613 A1

Zihao Zhang, zz247

21/01/2022

```
library(xlsx)
library(dplyr)
library(tidyverse)
library(lubridate)
library(tidyr)
library(magrittr)
library(data.table)
library(gdata)

# Data
path <- "/Users/zhangzihao/Desktop/duke/613/HW/a1/Data"
filenames <- dir(path)
filepath <- sapply(filenames, function(x){
  paste(path,x,sep='/')})
data <- lapply(filepath, function(x){
  fread(x, colClasses=c(idind="character",idmen="character"), header=T)})

# Exercise 1

# 1
library(dplyr)
N_Household =
  data[["dathh2007.csv"]] %>%
  group_by(idmen) %>%
  count %>%
  ungroup
N_Household %>% nrow
#10498

# 2
data[["dathh2005.csv"]] %>% group_by(mstatus) %>% count #3374

# 3
data[["datind2008.csv"]] %>% group_by(idind) %>% nrow #25510

# 4
sum(between(data[["datind2016.csv"]]$age,25,35)) #2765
```

 # 5

```
table(data[["datind2009.csv"]][,c("profession","gender")])
```

```
          gender
profession Female Male
        0      11   19
        11     30   57
        12      8   19
        13     29   78
        21     63  213
        22     65  114
        23      8   48
        31     68   98
        33     85  107
        34    184  142
        35     50   59
```

 # 6

```
wage2005 <- data[["datind2005.csv"]]$wage
wage2019 <- data[["datind2019.csv"]]$wage
```

 # mean

```
mean(wage2005,na.rm = T) #11992.26
mean(wage2019,na.rm = T) #15350.47
```

 # sd

```
sd(wage2005,na.rm = T) #17318.56
sd(wage2019,na.rm = T) #23207.18
```

 # D9/D1

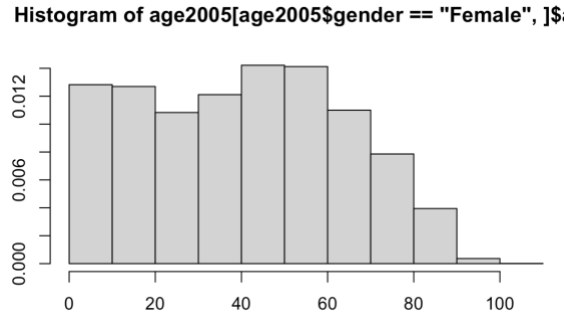
```
quantile(wage2005,0.1,na.rm = T) #0
quantile(wage2005,0.9,na.rm = T) #32340
quantile(wage2019,0.1,na.rm = T) #0
quantile(wage2019,0.9,na.rm = T) #40267
```

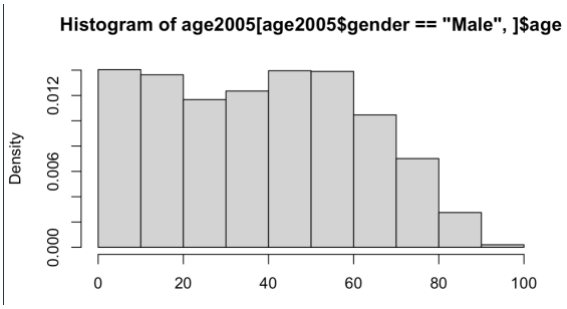
 # Gini coefficent

```
gini_wages2005 <- data.frame(data[["datind2005.csv"]][,10]) %>% na.omit() %>%
  arrange(wage) %>%
  mutate(R = rank(wage)/n()) %>%
  mutate(RI = cumsum(wage)/sum(wage)) %>%
  mutate(gini = sum(2*(R- RI)/n()))
gini2005 <- gini_wages2005 %>% select(gini) %>% distinct()
gini2005 #0.6671654

gini_wages2019 <- data.frame(data[["datind2019.csv"]][,10]) %>% na.omit() %>%
  arrange(wage) %>%
  mutate(R = rank(wage)/n()) %>%
  mutate(RI = cumsum(wage)/sum(wage)) %>%
  mutate(gini = sum(2*(R- RI)/n()))
gini2019 <- gini_wages2019 %>% select(gini) %>% distinct()
gini2019 #0.6655301
```

 # 7

```
hist(data[["datind2010.csv"]]$age, breaks = 12, freq = F)
age2005 <- data[["datind2010.csv"]][,c(8,9)]
hist(age2005[age2005$gender == "Male",]$age, breaks = 12, freq = F)
hist(age2005[age2005$gender == "Female",]$age, breaks = 12, freq = F)
```






```
# In general, males have a higher proportion of younger groups than females
```

 # 8

```
data2011 <- left_join(data[["datind2011.csv"]],data[["dathh2011.csv"]],by = c('year','idmen'))
sum(data2011$location == "Paris",na.rm = T) #3514
```

 # Exercise 2

 # 1

```
dathh <- data.frame()
for(i in 1:16){
   dathh <- rbind(dathh,data[[i]])
}
```

 # 2

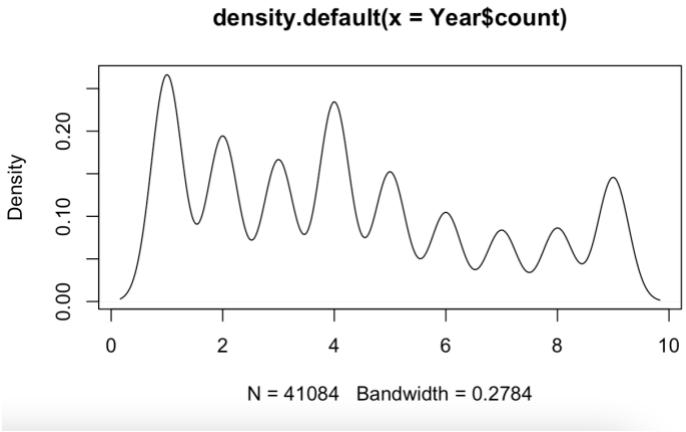
```
datind <- data.frame()
for(i in 17:32){
   datind <- rbind(datind,data[[i]])
}
```

```
# 3
variables.dathh <- colnames(dathh)
variables.datind <- colnames(datind)
variables.dathh <- variables.dathh[-1]
variables.datind <- variables.datind[-1]
intersect(variables.dathh,variables.datind) #"idmen" "year"

# 4
data2 <- left_join(datind[,-1], dathh[,-1], by = c('year','idmen'))

# 5
datind %>% group_by(idmen, year) %>% summarize(count = n()) %>%
   filter(count>4) %>% nrow() #12436

# 6
datind %>% group_by(idmen,empstat) %>% summarize(count = n()) %>%
   filter(empstat == "Unemployed", count>0) %>% nrow() #8162

# 7
datind[which(datind$profession>0),] %>%
   group_by(idmen,year,profession) %>% summarize(count = n()) %>%
   filter(count>1) %>% nrow() # 7615

# 8
data2 %>% filter(mstatus == "Couple, with Kids") %>% nrow() #209382

# 9
data2 %>% filter(location == "Paris") %>% nrow() #51904

# 10
biggest_family <- data2 %>% group_by(idmen,year) %>% summarise(count=n()) %>% arrange(desc(count))
biggest_family[1,1] #2207811124040100

# 11
n_2010 <- data2 %>% filter(year == 2010)
length(unique(n_2010$idmen)) #11050
n_2011 <- data2 %>% filter(year == 2011)
length(unique(n_2011$idmen)) #11360
length(unique(n_2011$idmen)) + length(unique(n_2010$idmen)) #22410
```

Exercise 3

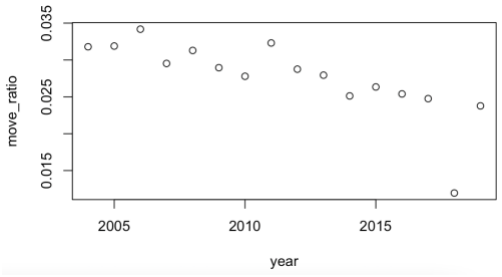
1

```
dathh %>% group_by(idmen,year) %>% arrange(idmen)
Year <- dathh %>% group_by(idmen) %>% summarise(count = n())
Year <- Year[,2]
summary(Year)
plot(density(Year$count))
```



2

```
dathh$move_in = dathh$year - dathh$datent == 0
head(dathh$move_in,10)

sameyear <- data_frame()
function_sameyear <- function(x){
   number <- dathh %>% filter(year == x) %>% filter(datent == year) %>% summarise(count=n())
   total <- dathh %>% filter(year == x) %>% nrow()
   return(number/total)
}
for (i in 2004:2019){
   sameyear[c(i-2003),1] = c(i)
   sameyear[c(i-2003),2] = as.numeric(function_sameyear(i))
}
colnames(sameyear)
sameyear <- rename(sameyear, year = ...1 , move_ratio = ...2)
plot(sameyear, xlab = "year", ylab = "move_ratio")
```



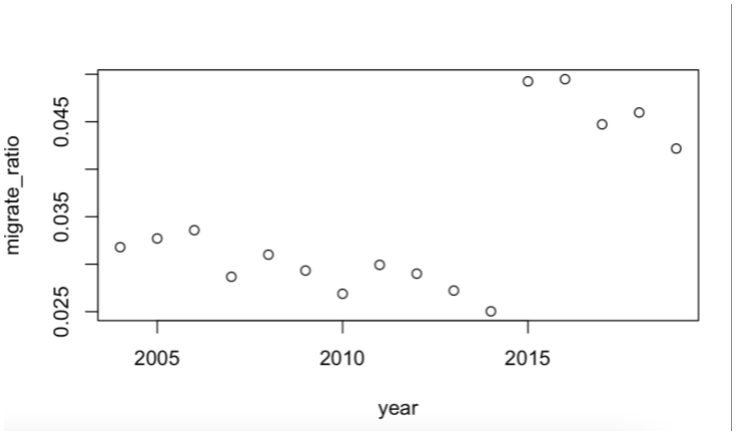
 # 3

 # before 2014

```
dathh %>% filter(myear == year) %>% head(10)
myear <- data_frame()
function_myear <- function(x){
   number <- dathh %>% filter(year == x) %>% filter(myear == year) %>% summarise(count=n())
   total <- dathh %>% filter(year == x) %>% nrow()
   return(number/total)
}
for (i in 2004:2014){
   myear[c(i-2003),1] = c(i)
   myear[c(i-2003),2] = as.numeric(function_myear(i))
}
```

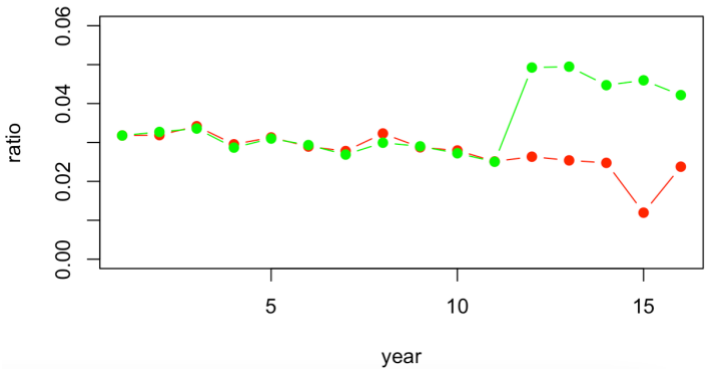
 # after 2014

```
dathh %>% filter(move == 2) %>% head(10)
dathh %>% filter(year >= 2015)
function_move <- function(x){
   number <- dathh %>% filter(year == x) %>% filter(move == 2) %>% summarise(count=n())
   total <- dathh %>% filter(year == x) %>% nrow()
   return(number/total)
}
for (i in 2015:2019){
   myear[c(i-2003),1] = c(i)
   myear[c(i-2003),2] = as.numeric(function_move(i))
}
colnames(myear)
myear <- rename(myear, year = ...1 , migrate_ratio = ...2)
plot(myear, xlab = "year", ylab = "migrate_ratio")
```



 # 4

```
plot_data <- left_join(myear,sameyear, by = c("year"))
plot(plot_data$move_ratio, type = "b", pch = 19, col = "red", xlab = "year", ylab = "ratio", ylim=c(0,0.06))
lines(plot_data$migrate_ratio, type = "b", pch = 19, col = "green")
```



I prefer the first one, the second one is not clear considering the data in 2015. In contrast, the first statistical method is more continuous

 # 5

```
migrate_before2014 <- data2 %>% filter(myear==year)
migrate_after2014 <- data2 %>% filter(move==2)
migrate <- rbind(migrate_before2014,migrate_after2014) %>% select(idmen,idind,year,empstat,profession)
migrate <- migrate %>% group_by(idind) %>% mutate(empstat_change=length(unique(empstat)) >= 2)
migrate <- migrate %>% group_by(idind) %>% mutate(profession_change=length(unique(profession)) >= 2)
migrate %>% filter(profession_change==T|empstat_change==T) %>% nrow()
#1407
```

 # Exercise 4

```
Attrition <- migrate %>% group_by(idind) %>% arrange(year) %>%
summarise(entry=head(year,1),exit=tail(year,1))
# the lists of "entry" and "exit" mean the years that specific individual entered or exited the penal
# if one individual exits in 2017, it is "attrition" in 2018, which means he won't appear in 2018.
attrition_data <- left_join(migrate,attrition,by = c("idind")) %>% select(idmen,idind,year,entry,exit)
attrition_data
attrition_data[,5] <- as.numeric(unlist(attrition_data[,5]))
attrition_data[,3] <- as.numeric(unlist(attrition_data[,3]))
num_attrition <- data_frame()
function_attrition <- function(x){
   num <- attrition_data %>% filter(year == x) %>% filter(exit == year) %>% nrow() %>% as.numeric()
   tot <- attrition_data %>% filter(year == x) %>% nrow() %>% as.numeric()
   return(num/tot)
}
num_attrition <- list()
for (i in 2004:2019){
```

```
    num_attrition$ratio[c(i-2003)] = as.numeric(function_attrition(i))
}
num_attrition <- data.frame(num_attrition)
num_attrition$year <- c(2004:2019)
num_attrition$year <- num_attrition[,1]
num_attrition <- rename(num_attrition, percentage = year, time = ratio)
num_attrition$time <- c(2004:2019)
plot(num_attrition)
# it seems that not much person will stay in the penal for a long time
```