GSS

Arshnoor Gill Samantha Wong Zhendong Zhang

2020/10/13

Abstract

We wanted to investigate the social causes that could affect one's outlook on life, especially during such a tumultuous time. In this paper we investigated general feelings about life, on a scale from 1 to 10, that Canadians harboured in 2017, particularly as it relates to key variables of interest which socially and culturally are considered to impact people's outlooks—income, the number of children one has, and age. We found there is a definite positive relationship between income, age and number of children on one's general outlook on life (though there were more peaks in the age variable, suggesting individual stages in life could affect outlook) through graphical representations of scatterplots and the construction of a linear regression model that took all three variables into consideration with the explanatory variable being the general outlook on life. That being said, much of the data is created from manipulations on discrete information so it is not as accurate as raw continuous data would be, and it is possible that feelings of life could impact conversely how many children one has, for instance—it is ambiguous what is the explanatory and response variable.

Introduction

Data

The dataset was obtained from the Canadian General Social Survey (GSS) conducted on an annual basis by Statistics Canada for the purpose of monitoring changes in living conditions and well-being of individuals across Canada as well as providing information on social policy issues of interest during the year the study was conducted. Data from the 2017 iteration of the GSS (Cycle 31) was analyzed. The content of the survey focuses on different aspects over time - the 2017 GSS featured notable changes involving redesigning childcare services, childcare arrangements, child custody and financial support and programs after a separation or a divorce, expanding modules on parents and grandparents and removing module on work history. The targeted population in the survey was all non-institutionalized persons older than the age of 15 living in any of the 10 provinces of Canada. The survey design employed was stratified cross-sectional sampling. Stratification was conducted at the province/census metropolitan area level and the frames were constructed from telephone numbers. In order to increase response-rates on income, personal income questions were not included in the 2017 GSS. Instead, income information was obtained through tax data for those who chose to respond. Based on given personal tax records, household information, social insurance number and other key variables, linkages were formed to obtain income information. The variables of interest in this study:

- Age
- Income
- Number of children

Modelling

In terms of modelling, we've decided to use linear regression models in order to attempt to derive a relation between the general feelings of life the respondents of the survey feel, from 1-10, with the other data we've

picked as possible explanatory variables: age, income and number of children.

The reasonings are as follows: in terms of age, there is often more uncertainty and fear regarding the future that comes when you are relatively younger, and less so when you are older and more established professionally and interpersonally. We wanted to see if people who are older tend to have a better general feeling of life as described by the survey. In terms of income, it follows that the more financially stable you feel, the greater you might feel towards life in general, and in terms of children, we wanted to know if more kids made life in some ways more fulfilling (or the opposite).

For our preliminary findings, we've decided to examine the distribution of each of the four variables both graphically and also numerically. This is through making histograms of all the variables, which will let us know the general trends the respondents feel towards their general feeling of life, their age, their number of children, and their financial status.

We chose to use a linear regression model given that the feelings of life variable, despite being discrete (in that you cannot pick a value between the integers), is still ordered, and that this way any relation between the two variables could be mapped out graphically and in terms of numerics. We used an alpha of 0.05 to determine if a result is statistically significant—thus, the p-value will test if the null hypothesis in each of these scenarios (that there is in fact no relation), which if it's under 0.05 will be verified.

That being said, in order to conduct this analysis, we had to make several decisions while modelling. First and foremost, we had to take out responses due to the issue of non-response, which decreases our overall sample size. Moreover, though age and number of children can be reported as is, the income was a categorical variable that needed to be redefined in order to fit the context of our analysis. For income, in categories "Less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000" to "99,999" and $100,000 to $124,000", we replaced each of those results with the midpoint between the ranges of each category. For instance, "$25,000 to $49,999" was replaced by 37,499.50. The final category, "$125,000 and more", was replaced simply by 125,000.

This choice of course means that the graph is not valid for income greater than $125,000, but we decided that based on a certain assumption that after a certain level of money perhaps income does not increase quality of life, we decided it was the best way to still represent those respondents in that income bracket without making further assumptions about how much money they have.

Results

Exploratory Data

```
gss <- read_csv("./Output/gss.csv")
head(gss)
```

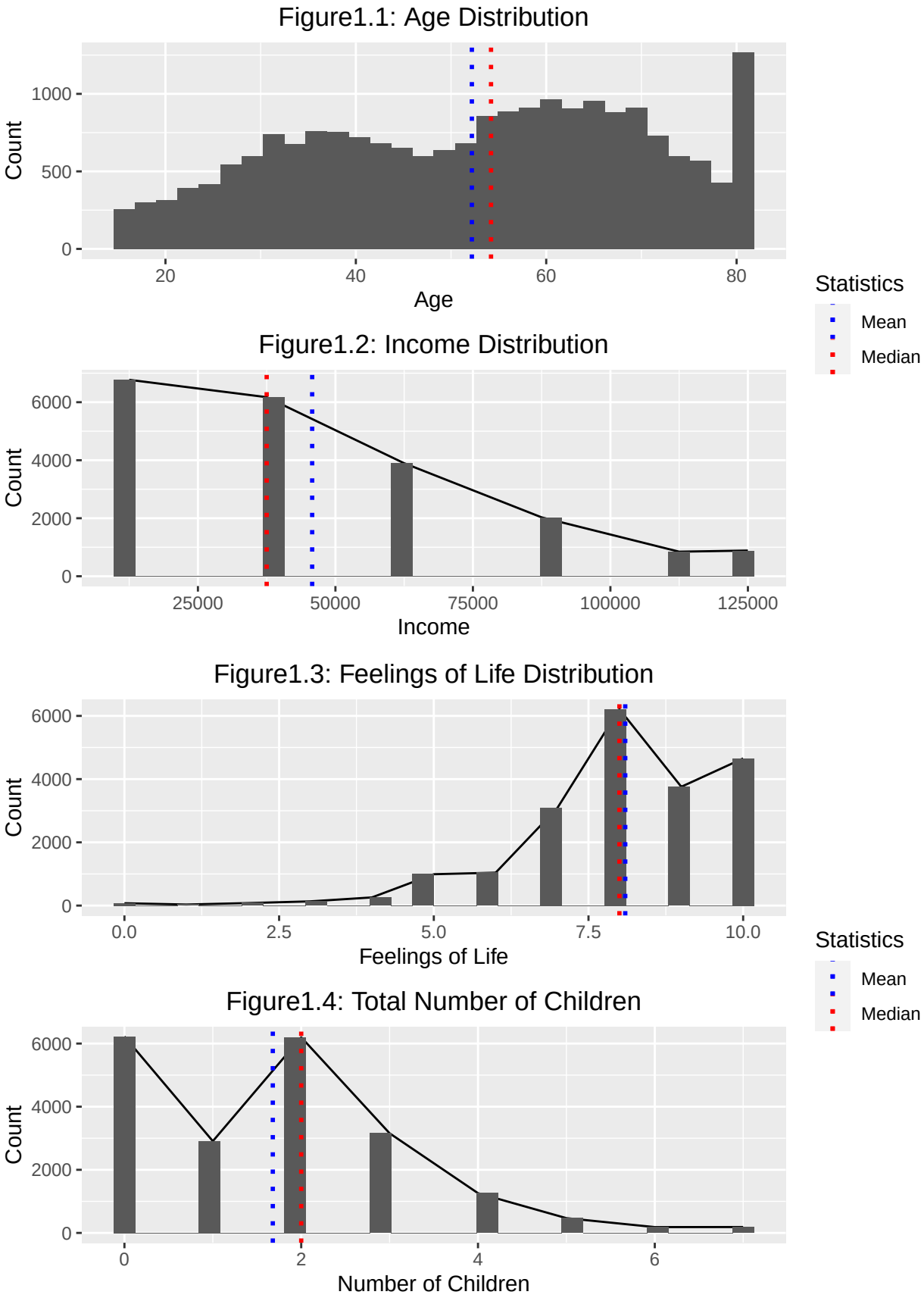
```
## # A tibble: 6 x 5
##   caseid   age income_respondent feelings_life total_children
##    <dbl> <dbl>             <dbl>         <dbl>          <dbl>
## 1      1  52.7             37500             8              1
## 2      2  51.1             12500            10              5
## 3      3  63.6             37500             8              5
## 4      4  80               62500            10              1
## 5      5  28               12500             8              0
## 6      6  63               12500             9              2
```

Summary Table

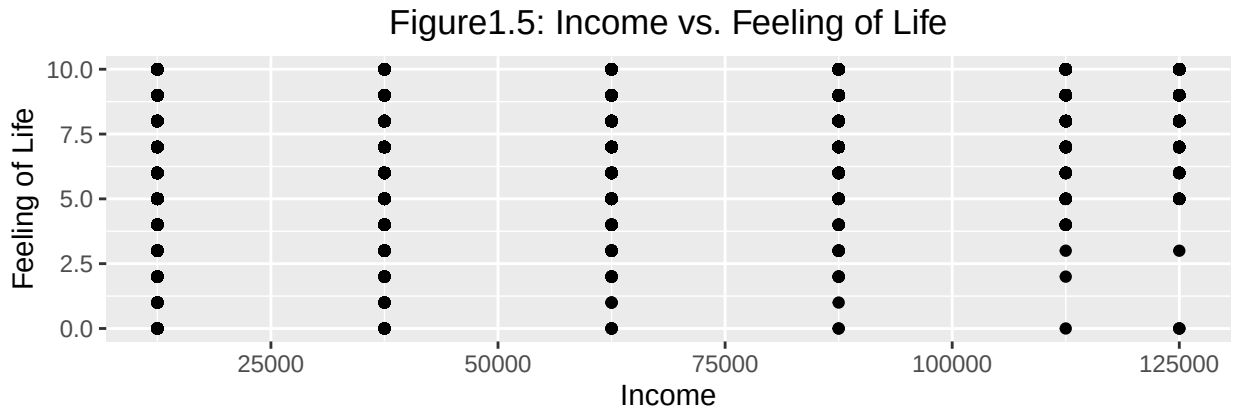
Table 1: Summary of Variables

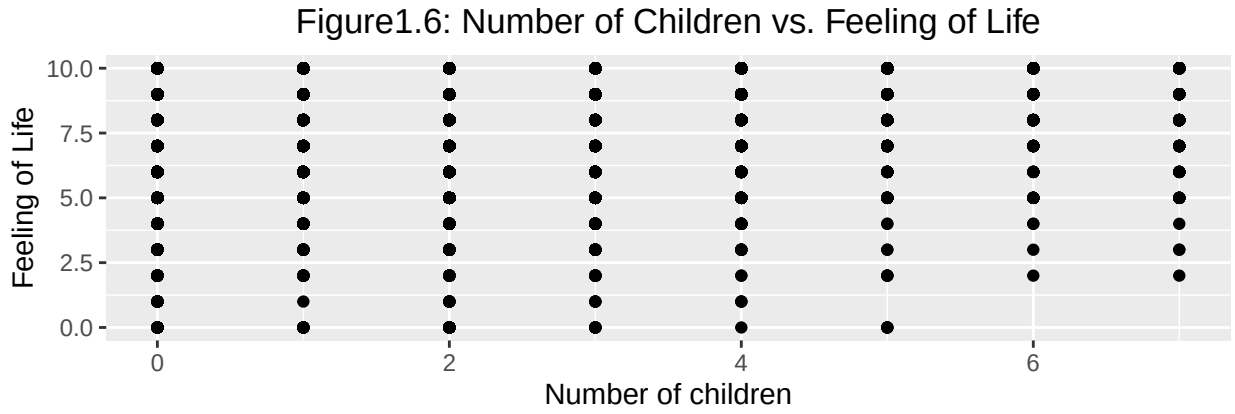
	Age	Income	Feelings of Life	Number of Children
Mean	52.2	45775.29	8.1	1.68
Median	54.2	37500	8	2
Minimum	15	12500	0	0
Maximum	80	125000	10	7

Histograms of Variables of Interest



Scatterplots



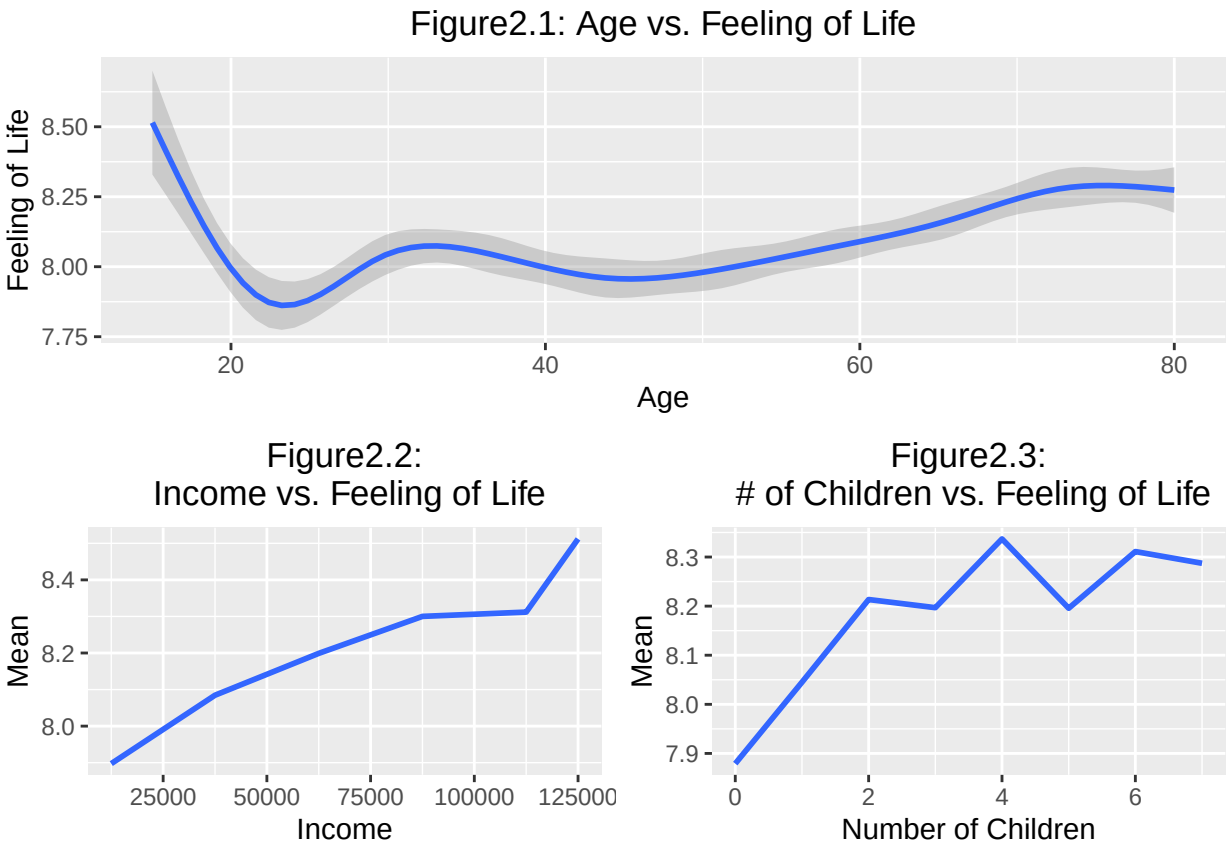


Further Data

Mean versus Income/Number of Children

```
gss_income <- gss %>%
  group_by(income_respondent) %>%
  summarise(mean_feel = mean(feelings_life, na.rm=T), .groups = 'drop')

gss_child <- gss %>%
  group_by(total_children) %>%
  summarise(mean_feel = mean(feelings_life, na.rm=T), .groups = 'drop')

#Omitted NA values for calcuating Means.

fig2.1 <- gss %>%
  ggplot(aes(y = feelings_life, x = age)) + geom_smooth() +
  labs(x = "Age", y = "Feeling of Life", title = "Figure2.1: Age vs. Feeling of Life") +
  theme(plot.title = element_text(hjust = 0.5))

#Age can be plotted without mean, because ages are more consistent. Other variables
#would produce error when plotting smooth line.

#So we plotted these two plot by using mean for each group, might be useful.
```

```
fig2.2 <- gss_income %>%
  ggplot(aes(y = mean_feel, x = income_respondent)) + geom_smooth(stat='identity') +
  labs(x = "Income", y = "Mean", title = "Figure2.2:\n Income vs. Feeling of Life") +
  theme(plot.title = element_text(hjust = 0.5))

fig2.3 <- gss_child %>%
  ggplot(aes(y = mean_feel, x = total_children)) + geom_smooth(stat='identity') +
  labs(x = "Number of Children", y = "Mean",
       title = "Figure2.3:\n # of Children vs. Feeling of Life") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(fig2.1, fig2.2, fig2.3, layout_matrix = rbind(c(1), c(2,3)))
```



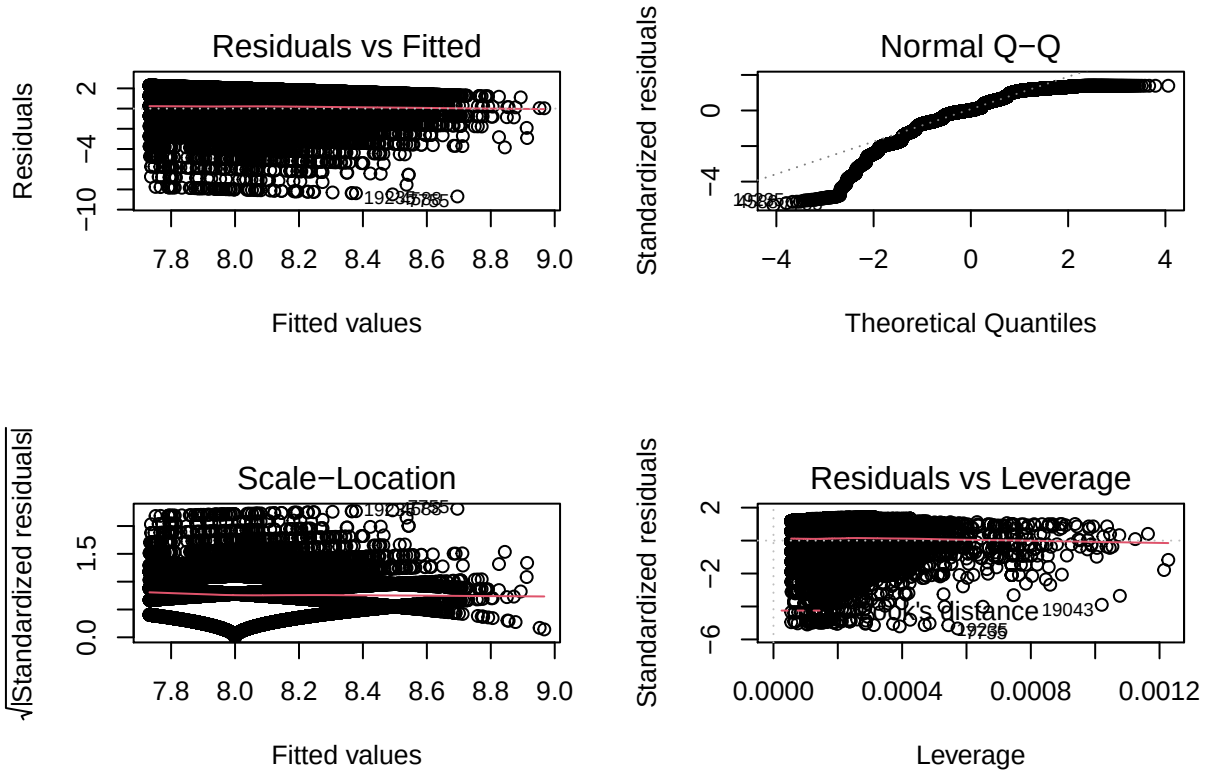
Linear Regression Model

```
mod <- lm(feelings_life ~ age + income_respondent + total_children, gss)
```

Table 2: Regression Statistics

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	141.38	141.38	53.16	3.18214369163231e−13
income_respondent	1	552.4	552.4	207.72	7.37360433804006e−47
total_children	1	243.11	243.11	91.42	1.2935341689774e−21
Residuals	20308	54007.19	2.66	NA	NA

	Value of Regression Residuals
Min	−8.6956
1st Quantile	−0.879
Median	0.0333
3rd Quantile	1.1291
Max	2.2677



Discussion

Before we investigated the specifics in terms of causal relationships between variables, we wanted to example the spread of each of the variables that we decided to investigate, in order to get a better sense of the data we were working with.

Figure 1.1, the distribution of age, is bimodal, with peaks around the 30s and 60s. Outside this spread, there's a huge population of respondents who are in their 80s, strangely not following the general trend of the distribution. The mean and medians are 52.2 and 54.2 years of age respectively, and the data appears to be, in a very broad sense, left-skewed.

Figure 1.2 is the distribution of income, and as you can see the data appears to be like a bar graph despite being a histogram due to the fact that we created this variable by finding the midpoints of each income bracket in the GSS survey data. For this reason our analysis of this section will be pointedly less accurate than say, age, but we can notice from the histogram that income is broadly right-skewed, the great majority of people making less than $50,000 a year. It was for this reason that we decided it would be okay to relabel the income bracket of people who made more than $125,000 as $125,000 because it does not apply to a huge part of our population and most of our analysis would be centred lower in that particular explanatory variable. The mean and medians are $45,775.29 and $37,500 respectively, reflecting that right skew.

Figure 1.3 is our chosen response variable, feelings of life. The mean and median are 8.1 and 8, and graphically there appears to a left skew to the data, meaning a great majority of the respondents feel generally good about life. This motivates the question of whether any of the previous variables motivate this distribution—if people with a higher income tend to have a better outlook than those who don't, whether children influence one's influence, and if getting older gives one a more charitable and understanding perspective on life.

Figure 1.4 is the distribution of number of children, and there is also a right skew to this data, reflected in the

means and medians being 1.68 and 2 respectively. This does fit with the consensus that in more developed countries people tend to have less children (due to their being a less economical need for a larger family).

Conclusion

References

- Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816