HW2 Part B

Omair Shafi Ahmed

10/10/2017

Part B

Importing Libraries

```
library('ggplot2')
library('tidyverse')
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library('dplyr')
```

Importing Data

```
CRDC = read_csv('./crdc201314csv/CRDC2013_14_SCH.csv' ,na = c("-2", "-5", "-9"))
```

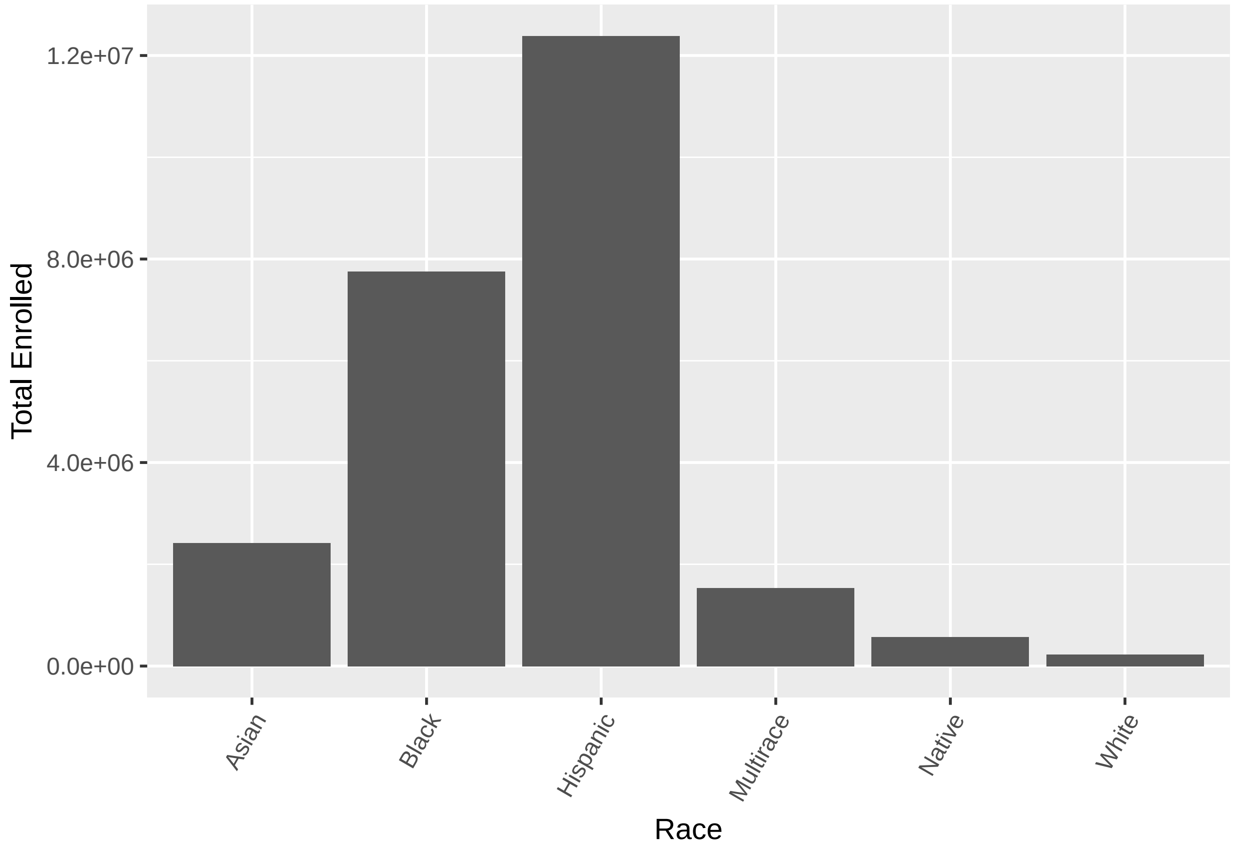
```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   LEA_STATE = col_character(),
##   LEA_NAME = col_character(),
##   SCH_NAME = col_character(),
##   COMBOKEY = col_character(),
##   LEAID = col_character(),
##   SCHID = col_character(),
##   JJ = col_character(),
##   CCD_LATCOD = col_double(),
##   CCD_LONCOD = col_double(),
##   NCES_SCHOOL_ID = col_character(),
##   MATCH_FLAG = 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()
##   # ... with 75 more columns
## )
```

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

Selecting

```
CRDC_3        <- select(CRDC, starts_with('TOT_ENR_'),
                        starts_with('SCH_ENR_BL_'),
                        starts_with('SCH_ENR_HI_'),
                        starts_with('SCH_ENR_AS_'),
                        starts_with('SCH_ENR_AM_'),
                        starts_with('SCH_ENR_TR_'),
                        starts_with('SCH_ENR_HP_'))

#CRDC_3
CRDC_3 <- mutate(CRDC_3, `Total Enrolled` =  `TOT_ENR_M` + `TOT_ENR_F`,
                `Black` = `SCH_ENR_BL_M` + `SCH_ENR_BL_F`,
                `Hispanic` = `SCH_ENR_HI_M` + `SCH_ENR_HI_F`,
                `Asian` = `SCH_ENR_AS_M` + `SCH_ENR_AS_F`,
                `Native` = `SCH_ENR_AM_M` + `SCH_ENR_AM_F`,
                `Multirace` = `SCH_ENR_TR_M` + `SCH_ENR_TR_F`,
                `White` = `SCH_ENR_HP_M` + `SCH_ENR_HP_F`,)

CRDC_3 <- CRDC_3[c('Black', 'Hispanic',
                   'Asian', 'Native',
                   'Multirace', 'White')]

CRDC_3 <- gather(CRDC_3, `Black`, `Hispanic`,
                   `Asian`, `Native`,
                   `Multirace`, `White`, key='Race', value='Total Enrolled')


#CRDC_3
ggplot(CRDC_3) + geom_col(aes(x=`Race`, y=`Total Enrolled`)) +
                theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

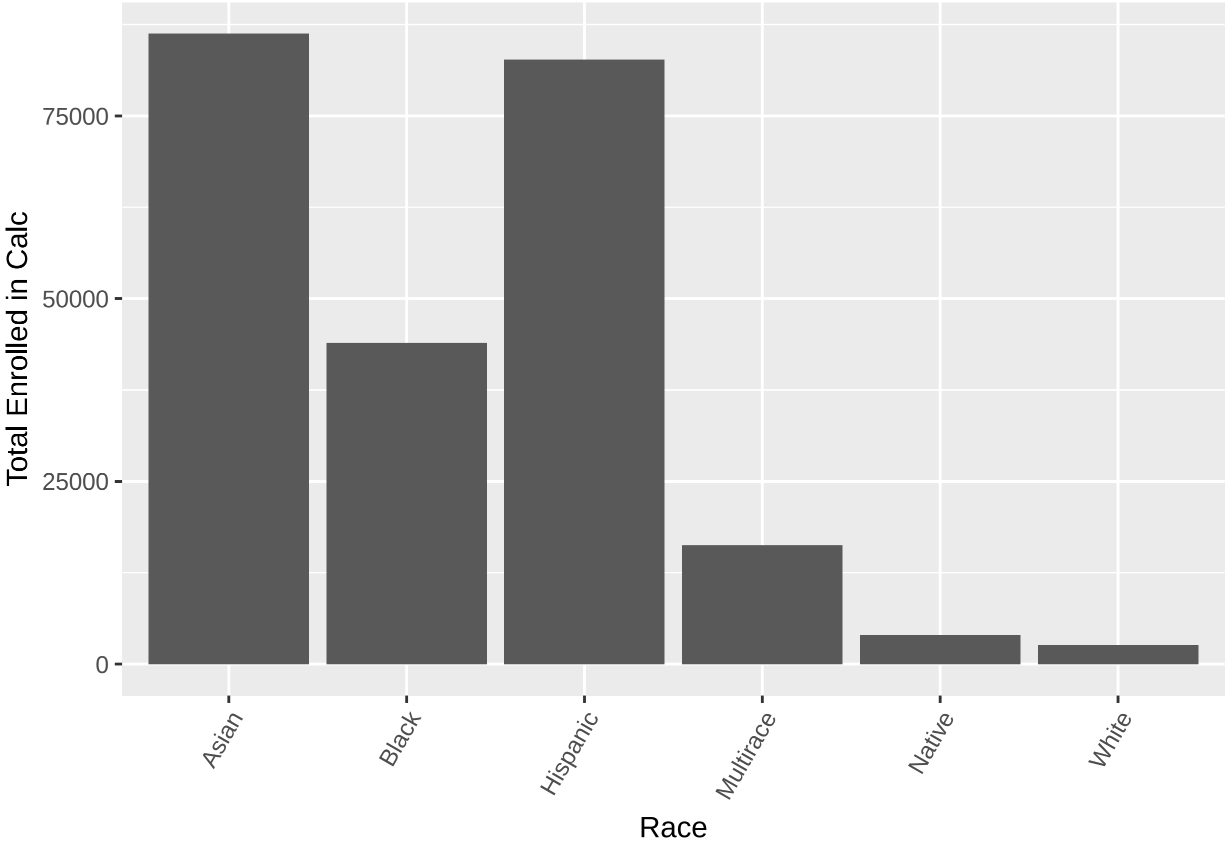
```
## Warning: Removed 52 rows containing missing values (position_stack).
```



```r
CRDC_3        <- select(CRDC, starts_with('TOT_ENR_'),
                        starts_with('SCH_MATHENR_CALC_BL_'),
                        starts_with('SCH_MATHENR_CALC_HI_'),
                        starts_with('SCH_MATHENR_CALC_AS_'),
                        starts_with('SCH_MATHENR_CALC_AM_'),
                        starts_with('SCH_MATHENR_CALC_TR_'),
                        starts_with('SCH_MATHENR_CALC_HP_'))

#CRDC_3
CRDC_3 <- mutate(CRDC_3, `Enrolled` =  `TOT_ENR_M` + `TOT_ENR_F`,
                `Black` = `SCH_MATHENR_CALC_BL_M` + `SCH_MATHENR_CALC_BL_F`,
                `Hispanic` = `SCH_MATHENR_CALC_HI_M` + `SCH_MATHENR_CALC_HI_F`,
                `Asian` = `SCH_MATHENR_CALC_AS_M` + `SCH_MATHENR_CALC_AS_F`,
                `Native` = `SCH_MATHENR_CALC_AM_M` + `SCH_MATHENR_CALC_AM_F`,
                `Multirace` = `SCH_MATHENR_CALC_TR_M` + `SCH_MATHENR_CALC_TR_F`,
                `White` = `SCH_MATHENR_CALC_HP_M` + `SCH_MATHENR_CALC_HP_F`,)

CRDC_3 <- CRDC_3[c('Black', 'Hispanic',
                   'Asian', 'Native',
                   'Multirace', 'White')]

CRDC_3 <- gather(CRDC_3, `Black`, `Hispanic`,
                   `Asian`, `Native`,
                   `Multirace`, `White`, key='Race', value='Total Enrolled in Calc')


#CRDC_3
```

```
ggplot(CRDC_3) + geom_col(aes(x=`Race`, y=`Total Enrolled in Calc`)) +
                theme(axis.text.x = element_text(angle = 60, hjust = 1))
```

```
## Warning: Removed 496723 rows containing missing values (position_stack).
```



It appears as though Asians are overrepresented in the Calc class.

Part C

Importing Data

```
if(!require('ggplot2')) install.packages("ggplot2",repos = "http://cran.us.r-project.org")
if(!require('dplyr')) install.packages("dplyr",repos = "http://cran.us.r-project.org")
if(!require('RMySQL')) install.packages("RMySQL",repos = "http://cran.us.r-project.org")
```

```
## Loading required package: RMySQL

## Loading required package: DBI
```

```
if(!require('dbplyr')) install.packages("dbplyr",repos = "http://cran.us.r-project.org")
```

```
## Loading required package: dbplyr

##
## Attaching package: 'dbplyr'

## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
if(!require('rstudioapi')) install.packages("rstudioapi",repos = "http://cran.us.r-project.org")
```

```
## Loading required package: rstudioapi
```

```
library('ggplot2')
library('dplyr')
library ('RMySQL')
library('dbplyr')
library('rstudioapi')
```

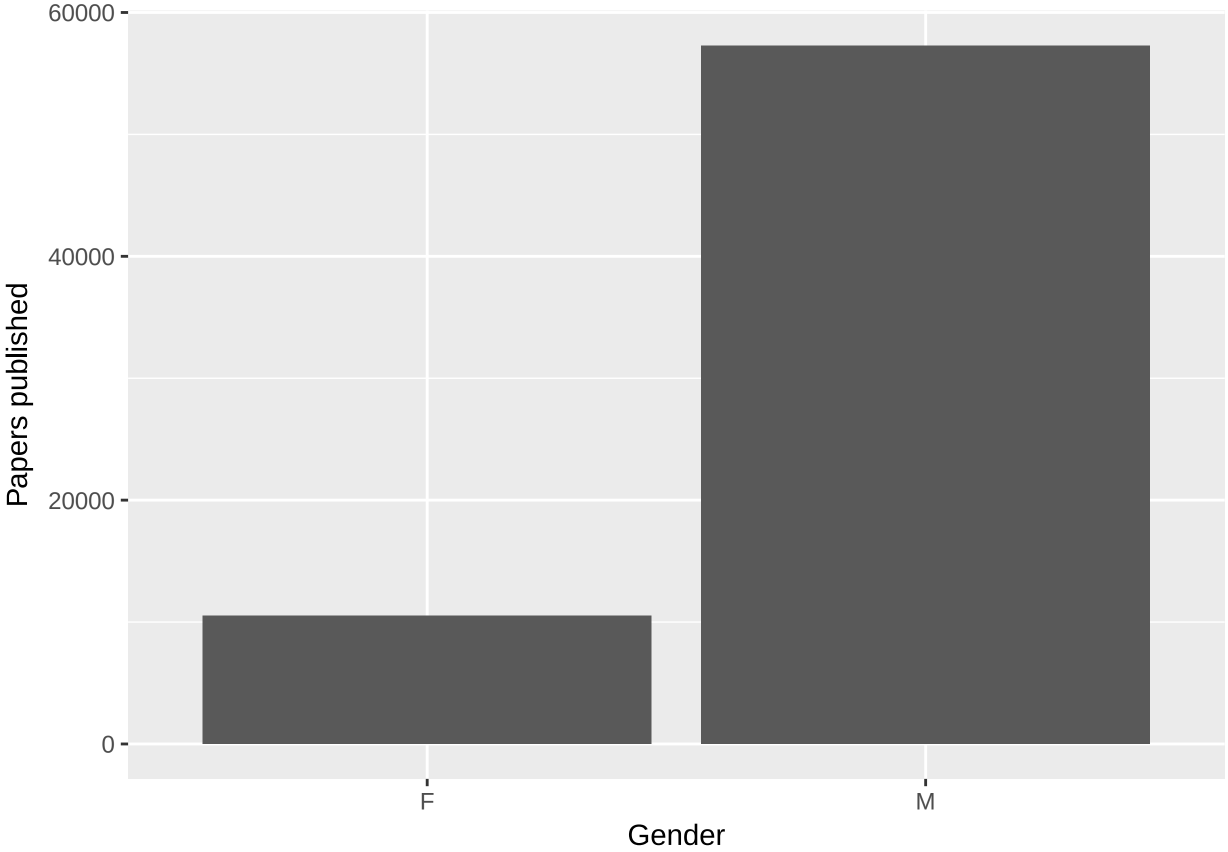
Connecting to DB

```
con <- DBI::dbConnect(RMySQL::MySQL(),
  host = "localhost",
  socket = "/Applications/MAMP/tmp/mysql/mysql.sock",
  port= 8889,
  user = "root",
  #password = rstudioapi::askForPassword("Database password")
  password = "root",
  db = "DBLP"
)

authors <- tbl(con, "authors")
general <- tbl(con, "general")
```

Problem 5

Filter the data to include only the authors for whom a gender was predicted with a probability of 0.95 or greater, and then create a bar plot showing the number of distinct male and female authors in the dataset.

```
filtered_authors <- authors %>% filter(prob > 0.95, prob < 1) %>%
                    collect()

filtered_authors %>%
  group_by(gender) %>%
  count() %>%
  ggplot() +
    geom_col(mapping = aes(x= gender, y= n)) +
    xlab("Gender") +
    ylab("Papers published")
```



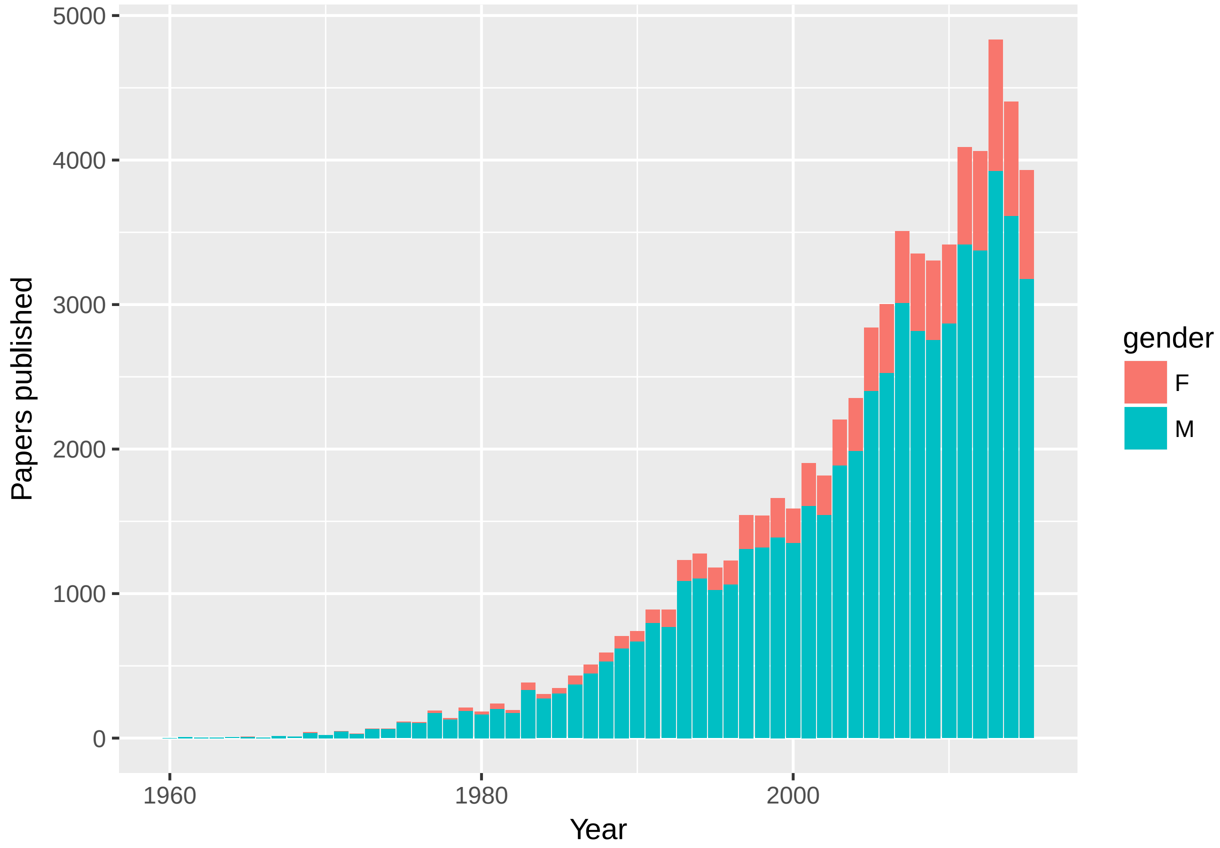
It appears as if males authors generally dominate in numbers of papers published.

Problem 6

Again including only the authors for whom a gender was predicted with a probability of 0.95 or greater, create a stacked bar plot showing the number of distinct male and female authors published each year.

```
filtered_authors %>%
  inner_join(general, key = 'k', copy=TRUE) %>%
  select(gender, year) %>%
  group_by(gender, year) %>%
  count() %>%
  ggplot() +
    geom_col(mapping = aes(x = year, y = n, fill = gender)) +
    xlab("Year") +
    ylab("Papers published")
```

```
## Joining, by = "k"
```

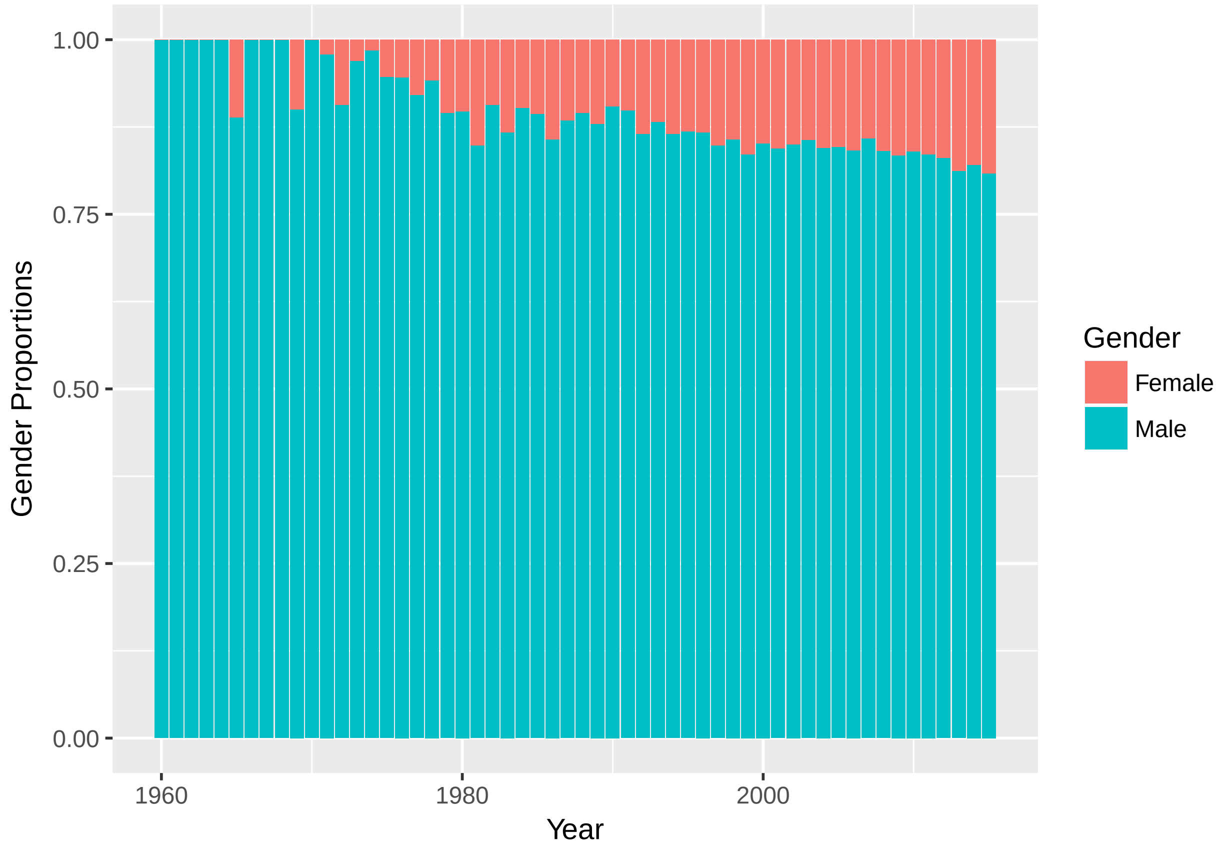


The proportion of female scientists publishing papers have increased in recent years

Problem 7

```
filtered_authors %>%
  inner_join(general, key = 'k', copy = TRUE) %>%
  select(gender, year) %>%
  group_by(year) %>%
  summarise(`Male` = mean(gender == "M"),
            `Female` = mean(gender == "F")) %>%
  gather(-year, key = "Gender", value = "val") %>%
  ggplot() +
    geom_col(mapping = aes(x = year, y = val, fill = Gender)) +
    xlab("Year") +
    ylab("Gender Proportions")
```

```
## Joining, by = "k"
```



Here, the increase in the proportion of female authors publishing papers is clearly visible.

Part D

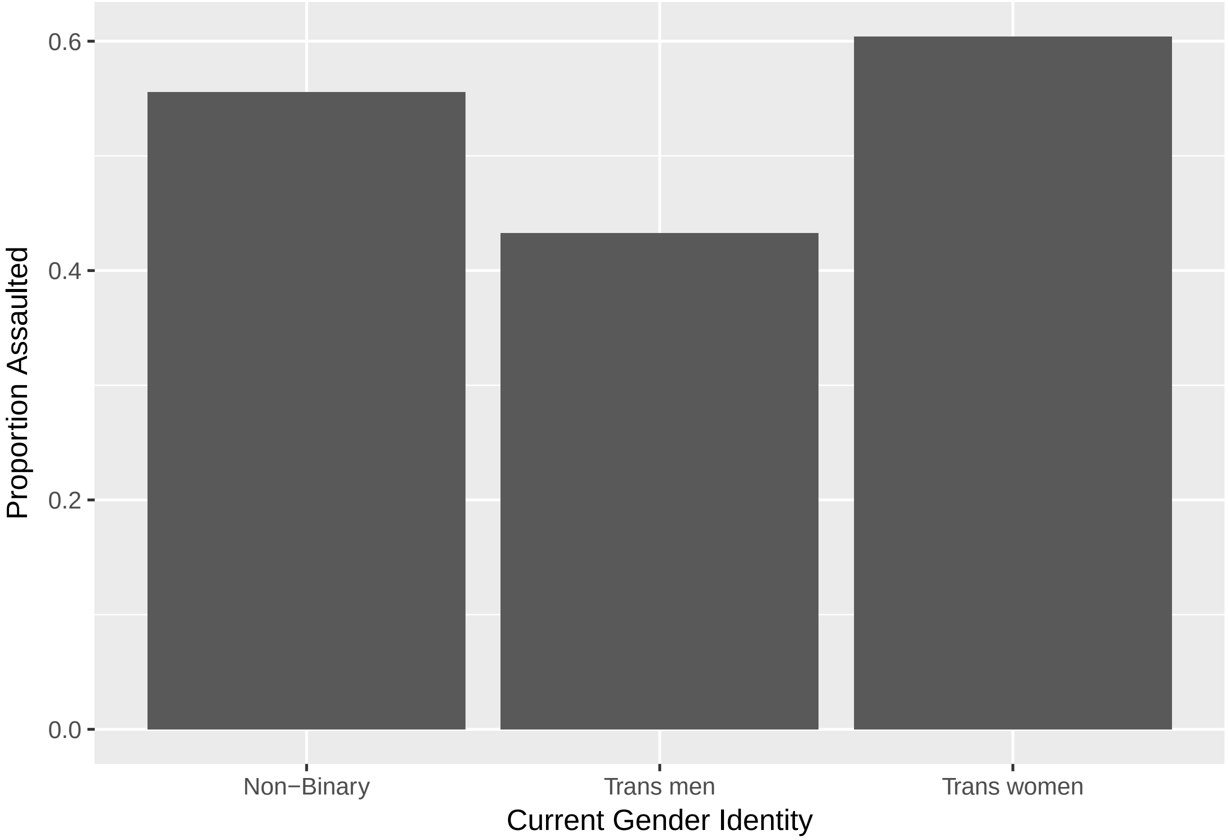
Importing the data

```
load("/Users/omairs/Documents/Masters/DS 5110/HW2/ICPSR_31721/DS0001/31721-0001-Data.rda")
ICPSR <- as.tibble(da31721.0001)
```

Problem 8

Recode gender to create a category for "Non-binary" identities which includes the "Androgynous" and "Gender Queer" categories. Then create bar plots showing the proportions of participants who have been a victim of a violent assault since age 13 (either a sexual assault or another type of physical assault) for trans men, for trans women, and for non-binary people.

```
ICPSR %>%
        transmute(`Current Gender Identity` = recode(Q6,
                                              `(4) Androgynous` = "Androgynous",
                                              `(6) Gender Queer` = "Gender Queer",
                                              `(3) Transgender` = "Transgender",
                                              `(1) Man` = "Man",
                                              `(2) Woman` = "Woman",
                                              .default = NA_character_),
                  `Birth Sex` = recode(Q5,
                             `(1) Male` = "Trans men",
                             `(2) Female` = "Trans women"),

                  `Sexually Assaulted` = recode(Q96,
                               `(1) Yes` = "Yes",
                               `(2) No` = "No"),

                  `Physically Assaulted` = recode(Q106,
                                 `(1) Yes` = "Yes",
                                 `(2) No` = "No"),

                  `Assault` = if_else(`Sexually Assaulted` == "Yes" |
                                      `Physically Assaulted` == "Yes", "Yes", "No")) %>%

  mutate(`Gender` = if_else(`Current Gender Identity` %in% c("Androgynous", "Gender Queer"),
                        "Non-Binary", as.character(`Birth Sex`))) %>%
  select(`Gender`, `Assault`) %>%
  filter(!is.na(`Assault`)) %>%
  group_by(`Gender`) %>%
  summarise(`Proportion Assaulted` = mean(`Assault` == "Yes")) %>%
  ggplot() +
    geom_col(mapping = aes(x = `Gender`, y = `Proportion Assaulted`)) +
    xlab("Current Gender Identity") +
    ylab("Proportion Assaulted")
```

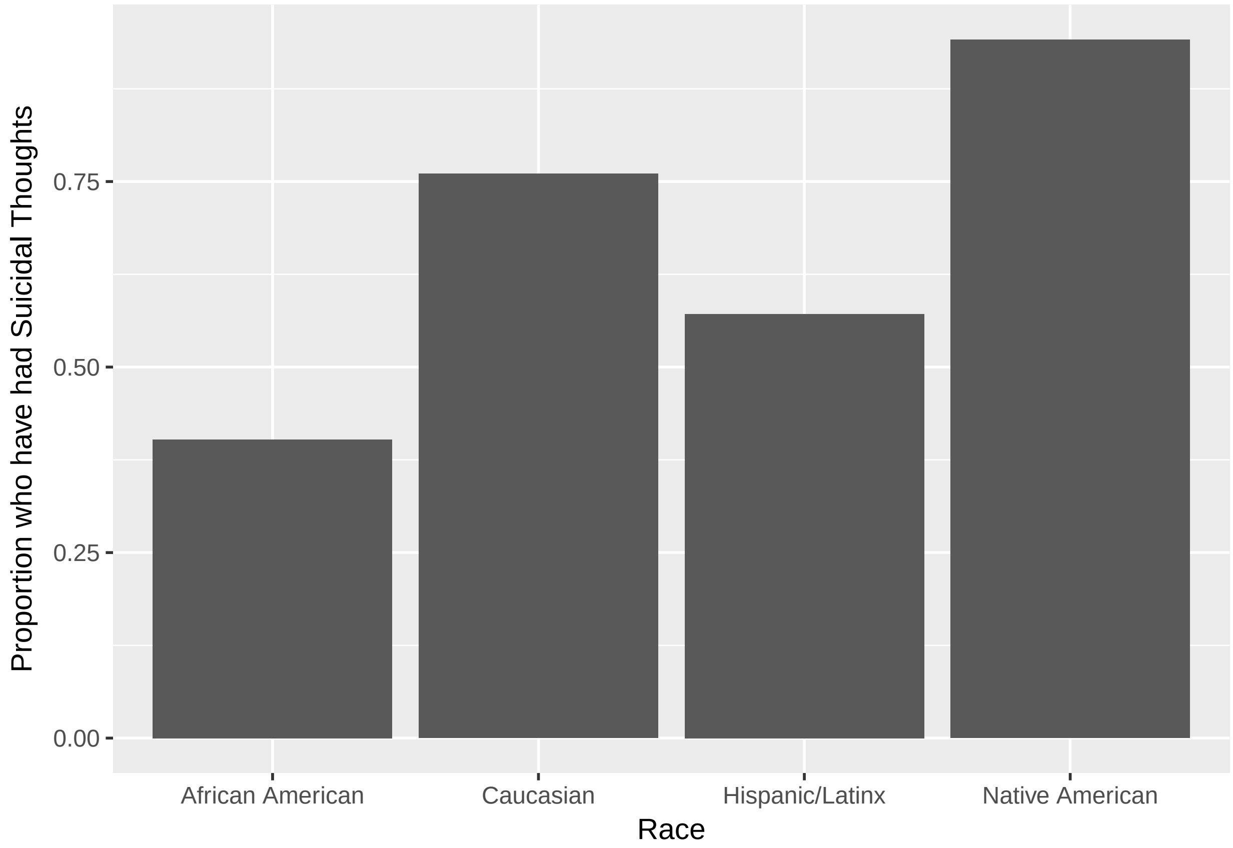


There are over 10% more Trans women and Non-binaries being assualted than Trans men.

Problem 9

Create bar plots showing the proportions of participants who have thought about killing themselves for African American, Caucasian, Hispanic/Latinx, and Native American demographics. (Do not include participants who declined to answer.)

```
`Suicidal proportion` <-
                        ICPSR %>%
                          transmute(`African American` = D9_1,
                                      `Caucasian` = D9_2,
                                      `Hispanic/Latinx` = D9_3,
                                      `Native American` = D9_4,
                                      KillingSelf = Q131) %>%
                                      mutate_all(funs(recode),
                                      `(0) Not selected` = "No",
                                      `(1) Selected` = "Yes",
                                      `(1) Yes` = "Yes",
                                      `(2) No` = "No") %>%
                            filter(!is.na(KillingSelf)) %>%
                            gather(key = "race", value = "is_race", -KillingSelf) %>%
                            filter(is_race == "Yes") %>%
                            select(-is_race)

`Suicidal proportion` %>%
                            group_by(race) %>%
                            summarise(Prop = mean(KillingSelf == "Yes")) %>%
                            ggplot() +
                            geom_col(mapping = aes(x = race, y = Prop)) +
                            xlab("Race") +
                            ylab("Proportion who have had Suicidal Thoughts")
```



The proportion of Native Americans who have had suicidal thoughts are over 12.5% higher than the next highest demographic, ie. Caucasian. The proportion of Caucasians who have had suicidal thoughts are in turn over 12.5% higher than the next highest demohgraphic, ie. Hispanic/Latinx. This shows a significant amount of variance between races, for participans who have had suicidal thoughts. Could this be due differences in to the acceptance (or lack of) of non-binary, traditional sexual identities between communities?

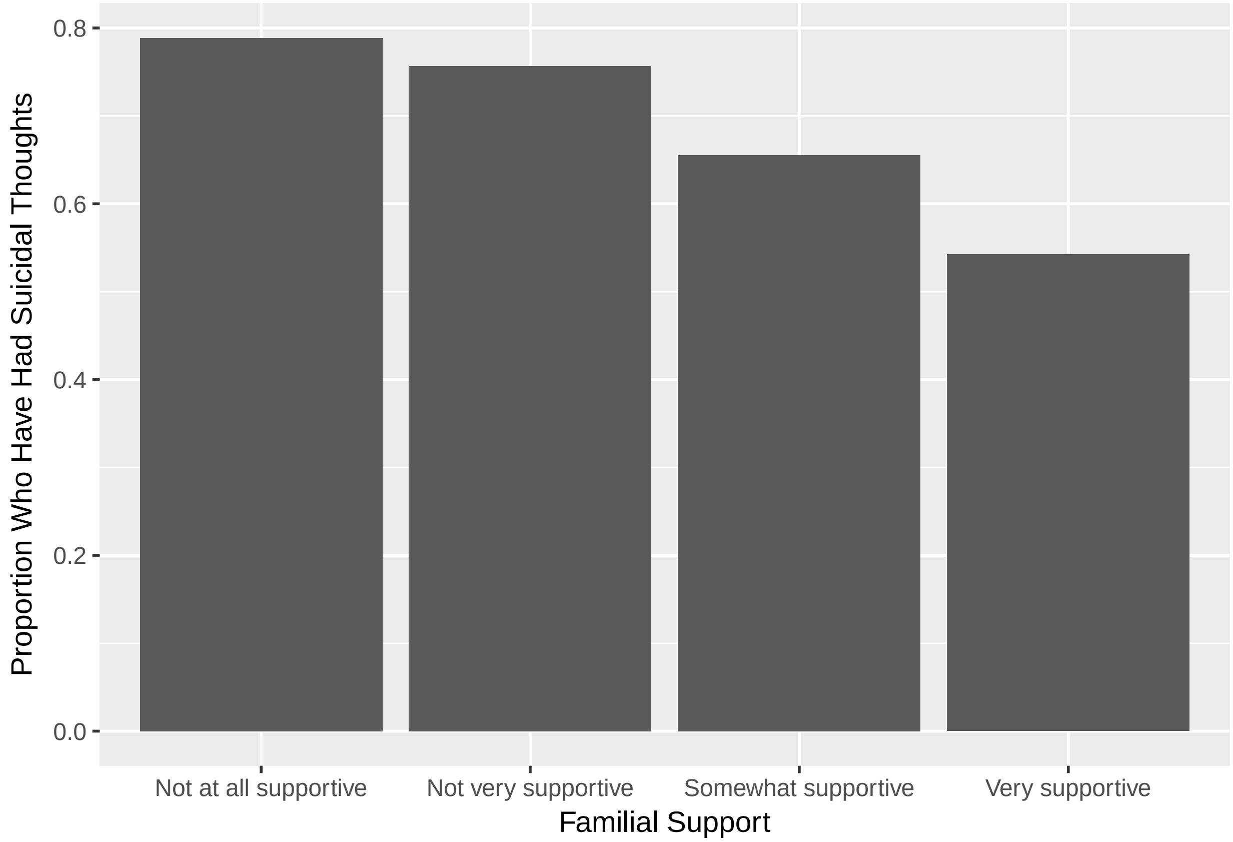
```
`Suicidal proportion` %>%
  summarise(Prop = mean(KillingSelf == "Yes"))
```

```
## # A tibble: 1 x 1
##        Prop
##       <dbl>
## 1 0.6713092
```

The number of participants with suicidal thoughts are alarmingly high. For the participants of this survey it's 67.13%, which is higher than the national avergae 41%, which in turn is higher than the general population, 1.6%. These are staggeringly high figures.

Problem 10

```
ICPSR %>%
        transmute(Support = recode(Q119,
                                   `(1) Not at all supportive` = "Not at all supportive",
                                   `(2) Not very supportive` = "Not very supportive",
                                   `(3) Somewhat supportive` = "Somewhat supportive",
                                   `(4) Very supportive` = "Very supportive",
                                     .default = NA_character_),
                  KillingSelf = recode(Q131,
                                        `(1) Yes` = "Yes",
                                        `(2) No` = "No")) %>%
          filter(!is.na(Support) & !is.na(KillingSelf)) %>%
          group_by(Support) %>%
          summarise(Proportion = mean(KillingSelf == "Yes")) %>%
          ggplot() +
            geom_col(mapping = aes(x = Support, y = Proportion)) +
            xlab("Familial Support") +
            ylab("Proportion Who Have Had Suicidal Thoughts")
```



Here, the charts show a clear linear relationship between families being supportive of the offsprings identity and the proportion of offsprings who have had suicidal thoughts.