The Effects of Income, Working Hours, and House Ownership on Happiness

Benjamin Zhang, Daniel Leung, Diane Kim, Pablo Mercado

10/19/2020

Abstract

The purpose of this report is the use of multivariate linear regression (MLR) to model out how indicators of success can predict the happiness of the person. MLR is a model that takes multiple explanatory variables and uses them to predict the outcome of a response variable. Specifically, by testing if using Income, House Ownership, and Working hours can predict the happiness level of a person. This was done with the use of the 2017 GSS dataset which collected data on families including the variables needed for the model. The results of model can simplified to finding that Income and Working Hours are good indicators for Happiness, but not Home Ownership.

Introduction

In our society, the perception of one's success is often associated with possession and power. However is the social perception of success truly correlated with individual happiness?

In 2018, a study from Harvard University investigated the overall happiness of over 4000 millionaires (Donnelly et al.) From the study, a moderate positive correlation between wealth and overall millionaire happiness was found. However, the happiness of millionaires was dependent on how their money was obtained. Individuals who both earned a lot and obtained their money through self merit often showed the greatest amount of self satisfaction.

While correlation between wealth and happiness may be significant for millionaires, it is erroneous to extrapolate this claim to the rest of society. From the 2019 world happiness report, Canadians currently rank 9th internationally (22-23). The majority of Canadians, despite being vastly poorer than millionaires or billionaires, experience happiness and content to a greater degree than normal. So therefore, to what degree do monetary gains and other social images of success contribute to the overall happiness of the average Canadian?

By investigating the Canadian GSS database, this study seeks to: 1. Develop a multivariate linear regression model using quantifications of "success": Family Income, House Ownership, and Average Hours Worked per week as a set of predictors for happiness within individual Canadians. 2. Using the model, see whether these stereotypes hold any true correlation or meaning.

Data

The data selected for this study is from the 2017 GSS, the latest version of the survey that is publicly available online. The GSS collects a wide range of information from its participants including age, work/workplace, religion, and house ownership. As a result the dataset is a treasure trove of possible predictor variables for our model. Specific to this study we have chosen to focus on family income, house ownership, and average hours worked per week.

The 2017 GSS has a strong focus on family and aims to study the future of families as they become increasingly diverse. They answer questions like: How many families are there in Canada? What are their characteristics

and socio-economic conditions? What do families at different stages of life look like? How common are step or single-parent families?

The GGS data was collected via stratified random sampling where each province was a stratum. The population of interest are all Canadians and the population frame were the participants that qualified to participate in the survey. There was a record of phone numbers from a census and other files where participants were all contacted. Households that did not own a phone or did not have at least one person over the age of 15 were excluded. From those who did meet the requirements, participants were randomly selected and interviewed.

One of the strengths of this dataset is that it managed to collect a large amount of data for each participant. This is seen from the sheer number of variables collected and the variety of them as well. The goal of the GSS was to gather data on social trends in order to monitor changes in the living conditions and well-being of Canadians over time.

Several restrictions to this data include the overall sample size. Despite collecting information on 20,602 individuals, this quantity is considerably smaller compared to the entire population of Canada. Questions may arise on whether the sampled population is an accurate representation of the target population. Biases in the data may favor individuals with frequent telephone accessibility, underepreresting individuals who live in poverty or dislike phone calls from government funded institutions, such as First nations individuals.

Furthermore, there are several missing or unusable values within our data set. About half of the rows from our selected columns contain null values, making these values unusable for data analysis. Furthermore, the columns of "house ownership" and "average hours worked" contain several 'I don't know' responses, which are difficult when attempting to place meaning to it.

However, if we choose to ignore these values, we lose 35% of our entire data (from 20k to 12,900) making the data set even smaller. Most of the recorded answers are in categorical form, leading to discrete distributions which often increases variance when analyzing correlations. Specifically a major issue with family income values are the wide ranges that are covered. The data categorizes the total family income values per year into 25 thousand dollar intervals. This range is considerably large and is a major source of error.

Figure 1: Count of feelings of life categorized by whether they own or rent

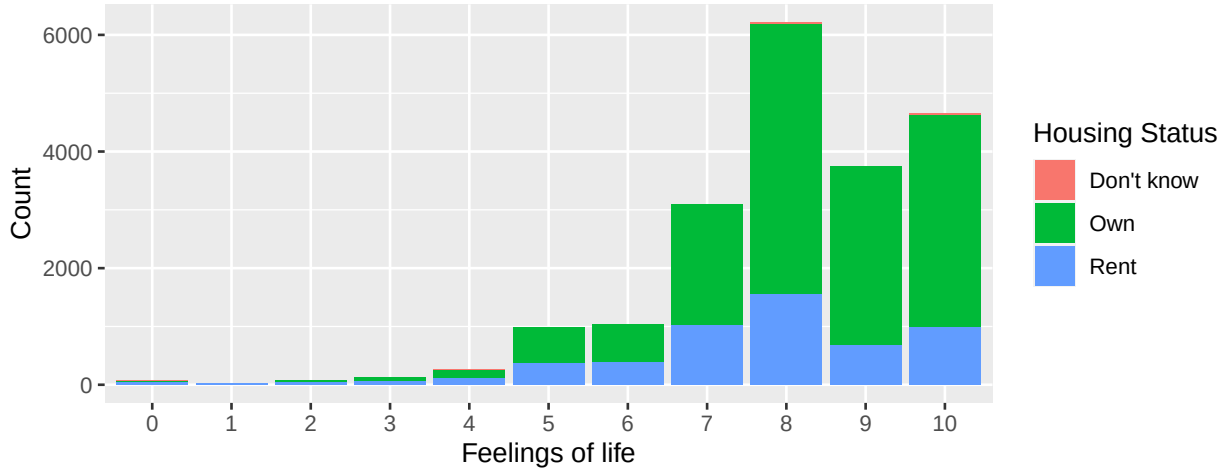


Figure 2: Count of feelings of life categorized income of the household

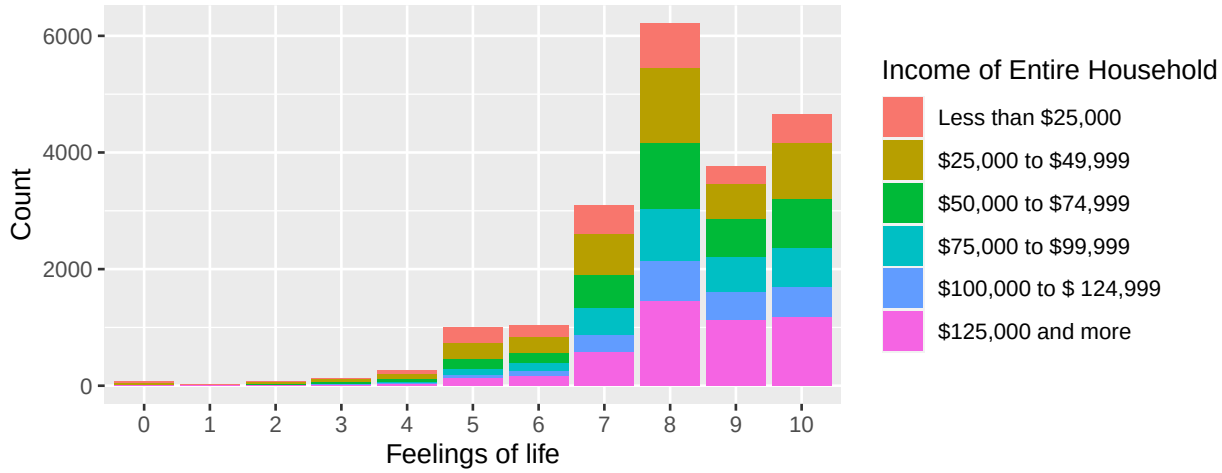
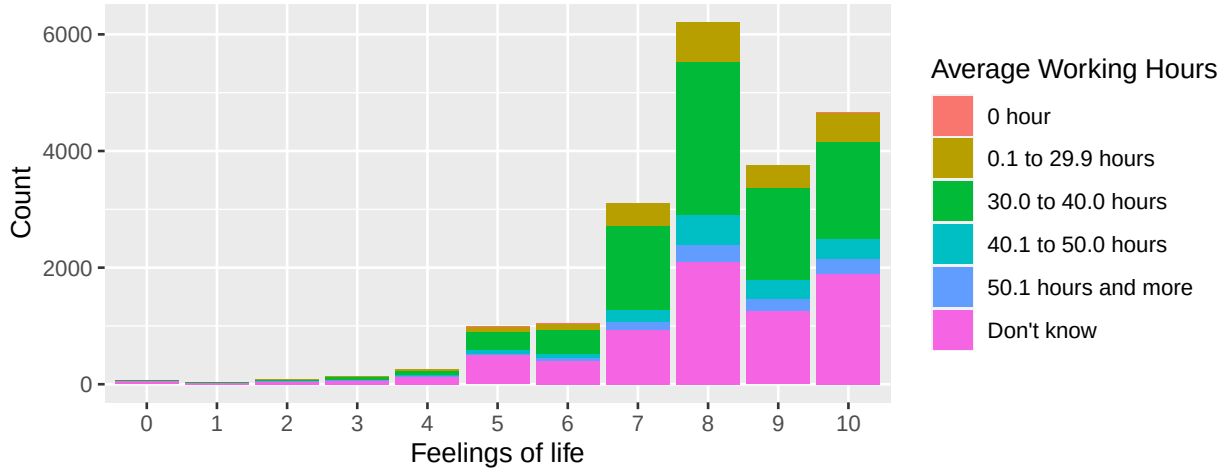


Figure 3: Count of feelings of life categorized by average working hours



Looking at home ownership (Figure 1), most people buy a house instead of renting among people who have a satisfaction level of 7 and above. With a happiness level below 7, the split between homeowners and landlords is nearly equal. The number of persons in each income group (Figure 2) seems reasonably stable in comparison to their level of satisfaction. In each of the segments, an vast portion of an specific income group does not seem to exist.

Model

Our goal is to attempt to predict which variables correlate with overall happiness for the citizens of Canada.The model we have currently selected is the multivariate linear regression model, meaning that we are creating an equation using many explanatory variables to predict the response variable. In our case, we are using Household Income, Housing Status, and Working Hours to predict Happiness.

We chose to use multivariate linear regression over simple linear regression, logistic regression, or bayesian inference. Simple linear regression would not be appropriate for the purpose of our paper since we are interested in the effects of multiple variables on an outcome. We could have used logistic regression however, that would mean mutating our outcome variable, happiness (an ordinal scale from 0-10), into a binary outcome. For example, participants that respond 5 and below become one category (unhappy) and 6 and above becomes another (happy). However, it would be subjective to decide at what point a person is happy. Lastly, we did not use bayesian inference because that requires prior knowledge about the parameters. Although similar

studies have been done by the GSS, there were not many comparable variables between data sets.

$$Happiness = \beta + \beta_1 x_{income_i} + \beta_2 x_{own_i} + \beta_3 x_{work_i}$$

This general equation represents the model we are going to create. Where Happiness is equal to the combination of Income, Housing Status, and Working Hours. Since all of the predictor variables are categorical, the different levels of each variable is denoted as i in the general model. Also for that same reason, the base case of the equation is the scenario that you make less than $25k in a year, you do not know your current housing status, and on average you work 0 hours a week.

For Household Income, we are using the range of less than $25k a year to over more than $125k a year. For Housing Status, we are using the statuses of owning, renting, and I don't know. For Average Working Hours, we are using the range from 0 hours worked to 50.1 or more hours worked. In the scenario where the cell was NA, we mutated those cells to match the "I don't know" cells which were already present. Our interpretation of I don't know depends on the variable. In the case of Housing Status, it could indicate homelessness or they are reluctant to share that information. Similarly, the "I don't know" in Working Hours represents people with no jobs or reluctance to share that information.

We specifically chose these values because they are qualitative measurements of success. Family income can be seen as an access to wealth a person has, meaning the more money a person has access to the more successful that person is. Home Status has often been associated with sucess as a major milestone in society is finally being able to "pay off the house." For Working Hours people usually don't like working. People associate success as working when you want to for how long you want to. Greater hours worked generally demonstrates a greater need for money, or even multiple jobs.

This model is best suited for our study as It allows us to see the overall impacts of each variable on happiness. It also allows us to compare and isolate variables that are most influential by viewing them all at once.

As aforementioned, the independent variables of our linear regression line are recorded categorically within the survey, dummy variables were used instead. In this method each category would be associated with its own weight parameter Beta which would correlate with its effect on a person's overall perception of happiness.

Results

Table 1: Table 3

Variable	Coefficients	P.values
Intercept	6.28541	< 2e-16
Income 1: $25,000 to $49,999	0.37389	3.46e-15
Income 2: $50,000 to $74,999	0.54493	< 2e-16
Income 3: $75,000 to $99,999	0.62581	< 2e-16
Income 4: $100,000 to $ 124,999	0.75909	< 2e-16
Income 5: $125,000 and more	0.80657	< 2e-16
Own	0.44886	0.0859
Rent	0.09092	0.7290
work 1: 0.1 to 29.9 hours	0.90671	0.0205
work 2: 30.0 to 40.0 hours	0.86671	0.0263
work 3: 40.1 to 50.0 hours	0.95304	0.0150
work 4: 50.1 hours and more	1.01095	0.0101
work 5: Don't know	0.97984	0.0121

Because our data was taken from a survey, we must use survey weights by using the values of the target and sample populations of 30,538,825 and 20,331 respectively. The target population was calculated by taking using Stats Canada and filtering out all territories and well as ages 14 and below. The sample population was

the number of people taking the survey. With the population correction, we used survey-weighted generalised linear model funtion to create our MLR. Table 1 is the summary of that MLR which includes the coefficients which will be used for the equation, and the p-values which will be used to test if the variable is a good indicator for Happiness.

Table 2: Table 2

Statistics	Numbers
R-Squared	0.044000
Adjusted R-Squared	0.043000
RSE	2.587778

To test how well the model fits, we also found R-squared, adjusted R-squared, and the mean squared error (MSE) which can be found in table 2. The R-squared numbers represent the variance for the dependent variables for the MLR and MSE is an estimator for the difference between the actual values and the fitted values. Both R-squared values are very small at around 0.4 and the MSE is at 2.59.

$$Happiness = 6.29 + 0.37x_{income_1} + 0.54x_{income_2} + 0.62x_{income_3} + 0.62x_{income_4} + 0.62x_{income_5}$$
$$+0.45x_{own} + 0.09x_{rent}$$
$$+0.91x_{work_1} + 0.87x_{work_2} + 0.95x_{work_3} + 1.01x_{work_4} + 0.98x_{work_5}$$

This is the full model's equation based on the survey weighted linear regression that was done. This equation showcases all levels within each variable and it was created by taking the coefficients from the regression found from table 1. The representation of what each specific level means can also be found in Table 1.

Figure 4: Comparison between the coefficients of the income categories

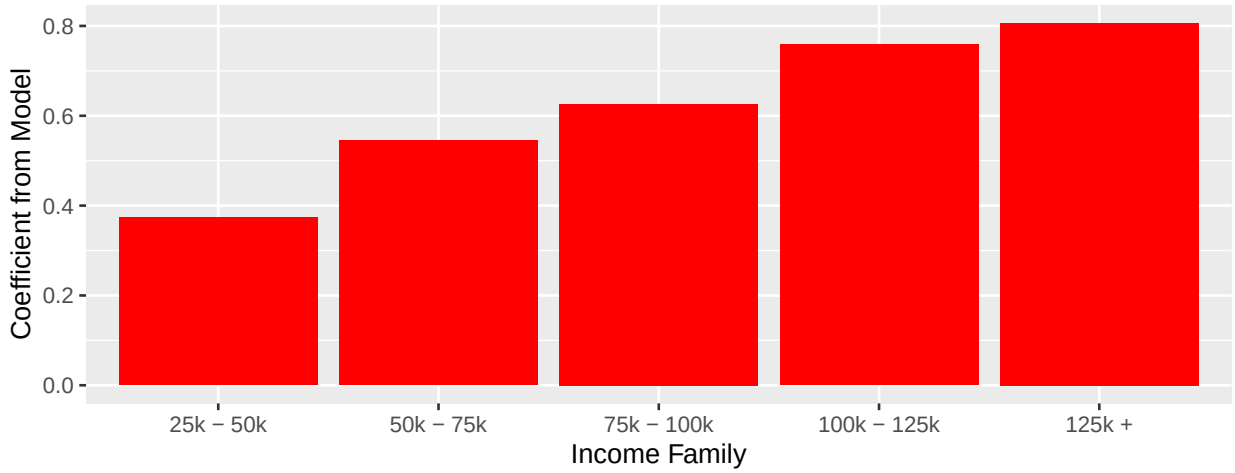


Figure 5: Comparison between the coefficients of owning and renting

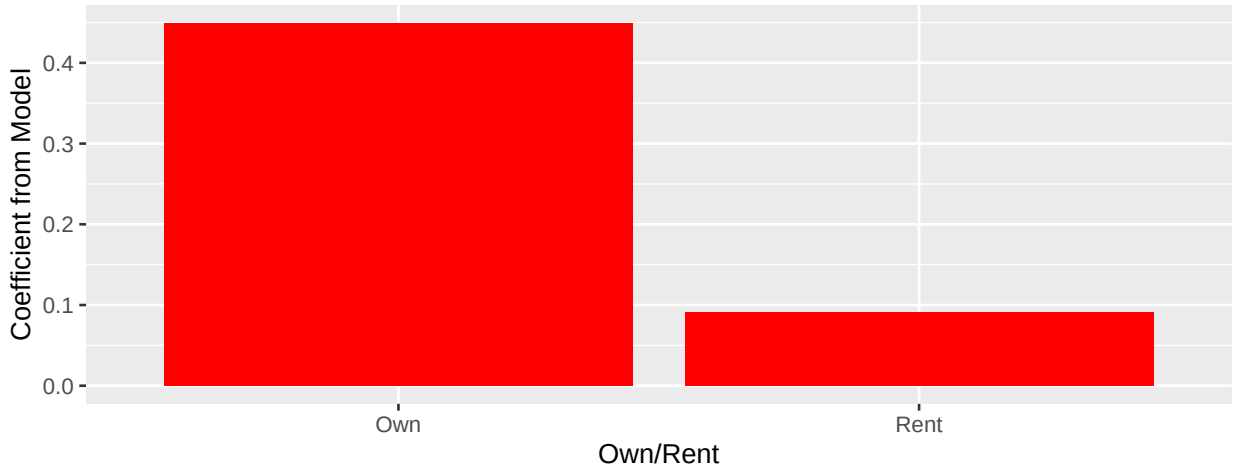
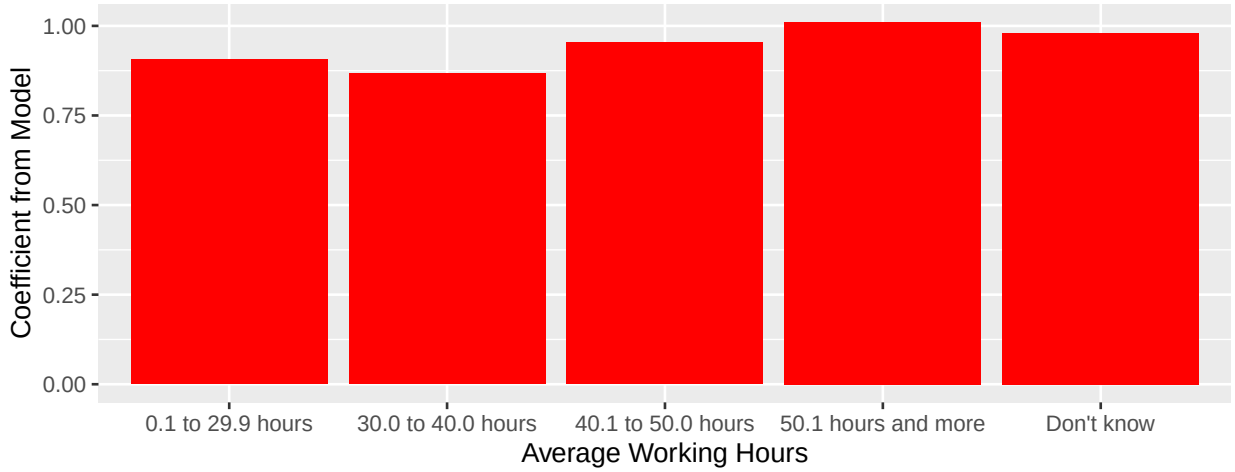


Figure 6: Comparison between the coefficients of the working hours categories



Figures 4-6 showcase the differences between the coefficients of the category for each variable. This showcases how much more happy you are depnding on which category you fall under for each variable. In Figure 4, it shows that the more income your household has, the more happy you are. In Figure 5, it shows that owning your home will make you more happier than renting your home. In Figure 6, that there not much difference on how happy a person is depending on how many hours they work.

Discussion

Null Hypothesis: that happiness is NOT correlated with high income, home ownership and low working hours

Alt Hypothesis: that happiness IS correlated with high income, home ownership and low working hours

The goal of this report is to investigate whether or not our social measurements of success correlate with general happiness using a multivariable linear regression. These measures are family income, house ownership, and amount of working hours. We assume a significance value of 0.05. Based on the p-values in the summary of the svyglm function, we can see that owning a home and renting a home is not significant to one's happiness. However, what's interesting is that there is a large difference between the coefficients as seen in Figure 5: Comparison between coefficients of owning and renting graph. This suggests that owning a house does indeed have a much stronger positive effect on happiness than renting. But again, the p-values for both suggest that there's not a correlation. So perhaps home ownership may not be a good measurement towards feelings of life. Perhaps as long as someone has a place to stay, the type of ownership does not have much effect on their

happiness.

What's noteworthy is that every category within average hours worked supports our hypothesis, with working 1-29.9 hours and working 50.1 hours or more being the two strongest effects (looking at their coefficients in Figure 6.). However, there seems to be only slight differences between the rest of the working hour categories based on the bars.

We see a similar case with every category within family income. As we see in the Figure 4: Comparison between the coefficients of income categories graph, the coefficients get larger as family income increases. So making more money could equate to a higher happiness rating.

Next, we have segmented bar graphs (refer to figures 1-3), so we can get a quick look into the frequency of incomes, home ownership and average working hours within each feelings of life rating. We can also distinguish any patterns that may be present in the data.

Looking at home ownership (figure 1), among the people who have a happiness rating of 7 and above, most people own a house rather than renting. With a happiness rating below 7, there's approximately an even split between homeowners and renters. So a majority of happier people are homeowners or at least members of a household. But as our multivariable linear analysis showed, there doesn't seem to be any correlation between home ownership and happiness. This can mean that owning a house can contribute to happiness, but is not necessary and just simply renting a house is enough for a good amount of happiness.

The amount of people in each income category (figure 2) in relation to their happiness level seems fairly consistent. There doesn't seem to be an overwhelming majority of an income group in any of the segments. This suggests that level of income does not contribute to the overall level of happiness. This finding agrees with what Donnelly et al. said about people being happier when they earn their wealth, rather than inheriting it.

Based on the average working hours plot (figure 3), happier people work around 30-40 hours or don't know how many hours that they work. This suggests that working a normal day job or not keeping track of the hours that they work contributes to happiness.

Overall, the indicators of Family Income and Average Hours Worked are good predictors for Happiness based on the rejection of the null hypothesis. Home Ownership, however failed to reject the null hypothesis so it may not be a good indicator for Happiness.

Weaknesses

A major issue with the data is that the variables are discrete and are defined categorically which increases error size when attempting to fit a linear regression model. This error can be specifically seen within the family income variable, especially between the 25 000-50 000 annual income range, which can define the difference between lower and middle class individuals. It would have been beneficial if exact values for income could be provided, allowing us to better analyse the distribution of this data to better fit our model.

Another issue, specific to the use of the family income variable, is its use to measure and predict the financial state of the respondent's household. An individual who lives alone is the single contributor to the family income value. Therefore, a single person with a higher paying job may possess the same family income value as that of two working parents taking care of a large family. The family income value does not necessarily reveal the nature of financial strain and actual available wealth. This financial strain is the true pressure that affects the overall happiness of individuals. People who have more money to spare can afford to spend more on luxury goods and services, such as additional cars and vacations. While family income is a major factor of determining the amount of financial strain an individual is experiencing it doesn't necessarily give the entire picture. As a result, trying to predict happiness from this family income may prove difficult and inaccurate. A possible solution to this could be to create a new variable such as "average house income per member of family". This way we can incorporate the stress of providing for additional individuals within a family and better quantify financial strains.

Survey data used in this paper was limited to the provinces in Canada, excluding the territories. This means that our sample is actually not representative of the population because it completely leaves out an important group of people. For example, First Nations people are excluded from the data but they take up a significant portion of the population of Canada. Additionally, extrapolations to our findings may be limited to other countries, especially to countries that are not similar to Canada. For example, countries with different cultures, wealth, government, etc.

The data had a large amount of missing data which negatively affected our ability to perform our regression analysis. This could be attributed to the following: The survey itself was very extensive and asked many questions which is good for us analysing that data. But for the participants answering these questions, it could be too time consuming and it's possible that the length of the survey contributed to some (or much) of the missing data; The survey also asks about personal information that some participants don't feel comfortable disclosing such sensitive information or simply don't want to answer. This also allows for people to leave empty responses or respond differently than the truth.

Next Steps

For our next steps, we could collect the same data including the same variables but in a slightly different setting. For example, use the United States as our population. Then, we can use a Bayesian inference to perform analysis on the effects of family income, working hours, and home ownership. This is because we will have a prior distribution that we can use. Additionally, this would allow us to compare the results from Canada with other countries and see how well our model represents them.

We found that the number of work hours had correlation with happiness and so it may be beneficial to do a follow up survey focusing on occupation. Collecting data about each participant's occupations, workplace conditions, job satisfaction, etc.

Although there was a focus on family in this GSS, it would also be nice to explore the levels of happiness among people that don't live with or have families. In other words, it would be interesting to determine how large of a role that family plays in happiness and how it compares to the happiness of people without them.

References

Donnelly, G. E., Zheng, T., Haisley, E., & Norton, M. I. (2018). The Amount and Source of Millionaires' Wealth (Moderately) Predict Their Happiness. Personality and Social Psychology Bulletin, 44(5), 684-699. doi:10.1177/0146167217744766

Helliwell, J. F., Huang, H., Wang, S., & Norton, M. (2020). Social Environments for World Happiness. In World Happiness Report 2020 (8th ed., pp. 22-23). Sustainable Development Solutions Network.

Government of Canada, Statistics Canada. Population Estimates on July 1st, by Age and Sex, Government of Canada, Statistics Canada, 29 Sept. 2020, www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000501.

Holtz, Yan. "Grouped, Stacked and Percent Stacked Barplot in ggplot2." – The R Graph Gallery, 2018, www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2.html.

Technology, Advancing Knowledge through. Computing in the Humanities and Social Sciences, 2010, www.chass.utoronto.ca/.

Yihui Xie, J. J. Allaire. "R Markdown: The Definitive Guide." 2.5 Markdown Syntax, 14 Oct. 2020, bookdown.org/yihui/rmarkdown/markdown-syntax.html.