Men and past smokers from ontario are potentially more physically active than women and non-smokers*

Women are engaging in lesser physical activities than men. Furthermore, individuals who smoked before are more likely to engage in physical activity. Consumption of stimulant-related drugs could have also increased as well, resulting in increased overall physical activities in Ontario.

Chris Yong Hong Sen

December 4, 2024

This study explores factors influencing physical activity levels in Ontario, focusing on demographic and substance abuse. Our findings reveal that women engage in less physical activity than men, while individuals who have smoked in the past and those consuming stimulant-related drugs are more likely to be physically active. Additionally, age, highest educational attainment, and health region also play significant roles in determining physical activity levels. These insights highlight the complex interplay of gender, lifestyle behaviors, and sociodemographic factors, providing valuable information for designing targeted public health interventions to promote physical activity.

1 Introduction

Physical activity is a cornerstone of public health, yet participation rates remain suboptimal across various populations in Ontario. This is especially concerning as physical inactivity is a leading contributor to chronic diseases such as heart disease, diabetes, and mental health disorders. Factors such as age, sex, educational attainment, alcohol consumption, smoking habits, and regional disparities in health access play a critical role in shaping physical activity behaviors. Understanding these factors and their complex interactions is crucial for developing targeted interventions that can improve health outcomes and reduce health inequalities in Ontario.

*Code and data are available at: https://github.com/Monoji77/Alcohol_Use_Ontario.

Recent studies suggest that certain demographic groups, such as older adults and individuals with lower educational levels, are more likely to experience barriers to physical activity, while alcohol consumption and smoking are associated with lower physical activity levels (Bauman et al. 2012). However, limited research has explored how these factors specifically interact in Ontario, a province with diverse populations and health regions. As the burden of chronic disease grows, addressing these disparities becomes even more urgent (Warburton, Nicol, and Bredin 2006) .

This study seeks to address the gaps in understanding by exploring how age, sex, educational attainment, alcohol consumption, smoking, and health regions influence physical activity patterns in Ontario. By examining these relationships, we aim to provide actionable insights that can inform public health policies and interventions tailored to the needs of specific population groups. Our findings will be pivotal in developing strategies to increase physical activity participation and improve health equity in Ontario.

The estimand of this paper would be the actual effect of socio-demographic factors, substance abuse, and health factors on how much time is spent on strenuous activities among individuals in Ontario who engage in physical activities.

The remainder of this paper is structured as follows. Section 2 would provide a sense of the chosen Computing in the Humanities and Social Sciences dataset as well as data exploration of interested variables. Section 3 goes into the details of the linear regression model fitted on the predictors of interest variables, for which it's results would be showcased in Section 4. Finally Section 5 would interpret the results of the fitted linear regression model, which includes limitations of the paper. At the end of the paper in Appendix A, this contains diagnostic checks for all models considered.

Overall, this paper was generated using R programming language (R Core Team 2023). Reading of raw data and data manipulation was performed using `Tidyverse` package (Wickham et al. 2019) and parquet file handling was done using functions in `arrow` package (Richardson et al. 2024). Generation of maps were done using functions in `tmap` package (Tennekes 2018). Any other visual plots were using `ggplot`-related functions in Tidyverse (Wickham et al. 2019) and tables were generated using `kable` function in `knitr` package (Xie 2015).

2 Data

2.1 Context of dataset

The dataset chosen is obtained from Computing in the Humanities and Social Sciences/CHASS at the university of Toronto (Data Centre, Faculty of Arts & Science, University of Toronto 2018), specifically the Canadian Community Health Survey (CCHS) annual component 2017-2018 which was collected by Statistics Canada (2018). The goal of the CCHS survey is to collect comprehensive and reliable health-related data on the Canadian population to support

health surveillance, research, and program evaluation. It aims to provide timely, accessible, and flexible data for monitoring health trends, studying small populations and rare characteristics, and addressing emerging health issues, ultimately helping improve the health and well-being of Canadians.

Other datasets explored include sources from World Health Organisation and ParticipACTION. World Health Organisation provides aggregated data informing certain indicators of interest such as obesity levels and alcohol consumption. This can be found in the data section for Canada(World Health Organization 2024). Also, ParticipACTION provides key statistics regarding proportion of adult population in canada meeting the national guidelines of 150 minutes of moderate to vigorous physical activites, as well as how sedentary canadians are (ParticipACTION 2024). However, the availability of raw individual level data is absent from these sources. Although the aggregated data gives us a sensing about general Canadian health, we are not able to focus our attention to the Ontario population. The dataset chosen from Canadian Community Health Survey (CCHS) annual component is ideal as it contains individual level data, where we are able to obtain the province the respondents are from.

2.2 Original Dataset

The dataset provides 113,290 individual response from Canadians all throughout the country. Surveys were conducted to collect health-related data from respondents. The relevant data collected includes…

- Demographics and Identity: Age, gender, marital status, citizenship, Aboriginal identity, visible minority status, immigrant status, language (mother tongue, spoken at home, first official), sexual orientation.
- Health Status and Conditions: Chronic diseases, neurological conditions, mental health indicators, perceived physical and mental health, pain, discomfort, activity limitations, and disabilities.
- Health Behaviors and Lifestyle: Smoking, drinking, cannabis use, physical activity, helmet use, dietary practices (e.g., fruits, vegetables), breastfeeding, and BMI.
- Health Care Access and Utilization: Contact with health professionals, access to health services, immunizations (e.g., flu shots), cancer screenings (e.g., mammograms, pap smears), waiting times, and satisfaction with services.
- Socio-Economic Factors: Educational attainment, work activity, household income, living arrangements, presence of children in the household, food insecurity.
- Life Quality and Perceptions: Life satisfaction, life stress, sense of belonging, health-adjusted life expectancy.
- Specialized Health Indicators: Two-week disability days, quality and ratings of health care services, functional health status.

- Specific Population Measures: Aboriginal group, immigrant duration, visible minority, minority health disparities.
- Geographic and Household Data: Geographic location, number of persons or households, household demographics.

This dataset is highly suitable in determining both quantitative factors such as c qualitative factors

2.3 Measurement

Our goal is to explore the actual effect of socio-demographic factors, substance abuse, and health factors on how much time is spent on physical activities among active individuals in Ontario. The instrument of measurement would be surveys. Through the cross-sectional survey conducted by CCHS, we are able to get the number of hours spent on physical activities per individual, our response variable.

According to the questionnaire, respondents were asked in a phone survey to indicate whether they had engaged in any strenuous activity that lasted at least 10 minutes and made them sweat and breathe a little harder in the past 7 days. Afterwards, they were asked to indicate the total time spent on those activities. To address the concern of recall bias where an individual may not recall the specific timing in engaging with physical activity, surveyors sectioned physical activity into commuting, sports, work and volunteering. Then within each section, surveyors asked which day (Monday to Sunday) in the past week had they engaged in such activities that made them sweat and breathe a little harder. By doing so, the activities individuals did during the past week would be anchored on the days of the week. Following this would be for the respondents to indicate the total number of hours they engaged in each section of physical activities for the past 7 days.

In addition, the explanatory variables we are interested in were also obtained in the same survey.

For socio-demographic factors possibly affecting time spent on physical activites, the surveyors asked the respondent for their age (DHH_AGE) from 0 years old to 121 years old), sex (DHH_SEX) being either male or female, personal income obtained from tax forms submitted to the Canada Revenue Agency upon consent from respondent, geographical health region (ADM_Q037) that was provided by respondents when they agreed to the phone survey, and highest educational attainment (EHG2_04) ranging from less than high school diploma to master's degree and above.

For substance abuse, respondents were asked how frequent they consumed alcoholic drinks in the past 12 months (ALC_Q015), ranging from less than once per month to everyday. Surveyors also asked respondents for usage of drugs including one-time marijuana or hashish in the past 12 months (DRG_Q005 to DRG_Q075), and whether they have smoked hundred

cigarettes in the past 12 months (SMK_Q020). The last 2 questions allowed respondents to indicate a simple yes or no flag to express their answer.

For health factors, questions were asked to determine an individual's body mass index ranging from underweight to obese and perceived life/work stress ranging from not at all stressful to extremely stressful. Additionally, surveyors asked respondents if they have mood disorders (E.g. depression, bipolarism, mania), for which respondents replied with a simple yes or no.

Overall, each entry obtained by respondents assigns quantitative values to each factor and the the response for which we can explore subsequently using linear regression to determine which factors have greater influence in the time spent on physical activities amongst active individuals.

2.4 Response variable: Time spent on physical activity in the past 7 days

From the CCHS dataset, we look at paadvmva variable. This variable is continuous with values ranging from 0 to 9902 representing the number of minutes spent on physical activity in the past 7 days. For easier readability, the variable will be renamed as $time\_spent\_vigorous\_exercise\_7d$. As explained in section Section 2.3, this variable measures the time spent on any activities lasting more than 10 minutes that makes an individual sweat more than they normally would. This would include individuals undertaking laborious employment or volunteering efforts, or it could simply include individuals who do cardio and/or strength training.

By using this variable, we are assuming that the routine of an individual from the past 7 days is indicative of their general routine beyond the past 7 days. This variable may not account for individuals who only began a physical routine involving sweat within the past 7 days and were sedentary before, which may not be helpful if included in subsequent model fitting.

Since we are interested in individuals that are already engaging in strenuous activities, we can filter $time\_spent\_vigorous\_exercise\_7d$ to exclude 0. After removing such entries, we obtain the following summary statistics as shown in Table 1. The median time spent on physical activities the past 7 days would be 300 minutes, which averages out to 43 minutes per day.

Table 1: Summarised statistics for time spent on physical activities in the past 7 days

Min.	1st Qu.	Median	3rd Qu.	Max.
10	135	300	600	9902

Interestingly, the maximum time spent on physical activities the past 7 days was 9902 minutes, and this averages to 1415 minutes per day, which is equivalent to 23.5 hours in a day. This is very unlikely as this suggests that an individual is constantly engaging in physical strenuous

activity without sleeping. Logically, this suggests the possibility of outlier values. By observing Figure 1, there are many outlier observations for this variable. Roughly 10% of the observed data are considered outliers as they fall outside 1.5 times the interquartile range from the 25% and 75% quartile values. We would be performing log transformation on our response variable. The motivation for this is detailed in Appendix A.1.

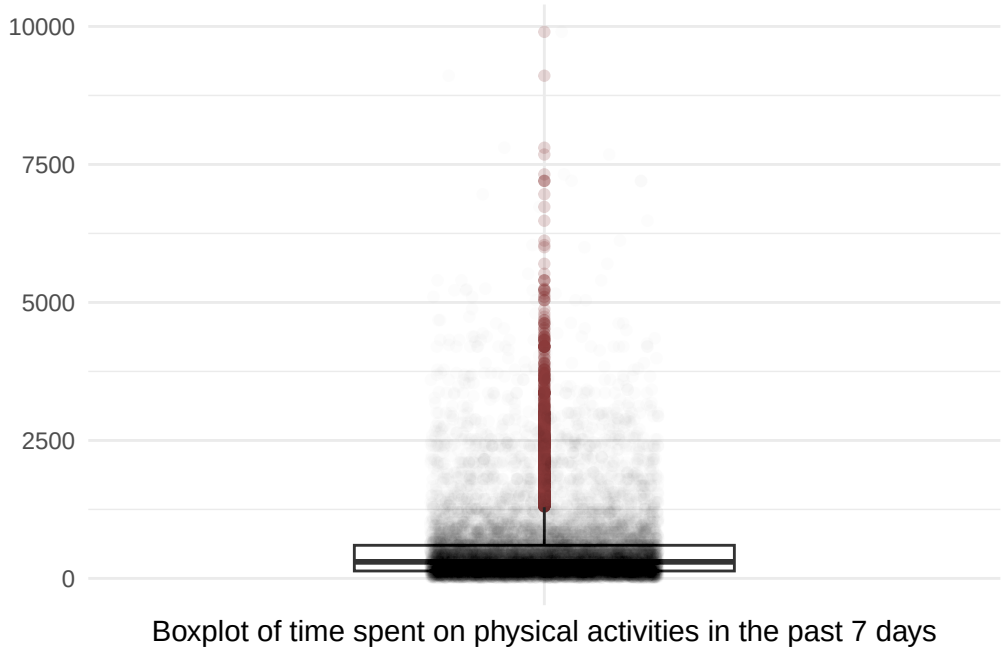

Figure 1: Boxplot of time spent on physical activities in the past 7 days

Subsequently, we would have to remove observations that are very unlikely as they do not represent the general population and could skew the distribution of time spent on physical activities.

By observing the histogram generated in Figure 2, we obtain a severely right skewed response histogram where most observations lie between 0 to 500 minutes. For the building of linear models in subsequent sections of the paper, this would not be ideal as linear regression assumes a normal distribution of the response variable, suggesting that a transformation would be required to fit a linear regression model.

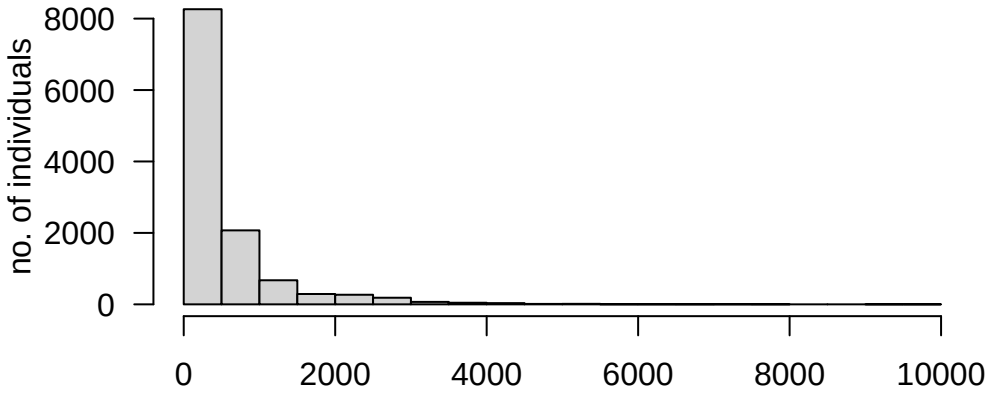

Figure 2: Histogram of time spent on physical activities the past 7 days

2.5 Interested explanatory variables

There are a total of 13 predictors of interest that is seemingly related to the time spent on physical activities by an individual in Ontario.

2.5.1 Variable 1: Age

This is a variable of interest because age is seemingly correlated to The age groups of respondents were obtained and they exist in bins that spans ages 18 years to 74 years old.

- 3: 18-19 years old
- 4: 20-24 years old
- 5: 25-29 years old
- 6: 30-34 years old
- 7: 35-39 years old
- 8: 40-44 years old
- 9: 45-49 years old
- 10: 50-54 years old
- 11: 55-59 years old
- 12: 60-64 years old

- 13: 65-69 years old
- 14: 70-74 years old

From Figure 3, we can immediately see that most observations fall within age groups 30 to 34 years old (bin 6), and lesser observations fall within age groups 70-74 years old (bin 14).

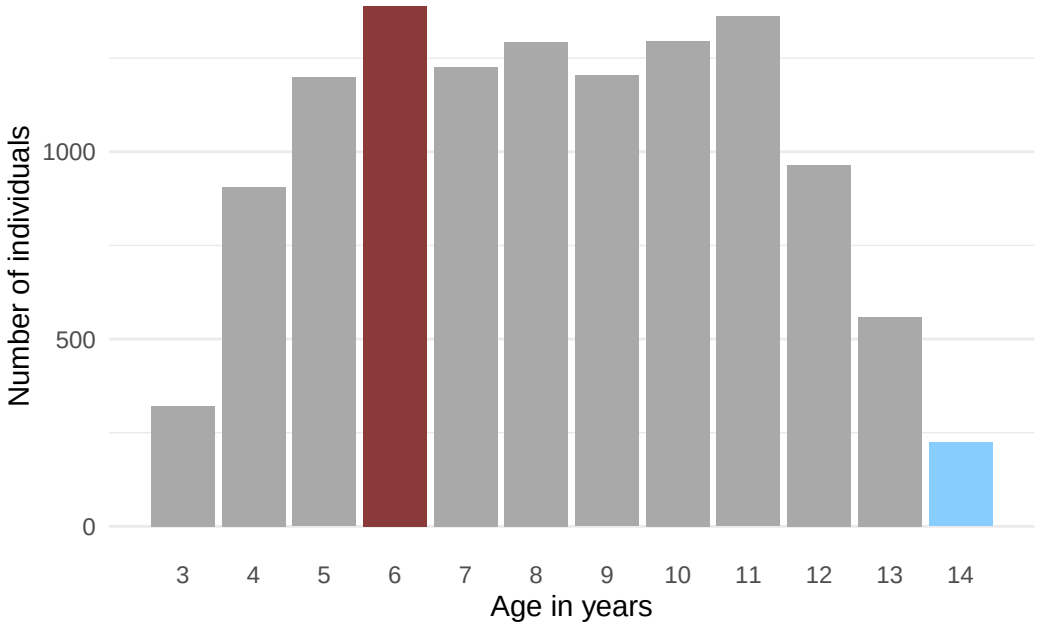

Figure 3: Histogram of time spent on physical activites the past 7 days

From Table 2 and Figure 4, we are able to see that the the median age bin belongs to age group 40-44 years old. We are able to see that there are no outlier values from Figure 4, which means that the survey has obtained an acceptable number of respondents across all age bins.

Table 2: Summarised statistics of respondent ages

Min.	1st Qu.	Median	3rd Qu.	Max.
3	6	8	11	14

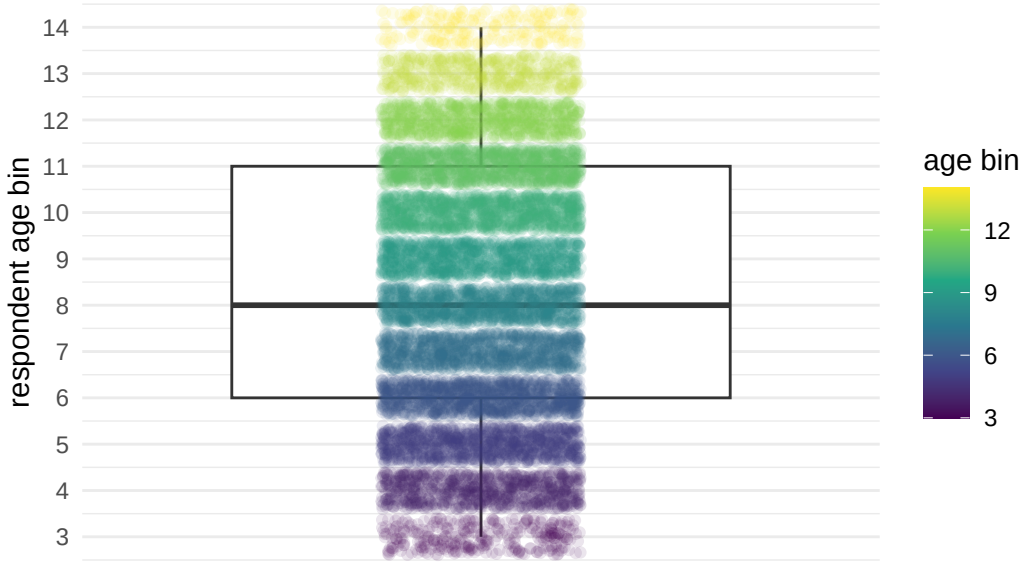

Figure 4: Boxplot of respondent ages

2.5.2 Variable 2: Sex

Sex is a possible influential predictor for total time spent on physical activity. Historically, men are associated with exercise and involved in more labour intensive roles such as construction which involves heavy lifting. According to World Health Organisation (2024), women are less active than men by 5% since 2000. However, with the advent of female only establishments such as female gyms, and 32% increase in memberships at fitness and health clubs from 2010 to 2019, it would be interesting to see if the sex of an individual still affects the time spent on physical activites among physically active individuals.

According to Figure 5, there is roughly an equal proportion of sexes from our dataset, with slightly more males (6082 respondents) than females (5859 respondents). Since this was a health survey conducted over phone call, there is an equal chance of a respondent to be male or female, which is roughly reflected ing Figure 5.

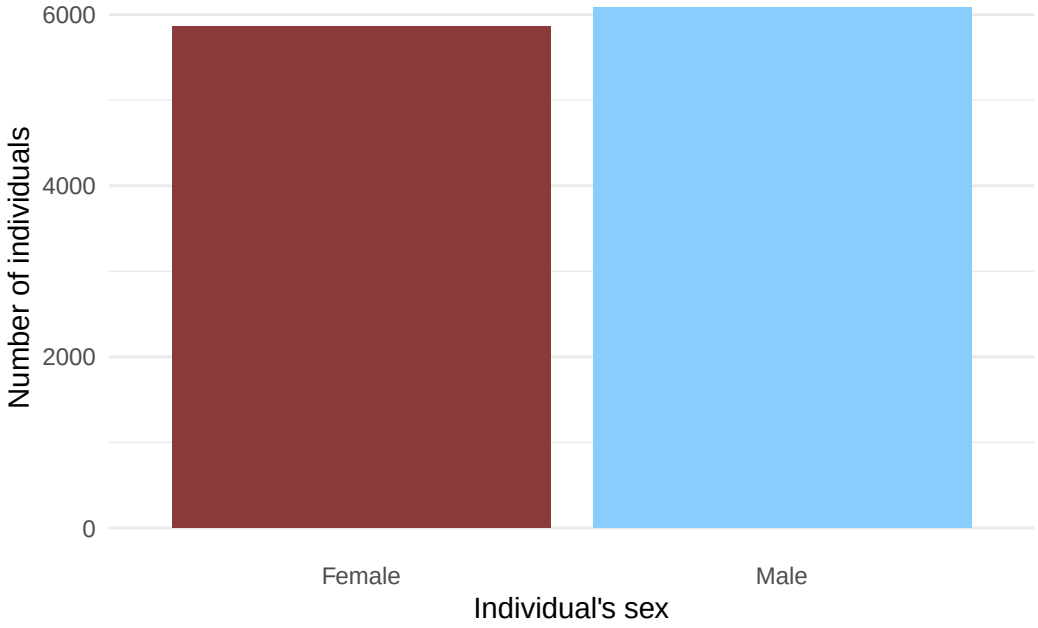

Figure 5: Barplot of sex

2.5.3 Variable 3: Highest level of education

In 2023, Kari et al. (2020) has done a study to show that highest educational attainment may be a leading factor for physical activity. Therefore, it would be worth exploring the effects of highest levels of education with the time spent on physical activities among the Ontario residents.

According to Table 3, respondents who did not graduate secondary school account for only 5% of the population, and respondents who have at least a bachelor's degree account for 72% of the population. According to Figure 6, there seems to be class imbalance for this variable.

However, as seen from post-secondary institution enrollment data from Statista Research Department (2024), there has been an increase in total number of Canadians enrolled in post secondary education from 2000 to 2020. Accounting for population changes from 2000 (The Daily, Statistics Canada 2000) to 2022 (Statistics Canada 2023), there is still a real increase in the number of real enrollments into post secondary education. Therefore, it is very possible that this is representative of the Ontario's population where majority of the population has at least post secondary education.

Table 3: Proportion of respondents grouped by highest educational level

highest_educational_attainment	proportion_of_respondents
less than sec sch grad	5%
sec sch grad	23%
post sec certification and above	72%

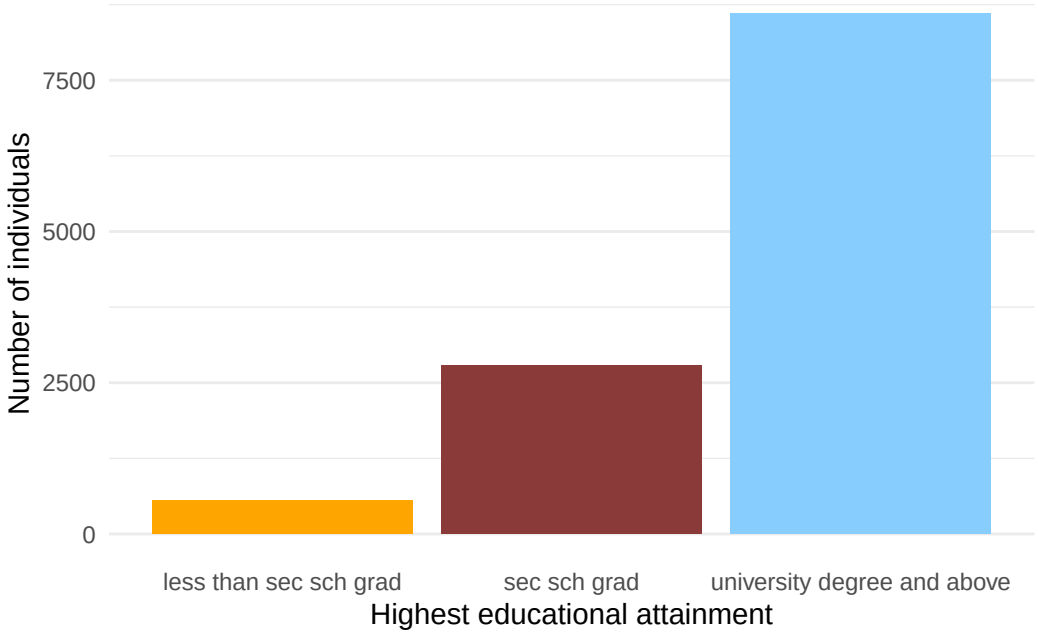

Figure 6: Barplot of highest_educational_attainment

2.5.4 Variable 4: Health Region

Health region is a categorical variable which tells us the geographical location an individual belongs to defined by Provincial Ministries of Health (as detailed by documentation provided by CHASS). It could be possible that certain health regions have more funding that are projected towards increasing awareness towards having an active lifestyle. Therefore, it is worth exploring the relationship between health region and time spent on physical activities.

There are a total of 34 different health regions in the Ontario province. They are encoded using 5 digits. Figure 7 shows us the various health regions in Ontario. The blue regions shows us health regions for which lesser respondents belong to while red regions show us health regions for which more respondents belong to.

Highest occurring health regions

- 35951: City of Ottawa Health Unit
- 35953: Peel Regional Health Unit
- 35960: Simcoe Muskoka District Health Unit
- 35970: York Regional Health Unit
- 35995: City of Toronto Health Unit *(highest number of observations)

Lowest occurring health regions

- 35931: Elgin-St Thomas Health Unit
- 35935: Haliburton, Kawartha, Pine Ridge District Health Unit
- 35940: Chatham-Kent Health Unit
- 35942: Lambton Health Unit
- 35956: Porcupine Health Unit *(lowest number of observations)

By using a GIS focused library in R called tmap (Tennekes 2018), we are able to generate the map shown in Figure 7.

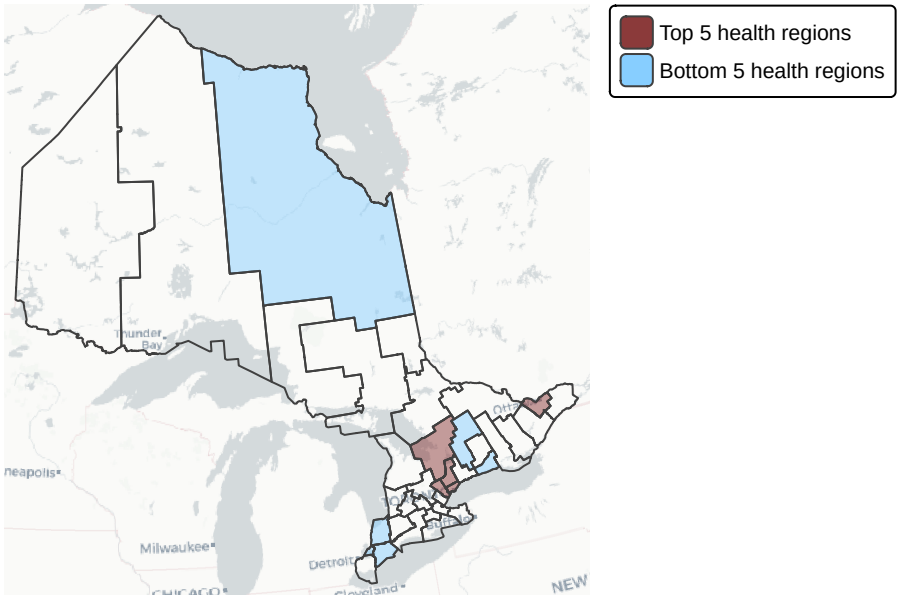

Figure 7: Map of top five and bottom five occuring Ontario health

Overall, Figure 8 shows us the actual number of observations belonging to the top five and bottom 5 occurring health regions. We can see that most respondents come from the City of

Ontario Health Unit (Health_Region = 35995) and lesser respondents belong to Porcupine Health Unit (Health_Region = 35956).

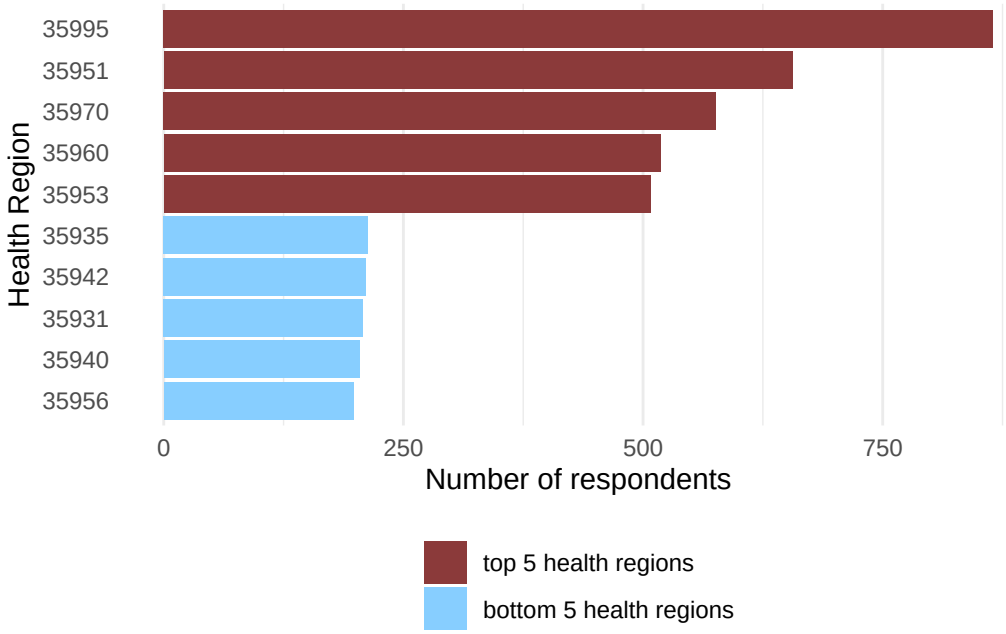

Figure 8: Barplot of top and bottom five occuring health regions

2.5.5 Variable 5: Personal Income

In the dataset, personal income (INCDGPER) is a categorical variable that places an individuals annual income into a range of income values. Below summarises the encoded number and the corresponding meaning of each encoded number. All income values are measured per year.

- 1: No income or income loss
- 2: Less than $20,000
- 3: $20,000 to $39,999
- 4: $40,000 to $59,999
- 5: $60,000 to $79,999
- 6: $80,000 or more

According to Figure 9, most of our respondents (25% of the sample) in Ontario are earning between $20,000 to $39,000 as part of their annual personal income. It is also worth mentioning that there are very few respondents who are not earning any income in that year. They account for less than 0.5% of the sample. In the 2018 annual wages report provided by Statistics

Canada (2020), the annual personal income reported from tax filing in Ontario that year was $39,510. Therefore, the sampling population's personal income data is representative of Ontario individuals.

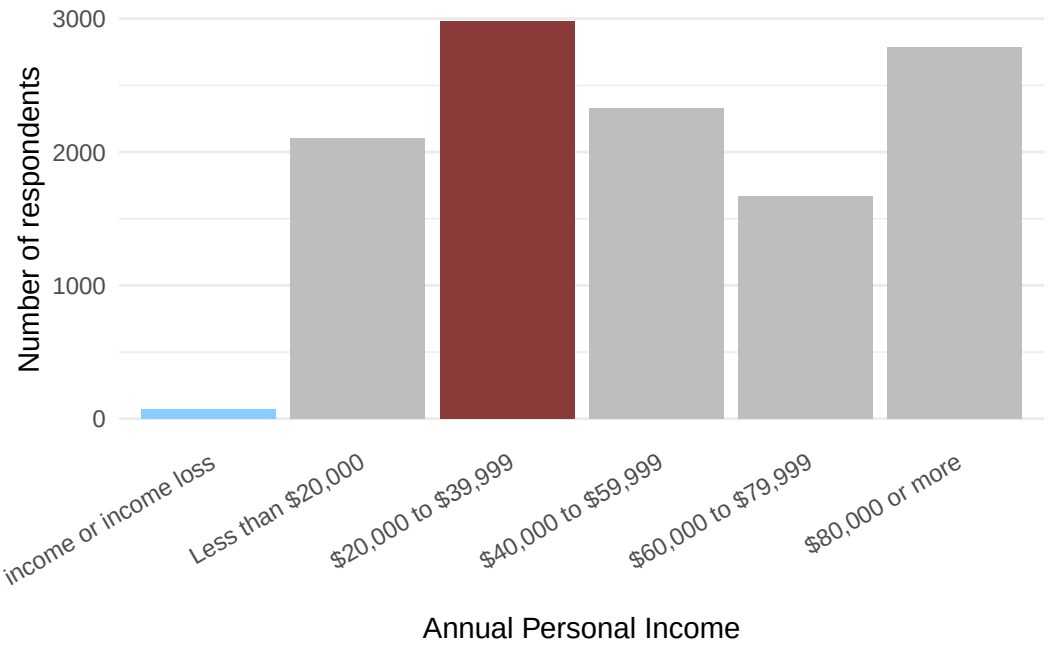

Figure 9: Barplot of personal income

2.5.6 Variable 6: Frequency of alcohol consumption

This variable is a categorical variable with an implied order.

- 1: Less than once a month
- 2: Once a month
- 3: two to three times a month
- 4: once a week
- 5: two to three times a week
- 6: four to six times a week
- 7: every day

From Figure 10, we are able to see that most respondents from this dataset drink about 2-3 times per week (2973 respondents) as seen from the red bar. Also, we are seeing fewer respondents drinking about 4-6 times per week (887 respondents) as seen from the blue bar.

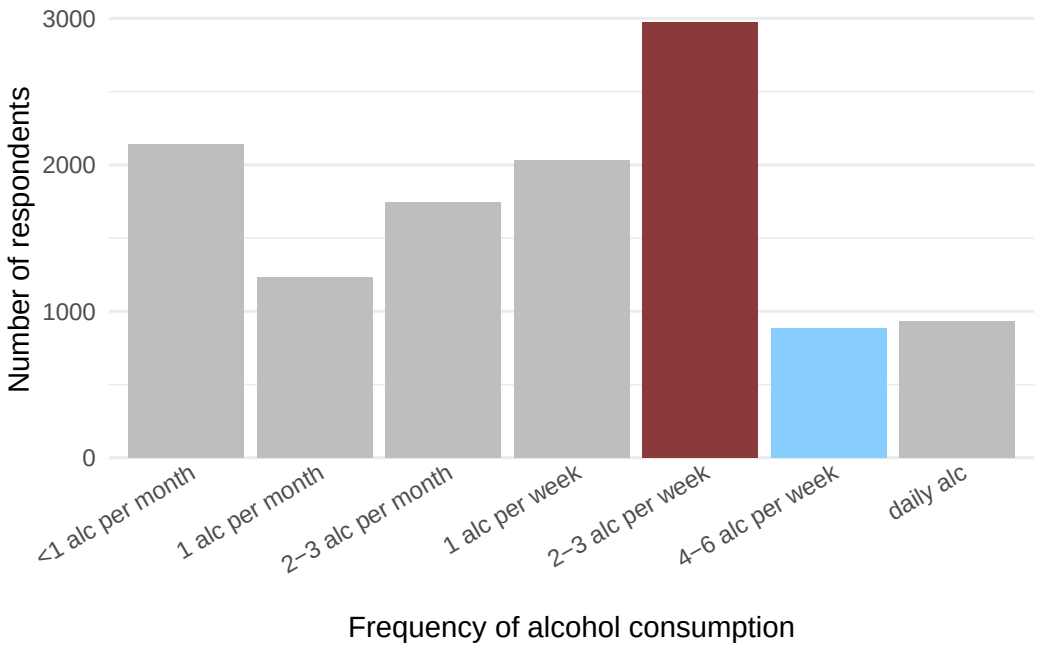

Figure 10: Barplot of frequency of alcohol consumption

2.5.7 Variable 7: Illicit drug use the past 12 months

This is a binary variable which indicates whether an individual has used drugs in the past 12 months prior to the interview. Drugs include marijuana, hashish, cocaine, amphetamines, speed, methamphetamine, crystal meth, ecstasy, hallucinogens, and sniffing glue, gasoline or other solvents. If they have attempted any of these drugs, this variable would be flagged as '1' and '0' otherwise.

It would be interesting to see if drug consumption has an effect on physical activeness. Brellenthin and Lee (2018) exerts that individuals who are more active tend to have lower substance abuse, suggesting a correlation between these two. Having this variable would be interesting to explore the relationship of an individuals drug usage and their time spent on physical activities.

Figure 11 shows us the number of respondents who have (or haven't) consumed drugs 12 months prior to their interviews. We can see that majority of respondents have not consumed drugs in the given timeframe. However, from Table 4, we can see that there are still a number of respondents who had consumed drugs in the past 12 months, particularly a fifth of the sampling population. According to Canadian Mental Health Association (2024), they estimate that about 21% of canadians, or 6 million people, will meet the criteria for substance use in their lifetime, and this is inclusive of alcohol and that there is a growing opiod crisis in Ontario.

Therefore the obtained sample is somewhat representative of usage of drugs among the Ontario population.

Table 4: Summary of illicit drug use the past 12 months

illicit_drug_use	num_of_respondents	prop_of_population
never consumed drugs	9328	78%
consumed at least once	2613	22%

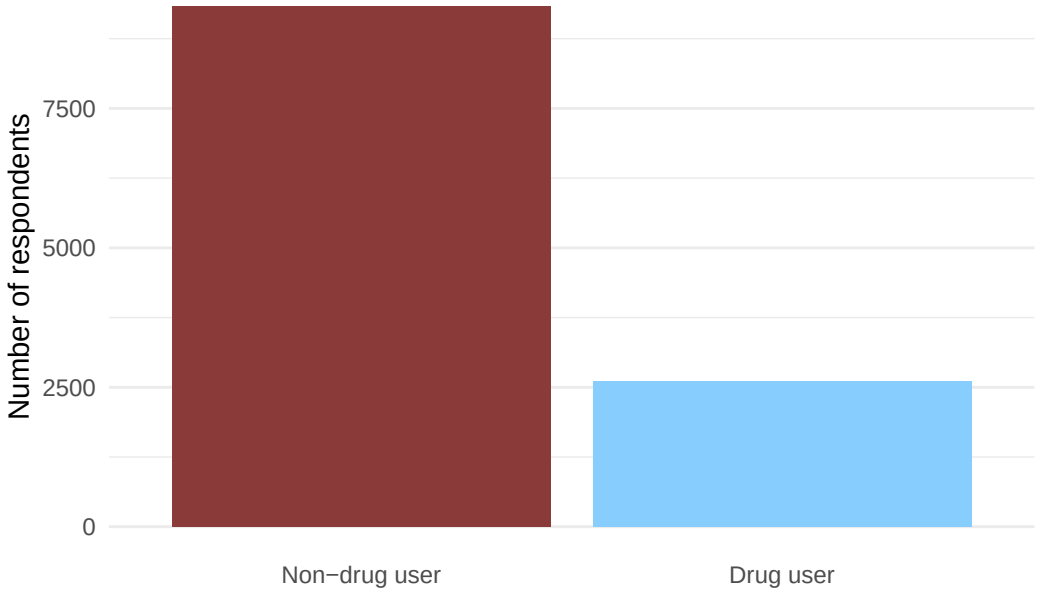

Figure 11: Barplot of illicit drug use the past 12 months

2.5.8 Variable 8: Cigarette consumption

Respondents were asked whether they had smoked more than 100 cigarettes in their lifetimes. This variable is therefore binary, with '1' representing respondents that meet the critera and '0' otherwise.

It is well known that excessive smoking has negative health impacts on individuals. A cross sectional study conducted by Heydari et al. (2015) found that smokers are less likely to be physically active compared to non-smokers. Hence, it would be interesting to discover if there is an effect of smoking more than 100 cigarettes on the time spent on physical activites.

Figure 12 shows us that within the sample, the visual difference between the number of respondents who smoked more than 100 cigarettes (blue column) and respondents who smoked lesser than 100 cigarettes are not that big. Specifically, 57% of respondents have smoked less than 100 cigarettes while the remaining 43% of respondents smoked more than 100 cigarettes. As the classes in this variable is almost split half,

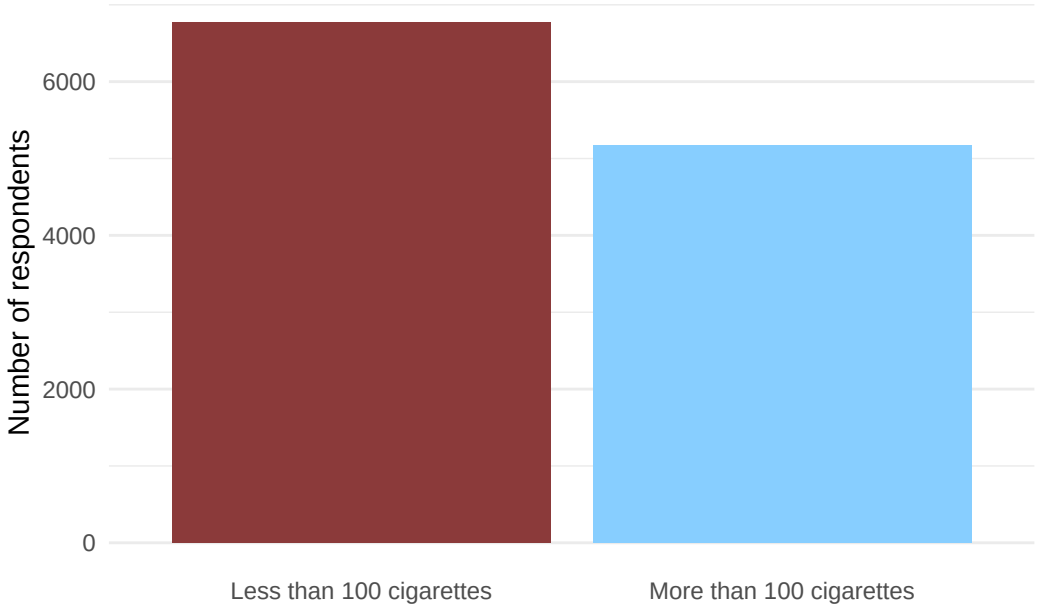

Figure 12: Barplot of total cigarettes smoked in a respondent's lifetime

Figure 13 shows us cigarette consumption by age bins among out sample population. As respondent age increases from the first bin (18 to 19 year old) to the fourth bin (30 to 34 year old), we can see the number of respondents who have smoked more than 100 cigarettes in their lifetime increased from less than 125 respondents to more than 500 respondents (about 3 times more respondents between the two age groups).

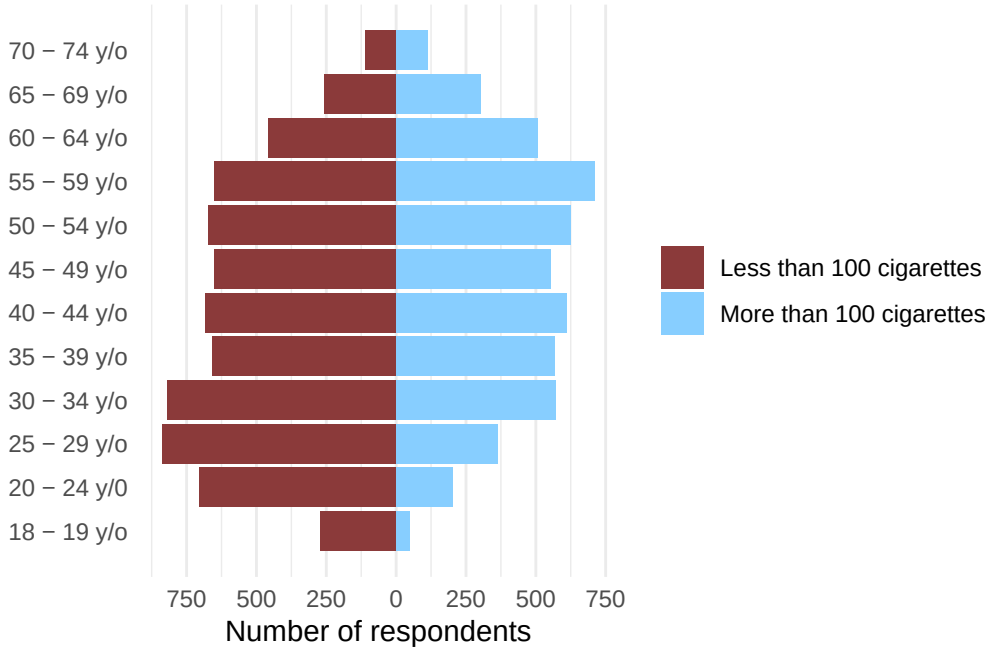

Figure 13: Pyramid of cigarettes consumption in respondent's lifetime by age groups

3 Model

The goal of our model is to explore which factors affect time spent on physical activities. Therefore a simple exploratory linear regression would be the chosen modelling algorithm to determine it. A full model where our response $time\_spent\_vigorous\_exercise\_7d$ would be fitted on all our interested variables outlined in Section 2.5.

3.1 Model set-up

Let k be the individual from CCHS dataset. Define y_k as the time spent on physical activities for individual k. Then Age_k, Sex_k, $Educ_k$ and Alc_k represents individual i's age, sex, highest educational attainment, and frequency of alcohol consumption the past 12 months respectively. Also, $Drug_k$ and $Cigar_k$ would be indicator explanatory variables derived from the binary variables in our dataset representing if individual k had consumed drugs in the past 12 months and if he/she has smoked 100 cigars in their lifetime respectively. Finally, $HealthRegion_k = 35953$ and $HealthRegion_k = 35970$ are also indicator explanatory variables derived from the categorical variable health_region in our dataset, which represents if individual k belongs to Health Region 35953 (Peel Regional Health Unit) and if he/she belongs to Health Region

35970 (York Regional Health Unit). Note that the reference with respect to Health Region would be individuals that are not in both health region 35953 and 35970.

Therefore the final model would be…

$$\ln(y_k) = \beta_0 + \beta_1 Age_k + \beta_2 Sex_k + \beta_3 Educ_k + \beta_4 Alc_k + \beta_5 I*(Drug_k) + \beta_6 I*(Cigar_k) + \\ \beta_7 I*(HealthRegion_k = 35953) + \beta_8 * I(HealthRegion_k = 35970) + \epsilon_k$$

$$\ln(y_k)|\mu_k, \sigma^2 \sim \text{Normal}(\mu_k, \sigma^2) \tag{1}$$

$$\begin{aligned} \mu_k = {} & \beta_0 + \beta_1 Age_k + \beta_2 Sex_k + \beta_3 Educ_k + \beta_4 Alc_k + \\ & \beta_5 I*(Drug_k) + \beta_6 I*(Cigar_k) + \\ & \beta_7 I*(HealthRegion_k = 35953) + \\ & \beta_8 * I(HealthRegion_k = 35970) \end{aligned} \tag{2}$$

$$\epsilon_k \sim \text{Normal}(0, \sigma^2) \tag{3}$$

We run the model in R (R Core Team 2023) using the `stats` package which is a base library offered by the R Core Team, specifically the `lm()` function to create our simple linear regression. The default parameters of the `lm()` function would be used.

3.1.1 Model justification

Initially, we fitted a full model containing every variable of interest. Since health region is a categorical variable with no implied order, To handle health region, we will be creating dummy variables for each health region, specifically $I*(HealthRegion_k = i)$ for $i \in health\_region$, and the reference chosen for health region would be the health_region = 35926 (The District of Algoma HU). We define $Income_k$ to be the the personal income of individual k in our dataset.

Then, for $8 \leq c \leq 40$ and $i \in health\_region$, our original full model was defined as…

$$\ln(y_k) = \beta_0 + \beta_1 Age_k + \beta_2 Sex_k + \beta_3 Educ_k + \beta_4 Income_k + \beta_5 Alc_k + \beta_6 I*(Drug_k) + \\ \beta_7 I*(Cigar_k) + \beta_c I*(HealthRegion = i) + \epsilon_k$$

We then conducted model reduction using backward feature elimination using Bayesian Information Criterion (BIC) as our scoring metric to obtain the least number of predictors that minimises the residual errors from the reduced model. This model reduction was performed

using `stats` package from the R Core Team (2023) as well. The resulting BIC value was obtained from the `lme4` package (Bates et al. 2015).

Table 5 shows us the BIC values for both the full model and the reduced model. After removing certain explanatory variables, the BIC score has decreased, indicating that the reduced model is has stronger explanatory power while having a less complex model. Therefore, the reduced model is chosen from the original full model.

Table 5: Table of BIC values for original and final model

Model	BIC value
Original full model	37082
Final reduced model	36841

4 Results

The final model with our estimated coefficients are as follows

$$\ln(y_k) = 5.84 - 0.03Age_k + 0.23Sex_k - 0.08Educ_k + 0.04Alc_k + 0.15I * (Drug_k) + 0.12I * (Cigar_k) + \beta_c I * (HealthRegion = i) + \epsilon_k$$

Overall, the individual t-test for our predictors were found to be below our set threshold of 95% alpha. The confidence interval for each predictor is listed in Table 6 shown below.

Table 6: Confidence Interval of linear model

Variable Name	2.5 %	97.5 %
(Intercept)	5.72	5.97
num_alc_drank_12m	0.03	0.05
age	-0.04	-0.02
sex	0.19	0.27
illicit_drug_use	0.10	0.21
highest_educational_attainment	-0.12	-0.05
smoked_hundred_cigarettes	0.08	0.17
health_region_35953	-0.33	-0.13
health_region_35970	-0.27	-0.08

Our results are summarized in Table 7.

Table 7: Linear model of time spent on physically taxing activities based on explanatory variables

	Log(Time_spent_on_physical_activity)
(Intercept)	5.842
	(0.063)
num_alc_drank_12m	0.038
	(0.006)
age	−0.033
	(0.004)
sex	0.232
	(0.021)
illicit_drug_use	0.155
	(0.027)
highest_educational_attainment	−0.087
	(0.019)
smoked_hundred_cigarettes	0.127
	(0.022)
health_region_35953	−0.233
	(0.051)
health_region_35970	−0.174
	(0.048)
Num.Obs.	11 941
R2	0.039
R2 Adj.	0.038
AIC	36 767.1
BIC	36 841.0
Log.Lik.	−18 373.547
RMSE	1.13

5 Discussion

5.1 Males are more likely to engage in physical activities

According to Figure 14, we can observe that amongst individuals that are active, men spend (on average) more time on physical activity than women. In particular, according to our sample men spend on average 639 minutes a week on physical activites while women spend on average 437 minutes a week on physical activities.

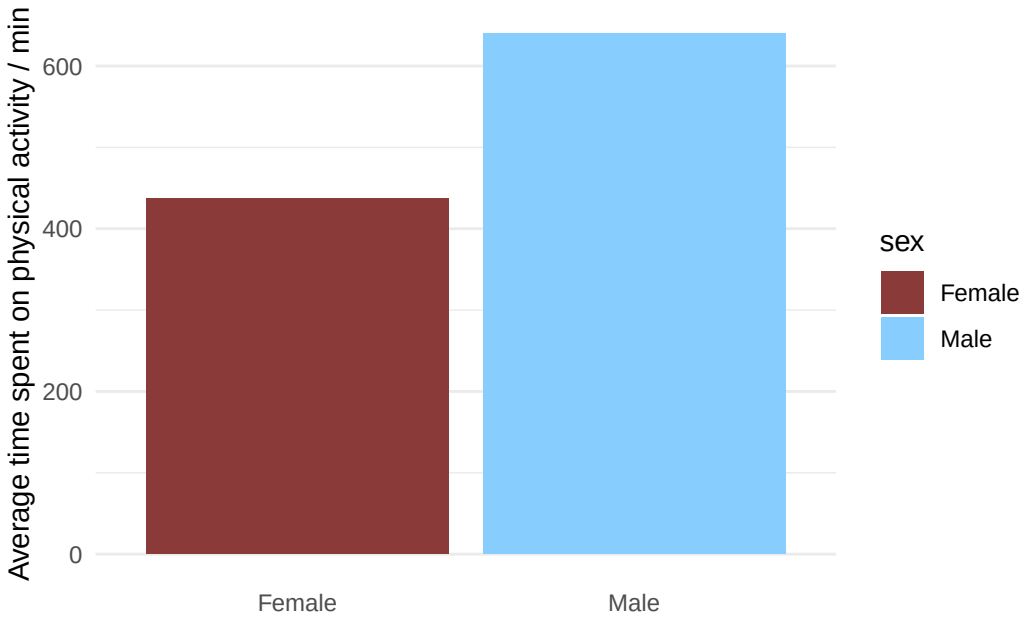

Figure 14: Bar plot of average time spent on physical activity by sex

For the explanatory variable `sex`, the corresponding coefficient from the fitted linear model is 0.232. This means that if an individual is a male, this increases the time spent on physical activities by 26%. While we were expecting increase in the overall engagement in physical activity among women as explained in Section 2.5.2, men are still more likely to engage in physical activities than women. In 2018, Global News Canada reported that 31% of women and 25% of men have insufficient physical activity (Global News, a division of Corus Entertainment Inc. Corus News. 2018). Therefore, our linear regression model discovery was not surprising.

In fact, in a study published in the Journal of American College of Cardiology, they discovered that women who were physically active would experience greater health benefits than men who were also physically active (Ji et al. 2024). Therefore, this means that more can be done to address the gender gap in the engagement of physical activities.

5.2 Ontario residents who have smoked before might be more likely to engage in physical activities.

From elementary school, Ontario students are exposed to health education where schools would educate about the harms of smoking. It is well known that smoking has negative effect's on an individual's health. Therefore it is unexpected that the linear regression model showed a positive coefficient for the indicator variable for 'having smoked 100 cigarettes' in an individual's lifetime.

The linear model is telling us that if an individual has smoked 100 cigarettes in their lifetime, they are 13% more likely to spend more time being physically active than an individual who has not smoked more than 100 cigarettes in the same time frame. While it may seem contradicting to what we believe, there may be merit to this value. From Figure 15, we can see that individuals who had smoked 100 cigarettes in their lifetime spent more time on physical activity than individuals who have not smoked 100 cigarettes. In particular, our data shows that on average, individuals who smoked 100 cigarettes spend 618 minutes on physical activities in a week while those who don't spend only 418 minute on physical activities.

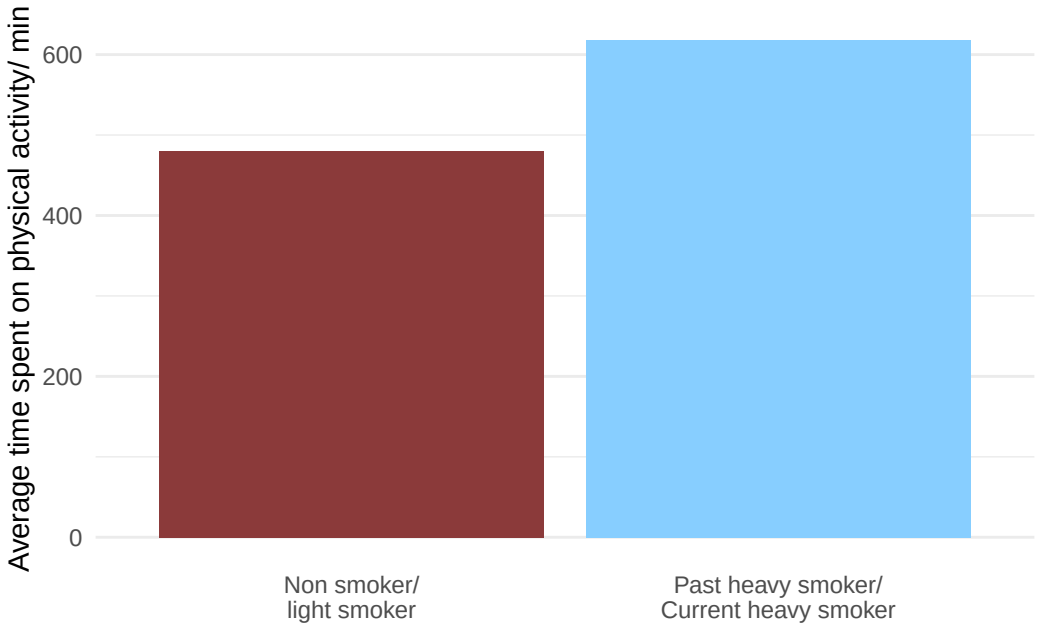

Figure 15: Bar plot of average time spent on physical activity by smoking status

A team from the school of Public Health Sciences at University of Waterloo had explored patterns and trends for tobbaco use in canada. What they found was that 70% of Canadians who had smoked before have quit as of 2020 Reid JL and Hammond D and Burkhalter R and

Rynard VL (2022). While smoking may have been affected an individual's physical capabilities, it could have also increased awareness regarding the importance of their health. Since most Canadians who had a history of smoking had quit smoking, it is very possible that they have greater health conscious. Coupled with increased government support for programs with the sole purpose of increasing physical activeness in Ontario (Public Health Agency of Canada 2024), it is no longer surprising why individuals who smoked 100 cigarettes in their lifetime are on average 13% more likely to engage in physical activities.

5.3 Possible increase in stimulant use

According to the model, the coefficient for illicit drug use is a positive value. This means that if an individual from Ontario has consumed drugs listed in Section 2.5.7, they would spend on average 17% more time on physical activities. This is also another surprising finding since we have also discussed in Section 2.5.7 that there is a negative correlation between drug use and physical activeness, but our findings for our Ontario population suggests a positive correlation. From Figure 16, we can see that individuals who consumed drugs listed in Section 2.5.7 in their lifetime spent more time on physical activity than individuals who have not consumed drugs. In particular, our data shows that on average, individuals are drug abusers spend 670 minutes on physical activities in a week while those who don't spend only 503 minute on physical activities.

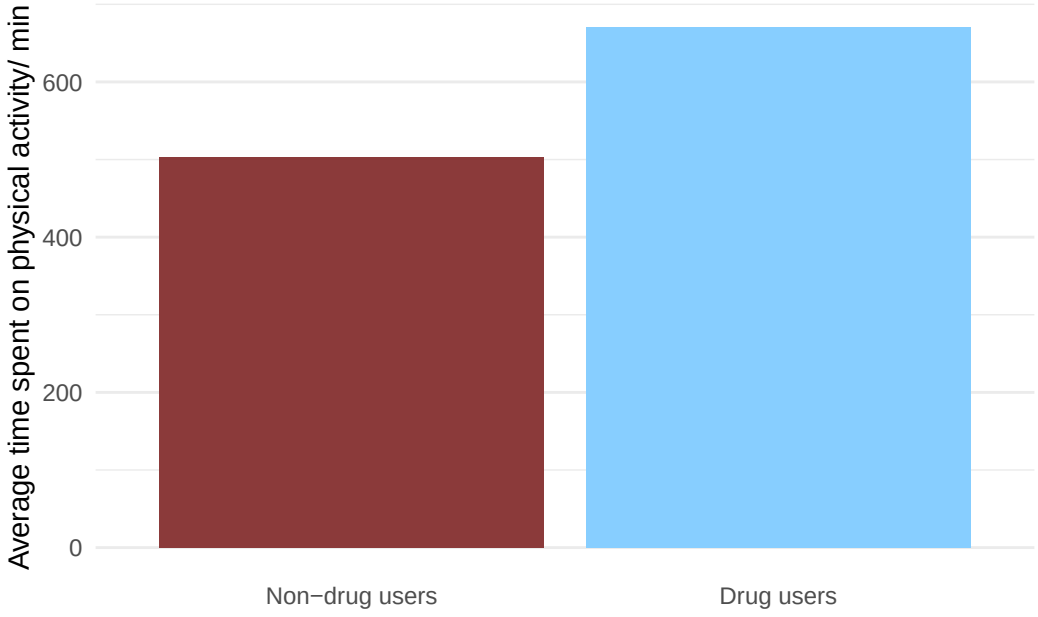

Figure 16: Bar plot of average time spent on physical activity by illicit drug use

Of the listed drugs, amphetamine and methamphetamine were identified as well. According to Drug Free Kids Canada (2024), they mentioned that the use of meth (short term for methamphetamine) leads to increased concentrations of dopamine, which is a mood enhancer. This is coupled with the increase of body movement. Likewise, in a study conducted by the Indiana University (Zaretsky et al. 2014), they found that the use of Amphetamine masks fatigue, allowing for individuals to exercise longer. This shows that it is possible that the inclusion of these drugs could have explained the surprising increase in time spent on physical activities despite the expectation that an individual who abuses drugs would have lower exercise.

The inclusion of these drugs in the survey could explain the positive coefficient seen from our model. However, this is just a hypothesis and an experiment would need to be conducted to explore if this finding is true, where we obtain a sample of drug users who consumed amphetamine and methamphetamine and drug users who did not consume the two drugs.

5.4 Weaknesses

Self-reported bias is a common limitation in health data collection, particularly when examining behaviors like physical activity, alcohol consumption, and smoking. Participants may overestimate or underestimate their activities due to memory recall issues, social desirability bias, or personal perceptions of what constitutes "healthy" behavior. For example, individuals might inflate their reported exercise levels or downplay unhealthy habits like smoking. Such inaccuracies can introduce systematic error, leading to biased estimates of associations and potentially undermining the validity of the study's conclusions. Mitigating this requires complementary objective measures, such as wearable activity trackers or biomarkers.

Appendix

A Diagnostic Checks

A.1 Motivation to transform response variable for linear regression model

Our response variable was log transformed to fulfill diagnostic checks for linear regression model. Observe the change in qq plot of the fitted linear regression model before our response ($time\_spent\_on\_vigorous\_exercise\_7d$) was log-transformed and after it was log transformed. From the left plot of Figure 17, we can see that values of standardised residuals are a lot larger than expected at higher quantile values. This suggests that the distribution of the residuals is severely right skewed. This contradicts the assumption of normality of residuals, which would make our linear regression model unreliable.

From the right plot of Figure 17, after we transform our response using the natural logarithm function (ln), we can see that most of the residuals lie on the y=x line in the qqplot, suggesting that the sampling distribution is approximately normal.

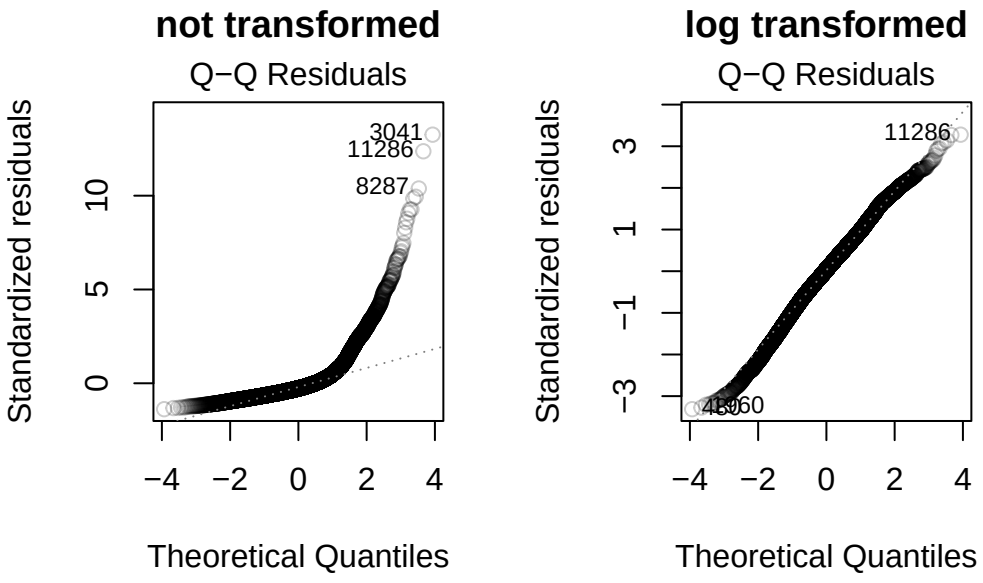

Figure 17: Diagnostic plot: QQ Plot

Now, we observe the change in the scale location plot of the fitted linear regression model before our response ($time\_spent\_on\_vigorous\_exercise\_7d$) was log-transformed and after it was

log transformed. From the left plot of Figure 18, we can see a weak but visible trend that as the fitted values increase, so does the square root of the standardized residuals. This suggests that the residuals are not randomly distributed, which goes against the assumption of linear regression that residuals are independent from each other. Furthermore, the mean red line is not horizontally flat. This is suggesting a positive trend. This tells us that the variance of the residuals is increasing, therefore implying that our residuals do not satisfy homoscedasticity assumption of linear regression, which would lead to an unreliable linear regression model.

From the right plot of Figure 18, we can see that the standardised residuals are randomly scattered across the plot across all fitted values. This suggests that residuals are independant from each other. Furthermore, the mean red line is (almost) horizontally flat, for which suggests that the variance of residuals across all fitted values are constant. Therefore, homoscedasticity assumption for linear regression is also satisfied, allowing us to interpret the coefficient estimates of our fitted linear model with greater reliability.

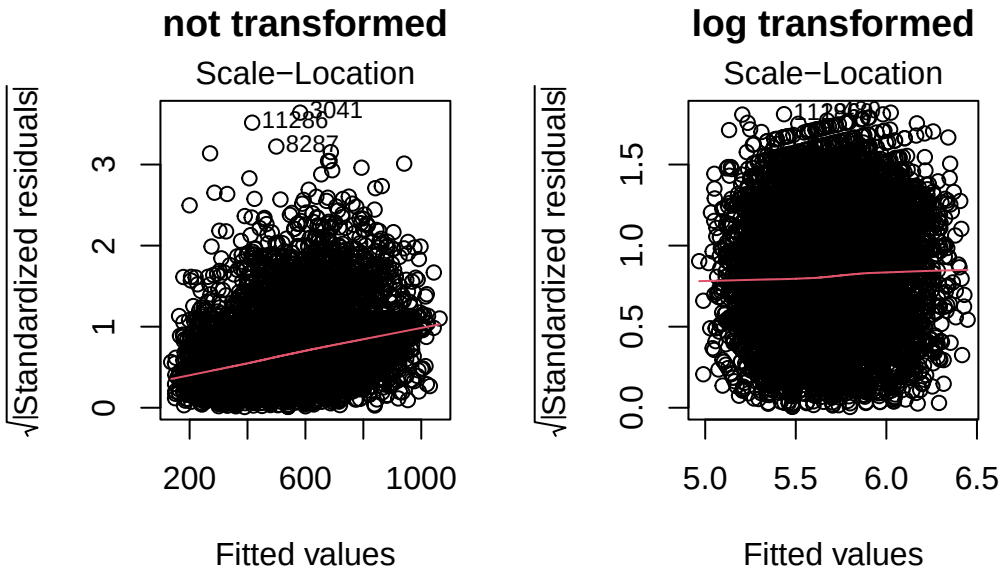

Figure 18: Diagnostic plot: Scale location

By performing log transformation of my response, we have satisfied the linear regression assumptions, allowing us to interpret from a reliable linear regression model, therefore explaining the need to transform our response variable in Section 3. This would also allow us to perform model reduction outlined in Section 3.1.1

A.2 Diagnostic check for final linear model after BIC

The final linear model was obtained after performing backward feature elimination using BIC score as the criterion. In this section, we will be showing that the final model satisfies linear regression assumptions.

Figure 19 shows us that the values of standardised residuals are approximately equal to the theoretical quantiles, which suggests that the residuals are normally distributed.

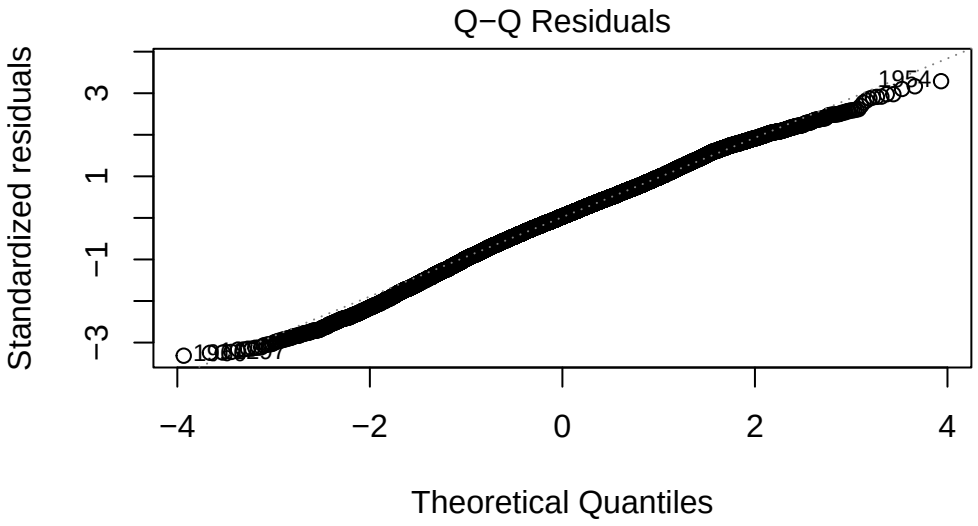

Figure 19: Diagnostic plot: QQ Plot

From Figure 20, we can see that the standardised residuals are randomly scattered across the scale location plot, which suggests that residuals are independant from each other. Furthermore the mean red line is (almost) horizontal which suggests that variance of residuals are constant across all fitted values. This means that the residuals obtained from the final model satisfies homoscedasticity assumption in linear regression

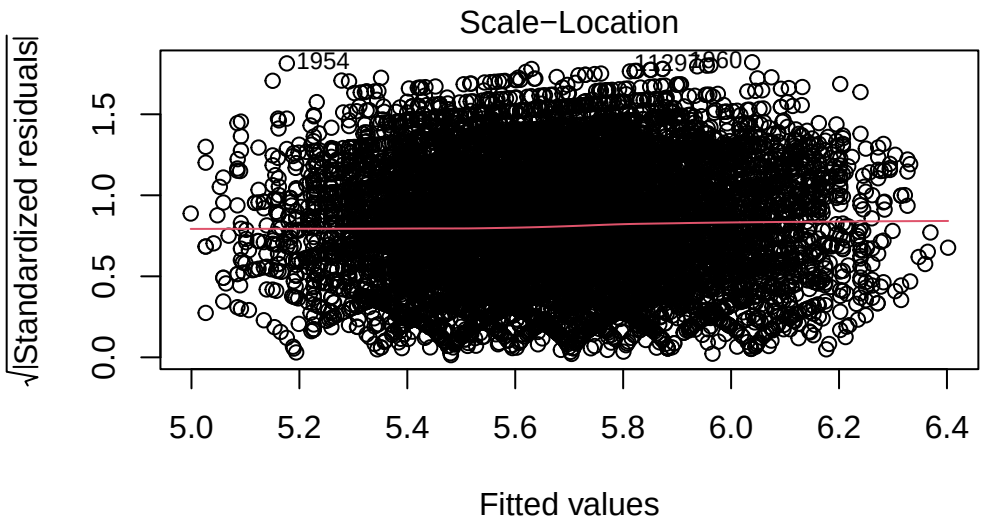

Figure 20: Diagnostic plot: Scale location

Overall, since the residual assumptions in linear regression are satisfied, we are able to use the coefficient estimates obtained from our final fitted linear model to gain insights into relationship between our predictors and the response as discussed in Section 5.

B Survey: Canadian Community Health Survey (CCHS) Annual component

B.1 CCHS Background

The purpose of CCHS is to provide cross sectional estimates of factors related to different health determinants in Canada. The use case for the Canadian health dataset is non-exhaustive and many have used it for the betterment of Canada. In the department of sociology and social work, University of Toronto, researchers have used the weight and weight perception data from the 2017-2018 CCHS to predict adolescents mental health, and they were able to find out that normal weighted girls are not immune from poor mental health arising from poor weight perceptions, hence urging specialists in the mental health field to increase support for these girls that have not been receiving enough avenues (Chai and Xue 2022). The Canadian Partnership Against Cancer also uses CCHS data to access the severity of the cervical cancer among Canadians, streamlining their cancer screening process as a result and urging the country to focus resources to address inequalities in access to cervical cancer prevention and treatment options between indigenous communities and poor regions (Canadian Partnership Against Cancer 2024b, 2024a). The one thing that ties many of the health-related studies arising from the use of CCHS studies is to address underrepresented communities in Canada.

With the purpose of unifying data across health sources, the Statistics of Canada had began conducting surveys yearly since 2007. From 2015, there was a redesign to their survey methodology that is used for the dataset that was used in this paper.

B.2 CCHS Target Population

Their target population covers individuals aged 12 and above living in all ten provinces of Canada. The target population excludes the following…

- individuals living on reserves, more commonly known as indigenous populations with their own governance within Canada ()
- full time servicemen in the Canadian Armed Forces
- youths aged 12 to 17 residing in foster homes
- individuals residing in jail, nursing homes, mental health institutions and shelters
- individuals in quebec health region of Nunavik and Terres-Cries-de-la-Baie-James

According to CCHS, they account for lesser than 3% of the population

B.3 CCHS Sampling Frame

The sampling frame for the Canadian Community Health Survey (CCHS) consists of two distinct components based on the target population. For the adult population (18 years and older), the sampling frame is derived from an area frame used by the Canadian Labour Force Survey (LFS). This area frame includes all primary sampling units (PSUs), which are geographic clusters of dwellings within health regions across Canada. For the youth population

(12 to 17 years old), the sampling frame is based on a list derived from the Canadian Child Benefit (CCB) files, which includes a list of all program beneficiaries with their names, addresses, and phone numbers.

B.4 CCHS Sample

The sample for the CCHS is drawn from the respective sampling frames for adults and youth. Approximately 130,000 individuals are selected for the survey, with around 120,000 respondents allocated to the adult population and 10,000 to the youth population. The sample is stratified by health region, ensuring that each health region has a minimum of 500 respondents. The sample for adults is selected from the area frame using a multi-stage sampling process, while the youth sample is directly selected from the CCB list. The allocation for provinces and territories is based on population size, with a focus on achieving representativeness within each health region.

B.5 CCHS Data Collection

Sample Recruitment for the 2018 Canadian Community Health Survey (CCHS) The sample for the 2018 CCHS was recruited through a multi-stage, stratified process designed to obtain a representative sample of the Canadian population, aged 12 years and older. The sampling process was divided into two main categories based on the target age groups: individuals aged 18 and older and youths aged 12 to 17.

The sample was drawn from two distinct frames:

- Area Frame: This frame was used to select households for respondents aged 18 and older. The area frame covers the entire Canadian population and is designed to represent geographic areas across the country.
- Canadian Child Benefit (CCB) Frame: This frame was used to select households with youths aged 12 to 17. The CCB frame was specifically used to identify youth in households receiving the Canadian Child Benefit.

The data collection for the CCHS was carried out using two primary methods:

- Computer-Assisted Telephone Interviewing (CATI): This method was used for approximately 75% of respondents selected from the area frame and for all respondents selected from the CCB frame (aged 12-17). Interviews were conducted over the phone by trained interviewers in centralized call centers.
- Computer-Assisted Personal Interviewing (CAPI): This method was used for about 25% of the area frame respondents. These interviews were conducted face-to-face by decentralized field interviewers using laptops. Interviewers primarily visited households in person, though telephone follow-ups were permitted if necessary.

The sampling process was designed to maximize coverage and minimize non-response. Each regional office of Statistics Canada was responsible for assigning cases to interviewers, with careful consideration of the geographic and demographic distribution of respondents. The sample was stratified by health region to ensure representation from each geographic area. Interviewers were assigned cases based on their regional collection office, and the assignments typically included no more than 15 cases per interviewer to ensure manageable workloads. These assignments were planned to be completed in four non-overlapping three-month collection periods, with different phases for initial and follow-up attempts.

To optimize the contact process, the interviewers first initiated contact with the selected households by either telephone or personal visit, depending on the sampling frame and mode of data collection. If initial contact was unsuccessful, additional efforts were made to reach the household through scheduled callbacks, including multiple attempts at different times and days. If necessary, personal visits were conducted to obtain the interview.

For youths aged 12-17 selected for interviews, interviewers were required to obtain verbal consent from a parent or guardian before interviewing youths aged 12-14. If a parent or guardian requested to see the questionnaire, they were provided with a copy before the interview. In cases where the selected youth was unable to complete the interview, a "Person Most Knowledgeable" (PMK) within the household was interviewed instead. This procedure ensured that household-level information was collected even if the selected youth was unable to provide it.

If the selected respondent was unable to participate due to physical or mental health conditions, a proxy interview was conducted. A knowledgeable household member would provide information on behalf of the selected respondent. However, the more personal and sensitive questions were omitted from these interviews, as the proxy could not reliably answer them.

Throughout the recruitment process, several quality control measures were in place to ensure the accuracy and validity of the data. A random selection of interviews was validated to ensure the integrity of the data. These validations involved contacting households to confirm that the interview took place and checking the quality of the responses. Cumulative reports were generated at the end of each collection period, showing response rates by region, to help regional offices identify areas that required additional follow-up. Specific sections of interviews were recorded for quality control purposes and analyzed to ensure proper procedures were followed by interviewers.

B.6 CCHS Nonresponse and Nonresponse handling

The Canadian Community Health Survey (CCHS) employs a systematic approach to address nonresponse at both the household and person levels. This ensures that the survey results remain representative of the general population, even when certain individuals or households do not respond. The nonresponse handling methodology is as follows:

1. Household Nonresponse

Nonresponse at the household level occurs when a sampled household either refuses to participate, is unreachable, or provides unusable data. To adjust for nonresponse, the survey sample is divided into response homogeneity groups (RHGs), which group households based on their likelihood of responding. This helps ensure that the adjustment for nonresponse is made based on similar response propensities. A logistic regression model is used to predict the likelihood of a household responding to the survey. Based on the predicted response probabilities, an adjustment factor is computed within each RHG. This factor helps correct for nonresponse by redistributing the weight of nonresponding households to those that did respond. The final adjustment is made by multiplying the weight of responding households by the adjustment factor, resulting in a new weight. Nonresponding households are excluded from the weighting process after this adjustment.

2. Person-Level Nonresponse

The interviewer first collects a complete roster of household members. A selected individual from the roster is then interviewed. In cases where the selected individual cannot be contacted or refuses to participate, person-level nonresponse occurs. Similar to household nonresponse, an adjustment is applied to the weights of responding individuals to account for those who did not respond. This ensures that the survey remains representative of the entire population. A logistic regression model is applied to estimate the likelihood that a selected individual will respond to the survey. This model uses information gathered during the first part of the interview (e.g., age, sex, etc.) and the location of the household. The selected individuals are then grouped into response homogeneity groups (RHGs) based on their calculated response probabilities. Within each RHG, an adjustment factor is calculated to correct for person-level.

By including factors such as geographic location and household characteristics in the nonresponse adjustment models, CCHS enhances the likelihood that the adjusted sample is representative of diverse populations across different regions and demographic groups. Also, the distinct handling of household nonresponse (via weight adjustment A6) and person nonresponse (via weight adjustment A9) ensures that both levels of nonresponse are addressed separately, which helps maintain the accuracy of estimates at both the household and individual levels.

However, the use of multiple weighting adjustments, while intended to improve representativeness, can sometimes lead to overfitting or over-adjustment. This may result in a dataset that over-represents certain characteristics at the expense of others, especially if the logistic regression models are not perfectly specified or if important variables are omitted. Furthermore, even with sophisticated adjustments, response bias — such as social desirability bias or recall bias — can still affect the dataset. Respondents may provide inaccurate or socially desirable answers, and the weighting adjustments cannot fully correct for these biases, especially if they are not uniform across all demographic groups.

References

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.

Bauman, Adrian E, Rodrigo S Reis, James F Sallis, Jonathan C Wells, Ruth JF Loos, and Brian W Martin. 2012. "Correlates of Physical Activity: Why Are Some People Physically Active and Others Not?" *The Lancet* 380 (9838): 258–71. https://doi.org/https://doi.org/10.1016/S0140-6736(12)60735-1.

Brellenthin, Angelique G, and Duck-chul Lee. 2018. "Physical Activity and the Development of Substance Use Disorders: Current Knowledge and Future Directions." *Prog Prev Med (N Y)* 3 (3): e0018. https://doi.org/10.1097/pp9.0000000000000018.

Canadian Mental Health Association. 2024. *Substance Use and Addiction*. https://ontario.cmha.ca/addiction-and-substance-use-and-addiction/#_edn5.

Canadian Partnership Against Cancer. 2024a. *Data Tables for the Eliminating Cervical Cancer in Canada Digital Report*. https://www.partnershipagainstcancer.ca/ecc-data/.

———. 2024b. *Eliminating Cervical Cancer in Canada*. https://www.partnershipagainstcancer.ca/topics/eliminating-cervical-cancer/.

Chai, Lei, and Jia Xue. 2022. "Weight, Weight Perceptions, and Health and Well-Being Among Canadian Adolescents: Evidence from the 2017-2018 Canadian Community Health Survey." *American Journal of Health Promotion* 36 (1): 55–63. https://doi.org/10.1177/08901171211031064.

Data Centre, Faculty of Arts & Science, University of Toronto. 2018. *Canadian Community Health Survey (CCHS)*. https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/index.html/sda.htm.

Drug Free Kids Canada. 2024. *Effects & Risks of Meth*. https://www.drugfreekidscanada.org/drug-spotlights/meth/effects-risks-of-meth/.

Global News, a division of Corus Entertainment Inc. Corus News. 2018. *Nearly One Third of Canadian Women Don't Get Enough Physical Activity — Worse Than Men*. https://globalnews.ca/news/4428978/women-physical-activity-gender-gap/.

Heydari, Gholamreza, Mostafa Hosseini, Mahmoud Yousefifard, Hadi Asady, Masoud Baikpour, and Atena Barat. 2015. "Smoking and Physical Activity in Healthy Adults: A Cross-Sectional Study in Tehran." *Prog Prev Med (N Y)* 14 (4): 238–45.

Ji, Hongwei, Martha Gulati, Tzu Yu Huang, Alan C. Kwan, David Ouyang, Joseph E. Ebinger, Kaitlin Casaletto, Kerrie L. Moreau, Hicham Skali, and Susan Cheng. 2024. "Sex Differences in Association of Physical Activity with All-Cause and Cardiovascular Mortality." *Journal of the American College of Cardiology* 83 (8): 783–93. https://doi.org/https://doi.org/10.1016/j.jacc.2023.12.019.

Kari, Jaana T, Jutta Viinikainen, Petri Böckerman, Tuija H Tammelin, Niina Pitkänen, Terho Lehtimäki, Katja Pahkala, Mirja Hirvensalo, Olli T Raitakari, and Jaakko Pehkonen. 2020. "Education Leads to a More Physically Active Lifestyle: Evidence Based on Mendelian Randomization." *Scand J Med Sci Sports* 30 (7): 1194–204. https://doi.org/10.1111/sms.

13653.

ParticipACTION. 2024. *Key Statistics*. https://www.participaction.com/the-science/key-facts-and-stats/.

Public Health Agency of Canada. 2024. *Government of Canada Invests in Programming to Promote Physical Activity and Healthy Living*. https://www.canada.ca/en/public-health/news/2024/06/government-of-canada-invests-in-programming-to-promote-physical-activity-and-healthy-living.html.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reid JL and Hammond D and Burkhalter R and Rynard VL. 2022. *Tobacco Use in Canada: Patterns and Trends*. https://uwaterloo.ca/tobacco-use-canada/tobacco-use-canada-patterns-and-trends.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Statista Research Department. 2024. *Number of Students Enrolled in Postsecondary Institutions in Canada from 2000 to 2022*. https://www.statista.com/statistics/447739/enrollment-of-postsecondary-students-in-canada/.

Statistics Canada. 2018. *Canadian Community Health Survey - Annual Component (CCHS)*. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=329241.

———. 2020. *Table 2: Median Wages, Salaries and Commissions, by Province and Territory*. https://www150.statcan.gc.ca/n1/daily-quotidien/200218/t002c-eng.htm.

———. 2023. *Canada's Population Estimates: Record-High Population Growth in 2022*. https://www150.statcan.gc.ca/n1/daily-quotidien/230322/dq230322f-eng.htm.

Tennekes, Martijn. 2018. "tmap: Thematic Maps in R." *Journal of Statistical Software* 84 (6): 1–39. https://doi.org/10.18637/jss.v084.i06.

The Daily, Statistics Canada. 2000. *Population Estimates*. https://www150.statcan.gc.ca/n1/daily-quotidien/000926/dq000926a-eng.htm.

Warburton, Darren E. R., Crystal Whitney Nicol, and Shannon S. D. Bredin. 2006. "Health Benefits of Physical Activity: The Evidence." *CMAJ* 174 (6): 801–9. https://doi.org/10.1503/cmaj.051351.

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.

World Health Organisation. 2024. *Physical Activity*. https://www.who.int/news-room/fact-sheets/detail/physical-activity.

World Health Organization. 2024. *Canada, Health Data Overview for Canada*. https://data.who.int/countries/124.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

Zaretsky, Dmitry V, Mary Beth Brown, Maria V Zaretskaia, Pamela J Durant, and Daniel E Rusyniak. 2014. "The Ergogenic Effect of Amphetamine." *Temperature (Austin)* 1 (3):

242–47. https://doi.org/10.4161/23328940.2014.987564.