Econ613 HW1 Yida Xu

Exercise1

1、

```
> Datind2007<-read.csv("datind2007.csv")
> library(dplyr)
> numberhousehold<-group_by(Datind2007,idmen)
> count<-count(numberhousehold)
> nrow(count)
[1] 10498
```

2、

```
> Dathh2005<-read.csv("dathh2005.csv")
> mstatus<-group_by(Dathh2005,mstatus)
> count(mstatus)
# A tibble: 5 x 2
# Groups:   mstatus [5]
  mstatus                n
  <chr>              <int>
1 Couple, No kids     2656
2 Couple, with Kids   3374
3 Other                275
4 Single              2663
5 Single Parent        785
```

Number of households with couple with kids is 3374

3、

```
> Datind2008<-read.csv("datind2008.csv")
> numberindividual<-group_by(Datind2008,idind)
> nrow(numberindividual)
[1] 25510
```

4、

```
> Datind2016<-read.csv("datind2016.csv")
> sum(between(Datind2016$age,25,35))
[1] 2765
```

5、

```
#5
Datind2009<-read.csv("datind2009.csv")
CrossTable(Datind2009$gender,Datind2009$profession,prop.chisq = FALSE)
table(Datind2009[,c("gender","profession")])
```

```
> table(Datind2009[,c("gender","profession")])
        profession
gender     0  11  12  13  21  22  23  31  33  34  35  37  38  42  43  44  45  46
  Female  11  30   8  29  63  65   8  68  85 184  50 179  78 258 437   1 153 410
  Male    19  57  19  78 213 114  48  98 107 142  59 260 368 110 117   2  95 340
        profession
gender    47  48  52  53  54  55  56  62  63  64  65  67  68  69
  Female  82  22 782  27 584 353 696  64  35  29  19 147 120  40
  Male   429 215 169 182  98 101  74 443 520 246 159 237 177  82
```

6、

```
> Datind2005<-read.csv("datind2005.csv")
> summarise(Datind2005,mean(wage,na.rm=T),sd(wage,na.rm=T))
  mean(wage, na.rm = T) sd(wage, na.rm = T)
1              11992.26             17318.56
> quantile(Datind2005$wage,0.1,na.rm = T)
10%
  0
> quantile(Datind2005$wage,0.9,na.rm = T)
    90%
32340.4
```

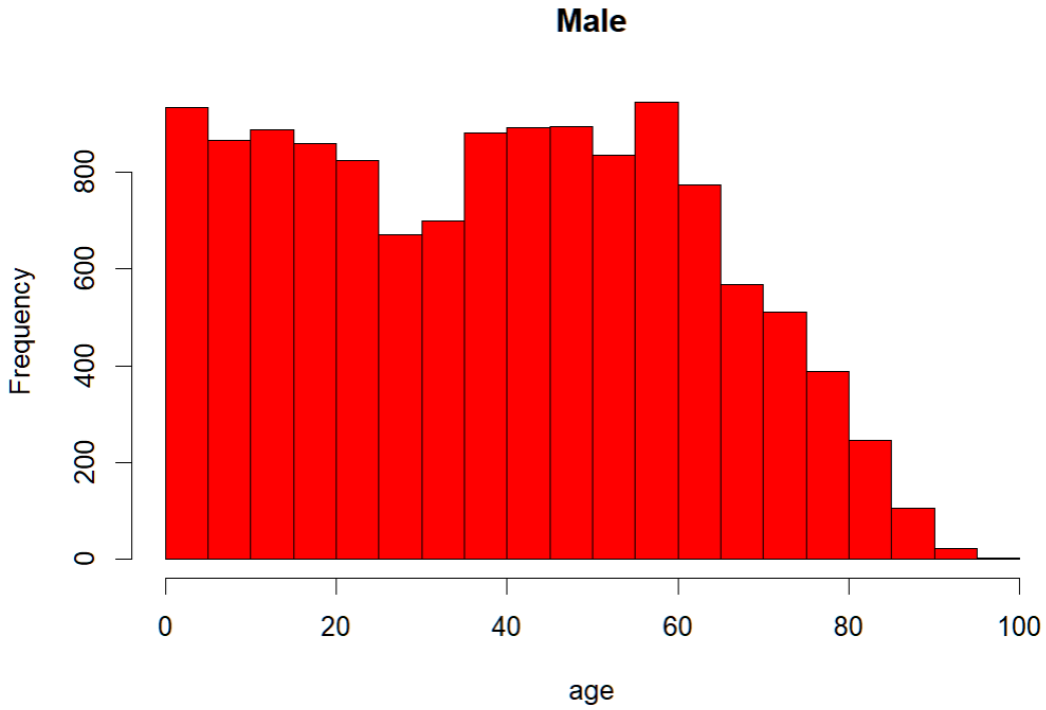
```
D <- na.omit(Datind2005)
D <- arrange(D,wage)
D <- mutate(D,Rect = rank(wage)/n())
D <- mutate(D,Rect_Income = cumsum(wage)/sum(wage))
D <- mutate(D,gini = sum(2*(Rect- Rect_Income)/n()))
D$gini
```

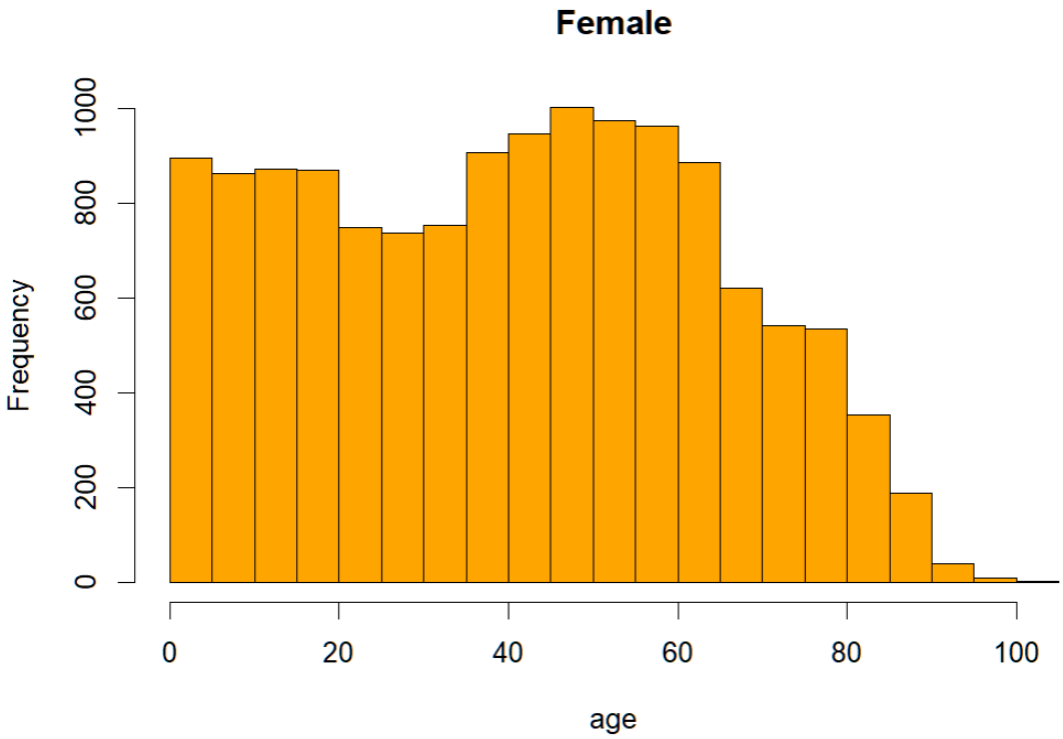
```
> D$gini
   [1] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
   [8] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
  [15] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
  [22] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
  [29] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
  [36] 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654 0.6671654
> Datind2019<-read.csv("datind2019.csv")
> summarise(Datind2019,mean(wage,na.rm=T),sd(wage,na.rm=T))
  mean(wage, na.rm = T) sd(wage, na.rm = T)
1              15350.47             23207.18
> quantile(Datind2019$wage,0.1,na.rm = T)
10%
  0
> quantile(Datind2019$wage,0.9,na.rm = T)
  90%
40267
> D1 <- na.omit(Datind2019)
> D1 <- arrange(D1,wage)
> D1 <- mutate(D1,Rect = rank(wage)/n())
> D1 <- mutate(D1,Rect_Income = cumsum(wage)/sum(wage))
> D1 <- mutate(D1,gini = sum(2*(Rect- Rect_Income)/n()))
> D1$gini
   [1] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
   [8] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
  [15] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
  [22] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
  [29] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
  [36] 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533 0.3864533
```

7、

```
> Datind2010<-read.csv("datind2010.csv")
> summarise(Datind2010,mean(age,na.rm=T),sd(age,na.rm=T))
  mean(age, na.rm = T) sd(age, na.rm = T)
1             39.87893            23.42486
> quantile(Datind2010$age,0.1,na.rm = T)
10%
  8
> quantile(Datind2010$age,0.9,na.rm = T)
90%
 72
```

```
> Female2010<-filter(Datind2010,gender=="Female")
> Male2010<-filter(Datind2010,gender=="Male")
> hist(Female2010$age,xlab = "age",freq = TRUE,main = "Female",col = "orange")
> hist(Male2010$age,xlab = "age",freq = TRUE,main = "Male",col = "red")
>
```





```
> quantile(Female2010$age,0.1,na.rm = T)
10%
  8
> quantile(Female2010$age,0.9,na.rm = T)
90%
 73
> summarise(Female2010,mean(age,na.rm=T),median(age,na.rm=T),sd(age,na.rm=T))
  mean(age, na.rm = T) median(age, na.rm = T) sd(age, na.rm = T)
1             40.81651                    42           23.62735
> summarise(Male2010,mean(age,na.rm=T),median(age,na.rm=T),sd(age,na.rm=T))
  mean(age, na.rm = T) median(age, na.rm = T) sd(age, na.rm = T)
1             38.87362                    39           23.16458
> quantile(Male2010$age,0.1,na.rm = T)
10%
  7
> quantile(Male2010$age,0.9,na.rm = T)
90%
 70
```

8、

```
> Dathh2011<-fread("dathh2011.csv")
> Datind2011<-fread("datind2011.csv")
> newparis<-Dathh2011[which(Dathh2011$location=="Paris")]
> joinparis<-inner_join(Datind2011,newparis,c("idmen"))
> numberparis<-group_by(joinparis,location)
> count(numberparis)
# A tibble: 1 x 2
# Groups:   location [1]
  location     n
  <chr>    <int>
1 Paris     3514
```

Exercise2

1、

```
> datind04<-fread("datind2004.csv",colClasses=c(idind="character",idmen="character"))
> datind05<-fread("datind2005.csv",colClasses=c(idind="character",idmen="character"))
> datind06<-fread("datind2006.csv",colClasses=c(idind="character",idmen="character"))
> datind07<-fread("datind2007.csv",colClasses=c(idind="character",idmen="character"))
> datind08<-fread("datind2008.csv",colClasses=c(idind="character",idmen="character"))
> datind09<-fread("datind2009.csv",colClasses=c(idind="character",idmen="character"))
> datind10<-fread("datind2010.csv",colClasses=c(idind="character",idmen="character"))
> datind11<-fread("datind2011.csv",colClasses=c(idind="character",idmen="character"))
> datind12<-fread("datind2012.csv",colClasses=c(idind="character",idmen="character"))
> datind13<-fread("datind2013.csv",colClasses=c(idind="character",idmen="character"))
> datind14<-fread("datind2014.csv",colClasses=c(idind="character",idmen="character"))
> datind15<-fread("datind2015.csv",colClasses=c(idind="character",idmen="character"))
> datind16<-fread("datind2016.csv",colClasses=c(idind="character",idmen="character"))
> datind17<-fread("datind2017.csv",colClasses=c(idind="character",idmen="character"))
> datind18<-fread("datind2018.csv",colClasses=c(idind="character",idmen="character"))
> datind19<-fread("datind2019.csv",colClasses=c(idind="character",idmen="character"))
> append1<-rbind(datind04,datind05,datind06,datind07,datind08,datind09,datind10,datind
11,datind12,datind13,datind14,datind15,datind16,datind17,datind18,datind19)
```

2、

```
> #2
> dathh04<-fread("dathh2004.csv",colClasses=c(idmen="character"))
> dathh05<-fread("dathh2005.csv",colClasses=c(idmen="character"))
> dathh06<-fread("dathh2006.csv",colClasses=c(idmen="character"))
> dathh07<-fread("dathh2007.csv",colClasses=c(idmen="character"))
> dathh08<-fread("dathh2008.csv",colClasses=c(idmen="character"))
> dathh09<-fread("dathh2009.csv",colClasses=c(idmen="character"))
> dathh10<-fread("dathh2010.csv",colClasses=c(idmen="character"))
> dathh11<-fread("dathh2011.csv",colClasses=c(idmen="character"))
> dathh12<-fread("dathh2012.csv",colClasses=c(idmen="character"))
> dathh13<-fread("dathh2013.csv",colClasses=c(idmen="character"))
> dathh14<-fread("dathh2014.csv",colClasses=c(idmen="character"))
> dathh15<-fread("dathh2015.csv",colClasses=c(idmen="character"))
> dathh16<-fread("dathh2016.csv",colClasses=c(idmen="character"))
> dathh17<-fread("dathh2017.csv",colClasses=c(idmen="character"))
> dathh18<-fread("dathh2018.csv",colClasses=c(idmen="character"))
> dathh19<-fread("dathh2019.csv",colClasses=c(idmen="character"))
> append2<-rbind(dathh04,dathh05,dathh06,dathh07,dathh08,dathh09,dathh10,dathh11,dathh
12,dathh13,dathh14,dathh15,dathh16,dathh17,dathh18,dathh19)
```

3、

```
> #3
> colnames(append1)
 [1] "V1"         "idind"      "idmen"      "year"       "empstat"    "respondent"
 [7] "profession" "gender"     "age"        "wage"
> colnames(append2)
[1] "V1"       "idmen"    "year"     "datent"   "myear"    "mstatus"  "move"
[8] "location"
> intersect(colnames(append1),colnames(append2))
[1] "V1"    "idmen" "year"
```

4、

```
> #4
> mergeappend<-left_join(append1,append2,c("idmen","year"))
```

5、

```
> #5
> newdata<-group_by(mergeappend,idmen,year)
> countofnewdata<-summarise(newdata,count=n())
`summarise()` has grouped output by 'idmen'. You can override using the `.groups` argu
ment.
> filter(countofnewdata,count>4)
# A tibble: 12,436 x 3
# Groups:   idmen [3,622]
   idmen             year count
   <chr>            <int> <int>
 1 1200177087500100  2004     6
 2 1200177087500100  2005     6
 3 1200339082030100  2004     5
 4 1200339082030100  2005     5
 5 1200339108800100  2004     5
 6 1200496078730100  2004     5
 7 1200496078730100  2005     5
 8 1200597103840100  2004     5
 9 1200597103840100  2005     5
10 1200597118450100  2004     5
# ... with 12,426 more rows
> nrow(filter(countofnewdata,count>4))
[1] 12436
```

6、

```
> #6
> append3<-filter(append1,empstat=="Unemployed")
> length(unique(append3$idmen))
[1] 8162
```

7、

```
> #7
> append4<-filter(append1,profession>0)
> append5<-group_by(append4,idmen,year,profession)
> number<-summarise(append5,count=n())
`summarise()` has grouped output by 'idmen', 'year'. You can override using the `.grou
ps` argument.
> nrow(filter(number,count>1))
[1] 7615
```

8、

```
> #8
> number8<-filter(mergeappend,mstatus == "Couple, with Kids")
> length(unique(number8$idind))
[1] 55094
```

9、

```
> #9
> number9<-filter(mergeappend,location == "Paris")
> length(unique(number9$idind))
[1] 14563
```

10、

```
#10
append10<-group_by(append1,year,idmen)
max1<-table(append10$idmen,append10$year)
max2<-as.data.frame(max1)
max2[which.max(max2$Freq),]
max2[which(max2$Freq=="14"),]
Var1 Var2 Freq
2207811124040100 2007   14
2510263102990100 2010   14
```

11、

```
#11
> idem2010<-filter(append1,year=="2010")
> length(unique(idem2010$idmen))
[1] 11050
> idem2011<-filter(append1,year=="2011")
> length(unique(idem2011$idmen))
[1] 11360
```

Exercise 3

1、

```
#1
yearmin=aggregate(mergeappend$year,by=list(mergeappend$idmen),FUN=min)
yearmax=aggregate(mergeappend$year,by=list(mergeappend$idmen),FUN=max)
yearlength<-merge(yearmax,yearmin,by="Group.1")
yearlength[,4]=yearlength[,2]-yearlength[,3]
#v4is the year each household enters and exists the panel.
```

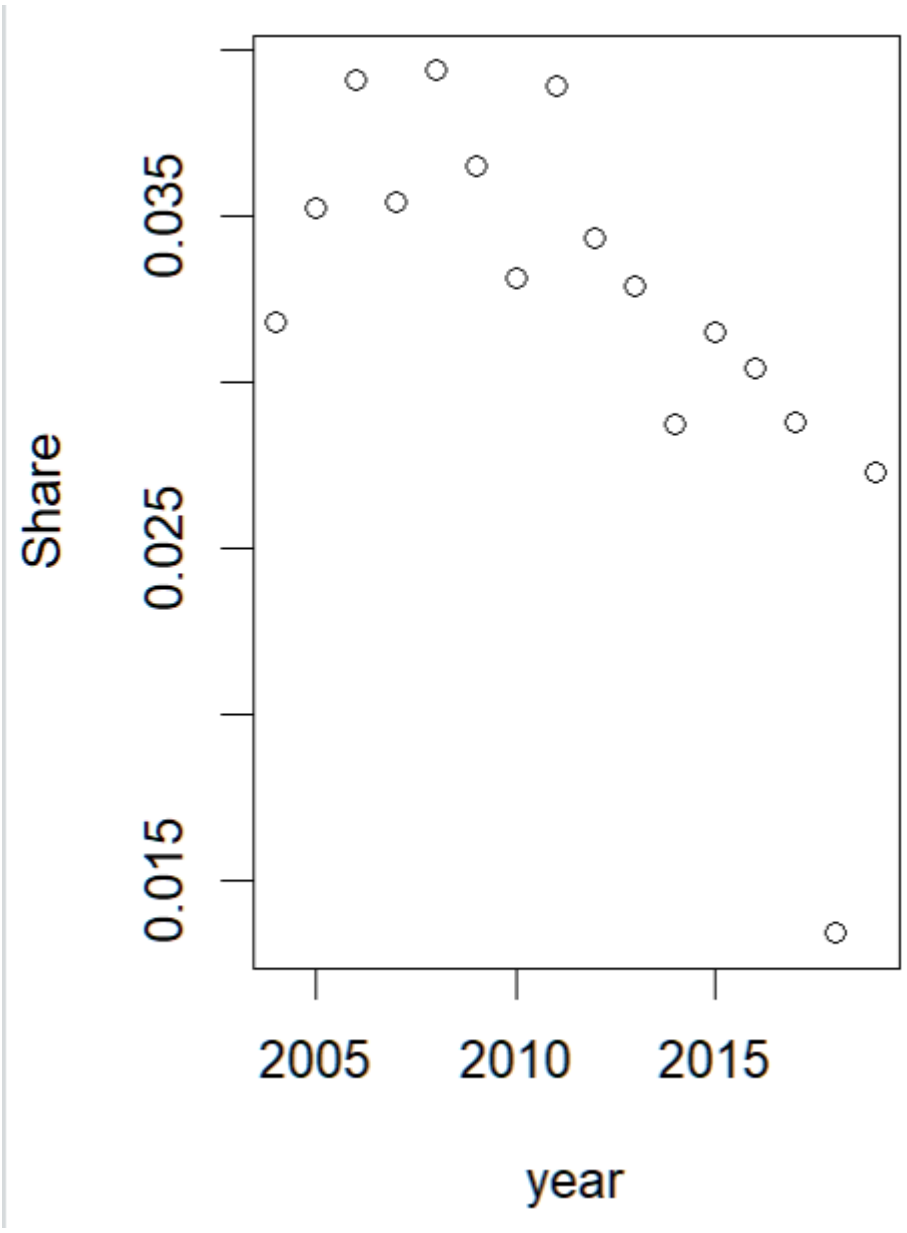
	Group.1	x.x	x.y	V4
1	1200010012930100	2004	2004	0
2	1200010040580100	2005	2004	1
3	1200010066630100	2005	2004	1
4	1200010082450100	2005	2004	1
5	1200010086440100	2005	2004	1
6	1200010102990100	2005	2004	1
7	1200010118450100	2005	2004	1
8	1200020012930100	2005	2004	1
9	1200020017390100	2005	2004	1
10	1200020026420100	2005	2004	1
11	1200020045130100	2005	2004	1
12	1200020094370100	2005	2004	1
13	1200020118450100	2005	2004	1
14	1200020122680100	2005	2004	1
15	1200149012930100	2005	2004	1
16	1200149034710100	2005	2004	1
17	1200149057530100	2005	2004	1
18	1200149073620100	2005	2004	1
19	1200149099400100	2005	2004	1

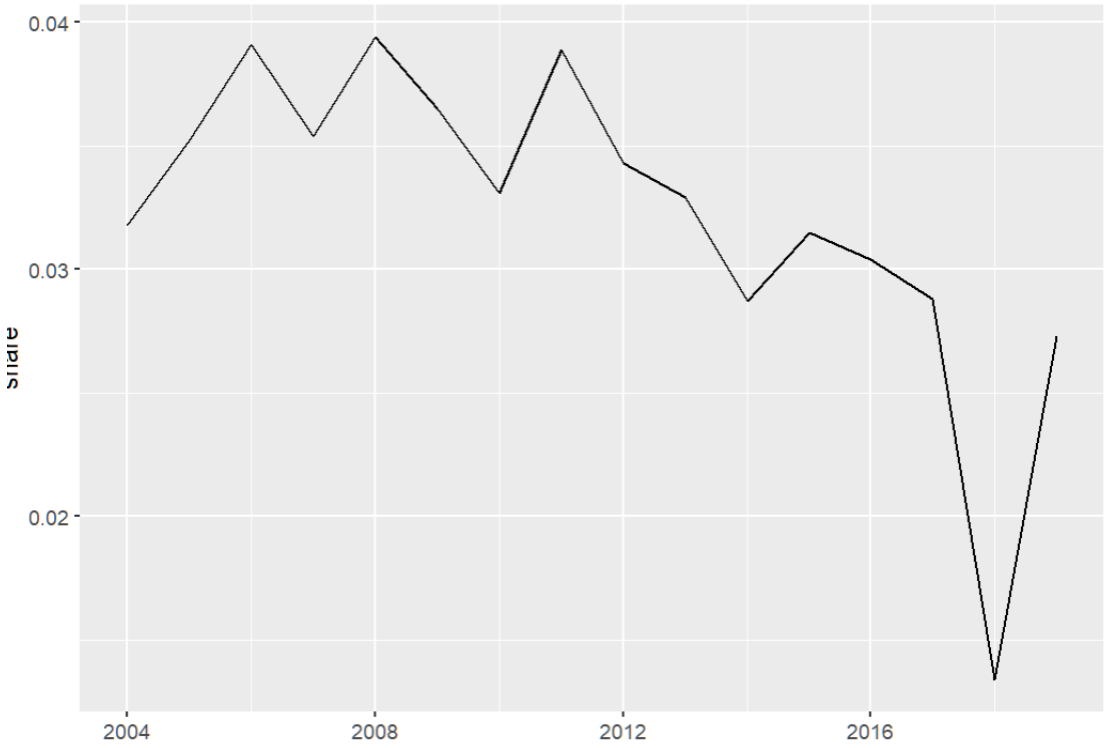
The V4 is the time spent in the survey.

2、

```r
#2
append2edited<-append2
append2edited$movein<-append2edited$year-append2edited$datent==0
head(append2edited$movein,10)
#[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
mergeappededited<-mergeappend
mergeappededited=as.data.frame(mergeappededited)
mergeappededited$year<-as.numeric(as.character(mergeappededited$year))
mergeappededited$idind<-as.numeric(as.character(mergeappededited$idind))
mergeappededited$year = factor(mergeappededited$year)
individualmove=c()
individualmovein=c()
individualnumber=c()
individualn=c()
for (i in 2004:2019) {
  individualmove=length(na.omit(unique(mergeappededited[mergeappededited$datent==i&mergeappededited$year==i,'idind'])))
individualmovein=c(individualmovein,individualmove)
}
for (i in 2004:2019) {
  individualn=length(na.omit(unique(mergeappededited[mergeappededited$year==i,'idind'])))
individualnumber=c(individualnumber,individualn)
}
shareofindividual<-data.frame(year=2004:2019,share=round(individualmovein/individualnumber,4))
shareofindividual
x1<-plot(shareofindividual$year,shareofindividual$share, xlab = "year", ylab = "Share")
y1<-ggplot(select(shareofindividual,year,share),aes(x=year,y=share))+geom_line()
```

```
> shareofindividual
   year  share
1  2004 0.0318
2  2005 0.0352
3  2006 0.0391
4  2007 0.0354
5  2008 0.0394
6  2009 0.0365
7  2010 0.0331
8  2011 0.0389
9  2012 0.0343
10 2013 0.0329
11 2014 0.0287
12 2015 0.0315
13 2016 0.0304
14 2017 0.0288
15 2018 0.0134
16 2019 0.0273
```


Share
0.035
0.025
0.015
2005
2010
2015
year



All false for the first ten rows, and the plot is showed above.

3、

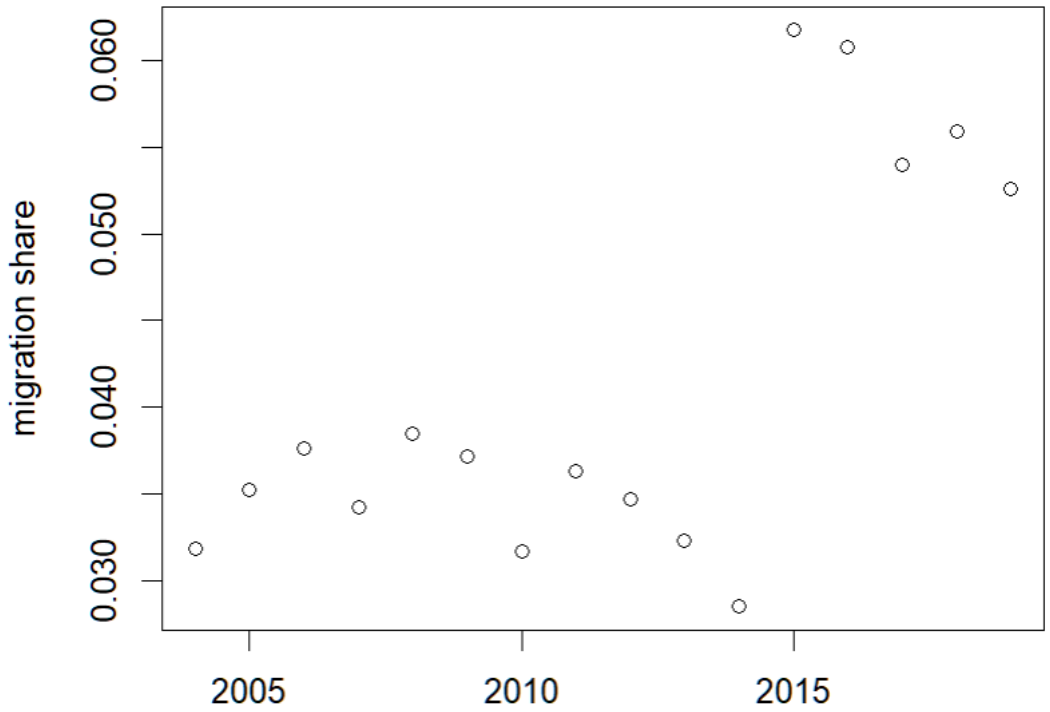
```
#3
Q33<-append2
Q33<-group_by(Q33,idmen,year,myear)
Q33<-filter(Q33,year==myear,year<2015)
Q331<-append2
Q331<-group_by(Q331,idmen,year,myear)
Q331<-filter(Q331,move=="2",year>=2015)
Q33t<-rbind(Q33,Q331)
head(Q33t,10)
```

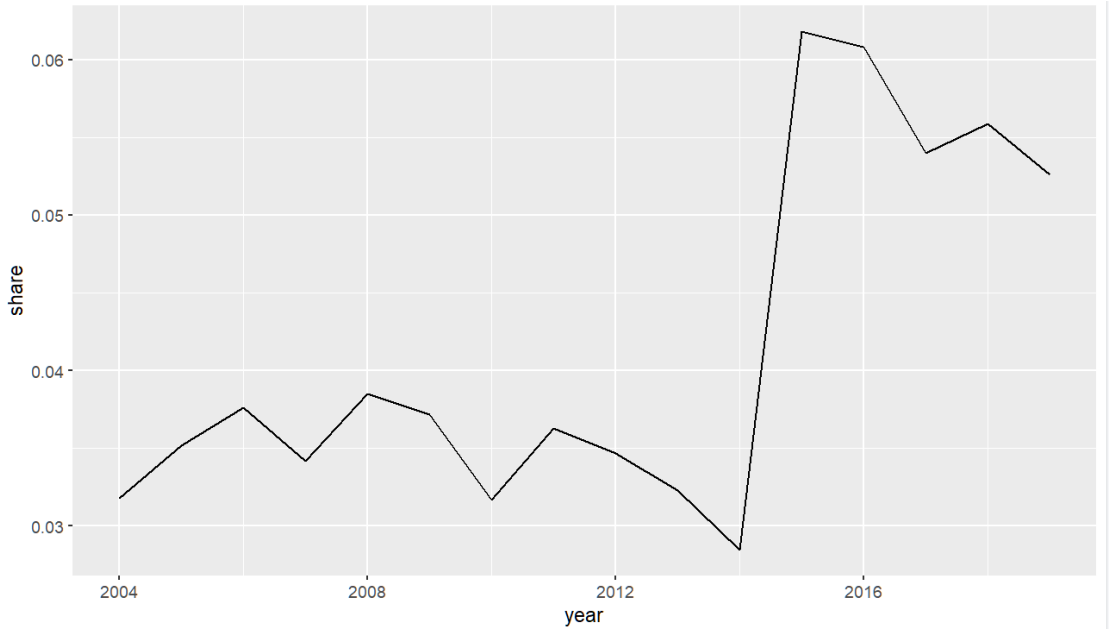
```
> head(Q33t,10)
# A tibble: 10 x 8
# Groups:   idmen, year, myear [10]
      V1 idmen             year datent myear mstatus            move location
   <int> <chr>            <int>  <int> <int> <chr>              <int> <chr>
 1    43 1200493010270100  2004   2004  2004 Couple, with Kids    NA Rural
 2    85 1200742020540100  2004   2004  2004 Couple, No kids      NA Urban 10000 t~
 3   115 1200896012620100  2004   2004  2004 Single               NA Paris
 4   164 1201386067860100  2004   2004  2004 Single               NA Paris
 5   167 1201386106580100  2004   2004  2004 Single Parent        NA Paris
 6   233 1202243012930100  2004   2004  2004 Couple, No kids      NA Rural
 7   308 1202839101420100  2004   2004  2004 Couple, with Kids    NA Paris
 8   310 1202969015750100  2004   2004  2004 Couple, No kids      NA Urban 100000 ~
 9   338 1203431075690100  2004   2004  2004 Couple, No kids      NA Paris
10   347 1203516057590100  2004   2004  2004 Couple, with Kids    NA Urban 2000 to~
```

This is the first ten row of house that migrated at the year of survey.

```
individualmigration = c()
for (i in 2004:2019){
  individualmig = length(na.omit(unique(mergeappededited[(mergeappededited$myear == i
   | mergeappededited$move == 2)
  & mergeappededited$year == i,'idind'])))
  individualmigration = c(individualmigration, individualmig)
}
shareofmigration = data.frame(year = 2004:2019,
   share = round(individualmigration / individualnumber,4))
x2<-plot(shareofmigration$year,shareofmigration$share, xlab = "year", ylab = "migration share")
y2<-ggplot(select(shareofmigration,year,share),aes(x=year,y=share))+geom_line()
```

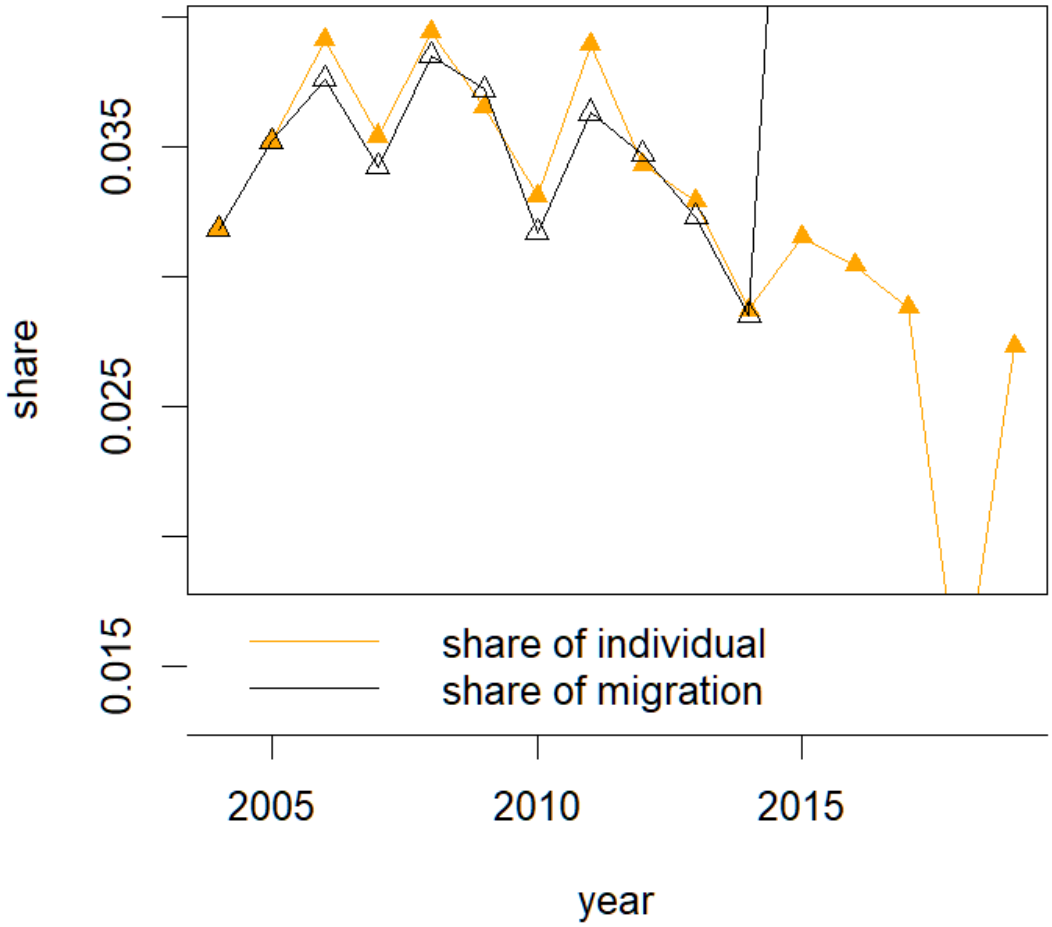
```
> shareofmigration
    year  share
1   2004 0.0318
2   2005 0.0352
3   2006 0.0376
4   2007 0.0342
5   2008 0.0385
6   2009 0.0372
7   2010 0.0317
8   2011 0.0363
9   2012 0.0347
10  2013 0.0323
11  2014 0.0285
12  2015 0.0618
13  2016 0.0608
14  2017 0.0540
15  2018 0.0559
16  2019 0.0526
```





4、

```
#4
plot(shareofindividual, type = "o", pch = 17, col = "orange", xlab = "year", ylab = "share")
lines(shareofmigration, type = "o", pch = 24, col = "black")
legend("bottomleft", c("share of individual", "share of migration"), lty = c(1,1), col = c("orange", "black"))
```



I prefer the method of 3.2 because the data of 3.3 is not complete, and we need additional method to deal with the gap of the “myear” and “move”.