HW_1
Shiqi Zhou

Exercise 1
#1.1 Number of households surveyed in 2007

```
> nrow(dathh2007)
[1] 10498
```

#1.2 Number of households with marital status “Couple with kids” in 2005

```
> typeof(dathh2005$mstatus)
[1] "character"
> nrow(filter(dathh2005,mstatus=='Couple, with Kids'))
[1] 3374
```

#1.3 Number of individuals surveyed in 2008

```
> nrow(datind2008)
[1] 25510
```

#1.4 Number of individuals aged between 25 and 35 in 2016

```
> nrow(filter(datind2016,age >= 25, age <= 35))
[1] 2765
```

#1.5 Cross-table gender/profession in 2009

#way 1: Use CrossTable

```
   Cell Contents
|-------------------------|
|                       N |
|-------------------------|


Total Observations in Table:  10495

                     | datind2009$gender
datind2009$profession |    Female |      Male | Row Total |
----------------------|-----------|-----------|-----------|
                    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 |
----------------------|-----------|-----------|-----------|
         Column Total |      5117 |      5378 |     10495 |
----------------------|-----------|-----------|-----------|
```

#way 2: Count while grouping by gender & profession

	gender	profession	n
1:	Female	0	11
2:	Female	11	30
3:	Female	12	8
4:	Female	13	29
5:	Female	21	63
6:	Female	22	65
7:	Female	23	8
8:	Female	31	68
9:	Female	33	85
10:	Female	34	184
11:	Female	35	50
12:	Female	37	179
13:	Female	38	78
14:	Female	42	258
15:	Female	43	437
16:	Female	44	1
17:	Female	45	153
18:	Female	46	410
19:	Female	47	82
20:	Female	48	22
21:	Female	52	782
22:	Female	53	27
23:	Female	54	584
24:	Female	55	353
25:	Female	56	696
26:	Female	62	64
27:	Female	63	35
28:	Female	64	29
29:	Female	65	19
30:	Female	67	147
31:	Female	68	120
32:	Female	69	40
33:	Female	NA	8167
34:	Male	0	19
35:	Male	11	57
36:	Male	12	19
37:	Male	13	78
38:	Male	21	213
39:	Male	22	114
40:	Male	23	48
41:	Male	31	98
42:	Male	33	107
43:	Male	34	142
44:	Male	35	59
45:	Male	37	260
46:	Male	38	368
47:	Male	42	110
48:	Male	43	117
49:	Male	44	2
50:	Male	45	95
51:	Male	46	340
52:	Male	47	429
53:	Male	48	215
54:	Male	52	169
55:	Male	53	182
56:	Male	54	98
57:	Male	55	101
58:	Male	56	74
59:	Male	62	443
60:	Male	63	520
61:	Male	64	246
62:	Male	65	159
63:	Male	67	237
64:	Male	68	177
65:	Male	69	82
66:	Male	NA	6949
	gender	profession	n

#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

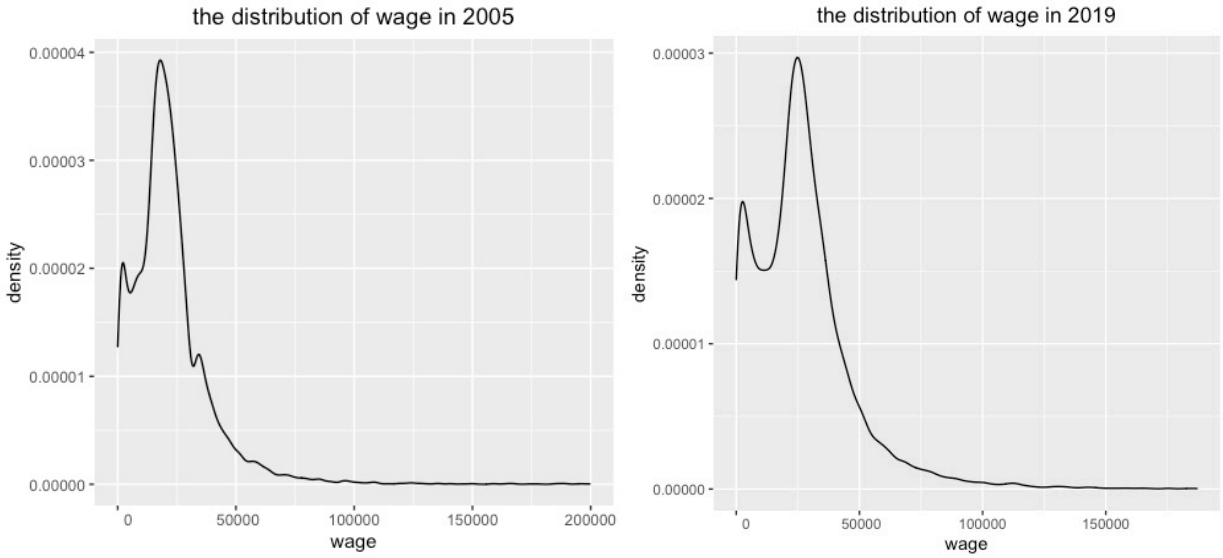
#exclude NA, and get mean, sd, decile

```
> mean(datind2005$wage,na.rm=T)
[1] 11992.26
> mean(datind2019$wage,na.rm=T)
[1] 15350.47
> sd(datind2005$wage,na.rm=T)
[1] 17318.56
> sd(datind2019$wage,na.rm=T)
[1] 23207.18
> quantile(datind2005$wage, probs = seq(0, 1, 0.1), na.rm =T,names = TRUE)
     0%      10%      20%      30%      40%      50%      60%      70%      80%      90%     100%
    0.0      0.0      0.0      0.0      0.0   2444.0  12503.2  18079.2  23084.0  32340.4 271962.0
> quantile(datind2019$wage, probs = seq(0, 1, 0.1), na.rm =T,names = TRUE)
    0%     10%     20%     30%     40%     50%     60%     70%     80%     90%    100%
     0       0       0       0       0    3710   15369   23550   29744   40267 1068556
```

#with the results below, there are many wage==0, exclude those.
#redo this part, exclude wage==0, wage==NA

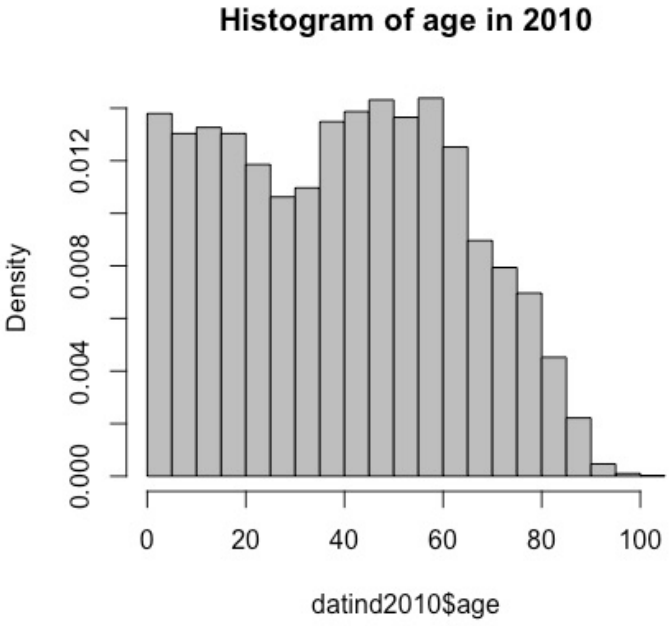
```
> # mean
> mean(wageind_2005)
[1] 22443.03
> mean(wageind_2019)
[1] 27578.84
> # sd
> sd(wageind_2005)
[1] 18076.71
> sd(wageind_2019)
[1] 25107.19
> # decile, and D9/D1
> wageq_2005 <- quantile(wageind_2005, probs = seq(0, 1, 0.1), na.rm =T,names = TRUE)
> wageq_2019 <- quantile(wageind_2019, probs = seq(0, 1, 0.1), na.rm =T,names = TRUE)
> wageq_2005[[10]]/wageq_2005[[2]]
[1] 8.896525
> wageq_2019[[10]]/wageq_2019[[2]]
[1] 13.8623
> gini_eff(wageind_2005)
[1] 0.3771135
> gini_eff(wageind_2019)
[1] 0.3990875
```

#draw the distribution, exclude wage==0, wage==NA, and extremely large values

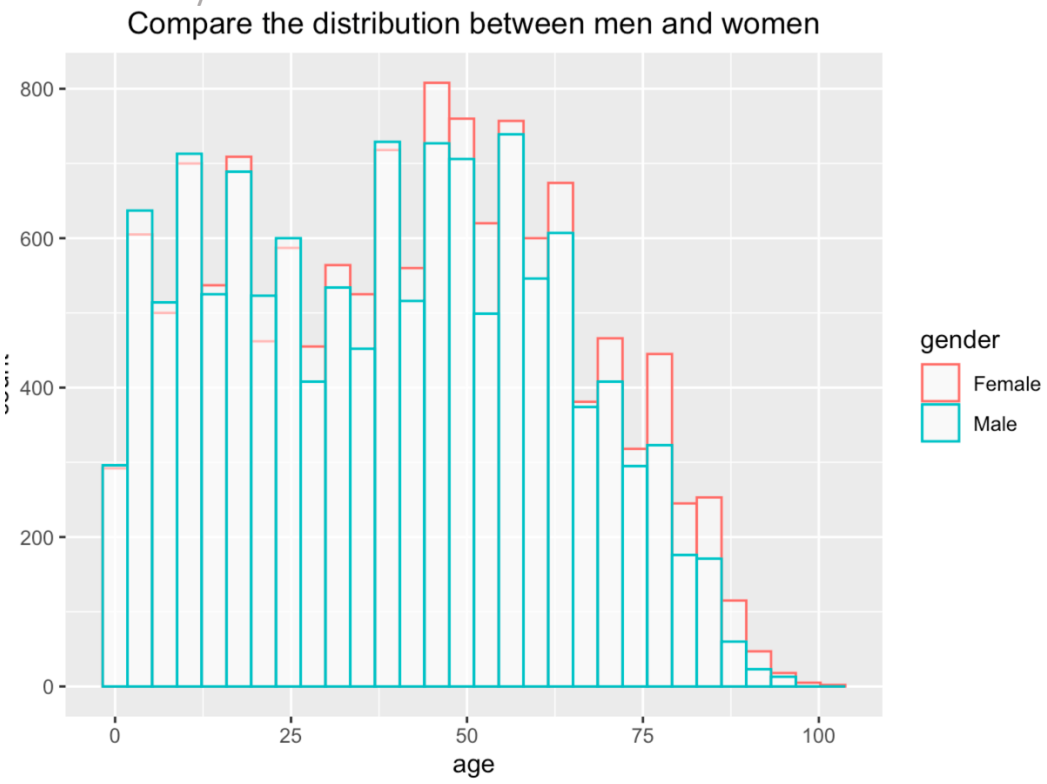


#1.7 Distribution of age in 2010. Plot a histogram.

```
> mean(datind2010$age,na.rm=T)
[1] 39.87893
> sd(datind2010$age,na.rm=T)
[1] 23.42486
```



Is there any difference between men and women?



#1.8 Number of individuals in Paris in 2011

```
> nrow(filter(mer.2011,location=='Paris'))
[1] 3514
```

Exercise 2

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

```
> intersect(ind_name,hh_name)
[1] "idmen" "year"
```

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

#count by year

```
# A tibble: 16 × 2
# Groups:   year [16]
    year     n
   <int> <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
```

total number

```
> sum(a1$n)
[1] 12436
```

2.6 Number of households in which at least one member is unemployed

#count by year

```
# A tibble: 16 × 2
# Groups:   year [16]
    year     n
   <int> <int>
 1  2004   950
 2  2005  1039
 3  2006  1030
 4  2007   975
 5  2008   909
 6  2009  1045
 7  2010  1110
 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
```

total number

```
> sum(b1$n)
[1] 17242
```

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

#count by year

	year	n
	<int>	<int>
1	2004	445
2	2005	497
3	2006	485
4	2007	492
5	2008	460
6	2009	453
7	2010	477
8	2011	492
9	2012	517
10	2013	460
11	2014	477
12	2015	469
13	2016	475
14	2017	459
15	2018	457
16	2019	500

total number

```
> sum(c1$n)
[1] 7615
```

2.8 Number of individuals in the panel that are from household-Couple with kids

#count by year

	year	mstatus	nmem
1	2004	Couple, with Kids	11993
2	2005	Couple, with Kids	13217
3	2006	Couple, with Kids	13637
4	2007	Couple, with Kids	13963
5	2008	Couple, with Kids	13481
6	2009	Couple, with Kids	13286
7	2010	Couple, with Kids	13726
8	2011	Couple, with Kids	13801
9	2012	Couple, with Kids	14403
10	2013	Couple, with Kids	13114
11	2014	Couple, with Kids	13228
12	2015	Couple, with Kids	13008
13	2016	Couple, with Kids	12967
14	2017	Couple, with Kids	11963
15	2018	Couple, with Kids	11444
16	2019	Couple, with Kids	12151

total number

```
> sum(d1$nmem)
[1] 209382
```

2.9 Number of individuals in the panel that are from Paris

#count by year

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

total number

```
> sum(e1$nmem)
[1] 51904
```

2.10 Find the household with the most number of family members.

#report those household that have the most number of family members by year

```
# A tibble: 34 × 3
# Groups:   year [16]
    year idmen             nmem
   <int> <chr>            <int>
 1  2004 1208045118450100    10
 2  2004 1607839058220100    10
 3  2005 1607839058220100    11
 4  2006 1607839058220100    10
 5  2004 1610263040580100    10
 6  2008 1700707001000100    10
 7  2009 1700707001000100    11
 8  2004 1804363114960100    10
 9  2006 1811109095380100    10
10  2008 1811109095380100    10
# … with 24 more rows
```

not by year

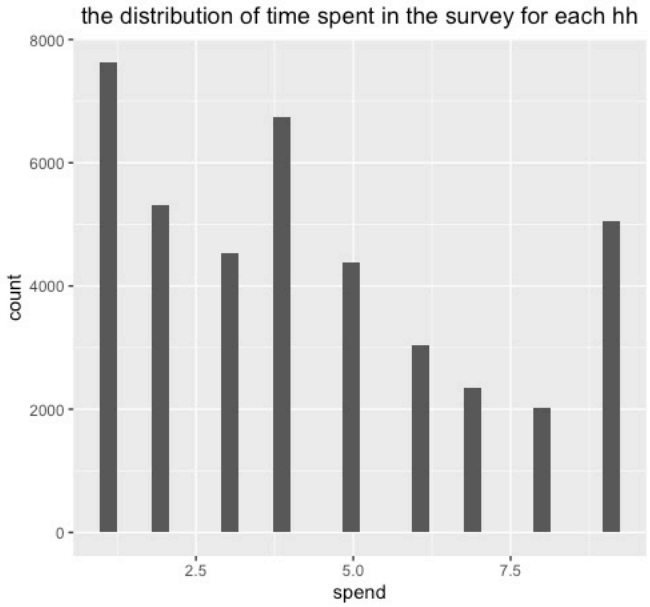
```
  year             idmen nmem
1 2007 2207811124040100   14
2 2010 2510263102990100   14
```

2.11 Number of households present in 2010 and 2011

```
> nrow(g1)
[1] 8984
```

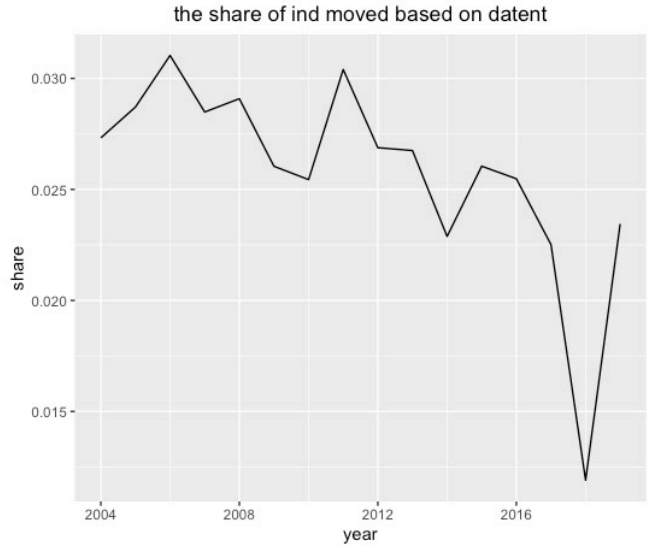
Exercise 3 Migration

3.1 Find out the year each household enters and exit the panel. Report the distribution of the time spent in the survey for each household.



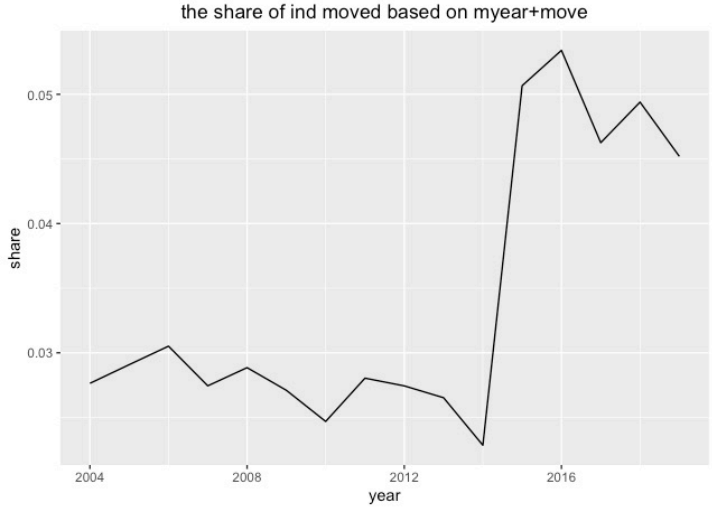
3.2 Based on datent, identify whether or not a household moved into its dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

```
> head(filter(hhind,ymd1==1),n=10)
              idind            idmen year    empstat respondent profession gender age  wage datent myear          mstatus move             location m_d1 ymd1
 1: 1120049301027010001 1200493010270100 2004 Unemployed          1            Female  36     0   2004  2004 Couple, with Kids   NA                Rural    0    1
 2: 1120049301027010002 1200493010270100 2004   Employed          0         68   Male  31     0   2004  2004 Couple, with Kids   NA                Rural    0    1
 3: 1120049301027010003 1200493010270100 2004   Inactive          0            Female   8    NA   2004  2004 Couple, with Kids   NA                Rural    0    1
 4: 1120049301027010004 1200493010270100 2004   Inactive          0            Female   8    NA   2004  2004 Couple, with Kids   NA                Rural    0    1
 5: 1120074202054010001 1200742020540100 2004   Employed          1         67   Male  29 16106   2004  2004   Couple, No kids   NA Urban 10000 to 19999    0    1
 6: 1120074202054010002 1200742020540100 2004   Employed          0         56 Female  23 15180   2004  2004   Couple, No kids   NA Urban 10000 to 19999    0    1
 7: 1120089601262010001 1200896012620100 2004   Employed          1         55   Male  36 31783   2004  2004            Single   NA                Paris    0    1
 8: 1120089808968010001 1200898089680100 2004    Retired          1            Female  55 24258   2004  1977   Couple, No kids   NA                Rural    0    1
 9: 1120089808968010002 1200898089680100 2004    Retired          0            Male    56  7453   2004  1977   Couple, No kids   NA                Rural    0    1
10: 1120138606786010001 1201386067860100 2004   Employed          1         43 Female  44 27051   2004  2004            Single   NA                Paris    0    1
>
```



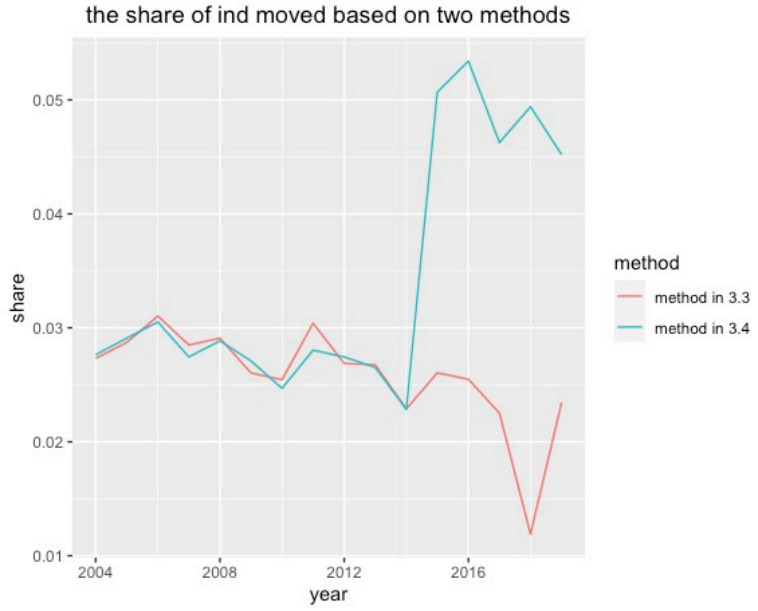
3.3 Based on myear and move, identify whether or not household migrated at the year of survey. the first 10 rows of your result and plot the share of individuals in that situation across years.

```
> head(filter(hhind,ymd4==1),n=10)
           idind          idmen year    empstat respondent profession gender age  wage datent myear          mstatus move             location m_d1 ymd1 ymd2 ymd3 ymd4
 1: 1120049301027010001 1200493010270100 2004 Unemployed          1            Female  36     0   2004  2004 Couple, with Kids   NA                Rural    0    1    1   NA    1
 2: 1120049301027010002 1200493010270100 2004   Employed          0         68   Male  31     0   2004  2004 Couple, with Kids   NA                Rural    0    1    1   NA    1
 3: 1120049301027010003 1200493010270100 2004   Inactive          0            Female   8    NA   2004  2004 Couple, with Kids   NA                Rural    0    1    1   NA    1
 4: 1120049301027010004 1200493010270100 2004   Inactive          0            Female   8    NA   2004  2004 Couple, with Kids   NA                Rural    0    1    1   NA    1
 5: 1120074202054010001 1200742020540100 2004   Employed          1         67   Male  29 16106   2004  2004   Couple, No kids   NA Urban 10000 to 19999    0    1    1   NA    1
 6: 1120074202054010002 1200742020540100 2004   Employed          0         56 Female  23 15180   2004  2004   Couple, No kids   NA Urban 10000 to 19999    0    1    1   NA    1
 7: 1120089601262010001 1200896012620100 2004   Employed          1         55   Male  36 31783   2004  2004           Single   NA                Paris    0    1    1   NA    1
 8: 1120138606786010001 1201386067860100 2004   Employed          1         43 Female  44 27051   2004  2004           Single   NA                Paris    0    1    1   NA    1
 9: 1120138610658010001 1201386106580100 2004   Employed          1         53   Male  45 13825   2004  2004     Single Parent   NA                Paris    0    1    1   NA    1
10: 1120138610658010002 1201386106580100 2004   Inactive          0            Male  15    NA   2004  2004     Single Parent   NA                Paris    0    1    1   NA    1
```



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

I prefer the method in 3.3. It used the same variable to estimate the share, but in 3.4, there are two variables used to measure the moved share and it increase sharply in 2014 where the used variable changed.



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

	year	nmem
1	2005	522
2	2006	585
3	2007	522
4	2008	547
5	2009	478
6	2010	461
7	2011	578
8	2012	563
9	2013	503
10	2014	463
11	2015	1134
12	2016	1178
13	2017	1004
14	2018	1002
15	2019	994

Exercise 4 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. (Hint: Construct a year of entry and exit for each individual.)

	exit	`2004`	`2005`	`2006`	`2007`	`2008`	`2009`	`2010`	`2011`	`2012`	`2013`	`2014`	`2015`	`2016`	`2017`	`2018`
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2004	0.104	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	2005	0.165	0.169	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	2006	0.135	0.131	0.150	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	2007	0.188	0.180	0.184	0.197	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	2008	0.127	0.142	0.150	0.152	0.176	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	2009	0.0912	0.0969	0.119	0.126	0.138	0.156	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	2010	0.0777	0.0838	0.0985	0.121	0.133	0.141	0.158	NA	NA	NA	NA	NA	NA	NA	NA
8	2011	0.0588	0.0598	0.0676	0.0809	0.103	0.113	0.121	0.142	NA	NA	NA	NA	NA	NA	NA
9	2012	0.0547	0.0686	0.0810	0.0980	0.125	0.158	0.168	0.174	0.193	NA	NA	NA	NA	NA	NA
10	2013	NA	0.0694	0.0773	0.0830	0.0993	0.115	0.138	0.143	0.147	0.177	NA	NA	NA	NA	NA
11	2014	NA	NA	0.0729	0.0762	0.0871	0.0991	0.112	0.135	0.138	0.147	0.177	NA	NA	NA	NA
12	2015	NA	NA	NA	0.0657	0.0733	0.0818	0.0926	0.108	0.130	0.141	0.151	0.178	NA	NA	NA
13	2016	NA	NA	NA	NA	0.0663	0.0728	0.0804	0.0958	0.113	0.147	0.160	0.170	0.200	NA	NA
14	2017	NA	NA	NA	NA	NA	0.0632	0.0691	0.0771	0.0872	0.111	0.141	0.155	0.165	0.198	NA
15	2018	NA	NA	NA	NA	NA	NA	0.0605	0.0715	0.0816	0.0981	0.116	0.149	0.162	0.185	0.226
16	2019	NA	NA	NA	NA	NA	NA	NA	0.0542	0.110	0.179	0.256	0.349	0.474	0.617	0.774