My title*

My subtitle if needed

Jiaxuan Song

November 25, 2024

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

1 Introduction

Depression is a global mental health challenge that significantly affects the quality of life and economic productivity. According to the World Health Organization (WHO), over 280 million people worldwide experience depression, making it a leading cause of mental disability (World Health Organization 2024). The multifactorial nature of depression—spanning biological, psychological, social, and environmental dimensions—makes it complex to study and address effectively. (Stringaris 2017) Understanding the key factors influencing depression is critical for developing targeted interventions and policies to mitigate its burden. Depression is understood to result from a combination of factors across biological, psychological, social, and environmental dimensions (**epine2011increasing?**). For instance, genetic studies indicate that certain genes may increase an individual's susceptibility to depression, while sociological research highlights the importance of environmental factors, such as poverty, unemployment, or a lack of social support (Addis, Truax, and Jacobson 1995). However, most existing studies focus on isolated variables, overlooking the intricate interplay between these factors. Furthermore, much of the current research is limited to Western contexts, leaving significant gaps in understanding depression across different cultural settings (Chentsova-Dutton and Tsai 2009).

Although extensive research has been conducted on depression, most studies focus on individual factors such as socioeconomic status, genetics, or social support, without fully exploring the interactions between these variables. Furthermore, much of the existing literature is region-specific, leaving gaps in understanding cross-cultural or generalized patterns. In the Canadian context, while research on depression has increased, there is still a lack of systematic analysis of specific social and economic factors contributing to depression. This project aims to address

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

this gap by analyzing data from the Canadian Community Health Survey (CCHS) for the years 2019 and 2020. (Statistics Canada 2024) The dataset includes various potential predictors of depression, such as age, gender, income, employment status, social support, and health behaviors. By applying statistical techniques, this study seeks to identify the most critical factors influencing depression risk and their relative contributions, with a specific focus on trends and disparities evident during this period in Canada.

Results paragraph

Why it matters paragraph

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

2 Data

2.1 Overview

The dataset used in this study was sourced from the public use microdata file (PUMF) of the Canadian Community Health Survey (CCHS), which provides comprehensive data for health regions and combinations of health regions across Canada. Covering a two-year period, the dataset includes interviews with approximately 130,000 respondents aged 12 or older residing in households across all provinces and territories (Statistics Canada 2024).

The dataset contains information on a wide range of topics, including physical activity, height and weight, smoking, exposure to second-hand smoke, alcohol consumption, general health, chronic health conditions, injuries, and use of healthcare services. Additionally, it offers insights into socio-demographic, income, and labour force characteristics of the population. For this study, the dataset underwent preprocessing to focus specifically on variables relevant to depression analysis, such as socio-demographic factors, health behaviors, and chronic health conditions. These records were retained to form the basis for the analysis, enabling a detailed examination of factors influencing depression in Canada.

In this project, we employ `R` (R Core Team 2023) as the primary statistical programming language, integrating several key packages to enhance analysis and visualization. These include the `tidyverse` (**tidyverse?**) suite for data manipulation and cleaning, `palmerpenguins` (Horst, Hill, and Gorman 2020) for illustrative data examples, and `knitr` (Xie 2014) for dynamic report generation. Additionally, we use `rstanarm` (Goodrich et al. 2022) for linear regression modeling, `modelsummary` (Arel-Bundock 2023) for streamlined model summaries, `janitor` (Firke 2023) for data cleaning utilities, `arrow` (Richardson and Developers 2023) for efficient data access and processing, and `ggplot2` (Wickham 2016) for advanced data visualization.

2.2 Measurement

In mt analysis of depression and its associated factors in Canada, I translate complex health, behavioral, and socio-demographic data into a structured dataset. The dataset, derived from the 2019–2020 Canadian Community Health Survey (CCHS), provides a nationally representative sample of respondents aged 12 and older, enabling us to analyze depression severity and its potential predictors. While variables such as depression severity score (DEPDVPHQ) and classification of depression levels (DEPDVSEV) simplify the nuanced experiences of respondents, they serve as practical outcome measures, allowing us to examine patterns of mental health across diverse populations.

To ensure consistency and accuracy, we implemented a systematic data-cleaning process, including handling missing values, removing duplicates, and verifying variable formats. For example, we standardized key variables such as marital status (DHHGMS), categorizing respondents as either "Married/Common-law" or "Other," and grouped age (DHHGAGE) into meaningful ranges (e.g., "12-17," "18-34"). Weekly alcohol consumption (ALWDVWKY) was retained as a measure of behavioral health, while household income (INCDGHH) and education level (EHG2DVH3) provided socio-economic context. By focusing on these relevant variables, we streamlined the dataset to emphasize the factors most likely to influence depression outcomes.

To specifically examine depression, we utilized two key variables: the depression severity score (DEPDVPHQ) and the classification of depression levels (DEPDVSEV), enabling a detailed analysis of both continuous and categorical measures of depression. Data standardization efforts included recoding missing or ambiguous information to ensure clarity. This systematic preparation provides a robust foundation for exploring relationships between predictors and depression severity, allowing us to identify key factors influencing mental health outcomes in Canada and develop actionable insights for intervention and policy.

2.3 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 1), from Horst, Hill, and Gorman (2020).

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.

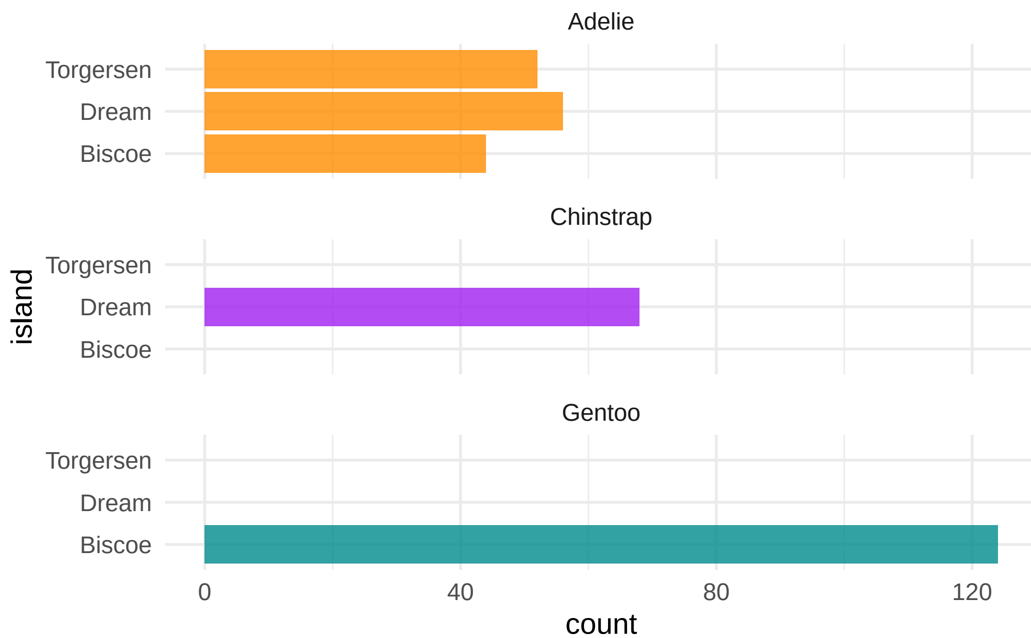


Figure 1: Bills of penguins

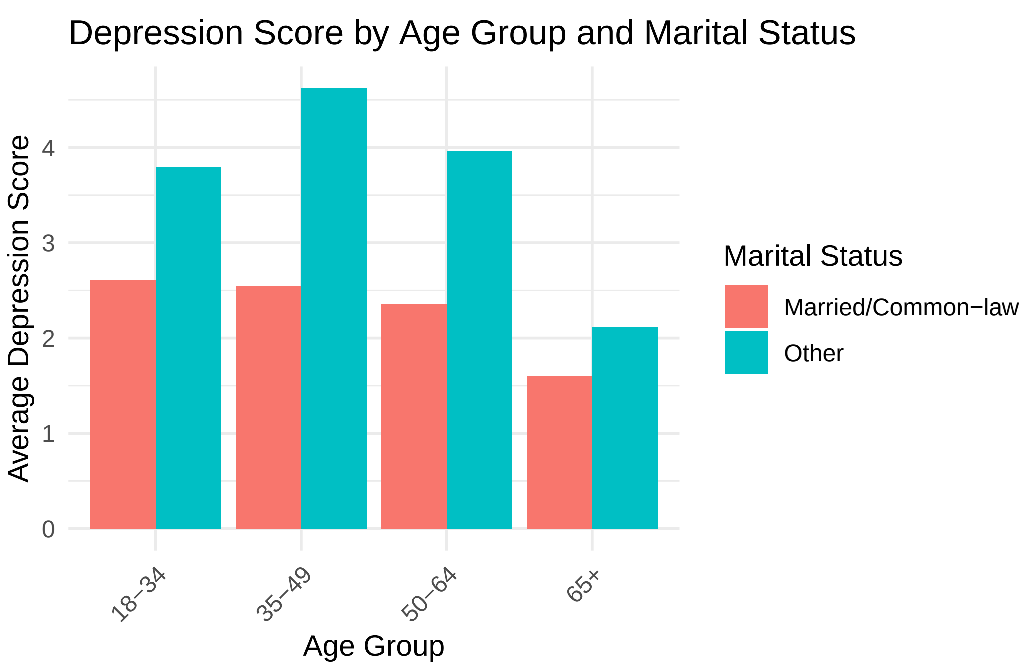
2.4 Predictor variables

```
library(ggplot2)
library(dplyr)

# Assuming the dataset is named 'data'
# Select and clean relevant columns
cleaned_data <- data %>%
  select(DHHGAGE, DEPDVPHQ, DHHGMS) %>%
  drop_na() %>%
  rename(
    `Age Group` = DHHGAGE,
    `Depression Score` = DEPDVPHQ,
    `Marital Status` = DHHGMS
  ) %>%
  mutate(
    `Marital Status` = case_when(
      `Marital Status` == 1 ~ "Married/Common-law",
      `Marital Status` == 2 ~ "Other",
      TRUE ~ "Unknown"
    ),
```

```r
    `Age Group` = case_when(
      `Age Group` == 1 ~ "12-17",
      `Age Group` == 2 ~ "18-34",
      `Age Group` == 3 ~ "35-49",
      `Age Group` == 4 ~ "50-64",
      `Age Group` == 5 ~ "65+",
      TRUE ~ as.character(`Age Group`)
    )
  )

# Create the grouped bar chart
ggplot(cleaned_data, aes(x = `Age Group`, y = `Depression Score`, fill = `Marital Status`)) +
  stat_summary(fun = mean, geom = "bar", position = "dodge") +
  labs(
    title = "Depression Score by Age Group and Marital Status",
    x = "Age Group",
    y = "Average Depression Score",
    fill = "Marital Status"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



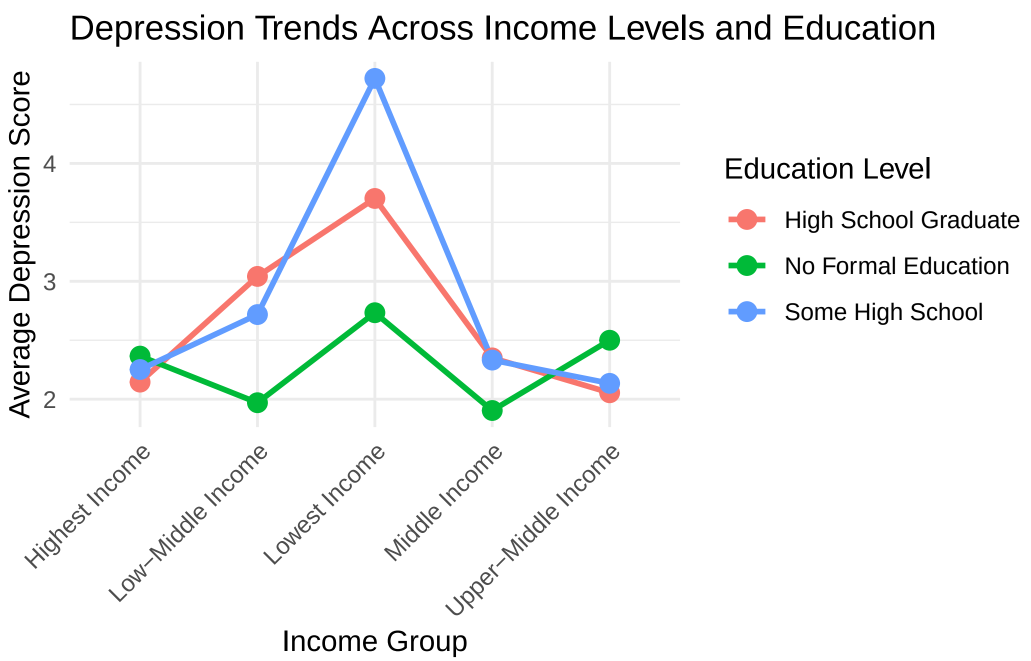
```r
library(ggplot2)
library(dplyr)

# Assuming the dataset is named 'data'
# Clean and process the data
cleaned_data <- data %>%
  select(DEPDVPHQ, INCDGHH, EHG2DVH3) %>%
  drop_na() %>%
  rename(
    `Depression Score` = DEPDVPHQ,
    `Income Group` = INCDGHH,
    `Education Level` = EHG2DVH3
  ) %>%
  mutate(
    `Income Group` = case_when(
      `Income Group` == 1 ~ "Lowest Income",
      `Income Group` == 2 ~ "Low-Middle Income",
      `Income Group` == 3 ~ "Middle Income",
      `Income Group` == 4 ~ "Upper-Middle Income",
      `Income Group` == 5 ~ "Highest Income",
      TRUE ~ "Unknown"
    ),
    `Education Level` = case_when(
      `Education Level` == 1 ~ "No Formal Education",
      `Education Level` == 2 ~ "Some High School",
      `Education Level` == 3 ~ "High School Graduate",
      `Education Level` == 4 ~ "Some College",
      `Education Level` == 5 ~ "College Graduate",
      `Education Level` == 6 ~ "Postgraduate",
      TRUE ~ "Unknown"
    )
  ) %>%
  group_by(`Income Group`, `Education Level`) %>%
  summarise(Average_Depression_Score = mean(`Depression Score`, na.rm = TRUE)) %>%
  arrange(match(`Income Group`, c(
    "Lowest Income", "Low-Middle Income", "Middle Income",
    "Upper-Middle Income", "Highest Income"
  )))
```

```
`summarise()` has grouped output by 'Income Group'. You can override using the
`.groups` argument.
```

```r
# Create the line graph with Education Level as a color factor
ggplot(cleaned_data, aes(x = `Income Group`, y = Average_Depression_Score, group = `Educatio
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(
    title = "Depression Trends Across Income Levels and Education",
    x = "Income Group",
    y = "Average Depression Score",
    color = "Education Level"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



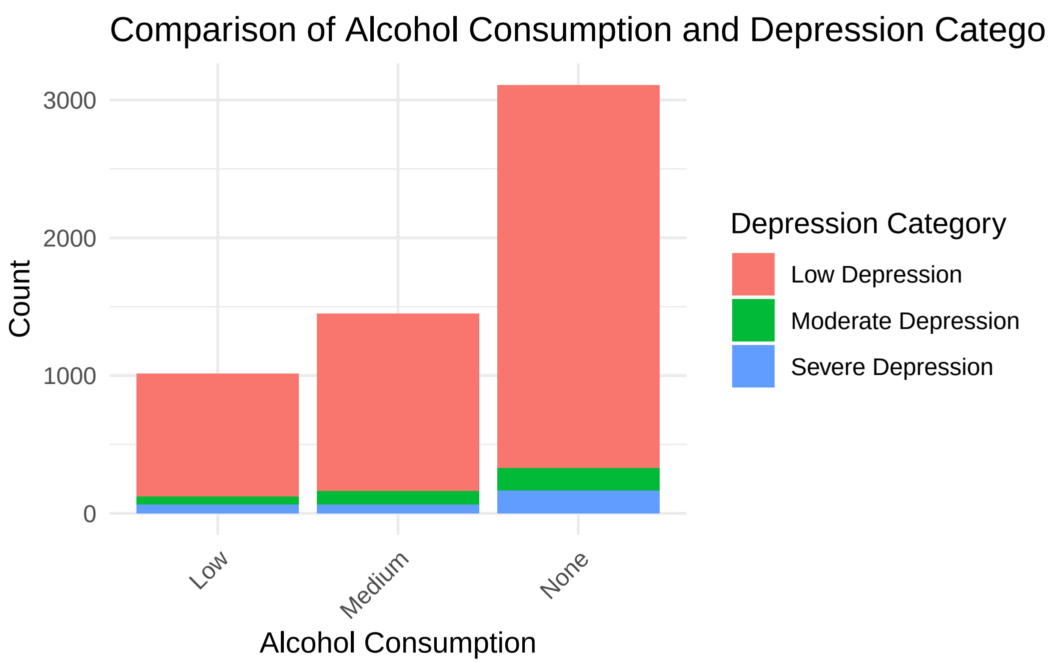

```r
library(ggplot2)
library(dplyr)

# Assuming the dataset is named 'data'
# Select and clean relevant columns for alcohol consumption and depression
cleaned_data <- data %>%
```

```r
  select(ALCDVTTM, DEPDVPHQ) %>%
  drop_na() %>%
  rename(
    `Alcohol Consumption` = ALCDVTTM,
    `Depression Score` = DEPDVPHQ
  ) %>%
  mutate(
    `Alcohol Consumption` = case_when(
      `Alcohol Consumption` == 1 ~ "None",
      `Alcohol Consumption` == 2 ~ "Low",
      `Alcohol Consumption` == 3 ~ "Medium",
      `Alcohol Consumption` == 4 ~ "High",
      TRUE ~ as.character(`Alcohol Consumption`)
    ),
    `Depression Category` = case_when(
      `Depression Score` <= 5 ~ "Low Depression",
      `Depression Score` > 5 & `Depression Score` <= 10 ~ "Moderate Depression",
      `Depression Score` > 10 ~ "Severe Depression",
      TRUE ~ as.character(`Depression Score`)
    )
  )

# Summarize the data to count occurrences of each alcohol consumption and depression categor
alcohol_depression_summary <- cleaned_data %>%
  count(`Alcohol Consumption`, `Depression Category`) %>%
  mutate(percentage = n / sum(n) * 100)

# Create the bar chart comparing alcohol consumption and depression categories
ggplot(alcohol_depression_summary, aes(x = `Alcohol Consumption`, y = n, fill = `Depression C
  geom_bar(stat = "identity", position = "stack") +  # Stacked bar chart
  labs(
    title = "Comparison of Alcohol Consumption and Depression Categories",
    x = "Alcohol Consumption",
    y = "Count",
    fill = "Depression Category"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



2.5 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 algorithm

References

Addis, Michael E, Paula Truax, and Neil S Jacobson. 1995. "Why Do People Think They Are Depressed?: The Reasons for Depression Questionnaire." *Psychotherapy: Theory, Research, Practice, Training* 32 (3): 476.

Arel-Bundock, Vincent. 2023. *Modelsummary: Beautiful and Customizable Model Summaries in r*. https://vincentarelbundock.github.io/modelsummary/.

Chentsova-Dutton, Yulia E, and Jeanne L Tsai. 2009. "Understanding Depression Across Cultures." *Handbook of Depression* 2: 363–85.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://sfirke.github.io/janitor/.

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.

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

Richardson, Neal, and Apache Arrow Developers. 2023. *Arrow: Interoperable Data Frames for r and Python with Apache Arrow*. https://arrow.apache.org/.

Statistics Canada. 2024. "Canadian Community Health Survey (CCHS) - 2024: Public Use Microdata File (PUMF)." https://www150.statcan.gc.ca/n1/pub/82m0013x/82m0013x2024001-eng.htm.

Stringaris, Argyris. 2017. "What Is Depression?" *Journal of Child Psychology and Psychiatry*. Wiley Online Library.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. 2nd ed. Springer-Verlag New York. https://ggplot2.tidyverse.org.

World Health Organization. 2024. "Depression." https://www.who.int/news-room/fact-sheets/detail/depression.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in r." *Implementing Reproducible Computational Research*.