Investigation of the Relationship between Number of Marriages and Several Other Variables

Ying Tian (1005174240), Zhaowei Yao (1005333355),

Yuqing Wu (1004725737), Baoying Xuan (1004808149)

October 19, 2020

Abstract

This project aims to investigate the relationship between the number of marriages and some other variables, including feelings of life, age at first marriage, household size, and sex. For data analysis, we used two different models for seeking the relationship among variables, namely the Linear Regression Model, and Logistics Regression Model. The result suggests that no single linear correlation between data is clearly shown in the dataset. However, by conducting the multiple linear regression model and the logistic regression model estimations, the outcomes indicate that age at first marriage is the opposite of the number of marriages, while the other three variables have a positive correlation respectively. In addition, males get married more often than females on average.

Introduction

With the improvement of people's life quality, people have higher demands for life satisfaction. Marital status has also become a topic of great concern, so what is the relationship between the number of marriages and people's first marriage ages, feelings of life, household size, and sex? Will people feel happier if they get married more often? This article from the statistical point of view, to carry on the analysis.

Data

The data of this study is obtained from the 2017 General Social Survey (GSS). We have selected five variables:

- *"feelings_life"* (respondents' feelings of life when taking the interview)
- *"is_male"* (gender of respondents: male or female),
- *"number_marriages"* (number of marriages that the respondents have so far),
- *"hh_size"* (household size of the respondents' family)
- *"age_at_first_marriage"* (respondents' age of their first marriage)

Among these, *"is_male"* is a binary variable. In this study, when the respondent is a male, it is equal to 1; and when the respondent is a female, it is equal to 0.

Moreover, these variables are chosen because our theme is mainly about what factors could be related to the number of marriages. The chosen factors are what we are interested in, and they are also the factors that come to mind for the first time when ordinary people think about this. In this way, we can be more able to resonate with respondents when they do a survey and make people think more deeply.

Strength and limitation

The questionnaire enables a large amount of information to be gathered from many people in a relatively cost-effective manner within a short period of time. There are many ways to distribute so that the operation is more practical, and the results of the questionnaire could usually be quantified quickly and easily by the researchers.

However, the answers are considered inadequate to understand certain forms of information, such as life feelings. People's emotions are complex and very changeable that we cannot simply guarantee that the respondents will be in the same emotional state and feeling that he or she is most of the time.

Also, people may have different interpretations of each question, so their answers would be mostly based on their own understanding of the question, that is, what is "good" for one person may be "bad" for the other, so there is an unrecognized subjectivity.

In addition, in the *"is_male"* variable, the respondents are only divided into males and females, but transgender is not considered. The data obtained from this questionnaire has many missing values, which means that many people skip the question or do not answer. Considering this, we chose the variables with relatively less missing value to conduct the study.

Methodology Discussion

The target population of the 2017 General Social Survey includes all Canadians aged 15 and above, excluding full-time residents of institutions and residents of the Yukon, Northwest Territories, and Nunavut Territories.

They used the Stratified Random Sampling to divide the population into strata based on different locations. Besides being considered separate strata as many Census Metropolitan Area (CMA), the CMA concentrated in Quebec, Ontario, and British Columbia constitutes three strata. The non-CMA areas in each province are also grouped into another ten levels, a total of 27 strata. Then, each of the strata would be sampling by a simple random sample without replacement.

As for the frame of the survey, they used two different approaches. One is that Statistics Canada has a list of telephone numbers in use from various sources such as census and telephone companies. The other one is Address register: A list of all residences in the provinces.

Clearly, it would be a relatively large population. Thus, the workload is still heavy even if they take sampling by groups. Compared with simple random sampling, the organization and analysis of the results are more complicated. In addition, the survey sent by us may not receive a response. In terms of non-response, they try to look for auxiliary information related to the phone number or address for partial incomplete replies. For example, the address register information associated with the phone number has marked household size, and the information not obtained can be retrieved. If it is a complete non-response and no relevant information can be found, such telephone numbers were then dropped.

Model

Two models are generated for our analysis, one is the Multiple Linear Regression and the other is the Logistics Regression.

[Model 1 - Multiple Linear Regression]

```
## 
## Call:
## lm(formula = number_marriages ~ age_at_first_marriage + feelings_lif
e + 
##     hh_size + is_male, data = dataframe)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -1.2691 -0.3514 -0.2174  0.4200  2.7367 
## 
```

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            1.486667   0.047312  31.422  < 2e-16 ***
## age_at_first_marriage -0.022438   0.001358 -16.520  < 2e-16 ***
## feelings_life          0.013404   0.003989   3.361 0.000783 ***
## hh_size                0.153170   0.007427  20.623  < 2e-16 ***
## is_male                0.207281   0.015454  13.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.5179 on 5269 degrees of freedom
## Multiple R-squared:  0.1407, Adjusted R-squared:   0.14 
## F-statistic: 215.7 on 4 and 5269 DF,  p-value: < 2.2e-16
```

[Model 2 - Logistics Regression]

```
## Warning: glm.fit: algorithm did not converge

## 
## Call:
## glm(formula = low_stability ~ age_at_first_marriage + feelings_life 
+ 
##     hh_size + is_male, family = binomial, data = dataframe)
## 
## Deviance Residuals: 
##       Min         1Q     Median         3Q        Max  
## 2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06  
## 
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.657e+01  3.253e+04   0.001    0.999
## age_at_first_marriage  1.078e-11  9.340e+02   0.000    1.000
## feelings_life          1.934e-08  2.743e+03   0.000    1.000
## hh_size               -1.022e-06  5.107e+03   0.000    1.000
## is_male                1.247e-07  1.063e+04   0.000    1.000
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 0.0000e+00  on 5273  degrees of freedom
## Residual deviance: 3.0598e-08  on 5269  degrees of freedom
## AIC: 10
## 
## Number of Fisher Scoring iterations: 25
```

Results

Here are the results with general explanations for our investigation, more deep analysis would be given in the Discussion Part.

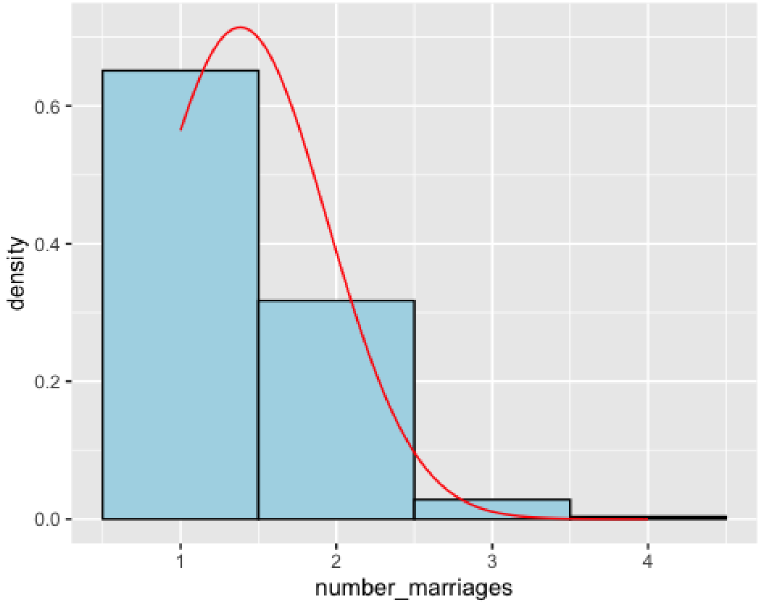

Figure 1: The Histogram of Number of Marriages

The histogram of number of marriages is skewed to the right (i.e. positively skewed). Also, the right tail is much longer, and the mass of the distribution is concentrating on the left of the figure.

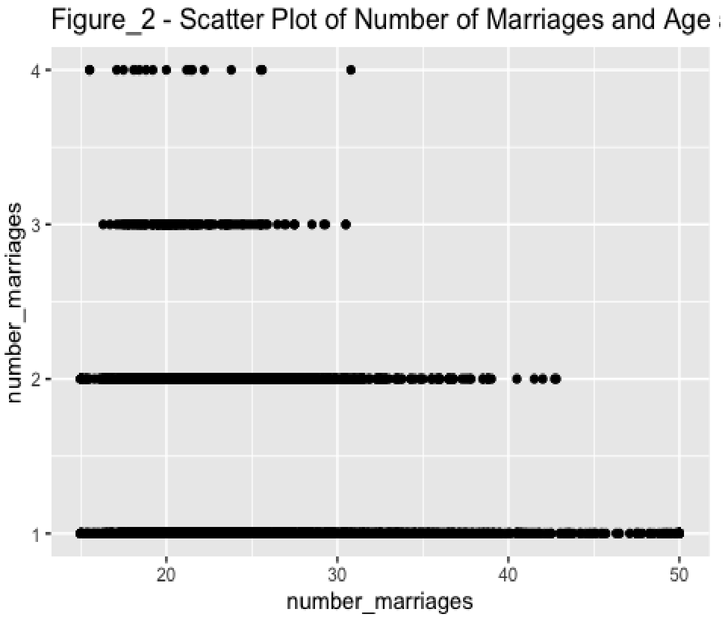

Figure 2: The Scatter Plot of Number of Marriages and Age at First Marriage

The scatter plot of Number of Marriages and Age at First Marriage is shown above, and it can be seen that there is no linear relationship between number of marriages and respondents' age at first marriage.

```
## Call:
## lm(formula = number_marriages ~ age_at_first_marriage + feelings_lif
e +
##     hh_size + is_male, data = dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2691 -0.3514 -0.2174  0.4200  2.7367
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            1.486667   0.047312  31.422  < 2e-16 ***
## age_at_first_marriage -0.022438   0.001358 -16.520  < 2e-16 ***
## feelings_life          0.013404   0.003989   3.361 0.000783 ***
## hh_size                0.153170   0.007427  20.623  < 2e-16 ***
## is_male                0.207281   0.015454  13.413  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5179 on 5269 degrees of freedom
## Multiple R-squared:  0.1407, Adjusted R-squared:   0.14
## F-statistic: 215.7 on 4 and 5269 DF,  p-value: < 2.2e-16

##
## Call:
## glm(formula = low_stability ~ age_at_first_marriage + feelings_life
+
##     hh_size + is_male, family = binomial, data = dataframe)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## 2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            2.657e+01  3.253e+04   0.001    0.999
## age_at_first_marriage  1.078e-11  9.340e+02   0.000    1.000
## feelings_life          1.934e-08  2.743e+03   0.000    1.000
## hh_size               -1.022e-06  5.107e+03   0.000    1.000
## is_male                1.247e-07  1.063e+04   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.0000e+00  on 5273  degrees of freedom
## Residual deviance: 3.0598e-08  on 5269  degrees of freedom
## AIC: 10
##
## Number of Fisher Scoring iterations: 25
```

Table 1: Multiple Linear Regression Model

The above results are from the Multiple Linear Regression Model, where all predictor variables are regular, but only "*is_male*" being an indicator variable, with *"is_male"* = 1 as a male and "*is_male*" = 0 as a female. Besides, the relationship can be expressed as:

$$Num\ of\ Marriages =$$

$$\beta_0\ + \beta_1 \times age\_at\_first\_marriage + \beta_2 \times feelings\_life + \beta_3 \times hh\_size$$

$$+\beta_4 \times is\_male + \epsilon_i$$

Discussion

1.1 [Interpretations - Model 1- Multiple Linear Regression Model]

As shown in Figure 1, the histogram of number of marriages is skewed to the right (i.e. positively skewed), with the right tail being much longer and the mass of the distribution concentrating on the left of the figure, with more than 60% of the total respondents have married once until when completing of the survey.

In the Figure 2 - Scatter Plot of Number of Marriages and Age at First Marriage, no linear relationship is shown between number of marriages and respondents' age at first marriage. However, it is interesting to find out the trend that, people who get married at a younger age are more likely to have more marriages, which is proven by the data points in more than 3 times of marriages locating on the left-side of the plot (i.e. gets married at the age younger than 30).

The first model used here is Multiple Linear Regression Model, where all predictor variables are regular, but only *"is_male"* being an indicator variable, with *"is_male"* = 1 as a male and *"is_male"* = 0 as a female.

$$Num\ of\ Marriages =$$

$$\beta_0\ + \beta_1 \times age\_at\_first\_marriage + \beta_2 \times feelings\_life + \beta_3 \times hh\_size$$

$$+\beta_4 \times is\_male + \epsilon_i$$

The first step to interpret the multiple regression analysis is to examine the F-statistic and the associated p-value in the summary table, at the bottom of table_1. It shows that p-value of the F-statistic is < 2.2e-16(i.e. 2.2*10^(-16)), meaning that it is highly significant, and at least one of the predictor variables is significantly related to the outcome variable.

Secondly, by examining the coefficients table, it shows the estimate of regression beta coefficients and the associated t-statistic as well as p-values.

1.1(a). Betas

<u>[beta 0 hat is 1.484385]:</u> it is the estimated intercept with standard error of 0.047347.

<u>[beta1 hat is -0.02238019]:</u> meaning that holding all other predictor variables - feelings of life, household size(i.e. *"hh_size"*) and gender(i.e. *"is_male"*) unchanged, the

increase of age at first marriage by 1 year, on average, will lead to the number of marriages decrease by 0.02238019.

[beta2_hat is 0.01348722]: meaning that holding all other predictor variables - age at first marriage, household size and gender unchanged, the increase of feelings of life by 1 rating unit, on average, will lead to the number of marriages increase by 0.01348722.

[beta3_hat is 0.15315004]: meaning that holding all other predictor variables - age at first marriage, feelings of life and gender unchanged, the increase of household size by 1 unit（1 more person in household), on average, will lead to the number of marriages will increase by 0.15315004.

[beta4_hat is 0.20742660]: since "*is_male*" is an indicator variable, the coefficient means the difference in average number of marriages between male respondents and females, which is 0.20742660.

1.1(b). Standard Errors

Each estimates of the regression contains relatively small uncertainty and is considered to be good estimates, since all of the corresponding standard errors of the betas are relatively small.

1.1(c). t-statistic & p-values

Assuming the significance level equal to 0.05, we set the hypothesis here for each coefficient as the following:

[Null Hypothesis *H0*]: the corresponding predictor has no linear correlation to the dependent variable (i.e. *"number_marriages"*).

[Alternative Hypothesis *Ha*]: the corresponding predictor does have linear correlation to *"number_marriages"*.

As shown in the table, all of the p-values of the corresponding coefficients are much less than the significance value of 0.05 (as proven by *** on the right of the p-values, meaning that they are all between 0 and 0.0001),thus **reject** the null hypothesis H0 for all the predictors that, the predictor variables stated in the model have no linear correlation to the number of marriages.

1.1(d). Findings & Results

Lastly, we can conclude the result of the estimated model: Since "is_male" is an indicator variable, meaning that when the respondent is male, it is equal to 1; and when the respondent is female, it is equal to 0. Thus, if we are predicting the number of marriages for a female respondent, the estimated multiple linear regression model on the *"number_marriages"* would be:

$$\boldsymbol{For\ Males} : \hat{} Num\ of\ Marriages\mathrm{s}$$

$$= 1.484385\ - 0.022380 \times age\_at\_first\ marriage + 0.013487 \times feelings\_life$$

$$+0.153150 \times hh\_size + 0.207427 \times 1 + \epsilon_i$$

$$\boldsymbol{For\ Females} : \hat{} Num\ of\ Marriages\mathrm{s}$$

$$=\ 1.484385\ - 0.022380 \times age\_at\_first\ marriage + 0.013487 \times feelings\_life$$

$$+0.153150 \times hh\_size + 0.207427 \times 0 + \epsilon_i$$

1.2 [Interpretations - Model2 - Logistic Regression]

For the second model, we are going to interpret the logistic model data by examining Deviance Residuals and coefficients table.

First of all, as it given by the summary of table_2, we notice the Deviance Residuals look good since they are closed to being centered on '0' and roughly symmetrical.

Secondly, we interpret the coefficients table by examining the coefficients table, which shows the estimated regression beta and the associated t-statistic p-values.

1.2(a). Betas

[beta_0_hat is -0.006113]: it shows that the estimated intercept is -0.006113.

[beta1_hat is -0.117605 (i.e. the coefficient of "age_at_first_marriage")]: meaning that holding all other predictor variables - feelings of life, household size(i.e. *"hh_size"*) and gender(i.e. *"is_male"*) unchanged, the increase of age at first marriage by 1 year, on average, will lead to the number of marriages decrease by 0.117605.

[beta2_hat (i.e. "feelings_life") is 0.069055]: meaning that holding all other predictor variables - age at first marriage, household size and gender unchanged, the increase of feelings of life by 1 rating unit, on average, will lead to the number of marriages increase by 0.069055. The p-value is < 2e-16(i.e. 2*10^(-16)), much less than 0.01, thus reject H0 and support Ha that age at first marriage has correlation to the number of marriages.

[beta3 hat (i.e. *"hh_size"*) is 0.73960]: meaning that holding all other predictor variables - age at first marriage, feelings of life and gender unchanged, the increase of household size by 1 unit（1 more person in household), on average, will lead to the number of marriages increase by 0.739606.

[beta4 hat is (i.e. *"is_male"*) 0.892004]: since *"is_male"* is an indicator variable, the coefficient means the difference in average number of marriages between male respondents and females, which is 0.892004. This number indicates that on average, the number of marriages for a male will be 0.892004 times more than that of a female.

1.2(b). Standard Errors

Each estimates of the regression contains relatively small uncertainty and is considered to be good estimates, since all of the corresponding standard errors of the betas are relatively small.

1.2(c). t-statistic & p-values

Assuming the significance level equal to 0.05, we set the hypothesis here for each coefficient as the following:

[Null Hypothesis *H0*]: the corresponding predictor has no correlation to the dependent variable (i.e. *"number_marriages"*).

[Alternative Hypothesis *Ha*]: the corresponding predictor does have correlation to *"number_marriages"*.

As it given by table_2, all p-values are well-below 0.05 and thus, all of the predictors are significantly related to the dependent variable - *"number_marriages"*. Also, there are *** signs right next to all p-values, which also proves that they are all **statistically significant**.

1.2(d). AIC

We have AIC which in this context, is just the Residual Deviance adjusted for the number of parameters in the model.

1.2(e). Number of Fisher Scoring Iterations

We have the number of Fisher Scoring iterations which is 4, telling that how quickly the glm() function converged on the maximum odds estimates for the coefficients.

1.2(f). Findings & Results

Since we have a sample includes both female and male. Then a dummy variable *"in_male"* can be defined as male = 1, for female = 0. Thus, if we are predicting the number of marriages for a male respondent, the estimated logistic regression model on the *"number_marriages"* would be:

$$\boldsymbol{For\ Males} : \hat{}Num\ of\ Marriages$$
$$= -0.006113 - 0.117605 \times age\_at\_first\ marriage + 0.069055 \times feelings\_life$$
$$+0.739606 \times hh\_size + 0.892004 \times 1 + \epsilon_i$$

If we are predicting the number of marriages for a female respondent, then the estimated logistic regression model on the "number_marriages" would be:

$$\boldsymbol{For\ Females} : \hat{}Num\ of\ Marriages$$
$$= -0.006113 - 0.117605 \times age\_at\_first\ marriage + 0.069055 \times feelings\_life$$
$$+0.739606 \times hh\_size + 0.892004 \times 0 + \epsilon_i$$

1.3 [Relation to the Real World]

As stated, the central role of the family plays in people's lives is indisputable. Since families are becoming more and more diverse, the investigation of how to maintain a good and healthy family future is one of the most essential features to keep social stability and development. From the previous analysis, there are positive relationships between the number of marriages and both "feeling of life" and "household size" for both males and females separately. There is also a negative relationship found between the number of marriages and the "age at first marriage".

In the real world, the number of marriages is considered to be an important feature to "measure" the family health. Generally, the family would be healthier and more stable if the number of marriages is low.

However, some of the findings from our model could not accurately demonstrate this association. For example, life's feelings are positively correlated to the number of marriages, which goes against the grain as the families would be healthier when people are satisfied with their lives. This wired finding may come from the bias of the data, and we are going to explain more in the weaknesses and limitations session.

Weaknesses

In general, there are a variety of weaknesses and limitations existing in our investigation. More specific details and thoughts are expressed below:

Data Limitations

The data that we use stated that "the data for the 2017 GSS was collected via computer-assisted telephone interviews (CATI)". Since there would be no supervision for the interviewees while doing the interview, someone may lie on the information they provide, causing the response error for the data collected.

Moreover, for the variable of "feelings of life", the data are collected using a scale of 1 to 10 where 0 means "very dissatisfied" and 10 means "very satisfied" through a question asking how the interviewees feel about their life as a whole at the time of doing the interview. As known, people's thoughts and feelings often change under different circumstances. For example, if someone is taking the interview when he/she just achieve a specific goal and receive a bonus payment, then this person would definitely rate him/herself a higher score.

On the contrary, when someone is taking the interview in a sad mood (ie. due to losing jobs, just arguing with someone else, etc.), a low rating score about feelings of life would be provided to the interviewer. Therefore, it may cause bias for the data collected, which may make the results not precise enough.

Methodological Weaknesses

In terms of methodological weaknesses, our model is mainly based on past results. Generally, situations are changing every year, even every moment. In our investigation, the previous analysis indicates that the number of marriages is positively correlated to “household size” and “feelings of life” but negatively corrected to “age at first marriage” for both males and females in 2017.

One example that the model result may not apply for now is the coronavirus pandemic, one of the worlds’ greatest challenges happening in 2020. There are a variety of articles and reports mentioned that the pandemic would significantly decrease the divorce rate as it makes families more stable. Therefore, the correlation between the number of marriages and variables like age at first marriage may not be what it is supposed to be in 2017. (Collins, 2020)

In addition, it is mentioned that those interviewees who “at first refused to participate would be re-contacted up to two more times”, “for cases in which the timing of the interviewer’s call was inconvenient, an appointment was arranged to call back at a more convenient time” and “for cases in which there was no one home, numerous call backs were made”. Hence, these quotes jointly indicate that the data collecting procedure is complicated and time-consuming, which is one of the weaknesses of the methodological perspective.

Communication Failings

From the communication perspective, our research and investigation are merely in the theoretical stage, where the real-world situations are more complex and might be different from what we found. Therefore, it is required for us to practically do more field studies and surveys to avoid communication failings and get a better understanding of our investigation.

Next Steps

As known, our research and investigation mainly focus on what variables would impact the number of marriages, and we choose only several variables like “age for the first marriage”, “feeling of life” and so on. However, these variables are mainly internal factors, and it is believed that adding some external factors would be beneficial.

For example, economic conditions would be the one that could impact the number of marriages. More specifically, when the economy is booming, the number of marriages would be relatively higher since people’s income is higher, making them economically independent. Thus, people don’t need to rely on the marriage partners much to maintain good living standards, and the number of marriages would generally increase.

Moreover, it is necessary to upgrade the ways of collecting the data to ensure that the researcher could get the data as precise as possible but also save time and money. Since the amount of data after filtering out the missing values is only about 5274, which is too small to represent the whole population. Thus, it is also important to collect more data that could cover all regions to represent the entire population that we are going to investigate. (Looney & Project, 2016)

All in all, social problems of family investigation such as the number of marriages are inextricably related to many other external factors such as the society rules, economic conditions and so on. Thus, there is actually much to desire in its completion method. For now, we have just created and implemented a simple direction of investigating the relationship between the number of marriages and several variables like age of first marriage, household size, etc. and we are looking forward to others exploring more and improving it.

References

1. Alexander, R., &amp; Caetano, S. (2020, October 7). Gss_clean.R.

2. Bache, S. M., &amp; Wickham, H. (2014). Magrittr: A Forward-Pipe Operator for R.

3. Collins, L. (2020, May 08). Is this the end of 'soulmate' marriages? COVID-19 may change relationships forever. Retrieved October 18, 2020, from https://www.deseret.com/indepth/2020/5/7/21246596/covid-19-marriage-stabilize-family-brad-wilcox-eli-finkel-family-studies-coronavirus-divorce

4. Frike, S. (2020). Janitor: Simple Tools for Examining and Cleaning Dirty Data.

5. Greenstone, M. & Looney P. (2020, Feb 03). The Marriage Gap: The Impact of Economic and Technological Change on Marriage Rates. Retrieved from https://www.brookings.edu/blog/jobs/2012/02/03/the-marriage-gap-the-impact-of-economic-and-technological-change-on-marriage-rates/

6. Statistics Canada. (2020, April). General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. Retrieved from https://sda-artsci-utoronto ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

7. Statistics Canada. (2020, April). [General Social Survey on social identity (cycle 27)]. Unpublished raw data.

8. Statistics Canada. (2020, April). 2017 General Social Survey: Families Cycle 31 Public Use Microdata File PUMF. Retrieved from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf

9. Wickham, H. (n.d.). Welcome to the {tidyverse}. Journal of Open Source Software, 4, 1686.

10. Wickham, H., François, R., &amp; Müller, K. (2020). Dplyr: A Grammar of Data Manipulation.