HW2-Aishwarya Vantipuli

Aishwarya

January 31, 2019

```
#install.packages("maps")
#install.packages("mapproj")
#install.packages("measurements")
```

PART-A

PROBLEM 1

```
data <- read_csv("I:/Data Science/NEU/SEM1/Introduction to Data management/hw2/NavajoWaterExport.csv" )
```

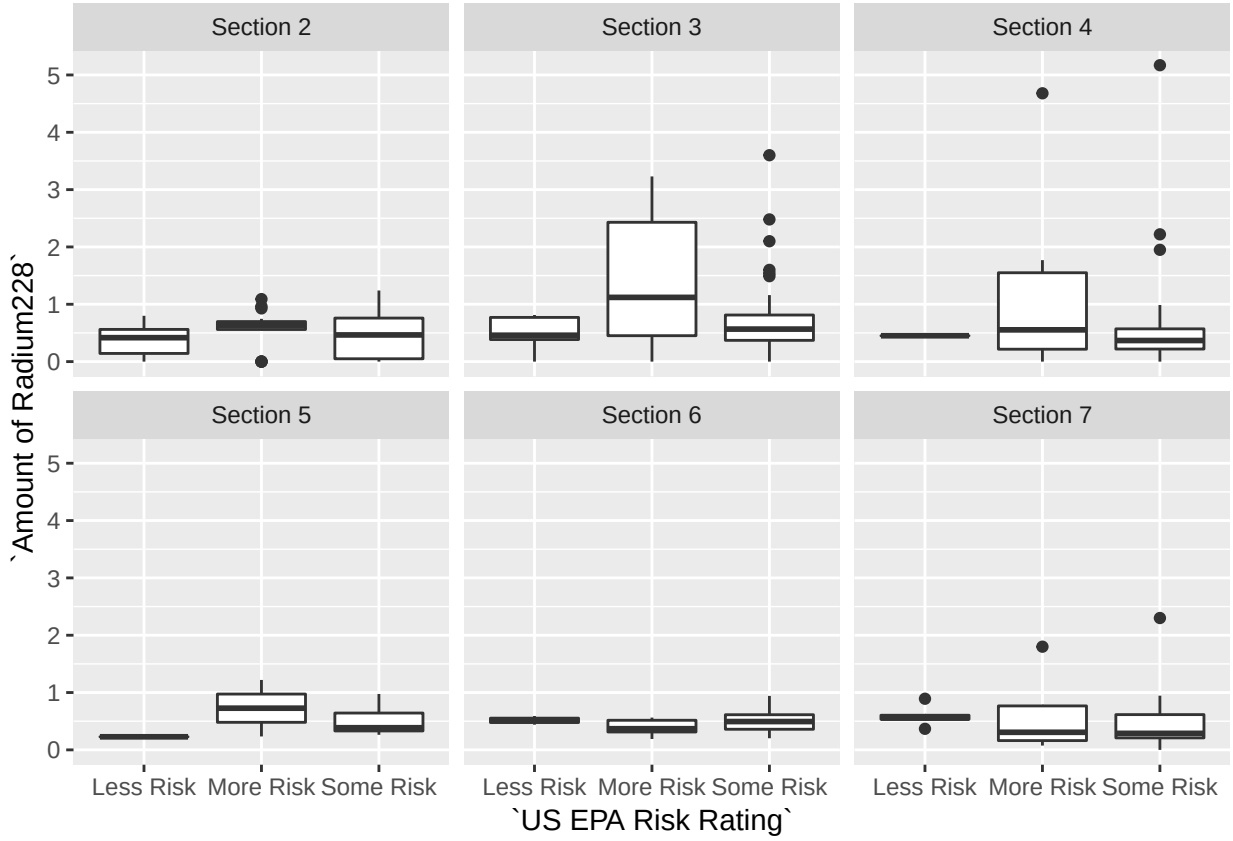
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   `Amount of Aluminum (Al)` = col_number(),
##   `Amount of Antimony (Sb)` = col_double(),
##   `Amount of Arsenic (As)` = col_double(),
##   `Amount of Barium (Ba)` = col_number(),
##   `Amount of Beryllium (Be)` = 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.
```

```
#Removing Negative Values in Amount of Radium228 Coloumn
old_col <- data$`Amount of Radium228`
new_col <- ifelse(old_col < 0, 0, old_col)
data <- mutate(data,'Amount of Radium228'= new_col)

#Printing head
head(data$`Amount of Radium228`)
```

```
## [1] 0.500 1.540 0.591 0.183 0.439 0.892
```

```
#Removing row containing 'Unknown Risk'
data <- data[-(which(data$`US EPA Risk Rating` == 'Unknown Risk')), ]

#Plotting Boxplots
ggplot(data = data, mapping = aes( x =`US EPA Risk Rating`, y =`Amount of Radium228`)) + geom_boxplot()
```



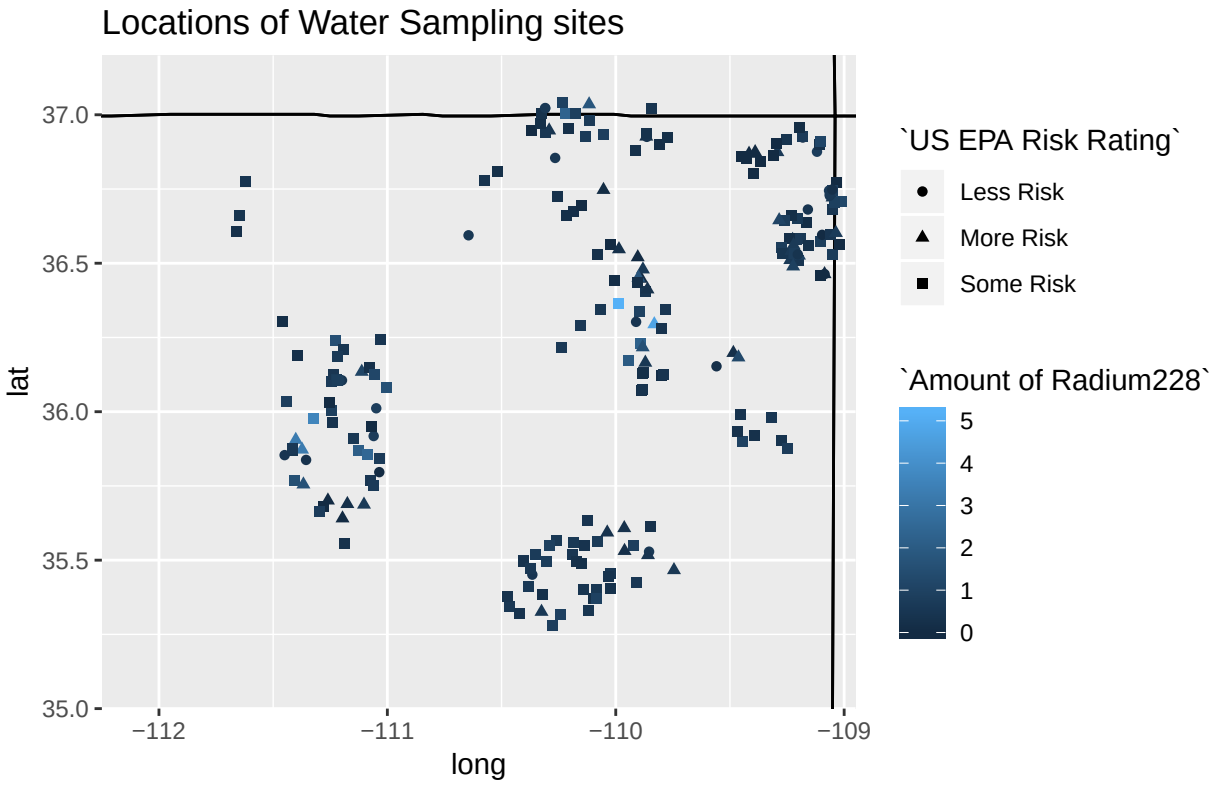
OBSERVATION:-

Section 3,4,7 have comparatively more amount of radium228 which are of higher risk than other sections.

PROBLEM - 2

```
#PLOTING FOUR CORNERS
four_corners <- ggplot2::map_data("state", region = c("arizona", "New Mexico", "Utah", "Colorado"))

#CONVERTING LAT AND LONG DATA
data$Longitude <- measurements::conv_unit(data$Longitude, "deg_min_sec", "dec_deg")
data$Latitude <- measurements::conv_unit(data$Latitude, "deg_min_sec", "dec_deg")

#PLOTING POINTS ON FOUR CORNERS
ggplot(four_corners) +
  geom_polygon(mapping=aes(x=long,
                           y=lat,
                           group = group),
               fill=NA,
```

```
              color="black") +
 geom_point(data, mapping = aes(x = -as.double(Longitude),
                                y = as.double(Latitude),
                                shape = `US EPA Risk Rating`,
                                color = `Amount of Radium228`)) +
#ZOOMING
  coord_fixed(xlim = c(-112.1, -109.1),  ylim = c(35.1,37.1), ratio = 1.3) +
  ggtitle("Locations of Water Sampling sites")
```



PART-B

PROBLEM-3

```
school <- read_csv("I:/Data Science/NEU/SEM1/Introduction to Data management/hw2/Data Files and Layouts/
```

```
## 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 42 more columns
## )

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

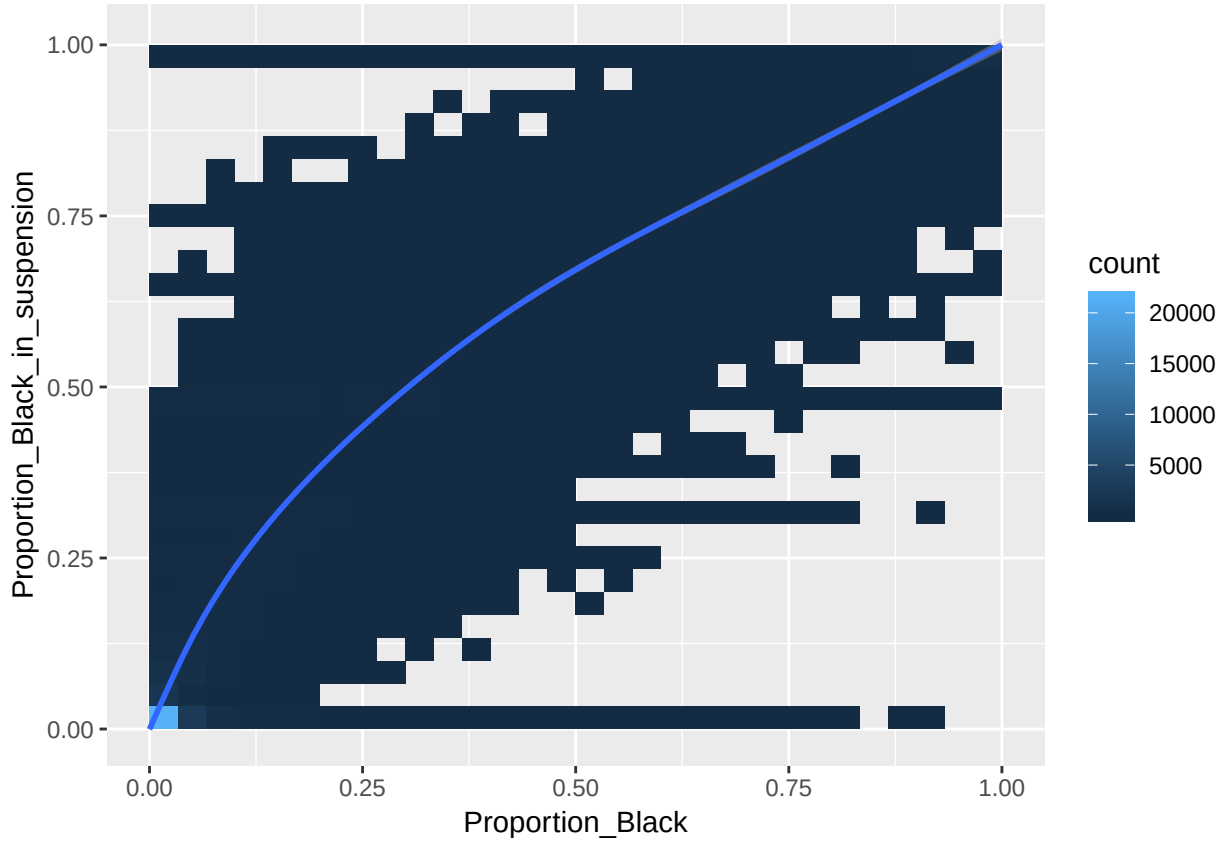
```
#CHECKING FOR PRESCENCE OF RESERVE CODES
any(school$TOT_ENR_F < 0)
```

```
## [1] TRUE
```

```
#REPLACING EACH REQUIRIED COLOUMN CONTAINING RESERVE CODES WITH NA
school$TOT_ENR_M<-   ifelse(school$TOT_ENR_M < 0, NA, school$TOT_ENR_M)
school$TOT_ENR_F<-   ifelse(school$TOT_ENR_F < 0, NA, school$TOT_ENR_F)


school$SCH_ENR_BL_F<-   ifelse(school$SCH_ENR_BL_F< 0, NA, school$SCH_ENR_BL_F)
school$SCH_ENR_BL_M<-   ifelse(school$SCH_ENR_BL_M< 0, NA, school$SCH_ENR_BL_M)


school$TOT_DISCWDIS_ISS_IDEA_F <-   ifelse(school$TOT_DISCWDIS_ISS_IDEA_F < 0, NA, school$TOT_DISCWDIS_I
school$TOT_DISCWDIS_ISS_IDEA_M<-   ifelse(school$TOT_DISCWDIS_ISS_IDEA_M < 0, NA, school$TOT_DISCWDIS_IS
school$TOT_DISCWODIS_ISS_M <-   ifelse(school$TOT_DISCWODIS_ISS_M < 0, NA, school$TOT_DISCWODIS_ISS_M )
school$TOT_DISCWODIS_ISS_F <-   ifelse(school$TOT_DISCWODIS_ISS_F < 0, NA, school$TOT_DISCWODIS_ISS_F)


school$SCH_DISCWODIS_ISS_BL_M <-   ifelse(school$SCH_DISCWODIS_ISS_BL_M < 0, NA, school$SCH_DISCWODIS_IS
school$SCH_DISCWODIS_ISS_BL_F <-   ifelse(school$SCH_DISCWODIS_ISS_BL_F < 0, NA, school$SCH_DISCWODIS_IS
school$SCH_DISCWDIS_ISS_IDEA_BL_F <-   ifelse(school$SCH_DISCWDIS_ISS_IDEA_BL_F < 0, NA, school$SCH_DISC
school$SCH_DISCWDIS_ISS_IDEA_BL_M <-   ifelse(school$SCH_DISCWDIS_ISS_IDEA_BL_M < 0, NA, school$SCH_DISC


#CREATING NEW DATAFRAME AND INSERTING NEW COMPUTED VARIABLES
d1 <- data.frame(transmute(school,
      Total_students = TOT_ENR_M + TOT_ENR_F,

      Total_Black_students = SCH_ENR_BL_M + SCH_ENR_BL_F,

      Total_in_suspension = TOT_DISCWDIS_ISS_IDEA_M + TOT_DISCWDIS_ISS_IDEA_F +TOT_DISCWODIS_ISS_M + TOT

      Total_Black_in_suspension = SCH_DISCWODIS_ISS_BL_M + SCH_DISCWODIS_ISS_BL_F + SCH_DISCWDIS_ISS_IDE
```

```
      Proportion_Black = Total_Black_students/Total_students,

      Proportion_Black_in_suspension = Total_Black_in_suspension /Total_in_suspension))


#PLOTTING USING GEOM_BIN2D
ggplot(d1) + geom_bin2d(mapping = aes(x =Proportion_Black, y = Proportion_Black_in_suspension)) + geom_
```

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



OBSERVATION:-

The plot describes a positive relation between proportion of blacks and proprtion of suspended blacks but we can observe that at a certain point on x-axis there are HIGHER values of blacks in suspension which indicates an OVER REPRESENTATION of Blacks in school suspensions.

```
#CALCULATING OVERALL PROPORTIONS
overall_prop_black <- mean(d1$Proportion_Black, na.rm = TRUE)
overall_prop_black_susp  <- mean(d1$Proportion_Black_in_suspension, na.rm = TRUE)

overall_prop_black
```

```
## [1] 0.1537564
```

```
overall_prop_black_susp
```

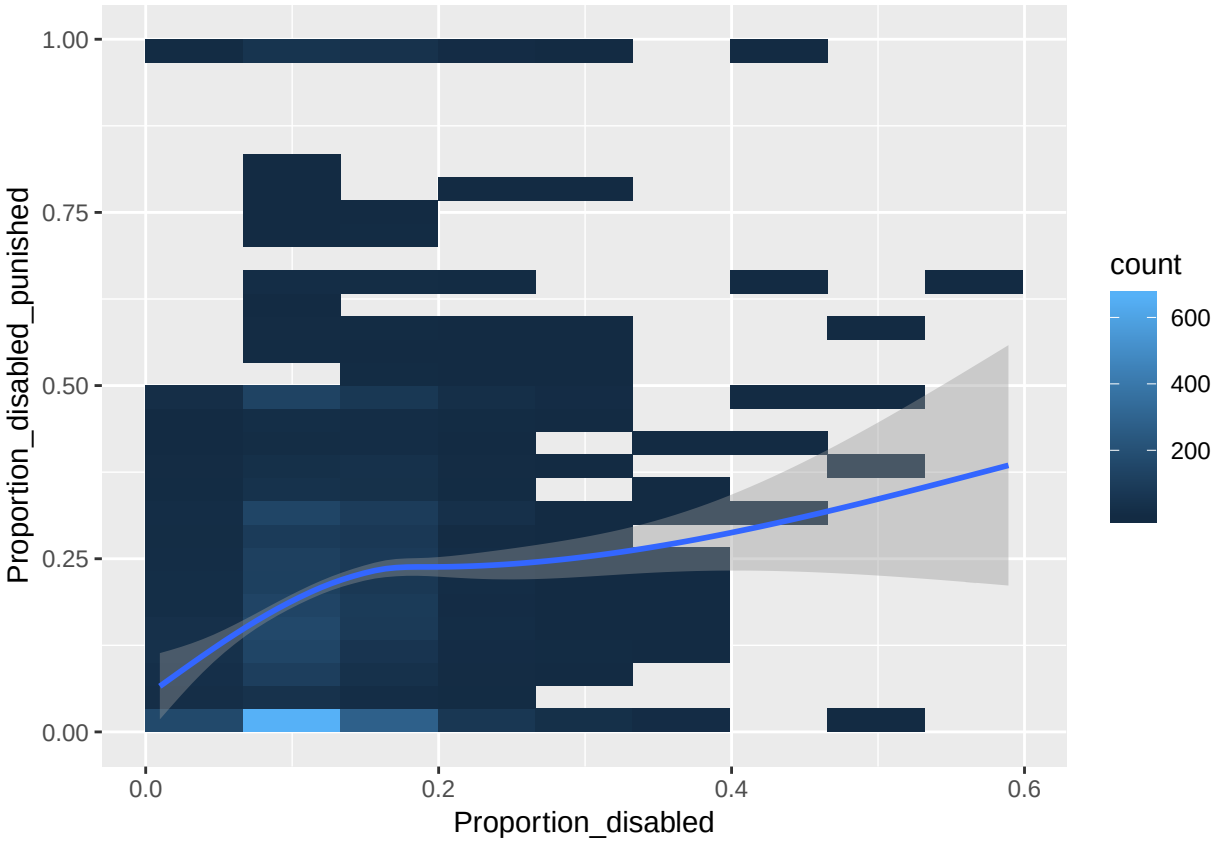
```
## [1] 0.2384816
```

OBSERVATION:- From calculations we can say that black students are OVER REPRESENTED.

PROBLEM-4

```
#REPLACING EACH REQUIRIED COLOUMN CONTAINING RESERVE CODES WITH NA
school$TOT_IDEAENR_M <-   ifelse(school$TOT_IDEAENR_M < 0, NA, school$TOT_IDEAENR_M)
school$TOT_IDEAENR_F <-   ifelse(school$TOT_IDEAENR_F < 0, NA, school$TOT_IDEAENR_F)


school$TOT_DISCWODIS_CORP_M <-   ifelse(school$TOT_DISCWODIS_CORP_M < 0, NA, school$TOT_DISCWODIS_CORP_
school$TOT_DISCWODIS_CORP_F <-   ifelse(school$TOT_DISCWODIS_CORP_F < 0, NA, school$TOT_DISCWODIS_CORP_
school$TOT_DISCWDIS_CORP_IDEA_M <-   ifelse(school$TOT_DISCWDIS_CORP_IDEA_M < 0, NA, school$TOT_DISCWDI
school$TOT_DISCWDIS_CORP_IDEA_F <-   ifelse(school$TOT_DISCWDIS_CORP_IDEA_F < 0, NA, school$TOT_DISCWDI

#CREATING NEW DATAFRAME AND INSERTING NEW COMPUTED VARIABLES
d2 <- data.frame(transmute(school,
      Total_students = TOT_ENR_M + TOT_ENR_F,

      Total_disable_students = TOT_IDEAENR_M + TOT_IDEAENR_F,

      Total_students_punishment =TOT_DISCWODIS_CORP_M+ TOT_DISCWODIS_CORP_F+TOT_DISCWDIS_CORP_IDEA_M+ T

      Disabled_punished = TOT_DISCWDIS_CORP_IDEA_M+ TOT_DISCWDIS_CORP_IDEA_F,

      Proportion_disabled = Total_disable_students/Total_students,

      Proportion_disabled_punished = Disabled_punished/ Total_students_punishment))


#PLOTTING USING GEOM_BIN2D
ggplot(d2) + geom_bin2d(mapping = aes(x =  Proportion_disabled, y = Proportion_disabled_punished )) + g
```

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



OBSERVATION:-

The relationship seems to be almost constant

```
#CALCULATING OVERALL PROPORTIONS
overall_prop_disabled = mean(d2$Proportion_disabled, na.rm = TRUE)
overall_prop_disabled_punished = mean(d2$Proportion_disabled_punished, na.rm = TRUE)


overall_prop_disabled
```

```
## [1] 0.1437109
```

```
overall_prop_disabled_punished
```

```
## [1] 0.1929799
```

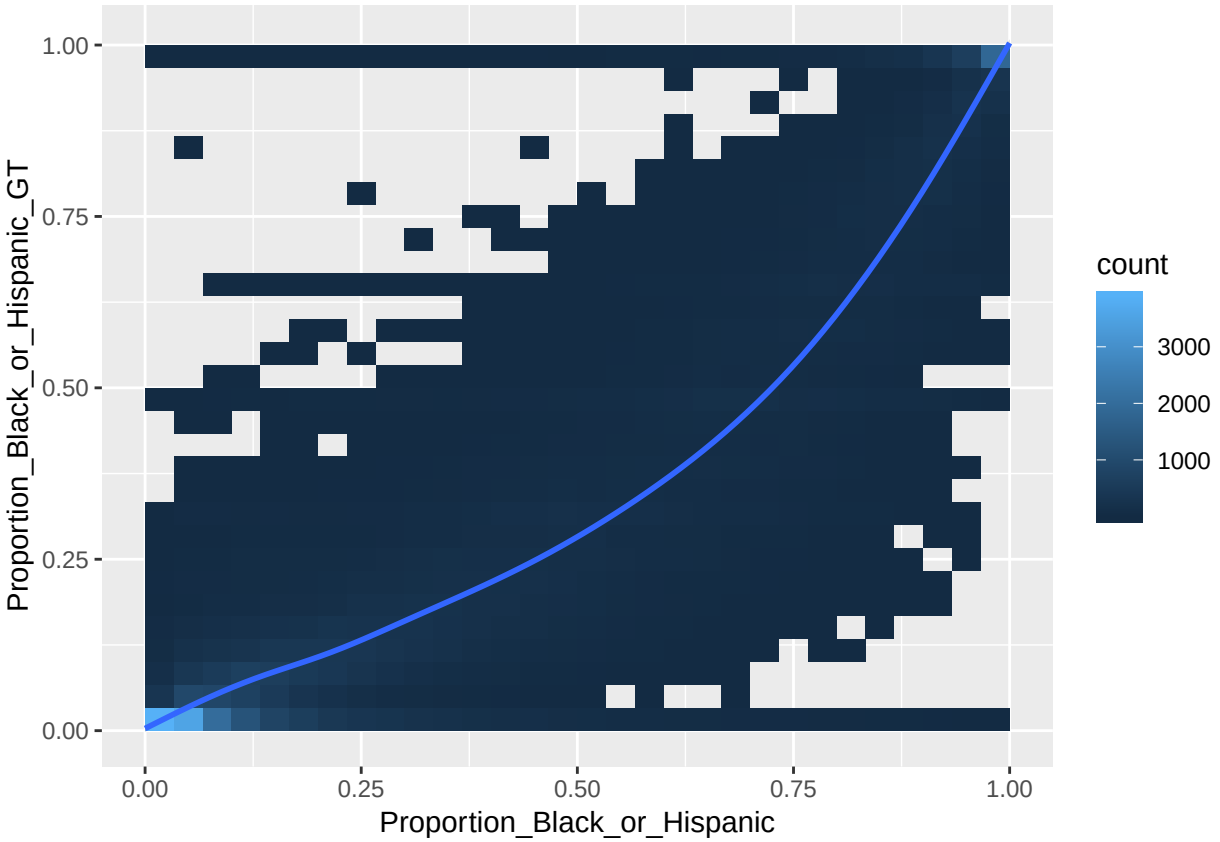
OBSERVATION:-

As said before, the variation slightly differs. so we can say that disabled students are equally treated with others when it comes to corporal punishment.

PROBLEM-5

```r
#REPLACING EACH REQUIRIED COLOUMN CONTAINING RESERVE CODES WITH NA
school$SCH_ENR_HI_M <-   ifelse(school$SCH_ENR_HI_M < 0, NA, school$SCH_ENR_HI_M)
school$SCH_ENR_HI_F <-   ifelse(school$SCH_ENR_HI_F < 0, NA, school$SCH_ENR_HI_F)


school$TOT_GTENR_M <-   ifelse(school$TOT_GTENR_M < 0, NA, school$TOT_GTENR_M)
school$TOT_GTENR_F <-   ifelse(school$TOT_GTENR_F < 0, NA, school$TOT_GTENR_F)

school$SCH_GTENR_BL_M <-   ifelse(school$SCH_GTENR_BL_M < 0, NA, school$SCH_GTENR_BL_M)
school$SCH_GTENR_BL_F <-   ifelse(school$SCH_GTENR_BL_F < 0, NA, school$SCH_GTENR_BL_F)

school$SCH_GTENR_HI_M <-   ifelse(school$SCH_GTENR_HI_M< 0, NA, school$SCH_GTENR_HI_M)
school$SCH_GTENR_HI_F <-   ifelse(school$SCH_GTENR_HI_F< 0, NA, school$SCH_GTENR_HI_F)


#CREATING NEW DATAFRAME AND INSERTING NEW COMPUTED VARIABLES
d3 <- data.frame(transmute(school,
      Total_students= TOT_ENR_M + TOT_ENR_F,
      Black_and_Hispanic = SCH_ENR_HI_M + SCH_ENR_HI_F + SCH_ENR_BL_M + SCH_ENR_BL_F,
      Total_students_GT = TOT_GTENR_M + TOT_GTENR_F,
      Black_or_Hispanic_GT= SCH_GTENR_BL_M + SCH_GTENR_BL_F+ SCH_GTENR_HI_M + SCH_GTENR_HI_F,
      Proportion_Black_or_Hispanic= Black_and_Hispanic/Total_students,
      Proportion_Black_or_Hispanic_GT= Black_or_Hispanic_GT/Total_students_GT))


#PLOTTING WITH GEOM_BIN2D
ggplot(d3) +
  geom_bin2d(mapping = aes(x =  Proportion_Black_or_Hispanic, y = Proportion_Black_or_Hispanic_GT)) +
  geom_smooth(mapping = aes(x =  Proportion_Black_or_Hispanic, y = Proportion_Black_or_Hispanic_GT))
```

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



OBSERVATION:-

Clearly, there is a positive relation but with negative slope between both proportions. At any point on Proportion of Black/Hispanic there exists corresponding point on Proportion of gifted Black/Hispanic with LOWER VALUE which is UNDER REPRESENTATION.

```
#CALCULATING OVERALL PROPORTIONS
overall_prop_B_H= mean(d3$Proportion_Black_or_Hispanic, na.rm = TRUE)
overall_prop_B_H_GT = mean(d3$Proportion_Black_or_Hispanic_GT, na.rm = TRUE)

overall_prop_B_H
```

```
## [1] 0.3785035
```

```
overall_prop_B_H_GT
```

```
## [1] 0.2794983
```

OBSERVATION:-

As, overall proportion of Gifted Black and Hispanic is LOWER than overall proportion of Black and Hispanic, hence it is UNDER REPRESENTATION.