# Midterm Exam: Take-Home Component

Taylor Blair

Math 141, Week 7

## Instructions

- You have two consecutive hours to complete this exam. **Make sure to fill-in your start day/time and end day/time:**

**Start day/time:**

**End day/time:**

- Once you have completed the exam, turn in the knitted pdf on Gradescope.
- For this component, you are allowed to consult any materials from this course: the textbook, your notes, class handouts, lab assignments, worksheets, project assignments, RStudio cheatsheets, RStudio help files.
- You are not allowed to consult anybody (classmates, friends, family, teachers, random people on the street, etc.) for help on this exam. You must work on it alone.
- Do **NOT** post questions about the take-home exam to any Slack channels or anywhere else on the internet.
- Send technical questions to Kelly, Tom, and Jonathan as a joint Direct Message. For equity reasons, we will not answer conceptual or clarifying questions.
- Remember the Honor Principle.

## The Data

In April of 2014, the water source for the city of Flint, Michigan was changed from Lake Huron to the Flint River. The river water ate through the protective film layer inside the city water pipes. As a result, lead from the pipes leached into the Flint drinking water.

To better understand the levels of lead content in the drinking water, researchers from Virginia Tech worked with residents to collect water samples from households in Flint. The researchers then tested the water for its lead content. For each residence, here are the variables we have access to:

- `sample_id`
- `zip_code`
- `ward`: Wards are subsections of the city
- `lead_initial`: Amount of lead, in parts per billion (ppb), for the first liter of water from the faucet at a normal flow
- `lead_45_sec_flushing`: Amount of lead, in parts per billion (ppb), in 1 liter of water from the faucet at a normal flow **after running the water for 45 seconds**
- `lead_2_min_flushing`: Amount of lead, in parts per billion (ppb), in 1 liter of water from the faucet at a normal flow **after running the water for 2 minutes**

Note: High risk homes are defined by the EPA as homes where the water samples show lead higher than 15 ppb.

```
library(tidyverse)
flint_mi_water <- read_csv("/home/courses/math141f19/Data/flint_mi_water.csv")
```

# Problems

Load the libraries you will need here (beyond `tidyverse` which is provided in the previous chunk).

```
library(dplyr)
library(ggrepel)
```

### Problem 1

a. Remove the row from the dataset that is in Ward 0. **For all future problems, use this subsetted data.**

```
flint_mi_water <- flint_mi_water[flint_mi_water["ward"]!=0, ]

flint_mi_water
```

```
## # A tibble: 270 x 6
##    sample_id zip_code  ward lead_initial lead_45_sec_flushing lead_2_min_flushi~
##        <dbl>    <dbl> <dbl>        <dbl>                <dbl>              <dbl>
##  1         1    48504     6        0.344                0.226              0.145
##  2         2    48507     9        8.13                10.8                2.76
##  3         4    48504     1        1.11                 0.11               0.123
##  4         5    48507     8        8.01                 7.45               3.38
##  5         6    48505     3        1.95                 0.048              0.035
##  6         7    48507     9        7.2                  1.4                0.2
##  7         8    48507     9       40.6                  9.73               6.13
##  8         9    48503     5        1.1                  2.5                0.1
##  9        12    48507     9       10.6                  1.04               1.29
## 10        13    48505     3        6.2                  4.2                2.3
## # ... with 260 more rows
```

b. What does each row in the dataset represent?

---

- A high risk house
- Does not include subsection 0 of the city

---

c. For each variable in the dataset, identify its type.

---

- `sample_id`
  - Quanitative
  - represents a data point
- `zip_code`
  - Categorical
  - We have a limited set, can be treated as quanitative

- `ward`:
  - Categorical
  - Represents a subsection of the city
- `lead_initial`:
  - Quanitative
  - First sample parts per million
- `lead_45_sec_flushing`:
  - Quanatative
  - Represents PPM
- `lead_2_min_flushing`:
  - Quanatative

---

**Problem 2**

a. Create an `epa_lead_initial` column that denotes whether or not a household exceeded the EPA limit of 15 ppb of lead during the initial water sample (given in `lead_initial`).

```
flint_mi_water <- flint_mi_water %>%
  mutate(epa_lead_initial= lead_initial>15)
```

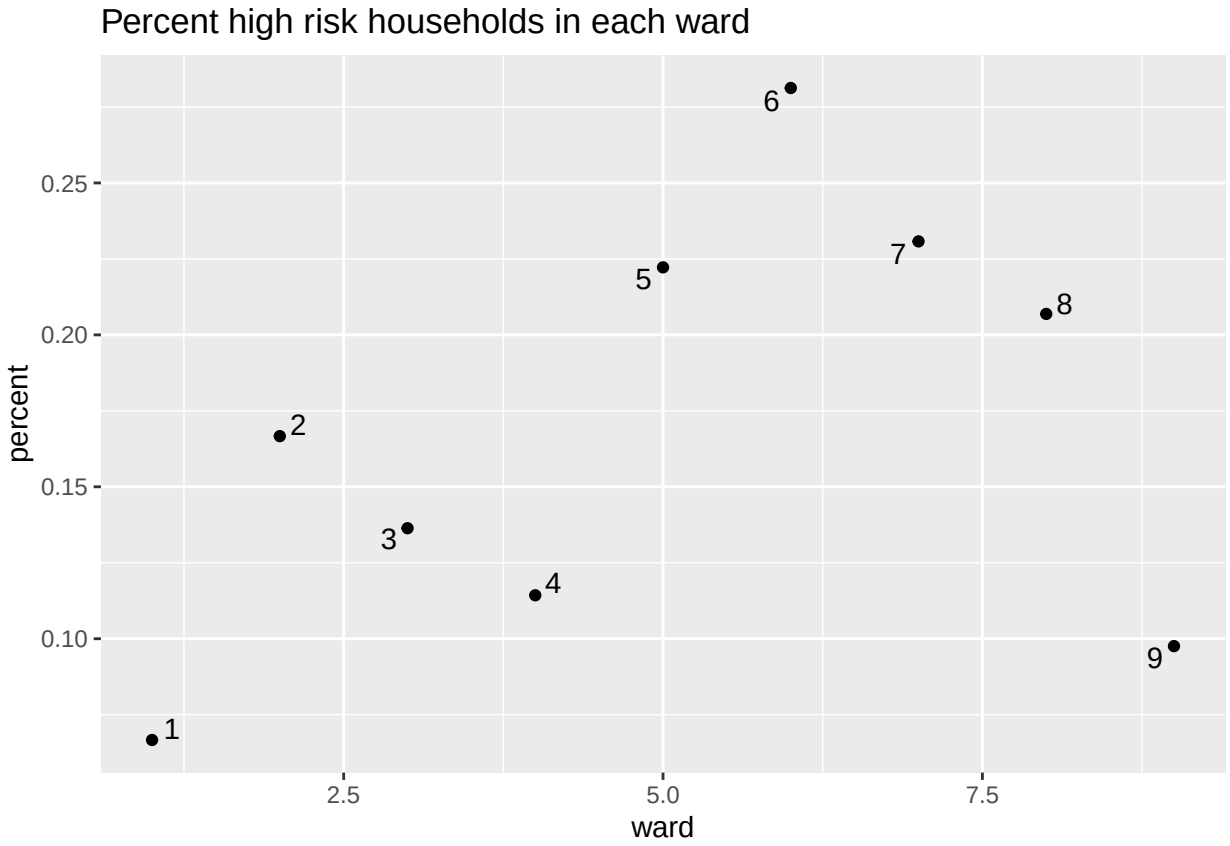
b. Create a data frame (or table) that answers the following question: Within each ward, what proportion of households are high risk (in terms of lead content in the initial water sample) and what households are not high risk based on the EPA guidelines?

```
high_risk <- flint_mi_water %>%
  group_by(ward) %>%
  summarise(percent = mean(as.numeric(factor(epa_lead_initial))-1))


high_risk
```

```
## # A tibble: 9 x 2
##    ward percent
##   <dbl>   <dbl>
## 1     1  0.0667
## 2     2  0.167
## 3     3  0.136
## 4     4  0.114
## 5     5  0.222
## 6     6  0.281
## 7     7  0.231
## 8     8  0.207
## 9     9  0.0976
```

c. Also create a graph that will help you examine how the proportion of high risk versus not high risk households varies by ward. Make sure to give the graph nice labels and a title.

```
bar_labels = as.character(high_risk$ward)

ggplot(high_risk, aes(ward, percent, label=ward)) +
  geom_text_repel() + geom_point() +
  labs(title = "Percent high risk households in each ward")
```




d. Which three wards have the highest proportion of high risk households (in terms of lead content in the initial water sample)?

---

- **Ranking**
  - *Ward 6:* 28.1
  - *Ward 7:* 23.1
  - *Ward 5:* 22.2

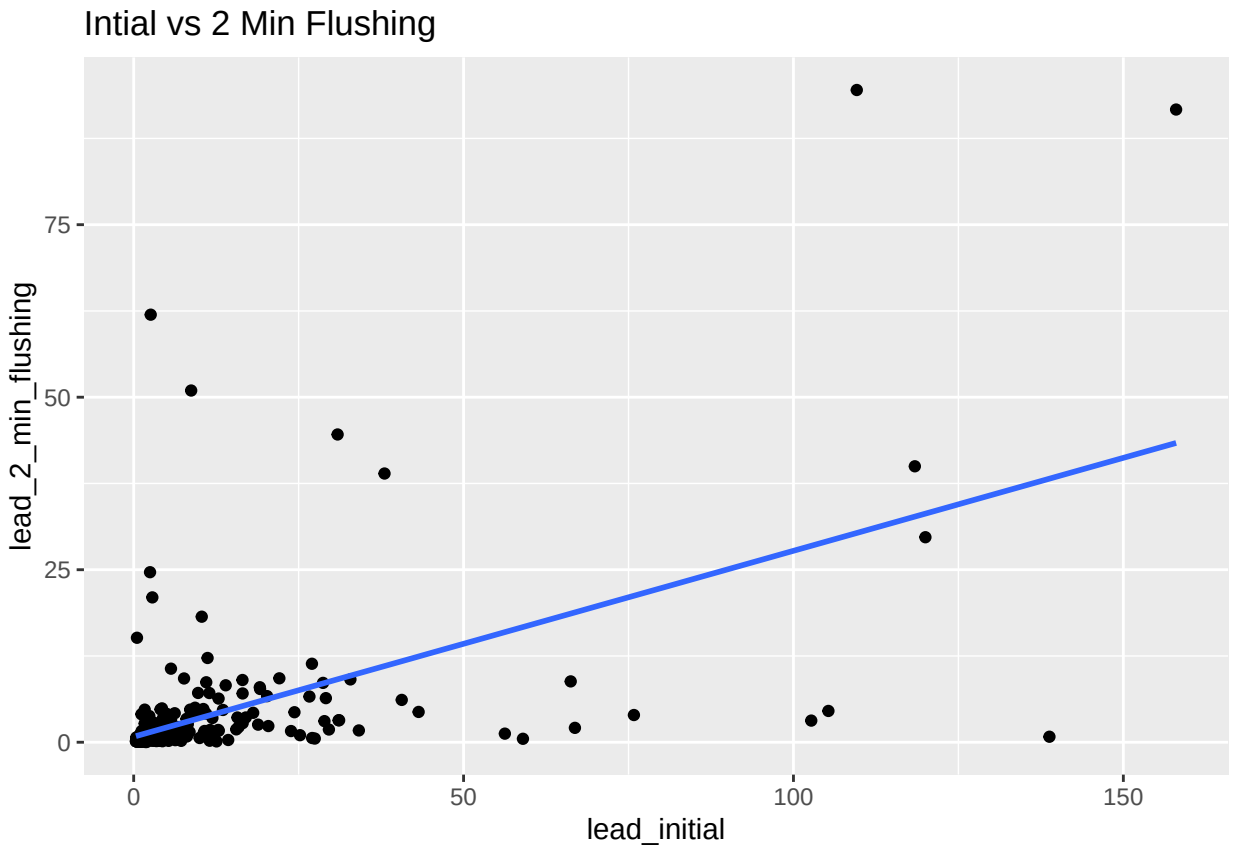
---

**Problem 3**

In this problem, we will build a model for the lead content after two minutes of flushing through the faucet tap.

a. Create a graph that compares the lead content after two minutes of flushing (`lead_2_min_flushing`) to the lead content in the initial water sample (`lead_initial`). Include the line of best fit on your graph.

```
ggplot(flint_mi_water, aes(lead_initial, lead_2_min_flushing)) +
  geom_point() + geom_smooth(method='lm', formula= y~x, se=F) +
  labs(title = "Intial vs 2 Min Flushing")
```




b. Identify and provide the `sample_id` value for a household that is an outlier and **is** an influential point. The slope of different regression lines must be included in the justification of your selection.

```
flint_mi_water$abs_diff_flush <- abs(flint_mi_water$lead_initial-flint_mi_water$lead_2_min_flushing)


# I looked up the max difference in the other view and found smaple_id 95
# had a difference of 138.003
no_house_95 <- flint_mi_water[flint_mi_water["sample_id"]!=95, ]

# With point
lm(lead_2_min_flushing ~ lead_initial, flint_mi_water)
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ lead_initial, data = flint_mi_water)
##
## Coefficients:
##  (Intercept)  lead_initial
##       0.7921        0.2695
```

```
#Without house 95
lm(lead_2_min_flushing ~ lead_initial, no_house_95)
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ lead_initial, data = no_house_95)
##
```

```
## Coefficients:
##  (Intercept)  lead_initial
##       0.4805        0.3136
```

---

- Major difference in both slope and intercept

---

c. Identify and provide the `sample_id` value for a household that is an outlier and is **not** an influential point. The slope of different regression lines must be included in the justification of your selection.

```
flint_mi_water$abs_diff_flush <- abs(flint_mi_water$lead_initial-flint_mi_water$lead_2_min_flushing)

# same method as above for finding the min house
# House 264
no_house_264 <- flint_mi_water[flint_mi_water["sample_id"]!=264, ]


# With point
print(lm(lead_2_min_flushing ~ lead_initial, flint_mi_water))
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ lead_initial, data = flint_mi_water)
##
## Coefficients:
##  (Intercept)  lead_initial
##       0.7921        0.2695
```

```
#Without house 95
lm(lead_2_min_flushing ~ lead_initial, no_house_264)
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ lead_initial, data = no_house_264)
##
## Coefficients:
##  (Intercept)  lead_initial
##       0.7940        0.2694
```

---

- Difference between the two intercepts is 0.0019 higher
- Difference between the two is an intercept 0.0001 flatter

---

d. Add two new variables to your data frame that represent the (natural) logged versions of `lead_initial` and `lead_2_min_flushing`.

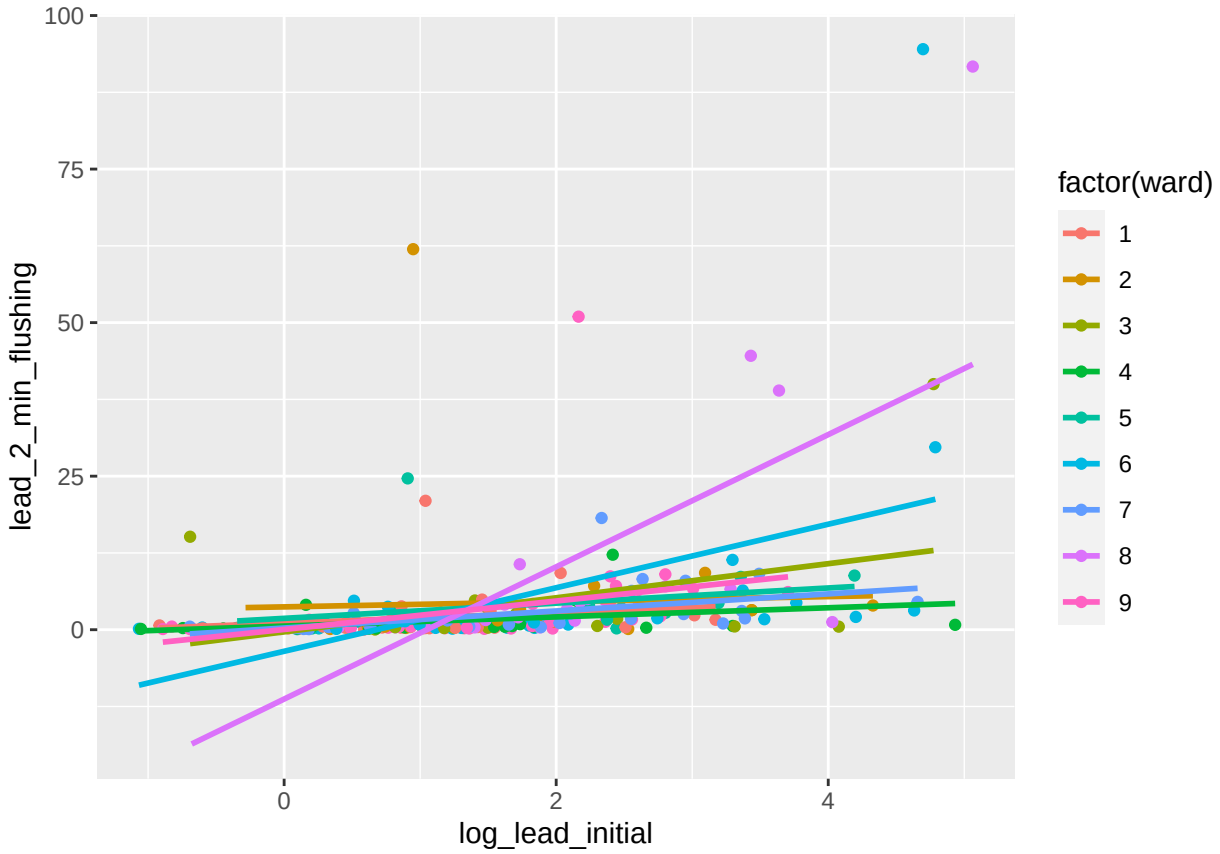
```
flint_mi_water$log_lead_initial <- log(flint_mi_water$lead_initial)
flint_mi_water$log_lead_2_min_flushing <- log(flint_mi_water$lead_2_min_flushing)
```

e. Build a linear regression model for the log of the lead content after two minutes of flushing using the log of the lead content of the initial sample and the ward. Allow the slopes in the model to vary. What is the estimated slope of the regression line for Ward 1? What is the estimated slope of the regression line for Ward 7?

Hints:

- Consider using `factor()` to ensure R properly handles the categorical variable.
- Use `print = TRUE` within `get_regression_table()` so that R prints the whole term name.

```
ggplot(flint_mi_water, aes(log_lead_initial, lead_2_min_flushing, color=factor(ward))) +
  geom_point() + geom_smooth(method='lm', se=F, formula= y~x)
```




```
ward_1 <-flint_mi_water[flint_mi_water["ward"]==1, ]
ward_7 <-flint_mi_water[flint_mi_water["ward"]==7, ]

#Ward 1
print(lm(lead_2_min_flushing~log_lead_initial, ward_1))
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ log_lead_initial, data = ward_1)
##
## Coefficients:
##      (Intercept)  log_lead_initial
##           1.0703            0.8513
```

```
#Ward 7
lm(lead_2_min_flushing~log_lead_initial, ward_7)
```

```
##
## Call:
## lm(formula = lead_2_min_flushing ~ log_lead_initial, data = ward_7)
##
## Coefficients:
##      (Intercept)  log_lead_initial
```

```
##             0.2777              1.3887
```

---

- What should I write?

---