Result

Exercise1

1. Number of households surveyed in 2007.

```
> data1=read.csv("dathh2007.csv")
> max(data1$X)
[1] 10498
```

2. Number of households with marital status “Couple with kids” in 2005.

```
> data2=read.csv("dathh2005.csv")
> table1=table(factor(data2$mstatus))
> table1[names(table1)=="Couple, with Kids"]
Couple, with Kids
             3374
```

3. Number of individuals surveyed in 2008.

```
> #3
> data3=read.csv("datind2008.csv")
> length(unique(data3$idind))
[1] 10828
```

4. Number of individuals aged between 25 and 35 in 2016.

```
> #4
> data4=read.csv("datind2016.csv")
> a=data4$age>=25 & data4$age<=35
> table2=summary(a)
> num1=table2[3]
> num1
TRUE
2765
```

5. Cross-table gender/profession in 2009.

```
> data5=read.csv("datind2009.csv")
> crosstable = table(data5$gender, data5$profession)
> crosstable

          0  11  12  13  21  22  23  31  33  34  35  37  38  42  43  44  45  46  47  48
  Female  11  30   8  29  63  65   8  68  85 184  50 179  78 258 437   1 153 410  82  22
  Male    19  57  19  78 213 114  48  98 107 142  59 260 368 110 117   2  95 340 429 215

         52  53  54  55  56  62  63  64  65  67  68  69
  Female 782  27 584 353 696  64  35  29  19 147 120  40
  Male   169 182  98 101  74 443 520 246 159 237 177  82
```

6. Distribution of wages in 2005 and 2019. Report the mean, the standard deviation, the inter-decile ratio D9/D1 and the Gini.

For 2005

```
> data6=read.csv("datind2005.csv")
> data7=read.csv("datind2019.csv")
> #for 2005
> #mean
> mean(data6$wage,na.rm=T)
[1] 11992.26
> #The standard deviation
> sd(data6$wage, na.rm=TRUE)
[1] 17318.56
> #IQR
> IQR(data6$wage, na.rm=T)
[1] 20453
> #Gini
> data6_1=data6%>%
+   drop_na(wage)
> u=sum(data6_1$wage*1/18767)
> x = data6_1$wage
> y = t(data6_1$wage)
> 1/(2*u)*1/18767*1/18767*sum(abs(outer(x,y,FUN="-")))
[1] 0.6671654
```

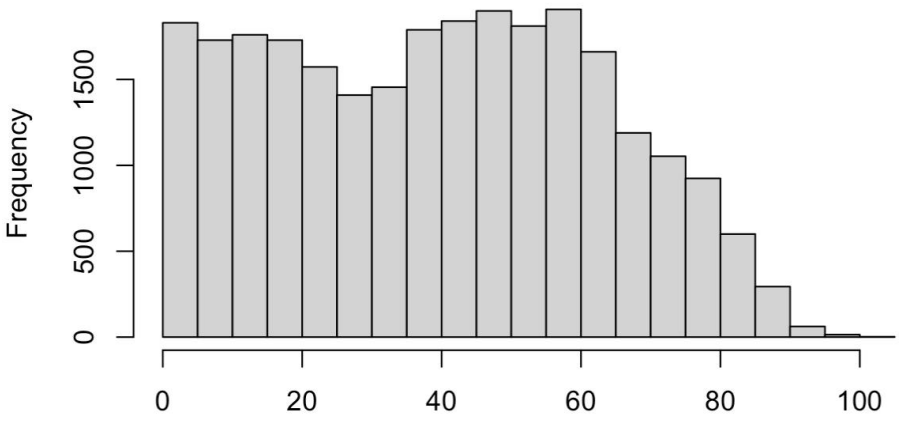
For 2019

```
> ##for 2019
> #mean
> mean(data7$wage,na.rm=T)
[1] 15350.47
> #The standard deviation
> sd(data7$wage,na.rm=T)
[1] 23207.18
> #IQR
> IQR(data7$wage, na.rm=T)
[1] 26428
> #Gini
> data7_1=data7 %>%
+ drop_na(wage)
> u=sum(data7_1$wage*1/21421)
> x = data7_1$wage
> y = t(data7_1$wage)
> 1/(2*u)*1/21421*1/21421*sum(abs(outer(x,y,FUN="-")))
[1] 0.6655301
```

7. Distribution of age in 2010. Plot an histogram. Is there any difference between men and women?

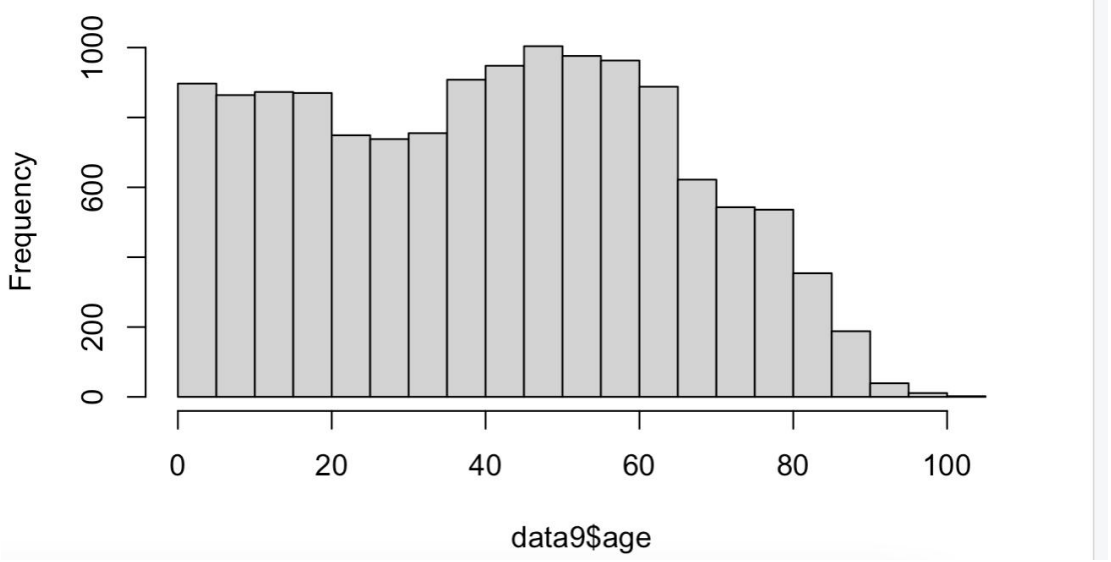
For the whole

```
> data8=read.csv("datind2010.csv")
> hist(data8$age)
```



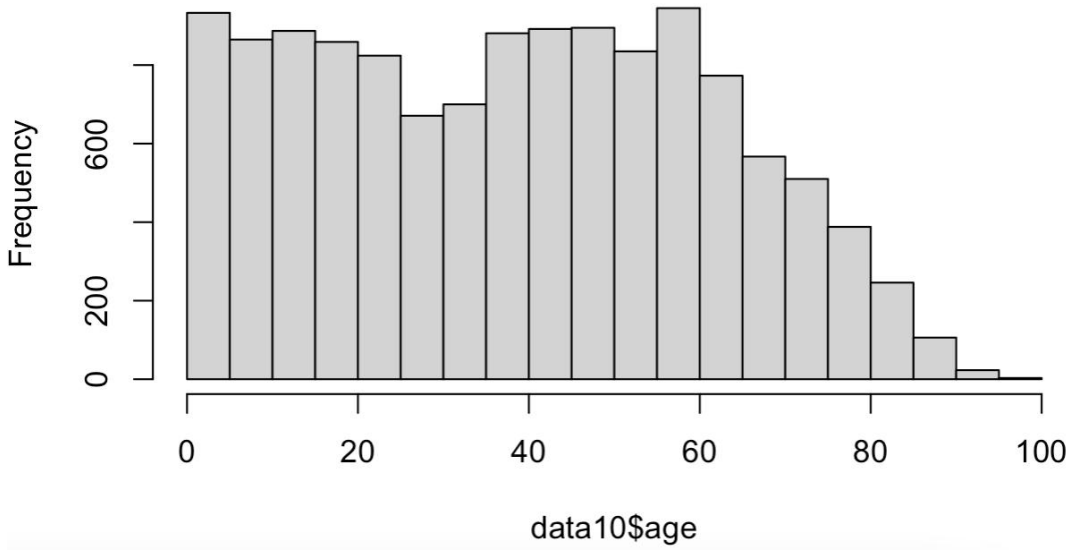
For female

```
> #7 Plot an histogram of 2010
> data8=read.csv("datind2010.csv")
>
> #for female
> data9=filter(data8,data8$gender=="Female")
> hist(data9$age)
```



For male

```
> ##for male
> data10=filter(data8,data8$gender=="Male")
> hist(data10$age)
```



Thus, there is difference between males and females

8. Number of individuals in Paris in 2011.

```
> data8=read.csv("datind2010.csv")
> hist(data8$age)
> datind2011=read.csv("datind2011.csv")
> dathh2011=read.csv("dathh2011.csv")
> merge2011=merge(datind2011, dathh2011,by="idmen")
> merge2011_1=merge2011%>%filter(location=="Paris")
> length(unique(merge2011_1$idind))
[1] 1552
```

Number of individuals in Paris in 2011 is 1552.

Exercise 2

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

```
> #============================================
> # Exercise 2:
> #============================================
> #1 Read individual datasets
> options(scipen = 200)
> datind2004=read.csv("datind2004.csv")
> datind2005=read.csv("datind2005.csv")
> datind2006=read.csv("datind2006.csv")
> datind2007=read.csv("datind2007.csv")
> datind2008=read.csv("datind2008.csv")
> datind2009=read.csv("datind2009.csv")
> datind2010=read.csv("datind2010.csv")
> datind2011=read.csv("datind2011.csv")
> datind2012=read.csv("datind2012.csv")
> datind2013=read.csv("datind2013.csv")
> datind2014=read.csv("datind2014.csv")
> datind2015=read.csv("datind2015.csv")
> datind2016=read.csv("datind2016.csv")
> datind2017=read.csv("datind2017.csv")
> datind2018=read.csv("datind2018.csv")
> datind2019=read.csv("datind2019.csv")
> ##append all the data
> Append1=rbind(datind2004,datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,da
tind2011,datind2012,datind2013,datind2014,datind2015,datind2016,datind2017,datind2018,datind201
9)
```

	X	idind	idmen	year	empstat	respondent	profession	gend
1	1	1120001001293010048	1200010012930100	2004	Employed	1	67	Male
2	2	1120001004058009984	1200010040580100	2004	Employed	1	56	Fema
3	3	1120001004058009984	1200010040580100	2004	Inactive	0		Fema
4	4	1120001006663010048	1200010066630100	2004	Employed	1	38	Male
5	5	1120001006663010048	1200010066630100	2004	Employed	0	45	Fema
6	6	1120001008245010048	1200010082450100	2004	Retired	1		Fema
7	7	1120001008644009984	1200010086440100	2004	Employed	1	34	Male
8	8	1120001008644009984	1200010086440100	2004	Employed	0	42	Fema
9	9	1120001010299010048	1200010102990100	2004	Employed	1	46	Fema
10	10	1120001010299010048	1200010102990100	2004	Inactive	0		Fema
11	11	1120001011845010048	1200010118450100	2004	Employed	1	37	Male
12	12	1120001011845010048	1200010118450100	2004	Employed	0	54	Fema
13	13	1120002001293010048	1200020012930100	2004	Employed	1	11	Male
14	14	1120002001293010048	1200020012930100	2004	Employed	0	11	Fema
15	15	1120002001293010048	1200020012930100	2004	Retired	0		Fema
16	16	1120002001293010048	1200020012930100	2004	Employed	0	63	Male
17	17	1120002001739010048	1200020017390100	2004	Employed	1	11	Male
18	18	1120002002642009984	1200020026420100	2004	Employed	1	11	Male
19	19	1120002002642009984	1200020026420100	2004	Employed	0	11	Male

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

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

```
> #2 Read all household datasets from 2004 to 2019. Append all these datasets.
> #Read household datasets
> dathh2004=read.csv("dathh2004.csv")
> dathh2005=read.csv("dathh2005.csv")
> dathh2006=read.csv("dathh2006.csv")
> dathh2007=read.csv("dathh2007.csv")
> dathh2008=read.csv("dathh2008.csv")
> dathh2009=read.csv("dathh2009.csv")
> dathh2010=read.csv("dathh2010.csv")
> dathh2011=read.csv("dathh2011.csv")
> dathh2012=read.csv("dathh2012.csv")
> dathh2013=read.csv("dathh2013.csv")
> dathh2014=read.csv("dathh2014.csv")
> dathh2015=read.csv("dathh2015.csv")
> dathh2016=read.csv("dathh2016.csv")
> dathh2017=read.csv("dathh2017.csv")
> dathh2018=read.csv("dathh2018.csv")
> dathh2019=read.csv("dathh2019.csv")
> Append2=rbind(dathh2004,dathh2005,dathh2006,dathh2007,dathh2008,dathh2009,dathh2010,dathh2011,d
athh2012,dathh2013,dathh2014,dathh2015,dathh2016,dathh2017,dathh2018,dathh2019)
```

	X	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

Showing 1 to 12 of 173,851 entries, 8 total columns

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

```
> ls(dathh2019)
[1] "datent"   "idmen"    "location" "move"     "mstatus"  "myear"    "X"
[8] "year"
> ls(datind2019)
 [1] "age"        "empstat"    "gender"     "idind"      "idmen"      "profession"
 [7] "respondent" "wage"       "X"          "year"
> #Then find the same variables
> ##The variables are X, year, idmen
```

4.Merge the appended individual and household datasets.

```
> #4Merge the individual and household datasets. Then append them in a dataset called merge1
> merge2004=merge(datind2004, dathh2004,by=c('idmen','year'))
> merge2004=merge(datind2004, dathh2004,by=c('idmen','year'))
> merge2005=merge(datind2005, dathh2005,by=c('idmen','year'))
> merge2006=merge(datind2006, dathh2006,by=c('idmen','year'))
> merge2007=merge(datind2007, dathh2007,by=c('idmen','year'))
> merge2008=merge(datind2008, dathh2008,by=c('idmen','year'))
> merge2009=merge(datind2009, dathh2009,by=c('idmen','year'))
> merge2010=merge(datind2010, dathh2010,by=c('idmen','year'))
> merge2011=merge(datind2011, dathh2011,by=c('idmen','year'))
> merge2012=merge(datind2012, dathh2012,by=c('idmen','year'))
> merge2013=merge(datind2013, dathh2013,by=c('idmen','year'))
> merge2014=merge(datind2014, dathh2014,by=c('idmen','year'))
> merge2015=merge(datind2015, dathh2015,by=c('idmen','year'))
> merge2016=merge(datind2016, dathh2016,by=c('idmen','year'))
> merge2017=merge(datind2017, dathh2017,by=c('idmen','year'))
> merge2018=merge(datind2018, dathh2018,by=c('idmen','year'))
> merge2019=merge(datind2019, dathh2019,by=c('idmen','year'))
> merge1=merge(Append1,Append2, by=c("idmen","year"))
```

	idmen	year	X.x	idind	empstat	respondent	profession	gende
1	1200010012930100	2004	1	1120001001293010048	Employed	1	67	Male
2	1200010040580100	2004	2	1120001004058009984	Employed	1	56	Femal
3	1200010040580100	2004	3	1120001004058009984	Inactive	0		Femal
4	1200010040580100	2005	1	1120001004058009984	Inactive	1		Femal
5	1200010040580100	2005	2	1120001004058009984	Inactive	0		Femal
6	1200010066630100	2004	4	1120001006663010048	Employed	1	38	Male
7	1200010066630100	2004	5	1120001006663010048	Employed	0	45	Femal
8	1200010066630100	2005	4	1120001006663010048	Employed	0	45	Femal
9	1200010066630100	2005	3	1120001006663010048	Employed	1	38	Male
10	1200010082450100	2004	6	1120001008245010048	Retired	1		Femal
11	1200010082450100	2005	5	1120001008245010048	Retired	1		Femal
12	1200010086440100	2004	7	1120001008644009984	Employed	1	34	Male
13	1200010086440100	2004	8	1120001008644009984	Employed	0	42	Femal
14	1200010086440100	2005	6	1120001008644009984	Employed	1	34	Male
15	1200010086440100	2005	7	1120001008644009984	Employed	0	42	Femal
16	1200010102990100	2004	10	1120001010299010048	Inactive	0		Femal
17	1200010102990100	2004	9	1120001010299010048	Employed	1	46	Femal
18	1200010102990100	2005	9	1120001010299010048	Inactive	0		Femal

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

```
> a1=merge2004 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a2=merge2005 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a3=merge2006 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a4=merge2007 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a5=merge2008 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a6=merge2009 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a7=merge2010 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a8=merge2011 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a9=merge2012 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a10=merge2013 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a11=merge2014 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a12=merge2015 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a13=merge2016 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a14=merge2017 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a15=merge2018 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a16=merge2019 %>% select(idmen) %>% group_by(idmen) %>% summarise(count = n()) %>% filter(count > 4)
> a0=rbind(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10,a11,a12,a13,a14,a15,a16)
> length(unique(a0$idmen))
[1] 3622
```

6.Number of households in which at least one member is unemployed

```
> #5 Number of households in which at least one member is unemployed
> #find the individuals that are unemployed
> #get the idmen of these individuals
> #Remove duplicate idmen
> data_unemployed=filter(merge1,merge1$empstat=="Unemployed")
> length(unique(data_unemployed$idmen))
[1] 8161
```

7.Number of households in which at least two members are of the same profession

```
> #7 Number of households in which at least two members are of the same profession
> #first drop if profession=NA
> #group by year, profession, idmen
> #select households in which at least two members are of the same profession
> #Then get the number
> c=merge1 %>%
+   drop_na(profession) %>%
+   group_by(year, profession,idmen)%>%
+   summarise(count = n())%>%
+   filter(count >= 2)
`summarise()` has grouped output by 'year', 'profession'. You can override using the `.groups` argument.
>   length(unique(c$idmen))
[1] 8752
```

8.Number of individuals in the panel that are from household–Couple with kids

```
> #8 Number of individuals in the panel that are from household-Couple with kids
> data_kids=filter(merge1,merge1$mstatus=="Couple, with Kids")
> length(unique(data_kids$idind))
[1] 15567
```

9.Number of individuals in the panel that are from Paris.

```
> #8 Number of individuals in the panel that are from Paris.
> data_Paris =filter(merge1,merge1$location=="Paris")
> length(unique(data_Paris$idind))
[1] 6177
```

10.Find the household with the most number of family members. Report its idmen.

```
> #9 Find the household with the most number of family members. Report its idmen.
> #first find the most number of family members is 14
> Most_num=merge1%>%
+   group_by(idmen, year) %>%
+   count(idind)
> max(Most_num[,4])
[1] 14
> #then Report the  idme of the family members that is 14
> Most_num %>%
+   filter(n ==14) %>%
+   arrange(idmen)
# A tibble: 2 × 4
# Groups:   idmen, year [2]
    idmen  year   idind     n
    <dbl> <int>   <dbl> <int>
1 2.21e15  2007 1.22e18    14
2 2.51e15  2010 1.25e18    14
```

	idmen	year	idind	n
1	2207811124040100	2007	1220781112404009984	14
2	2510263102990100	2010	1251026310299010048	14

10. Number of households present in 2010 and 2011.

```
> #10Number of households present in 2010 and 2011.
> household2010_2011=rbind(dathh2010,dathh2011)
> length(unique(household2010_2011$idmen))
[1] 13424
```

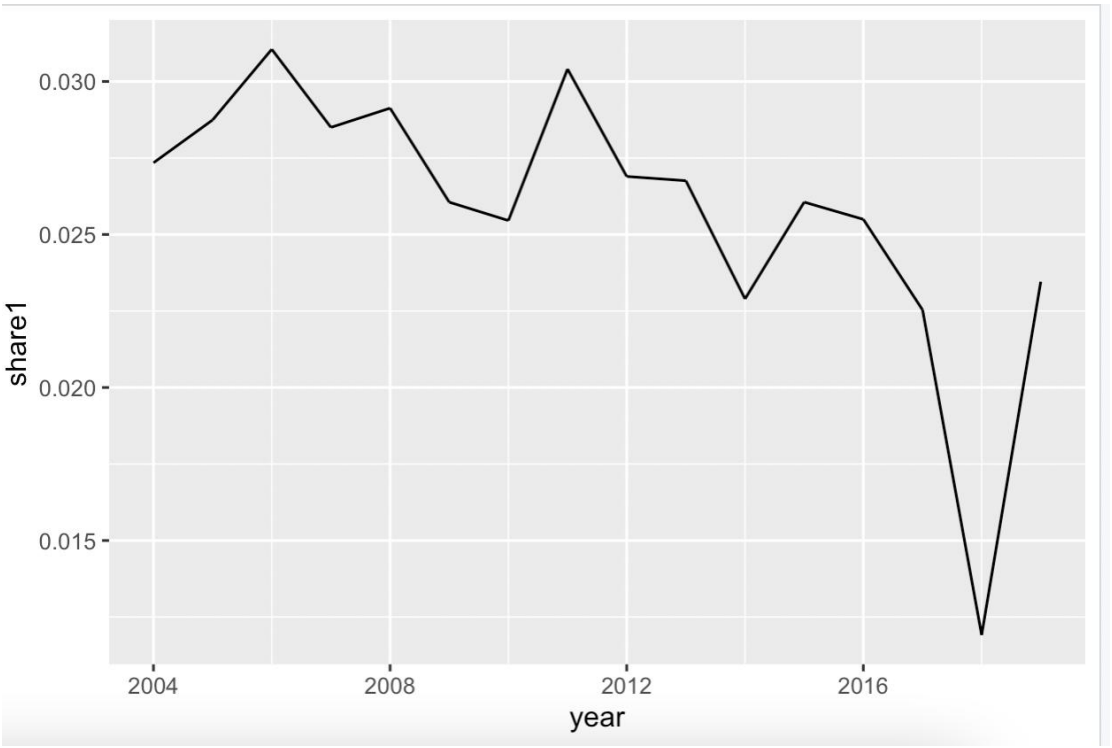
Exercise 3

1.Based on *datent*, identify whether or not a household moved into its current dwelling at the year of survey. Report the fifirst 10 rows of your result and plot the share of individuals in that situation across years.

```
> #============================================
> # Exercise 3:
> #============================================
>
> #1
> #find the earliest and latest year that each household was surveyed
> move1=aggregate(merge1[,2], by=list(merge1$idmen), FUN="min")
> move2=aggregate(merge1[,2], by=list(merge1$idmen), FUN="max")
> move3=merge(move1,move2,by="Group.1")
> #calculate the difference as the distribution of the time spent in the survey for each househol
d.
> move4=cbind(move3,move3$x.y-move3$x.x)
> move4
            Group.1  x.x  x.y move3$x.y - move3$x.x
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
```

```
> #if "year" and "datent" are the same,  a household moved into its current dwelling at the year of survey
> merge1 =merge1 %>%
+   drop_na(datent)
> merge1$live = ifelse(merge1$year == merge1$datent, 1, 0)
> #Report the first 10 rows
> merge1[1:10,]
            idmen year X.x              idind  empstat respondent profession gender age  wage X.y datent myear
1  1200010012930100 2004   1 1120001001293010048 Employed          1         67   Male  31 19187   1   2000  2000
2  1200010040580100 2004   2 1120001004058009984 Employed          1         56 Female  30 11586   2   2001  2001
3  1200010040580100 2004   3 1120001004058009984 Inactive          0            Female   9    NA   2   2001  2001
4  1200010040580100 2005   1 1120001004058009984 Inactive          1            Female  31 12334   1   2001  2001
5  1200010040580100 2005   2 1120001004058009984 Inactive          0            Female  10    NA   1   2001  2001
6  1200010066630100 2004   4 1120001006663010048 Employed          1         38   Male  31 44656   3   2000  2000
7  1200010066630100 2004   5 1120001006663010048 Employed          0         45 Female  27 20413   3   2000  2000
8  1200010066630100 2005   4 1120001006663010048 Employed          0         45 Female  28 19231   2   2005  2005
9  1200010066630100 2005   3 1120001006663010048 Employed          1         38   Male  32 50659   2   2005  2005
10 1200010082450100 2004   6 1120001008245010048  Retired          1            Female  89     0   4   1957  1957
           mstatus move location live
1           Single   NA    Paris    0
2    Single Parent   NA    Paris    0
3    Single Parent   NA    Paris    0
4    Single Parent   NA    Paris    0
5    Single Parent   NA    Paris    0
6  Couple, No kids   NA    Paris    0
7  Couple, No kids   NA    Paris    0
8  Couple, No kids   NA    Paris    1
9  Couple, No kids   NA    Paris    1
10          Single   NA    Paris    0
```

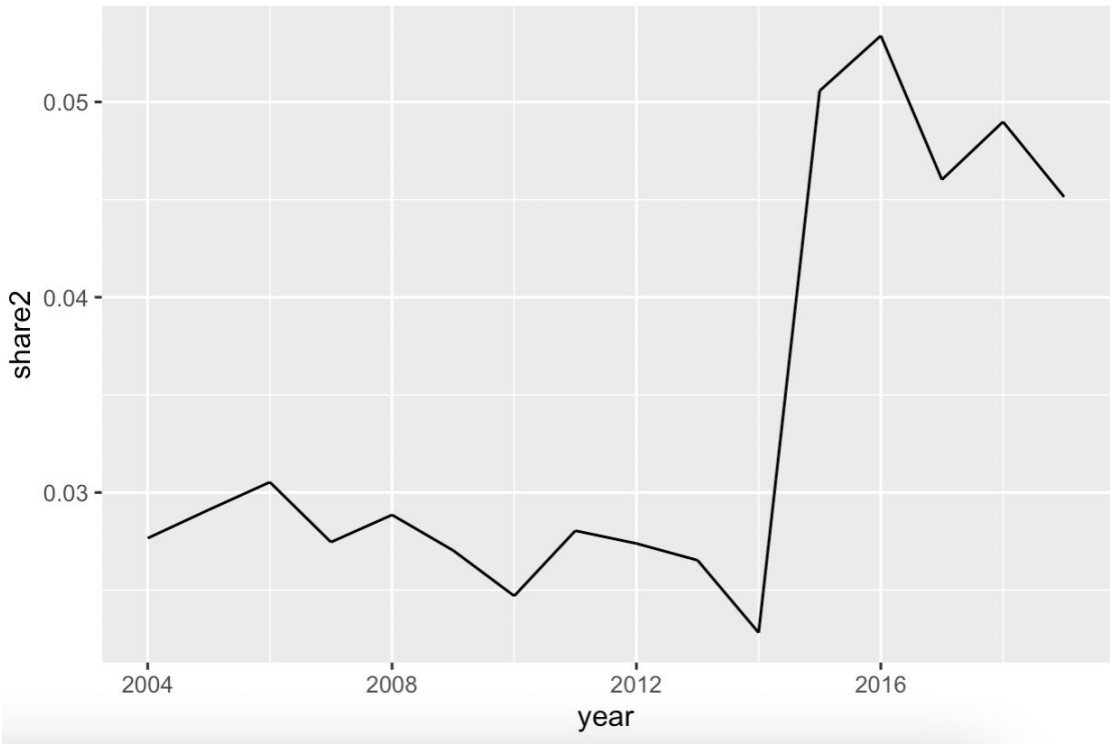
```
> # calculate the share of individuals
> merge1 = merge1 %>%
+   group_by(year) %>%
+   mutate(share1 = sum(live == 1)/sum(live == 0|1))
> #plot across the year
> plot1=select(merge1,year,share1)
> ggplot(plot1, aes(x=year, y=share1)) + geom_line()
```



2.Based on *myear* and *move*, identify whether or not household migrated at the year of survey.

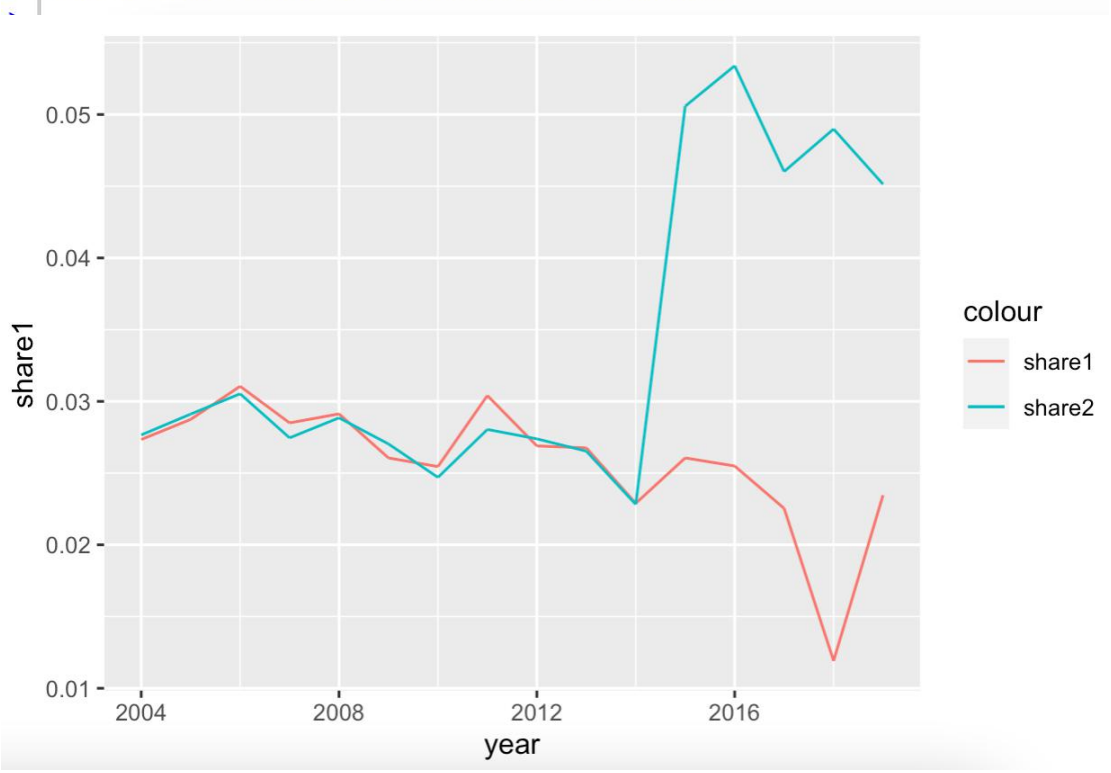
Report the first 10 rows of your result and plot the share of individuals in that situation across years.

```
> merge2=merge1
> #for 2009-2014,use myear to get individuals who have moved,
> merge2$c=ifelse(merge2$year == merge2$myear,  1, 0)
> #for 2015-2019,use mive to get individuals who have moved,
> merge2$d =ifelse(merge2$move == 2,  1, 0)
> #replace NA=0
> merge2$d=replace(merge2$d,is.na(merge2$d),0)
> merge2$c=replace(merge2$c,is.na(merge2$c),0)
> #plus c and d to get individuals who have moved across year
> merge2$e=merge2$d+merge2$c
> #Report the first 10 rows
> merge2[1:10,]
# A tibble: 10 x 21
# Groups:   year [2]
     idmen  year   X.x   idind empstat respondent profession gender   age  wage   X.y datent
     <dbl> <int> <int>   <dbl> <chr>        <int> <chr>      <chr>  <int> <int> <int>  <int>
 1 1.20e15  2004     1 1.12e18 Employ…          1 "67"       Male      31 19187     1   2000
 2 1.20e15  2004     2 1.12e18 Employ…          1 "56"       Female    30 11586     2   2001
 3 1.20e15  2004     3 1.12e18 Inacti…          0 ""         Female     9    NA     2   2001
 4 1.20e15  2005     1 1.12e18 Inacti…          1 ""         Female    31 12334     1   2001
 5 1.20e15  2005     2 1.12e18 Inacti…          0 ""         Female    10    NA     1   2001
 6 1.20e15  2004     4 1.12e18 Employ…          1 "38"       Male      31 44656     3   2000
 7 1.20e15  2004     5 1.12e18 Employ…          0 "45"       Female    27 20413     3   2000
 8 1.20e15  2005     4 1.12e18 Employ…          0 "45"       Female    28 19231     2   2005
 9 1.20e15  2005     3 1.12e18 Employ…          1 "38"       Male      32 50659     2   2005
10 1.20e15  2004     6 1.12e18 Retired          1 ""         Female    89     0     4   1957
# … with 9 more variables: myear <int>, mstatus <chr>, move <int>, location <chr>,
#   live <dbl>, share1 <dbl>, c <dbl>, d <dbl>, e <dbl>
> #calculate the share of individuals who have migrated
> merge2 = merge2 %>% group_by(year) %>%
+   mutate(share2 =sum(e ==1)/sum(e ==0|1))
> #plot across the year
> plot2=select(merge2,year,share2)
> ggplot(merge2, aes(x=year, y=share2)) + geom_line()
```



3.Mix the two plots you created above in one graph, clearly label the graph. Do you prefer one method over the other? Justify.

```
> #3
> #Mix the two plots
> plot3=select(merge2,year,share1,share2)
> ggplot(plot3, aes(x=year, y=share2)) + geom_line()
> plot4= ggplot(plot3) +
+   geom_line(mapping = aes(x=year, y=share1, color = "share1")) +
+   geom_line(mapping = aes(x=year, y=share2, color = "share2"))
> plot4
> #I think the method based on datent is better.
```



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

```
> #4
> merge2=merge2%>%
+   drop_na(datent)%>%
+   filter(merge2$e>0)
```

```
> merge2=merge2%>%
+   mutate(l_profession = lag(profession, 1, order_by = year),
+          l_empstat = lag(empstat, 1, order_by = year))
> count=merge2%>%
+ #if the year the individual moves is the same year the people profession changes, then the value is 1
+     mutate(change = ifelse((l_empstat != empstat|l_profession != profession)& year==datent,1, 0))
> z=count %>% select(change,idind,year) %>% filter(change == 1)
> z%>%group_by(year)
# A tibble: 8,224 x 3
# Groups:   year [16]
   change   idind  year
    <dbl>   <dbl> <int>
 1      1 1.12e18  2005
 2      1 1.12e18  2005
 3      1 2.12e18  2005
 4      1 1.12e18  2005
 5      1 1.12e18  2005
 6      1 1.12e18  2005
 7      1 1.12e18  2005
 8      1 1.12e18  2005
 9      1 1.12e18  2005
10      1 2.12e18  2005
# … with 8,214 more rows
> table(z$year)

2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
 489  545  626  569  567  530  533  628  588  520  486  526  498  434  232  453
> x=unique(z$idind)
> length(x)
[1] 4518
```

Exercise4

```
> #position20xx is to filter the same elements in the ID of the first year and next year
> position2004=Reduce(intersect,list(datind2004$idind,datind2005$idind))
> #calculate the number of the same elements in the ID of 2004 and 2005,the number is 7842
> length(unique(position2004))
[1] 7842
> #(z1)use the total number of the individuals in 2004 minus the number of the same elements to get the individuals who leave
> z1=length(datind2004$idind)-7842
> #(w1)use the total number of the individuals in 2005 minus the number of the same elements to get the individuals who enter
> w1=length(datind2005$idind)-7842
> position2005=Reduce(intersect,list(datind2005$idind,datind2006$idind))
> length(unique(position2005))
[1] 7997
> z2=length(datind2005$idind)-7997
> w2=length(datind2006$idind)-7997
> position2006=Reduce(intersect,list(datind2006$idind,datind2007$idind))
> length(unique(position2006))
[1] 8518

> z3=length(datind2006$idind)-8518
> w3=length(datind2007$idind)-8518
>
> position2007=Reduce(intersect,list(datind2007$idind,datind2008$idind))
> length(unique(position2007))
[1] 8472
> z4=length(datind2007$idind)-8472
> w4=length(datind2008$idind)-8472
>
> position2008=Reduce(intersect,list(datind2008$idind,datind2009$idind))
> length(unique(position2008))
[1] 8678
```

```
> z5=length(datind2008$idind)-8678
> w5=length(datind2009$idind)-8678
>
> position2009=Reduce(intersect,list(datind2009$idind,datind2010$idind))
> length(unique(position2009))
[1] 9058
> z6=length(datind2008$idind)-9058
> w6=length(datind2009$idind)-9058
>
> position2010=Reduce(intersect,list(datind2010$idind,datind2011$idind))
> length(unique(position2010))
[1] 9313
> z7=length(datind2010$idind)-9313
> w7=length(datind2011$idind)-9313
>
> position2011=Reduce(intersect,list(datind2011$idind,datind2012$idind))
> length(unique(position2011))
[1] 9822
> z8=length(datind2011$idind)-9822
> w8=length(datind2012$idind)-9822
>
> position2012=Reduce(intersect,list(datind2012$idind,datind2013$idind))
> length(unique(position2012))
[1] 9405
> z9=length(datind2012$idind)-9405
> w9=length(datind2013$idind)-9405
>
> position2013=Reduce(intersect,list(datind2013$idind,datind2014$idind))
> length(unique(position2013))
[1] 9179

> z10=length(datind2013$idind)-9179
> w10=length(datind2014$idind)-9179
>
> position2014=Reduce(intersect,list(datind2014$idind,datind2015$idind))
> length(unique(position2014))
[1] 9351
> z11=length(datind2013$idind)-9351
> w11=length(datind2014$idind)-9351
>
> position2015=Reduce(intersect,list(datind2015$idind,datind2016$idind))
> length(unique(position2015))
[1] 9385
> z12=length(datind2015$idind)-9385
> w12=length(datind2016$idind)-9385
>
> position2016=Reduce(intersect,list(datind2016$idind,datind2017$idind))
> length(unique(position2016))
[1] 9134
> z13=length(datind2016$idind)-9134
> w13=length(datind2017$idind)-9134
>
> position2017=Reduce(intersect,list(datind2017$idind,datind2018$idind))
> length(unique(position2017))
[1] 8846
> z14=length(datind2017$idind)-8846
> w14=length(datind2018$idind)-8846
>
> position2018=Reduce(intersect,list(datind2018$idind,datind2019$idind))
> length(unique(position2018))
[1] 8683


> attrition=cbind(c(2004:2019),c(z1/w1,z2/w2,z3/w3,z4/w4,z5/w5,z6/w6,z7/w7,z8/w8,z9/w9,z10/w10,z11/w11,z12/w12,z13/w13,z14/w14,z1
5/w15,0))
> attrition
      [,1]      [,2]
 [1,] 2004 0.8721263
 [2,] 2005 0.9587440
 [3,] 2006 0.9443901
 [4,] 2007 1.0233009
 [5,] 2008 0.9940353
 [6,] 2009 0.9938984
 [7,] 2010 0.9695912
 [8,] 2011 0.9218149
 [9,] 2012 1.1286878
[10,] 2013 0.9753521
[11,] 2014 0.9751090
[12,] 2015 0.9998262
[13,] 2016 1.0765306
[14,] 2017 1.0444108
[15,] 2018 0.8996686
[16,] 2019 0.0000000
```