My title*

My subtitle if needed

Chris Yong Hong Sen

November 24, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph The estimand 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 regular exercise.

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

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,

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

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 exercise, 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 exercise among individuals in Ontario who engage in regular exercise. 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 vigorous exercise 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 exercise, surveyors sectioned vigorous exercise 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 exercise 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 vigorous exercise 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 vigorous exercise, 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 exercising amongst individuals that are engaged in exercising.

2.4 Response variable: Time spent on vigorous exercise 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 vigorous 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 vigorous sweating routine 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 vigorous 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 vigorous exercise in the past 7 days

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

Interestingly, the maximum time spent on vigorous 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 which would be discussed in section (INSERT SECTION HERE PLEASE DONT FORGET) which would address these outlier values.

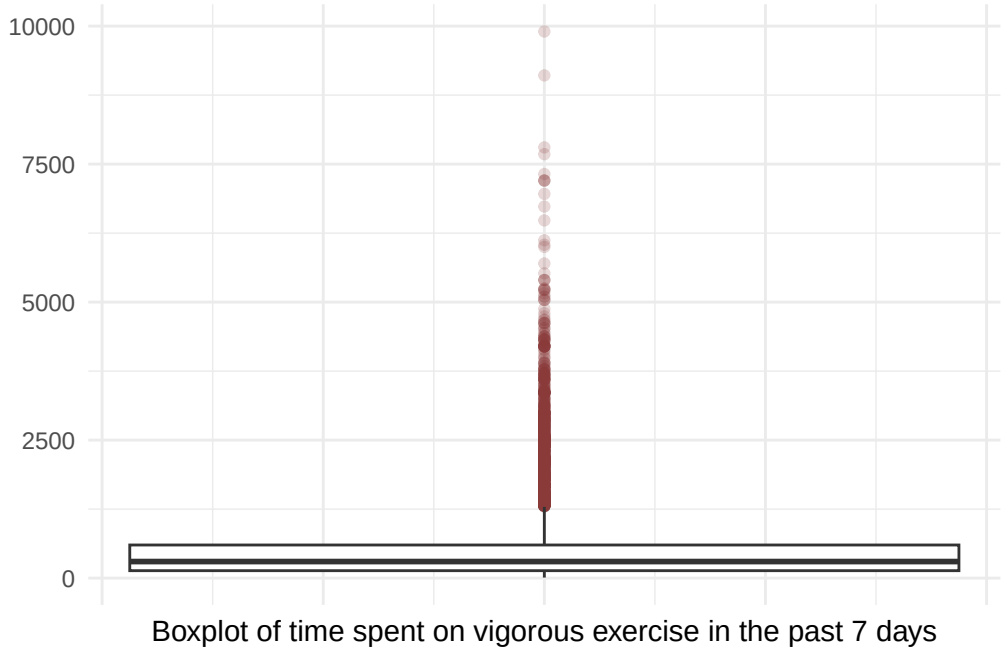

Figure 1: Boxplot of time spent on vigorous exercise in the past 7 days

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 vigorous exercise.

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.

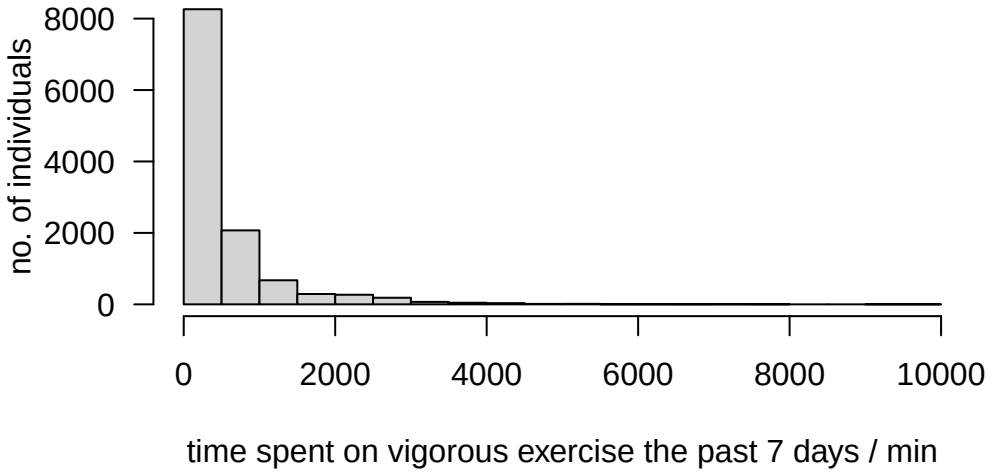

Figure 2: Histogram of time spent on vigorous exercise the past 7 days

2.5 Interested explanatory variables

There are a total of 13 predictors of interest that is seemingly related to the time spent in vigorous exercise an exercising individual gets.

2.5.1 Explanatory 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).

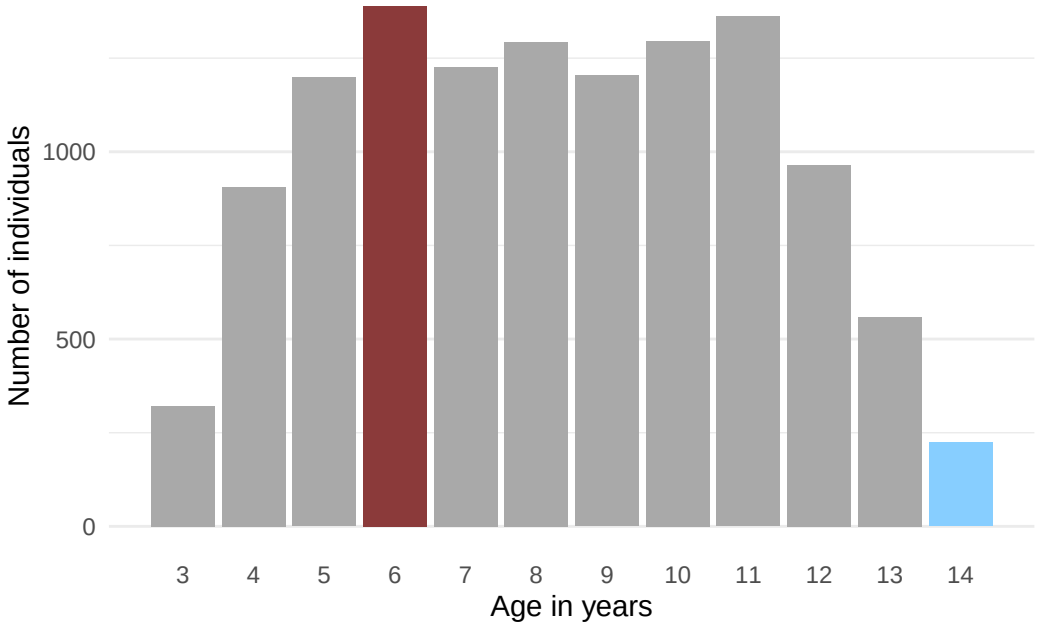

Figure 3: Histogram of time spent on vigorous exercise the past 7 days

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

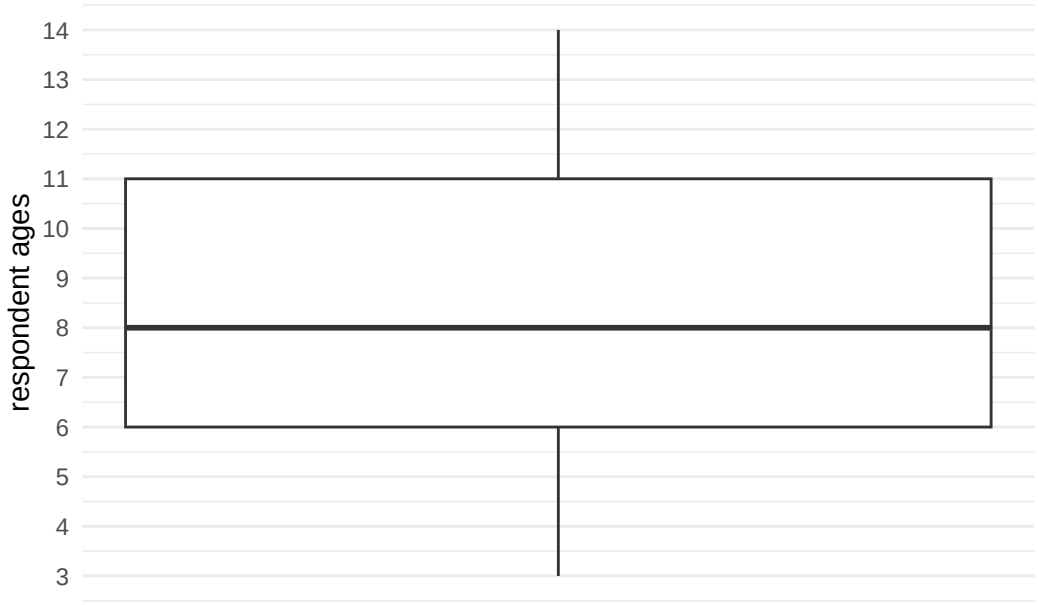

Figure 4: Boxplot of respondent ages

2.5.2 Explanatory Variable 2: Sex

Sex is a possible influential predictor for total time spent on vigorous 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 vigorous 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.

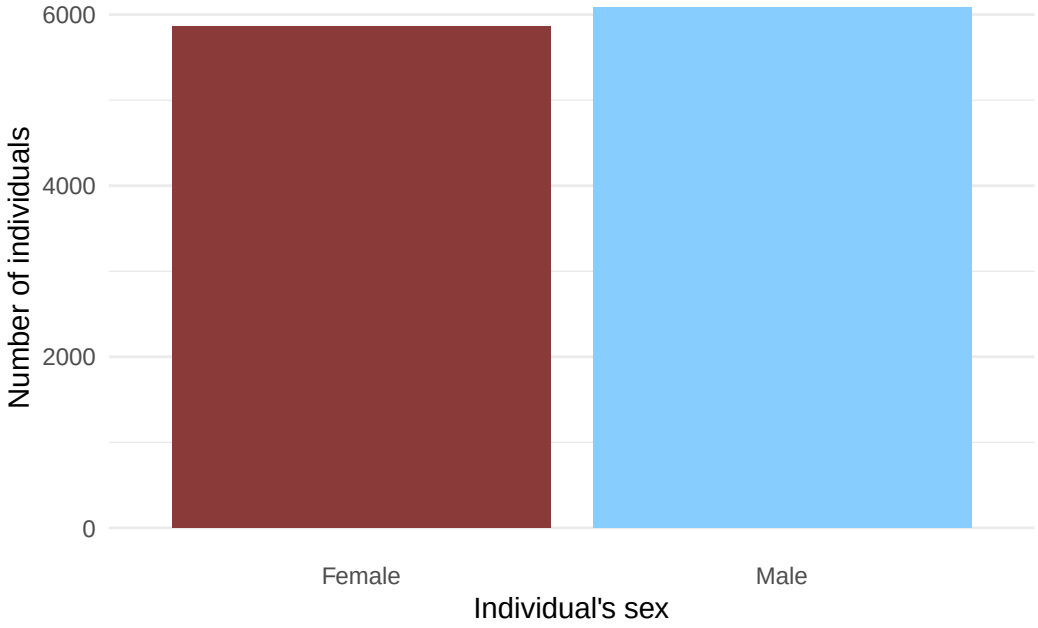

Figure 5: Barplot of sex

2.5.3 Explanatory 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 vigorous 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, we are immediately alerted to a class imbalance in our dataset.

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%
university degree and above	72%

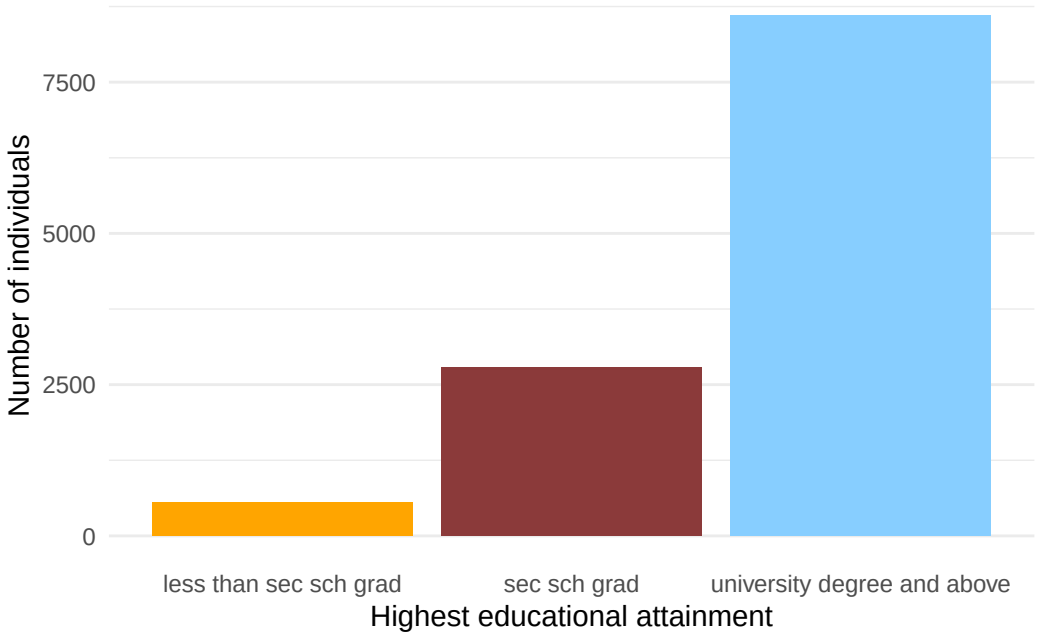

Figure 6: Barplot of highest_educational_attainment

2.6 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (Figure 7), from Horst, Hill, and Gorman (2020).

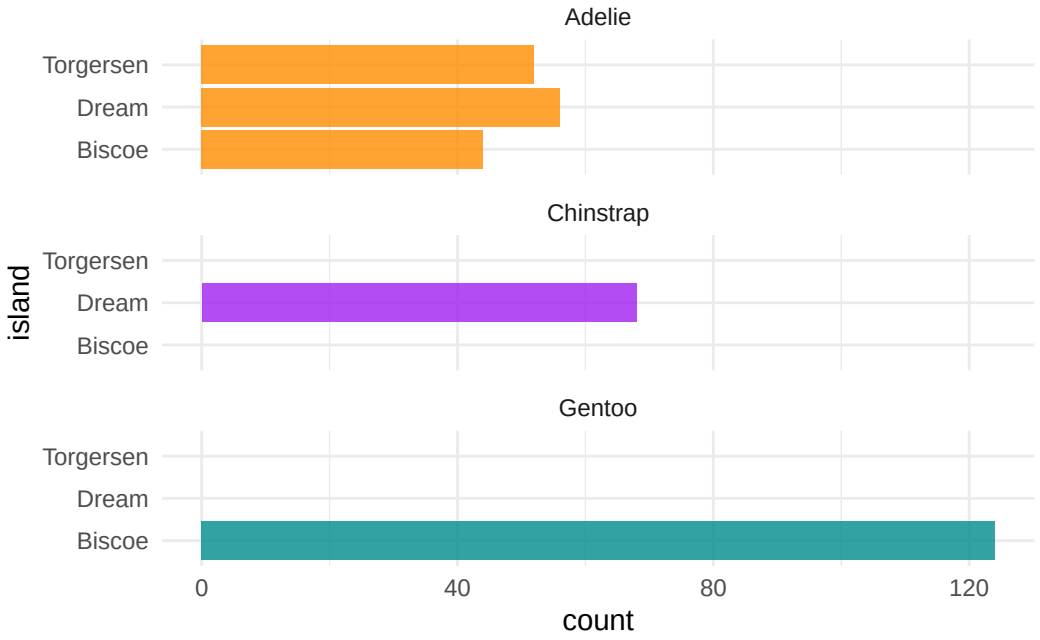

Figure 7: Bills of penguins

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.7 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 4.

5 Discussion

5.1 First discussion point

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

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

Table 4: Explanatory models of flight time based on wing width and wing length

	Final model
(Intercept)	6.302
	(0.143)
num_alc_drank_12m	0.038
	(0.006)
age	−0.088
	(0.016)
sex	0.230
	(0.021)
illicit_drug_use	0.152
	(0.027)
highest_educational_attainment	−0.267
	(0.054)
smoked_hundred_cigarettes	0.128
	(0.022)
health_region_35953	−0.232
	(0.051)
health_region_35970	−0.178
	(0.048)
age × highest_educational_attainment	0.021
	(0.006)
Num.Obs.	11 941
R2	0.040
R2 Adj.	0.039
AIC	36 756.2
BIC	36 837.4
Log.Lik.	−18 367.090
RMSE	1.13

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

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.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

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/.

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

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

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.