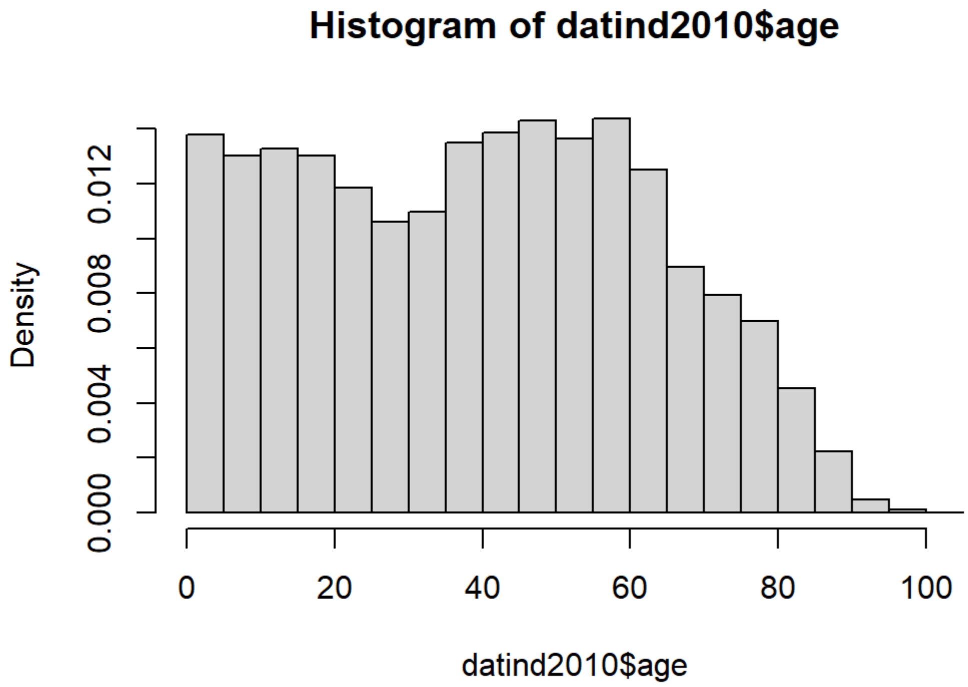
```
# Homework 1
# Exercise 1
install.packages("tidyverse")
# Q1 There are 10498 households surveyed.
dathh2007 = read_csv("Data/dathh2007.csv")
dimnames(dathh2007)
class(dathh2007)
nrow(dathh2007)
# Q2 There are 3374 households with marital status "Couple with Kids" in 2005.
dathh2005 = read_csv("Data/dathh2005.csv")
Q2 = dathh2005[dathh2005$mstatus == "Couple, with Kids", c("year","mstatus")]
nrow(Q2)
# Q3 There are 25510 individuals surveyed in 2008
datind2008 = read_csv("Data/datind2008.csv")
nrow(datind2008)
# Q4 There are 255 individuals aged between 25 and 35
datind2016 = read_csv("Data/datind2016.csv")
Q4 = datind2016[datind2016$age == c(25:35), c("year", "age")]
nrow(Q4)
# Q5
datind2009 = read_csv("Data/datind2009.csv")
install.packages("gmodels")
library(gmodels)
Q5 = CrossTable(datind2009$gender,datind2009$profession,prop.chisq = FALSE)
# Q6
datind2005 = read_csv("Data/datind2005.csv")
datind2019 = read_csv("Data/datind2019.csv")
Q6a = datind2005[,10]
na_Q6a = which(!complete.cases(Q6a))
```

```
Q6a = Q6a[-na_Q6a,]
mean(Q6a$wage) #in 2005, mean is 11992.26
sd(Q6a$wage) #standard deviation is 17318.56
quantile(Q6a$wage,prob=c(0.1,0.9)) # D9 is 32340.4 and D1 is 0.0, we cannot get a valid result
install.packages("REAT")
library(REAT)
gini05 = gini(Q6a$wage) # Gini coefficient is 0.667165445253239 for 2005
Q6b = datind2019[,10]
na_Q6b = which(!complete.cases(Q6b))
Q6b = Q6b[-na_Q6b,]
mean(Q6b$wage) # Mean wage is 15350.47 in 2019
sd(Q6b$wage) # Standard deviation is 23207.18 in 2019
quantile(Q6b$wage, prob = c(0.1,0.9)) # In 2019, D1 is 0, D9 is 40267, no valid result of the ratio
gini19 = gini(Q6b$wage) # In 2019, Gini coefficient is 0.66553009302
# Q7
datind2010 = read.csv("Data/datind2010.csv")
hist(datind2010$age,freq = F)
Q7a = datind2010[datind2010$gender == "Male", c("gender","age")]
mean(Q7a$age) # Mean of male's age is 38.8736
Q7b = datind2010[datind2010$gender == "Female", c("gender","age")]
mean(Q7b$age) # Mean of Female's age is 40.8165
```



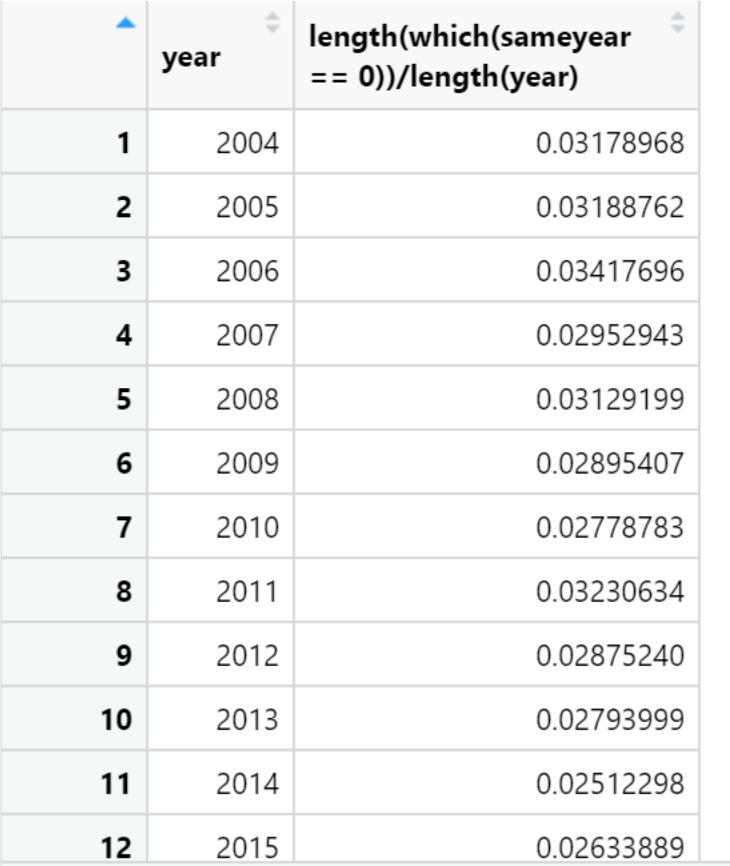
```
# Q8

dathh2011 = read.csv("Data/dathh2011.csv")

datind2011 = read.csv("Data/datind2011.csv")

Q8a = dathh2011[dathh2011$location == "Paris", c("idmen","location")]

Q8b = merge(datind2011, Q8a, by = "idmen")

nrow(Q8b) # There are 3514 individuals in Pris in 2011 dataset


# Exercise 2

# Q1

datind2004 = read.csv("Data/datind2004.csv")

datind2006 = read.csv("Data/datind2006.csv")

datind2007 = read.csv("Data/datind2007.csv")

datind2012 = read.csv("Data/datind2012.csv")

datind2013 = read.csv("Data/datind2013.csv")

datind2014 = read.csv("Data/datind2014.csv")
```

```
datind2015 = read.csv("Data/datind2015.csv")
datind2017 = read.csv("Data/datind2017.csv")
datind2018 = read.csv("Data/datind2018.csv")
datind = bind_rows(datind2004,datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,datind2011,datind2012,datind2013,datind2014,datind2015,datind2016,datind2017,datind2018,datind2019)
datind2007$profession = as.character(as.integer(datind2007$profession))
typeof(datind2007$profession)
datind2007$profession = as.character(as.integer(datind2007$profession))
typeof(datind2008$profession)
datind2008$profession = as.character(as.double(datind2008$profession))
typeof(datind2009$profession)
datind2009$profession = as.character(as.double(datind2009$profession))
typeof(datind2010$profession)
datind2010$profession = as.character(as.integer(datind2010$profession))
typeof(datind2011$profession)
typeof(datind2012$profession)
typeof(datind2013$profession)
typeof(datind2014$profession)
typeof(datind2015$profession)
typeof(datind2016$profession)
datind2016$profession = as.character(as.double(datind2016$profession))
typeof(datind2017$profession)
typeof(datind2018$profession)
typeof(datind2019$profession)
datind2019$profession = as.character(as.double(datind2019$profession))
datind2011$profession = as.character(as.integer(datind2011$profession))
datind2012$profession = as.character(as.integer(datind2012$profession))
datind2013$profession = as.character(as.integer(datind2013$profession))
```

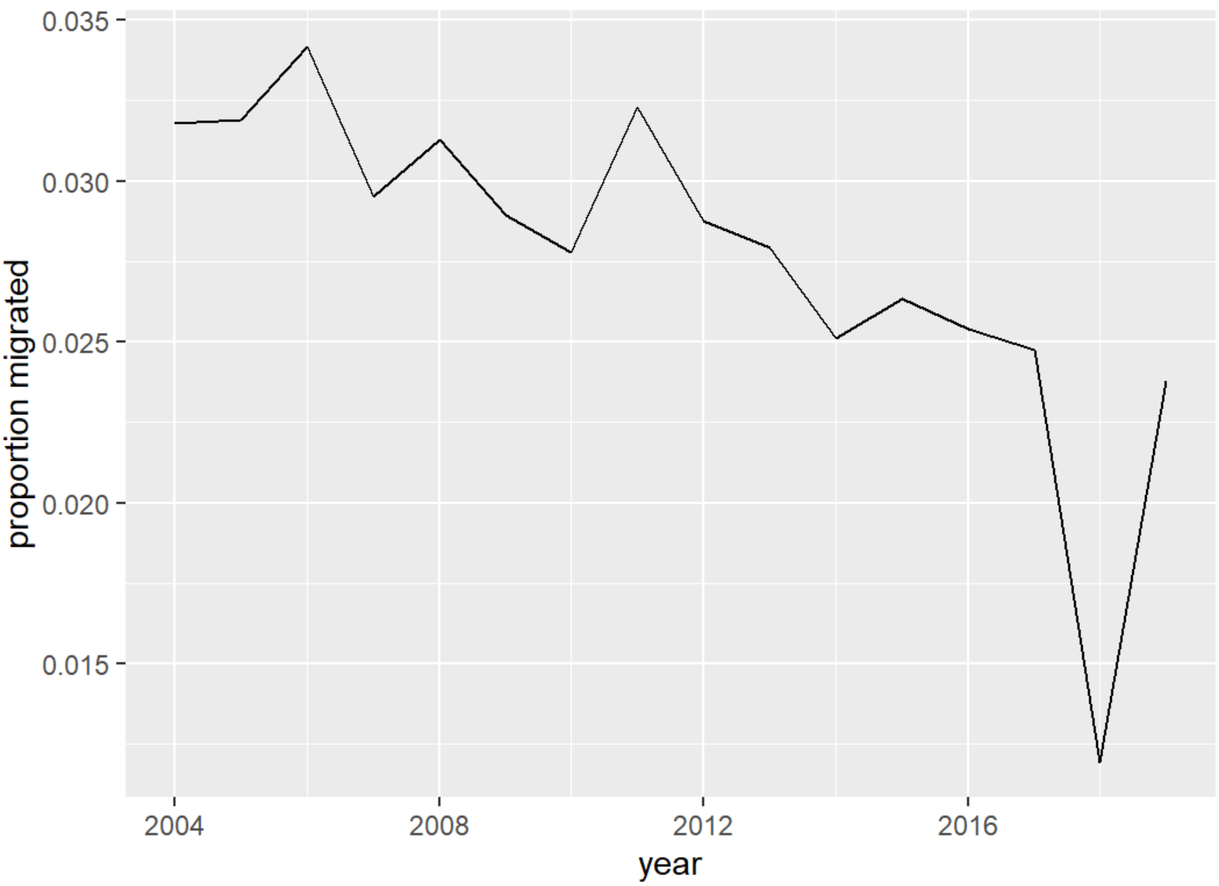
```
datind2014$profession = as.character(as.integer(datind2014$profession))
datind2015$profession = as.character(as.integer(datind2015$profession))
datind2017$profession = as.character(as.integer(datind2017$profession))
datind2018$profession = as.character(as.integer(datind2018$profession))
datind = bind_rows(datind2004,datind2005,datind2006,datind2007,datind2008,datind2009,datind2010,datind2011,datind2012,datind2013,datind2014,datind2015,datind2016,datind2017,datind2018,datind2019)
# Q2
dathh2004 = read_csv("Data/dathh2004.csv")
dathh2006 = read_csv("Data/dathh2006.csv")
dathh2008 = read_csv("Data/dathh2008.csv")
dathh2009 = read_csv("Data/dathh2009.csv")
dathh2010 = read_csv("Data/dathh2010.csv")
dathh2012 = read_csv("Data/dathh2012.csv")
dathh2013 = read_csv("Data/dathh2013.csv")
dathh2014 = read_csv("Data/dathh2014.csv")
dathh2015 = read_csv("Data/dathh2015.csv")
dathh2016 = read_csv("Data/dathh2016.csv")
dathh2017 = read_csv("Data/dathh2017.csv")
dathh2018 = read_csv("Data/dathh2018.csv")
dathh2019 = read_csv("Data/dathh2019.csv")
dathh = bind_rows(dathh2004,dathh2005,dathh2006,dathh2007,dathh2008,dathh2009,dathh2010,dathh2011,dathh2012,dathh2013,dathh2014,dathh2015,dathh2016,dathh2017,dathh2018,dathh2019)
# Q3
names(dathh)
names(datind) # "idmen" and "year" are simultaneously present in the two datasets
# Q4
dathhind = merge(dathh,datind,by = "idmen")
# Q5
```

```
dathhind = bind_cols(dathhind,matrix(1,nrow = nrow(dathhind), ncol = 1))
colnames(dathhind)[colnames(dathhind) == "...20"] = "num"
hhmem = dathhind%>%group_by(idmen)%>%summarize(household_mem = sum(num))
matrix1 = matrix(hhmem$household_mem, nrow = nrow(hhmem), ncol = 1)
matrix1 = as.vector(matrix1)
matrix1_new = matrix1[!matrix1 %in% c("1","2","3","4")]
length(matrix1_new) #there are 32122 households with more than 4 members
#Q6
unemp = dathhind[,c(1,13)]
unemp = unemp[(unemp$empstat == "Unemployed"),]
uniqueunemp = rapply(unemp, function(x) length(unique(x))) # 8161 households with at least one unemployed
# Q7 ???
twopro = dathhind[,c("idmen","profession")]
twopro = twopro[!is.na(twopro$profession),]
# Q8
couplekids = dathhind[,6]
as.vector(couplekids)
couplekids_new = couplekids[couplekids %in% "Couple, with Kids"]
length(couplekids_new) # There are 1200018 individuals that are from household-couple with kids
# Q9
paris = dathhind[,8]
paris = as.vector(paris)
paris_new = paris[paris %in% "Paris"]
length(paris_new) # there are 280463 individuals from Paris
# Q10
max(hhmem$household_mem)
hhmem_new = hhmem[hhmem$household_mem == 729,]
hhmem_new[1,] # idmen 2.202243e+15 has the most family member
```

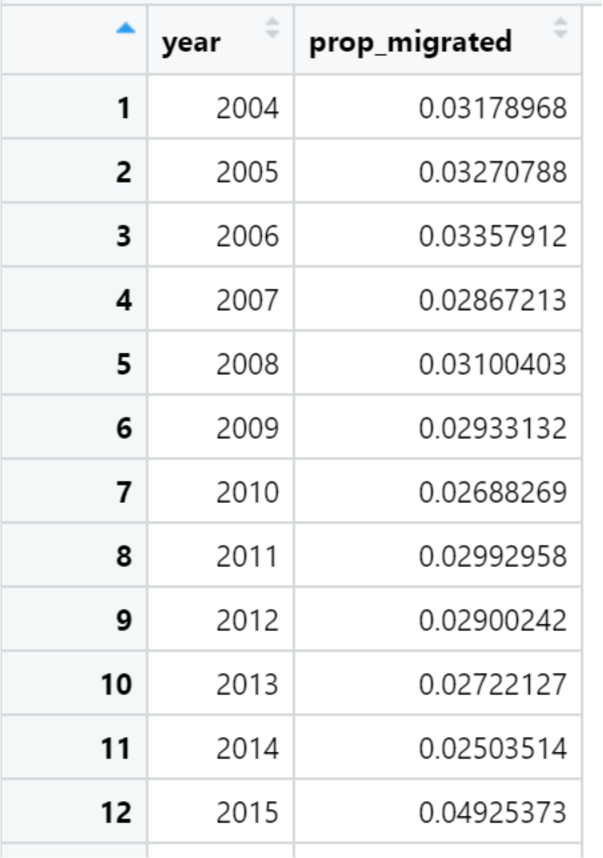
```
# Q11
hh2010 = dathhind[,c(1,3)]
hh2010_1 = hh2010[hh2010$year.x == 2010,]
hh2010_2 = rapply(hh2010_1,function(x) length(unique(x)))
hh2010_2 # There are 11048 households present in 2010
hh2011 = hh2010[hh2010$year.x == 2011,]
hh2011_1 = rapply(hh2011, function(x) length(unique(x)))
hh2011_1 # There are 11360 households present in 2011

# Exercise 3
# Q1
hhpanel_1 = dathh %>% group_by(idmen) %>% summarize(enter = min(year))
hhpanel_2 = dathh %>% group_by(idmen) %>% summarize(exit = max(year))
hhpanel_3 = dathh %>% group_by(idmen) %>% summarize(time = max(year)-min(year))
hist(hhpanel_3$time)
plot(density(hhpanel_3$time))
# Q2
datent = dathh[,c(2,3,4)]
datent["sameyear"] = datent$year-datent$datent
typeof(datent$sameyear)
datent$sameyear = as.character(datent$sameyear)
sameyear = datent %>% group_by(year) %>% summarize(prop_migrated =
length(which(sameyear==0))/length(year))
ggplot(sameyear,aes(x=year,y=prop_migrated))+geom_line()+ylab("proportion migrated")
```



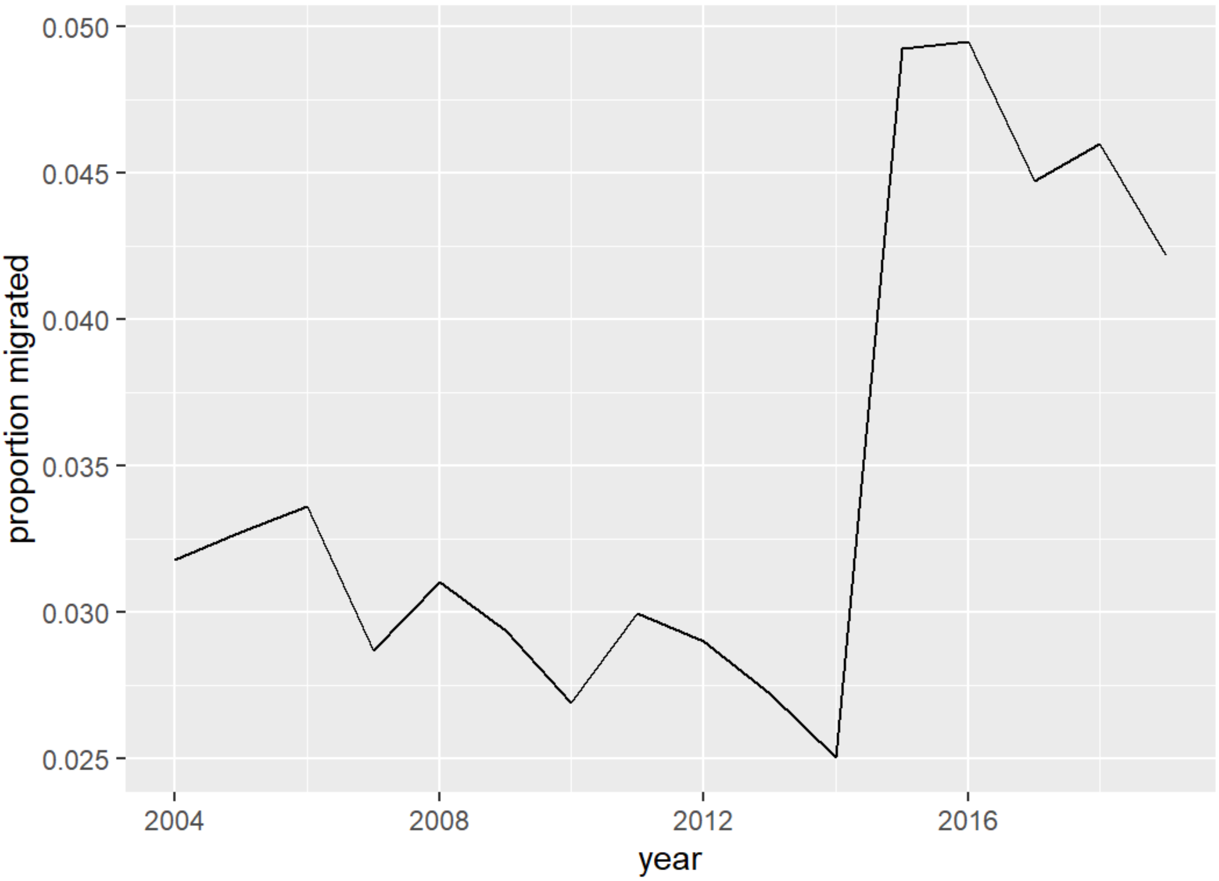
	year	length(which(sameyear == 0))/length(year)
1	2004	0.03178968
2	2005	0.03188762
3	2006	0.03417696
4	2007	0.02952943
5	2008	0.03129199
6	2009	0.02895407
7	2010	0.02778783
8	2011	0.03230634
9	2012	0.02875240
10	2013	0.02793999
11	2014	0.02512298
12	2015	0.02633889



```
# Q3
myyear = dathh[,c(2,3,5,7)]
myyear["same"] = myyear$year-myyear$myear
myyear_same = myyear[,c(1,2,3,5)]
myyear_same = myyear_same[myyear_same$year != 2015,]
myyear_same = myyear_same[myyear_same$year != 2016,]
myyear_same = myyear_same[myyear_same$year != 2017,]
myyear_same = myyear_same[myyear_same$year != 2018,]
myyear_same = myyear_same[myyear_same$year != 2019,]
sameyear0414 = myyear_same %>% group_by(year) %>% summarize(prop_migrated = length(which(same==0))/length(year))
move_same = myyear[,c(1,2,4)]
move_same = move_same[move_same$year != 2004,]
move_same = move_same[move_same$year != 2005,]
move_same = move_same[move_same$year != 2006,]

move_same = move_same[move_same$year != 2007,]

move_same = move_same[move_same$year != 2008,]

move_same = move_same[move_same$year != 2009,]

move_same = move_same[move_same$year != 2010,]

move_same = move_same[move_same$year != 2011,]

move_same = move_same[move_same$year != 2012,]

move_same = move_same[move_same$year != 2013,]

move_same = move_same[move_same$year != 2014,]

sameyear1419 = move_same %>% group_by(year) %>% summarize(prop_migrated = 
length(which(move==2))/length(year))

Ex3Q3 = bind_rows(sameyear0414,sameyear1419)

ggplot(Ex3Q3, aes(x=year,y=prop_migrated))+geom_line()+ylab("proportion migrated")
```

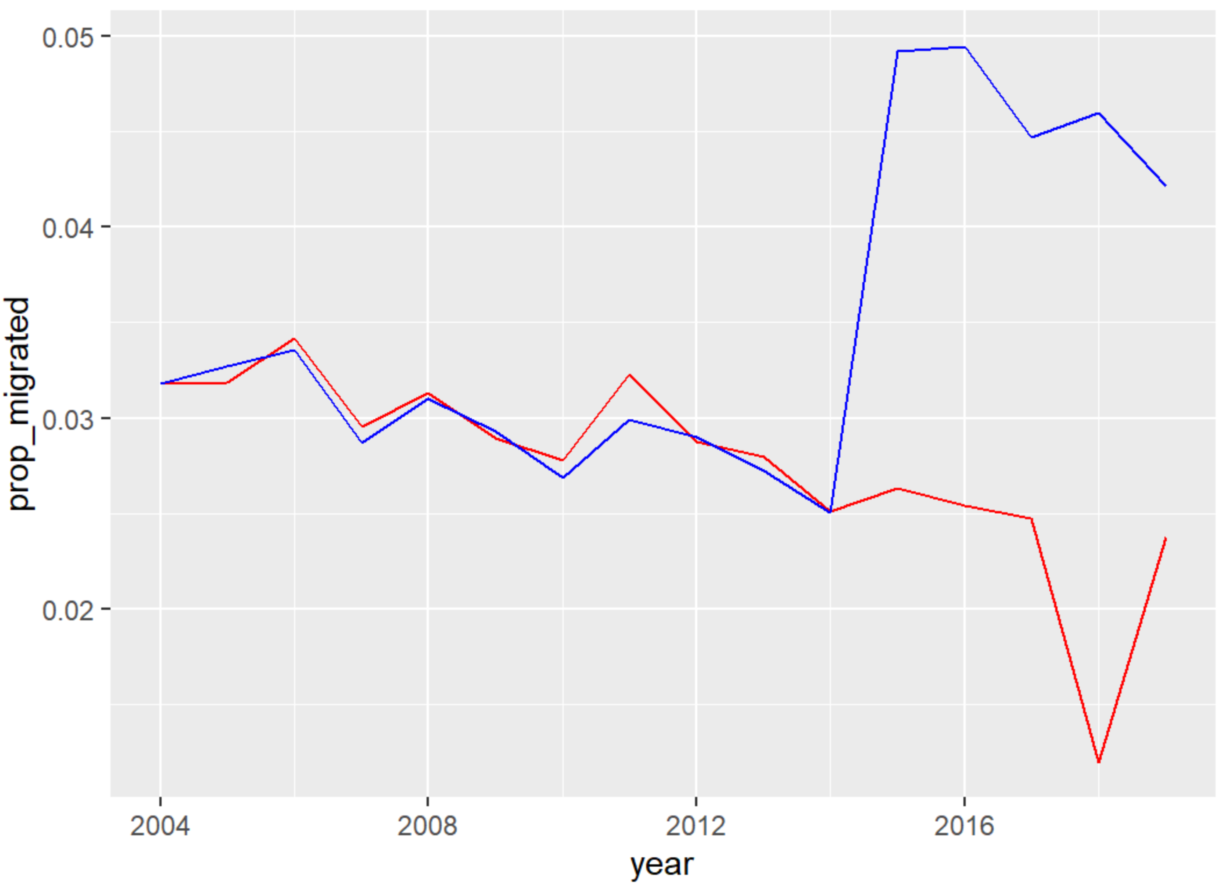


	year	prop_migrated
1	2004	0.03178968
2	2005	0.03270788
3	2006	0.03357912
4	2007	0.02867213
5	2008	0.03100403
6	2009	0.02933132
7	2010	0.02688269
8	2011	0.02992958
9	2012	0.02900242
10	2013	0.02722127
11	2014	0.02503514
12	2015	0.04925373

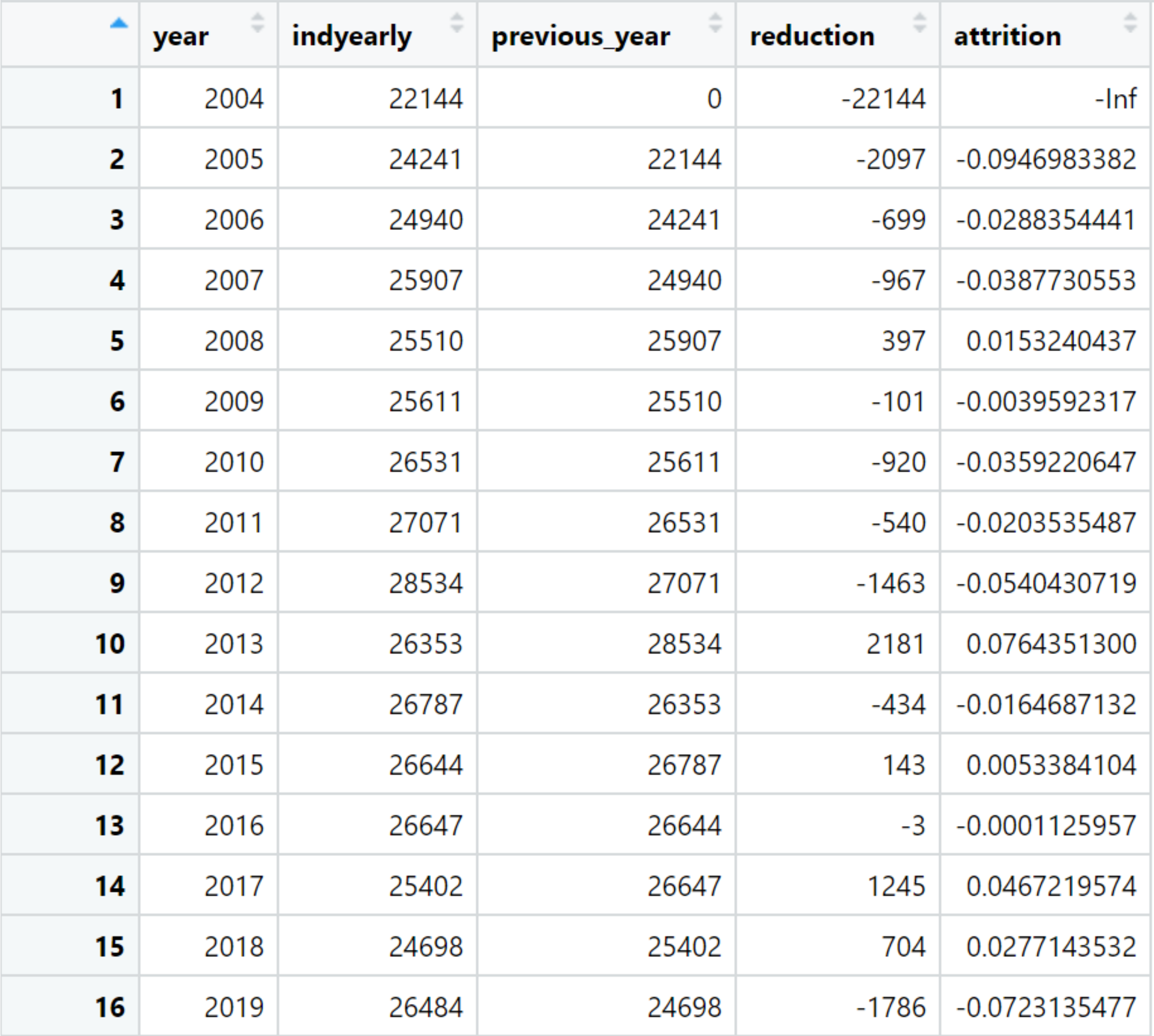


# Q4 I prefer the method in Q2 because it tells us more about what happens in the most recent year. Whereas method in Q3 tells us movement since last survey.

```
ggplot(NULL,aes(x=year,y=prop_migrated))+geom_line(data=sameyear,col="red")+geom_line(data=Ex3Q3,col="blue")
```



```
# Q5????

mighh = datent[datent$sameyear==0,]

mighh = mighh[!is.na(mighh$sameyear),]

migindpro = merge(mighh,dathhind,by="idmen")


# Exercise 4

indeachyear = datind %>% group_by(year) %>% summarize(indyearly = length(year))

vec1 = c(0,22144,24241,24940,25907,25510,25611,26531,27071,28534,26353,26787,26644,26647,25402,24698)

indeachyear["previous_year"] = vec1

indeachyear["reduction"]=indeachyear$previous_year-indeachyear$indyearly

indeachyear["attrition"]=indeachyear$reduction/indeachyear$previous_year
```



	year	indyearly	previous_year	reduction	attrition
1	2004	22144	0	-22144	-Inf
2	2005	24241	22144	-2097	-0.0946983382
3	2006	24940	24241	-699	-0.0288354441
4	2007	25907	24940	-967	-0.0387730553
5	2008	25510	25907	397	0.0153240437
6	2009	25611	25510	-101	-0.0039592317
7	2010	26531	25611	-920	-0.0359220647
8	2011	27071	26531	-540	-0.0203535487
9	2012	28534	27071	-1463	-0.0540430719
10	2013	26353	28534	2181	0.0764351300
11	2014	26787	26353	-434	-0.0164687132
12	2015	26644	26787	143	0.0053384104
13	2016	26647	26644	-3	-0.0001125957
14	2017	25402	26647	1245	0.0467219574
15	2018	24698	25402	704	0.0277143532
16	2019	26484	24698	-1786	-0.0723135477