```r
#Class: Week 03
#Course: Big Data and Social Analysis
#Semester: Spring 2021
#Lesson: Dataframe and Import Data
#Instructor: Chung-pei Pien
#Organization: ICI, NCCU

### Student Information --------

#Chinese Name: 辛鐘成
#English First Name: Jongsung
#UID: 110ZU1038
#E-mail: sjongsung97@gmail.com

### Questions --------

#Please answer the following questions. Remember, No Comments, No Points!!!!!!!

#Read 116th and 117th US House lawmaker table: house_116.csv and house_117.csv
setwd("C:/Users/sung/Desktop/Big data with R/0303 US HOUSE") # set the working directory so that I can bring two files
house_115 <- read.csv("house_115.csv") # it is from our last class but I put it here to make sure
colnames(house_115)[c(7,10:11)] <- c('brith','twitter','facebook')
house_115_2016 <- house_115[!(house_115$successor == 1) & !(house_115$non_voting == 1),] # it is ([i,j])
house_116 <- read.csv("house_116.csv") # read csv file
house_117 <- read.csv("house_117.csv")

#Clean and create the following tables:

#Question 1: (6 points)

#Create a table object house_116_2019 which shows 116th house of lawmakers in the beginning of the 116th term (Jan. 3, 2019).
nrow(house_116) # check if the nrow is equal to the real data = 451. so it's not the same. We need to clean it.
colnames(house_116) # find the names of column

house_116_c4 <- house_116[house_116$change_party == '1',] # I found out there are 3 people who have duplicated data. So change_party indicates the people who changed their party
nrow(house_116_c4) #6
house_116_c5 <- house_116_c4[-c(2, 4, 6), ] # it should be like this in house_116_c5 but we will remove it from house_116_2019 like below
nrow(house_116_c5) #3
```

```
house_116 <- house_116[-c(8, 283, 418), ] # I removed the same data that indicates newer update
nrow(house_116) # 448

house_116_c1 <- house_116[house_116$vacate == '1',] # 12
nrow(house_116_c1)

house_116_c2 <- house_116[house_116$successor == '1',] # 8
nrow(house_116_c2)

house_116_c3 <- house_116[house_116$non_voting == '1',] # 6
nrow(house_116_c3)

house_116_2019 <- house_116[!(house_116$successor == 1) & !(house_116$non_voting == 1),]
nrow(house_116_2019) # 434 which is what wiki shows

# When the term begins(eg.2016),the reason why we have to minus the numbers of successor is that, the successor will be in office after the midterm election.
# They are not in office at the beginning of the term. If we want to count the number of lawmaker before midterm election(2018),
# We have to minus the numbers of vacate as there is still no successor after the election.
# https://en.wikipedia.org/wiki/116th_United_States_Congress
# In the table it showed 451 while Wiki shows total 434 in Jan. 3, 2019 and total 430 in Dec. 14, 2020
# https://en.wikipedia.org/wiki/Jeff_Van_Drew
# Van Drew was a Democrat in 2019
# Van Drew officially switched his party affiliation on January 7, 2020
# Paul changed his party R to I

#Be careful! house_116 has 3 + 1 variables to identify an excessive number of lawmakers. The plus 1 variable is change_party.
#You need to analyze what change_party's meaning is and try to use the data of change_party to create the table I ask.
```

#Question 2: (5 points)

#Create a table object house_116_2020 which shows 116th house of lawmakers in the end of the 116th term (Dec. 14, 2020).

```
house_116 <- read.csv("house_116.csv") # read csv file again to refresh
house_116 <- house_116[-c(7, 284, 417), ] #  I removed the same data that indicates older
update
nrow(house_116) # 448
house_116_2020 <- house_116[!(house_116$vacate == 1) & !(house_116$non_voting == 1),]
nrow(house_116_2020) # 430 which is what wiki shows
```

#Question 3: (5 points)

#Create a table object house_117_2021 which shows 117th house of lawmakers in the beginning of the 117th term (Jan. 3, 2021).

```
nrow(house_117) # 447

house_117_2021 <- house_117[!(house_117$successor == 1) & !(house_117$non_voting ==
1),] # According to the info above, I applied the same code.
nrow(house_117_2021) # 433

# https://en.wikipedia.org/wiki/117th_United_States_Congress
# In the table, it showed 447 while Wiki shows total 434 in Jan. 3, 2021 and total 433 in FEB.
17, 2022
```

#Question 4: (5 points)

#Create a table object house_117_2022 which shows 117th house of lawmakers in the end of this February.

```
house_117_2022 <- house_117[!(house_117$vacate == 1) & !(house_117$non_voting == 1),]
nrow(house_117_2022) # 433
```

```
#Question 5: (3 points)

#Using house_115_2016, house_116_2019, house_117_2021 tables to calculate every term's gender ratio.

a <- nrow(house_115_2016[house_115_2016$gender == 'F',]) # 83 female has in house_115_2016 and I assigned it as variable 'a'
b <- nrow(house_115_2016[house_115_2016$gender == 'M',]) # 352 male has in house_115_2016 and I assigned it as variable 'b'
ab <- a+b # Total 435
ab
a/ab # 0.1908046 => 19% is the ratio of female
b/ab # 0.8091954 => 81% is the ratio of male

c <- nrow(house_116_2019[house_116_2019$gender == 'F',]) # 102 female has in house_115_2016 and I assigned it as variable 'c'
d <- nrow(house_116_2019[house_116_2019$gender == 'M',]) # 332 male has in house_115_2016 and I assigned it as variable 'd'
cd <- c+d # Total 434
cd
c/cd # 0.235023 => 24% is the ratio of female
d/cd # 0.764977 => 76% is the ratio of male

e <- nrow(house_117_2021[house_117_2021$gender == 'F',]) # 118 female has in house_115_2016 and I assigned it as variable 'e'
f <- nrow(house_117_2021[house_117_2021$gender == 'M',]) # 315 male has in house_115_2016 and I assigned it as variable 'f'
ef <- e+f # Total 433
ef
e/ef # 0.2725173 => 27% is the ratio of female
f/ef # 0.7274827 => 73% is the ratio of male

#Question 6: (3 points)
#Using house_115_2016, house_116_2019, house_117_2021 tables to calculate every term's mean of age.
lawmaker_age_2016 <- substr(house_115_2016$brith, 1, 4) # extract the year of the birthday by using substr
lawmaker_age_2016 <- as.numeric(lawmaker_age_2016) # make it as numeric
age_lawmaker_2016 <- 2022 - lawmaker_age_2016 # get an age at the moment
mean(age_lawmaker_2016) # The average age is 63.55862

colnames(house_116_2019)
lawmaker_age_2019 <- substr(house_116_2019$date_of_birth, 1, 4) # extract the year of the birthday by using substr
```

```
lawmaker_age_2019 <- as.numeric(lawmaker_age_2019) # make it as numeric
age_lawmaker_2019 <- 2022 - lawmaker_age_2019 # get an age at the moment
mean(age_lawmaker_2019) # The average age is 60.96774

lawmaker_age_2021 <- substr(house_117_2021$date_of_birth, 1, 4) # extract the year of the birthday by using substr
lawmaker_age_2021 <- as.numeric(lawmaker_age_2021) # make it as numeric
age_lawmaker_2021 <- 2022 - lawmaker_age_2021 # get an age at the moment
mean(age_lawmaker_2021) # The average age is 59.67436

#Question 7: (3 points)

#Using house_115_2016, house_116_2019, house_117_2021 tables to calculate every term's partisan ratio.

# 3. which party dominate the house?
unique(house_115_2016$party) # check if there is an another party
Ip_2016_R <- nrow(house_115_2016[house_115_2016$party == 'R',]) # get the number of Republican in house_115_2016 and assign it as ip_2016_R
Ip_2016_D <- nrow(house_115_2016[house_115_2016$party == 'D',]) # get the number of Democrat in house_115_2016 and assign it as ip_2016_D
Ip_2016_Total <- Ip_2016_R+Ip_2016_D # Total 435
Ip_2016_R / Ip_2016_Total # 0.554023 => Republican lawmakers has the ratio of 55.40%
Ip_2016_D / Ip_2016_Total # 0.445977 => Democrat lawmaker has the ratio of 44.59%

unique(house_116_2019$party) # check if there is another party. There is I
Ip_2019_R <- nrow(house_116_2019[house_116_2019$party == 'R',]) # get the number of Republican in house_116_2019 and assign it as ip_2019_R
Ip_2019_R # 199
Ip_2019_D <- nrow(house_116_2019[house_116_2019$party == 'D',]) # get the number of Democrat in house_116_2019 and assign it as ip_2019_D
Ip_2019_Total <- Ip_2019_R+Ip_2019_D # Total 433
Ip_2019_R / Ip_2019_Total # 0.4585253 => Republican lawmakers has the ratio of 45.85%
Ip_2019_D / Ip_2019_Total # 0.5414747 => Democrat lawmaker has the ratio of 54.15%

unique(house_117_2021$party) # check if there is an another party
Ip_2021_R <- nrow(house_117_2021[house_117_2021$party == 'R',]) # get the number of Republican in house_117_2021 and assign it as ip_2021_R
Ip_2021_D <- nrow(house_117_2021[house_117_2021$party == 'D',]) # get the number of Democrat in house_117_2021 and assign it as ip_2021_D
Ip_2021_Total <- Ip_2021_R+Ip_2021_D # Total 433
Ip_2021_R / Ip_2021_Total # 0.4872979 => Republican lawmakers has the ratio of 48.73%
Ip_2021_D / Ip_2021_Total # 0.5127021 => Democrat lawmaker has the ratio of 51.27%
```