VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

FALL SEMESTER 2025-26

FOUNDATION OF DATA SCIENCE

SHIULI- 22BCE0342

PROJECT TITLE

Socioeconomic and Demographic Factors Influencing Fertility: A Regression and Time Series Analysis of General Social Survey Data (1974-2002)

02

PROJECT ROADMAP

A Problem Identification & Objective Definition

B Data Collection & Understanding

C Data Cleaning & Preprocessing

D Clustering & Dimensionality Reduction

E Evaluation & Insights

F Visualization & Final Reporting

INTRODUCTION

Why Study Fertility?

- **Fertility Behavior:** A Key Social Driver
  - Fertility patterns determine the long-term demographic dynamics of a population.
  - They define:
    i. How many children people have.
    ii. When they have them.
    iii. Which socio-economic groups remain childless.
- **Why It Matters:** Policy & Economics
  - Changes in fertility directly impact:
    - Labor Supply & Economic Growth
    - Dependency Ratios (e.g., ratio of workers to retirees)
    - Demand for Healthcare, Childcare, and Schooling
    - Public Finances (Pensions, Tax Incentives)
- The **GSS7402 Dataset**
  - Provides a rich basis for exploring these questions from 1974-2002.
  - Contains key variables on:
    - **Fertility Outcomes:** kids, agefirstbirth
    - **Demographics:** age, education, ethnicity, immigrant
    - **Family Background:** siblings, city16, lowincome16

PROBLEM DESCRIPTION & PROJECT OBJECTIVES

04

1. **EXPLAIN:** Which socio-demographic factors are associated with family size and childlessness.
2. **PREDICT:** Individual-level fertility outcomes to inform planning and targeted interventions.
3. **ASSESS:** Temporal and cohort trends to understand how fertility behavior has evolved.

Specific Analysis Plan

1. **Fertility and Family Analysis**
   - Investigate Factors: Use regression to model how age, education, ethnicity, and lowincome16 are associated with the number of kids.
   - Study Timing: Analyze the distribution and predictors of agefirstbirth for individuals with at least one child.
2. **Socioeconomic and Demographic Analysis**
   - Background vs. Outcomes: Compare average education and kids for individuals from lowincome16backgrounds versus those who were not.
   - Impact of Immigration: Compare family characteristics (kids, education) between immigrant and non-immigrant groups.
   - Ethnic Differences: Compare distributions of education, kids, and agefirstbirth between ethnicity groups.
3. T**ime Series Analysis**
   - Explore Trends Over Time: Use the year variable to plot changes in average kids and education from 1974 to 2002.

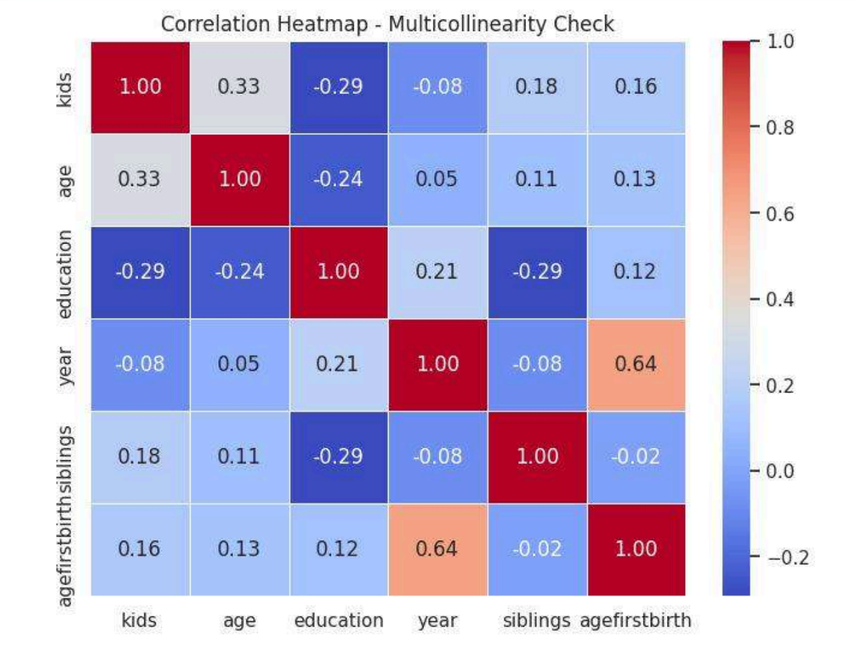
05

DATASET

The GSS7402 dataset contains approximately 9,120 cross-sectional observations collected from the General Social Survey between 1974 and 2002. The dataset includes the following variables:

- **rownames** – record identifier
- **kids** – number of children
- **age** – respondent's age (in years)
- **education** – years of formal education
- **year** – survey year
- **siblings** – number of siblings
- **agefirstbirth** – age at first birth (missing for those without children)
- **ethnicity** – self-reported ethnicity
- **city16** – lived in a city at age 16 (yes/no)
- **lowincome16** – low-income household at age 16 (yes/no)
- **immigrant** – immigrant status (yes/no)

CORRELATION, COVARIANCE & MULTICOLLINEARITY SUMMARY



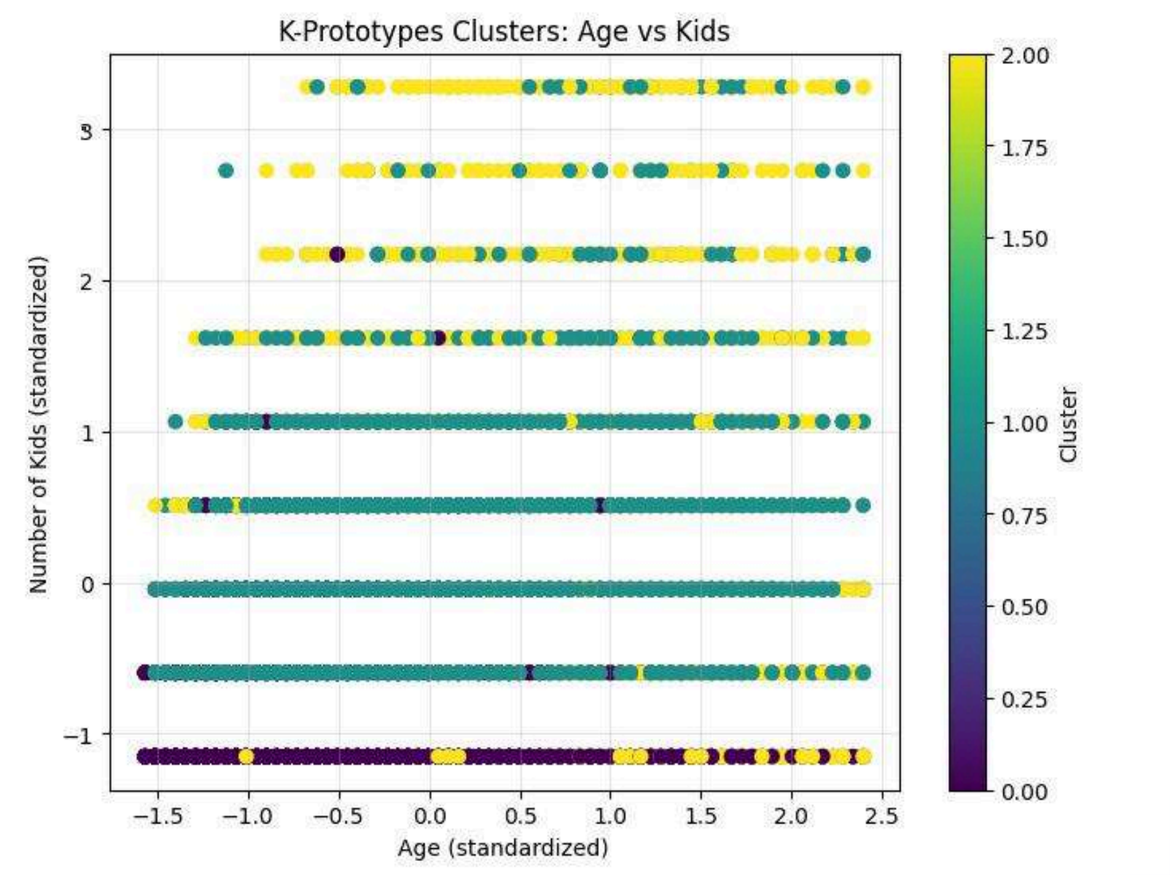
Variance Inflation Factor (VIF):

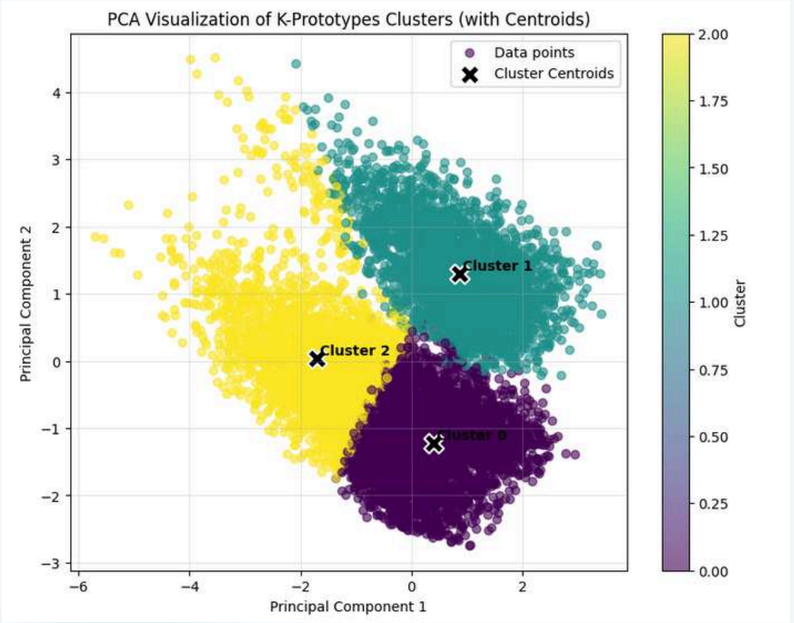
	Feature	VIF
0	kids	1.277651
1	age	1.168342
2	education	1.252934
3	year	1.816908
4	siblings	1.105030
5	agefirstbirth	1.833897

- Education shows a negative correlation with number of children → higher education → fewer kids.
- Age positively correlated with kids → older respondents have more children.
- Year & Age at First Birth strongly correlated → newer cohorts delay childbirth.
- Covariance values confirm consistent variable relationships (expected directions).
- All VIF < 2 → no multicollinearity; predictors are independent and reliable.
- **Conclusion:** Data is statistically sound for regression and predictive modeling.

INITIAL CLUSTERING AND ANALYSIS OF PRINCIPAL COMPONENTS

07





Cluster Overview

Cluster	Description
0	Young, educated, urban, early parents, fewer kids
1	Middle-aged, educated, delayed childbirth, moderate family size
2	Older, low education, early parents, larger families, lower income

LINEAR REGRESSION ANALYSIS: FACTORS INFLUENCING NUMBER OF KIDS

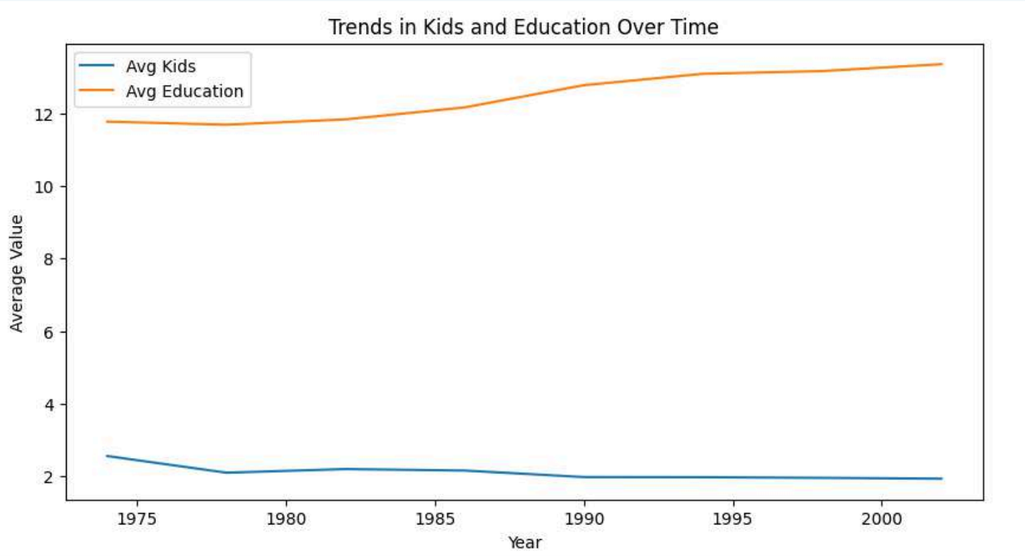
The model explains only about 13% of the variation in the number of kids, indicating a weak predictive relationship.
Higher education slightly reduces predicted kids, while age, ethnicity, and low income at 16 are positively associated.

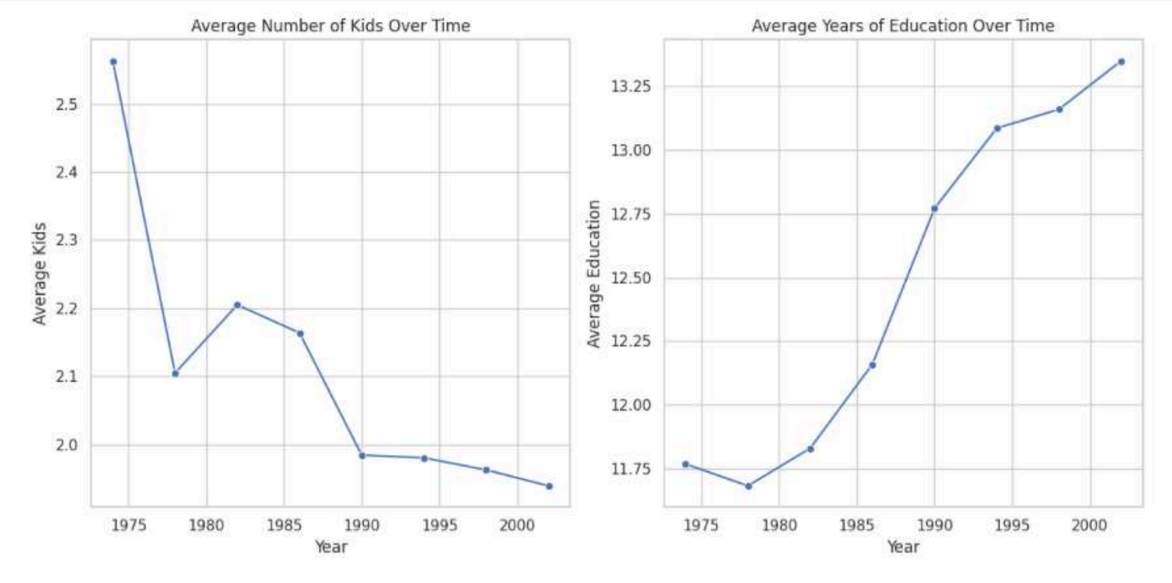
```
Coefficients:
             Coefficient
age             0.028543
education      -0.134021
ethnicity       0.378149
lowincome16     0.161754

Mean Squared Error: 2.8908235773804556
R-squared: 0.1330136499901139
```

09

INTERPRETATION

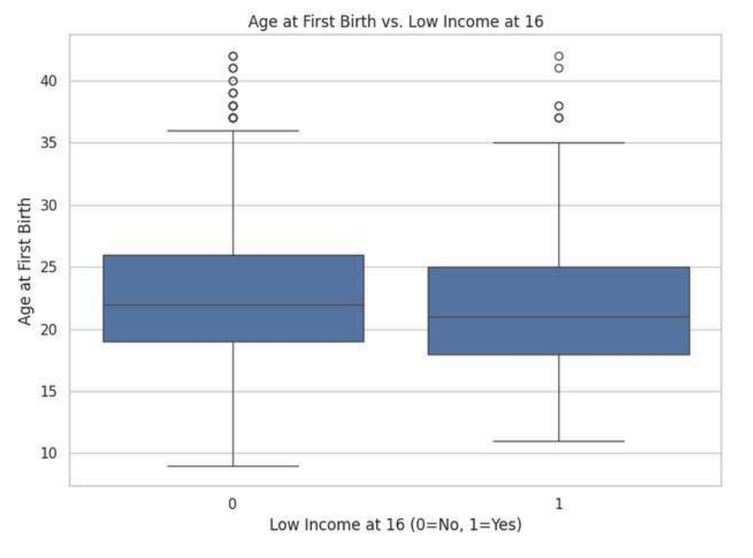


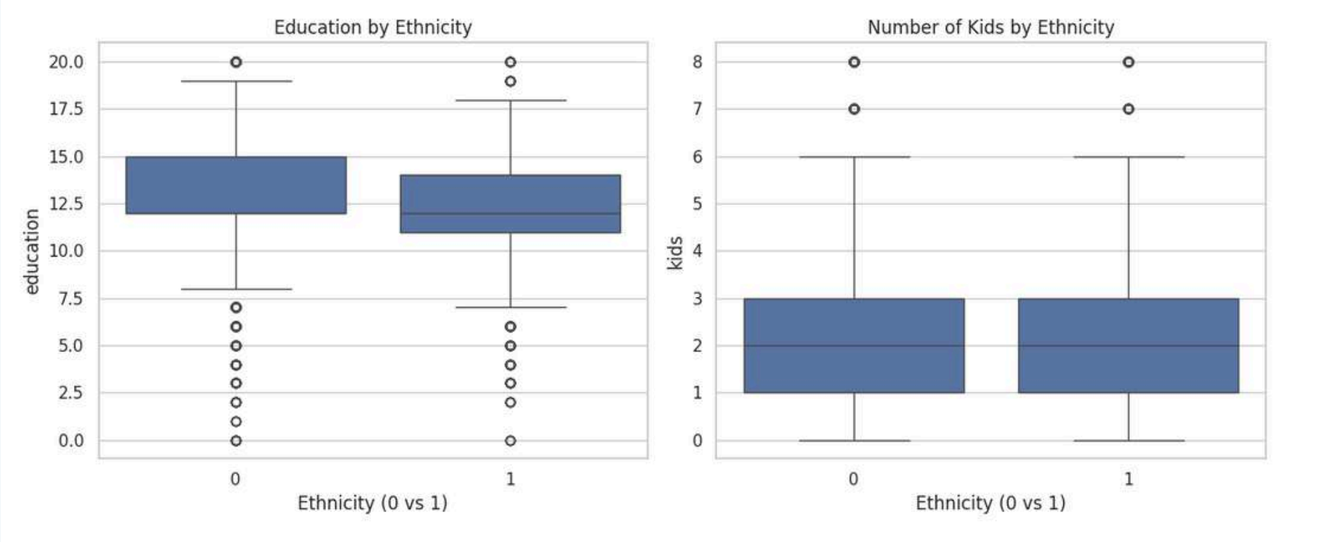


- From 1974 → 2002, the average education level increased steadily (from about 11.7 to 13.3 years).
- During the same period, the average number of kids declined (from about 2.56 to 1.94).

Over time, education levels have risen, while family sizes have decreased , suggesting that higher education correlates with smaller families in later years.

INTERPRETATION



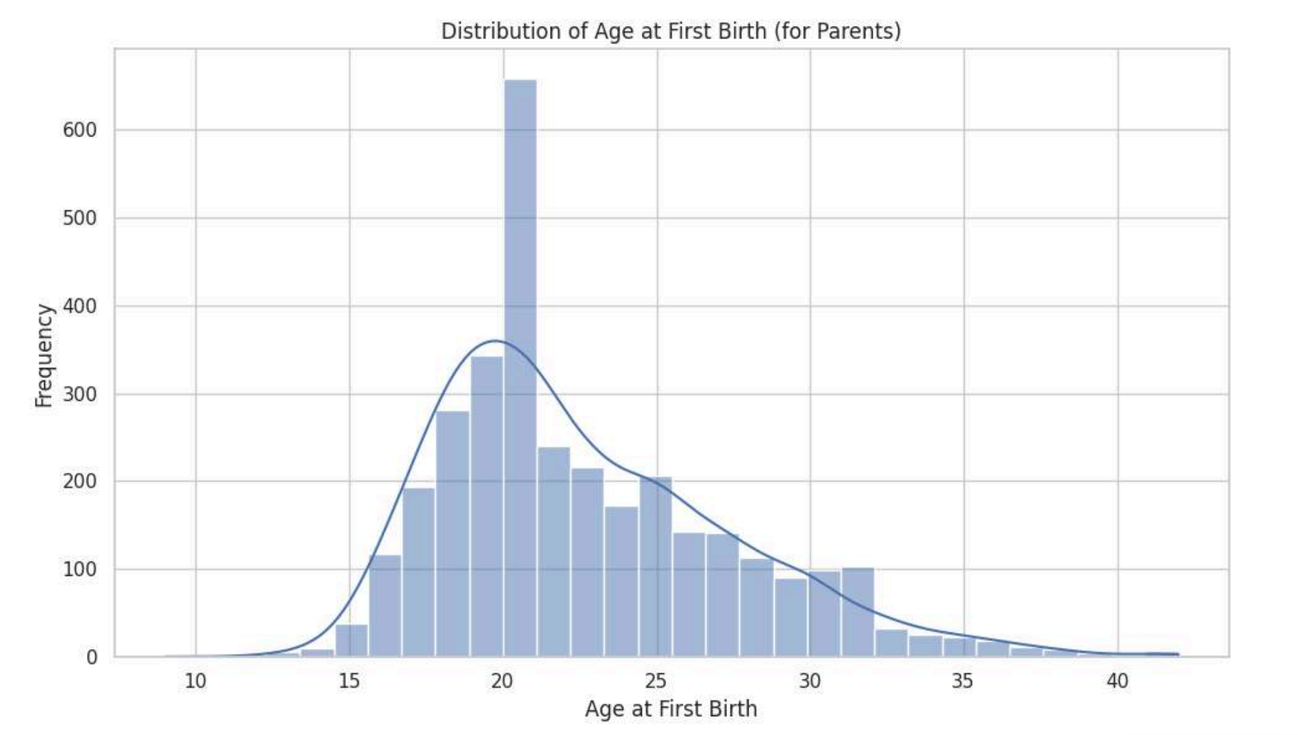


- Individuals from low-income backgrounds at 16 tend to have their first child at a younger age.
- Economic background in adolescence appears linked to earlier family formation.

- Education levels are slightly higher for Ethnicity 0, while family size is similar across both groups.
- Minor differences suggest ethnicity may modestly influence education, but not strongly affect number of kids.

INTERPRETATION

- The peak occurs between 18 and 22 years, meaning most individuals had their first child in their early 20s.
- The distribution is right-skewed, indicating fewer people had their first child at older ages (30+).
- The spread shows that while early adulthood is the most common time, some individuals have children well into their 30s and 40s.



12

SUMMARY OF ANALYSIS RESULTS

Here is a high-level summary of the findings produced by the code:

1. **Fertility and Family Analysis**
   - Regression Model: All selected variables (age, education, ethnicity, lowincome16) are statistically significant predictors of the number of kids.
     - Age: Positively associated (older individuals tend to have more kids).
     - Education: Negatively associated (more education is linked to fewer kids).
     - Ethnicity (1='other'): Positively associated (the 'other' group is linked to more kids than the 'cauc' group).
     - Low Income at 16: Positively associated (growing up low income is linked to more kids).
   - Age at First Birth:
     - The distribution is skewed to the right, with a large number of births occurring in the late teens and early twenties.
     - Those who were not low income at age 16 had their first child about half a year later, on average (mean age 22.7 vs. 22.1).
2. **Socioeconomic and Demographic Analysis**
   - Background vs. Outcomes: Individuals who grew up in a low-income household (lowincome16 = 1) have, on average:
     - Lower education (11.6 years vs. 12.9 years).
     - More children (2.4 kids vs. 2.0 kids).

SUMMARY OF ANALYSIS RESULTS

13

- Immigration: The differences between immigrant and non-immigrant groups in this dataset are modest:
  - Non-immigrants have slightly more kids (2.08 vs. 2.00).
  - Non-immigrants have slightly higher education (12.65 vs. 12.50 years).
  - Immigrants have slightly more siblings (4.5 vs. 4.0).
- Ethnic Differences: The 'other' group (ethnicity = 1), compared to the 'cauc' group (ethnicity = 0), has on average:
  - Lower education (12.1 vs. 12.8 years).
  - More children (2.3 vs. 2.0).
  - A younger age at first birth (20.8 vs. 23.2 years).

3. Time Series Analysis

- The dataset spans 8 unique years, from 1974 to 2002.
- Trend 1: The average number of kids has steadily decreased over this period, from ~2.56 in 1974 to ~1.94 in 2002.
- Trend 2: The average years of education have steadily increased, from ~11.77 in 1974 to ~13.35 in 2002.

THANK YOU