The Puzzle of Falling Birthrates in USA

Venkatesh Kannan
Email: kannan.venkatesh@stud.hs-fresenius.de
ID: 400370901

2025-02-02

Abstract

In this analysis, I work on the extreme decline in US birth rates since the Great Recession using (Kearney et al., 2022) for a guide. I intend to reproduce and extend empirical results of the original paper with an aim of deeper understanding of diversified factors driving birth rates. With R Studio being the main instrument, I dwell on the interrelationship of economic, social, and policy influences that have led to the sustained decline in fertility rates. This report is written in R Markdown, which provides a full integration of analysis and documentation; presentation developed in Quarto effectively communicates the findings. Throughout the whole work, Git and GitHub functions will be very useful in replicating and managing the workflow, structuring the handling of updates and revisions that will be necessary. This replication not only serves to validate the robustness of the initial results but deepens my insight into how economic conditions and societal shifts are framing reproductive decisions. The complexity of my findings is indicative that many factors intermingled beyond simple economic ones are at play, shaping reproductive choices in the United States.

1 Introduction

The United States experienced a dramatic decline in birth rates since 2007, not only during the Great Recession but well into the subsequent recovery and beyond.(Kearney

et al., 2022) proceed with an inquiry into this idea, taking each of various possible economic, social, and policy reasons that can influence fertility decisions. Their work has emphasized the ways in which changing economic fortunes, evolving social norms, and shifting policy climates may play into emerging patterns of fertility.

This decline in birth rates is multi-factorial and complex in nature, impacting the demographic pattern of the country along with economic and social systems. Understanding what accounts for such a change is imperative for policy analysts, healthcare managers, and social scientists, given that long-range planning demands anticipating future changes and their effects on demographic structures. As birth rates continue shaping up economic outcomes and configurations of society, the analyses of such trends become ever so important.

Data for this project was acquired with from the repository linked to (Kearney et al., 2022), available at https://www.aeaweb.org/articles?id=10.1257/jep.36.1.151 This.zip file contains all the '.dta' files, which are the data files for Stata, a statistical software tool, and all the '.do' files, which are the script files containing commands that Stata executes and '.csv' file that can be analysed in Excel file. The raw data needed for analysis, as well as all variables on birth rates, demographic factors, and socioeconomic indicators across different states and time periods, are contained in the '.dta' files. All the code necessary to replicate the analyses of the original paper can be found in the '.do' files, allowing for an accurate reconstruction of the findings of the study.

2 Objective

The aim of the project is to reproduce and extend the analysis of (Kearney et al., 2022) about the decline in birth rates since the Great Recession in the United States. This project will be specifically devoted to reproducing some of the key graphical displays of the original paper: time series of the total birth rate, state-by-state decline in birth rates, time series of birth rates by age group, and racial composition. It also tends to display the reproduction rates of different cohorts of mothers throughout their lifetime.

3 Methodology

The review of R and R Markdown at the outset of the methodology ensures robustness in analysis. Further, this study discusses the collection of data from the '.dta' and '.do' files availed from the repository of the American Economic Association. The next section elaborates on a replication strategy aimed at reconstructing the original findings

in (Kearney et al., 2022), with extensions aimed at enhancing comprehension of the dynamic nature of U.S. birth rates

3.1 R

R is basically designed for statistical computing and plotting. Ross Ihaka and Robert Gentleman from Auckland univeristy, New Zealand, initiated the R programming language in the early 1990s. Since then R has matured into a comprehensive platform that is integrated with several packages and libraries. RStudio is one such distribution based on it. Among the goals for the language designers is to make syntax easy to use and powerful especially for tasks that pertain to data manipulation and Statistical calculations. This structure also facilitates the persons who enter into the field of data analysis and replication to easily work with the system.

3.2 R Markdown

R Markdown is a generic documentation tool that allows R users to produce high-quality dynamic reports and presentations. It allows one to embed narrative text, code, and its output in one document, rendering into various formats: HTML, PDF, Word. It thus gracefully meets analysis and reporting. This tool is crucial for all researchers and analysts who want to share their findings in a comprehensive and reproducible way: it keeps data analysis and its textual description tightly coupled and easily accessible.

3.3 Data Collection

This is a replication study for which data is downloaded from the American Economic Association web portal https://www.aeaweb.org/articles?id=10.1257/jep.36.1.151. In the given data with this replication package, there is a number of '.dta' and '.do' files included. They contain a complete set of data such as state-year-level aggregate birth data, birth rates by mother's education level,birth order, and total population retrieved from multiple reliable sources. The data underlying the statistical analysis are stored in the '.dta' files, and the '.do' files store Stata commands used to process and analyze data in a thorough, structured way in order to reproduce the original study's analyses or extend them wherever possible.

Data File: The replication package contains several datasets, most relevant for the analysis of trends and determinants of birth rates in the United States. In detail, the list of the included datasets is the following: 1.data/nchs/nchs_1990_2020.dta

2.data/nchs/nchs_cohort_analysis.dta
3.data/pop/us.1990_2019.singleages.dta 4.data/sec_2a/hisp_nativity.dta 5.data/annual_policy/policy
6.data/decomp/cps_19902019.dta 7.data/decomp/state_year_age6grp_educ_raceeth_pop1544_aggte
8.data/long_term/longdiff-RHS-0408 9.data/long_term/longdiff-RHS-1519

3.4 Process Of Replication

I started the replication process by opening the working paper on the official webpage of the American Economic Association. After obtaining this information, I downloaded the working paper and the complete replication package. In the replication package, there is a Zip file named 144981-V1, containing few '.dta' files, which is a Stata file format used to conduct the analysis, and '.do'file which is the code or directions to be followed for replicating the figures in the (Kearney et al., 2022) paper. The replication package contains a readme file just to inform about the data set and provide links to download them. The analysis in'.dta' format was done through R studios.

The research provided a systematic approach toward the generation of initial replication graphs. From Figure 1, the data taken from "nchs_births_pop_1990_2019.dta". Libraries for "ggplot" in "tidyverse" were used for data transformation, first by installing using install.packages("ggplot2") and install.packages("tidyverse"). Following that, I now import their functions into R studios using library(ggplot) and library(tidyverse), as well as other functions such as "tidyr" for manipulating data. "sf" for spatial data and most importantly "haven" for reading '.dta' files.

```
# Load required packages
library(ggplot2)
library(dplyr)
library(sf)          # For spatial data
library(tidyr)       # For data manipulation (pivot_wider)
library(haven)       # For .dta files
library(tidyverse)

# Set output file path
setwd("C:/Users/Venka/Desktop/Uni subjects/Data science in business/The Puzzle of Fall
output_path <- "C:/Users/Venka/Desktop/Uni subjects/Data science in business/144981-V1
```

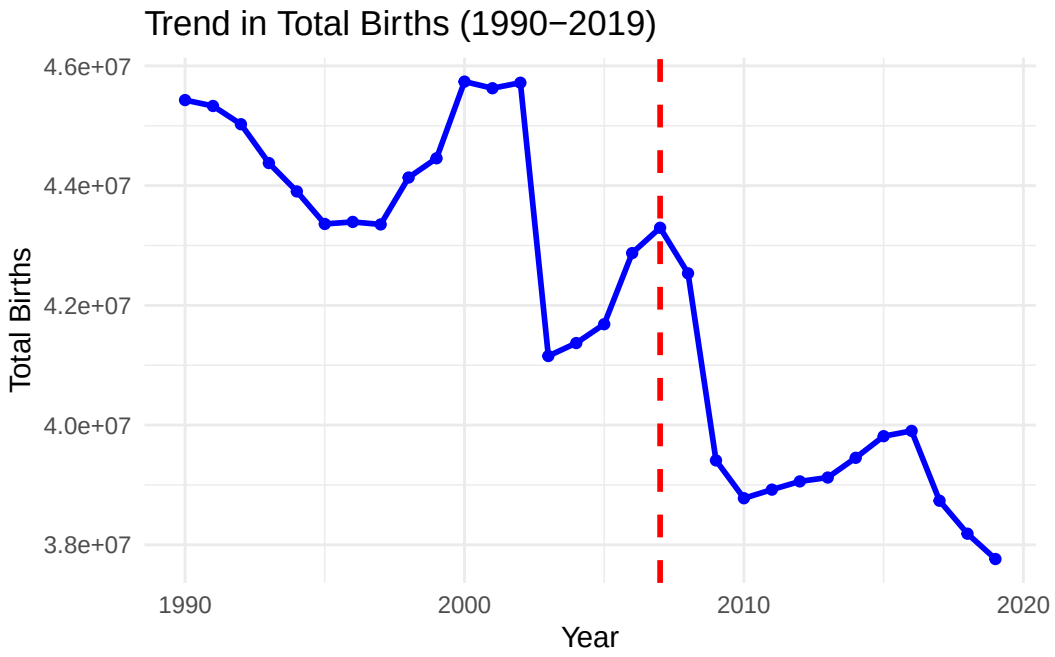
I set the working directory path and output file path based on the instruction given in the fig1.do file. I imported the '.dta' file "nchs_births_pop_1990_2019.dta" to visualise the total trend in Total Births in the USA using the following code

```r
# Load the dataset
birth_data <- read_dta("nchs_births_pop_1990_2019.dta")
```

4 Graphical Representation

4.1 Trend in Total Births

```r
# Replace missing values with zero
birth_data[is.na(birth_data)] <- 0

# Aggregate total births per year
birth_yearly <- birth_data %>%
  group_by(year) %>%
  summarise(total_births = sum(across(starts_with("numbirth_")), na.rm = TRUE))

# Plot the trend in total births with a vertical line at 2007
fig_1 <- ggplot(birth_yearly, aes(x = year, y = total_births)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue") +
  geom_vline(xintercept = 2007, linetype = "dashed", color = "red", size = 1) +
  labs(title = "Trend in Total Births (1990-2019)",
       x = "Year",
       y = "Total Births") +
  theme_minimal()

# Save the plot as a PNG file
ggsave(output_path, plot = fig_1, width = 8, height = 5, dpi = 300)

# Display the graph
print(fig_1)
```



In this R script, I first loaded a dataset from a Stata file called `nchs_births_pop_1990_2019.dta` containing detailed information on U.S. births between 1990 and 2019. Any inconsistencies in the data set were fixed by replacing the missing values with zeros for a smooth process in handling data downstream. Next, I used the `group_by` in the `dplyr` package to get the total birth per year; this was by summing the columns that start with "numbirth_" in each group and ignoring residual missing values so that the summation would not get affected.

I have used the `ggplot2` package to visualize it, creating a detailed line plot of the total number of births for each year. I charted and labeled all data points to enhance readability. To identify a notable year within the data set, I applied a red dashed vertical line at the year 2007. This is a visual marker that flags a potential inflection point or notable event driving birth trends. I then saved the resultant plot in a high-resolution PNG format to a specified directory and displayed the plot on-screen, allowing both the preservation and immediate review of the visual data analysis. This will ensure that birth trends over the period specified would be clearly and effectively presented, especially those that have captured the important moments in history.

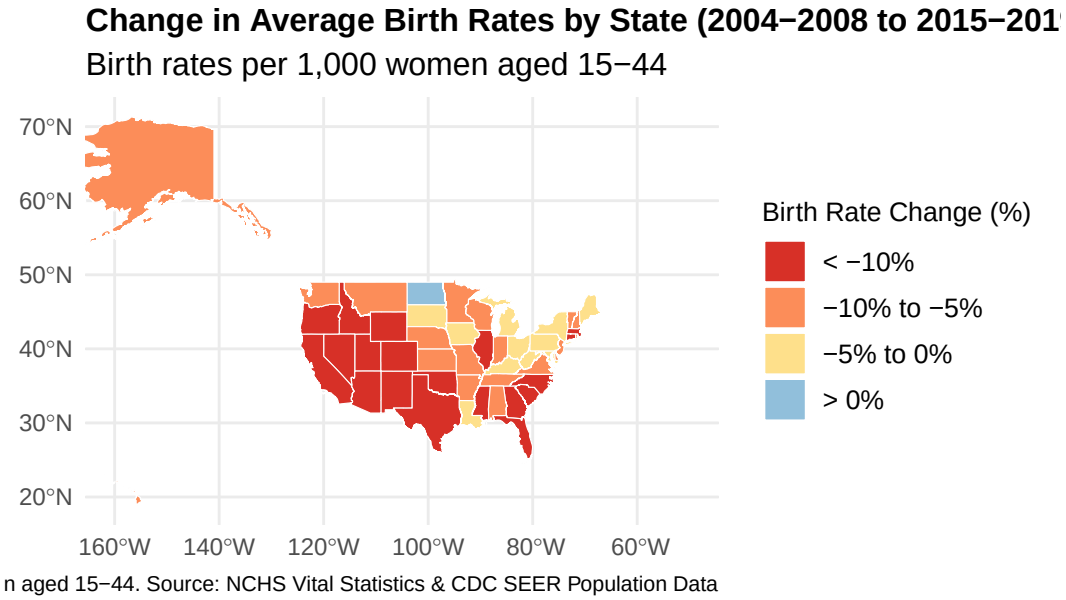
4.2 Decline in Birth Rates by State

```r
library(sf)
library(tigris)
options(tigris_use_cache = TRUE)

# Set File Paths
setwd("C:/Users/Venka/Desktop/Uni subjects/Data science in business/The Puzzle of Fall
output_path <- "C:/Users/Venka/Desktop/Uni subjects/Data science in business/144981-V1
output_csv_path <- "C:/Users/Venka/Desktop/Uni subjects/Data science in business/14498

pop_data <- read_dta("age_race_comp_seer.dta")
pop_data_processed <- pop_data %>%
  mutate(
    pop1544 = pop1519 + pop2034 + pop3544,  # Summing population age groups
    stname = toupper(trimws(stname))  # Normalize state names
  ) %>%
  filter((year >= 2004 & year <= 2008) | (year >= 2015 & year <= 2019)) %>%
  mutate(year2 = ifelse(year <= 2008, 2004, 2019)) %>%  # Collapse years
  group_by(stname, year2) %>%
  summarise(pop1544 = sum(pop1544, na.rm = TRUE), .groups = 'drop')

# Process birth data
birth_data_processed <- birth_data %>%
  mutate(stname = toupper(trimws(stname))) %>%
  filter((year >= 2004 & year <= 2008) | (year >= 2015 & year <= 2019)) %>%
  mutate(year2 = ifelse(year <= 2008, 2004, 2019)) %>%  # Collapse years
  group_by(stname, year2) %>%
  summarise(numbirth1544 = sum(numbirth1544, na.rm = TRUE), .groups = 'drop')

# Merge birth & population data
merged_data <- left_join(birth_data_processed, pop_data_processed, by = c("stname", "y

# Pivot years into separate columns for proper calculations
wide_data <- merged_data %>%
  pivot_wider(
    names_from = year2,
    values_from = c(numbirth1544, pop1544),
    names_prefix = "year_"
  )
```

```r
# Compute birth rates and changes
wide_data <- wide_data %>%
  mutate(
    brate_2004 = (numbirth1544_year_2004 / pop1544_year_2004) * 1000,
    brate_2019 = (numbirth1544_year_2019 / pop1544_year_2019) * 1000,
    brate1544_thsnds_ch = brate_2019 - brate_2004,
    brate1544_thsnds_ch_pct = 100 * (brate_2019 - brate_2004) / brate_2004
  )
# Load U.S. state shapes using `tigris::states()`
us_states <- tigris::states(cb = TRUE) %>%
  mutate(
    stname = toupper(NAME),  # Normalize state names
    state_abbr = case_when(
      NAME == "American Samoa" ~ "AS",
      NAME == "Guam" ~ "GU",
      NAME == "District of Columbia" ~ "DC",
      NAME == "Commonwealth of the Northern Mariana Islands" ~ "MP",
      NAME == "United States Virgin Islands" ~ "VI",
      NAME == "Puerto Rico" ~ "PR",
      TRUE ~ state.abb[match(NAME, state.name)]  # Use state.abb for other states
    )
  )

# Check for remaining missing abbreviations
if (any(is.na(us_states$state_abbr))) {
  print("Some state names still do not have corresponding abbreviations:")
  print(us_states %>% filter(is.na(state_abbr)) %>% select(NAME))
}
# Debug: Check unmatched rows
print("State names in wide_data:")
print(unique(wide_data$stname))

print("State names in us_states:")
print(unique(us_states$stname))

unmatched_states <- anti_join(wide_data, us_states, by = "stname")
print("Unmatched states:")
print(unmatched_states)
# Load U.S. state shapes using `tigris::states()`
us_states <- tigris::states(cb = TRUE) %>%
```

```r
  filter(!NAME %in% c(
    "American Samoa",
    "Guam",
    "Commonwealth of the Northern Mariana Islands",
    "United States Virgin Islands",
    "Puerto Rico"
  )) %>%
  mutate(
    stname = toupper(NAME),  # Normalize state names
    state_abbr = case_when(
      NAME == "Hawaii" ~ "HI",  # Include Hawaii
      TRUE ~ state.abb[match(NAME, state.name)]  # Use state.abb for other states
    )
  )

# Merge state shapes with birth rate changes
us_states <- left_join(us_states, wide_data, by = c("state_abbr" = "stname"))

# Fix missing values in percentage change
us_states$brate1544_thsnds_ch_pct <- replace_na(us_states$brate1544_thsnds_ch_pct, 0)


# Define bins and colors for percentage change
colors <- c("#d73027", "#fc8d59", "#fee08b", "#91bfdb")  # Red to blue gradient
breaks <- c(-Inf, -10, -5, 0, Inf)                       # Define bins
labels <- c("< -10%", "-10% to -5%", "-5% to 0%", "> 0%")  # Bin labels

# Plot the map with longitude range
fig_3 <- ggplot(us_states) +
  geom_sf(aes(fill = cut(brate1544_thsnds_ch_pct, breaks = breaks, labels = labels)),
  scale_fill_manual(values = setNames(colors, labels), name = "Birth Rate Change (%)")
  coord_sf(xlim = c(-160,-50)) +
  labs(
    title = "Change in Average Birth Rates by State (2004-2008 to 2015-2019)",
    subtitle = "Birth rates per 1,000 women aged 15-44",
    caption = "Notes: Birth Rates are calculated among women aged 15-44. Source: NCHS
  ) +
  theme_minimal() +
  theme(
    legend.position = "right",
```

```
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 12),
    plot.caption = element_text(size = 8),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 10)
  )

# Save the plot with larger dimensions and Hawaii included
ggsave(output_path, plot = fig_3, width = 20, height = 12, dpi = 300)

# Display the plot
print(fig_3)
```



In the R code, I used two very crucial data sets to achieve an in-depth analysis of births across the U.S. through selected periods of time. The first data set taken from the "age_race_comp_seer.dta" includes demographic information-precise by age population at various states' levels in a number of consequent years in the United States. This is crucial data in building an understanding of the population base that the birth rate will be calculated on and narrows the focus down to women between ages 15-44. Cleaning this data was done by aggregating populations by relevant age group and standardizing state names in all uppercase; this ensures accurate merges during later steps.

To complement the demographic data, the second important data used is from "nchs_births_pop_1990_2019.dta," which shows birth numbers across states and years. This birth dataset required similar preprocessing steps to perfectly align with the demographic data, focusing on the same age demographic and making sure state names were standardized across both datasets. This would ensure that the data sets align on key variables such as state names and year ranges for analysis, ensuring birth data perfectly matches demographic data for calculations.

Both of these data sets preprocessed and aligned, I merged them using a left join operation from 'state name' and a new variable 'year2'. This variable 'year2' divided the time in two groups, 2004-2008 and 2015-2019, to clearly separate the birth rates before and after that period. This was an important merge that joined the detailed birth numbers with corresponding population figures, thus enabling the exact calculation of birth rates per 1,000 women in the specified age bracket.

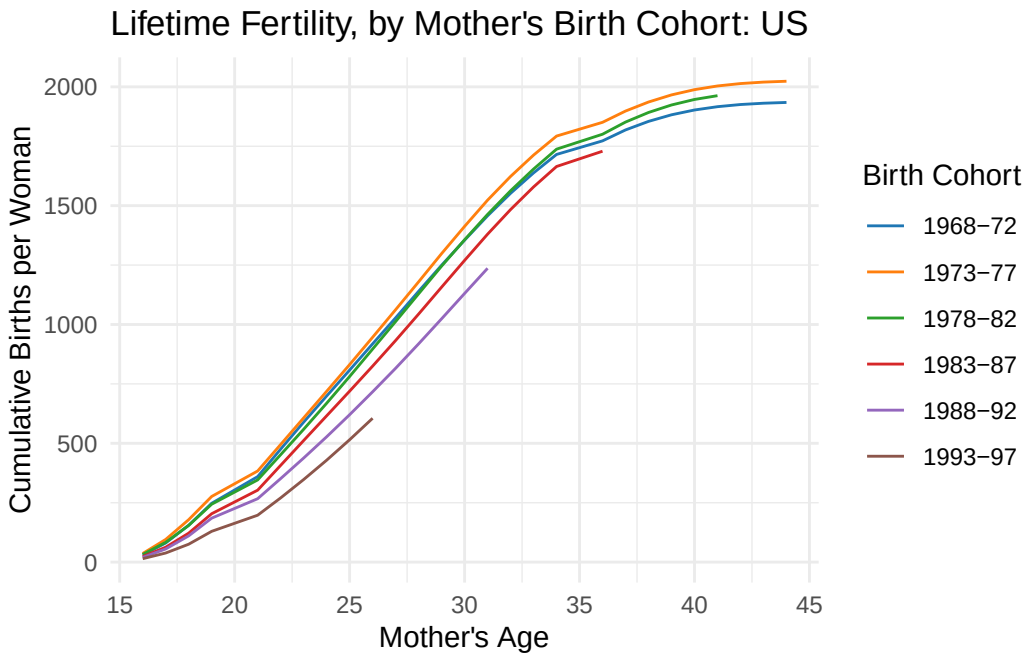
After merging, the combined dataset allowed computation of birth rates for each state and period, followed by calculation of changes in these rates across the years. These were then visualized to show graphically the change in birth rates across different states in the U.S. from one period to the next. The reason for this was to ensure that rigorous preparation, alignment, and merging of both demographic and birth datasets would provide valid and insightful analyses into trends of birth rates, highlighting regional differences and temporal changes effectively.

4.3 Lifetime Fertility by Mothers Birth Cohorts

```
nchs_data <- read_dta("nchs_cohort_analysis.dta")
seer_data <- read_dta("agecomp-seer.dta")
seer_data <- seer_data %>%
  select(-starts_with("fem15"), -starts_with("fem20"), -starts_with("fem35")) %>%
  pivot_longer(cols = starts_with("fem"), names_to = "mage", values_to = "pop", names_
  mutate(
    mage = as.numeric(mage),
    year = as.numeric(year),
    cohort = year - mage,
    cohort2 = cut(cohort, breaks = c(1967, 1972, 1977, 1982, 1987, 1992, 1997), labels
  ) %>%
  mutate(cohort2 = as.character(cohort2)) %>%
  group_by(cohort2, mage) %>%
  summarize(pop = sum(pop, na.rm = TRUE), .groups = 'drop')
```

```r
nchs_data <- nchs_data %>%
  mutate(
    mage = as.numeric(mage),
    cohort2 = as.character(cohort2)
  )
print(unique(nchs_data$mage))
print(unique(seer_data$mage))
print(unique(nchs_data$cohort2))
print(unique(seer_data$cohort2))
nchs_data_adjusted <- nchs_data %>%
  filter(mage %in% unique(seer_data$mage))
seer_data <- seer_data %>%
  filter(!is.na(cohort2))  # Exclude rows where cohort2 is NA
matched_data <- full_join(nchs_data_adjusted, seer_data, by = c("mage", "cohort2"))


print(sum(is.na(matched_data$pop)))
print(sum(is.na(matched_data$numbirth)))



matched_data <- matched_data %>%
  group_by(cohort2) %>%
  arrange(mage) %>%
  mutate(
    brate = (numbirth / pop) * 1000,
    cum_brate = cumsum(brate)
  ) %>%
  ungroup()

ggplot(matched_data, aes(x = mage, y = cum_brate, group = cohort2, color = cohort2)) +
  geom_line() +
  scale_color_manual(
    values = c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd", "#8c564b"),  # C
    labels = c("1968-72", "1973-77", "1978-82", "1983-87", "1988-92", "1993-97")  # Cu
  ) +
  labs(
    title = "Lifetime Fertility, by Mother's Birth Cohort: US",
    x = "Mother's Age",
    y = "Cumulative Births per Woman",
```

```
    color = "Birth Cohort"  # Label for the color legend
  ) +
  theme_minimal()
```




```
ggsave("lifetime_fertility_plot.png", width = 10, height = 6, dpi = 300)
```

In this R script, I will be working with two different datasets: `nchs_cohort_analysis.dta` and `agecomp-seer.dta`, which I have loaded and named `nchs_data` and `seer_data`, respectively. First, I cleaned up `seer_data` by removing columns that had specific prefixes corresponding to age groups irrelevant to my analysis. Then I transformed this dataset into a long-format dataset, where each row represented the population count of a particular age using the `pivot_longer` function. During this transformation, I also calculated each individual's birth cohort by subtracting their age from the recorded year, further grouping these into labeled cohort bins.

First, I standardized age and cohort labels in both the data sets for completeness and to effectively merge them correctly. Following the standardization of the aforementioned variables, I filtered nchs_data for only those ages which exist within seer_data, making both data sets compatible with each other and I excluded the rows where the cohort was missing in the case of `seer_data` to keep data integrity. Next, a full join would combine the cleaned `nchs_data` and `seer_data` based on age and cohort group for one data set

called `matched_data`. This will associate the numbers of births to the population across cohorts.

Then, I also checked the missing values in the key columns after merging to ascertain that further analysis would not be compromised. Now, birth rates per 1,000 women were calculated from the merged data grouped by cohorts and ordered in maternal age and corresponding cumulative birth rates. Indeed, these played an important part in the different generations' analyses of fertility trends.

Finally, I visualized these trends using a line plot via `ggplot2`, assigning unique colors and labels to each cohort group for clear differentiation. This plot presents the cumulative birth rates as a function of mother's age, divided into birth cohorts. I saved this plot to a PNG file, ensuring it was saved for presentation and further analysis.

This holistic approach allowed me to analyze and visualize longitudinal fertility trends by maternal age cohort effectively, underlining generational changes in fertility behaviors within the U.S. population.

5Various Facets of the process of Replication

5.1 Tools Used

5.1 Tool Used The study mainly depended on the lecture notes provided by Prof. Dr. Stephan Huber, which can be accessed at https://hubchev.github.io/ds/.Additional tools were used for this job were. 1. R Studio: it served as the main interface for coding in R-script and preparing the report using R-Markdown. 2. Online tools : Several internet-based tools such as Chatgpt and bard Use Google's AI for better coding and to visualize things nicely.

5.2 Issues Faced

5.2.1 R Language

Knowing R and R markdown: I needed to understand the concepts of different functions and many parameters in order to complete the replication project and produce a report.

5.2.2 Data Aquisition

Some challenges were observed during the acquiring of data. Acquiring `seer_data` dataset while trying to download this `seer_data` dataset by following instructions laid down in this project's README file at the following link - https://seer.cancer.gov/popdata/download.html. The limitation is that the download came as an executable file (.exe), which can't be imported directly into RStudio for data analysis.

Since the.exe file could not be used within RStudio, my approach needed to change. Rather than try to extract data from the executable file a process that would have required using other tools and several steps outside of R, I used already available data from the replication package. The package provided several '.dta' files that could hold segments of data different from what I would need for this analysis.

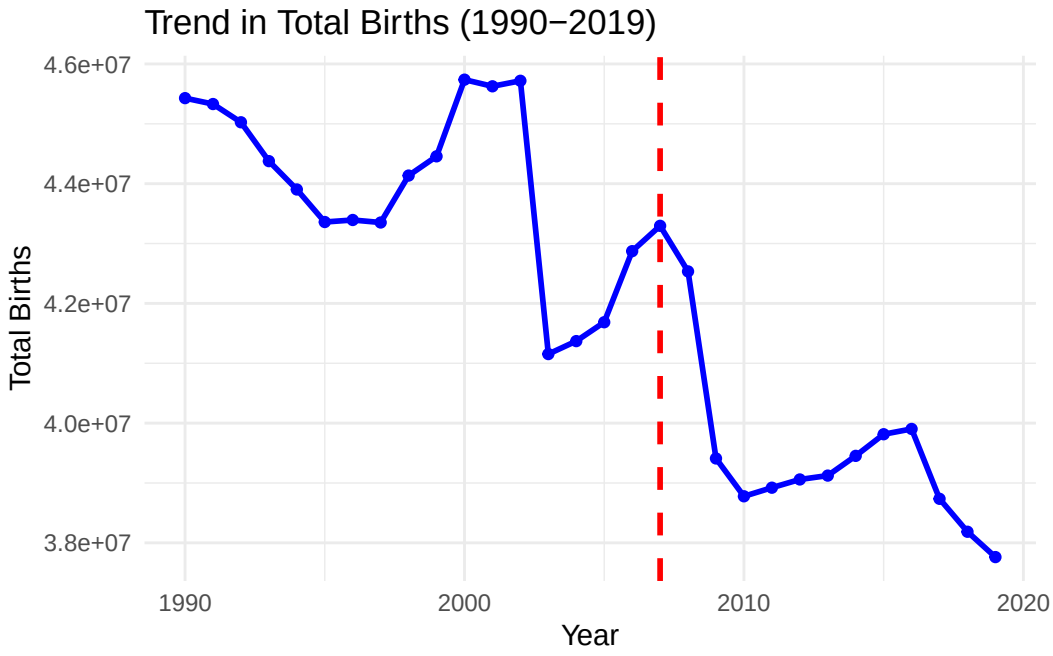
To proceed with that effectively, I imported these '.dta' files into R to take a look at what was inside. This examination included reading the columns of each file to understand the types of data they contained and to determine which dataset was necessary for each specific figure I intended to reproduce. In general, this decision was taken to avoid complexity with the '.exe' file in generating the analysis and rely on the already available '.dta' files available within the replication package. This way, I am able to continue the analysis using reliable and accessible data formats suitable for RStudio.

This approach emphasizes flexibility in data management strategies, especially in terms of unexpected technical problems during data collection. I could maintain the integrity and progress of the analysis, given the circumstance and using whatever resources were available at that time.

5.2.3 Graphical Representations

```
# Load the dataset
birth_data <- read_dta("nchs_births_pop_1990_2019.dta")
birth_data[is.na(birth_data)] <- 0

# Aggregate total births per year
birth_yearly <- birth_data %>%
  group_by(year) %>%
  summarise(total_births = sum(across(starts_with("numbirth_")), na.rm = TRUE))

# Plot the trend in total births with a vertical line at 2007
fig_1 <- ggplot(birth_yearly, aes(x = year, y = total_births)) +
```

```r
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue") +
  geom_vline(xintercept = 2007, linetype = "dashed", color = "red", size = 1) +
  labs(title = "Trend in Total Births (1990-2019)",
       x = "Year",
       y = "Total Births") +
  theme_minimal()

# Save the plot as a PNG file
ggsave(output_path, plot = fig_1, width = 8, height = 5, dpi = 300)

# Display the graph
print(fig_1)
```



One off the difficult problem I encountered in trying to reproduce the trend analysis of total births using the R script was that of the availability of data. The original graph presented in the paper ran from 1980 to 2022, thus giving an extensive view of the birth trends for over four decades. However, the replication package dataset, `nchs_births_pop_1990_2019.dta`, contains data only up to 2019. In this way, the first decade, from 1980 to 1989, and the last three years, from 2020 to 2022, were missing in the analysis.

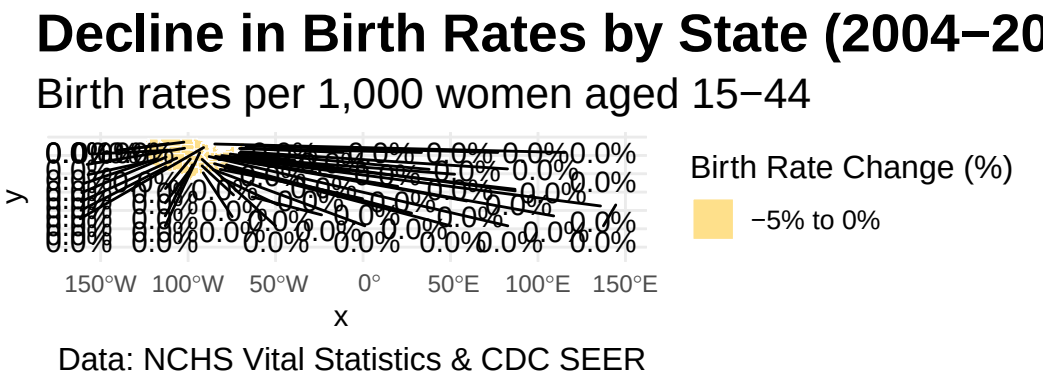
This mismatch in the period range raised some difficulties in obtaining an exact replication of the original graph. These missing years reduced the temporal scope of the analysis, apart from possibly affecting the trends leading up to 1990 and after 2019, which could have included major demographic shifts or policy impacts on birth rates. Accordingly, although I was able to plot and analyze the data available within the period 1990 to 2019, the graph that I created was not able to capture either the historical breadth or the most recent trends that the original study portrayed.

Considering this to be a limiting problem, the presentation of the trends should be strictly focused on accurately mimicking what can be depicted based on data trends from 1990 to 2019. I key events and shifts, such as a dash vertical line set at 2007 in relation to marking certain events or breaks in birth tendencies. However, the inability to incorporate data from 1980-1989 and 2020-2022 was an important limitation in fully replicating the insights and implications of the original graph. This experience also reminded me of the need to have full access to complete data sets for doing historical analysis and flexibility in adapting the techniques of analysis to what the data allow.

```
library(tidyverse)
library(haven)
library(sf)
library(tigris)
library(ggrepel)
#load the data set
pop_data <- read_dta("age_race_comp_seer.dta")
pop_data_processed <- pop_data %>%
  mutate(
    pop1544 = pop1519 + pop2034 + pop3544,  # Summing population age groups
    stname = toupper(trimws(stname))  # Normalize state names
  ) %>%
  filter((year >= 2004 & year <= 2008) | (year >= 2015 & year <= 2019)) %>%
  mutate(year2 = ifelse(year <= 2008, 2004, 2019)) %>%  # Collapse years
  group_by(stname, year2) %>%
  summarise(pop1544 = sum(pop1544, na.rm = TRUE), .groups = 'drop')

# Process birth data
if (!'numbirth1544' %in% names(birth_data)) {
  stop("The column 'numbirth1544' does not exist in the birth data.")
}

birth_data_processed <- birth_data %>%
  mutate(stname = toupper(trimws(stname))) %>%
```

```r
  filter((year >= 2004 & year <= 2008) | (year >= 2015 & year <= 2019)) %>%
  mutate(year2 = ifelse(year <= 2008, 2004, 2019)) %>%  # Collapse years
  group_by(stname, year2) %>%
  summarise(numbirth1544 = sum(numbirth1544, na.rm = TRUE), .groups = 'drop')

# Merge birth & population data
merged_data <- left_join(birth_data_processed, pop_data_processed, by = c("stname", "y

# Check for merge issues
if (nrow(merged_data) == 0 || any(is.na(merged_data$pop1544))) {
  stop("Merge failed or 'pop1544' data is missing.")
}



merged_data <- merged_data %>%
  mutate(
    brate1544_thsnds = (numbirth1544 / pop1544) * 1000  # Birth rate per 1,000 women a
  ) %>%
  pivot_wider(
    names_from = year2, values_from = brate1544_thsnds,
    names_prefix = "brate_",
    values_fill = list(brate1544_thsnds = NA)  # Fill missing values as NA
  ) %>%
  mutate(
    brate1544_thsnds_ch = brate_2019 - brate_2004,  # Calculate birth rate change
    brate1544_thsnds_ch_pct = 100 * (brate_2019 - brate_2004) / brate_2004  # Percenta
  )

# Load and prepare geographic data
states <- st_as_sf(tigris::states(cb = TRUE), crs = 4326) %>%
  filter(!NAME %in% c("Alaska", "Hawaii", "Puerto Rico")) %>%
  mutate(stname = NAME)

# Merge geographic data with birth rate changes
geo_data <- left_join(states, merged_data, by = "stname")

# Fix missing rate_change values before plotting
geo_data$brate1544_thsnds_ch <- replace_na(geo_data$brate1544_thsnds_ch, 0)  # Avoid N

#Define Bins and Colors for Population Decline**
```

```r
colors <- c("#d73027", "#fc8d59", "#fee08b", "#91bfdb")  # Red to blue color scheme
breaks <- c(-Inf, -10, -5, 0, Inf)                      # Define bins for change
labels <- c("< -10%", "-10% to -5%", "-5% to 0%", "> 0%")  # Labels for bins


fig_3 <- ggplot(geo_data) +
  geom_sf(aes(fill = cut(brate1544_thsnds_ch, breaks = breaks, labels = labels)), colo
  geom_text_repel(
    aes(label = sprintf("%.1f%%", brate1544_thsnds_ch), geometry = geometry),
    stat = "sf_coordinates",
    size = 4, color = "black",
    max.overlaps = 100
  ) +
  scale_fill_manual(values = setNames(colors, labels), name = "Birth Rate Change (%)")
  labs(
    title = "Decline in Birth Rates by State (2004-2008 to 2015-2019)",
    subtitle = "Birth rates per 1,000 women aged 15-44",
    caption = "Data: NCHS Vital Statistics & CDC SEER"
  ) +
  theme_minimal() +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 20, face = "bold"),
    plot.subtitle = element_text(size = 16),
    plot.caption = element_text(size = 12),
    legend.title = element_text(size = 12),
    legend.text = element_text(size = 10)
  )

#Save High-Resolution Map
ggsave(output_path, plot = fig_3, width = 14, height = 8, dpi = 300)

# Display the plot
print(fig_3)
```



In this R script, while integrating and analyzing state-level birth and population data to examine trends in birth rates across different U.S. states, I encountered a significant issue related to missing values during the data merging process. This challenge arose from the merging of processed birth data and population data into a single dataset.

After processing the birth and population data separately, making sure each of the datasets included total births and population of women aged 15-44 for specific year groups (2004-2008 and 2015-2019), I merged these two sets. The merge was to combine births by numbers with the respective population by state and year group for easy computation of birth rates per 1,000 women. In reality, this was not the case in the actual merging.

The main complication arose with the discrepancies in the data, such as different state names or missing entries in one of the datasets, leading to a lot of NA values in the merged dataset. Specifically, the 'pop1544' column, which was supposed to contain the total female population aged 15-44, had missing values for some states or years. This was critical because the birth rate calculations required both the number of births and the population data to be present and correctly matched.

These missing values mean that when I compute changes in the birth rate across the two time periods, states with a missing population data get a default 0% change in birth rate. This heavily biased the analysis toward showing, in general, no change in birth rates across all states, which was not at all representative of the underlying data.

5.3 Future Scope

Moving forward, I would put in a lot of effort to handle the executable file with the refined and perhaps more detailed population data needed for an in-depth analysis. Firstly, I will make use of a software package or scripting to extract the dataset embedded in this executable file. This would be a very important process in that, through getting hold of the latest and most complete data, a trend analysis on birth rates could be done more comprehensively.

After extracting and verifying the data from the.exe file, I will integrate this new information into our existing datasets. Such integration will most likely involve the revision of our data processing routines to accommodate any new data formats or additional variables that come with the updated dataset.

With the newly integrated data, I will focus extensively on enhancing the data visualization components of the project. This will be based on the generation of more intelligent graphical representations for complex trends in an attempt at deeper insights towards factors affecting births across different demographic and regional trends. I might consider a range of visuals that are more interactive; given that this will let one interact more dynamically with visualizations through online interfaces or dashboard interactions. These will include, among other analysis tools, the ability for stakeholders to consider different scenarios and pull out tailored insights relevant to specific interests or areas of policy interest.

Besides, I intend to leverage some more advanced techniques, including predictive modeling and machine learning that may use historic data coupled with future socio-economic projected changes to estimate the trends of birth rates into the future. Predictive insight will thus be invaluable in planning by policymakers and healthcare providers for future needs.

Ultimately, this extension of the scope of the project will help in providing richer analysis with more comprehensive data, but also, more importantly, enhance access and usability through superior data visualization techniques. This would ensure that the findings of the project can actually inform decisions and policies that seek to address the dynamics of population growth and demographic changes.

6 Conclusion

The project, through the integration of highly detailed datasets and hi-tech methodologies, places current understanding of birth rate trends across the United States at a whole new level. While there were initial issues in acquiring data, for the most part-key

visualizations, which were imperative in developing critical insight into the dynamics of U.S. birth rates, were able to be created through the use of alternative data from the replication package.

Figure 1, "Trend in Total Births (1990-2019)," showed the long-term trends and fluctuations in national birth rates over almost three decades, despite missing data for earlier and more recent years. This visualization emphasized some of the major temporal shifts, such as the dramatic decline around the time of the 2007 economic downturn, in a clear historical context of how economic factors can influence birth rates.

Figure 3, "Decline in Birth Rates by State (2004-2008 to 2015-2019)," further broke down the regional variation in birth rates for different states in two separate periods. This was very helpful in locating state-to-state differences and trends, hence showing how local policy, economic conditions, and demographic influences produced different impacts on the trend in birth rates across major parts of the country.

Figure 5, "Lifetime Fertility, by Mother's Birth Cohort: US," puts this in a more longitudinal perspective to examine the change over time in fertility rates across different generations of women. Tracing cumulative births per woman in successive birth cohorts, generational changes in reproductive behavior became vividly apparent in this graph, and it illustrated how societal changes and policy reversals across several decades have played into family-size decisions.

Put together, they form a whole picture of influences that shape the birth rates of the United States. The different economic, regional, and generational shifts to reproductive choices serve to provide ample insights for thinkers into policy issues, health professionals, and social planners. Instructive not only in illustrating past and current trends but, indeed, predict future demographic shifts.

In the future, the project will be further improved by updating with more recent and complete data, enhancing the level of sophistication in data visualization, and expanding the predictive analytics. These steps will ensure that the project continues to provide valuable insights into birth rate trends, supporting informed decision-making that can adapt to changing societal needs. By deepening the analysis and broadening the scope, this project has the potential to contribute even more significantly to our understanding of demographic dynamics and inform effective public health strategies and policies.

7 Presentation using Github

GitHub is a good website for those who want to become programmers since it fosters collaboration and knowledge sharing. For further information on the exercise, please

check out Professor Dr. Stephan Huber's GitHub page: hubchev/ds_summer23

To control and display the project at once we have created a repository called "The_Puzzle_of_falling_birthrates_in_USA" in which you can upload standalone HTML files and other documents of interest. After making changes into the GitHub repo, after reviewing all updates, I used my official username in GitHub-https://github.com/VenkateshKannan1999/The_Puzzle_of_falling_birthrates_in_USA.git,the website is stored as an HTML format. The final step is pushing all relevant project files to this repository so that it is easily accessible.

8 Word count

Word Count The wordcount was made in Rmd file uploaded to my Github using Install the 'wordcountaddin' package from GitHub devtools::install_github("benmarwick/wordcountaddin", type = "source") Load the 'wordcountaddin' library library(wordcountaddin)

The result of the code above showed the Word Count of this paper to 4403

8 Affidavit

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I have read the Handbook of Academic Writing by Hildebrandt & Nelke (2019) and have endeavored to comply with the guidelines and standards set forth therein.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

The report includes:

- ☒ About 4000 words (+/- 500).
- ☒ A title page with personal details (name, email, matriculation number).
- ☒ An abstract.

☒ A bibliography, created using BibTeX with APA citation style.

☒ The complete R code required to reproduce the results.

☒ Detailed instructions on data acquisition and importation into

R.

☒ An introduction to guide the reader and a conclusion summarizing the work and discussing potential future extensions.

☒ All significant resources used in the report and R code development.

☒ The filled out Affidavit.

☐ A concise description of the successful use of Git and GitHub, as detailed here: -- make_a_pull_request.

☒ A concise description of the presentation published on GitHub.

☒ The project submission includes:

☒ … The .qmd file(s) of the report.

☐ … The quarto.yml file of the report.

☒ … The .pdf file of the report.

☒ … The standalone .html file of the report.

☒ … All necessary files (not available online) to reproduce the report and the R code.

☒ … The standalone .html file of the presentation.

$$Venkatesh Kannan$$

$$02, 02, 2025$$

$$Cologne, Germany$$

References

Kearney, M. S., Levine, P. B., & Pardue, L. (2022). The puzzle of falling US birth rates since the great recession. *Journal of Economic Perspectives*, *36*(1), 151–176. https://doi.org/10.1257/jep.36.1.151