Homework Assignment 1

Greg Forkutza
Student ID: 400277514

28 September, 2023

Contents

Report

Materials and Sources can be viewed at my GitHub repository

Question(i)

The third goal of the UN sustainable development project is to ensure healthy lives and to promote well-being for all at all ages. This includes ensuring that women are able to have healthy pregnancies and raise their children. Being pregnant at an age for which they are better able to support themselves. On the other hand, if the number of older pregnancies increases that may say something about a down turned economic situation in Canada. This could mean that the cost of living is too high and therefore women have to wait much longer before they are financially comfortable to have children. Lastly as the population of Canada increases, we should expect the birth rate to increase. This may not be the case if the above is true. This might indicate also a downturn in the health and well being of women in Canada.

This data set is called Live births, by age of mother It is sourced from the Statistics Canada, Canadian Vital Statistics, Birth Database (CVSB) This is an administrative survey that collects demographic information annually for any Canadian resident anywhere in the world. The birth is assigned to the mothers province of resident.

The data covers the years 1991 to 2021. There is one categorical variable. This category is the age of the mother. There are two quantitative variables, the percentage (relative to that province and time period) and the number of live births in that age category at the time period.

By exploring size and distribution of the age of live births across Canada we can better understand how the health and well being of women and motherhood is changing over time.

Question (ii)

The category `GEO` records the province of residence of the mother. If the province of residence is unknown it is given the level of `Unknown province or territory`. However there are no non zero values over the entire time period so we can safely drop this level Wickham [2019].

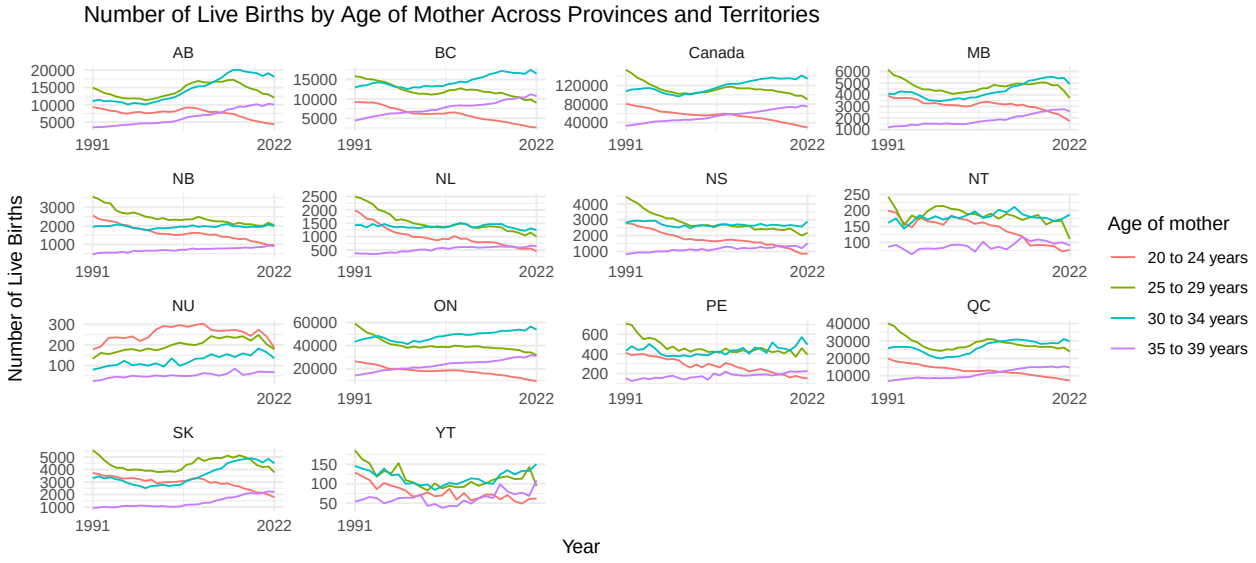
For each province of the mother's residence and each time period there are two numerical values; the number of live births and the percentage of live births within that time period over all levels of `Age of mother`. Since we want to better understand not only the distribution of the age of the mother but the change in number of births over time, we will drop the `Percentage of live births` value and keep only the `Number of live births` value. The time series of the number of live births plotted simultaneously across all levels will convey both the change in distribution and magnitude over time Wickham [2016].

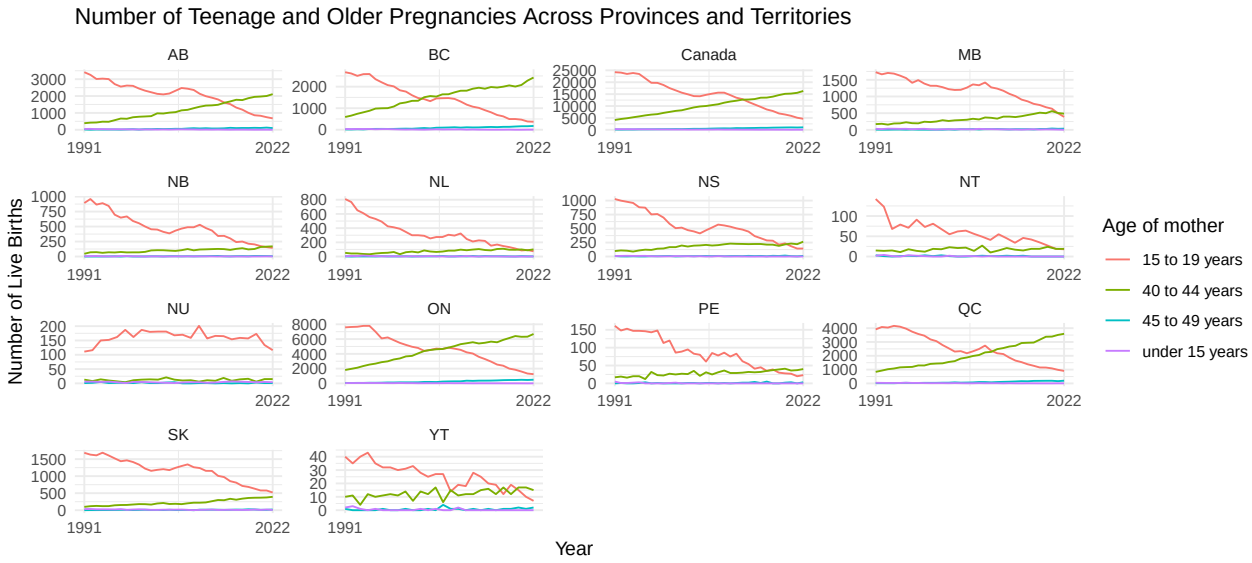
Fortunately, because records of birth are non-voluntary, the data set does not have any missing values. This is count data from birth records over levels of a single category.

There is probably not any outliers for the same reason as above. If there was some other sociological or economic information available to look at why unexpected changes in the gradient of the birth rate for that level occurred for a particular time period, then we might have reason to smooth these. However, there is no reason with the information given to indicate the presence of any outliers as the survey nature of the data is very high quality and reliable.

Question (iii)

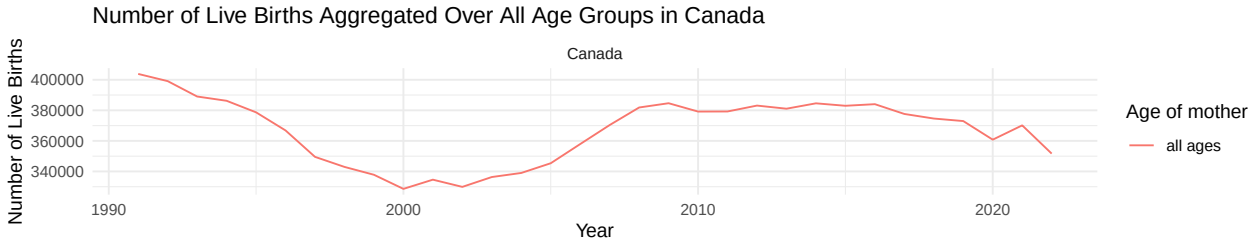
We can visualize across each province, the number of births in each age group. However the number of teen pregnancies($age \leq 19$) is small relative to the other age groups. This is the same with older pregnancies ($40 \leq age \leq 49$). Therefore we plot these four age categories separately in order to be able to better visualize their trends. Using a log transformation was considered but it did not provide clarity of the trends across all age categories as it caused the slope of the curves to decrease towards horizontal and lose visual information.





We see in the majority of provinces, the mean level of `Age of mother` goes from 25 to 29 years old to 30 to 34 years old. We observe a decrease in the number of births for women less than 30 years old across all provinces over the 31 year time period. Interestingly we see large positive gradients in the curves for women 35 to 39 years old and 40 to 44 years old in the provinces with the largest economies: Ontario, BC, AB, QC. We also notice that teen pregnancy decreases dramatically everywhere except in Nunavut. In conclusion we see that women on average are having babies at an older age across all provinces. This can be interpreted both positively and negatively: older mothers can mean both that women are not pressured into motherhood early and that they lack the sufficient resources for motherhood early. We need to know the distribution of the age of mother as a function of economic and ethnic characteristics. Then, we can say whether the health and well being of mothers and motherhood is increasing or decreasing and for whom specifically.

Question (iv)



We see that the number of live births, aggregating over all levels of age, provinces and territories is decreasing. This period of a sharp negative gradient begins in 2016. This is when many would remember the cost of living in Canada to begun to have increased dramatically. So even as the population increases in Canada we have a decrease in the birth rate. Again, demographic data of the mothers and households would make this exploration more interesting. This suggests a decrease in the health and well being of the average women in

Canada.

Question (v)

The link to the shiny app is HERE. This app allows you to view the time series of each level of the age of mother for a particular province. The interpretation of these plots is the same as in sections (iii) and (iv).

Supplementary Material

```r
#  Download the zip file from the URL and convert to CSV
url <- "https://www150.statcan.gc.ca/n1/tbl/csv/13100416-eng.zip"
destfile <- "13100416-eng.zip"
download.file(url, destfile)

unzip(destfile)

# Change province columns to be more compact
data3 <- read_csv("13100416.csv")  %>%
  mutate(GEO = str_extract(GEO, "^[^,]+")) %>%
  mutate(GEO = case_when(
    GEO == "Newfoundland and Labrador" ~ "NL",
    GEO == "Prince Edward Island" ~ "PE",
    GEO == "Nova Scotia" ~ "NS",
    GEO == "New Brunswick" ~ "NB",
    GEO == "Quebec" ~ "QC",
    GEO == "Ontario" ~ "ON",
    GEO == "Manitoba" ~ "MB",
    GEO == "Saskatchewan" ~ "SK",
    GEO == "Alberta" ~ "AB",
    GEO == "British Columbia" ~ "BC",
    GEO == "Yukon" ~ "YT",
    GEO == "Northwest Territories" ~ "NT",
    GEO == "Nunavut" ~ "NU",
    TRUE ~ GEO
  )) %>%
  mutate(`Age of mother` = sub("Age of mother, ", "", `Age of mother`))

# remove "Age of mother, all ages" and  "Canada"
filtered_data2 <- data3 %>%
  filter(grepl("Number of live births", `Characteristics`))  %>%
  filter(GEO != "Unknown province or territory") %>%
  filter(GEO != "Northwest Territories including Nunavut") %>%
  filter(!(`Age of mother` %in% c("15 to 19 years",
```

```
                                    "all ages",
                                    "under 15 years",
                                    "not stated",
                                    "40 to 44 years",
                                    "45 to 49 years"
  ))) %>%
  group_by(`Age of mother`)

# Plot
ggplot(filtered_data2, aes(x = REF_DATE, y = VALUE, group = `Age of mother`,
                           color = `Age of mother`)) +
  geom_line() +
  facet_wrap(~GEO, scales = "free", ncol = 4) +
  labs(title = "Number of Live Births by Age of Mother Across Provinces
               and Territories",
       x = "Year",
       y = "Number of Live Births") +
  theme_minimal() +
  scale_x_continuous(breaks = c(min(filtered_data2$REF_DATE),
                                max(filtered_data2$REF_DATE)))

# Remove teen preganancies and older pregnancies
filtered_data3 <- data3 %>%
  filter(grepl("Number of live births", `Characteristics`))  %>%
  filter(GEO != "Unknown province or territory") %>%
  filter(GEO != "Northwest Territories including Nunavut") %>%
  filter((`Age of mother` %in% c("under 15 years",
                                 "15 to 19 years",
                                 "40 to 44 years",
                                 "45 to 49 years"
  ))) %>%
  group_by(`Age of mother`)

# Plot
ggplot(filtered_data3, aes(x = REF_DATE, y = VALUE, group = `Age of mother`,
                           color = `Age of mother`)) +
```

```
  geom_line() +
  facet_wrap(~GEO, scales = "free", ncol = 4) +
  labs(title = "Number of Teenage and Older Pregnancies
                Across Provinces and Territories",
       x = "Year",
       y = "Number of Live Births") +
  theme_minimal()+
  scale_x_continuous(breaks = c(min(filtered_data3$REF_DATE),
                                max(filtered_data3$REF_DATE)))


# Include only Canada
filtered_data4 <- data3 %>%
  filter(grepl("Number of live births", `Characteristics`))  %>%
  filter(GEO == "Canada") %>%
  filter(`Age of mother` %in% "all ages") %>%
  group_by(`Age of mother`)


# Plot
ggplot(filtered_data4, aes(x = REF_DATE, y = VALUE, group =
                             `Age of mother`, color = `Age of mother`)) +
  geom_line() +
  facet_wrap(~GEO, scales = "free", ncol = 4) +
  labs(title = "Number of Live Births Aggregated Over All Age
                Groups in Canada ",
       x = "Year",
       y = "Number of Live Births") +
  theme_minimal()

#Shiny App code

data4 <- data3 %>%
  arrange(REF_DATE)

data4$REF_DATE <- as.numeric(as.character(data3$REF_DATE))
```

```r
# UI
ui <- fluidPage(
  selectInput("province", "Choose a Province:",
              choices = unique(data4$GEO)),
  selectInput("age", "Choose an Age Group:",
              choices = unique(data4$`Age of mother`)),
  plotOutput("plot")
)

# Server logic
server <- function(input, output) {

  output$plot <- renderPlot({
    filtered_data <- data4 %>%
      filter(GEO == input$province & `Age of mother` == input$age) %>%
      filter(grepl("Number of live births", `Characteristics`))
    ggplot(filtered_data, aes(x = REF_DATE, y = VALUE,
                              group = `Age of mother`,
                              color = `Age of mother`)) +
      geom_line() +
      labs(title = "Number of Live Births by Age of Mother",
           x = "Year",
           y = "Number of Live Births") +
      theme_minimal()
  })
}

# Run the app
shinyApp(ui = ui, server = server)
```

References

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham. *Advanced R, Second Edition (2nd ed)*. Chapman and Hall/CRC, 2019. URL https://adv-r.hadley.nz/.