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/cindyzhang99/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). More explicitly, here is what information the chosen variables in the dataset represent:

- age (agedc in the original dataset): 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"
- marital_status (marstat): marital status of the respondent, the options being "Single, never married", "Married", "Living common-law", "Separated" (but still legally married), "Divorced", or "Widowed"
- self_rated_health (srh_110): self-rated physical health, the options being "Excellent", "Very good", "Good", "Fair", and "Poor"
- self_rated_mental_health (srh_115): 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 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. 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.

Data Visualization

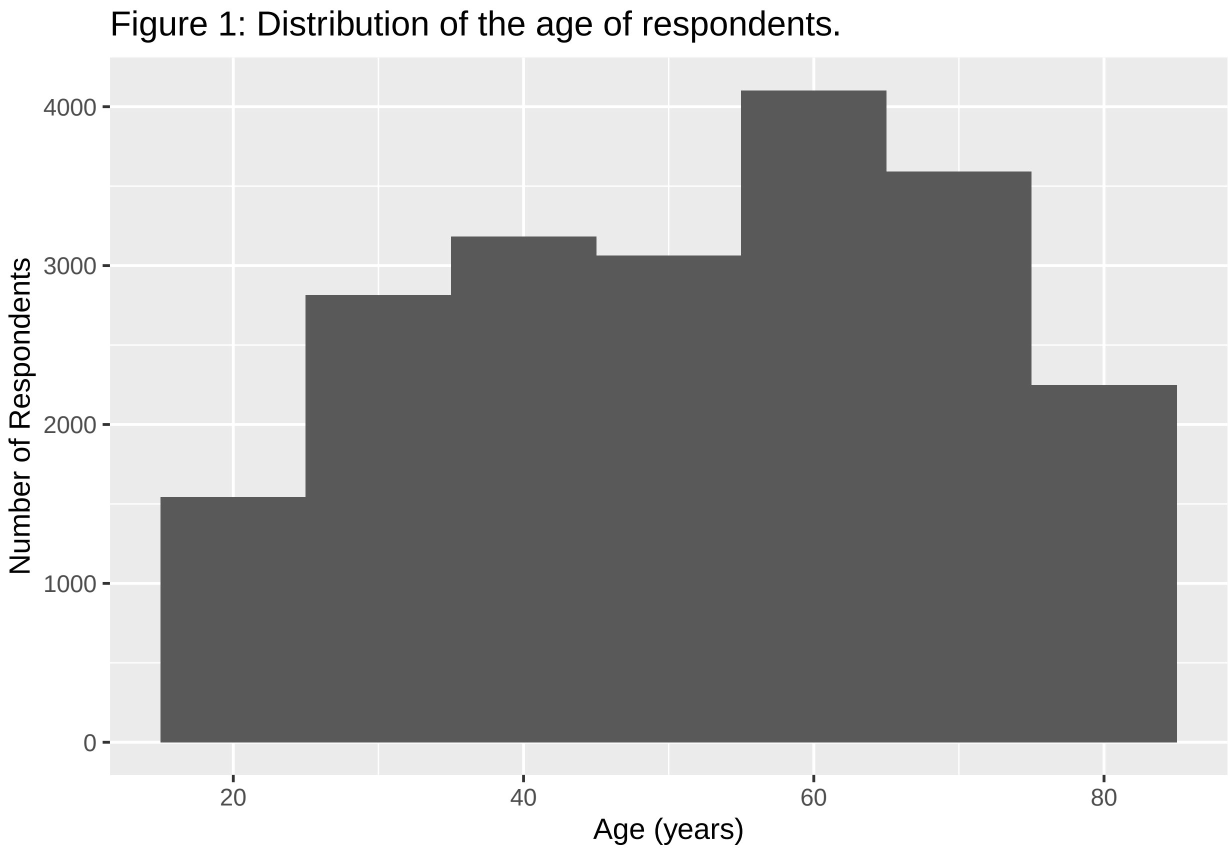
Specific instructions on how to download the 2017 GSS dataset can be found in the header of the gss_cleaning.R file found here (https://github.com/cindyzhang99/sta304_ps3/blob/main/gss_cleaning/gss_cleaning.R).

We cleaned the original 2017 GSS dataset using gss_cleaning.R, producing gss_cleaned.csv.

```r
library(tidyverse)

# load the cleaned dataset
data <- read_csv("gss_cleaned.csv")

# choose pertinent variables
data <- data %>% select(age, sex, marital_status, self_rated_health,
                        self_rated_mental_health)

# choose subset of data with valid mental health self-rating
data<-data[!grepl("Don't know", data$self_rated_mental_health),]

# view distributions of the variables of interest
# summary table with statistics for age
summary(data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   37.30   54.20   52.17   66.70   80.00
```

```
# distribution of age as a histogram
ggplot(data, aes(x=age)) + geom_histogram(position="identity", binwidth = 10) +
  xlab("Age (years)") +
  ylab("Number of Respondents") +
  ggtitle("Figure 1: Distribution of the age of respondents.")
```

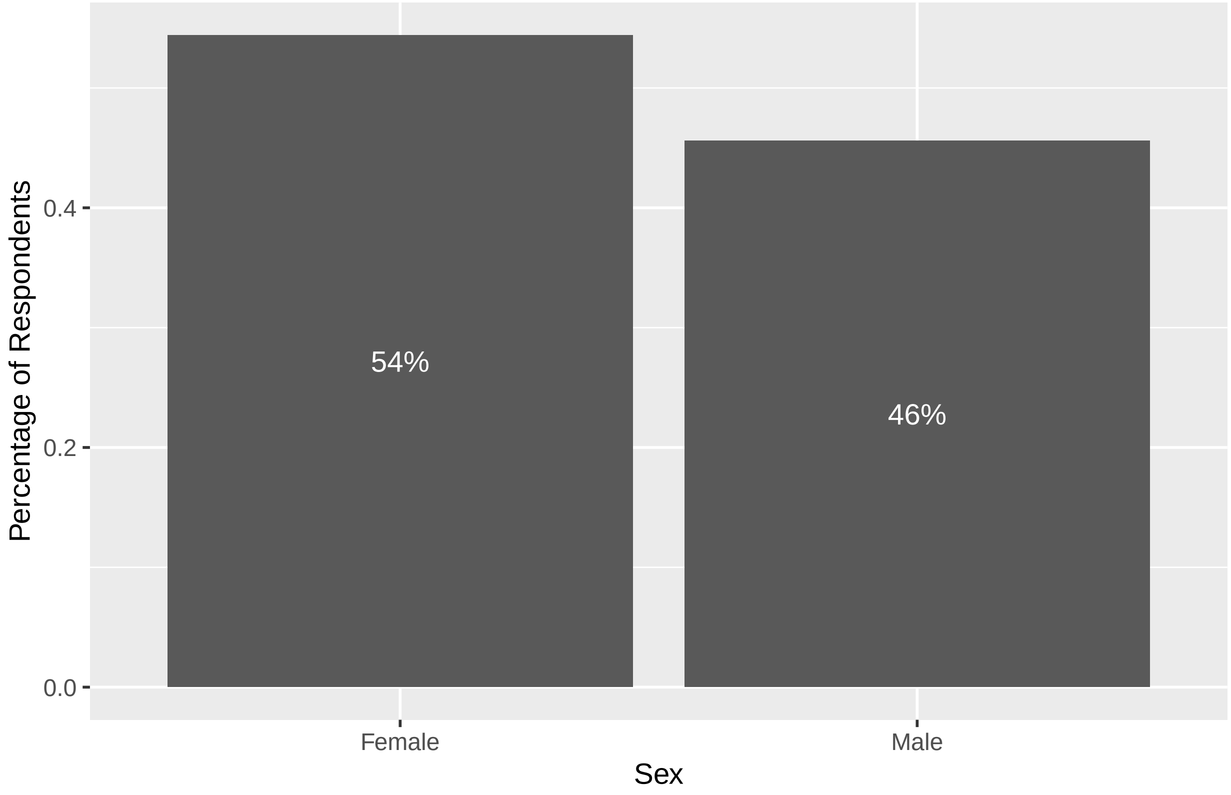


```
# total number of observations in the dataset
total_count = nrow(data)

# summary table with counts for sex
tibble_sex <- data.frame(table(data$sex)) %>%
  rename(
    sex = Var1,
    count = Freq
  )
tibble_sex
```

```
##      sex count
## 1 Female 11177
## 2   Male  9368
```

```
# distribution of sex as a bar graph in percentages
ggplot(tibble_sex, aes(x = sex, y = count/total_count)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(count/total_count*100), "%")),
            color="white",
            position = position_stack(vjust = 0.5)) +
  xlab("Sex") +
  ylab("Percentage of Respondents") +
  ggtitle("Figure 2: Distribution of the sex of respondents in percentages.")
```

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

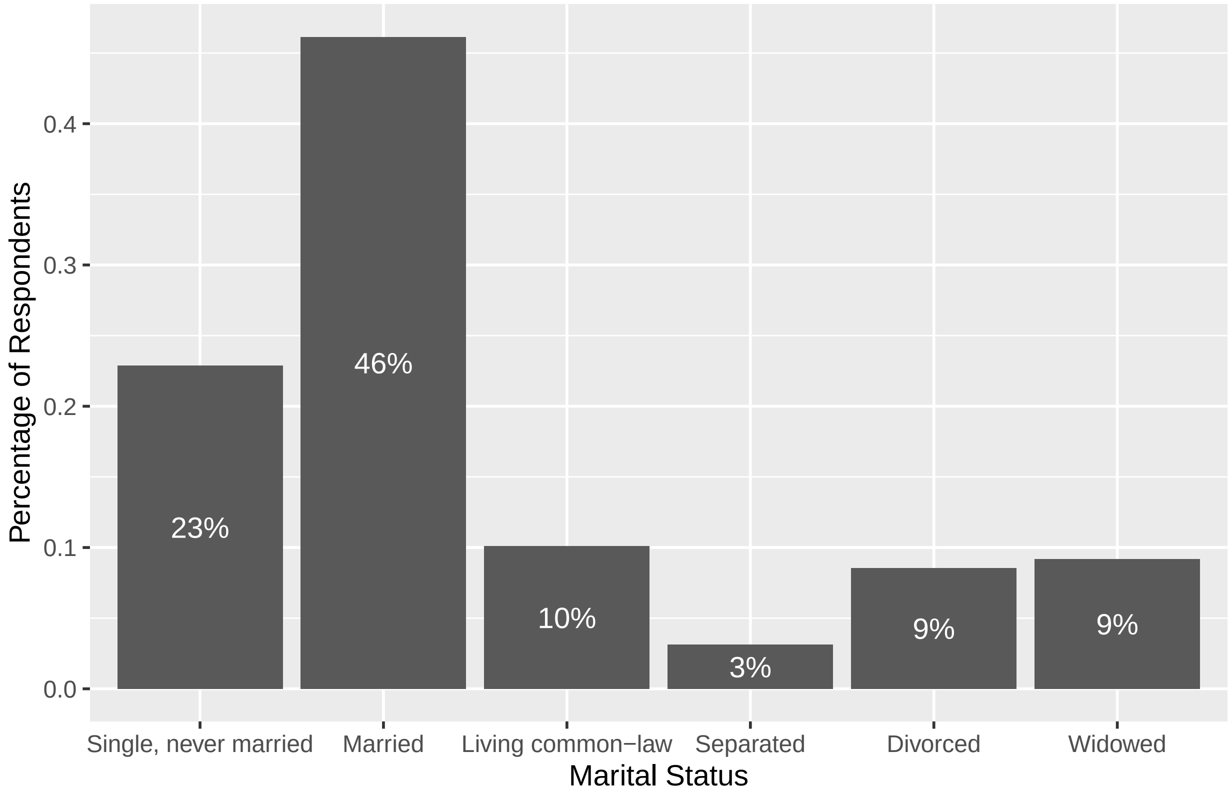


```
# summary table with counts for marital status
tibble_marital_status <- data.frame(table(data$marital_status)) %>%
  rename(
    marital_status = Var1,
    count = Freq
  )
# manually changing the order of factors in the following output table and graph
tibble_marital_status$marital_status <- factor(tibble_marital_status$marital_status,
                                               levels = c("Single, never married",
                                                          "Married", "Living common-law",
                                                          "Separated", "Divorced",
                                                          "Widowed"))
tibble_marital_status
```

```
##          marital_status count
## 1              Divorced  1759
## 2     Living common-law  2073
## 3               Married  9481
## 4             Separated   641
## 5 Single, never married  4698
## 6               Widowed  1887
```

```
# distribution of marital status as a bar graph in percentages
ggplot(tibble_marital_status, aes(x = marital_status, y = count/total_count)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(count/total_count*100), "%")),
            color="white",
            position = position_stack(vjust = 0.5)) +
  xlab("Marital Status") +
  ylab("Percentage of Respondents") +
  ggtitle("Figure 3: Distribution of the marital status of respondents in percentages.")
```

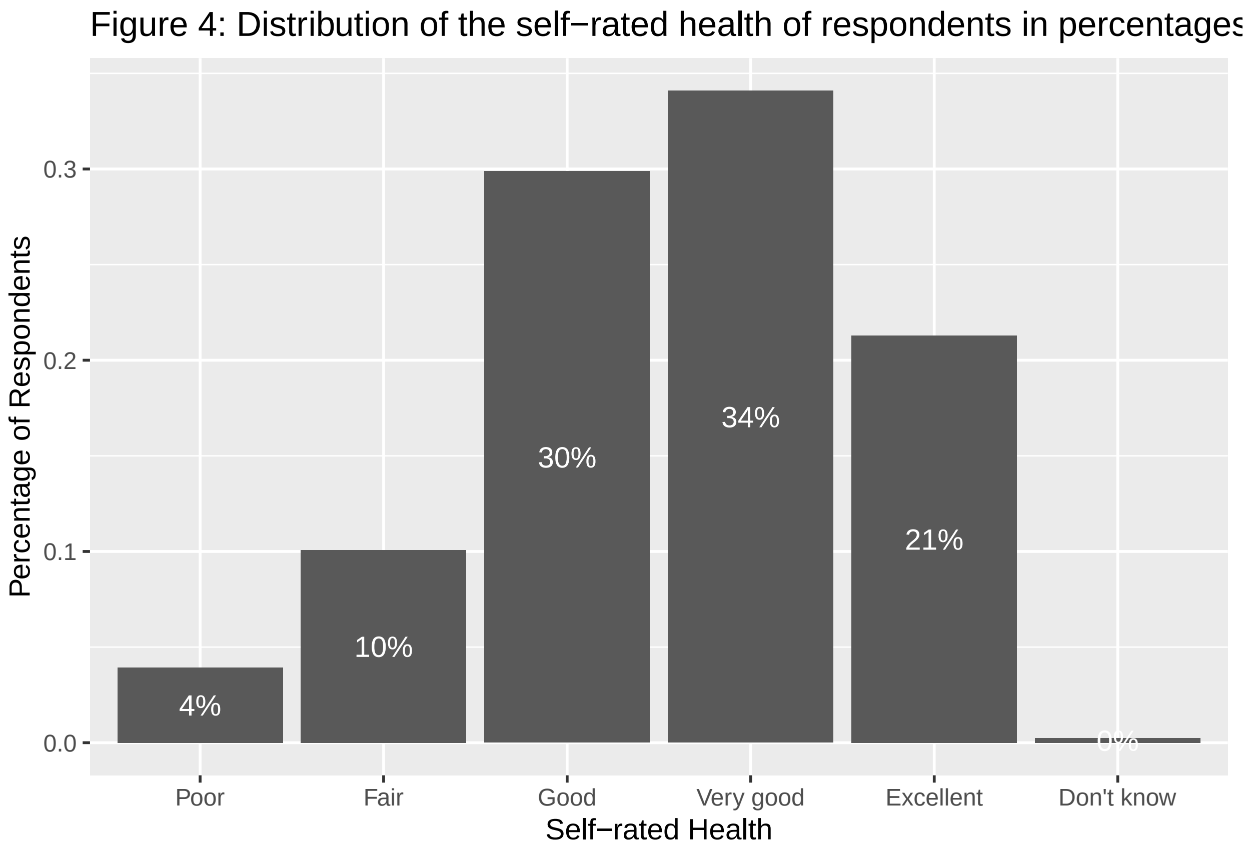
Figure 3: Distribution of the marital status of respondents in percentages.



```
# summary table with counts for self-rated health
tibble_health <- data.frame(table(data$self_rated_health)) %>%
  rename(
    health = Var1,
    count = Freq
  )
# manually changing the order of factors in the following output table and graph
tibble_health$health <- factor(tibble_health$health,
                               levels = c("Poor", "Fair", "Good", "Very good",
                                          "Excellent", "Don’t know"))
tibble_health
```

```
##       health count
## 1 Don’t know    51
## 2  Excellent  4373
## 3       Fair  2070
## 4       Good  6139
## 5       Poor   809
## 6  Very good  7004
```

```
# distribution of self-rated health as a bar graph in percentages
ggplot(tibble_health, aes(x = health, y = count/total_count)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(count/total_count*100), "%")),
            color="white",
            position = position_stack(vjust = 0.5)) +
  xlab("Self-rated Health") +
  ylab("Percentage of Respondents") +
  ggtitle("Figure 4: Distribution of the self-rated health of respondents in percentages.")
```

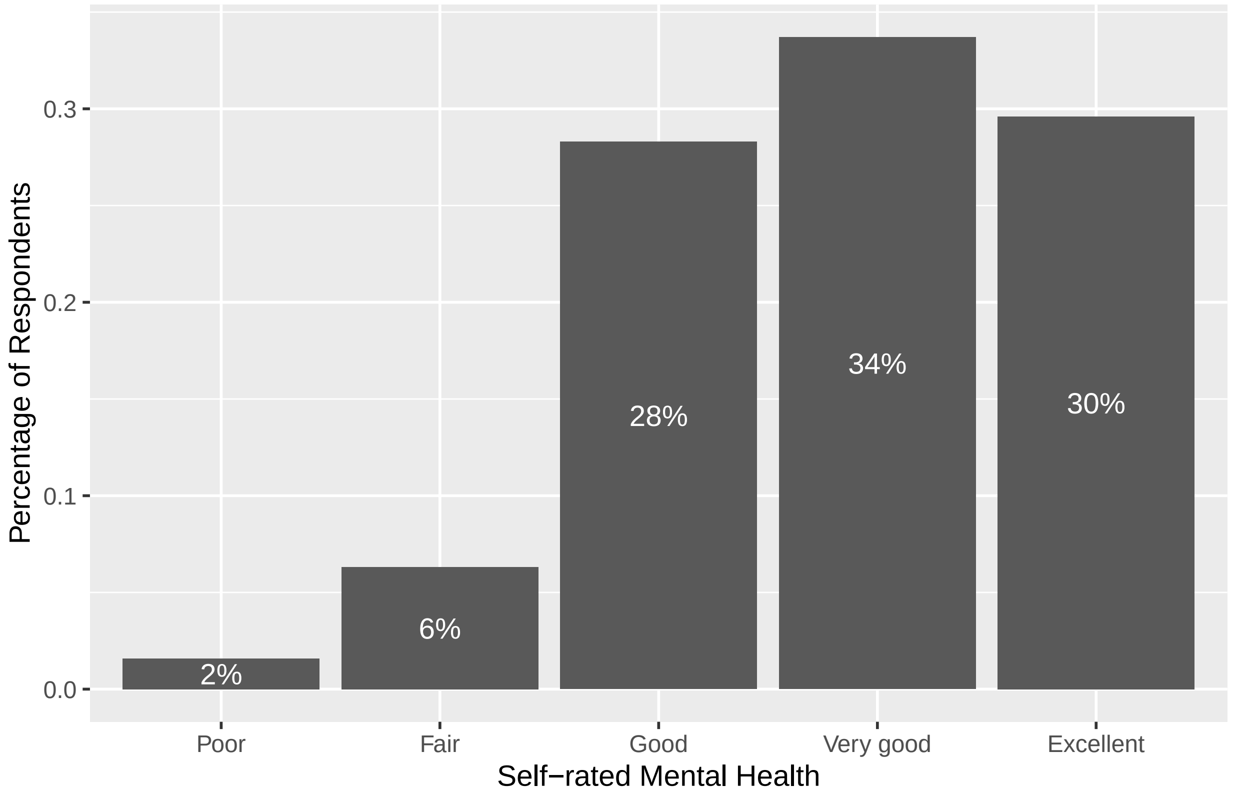


```
# summary table with counts for self-rated mental health
tibble_mental_health <- data.frame(table(data$self_rated_mental_health)) %>%
  rename(
    mental_health = Var1,
    count = Freq
  )
# manually changing the order of factors in the following output table and graph
tibble_mental_health$mental_health <- factor(tibble_mental_health$mental_health,
                                             levels = c("Poor", "Fair", "Good",
                                                        "Very good", "Excellent",
                                                        "Don't know"))
tibble_mental_health
```

```
##   mental_health count
## 1     Excellent  6080
## 2          Fair  1296
## 3          Good  5813
## 4          Poor   326
## 5     Very good  6924
```

```r
# distribution of self-rated mental health as a bar graph in percentages
ggplot(tibble_mental_health, aes(x = mental_health, y = count/total_count)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(count/total_count*100), "%")),
            color="white",
            position = position_stack(vjust = 0.5)) +
  xlab("Self-rated Mental Health") +
  ylab("Percentage of Respondents") +
  ggtitle("Figure 5: Self-rated mental health of respondents in percentages.")
```

Figure 5: 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 respondents, 57 individuals declined to provide a self-rating of their mental health. 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 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.

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.

```r
# install.packages("nnet")
require(nnet)
library(tidyverse)
model <- nnet::multinom(self_rated_mental_health ~ age + sex + marital_status + self_rated_health,
                  data = data)
```

```
## # weights:  70 (52 variable)
## initial  value 32875.988237
## iter  10 value 23611.931535
## iter  20 value 23441.438461
## iter  30 value 23049.583322
## iter  40 value 22914.075187
## iter  50 value 22822.824806
## iter  60 value 22775.879430
## final  value 22775.875150
## converged
```

```r
summary(model)
```

```
## Call:
## nnet::multinom(formula = self_rated_mental_health ~ age + sex + 
##     marital_status + self_rated_health, data = data)
## 
## Coefficients:
##           (Intercept)          age    sexMale marital_statusLiving common-law
## Fair        1.0217425 -0.032798399 -0.3224026                     -0.16949105
## Good        1.5874823 -0.011040378 -0.2400994                      0.09958731
## Poor        1.8561011 -0.044664300 -0.2137392                     -0.62177910
## Very good   0.9413021 -0.005288015 -0.2162900                      0.04000497
##           marital_statusMarried marital_statusSeparated
## Fair                -0.29251444               0.7299606
## Good                -0.01207268               0.4443427
## Poor                -0.92293929               0.5025271
## Very good           -0.03471334               0.1393385
##           marital_statusSingle, never married marital_statusWidowed
## Fair                              0.224221668            0.17670603
## Good                              0.130575216            0.22994484
## Poor                              0.033859427            0.01458883
## Very good                        -0.002477955            0.17862358
##           self_rated_healthExcellent self_rated_healthFair
## Fair                       -3.360131             1.5800364
## Good                       -3.050227             0.1066989
## Poor                       -5.162473            -0.2307758
## Very good                  -1.813127            -0.3809237
##           self_rated_healthGood self_rated_healthPoor
## Fair                 -0.1643287             1.9008484
## Good                  0.1545703             0.1368938
## Poor                 -1.8749204             1.6135608
## Very good            -0.1348050            -0.1311144
##           self_rated_healthVery good
## Fair                      -1.6225589
## Good                      -1.2432679
## Poor                      -3.5250590
## Very good                  0.3213029
## 
## Std. Errors:
##           (Intercept)         age    sexMale marital_statusLiving common-law
## Fair        0.4695617 0.002429496 0.06897229                      0.16272558
## Good        0.3101351 0.001459802 0.04227380                      0.09946604
## Poor        0.5122282 0.004525478 0.12350622                      0.28565061
## Very good   0.3325459 0.001338815 0.03893398                      0.09150798
##           marital_statusMarried marital_statusSeparated
## Fair                 0.12358331               0.1975624
## Good                 0.07773531               0.1410396
## Poor                 0.20958816               0.3020509
## Very good            0.07207928               0.1364599
##           marital_statusSingle, never married marital_statusWidowed
## Fair                               0.13684546            0.15750158
## Good                               0.08986249            0.10021966
## Poor                               0.21528134            0.25746406
## Very good                          0.08379589            0.09385189
##           self_rated_healthExcellent self_rated_healthFair
## Fair                       0.4574703             0.4452440
## Good                       0.2951180             0.2967677
## Poor                       0.5107378             0.4420195
## Very good                  0.3189361             0.3248795
##           self_rated_healthGood self_rated_healthPoor
## Fair                  0.4437937             0.4540376
## Good                  0.2921365             0.3112635
## Poor                  0.4407436             0.4439970
## Very good             0.3188215             0.3405639
##           self_rated_healthVery good
## Fair                       0.4472054
## Good                       0.2925541
## Poor                       0.4643552
## Very good                  0.3178609
##
## Residual Deviance: 45551.75
## AIC: 45655.75
```

```
head(fitted(model))
```

```
##   Excellent        Fair       Good         Poor Very good
## 1 0.6719425 0.014404347 0.09910454 0.0024202042 0.2121284
## 2 0.1951443 0.046536541 0.49248128 0.0062707489 0.2595671
## 3 0.2389554 0.012145612 0.16505036 0.0010447006 0.5828039
## 4 0.2601140 0.007720795 0.14990947 0.0005466639 0.5817091
## 5 0.1437139 0.082680958 0.52334153 0.0175120007 0.2327516
## 6 0.7074104 0.006452120 0.08074126 0.0006178245 0.2047784
```

```
input <- data.frame(self_rated_health = c("Excellent"), age = c(21.5), sex = c("Male"), marital_status =
predict(model, newdata = input, "probs")
```

```
##   Excellent        Fair        Good        Poor   Very good
## 0.658524299 0.028453356 0.107809344 0.007717323 0.197495678
```

Results

Discussion

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

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 Mental health statistics: https://www.camh.ca/en/Driving-Change/The-Crisisis-Real/Mental-Health-Statistics Discriminative vs Generative Classifiers (Naive Bayes vs lostic 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