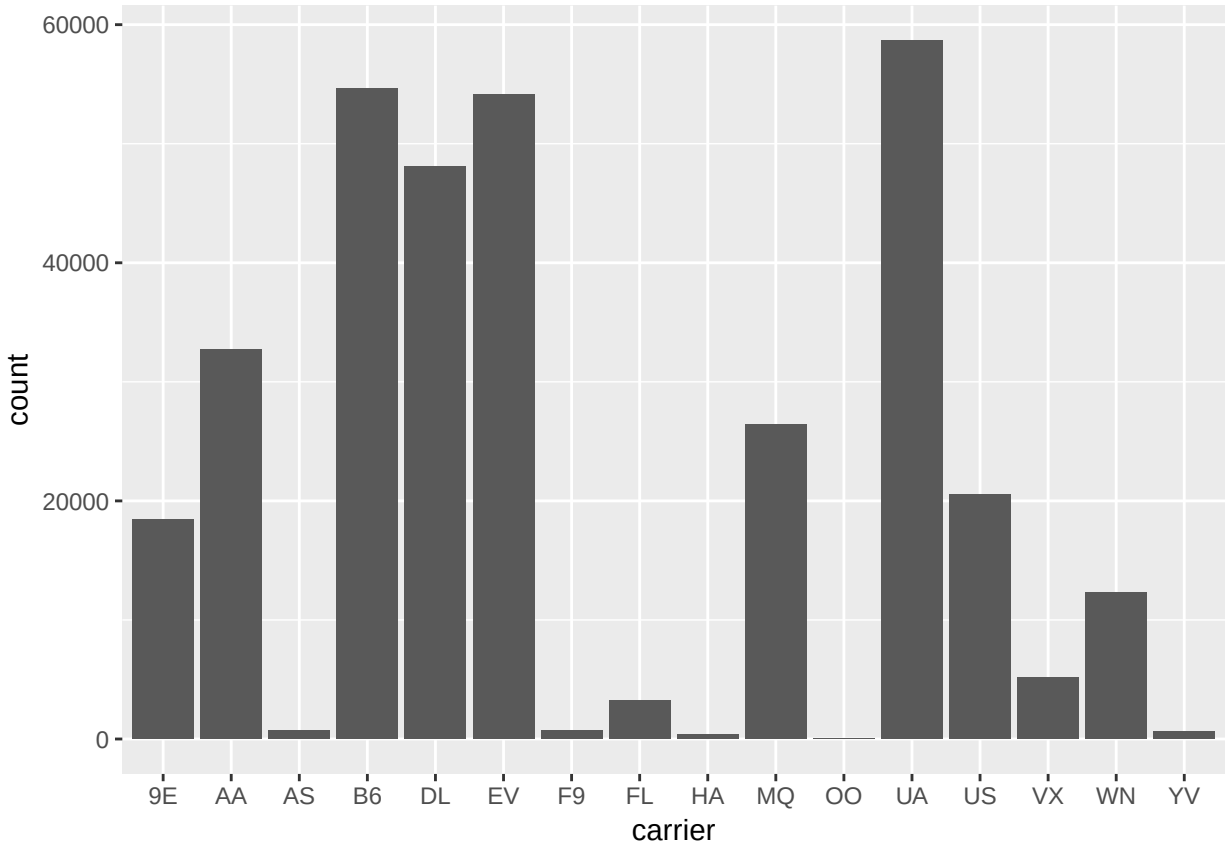
Assignment1

Question 1

```
library(tidyverse)
library(nycflights13)
ggplot(data=flights,mapping=aes(x=carrier))+
  geom_bar()
```



Based on the graph, we could know that UA flew the most flight.

Question 2

```
flights %>%
  group_by(dest) %>% summarise(prop_earlier=mean(arr_delay<0, na.rm=TRUE),
                               Median_distance=median(distance,na.rm=TRUE))
```

```
## # A tibble: 105 x 3
##     dest prop_earlier Median_distance
##    <chr>        <dbl>           <dbl>
##  1   ABQ    0.5708661            1826
##  2   ACK    0.5871212             199
##  3   ALB    0.5478469             143
##  4   ANC    0.3750000            3370
```

```
##  5   ATL    0.5072756             762
##  6   AUS    0.5810867            1521
##  7   AVL    0.5134100             583
##  8   BDL    0.6407767             116
##  9   BGR    0.6033520             378
## 10   BHM    0.5353160             866
## # ... with 95 more rows
```

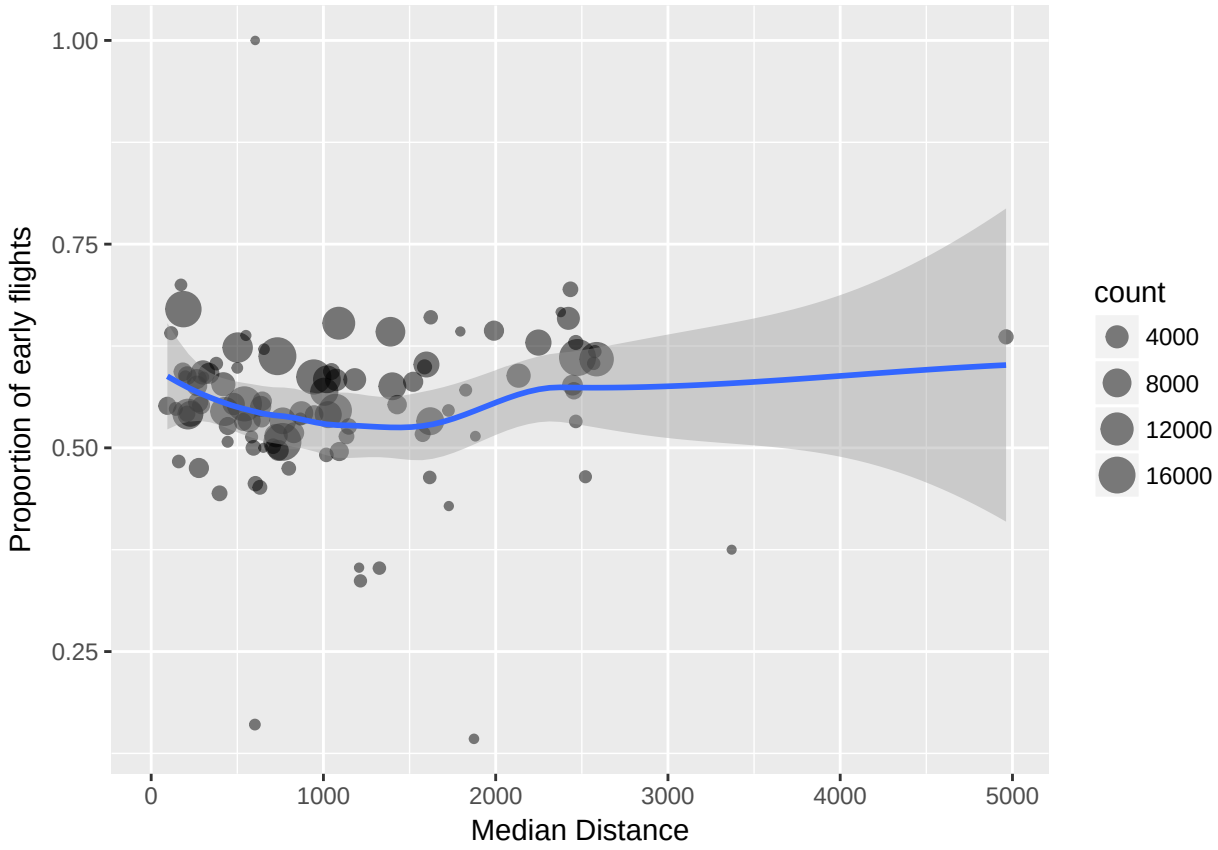
The table above shows, for each destination, the proportion of flights that arrived at their destination earlier than scheduled and the median distance flown to each destination.

```
flights %>%
  group_by(dest) %>%
  summarise(Prop_early = mean(arr_delay < 0, na.rm=TRUE),
            Median_Distance = median(distance, na.rm=TRUE),
            count = n()) %>%
  ggplot(aes(x=Median_Distance,y=Prop_early)) + geom_point(aes(size=count), alpha=1/2) +
  geom_smooth()+
  xlab("Median Distance")+ylab("Proportion of early flights")
```

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

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

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

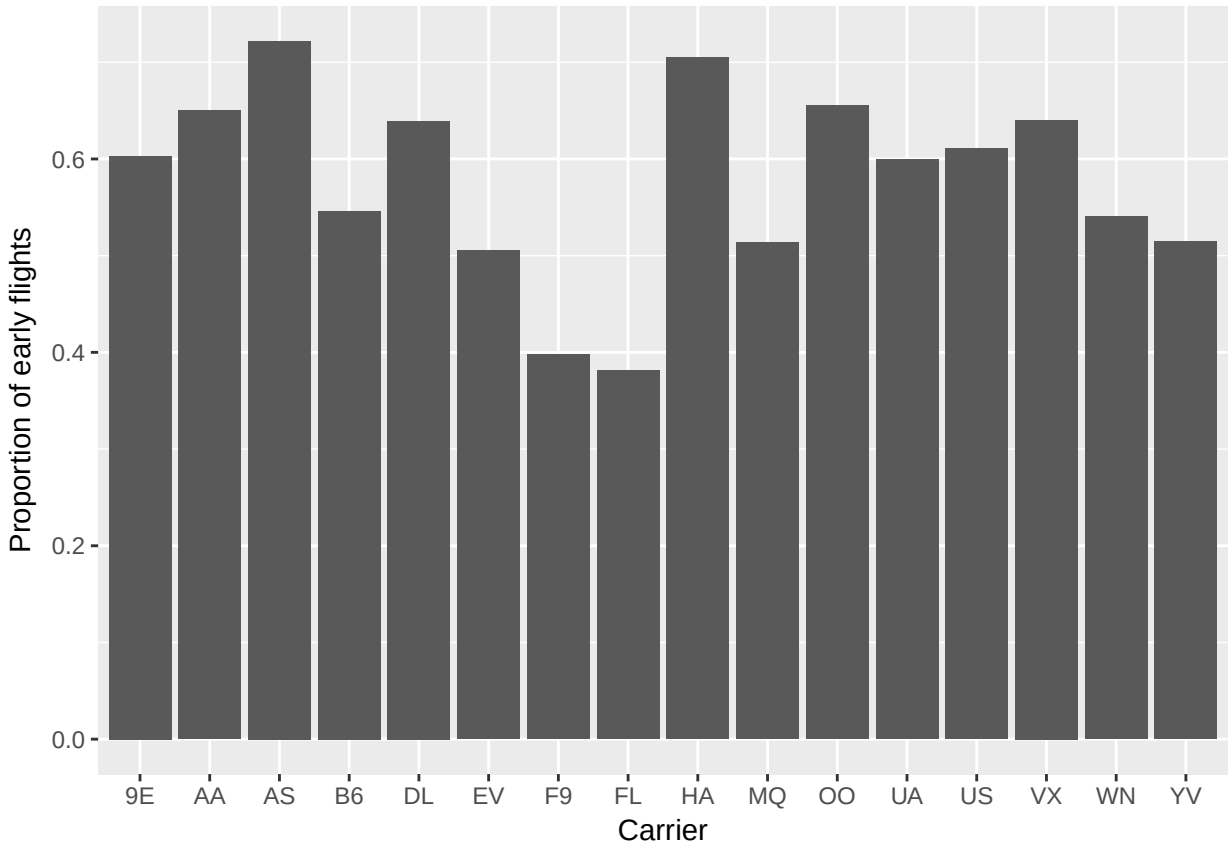


The graph tells that the flights are most likely to arrive early at about 200 miles; then, the probability of early flight decreases along with the increase of median distance. The probability of early flight reaches its minimum when median distance is around 1500 miles. After 1500 miles, the probability bounce back a bit

and finally is stable after 2500 miles.

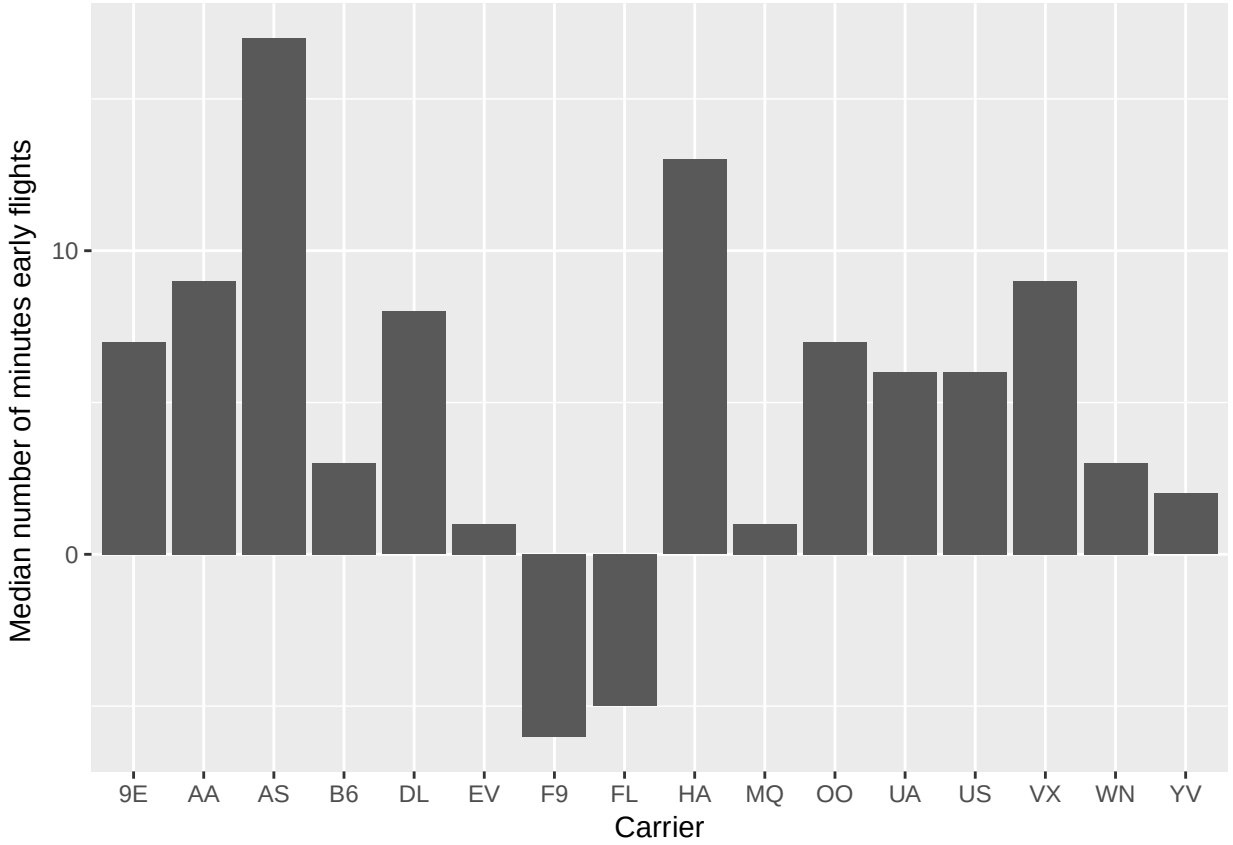
Question 3

```
flights %>%
  group_by(carrier) %>% summarise(prop_earlier=mean(arr_delay<0,na.rm=TRUE)) %>%
  ggplot(aes(x=carrier,y=prop_earlier)) + geom_col()+
  xlab("Carrier")+ylab("Proportion of early flights")
```



The graph shows the proportion of early flights for each carrier.

```
flights %>%
  group_by(carrier) %>%
  summarise(median_earlier=-median(arr_delay,na.rm=TRUE)) %>%
  ggplot(aes(x=carrier,y=median_earlier)) + geom_col()+
  xlab("Carrier")+ylab("Median number of minutes early flights ")
```

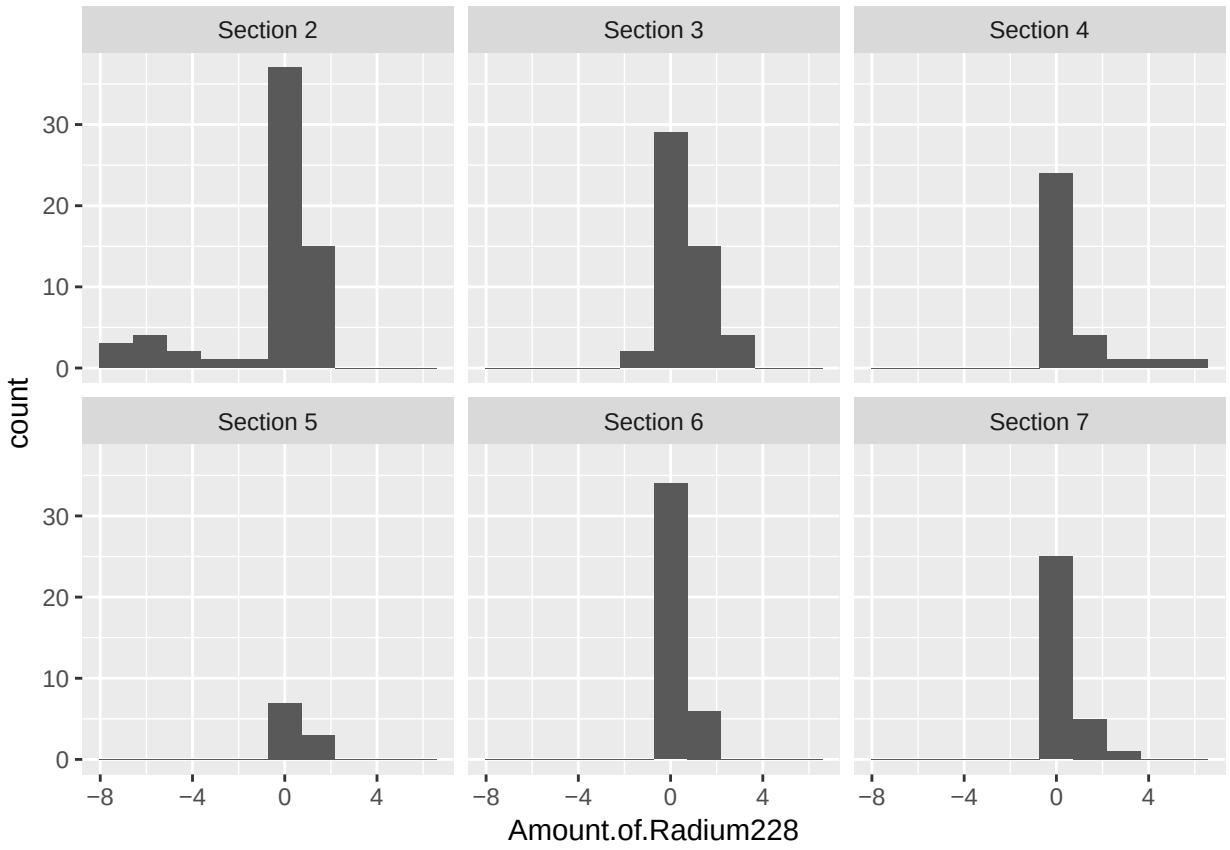


The graph indicates the median number of minutes early that flights arrivee for each carrier. From the graphes, we can know that AS is the most consistently ahead of schedule, and AS also arrives the most early.

However, FL is most consistenly behind schedule, and F9 arrives the latest.

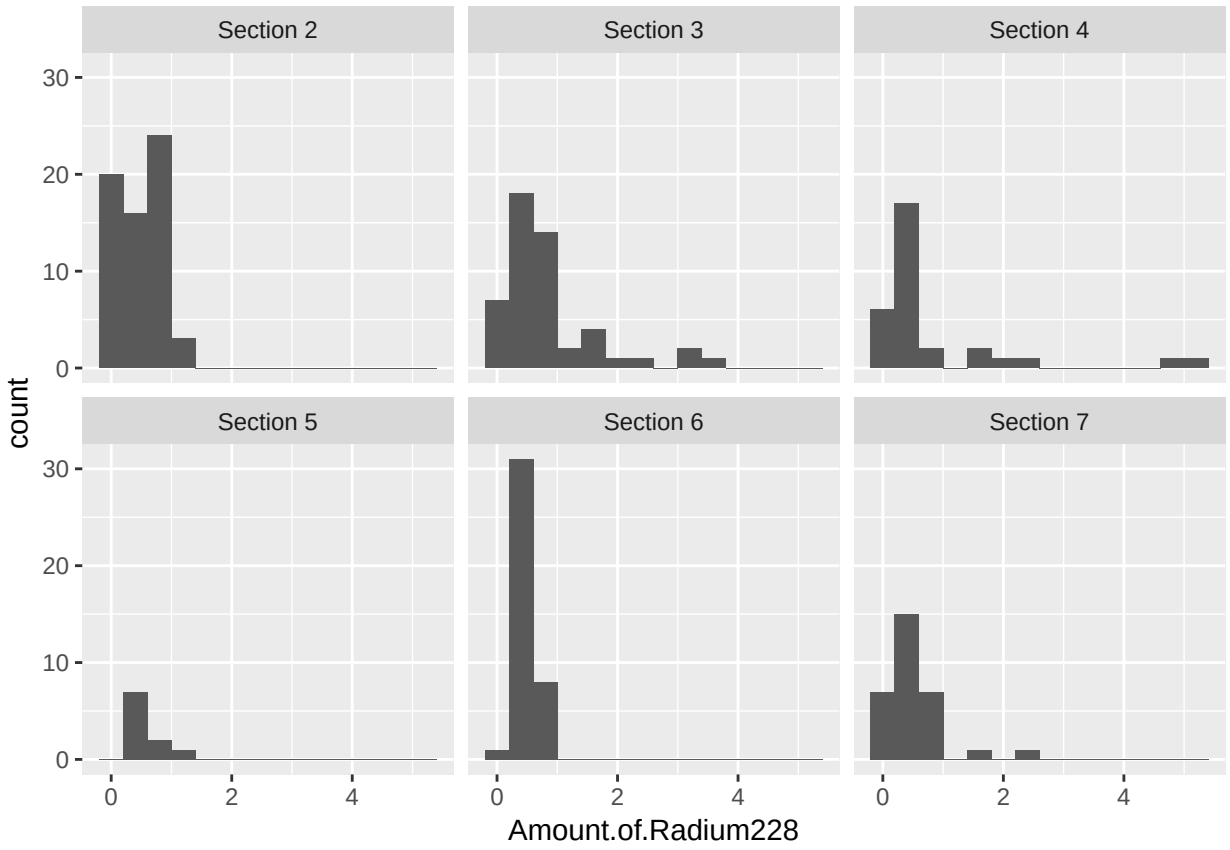
Question 4

```
data1<-read.csv("NavajoWaterExport.csv",header=T)
data2 <- data1
ggplot(data1, aes(x=Amount.of.Radium228)) + geom_bar(stat="bin",bins=10) +
  facet_wrap(~ Which.EPA.Section.is.This.From.)
```



These histograms show the distribution of the amount of Radium-228 in water samples for each EPA section. And, the odd thing is that amount of radium-228 is negative for some sections, which makes no sense.

```
data1 %>%
  mutate(Amount.of.Radium228=ifelse(Amount.of.Radium228<0,0,Amount.of.Radium228))%>%
  ggplot(aes(x=Amount.of.Radium228)) + geom_histogram(binwidth = 0.4) +
  facet_wrap(~ Which.EPA.Section.is.This.From.)
```



These histograms show the distribution of the amount of Radium-228 in water samples for each EPA section after mutating any negative numbers into zero.

Question 5

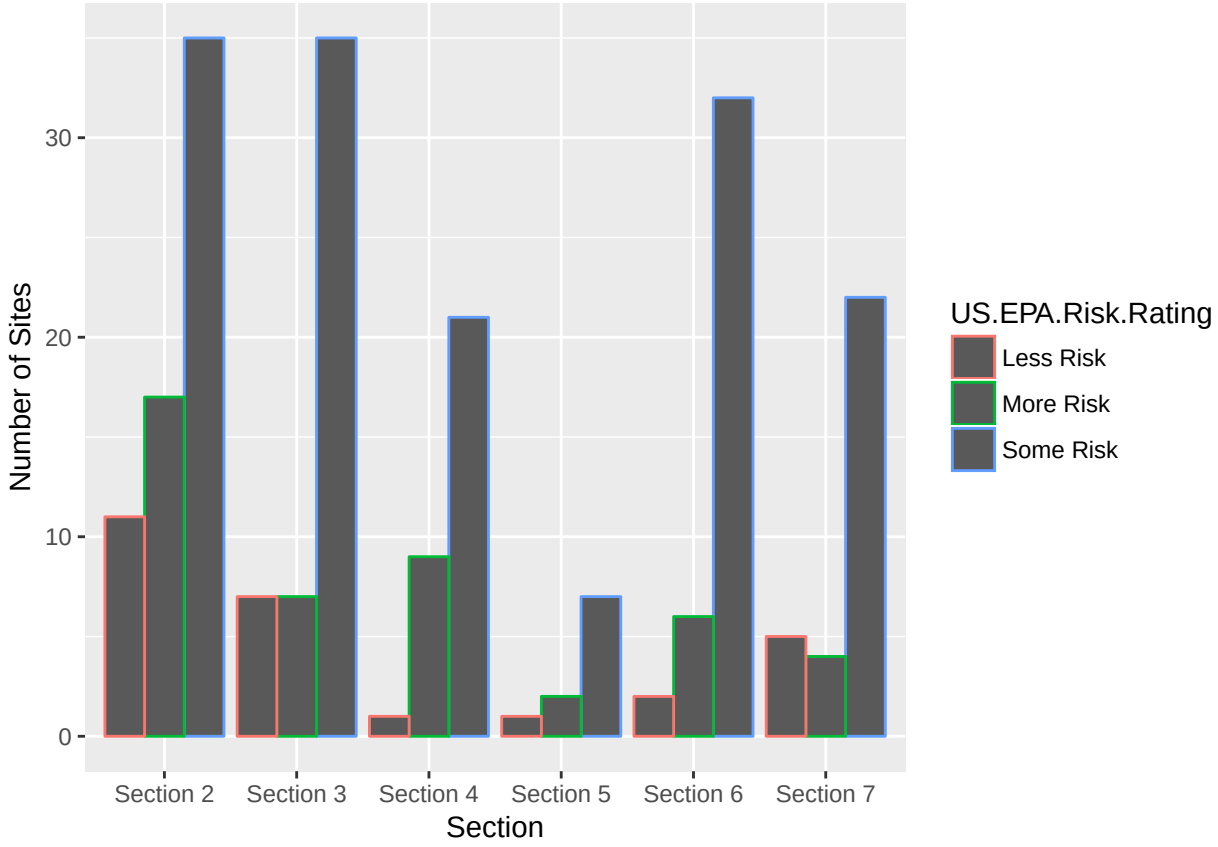
```
(df1 <- data1%>%
  select(Which.EPA.Section.is.This.From.,US.EPA.Risk.Rating,Amount.of.Uranium238)%>%
  filter(US.EPA.Risk.Rating!="Unknown Risk")%>%
  group_by(Which.EPA.Section.is.This.From.,US.EPA.Risk.Rating)%>%
  summarise(number.of.sites=n(),Mean_U238=mean(Amount.of.Uranium238,na.rm=TRUE)))
```

```
## # A tibble: 18 x 4
## # Groups:   Which.EPA.Section.is.This.From. [?]
##    Which.EPA.Section.is.This.From. US.EPA.Risk.Rating number.of.sites
##                                <fctr>             <fctr>           <int>
##  1                       Section 2          Less Risk              11
##  2                       Section 2          More Risk              17
##  3                       Section 2          Some Risk              35
##  4                       Section 3          Less Risk               7
##  5                       Section 3          More Risk               7
##  6                       Section 3          Some Risk              35
##  7                       Section 4          Less Risk               1
##  8                       Section 4          More Risk               9
##  9                       Section 4          Some Risk              21
## 10                       Section 5          Less Risk               1
## 11                       Section 5          More Risk               2
```

```
## 12                        Section 5          Some Risk              7
## 13                        Section 6          Less Risk              2
## 14                        Section 6          More Risk              6
## 15                        Section 6          Some Risk             32
## 16                        Section 7          Less Risk              5
## 17                        Section 7          More Risk              4
## 18                        Section 7          Some Risk             22
## # ... with 1 more variables: Mean_U238 <dbl>
```

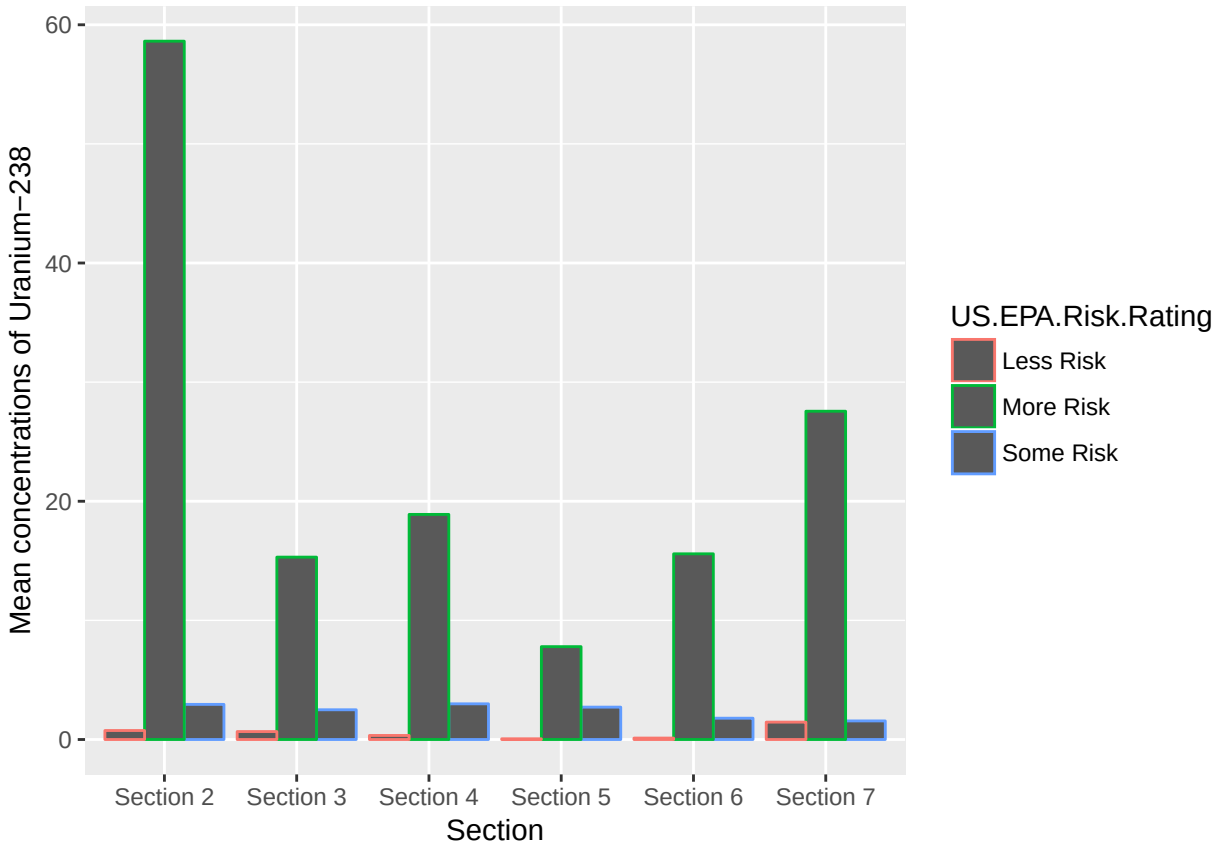
The graph show the the number of sites of each EPA risk rating in each EPA section, and also the mean concentration of Uranium-238 in the water samples for each EPA risk rating in each EPA section.

```
df1%>%
  group_by(US.EPA.Risk.Rating)%>%
  ggplot(aes(x=Which.EPA.Section.is.This.From.,y=number.of.sites,color=US.EPA.Risk.Rating)) +
           geom_bar(stat="identity",position = "dodge")+
  xlab("Section")+ylab("Number of Sites")
```



The graph demonstrates the number of sites at each EPA section with different risk rating. It tells that section 2 have the most sites with "more risk".

```
df1%>%
  ggplot(aes(x=Which.EPA.Section.is.This.From.,y=Mean_U238,color=US.EPA.Risk.Rating)) +
    geom_bar(stat="identity",position = "dodge")+
  xlab("Section")+ylab("Mean concentrations of Uranium-238")
```



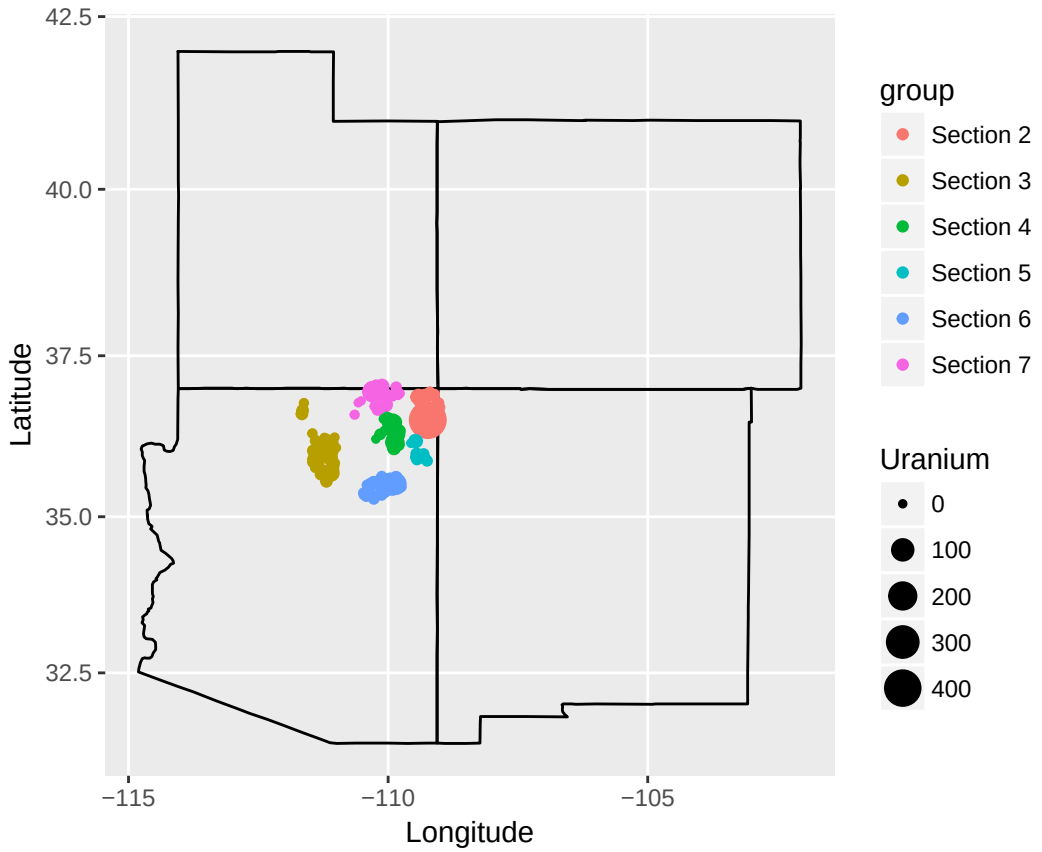
The graph shows the mean concentration of Uranium-238 for each EPA section and color-code the different risk rating. From the graph, we can know that section 2 has the sites with the highest concentration of Uranium-238 on average.

Question 6

```
library(maps)
library(measurements)
four_corners <- map_data("state",
                         region=c("arizona", "new mexico",
                                  "utah",
                                  "colorado"))
Sites=data1%>%
  transmute(long=conv_unit(Longitude,"deg_min_sec","dec_deg"),
            lat=conv_unit(Latitude,"deg_min_sec","dec_deg"),
          group=Which.EPA.Section.is.This.From.,
          Uranium=ifelse(Amount.of.Uranium238<0,0,Amount.of.Uranium238))
Sites$long <- -abs(as.numeric(Sites$long))
Sites$lat <- round(as.numeric(Sites$lat),digits=5)

ggplot(four_corners) + geom_polygon(mapping=aes(x=long,
                                                y=lat,
                                                group=group),fill=NA,
                                    color="black")+
  coord_map()+
```

```
  geom_point(data=Sites,mapping=aes(x=long,
                                    y=lat,color=group,size=Uranium))+
  xlab("Longitude")+ylab("Latitude")
```



The map indicates the the location of EPA section and the average amount of Uranium-238 at that site.

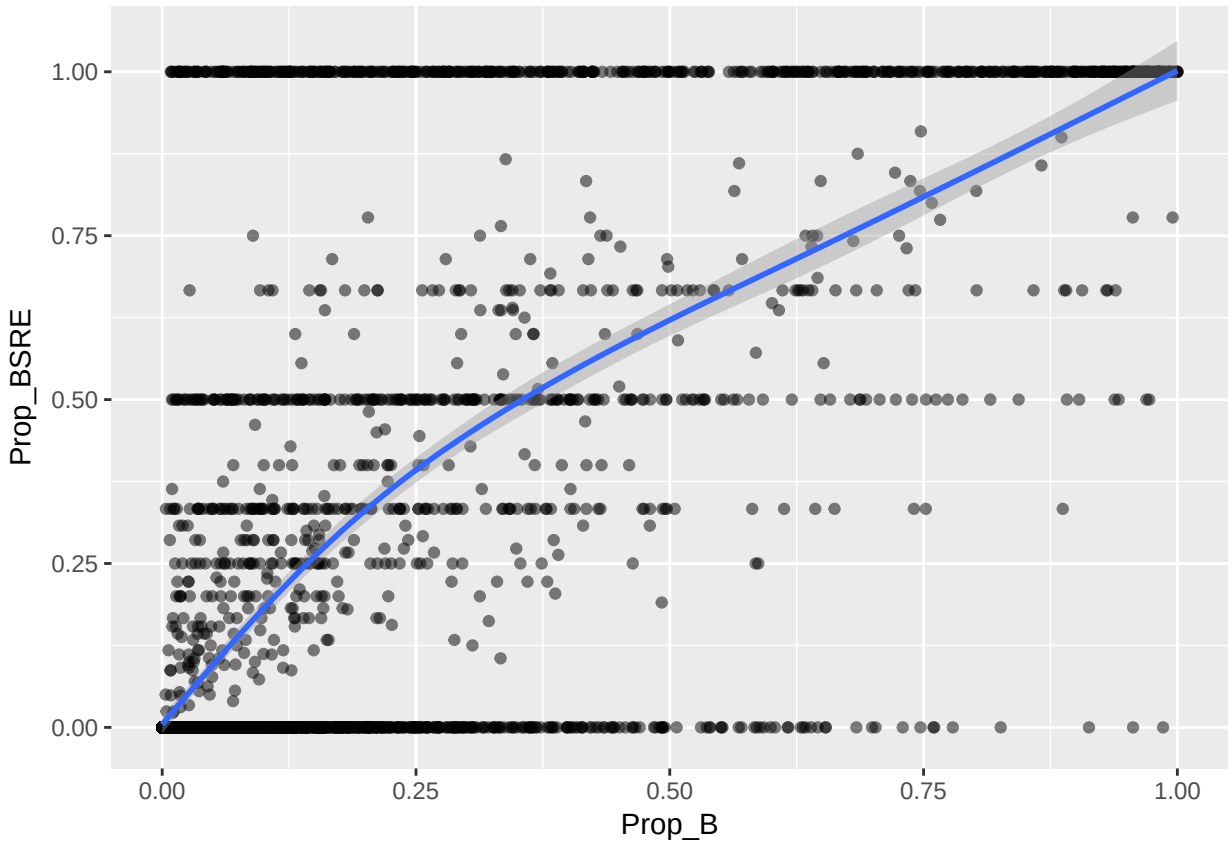
Question 7

```
df3<- read_csv("CRDC2013_14_SCH.csv",na=c("-2","-5","-9"))
dfB=df3%>%
  transmute(TS=TOT_ENR_M+TOT_ENR_F,TBS=SCH_ENR_BL_M+SCH_ENR_BL_F,
  SRE=TOT_DISCWODIS_EXPZT_M+TOT_DISCWODIS_EXPZT_F+
    TOT_DISCWDIS_EXPZT_IDEA_M+TOT_DISCWDIS_EXPZT_IDEA_F,
  BSRE=SCH_DISCWDIS_EXPZT_IDEA_BL_M+SCH_DISCWDIS_EXPZT_IDEA_BL_F+
    SCH_DISCWODIS_EXPZT_BL_M+SCH_DISCWODIS_EXPZT_BL_F,
  Prop_B=TBS/TS,Prop_BSRE=BSRE/SRE)

dfB%>%
  ggplot()+geom_point(aes(x=Prop_B,y=Prop_BSRE),alpha=1/2)+geom_smooth(aes(x=Prop_B,y=Prop_BSRE))
```

```
## Warning: Removed 91683 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 91683 rows containing missing values (geom_point).
```



The graph shows proportion of Black students at each school (on the x-axis) versus the proportion of expelled students who are Black (on the y-axis). The smooth line indicates that Black students is over-represented in expulsions under zero-tolerance policies because the smooth line is above the y=x line.

```
dfB%>%
  summarise(Prop_overallB=sum(TBS,na.rm=TRUE)/sum(TS,na.rm=TRUE),
            Prop_overallBRE=sum(BSRE,na.rm=TRUE)/sum(SRE,na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   Prop_overallB Prop_overallBRE
##           <dbl>           <dbl>
## 1     0.1549763       0.2304784
```

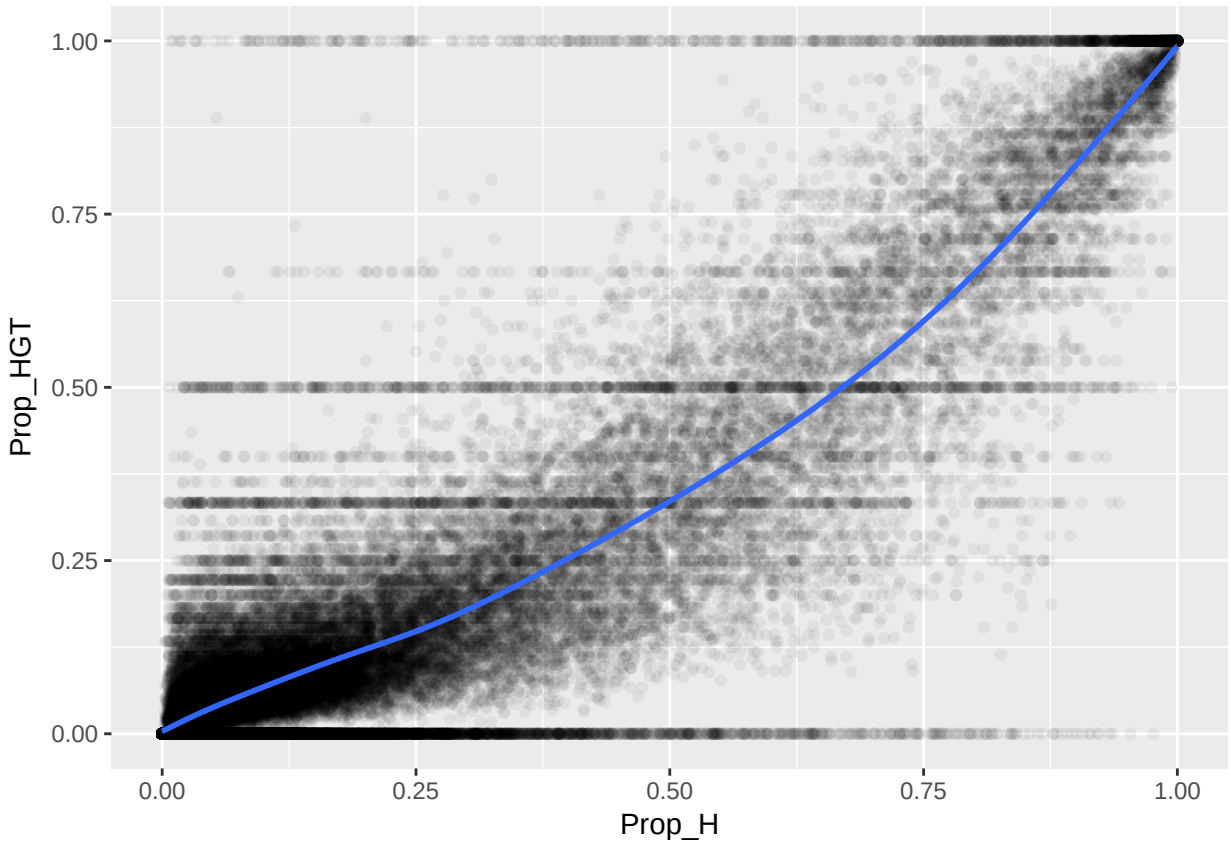
The table shows the overall proportion of Black students across all schools and the overall proportion of students expelled under zero-tolerance policies who are Black across all schools.

Question 8

```
dfH=df3%>%
  transmute(TS=TOT_ENR_M+TOT_ENR_F,TSH=SCH_ENR_HI_M+SCH_ENR_HI_F,
            SGT=TOT_GTENR_M+TOT_GTENR_F,HSGT=SCH_GTENR_HI_M+SCH_GTENR_HI_F,
            Prop_H=TSH/TS,Prop_HGT=HSGT/SGT)
dfH%>%
  ggplot()+geom_point(aes(x=Prop_H,y=Prop_HGT),alpha=1/20)+geom_smooth(aes(x=Prop_H,y=Prop_HGT))
```

```
## Warning: Removed 40556 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 40556 rows containing missing values (geom_point).
```



The graph shows the proportion of Hispanic students at each school (on the x-axis) versus the proportion of GT students who are Hispanic (on the y-axis). In addition, the smooth line indicates an under-representation of Hispanic students in Gifted & Talented programs as the it is belowed the line x=y.

```
dfH%>%
  summarise(Prop_overallH=sum(TSH,na.rm=TRUE)/sum(TS,na.rm=TRUE),
            Prop_overallHSGT=sum(HSGT,na.rm=TRUE)/sum(SGT,na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   Prop_overallH Prop_overallHSGT
##           <dbl>            <dbl>
## 1      0.247396        0.1806664
```

The table indicates the overall proportion of Hispanic students across all schools and the overall proportion of GT students who are Hispanic.

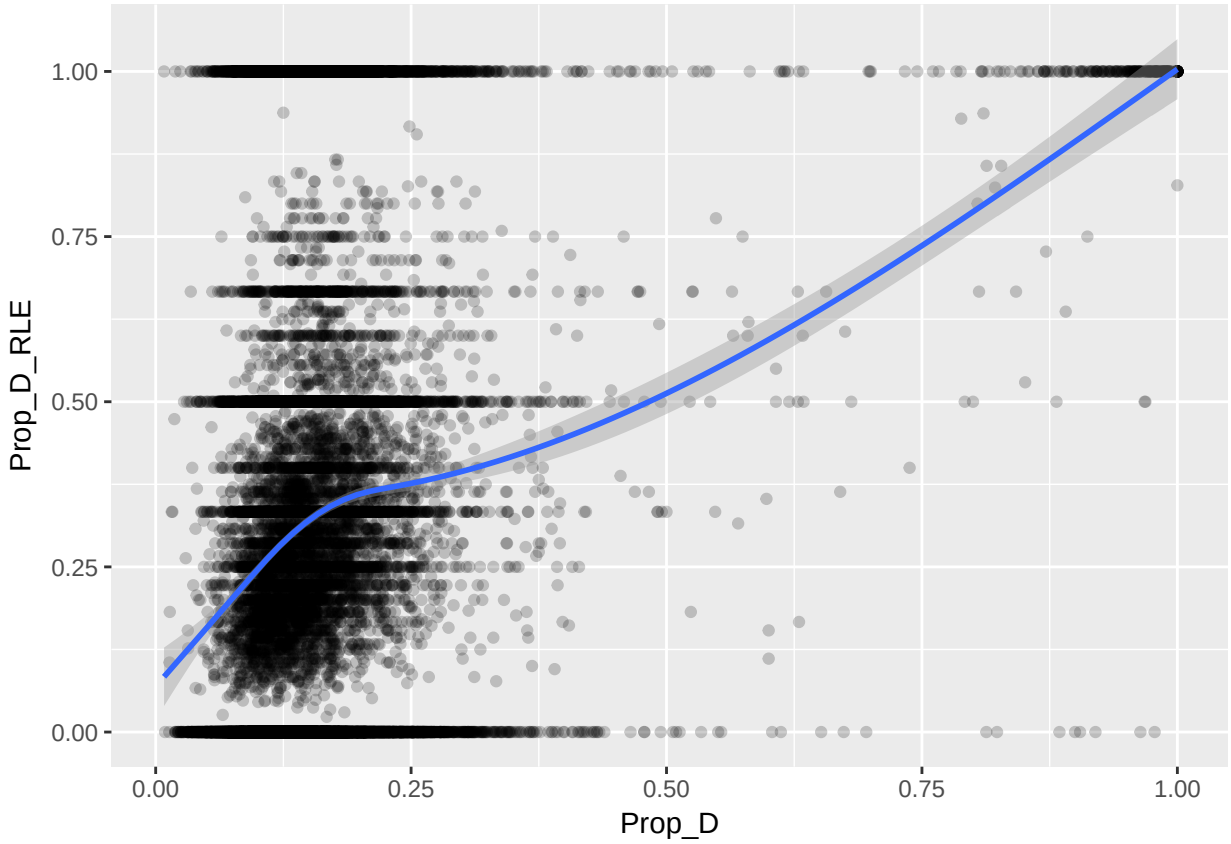
Question 9

```
dfD=df3%>%
  filter(TOT_ENR_M+TOT_ENR_F>=SCH_ENR_IDEA_M+SCH_ENR_IDEA_F+
           SCH_ENR_504_M+SCH_ENR_504_F)%>%
  transmute(TS=TOT_ENR_M+TOT_ENR_F,TIDEA=SCH_ENR_IDEA_M+SCH_ENR_IDEA_F+SCH_ENR_504_M+SCH_ENR_504_F,
            DS_RLE=TOT_DISCWDIS_REF_IDEA_M+TOT_DISCWDIS_REF_IDEA_F,
            TS_RLE=TOT_DISCWDIS_REF_IDEA_M+TOT_DISCWDIS_REF_IDEA_F+
              TOT_DISCWODIS_REF_M+TOT_DISCWODIS_REF_F,
            Prop_D=TIDEA/TS,Prop_D_RLE=DS_RLE/TS_RLE)
dfD%>%
```

```
  ggplot()+geom_point(aes(x=Prop_D,y=Prop_D_RLE),alpha=1/5)+
  geom_smooth(aes(x=Prop_D,y=Prop_D_RLE))
```

```
## Warning: Removed 65435 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 65435 rows containing missing values (geom_point).
```



The graph shows the proportion of disabled students at each school (on the x-axis) versus the proportion of students referred to law enforcement who are disabled (on the y-axis).The smooth line indicates an over-representation of disabled students referred to law enforcement as the it is belowed the line x=y

```
dfD%>%
  summarise(Prop_overall_Dis=sum(TIDEA,na.rm=TRUE)/sum(TS,na.rm=TRUE),
            Prop_overall_Dis_CP=sum(DS_RLE,na.rm=TRUE)/sum(TS_RLE,na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   Prop_overall_Dis Prop_overall_Dis_CP
##              <dbl>               <dbl>
## 1        0.1427554           0.2734696
```

The table indicates the overall proportion of disabled students across all schools and the overall proportion of students referred to law enforcement who are disabled across all schools.

Question 10

The Question is about whether Asian students are over- or under-represented in Gifted & Talented programs.

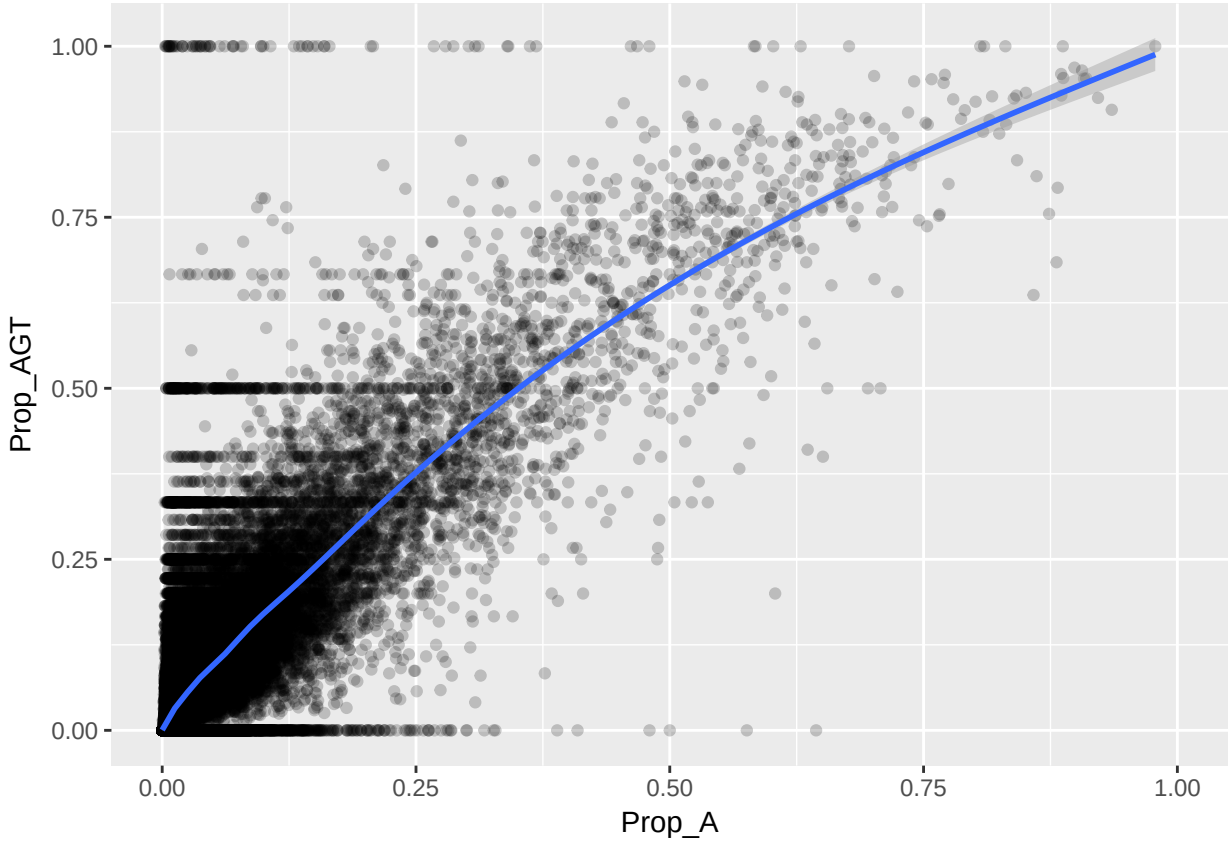
A new data is created, containing only schools with a Gifted & Talented program with the following columns:

• The total number of students enrolled at each school • The total number of Asian 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 Asian • The proportion of students at each school who are Asian • The proportion of students in the GT program who are Asian

```
dfA=df3%>%
transmute(TS=TOT_ENR_M+TOT_ENR_F,TSA=SCH_ENR_AS_M+SCH_ENR_AS_F,
            SGT=TOT_GTENR_M+TOT_GTENR_F,
          ASGT=SCH_GTENR_AS_M+SCH_GTENR_AS_F,
            Prop_A=TSA/TS,Prop_AGT=ASGT/SGT)
dfA%>%
  ggplot()+geom_point(aes(x=Prop_A,y=Prop_AGT),alpha=1/5)+geom_smooth(aes(x=Prop_A,y=Prop_AGT))
```

```
## Warning: Removed 40555 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 40555 rows containing missing values (geom_point).
```



The graph shows the proportion of Asian students at each school (on the x-axis) versus the proportion of GT students who are Asian (on the y-axis). The smooth line indicates an over-representation of Asian students in Gifted & Talented programs as the it is above the line x=y.

```
dfA%>%
  summarise(Prop_overall_Asian=sum(TSA,na.rm=TRUE)/sum(TS,na.rm=TRUE),
            Prop_overall_AsianinGT=sum(ASGT,na.rm=TRUE)/sum(SGT,na.rm=TRUE))
```

```
## # A tibble: 1 x 2
##   Prop_overall_Asian Prop_overall_AsianinGT
##                <dbl>                  <dbl>
## 1         0.04826402             0.09762055
```

The table indicates the overall proportion of Asian students across all schools and the overall proportion of GT students who are Asian.