# Data-412 HW6

Ashley Totten

April 17, 2025

### Cleaning Flint data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(janitor)
```

```
Attaching package: 'janitor'

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(dplyr)
library(stringr)
library(ggplot2)

Flint <- read.csv("/Users/tottena17/Downloads/AU 24-25/Data-412 R/Flint_Facilities_Testing_2(
```

```
clean_names(Flint)->Flint1
```

### Cleaning facility name

```
#checking for unique values in facility name
Flint1 %>%
  distinct(facility_name)
```

```
                    facility_name
1                 ALLEREE BILLINGS
2                     ANGIE MCNEAL
3                    BETTY JOE PEA
4                     BRIDON`S CDC
5    CATHEDRAL OF FAITH HEAD START
6     CUMMINGS/ GREAT EXPECTATIONS
7      CUMMINGS/GREAT EXPECTATIONS
8                      GAIL SEWELL
9         GENESEE COUNTY JOB CORPS
10          GLORIA`S LITTLE ANGELS
11  HEAVENLY ANGELS (LATISH SMITH)
12     HONEY BEE PALACE CHILD CARE
13                JANE ZITTERKOPH
14        JUST FOR KIDS GROUP HOME
15          KIDDIE TIME CHILD CARE
16              KINGDOM KAMPUS CDC
17                 LEALI ALEXANDER
18                    LEO ADAMS JR
19                       LORI HILL
20         LULLIBY (JANICE MOBLEY)
21                   MANLEY SCHOOL
22                   MONICA WALKER
23                MONIQUE HENDRIX
24   MOTT EARLY CHILDHOOD LEARNING
25                     REACH DHHS
26            SAND CASTLE DAY CARE
27                     SUNNY PATCH
28      TEDDY BEARS/ PATRICE MOORE
29       TEDDY BEARS/PATRICE MOORE
30        WHALEY CHILDREN`S CENTER
31 WHALEY CHILDREN`S CENTER (MOTT)
```

```
32         WHALEY CHILDRENS CENTER
33     CUMMING/ GREAT EXPECTATIONS
```

```
Flint1 %>% #replacing facility names that were duplicates
  mutate(facility_name = str_replace(facility_name, "CUMMINGS/ GREAT EXPECTATIONS", "CUMMINGS
  mutate(facility_name = str_replace(facility_name, "CUMMING/ GREAT EXPECTATIONS", "CUMMINGS/
  mutate(facility_name = str_replace(facility_name, "WHALEY CHILDRENS CENTER", "WHALEY CHILDR
  mutate(facility_name = str_replace(facility_name, "WHALEY CHILDREN`S CENTER (MOTT)", "WHALE
  mutate(facility_name = str_replace(facility_name, "TEDDY BEARS/ PATRICE MOORE", "TEDDY BEAR
Flint2 %>%
  distinct(facility_name)
```

```
                    facility_name
1                 ALLEREE BILLINGS
2                     ANGIE MCNEAL
3                    BETTY JOE PEA
4                     BRIDON`S CDC
5    CATHEDRAL OF FAITH HEAD START
6      CUMMINGS/GREAT EXPECTATIONS
7                      GAIL SEWELL
8         GENESEE COUNTY JOB CORPS
9           GLORIA`S LITTLE ANGELS
10  HEAVENLY ANGELS (LATISH SMITH)
11    HONEY BEE PALACE CHILD CARE
12                 JANE ZITTERKOPH
13        JUST FOR KIDS GROUP HOME
14          KIDDIE TIME CHILD CARE
15              KINGDOM KAMPUS CDC
16                 LEALI ALEXANDER
17                    LEO ADAMS JR
18                       LORI HILL
19         LULLIBY (JANICE MOBLEY)
20                   MANLEY SCHOOL
21                   MONICA WALKER
22                 MONIQUE HENDRIX
23   MOTT EARLY CHILDHOOD LEARNING
24                      REACH DHHS
25            SAND CASTLE DAY CARE
26                     SUNNY PATCH
27       TEDDY BEARS/PATRICE MOORE
28        WHALEY CHILDREN`S CENTER
29 WHALEY CHILDREN`S CENTER (MOTT)
```

### Pivoting the data frame

```
Flint2 %>%  #removing unnecessary columns
  mutate(city = NULL) %>%
  #pivoting the data set so Lead and Copper have their own columns
  pivot_wider(
    names_from = "lead",
    values_from = "result_ppb_3"
  ) %>%
  pivot_wider(
    names_from = "copper",
    values_from = "result_ppb_6"
  ) %>% #I want to simplify the name of the column instead of just cleaning it
  rename(lead = `Lead 250 mL Sample`) %>%
  rename(copper = `Copper 250 mL Sample`)->Flint3


Flint3
```

```
# A tibble: 264 x 8
   sample_number sample_description   subdate street_name zip_code facility_name
   <chr>         <chr>                <chr>   <chr>          <int> <chr>
 1 LH59064       01KC003 KITCHEN      2017-1~ DAMON ST       48505 ALLEREE BILL~
 2 LH55198       01KC001 KITCHEN      2017-1~ BALDWIN BL~    48505 ANGIE MCNEAL
 3 LH55197       01BF002 BATHROOM     2017-1~ BALDWIN BL~    48505 ANGIE MCNEAL
 4 LH58462       001KC004 KITCHEN     2017-1~ CLEMENT ST     48504 BETTY JOE PEA
 5 LH58463       001BF002 1ST FLOOR ~ 2017-1~ CLEMENT ST     48504 BETTY JOE PEA
 6 LH58464       001BF001 1ST FLOOR ~ 2017-1~ CLEMENT ST     48504 BETTY JOE PEA
 7 LH55194       LLBF001 BOY`S RESTR~ 2017-1~ KEARSLEY       48503 BRIDON`S CDC
 8 LH55196       LLKC004 DAY CARE KI~ 2017-1~ EAST KEARS~    48503 BRIDON`S CDC
 9 LH55193       LLBF002 GIRLS RESTR~ 2017-1~ EAST KEARS~    48503 BRIDON`S CDC
10 LH55195       LLCF005 DAY CARE CL~ 2017-1~ EAST KEARS~    48503 BRIDON`S CDC
# i 254 more rows
# i 2 more variables: lead <int>, copper <int>
```

### Creating data frame for 2

```
Flint3 %>%
  count(facility_name, name = "samples") -> Flint4
```

```r
Flint3 %>%
  group_by(facility_name) %>%
  count(lead > 15, name = "exceeds_lead") %>%
  filter(`lead > 15` == TRUE)-> Flint5

Flint3 %>%
  group_by(facility_name) %>%
  count(copper > 1300, name = "exceeds_copper") %>%
  filter(`copper > 1300` == TRUE)-> Flint6

Flint3 %>%
  group_by(facility_name) %>%
  summarize(max_lead = max(lead), min_lead = min(lead), max_copper = max(copper), min_copper

Flint_merge <- full_join(Flint4, Flint5, by = "facility_name")
Flint_merge1 <- full_join(Flint_merge, Flint6, by = "facility_name")
Flint_merge2 <- full_join(Flint_merge1, Flint7, by = "facility_name")

Flint_merge2 %>%
  mutate(`lead > 15` = NULL, `copper > 1300` = NULL) %>%
  mutate(across(everything(), ~replace(., is.na(.), 0)))->Flint_data

Flint_data
```

```
# A tibble: 29 x 8
   facility_name           samples exceeds_lead exceeds_copper max_lead min_lead
   <chr>                     <dbl>        <dbl>          <dbl>    <dbl>    <dbl>
 1 ALLEREE BILLINGS              3            0              0        5        0
 2 ANGIE MCNEAL                  2            0              0        0        0
 3 BETTY JOE PEA                 4            0              0        2        0
 4 BRIDON`S CDC                  5            0              0        0        0
 5 CATHEDRAL OF FAITH HEA~       9            0              0        3        0
 6 CUMMINGS/GREAT EXPECTA~      17            0              0        2        0
 7 GAIL SEWELL                   4            0              0        7        0
 8 GENESEE COUNTY JOB COR~      17            0              0        2        0
 9 GLORIA`S LITTLE ANGELS        9            0              0        1        0
10 HEAVENLY ANGELS (LATIS~       3            0              0        2        0
# i 19 more rows
# i 2 more variables: max_copper <dbl>, min_copper <dbl>
```

### Filter variable for 3

```
Flint3 %>%
  filter(facility_name == "HONEY BEE PALACE CHILD CARE" | (facility_name == "MONIQUE HENDRIX'

filter1 %>%
  mutate(filter = "filtered")->filter2

filter3 <- right_join(filter2, Flint3)
```

```
Joining with `by = join_by(sample_number, sample_description, subdate,
street_name, zip_code, facility_name, lead, copper)`
```

```
filter3 %>%
  mutate(across(everything(), ~replace(., is.na(.), "unfiltered"))) %>%
  mutate(filter = as.factor(filter)) %>%
  mutate(lead = as.numeric(lead)) %>%
  mutate(copper = as.numeric(copper))-> filter_data

head(filter_data)
```

```
# A tibble: 6 x 9
  sample_number sample_description    subdate street_name zip_code facility_name
  <chr>         <chr>                 <chr>   <chr>       <chr>    <chr>
1 LH53585       WC005 OUTSIDE WOMEN`~ 2017-1~ DUPONT STR~ 48505    CATHEDRAL OF~
2 LH52343       01WC003 OFF GYM       2017-1~ WALTON      48532    CUMMINGS/GRE~
3 LH52334       01DW002 GYM           2017-1~ WALTON      48532    CUMMINGS/GRE~
4 LH52340       01WC004 OFF GYM       2017-1~ WALTON      48532    CUMMINGS/GRE~
5 LH54764       01DW016               2017-1~ N SAGINAW   48505    GENESEE COUN~
6 LH54769       01DW003               2017-1~ N SAGINAW   48505    GENESEE COUN~
# i 3 more variables: lead <dbl>, copper <dbl>, filter <fct>
```

### Location variable for 4

```
Flint3 %>%
  mutate(location = if_else(str_detect(sample_description, "KC"),"kitchen", false = "other"))
  mutate(location = if_else(str_detect(sample_description, "BF"),"restroom", location)) %>%
  mutate(location = if_else(str_detect(sample_description, "NS"),"nurses station", location))
```

```
  mutate(location = if_else(str_detect(sample_description, "WC"),"water cooler", location)) %
  mutate(location = if_else(str_detect(sample_description, "DW"),"bubbler", location)) %>%
  mutate(location  = as.factor(location))->location1

levels(location1$location)
```

```
[1] "bubbler"        "kitchen"        "nurses station" "other"
[5] "restroom"       "water cooler"
```

```
head(location1)
```

```
# A tibble: 6 x 9
  sample_number sample_description    subdate street_name zip_code facility_name
  <chr>         <chr>                 <chr>   <chr>          <int> <chr>
1 LH59064       01KC003 KITCHEN       2017-1~ DAMON ST       48505 ALLEREE BILL~
2 LH55198       01KC001 KITCHEN       2017-1~ BALDWIN BL~    48505 ANGIE MCNEAL
3 LH55197       01BF002 BATHROOM      2017-1~ BALDWIN BL~    48505 ANGIE MCNEAL
4 LH58462       001KC004 KITCHEN      2017-1~ CLEMENT ST     48504 BETTY JOE PEA
5 LH58463       001BF002 1ST FLOOR B~ 2017-1~ CLEMENT ST     48504 BETTY JOE PEA
6 LH58464       001BF001 1ST FLOOR B~ 2017-1~ CLEMENT ST     48504 BETTY JOE PEA
# i 3 more variables: lead <int>, copper <int>, location <fct>
```

Some of the entries with a sample description of other include descriptions for labelled locations. For example, there are some sample descriptions that include the word "kitchen" but don't include the KC indicator that the description listed as being associated with a kitchen sink.

### Time and Date for 5

```
Flint3%>%
  mutate(subdate = str_replace_all(Flint3$subdate, "T", " ")) %>%
  mutate(subdate = str_remove_all(subdate, "Z"))->time1

time1 %>%
  mutate(subdate = str_remove_all(time1$subdate, "....-..-.. "))->time2

time2 %>%
  mutate(subdate = parse_time(time2$subdate, "%H:%M:%S"))->time3
```

```r
time1 %>%
  mutate(subdate = str_remove_all(time1$subdate, "..:..:.."))->time4

time4 %>%
  mutate(subdate = parse_date(time4$subdate, "%Y-%m-%d"))->time5

head(time3)
```

```
# A tibble: 6 x 8
  sample_number sample_description   subdate  street_name zip_code facility_name
  <chr>         <chr>                <time>   <chr>          <int> <chr>
1 LH59064       01KC003 KITCHEN      14:05:00 DAMON ST       48505 ALLEREE BILL~
2 LH55198       01KC001 KITCHEN      11:11:07 BALDWIN BL~    48505 ANGIE MCNEAL
3 LH55197       01BF002 BATHROOM     11:11:06 BALDWIN BL~    48505 ANGIE MCNEAL
4 LH58462       001KC004 KITCHEN     14:05:42 CLEMENT ST     48504 BETTY JOE PEA
5 LH58463       001BF002 1ST FLOOR ~ 14:05:43 CLEMENT ST     48504 BETTY JOE PEA
6 LH58464       001BF001 1ST FLOOR ~ 14:05:44 CLEMENT ST     48504 BETTY JOE PEA
# i 2 more variables: lead <int>, copper <int>
```

```r
head(time5)
```

```
# A tibble: 6 x 8
  sample_number sample_description subdate    street_name zip_code facility_name
  <chr>         <chr>              <date>     <chr>          <int> <chr>
1 LH59064       01KC003 KITCHEN    2017-12-21 DAMON ST       48505 ALLEREE BILL~
2 LH55198       01KC001 KITCHEN    2017-11-29 BALDWIN BL~    48505 ANGIE MCNEAL
3 LH55197       01BF002 BATHROOM   2017-11-29 BALDWIN BL~    48505 ANGIE MCNEAL
4 LH58462       001KC004 KITCHEN   2017-12-19 CLEMENT ST     48504 BETTY JOE PEA
5 LH58463       001BF002 1ST FLOO~ 2017-12-19 CLEMENT ST     48504 BETTY JOE PEA
6 LH58464       001BF001 1ST FLOO~ 2017-12-19 CLEMENT ST     48504 BETTY JOE PEA
# i 2 more variables: lead <int>, copper <int>
```

### Exploratory Analysis

```r
filter_data %>%
  summarize(mean = mean(lead), sd = sd(lead), median = median(lead), IQR = IQR(lead), .by = f
```

```
# A tibble: 2 x 5
  filter      mean    sd median   IQR
```

```
  <fct>      <dbl> <dbl>  <dbl> <dbl>
1 filtered   10.2   50.1      0  1
2 unfiltered  2.08  11.6      0  0.75
```

```
filter_data %>%
  count(filter)
```

```
# A tibble: 2 x 2
  filter         n
  <fct>      <int>
1 filtered      66
2 unfiltered   198
```

```
filter_data %>%
  filter(lead > 15, .by = filter)
```

```
# A tibble: 6 x 9
  sample_number sample_description    subdate street_name zip_code facility_name
  <chr>         <chr>                 <chr>   <chr>       <chr>    <chr>
1 LH54656       01WC013 OUTSIDE ROOM~ 2017-1~ E COURT ST  48503    MOTT EARLY C~
2 LH57760       01DW003               2017-1~ DONALDSON ~ 48504    SUNNY PATCH
3 LH57757       01DW001 NEXT TO REST~ 2017-1~ DONALDSON ~ 48504    SUNNY PATCH
4 LH53780       KC040A ROOM 105       2017-1~ FARLEY STR~ 48507    MANLEY SCHOOL
5 LH53777       KC038A ROOM 106       2017-1~ FARLEY STR~ 48507    MANLEY SCHOOL
6 LH56740       01KC003               2017-1~ ODETTE      48503    JUST FOR KID~
# i 3 more variables: lead <dbl>, copper <dbl>, filter <fct>
```

The mean of lead levels of samples from filtered fixtures is greater than the mean lead levels of samples from unfiltered fixtures, samples filtered fixtures also have a greater standard deviation than that of unfiltered fixtures. However, of there are six entries of samples that have concerning levels of lead, three are from filtered samples and three are from unfiltered samples. I would be curious to see if there is a statistically significant difference in the levels of lead from samples from filtered fixtures and samples from unfiltered fixtures. I would hypothesize that there would be a statistically significant difference in the mean sample of lead, but there is not enough evidence to say whether there is a statistically significant difference in samples from fixtures that are above 15ppb, or the level which indicates concern.