**ECON613**
**Prof. Modibo Sidibe**
**Spring 2022**

# Homework 1: Data(1)

Runling Wu

## 1 Exercise 1: Basic Statistics

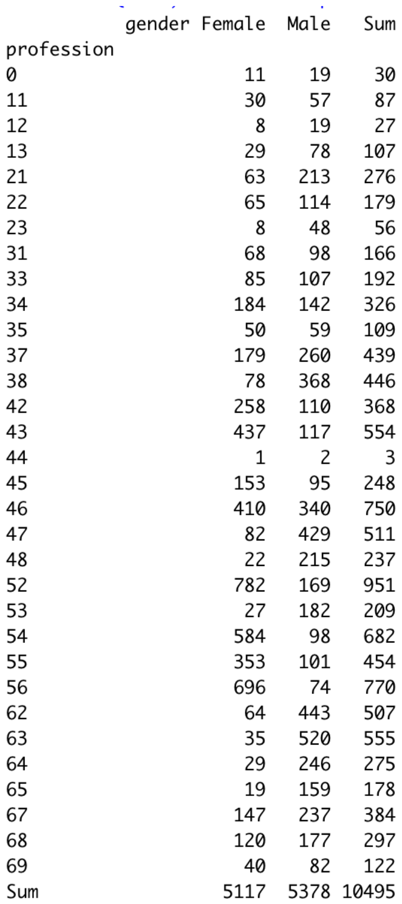
1. The number of households surveyed in 2007 is 10498.

2. The Number of households with marital status “Couple with kids” in 2005 is 3374.

3. The number of individuals surveyed in 2008 is 25510.

4. The number of individuals aged between 25 and 35 in 2016 is 2765.

5. Cross-table gender/profession in 2009.



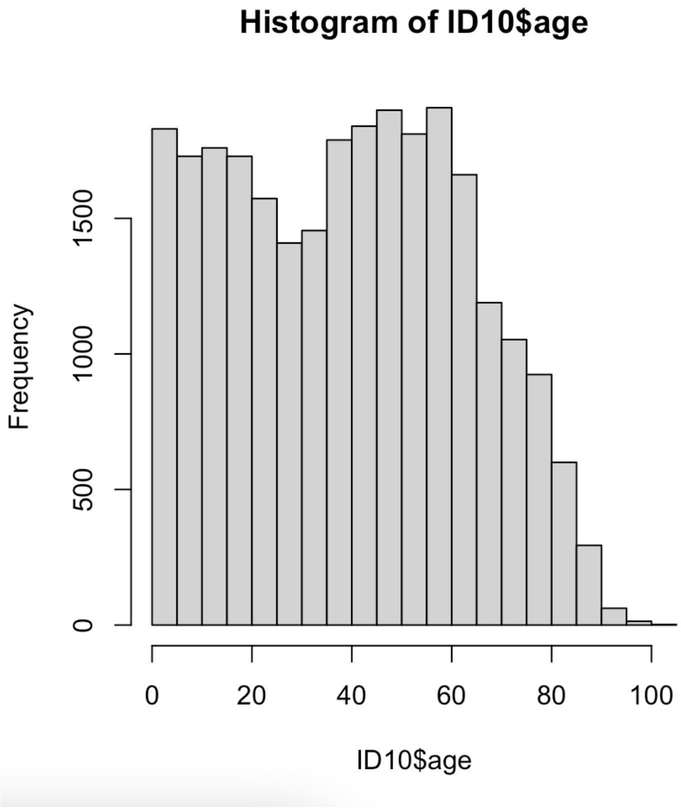
```
           gender Female  Male   Sum
profession
0                     11    19    30
11                    30    57    87
12                     8    19    27
13                    29    78   107
21                    63   213   276
22                    65   114   179
23                     8    48    56
31                    68    98   166
33                    85   107   192
34                   184   142   326
35                    50    59   109
37                   179   260   439
38                    78   368   446
42                   258   110   368
43                   437   117   554
44                     1     2     3
45                   153    95   248
46                   410   340   750
47                    82   429   511
48                    22   215   237
52                   782   169   951
53                    27   182   209
54                   584    98   682
55                   353   101   454
56                   696    74   770
62                    64   443   507
63                    35   520   555
64                    29   246   275
65                    19   159   178
67                   147   237   384
68                   120   177   297
69                    40    82   122
Sum                 5117  5378 10495
```

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

   - There are two possibilities We exclude 0 and missing wages before calculating the statistics and plotting the distribution.
   - The distribution of wages in 2005 is as follows: mean is 22443, the standard deviation is 18076.1, the inter-decile is 8.8965, the gini coefficient is 0.377.
   - The distribution of wages in 2019 is as follows: mean is 27529 , the standard deviation is 25107.19, the inter-decile is 13.8623, the gini coefficient is 0.399.

7. Distribution of age in 2010. We first plot the overall distribution of age in 2010 ( for all females and males).

Figure 1. The distribution of age for females and males



   - The distribution of age for all males and females, respectively. There are slightly differences between the age distribution of males and females. However, we also notice that the sample tends to have more proportion of young males, middle-age females and senior females.

Figure 2. The distribution of age for all males

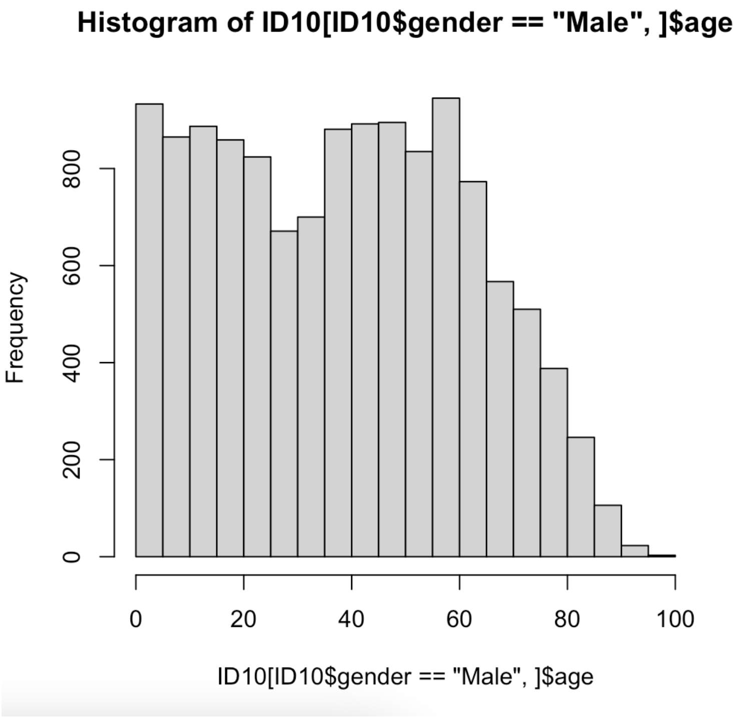
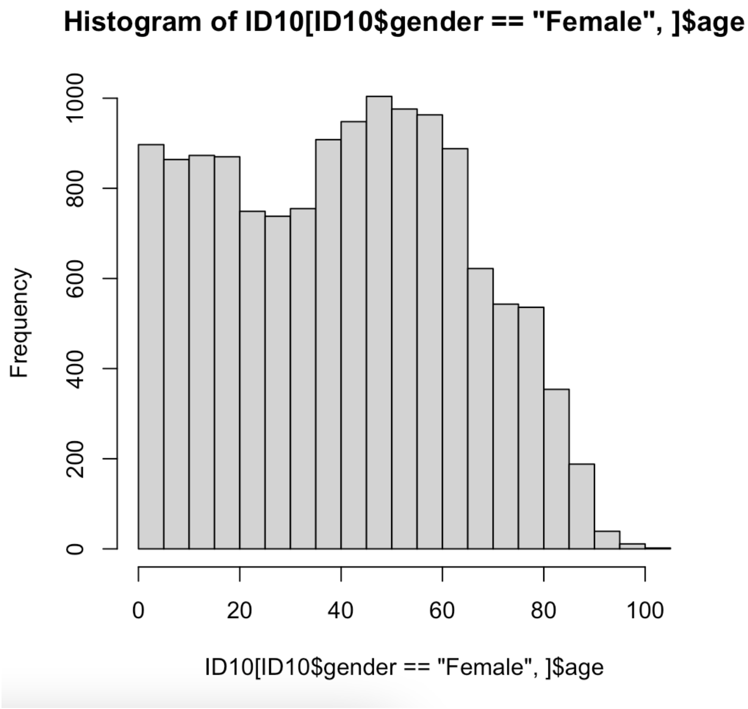


Figure 3. The distribution of age for all females



8. The number of individuals in Paris in 2011 is 3514.

## 2 Exercise 2: Merge Datasets

1. see R scripts for more details.
2. see R scripts for more details.
3. List the variables that are simultaneously present in the individual and household datasets: idmen and year.
4. see R scripts for more details.

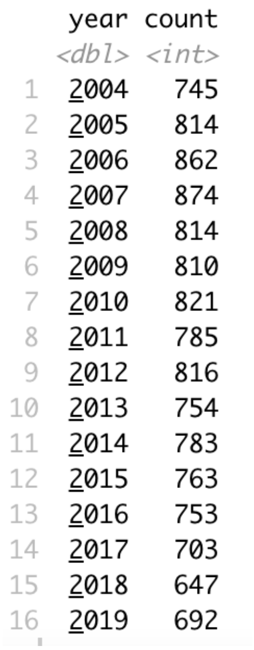
### 2.1 Version I: we consider the dataset as a whole.

5. There are 27195 households in which there are more than four family members.
6. There are 8161 households in which at least one member is unemployed across years.
7. There are 36232 households in which at least two members are of the same profession.
8. If we consider the union, there are total 209371 individuals in the panel that are from household-Couple with kids. If we consider the intersection, there are total 55094 individuals in the panel that are from household-Couple with kids.
9. If we consider the union, there are 51904 individuals in the panel that are from Paris. If we consider the intersection, there are 14563 individuals in the panel that are from Paris.
10. Find the household with the most number of family members. Report its idmen: Two households with 14 family members 2207811124040100 and 2510263102990100.
11. If we consider the union, there are 13424 distinct households present in either 2010 or 2011. If we consider the intersection, there are 8984 households present both in 2010 and 2011.

### 2.2 Version II: we consider it as a highly unbalanced panel, thus we will answer the below questions for every year.

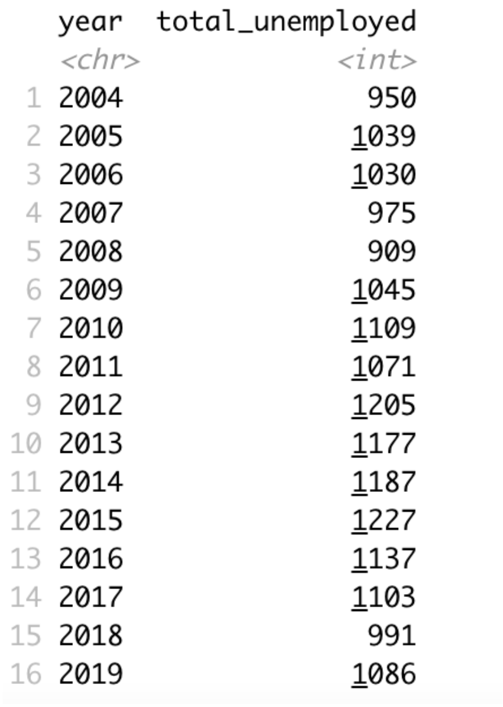
There are only 32 duplicates in individual level data in year 2013. Consider the size of the dataset, this amount of error should be allowed. It is a unbalanced panel, thus we will answer the below questions for every year.

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



```
     year count
    <dbl> <int>
 1   2004   745
 2   2005   814
 3   2006   862
 4   2007   874
 5   2008   814
 6   2009   810
 7   2010   821
 8   2011   785
 9   2012   816
10   2013   754
11   2014   783
12   2015   763
13   2016   753
14   2017   703
15   2018   647
16   2019   692
```

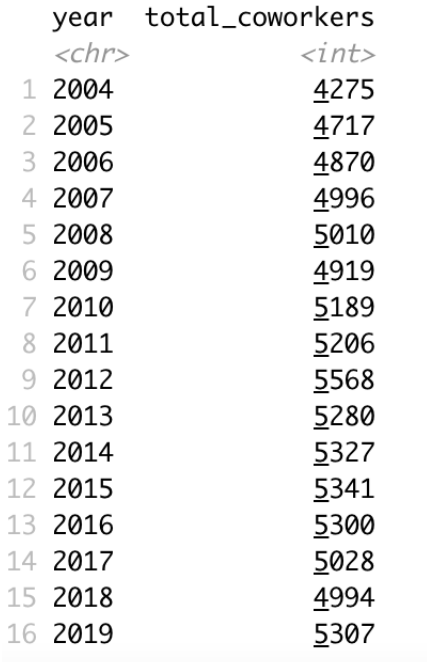
6. The following table shows number of households in which at least one member is unemployed each year.



```
   year  total_unemployed
   <chr>            <int>
 1 2004               950
 2 2005              1039
 3 2006              1030
 4 2007               975
 5 2008               909
 6 2009              1045
 7 2010              1109
 8 2011              1071
 9 2012              1205
10 2013              1177
11 2014              1187
12 2015              1227
13 2016              1137
14 2017              1103
15 2018               991
16 2019              1086
```

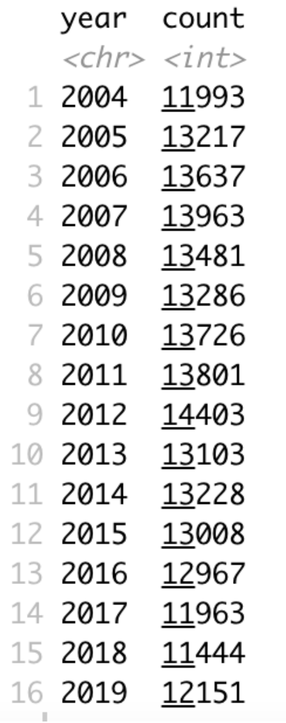
7. The following table shows number of households in which at least two members are of

the same profession each year.



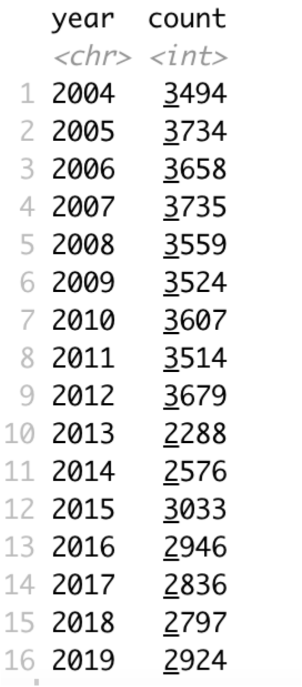
```
   year  total_coworkers
   <chr>           <int>
 1 2004             4275
 2 2005             4717
 3 2006             4870
 4 2007             4996
 5 2008             5010
 6 2009             4919
 7 2010             5189
 8 2011             5206
 9 2012             5568
10 2013             5280
11 2014             5327
12 2015             5341
13 2016             5300
14 2017             5028
15 2018             4994
16 2019             5307
```

8. The following table shows number of individuals in the panel that are from household-Couple with kids each year.



```
   year  count
   <chr> <int>
 1 2004  11993
 2 2005  13217
 3 2006  13637
 4 2007  13963
 5 2008  13481
 6 2009  13286
 7 2010  13726
 8 2011  13801
 9 2012  14403
10 2013  13103
11 2014  13228
12 2015  13008
13 2016  12967
14 2017  11963
15 2018  11444
16 2019  12151
```

9. The following table shows number of individuals in the panel that are from Paris each year.



```
   year  count
   <chr> <int>
 1 2004   3494
 2 2005   3734
 3 2006   3658
 4 2007   3735
 5 2008   3559
 6 2009   3524
 7 2010   3607
 8 2011   3514
 9 2012   3679
10 2013   2288
11 2014   2576
12 2015   3033
13 2016   2946
14 2017   2836
15 2018   2797
16 2019   2924
```

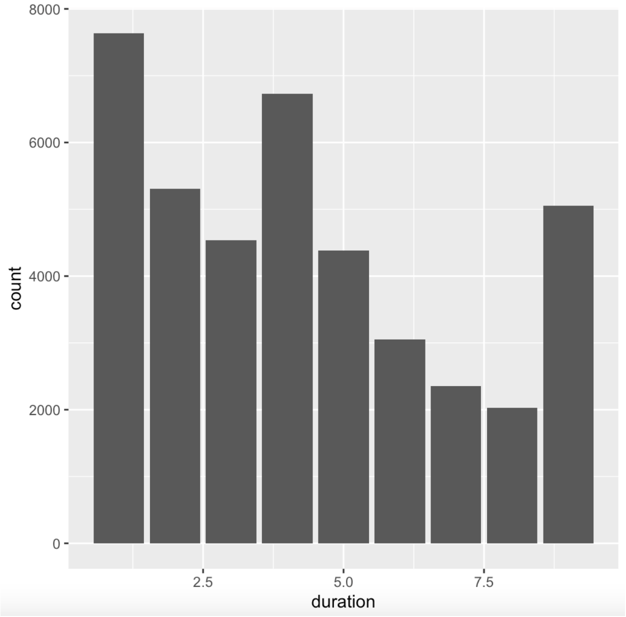
10. Find the household with the most number of family members each year. Report its idmen:

| year | idmen | HH_num |
|---|---:|---:|
| 2004 | 1208045118450100 | 10 |
| 2004 | 1607839058220100 | 10 |
| 2004 | 1610263040580100 | 10 |
| 2004 | 1804363114960100 | 10 |
| 2005 | 1607839058220100 | 11 |
| 2006 | 1607839058220100 | 10 |
| 2006 | 1811109095380100 | 10 |
| 2007 | 2207811124040100 | 14 |
| 2008 | 1700707001000100 | 10 |
| 2008 | 1811109095380100 | 10 |
| 2008 | 2006865025180100 | 10 |
| 2009 | 1700707001000100 | 11 |
| 2010 | 2510263102990100 | 14 |
| 2011 | 1905191114960100 | 10 |
| 2011 | 2202243098040100 | 10 |
| 2012 | 1905191114960100 | 10 |
| 2012 | 2202243098040100 | 10 |
| 2013 | 2202243098040100 | 10 |
| 2014 | 2106457101960100 | 9 |
| 2014 | 2200896118640100 | 9 |
| 2014 | 2209201025180100 | 9 |
| 2014 | 2701042078730100 | 9 |
| 2014 | 2707811117610100 | 9 |
| 2014 | 2710263020060100 | 9 |
| 2014 | 2905191059550100 | 9 |
| 2014 | 2905459051770100 | 9 |
| 2015 | 3000896115750100 | 12 |
| 2016 | 3000896115750100 | 12 |
| 2017 | 3000896115750100 | 12 |
| 2018 | 3000896115750100 | 11 |
| 2019 | 2806477001000100 | 9 |
| 2019 | 3200528124040100 | 9 |
| 2019 | 3300896124060100 | 9 |
| 2019 | 3402178051020100 | 9 |

11. If we consider the union, there are 13424 distinct households present in either 2010 or 2011. If we consider the intersection, there are 8984 households present both in 2010 and 2011.

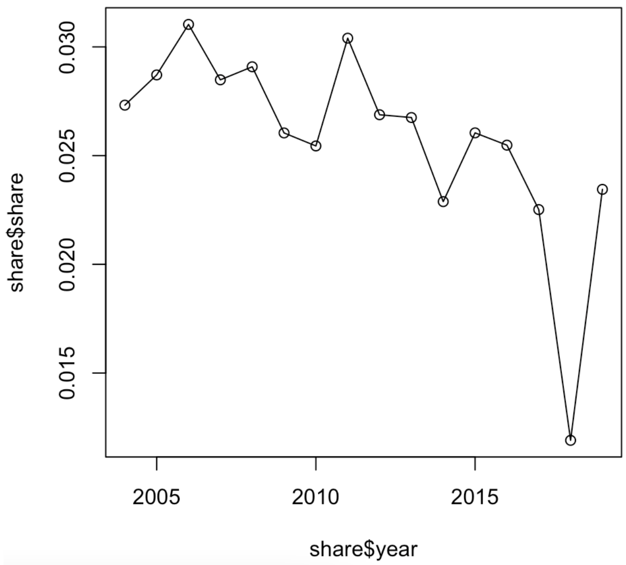
## 3 Exercise 3: Migration

1. Report the distribution of the time spent in the survey. The max value of duration is 9 years and the minimum is 1 year. If respondent enter or exit the survey in the same year, we consider the minimum duration as 1 year.



2. Report the first 10 rows of your result of whether or not a household moved into its current dwelling at the year of survey.

| | idmen | year | moved |
|---|---|---|---|
| 1 | 1200010012930100 | 2004 | 0 |
| 2 | 1200010040580100 | 2004 | 0 |
| 3 | 1200010040580100 | 2004 | 0 |
| 4 | 1200010040580100 | 2005 | 0 |
| 5 | 1200010040580100 | 2005 | 0 |
| 6 | 1200010066630100 | 2004 | 0 |
| 7 | 1200010066630100 | 2004 | 0 |
| 8 | 1200010066630100 | 2005 | 1 |
| 9 | 1200010066630100 | 2005 | 1 |
| 10 | 1200010082450100 | 2004 | 0 |

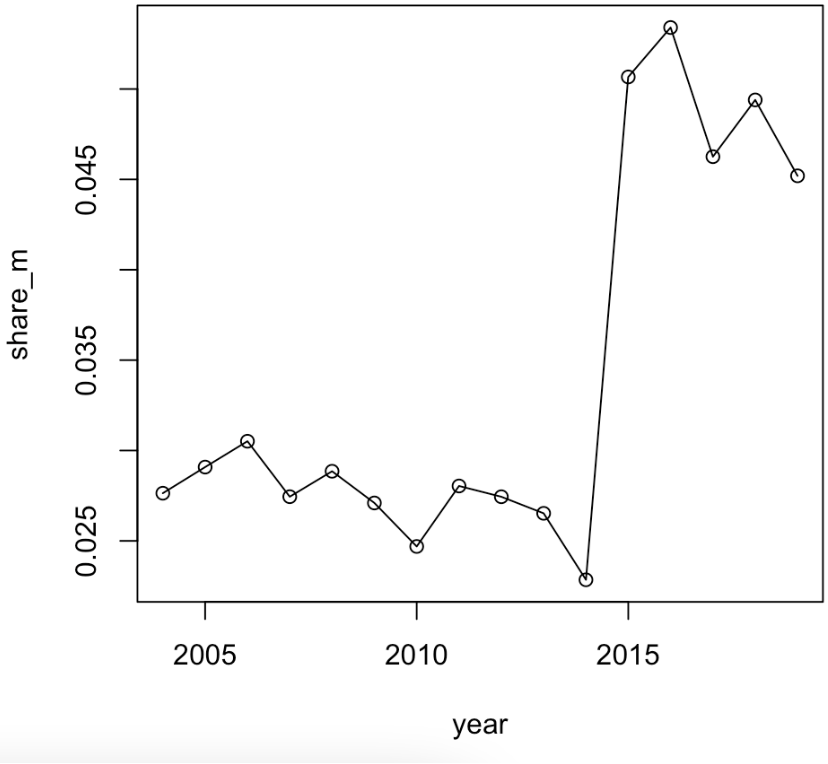


3. Based on myear and move, identify whether or not household migrated at the year of survey. Report the first 10 rows of your result. The last column migrate is true, which means the household migrated at the year of survey. I define migration as household either myear equals to year before 2014 and moved equals to 2 after 2014.

| | idmen | year | migrate |
|---|---|---|---|
| 1 | 1200010012930100 | 2004 | FLASE |
| 2 | 1200010040580100 | 2004 | FLASE |
| 3 | 1200010040580100 | 2004 | FLASE |
| 4 | 1200010040580100 | 2005 | FLASE |
| 5 | 1200010040580100 | 2005 | FLASE |
| 6 | 1200010066630100 | 2004 | FLASE |
| 7 | 1200010066630100 | 2004 | FLASE |
| 8 | 1200010066630100 | 2005 | TRUE |
| 9 | 1200010066630100 | 2005 | TRUE |
| 10 | 1200010082450100 | 2004 | FLASE |

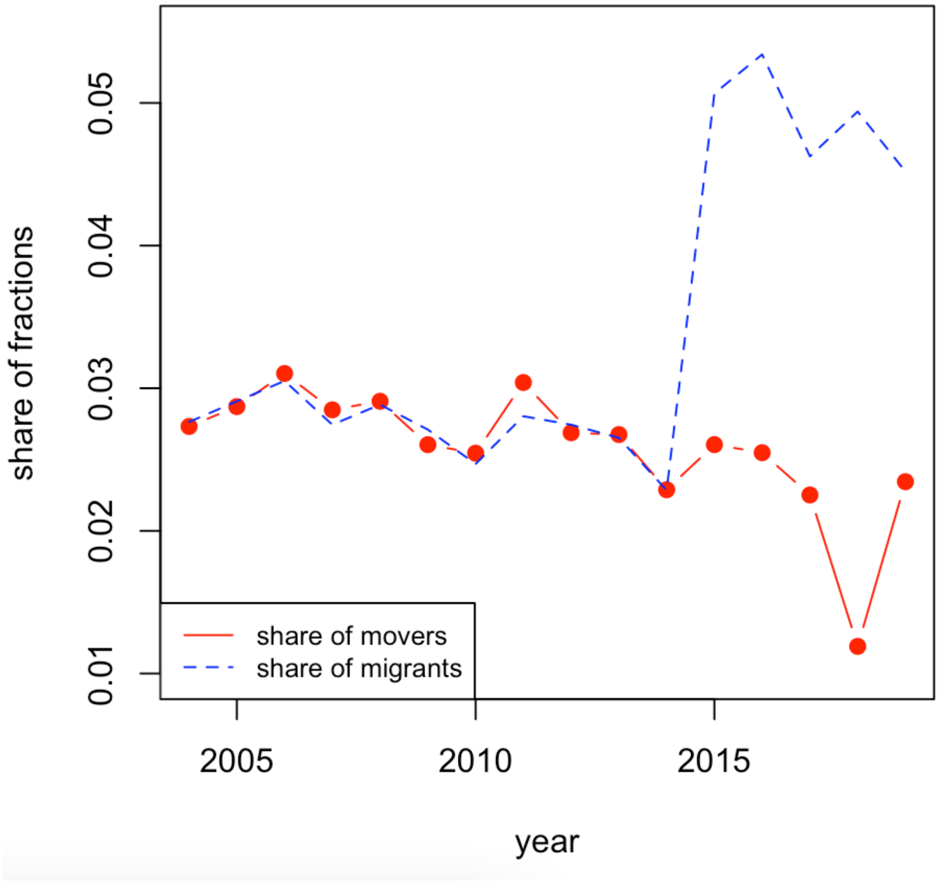
plot the share of individuals in that situation across years.

Figure 4. Share of migrants



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

   Personally, I prefer mixing the two plots together in one graph. I could observe the disparities and similarities of these two share. Also, their overall trends are very clear in the combined graph.



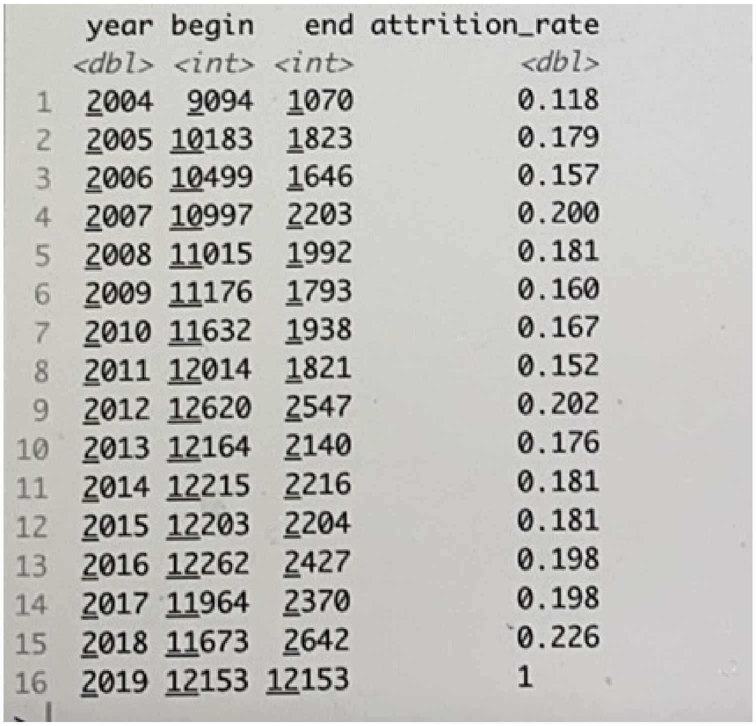
5. There are 227 distinct households had at least one family member changed his/her profession or employment status over the whole sample period.. We don't want to consider missing value in either profession or empstat. Thus we drop any rows with missing in ourdata. The reason for that is we do not consider missing value as the same profession/empstat.

# 4 Exercise 4: Attrition

## 4.1 Version I: individual ID as a list, counting for reentry

| | year | att_rate |
|---|---|---|
| 1 | 2004.00 | 13.53 |
| 2 | 2005.00 | 20.01 |
| 3 | 2006.00 | 17.47 |
| 4 | 2007.00 | 22.57 |
| 5 | 2008.00 | 20.73 |
| 6 | 2009.00 | 18.82 |
| 7 | 2010.00 | 19.58 |
| 8 | 2011.00 | 16.78 |
| 9 | 2012.00 | 25.53 |
| 10 | 2013.00 | 22.21 |
| 11 | 2014.00 | 21.24 |
| 12 | 2015.00 | 21.75 |
| 13 | 2016.00 | 25.09 |
| 14 | 2017.00 | 24.43 |
| 15 | 2018.00 | 24.30 |

## 4.2 Version II: exit and entry year



```
    year begin   end attrition_rate
   <dbl> <int> <int>          <dbl>
 1  2004  9094  1070          0.118
 2  2005 10183  1823          0.179
 3  2006 10499  1646          0.157
 4  2007 10997  2203          0.200
 5  2008 11015  1992          0.181
 6  2009 11176  1793          0.160
 7  2010 11632  1938          0.167
 8  2011 12014  1821          0.152
 9  2012 12620  2547          0.202
10  2013 12164  2140          0.176
11  2014 12215  2216          0.181
12  2015 12203  2204          0.181
13  2016 12262  2427          0.198
14  2017 11964  2370          0.198
15  2018 11673  2642          0.226
16  2019 12153 12153          1
```