# Major factors influencing the physical and mental health of Canadians[*]

Yuxuan Yang

26 April 2022

**Abstract**

The general wellbeing of Canadians, in terms of both physical and mental health, can greatly reflect the development level of Canada as a country, as well as serving as accurate measures of living conditions of Canadians; yet wellbeing of Canadians differ to a large extent between groups of varying demographic characteristics. In this paper, we used data from "General social survey on Canadians at Work and Home (cycle 30), 2016" to analyze potential factors affecting the wellbeing of Canadians. We found that smoking, drinking, and eating behaviors could all affect people's health level, while sex, income, and job satisfactions as well as other factors are related to people's mental stress. As factors influencing Canadian wellbeing are becoming clear, we hope the Canadian government could pay attention to and improve Canadians' living qualities, especially for minority groups.

## 1 Introduction

Improving the Canadian wellbeing and living standards has always been a main goal of the Canadian government. As stated by Statistics Canada, measures such as the Canadian Index of Well-being (CIW) were designed and put into use in as early as the 2000s in order to provide more accurate measures of Canadian people's wellbeing, from aspects including health conditions, economic status, and social status (Sanmartin et al., 2021). In the meantime, with the fast-paced development of technology and medical benefits, Canadians are experiencing longer lives and more healthy lifestyles. Nevertheless, contrary to the longer lifespan, the self-rated happiness level of Canadians is deteriorating (2017). In this paper, our goal is to examine which factors would affect Canadian wellbeing and give suggestions to the government to improve both physical and mental health conditions of Canadians, especially those who belong to minority groups.

Factors such as age, gender, and income are all related to happiness of Canadians; in specific, more than 60% of Canadians with household incomes greater than 80,000 dollars per year reported to be in excellent health conditions, while less than 50% of Canadians with household incomes less than 40,000 dollars per year reported to be in good health. There are also less common variables that have not been widely examined before, including smoking, drinking, and eating behaviors of Canadian people, as well as frequency of doing sports or exercises.

A survey was created in 2016 to measure both the physical and mental health of Canadians in 10 provinces in Canada aged 15 and above. To facilitate sampling procedure, each of the 10 provinces were divided into strata, and each of the respondent was reached via telephone. Online survey and telephone survey are also used in order to decrease the non-response rate. From the survey responses, age, sex and income are all related to people's mental health; smoking, drinking, and eating habits are related to physical health. Particularly, people with unhealthy life styles and lower incomes experience worse levels of wellbeing.

The paper is structured as follows: first, we talk about the survey methodology and sample and population frames. Second, we manipulated collected survey data to analyze potential factors influencing general wellbeing of Canadians. Finally, we give some advice to the government on how to improve the physical and mental health of Canadian people.

[*] Code and data are available at: https://github.com/Yuxuan-Yang-Maggie/Canadian-Wellbeing.

# 2 Data

## 2.1 Data Collection

The data we used in this paper was retrieved from the Canadian General Social Survey (GSS). The Canadian GSS was designed to be "a series of independent, annual, cross-sectional surveys, each covering one topic in-depth" that collects survey responses from Canadian citizens and permanent residents (Government of Canada 2016). It serves a purpose of analyzing the social trends of Canadians' well-beings from varying aspects. In this paper, we will focus on the "General social survey on Canadians at Work and Home (cycle 30), 2016", which gathers information on various aspects of Canadians through phone calls and interviews, throughout August to December, 2016.

Note that we could have used other datasets since many people have conducted relevant researches on factors associated with people's well-beings. However, the Canadian GSS is a government-lead platform, so we trust in the integrity and authenticity of its information.

## 2.2 Data Processing

We conducted all data analysis using the R Programming Language (R Core Team 2020). First, we used the "readxl" (Wickham and Bryan 2022) package to read in the data in excel format. Then we used the "janitor" package (Firke 2021) and "tidyverse" package (Wickham et al. 2019) to perform high-level cleaning and manipulation on the data. In specific, we filtered out all observations with missing responses and NAs and created new variables recording the gender, health level and stress level in categorical formats. We stored the cleaned data as csv file in the inputs sub-folder of our project. We also used the "ggplot2" package (Wickham 2016) to plot graphs and images in Data and Results Sections, and used "stats" package (R Core Team 2022) and "jtools" package (Long 2020) to build models.

## 2.3 Survey Method

Stratified sampling method was used in the collection procedure of "General social survey on Canadians at Work and Home (cycle 30), 2016". In specific, the Canadian government divided the target population (residents) who live in 10 provinces in Canada (Newfoundland and Labrador, Prince Edward Island, Nova Scotia, New Brunswick, Quebec, Ontario, Manitoba, Saskatchewan, Alberta, and British Columbia) in to 27 strata by Metropolitan areas, and randomly sampled respondents from each strata to collect information. Note that regions including Yukon, Northwest Territories, and Nunavut are not included in the survey.

The data were gathered electronically via computer-assisted telephone interviews (CATI), and all the survey question responses were self-reported by participants. Specifically, the government first sent letters to randomly selected households in each strata to invite them to participate in this survey. Then one member from each household was randomly selected to fill in the specific questions in the questionnaire. There turned out to be 19,609 respondents in this survey.

### 2.3.1 Strengths

The most notable strength of this survey is that it used stratified sampling; dividing the target population into different subgroups and randomly selecting participants from each strata could significantly improve the representativeness of a survey compared to simple random sampling.

### 2.3.2 Weaknesses

Since the overall response rate was only approximately 50% after the government invited households to participate, there would be significant non-response bias associated with the results collected by the survey. In order to make up for this weakness, we filtered out all observations with missing values in the variables we are interested in. Nevertheless, this action of filtering out missing values would decrease the total number of usable observations drastically from 19,609 to no more than 2000, which led to another problem of a too small actual sample size in our further analysis. What's more, the complex nature of stratified sampling procedure

Table 1: Glimpse: Canadians and their well-being measured in various aspects in 2016

| caseid | smoke_status | drink_status | stress_level | health_level |
|---:|---:|---:|---:|---:|
| 5 | 3 | 7 | 2 | 10 |
| 8 | 3 | 4 | 3 | 7 |
| 10 | 3 | 7 | 3 | 8 |
| 20 | 3 | 5 | 1 | 10 |
| 21 | 3 | 7 | 2 | 8 |
| 36 | 3 | 6 | 3 | 7 |
| 41 | 3 | 6 | 2 | 10 |
| 59 | 3 | 6 | 3 | 10 |
| 70 | 3 | 5 | 4 | 10 |
| 74 | 1 | 6 | 2 | 8 |

Table 2: Number and proportion of respondents by self-rated health level

| health | count | proportion |
|:---|---:|---:|
| Not at all satisfied | 10 | 0.0063052 |
| level 1 | 7 | 0.0044136 |
| level 2 | 12 | 0.0075662 |
| level 3 | 25 | 0.0157629 |
| level 4 | 42 | 0.0264817 |
| level 5 | 134 | 0.0844893 |
| level 6 | 136 | 0.0857503 |
| level 7 | 299 | 0.1885246 |
| level 8 | 467 | 0.2944515 |
| level 9 | 229 | 0.1443884 |
| Completely satisfied | 225 | 0.1418663 |

would make it difficult to interpret analysis results, compared with simple random sampling directly from the entire population.

## 2.4 Data Characteristics

The original raw dataset we extracted from the CHASS data center contains 19,609 observations, yet the cleaned dataset we used in our analysis would only contain 1,586 observations of respondents (Canadian residents) and their demographic information (such as sex, age group, and province of residence), as well as information on their well-beings from different aspects. Using the cleaned dataset, we created Table 1 to take a glimpse and get a sense of what our actual data looks like. From Table 1, we observed the stress level, health level, and alcohol consumption status of 10 respondents. In specific, each row is an individual respondent, and each column represents a specific aspect associated with well-being of a person.

We also created Table 2 to demonstrate the number and proportion of respondents at each self-rated health level. From Table 2, we observed that the proportion of participants is highest for health level 7-10 (18.9%, 29.4%, 14.4%, and 14.2%), indicating that most people seem to be very satisfied with their health conditions in 2016 on a scale from 0 to 10 (0 is not at all satisfied with health condition, and 10 is completely satisfied). In addition, we created a bar plot to demonstrate the proportions of people by stress level.

Figure 1 plots proportion (y axis) against stress level (x-axis) and shows that almost half of participants reported to be a bit stressful in daily lives, and more than 20 percent of people reported to be not very stressful. Figure 1 shows that most Canadians experienced appropriate levels of stress in 2016.

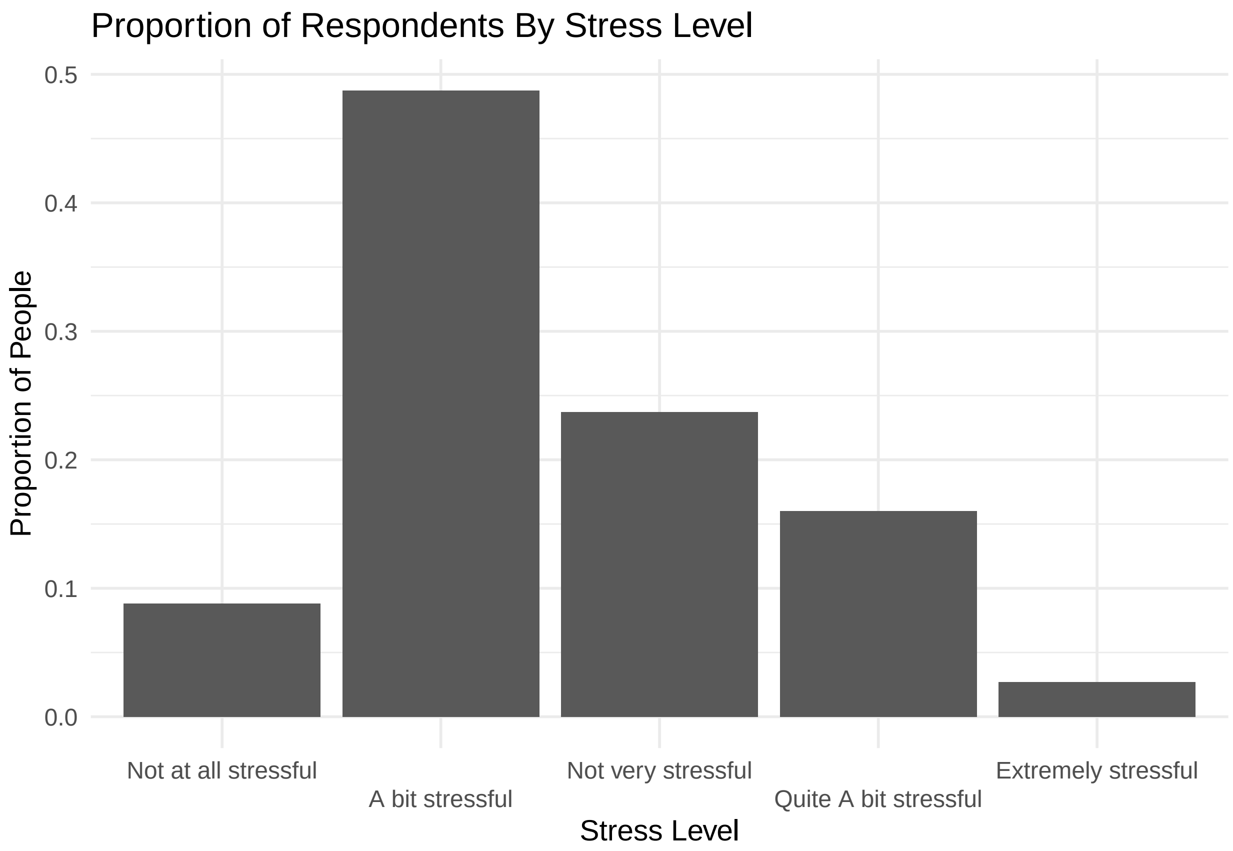

Figure 1: Proportions of Canadians Who Responded to the Work and Life General Social Survey in 2016, By Stress Level

# 3 Model

We will build one OLS (ordinary least squares) regression model to examine about the potential correlations between eating habits, drinking habits, job satisfaction, and living standards of people and their self-rated health conditions. The model expression is as follows:

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \beta_3 * X_{3i} + \beta_4 * X_{4i} \tag{1}$$

Where $Y_i$ is the expected health level (scaled from 0 to 10; 0 means not at all satisfied and 10 means completely satisfied) of an individual person, $X_{1i}$ is the eat habits (scaled from 1 to 5; 1 means excellent and 5 means poor), $X_{2i}$ is the frequency of alcohol consumption (scaled from 1 to 7; 1 means drink alcohol daily and 7 means never consumed alcohol), $X_{3i}$ is the job satisfaction level (scaled from 1 to 5; 1 means very satisfied and 5 means not at all satisfied), and $X_{4i}$ is the living standard of the person (scaled from 0 to 10; 0 means not at all satisfied and 10 means completely satisfied).

Additionally, $\beta_0$ is the expected intercept given everything else equals to 0 (expected health level when a person has excellent eat habit, drink daily, feels very satisfied with job but very dissatisfied about living conditions); $\beta_1$ is the estimated coefficient of $X_{1i}$ (expected change in health level when eat habit changes by 1 level, given all else constant); $\beta_2$ is the estimated coefficient of $X_{2i}$ (expected change in health level when alcohol consumption changes by 1 level, given all else constant); $\beta_3$ is the estimated coefficient of $X_{3i}$ (expected change in health level when job satisfaction changes by 1 level, given all else constant); $\beta_4$ is the estimated coefficient of $X_{4i}$ (expected change in health level when living condition changes by 1 level, given all else constant);

We will also build another linear regression model to examine about a few factors that could possibly influence people's mental stress conditions. The model is as follows:

$$y_i = \alpha_0 + \alpha_1 * x_{1i} + \alpha_2 * x_{2i} + \alpha_3 * x_{3i} \tag{2}$$

Where $y_i$ is the expected stress level (scaled from 1 to 5; 1 means not at all stressful and 5 means extremely stressful) of an individual person, $x_{1i}$ is the job satisfaction level (scaled from 1 to 5; 1 means very satisfied and 5 means not at all satisfied), $x_{2i}$ is the sex (1 if male and 0 if female), and $x_{3i}$ is the income level (scaled from 1 to 6; 1 means less than 25,000 dollars per year and 5 means 125,000 or more dollars per year).

Additionally, $\alpha_0$ is the expected intercept given everything else equals to 0 (expected stress level when a person is female, very satisfied with job, and earns less than 25,000 dollars per year); $\alpha_1$ is the estimated coefficient of $x_{1i}$ (expected change in stress level when job satisfaction changes by 1 level, given all else constant); $\alpha_2$ is the estimated coefficient of $x_{2i}$ (expected change in stress level when a person is male, given all else constant); $\alpha_3$ is the estimated coefficient of $x_{3i}$ (expected change in stress level when income changes by 1 level, given all else constant).

# 4 Results

From Figure 2 where we generated a pie chart to show the percentages of respondents with different smoking behaviors, we observed that only around 10% of total respondents smoke daily or occasionally, while most respondents don't smoke at all. From Figure 3 where we also drew a pie chart to demonstrate the percentages of people at each level of alcohol consumption, we observed that 27% respondents didn't drink in the past month and 18% respondents never had the habit of drinking, while almost no respondents (0%) drink every day. Figure 2 and Figure 3 together shows that respondents in this survey seems to have healthy, sustainable drinking and smoking habits.

Then we drew a bar plot called Figure 4 to demonstrate the distribution of number of people at different health levels, with eating habits highlighted in different colors at each health level. Figure 4 shows that the mode of this distribution is more than 400 people at health level 8. Since 0 means not at all satisfied and 10 means completely satisfied on a scale from 0 to 10, we know that most people are quite satisfied with their

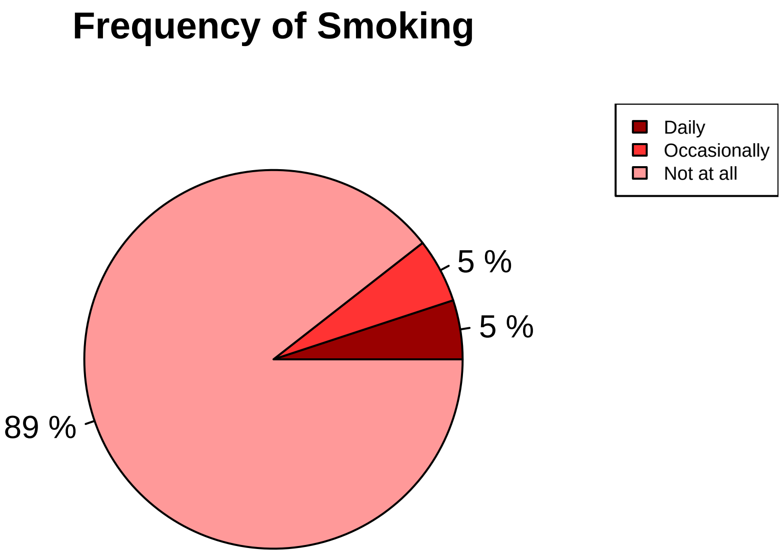

Figure 2: Smoking Frequency

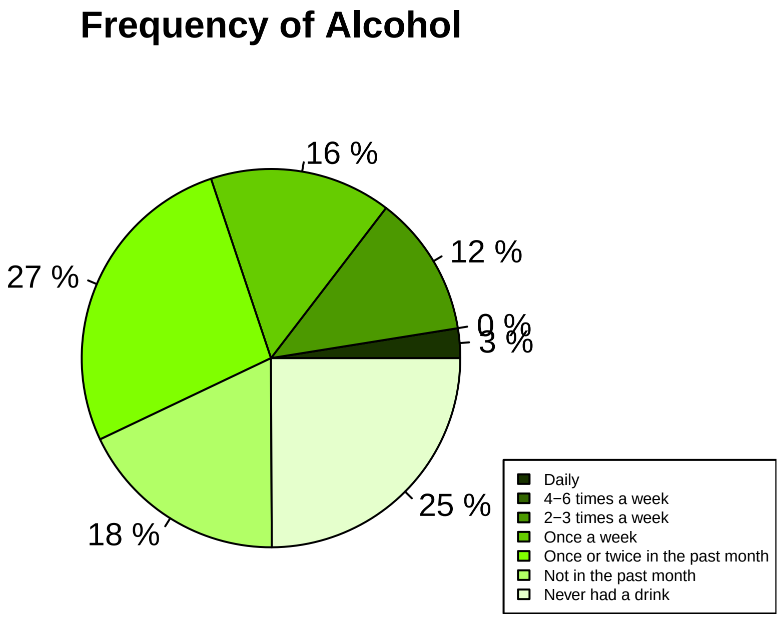

Figure 3: Alcohol Consumption Frequency

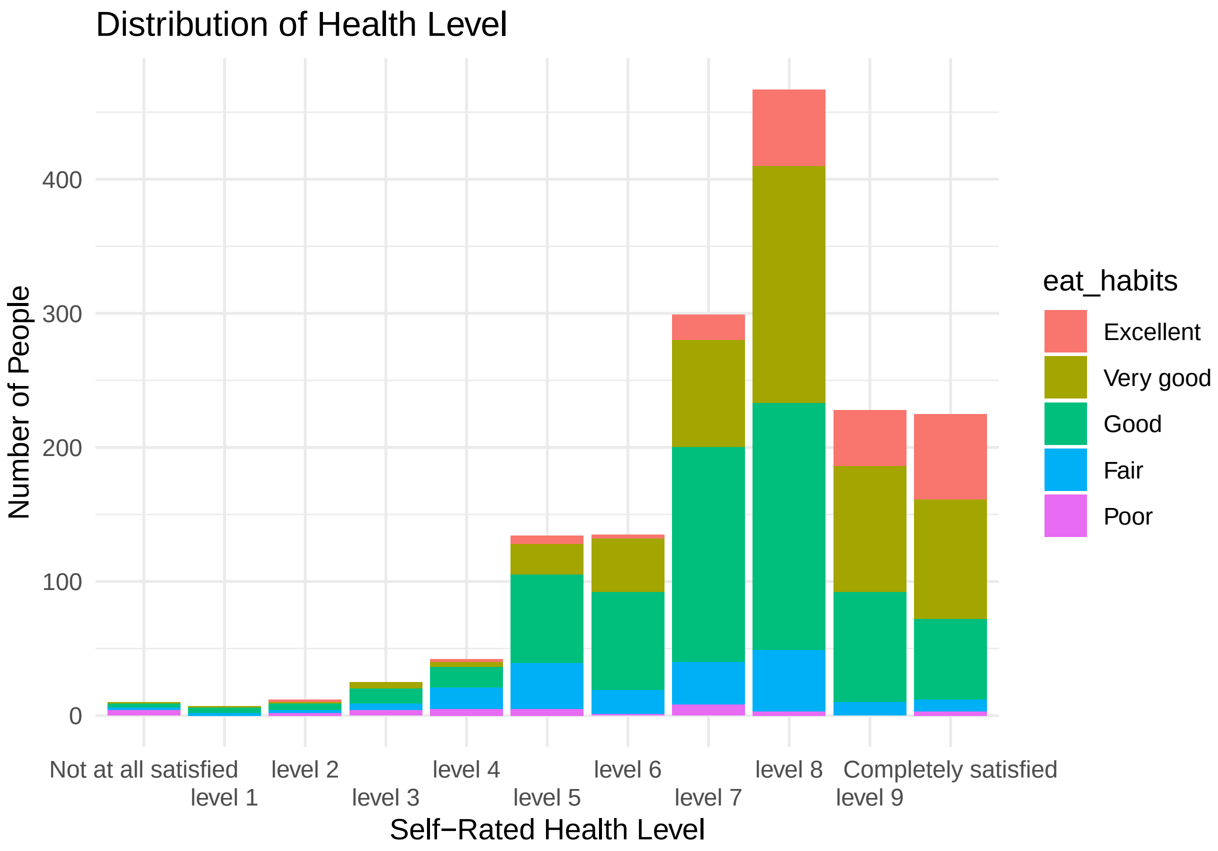

Figure 4: Number of People Under Each Self-Rated Health Level, with Eat Habits Highlighted in Colors

health conditions. What's more, people with higher health levels (better self-rated health conditions) tend to have better overall eating habits compared with those with lower health levels (worse health conditions).

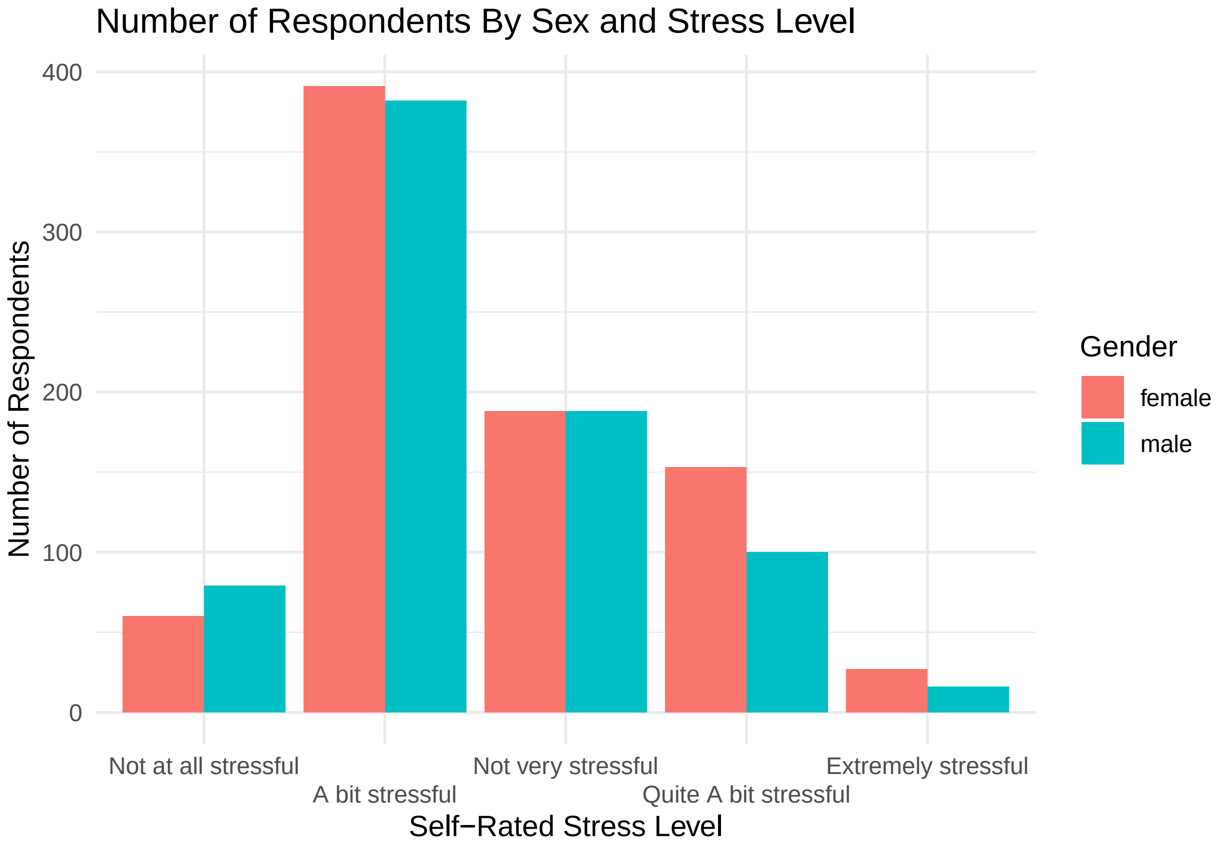

Figure 5: Number of People Under Each Self-Rated Stress Level, By Sex

We also drew Figure 5 to show the distribution of number of people at different stress levels. Figure 5 shows that for both males and females, the number of respondents is the largest (nearly 400) under the level of "a bit stressful" and the second largest (nearly 200) under the level "not very stressful". This indicates that most respondents are experiencing appropriate, moderate mental stress regardless of sex.

Then we created Figure 6 to also illustrate the distribution of number of people at different stress levels, but this time with number of people with different job satisfaction levels highlighted in colors at each stress level. Same from Figure 5, we observed from Figure 6 that most people experience normal stress conditions. Additionally, we observed that people with lower levels of mental stress tend to be more satisfied with their jobs compared with those with high levels of mental stress, indicating that job satisfaction could be correlated with people's mental health conditions.

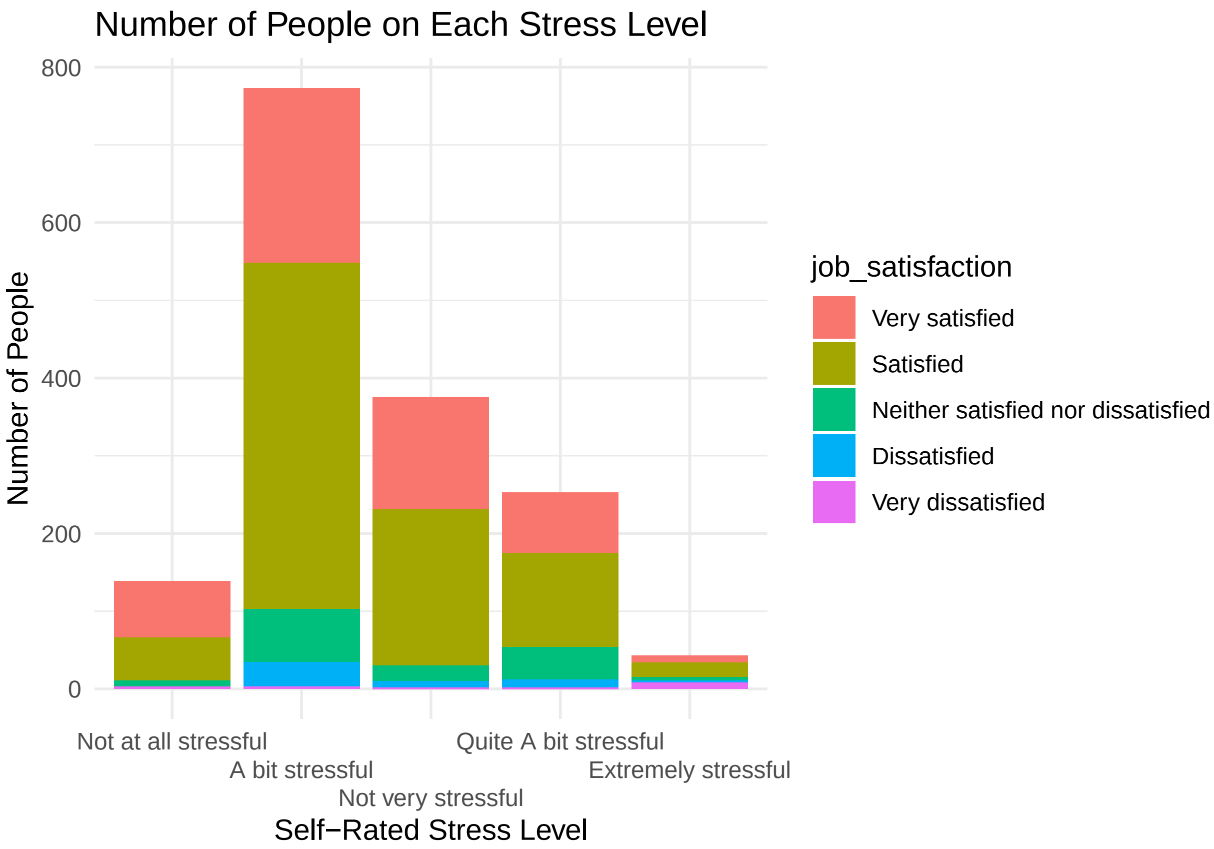

Figure 6: Number of People Under Each Self-Rated Stress Level, with Job Satisfaction Highlighted

| | |
|---|---|
| Observations | 1584 |
| Dependent variable | health_level |
| Type | OLS linear regression |

| | |
|---|---|
| F(4,1579) | 168.80 |
| $R^2$ | 0.30 |
| Adj. $R^2$ | 0.30 |

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 5.99 | 0.28 | 21.40 | 0.00 |
| eat_habits | -0.62 | 0.04 | -14.33 | 0.00 |
| drink_status | 0.07 | 0.02 | 2.77 | 0.01 |
| job_satisfaction | -0.09 | 0.05 | -1.68 | 0.09 |
| living_standard | 0.40 | 0.02 | 17.87 | 0.00 |

Standard errors: OLS

| | |
|---|---|
| Observations | 1584 |
| Dependent variable | stress_level |
| Type | OLS linear regression |

| | |
|---|---|
| F(3,1580) | 27.49 |
| $R^2$ | 0.05 |
| Adj. $R^2$ | 0.05 |

| | Est. | S.E. | t val. | p |
|---|---|---|---|---|
| (Intercept) | 2.32 | 0.08 | 30.24 | 0.00 |
| job_satisfaction | 0.22 | 0.03 | 7.94 | 0.00 |
| Gendermale | -0.19 | 0.05 | -4.23 | 0.00 |
| income | 0.06 | 0.02 | 3.67 | 0.00 |

Standard errors: OLS

Lastly, we created two regression tables to show the results of performing OLS (ordinary least squares) regression analyses on health level and stress level. The first regression table shows that with health level as the dependent variable, the p-values of eating habits, alcohol drinking behaviors, and living standards are all significantly smaller than 0.05, while the p-value of job satisfaction is not very small. This indicates that given all else equal, eat habits, alcohol consumption, and living standards seem to be highly correlated with people's self-rated health conditions. Since the estimate for eat habits is -0.62, given a scale of 1 to 5 (higher score means worse eat habits), the worse the eat habits, the worse the expected health condition of people would be. Similarly, better alcohol consumption habits and living standards correlate with better health conditions.

The second regression summary table shows that with stress level as the dependent variable, the p-values of job satisfaction, gender male, and income are all significantly smaller than 0.05, indicating that income, gender, and job satisfaction all seem to be correlated with people's mental stress. Specifically, since higher scores mean higher stress levels given a scale from 1-5 (1 mean not at all stressful and 5 means extremely stressful), and higher scores for job satisfactions mean less satisfaction (1 means very satisfied and 5 means very dissatisfied), the estimate of 0.22 for job satisfaction indicates as people become less satisfied with their jobs, their mental stress would increase. Similarly, males tend to be less stressful than females in general, while income doesn't seem to influence people's stress level to a great extent.

# 5 Discussion

## 5.1 What was done in this paper

Canadians nowadays are facing great pressure from enormous sources, including but not limited to rising house prices, rising inflations, and COVID pandemic. Excessive amounts of stress can lead to inappropriate pressure-relieving behaviors such as consuming too much alcohol or smoking that could negatively affect people's physical health. In this paper, we aimed to examine about both the physical health and mental health of Canadians in order to provide suggestions and solutions on what to focus in the future to improve their mental stress and physical health conditions.

In this paper, we used the 2016 Canadian GSS (General Social Survey) and its "Canadians at Work and Home" aspect to investigate both the physical and mental health conditions of Canadian people, targeting at the total population of 10 provinces in Canada (as outlined in the Data Section, excluding regions of Yukon, Northwest Territories, and Nunavut). Specifically, we selected 22 variables from the original raw survey data and filtered out all the observations with missing values in the columns we are interested in. We then briefly looked at the data through a few tables and graphs, and conducted two OLS (ordinary least squares) regression models to examine about variables that could influence mental stress and physical health conditions of Canadians. Both the Canadian GSS survey and our report could provide meaningful insights to future researches on what affects and how to improve Canadian people's well-beings.

## 5.2 What we learned about the world

From the tables, graphs, and regression models we conducted in this paper, we learned that the overall smoking and alcohol consuming behaviors of Canadians are quite moderate and healthy in 2016, as suggested by the statistics that only 5% respondents to the survey smoked daily and almost nobody (nearly 0%) consumed alcohol daily. We also learned that among people with better health conditions, the proportion of them possessing good eat habits tend to be larger than those with worse health conditions. Given the overall good smoking, drinking, and eating habits of most respondents, we suggest that the Canadian government could release promotional videos and advertisements to help people learn healthy lifestyles and maintain good habits.

We also learned that females tend to be a bit more stressful than males in general, as suggested by the data in our filtered observations that more than 150 women felt "quite a bit stressful", yet almost no men felt "extremely stressful". In addition, we noticed that the proportion of Canadians feeling very satisfied with their jobs is larger among groups with lower stress levels, indicating that high satisfaction with jobs correlate with good mental health conditions. Therefore, we suggest that the government could pay more attention to plights of women in workplaces and families and introduce polices that could reduce gender inequalities in society.

Lastly, we learned that linear correlations do seem to exist between physical health and drinking, eating habits, and living standards of Canadian people; and between mental stress and sex, income, and job satisfaction. Healthy living styles, high income, and good living conditions all indicate better conditions for human bodies. Therefore, our advice is that the Canadian government could open more free counseling programs for Canadian residents to facilitate the improvement in national health conditions.

## 5.3 Weaknesses and how to proceed in the future

One of the most notable weaknesses of our analyses lies in the nature of the survey method of using telephone interviews to collect information on Canadian residents' demographic information and answers on survey questions. In short, since the survey randomly selected one person from each household, that person might not actually precisely represent the whole household's opinions, which could lead to significant bias in the information we collected. Besides the fact that non-response bias occurs widely in the sampling procedure, we also face the risk of sampling bias, since using CATI and telephone to collect answers automatically ignores people who didn't possess telephones or those who failed to fluently use telephones, such as elder people.

However, elder people are important to our society, and they are the group to be most likely suffering from mental stress due to the inability to catch up with the rapidly developing society.

The fact that we asked the respondents to self-report their stress levels and health conditions also indicate potential bias in the answers we got; people might not be fully aware of their true health conditions. What's more, our procedure of filtering out all observations with missing values in the columns we are interested in also caused our sample size to significantly decrease, due to the large size of "don't know" and "refused to answer" in the survey answers we got. All these different weaknesses could cause our analyses to be biased and under-representative of the target population.

In the future, in order to mitigate all the negative effects on the accuracy and representativeness of our survey and analyses brought by these weaknesses, we could send invitations to participate in the survey to more people instead of just one person from each selected household. We could also hire highly-skilled interviewers to involve in the process of conducting the CATI, and hopefully, the rate of non-response would decrease. For people who got ignored by telephone interviews, we could use random interviews on the street to make up for potential sampling bias. More importantly, we need to do further research on things that could affect people's health and stress levels, and raise social concern on the importance of maintaining a good national health condition. Only when Canadians become fully aware of their health conditions could the nation become better.

# Appendix

## Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to record respondents' answers to the survey questions of the 2016 Canadian GSS. It was created for the purpose of getting a sense of Canadians' well-beings. There is no obvious specific gap needed to be filled.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Yuxuan Yang created the dataset on behalf of University of Toronto.
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - There is no funding for this dataset.
4. *Any other comments?*
   - No.

## Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - The instances that comprise this dataset represent individual respondents who participated in the 2016 Canadian GSS and their answers to the survey questions. There are no multiple types of instances.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 19,609 observations and hundreds of columns in the original raw data; there are 1,508 observations and 28 columns in the filtered, transformed dataset we actually used.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - The dataset we used is a sample of a larger set of all observations collected in the 2016 Canadian GSS. We filtered out this sample dataset by removing all observations with missing values in variables we are interested in, so it might not be fully representative of the original larger set.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Both the raw data and the sample data contains information on survey responses of the 2016 Canadian GSS. Each instance consists of responses to different survey questions of an individual person, including but not limited to sex, province, job satisfaction, health level, etc.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Yes, there are demographic information such as sex and province of residence of each individual respondent.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - Yes, there are instances which have missing values in different columns, such as income (some people might not be willing to provide income information).
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - Yes. We know each instance represents a person from a different household, so no two people can be from the same household in the data.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

- No.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - No.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - It linked to the Canadian GSS website; they are guaranteed to exist over time, and the complete version is on the website. The restriction is that you need an UofT account to log into CHASS to extract the columns.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Yes. The dataset has sub-populations such as females and males, or people of different income levels. The distribution of these sub-populations is quite even.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - No.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - Yes, it provides information on respondents' province of residence and sex.
16. *Any other comments?*
    - No.

**Collection process**

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

   - The data was acquired from CHASS. The data is not directly observable, and there is no obvious error in data, except missing values in certain observations and columns.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

   - The original dataset was collected through CATI (computer-assisted telephone interview), and we extracted data set we used from this original set.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

- The sampling strategy is stratified sampling.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - Government employees who conducted the CATI were involved; respondents randomly selected from chosen households were also involved.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

   - The data was collected over the timeframe of September to December, 2016. Yes, the timeframe matched the creation timeframe of data associated with the instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - No.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - We obtained the data from the CHASS (data was provided by Canadian government).

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - Yes. The respondents knew that they were doing survey questions provided by the Canadian government and that the information they provided will be accessible to the public.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.* -No.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - The survey didn't mention whether respondents can revoke their consent in the future.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

12. *Any other comments?*

    - No.

**Preprocessing/cleaning/labeling**

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - Yes. We cleaned the data by removing all observations with missing values in columns we are interested in, and renamed some columns.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

- Yes. It is in the “inputs” subfolder of the GitHub repo we provided.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - We used R.
4. *Any other comments?*
    - No.

**Uses**

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - No.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - The dataset can also be used to analyze different aspects of Canadians’ well-beings, which could help the Canadian government enact better policies in the future.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No.
6. *Any other comments?*
    - No.

**Distribution**

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The dataset and report will be provided through GitHub. Code and data are available at: https://github.com/Yuxuan-Yang-Maggie/Canadian-Wellbeing
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset is distributed on GitHub and does not have a DOI.
3. *When will the dataset be distributed?*
    - Already in GitHub.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - No, but the data set is licensed under the MIT license.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - No.

**Maintenance**

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Yuxuan Yang.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - yuxuanmaggie.yang@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - No.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - No.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - No.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - They can get data from the GitHub link we provided; we will verify their distributions through crediting to their names.
8. *Any other comments?*
   - No.

# References

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Government of Canada, S. C. 2016. *The General Social Survey: An Overview*. https://www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2013001-eng.htm.

Long, Jacob A. 2020. *Jtools: Analysis and Presentation of Social Scientific Data*. https://cran.r-project.org/package=jtools.

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

———. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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

Wickham, Hadley, and Jennifer Bryan. 2022. *Readxl: Read Excel Files*.