

SS21 STT 180 Homework 2

Feb 27-March 13, 2021

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, 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...

```

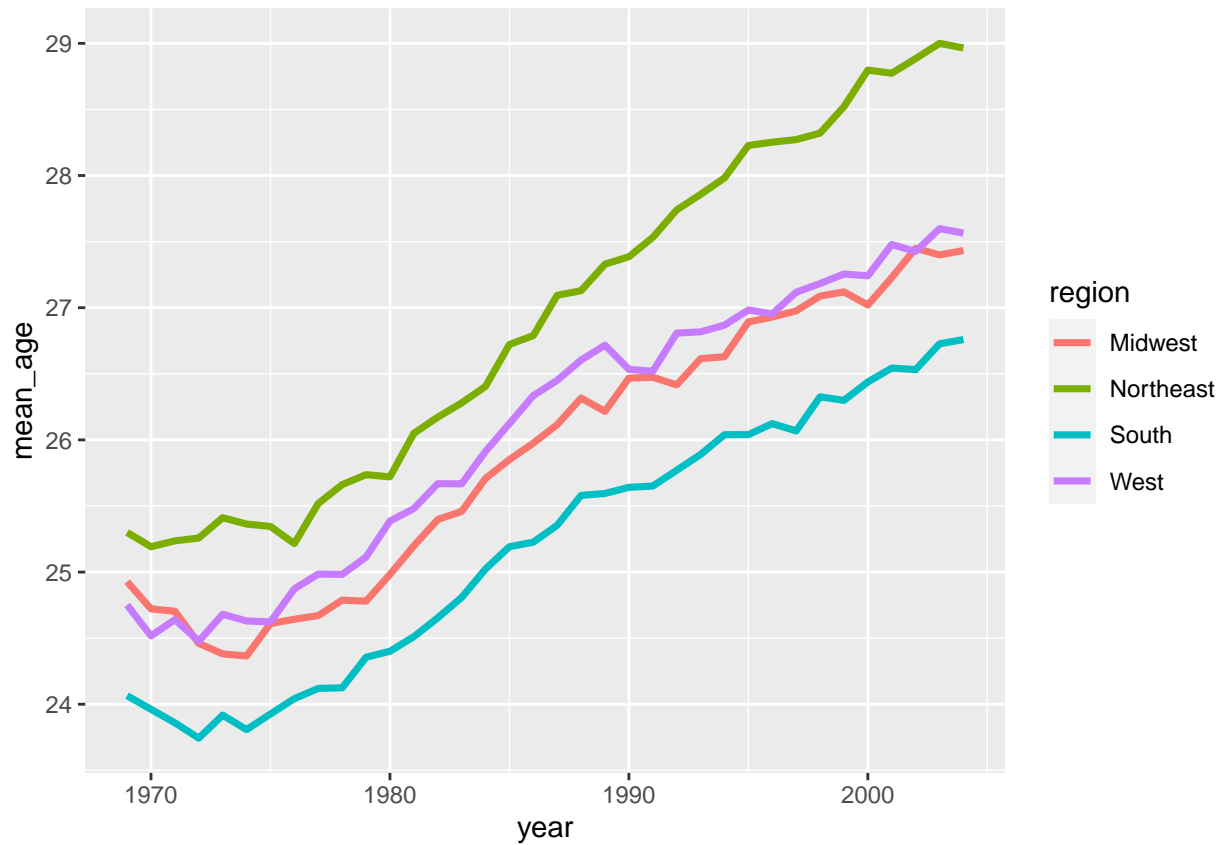
The variables in the data set are:

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.

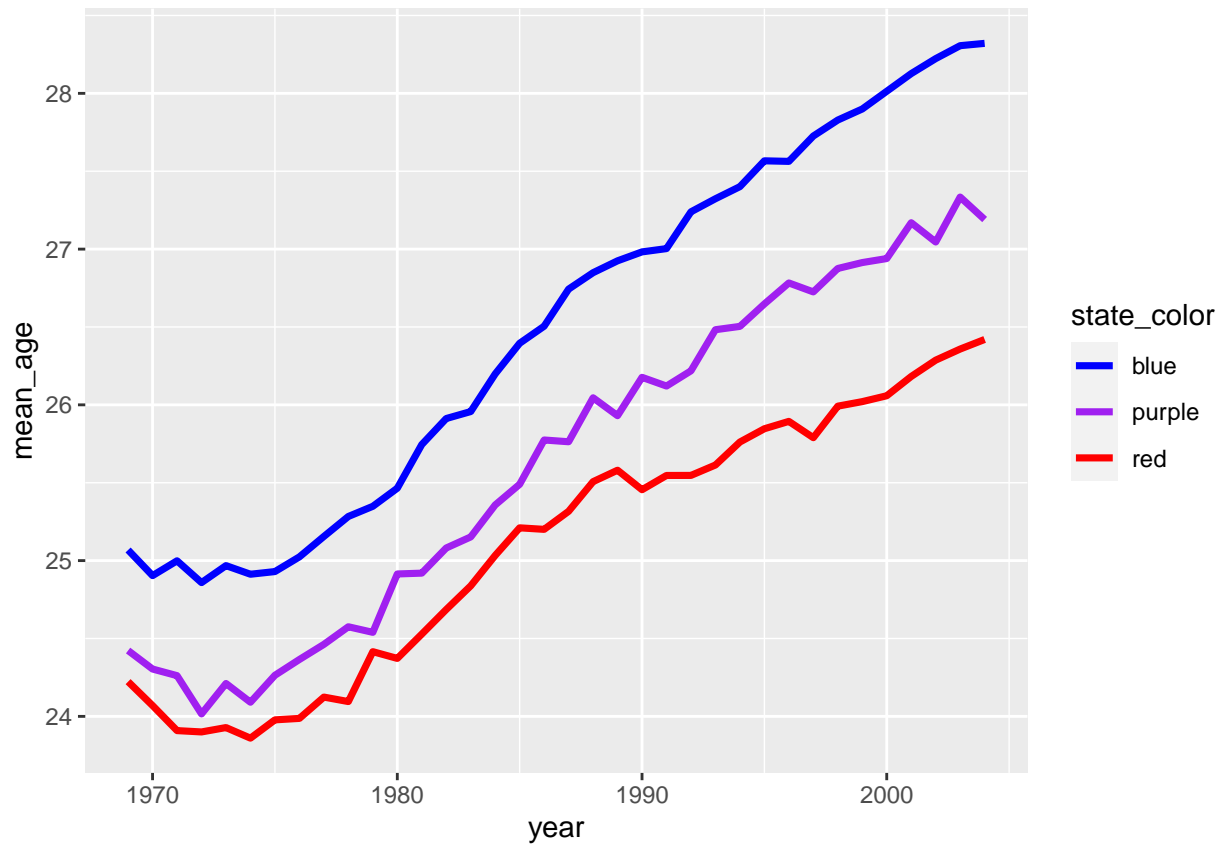
Question 1

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

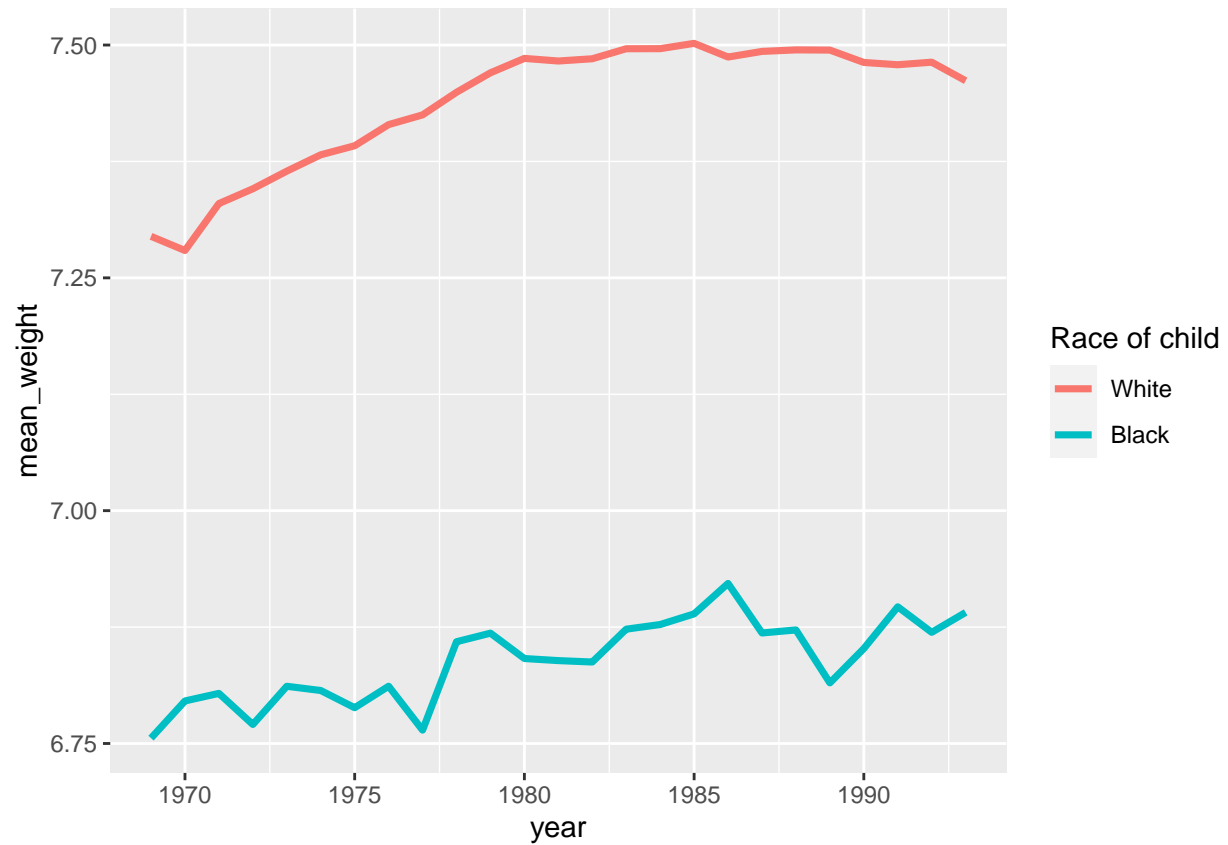
Create a graphic of mean age of mother versus year, which includes separate lines for each of the three values of `state_color`. (Don't include data for which `state_color` is missing.) The graphic should look like the following. Notice that the colors are different from the default colors.



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

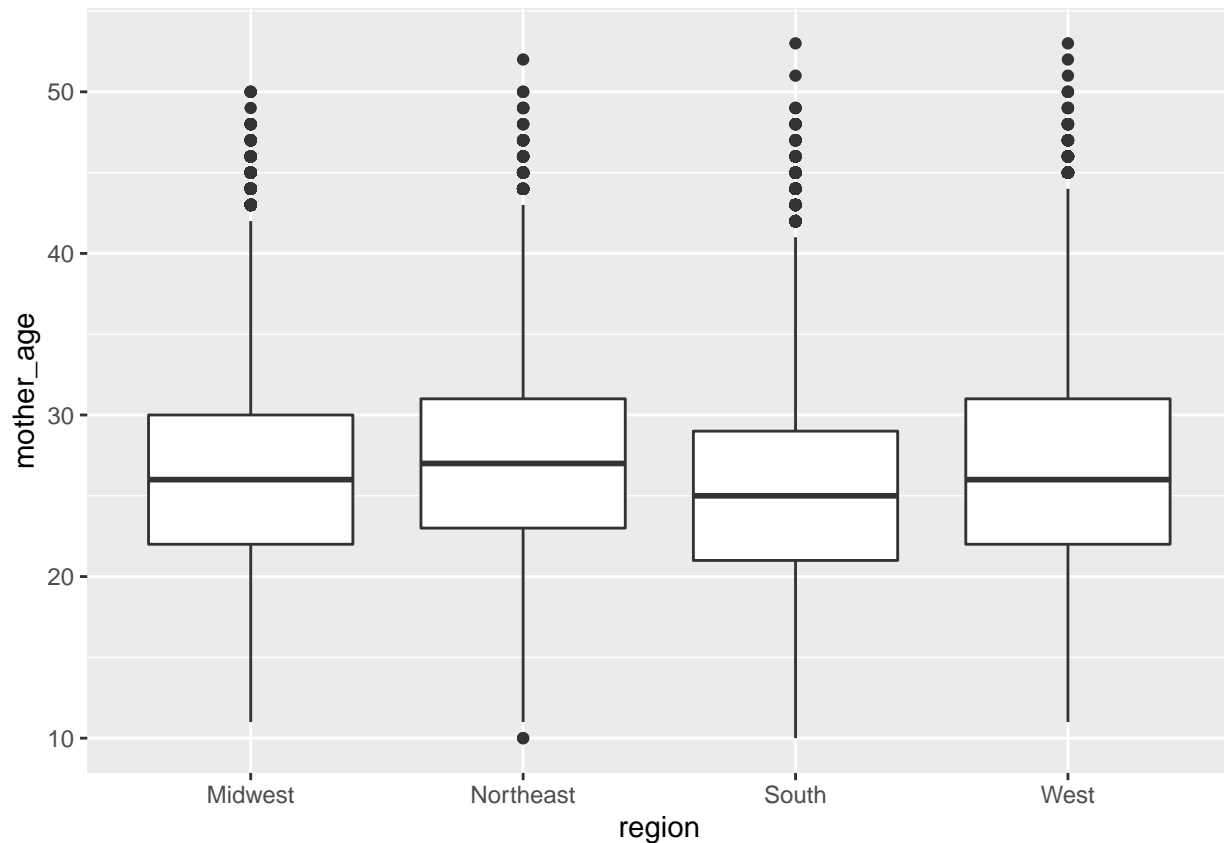
Question 3

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.

```
quantitative_summary(birth_data$weight_pounds, birth_data$is_male)
```

```
$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

`sds`: The standard deviations of `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):

```
binary_summary(birth_data$plurality1, birth_data$is_male)
```

```
$missing  
[1] 27869
```

```
$prop  
  groups  
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,...
```

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.

Question 2

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

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

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.

Question 4

Langkjaer-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?

Question 5

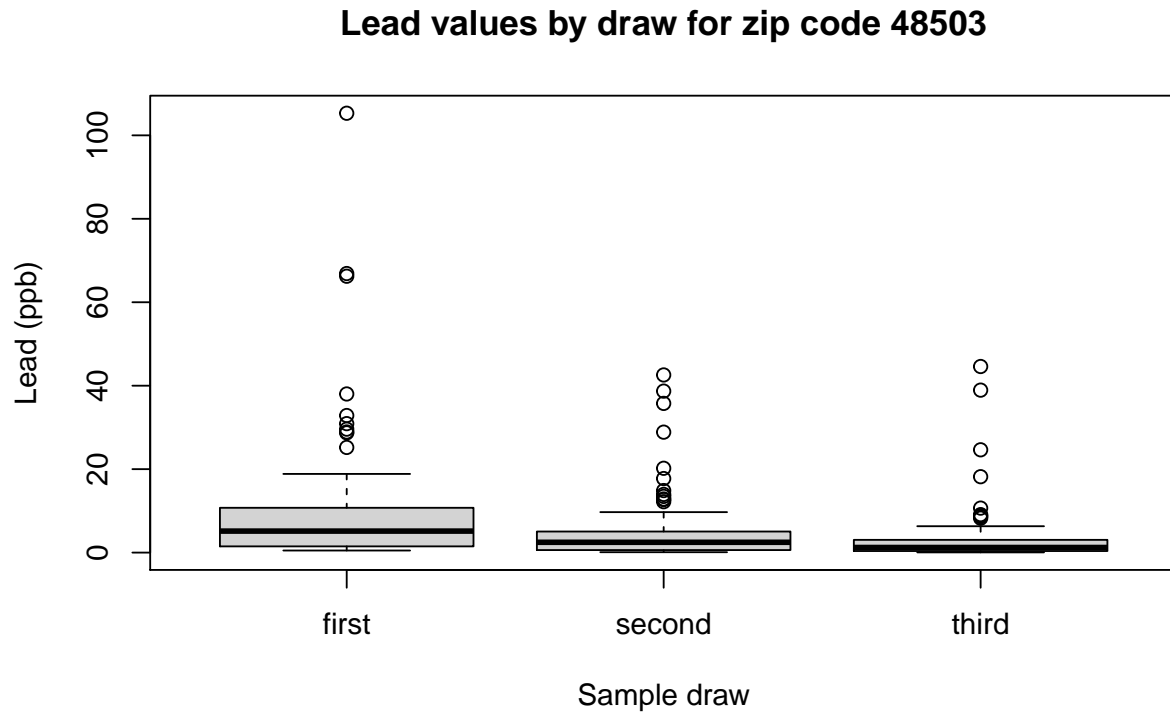
Compute the sample quantile for the 85th percentile of lead values for each draw. Comment on what you observe. Is any draw above the EPA action threshold level?

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 Langkjaer-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”



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.

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

Question 8

One way to standardize 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?

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.

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. Langkjaer-Bain, R. (2017), The murky tale of Flint's deceptive water data. *Significance*, 14: 16-21.
3. <https://holtzy.github.io/Pimp-my-rmd/>