SS21 STT 180 Homework 2

Emani Hunter

Feb 27-March 13, 2021

Contents

General Instructions:

- This is an individual assignment. You may consult with others as you work on the assignment, but each student should write up a separate set of solutions.
- Rather than creating a new Rmd file, just add your solutions to the supplied Rmd file. Submit **both** the Rmd file and the resulting HTML/PDF file to D2L.
- Except for questions, or parts of questions, that ask for your commentary, use R in a code chunk to answer the questions.
- The code chunk option `echo = TRUE` is specified in the setup code chunk at the beginning of the document. Please do not override this in your code chunks.
- A solution will lose points if the Rmd file does not compile. If one of your code chunks is causing your Rmd file to not compile, you can use the `eval = FALSE` option. Another possibility is to use the `error = TRUE` option in the code chunk.
- This Homework is due on **Saturday, March 13, 2021 on or before 11 pm.**
- Kindly submit **both** the .rmd and the HTML/PDF output files. If you submit the output in html format, zip the files while submitting.

Setting up:

Load `tidyverse`, which includes `dplyr`, `ggplot2`, `tidyr`, and other packages, and the load ‘knitr.

```
library(tidyverse)
library(knitr)
```

Homework 2 has two sections. In Section 1 you will use data visualization and write function to analyze a dataset.For Section 2 you will read an article, explore the data, validate the claims and come to own conclusions.

Section 1

For the first section of this homework will use the same birth dataset you used for Homework 1. Please use the `BirthDataWithRegionColors.csv` file for this HW. The dataset contains information about births in the United States. The full data set is from the Centers for Disease Control. The data for this homework assignment is a “small” sample (chosen at random) of slightly over one million records from the full data set. The data for this homework assignment also only contain a subset of the variables in the full data set.

Introduction

Read in the data and convert the data frame to a tibble.

```
birth_data <- read.csv("BirthDataWithRegionColors.csv", header = TRUE)
birth_data <- as_tibble(birth_data)
```

A glimpse of the data:

```
Rows: 1,048,575
Columns: 10
$ year          <int> 1969, 1969, 1969, 1969, 1969, 1969, 1969, 1969, 1969,...
$ month         <int> 2, 3, 5, 5, 5, 6, 8, 8, 11, 11, 11, 3, 3, 6, 7, 10, 3...
$ state         <chr> "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA", "CA",...
$ is_male       <lgl> TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE, FALSE, F...
$ weight_pounds <dbl> 6.999677, 6.876218, 7.187070, 7.749249, 7.374463, 9.6...
$ mother_age    <int> 20, 25, 30, 17, 22, 26, 26, 19, 25, 26, 26, 31, 28, 2...
$ child_race    <int> 1, 2, 1, 1, 4, 2, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 2,...
$ plurality     <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
$ region        <chr> "West", "West", "West", "West", "West", "West", "West...
$ state_color   <chr> "blue", "blue", "blue", "blue", "blue", "blue", "blue...
```

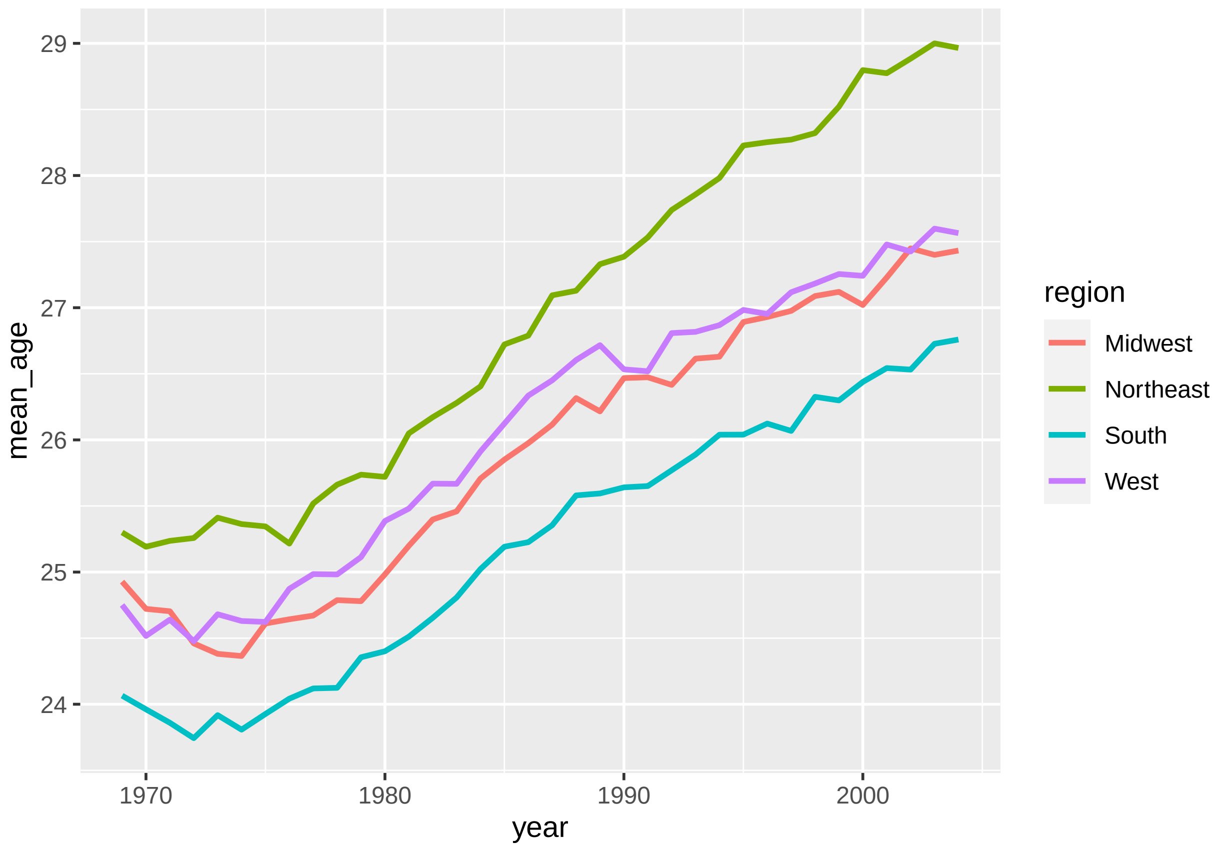
The variables in the data set are:

Variable	Description
`year`	the year of the birth
`month`	the month of the birth
`state`	the state where the birth occurred, including “DC” for Washington D.C.
`is_male`	which is `TRUE` if the child is male, `FALSE` otherwise
`weight_pounds`	the child’s birth weight in pounds
`mother_age`	the age of the mother
`child_race`	race of the child.
`plurality`	the number of children born as a result of the pregnancy, with 1 representing a single birth, 2 representing twins, etc.

Combine `dplyr`with `ggplot2` to create graphical displays of the data. Use `filter`, `group_by`, and `summarize` build the required data frame.

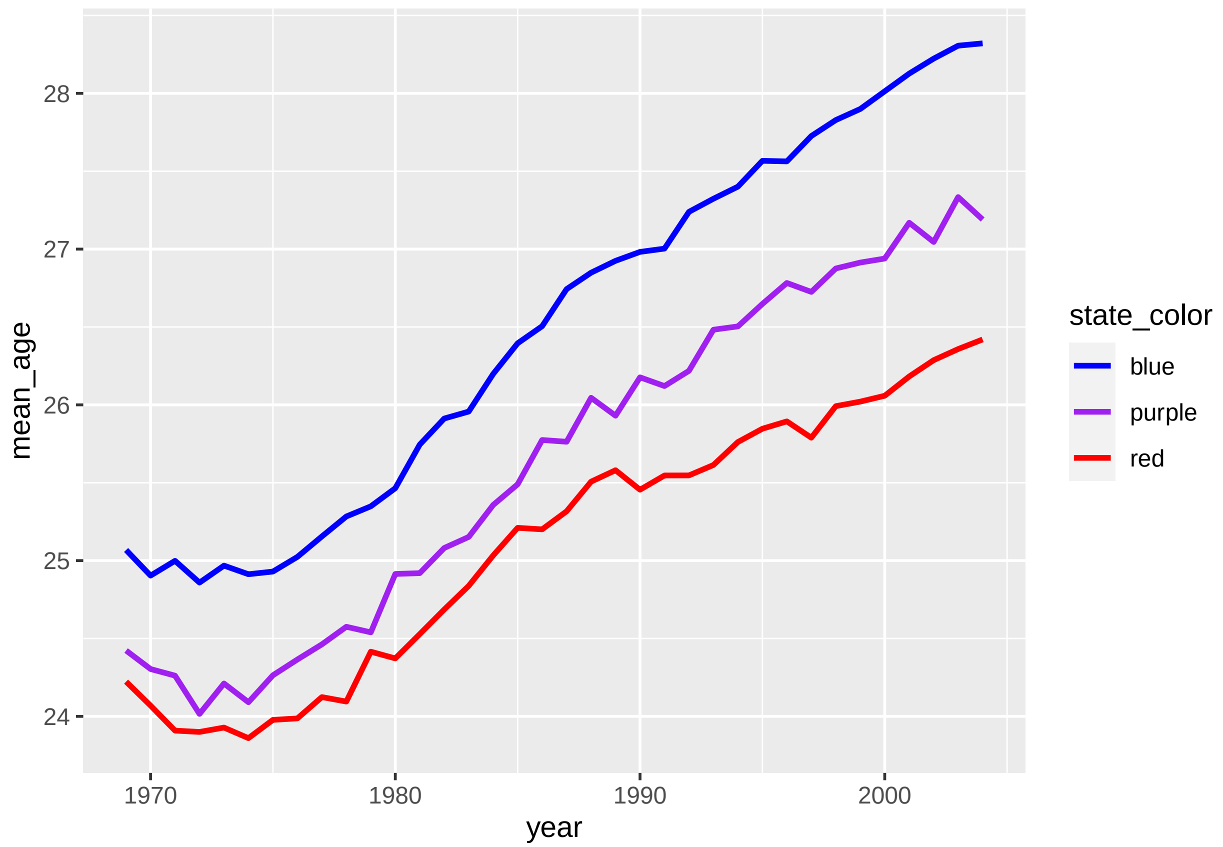
Question 1

Create a plot of mean age of mother versus year, which includes separate lines for each of the four regions of the country. (Don’t include data for which the region is missing.) The graphic should look like the following.



Question 2

Create a graphic of mean age of mother versus year, which includes separate lines for each of the three values of `state_color`. (Don't include data for which `state_color` is missing.) The graphic should look like the following. Notice that the colors are different from the default colors.

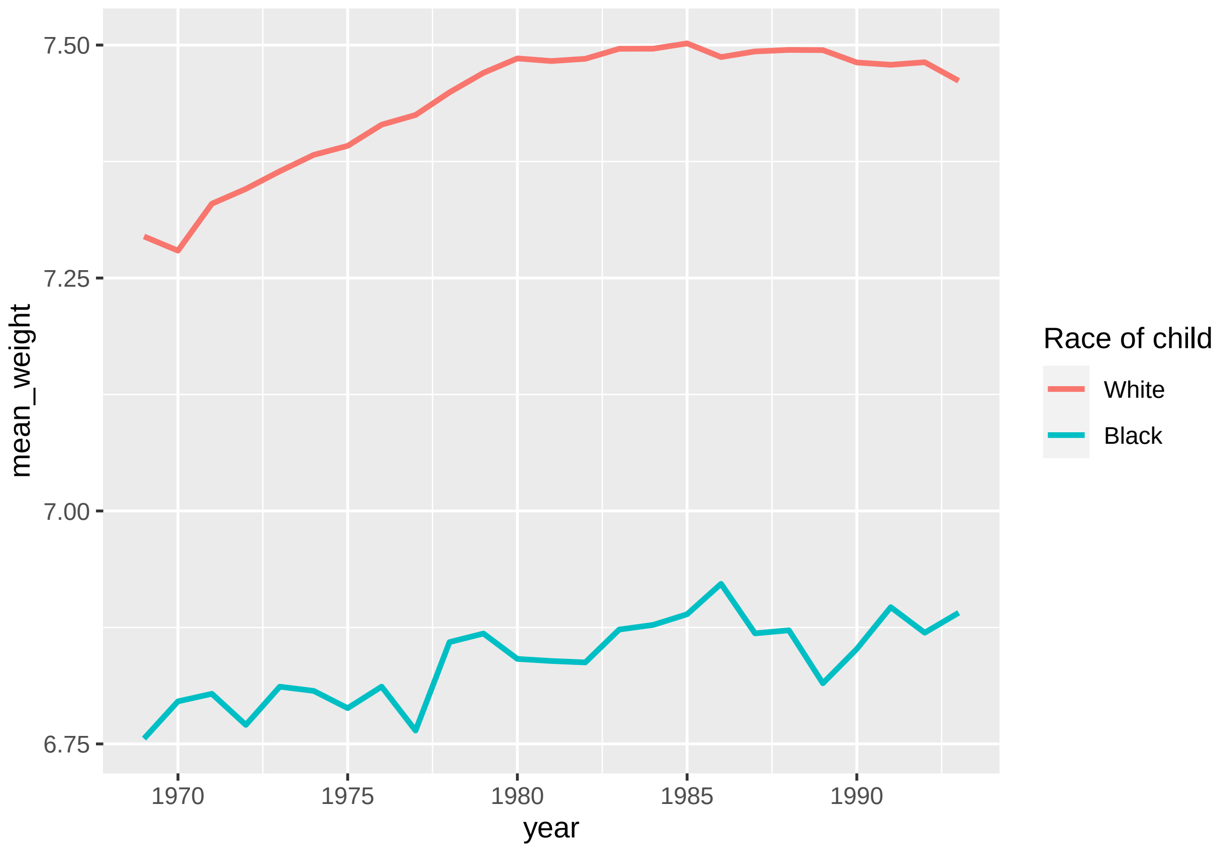


Write 2-3 sentences comparing Question 1 and Question 2.

The mean mother age per year with respect to region plot and the mean mother age per year with respect to state color plot are similar in distribution. It makes sense that the distribution would be about the same due to each corresponding region being in its respective state color. The midwest and northeast regions have a state color of blue, so looking at the blue line distribution, it is apparent that the line is a culmination of mean ages for both regions. Most of the south regions has a state color of red and most of the west (not all) region has a state color of purple.

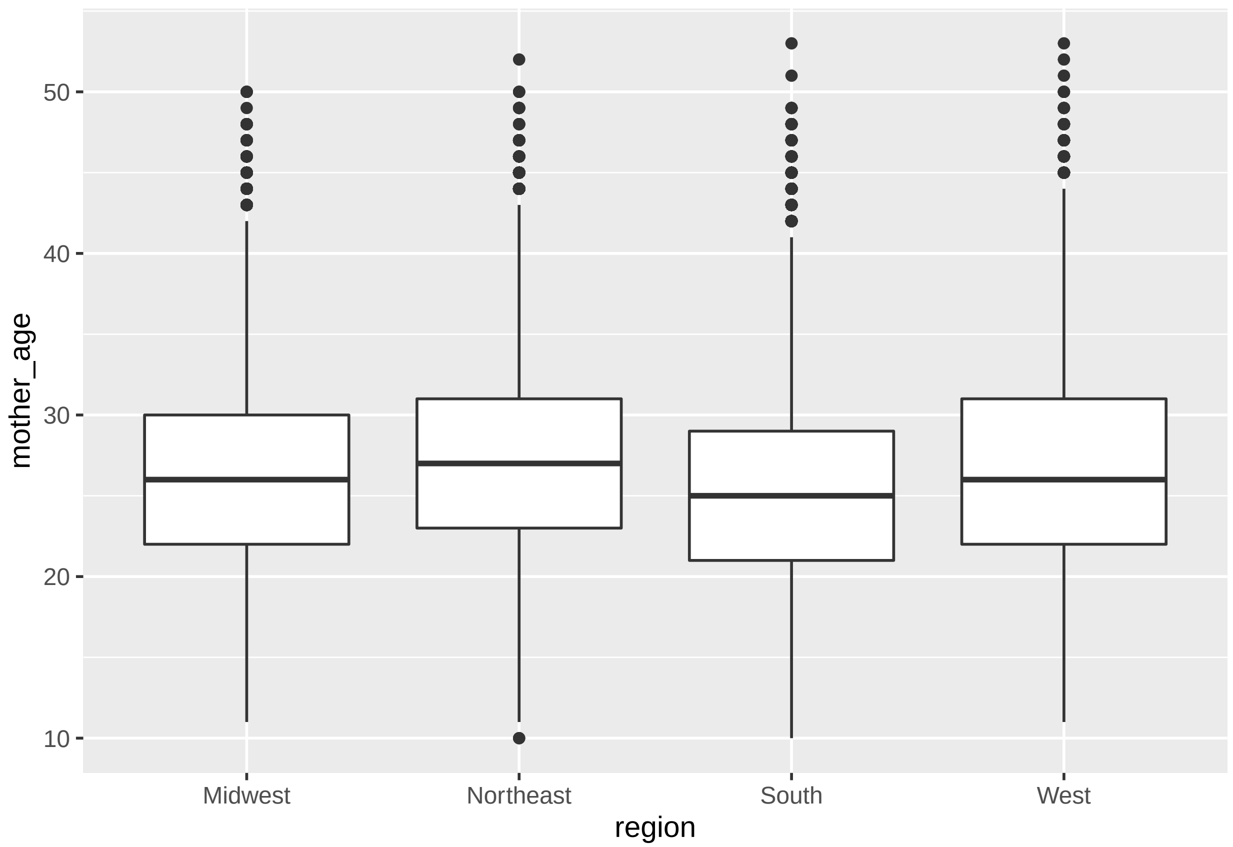
Question 3

Create a graphic of mean weight of the child versus year, which includes separate lines for the two top race categories, white and black. The graphic should look like the following. Notice that the legend is different from the default legend. You'll want to investigate `scale_color_discrete` to change the legend.



Question 4

Create a graphic showing side-by-side boxplots of the age of the mother for the four regions. (Don't include data for which `region` is missing.) The graphic should look like the following.



Question 5

Write a function called `quantitative_summary` which takes two inputs:
`x`: A numeric vector
`group`: A factor vector of the same length as x

and produces a **list** as output which contains the following elements:

`missing`: The number of missing values in x
`means`: The means of x for each level of groups.
`sds`: The standard deviations of x for each level of groups
`is.binary`: Set to FALSE for for this function

Here is an example of the function in action.

```
$missing
[1] 1583

$means
   FALSE     TRUE 
7.178759 7.438649 

$sds
   FALSE     TRUE 
1.283197 1.353845 
```

```
$is.binary
[1] FALSE
```

Hint:

- When computing the means and standard deviations, you need to exclude missing values using `na.rm`.

Question 6

Write a function called `binary_summary` which takes two inputs:

`x`: A vector containing the values 0 and 1 (possibly NA)
`group`: A factor vector of the same length as x

and produces a **list** as output which contains the following elements:

`missing`: The number of missing values in x
`prop`: The proportion of 1s in x for each level of groups
`is.binary`: Set to TRUE for for this function.

Here is an example of the function in action using plurality defined as a binary variable (single vs multiple births):

```
$missing
[1] 27869

$prop
   group
x        FALSE       TRUE
  1 0.47408950 0.50015186
  2 0.01280192 0.01295672

$is.binary
[1] TRUE
```

Section 2

Flint is the second poorest city of its size in the United States and has spent six of the past 15 years in a state of financial emergency. One of the cost-cutting measures taken by emergency managers was to stop buying water, sourced from Lake Huron, from the Detroit Water and Sewerage Department. Instead, Flint would use the Flint River for its water supply while waiting for a new pipeline to Lake Huron to be opened. The move was expected to save roughly $5 million over a period of two years.

The Flint River supply was switched on in April 2014. Not long after, problems arose.Flint resident and mother of four LeeAnne Walters noticed that the water coming out of her taps was orange. More worryingly, her family's hair was falling out, her preschool sons had broken out in rashes and one of them had stopped growing.

The orange colour was from iron, but the family's symptoms pointed to a far more dangerous contaminant: lead. (Langkjaer - Bain 2017)

Introduction

The data set consists of 271 homes sampled with three water lead contaminant values at designated time points. The lead content is in parts per billion (ppb). Additionally, some location data is given about each home.

To get started, read in the `flint.csv` file using the function `read.csv`, as was done in the ICA with the cereal data. However, you do not need to use the `attach` function. The data set has five variables:

- **id**: sample id number
- **zip**: zip code in Flint as to the water sample's location
- **ward**: ward in Flint as to the water sample's location
- **draw**: water sample at one of three time points
- **lead**: lead content in parts per billion

Before you get started, read *The murky tale of Flint's deceptive water data* by Langkjaer - Bain (2017).

```
flint <- read.csv("flint.csv")
glimpse(flint)
```

```
Rows: 813
Columns: 6
$ id                        <int> 1, 2, 4, 5, 6, 7, 8, 9, 12, 13, 15, 16, 17, ...
$ zip                       <int> 48504, 48507, 48504, 48507, 48505, 48507, 48...
$ ward                      <int> 6, 9, 1, 8, 3, 9, 9, 5, 9, 3, 9, 5, 2, 7, 9,...
$ draw                      <chr> "first", "first", "first", "first", "first",...
$ lead                      <dbl> 0.344, 8.133, 1.111, 8.007, 1.951, 7.200, 40...
$ flushing_time..in.sec. <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

Question 1

Select one passage that you found particularly striking (perhaps you strongly agreed or disagreed with it, perhaps it made you question an assumption or seemed unclear) in the article and write a 4-5 sentence paragraph commenting on it.

One particular passage, "Sampling Errors" caught my attention a lot. It was stated that the MDEQ allowed for procedures for sampling the water to underestimate the amount of lead that was in the water. It's quite disappointing that the city and MDEQ would allow this and try to falsely claim that the water supply "isn't that bad". I strongly disagree with the way the MDEQ collected and handled their data. The MDEQ also seemed to dismiss concerns that were raised by scientists and tried to make it seem as though what they were doing was perfectly fine. The MDEQ was fully aware of what they were doing and it is truly sad to know that they allowed the water supply be as contaminated as it was for so long.

Question 2

How many unique zip codes are in the data set? How many unique wards are in the data set?

```
[1] 8
```

```
[1] 10
```

There are 8 unique zip codes and 10 unique wards in the dataset.

Do the number of wards in the data set match how many wards Flint has? If not, suggest a way to handle this discrepancy.

No, the number of wards in the data set do not match the number of wards Flint has. Flint has 9 wards. A way to handle this discrepancy is to assume that ward 0 is not a ward. There can only be between 1 and 9 wards, so, we must dismiss/ drop off the values that correspond to ward 0 in the dataset as this may be a typo within the dataset.

Question 3

Which ward appears to have the worst water quality? Note that your answer should consider mean, median, and maximum lead levels. Your choice of 'worst ward' should include justification for why some of these statistics are more important to consider than others.

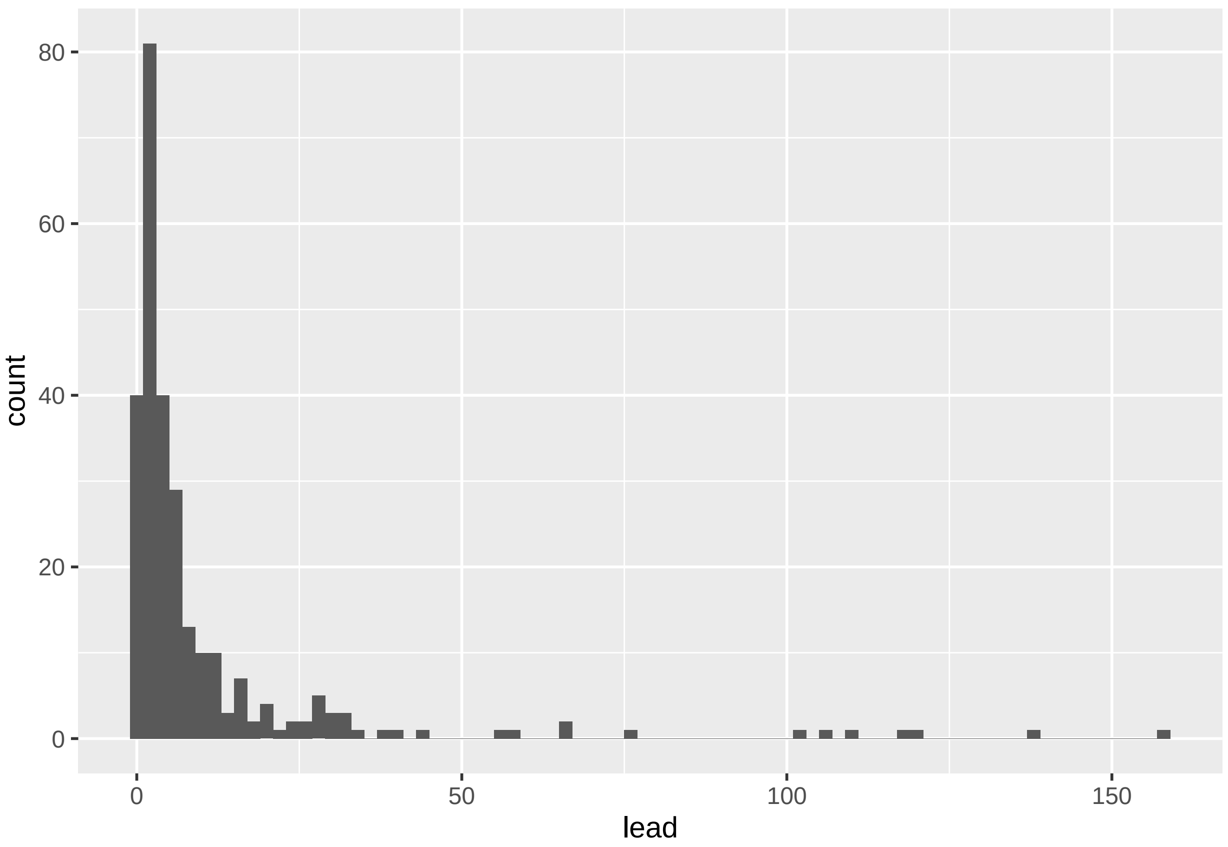
```
# A tibble: 10 x 4
    ward mean_lead med_lead max_lead
 * <int>     <dbl>    <dbl>    <dbl>
 1     0      2.19    0.953     4.88
 2     1      2.58    1.17     23.8
 3     2     24.2     2.49   1051
 4     3      6.71    1.87    118.
 5     4      3.88    0.896   139.
 6     5      5.97    2.25     66.2
 7     6     12.9     2.11    240.
 8     7      6.55    2.48    105.
 9     8     10.5     3.1     158
10     9      4.83    1.93     51.0
```

Ward 2 appears to have the worst water quality. Ward 2 has the largest maximum lead value and the largest mean lead value compared to all other wards. The maximum lead value, in my opinion, is the most important value to consider as it showcases the highest lead content found in the water supply within the ward. A maximum lead value of 1051 is extremely high compared to the second highest value of 239.7 for ward 6.

Question 4

Langkjear-Bain (2017) writes at length about the practice of 'drawing' water before sampling it for lead levels. Compute the median and mean lead values for each draw. How do they compare? Create a histogram of the lead values for just the first draw and comment on the histogram's shape – does it confirm the earlier relationships between mean and median?

```
# A tibble: 3 x 3
  draw   mean_draw med_draw
* <chr>      <dbl>    <dbl>
1 first      10.6     3.52
2 second     10.3     1.4
3 third       3.66    0.831
```



The mean draw value for the first and second draw are fairly close, however, the third draw mean value is extremely low compared to the first and second. The median is smaller than the mean in all “drawing” cases, so, it could be assumed that the data will be skewed. The median is also quite low for both the second and third draw.

Looking at the histogram for the first draw, we can see that the data is strongly skewed to the right. It is also apparent that the median is less than the mean and that there are some fairly large values that drive the mean upward. This confirms the relationship between the mean and median computed above.

Question 5

Compute the sample quantile for the *85th percentile* of lead values for each draw. Comment on what you observe. Is any draw above the EPA action threshold level?

```
 85%
16.5
```

```
  85%
8.837
```

```
  85%
4.517
```

Looking at the 85th percentile quantiles, it can be seen that the first draw has the largest lead sample quantile value. Going from first draw to third draw, the values drop by approximately 2 times the previous draw amount. The first draw was well above the 10% EPA action threshold level, being 16.5%.

Question 6

Recreate the below plot based on data from zip code **48505**.

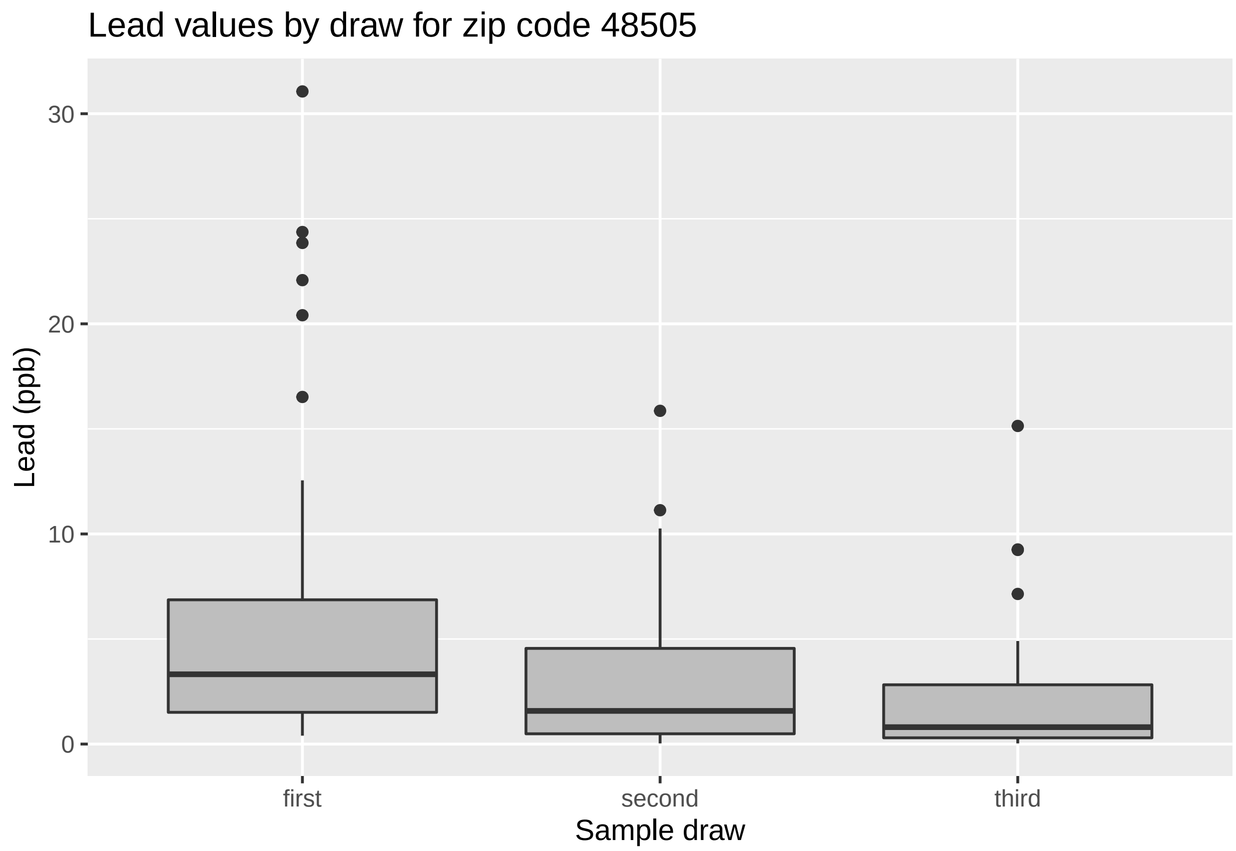
In 1-2 sentences, comment on whether the plot confirms or contradicts the statement below, pulled from Langkjear-Bain (2017)

“Pre-stagnation flushing” – as it is known – “may potentially lower” lead levels as flushing “removes water that may have been in contact with the lead service line for extended periods”

```r
unique_zip <- flint %>%
              filter(zip == 48505)

plt_zip <- ggplot(data = unique_zip, mapping = aes(x = draw, y= lead)) + geom_boxplot(fill = "gray") +

plt_zip
```



The plot confirms the statement. We can see that after each flush during a 5 minute time period, the lead levels decrease. ### Question 7

What is the largest lead value? What draw and zip code does it belong to? Comment on how we should handle this value if further statistical analysis were to be performed.

```
    id   zip ward   draw lead flushing_time..in.sec.
356 97 48504    2 second 1051                     45
```

The largest lead value is 1051 ppb. The draw is the second draw and the value corresponds to zip code 48504. This value may be handled as an outlier if further statistical analysis were to be performed. If no values exceed 500 ppb, it should be assumed that this one value is an outlier deviating from the mean by a significant amount.

What is the smallest lead value? What draw and zip code does it belong to?

```
     id   zip ward  draw  lead flushing_time..in.sec.
670 141 48505    1 third 0.031                    120
```

The smallest lead value is 0.031 ppb. The draw is the third draw and the value corresponds to zip code 48505.

Question 8

One way to standarize the data is to use z-scores. Based on each draw, compute z-scores for the lead values. How many z-scores exceed three in absolute value for each draw?

```
[1] 5.071895 4.589538 4.390102 4.269512 3.022804 5.943849 4.997687 6.834355
```

```
[1]  3.396929 15.410626  3.694569
```

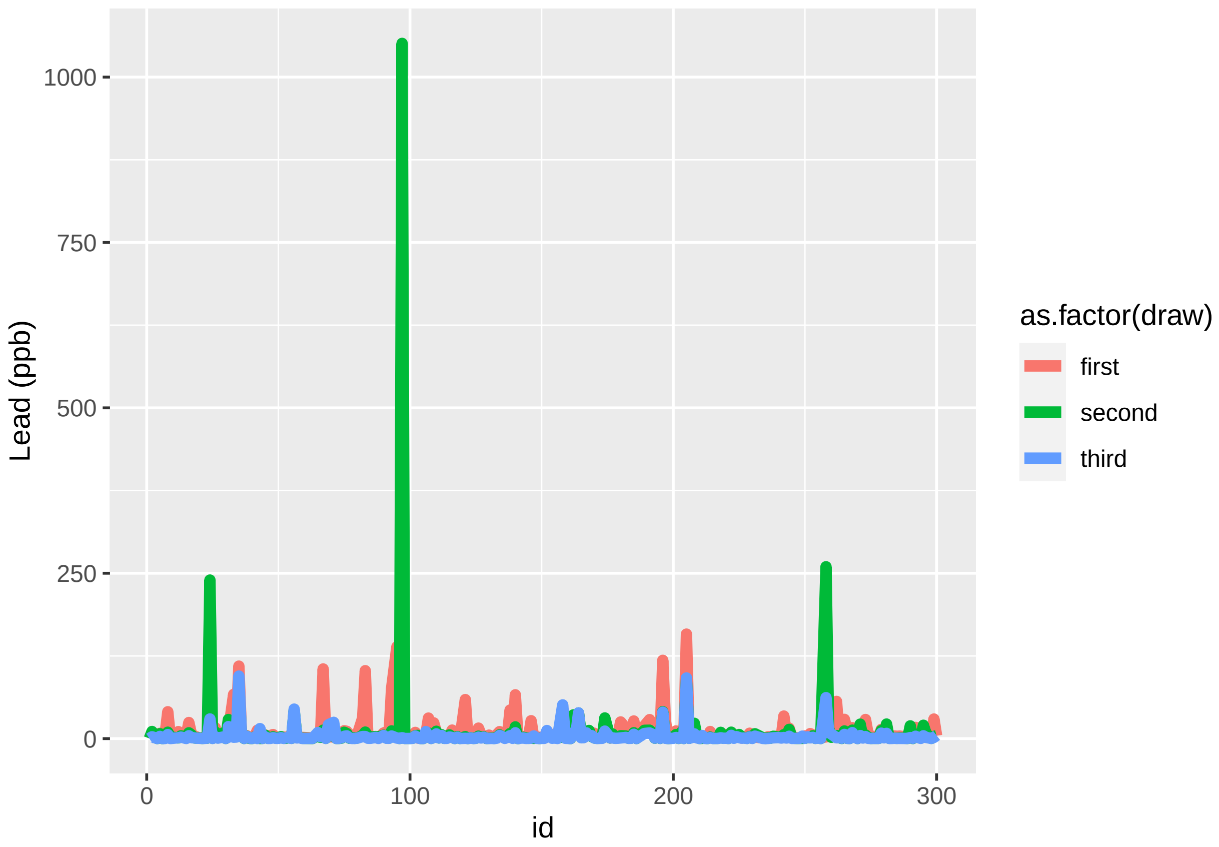
```
[1] 8.621653 3.884736 4.489187 3.347658 3.447293 8.353114 5.532030
```

```
[1] 18
```

18 z-scores exceed three in absolute value for each draw.

Question 9

Based on your analysis in questions 1-8, does it seem that flushing the water decreases the lead content? You may include further code and visualizations.



Based on my analysis in questions 1-8. It seems as though flushing the water decreasse the lead content. As we observe each draw from first to third, the lead values drop by a reasonable amount (though there are some outliers within the data).

Essential details

Deadline and submission instructions

- The deadline to submit Homework 2 is **11:00pm on Saturday, 13 March, 2021.**
- This is a individual assignment.
- Submit your work by uploading **both** your RMD and HTML/PDF files through D2L. Kindly double check your submission to note whether the everything is displayed in the uploaded version of the output in D2L or not. If submitting HTML outputs, please zip the files for submission.
- Kindly ensure that **the `echo=TRUE` is set in the every chunk option.**
- Late work **will not be accepted** except under certain extraordinary circumstances.

Help

- Post general questions in the Teams HW 1 channel. If you are trying to get help on a code error, explain your error in detail
- Feel free to visit us in during our virtual office hours or make an appointment.

- Communicate with your classmates, but do not share snippets of code.
- **The instructional team will not answer any questions within the first 24 hours of this homework being assigned, and we will not answer any questions after 6 P.M of the due date}.**

Academic integrity

This is an individual assignment. You may discuss ideas, how to debug code, or how to approach a problem with your classmates.You may also post your general questions in the HW2 channel in Teams.But you may not copy-and-paste another individual's code from this class. As a reminder, below is the policy on sharing and using other's code.

> Similar reproducible examples (reprex) exist online that will help you answer many of the questions posed on in-class assignments, pre-class assignments, homework assignments, and midterm exams. Use of these resources is allowed unless it is written explicitly on the assignment. You must always cite any code you copy or use as inspiration. Copied code without citation is plagiarism and will result in a 0 for the assignment.

Grading

Use the R Markdown blank file that is provided. If you want, you can use your own formatting. Self-formatting is at your discretion but is graded. Use the in-class assignments and resources available online for inspiration. Another useful resource for R Markdown formatting is available at: https://holtzy.github.io/Pimp-my-rmd/

Topic	**Points**
Questions(total 15)	75
R Markdown formatting and knitting	7
Communication of results	10
Code style	8
Total	**100**

Please note: Code style includes code efficiency.

Reference

1. http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm
2. Langkjr-Bain, R. (2017), The murky tale of Flint's deceptive water data. Significance, 14: 16-21.
3. https://holtzy.github.io/Pimp-my-rmd/