A statistical suggestion on living happier for Canadians

Zhiming Huang, Yidou Wang

2020-10-19

Abstract

This study mainly invesigated the factors affect living quality for Canadians based on a reliable national survey data using statistical methods including linear regression models. The finding shows that keep good body and mental health, be married or living common with loevers. have more total children and has grandchildren as well as work more than 50 hours weekly with high income could make people living much happier. Also, there are lots of other suggestions on living happier for Canadians could be made based on the findings of this study. The government could make better decisions to improve the quality of life for the citizens by making better conditions for the suggestions of the study.

Introduction

In recent years, life quality becomes a very important indicator for people. A high quality of life is now becoming a basic requirement for lots of people and countries. However, there are many factors affect quality of life such as health, income, education levels and so on. This study is aimed to invesigate the issues of effects of different factors on the quality of life measured by people's feeling score of life. This study mainly used graphical and model summarises to investigate the topic. It was found that male, education level not below bachelor, poor body health, poor mental health, citizenship by naturalization not by birth, single martial status, rented house not owned all leading to show lower feeling score of life while not living alone, have more total children and has grandchildren make people living happier with higher feeling score of life. Also, average hours worked 50.1 hours and more as well as family income with 125,000 dollars and more make people living happier. These findings are important because it would help the government to understand the profiles of groups with high or low feeling score of life and to make better decisions in improving the quality of life for the citizens. At last, the study first give an overall introduction followed by data and model dicussions, then results and dicussions of findings are given. The link to the study is: https://github.com/Zhiming-Huang8/PS2_304/blob/main/ps2_304.pdf.

Data

The source of data is the 2017 Canadian General Social Survey Data and the source code to clean the data provided by Alexander and Sam Caetano (2020). A subset of the cleaned data is used in this study which includes about 12000 instances with over 10 attributes. The response variable is the feeling score of life scaled from 0 to 10, the factors are mianly including age, gender, education, martial status, family income, weekly working hours, number of children, number of grandchildren, body and mental health. Some of the categorical factors are recoded in this study to simplify the original variables. For example, education level is only recoded into not below bachelor and below bachelor, more details of data cleaning procedure could check the source files of the study which is hold in the Github repo link https://github.com/Zhiming-Huang8/PS2_304.
.

The questionnaire of the survey has both good and bad aspects, the good ones including that it is a well-designed, well-tested survey which has a very high quality. The bad one is that there are lots of questions in

the survey which is easily leading non-responses.

The target population of the survey is all of the non-Canadians older than or at least equal to 15 years old. The frame is a list of telephone numbers plus dwelling frame. The samples are units collceted in the survey. This survey applied a stratified sampling approach, the strata are the geographic areas of 10 provinces in Canada.

There are non-reponse bias in the survey. First, there are population with no telephone numbers that could not be studied. Second, there are no answers for some questions but have answers for most of other questions. Because of the cost of the survey is huge due to it is a national wide one, so a trade-off is to keep the answered questions not changed but questions without answers imputed by estimations.

Model

The model used in this study is the linear regression model which has the following form:

$$Y = \alpha_0 + \alpha_1 x_1 + ... + \alpha_k x_k + \epsilon$$

where Y is the response variable feeling score of life in this study. X_i is factor such as age, gender and etc. Note that for categorical factors with more than two levels, there are more than one dummy variables in the model. And ϵ is the error term which is assumed to be i.i.d. $\sim$ N(0,σ^2).

The linear model is choosen because although the response feeling score of life is not continuous, it is ordinal response. Using linear regression, the results could be directly interpreted which matchs the goal of the study. Other models like bayes models could also be used, but as the data is large, estimations of these complicated models are big problems. This study uses R software to run the linear model and perform model diagnostics. The model diagnostics in this study mainly performed by using model diagnostic plots. The assumptions to be checked mainly including: independence, linearity, normality and constant variance assumptions. At last, for this study, there is no covengence problem.

Results

The results section mainly include graphical summarises and model summarises. Figure 1 shows the distributions of feelings score of life grouped by gender, it seems no difference between female and male. Figure 2 shows the distributions of feelings score of life grouped by education level, it seems lower education level living happier. Figure 3 shows the distributions of feelings score of life grouped by whether living alone, it seems not alone living happier. Figure 4 shows the distributions of feelings score of life grouped by citizenship status, it seems by birth living happier. Figure 5 shows the distributions of feelings score of life grouped by own or rent house, it seems owned house living happier.

Figure 6 shows the distributions of feelings score of life grouped by whether has grandchildren, it seems has grandchildren living happier. Figure 7 shows the distributions of feelings score of life grouped by marital status, it seems married, living common living happier. Figure 8 shows the distributions of feelings score of life grouped by body health, it seems Very good and Excellent living happier. Figure 9 shows the distributions of feelings score of life grouped by mental health, it seems Very good and Excellent living happier. Figure 10 shows the distributions of feelings score of life grouped by average hours worked per week, it seems 30 to 40 hours living happier. Figure 11 shows the distributions of feelings score of life grouped by family income level, it seems 125,000 dollars and more living happier.

Table 1 shows the estimates of full model including all of the interested factors. However, due to the findings in graphical summarises, there are some factors might not be important in explanning feeling score of life. Thus, based on the model, the AIC backward model selection is performed and the results are shown in table 2. It can be found factors like age are dropped which is consistent with findings in the graphical summarises.

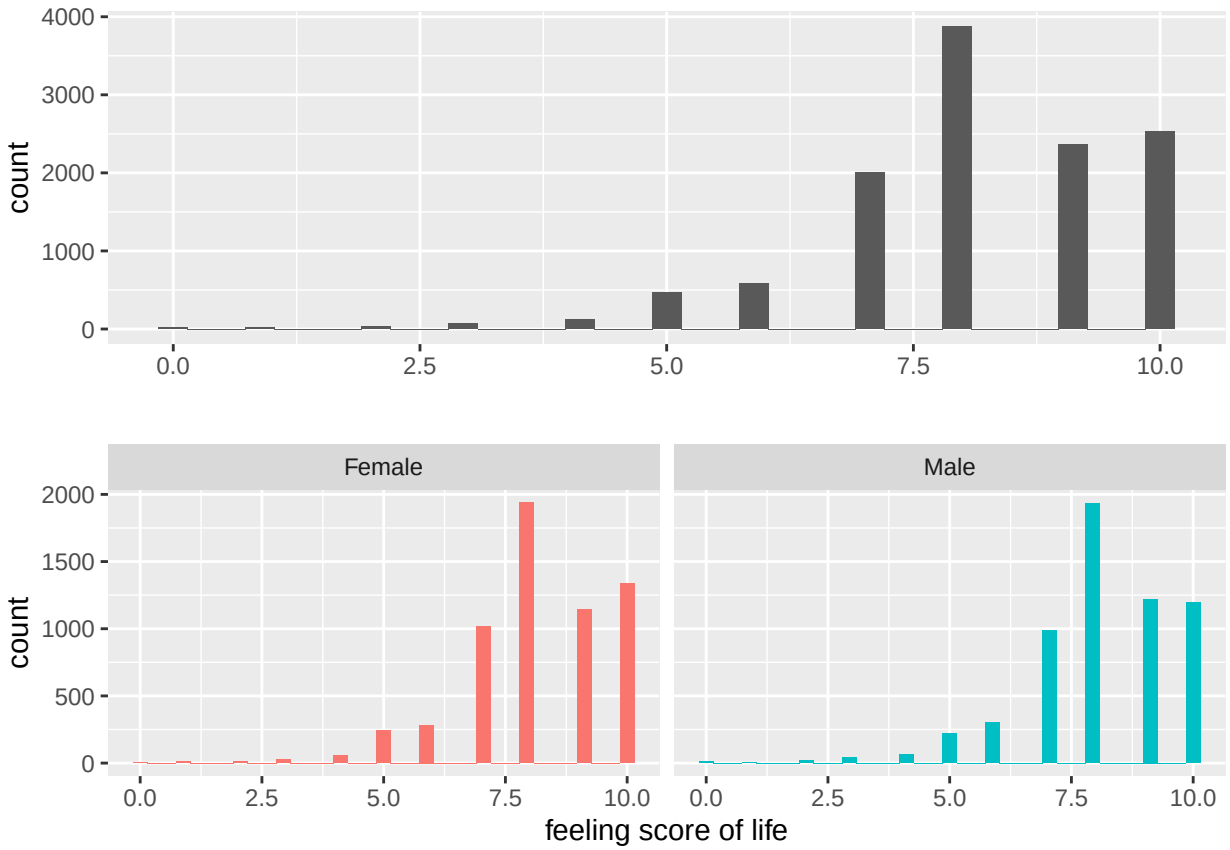

Figure 1: Distributions of feelings score of life grouped by gender

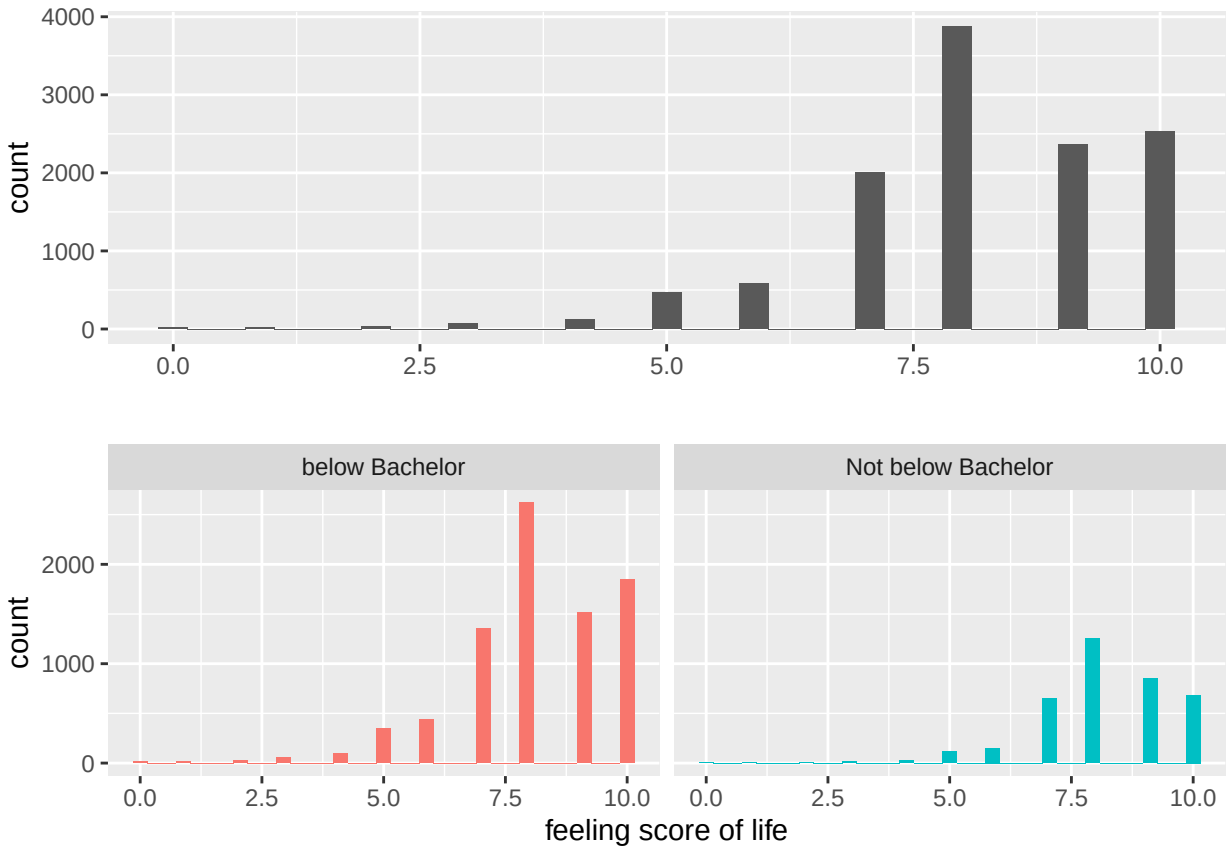

Figure 2: Distributions of feelings score of life grouped by education level

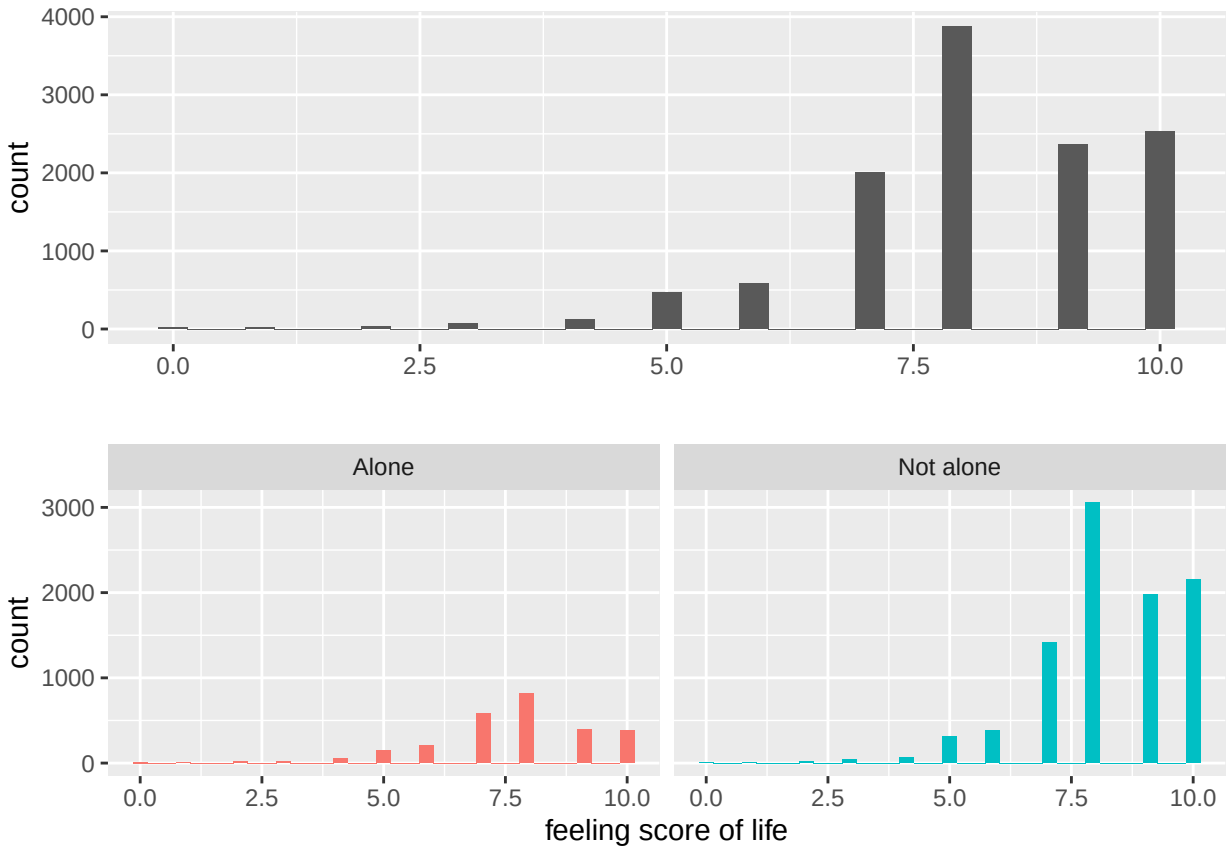

Figure 3: Distributions of feelings score of life grouped by whether living alone

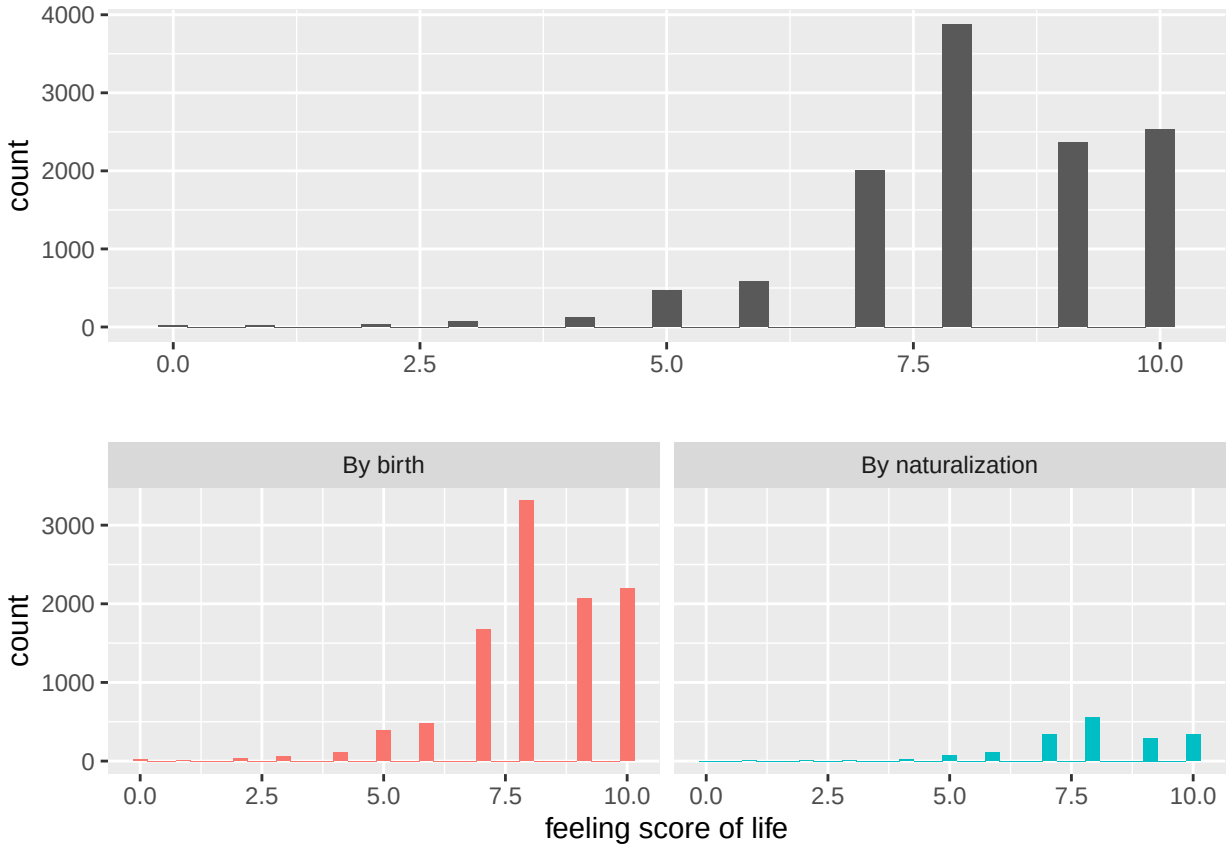

Figure 4: Distributions of feelings score of life grouped by citizenship status

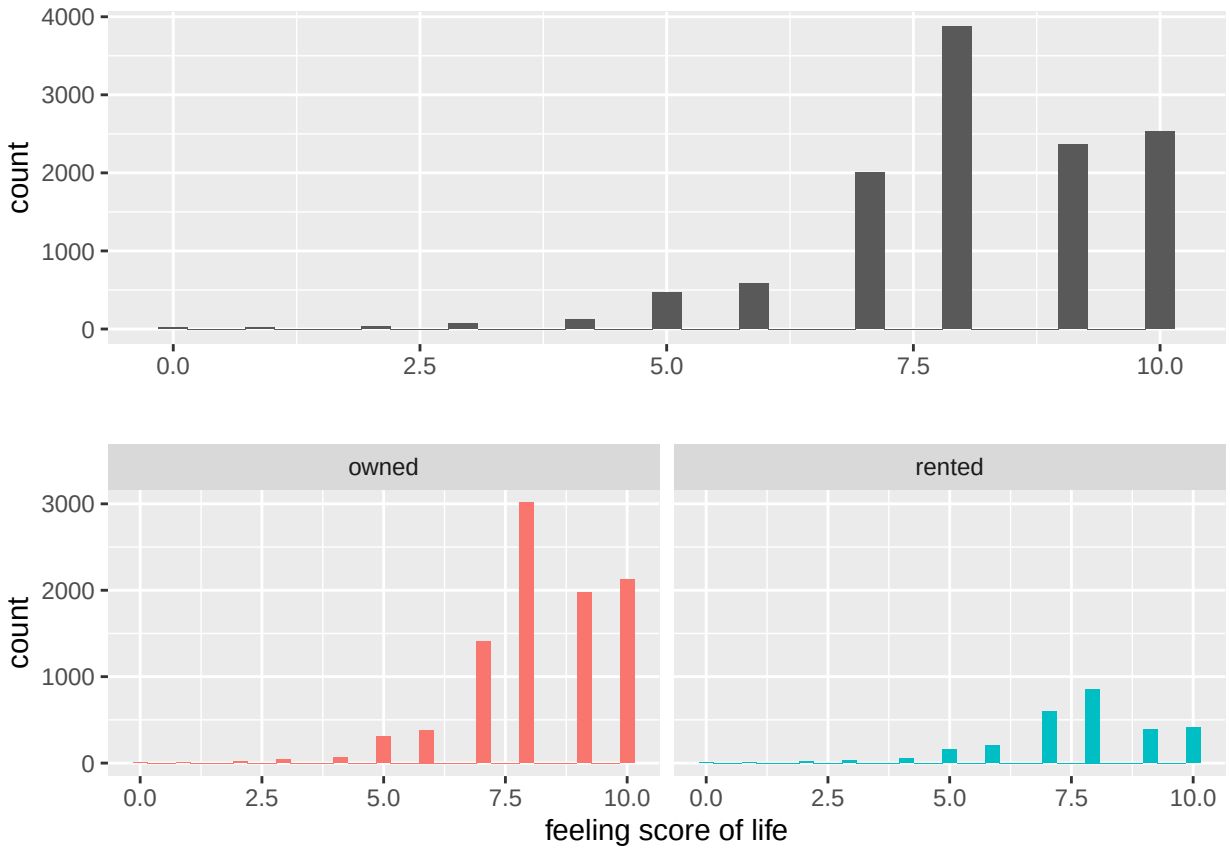

Figure 5: Distributions of feelings score of life grouped by own or rent house

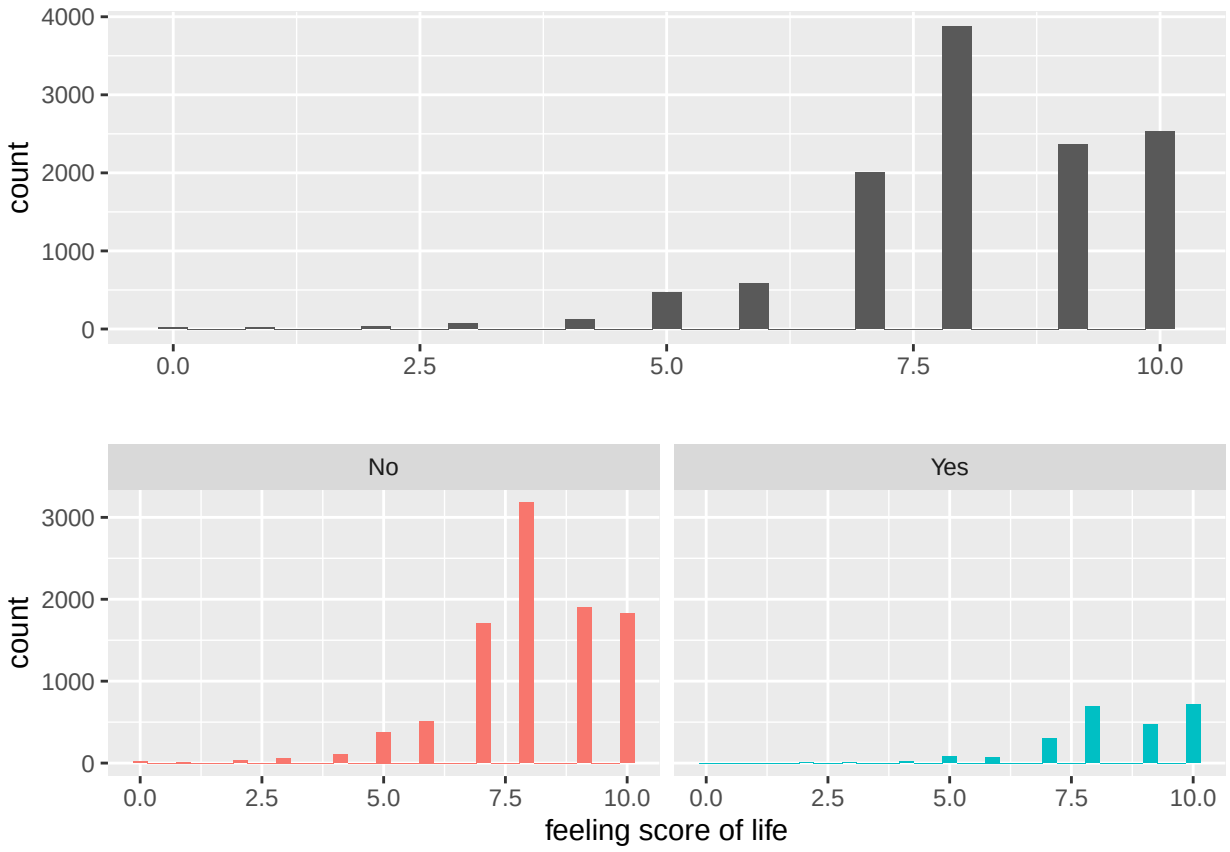

Figure 6: Distributions of feelings score of life grouped by whether has grandchildren

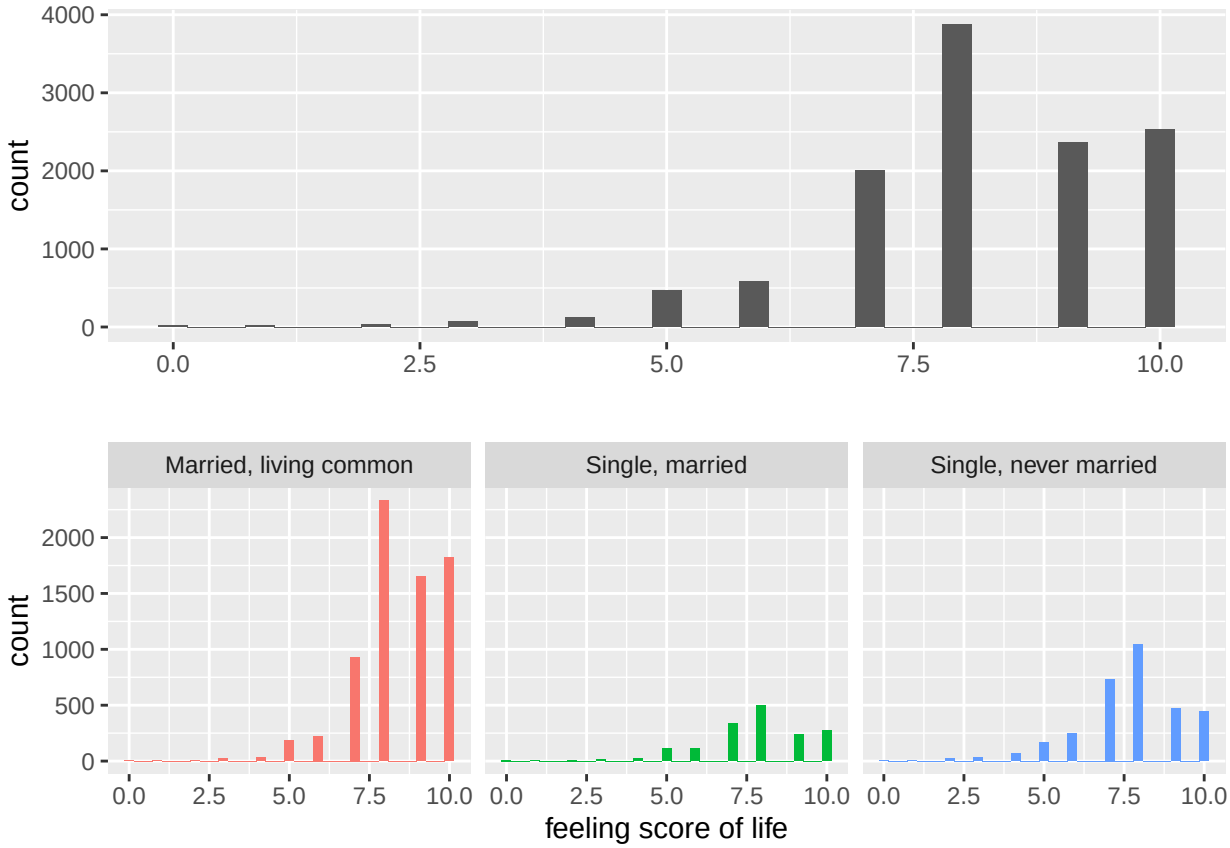

Figure 7: Distributions of feelings score of life grouped by marital status

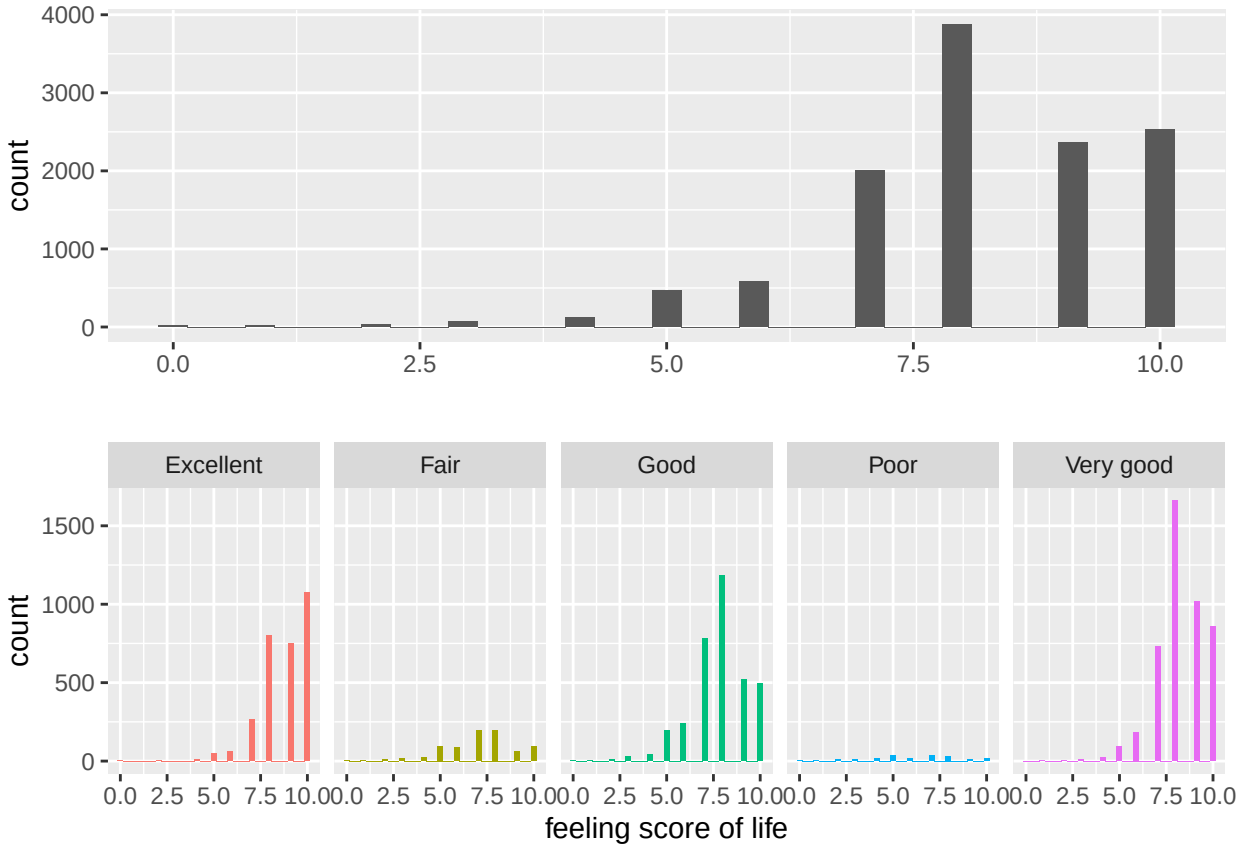

Figure 8: Distributions of feelings score of life grouped by body health

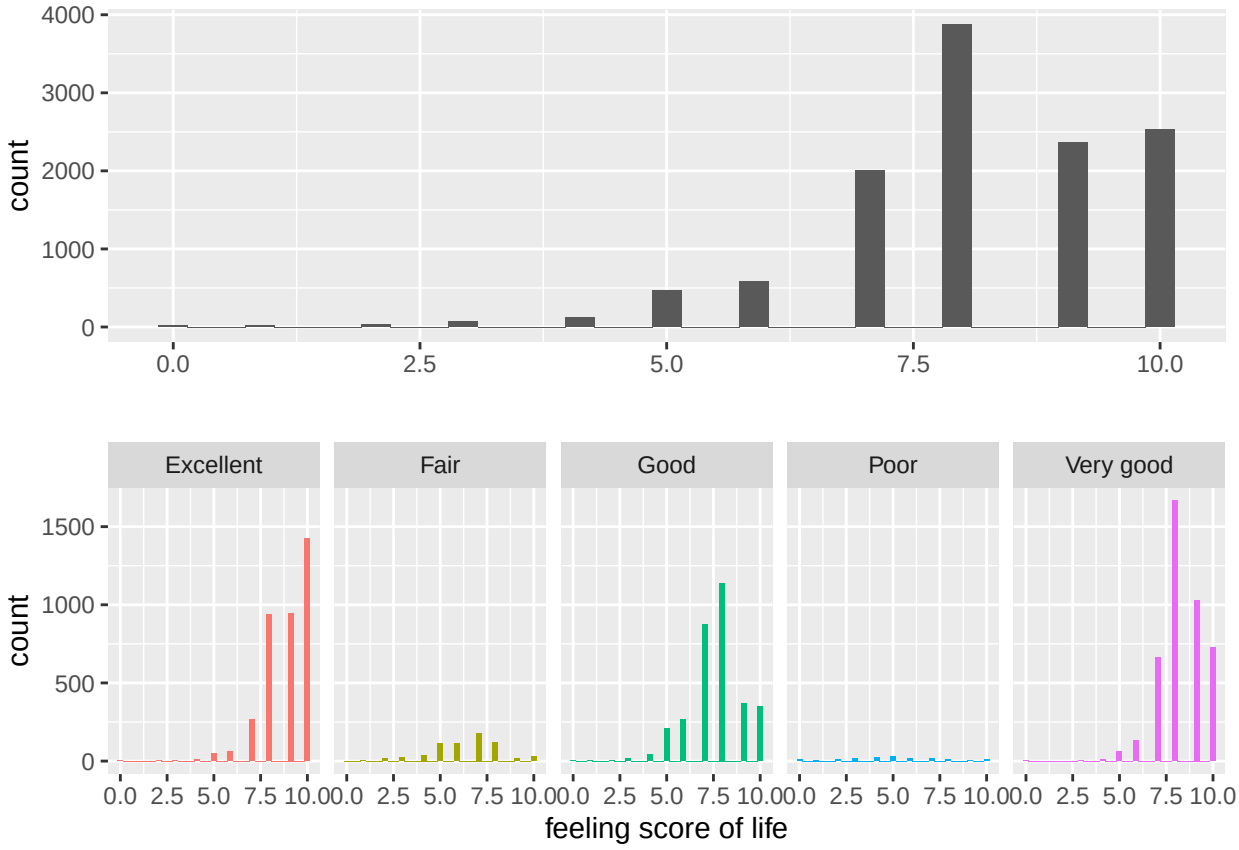

Figure 9: Distributions of feelings score of life grouped by mental health

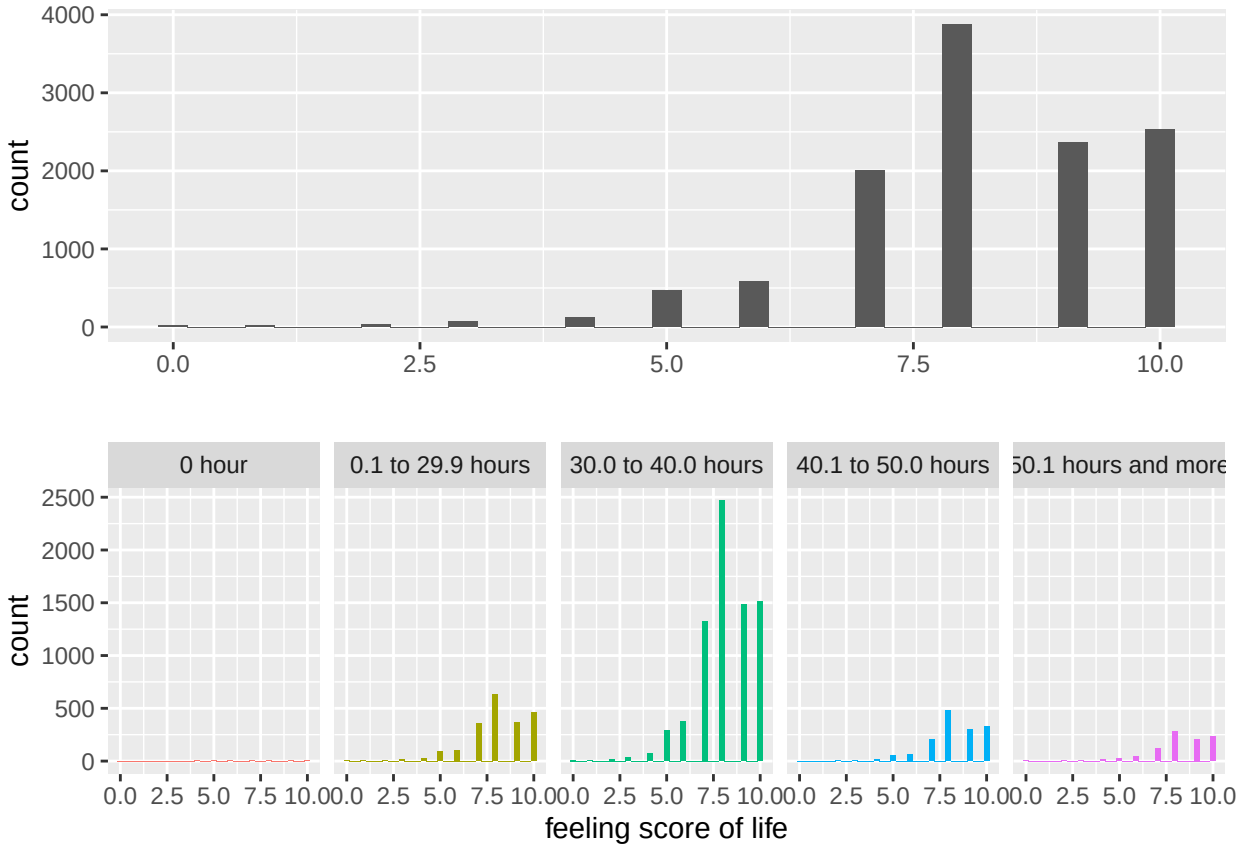

Figure 10: Distributions of feelings score of life grouped by average hours worked per week

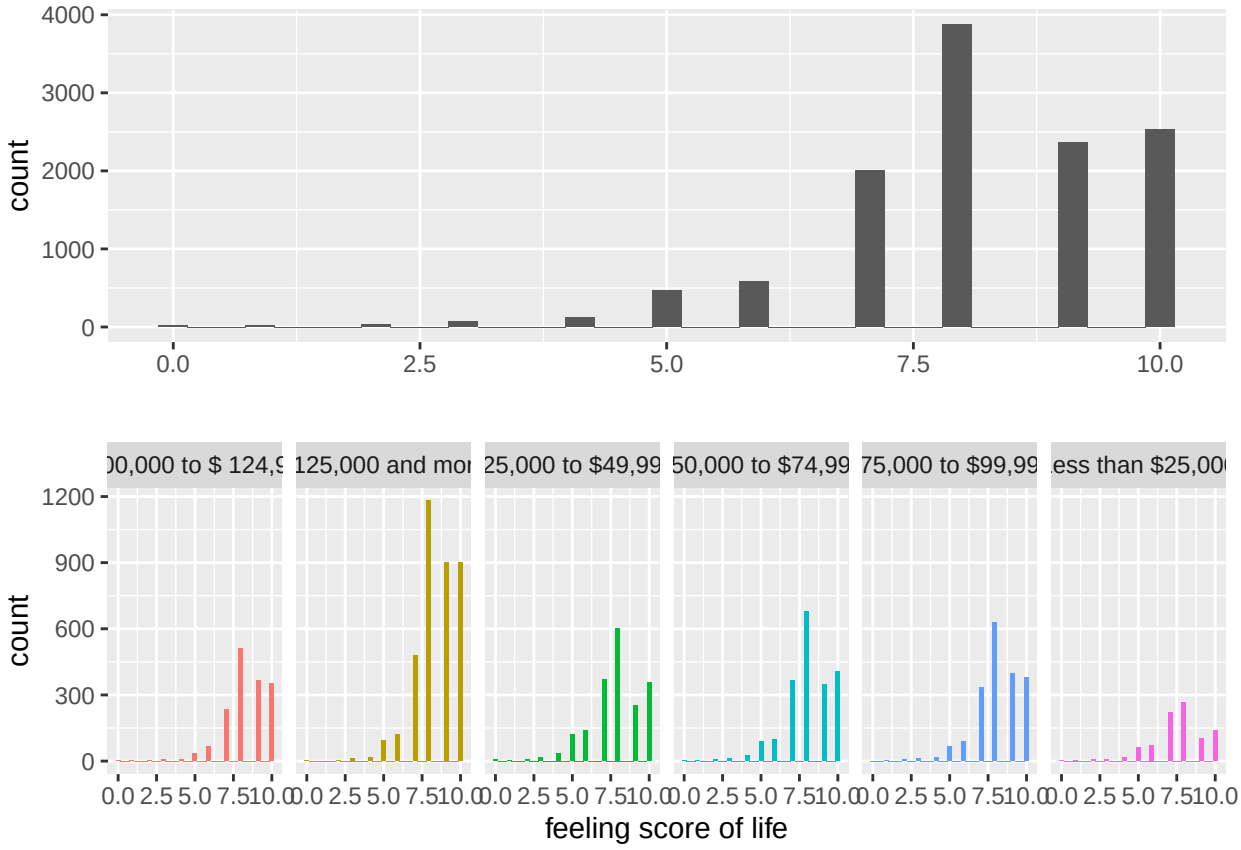

Figure 11: Distributions of feelings score of life grouped by family income level

Table 3 shows the Variance Inflation Factors of the factors in the model, it can be found all of the factors left in the model show VIF values lower than 5 which means there is no mutli-collinearity issues in the model. However, figure 12 shows the model diagnostics plots. Clearly, it can be found from the normal QQ plot that there are lots of points far from the straight line at the head which means the normality assumption is not true. By dropping these outliers, the table 4 shows the estimates and figure 13 shows the model diagnostics plots for the new model. It can be found from the model diagnostics plots that there are no serious problems for the model that linearity, independence, normality and constant variance assumptions are all satisfied, so the linear model obtained is valid.

Table 1: Estimates for full model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.77	0.28	30.94	0.00
age	0.00	0.00	-0.53	0.60
total_children	0.02	0.01	1.74	0.08
sexMale	-0.15	0.02	-6.45	0.00
educationNot below Bachelor	-0.11	0.03	-4.08	0.00
self_rated_healthFair	-0.73	0.05	-13.21	0.00
self_rated_healthGood	-0.36	0.04	-10.08	0.00
self_rated_healthPoor	-1.18	0.10	-11.87	0.00
self_rated_healthVery good	-0.20	0.03	-6.16	0.00
citizenship_statusBy naturalization	-0.16	0.03	-4.84	0.00
self_rated_mental_healthFair	-2.01	0.06	-34.62	0.00
self_rated_mental_healthGood	-1.02	0.03	-29.77	0.00
self_rated_mental_healthPoor	-3.48	0.11	-31.56	0.00
self_rated_mental_healthVery good	-0.46	0.03	-14.99	0.00
marital_statusSingle, married	-0.51	0.04	-11.93	0.00
marital_statusSingle, never married	-0.38	0.04	-10.09	0.00
own_rentrented	-0.15	0.03	-4.95	0.00
living_arrangementNot alone	0.05	0.04	1.08	0.28
hh_size	0.00	0.01	-0.06	0.95
has_grandchildrenYes	0.28	0.04	7.38	0.00
average_hours_worked0.1 to 29.9 hours	0.49	0.27	1.81	0.07
average_hours_worked30.0 to 40.0 hours	0.46	0.27	1.72	0.08
average_hours_worked40.1 to 50.0 hours	0.58	0.27	2.14	0.03
average_hours_worked50.1 hours and more	0.59	0.27	2.16	0.03
income_family$125,000 and more	0.02	0.04	0.54	0.59
income_family$25,000 to $49,999	-0.12	0.05	-2.59	0.01
income_family$50,000 to $74,999	-0.04	0.04	-0.91	0.36
income_family$75,000 to $99,999	-0.09	0.04	-2.20	0.03
income_familyLess than $25,000	-0.13	0.06	-2.33	0.02

Table 2: Estimates for model selected by AIC

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.74	0.27	31.80	0.00
total_children	0.02	0.01	1.90	0.06
sexMale	-0.15	0.02	-6.48	0.00
educationNot below Bachelor	-0.11	0.03	-4.11	0.00
self_rated_healthFair	-0.73	0.05	-13.32	0.00
self_rated_healthGood	-0.36	0.04	-10.16	0.00
self_rated_healthPoor	-1.19	0.10	-11.94	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
self_rated_healthVery good	-0.20	0.03	-6.18	0.00
citizenship_statusBy naturalization	-0.16	0.03	-4.96	0.00
self_rated_mental_healthFair	-2.01	0.06	-34.68	0.00
self_rated_mental_healthGood	-1.02	0.03	-29.79	0.00
self_rated_mental_healthPoor	-3.47	0.11	-31.60	0.00
self_rated_mental_healthVery good	-0.46	0.03	-14.98	0.00
marital_statusSingle, married	-0.51	0.04	-11.92	0.00
marital_statusSingle, never married	-0.38	0.04	-10.50	0.00
own_rentrented	-0.15	0.03	-4.96	0.00
living_arrangementNot alone	0.05	0.04	1.45	0.15
has_grandchildrenYes	0.27	0.03	8.22	0.00
average_hours_worked0.1 to 29.9 hours	0.49	0.27	1.82	0.07
average_hours_worked30.0 to 40.0 hours	0.47	0.27	1.73	0.08
average_hours_worked40.1 to 50.0 hours	0.58	0.27	2.15	0.03
average_hours_worked50.1 hours and more	0.59	0.27	2.17	0.03
income_family$125,000 and more	0.02	0.04	0.53	0.59
income_family$25,000 to $49,999	-0.12	0.05	-2.61	0.01
income_family$50,000 to $74,999	-0.04	0.04	-0.92	0.36
income_family$75,000 to $99,999	-0.09	0.04	-2.20	0.03
income_familyLess than $25,000	-0.13	0.06	-2.33	0.02

Table 3: Variance Inflation Factors

	GVIF	Df	GVIFˆ(1/(2*Df))
total_children	1.54	1	1.24
sex	1.09	1	1.05
education	1.12	1	1.06
self_rated_health	1.78	4	1.07
citizenship_status	1.04	1	1.02
self_rated_mental_health	1.76	4	1.07
marital_status	2.41	2	1.25
own_rent	1.22	1	1.11
living_arrangement	1.81	1	1.34
has_grandchildren	1.34	1	1.16
average_hours_worked	1.13	4	1.02
income_family	1.58	5	1.05

Table 4: Estimates for model selected by AIC with outliers dropped

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.71	0.24	35.89	0.00
total_children	0.03	0.01	3.18	0.00
sexMale	-0.15	0.02	-6.96	0.00
educationNot below Bachelor	-0.12	0.02	-5.23	0.00
self_rated_healthFair	-0.70	0.05	-14.30	0.00
self_rated_healthGood	-0.36	0.03	-11.25	0.00
self_rated_healthPoor	-1.13	0.09	-12.29	0.00
self_rated_healthVery good	-0.22	0.03	-7.52	0.00
citizenship_statusBy naturalization	-0.16	0.03	-5.30	0.00

	Estimate	Std. Error	t value	Pr(>|t|)
self_rated_mental_healthFair	-2.02	0.05	-38.38	0.00
self_rated_mental_healthGood	-1.02	0.03	-33.36	0.00
self_rated_mental_healthPoor	-3.64	0.10	-34.63	0.00
self_rated_mental_healthVery good	-0.49	0.03	-17.66	0.00
marital_statusSingle, married	-0.43	0.04	-11.34	0.00
marital_statusSingle, never married	-0.34	0.03	-10.74	0.00
own_rentrented	-0.13	0.03	-4.67	0.00
living_arrangementNot alone	0.05	0.03	1.45	0.15
has_grandchildrenYes	0.25	0.03	8.58	0.00
average_hours_worked0.1 to 29.9 hours	0.56	0.24	2.36	0.02
average_hours_worked30.0 to 40.0 hours	0.53	0.24	2.23	0.03
average_hours_worked40.1 to 50.0 hours	0.64	0.24	2.69	0.01
average_hours_worked50.1 hours and more	0.70	0.24	2.93	0.00
income_family$125,000 and more	0.02	0.03	0.55	0.58
income_family$25,000 to $49,999	-0.11	0.04	-2.65	0.01
income_family$50,000 to $74,999	-0.02	0.04	-0.58	0.56
income_family$75,000 to $99,999	-0.08	0.04	-2.07	0.04
income_familyLess than $25,000	-0.10	0.05	-1.90	0.06

Discussion

From the estimates of the final model. Useful infernece could be made. Table 4 shows that male, education level not below bachelor, poor body health, poor mental health, citizenship by naturalization not by birth, single martial status, rented house not owned all leading to show lower feeling score of life while not living alone, have more total children and has grandchildren make people living happier with higher feeling score of life. Also, average hours worked 50.1 hours and more as well as family income with 125,000 dollars and more make people living happier fixed other factors.

The findings are mainly for Canadians, however, the methods could be applied on any other similar data sets for other countries. People might draw different conclusions based on different data sets or even same data sets with different subsets. But, the findings should be consistent under similar situations for inviduals. The findings are important which could also be used for government which wants to improve the happiness of citizens. It would help the government understand the profiles of groups with high or low feeling score of life, for example, one of the most easiest way is to encourage single pearsons to get married or living common with their lovers. There are many different ways could be done for improving happiness.

At last, besides the findings of the study. There are some weaknesses. First of all, the whole study is performed based on the subset of 2017 GSS data for Canadias, different subsets of the data could lead to different results. Also, there are non-response bias in the original survey data which could lead to bias problems for this study. Second, the response feeling score of life is scaled from 0 to 10 which is a small range for using a linear regression model under the context of the study. Finally, because the survey was performed for inviduals within households, couples might be correlated in responses which lead to the independence assumption of linear model to be questionable, this kind of work could be left in future work to improve the findings in this study.

References

1. Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
2. Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

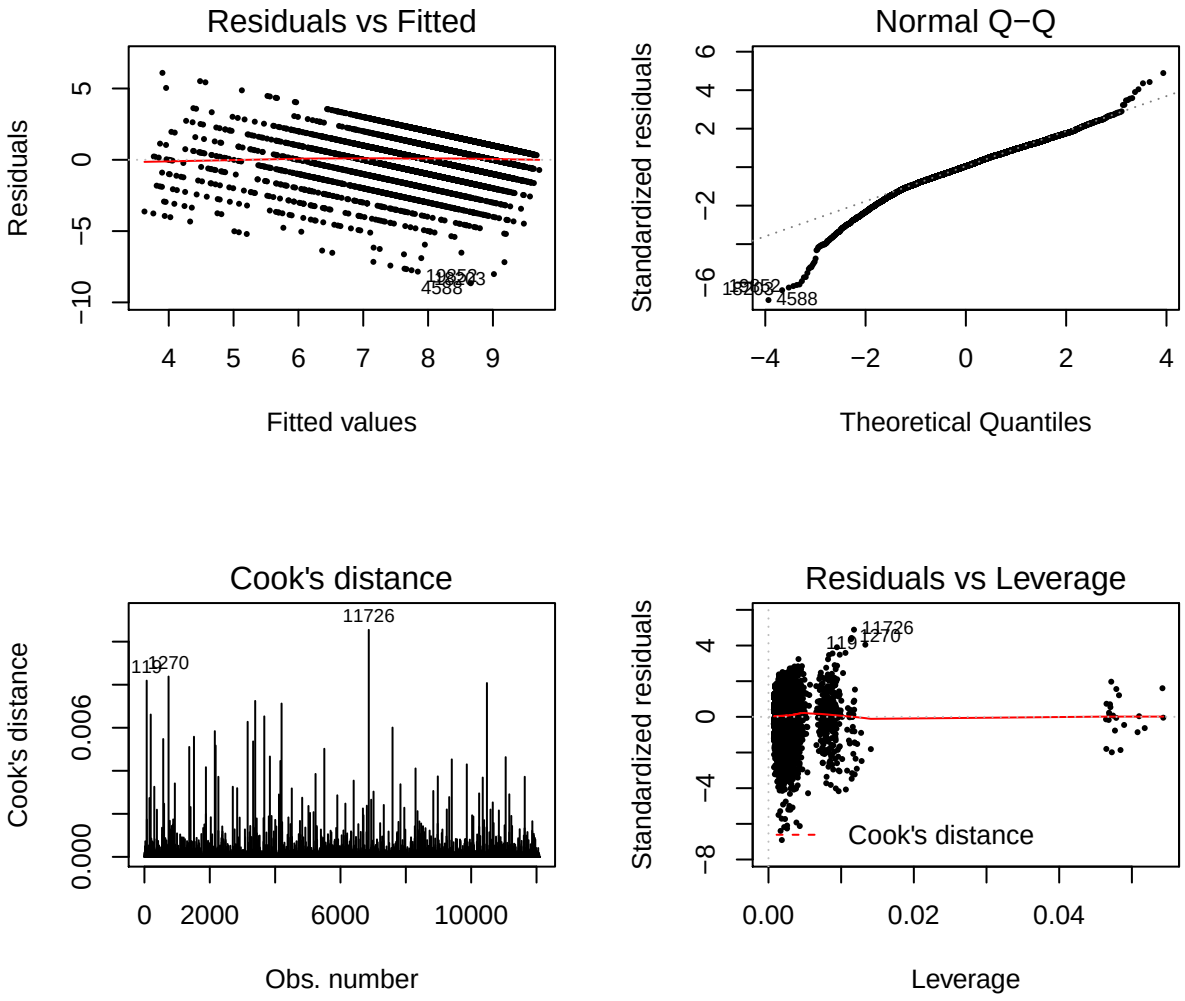

Figure 12: Model diagnostics plots for the model selected by AIC

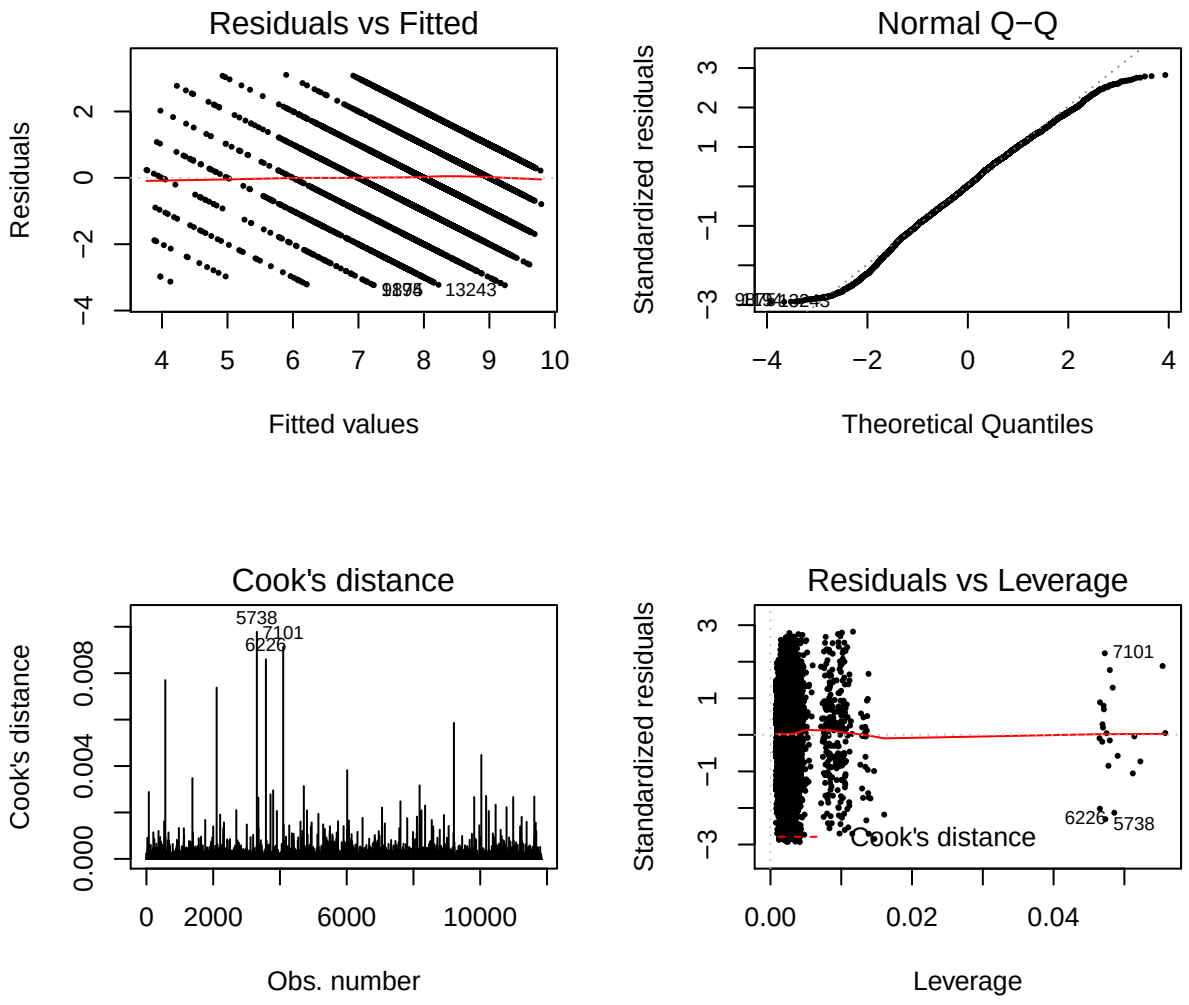

Figure 13: Model diagnostics plots for the model selected by AIC and outliers dropped

3. Hadley Wickham, Romain Franois, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

4. John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression,Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

5. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

6. Rohan Alexander and Sam Caetano (2020). 2017 Canadian General Social Survey data cleaning source code.

7. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.27.