ECON 613 Assignment 1

Xin Lin

```r
# Load packages
library(tidyverse)
library(magrittr)
library(janitor)
library(knitr)
library(kableExtra)
library(DescTools)
library(reshape)

# Import household datasets
dat04hh <- read_csv("data/dathh2004.csv")
dat05hh <- read_csv("data/dathh2005.csv")
dat06hh <- read_csv("data/dathh2006.csv")
dat07hh <- read_csv("data/dathh2007.csv")
dat08hh <- read_csv("data/dathh2008.csv")
dat09hh <- read_csv("data/dathh2009.csv")
dat10hh <- read_csv("data/dathh2010.csv")
dat11hh <- read_csv("data/dathh2011.csv")
dat12hh <- read_csv("data/dathh2012.csv")
dat13hh <- read_csv("data/dathh2013.csv")
dat14hh <- read_csv("data/dathh2014.csv")
dat15hh <- read_csv("data/dathh2015.csv")
dat16hh <- read_csv("data/dathh2016.csv")
dat17hh <- read_csv("data/dathh2017.csv")
dat18hh <- read_csv("data/dathh2018.csv")
dat19hh <- read_csv("data/dathh2019.csv")

# Import individual datasets
dat04ind <- read_csv("data/datind2004.csv")
dat05ind <- read_csv("data/datind2005.csv")
dat06ind <- read_csv("data/datind2006.csv")
dat07ind <- read_csv("data/datind2007.csv")
dat08ind <- read_csv("data/datind2008.csv")
dat09ind <- read_csv("data/datind2009.csv")
dat10ind <- read_csv("data/datind2010.csv")
dat11ind <- read_csv("data/datind2011.csv")
dat12ind <- read_csv("data/datind2012.csv")
dat13ind <- read_csv("data/datind2013.csv")
dat14ind <- read_csv("data/datind2014.csv")
dat15ind <- read_csv("data/datind2015.csv")
dat16ind <- read_csv("data/datind2016.csv")
dat17ind <- read_csv("data/datind2017.csv")
dat18ind <- read_csv("data/datind2018.csv")
dat19ind <- read_csv("data/datind2019.csv")
```

```
# abadon scientific notation in R
options(scipen = 999)
```

Exercise 1 Basic Statistics

1. Number of households surveyed in 2007

```
nrow(dat07hh)
```

```
## [1] 10498
```

Answer: The number of households surveyed in 2007 is 10,498

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

```
nrow(dat05hh[which(dat05hh$mstatus == "Couple, with Kids"),])
```

```
## [1] 3374
```

Answer: The number of household with marital status "Couple with kids" in 2005 is 3,374

3. Number of individuals surveyed in 2008

```
# number of individuals
nrow(dat08ind)
```

```
## [1] 25510
```

```
# number of individuals with unique idind
length(unique(dat08ind$idind))
```

```
## [1] 10828
```

Answer: The number of individuals surveyed in 2008 is 25,510, and the number of individuals with unique idind surveyed in 2008 is 10,828.

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

```
nrow(dat16ind[which(dat16ind$age %in% c(25:35)),])
```

```
## [1] 2765
```

Answer: The number of individuals aged between 25 and 35 in 2016 is 2,765

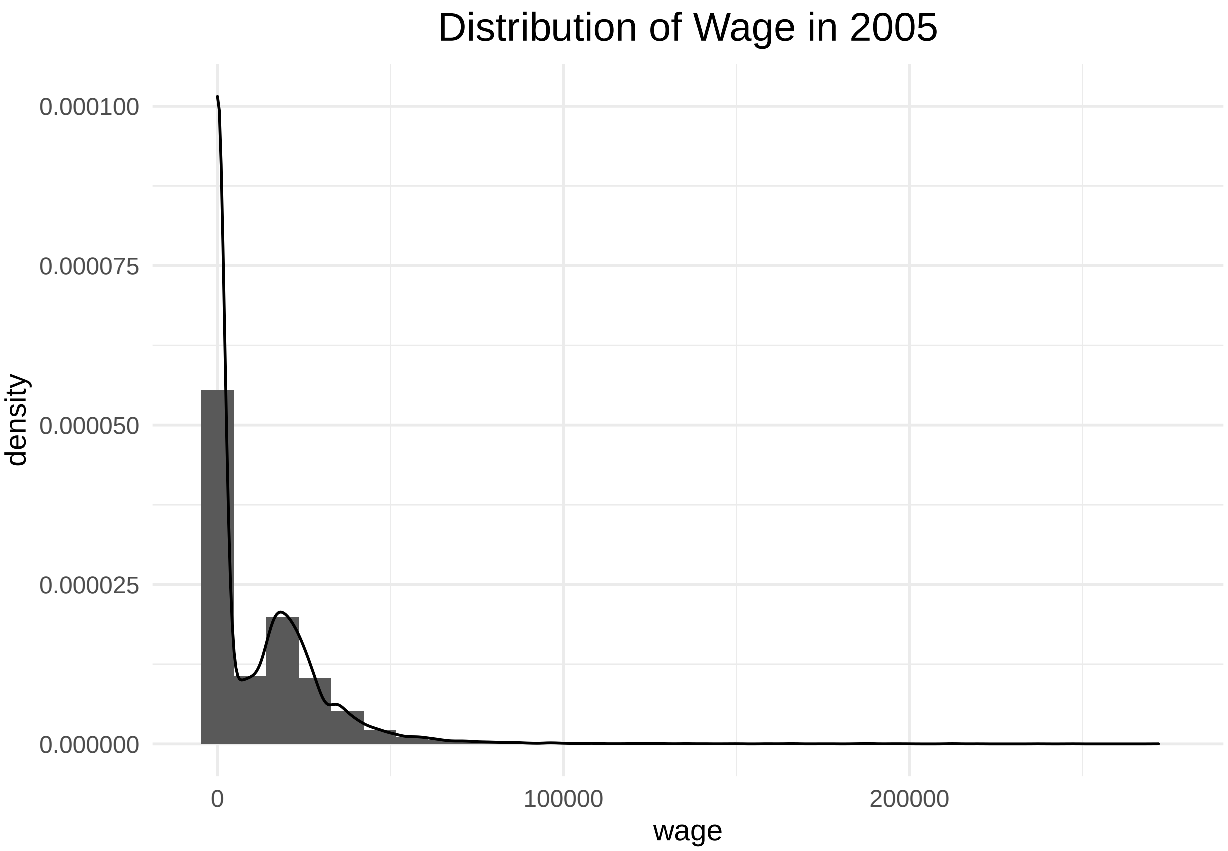
5. Cross-table gender/profession in 2009

```
crosstable <- dat09ind %>%
  tabyl(profession, gender)
kable(crosstable) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

profession	Female	Male
0	11	19
11	30	57
12	8	19
13	29	78
21	63	213
22	65	114
23	8	48
31	68	98
33	85	107
34	184	142
35	50	59
37	179	260
38	78	368
42	258	110
43	437	117
44	1	2
45	153	95
46	410	340
47	82	429
48	22	215
52	782	169
53	27	182
54	584	98
55	353	101
56	696	74
62	64	443
63	35	520
64	29	246
65	19	159
67	147	237
68	120	177
69	40	82
NA	8167	6949

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

```
# 2005 distribition plot
ggplot(dat05ind, aes(x=wage)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density() +
  theme_minimal() +
  labs(title = "Distribution of Wage in 2005") +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



```
# drop NA and 0's to get the statistics
dat05ind_new <- dat05ind[-which(dat05ind$wage %in% c(0, NA)),]
# 2005 mean
mean(dat05ind_new$wage)
```

```
## [1] 22443.03
```

```
# 2005 sd
sd(dat05ind_new$wage)
```

```
## [1] 18076.71
```

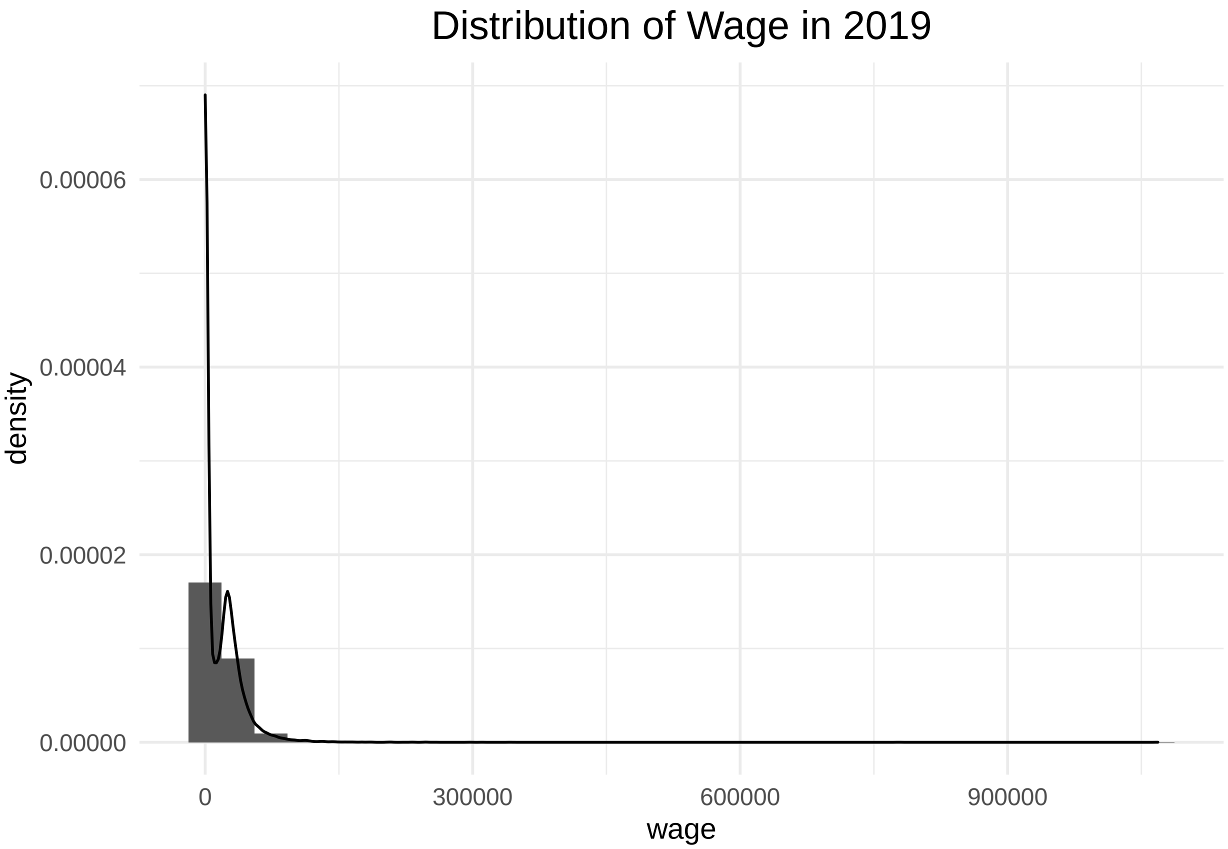
```
# 2005 inter-decile ratio D9/D1
quantile(dat05ind_new$wage, probs = 0.9)/quantile(dat05ind_new$wage, probs = 0.1)
```

```
##      90%
## 8.896525
```

```r
# 2005 Gini coefficient
Gini(dat05ind_new$wage)
```

```
## [1] 0.3771511
```

```r
# 2019 distribution plot
ggplot(dat19ind, aes(x=wage)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density() +
  theme_minimal() +
  labs(title = "Distribution of Wage in 2019") +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



```r
# drop NA and 0's to get the statistics
dat19ind_new <- dat19ind[-which(dat19ind$wage %in% c(0, NA)),]
# 2019 mean
mean(dat19ind_new$wage)
```

```
## [1] 27578.84
```

```r
# 2019 sd
sd(dat19ind_new$wage)
```

```
## [1] 25107.19
```

```
# 2019 inter-decile ratio D9/D1
quantile(dat19ind_new$wage, probs = 0.9) / quantile(dat19ind_new$wage, probs = 0.1)
```

```
##     90%
## 13.8623
```

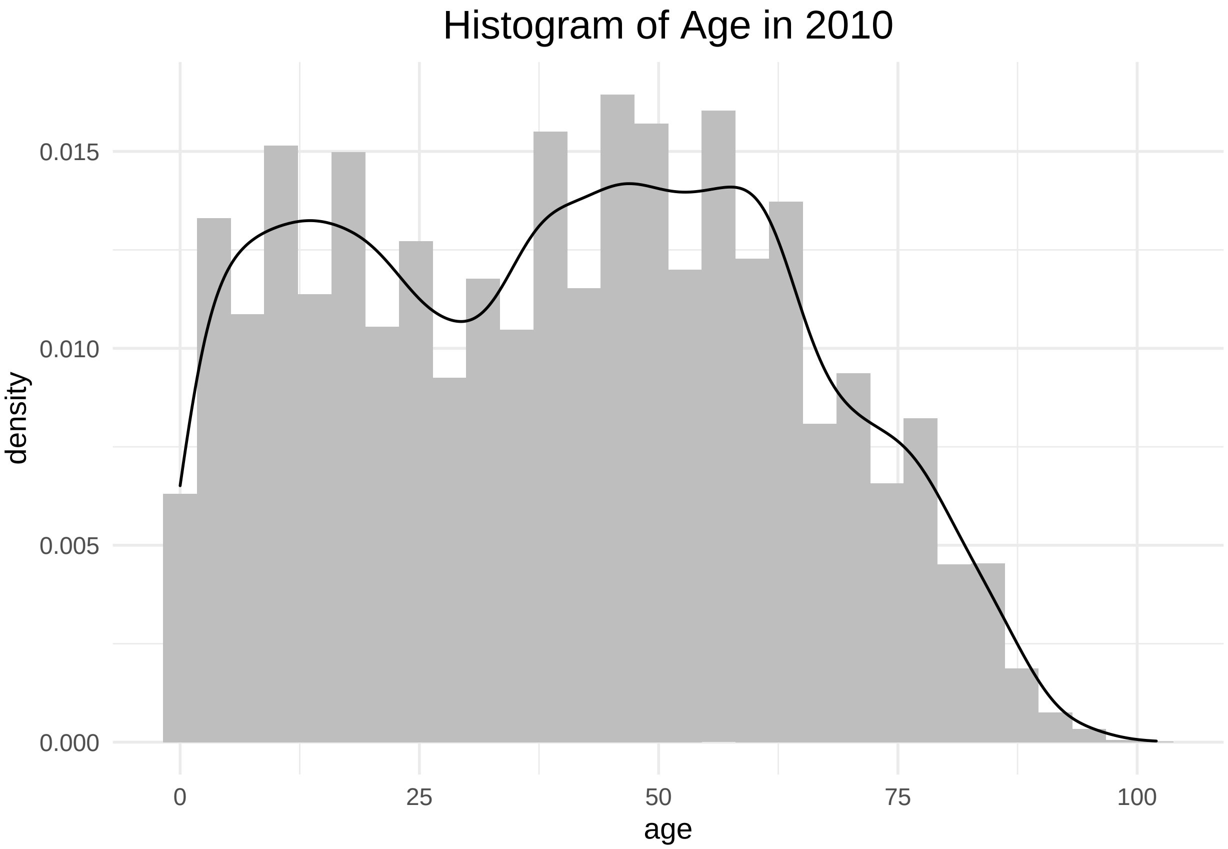
```
# 2019 Gini coefficient
Gini(dat19ind_new$wage)
```

```
## [1] 0.399121
```

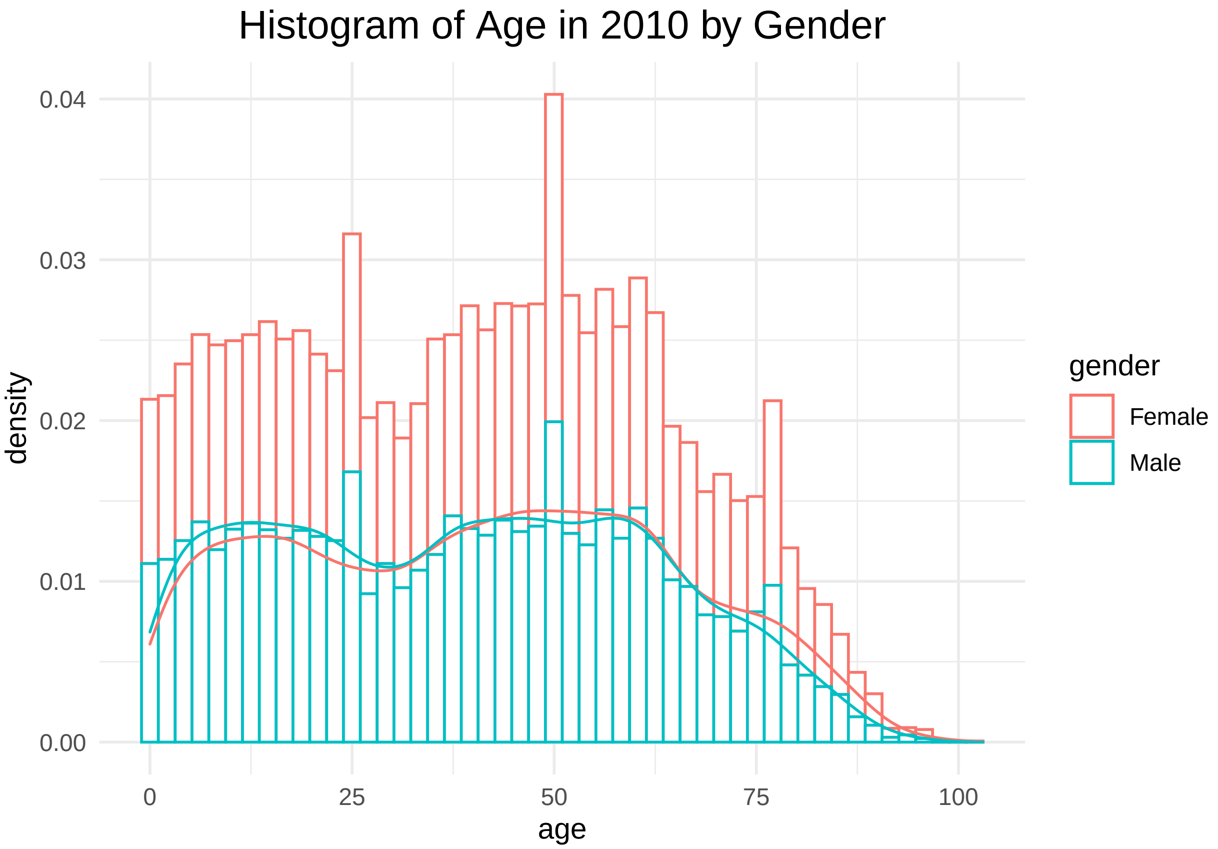
Answer: In order to get the statistics, I dropped all missing values and 0's in the wage. In 2005, the mean of wage is $22,443.03, the standard deviation is $18,076.71, the inter-decile ratio D9/D1 is 8.896525, and the Gini coefficient is 0.3771135. In 2019, the mean of wage is $27,578.84, the standard deviation is $25,107.19, the inter-decile ratio D9/D1 is 13.8623, and the Gini coefficient is 0.399121.

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

```
# histogram
ggplot(dat10ind, aes(x=age)) +
  geom_histogram(aes(y=..density..), fill = ’grey’, bins = 30) +
  geom_density() +
  theme_minimal() +
  labs(title = "Histogram of Age in 2010") +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



```
# histogram by age
ggplot(dat10ind, aes(x=age, color=gender)) +
  geom_histogram(aes(y=..density..), fill = 'white', bins = 50) +
  geom_density() +
  theme_minimal() +
  labs(title = "Histogram of Age in 2010 by Gender") +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



Answer: From the histogram by gender below, we can observe that the percentage of under 30 is higher in males and the percentage of above 30 is higher in females.

8. Number of individuals in Paris in 2011

```
# merge the data sets
dat11 <- as.data.frame(merge(x = dat11ind, y = dat11hh, by = "idmen"))
length(which(dat11$location == "Paris"))
```

```
## [1] 3514
```

Answer: The number of individuals in Paris in 2011 is 3,514.

Exercise 2 Merge Datasets

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

```
dathh <- rbind(dat04hh, dat05hh, dat06hh, dat07hh, dat08hh, dat09hh, dat10hh, dat11hh,
               dat12hh, dat13hh, dat14hh, dat15hh, dat16hh, dat17hh, dat18hh, dat19hh)
```

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

```
datind <- rbind(dat04ind, dat05ind, dat06ind, dat07ind, dat08ind, dat09ind, dat10ind, dat11ind,
                dat12ind, dat13ind, dat14ind, dat15ind, dat16ind, dat17ind, dat18ind, dat19ind)
```

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

```
dathh_col_name <- names(dathh)
datind_col_name <- names(datind)
intersect(dathh_col_name, datind_col_name)
```

```
## [1] "X1"    "idmen" "year"
```

Answer: The variables that are simultaneously present in the individual and household datasets are “X1”, “idmen”, and “year”.

4. Merge the appended individual and household datasets

```
dat <- merge(x = datind, y = dathh, by = c("idmen", "year"), all.x = TRUE)
```

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

```
# frequency table
n_occur <- data.frame(table(dat$idmen, dat$year))

# find all id’s of households satisfying the condition
n_occur <- n_occur %>%
  filter(Freq > 4) %>%
  group_by(Var2) %>%
  summarise(n = n())

# change the name of columns and print the number of households satisfying the condition for each year
colnames(n_occur) <- c("Year", "Number of Households")
kable(n_occur) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
              full_width=FALSE)
```

Year	Number of Households
2004	745
2005	814
2006	862
2007	874
2008	814
2009	810
2010	821
2011	785
2012	816
2013	754
2014	783
2015	763
2016	753
2017	703
2018	647
2019	692

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

```r
# frequency table
n_occur2 <- data.frame(table(dat$idmen, dat$year, dat$empstat))

# find all id's of households satisfying the condition
n_occur2 <- n_occur2 %>%
  filter(Var3 == "Unemployed") %>%
  filter(Freq >= 1)%>%
  group_by(Var2) %>%
  summarise(n = n())

# change the name of columns and print the number of households satisfying the condition for each year
colnames(n_occur2) <- c("Year", "Number of Households")
kable(n_occur2) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year	Number of Households
2004	950
2005	1039
2006	1030
2007	975
2008	909
2009	1045
2010	1110
2011	1071
2012	1205
2013	1177
2014	1187
2015	1227
2016	1137
2017	1103
2018	991
2019	1086

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

```r
# frequency table
n_occur3 <- data.frame(table(dat$idmen, dat$year, dat$profession))

# find all id's of households satisfying the condition
n_occur3 <- n_occur3 %>%
  filter(Freq >= 2) %>%
  group_by(Var2) %>%
  summarise(n = n())

# change the name of columns and print the number of households satisfying the condition for each year
colnames(n_occur3) <- c("Year", "Number of Households")
kable(n_occur3) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year	Number of Households
2004	445
2005	497
2006	485
2007	492
2008	460
2009	453
2010	477
2011	492
2012	517
2013	460
2014	477
2015	469
2016	475
2017	459
2018	457
2019	500

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

```
# frequency table
n_occur4 <- data.frame(table(dat$idmen, dat$year, dat$mstatus))

# find all id's of households satisfying the condition
n_occur4 <- n_occur4 %>%
  filter(Var3 == "Couple, with Kids") %>%
  filter(Freq >= 1) %>%
  group_by(Var2) %>%
  summarise(n = sum(Freq))

# change the name of columns and print the number of individuals satisfying the condition for each year
colnames(n_occur4) <- c("Year", "Number of Individuals")
kable(n_occur4) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year	Number of Individuals
2004	11993
2005	13217
2006	13637
2007	13963
2008	13481
2009	13286
2010	13726
2011	13801
2012	14403
2013	13114
2014	13228
2015	13008
2016	12967
2017	11963
2018	11444
2019	12151

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

```
# frequency table
n_occur5 <- data.frame(table(dat$idmen, dat$year, dat$location))

# find all id's of households satisfying the condition
n_occur5 <- n_occur5 %>%
  filter(Var3 == "Paris") %>%
  filter(Freq >= 1) %>%
  group_by(Var2) %>%
  summarise(n = sum(Freq))

# change the name of columns and print the number of individuals satisfying the condition for each year
colnames(n_occur5) <- c("Year", "Number of Individuals")
kable(n_occur5) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year	Number of Individuals
2004	3494
2005	3734
2006	3658
2007	3735
2008	3559
2009	3524
2010	3607
2011	3514
2012	3679
2013	2288
2014	2576
2015	3033
2016	2946
2017	2836
2018	2797
2019	2924

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

```r
# frequency table
n_occur6 <- data.frame(table(dat$idmen, dat$year))

# find all id's of households satisfying the condition
idmen_most_fm <- n_occur6 %>%
  filter(Freq == max(n_occur6$Freq)) %>%
  select(Var1, Var2, Freq)

# change the name of columns and print the number of households satisfying the condition for each year
colnames(idmen_most_fm) <- c("idmen", "Year", "Family Numbers")
kable(idmen_most_fm) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

idmen	Year	Family Numbers
2207811124040100	2007	14
2510263102990100	2010	14

Answer: The idmen of the household with the most number of family members are 2207811124040100 and 2510263102990100. Also, the number of their family member is 14.

11. Number of households present in 2010 and 2011.

```r
# number of households present in 2010
hh_10 <- dat %>%
  filter(year %in% c(2010))
hh_10 <- unique(hh_10$idmen)
length(hh_10)
```

```
## [1] 11050
```

```
# number of households present in 2011
hh_11 <- dat %>%
  filter(year %in% c(2011))
hh_11 <- unique(hh_11$idmen)
length(hh_11)
```

```
## [1] 11360
```

```
# number of households present in 2010 and 2011
hh_10_11 <- intersect(hh_10, hh_11)
length(hh_10_11)
```

```
## [1] 8984
```

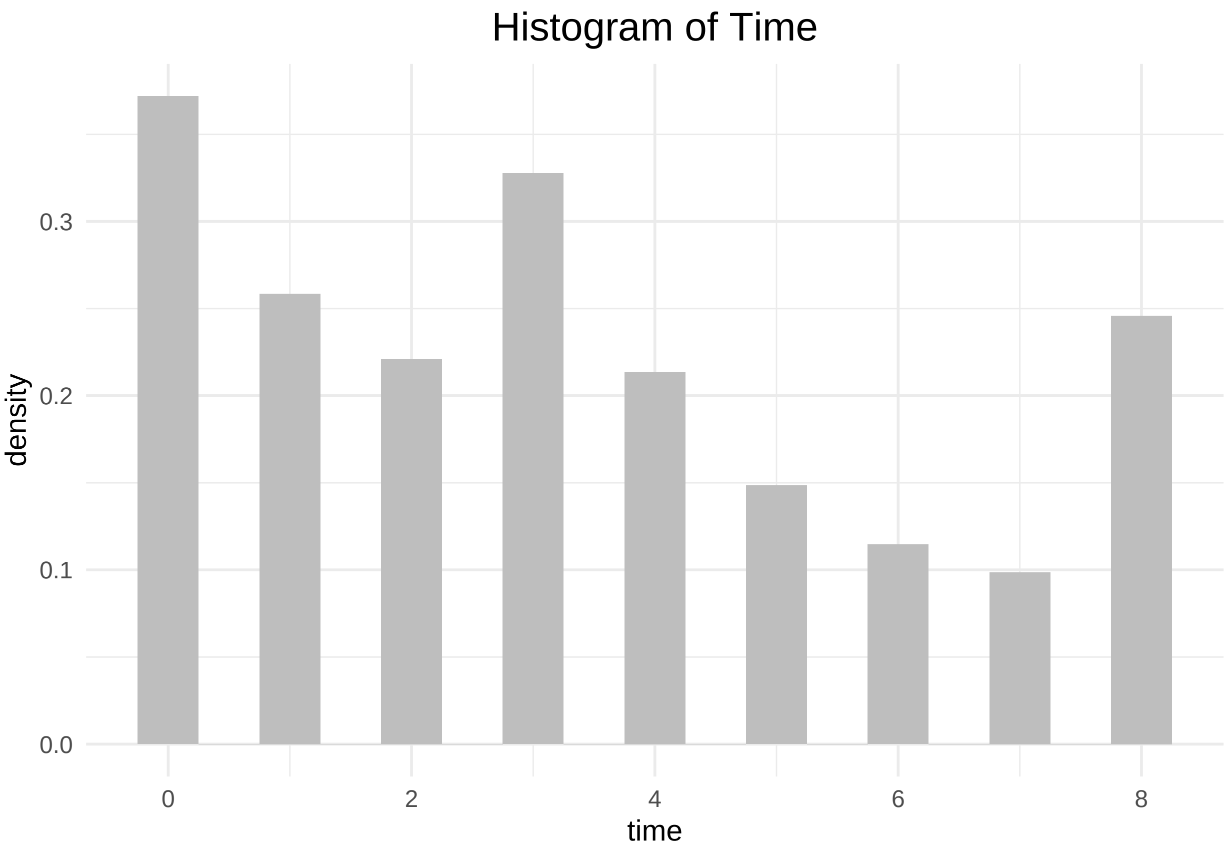
Answer: There are 11,050 households present in 2010 and 11,360 households present in 2011. Also, there are 8,984 households present in both 2010 and 2011.

Exercise 3 Migration

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

```r
# create three new variables: year enter, year exit, and time
dat_time <- dat %>%
  group_by(idmen) %>%
  summarise(year_enter = min(year), # year enter
            year_exit = max(year), # year exit
            time = year_exit - year_enter) %>%
  ungroup()
```

```r
# distrbution of time spent in the survey for each household
ggplot(dat_time, aes(x=time)) +
  geom_histogram(aes(y=..density..), fill = 'grey', bins = 17) +
  theme_minimal() +
  labs(title = "Histogram of Time") +
  theme(plot.title = element_text(face = "plain", size = 15, hjust = 0.5, color = "black"))
```



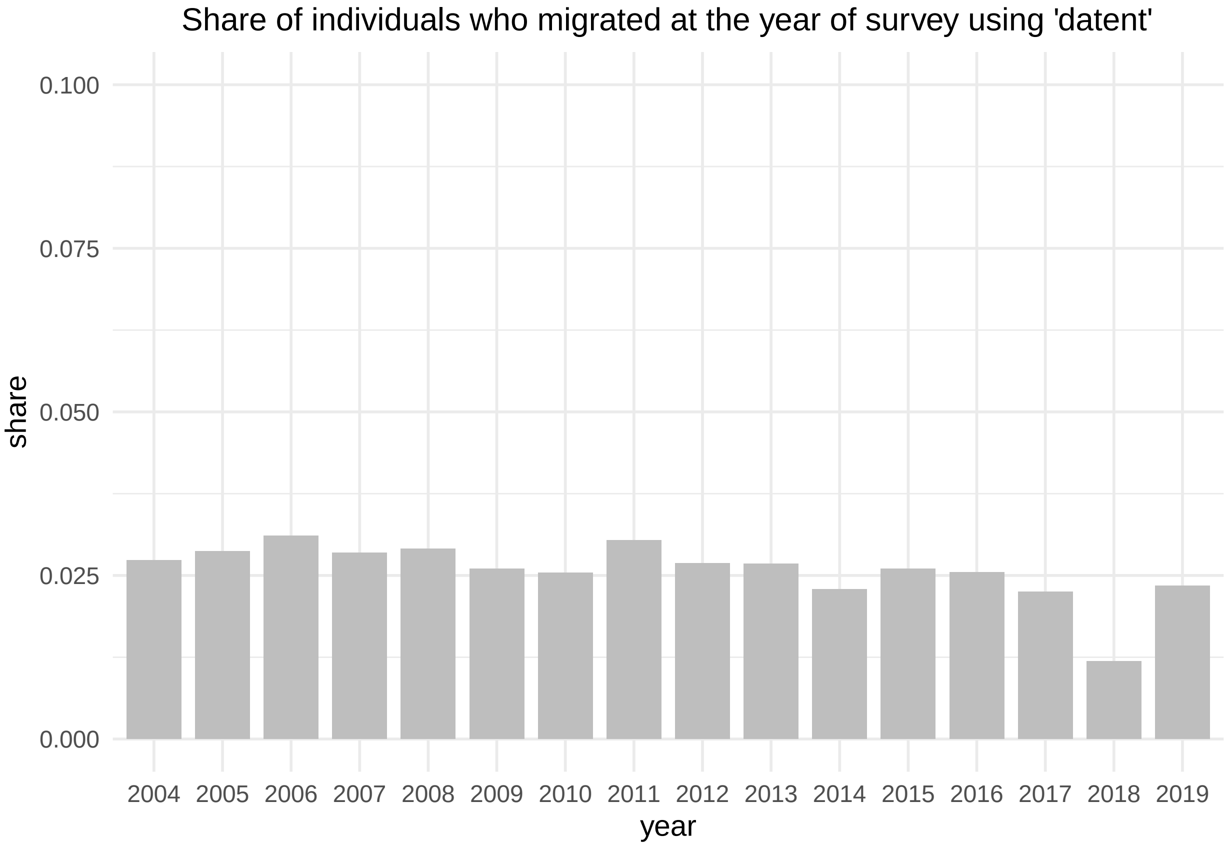
2. Based on datent, identify whether or not a household moved into its current dwelling at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years

```
# create a new variable: whether_move, if move: 1, if not move: 0
dat_whether <- dat %>%
  mutate(whether_move = ifelse(year == datent, 1, 0))

# print the first ten rows
head(dat_whether,10)
```

```
##                idmen year X1.x               idind  empstat respondent
## 1  1200010012930100 2004    1 1120001001293010048 Employed          1
## 2  1200010040580100 2004    2 1120001004058009984 Employed          1
## 3  1200010040580100 2004    3 1120001004058009984 Inactive          0
## 4  1200010040580100 2005    1 1120001004058009984 Inactive          1
## 5  1200010040580100 2005    2 1120001004058009984 Inactive          0
## 6  1200010066630100 2004    4 1120001006663010048 Employed          1
## 7  1200010066630100 2004    5 1120001006663010048 Employed          0
## 8  1200010066630100 2005    4 1120001006663010048 Employed          0
## 9  1200010066630100 2005    3 1120001006663010048 Employed          1
## 10 1200010082450100 2004    6 1120001008245010048  Retired          1
##    profession gender age  wage X1.y datent myear         mstatus move location
## 1          67   Male  31 19187    1   2000  2000          Single   NA    Paris
## 2          56 Female  30 11586    2   2001  2001   Single Parent   NA    Paris
## 3        <NA> Female   9    NA    2   2001  2001   Single Parent   NA    Paris
## 4        <NA> Female  31 12334    1   2001  2001   Single Parent   NA    Paris
## 5        <NA> Female  10    NA    1   2001  2001   Single Parent   NA    Paris
## 6          38   Male  31 44656    3   2000  2000 Couple, No kids   NA    Paris
## 7          45 Female  27 20413    3   2000  2000 Couple, No kids   NA    Paris
## 8          45 Female  28 19231    2   2005  2005 Couple, No kids   NA    Paris
## 9          38   Male  32 50659    2   2005  2005 Couple, No kids   NA    Paris
## 10       <NA> Female  89     0    4   1957  1957          Single   NA    Paris
##    whether_move
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
## 6             0
## 7             0
## 8             1
## 9             1
## 10            0
```

```
# create a table of shares accross years
move_share <- dat_whether %>%
  tabyl(year, whether_move) %>%
  adorn_totals(where = c("row", "col"))

# create a new variable: share, representing the share of inds in that situation
move_share <- as.data.frame(move_share)
```

```r
move_share$share <- move_share$'1' / move_share$Total
move_share <- move_share[-17,]

# plot the share of individuals in that situation across years
ggplot(move_share, aes(x=year, y=share)) +
  geom_bar(stat = "identity", width = 0.8, fill = 'grey') +
  lims(y = c(0,0.1)) +
  theme_minimal() +
  labs(title = "Share of individuals who migrated at the year of survey using 'datent'") +
  theme(plot.title = element_text(face = "plain", size = 12, hjust = 0.5, color = "black"))
```

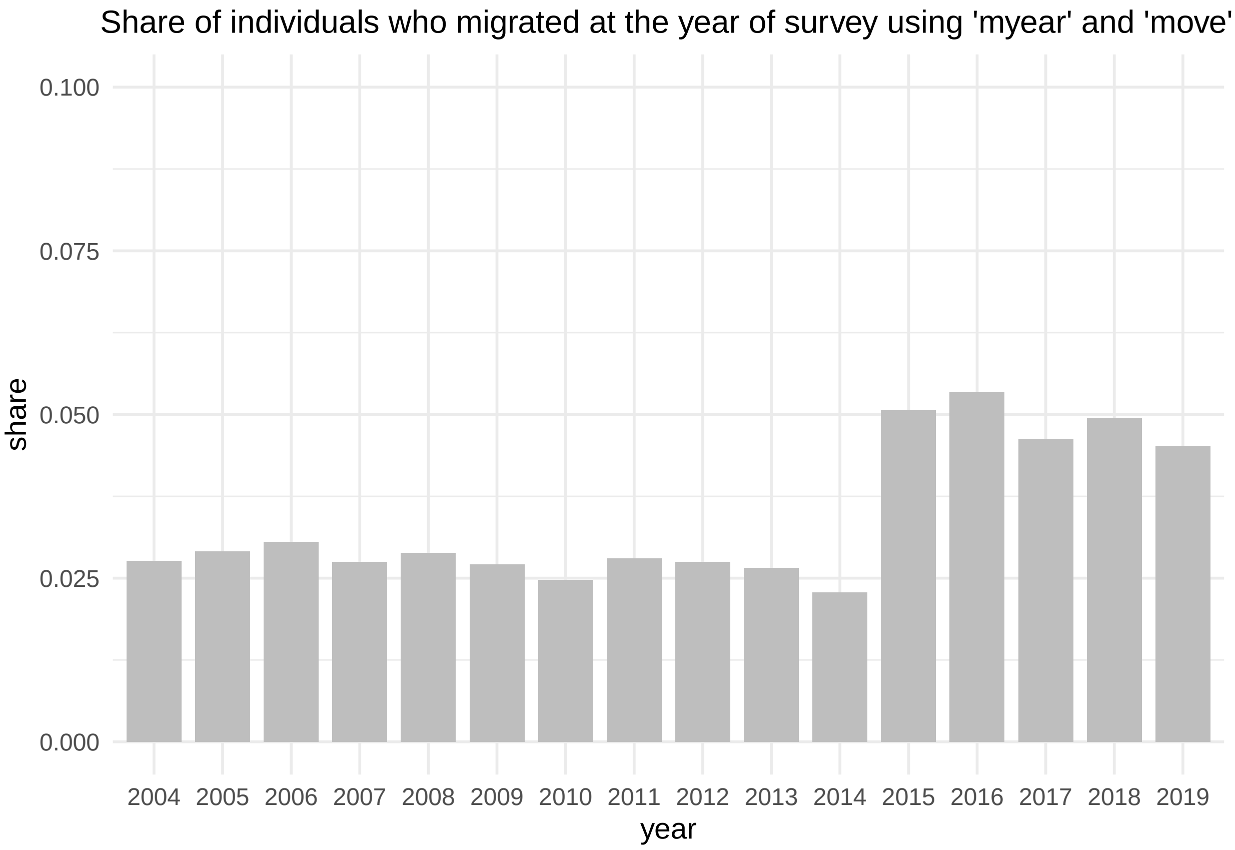


3. Based on myear and move, identify whether or not household migrated at the year of survey. Report the first 10 rows of your result and plot the share of individuals in that situation across years.

```r
# create a new variable: whether_migrate, if migrate: 1, if not migrate: 0
dat_whether <- dat_whether %>%
  mutate(whether_migrate = ifelse(year<=2014, ifelse(year==myear,1,0), ifelse(move==2,1,0)))

# print the first ten rows
head(dat_whether,10)
```

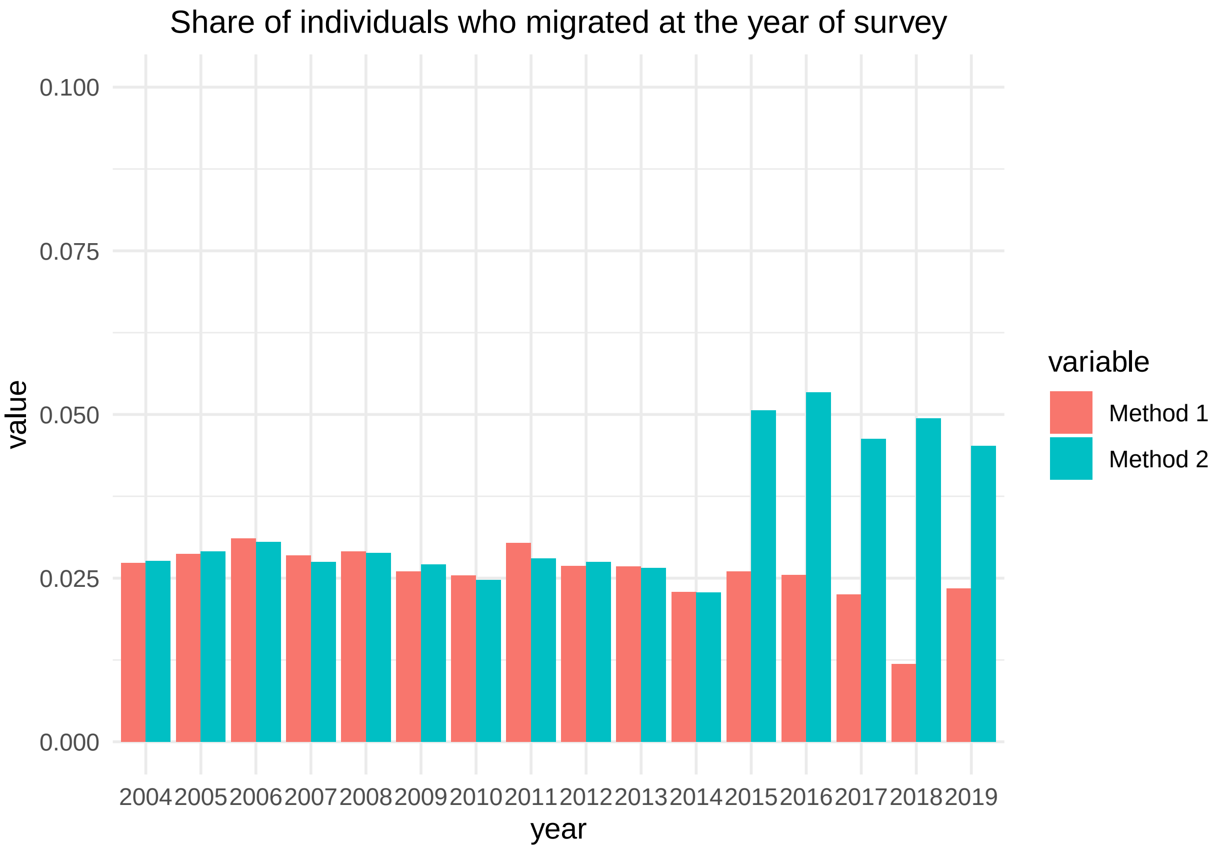
```
##                idmen year X1.x               idind  empstat respondent
## 1  1200010012930100 2004    1 1120001001293010048 Employed          1
## 2  1200010040580100 2004    2 1120001004058009984 Employed           1
## 3  1200010040580100 2004    3 1120001004058009984 Inactive           0
## 4  1200010040580100 2005    1 1120001004058009984 Inactive           1
## 5  1200010040580100 2005    2 1120001004058009984 Inactive           0
## 6  1200010066630100 2004    4 1120001006663010048 Employed           1
## 7  1200010066630100 2004    5 1120001006663010048 Employed           0
## 8  1200010066630100 2005    4 1120001006663010048 Employed           0
## 9  1200010066630100 2005    3 1120001006663010048 Employed           1
## 10 1200010082450100 2004    6 1120001008245010048  Retired           1
##    profession gender age  wage X1.y datent myear         mstatus move location
## 1          67   Male  31 19187    1   2000  2000          Single   NA    Paris
## 2          56 Female  30 11586    2   2001  2001   Single Parent   NA    Paris
## 3        <NA> Female   9    NA    2   2001  2001   Single Parent   NA    Paris
## 4        <NA> Female  31 12334    1   2001  2001   Single Parent   NA    Paris
## 5        <NA> Female  10    NA    1   2001  2001   Single Parent   NA    Paris
## 6          38   Male  31 44656    3   2000  2000 Couple, No kids   NA    Paris
## 7          45 Female  27 20413    3   2000  2000 Couple, No kids   NA    Paris
## 8          45 Female  28 19231    2   2005  2005 Couple, No kids   NA    Paris
## 9          38   Male  32 50659    2   2005  2005 Couple, No kids   NA    Paris
## 10       <NA> Female  89     0    4   1957  1957          Single   NA    Paris
##    whether_move whether_migrate
## 1             0               0
## 2             0               0
## 3             0               0
## 4             0               0
## 5             0               0
## 6             0               0
## 7             0               0
## 8             1               1
## 9             1               1
## 10            0               0
```

```
# create a table of shares accross years
move_share2 <- dat_whether %>%
  tabyl(year, whether_migrate) %>%
  adorn_totals(where = c("row", "col"))

# create a new variable: share, representing the share of inds in that situation
move_share2 <- as.data.frame(move_share2)
move_share2$share <- move_share2$’1’ / move_share2$Total
move_share2 <- move_share2[-17,]

# plot the share of individuals in that situation across years
ggplot(move_share2, aes(x=year, y=share)) +
  geom_bar(stat = "identity", width = 0.8, fill = ’grey’) +
  lims(y = c(0,0.1)) +
  theme_minimal() +
  labs(title = "Share of individuals who migrated at the year of survey using ’myear’ and ’move’") +
  theme(plot.title = element_text(face = "plain", size = 12, hjust = 0.5, color = "black"))
```



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

```
# mix move_share and migrate_share in one dataset
move_share3 <- data.frame(move_share$year, move_share$share, move_share2$share)
colnames(move_share3) <- c("year", "Method 1", "Method 2")
move_share3 <- melt(move_share3, id.vars='year')

# mix move_share and migrate_share in one plot
ggplot(move_share3, aes(x=year, y=value, fill=variable)) +
    geom_bar(stat='identity', position='dodge') +
    lims(y = c(0,0.1)) +
    theme_minimal() +
    labs(title = "Share of individuals who migrated at the year of survey") +
    theme(plot.title = element_text(face = "plain", size = 12, hjust = 0.5, color = "black"))
```



Answer: I prefer the first method: using the variable ‘datent’ to determine whether the surveyed individuals migrated in the year of survey (Exercise 3.2). Since the second method (Exercise 3.3) uses two meausres ‘myear’ and ‘move’ to determine the share of individuals who migrated at the year of survey, these two measures are inconsistent as shown in the above graph: the share after 2014 are much higher. Therefore, I prefer the first method using ‘datent’ since it is more consistent without change of measures acrross years.

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

```
# find id and year of households who migrated
idmen_migrate <- dat_whether %>%
  filter(whether_move == 1) %>%
  group_by(idmen, year) %>%
  summarise(move_when_survey = n())

# keep only migrated household and useful variables
dat_migrate <- merge(idmen_migrate, dat, by = c(’year’, ’idmen’)) %>%
  select(year, idmen, idind, profession, empstat)

# find the new empstat and profession for each migrated households
dat_migrate_next_year <- dat_migrate %>%
  select(idind, year, idmen) %>%
  mutate(year = year + 1) # ’year’ is next year but not current year
dat_migrate_next_year <- merge(dat_migrate_next_year, dat,
                               by = c("year", "idind", "idmen"))
```

```
dat_migrate_next_year <- dat_migrate_next_year %>%
  mutate(year = year - 1) %>% # ’year’ is current year
  select(year, idmen, idind, profession, empstat)
colnames(dat_migrate_next_year)[4:5] <- c("new_profession", "new_empstat")

# add new empstat and profession to migrated household dataset
dat_migrate2 <- merge(dat_migrate_next_year, dat_migrate,
                     by = c(’year’, ’idind’, ’idmen’)) %>%
  mutate(year_of_change = year + 1)

# find whether the migrated households change their profession or empstat
# add a new variable ’whether_change’: 1 if change and 0 if not change
dat_migrate2 <- dat_migrate2 %>%
  mutate(whether_change = ifelse(empstat != new_empstat | profession != new_profession, 1, 0)) %>%
  group_by(year_of_change, idmen) %>%
  filter(whether_change == 1) %>%
  group_by(year_of_change) %>%
  summarise(length(unique(idmen)))

# print the number of migrated households who change empstat or profession
colnames(dat_migrate2) <- c("Year of Change", "Number of Households")
kable(dat_migrate2) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year of Change	Number of Households
2005	158
2006	151
2007	164
2008	149
2009	143
2010	164
2011	159
2012	191
2013	142
2014	140
2015	125
2016	156
2017	119
2018	111
2019	67

Exercise 4 Attrition

Compute the attrition across each year, where attrition is defined as the reduction in the number of indi- viduals staying in the data panel. Report your final result as a table in proportions

```
# add a year of entry and year of exit for all individuals
dat_attrition <- merge(x = dat, y = dat_time, by = "idmen", all.x = TRUE)
```

```r
# add a variable ’stay’: whether the individual stays in the panel for the year surveyed
# 1 if ’year’ != ’year_exit’ and 0 if ’year’ == ’year_exit’
dat_attrition <- dat_attrition %>%
  mutate(stay = ifelse(year == year_exit, 0, 1))

# stay = 0 means reduction in the number of individuals staying in the data panel
# attrition rate = reduction in the number of individuals / total number of individuals
attrition_rate <- dat_attrition %>%
  group_by(year) %>%
  summarise(Reduction = length(which(stay == 0)),
            Total = n(),
            ’Attrition Rate’ = Reduction / Total)
colnames(attrition_rate)[1] = "Year"
kable(attrition_rate) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                full_width=FALSE)
```

Year	Reduction	Total	Attrition Rate
2004	2384	22144	0.1076590
2005	4098	24241	0.1690524
2006	3798	24940	0.1522855
2007	5236	25907	0.2021075
2008	4557	25510	0.1786358
2009	4070	25611	0.1589161
2010	4305	26531	0.1622630
2011	3931	27071	0.1452107
2012	5635	28534	0.1974837
2013	4733	26353	0.1796000
2014	4819	26787	0.1799007
2015	4816	26644	0.1807536
2016	5407	26647	0.2029121
2017	5133	25402	0.2020707
2018	5681	24698	0.2300186
2019	26484	26484	1.0000000