

# Predicting Self-Rated Mental Health Based on Demographic and Family Traits

James Bao, Alan Chen, Xinyi Zhang, Zidong Yang

10/19/2020

The code used to generate this RMarkdown file can be found at [https://github.com/cindyzzhang99/sta304\\_ps3/blob/main/sta304\\_ps3.Rmd](https://github.com/cindyzzhang99/sta304_ps3/blob/main/sta304_ps3.Rmd).

## Abstract

- An abstract is included and appropriately pitched to a general audience.
- The abstract answers: what was done, what was found, and why this matters (all at a high level).
- If your abstract is longer than four sentences then you need to think a lot about whether it is too long. It may be fine (there are always exceptions) but you should probably have a good reason.

## Introduction

- The introduction is self-contained and tells a reader everything they need to know, including putting it into a broader context.
- Your introduction should provide a bit of broader context to motivate the reader, as well as providing a bit more detail about what you're interested in, what you did, what you found, why it's important, etc.
- A reader should be able to read only your introduction and have a good idea about the research that you carried out.
- It would be rare that you would have tables or figures in your introduction (again there are always exceptions but think deeply about whether yours is one).
- It must outline the structure of the report.

## Data

The dataset we used in our modeling is the 2017 General Social Survey (Family cycle). The following sections will discuss how the data was collected, what the key features of the dataset are, and what the data looks like.

### Data Collection

From February 1, 2017 to November 30, 2017, Statistics Canada gathered data on the Canadian family unit by conducting voluntary telephone interviews. Their target population was all non-institutionalized individuals living in Canada, aged 15 or older. Cross-sectional sampling was conducted in a two-stage design.

The stratified simple random sampling method was used in the first stage. Here, the sampling frame consisted of telephone numbers from the Census grouped as households using data from Statistic Canada’s dwelling frame. Strata were formed at the census metropolitan area (CMA) level and at the province level (i.e., large CMAs formed their own stratum, smaller CMAs were grouped together, and the non-CMA regions of each province were grouped together), forming a total of 27 non-overlapping strata. Finally, households were sampled randomly from each stratum such that the number sampled units from each stratum corresponded to the population sizes of each stratum. To reiterate, the sampled population for this first stage was the chosen households from each stratum. The stratified simple random sampling method was also used in the second stage. Here, the sampling frame was a list of household members, aged 15 and older, from the households selected in the first stage. Then, one individual was randomly selected from each household, forming the sampled population. Approximately 43,000 individuals were contacted to participate in the survey.

Overall, the surveying method using two-stage simple random stratified sampling is effective in generating a sample that is geographically representative of the Canadian population. In addition to estimates about the Canadian population at large, the stratified sampling method also allows estimates to be made about subpopulations (at the province level).

Statistics Canada reported that the non-response rate was 47.6%. To reduce the effects of non-response bias, survey responses were adjusted based on the demographic characteristics of households that were non-responsive (by pulling their information from the 2016 Census). This ensures that the discrepancy between the target population and survey responses resulting from non-response is minimized. Furthermore, for the Family cycle of the GSS, responses were also adjusted for income and household size to make more accurate survey estimates for the variables of interest.

Statistics Canada did not disclose the true cost of conducting the survey but we can make some speculations based on the available information about their field work methodology. Surveying was conducted using Computer Assisted Telephone Interviewing (CATI) wherein interviewers read aloud the computerized questionnaire and immediately record the respondent’s answers. Although this allows for a reduction in costs compared to traditional in-person surveying, labor costs still include time spent computerizing the survey, training interviewers, and having interviewers administer the questionnaire. Other labor costs include designing the questionnaire and surveying methodology as well as conducting quality control (data consistency was checked by the CATI system during surveying and unresolved inconsistencies were handled afterwards by support staff). Non-labor costs likely included paying for equipment, phone service, offices, and so forth. Again, although we don’t have exact costs, we can conclude that the time and costs associated with conducting the GSS is a clear reason why it is only administered once a year.

Per the report on the 2017 GSS from Statistics Canada, extensive research and testing was conducted when designing the questionnaire. Consequently, a major strength of the questionnaire is that it contains focused questions that comprehensively and extensively capture the subject of interest (the Canadian family). Upon reading through the questionnaire made available by Statistics Canada, the wording of each question is precise and clear, leaving little room for ambiguity. Additionally, another strength of the survey is that a vast majority of questions were objective (dates, events, counts) removing potential response biases that occur with subjective questions. (Not all questions were objective however, in fact the variable of interest we will model in subsequent sections consists of subjective responses.) On the other hand, because of the specificity of the questions, the survey is very long with several dozens sections and several questions per section. Furthermore, as a result of the large scope of the target population, many questions in the survey did not apply to a large majority of respondents (e.g., number of grandchildren, questions about additional marriages, etc.). The data collected is also incomplete because participants were given the option to refuse to answer or answer “I don’t know” to each question since participation was voluntary.

Overall, the surveying methodology and distributed questionnaire were carefully designed in the interest of collecting accurate, representative data wherever possible.

## Data Characteristics

The full dataset of responses to the 2017 General Social Survey (Family cycle) contains 20,602 observations for over 400 variables relating to the Canadian family. A large reason for our choice to use this dataset is because it is the most recent GSS cycle available for modeling. Other benefits of this dataset have been previously touched upon in the previous section. Namely, the data was checked for consistency in real time by the CATI system (as well as by survey support staff) so there is a certain measure of accuracy that other survey results lack. Additionally, the stratified simple random sampling method used to distribute the survey suggests that the results are representative of the Canadian population to some degree (in the geographical sense at the very least). A major weakness of the data is that it is not complete because of the voluntary nature of the surveying.

In the interest of space, we will only discuss the variables in the dataset that are relevant to our model. The variable we aim to predict is `self_rated_mental_health` while the factors that we chose to inform this prediction are age, sex, marital\_status, and self\_rated\_health. We chose these factors based on the demographic information mentioned in mental health statistics (age, sex, and health) and based on what we suspected might contribute to mental health in the context of family composition (marital status, has children). More explicitly, here are the chosen variables we used from the original dataset:

- `agedc` (renamed to `age`): the exact age of the respondent (in decimals) at the time of the survey
- `sex` (`sex`): sex of the respondent, the options being “Male” or “Female”
- `marstat` (`marital_status`): marital status of the respondent, the options being “Single, never married”, “Married”, “Living common-law”, “Separated” (but still legally married), “Divorced”, or “Widowed”
- `totchdc` (`total_children`): total number of children reported by respondent
- `srh_110` (`self_rated_health`): self-rated physical health, the options being “Excellent”, “Very good”, “Good”, “Fair”, and “Poor”
- `srh_115` (`self_rated_mental_health`): self-rated mental health, the options being “Excellent”, “Very good”, “Good”, “Fair”, and “Poor”

For age, there is a similar variable in the original dataset that uses only natural numbers (`agec`), however, we chose to use `agedc` (and renamed it to “age”) in the interest of accuracy. There are many variables related to marriage in the original dataset (`totunc`: total number of marriage and common-law unions, `nmarevrc`: number of marriages the respondent has had, etc.) but they don’t capture the same scope of information as `marstat` (renamed to “marital\_status”) does (for example, being divorced or widowed is not reflected in those variables). Consequently, we chose `marital_status` as opposed to the other available variables related to marriage. We transformed `totchdc` (renamed to “total\_children”) into a binary variable `has_children` (if total children is 0, `has_children` equals FALSE, if total children is greater than 0, `has_children` equals TRUE). For the other three variables we use (`sex`, `self_rated_health` or `srh_110` in the original dataset, and `self_rated_mental_health` or `srh_115` in the original dataset), there are no similar equivalents.

We will refer to the variables by their renamed identifiers for the rest of the paper.

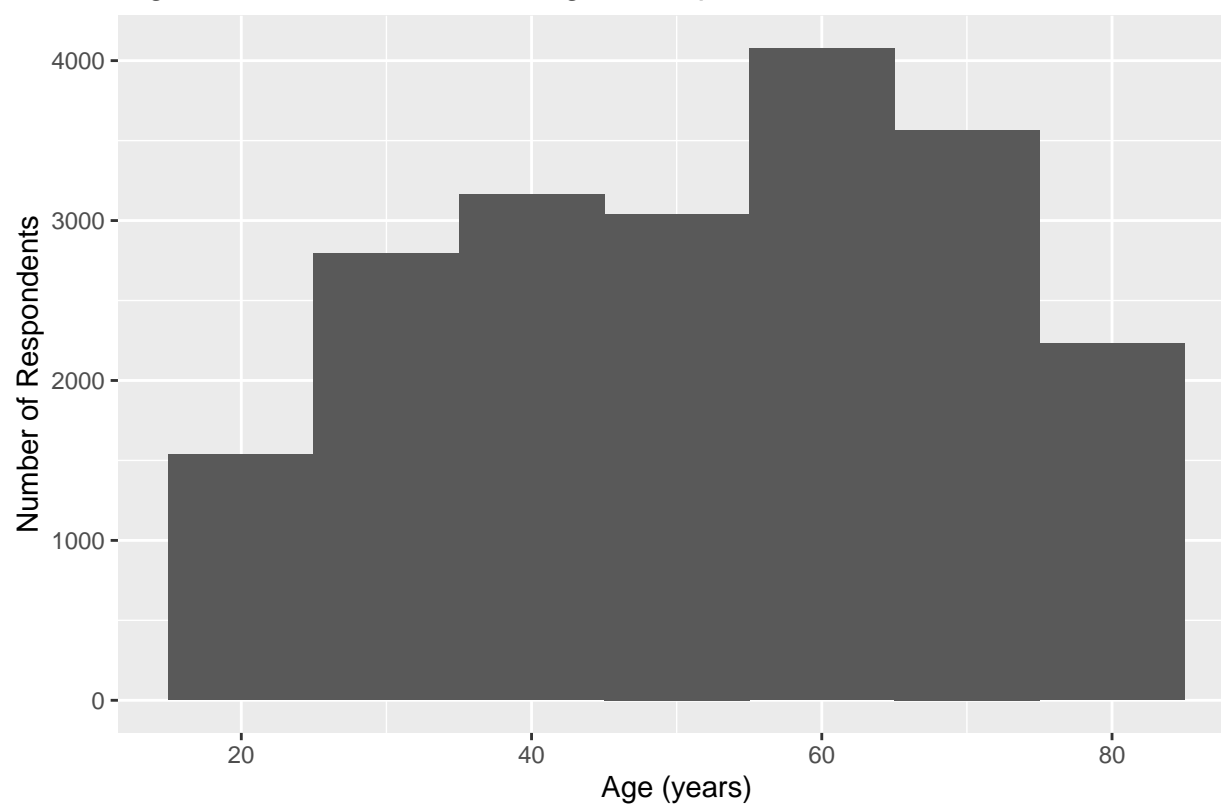
## Data Visualization

Specific instructions on how to download the 2017 GSS dataset can be found in the header of the `gss_cleaning.R` file found here ([https://github.com/cindyzhang99/sta304\\_ps3/blob/main/gss\\_cleaning/gss\\_cleaning.R](https://github.com/cindyzhang99/sta304_ps3/blob/main/gss_cleaning/gss_cleaning.R)).

We cleaned the original 2017 GSS dataset using `gss_cleaning.R`, producing `gss_cleaned.csv` (with variables renamed accordingly).

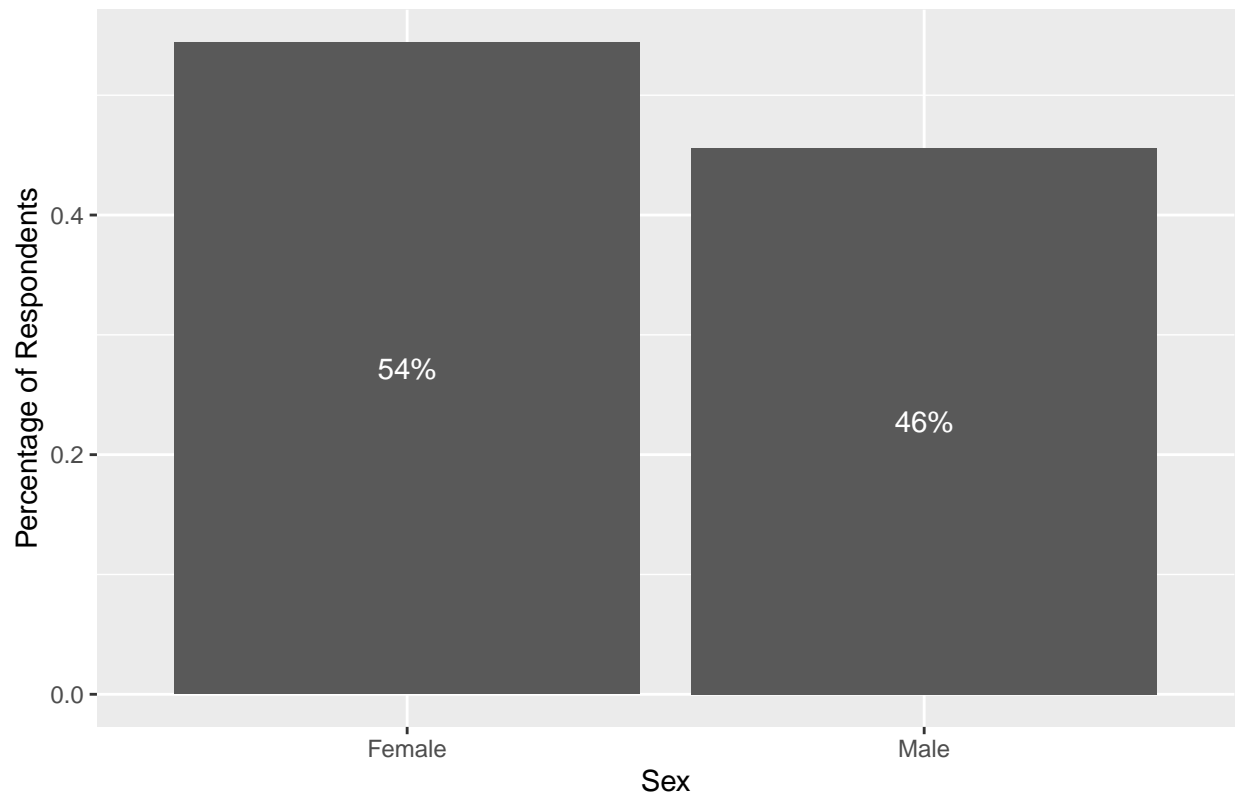
##	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	15.00	37.30	54.20	52.16	66.70	80.00

Figure 1: Distribution of the age of respondents.



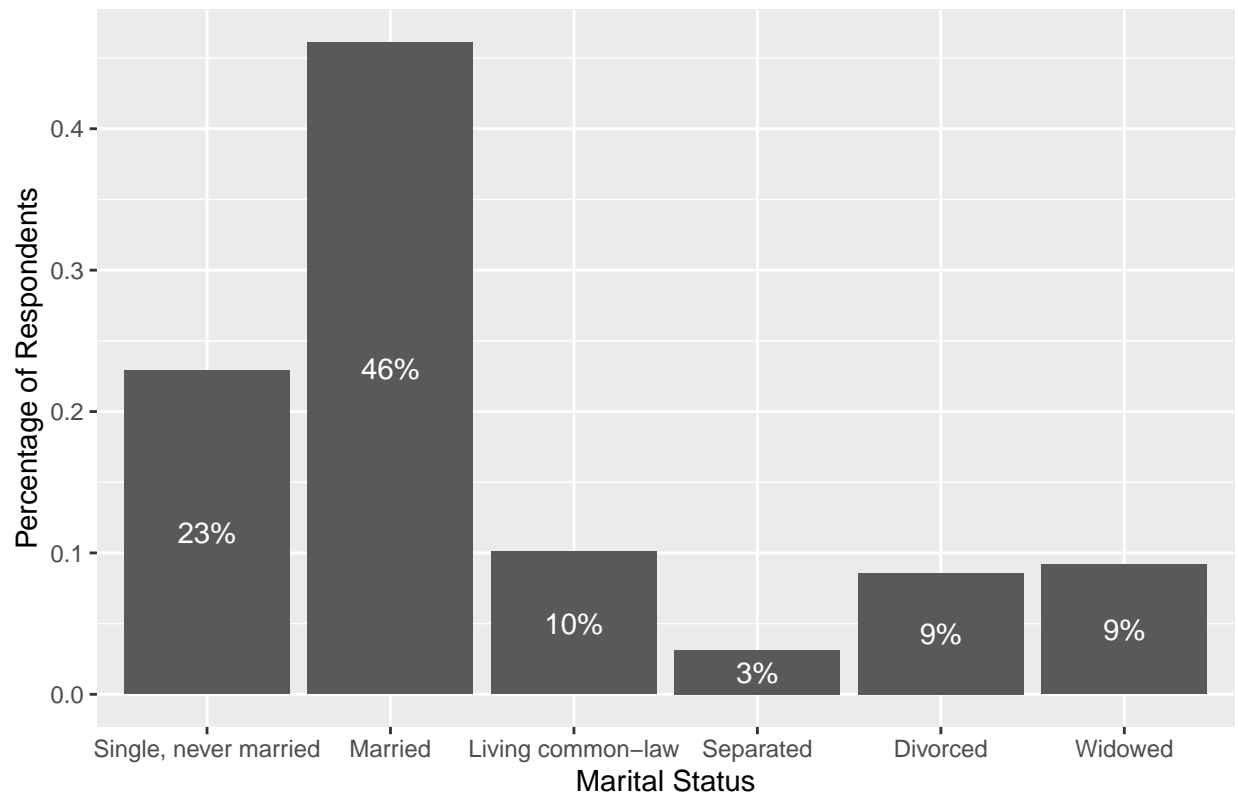
```
##      sex count
## 1 Female 11105
## 2  Male  9303
```

Figure 2: Distribution of the sex of respondents in percentages.



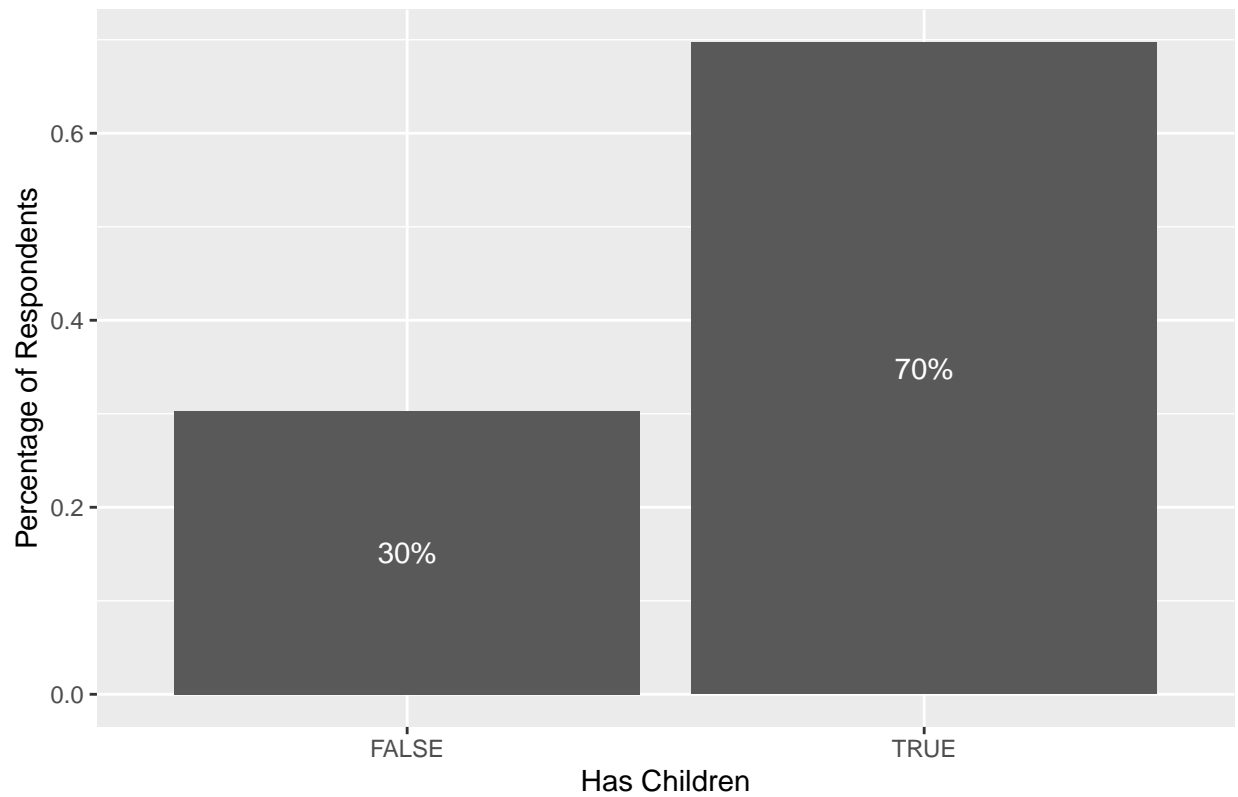
```
##      marital_status count
## 1      Divorced    1749
## 2  Living common-law 2059
## 3      Married    9411
## 4      Separated     639
## 5 Single, never married 4674
## 6      Widowed    1876
```

Figure 3: Distribution of the marital status of respondents in percentages.



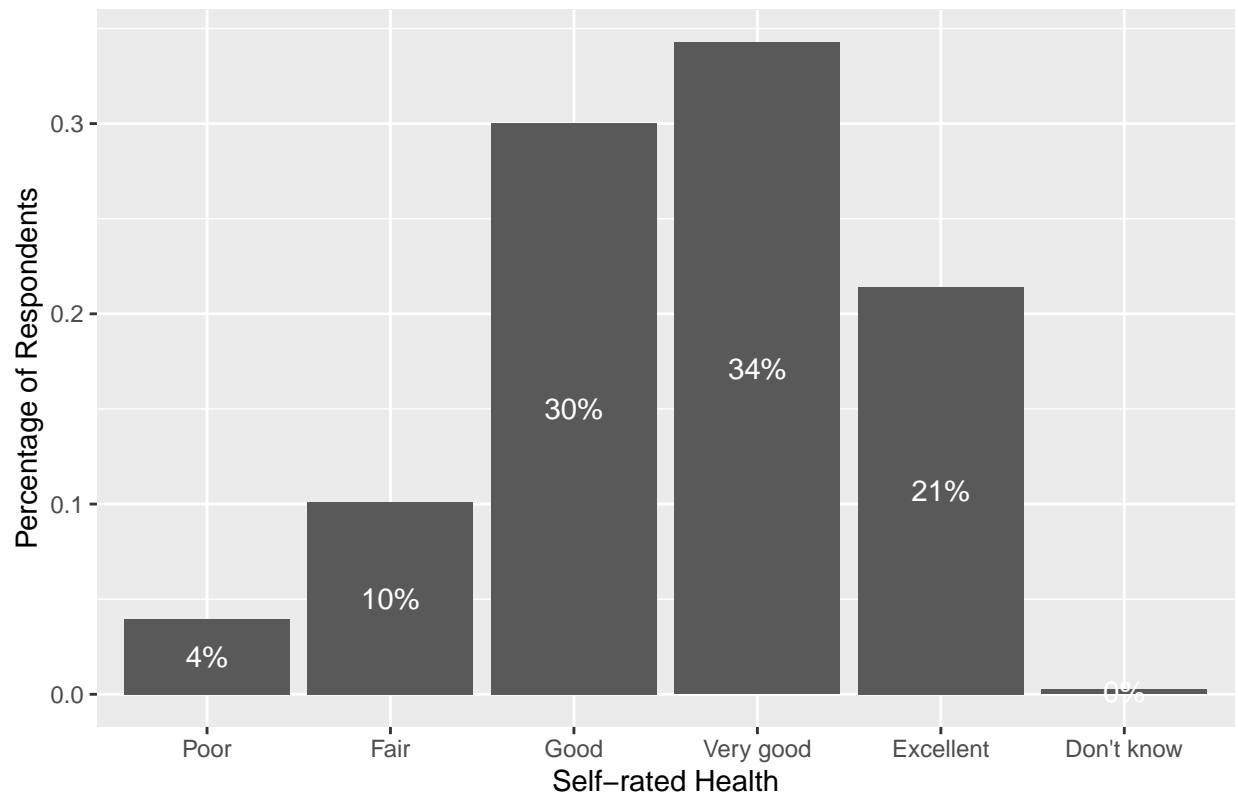
```
## has_children count
## 1 FALSE 6179
## 2 TRUE 14229
```

Figure 4: Distribution of respondents with and without children.



```
##      health count
## 1 Don't know    51
## 2 Excellent  4369
## 3      Fair   2062
## 4      Good   6126
## 5      Poor    806
## 6 Very good  6994
```

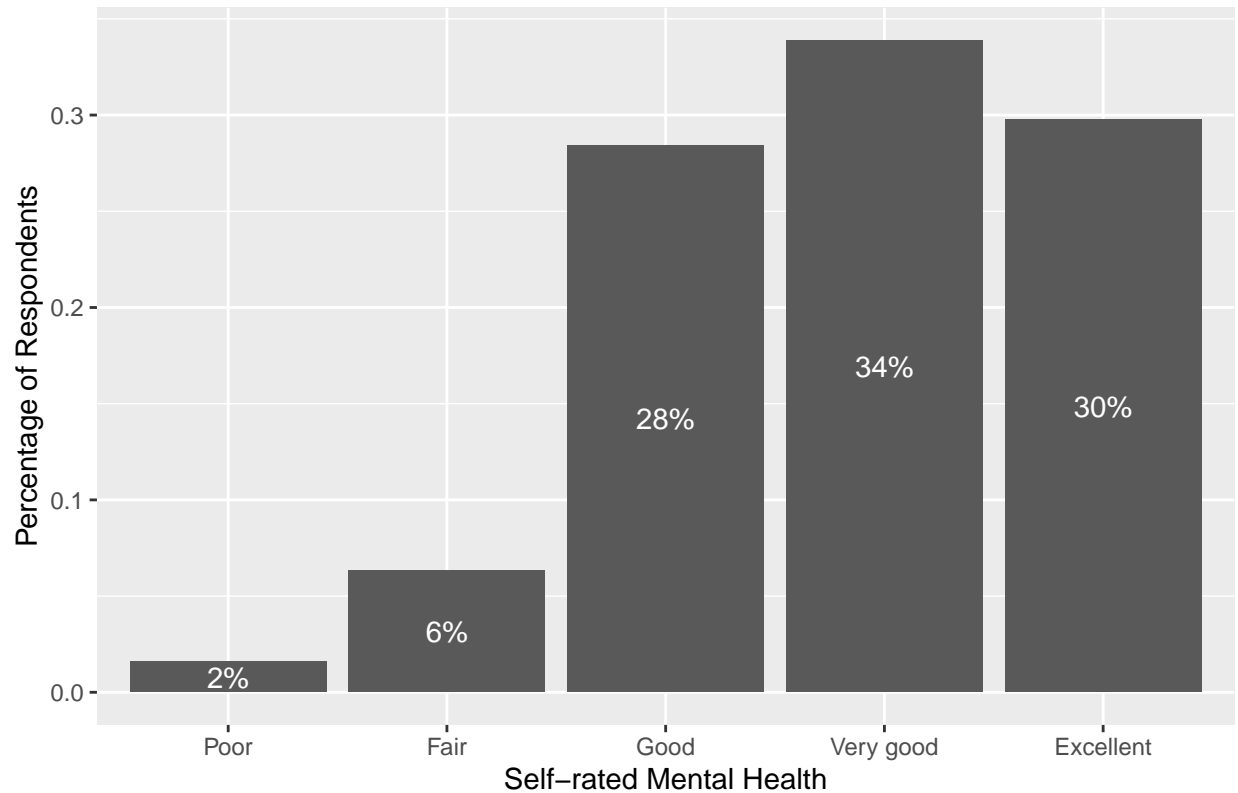
Figure 5: Distribution of the self-rated health of respondents in percentages



```
##   mental_health count
## 1    Excellent  6073
## 2     Fair    1294
## 3     Good    5801
## 4     Poor     325
## 5   Very good  6915
```



Figure 6: Self-rated mental health of respondents in percentages.



Out of the respondents, approximately 54% were female and 46% were male (Figure 2). Most of the respondents were in their 60's and 70's, while the next most common demographic were respondents in their 40's and 50's (Figure 1). This preliminary look at the dataset is fairly consistent with Canadian demographics according to the 2016 Census, with the female response being approximately 3% higher than expected and the average age being approximately 11 years older than expected (the average Canadian age is 41 while the average respondent age was 52). This older demographic makes sense as individuals less than 15 years of age were not eligible to respond to the survey and are therefore not represented here.

Of the 20,602 responses, 194 rows were dropped if the value for a variable was not available. For our purposes of attempting to model mental health, we consequently removed these individuals from the dataset we used in generating our model. Furthermore, according to Figure 5, the responses are heavily skewed towards positive responses, with 30% of respondents replying with 'Excellent' and 34% replying 'Very good'. 28% rated their mental health as 'Good' with the remaining 8% split 6 to 2 with regards to 'Fair' and 'Poor', respectively. These results overwhelmingly indicate that a large proportion of the sampled population feel that their mental is very strong. However, we proceed with modeling in the next section of this paper to better understand the contribution of the chosen demographic and family factors on self-rated mental. Is there a pattern of traits that separate "Excellent", "Very good", and "Good" ratings? What are the biggest distinctions between an individual with good mental health and poor mental health? These are some of the motivating questions we strive to answer with our model.

## Model

The purpose of the model is to predict a person's self-rated mental health based on our selected factors of age, sex, marital status, and self-rated physical health. Since self-rated mental health is a categorical data type in this dataset, the task at its core is a classification problem.

## Model Selection

Some models that we considered were linear regression, naive Bayes, binary logistic regression, and multinomial logistic regression. To begin, linear regression is not suitable because it is often difficult to find an accurate linear relationship between predictors and categories, not to mention the fact that linear regression is more suitable when the dependent variable is continuous. Naive Bayes is a viable option since it is able to handle classification of more than two categories using joint probability and Bayes' Theorem. However, naive Bayes only works well under the assumption that the explanatory variables are independent, but this is often not the case. Given the context of the data, it is highly anticipated that the characteristics of a person are correlated in some way or another (e.g., age showing a correlation with self-rated physical health).

In addition, generative models (e.g., Naive Bayes) have a higher asymptotic error than discriminative models (e.g., logistic regression), but they approach the asymptotic error faster. In other words, discriminative models tend to perform better given a large enough dataset while generative models will perform better on small dataset as they learn faster. Since the dataset is large, choosing a discriminative model would be more appropriate for this task.

It is also important to note that the dependent variable has more than two categories. One way to handle this is to group multiple categories together so that there are only two categories. Then we would be able to use binary logistic regression to model the relationship. While this simplifies the complexity of implementing the model itself, there will be a loss in information from merging classes, impacting the strength of the conclusions we can draw. Another option to handle multi-classification is to use multinomial logistic regression which is an extension of binary logistic regression. The basic idea of this approach is to create a binary logistic regression model for each class. Each binary logistic regression model will do a one-versus-rest prediction for the corresponding class. The class with the highest probability will be the output prediction of the multinomial model. One of the caveats of this approach is that even more data is required to provide enough information for all binary logistic regression models for each class; otherwise, this method is prone to overfitting. This is less of a concern in this case because the size of the dataset is large enough to predict 5 classes. Thus, we chose multinomial logistic regression as our model for this task.

As described in the Data Characteristics section of this report, we chose four features (age, sex, marital\_status, and self Rated Health) based on their perceived impact on self-rated mental health (self Rated Mental Health). We chose age as a continuous factor as opposed to the discrete version (where age is only given in whole numbers) in the interest of accuracy and being able to provide the model with the most fine-grained data. Sex and marital status are categorical variables due to the nature of the information they provide (you're either single, married, in a common-law partnership, separated, divorced, or widowed, there is no status that exists in-between any of these options). Although self-rated health and self-rated mental health exists in real life on a continuous scale, due to the nature of responses gathered by the 2017 GSS, self Rated Health and self Rated Mental Health are categorical variables with possible values such as Poor, Fair, Good, Very good, and Excellent. There is no way for us to transform the categorical responses into a continuous variable that remains true to the data collected and there isn't any need to do so because we adjusted the selection of our model accordingly.

## Mathematical Model

We fully derive the mathematical representation of the multinomial logistic regression model in the Appendix. In the interest of space, here we will just provide the equations we derived for the probability of each outcome:

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot X_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot X_i}}$$
$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot X_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot X_i}}$$

$$\Pr(Y_i = 3) = \frac{e^{\beta_3 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 4) = \frac{e^{\beta_4 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

We will explain each variable that appears in the equations above (again, please see the Appendix for a more natural definition of each variable as we derive these equations).  $Y_i$  represents the outcome of the response variable for the  $i$ th observation. Since we have a total of 5 possible outcomes (Poor, Fair, Good, Very Good, and Excellent), we represent each of them as 1, 2, 3, 4, and 5 in the interest of space.

$\beta_k$  is the row vector with elements that are the coefficients for the explanatory variables for the  $k$ th outcome. More explicitly,  $\beta_k = [\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}]$  where  $\beta_{0,k}$  is the intercept for the  $k$ th outcome,  $\beta_{1,k}$  is the coefficient for the first explanatory variable (age) for the  $k$ th outcome,  $\beta_{2,k}$  is the coefficient for the second explanatory variable (sex) for the  $k$ th outcome,  $\beta_{3,k}$  is the coefficient for the third explanatory variable (marital\_status) for the  $k$ th outcome,  $\beta_{4,k}$  is the coefficient for the fourth explanatory variable (has\_children) for the  $k$ th outcome, and  $\beta_{5,k}$  is the coefficient for the fifth explanatory variable (self-rated\_health) for the  $k$ th outcome.

$\mathbf{x}_i$  is the row vector of explanatory variables for the  $i$ th observation. Specifically,  $\mathbf{x}_i = [1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}]$  where  $x_{1,i}$  is the value of the first explanatory variable (age) for the  $i$ th observation,  $x_{2,i}$  is the value of the second explanatory variable (sex) for the  $i$ th observation, and so on.

Therefore, the first equation represents the probability that the self-rated\_mental\_health of the  $i$ th observation is Poor given values for age, sex, marital\_status, has\_children, and self-rated\_health. The other 4 equations can be interpreted in a similar manner.

## Running the Model

To estimate the model, we use the multinom function from the library nnet written in the programming language R. Our script is run using the software RStudio.

```
# install.packages("nnet")
library(nnet)
# install.packages("Metrics")
library(Metrics)

# shuffle rows of data tibble for cross-validation
# for reproducibility
set.seed(42)
# generate randomized indices for rows
shuffled_indices <- sample(total_count)
# shuffle data tibble accordingly
data <- data[shuffled_indices,]
# set boundary between training and testing
boundary = as.integer(total_count * 0.8)
# take subset of dataset to form training dataset
training_dataset = data[0:boundary,]

# run model where self-rated_mental_health is response variable and age, sex,
# marital status, has children, and self rated health are explanatory variables
model <- nnet::multinom(self-rated_mental_health ~ age + sex + marital_status + has_children +
  self-rated_health, data = training_dataset)
```

```
## # weights: 75 (56 variable)
## initial value 26275.683358
## iter 10 value 19588.463000
## iter 20 value 19283.841204
## iter 30 value 18586.312225
## iter 40 value 18332.871105
## iter 50 value 18246.392191
## iter 60 value 18182.094276
## final value 18179.462642
## converged
```

```
#summary(model)
#head(fitted(model))
```

Our model successfully converges after 60 iterations on the training dataset to a final negative log-likelihood value of 18179. Although this is a really large number, this is not too bad considering we have over 16300 observations in the training dataset and the negative log-likelihood is equal to the likelihood we would observe the specific dataset (after taking the log and multiplying by -1).

```
# take subset of dataset to form testing dataset
testing_dataset = data[boundary:total_count,]
testing_dataset = testing_dataset[complete.cases(testing_dataset), ]
# predict what is the likelihood of each outcome for each observation in the testing dataset
testing_probabilities <- predict(model, newdata = testing_dataset, "probs")

# identify the predicted outcome based on which probability is the largest
# find the column index with the largest probability
column_index <- max.col(testing_probabilities, tie="random")
# form a list of outcomes with column names for testing dataset
testing_predictions <- colnames(testing_probabilities)[column_index]
testing_ground_truth <- testing_dataset %>% pull(self_rated_mental_health)
accuracy(testing_ground_truth, testing_predictions)
```

```
## [1] 0.5486162
```

We performed a check on the model by conducting cross validation. We split the shuffled dataset into training and testing subsets (an 80-20 split, respectively). Then, we trained the model using only the training data and made predictions on the testing data. Finally, we compared the model's predictions to the ground truth values for self-rated mental health for the test set. Our predictions for the test set had an accuracy of 54.9%. Compared to random predictions which would have an accuracy of only 20% (since there are five possible outcomes for the response variable), our model performs at least twice as well. Furthermore, this accuracy is calculated based on exact matches between predicted and ground truth values. If we considered predictions of Very good to be close enough to ground truth values of Excellent, the accuracy of our model will be much, much higher.

```
map <- data.frame(find=c('Excellent', 'Very good', 'Good', 'Fair', 'Poor'),replace=c('Excellent', 'Excellent', 'Very good', 'Good', 'Fair', 'Poor'))
testing_ground_truth_simplified <- as.character(map[match(testing_ground_truth, map$find), "replace"])
testing_predictions_simplified <- as.character(map[match(testing_predictions, map$find), "replace"])
accuracy(testing_ground_truth_simplified, testing_predictions_simplified)
```

```
## [1] 0.671565
```

We gain a sense of the relative accuracy of our model by mapping Very good values to Excellent values and Fair values to Poor values. This has a 67.2% accuracy which again suggests the model has identified relationships between the explanatory and response variables.

Consequently, we can proceed to the results section satisfied our model has a considerable prediction power.

## Results

## Discussion

Predicting mental health in the context of the nuclear family. Does being married have a positive or negative impact?

## Weaknesses

## Next Steps

## Appendix

### Derivation of the mathematical representation of the multinomial logistic regression model

Recall that the multinomial logistic regression consists of several binary logistic regression models. Like binary logistic regression, multinomial logistic regression predicts the probability that the  $i$ th observation has outcome  $k$  using the following function:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \dots + \beta_{M,k}x_{M,i}$$

where  $\beta_{m,k}$  is the coefficient for the  $m$ th explanatory variable and the  $k$ th outcome while  $x_{m,i}$  is the value of the  $m$ th explanatory variable for the  $i$ th observation. In our case, we have  $M = 5$  (age, sex, marital\_status, has\_children self Rated health) explanatory variables so the function as applicable to our model is:

$$f(k, i) = \beta_{0,k} + \beta_{1,k}x_{1,i} + \beta_{2,k}x_{2,i} + \beta_{3,k}x_{3,i} + \beta_{4,k}x_{4,i} + \beta_{5,k}x_{5,i}$$

Note that we can represent  $\beta_{0,k}, \beta_{1,k}, \beta_{2,k}, \beta_{3,k}, \beta_{4,k}, \beta_{5,k}$  and  $1, x_{1,i}, x_{2,i}, x_{3,i}, x_{4,i}, x_{5,i}$  as row vectors  $\beta_k$  and  $\mathbf{x}_i$ , respectively. Then, the function can be simplified as follows:

$$f(k, i) = \beta_k \cdot \mathbf{x}_i$$

where we take the dot product of the two row vectors we just defined.

As previously mentioned, the multinomial logistic regression model is a series of binary logistic regressions where the probability of each outcome of the response variable (self-rated mental health) is regressed against a chosen pivot outcome. Let  $Y_i$  represent the outcome of the response variable for the  $i$ th observation. We have a total of 5 possible outcomes (Poor, Fair, Good, Very Good, and Excellent represented as 1, 2, 3, 4, and 5, respectively). Let's choose the last outcome (Excellent or 5) as the pivot. In mathematical notation, this is:

$$\ln \frac{\Pr(Y_i = 1)}{\Pr(Y_i = 5)} = \beta_1 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 2)}{\Pr(Y_i = 5)} = \beta_2 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 3)}{\Pr(Y_i = 5)} = \beta_3 \cdot \mathbf{X}_i$$

$$\ln \frac{\Pr(Y_i = 4)}{\Pr(Y_i = 5)} = \beta_4 \cdot \mathbf{X}_i$$

Then, we solve for the probabilities by exponentiating both sides:

$$\Pr(Y_i = 1) = \Pr(Y_i = 5) \cdot e^{\beta_1 \cdot \mathbf{X}_i}$$

$$\Pr(Y_i = 2) = \Pr(Y_i = 5) \cdot e^{\beta_2 \cdot \mathbf{X}_i}$$

$$\Pr(Y_i = 3) = \Pr(Y_i = 5) \cdot e^{\beta_3 \cdot \mathbf{X}_i}$$

$$\Pr(Y_i = 4) = \Pr(Y_i = 5) \cdot e^{\beta_4 \cdot \mathbf{X}_i}$$

The probability of the pivot outcome can be calculated because we know that the probability of all outcomes must sum to 1:

$$\Pr(Y_i = 5) = 1 - (\Pr(Y_i = 5) \cdot e^{\beta_1 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_2 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_3 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_4 \cdot \mathbf{X}_i})$$

$$1 = \Pr(Y_i = 5) + \Pr(Y_i = 5) \cdot e^{\beta_1 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_2 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_3 \cdot \mathbf{X}_i} + \Pr(Y_i = 5) \cdot e^{\beta_4 \cdot \mathbf{X}_i}$$

$$1 = \Pr(Y_i = 5) (1 + e^{\beta_1 \cdot \mathbf{X}_i} + e^{\beta_2 \cdot \mathbf{X}_i} + e^{\beta_3 \cdot \mathbf{X}_i} + e^{\beta_4 \cdot \mathbf{X}_i})$$

$$1 = \Pr(Y_i = 5) \left( 1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i} \right)$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

Having the expression for  $\Pr(Y_i = 4)$ , we can represent the probabilities of the other outcomes as follows:

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 3) = \frac{e^{\beta_3 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 4) = \frac{e^{\beta_4 \cdot \mathbf{X}_i}}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

$$\Pr(Y_i = 5) = \frac{1}{1 + \sum_{k=1}^4 e^{\beta_k \cdot \mathbf{X}_i}}$$

## References

Interview method/survey size: <https://www.statcan.gc.ca/eng/survey/household/4501> Detailed information about GSS 2017: <https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816> Questionnaire: [https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item\\_Id=335815#qb345205](https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=assembleInstr&lang=en&Item_Id=335815#qb345205) Mental health statistics: <https://www.camh.ca/en/Driving-Change/The-Crisis-is-Real/Mental-Health-Statistics> Discriminative vs Generative Classifiers (Naive Bayes vs logistic regression): <https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf> Age and sex <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/as/Table.cfm?Lang=E&T=21>