DS5110 Homework 3

Parth Shah

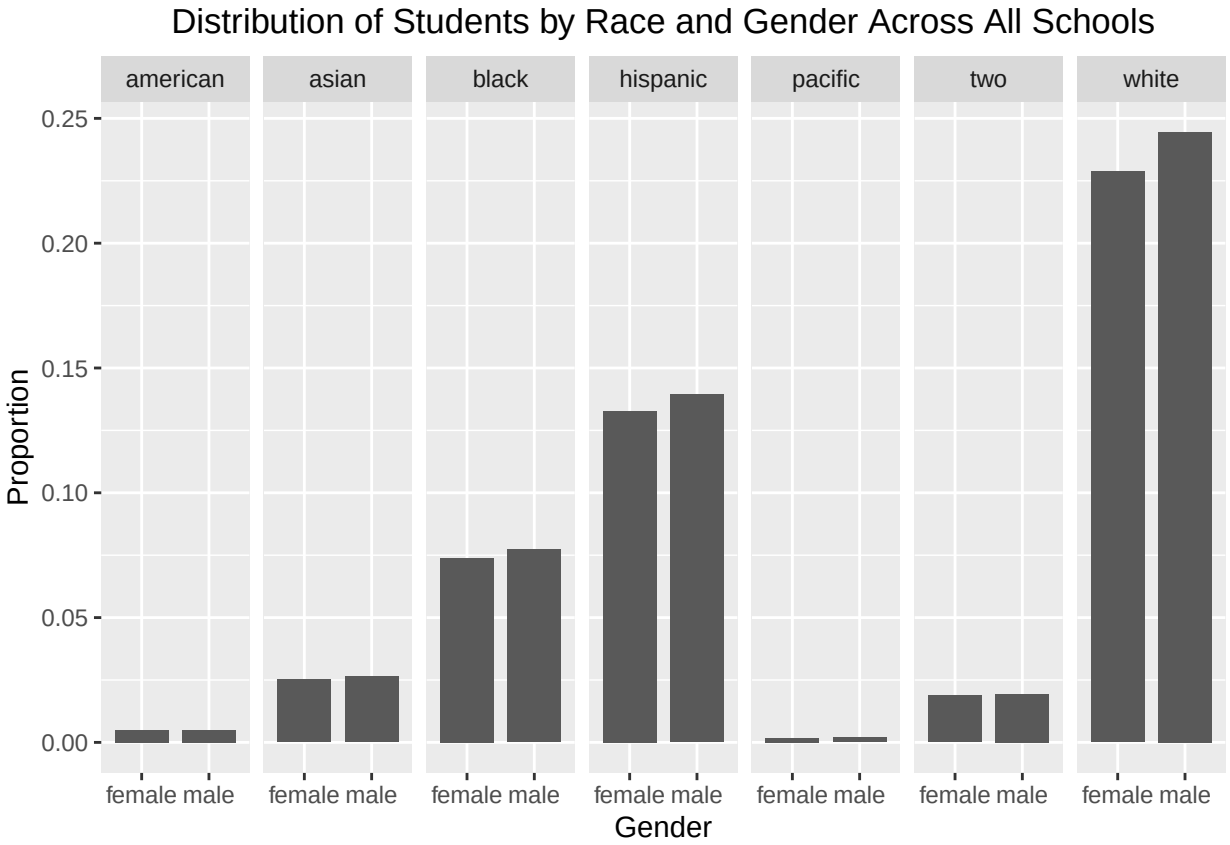
2023-02-27

Part A

Problem 1

```
# Load required libraries
suppressPackageStartupMessages(library(dplyr))
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
data <- read.csv("Enrollment.csv")

# Replace negative values with NA
neg_indices <- which(data < 0, arr.ind = TRUE)
data[neg_indices] <- NA

# Summarize the data to calculate total enrollment and enrollment by race and gender
data <- data %>%
  summarize(
    total_enrollment = sum(TOT_ENR_M + TOT_ENR_F, na.rm = TRUE),
    hispanic_male_full = sum(SCH_ENR_HI_M, na.rm = TRUE),
    hispanic_female_full = sum(SCH_ENR_HI_F, na.rm = TRUE),
    american_indian_male_full = sum(SCH_ENR_AM_M, na.rm = TRUE),
    american_indian_female_full = sum(SCH_ENR_AM_F, na.rm = TRUE),
    asian_male_full = sum(SCH_ENR_AS_M, na.rm = TRUE),
    asian_female_full = sum(SCH_ENR_AS_F, na.rm = TRUE),
    pacific_islander_male_full = sum(SCH_ENR_HP_M, na.rm = TRUE),
    pacific_islander_female_full = sum(SCH_ENR_HP_F, na.rm = TRUE),
    black_male_full = sum(SCH_ENR_BL_M, na.rm = TRUE),
    black_female_full = sum(SCH_ENR_BL_F, na.rm = TRUE),
    white_male_full = sum(SCH_ENR_WH_M, na.rm = TRUE),
    white_female_full = sum(SCH_ENR_WH_F, na.rm = TRUE),
    two_or_more_male_full = sum(SCH_ENR_TR_M, na.rm = TRUE),
    two_or_more_female_full = sum(SCH_ENR_TR_F, na.rm = TRUE),
  )

# Calculate the proportion of students by race and gender
total_enrollment_prop <- data %>%
  mutate(
```

```
    hispanic_male = hispanic_male_full / total_enrollment,
    hispanic_female = hispanic_female_full / total_enrollment,
    american_male = american_indian_male_full / total_enrollment,
    american_female = american_indian_female_full / total_enrollment,
    asian_male = asian_male_full / total_enrollment,
    asian_female = asian_female_full / total_enrollment,
    pacific_male = pacific_islander_male_full / total_enrollment,
    pacific_female = pacific_islander_female_full / total_enrollment,
    black_male = black_male_full / total_enrollment,
    black_female = black_female_full / total_enrollment,
    white_male = white_male_full / total_enrollment,
    white_female = white_female_full / total_enrollment,
    two_male = two_or_more_male_full / total_enrollment,
    two_female = two_or_more_female_full / total_enrollment,
  ) %>%
  select(ends_with('male'), ends_with('female'))

# Transform data for plotting
total_enrollment_prop %>%
  gather(key = "race_gender", value = "proportion") %>%
  separate(race_gender, into = c("race", "gender"), sep = "_") %>%
  ggplot(aes(x = gender, y = proportion)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(x = "Gender", y = "Proportion",
       title = "Distribution of Students by Race and Gender Across All Schools") +
  scale_fill_brewer(palette = "Set1") +
  theme(plot.title = element_text(hjust = 0.5)) +
  facet_grid(. ~ race)
```

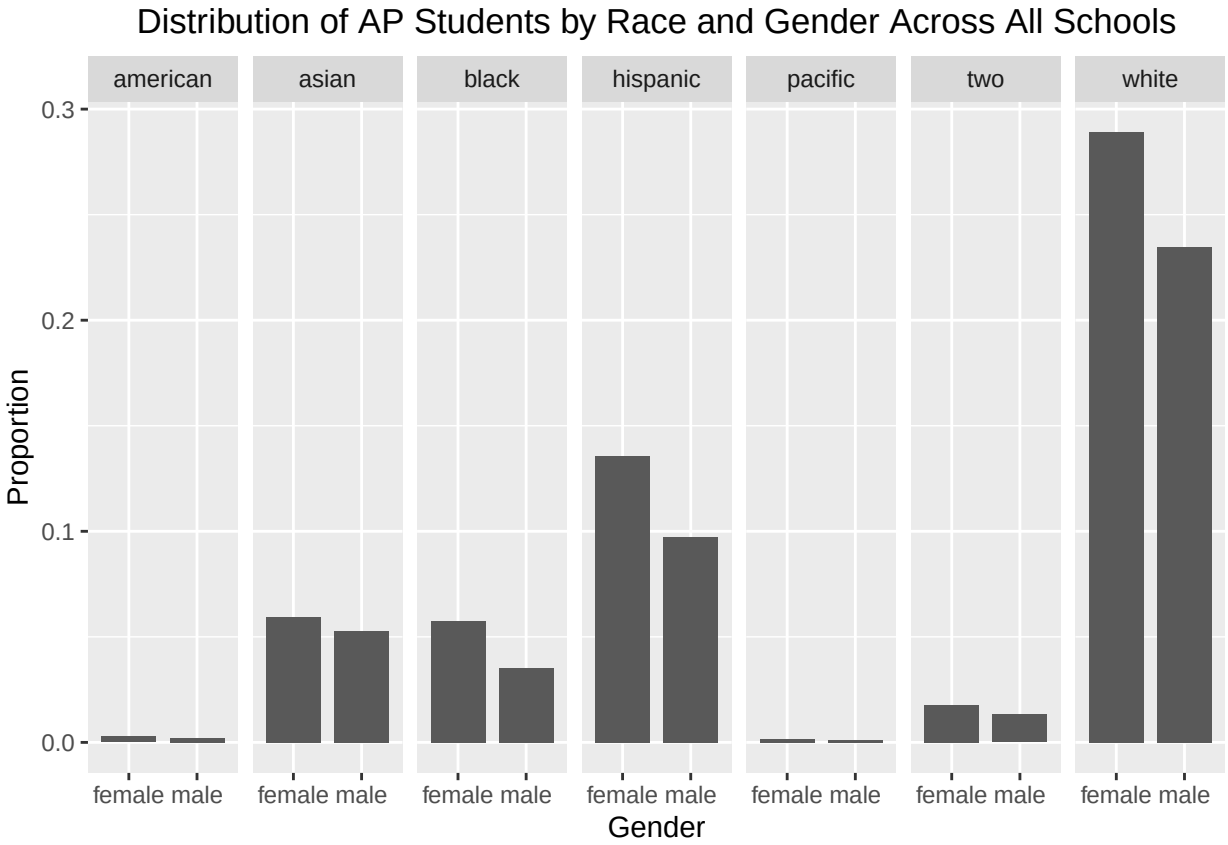


The plot indicates that White students have the highest number of enrollments among all races, regardless of gender. Male enrollment either surpasses or is comparable to female enrollment across all races. Conversely, Pacific Islanders and American Indians have notably lower enrollment rates compared to other races. White enrollment is almost twice as high as the second most enrolled race, which is Hispanic. Generally, there is considerable inequality in enrollments across races, while similarities exist in gender enrollment rates across all races.

Problem 2

```r
# Load required libraries
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
data <- read.csv("Advanced Placement.csv")

# Replace negative values with NA
neg_indices <- which(data < 0, arr.ind = TRUE)
data[neg_indices] <- NA

# Filter out the schools where AP courses are conducted
data <- data[!is.na(data$SCH_APENR_IND) & (data$SCH_APENR_IND == "Yes"),]

# Summarize the data to calculate total enrollment and enrollment by race
```

```r
# and gender for AP courses
data <- data %>%
  summarize(
    total_enrollment = sum(TOT_APENR_M + TOT_APENR_F, na.rm = TRUE),
    hispanic_male_full = sum(SCH_APENR_HI_M, na.rm = TRUE),
    hispanic_female_full = sum(SCH_APENR_HI_F, na.rm = TRUE),
    american_indian_male_full = sum(SCH_APENR_AM_M, na.rm = TRUE),
    american_indian_female_full = sum(SCH_APENR_AM_F, na.rm = TRUE),
    asian_male_full = sum(SCH_APENR_AS_M, na.rm = TRUE),
    asian_female_full = sum(SCH_APENR_AS_F, na.rm = TRUE),
    pacific_islander_male_full = sum(SCH_APENR_HP_M, na.rm = TRUE),
    pacific_islander_female_full = sum(SCH_APENR_HP_F, na.rm = TRUE),
    black_male_full = sum(SCH_APENR_BL_M, na.rm = TRUE),
    black_female_full = sum(SCH_APENR_BL_F, na.rm = TRUE),
    white_male_full = sum(SCH_APENR_WH_M, na.rm = TRUE),
    white_female_full = sum(SCH_APENR_WH_F, na.rm = TRUE),
    two_or_more_male_full = sum(SCH_APENR_TR_M, na.rm = TRUE),
    two_or_more_female_full = sum(SCH_APENR_TR_F, na.rm = TRUE),
  )

# Calculate the proportion of students by race and gender
total_enrollment_prop <- data %>%
  mutate(
    hispanic_male = hispanic_male_full / total_enrollment,
    hispanic_female = hispanic_female_full / total_enrollment,
    american_male = american_indian_male_full / total_enrollment,
    american_female = american_indian_female_full / total_enrollment,
    asian_male = asian_male_full / total_enrollment,
    asian_female = asian_female_full / total_enrollment,
    pacific_male = pacific_islander_male_full / total_enrollment,
    pacific_female = pacific_islander_female_full / total_enrollment,
    black_male = black_male_full / total_enrollment,
    black_female = black_female_full / total_enrollment,
    white_male = white_male_full / total_enrollment,
    white_female = white_female_full / total_enrollment,
    two_male = two_or_more_male_full / total_enrollment,
    two_female = two_or_more_female_full / total_enrollment,
  ) %>%
  select(ends_with('male'), ends_with('female'))

# Transform data for plotting
total_enrollment_prop %>%
  gather(key = "race_gender", value = "proportion") %>%
  separate(race_gender,
           into = c("race", "gender"),
           sep = "_") %>%
  ggplot(aes(x = gender, y = proportion)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(x = "Gender", y = "Proportion",
       title = "Distribution of AP Students by Race and Gender Across All Schools") +
  scale_fill_brewer(palette = "Set1") +
  theme(plot.title = element_text(hjust = 0.5)) +
  facet_grid(. ~ race)
```



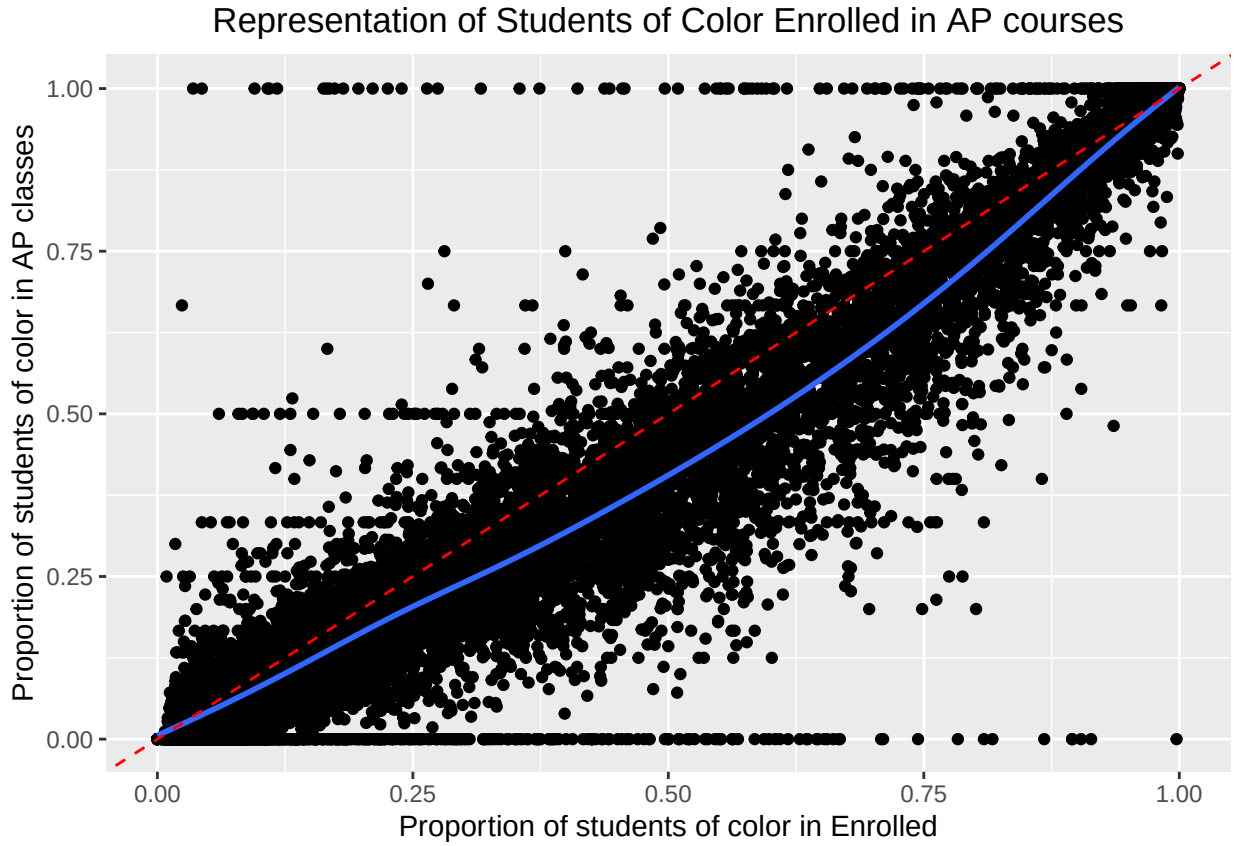
The plot is concerning because it highlights that the White race is disproportionately over represented in enrollments compared to other races. Another noteworthy finding is that for AP courses, females are either enrolled more or equally as males, regardless of race. Enrollments from American Indians and Pacific Islanders are particularly low compared to other races. There is a substantial decline in enrollments for AP courses among all other races except Whites (who, once again, have more than twice the enrollment rate of the second-highest enrolled race in AP courses).

Problem 3

```
# Load the packages
library(tidyr)
library(dplyr)
library(ggplot2)

# Load data from csv file
ap_data <- read.csv("Advanced Placement.csv")

# Replace negative values with NA
neg_indices <- which(ap_data < 0, arr.ind = TRUE)
ap_data[neg_indices] <- NA

# Filter out the schools where AP courses are conducted
ap_data <- ap_data[!is.na(ap_data$COMBOKEY) & (ap_data$SCH_APENR_IND == "Yes"),]

# Select columns of students of color from the dataset
```

```r
ap_data <- ap_data %>%
  select(COMBOKEY, TOT_APENR_M, TOT_APENR_F, SCH_APENR_HI_M, SCH_APENR_HI_F,
         SCH_APENR_AM_M, SCH_APENR_AM_F, SCH_APENR_AS_M, SCH_APENR_AS_F,
         SCH_APENR_HP_M, SCH_APENR_HP_F, SCH_APENR_BL_M, SCH_APENR_BL_F,
         SCH_APENR_TR_M, SCH_APENR_TR_F)

# Group the data by "COMBOKEY" and calculate the total number of students,
# total number of students of color, and proportion of students of color in AP classes
ap_data <- ap_data %>%
  group_by(COMBOKEY) %>%
  summarise(total_students_ap = sum(TOT_APENR_M, TOT_APENR_F, na.rm = TRUE),
            total_students_of_color_ap = sum(SCH_APENR_HI_M, SCH_APENR_HI_F,
                                             SCH_APENR_AM_M, SCH_APENR_AM_F,
                                             SCH_APENR_AS_M, SCH_APENR_AS_F,
                                             SCH_APENR_HP_M, SCH_APENR_HP_F,
                                             SCH_APENR_BL_M, SCH_APENR_BL_F,
                                             SCH_APENR_TR_M, SCH_APENR_TR_F,
                                             na.rm = TRUE),
            prop_students_of_color_ap = total_students_of_color_ap/total_students_ap)

# The code removes any rows with missing values from the dataset
ap_data <- na.omit(ap_data)

# Load data from csv file
enrollment <- read.csv("Enrollment.csv")

# Replace negative values with NA
neg_indices <- which(enrollment < 0, arr.ind = TRUE)
enrollment[neg_indices] <- NA

# Filter out the schools where AP courses are conducted
enrollment <- enrollment[enrollment$COMBOKEY %in% ap_data$COMBOKEY, ]

# Select columns of students of color from the dataset
enrollment <- enrollment %>%
  select(COMBOKEY, TOT_ENR_M, TOT_ENR_F, SCH_ENR_HI_M, SCH_ENR_HI_F, SCH_ENR_AM_M,
         SCH_ENR_AM_F, SCH_ENR_AS_M, SCH_ENR_AS_F, SCH_ENR_HP_M, SCH_ENR_HP_F,
         SCH_ENR_BL_M, SCH_ENR_BL_F, SCH_ENR_TR_M, SCH_ENR_TR_F)

# Group the data by "COMBOKEY" and calculates the total number of students,
# total number of students of color, and proportion of students of color in
# enrolled classes
enrollment <- enrollment %>%
  group_by(COMBOKEY) %>%
  summarise(total_students = sum(TOT_ENR_M, TOT_ENR_F, na.rm = TRUE),
            total_students_of_color = sum(SCH_ENR_HI_M, SCH_ENR_HI_F,
                                          SCH_ENR_AM_M, SCH_ENR_AM_F,
                                          SCH_ENR_AS_M, SCH_ENR_AS_F,
                                          SCH_ENR_HP_M, SCH_ENR_HP_F,
                                          SCH_ENR_BL_M, SCH_ENR_BL_F,
                                          SCH_ENR_TR_M, SCH_ENR_TR_F, na.rm = TRUE),
            prop_students_of_color = total_students_of_color/total_students)
```

```r
# Filter out the COMBOKEY and Proportion of students of color enrolled overall
enrollment_prop_students_of_color <- enrollment %>%
  select(COMBOKEY, prop_students_of_color)

# Filter out the COMBOKEY and Proportion of students of color enrolled in AP courses
ap_data_prop_students_of_color_ap <- ap_data %>%
  select(COMBOKEY, prop_students_of_color_ap)

# Merge the data on COMBOKEY
merged_data <- merge(enrollment_prop_students_of_color,
                     ap_data_prop_students_of_color_ap, by = "COMBOKEY")

# Plot the distribution graph
ggplot(merged_data, aes(x = prop_students_of_color, y = prop_students_of_color_ap)) +
  geom_point() +
  geom_smooth(se = FALSE, method = 'gam', formula = y ~ s(x, bs = "cs")) +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color = "red") +
  labs(x = "Proportion of students of color in Enrolled",
       y = "Proportion of students of color in AP classes",
       title = "Representation of Students of Color Enrolled in AP courses") +
  theme(plot.title = element_text(hjust = 0.5))
```



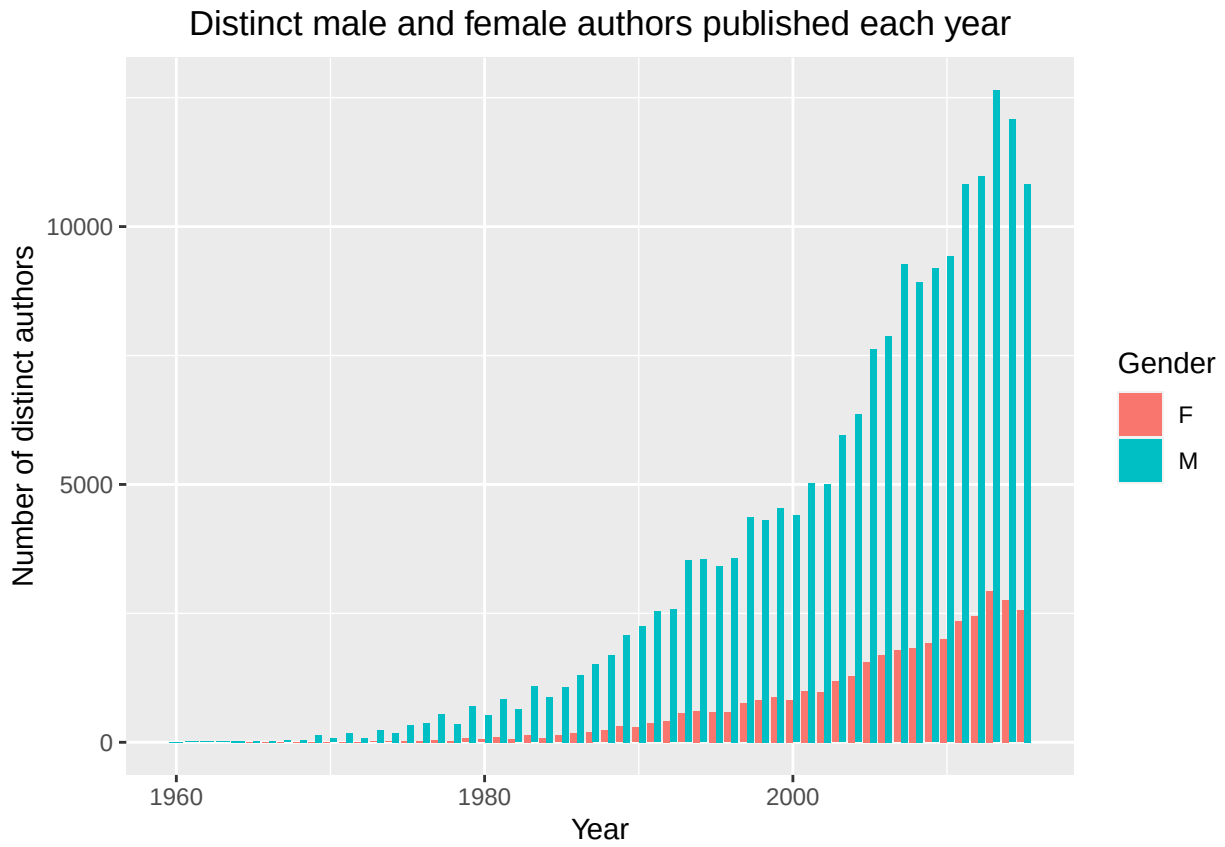
The plot represents the proportion of students of color in Advanced Placement (AP) courses compared to the enrolled population. It shows a positive correlation between the two, but many schools have a lower proportion of students of color in AP courses, indicating under-representation. The smoothing line further highlights this trend, suggesting a need to address the under-representation and promote equitable access to

advanced coursework. Some points are above or below the red line, indicating variation in the representation of students of color in AP courses across schools. The few points lined along the upper and lower limits of the x-axis may indicate significantly higher or lower proportions of students of color in enrolled classes than in AP courses. It is essential to address these disparities and promote equal access and opportunities for all students.

Part B

Problem 4

```r
# Load required packages
library(RSQLite)
library(tidyr)
library(dplyr)
library(ggplot2)

# Connect to the SQLite database file
db <- dbConnect(RSQLite::SQLite(), "dblp.db")

# Filter out non-male and non-female authors with prediction probability less than 0.9
query <- "SELECT * FROM general JOIN authors ON general.k = authors.k"
df <- dbGetQuery(db, query)
df <- df[df$gender %in% c("M", "F") & df$prob >= 0.9,]

# Remove the "k" column as its duplicate and not needed
df <- select(df, -k)

# Disconnect from the database
dbDisconnect(db)

# Aggregate the number of distinct authors by year and gender
author_counts <- aggregate(name ~ year + gender, data = df, FUN = function(x)
  length(unique(x)))

# Create a bar plot of the number of distinct male and female authors published each year
ggplot(author_counts, aes(x = year, y = name, fill = gender)) +
  geom_col(position = "dodge") +
  xlab("Year") +
  ylab("Number of distinct authors") +
  ggtitle("Distinct male and female authors published each year") +
  labs(fill = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```

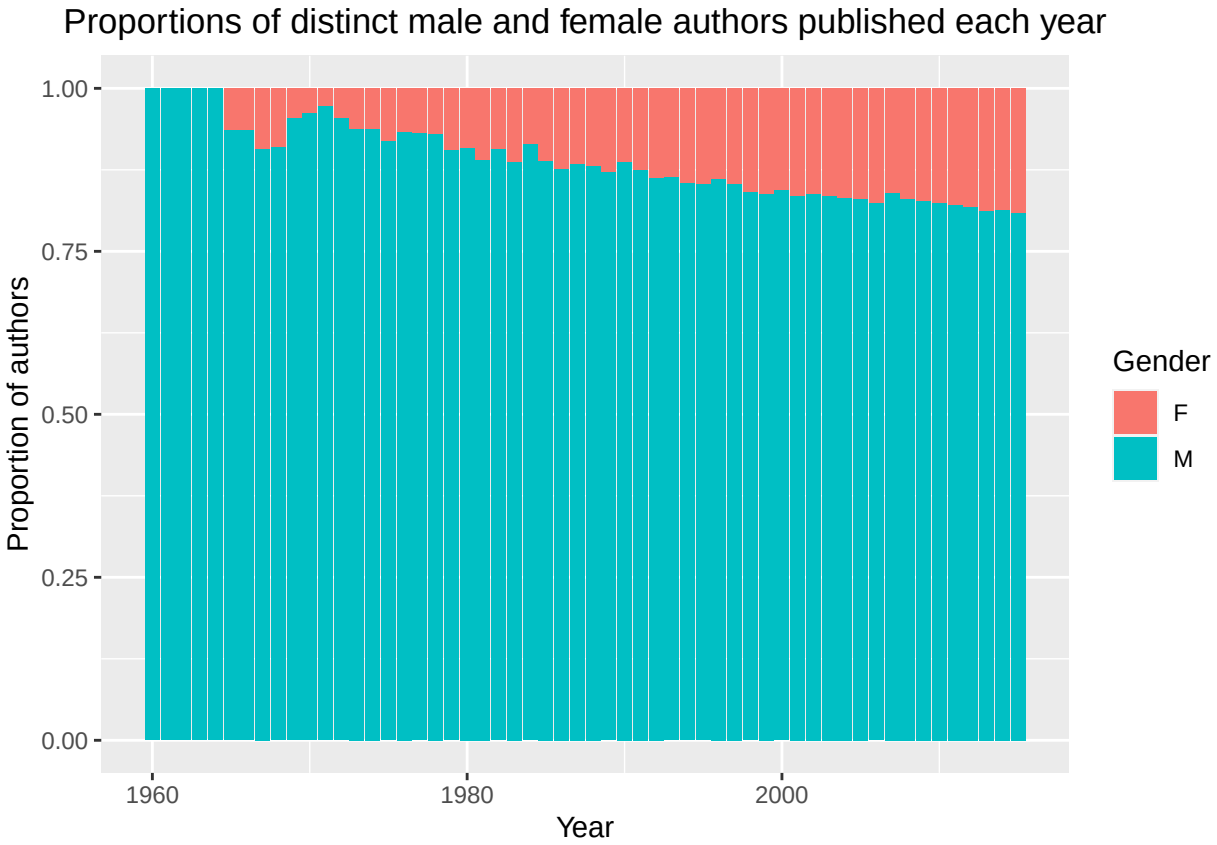


The visualization depicts the trend of distinct male and female authors published each year, indicating a steady increase in the number of authors over time. However, it also highlights the widening gap between male and female authors, with the number of male authors always being higher than the number of female authors. This suggests the existence of gender bias or barriers that prevent women from publishing at the same rate as men and highlights the need for addressing this issue to ensure equal opportunities for male and female researchers.

Problem 5

```
# Load required packages
library(RSQLite)
library(tidyr)
library(dplyr)
library(ggplot2)

# Connect to the database
db <- dbConnect(RSQLite::SQLite(), "dblp.db")

# Create a query to join the "general" and "authors" tables
query <- "SELECT * FROM general JOIN authors ON general.k = authors.k"

# Retrieve the query result into a dataframe
df <- dbGetQuery(db, query)

# Filter out non-male and non-female authors with prediction probability less than 0.9
```

```r
df <- df[df$gender %in% c("M", "F") & df$prob >= 0.9,]

# Remove the "k" column
df <- select(df, -k)

# Disconnect from the database
dbDisconnect(db)

# Aggregate the number of distinct authors by year and gender
author_counts <- aggregate(name ~ year + gender, data = df, FUN = function(x)
  length(unique(x)))

# Group the author_counts by year and gender, calculate the proportion of
# authors and select only relevant columns
author_props <- author_counts %>%
  group_by(year) %>%
  mutate(prop = name/sum(name)) %>%
  select(year, gender, prop)

# Plot the proportion of distinct male and female authors published each year
ggplot(author_props, aes(x = year, y = prop, fill = gender)) +
  geom_col(position = "stack") +
  xlab("Year") +
  ylab("Proportion of authors") +
  ggtitle("Proportions of distinct male and female authors published each year") +
  labs(fill = "Gender") +
  theme(plot.title = element_text(hjust = 0.5))
```



The visualization shows the proportions of distinct male and female authors published each year. The graph indicates that the proportion of female authors has been increasing over time, but the proportion of male authors is still higher. The graph shows that the gap between male and female authorship is slowly decreasing over time, indicating that efforts to promote gender equality in authorship may be having a positive impact. The visualization highlights an important issue of gender inequality in authorship and suggests that continued efforts to promote gender diversity in academic publishing are needed.