hw1-Zhongheng Yang

To automatically wrap the lines in output PDF:

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=25),tidy=TRUE)
```

!!!ATTENTION 1 if something's wrong, please first input:install.packages(c("maps","tidyverse","nycflights13","measurements"))
2 please run by order of questions, due to some shared variabled and libraried 3. Since the homework instructions says no data in directory, inorder to input the data, you will need the CRDC2013_14_SCH.csv and NavajoWaterExport.csv in the working directory–which is this Rmd's directory prob1_1

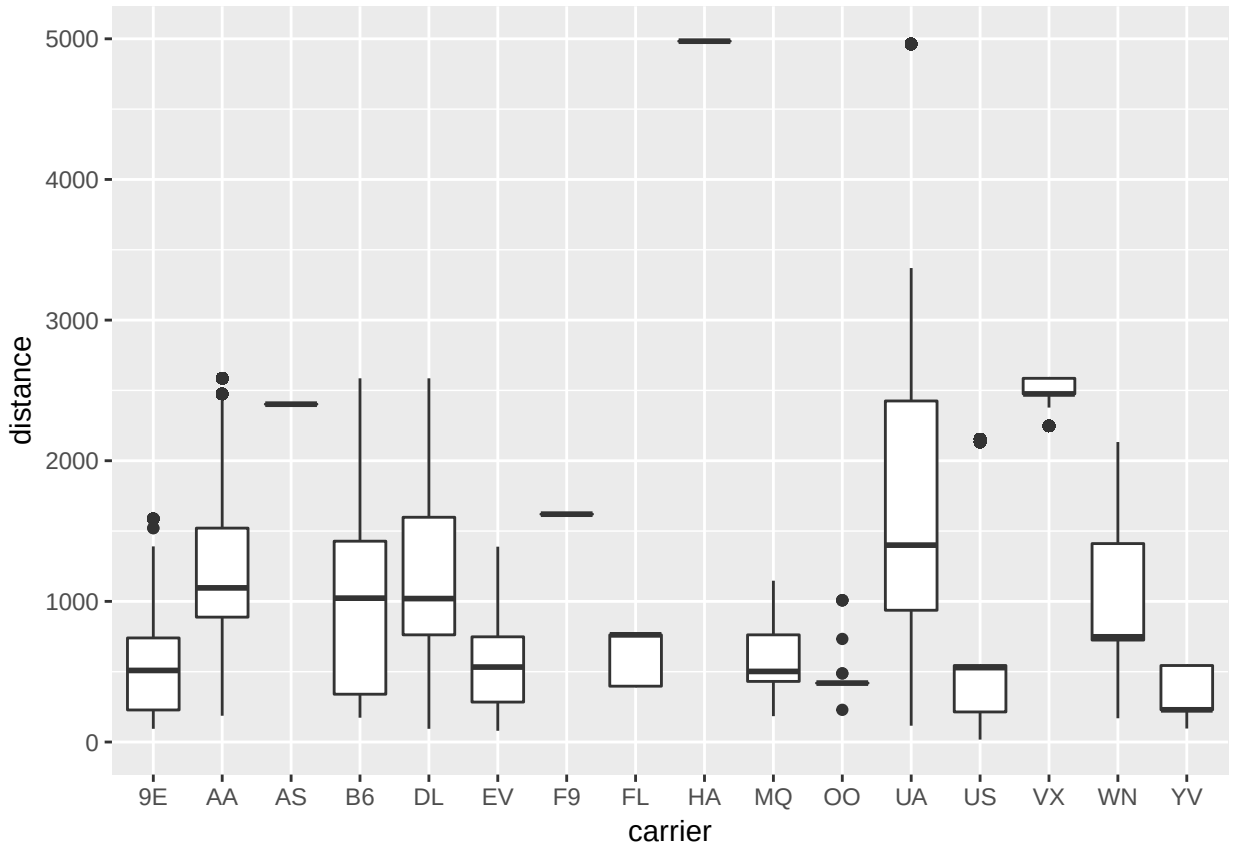
```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

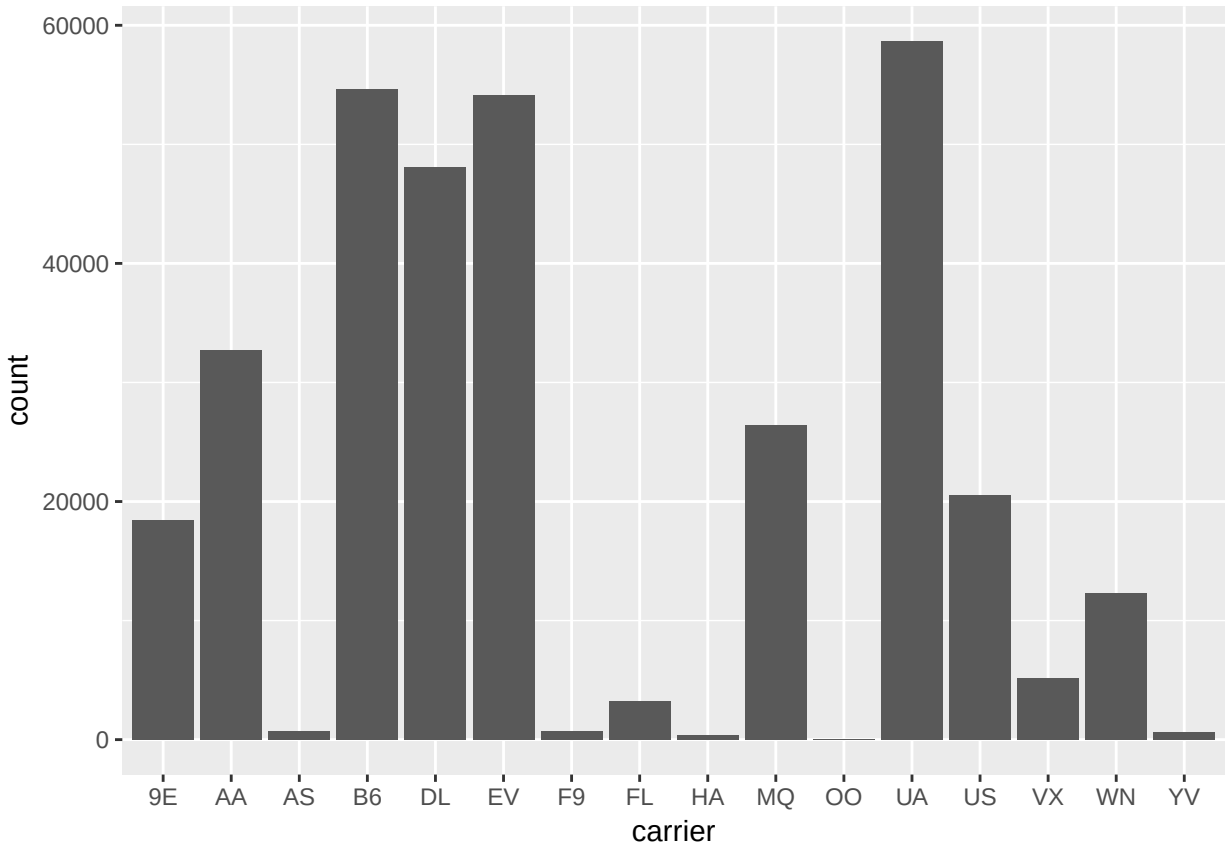
```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(nycflights13)
ggplot(data = nycflights13::flights,
    mapping = aes(y = distance,
        x = carrier)) + geom_boxplot()
```




```
ggplot(data = nycflights13::flights) +
    geom_bar(mapping = aes(x = carrier))
```



UA and VX typically fly longer routes than others, others like AS and F9 fly longer but have limited flighted compared to others, according to the bar plot

---

prob1_2

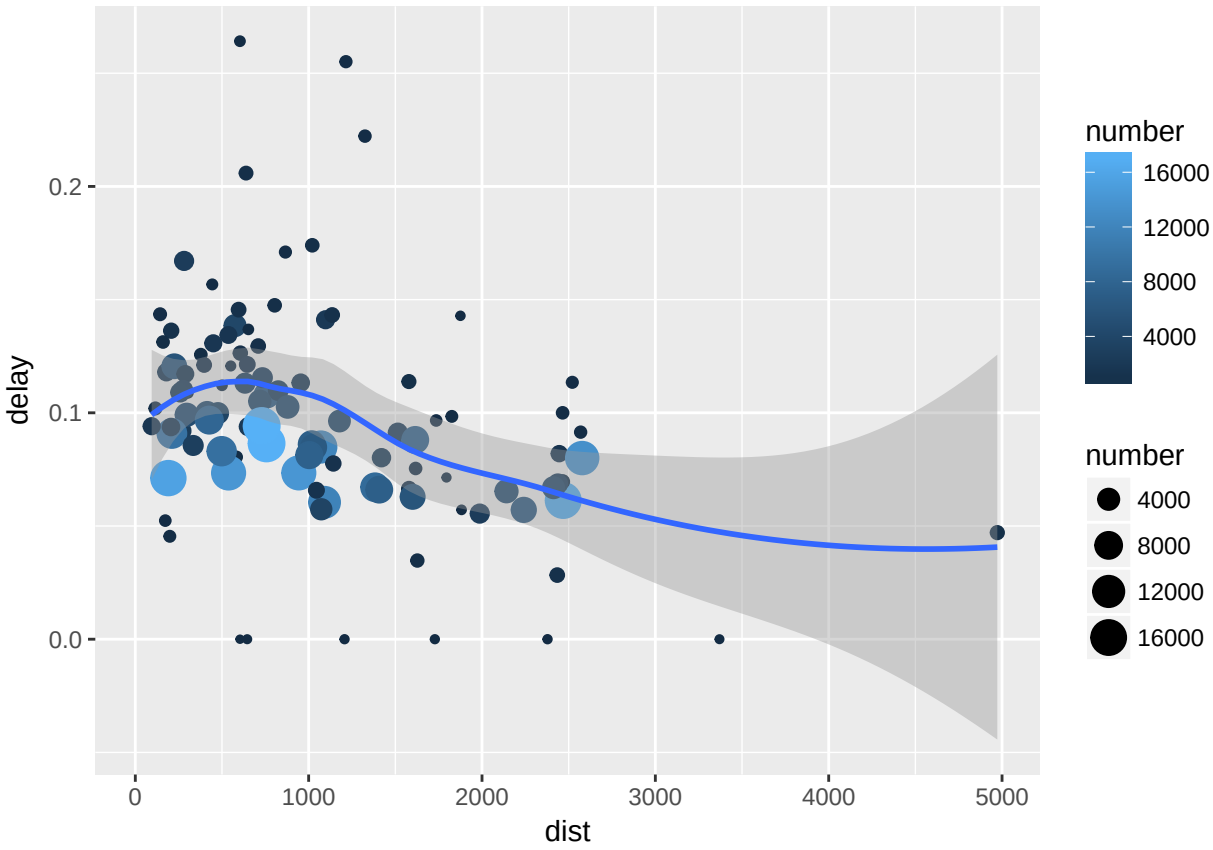
```
fly_tbl <- summarize(group_by(flights,
    dest), delay = mean(arr_delay >=
    60, na.rm = T), dist = mean(distance,
    na.rm = T), number = sum(!is.na(dest)))
ggplot(data = fly_tbl) + geom_point(aes(x = dist,
    y = delay, size = number,
    color = number, position = "jitter")) +
    geom_smooth(aes(x = dist,
        y = delay))
```

```
## Warning: Ignoring unknown aesthetics: position
```

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

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

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

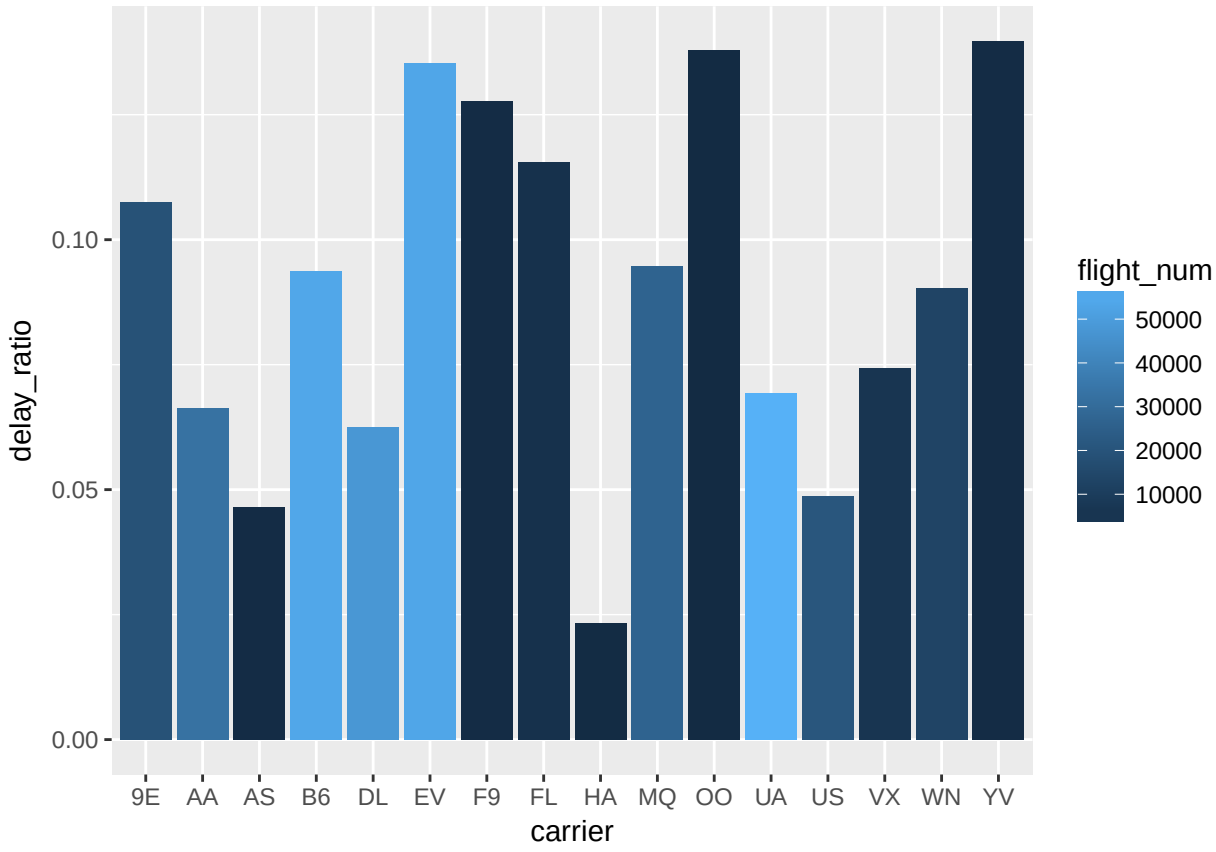


Conclusion:Initially in short distances, the longer they travel, the more they delay. But after a peak, in longer distances , maybe longer tours can decrease the delay due to the possible reason that there's more time to fly quicker to compensate the departure delay.

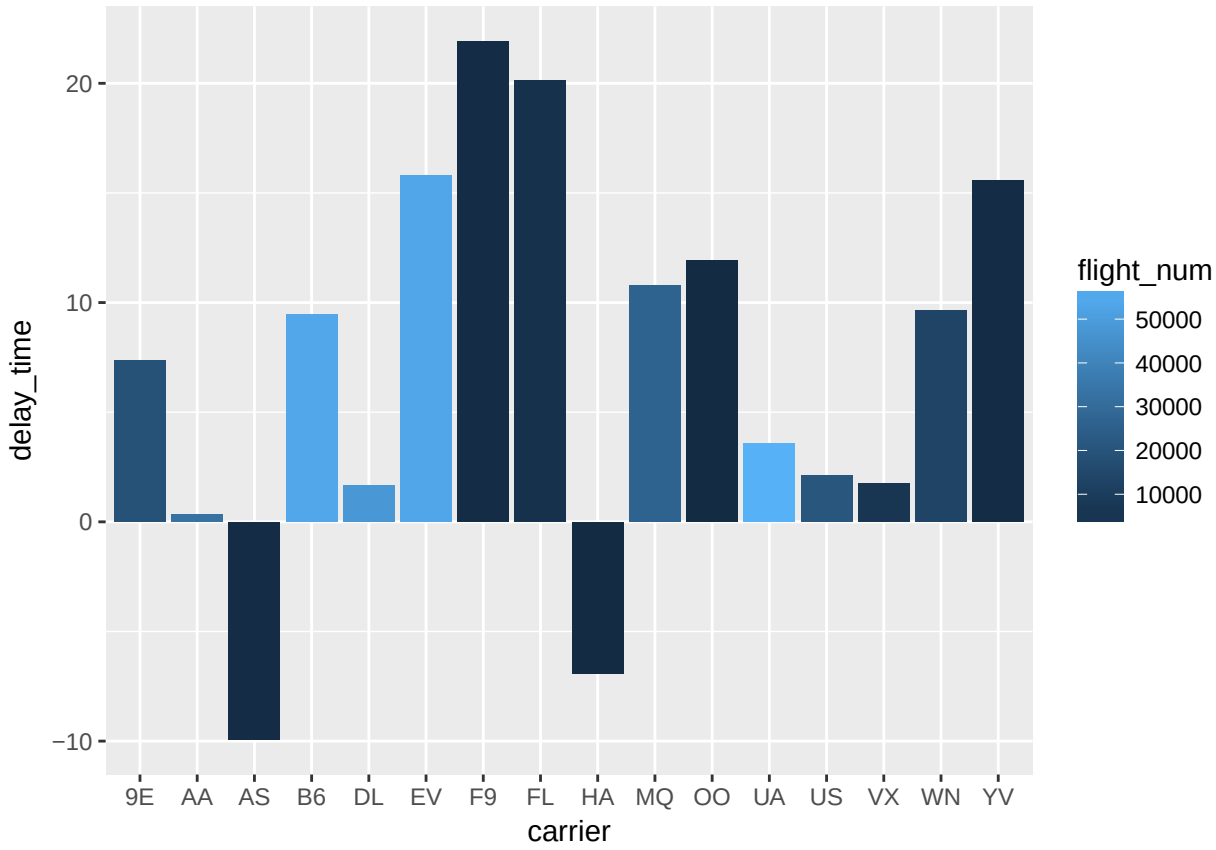
---

prob1_3

```
fly_tbl2 <- summarize(group_by(flights,
    carrier), delay_ratio = mean(arr_delay >=
    60, na.rm = T), delay_time = mean(arr_delay,
    na.rm = T), flight_num = sum(!is.na(carrier)))
ggplot(data = fly_tbl2) +
    geom_col(aes(x = carrier,
        y = delay_ratio, fill = flight_num))
```




```
ggplot(data = fly_tbl2) +
    geom_col(aes(x = carrier,
        y = delay_time, fill = flight_num))
```



Conclusion: YV have the worst delay. AS is the most consistently on time or ahead of schedule(also the filght number is filled as color in the plot for comparison)

prob1_4

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

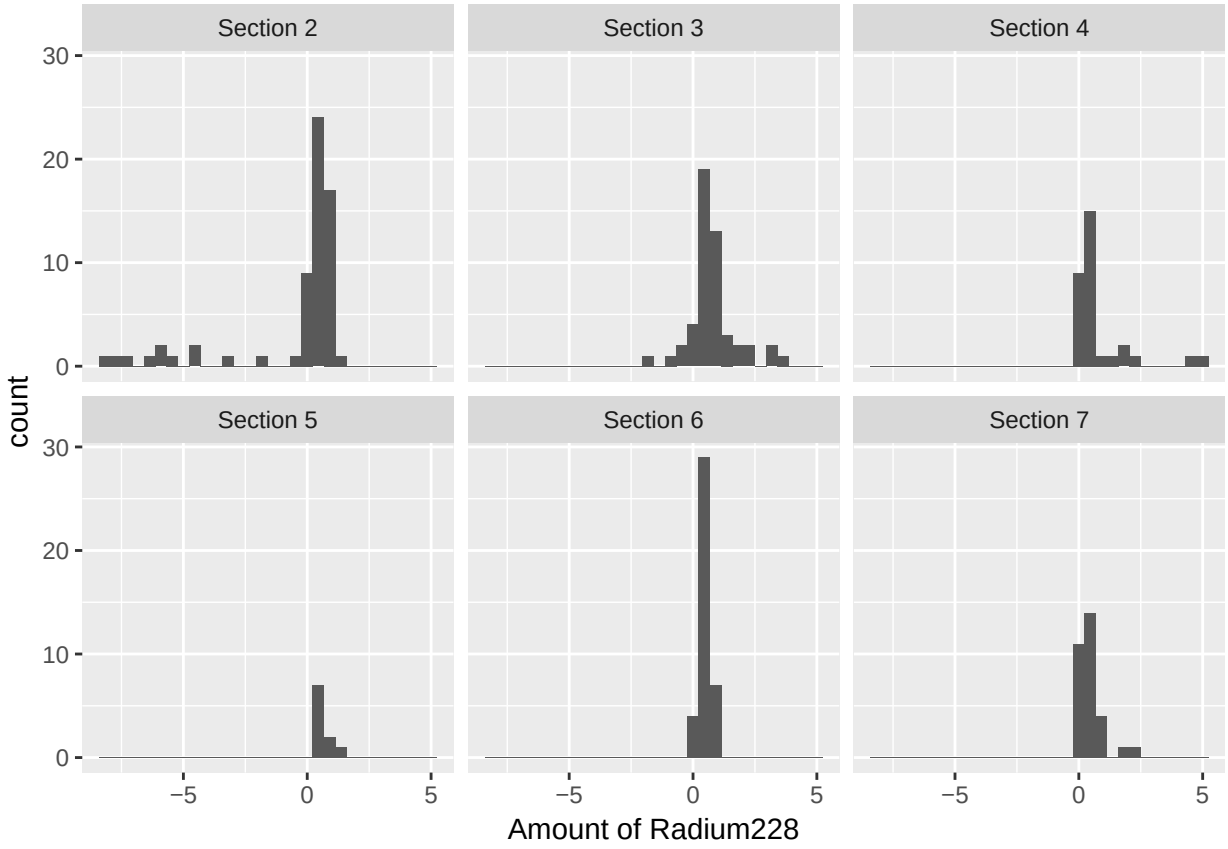
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Amount of Aluminum (Al)` = col_number(),
##   `Amount of Antimony (Sb)` = col_double(),
##   `Amount of Arsenic (As)` = col_double(),
##   `Amount of Barium (Ba)` = col_number(),
##   `Amount of Beryllium (Be)` = col_double(),
##   `Amount of Cadmium (Cd)` = col_double(),
##   `Amount of Chromium (Cr)` = col_double(),
##   `Amount of Copper (Cu)` = col_double(),
##   `Amount of Iron (Fe)` = col_number(),
##   `Amount of Lead (Pb)` = col_double(),
##   `Amount of Manganese (Mn)` = col_number(),
##   `Amount of Mercury (Hg)` = col_double(),
##   `Amount of Nickel (Ni)` = col_double(),
##   `Amount of Selenium (Se)` = col_double(),
##   `Amount of Silver (Ag)` = col_double(),
##   `Amount of Thallium (TI)` = col_double(),
##   `Amount of Vanadium (V)` = col_double(),
```

```
##   `Amount of Zinc (Zn)` = col_number(),
##   `Amount of Alpha Particles` = col_double(),
##   `Amount of Beta Particles` = col_double()
##   # ... with 9 more columns
## )
```

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

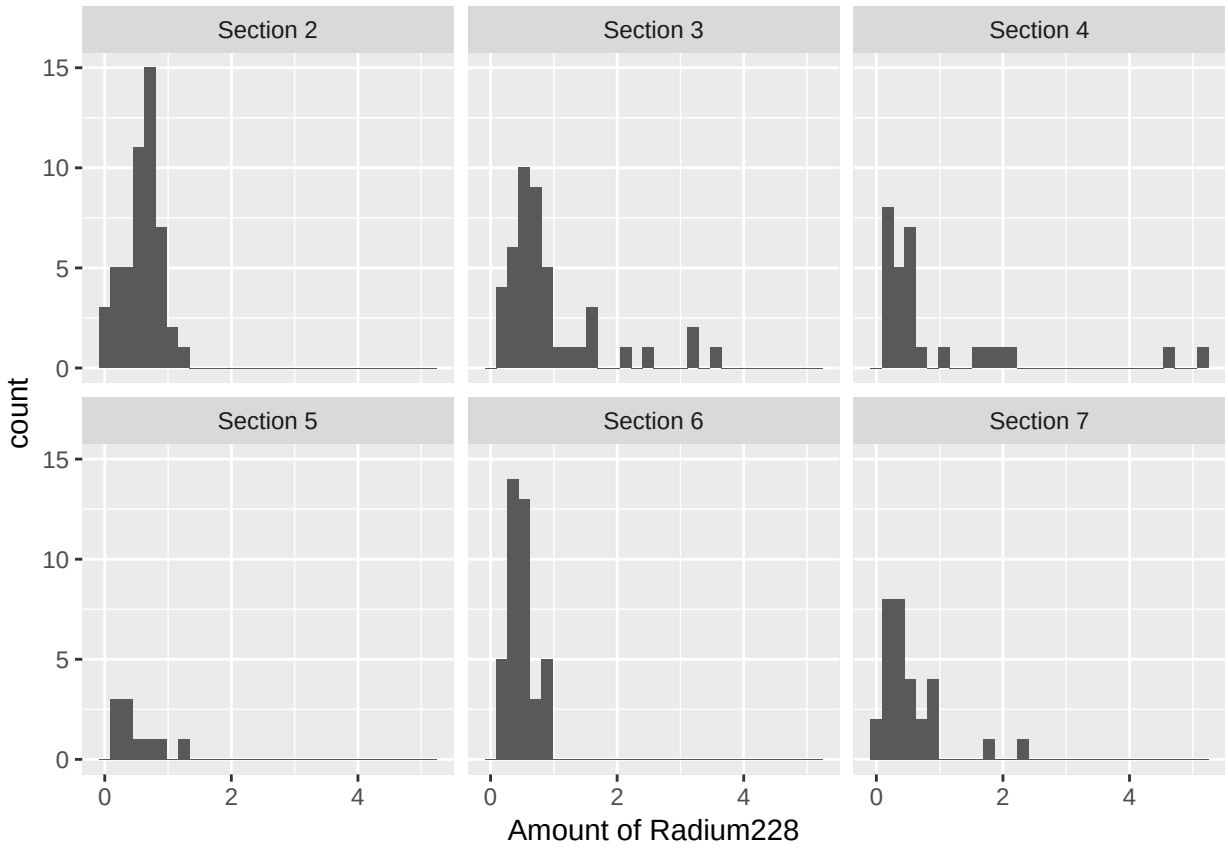
```
ggplot(data = report) + geom_histogram(aes(x = `Amount of Radium228`)) +
    facet_wrap(~`Which EPA Section is This From?`)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
no_NA <- filter(report, !is.na(`Amount of Radium228`))
no_zero <- filter(no_NA, `Amount of Radium228` >
    0)
ggplot(data = no_zero) + geom_bar(aes(x = `Amount of Radium228`),
    stat = "bin") + facet_wrap(~`Which EPA Section is This From?`)
```

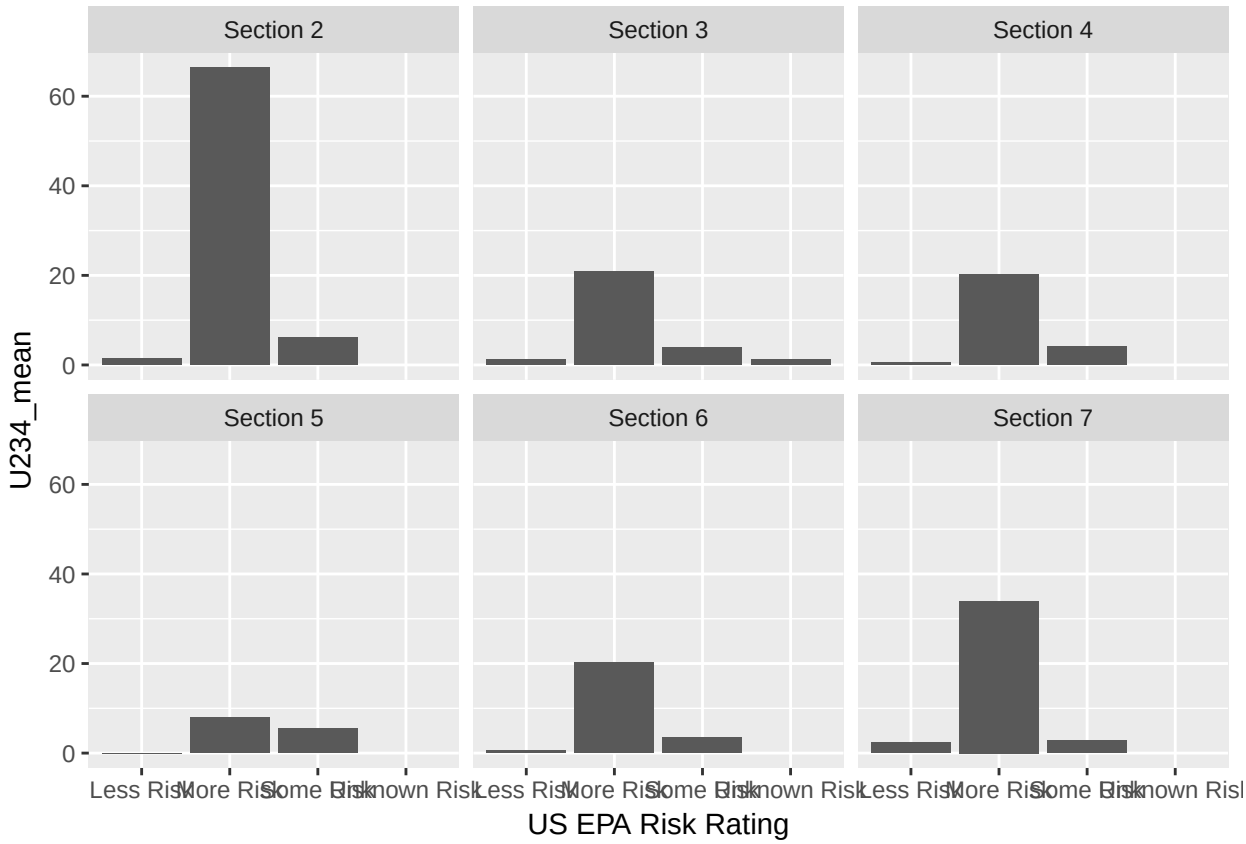
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



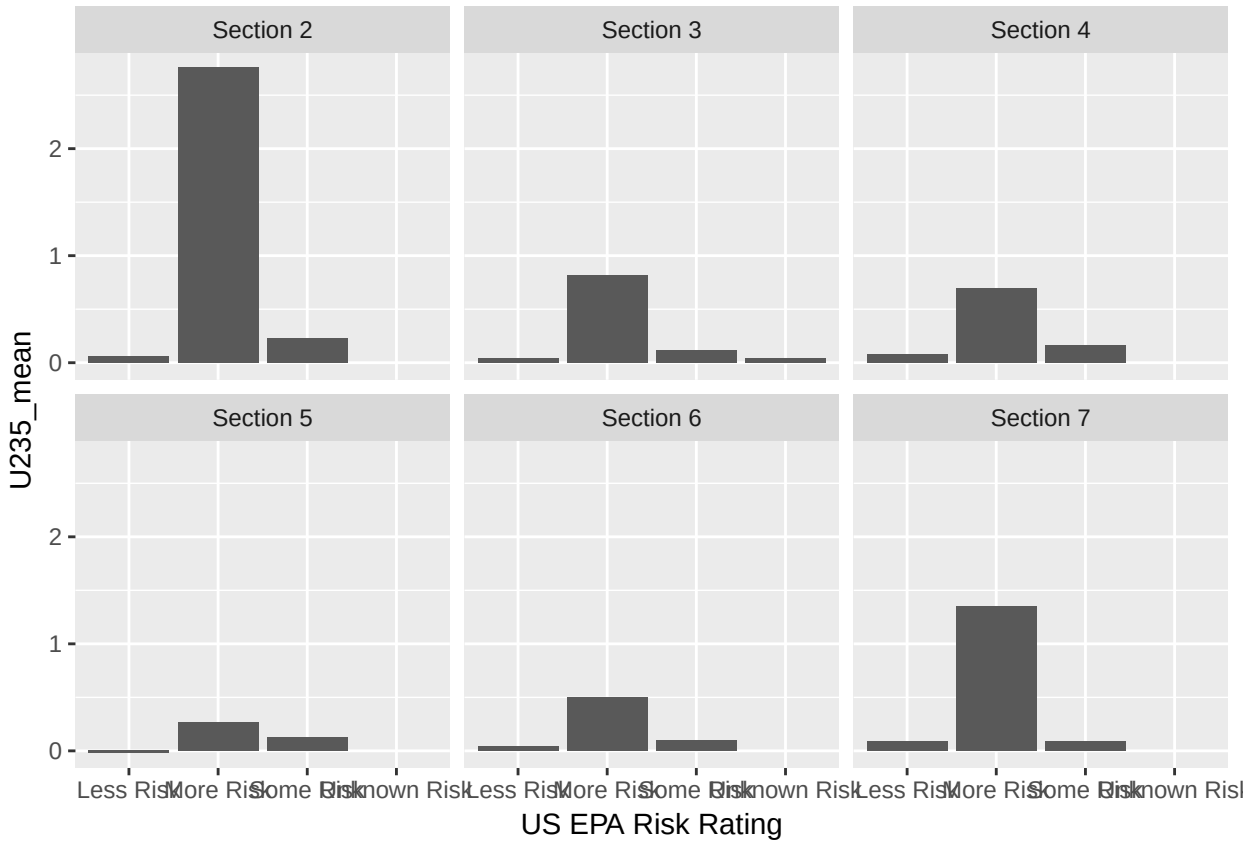
---

prob1_5

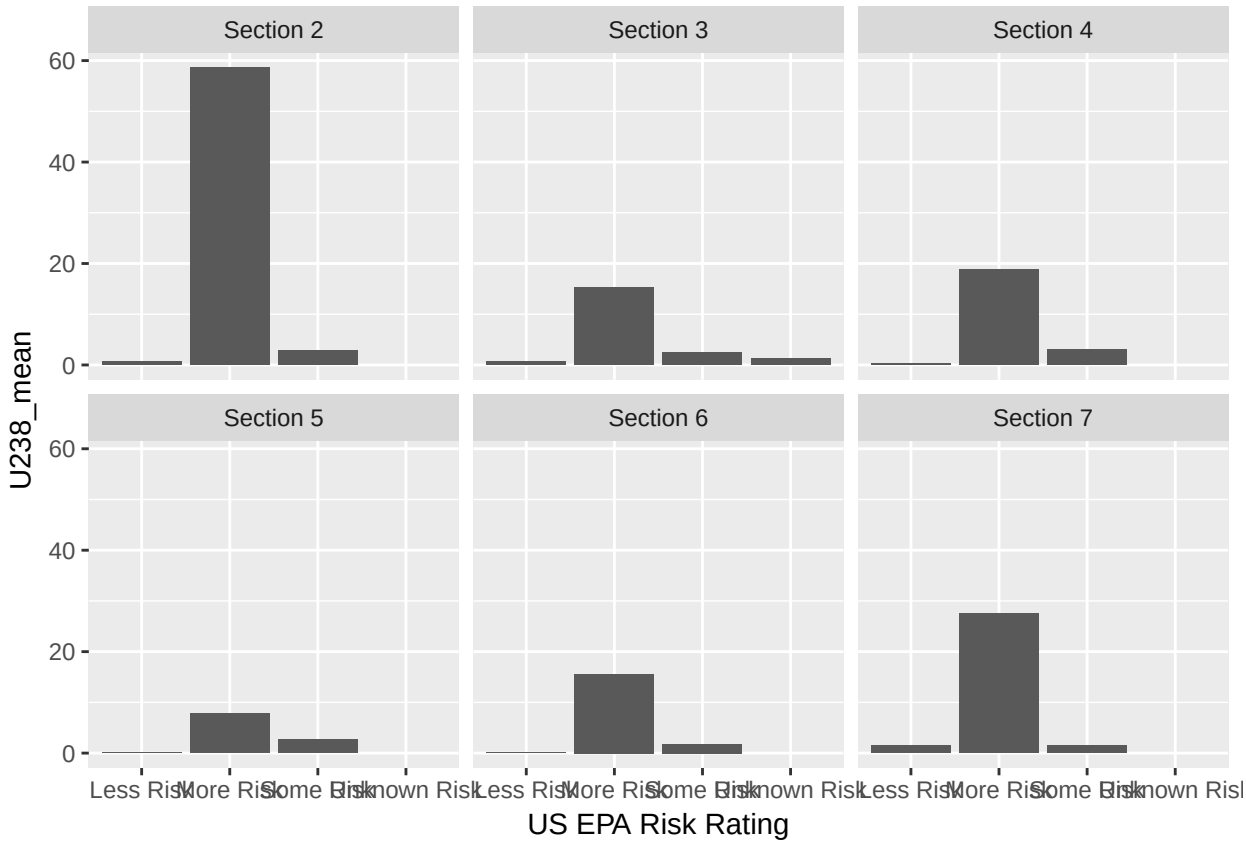
```
EPAU234_tbl <- summarize(group_by(report,
    `Which EPA Section is This From?`,
    `US EPA Risk Rating`),
    U234_mean = mean(`Amount of Uranium234`,
        na.rm = T))
EPAU235_tbl <- summarize(group_by(report,
    `Which EPA Section is This From?`,
    `US EPA Risk Rating`),
    U235_mean = mean(`Amount of Uranium235`,
        na.rm = T))
EPAU238_tbl <- summarize(group_by(report,
    `Which EPA Section is This From?`,
    `US EPA Risk Rating`),
    U238_mean = mean(`Amount of Uranium238`,
        na.rm = T))
ggplot(data = EPAU234_tbl) +
    geom_col(aes(x = `US EPA Risk Rating`,
        y = U234_mean)) +
    facet_wrap(~`Which EPA Section is This From?`)
```



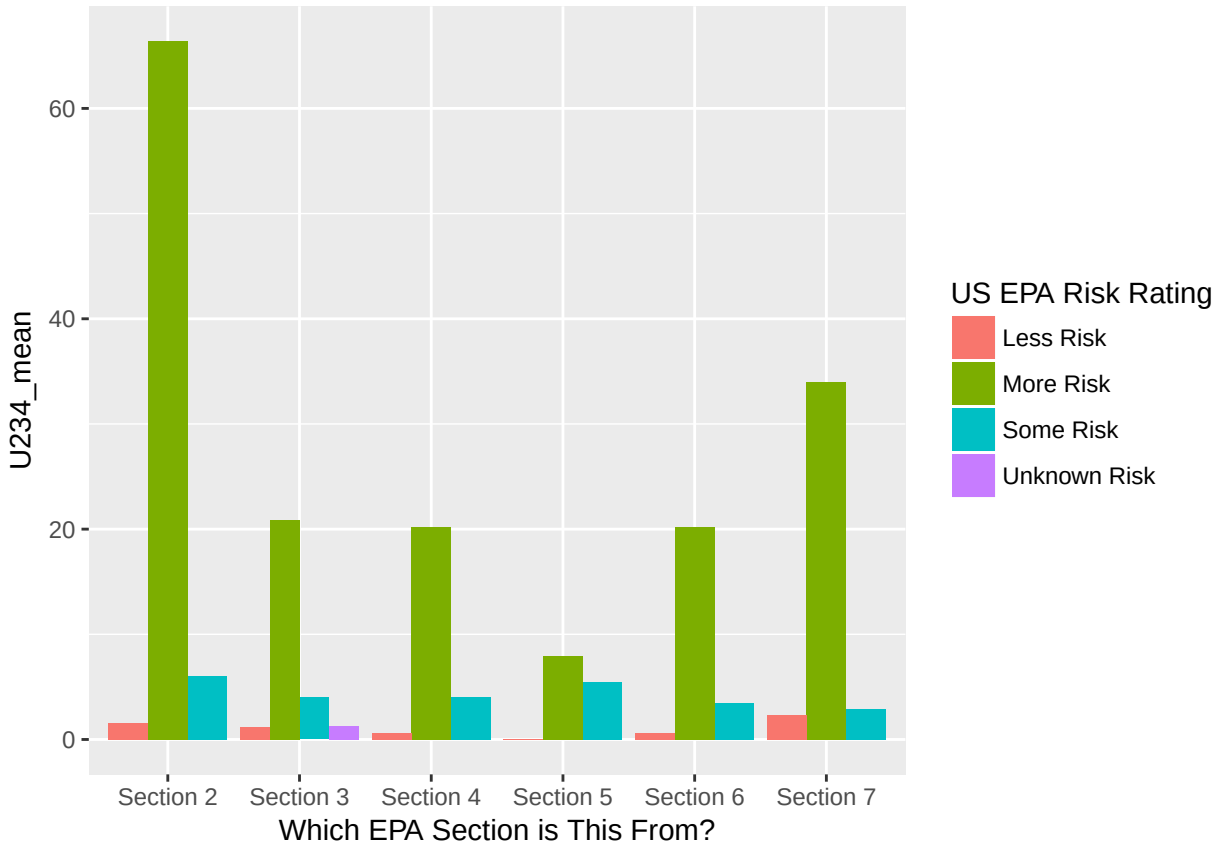

```
ggplot(data = EPAU235_tbl) +
    geom_col(aes(x = `US EPA Risk Rating`,
        y = U235_mean)) +
    facet_wrap(~`Which EPA Section is This From?`)
```



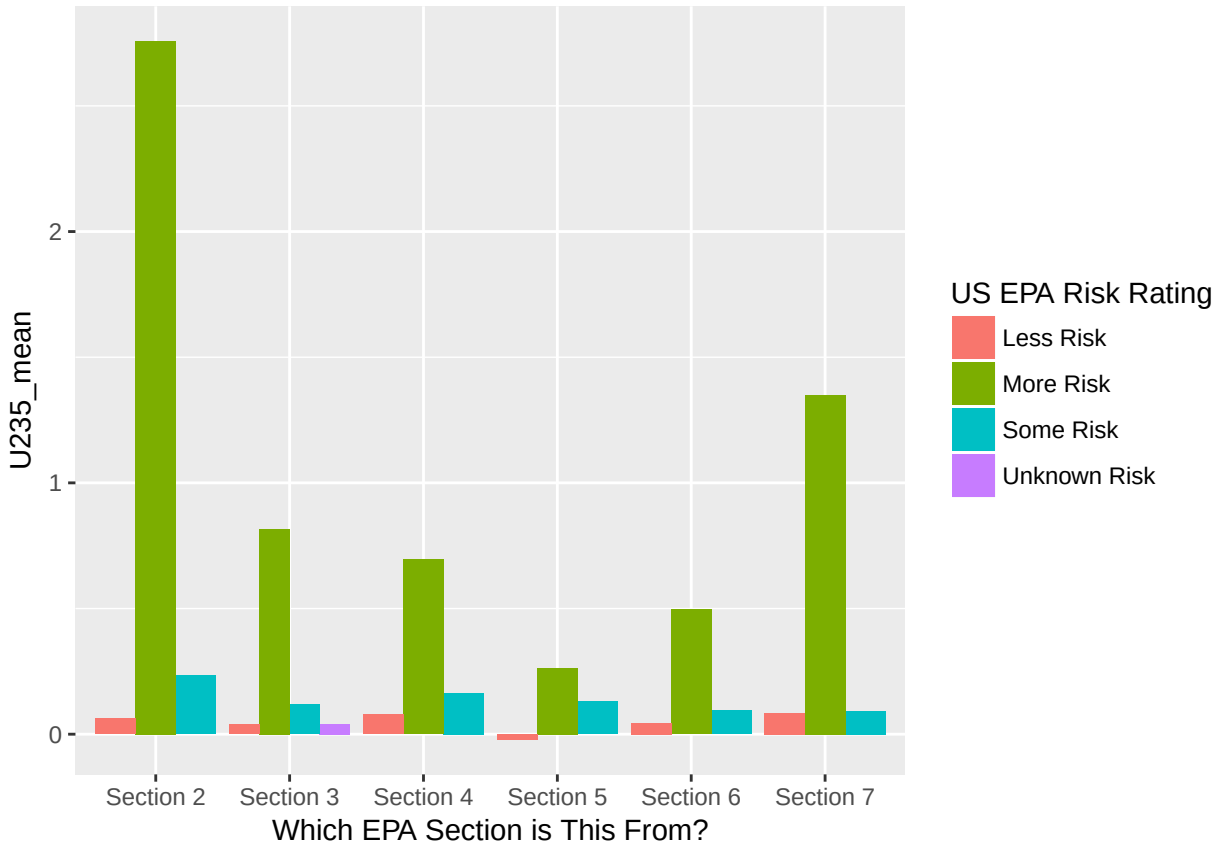

```
ggplot(data = EPAU238_tbl) +
    geom_col(aes(x = `US EPA Risk Rating`,
        y = U238_mean)) +
    facet_wrap(~`Which EPA Section is This From?`)
```



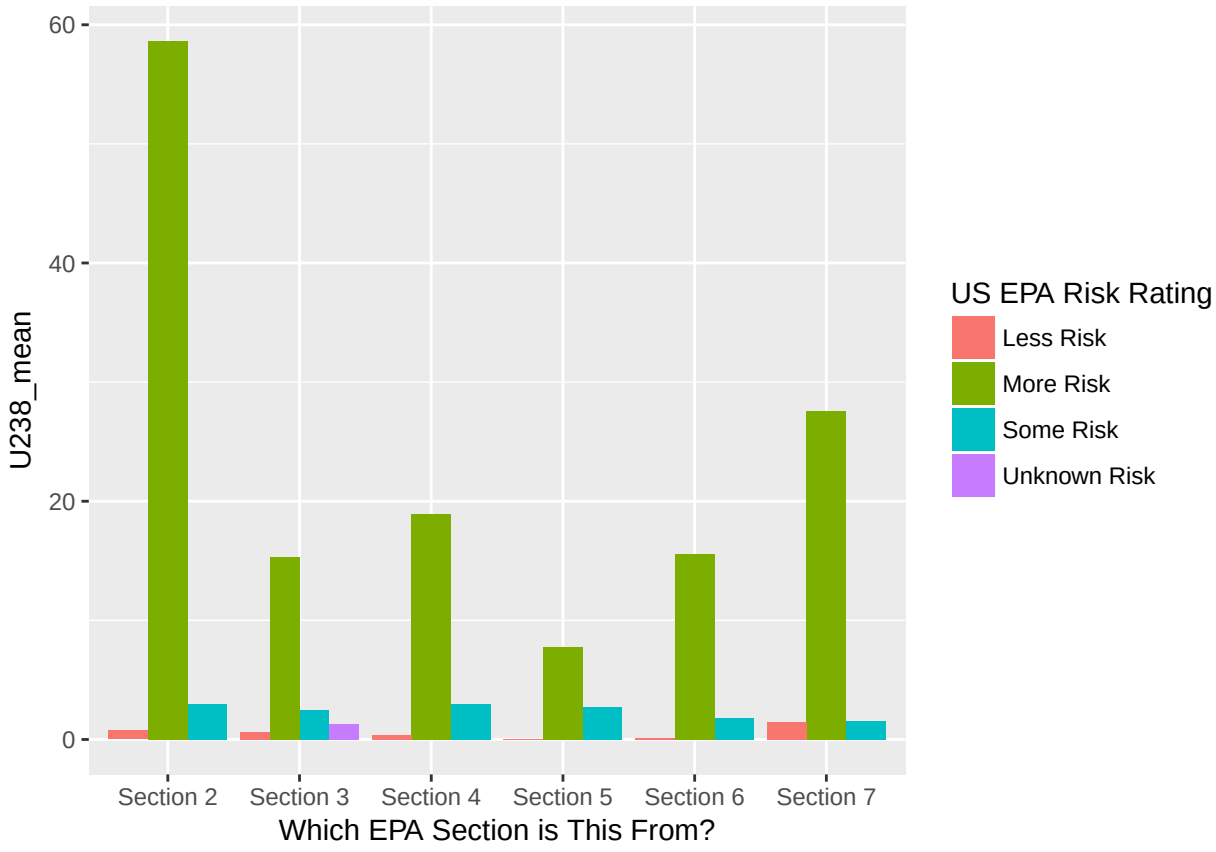

```
ggplot(data = EPAU234_tbl) +
    geom_col(aes(x = `Which EPA Section is This From?`,
        y = U234_mean, fill = `US EPA Risk Rating`),
        position = "dodge")
```




```
ggplot(data = EPAU235_tbl) +
    geom_col(aes(x = `Which EPA Section is This From?`,
        y = U235_mean, fill = `US EPA Risk Rating`),
        position = "dodge")
```




```
ggplot(data = EPAU238_tbl) +
    geom_col(aes(x = `Which EPA Section is This From?`,
        y = U238_mean, fill = `US EPA Risk Rating`),
        position = "dodge")
```



---

prob1_6

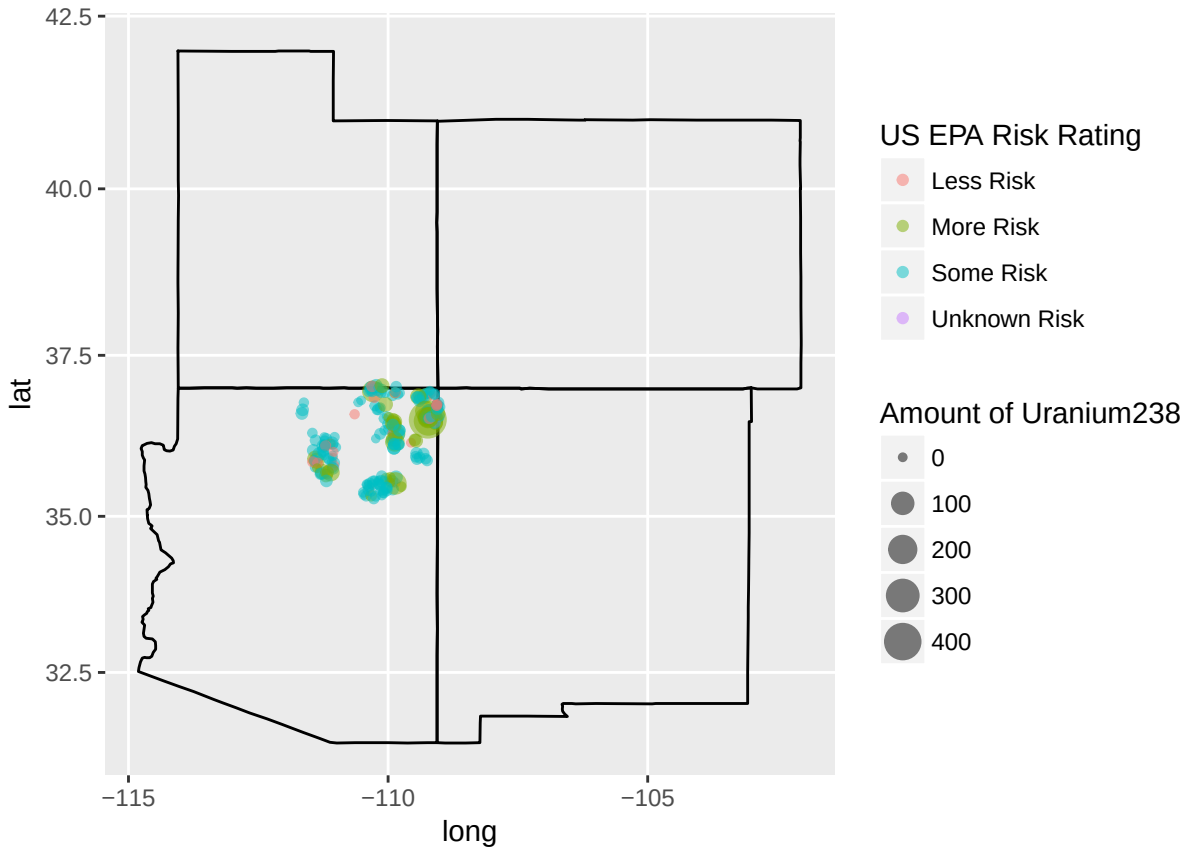
```
library(maps)
```

```
##
## Attaching package: 'maps'
```

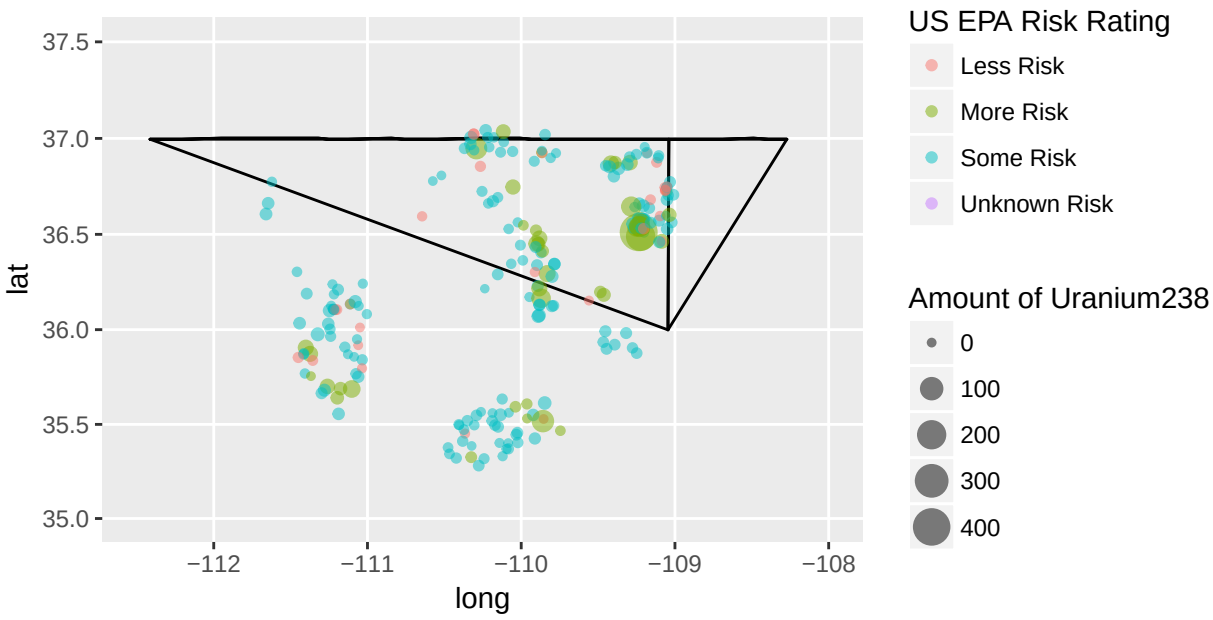
```
## The following object is masked from 'package:purrr':
##
##     map
```

```
library(measurements)
four_corners <- map_data("state",
    region = c("arizona",
        "new mexico", "utah",
        "colorado"))
report_x <- mutate(report,
    long = as.numeric(conv_unit(Longitude,
        "deg_min_sec", "dec_deg")),
    lati = as.numeric(conv_unit(report$Latitude,
        "deg_min_sec", "dec_deg")))
ggplot(four_corners) + geom_polygon(mapping = aes(x = long,
    y = lat, group = group),
    fill = NA, color = "black") +
    geom_point(data = report_x,
        aes(x = -long, y = lati,
```

```
            color = `US EPA Risk Rating`,
            size = `Amount of Uranium238`),
        alpha = 0.5, position = "jitter") +
    coord_map()
```




```
ggplot(four_corners) + geom_polygon(mapping = aes(x = long,
    y = lat, group = group),
    fill = NA, color = "black") +
    geom_point(data = report_x,
        aes(x = -long, y = lati,
            color = `US EPA Risk Rating`,
            size = `Amount of Uranium238`),
        alpha = 0.5, position = "jitter") +
    coord_map() + xlim(-112.5,
    -108) + ylim(35, 37.5)
```



In following questions, red lines in graphs are y=x reference lines The nationwide schools' proportion of black people in suspension VS nationwide shcool's black's proportion is also plotted, and summary table is given.So do other questions

Prob1.7

data from: SCH_ENR_BL_M,70,1,Overall Student Enrollment: Black Male,I,7 SCH_ENR_BL_F,71,1,Overall Student Enrollment: Black Female,I,7

TOT_ENR_M,76,1,Total number of students enrolled: Male,I,7 TOT_ENR_F,77,1,Total number of students enrolled: Female,I,7

TOT_DISCWODIS_ISS_M,1255,1,Total number of students without disabilities who received one or more in-school suspensions: Male,II,17 TOT_DISCWODIS_ISS_F,1256,1,Total number of students without disabilities who received one or more in-school suspensions: Female,II,17

TOT_DISCWDIS_ISS_IDEA_M,1309,1,Total number of students with disabilities who received one or more in-school suspensions: IDEA Male,II,18 TOT_DISCWDIS_ISS_IDEA_F,1310,1,Total number of students with disabilities who received one or more in-school suspensions: IDEA Female,II,18

SCH_DISCWDIS_ISS_504_M,1313,1,Students with disabilities who received one or more in-school suspensions: Section 504 Only Male,II,18 SCH_DISCWDIS_ISS_504_F,1314,1,Students with disabilities who received one or more in-school suspensions: Section 504 Only Female,II,18

SCH_DISCWODIS_ISS_BL_M,1249,1,Students without disabilities who received one or more in-school suspensions: Black Male,II,17 SCH_DISCWODIS_ISS_BL_F,1250,1,Students without disabilities who received one or more in-school suspensions: Black Female,II,17

SCH_DISCWDIS_ISS_IDEA_BL_M,1303,1,Students with disabilities who received one or more in-school suspensions: IDEA Black Male,II,18 SCH_DISCWDIS_ISS_IDEA_BL_F,1304,1,Students with disabilities who received one or more in-school suspensions: IDEA Black Female,II,18 ??504 black

```
sch_rp <- read_csv("CRDC2013_14_SCH.csv")
```

```
## 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 57 more columns
## )
```

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

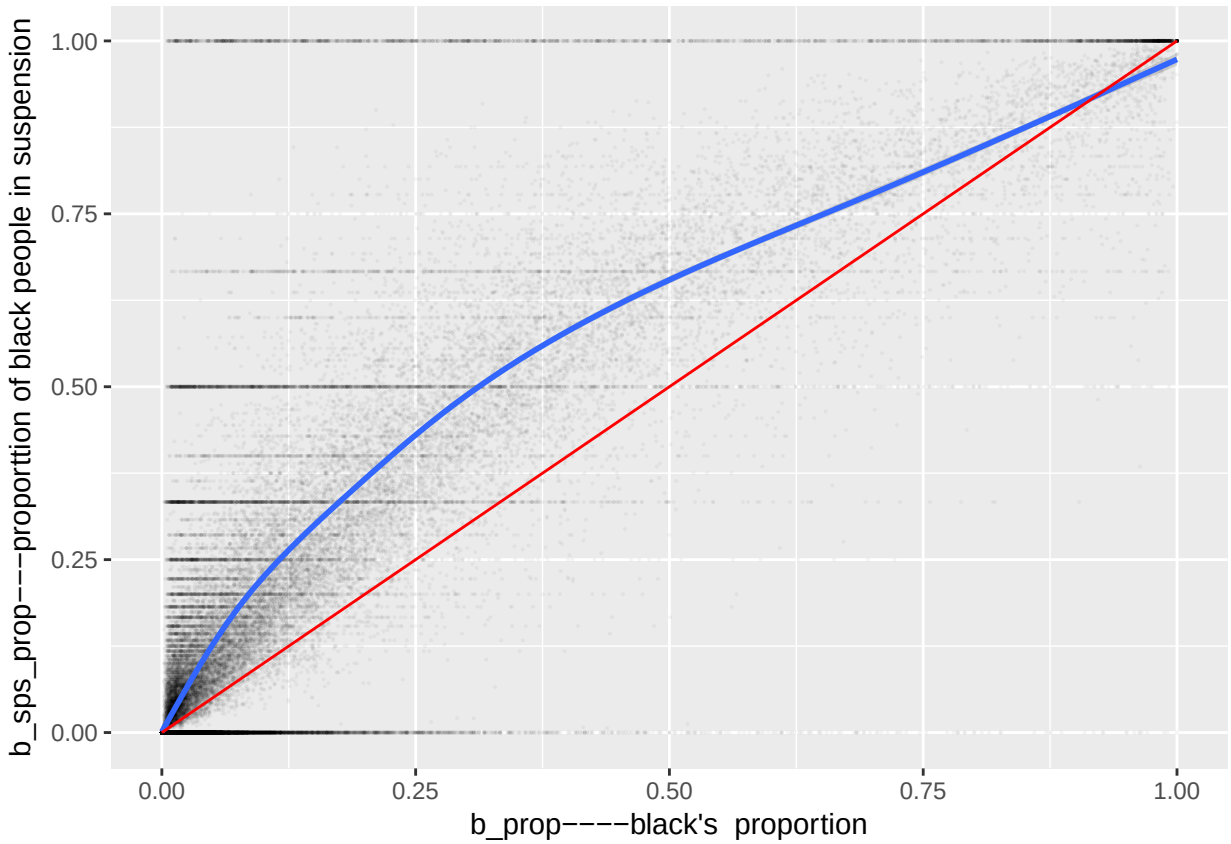
```
library(tidyverse)
sch_rpBB = transmute(sch_rp,
    SCH_ENR_BL_M = ifelse(SCH_ENR_BL_M >=
        0, SCH_ENR_BL_M, NA),
    SCH_ENR_BL_F = ifelse(SCH_ENR_BL_M >=
        0, SCH_ENR_BL_F, NA),
    TOT_ENR_M = ifelse(TOT_ENR_M >=
        0, TOT_ENR_M, NA),
    TOT_ENR_F = ifelse(TOT_ENR_F >=
        0, TOT_ENR_F, NA),
    TOT_DISCWODIS_ISS_M = ifelse(TOT_DISCWODIS_ISS_M >=
        0, TOT_DISCWODIS_ISS_M,
        NA), TOT_DISCWODIS_ISS_F = ifelse(TOT_DISCWODIS_ISS_F >=
        0, TOT_DISCWODIS_ISS_F,
        NA), TOT_DISCWDIS_ISS_IDEA_M = ifelse(TOT_DISCWDIS_ISS_IDEA_M >=
        0, TOT_DISCWDIS_ISS_IDEA_M,
        NA), TOT_DISCWDIS_ISS_IDEA_F = ifelse(TOT_DISCWDIS_ISS_IDEA_F >=
        0, TOT_DISCWDIS_ISS_IDEA_F,
        NA), SCH_DISCWDIS_ISS_504_M = ifelse(SCH_DISCWDIS_ISS_504_M >=
        0, SCH_DISCWDIS_ISS_504_M,
        NA), SCH_DISCWDIS_ISS_504_F = ifelse(SCH_DISCWDIS_ISS_504_F >=
        0, SCH_DISCWDIS_ISS_504_F,
        NA), SCH_DISCWODIS_ISS_BL_M = ifelse(SCH_DISCWODIS_ISS_BL_M >=
        0, SCH_DISCWODIS_ISS_BL_M,
        NA), SCH_DISCWODIS_ISS_BL_F = ifelse(SCH_DISCWODIS_ISS_BL_F >=
        0, SCH_DISCWODIS_ISS_BL_F,
        NA), SCH_DISCWDIS_ISS_IDEA_BL_M = ifelse(SCH_DISCWDIS_ISS_IDEA_BL_M >=
```

```
        0, SCH_DISCWDIS_ISS_IDEA_BL_M,
        NA), SCH_DISCWDIS_ISS_IDEA_BL_F = ifelse(SCH_DISCWDIS_ISS_IDEA_BL_F >=
        0, SCH_DISCWDIS_ISS_IDEA_BL_F,
        NA))


ref_line <- data.frame(A = c(0,
    1), B = c(0, 1))

black_sps <- transmute(sch_rpBB,
    tot_enr = TOT_ENR_M +
        TOT_ENR_F, tot_benr = (SCH_ENR_BL_M) +
        abs(SCH_ENR_BL_F),
    tot_sps = (TOT_DISCWODIS_ISS_M) +
        (TOT_DISCWODIS_ISS_F) +
        (TOT_DISCWDIS_ISS_IDEA_M) +
        (TOT_DISCWDIS_ISS_IDEA_F) +
        (SCH_DISCWDIS_ISS_504_M) +
        (SCH_DISCWDIS_ISS_504_F),
    tot_b_sps = (SCH_DISCWODIS_ISS_BL_M) +
        (SCH_DISCWODIS_ISS_BL_F) +
        (SCH_DISCWDIS_ISS_IDEA_BL_M) +
        (SCH_DISCWDIS_ISS_IDEA_BL_F),
    b_prop = tot_benr/tot_enr,
    b_sps_prop = tot_b_sps/tot_sps)

ggplot(data = black_sps) +
    geom_point(aes(x = b_prop,
        y = b_sps_prop), size = 0.01,
        alpha = 0.045) + geom_smooth(aes(x = b_prop,
    y = b_sps_prop)) + geom_path(data = ref_line,
    aes(x = A, y = B), color = "red") +
    labs(x = "b_prop----black's  proportion",
        y = "b_sps_prop---proporttion of black people in suspension")
```

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

```
## Warning: Removed 35525 rows containing non-finite values (stat_smooth).
```

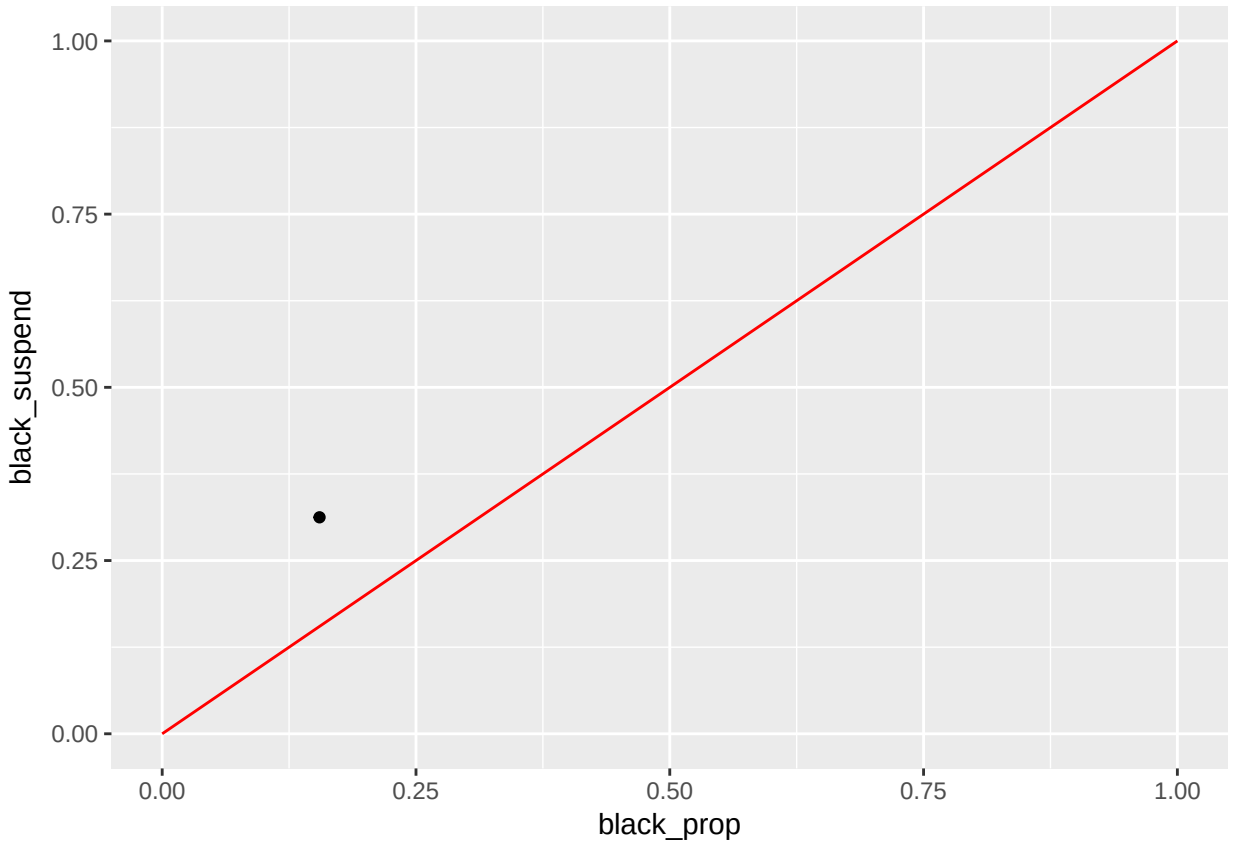
```
## Warning: Removed 35525 rows containing missing values (geom_point).
```




```
black_cal <- summarize(black_sps,
    black_prop = sum(tot_benr,
        na.rm = T)/sum(tot_enr,
        na.rm = T), black_suspend = sum(tot_b_sps,
        na.rm = T)/sum(tot_sps,
        na.rm = T))
black_cal
```

```
## # A tibble: 1 x 2
##   black_prop black_suspend
##        <dbl>         <dbl>
## 1  0.1549763      0.312431
```

```
ggplot(data = black_cal) +
    geom_point(aes(x = black_prop,
        y = black_suspend)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red")
```



It's seems blackstudents are a little more suspended, both through the smooth line and the last summarize(overall proportion of Black students across all schools and the overall proportion of suspended students who are Black across all schools). So, according to red reference line, black students are over-presented in suspensions.

Prob1.8 data from:

TOT_ENR_M,76,1,Total number of students enrolled: Male,I,7 TOT_ENR_F,77,1,Total number of students enrolled: Female,I,7

TOT_IDEAENR_M,132,1,"Total number of students with disabilities, served under IDEA: Male",I,10 TOT_IDEAENR_F,133,1,"Total number of students with disabilities, served under IDEA: Female",I,10

TOT_504ENR_M,150,1,"Total number of students with disabilities, served under Section 504 only: Male",I,10 TOT_504ENR_F,151,1,"Total number of students with disabilities, served under Section 504 only: Female",I,10

SCH_CORPINSTANCES_IND,1162,2,Corporal Punishment Indicator: Does this school use corporal punishment to discipline students?,II,16

TOT_DISCWODIS_CORP_M,1177,1,Total number of students without disabilities who received corporal punishment: Male,II,17 TOT_DISCWODIS_CORP_F,1178,1,Total number of students without disabilities who received corporal punishment: Female,II,17

TOT_DISCWDIS_CORP_IDEA_M,1195,1,Total number of students with disabilities who received corporal punishment: IDEA Male,II,18 TOT_DISCWDIS_CORP_IDEA_F,1196,1,Total number of students with disabilities who received corporal punishment: IDEA Female,II,18

SCH_DISCWDIS_CORP_504_M,1199,1,Students with disabilities who received corporal punishment: Section 504 Only Male,II,18 SCH_DISCWDIS_CORP_504_F,1200,1,Students with disabilities who received

corporal punishment: Section 504 Only Female,II,18

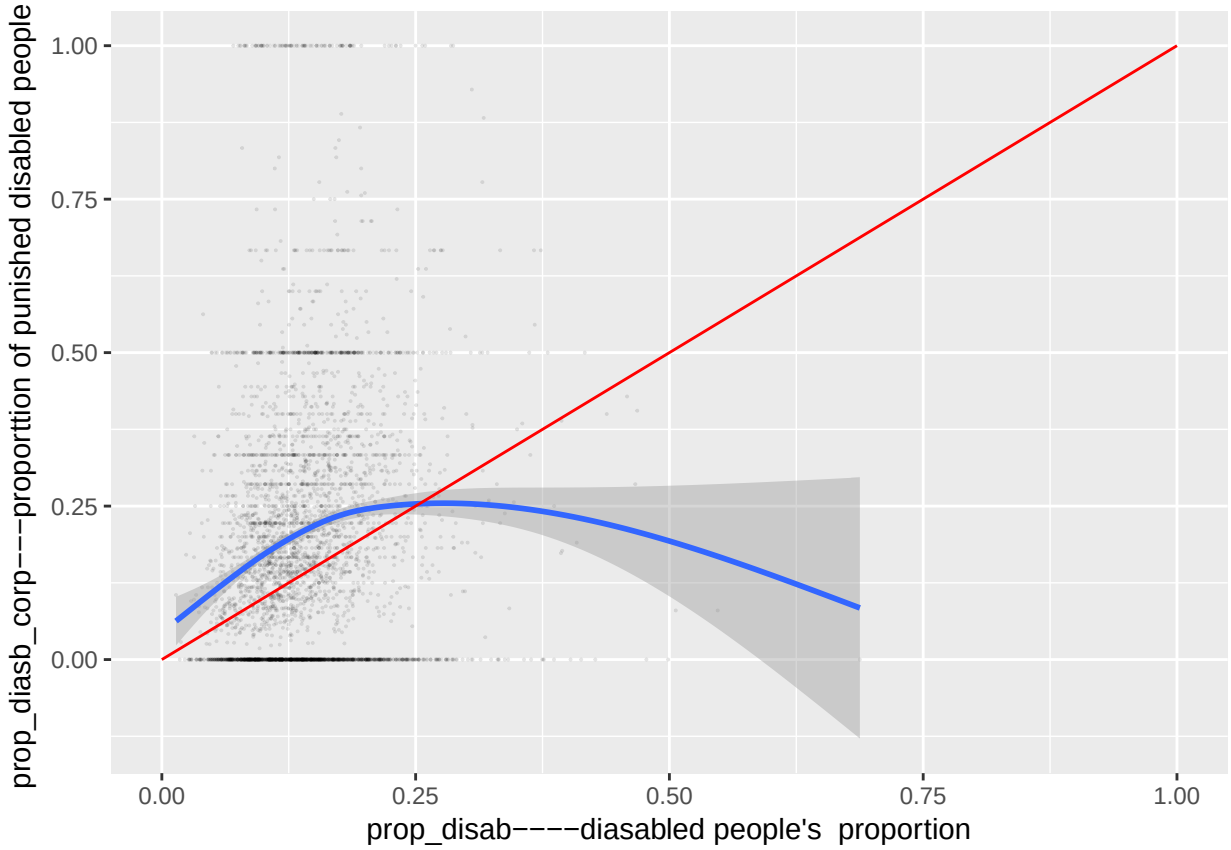
```
library(tidyverse)

sch_rpDD = transmute(sch_rp,
    SCH_CORPINSTANCES_IND = SCH_CORPINSTANCES_IND,
    TOT_ENR_M = ifelse(TOT_ENR_M >=
        0, TOT_ENR_M, NA),
    TOT_ENR_F = ifelse(TOT_ENR_F >=
        0, TOT_ENR_F, NA),
    TOT_IDEAENR_M = ifelse(TOT_IDEAENR_M >=
        0, TOT_IDEAENR_M,
        NA), TOT_IDEAENR_F = ifelse(TOT_IDEAENR_F >=
        0, TOT_IDEAENR_F,
        NA), TOT_504ENR_M = ifelse(TOT_504ENR_M >=
        0, TOT_504ENR_M, NA),
    TOT_504ENR_F = ifelse(TOT_504ENR_F >=
        0, TOT_504ENR_F, NA),
    TOT_DISCWODIS_CORP_M = ifelse(TOT_DISCWODIS_CORP_M >=
        0, TOT_DISCWODIS_CORP_M,
        NA), TOT_DISCWODIS_CORP_F = ifelse(TOT_DISCWODIS_CORP_F >=
        0, TOT_DISCWODIS_CORP_F,
        NA), TOT_DISCWDIS_CORP_IDEA_M = ifelse(TOT_DISCWDIS_CORP_IDEA_M >=
        0, TOT_DISCWDIS_CORP_IDEA_M,
        NA), TOT_DISCWDIS_CORP_IDEA_F = ifelse(TOT_DISCWDIS_CORP_IDEA_F >=
        0, TOT_DISCWDIS_CORP_IDEA_F,
        NA), SCH_DISCWDIS_CORP_504_M = ifelse(SCH_DISCWDIS_CORP_504_M >=
        0, SCH_DISCWDIS_CORP_504_M,
        NA), SCH_DISCWDIS_CORP_504_F = ifelse(SCH_DISCWDIS_CORP_504_F >=
        0, SCH_DISCWDIS_CORP_504_F,
        NA))


corp_tbl_raw <- transmute(sch_rpDD,
    SCH_CORPINSTANCES_IND = SCH_CORPINSTANCES_IND,
    tot_enr = (TOT_ENR_M) +
        (TOT_ENR_F), tot_disab = (TOT_IDEAENR_M) +
        (TOT_IDEAENR_F) +
        (TOT_504ENR_M) + (TOT_504ENR_F),
    tot_corp = (TOT_DISCWODIS_CORP_M) +
        (TOT_DISCWODIS_CORP_F) +
        (TOT_DISCWDIS_CORP_IDEA_M) +
        (TOT_DISCWDIS_CORP_IDEA_F) +
        (SCH_DISCWDIS_CORP_504_M) +
        (SCH_DISCWDIS_CORP_504_F),
    tot_disab_corp = (TOT_DISCWDIS_CORP_IDEA_M) +
        (TOT_DISCWDIS_CORP_IDEA_F) +
        (SCH_DISCWDIS_CORP_504_M) +
        (SCH_DISCWDIS_CORP_504_F),
    prop_disab = tot_disab/tot_enr,
    prop_diasb_corp = tot_disab_corp/tot_corp)

corp_tbl <- filter(corp_tbl_raw,
    SCH_CORPINSTANCES_IND ==
```

```
        "YES")
ref_line <- data.frame(A = c(0,
    1), B = c(0, 1))
ggplot(data = corp_tbl) +
    geom_point(aes(x = prop_disab,
        y = prop_diasb_corp),
        size = 0.01, alpha = 0.1) +
    geom_smooth(aes(x = prop_disab,
        y = prop_diasb_corp)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red") + labs(x = "prop_disab----diasabled people's  proportion",
    y = "prop_diasb_corp---proporttion of punished disabled people ")
```

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

```
## Warning: Removed 1270 rows containing non-finite values (stat_smooth).
```

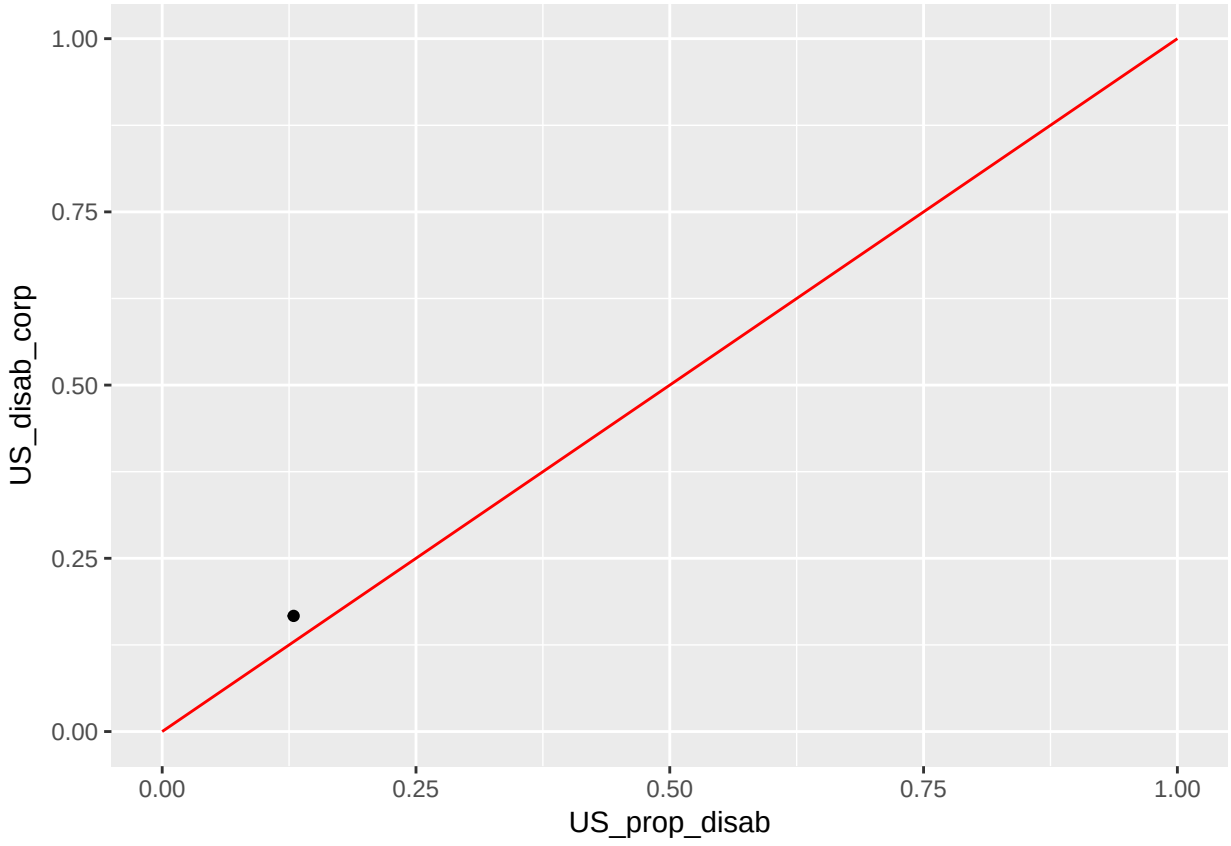
```
## Warning: Removed 1270 rows containing missing values (geom_point).
```




```
corp_total <- summarize(corp_tbl,
    US_prop_disab = sum(tot_disab,
        na.rm = T)/sum(tot_enr,
        na.rm = T), US_disab_corp = sum(tot_disab_corp,
        na.rm = T)/sum(tot_corp,
        na.rm = T))
corp_total
```

```
## # A tibble: 1 x 2
##   US_prop_disab US_disab_corp
##           <dbl>         <dbl>
## 1     0.1294762     0.1668854
```

```
ggplot(data = corp_total) +
    geom_point(aes(x = US_prop_disab,
        y = US_disab_corp)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red")
```



It seems that more disbled students suffer from a little more corporal punishments than common(the red reference line) ,meaning over-presentation in punishment But when proportion of disabilities increase to a certain value, they seems to be less punished.

Prob1.9 data from: TOT_ENR_M,76,1,Total number of students enrolled: Male,I,7 TOT_ENR_F,77,1,Total number of students enrolled: Female,I,7

SCH_ENR_HI_M,62,1,Overall Student Enrollment: Hispanic Male,I,7 SCH_ENR_HI_F,63,1,Overall Student Enrollment: Hispanic Female,I,7 SCH_ENR_BL_M,70,1,Overall Student Enrollment: Black Male,I,7 SCH_ENR_BL_F,71,1,Overall Student Enrollment: Black Female,I,7

SCH_GT_IND,215,2,Gifted and Talented Education Program Indicator,I,11

TOT_GTENR_M,230,1,Total number of students enrolled in gifted/talented programs: Male,I,12

TOT_GTENR_F,231,1,Total number of students enrolled in gifted/talented programs: Female,I,12

SCH_GTENR_BL_M,224,1,Gifted and Talented Students Enrollment: Black Male,I,12 SCH_GTENR_BL_F,225,1,Gifted and Talented Students Enrollment: Black Female,I,12 SCH_GTENR_HI_M,216,1,Gifted and Talented Students Enrollment: Hispanic Male,I,12 SCH_GTENR_HI_F,217,1,Gifted and Talented Students Enrollment: Hispanic Female,I,12

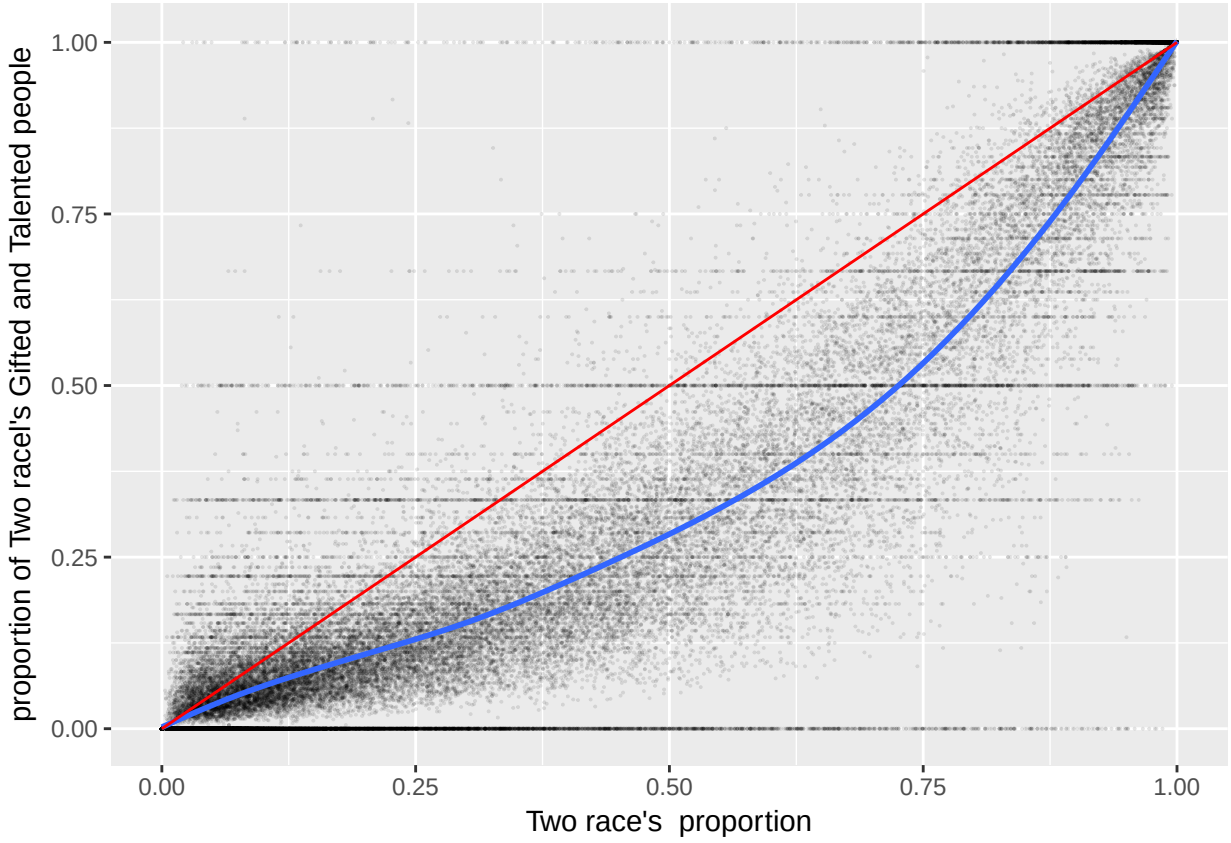
```
library(tidyverse)

sch_rpG = transmute(sch_rp,
    SCH_GT_IND = SCH_GT_IND,
    TOT_ENR_M = ifelse(TOT_ENR_M >=
        0, TOT_ENR_M, NA),
    TOT_ENR_F = ifelse(TOT_ENR_F >=
        0, TOT_ENR_F, NA),
    SCH_ENR_HI_M = ifelse(SCH_ENR_HI_M >=
        0, SCH_ENR_HI_M, NA),
    SCH_ENR_HI_F = ifelse(SCH_ENR_HI_F >=
        0, SCH_ENR_HI_F, NA),
    SCH_ENR_BL_M = ifelse(SCH_ENR_BL_M >=
        0, SCH_ENR_BL_M, NA),
    SCH_ENR_BL_F = ifelse(SCH_ENR_BL_F >=
        0, SCH_ENR_BL_F, NA),
    TOT_GTENR_M = ifelse(TOT_GTENR_M >=
        0, TOT_GTENR_M, NA),
    TOT_GTENR_F = ifelse(TOT_GTENR_F >=
        0, TOT_GTENR_F, NA),
    SCH_GTENR_BL_M = ifelse(SCH_GTENR_BL_M >=
        0, SCH_GTENR_BL_M,
        NA), SCH_GTENR_BL_F = ifelse(SCH_GTENR_BL_F >=
        0, SCH_GTENR_BL_F,
        NA), SCH_GTENR_HI_M = ifelse(SCH_GTENR_HI_M >=
        0, SCH_GTENR_HI_M,
        NA), SCH_GTENR_HI_F = ifelse(SCH_GTENR_HI_F >=
        0, SCH_GTENR_HI_F,
        NA))


corp_tbl_raw <- transmute(sch_rpG,
    SCH_GT_IND = SCH_GT_IND,
    tot_enr = (TOT_ENR_M) +
        (TOT_ENR_F), tot_bh = (SCH_ENR_HI_M) +
        (SCH_ENR_HI_F) + (SCH_ENR_BL_M) +
        (SCH_ENR_BL_F), tot_gt = (TOT_GTENR_M) +
        (TOT_GTENR_F), tot_bh_gt = (SCH_GTENR_BL_M) +
        (SCH_GTENR_BL_F) +
        (SCH_GTENR_HI_M) +
        (SCH_GTENR_HI_F),
    prop_bh = tot_bh/tot_enr,
    prop_bh_gt = tot_bh_gt/tot_gt)

corp_tbl <- filter(corp_tbl_raw,
    SCH_GT_IND == "YES")
ref_line <- data.frame(A = c(0,
    1), B = c(0, 1))
```

```
ggplot() + geom_point(data = corp_tbl,
    aes(x = tot_bh/tot_enr,
        y = tot_bh_gt/tot_gt),
    size = 0.01, alpha = 0.1) +
    geom_smooth(data = corp_tbl,
        aes(x = tot_bh/tot_enr,
            y = tot_bh_gt/tot_gt)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red") + labs(x = "Two race's  proportion",
    y = "proportion of Two racel's Gifted and Talented people")
```

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

```
## Warning: Removed 3 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

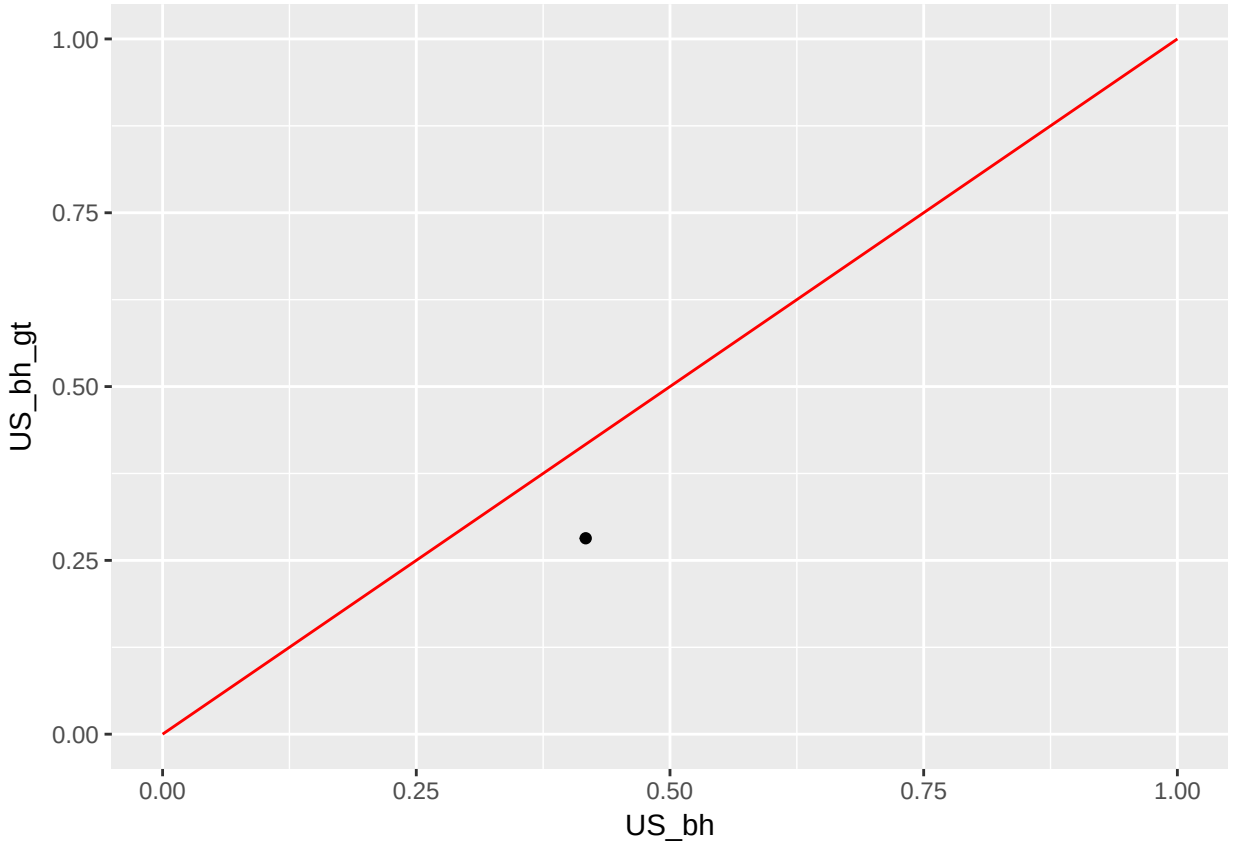



```
corp_total <- summarize(corp_tbl,
    US_bh_gt = sum(tot_bh_gt,
        na.rm = T)/sum(tot_gt,
        na.rm = T), US_bh = sum(tot_bh,
        na.rm = T)/sum(tot_enr,
        na.rm = T))
corp_total
```

```
## # A tibble: 1 x 2
##   US_bh_gt     US_bh
```

```
##       <dbl>     <dbl>
## 1 0.281827 0.4169325
```

```
ggplot(data = corp_total) +
    geom_point(aes(x = US_bh,
        y = US_bh_gt)) + geom_path(data = ref_line,
    aes(x = A, y = B), color = "red")
```



Obviously, black and hispanic people have a little less GT(Gift and Talent. . .), meaning under-representation

---

Prob1,10

Calculate The proportion of asian ACT or SAT takers among all takers VS The proportion of asian students , in every school, and plot that

Also, summarize the nationwide comparison

Data form: SCH_ENR_AS_M,66,1,Overall Student Enrollment: Asian Male,I,7 SCH_ENR_AS_F,67,1,Overall Student Enrollment: Asian Female,I,7

TOT_ENR_M,76,1,Total number of students enrolled: Male,I,7 TOT_ENR_F,77,1,Total number of students enrolled: Female,I,7

SCH_SATACT_AS_M,666,1,SAT or ACT Test Participation: Asian Male,II,7 SCH_SATACT_AS_F,667,1,SAT or ACT Test Participation: Asian Female,II,7

TOT_SATACT_M,676,1,Total number of students participating in the SAT and ACT tests: Male,II,7 TOT_SATACT_F,677,1,Total number of students participating in the SAT and ACT tests: Female,II,7

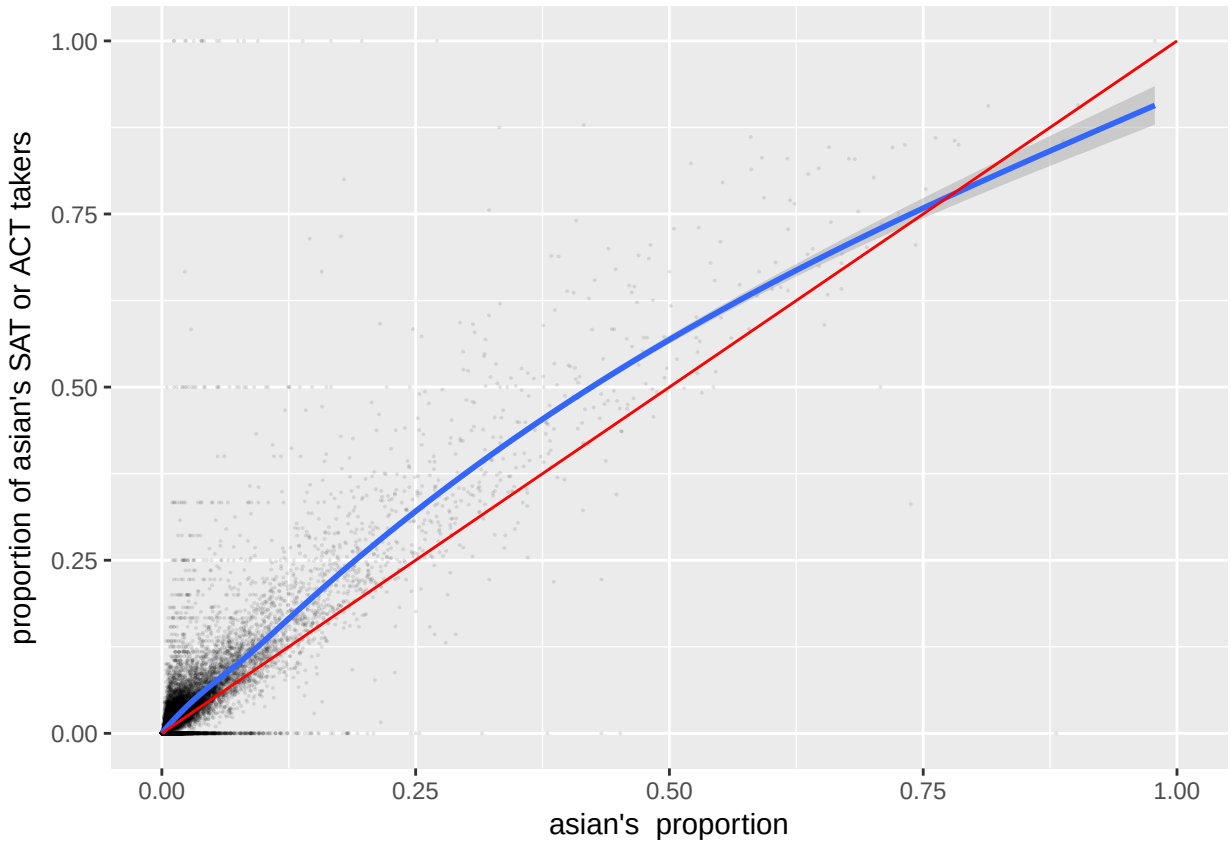
SCH_SATACT_AS_M,666,1,SAT or ACT Test Participation: Asian Male,II,7 SCH_SATACT_AS_F,667,1,SAT or ACT Test Participation: Asian Female,II,7

```
library(tidyverse)

sch_rpAS = transmute(sch_rp,
    TOT_ENR_M = ifelse(TOT_ENR_M >=
        0, TOT_ENR_M, NA),
    TOT_ENR_F = ifelse(TOT_ENR_F >=
        0, TOT_ENR_F, NA),
    SCH_ENR_AS_M = ifelse(SCH_ENR_AS_M >=
        0, SCH_ENR_AS_M, NA),
    SCH_SATACT_AS_M = ifelse(SCH_SATACT_AS_M >=
        0, SCH_SATACT_AS_M,
        NA), SCH_SATACT_AS_F = ifelse(SCH_SATACT_AS_F >=
        0, SCH_SATACT_AS_F,
        NA), SCH_ENR_AS_F = ifelse(SCH_ENR_AS_F >=
        0, SCH_ENR_AS_F, NA),
    TOT_SATACT_M = ifelse(TOT_SATACT_M >=
        0, TOT_SATACT_M, NA),
    TOT_SATACT_F = ifelse(TOT_SATACT_F >=
        0, TOT_SATACT_F, NA),
    SCH_SATACT_AS_M = ifelse(SCH_SATACT_AS_M >=
        0, SCH_SATACT_AS_M,
        NA), SCH_SATACT_AS_F = ifelse(SCH_SATACT_AS_F >=
        0, SCH_SATACT_AS_F,
        NA))



corp_tbl_raw <- transmute(sch_rpAS,
    tot_enr = (TOT_ENR_M) +
        (TOT_ENR_F), tot_asx = (SCH_ENR_AS_M) +
        (SCH_ENR_AS_F), tot_sat = (TOT_SATACT_M) +
        (TOT_SATACT_F), tot_sat_as = (SCH_SATACT_AS_M) +
        (SCH_SATACT_AS_F))

ref_line <- data.frame(A = c(0,
    1), B = c(0, 1))
ggplot() + geom_point(data = corp_tbl_raw,
    aes(x = tot_asx/tot_enr,
        y = tot_sat_as/tot_sat),
    size = 0.01, alpha = 0.1) +
    geom_smooth(data = corp_tbl_raw,
        aes(x = tot_asx/tot_enr,
            y = tot_sat_as/tot_sat)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red") + labs(x = "asian's  proportion",
    y = "proportion of asian's SAT or ACT takers")
```

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

```
## Warning: Removed 76554 rows containing non-finite values (stat_smooth).
```

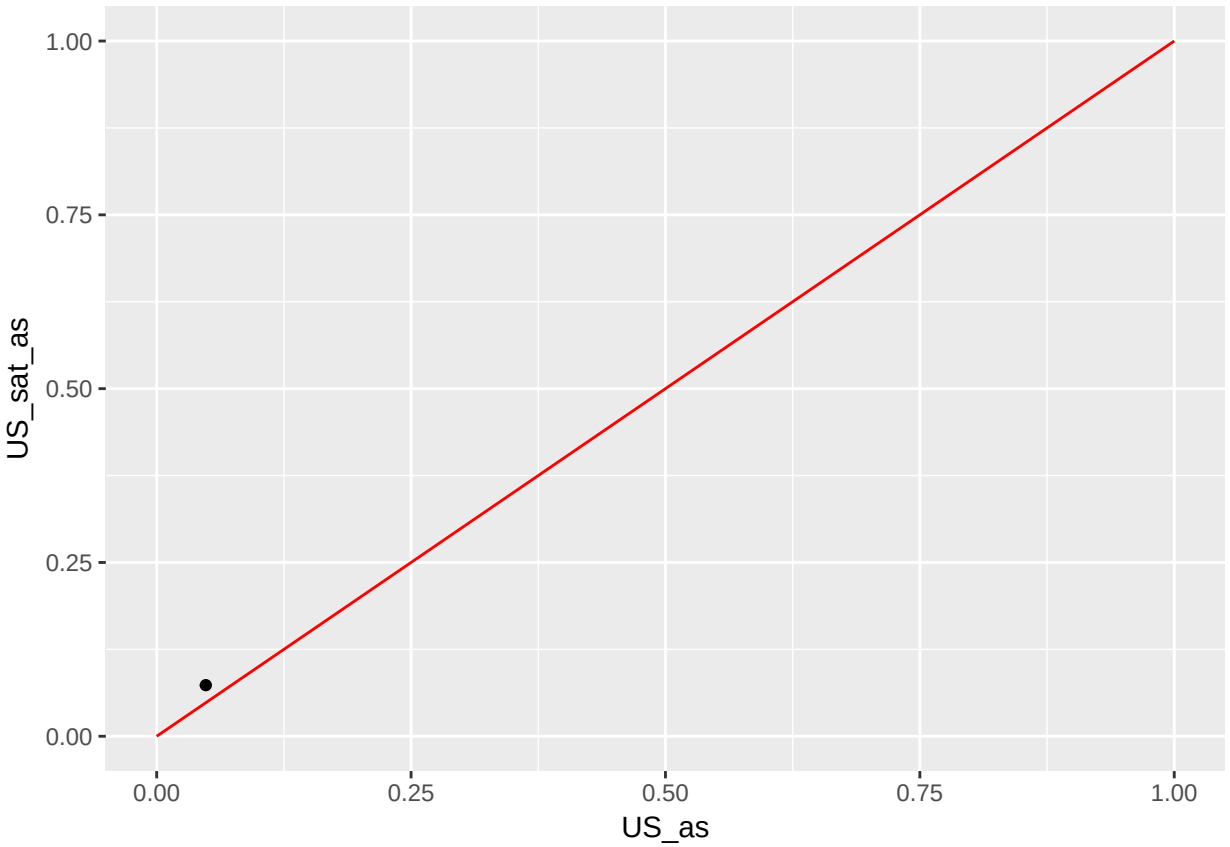
```
## Warning: Removed 76554 rows containing missing values (geom_point).
```




```
corp_total <- summarize(corp_tbl_raw,
    US_sat_as = sum(tot_sat_as,
        na.rm = T)/sum(tot_sat,
        na.rm = T), US_as = sum(tot_asx,
        na.rm = T)/sum(tot_enr,
        na.rm = T))
corp_total
```

```
## # A tibble: 1 x 2
##   US_sat_as      US_as
##       <dbl>      <dbl>
## 1 0.0735271 0.04826402
```

```
ggplot(data = corp_total) +
    geom_point(aes(x = US_as,
        y = US_sat_as)) +
    geom_path(data = ref_line,
        aes(x = A, y = B),
        color = "red")
```



Conclusion: The proportion of asian ACT or SAT takers among all takers is a littlt bigger than The proportion of asian students, meaning somewhat over-presentation. So the nationwide data also represent this