## Exploratory Data Analysis

```
library(readr)
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
```

```
library(tidyr)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.1.8     v stringr 1.5.0
## v purrr   1.0.1     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RSQLite)
```

Using data from the US Department of Education's Civil Rights Data Collection. It was downloaded from the zipped 2017-2018 data available at https://www2.ed.gov/about/offices/list/ocr/docs/crdc-2017-18.html.

```
setwd('..')
dir <- getwd()
```

1. The distribution of students by race and gender across all schools

```
# load enrollment data

dir1 <- "2017-18-crdc-data-corrected-publication 2/2017-18 Public-Use Files/Data/SCH/CRDC/CSV/Enrollmen
path <- paste(dir, dir1, sep="/")
enrollment <- read_csv(file=path)
```

```
## Rows: 97632 Columns: 123
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (11): LEA_STATE, LEA_STATE_NAME, LEAID, LEA_NAME, SCHID, SCH_NAME, COMB...
## dbl (112): SCH_PSENR_HI_M, SCH_PSENR_HI_F, SCH_PSENR_AM_M, SCH_PSENR_AM_F, S...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# there are negative values for enrollment count (CRDC reserve codes)
# replacing them with 0

numeric_cols <- sapply(enrollment, is.numeric)
enrollment[numeric_cols][enrollment[numeric_cols] < 0] <- 0
enrollment
```

```
## # A tibble: 97,632 x 123
##    LEA_STATE LEA_STA~1 LEAID LEA_N~2 SCHID SCH_N~3 COMBO~4 JJ    SCH_P~5 SCH_P~6
##    <chr>     <chr>     <chr> <chr>   <chr> <chr>   <chr>   <chr> <chr>   <chr>
##  1 AL        ALABAMA   0100~ Alabam~ 01705 Wallac~ 010000~ Yes   -9      -9
##  2 AL        ALABAMA   0100~ Alabam~ 01706 McNeel~ 010000~ Yes   -9      -9
##  3 AL        ALABAMA   0100~ Alabam~ 01876 Alabam~ 010000~ No    -9      -9
##  4 AL        ALABAMA   0100~ Alabam~ 99995 AUTAUG~ 010000~ Yes   -9      -9
##  5 AL        ALABAMA   0100~ Albert~ 00870 Albert~ 010000~ No    -9      -9
##  6 AL        ALABAMA   0100~ Albert~ 00871 Albert~ 010000~ No    -9      -9
##  7 AL        ALABAMA   0100~ Albert~ 00879 Evans ~ 010000~ No    -9      -9
##  8 AL        ALABAMA   0100~ Albert~ 00889 Albert~ 010000~ No    -9      -9
##  9 AL        ALABAMA   0100~ Albert~ 01616 Big Sp~ 010000~ No    -9      -9
## 10 AL        ALABAMA   0100~ Albert~ 02150 Albert~ 010000~ No    Yes     Yes
## # ... with 97,622 more rows, 113 more variables: 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>, ...
```

```r
# total enrollment across all schools
enr_totals <- select(enrollment, starts_with('TOT_ENR')) %>% colSums()
total_enrollment <- enr_totals['TOT_ENR_M'] + enr_totals['TOT_ENR_F']
```

```r
# get all ethnicity enrollments
enr_req <- select(enrollment, starts_with('SCH_ENR'))

cols <- names(enr_req)[grep("^SCH_ENR_", names(enr_req))]
cols_m <- names(enr_req)[grep("_M$", names(enr_req))]
cols_f <- names(enr_req)[grep("_F$", names(enr_req))]

# sum all enrollments for each ethnicity and gender
enr_req <- enr_req %>% pivot_longer(cols, names_to = "ethnicity", values_to = "count")
```

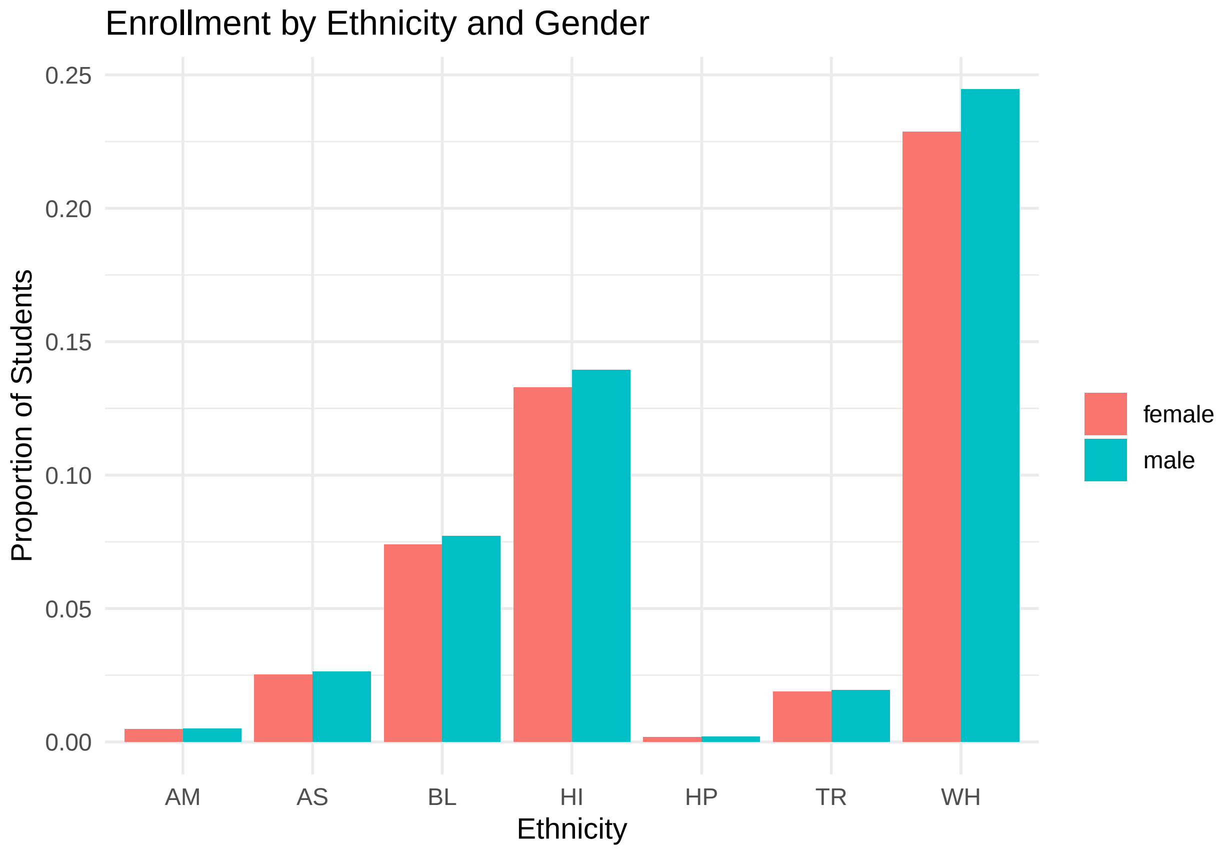
```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(cols)
##
##   # Now:
##   data %>% select(all_of(cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

```r
enr_req <- summarise(group_by(enr_req, ethnicity), count = sum(count))

enr_req$gender <- ifelse(grepl("_M$", enr_req$ethnicity),  "male" , "female")

# remove prefix and sufix from ethnicity
enr_req$ethnicity <- sub("_M$", "", enr_req$ethnicity)
enr_req$ethnicity <- sub("_F$", "", enr_req$ethnicity)
enr_req$ethnicity <- sub("^SCH_ENR_", "", enr_req$ethnicity)
enr_req$ethnicity <- sub("^SCH_ENR_", "", enr_req$ethnicity)
```

```
# remove extra columns
enr_req <- subset(enr_req, !(ethnicity %in% c("504", "IDEA", "LEP")))

# calculate proportion
enr_req <- enr_req %>% mutate(prop = count / total_enrollment)

enr_req
```

```
## # A tibble: 14 x 4
##    ethnicity    count gender    prop
##    <chr>        <dbl> <chr>    <dbl>
##  1 AM          245129 female 0.00481
##  2 AM          257342 male   0.00505
##  3 AS         1281702 female 0.0252
##  4 AS         1344407 male   0.0264
##  5 BL         3763447 female 0.0739
##  6 BL         3933267 male   0.0772
##  7 HI         6763088 female 0.133
##  8 HI         7099395 male   0.139
##  9 HP           93838 female 0.00184
## 10 HP           99586 male   0.00196
## 11 TR          957267 female 0.0188
## 12 TR          987608 male   0.0194
## 13 WH        11646416 female 0.229
## 14 WH        12449909 male   0.244
```

```
ggplot(enr_req, aes(x = ethnicity, y = prop, fill = gender)) +
  geom_bar(stat = "identity", position=position_dodge()) +
  labs(title = "Enrollment by Ethnicity and Gender",
       x = "Ethnicity",
       y = "Proportion of Students",
       fill = "") +
  theme_minimal()
```



We can see that the proportion of white students is very high in comparison to other races. Lowest are Hawaiian/Pacific Islanders, American Indian/Alaska Native, Asians and students belonging to two or more races (TR) - all less than 5%. Male - female proportions are almost equal in all races (female enrollment is slightly lower).

2. The distribution of Advanced Placement (AP) students (i.e., students enrolled in at least one AP course) by race and gender across all schools.

```
# read AP dataset
dir2 <- "2017-18-crdc-data-corrected-publication 2/2017-18 Public-Use Files/Data/SCH/CRDC/CSV/Advanced
path <- paste(dir, dir2, sep="/")
ap <- read_csv(file=path, na="-9")
```

```
## Rows: 97632 Columns: 134
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (13): LEA_STATE, LEA_STATE_NAME, LEAID, LEA_NAME, SCHID, SCH_NAME, COMB...
## dbl (121): SCH_APCOURSES, SCH_APENR_HI_M, SCH_APENR_HI_F, SCH_APENR_AM_M, SC...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

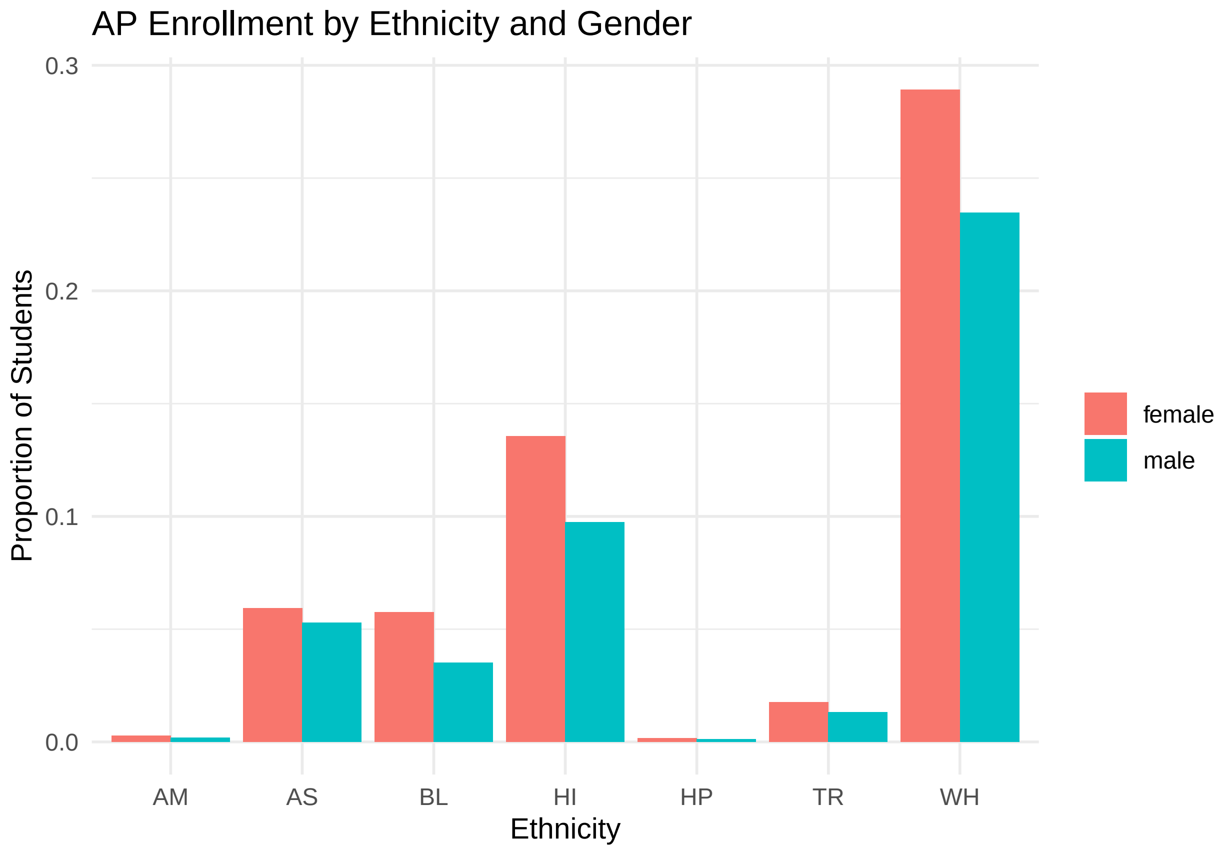
```
# there are negative values (CRDC reserve codes)
# replacing them with 0
numeric_cols <- sapply(ap, is.numeric)
ap[numeric_cols][ap[numeric_cols] < 0] <- 0
ap
```

```
## # A tibble: 97,632 x 134
##    LEA_STATE LEA_STA~1 LEAID LEA_N~2 SCHID SCH_N~3 COMBO~4 JJ    SCH_A~5 SCH_A~6
##    <chr>     <chr>     <chr> <chr>   <chr> <chr>   <chr>   <chr> <chr>     <dbl>
##  1 AL        ALABAMA   0100~ Alabam~ 01705 Wallac~ 010000~ Yes   <NA>         NA
##  2 AL        ALABAMA   0100~ Alabam~ 01706 McNeel~ 010000~ Yes   <NA>         NA
##  3 AL        ALABAMA   0100~ Alabam~ 01876 Alabam~ 010000~ No    No           NA
##  4 AL        ALABAMA   0100~ Alabam~ 99995 AUTAUG~ 010000~ Yes   <NA>         NA
##  5 AL        ALABAMA   0100~ Albert~ 00870 Albert~ 010000~ No    <NA>         NA
##  6 AL        ALABAMA   0100~ Albert~ 00871 Albert~ 010000~ No    Yes           8
##  7 AL        ALABAMA   0100~ Albert~ 00879 Evans ~ 010000~ No    <NA>         NA
##  8 AL        ALABAMA   0100~ Albert~ 00889 Albert~ 010000~ No    <NA>         NA
##  9 AL        ALABAMA   0100~ Albert~ 01616 Big Sp~ 010000~ No    <NA>         NA
## 10 AL        ALABAMA   0100~ Albert~ 02150 Albert~ 010000~ No    <NA>         NA
## # ... with 97,622 more rows, 124 more variables: SCH_APSEL <chr>,
## #   SCH_APENR_HI_M <dbl>, SCH_APENR_HI_F <dbl>, SCH_APENR_AM_M <dbl>,
## #   SCH_APENR_AM_F <dbl>, SCH_APENR_AS_M <dbl>, SCH_APENR_AS_F <dbl>,
## #   SCH_APENR_HP_M <dbl>, SCH_APENR_HP_F <dbl>, SCH_APENR_BL_M <dbl>,
## #   SCH_APENR_BL_F <dbl>, SCH_APENR_WH_M <dbl>, SCH_APENR_WH_F <dbl>,
## #   SCH_APENR_TR_M <dbl>, SCH_APENR_TR_F <dbl>, TOT_APENR_M <dbl>,
## #   TOT_APENR_F <dbl>, SCH_APENR_LEP_M <dbl>, SCH_APENR_LEP_F <dbl>, ...
```

```r
# filter only schools with at least one AP course
ap <- filter(ap, SCH_APENR_IND == 'Yes')

# calculate AP total enrollment
ap_enr_totals <- select(ap, starts_with('TOT_APENR_')) %>% colSums()
ap_total_enrollment <- ap_enr_totals['TOT_APENR_M'] + ap_enr_totals['TOT_APENR_F']

# take only necesarry columns
ap_enr_req <- select(ap, starts_with('SCH_APENR_'))
ap_enr_req <- subset(ap_enr_req, select = -SCH_APENR_IND)

cols <- names(ap_enr_req)[grep("^SCH_APENR_", names(ap_enr_req))]
cols_m <- names(ap_enr_req)[grep("_M$", names(ap_enr_req))]
cols_f <- names(ap_enr_req)[grep("_F$", names(ap_enr_req))]

# sum all enrollments for each ethnicity and gender combination
ap_enr_req <- ap_enr_req %>% pivot_longer(cols, names_to = "ethnicity", values_to = "count")
ap_enr_req <- summarise(group_by(ap_enr_req, ethnicity), count = sum(count))
# separate male and female from ethnicity
ap_enr_req$gender <- ifelse(grepl("_M$", ap_enr_req$ethnicity),  "male" , "female")

# remove suffix and prefix from names
ap_enr_req$ethnicity <- sub("_M$", "", ap_enr_req$ethnicity)
ap_enr_req$ethnicity <- sub("_F$", "", ap_enr_req$ethnicity)
ap_enr_req$ethnicity <- sub("^SCH_APENR_", "", ap_enr_req$ethnicity)
ap_enr_req$ethnicity <- sub("^SCH_APENR_", "", ap_enr_req$ethnicity)

ap_enr_req <- subset(ap_enr_req, !(ethnicity %in% c("504", "IDEA", "LEP")))

# calculate proportion
ap_enr_req <- ap_enr_req %>% mutate(prop = count / ap_total_enrollment)

ap_enr_req
```

```
## # A tibble: 14 x 4
##    ethnicity  count gender    prop
##    <chr>      <dbl> <chr>    <dbl>
##  1 AM          8475 female 0.00280
##  2 AM          5811 male   0.00192
##  3 AS        179533 female 0.0592
##  4 AS        160350 male   0.0529
##  5 BL        174634 female 0.0576
##  6 BL        106512 male   0.0351
##  7 HI        410834 female 0.136
##  8 HI        295258 male   0.0974
##  9 HP          5118 female 0.00169
## 10 HP          3599 male   0.00119
## 11 TR         53437 female 0.0176
## 12 TR         40209 male   0.0133
## 13 WH        876243 female 0.289
## 14 WH        710978 male   0.235
```

```
ggplot(ap_enr_req, aes(x = ethnicity, y = prop, fill = gender)) +
  geom_bar(stat = "identity", position=position_dodge()) +
  labs(title = "AP Enrollment by Ethnicity and Gender",
       x = "Ethnicity",
       y = "Proportion of Students",
       fill = "") +
  theme_minimal()
```



We can see that the proportion of white students taking at least one AP course is high in comparison to other races. Lowest are Hispanic, Indian Americans and students belonging to two or more races (TR) - all less than 5%. This is similar to overall enrollment plot in problem 1. Other than that, Asians and Black ethnic groups have similar number of enrollments. There is more female enrollment in AP courses for all races which is different compared to previous plot.

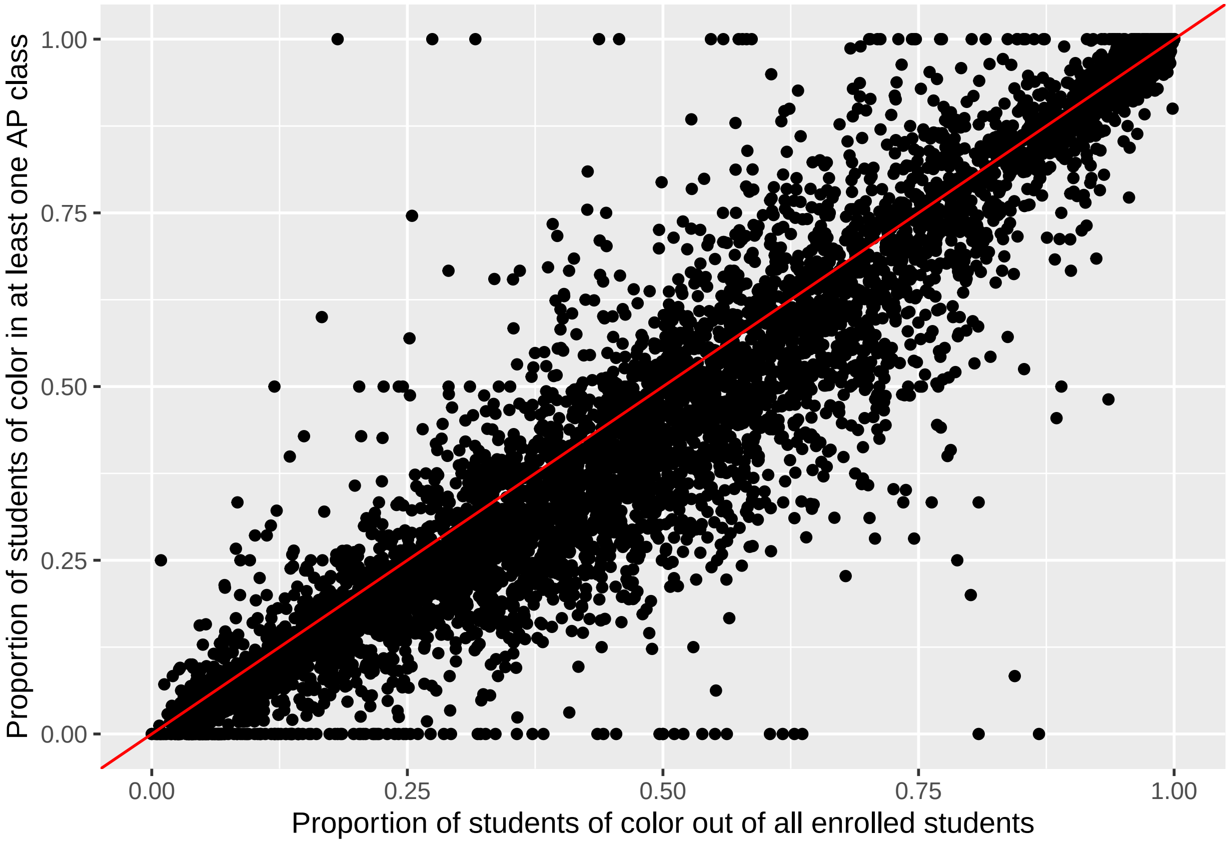
3. Visualize whether there is a trend of students of color (i.e., non-white students) being underrepresented in AP programs at schools.

```
# merge enrollment data and filtered ap data
cols_enr <- names(enrollment)[grep("^SCH_ENR_", names(enrollment))]
cols_enr <- cols_enr[!(cols_enr %in% c("SCH_ENR_LEP_M","SCH_ENR_LEP_F",
                                       "SCH_ENR_504_M","SCH_ENR_504_F",
                                       "SCH_ENR_IDEA_M","SCH_ENR_IDEA_F"))]

cols_ap_enr <- names(ap)[grep("^SCH_APENR_", names(ap))]
cols_ap_enr <- cols_ap_enr[!(cols_ap_enr %in% c("SCH_APENR_LEP_M",
                                                "SCH_APENR_LEP_F",
                                                "SCH_APENR_IDEA_M",
                                                "SCH_APENR_IDEA_F",
                                                "SCH_APENR_IND"))]

cols <- c('SCHID',cols_enr, cols_ap_enr, 'TOT_ENR_M','TOT_ENR_F',
          'TOT_APENR_M','TOT_APENR_F')

merged_df <- select(merge(x=enrollment, y=ap, by=c('COMBOKEY', 'SCHID')), cols)

# group by school
df_summary <- merged_df %>%
  group_by(SCHID) %>%
  summarize(across(everything(), sum))

#calculate color students proportions for each school
df_summary <- df_summary %>% mutate(TOT_COLOR =
                                      rowSums(df_summary[, cols_enr[!(cols_enr
                                      %in% c("SCH_ENR_WH_M","SCH_ENR_WH_F"))]]),
                                    TOT_AP_COLOR =
                                      rowSums(df_summary[, cols_ap_enr[!(cols_ap_enr
                                      %in% c("SCH_APENR_WH_M","SCH_APENR_WH_F"))]]),
                                    PROP_COLOR = TOT_COLOR/(TOT_ENR_F + TOT_ENR_M),
                                    PROP_AP_COLOR = TOT_AP_COLOR/(TOT_APENR_F + TOT_APENR_M))

df_summary <- select(df_summary, c('SCHID','PROP_COLOR', 'PROP_AP_COLOR'))
df_summary
```

```
## # A tibble: 5,925 x 3
##    SCHID PROP_COLOR PROP_AP_COLOR
##    <chr>      <dbl>         <dbl>
##  1 00001      0.721         0.737
##  2 00002      0.374         0.370
##  3 00003      0.428         0.401
##  4 00004      0.397         0.555
##  5 00005      0.606         0.642
##  6 00006      0.578         0.585
##  7 00007      0.596         0.500
##  8 00008      0.475         0.620
##  9 00009      0.474         0.417
## 10 00010      0.721         0.732
## # ... with 5,915 more rows
```

```
ggplot(df_summary, aes(x = PROP_COLOR, y = PROP_AP_COLOR)) +
  geom_point() +
  geom_abline(slope=1, color='red') +
  xlab("Proportion of students of color out of all enrolled students") +
  ylab("Proportion of students of color in at least one AP class")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



The points above the reference line are less dense than the points below. Most classes have almost equal representation. Some classes have little to no color representation. It can be said that color students are somewhat underrepresented.