DS5110 Homework 1 - Solutions

Kylie Ariel Bemis

1/14/2018

Instructions

Create a directory with the following structure:

- `hw1-your-name/hw1-your-name.Rmd`
- `hw1-your-name/hw1-your-name.pdf`

where `hw1-your-name.Rmd` is an R Markdown file that compiles to create `hw1-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder "hw1", go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw1 solutions] - your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Your code and answers may differ and still be correct.

Make sure you have following packages installed

```
install.packages("tidyverse")
install.packages("nycflights13")
```

```
library(tidyverse)
```

Part A

Problems 1–3 use the `flights` dataset from the `nycflights13` package, which includes data for all flights that departed New York City (JFK, LGA, or EWR) in 2013.

```
library(nycflights13)
flights
```

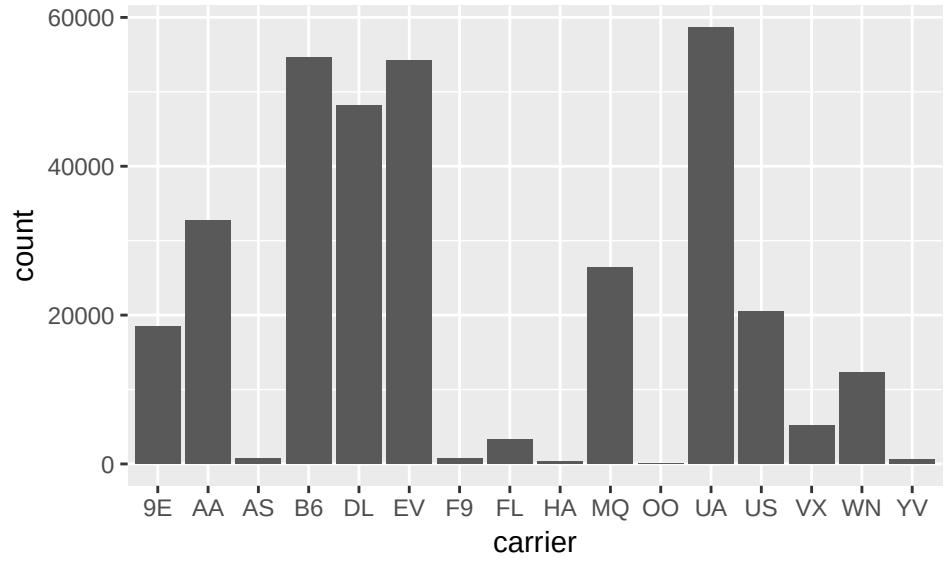
```
## # A tibble: 336,776 x 19
##     year month   day dep_time sched_dep_time dep_delay arr_time
##    <int> <int> <int>    <int>          <int>     <dbl>    <int>
##  1  2013     1     1      517            515         2      830
##  2  2013     1     1      533            529         4      850
##  3  2013     1     1      542            540         2      923
##  4  2013     1     1      544            545        -1     1004
##  5  2013     1     1      554            600        -6      812
##  6  2013     1     1      554            558        -4      740
##  7  2013     1     1      555            600        -5      913
##  8  2013     1     1      557            600        -3      709
##  9  2013     1     1      557            600        -3      838
```

```
## 10  2013     1     1      558            600        -2      753
## # ... with 336,766 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dttm>
```

Problem 1

> Create a bar plot showing the number of flights flown out of New York airports by each carrier in 2013. Which airline carrier flew the most flights?

```
ggplot(data=flights, mapping=aes(x=carrier)) + geom_bar()
```



United Airlines (UA) flew the most flights out of New York airports in 2013.

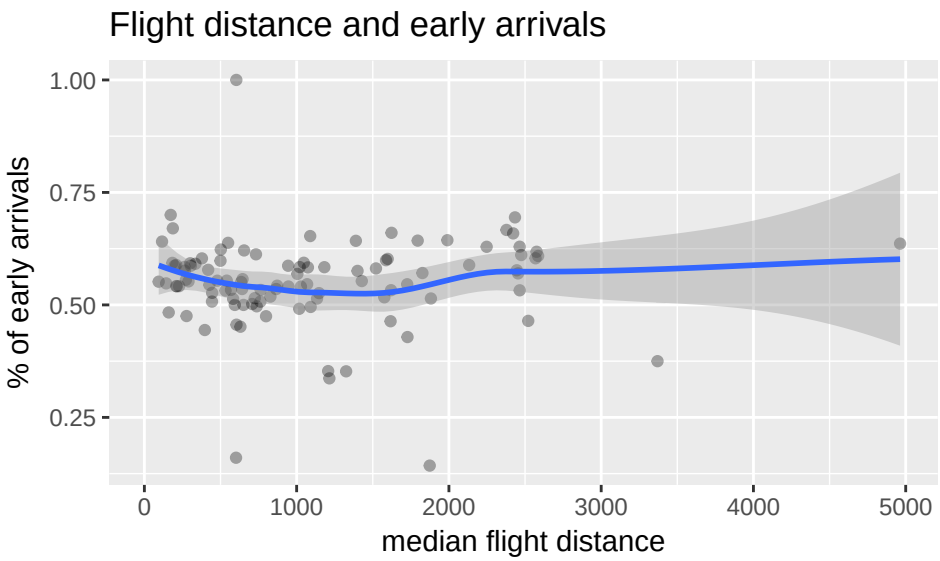
Tip: you can use `filter(airlines, carrier == 'UA')` to find the fullname of an airline.

Problem 2

> For each destination, calculate the proportion of flights that arrived at their destination earlier than scheduled. Also calculate the median distance flown to each destination.
>
> Plot the proportion of early arrivals (on the y-axis) against the median distance flown (on the x-axis) for each destination. Add a smooth line to the plot. Based on the smooth line, at what distances are flights most likely to arrive early? Describe the relationship between early arrivals and flight distance.

```
early_dest <- flights %>%
  group_by(dest) %>%
  summarise(dist=median(distance, na.rm=TRUE),
            prop_early=mean(arr_delay < 0, na.rm=TRUE))

ggplot(early_dest, mapping=aes(x=dist, y=prop_early)) +
  geom_point(alpha=1/3) +
```

```
  geom_smooth() +
  labs(
    title='Flight distance and early arrivals',
    x='median flight distance',
    y='% of early arrivals',
    size='# of flights')
```

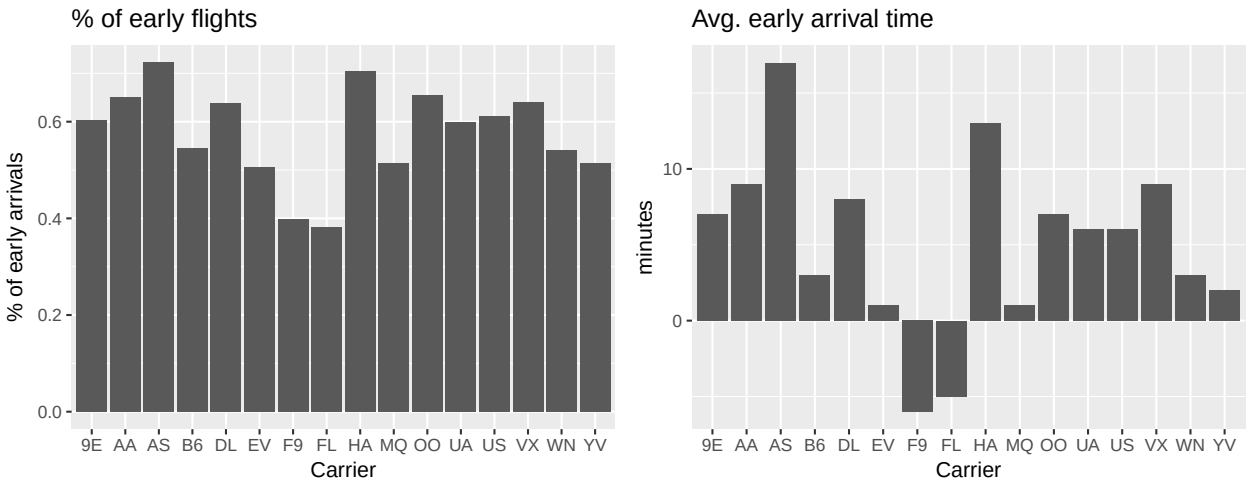


- Up to about ~1000 miles, the proportion of flights arriving early appears to decrease with distance, and then the relationship flattens out.
- However, after about ~1500 miles, increasing distance generally means more flights are arriving early. This may be due to being able to make up more time in the air during much longer flights.

Problem 3

> Create two bar plots that characterize each carrier by how early their flights arrive. One should show the proportion of flights that arrive early for each carrier, and the other should show the median number of minutes early that flights arrivee for each carrier.
>
> Which airlines are the most consistently ahead of schedule? Which airlines arrive the most early?
>
> Which airlines are most consistently behind schedule? Which airlines arrive the latest?

```
early_carrier <- flights %>%
  group_by(carrier) %>%
  summarise(dist=median(distance, na.rm=TRUE),
            prop_early=mean(arr_delay < 0, na.rm=TRUE),
            avg_early=median(-arr_delay, na.rm=TRUE))

p1 <- ggplot(early_carrier, aes(x=carrier, y=prop_early)) +
  geom_col() +
  labs(title='% of early flights', x='Carrier', y='% of early arrivals')

p2 <- ggplot(early_carrier, aes(x=carrier, y=avg_early)) +
  geom_col() +
  labs(title='Avg. early arrival time', x='Carrier', y='minutes')
```



Hawaiian Airlines (HA) and *Alaska Airlines* (AS) are the carriers that are most consistently ahead of schedule (most flights arriving early) and arrive the earliest (>10 min ahead of schedule).

Frontier Airlines (F9) and *AirTran Airways* (FL) are the carriers that are most consistently behind schedule (most flights arriving late) and arrive the latest (delayed on average).

Part B

Problems 4–6 use data from the Navajo Nation Water Quality Project. Download the CSV file from http://navajowater.org/export-raw-data/.

Water quality is a major issue on American Indian reservations in the southwestern United States. The prevalence of uranium mines and uranium mill accidents mean that much of the water in the Navajo Nation is irradiated, and many homes are left without clean, drinkable water. Multiple environmental agencies routinely sample water in the region and report on contaminants.

Read the documentation for the `tidyverse` function `read_csv`, and use it to import the dataset into R.

```
water_raw <- read_csv("../../../data/NavajoWaterExport.csv")
library(measurements)

water <- transmute(water_raw,
       section=`Which EPA Section is This From?`,
       name=`Name of Water Source`,
       date=`Date of Water Sampling`,
       long=Longitude,
       lat=Latitude,
       risk=`US EPA Risk Rating`,
       alpha=`Amount of Alpha Particles`,
       radium226=`Amount of Radium226`,
       radium228=`Amount of Radium228`,
       uranium234=`Amount of Uranium234`,
       uranium235=`Amount of Uranium235`,
       uranium238=`Amount of Uranium238`)
water
```

```
## # A tibble: 225 x 12
##      section                          name     date          long
##        <chr>                         <chr>    <chr>         <chr>
##  1 Section 3                   Gold Spring  1/19/00 111 4 28.4861
##  2 Section 3                   Tank 3K-331  7/27/98 111 24 24.948
##  3 Section 6 Lower Greasewood Chapter House  4/14/99 109 51 14.651
##  4 Section 7                   Tank 8T-549  10/9/98 110 12 49.674
##  5 Section 6                  Cedar Spring  7/13/98 110 21 54.736
##  6 Section 7                    Tank 8AI-1  9/21/98 110 18 34.589
##  7 Section 6                 Coyote Spring   7/8/98 110 27 58.443
##  8 Section 2                         9T-523  3/18/99 109 10 51.933
##  9 Section 6           Chimney Butte Spring  7/14/98 110 25 20.890
## 10 Section 5          Nazlini Chapter House 11/17/98 109 26 41.176
## # ... with 215 more rows, and 8 more variables: lat <chr>, risk <chr>,
## #   alpha <dbl>, radium226 <dbl>, radium228 <dbl>, uranium234 <dbl>,
## #   uranium235 <dbl>, uranium238 <dbl>
```

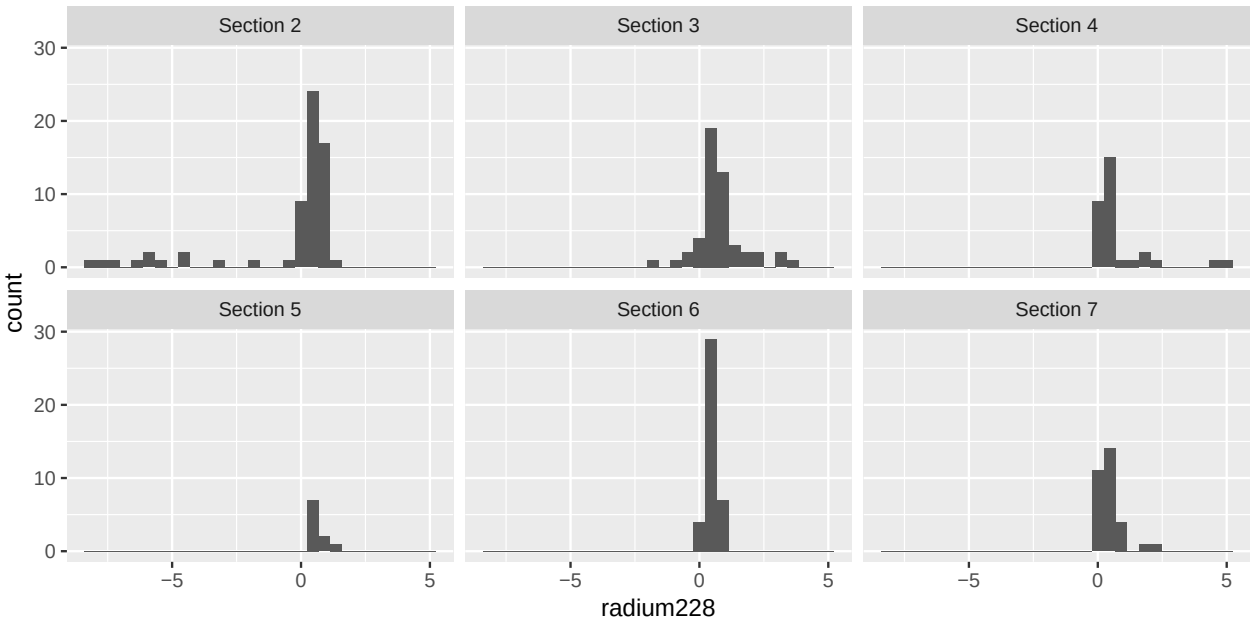
Problem 4

Create histograms showing the distribution of the amount of Radium-228 in water samples for each EPA section (use faceting). Do you notice anything odd? (Besides the fact that the water samples are radioactive in the first place?)

The concentration of radiactive elements in a sample is measured in rate of atomic disintegrations per volume, rather than mass per volume, as used for stable isotopes. This is done by counting the number of atomic disintegrations per minute and comparing it to the mass of the material involved. However, laboratory environments and instruments used for detection create some number of atomic emissions on their own, so background correction must be performed. Because this process involves sampling many times, and the background can be inconsistent, resulting in over-correction, sometimes negative values are reported for the concentration. For practical purposes, these values can be considered zero.

Mutate the dataset to replace the negative values with 0, and then create the histograms again, using a different combination of `ggplot2` functions this time.

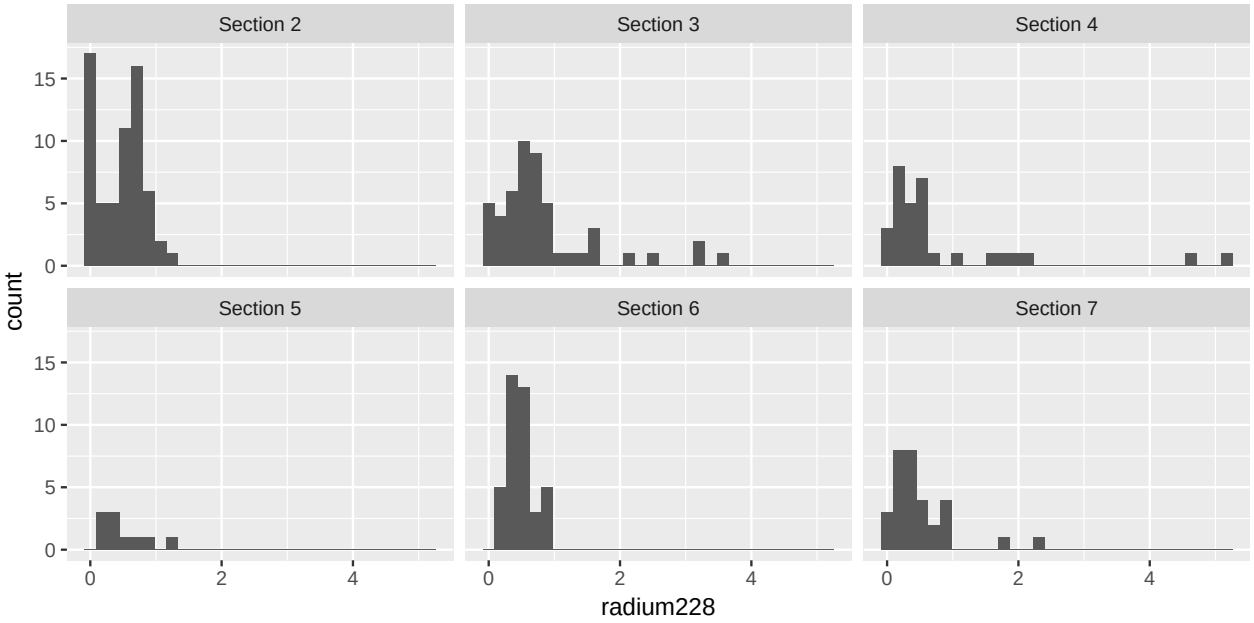
```
ggplot(water, mapping=aes(x=radium228)) +
  geom_histogram() +
  facet_wrap(~section)
```



When first plotting the amount of radium-228 for each EPA section, we notice that many of the sections have negative values, which does not make sense for concentration.

We would then make another plot after replacing the negative values with 0. This time instead of `geom_histogram()`, we use `geom_bar()` with `stat="bin"`.

```
water2 <- mutate(water, radium228=ifelse(radium228 < 0, 0, radium228))

ggplot(water2, mapping=aes(x=radium228)) +
  geom_bar(stat="bin") +
  facet_wrap(~section)
```



Problem 5

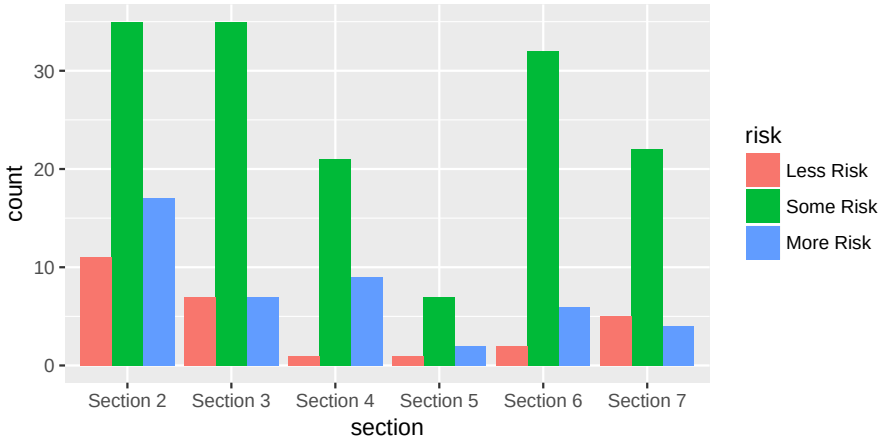
> Filter the dataset to remove any sites with "Unknown Risk" for the EPA risk rating.
>
> Count the number of sites of each EPA risk rating in each EPA section, and then calculate the mean concentration of Uranium-238 in the water samples for each EPA risk rating in each EPA section.
>
> Plot the number of sites at each EPA section using a bar plot, using the fill color of the bars to indicate the risk rating, and then plot the mean concentrations of Uranium-238 for each EPA section using a bar plot, using the fill color of the bars to indicate the risk rating.
>
> Which EPA section(s) have the most sites with "More Risk"? Which EPA section(s) have the sites with the highest concentration of Uranium-238 on average?

```
library(forcats)

water3 <- water2 %>%
  filter(risk != "Unknown Risk") %>%
  mutate(risk=fct_relevel(risk, "Less Risk", "Some Risk", "More Risk"))

water_u <- water3 %>%
  group_by(risk, section) %>%
  summarise(count=n(),
            u238=mean(uranium238))
```

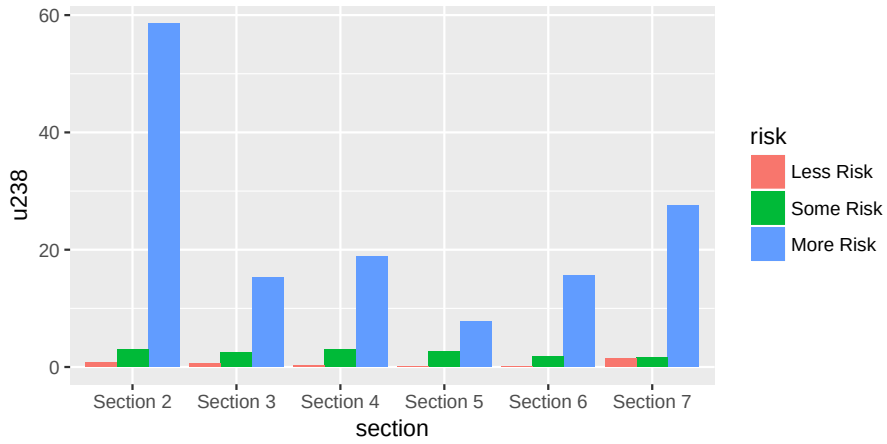
First, we filter the data to remove the sites with "Unknown Risk". Additionally, we use `forcats::fct_relevel` to reordered the levels of the `risk` factor, so that the graph will display the risk categories in a more intuitive order.

Then we summarize the data grouped on both EPA risk and EPA section, getting the counts and the mean of each uranium isotope for each group.

```
ggplot(water_u, aes(x=section, y=count, fill=risk)) +
  geom_col(position="dodge")
```




```
ggplot(water_u, aes(x=section, y=u238, fill=risk)) +
  geom_col(position="dodge")
```



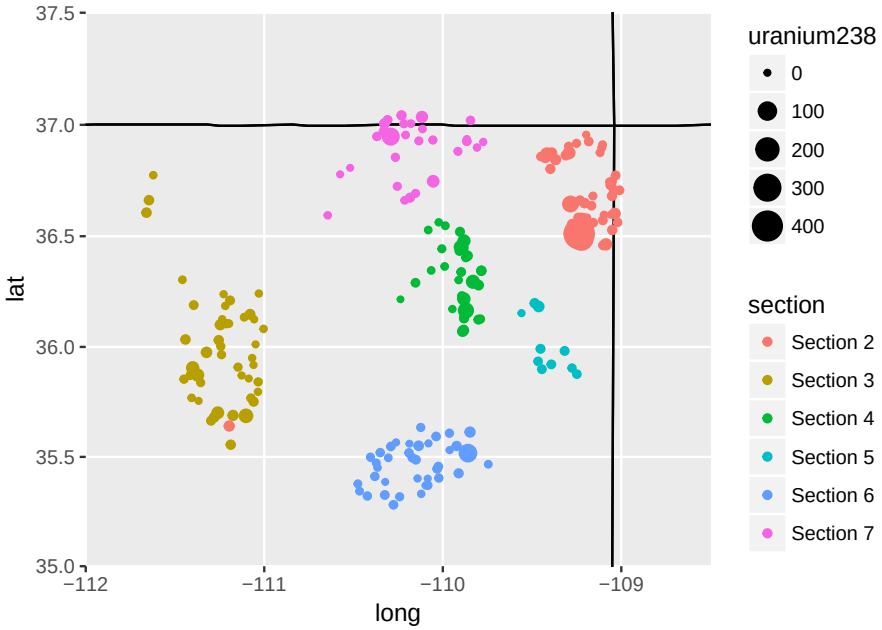
Here, we plot the site counts and the mean concentrations for Uranium-238 using `geom_col` (which is a shortcut for `geom="bar"` with `stat="identity"`), using the fill color aesthetic for the risk rating. We use `position="dodge"` so we can easily compare the heights of the bars between sites and risk ratings.

From the plot of the site counts, we see that Sections 2 and 3 have the most sites rated "More Risk", and Section 2 has the sites with the highest concentration of Uranium-238 on average.

Problem 6

> Install the `maps` package (you do not need to load it) and use the `ggplot2::map_data` function to get data for drawing the "Four Corners" region of the United States (i.e., Arizona, New Mexico, Utah, and Colorado).
>
> Install the `measurements` package and use the `measurements::conv_unit` function to convert the latitude and longitude information in the dataset to decimal degrees suitable to be used for plotting.
>
> Plot a map of the region (you may want to adjust the plotting limits to an appropriate "zoom" level), and overlay the locations of the water sampling sites on the map. Use color to indicate the EPA Section and size to indicate the amount of Uranium-238 measured at each site.

```
water_map <- water3 %>%
  mutate(long=-as.numeric(conv_unit(long, from="deg_min_sec", to="dec_deg")),
         lat=as.numeric(conv_unit(lat, from="deg_min_sec", to="dec_deg")))

four_corners <- map_data("state",
                  region=c("arizona", "new mexico", "utah", "colorado"))

ggplot(water_map) +
  geom_polygon(mapping=aes(x=long, y=lat, group=group),
               data=four_corners,
               fill=NA, color="black") +
  geom_point(mapping=aes(x=long, y=lat,
                         color=section,
                         size=uranium238)) +
  coord_map(xlim=c(-112, -108.5), ylim=c(35,37.5))
```



We use `geom_polygon` with its own data and aesthetics to plot the map, and `geom_point` to plot the EPA sites. We use color to indicate the section, and size of the points to indicate the amount of Uranium-238 at the site.

We notice that one site in Section 3 has been mislabeled as Section 2 in the data. However, there does not appear to be a very large concentration of uranium at the site, it is unlikely to have had a large influence in our earlier analysis.

Part C

> Problems 7–10 use data from the US Department of Education's Civil Rights Data Collection. Download the zipped 2013-2014 data from https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2013-14.html. Download the Public Use Data File User's Manual at the same location.
>
> Read the documentation for the `tidyverse` function `read_csv`, and use it to import the dataset into R. Check the User's Manual for how missing values were reported, and handle them appropriately.

```
crdc <- read_csv("../../../data/crdc201314csv/CRDC2013_14_SCH.csv",
                 na=c("-9", "-5", "-2"))
crdc
```

```
## # A tibble: 95,507 x 1,929
##    LEA_STATE               LEA_NAME                        SCH_NAME
##        <chr>                  <chr>                           <chr>
##  1        AL ALABAMA YOUTH SERVICES WALLACE SCH - MT MEIGS CAMPUS
##  2        AL ALABAMA YOUTH SERVICES     MCNEEL SCH - VACCA CAMPUS
##  3        AL           ALA YOUTH SER                AUTAUGA CAMPUS
##  4        AL       ALBERTVILLE CITY         ALA AVENUE MIDDLE SCH
##  5        AL       ALBERTVILLE CITY           ALBERTVILLE HIGH SCH
##  6        AL       ALBERTVILLE CITY                EVANS ELEM SCH
##  7        AL       ALBERTVILLE CITY           ALBERTVILLE ELEM SCH
##  8        AL       ALBERTVILLE CITY   BIG SPRING LAKE KINDERG SCH
##  9        AL       ALBERTVILLE CITY       ALBERTVILLE PRIMARY SCH
```

```
## 10        AL        MARSHALL COUNTY  KATE DUNCAN SMITH DAR MIDDLE
## # ... with 95,497 more rows, and 1926 more variables: COMBOKEY <chr>,
## #   LEAID <chr>, SCHID <chr>, JJ <chr>, CCD_LATCOD <dbl>,
## #   CCD_LONCOD <dbl>, NCES_SCHOOL_ID <chr>, MATCH_FLAG <chr>,
## #   SCH_GRADE_PS <chr>, SCH_GRADE_KG <chr>, SCH_GRADE_G01 <chr>,
## #   SCH_GRADE_G02 <chr>, SCH_GRADE_G03 <chr>, SCH_GRADE_G04 <chr>,
## #   SCH_GRADE_G05 <chr>, SCH_GRADE_G06 <chr>, SCH_GRADE_G07 <chr>,
## #   SCH_GRADE_G08 <chr>, SCH_GRADE_G09 <chr>, SCH_GRADE_G10 <chr>,
## #   SCH_GRADE_G11 <chr>, SCH_GRADE_G12 <chr>, SCH_GRADE_UG <chr>,
## #   SCH_UGDETAIL_ES <chr>, SCH_UGDETAIL_MS <chr>, SCH_UGDETAIL_HS <chr>,
## #   SCH_STATUS_SPED <chr>, SCH_STATUS_MAGNET <chr>,
## #   SCH_STATUS_CHARTER <chr>, SCH_STATUS_ALT <chr>,
## #   SCH_MAGNETDETAIL <chr>, SCH_ALTFOCUS <chr>,
## #   SCH_PSENR_NONIDEA_A3 <chr>, SCH_PSENR_NONIDEA_A4 <chr>,
## #   SCH_PSENR_NONIDEA_A5 <chr>, SCH_PSENR_HI_M <int>,
## #   SCH_PSENR_HI_F <int>, SCH_PSENR_AM_M <int>, SCH_PSENR_AM_F <int>,
## #   SCH_PSENR_AS_M <int>, SCH_PSENR_AS_F <int>, SCH_PSENR_HP_M <int>,
## #   SCH_PSENR_HP_F <int>, SCH_PSENR_BL_M <int>, SCH_PSENR_BL_F <int>,
## #   SCH_PSENR_WH_M <int>, SCH_PSENR_WH_F <int>, SCH_PSENR_TR_M <int>,
## #   SCH_PSENR_TR_F <int>, TOT_PSENR_M <int>, TOT_PSENR_F <int>,
## #   SCH_PSENR_LEP_M <int>, SCH_PSENR_LEP_F <int>, SCH_PSENR_IDEA_M <int>,
## #   SCH_PSENR_IDEA_F <int>, DSO_SCH_PSENR_HI_F <int>,
## #   DSO_SCH_PSENR_LEP_F <int>, DSO_SCH_PSENR_LEP_M <int>,
## #   SCH_ENR_HI_M <int>, SCH_ENR_HI_F <int>, SCH_ENR_AM_M <int>,
## #   SCH_ENR_AM_F <int>, SCH_ENR_AS_M <int>, SCH_ENR_AS_F <int>,
## #   SCH_ENR_HP_M <int>, SCH_ENR_HP_F <int>, SCH_ENR_BL_M <int>,
## #   SCH_ENR_BL_F <int>, SCH_ENR_WH_M <int>, SCH_ENR_WH_F <int>,
## #   SCH_ENR_TR_M <int>, SCH_ENR_TR_F <int>, TOT_ENR_M <int>,
## #   TOT_ENR_F <int>, SCH_ENR_LEP_M <int>, SCH_ENR_LEP_F <int>,
## #   SCH_ENR_504_M <int>, SCH_ENR_504_F <int>, SCH_ENR_IDEA_M <int>,
## #   SCH_ENR_IDEA_F <int>, SCH_LEPENR_HI_M <int>, SCH_LEPENR_HI_F <int>,
## #   SCH_LEPENR_AM_M <int>, SCH_LEPENR_AM_F <int>, SCH_LEPENR_AS_M <int>,
## #   SCH_LEPENR_AS_F <int>, SCH_LEPENR_HP_M <int>, SCH_LEPENR_HP_F <int>,
## #   SCH_LEPENR_BL_M <int>, SCH_LEPENR_BL_F <int>, SCH_LEPENR_WH_M <int>,
## #   SCH_LEPENR_WH_F <int>, SCH_LEPENR_TR_M <int>, SCH_LEPENR_TR_F <int>,
## #   TOT_LEPENR_M <int>, TOT_LEPENR_F <int>, SCH_LEPPROGENR_HI_M <int>,
## #   SCH_LEPPROGENR_HI_F <int>, SCH_LEPPROGENR_AM_M <int>,
## #   SCH_LEPPROGENR_AM_F <int>, ...
```

Problem 7

We would like to investigate whether Black students receive a disproportionate number of expulsions under zero-tolerance policies.

Create a new `data.frame` or `tibble` with the following columns:

- The total number of students enrolled at each school
- The total number of Black students enrolled at each school
- The total number of students who received an expulsion under zero-tolerance policies
- The number of Black students who received an expulsion under zero-tolerance policies
- The proportion of students at each school who are Black

- The proportion of students expelled under zero-tolerance policies who are Black

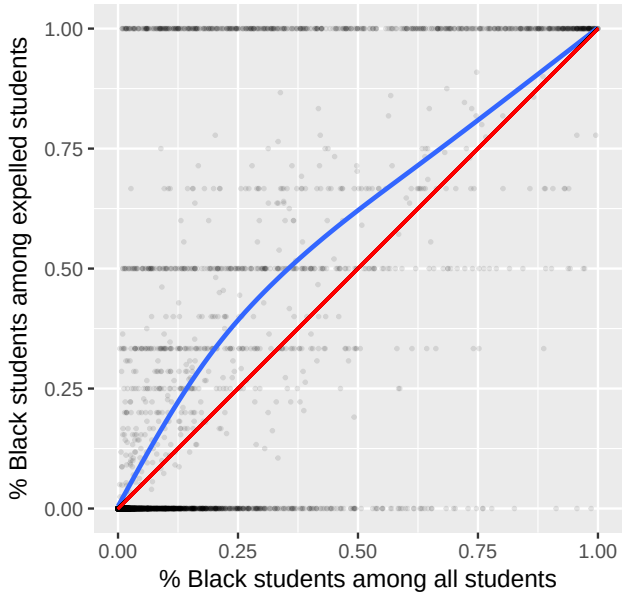
Filter the data to include only those schools in which at least one student received an expulsion under zero-tolerance policies.

Plot the proportion of Black students at each school (on the x-axis) versus the proportion of expelled students who are Black (on the y-axis). Include a smoothing line on the plot.

What do you observe in the plot? Does the plot indicate an over- or under-representation of Black students in expulsions under zero-tolerance policies?

Calculate the overall proportion of Black students across all schools and the overall proportion of students expelled under zero-tolerance policies who are Black across all schools.

```
crdc_exp <- transmute(crdc,
                      enr_tot = TOT_ENR_M + TOT_ENR_F,
                      enr_bl = SCH_ENR_BL_M + SCH_ENR_BL_F,
                      exp_tot = TOT_DISCWODIS_EXPZT_M +
                        TOT_DISCWODIS_EXPZT_F +
                        TOT_DISCWDIS_EXPZT_IDEA_M +
                        TOT_DISCWDIS_EXPZT_IDEA_F,
                      exp_bl = SCH_DISCWODIS_EXPZT_BL_M +
                        SCH_DISCWODIS_EXPZT_BL_F +
                        SCH_DISCWDIS_EXPZT_IDEA_BL_M +
                        SCH_DISCWDIS_EXPZT_IDEA_BL_F,
                     prop_bl = enr_bl / enr_tot,
                     prop_exp_bl = exp_bl / exp_tot) %>%
  filter(exp_tot >= 1)

crdc_exp %>%
  ggplot(aes(x=prop_bl, y=prop_exp_bl)) +
  geom_point(alpha=1/10, size=0.5) + geom_smooth(se=FALSE) +
  geom_segment(aes(x=0, y=0, xend=1, yend=1), color="red") +
  coord_fixed() +
  labs(x='% Black students among all students',
       y='% Black students among expelled students')
```



We use `coord_fixed` to make fixed scale coordinates in which the x- and y-axis have the same length for one unit. This makes it easier to interpret the plot. We also draw a reference line using `geom_segment` to represent the case when the two proportions are the same.

In an world with equally-administered consequences, the proportion of expelled students who are black should be approximately the same as the proportion of black students in the whole students body (shown by the red reference line). But the proportion of expelled students who are Black is actually typically greater than the overall proportion of Black students (shown by the blue smooth line), indicating an over-representation of Black students among students expelled under zero-tolerance policies

The overall proportions are as follows:

```
summarise(crdc_exp,
          prop_bl=sum(enr_bl, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
          prop_exp_bl=sum(exp_bl, na.rm=TRUE) / sum(exp_tot, na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##     prop_bl prop_exp_bl
##       <dbl>       <dbl>
## 1 0.1680485   0.2304784
```

Roughly 15% of the overall student population is Black, but Black students represent roughly 23% of all students expelled under zero-tolerance policies.

Problem 8

We would like to investigate whether Hispanic students are over- or under-represented in Gifted & Talented programs.

Create a new ‘data.frame‘ or ‘tibble‘ containing only schools with a Gifted & Talented program with the following columns:

- The total number of students enrolled at each school
- The total number of Hispanic students at each school
- The total number of students in the school’s GT program

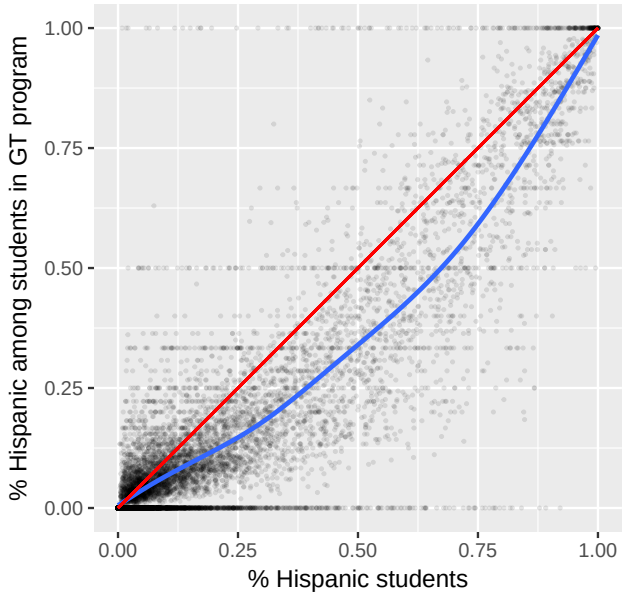
- The number of students in the GT program who are Hispanic
- The proportion of students at each school who are Hispanic
- The proportion of students in the GT program who are Hispanic

Plot the proportion of Hispanic students at each school (on the x-axis) versus the proportion of GT students who are Hispanic (on the y-axis). Include a smooth line on the plot.

What do you observe in the plot? Does the plot indicate an over- or under-representation of Hispanic students in Gifted & Talented programs?

Calculate the overall proportion of Hispanic students across all schools and the overall proportion of GT students who are Hispanic.

```
crdc_gt <- filter(crdc, SCH_GT_IND=="YES") %>%
                transmute(
                       enr_tot = TOT_ENR_M + TOT_ENR_F,
                       enr_hi = SCH_ENR_HI_M + SCH_ENR_HI_F,
                       gt_tot = TOT_GTENR_M + TOT_GTENR_F,
                       gt_hi = SCH_GTENR_HI_M + SCH_GTENR_HI_F,
                       prop_hi = enr_hi / enr_tot,
                       prop_gt_hi = gt_hi / gt_tot)

crdc_gt %>% sample_n(10000) %>%
  ggplot(aes(x=prop_hi, y=prop_gt_hi)) +
  geom_point(alpha=1/10, size=0.4) + geom_smooth(se=FALSE) +
  geom_segment(aes(x=0, y=0, xend=1, yend=1), color="red") +
  coord_fixed() +
  labs(x='% Hispanic students', y='% Hispanic among students in GT program')
```



The fitted smooth lines shows that the proportion of Gifted & Talented students who are Hispanic is typically lower than the proportion of Black and Hispanic students at each school. This indicates an under-representation of Black and Hispanic students in Gifted & Talented programs.

```
summarise(crdc_gt,
    pr_hi=sum(enr_hi, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
    pr_gt_hi=sum(gt_hi, na.rm=TRUE) / sum(gt_tot, na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##       pr_hi pr_gt_hi
##       <dbl>    <dbl>
## 1 0.2645159 0.180666
```

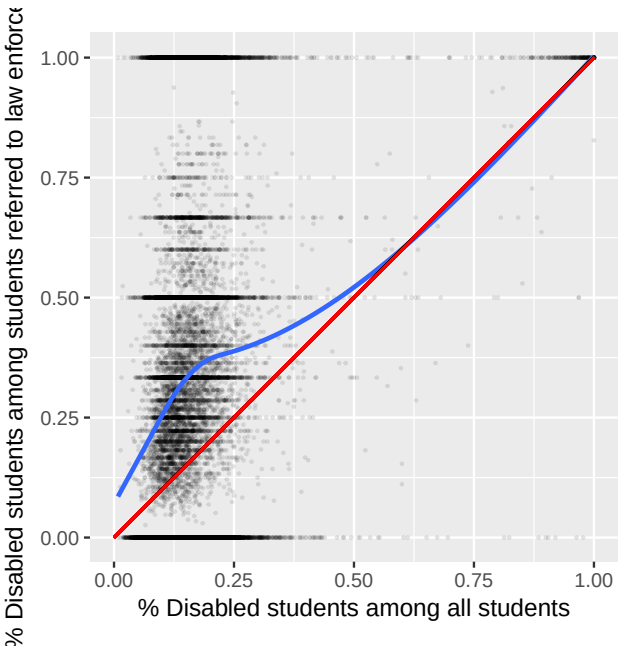
Hispanic students represent roughly 26% of the overall student population, but only 18% of GT students.

Problem 9

> We would like to investigate whether disabled students are more often referred to a law enforcement agency or official.
>
> Create a new `data.frame` or `tibble` containing only schools that use corporal punishment with the following columns:
>
> - The total number of students enrolled at each school
> - The total number of disabled students (under IDEA and/or 504) at each school
> - The total number of students who were referred to law enforcement
> - The number of disabled students who were referred to law enforcement
> - The proportion of students at each school who are disabled
> - The proportion of students who were referred to law enforcement who are disabled
>
> Filter the data to include only those schools without errors in data entry (i.e., remove all schools with more disabled students enrolled than the total number of enrolled students).
>
> Plot the proportion of disabled students at each school (on the x-axis) versus the proportion of students referred to law enforcement who are disabled (on the y-axis). Include a smoothing line on the plot.
>
> What do you observe in the plot? Does the plot indicate an over- or under-representation of disabled students among students who are referred to law enforcement?
>
> Calculate the overall proportion of disabled students across all schools and the overall proportion of students referred to law enforcement who are disabled across all schools.

```
crdc_ref <- transmute(crdc,
                      enr_tot = TOT_ENR_M + TOT_ENR_F,
                      enr_dis = SCH_ENR_IDEA_M +
                        SCH_ENR_IDEA_F +
                        SCH_ENR_504_M +
                        SCH_ENR_504_F,
                       ref_dis = TOT_DISCWDIS_REF_IDEA_M +
                         TOT_DISCWDIS_REF_IDEA_F +
                         SCH_DISCWDIS_REF_504_M +
                         SCH_DISCWDIS_REF_504_F,
                       ref_tot = ref_dis +
                         TOT_DISCWODIS_REF_M +
                         TOT_DISCWODIS_REF_F,
                       prop_dis=enr_dis / enr_tot,
                       prop_ref_dis=ref_dis / ref_tot) %>%
```

```
  filter(ref_tot >= 1, enr_dis <= enr_tot)

crdc_ref %>%
  ggplot(mapping=aes(x=prop_dis, y=prop_ref_dis)) +
  geom_point(alpha=1/10, size=0.3) + geom_smooth(se=FALSE) +
  geom_segment(aes(x=0, y=0, xend=1, yend=1), color="red") +
  coord_fixed() +
  labs(x='% Disabled students among all students',
       y='% Disabled students among students referred to law enforcement')
```



The fitted smooth line suggests that, until the schools reach roughly 50% disabled students, disabled students are over-represented among students who are referred to law enforcement. This is indicated by the proportion of students referred to law enforcement who are disabled typically being greater than the proportion of disabled students at the school for `prop_dis < 0.50`. However, this relationship ceases as the proportion of disabled students at the school increases. But the second claim should be taken with a grain of salt, as we have much less data points where `prop_dis > 0.50`.

```
summarise(crdc_ref,
    prop_dis=sum(enr_dis, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
    prop_ref_dis=sum(ref_dis, na.rm=TRUE) / sum(ref_tot, na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##    prop_dis prop_ref_dis
##       <dbl>        <dbl>
## 1 0.1454699    0.2881855
```

Disabled students represent roughly 15% of the overall student population but nearly 29% of students referred to law enforcement agencies or officials for discipline.

Problem 10

Develop your own question about whether a particular demographic is over- or under-represented in a particular aspect of the education system.

State your question.

Process, plot, and summarise the data to answer your question. State what you observe in the plot and your conclusions based on the plot and the summary statistics.

```
## Solutions will vary
```