```
#ECON613 A1 Mengzhi Chen
#================================================

#Exercise1 Basic Statistics
#================================================

#Set working directory
setwd("/Users/halcyonchan/Desktop/Econ613/A1/Data")
install.packages("data.table")
library(data.table)

#1.1
#Number of households surveyed in 2007
#Read the file using fread
dathh2007 = fread ("dathh2007.csv")
print(dathh2007)
#The number of X represents the number of household surveyed
max(dathh2007$V1) #10498

#1.2
#Number of households with a marital status “Couple with kids” in 2005
dathh2005 = fread("dathh2005.csv")
print(dathh2005)
#Filter the households that are Couple, with Kids
dathh2005_Couple = dathh2005[which(dathh2005$mstatus=="Couple, with Kids"),]
#Count the number
length(dathh2005_Couple$idmen) #3374

#1.3
#Number of individuals surveyed in 2008
datind2008 = fread("datind2008.csv")
print(datind2008)
#The number of X represents the number of individual surveyed
length(datind2008$idind) #25510

#1.4
#Number of individuals aged between 25 and 35 in 2016
datind2016 = read.csv("datind2016.csv")
print(datind2016)
#Which() shows the condition and Length() compute the number that satisfy the condition
length(which(datind2016$age<=35 & datind2016$age>=25)) #2765

#1.5
```

#Cross-table gender/profession in 2009

```
datind2009 = read.csv("datind2009.csv")
print(datind2009)
CrossTable = table(datind2009$gender,datind2009$profession)
CrossTable
```

```
> CrossTable

          0  11  12  13  21  22  23  31  33  34  35  37  38  42  43  44  45  46  47  48  52  53  54  55  56  62  63  64  65  67  68
Female   11  30   8  29  63  65   8  68  85 184  50 179  78 258 437   1 153 410  82  22 782  27 584 353 696  64  35  29  19 147 120
Male     19  57  19  78 213 114  48  98 107 142  59 260 368 110 117   2  95 340 429 215 169 182  98 101  74 443 520 246 159 237 177

         69
Female   40
Male     82
```

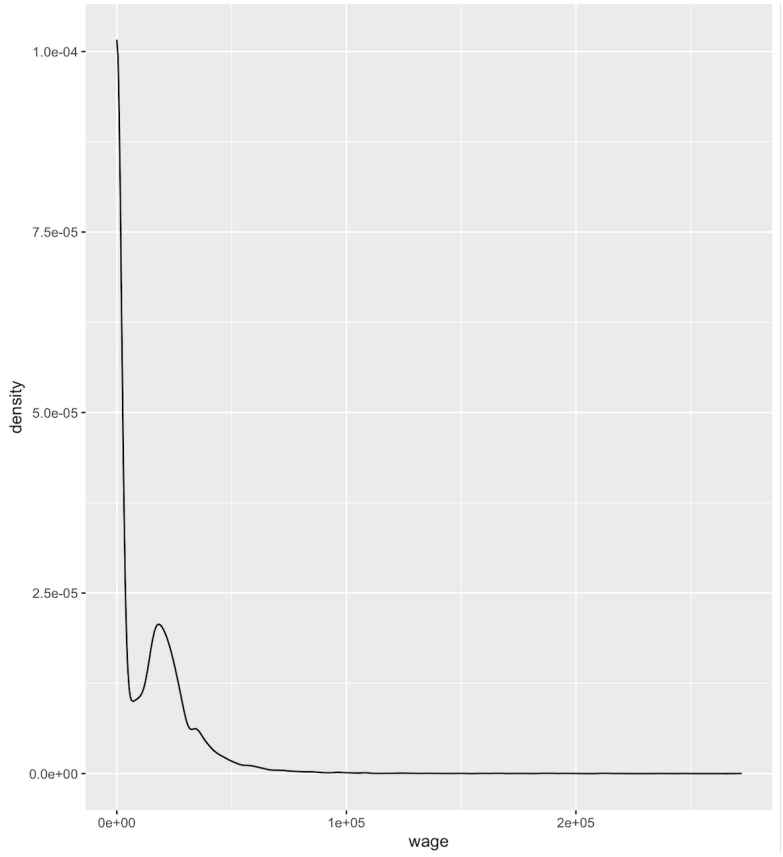
#1.6

#Distribution of wages in 2005 and 2019. Report the mean, the standard deviation,

#the inter-decile ratio D9/D1 and the Gini coefficient

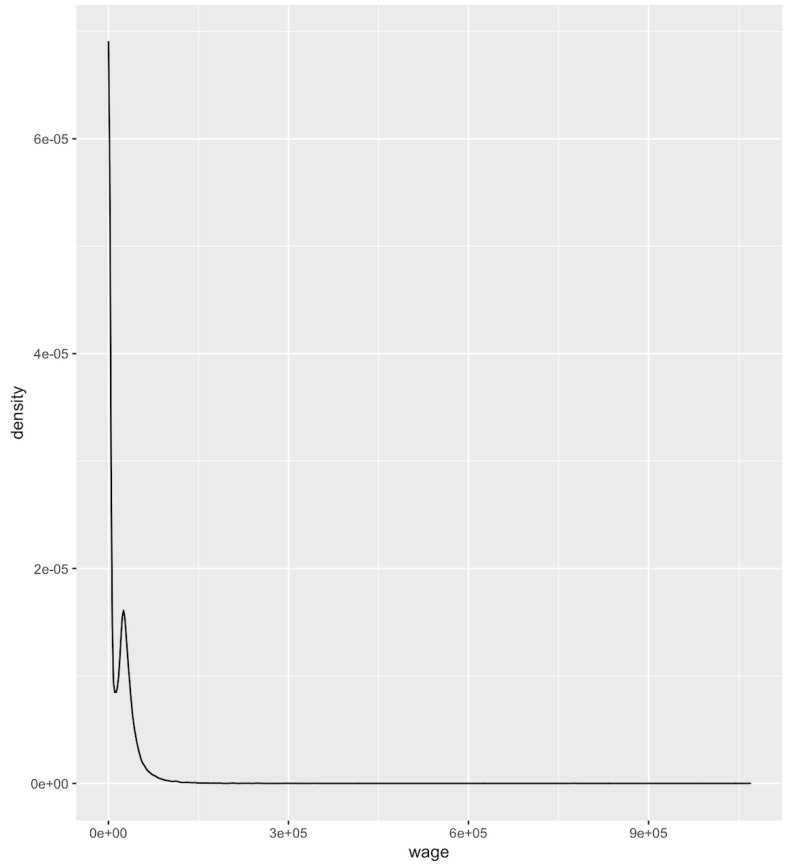
```
datind2005 = fread ("datind2005.csv")
datind2019 = fread("datind2019.csv")
#Dist of wage
install.packages("ggplot2")
library(ggplot2)
ggplot(datind2005,aes(x=wage),na.rm=TRUE)+geom_density(color="black")
```

#Below is the distribution of wage in 2005.



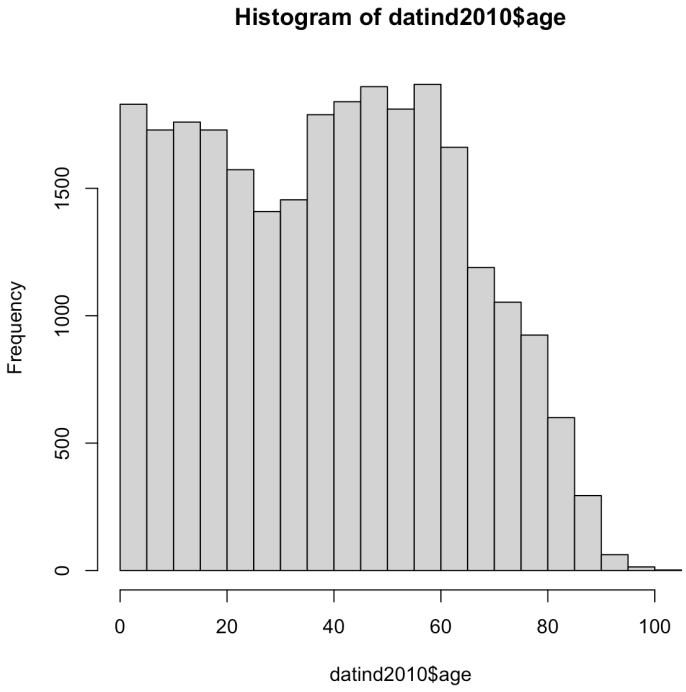

```
ggplot(datind2019,aes(x=wage),na.rm=TRUE)+geom_density(color="black")
```

#Below is the distribution of wage in 2019

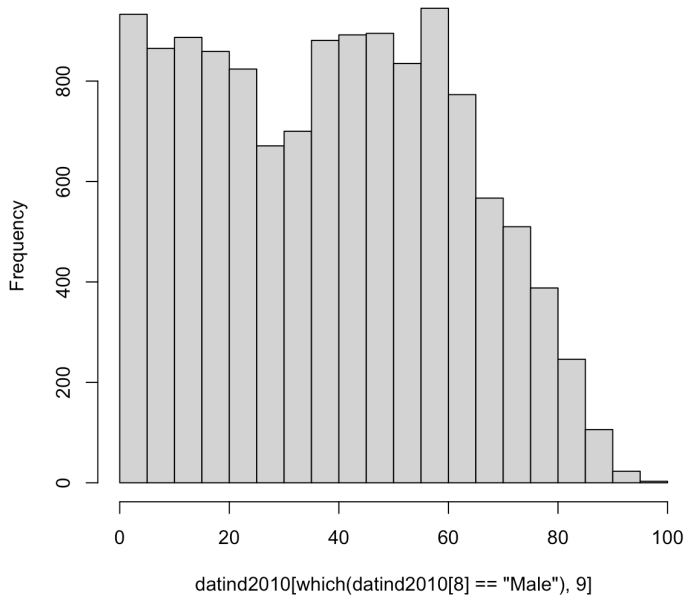



```
#Report statistical indicators
#First, write the function of Gini
install.packages("tidyverse")
library(tidyverse)
wage2005 = datind2005%>%drop_na(wage)
w = sum(wage2005$wage/18767)
x = wage2005$wage
y = t(wage2005$wage)
Gini2005 = 1/(2*w*18767^2)*sum(abs(outer(x,y,FUN="-")))

wage2019 = datind2019%>%drop_na(wage)
v = sum(wage2019$wage/21421)
m = wage2019$wage
n = t(wage2019$wage)
Gini2019 = 1/(2*v*21421^2)*sum(abs(outer(m,n,FUN="-")))

#Construct a function to output mean ,sd and dec9/dec1
dsummary = function(datind) {
   mean = summary(datind)[[4]]
   sd = sd(datind,na.rm=TRUE)
   dec1 = quantile(datind,prob=c(0.1,0.9),na.rm=TRUE)[[1]]
   dec9 = quantile(datind,prob=c(0.1,0.9),na.rm=TRUE)[[2]]
   D9D1 = dec9/dec1
```

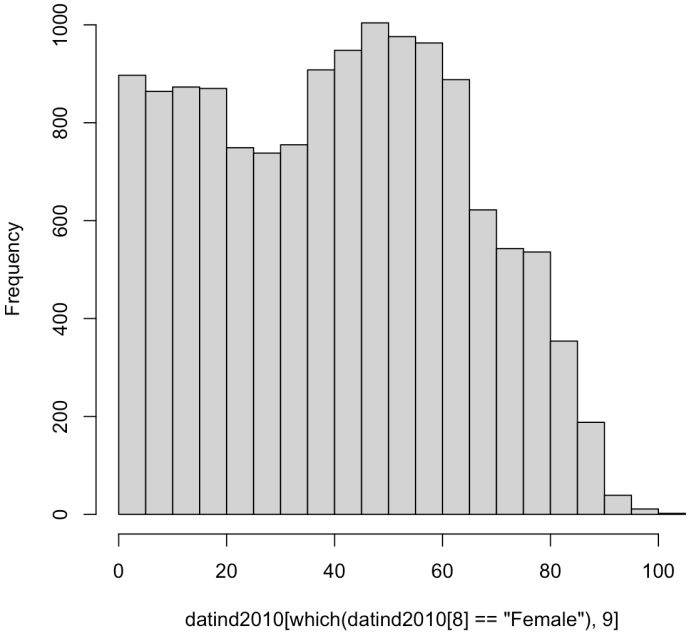
```
  return(c(mean,sd,D9D1))
}
#The answers are below
dsummary(datind2005$wage)
mean #11992.26
sd #17318.56
D9D1 #inf
Gini2005 #0.6671654

dsummary(datind2019$wage)
mean #15350.47
sd #23207.18
D9D1 #inf
Gini2019 #0.6655301

#1.7
#Distribution of age in 2010. Plot an histogram. Is there any difference between men and women?
datind2010 = fread("datind2010.csv")
print(datind2010)
hist(datind2010$age)
#Plat the histogram of age in 2010
```




```
male = which(datind2010[8]=="Male")
#Plot age of male and female separately to see the difference We can see that their distributions are similar.
hist(datind2010[which(datind2010$gender=="Male"),age])
```




```
hist(datind2010[which(datind2010$gender=="Female"),age])
```




```
#1.8
#Number of individuals in Paris in 2011
dathh2011 = fread("dathh2011.csv")
datind2011 = fread("datind2011.csv")
dat2011 = merge(datind2011,dathh2011,by =c("idmen"))
ind_paris = dat2011[which(dat2011$location == "Paris"),]
length(unique(ind_paris$idind)) #3514
```

#Exercise2 Merge Datasets

#==========

#2.1

#Read all individual datasets from 2004 to 2019. Append all these datasets

```
datind2004 = fread("datind2004.csv",colClasses=c(idmen = "character",idind = "character"))
datind2005 = fread("datind2005.csv",colClasses=c(idmen = "character",idind = "character"))
datind2006 = fread("datind2006.csv",colClasses=c(idmen = "character",idind = "character"))
datind2007 = fread("datind2007.csv",colClasses=c(idmen = "character",idind = "character"))
datind2008 = fread("datind2008.csv",colClasses=c(idmen = "character",idind = "character"))
datind2009 = fread("datind2009.csv",colClasses=c(idmen = "character",idind = "character"))
datind2010 = fread("datind2010.csv",colClasses=c(idmen = "character",idind = "character"))
datind2011 = fread("datind2011.csv",colClasses=c(idmen = "character",idind = "character"))
datind2012 = fread("datind2012.csv",colClasses=c(idmen = "character",idind = "character"))
datind2013 = fread("datind2013.csv",colClasses=c(idmen = "character",idind = "character"))
datind2014 = fread("datind2014.csv",colClasses=c(idmen = "character",idind = "character"))
datind2015 = fread("datind2015.csv",colClasses=c(idmen = "character",idind = "character"))
datind2016 = fread("datind2016.csv",colClasses=c(idmen = "character",idind = "character"))
datind2017 = fread("datind2017.csv",colClasses=c(idmen = "character",idind = "character"))
datind2018 = fread("datind2018.csv",colClasses=c(idmen = "character",idind = "character"))
datind2019 = fread("datind2019.csv",colClasses=c(idmen = "character",idind = "character"))
# Use rbind() to append all these datasets
datind=rbind(datind2004,datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,datind2011,datind2012,datind2013,datind2014,datind2015,datind2016,datind2017,datind2018,datind2019)
```

Filter

	V1	idind	idmen	year	empstat	respondent	profession	gender	age	wage
1	1	1120001001293010001	1200010012930100	2004	Employed	1	67	Male	31	19187
2	2	1120001004058010001	1200010040580100	2004	Employed	1	56	Female	30	11586
3	3	1120001004058010002	1200010040580100	2004	Inactive	0		Female	9	NA
4	4	1120001006663010001	1200010066630100	2004	Employed	1	38	Male	31	44656
5	5	1120001006663010002	1200010066630100	2004	Employed	0	45	Female	27	20413
6	6	1120001008245010001	1200010082450100	2004	Retired	1		Female	89	0
7	7	1120001008644010001	1200010086440100	2004	Employed	1	34	Male	36	30702
8	8	1120001008644010002	1200010086440100	2004	Employed	0	42	Female	34	24650
9	9	1120001010299010001	1200010102990100	2004	Employed	1	46	Female	40	29604
10	10	1120001010299010002	1200010102990100	2004	Inactive	0		Female	15	NA
11	11	1120001011845010001	1200010118450100	2004	Employed	1	37	Male	54	39851
12	12	1120001011845010002	1200010118450100	2004	Employed	0	54	Female	54	20422
13	13	1120002001293010001	1200020012930100	2004	Employed	1	11	Male	56	0
14	14	1120002001293010002	1200020012930100	2004	Employed	0	11	Female	51	0
15	15	1120002001293010003	1200020012930100	2004	Retired	0		Female	81	0
16	16	1120002001293010004	1200020012930100	2004	Employed	0	63	Male	49	19372
17	17	1120002001739010001	1200020017390100	2004	Employed	1	11	Male	51	0

Showing 1 to 18 of 413,504 entries, 10 total columns

#2.2

#Read all household datasets from 2004 to 2019. Append all these datasets

```
dathh2004 = fread("dathh2004.csv",colClasses=c(idmen = "character"))
```

```
dathh2005 = fread("dathh2005.csv",colClasses=c(idmen = "character"))
dathh2006 = fread("dathh2006.csv",colClasses=c(idmen = "character"))
dathh2007 = fread("dathh2007.csv",colClasses=c(idmen = "character"))
dathh2008 = fread("dathh2008.csv",colClasses=c(idmen = "character"))
dathh2009 = fread("dathh2009.csv",colClasses=c(idmen = "character"))
dathh2010 = fread("dathh2010.csv",colClasses=c(idmen = "character"))
dathh2011 = fread("dathh2011.csv",colClasses=c(idmen = "character"))
dathh2012 = fread("dathh2012.csv",colClasses=c(idmen = "character"))
dathh2013 = fread("dathh2013.csv",colClasses=c(idmen = "character"))
dathh2014 = fread("dathh2014.csv",colClasses=c(idmen = "character"))
dathh2015 = fread("dathh2015.csv",colClasses=c(idmen = "character"))
dathh2016 = fread("dathh2016.csv",colClasses=c(idmen = "character"))
dathh2017 = fread("dathh2017.csv",colClasses=c(idmen = "character"))
dathh2018 = fread("dathh2018.csv",colClasses=c(idmen = "character"))
dathh2019 = fread("dathh2019.csv",colClasses=c(idmen = "character"))
# Use rbind() to append all these datasets
dathh=rbind(dathh2004,dathh2005,dathh2006,dathh2007,dathh2008,dathh2009,dathh2010,dathh2011,dathh2012,dathh2013,dathh2014,dathh2015,dathh2016,dathh2017,dathh2018,dathh2019)
```

	V1	idmen	year	datent	myear	mstatus	move	location
1	1	1200010012930100	2004	2000	2000	Single	NA	Paris
2	2	1200010040580100	2004	2001	2001	Single Parent	NA	Paris
3	3	1200010066630100	2004	2000	2000	Couple, No kids	NA	Paris
4	4	1200010082450100	2004	1957	1957	Single	NA	Paris
5	5	1200010086440100	2004	2001	2001	Couple, No kids	NA	Paris
6	6	1200010102990100	2004	1990	1990	Single Parent	NA	Paris
7	7	1200010118450100	2004	2000	2000	Couple, No kids	NA	Paris
8	8	1200020012930100	2004	1948	1988	Other	NA	Rural
9	9	1200020017390100	2004	1979	1979	Single	NA	Rural
10	10	1200020026420100	2004	1984	1981	Other	NA	Rural
11	11	1200020045130100	2004	2001	2001	Single Parent	NA	Urban 10000 to 19999
12	12	1200020094370100	2004	1998	1998	Couple, with Kids	NA	Urban 50000 to 99999
13	13	1200020118450100	2004	1925	1973	Single	NA	Rural
14	14	1200020122680100	2004	2002	2002	Couple, with Kids	NA	Urban 10000 to 19999
15	15	1200149012930100	2004	1993	1992	Couple, with Kids	NA	Rural
16	16	1200149034710100	2004	1971	1968	Single	NA	Rural
17	17	1200149057530100	2004	1976	1996	Couple, No kids	NA	Rural

Showing 1 to 18 of 173,851 entries, 8 total columns

#2.3

#List the variables that are simultaneously present in the individual and household datasets

ls(dathh)

ls(datind)

```
> ls(dathh)
[1] "datent"   "idmen"    "location" "move"     "mstatus"  "myear"    "X"        "year"
> ls(datind)
 [1] "age"        "empstat"    "gender"     "idind"      "idmen"      "profession" "respondent" "V1"         "wage"       "year"
```

#The variables that are simultaneously present are "idmen","X" and "year"

#2.4

#Merge the appended individual and household datasets

#group by idmen and year

```
dat = merge(datind,dathh,by =c("idmen","year"))
```

dat

Filter

	idmen	year	X.x	idind	empstat	respondent	profession	gender	age	wage	X.y	datent	myear	mstatus	move	location
1	1200010012930100	2004	1	1120001001293010048	Employed	1	67	Male	31	19187	1	2000	2000	Single	NA	Paris
2	1200010040580100	2004	2	1120001004058009984	Employed	1	56	Female	30	11586	2	2001	2001	Single Parent	NA	Paris
3	1200010040580100	2004	3	1120001004058009984	Inactive	0		Female	9	NA	2	2001	2001	Single Parent	NA	Paris
4	1200010040580100	2005	1	1120001004058009984	Inactive	1		Female	31	12334	1	2001	2001	Single Parent	NA	Paris
5	1200010040580100	2005	2	1120001004058009984	Inactive	0		Female	10	NA	1	2001	2001	Single Parent	NA	Paris
6	1200010066630100	2004	4	1120001006663010048	Employed	1	38	Male	31	44656	3	2000	2000	Couple, No kids	NA	Paris
7	1200010066630100	2004	5	1120001006663010048	Employed	0	45	Female	27	20413	3	2000	2000	Couple, No kids	NA	Paris
8	1200010066630100	2005	4	1120001006663010048	Employed	0	45	Female	28	19231	2	2005	2005	Couple, No kids	NA	Paris
9	1200010066630100	2005	3	1120001006663010048	Employed	1	38	Male	32	50659	2	2005	2005	Couple, No kids	NA	Paris
10	1200010082450100	2004	6	1120001008245010048	Retired	1		Female	89	0	4	1957	1957	Single	NA	Paris
11	1200010082450100	2005	5	1120001008245010048	Retired	1		Female	90	0	3	1957	1957	Single	NA	Paris
12	1200010086440100	2004	7	1120001008644009984	Employed	1	34	Male	36	30702	5	2001	2001	Couple, No kids	NA	Paris
13	1200010086440100	2004	8	1120001008644009984	Employed	0	42	Female	34	24650	5	2001	2001	Couple, No kids	NA	Paris
14	1200010086440100	2005	6	1120001008644009984	Employed	1	34	Male	37	31511	4	2001	2001	Couple, No kids	NA	Paris
15	1200010086440100	2005	7	1120001008644009984	Employed	0	42	Female	35	24873	4	2001	2001	Couple, No kids	NA	Paris
16	1200010102990100	2004	10	1120001010299010048	Inactive	0		Female	15	NA	6	1990	1990	Single Parent	NA	Paris
17	1200010102990100	2004	9	1120001010299010048	Employed	1	46	Female	40	29604	6	1990	1990	Single Parent	NA	Paris
18	1200010102990100	2005	9	1120001010299010048	Inactive	0		Female	16	0	5	1990	1990	Single Parent	NA	Paris
19	1200010102990100	2005	8	1120001010299010048	Employed	1	55	Female	41	30080	5	1990	1990	Single Parent	NA	Paris
20	1200010118450100	2004	11	1120001011845010048	Employed	1	37	Male	54	39851	7	2000	2000	Couple, No kids	NA	Paris
21	1200010118450100	2004	12	1120001011845010048	Employed	0	54	Female	54	20422	7	2000	2000	Couple, No kids	NA	Paris
22	1200010118450100	2005	10	1120001011845010048	Employed	1	37	Male	55	43296	6	2000	2000	Couple, No kids	NA	Paris
23	1200010118450100	2005	11	1120001011845010048	Employed	0	54	Female	55	20426	6	2000	2000	Couple, No kids	NA	Paris
24	1200020012930100	2004	14	1120002001293010048	Employed	0	11	Female	51	0	8	1948	1988	Other	NA	Rural
25	1200020012930100	2004	13	1120002001293010048	Employed	1	11	Male	56	0	8	1948	1988	Other	NA	Rural
26	1200020012930100	2004	15	1120002001293010048	Retired	0		Female	81	0	8	1948	1988	Other	NA	Rural
27	1200020012930100	2004	16	1120002001293010048	Employed	0	63	Male	49	19372	8	1948	1988	Other	NA	Rural
28	1200020012930100	2005	13	1120002001293010048	Employed	0	11	Female	52	0	7	1948	1988	Other	NA	Rural
29	1200020012930100	2005	12	1120002001293010048	Employed	1	11	Male	57	0	7	1948	1988	Other	NA	Rural
30	1200020012930100	2005	14	1120002001293010048	Retired	0		Female	82	NA	7	1948	1988	Other	NA	Rural

Showing 1 to 30 of 413,501 entries, 16 total columns

#2.5

#In the second part, we use the newly created dataset from the previous to answer the following questions:

#Number of households in which there are more than four family members

```
count = matrix(rep(1,413501))
#Add a column of 1's to dat
dat_count = cbind(dat,count)
#Sum up the number of individuals with same year, idmen
#And find out which household has more than four members
dat_count1 = aggregate(x=dat_count[c('count')],by=list(dat$year,dat$idmen),FUN=sum)
count2004 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2004"),]
count2005 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2005"),]
count2006 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2006"),]
count2007 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2007"),]
count2008 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2008"),]
count2009 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2009"),]
count2010 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2010"),]
count2011 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2011"),]
count2012 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2012"),]
```

```
count2013 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2013"),]
count2014 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2014"),]
count2015 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2015"),]
count2016 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2016"),]
count2017 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2017"),]
count2018 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2018"),]
count2019 = dat_count[which(dat_count1$count>4 & dat_count1$Group.1=="2019"),]
#combine the data of each year
count_all =
rbind(count2004,count2005,count2006,count2007,count2008,count2009,count2010,count2011,count2012,count201
3,count2014,count2015,count2016,count2017,count2018,count2019)
#remove the repeat data and count the number
length(unique(count_all$idmen)) #3734

#2.6
#Number of households in which at least one member is unemployed
#Sum up the number of individuals with same year, idmen and empstat
dat_count2 =
aggregate(x=dat_count[c('count')],by=list(dat_count$year,dat_count$idmen,dat_count$empstat),FUN=sum)
#And then count the number of households in which at least one member is unemployed
unemp2004 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2004"),]
unemp2005 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2005"),]
unemp2006 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2006"),]
unemp2007 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2007"),]
unemp2008 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2008"),]
unemp2009 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2009"),]
unemp2010 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2010"),]
unemp2011 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2011"),]
unemp2012 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2012"),]
unemp2013 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2013"),]
unemp2014 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2014"),]
unemp2015 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2015"),]
unemp2016 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2016"),]
unemp2017 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2017"),]
unemp2018 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2018"),]
unemp2019 = dat_count2[which(dat_count2$Group.3=="Unemployed" & dat_count2$Group.1=="2019"),]
#combine the data of each year
unemp_all =
rbind(unemp2004,unemp2005,unemp2006,unemp2007,unemp2008,unemp2009,unemp2010,unemp2011,unemp201
2,unemp2013,unemp2014,unemp2015,unemp2016,unemp2017,unemp2018,unemp2019)
#remove the repeat data and count the number
length(unique(unemp_all$Group.2)) #8161

#2.7
```

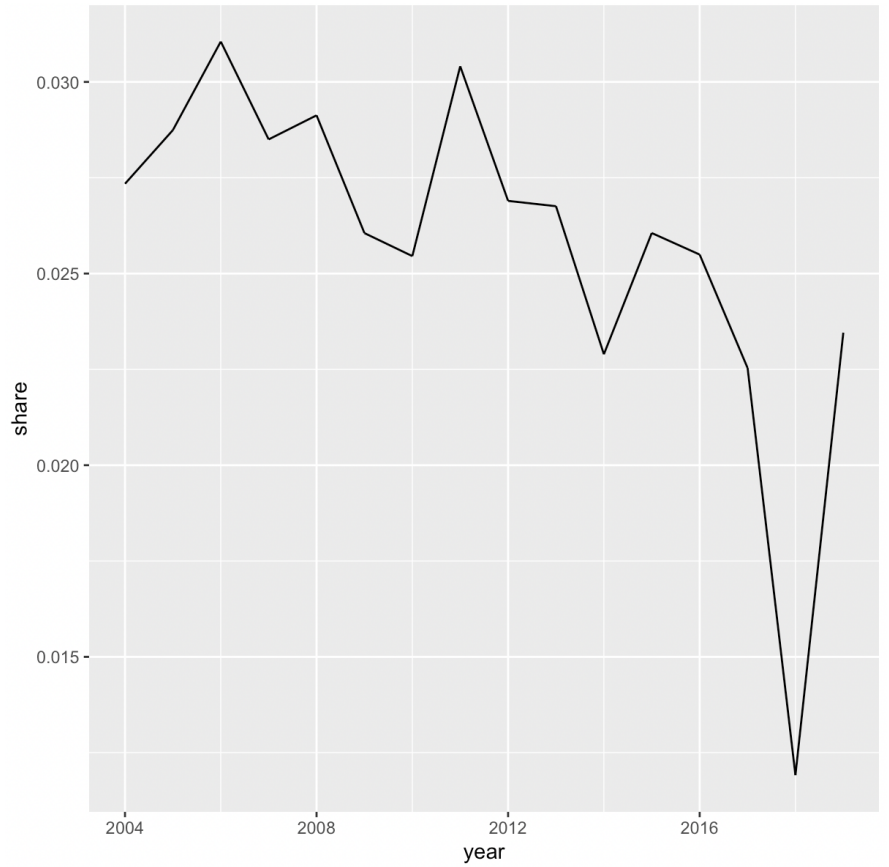
```
#Number of households in which at least two members are of the same profession
#Sum up the number of individuals with same year, idmen and profession
dat_count3 =
aggregate(x=dat_count[c('count')],by=list(dat_count$year,dat_count$idmen,dat_count$profession),FUN=sum)
#And then count the number of households that in which at least two members are of the same profession
prof2004 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2004"),]
prof2005 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2005"),]
prof2006 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2006"),]
prof2007 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2007"),]
prof2008 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2008"),]
prof2009 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2009"),]
prof2010 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2010"),]
prof2011 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2011"),]
prof2012 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2012"),]
prof2013 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2013"),]
prof2014 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2014"),]
prof2015 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2015"),]
prof2016 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2016"),]
prof2017 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2017"),]
prof2018 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2018"),]
prof2019 = dat_count3[which(dat_count3$count>=2 & dat_count3$Group.1=="2019"),]
#combine the data of each year
prof_all =
rbind(prof2004,prof2005,prof2006,prof2007,prof2008,prof2009,prof2010,prof2011,prof2012,prof2013,prof2014,pr
of2015,prof2016,prof2017,prof2018,prof2019)
#remove the repeat data and count the number
length(unique(prof_all$Group.2)) #8752

#2.8
#Number of individuals in the panel that are from household-Couple with kids
#Sum up the number of individuals with same year, idind and mstatus
dat_count4 =
aggregate(x=dat_count[c('count')],by=list(dat_count$year,dat_count$idind,dat_count$mstatus),FUN=sum)
#And then count the number of individuals that are from household-Couple with kids in each year
Couple2004 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2004"),]
Couple2005 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2005"),]
Couple2006 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2006"),]
Couple2007 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2007"),]
Couple2008 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2008"),]
Couple2009 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2009"),]
Couple2010 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2010"),]
Couple2011 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2011"),]
Couple2012 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2012"),]
Couple2013 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2013"),]
```

```
Couple2014 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2014"),]
Couple2015 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2015"),]
Couple2016 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2016"),]
Couple2017 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2017"),]
Couple2018 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2018"),]
Couple2019 = dat_count4[which(dat_count4$Group.3=="Couple, with Kids" & dat_count4$Group.1=="2019"),]
#combine the data of each year
Couple_all =
rbind(Couple2004,Couple2005,Couple2006,Couple2007,Couple2008,Couple2009,Couple2010,Couple2011,Couple
2012,Couple2013,Couple2014,Couple2015,Couple2016,Couple2017,Couple2018,Couple2019)
#remove the repeat data and count the number
length(unique(Couple_all$Group.2)) #15567

#2.9
#Number of individuals in the panel that are from Paris
#Sum up the number of individuals with same year, idind and locaiton
dat_count5 =
aggregate(x=dat_count[c('count')],by=list(dat_count$year,dat_count$idind,dat_count$location),FUN=sum)
#And then count the number of individuals that are from Paris in each year
Paris2004 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2004"),]
Paris2005 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2005"),]
Paris2006 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2006"),]
Paris2007 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2007"),]
Paris2008 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2008"),]
Paris2009 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2009"),]
Paris2010 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2010"),]
Paris2011 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2011"),]
Paris2012 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2012"),]
Paris2013 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2013"),]
Paris2014 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2014"),]
Paris2015 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2015"),]
Paris2016 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2016"),]
Paris2017 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2017"),]
Paris2018 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2018"),]
Paris2019 = dat_count5[which(dat_count5$Group.3=="Paris" & dat_count5$Group.1=="2019"),]
#combine the data of each year
Paris_all =
rbind(Paris2004,Paris2005,Paris2006,Paris2007,Paris2008,Paris2009,Paris2010,Paris2011,Paris2012,Paris2013,Par
is2014,Paris2015,Paris2016,Paris2017,Paris2018,Paris2019)
#remove the repeat data and count the number
length(unique(Paris_all$Group.2)) #6177

#2.10
#Find the household with the most number of family members. Report its idmen
```

```
#Find the maximum number of family numbers
max(dat_count1$count) #14
#Find the according idmen use the condition of "14" family members
dat_count1[which(dat_count1$count=="14"),2]
#Report idmen: 2207811124040100 2510263102990100

#2.11
#Number of households present in 2010 and 2011
dathh_num = rbind(dathh2010,dathh2011)
#Use Unique() to remove the repeat idmen
length(unique(dathh_num$idmen)) #13424


#Exercise3 Migration
#================================================

#3.1
#Find out the year each household enters and exit the panel.
#First, find out the year enters and exit separately
dat_min = aggregate(x=dat_count[c('year')],by=list(dat_count$idmen),FUN=min)
dat_max = aggregate(x=dat_count[c('year')],by=list(dat_count$idmen),FUN=max)
#Merge them into one dataset
dat_length = merge(dat_min,dat_max,by="Group.1")
#Report the length of years each household stay in the panel.
#The 4th column “length_year” in dat_length shows the length of years each household stay in the panel
dat_length[,4]=dat_length[,3]-dat_length[,2]
#rename the columns
colnames(dat_length)[1]="idmen"
colnames(dat_length)[2]="enter_year"
colnames(dat_length)[3]="exit_year"
colnames(dat_length)[4]="length_year"
```

	idmen	enter_year	exit_year	length_year
1	1200010012930100	2004	2004	0
2	1200010040580100	2004	2005	1
3	1200010066630100	2004	2005	1
4	1200010082450100	2004	2005	1
5	1200010086440100	2004	2005	1
6	1200010102990100	2004	2005	1
7	1200010118450100	2004	2005	1
8	1200020012930100	2004	2005	1
9	1200020017390100	2004	2005	1
10	1200020026420100	2004	2005	1
11	1200020045130100	2004	2005	1
12	1200020094370100	2004	2005	1
13	1200020118450100	2004	2005	1
14	1200020122680100	2004	2005	1

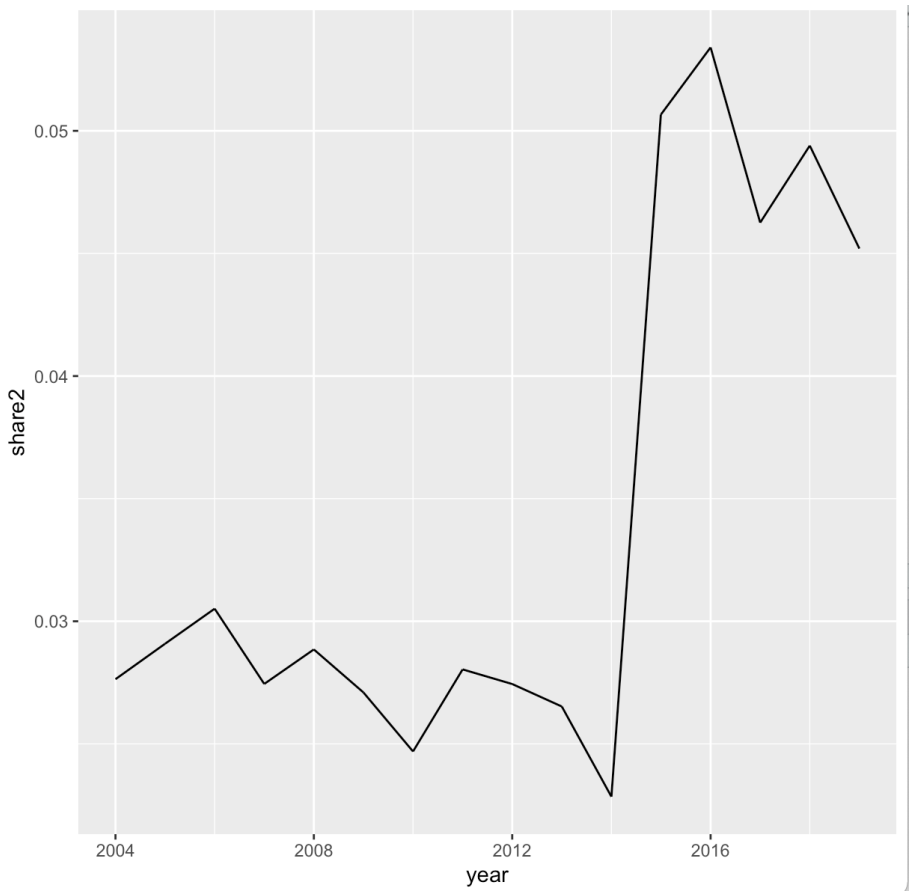
```
#3.2
#Base on datent,identify whether or not household moved into its current dwelling at the year of survey.
#Report the first 10 rows of your result and plot the share of individuals in that situation across years.
dat[17] = dat[2]==dat[12] #V17 identify whether or not household moved into its current dwelling at the year of survey
dat[1:10,] #Report the first 10 rows
#Convert the logical varibles into 0-1
dat[18] = ifelse(dat[17] == TRUE,1,0)
#Compute the share of each year
dat_share = dat %>% group_by(year) %>% mutate(share = length(which(V18 == 1))/length(which(V18 >= 0)))
#Plot the share of individuals in that situation across years
ggplot(select(dat_share,year,share),aes(x=year,y=share))+geom_line()
```




```
#3.3
#Base on myear and move, identify whether or not household migrated at the year of survey.
#Report the first 10 rows of your result and plot the share of individuals in that situation across year
dat_share[20] = ifelse(dat_share$myear == dat_share$year,1,0)
dat_share[21] = ifelse(dat_share$move == 2,1,0)
dat_share[20] = replace(dat_share[,20],is.na(dat_share[,20]),0)
dat_share[21] = replace(dat_share[,21],is.na(dat_share[,21]),0)
#The column migration shows whether or not household migrated at the year of survey
#where 0 reprent not and 1 reprent migration at the year of survey
dat_share[,"migration"] = dat_share[,20]+dat_share[,21]
dat_share[1:10,c(1,2,20,21,22)] #Report the first 10 rows
```

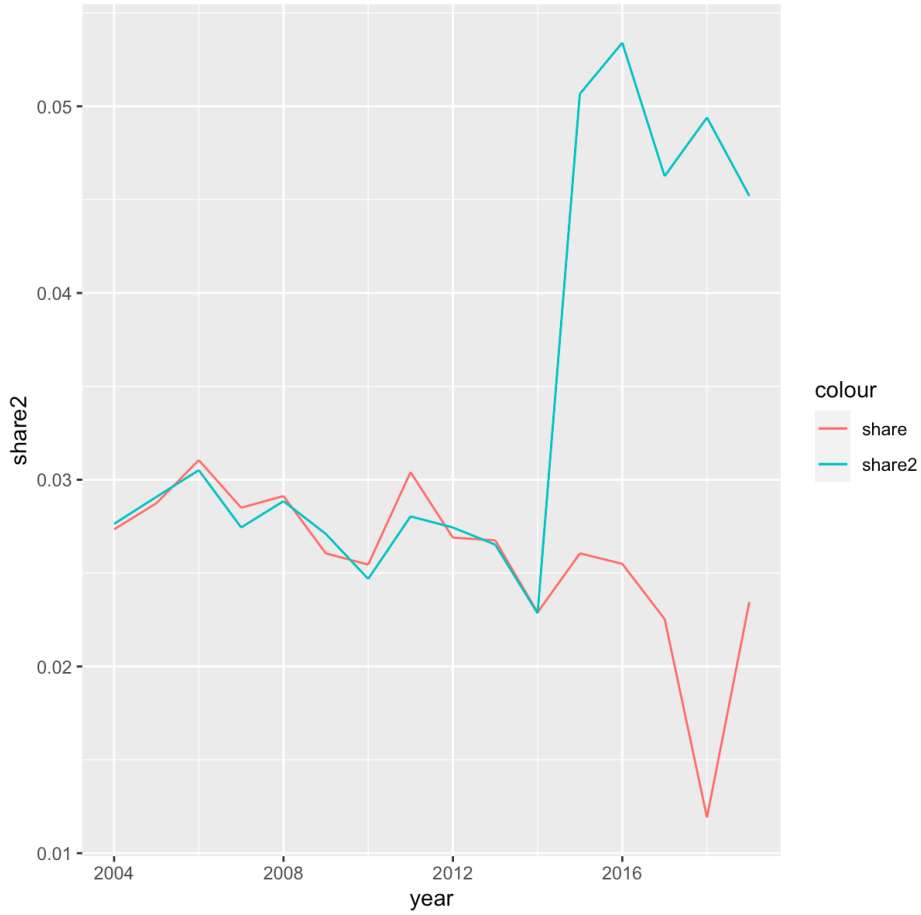
```
# A tibble: 10 x 5
# Groups:   year [2]
     idmen  year ...20 ...21 migration
     <dbl> <int> <dbl> <dbl>     <dbl>
 1 1.20e15  2004     0     0         0
 2 1.20e15  2004     0     0         0
 3 1.20e15  2004     0     0         0
 4 1.20e15  2005     0     0         0
 5 1.20e15  2005     0     0         0
 6 1.20e15  2004     0     0         0
 7 1.20e15  2004     0     0         0
 8 1.20e15  2005     1     0         1
 9 1.20e15  2005     1     0         1
10 1.20e15  2004     0     0         0
```

```
#Compute the share of each year
dat_share = dat_share %>% group_by(year) %>%
   mutate(share2 = length(which(migration == 1))/length(which(migration >= 0)))
#Plot the share of individuals in that situation across years
ggplot(select(dat_share,year,share2),aes(x=year,y=share2))+geom_line()
```




```
#3.4
#Mix the two plots you created above in one graph, clearly label the graph.
ggplot(select(dat_share,year,share,share2),aes(x=year,y=share2))+
   geom_line(mapping=aes(x=year,y=share,color="share"))+
```

geom_line(mapping=aes(x=year,y=share2,color="share2"))



#Do you prefer one method over the other? Justify.
#I prefer the first method, because the graph shows less volatility. Since we use two different variables "myear" and "move" to compute in the second method, there might be more biases according to the changing of statistical caliber. Thus, I think the first method of using "datent" is better.

#3.5
#For households who migrate, find out how many households had at least one family member
#changed his/her profession or employment status.

```
install.packages("tidyverse")
library(tidyverse)
#Remove NA from datent, and the rest households all have migrated.
dat %>% drop_na(datent)
#Add column lagprof and lagemp
dat = dat %>% mutate(lagprof = lag(profession,1,order_by=year),lagemp=lag(empstat,1,order_by=year))
#Identify whether they change their profession and empstat
dat = dat %>% mutate(change = ifelse(lagprof != profession | lagemp != empstat,1,0))
#Filter the data that change equals to 1, which means there exits changes between years
change2 = dat[which(dat["change"]==1),]
#Count the number of households who migrate had at least one family member changed his/her profession or
employment status
#Unique() remove the repeat data.
```

```
length(unique(change2$idind)) #39849


#Exercise4 Attrition
#================================================

#Compute the attrition across each year, where attrition is defined as the reduction
#in the number of individuals staying in the data panel. Report your final result as a table in proportions.
i=0
z=0
vec=1:15
for (y in 2004:2018) {
   dat1 = dat[which(dat["year"]==y),]
   dat2 = dat[which(dat["year"]==y+1),]
   #find the individuals in both panel
   attrition = Reduce(intersect,list(dat1$idind,dat2$idind))
   i=i+1
   #the number of people exit year y = length(dat1$idind)-length(unique(attrition))
   #the number of people enter year y+1 = length(dat2$idind)-length(unique(attrition)
   #define proportion = (people who leave the panel of this year/people who enter the next year)
   vec[i]=(length(dat1$idind)-length(unique(attrition)))/(length(dat2$idind)-length(unique(attrition)))
}
table = cbind(c(2004:2018),vec)
colnames(table)[1]="year"
colnames(table)[2]="proportion"
table
```

```
> table
      year proportion
 [1,] 2004  0.8721263
 [2,] 2005  0.9587440
 [3,] 2006  0.9443901
 [4,] 2007  1.0233009
 [5,] 2008  0.9940353
 [6,] 2009  0.9475100
 [7,] 2010  0.9694240
 [8,] 2011  0.9218149
 [9,] 2012  1.1286878
[10,] 2013  0.9753521
[11,] 2014  1.0082692
[12,] 2015  0.9998262
[13,] 2016  1.0765306
[14,] 2017  1.0444108
[15,] 2018  0.8996686
```