Linear Regression Modle on Individuals Self rating About Feelings Of Their Life

Wenxi Li, Tianyi Jiang, Jiatong Li

Oct.18,2020

Abstract

An Individual's well-being and living conditions have always been a major topic for the government. General Social Survey(GSS) is a survey carried out since 1985. It has a great collection of cross-sectional data that allows for many analyses about the living condition of Canadians. In this study, we used data from the 2017 GSS data and aimed to study the Canadian's overall well-being. By building a linear model, we found a linear trend between one's self feelings rate about their life and many explanatory variables, ie. family income, age, etc. From the model, we can predict one's well-being, which could be useful in many ways.

Introduction

In the GSS 2017, a question 'On a level from 1 to 10, in general, would you say your mental health is...?' was asked. We want to find what factors in the data determine the rating. To do so, we tried to find a model that predicts it. Obviously, the score could be affected by many factors, so in this study, we chose three that we are most interested in, which are one's household size, age, and income of the family. Clearly, an individual's family condition as well as their personal living condition play an important role in determining their own feelings about life. Therefore, we chose the household size and the income of the family as the criteria to measure an individual's family condition, and his/her age as the standard to measure the personal living conditions.

Data

We carried out the study using the GSS 2017 (General Social Study) data available from CHASS (a computing facility within the Faculty of Arts & Science, University of Toronto). The data itself is published by Statistics Canada. The survey that is used to collect the data practiced a stratification sampling technique; there are a total of 10 strata, where each province of Canada counts as one. The survey estimates are weighted according to the population size of each province(strata).

The GSS 2017 has a target population of all persons who are 15 years old or older in Canada excluding residents of the Yukon, Northwest Territories, and Nunavut and full-time residents of institutions. The survey frame are a list of telephone numbers in use available to Statistics Canada and the Address Register(AR). The final survey sample is those who respond to the telephone call. Noticed that if the phone is not being picked up or the person wasn't available at that time, several more phone calls were made to that telephone number to maximize the respondent's response rate(maximize the survey sample). It is also worth noticing that for the phone numbers that didn't respond, a three-stage adjustment was made; it is adjusted by how much auxiliary information was available to the Statistics Canada or how much auxiliary information were collected from a partial non-response. All the auxiliary information is used to model propensity to respond.

The data are great to perform a regression model since it has a great number of data sets which analyzes a model more convincing and practical. Besides, with each data set, there are also many variables that we can look into and analyze, i.e the person's marital status, number of children, age, etc. We also noticed that the telephone number belonging to the same address were grouped together, which would make the model even more accurate since no duplicate or similar data were obtained.

However, there do have some drawbacks. For instance, there is missing information that could influence the final model; since we are omitting the data with missing information. Additionally, all the data collected are from those who respond to the telephone interview, but those households or individuals who did not have a phone number were clearly excluded from the survey. Therefore, the final model we built might be biased, since we are not considering the living conditions or well-being of those people.

When deciding what variables needed to be included in our model, we chose the variables that have a representative meaning of an individual's personal and family life. Here we chose the age of the individual, household size, and family income as the three explanatory variables that might be able to explain their rating of the life feelings. Notice here, we chose household size rather than the number of children since the latter is less meaningful than the first; household size has a more general meaning since the number of children is included in it. Additionally, family income is more preferable than an individual's income as it better represents the family conditions of that individual. We also included age since it is the most noticeable variable or factor of respondents themselves.

Model

We used R software to do the linear regression model. Specifically, we used a finite population correction to make the model more accurate.

From the linear regression model we carried out, we can predict the individual's rate of the feelings of his/her life using their household size, age, and income of the family, where the income of the family is categorical variables. Here, a range of family income is preferred since the number could vary greatly. Therefore, to reduce the number of outliers (unusual or extreme values) and to make the model more accurate, a categorical variable of family income is more ideal. On the other hand, it is reasonable to use age and household size as numerical variables since they are generally easy to be counted or reported, and they are less likely to have too many outliers in them.

The model follows the formula:

rating of one's feelings about their life = 7.5661716 + 0.017026 * age - 0.8310879 * (income_familyLess than $25,000) + -0.1475670 * (income_family$75,000 to $99,999) - 0.4615397 * (income_family$25,000 to $49,999) - 0.2493725 * (income_family$50,000 to $74,999)

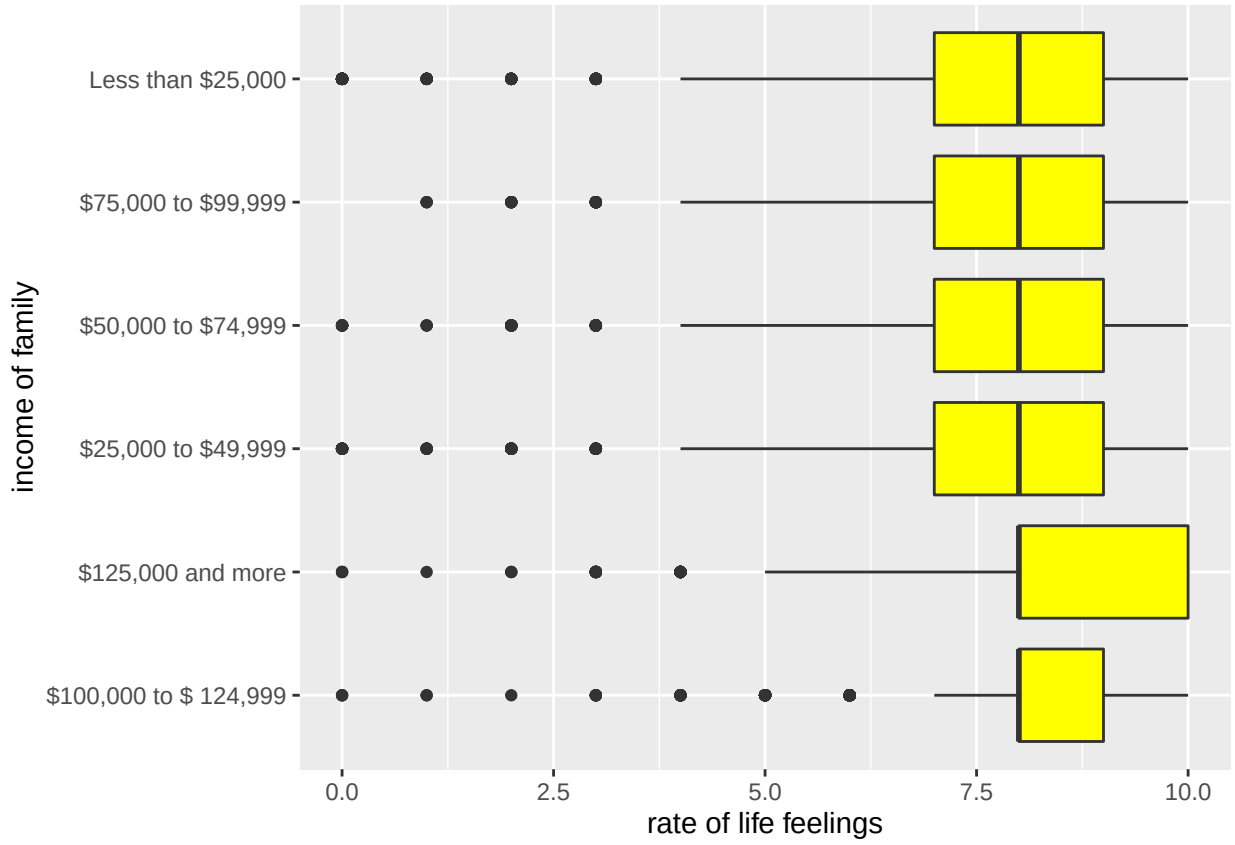
The model indicates that for every one extra home member, the rating increases by 0.099 on average, the rating increases by 0.0107 on average when individuals are one year older, and the rating decreases less when their family income increases.

Where (income_familyLess than $25,000) = 1 if one's family income is less than $25,000 and 0 otherwise. (income_family$75,000 to $99,999) = 1 if one's family income is less than $99,999 greater than $75,000 and 0 otherwise. (income_family$25,000 to $49,999) = 1 if one's family income is less than $99,999 greater than $75,000 and 0 otherwise. (income_family$50,000 to $79,999) = 1 if one's family income is less than $99,999 greater than $75,000 and 0 otherwise. If the family income group is $100,000 to $ 124,999 then all the other components of family income = 0.

Here, group $100,000 to $ 124,999 is a base case.

Caveats: this model only describes the pattern found in the 2017GSS data, hence, it may vary from time. The model has several limitations in predicting one's rating about their life feeling. In the model, we are only considering three aspects of the individual, but clearly, there are a great many other factors that determine it. Therefore, a perfect model is not likely to be achieved, instead, we could add more explanatory variables to the model such as marital status, educational background, etc.

Results



Variables	Estimates	p-value
intercept	7.5661716	< 2e-16
age	0.0107026	< 2e-16
income_familyLess than $25,000	-0.8310879	< 2e-16
income_family$50,000 to $74,999	-0.2493725	1.63e-08
income_family$25,000 to $49,999	-0.4615397	< 2e-16
income_family$75,000 to $99,999	-0.1475670	0.00131
income_family$125,000 and more	-0.2493725	1.63e-08
hh_size	0.0996582	< 2e-16
adjusted R-Squared value = 0.04468		

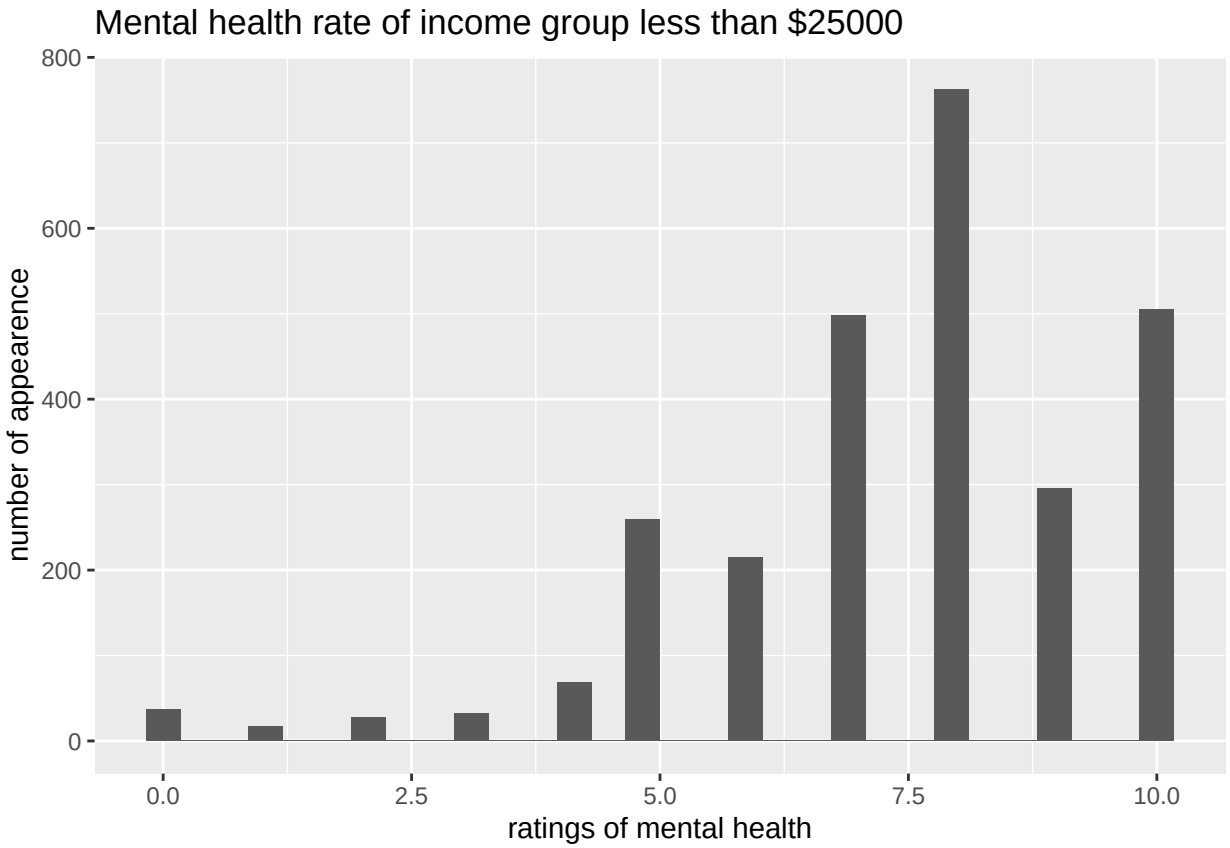
From the box plot above, we see that individuals with higher family income generally have a higher minimum of rating (the left side of the whisker). However, we can clearly see that the rating has to be explained by some other factors since the span of the rating among different income groups are huge, basically all the way from 0 to 10.

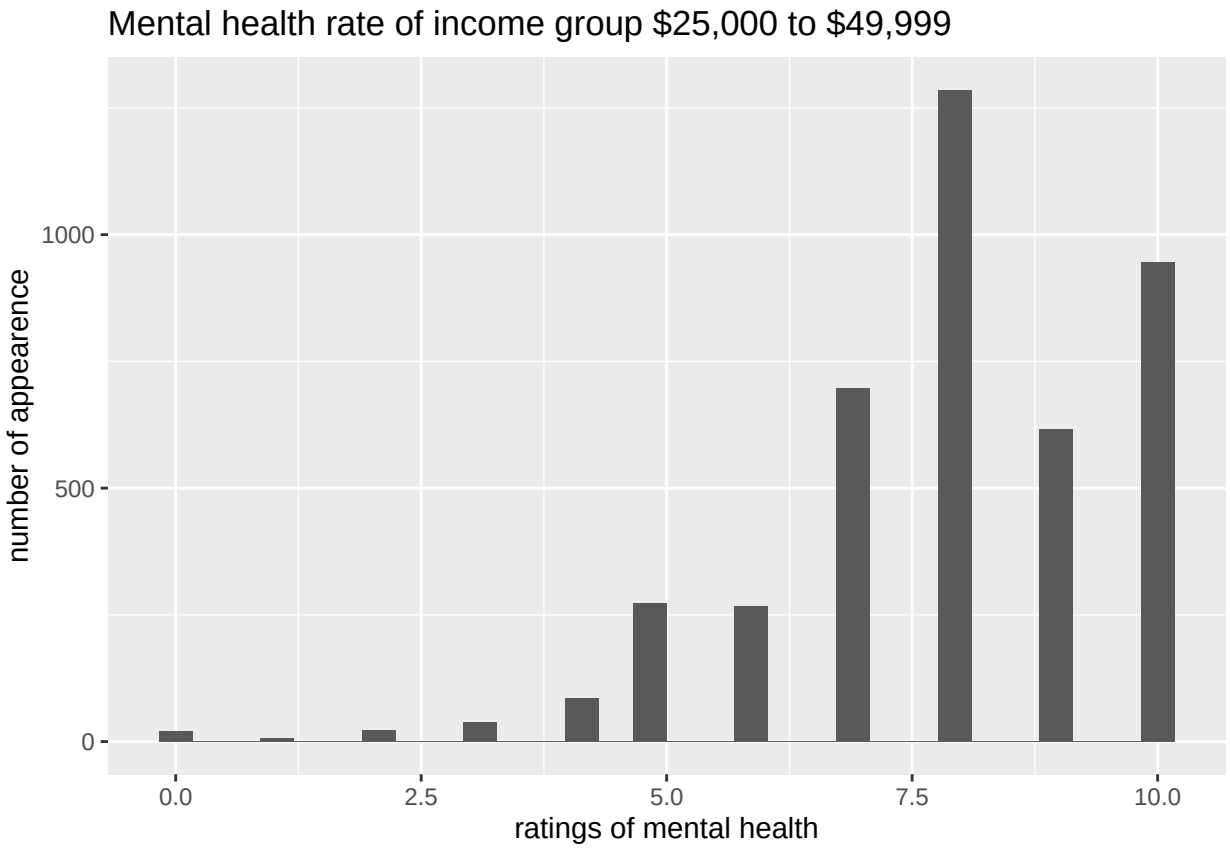
We then look at the summary table and see the p-value of each explanatory variable. We see that most of the variables' p-value are less than 0.05, indicating that there's a linear relationship between the rating and these variable. The only exception occurs in the group of family income greater than $125,000, which is quite interesting and will be discussed later. Additionally, the intercept has a value of 7.566 and its p-value is less than 0.05, meaning that we predict one's rating about their life feelings by adding or subtracting from 7.566 given the explanatory variables we know.

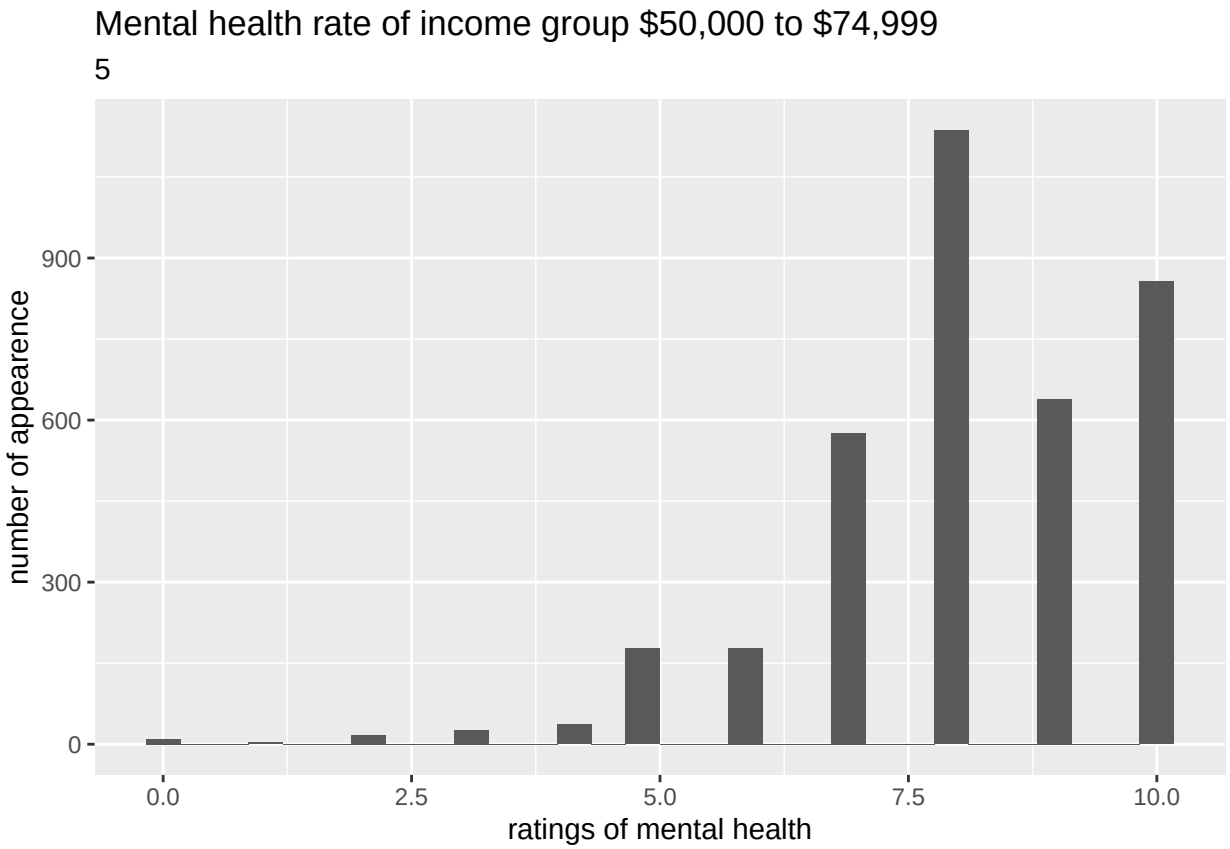
Lastly, we calculate the R-squared value to be 0.04468, which is quite small. R-squared value shows the

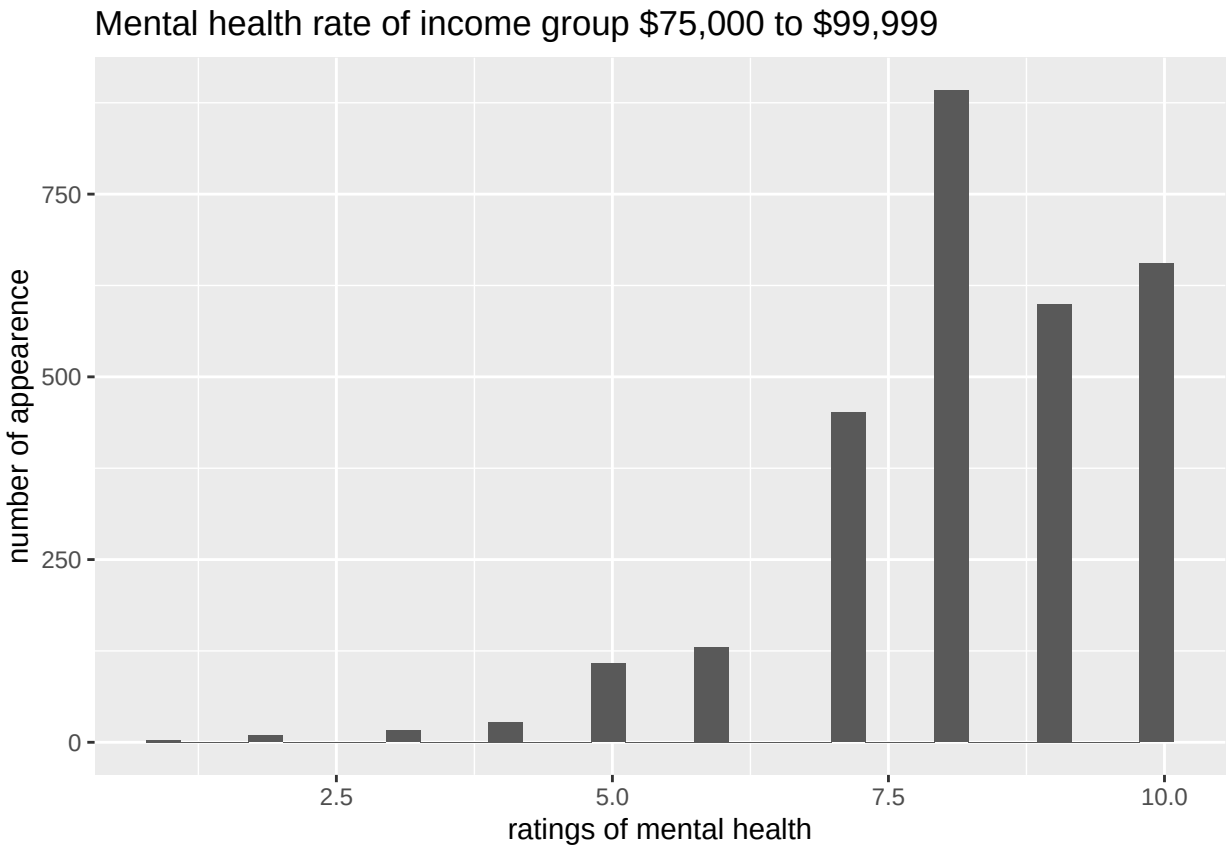
percentage of the variance of the rating explained by the three chosen explanatory variables(age, family income, number of households). The small R-squared value shows that our model doesn't fully predict the expectation of one's rating about their feeling of life.

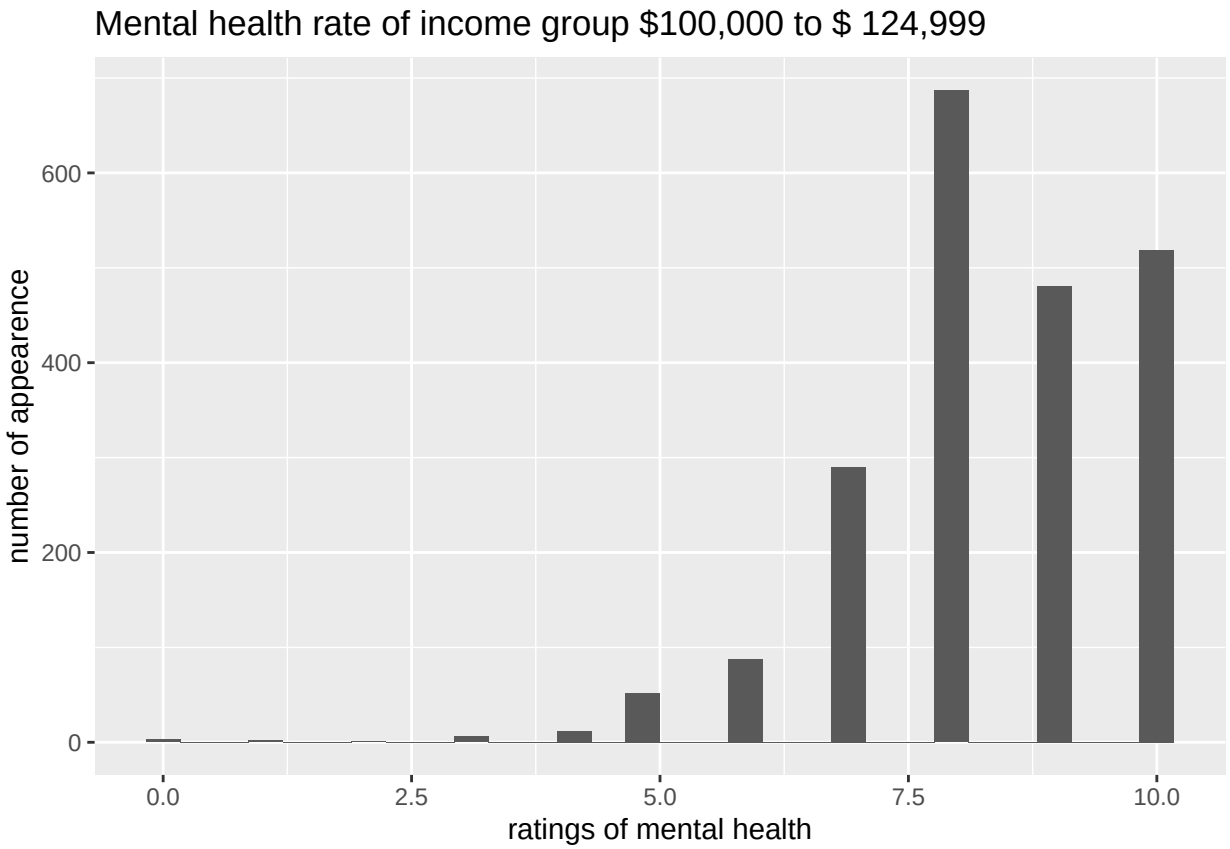
Discussion

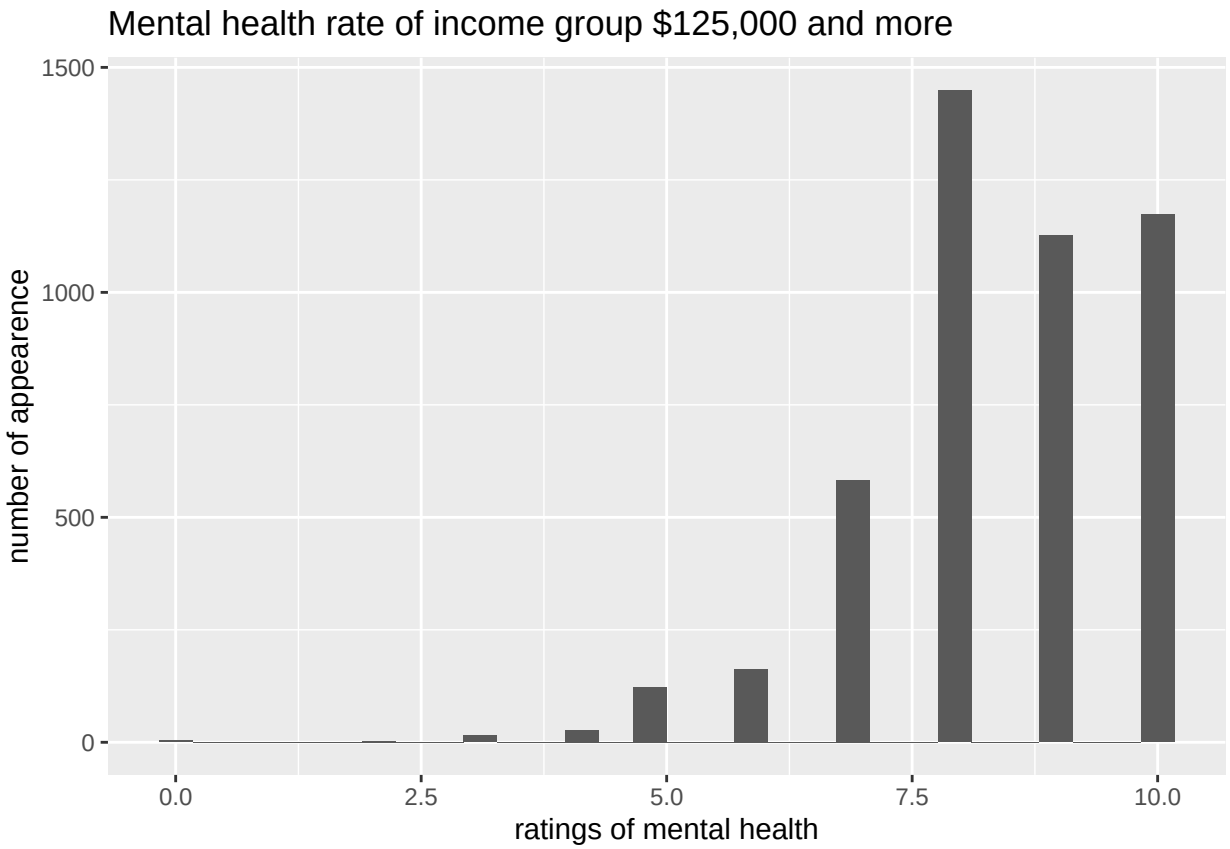

Mental health rate of income group less than $25000
number of appearence
800
600
400
200
0
0.0
2.5
5.0
7.5
10.0
ratings of mental health


Mental health rate of income group $25,000 to $49,999
number of appearence
1000
500
0
0.0
2.5
5.0
7.5
10.0
ratings of mental health

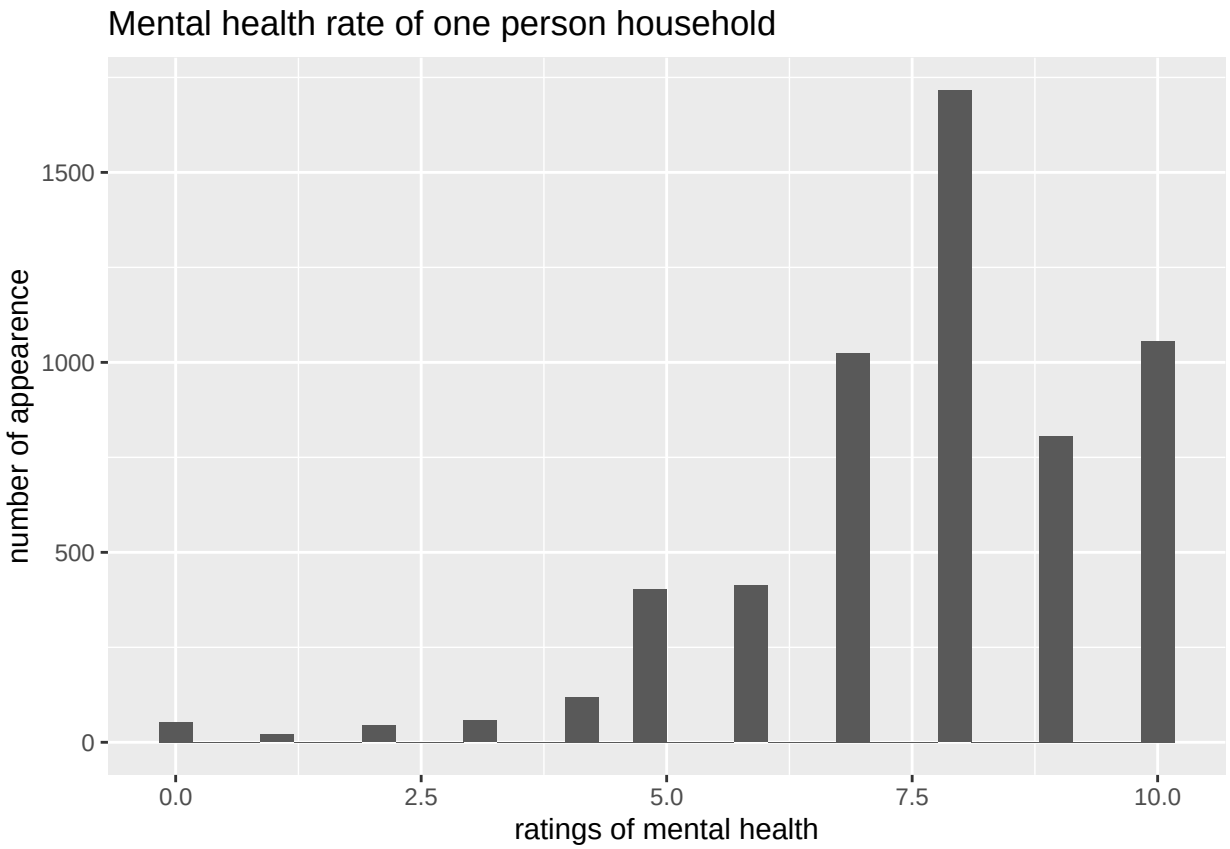

Mental health rate of income group $50,000 to $74,999
5
number of appearence
900
600
300
0
0.0
2.5
5.0
7.5
10.0
ratings of mental health


Mental health rate of income group $75,000 to $99,999
number of appearence
750
500
250
0
2.5
5.0
7.5
10.0
ratings of mental health

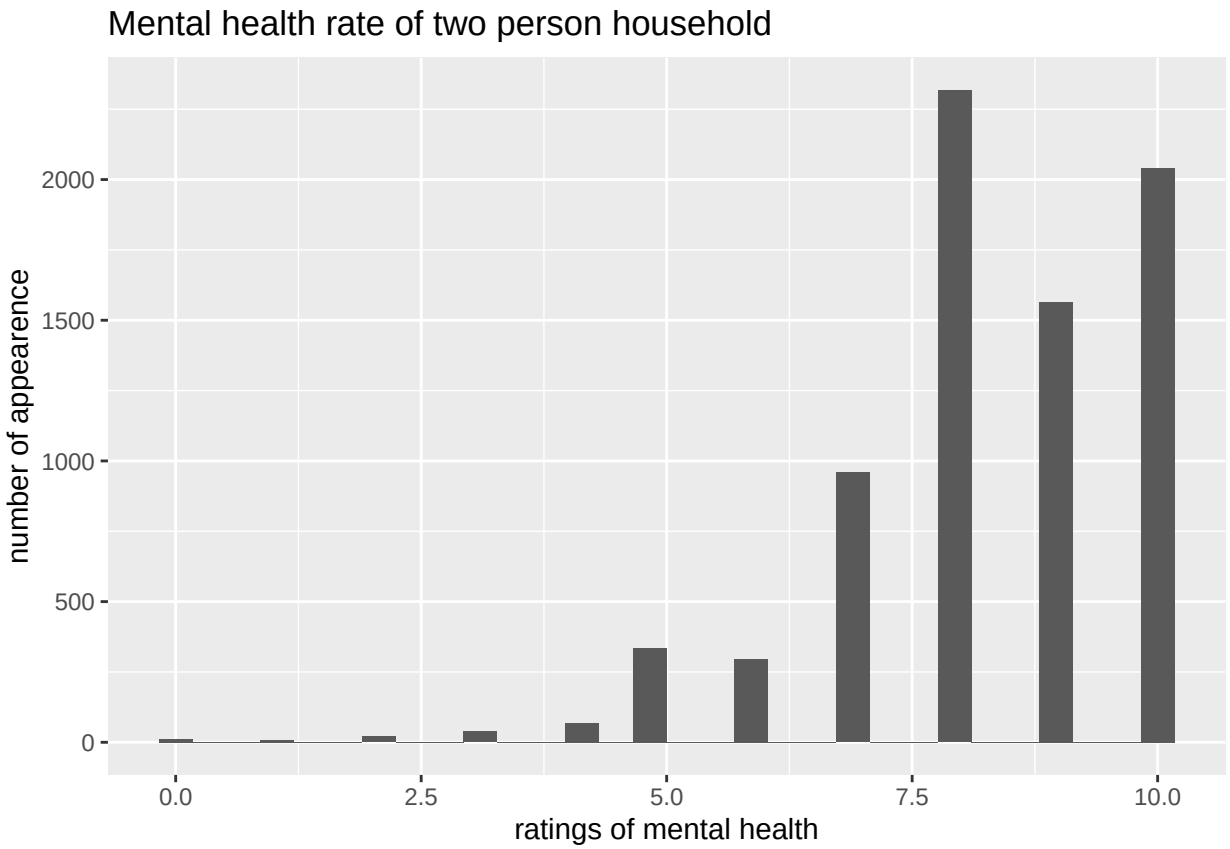

Mental health rate of income group $100,000 to $ 124,999
number of appearence
600
400
200
0
0.0
2.5
5.0
7.5
10.0
ratings of mental health



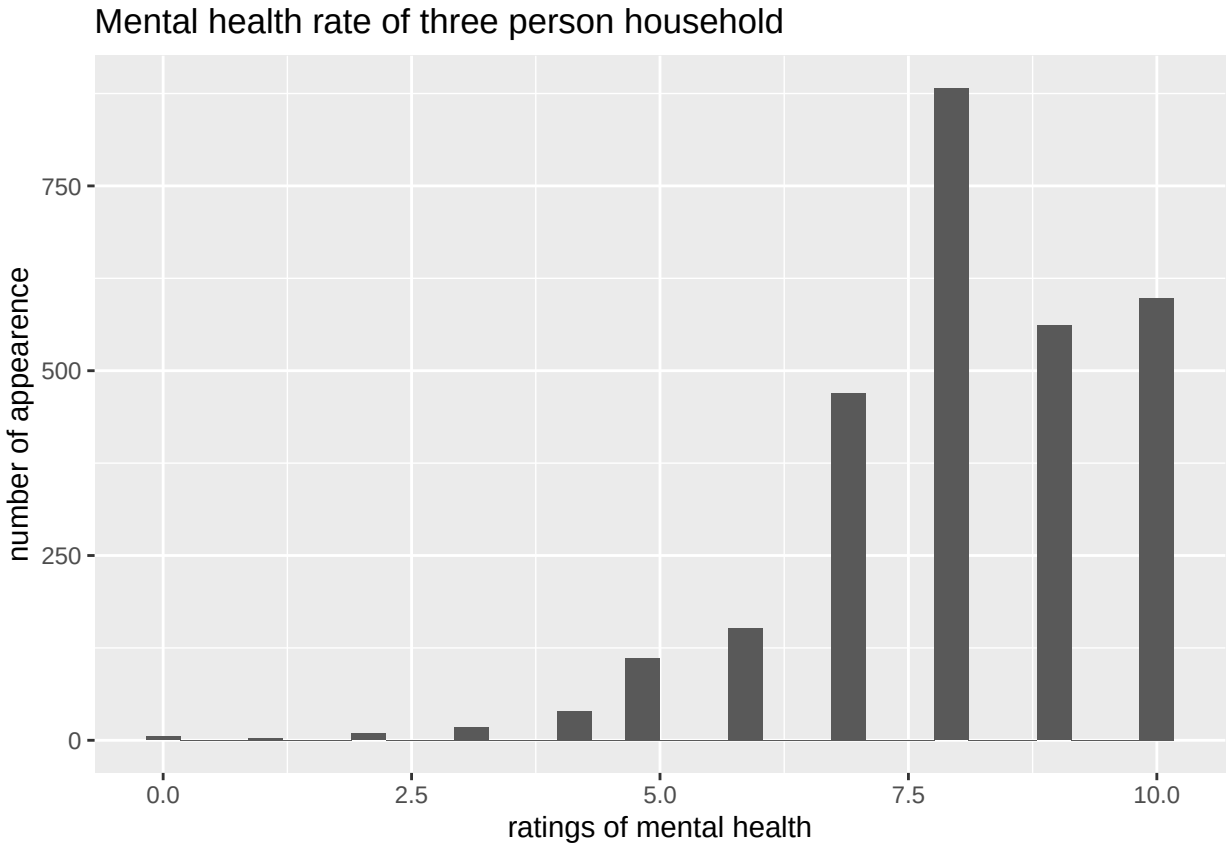

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



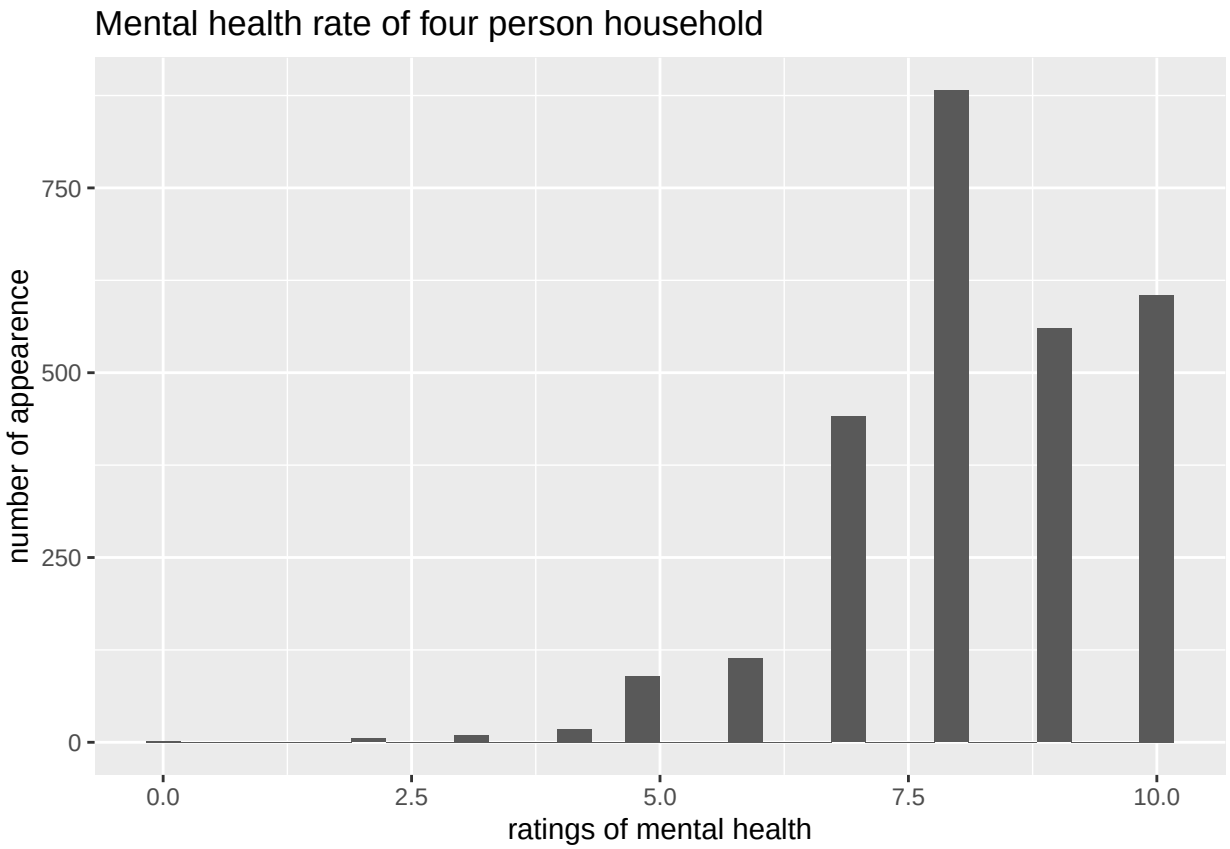

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



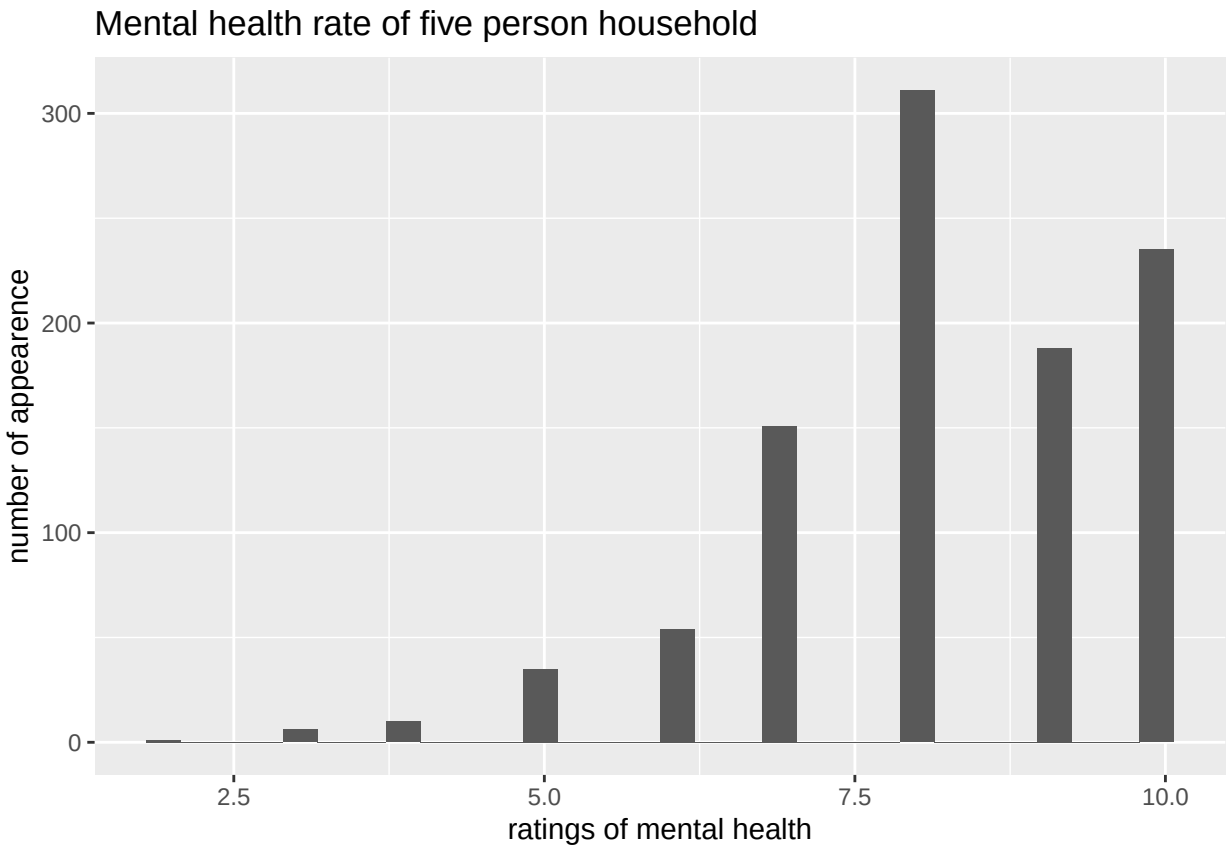

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



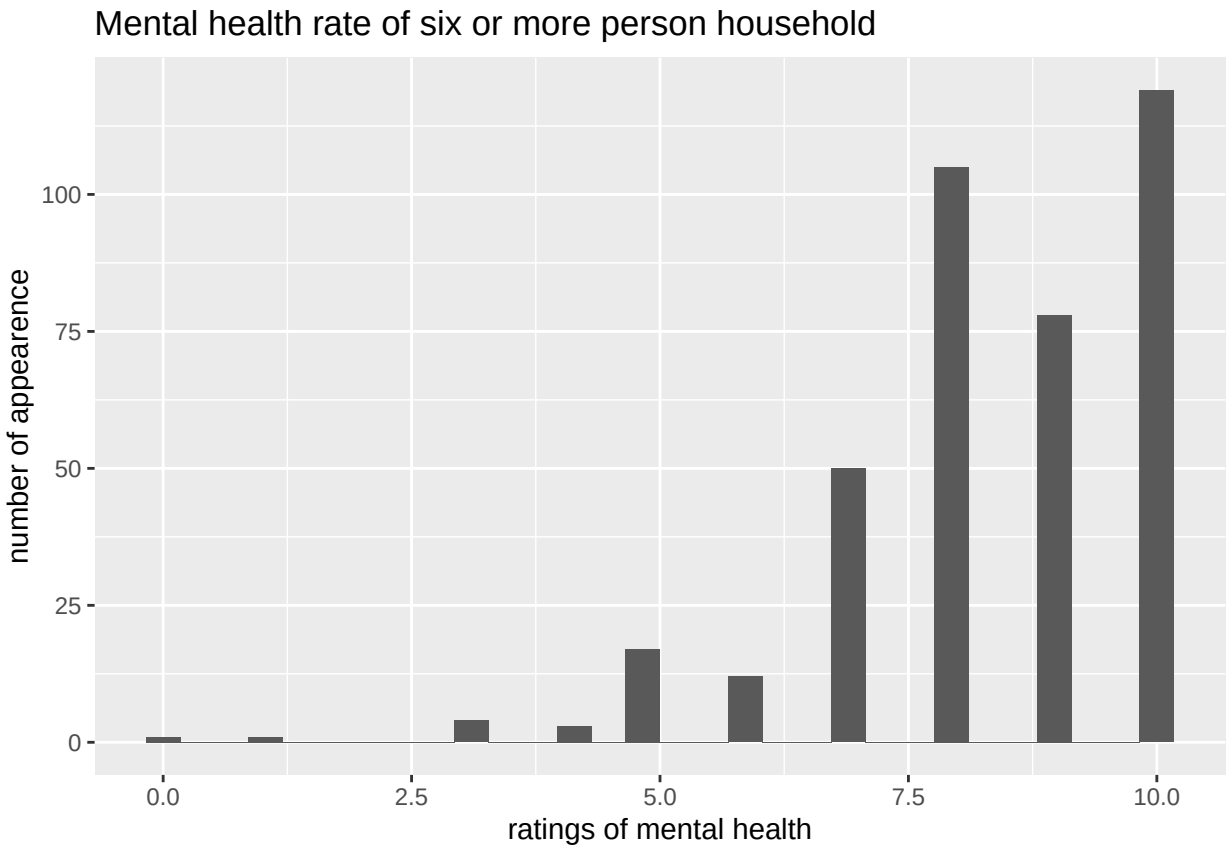

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## [1] 7.693963
```

```
## [1] 8.307743
```

```
## [1] 8.100035
```

```
## [1] 8.243124
```

```
## [1] 8.21998
```

```
## [1] 8.369231
```

The mental health rate is related to all three variables that we are interested in according to the model. Generally speaking, the older an individual, the more family member the individual has, and the more income will lead to a higher mental health rating. However, both age and number of a family member have a relatively small impact on the rating. With the rating increases by 0.0107 on average when individuals are one year older, age is not much of an influence because human life expectancy is around 80 years old and from 0 years old to 80 years old the rating will increase by less than 1 on average. So the effect of age on mental health rating is pretty small according to our model.

The same goes for the number of family members. Because in reality, a family can only have a certain number of members unless when it comes to a newborn baby. But that can't affect the rating much because according to the model, for every extra member the rating increases by 0.099, which is relatively small even after it times the number of kids the family is going to have. The histograms of mental health rating with

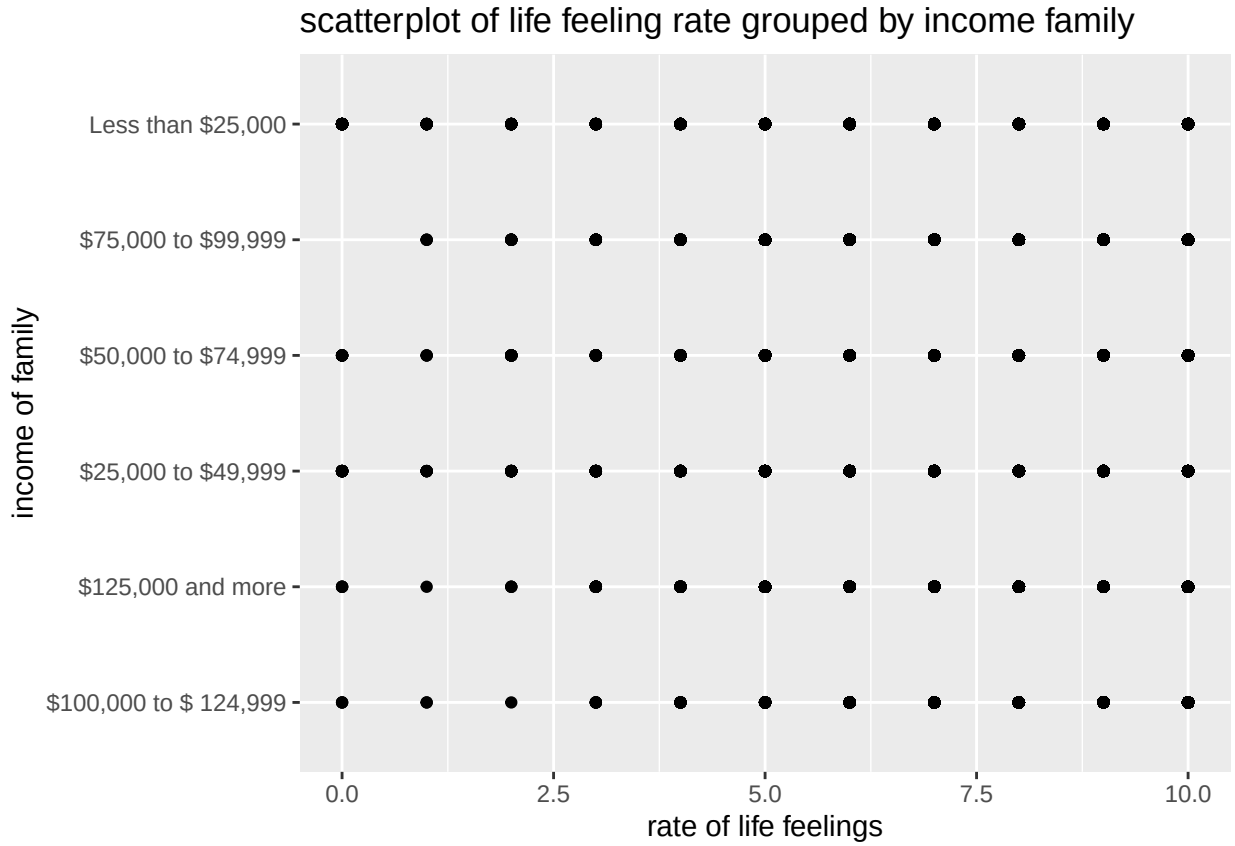
different number of persons in household also shows that. There is no obvious trend of people rating it higher as family member increases. Actually according to the mean rating of different groups, we saw a drop in rating from two person household to three person household and from four person household to five person household. But one obvious conclusion is that one person house hold has lower rating than other household groups. This can be seen from both the histogram and the mean. We can see an obvious increase in high ratings and decrease in low rating when comparing one person household to others. And one person household is the only group with mean rating less than 8 with only 7.69 in average.

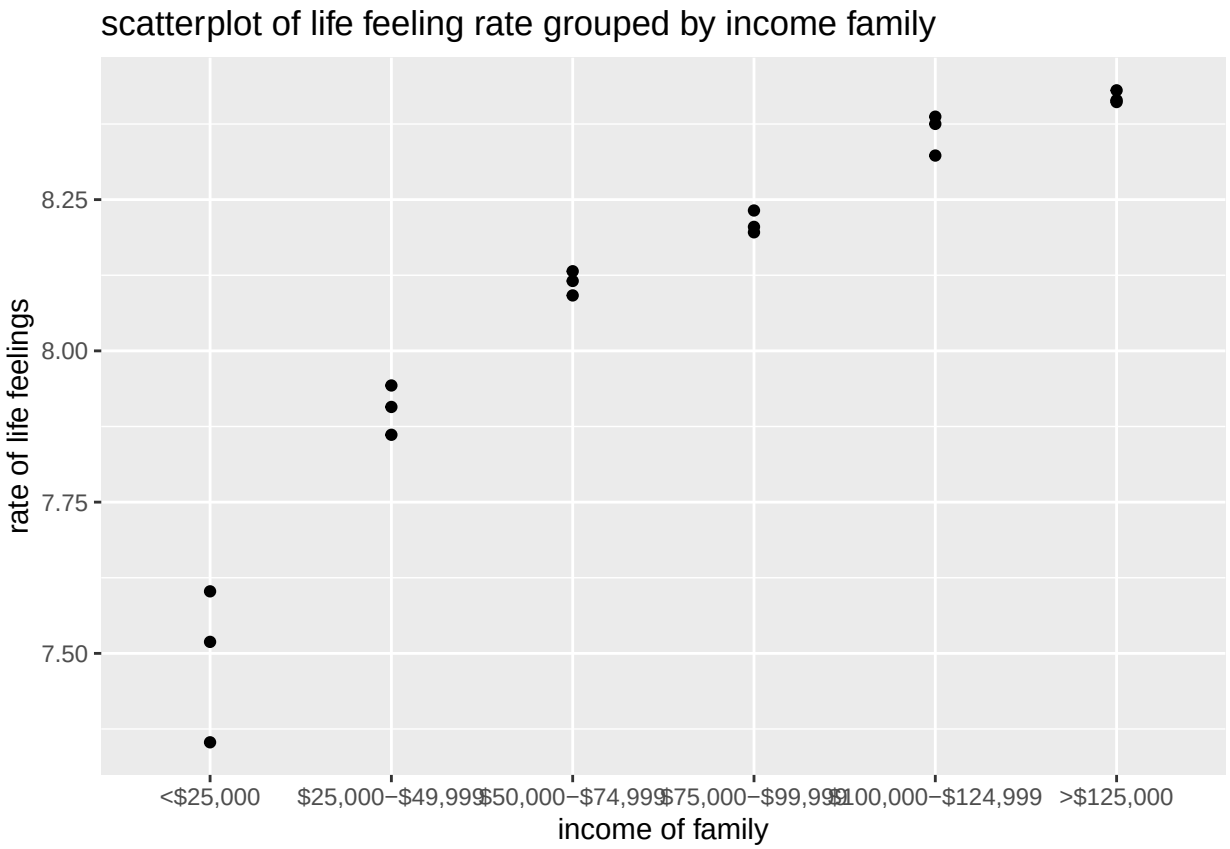
As for income, from the histograms above we can see clearly that as income gets more, the percentage of high ratings(7.5 and above) steadily increases, and low ratings(less than 5) steadily decrease. So income and mental health ratings are positively related. The more income people have better mental health they tend to have.

However, our model can still serve as a reference for people on certain issues. Keeping the mental health rate high is very important for an individual. With better mental health an individual can be more productive and happier in life. So our model can be helpful for people who are trying to make a life decision such as whether or not to have another kid(another family member) or is it worth accepting this job with a higher salary but need to move to another city.

For example, from our model people can know that with a reasonable income, it is a good idea to have another kid without needing to worry about more stress. Because the mental health rating will generally increase with one more family member.

Weaknesses





The model has its flaws. First of all, mental health is a very subjective topic. Whether it can simply be mathematically measured using numbers is in doubt. This is the problem about the original survey, which we will talk about how to improve in Next Steps.

Second, as the first scatterplot shows, the number of ratings is all integers, and there are a lot of samples so every possible answer is covered. Thus the scatterplot for this survey barely tells any information about our topic of interest. This is the problem with all integer rating survey. When the sample is fitted into the model, we have coefficients such as 0.099 and 0.0107. But what do they represent when all the answers are integers. To improve this, I did some data analysis. I separated every income group into 3 subgroups with an approximately equal number of samples and get the mean of each group. I then plotted the scatterplot for these sample means and arranged them according to value on the y axis which is the rating. This way we can see the relationship between income and rating because the numbers are not integers anymore and now we see a clear pattern on the scatterplot that the higher income, the higher mental health rating. So income and rating are positively related, we can now fit a regression line to it.

Next Steps

Next, we can try different models on our data, maybe conduct a Bayesian analysis on the old model by survey more people with the same questions and get more data to correct our model. The frame population has not covered the target population and in fact, is far off. So we could get a lot more data if we find the correct method. And since the old data is not very equally distributed among different age groups, we can adjust it through the new data to fit the percentage of data in each age group the same as the actual Canadian population's age structure.

To better analyze the relationship between the number of family members and the rating, we could separate the number of family members by groups such as 2 and under, 3 to 4, 5, and above. Ask them more direct

questions about the relationship such as:do you think you are stressed about supporting your household? Do you think you have enough personal space at home? And then conduct the same statistical analysis to all different groups. Because it is very possible that for fewer member families, one more family member will increase the mental health rating. But what about a family with already 6 people? Then one more member may be more stressed and cause a lower rating. So it is a good thing to separate them into groups and then discuss them.

For the next steps, we could also conduct another survey, this time focus less subjective and more objective ways to measure mental health. Such as how often do you have insomnia and how often do you see a psychologist. These questions have more quantitative and objective answers that are not influenced by subject feelings when one 'rates' his/her own mental health. We can also ask are you more stressed now compare to a year before and 5 years before. This question is more direct about the relationship between age and mental health because it is a yes or no question, we can better know each age group feels about getting old and to see if they are actually related. Because from the survey now we get that the older the higher one's mental health rating is. But the number of the coefficient is very small. So we want to find out are they actually related to the new survey.

References

1. "Age distribution in Canada 2009-2019", Published by H. Plecher, Oct 7, 2020, 2019https://www.statista.com/statistics/ distribution-in-canada/ Accesed on 10/18, 2020

2. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

3. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686.

4. Lumley T (2020). "survey: analysis of complex survey samples." R package version 4.0. Lumley T (2004). "Analysis of Complex Survey Samples." Journal of Statistical Software, 9(1), 1-19. R package verson 2.2. Lumley T (2010). Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R. John Wiley and Sons.

5. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

6. Guide book of 2017 General Social Survey (GSS): Families Cycle 31 https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

7. "Canadian general social surveys (GSS)", By Statistics Canada under the terms of the Data Liberation Initiative (DLI), https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

appendix

All the code and the report can be found at https://github.com/wethanl/STA304A2