datasetxxx

```
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=25),tidy=TRUE)
```

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

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

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

```
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

```
hire <- read_csv("/Users/yzh/Downloads/h1b_kaggle.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   CASE_STATUS = col_character(),
##   EMPLOYER_NAME = col_character(),
##   SOC_NAME = col_character(),
##   JOB_TITLE = col_character(),
##   FULL_TIME_POSITION = col_character(),
##   PREVAILING_WAGE = col_double(),
##   YEAR = col_integer(),
##   WORKSITE = col_character(),
##   lon = col_double(),
##   lat = col_double()
## )
```

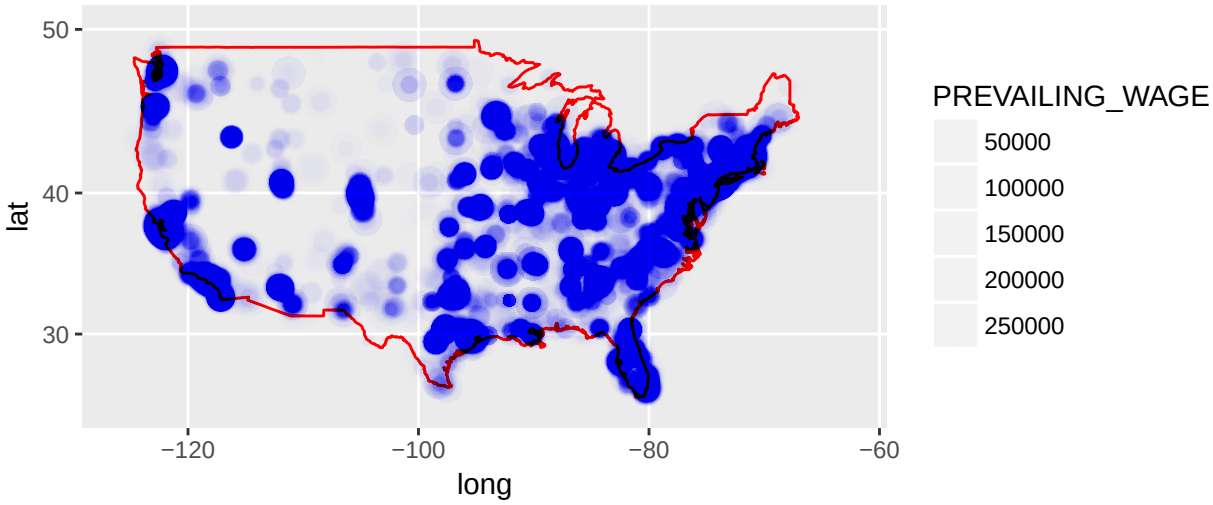
```
hire[1:10, ]
```

```
## # A tibble: 10 x 11
##       X1         CASE_STATUS
##    <int>               <chr>
##  1     1 CERTIFIED-WITHDRAWN
##  2     2 CERTIFIED-WITHDRAWN
##  3     3 CERTIFIED-WITHDRAWN
##  4     4 CERTIFIED-WITHDRAWN
##  5     5           WITHDRAWN
```

```
##  6     6 CERTIFIED-WITHDRAWN
##  7     7 CERTIFIED-WITHDRAWN
##  8     8 CERTIFIED-WITHDRAWN
##  9     9 CERTIFIED-WITHDRAWN
## 10    10           WITHDRAWN
## # ... with 9 more variables: EMPLOYER_NAME <chr>, SOC_NAME <chr>,
## #   JOB_TITLE <chr>, FULL_TIME_POSITION <chr>, PREVAILING_WAGE <dbl>,
## #   YEAR <int>, WORKSITE <chr>, lon <dbl>, lat <dbl>
```

```
hire2016certi <- filter(hire,
    CASE_STATUS == "CERTIFIED",
    YEAR == 2016, PREVAILING_WAGE <=
        250000, 40000 <= PREVAILING_WAGE)
us <- map_data("usa")
ggplot(us) + geom_polygon(aes(x = long,
    y = lat, group = group),
    fill = NA, color = "red") +
    geom_point(data = hire2016certi,
        aes(x = lon, y = lat,
            size = PREVAILING_WAGE),
        color = "blue", alpha = 0.002,
        position = "jitter") +
    coord_map() + xlim(c(-126,
    -62.5)) + ylim(c(24, 50))
```

```
## Warning: Removed 14917 rows containing missing values (geom_point).
```



Prob 3

Create a bar plot showing the total number of enrolled students of each race. missing values are represented in CSV file as: -2, -5 -9

Data from: SCH_ENR_HI_M,62,1,Overall Student Enrollment: Hispanic Male,I,7 SCH_ENR_HI_F,63,1,Overall Student Enrollment: Hispanic Female,I,7 SCH_ENR_AM_M,64,1,Overall Student Enrollment: American Indian/Alaska Native Male,I,7 SCH_ENR_AM_F,65,1,Overall Student Enrollment: American Indian/Alaska Native Female,I,7 SCH_ENR_AS_M,66,1,Overall Student Enrollment: Asian Male,I,7

SCH_ENR_AS_F,67,1,Overall Student Enrollment: Asian Female,I,7 SCH_ENR_HP_M,68,1,Overall Student Enrollment: Native Hawaiian/Pacific Islander Male,I,7 SCH_ENR_HP_F,69,1,Overall Student Enrollment: Native Hawaiian/Pacific Islander Female,I,7 SCH_ENR_BL_M,70,1,Overall Student Enrollment: Black Male,I,7 SCH_ENR_BL_F,71,1,Overall Student Enrollment: Black Female,I,7 SCH_ENR_WH_M,72,1,Overall Student Enrollment: White Male,I,7 SCH_ENR_WH_F,73,1,Overall Student Enrollment: White Female,I,7 SCH_ENR_TR_M,74,1,Overall Student Enrollment: Two or More Races Male,I,7 SCH_ENR_TR_F,75,1,Overall Student Enrollment: Two or More Races Female,I,7

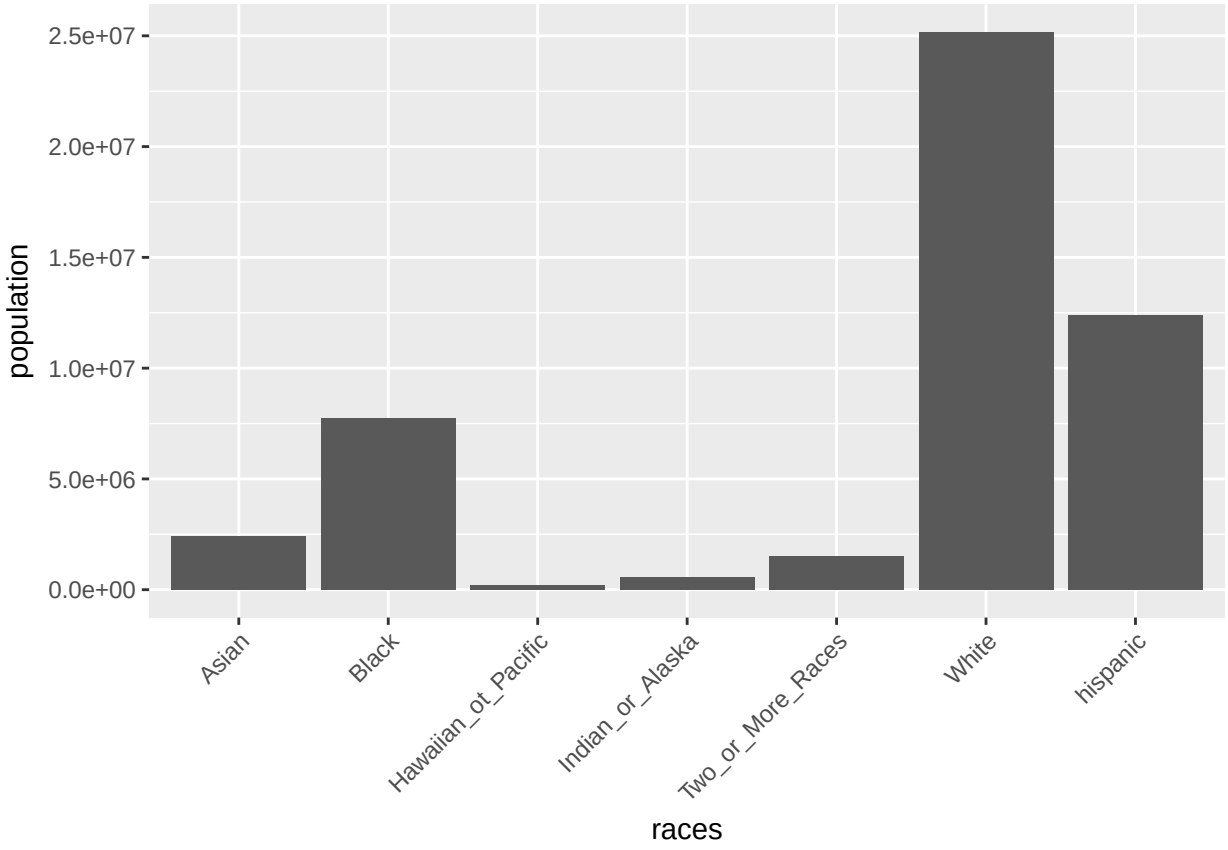
```
library(tidyverse)
crdc <- read_csv("CRDC2013_14_SCH.csv",
    na = c("-2", "-5", "-9"))
```

```
## 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.
```

```
race <- transmute(crdc, hispanic = SCH_ENR_HI_M +
    SCH_ENR_HI_F, Indian_or_Alaska = SCH_ENR_AM_M +
    SCH_ENR_AM_F, Asian = SCH_ENR_AS_M +
    SCH_ENR_AS_F, Hawaiian_ot_Pacific = SCH_ENR_HP_M +
    SCH_ENR_HP_F, Black = SCH_ENR_BL_M +
    SCH_ENR_BL_F, White = SCH_ENR_WH_M +
    SCH_ENR_WH_F, Two_or_More_Races = SCH_ENR_TR_M +
    SCH_ENR_TR_F)
stats <- summarize(race, sum(hispanic,
    na.rm = T), sum(Indian_or_Alaska,
    na.rm = T), sum(Asian,
    na.rm = T), sum(Hawaiian_ot_Pacific,
    na.rm = T), sum(Black,
    na.rm = T), sum(White,
    na.rm = T), sum(Two_or_More_Races,
    na.rm = T))
```

```
plotdata <- data.frame(races = c("hispanic",
    "Indian_or_Alaska", "Asian",
    "Hawaiian_ot_Pacific",
    "Black", "White", "Two_or_More_Races"),
    population = c(stats[[1]],
        stats[[2]], stats[[3]],
        stats[[4]], stats[[5]],
        stats[[6]], stats[[7]]))

ggplot(data = plotdata) +
    geom_col(aes(x = races,
        y = population)) +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
```

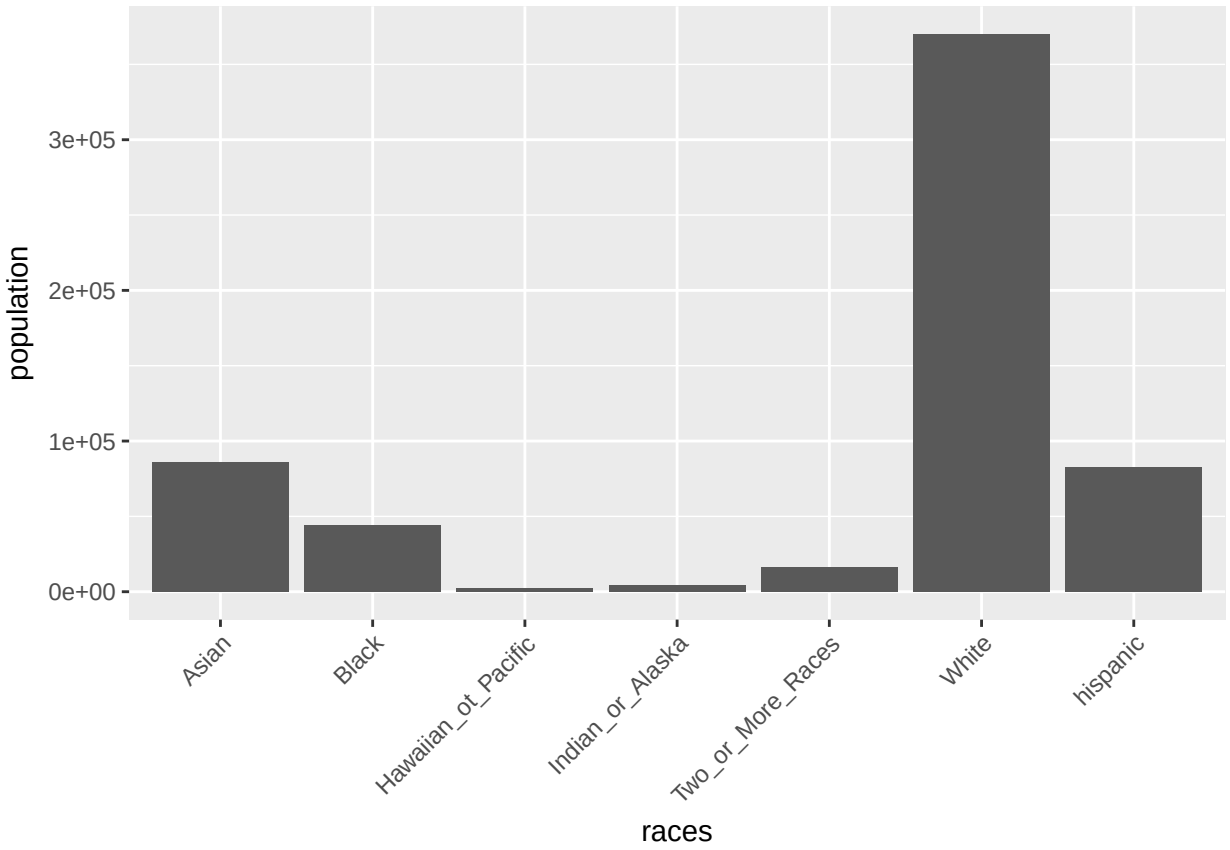


Prob 4 Create a bar plot showing the number of students of each race enrolled in a Calculus class. Comment on any similarities or differences between this distribution and the one you plotted in Problem 3.

Data from: SCH_MATHENR_CALC_HI_M,426,1,Students Enrolled in Calculus: Hispanic Male,I,19 SCH_MATHENR_CALC_HI_F,427,1,Students Enrolled in Calculus: Hispanic Female,I,19 SCH_MATHENR_CALC_AM_M Enrolled in Calculus: American Indian/Alaska Native Male,I,19 SCH_MATHENR_CALC_AM_F,429,1,Students Enrolled in Calculus: American Indian/Alaska Native Female,I,19 SCH_MATHENR_CALC_AS_M,430,1,Students Enrolled in Calculus: Asian Male,I,19 SCH_MATHENR_CALC_AS_F,431,1,Students Enrolled in Calculus: Asian Female,I,19 SCH_MATHENR_CALC_HP_M,432,1,Students Enrolled in Calculus: Native

Hawaiian/Pacific Islander Male,I,19 SCH_MATHENR_CALC_HP_F,433,1,Students Enrolled in Calculus: Native Hawaiian/Pacific Islander Female,I,19 SCH_MATHENR_CALC_BL_M,434,1,Students Enrolled in Calculus: Black Male,I,19 SCH_MATHENR_CALC_BL_F,435,1,Students Enrolled in Calculus: Black Female,I,19 SCH_MATHENR_CALC_WH_M,436,1,Students Enrolled in Calculus: White Male,I,19 SCH_MATHENR_CALC_WH_F,437,1,Students Enrolled in Calculus: White Female,I,19 SCH_MATHENR_CALC_TR_M,438,1,Students Enrolled in Calculus: Two or More Races Male,I,19 SCH_MATHENR_CALC_TR_F,439,1,Students Enrolled in Calculus: Two or More Races Female,I,19

```
race_cal <- transmute(crdc,
    hispanic = SCH_MATHENR_CALC_HI_M +
        SCH_MATHENR_CALC_HI_F,
    Indian_or_Alaska = SCH_MATHENR_CALC_AM_M +
        SCH_MATHENR_CALC_AM_F,
    Asian = SCH_MATHENR_CALC_AS_M +
        SCH_MATHENR_CALC_AS_F,
    Hawaiian_ot_Pacific = SCH_MATHENR_CALC_HP_M +
        SCH_MATHENR_CALC_HP_F,
    Black = SCH_MATHENR_CALC_BL_M +
        SCH_MATHENR_CALC_BL_F,
    White = SCH_MATHENR_CALC_WH_M +
        SCH_MATHENR_CALC_WH_F,
    Two_or_More_Races = SCH_MATHENR_CALC_TR_M +
        SCH_MATHENR_CALC_TR_F)
stats_cal <- summarize(race_cal,
    sum(hispanic, na.rm = T),
    sum(Indian_or_Alaska,
        na.rm = T), sum(Asian,
        na.rm = T), sum(Hawaiian_ot_Pacific,
        na.rm = T), sum(Black,
        na.rm = T), sum(White,
        na.rm = T), sum(Two_or_More_Races,
        na.rm = T))
plotdata_cal <- data.frame(races = c("hispanic",
    "Indian_or_Alaska", "Asian",
    "Hawaiian_ot_Pacific",
    "Black", "White", "Two_or_More_Races"),
    population = c(stats_cal[[1]],
        stats_cal[[2]], stats_cal[[3]],
        stats_cal[[4]], stats_cal[[5]],
        stats_cal[[6]], stats_cal[[7]]))

ggplot(data = plotdata_cal) +
    geom_col(aes(x = races,
        y = population)) +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
```



Asian are over-represented in calculus classes.Black and hispanic are under-represented.

Prob 5

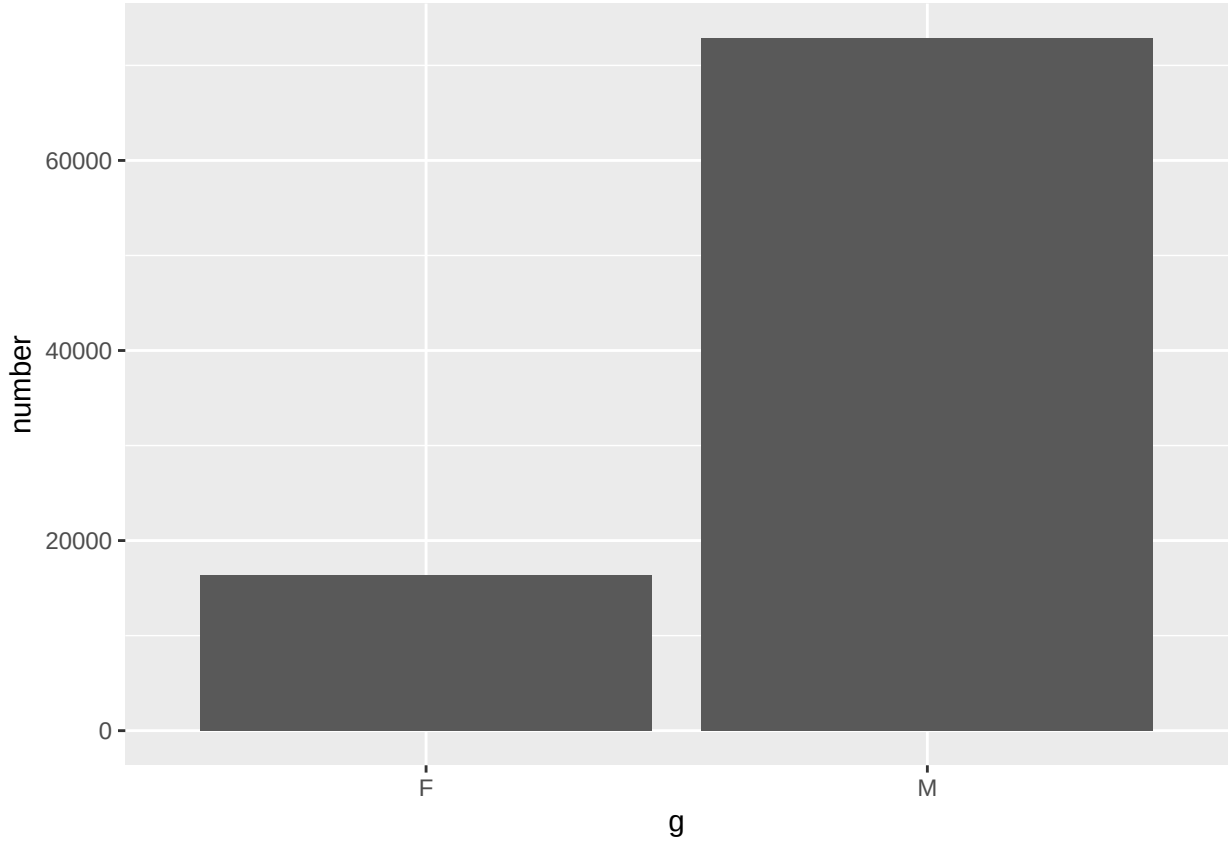
```
library(DBI)
library(RMySQL)
library(dbplyr)
```

```
##
## Attaching package: 'dbplyr'
```

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

```
con <- dbConnect(MySQL(),
    dbname = "dblp", username = "root",
    password = "mysql")
author <- tbl(con, "authors")
general <- tbl(con, "general")
author %>% filter(prob > 0.95,
    prob <= 1) %>% group_by(name) %>%
    mutate(G = ifelse(gender ==
        "M", 1, 0)) %>% summarize(Gx = sum(G)) %>%
    transmute(g = ifelse(Gx >
        0, "M", "F")) %>%
```

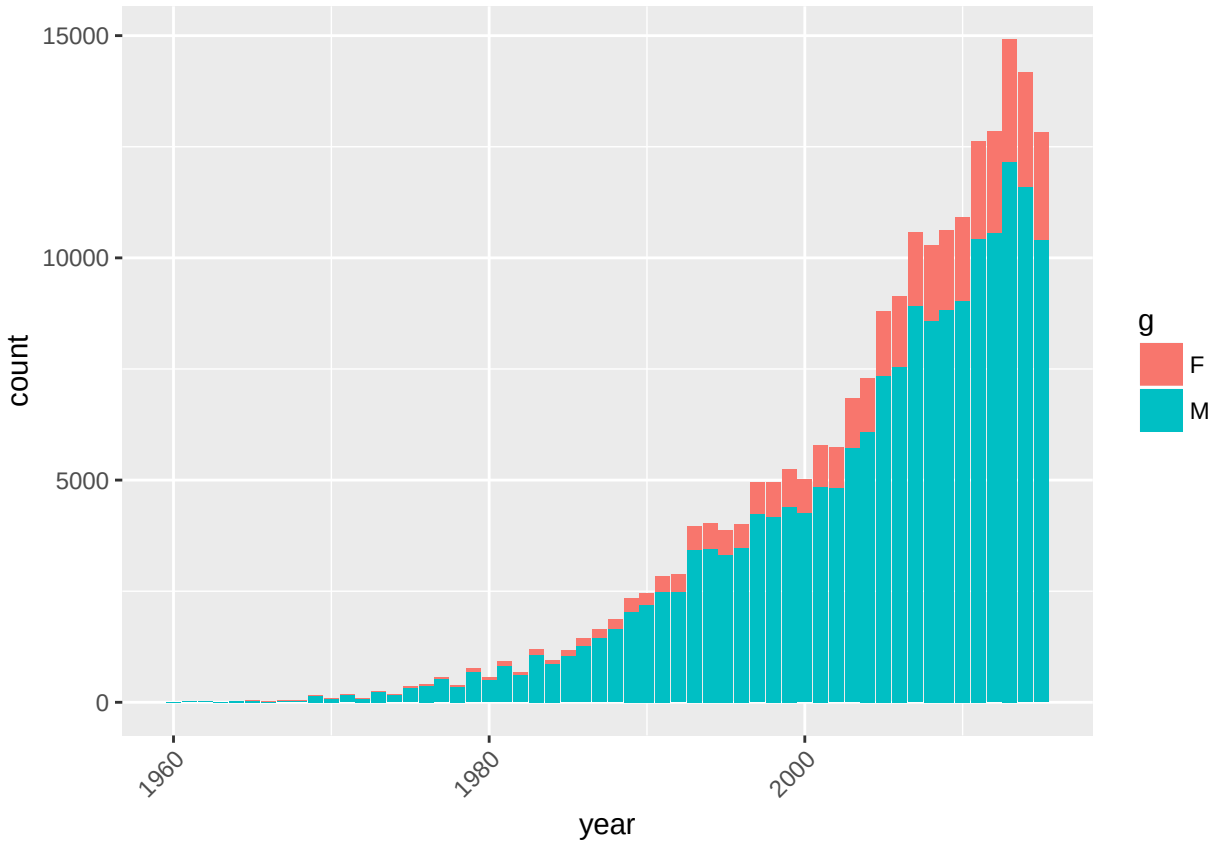
```
    group_by(g) %>% summarize(number = n()) %>%
    collect() %>% ggplot() +
    geom_col(aes(x = g, y = number))
```



Prob6 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.

```
authorinfo <- left_join(author,
    general, by = "k")
pd <- authorinfo %>% filter(prob >
    0.95, prob <= 1) %>% mutate(G = ifelse(gender ==
    "M", 1, 0)) %>% group_by(name,
    year) %>% summarize(Gx = sum(G)) %>%
    mutate(g = ifelse(Gx >
        0, "M", "F"))

ggplot(collect(pd)) + geom_bar(aes(x = year,
    fill = g), position = "stack") +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
```

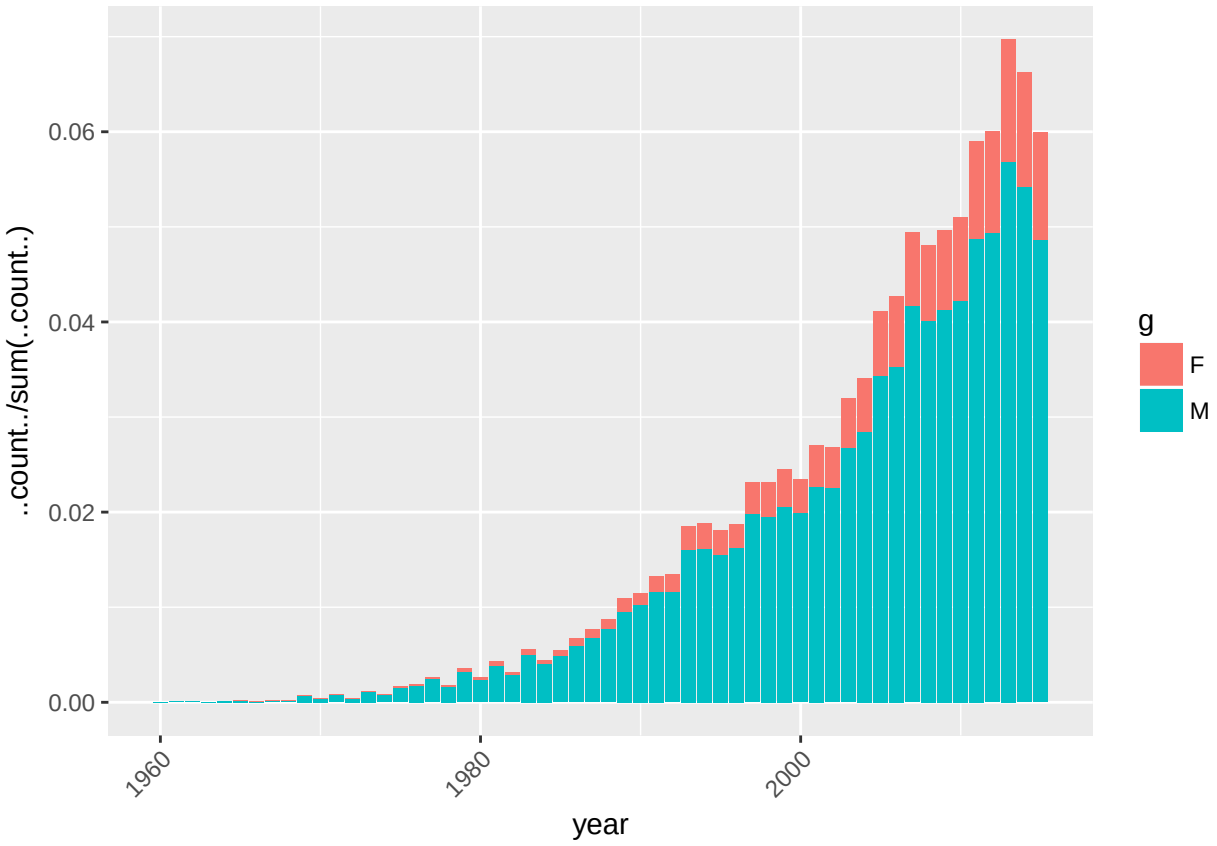
```
## Warning in .local(conn, statement, ...): Decimal MySQL column 2 imported as
## numeric
```



Prob 7

```
ggplot(collect(pd)) + geom_bar(aes(x = year,
    y = ..count../sum(..count..),
    fill = g), position = "stack") +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
```

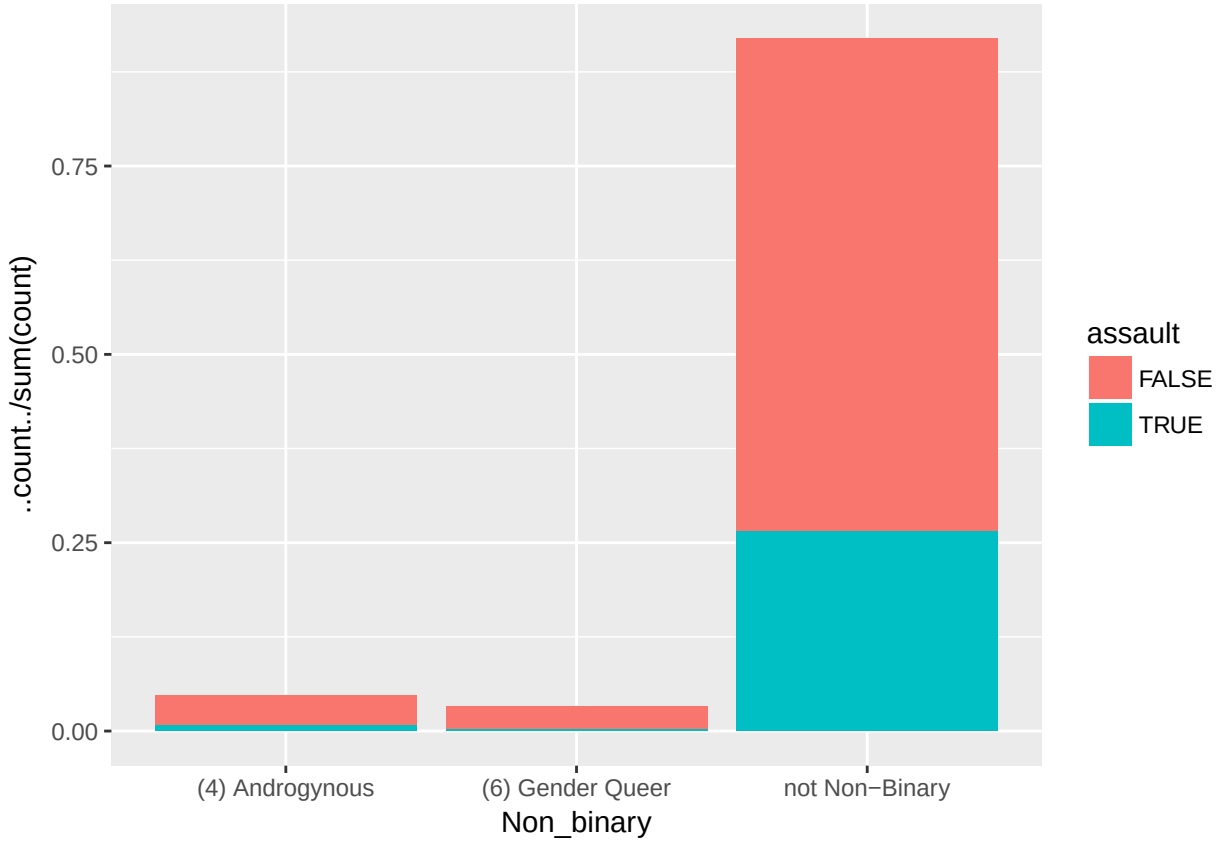
```
## Warning in .local(conn, statement, ...): Decimal MySQL column 2 imported as
## numeric
```



Prob 8

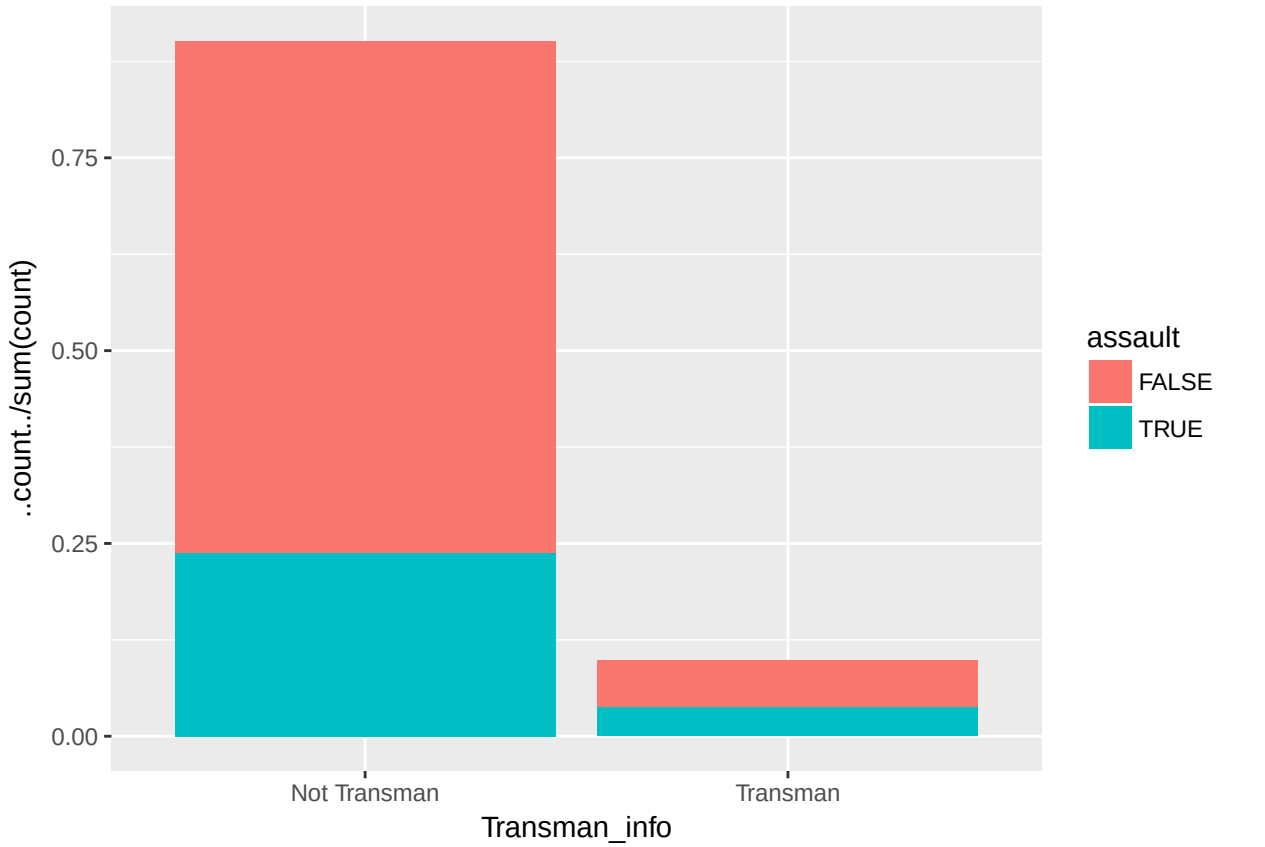
```
load("/Users/yzh/Desktop/R data/HW2 data/ICPSR_31721/DS0001/31721-0001-Data.rda")
nofactor <- data.frame(lapply(da31721.0001,
    as.character), stringsAsFactors = FALSE)
health <- filter(nofactor,
    is.na(Q6) == F, is.na(Q96) ==
        F, is.na(Q106) ==
        F)
nonbi <- mutate(health, Non_binary = ifelse(Q6 ==
    "(4) Androgynous" | Q6 ==
    "(6) Gender Queer", Q6,
    "not Non-Binary"), Transwoman_info = ifelse(Q5 ==
    "(1) Male" & Q6 == "(3) Transgender",
    "Transwoman", "Not Transwoman"),
    Transman_info = ifelse(Q5 ==
        "(2) Female" & Q6 ==
        "(3) Transgender",
        "Transman", "Not Transman"),
    assault = ifelse(Q96 ==
        "(1) Yes" | Q106 ==
        "(1) No", T, F))
ggplot(nonbi) + geom_bar(aes(x = Non_binary,
    y = ..count../sum(count),
```

```
    fill = assault), position = "stack")
```




```
ggplot(nonbi) + geom_bar(aes(x = Transwoman_info,
    y = ..count../sum(count),
    fill = assault), position = "stack")
```


```
ggplot(nonbi) + geom_bar(aes(x = Transman_info,
    y = ..count../sum(count),
    fill = assault), position = "stack")
```



Note: I plotted 2 barplots since in problem "barplots" indicates multiple plots

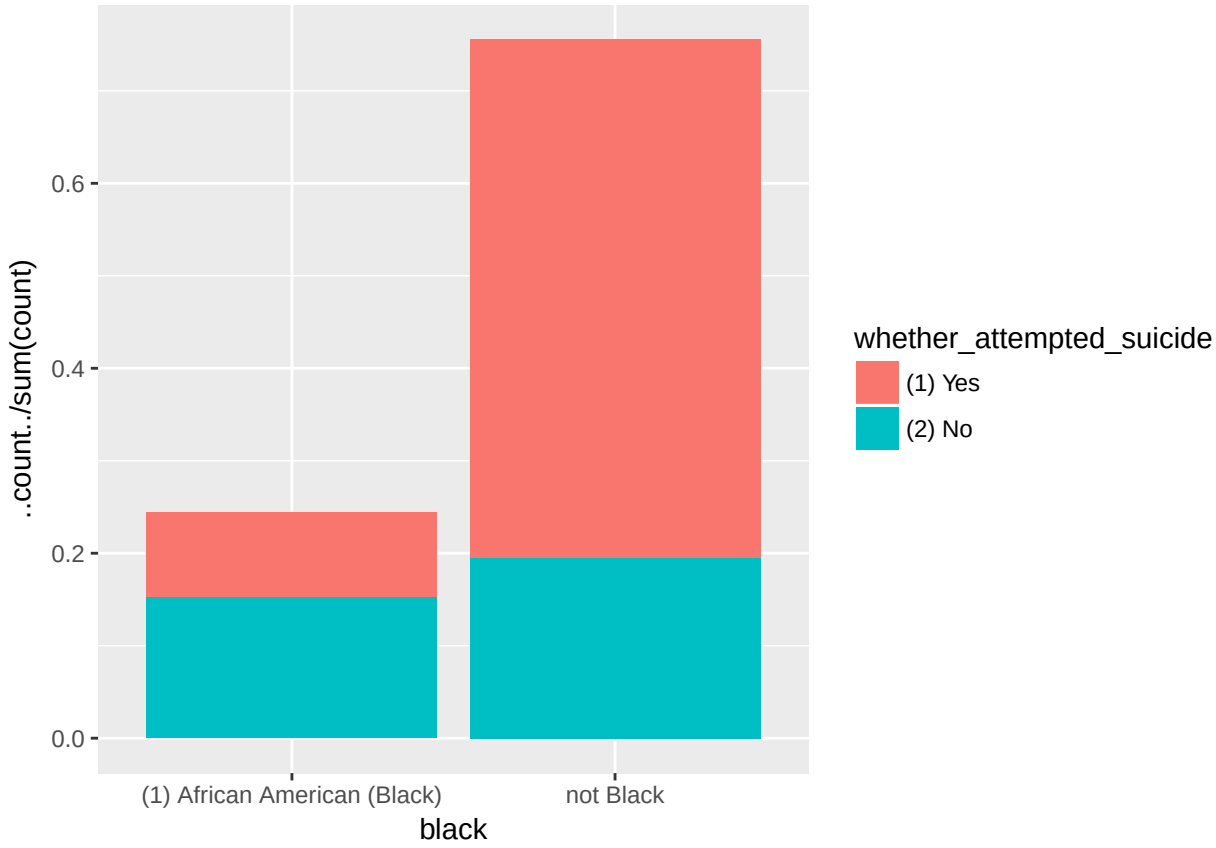
Prob9 Q131: Have you ever thought about killing yourself?

```
sui <- filter(nofactor, is.na(RACE) ==
    F, is.na(Q131) == F)
racesui <- mutate(sui, black = ifelse(RACE ==
    "(1) African American (Black)",
    RACE, "not Black"), white = ifelse(RACE ==
    "(2) White (Caucasian)",
    RACE, "not white"), Hispanic_Latinx = ifelse(RACE ==
    "(3) Hispanic or Latino/Latina",
    RACE, "not hispanic"),
    Native_American = ifelse(RACE ==
        "(4) Native American/American Indian",
        RACE, "not native american"))
library(plyr)
```

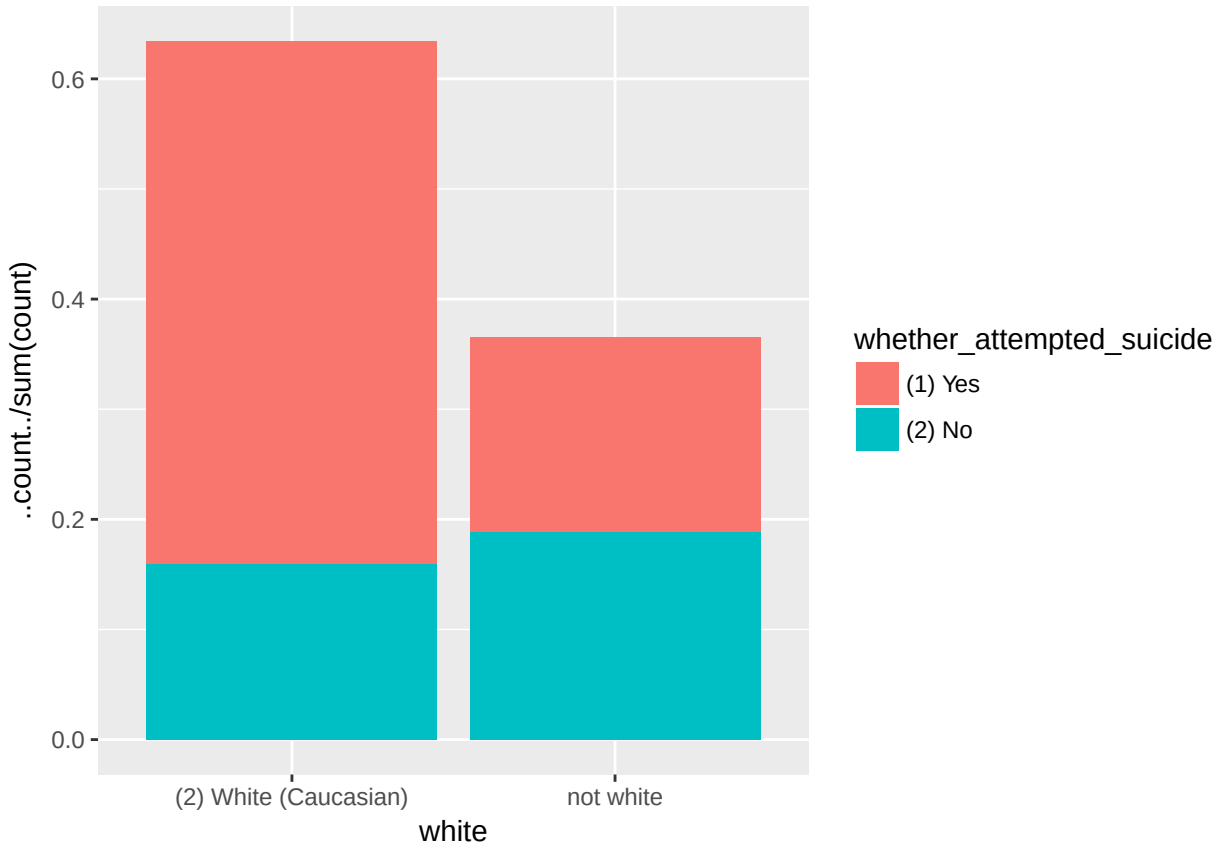
```
## ------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following object is masked from 'package:maps':
```

```
##
##     ozone

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact
```

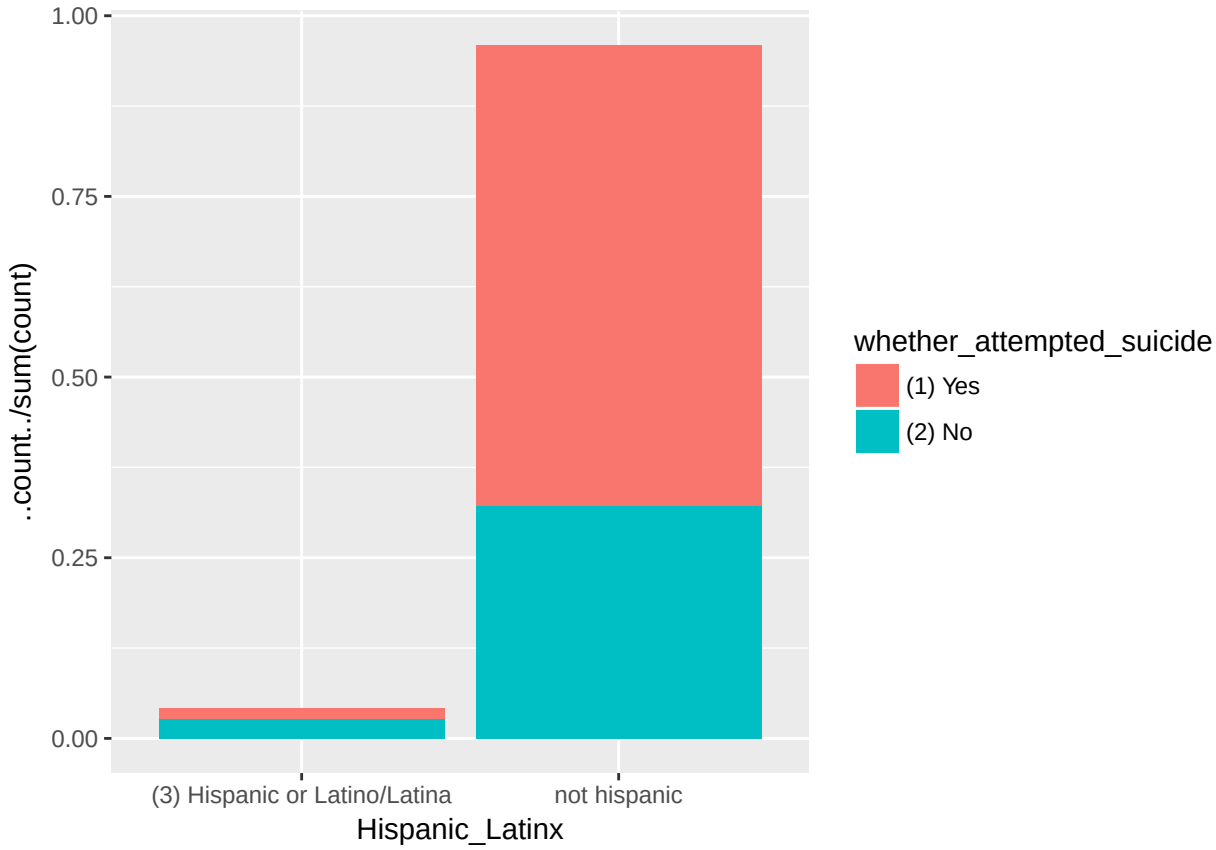
```
racesui <- rename(racesui,
    c(Q131 = "whether_attempted_suicide"))
library(dplyr)
ggplot(racesui) + geom_bar(aes(x = black,
    y = ..count../sum(count),
    fill = whether_attempted_suicide),
    position = "stack")
```



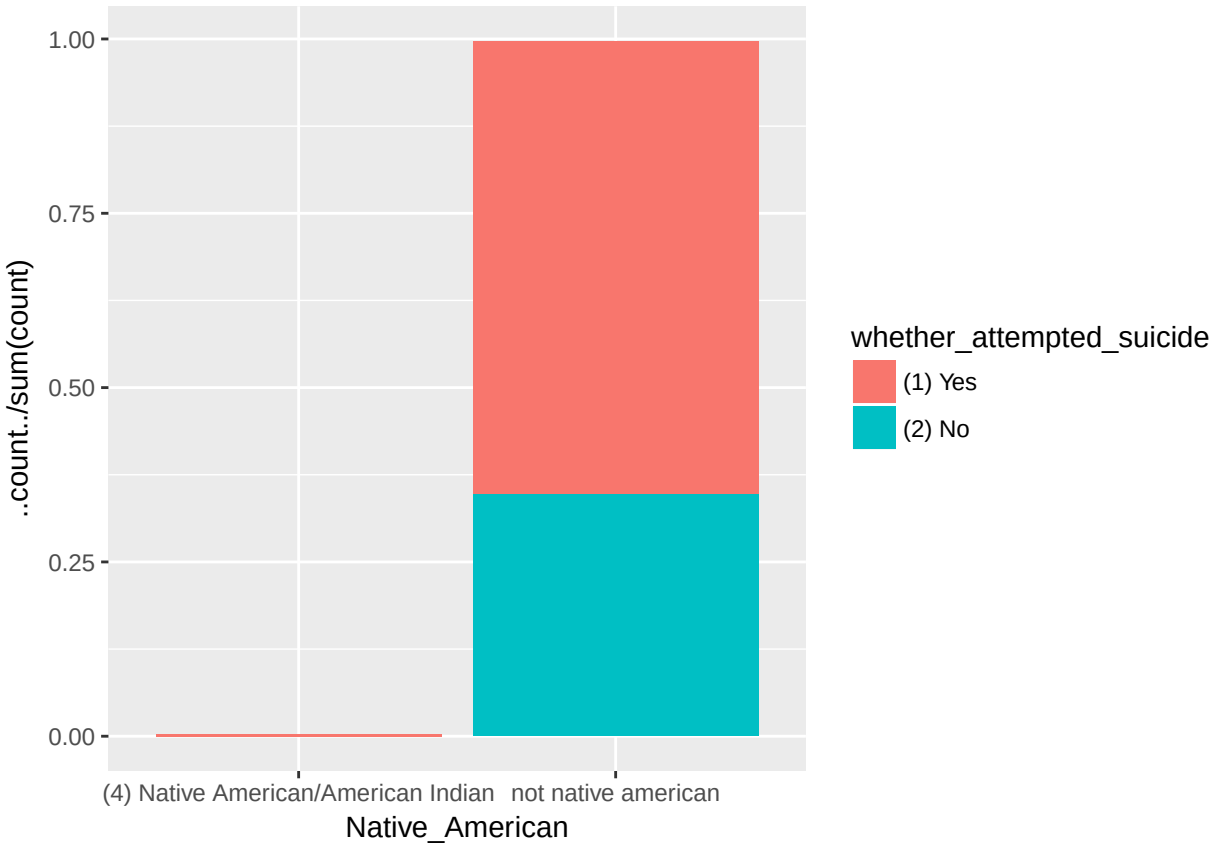

```
ggplot(racesui) + geom_bar(aes(x = white,
    y = ..count../sum(count),
    fill = whether_attempted_suicide),
    position = "stack")
```




```
ggplot(racesui) + geom_bar(aes(x = Hispanic_Latinx,
    y = ..count../sum(count),
    fill = whether_attempted_suicide),
    position = "stack")
```




```
ggplot(racesui) + geom_bar(aes(x = Native_American,
    y = ..count../sum(count),
    fill = whether_attempted_suicide),
    position = "stack")
```




```
suitotal <- filter(nofactor,
    is.na(Q131) == F)
totalsui <- transmute(suitotal,
    suicide = ifelse(Q131 ==
        "(1) Yes", T, F))
summarize(totalsui, proportion = mean(suicide,
    na.rm = T))
```

```
##   proportion
## 1  0.6520468
```

It is higher than the national average.

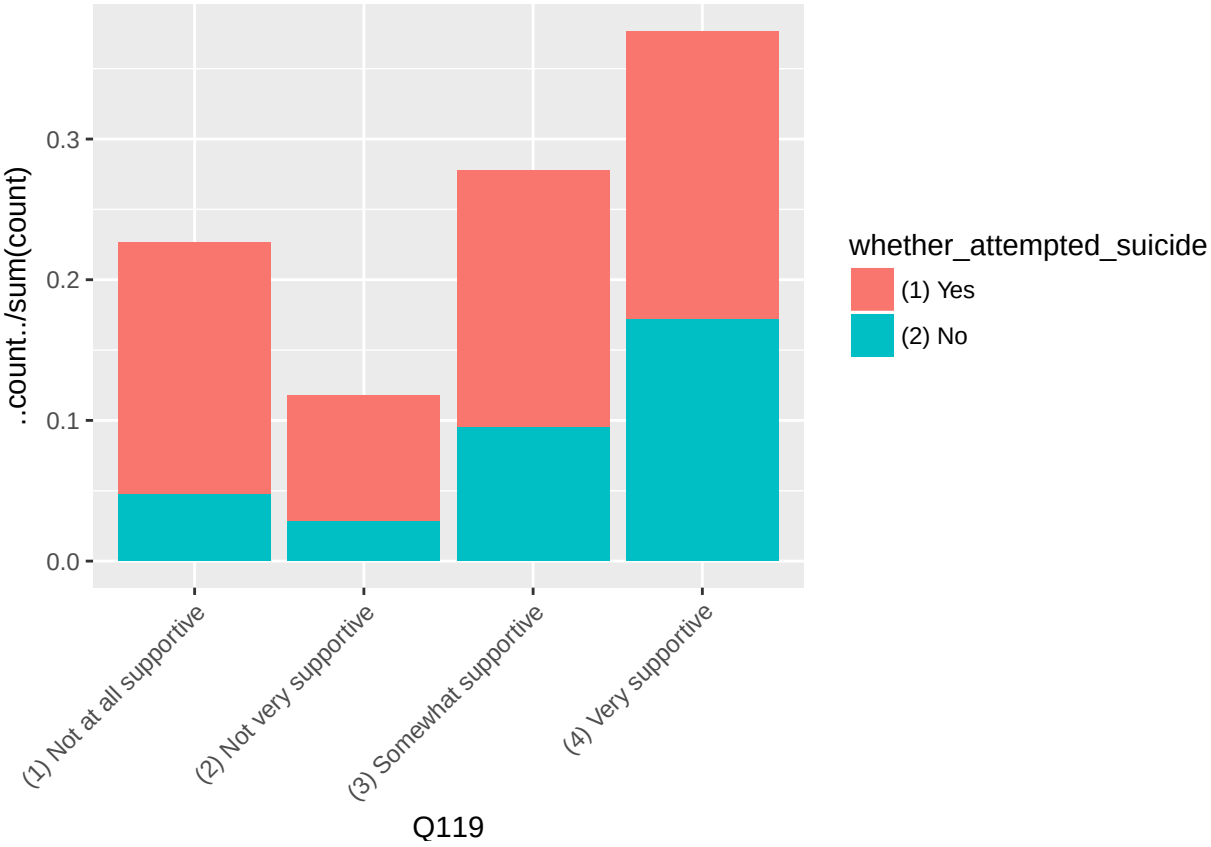
Note: I plotted 4 barplots since in problem "barplots" indicates multiple plots

---

Prob10

```
suifam <- filter(nofactor,
    is.na(Q119) == F, is.na(Q131) ==
        F)
suifam2 <- filter(suifam,
    Q119 != "(5) Not applicable to me")
library(plyr)
suifam2 <- rename(suifam2,
    c(Q131 = "whether_attempted_suicide"))
library(dplyr)
```

```
ggplot(suifam2) + geom_bar(aes(x = Q119,
    y = ..count../sum(count),
    fill = whether_attempted_suicide),
    position = "stack") +
    theme(axis.text.x = element_text(angle = 45,
        hjust = 1))
```



Note: Although in problem "barplots" indicates multiple plots , I feel one barplot is enough and clear