hw2-Sayan-Biswas

Sayan Biswas

31 January 2019

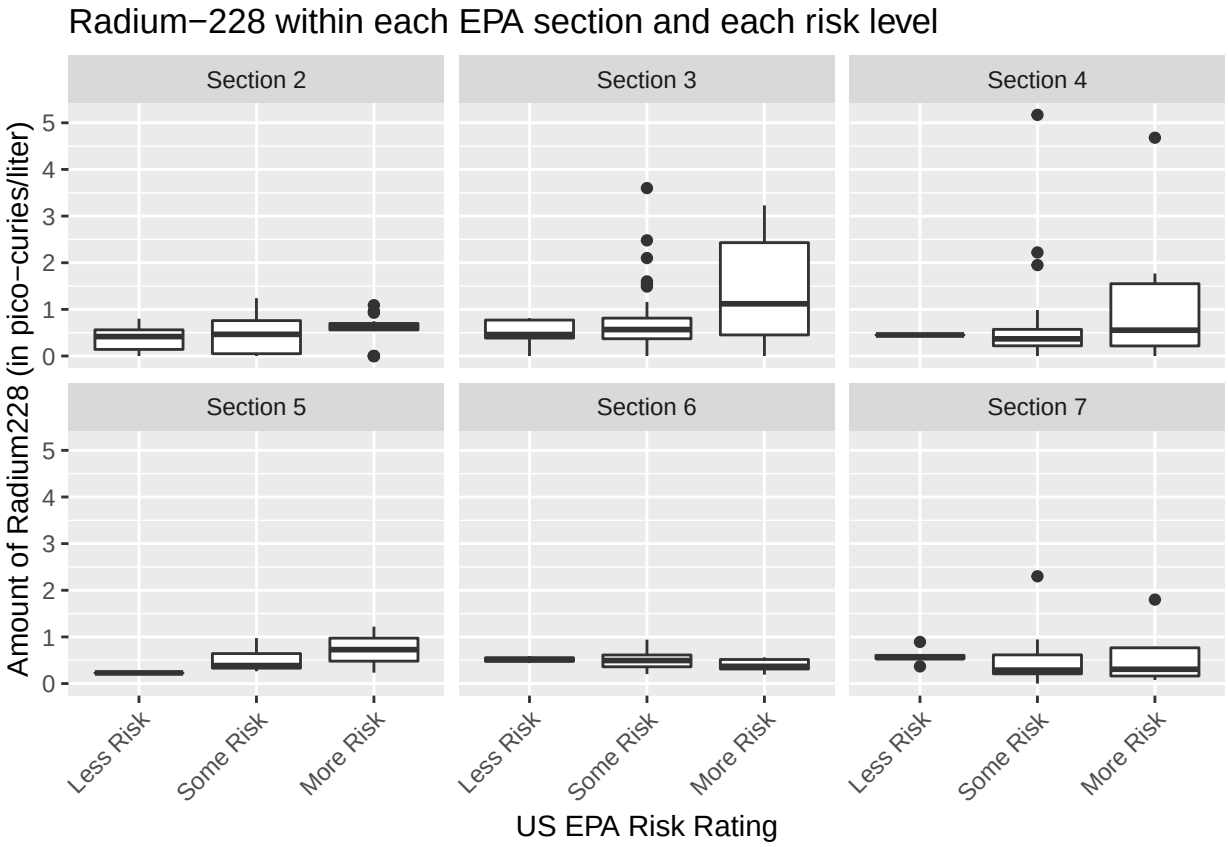
Part A

Problem 1

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

```
nvw_filtered <- navajowaterdata%>%
  mutate(`Amount of Radium228`=ifelse((`Amount of Radium228`<0),0,`Amount of Radium228`))%>%
  filter(`US EPA Risk Rating`!="Unknown Risk")%>%
  mutate(`US EPA Risk Rating`=factor(
    `US EPA Risk Rating`,
    levels=c("Less Risk","Some Risk","More Risk"),
    ordered = TRUE))

nvw_filtered %>%
  ggplot(mapping = aes(x=`US EPA Risk Rating`,
                       y=`Amount of Radium228`))+
  geom_boxplot()+
  facet_wrap(~`Which EPA Section is This From?`)+
  labs(x="US EPA Risk Rating",
       y="Amount of Radium228 (in pico-curies/liter)",
       title = "Radium-228 within each EPA section and each risk level")+
  theme(axis.text.x = element_text(angle=45, hjust = 1))
```



1. The plot indicates that Section 3 has a larger Radium228 distribution and also a higher median for Radium228 content than the other sections under the more risk section, which maybe the reason for it be more risky. Section 6 has the least amount of median radium content and maybe the less riskier.

2. Section 6 and Section 7 has a lower median Radium228 contents in them and maybe the presence of other radioacactive material is making the water of those sections more risky. The median radium content in more risk samples is less than the less risk samples for these two sections which somewhat supports the above statement.

3. For Section 4 and 5,the distribution of Radium 228 increases as the risk of the sample increases. Section 4 has one sampling site where the amount of Radium is greater than 5 pico-curies/liter under the some risk section and another sampling site where the amount of Radium is greater than 4.5 pico-curies/liter.

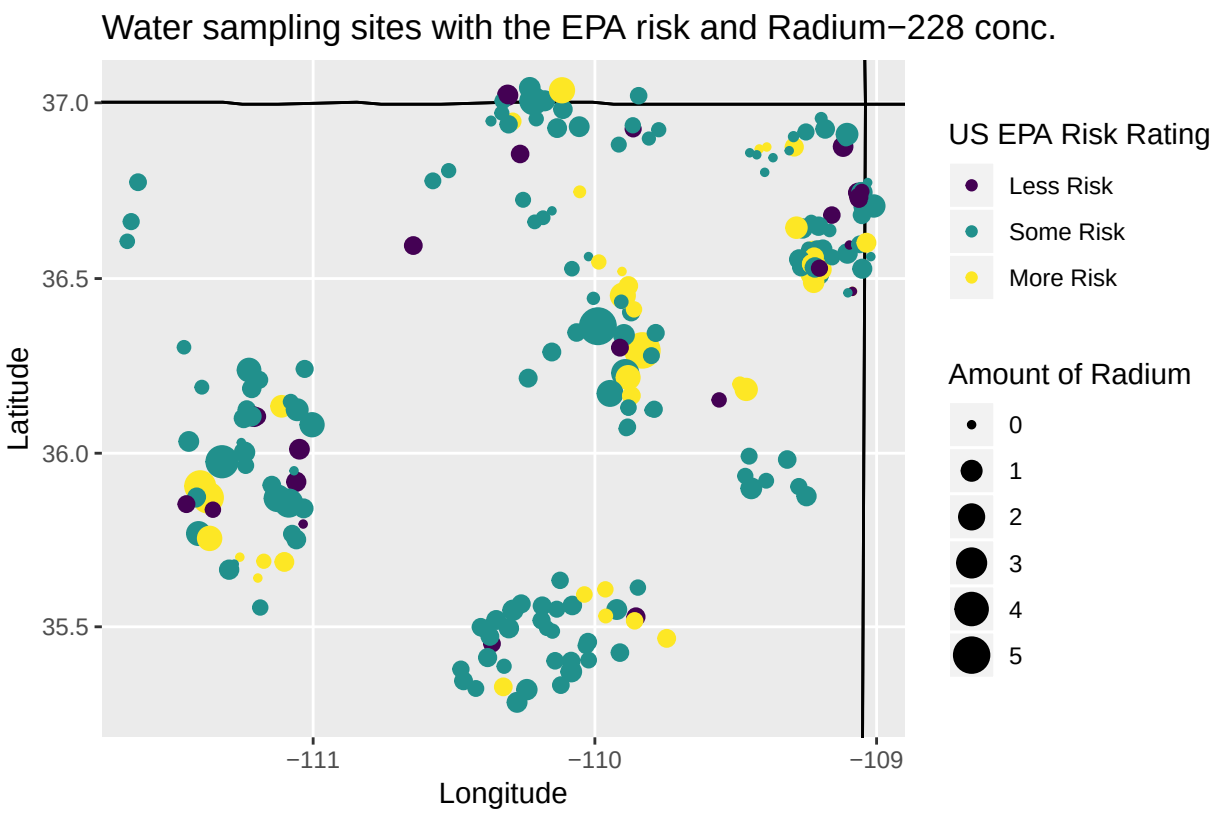
4. Section 2 has the least distribution of radium228 in the more risk samples.For Sections 2,3 and 5 the median amount of radium increases as the risk of the samples increases from some to more.

5. Section 3 has more outliers in the some risk section.

Problem 2

```
nvw_filtered <- nvw_filtered %>%
  mutate(
    Longitude=conv_unit(Longitude, "deg_min_sec", "dec_deg"),
    Latitude=conv_unit(Latitude, "deg_min_sec", "dec_deg"))%>%
  mutate(Longitude=-(as.numeric(Longitude)),
         Latitude=(as.numeric(Latitude)))

four_corners <- map_data("state",
                         region=c("arizona","new mexico","utah","colorado"))
ggplot(four_corners) +
  geom_polygon(mapping=aes(x=long,
                           y=lat,
                           group=group),
               fill=NA,
               color="black") +
  geom_point(data=nvw_filtered,
             mapping=aes(x=Longitude,
                         y=Latitude,
                         size=`Amount of Radium228`,
                         color=`US EPA Risk Rating`))+
  labs(x="Longitude",
       y="Latitude",
       size="Amount of Radium",
       color="US EPA Risk Rating",
       title = "Water sampling sites with the EPA risk and Radium-228 conc.")+
  coord_map(xlim = c(-111.75, -108.9),  ylim = c(35.18, 37.12))
```


Water sampling sites with the EPA risk and Radium−228 conc.
37.0
36.5
36.0
35.5
Latitude
−111
−110
−109
Longitude
US EPA Risk Rating
Less Risk
Some Risk
More Risk
Amount of Radium
0
1
2
3
4
5

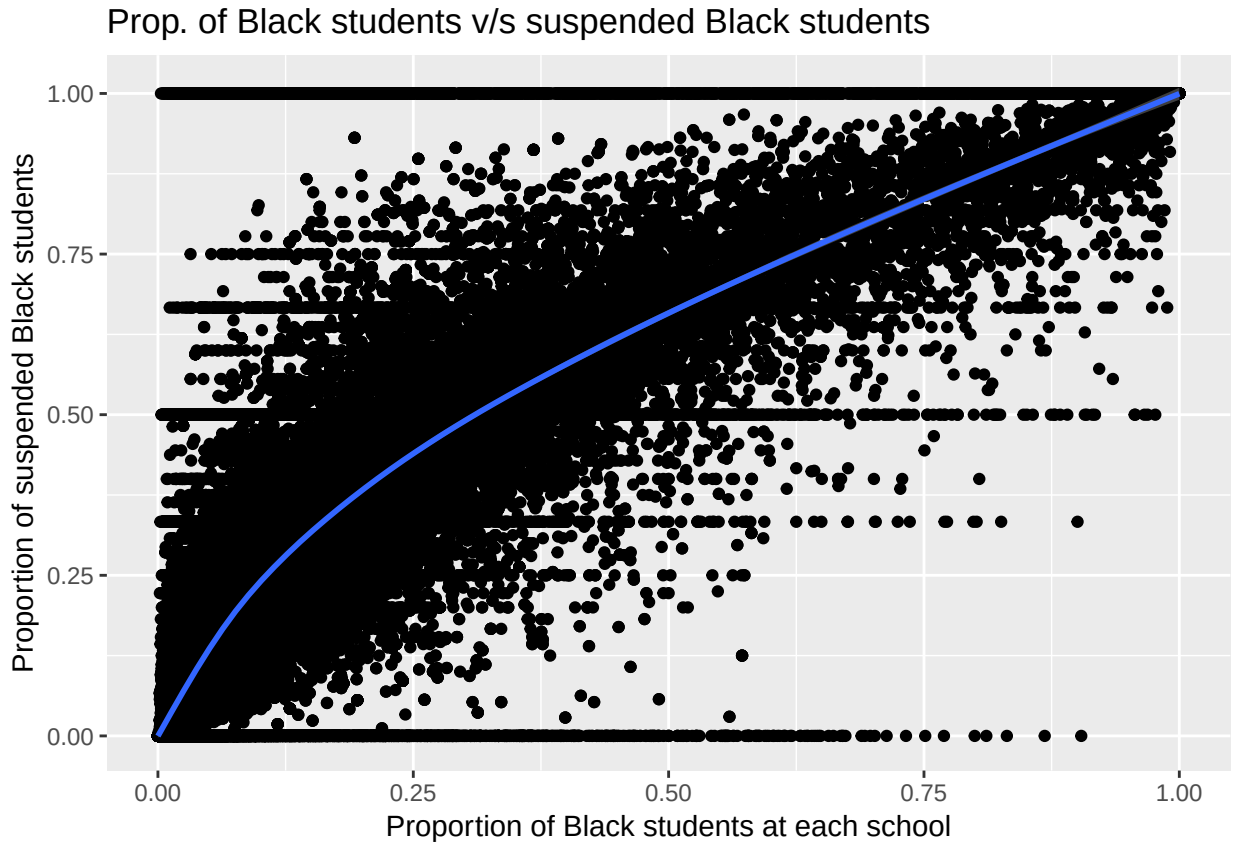
Part B

Problem 3

```
school<- read_csv("CRDC 2015-16 School Data.csv",
                  col_types = cols(.default = col_character()),
                  na=c("-2","-5","-6","-7","-8","-9"))
```

```
BL_ISS <- school%>%
  group_by(SCH_NAME)%>%
  transmute(
    TOT_ENR = sum(
      as.double(TOT_ENR_M),
      as.double(TOT_ENR_F),
      na.rm = TRUE),
    SCH_ENR_BL = sum(
      as.double(SCH_ENR_BL_M),
      as.double(SCH_ENR_BL_F),
      na.rm=TRUE),
    TOT_DISCW_ISS = sum(
      as.double(TOT_DISCWODIS_ISS_M),
      as.double(TOT_DISCWODIS_ISS_F),
      as.double(TOT_DISCWDIS_ISS_IDEA_M),
      as.double(TOT_DISCWDIS_ISS_IDEA_F),
      na.rm = TRUE),
    SCH_DIS_ISS_BL = sum(
      as.double(SCH_DISCWODIS_ISS_BL_M),
      as.double(SCH_DISCWODIS_ISS_BL_F),
      as.double(SCH_DISCWDIS_ISS_IDEA_BL_M),
      as.double(SCH_DISCWDIS_ISS_IDEA_BL_F),
      na.rm = TRUE),
    PROP_BL_SCH = SCH_ENR_BL/TOT_ENR,
    PROP_BL_ISS = SCH_DIS_ISS_BL/TOT_DISCW_ISS)

BL_ISS%>%
  ggplot(aes(x=PROP_BL_SCH,y=PROP_BL_ISS))+
  geom_point()+
  geom_smooth()+
  labs(x = "Proportion of Black students at each school",
       y = "Proportion of suspended Black students",
       title = "Prop. of Black students v/s suspended Black students")
```



The plot shows that the proportion of suspended students who are black has a positive co-relation with proportion of black students at each school. The plot indicates an over-representation of Black students in in-school suspensions as the proportion of suspended Black students is greater than the proportion of Black students at each school.

```
BL_ISS%>%
  ungroup(SCH_NAME)%>%
  summarise("Overall prop. of Black students" = sum(SCH_ENR_BL)/sum(TOT_ENR),
            "Overall prop. of suspended Black students" =
              sum(SCH_DIS_ISS_BL)/sum(TOT_DISCW_ISS))
```

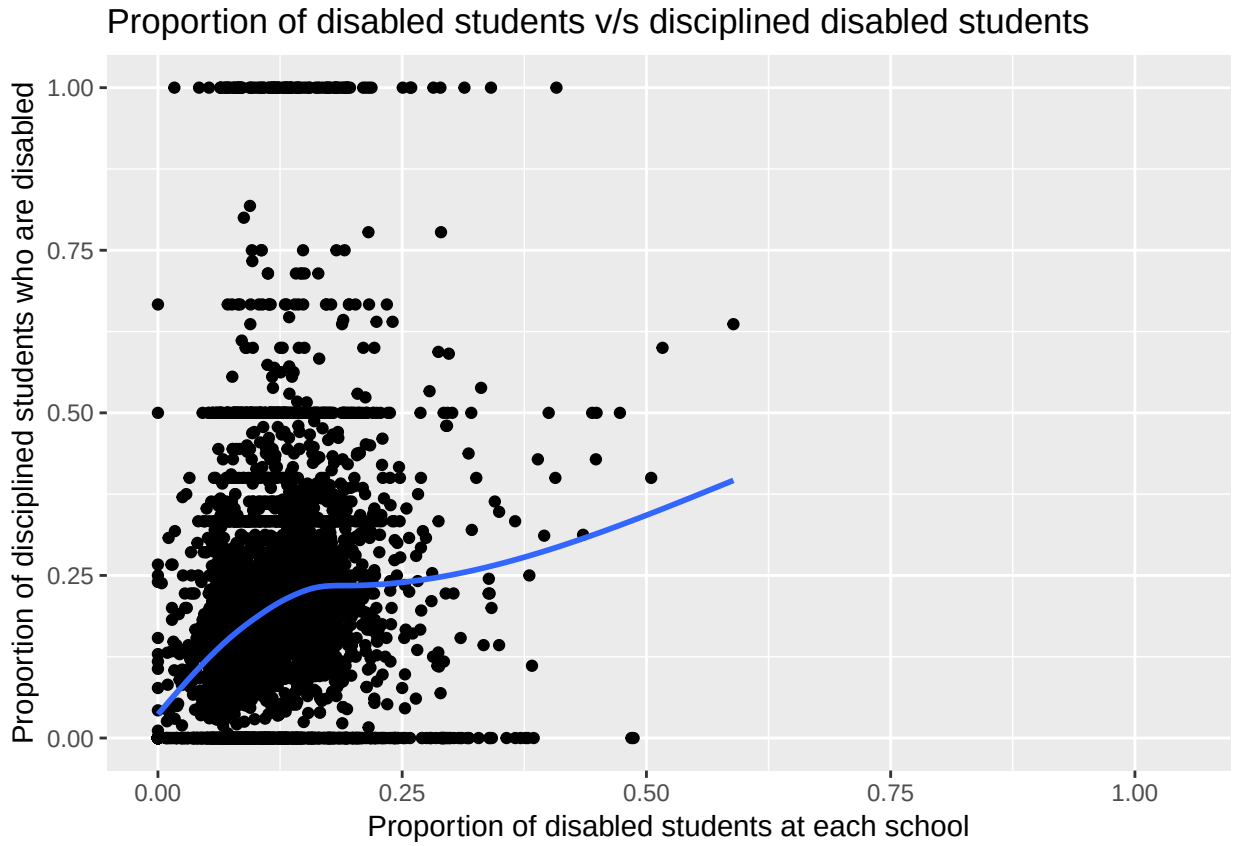
```
## # A tibble: 1 x 2
##   `Overall prop. of Black student~ `Overall prop. of suspended Black stude~
##                              <dbl>                                     <dbl>
## 1                            0.137                                     0.320
```

Since the overall proportion of suspended Black students is greater than the overall proportion of Black students, the black students are represented in a higher proportion in the sub-population of suspended students. Hence the Black students are over represented in in-school suspensions.

Problem 4

```
DIS_CORP <- school%>%
  filter(SCH_CORPINSTANCES_IND=="Yes")%>%
  group_by(SCH_NAME)%>%
  transmute(
    TOT_ENR = sum(
      as.double(TOT_ENR_M),
      as.double(TOT_ENR_F),
      na.rm = TRUE),
    TOT_IDEAENR = sum(
      as.double(TOT_IDEAENR_M),
      as.double(TOT_IDEAENR_F),
      na.rm = TRUE),
    TOT_CORP = sum(
      as.double(TOT_DISCWODIS_CORP_M),
      as.double(TOT_DISCWODIS_CORP_F),
      as.double(TOT_DISCWDIS_CORP_IDEA_M),
      as.double(TOT_DISCWDIS_CORP_IDEA_F),
      as.double(TOT_PSDISC_CORP_M),
      as.double(TOT_PSDISC_CORP_F),
      na.rm = TRUE),
    TOT_DISCWDIS_CORP_IDEA = sum(
      as.double(TOT_DISCWDIS_CORP_IDEA_M),
      as.double(TOT_DISCWDIS_CORP_IDEA_F),
      na.rm = TRUE),
    PROP_DIS_SCH = TOT_IDEAENR/TOT_ENR,
    PROP_DISC_DIS=TOT_DISCWDIS_CORP_IDEA/TOT_CORP)

DIS_CORP%>%
  ggplot(aes(x=PROP_DIS_SCH,y=PROP_DISC_DIS))+
  geom_point()+
  geom_smooth(se=FALSE)+
  labs(x = "Proportion of disabled students at each school",
       y = "Proportion of disciplined students who are disabled",
       title="Proportion of disabled students v/s disciplined disabled students")
```



The plot shows a positive co-relation between proportion of disabled students at each school and the proportion of disciplined students who are disabled. The plot indicates an over-representation of disabled students among students who are disciplined with corporal punishment when the proportion of disabled students at each school is close to 0.19. Post this when the proportion of disabled students at each school increases then the plot indicates an under-representation of disabled students among students who are disciplined with corporal punishment.

Since the majority of the data points lie when the proportion of disabled students is less than 0.19 and there are very less school with prop. of disabled students greater than 0.25, we can consider that the plot indicates an over-representation of disabled students among students who are disciplined with corporal punishment.

```
DIS_CORP%>%
  ungroup(SCH_NAME)%>%
  summarise("Overall prop. of disabled students" =
              sum(TOT_IDEAENR)/sum(TOT_ENR),
            "Overall prop. of disciplined disabled students" =
              sum(TOT_DISCWDIS_CORP_IDEA)/sum(TOT_CORP))
```

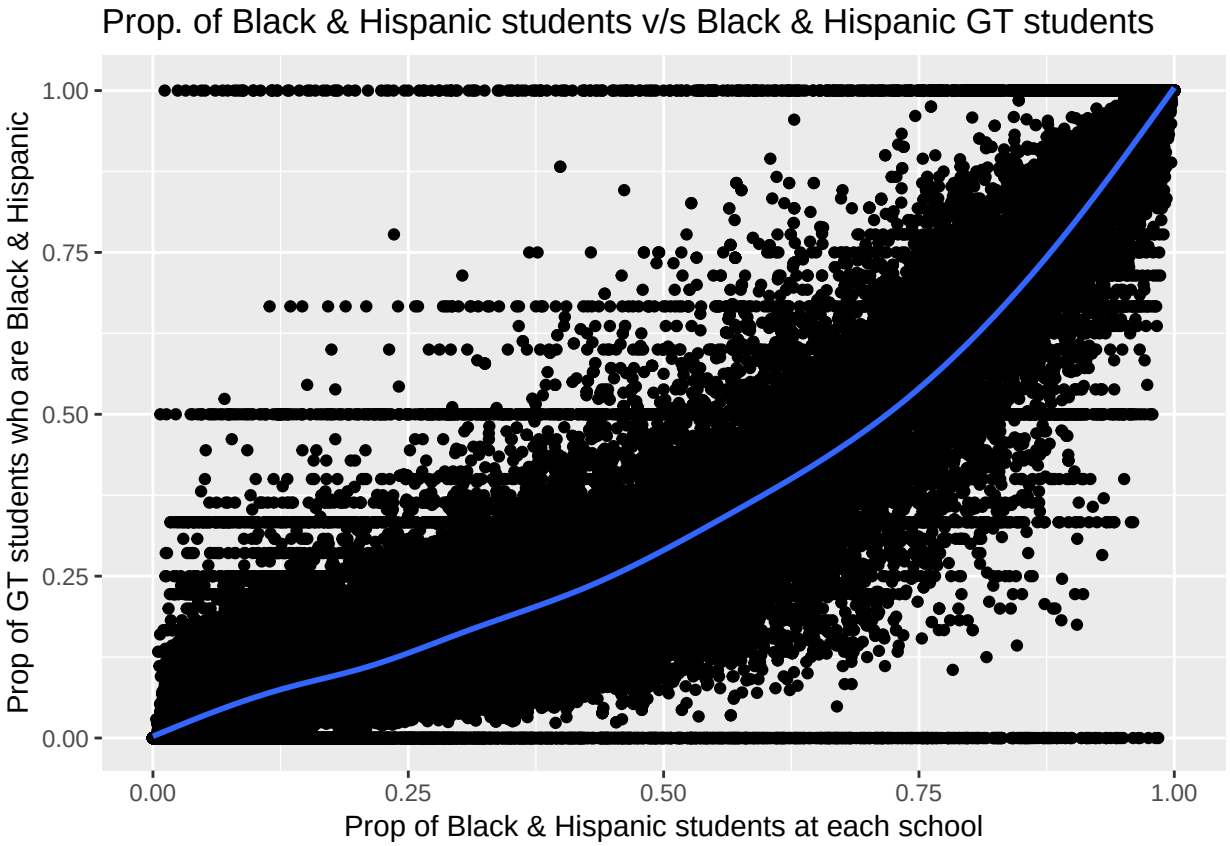
```
## # A tibble: 1 x 2
##   `Overall prop. of disabled stud~ `Overall prop. of disciplined disabled ~
##                              <dbl>                                    <dbl>
## 1                            0.118                                    0.171
```

Since the overall proportion of disciplined students who are disabled is slightly greater than the overall proportion of disabled students across all schools, the disabled students are slightly over-represented in the sub population of disciplined students.

Problem 5

```
BH_GT <- school%>%
  filter(SCH_GT_IND=="Yes")%>%
  group_by(SCH_NAME)%>%
  transmute(
    TOT_ENR = sum(
      as.double(TOT_ENR_M),
      as.double(TOT_ENR_F),
      na.rm = TRUE),
    SCH_ENR_HI_BL = sum(
      as.double(SCH_ENR_HI_M),
      as.double(SCH_ENR_HI_F),
      as.double(SCH_ENR_BL_M),
      as.double(SCH_ENR_BL_F),
      na.rm = TRUE),
    TOT_GTENR = sum(
      as.double(TOT_GTENR_M),
      as.double(TOT_GTENR_F),
      na.rm=TRUE),
    SCH_GTENR_HI_BL = sum(
      as.double(SCH_GTENR_HI_M),
      as.double(SCH_GTENR_HI_F),
      as.double(SCH_GTENR_BL_M),
      as.double(SCH_GTENR_BL_F),
      na.rm=TRUE),
    PROP_BL_HI = SCH_ENR_HI_BL/TOT_ENR,
    PROP_BL_HI_GT=SCH_GTENR_HI_BL/TOT_GTENR)

BH_GT%>%
  ggplot(aes(x=PROP_BL_HI,y=PROP_BL_HI_GT))+
  geom_point()+
  geom_smooth(se=FALSE)+
  labs(x = "Prop of Black & Hispanic students at each school",
       y = "Prop of GT students who are Black & Hispanic",
       title = "Prop. of Black & Hispanic students v/s Black & Hispanic GT students")
```



The plot shows that the proportion of Gifted and Talented students who are Black & Hispanic has a positive co-relation with proportion of Black and Hispanic students at each school. The plot indicates an under-representation of Black and Hispanic students in in Gifted & Talented program as the proportion of proportion of Gifted and Talented students who are Black & Hispanic is less than the proportion of Black and Hispanic students at each school.

```
BH_GT%>%
  ungroup(SCH_NAME)%>%
  summarise("Overall prop of Black & Hispanic" =
              sum(SCH_ENR_HI_BL)/sum(TOT_ENR),
            "Overall Prop of Black & Hispanic GT" =
              sum(SCH_GTENR_HI_BL)/sum(TOT_GTENR))
```

```
## # A tibble: 1 x 2
##   `Overall prop of Black & Hispanic` `Overall Prop of Black & Hispanic GT`
##                                <dbl>                                 <dbl>
## 1                              0.425                                 0.267
```

Since the overall proportion of Black and Hispanic students under the Gifted & Talented program is less than the overall proportion of Black and Hispanic students across all schools, the Black and Hispanic students are represented in a lower proportion in the sub-population of Gifted and Talented students. Hence the Black and Hispanic students are under-represented in Gifted & Talented program.