Hamzah Ahmed
Econ 613
Assignment 1

General Assumption Made: Every data entry in the given data files (datind and dathh) are unique (TAs also specified this).Later on, I extracted the unique id characters but duplicates (which should be few and far between) are errors of the data collector. This is important to point out because, due to error in data procurement, there were issues of uniqueness in id numbers (in some cases). Duplicated ID numbers (such as idind) won't be removed because the characteristics of each are, generally different (ie have different values for other variables) so it has most likely been an error on the part of those gathering the data (although this is not always the case). Also skip missing values where applicable.

Also had to read in id's as characters so that I could get the accurate ID. This was fine because the IDs were mostly used for matching.

Also just an fyi, I initially did a lot of this using many nested for loops and that took a while. I tried to simplify as much as I could using matrices and other techniques and I just commented out my other code and there are instances where I give options for answers and some parts of the code may need to be uncommented to have it run properly!

Exercise 1

a. Looked at number of rows in dathh2007. In Slack, people were talking about redundancy issues so I also used the pipeline method to check for uniqueness as well. In either case, I had gotten:

   [1] 10498: number of households surveyed in '07

   In general, as said above, I will be assuming that each data entry within a given dataset is unique and any id duplicates are the fault of the data collector.

b. Used dathh2005. Used table function and just looked at occurrences of Couples, with Kids

```
 Couple, No kids Couple, with Kids            Other           Single
            2656              3374              275             2663
   Single Parent
             785
```

Here, the output was 3374 households that were couples with kids in 2005 (according to survey data).

Hamzah Ahmed
Econ 613
Assignment 1

c. Used datind2008. Again, double checked with pipeline method to insure uniqueness as well as number of rows. Output was:

   [1] 25510: number of individuals surveyed in '08

d. Used datind2016 and used inclusive bounds for ages (in a condition). Output was:
   [1] 2765: individuals surveyed between 25 and 35 (inclusive) in '16

e. Used table function and datind2009. Professions were categorical variables so they are just numbers (mapped to certain profession. Output was (in 2009):

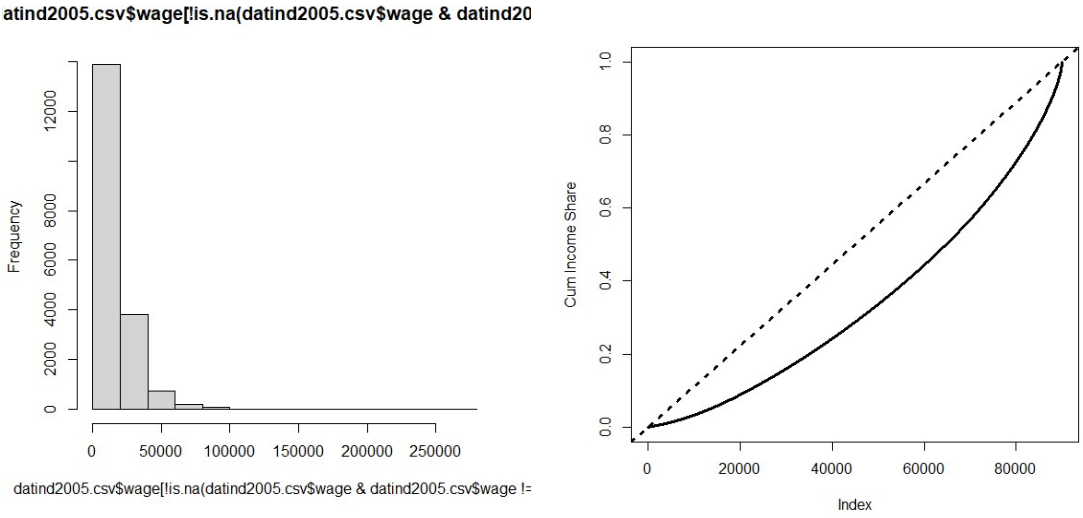
	0	11	12	13	21	22	23	31	33	34	35	37	38	42	43	44	45	46	47	48	52	53	54	55	56	62	63	64	65	67	68	69
Female	11	30	8	29	63	65	8	68	85	184	50	179	78	258	437	1	153	410	82	22	782	27	584	353	696	64	35	29	19	147	120	40
Male	19	57	19	78	213	114	48	98	107	142	59	260	368	110	117	2	95	340	429	215	169	182	98	101	74	443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

f. Unclear if we were supposed to merge the data or not so I did it both ways.
   General Note: I omitted the missing data (!is.na()) as, mostly, you can't do many function operators with that data but also, we are looking at the datasets individually and looking at distributions – NAs can't cleanly be implemented in them. I also omitted 0's because the TA told us to.

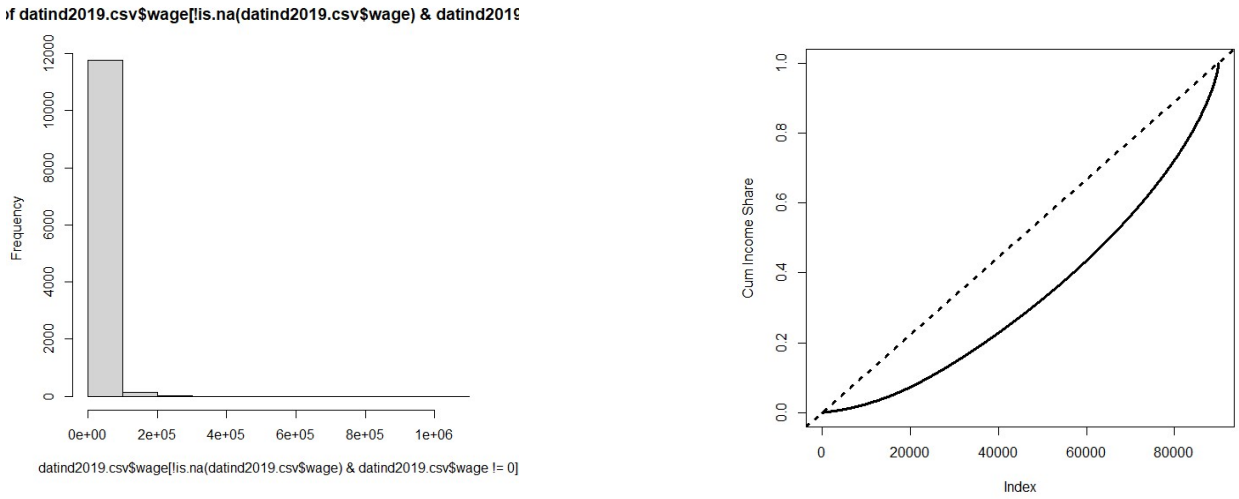
   For Gini function, for the sake of simplicity, I found the decile values and then assumed there was just 1 person in each "bucket" so that finding total income was straight forward and thus the cumulative sum ratios along the way were easy to calculate (just the decile values summed together). I also made the jumps relatively small for accuracy sake – could make smaller but takes too much time. Coefficient was area between curves over area under full equality. These Gini Coefficient values are most likely, not entirely accurate, but sufficiently close. Also plot a rough Lorenz Curve and Perfect Equality Curve so we can see inequality. Results begin on next page.

Hamzah Ahmed
Econ 613
Assignment 1

a. 2005 Separate:
   mean:[1] 22443.02911846829
   sd:[1] 18076.708881794755
   Inter-Decile Ratio: 8.8965251814383119
   D1: 4547 and D9: 40452.500000000007
   Gini: [1] 0.3201291433600677

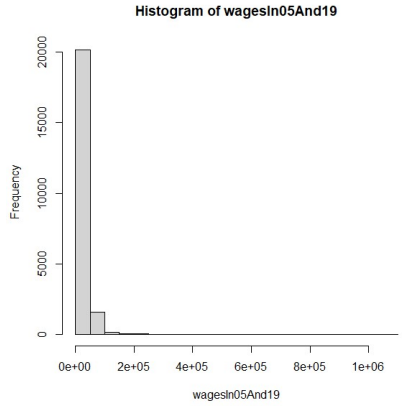


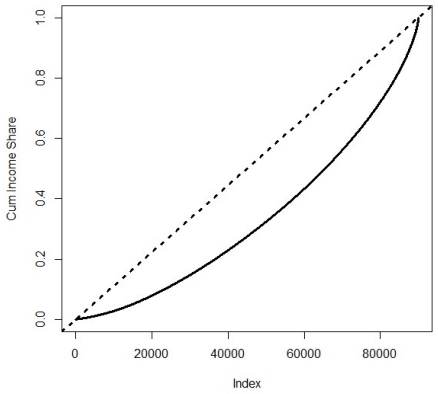
b. 2019 Separate:
   mean: [1] 27578.839302189048
   sd: [1] 25107.187195539096
   Inter-Decile Ratio: 13.862300495321959
   D1: 3634.0000000000005 and D9: 50375.600000000006
   Gini: [1] 0.33903644704838326



Hamzah Ahmed
Econ 613
Assignment 1

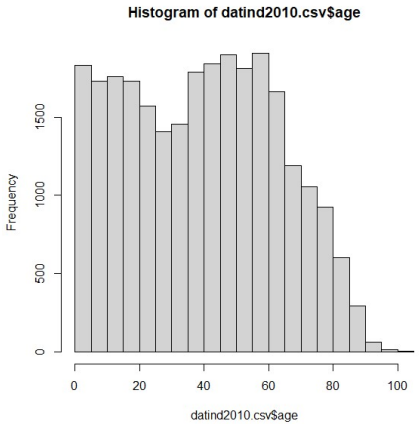
c. Together (combining 2005 and 2019)
mean: [1] 25232.617967290786
sd: [1] 22320.350540064475
Inter-Decile Ratio: 11.580806815334503
D1: 3991 and D9: 46219
Gini: [1] 0.33725692054417167





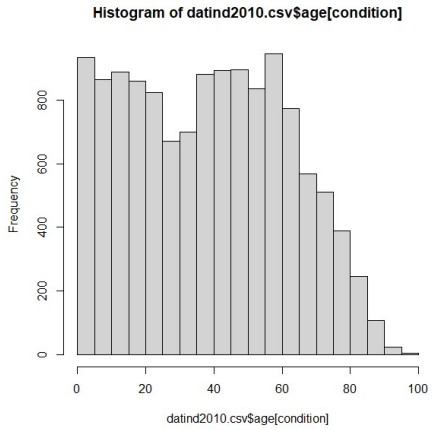
g. Used hist for aggregate and then looked at histogram for males and females (to see if there is a difference). Also did summary to get summary statistics

Aggregate



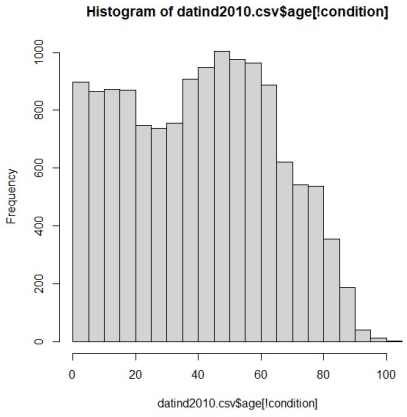
min = 0, 1st Q = 19, mean = 39.88, median = 40, 3rd Q = 58, max = 102

Males



min = 0, 1st Q = 19, mean = 38.87, median = 39, 3rd Q = 57, max = 96

Females



min = 0, 1st Q = 20, mean = 40.82, median = 42, 3rd Q = 59, max = 102

Hamzah Ahmed
Econ 613
Assignment 1

Similar, but not EXACT, distributions so there is a difference between male and female (important to notice left side frequencies are different (more females) but overall trends seem similar to each other and to the aggregate but females are more skewed and have a higher age frequency (higher mean, median, max, and IQR) so there is a difference between men and women.

h. Location information in households so looking at individual data, connecting them to household data, and then counting number of those in Paris (first checking that data is not missing on location). Assuming each individual is associated with only one household (ie household may have multiple individuals but individuals should only belong to one household). Also did it through pipeline, counting number of individuals in each household and then adding if that household was in Paris – seemed more straightforward. Either way, Output was:
   [1] 3514: Individuals in Paris in 2011

Hamzah Ahmed
Econ 613
Assignment 1

Exercise 2

a. Previously read in all the individual data. Aggregated them together using rbind. Aggregate called indData.
b. Same process as above but for household data. Aggregate called hhData.

Weird Issue I noticed: the number of unique idmen in the individual dataset was 41086 but the number of unique idmen in the household dataset was 41084 (2 less households). I think this may lead to some differences of answers based on what methods were used. Number of unique idind was 100160 (was much less when I initially did not read it in as characters).

c. Looking at column names for each dataset

```
>   colnames(indData)
 [1] "X"          "idind"      "idmen"      "year"
"empstat"    "respondent" "profession"
 [8] "gender"     "age"        "wage"
>   colnames(hhData)
[1] "X"        "idmen"    "year"     "datent"   "myear"
"mstatus"  "move"     "location"
```

Variables simultaneously in both: "X" "idmen" "year"
(X is in it as a byproduct of how R reads in the data – not used when merging) was found using intersect. 413501 rows of Data.

d. Aggregated datasets together and merged by idmen and year (not X because it would mess things up).

For the following problems, I mainly used pipelines to group by idmen and year and then went about finding the values asked for by specific summary() stuff. For me, it made sense to also group by year for households, otherwise, we would be looking at aggregate data at household level (ie n() would give us, in total, how many people were ultimately associated with a given household and repeats were allowed over years) and this seemed less helpful (given the questions asked). In general, when asked for the number of something, I counted occurrences that happened for the same individual/household across years (so if something happened 3 years in a row, it was counted 3 times).

e. Used pipeline to group by idmen and year and counted number in each grouping. Checked for households that were repeated more than 4 times (had more than 4 associated individuals) in a given year to find number of households with at least four

members in a given year. Output was:
[1] 12436

f. Using pipeline methods, grouped by idmen and year and then checked if "Unemployed" was in empstat variable of each group (using %in%) in a given year and then counted occurrences of TRUE (would be TRUE if at least one member was "Unemployed" in household in a given year. Output was:
[1] 17241

g. Similar pipeline method used, this time, used anyDuplicated to see if there were any duplicate professions in a household in a given year. Output would be 0 if all unique and >=2 otherwise (position of first duplicate). I dropped the missing data but I also noticed a lot of blank data "" which I also dropped as it was different than zero and probably meant to be missing data. I checked for instances of being non-zero and added those instances together. Output was:
[1] 47501

h. Each row, in theory, should be a unique individual in aggData. As we merged data, we can now see if each individual was in a household that was a "Couple, with Kids" and skipped missing data. I simply added up these occurrences. I assumed we were looking at data per year. So if an individual was in a house "Couple, with Kids" for 3 consecutive years, they would be counted three times (once for each year). Output was:
[1] 209382

i. Pretty much identical treatment as above. Cycled through rows and checked if each (presumably unique) individual in aggData was in Paris at the time of their survey. This time, I just checked if the household was in Paris in a given year and then added all the people. Output was:
[1] 51904.
If, instead, the question meant to ask how many people were in Paris at some point in time in the survey, I grouped the aggData by idind and checked if Paris was in set of their responses to location. The output was:
[1] 14563 unique individuals in Paris at some point in time in this survey.

j. Same pipeline as earlier (grouping by idmen and year). Then found max count in a household (14) in a given year. From there, identified which household(s) it was (in data order by

index). There were 2 and the rows were 69653 & 107200.
Corresponding idmen were:
[1] "2207811124040100" in 2007
[2] "2510263102990100" in 2010
2 different households had maximum individuals corresponding to household (14) in 2 different years.

k. There were a total of 22408 households in the 2010 and 2011 datasets combined (by adding the number of rows in the corresponding household datasets).

However, the question, clarified on Slack, tells us to look for how many households were present both in 2010 and 2011. I grouped by idmen and checked if these households were present in the data in 2010 **and** 2011. The Output was:
[1] 8984 households present in 2010 and 2011

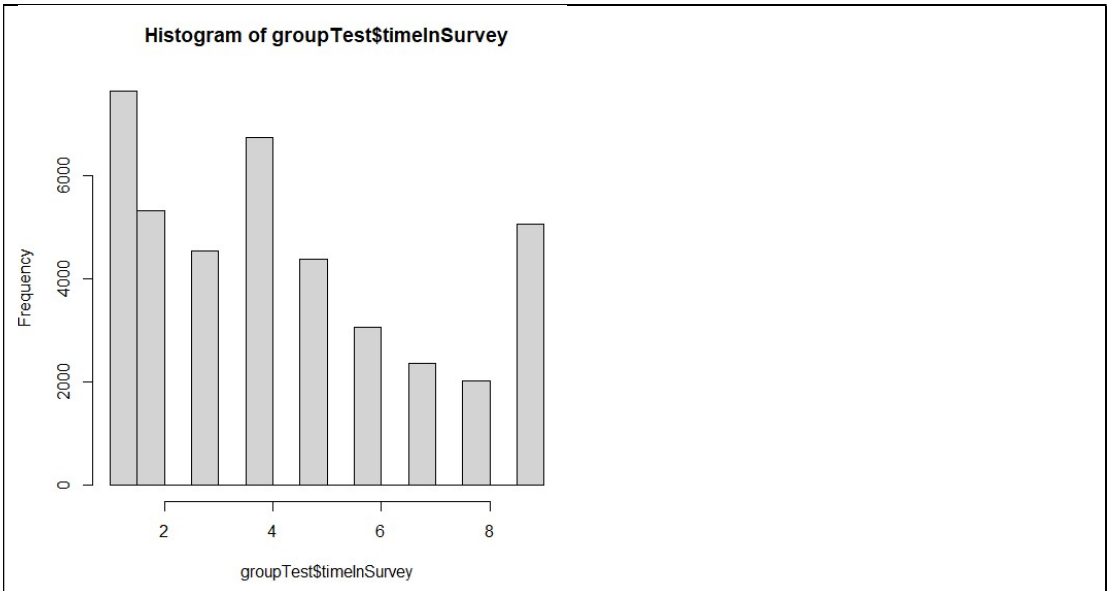
Hamzah Ahmed
Econ 613
Assignment 1

Exercise 3

a. First, I created a vector with the unique household numbers (idmen) just to see. There were 41084 in total. I used the pipeline method and grouped by household. I summarized the timeInSurvey by calculating the difference between the largest year (exit year) and the smallest year (entry year) and adding 1. A key assumption I made was that the last occurrence of the household in the survey data was their last year (so if they, for some reason, were omitted for an intermediate year but came back later, they were still counted). More explicitly, if household j was counted in 2004-2006, and 2008 (not in 2007), their duration in the survey would be 2008-2004 + 1 = 5 years. The TA said not to worry about cases of exit and reentry so I feel as though my assumption is valid.
I also created a code (that is commented out) that stores the exit and entry years in a matrix but the question did not sound like we needed to report/store them but just find the difference using them. Using which() function, I found the first and last occurrences and then added them to RHS by extracting the entrance and exit years and finding the difference and adding 1 (which is time in survey). The output is very long and is stored in timeInSurveyMat.

   For a visual, I also plotted the histogram to see the general distribution. To try to write more efficient code, I also grouped by idmen and then summarized the time in survey as the max(year) – min(year) + 1. I was able to get the same histogram and summary stats much quicker. 1Output of distribution was:

Hamzah Ahmed
Econ 613
Assignment 1



Min = 1, 1st Q = 2, Med = 4, mean = 4.3123, 3rd Q = 6, Max = 9

Hamzah Ahmed
Econ 613
Assignment 1

b. This question asks us to determine whether a household (in a given year) moved into its current dwelling at the year of survey (based on datent). It also asks us to plot the share of individuals in that situation across years. I found the first part of this question more straightforward to do just using the aggregated household data and grouping by idmen and year and checking if year == datent. Results would be TRUE if they were equal and FALSE otherwise (ie TRUE if they moved into the dwelling at the year of survey). The first 10 rows (idmen and year) can be seen below:

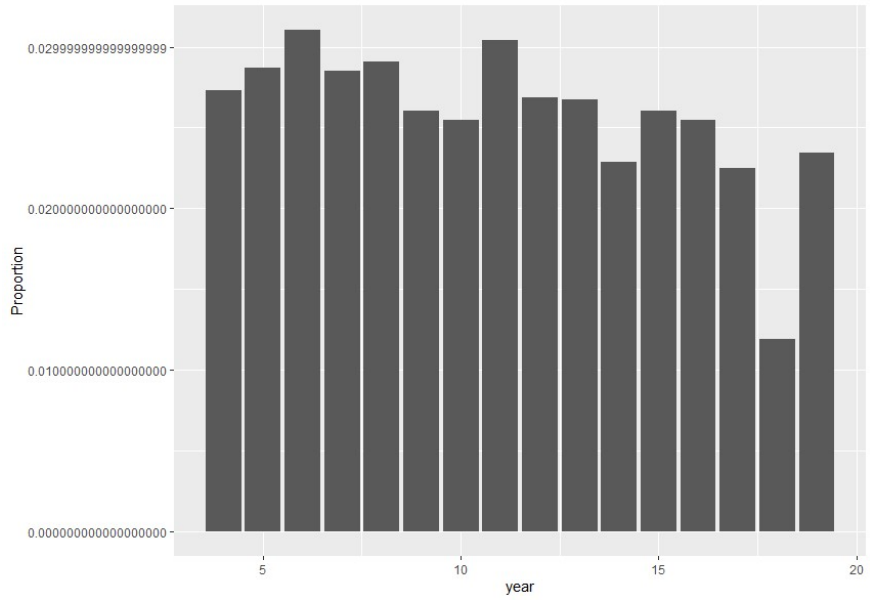
We can see, from the first ten (grouped by idmen and year), only 1 household had datent == year (moved at year of survey)

	idmen	year	movedAtYearOfSurvey
1	1200010012930100	2004	FALSE
2	1200010040580100	2004	FALSE
3	1200010040580100	2005	FALSE
4	1200010066630100	2004	FALSE
5	1200010066630100	2005	TRUE
6	1200010082450100	2004	FALSE
7	1200010082450100	2005	FALSE
8	1200010086440100	2004	FALSE
9	1200010086440100	2005	FALSE
10	1200010102990100	2004	FALSE

The second part of the question asks us to plot the share of individuals in that situation across years. I made 2 groupings(by year) one which summarized the the number of people who migrated each year and another which summarized the total number of people each year and divided the two output to get share of movement over year by individuals (assumed each row was a unique individual (duplicate idind due to errors in data collection) so I didn't drop them before doing this.

Hamzah Ahmed
Econ 613
Assignment 1

Share of Individuals that had datent = year (notice highest proportion is around .03 so very small). Fairly uniform



c. Assuming Migration is TRUE if myear == year (moved in that year) OR move == 2 (assuming 2 codes for moved that year). myear exists for data up to 2014 and move exists afterwards.

Question asks us to do similar process as before, but using 2 other variables to check. I again broke this down into two parts with very similar code with similar logic/processes as before – the checks were just different.

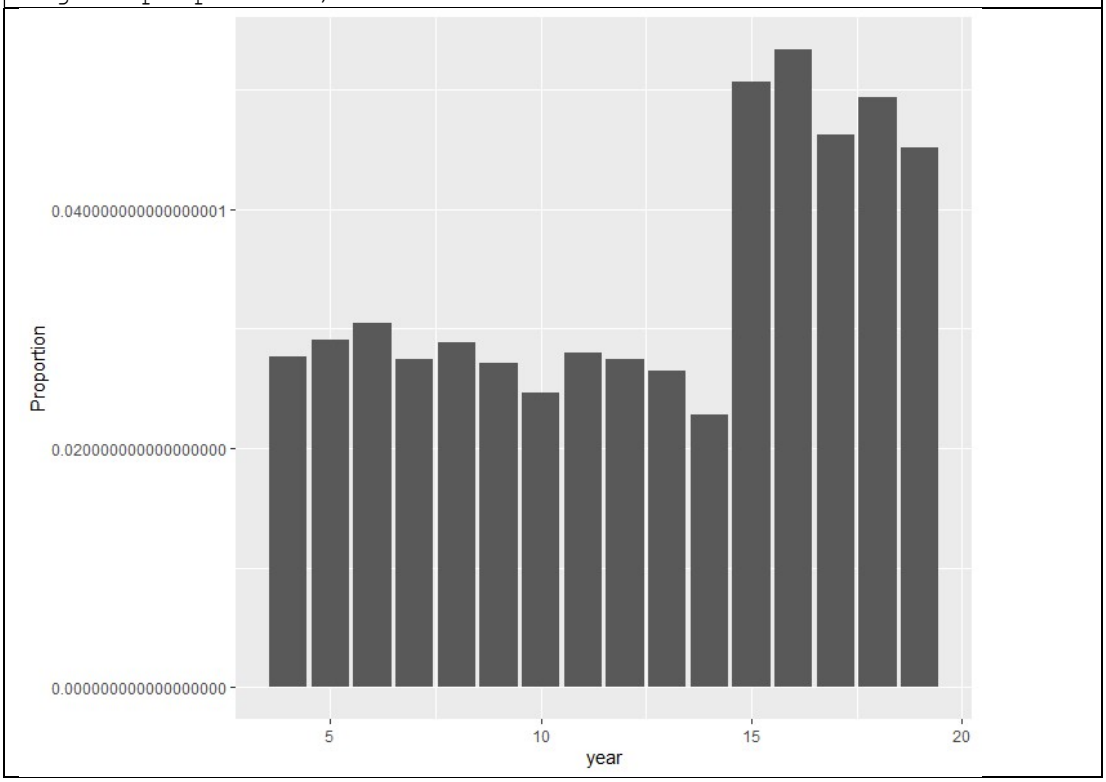
Similar results as before

	idmen	year	migration
1	1200010012930100	2004	FALSE
2	1200010040580100	2004	FALSE
3	1200010040580100	2005	FALSE
4	1200010066630100	2004	FALSE
5	1200010066630100	2005	TRUE
6	1200010082450100	2004	FALSE
7	1200010082450100	2005	FALSE
8	1200010086440100	2004	FALSE
9	1200010086440100	2005	FALSE
10	1200010102990100	2004	FALSE

Hamzah Ahmed
Econ 613
Assignment 1

Again, using individual data, but checks are different now. Before 2014, Migration is when year == myear and, after 2014, Migration is when move == 2 (my assumption) and similar process of 2 groupings (1 for migration and 1 for total). Barplot was:

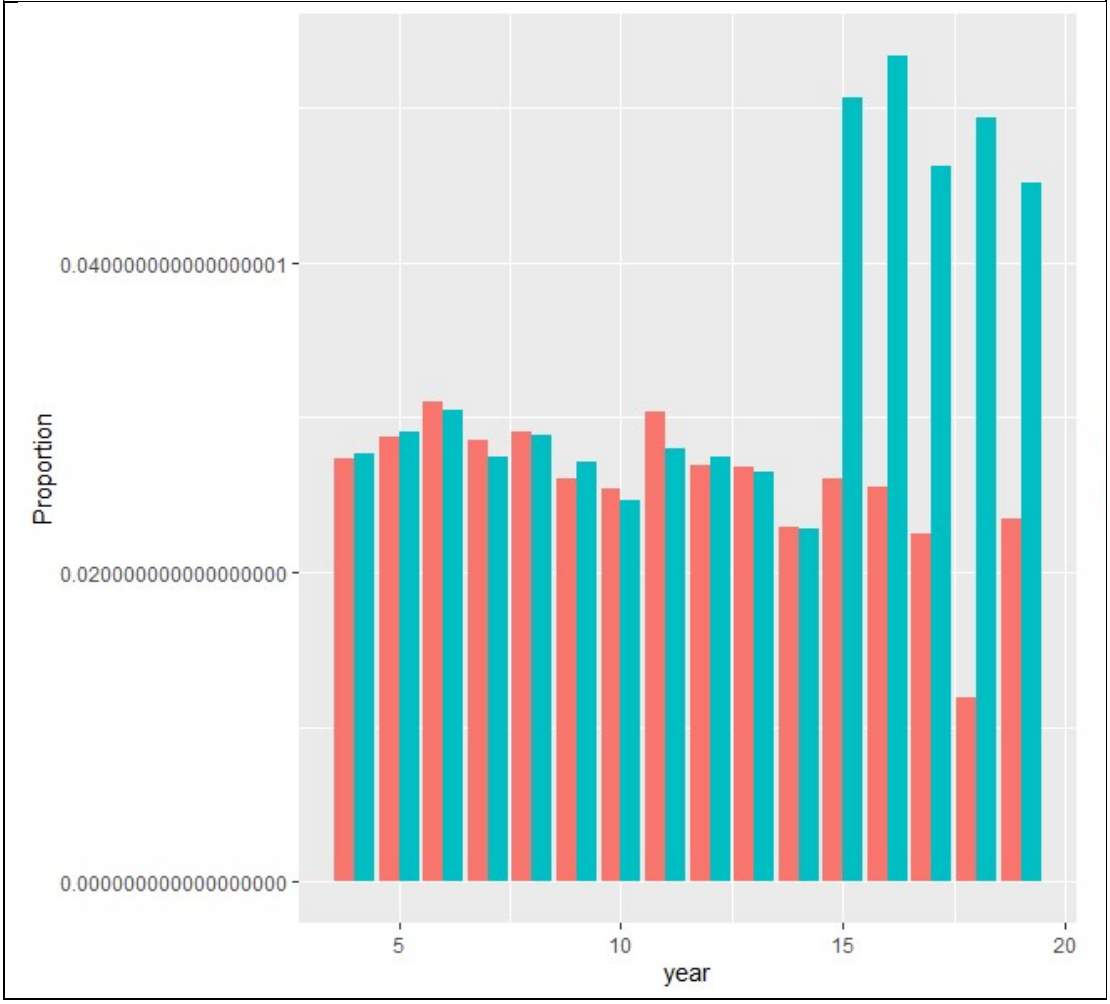
Much more skewness. More people moved in recent years (and a higher proportion).



Hamzah Ahmed
Econ 613
Assignment 1

d.

Method1 was using datent and Method2 was using myear or move (depending on year of survey)



Interesting that the methods had, almost identical, migration proportions until 2015 but they drastically alter (method2 showing higher proportions) when we switched to move. I would lean towards method1 for 1 large reason: the amount of missing data for datent is much smaller than that of myear(until 2014) and move(after 2014). It is 245 vs 6446 + 25250 so this lack of evidence, for this statistic, makes me prefer method1 as there is more data available. In theory, these migration values shouldn't be different at all which also highlights the real likelihood of data collection errors which drastically alters results depending on which method is used as there were time inconsistency issues in the data.

e. This process took multiple steps and I will try to briefly explain what I did. Firstly, I used the migration data from method2 and merged those results with my aggregate data (by idmen and year) to get a newAggData and added a column of 0's that would indicate job change for that individual in that year. I did this because method1 was described as moved and method2 was described as migrated so I assumed that method was meant for migration (as stated in the problem) – I also ran the code with method1 at the end because I believed it was better (less missing). I took a subset of the newAggData that only held all the individuals who were a part of a household that migrated and called it migratedHouseData. I extracted the unique idind's in this new subset. Then, I went through each ID and checked empstat and profession compared to the previous year (if this individual was surveyed multiple times). I made sure to skip people who had missing empstat/professions in a two year span (ie they had no data for at least one of the years in which they were in migration). When there was a profession or empstat change in a year (compared to the profession/empstat of the individual in their last survey), I updated the last column in the migratedHouseData to 1. One thing I should point out is that I made my decisions based on if the individual had a change since their last survey; I did not check that their last survey was the previous year as the question didn't specify something like that as another condition. Using the pipeline method, I grouped by idmen and year and summarized the sum of the Change variable in that group (household in a year) and if a migration occurred. If nobody changed profession/empstat in a household in a year, the sum should be zero. I counted the instances where the sum was NOT zero (ie at least one person changed a job from the previous year in that household in that year) and a migration at the same time. The output was:

[1] 518 if using method2 (for migration)
or
[1] 312 if using method1 (for moving) which I thought was better as it had less missing data. Makes sense that its larger as there are less missing data skipped.

Hamzah Ahmed
Econ 613
Assignment 1

Exercise 4

a. By the hint, I created a table that had: idind, entry year, and exit year via grouping by idind and year and said the first year was their entry year and the last year was their exit year (didn't account for leaving and coming back). Essentially, I added up the total number of people in the data per year and created totalInd. Concurrently, I checked if, at each year, this was the last time an individual was in the data (ie if the year was their exit year). Adding all the people who were at their last year (and then would leave) gave me attritionPerYear (ie people who left after that year) as this would show the people leaving each year. The ratio of those two gave me the proportions of attrition per year where I viewed attrition as the people who, once in this current data panel, would not return (ie that year was there last year) and thus would leave. By using the hint, I presumed that people did not exit and then reenter and then exit the data (once you leave, you are gone for good). The table of proportions is below:

Hamzah Ahmed
Econ 613
Assignment 1

	years	propAtrrition
1	2004	0.11596820809248555
2	2005	0.18002351128947261
3	2006	0.16151297625621203
4	2007	0.20907050184529924
5	2008	0.18760132787771172
6	2009	0.16840324503056633
7	2010	0.17346976744186046
8	2011	0.15602553870710295
9	2012	0.21096051856216852
10	2013	0.18614624218636430
11	2014	0.18920394569213408
12	2015	0.18961038961038962
13	2016	0.21249587261987746
14	2017	0.20849655801924466
15	2018	0.23640940917315711
16	2019	1.00000000000000000