General Social Trends Within a Canadian Family from the 2017 GSS Survey

Allison Li, Bianca Pokhrel, Yitian Zhao, Zikun (Alex) Xu

10/19/2020

Abstract

The General Social Survey is designed to gather data on and analyze social trends in Canada to serve as an important insight on social policy issues. We are using the 2017 Canadian General Social Survey to explore the the different social factors affecting the number of children Canadians are having. In order to analyze the various factors that affect starting a family, we have decided on looking into variables such as age, population centers, marital status, and education level to analyze the sort of living conditions Canadians hold themselves to. The conclusions made in this report show a glimpse of the living standards in Canada in 2017 and what areas need further support when considering social programs.

Introduction

In our analysis, we used the 2017 Canadian General Social Survey which contains data for various social topics such as family life, social identity, care-giving, and volunteering. It is purposed for analyzing social trends within Canadians to observe changes in living conditions over time and possibly introduce or adjust social policies. Its data collection methodology mainly consists of Random Digit Dialing collecting cross-sectional data from Canadians over the age of 15 throughout 10 provinces. This sort of interview typically lasts 40 min is collected through a 6-12 month period.

Given the wide array of data available from the GSS, we have decided to focus specifically on factors forming the average Canadian family. There are several variables we have looked into which possibly form a relationship with the number of children Canadians are having such as age, population center, marital status, and education level. With age and population centers, we see there is a general trend of older Canadians living rural areas typically having more children. We figure this is attributed to the fact that the younger generation of Canadians living in dense urban areas generally have different values with starting a family since they grew up with more social reform changes for womens' rights and may be more career-oriented in a dense population center. This ties in with educational factors as we have seen that education above a bachelor's degree/diploma does not seem to have a large effect either as individuals with higher level education may be likely to pursue research or a career over starting a family. Marital status also is not shown to have a large impact on predicting the number of children although an interesting observation of the regression intercept suggests that on average a married couple will have 2 children.

This study will analyze the relationships of various factors with the number of children an individual has. While we provide some surface level insight into how families are affected through our chosen variables, there is room for further analysis in each area. For example, a better study of age demographics could be analyzed through the birth rates in Canada throughout the years instead of being limited to GSS 2017. Interesting factors such as the increase in womens' education which is suggested to play a role in declining birth rates could be measured through womens' university graduation rates compared to birth rates in Canada. Couple dynamics especially with gender wage differences and parental leave definitely contribute to a couple's

decision to start a family and we can analyze this by looking at data from Canadian social programs and pay differences at large companies.

In this paper we found that certain factors are more significant than others in terms of modeling; for example, predicting the number of children an individual decides to have. We have concluded that based on the responses of the GSS survey, the amount of income an individual has does not affect how many children that individual has, which tells us income levels do not determine the number of children they decide to have even though they can afford to raise more children, other factors such as education, living environment, and martial status are more significant when it comes to starting a family.

Data

A general strength of the survey is that the data resulting from the questionnaire is detailed and covers many areas of interest. Furthermore, since the GSS is government run they have access to reliable information/sources for Canadians and their phone numbers for their frame. However, a noticeable weakness was that the survey used a stratified design where they reported in the documentation that between strata, there were significant differences in sampling fractions. Meaning that some areas could have been over-represented or under-represented making their "un-weighed" sample an inaccurate representative of the population.

The questionnaire starts off with basic questions to obtain more information about the demographic. A large benefit was that these questions had little non-response answers due to their survey methodology. Moreover, the questionnaire contained many areas of interest, which were helpful for research. A weakness of the survey that was unavoidable is that many questions were open ended, and as a result a lot of work had to be done to sort and can the data retrieved. A major strength of this survey is that it contained many questions, 86 variables to be exact, which is useful to run regression on various coefficients to see which social factors are more important than others for the topic of Canadian family's. The cleaned data that we used for analysis was referenced from Prof. Rohan Alexander and Prof. Sam Caetano's cleaned data file of the GSS 2017 survey. The GSS conducts phone surveys in the Canadian provinces, and their population are people 15+ years old in Canada, excluding residents of the territories. Each of the 10 provinces were divided into strata based on geographic areas. Many of the Census Metropolitan Areas were considered their own strata. Overall, there were a total of 27 strata.

The frame of the survey consists of the list of phone numbers in use that are available to Statistics Canada and the Address Register, which is a list of dwellings. Each record in the survey frame was assigned to a stratum in its province, and records were chosen using simple random sampling without replacement within each stratum. When minimum sample sizes targets were met in a province, they allocated the remaining samples to remaining strata to maintain balance. An eligible respondent was selected to participate from the household represented by the chosen record. Selected household were terminated if they did not meet the eligible criteria. Their final population sample consisted of 20602 respondents.

In addition, their survey methodology did not permit for non-responses on questions that were required for weighting. Thus values were imputed based on existing data when a respondent was not comfortable providing details. According to their reported non-response rates, they had very few in person non-response rates. However, they reported that a non-response rate of 37.4% at the household level. To handle their non-responses, they had to do a three-stage non-response adjustment where they used auxiliary data from Statistics Canada. Non-response rates contributed to the weighting assigned to each respondent since non-response rates varied by demographic groups, making the unweighted sample less representative.

The dataset we are using itself contains 81 variables that covers many domains of interest. However, the main variable we are concerned with predicting is total_children while the predictors are age, pop_center, marital_status, and education. When picking out the variables as predictors, we also considered age, which was really similar to age_at_first_birth. However, age was chosen instead since it seemed like the age of a person overall would have more impact.

Dataset

The data-set was obtained from the data used for the GSS in 2017 which was provided by Statistics Canada. Statistics Canada is government run and mostly seems to be a reputable source of data, meaning that it is unlikely for them to modify collected data to show more favorable statistics.

Fig 1. Age of Respondent

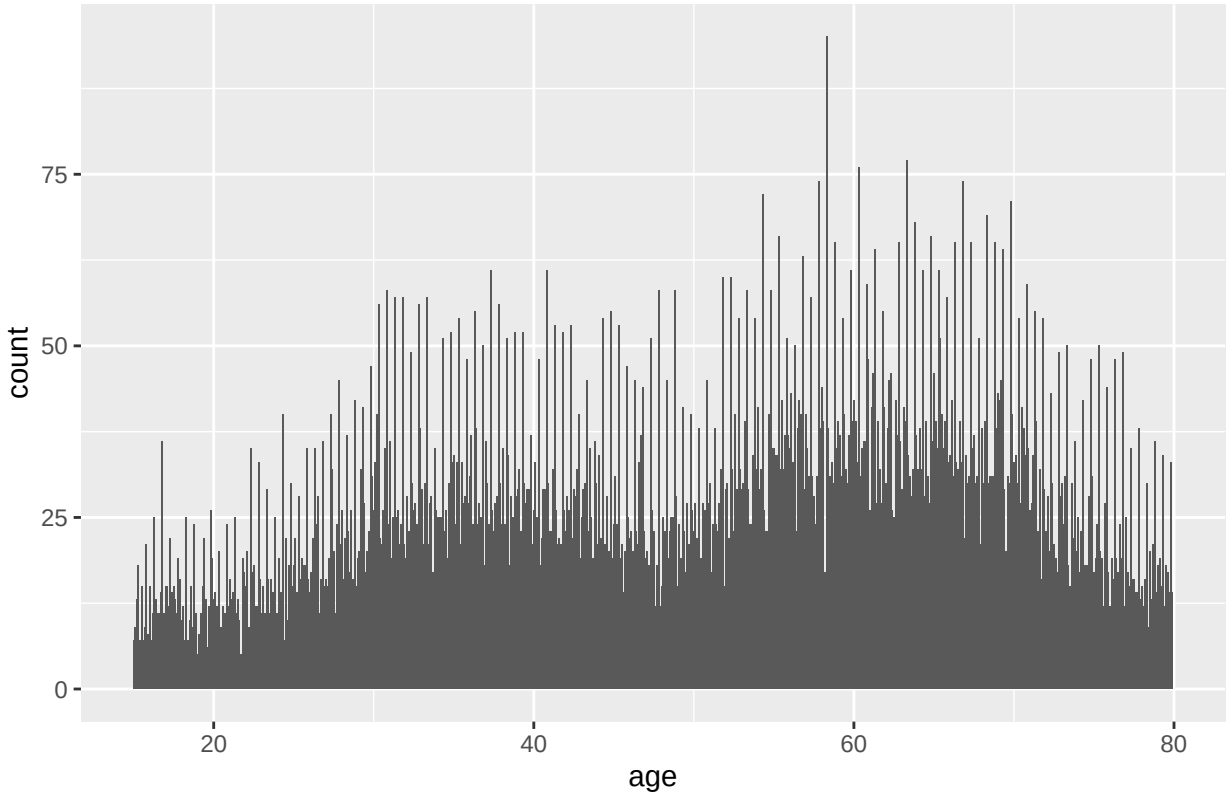


Fig 2. Marital Status of Respondent

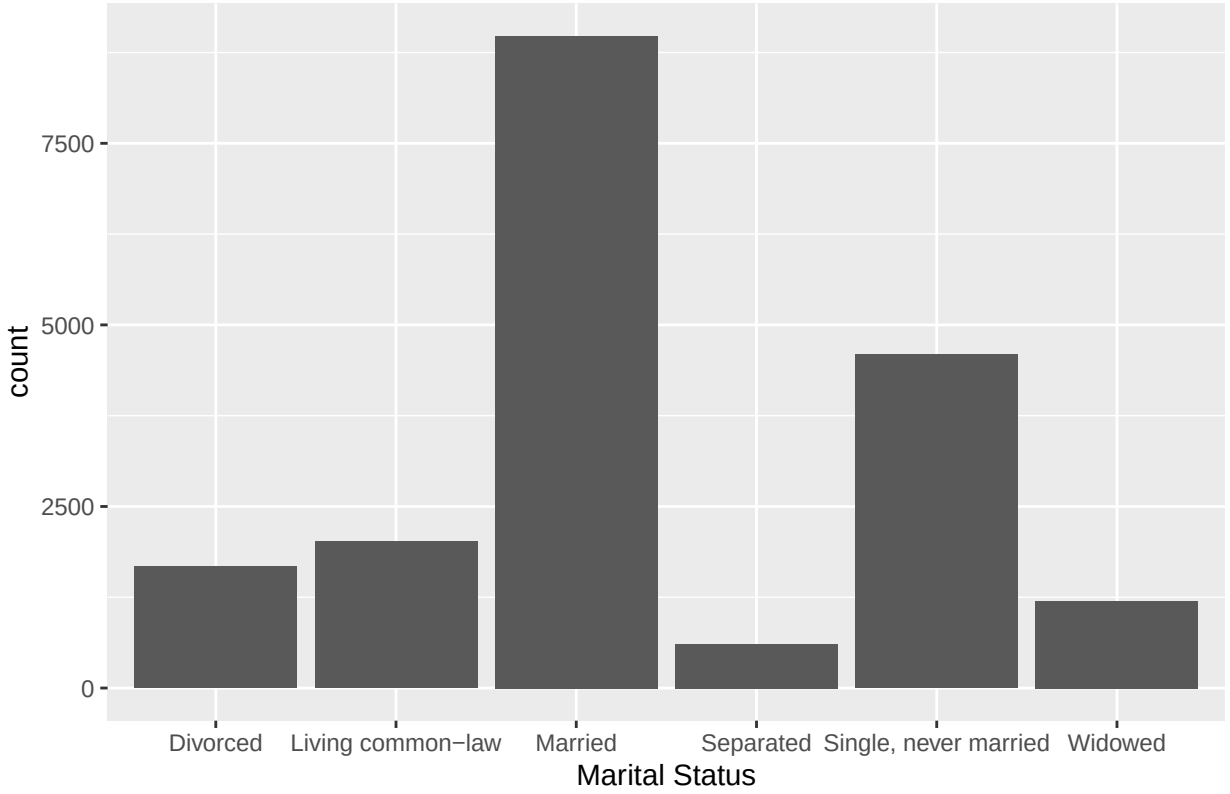


Fig 3. Education Level of Respondent

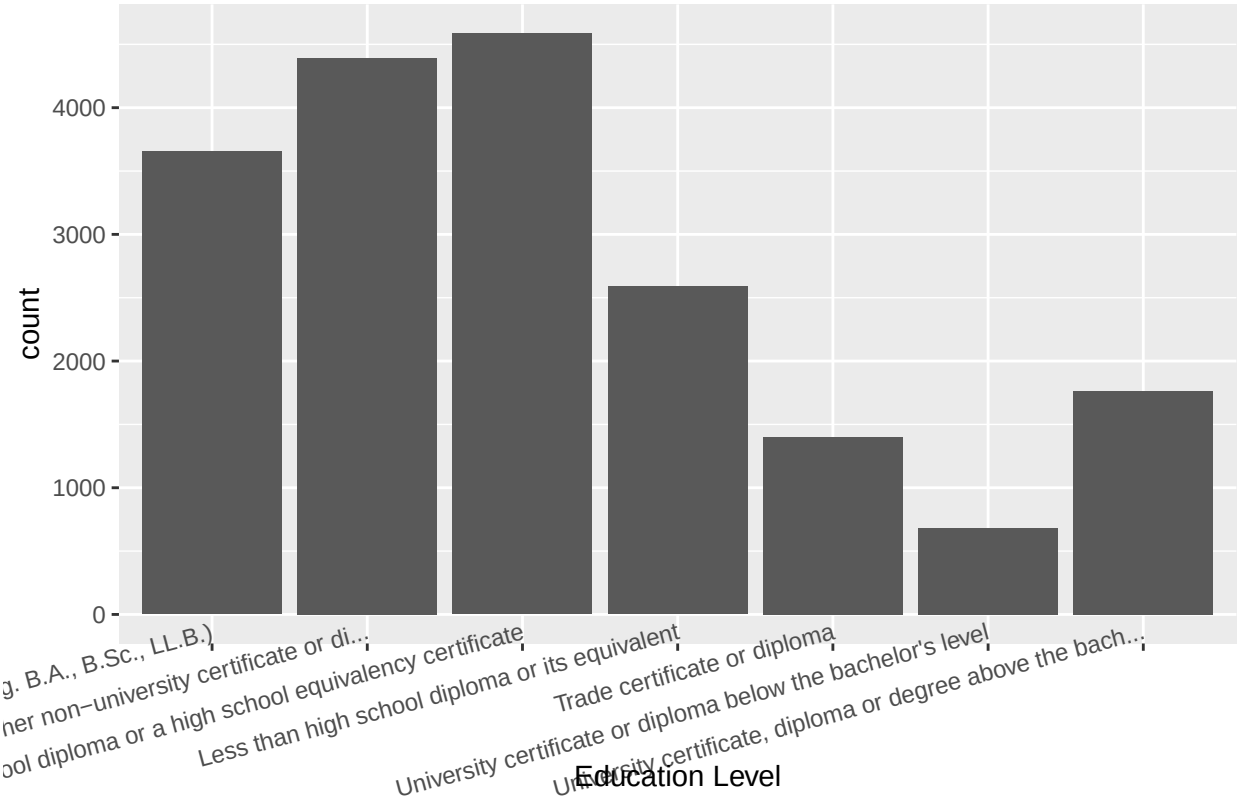


Fig 4. Area of Living of Respondent

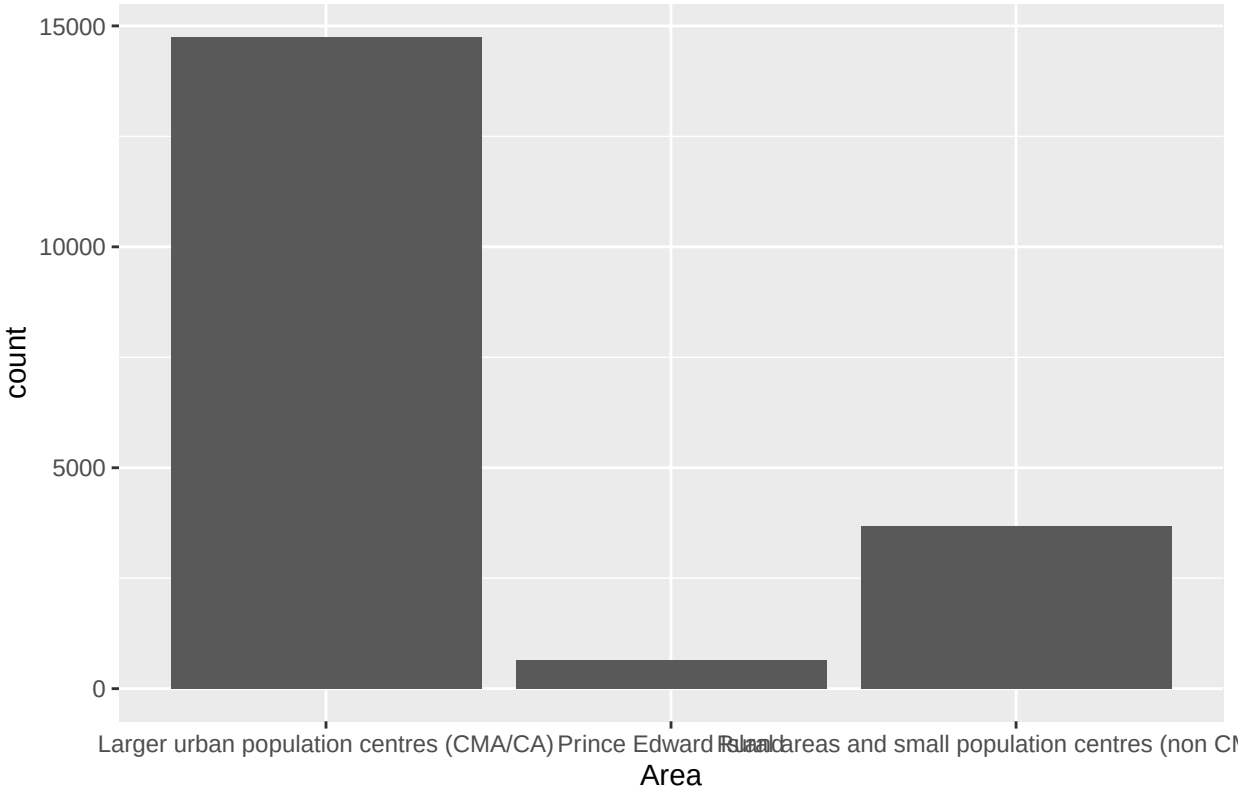
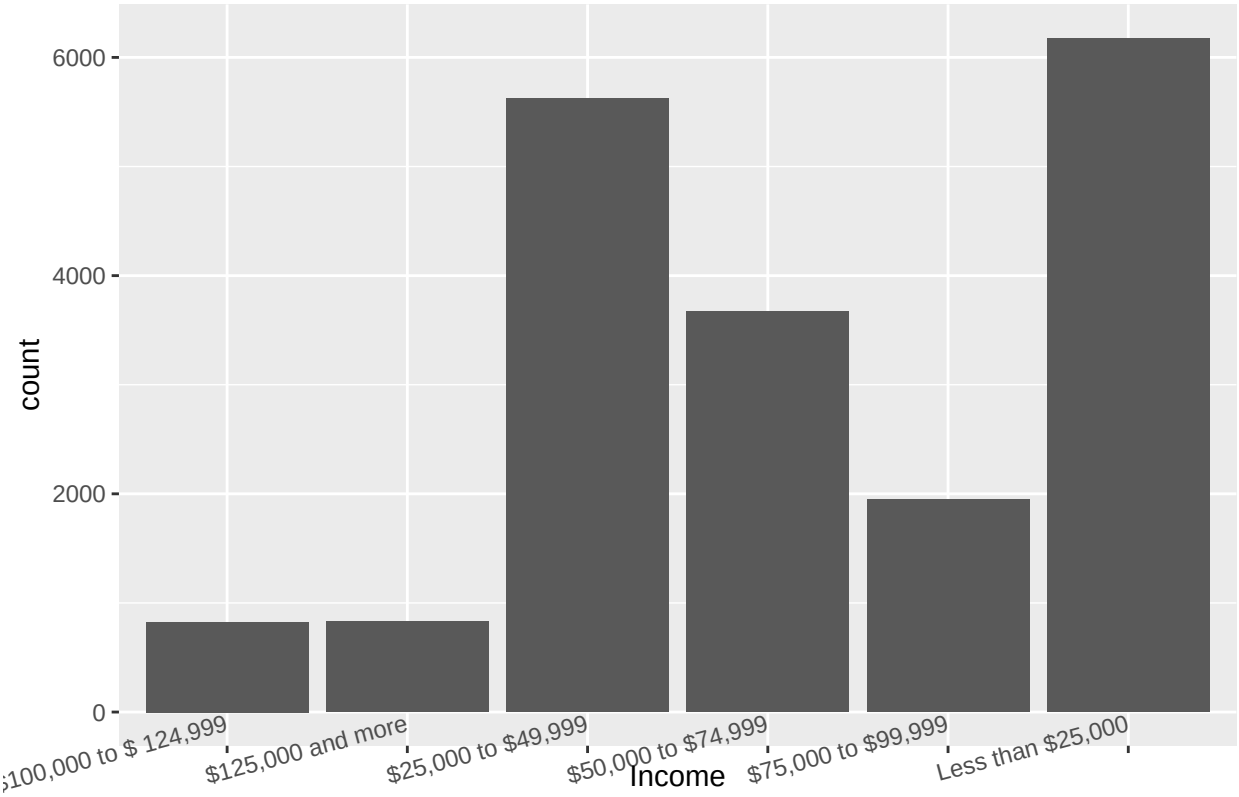


Fig 5. Income of Respondent



From the figure 1-5, you can see and get a better understanding of the distributions of age when concerning our predictors. Moreover, we can see that marital_status, education, and pop_center are qualitative variables while age is quantitative. We can also see that total_children only has values with whole numbers and lies between the range of (0, 7). An unique observation that we can observe is that there seems to be a large amout of respondents who are 80 years old, which can skew the answers.

Models/Graphs

When deciding on the overall model for our data, we decided to approach it a single variable at a time. We fitted each of our predictor variables to a linear model and then looked at their relationships. Thus, if all of the predictor variables can be fitted to a linear model, then the overall model is linear as well. This is similar to step-wise regression, however we looked specifically at the p-values of each model rather than the total error. This allowed us to see which categories of each predictor variable had the most effect on the overall regression.

Age distributions corresponding with number of children he/she has

Fig 6. Age Distribution of Canadians with No Children

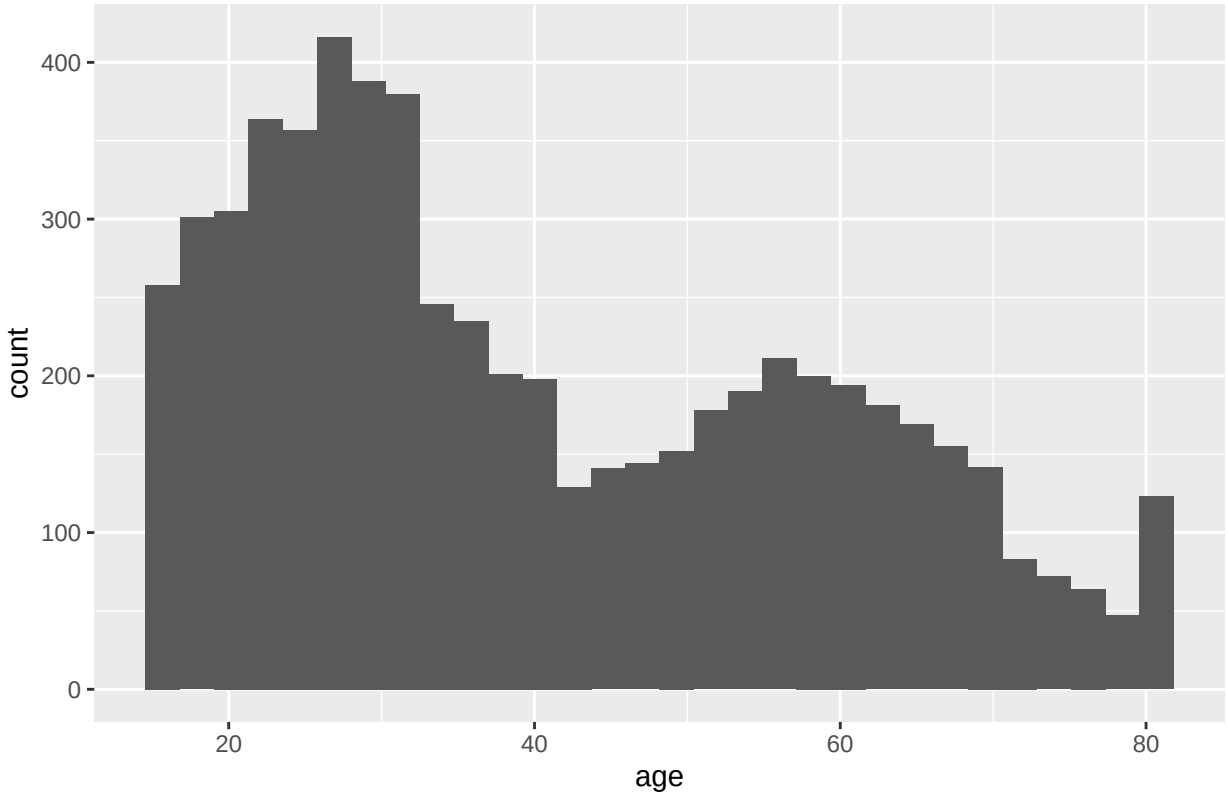


Fig 7. Age Distribution of Canadians with 1 Child

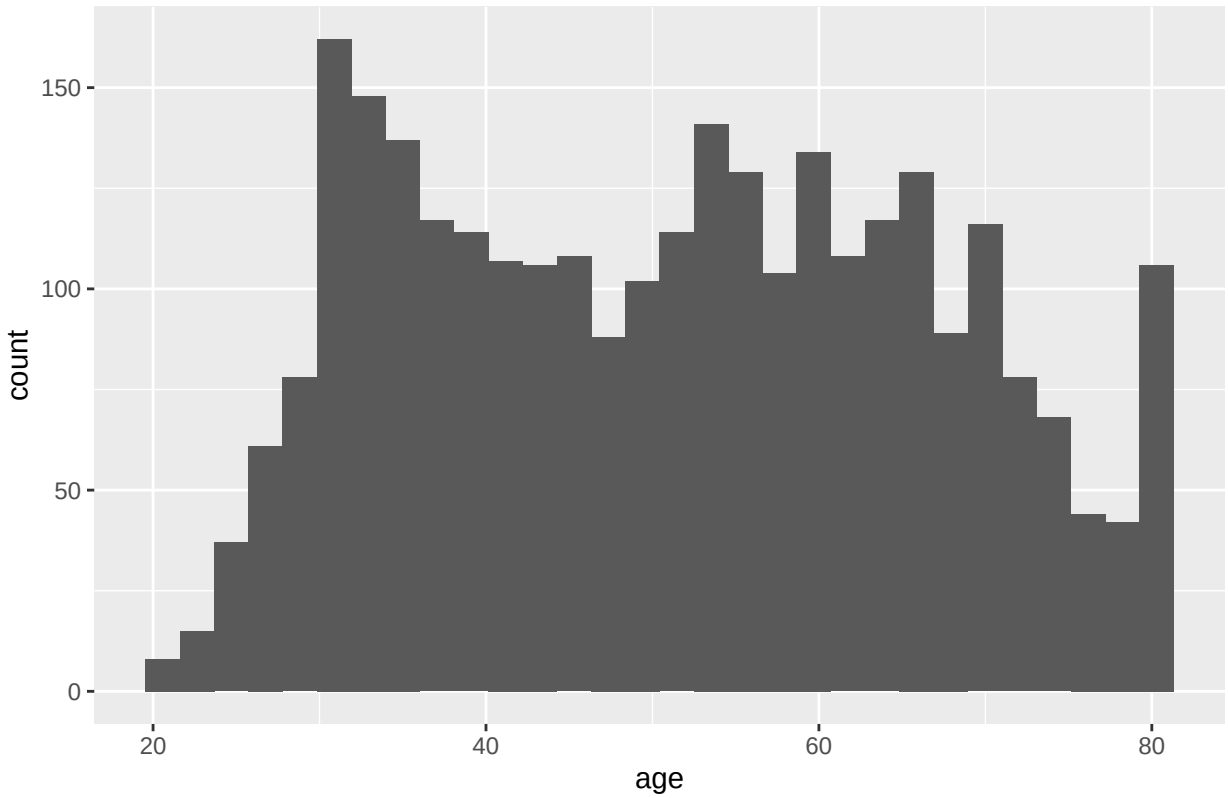


Fig 8. Age Distribution of Canadians with 2 Children

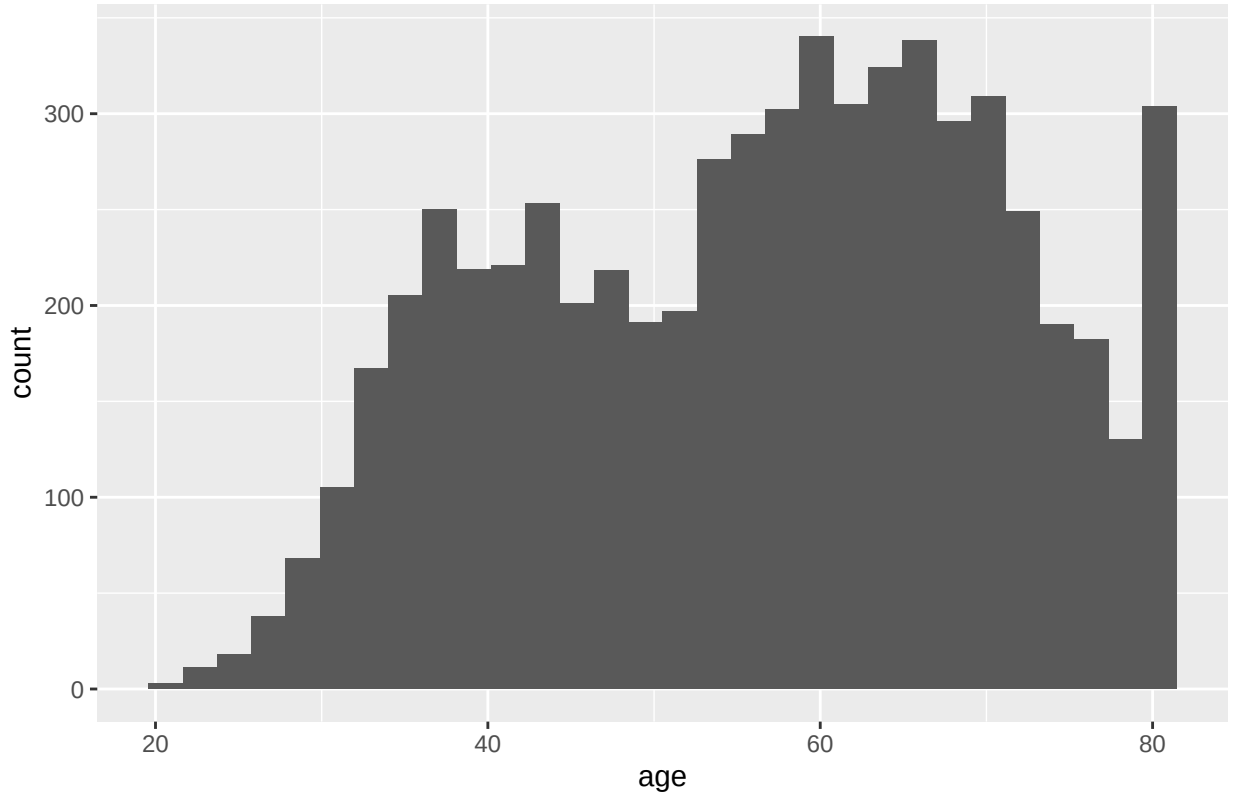


Fig 9. Age Distribution of Canadians with 3 Children

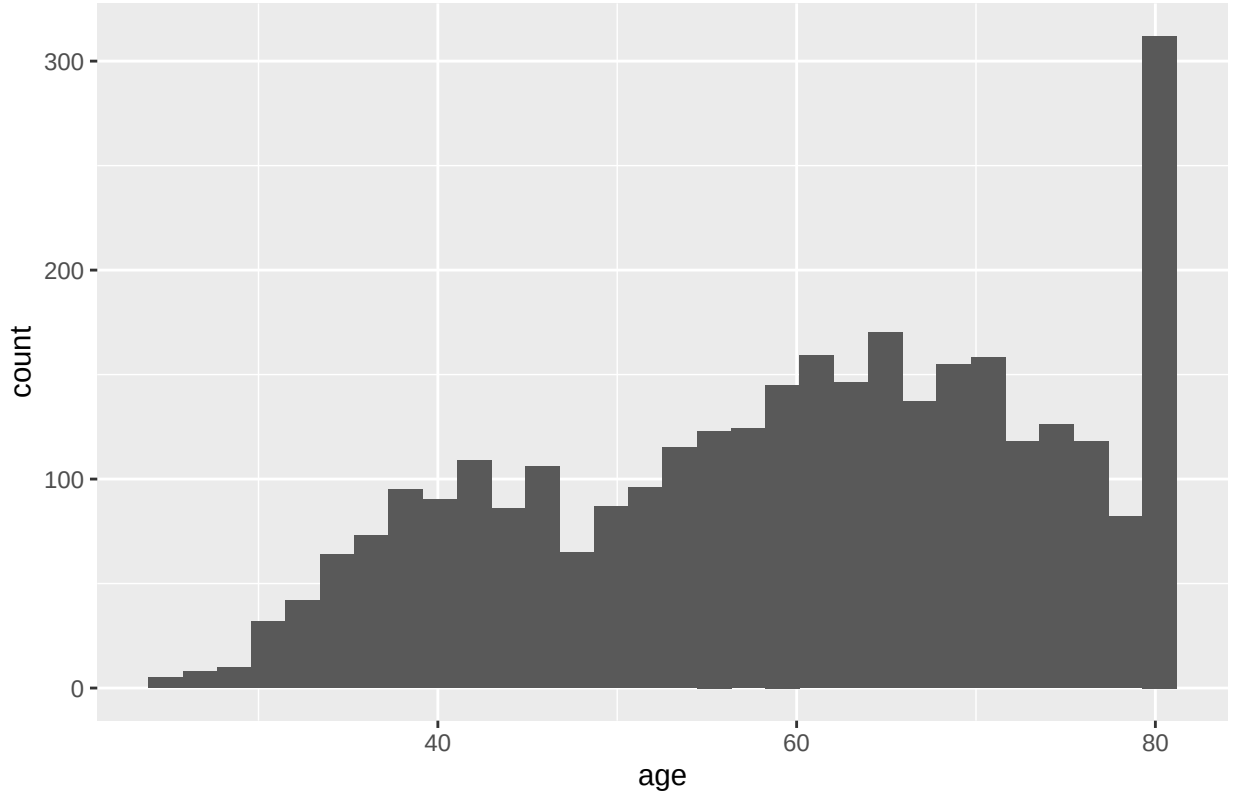


Fig 10. Age Distribution of Canadians with 4 Children

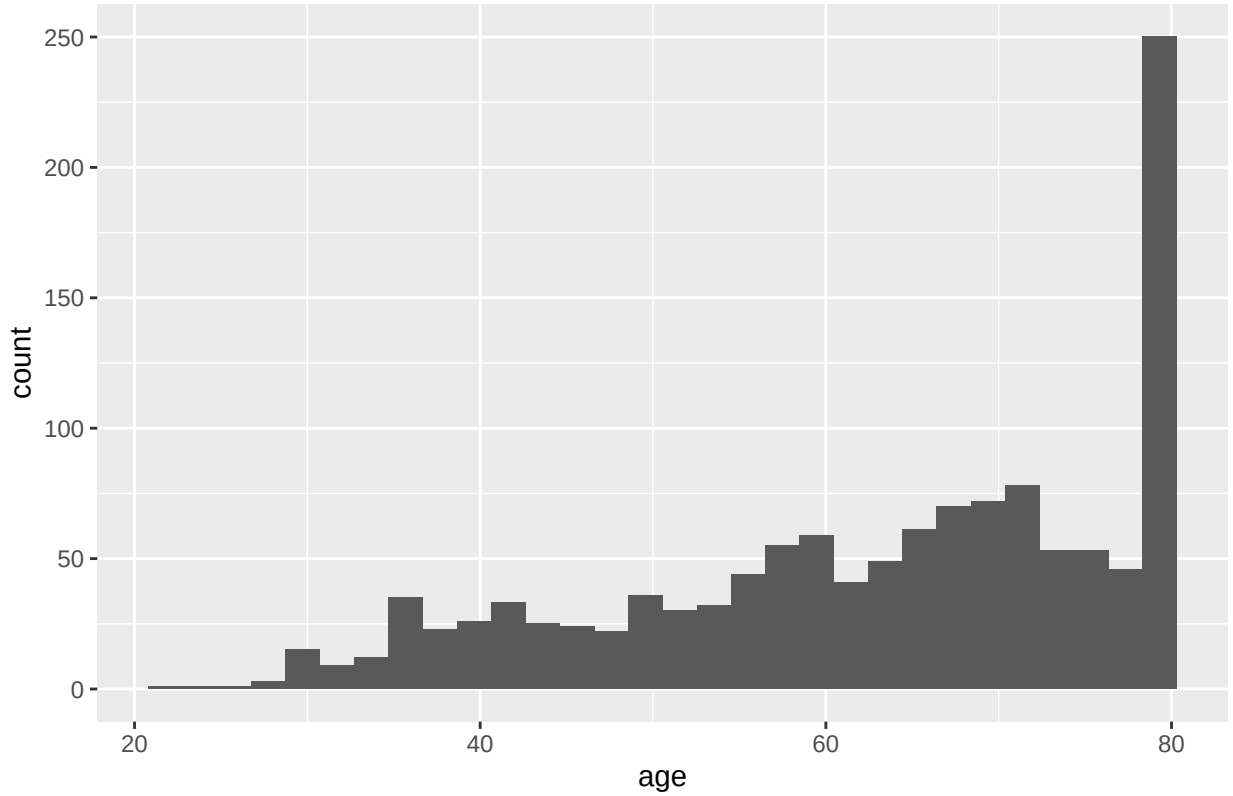


Fig 11. Age Distribution of Canadians with 5 Children

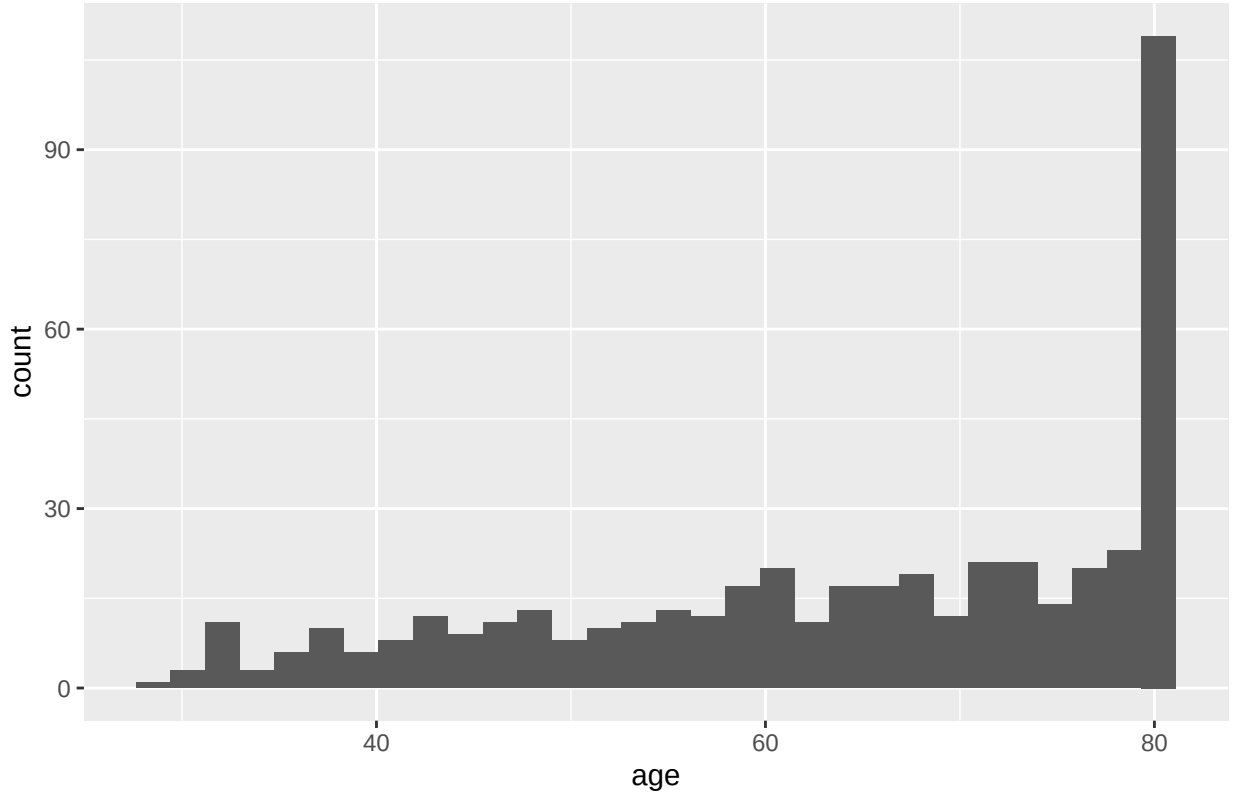


Fig 12. Age Distribution of Canadians with 6 Children

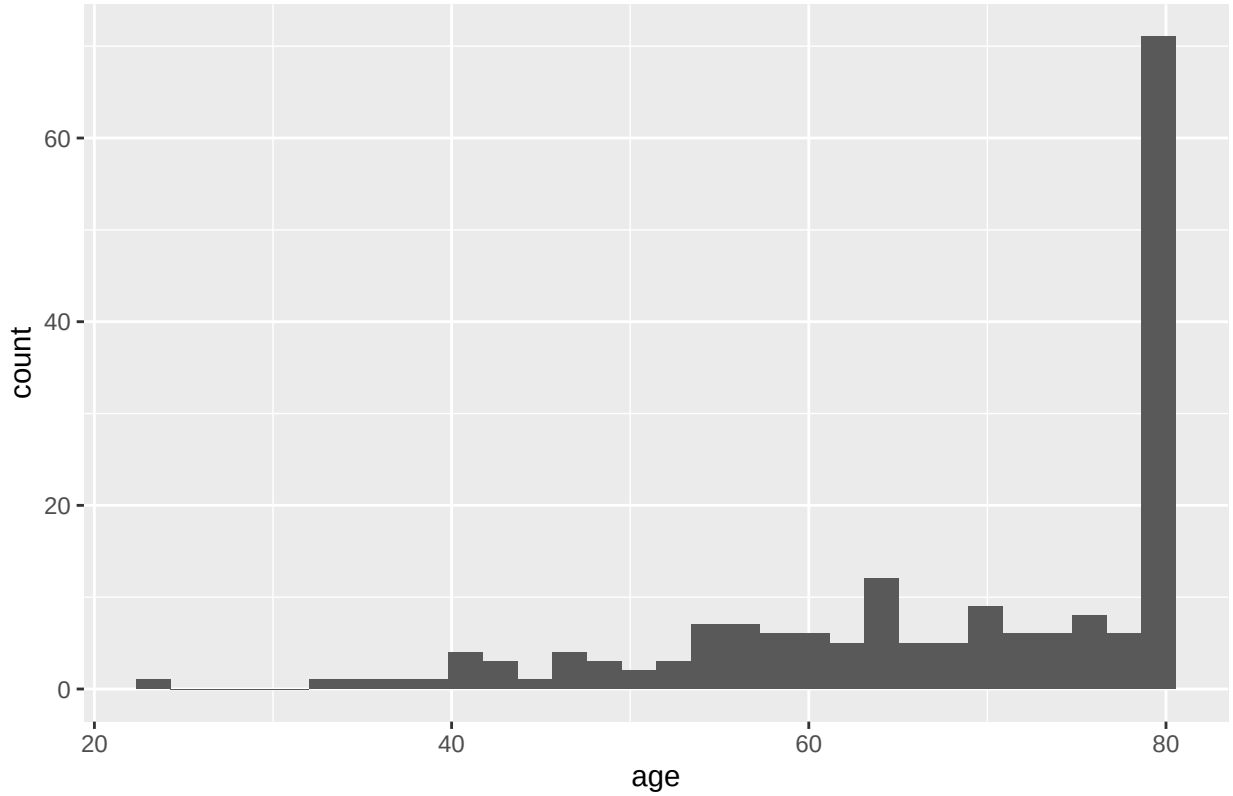
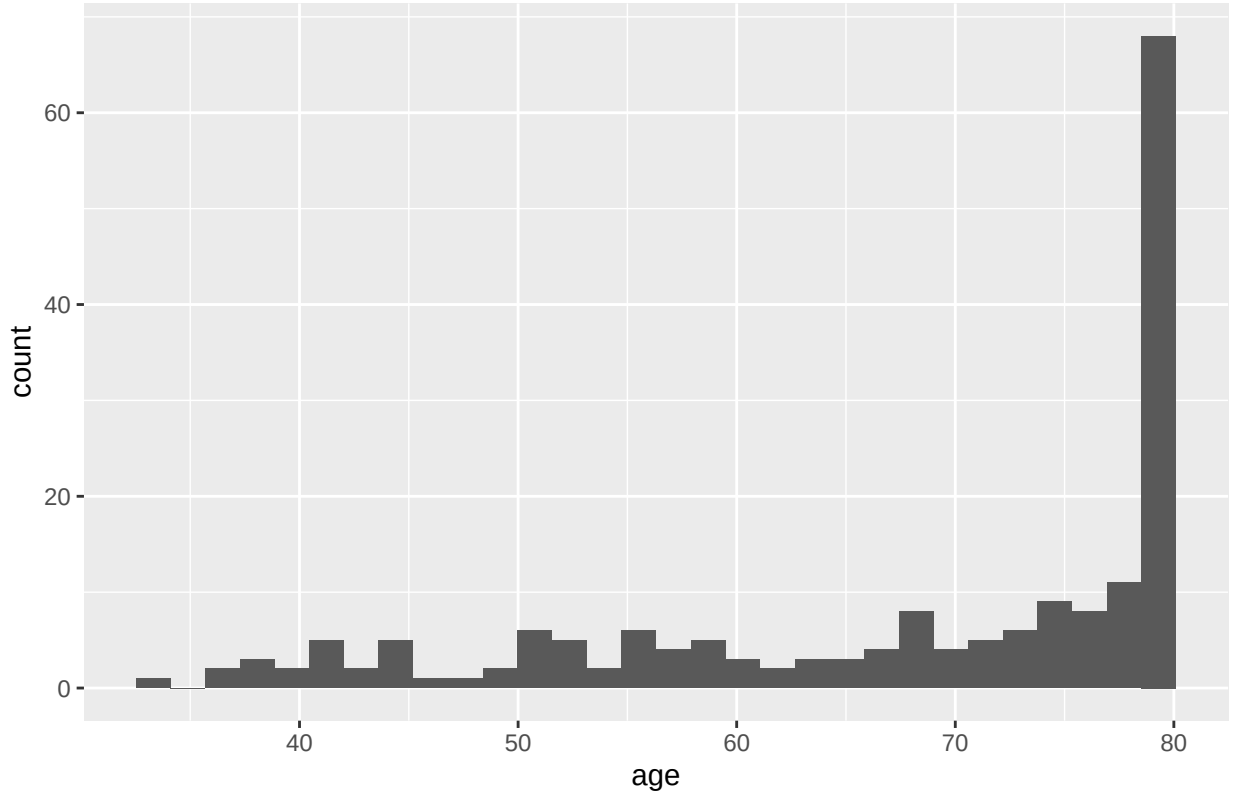


Fig 13. Age Distribution of Canadians with 7 Children



Modelling the area of living vs the total number of children in the family

Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.5762669	0.0164657	95.73035	0
pop_centerRural areas and small population centres (non CMA/CA)	0.5059044	0.0367082	13.78177	0

Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0821713	0.0328082	63.46504	0
pop_centerLarger urban population centres (CMA/CA)	- 0.5059044	0.0367082	- 13.78177	0

Predict the total number of children by marital status

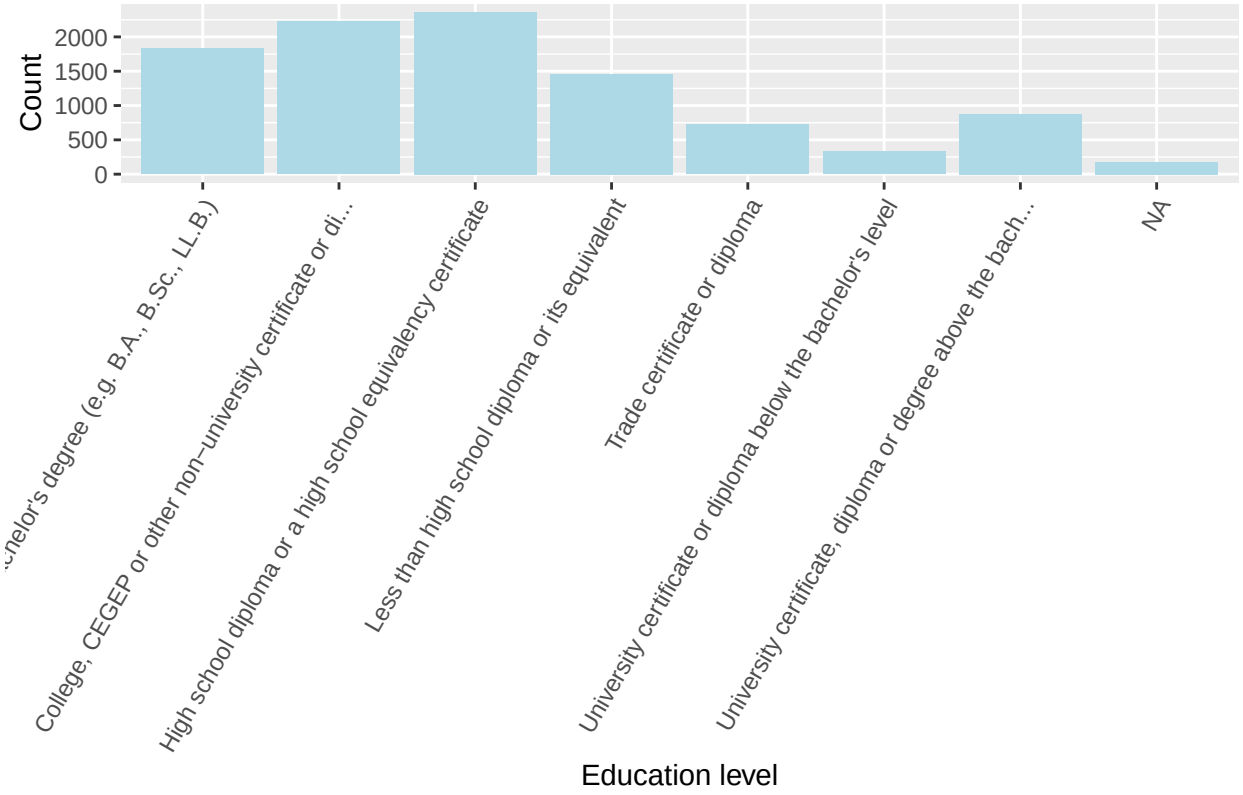
Model 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9877642	0.0426784	46.575391	0.0000000
marital_statusLiving common-law	-0.4926375	0.0584588	-8.427086	0.0000000
marital_statusMarried	0.1058631	0.0466778	2.267951	0.0233535
marital_statusSeparated	0.2064106	0.0843845	2.446073	0.0144593
marital_statusSingle, never married	-1.6256188	0.0504706	-32.209221	0.0000000
marital_statusWidowed	0.5892015	0.0602895	9.772866	0.0000000

Education levels vs total number of children

Model 4

Fig 14. Education Level of Respondents

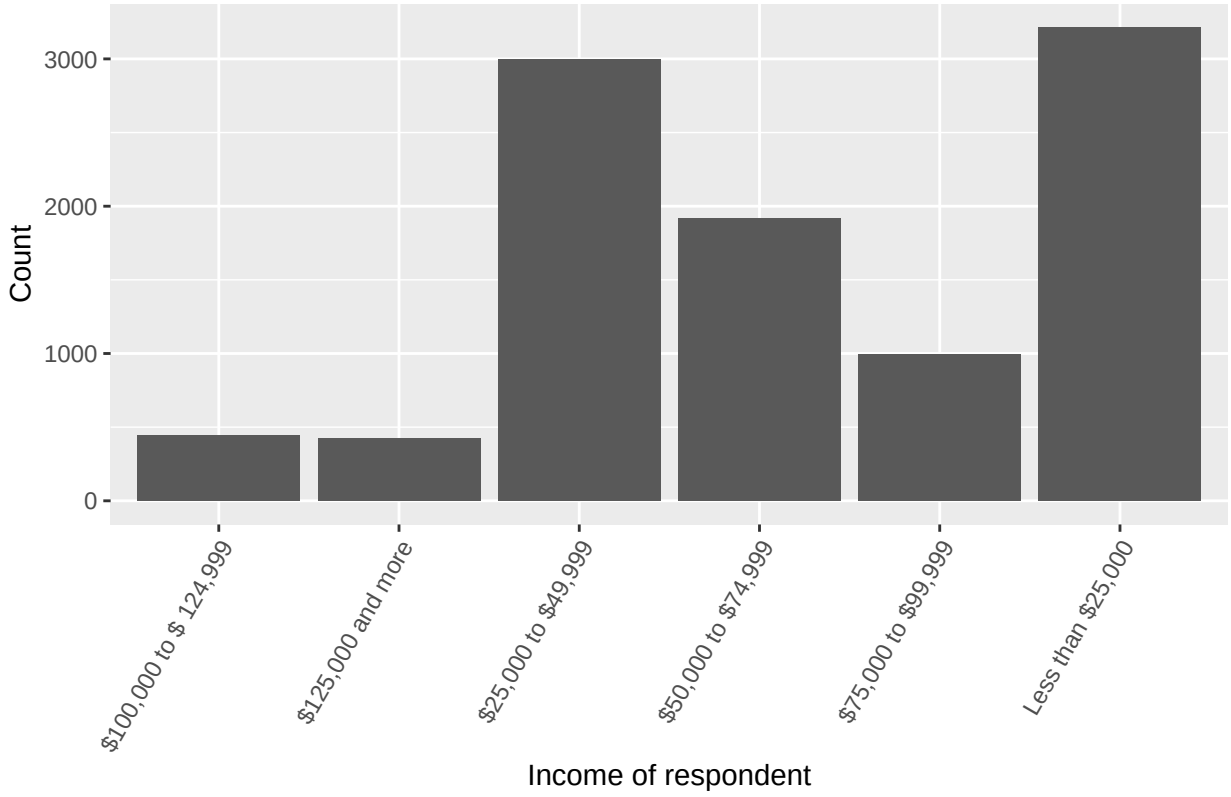


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.4153930	0.0343072	41.256441	0.0000000
educationCollege, CEGEP or other non-university certificate or di...	0.2271104	0.0463446	4.900469	0.0000010
educationHigh school diploma or a high school equivalency certificate	0.2760016	0.0457278	6.035754	0.0000000
educationLess than high school diploma or its equivalent	0.6536904	0.0514960	12.694007	0.0000000
educationTrade certificate or diploma	0.3703213	0.0643339	5.756242	0.0000000
educationUniversity certificate or diploma below the bachelor's level	0.3466552	0.0875881	3.957791	0.0000762
educationUniversity certificate, diploma or degree above the bach...	0.0725519	0.0604365	1.200465	0.2299879

Income vs number of children

Model 5

Figure 15. Income of Respondents



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.7142857	0.0701182	24.4485117	0.0000000
income_respondent$125,000 and more	-0.0074299	0.1006166	-0.0738439	0.9411361
income_respondent$25,000 to $49,999	-0.0097403	0.0751847	-0.1295512	0.8969242
income_respondent$50,000 to $74,999	-0.0241081	0.0778932	-0.3095015	0.7569465
income_respondent$75,000 to $99,999	-0.0529954	0.0844804	-0.6273096	0.5304707
income_respondentLess than $25,000	-0.0718067	0.0748500	-0.9593418	0.3374099

Overall model with all factors

In this final model, we ran a generalized linear regression on the various variables that we felt were overall most important to predict the total number of children that a family has depending on the various social factors of that individual. In model 5, we concluded that the income of the respondent was not significant enough to be included in the overall model. In this overall regression model, we see that the predicted number of children in an average household would be the sum of the estimated coefficients. The generalized linear model can be interpreted as total number of children would equal the sum of the estimates of population centers (rural or urban area), education level (High school, College, University, or above Bachelor's degree), and marital status (being single, married, divorced etc), with the exception of income since in model 5, it is concluded that it is not significant enough to be included in the overall model.

We chose these three main factors because, in discussion, we talked about the impact that living area, education, and marital status has on whether a family will choose to have multiple children or not. Living in a rural area or urban area affects the amount of space a family has in terms of raising a child and we concluded in discussion that people living in urban areas tend to have less children compared to people living in rural areas. Education is also a significant factor in this model because we concluded in discussion that respondents with higher education tend to have less children. We chose marital status to be included in this model as well to since we concluded in discussion that marital status does affect the number of children respondent's tend to have because for example respondents in a common-law relationship tend to have fewer children than respondents that are divorced. These factors are all categorical because given GSS survey, the answers only fit in different categories similar to how gender is divided into two categories. We chose to use total number of children as our independent variable since we want to analyze social characteristics that households in Canada have when it comes to raising children which is our main focus in this paper.

According to the summary statistic of this overall model, the coefficients are all categorical variables, this model would regress the dependent variable as factors in this case Y = b0 + b1X1 + b2X2+ b3X3 where Y would be the number of children; b0 equals the intercept; b1, b2, and b3 would be coefficient estimates; and X1, X2, and X3 would be the dependent variable which in this case would equal to 1 since we regressed it as factors. To further analyze this general linear model, we interpret the summary statistics as for example if the respondent lives in a rural area, with a high school diploma, and is currently married, we can use the estimates of these coefficients to show many children he/she has; in this case it would be 1.7172+0.2862+0.1094 = 2.11 children on average.

We chose to do a simple linear regression instead of other models such as logistic regression or bayesian because our dependent variable, total number of children, is quantitative and logistic regression is only used to model a binary dependent variable. On the other hand, we do not know the priors for our variables therefore we would have used non-informative priors thus using bayesian linear regression would not be as helpful as if we had priors.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.0053918	0.0824919	24.3101731	0.0000000
pop_centerLarger urban population centres (CMA/CA)	- 0.3105260	0.0450572	- 6.8918197	0.0000000
educationCollege, CEGEP or other non-university certificate or di...	0.1571865	0.0560541	2.8041910	0.0050640
educationHigh school diploma or a high school equivalency certificate	0.2576415	0.0551049	4.6754732	0.0000030
educationLess than high school diploma or its equivalent	0.5479591	0.0642645	8.5266220	0.0000000
educationTrade certificate or diploma	0.3272292	0.0776688	4.2131344	0.0000256
educationUniversity certificate or diploma below the bachelor's level	0.3611535	0.1054174	3.4259383	0.0006177

	Estimate	Std. Error	t value	Pr(>|t|)
educationUniversity certificate, diploma or degree above the bach...	-0.0071332	0.0727838	-0.0980047	0.9219325
marital_statusLiving common-law	-0.4901199	0.0804170	-6.0947281	0.0000000
marital_statusMarried	0.1349706	0.0643999	2.0958198	0.0361489
marital_statusSeparated	0.1751935	0.1157982	1.5129214	0.1303635
marital_statusSingle, never married	-1.6298613	0.0691982	-23.5535102	0.0000000
marital_statusWidowed	0.5159092	0.0826668	6.2408267	0.0000000

Discussion/Results

Comparing age distributions with different total numbers of children

Through the Canadian General Social Survey, we can see there is a trend within the age demographics when comparing Canadians by the total number of children they have. Canadians with no children have slightly right skewed distribution which is to be expected as young adults typically don't have the financial stability and mindset to be starting a family.

As we observe figures 6-13 for age distributions of Canadians with more children, we see there is a trend of the graphs to be increasingly more left-skewed, especially with a large number of Canadians 80 and above with 3 or more children. This sort of trend is most likely attributed to generational differences in family values and lifestyle as modernization typically entails lower fertility.

According to Statistics Canada, there is a strong link between fertility rates and social and legislative changes in the mid-1900s where the high marriage rates and stable economy post WWII led to Canada's baby boom peaking at a fertility rate of 3.94 children per woman. This helps explain the stand out numbers in older Canadians past 80 years of age with a high total number of children in our data. We see these numbers lower for younger age groups as less influence of religion, more accessible contraception, and more accessible education for women all became significant factors for the declining birth rate beginning in the 1970s. The effects of such factors is visible in the 2017 GSS data and provides a glimpse of how different age demographics are shaped today.

Comparing number of children with rural vs urban areas in the city

In model 1, we look at the correlation between number of children and different regions in a city, specifically rural and urban areas. We run a linear regression model on these two variables in model 1 to see if the number of children differs between rural and urban areas in a city. The function that we used for this regression is lm (Chambers, 1992) and ran two different regression models to depict a difference between how many children adults have whom are living in urban areas vs rural areas.

As we can see from the summary statistics of model 1, the p-values for the coefficients are very small therefore meaning that we can reject the null-hypothesis and that the predictor is very significant and is likely a meaningful addition to our model. Also, given that the p-value for the dummy variable "Rural areas and small population centers" is very significant, this suggests that there is a statistical evidence of a difference in number of children between rural and urban areas. The linear regression model would be y = b0 + b1*x where y is the number of children, b0 being the intercept, b1 being the slope estimate, and x being the categorical variable 'areas'. Since we modeled the categorical variables by factors with two levels, 1 and 0, the x value for both model 1 and model 2 would be 1. Therefore we can interpret from the coefficients in model 1 that the number of children would be about 2.049 = 1.5579 + 0.4915 or (y = b0 + b1), if the respondent lived in a rural community with small populations; and on average 1.5579 or (b0) for the number of children for respondents living in urban areas with larger populations.

A variant of model 1, we ran a linear regression with urban areas as the dependent variable called model 2. Comparing the coefficients in the summary statistics of model 2, we can see that the "Larger urban population centers" coefficient is negative meaning that this regression model has a negative slope, compared to model 1, in general we can interpret that as living in urban areas with larger populations is associated with a decrease in the number of children that people decide to have compared to model 1. The results in model 2 would be on average, respondents would have 1.5579 children = 2.0494 - 0.4915 (b0 + b1) for those living in urban areas and 2.0494 children for those living in rural areas. The results in both models are the same but the difference in slopes concludes that there is generally a decrease in the number of children people tend to have if they were to live in an urban area compared to those whom lives in rural areas.

Predict the total number of children from marital status

In model 3, we can see that the intercept starts at around 2. This means that on average, for the observations used for the model, households had 2 children. Based on marital status such as single, some estimates of the coefficients were negative.From the p-values that were generated from the model, we can see that a majority of them are small meaning that they are statistically significant (at p=0.05), meaning that it is likely that marital status has a considerable impact on the total number of children.

This result doesn't come as a shock to most. However, it is interesting to take this result and look further into the relationship between marital status and children- specifically, living with a common-law partner. Looking at our model, we know that having a common-law partner does not necessarily have a large impact on the number of children that people have. However, if we investigate further, we see that children that come from common-law unions, are disproportionately affected by divorced parents. As a study by the Canadian Department of Justice, children born to common-law unions were clearly over-represented among children who experienced the break-up of their families (Government of Canada, 2015).

Divorce rates in Canada have been increasing since the 1980s however this is also due to the decrease in marriages that have been happening in Canada recently. In Canada almost half (47 percent) of the children from broken marriages had not seen their parents divorce after three years of separation and this percentage was still 28 percent after five years of separation (Russell, 2017). Thus, when we look at the results of this study, it is important to open the door to further investigations so we as a society can ask more probing questions.

Education levels vs total number of children

When developing a model for investigating the relationship between education levels and the total number of children, we notice that the response for education levels is categorical. Thus, to use a model to model the relationship between these two variables, it is necessary to convert education levels to integer factors to develop a count that we can then graph.

From model 4, figure 14, we can see that the number of total children per person seems to have a linear shape indicating that a linear regression can fit as a potential model. Looking at the p-values of the predictor model created, we notice that it is very small for a university certificate, diploma above a bachelor's degree but not significant at 0.27 (alpha = 0.05). This indicates that more education above a bachelor's degree does not play a significant role in how many children someone has. Additionally, a bachelor's degree or relevant certificate seems to have the smallest p-value and thus seems to be the most significant factor in the number of children had by a person.

Looking at these results, we can use our common knowledge to confirm that this makes sense. Generally speaking, most jobs that require higher education, pay a sustainable wage that allows families to have children and care for them as well. Looking into this further, we can start to see trends between the results in this study and relate them to issues in other areas of our society such as income, and parental leave. One area that isn't discussed in this analysis due to insufficient data is where women and men fall in this. It is perhaps for a future study to dissect this however, a StatCan article states that "one cost is the so-called 'family gap,' also referred to as the 'child penalty' or 'motherhood earnings gap.' It measures how much the earnings of women with children fall below those of women without children, other factors being equal (Earnings of women with and without children 2015).

Income vs number of children

As we interpret model 5, in figure 15, we compare total number of children to income, our results show that the p-values of all the predictor variables are not significant thus not rejecting the null-hypothesis. This means that the income of respondents, no matter how much he/she makes, does not affect the total number of children that they have. The p-values of all the income sections were above the 5% threshold thus is not significant towards the overall regression model. Justified by the model regressing income of the respondent over the total number of children, we can conclude that how much money an individual makes does not help predict how many children he/she will have.

Weaknesses

Since the survey was conducted over the phone, the survey itself, which is quite long, would take a while to conduct. This could have an impact on the respondent's non-response rate and the detail of their answers where they are unlikely to respond if they did not remember. Moreover, the survey has a lot of personal questions that people usually would not like to answer, such as income. Furthermore, open ended questions seemed unavoidable due to the nature of the question and a phone survey. The common presence of open ended questions meant that a lot had to be sorted afterwards. Meaning they are relying on human interpretation, which is not always clear. Finally, although not used in our model, some questions relied for answers of self-interpretation. An example of this would be feelings_life where people answered on a scale of 1-10. In addition, the GSS used a stratified sampling approach in which they separated geographical area in the 10 provinces into 27 strata. Although they divided the areas using judgment on areas that were more populated or rural, the area covered by these strata is still quite wide. Meaning that a small town could still have the chance of being under-represented in comparison to a mid sized town. Moreover, certain demographics of people have preferences towards where they live, which could further influence the accuracy of the sample as a reflection of the population.

Appendix

Code and data supporting this analysis is available at: https://github.com/xuzi9/GSS-2017-Family/blob/main/GSS-2017.Rmd

References

- Alexander, R. and Caetano S. (2020). GSS-Cleaning.rmd. Cleaning of the GSS 2017 survey. https://rohanalexander.com/sta304.html
- Bohnert, N. A. Milan and H. Lathe. (2014) Enduring diversity: Living arrangements of children in Canada over 100 years of the census. Demographic documents, Statistics Canada. Catalogue no. 91-0015 — No. 11.
- Canadian Demographics at a Glance. (2014). Statistics Canada. Catalogue no. 91-003-X.
- Chambers, J. M. (1992). Linear models. Chapter 4 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
- David Robinson, Alex Hayes and Simon Couch (2020). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.
- Earnings of women with and without children. (2015) https://www150.statcan.gc.ca/n1/pub/75-001-x/2009103/article/10823-eng.htm.
- Gee, E.M. (1987). "Historical Change in the Family Life Course of Canadian Men and Women", Aging in Canada: Social Perspectives, editor V.W. Marshall,
  Fitzhenry and Whiteside, Table 4.
- General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. (2020). Retrieved October 19, 2020, from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf
- Government of Canada, D. of J. (2015). Custody, Access and Child Support: Findings from The National Longitudinal Survey of Children and Youth.
  https://www.justice.gc.ca/eng/rp-pr/fl-lf/famil/anlsc-elnej/p3_01.html.
- Kirill Müller (2017). here: A Simpler Way to Find Your Files. https://github.com/krlmlr/here, http://krlmlr.github.io/here.
- Milan, A. (2000). "One hundred years of families", Canadian Social Trends, Statistics Canada. Catalogue no. 11-008.
- Milan, A. (2013). "Fertility: Overview, 2011", Report on the Demographic Situation in Canada, Statistics Canada. Catalogue no. 91-209-X.
- Milan, A. (2013). "Marital status: Overview, 2011", Report on the Demographic Situation in Canada, Statistics Canada. Catalogue no. 91-209-X.
- Milan, A. and L. Martel. (2008). "Fertility and induced abortions, 2011", Part I of Report on the Demographic Situation in Canada, Statistics Canada. Catalogue no. 91-209-X.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Romaniuc, A. (1984). Fertility in Canada: From Baby-boom to Baby-bust. Statistics Canada. Catalogue no. 91-524.
- Russell, A. (2017). Here's why Canadians are having fewer children. https://globalnews.ca/news/3429950/canada-fewer-children-census-216/.
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Wilkinson, G. N. and Rogers, C. E. (1973). Symbolic descriptions of factorial models for analysis of variance. Applied
  Statistics, 22, 392–399. doi: 10.2307/2346786.