Factors that affect Canadian's attitude towards life

Shuyu Duan,Fanxi Zhou, Feixue Han, Zhiang Chen

2020-10-18

Contents

1 Abstract

We investigated the factors that significantly impact Canadian citizens' feelings of life, and then we explored the data set collected by the 2017 General Social Survey (GSS). We found that age, the total number of children, sex, marital status, health condition and family income are the factors that significantly affect people's feelings of life. These results are important because our findings indicated what factors have affected people's feelings of life and understood Canadians' attitudes towards life.

2 Introduction

The World Happiness Report did a survey about the world's happiness scores and the survey has been undertaken in more than 160 countries. The results show that Canada's life satisfaction score has ranked in the top 20 in 160 countries (Ortiz-Ospina 2013). There are a variety of interior or exterior factors that affect Canadian citizens' life satisfaction rates. In this study, we want to explore what factors significantly impact Canadian citizens' feelings of life.

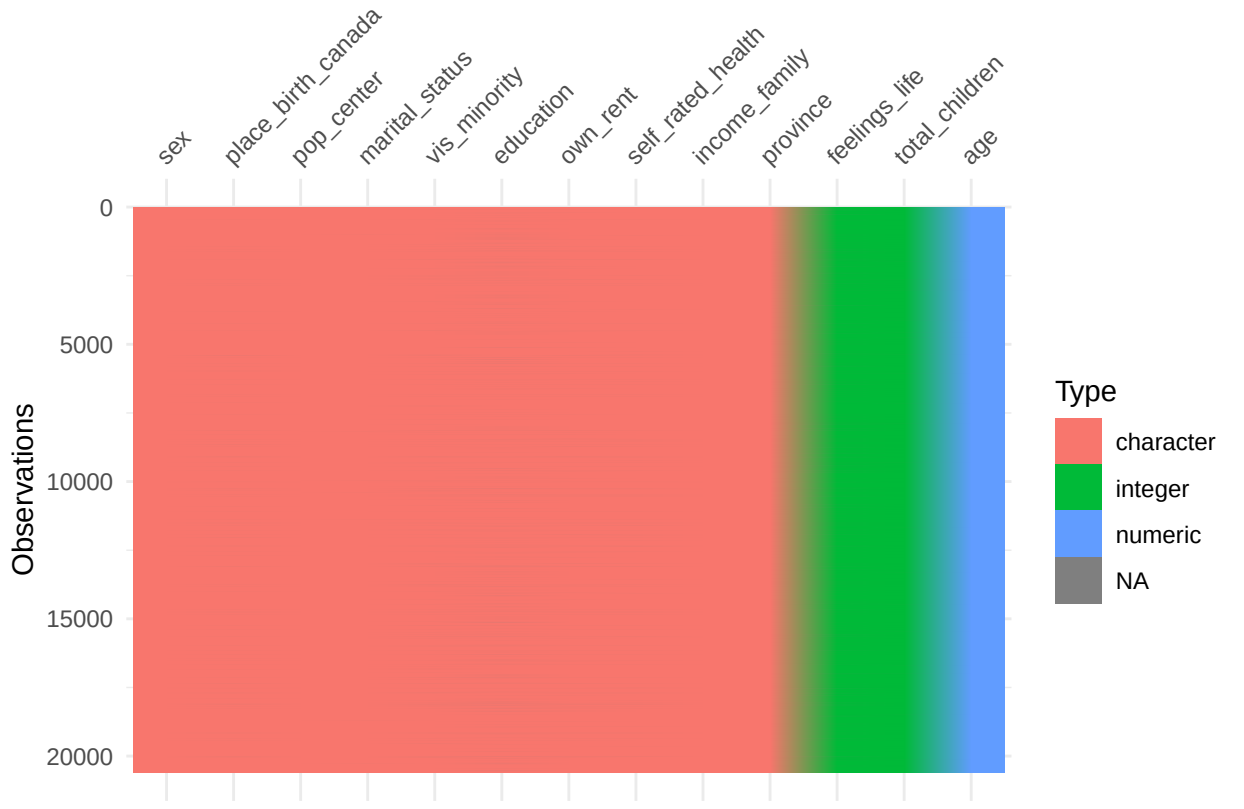
We analyze Canadian's feelings of life through several factors by using binary logistic regression models. The goal was to familiarize users with the content of the survey, and to describe procedures and concepts related to data quality, estimation, collection, processing and methodology through the provided context and background information. The guide was provided by the Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family.

In our study, we found that the total number of children, sex, marital status, health condition and family income significantly effect people's feelings of life. As a Canadian, people feel satisfied with less children in their life. Male have lower satisfaction towards life. Furthermore, people who live in common-law or married have a better life than the people who separated and widowed. Also, Canadian with good health conditions were more likely to have a satisfied life. People with the annual income of $\$125,000$ and more have a more satisfied life than the people with less than $\$100,000$ annual income.

3 Data

The GSS dataset is based on the General Social Survey (GSS), which aimed to gather information about the changes in living conditions of Canadians and to provide information on specific policy issues. The original data contains 20602 observations and 81 variables. In this dataset, the target population includes the person who is 15 years old and older,living in 10 provinces in Canada. we tried to analysis the distribution of age, sex, education and so on. The frame of the survey was to combine the telephone numbers (landline and cellular) with Statistics Canada's Address Register. During the survey, 91.8% of the telephone numbers reached the household. The overall response rate of the survey is 52.4% (Gagné and Keown 2014).

We selected a group of data which is a subset of the 2017 General Social Survey (GSS) which contains 19949 observations and 13 variables . The purpose of our study is to investigate the factors that affect people's attitude towards life. From the original dataset, we conducted that people's feeling towards life may be related to factors such as age, education, income and so on. Based on these conjectures, we built our dataset.



In the figure above, the factors such as sex, birthplace, age, education and so on are described as character, while feeling of life, number of children and age are described as integer and numeric correspondingly.

4 model

Binary logistic regression is used to predict a categorical variable based on a set of independent variables. For example, when we predict variables with yes or no or estimate the probability of "success". In our analysis, the response variable only has value 0 and value 1, so we decide to use this model for analyzing. The formula for binary logistic regression is:

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

or

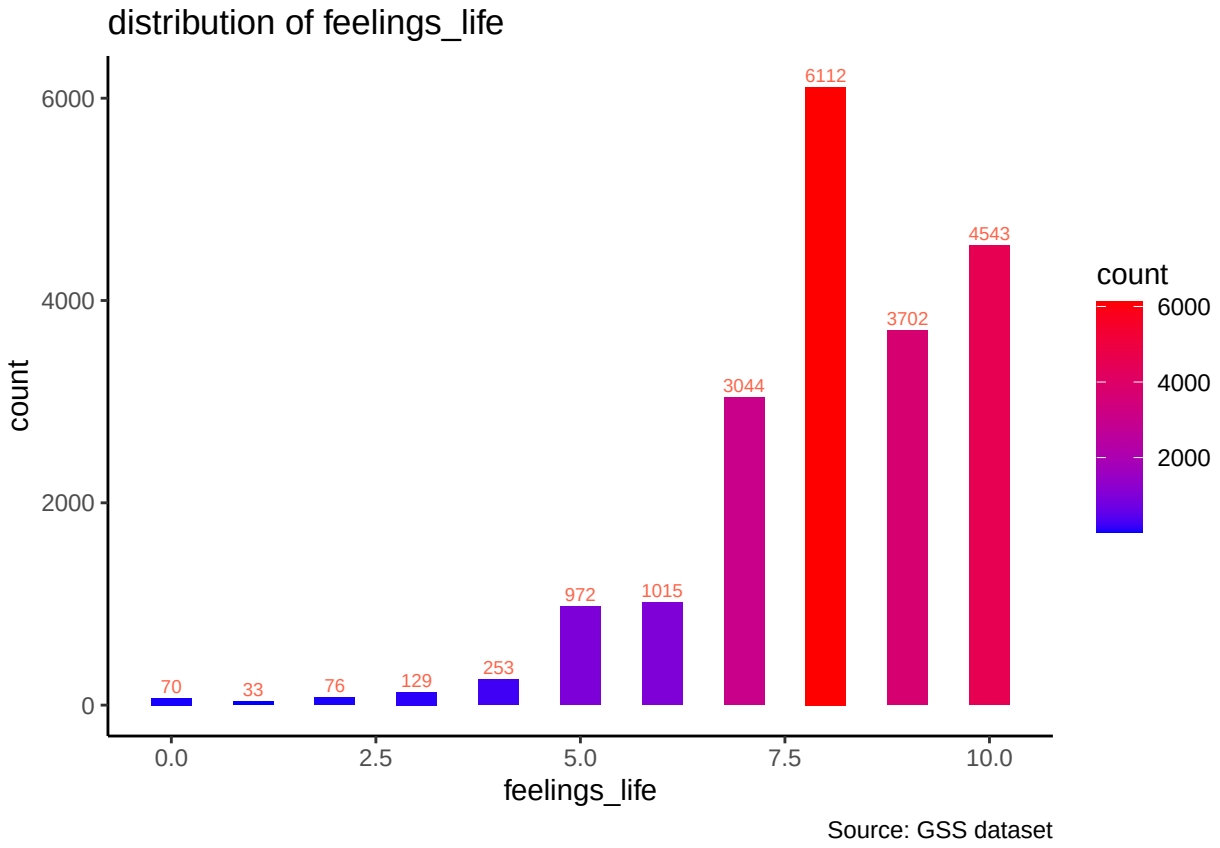
$$logit(\pi_i) = log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

We assume that Y_i is a binary response variable for i = 1,...n and takes on value 0 or 1 with P(Y_i = 1) = pi_i. Suppose X is a set of explanatory variables, x_i is the observed value of the explanatory variables for observation i = 1,...q. From the above formula, we can also get:

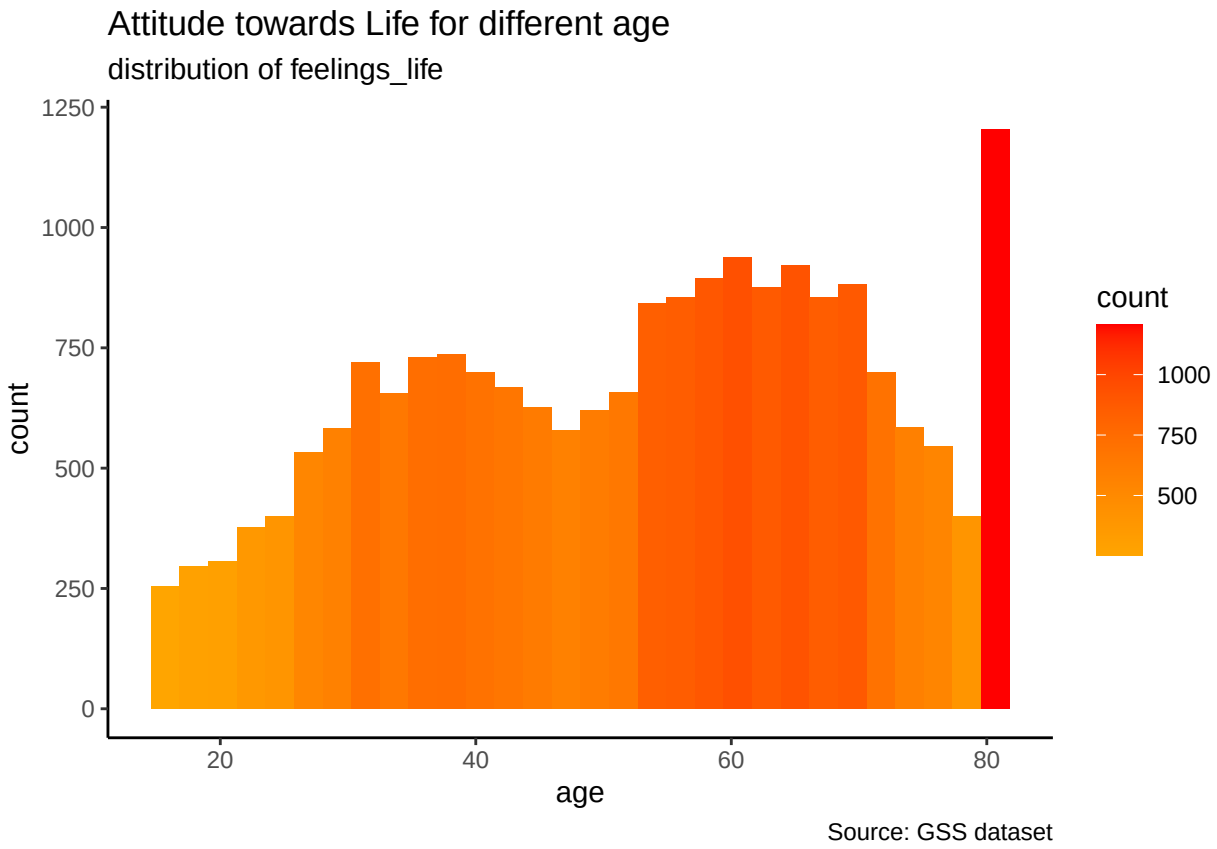
$$\frac{\pi}{1-\pi} = e^{\beta_0} e^{\beta_1 x_1} ... e^{\beta_q x_q}$$

Then the β_0 is the baseline odds and β_1 can be interpreted as holding predictors constant, a one-unit increase in x_1 increases the odds of success by a factor of e^{β_1}.

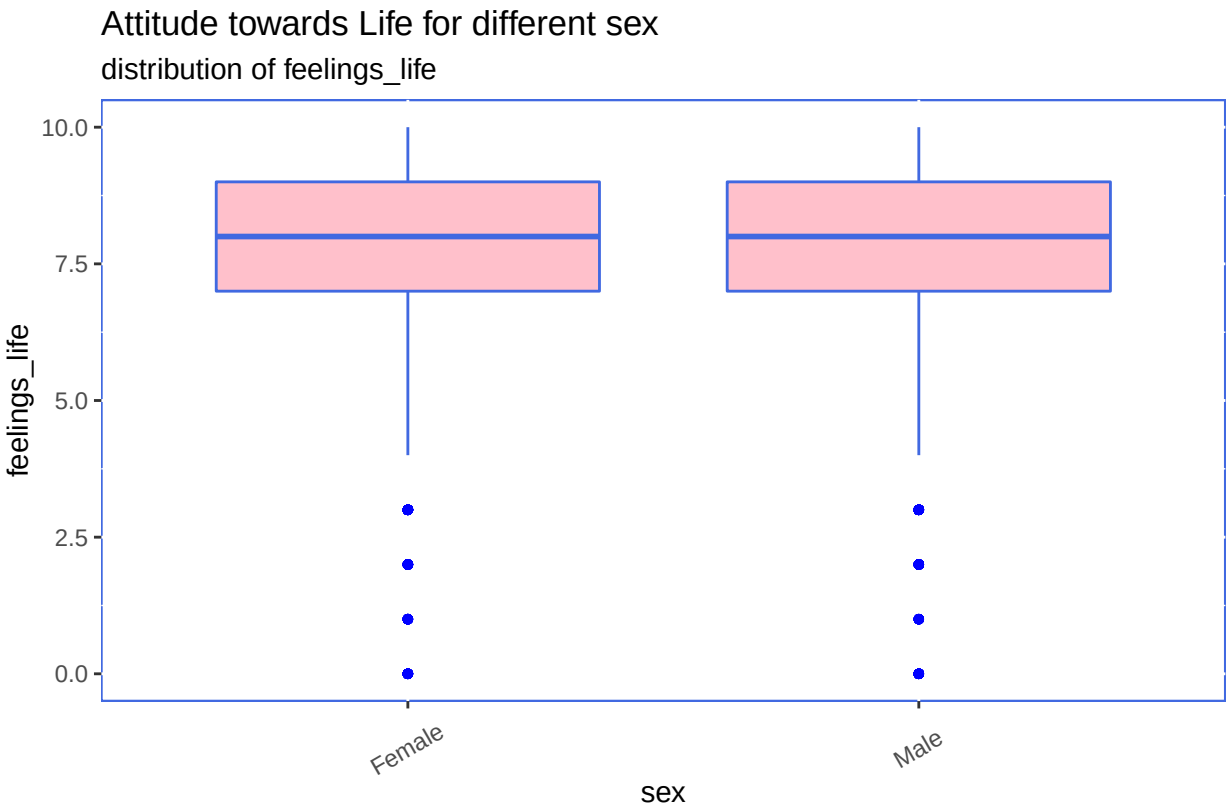
5 Results



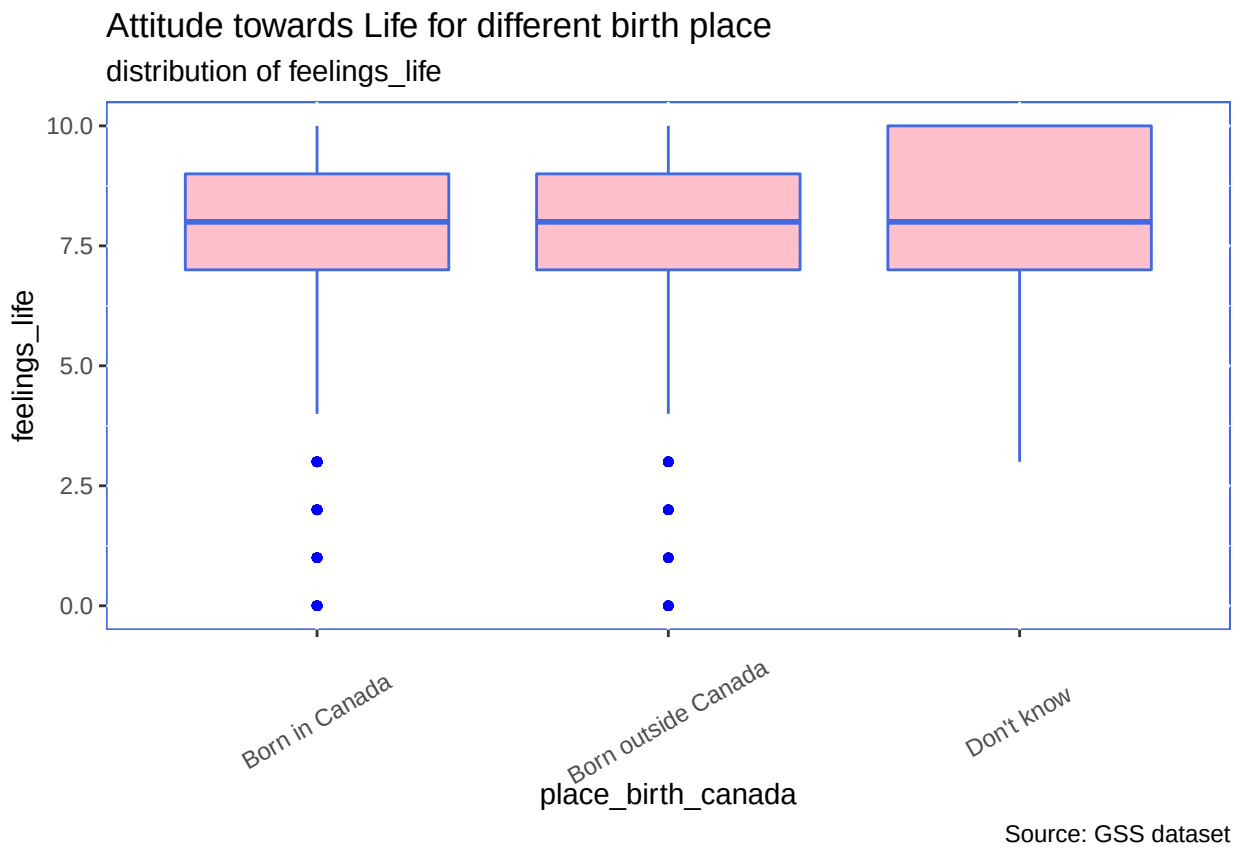
In the bar plot of distribution of feelings of life, people are mainly distributed in the 7.5-10 points range. There is a small number of people distributed under 5.



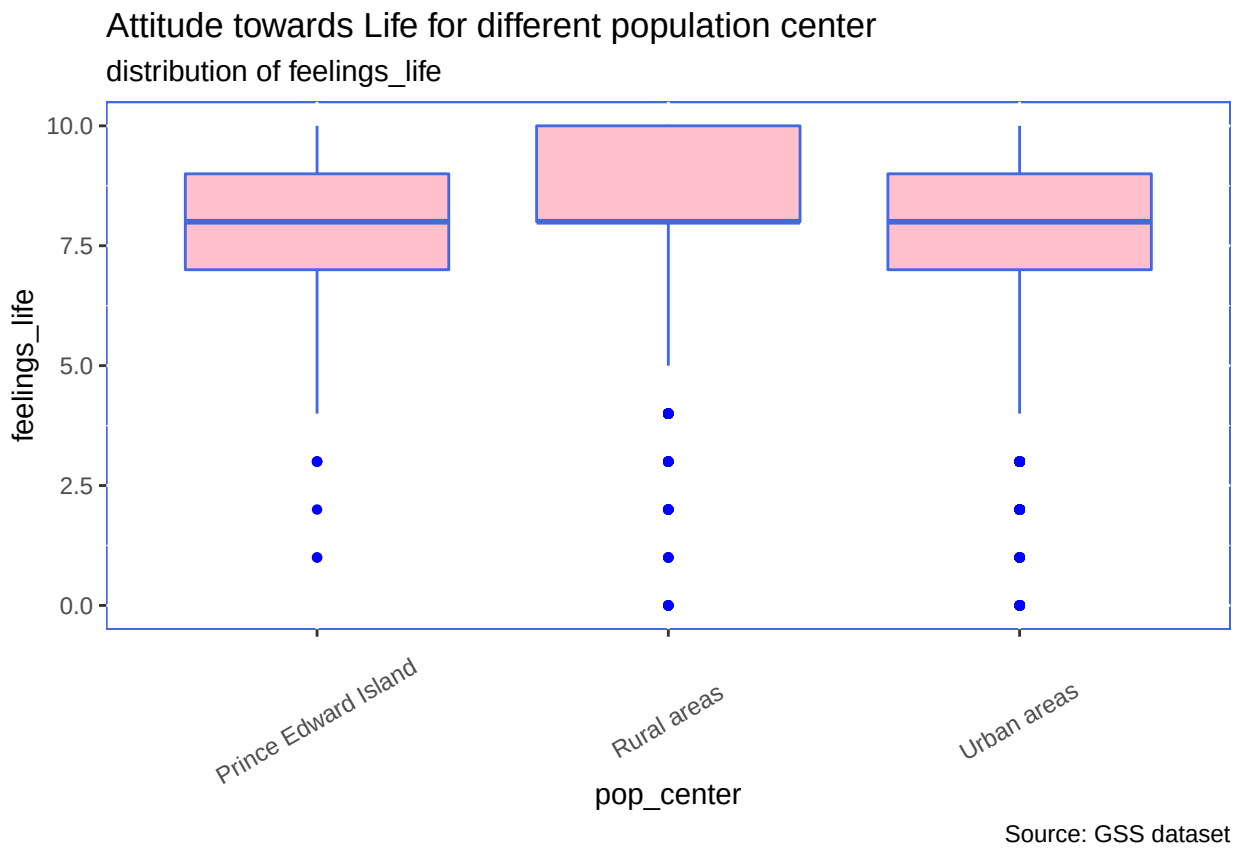
In the historic diagram of distribution of age, the amount of people are mainly distributed in the 35-45 and 55-65 age range. Furthermore, people over 80 years old had the largest amount.



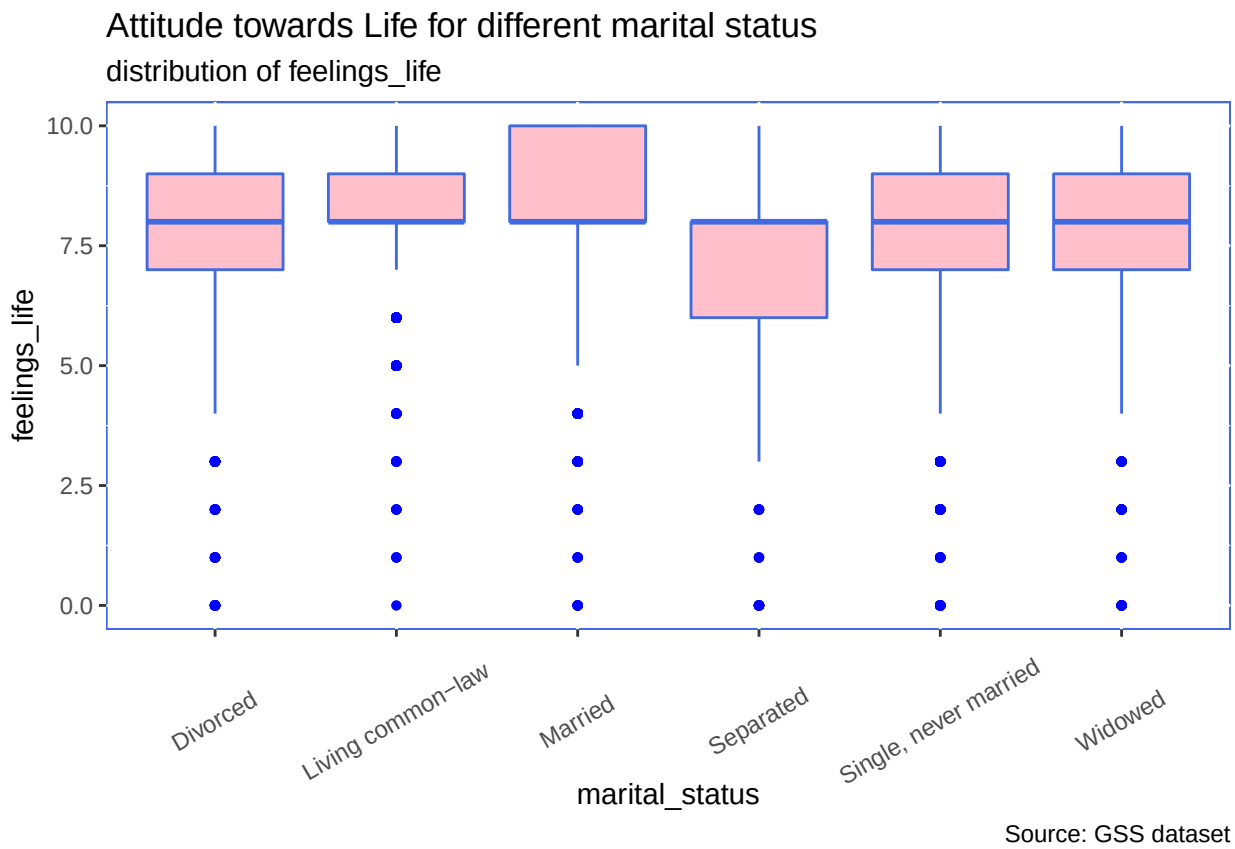
In the boxplot of distribution of feeling, the quantities of satisfaction of male and female are similar.



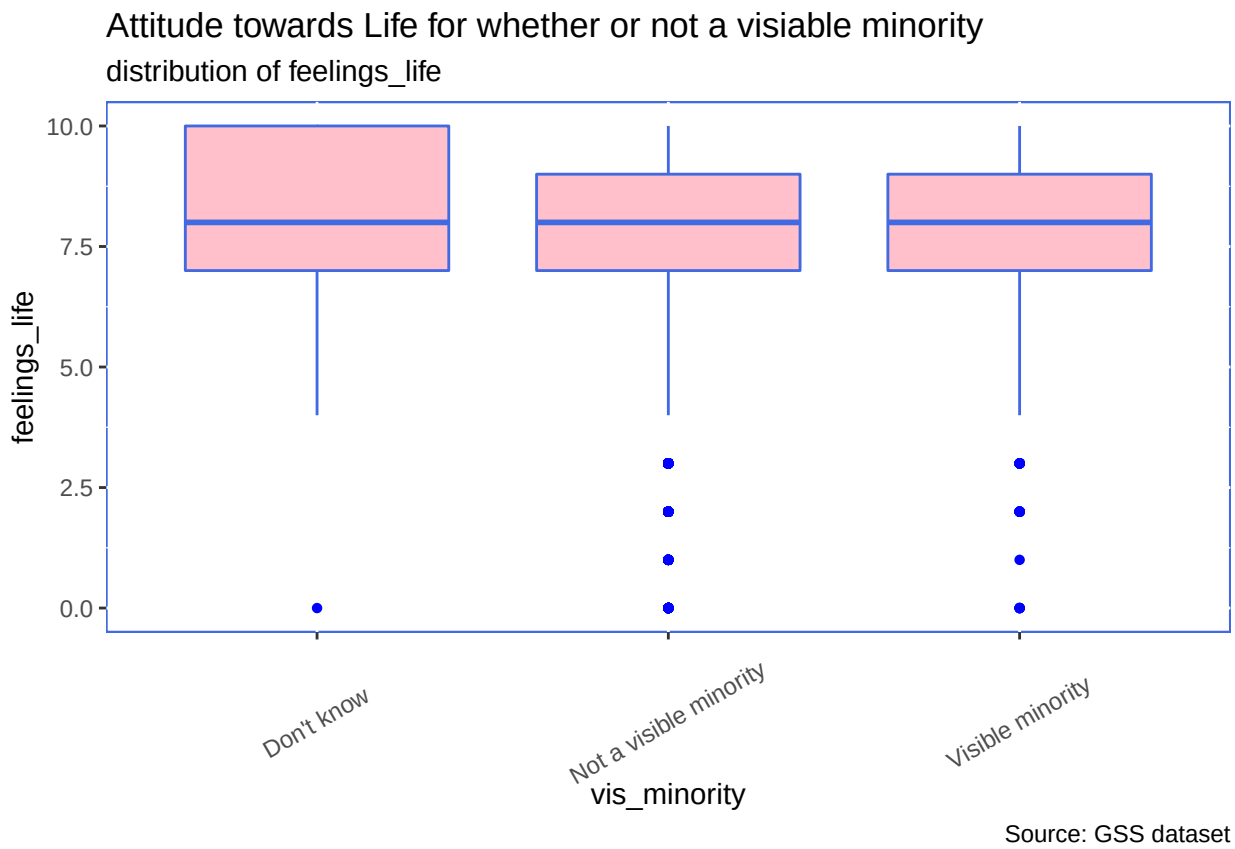
In the box plot of attitude towards Life for different birth places, people who are born inside or outside of Canada have similar feelings of life.



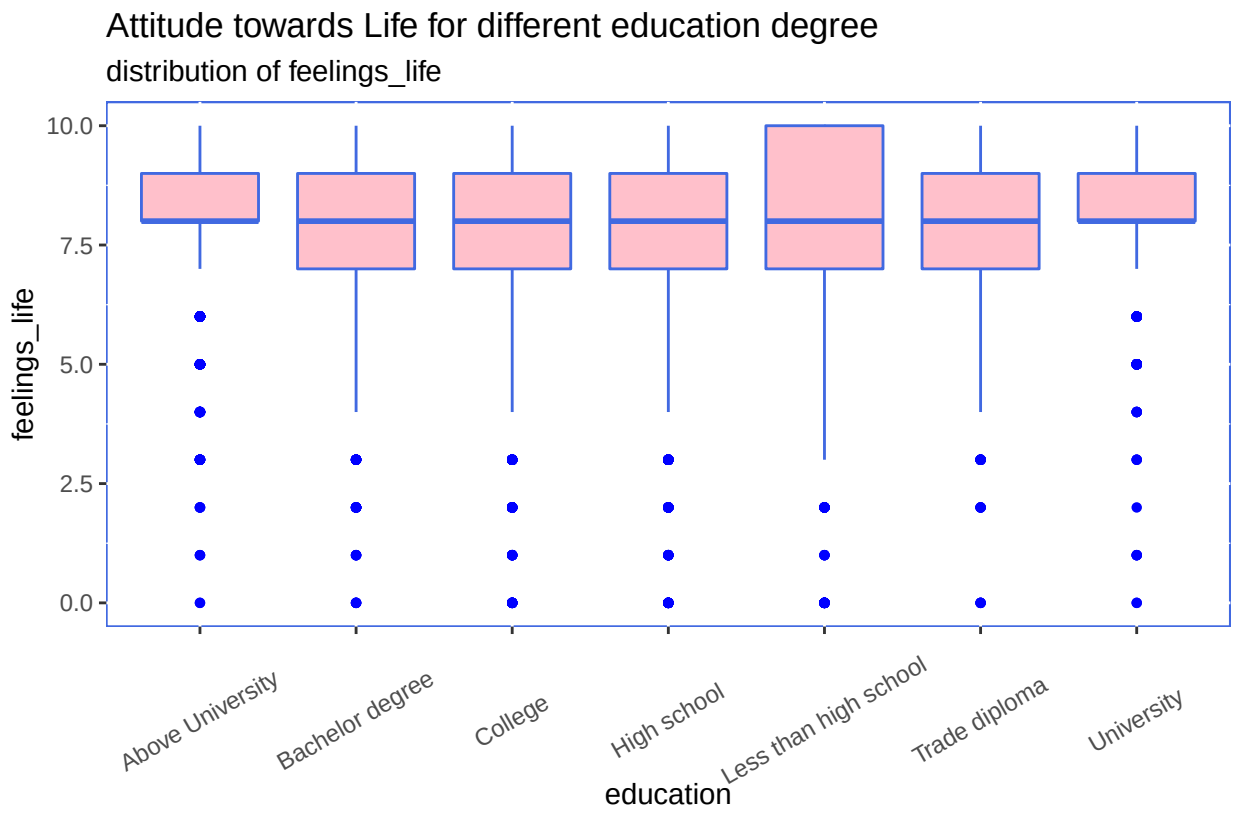
In the box plot of attitude towards Life for different population centers, the median of people's satisfaction in Prince Edward Island, rural area and urban area are almost the same.



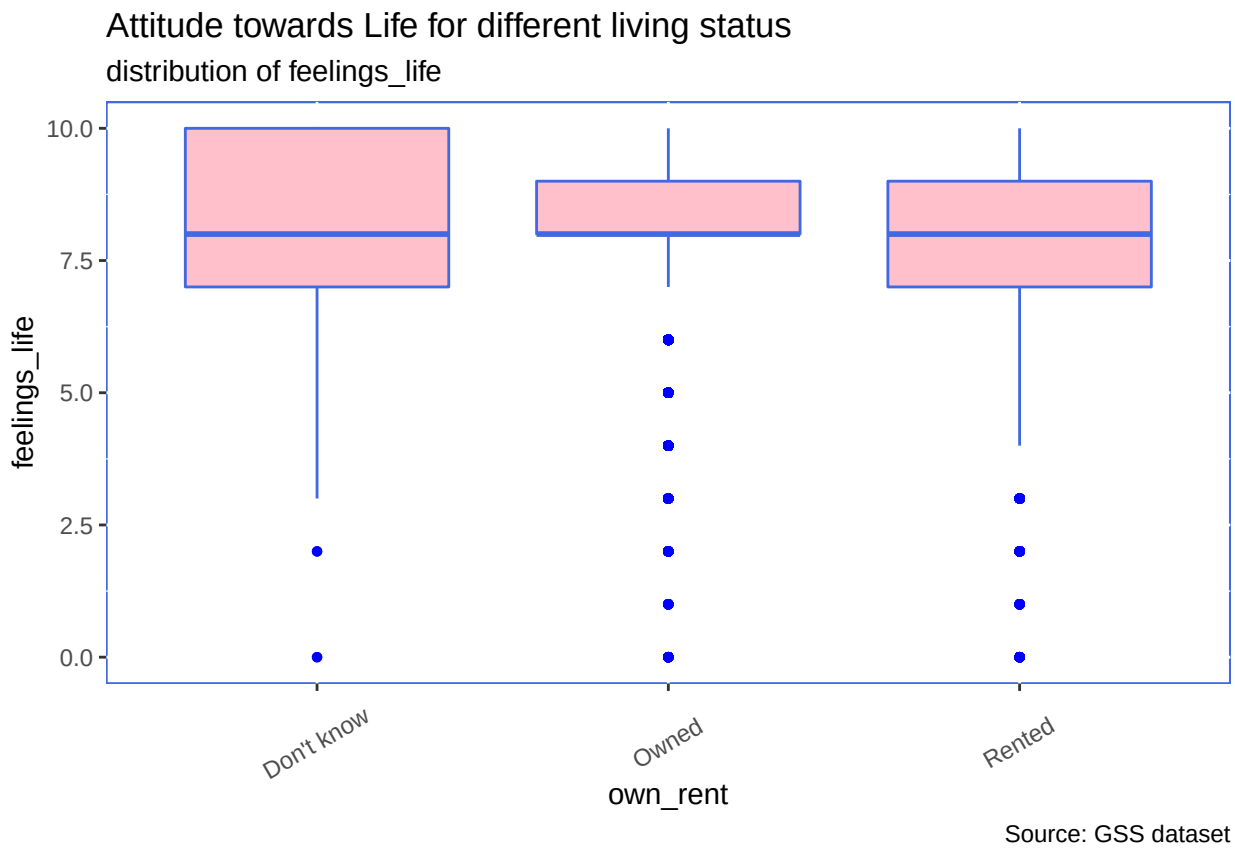
According to the attitude towards Life for different marital status boxplot, the living common-law has the most outliers while the single, widowed and divorced rate are almost the same.



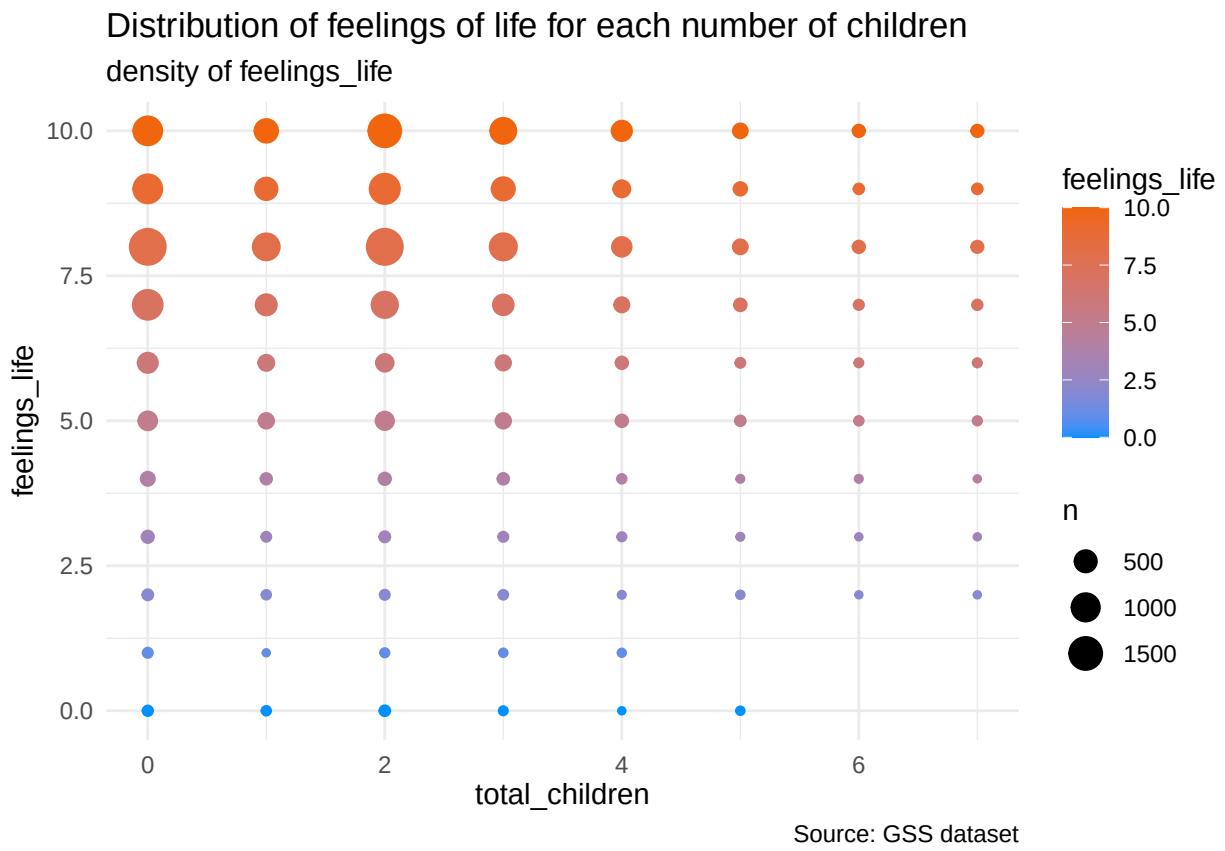
In the boxplot for attitude towards Life for whether or not a visible minority, it shows that most of the people do not know the minority.



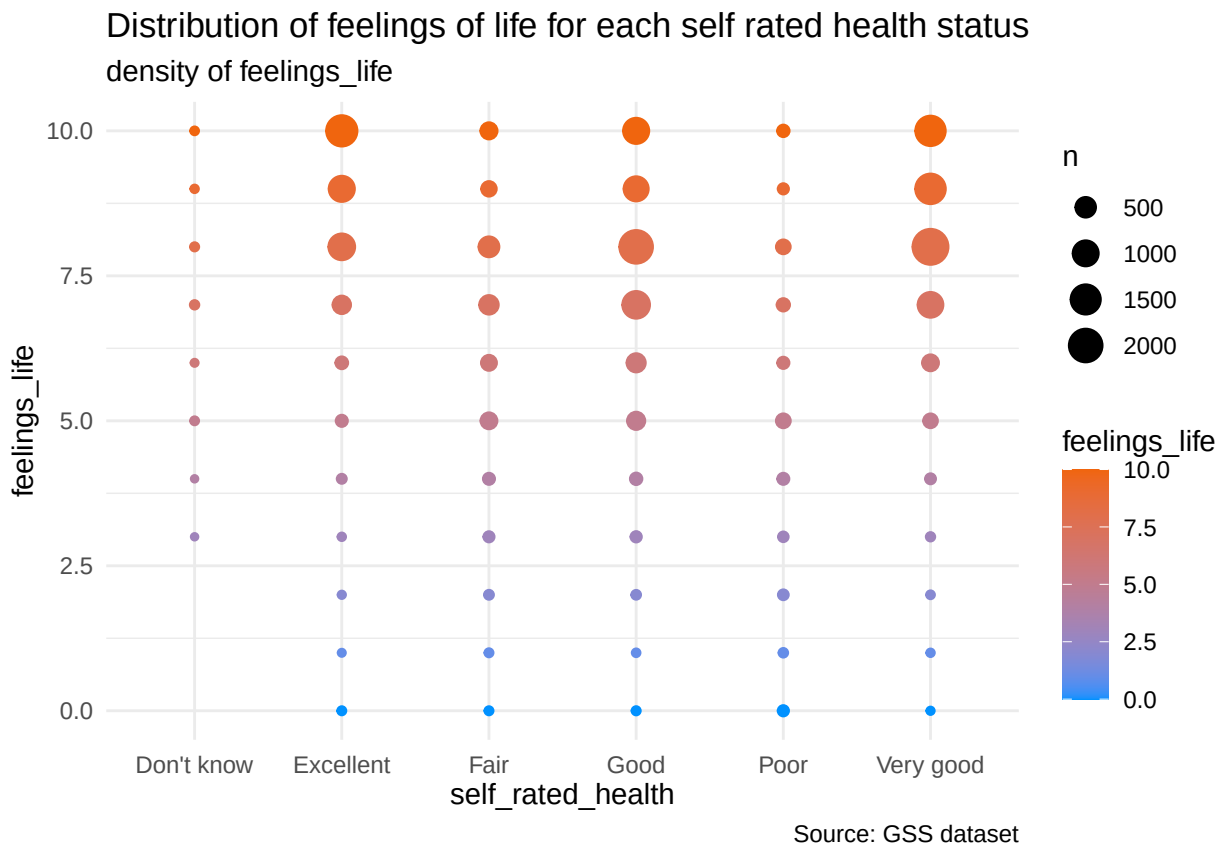
In the boxplot of attitude towards Life for different education degrees, the category shows that people with less than high school education level are more satisfied with their life.



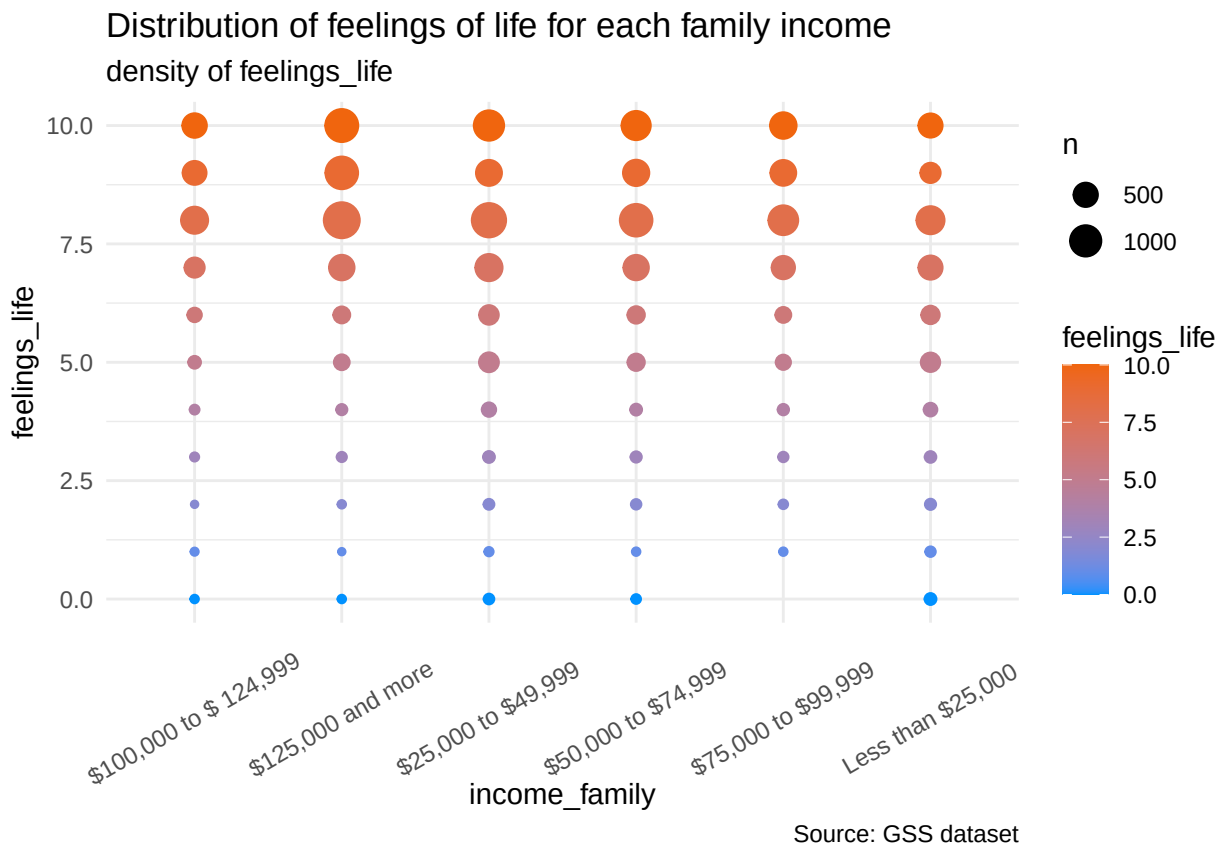
In the box plot of attitude towards Life for different living status, the median of the category of people who rented, owned a house and unclear are almost the same. The owned category has the most number of outliers.



In the mapping of distribution of feelings of life for each number of children, most of people have fewer children. Whether they have fewer or more children, their life satisfaction is relatively high.



In the mapping of distribution of feelings of life for each health rate status, people with good health status show higher satisfaction towards life.



In the mapping of distribution of feelings of life for each family income, people with $50000 to $$1250000 $ income are more satisfied with the feeling of life. However, the drawbacks of data is that most of the points lie between 7 to 10. It can be improved by changing the type of diagram to historic diagram. The median will show the result better.

Table 1: Summary of feelings of life in each province

province	mean	median
Alberta	8.053	8
British Columbia	8.032	8
Manitoba	8.091	8
New Brunswick	8.204	8
Newfoundland and Labrador	8.240	8
Nova Scotia	8.043	8
Ontario	8.065	8
Prince Edward Island	8.160	8
Quebec	8.102	8
Saskatchewan	8.183	8

The above table shows the mean and median of people's satisfaction of life in different province.

Through the binary logistic regression model, the data was summarized in the table below. In the Pr category, the self-rated-health-fair has the highest value which is 0.877 while category age, sex and so on have the minimum value. Most of the Pr-value is lower than 0.1. The std.error value is lower than 0.5 and the maximum is 0.383. The maximum value of z is 11.141 which is from the age category while the minimum is -5.455 which is from the intercept. The coefficients for most of the categories are positive. The maximum

Table 2: Summary of Model Results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.143	0.393	-5.455	0.000
age	0.013	0.001	11.141	0.000
total_children	0.049	0.012	3.929	0.000
sexMale	-0.073	0.031	-2.318	0.020
pop_centerRural areas	0.086	0.089	0.960	0.337
pop_centerUrban areas	-0.119	0.084	-1.415	0.157
marital_statusLiving common-law	0.397	0.076	5.261	0.000
marital_statusMarried	0.600	0.062	9.708	0.000
marital_statusSeparated	-0.371	0.113	-3.295	0.001
marital_statusSingle, never married	0.080	0.070	1.139	0.255
marital_statusWidowed	0.174	0.076	2.300	0.021
self_rated_healthExcellent	1.811	0.371	4.879	0.000
self_rated_healthFair	-0.058	0.373	-0.155	0.877
self_rated_healthGood	0.452	0.371	1.221	0.222
self_rated_healthPoor	-0.531	0.383	-1.387	0.166
self_rated_healthVery good	1.029	0.371	2.777	0.005
income_family$125,000 and more	0.046	0.056	0.832	0.405
income_family$25,000 to $49,999	-0.167	0.060	-2.781	0.005
income_family$50,000 to $74,999	-0.104	0.059	-1.754	0.079
income_family$75,000 to $99,999	-0.088	0.061	-1.439	0.150
income_familyLess than $25,000	-0.209	0.070	-2.999	0.003

value is 1.811 which is from the health excellent category. The minimum value for the coefficient is -0.531 which is from the health poor category.

6 Discussion

We used the binary logistic regression model to do this analysis. In our model, we divided our observations of feeling_life into 2 groups to determine the relationship between feelings_life and other factors. Most of the people give the feelings_life a high score, with a mean of 8.09 and a median of 8. Under ideal condition, we should divide people into 2 groups with the same population, but in order to form a relationship of function which is that one independent variable should only have one corresponding dependent variable, we should not divide people giving a score of 8 into different groups, thus we divide people into 2 groups by the mean. We put people with a really high score (9 and 10) into group one, and the dummy variable for this group is 1, and put people without a really high score (8 and under) into group 2, the dummy variable for this group is 0.

Then, the binary response variable of our model is feeling_life which has dummy variables 0 and 1. We also have a set of explanatory variables: age, number of total children, sex, the center of population, marital status, self-rated health, family income. In this model, the p-value helps us to test the null hypothesis so that we can indicate whether the factors have a correlation with our response variable in the whole population. The null hypothesis for our model is that there is no correlation with our response variable. If the p-value smaller than 0.05, it rejects the null hypothesis so there may be a correlation between that factor and our response variable. The smaller the p-value, the stronger evidence for rejecting the null hypothesis. However, if the p-value greater than 0.05, it supports the null hypothesis, which means that there may not be a correlation between the factor and our response variable.

By observing the model results, we can see that age, the total number of children, sex, marital status, health condition and family income are the factors that significantly affect people's feelings of life. As Canadian citizens' age and the total number of children in their family increases, people feel more satisfied with their lives. In the p-value table, sex(male) is negatively correlated with the feelings of their life, which means

low life satisfaction was found among males. People who live common-law or married tend to feel better about their lives and worse with the people who separated. Moreover, widowed people also tend to be more satisfied with their lives. From the table, we can also conclude that Canadians with good health conditions were more likely to report satisfied with their lives. High life satisfaction can be found among the families with an annual income of $\$125,000$ and more. The families with a yearly income of less than $\$100,000$ start to show a negative correlation with their life feelings, and the result is more significant for the families with an annual income that is less than $\$25,000$, which means the families with a lower annual income are more likely to be dissatisfied.

7 Weakness and areas for future

7.1 Weaknesses

- According to the 2017 General Social Survey (GSS), 91.8% of the telephone numbers reached the household during the survey, but the survey's response rate is only 52.4%. A low response rate can produce sampling bias if the non-response outcomes are unequal among the participants.
- People who completed the survey were mainly distributed in the 30-45 and 50-70 age range, which means older people are more likely to complete the survey. From the result, we concluded that people feel more satisfied with their lives as their age increases. However, the result may not be accurate since our sample can not represent the whole population.
- The distribution of feeling_life is flawed. In our data, most of people have a higher score for feeling_life, which could be harder for us to turn it into a binary observation. We used "8" as the boundary to divided feeling_life into two groups. However, in the real life, "6","7" also represent a positive attitude towards life. It would affect our conclusion in terms of people's attitudes towards life in Canada.

7.2 Next step

- For the further study, we can study the components of the feelings of life and do a further survey to collect the data on the detail of each specific parts of feelings of life to increase the accuracy of the relation between the independent variables and dependent variable and increase the liability of the regression model.
- Also since around one-third of the respondents give 8 on the feeling of life, thus we can develop a further study focus on this part of data to study the common characteristics within the group and difference from the whole data set. We can create a new model to forecast for any observation whether or not he or she is in the "8" group. Also, we can do the same step for the group of peoples who give the extremely low mark, like lower than 5.
- There might be some other facts which might infect the results but not in the model, for the further study we can conduct a series of further survey focus on these variables. We can give the respondents a question that what might infect your feelings for life and according to the answer to modify the questionnaire. Through a series of the survey, we can have a more accurate model for the relationship between variables and feelings of life.

8 Appendix

Code and data supporting this analysis is available at: "https://github.com/zhoufanx/STA304-problem-set3".

We use R (R Core Team 2020) and packages tidyverse (Wickham et al. 2019), visdat (Tierney 2017), dplyr (Wickham et al. 2020), kableExtra (Zhu 2020) and GSS dataset(Gagné and Keown 2014) for this analysis.

9 References

Gagné, C., Roberts, G. and Keown, L.-A. (2014) "Weighted estimation and bootstrap variance estimation for analyzing survey data: How to implement in selected software". The Research Data Centres Information and Technical Bulletin. (Winter) 6(1):5-70. Statistics Canada Catalogue no. 12-002-X. http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002- X20040027032&lang=eng

Ortiz-Ospina, E., & Roser, M. (2013, May 14). Happiness and Life Satisfaction. Retrieved October 18, 2020, from https://ourworldindata.org/happiness-and-life-satisfaction

6.2 - Binary Logistic Regression with a Single Categorical Predictor. (n.d.). PennState Eberly College of Science. Retrieved October 18, 2020, from https://online.stat.psu.edu/stat504/node/150/

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Gagné, Roberts, C., and L.-A. Keown. 2014. "Weighted Estimation and Bootstrap Variance Estimation for Analyzing Survey Data: How to Implement in Selected Software." http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-002-%20X20040027032&lang=eng.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tierney, Nicholas. 2017. "Visdat: Visualising Whole Data Frames." *JOSS* 2 (16): 355. https://doi.org/10.21105/joss.00355.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain Fran?ois, Lionel Henry, and Kirill Müller. 2020. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.