Final Project

Yigit Tahmisoglu

2023-04-02

Github repository: https://github.com/tahmisoglu-yigit/R_Project.git

Requirements: 1. 1-page executive summary (can include 1 graph or image) \ 4. Data exploration, questions you tried to answer, interesting things. This should be similar to the data exploration in R for Data Science (Chapter 7) and most other parts of the book. No hard results needed. (3-5 pages)\ 5. Code: Include one piece of code that you describe in some detail what it does. Pick your favorite graph or table, and explain the choices you made and what each step does. A pipeline (|>) of 4 to 7 steps is sufficient. The goal is to have you write and explain code.

Summary of the raw data set

The data set is the one I plan to use for my masters thesis. It is a cross-sectional survey data from the Demographic and Health Surveys for Turkey from the year 2018, which is a representative household survey providing detailed information on birth recodes and individual records for women of each household for developing countries. The program conducted its first survey in Turkey in 1993 and it has been conducted once in every five years. The dataset provides nationally representative individual and household level categorical data on household characteristics, living conditions, respondent background characteristics (e.g. fertility rates, education levels, child mortality rates, employment status), child health, family planning, domestic violence to name but a few. In this analysis, I will try to show the relationship between age at first childbirth and different variables such as education, type of residence, parental education, employment status, consent for marriage etc.

Explain how you cleaned the data, or if it was clean, how you ensured that it was. Clearly mention any R scripts that did the checking. (It should be automated, that means no "I printed it a few times to the screen and stared at it.") (1-2 pages, less if data was clean) :

Install libraries:

Variable names are not descriptive. But taking the DHS Manual as reference, we can rename the variable of interests to make further analyses easier.

```
df1 <- df |>
  rename(birth_month=V009,
         birth_year=V010,
         age = V012,
         region = V024,
         residence = V025,
         residence_childhood=V103,
         educ_level=V106,
         religion=V130,
         educ = V133,
```

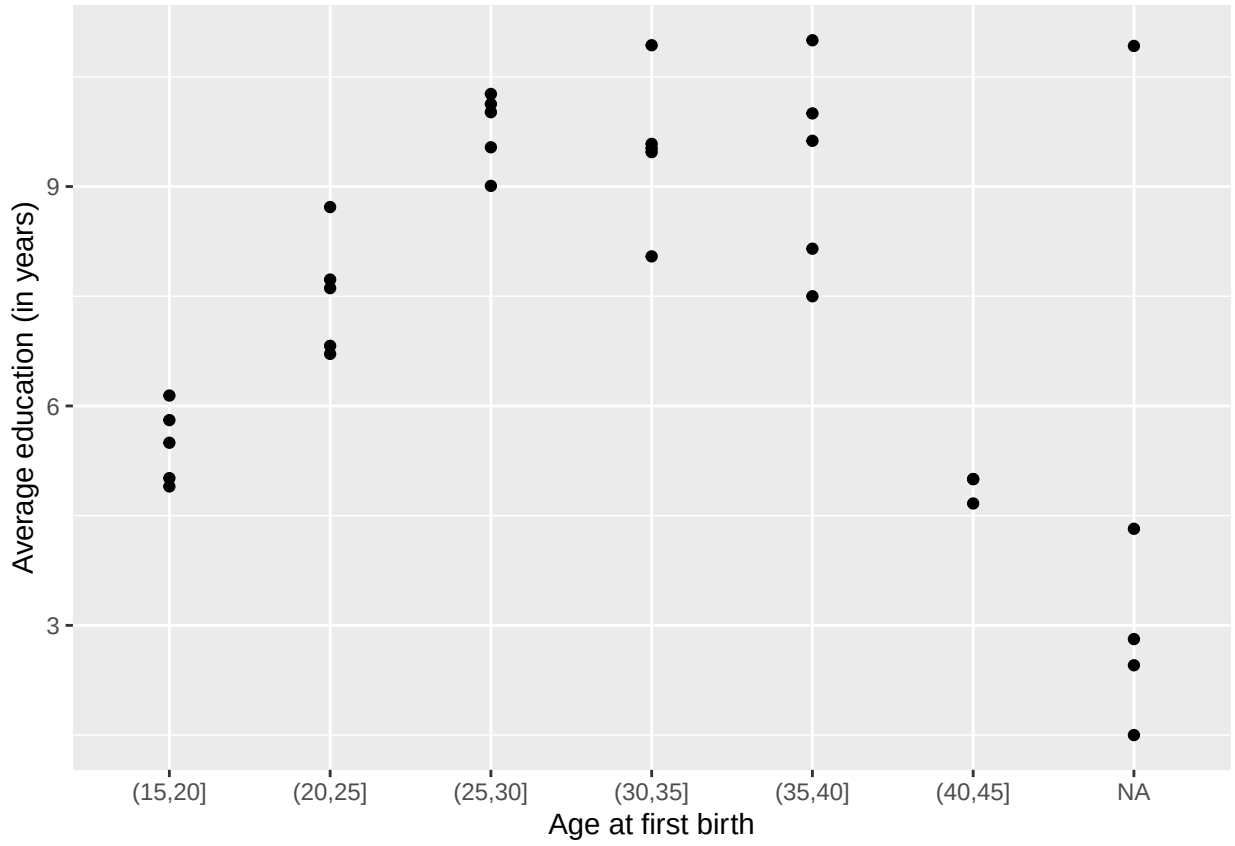
```
         age_at_birth = V212,
         educ_mother = S119,
         educ_father = S121
         )
# mean(df1$age_at_birth, na.rm = TRUE)

# Feature engineering
df1$region <- recode(df1$region,
                     "1" = "west",
                     "2" = "south",
                     "3" = "central",
                     "4" = "north",
                     "5" = "east")
df1$educ_level <- recode(df1$educ_level,
                     "1" = "no educ",
                     "2" = "primary",
                     "3" = "secondary",
                     "4" = "higher")
```

```
## Warning: Unreplaced values treated as NA as `.x` is not compatible.
## Please specify replacements exhaustively or supply `.default`.
```

```
df1$residence <- recode(df1$residence,
                     "1" = "urban",
                     "2" = "rural")
```

Plotting the relationships:




```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```