Number of COVID-19 Cases in Canada By Region, Age, and Sex*

Predicting the Factors that Influence the Number of COVID-19 Cases and Deaths

Bruce Zhang

November 23, 2024

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

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

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

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to analyze the data and to create graphs and models. Th packages that were used include tidyverse (Wickham, Averick, et al. 2023), tidyr (Wickham, Henry, and Vaughan 2023), dplyr (Wickham, François, et al. 2023), and rstanarm (Gabry et al. 2023). Our data (Open Science Framework (OSF) 2022) was obtained from COVerAGE-DB. The data anlysis was conducted based on the guidance of Alexander (2023).

*Code and data are available at: https://github.com/brucejczhang/covid_data.

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

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 (**palmerpenguins?**).

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

Region	Sex	Mean Age	Total Cases	Total Deaths	Mean CFR (%)
Rural	f	50.05	6763540	156814	4.69
Rural	m	49.83	6869820	163430	5.37
Urban	f	50.02	6769033	165648	5.18
Urban	m	50.12	6630777	160191	5.44

Figure 1: 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.

2.4 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

A logistic model has been selected as part of this analysis to examine the relationship between region, sex, age, and the number of cases and deaths of COVID-19.

The logistic regression model predicts the probability of a region or individual being classified as "high risk" based on key predictors. In this case, "high risk" is defined as regions where the death rate (CFR) exceeds the median death rate across the dataset.

The CFR is calculated as:

$$\text{CFR} = \frac{\text{Deaths}}{\text{Cases} + 10^{-6}} \times 100 \tag{1}$$

The **box-and-whisker plot** visualizes the distribution of predicted probabilities ((p)) across different **age intervals**. Each box represents the interquartile range (IQR) of predicted probabilities, with the line inside the box indicating the median predicted probability for that age interval. Whiskers extend to 1.5 times the IQR, and any points outside this range are considered outliers. This visualization highlights the variability of predicted probabilities within each age group.

The logistic regression model equation is:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Tests} + \beta_2 \cdot \text{Age Interval} + \epsilon \tag{2}$$

where:

- p: Predicted probability of being high risk (p = Predicted Probability).
- $\text{logit}(p)$: Log-odds of the high-risk classification.
- β_0: Intercept of the model.
- β_1, β_2: Coefficients for the predictors (e.g., number of tests, age interval).
- ϵ: Error term.

This model explains how testing rates and age intervals affect the likelihood of a region or demographic being classified as high risk. The box-and-whisker plot shows the variability of predicted probabilities across age groups, helping to identify specific age intervals with consistently higher or lower risks.

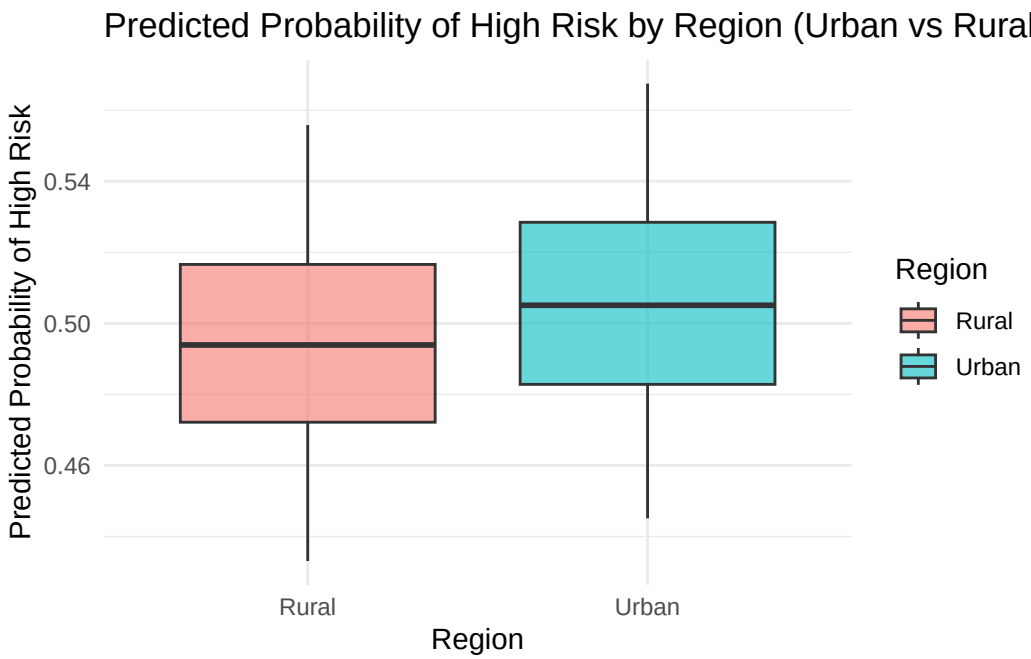
We run the model in R (R Core Team 2023) using the `rstanarm` package of Gabry et al. (2023). We use the default priors from `rstanarm`.

```
# Load necessary libraries
library(dplyr)
library(ggplot2)
set.seed(21)

# Step 1: Filter for Urban and Rural Regions Only
filtered_data <- analysis_data %>%
  mutate(
    Death_Rate = Deaths / (Cases + 1e-6), # Calculate death rate (avoid division by zero)
    High_Risk = ifelse(Death_Rate > median(Death_Rate, na.rm = TRUE), 1, 0) # Classify high-r
  )

# Step 2: Logistic Regression Model
# Fit a logistic regression model
logistic_model <- glm(
  High_Risk ~ Tests + Region + Age,
  family = binomial(),
  data = filtered_data
)

# Summary of the model
summary(logistic_model)
```

```
Call:
glm(formula = High_Risk ~ Tests + Region + Age, family = binomial(),
    data = filtered_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.571e-02  1.076e-01   0.796  0.42580
Tests       -7.161e-06  2.738e-06  -2.615  0.00892 **
RegionUrban  5.081e-02  7.839e-02   0.648  0.51690
Age          1.407e-03  1.238e-03   1.136  0.25576
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3622.4  on 2612  degrees of freedom
Residual deviance: 3613.9  on 2609  degrees of freedom
AIC: 3621.9

Number of Fisher Scoring iterations: 3
```

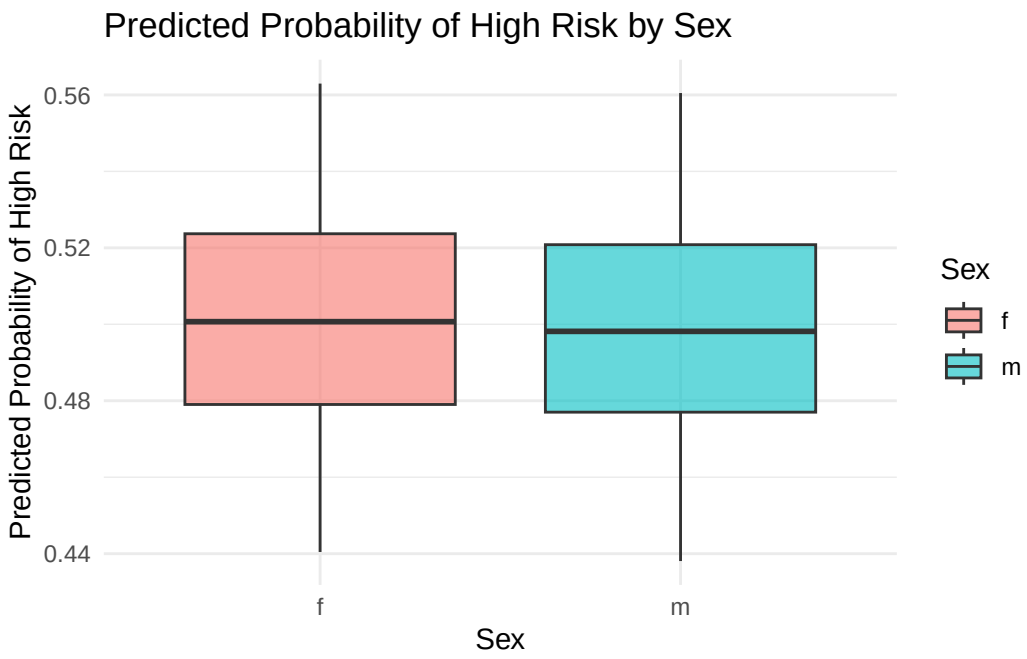
```r
# Step 3: Add Predictions to the Dataset
filtered_data <- filtered_data %>%
  mutate(Predicted_Probability = predict(logistic_model, type = "response"))

# Step 4: Visualize the Results
# Plot predicted probabilities by region
ggplot(filtered_data, aes(x = Region, y = Predicted_Probability, fill = Region)) +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Predicted Probability of High Risk by Region (Urban vs Rural)",
    x = "Region",
    y = "Predicted Probability of High Risk"
  ) +
  theme_minimal()
```




```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
set.seed(21)

# Step 1: Filter for only 'f' and 'm', and define High Risk
filtered_data <- analysis_data %>%
  mutate(
    Death_Rate = Deaths / (Cases + 1e-6), # Calculate death rate (avoid division by zero)
    High_Risk = ifelse(Death_Rate > median(Death_Rate, na.rm = TRUE), 1, 0) # Classify high-r
  )

# Step 2: Logistic Regression Model
# Fit a logistic regression model
logistic_model <- glm(
  High_Risk ~ Tests + Sex + Age,
  family = binomial(),
  data = filtered_data
)

# Summary of the model
summary(logistic_model)
```

```
Call:
glm(formula = High_Risk ~ Tests + Sex + Age, family = binomial(),
    data = filtered_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.161e-01  1.082e-01   1.073  0.28328
Tests       -7.132e-06  2.737e-06  -2.605  0.00918 **
Sexm        -1.157e-02  7.837e-02  -0.148  0.88265
Age          1.409e-03  1.238e-03   1.138  0.25531
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3622.4  on 2612  degrees of freedom
Residual deviance: 3614.3  on 2609  degrees of freedom
AIC: 3622.3

Number of Fisher Scoring iterations: 3
```

```r
# Step 3: Add Predictions to the Dataset
filtered_data <- filtered_data %>%
  mutate(Predicted_Probability = predict(logistic_model, type = "response"))

# Step 4: Visualize the Results
# Plot predicted probabilities by sex
ggplot(filtered_data, aes(x = Sex, y = Predicted_Probability, fill = Sex)) +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Predicted Probability of High Risk by Sex",
    x = "Sex",
    y = "Predicted Probability of High Risk"
  ) +
  theme_minimal()
```



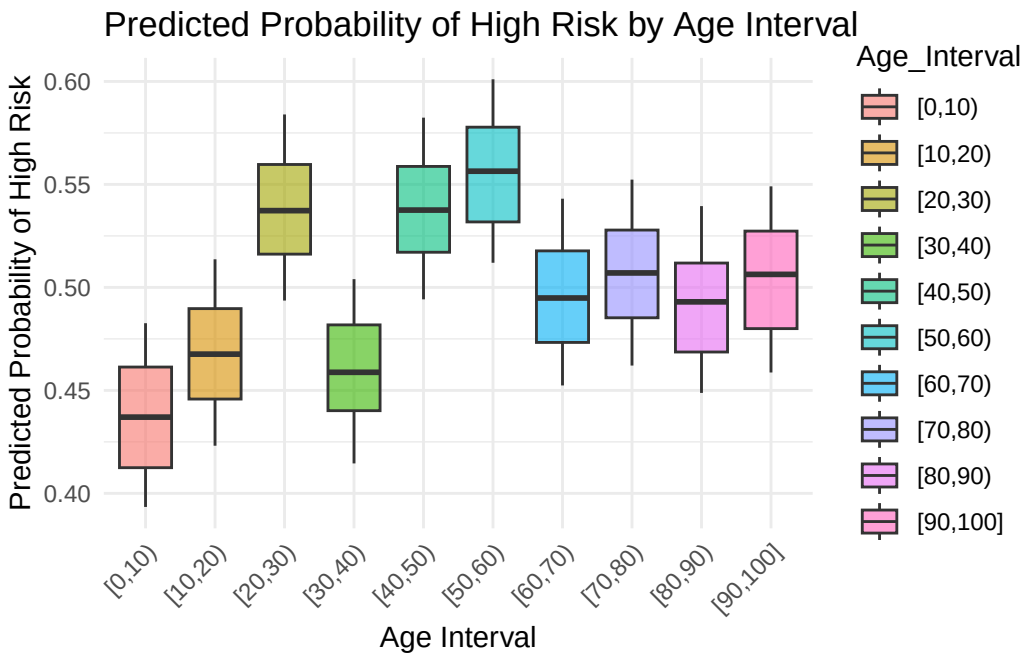

```r
# Load necessary libraries
library(dplyr)
library(ggplot2)
set.seed(21)

# Step 1: Create Age Intervals and Define High Risk
filtered_data <- analysis_data %>%
  mutate(
    Age_Interval = cut(Age, breaks = seq(0, 100, by = 10), right = FALSE, include.lowest = TR
    Death_Rate = Deaths / (Cases + 1e-6), # Calculate death rate (avoid division by zero)
    High_Risk = ifelse(Death_Rate > median(Death_Rate, na.rm = TRUE), 1, 0) # Classify high-r
  ) %>%
  filter(!is.na(Age_Interval)) # Remove any rows with missing age intervals

# Step 2: Logistic Regression Model
# Fit a logistic regression model
logistic_model <- glm(
  High_Risk ~ Tests + Age_Interval,
  family = binomial(),
  data = filtered_data
)

# Summary of the model
```

```r
summary(logistic_model)
```

```
Call:
glm(formula = High_Risk ~ Tests + Age_Interval, family = binomial(),
    data = filtered_data)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -6.803e-02  1.481e-01  -0.459  0.64592
Tests               -7.324e-06  2.746e-06  -2.667  0.00765 **
Age_Interval[10,20)  1.235e-01  1.847e-01   0.669  0.50363
Age_Interval[20,30)  4.080e-01  1.842e-01   2.215  0.02676 *
Age_Interval[30,40)  8.492e-02  1.848e-01   0.459  0.64595
Age_Interval[40,50)  4.056e-01  1.850e-01   2.193  0.02833 *
Age_Interval[50,60)  4.817e-01  1.853e-01   2.599  0.00934 **
Age_Interval[60,70)  2.431e-01  1.844e-01   1.319  0.18733
Age_Interval[70,80)  2.818e-01  1.845e-01   1.527  0.12675
Age_Interval[80,90)  2.277e-01  1.844e-01   1.235  0.21675
Age_Interval[90,100] 2.672e-01  1.599e-01   1.671  0.09464 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3622.4  on 2612  degrees of freedom
Residual deviance: 3603.2  on 2602  degrees of freedom
AIC: 3625.2

Number of Fisher Scoring iterations: 3
```

```r
# Step 3: Add Predictions to the Dataset
filtered_data <- filtered_data %>%
  mutate(Predicted_Probability = predict(logistic_model, type = "response"))

# Step 4: Visualize the Results
# Plot predicted probabilities by age interval
ggplot(filtered_data, aes(x = Age_Interval, y = Predicted_Probability, fill = Age_Interval)) +
  geom_boxplot(alpha = 0.6) +
  labs(
    title = "Predicted Probability of High Risk by Age Interval",
```

```
    x = "Age Interval",
    y = "Predicted Probability of High Risk"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # Rotate x-axis labels for readabi
```



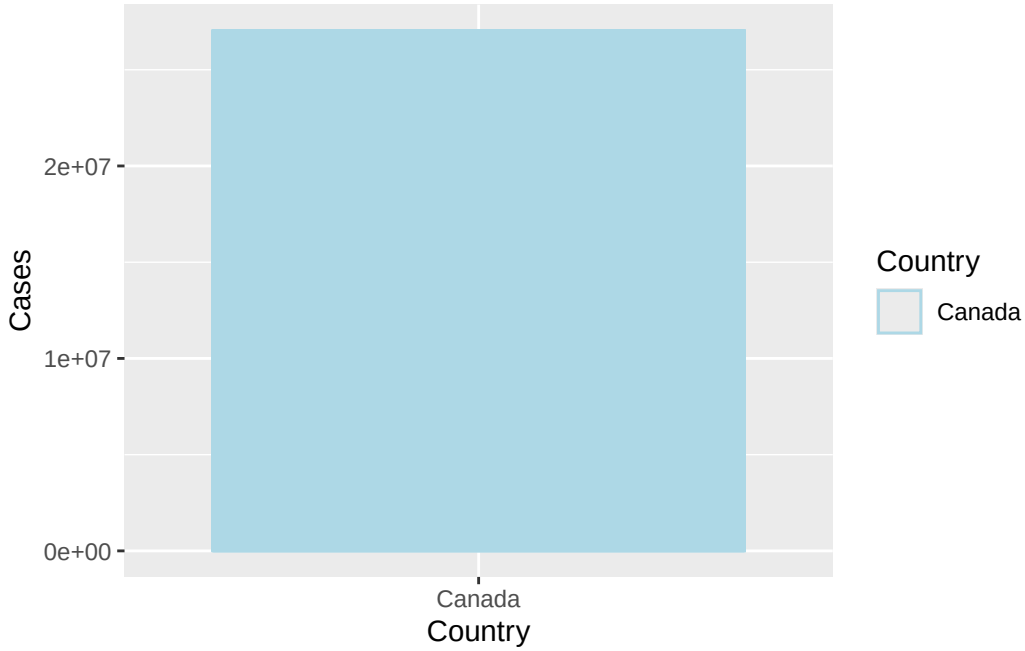
3.1.1 Model justification

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

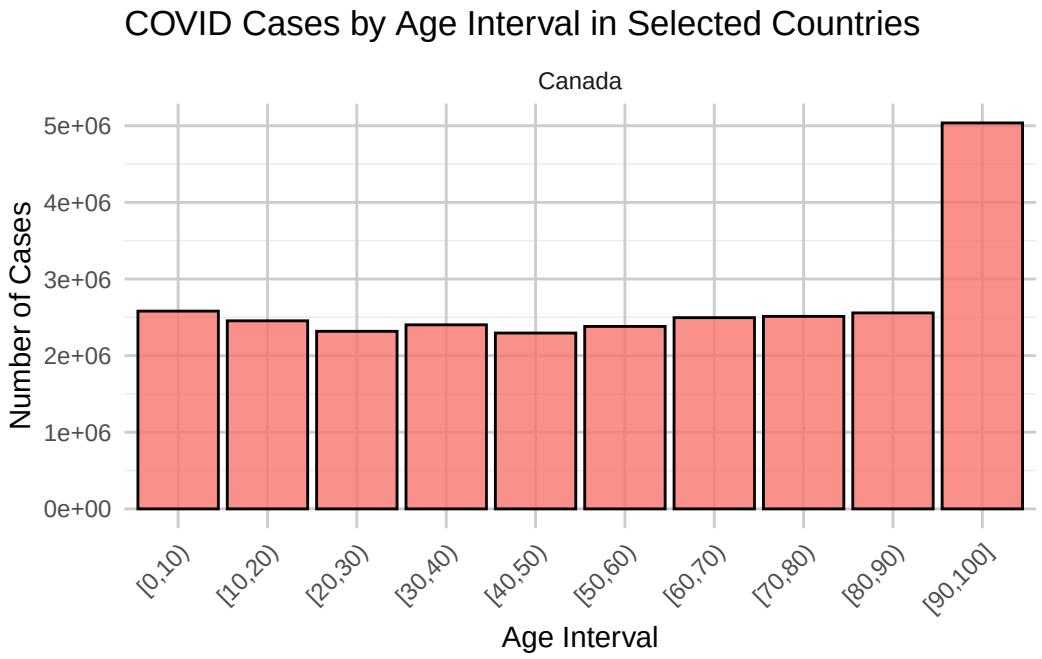
Our results are summarized in **?@tbl-modelresults**.

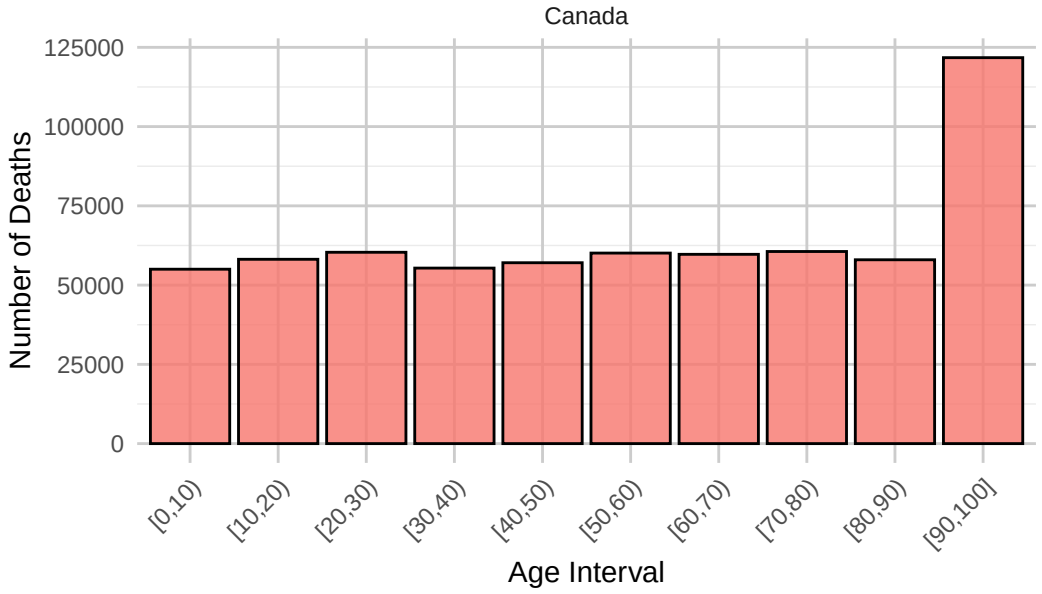


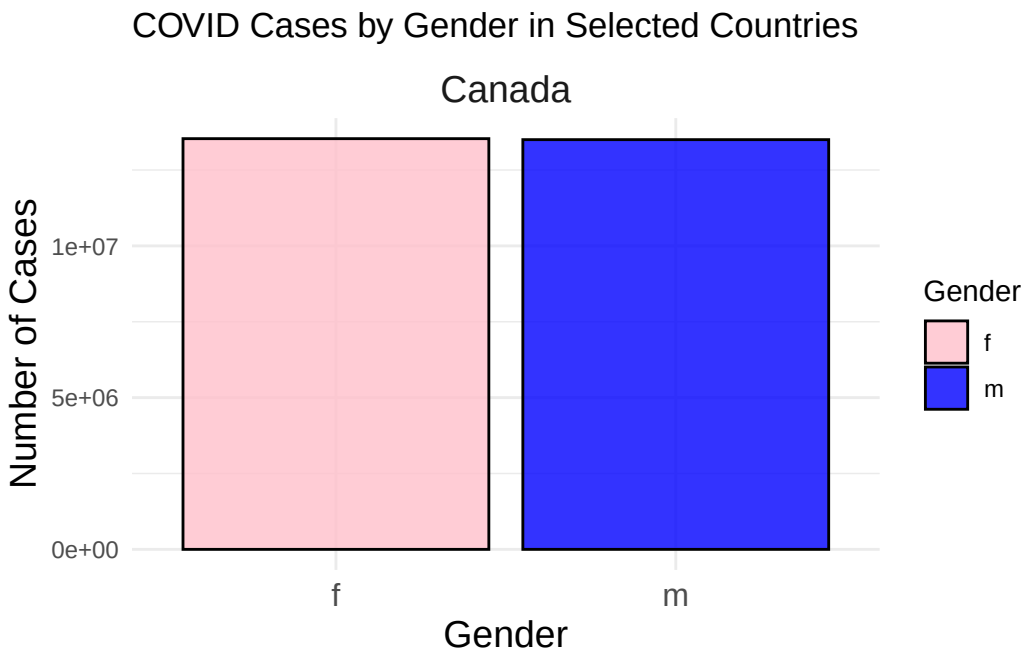
Monthly COVID Cases in Canada (Urban Region)

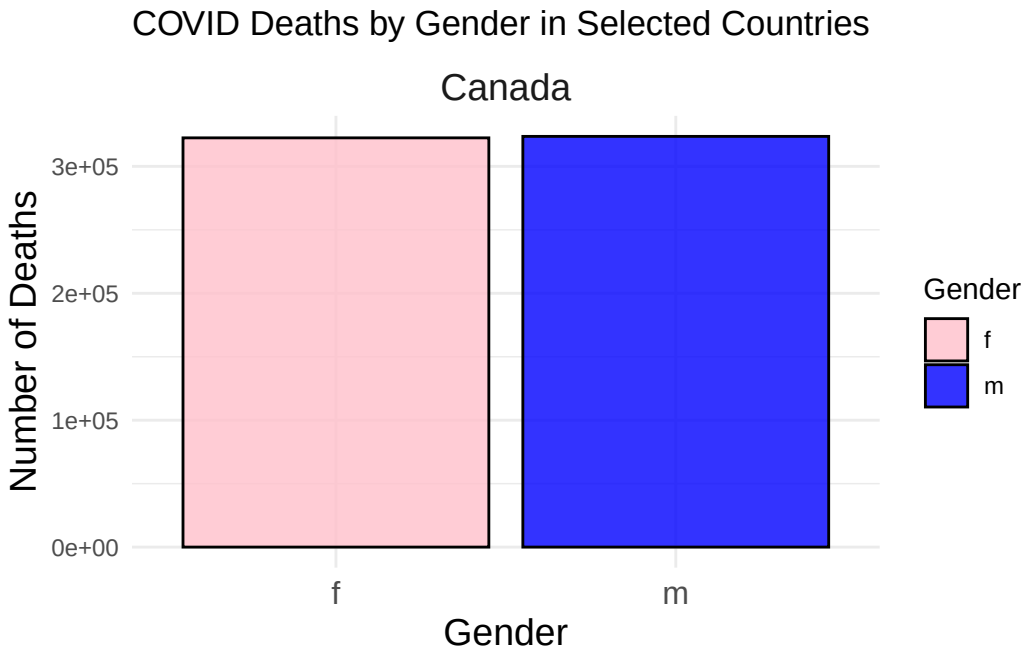
Total Number of Cases

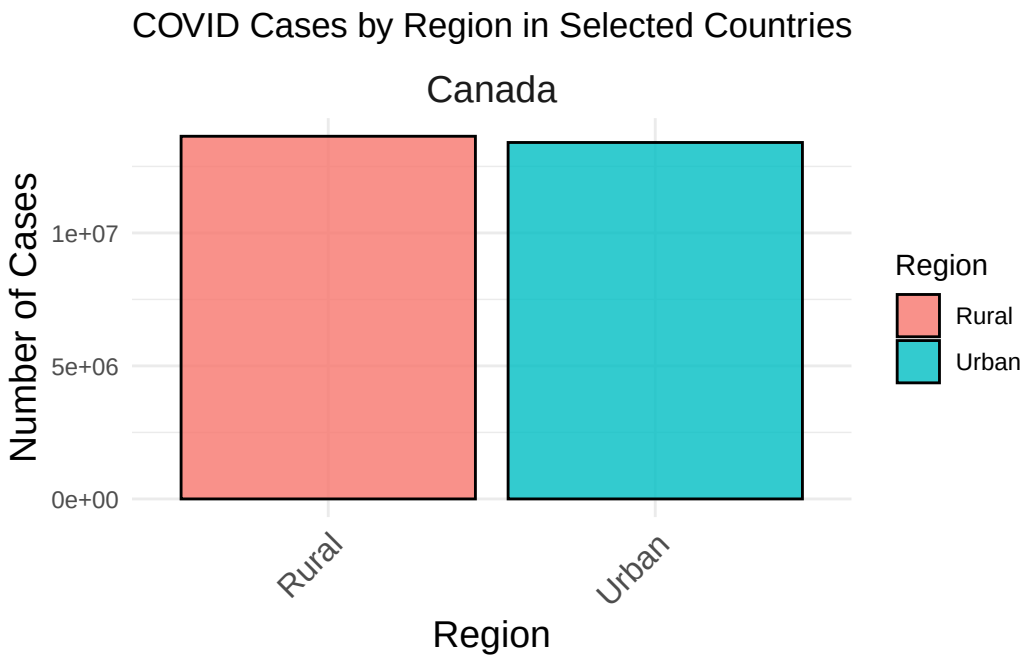
Month

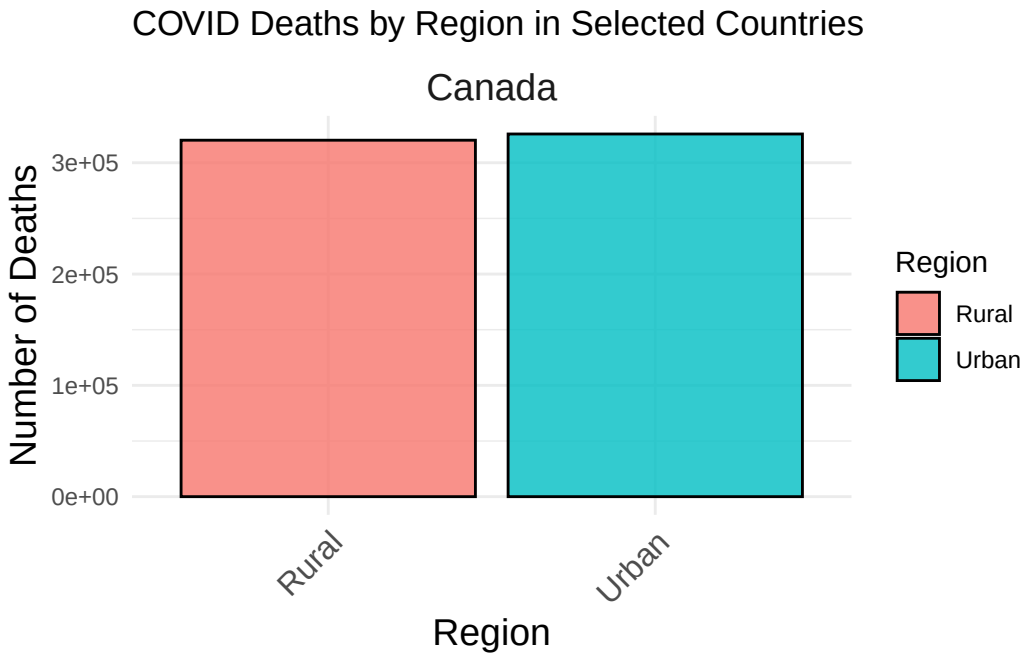

COVID Cases by Age Interval in Selected Countries
Canada
Number of Cases
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
[0,10)
[10,20)
[20,30)
[30,40)
[40,50)
[50,60)
[60,70)
[70,80)
[80,90)
[90,100]
Age Interval


COVID Deaths by Age Interval in Selected Countries
Canada
Number of Deaths
125000
100000
75000
50000
25000
0
[0,10)
[10,20)
[20,30)
[30,40)
[40,50)
[50,60)
[60,70)
[70,80)
[80,90)
[90,100]
Age Interval


COVID Cases by Gender in Selected Countries
Canada
Number of Cases
1e+07
5e+06
0e+00
f
m
Gender
Gender
f
m


COVID Deaths by Gender in Selected Countries
Canada
Number of Deaths
3e+05
2e+05
1e+05
0e+00
f
m
Gender
Gender
f
m


COVID Cases by Region in Selected Countries
Canada
Number of Cases
1e+07
5e+06
0e+00
Rural
Urban
Region
Region
Rural
Urban


COVID Deaths by Region in Selected Countries
Canada
Number of Deaths
3e+05
2e+05
1e+05
0e+00
Rural
Urban
Region
Region
Rural
Urban

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.

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

B.2 Diagnostics

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Gabry, Jonah, Ben Goodrich, Andrew Gelman, Aki Vehtari, and Daniel Simpson. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Open Science Framework (OSF). 2022. "COVID-19 Cases, Deaths, and Tests by Age, Gender, and Region." Open Science Framework (OSF). https://osf.io/43ucn.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Lionel Henry, and Davis Vaughan. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.