hw3-Aishwarya Vantipuli

Aishwarya

February 12, 2019

PART-A

PROBLEM-1

```
#Loading Libraries
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(ggplot2)
library(readr)
library(dplyr)


#reading data as data frame
data <- as.data.frame(read_csv("I:/Data Science/NEU/SEM1/Introduction to Data management/hw3/master.csv
```

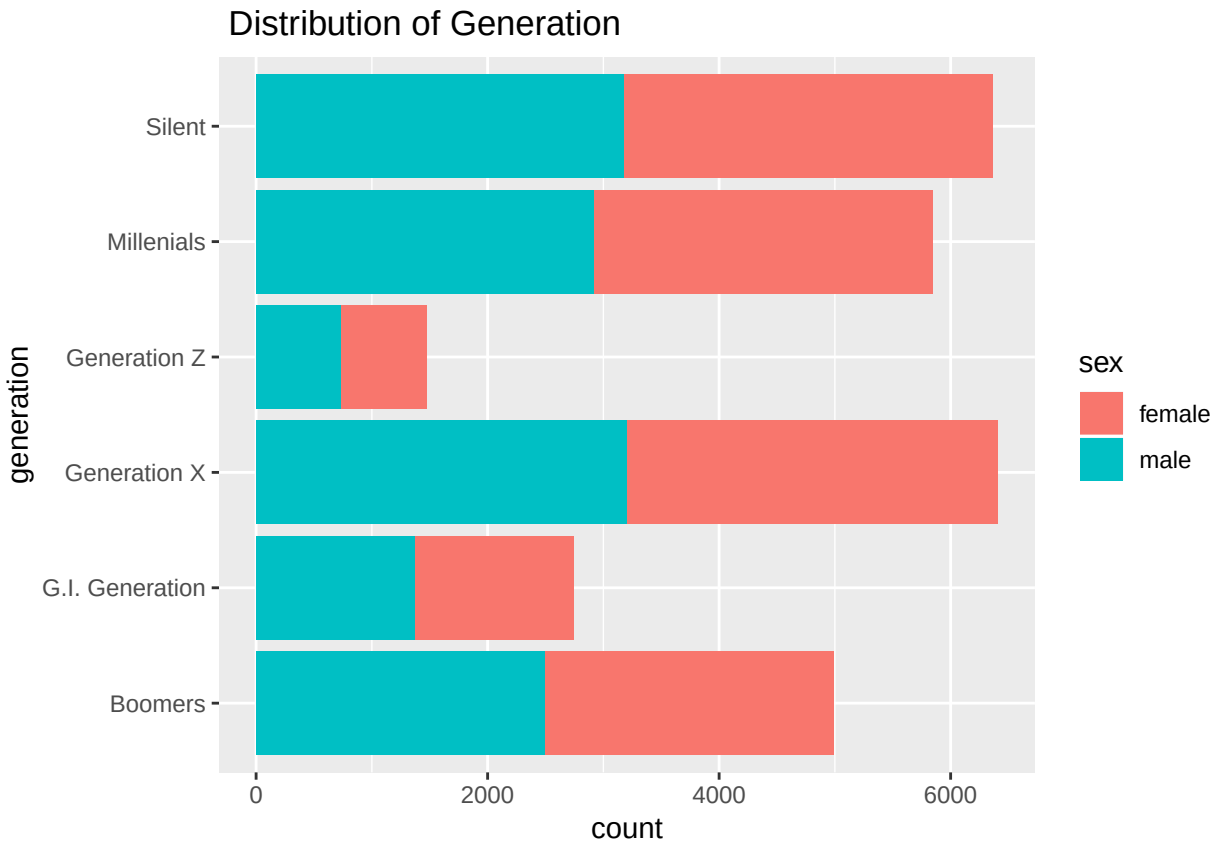
```
## Parsed with column specification:
## cols(
##   country = col_character(),
##   year = col_double(),
##   sex = col_character(),
##   age = col_character(),
##   suicides_no = col_double(),
##   population = col_double(),
##   `suicides/100k pop` = col_double(),
##   `country-year` = col_character(),
##   `HDI for year` = col_double(),
##   `gdp_for_year ($)` = col_number(),
##   `gdp_per_capita ($)` = col_double(),
##   generation = col_character()
## )
```

```r
#Tidying
#Data is downloaded from kaggle.
#As the data is almost clean,
#basic transformation would suffice the current requirement.


#Normalising GDP variable
data <- mutate(data,
               `gdp in $100k` = `gdp_for_year ($)`/(100*1000))

#Excluding unwanted variables
data <- select(data, -`country-year`, -`gdp_for_year ($)`)



#Displaying first 10 observations
data[1:10,]
```

```
##    country year    sex         age suicides_no population
## 1  Albania 1987   male 15-24 years          21     312900
## 2  Albania 1987   male 35-54 years          16     308000
## 3  Albania 1987 female 15-24 years          14     289700
## 4  Albania 1987   male   75+ years           1      21800
## 5  Albania 1987   male 25-34 years           9     274300
## 6  Albania 1987 female   75+ years           1      35600
## 7  Albania 1987 female 35-54 years           6     278800
## 8  Albania 1987 female 25-34 years           4     257200
## 9  Albania 1987   male 55-74 years           1     137500
## 10 Albania 1987 female  5-14 years           0     311000
##    suicides/100k pop HDI for year gdp_per_capita ($)      generation
## 1               6.71           NA               796    Generation X
## 2               5.19           NA               796          Silent
## 3               4.83           NA               796    Generation X
## 4               4.59           NA               796 G.I. Generation
## 5               3.28           NA               796         Boomers
## 6               2.81           NA               796 G.I. Generation
## 7               2.15           NA               796          Silent
## 8               1.56           NA               796         Boomers
## 9               0.73           NA               796 G.I. Generation
## 10              0.00           NA               796    Generation X
##    gdp in $100k
## 1      21566.25
## 2      21566.25
## 3      21566.25
## 4      21566.25
## 5      21566.25
## 6      21566.25
## 7      21566.25
## 8      21566.25
## 9      21566.25
## 10     21566.25
```

PROBLEM 2

```
#obs 1
ggplot(data) +
  geom_bar(data,mapping = aes(x = generation,fill = sex)) +
  ggtitle(" Distribution of Generation") +
  coord_flip()
```



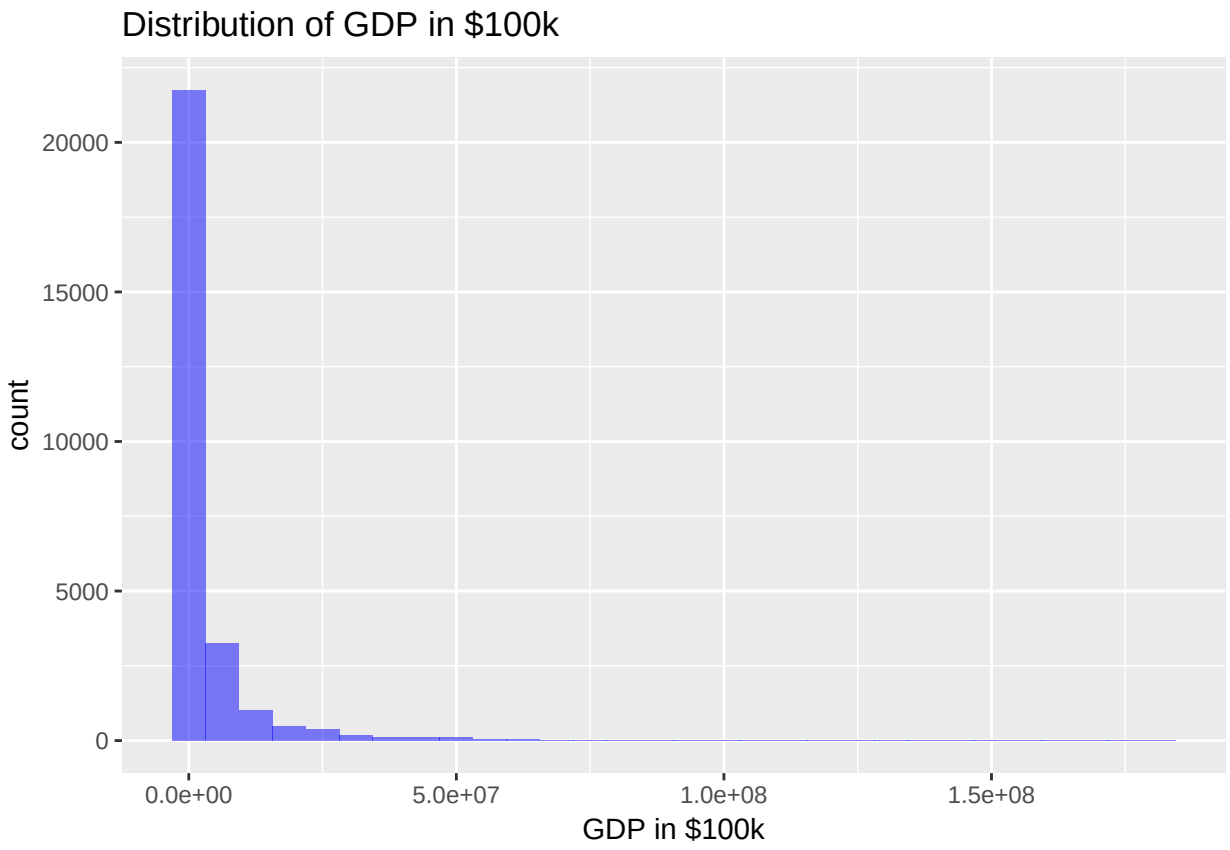

```
data %>% count(generation)
```

```
## # A tibble: 6 x 2
##   generation          n
##   <chr>           <int>
## 1 Boomers          4990
## 2 G.I. Generation  2744
## 3 Generation X     6408
## 4 Generation Z     1470
## 5 Millenials       5844
## 6 Silent           6364
```

Based on graph, Genration X and Silent have higher number of suicide rates. Calculated results also shows the same.

```
#OBS 2

ggplot(data) + geom_histogram(aes(x = `gdp in $100k`),
                              fill = 'blue', alpha= 0.5) +
  ggtitle("Distribution of GDP in $100k") +
  xlab("GDP in $100k")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

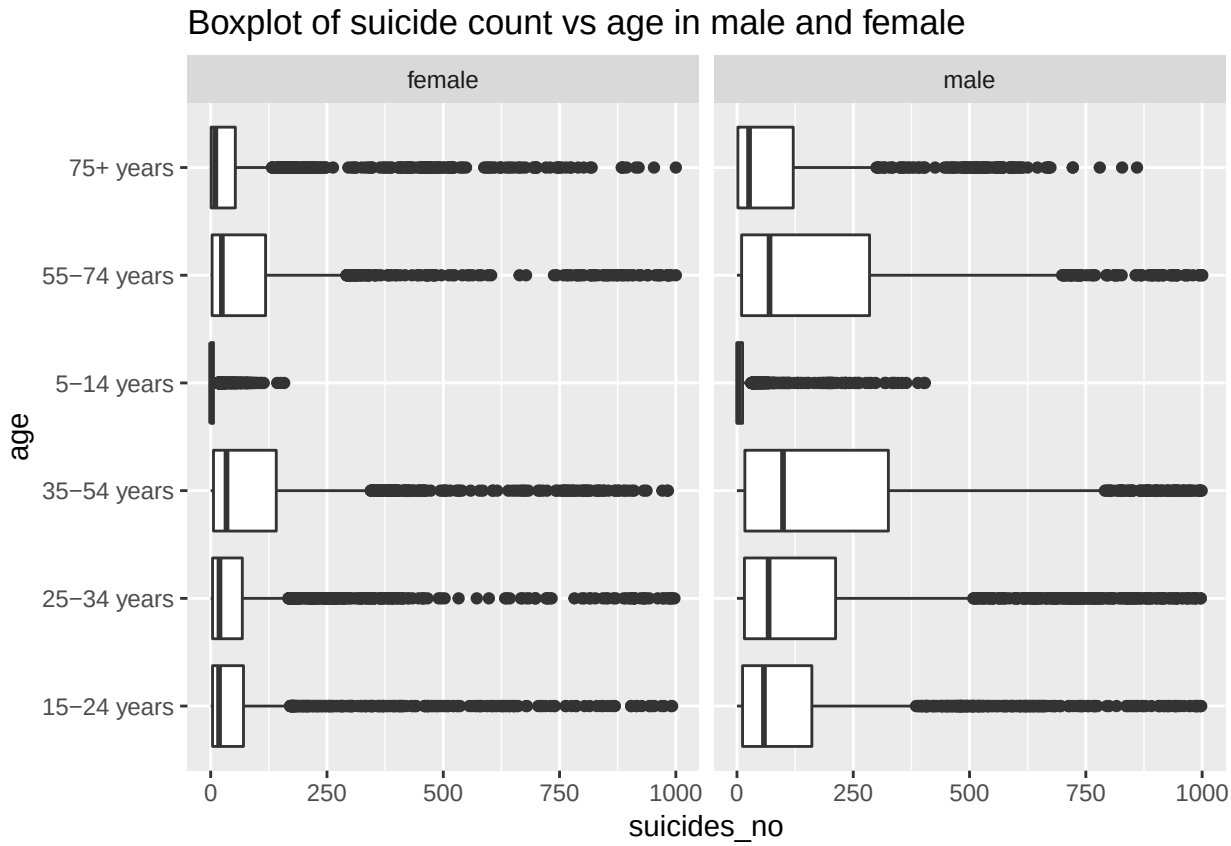



```
data %>% count(cut_interval(`gdp in $100k`, n = 10))
```

```
## # A tibble: 10 x 2
##    `cut_interval(\`gdp in $100k\`, n = 10)`     n
##    <fct>                                    <int>
##  1 [469,1.81e+07]                           26224
##  2 (1.81e+07,3.62e+07]                        888
##  3 (3.62e+07,5.44e+07]                        336
##  4 (5.44e+07,7.25e+07]                        108
##  5 (7.25e+07,9.06e+07]                         48
##  6 (9.06e+07,1.09e+08]                         48
##  7 (1.09e+08,1.27e+08]                         36
##  8 (1.27e+08,1.45e+08]                         48
##  9 (1.45e+08,1.63e+08]                         48
## 10 (1.63e+08,1.81e+08]                         36
```

```
#obs 3

ggplot(data) + geom_boxplot(aes(y = suicides_no,  x = age)) +
  facet_grid(~sex) +
 coord_flip() +
  ylim(c(0,1000)) +
  ggtitle("Boxplot of suicide count vs age in male and female")
```

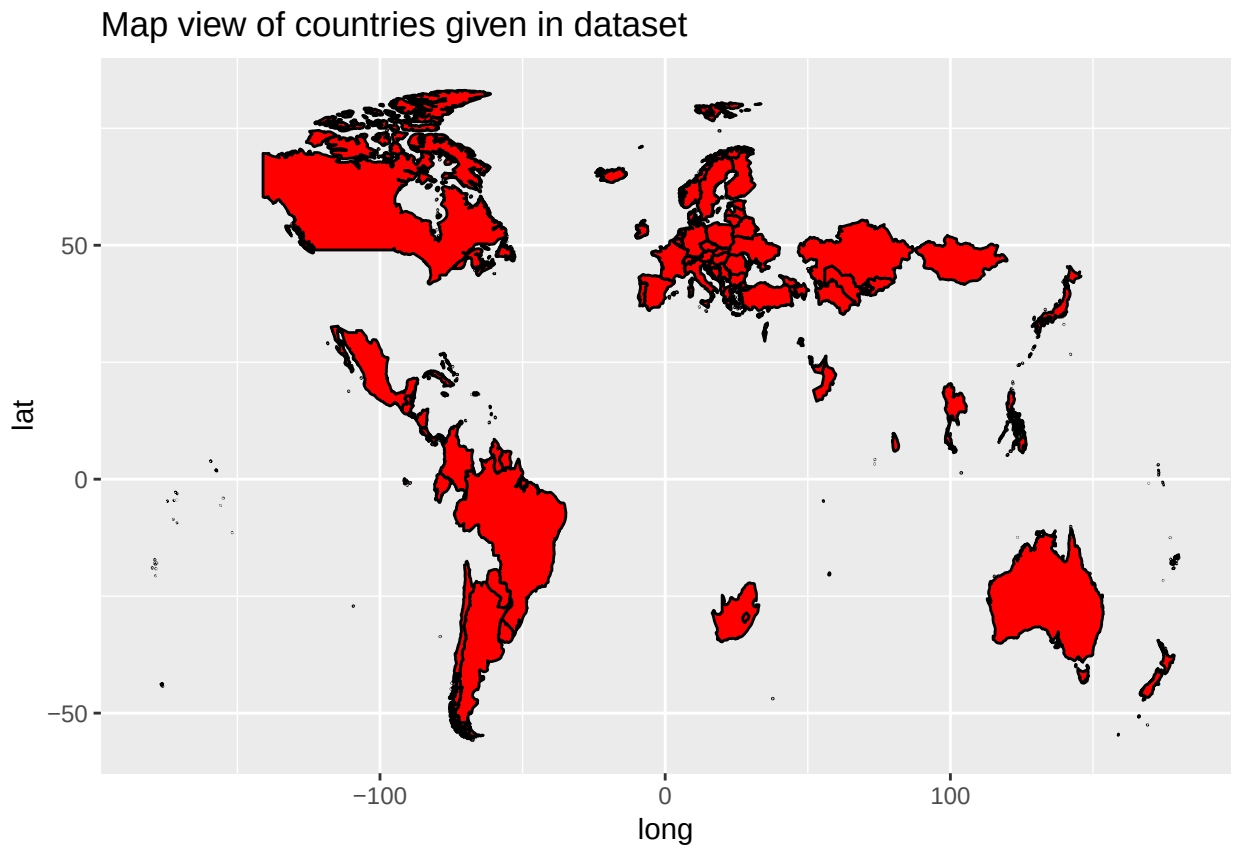
```
## Warning: Removed 1467 rows containing non-finite values (stat_boxplot).
```



There are comparetively higher no. of suicides recorded in male than female. Women are undergoing higher levels of stress at the age in between 35-74 years whereas in men higher rate is observed in between 35-54 years.

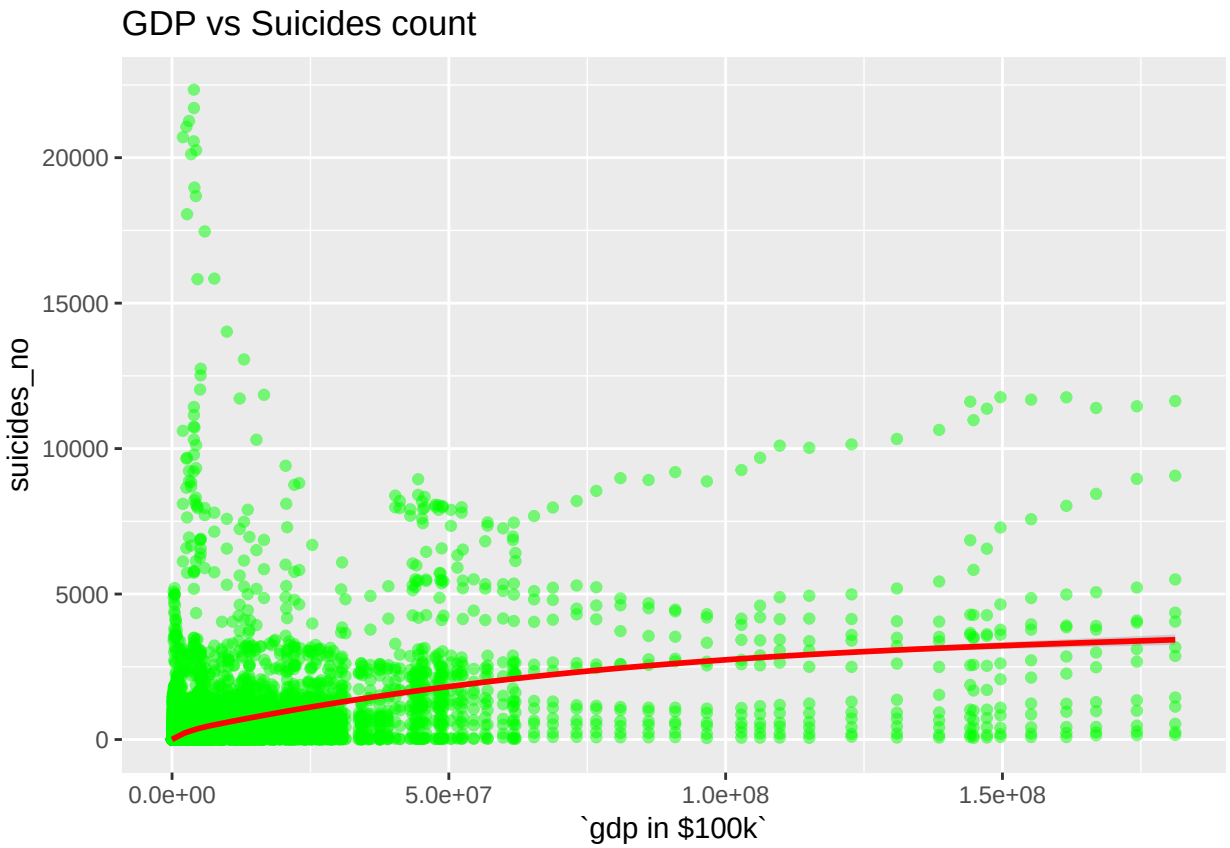
```
#obs 4

world <- ggplot2::map_data("world")

df <- data.frame(region = c(data$country))


world_new <- world[world$region %in% df$region, ]

ggplot(world_new) +
  geom_polygon(mapping=aes(x=long, y=lat, group = group),
```

```
              fill= 'red', color = "black") +
  ggtitle("Map view of countries given in dataset")
```




```
#obs 5
ggplot(data, mapping = aes(x = `gdp in $100k`, y = suicides_no)) +
  geom_point( position = "jitter", color = "green", alpha = 0.5) +
  geom_smooth(color = 'red') +
  ggtitle("GDP vs Suicides count")
```

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



Although there are higher no. of sucides in less earning countries, the trend seems to decrease at first and then increase proportionately with increase in gdp

PART B

PROBLEM-3

```
#install.packages(c("DBI", "RSQLite", "dbplyr"))
#install.packages("RMySQL")


library("dbplyr")
```

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

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

```
library("RMySQL")
```

```
## Loading required package: DBI
```

```
library("RSQLite")
```

```
##
## Attaching package: 'RSQLite'
```

```
## The following object is masked from 'package:RMySQL':
##
##     isIdCurrent
```

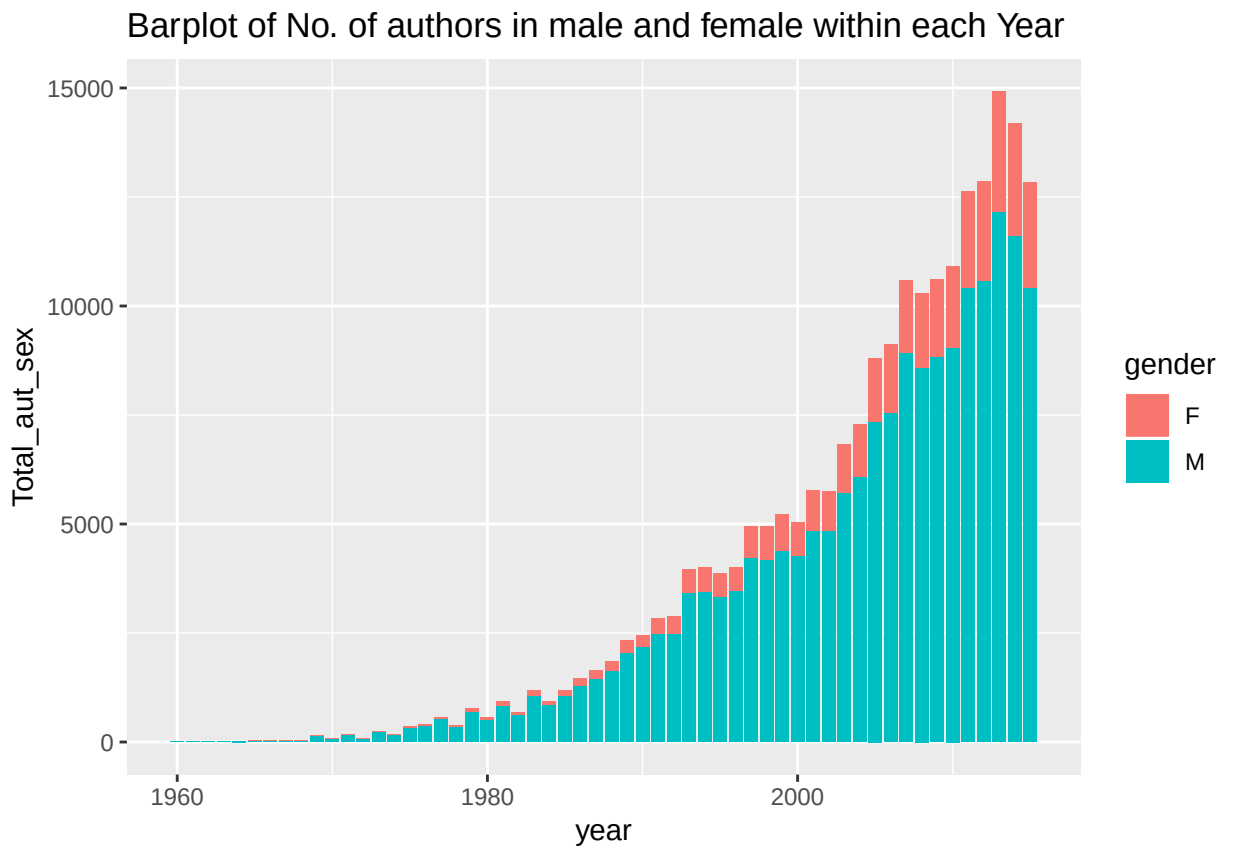
```
library("DBI")



#cREATING cONNECTION
con <- dbConnect(MySQL(),
                 user = 'root',
                 password = 'Premkumar007',
                 host = 'localhost',
                 dbname = 'dblp')
#LOADING TABLES
dblp_gen <- tbl(con, "general")

dblp_aut <- tbl(con, "authors")



total <-dblp_aut %>% left_join(dblp_gen) %>% group_by(year) %>%
  filter(gender %in% c("M", "F") & prob >= 0.95)   %>%
  summarise(Total_aut = n_distinct(name))
```

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

```
Total_sex <- dblp_aut %>% left_join(dblp_gen) %>% group_by(year, gender) %>%
  filter(gender %in% c("M", "F") & prob >= 0.95)   %>%
  summarise(Total_aut_sex = n_distinct(name))
```

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

```
#Prob 3
Total_sex %>%collect() %>%

  ggplot() +geom_col(aes(x = year,y = Total_aut_sex, fill = gender)) +

   ggtitle('Barplot of No. of authors in male and female within each Year')
```

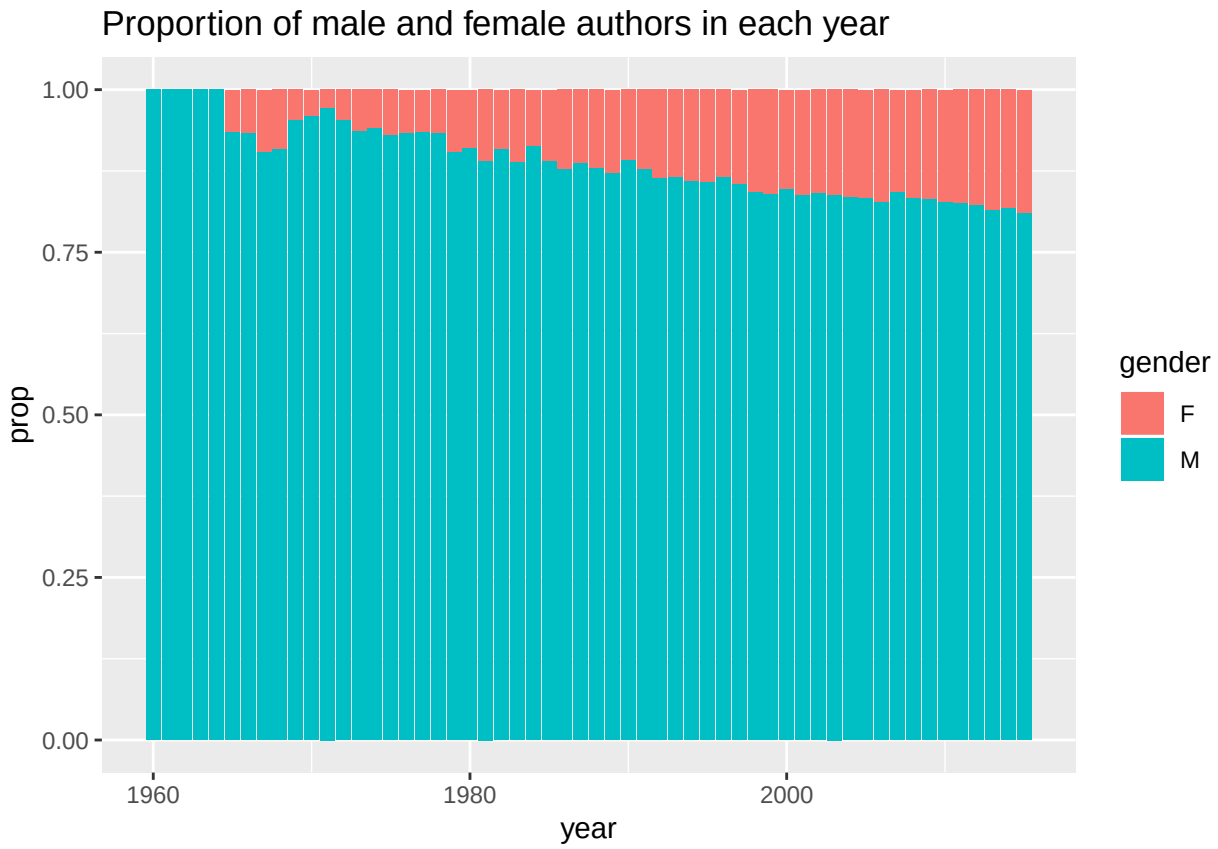



```
#prob 5
```

There are no female authors untill arround 1970's. Female authors are increasing at lower rates comapared to male. There seems to be decrease in authors publications in recent years.

```
#prob 4

right_join(Total_sex,total) %>%
  group_by(year, gender) %>%
  summarise(prop = Total_aut_sex/Total_aut) %>%
  ggplot() + geom_col(aes(x= year, y = prop, fill = gender)) +
  ggtitle("Proportion of male and female authors in each year")
```

```
## Joining, by = "year"
```

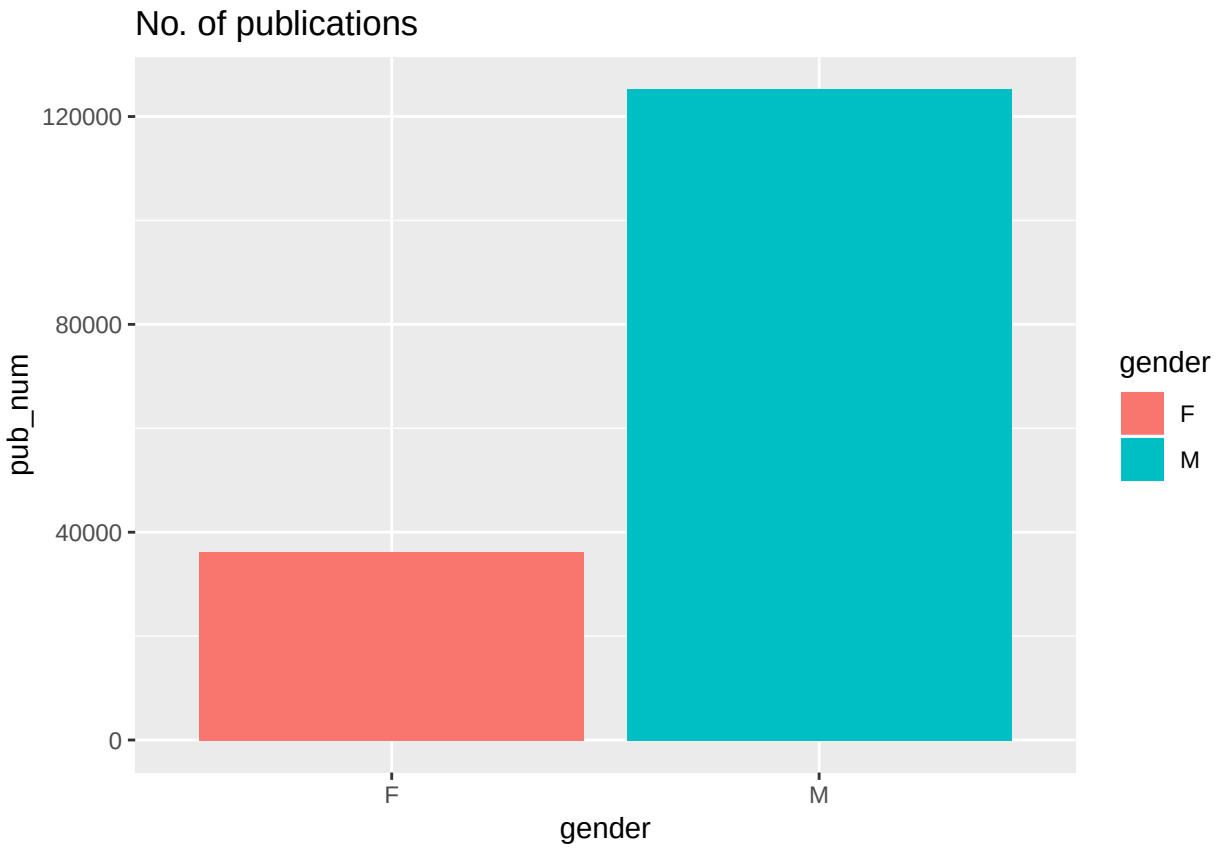


Proportion of male is way higher than female and there are no female authors untill around 1965

PROBLEM 5

```
#prob 5


dblp_aut %>% left_join(dblp_gen) %>% group_by(year, gender, pos = 0) %>%
  filter(gender %in% c("M", "F") & prob > 0.95) %>%
  summarise(pub_num = n_distinct(k)) %>%
  ggplot() + geom_col(aes(x = gender,y = pub_num, fill = gender)) +
  ggtitle('No. of publications')
```

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



There are lower number of female publications than male