

# Mental Health Patterns in Canada and the United States: A Multivariate Analysis

Abbas Ali

2026-08-17

## Research Question

How do mental-health and substance-use patterns differ between Canada and the United States, and has the multivariate difference between the two countries changed significantly over time from 2000 to 2021?

## Data Source

The data used in this analysis were obtained from Our World in Data.

## Data Import

```
# Store the data URL
drug_use_url <- paste0(
  "https://ourworldindata.org/grapher/",
  "rate-of-disease-burden-from-drug-use-disorders-who.csv",
  "?v=1&csvType=full&useColumnShortNames=true"
)

# Import the drug-use data
drug_use_data <- readr::read_csv(drug_use_url)

# Rename the DALY variable
drug_use_data <- drug_use_data %>%
  dplyr::rename(
    daly_rate =
      daly_rate100k__age_group_allages__sex_both_sexes__cause_drug_use_disorders
  )
```

## Data Description

The dataset contains country-level drug-use disorder data across multiple years. It includes the country name, country code, year, and the DALY rate from drug-use disorders per 100,000 people.

## Dataset Structure

```
# Create a summary of the dataset
dataset_summary <- tibble(
  Measure = c("Observations", "Variables"),
  Value = c(
    nrow(drug_use_data),
    ncol(drug_use_data)
  )
)

knitr::kable(
  dataset_summary,
  caption = "Dataset Summary"
)
```

Table 1: Dataset Summary

| Measure      | Value |
|--------------|-------|
| Observations | 4417  |
| Variables    | 4     |

```
# Check the variable types
variable_summary <- tibble(
  Variable = names(drug_use_data),
  Type = sapply(drug_use_data, typeof)
)

knitr::kable(
  variable_summary,
  col.names = c("Variable", "Type"),
  caption = "Variable Structure"
)
```

Table 2: Variable Structure

| Variable  | Type      |
|-----------|-----------|
| entity    | character |
| code      | character |
| year      | double    |
| daly_rate | double    |

## Interpretation

The source dataset contains 4,417 observations and four variables.

## Data Preparation

I restricted the analysis to Canada and the United States because these are the two countries being compared throughout the study.

```
# Keep only Canada and United States
can_us_data <- drug_use_data %>%
  filter(entity %in% c("Canada", "United States"))
```

## Do Canada and the United States Have the Same Number of Observations?

I first checked whether both countries contribute the same number of observations.

```
# Count the observations for each country
country_counts <- dplyr::count(
  can_us_data,
  entity,
  name = "Observations"
)

knitr::kable(
  country_counts,
  col.names = c("Country", "Observations"),
  caption = "Number of Observations by Country"
)
```

Table 3: Number of Observations by Country

| Country       | Observations |
|---------------|--------------|
| Canada        | 22           |
| United States | 22           |

## Interpretation

Canada and the United States each have 22 observations, so both countries are equally represented in the dataset.

## Do Both Countries Cover the Same Time Period?

I then verified that the observations for both countries cover the same years.

```
# Check the time period for each country
time_period <- can_us_data %>%
  group_by(entity) %>%
  summarise(
    first_year = min(year),
    last_year = max(year),
    .groups = "drop"
  )

knitr::kable(
  time_period,
  col.names = c("Country", "First Year", "Last Year"),
  caption = "Time Period Covered by Each Country"
)
```

Table 4: Time Period Covered by Each Country

| Country       | First Year | Last Year |
|---------------|------------|-----------|
| Canada        | 2000       | 2021      |
| United States | 2000       | 2021      |

### Interpretation

Both countries cover the period from 2000 to 2021, so the comparison is based on the same years.

## Data Quality Checks

### Are There Any Missing Values?

Before continuing with the analysis, I checked whether any values were missing from the Canada and United States data.

```
# Count missing values in each variable
missing_values <- can_us_data %>%
  summarise(
    across(everything(), ~ sum(is.na(.)))
  ) %>%
  pivot_longer(
    everything(),
    names_to = "Variable",
    values_to = "Missing_Values"
  )

knitr::kable(
  missing_values,
  col.names = c("Variable", "Missing Values"),
  caption = "Missing Values by Variable"
)
```

Table 5: Missing Values by Variable

| Variable  | Missing Values |
|-----------|----------------|
| entity    | 0              |
| code      | 0              |
| year      | 0              |
| daly_rate | 0              |

### Interpretation

No missing values were found, so all 44 observations can be retained for the analysis.

### Are There Any Duplicate Observations?

I also checked whether any country-year combination appears more than once in the dataset.

```
# Check for duplicate country-year observations
duplicate_check <- can_us_data %>%
  dplyr::group_by(entity, year) %>%
  dplyr::summarise(
```

```

    n = dplyr::n(),
    .groups = "drop"
  ) %>%
  dplyr::filter(n > 1)

# Summarize the duplicate check
duplicate_summary <- tibble(
  Check = "Duplicate country-year observations",
  Number_Found = nrow(duplicate_check)
)

knitr::kable(
  duplicate_summary,
  col.names = c("Check", "Number Found"),
  caption = "Duplicate Observation Check"
)

```

Table 6: Duplicate Observation Check

| Check                               | Number Found |
|-------------------------------------|--------------|
| Duplicate country-year observations | 0            |

## Interpretation

No duplicate country-year observations were found, so each country has one observation for every year in the study period.

## Descriptive Statistics

### How Does the Drug-Use Burden Compare Between Canada and the United States?

Before fitting statistical models, I summarized the DALY rates for each country to compare their typical levels, variation, and observed range.

```

# Calculate summary statistics for each country
drug_summary <- can_us_data %>%
  group_by(entity) %>%
  summarise(
    mean_rate = mean(daly_rate),
    median_rate = median(daly_rate),
    sd_rate = sd(daly_rate),
    min_rate = min(daly_rate),
    max_rate = max(daly_rate),
    .groups = "drop"
  )

# Display the summary statistics
knitr::kable(
  drug_summary,
  digits = 2,
  col.names = c(
    "Country",
    "Mean",
    "Median",

```

```

    "SD",
    "Minimum",
    "Maximum"
  ),
  caption = "Summary Statistics for Drug-Use DALY Rates"
)

```

Table 7: Summary Statistics for Drug-Use DALY Rates

| Country       | Mean    | Median | SD     | Minimum | Maximum |
|---------------|---------|--------|--------|---------|---------|
| Canada        | 658.76  | 524.02 | 277.35 | 436.73  | 1323.58 |
| United States | 1171.02 | 978.66 | 542.71 | 542.60  | 2442.41 |

## Interpretation

The United States had a higher average drug-use DALY rate than Canada (1171.02 versus 658.76 per 100,000). It also had greater variation over the study period, with a standard deviation of 542.71 compared with 277.35 in Canada.

## How Have Drug-Use DALY Rates Changed Over Time?

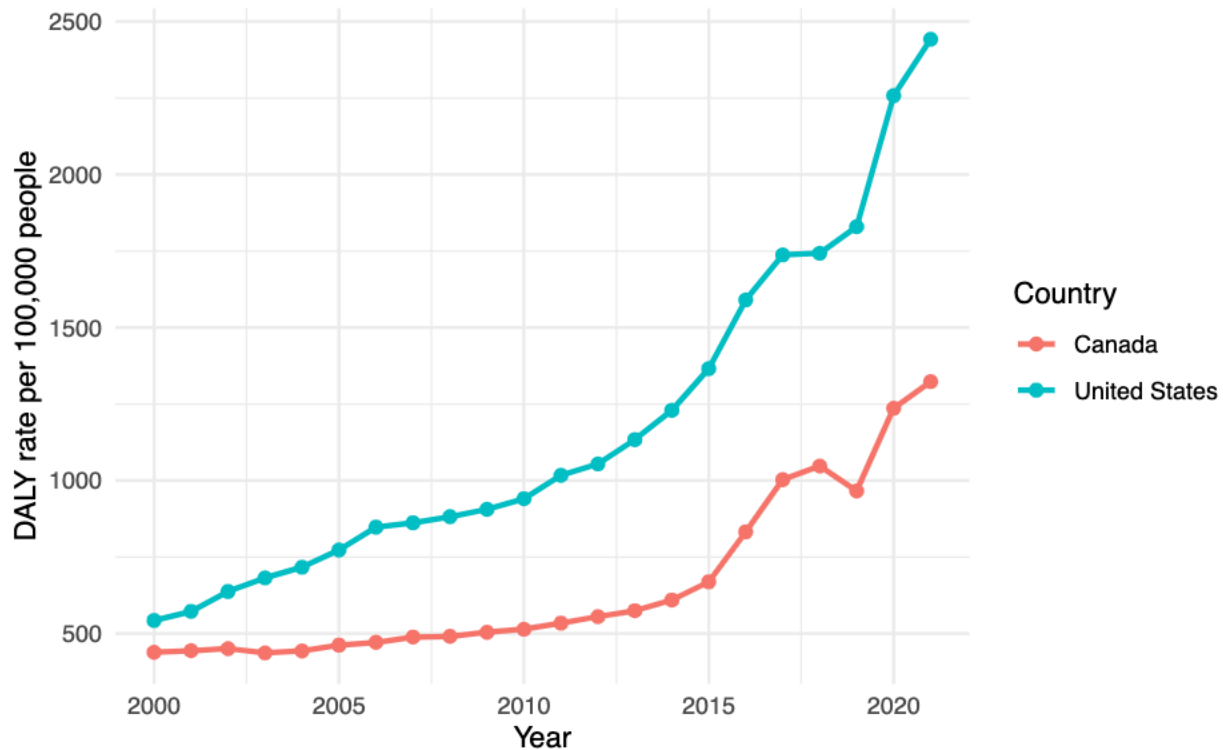
I plotted the yearly DALY rates to examine how the pattern changed in each country from 2000 to 2021.

```

# Plot DALY rates over time
ggplot(
  can_us_data,
  aes(x = year, y = daly_rate, color = entity)
) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Drug-Use DALY Rates in Canada and the United States",
    subtitle = "2000-2021",
    x = "Year",
    y = "DALY rate per 100,000 people",
    color = "Country"
  ) +
  theme_minimal()

```

## Drug-Use DALY Rates in Canada and the United States 2000–2021



### Interpretation

Drug-use DALY rates were consistently higher in the United States. The gap between the two countries became more pronounced after the mid-2010s, suggesting that the difference changed over time.

### How Large Is the Difference Between the Two Countries Each Year?

I calculated the U.S.–Canada difference in DALY rates for every year. Selected years are shown below to summarize how the gap changed across the study period.

```
# Calculate the yearly difference
yearly_difference <- can_us_data %>%
  dplyr::select(entity, year, daly_rate) %>%
  tidyr::pivot_wider(
    names_from = entity,
    values_from = daly_rate
  ) %>%
  dplyr::mutate(
    difference = `United States` - Canada
  )

# Select representative years for the table
selected_differences <- yearly_difference %>%
  dplyr::filter(year %in% c(2000, 2005, 2010, 2015, 2020, 2021)) %>%
  dplyr::rename(
    Year = year,
    Difference = difference
  )
```

```

# Display the table
knitr::kable(
  selected_differences,
  digits = 2,
  col.names = c(
    "Year",
    "Canada",
    "United States",
    "Difference"
  ),
  caption = "Selected Yearly Differences in Drug-Use DALY Rates"
)

```

Table 8: Selected Yearly Differences in Drug-Use DALY Rates

| Year | Canada  | United States | Difference |
|------|---------|---------------|------------|
| 2000 | 438.56  | 542.60        | 104.04     |
| 2005 | 461.97  | 772.88        | 310.91     |
| 2010 | 514.08  | 940.33        | 426.25     |
| 2015 | 668.93  | 1365.76       | 696.83     |
| 2020 | 1236.39 | 2257.78       | 1021.39    |
| 2021 | 1323.58 | 2442.41       | 1118.83    |

## Interpretation

The difference was positive in every year, confirming that the United States had the higher drug-use DALY rate throughout the study period.

## How Did the Difference Change Over Time?

I plotted the yearly difference to examine how the gap changed across the study period.

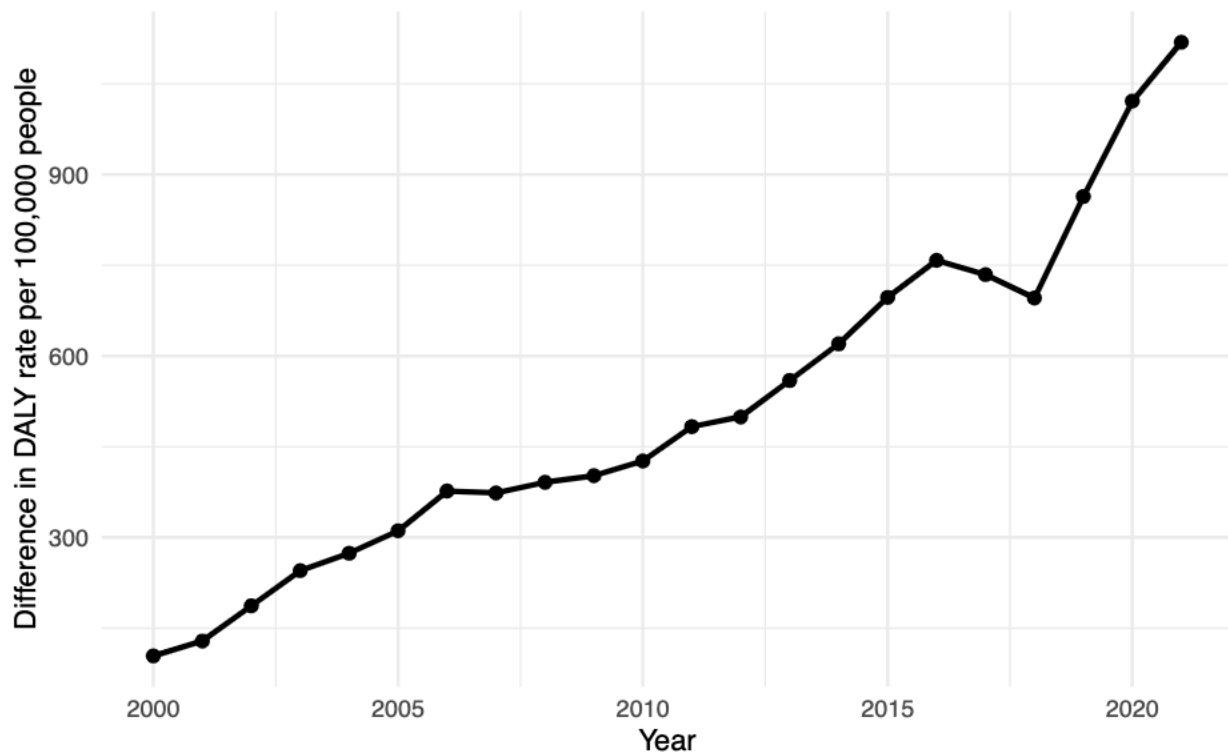
```

# Plot the yearly difference
ggplot(
  yearly_difference,
  aes(x = year, y = difference)
) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Yearly Difference in Drug-Use DALY Rates",
    subtitle = "United States minus Canada, 2000-2021",
    x = "Year",
    y = "Difference in DALY rate per 100,000 people"
  ) +
  theme_minimal()

```

## Yearly Difference in Drug-Use DALY Rates

United States minus Canada, 2000–2021



### Interpretation

The gap generally widened over time, and the increase became steeper in the later years. This suggests that a simple straight-line trend may not fully describe how the difference changed over time.

### Trend Analysis

#### How Quickly Did the Gap Change Over Time?

To quantify how the difference changed over time, I fitted a simple linear trend model to the yearly DALY-rate gap.

```
# Create a time variable starting at 0 in the year 2000
yearly_difference <- yearly_difference %>%
  dplyr::mutate(
    year_since_2000 = year - 2000
  )

# Fit a linear trend model
gap_model <- lm(
  difference ~ year_since_2000,
  data = yearly_difference
)

# Display the model results
summary(gap_model)
```

```
##
```

```
## Call:
## lm(formula = difference ~ year_since_2000, data = yearly_difference)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.540  -47.848    2.901   26.159  172.260
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      77.942     27.463   2.838  0.0102 *
## year_since_2000   41.363       2.239  18.477 4.87e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.62 on 20 degrees of freedom
## Multiple R-squared:  0.9447, Adjusted R-squared:  0.9419
## F-statistic: 341.4 on 1 and 20 DF,  p-value: 4.871e-14
```

### Interpretation

The estimated slope is 41.36, meaning that the U.S.–Canada DALY-rate gap increased by about 41 DALYs per 100,000 people per year on average.

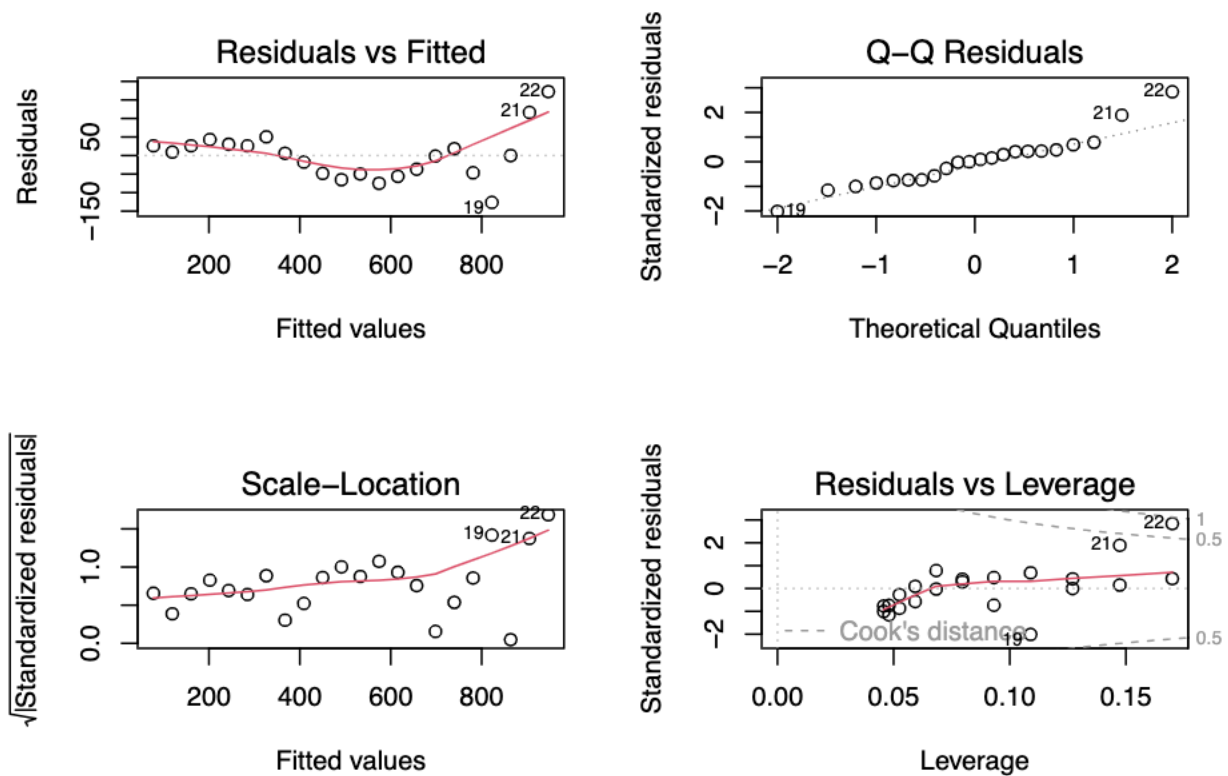
The model explains about 94.5% of the variation in the yearly gap. Before relying on the model for inference, however, its residual assumptions need to be checked.

### Does the Linear Trend Model Meet Its Assumptions?

Before relying on the regression results, I examined the residual diagnostics to check whether the linear model assumptions were reasonable.

```
# Check the linear model diagnostics
par(mfrow = c(2, 2))

plot(gap_model)
```



```
par(mfrow = c(1, 1))
```

### Interpretation

The residuals-versus-fitted plot shows noticeable curvature, indicating that the straight-line model does not fully capture the time pattern.

The Q-Q plot is reasonably close to linear for most observations, although some departures remain in the tails. The scale-location plot also suggests greater residual spread at higher fitted values.

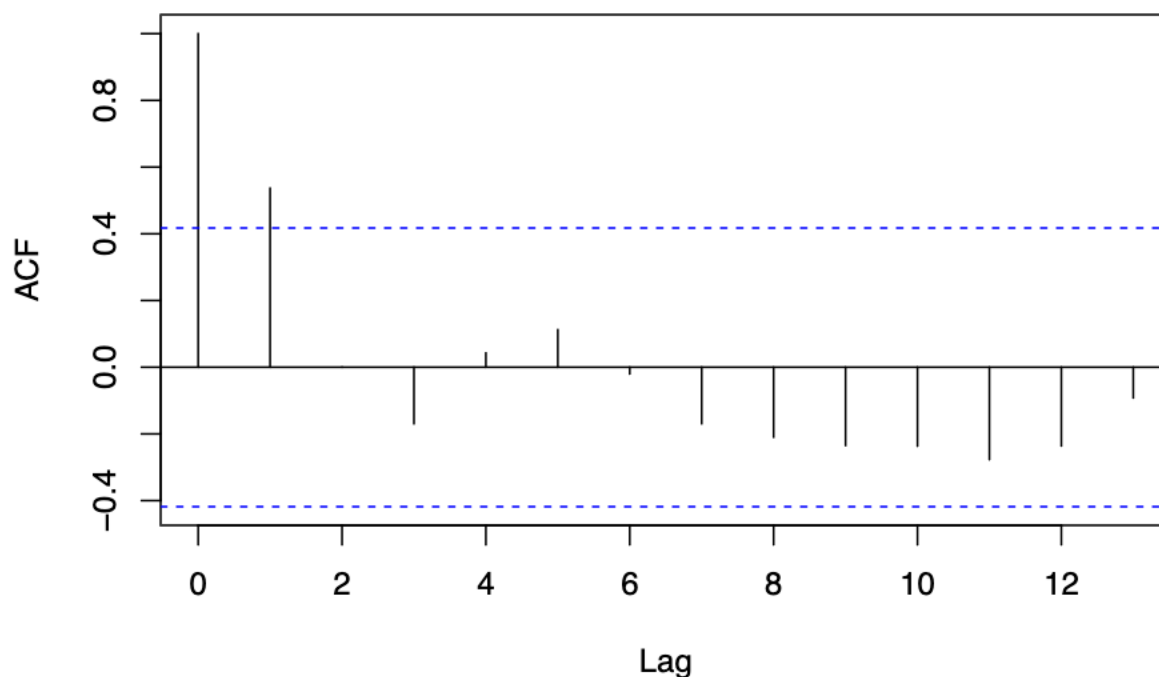
A few observations have relatively high leverage, so their influence should be examined separately.

### Are the Regression Residuals Correlated Across Years?

Because the observations are recorded in consecutive years, I checked whether residuals from nearby years were related.

```
# Check residual autocorrelation
acf(residuals(gap_model))
```

## Series residuals(gap\_model)



### Interpretation

The lag-1 autocorrelation exceeds the approximate confidence limit, indicating that residuals from adjacent years are not fully independent.

This means the independence assumption of the simple linear model is not satisfied, so its standard errors and p-values should be interpreted with caution.

### Would a Curved Trend Fit the Yearly Gap Better?

The linear-model diagnostics suggest that the DALY-rate gap does not change at a constant rate over time. I therefore fitted a quadratic trend model to allow the relationship to curve.

The linear model is

$$\text{Gap} = \beta_0 + \beta_1(\text{Time})$$

where  $\beta_1$  represents a constant yearly change.

The quadratic model is

$$\text{Gap} = \beta_0 + \beta_1(\text{Time}) + \beta_2(\text{Time}^2)$$

The squared term allows the rate of change in the gap to vary over time.

```
# Fit a quadratic trend model
gap_model_quadratic <- lm(
  difference ~ year_since_2000 + I(year_since_2000^2),
  data = yearly_difference
)
```

```
# Display the model results
summary(gap_model_quadratic)
```

```
##
## Call:
## lm(formula = difference ~ year_since_2000 + I(year_since_2000^2),
##     data = yearly_difference)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -143.97  -30.70   -6.38   29.02   95.98
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    154.2168    31.5363   4.890 0.000102 ***
## year_since_2000    18.4808     6.9582   2.656 0.015600 *
## I(year_since_2000^2)  1.0896     0.3199   3.406 0.002965 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.86 on 19 degrees of freedom
## Multiple R-squared:  0.9656, Adjusted R-squared:  0.962
## F-statistic: 267 on 2 and 19 DF, p-value: 1.238e-14
```

### Interpretation

The quadratic term is statistically significant ( $p = 0.003$ ), providing evidence that the yearly gap follows a curved rather than purely linear pattern.

The positive quadratic coefficient ( $\beta_2 = 1.09$ ) indicates that the gap increased at a faster rate as the study period progressed. The model explains approximately 96.6% of the variation in the yearly gap.

### Does the Quadratic Model Fit Better Than the Linear Model?

To determine whether adding the quadratic time term improves the model, I compared the linear and quadratic models directly.

```
# Compare the linear and quadratic models
anova(gap_model, gap_model_quadratic)
```

```
## Analysis of Variance Table
##
## Model 1: difference ~ year_since_2000
## Model 2: difference ~ year_since_2000 + I(year_since_2000^2)
##   Res.Df  RSS Df Sum of Sq    F    Pr(>F)
## 1      20 88754
## 2      19 55110  1    33644 11.599 0.002965 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### Interpretation

The model comparison is statistically significant ( $F = 11.60$ ,  $p = 0.003$ ). This provides evidence that adding the quadratic time term improves the fit compared with the simple linear model.

The quadratic model will therefore be used for the next stage of the analysis.

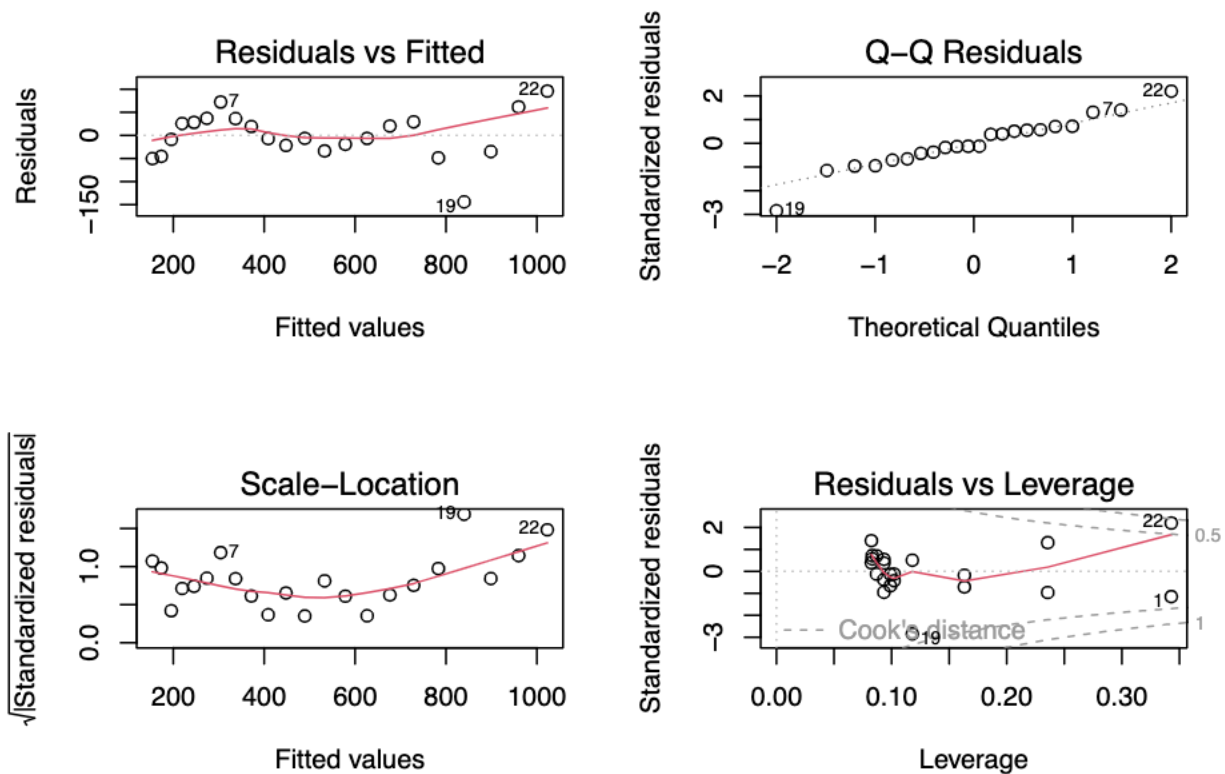
## Does the Quadratic Model Meet the Regression Assumptions?

After selecting the quadratic model, I checked whether its residuals satisfied the assumptions required for reliable inference.

```
# Check the quadratic model diagnostics
```

```
par(mfrow = c(2, 2))
```

```
plot(gap_model_quadratic)
```

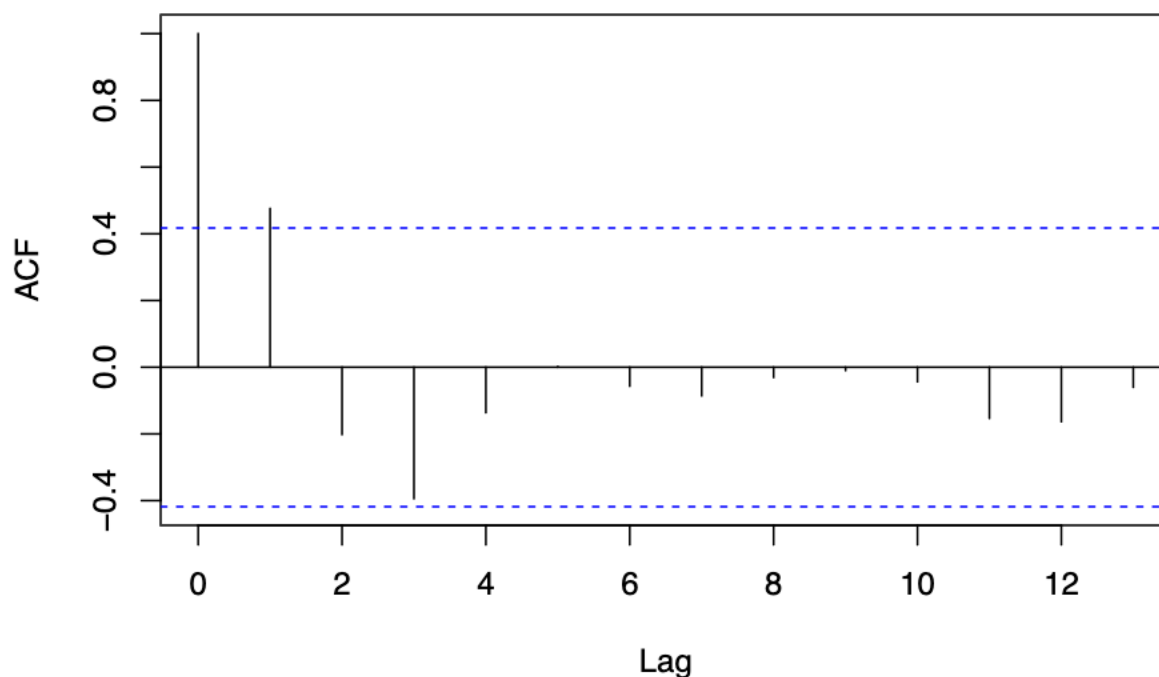


```
par(mfrow = c(1, 1))
```

```
# Check residual autocorrelation
```

```
acf(residuals(gap_model_quadratic))
```

## Series residuals(gap\_model\_quadratic)



### Interpretation

The quadratic model shows less systematic residual structure than the linear model. The Q-Q plot is reasonably linear for most observations, although some departure remains in the tails.

The scale-location plot suggests that residual variation may be somewhat larger at higher fitted values, and a few observations have relatively high leverage.

All non-zero lags in the ACF remain within the approximate confidence limits, so there is no clear residual autocorrelation after including the quadratic time term.

### Are Any Observations Strongly Influencing the Quadratic Model?

The diagnostic plots identified a few observations with relatively high leverage. I therefore examined Cook's distance to check whether any individual year has a strong influence on the fitted model.

```
# Calculate Cook's distance
cooks_distance <- cooks.distance(gap_model_quadratic)

# Identify observations above the screening threshold
which(
  cooks_distance > 4 / nrow(yearly_difference)
)
```

```
## 1 19 22
```

```
## 1 19 22
```

### Interpretation

Observations 1, 19, and 22 exceed the Cook's distance screening threshold. These correspond to the years 2000, 2018, and 2021 and may have greater influence on the fitted model.

Their influence should therefore be checked before deciding how stable the quadratic trend is.

### Is the Quadratic Model Sensitive to the Influential Years?

To check whether the fitted trend depends heavily on the influential observations, I refitted the quadratic model after excluding the years 2000, 2018, and 2021.

```
# Remove the influential years
sensitivity_data <- yearly_difference %>%
  dplyr::filter(!year %in% c(2000, 2018, 2021))

# Refit the quadratic model
sensitivity_model <- lm(
  difference ~ year_since_2000 + I(year_since_2000^2),
  data = sensitivity_data
)

# Display the sensitivity-model results
summary(sensitivity_model)
```

```
##
## Call:
## lm(formula = difference ~ year_since_2000 + I(year_since_2000^2),
##     data = sensitivity_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -55.993 -25.037  -9.182   23.560   64.132
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    165.3413    27.5273   6.006 1.83e-05 ***
## year_since_2000    18.0647     6.0659   2.978  0.00888 **
## I(year_since_2000^2)  1.0766     0.2839   3.792  0.00160 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.96 on 16 degrees of freedom
## Multiple R-squared:  0.9787, Adjusted R-squared:  0.9761
## F-statistic: 368.2 on 2 and 16 DF,  p-value: 4.179e-14
```

### Interpretation

After removing the influential years, the estimated quadratic trend changed very little and remained statistically significant.

This suggests that the curved pattern is not being driven by only those three observations.

### How Well Does the Quadratic Model Follow the Observed Gap?

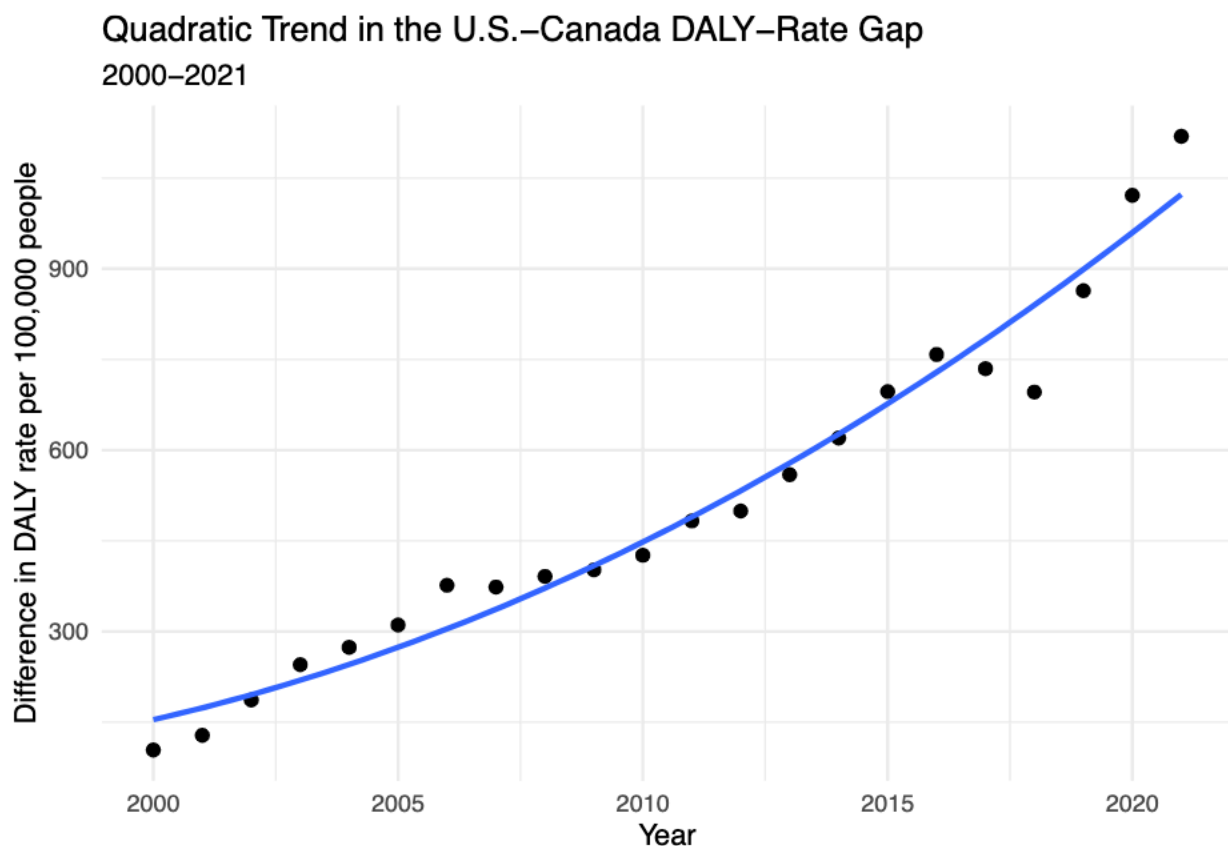
After checking the model assumptions and sensitivity to influential years, I plotted the fitted quadratic trend together with the observed yearly differences.

```
# Plot the observed gap and fitted quadratic trend
ggplot(
  yearly_difference,
  aes(x = year, y = difference)
```

```

) +
  geom_point(size = 2) +
  geom_smooth(
    method = "lm",
    formula = y ~ poly(x, 2, raw = TRUE),
    se = FALSE,
    linewidth = 1
  ) +
  labs(
    title = "Quadratic Trend in the U.S.–Canada DALY–Rate Gap",
    subtitle = "2000–2021",
    x = "Year",
    y = "Difference in DALY rate per 100,000 people"
  ) +
  theme_minimal()

```



### Interpretation

The fitted quadratic curve follows the overall pattern of the observed yearly differences well. The upward curvature shows that the gap between the two countries increased more rapidly in the later years.

### What Do We Conclude About Drug-Use Disorders?

The United States had a higher drug-use DALY burden than Canada throughout 2000–2021, and the gap between the two countries increased over time.

A quadratic trend described this change better than a simple linear trend. The result remained stable after checking model assumptions and influential years.

## Second Indicator: Mental Health and Substance-Use Disorder Death Rate

To expand the analysis beyond drug-use burden, I added the death rate from mental-health and substance-use disorders as a second indicator.

### Does the Second Indicator Cover the Same Countries and Years?

Before combining the two indicators, I checked whether the second dataset provides complete observations for Canada and the United States over the same 2000–2021 period.

```
# Import the second indicator
mental_health_death_data <- readr::read_csv(
  "https://ourworldindata.org/grapher/death-rate-from-mental-health-and-substance-use-disorders-who.csv"
)

# Rename variables for consistency
mental_health_death_data <- mental_health_death_data %>%
  dplyr::rename(
    entity = Entity,
    code = Code,
    year = Year,
    mental_health_death_rate =
      `Death rate from mental and substance use disorders among both sexes`
  )

# Keep only Canada and the United States
mental_health_death_can_us <- mental_health_death_data %>%
  dplyr::filter(
    entity %in% c("Canada", "United States")
  )

# Summarize coverage by country
mental_health_coverage <- mental_health_death_can_us %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    Observations = dplyr::n(),
    First_Year = min(year),
    Last_Year = max(year),
    .groups = "drop"
  )

# Display the coverage table
knitr::kable(
  mental_health_coverage,
  col.names = c(
    "Country",
    "Observations",
    "First Year",
    "Last Year"
  ),
  caption = "Coverage of the Second Health Indicator"
)
```

Table 9: Coverage of the Second Health Indicator

| Country       | Observations | First Year | Last Year |
|---------------|--------------|------------|-----------|
| Canada        | 22           | 2000       | 2021      |
| United States | 22           | 2000       | 2021      |

### Interpretation

Canada and the United States each have 22 observations covering 2000–2021. The second indicator therefore matches the country and time coverage of the drug-use DALY data and can be combined with it directly.

### Did the Two Indicators Join Correctly?

Because both datasets cover the same countries and years, I combined them using country, code, and year. I then checked whether the join preserved all 44 observations and whether any mental-health death-rate values were missing.

```
# Combine the two indicators
combined_data <- can_us_data %>%
  dplyr::left_join(
    mental_health_death_can_us,
    by = c("entity", "code", "year")
  )

# Check whether the join was complete
join_check <- tibble(
  Check = c(
    "Expected observations",
    "Observations after join",
    "Missing death-rate values"
  ),
  Result = c(
    44,
    nrow(combined_data),
    sum(is.na(combined_data$mental_health_death_rate))
  )
)

# Display the join check
knitr::kable(
  join_check,
  col.names = c("Check", "Result"),
  caption = "Validation of the Two-Indicator Join"
)
```

Table 10: Validation of the Two-Indicator Join

| Check                     | Result |
|---------------------------|--------|
| Expected observations     | 44     |
| Observations after join   | 44     |
| Missing death-rate values | 0      |

## Interpretation

The combined dataset contains all 44 expected observations, and no mental-health death-rate values are missing. This confirms that the two indicators matched correctly by country and year.

## How Does the Mental-Health and Substance-Use Disorder Death Rate Compare Between the Two Countries?

Before examining the relationship between the indicators, I summarized the mental-health and substance-use disorder death rate for each country.

```
# Calculate summary statistics for the second indicator
mental_health_summary <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    mean_rate = mean(mental_health_death_rate),
    median_rate = median(mental_health_death_rate),
    sd_rate = sd(mental_health_death_rate),
    min_rate = min(mental_health_death_rate),
    max_rate = max(mental_health_death_rate),
    .groups = "drop"
  )

# Display the summary statistics
knitr::kable(
  mental_health_summary,
  digits = 2,
  col.names = c(
    "Country",
    "Mean",
    "Median",
    "SD",
    "Minimum",
    "Maximum"
  ),
  caption = "Summary Statistics for Mental-Health and Substance-Use Disorder Death Rates"
)
```

Table 11: Summary Statistics for Mental-Health and Substance-Use Disorder Death Rates

| Country       | Mean  | Median | SD   | Minimum | Maximum |
|---------------|-------|--------|------|---------|---------|
| Canada        | 10.15 | 7.56   | 5.32 | 5.81    | 23.71   |
| United States | 16.91 | 14.19  | 7.87 | 7.84    | 38.09   |

## Interpretation

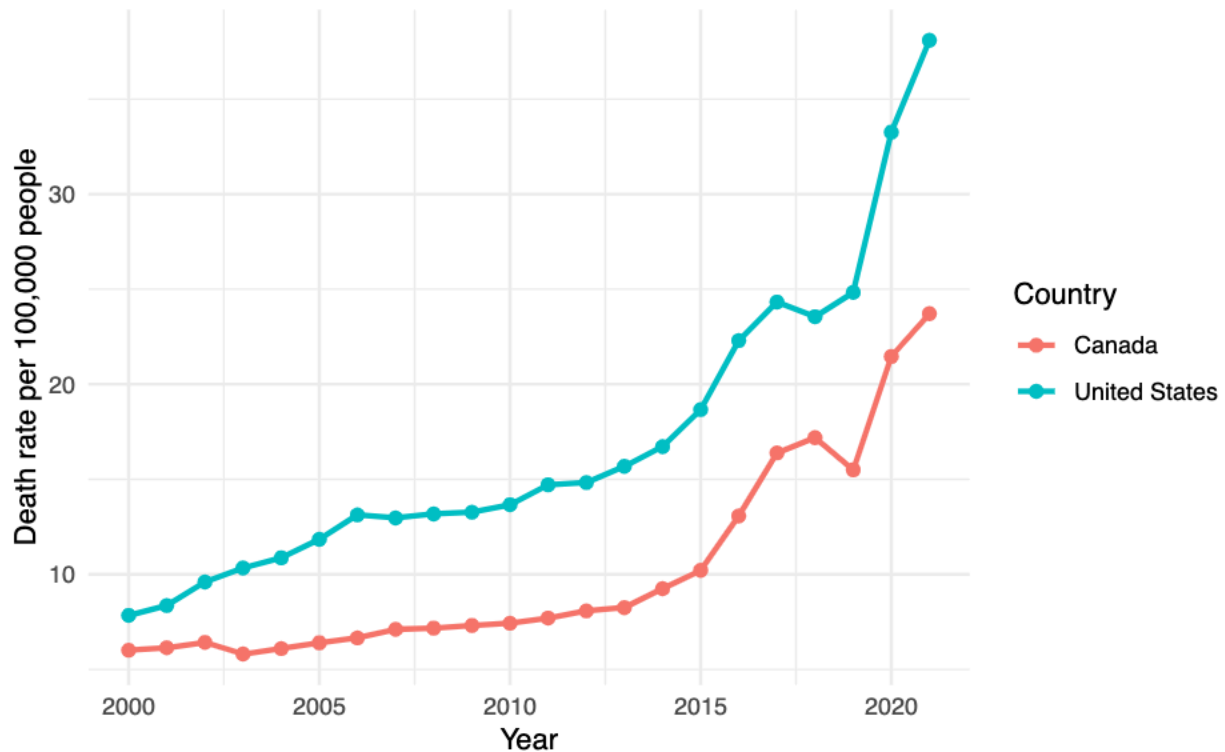
The United States had a higher average death rate than Canada (16.91 versus 10.15). It also showed greater variation over the study period, with a standard deviation of 7.87 compared with 5.32 in Canada.

## How Did the Mental-Health and Substance-Use Disorder Death Rate Change Over Time?

I plotted the yearly death rates to compare how the pattern changed in Canada and the United States from 2000 to 2021.

```
# Plot the death rate over time
ggplot(
  combined_data,
  aes(
    x = year,
    y = mental_health_death_rate,
    color = entity
  )
) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Mental-Health and Substance-Use Disorder Death Rates",
    subtitle = "Canada and United States, 2000-2021",
    x = "Year",
    y = "Death rate per 100,000 people",
    color = "Country"
  ) +
  theme_minimal()
```

**Mental-Health and Substance-Use Disorder Death Rates**  
Canada and United States, 2000-2021



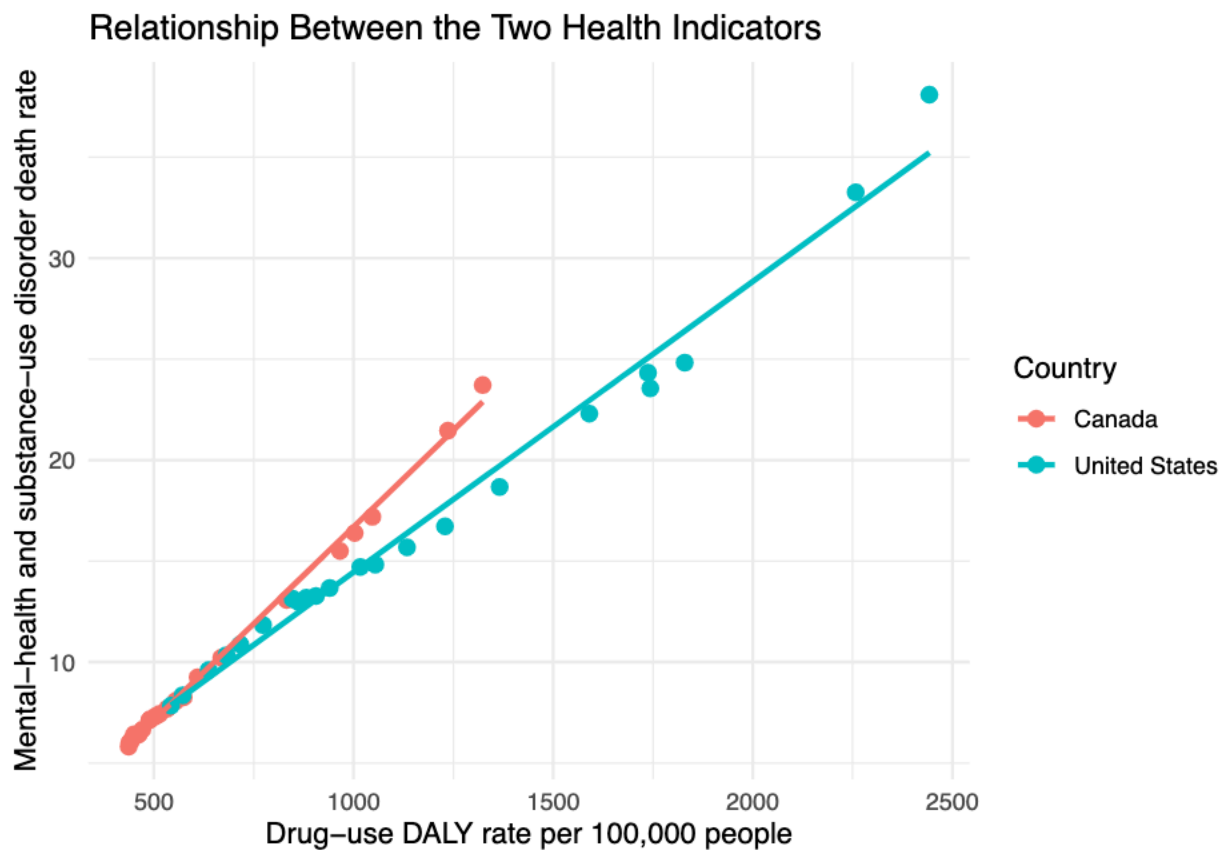
### Interpretation

Death rates increased in both countries over the study period, with a steeper rise after the mid-2010s. The United States remained above Canada, and the gap between the two countries became more pronounced in the later years.

## Are the Two Health Indicators Related?

Because the two indicators measure different aspects of health burden, I examined whether higher drug-use DALY rates are associated with higher mental-health and substance-use disorder death rates.

```
# Plot the relationship between the two indicators
ggplot(
  combined_data,
  aes(
    x = daly_rate,
    y = mental_health_death_rate,
    color = entity
  )
) +
  geom_point(size = 2.5) +
  geom_smooth(
    method = "lm",
    se = FALSE
  ) +
  labs(
    title = "Relationship Between the Two Health Indicators",
    x = "Drug-use DALY rate per 100,000 people",
    y = "Mental-health and substance-use disorder death rate",
    color = "Country"
  ) +
  theme_minimal()
```



## Interpretation

The two indicators move in the same direction in both countries: higher drug-use DALY rates are associated with higher mental-health and substance-use disorder death rates.

However, both indicators also increase over time, so the apparent relationship may partly reflect their shared time trend. A correlation measure is therefore useful, but it should be interpreted cautiously.

## How Strong Is the Relationship Between the Two Indicators?

To quantify the relationship, I calculated the Pearson correlation between drug-use DALY rates and mental-health and substance-use disorder death rates separately for each country.

```
# Calculate the correlation for each country
indicator_correlation <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    correlation = cor(
      daly_rate,
      mental_health_death_rate
    ),
    .groups = "drop"
  )

# Display the correlations
knitr::kable(
  indicator_correlation,
  digits = 3,
  col.names = c(
    "Country",
    "Correlation"
  ),
  caption = "Correlation Between the Two Health Indicators"
)
```

Table 12: Correlation Between the Two Health Indicators

| Country       | Correlation |
|---------------|-------------|
| Canada        | 0.998       |
| United States | 0.992       |

## Interpretation

The two indicators are very strongly positively correlated in both countries, with  $r = 0.998$  for Canada and  $r = 0.992$  for the United States.

Because both indicators also rise over time, these correlations may partly reflect their shared time trend rather than only a direct relationship between the indicators.

## Does the Relationship Remain After Accounting for Time?

Because both indicators increase over the study period, their strong correlation may partly reflect a shared time trend. I therefore removed the time trend from each indicator within each country and then examined the correlation between the remaining variation.

```

# Create a time variable
combined_data <- combined_data %>%
  dplyr::mutate(
    year_since_2000 = year - 2000
  )

# Remove the quadratic time trend within each country
detrended_data <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::mutate(
    daly_residual = residuals(
      lm(
        daly_rate ~ year_since_2000 + I(year_since_2000^2)
      )
    ),
    death_residual = residuals(
      lm(
        mental_health_death_rate ~
          year_since_2000 + I(year_since_2000^2)
      )
    )
  ) %>%
  dplyr::ungroup()

# Calculate the correlation after removing the time trend
adjusted_correlation <- detrended_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    correlation = cor(
      daly_residual,
      death_residual
    ),
    .groups = "drop"
  )

# Display the adjusted correlations
knitr::kable(
  adjusted_correlation,
  digits = 3,
  col.names = c(
    "Country",
    "Adjusted Correlation"
  ),
  caption = "Correlation After Removing the Time Trend"
)

```

Table 13: Correlation After Removing the Time Trend

| Country       | Adjusted Correlation |
|---------------|----------------------|
| Canada        | 0.976                |
| United States | 0.968                |

## Interpretation

The relationship remained very strong after accounting for the time trend, with adjusted correlations of  $r = 0.976$  for Canada and  $r = 0.968$  for the United States.

This indicates that the association between the two indicators is not explained only by their common increase over time.

## Are the Detrended Residuals Still Correlated Over Time?

After removing the time trend, I checked whether the remaining residuals still showed year-to-year dependence. This helps determine how cautiously the adjusted correlations should be interpreted.

*# Check autocorrelation in the detrended residuals*

```
par(mfrow = c(2, 2))

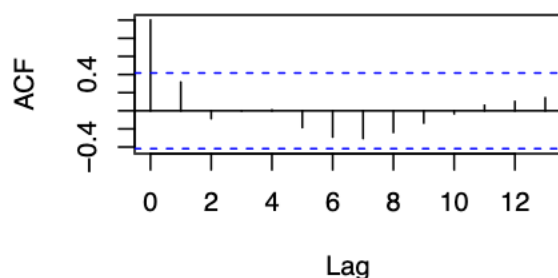
acf(
  detrended_data$daly_residual[
    detrended_data$entity == "Canada"
  ],
  main = "Canada: DALY Residuals"
)

acf(
  detrended_data$death_residual[
    detrended_data$entity == "Canada"
  ],
  main = "Canada: Death-Rate Residuals"
)

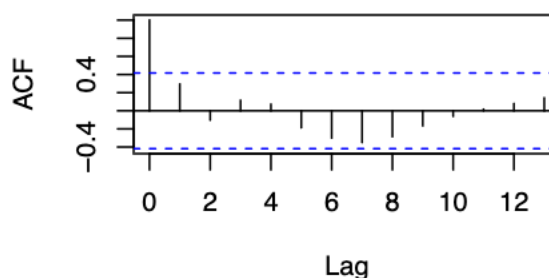
acf(
  detrended_data$daly_residual[
    detrended_data$entity == "United States"
  ],
  main = "United States: DALY Residuals"
)

acf(
  detrended_data$death_residual[
    detrended_data$entity == "United States"
  ],
  main = "United States: Death-Rate Residuals"
)
```

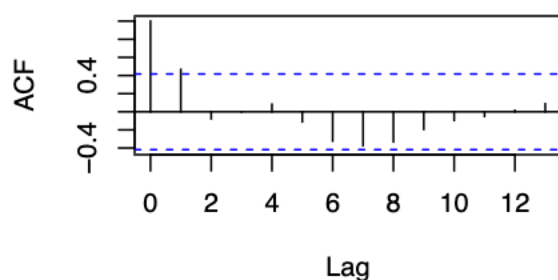
### Canada: DALY Residuals



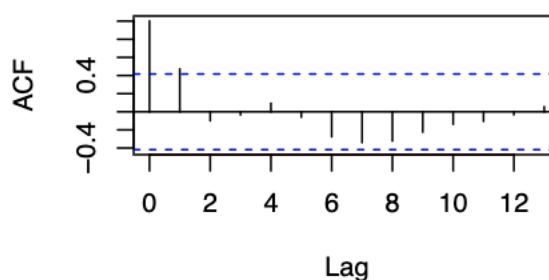
### Canada: Death-Rate Residuals



### United States: DALY Residuals



### United States: Death-Rate Residuals



```
par(mfrow = c(1, 1))
```

### Interpretation

The Canadian residuals show no clear autocorrelation after removing the time trend.

For the United States, the lag-1 correlations are close to or slightly above the confidence limits, suggesting that some year-to-year dependence may remain. The adjusted correlations should therefore be treated mainly as descriptive evidence rather than as fully independent inferential results.

### Third Indicator: Suicide Rate

To broaden the analysis, I added the suicide rate as a third health indicator.

#### Does the Suicide-Rate Indicator Cover the Same Countries and Years?

Before combining the suicide-rate data with the existing indicators, I checked whether Canada and the United States both had complete coverage over the 2000–2021 study period.

```
# Import the suicide-rate data
suicide_data <- readr::read_csv(
  "https://ourworldindata.org/grapher/death-rate-from-suicides-ghe.csv?v=1&csvType=full&useColumnShortN
)

# Rename variables for consistency
suicide_data <- suicide_data %>%
  dplyr::rename(
    entity = Entity,
    code = Code,
    year = Year,
```

```

    suicide_rate =
      `Death rate from self-harm among both sexes`
  )

# Keep only Canada and the United States
suicide_can_us <- suicide_data %>%
  dplyr::filter(
    entity %in% c("Canada", "United States")
  )

# Summarize coverage by country
suicide_coverage <- suicide_can_us %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    Observations = dplyr::n(),
    First_Year = min(year),
    Last_Year = max(year),
    .groups = "drop"
  )

# Display the coverage table
knitr::kable(
  suicide_coverage,
  col.names = c(
    "Country",
    "Observations",
    "First Year",
    "Last Year"
  ),
  caption = "Coverage of the Suicide-Rate Indicator"
)

```

Table 14: Coverage of the Suicide-Rate Indicator

| Country       | Observations | First Year | Last Year |
|---------------|--------------|------------|-----------|
| Canada        | 22           | 2000       | 2021      |
| United States | 22           | 2000       | 2021      |

## Interpretation

Canada and the United States each have 22 suicide-rate observations covering 2000–2021. The third indicator therefore matches the country and time coverage of the existing dataset and can be added directly.

## Did the Suicide-Rate Indicator Join Correctly?

Because the suicide-rate data cover the same countries and years, I added them to the existing dataset using country, code, and year. I then checked whether all 44 observations matched correctly.

```

# Add the suicide-rate indicator
combined_data <- combined_data %>%
  dplyr::left_join(
    suicide_can_us,
    by = c("entity", "code", "year")
  )

```

```

# Check whether the join was complete
suicide_join_check <- tibble(
  Check = c(
    "Expected observations",
    "Observations after join",
    "Missing suicide-rate values"
  ),
  Result = c(
    44,
    nrow(combined_data),
    sum(is.na(combined_data$suicide_rate))
  )
)

# Display the join check
knitr::kable(
  suicide_join_check,
  col.names = c("Check", "Result"),
  caption = "Validation of the Suicide-Rate Join"
)

```

Table 15: Validation of the Suicide-Rate Join

| Check                       | Result |
|-----------------------------|--------|
| Expected observations       | 44     |
| Observations after join     | 44     |
| Missing suicide-rate values | 0      |

## Interpretation

The combined dataset still contains all 44 expected observations, and no suicide-rate values are missing. This confirms that the third indicator matched correctly by country and year.

## How Does the Suicide Rate Compare Between the Two Countries?

Before using all three indicators together, I summarized the suicide rate for Canada and the United States.

```

# Calculate summary statistics for suicide rate
suicide_summary <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    mean_rate = mean(suicide_rate),
    median_rate = median(suicide_rate),
    sd_rate = sd(suicide_rate),
    min_rate = min(suicide_rate),
    max_rate = max(suicide_rate),
    .groups = "drop"
  )

# Display the summary statistics
knitr::kable(
  suicide_summary,
  digits = 2,

```

```
col.names = c(
  "Country",
  "Mean",
  "Median",
  "SD",
  "Minimum",
  "Maximum"
),
caption = "Summary Statistics for Suicide Rates"
)
```

Table 16: Summary Statistics for Suicide Rates

| Country       | Mean  | Median | SD   | Minimum | Maximum |
|---------------|-------|--------|------|---------|---------|
| Canada        | 12.36 | 12.51  | 0.83 | 9.44    | 13.42   |
| United States | 13.20 | 13.20  | 1.43 | 11.03   | 15.46   |

## Interpretation

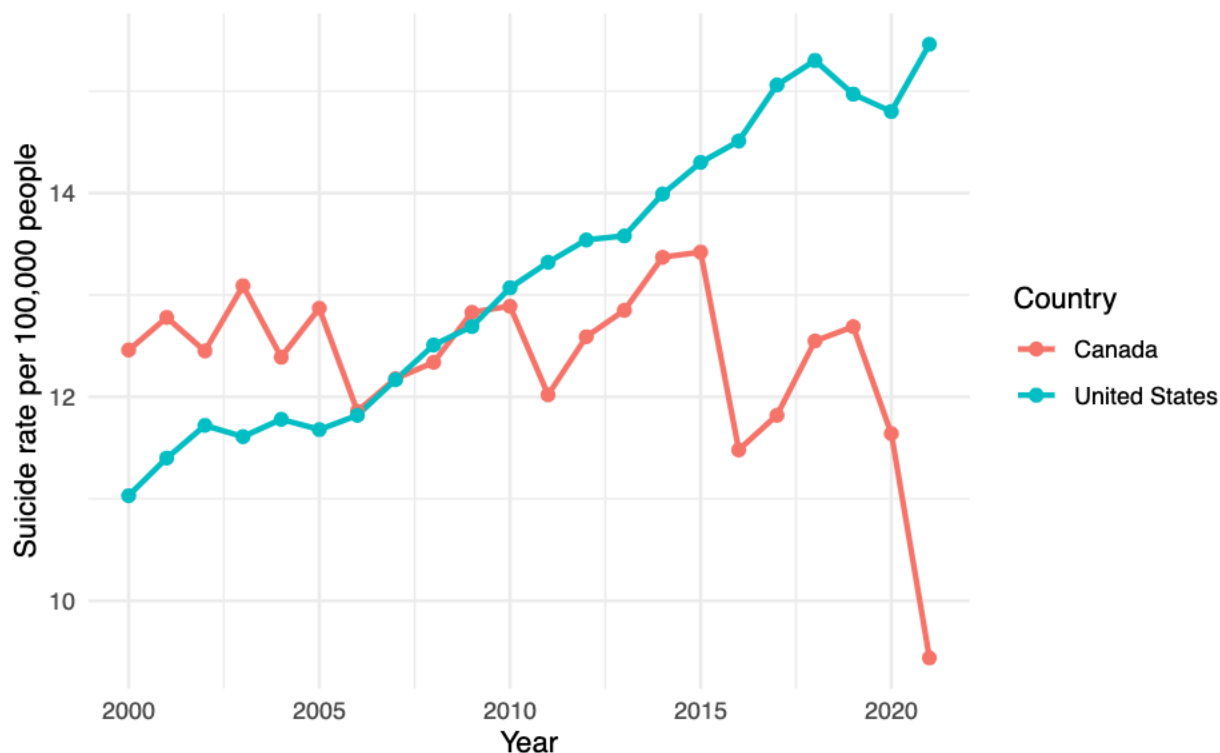
The United States had a slightly higher average suicide rate than Canada (13.20 versus 12.36 per 100,000). It also showed greater variation over the study period, with a standard deviation of 1.43 compared with 0.83 in Canada.

## How Did Suicide Rates Change Over Time?

I plotted the yearly suicide rates to compare how the pattern changed in Canada and the United States from 2000 to 2021.

```
# Plot suicide rates over time
ggplot(
  combined_data,
  aes(x = year, y = suicide_rate, color = entity)
) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Suicide Rates in Canada and the United States",
    subtitle = "2000-2021",
    x = "Year",
    y = "Suicide rate per 100,000 people",
    color = "Country"
  ) +
  theme_minimal()
```

## Suicide Rates in Canada and the United States 2000–2021



### Interpretation

Suicide rates followed different patterns in the two countries. The United States showed a gradual increase over most of the period and moved above Canada after the late 2000s.

Canada showed more year-to-year variation, and the sharp decline in 2021 stands out as an unusual change that should be examined before including suicide rate in the multivariate analysis.

### Is Canada's 2021 Suicide Rate Unusual?

The time-series plot shows a sharp decline in Canada's suicide rate in 2021. I first compared the most recent year-to-year changes to see how large this decline was relative to nearby years.

```
# Calculate yearly changes in Canada's suicide rate
canada_suicide <- combined_data %>%
  dplyr::filter(entity == "Canada") %>%
  dplyr::arrange(year) %>%
  dplyr::mutate(
    yearly_change = suicide_rate - dplyr::lag(suicide_rate),
    percent_change = 100 * (
      suicide_rate / dplyr::lag(suicide_rate) - 1
    )
  )

# Select recent years for display
recent_changes <- canada_suicide %>%
  dplyr::filter(year >= 2016) %>%
  dplyr::select(
```

```

    year,
    suicide_rate,
    yearly_change,
    percent_change
  )

# Display recent changes
knitr::kable(
  recent_changes,
  digits = 2,
  col.names = c(
    "Year",
    "Suicide Rate",
    "Yearly Change",
    "Change (%)"
  ),
  caption = "Recent Changes in Canada's Suicide Rate"
)

```

Table 17: Recent Changes in Canada's Suicide Rate

| Year | Suicide Rate | Yearly Change | Change (%) |
|------|--------------|---------------|------------|
| 2016 | 11.48        | -1.94         | -14.46     |
| 2017 | 11.82        | 0.34          | 2.96       |
| 2018 | 12.55        | 0.73          | 6.18       |
| 2019 | 12.69        | 0.14          | 1.12       |
| 2020 | 11.64        | -1.05         | -8.27      |
| 2021 | 9.44         | -2.20         | -18.90     |

## Interpretation

Canada's suicide rate fell from 11.64 in 2020 to 9.44 in 2021, a decrease of about 18.9%. This is a large one-year decline, so the next step is to compare it with changes across the full 2000–2021 series.

## Is the 2021 Decline Unusual Compared With the Full Series?

To place the 2021 decline in context, I compared the yearly percentage changes in Canada's suicide rate across the full 2000–2021 period.

```

# Compare yearly percentage changes across the full series
canada_change_table <- canada_suicide %>%
  dplyr::filter(!is.na(percent_change)) %>%
  dplyr::arrange(percent_change) %>%
  dplyr::select(
    year,
    suicide_rate,
    yearly_change,
    percent_change
  )

# Display the yearly changes
knitr::kable(
  canada_change_table,
  digits = 2,

```

```
col.names = c(
  "Year",
  "Suicide Rate",
  "Yearly Change",
  "Change (%)"
),
caption = "Yearly Changes in Canada's Suicide Rate, 2001-2021"
)
```

Table 18: Yearly Changes in Canada's Suicide Rate, 2001-2021

| Year | Suicide Rate | Yearly Change | Change (%) |
|------|--------------|---------------|------------|
| 2021 | 9.44         | -2.20         | -18.90     |
| 2016 | 11.48        | -1.94         | -14.46     |
| 2020 | 11.64        | -1.05         | -8.27      |
| 2006 | 11.86        | -1.01         | -7.85      |
| 2011 | 12.02        | -0.87         | -6.75      |
| 2004 | 12.39        | -0.70         | -5.35      |
| 2002 | 12.45        | -0.33         | -2.58      |
| 2015 | 13.42        | 0.05          | 0.37       |
| 2010 | 12.89        | 0.06          | 0.47       |
| 2019 | 12.69        | 0.14          | 1.12       |
| 2008 | 12.34        | 0.16          | 1.31       |
| 2013 | 12.85        | 0.26          | 2.07       |
| 2001 | 12.78        | 0.32          | 2.57       |
| 2007 | 12.18        | 0.32          | 2.70       |
| 2017 | 11.82        | 0.34          | 2.96       |
| 2005 | 12.87        | 0.48          | 3.87       |
| 2009 | 12.83        | 0.49          | 3.97       |
| 2014 | 13.37        | 0.52          | 4.05       |
| 2012 | 12.59        | 0.57          | 4.74       |
| 2003 | 13.09        | 0.64          | 5.14       |
| 2018 | 12.55        | 0.73          | 6.18       |

## Interpretation

The 2021 decline was the largest one-year decrease in Canada's suicide rate during the study period, at approximately 18.9%.

The next largest decline occurred in 2016 at about 14.5%, so 2021 stands out as an unusual observation and its influence should be checked before suicide rate is used in the multivariate analysis.

## Are the Three Indicators on Comparable Scales?

The three health indicators measure different outcomes and may have very different numerical ranges. I therefore compared their means and standard deviations before combining them in a multivariate analysis.

```
# Compare the scale of the three indicators
scale_summary <- combined_data %>%
  dplyr::summarise(
    `Drug-Use DALY Mean` = mean(daly_rate),
    `Drug-Use DALY SD` = sd(daly_rate),
    `Mental-Health Death Mean` = mean(mental_health_death_rate),
    `Mental-Health Death SD` = sd(mental_health_death_rate),
```

```

`Suicide Mean` = mean(suicide_rate),
`Suicide SD` = sd(suicide_rate)
) %>%
tidyr::pivot_longer(
  everything(),
  names_to = "Measure",
  values_to = "Value"
)

# Display the scale comparison
knitr::kable(
  scale_summary,
  digits = 2,
  col.names = c(
    "Measure",
    "Value"
  ),
  caption = "Scale Comparison of the Three Health Indicators"
)

```

Table 19: Scale Comparison of the Three Health Indicators

| Measure                  | Value  |
|--------------------------|--------|
| Drug-Use DALY Mean       | 914.89 |
| Drug-Use DALY SD         | 498.53 |
| Mental-Health Death Mean | 13.53  |
| Mental-Health Death SD   | 7.47   |
| Suicide Mean             | 12.78  |
| Suicide SD               | 1.23   |

## Interpretation

The three indicators are on very different numerical scales. Drug-use DALY rate has much greater variability ( $SD = 498.53$ ) than the mental-health death rate ( $SD = 7.47$ ) or suicide rate ( $SD = 1.23$ ).

Without standardization, the DALY variable could dominate the multivariate analysis simply because of its larger scale. I will therefore standardize the indicators before combining them in methods such as principal component analysis.

## How Are the Three Health Indicators Related to One Another?

Before combining the indicators in a multivariate model, I examined their pairwise relationships within each country. This helps identify whether some variables contain very similar information.

```

# Calculate pairwise correlations within each country
three_indicator_correlations <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    `DALY vs Mental-Health Death` =
      cor(daly_rate, mental_health_death_rate),

    `DALY vs Suicide` =
      cor(daly_rate, suicide_rate),

```

```

`Mental-Health Death vs Suicide` =
  cor(mental_health_death_rate, suicide_rate),

  .groups = "drop"
)

# Display the correlations
knitr::kable(
  three_indicator_correlations,
  digits = 3,
  col.names = c(
    "Country",
    "DALY vs Mental-Health Death",
    "DALY vs Suicide",
    "Mental-Health Death vs Suicide"
  ),
  caption = "Pairwise Correlations Among the Three Health Indicators"
)

```

Table 20: Pairwise Correlations Among the Three Health Indicators

| Country       | DALY vs Mental-Health Death | DALY vs Suicide | Mental-Health Death vs Suicide |
|---------------|-----------------------------|-----------------|--------------------------------|
| Canada        | 0.998                       | -0.619          | -0.645                         |
| United States | 0.992                       | 0.927           | 0.885                          |

## Interpretation

The three indicators do not have the same relationship in Canada and the United States.

In Canada, drug-use DALY rate and mental-health and substance-use disorder death rate are almost perfectly positively correlated ( $r = 0.998$ ). In contrast, suicide rate is moderately negatively correlated with both drug-use DALY rate ( $r = -0.619$ ) and the mental-health death rate ( $r = -0.645$ ).

In the United States, all three indicators are strongly positively related. Drug-use DALY rate is strongly correlated with both the mental-health death rate ( $r = 0.992$ ) and suicide rate ( $r = 0.927$ ), while the correlation between the two death-rate indicators is also high ( $r = 0.885$ ).

These contrasting patterns suggest that the indicators do not move together in the same way in the two countries. However, because all three variables are measured repeatedly over time, the raw correlations may partly reflect their underlying time trends. I therefore examine the relationships again after accounting for time before using the indicators together in the multivariate analysis.

## Do the Relationships Among the Three Indicators Remain After Accounting for Time?

Because the three indicators change over time, the raw correlations may partly reflect shared time trends. I therefore removed the time trend from each indicator within each country before examining the relationships again.

```

# Create detrended residuals for all three indicators
three_indicator_detrended <- combined_data %>%
  dplyr::group_by(entity) %>%
  dplyr::mutate(
    daly_residual = residuals(
      lm(
        daly_rate ~ year_since_2000 + I(year_since_2000^2)

```

```

    )
  ),

  mental_health_residual = residuals(
    lm(
      mental_health_death_rate ~
        year_since_2000 + I(year_since_2000^2)
    )
  ),

  suicide_residual = residuals(
    lm(
      suicide_rate ~ year_since_2000 + I(year_since_2000^2)
    )
  )
) %>%
dplyr::ungroup()

# Calculate pairwise correlations after removing time trends
detrended_correlations <- three_indicator_detrended %>%
dplyr::group_by(entity) %>%
dplyr::summarise(
  `DALY vs Mental-Health Death` =
    cor(daly_residual, mental_health_residual),

  `DALY vs Suicide` =
    cor(daly_residual, suicide_residual),

  `Mental-Health Death vs Suicide` =
    cor(mental_health_residual, suicide_residual),

  .groups = "drop"
)

# Display the adjusted correlations
knitr::kable(
  detrended_correlations,
  digits = 3,
  col.names = c(
    "Country",
    "DALY vs Mental-Health Death",
    "DALY vs Suicide",
    "Mental-Health Death vs Suicide"
  ),
  caption = "Pairwise Correlations After Removing Time Trends"
)

```

Table 21: Pairwise Correlations After Removing Time Trends

| Country       | DALY vs Mental-Health Death | DALY vs Suicide | Mental-Health Death vs Suicide |
|---------------|-----------------------------|-----------------|--------------------------------|
| Canada        | 0.976                       | -0.577          | -0.657                         |
| United States | 0.968                       | -0.485          | -0.527                         |

## Interpretation

After removing the time trends, drug-use DALY rate and mental-health and substance-use disorder death rate remained very strongly positively correlated in both Canada ( $r = 0.976$ ) and the United States ( $r = 0.968$ ).

The relationship with suicide rate was different. In Canada, suicide rate remained negatively related to both drug-use DALY rate ( $r = -0.577$ ) and the mental-health death rate ( $r = -0.657$ ). In the United States, the correlations with suicide rate also became negative after detrending ( $r = -0.485$  and  $r = -0.527$ ).

The change from strong positive raw correlations with suicide rate in the United States to moderate negative correlations after detrending shows that the original positive relationships were strongly influenced by their shared upward trends over time. This makes the detrended correlation structure more informative for deciding how the three indicators should be used together in the multivariate analysis.

## Are the Detrended Residuals Independent Across Years?

Before using the detrended relationships in the multivariate analysis, I checked whether the remaining residuals still showed autocorrelation. If strong year-to-year dependence remains, the adjusted relationships should be interpreted more cautiously.

```
# Check autocorrelation for all three detrended indicators
par(mfrow = c(3, 2))

acf(
  three_indicator_detrended$daly_residual[
    three_indicator_detrended$entity == "Canada"
  ],
  main = "Canada: DALY Residuals"
)

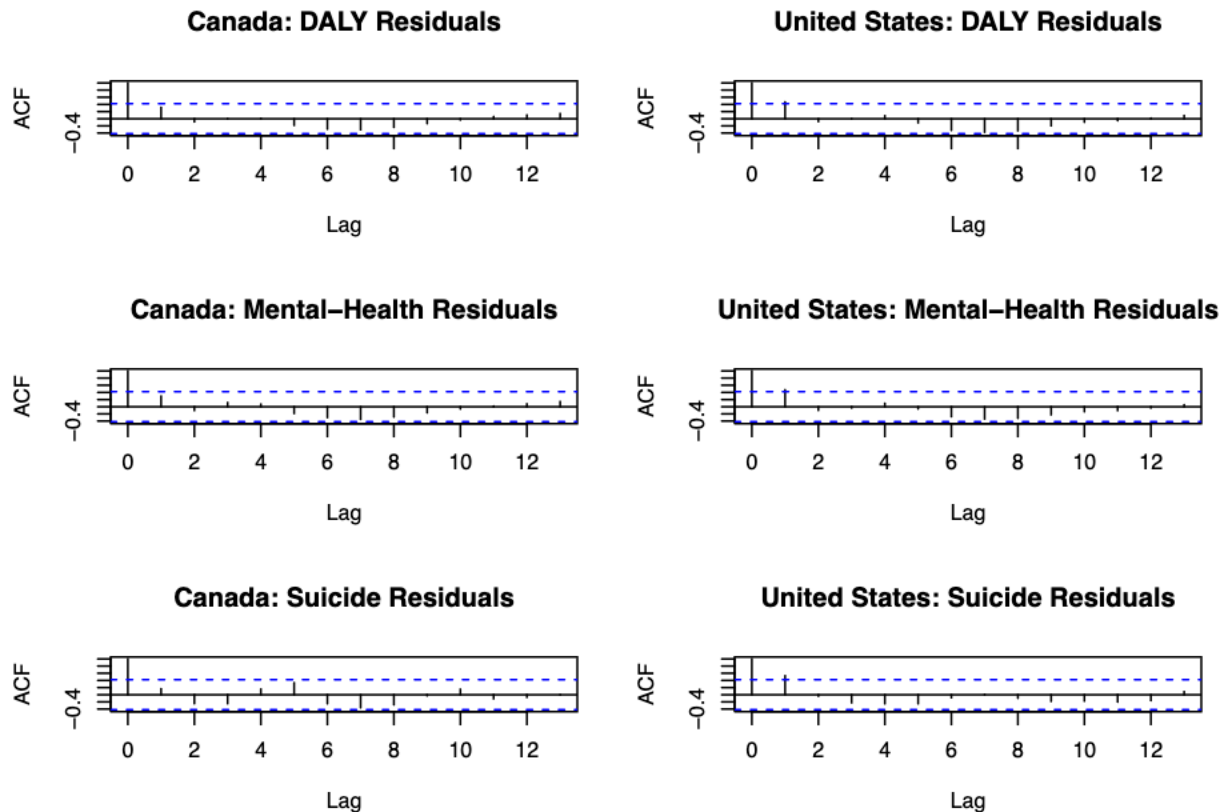
acf(
  three_indicator_detrended$daly_residual[
    three_indicator_detrended$entity == "United States"
  ],
  main = "United States: DALY Residuals"
)

acf(
  three_indicator_detrended$mental_health_residual[
    three_indicator_detrended$entity == "Canada"
  ],
  main = "Canada: Mental-Health Residuals"
)

acf(
  three_indicator_detrended$mental_health_residual[
    three_indicator_detrended$entity == "United States"
  ],
  main = "United States: Mental-Health Residuals"
)

acf(
  three_indicator_detrended$suicide_residual[
    three_indicator_detrended$entity == "Canada"
  ],
  main = "Canada: Suicide Residuals"
)
```

```
)
acf(
  three_indicator_detrended$suicide_residual[
    three_indicator_detrended$entity == "United States"
  ],
  main = "United States: Suicide Residuals"
)
```



```
par(mfrow = c(1, 1))
```

## Interpretation

After removing the quadratic time trends, none of the non-zero autocorrelations clearly exceed the approximate confidence limits for the three indicators in either country.

The United States shows somewhat larger lag-1 correlations for some residual series, but these remain within the confidence bounds. Overall, there is no strong evidence of remaining year-to-year autocorrelation after detrending.

This supports using the detrended relationships as a more appropriate description of how the three indicators move together beyond their common time trends.

## How Do Canada and the United States Compare Across the Three Indicators on a Common Scale?

Because the three indicators have very different numerical scales, I standardized each variable before comparing their overall country profiles. Standardization expresses each value in standard-deviation units, allowing the indicators to be compared on the same scale.

```

# Standardize the three health indicators
standardized_data <- combined_data %>%
  dplyr::mutate(
    daly_z = as.numeric(scale(daly_rate)),
    mental_health_z = as.numeric(scale(mental_health_death_rate)),
    suicide_z = as.numeric(scale(suicide_rate))
  )

# Calculate the average standardized value for each country
standardized_profile <- standardized_data %>%
  dplyr::group_by(entity) %>%
  dplyr::summarise(
    `Drug-Use DALY` = mean(daly_z),
    `Mental-Health Death Rate` = mean(mental_health_z),
    `Suicide Rate` = mean(suicide_z),
    .groups = "drop"
  )

# Display the standardized country profiles
knitr::kable(
  standardized_profile,
  digits = 3,
  col.names = c(
    "Country",
    "Drug-Use DALY",
    "Mental-Health Death Rate",
    "Suicide Rate"
  ),
  caption = "Average Standardized Health Indicators by Country"
)

```

Table 22: Average Standardized Health Indicators by Country

| Country       | Drug-Use DALY | Mental-Health Death Rate | Suicide Rate |
|---------------|---------------|--------------------------|--------------|
| Canada        | -0.514        | -0.452                   | -0.337       |
| United States | 0.514         | 0.452                    | 0.337        |

## Interpretation

After standardization, the United States had higher average values across all three health indicators, while Canada had lower average values.

The largest standardized separation between the countries was observed for drug-use DALY rate ( $\pm 0.514$ ), followed by the mental-health and substance-use disorder death rate ( $\pm 0.452$ ) and suicide rate ( $\pm 0.337$ ).

This shows that the country difference is present across all three indicators even after placing them on a common scale, with the strongest separation occurring in drug-use burden.

## Can the Three Standardized Indicators Be Summarized by Fewer Components?

Because the three indicators are related, I used principal component analysis (PCA) to determine whether most of their variation can be summarized by a smaller number of components. The indicators have already been standardized, so each variable contributes on a comparable scale.

```

# Select the standardized indicators for PCA
pca_data <- standardized_data %>%
  dplyr::select(
    daly_z,
    mental_health_z,
    suicide_z
  )

# Fit the principal component analysis
pca_model <- prcomp(
  pca_data,
  center = FALSE,
  scale. = FALSE
)

# Calculate the variance explained by each component
pca_variance <- tibble(
  Component = c("PC1", "PC2", "PC3"),
  Eigenvalue = pca_model$sdev^2,
  `Variance Explained (%)` =
    100 * (pca_model$sdev^2 / sum(pca_model$sdev^2)),
  `Cumulative Variance (%)` =
    100 * cumsum(pca_model$sdev^2 / sum(pca_model$sdev^2))
)

# Display the PCA variance table
knitr::kable(
  pca_variance,
  digits = 2,
  caption = "Variance Explained by the Principal Components"
)

```

Table 23: Variance Explained by the Principal Components

| Component | Eigenvalue | Variance Explained (%) | Cumulative Variance (%) |
|-----------|------------|------------------------|-------------------------|
| PC1       | 2.46       | 81.95                  | 81.95                   |
| PC2       | 0.54       | 17.90                  | 99.85                   |
| PC3       | 0.00       | 0.15                   | 100.00                  |

## Interpretation

The first principal component explains 81.95% of the total variation in the three standardized health indicators, so most of the multivariate structure can be summarized along a single dominant dimension.

The second component explains an additional 17.90%, bringing the cumulative variance explained by the first two components to 99.85%. In contrast, PC3 explains only 0.15% of the variation, indicating that very little additional information remains in the third dimension.

Before deciding what these components represent, the next step is to examine the PCA loadings to determine which indicators contribute most strongly to each component.

## Which Indicators Define Each Principal Component?

To understand what each principal component represents, I examined the PCA loadings. The loadings show how strongly each standardized health indicator contributes to each component.

```
# Extract the PCA loadings
pca_loadings <- as.data.frame(pca_model$rotation) %>%
  tibble::rownames_to_column("Indicator") %>%
  dplyr::mutate(
    Indicator = dplyr::recode(
      Indicator,
      daly_z = "Drug-Use DALY Rate",
      mental_health_z = "Mental-Health Death Rate",
      suicide_z = "Suicide Rate"
    )
  )

# Display the PCA loadings
knitr::kable(
  pca_loadings,
  digits = 3,
  col.names = c(
    "Indicator",
    "PC1",
    "PC2",
    "PC3"
  ),
  caption = "Principal Component Loadings"
)
```

Table 24: Principal Component Loadings

| Indicator                | PC1   | PC2    | PC3    |
|--------------------------|-------|--------|--------|
| Drug-Use DALY Rate       | 0.625 | -0.264 | 0.735  |
| Mental-Health Death Rate | 0.604 | -0.433 | -0.669 |
| Suicide Rate             | 0.494 | 0.862  | -0.111 |

## Interpretation

PC1 has positive loadings for all three indicators, with the largest contributions from drug-use DALY rate (0.625) and mental-health death rate (0.604), followed by suicide rate (0.494). Because PC1 also explains 81.95% of the total variance, it can be interpreted as a broad dimension of overall mental-health and substance-use burden.

PC2 is driven mainly by suicide rate (0.862), while drug-use DALY rate (-0.264) and mental-health death rate (-0.433) load in the opposite direction. This component therefore captures a contrast between suicide patterns and the other two indicators that is not represented by the general burden dimension.

PC3 explains only 0.15% of the total variance, so its loadings provide little additional information and do not need substantive interpretation.

## Do Canada and the United States Follow Different Multivariate Trajectories Over Time?

The PCA loadings describe what the components represent, but they do not show where individual country-year observations lie. I therefore examined the PC1 and PC2 scores to see how the multivariate health profiles of Canada and the United States changed from 2000 to 2021.

```

# Extract PC1 and PC2 scores
pca_scores <- as.data.frame(pca_model$x[, 1:2]) %>%
  dplyr::bind_cols(
    standardized_data %>%
      dplyr::select(entity, year)
  ) %>%
  dplyr::arrange(entity, year)

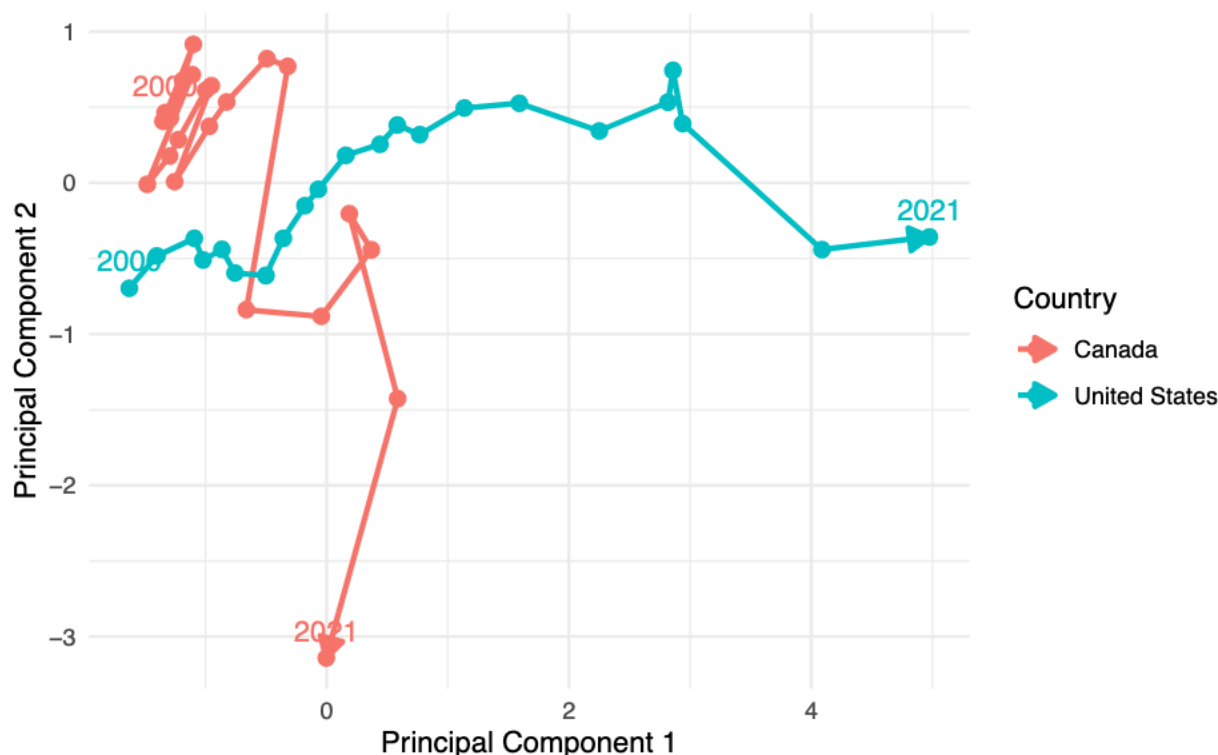
# Select the first and last year for labels
trajectory_labels <- pca_scores %>%
  dplyr::filter(year %in% c(2000, 2021))

# Plot each country's movement through PCA space
ggplot(
  pca_scores,
  aes(
    x = PC1,
    y = PC2,
    color = entity,
    group = entity
  )
) +
  geom_path(
    linewidth = 1,
    arrow = arrow(
      type = "closed",
      length = grid::unit(0.12, "inches")
    )
  ) +
  geom_point(size = 2.5) +
  geom_text(
    data = trajectory_labels,
    aes(label = year),
    vjust = -0.8,
    show.legend = FALSE
  ) +
  labs(
    title = "Multivariate Health Trajectories in PCA Space",
    subtitle = "Canada and United States, 2000-2021",
    x = "Principal Component 1",
    y = "Principal Component 2",
    color = "Country"
  ) +
  theme_minimal()

```

## Multivariate Health Trajectories in PCA Space

### Canada and United States, 2000–2021



### Interpretation

The two countries follow clearly different trajectories through the multivariate health space.

The United States moves strongly toward higher PC1 values over time. Because PC1 represents a broad mental-health and substance-use burden, this movement indicates that the overall combined burden increased substantially in the United States, particularly in the later years.

Canada shows a much smaller shift along PC1, but its 2021 observation moves sharply downward on PC2. Since PC2 is driven mainly by suicide rate, this unusual movement is consistent with the large decline in Canada's suicide rate observed in 2021.

The trajectories therefore show that the growing difference between the countries is not limited to a single indicator. Their overall multivariate health profiles became increasingly different over the study period, while Canada's unusual 2021 suicide rate mainly affected the secondary PCA dimension.

### Did the Multivariate Distance Between Canada and the United States Increase Over Time?

The PCA trajectory suggests that the overall health profiles of Canada and the United States became more separated over time. To quantify this separation directly, I calculated the Euclidean distance between the two countries each year using all three standardized indicators.

Because the variables are standardized, each indicator contributes on a comparable scale to the distance measure.

```
# Create Canada's standardized yearly profile
canada_profile <- standardized_data %>%
  dplyr::filter(entity == "Canada") %>%
  dplyr::select(
    year,
```

```

    daly_z,
    mental_health_z,
    suicide_z
  )

# Create the United States standardized yearly profile
us_profile <- standardized_data %>%
  dplyr::filter(entity == "United States") %>%
  dplyr::select(
    year,
    daly_z,
    mental_health_z,
    suicide_z
  )

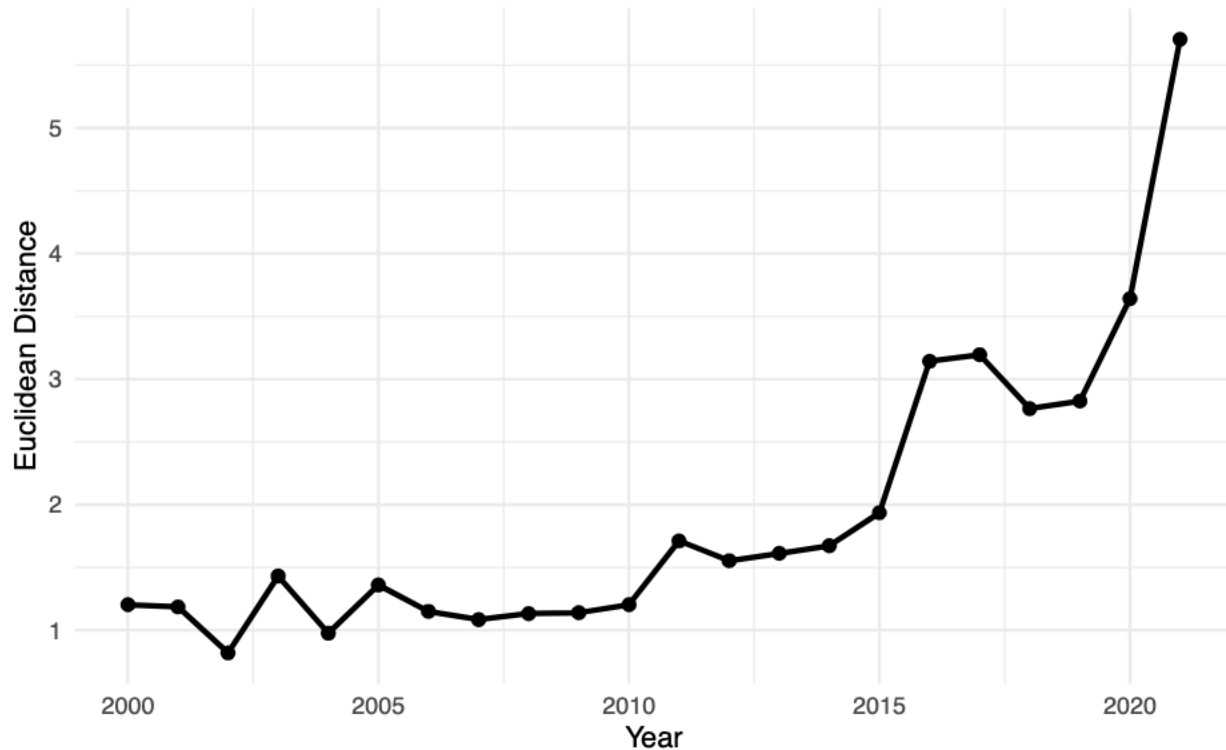
# Match both countries by year and calculate multivariate distance
multivariate_distance <- canada_profile %>%
  dplyr::inner_join(
    us_profile,
    by = "year",
    suffix = c("_canada", "_us")
  ) %>%
  dplyr::mutate(
    distance = sqrt(
      (daly_z_us - daly_z_canada)^2 +
      (mental_health_z_us - mental_health_z_canada)^2 +
      (suicide_z_us - suicide_z_canada)^2
    )
  )

# Plot the multivariate distance over time
ggplot(
  multivariate_distance,
  aes(x = year, y = distance)
) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "Multivariate Distance Between Canada and the United States",
    subtitle = "Based on three standardized health indicators, 2000-2021",
    x = "Year",
    y = "Euclidean Distance"
  ) +
  theme_minimal()

```

## Multivariate Distance Between Canada and the United States

Based on three standardized health indicators, 2000–2021



### Interpretation

The multivariate distance between Canada and the United States remained relatively stable from 2000 through about 2010, indicating that their combined health profiles were fairly similar during the earlier part of the study period.

The separation increased after 2010 and became much larger after 2015. By 2021, the distance reached its highest value in the series, showing that the two countries had become substantially more different when all three standardized indicators were considered together.

This provides multivariate evidence that the Canada–United States difference widened over time rather than remaining constant. Because the distance combines all three indicators, the next step is to determine which indicators contributed most to this growing separation.

### Which Health Indicators Drove the Growing Multivariate Distance?

The overall multivariate distance shows that Canada and the United States became more separated over time, but it does not show which indicators produced that separation. I therefore decomposed the squared Euclidean distance into the contribution from each standardized health indicator.

```
# Calculate each indicator's contribution to the squared distance
distance_contributions <- multivariate_distance %>%
  dplyr::mutate(
    daly_contribution =
      (daly_z_us - daly_z_canada)^2,

    mental_health_contribution =
      (mental_health_z_us - mental_health_z_canada)^2,
```

```

    suicide_contribution =
      (suicide_z_us - suicide_z_canada)^2,

    total_squared_distance =
      daly_contribution +
      mental_health_contribution +
      suicide_contribution
  ) %>%
  dplyr::mutate(
    daly_percent =
      100 * daly_contribution / total_squared_distance,

    mental_health_percent =
      100 * mental_health_contribution / total_squared_distance,

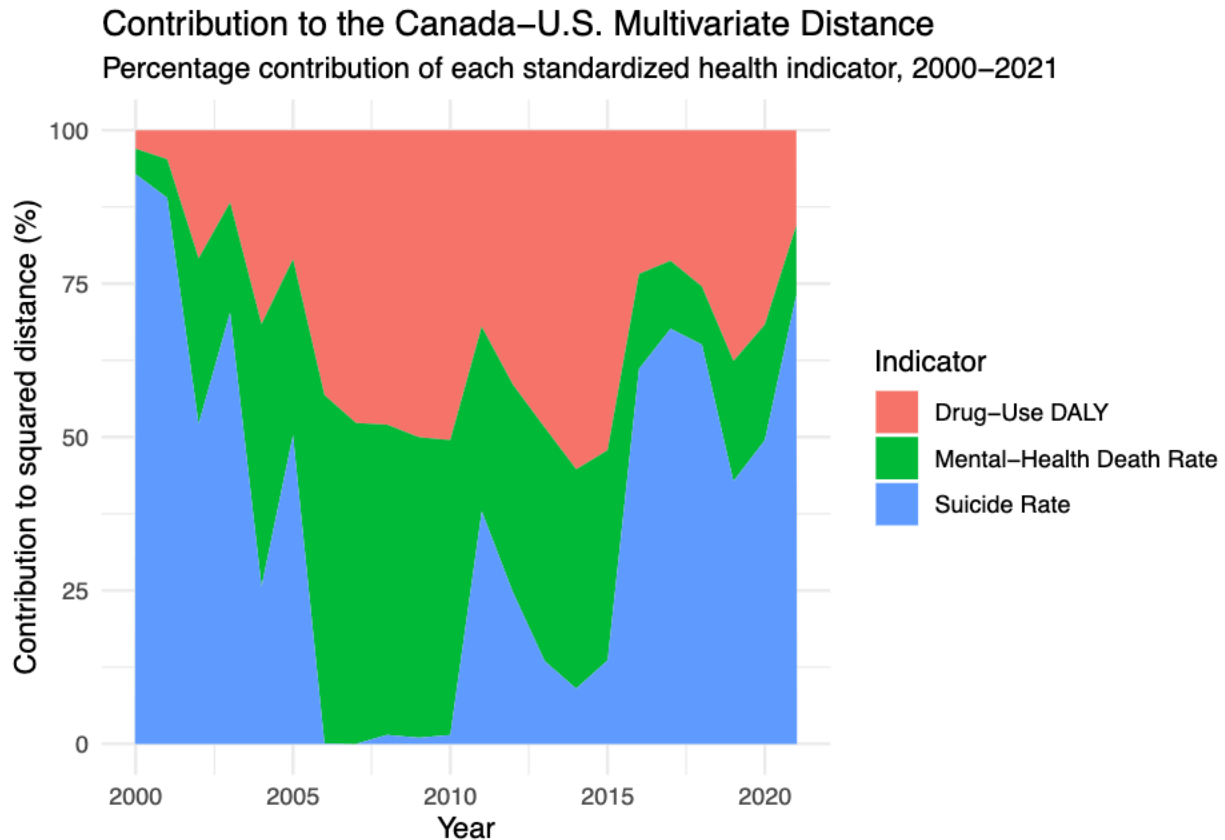
    suicide_percent =
      100 * suicide_contribution / total_squared_distance
  )

# Convert contribution percentages to long format
contribution_plot_data <- distance_contributions %>%
  dplyr::select(
    year,
    daly_percent,
    mental_health_percent,
    suicide_percent
  ) %>%
  tidyr::pivot_longer(
    cols = -year,
    names_to = "Indicator",
    values_to = "Contribution"
  ) %>%
  dplyr::mutate(
    Indicator = dplyr::recode(
      Indicator,
      daly_percent = "Drug-Use DALY",
      mental_health_percent = "Mental-Health Death Rate",
      suicide_percent = "Suicide Rate"
    )
  )

# Plot each indicator's percentage contribution
ggplot(
  contribution_plot_data,
  aes(
    x = year,
    y = Contribution,
    fill = Indicator
  )
) +
  geom_area(
    position = "stack"
  ) +

```

```
labs(
  title = "Contribution to the Canada-U.S. Multivariate Distance",
  subtitle = "Percentage contribution of each standardized health indicator, 2000-2021",
  x = "Year",
  y = "Contribution to squared distance (%)",
  fill = "Indicator"
) +
theme_minimal()
```



```
# Show contributions for selected years
selected_contributions <- distance_contributions %>%
  dplyr::filter(
    year %in% c(2000, 2010, 2015, 2020, 2021)
  ) %>%
  dplyr::select(
    year,
    daly_percent,
    mental_health_percent,
    suicide_percent
  )

knitr::kable(
  selected_contributions,
  digits = 1,
  col.names = c(
    "Year",
    "Drug-Use DALY (%)",
```

```

    "Mental-Health Death (%)",
    "Suicide (%)"
  ),
  caption = "Indicator Contributions to Multivariate Distance in Selected Years"
)

```

Table 25: Indicator Contributions to Multivariate Distance in Selected Years

| Year | Drug-Use DALY (%) | Mental-Health Death (%) | Suicide (%) |
|------|-------------------|-------------------------|-------------|
| 2000 | 3.0               | 4.1                     | 92.8        |
| 2010 | 50.5              | 48.0                    | 1.5         |
| 2015 | 52.2              | 34.3                    | 13.6        |
| 2020 | 31.7              | 18.8                    | 49.5        |
| 2021 | 15.5              | 11.4                    | 73.2        |

## Interpretation

The sources of the Canada–U.S. multivariate separation changed substantially over time rather than being driven by one indicator throughout the study period.

In 2000, suicide rate accounted for 92.8% of the squared distance. By 2010, the separation was instead divided almost equally between drug-use DALY rate (50.5%) and the mental-health death rate (48.0%), with suicide contributing very little.

The pattern changed again in the later years. Suicide accounted for 49.5% of the separation in 2020 and 73.2% in 2021. Because the overall multivariate distance was also largest in 2021, the unusual Canadian suicide-rate decline contributed substantially to the extreme separation observed at the end of the series.

Overall, the widening difference between the countries reflects changing contributions from all three indicators rather than a single persistent source.

## Is the Growing Multivariate Separation Still Present Without Suicide Rate?

Because suicide rate contributed strongly to the multivariate distance in some years, especially 2021, I repeated the distance analysis using only drug-use DALY rate and the mental-health and substance-use disorder death rate.

This sensitivity check shows whether the broader Canada–U.S. separation remains even when suicide rate is excluded.

```

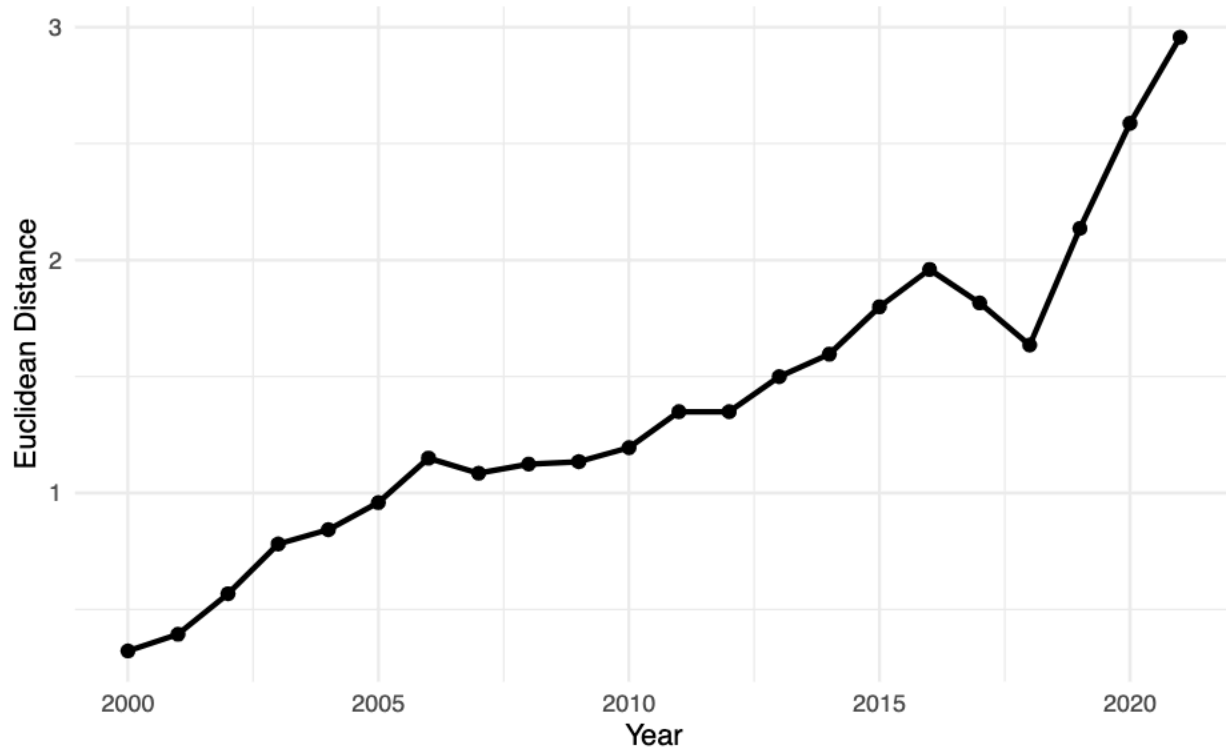
# Calculate the two-indicator distance without suicide rate
distance_without_suicide <- multivariate_distance %>%
  dplyr::mutate(
    distance_without_suicide = sqrt(
      (daly_z_us - daly_z_canada)^2 +
      (mental_health_z_us - mental_health_z_canada)^2
    )
  )

# Plot the distance without suicide rate
ggplot(
  distance_without_suicide,
  aes(x = year, y = distance_without_suicide)
) +
  geom_line(linewidth = 1) +

```

```
geom_point(size = 2) +
labs(
  title = "Canada-U.S. Multivariate Distance Without Suicide Rate",
  subtitle = "Based on standardized drug-use DALY and mental-health death rates, 2000-2021",
  x = "Year",
  y = "Euclidean Distance"
) +
theme_minimal()
```

Canada-U.S. Multivariate Distance Without Suicide Rate  
Based on standardized drug-use DALY and mental-health death rates, 2000-2021



```
# Compare the full and reduced distances in selected years
distance_sensitivity_table <- distance_without_suicide %>%
  dplyr::filter(
    year %in% c(2000, 2010, 2015, 2020, 2021)
  ) %>%
  dplyr::select(
    year,
    distance,
    distance_without_suicide
  )

knitr::kable(
  distance_sensitivity_table,
  digits = 2,
  col.names = c(
    "Year",
    "Three-Indicator Distance",
    "Distance Without Suicide"
  )
)
```

```
),
caption = "Sensitivity of Multivariate Distance to Excluding Suicide Rate"
)
```

Table 26: Sensitivity of Multivariate Distance to Excluding Suicide Rate

| Year | Three-Indicator Distance | Distance Without Suicide |
|------|--------------------------|--------------------------|
| 2000 | 1.20                     | 0.32                     |
| 2010 | 1.20                     | 1.19                     |
| 2015 | 1.94                     | 1.80                     |
| 2020 | 3.64                     | 2.59                     |
| 2021 | 5.71                     | 2.96                     |

### Interpretation

Excluding suicide rate reduces the size of the Canada–U.S. multivariate distance, especially in the later years, but it does not remove the overall pattern of increasing separation.

The two-indicator distance rises from 0.32 in 2000 to 2.96 in 2021, with the largest value still occurring at the end of the study period. This shows that the growing difference between the countries is also strongly supported by drug-use DALY rate and the mental-health and substance-use disorder death rate.

The much larger three-indicator distance in 2021 (5.71 versus 2.96 without suicide) confirms that suicide rate amplified the final-year separation, but it was not solely responsible for the broader multivariate divergence.

### Are the Yearly Multivariate Differences Correlated Over Time?

Before formally testing whether the overall health profiles of Canada and the United States differ, I checked whether the yearly differences in the three standardized indicators still show autocorrelation.

```
# Calculate the standardized U.S.-Canada difference for each indicator
yearly_indicator_differences <- multivariate_distance %>%
  dplyr::mutate(
    daly_difference =
      daly_z_us - daly_z_canada,

    mental_health_difference =
      mental_health_z_us - mental_health_z_canada,

    suicide_difference =
      suicide_z_us - suicide_z_canada
  ) %>%
  dplyr::select(
    year,
    daly_difference,
    mental_health_difference,
    suicide_difference
  )

# Check autocorrelation in the three yearly difference series
par(mfrow = c(3, 1))

acf(
  yearly_indicator_differences$daly_difference,
```

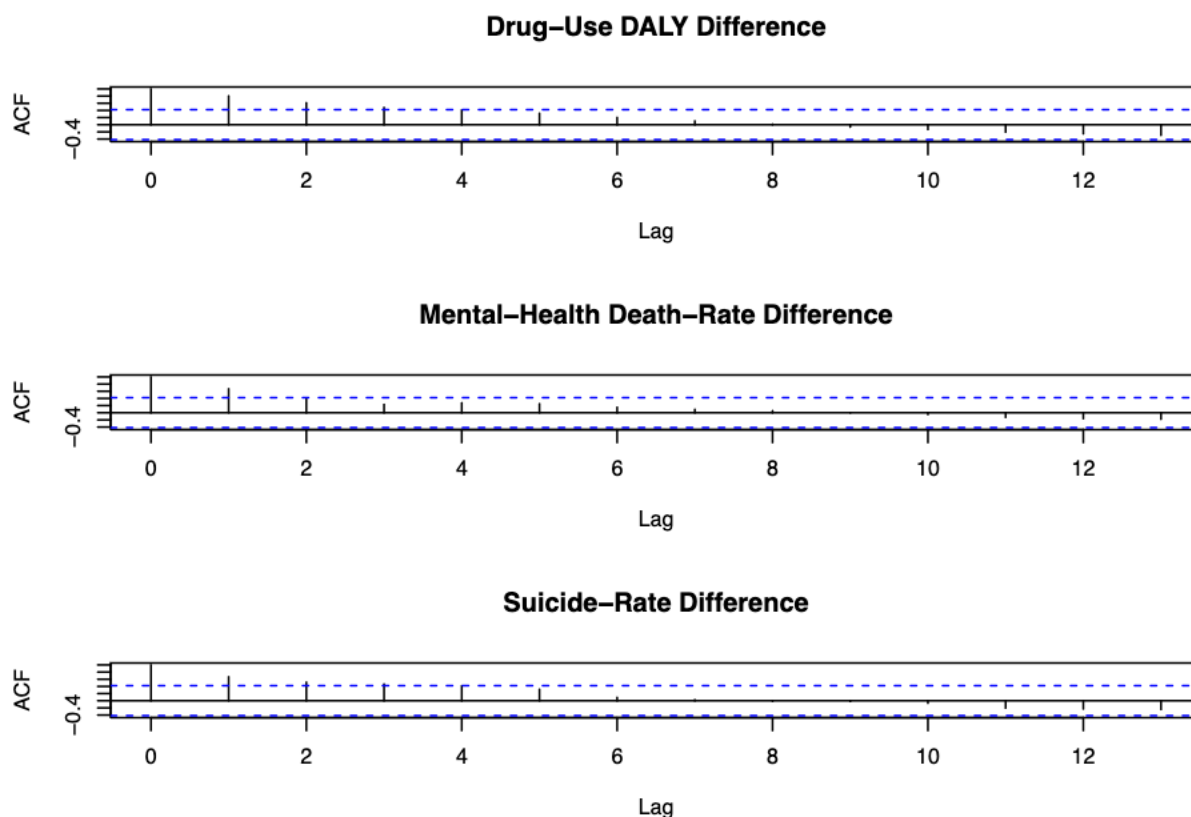
```

    main = "Drug-Use DALY Difference"
  )

  acf(
    yearly_indicator_differences$mental_health_difference,
    main = "Mental-Health Death-Rate Difference"
  )

  acf(
    yearly_indicator_differences$suicide_difference,
    main = "Suicide-Rate Difference"
  )

```



```

par(mfrow = c(1, 1))

```

## Interpretation

Ignoring lag 0, which is always equal to 1, the ACF plots show clear positive autocorrelation in all three yearly country-difference series. The strongest dependence occurs at the first few lags, with several correlations exceeding the approximate confidence limits.

This means that differences observed in one year are related to differences in nearby years. Therefore, the 22 yearly difference vectors should not be treated as independent observations in a standard multivariate test.

Before carrying out formal multivariate inference, the time dependence should be modeled or removed so that the country comparison does not rely on an independence assumption that the data clearly violate.

## Does the Multivariate Country Difference Still Increase After Accounting for Time Dependence?

The yearly multivariate distances are correlated across nearby years, so an ordinary regression would violate the independence assumption. I therefore fitted a quadratic time trend with AR(1) errors, which allows consecutive years to remain correlated while estimating how the overall Canada–U.S. separation changed over time.

```
# Create time variables for the multivariate distance
distance_time_data <- multivariate_distance %>%
  dplyr::mutate(
    year_since_2000 = year - 2000,
    year_squared = year_since_2000^2
  )

# Fit a quadratic trend with AR(1) errors
distance_ar1_model <- stats::arima(
  distance_time_data$distance,
  order = c(1, 0, 0),
  xreg = cbind(
    year_since_2000 = distance_time_data$year_since_2000,
    year_squared = distance_time_data$year_squared
  ),
  include.mean = TRUE,
  method = "ML"
)

# Display the model results
distance_ar1_model

##
## Call:
## stats::arima(x = distance_time_data$distance, order = c(1, 0, 0), xreg = cbind(year_since_2000 = dis
##      year_squared = distance_time_data$year_squared), include.mean = TRUE, method = "ML")
##
## Coefficients:
##          ar1 intercept  year_since_2000  year_squared
##      0.1238    1.4424        -0.1721         0.0153
## s.e.  0.3010    0.2700         0.0617         0.0030
##
## sigma^2 estimated as 0.1722:  log likelihood = -11.87,  aic = 33.74

# Create a clean coefficient table
ar1_coefficients <- tibble(
  Term = names(distance_ar1_model$coef),
  Estimate = as.numeric(distance_ar1_model$coef),
  Standard_Error = sqrt(diag(distance_ar1_model$var.coef))
) %>%
  dplyr::mutate(
    z_value = Estimate / Standard_Error,
    p_value = 2 * pnorm(-abs(z_value))
  )

knitr::kable(
  ar1_coefficients,
  digits = 4,
```

```

col.names = c(
  "Term",
  "Estimate",
  "Standard Error",
  "z",
  "p-value"
),
caption = "Quadratic Trend in Multivariate Distance With AR(1) Errors"
)

```

Table 27: Quadratic Trend in Multivariate Distance With AR(1) Errors

| Term            | Estimate | Standard Error | z       | p-value |
|-----------------|----------|----------------|---------|---------|
| ar1             | 0.1238   | 0.3010         | 0.4113  | 0.6808  |
| intercept       | 1.4424   | 0.2700         | 5.3423  | 0.0000  |
| year_since_2000 | -0.1721  | 0.0617         | -2.7892 | 0.0053  |
| year_squared    | 0.0153   | 0.0030         | 5.1473  | 0.0000  |

### Interpretation

After accounting for AR(1) time dependence, the quadratic time term remains strongly significant ( $\beta = 0.0153$ ,  $p < 0.001$ ). This provides evidence that the multivariate distance between Canada and the United States changed nonlinearly over time and increased more rapidly in the later years.

The negative linear term ( $\beta = -0.1721$ ,  $p = 0.005$ ) together with the positive quadratic term indicates a shallow decline in separation early in the study period, followed by an increasingly strong rise thereafter.

The estimated AR(1) coefficient is small and not statistically significant ( $\phi = 0.124$ ,  $p = 0.681$ ). This suggests that, once the quadratic time trend is included, there is no clear remaining first-order autocorrelation in the multivariate-distance series.

Overall, the increasing Canada–U.S. multivariate separation is not explained simply by year-to-year dependence; the curved upward trend remains after explicitly accounting for temporal correlation.

### Does the AR(1) Quadratic Model Leave Any Residual Time Dependence?

After fitting the quadratic trend with AR(1) errors, I checked whether any meaningful autocorrelation remained in the residuals. A well-specified time-series model should leave residuals with little systematic dependence across years.

```

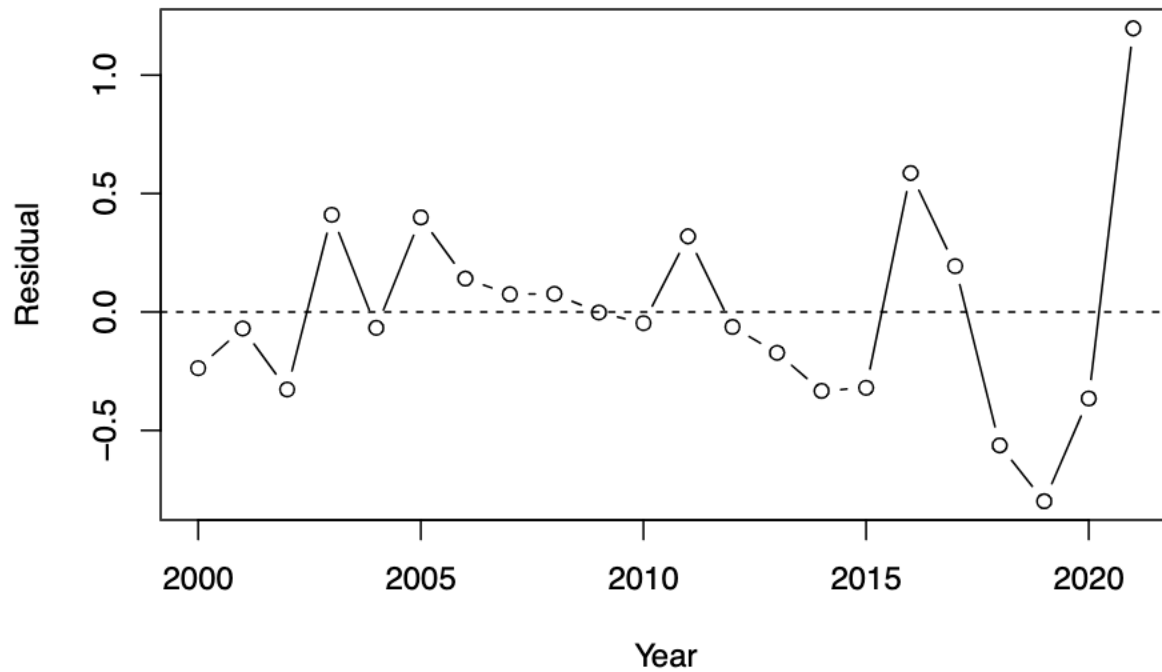
# Extract residuals from the AR(1) quadratic model
distance_ar1_residuals <- residuals(distance_ar1_model)

# Plot the residuals over time
plot(
  distance_time_data$year,
  distance_ar1_residuals,
  type = "b",
  xlab = "Year",
  ylab = "Residual",
  main = "Residuals From the AR(1) Quadratic Distance Model"
)

abline(h = 0, lty = 2)

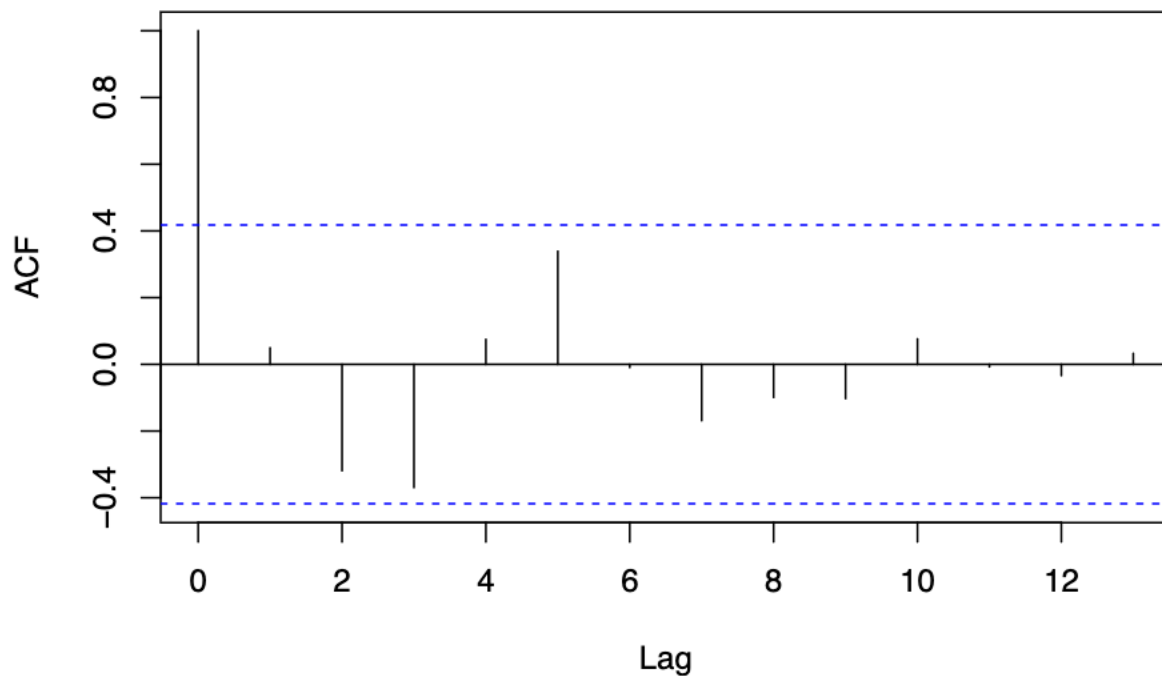
```

## Residuals From the AR(1) Quadratic Distance Model



```
# Check residual autocorrelation
acf(
  distance_ar1_residuals,
  main = "ACF of AR(1) Quadratic Model Residuals"
)
```

## ACF of AR(1) Quadratic Model Residuals



```

# Ljung-Box test for remaining autocorrelation
ljung_box_result <- Box.test(
  distance_ar1_residuals,
  lag = 5,
  type = "Ljung-Box",
  fitdf = 1
)

# Display the test result
ljung_box_table <- tibble(
  Statistic = as.numeric(ljung_box_result$statistic),
  `Degrees of Freedom` = as.numeric(ljung_box_result$parameter),
  `p-value` = ljung_box_result$p.value
)

knitr::kable(
  ljung_box_table,
  digits = 4,
  caption = "Ljung-Box Test for Remaining Residual Autocorrelation"
)

```

Table 28: Ljung-Box Test for Remaining Residual Autocorrelation

| Statistic | Degrees of Freedom | p-value |
|-----------|--------------------|---------|
| 10.2531   | 4                  | 0.0364  |

## Interpretation

The residual ACF does not show any single non-zero lag clearly exceeding the approximate confidence limits. However, the Ljung-Box test gives  $\chi^2 = 10.25$  with  $p = 0.036$ .

Because this p-value is below 0.05, there is evidence that some residual autocorrelation remains when the first five lags are considered jointly. The residual plot also shows a noticeable late-period pattern, including a relatively large positive residual in 2021.

Therefore, the quadratic AR(1) model captures much of the time structure, but the residuals are not yet fully independent. I would not treat this model as the final inferential model. The next step is to test whether a slightly richer error structure, such as AR(2), better accounts for the remaining dependence.

## Does an AR(2) Error Structure Better Capture the Remaining Time Dependence?

The AR(1) quadratic model still showed evidence of residual autocorrelation in the Ljung-Box test. I therefore fitted the same quadratic time trend with an AR(2) error structure and compared it with the AR(1) model.

```

# Fit the quadratic trend with AR(2) errors
distance_ar2_model <- stats::arima(
  distance_time_data$distance,
  order = c(2, 0, 0),
  xreg = cbind(
    year_since_2000 = distance_time_data$year_since_2000,
    year_squared = distance_time_data$year_squared
  ),
  include.mean = TRUE,
  method = "ML"
)

```

```

)

# Display the AR(2) model
distance_ar2_model

##
## Call:
## stats::arima(x = distance_time_data$distance, order = c(2, 0, 0), xreg = cbind(year_since_2000 = dis
##   year_squared = distance_time_data$year_squared), include.mean = TRUE, method = "ML")
##
## Coefficients:
##          ar1          ar2  intercept  year_since_2000  year_squared
##      0.2062  -0.6484      1.4295      -0.1551      0.0141
## s.e.  0.2740   0.2171      0.1709      0.0392      0.0019
##
## sigma^2 estimated as 0.1238:  log likelihood = -8.79,  aic = 29.59

# Create a coefficient table for the AR(2) model
ar2_coefficients <- tibble(
  Term = names(distance_ar2_model$coef),
  Estimate = as.numeric(distance_ar2_model$coef),
  Standard_Error = sqrt(diag(distance_ar2_model$var.coef))
) %>%
  dplyr::mutate(
    z_value = Estimate / Standard_Error,
    p_value = 2 * pnorm(-abs(z_value))
  )

knitr::kable(
  ar2_coefficients,
  digits = 4,
  col.names = c(
    "Term",
    "Estimate",
    "Standard Error",
    "z",
    "p-value"
  ),
  caption = "Quadratic Trend in Multivariate Distance With AR(2) Errors"
)

```

Table 29: Quadratic Trend in Multivariate Distance With AR(2) Errors

| Term            | Estimate | Standard Error | z       | p-value |
|-----------------|----------|----------------|---------|---------|
| ar1             | 0.2062   | 0.2740         | 0.7524  | 0.4518  |
| ar2             | -0.6484  | 0.2171         | -2.9868 | 0.0028  |
| intercept       | 1.4295   | 0.1709         | 8.3659  | 0.0000  |
| year_since_2000 | -0.1551  | 0.0392         | -3.9601 | 0.0001  |
| year_squared    | 0.0141   | 0.0019         | 7.3784  | 0.0000  |

```

# Compare AR(1) and AR(2) models using AIC
model_comparison <- tibble(

```

```

Model = c(
  "Quadratic + AR(1)",
  "Quadratic + AR(2)"
),
AIC = c(
  AIC(distance_ar1_model),
  AIC(distance_ar2_model)
)
)

knitr::kable(
  model_comparison,
  digits = 2,
  caption = "AIC Comparison of AR(1) and AR(2) Models"
)

```

Table 30: AIC Comparison of AR(1) and AR(2) Models

| Model             | AIC   |
|-------------------|-------|
| Quadratic + AR(1) | 33.74 |
| Quadratic + AR(2) | 29.59 |

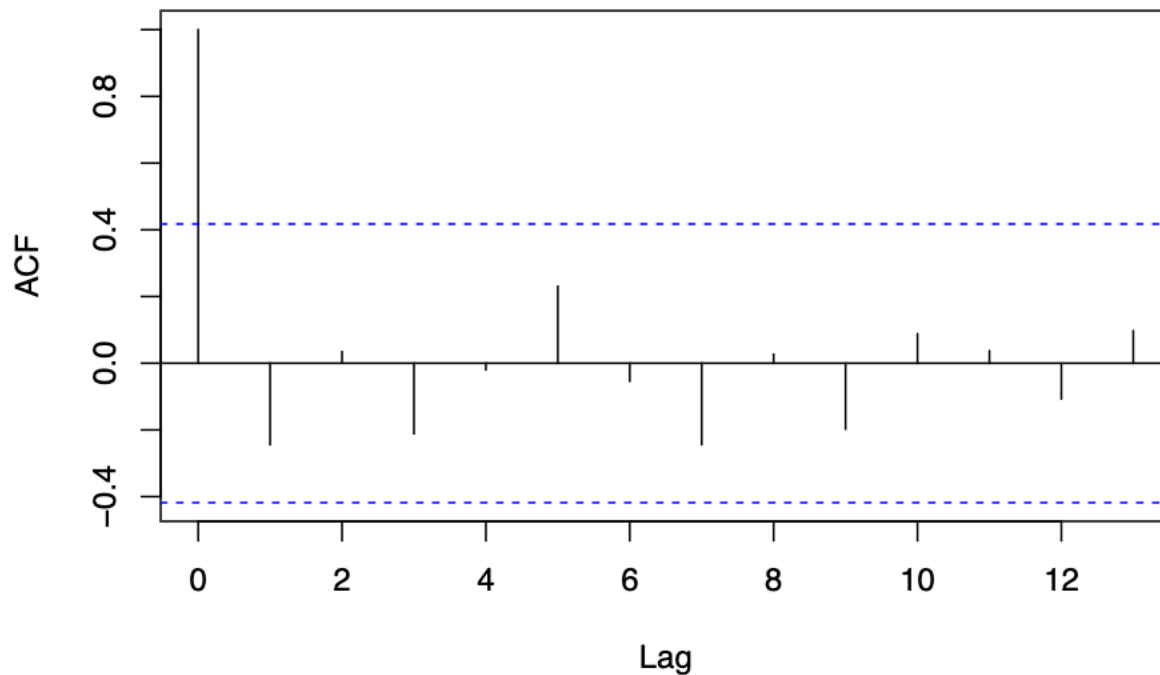
```

# Extract AR(2) residuals
distance_ar2_residuals <- residuals(distance_ar2_model)

# Check remaining residual autocorrelation
acf(
  distance_ar2_residuals,
  main = "ACF of AR(2) Quadratic Model Residuals"
)

```

## ACF of AR(2) Quadratic Model Residuals



```
# Ljung-Box test
ar2_ljung_box <- Box.test(
  distance_ar2_residuals,
  lag = 5,
  type = "Ljung-Box",
  fitdf = 2
)

ar2_ljung_box_table <- tibble(
  Statistic = as.numeric(ar2_ljung_box$statistic),
  `Degrees of Freedom` = as.numeric(ar2_ljung_box$parameter),
  `p-value` = ar2_ljung_box$p.value
)

knitr::kable(
  ar2_ljung_box_table,
  digits = 4,
  caption = "Ljung-Box Test for AR(2) Model Residuals"
)
```

Table 31: Ljung-Box Test for AR(2) Model Residuals

| Statistic | Degrees of Freedom | p-value |
|-----------|--------------------|---------|
| 4.434     | 3                  | 0.2183  |

## Interpretation

The AR(2) model provides a better description of the multivariate-distance series than the AR(1) model. Its AIC is lower (29.59 versus 33.74), indicating improved fit after allowing dependence across two yearly lags.

The second autoregressive coefficient is statistically significant ( $\phi_2 = -0.648$ ,  $p = 0.003$ ), while the first-lag coefficient is not ( $p = 0.452$ ). This shows that the remaining temporal structure was not adequately captured by an AR(1) process alone.

Importantly, the quadratic time term remains strongly significant ( $\beta_2 = 0.0141$ ,  $p < 0.001$ ). Therefore, the conclusion that the Canada–U.S. multivariate separation accelerated over time remains even after using the richer AR(2) error structure.

The residual ACF shows no clear non-zero lag outside the approximate confidence limits, and the Ljung–Box test is not significant ( $p = 0.218$ ). There is therefore no clear evidence of remaining residual autocorrelation.

Taken together, the quadratic AR(2) model provides the strongest time-series fit considered so far and supports the conclusion that the multivariate health profiles of Canada and the United States became increasingly separated over the study period.

## How Closely Does the Preferred AR(2) Model Follow the Observed Multivariate Distance?

After selecting the quadratic AR(2) model, I compared its fitted values with the observed Canada–U.S. multivariate distances. This provides a visual check of how well the preferred model captures the changing separation over time.

```
# Calculate fitted values from the AR(2) model
distance_fitted_ar2 <- distance_time_data$distance -
  residuals(distance_ar2_model)

# Create data for observed-versus-fitted plot
distance_model_plot <- dplyr::bind_rows(
  tibble::tibble(
    year = distance_time_data$year,
    Distance = distance_time_data$distance,
    Series = "Observed"
  ),
  tibble::tibble(
    year = distance_time_data$year,
    Distance = as.numeric(distance_fitted_ar2),
    Series = "Fitted AR(2)"
  )
)

# Plot observed and fitted multivariate distance
ggplot2::ggplot(
  distance_model_plot,
  ggplot2::aes(
    x = year,
    y = Distance,
    color = Series,
    linetype = Series,
    shape = Series
  )
) +
  ggplot2::geom_line(linewidth = 1.1) +
  ggplot2::geom_point(size = 2.8) +
```

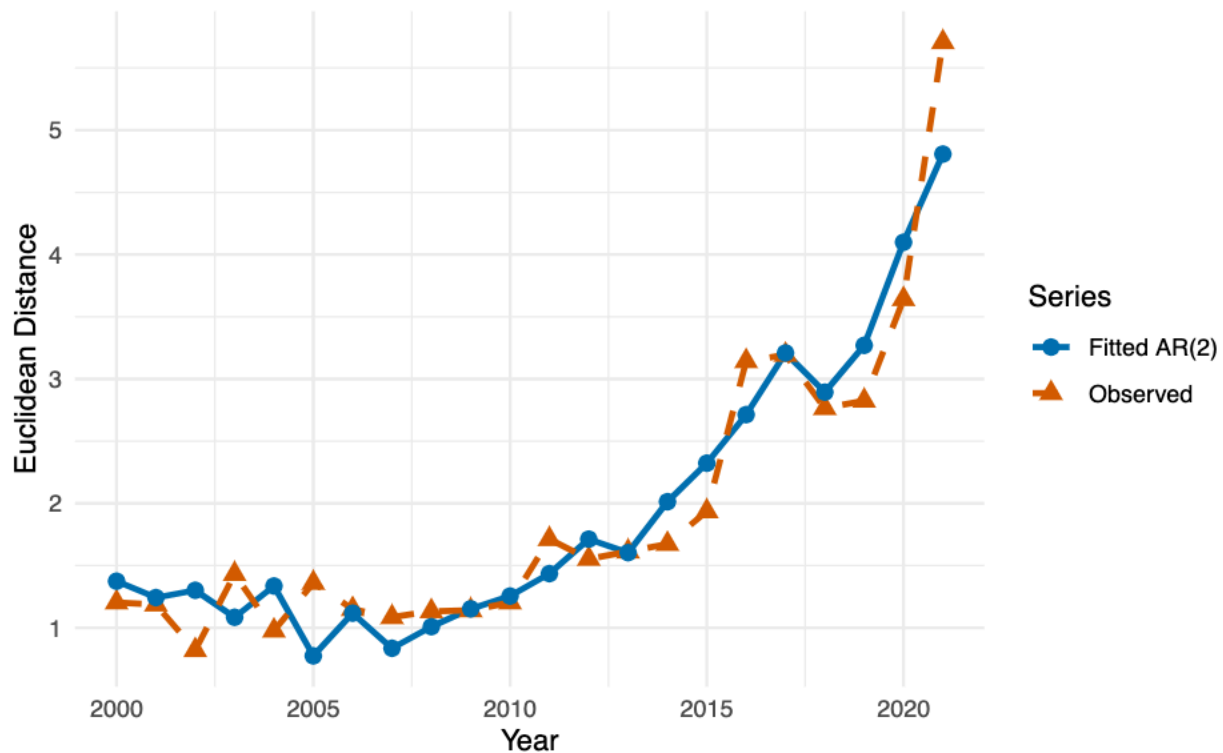
```

ggplot2::scale_color_manual(
  values = c(
    "Observed" = "#D55E00",
    "Fitted AR(2)" = "#0072B2"
  )
) +
ggplot2::scale_linetype_manual(
  values = c(
    "Observed" = "dashed",
    "Fitted AR(2)" = "solid"
  )
) +
ggplot2::scale_shape_manual(
  values = c(
    "Observed" = 17,
    "Fitted AR(2)" = 16
  )
) +
ggplot2::labs(
  title = "Observed and Fitted Canada-U.S. Multivariate Distance",
  subtitle = "Quadratic trend with AR(2) errors, 2000-2021",
  x = "Year",
  y = "Euclidean Distance",
  color = "Series",
  linetype = "Series",
  shape = "Series"
) +
ggplot2::theme_minimal() +
ggplot2::theme(
  legend.position = "right"
)

```

## Observed and Fitted Canada–U.S. Multivariate Distance

Quadratic trend with AR(2) errors, 2000–2021



### Interpretation

The fitted AR(2) model follows the observed multivariate distance closely across most of the study period. It captures both the relatively stable separation in the early years and the much stronger increase after the mid-2010s.

The largest discrepancy occurs in 2021, when the observed distance rises more sharply than the model predicts. This suggests that 2021 contains an unusually large additional separation beyond the broader quadratic trend.

Overall, the close agreement between the observed and fitted series supports the quadratic AR(2) model as a useful summary of the long-term Canada–U.S. multivariate divergence.

### Is the Preferred Multivariate Trend Sensitive to the Extreme 2021 Observation?

The observed-versus-fitted plot shows that 2021 has the largest deviation from the preferred AR(2) model. I therefore refitted the same quadratic AR(2) model after excluding 2021 to check whether the main time trend depends heavily on that final observation.

```
# Remove the 2021 observation
distance_sensitivity_2021 <- distance_time_data %>%
  dplyr::filter(year != 2021)

# Refit the quadratic AR(2) model
distance_ar2_without_2021 <- stats::arima(
  distance_sensitivity_2021$distance,
  order = c(2, 0, 0),
  xreg = cbind(
    year_since_2000 =
      distance_sensitivity_2021$year_since_2000,
```

```

    year_squared =
      distance_sensitivity_2021$year_squared
  ),
  include.mean = TRUE,
  method = "ML"
)

# Create coefficient table for the sensitivity model
sensitivity_2021_coefficients <- tibble(
  Term = names(distance_ar2_without_2021$coef),
  Estimate = as.numeric(distance_ar2_without_2021$coef),
  Standard_Error =
    sqrt(diag(distance_ar2_without_2021$var.coef))
) %>%
  dplyr::mutate(
    z_value = Estimate / Standard_Error,
    p_value = 2 * pnorm(-abs(z_value))
  )

knitr::kable(
  sensitivity_2021_coefficients,
  digits = 4,
  col.names = c(
    "Term",
    "Estimate",
    "Standard Error",
    "z",
    "p-value"
  ),
  caption = "Quadratic AR(2) Model After Excluding 2021"
)

```

Table 32: Quadratic AR(2) Model After Excluding 2021

| Term            | Estimate | Standard Error | z       | p-value |
|-----------------|----------|----------------|---------|---------|
| ar1             | 0.1105   | 0.2012         | 0.5494  | 0.5827  |
| ar2             | -0.3759  | 0.2057         | -1.8271 | 0.0677  |
| intercept       | 1.2854   | 0.1362         | 9.4355  | 0.0000  |
| year_since_2000 | -0.1008  | 0.0319         | -3.1641 | 0.0016  |
| year_squared    | 0.0108   | 0.0016         | 6.8751  | 0.0000  |

```

# Compare the key time-trend coefficients
trend_sensitivity_comparison <- tibble(
  Model = c(
    "Full series",
    "Without 2021"
  ),
  Linear_Term = c(
    distance_ar2_model$coef["year_since_2000"],
    distance_ar2_without_2021$coef["year_since_2000"]
  ),
  Quadratic_Term = c(
    distance_ar2_model$coef["year_squared"],

```

```

    distance_ar2_without_2021$coef["year_squared"]
  )
)

knitr::kable(
  trend_sensitivity_comparison,
  digits = 4,
  col.names = c(
    "Model",
    "Linear Term",
    "Quadratic Term"
  ),
  caption = "Sensitivity of the Multivariate Time Trend to 2021"
)

```

Table 33: Sensitivity of the Multivariate Time Trend to 2021

| Model        | Linear Term | Quadratic Term |
|--------------|-------------|----------------|
| Full series  | -0.1551     | 0.0141         |
| Without 2021 | -0.1008     | 0.0108         |

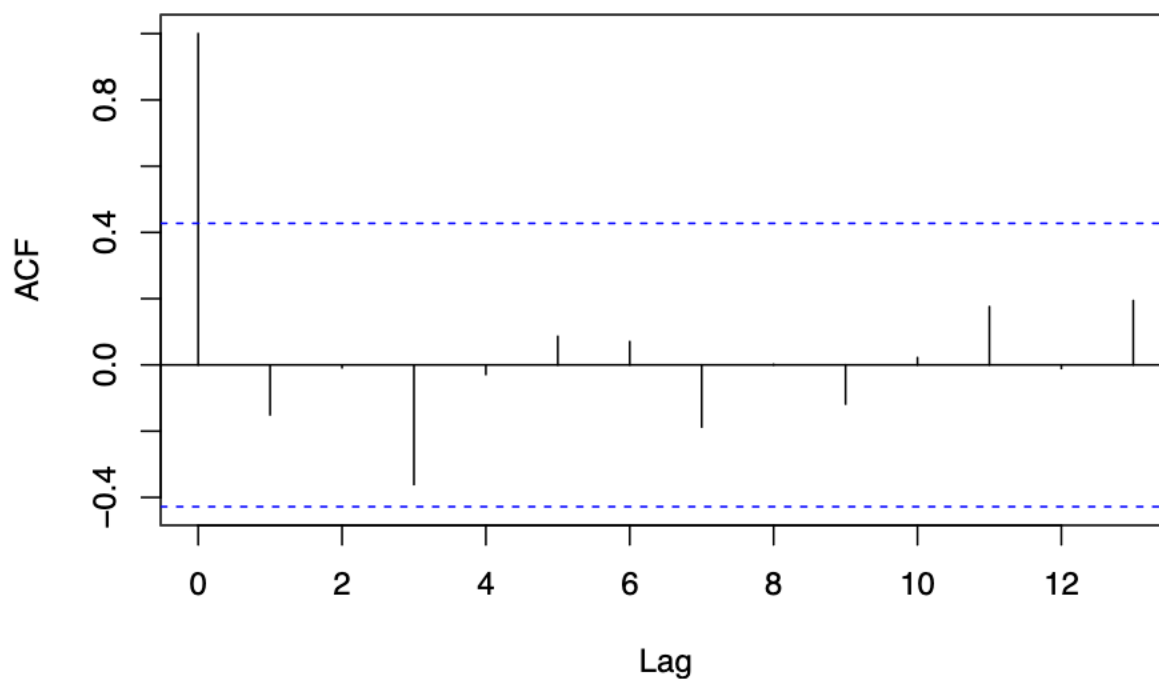
```

# Check residual autocorrelation after excluding 2021
sensitivity_residuals <-
  residuals(distance_ar2_without_2021)

acf(
  sensitivity_residuals,
  main = "ACF of AR(2) Residuals Without 2021"
)

```

## ACF of AR(2) Residuals Without 2021



```
# Ljung-Box test
sensitivity_ljung_box <- Box.test(
  sensitivity_residuals,
  lag = 5,
  type = "Ljung-Box",
  fitdf = 2
)

sensitivity_ljung_table <- tibble(
  Statistic =
    as.numeric(sensitivity_ljung_box$statistic),
  `Degrees of Freedom` =
    as.numeric(sensitivity_ljung_box$parameter),
  `p-value` =
    sensitivity_ljung_box$p.value
)

knitr::kable(
  sensitivity_ljung_table,
  digits = 4,
  caption = "Ljung-Box Test After Excluding 2021"
)
```

Table 34: Ljung-Box Test After Excluding 2021

| Statistic | Degrees of Freedom | p-value |
|-----------|--------------------|---------|
| 4.2915    | 3                  | 0.2317  |

## Interpretation

The sensitivity analysis shows that the main multivariate trend remains strong after excluding 2021.

The quadratic term stays highly significant ( $\beta_2 = 0.0108$ ,  $p < 0.001$ ), indicating that the Canada–U.S. multivariate separation still increases at an accelerating rate even without the unusually large 2021 distance.

The estimated quadratic effect is somewhat smaller than in the full model (0.0108 versus 0.0141), which shows that 2021 strengthens the late-period divergence but does not create it.

The residual diagnostics also remain acceptable. No non-zero lag clearly exceeds the ACF confidence limits, and the Ljung–Box test is not significant ( $p = 0.232$ ), providing no clear evidence of remaining residual autocorrelation.

Overall, the increasing multivariate separation is robust to the exclusion of 2021 and therefore does not depend on that single extreme observation.

## Is the PCA Structure Sensitive to the Unusual 2021 Observation?

Because 2021 produced the largest multivariate distance and an unusual suicide-rate pattern in Canada, I checked whether it also had a strong effect on the PCA results. I repeated the standardization and PCA after excluding 2021 and compared the variance explained and component loadings with the full-sample PCA.

```
# Remove 2021 before repeating the PCA
pca_sensitivity_data <- combined_data %>%
  dplyr::filter(year != 2021) %>%
  dplyr::mutate(
    daly_z = as.numeric(scale(daly_rate)),
    mental_health_z = as.numeric(scale(mental_health_death_rate)),
    suicide_z = as.numeric(scale(suicide_rate))
  )

# Fit PCA without 2021
pca_without_2021 <- prcomp(
  pca_sensitivity_data %>%
    dplyr::select(
      daly_z,
      mental_health_z,
      suicide_z
    ),
  center = FALSE,
  scale. = FALSE
)

# Compare variance explained
full_variance <- 100 *
  (pca_model$sdev^2 / sum(pca_model$sdev^2))

reduced_variance <- 100 *
  (pca_without_2021$sdev^2 /
   sum(pca_without_2021$sdev^2))

pca_variance_sensitivity <- tibble(
  Component = c("PC1", "PC2", "PC3"),
  `Full Sample (%)` = full_variance,
  `Without 2021 (%)` = reduced_variance
)
```

```
knitr::kable(
  pca_variance_sensitivity,
  digits = 2,
  caption = "Sensitivity of PCA Variance Explained to Excluding 2021"
)
```

Table 35: Sensitivity of PCA Variance Explained to Excluding 2021

| Component | Full Sample (%) | Without 2021 (%) |
|-----------|-----------------|------------------|
| PC1       | 81.95           | 86.08            |
| PC2       | 17.90           | 13.76            |
| PC3       | 0.15            | 0.16             |

```
# Compare absolute loadings to avoid arbitrary PCA sign changes
full_loadings <- abs(pca_model$rotation[, 1:2])

reduced_loadings <- abs(
  pca_without_2021$rotation[, 1:2]
)

pca_loading_sensitivity <- tibble(
  Indicator = c(
    "Drug-Use DALY Rate",
    "Mental-Health Death Rate",
    "Suicide Rate"
  ),
  `PC1 Full` = full_loadings[, 1],
  `PC1 Without 2021` = reduced_loadings[, 1],
  `PC2 Full` = full_loadings[, 2],
  `PC2 Without 2021` = reduced_loadings[, 2]
)

knitr::kable(
  pca_loading_sensitivity,
  digits = 3,
  caption = "Sensitivity of PCA Loadings to Excluding 2021"
)
```

Table 36: Sensitivity of PCA Loadings to Excluding 2021

| Indicator                | PC1 Full | PC1 Without 2021 | PC2 Full | PC2 Without 2021 |
|--------------------------|----------|------------------|----------|------------------|
| Drug-Use DALY Rate       | 0.625    | 0.611            | 0.264    | 0.284            |
| Mental-Health Death Rate | 0.604    | 0.595            | 0.433    | 0.451            |
| Suicide Rate             | 0.494    | 0.522            | 0.862    | 0.846            |

### Interpretation

The PCA structure remains very stable after excluding 2021.

PC1 explains 86.08% of the variance without 2021 compared with 81.95% in the full sample, while PC2 decreases from 17.90% to 13.76%. PC3 remains negligible in both analyses.

The loading pattern is also nearly unchanged. Drug-use DALY rate and mental-health death rate continue to contribute most strongly to PC1, while suicide rate remains the dominant contributor to PC2.

This shows that the main PCA interpretation is not driven by the unusual 2021 observation. The first component still represents a broad overall burden dimension, while the second component continues to capture a suicide-related contrast.

### How Did the Direction and Magnitude of Each Country Difference Change Over Time?

The multivariate distance measures the overall separation between the countries, but it does not show whether each individual indicator was higher or lower in the United States relative to Canada. I therefore examined the standardized U.S.–Canada difference for each indicator across all years.

Positive values indicate higher standardized rates in the United States, while negative values indicate higher standardized rates in Canada.

```
# Convert the yearly standardized differences to long format
difference_heatmap_data <- yearly_indicator_differences %>%
  tidyr::pivot_longer(
    cols = c(
      daly_difference,
      mental_health_difference,
      suicide_difference
    ),
    names_to = "Indicator",
    values_to = "Difference"
  ) %>%
  dplyr::mutate(
    Indicator = dplyr::recode(
      Indicator,
      daly_difference = "Drug-Use DALY",
      mental_health_difference = "Mental-Health Death Rate",
      suicide_difference = "Suicide Rate"
    )
  )

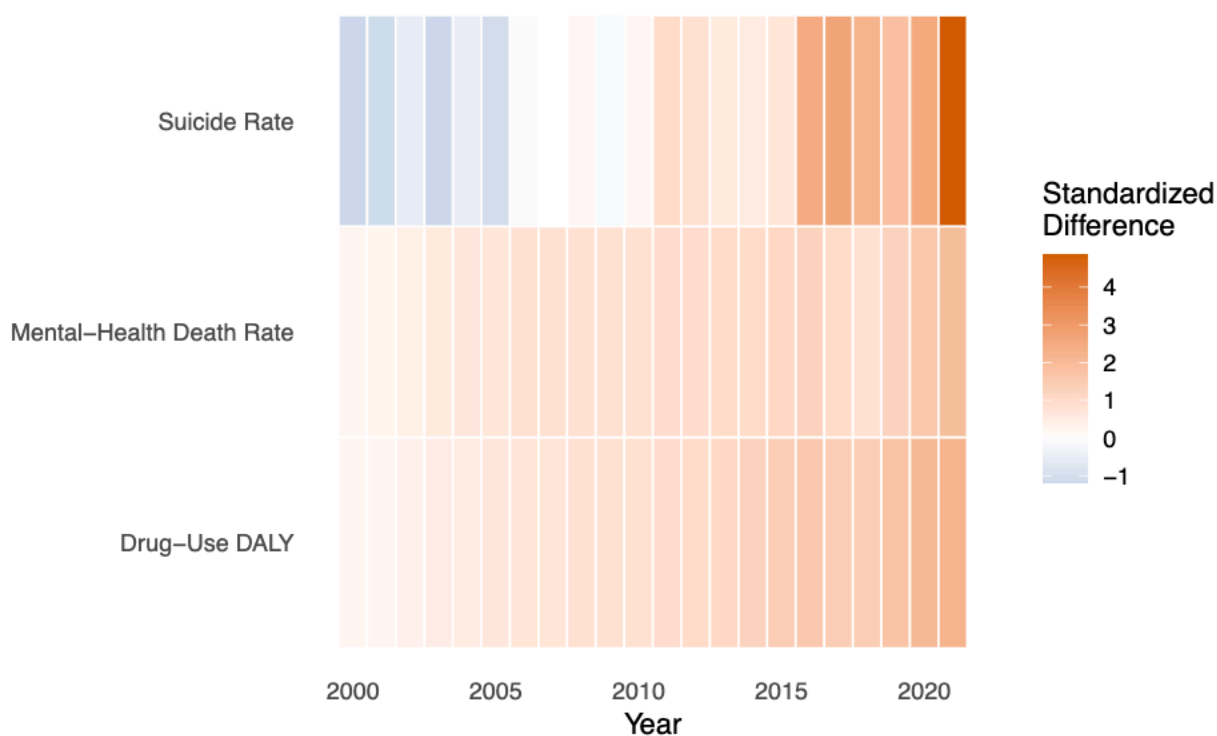
# Plot standardized U.S.–Canada differences over time
ggplot(
  difference_heatmap_data,
  aes(
    x = year,
    y = Indicator,
    fill = Difference
  )
) +
  geom_tile(
    color = "white",
    linewidth = 0.4
  ) +
  scale_fill_gradient2(
    low = "#0072B2",
    mid = "white",
    high = "#D55E00",
    midpoint = 0
  ) +
  labs(
```

```

title = "Standardized U.S.–Canada Differences Across the Three Indicators",
subtitle = "Positive values indicate higher rates in the United States, 2000–2021",
x = "Year",
y = NULL,
fill = "Standardized\nDifference"
) +
scale_x_continuous(
  breaks = seq(2000, 2020, by = 5)
) +
theme_minimal() +
theme(
  panel.grid = element_blank(),
  legend.position = "right"
)

```

Standardized U.S.–Canada Differences Across the Three Ind  
Positive values indicate higher rates in the United States, 2000–2021



### Interpretation

The heatmap shows that the direction and size of the Canada–U.S. differences changed differently across the three indicators.

Drug-use DALY rate and the mental-health death rate were generally higher in the United States throughout the study period, and the standardized differences became larger in the later years.

Suicide rate followed a different pattern. Canada had the higher standardized suicide rate in the early 2000s, the difference became small around the late 2000s, and the United States moved increasingly above Canada after about 2011. The strongest separation appears in 2021.

Overall, the widening multivariate distance is not simply an increase in all three indicators at the same

rate. The composition of the Canada–U.S. difference changed over time, with suicide rate contributing a particularly distinct pattern.

### When Did the Multivariate Separation Shift From Declining to Increasing?

The preferred quadratic AR(2) model indicates that the Canada–U.S. multivariate distance follows a curved trajectory. I therefore calculated the turning point of the fitted quadratic trend and the estimated rate of change at selected years.

```
# Extract the quadratic time-trend coefficients
linear_term <- distance_ar2_model$coef["year_since_2000"]
quadratic_term <- distance_ar2_model$coef["year_squared"]

# Calculate the turning point of the quadratic trend
turning_time <- -linear_term / (2 * quadratic_term)
turning_year <- 2000 + turning_time

# Create a table showing the estimated turning point
turning_point_table <- tibble(
  Measure = c(
    "Years after 2000",
    "Estimated turning year"
  ),
  Value = c(
    turning_time,
    turning_year
  )
)

knitr::kable(
  turning_point_table,
  digits = 2,
  col.names = c(
    "Measure",
    "Estimate"
  ),
  caption = "Estimated Turning Point of the Multivariate Distance Trend"
)
```

Table 37: Estimated Turning Point of the Multivariate Distance Trend

| Measure                | Estimate |
|------------------------|----------|
| Years after 2000       | 5.5      |
| Estimated turning year | 2005.5   |

```
# Calculate the estimated rate of change at selected years
rate_of_change <- tibble(
  Year = c(2000, 2005, 2010, 2015, 2020, 2021)
) %>%
  dplyr::mutate(
    year_since_2000 = Year - 2000,
    estimated_rate =
      linear_term +
```

```

    2 * quadratic_term * year_since_2000
) %>%
dplyr::select(
  Year,
  estimated_rate
)

knitr::kable(
  rate_of_change,
  digits = 3,
  col.names = c(
    "Year",
    "Estimated Annual Change in Distance"
  ),
  caption = "Estimated Rate of Change in Multivariate Distance"
)

```

Table 38: Estimated Rate of Change in Multivariate Distance

| Year | Estimated Annual Change in Distance |
|------|-------------------------------------|
| 2000 | -0.155                              |
| 2005 | -0.014                              |
| 2010 | 0.127                               |
| 2015 | 0.268                               |
| 2020 | 0.409                               |
| 2021 | 0.437                               |

## Interpretation

The fitted quadratic trend reaches its estimated minimum around mid-2005. Before this point, the Canada–U.S. multivariate distance was decreasing slightly; after 2005, the direction reversed and the separation began to increase.

The estimated annual change rises from approximately  $-0.014$  in 2005 to  $0.127$  in 2010,  $0.268$  in 2015, and  $0.437$  by 2021. This shows that the divergence did not simply increase at a constant rate—the pace of separation became progressively faster over time.

Within the preferred quadratic AR(2) model, the results therefore identify the mid-2000s as the approximate turning point from modest convergence to sustained and accelerating multivariate divergence.

## Does the Multivariate Structure Remain After Removing Time Trends?

The original PCA was based on the observed indicator levels, so part of its structure may reflect the strong long-term trends in the data. To determine whether the same multivariate pattern remains beyond those trends, I repeated the PCA using the detrended residuals for all three indicators.

```

# Standardize the detrended residuals
detrended_pca_data <- three_indicator_detrended %>%
  dplyr::mutate(
    daly_residual_z =
      as.numeric(scale(daly_residual)),

    mental_health_residual_z =
      as.numeric(scale(mental_health_residual)),

```

```

    suicide_residual_z =
      as.numeric(scale(suicide_residual))
  )

# Fit PCA to the detrended standardized residuals
detrended_pca_model <- prcomp(
  detrended_pca_data %>%
    dplyr::select(
      daly_residual_z,
      mental_health_residual_z,
      suicide_residual_z
    ),
  center = FALSE,
  scale. = FALSE
)

# Calculate variance explained by the detrended PCA
detrended_pca_variance <- tibble(
  Component = c("PC1", "PC2", "PC3"),
  Eigenvalue =
    detrended_pca_model$sdev^2,
  `Variance Explained (%)` =
    100 * (
      detrended_pca_model$sdev^2 /
      sum(detrended_pca_model$sdev^2)
    ),
  `Cumulative Variance (%)` =
    100 * cumsum(
      detrended_pca_model$sdev^2 /
      sum(detrended_pca_model$sdev^2)
    )
)

knitr::kable(
  detrended_pca_variance,
  digits = 2,
  caption = "Variance Explained by PCA After Removing Time Trends"
)

```

Table 39: Variance Explained by PCA After Removing Time Trends

| Component | Eigenvalue | Variance Explained (%) | Cumulative Variance (%) |
|-----------|------------|------------------------|-------------------------|
| PC1       | 2.29       | 76.22                  | 76.22                   |
| PC2       | 0.68       | 22.75                  | 98.97                   |
| PC3       | 0.03       | 1.03                   | 100.00                  |

```

# Extract loadings from the detrended PCA
detrended_pca_loadings <-
  as.data.frame(detrended_pca_model$rotation) %>%
  tibble::rownames_to_column("Indicator") %>%
  dplyr::mutate(
    Indicator = dplyr::recode(

```

```

Indicator,
daly_residual_z =
  "Drug-Use DALY Rate",
mental_health_residual_z =
  "Mental-Health Death Rate",
suicide_residual_z =
  "Suicide Rate"
)
)

knitr::kable(
  detrended_pca_loadings,
  digits = 3,
  col.names = c(
    "Indicator",
    "PC1",
    "PC2",
    "PC3"
  ),
  caption = "PCA Loadings After Removing Time Trends"
)

```

Table 40: PCA Loadings After Removing Time Trends

| Indicator                | PC1    | PC2   | PC3    |
|--------------------------|--------|-------|--------|
| Drug-Use DALY Rate       | 0.630  | 0.334 | 0.701  |
| Mental-Health Death Rate | 0.636  | 0.295 | -0.713 |
| Suicide Rate             | -0.445 | 0.895 | -0.027 |

## Interpretation

After removing the long-term time trends, the first two principal components still explain 98.97% of the remaining multivariate variation. This shows that the three indicators continue to have a strong low-dimensional structure even beyond their shared movement over time.

PC1 explains 76.22% of the detrended variation. Drug-use DALY rate (0.630) and mental-health death rate (0.636) load strongly in the same direction, while suicide rate loads in the opposite direction (-0.445). PC1 therefore represents a contrast between drug-use/mental-health burden and suicide-rate variation after the common time trends have been removed.

PC2 explains a further 22.75% and is dominated by suicide rate (0.895). This indicates that suicide contributes an additional dimension of variation that is not captured by the close relationship between drug-use DALY rate and mental-health death rate.

Compared with the original PCA, the detrended PCA gives a more nuanced interpretation: the apparent overall-burden dimension was partly influenced by shared time trends, while the underlying relationship between drug-use DALY and mental-health death rate remains very strong. PC3 explains only 1.03% and adds little additional information.

## Does the Two-Component PCA Preserve the Multivariate Country Difference?

Because PCA reduces the three standardized indicators to two main components, I checked whether the two-dimensional PCA representation preserves the Canada–U.S. separation observed in the full three-variable space.

```

# Add PCA scores to the standardized dataset
pca_score_data <- standardized_data %>%
  dplyr::bind_cols(
    as.data.frame(pca_model$x) %>%
      dplyr::select(PC1, PC2, PC3)
  )

# Create Canada PCA scores
canada_pca_scores <- pca_score_data %>%
  dplyr::filter(entity == "Canada") %>%
  dplyr::select(
    year,
    PC1,
    PC2
  )

# Create United States PCA scores
us_pca_scores <- pca_score_data %>%
  dplyr::filter(entity == "United States") %>%
  dplyr::select(
    year,
    PC1,
    PC2
  )

# Match countries by year and calculate PCA-space distance
pca_distance <- canada_pca_scores %>%
  dplyr::inner_join(
    us_pca_scores,
    by = "year",
    suffix = c("_canada", "_us")
  ) %>%
  dplyr::mutate(
    pc12_distance = sqrt(
      (PC1_us - PC1_canada)^2 +
      (PC2_us - PC2_canada)^2
    )
  )

# Compare full three-indicator distance with PC1-PC2 distance
distance_preservation <- multivariate_distance %>%
  dplyr::select(
    year,
    full_distance = distance
  ) %>%
  dplyr::left_join(
    pca_distance %>%
      dplyr::select(
        year,
        pc12_distance
      ),
    by = "year"
  ) %>%

```

```

dplyr::mutate(
  percent_preserved =
    100 * pc12_distance / full_distance
)

# Show selected years
distance_preservation_table <- distance_preservation %>%
  dplyr::filter(
    year %in% c(2000, 2005, 2010, 2015, 2020, 2021)
  )

knitr::kable(
  distance_preservation_table,
  digits = 2,
  col.names = c(
    "Year",
    "Full 3-Indicator Distance",
    "PC1-PC2 Distance",
    "Distance Preserved (%)"
  ),
  caption = "Multivariate Distance Preserved by the First Two Principal Components"
)

```

Table 41: Multivariate Distance Preserved by the First Two Principal Components

| Year | Full 3-Indicator Distance | PC1-PC2 Distance | Distance Preserved (%) |
|------|---------------------------|------------------|------------------------|
| 2000 | 1.20                      | 1.20             | 99.52                  |
| 2005 | 1.36                      | 1.36             | 99.84                  |
| 2010 | 1.20                      | 1.20             | 99.90                  |
| 2015 | 1.94                      | 1.93             | 99.52                  |
| 2020 | 3.64                      | 3.64             | 99.90                  |
| 2021 | 5.71                      | 5.70             | 99.95                  |

```

# Calculate overall agreement between the two distance measures
distance_agreement <- tibble(
  Measure = c(
    "Correlation",
    "Mean distance preserved (%)"
  ),
  Value = c(
    cor(
      distance_preservation$full_distance,
      distance_preservation$pc12_distance
    ),
    mean(
      distance_preservation$percent_preserved
    )
  )
)

knitr::kable(

```

```

distance_agreement,
digits = 3,
col.names = c(
  "Measure",
  "Value"
),
caption = "Agreement Between Full and Two-Component Multivariate Distances"
)

```

Table 42: Agreement Between Full and Two-Component Multivariate Distances

| Measure                     | Value  |
|-----------------------------|--------|
| Correlation                 | 1.000  |
| Mean distance preserved (%) | 99.813 |

## Interpretation

The first two principal components preserve almost all of the Canada–U.S. multivariate separation observed in the original three-indicator space.

Across the selected years, the PC1–PC2 distance retains more than 99.5% of the full three-indicator distance. On average, 99.81% of the distance is preserved, and the correlation between the full and two-component distance measures is essentially perfect ( $r = 1.000$ ).

This confirms that reducing the three indicators to PC1 and PC2 results in almost no loss of information about the separation between the two countries. The two-dimensional PCA representation can therefore be used confidently to visualize the main multivariate structure, while PC3 adds negligible information.

## Which Years Show the Greatest Multivariate Separation Between Canada and the United States?

The multivariate distance changed considerably over time. To identify the years with the strongest overall separation, I ranked all country-year distances from largest to smallest.

```

# Rank years by multivariate distance
largest_distances <- multivariate_distance %>%
  dplyr::arrange(dplyr::desc(distance)) %>%
  dplyr::mutate(
    Rank = dplyr::row_number()
  ) %>%
  dplyr::select(
    Rank,
    year,
    distance
  )

# Display the five years with the largest separation
top_distance_years <- largest_distances %>%
  dplyr::slice_head(n = 5)

knitr::kable(
  top_distance_years,
  digits = 2,
  col.names = c(

```

```

  "Rank",
  "Year",
  "Multivariate Distance"
),
caption = "Years With the Greatest Canada-U.S. Multivariate Separation"
)

```

Table 43: Years With the Greatest Canada-U.S. Multivariate Separation

| Rank | Year | Multivariate Distance |
|------|------|-----------------------|
| 1    | 2021 | 5.71                  |
| 2    | 2020 | 3.64                  |
| 3    | 2017 | 3.19                  |
| 4    | 2016 | 3.14                  |
| 5    | 2019 | 2.82                  |

### Interpretation

The years with the greatest multivariate separation are concentrated in the later part of the study period. The largest distance occurred in 2021 (5.71), followed by 2020 (3.64), 2017 (3.19), 2016 (3.14), and 2019 (2.82).

Four of the five most separated years occurred between 2016 and 2021, reinforcing the earlier evidence that the Canada-U.S. multivariate difference became substantially larger in the later years.

The exceptionally large distance in 2021 stands out, but the presence of several other late-period years among the highest-ranked observations shows that the growing separation is not attributable to 2021 alone.

### Which Years Show the Greatest Multivariate Separation Between Canada and the United States?

I ranked all study years by their multivariate distance to identify when the overall health profiles of Canada and the United States were most different.

```

# Rank years by multivariate distance
largest_distances <- multivariate_distance %>%
  dplyr::arrange(dplyr::desc(distance)) %>%
  dplyr::mutate(
    Rank = dplyr::row_number()
  )

# Plot years ranked by multivariate separation
ggplot(
  largest_distances,
  aes(
    x = reorder(as.factor(year), distance),
    y = distance,
    fill = distance
  )
) +
  geom_col(
    width = 0.78
  ) +
  geom_text(

```

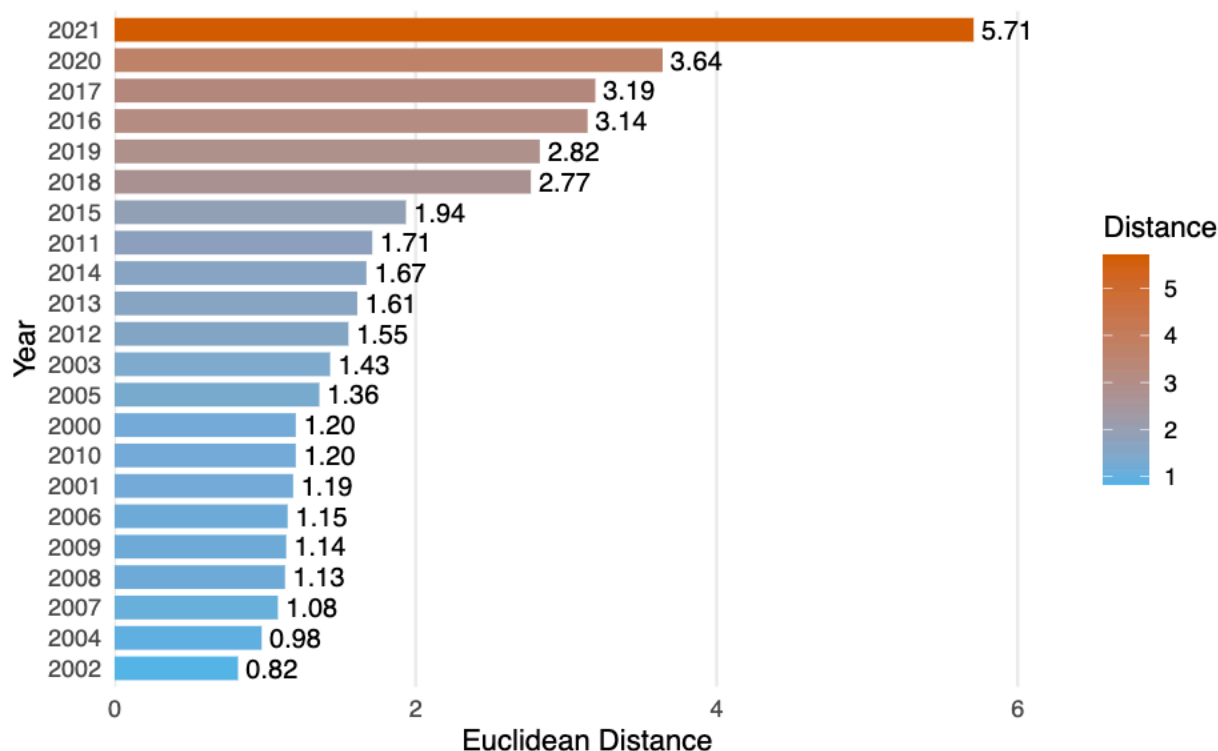
```

aes(
  label = sprintf("%.2f", distance)
),
hjust = -0.15,
size = 3.5
) +
coord_flip() +
scale_fill_gradient(
  low = "#56B4E9",
  high = "#D55E00"
) +
scale_y_continuous(
  expand = expansion(
    mult = c(0, 0.10)
  )
) +
labs(
  title = "Ranking of Canada-U.S. Multivariate Separation by Year",
  subtitle = "Larger distances indicate greater separation across the three health indicators",
  x = "Year",
  y = "Euclidean Distance",
  fill = "Distance"
) +
theme_minimal() +
theme(
  legend.position = "right",
  panel.grid.major.y = element_blank(),
  panel.grid.minor = element_blank(),
  plot.title = element_text(face = "bold"),
  axis.text.y = element_text(size = 9)
)

```

## Ranking of Canada–U.S. Multivariate Separation by Year

Larger distances indicate greater separation across the three health indicators



### Interpretation

The ranking shows that the greatest multivariate separation between Canada and the United States occurred in the later years of the study period.

The largest distance occurred in 2021 (5.71), followed by 2020 (3.64), 2017 (3.19), and 2016 (3.14). In contrast, most of the smallest distances occurred during the 2000s.

This concentration of the highest-ranked distances in the later years supports the broader finding that the two countries became increasingly different across the combined health indicators over time. The exceptionally large 2021 value stands out, but the strong separation in several earlier late-period years shows that the pattern is not driven by 2021 alone.

### Does the Result Change With a Different Standardization Method?

The main analysis used z-scores to put the three indicators on the same scale. Since unusual values can affect the mean and standard deviation, I repeated the calculation using the median and MAD.

I want to check whether the same Canada–U.S. pattern still appears.

```
# Standardize the indicators using the median and MAD
robust_data <- combined_data %>%
  dplyr::mutate(
    daly_robust =
      (daly_rate - median(daly_rate)) / mad(daly_rate),

    mental_health_robust =
      (mental_health_death_rate -
       median(mental_health_death_rate)) /
```

```

    mad(mental_health_death_rate),

    suicide_robust =
      (suicide_rate - median(suicide_rate)) /
      mad(suicide_rate)
  )

# Separate Canada
canada_robust <- robust_data %>%
  dplyr::filter(entity == "Canada") %>%
  dplyr::select(
    year,
    daly_robust,
    mental_health_robust,
    suicide_robust
  )

# Separate the United States
us_robust <- robust_data %>%
  dplyr::filter(entity == "United States") %>%
  dplyr::select(
    year,
    daly_robust,
    mental_health_robust,
    suicide_robust
  )

# Calculate the robust distance for each year
robust_distance <- canada_robust %>%
  dplyr::inner_join(
    us_robust,
    by = "year",
    suffix = c("_canada", "_us")
  ) %>%
  dplyr::mutate(
    robust_distance = sqrt(
      (daly_robust_us - daly_robust_canada)^2 +
      (mental_health_robust_us -
        mental_health_robust_canada)^2 +
      (suicide_robust_us - suicide_robust_canada)^2
    )
  )

# Compare with the original z-score distance
standardization_comparison <- multivariate_distance %>%
  dplyr::select(
    year,
    z_score_distance = distance
  ) %>%
  dplyr::left_join(
    robust_distance %>%
      dplyr::select(
        year,

```

```

      robust_distance
    ),
    by = "year"
  )

# Show selected years
standardization_comparison %>%
  dplyr::filter(
    year %in% c(2000, 2005, 2010, 2015, 2020, 2021)
  ) %>%
  knitr::kable(
    digits = 2,
    col.names = c(
      "Year",
      "Z-Score Distance",
      "Median-MAD Distance"
    ),
    caption = "Comparison of Multivariate Distance Using Two Standardization Methods"
  )

```

Table 44: Comparison of Multivariate Distance Using Two Standardization Methods

| Year | Z-Score Distance | Median-MAD Distance |
|------|------------------|---------------------|
| 2000 | 1.20             | 1.29                |
| 2005 | 1.36             | 1.50                |
| 2010 | 1.20             | 1.38                |
| 2015 | 1.94             | 2.22                |
| 2020 | 3.64             | 4.07                |
| 2021 | 5.71             | 6.23                |

```

# Check how closely the two distance series agree
standardization_agreement <- tibble(
  Measure = c(
    "Pearson correlation",
    "Spearman correlation"
  ),
  Value = c(
    cor(
      standardization_comparison$z_score_distance,
      standardization_comparison$robust_distance
    ),
    cor(
      standardization_comparison$z_score_distance,
      standardization_comparison$robust_distance,
      method = "spearman"
    )
  )
)

knitr::kable(
  standardization_agreement,

```

```

digits = 3,
col.names = c("Measure", "Value"),
caption = "Agreement Between the Two Distance Series"
)

```

Table 45: Agreement Between the Two Distance Series

| Measure              | Value |
|----------------------|-------|
| Pearson correlation  | 1.000 |
| Spearman correlation | 0.977 |

## Interpretation

The result stays almost the same when the indicators are standardized using the median and MAD instead of z-scores.

The robust distances are somewhat larger, but the two distance series are almost perfectly related ( $r = 1.000$ ). The Spearman correlation is also very high ( $r_s = 0.977$ ), showing that the ranking of years changes very little.

Most importantly, both methods show the same pattern: the Canada–U.S. separation is relatively small in the earlier years and becomes much larger toward the end of the study period, with the largest separation in 2021.

This means that the main result does not depend on the choice of standardization method.

## Does AR(2) Still Fit Better After Correcting for the Small Sample?

The time series contains only 22 years. Since ordinary AIC can favor more complicated models in small samples, I also compared the AR(1) and AR(2) models using AICc.

```

# Number of yearly observations
n_time <- nrow(distance_time_data)

# Compare AR(1) and AR(2) using AICc
aicc_comparison <- tibble(
  Model = c(
    "Quadratic + AR(1)",
    "Quadratic + AR(2)"
  ),
  AIC = c(
    AIC(distance_ar1_model),
    AIC(distance_ar2_model)
  ),
  Parameters = c(
    length(distance_ar1_model$coef) + 1,
    length(distance_ar2_model$coef) + 1
  )
) %>%
dplyr::mutate(
  AICc = AIC +
    (2 * Parameters * (Parameters + 1)) /
    (n_time - Parameters - 1),

  Delta_AICc = AICc - min(AICc)
)

```

```

# Display the small-sample model comparison
knitr::kable(
  aicc_comparison,
  digits = 2,
  col.names = c(
    "Model",
    "AIC",
    "Parameters",
    "AICc",
    "Delta AICc"
  ),
  caption = "Small-Sample Comparison of the AR(1) and AR(2) Models"
)

```

Table 46: Small-Sample Comparison of the AR(1) and AR(2) Models

| Model             | AIC   | Parameters | AICc  | Delta AICc |
|-------------------|-------|------------|-------|------------|
| Quadratic + AR(1) | 33.74 | 5          | 37.49 | 2.31       |
| Quadratic + AR(2) | 29.59 | 6          | 35.19 | 0.00       |

## Interpretation

After correcting AIC for the small sample, the AR(2) model still has the better fit. Its AICc is 35.19 compared with 37.49 for the AR(1) model.

The difference in AICc is 2.31, which gives some support to the AR(2) model, although the advantage is not overwhelming.

Together with the earlier residual checks, this supports keeping the quadratic AR(2) model as the preferred model for the multivariate-distance trend.

## Does the Quadratic Trend Really Improve the Preferred AR(2) Model?

The preferred model uses AR(2) errors, but I also want to check whether the curved time trend is actually needed. I therefore compared a linear AR(2) model with the quadratic AR(2) model using the same error structure.

```

# Fit a linear trend with AR(2) errors
distance_linear_ar2 <- stats::arima(
  distance_time_data$distance,
  order = c(2, 0, 0),
  xreg = cbind(
    year_since_2000 =
      distance_time_data$year_since_2000
  ),
  include.mean = TRUE,
  method = "ML"
)

# Calculate small-sample AICc for both models
n_time <- nrow(distance_time_data)

linear_k <- length(distance_linear_ar2$coef) + 1
quadratic_k <- length(distance_ar2_model$coef) + 1

```

```

trend_model_comparison <- tibble(
  Model = c(
    "Linear + AR(2)",
    "Quadratic + AR(2)"
  ),
  AIC = c(
    AIC(distance_linear_ar2),
    AIC(distance_ar2_model)
  ),
  AICc = c(
    AIC(distance_linear_ar2) +
      (2 * linear_k * (linear_k + 1)) /
      (n_time - linear_k - 1),

    AIC(distance_ar2_model) +
      (2 * quadratic_k * (quadratic_k + 1)) /
      (n_time - quadratic_k - 1)
  )
) %>%
dplyr::mutate(
  Delta_AICc = AICc - min(AICc)
)

knitr::kable(
  trend_model_comparison,
  digits = 2,
  col.names = c(
    "Model",
    "AIC",
    "AICc",
    "Delta AICc"
  ),
  caption = "Comparison of Linear and Quadratic AR(2) Models"
)

```

Table 47: Comparison of Linear and Quadratic AR(2) Models

| Model             | AIC   | AICc  | Delta AICc |
|-------------------|-------|-------|------------|
| Linear + AR(2)    | 45.80 | 49.55 | 14.36      |
| Quadratic + AR(2) | 29.59 | 35.19 | 0.00       |

### Interpretation

The quadratic AR(2) model fits the multivariate-distance trend much better than the linear AR(2) model. Its AICc is 35.19, compared with 49.55 for the linear model. The difference in AICc is 14.36, which provides strong support for keeping the quadratic time term.

This confirms that the Canada–U.S. multivariate separation did not change at a constant rate. The curved trend is needed to describe the faster increase in separation during the later years.

```

# Likelihood-ratio test for the quadratic term
loglik_linear <- as.numeric(logLik(distance_linear_ar2))
loglik_quadratic <- as.numeric(logLik(distance_ar2_model))

```

```

lr_statistic <- 2 * (
  loglik_quadratic - loglik_linear
)

lr_p_value <- pchisq(
  lr_statistic,
  df = 1,
  lower.tail = FALSE
)

quadratic_test <- tibble(
  Statistic = lr_statistic,
  `Degrees of Freedom` = 1,
  `p-value` = lr_p_value
)

knitr::kable(
  quadratic_test,
  digits = 4,
  caption = "Likelihood-Ratio Test for Adding the Quadratic Time Term"
)

```

Table 48: Likelihood-Ratio Test for Adding the Quadratic Time Term

| Statistic | Degrees of Freedom | p-value |
|-----------|--------------------|---------|
| 18.2071   | 1                  | 0       |

### Interpretation

The likelihood-ratio test is significant ( $\chi^2 = 18.21$ ,  $df = 1$ ,  $p < 0.001$ ). This shows that adding the quadratic time term improves the AR(2) model compared with using a straight-line trend.

Together with the AICc comparison, this gives strong evidence that the Canada–U.S. multivariate separation follows a curved pattern and increases more quickly in the later years.

### How Certain Is the Estimated Turning Year?

The preferred quadratic AR(2) model places the turning point around 2005.5. I also want to measure the uncertainty around that estimate rather than reporting a single year only.

I used the estimated coefficients and their covariance matrix to simulate plausible values of the linear and quadratic terms and then recalculated the turning year for each simulation.

```

# Select the two time-trend coefficients
trend_coef <- distance_ar2_model$coef[
  c("year_since_2000", "year_squared")
]

# Select their covariance matrix
trend_vcov <- distance_ar2_model$var.coef[
  c("year_since_2000", "year_squared"),
  c("year_since_2000", "year_squared")
]

```

```

# Simulate plausible coefficient values
set.seed(3690)

turning_simulations <- MASS::mvrnorm(
  n = 10000,
  mu = trend_coef,
  Sigma = trend_vcov
)

# Calculate the turning year for each simulation
turning_year_sim <- 2000 +
  (
    -turning_simulations[, "year_since_2000"] /
    (2 * turning_simulations[, "year_squared"])
  )

# Keep simulations with positive curvature
turning_year_sim <- turning_year_sim[
  turning_simulations[, "year_squared"] > 0
]

# Summarize the uncertainty
turning_year_ci <- tibble(
  Estimate = turning_year,
  Lower_95 = quantile(turning_year_sim, 0.025),
  Upper_95 = quantile(turning_year_sim, 0.975)
)

knitr::kable(
  turning_year_ci,
  digits = 2,
  col.names = c(
    "Estimated Turning Year",
    "Lower 95% Limit",
    "Upper 95% Limit"
  ),
  caption = "Uncertainty Around the Estimated Turning Year"
)

```

Table 49: Uncertainty Around the Estimated Turning Year

| Estimated Turning Year | Lower 95% Limit | Upper 95% Limit |
|------------------------|-----------------|-----------------|
| 2005.5                 | 2003.74         | 2006.65         |

### Interpretation

The estimated turning point is 2005.5, with a 95% interval from about 2003.7 to 2006.7.

This means the shift from a small early decline in Canada–U.S. multivariate distance to an increasing trend most likely occurred in the mid-2000s. The interval is fairly narrow, so the location of the turning point is reasonably stable rather than being tied to one exact year.

## Conclusion and Answer to the Research Question

Canada and the United States showed different mental-health and substance-use patterns over 2000–2021. The United States generally had higher levels across the three health indicators, but the size and composition of the difference changed over time.

PCA showed that most of the multivariate variation could be summarized by two components. The first mainly reflected the shared pattern in drug-use DALY rate and mental-health death rate, while the second was driven more strongly by suicide rate.

The multivariate distance between the countries was relatively small in the earlier years and increased more rapidly after the mid-2000s. This curved increase remained statistically significant after accounting for serial dependence with the quadratic AR(2) model ( $p < 0.001$ ).

The main result also remained stable when 2021 was excluded, when suicide rate was removed, when a different standardization method was used, and when the PCA was repeated without 2021.

Overall, the evidence shows that the two countries had different health profiles and that the separation between them became substantially larger over the study period.