Assignment 1

Mounica Subramani

January 25, 2018

```
# Import required library files

library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v tibble  1.4.1     v purrr   0.2.4
## v tidyr   0.7.2     v dplyr   0.7.4
## v readr   1.1.1     v stringr 1.2.0
## v tibble  1.4.1     v forcats 0.2.0
## -- Conflicts ------------------------------------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(rmarkdown)
library(nycflights13)
library(maps)
```

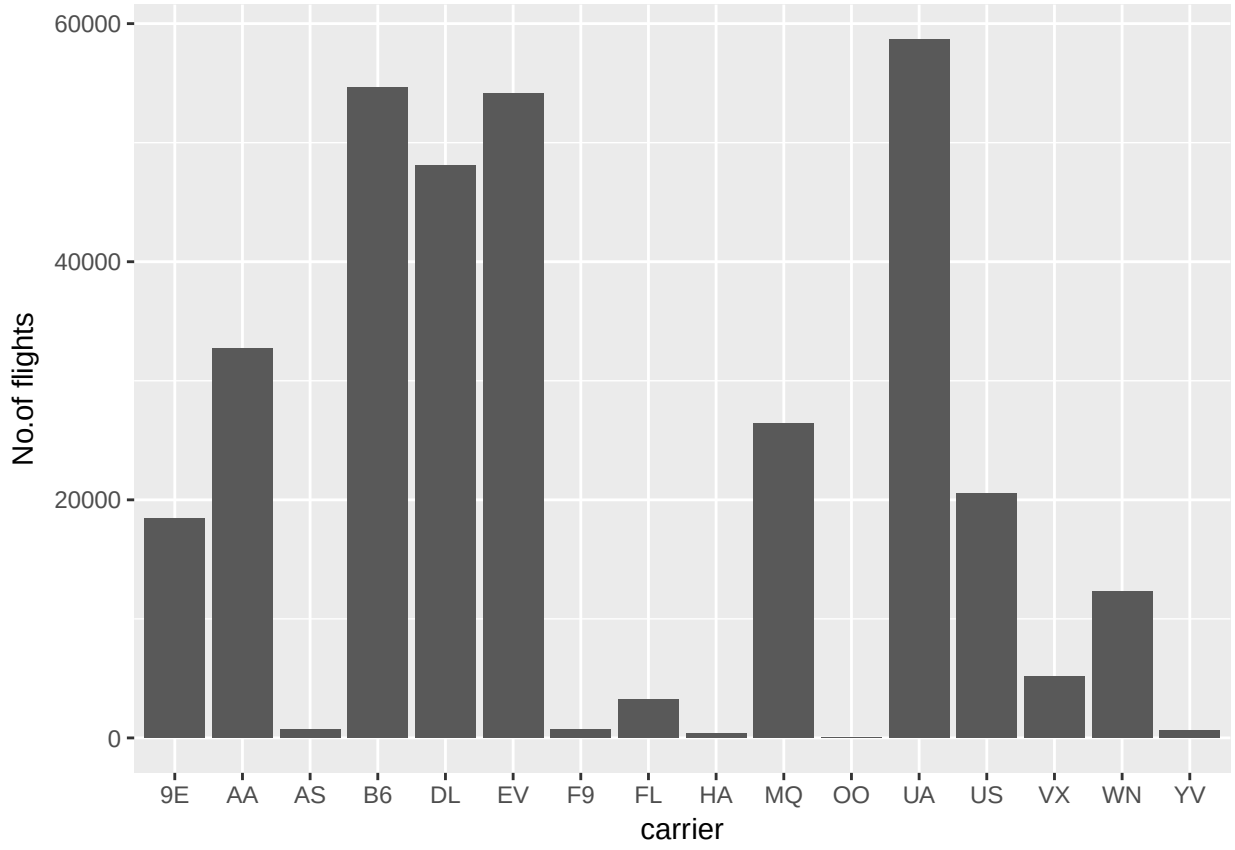
```
##
## Attaching package: 'maps'
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(measurements)
options(width = 90)
```

Part A

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?

```
# X axis featuring Carrier and Y axis representing count of
#flights flown out of New York airports in 2013.

ggplot(data=flights,
       mapping = aes(x=carrier))+layer(geom = "bar",
                                       stat="count",
                                       position = "identity") + labs(y="No.of flights")
```



United Airlines "UA" carrier flew the most flights in 2013 from New York airport.

Problem 2

For each destination, calculate the proportion of flights that arrived at their destination earlier than scheduled.

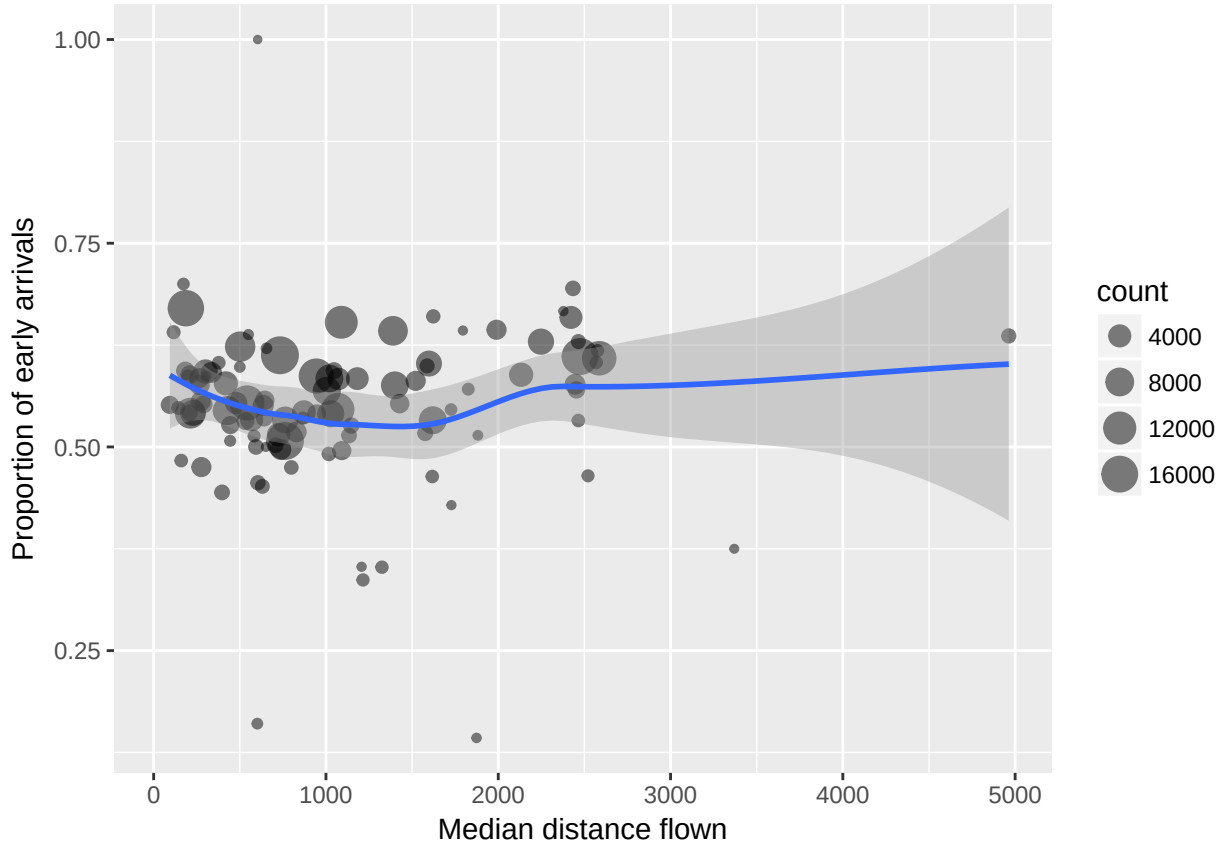
Also calculate the median distance flown to each destination. Plot the proportion of early arrivals (on the y-axis) against the median distance flown (on the x-axis) for each destination. Add a smooth line to the plot. Based on the smooth line, at what distances are flights most likely to arrive early?

```
flights %>%
    group_by(dest) %>%
    summarise(prop_early = mean(arr_delay < 0, na.rm=TRUE),
              m_distance = median(distance, na.rm=TRUE),
              count = n()) %>%
  ggplot(aes(x=m_distance,
             y=prop_early)) + geom_point(aes(size=count),
alpha =1/2) + geom_smooth() + labs(x="Median distance flown",
                                   y="Proportion of early arrivals")
```

```
## `geom_smooth()` using method = 'loess'
```

```
## Warning: Removed 1 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
#prop_early defines the calculation of proportion of flights that arrived at
#-their destination earlier than scheduled.

## m_distance gives median distance of flights flown to each destination
## At the distance 1000-1500 the flights are most likely to arrive early.
```

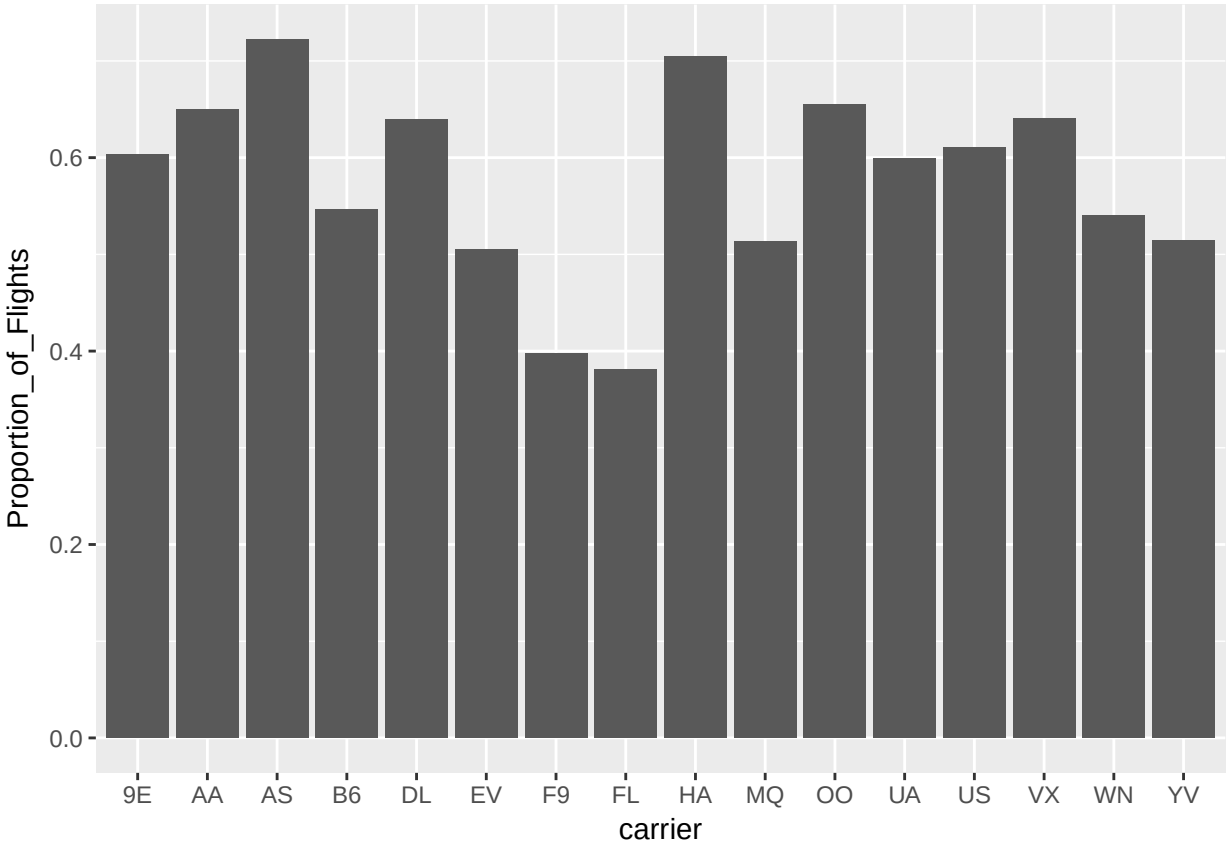
Problem 3

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 arrive for each carrier.

Which airlines are the most consistently ahead of schedule? Which airlines arrive the most early?

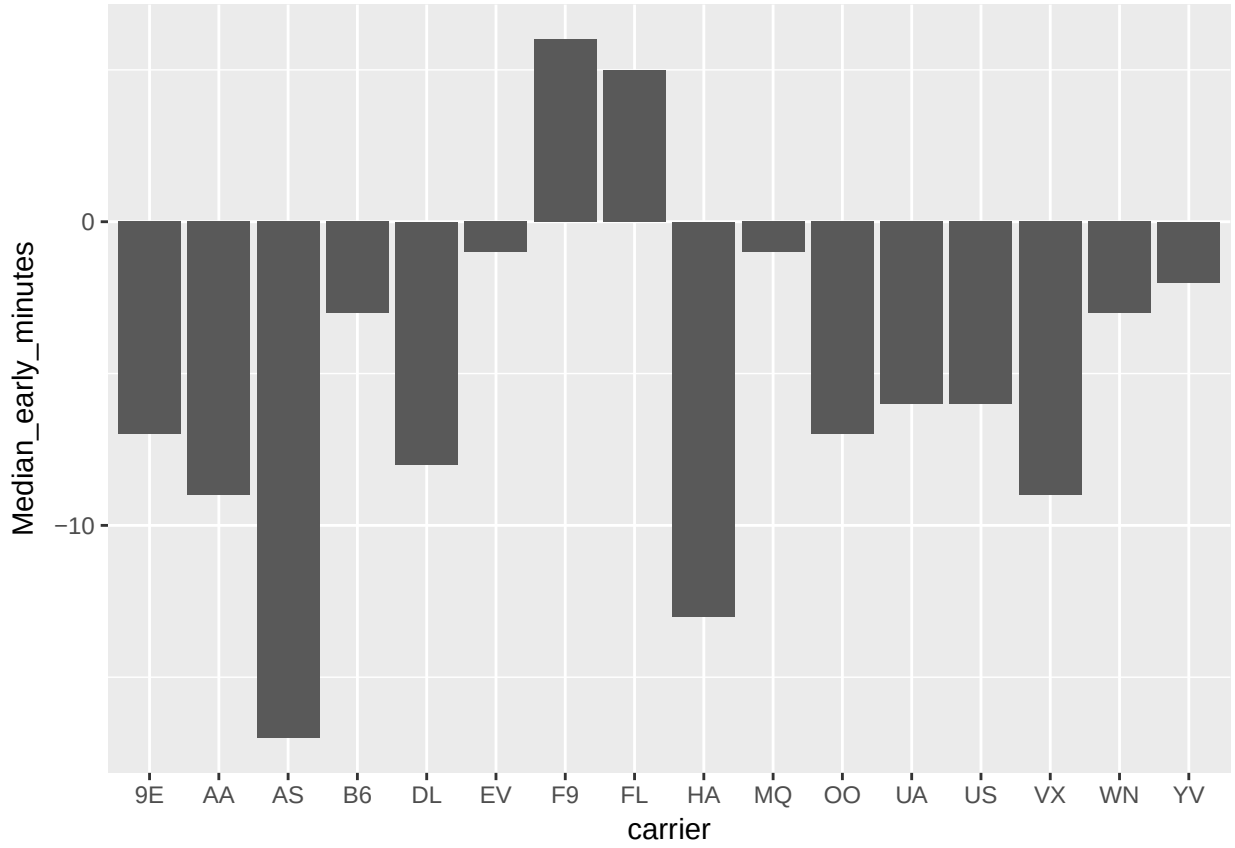
Which airlines are most consistently behind schedule? Which airlines arrive the latest?

```
## Proportion of flights that arrive early for each carrier.

flights %>%
    group_by(carrier) %>%
    summarise(Proportion_of_Flights = mean(arr_delay <0 , na.rm=TRUE)) %>%
    ggplot(aes(x=carrier, y=Proportion_of_Flights)) + geom_col()
```



Alaska Airline "AS" and Hawaiian Airline "HA" are more consistently ahead of the schedule. AS airlines arive the most early.

```
## Median number of minutes early that flights arrive for each carrier.

flights %>%
    group_by(carrier) %>%
    summarise(Median_early_minutes = median(arr_delay , na.rm=TRUE)) %>%
    ggplot(aes(x=carrier, y=Median_early_minutes)) + geom_col()
```



F9 and AirTran "FL" airlines are most consistently behind the schedule. F9 ariline arrive the latest.

PART B

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?)

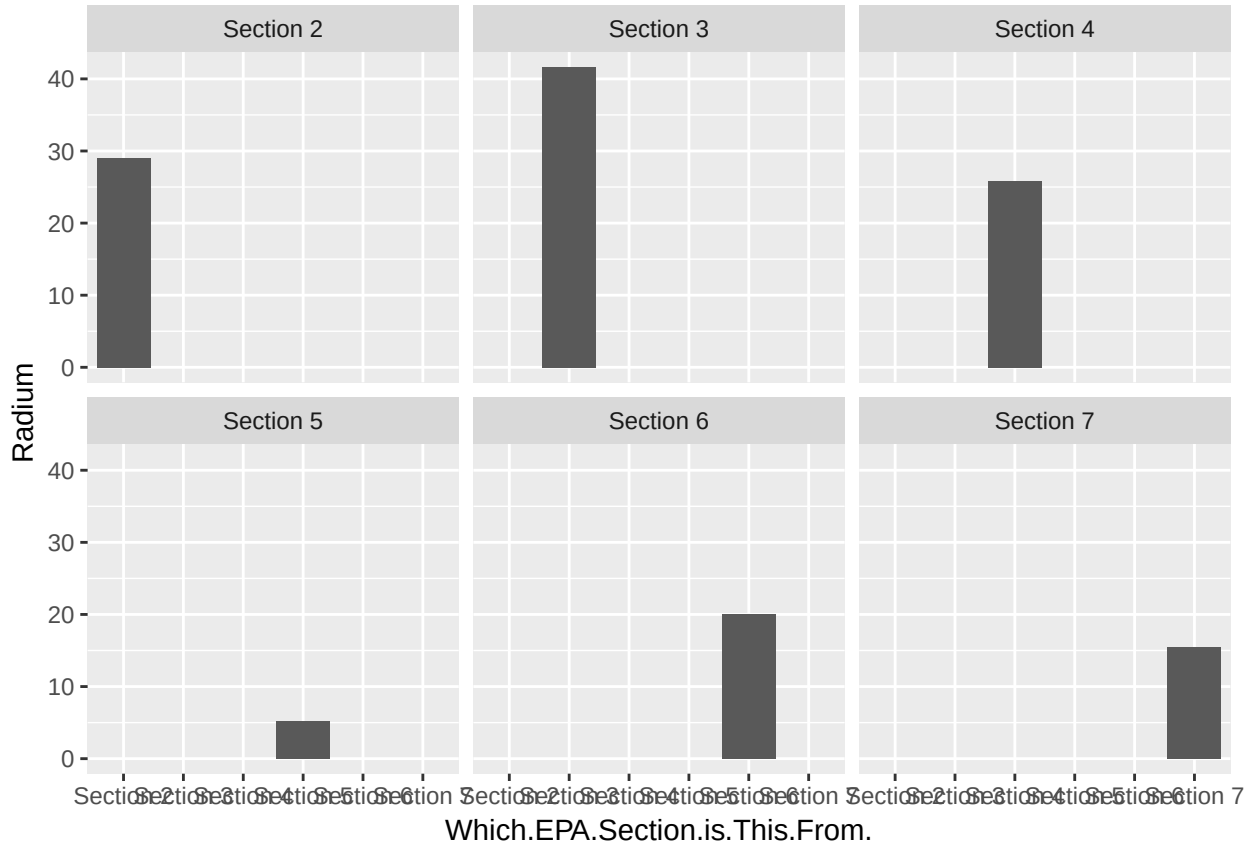
The concentration of radiactive elements in a sample is measured in rate of atomic disintegrations per volume, rather than mass per volume, as used for stable isotopes. This is done by counting the number of atomic disintegrations per minute and comparing it to the mass of the material involved. However, 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.

```
## Read the CSV file "NavajoWaterExport" into "PartB_Data" data frame.

PartB_Data <- read.csv("C://users//mouni//Downloads//NavajoWaterExport.csv")
```

```
## Mutating the dataset to replace the negative values with 0.

mydata<-mutate(PartB_Data, Radium=ifelse(Amount.of.Radium228 < 0, 0, Amount.of.Radium228))
```

```
## Plot histograms showing the distribution of the amount of Radium-228
#-in water samples for each EPA section (use faceting).

ggplot(mydata, mapping = aes(x=Which.EPA.Section.is.This.From.,
y=Radium)) + geom_col() + facet_wrap(~Which.EPA.Section.is.This.From.)
```



Besides the fact that water samples are radioactive, the concentration of radium is high in Section 3 and Section 2, whereass they are considerably less in Section 5 and Section 7.

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?

```
## Filtering the dataset to remove any sites with "Unknown Risk Rating" for the EPA risk rating.

mydata <- mydata %>%
filter(!mydata$US.EPA.Risk.Rating == "Unknown Risk")
```

```
## Counting number of sites of each EPA risk rating in each EPA section.

summarise(group_by(mydata, Which.EPA.Section.is.This.From.,US.EPA.Risk.Rating), n())
```

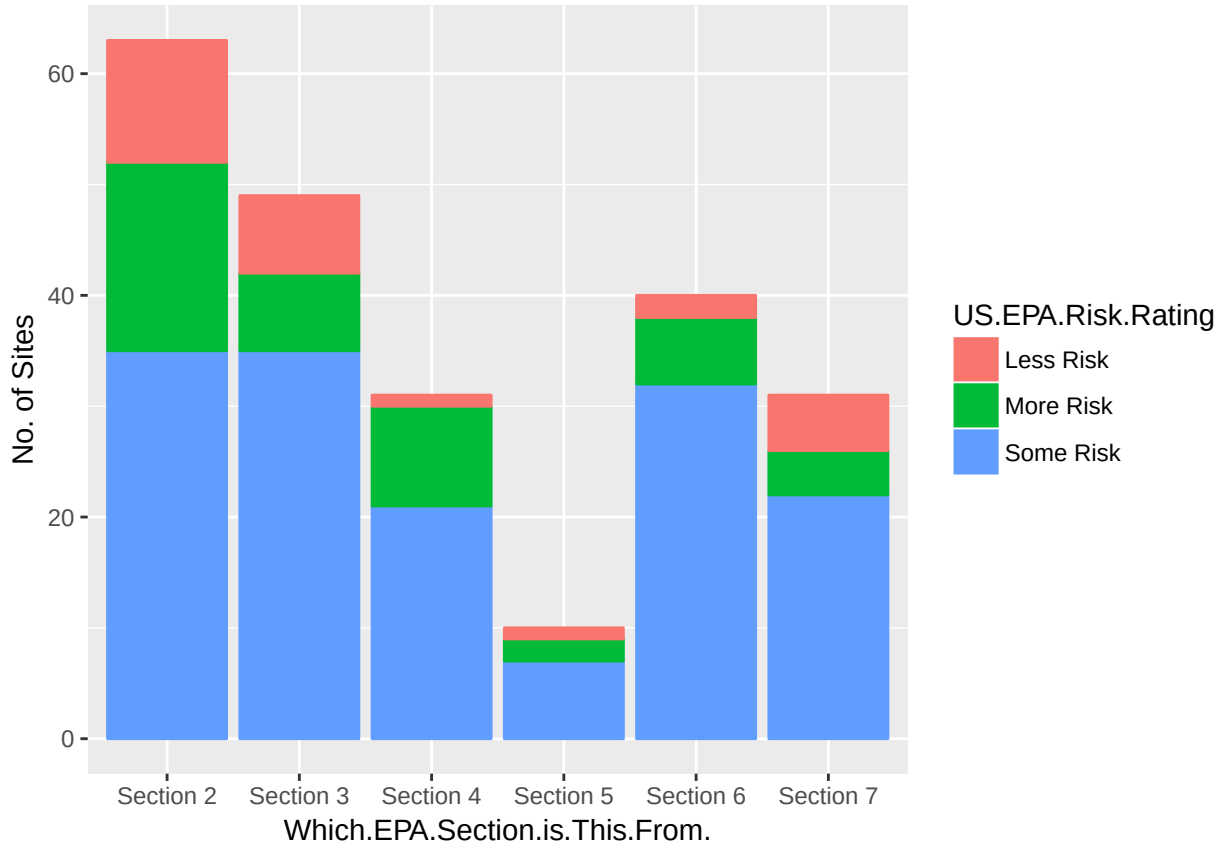
```
## # A tibble: 18 x 3
## # Groups: Which.EPA.Section.is.This.From. [?]
##    Which.EPA.Section.is.This.From. US.EPA.Risk.Rating `n()`
##    <fct>                           <fct>              <int>
##  1 Section 2                       Less Risk             11
##  2 Section 2                       More Risk             17
##  3 Section 2                       Some Risk             35
##  4 Section 3                       Less Risk              7
##  5 Section 3                       More Risk              7
##  6 Section 3                       Some Risk             35
##  7 Section 4                       Less Risk              1
##  8 Section 4                       More Risk              9
##  9 Section 4                       Some Risk             21
## 10 Section 5                       Less Risk              1
## 11 Section 5                       More Risk              2
## 12 Section 5                       Some Risk              7
## 13 Section 6                       Less Risk              2
## 14 Section 6                       More Risk              6
## 15 Section 6                       Some Risk             32
## 16 Section 7                       Less Risk              5
## 17 Section 7                       More Risk              4
## 18 Section 7                       Some Risk             22
```

```
## Calcuating the mean concentration of Uranium 238 in the water
#-samples for each EPA risk rating in each EPA section.

summarise(group_by(mydata, Which.EPA.Section.is.This.From.,US.EPA.Risk.Rating),
          mean(Amount.of.Uranium238, na.rm = TRUE))
```

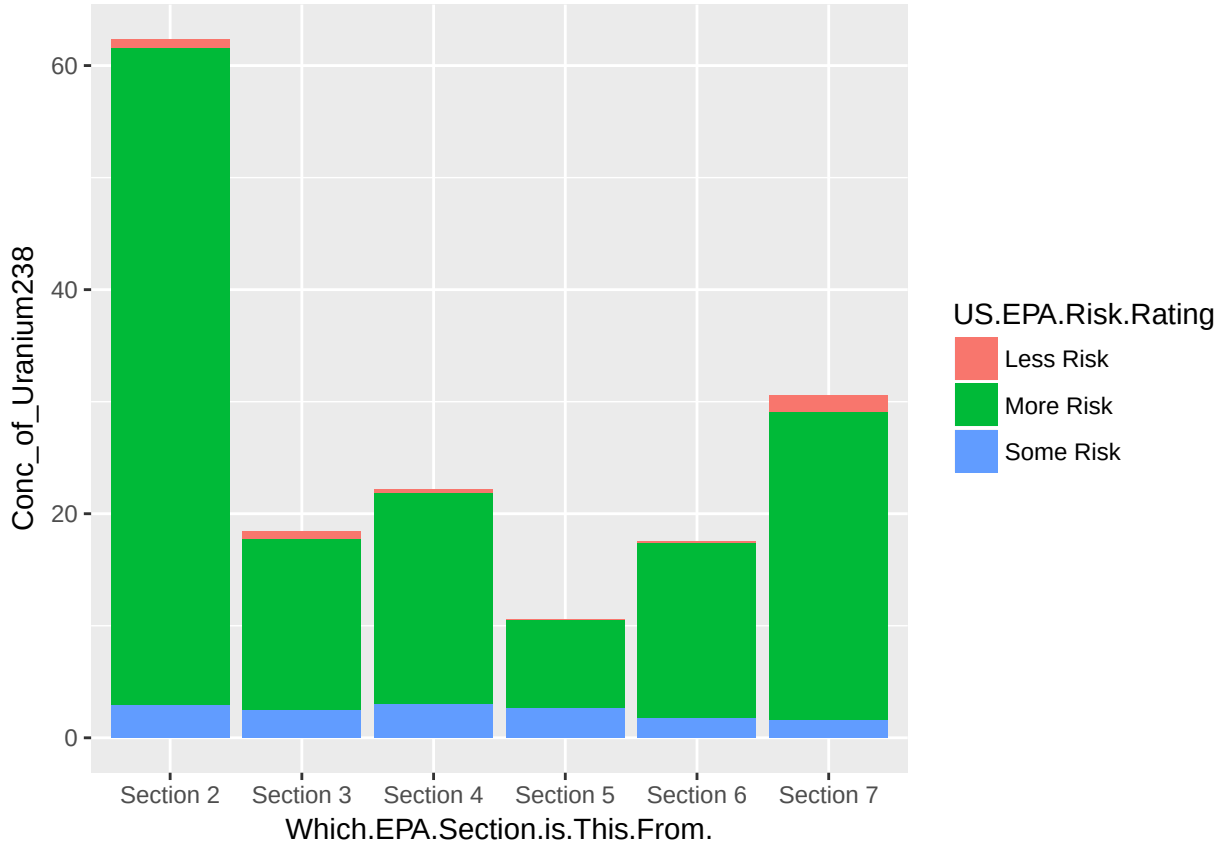
```
## # A tibble: 18 x 3
## # Groups: Which.EPA.Section.is.This.From. [?]
##    Which.EPA.Section.is.This.From. US.EPA.Risk.Rating `mean(Amount.of.Uranium238, na.rm ~
##    <fct>                           <fct>                                               <dbl>
##  1 Section 2                       Less Risk                                           0.755
##  2 Section 2                       More Risk                                          58.6
##  3 Section 2                       Some Risk                                           2.95
##  4 Section 3                       Less Risk                                           0.667
##  5 Section 3                       More Risk                                          15.3
##  6 Section 3                       Some Risk                                           2.50
##  7 Section 4                       Less Risk                                           0.340
##  8 Section 4                       More Risk                                          18.9
##  9 Section 4                       Some Risk                                           2.99
## 10 Section 5                       Less Risk                                           0.0600
## 11 Section 5                       More Risk                                           7.80
## 12 Section 5                       Some Risk                                           2.72
## 13 Section 6                       Less Risk                                           0.120
```

```
## 14 Section 6                        More Risk                                15.6
## 15 Section 6                        Some Risk                                 1.79
## 16 Section 7                        Less Risk                                 1.46
## 17 Section 7                        More Risk                                27.6
## 18 Section 7                        Some Risk                                 1.56
```

```r
## Plotting number of sites at each EPA section using a bar plot.

ggplot(mydata) + geom_bar(aes(x=Which.EPA.Section.is.This.From. ,
                              color=US.EPA.Risk.Rating,
                              fill=US.EPA.Risk.Rating)) + labs(y="No. of Sites")
```




```r
## Plotting mean concentration of Uranium-238 for each EPA section using a bar plot.

mydata %>%
    group_by(Which.EPA.Section.is.This.From., US.EPA.Risk.Rating) %>%
    summarise(Conc_of_Uranium238 = mean(Amount.of.Uranium238, na.rm = TRUE)) %>%
    ggplot(aes(x=Which.EPA.Section.is.This.From. ,
               y=Conc_of_Uranium238)) + geom_col(aes(fill=US.EPA.Risk.Rating))
```



EPA section "Section 2" have the most sites with "More Risk" and also has the highest concentration of Uranium 238 on average.

Problem 6

Install the maps package (you do not need to load it) and use the ggplo2::map_data function to get data for drawing the "Four Corners" region of the United States (i.e., Arizona, New Mexico, Utah, and Colorado).

```
four_corners <- map_data("state", region=c("arizona", "new mexico", "utah", "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.

```
mydata$Latitude = measurements::conv_unit(mydata$Latitude, 'deg_dec_min', 'dec_deg')
```

```
## Warning in split.default(as.numeric(unlist(strsplit(x, " "))) * c(3600, : data length is
## not a multiple of split variable
```

```
mydata$Longitude = measurements::conv_unit(mydata$Longitude, 'deg_dec_min', 'dec_deg')
```

```
## Warning in split.default(as.numeric(unlist(strsplit(x, " "))) * c(3600, : data length is
## not a multiple of split variable
```

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.

```
# ggplot() + geom_polygon(four_corners,
# aes(x=long, y=lat, group=group),
```

```
# fill=NA,color="black")+coord_map()+geom_point(mydata,
# aes(x=Longitude,
#     y=Latitude,
#     fill=Which.EPA.Section.is.This.From.))
```

Part C

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 smooth 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.

```
## Read the CSV file.

C_Data <- read_csv("C:/users/mouni/Downloads/crdc201314/CRDC2013_14_SCH.csv",
                   na=c("-9", "-5", "-2"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   LEA_STATE = col_character(),
##   LEA_NAME = col_character(),
##   SCH_NAME = col_character(),
##   COMBOKEY = col_character(),
##   LEAID = col_character(),
##   SCHID = col_character(),
##   JJ = col_character(),
##   CCD_LATCOD = col_double(),
##   CCD_LONCOD = col_double(),
##   NCES_SCHOOL_ID = col_character(),
##   MATCH_FLAG = col_character(),
##   SCH_GRADE_PS = col_character(),
##   SCH_GRADE_KG = col_character(),
##   SCH_GRADE_G01 = col_character(),
##   SCH_GRADE_G02 = col_character(),
##   SCH_GRADE_G03 = col_character(),
##   SCH_GRADE_G04 = col_character(),
```

```
##   SCH_GRADE_G05 = col_character(),
##   SCH_GRADE_G06 = col_character(),
##   SCH_GRADE_G07 = col_character()
##   # ... with 75 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
#A tibble containing required columns.

CRDC_BL_Prop <- transmute(C_Data,
                      Enr_tot = TOT_ENR_M + TOT_ENR_F,
                      Enr_Black_tot = SCH_ENR_BL_M + SCH_ENR_BL_F,
                      Enr_Tot_Tolerance = TOT_DISCWODIS_EXPZT_M + TOT_DISCWODIS_EXPZT_F
                      + TOT_DISCWDIS_EXPZT_IDEA_M + TOT_DISCWDIS_EXPZT_IDEA_F
                      +SCH_DISCWDIS_EXPZT_LEP_M +SCH_DISCWDIS_EXPZT_LEP_F
                      + SCH_DISCWDIS_EXPZT_504_M,
                      Enr_Tot_Tolerance_BL = SCH_DISCWODIS_EXPZT_BL_M + SCH_DISCWODIS_EXPZT_BL_F
                      + SCH_DISCWDIS_EXPZT_IDEA_BL_M + SCH_DISCWDIS_EXPZT_IDEA_BL_F,
                      Black_Prop=Enr_Black_tot/Enr_tot,
                      Black_Tolerance_Prop = Enr_Tot_Tolerance_BL/Enr_tot)
CRDC_BL_Prop
```

```
## # A tibble: 95,507 x 6
##    Enr_tot Enr_Black_tot Enr_Tot_Tolerance Enr_Tot_Toleranc~ Black_Prop Black_Tolerance_~
##      <int>         <int>             <int>             <int>      <dbl>             <dbl>
##  1    1798          1001                 0                 0    0.557                 0
##  2     994           599                 0                 0    0.603                 0
##  3     910           500                 0                 0    0.549                 0
##  4     635            13                 0                 0    0.0205                0
##  5    1114            25                 0                 0    0.0224                0
##  6     680            10                 0                 0    0.0147                0
##  7     783            19                 0                 0    0.0243                0
##  8     479            10                 0                 0    0.0209                0
##  9    1032            19                 0                 0    0.0184                0
## 10     423             4                 0                 0    0.00946               0
## # ... with 95,497 more rows
```

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

PartC_Data <- CRDC_BL_Prop %>%
  filter(!CRDC_BL_Prop$Enr_Tot_Tolerance == "0")
PartC_Data
```

```
## # A tibble: 3,845 x 6
##    Enr_tot Enr_Black_tot Enr_Tot_Tolerance Enr_Tot_Toleranc~ Black_Prop Black_Tolerance_~
##      <int>         <int>             <int>             <int>      <dbl>             <dbl>
##  1     871           238                 2                 0      0.273           0
##  2    1605           472                 2                 2      0.294           0.00125
##  3     329           310                 4                 4      0.942           0.0122
##  4     885           871                 5                 5      0.984           0.00565
##  5     670           322                 4                 2      0.481           0.00299
##  6    1850           241                60                15      0.130           0.00811
##  7    1744           292                42                30      0.167           0.0172
##  8     532           166                 2                 2      0.312           0.00376
```

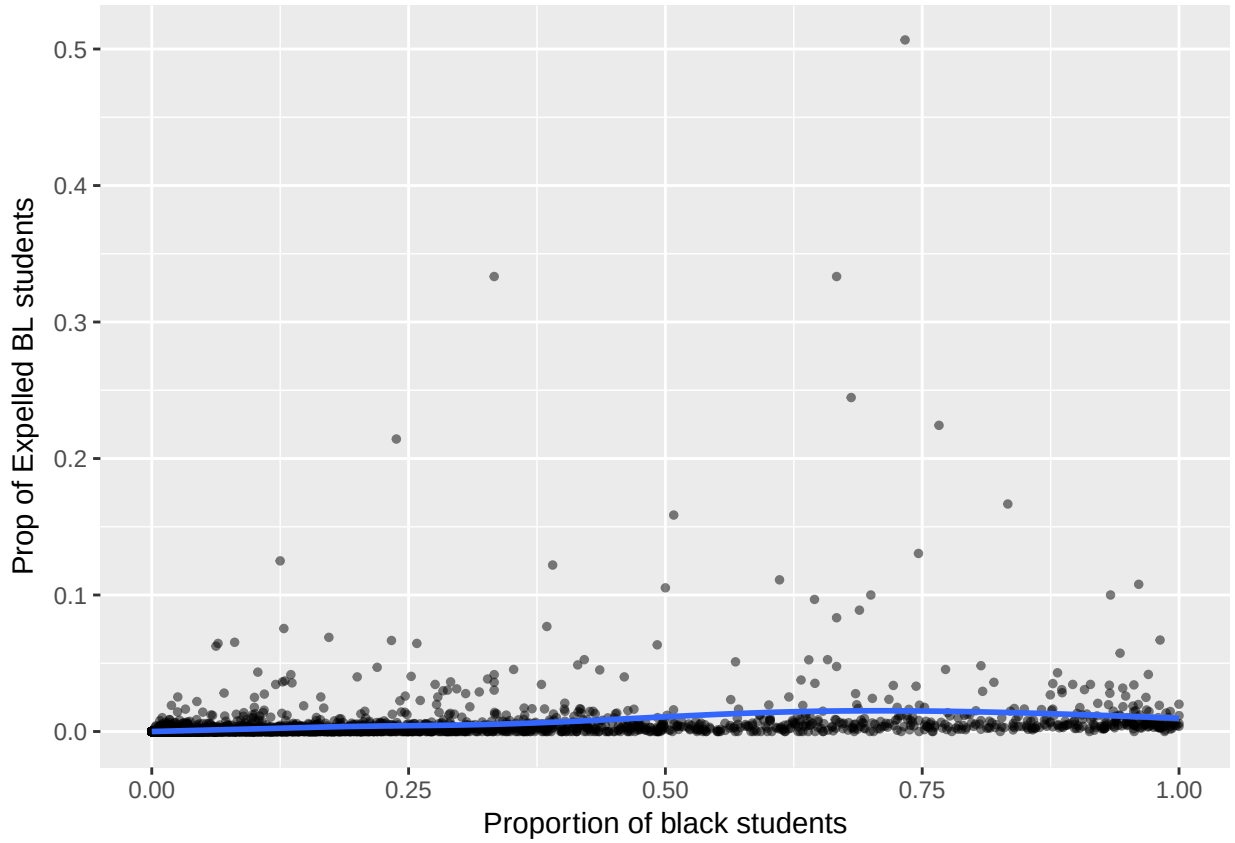
```
##  9     566           292                 8                 4      0.516          0.00707
## 10     360           148                 4                 4      0.411          0.0111
## # ... with 3,835 more rows
```

```
## Plotting 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).

ggplot(PartC_Data,
mapping = aes(x=Black_Prop,
              y=Black_Tolerance_Prop)) + geom_point(alpha=1/2,
              size=1) + geom_smooth() + labs(x="Proportion of black students",
                                             y="Prop of Expelled BL students")
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



The proportion of black students who are expelled under zero tolerance policies are under represented in the plot. There is no much expellsion is observed.

```
## Calculating the overall proportion of Black students across all schools.

summarise(PartC_Data, Overall_Prop_BL= mean(Enr_Black_tot>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_BL
##             <dbl>
```

```
## 1           0.945
```

```
## Calculating the overall proportion of students expelled under
#-zero-tolerance policies who are Black across all schools.

summarise(PartC_Data, Overall_Prop_BL_Tol= mean(Enr_Tot_Tolerance_BL>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_BL_Tol
##                 <dbl>
## 1               0.353
```

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.

```
## A tibble containing required columns

CRDC_Hisp_Prop <- transmute(C_Data,
                        Enr_tot = TOT_ENR_M + TOT_ENR_F,
                        Enr_Hisp_Tot = SCH_ENR_HI_M+ SCH_ENR_HI_F,
                        Enr_GT_Tot = TOT_GTENR_M + TOT_GTENR_F,
                        Enr_GT_Hisp_Tot = SCH_GTENR_HI_M + SCH_GTENR_HI_F,
                        Hisp_Prop = Enr_Hisp_Tot/Enr_tot,
                        GT_Hisp_Prop = Enr_GT_Hisp_Tot/Enr_tot)
CRDC_Hisp_Prop
```

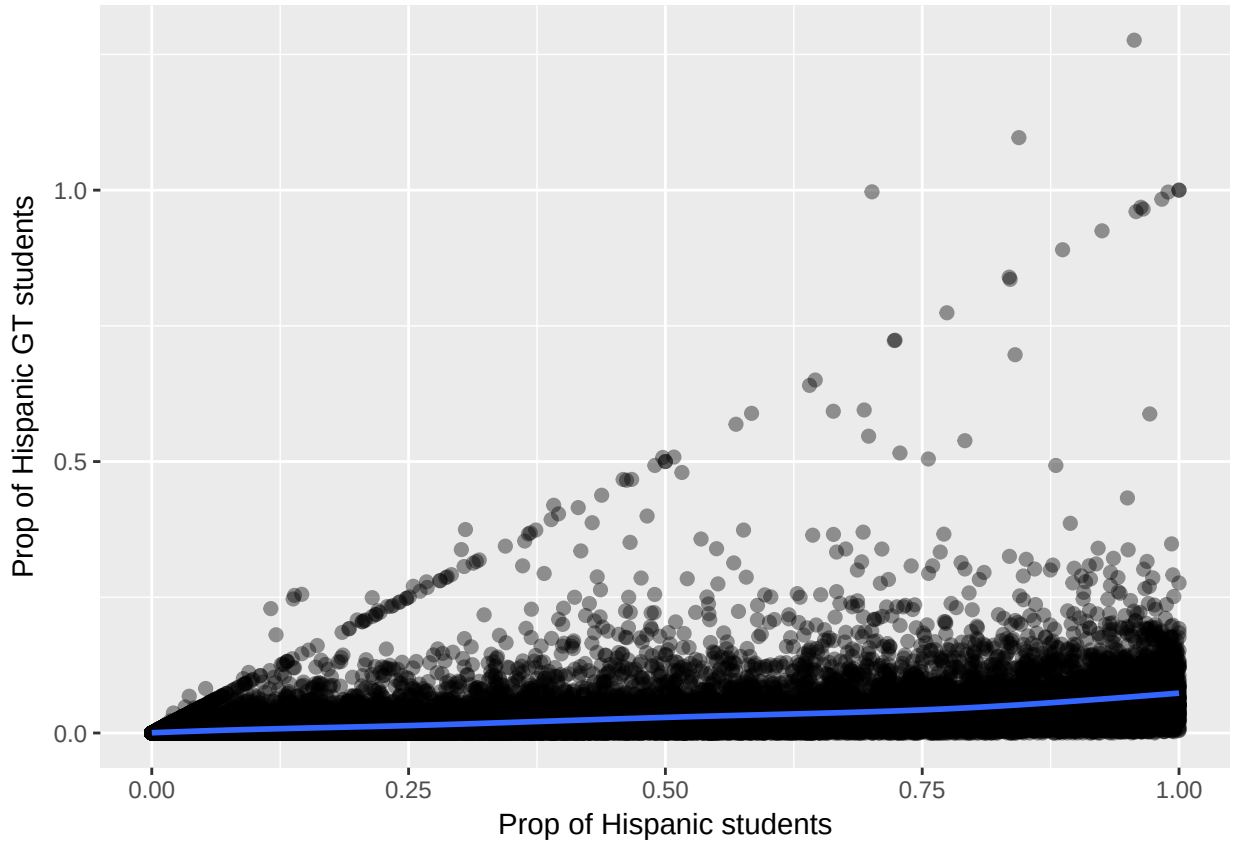
```
## # A tibble: 95,507 x 6
##    Enr_tot Enr_Hisp_Tot Enr_GT_Tot Enr_GT_Hisp_Tot Hisp_Prop GT_Hisp_Prop
##      <int>        <int>      <int>           <int>     <dbl>        <dbl>
##  1    1798            0         NA              NA    0           NA
##  2     994            0         NA              NA    0           NA
##  3     910            0         NA              NA    0           NA
##  4     635          235         NA              NA    0.370       NA
##  5    1114          310         NA              NA    0.278       NA
##  6     680          256         92               7    0.376        0.0103
##  7     783          322         63              13    0.411        0.0166
##  8     479          232         NA              NA    0.484       NA
##  9    1032          517         NA              NA    0.501       NA
## 10     423           10         46               0    0.0236       0
## # ... with 95,497 more rows
```

```
## Plotting 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).

ggplot(CRDC_Hisp_Prop,
mapping = aes(x=Hisp_Prop,
              y=GT_Hisp_Prop,
              na.rm=TRUE)) + geom_point(alpha=1/2.5,
              size=2) + geom_smooth(se=FALSE) + labs(x="Prop of Hispanic students",
              y="Prop of Hispanic GT students")
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 40551 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 40551 rows containing missing values (geom_point).
```



The plot indicates an under-representation of Hispanic students in Gifted & Talented programs.

```
## Calculating the overall proportion of Hispanic students across all schools.

summarise(CRDC_Hisp_Prop, Overall_Prop_Hisp= mean(Enr_Hisp_Tot>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_Hisp
##               <dbl>
## 1             0.938
```

```
## Calculating the overall proportion of GT students who are Hispanic.
```

```
summarise(CRDC_Hisp_Prop, Overall_Prop_Hisp= mean(Enr_GT_Hisp_Tot>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_Hisp
##               <dbl>
## 1             0.654
```

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 disciplined with corporal punishment . The number of disabled students who were disciplined with corporal punishment . The proportion of students at each school who are disabled . The proportion of students who were disciplined with corporal punishment 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.

```
## A Tibble containing required columns.

CRDC_Corp_Prop <- transmute(C_Data,
                        Enr_Tot = TOT_ENR_M + TOT_ENR_F,
                        WDIS_IDEA_504_Tot = TOT_DISCWDIS_REF_IDEA_M
                        + TOT_DISCWDIS_REF_IDEA_F + TOT_504ENR_M + TOT_504ENR_F,
                        Corp_Tot = TOT_DISCWODIS_CORP_M + TOT_DISCWODIS_CORP_F
                        + TOT_DISCWDIS_CORP_IDEA_M + TOT_DISCWDIS_CORP_IDEA_F,
                        WDIS_IDEA_504_Prop = WDIS_IDEA_504_Tot/Enr_Tot,
                        WDIS_Corp_Tot = TOT_DISCWDIS_CORP_IDEA_M + TOT_DISCWDIS_CORP_IDEA_F,
                        WDIS_Corp_Prop= WDIS_Corp_Tot/Enr_Tot,
                        WDIS_REF_Tot = TOT_DISCWDIS_REF_IDEA_M + TOT_DISCWDIS_REF_IDEA_F,
                        WDIS_REF_Prop= WDIS_REF_Tot/Enr_Tot)
CRDC_Corp_Prop
```

```
## # A tibble: 95,507 x 8
##    Enr_Tot WDIS_IDEA_504_Tot Corp_Tot WDIS_IDEA_504_Prop WDIS_Corp_Tot WDIS_Corp_Prop
##      <int>             <int>    <int>              <dbl>         <int>          <dbl>
##  1    1798                NA       NA           NA                  NA             NA
##  2     994                NA       NA           NA                  NA             NA
##  3     910                NA       NA           NA                  NA             NA
##  4     635                 6       NA            0.00945            NA             NA
##  5    1114                14       NA            0.0126             NA             NA
##  6     680                 5       NA            0.00735            NA             NA
##  7     783                 4       NA            0.00511            NA             NA
##  8     479                 2       NA            0.00418            NA             NA
##  9    1032                 4       NA            0.00388            NA             NA
```

```
## 10     423                 4        0            0.00946             0              0
## # ... with 95,497 more rows, and 2 more variables: WDIS_REF_Tot <int>, WDIS_REF_Prop
## #   <dbl>
```

```
## Filtering 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).

Filtered_CRDC_Corp_Prop <- CRDC_Corp_Prop %>%
  filter(CRDC_Corp_Prop$WDIS_IDEA_504_Tot < CRDC_Corp_Prop$Enr_Tot)
Filtered_CRDC_Corp_Prop
```

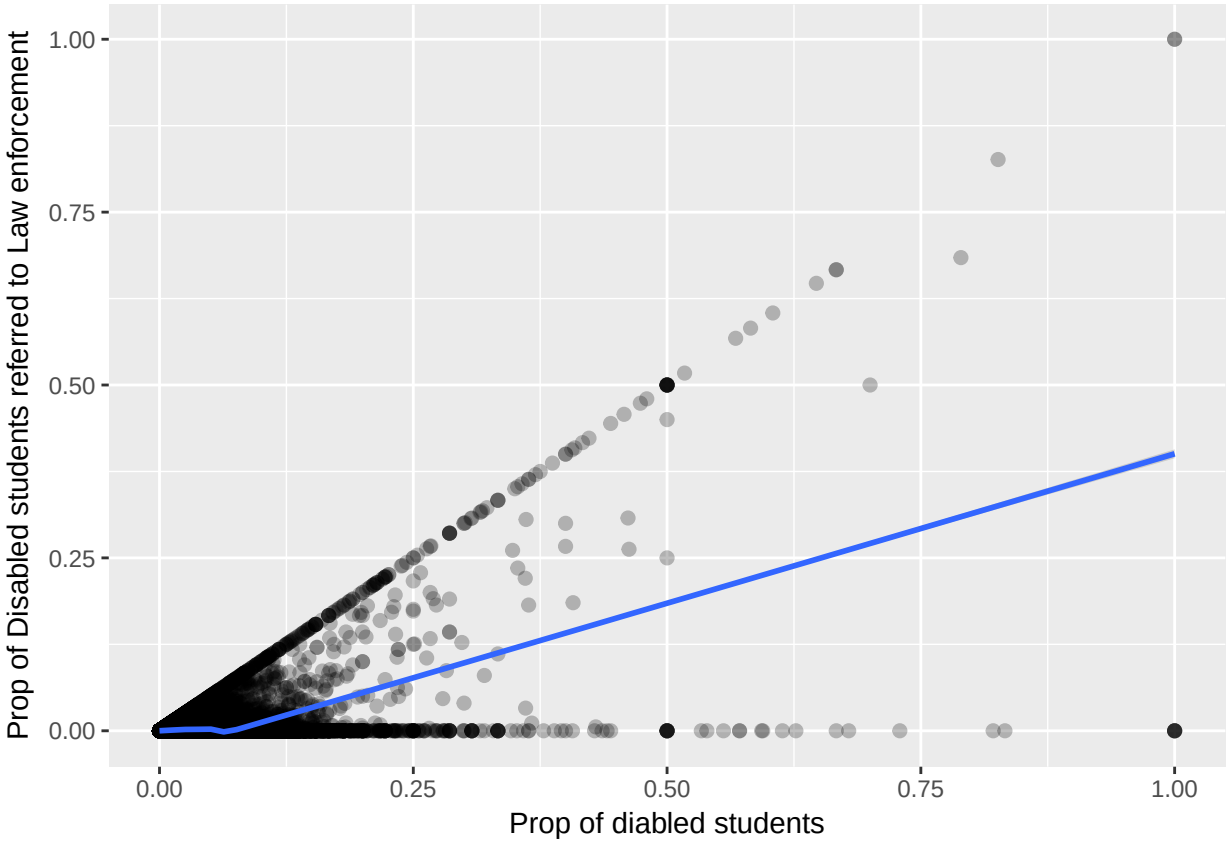
```
## # A tibble: 93,241 x 8
##    Enr_Tot WDIS_IDEA_504_Tot Corp_Tot WDIS_IDEA_504_Prop WDIS_Corp_Tot WDIS_Corp_Prop
##      <int>             <int>    <int>              <dbl>         <int>          <dbl>
##  1     635                 6       NA            0.00945            NA         NA
##  2    1114                14       NA            0.0126             NA         NA
##  3     680                 5       NA            0.00735            NA         NA
##  4     783                 4       NA            0.00511            NA         NA
##  5     479                 2       NA            0.00418            NA         NA
##  6    1032                 4       NA            0.00388            NA         NA
##  7     423                 4        0            0.00946             0          0
##  8     529                 0       22            0                   0          0
##  9     135                 2       NA            0.0148             NA         NA
## 10     478                 7       35            0.0146             10          0.0209
## # ... with 93,231 more rows, and 2 more variables: WDIS_REF_Tot <int>, WDIS_REF_Prop
## #   <dbl>
```

```
## Plotting 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.

ggplot(CRDC_Corp_Prop,
       mapping = aes(x=WDIS_IDEA_504_Prop,
       y=WDIS_REF_Prop)) + geom_point(alpha=1/4,
       size=2) + geom_smooth() + labs(x="Prop of diabled students",
                                      y="Prop of Disabled students referred to Law enforcement")
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 2258 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2258 rows containing missing values (geom_point).
```



The plot indicates an under-representation of disabled students among students who are referred to law enforcement.

```
## Calculating the overall proportion of disabled students across all schools.

summarise(CRDC_Corp_Prop, Overall_Prop_WDIS= mean(WDIS_IDEA_504_Tot>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_WDIS
##               <dbl>
## 1             0.713
```

```
## Calculating the overall proportion of students referred to
# -law enforcement who are disabled across all schools.

summarise(CRDC_Corp_Prop, Overall_Prop_REF= mean(WDIS_REF_Prop>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_REF
##              <dbl>
## 1            0.130
```

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.

We would like to investigate whether White students receive a proportionate number of suspensions from

school. Create a new data.frame or tibble with the following columns: . The total number of students enrolled at each school . The total number of White students enrolled at each school . The total number of students who received only one out of school suspension . The number of white students who received only one out of school suspension . The proportion of students at each school who are white . The proportion of students who received only one out of school suspension who are white

Plot the proportion of White students at each school (on the x-axis) versus the proportion of students who received only one out of school suspension who are white (on the y-axis). Include a smooth line on the plot.

Calculate the overall proportion of white students across all schools and the overall proportion of students who received only one out of school suspension who are white across all schools.

```
#A tibble containing required columns.

CRDC_WH_Prop <- transmute(C_Data,
                        Enr_tot = TOT_ENR_M + TOT_ENR_F,
                        Enr_White_tot = SCH_ENR_WH_M + SCH_ENR_WH_F,
                        SINGOOS_Tot = TOT_DISCWODIS_SINGOOS_M + TOT_DISCWODIS_SINGOOS_F
                        + TOT_DISCWDIS_SINGOOS_IDEA_M + TOT_DISCWDIS_SINGOOS_IDEA_F,
                        SINGOOS_Tot_WH = SCH_DISCWODIS_SINGOOS_WH_M + SCH_DISCWODIS_SINGOOS_WH_F
                        + SCH_DISCWDIS_SINGOOS_IDEA_WH_M + SCH_DISCWDIS_SINGOOS_IDEA_WH_F,
                        White_Prop=Enr_White_tot/Enr_tot,
                        White_SINGOOS_Prop = SINGOOS_Tot_WH/Enr_tot)
CRDC_WH_Prop
```

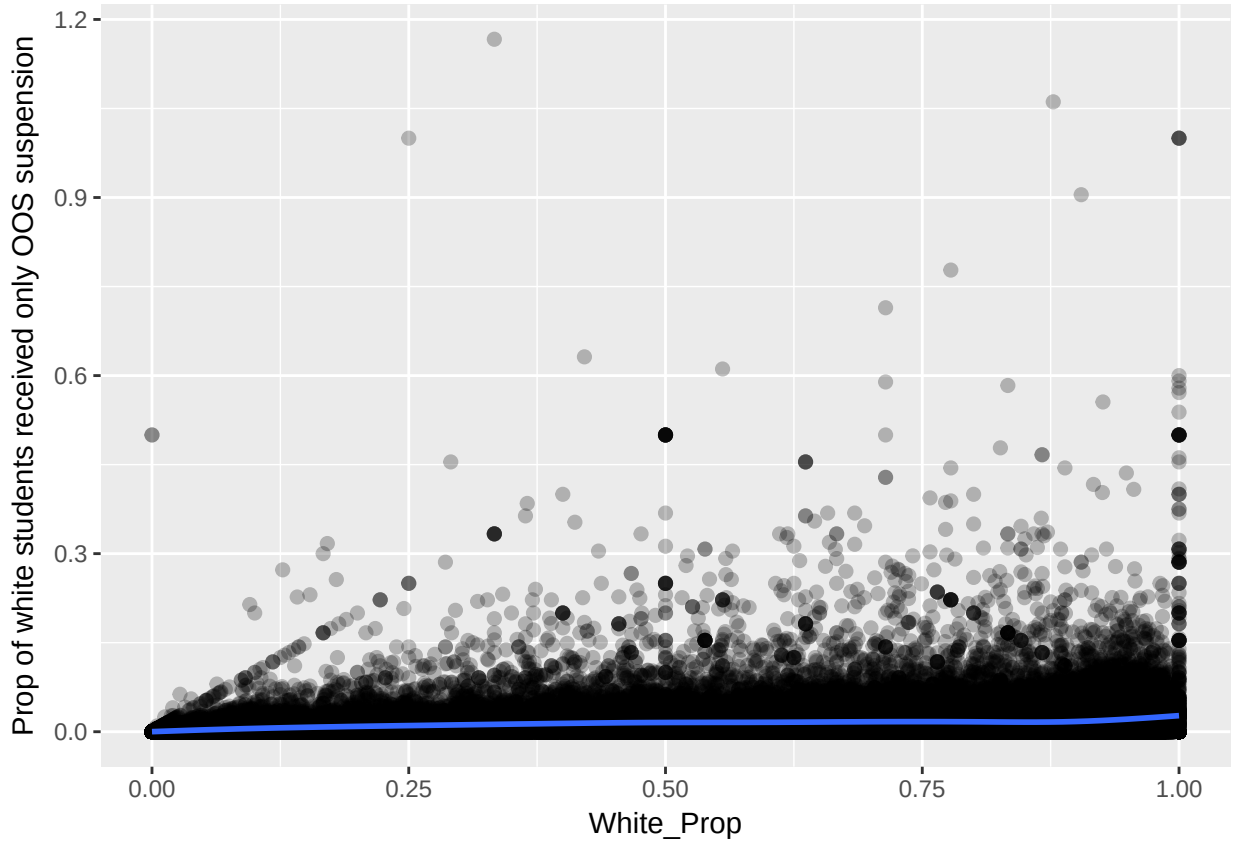
```
## # A tibble: 95,507 x 6
##    Enr_tot Enr_White_tot SINGOOS_Tot SINGOOS_Tot_WH White_Prop White_SINGOOS_Prop
##      <int>         <int>       <int>          <int>      <dbl>              <dbl>
##  1    1798           797           0              0      0.443            0
##  2     994           395           0              0      0.397            0
##  3     910           410           0              0      0.451            0
##  4     635           370          20             12      0.583            0.0189
##  5    1114           769          32             18      0.690            0.0162
##  6     680           385          13              6      0.566            0.00882
##  7     783           421          10              4      0.538            0.00511
##  8     479           217           0              0      0.453            0
##  9    1032           457           4              2      0.443            0.00194
## 10     423           397           0              0      0.939            0
## # ... with 95,497 more rows
```

```
## Plotting the proportion of White students at each school (on the x-axis)
#- versus the proportion of students who received only one out of school
# -suspension who are white (on the y-axis).

ggplot(CRDC_WH_Prop,
       mapping = aes(x=White_Prop,
       y=White_SINGOOS_Prop)) + geom_point(alpha=1/4,
       size=2) + geom_smooth() + labs(y="Prop of white students received only OOS suspension")
```

```
## `geom_smooth()` using method = 'gam'
```

```
## Warning: Removed 1762 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1762 rows containing missing values (geom_point).
```



The proportion of students who received only one out of school suspension who are white are not too high and they are under represented in the plot.

```
##Calculating the overall proportion of White students across all schools.

summarise(CRDC_WH_Prop, Overall_Prop_WH= mean(Enr_White_tot>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_WH
##             <dbl>
## 1           0.980
```

```
##Calculating the overall proportion of students who receieved only
# -one out of school suspension who are white across all schools.

summarise(CRDC_WH_Prop, Overall_Prop_WH_Tol= mean(SINGOOS_Tot_WH>0, na.rm = TRUE))
```

```
## # A tibble: 1 x 1
##   Overall_Prop_WH_Tol
##                 <dbl>
## 1               0.611
```