Major factors influencing the physical and mental health of Canadians*

Yuxuan Yang

25 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 (cite Canadian GSS website here). 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.
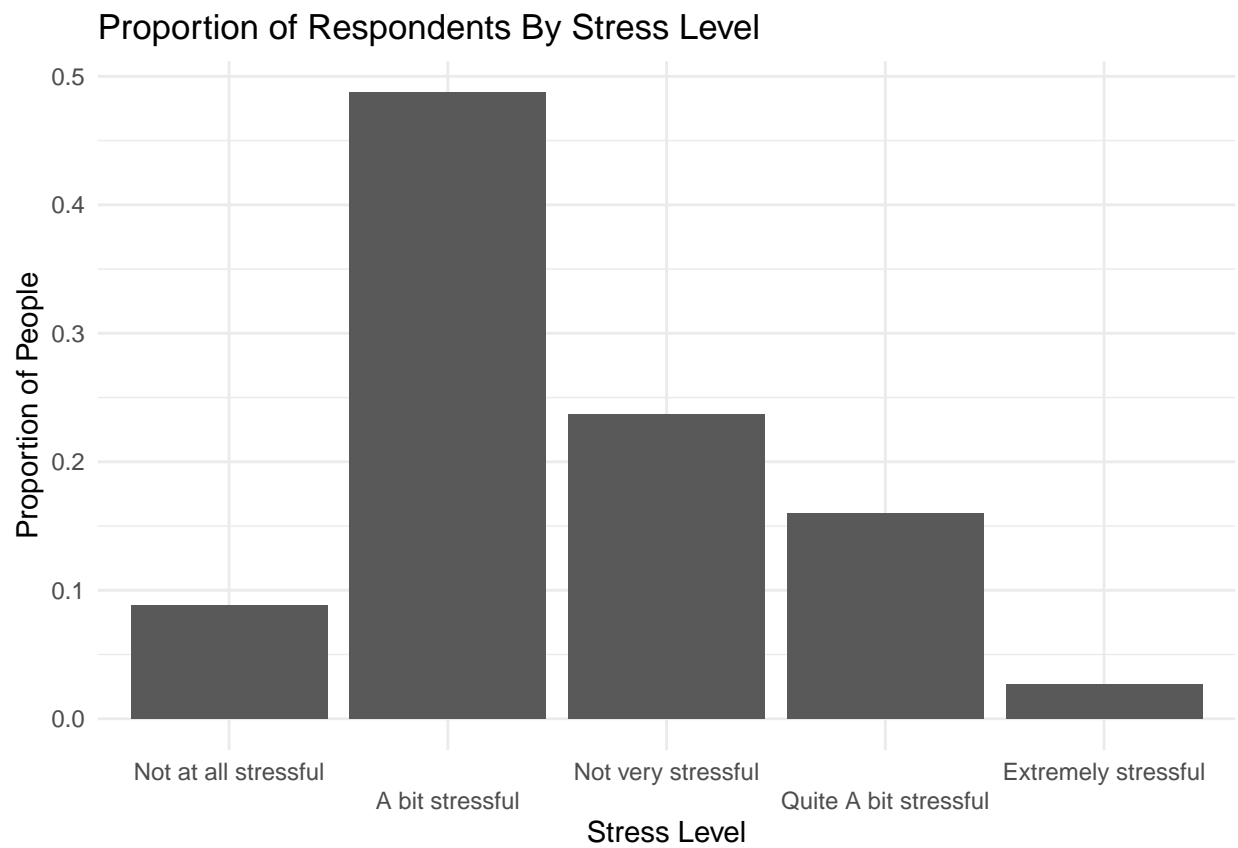
Figure 1: Proportions of Canadians Who Responded to the Work and Life General Social Survey in 2016, By Stress Level

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, β_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); β_1 is the estimated coefficient of X_{1i} (expected change in health level when eat habit changes by 1 level, given all else constant); β_2 is the estimated coefficient of X_{2i} (expected change in health level when alcohol consumption changes by 1 level, given all else constant); β_3 is the estimated coefficient of X_{3i} (expected change in health level when job satisfaction changes by 1 level, given all else constant); β_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, α_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); α_1 is the estimated coefficient of x_{1i} (expected change in stress level when job satisfaction changes by 1 level, given all else constant); α_2 is the estimated coefficient of x_{2i} (expected change in stress level when a person is male, given all else constant); α_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
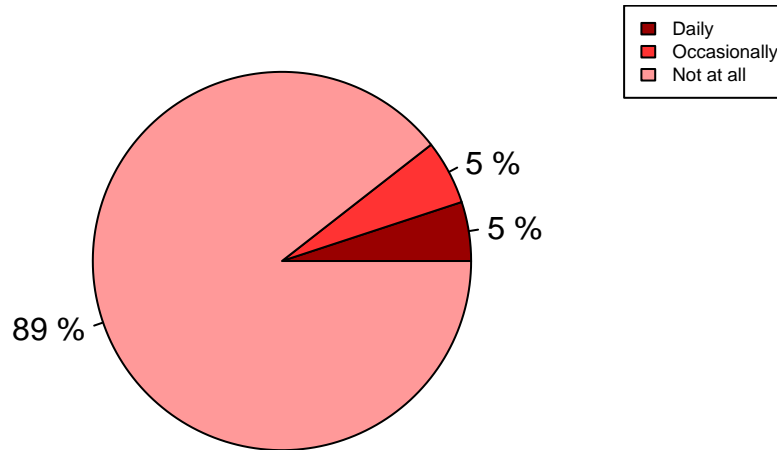
Frequency of Smoking

Legend:
- Daily
- Occasionally
- Not at all

5 %

5 %

89 %

Figure 2: Smoking Frequency

Frequency of Alcohol

16 %

12 %

27 %

0 %

3 %

25 %

18 %

Legend:
- Daily
- 4–6 times a week
- 2–3 times a week
- Once a week
- Once or twice in the past month
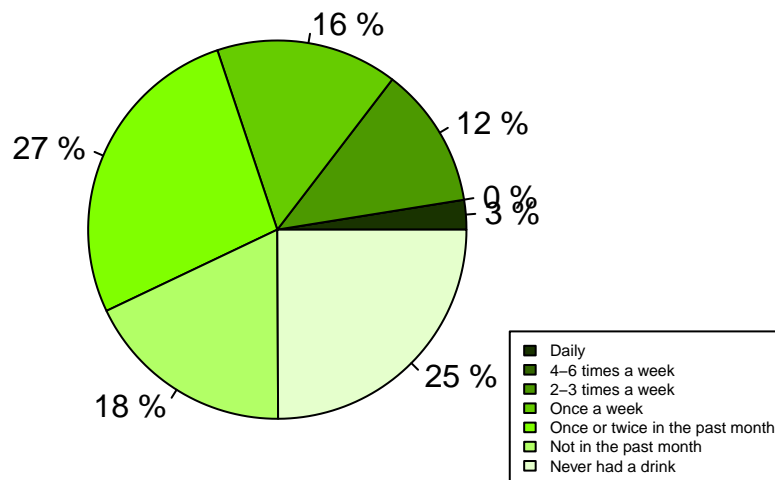- Not in the past month
- Never had a drink
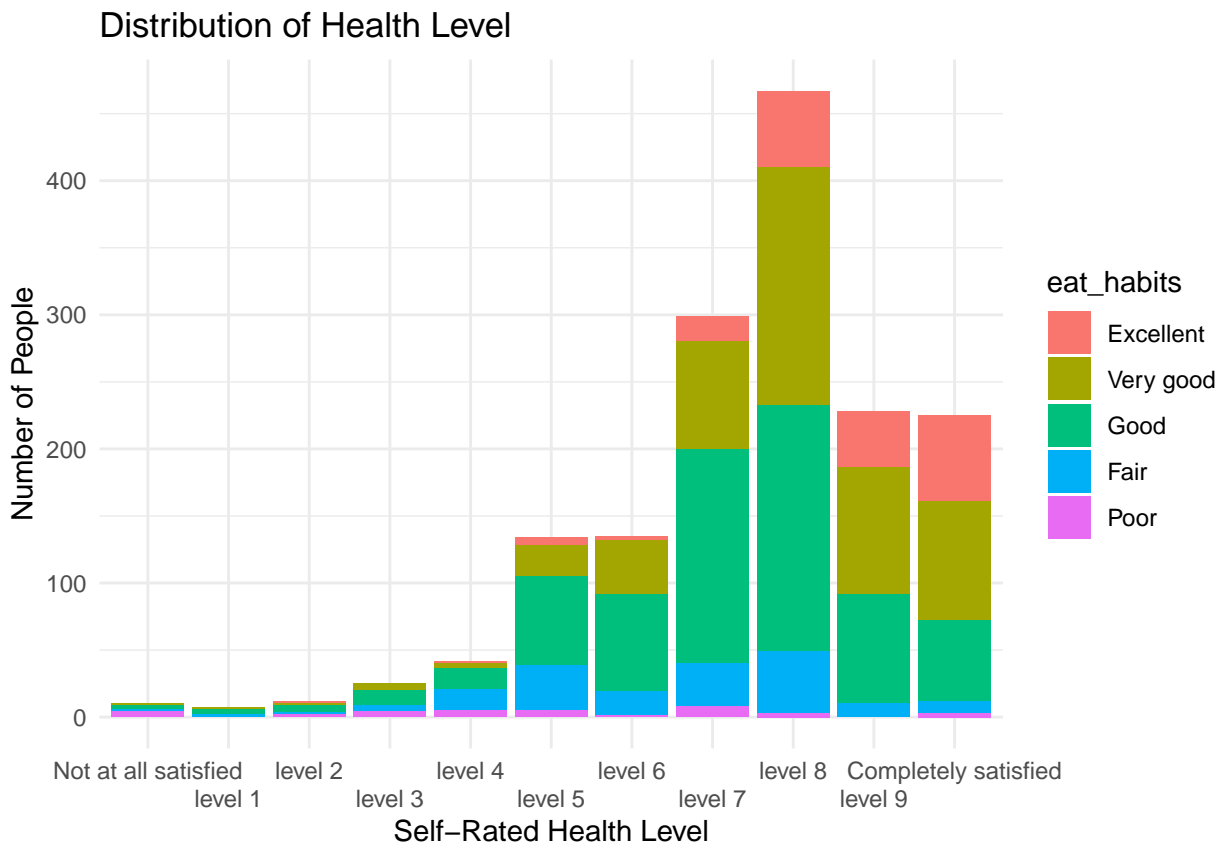
Figure 3: Alcohol Consumption Frequency

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).
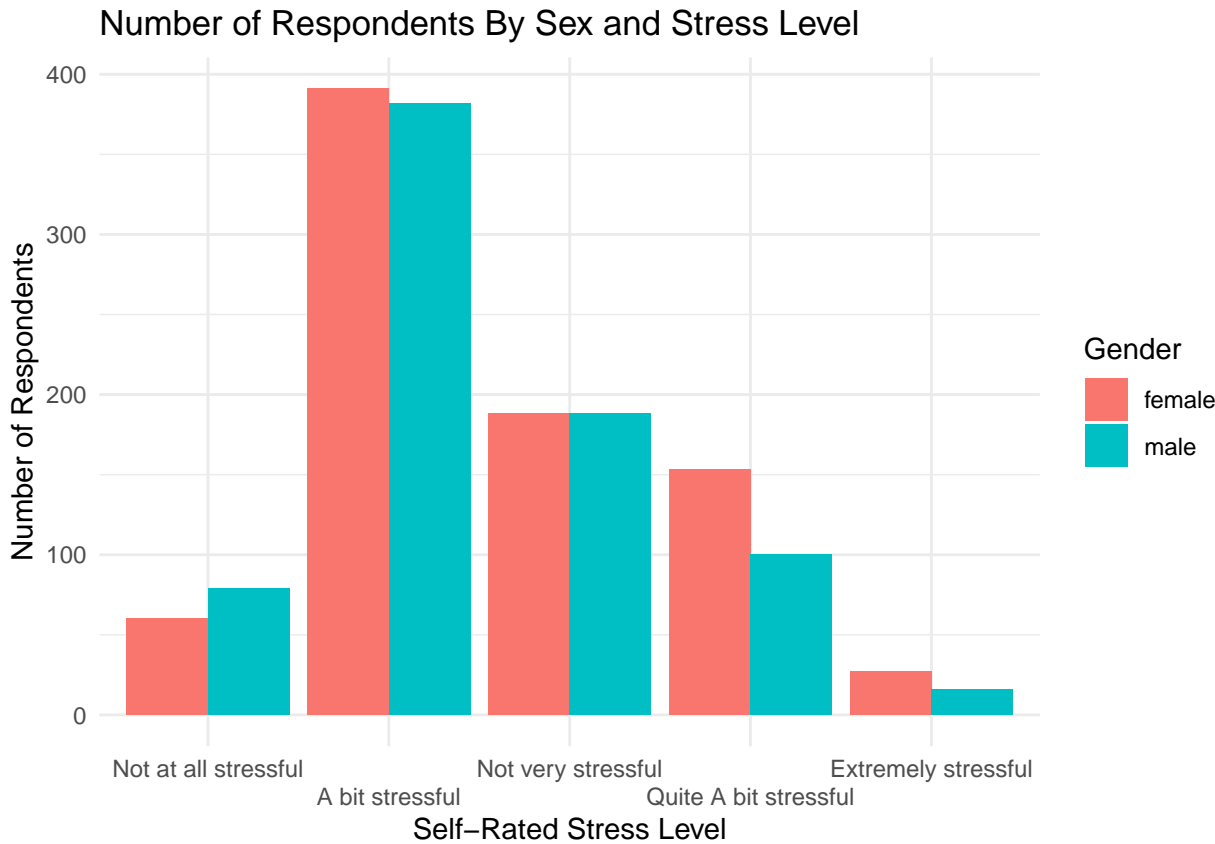

Figure 5: Number of People Under Each Self-Rated Stress Level, By Sex

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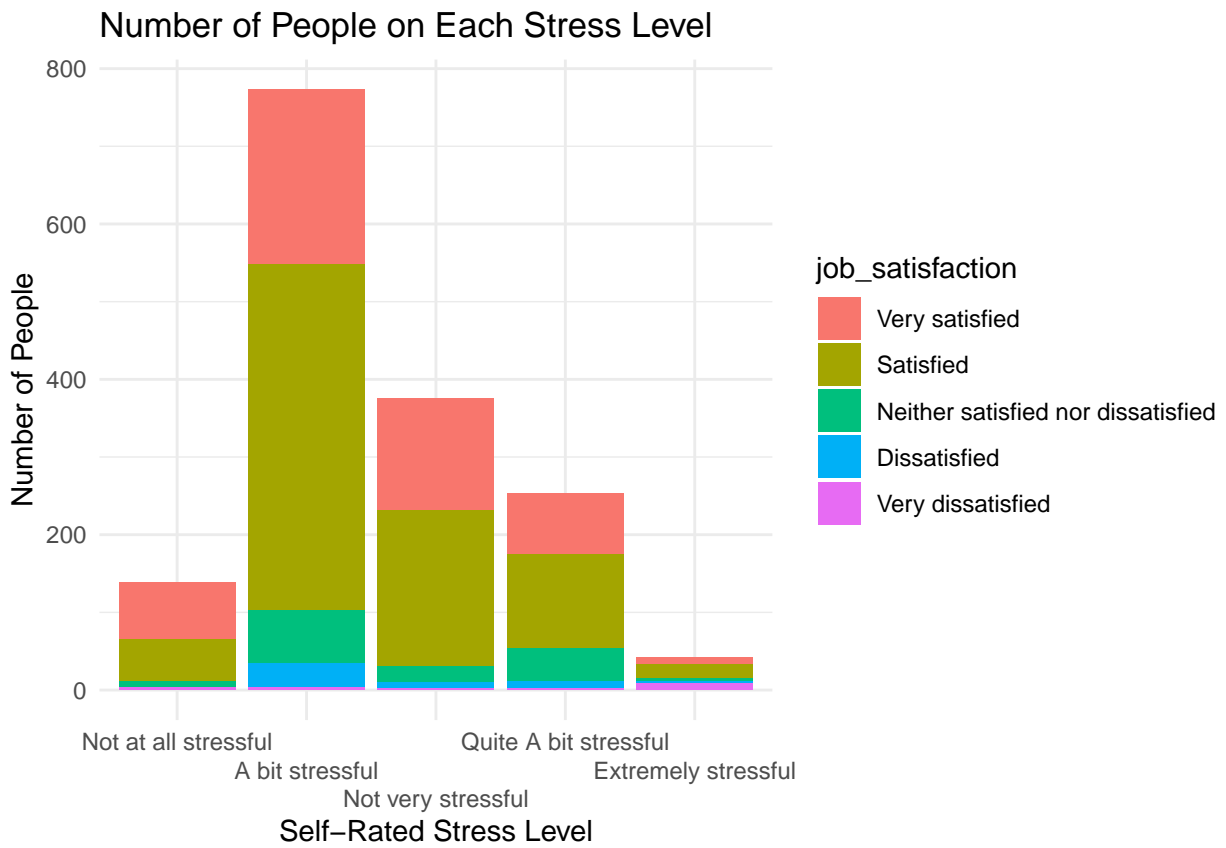
Figure 6: Number of People Under Each Self-Rated Stress Level, with Job Satisfaction Highlighted

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

Keywords: Extend the topic. Talk way more.

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what I know and what I learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional details

References

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

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.*