Part B

Problem 3

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 of Radium-228 with 0, then filter the dataset to remove any sites with "Unknown Risk" for the EPA risk rating.

Visualize the distribution of Radium-228 within each combination of EPA section and risk level. State your observations

Code

```
df <- read_csv("E:/FALL_2019/DS-5110-DM/HW2-DM/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.
```

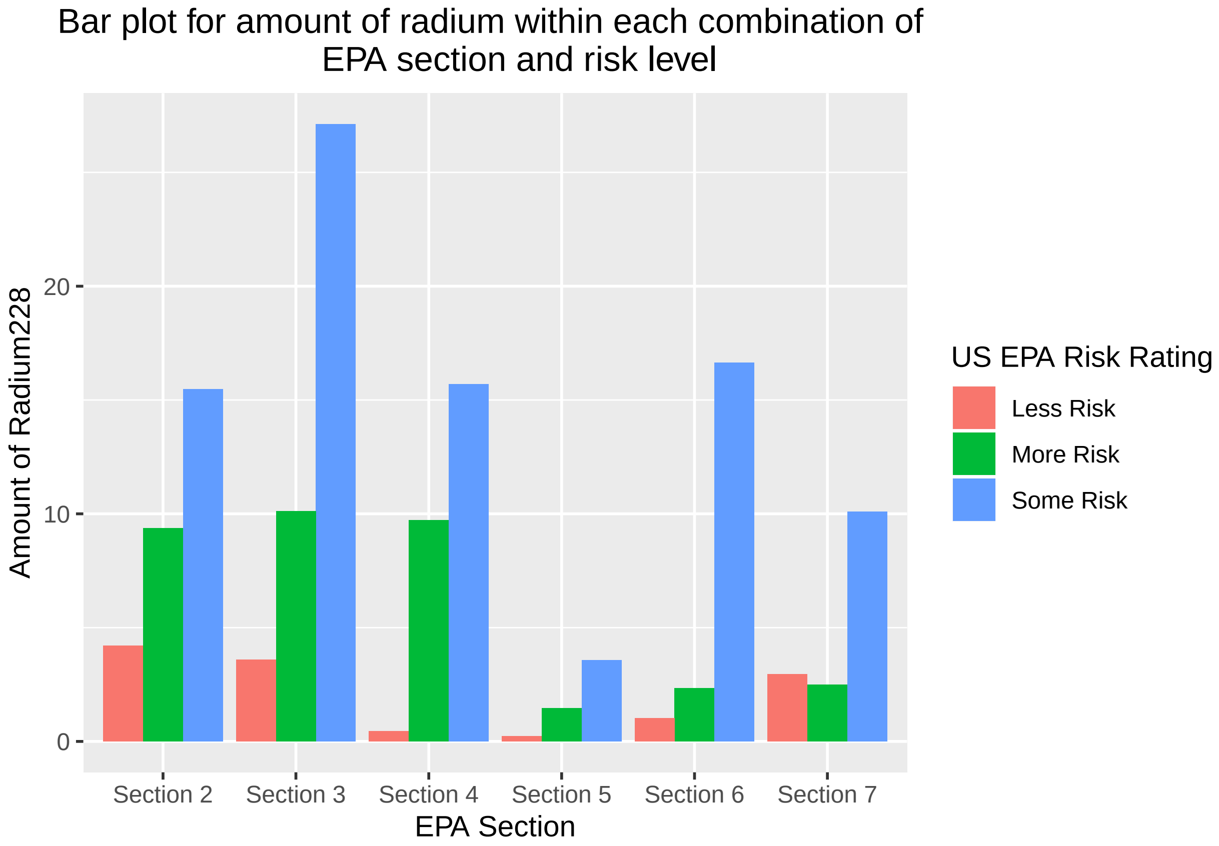
```
df1 <- mutate(df, `Amount of Radium228` = ifelse(`Amount of Radium228` >= 0 ,
                                                  `Amount of Radium228` ,0))
df2 <- filter(df1,!(`US EPA Risk Rating` == "Unknown Risk"))
#Print df2 .There is only one row corresponding to 'unknown risk' which is removed
print(df2)
```

```
## # A tibble: 224 x 64
##    `Which EPA Sect~ `Name of Water ~ `Date of Water ~ Longitude Latitude
##    <chr>            <chr>            <chr>            <chr>     <chr>
##  1 Section 3        Gold Spring      1/19/00          111 4 28~ 35 46 4~
##  2 Section 3        Tank 3K-331      7/27/98          111 24 2~ 35 46 8~
##  3 Section 6        Lower Greasewoo~ 4/14/99          109 51 1~ 35 31 4~
##  4 Section 7        Tank 8T-549      10/9/98          110 12 4~ 36 39 4~
##  5 Section 6        Cedar Spring     7/13/98          110 21 5~ 35 27 4~
##  6 Section 7        Tank 8AI-1       9/21/98          110 18 3~ 37 1 17~
##  7 Section 6        Coyote Spring    7/8/98           110 27 5~ 35 20 3~
##  8 Section 2        9T-523           3/18/99          109 10 5~ 36 55 2~
##  9 Section 6        Chimney Butte S~ 7/14/98          110 25 2~ 35 19 1~
## 10 Section 5        Nazlini Chapter~ 11/17/98         109 26 4~ 35 53 5~
## # ... with 214 more rows, and 59 more variables: `US EPA Risk
## #   Rating` <chr>, `Amount of Aluminum (Al)` <dbl>, `Exceedance of
## #   Aluminum (Al)?` <chr>, `Amount of Antimony (Sb)` <dbl>, `Exceedance of
## #   of Antimony (Sb)?` <chr>, `Amount of Arsenic (As)` <dbl>, `Exceedance
## #   of Arsenic (As)?` <chr>, `Amount of Barium (Ba)` <dbl>, `Exceedance of
## #   Barium (Ba)?` <chr>, `Amount of Beryllium (Be)` <dbl>, `Exceedance of
## #   Beryllium (Be)?` <chr>, `Amount of Cadmium (Cd)` <dbl>, `Exceedance of
## #   Cadmium (Cd)?` <chr>, `Amount of Chromium (Cr)` <dbl>, `Exceedance of
## #   Chromium (Cr)?` <chr>, `Amount of Copper (Cu)` <dbl>, `Exceedance of
## #   Copper (Cu)?` <chr>, `Amount of Iron (Fe)` <dbl>, `Exceedance of Iron
## #   (Fe)?` <chr>, `Amount of Lead (Pb)` <dbl>, `Exceedance of Lead
## #   (Pb)?` <chr>, `Amount of Manganese (Mn)` <dbl>, `Exceedance of
## #   Manganese (Mn)?` <chr>, `Amount of Mercury (Hg)` <dbl>, `Exceedance of
## #   Mercury (Hg)?` <chr>, `Amount of Nickel (Ni)` <dbl>, `Exceedance of
## #   Nickel (Ni)?` <chr>, `Amount of Selenium (Se)` <dbl>, `Exceedance of
## #   Selenium (Se)?` <chr>, `Amount of Silver (Ag)` <dbl>, `Exceedance of
## #   Silver (Ag)?` <chr>, `Amount of Thallium (TI)` <dbl>, `Exceedance of
## #   Thallium (TI)?` <chr>, `Amount of Vanadium (V)` <dbl>, `Exceedance of
## #   Vanadium (V)?` <chr>, `Amount of Zinc (Zn)` <dbl>, `Exceedance of Zinc
## #   (Zn)?` <chr>, `Amount of Alpha Particles` <dbl>, `Alpha Particle
## #   Exceedance?` <chr>, `Amount of Beta Particles` <dbl>, `Beta Particle
## #   Exceedance?` <chr>, `Amount of Lead210` <dbl>, `Exceedance of
## #   Lead210?` <chr>, `Amount of Radium226` <dbl>, `Exceedance of of
## #   Radium226?` <chr>, `Amount of Radium228` <dbl>, `Exceedance of
## #   Radium228?` <chr>, `Amount of Thorium228` <dbl>, `Exceedance of
## #   Thorium228?` <chr>, `Amount of Thorium230` <dbl>, `Exceedance of
## #   Thorium230?` <chr>, `Amount of Thorium232` <dbl>, `Exceedance of
## #   Thorium232?` <chr>, `Amount of Uranium234` <dbl>, `Exceedance of
## #   Uranium234?` <chr>, `Amount of Uranium235` <dbl>, `Exceedance of
## #   Uranium235?` <chr>, `Amount of Uranium238` <dbl>, `Exceedance of
## #   Uranium238?` <chr>
```

I have plotted 2 plot to visualize the distribution of amount of radium within each combination of EPA section and risk level.

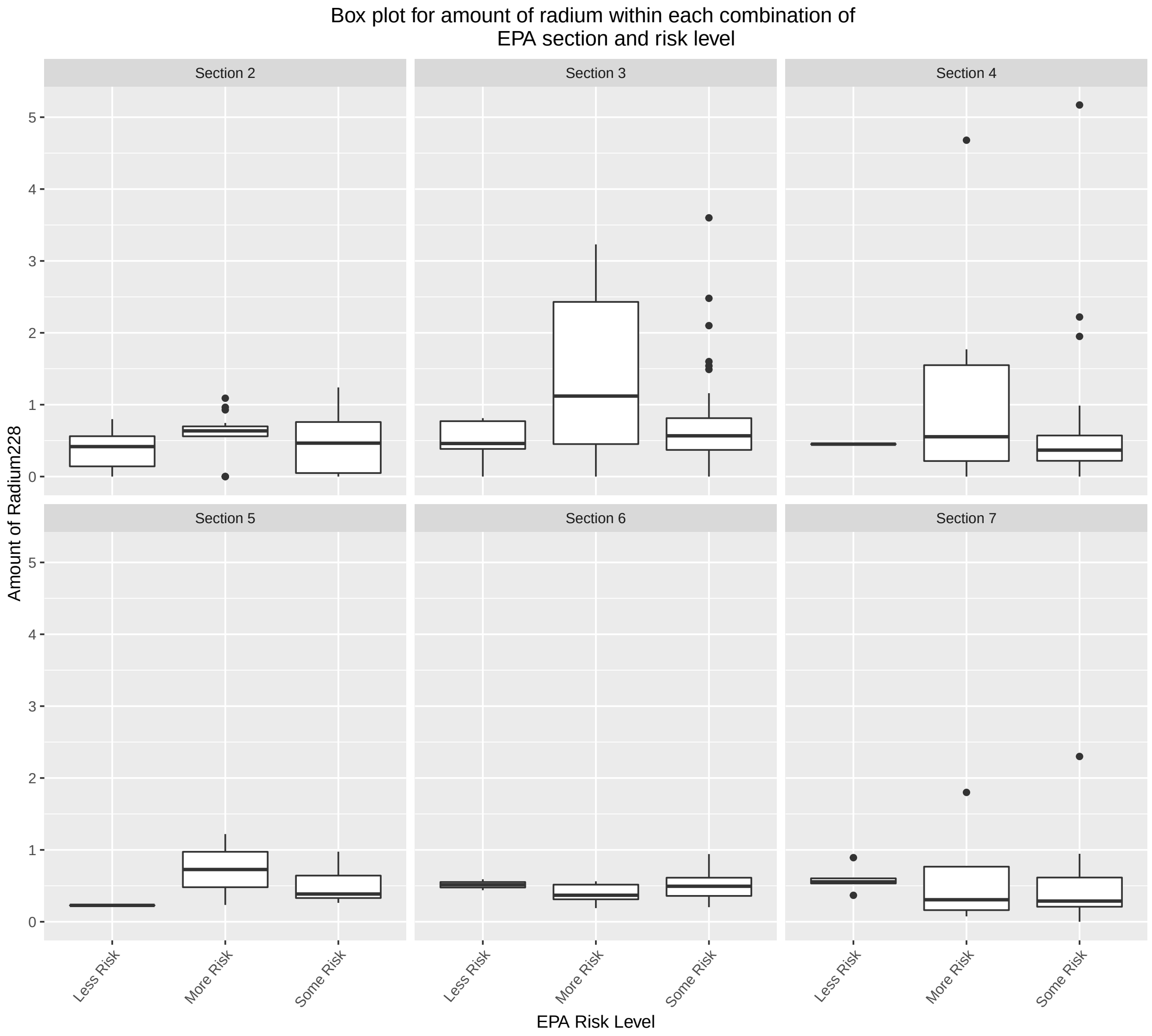
First graph shows overall distribution of amount of radium within each combination of EPA section and risk level.

```r
df3 <- group_by(df2,`Which EPA Section is This From?`,`US EPA Risk Rating`) %>%
                                     summarize(amount = sum(`Amount of Radium228`))
ggplot(df3,aes(x=`Which EPA Section is This From?`,y =amount)) +
                  geom_col(aes(fill=`US EPA Risk Rating`),position = position_dodge()) +
labs(title = "Bar plot for amount of radium within each combination of
     EPA section and risk level", x= "EPA Section", y = "Amount of Radium228") +
                                              theme(plot.title = element_text(hjust = 0.5))
```



Second graph shows boxplots of amount of radium within each combination of EPA section and risk level.

```
ggplot(df2, aes(x=`US EPA Risk Rating`,y=`Amount of Radium228`)) +
   facet_wrap(~`Which EPA Section is This From?`,nrow = 2) + geom_boxplot() +
  labs(title = "Box plot for amount of radium within each combination of
       EPA section and risk level", x= "EPA Risk Level", y = "Amount of Radium228") +
                                        theme(plot.title = element_text(hjust = 0.5))+
                                  theme(axis.text.x = element_text(angle = 50, hjust = 1))
```



Conclusion

As we can see from bar plot that water sources in section 3 have higher amount of radium228. Also from box plot we see that section 3(larger quartile range) has water sources having most amount of radium228.

Section 3 and 4 have large number of water sources with more risk and some risk.

Also we can see from bar plot and box plot that water sources in section 5 have lower amount of radium228.

We do have some outliers . The farthest outlier in Box plot under section 4 shows that it has the water source with highest amount of radium228.

As we can see from boxplot except section 5 and section 6 all other section have atleast one water source with zero amount of radium228.

Problem 4

Install the maps and mapproj packages (you do not need to load them) 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).

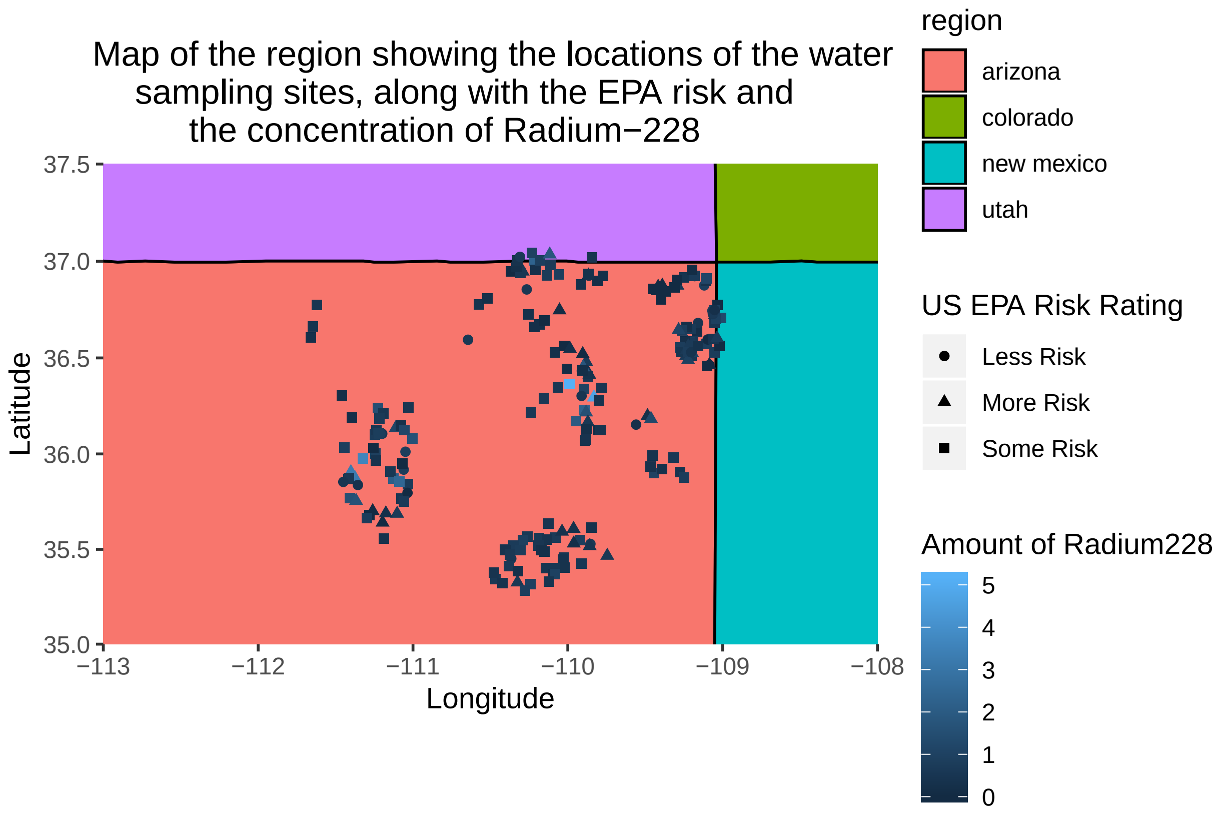
Install the measurements package and use the measurements::conv_unit() function to convert the latitude and longitude information in the dataset to decimal degrees suitable to be used for plotting.

Create a map of the region(you may want to adjust the plotting limits to an appropriate “zoom” level) showing the locations of the water sampling sites, along with the EPA risk and the concentration of Radium-228 for each location (mapped to an appropriate aesthetic).

Code

```
four_corners <- map_data("state",
                         region=c("arizona","new mexico", "utah","colorado"))
df4 <- df2
df4 <- mutate (df4, Longitude = -1*as.numeric(measurements::conv_unit(Longitude,
from = 'deg_min_sec', to = 'dec_deg')),Latitude = as.numeric(measurements::conv_unit
(Latitude,from = 'deg_min_sec', to = 'dec_deg')))

ggplot(four_corners) +
  geom_polygon(mapping=aes(x=long,y=lat,group=group,fill=region),color="black")+
    geom_point(df4,mapping = aes(x = Longitude, y = Latitude,
                          color =`Amount of Radium228`,shape = `US EPA Risk Rating`)) +
labs(title = "Map of the region showing the locations of the water
sampling sites, along with the EPA risk and \n the concentration of Radium-228",
                                                    x= "Longitude", y = "Latitude") +
                                            theme(plot.title = element_text(hjust = 0.3))+
                                             coord_map(xlim=c(-113,-108) , ylim=c(35,37.5))
```



Conclusion

We can see that most of water sampling sites are present in arizona , and few of them are in new mexico and utah.

As we can see from the map that most of water sources have amount of radium228 ranging from 0 to 2.

Part C

Problem 5

We would like to investigate whether Blacks tudents receive a disproportionate number of in-school suspensions.

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

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

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. What do you observe in the plot? Does the plot indicate an over- or under-representation of Black students in in-school suspensions?

Calculate the overall proportion of Black students across all schools and the overall proportion of suspended students who are Black across all schools. Are Black students over- or under-represented in in-school suspensions?

Code

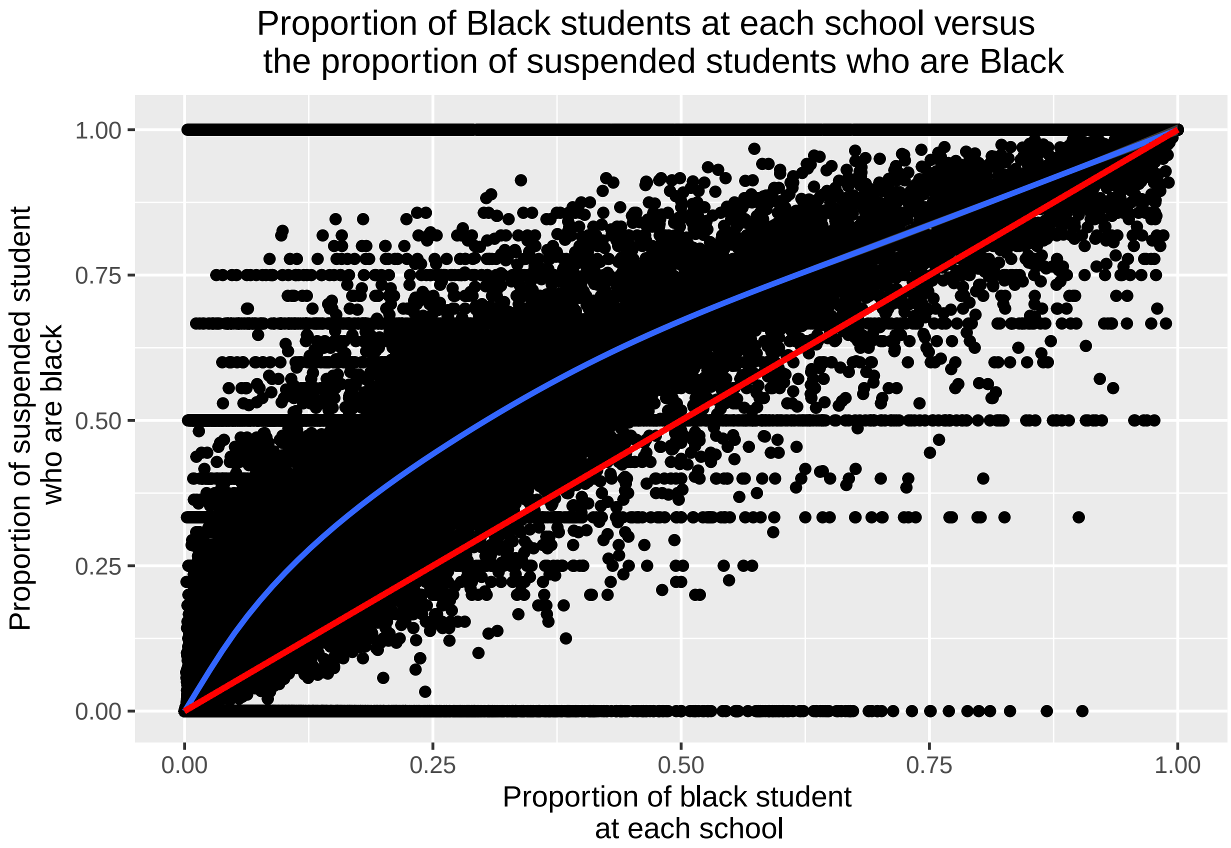
```r
dataframe <- read_csv("E:/FALL_2019/DS-5110-DM/HW2-DM/CRDC 2015-16 School Data.csv",
                      na = c("-2","-5","-6","-7","-8","-9"))
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   LEA_STATE = col_character(),
##   LEA_STATE_NAME = col_character(),
##   LEA_NAME = col_character(),
##   SCH_NAME = 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(),
##   SCH_GRADE_G12 = col_character(),
##   SCH_GRADE_UG = col_character()
##   # ... with 65 more columns
## )
```

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

```
new_df <- transmute(dataframe, TOT_ENR = TOT_ENR_M + TOT_ENR_F,
                        SCH_ENR_BL = SCH_ENR_BL_M + SCH_ENR_BL_F,
                        TOT_ISS = TOT_DISCWODIS_ISS_M + TOT_DISCWODIS_ISS_F+
                                  TOT_DISCWDIS_ISS_IDEA_M + TOT_DISCWDIS_ISS_IDEA_F,
                        SCH_ISS_BL = SCH_DISCWODIS_ISS_BL_M + SCH_DISCWODIS_ISS_BL_F +
                                  SCH_DISCWDIS_ISS_IDEA_BL_M + SCH_DISCWDIS_ISS_IDEA_BL_F,
                        PROP_BL = SCH_ENR_BL/TOT_ENR,
                        PROP_ISS_BL = SCH_ISS_BL/TOT_ISS)
ggplot(new_df, aes(x = PROP_BL, y = PROP_ISS_BL)) + geom_point() + geom_smooth() +
  labs(title = "Proportion of Black students at each school versus
  the proportion of suspended students who are Black", x= "Proportion of black student
  at each school", y = "Proportion of suspended student\n who are black") +
                                    theme(plot.title = element_text(hjust = 0.4)) +
  geom_segment(x=0,y=0,xend=1,yend=1,colour="red",size=1)
```

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



Conclusion

As we can see that our smoothing line(blue) is always over the straight line(red) with slope 1, plot indicates an over representation of Black students in in-school suspensions. As we can see from graph ,for almost every proportion of black student at each school there is more proportion of black student who are suspended in that school. For example, if we consider proportion of black student to be 0.25 we see that we have around 0.4 proportion of suspended student who were black student.

We also see a thick line parallel to X axis and passing throug Y=1. It tells that for some school irrespective of proportion of black student at that school , all suspended student were black.

```
OVR_BL = sum(new_df$SCH_ENR_BL, na.rm = TRUE)/sum(new_df$TOT_ENR, na.rm = TRUE)
OVR_BL_ISS = sum(new_df$SCH_ISS_BL, na.rm = TRUE)/sum(new_df$TOT_ISS, na.rm = TRUE)
cat("Overall proportion of black student across all school is", OVR_BL,"\n")
```

```
## Overall proportion of black student across all school is 0.1543446
```

```
cat("Overall proportion of suspended students who are Black across all schools",OVR_BL_ISS)
```

```
## Overall proportion of suspended students who are Black across all schools 0.3212122
```

Conclusion

Black student are over represented in in-school suspension. Because, as we see that overall proportion of black student across all school is only 0.1543446 ,but overall proportion of suspended students who are Black across all schools 0.3212122. That is even if black student are only 15 % of total student , there are 32% black student among all student that are suspended.