Understanding Happiness, an Analysis on Life Satisfaction of Canadians in 2017

Adel Boufama

October 19, 2020

Understanding Happiness

An Analysis on Life Satisfaction of Canadians in 2017

Adel Boufama

October 19, 2020

Abstract

Using the 2017 General Social Survey (GSS) of Canada which contains data on over 20000 people, an analysis of whether or not internal or external factors matter more for a person's life satisfaction is conducted. The internal factors studied here are self rated mental and self rated physical health, while the external factors studies are family income, religious affiliation and urban or rural residence. The data is analyzed and compared with life satisfaction being "good" or "bad" using a Logistical Regression model. Results show that internal factors have a greater effect on a person's life satisfaction compared to external factors.

Introduction

Everyone wants to be happy and satisfied with their lives and there no shortage of information out there about how to improve one's quality of life. The most popular tips tend to be health and wellness related and there is an old saying that money can't by happiness. Everyone has their own preferences, while some people enjoy the quiet and vast openness of living in a rural area, others enjoy the fast paced lifestyle of a city lifestyle. Having purpose in life is also very important for one's happiness and people get their purpose from different things, some get it from their religious affiliation, while others find purpose in other aspects of their lives. There have been many studies conducted on this particular topic to help pinpoint what areas people can put effort to change in order to improve life satisfaction. This time I have decided to conduct my own study on this topic.

The goal for this analysis is to figure out whether internal factors such as physical and mental health play a greater role in overall percieved life satisfaction compared to external factors such as family income, location and religious affiliation. This analysis can also allow

us to compare the magnitude of effect each factor studied has on a person's overall life satisfaction, this can bring about interesting findings.

Following this, we will discuss the dataset in detail. How the data was obtained, some advantages and potential flaws with the data. Then the model for this analysis will be introduced and finally there will be a discussion which covers the results from the model, some potential flaws and next steps.

Data

This data comes from the 2017 General Social Survey (GSS) of Canada which includes socio-demographic information such as age, sex, education, religion, immigration status, family income, etc. The year 2017 was a Family cycle of the GSS which specifically focused on monitoring changes in Canadian families. It contains information on family origins, marriages, family income, children and other socioeconomic characteristics. This dataset contains information from 20602 people 15 years of age and older, living in the 10 provinces.

The GSS data is cross-sectional which means that it takes a snapshot of data from the sample of the population at a specific point in time. The population is everyone residing in Canada at the time of the survey in 2017, approximately 36.54 million people (according to Google). The frame of this survey is every Canadian that has a landline or mobile phone number from the Census or residence/address information stored in Statistics Canada administrative sources. Statistics Canada used a form of stratified random sampling with 27 strata, each strata was a either a central metropolitan area (CMA) like Toronto, Montreal, Halifax, etc. or a non CMA part of one of the 10 provinces, this resulted in 17 CMA strata and 10 non-CMA parts of all provinces. One randomly selected individual from each family was selected to complete the survey questionnaire and data collection was done by "the computer-assisted telephone interviewing method". The field sample was 43000 units, 34000 surveys were sent to the randomly selected households and a 20000 questionnaires were expected to be completed. The final result was just a bit above the expected, at 20602 completed questionnaires.

The cross-sectional nature of this study has a main drawback of not being able to be used in logitudinal studies but otherwise the data is perfectly fine for analyzing correlations between different attributes. The frame population is fairly thorough, it was obtained using census data as well as phone numbers which will give them almost every person in the country. Issues might come up when dealing with people who are Canadian citizens with a Canadian address but are currently residing in another country. These cases have not been addressed on the Statistics Canada website so it is unclear if or how this is dealt with. In addition, undocumented immigrants who do not show up in the census will also be excluded.

Using stratified random sampling, the GSS is able to obtain an accurate reflection of the population because it ensures that each strata subgroup will recieve proper representation. The results will therefore be a more accurate model of the actual population when conducting studies compared with data obtained by simple random sampling or systematic

sampling. This sampling method does not come without its drawbacks though, since every unit in the frame population must be classified into only one strata (sub-population), it could cause problems when the classification is not very clear. With each strata representing a geographical area, this should not be a widespread issue but there could definitely be cases of people with multiple residences in different strata locations, especially if the person resides in an address not provided as their main residence.

Non-responses were dealt with using adjustments to survey weights. Most of the time, non-responses were just inputted as N/A with the exception basic household data such as income and household composition. The basic household data from 2016 was used to model and adjust the non-response sections accordingly.

Questionnaire

The questionnaire is very long and thorough, it covers a wide range of topics dealing with personal and domestic affairs. It does well covering the breadth of topics which can give a lot of rough information about many different aspects of a person's life. Where it falls short tends to be in the details of some categories and maybe some unnecessary questions.

Most or all of the possible answers to each question are categorical or discrete. This means it only captures what group each surveyed belongs to. For example, in the family income and personal income sections, they give several income brackets to choose from (less than 25k, 25k to 49k, 50k to 75k, 75k to 100k, etc..). As a result of this, you cannot do a accurate study into any of the categorical variables of the data such as income as there is no way to find an accurate average income, nor could you use the income data to find accurate correlations. Everything ends up being a rough estimate when recieving categorical answers to the survey questions. Other issues that could come up are that some answers to questions are very subjective. For example, someone with a more positive outlook on life might rate their physical health as 'good' even though they have a chronic disease while another person with similar health issues but a more negative outlook would rate their health as 'poor'.

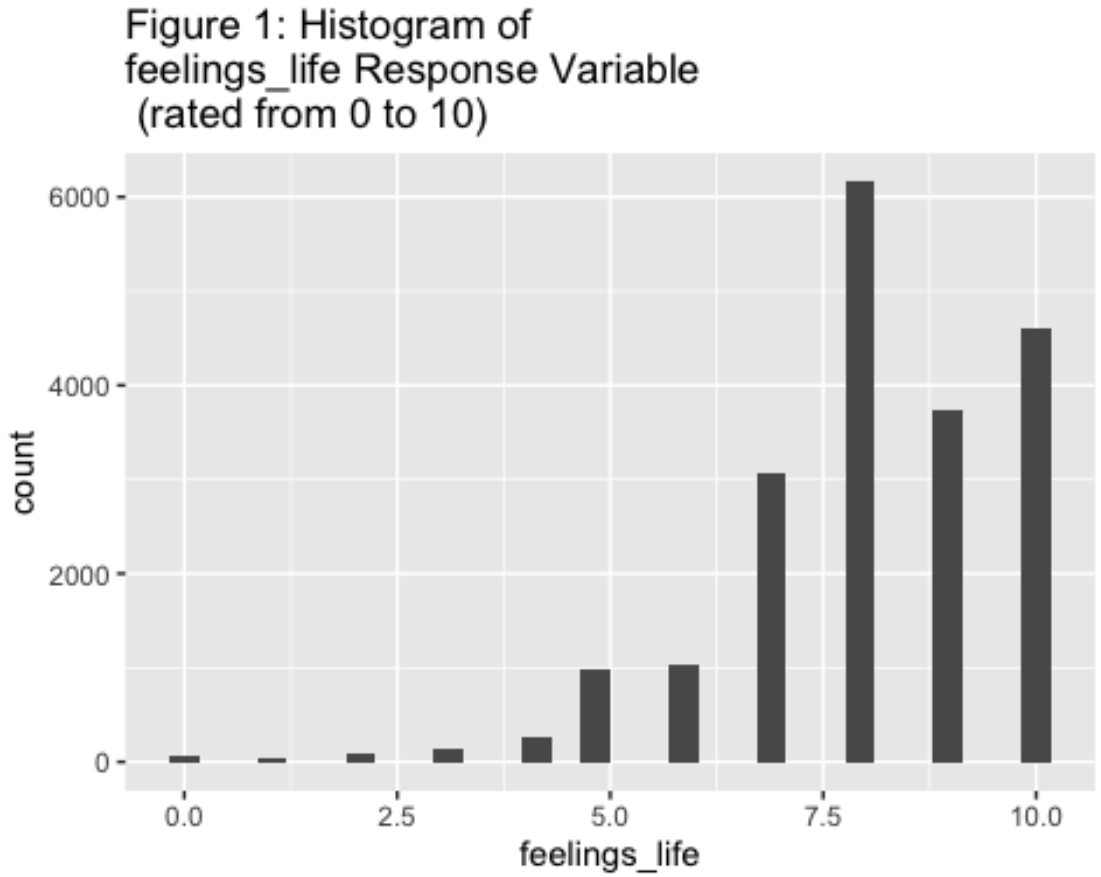
There are also some questions that can be screened out as they don't apply to everyone. This is especially apparent with the question asking if a person lives with their partner when they already answered that they are single, or a person who does not follow a religion being asked how often they attend religious services. This issue is relatively minor and might have been addressed already, but if it was not, it could definitely have helped save time or cost when conducting the surveys.

Variables for this Report

I will be using the variables: feelings_life, pop_center, income_family, religion_has_affiliation, self_rated_health and self_rated_mental_health. All of these things are discrete variables.

'feelings_life' and 'categorical_feelings_life'

The main variable that I will be focusing on and comparing everything to, is the 'feelings_life' question. This part of the questionnaire asks the survey taker to rate their feelings of life as a whole from a scale of 0 to 10 with 0 being the worst and 10 being the best. I will be finding if correlations exist with 'feelings_life' and the other 5 variables. Here is a closer look at the distribution of 'feelings_life'.



As you can see in Figure 1, it is not normally distributed and very lopsided in fact. I will discuss more on how I deal with it as my response variable in the next section.

Analyzing the response variable a bit more, we find that the mean is 8.095, median is 8 with a standard deviation of around 1.64.

I decided to make another variable called 'categorical_feelings_life', this is in order to fit the range of values in 'feelings_life' into a binary of either "good" or "bad". I organized them such that a person with a "bad" 'feelings_life' means they rated their general feeling of life within the range of 0 and 7 inclusive, while a person with a "good" 'feelings_life' gave a rating of 8 to 10 inclusive. I chose to seperate it as such because the median (midpoint value) of the 'feelings_life' variable is 8 and you can see in Figure 1 how the distribution

shows that this variable is split roughly around 8. I could have decided to consider 8 as a "bad" rating for life as it is the midpoint but using my own personal judgement, I would consider a rating of 8/10 quite good so I decided to do the split between 7 and 8. I will discuss the reasoning behind needing to use only 2 categories for this variable in the next section when talking about the model I used for this study.

'pop_center'

This variable respresents the population density (urban or rural type) of the area a person resides in. It has 3 possible categories: Larger urban population centres (CMA/CA), Rural areas and small population centres (non CMA/CA) and Prince Edward Island (PEI). I find it interesting how they put a specific category for PEI and don't fully understand why they wanted to categorize this but I will consider PEI as rural. I chose this variable to study its relation with 'feelings_life' because I want to see if there is any correlation between the two and it is an external factor that could affect a person's outlook on life.

The variable can be seen as having similarities to the 'region' variable where it categorizes people as residing in different regions of Canada which include Quebec, Praririe region, Ontario, Atlantic region and British Columbia. I decided to choose 'pop_center' because I believe it more clearly categorizes the different lifestyles of urban and rural.

'income_family'

This section is pretty self explanatory, it provides information on a person's family income. It is categorical in nature as discussed above briefly. The sections include: Less than $25,000, $25,000 to $49,999, $50,000 to $74,999, $50,000 to $74,999, $100,000 to $ 124,999, $125,000 and more. I chose this to compare with 'feelings_life' and find out more about the saying "money can't buy happiness" and to find out if it is has any merit. Unfortunately though, due to this variable being categorical, and ending at 125k, it is not easy to draw any detailed conclusions from this, especially for the individuals with very wealthy families. A rough correlation is the best we can get.

I chose this over the 'income_respondent' section which provides information on the individual person's income because family income provides more information than personal income in some cases, especially when dealing with general lifestyle. An example of this is if the person surveyed is in school and not working, they might be earning close to nothing but their lifestyle is that of someone earning 50k to 75k because their partner or parents are earning that much.

'religion_has_affiliation'

This section provides information about whether or not a person considers themselves as religious. It has 3 options: Has religious affiliation, No religious affiliation and Don't know. I chose this to also compare with 'feelings_life' because religion is very likely to have an

effect on a person's outlook in life. Some people get meaning from their religious beliefs while others feel more free being non-religious and some are in the middle or unsure.

There are two other sections such as 'religion_importance' and 'religion_participation'. I decided to choose 'religion_has_affiliation' because the other two only make sense if a person is religious so it would leave out the non religious. Combining all three of the religion sections will definitely give us more of a detailed analysis on this topic but since my focus is not soley on this topic, I am sticking to just the section with information on whether or not a person is religious.

'self_rated_health' and 'self_rated_mental_health'

These two sections are a person's self rated physical and mental health respectively. The selections for this are: poor, fair, good, very good, excellent. I chose these to be the internal factors that I would like to use to find a relationship with 'feelings_life'. This will help my analysis as I am trying to compare and contrast internal versus the external factors discussed above on a person's overall perceived quality of life.

Model

The model used for this analysis is a Logistical Regression. This type of statistical model that uses a logistic function to model the relationship between a binary response variable and one or more explanatory variables. This type of model is great for scenarios where one might need to predict a binary outcome given several factors that can be continuous or categorical. For example, when determining whether or not it will rain tomorrow in a city given factors that might influence the weather or in my case, whether or not a person is satisfied/happy with their life given their religious status, whether they live in a rural or urban location, their family income bracket, their self rated physical health and mental health. Putting this data in a logistic regression model can allow me to predict the probability of someone being satisfied or happy with their lives, given any specific factors covered in the model.

Mathematically, this model is represented as:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Notation Summary:

- p is the probability of the event of interest occurring. In my case, this is whether or not a person's life satisfaction is "good" or "bad".
- Beta 0 (B0) is the intercept of the y axis which represents the log odds of a person's life satisfaction being "good" given the default factors for every explanatory variable. x1 = x2 = ... = xk = 0.

- The rest of the coefficients B1,B2,...,Bk represent the change in log odds for every one unit increase in the respective x1,x2,...,xk. So for my case, every new category in the explanatory variables will be a new unit increase. For example having an income bracket of 25k to 49k being unit 1 increasing to 50k to 75k being unit 2.
- The predictor variables can be numerical or categorical

As we can see based on the mathematical formula, logistical regression does not fit a line on the data plot like linear regression does, instead it fits an 'S' shaped natural logarithmic based function to the data. This model fits the curve using maximum likelihood estimation.

The main reason I used this model is because it allows me to work with the categorical variables given in the GSS dataset. As mentioned in the Data Section, the vast majority of the survey answers were categorical so it would not suit a regression model which works best with continuous response and explanatory variables. There was an important variable I added to represent my response variable that was necessary for this model to work. It was the 'feelings_life' response variable, this was discussed above and I had to make another variable with a binary result, it is called 'binary_feelings_life'. I needed to do this because logistical regression requires the response variable to be binary. Finally, the software used to run this model is R run on RStudio.

Logistic regressions are great for probabilistic interpretation, is less prone to over-fitting when dealing with fewer parameters and can be updated easily. A weaknesses however, is that it tends to perform poorly when there dealing with a higher number of parameters.

Results

Figure 2: Summary Statistics from the Logistic Regression Model

Coefficients:

	Estimate	**Pr(>|z|)**
(Intercept)	-0.16246	0.739457
as.factor(pop_center)Rural areas and small population centres (non CMA/CA)	0.42609	< 2e-16
as.factor(income_family)$125,000 and more	0.06366	0.363060
as.factor(income_family)$25,000 to $49,999	-0.41775	8.85e-10
as.factor(income_family)$50,000 to $74,999	-0.26061	0.000213
as.factor(income_family)$75,000 to $99,999	-0.22288	0.002565
as.factor(income_family)Less than $25,000	-0.64336	< 2e-16
as.factor(religion_has_affiliation)Has religious affiliation	0.32545	0.091816
as.factor(religion_has_affiliation)No religious affiliation	0.02830	0.885050
as.factor(self_rated_health)Excellent	1.19213	0.000216
as.factor(self_rated_health)Fair	0.21519	0.503618
as.factor(self_rated_health)Good	0.62238	0.051209
as.factor(self_rated_health)Poor	-0.17462	0.595350
as.factor(self_rated_health)Very good	0.89226	0.005254
as.factor(self_rated_mental_health)Excellent	1.21329	0.000240
as.factor(self_rated_mental_health)Fair	-1.27955	0.000125
as.factor(self_rated_mental_health)Good	-0.21587	0.510845
as.factor(self_rated_mental_health)Poor	-1.72153	2.13e-06
as.factor(self_rated_mental_health)Very good	0.60250	0.066981

Figure 2, the summary statistics contain all the information we need for this analysis. The two sections that we will be focused on are 'Estimate' and 'Pr(>|z|)' which is the p-value. The 'Estimate' section tells us, for each answer to the survey questions, the factor of increase or decrease in log odds of a person rating their life satisfaction as 'good', given that they gave that specific answer. There is one answer that does not show up for three of the questions, like for a person who lives in an urban area, earns 100k to 125k or is not sure about their religious affiliation. These answers have an effect of 0 on the log odds as they are the default answers we compare other answers to.

Looking at the 'Estimate' column starting from the top, we can now see that living in a rural area increases the log odds of a person being satisfied with their lives by a factor of 0.42 on average compared with people living in urban areas which has a factor of 0 and having a higher family income up until around 100k, has a noticeable positive affect on a person's life satisfaction, after that it seems to level off. Being religious also seems to have a noticeable positive affect on a person's happiness. For the internal causes, we can clearly see now that mental and physical health both have a profound affect on life satisfaction, even more than any of the external factors. Excellent mental and physical health both increase the log odds by 1.21 and 1.19 respectively which is much higher than any other determinant. Poor mental health though, seems to have a significantly larger negative affect on a person's wellbeing than poor physical health does, having factors of -1.72 and -0.17 respectively.

Going over to the 'Pr(>|z|)' section, this can tell us whether or not any one of the factors in the 'Estimate' section is statistically significant. We will use a significance level of $p = 0.05$ so that means any 'Pr(>|z|)' entry of greater than 0.05 is not statistically significant and therefore we do not have enough evidence that a particular answer to a question with a p-value > 0.05 will give us any information about whether or not a person is satisfied with their life.

We can see that none of the answers dealing with religious affiliation have a p-value < 0.05 so therefore it seems that religious affiliation data doesn't give us enough evidence that it really affects life satisfaction even though the factor was large in the 'Estimates' section. Living in a rural area is highly statistically significant and therefore does have a profound positive effect on life satisfaction compared to living in an urban area. For family income levels, it seems that all factors are statistically significant except for the two categories above 100k, which is where the probability of life satisfaction levels out as discussed above. There are good reasons for this which will be discussed in the next section. Now, looking at the significance of the internal factors, we can see that self rated physical health only has evidence of affecting self rated life satisfaction (positively) if it is 'Very good' or 'Excellent', otherwise there is no proof of a relationship between the two variables. Self rated mental health on the other hand is statistically significant, affecting life satisfaction if it is rated as 'Excellent' (positive), 'Fair' (negative) or 'Poor' (negative)

I have decided against using any other visuals like graphs or charts because all my variables are categorical. It would be inefficient, as it would take many graphs/charts to display any meaningful information with the relationship/correlation of the variables.

Discussion

The goal of this study was to find out whether internal factors have a greater affect than external factors on a person's life satisfaction. The external variables tested were: 1) 'pop_center' whether they lived in an urban or rural area 2) 'income_family' their family income bracket 3) 'religion_has_affiliation' their religious affiliation

The internal variables tested were: 1) 'self_rated_health' self rated physical health 2) 'self_rated_mental_health' self rated mental health

It can be concluded that internal variables play a larger role in determining whether or not a person is satisfied with their lives because the factors in the 'Estimate' section of Figure 2 have a much greater magnitude as discussed in the last section. In addition, we also discovered that self rated physical health levels are do not affect a person's life satisfaction unless their health is 'Very good' or 'Excellent', while self rated mental health levels have an effect to a person's life satisfaction on both extremes. The mental health affect makes sense because it is pretty well researched how much mental health can affect a person's day to day mood and outlook for the future. On the other hand, physical health levels might have this behaviour because having great health is seen as a blessing and minor health issues are normalized in society. I was not expecting negative physical health scores to not be correlated with lower life satisfaction, but here are some of my best guesses as to why they are not. We can see that with the sheer number of goods and services dedicated to improving physical health, most people are not fully content with their physical health and have normalized it to the point that they don't consider it as a big negative in their lives while those who genuinely feel healthy might be more grateful and are less likely to take it for granted. There might be some flaws in the data and my model though that have led us to this result which I will discuss in more detail later.

For the external variables, we can see that although two of the three are related to a person's self rated life satisfaction, the magnitude of affect is smaller. People who live in a more sparsely populated rural area seem to be more content with their lives compared to their urban dwelling counterparts. This may be due to the fact that an urban lifestyle is more fast paced which is more likely to generate higher stress levels in people. It could also be due to the different demographics of the two areas, maybe people living in rural areas spend more time with their family or have bigger families compared to people in urban areas. Another thing affecting this could be the type of career people tend to have in an urban versus rural location. This analysis is not sufficient to provide enough information on this to prove any causation.

Money also seems to correlate with happiness, up until a certain point. As shown in my results section, family income has a steady positive relationship on life satisfaction up until the 100k bracket ranges. After that point, it levels off. This makes sense when applied to the real world because there is a big difference in the quality of life of people who don't have enough money for essentials compared to those who do. After reaching a certain income bracket (100k in this case), all or most of the necessities like good food, clothes, shelter, a vehicle, etc. are met. Any more income is just a bonus at that point, hence why it does not have a noticeable affect on self rated life satisfaction.

Lastly, whether or not a person is religion does not give any information about a person's life satisfaction according to this analysis, this is because the p-values for those answers were not statistically significant as discussed above. The result was not very surprising because although this had to do with where a person finds meaning in life, there are a lot of other places people find meaning in life outside of religion so therefore, they will be content in this regard.

Weaknesses

A general weakness of this study is that it is not as focused and deals with many variables. It could have been better to focus on a relationship between fewer variables so that the study could go into greater depth. In the real world, it would probably been better to split the study up into smaller studies looking at relationships between 2 variables and then put all the studies together to combine all the data. Another weakness which relates more to the data is the catagorical nature of most of the variables. This makes it inefficient to create meaningful plots when looking at the relationship between two variables or more. The data only allows us to make big picture analyses and leaves out some details as discussed in the Data section.

Another weakness with the data are the subjectivity of the survey answers. It is impossible to know whether the same answer from two people mean the same thing when dealing with subjective questions such as self rated mental or physical health. The response variable is also difficult to quantify, someone with arguably worse living conditions might rate themselves higher in life satisfaction than another person with arguably better living conditions. This subjectivity and lead us to misleading results and cause flaws in the data.

Lastly, this is a correlation study so there is no way to know the cause and effect of any of these results. We do not know cause or effect of life satisfaction versus mental health/physical health, income, religion. Is it more likely that a negative outlook in life causes bad mental or physical health or vice versa. Also does living in a rural area give people more satisfaction in life or do people with a better outlook in life tend to prefer the rural lifestyle. Some of these can be researched more to better understand the cause and effect relationship which brings us to the next steps.

Next Steps

The next steps are to tackle the weaknesses of my model and data. More surveys should be given that focus on the specific topics covered so that more detailed data will be attained. For example, Urban versus Rural should be made continuous and the data for every person should be the precise population density per square kilometer instead and family income should be continuous and give to as close to the exact income to the dollar as possible. For the internal variables, there should be more objectivity involved. For example: medical records, BMI and other stats for physical health evaluations, and also letting medical professionals rate a person's health on a scale to avoid subjectivity. A mental health evaluation by a trained professional could also be given for more objectivity.

Getting more continuous variables will allow us to run a linear regression model as well to analyze these relationships in greater depth. More research is also needed to help understand the cause and effect behind some of these factors. This will be very helpful in allowing us to know more about how to help improve everyone's life satisfaction.

References

Government of Canada, Statistics Canada. *Statistics Canada: Canada's National Statistical Agency*, 17 Oct. 2020, www.statcan.gc.ca/eng/start.

Grover, Khushnuma. "Advantages and Disadvantages of Logistic Regression." *OpenGenus IQ: Learn Computer Science*, OpenGenus IQ: Learn Computer Science, 23 June 2020, iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/.

"Tidyverse Packages." Tidyverse, www.tidyverse.org/packages/.

Appendix

Github link for project folder: https://github.com/AdelBoufama/PS2