Assignment One

Zhishuo Han

Exercise 1:

Q1: Number of households surveyed in 2007.

10498

Q2: Number of households with marital status “Couple with kids” in 2005.

3374

Q3: Number of individuals surveyed in 2008.

25510

Q4: Number of individuals aged between 25 and 35 in 2016.

2765

Q5: Cross-table gender/profession in 2009. (part)

```
> datind2009 %>% count(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
```

Q6: Distribution of wages in 2005 and 2019. Report the mean, the standard deviation, the inter-decile ratio D9/D1 and the Gini coefficient.

2005:

Mean: 11992.26

SD: 17318.56

Inter-decile ratio: Infinite

GINI coefficient: 0.6671654

2019:

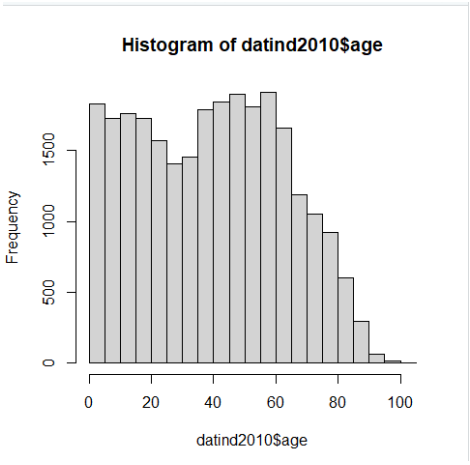
Mean: 15350.47

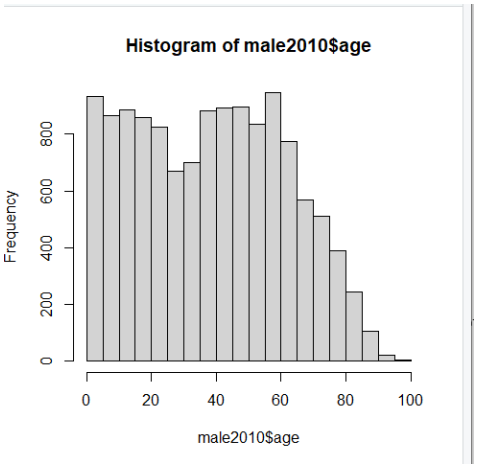
SD: 23207.18

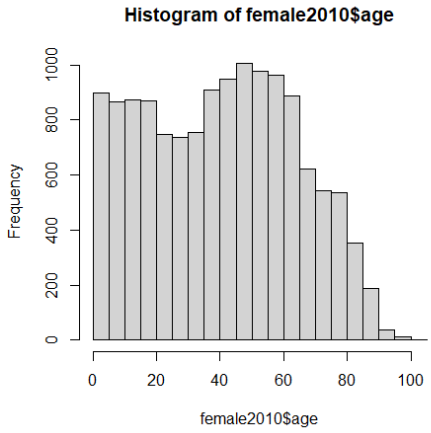
Inter-decile ratio: Infinite

GINI coefficient: 0.6655301

Q7: Distribution of age in 2010. Plot an histogram. Is there any difference between men and women?







Difference between men and women:

Looks like female has more observations in the interval between 40 and 60.

Q8: Number of individuals in Paris in 2011. 3514

Exercise 2:

Q1: Read all individual datasets from 2004 to 2019. Append all these datasets.

See the code. Dataset named “ind”

Q2: Read all household datasets from 2004 to 2019. Append all these datasets.

See the code. Dataset named “hh”

Q3: List the variables that are simultaneously present in the individual and household datasets.

"V1" "idmen" "year"

Q4: Merge the appended individual and household datasets.

See the code. Dataset named “ind_hh”

Q5: Number of households in which there are more than four family members

3622

Q6: Number of households in which at least one member is unemployed

8162

Q7: Number of households in which at least two members are of the same profession

8752

Q8: Number of individuals in the panel that are from household-Couple with kids

55094

Q9: Number of individuals in the panel that are from Paris.

14563

Q10: Find the household with the most number of family members. Report its idmen.

"2207811124040100" "2510263102990100"

Q11: Number of households present in 2010 and 2011.

22410

Exercise 3:

Q1: Find out the year each household enters and exit the panel. Report the distribution of the time spent in the survey for each household.

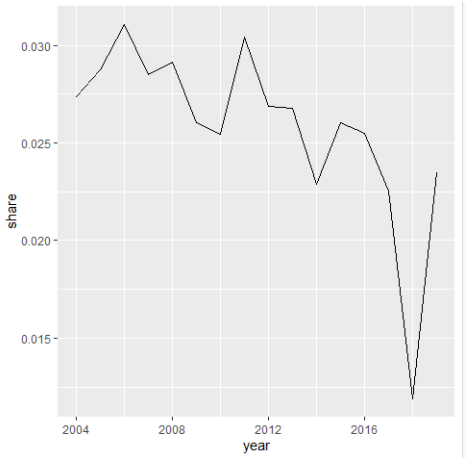
idmen	time_spent
1200010012930100	0
1200010040580100	1
1200010066630100	1
1200010082450100	1
1200010086440100	1
1200010102990100	1
1200010118450100	1
1200020012930100	1
1200020017390100	1
1200020026420100	1
1200020045130100	1
1200020094370100	1
1200020118450100	1
1200020122680100	1
1200149012930100	1
1200149034710100	1

(part)

See the code, table named “dist”

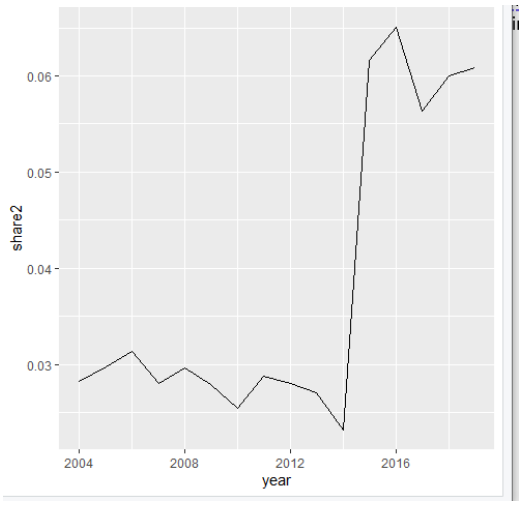
Q2: Based on *datent*, identify whether or not a household moved into its current dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

```
              idmen identify_dwelling
 1: 1200010012930100                0
 2: 1200010040580100                0
 3: 1200010040580100                0
 4: 1200010066630100                0
 5: 1200010066630100                0
 6: 1200010082450100                0
 7: 1200010086440100                0
 8: 1200010086440100                0
 9: 1200010102990100                0
10: 1200010102990100                0
```

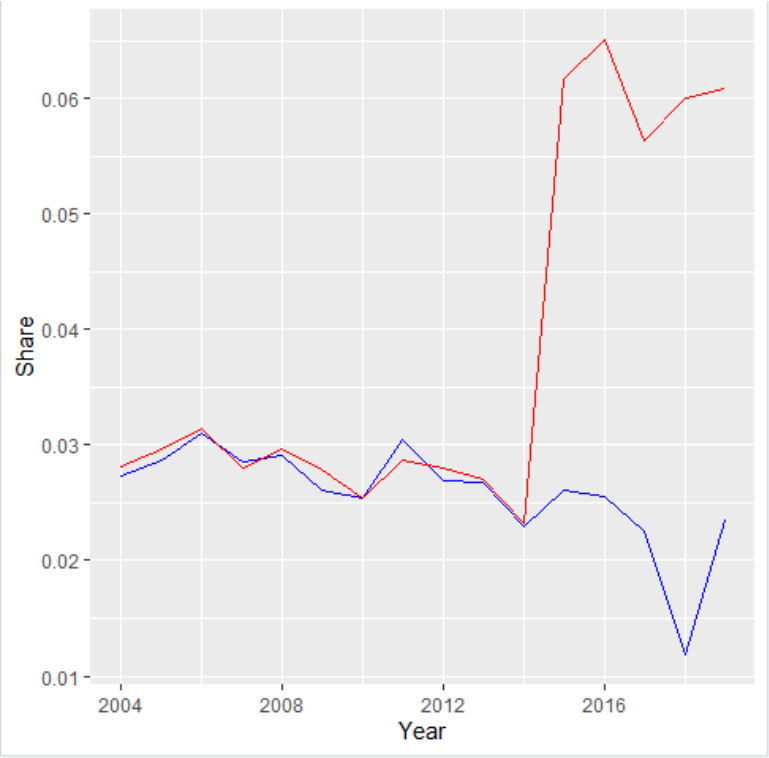


Q3: Based on *myear* and *move*, identify whether or not household migrated at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

```
> head(report_10_move, 10)
                idmen move
 1: 1200010012930100    1
 2: 1200010040580100    1
 3: 1200010040580100    1
 4: 1200010066630100    1
 5: 1200010066630100    1
 6: 1200010082450100    1
 7: 1200010086440100    1
 8: 1200010086440100    1
 9: 1200010102990100    1
10: 1200010102990100    1
```



Q4: 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 “datent” method. First, in the figure, we can see that, the red line (use move and myear) has a rapid increase in 2014; but the blue line (use datent) is relatively flat which implies a reasonable pattern of moving. Second, from the method, “move and myear” uses two different variables and combines them together. This may explain why the red line has a rapid increase after 2014: probably, these two variables include different information which cannot treat them as one variable.

Q5: For households who migrate, find out how many households had at least one family member changed his/her profession or employment status.

2245

Exercise 4:
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.

	year	proportion
	<int>	<dbl>
1	2005	0.296
2	2006	0.374
3	2007	0.698
4	2008	0.689
5	2009	0.704
6	2010	0.784
7	2011	0.733
8	2012	1.13
9	2013	1.28
10	2014	1.31
11	2015	1.44
12	2016	1.80
13	2017	2.29
14	2018	3.55