Factors that contribute towards a positive feeling towards life in Canada

Daniel Grant

Oct. 19th 2020

Factors that contribute towards a positive feeling towards life

Daniel Grant

Oct 19th 2020

Abstract

How one feels about the subject of life is something that everyone knows about themselves but sometimes it is hard to discover the factors that attribute to these feelings. It is important for us as statisticians to be able to identify the connection between these different data sets and figure out if there is a relationship between the variable or not. In this paper we will be looking at what are the most important factors to living a happy life. There is a clear relationship between some of these factors and the feeling towards life.

Introduction

In this study we are looking at a few key factors that effect the feelings towards life that a person has. We have decided to look towards the age, average hours worked, self-rated health, self-rated mental health, and the income of the respondent. Our goal with this assignment is to understand the relationship between these factors and the feelings towards life of the respondent of the survey. We can use the large amount of data available to us to infer a few facts about how income and other factors can affect one's feelings towards their life. From across Canada there are many answers that are given to this survey that we are able to use to create a model and derive inferences from. Along with the GSS data that is being used in this paper the government also provides documentation through a user guide on their website. This guide outlines various concepts that are useful to us as analyzers of the data. In the 3.2 Survey content section of this user guide the author, Pascale Beaupré of Statistics Canada, discusses the content of the survey and expands on what the raw data means. When looking at the Health and subjective well being section of 3.2 we are told "This module covers the respondent's life satisfaction, self-rated general health, and mental health. All are important factors in assessing the well-being of Canadians." (Beaupre,2020) . Assessing factors that lead towards the subjective well-being of Canadians will be the main goal of this paper.

As outlined in the user guide the goal for the sample size of the final 2017 GSS was 20,000 however, when we look at our raw data we are given 20,602 entries which means that the GSS was successful in it's mission and that we have plenty of data to analyze. There are seven factors that we are going to be looking at closely from the GSS data that we have retrieved off the CHASS website (https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm). The first step that we took was to acquire this GSS data, which was supplied for us through CHASS and the University of Toronto. We were able to download this data and then use code that was supplied to us by our course instructor Samantha-Jo Caetano to clean up the data and make it easier to work with and manipulate. This cleaning program was authored by both Professor Caetano and Rohan Alexander and without this R script it would be very difficult to finish this paper. After the data has been cleaned, we are able to take a proper look at it and decide what trends we would like to look for. When looking at the data for the first time a statistic that stood out to us was the satisfaction with life stat that was displayed. This number was self-reported, so it is seemingly accurate to what the person was feeling when the survey was taken. This factor will be the focus of our study and this paper. The other factors that we will be looking at specifically will be age, average hours worked, self-rated health, self-rated mental health, and the incomes of both the respondent and the income of the whole family of the respondent. In the User guide the author, Beaupré, discusses the objectives of the survey, a) To gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time; and b) To provide information on specific social policy issues of current or emerging interest(Beaupre, 2020). We will be sharing these objectives as we continue with this paper. This discussion will continue in the next section where will be looking at our raw data more in depth and figuring out what we are able to do with it.

Data

The data that was chosen to be analyzed in this report was the 2017 General Social Survey. The raw data that we retrieved was available at CHASS(https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm). This data was chosen due to the large number of variables available for analysis and the different factors that effect it. The user guide that is supplied with the data by Statistics Canada is very useful in trying to understand both the purpose of the data and how it was gathered. When we are looking at how this data was gathered in the first place, we are told that there are different ways that different categories are collected. For example, the income data was collected through tax information, assumingly through the CRA, however if one of the respondents of the survey didn't want their tax data linked to the survey they would be recorded as NA. Beaupré states that this is a plus to the quality of data that they received, "Linking to tax data diminishes respondent burden and increases data quality both in terms of accuracy and in response rates"(Beaupre, 2020). This allows for very accurate information as long as the respondents of the survey are submitting the correct information on their tax forms. This data was collected using "…A simple random sample without replacement of records was next performed in each stratum.(Beaupre, 2020)" Do to this sampling we are not left with the whole population of Canada as our sample size, instead we are left with 20,000 simple random samples of the Canadian population. This amount was chosen as it is balanced for the need to predict both national-level and stratum-level estimates(Beaupre, 2020) . It would be near impossible to survey the entire population of Canada in this way so statisticians at Statistics Canada needed to find an acceptable sampling proportion. When we look at the raw data, we can see that there are a multitude of NA entries in the data which can make it hard to work with. This would fall under the category of Item nonresponse, where the participant in the survey refuses to answer for the missing value(Wu, 2020) .

The key features of this data set are the fact that it has such a large dataset to pull from and it varied data from across the county. This makes it ideal for tracking social trends across the country, as was the goal of the survey. There are however weaknesses to the survey as well, which can be discussed in greater detail. For example, the survey ignored 3 of our territories, the Yukon, the Northwest Territories, and Nunavut(Beaupre, 2020) . This means that our raw data is not truly descriptive of the population of Canada. Another weakness that this data set has is that many of it's variables are categorical which can make it harder to apply a strong model to. We are able to look at the raw data between our factors to see if there could possibly be a relationship in it.

```
## -- Attaching packages --------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2          v purrr   0.3.4
## v tibble  3.0.3.9004     v dplyr   1.0.2
## v tidyr   1.1.2          v stringr 1.4.0
## v readr   1.4.0          v forcats 0.5.0

## -- Conflicts ------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: grid
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loading required package: survival

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Warning: package 'brms' was built under R version 4.0.3

## Loading required package: Rcpp

## Loading 'brms' package (version 2.14.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following object is masked from 'package:survival':
##
##     kidney

## The following object is masked from 'package:stats':
##
##     ar

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize

## The following object is masked from 'package:purrr':
##
##     compact

## Warning: Removed 271 rows containing non-finite values (stat_count).
```

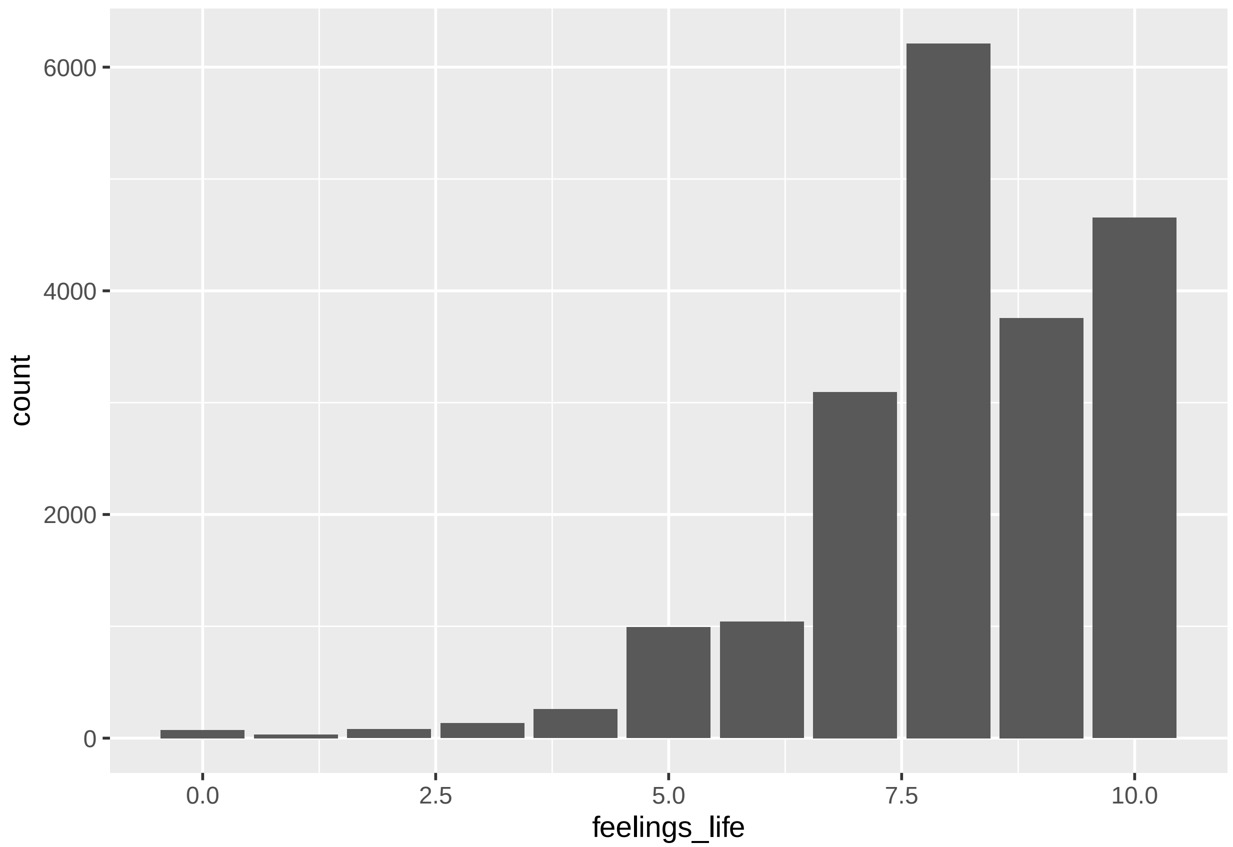

Fig 1.

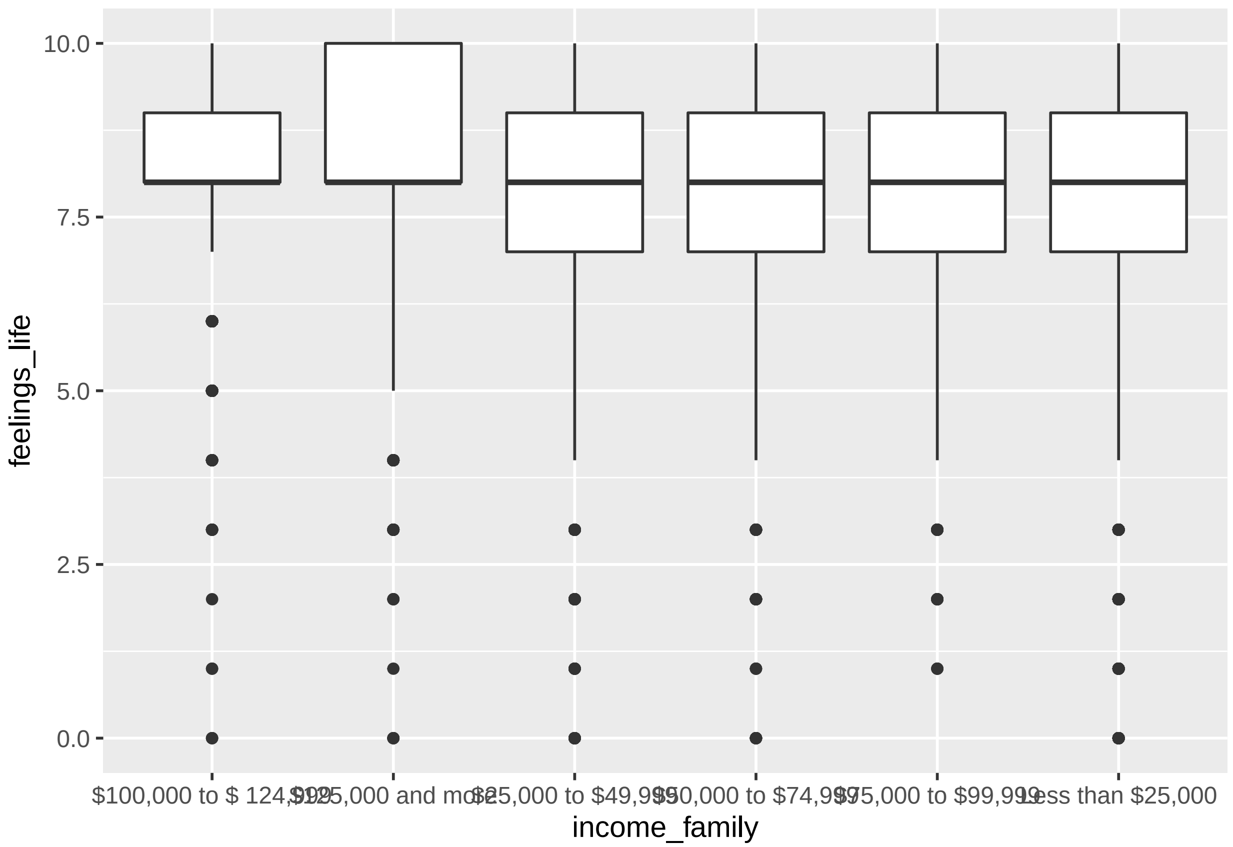
10.0
7.5
5.0
2.5
0.0
feelings_life
$100,000 to $ 124,999
$125,000 and more
$25,000 to $49,999
$50,000 to $74,999
$75,000 to $99,999
Less than $25,000
income_family

Fig 2.

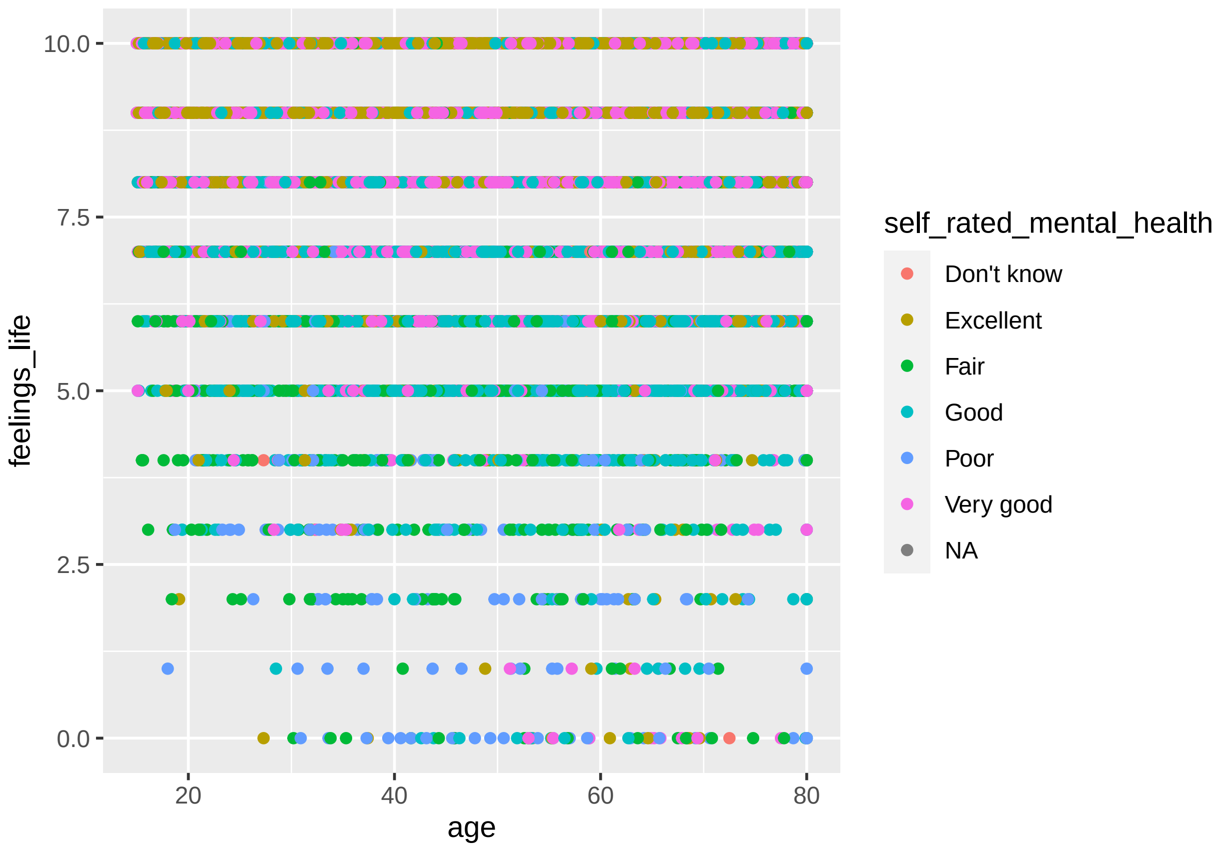

Fig 3.

Looking at Fig 1. we are able to see that the general trend of Canadians is to be very content with life, with most answers falling in above 7.Plotting the raw data can look confusing as the data hasn't been manipulated yet to make it more readable. This is seen in Fig. 3 where we can make out a trend in the colouring of the graph, where the better you feel about life the more likely you are to feel well mentally. In the following sections we will attempt to make more inferences out of the data and to present it more clearly. In the next section we will discuss our model that we have selected to work with.

Model

The model that we decided to work with for this paper is one that will help us understand the relationship between the satisfaction with life that the respondent is feeling at the time of the survey and different factors of their life that might effect this. Our hypothesis is that the older and more money that one has as an income the better one feels about life. This will mean that we will have to run a multiple linear regression with feelings_life as the response variable and age as well as income_family as the input variables. Our goal is to see if we can find a linear relationship between these variables thus proving that there is a relationship between the income, age, and the feelings that one has towards life. To start this modeling procedure, we first need to decide what type of model we are going to be using.

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\epsilon}_i$ Eq. 1

Equation 1 is the model that we will be using to attempt to discover a relationship between the previously discussed factors and one's feeling towards life. Our $\hat{Y}_i$ is equal to the estimate of the feelings towards life, our Xi1 will be the age of the respondent, and our Xi2 will be the income level of the respondent. At this point we realize that as the income_family variable is categorical and that we are unable to fit it to a single X. We must then use our income levels as a ranked as.factor(). Using the as.factor() command on the income_family data will allow us to reorganize our model and we are left with

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} + \hat{\beta}_4 X_{i4} + \hat{\beta}_5 X_{i5} + \hat{\beta}_6 X_{i6} + \hat{\beta}_7 X_{i7} + \hat{\beta}_8 X_{i8} + \hat{\beta}_9 X_{i9} \text{ Eq. 2}$$

$\hat{\beta}_2$ to $\hat{\beta}_9$

will be the parameters for the income levels of the respondent. Due to the income levels being categorical, there are only 5 parameters for 6 outcomes. We decided on using a multiple linear regression as it would allow us to see how different factors affect a single outcome variable at once. We are unable to use a simple linear regression because of the fact that the income is a categorical variable. Our model is not applicable to be Bayesian as we are not looking for the answer to a binary question, however with some manipulation of the data that would have been another option if we decided to as the question, "Is someone happy?". We would be able to find the probability that someone is happy based upon the economic factors and age that we discussed earlier. Therefore, we are approaching this model as a general linear regression model. The first step to find this model and to start our analysis will be to load the necessary packages,

after we have loaded the necessary packages we can begin out analysis. We must then attach the raw gss data to a variable, Data in this case,

```
Data <- read.csv("gss2.csv")
```

From Data we can extract the variables that we're interested in to make a cleaner data frame with the same information

```
SLM<- select(Data, caseid, age, feelings_life, average_hours_worked, self_rated_health, self_rated_menta
```

After this we notice that there are many NA entries for the categories that we are looking at. As discussed earlier in this paper this is due to some participants in the survey being not willing to connect their tax accounts with this survey. To deal with these entries we can use the following code in R,

```
SLM <- filter(SLM, !is.na(average_hours_worked))# Removes NA from average_hours_worked

SLM <- filter(SLM, !is.na(feelings_life))# Removes missing answers

SLM <- filter(SLM, !is.na(self_rated_mental_health))#Removes NA values
```

This code will remove the NA entries in the above 3 categories, average__hours__worked, feelings__life and self__rated__mental__health.OUr next goal is to find a smaller sample size to work with, to do this we are going to preform a simple random sample of 50 entries from each income class for the income family variable,

```
set.seed(1)#Set seed so we always get the same results from the following chunks

SRS<-ddply(SLM,.(SLM$income_family),function(x) x[sample(nrow(x),50),]) #simple random sampling 50 obse
```

This will allow us to work with a smaller sample of data to preform analysis on and allow us to apply our model to the whole population. We are now going to preform our general linear model, do do this we first need to find our population size, the size of the entire Data dataframe that we previously created and our sample size, the size of our SRS data frame,

```
N<- length(Data$caseid) #This will be our large N or, our population size 20,062
n<- length(SRS$caseid) #this will be our sample size, 300
```

Now we are also able to take this opportunity to clean up our data a little bit. when looking at the factors of both self__rated__mental__health and income__family we are able to see that R has them in the incorrect order, or rather, what R believes to be the right order, we are able to fix this quite easily,

```
SRS$self_rated_mental_health <- factor(SRS$self_rated_mental_health, levels = c("Excellent","Very good"
#Reorders the levels of the self reported mental health

SRS$income_family <- factor(SRS$income_family, levels = c("$125,000 and more","$100,000 to $ 124,999","$
#Reorders the levels on income to make data more readable
```

This manipulation of the data will make future graphs and regression models much easier to understand and more appealing to the eye. In this next section of code we will be creating a survey design and will be running a linear model on our data without replacement and using dummy variables for the categorical variables.

```
fpc.srs2 <- rep(N,n)

feel.design <- svydesign(id=~caseid, data=SRS, fpc= fpc.srs2)

SLM.ratio<- svyratio(~SRS$feelings_life, ~SRS$age, design=feel.design)

glm.feel <- svyglm(feelings_life ~ age + as.factor(income_family) + as.factor(self_rated_mental_health)

#runs a generalized linear model on our data without replacement and using dummy variables for both fam
```

This leaves us with a successfully modeled linear model. by using the summary() function on our model we are able to learn more about it and how the variables effect our response.

```
summary(glm.feel)
```

```
##
## Call:
## svyglm(formula = feelings_life ~ age + as.factor(income_family) +
##     as.factor(self_rated_mental_health), design = feel.design)
##
## Survey design:
## svydesign(id = ~caseid, data = SRS, fpc = fpc.srs2)
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                     8.853254   0.281875  31.408
## age                                             0.008067   0.005101   1.581
## as.factor(income_family)$100,000 to $ 124,999 -0.191852   0.204382  -0.939
## as.factor(income_family)$75,000 to $99,999    -0.306912   0.205248  -1.495
## as.factor(income_family)$50,000 to $74,999    -0.091450   0.196073  -0.466
## as.factor(income_family)$25,000 to $49,999    -0.148011   0.242409  -0.611
## as.factor(income_family)Less than $25,000     -0.398853   0.263128  -1.516
## as.factor(self_rated_mental_health)Very good  -0.901695   0.162704  -5.542
## as.factor(self_rated_mental_health)Good       -1.521154   0.190914  -7.968
## as.factor(self_rated_mental_health)Fair       -2.608488   0.398177  -6.551
## as.factor(self_rated_mental_health)Poor       -4.987146   0.764452  -6.524
##                                               Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## age                                               0.115
## as.factor(income_family)$100,000 to $ 124,999    0.349
## as.factor(income_family)$75,000 to $99,999       0.136
## as.factor(income_family)$50,000 to $74,999       0.641
## as.factor(income_family)$25,000 to $49,999       0.542
## as.factor(income_family)Less than $25,000        0.131
## as.factor(self_rated_mental_health)Very good  6.74e-08 ***
## as.factor(self_rated_mental_health)Good       3.76e-14 ***
## as.factor(self_rated_mental_health)Fair       2.61e-10 ***
## as.factor(self_rated_mental_health)Poor       3.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 1.505806)
##
## Number of Fisher Scoring iterations: 2
```

By calling upon this function we are able to see which variables are significant to the outcome of the model and which ones play a smaller role in the outputs. Thankfully R is able to help us with this, when looking at the output of summary() we can see that the state of someone's mental health effects their feelings about life very heavily.In the summary we also notice that the income level of 100,000 to 124,999 and the mental health response of Excellent are not present, that is because we ran these variables as as.factor(). If one has Excellent mental health and an income in the described range above then the rest of the responses after age would be zero. OUr equation for our multiple linear regression would look like

$$Feelings-Life = .8760408+0.011090X_{age}-0.043237X_{125k}-0.421014X_{25k}-0.824433X_{50k}-0.514348X_{75k}-0.486277X_{<25k}-1.982401X_{Fair}-1.564025X_{Good}-3.129485X_{Poor}-0.696082X_{VeryGood}$$

This is our final model and the last 9 variables are considered dummy variables since their x's will only ever be 0 or 1 depending on what the state of the participant's mental health is, or what income bracket they belong to.

Results

We will now be going over the results of our efforts. Throughout our process of modeling we are able to discover different aspects of our model and how it can succeed or fail at presenting us with a proper perspective of our data.

We will be looking at a few presentations of our Simple randomly selected data

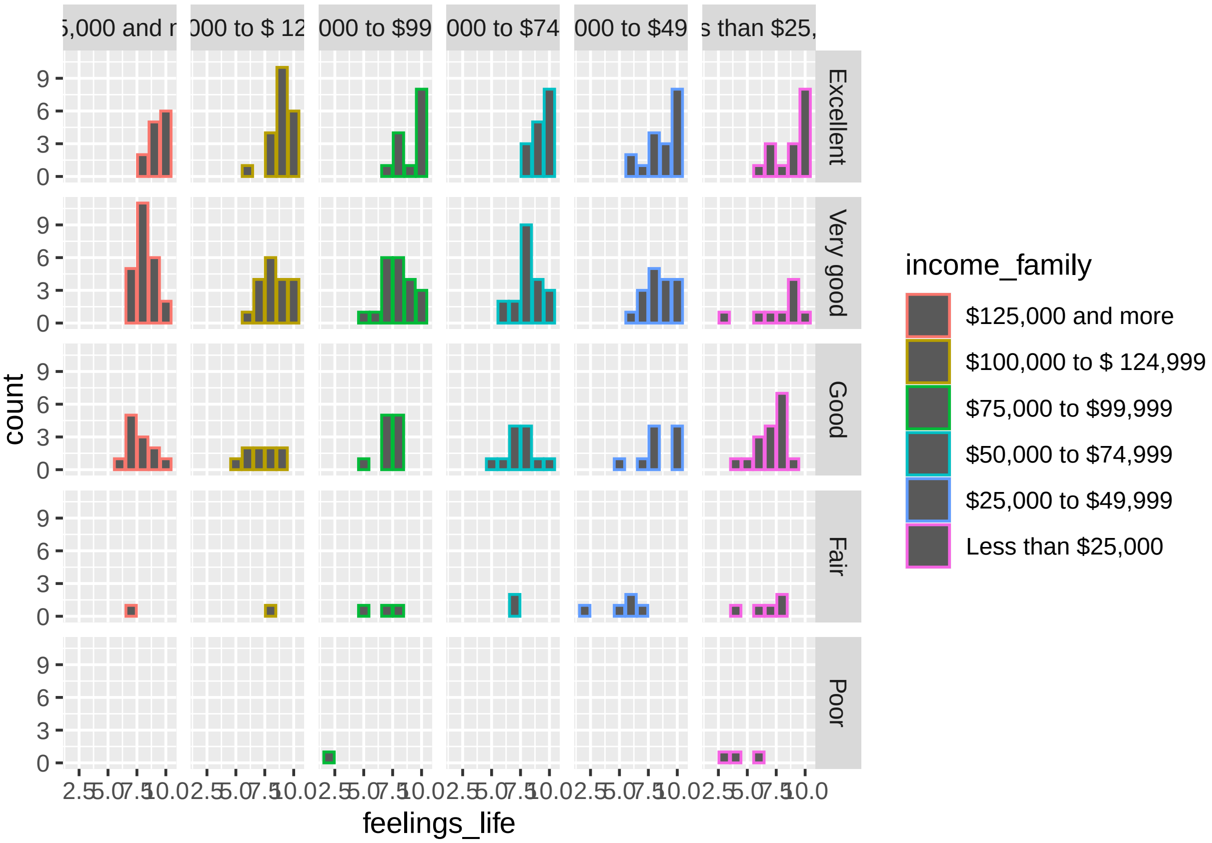

Fig. 4

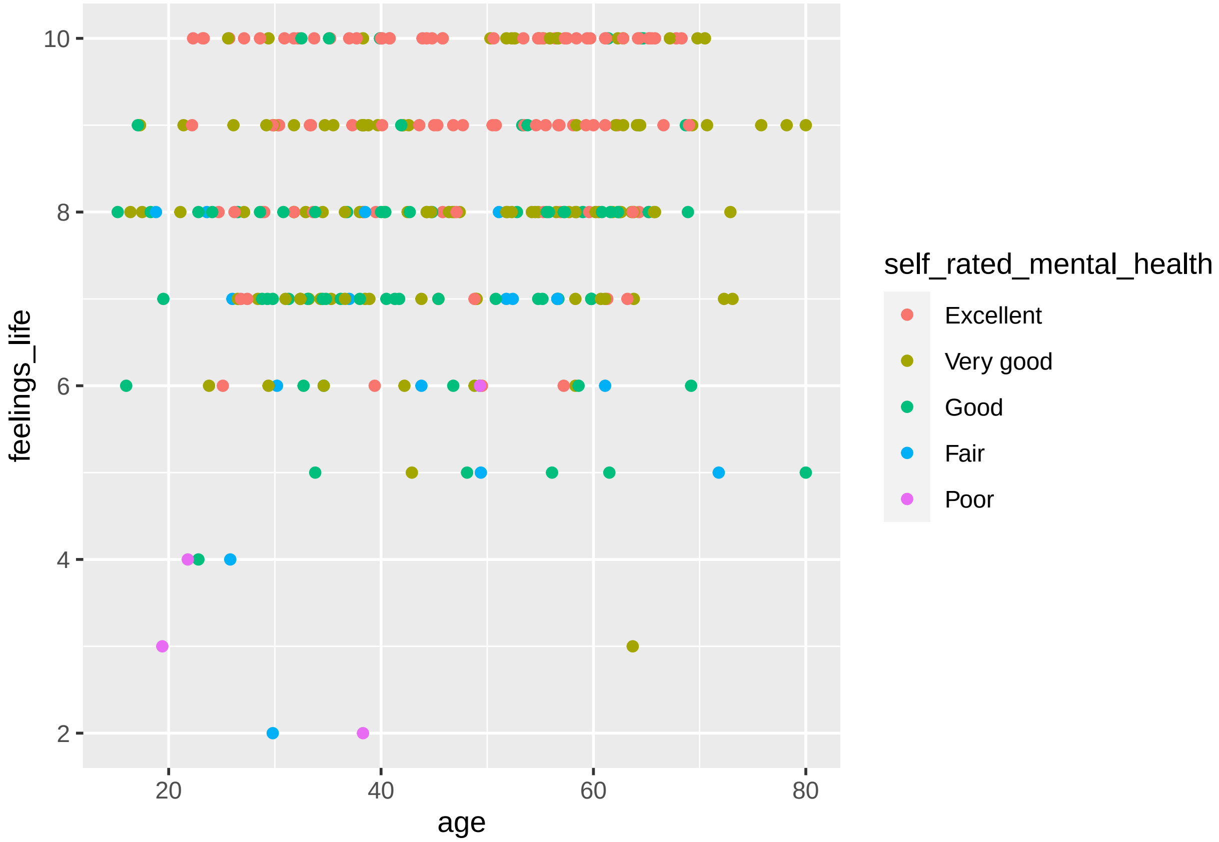

Fig. 5

```
##
## Call:
## svyglm(formula = feelings_life ~ age + as.factor(income_family) +
##     as.factor(self_rated_mental_health), design = feel.design)
##
## Survey design:
## svydesign(id = ~caseid, data = SRS, fpc = fpc.srs2)
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    8.853254   0.281875  31.408
## age                                            0.008067   0.005101   1.581
## as.factor(income_family)$100,000 to $ 124,999 -0.191852   0.204382  -0.939
## as.factor(income_family)$75,000 to $99,999    -0.306912   0.205248  -1.495
## as.factor(income_family)$50,000 to $74,999    -0.091450   0.196073  -0.466
## as.factor(income_family)$25,000 to $49,999    -0.148011   0.242409  -0.611
## as.factor(income_family)Less than $25,000     -0.398853   0.263128  -1.516
## as.factor(self_rated_mental_health)Very good  -0.901695   0.162704  -5.542
## as.factor(self_rated_mental_health)Good       -1.521154   0.190914  -7.968
## as.factor(self_rated_mental_health)Fair       -2.608488   0.398177  -6.551
## as.factor(self_rated_mental_health)Poor       -4.987146   0.764452  -6.524
##                                               Pr(>|t|)
## (Intercept)                                    < 2e-16 ***
## age                                              0.115
## as.factor(income_family)$100,000 to $ 124,999    0.349
## as.factor(income_family)$75,000 to $99,999       0.136    
## as.factor(income_family)$50,000 to $74,999       0.641    
## as.factor(income_family)$25,000 to $49,999       0.542    
## as.factor(income_family)Less than $25,000        0.131    
## as.factor(self_rated_mental_health)Very good  6.74e-08 ***
## as.factor(self_rated_mental_health)Good       3.76e-14 ***
## as.factor(self_rated_mental_health)Fair       2.61e-10 ***
## as.factor(self_rated_mental_health)Poor       3.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 1.505806)
## 
## Number of Fisher Scoring iterations: 2
```

Table 1.

We are now able to discuss our results comprehensively using our figures and the summary of our regression. When we look at Fig.1 we are able to see what the feelings are towards life for each income bracket based on their response to the self mental health assessment. We can infer that there is a connection from mental health to the feelings towards life that the participants have. by looking at Table 1 we are able to see that there is a strong negative relationship between respondents who answered Fair and Poor to the question on current mental health.We will be able to discuss these factors and results more indepth in the next section of this paper.

Discussion

We are now able to begin our discussion on our data and what the results mean.Through looking at all the information that is infront of us we are able to see that we were able to find a relationship between the feelings that a respondent has towards life and their age and income bracket. We found through a multiple linear regression that there is a weak positive relationship between age and your lookout on life, for each one year increase in your age there is estimated to be a 0.0009 increase in your outlook on life. However one of our other focuses of this paper is to see if there is any relationship between income and feelings. As we can see in the output of glm.feel shown in table 1. we are able to see that there is a strong negative relationship between living on less than $25,000 a year and the respondent's outlook on life. This makes sense and backs up the hypothesis that as you make more money the better your outlook on life is. This strong relationship is even documented as such by R in the output in table 1. While we were expecting a stronger positive relationship between income and feelings on life, the fact that there is still a relationship, abet a negative shows that here is a relationship between income and the respondent's feelings.

Weaknesses

There are many weaknesses that we can discuss in this paper, there are weaknesses with the survey, our sampling, and our model. Understanding these weaknesses are paramount to understanding how we can improve in the future. With regards to the survey we have already outlined a few weaknesses that it could be suffering from. For example, not surveying 3 territories of Canada can cause some issues in the reporting of a nationwide survey, these populations are part of Canada too. Hoiwever it is understandable that reaching the populations of these areas can be difficult. In Beaupre's documentation for the GSS data they discuss how the data was collected, they used a phone survey where they phoned households from 9am to 9:30pm on weekdays(Beaupre,2020). This would leave a small population unable to be reached by the surveyors. For example, if one was working nights and sleeping through the day it would be hard for them to be reached by this method. While these are some of the weaknesses that the survey might face there are other weaknesses that we can look at that are closer to this paper's subject. Our subset of data that we chose could have some weaknesses in that a relationship between poor mental health and a poor outlook on life doesn't require regression to understand. While the goal of this paper was to discover if there was a relationship between the income bracket of Canadians and their level of happiness we were not able to discover any strong relationships between being economically well off and the quality of their life.

Next Steps

The next steps that we hope to follow would be to go back and to see if there are any other features of the data set that would affect one's feelings towards life more than income. If we are able to find a stronger relationship between feelings of life and another variable it might be worth it to change the focus of our paper to see what truly makes someone more satisfied with life. We are also hopeful that the rest of the course will introduce even more methods that would allow us to analyze the data even better or perhaps introduce new methods so we can get new inferences out of the data. If we, the authors of this paper are able to have the infrastructure and surveying abilities of Statistics Canada we might be able to preform another survey that would reach into the territories that were missed as well as attempt to get more accurate income data from the CRA. if we are able to get more data from around Canada then that would allow for more varied data. If the issue for the surveyors that ran the first survey was that there wasn't proper phone infrastructure in the missing territories one perhaps could use a mail-in survey instead. This would take much longer and probably have a smaller response rate but we would then have data from these parts of Canada. The best next step that we can take is to continue learning about statistics and apply our learning to more and more data sets.

References

Beaupre, P. (2020). General Social Survey Cycle 31 : Families Public Use Microdata File Documentation and User's Guide. Ottawa, ON: Statistics Canada.

Wu, C., & Thompson, M. E. (2020). Sampling theory and practice. Cham, Switzerland: Springer.