Should You Be Rich? Using Age, Gender, and Happiness to Predict Income Level

James Bai, Pamela De Vera, Hyoeun Park, Shlok Somani

October 19, 2020

Abstract

Conducted by *Statistics Canada* every five years, the *General Social Survey* gathers socio-economic characteristics of Canadians across the board to monitor and improve their living conditions and well-being. With the help of this expansive data collection, the following paper utilizes regression analysis to determine the relationship between the interviewees' score of feelings towards life with their level of income. It further investigates the impact of other variables such as age and gender to analyse their effect on the aforementioned relationship.Taking a quantitative approach to answer a highly qualitative query, this research will capitalize on economic disparity to provide a fresh lens from which one can view the modern society.

I. Introduction

Just like beauty, satisfaction lies in the eye of the beholder. Based on one's perception of success and fulfillment, satisfaction may mean entirely different things. To demonstrate using purely stereotypical roles and dispositions, a Wall-Street banker may earn upwards of $400k and still be unhappy, while a sky-diving instructor with a $50k/year salary could have achieved everything they want and be extremely satisfied in life. This is precisely why it can be established that 'feelings towards life' is a fragile concept. It has no one, true definition. Our inherent recency and cognitive biases, anchoring effect and natural-human irrationality, leaves us often unable to judge our level of happiness in life at that very moment.

Leaving out those who are gifted with the ability to look past the 'now', we will define life satisfaction as an amalgamation of your past, present and your future. In today's world, much of your lifestyle depends on your income level. The choices you make – whether it be how you spend your Friday night or your investing tactics to save for retirement – are all dependent on the number of figures on your paycheque.

Diving down further, the relationship between one's income and life satisfaction is also largely impacted by personal factors such as age, gender and status of full-time employment. The research and our analysis suggests that there is a strong relationship between these variables allowing us to extract useful insights from this data. Utilizing logistic regression to construct a model which will give better understanding of how income affects the feelings of life based on sex and age, we deduced that:

As age and how an individual scores their feelings towards life increases, the probability the individual has a high level of income also increases. Additionally, if the individual reports as male, this would further increase the probability of a higher income.

This basic discovery was a necessary part of the journey that allowed us to further establish a general conclusion: these three factors are essential in establishing a core relationship between one's feelings towards life and their income level. The presence of these elements gives us the opportunity to incorporate further data and socio-economic variables including, but not limited to, level of education, location etc.

Given these outcomes, we can return to our original dilemma regarding the ambiguity of satisfaction. And even though these do not fully provide us with a comprehensive understanding of what life satisfaction truly means - we can leave that to the philosophy majors - we have taken the first step in identifying the composition of this convoluted metric.

II. Data

II-i. *Population and Sampling Techniques*

The data of the 2017 GSS (*General Social Survey*) was collected through CATI[1] (Computer Assisted Telephone Interviews). Respondents were interviewed in either English or French. Phone calls were made from 9:00 a.m. to 9:30 p.m. Mondays to Fridays, 10:00 a.m. to 5:00 p.m. on Saturday and 1:00 p.m. to 9:00 p.m. on Sunday.

The sampling strategy used for this survey was the stratified random sampling technique[2]. With a total of 27 strata (the groups in which the population is sorted), 14[3] of them are part of the CMAs(Census Metropolitan Areas), 3 of the stratum are the remaining CMAs[4] located in Quebec, Ontario, and British Columbia. The non-CMA areas of each province make up the remaining 10 strata. A SRSWR (simple random sample without replacement) would then be performed in each stratum.

The target sample size for this survey is 20,000. A minimum sample size for each stratum is needed to ensure an acceptable estimation for every stratum. Once that minimum is reached, the rest would be distributed to the strata that would balance the precision of both nation-level and stratum-level estimates. The amount of people who responded to this survey, which is the sample population, is 20,602. However, it is important to note that the response rate of this survey was 52.4%, so that means around 39,300 people were called and only 20,602 of them responded to this survey.

II-ii. *Data of Interest*

For our statistical analysis, we will use age, sex, income of respondents, average hours worked per week, and score of their feelings towards life. We will also generate a new variable defining the level of income each respondent earns (further mentioned in this section). Note that we define age and score of their feelings towards life as continuous variables, and the remaining variables as categorical.

The survey data provides the income of each respondents in 6 different ranges (hence a categorical variable) increasing in $25,000 increments. Similarly, the average hours worked per week are placed under 3 different ranges and a “Don’t know” response. This variable will be used for data cleaning purposes only. The score of feelings towards life was a scale from 0-10, with 0 being “very dissatisfied” and 10 being “very satisfied”. Although this can be treated as a categorical variable, we will treat it as quantitative for practical analysis purposes. We also have age recorded based on the month the survey was responded, which makes age values include up to one decimal place. Sex is simply defining whether the individual is a female or male. Later in our model, we will assign each sex as dummy variables, with male equaling to 1 and female equaling to 0.

The data gives the range of each respondent’s income when *Statistics Canada* actually has access to the exact income values. Placing each income under a range is adequate for privacy purposes, but creates difficulty for our analysis. Hence, for practicality, we will create a new variable by splitting the income of each respondent into two levels – high or low – based on their amount of income[5]. That is, each individual will be assigned to a homogeneous income stratum. Further simple random sampling has not been employed in this analysis,

[1] CATI is a system where the interviewer calls the respondents and asks them questions following the interviewer manual, then the interviewer would enter the response onto the computer.

[2] Stratified random sampling technique is splitting the total population into groups with the same characteristic. Ideally, we want the same ratio of a sample size to the stratum population for each stratum. For example, if a sample of 20 people from a company consisting of 80 men and 20 women needed to be collected using stratified random sampling techniques, the sample would consist of 16 men and 4 women.

[3] The 14 CMAs are St. John’s, Halifax, Saint John, Montreal, Quebec City, Toronto, Ottawa, Hamilton, Winnipeg, Regina, Saskatoon, Calgary, Edmonton and Vancouver

[4] with the exception of Moncton, which would be counted as part of the non-CMA stratum of New Brunswick.

[5] If an individual earns less than $50,000 (exclusive), they belong to the low-income level, and high-income level if otherwise. This criteria was based on the government’s announcement regarding OSAP in 2017. When Kathleen Wynne was the Premier in 2017, the government stated families earning less than $50,000 were eligible for “free tuition”.

since the larger number of observations we study, the more representative it becomes of the population we are interested in.

We are also specifically interested in analyzing those who are between the age of 18-52 (in 2017) and work on an average of 30 hours or more per week (i.e., full-time workers). We chose the age range as such since 18 is when one becomes a legal adult and 52 (born in 1965) is the maximum age in Generation X (based on 2017). Hence, we clean the data using RStudio and the tidyverse package in order to filter those who meet these criteria. We are now left with 6965 matches after data cleaning. This is the size of our study population[6].

In order to gain insight of the distributions of each explanatory variables (age, sex, score of feelings towards life) by income level, we created plots and tables. The figures and tables are demonstrated in Sections II-iii - v.

II-iii. *Age Distribution by Income Level*

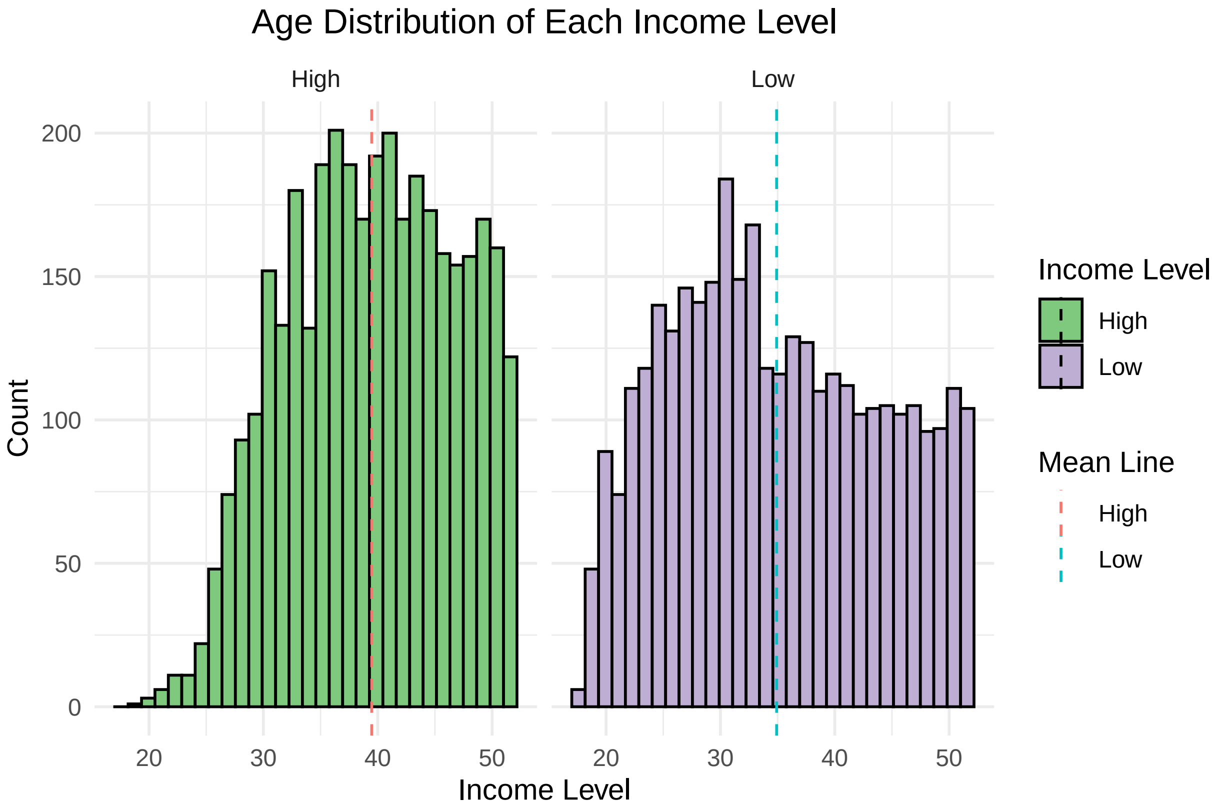

Figure 1

Table 1: Generations in Each Income Group

Income Level	Gen Z (18-20)	Millennials (21-36)	Gen X (37-52)	Mean Age	Median	Standard Deviation	Total
High	2	1201	2200	39.47	39.6	7.218311	3558
Low	110	1748	1391	34.91	34.0	9.154247	3407

Figure 1 is a histogram depicting the age distribution of the high and low-income groups. Table 1 gives some summary statistics of Figure 1. Overall, we can observe that the number of individuals increases as the age

[6] Equivalent term as sampled population.

increases in the high-income group, whereas the number decreases as the age increases in the low-income group. Table 1 shows that 61.26% of Generation X earn a high income, while only 40.73% of Millennials earn a high income. Another notable aspect is that there are only 2 respondents from Generation Z earning $50,000 or more in income, meaning only 1.8% of the respondents from Generation Z have a high level of income. We notice in the low-income group that the mean of 34.91 exceeds the median of 34, revealing that there are more people under the age of 34.91 earning a low level of income than those over the age of 34.91. Similarly, the mean age (39.47) is less than the median age (39.60) in the high-income group implying that the majority of individuals are older than the mean age in this group. In fact, the center (mean and median) of age in the high-income group is Generation X, and that in the low-income group is Millennial.

II-iv. *Satisfaction towards Life by Income Level*

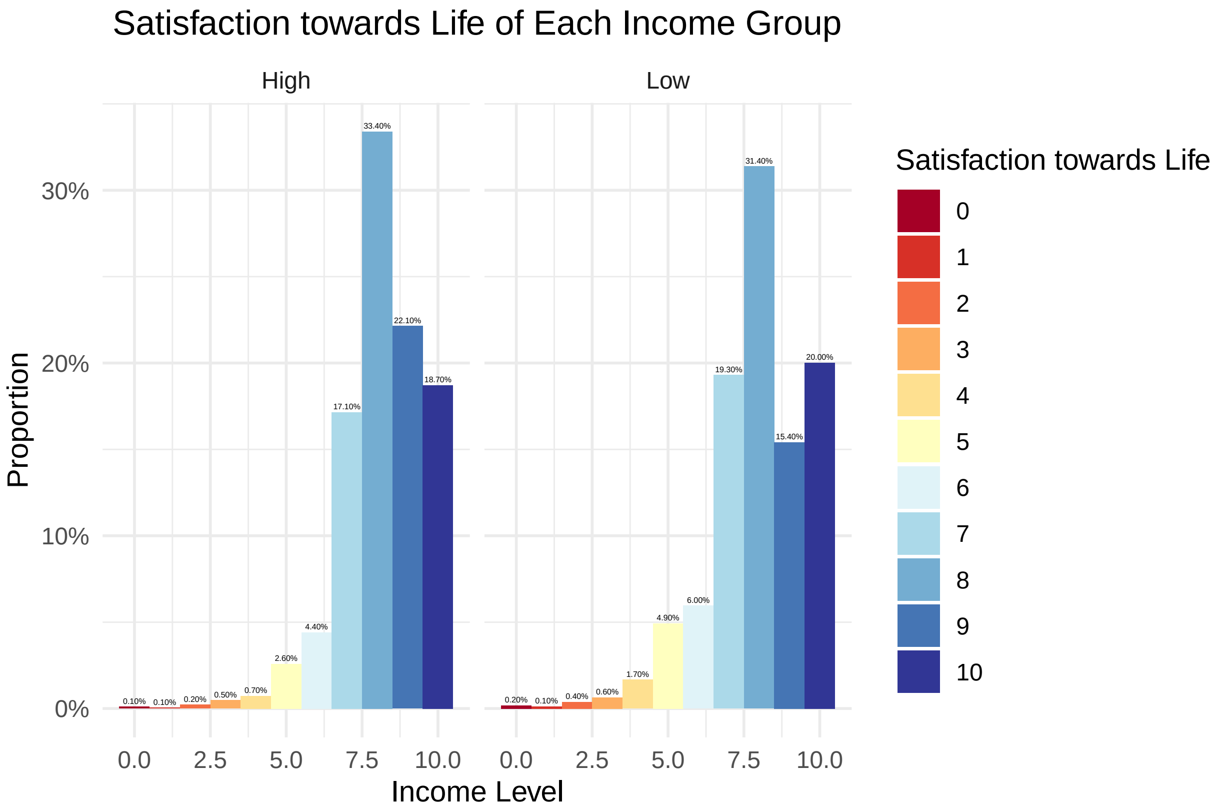

Figure 2

Table 2: Satisfaction towards Life (by Income Group)

Income Level	Mean	Q1	Median	Q3	IQR	Outliers	Standard Deviation	Total
High	8.1790	7	8	9	2	31	1.3915	3558
Low	7.9498	7	8	9	2	45	1.6016	3407

Figure 2 above is a bar plot that depicts how an individual scores their feelings towards life based on the different levels of income. From Table 2, we observe that the high and low income groups have 3558 and 3407 observations, respectively. Thus, the plot includes the proportion of satisfaction towards life (instead of counts) at the y-axis due to the different number of observations between the two income groups.

We can discern that the two groups share a common distribution shape. Both groups are left-skewed (or have a negative skew) with light tails on the right-hand side. This means the score of the satisfaction towards life from 0 to 3 was low in both income groups. Only 0.87% and 1.32% from the high and low-income groups, respectively, gave a response in this range. All of these responses are outliers for each group. The standard deviations explain why there are more outliers in the low-income group. The standard deviation for the low-income group (1.6016) is higher than the high-income group (1.3915), meaning the spread of data points is larger in the former. Another remarkable aspect is that the mode and median for both groups are both 8. Both have a peak at 8, but the higher income group having more proportion in 8 (33.39%) than it does in the low-income group (31.38%). We also note that the mean (8.18) for the high-income group is higher than the low-income group (7.95).

II-v. ***Sex by Income Level***

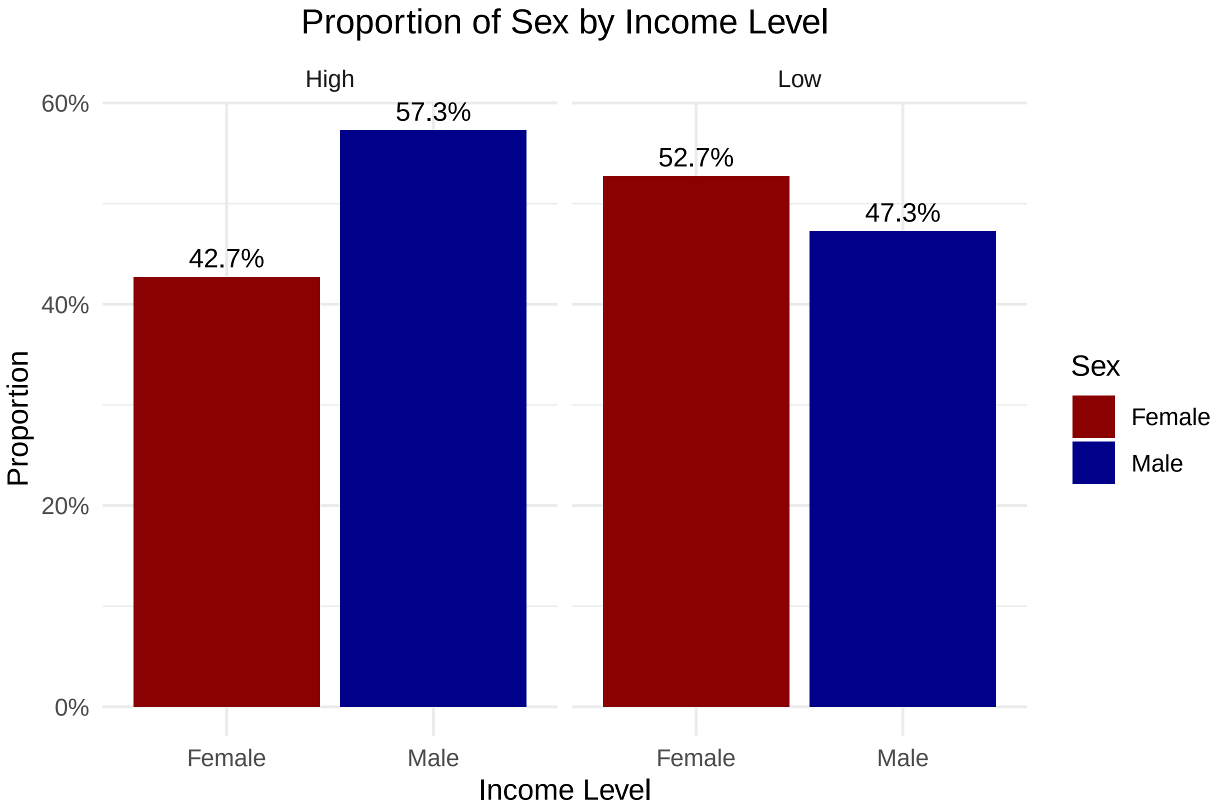

Figure 3

Table 3: Difference in Income Level by Sex

Sex	High Income	Low Income	High Income (%)	Low Income (%)	Total
Female	1520	1797	0.46	0.54	3317
Male	2038	1610	0.56	0.44	3648

Figure 3 is another bar plot visualizing the proportion of males and females for each income level. The proportion has been computed instead of the number of counts for the same reason mentioned in Figure 2.

We observe that the proportion of males is higher in the high-income group, whereas the proportion of females is higher in the low-income group. Specifically, there are 14.6% more males than females in the high-income group and 5.4% more females than males in the low-income group. Looking at Table 3, we also notice that there are 331 more male respondents than female in the data and that approximately 56% of them belong to the high-income group. On the other hand, more than half of the females are in the low-income group at 54%.

III. Model

For our model, we will be looking at how sex, age, and how an individual would score their feelings towards life can predict a person's income level. To do this, we will be using a logistic regression model with a finite population correction. This type of model will allow us to predict the probability of a specific event occurring, given other conditions. The conditions are known as the predictor variables, and the event in question is the response variable. A finite population correction will allow us to adjust the resulting variance, which measures how much the data deviates from its average, according to only those in the population who are not in the sample.

We will have two levels of income, low and high. We will only use two income levels as this model is best used for binary response variables. In the data provided by the 2017 GSS, income was published as a categorical variable. Since the income of each respondent was assigned as one of 6 different intervals increasing in $25,000 increments, it would be difficult to perform analysis on the actual value of the income as it is not provided.

Our predictor variables for the model, which we will use to predict how income level changes, will be age, sex, and how an individual ranks their feelings toward their life. We will use the respondent's age and their score on their feelings towards life as a continuous variable. We will define both as continuous because age is reported as numbers rounded to one decimal place on the range 15 to 80 and respondents score their feelings on a scale from 0 to 10.

This model is the optimal model to use compared to a multilinear regression model. In multilinear regression, which produces a linear equation relating changes in the predictor variables to changes in the response variable, we can have more than one predictor variable. However, this model works best when both the predictor variables and the response variables are continuous. This would not be ideal as we have two categorical variables in question.

The regression will give us the logarithmic equation of the approximate probability that a person has a high level of income given their age, sex, and score of how they feel towards life. The equation will follow the form below where p is the probability of the event in question occurring, x_i is a predictor variable, where i is age, sex, or the score of an individual's feelings towards life, $beta_0$ is the intercept value, and β_i is the coefficient weighing each x_i's change in the log probability of the event with a one unit increase in x_i.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{age}x_{age} + \beta_{male}x_{male} + \beta_{feel}x_{feel}$$

The survey package in RStudio will be used to run the model. We will also be calling on the tidyverse, and arm packages.

In order to assess the fit of the logistic model, we can use a binned residual plot. This graph takes bins based on the fitted equation obtained from the model and plots the average of the fitted values against the average value of their corresponding residuals in each bin. Two symmetrical lines on the plot represent $\pm(2SE)$ in which we should expect 95% of the data points to fall between. We will also be looking at the residual deviance of the model. The residual deviance measures how different our logistic model is relative to a hypothetical model that fits the data perfectly. We will look at the deviance of the model with all three predictor variables and compare it to the deviances of the model without one of the three predictor variables. If the deviance decreases by at least 1, then it would indicate that the remove predictor variable improves the fit of the model.

IV. Results

Based on the summary of the model, we arrive at the equation:

$$log(\frac{p}{1-p}) = 0.112663x_{feel} + 0.461070x_{male} + 0.067185x_{age} - 3.607194$$

Coefficient	Estimate	Std. Error	t-value	p-value
(Intercept)	-3.6071937	0.1837646	-19.6294292	$1.5249878 \times 10^{-83}$
Feelings towards life	0.1126631	0.0168723	6.6773922	$2.6198094 \times 10^{-11}$
If male	0.4610704	0.050464	9.1366182	8.308673×10^{-20}
Age	0.0671853	0.0030162	22.2751245	$3.0944462 \times 10^{-106}$

Table 4: Model Summary

For the intercept estimate and the estimates for the coefficients in regards to sex and age, we observe p-values less than $2 \cdot 10^{-16}$, indicating significance of the estimate values. Also noting that the p-value corresponding to the estimate for the coefficient of the score of feelings towards life $1.16 \cdot 10^{-15}$, we can recognize significance in this estimate value as well.

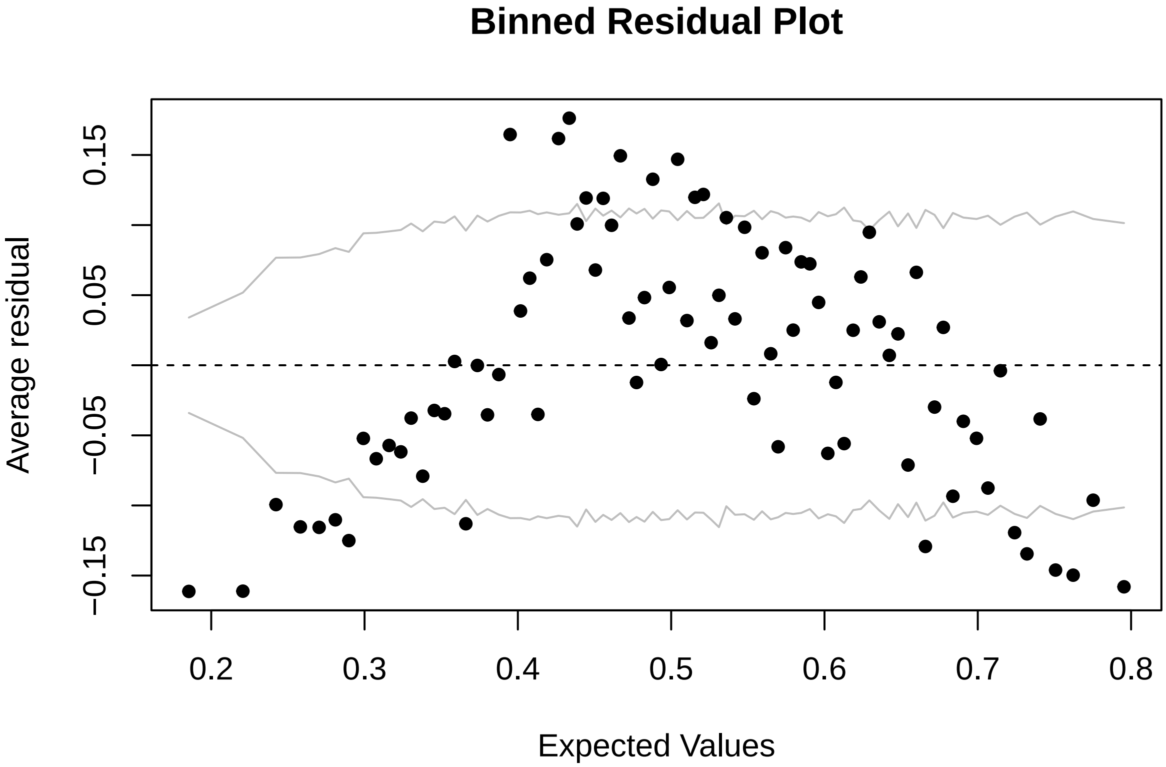

Figure 4

Deviance:	Null	With all Predictors	Without Age	Without Sex	Without Feelings
	9652.2663147	9017.6584358	9536.6754583	9101.9760189	9062.8523695

Table 5: Residual Deviance

Regarding the fit of the model, we observe a residual deviance of 9017.658. Noticing the null deviance of 9652.266, which is the deviance of the model without any of the predictor variables, we recognize that collectively, the predictor variables improve the fit of the model. Futhermore, the separate residual deviances of the model without age, sex, and score on feelings towards life are 9536.675, 9101.976, 9062.852, respectively. We can notice that as each variable is added as a predictor to income level, the residual deviance decreases, indicating that each predictor variable improves the fit of the model. Looking at the binned residual plot in Figure 4, we notice 29% of the points are not within the standard error band, which is evidence that this model is not the best fit as we can expect 95% of the points to fall within the bounds if the model has a reasonable fit.

V. Discussion

V-i. *Conclusion*

Based on the obtained equation from the regression model, we notice that all the coefficients for each predictor variable is positive. From this, we can see that as each predictor variable increases, then the log probability would also increase. That is, as the age and feelings a person has towards their own life increase, the probability that this person has a high level of income also increases. Additionally, if the person reports as male, their probability of having a high income also increases.

More specifically, suppose we have a 45 year old male who ranks their feelings towards their life as 10. Given this information, the probability that this individual has a high level of income would then be:

$$p = \frac{e^{\beta_0 + \beta_{age} x_{age} + \beta_{male} x_{male} + \beta_{feel} x_{feel}}}{1 + e^{\beta_0 + \beta_{age} x_{age} + \beta_{male} x_{male} + \beta_{feel} x_{feel}}}$$

$$p = \frac{e^{0.067185 * 45 + 0.461070 * 1 + 0.112663 * 10}}{1 + e^{0.067185 * 45 + 0.461070 * 1 + 0.112663 * 10}} = 0.990156$$

The intercept value would indicate the log probability of a person having a high level of income given they are female, are 0 years of age, and rank of 0 on their feelings towards their life. Since these conditions are out of the scope of the data, we cannot interpret the intercept value as such.

V-ii. *Weaknesses*

When we created the two high and low income strata with $50,000 as the baseline, we only took the income of respondents into account. The drawback of this is that there are other various factors that could have been considered when creating such strata. For instance, an individual is identified to be in low income if their household income is below Canada's Low Income Cut-Off (LICO). In other words, the government takes the economic status of one's family income into consideration when determining individuals that qualify as low income. Hence, dividing the strata solely based on the income of respondents might not be an accurate representation of the actual income division in Canada.

Another aspect we can touch on for improvements is the logistic regression model. The binned residual plot we have created for it revealed that approximately 29% of the data points were out of the standard error bounds. This gives evidence that the model might not be of best fit. However, from the deviance, we can still notice that the predictor variables do improve the model fit, since the deviance decreases as each predictor gets added to the equation of the logistic regression model. Thus, it is plausible that there could be more predictor variables that we have not considered in the model which could improve the model fit.

One of the variables we can possibly include as a predictor in our fitted model could be the provinces of Canada. Since the provinces across Canada differ in minimum wage, it is plausible to have it as a lurking variable in our data analysis. According to the official website of *Monster*, the minimum wage for Alberta in 2017 was $12.20 whereas it was 80 cents less in Ontario. Such circumstances can make differences in the income, and ultimately impact the validity of our fitted model.

Last notable limitation of this paper is that the *General Social Survey* was conducted in 2017. This means that the data we have employed for the main objective of this paper is outdated. Moreover, it is hard to predict the real world of 2020 based on what we have analyzed due to unprecedented situations of COVID-19.

V-iii. *Next Steps*

One thing we might be able to improve on next time we analyze a newer GSS is that we can try to find a better model that would fit the data. In the 2017 GSS, they didn't ask for specific incomes, that restricted us to using a logistic regression model as we didn't have the data to plot a scatter plot. We could also improve on our knowledge on stats to find a better model, as it is possible that there is a better model out there that would fit this model perfectly that we didn't learn yet.

Another thing we noticed while analysing this data is that the age distribution is that it heavily favours towards older people. There are significantly more elders who responded to the survey in comparison to younger generations. Also since the response rate is 52.4%, this might explain why the age distribution is skewed left. So we could suggest it to *Statistics Canada* to add another strata sampling technique when collecting data. That strata would divide the population into different age groups, so this way we will get a more accurate representation of the Population.

VI. References

1. Beaupré, P. (2020). General Social Survey Cycle 31: Families Public Use Microdata File Documentation and User's Guide. Ottawa, WA: Statistic Canada
2. Chow, J. (n.d.). *Ggplot2 Reference and Examples (Part 2) - Colours.* Retrieved October 18, 2020, from http://rstudio-pubs-static.s3.amazonaws.com/5312_98fc1aba2d5740dd849a5ab797cc2c8d.html
3. Crossman, E. (2020, January 03). *Low-income and Immigration: An Overview and Future Directions for Research.* Government of Canada. Retrieved October 18, 2020, from https://www.canada.ca/en/immigration-refugees-citizenship/corporate/reports-statistics/research/low-income-immigration-overview-future-directions-research.html
4. Faculty of Arts & Sciences, University of Toronto. (2017). *General social survey on Family (cycle 31), 2017.* Computing in the Humanities and Social Sciences. http://www.chass.utoronto.ca/index.html
5. Fry, R. (2020, August 28). *Millennials overtake Baby Boomers as America's largest generation.* Retrieved October 18, 2020, from https://www.pewresearch.org/fact-tank/2020/04/28/millennials-overtake-baby-boomers-as-americas-largest-generation/
6. Gelman, A., et al. (2020). *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models.* Retrieved October 18, 2020, from https://cran.r-project.org/package=arm
7. *Ggplot2 Quick Reference: Colour (and fill).* (n.d.). Retrieved October 18, 2020, from http://sape.inf.usi.ch/quick-reference/ggplot2/colour
8. *How to plot a 'percentage plot' with ggplot2.* (2016, November 03). Retrieved October 18, 2020, from https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/
9. Lumley T. (2020). *survey: analysis of complex survey samples.* R package version 4.0.
10. Portugués, E. (2018, January 20). *Lab notes for Statistics for Social Sciences II: Multivariate Techniques.* Retrieved October 18, 2020, from https://bookdown.org/egarpor/SSS2-UC3M/
11. Statistics Canada. (2017). *General social survey cycle 31 main survey - Family [Data set].* Statistics Canada. Retrieved October 18, 2020, from https://www23.statcan.gc.ca/imdb/p3Instr.pl?Function=getInstrumentList&Item_Id=335815&UL=1V&
12. Stephenson, L.B., Harell, A., Rubenson, D., Loewen, P.J. (2020). *2019 Canadian Election Study - Online Survey.* Harvard Dataverse, V1. https://doi.org/10.7910/DVN/DUS88V
13. Swartz, M. (2017). *Minimum Wages Across Canada In 2017.* Retrieved October 18, 2020, from https://www.monster.ca/career-advice/article/minimum-wages-across-canada-in-2017
14. Webb, J. (2017, September 03). *Course Notes for IS 6489, Statistics and Predictive Analytics.* Retrieved October 18, 2020, from https://bookdown.org/jefftemplewebb/IS-6489/logistic-regression.html
15. Wickham, H., et al. (2019, November 19). *Welcome to the Tidyverse.* Retrieved October 18, 2020, from https://tidyverse.tidyverse.org/articles/paper.html

16. Xie Y (2020). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.30, from https://yihui.org/knitr/

Appendix A

Code

Here is the link to our github repository: https://github.com/hynprk/Using-Age-Gender-and-Happiness-to-Predict-Income-Level