hw2-Deep-Bhalodia

Deep Bhalodia

1/27/2019

Loading required packages

```
library(ggplot2)
library(dplyr)
library(rlang)
library(readr)
library(forcats)
library(measurements)
```

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Part A

```
water_raw<-read_csv("NavajoWaterExport.csv")
```

```
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`,
  radium228=`Amount of Radium228`)

water
```

```
## # A tibble: 225 x 7
##    section  name               date   long      lat       risk    radium228
##    <chr>    <chr>              <chr>  <chr>     <chr>     <chr>       <dbl>
##  1 Section~ Gold Spring        1/19/~ 111 4 28~ 35 46 4.~ Some R~     0.5
##  2 Section~ Tank 3K-331        7/27/~ 111 24 2~ 35 46 8.~ Some R~     1.54
##  3 Section~ Lower Greasewood ~ 4/14/~ 109 51 1~ 35 31 42~ Less R~     0.591
##  4 Section~ Tank 8T-549        10/9/~ 110 12 4~ 36 39 41~ Some R~     0.183
##  5 Section~ Cedar Spring       7/13/~ 110 21 5~ 35 27 4.~ Less R~     0.439
##  6 Section~ Tank 8AI-1         9/21/~ 110 18 3~ 37 1 17.~ Less R~     0.892
##  7 Section~ Coyote Spring      7/8/98 110 27 5~ 35 20 37~ Some R~     0.565
##  8 Section~ 9T-523             3/18/~ 109 10 5~ 36 55 24~ Less R~     0.065
##  9 Section~ Chimney Butte Spr~ 7/14/~ 110 25 2~ 35 19 17~ Some R~     0.353
## 10 Section~ Nazlini Chapter H~ 11/17~ 109 26 4~ 35 53 56~ Some R~     0.975
```

```
## # ... with 215 more rows
```

Problem 1

Mutate the dataset to replace negative values of Radium-228 with 0

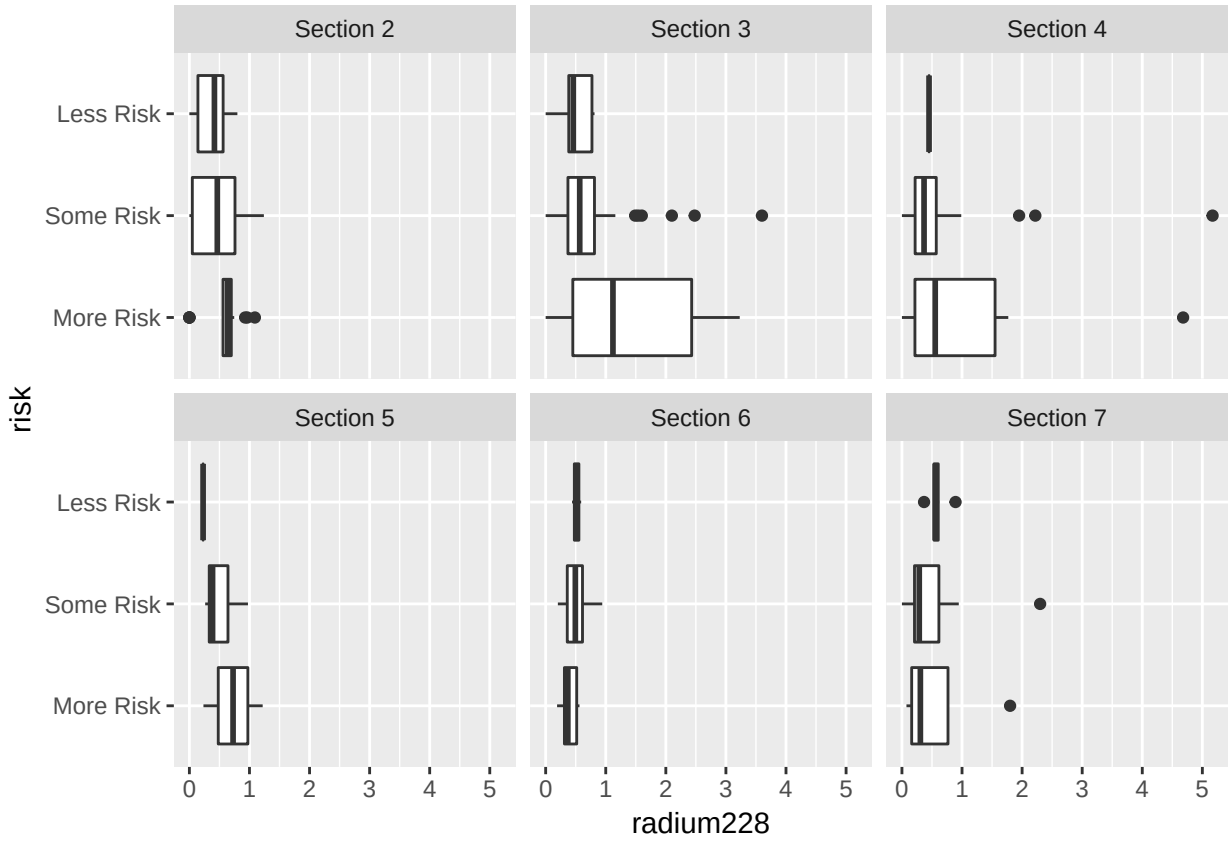
Filter the mutated dataset to remove any sites with "Unknown Risk" in EPA risk rating

```
water2 <- mutate(water, radium228=ifelse(radium228 < 0, 0, radium228)) %>%
  filter(risk != "Unknown Risk") %>%
  mutate(risk=fct_relevel(risk, "More Risk", "Some Risk", "Less Risk"))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

Visualize the distribution of Radium-228 with each EPA section and each risk level

```
ggplot(water2) +
  geom_boxplot(aes(x=risk, y=radium228)) +
  facet_wrap(~section) +
  coord_flip()
```



Observations

We choose to use side-by-side boxplots and faceting to visualize the distribution of Radium-228. We could also use histograms, but they may be somewhat more difficult to interpret in this case.

We notice that Section 3 and 4 tend to have the most sites with high concentrations of Radium-228. In general sites with higher EPA Risk are associated with higher levels of Radium-228, but this does not always seem to be the case, suggesting other radioactive isotopes contribute to EPA Risk as well. ### Problem 2

Get data for drawing the "Four Corners" region of the United States

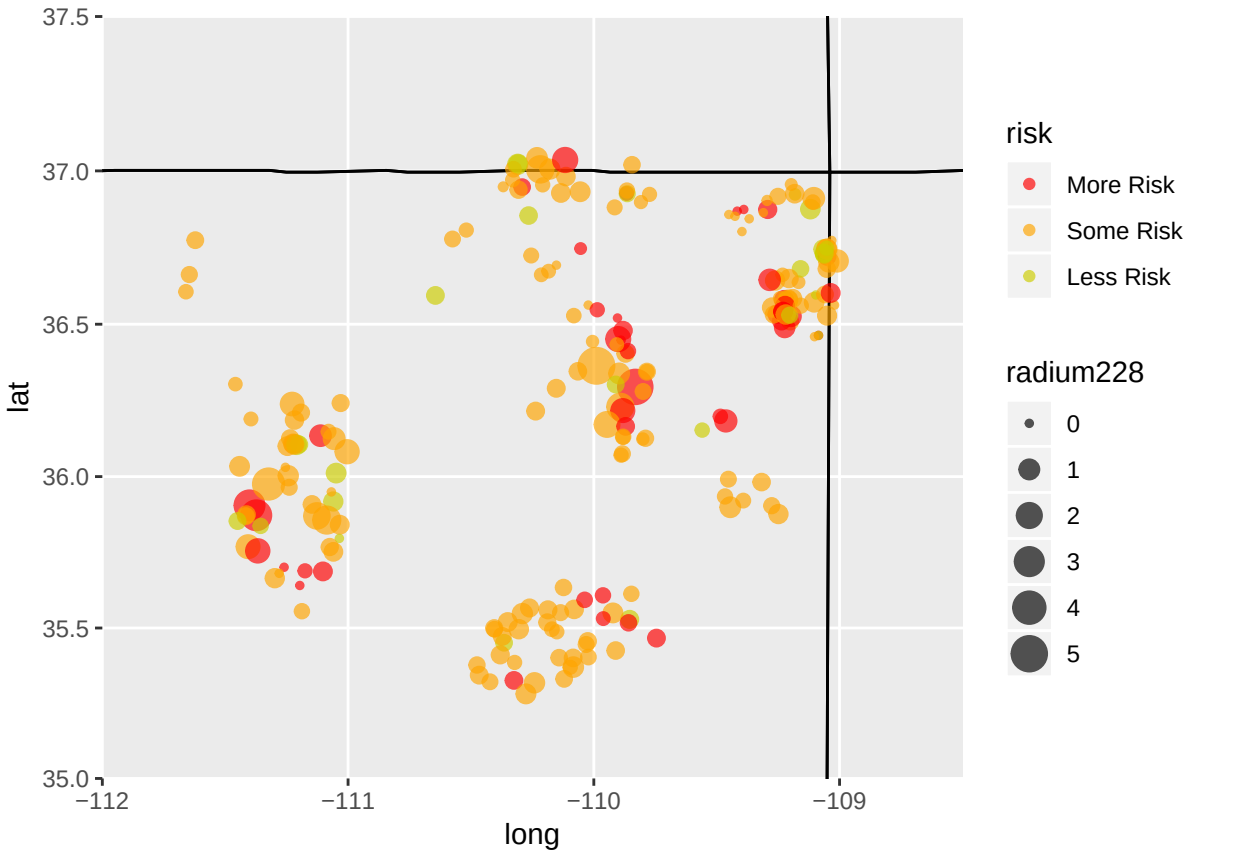
```
water_map <- water2 %>%mutate(long=-as.numeric(conv_unit(long,from="deg_min_sec",

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

```
## Warning: package 'maps' was built under R version 3.4.4
```

Create a map of the region showing the locations of the water sampling sites, along with the EPA risk and the concentration of Radium-228 for each location

```
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=risk,size=radium228),alpha=2/3) +
  scale_color_manual(values=c("red", "orange", "yellow3")) +
  coord_map(xlim=c(-112, -108.5), ylim=c(35, 37.5))
```



Part B

```
crdc <- read_csv("/Users/deep/downloads/CRDC 2015-16 School Data.csv",
na=c("-2", "-5", "-6", "-7", "-8", "-9"),guess=15000)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   LEA_STATE = col_character(),
##   LEA_STATE_NAME = col_character(),
##   LEAID = col_character(),
##   LEA_NAME = col_character(),
##   SCH_NAME = col_character(),
##   COMBOKEY = col_character(),
##   JJ = 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(),
##   SCH_GRADE_G08 = col_character(),
##   SCH_GRADE_G09 = col_character(),
##   SCH_GRADE_G10 = col_character(),
##   SCH_GRADE_G11 = col_character()
##   # ... with 45 more columns
## )
```

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

```
crdc
```

```
## # A tibble: 96,360 x 1,836
##    LEA_STATE LEA_STATE_NAME LEAID LEA_NAME SCHID SCH_NAME COMBOKEY JJ
##    <chr>     <chr>          <chr> <chr>    <dbl> <chr>    <chr>    <chr>
##  1 AL        ALABAMA        1000~ Alabama~  1705 Wallace~ 1000020~ Yes
##  2 AL        ALABAMA        1000~ Alabama~  1706 McNeel ~ 1000020~ Yes
##  3 AL        ALABAMA        1000~ Alabama~  1876 Alabama~ 1000020~ No
##  4 AL        ALABAMA        1000~ Alabama~ 99995 AUTAUGA~ 1000029~ Yes
##  5 AL        ALABAMA        1000~ Albertv~   870 Albertv~ 1000050~ No
##  6 AL        ALABAMA        1000~ Albertv~   871 Albertv~ 1000050~ No
##  7 AL        ALABAMA        1000~ Albertv~   879 Evans E~ 1000050~ No
##  8 AL        ALABAMA        1000~ Albertv~   889 Albertv~ 1000050~ No
##  9 AL        ALABAMA        1000~ Albertv~  1616 Big Spr~ 1000050~ No
## 10 AL        ALABAMA        1000~ Albertv~  2150 Albertv~ 1000050~ No
## # ... with 96,350 more rows, and 1,828 more variables: 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 <dbl>,
## #   SCH_PSENR_HI_F <dbl>, SCH_PSENR_AM_M <dbl>, SCH_PSENR_AM_F <dbl>,
## #   SCH_PSENR_AS_M <dbl>, SCH_PSENR_AS_F <dbl>, SCH_PSENR_HP_M <dbl>,
## #   SCH_PSENR_HP_F <dbl>, SCH_PSENR_BL_M <dbl>, SCH_PSENR_BL_F <dbl>,
## #   SCH_PSENR_WH_M <dbl>, SCH_PSENR_WH_F <dbl>, SCH_PSENR_TR_M <dbl>,
## #   SCH_PSENR_TR_F <dbl>, TOT_PSENR_M <dbl>, TOT_PSENR_F <dbl>,
## #   SCH_PSENR_LEP_M <dbl>, SCH_PSENR_LEP_F <dbl>, SCH_PSENR_IDEA_M <dbl>,
## #   SCH_PSENR_IDEA_F <dbl>, SCH_ENR_HI_M <dbl>, SCH_ENR_HI_F <dbl>,
## #   SCH_ENR_AM_M <dbl>, SCH_ENR_AM_F <dbl>, SCH_ENR_AS_M <dbl>,
## #   SCH_ENR_AS_F <dbl>, SCH_ENR_HP_M <dbl>, SCH_ENR_HP_F <dbl>,
## #   SCH_ENR_BL_M <dbl>, SCH_ENR_BL_F <dbl>, SCH_ENR_WH_M <dbl>,
## #   SCH_ENR_WH_F <dbl>, SCH_ENR_TR_M <dbl>, SCH_ENR_TR_F <dbl>,
## #   TOT_ENR_M <dbl>, TOT_ENR_F <dbl>, SCH_ENR_LEP_M <dbl>,
## #   SCH_ENR_LEP_F <dbl>, SCH_ENR_504_M <dbl>, SCH_ENR_504_F <dbl>,
## #   SCH_ENR_IDEA_M <dbl>, SCH_ENR_IDEA_F <dbl>, SCH_LEPENR_HI_M <dbl>,
## #   SCH_LEPENR_HI_F <dbl>, SCH_LEPENR_AM_M <dbl>, SCH_LEPENR_AM_F <dbl>,
## #   SCH_LEPENR_AS_M <dbl>, SCH_LEPENR_AS_F <dbl>, SCH_LEPENR_HP_M <dbl>,
## #   SCH_LEPENR_HP_F <dbl>, SCH_LEPENR_BL_M <dbl>, SCH_LEPENR_BL_F <dbl>,
## #   SCH_LEPENR_WH_M <dbl>, SCH_LEPENR_WH_F <dbl>, SCH_LEPENR_TR_M <dbl>,
## #   SCH_LEPENR_TR_F <dbl>, TOT_LEPENR_M <dbl>, TOT_LEPENR_F <dbl>,
## #   SCH_LEPPROGENR_HI_M <dbl>, SCH_LEPPROGENR_HI_F <dbl>,
## #   SCH_LEPPROGENR_AM_M <dbl>, SCH_LEPPROGENR_AM_F <dbl>,
## #   SCH_LEPPROGENR_AS_M <dbl>, SCH_LEPPROGENR_AS_F <dbl>,
## #   SCH_LEPPROGENR_HP_M <dbl>, SCH_LEPPROGENR_HP_F <dbl>,
## #   SCH_LEPPROGENR_BL_M <dbl>, SCH_LEPPROGENR_BL_F <dbl>,
## #   SCH_LEPPROGENR_WH_M <dbl>, SCH_LEPPROGENR_WH_F <dbl>,
## #   SCH_LEPPROGENR_TR_M <dbl>, SCH_LEPPROGENR_TR_F <dbl>,
## #   TOT_LEPPROGENR_M <dbl>, ...
```

Problem 3

Create new dataframe with following columns

1. The total number of students enrolled at each school 2.The number of Black students enrolled at each school 3.The total number of students who received one or more in-school suspension (including non-disabled students and disabled students served by IDEA) 4.The number of Black students who received one or more in-school suspension (including non-disabled students and disabled students served by IDEA) 5.The proportion of Black students at each school among all students 6.The proportion of students who received one or more in-school suspension who are Black among all suspended students

```
crdc_susp <- transmute(crdc,
                       enr_tot = TOT_ENR_M + TOT_ENR_F,
                       enr_bl = SCH_ENR_BL_M + SCH_ENR_BL_F,
                       susp_tot = TOT_DISCWODIS_ISS_M +
                       TOT_DISCWODIS_ISS_F +
                       TOT_DISCWDIS_ISS_IDEA_M +
                       TOT_DISCWDIS_ISS_IDEA_F,
                       susp_bl = SCH_DISCWODIS_ISS_BL_M +
                       SCH_DISCWODIS_ISS_BL_F +
                       SCH_DISCWDIS_ISS_IDEA_BL_M +
                       SCH_DISCWDIS_ISS_IDEA_BL_F,
```

```
                        pr_bl = enr_bl / enr_tot,
                        pr_susp_bl = susp_bl / susp_tot)
```

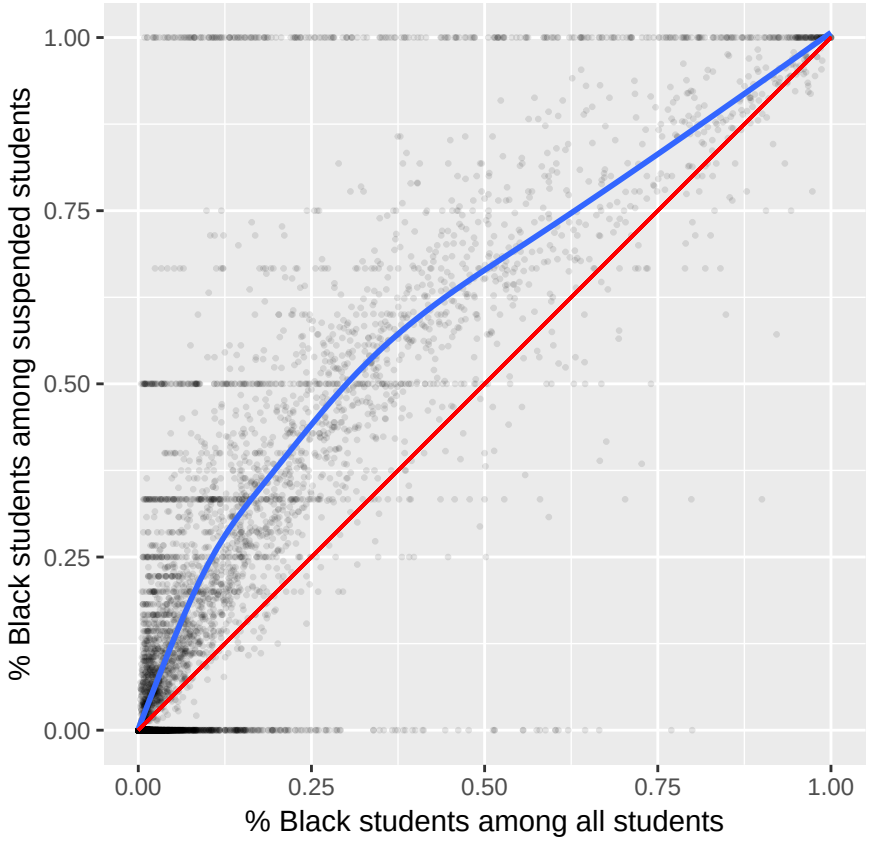
Plot the proportion of Black students at each school (on the x-axis) versus the proportion of suspended students who are Black (on the y-axis). Include a smoothing line on the plot.

```
crdc_susp %>%
  sample_n(10000) %>%
  ggplot(aes(x=pr_bl, y=pr_susp_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(x=c(0,1), y=c(0,1)) +
  labs(x='% Black students among all students', y='% Black students among suspended students')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 3668 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3668 rows containing missing values (geom_point).
```



Comments

Optionally, 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. If discipline is given fairly without regard to race, then the proportion of Black suspended students among all suspended students should be roughly the same as the proportion of black students in the whole student body, as shown by the reference line. But the former is actually typically greater than the latter, as shown by the fitted smooth line, indicating an over-representation of Black students among suspended students

Calculate the overall proportion of Black students across all schools and the overall proportion of suspended students who are Black across all schools.

```
summarise(crdc_susp,pr_bl=sum(enr_bl, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE), pr_susp_bl=sum(susp_bl, na
```

```
## # A tibble: 1 x 2
##   pr_bl pr_susp_bl
##   <dbl>      <dbl>
## 1 0.154      0.321
```

Are Black students over- or under-represented in in-school suspensions?

Black students tend to be over-represented among suspended students.

Problem 4

Create a new data.frame containing only schools that use corporal punishment with the following columns:

The total number of students enrolled at each school The number of disabled students (served by IDEA) at each school The total number of students who were disciplined with corporal punishment The number of disabled students (served by IDEA) who were disciplined with corporal punishment The proportion of disabled students (served by IDEA) at each school among of all students The proportion of students who were disciplined with corporal punishment who are disabled (served by IDEA) among all disciplined students

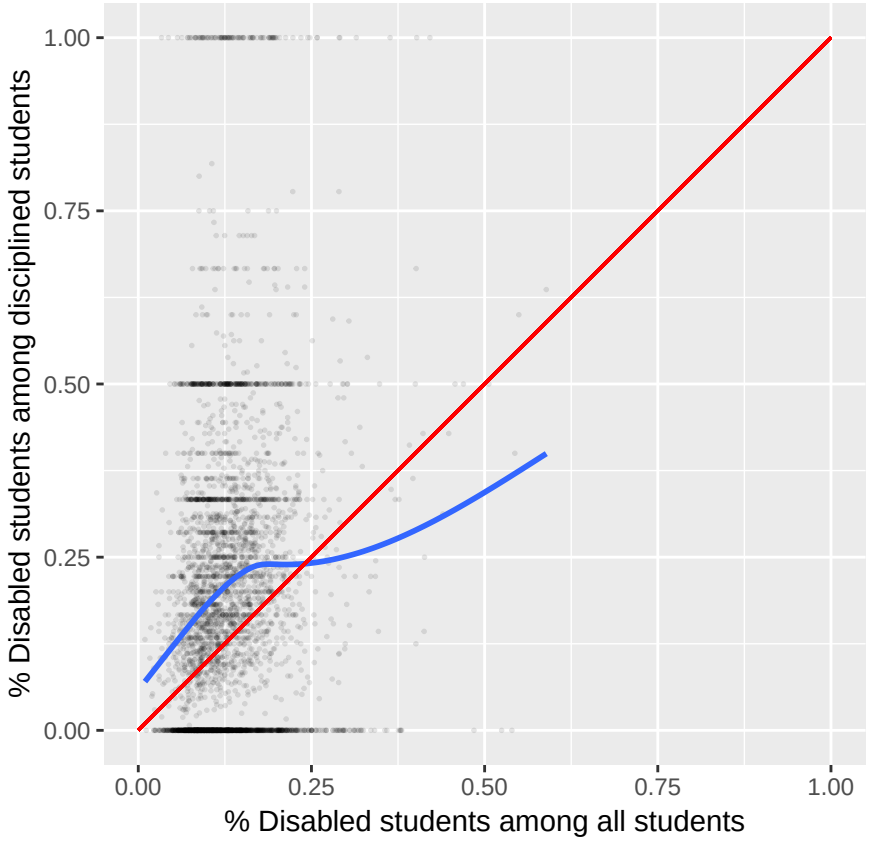
```
crdc_corp <- filter(crdc, SCH_CORPINSTANCES_IND=="Yes") %>%
  transmute(
  enr_tot = TOT_ENR_M + TOT_ENR_F,
  enr_dis = SCH_ENR_IDEA_M + SCH_ENR_IDEA_F,
  corp_dis = TOT_DISCWDIS_CORP_IDEA_M +
  TOT_DISCWDIS_CORP_IDEA_F,
  corp_tot = corp_dis +
  TOT_DISCWODIS_CORP_M +
  TOT_DISCWODIS_CORP_F,
  pr_dis=enr_dis / enr_tot,
  pr_corp_dis=corp_dis / corp_tot)
```

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

```
crdc_corp %>%ggplot(mapping=aes(x=pr_dis, y=pr_corp_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(x=c(0,1), y=c(0,1)) +
  labs(x='% Disabled students among all students', y='% Disabled students among disciplined students')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## Warning: Removed 912 rows containing non-finite values (stat_smooth).
## Warning: Removed 912 rows containing missing values (geom_point).
```



What do you observe in the plot? Does the plot indicate an over- or under-representation of disabled students among students who are disciplined with corporeal punishment?

The fitted smooth line suggests that, until the schools reach roughly 25% disabled students, disabled students are over-represented among students who are disciplined with corporal punishment. This is indicated by theproportion of disciplined students who are disabled typically being greater than the proportion of disabled students at the school for pr_dis < 0.25.

However, this relationship drops off as the proportion of disabled students at the school increases afer this point, suggesting under-representation for schools where pr_dis> 0.25. But the second claim should be taken with a grain of salt, as we have much less data points where pr_dis > 0.25.

Calculate the overall proportion of disabled students across all schools and the overall proportion of disciplined students who are disabled across all schools.

```
summarise(crdc_corp,pr_dis=sum(enr_dis, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
pr_corp_dis=sum(corp_dis, na.rm=TRUE) / sum(corp_tot, na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   pr_dis pr_corp_dis
##    <dbl>       <dbl>
## 1  0.123       0.174
```

Are disabled students over- or under-represented in corporal punishment?

Overall, disabled students appear to be over-represented among students disciplined with corporal punishment.

Problem 5

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

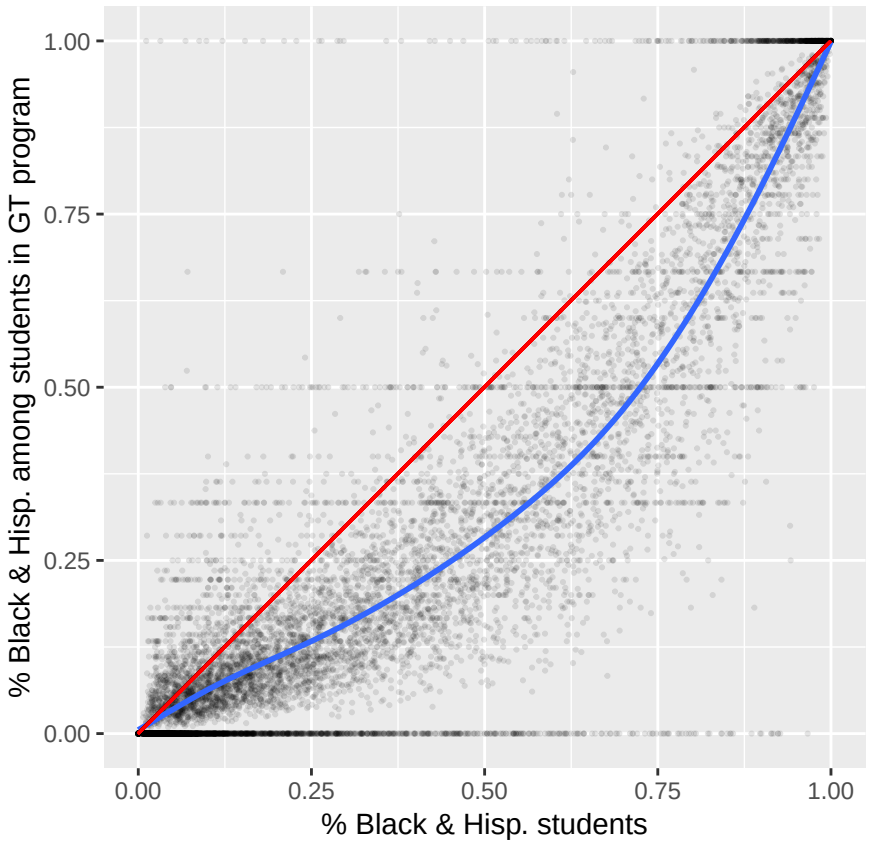
The total number of students enrolled at each school The number of Black and Hispanic students at each school The total number of students in the school's GT program The number of students in the GT program who are Black or Hispanic The proportion of students at each school who are Black or Hispanic among all students The proportion of students in the GT program who are Black or Hispanic among students in the GT program

```
crdc_gt <- filter(crdc, SCH_GT_IND=="Yes") %>%
  transmute(
  enr_tot = TOT_ENR_M + TOT_ENR_F,
  enr_hibl = SCH_ENR_HI_M + SCH_ENR_HI_F +
  SCH_ENR_BL_M + SCH_ENR_BL_F,
  pr_hibl = enr_hibl / enr_tot,
  gt_tot = TOT_GTENR_M + TOT_GTENR_F,
  gt_hibl = SCH_GTENR_HI_M + SCH_GTENR_HI_F +
  SCH_GTENR_BL_M + SCH_GTENR_BL_F,
  pr_gt_hibl = gt_hibl / gt_tot)
```

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

```
crdc_gt %>% sample_n(10000) %>%
  ggplot(aes(x=pr_hibl, y=pr_gt_hibl)) +
  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(x=c(0,1), y=c(0,1)) +
  labs(x='% Black & Hisp. students', y='% Black & Hisp. among students in GT program')
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



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

The fitted smooth lines shows that the proportion of Gifted & Talented students who are Black and 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.

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

```
summarise(crdc_gt,pr_hibl=sum(enr_hibl, na.rm=TRUE) / sum(enr_tot, na.rm=TRUE),
pr_gt_hibl=sum(gt_hibl, na.rm=TRUE) / sum(gt_tot, na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   pr_hibl pr_gt_hibl
##     <dbl>      <dbl>
## 1   0.421      0.268
```

Are Black and Hispanic students over- or under-represented in Gifted & Talented programs?

In general, Black and Hispanic students appear to be under-represented in Gifted & Talented programs.