Analysis of the 2018 Canadian Community Health Survey with a focus on sleep*

Mental health and stress have significant impact on sleep

Pascal Lee Slew

26 April 2022

Abstract

Sleep has become a luxury in modern times. There are now so many things to do but time is limited. In this paper, we used the 2018 Canadian Community Health Survey to focus on the sleeping trends and the factors affecting them. We found that mental health and stress have significant effect on sleep. Those with better mental health have higher odds of sleeping well. Similarly, how much stress we feel in life is associated with sleep. The analysis of sleeping data is important to get an insight of the current situation and identify ways to help Canadians live a better life.

keywords: sleep, canadian community health survey, sleep deficiency, age, mental health, stress

1 Introduction

Sleep is important for both our physical and mental health. In today's world, there are so much things we have or want to do but we only have 24 hours a day. Often, people reduce their sleeping hours for other priorities with the belief that the benefits outweigh the costs. This is particularly the case for many university students who are willing to sacrifice their sleep to study or complete assignments. Failing to get enough sleep over time can lead to other health complications. Therefore, having an idea of the current sleep habits of Canadians is important to formulate policies and implement programs to improve the situation of Canadians.

The Canadian Community Health Survey (CCHS), a joint initiative of Health Canada, the Public Health Agency of Canada, Statistics Canada and Canadian Institute for Health Information (CIHI), provides information on a broad range of topics around Canadians health lifestyles since 2000. The CCHS has been set up with the objective of collecting health-related data on Canadians. With the collected data, analysis can be conducted to identify any worrying trends and relay crucial information to the stakeholders. It seeks to draw attention to emerging health issues.

In this paper, we seek to investigate factors affecting the number of hours slept with a particular focus on mental health. We will be looking at the demographics and other factors that have an effect on sleep through visualizations. Then, we used a logistic regression model to gain more insights on the data. We found that mental health and stress are significantly associated with sleep. Better mental health is associated with higher odds of having a good sleep. Similarly, less stress is associated with higher odds of sleeping well.

The remaining part of the paper is divided as follows: Section 2 provides information on data collection and an overview of the variables of interest. Section 3 presents the model used in this paper. Section 4 explains the results of our model. Section 5 expands on what is found and why it is important. Section A contains a datasheet for our dataset.

*Code and data are available at: https://github.com/Pascal-304/canadian-sleep-2018.

2 Data

2.1 Data Source and Collection

The data was obtained from the Canadian Community Health Survey 2017-2018 (CCHS) (Canada 2020b). The Canadian Community Health Survey is a cross-sectional survey conducted by Health Canada, the Public Health Agency of Canada, Statistics Canada and the Canadian Institute for Health Information on a national level. Its objectives are to collected health-related data and provide crucial information around Canadian health lifestyle.

The survey was conducted by interviewing individuals aged 12 and over in Canada's ten provinces and territories excluding residents of Indian Reserves and Crown Lands, institutional residents, full-time members of the Canadian Forces, youth living in foster homes and certain remote regions. The CCHS's coverage is about 98% of the Canadian population of age 12 and over. The data was collected 4 times over a period of 3 months. (Canada 2020a)

The data collected consists of three components: core, theme and optional content. Core content is made up of health-related questions asked of every respondents. Theme content are modules that are introduced to better understand certain areas while optional content are more catered to selected provinces or territories, so should not be generalized across Canada.

2.2 Survey Frame and Sampling Strategy

The survey frame is different for the adults and population under 18. For the adult population, the sampling frame used is the area frame used by the Canadian Labour Force (LFS). It is a two-stage sample design. First, a sample of primary sampling units (PSU) corresponding to geographical regions is selected. Then, for each PSU which is formed of 100 to 600 dwellings, a sample of dwellings is drawn.

For the youth population, a list frame is used to select respondents. The list frame was created from the Canadian Child Benefit (CCB) files which contained address and phone numbers used to conduct phone interviews.

The target sample size for the 2017-2018 CCHS was about 130,000 and the actual number of respondents was 113,290. Each province is split into health regions (HR) and each territory is taken as one HR. In 2018, there was over 100 HRs in the ten provinces. For each health region, a minimum of 500 respondents was required to ensure reasonable data quality. To prevent the sampling of too many dwellings in smaller HRs, sampling was restricted to one out of twenty dwellings. The response rate for the 2018 CHHS was 58.8%.

2.3 Methodology

Introductory letters and brochures were sent to selected households to explain the purpose of the CCHS and how the data collected would be used. The CCHS used computer-assisted interviewing applications to collect data. Two options were available: telephone interviews and personal interviews.

2.4 Key features

The raw data contains a total of 1,051 variables and 113,290 observations. However, there were many missing observations for the number of hours usually spent sleeping which is our main variable of interest. So, we removed the missing values and we were left with a dataset of 55,051 observations.

This report will focus on a subset of variables that will be used to analyze the relationships of the number of hours usually spent sleeping with different variables such as perceived mental health, life satisfaction, and work stress. Table 1 shows a subset of key variables that will be discussed in this paper.

Table 1: An overview of key variables of the dataset

Age	Marital Status	Sex	Sleep hours	Perceived Health	Perceived Mental Health
Age between 70 and 74	Widowed/Divorced/Separated	Male	6 hours to less than 7 hours	Poor	Good
Age between 50 and 54	Widowed/Divorced/Separated	Female	5 hours to less than 6 hours	Very good	Very good
Age between 30 and 34	Common-law	Female	7 hours to less than 8 hours	Excellent	Excellent
Age between 45 and 49	Common-law	Female	9 hours to less than 10 hours	Very good	Very good
Age between 75 and 79	Widowed/Divorced/Separated	Female	7 hours to less than 8 hours	Good	Good
Age between 40 and 44	Married	Female	7 hours to less than 8 hours	Very good	Very good
Age between 45 and 49	Married	Male	6 hours to less than 7 hours	Very good	Good
Age between 12 and 14	Single	Female	8 hours to less than 9 hours	Very good	Good
Age between 60 and 64	Widowed/Divorced/Separated	Female	7 hours to less than 8 hours	Fair	Excellent
Age between 35 and 39	Common-law	Male	7 hours to less than 8 hours	Very good	Very good

The dataset was processed and analyzed using R (R Core Team 2020) along with the r packages tidyverse (Wickham et al. 2019), janitor (Firke 2021), dplyr (Wickham et al. 2022), forcats (Wickham 2021) and kableExtra (Zhu 2021).

2.4.1 Age and Sex

Figure 1 shows the age distribution of the respondents. For all age groups, there seems to be a higher number of female adult respondents than male respondents. The converse is true for the youth population; there is higher male respondents than female. About 53.7% of respondents are female and 46.3% are male.

2.4.2 Marital Status

Figure 2 shows the marital status distribution of the respondents. Most respondents are married (), followed by single. A fairly high number of respondents are widowed, divorced or separated while being in a common-law relationship is the least preferred.

2.4.3 Number of sleeping hours

Figure 3 shows the number of hours usually spent sleeping. Most respondents tend to sleep between 7 to 8 hours, followed by 6 to 7 hours and 8 to 9 hours. A fairly high number of people sleep more than 9 hours on a regular basis. It is also worthwhile to note that X% usually sleep for less than 5 hours. This suggests that we should take a closer look at the reasons why it is the case.

2.4.4 Income

Figure 4 shows the income distribution of the CCHS respondents. Most respondents earned between 20,000 dollars to 39,999 dollars, followed by those in the less than $20,000 income group. It is worth to note that most selected respondents in the Health regions are not financially stable. A small number of respondents (700) had no income or income loss; they should consists of mostly of youth.

2.4.5 Perceived Health

Figure 5 illustrates the distribution of the number of hours spent sleeping for different perceived health. For those who perceived their health as “Excellent”, “Very Good” and “Good”, we observe that the distribution of the number of sleeping hours to be fairly similar with 7 hours to less than 8 hours as the mode. For those who perceived their health as “Fair”, the distribution of their sleeping hours seems to be close to uniform. We can see that a higher proportion of respondents answered to have lower sleeping hours. Similarly for those who perceived their health as “Poor”, the distribution is almost uniform with most people sleeping less

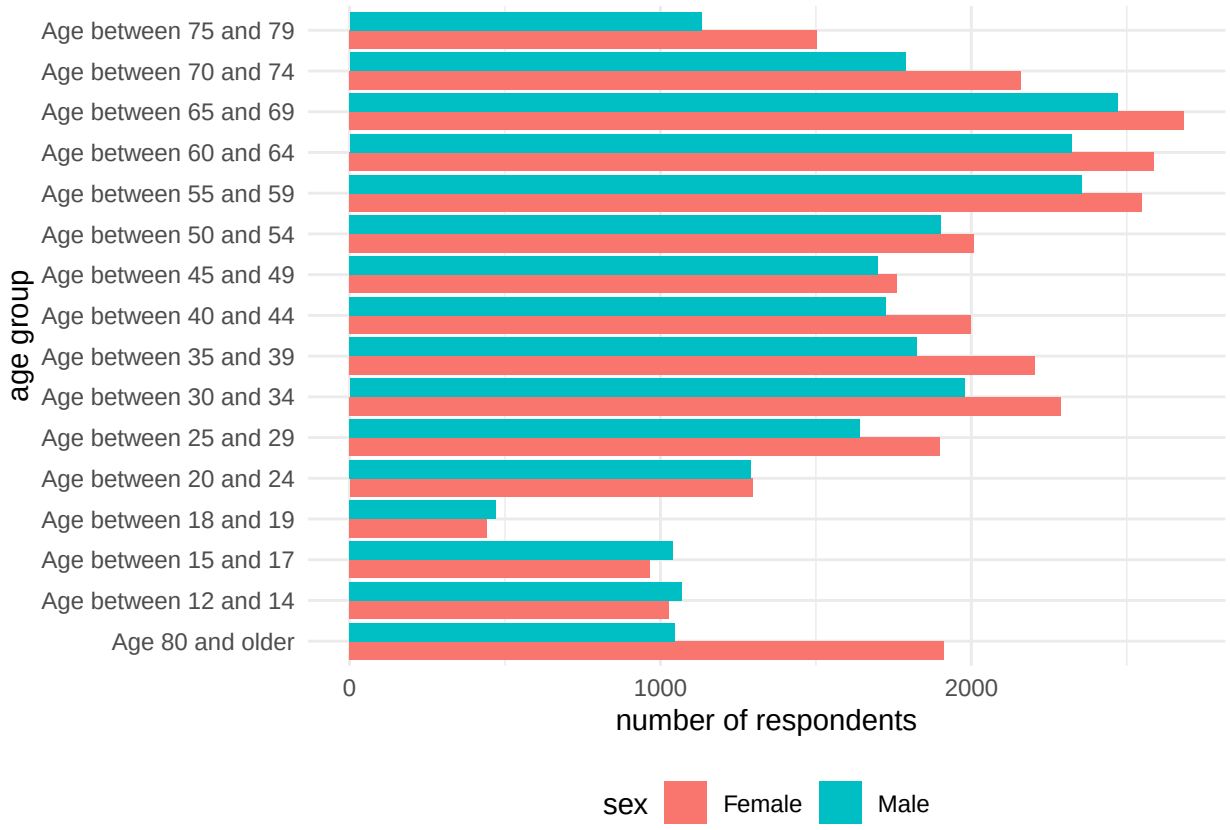

Figure 1: Distribution of survey respondents by age group and sex

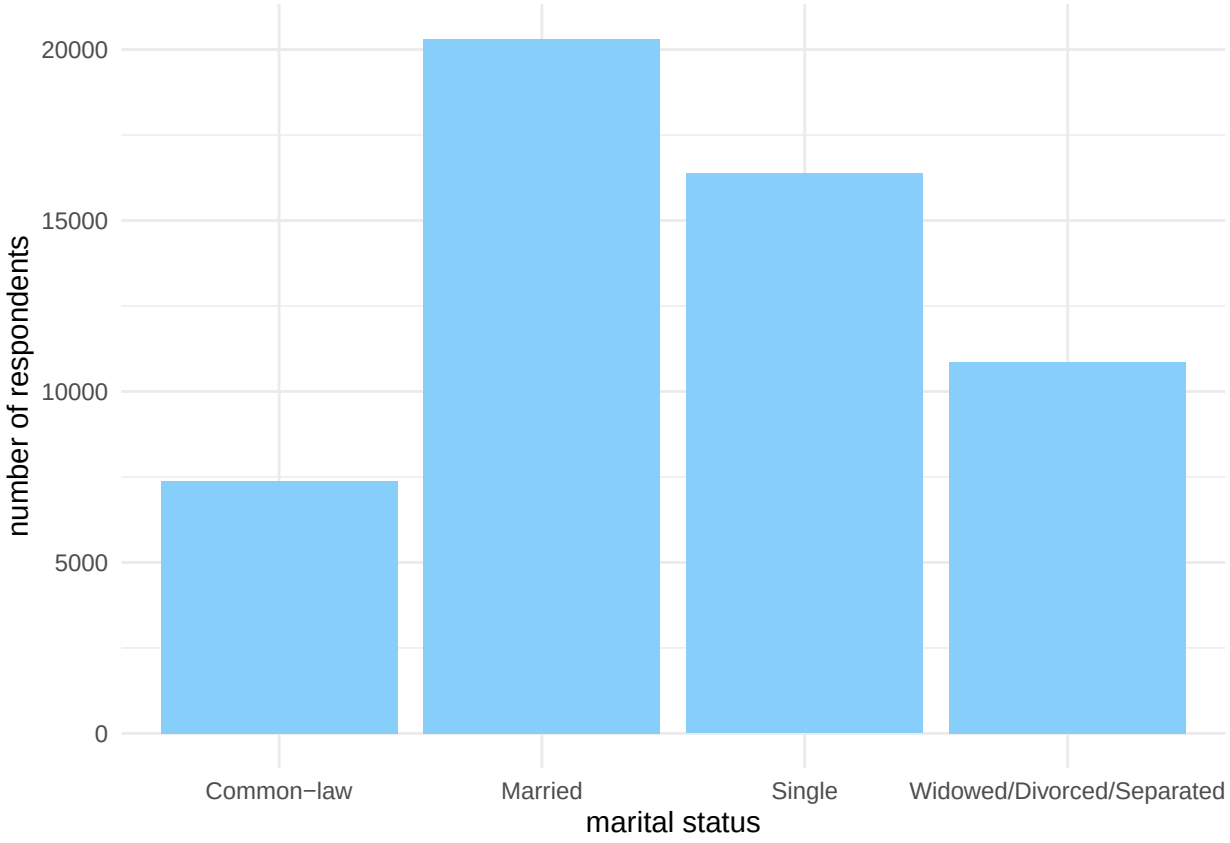

Figure 2: Marital status of respondents

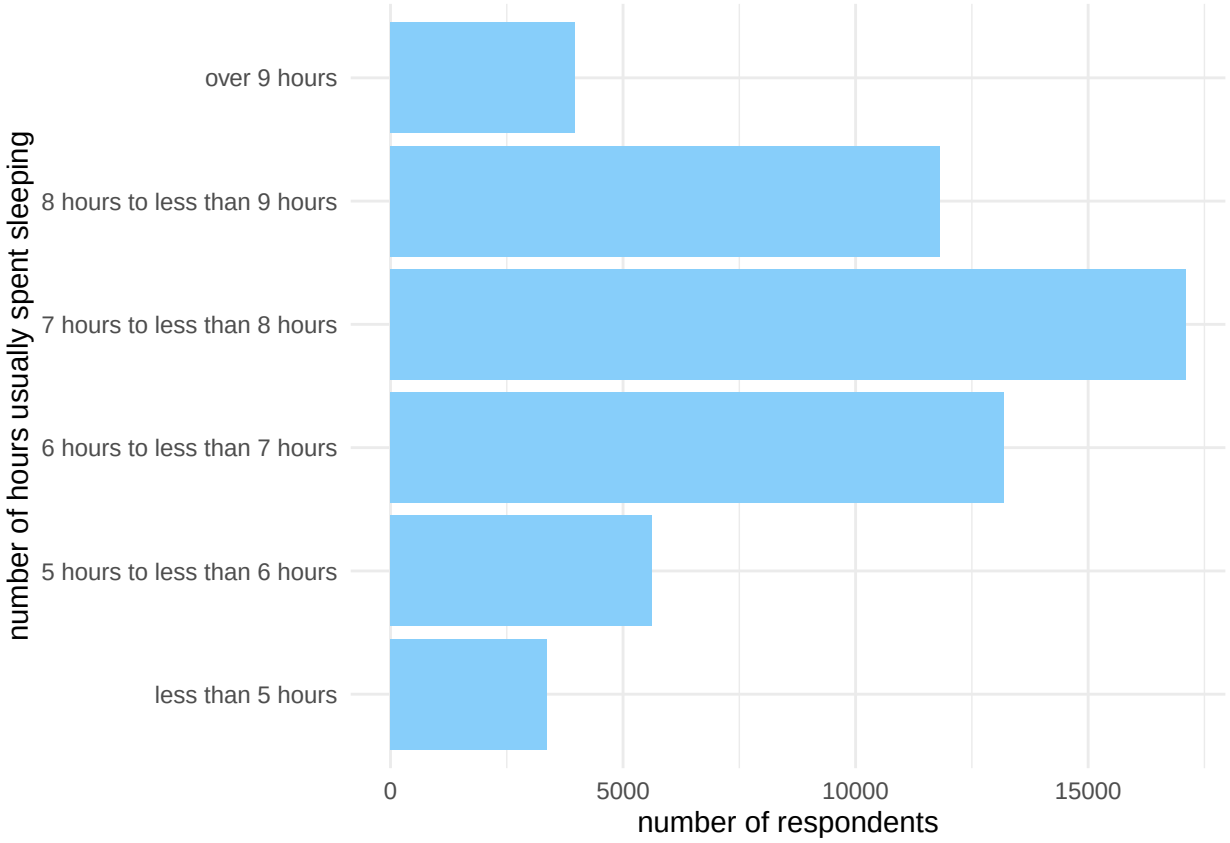

Figure 3: Distribution of the hours of sleep of respondents

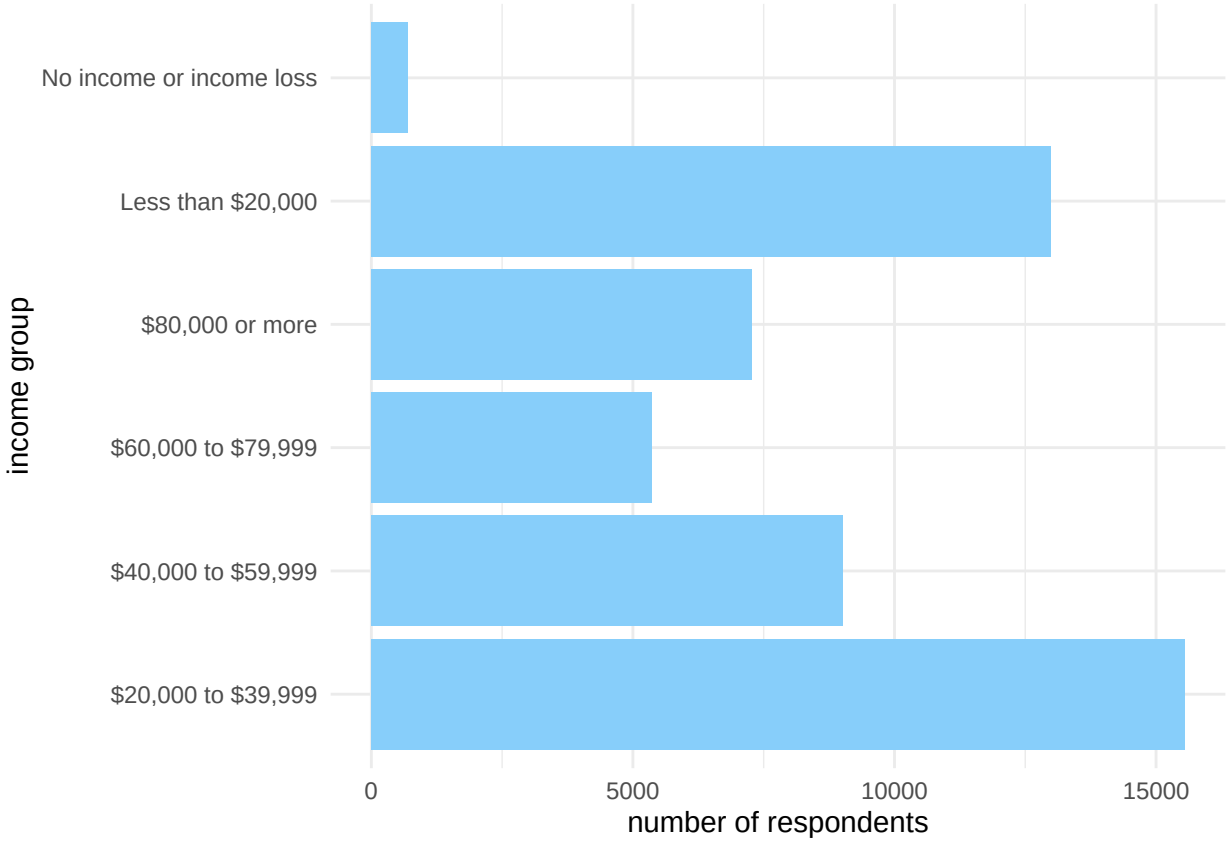

Figure 4: Income groups of respondents

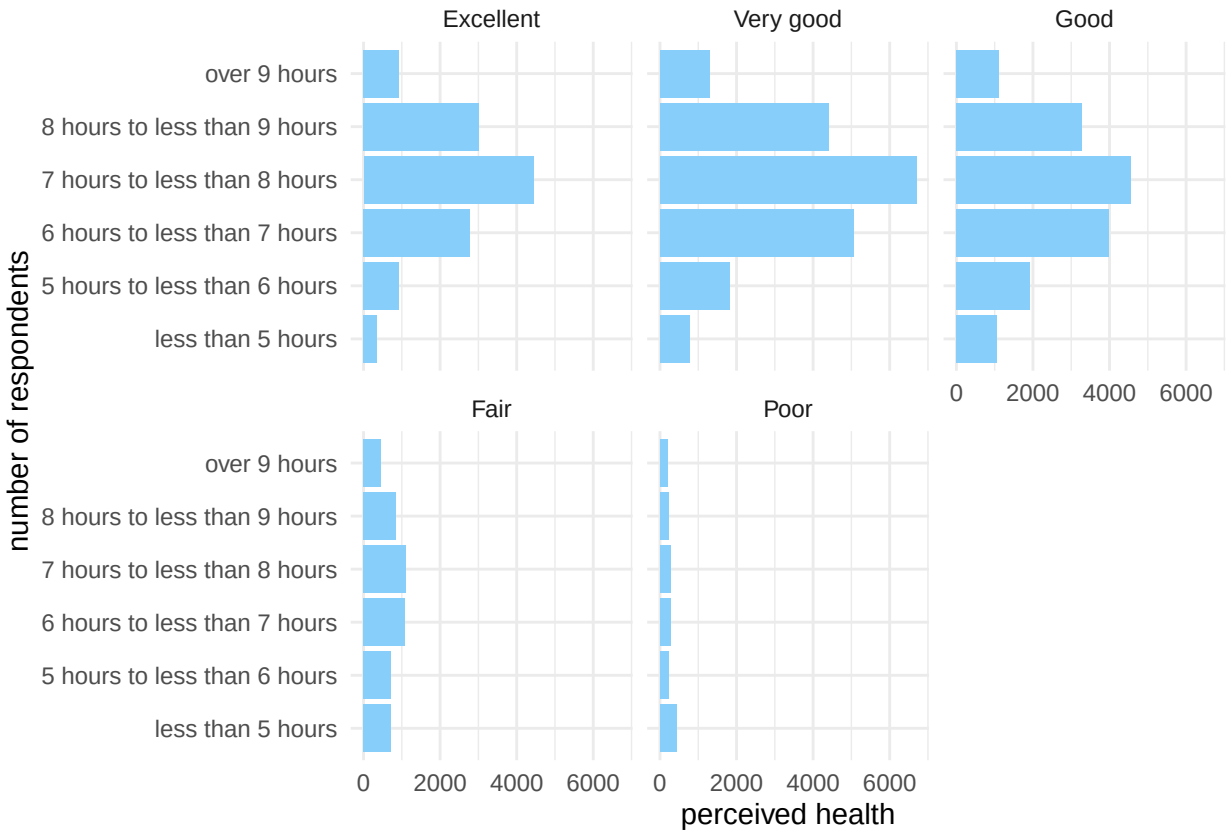

Figure 5: Comparison of the distribution of respondent based on their perceived health for sleeping hours

than 5 hours on a regular basis. This suggests that how people perceive their health may have an influence over the number of hours they usually sleep among other factors.

2.4.6 Perceived Mental Health

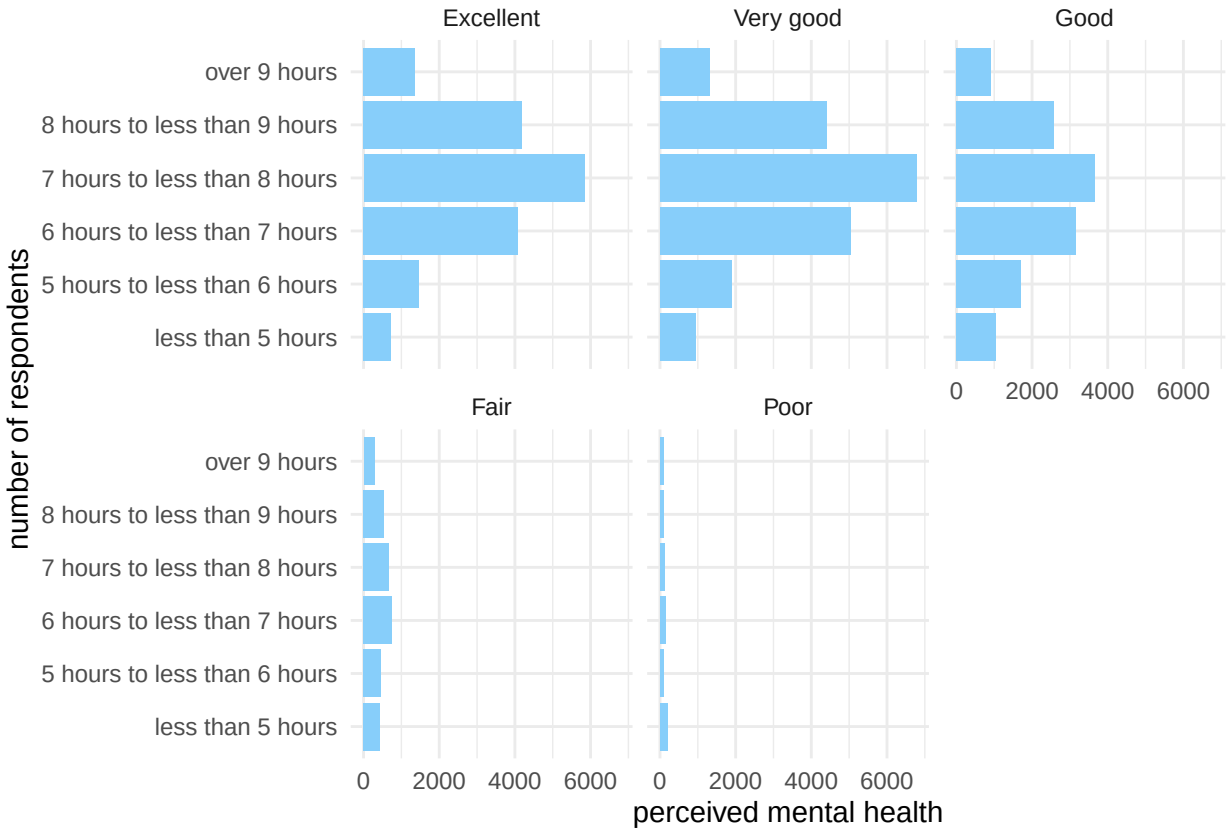

Figure 6: Comparison of the distribution of respondents' sleeping hours based on their perceived mental health

Figure 6 shows the distribution of the number of hours usually spent sleeping for different perceived mental health. For those who perceived their mental health as "Excellent", "Very Good" and "Good", the distribution of sleeping hours are similar. The higher people perceived their mental health, the lower the proportion of people sleeping too less. For "Fair" and "Poor", the proportion of respondents who reported lower sleeping hours is the highest. Again, the perception of mental health seems to have some effect on the hours spent sleeping.

2.4.7 Life Stress

Figure 7 shows how the number of hours usually spent sleeping is distributed for how the respondents perceived their stress in life. Except for those who feel that life is extremely stressful, the distributions of the number of hours spent sleeping are similar. The less stress people feel, the higher the concentration of people sleeping adequately. The proportion of people sleeping 8 to 9 hours is highest for those who feel life is not at all stressful.

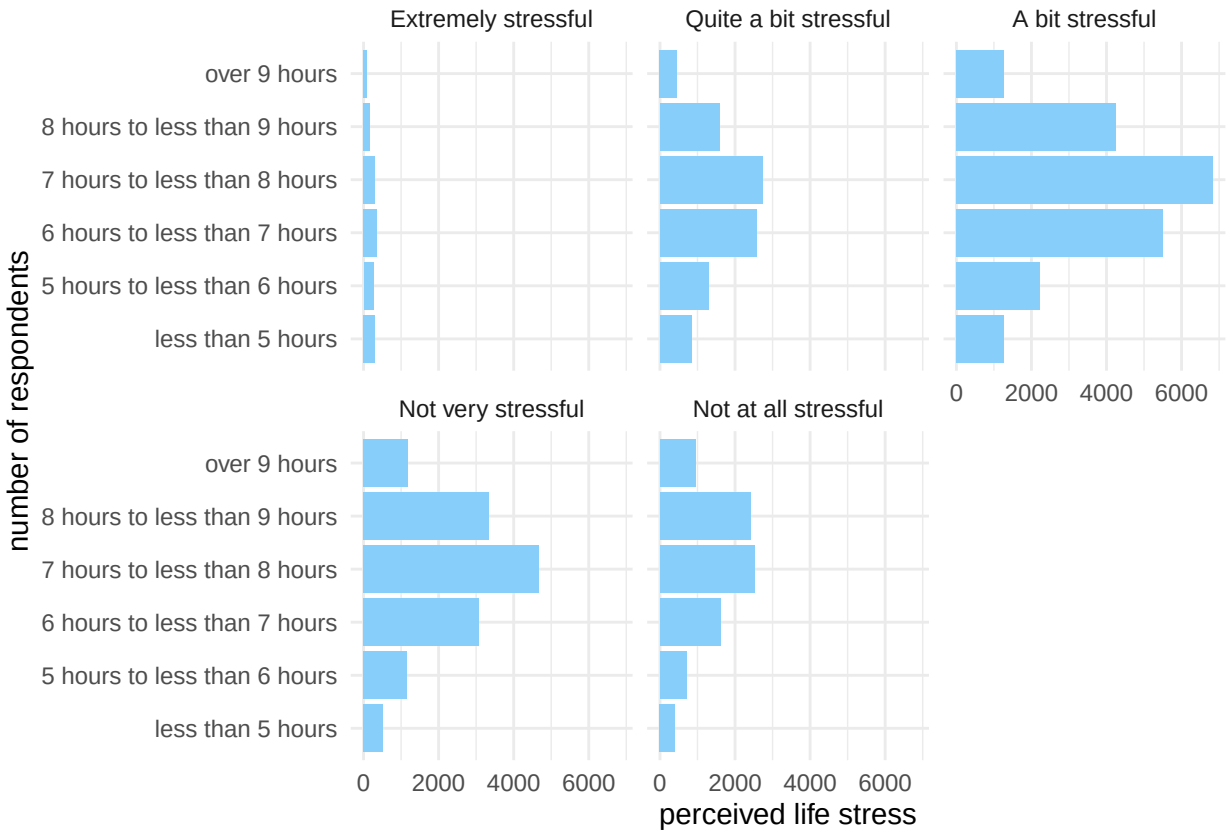

Figure 7: Comparing the number of hours of sleep across categories of stress in life

2.4.8 Perceived Stress at Work

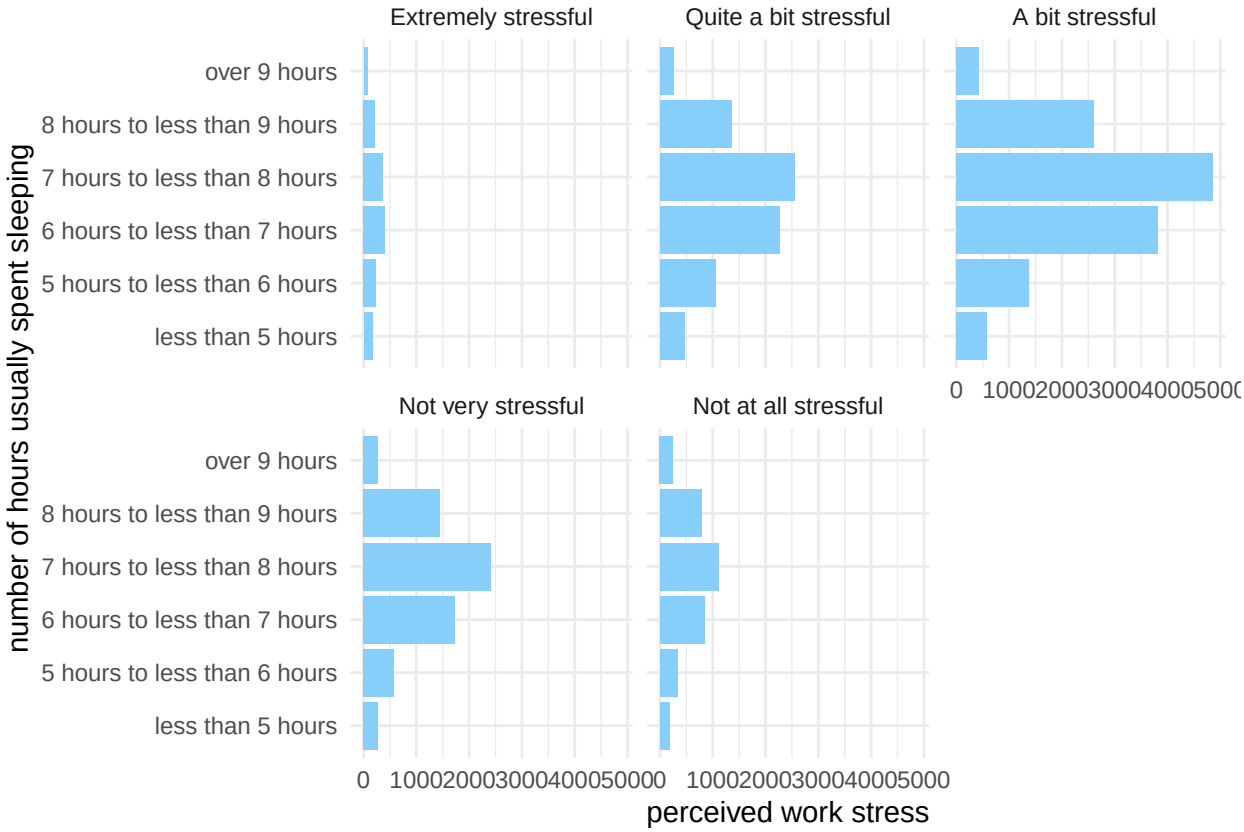

Figure 8: Comparison of the distribution of respondents based on their perceived stress at work

Figure 8 shows the comparison of the number of hours of sleep based on perceived work stress. Most people feel that their work is a bit stressful or not very stressful. Regardless of how they feel their work stress, people tend to sleep between 7 to 8 hours of sleep. It seems that work stress does not really have a noticeable effect on the number of hours usually spent sleeping.

2.4.9 Perceived Life Satisfaction

Figure 9 compares the number of hours slept based on life satisfaction. Life satisfaction is on a scale of 0-10, with 0 as being not satisfied at all and 10 as being fully satisfied with their life. Most people reported a 9 or 10 which is great.

3 Model

3.1 Logistic Regression

From the exploratory data analysis and research interest, we are interested in the effect of sex, marital status, age, income, perceived health, perceived mental health, life stress, work stress and life satisfaction on the number of hours spent sleeping. For some of the variables of interest, we have responses such as "Don't

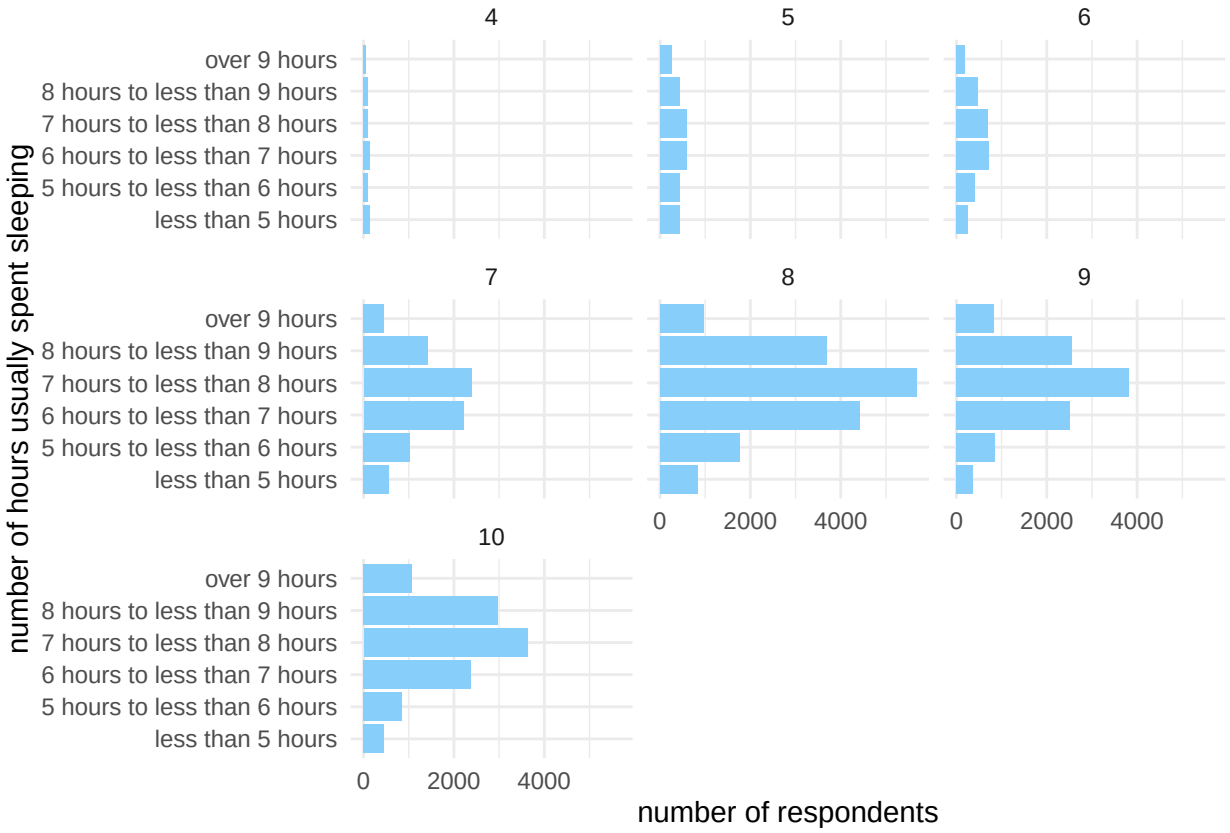

Figure 9: Comparing the number of hours usually spent sleeping against perceived life satisfaction

know", "Not stated" or "Refusal" for adolescents who did not receive parental approval for the survey; we removed them from the dataset which got narrowed down to 50,523 observations.

The number of hours usually spent sleeping, the dependent variable, is a categorical variable, so linear regression or poisson regression would not be appropriate models. 7 hours to 9 hours of sleep regularly is the healthy amount of sleep. (Ianzito 2014). It is considered an unhealthy habit to sleep too little or too much. Therefore, we assume that sleeping between 7 to 9 hours is considered sleeping "well". Based on that definition, we constructed a binary variable which takes 1 if the respondent reported sleeping between 7 hours to 9 hours and 0 otherwise. This new binary variable will be our dependent variable. The independent variables will be age, sex, marital status, income, perceived health, perceived mental health, perceived life stress, life satisfaction and perceived work stress.

Logistic regression is used due to its efficiency and popularity in assessing dichotomous outcomes. Since the observations come from different unique respondents, we will assume independence of observations.

We are interested in the following questions:

- Does the odds of having a good sleep depend upon whether the respondent is male or female?
- Was how people perceived their health and mental health related to them sleeping well or not?
- Is having more income associated with having a good sleep, regardless of sex?
- Is life satisfaction associated with having good sleep?

The logistic regression model was fitted in R (R Core Team 2020) using the lme4 r package (Bates et al. 2015). We fitted different possible models based on the variables of interest and compared the models using the drop-in-deviance test. The drop-in-deviance test is used to test the significance of model coefficients. The null hypothesis for the drop-in-deviance test is that the simple model explains the data as well as the complex model while the alternative hypothesis would be that the complex model explains the data better than the simple model. We used a 5% level of significance which means that if a p-value less than 0.05 is obtained, we have strong evidence supporting the complex model being better. The p-value provides a measure of strength of evidence.

The final model obtained is as follows:

$$log\ p/(1-p) = \beta_0 + \beta_1 age + \beta_2 income + \beta_3 sex + \beta_4 mental\ health + \beta_5 life\ stress + \beta_6 marital\ status$$

where:

- $log\ p/(1-p)$ represents the log-odds of sleeping well
- β_i ,for i = 1 to 6, represents the estimated coefficients of the predictors.

Finally, we performed a goodness of fit test by comparing the residual deviance (68560) to a χ^2 distribution with 50492 degrees of freedom. We found that our model with age, sex, income, mental health, life stress and marital status has statistically significant evidence of lack-of-fit ($p < .001$). The lack-of-fit could be the result of many reasons, namely: a) missing covariates, b) outliers, or c) over dispersion. We still choose to go with the model despite the seemingly lack-of-fit. As the statistician George Box said, all models are wrong, but some are useful. We are not trying to get the best possible model but one that we can interpret and provide insights on the data. With this in mind, the interpretation of the model results should not be taken as the only truth, rather as possible insights into the data.

4 Results

The reference group for the final model is female, aged 80 and over, with an income of 20,000 to 39,999 dollars, in a common-law relationship who feel life is extremely stressful and perceives her mental health as

Table 2: Summary of Coefficents

Variables	Estimated Coefficients	2.5%	97.5%
Reference level	0.661	0.567	0.771
mental health: Very Good	0.962	0.921	1.004
mental health: Good	0.764	0.728	0.803
mental health: Fair	0.590	0.542	0.641
mental health: Poor	0.469	0.394	0.555
Male	0.853	0.822	0.885
Age between 18 and 19	1.707	1.450	2.011
Age between 20 and 24	1.538	1.366	1.733
Age between 25 and 29	1.320	1.185	1.470
Age between 30 and 34	1.166	1.052	1.291
Age between 35 and 39	1.241	1.120	1.375
Age between 40 and 44	1.206	1.087	1.338
Age between 45 and 49	1.117	1.006	1.241
Age between 50 and 54	1.069	0.966	1.184
Age between 55 and 59	1.047	0.951	1.154
Age between 60 and 64	1.167	1.061	1.284
Age between 65 and 69	1.167	1.062	1.282
Age between 70 and 74	1.136	1.029	1.253
Age between 75 and 79	1.066	0.957	1.186
income: $40,000 to $59,999	1.074	1.018	1.132
income: $60,000 to $79,999	1.102	1.034	1.175
income: $80,000 or more	1.124	1.060	1.192
income: less than $20,000	0.899	0.856	0.945
No income or income loss	0.876	0.749	1.023
life stress: Quite a bit stressful	1.537	1.363	1.736
life stress: A bit stressful	1.902	1.692	2.140
life stress: Not very stressful	2.415	2.142	2.726
life stress: Not at all stressful	2.494	2.202	2.828
Married	0.895	0.846	0.946
Single	0.820	0.771	0.872
Widowed/Divorced/Separated	0.746	0.699	0.797

"Excellent". Since all independent variables are categorical, the results suggest how a change in one of the conditions affects the outcome of interest. The estimated coefficients indicate the effect of identifying in one of the categories.

The estimated coefficients and 95% confidence intervals in Table 2 are exponentiated and indicate the odds rather than the log odds. According to the model, males have 0.853 times higher odds of sleeping a healthy amount of hours compared to females, keeping all other conditions equal. Keeping all else equal, the odds of sleeping between 7 to 9 hours decreases as the perceived mental health get worse. Similarly, for how people feel how stressful their life is, the less stressful they feel about their life, the higher the odds of them sleeping an adequate amount of time, keeping all else constant. For example, the odds of sleeping well for someone not at all stressful about life is 2.494 time higher than someone who feel extremely stressful, keeping all else constant.

With 95% confidence, we may claim that for respondents aged 18 to 19, the odds of having adequate sleep is between 1.450 and 2.011 times higher, keeping all else equal. Another interpretation is with 95% confidence, we may claim that the odds of sleeping enough for those who are married is between 0.846 and 0.946 higher than those in a common-law relationship, keeping all else constant.

5 Discussion

5.1 Findings

This paper seeks to provide some insight on the sleeping trends of Canadians. To that end, we use logistic regression to model the data with a binary response which indicate whether someone slept well or not. We have defined sleeping well as having between 7 hours to 9 hours of sleep. We found that life stress and mental health both have significant effect on whether someone sleeps well. As Dr. Michael Wainberg said, sleep and mental health have a bi-directional relationship. (Addiction and Heath 2021) Poor sleep contributes to poor mental health and the converse is also true. 80% of people with mental health problems have possible problems with falling asleep and staying asleep. (Addiction and Heath 2021) The results in Table 2 indicate similar findings. Those with poor mental health have lower odds of sleeping well compared to those with excellent or very good mental health. The reason for which it is important to investigate such relationship is by identifying the association of the lack of sleep and poor mental health, we can better explain how inadequate sleep raises the risk of mental problems. Lack of sleep is strongly associated with higher odds of frequent mental distress. (Amanda Blackwelder 2021) Therefore, it may be better to take a closer look at the reasons why people do not sleep enough as possible way to help reduce mental distress.

Similarly, the less stressful people feel in life, the higher the odds of sleeping adequately. This is no surprise since stress is correlated with mental health. Chronic stress can be conducive to mental health problems such as anxiety and depression. (Addiction and Health, n.d.) This can also be supported by Figure 7, most people sleep between 8 to 9 hours, regardless of the amount of stress felt. However, the more stressful people feel, the proportion of people having a lack of sleep increases.

5.2 Limitations and weaknesses

The survey questionnaire was subject to a redesign in 2012 and changed in 2015. About 70% of previous content was retained. It is thus not appropriate to compare results with those prior to the redesign. The data is self-reported, so the way one respondent perceive his/her mental health differ. Since mental health is still a sensitive topic, not everybody may be willing to truthfully disclose how they feel their mental health are. Similarly, for other variables which depend on one's own perception, one issue is how do one associate with the categories; it is hard to properly define the categories since everyone may feel differently.

Since this survey is a cross-sectional one, the time difference in collecting data may cause some difference in responses. One example is if there is an event that make people feel happy in between 2 collection times, the answers to some questions of the survey may be slightly different.

Another weakness is the fact that we have assume that the healthy amount of sleep is between 7 to 9 hours. It is not account for some rare individuals who are fine with less than 6.5 hours of sleep. (Newman 2020) Different people may need different amount of sleep. The logistic regression model used showed significant lack of fit but that was the best model we have learned so far. The results should not be interpreted as unquestionnable truth.

This paper has a focus on the effect of mental health on sleep. Our findings along previous studies have indicated that inadequate sleep increases the odds of developing mental problems. Since mental health is a serious issue currently, one possible area we can investigate is the reasons why people do not have enough sleep so as to come up with possible measures to alleviate the situation.

A Appendix

A.1 Datasheet

Extract of the questions from Gebru et al. (2021)

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 enable analysis of sleep and variables affecting it. The survey dataset covers a wide range of areas and what we are interested in is only a small subset of it. The dataset is not publicly available.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset was created by Pascal Lee Slew from the 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.*
    - No funding was obtained/required for the creation of the dataset.
4. *Any other comments?*
    - None

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 dataset consists of categorical variables and an identifier. It contains 113,290 observations, so 113,290 rows and 18 columns.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 113,290 rows and 18 columns.
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 contained all the data possible.
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.*
    - The dataset has already been processed from the survey results.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - No
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, some respondents choose to not respond to some questions, leading to some missing data in the data collection process.

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.*
   - No
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.*
   - There may be possible sampling error.
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.*
    - The raw dataset was obtained through UofT CHASS.
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, the dataset has been processed such that all confidentiality information are suppresses
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, for sex, there are 2 categories: male and female. For age, there are multiple age groups of 5 years
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.*
    - No
16. *Any other comments?*
    - None

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.*
    - No, the raw dataset was created from survey responses in the 2018 CCHS.
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 survey responses were obtained by filling application using computer-assisted interviewing.
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 dataset is a subset of the 2018 CCHS dataset.
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)?*
    - The raw data was collected by trained interviewers.
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.*
    - Data was collected over three months period 4 times.
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.*
    - Unknown to the author
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
    - The raw data was collected directly while the dataset created was a subset of it.
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 individuals were informed how the information provided would be used.
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.*
    - Yes, the participants were informed of the purpose of the survey and gave consent to participate in it before collecting data.
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).*
    - Unknown to the author
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.*

- N/A

12. *Any other comments?*
    - None.

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, the data was cleaned using the tidyverse (Wickham et al. 2019) and janitor (Firke 2021) r packages.
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, but the data is not allowed to be made publicly available. The code used to obtain the dataset can be found at: https://github.com/Pascal-304/canadian-sleep-2018
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Yes, we used R software (R Core Team 2020)
4. *Any other comments?*

- None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset has been used to conduct an analysis of Canadian sleep. We created graphs and used a logistic regression model.
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.*
    - The repository containing all files pertaining to the dataset is available at: https://github.com/Pascal-304/canadian-sleep-2018/
3. *What (other) tasks could the dataset be used for?*
    - None
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?*
    - N/A
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - N/A
6. *Any other comments?*

- None

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.*
   - No
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - No
3. *When will the dataset be distributed?*
   - It will not be distributed but the code used to obtain the dataset is already available at: https://github.com/Pascal-304/canadian-sleep-2018/
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
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.*
   - Unknown to the author of the dataset
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.*
   - Unknown to the author of the dataset
7. *Any other comments?*
   - None

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Pascal Lee Slew
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - If anyone want to reach out to me, you can contact me at pascal.leeslew@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)?*
   - If there is any erro made in the creation of the dataset, it will be updated by Pascal Lee Slew. Any updates will be communicated as a note in the Readme file at the link below: https://github.com/Pascal-304/canadian-sleep-2018#readme

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.*
    - N/A
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.*
    - N/A
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.*
    - N/A
8. *Any other comments?*
    - None

References

Addiction, Centre for, and Mental Health. n.d. *Stress*. https://www.camh.ca/en/health-info/mental-illness-and-addiction-index/stress#:~:text=When%20stress%20becomes%20overwhelming%20and,complaints%20such%20as%20muscle%20tension.

Addiction, Centre for, and Mental Heath. 2021. *Mental Illness Associated with Poor Sleep Quality According to Largest Study of Its Kind*. https://www.camh.ca/en/camh-news-and-stories/mental-illness-associated-with-poor-sleep-quality.

Amanda Blackwelder, Larissa Huber, Mikhail Hoskins. 2021. *Effect of Inadequate Sleep on Frequent Mental Distress*. https://www.cdc.gov/pcd/issues/2021/20_0573.htm.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Canada, Statistics. 2020a. "CCHS 2017-2018 PUMF Complement User Guide.pdf." In *Canadian Community Health Survey - Annual Component (CCHS) 2017-2018*. Abacus Data Network. https://doi.org/11272.1/AB2/SEB16A/YCYAK0.

———. 2020b. "CCHS_annual_2017_2018.tab." In *Canadian Community Health Survey - Annual Component (CCHS) 2017-2018*. Abacus Data Network. https://doi.org/11272.1/AB2/SEB16A/J9QFAO.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Ianzito, Christina. 2014. *Regularly Sleeping Too Long May Indicate a Health Problem*. https://www.washingtonpost.com/national/health-science/regularly-sleeping-too-long-may-indicate-a-health-problem/2014/03/07/6ce03894-ade5-11e2-98ef-d1072ed3cc27_story.html.

Newman, Tim. 2020. *Medical Myths: How Much Sleep Do We Need?* https://www.medicalnewstoday.com/articles/medical-myths-how-much-sleep-do-we-need.

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

Wickham, Hadley. 2021. *Forcats: Tools for Working with Categorical Variables (Factors)*. https://CRAN.R-project.org/package=forcats.

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, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.