Predicting probability of life satisfaction level among adult Canadians with multiple factors in life

Steven Tran, Nayoung Kim

October 19th 2020

Abstract

This report analyzes how mature Canadians feel about their life, provided certain factors in life. Deriving data from the 2017 General Social Survey, this analysis utilizes a dependent variable that has two categorical levels- whether the respondent individual answered feeling generally happy, or the respondent was overall not content with life. Income of the respondent illustrated to play a role in influencing the life rating, whereas other factors in life did not portray a significance in affecting the rating. This was a meaningful discovery as higher income was found to be associated with higher life happiness level in general.

Introduction

Here is where you should give insight into the setting and introduce the goal of the analysis. Here you can introduce ideas and basic concepts regarding the study setting and the potential model. Again, this is the introduction, so you should be explaining the importance of the work that is ahead and hopefully build some suspense for the reader. You can also highlight what will be included in the subsequent sections.

The goal of this analysis is to discover a relationship between the outcome variable of contentness of life and the auxiliary variables of factors in life by utilizing a logistic regression model. In this analysis, the study variable is feelings about life, and five auxiliary variables were selected to test if they indeed have a relationship with the outcome variable. Feelings about life were measured as generally happy if the rating was five or higher out of 10, while they were grouped as overall not content if the rating was lower than 5. The chosen dependent variables are age at first birth, current age of the respondent, income of the respondent, place of birth whether it is within or ouside of Canada, and future intention of having any children. While there were numerous variables in the data that were available to select, these five specific variables were intuitively picked to check the hypothesis that these will have an association with life rating. The importance of the work is that through the logistic regression model, an analysis can be built about how adult individuals in Canada feel about their life, considering the input variables. More precisely, the probability of an adult Canadian having a contented life can be calculated. Ultimately, the main goal is to observe if any of these life factors affect life rating, and if so, how large the influence is. In the subsequent sections, the model section is covered and explained using tables and the results section through a graph, while weaknesses and next steps are also discussed.

Data

Introduce the data, explain why it was selected. Make sure to comment on important features and highlight any potential drawbacks to the data.

The selected output variable is feelings about life, and the auxiliary variables are age of the respondent, age at first birth, place of birth (whether it was in Canada or not), income of the respondent, and intention of having future children. We chose this study variable in the intent to scrutinize the life satisfaction rating of Canadians, whether this is dependent on certain factors in life. Also, we decided on the specific input variables because intuitively, they are the most suitable data for our hypothesis and to prove that indeed, there is a linear relationship between the y and x variables. A drawback to the data is that because of the numerous 'NA' responses, our population data size greatly decreased from the original sampled population number. Another drawback to the data is that there are numerous vague responses such as "Don't know" or "Unsure", which are difficult to place in order.

Figure 1: Distribution of feelings_life

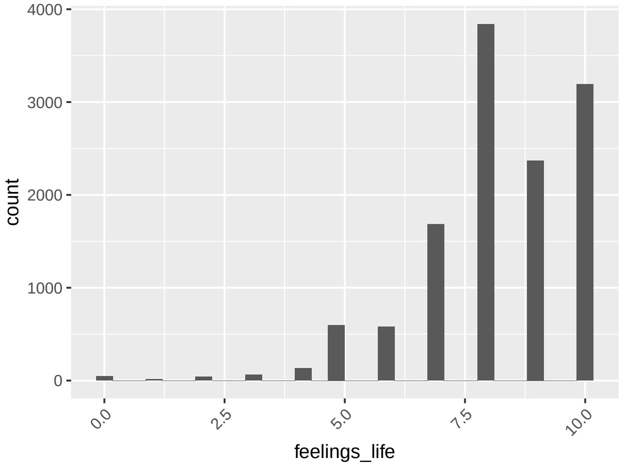


Figure 2: Distribution of age

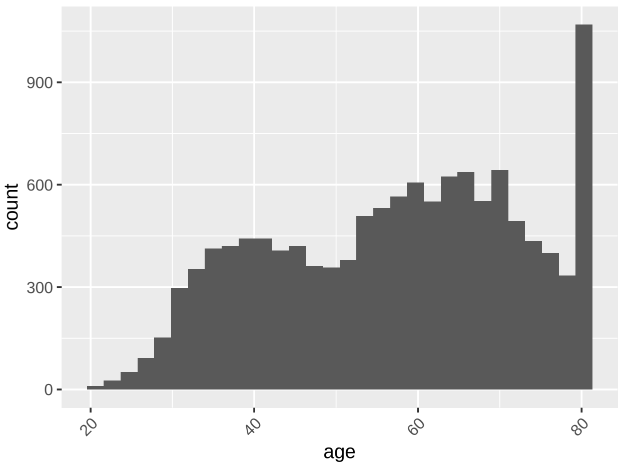


Figure 3: Distribution of age_at_first_birth

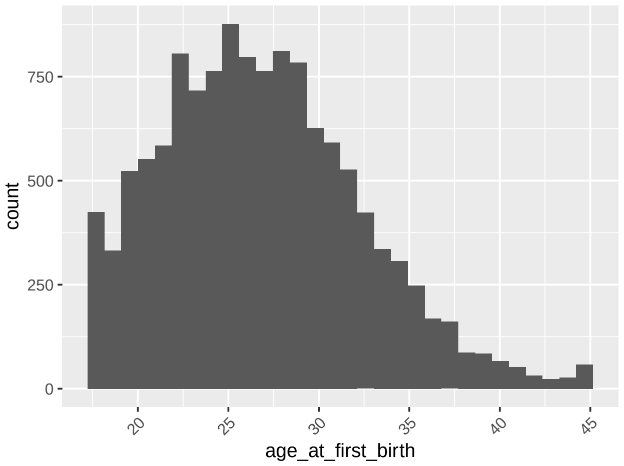


Figure 4: Distribution of income_respondent

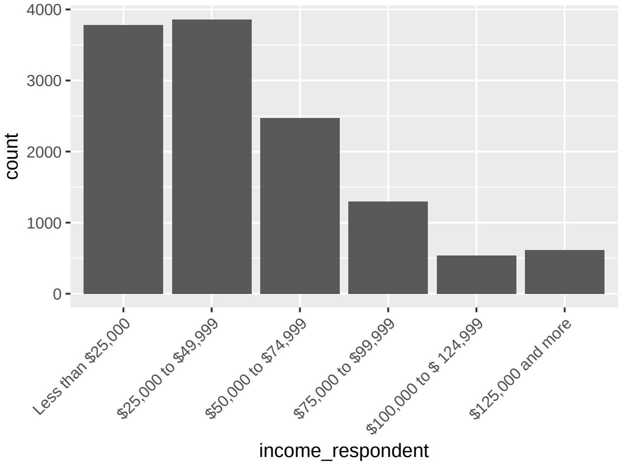


Figure 5: Distribution of place_birth_canada

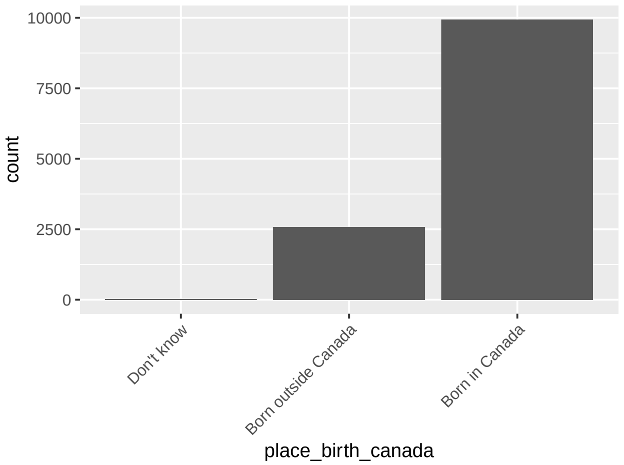


Figure 6: Distribution of future_children_intention

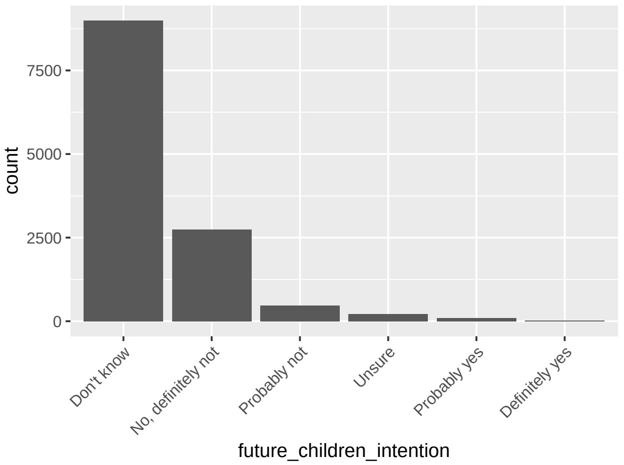


Figure 7: Mean feelings_life by province

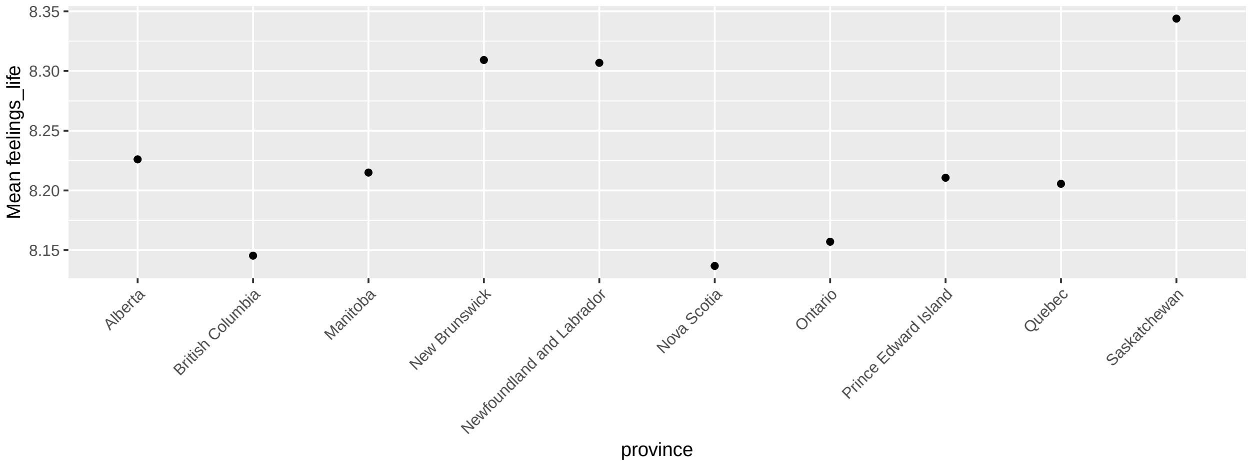


Figure 8: Mean age by province

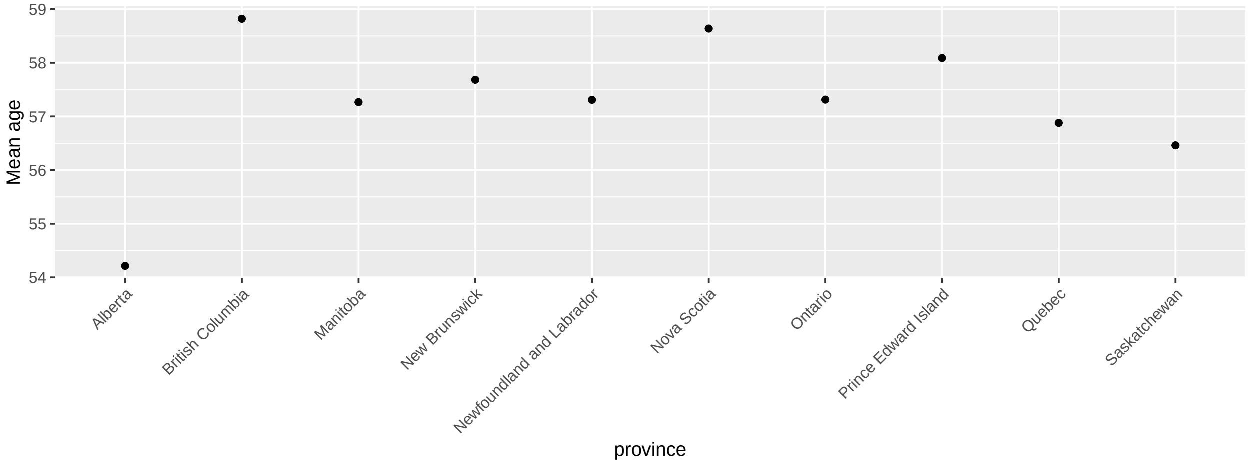


Figure 9: Mean age_at_first_birth by province

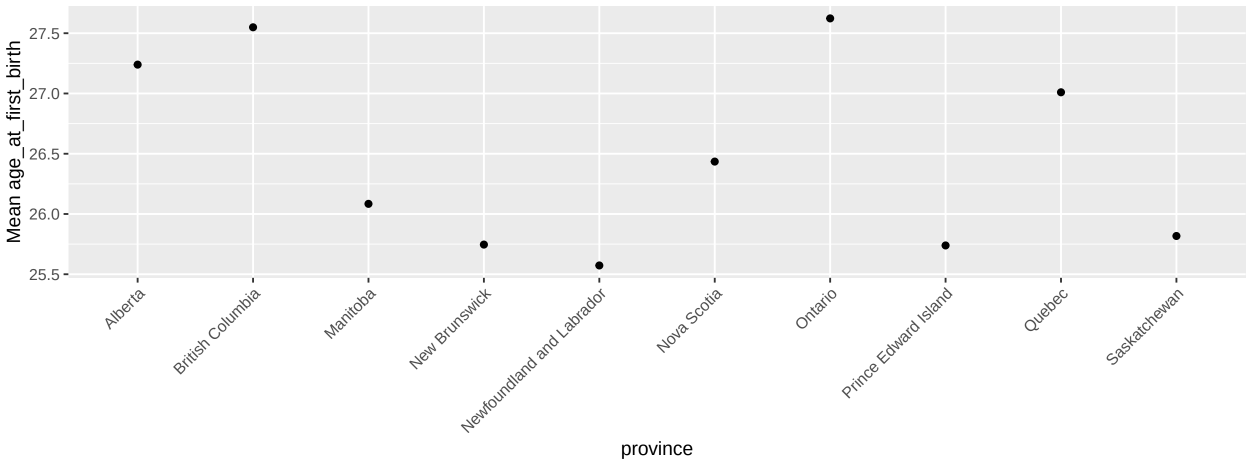
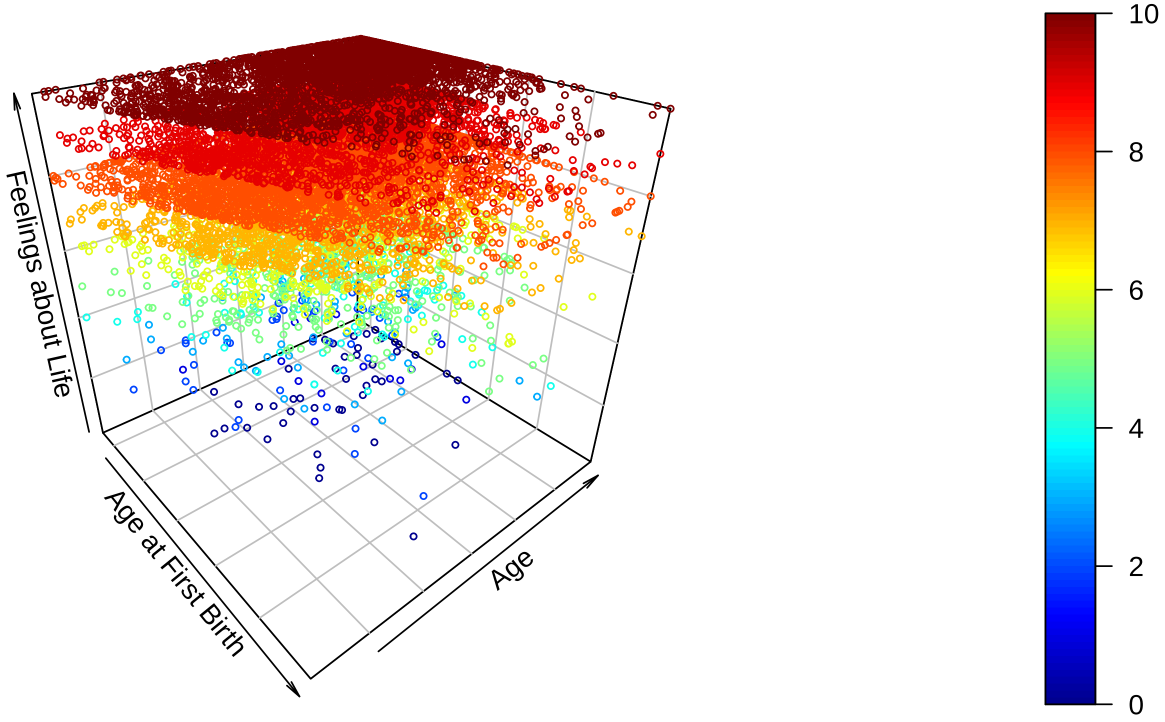


Figure 10: Age at First Birth & Age vs Feelings About Life



Model

Introduce the selected model here. It is expected that you will use some mathematical notation here. If you do please ensure that all notation is explained. You may also want to discuss any special (hypothetical) cases of your model here, as well as any caveats.

To predict the probability of a person being happy, which we define as:

$$\text{Prob}(h) := \begin{cases} 1, & \text{if feelings\_life} \geq 6 \\ 0, & \text{otherwise} \end{cases}$$

we fit a logistic regression model with some independent/predictor variables.

Table 1: Categorical Variables

Income of Respondent	Future Children Intention	Born in Canada?
Less than $25,000	Don't know	Don't know
$25,000 to $49,999	No, definitely not	Born outside Canada
$50,000 to $74,999	Probably not	Born in Canada
$75,000 to $99,999	Unsure	
$100,000 to $ 124,999	Probably yes	
$125,000 and more	Definitely yes	

The other predictor variables are `age_at_first_birth` and `age`.

Using the GSS data, we replicated the approach used in the original survey. A single-stage stratified sampling approach by applying finite population correction to the sample was employed, adjusting each observation by the corresponding provincial population to reduce the variation. Then, we fitted a logistic model to the survey design, yielding the following model:

$$\begin{aligned}\log\left(\frac{h}{1-h}\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ &+ \underbrace{\beta_{3a}x_{3a} + \beta_{3b}x_{3b} + \beta_{3c}x_{3c} + \beta_{3d}x_{3d} + \beta_{3e}x_{3e}}_{\text{dummy coding for income}} \\ &+ \underbrace{\beta_{4a}x_{4a} + \beta_{4b}x_{4b}}_{\text{dummy coding for place birth Canada}} \\ &+ \underbrace{\beta_{5a}x_{5a} + \beta_{5b}x_{5b} + \beta_{5c}x_{5c} + \beta_{5d}x_{5d} + \beta_{5e}x_{5e}}_{\text{dummy coding for future children intention}}\end{aligned}$$

The functional form of the logistic model gives the logarithm of the odds of the outcome variable – in this case, the binary variable `happy` (h). For each of the categorical variables in Table 1, we use dummy variable coding with the variable at the top representing the baseline, in order to be able to assess what effect, if any, moving to a category would have compared to the baseline.

Results

The coefficients fitted using the logistic regression model are given in Table 2. The *OR* column gives the odds ratio, *Beta (SE)* gives the logarithm of the odds ratio and corresponding standard error of the estimate. *P* gives the p-value for the significance test which evaluates the probability of encountering data as or more significant than the Wald test statistic.

Table 2: Logistic Regression Model Summary

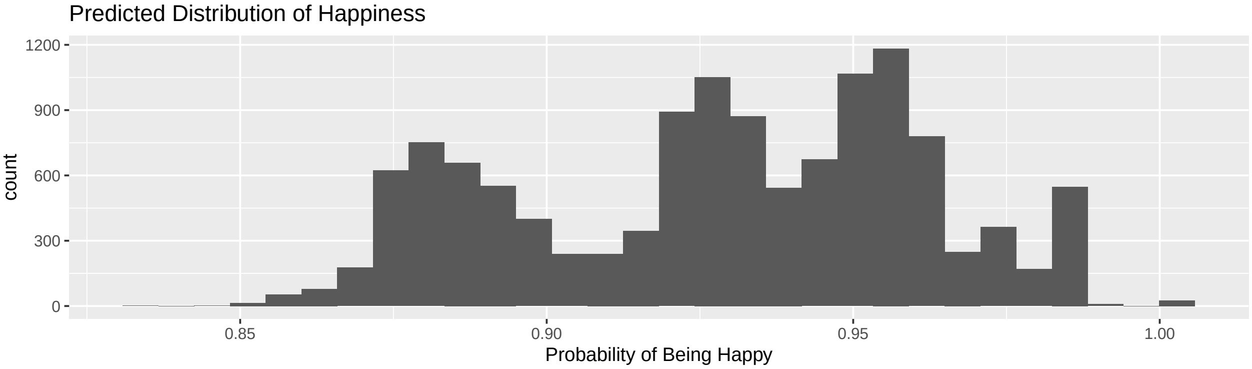
	Variable	OR	Beta (SE)	P
β_0	Intercept	-	2.12 (0.88)	0.02
β_1	age_at_first_birth	1.02	0.02 (0.01)	0.01
β_2	age	1.00	-0.00 (0.00)	0.35
income cat. var	as.factor(income_respondent)			
income baseline	Less than $25,000 (ref)	-	-	-
β_{3a}	$25,000 to $49,999	1.62	0.48 (0.09)	<0.001
β_{3b}	$50,000 to $74,999	2.37	0.86 (0.12)	<0.001
β_{3c}	$75,000 to $99,999	2.82	1.04 (0.17)	<0.001
β_{3d}	$100,000 to $ 124,999	4.39	1.48 (0.31)	<0.001
β_{3e}	$125,000 and more	7.65	2.03 (0.35)	<0.001
birthplace cat. var	as.factor(place_birth_canada)			
birthplace baseline	Don't know (ref)	-	-	-
β_{4a}	Born outside Canada	0.61	-0.50 (0.82)	0.55
β_{4b}	Born in Canada	0.69	-0.37 (0.82)	0.65
children int. cat. var	as.factor(future_children_intention)			
children int. baseline	Don't know (ref)	-	-	-
β_{5a}	No, definitely not	0.95	-0.05 (0.13)	0.67
β_{5b}	Probably not	0.95	-0.05 (0.24)	0.84
β_{5c}	Unsure	0.75	-0.28 (0.30)	0.34
β_{5d}	Probably yes	1.20	0.18 (0.45)	0.69
β_{5e}	Definitely yes	188128.04	12.14 (0.25)	<0.001

The model has the following interpretations:

- The intercept term has no meaningful interpretation and it is statistically significant with a very small p-value.

- The continuous variables `age_at_first_birth`, `age` very slightly affect the model with odds-ratios slightly off from 1.
- The categorical variables `place_birth_canada` and `future_children_intention` make a large difference to the odds ratio with odds-ratios near 0, but this is probably due to a flaw in the survey design which we will discuss in the Weaknesses section of this report.
- `income_respondent` is a meaningful variable in this model. For all six levels, the odds ratio is greater than 1 which implies that each increase in the respondent's income results in an increase in the log odds of being happy compared to the baseline.

Using a variety of parameters with this model, we obtain the following histogram that shows the class distribution of happiness.



Discussion

Weaknesses

Next Steps

References

Bibliography:

1. General Social Survey: An Overview, 2019. (2019, February 20). Retrieved October 12, 2020, from Statistics Canada, Canada website: https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm

2. General Social Survey - Family (GSS). (2019, February 7). Retrieved October 12, 2020, from Statistics Canada, Canada website: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501

3. Population and Dwelling Count Highlight Tables, 2016 Census. (2019, February 20). Retrieved October 17, 2020, from Statistics Canada, Canada website: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=101&S=50&O=A

4. Wu, Changbao, and Mary E. Thompson. "Basic Concepts in Survey Sampling." Sampling Theory and Practice. Springer, Cham, 2020. 3-160.