hw2-Rohit-Thakur

Rohit Thakur

1/24/2020

R Markdown

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
df<-read.csv("C:/Users/rthak/OneDrive/Desktop/crime.csv")
head(df,n=10)
```

```
##    INCIDENT_NUMBER OFFENSE_CODE              OFFENSE_CODE_GROUP
## 1       I182080058         2403              Disorderly Conduct
## 2       I182080053         3201                   Property Lost
## 3       I182080052         2647                           Other
## 4       I182080051          413              Aggravated Assault
## 5       I182080050         3122                        Aircraft
## 6       I182080049         1402                       Vandalism
## 7       I182080048         3803 Motor Vehicle Accident Response
## 8       I182080047         3301                 Verbal Disputes
## 9       I182080045          802                  Simple Assault
## 10      I182080044         3410                           Towed
##               OFFENSE_DESCRIPTION DISTRICT REPORTING_AREA SHOOTING
## 1            DISTURBING THE PEACE      E18            495
## 2                 PROPERTY - LOST      D14            795
## 3       THREATS TO DO BODILY HARM       B2            329
## 4  ASSAULT - AGGRAVATED - BATTERY       A1             92
## 5              AIRCRAFT INCIDENTS       A7             36
## 6                       VANDALISM      C11            351
## 7  M/V ACCIDENT - PERSONAL INJURY                      NA
## 8                  VERBAL DISPUTE       B2            603
## 9        ASSAULT SIMPLE - BATTERY      E18            543
## 10            TOWED MOTOR VEHICLE       D4            621
##       OCCURRED_ON_DATE YEAR MONTH DAY_OF_WEEK HOUR   UCR_PART           STREET
## 1  2018-10-03 20:13:00 2018    10   Wednesday   20   Part Two     ARLINGTON ST
## 2  2018-08-30 20:00:00 2018     8    Thursday   20 Part Three       ALLSTON ST
## 3  2018-10-03 19:20:00 2018    10   Wednesday   19   Part Two         DEVON ST
## 4  2018-10-03 20:00:00 2018    10   Wednesday   20   Part One     CAMBRIDGE ST
## 5  2018-10-03 20:49:00 2018    10   Wednesday   20 Part Three      PRESCOTT ST
## 6  2018-10-02 20:40:00 2018    10     Tuesday   20   Part Two   DORCHESTER AVE
## 7  2018-10-03 20:16:00 2018    10   Wednesday   20 Part Three
## 8  2018-10-03 19:32:00 2018    10   Wednesday   19 Part Three       TREMONT ST
## 9  2018-10-03 19:27:51 2018    10   Wednesday   19   Part Two         AVILA RD
## 10 2018-10-03 20:00:00 2018    10   Wednesday   20 Part Three COMMONWEALTH AVE
##          Lat      Long                    Location
## 1  42.26261 -71.12119 (42.26260773, -71.12118637)
## 2  42.35211 -71.13531 (42.35211146, -71.13531147)
## 3  42.30813 -71.07693 (42.30812619, -71.07692974)
## 4  42.35945 -71.05965 (42.35945371, -71.05964817)
## 5  42.37526 -71.02466 (42.37525782, -71.02466343)
## 6  42.29920 -71.06047 (42.29919694, -71.06046974)
## 7  42.32073 -71.05676 (42.32073413, -71.05676415)
## 8  42.33381 -71.10378 (42.33380683, -71.10377843)
## 9  42.25614 -71.12803 (42.25614494, -71.12802506)
## 10 42.34887 -71.08936 (42.34886600, -71.08936284)
```

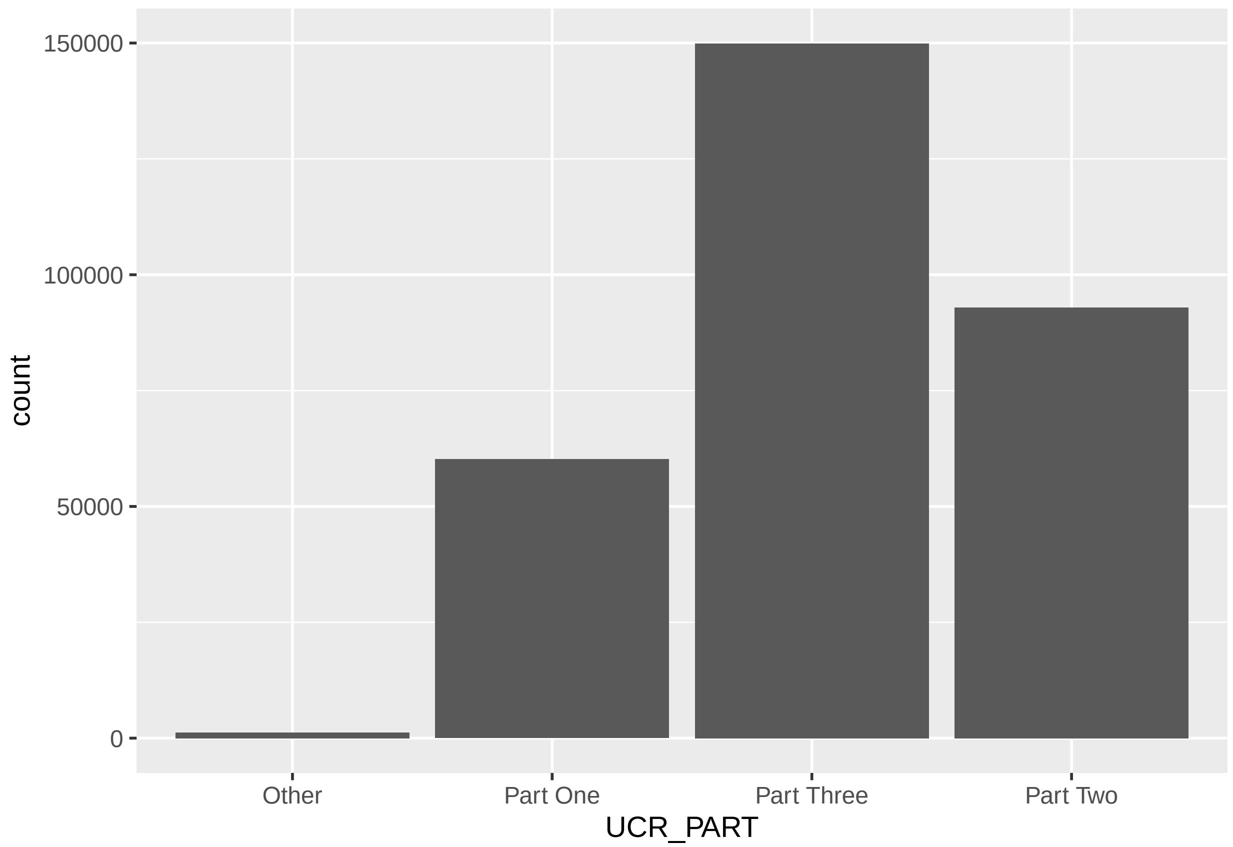
```r
df1<-select(df,-c("INCIDENT_NUMBER","OFFENSE_CODE","Location","OFFENSE_DESCRIPTION"))
df1<-separate(df1,col="OCCURRED_ON_DATE",into=c("Date","Time"),sep=" ")
df1<-filter(df1,df1$DISTRICT!="")
```

Observations made from below graph:UCR_PART three crimes are largest because they are non_violent and most occuring incidents(for ex. medical emergency). We will further investigate insights for part one crimes as they are most violent ones.

```r
library(ggplot2)
df1<-filter(df1,UCR_PART!=" ")
levels(df1$UCR_PART)
```

```
## [1] ""           "Other"      "Part One"   "Part Three" "Part Two"
```

```r
df1$UCR_PART<-droplevels(df1$UCR_PART,exclude="")
df1<-na.omit(df1)
ggplot(df1)+geom_bar(aes(x=UCR_PART),stat='count')
```



Focusing on Part one crimes

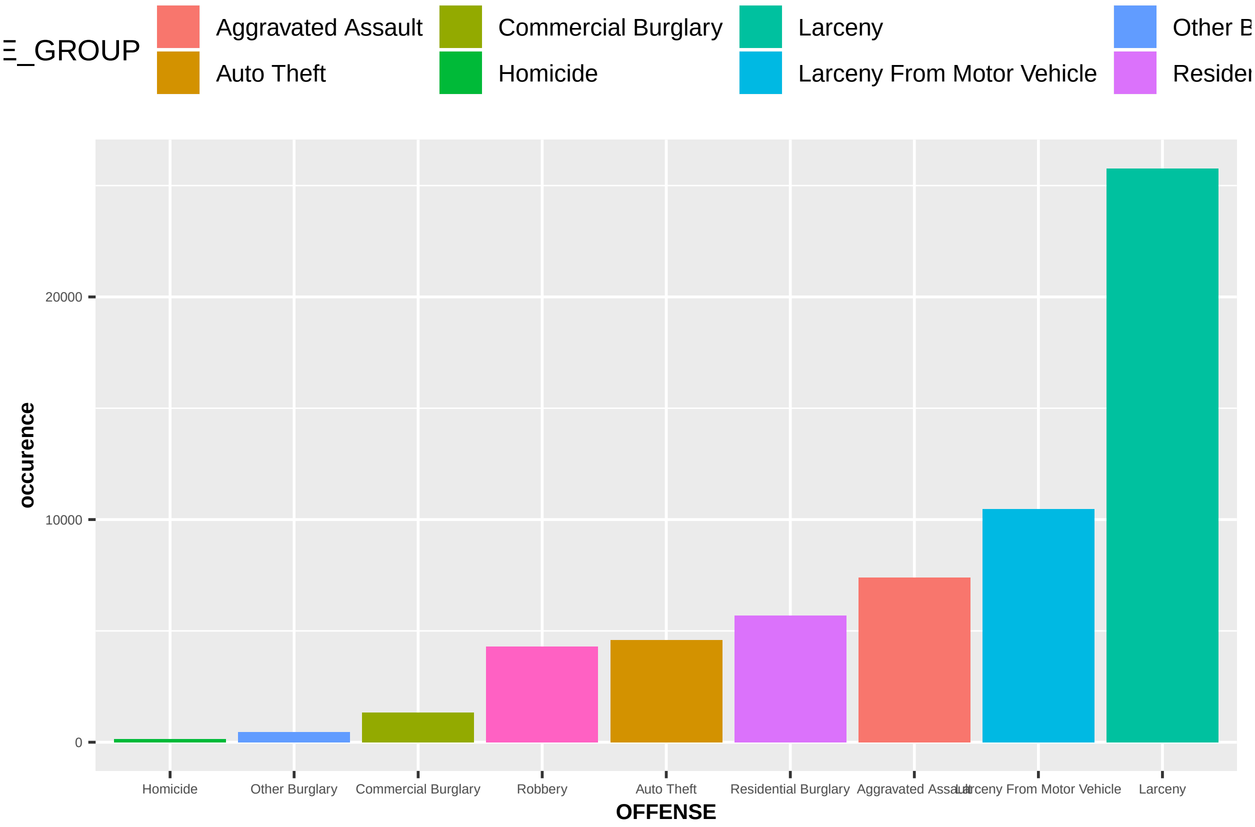
```
part1<-subset(df1,UCR_PART=="Part One")
```

```
bypart<-part1%>%group_by(OFFENSE_CODE_GROUP)%>%summarise(occurence=n())
bypart
```

```
## # A tibble: 9 x 2
##   OFFENSE_CODE_GROUP         occurence
##   <fct>                          <int>
## 1 Aggravated Assault              7395
## 2 Auto Theft                      4588
## 3 Commercial Burglary             1338
## 4 Homicide                         153
## 5 Larceny                        25772
## 6 Larceny From Motor Vehicle     10479
## 7 Other Burglary                   459
## 8 Residential Burglary            5691
## 9 Robbery                         4297
```

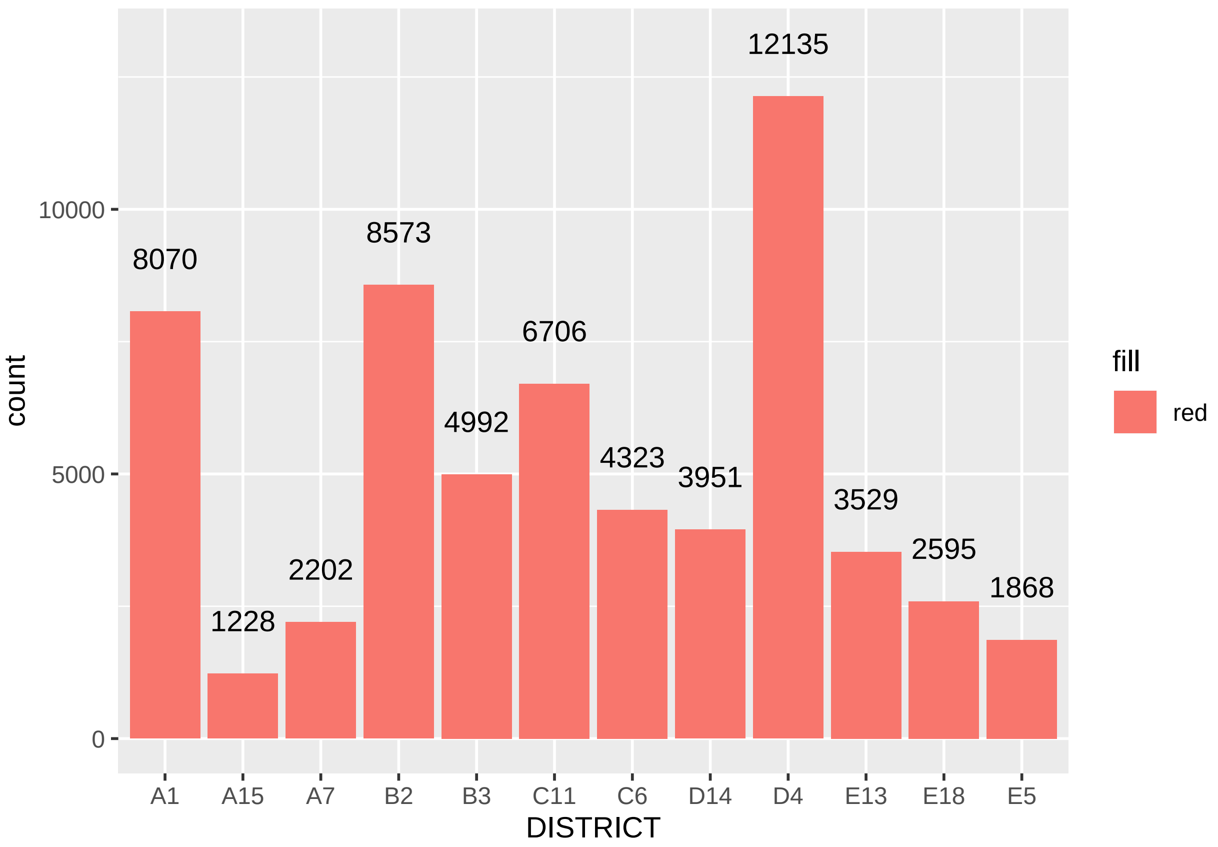
UCR PART ONE CRIMES

```
ggplot(bypart)+geom_bar(aes(reorder(x=OFFENSE_CODE_GROUP,occurence),y=occurence,fill=OFFENSE_CODE_GROUP
```



UCR PART_1 crime by district: D4 has highest number of crimes. D4 can be categorized as unsafe district.

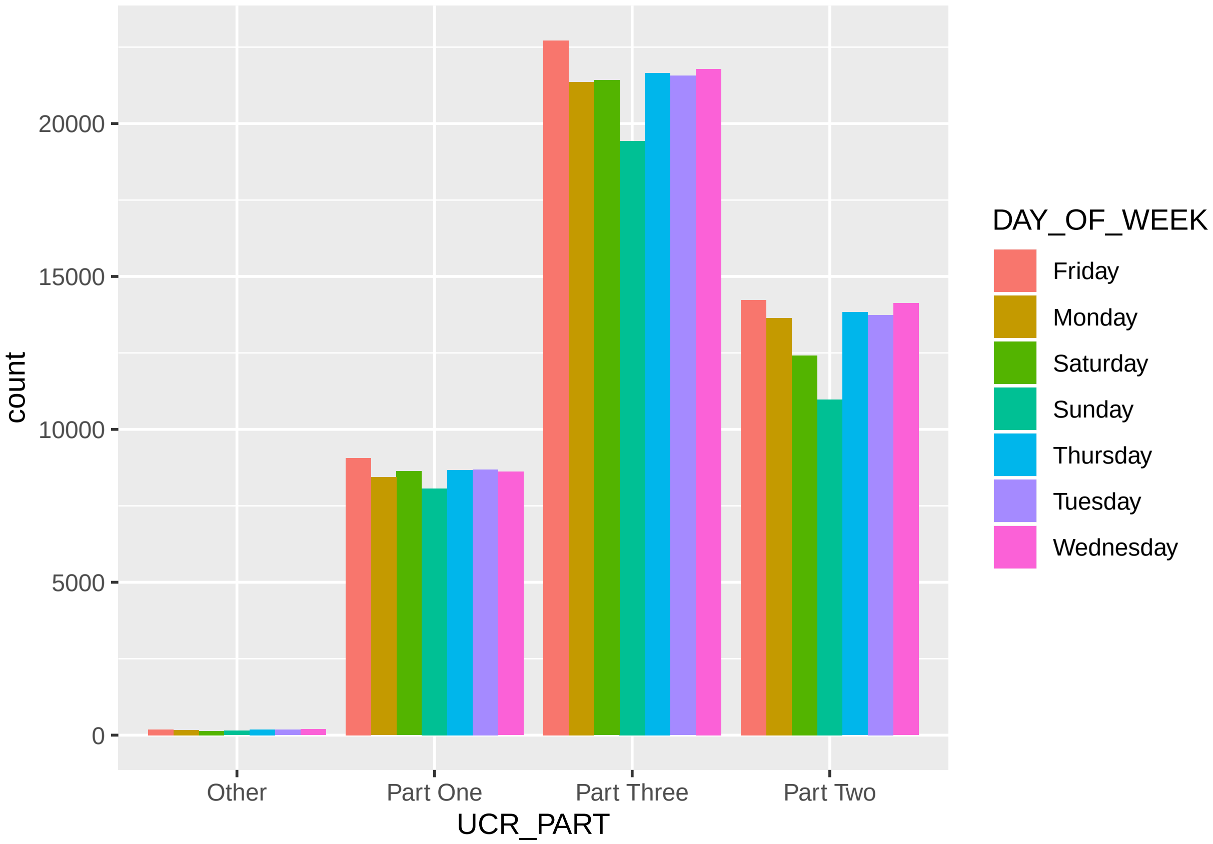
```
ggplot(part1)+geom_bar(aes(x=DISTRICT,fill="red"))+geom_text(aes(x=DISTRICT,label=stat(count)),stat='co
```



UCR_PART by day of week interesting observations: For every type of UCR_PART Fridays are having more crimes than other days.

```r
ggplot(df1)+geom_histogram(aes(x=UCR_PART,fill=DAY_OF_WEEK),stat='count',position="dodge")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



From bar graph below, larceny has greatest share in d4 crimes.

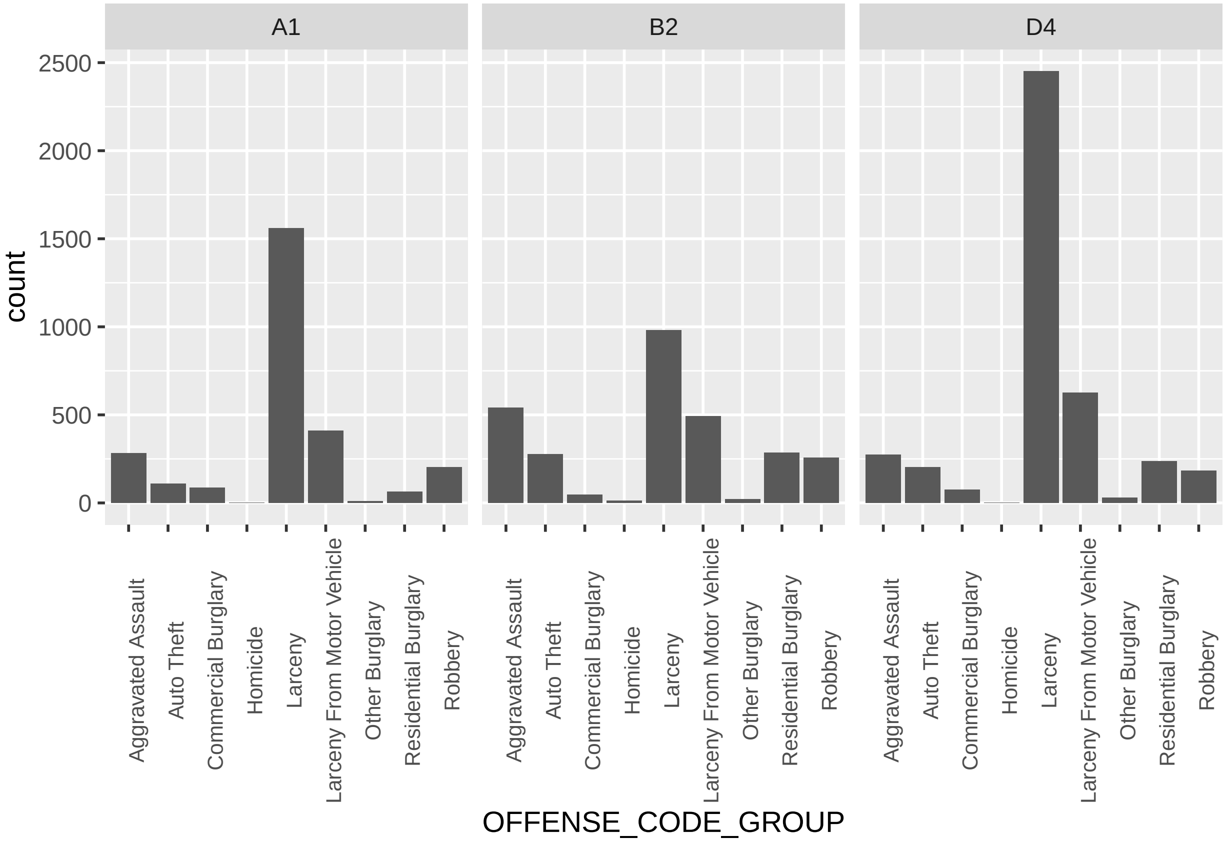
```
d4<-subset(part1,part1$DISTRICT==c("D4","B2","A1"))
```

```
## Warning in `==.default`(part1$DISTRICT, c("D4", "B2", "A1")): longer object
## length is not a multiple of shorter object length
```

```
## Warning in is.na(e1) | is.na(e2): longer object length is not a multiple of
## shorter object length
```

```
ggplot(d4)+geom_histogram(aes(x=OFFENSE_CODE_GROUP),stat='count')+theme(axis.text.x = element_text(size
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



```r
library(readr)
navajo<-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.
```

```r
navajo<-mutate(navajo,`Amount of Radium228`=ifelse(`Amount of Radium228`<0.0,0,`Amount of Radium228`))
navajo$`Amount of Radium228`
```

```
##   [1] 0.500 1.540 0.591 0.183 0.439 0.892 0.565 0.065 0.353 0.975 0.742 0.000
##  [13] 0.822 2.300 1.600 0.170 3.230 0.746 0.000 0.422 0.219 4.680 0.703 0.987
##  [25] 0.800 2.220 0.477 0.572 0.530 5.170 0.000 0.571 0.097 0.499 0.451 0.247
##  [37] 0.359 1.550 0.308 0.425 1.950 0.118 0.539 0.639 0.554 0.320 0.206 0.275
##  [49] 0.564 0.834 0.560 0.551 0.964 0.191 0.723 0.812 0.661 0.603 0.378 0.946
##  [61] 0.000 0.183 0.000 0.075 0.272 0.353 0.212 0.016 0.533 0.511 0.217 0.208
##  [73] 0.333 0.604 0.864 3.230 0.608 0.636 0.021 0.000 0.681 0.650 0.824 0.144
##  [85] 0.212 0.368 0.303 0.301 0.000 0.416 0.263 0.534 0.386 0.000 0.000 0.000
##  [97] 0.751 0.928 0.000 0.242 0.379 0.000 0.228 0.000 0.444 0.234 0.000 0.216
## [109] 0.237 0.000 1.220 0.000 1.770 1.490 0.358 0.000 0.475 1.800 0.000 0.172
## [121] 0.471 0.820 0.417 0.556 0.473 0.000 0.000 0.700 0.636 0.563 0.158 0.900
## [133] 0.534 0.309 0.000 0.759 1.090 0.413 0.591 3.600 0.190 0.837 0.668 0.308
## [145] 0.700 0.655 0.325 2.100 0.409 0.177 0.591 0.370 0.000 0.651 0.484 0.000
## [157] 0.249 0.759 0.190 0.215 0.465 0.504 0.362 0.677 0.631 0.337 1.120 0.644
## [169] 0.268 0.552 0.775 1.160 0.203 0.501 0.942 0.383 0.437 0.722 0.592 0.572
## [181] 0.898 0.460 0.297 0.699 0.768 0.799 0.483 0.803 0.346 0.444 0.559 0.362
## [193] 0.698 0.546 0.599 0.500 0.813 0.215 1.240 0.460 0.891 0.503 2.480 0.567
## [205] 0.458 0.833 0.311 0.584 0.400 1.630 0.672 0.219 0.756 0.689 0.215 0.799
## [217] 0.247 1.130 0.368 0.308 0.917 0.077 0.840 0.101 0.394
```

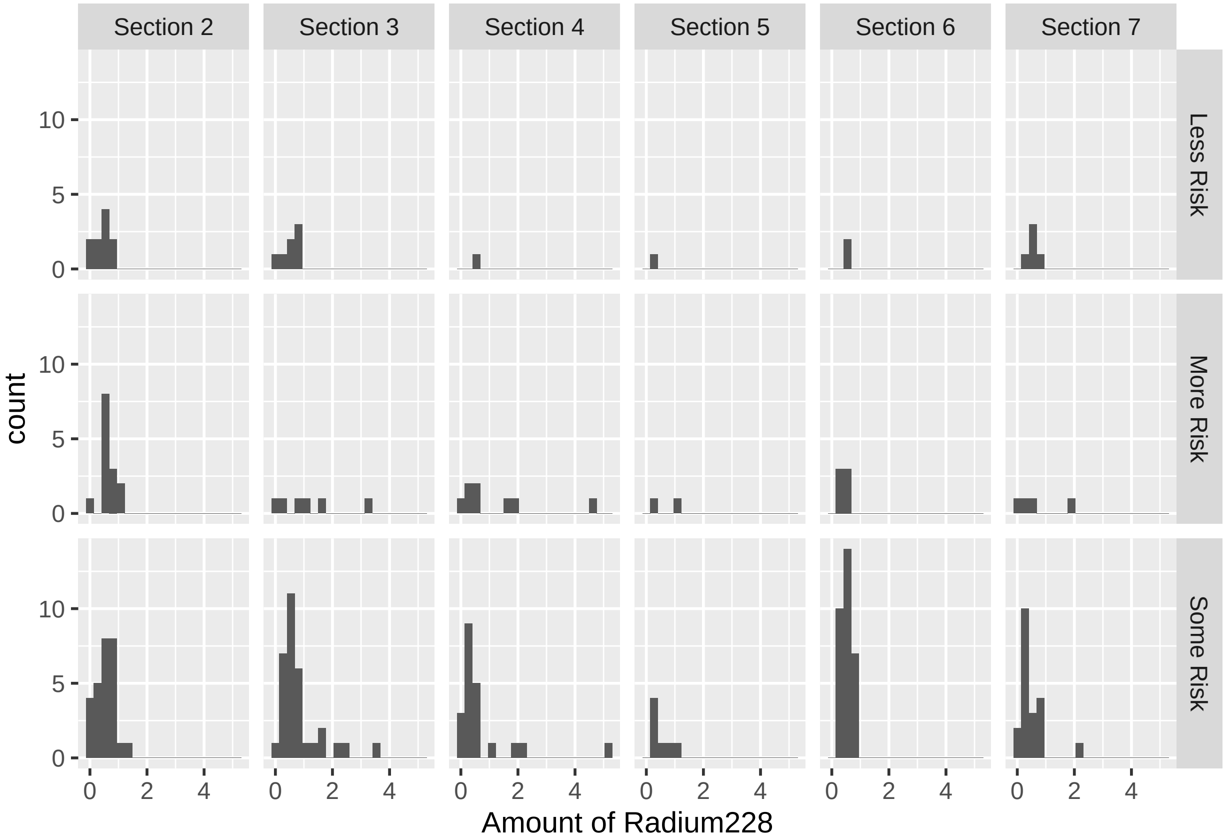
```r
nv1<-filter(navajo,`US EPA Risk Rating`!="Unknown Risk")
"Unknown Risk" %in% nv1$`US EPA Risk Rating`
```

```
## [1] FALSE
```

```r
head(nv1)
```

```
## # A tibble: 6 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~
## # ... with 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>
```

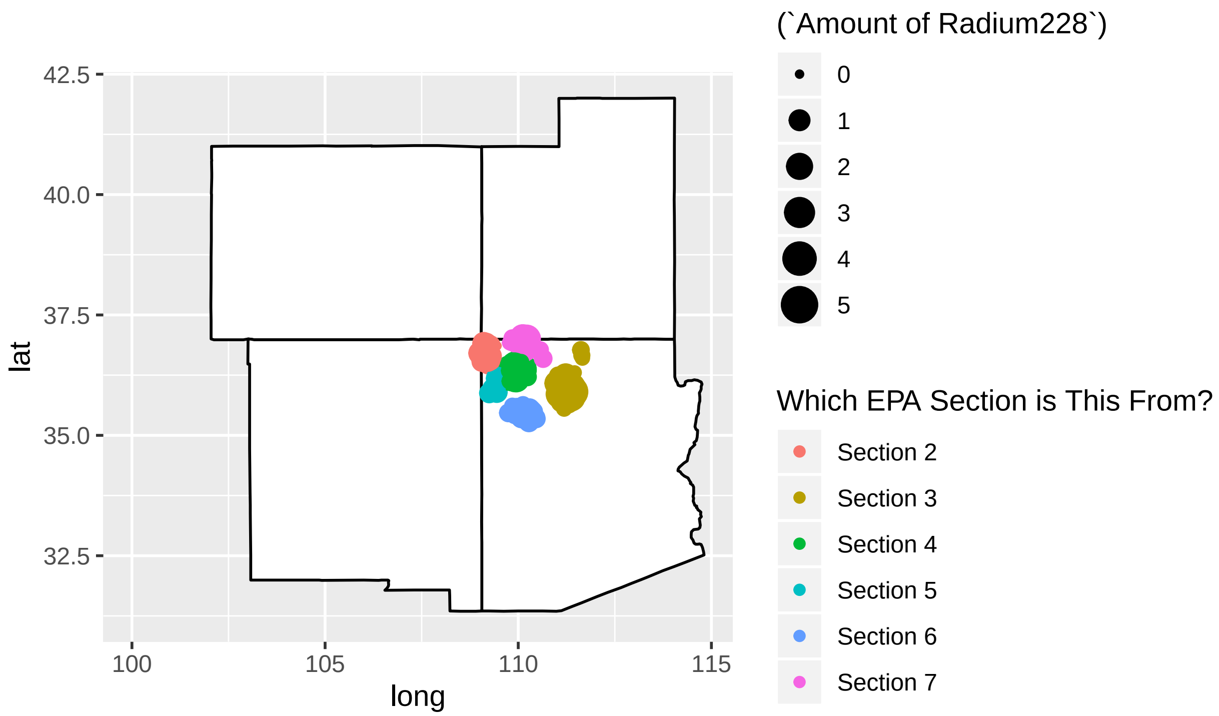
```
nv1_df <- summarise(group_by(nv1, `US EPA Risk Rating`, `Which EPA Section is This From?` ,`Amount of Ra
ggplot(nv1_df) +
  geom_histogram(aes(x=nv1_df$`Amount of Radium228`),bins=20) +
  facet_grid(nv1_df$`US EPA Risk Rating` ~nv1_df$`Which EPA Section is This From?`)+labs(x="Amount of Ra
```



```
library(ggplot2)
#install.packages("maps")
#install.packages("mapproj")
```

Observations: In plot above for less risk in each section, amount of radium228 is less along with its frequency for each section. For some risk there is high frequency of radium 228 in lower amount in every section. For more risk amount of radium is spread around larger spectrum with its frequency highest in section 2, that means more number of sites of section 2 come under more risk.

```
library(maps)
four_corners<-map_data("state",region = c("arizona","utah","new mexico","colorado"))
fc<-four_corners
#deg_dec_min, deg_min_sec
navajo$Longitude<-measurements::conv_unit(navajo$Longitude,"deg_min_sec","dec_deg")
navajo$Latitude<-measurements::conv_unit(navajo$Latitude,"deg_min_sec","dec_deg")
nv2<-navajo
nv2$Longitude<-as.numeric(nv2$Longitude)
nv2$Latitude<-as.numeric(nv2$Latitude)
fc$long<-abs(fc$long)
ggplot(fc)+geom_polygon(mapping=aes(x=long,y=lat,group=group),fill="white",color="black")+geom_point(da
```



```
lea<-read_csv("D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA 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(),
##   LEAID = col_character(),
##   LEA_NAME = col_character(),
##   LEA_ADDRESS = col_character(),
##   LEA_CITY = col_character(),
##   CJJ = col_character(),
##   LEA_CRCOORD_SEX_IND = col_character(),
##   LEA_CRCOORD_RAC_IND = col_character(),
##   LEA_CRCOORD_DIS_IND = col_character(),
##   LEA_CRCOORD_SEX_FN = col_character(),
##   LEA_CRCOORD_SEX_LN = col_character(),
##   LEA_CRCOORD_SEX_PH = col_character(),
##   LEA_CRCOORD_SEX_EM = col_character(),
##   LEA_CRCOORD_RAC_FN = col_character(),
##   LEA_CRCOORD_RAC_LN = col_character(),
##   LEA_CRCOORD_RAC_PH = col_character(),
##   LEA_CRCOORD_RAC_EM = col_character(),
##   LEA_CRCOORD_DIS_FN = col_character(),
##   LEA_CRCOORD_DIS_LN = col_character()
##   # ... with 37 more columns
## )

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

## Warning: 263 parsing failures.
##  row                col           expected actual                                              file
## 1133 LEA_GEDCRED_LEP_M 1/0/T/F/TRUE/FALSE      4  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA Data.csv'
## 1133 LEA_GEDCRED_LEP_F 1/0/T/F/TRUE/FALSE      4  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA Data.csv'
## 1186 LEA_GEDCRED_AM_F  1/0/T/F/TRUE/FALSE      4  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA Data.csv'
## 1186 LEA_GEDCRED_LEP_M 1/0/T/F/TRUE/FALSE      7  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA Data.csv'
## 1199 LEA_GEDCRED_AS_M  1/0/T/F/TRUE/FALSE      13 'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 LEA Data.csv'
## .... ................. .................. ...... ..................................................
## See problems(...) for more details.
```

```
school<-read_csv("D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School Data.csv",na=c("-2","-5","-6","-7","-8","-
```

```
## 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.

## Warning: 5577 parsing failures.
##  row                     col           expected actual
## 1401 SCH_ALGPASS_GS1112_AM_M  1/0/T/F/TRUE/FALSE      4  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School
## 1401 SCH_ALGPASS_GS1112_AM_F  1/0/T/F/TRUE/FALSE      7  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School
## 1403 SCH_ALGPASS_GS1112_AM_M  1/0/T/F/TRUE/FALSE      43 'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School
## 1403 SCH_ALGPASS_GS1112_AM_F  1/0/T/F/TRUE/FALSE      25 'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School
## 1403 SCH_ALGPASS_GS1112_LEP_F 1/0/T/F/TRUE/FALSE      4  'D:/Spring 20 Sem 2/DMP/CRDC 2015-16 School
## .... ........................ .................. ...... .............................................
## See problems(...) for more details.
```

```r
head(lea)
```

```
## # A tibble: 6 x 115
##   LEA_STATE LEA_STATE_NAME LEAID LEA_NAME LEA_ADDRESS LEA_CITY LEA_ZIP CJJ
##   <chr>     <chr>          <chr> <chr>    <chr>       <chr>      <dbl> <chr>
## 1 AL        ALABAMA        0100~ Alabama~ P O Box 66  Mt Meigs   36057 Yes
## 2 AL        ALABAMA        0100~ Albertv~ 107 West M~ Albertv~   35950 No
## 3 AL        ALABAMA        0100~ Marshal~ 12380 US H~ Gunters~   35976 No
## 4 AL        ALABAMA        0100~ Hoover ~ 2810 Metro~ Hoover     35243 No
## 5 AL        ALABAMA        0100~ Madison~ 211 Celtic~ Madison    35758 Yes
## 6 AL        ALABAMA        0100~ Al Inst~ P O Drawer~ Tallade~   35161 No
## # ... with 107 more variables: LEA_ENR <dbl>, LEA_ENR_NONLEAFAC <dbl>,
## #   LEA_SCHOOLS <dbl>, LEA_CRCOORD_SEX_IND <chr>, LEA_CRCOORD_RAC_IND <chr>,
## #   LEA_CRCOORD_DIS_IND <chr>, LEA_CRCOORD_SEX_FN <chr>,
## #   LEA_CRCOORD_SEX_LN <chr>, LEA_CRCOORD_SEX_PH <chr>,
## #   LEA_CRCOORD_SEX_EM <chr>, LEA_CRCOORD_RAC_FN <chr>,
## #   LEA_CRCOORD_RAC_LN <chr>, LEA_CRCOORD_RAC_PH <chr>,
## #   LEA_CRCOORD_RAC_EM <chr>, LEA_CRCOORD_DIS_FN <chr>,
## #   LEA_CRCOORD_DIS_LN <chr>, LEA_CRCOORD_DIS_PH <chr>,
## #   LEA_CRCOORD_DIS_EM <chr>, LEA_DESEGPLAN <chr>, LEA_HBPOLICY_IND <chr>,
## #   LEA_HBPOLICYURL_IND <chr>, LEA_HBPOLICY_URL <chr>, LEA_ECE_IND <chr>,
## #   LEA_ECE_NONIDEA <chr>, LEA_PS_IND <chr>, LEA_PS_FULLDAYFREE <chr>,
## #   LEA_PS_FULLDAYCOST <chr>, LEA_PS_PARTDAYFREE <chr>,
## #   LEA_PS_PARTDAYCOST <chr>, LEA_PSENR_NONIDEA_A3 <chr>,
## #   LEA_PSENR_NONIDEA_A4 <chr>, LEA_PSENR_NONIDEA_A5 <chr>, LEA_PSENR_A2 <dbl>,
## #   LEA_PSENR_A3 <dbl>, LEA_PSENR_A4 <dbl>, LEA_PSENR_A5 <dbl>,
## #   LEA_PSELIG_ALL <chr>, LEA_PSELIG_IDEA <chr>, LEA_PSELIG_TITLEI <chr>,
## #   LEA_PSELIG_LOWINC <chr>, LEA_KG_IND <chr>, LEA_KG_FULLDAYFREE <chr>,
## #   LEA_KG_FULLDAYCOST <chr>, LEA_KG_PARTDAYFREE <chr>,
## #   LEA_KG_PARTDAYCOST <chr>, LEA_GED_IND <chr>, LEA_GEDPART_HI_M <dbl>,
## #   LEA_GEDPART_HI_F <dbl>, LEA_GEDPART_AM_M <dbl>, LEA_GEDPART_AM_F <dbl>,
## #   LEA_GEDPART_AS_M <dbl>, LEA_GEDPART_AS_F <dbl>, LEA_GEDPART_HP_M <dbl>,
## #   LEA_GEDPART_HP_F <dbl>, LEA_GEDPART_BL_M <dbl>, LEA_GEDPART_BL_F <dbl>,
## #   LEA_GEDPART_WH_M <dbl>, LEA_GEDPART_WH_F <dbl>, LEA_GEDPART_TR_M <dbl>,
## #   LEA_GEDPART_TR_F <dbl>, TOT_GEDPART_M <dbl>, TOT_GEDPART_F <dbl>,
## #   LEA_GEDPART_LEP_M <dbl>, LEA_GEDPART_LEP_F <dbl>, LEA_GEDPART_IDEA_M <dbl>,
## #   LEA_GEDPART_IDEA_F <dbl>, LEA_GEDCRED_HI_M <dbl>, LEA_GEDCRED_HI_F <dbl>,
## #   LEA_GEDCRED_AM_M <dbl>, LEA_GEDCRED_AM_F <lgl>, LEA_GEDCRED_AS_M <lgl>,
## #   LEA_GEDCRED_AS_F <lgl>, LEA_GEDCRED_HP_M <lgl>, LEA_GEDCRED_HP_F <lgl>,
## #   LEA_GEDCRED_BL_M <dbl>, LEA_GEDCRED_BL_F <dbl>, LEA_GEDCRED_WH_M <dbl>,
## #   LEA_GEDCRED_WH_F <dbl>, LEA_GEDCRED_TR_M <lgl>, LEA_GEDCRED_TR_F <lgl>,
## #   TOT_GEDCRED_M <dbl>, TOT_GEDCRED_F <dbl>, LEA_GEDCRED_LEP_M <lgl>,
## #   LEA_GEDCRED_LEP_F <lgl>, LEA_GEDCRED_IDEA_M <dbl>,
## #   LEA_GEDCRED_IDEA_F <lgl>, LEA_DISTED_IND <chr>, LEA_DISTEDENR_HI_M <dbl>,
## #   LEA_DISTEDENR_HI_F <dbl>, LEA_DISTEDENR_AM_M <dbl>,
## #   LEA_DISTEDENR_AM_F <dbl>, LEA_DISTEDENR_AS_M <dbl>,
## #   LEA_DISTEDENR_AS_F <dbl>, LEA_DISTEDENR_HP_M <dbl>,
## #   LEA_DISTEDENR_HP_F <dbl>, LEA_DISTEDENR_BL_M <dbl>,
## #   LEA_DISTEDENR_BL_F <dbl>, LEA_DISTEDENR_WH_M <dbl>,
## #   LEA_DISTEDENR_WH_F <dbl>, LEA_DISTEDENR_TR_M <dbl>, ...
```

```r
dim(school)
```

```
## [1] 96360  1836
```

```
head(school)
```

```
## # A tibble: 6 x 1,836
##   LEA_STATE LEA_STATE_NAME  LEAID LEA_NAME SCHID SCH_NAME COMBOKEY JJ
##   <chr>     <chr>           <dbl> <chr>    <dbl> <chr>       <dbl> <chr>
## 1 AL        ALABAMA        100002 Alabama~  1705 Wallace~  1.00e10 Yes
## 2 AL        ALABAMA        100002 Alabama~  1706 McNeel ~  1.00e10 Yes
## 3 AL        ALABAMA        100002 Alabama~  1876 Alabama~  1.00e10 No
## 4 AL        ALABAMA        100002 Alabama~ 99995 AUTAUGA~  1.00e10 Yes
## 5 AL        ALABAMA        100005 Albertv~   870 Albertv~  1.00e10 No
## 6 AL        ALABAMA        100005 Albertv~   871 Albertv~  1.00e10 No
## # ... with 1,828 more variables: SCH_GRADE_PS <chr>, SCH_GRADE_KG <chr>,
## #   SCH_GRADE_G01 <chr>, SCH_GRADE_G02 <chr>, SCH_GRADE_G03 <chr>,
## #   SCH_GRADE_G04 <chr>, SCH_GRADE_G05 <chr>, SCH_GRADE_G06 <chr>,
## #   SCH_GRADE_G07 <chr>, SCH_GRADE_G08 <chr>, SCH_GRADE_G09 <chr>,
## #   SCH_GRADE_G10 <chr>, SCH_GRADE_G11 <chr>, SCH_GRADE_G12 <chr>,
## #   SCH_GRADE_UG <chr>, SCH_UGDETAIL_ES <chr>, SCH_UGDETAIL_MS <chr>,
## #   SCH_UGDETAIL_HS <chr>, SCH_STATUS_SPED <chr>, SCH_STATUS_MAGNET <chr>,
## #   SCH_STATUS_CHARTER <chr>, SCH_STATUS_ALT <chr>, SCH_MAGNETDETAIL <chr>,
## #   SCH_ALTFOCUS <chr>, SCH_PSENR_NONIDEA_A3 <chr>, SCH_PSENR_NONIDEA_A4 <chr>,
## #   SCH_PSENR_NONIDEA_A5 <chr>, SCH_PSENR_HI_M <dbl>, SCH_PSENR_HI_F <dbl>,
## #   SCH_PSENR_AM_M <dbl>, SCH_PSENR_AM_F <dbl>, SCH_PSENR_AS_M <dbl>,
## #   SCH_PSENR_AS_F <dbl>, SCH_PSENR_HP_M <dbl>, SCH_PSENR_HP_F <dbl>,
## #   SCH_PSENR_BL_M <dbl>, SCH_PSENR_BL_F <dbl>, SCH_PSENR_WH_M <dbl>,
## #   SCH_PSENR_WH_F <dbl>, SCH_PSENR_TR_M <dbl>, SCH_PSENR_TR_F <dbl>,
## #   TOT_PSENR_M <dbl>, TOT_PSENR_F <dbl>, SCH_PSENR_LEP_M <dbl>,
## #   SCH_PSENR_LEP_F <dbl>, SCH_PSENR_IDEA_M <dbl>, SCH_PSENR_IDEA_F <dbl>,
## #   SCH_ENR_HI_M <dbl>, SCH_ENR_HI_F <dbl>, SCH_ENR_AM_M <dbl>,
## #   SCH_ENR_AM_F <dbl>, SCH_ENR_AS_M <dbl>, SCH_ENR_AS_F <dbl>,
## #   SCH_ENR_HP_M <dbl>, SCH_ENR_HP_F <dbl>, SCH_ENR_BL_M <dbl>,
## #   SCH_ENR_BL_F <dbl>, SCH_ENR_WH_M <dbl>, SCH_ENR_WH_F <dbl>,
## #   SCH_ENR_TR_M <dbl>, SCH_ENR_TR_F <dbl>, TOT_ENR_M <dbl>, TOT_ENR_F <dbl>,
## #   SCH_ENR_LEP_M <dbl>, SCH_ENR_LEP_F <dbl>, SCH_ENR_504_M <dbl>,
## #   SCH_ENR_504_F <dbl>, SCH_ENR_IDEA_M <dbl>, SCH_ENR_IDEA_F <dbl>,
## #   SCH_LEPENR_HI_M <dbl>, SCH_LEPENR_HI_F <dbl>, SCH_LEPENR_AM_M <dbl>,
## #   SCH_LEPENR_AM_F <dbl>, SCH_LEPENR_AS_M <dbl>, SCH_LEPENR_AS_F <dbl>,
## #   SCH_LEPENR_HP_M <dbl>, SCH_LEPENR_HP_F <dbl>, SCH_LEPENR_BL_M <dbl>,
## #   SCH_LEPENR_BL_F <dbl>, SCH_LEPENR_WH_M <dbl>, SCH_LEPENR_WH_F <dbl>,
## #   SCH_LEPENR_TR_M <dbl>, SCH_LEPENR_TR_F <dbl>, TOT_LEPENR_M <dbl>,
## #   TOT_LEPENR_F <dbl>, SCH_LEPPROGENR_HI_M <dbl>, SCH_LEPPROGENR_HI_F <dbl>,
## #   SCH_LEPPROGENR_AM_M <dbl>, SCH_LEPPROGENR_AM_F <dbl>,
## #   SCH_LEPPROGENR_AS_M <dbl>, SCH_LEPPROGENR_AS_F <dbl>,
## #   SCH_LEPPROGENR_HP_M <dbl>, SCH_LEPPROGENR_HP_F <dbl>,
## #   SCH_LEPPROGENR_BL_M <dbl>, SCH_LEPPROGENR_BL_F <dbl>,
## #   SCH_LEPPROGENR_WH_M <dbl>, SCH_LEPPROGENR_WH_F <dbl>,
## #   SCH_LEPPROGENR_TR_M <dbl>, SCH_LEPPROGENR_TR_F <dbl>,
## #   TOT_LEPPROGENR_M <dbl>, ...
```

```
school1<-school
```

```
school1$tot_stud=school1$TOT_ENR_M+school1$TOT_ENR_F
school1$tot_stud_black=school1$SCH_ENR_BL_M+school1$SCH_ENR_BL_F
school1$tot_suspension=school1$TOT_DISCWDIS_ISS_IDEA_M + school1$TOT_DISCWDIS_ISS_IDEA_F+school1$TOT_DI
school1$tot_suspension_black=school1$SCH_DISCWODIS_ISS_BL_M + school1$SCH_DISCWODIS_ISS_BL_F + school1$
school1$proportion_black=school1$tot_stud_black/school1$tot_stud
school1$susp_prop_black = school1$tot_suspension_black/school1$tot_suspension
x<-select(school1, tot_stud,tot_stud_black,tot_suspension,tot_suspension_black,proportion_black,susp_pr
```

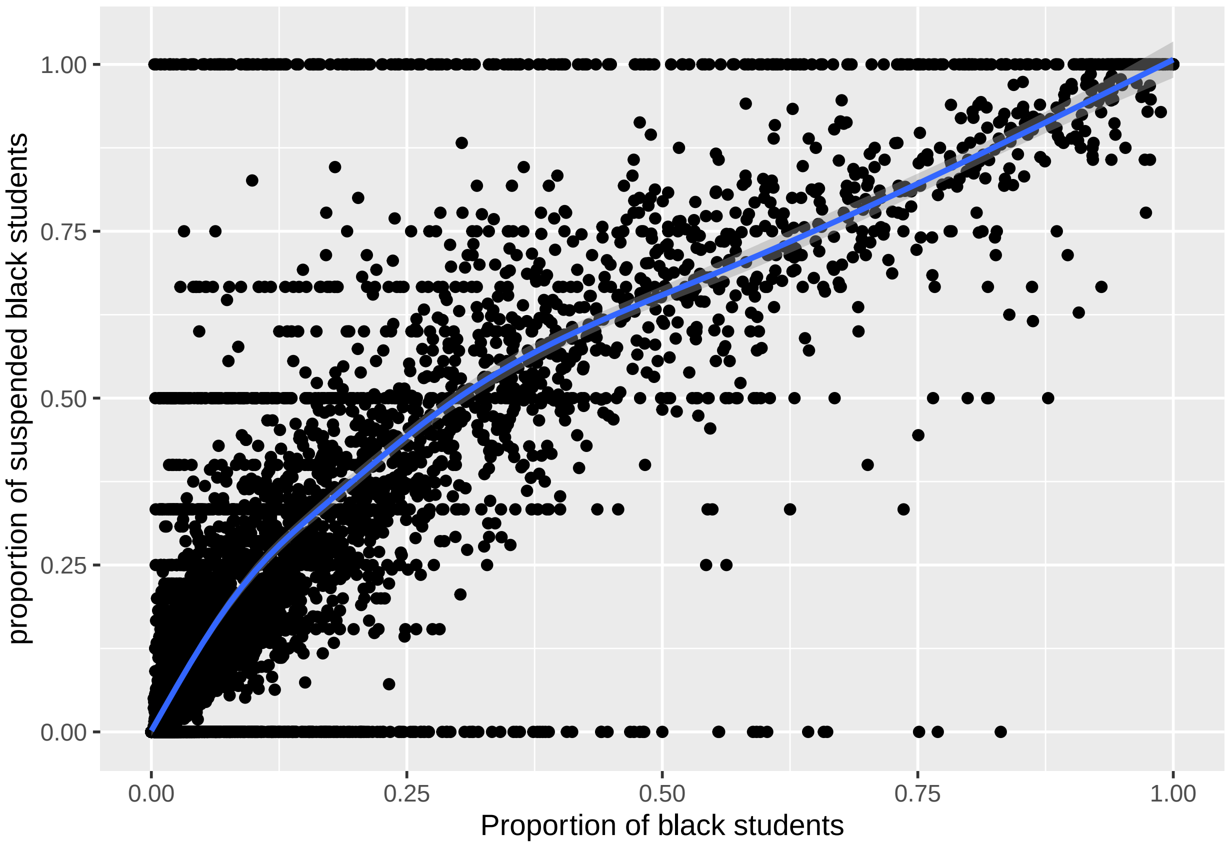
From the scatterplot below, we can see that proportion of suspended black students is more in schools where proportion of black students is less.

```
x%>%sample_n(10000)%>%ggplot(aes(x=proportion_black,y=susp_prop_black))+geom_point()+geom_smooth()+xlab
```

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

```
## Warning: Removed 3586 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3586 rows containing missing values (geom_point).
```



In graph above we can observe under representation of black students. Most of the suspended black students are within 50% of proportion of black students.Also as the proportion of black student increases proportion of suspended black students steadily decreases.

```
tot_student<-sum(x$tot_stud,na.rm=TRUE)
tot_black<-sum(x$tot_stud_black,na.rm=TRUE)
tot_black_suspended<-sum(x$tot_suspension_black,na.rm=TRUE)
tot_stud_suspended<-sum(x$tot_suspension,na.rm=TRUE)
prop_black<-tot_black/tot_student
prop_black_suspended<-tot_black_suspended/tot_stud_suspended
cat("Overall black student proportion:",prop_black)
```

```
## Overall black student proportion: 0.1543446
```

```
cat("Overall suspended black student proportion:",prop_black_suspended)
```

```
## Overall suspended black student proportion: 0.3212122
```

“‘ From above data, black students are under represented in school suspension as proportion of students suspended who are black holds smaller percentage of population.