ECON 613 HW1

Yonghan Shi

1/15/2022

Exercise 1 Basic Statistics

- Number of households surveyed in 2007.

The number of households surveyed in 2007 is 10498.

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

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

- Number of individuals surveyed in 2008.

The number of individuals surveyed in 2008 is 25510.

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

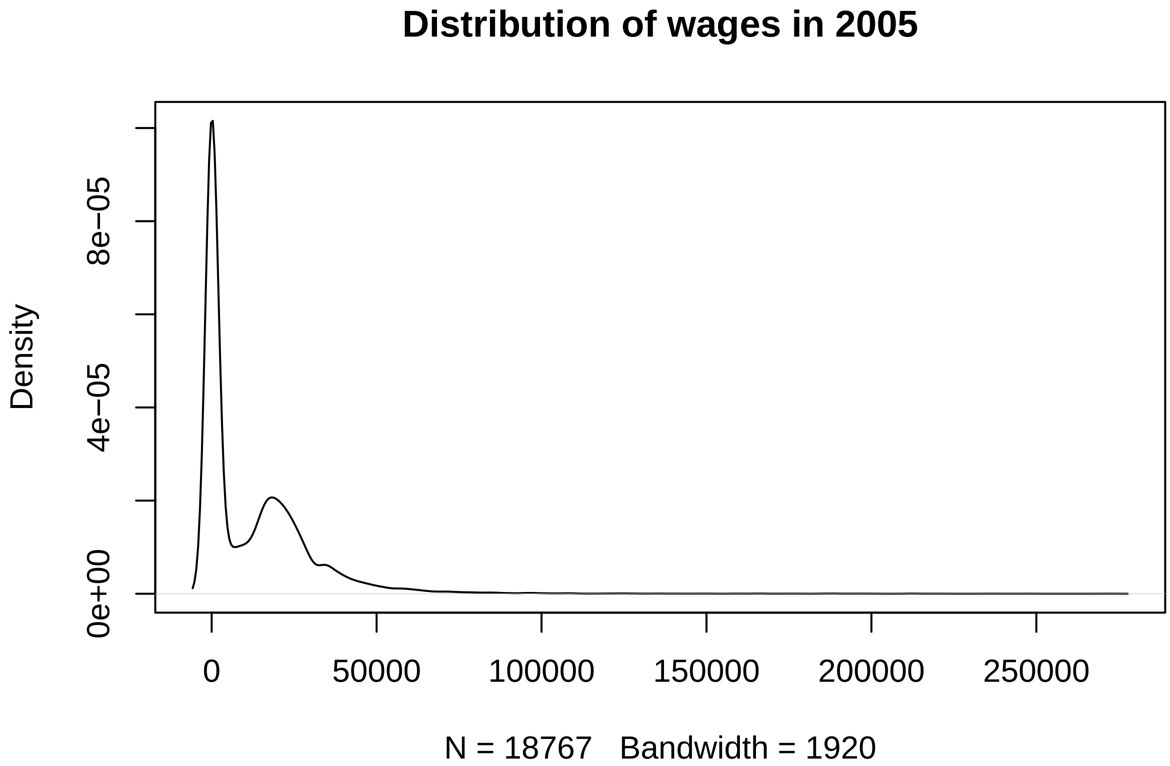
The number of individuals aged between 25 and 35 in 2016 is 2765.

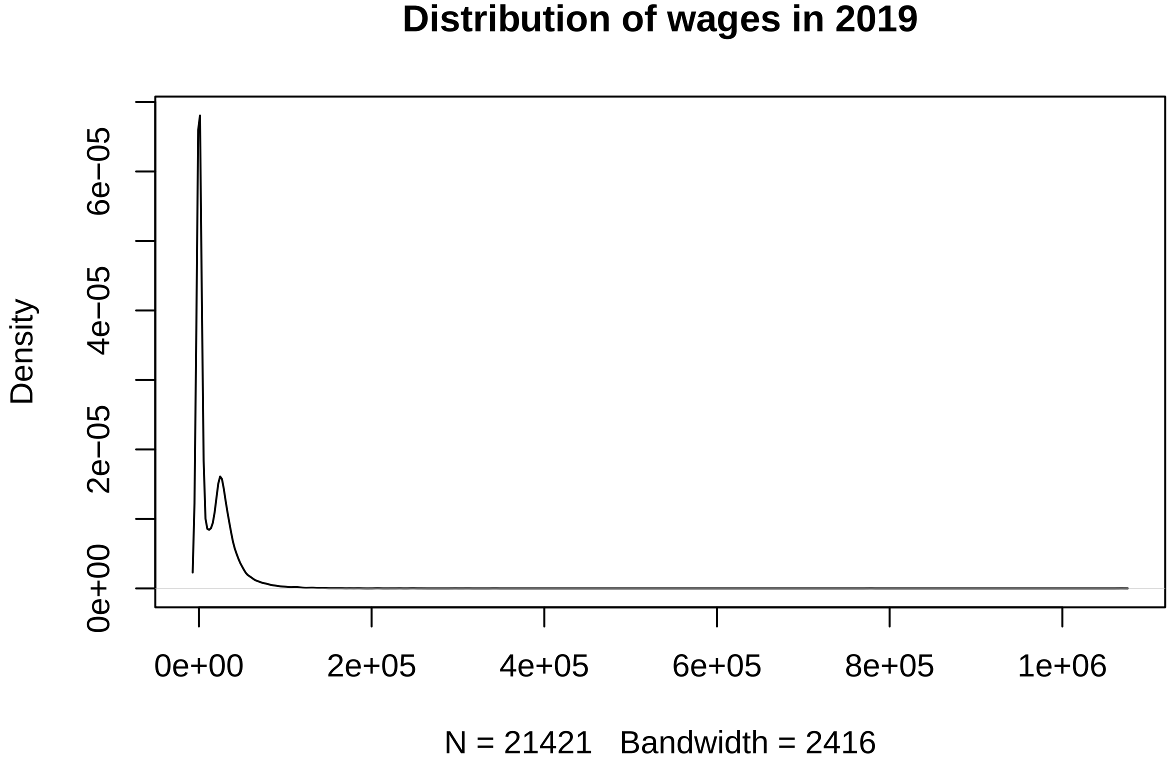
- Cross-table gender/profession in 2009.

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

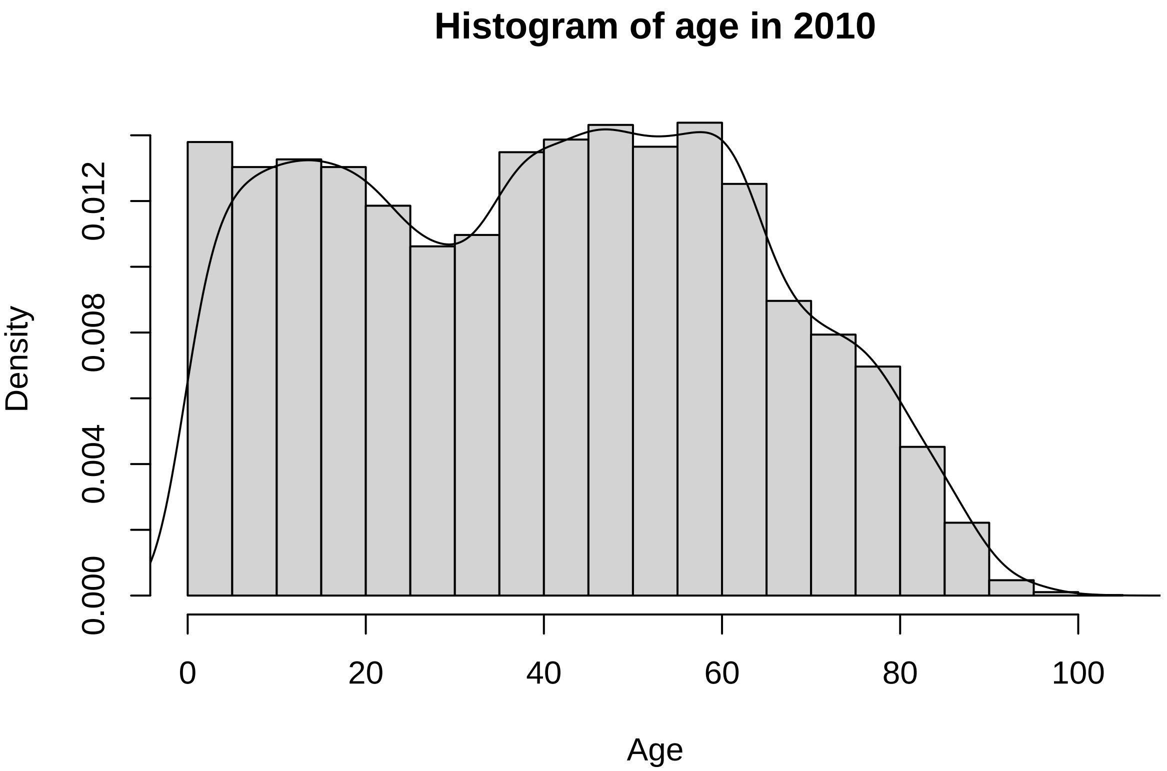
- Distribution of wages in 2005 and 2019. Report the mean, the standard deviation, the inter-decile ratio D9/D1 and the Gini coefficient.

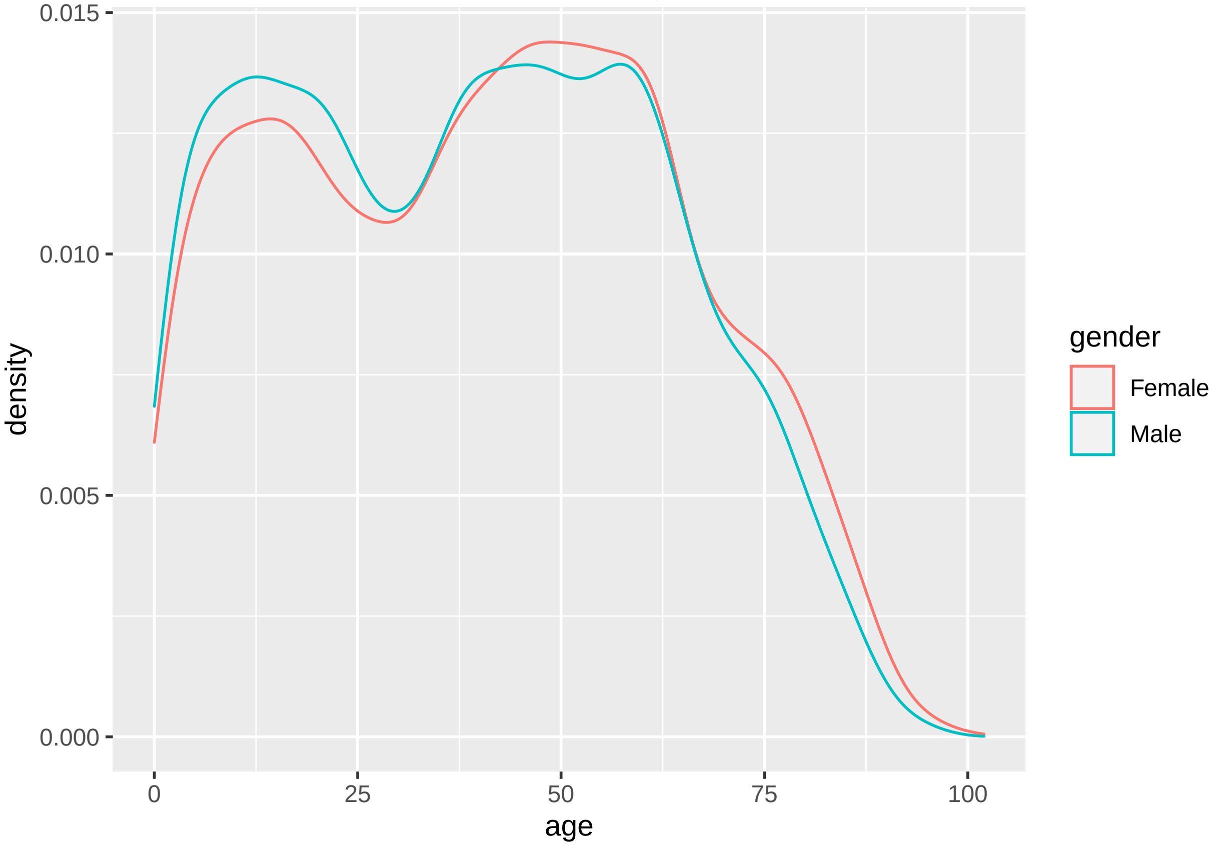
year	mean	sd	ratio	gini
2005	11992.26	17318.56	32340.4	0.6671654
2019	15350.47	23207.18	40267.0	0.6655301


Distribution of wages in 2005
Density
0e+00
4e−05
8e−05
0
50000
100000
150000
200000
250000
N = 18767   Bandwidth = 1920


Distribution of wages in 2019
Density
0e+00
2e−05
4e−05
6e−05
0e+00
2e+05
4e+05
6e+05
8e+05
1e+06
N = 21421   Bandwidth = 2416

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





The distribution is similar for men and women.

- Number of individuals in Paris in 2011.

The number of individuals in Paris in 2011 is 3514.

Exercise 2 Merge Datasets

- Read all individual datasets from 2004 to 2019. Append all these datasets.

```
> library(dplyr)
> library(plyr)
>
> filelist <- list.files("Data")
> filelist_sampleid <- as.matrix(gsub(".csv","", filelist))
>
> filelist2 <- filelist[17:32]
>
> files2 <- paste("./Data/", filelist2, sep = "")
>
> datind = ldply(files2, read_csv)
```

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

```
> filelist1 <- filelist[1:16]
>
> files1 <- paste("./Data/", filelist1, sep = "")
>
> dathh = ldply(files1, read_csv)
>
> detach("package:plyr", unload = T)
```

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

They are idmen, year.

- Merge the appended individual and household datasets.

```
> # delete column of serial number
> dathh <- dathh[2:8]
> datind <- datind[2:10]
>
> datall <- merge(dathh, datind, by = c("idmen", "year"), all = TRUE)
```

In the second part, we use the newly created dataset from the previous to answer the following questions:

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

The number of households in which there are more than four family members is 3622.

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

The number of households in which at least one member is unemployed is 35040.

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

Number of households in which at least two members are of the same profession is 3022.

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

year	2004	2005	2006	2007	2008	2009	2010	2011
number	3063	3523	3670	3826	3764	3740	3913	3934

year	2012	2013	2014	2015	2016	2017	2018	2019
number	4146	3782	3806	3754	3739	3445	3273	3435

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

year	2004	2005	2006	2007	2008	2009	2010	2011
number	1430	1541	1509	1561	1502	1529	1576	1552

year	2012	2013	2014	2015	2016	2017	2018	2019
number	1606	945	1090	1327	1306	1254	1225	1279

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

```
## [1] 2207811124040100 2510263102990100
```

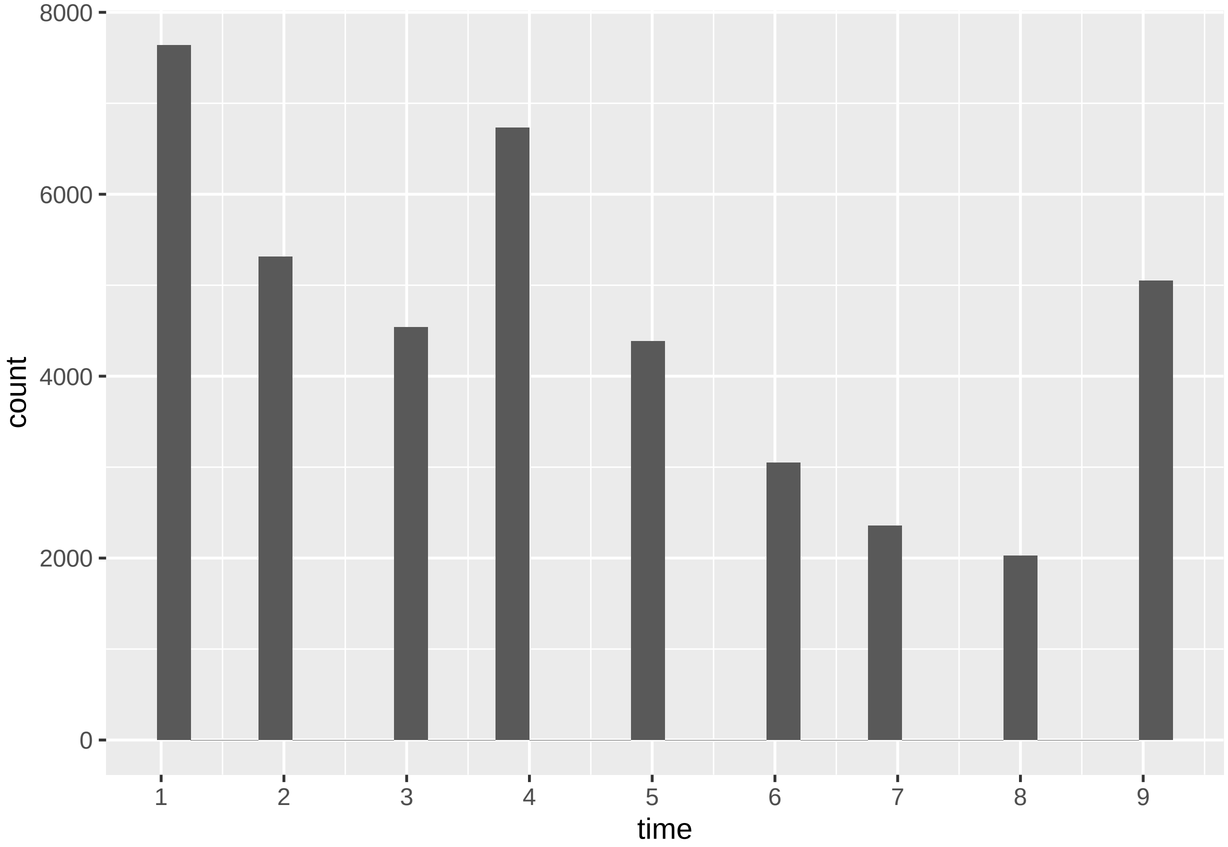
The households with the most number of family members are 2207811124040100, 2510263102990100.

- Number of households present in 2010 and 2011.

year	number
2010	9455
2011	9726

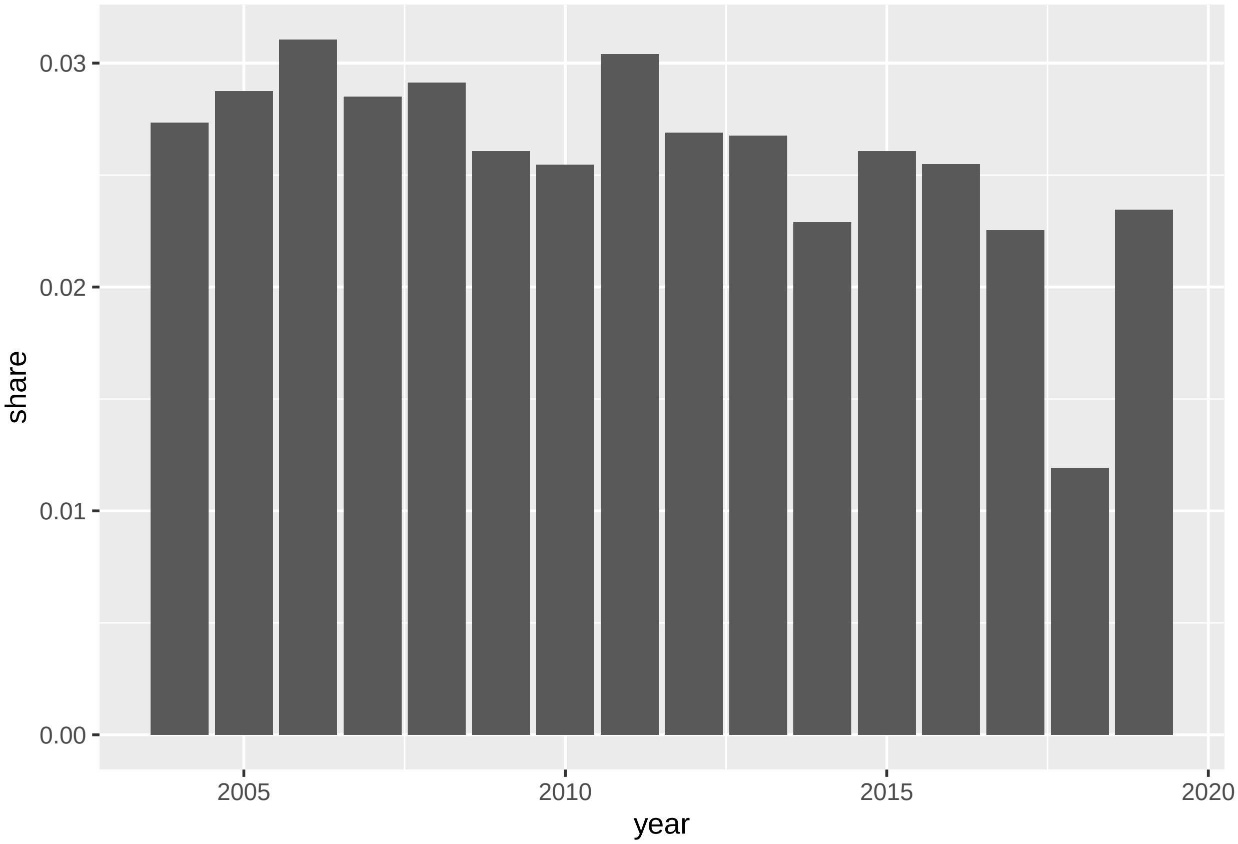
Exercise 3 Migration

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



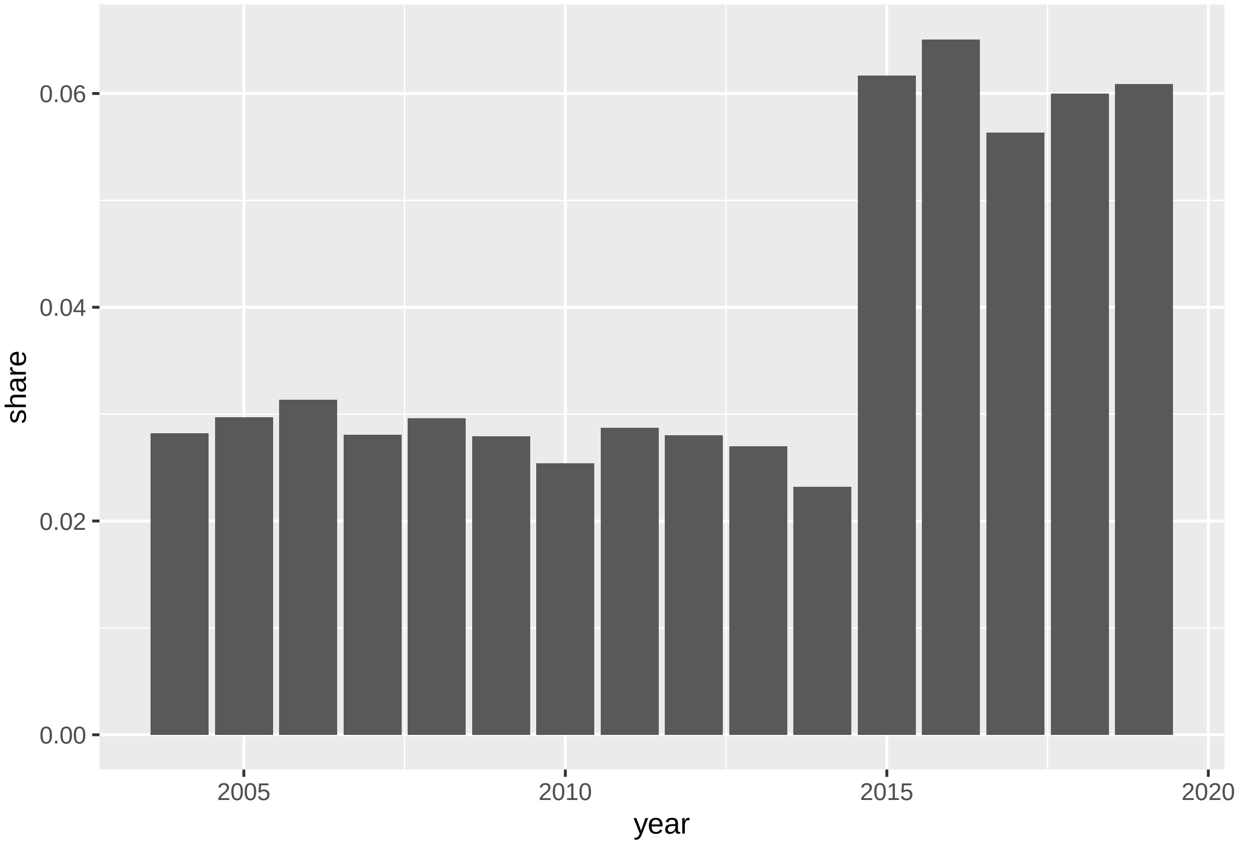
- Based on datent, identify whether or not a household moved into its current dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

idmen	year	datent	idind	move
1200010012930100	2004	2000	1120001001293009920	FALSE
1200010040580100	2004	2001	1120001004058009856	FALSE
1200010040580100	2004	2001	1120001004058009856	FALSE
1200010040580100	2005	2001	1120001004058009856	FALSE
1200010040580100	2005	2001	1120001004058009856	FALSE
1200010066630100	2004	2000	1120001006663009920	FALSE
1200010066630100	2004	2000	1120001006663009920	FALSE
1200010066630100	2005	2005	1120001006663009920	TRUE
1200010066630100	2005	2005	1120001006663009920	TRUE
1200010082450100	2004	1957	1120001008245009920	FALSE

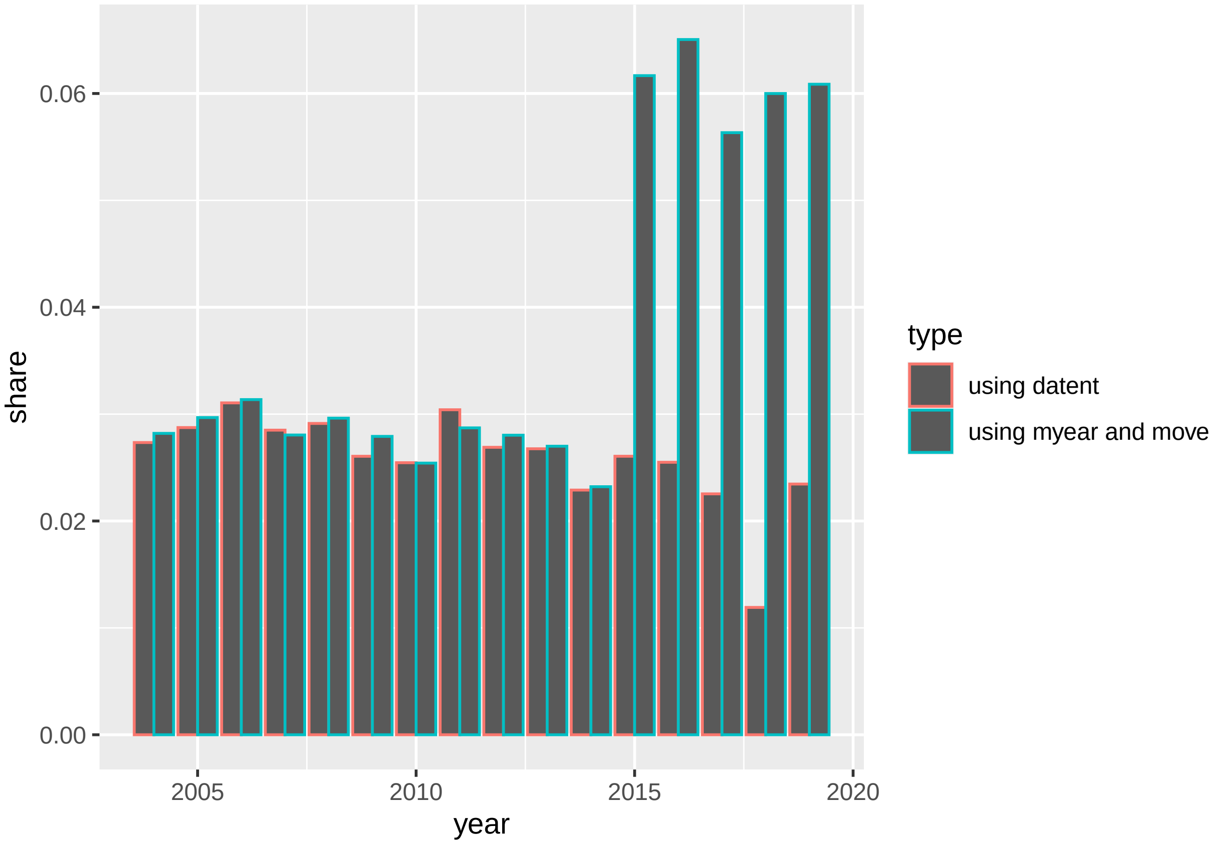


- Based on myear and move, identify whether or not household migrated at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

idmen	year	myear	idind	mig	move
1200010012930100	2004	2000	1120001001293009920	FALSE	NA
1200010040580100	2004	2001	1120001004058009856	FALSE	NA
1200010040580100	2004	2001	1120001004058009856	FALSE	NA
1200010040580100	2005	2001	1120001004058009856	FALSE	NA
1200010040580100	2005	2001	1120001004058009856	FALSE	NA
1200010066630100	2004	2000	1120001006663009920	FALSE	NA
1200010066630100	2004	2000	1120001006663009920	FALSE	NA
1200010066630100	2005	2005	1120001006663009920	TRUE	NA
1200010066630100	2005	2005	1120001006663009920	TRUE	NA
1200010082450100	2004	1957	1120001008245009920	FALSE	NA



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



As the method in 3.4 contains two question types in the questionnaire, which could be less consistent. The samples might have different understanding of variable ‘migration year’ and ‘moving’ in the last period. Therefore, I would choose the method in 3.3.

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

year	2004	2005	2006	2007	2008	2009	2010	2011
count	158	229	248	232	233	212	211	257

year	2012	2013	2014	2015	2016	2017	2018
count	249	207	206	210	204	177	87

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.

year	2005	2006	2007	2008	2009	2010	2011	2012
attrition	-1.0700879	-0.8165998	-0.6756861	-0.3684045	-0.3039319	-0.3316875	-0.2368586	-0.2532417

year	2013	2014	2015	2016	2017	2018	2019
attrition	0.3070998	0.2642700	0.3432668	0.4140053	0.6901819	0.8138311	0.8758118