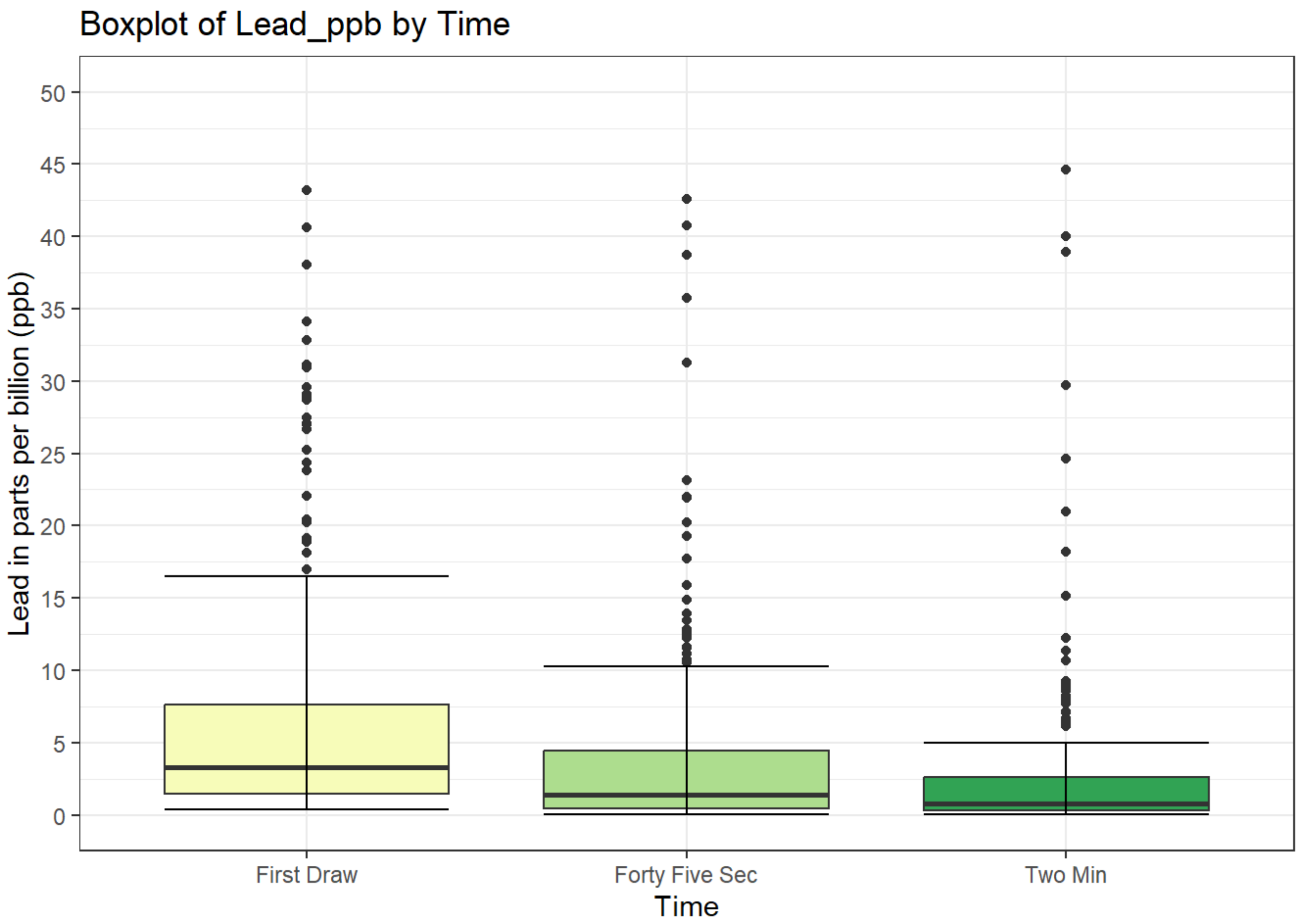
For sake of visualizing, let’s narrow in on where most of the data is. Create side by side **boxplots** of these same two variables and…

- Color (or fill) each boxplot differently and 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 in 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!

```
ggplot(data=Flint, aes(x=Time, y=Lead_ppb, fill=Time))+
  geom_boxplot()+
  stat_boxplot(geom = "errorbar") +
  scale_fill_brewer(palette = "YlGn")+
  scale_y_continuous(breaks = seq(0,50,5),limits = c(0,50))+
  theme_bw()+
  theme(legend.position = "none")+
  labs(title="Boxplot of Lead_ppb by Time", y="Lead in parts per billion (ppb)")
```

```
## Warning: Removed 21 rows containing non-finite values (`stat_boxplot()`).
## Removed 21 rows containing non-finite values (`stat_boxplot()`).
```



Question 5

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

```
Flint |>
  filter(Time==″First Draw″) |>
  group_by(Ward) |>
  summarise(Median_Lead_Level=round(median(Lead_ppb), digits = 3),
            Mean_Lead_Level=round(mean(Lead_ppb), digits = 3),
            Proportion_above5=round(mean(Lead_ppb>5), digits = 3),
            Proportion_above15=round(mean(Lead_ppb>15), digits = 3))
```

```
## # A tibble: 9 × 5
##   Ward  Median_Lead_Level Mean_Lead_Level Proportion_above5 Proportion_above15
##   <chr>             <dbl>           <dbl>             <dbl>              <dbl>
## 1 1                  2.36            4.07             0.129              0.065
## 2 2                  2.95            9.82             0.417              0.167
## 3 3                  3.44           12.5              0.318              0.136
## 4 4                  1.97            8.62             0.314              0.114
## 5 5                  2.38           10.3              0.389              0.222
## 6 6                  3.68           20.1              0.484              0.29
## 7 7                  5.22           11.1              0.526              0.237
## 8 8                  5.40           14.8              0.552              0.207
## 9 9                  3.83            6.62             0.439              0.098
```

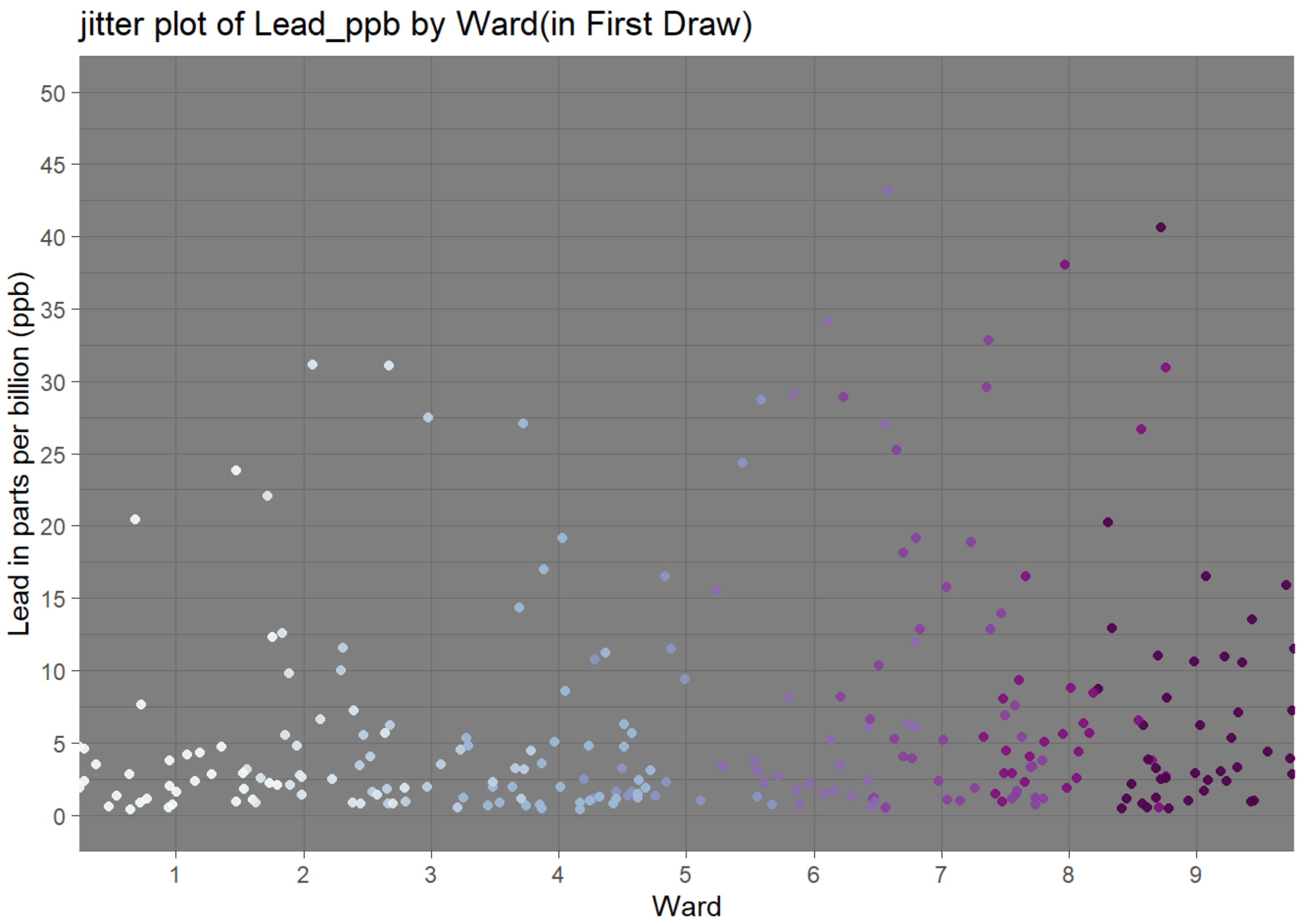
Question 6

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).

```
Flint |>
  filter(Time=="First Draw") |>
  ggplot(aes(x=Ward, y=Lead_ppb, color=Ward))+
  geom_jitter(width=0.8, alpha=0.9)+
  scale_color_brewer(palette="BuPu")+
  scale_y_continuous(breaks = seq(0,50,5),limits=c(0,50))+
  theme_dark()+
  theme(legend.position = "none")+
  labs(title = "jitter plot of Lead_ppb by Ward(in First Draw)", y="Lead in parts per billion
(ppb)")
```

```
## Warning: Removed 12 rows containing missing values (`geom_point()`).
```



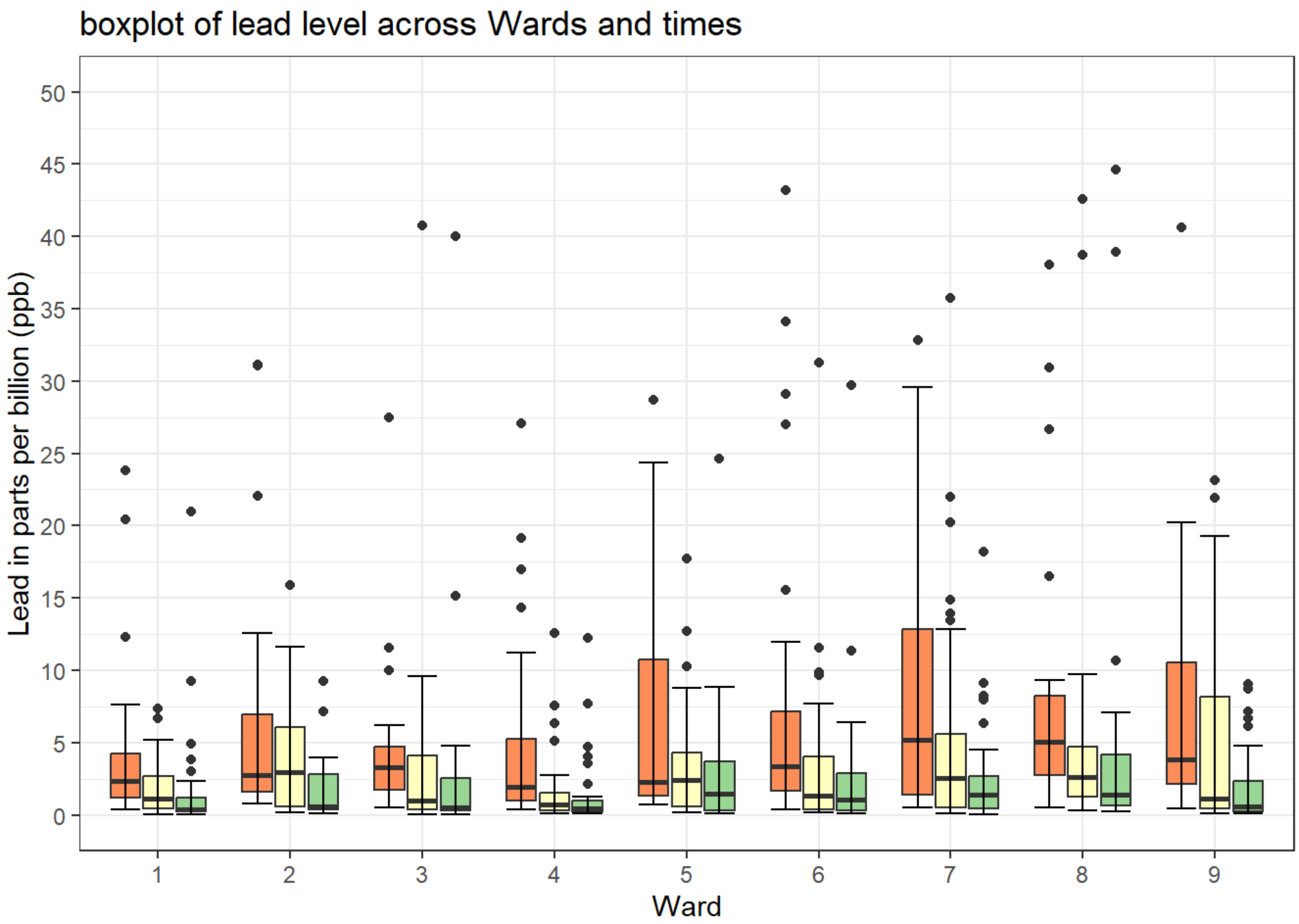
Question 7

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
- Remove the color legend (you can do this in the theme() function).

```
ggplot(data=Flint, aes(x=Ward, y=Lead_ppb, fill=Time))+
  geom_boxplot()+
  stat_boxplot(geom = "errorbar")+
  scale_fill_brewer(palette="Spectral")+
  scale_y_continuous(breaks = seq(0,50,5), limits = c(0,50))+
  labs(title = "boxplot of lead level across Wards and times", y="Lead in parts per billion (pp
b)")+
  theme_bw()+
  theme(legend.position = "none")
```

```
## Warning: Removed 21 rows containing non-finite values (`stat_boxplot()`).
## Removed 21 rows containing non-finite values (`stat_boxplot()`).
```



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?**

To maximize safety, based on the data collected across the 9 Wards, people should use water after running the faucet for 2 minutes or more. Though this may waste a lot of water, the data shows a tendency that if running the faucet longer, the lead remains will get lower.