datasciences research - R test

Adriana Cuppuleri

September, 2024

Contents

1 Introduction

This report analyzes survey data across three key areas: demographics, readiness to resume activities post-pandemic, and trust in media sources. In the demographics section, we examine gender and age distributions to identify any discrepancies from national statistics. The "when ready" section explores how Canadians across various age groups are prepared to resume different activities, providing insight into behavioral trends post-pandemic. Lastly, the report delves into Canadians' trust in media sources, segmented by both age and gender, to understand trust levels and media familiarity across demographic groups.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

```
# Load the data
data <- read.csv("C:/Users/ReDI/Downloads/ds_research_data.csv")
```

```
dim(data)
```

```
## [1] 1590  129
```

```
# Convert all column names to lowercase
colnames(data) <- tolower(colnames(data))

# Display the updated column names to confirm the change
colnames(data)
```

```
##   [1] "gender"         "birthyear"      "province"       "news_canada_1"
##   [5] "news_canada_2"  "news_canada_3"  "news_canada_4"  "news_canada_6"
##   [9] "news_canada_7"  "news_canada_8"  "news_canada_9"  "news_canada_10"
##  [13] "news_canada_11" "news_canada_12" "news_canada_13" "news_canada_14"
##  [17] "news_canada_15" "news_canada_16" "news_canada_17" "news_canada_18"
##  [21] "news_canada_19" "news_canada_20" "news_canada_21" "news_canada_22"
##  [25] "news_canada_23" "news_canada_24" "news_canada_25" "news_canada_26"
##  [29] "news_canada_27" "news_canada_28" "news_canada_29" "news_trust_1"
##  [33] "news_trust_2"   "news_trust_3"   "news_trust_4"   "news_trust_5"
##  [37] "news_trust_6"   "news_trust_7"   "news_trust_8"   "news_trust_9"
##  [41] "news_trust_10"  "news_trust_11"  "news_trust_12"  "news_trust_13"
##  [45] "news_trust_14"  "news_trust_15"  "news_trust_16"  "news_trust_17"
##  [49] "news_trust_18"  "news_trust_19"  "news_trust_20"  "news_trust_21"
##  [53] "news_trust_22"  "news_trust_23"  "news_trust_24"  "news_trust_25"
##  [57] "news_trust_26"  "news_trust_27"  "news_trust_28"  "q118_1"
##  [61] "q118_2"         "q118_3"         "q118_4"         "q118_5"
##  [65] "q118_6"         "q118_7"         "q118_8"         "q118_9"
##  [69] "q118_10"        "q118_11"        "q118_12"        "q118_13"
##  [73] "q118_14"        "q118_15"        "q118_16"        "q118_17"
##  [77] "q118_18"        "q118_19"        "q118_20"        "q118_21"
##  [81] "q118_22"        "q118_23"        "q118_24"        "q118_25"
##  [85] "q118_26"        "q118_27"        "q118_28"        "whenready_1"
##  [89] "whenready_2"    "whenready_3"    "whenready_4"    "whenready_5"
##  [93] "whenready_6"    "whenready_7"    "whenready_8"    "whenready_9"
##  [97] "whenready_10"   "q113"           "q92"            "q90"
## [101] "q108"           "q108_7_text"    "q104"           "q104_19_text"
## [105] "q104_20_text"   "q88"            "q98"            "q96_1"
## [109] "q96_4"          "q96_5"          "q96_6"          "q96_7"
## [113] "q96_8"          "q96_9"          "q96_10"         "q96_11"
## [117] "q96_12"         "q96_13"         "q96_14"         "q96_15"
## [121] "q96_16"         "q96_17"         "q96_17_text"    "q100"
## [125] "q100_6_text"    "q117.1"         "q102"           "q110"
## [129] "q112"
```

2 Test 1: Demographics

2.1 Exploratory Data Analysis

```
# Check the structure of the relevant columns: province, gender, birthyear
str(data[, c("province", "gender", "birthyear")])
```

```
## 'data.frame':    1590 obs. of  3 variables:
##  $ province : chr  "Province" "{\"ImportId\":\"Province\"}" "Ontario" "Ontario" ...
##  $ gender   : chr  "What is your gender?" "{\"ImportId\":\"QID4\"}" "Male" "Male" ...
##  $ birthyear: chr  "Please enter the year you were born:" "{\"ImportId\":\"QID5_TEXT\"}" "1986" "1978
```

Data Cleaning

the columns are stored as character strings. It looks like the first row of the dataset might be headers or metadata, not actual data values. To clean this up I will remove data rows and convert data types.

```
# Remove the first two rows that contain metadata
data <- data[-c(1, 2), ]

# Reset row names after removing rows
rownames(data) <- NULL
```

```
# Convert the 'birthyear' column to numeric
data$birthyear <- as.numeric(data$birthyear)
```

```
# Check the structure of the relevant columns again
str(data[, c("province", "gender", "birthyear")])
```

```
## 'data.frame':    1588 obs. of  3 variables:
##  $ province : chr  "Ontario" "Ontario" "Ontario" "BC" ...
##  $ gender   : chr  "Male" "Male" "Female" "Male" ...
##  $ birthyear: num  1986 1978 1999 1983 1986 ...
```

Checking for missing values

```
# Check for missing values in each column
num_missing_province <- sum(is.na(data$province))
num_missing_gender <- sum(is.na(data$gender))
num_missing_birthyear <- sum(is.na(data$birthyear))
```

```
# Print the number of missing values for each column
cat("Number of missing values in 'province':", num_missing_province, "\n")
```

```
## Number of missing values in 'province': 0
```

```
cat("Number of missing values in 'gender':", num_missing_gender, "\n")
```

```
## Number of missing values in 'gender': 0
```

```r
cat("Number of missing values in 'birthyear':", num_missing_birthyear, "\n")
```

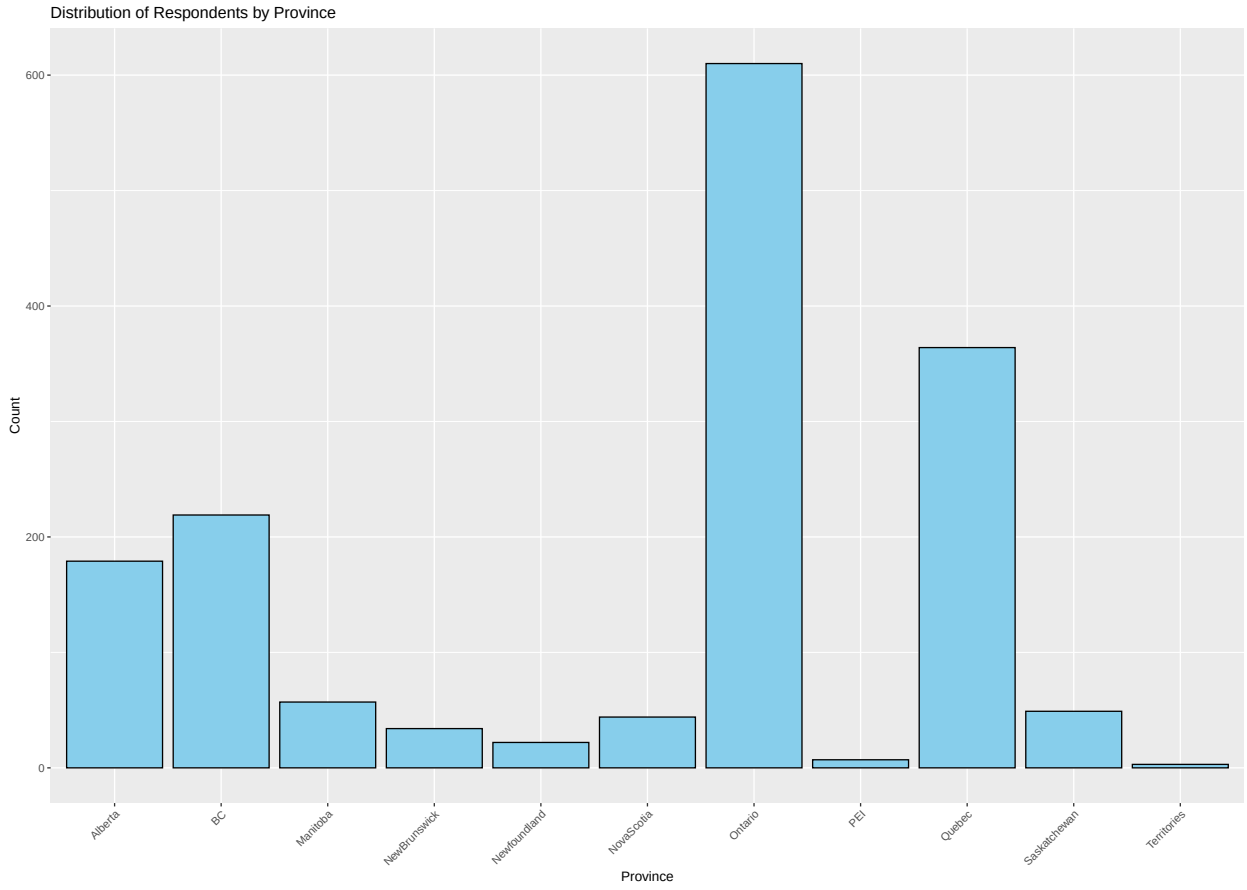
```
## Number of missing values in 'birthyear': 0
```

Summary Statistics

```r
# Summary statistics for 'province' column
province_summary <- table(data$province)
print(province_summary)
```

```
##
##      Alberta           BC     Manitoba NewBrunswick Newfoundland   NovaScotia
##          179          219           57           34           22           44
##      Ontario          PEI       Quebec Saskatchewan  Territories
##          610            7          364           49            3
```

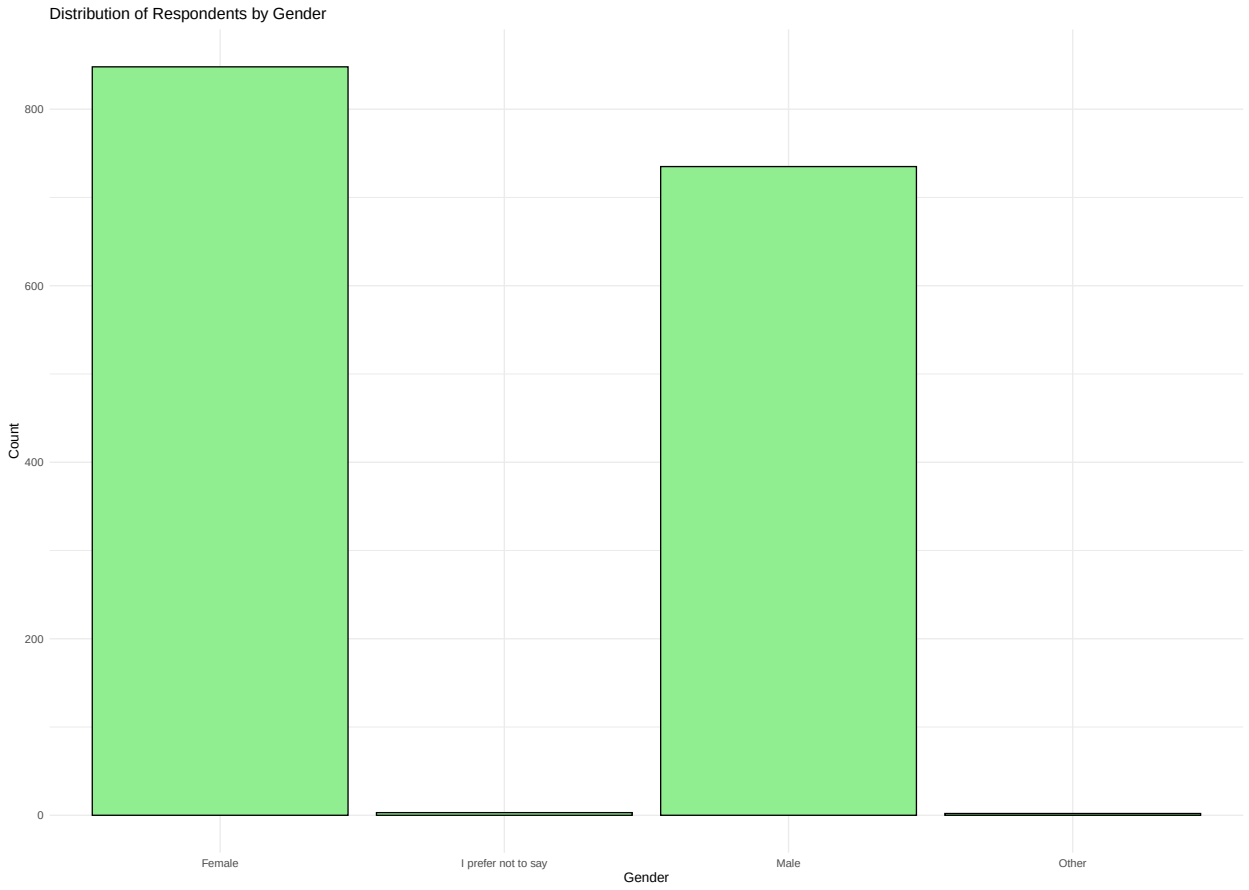
```r
# Plot distribution of province
library(ggplot2)
ggplot(data, aes(x = province)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(title = "Distribution of Respondents by Province", x = "Province", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
# Summary statistics for 'gender' column
gender_summary <- table(data$gender)
print(gender_summary)
```

```
##
##              Female I prefer not to say                 Male                Other
##                 848                   3                  735                    2
```

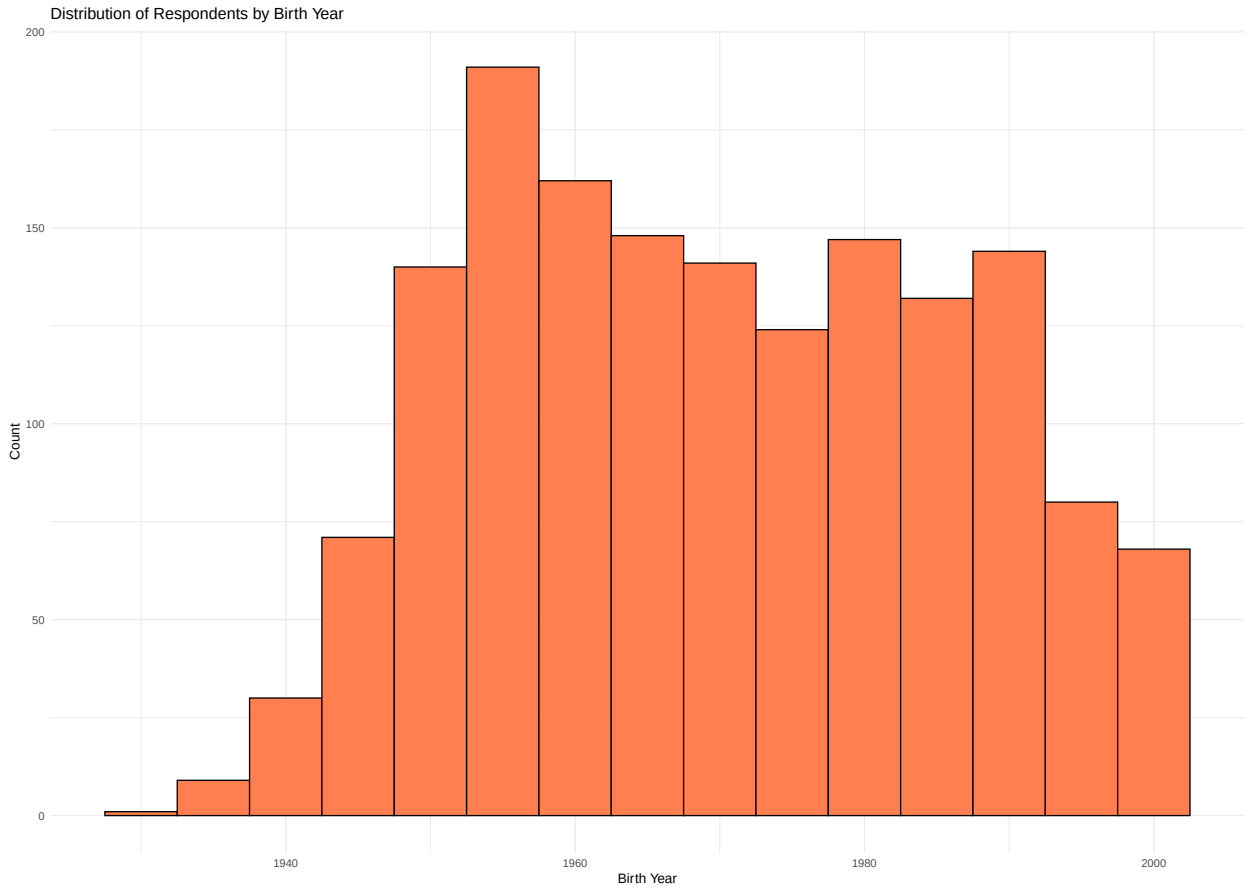
```r
# Plot distribution of gender
ggplot(data, aes(x = gender)) +
  geom_bar(fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Respondents by Gender", x = "Gender", y = "Count") +
  theme_minimal()
```




```r
# Summary statistics for 'birthyear' column
birthyear_summary <- summary(data$birthyear)
print(birthyear_summary)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1932    1956    1970    1970    1984    2002
```

```r
# Plot distribution of birth year
ggplot(data, aes(x = birthyear)) +
  geom_histogram(binwidth = 5, fill = "coral", color = "black") +
  labs(title = "Distribution of Respondents by Birth Year", x = "Birth Year", y = "Count") +
  theme_minimal()
```



2.2 Gender distribution across different Canadian provinces

```r
# Create a subset of the data for provinces (merge BC and Territories) and gender
# Calculate the proportion of male and female respondents by province

gender_province_filtered <- data %>%
  filter(gender %in% c("Male", "Female")) %>%
  mutate(province = ifelse(province %in% c("Territories","BC"), "BC and Territories",
                           province)) %>%
  group_by(province, gender) %>%
  summarise(count = n(), .groups = 'drop') %>%
  ungroup() %>%
  group_by(province) %>%
  mutate(prop = count / sum(count) * 100) %>%
  ungroup()
```

```r
# Adding a full sample row
full_sample <- gender_province_filtered %>%
  group_by(gender) %>%
  summarise(count = sum(count), .groups = 'drop') %>%
  mutate(province = "Full sample", prop = ceiling(count / sum(count) * 100))
```

```r
full_sample
```

```
## # A tibble: 2 x 4
##   gender count province     prop
##   <chr>  <int> <chr>       <dbl>
## 1 Female   848 Full sample    54
## 2 Male     735 Full sample    47
```

```r
# Combine the full sample with the original data
gender_province_combined <- bind_rows(full_sample, gender_province_filtered)
```

```r
gender_province_combined
```

```
## # A tibble: 22 x 4
##    gender count province            prop
##    <chr>  <int> <chr>              <dbl>
##  1 Female   848 Full sample         54
##  2 Male     735 Full sample         47
##  3 Female    91 Alberta             50.8
##  4 Male      88 Alberta             49.2
##  5 Female   113 BC and Territories  50.9
##  6 Male     109 BC and Territories  49.1
##  7 Female    28 Manitoba            49.1
##  8 Male      29 Manitoba            50.9
##  9 Female    20 NewBrunswick        60.6
## 10 Male      13 NewBrunswick        39.4
## # i 12 more rows
```

```r
# Reshape the dataframe: pivot wider to have Male and Female as columns
gender_province_wide <- gender_province_combined %>%
  select(province, gender, prop) %>%  # Select only the necessary columns
  pivot_wider(names_from = gender, values_from = prop) %>%
  arrange(province)  # Arrange by province for readability
```

```r
#sort it
gender_province_wide_sorted <- gender_province_wide %>%
  arrange(province)
```

```r
gender_province_wide_sorted
```

```
## # A tibble: 11 x 3
##    province           Female  Male
##    <chr>               <dbl> <dbl>
##  1 Alberta              50.8  49.2
```

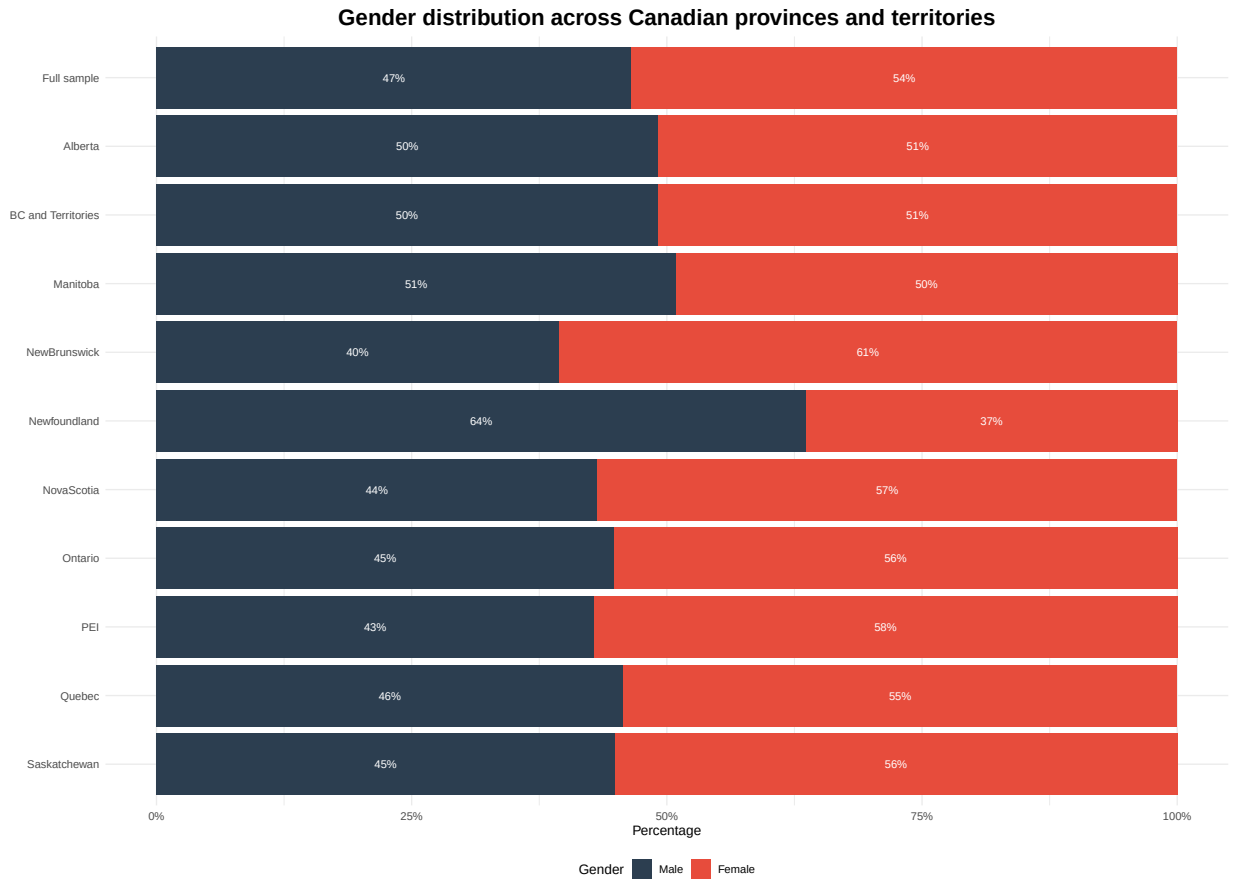
```
##  2 BC and Territories   50.9  49.1
##  3 Full sample          54    47
##  4 Manitoba             49.1  50.9
##  5 NewBrunswick         60.6  39.4
##  6 Newfoundland         36.4  63.6
##  7 NovaScotia           56.8  43.2
##  8 Ontario              55.2  44.8
##  9 PEI                  57.1  42.9
## 10 Quebec               54.3  45.7
## 11 Saskatchewan         55.1  44.9
```

```
gender_province_combined$province <- factor(gender_province_combined$province,
                                            levels = rev(unique(
                                              gender_province_combined$province)))
```

```
gender_province_combined
```

```
## # A tibble: 22 x 4
##    gender count province            prop
##    <chr>  <int> <fct>              <dbl>
##  1 Female   848 Full sample         54
##  2 Male     735 Full sample         47
##  3 Female    91 Alberta             50.8
##  4 Male      88 Alberta             49.2
##  5 Female   113 BC and Territories  50.9
##  6 Male     109 BC and Territories  49.1
##  7 Female    28 Manitoba            49.1
##  8 Male      29 Manitoba            50.9
##  9 Female    20 NewBrunswick        60.6
## 10 Male      13 NewBrunswick        39.4
## # i 12 more rows
```

```
# Plotting the gender distribution across provinces
ggplot(gender_province_combined, aes(x = province, y = prop, fill = gender)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(aes(label = paste0(ceiling(prop), "%")),
            position = position_fill(vjust = 0.5),
            color = "white", size = 3) +
  scale_y_continuous(labels = scales::percent_format()) +
  scale_fill_manual(values = c("Male" = "#2c3e50", "Female" = "#e74c3c"),
                    name = "Gender",
                    breaks = c("Male", "Female"),  # Ensures correct order in legend
                    labels = c("Male", "Female")) +  # Correct labels
  coord_flip() +
  labs(title = "Gender distribution across Canadian provinces and territories",
       x = "", y = "Percentage") +
  theme_minimal() +
  theme(legend.position = "bottom",
        plot.title = element_text(size = 18, face = "bold", hjust = 0.5))  # Adjust title
```



The gender distribution graph across Canadian provinces and territories highlights that, on average, there is a slight overrepresentation of females compared to males in the "Full sample" (54% female, 46% male). In Ontario and Quebec, the two most populated provinces in Canada, the gender distribution shows a relatively balanced representation. Some provinces, such as Newfoundland, show a higher proportion of male respondents (64% male), while others, like New Brunswick and Nova Scotia, have a noticeable female majority (61% and 57%, respectively). This variation suggests regional differences in gender distribution, which may reflect demographic characteristics, survey participation rates, or other social factors unique to each province.

1. According to Statistics Canada, the gender distribution in Canada is generally around 49% male and 51% female. However, in your survey: the full sample shows 47% male and 54% female, which is slightly off from the national average. Certain provinces, like Newfoundland (64% female) and New Brunswick (61% female), show much larger discrepancies compared to actual Canadian demographics, where a closer 50/50 split is expected.

   Possible Causes:

   - Sampling bias: Certain demographics (e.g., males in some provinces) may be underrepresented in the sample.

   Correction:

   - A weighting can be applied to adjust data and better reflect the actual Canadian population distribution.

2. The graph shows the distribution of survey respondents but does not account for the actual population size of each province. Without this context, it's difficult to assess how representative the survey

sample is relative to each province's population. Adding an additional layer to the analysis, such as a segmented bar chart by age group, or creating a separate age distribution graph alongside this gender distribution, could provide a more comprehensive demographic profile.

3 Test 2: when_ready

```
#select when_ready columns + gender and birthyear
when_ready_data <- data %>%
  select(gender, birthyear, whenready_1, whenready_2, whenready_3, whenready_4, whenready_5,
         whenready_6, whenready_7, whenready_8, whenready_9, whenready_10)
```

```
summary(when_ready_data)
```

```
##     gender            birthyear    whenready_1        whenready_2
##  Length:1588        Min.   :1932   Length:1588        Length:1588
##  Class :character   1st Qu.:1956   Class :character   Class :character
##  Mode  :character   Median :1970   Mode  :character   Mode  :character
##                     Mean   :1970
##                     3rd Qu.:1984
##                     Max.   :2002
##  whenready_3        whenready_4        whenready_5        whenready_6
##  Length:1588        Length:1588        Length:1588        Length:1588
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  whenready_7        whenready_8        whenready_9        whenready_10
##  Length:1588        Length:1588        Length:1588        Length:1588
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Creating Age Groups

```
when_ready_data <- when_ready_data %>%
  mutate(age = 2024 - birthyear,  # Assuming current year is 2024
         age_group = case_when(
           age >= 18 & age <= 34 ~ "18-34 years",
           age >= 35 & age <= 49 ~ "35-49 years",
           age >= 50 & age <= 64 ~ "50-64 years",
           age >= 65 ~ "65+ years",
           TRUE ~ "Unknown"  # Handles missing or incorrect values
         ))
```

Reshape Data

```r
#Add a unique ID for each respondent
when_ready_data <- when_ready_data %>%
  mutate(ID = row_number())
```

```r
when_ready_long <- when_ready_data %>%
  pivot_longer(
    cols = starts_with("whenready_"),
    names_to = "behavior",
    values_to = "readiness"
  )
```

```r
# Reorder columns to ensure 'ID' comes before 'behavior'
when_ready_long <- when_ready_long %>%
  select(ID, birthyear, age, age_group, behavior, readiness)
```

Convert Responses to Numeric Values

```r
unique(when_ready_long$readiness)
```

```
## [1] "\t\tIn more than 6 months"
## [2] "\t\tWhen a vaccine will be ready or the virus will have disappeared"
## [3] "I am already ready or doing it "
## [4] "\t\tIn 3-6 months"
## [5] "\t\tIn 1-3 months"
## [6] "\t\tI don’t foresee a time when I will be ready to do it again"
## [7] "Not applicable for me"
## [8] ""
```

```r
#Standardize the responses by trimming whitespace
when_ready_long <- when_ready_long %>%
  mutate(readiness = trimws(readiness))
```

```r
#Replace empty space/ NA values in 'readiness' with "Not applicable for me"
when_ready_long <- when_ready_long %>%
  mutate(readiness = ifelse(readiness == "" | is.na(readiness), "Not applicable for me", readiness))
```

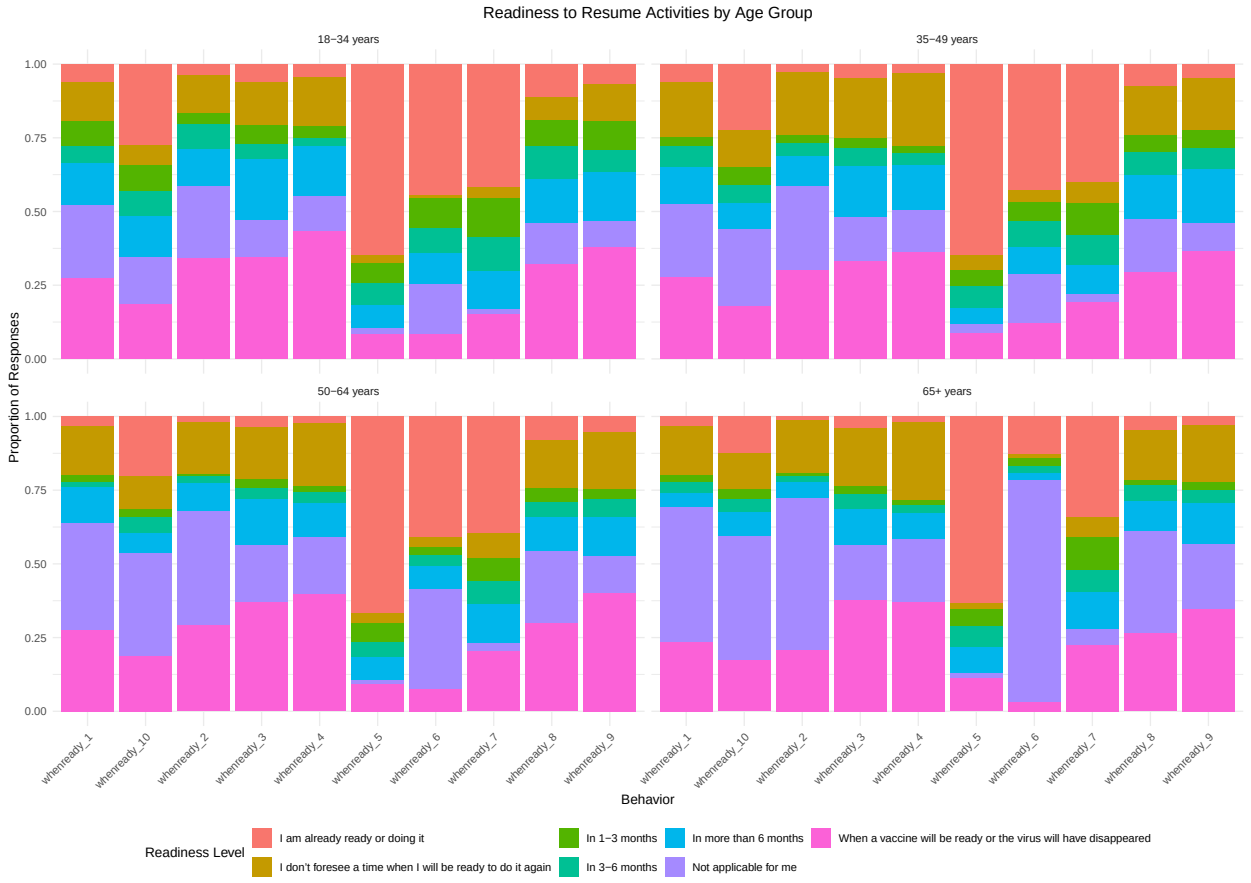
```r
# Convert the responses to numeric values
when_ready_long <- when_ready_long %>%
  mutate(readiness_numeric = case_when(
    readiness == "I am already ready or doing it" ~ 5,
    readiness == "In 1-3 months" ~ 4,
    readiness == "In 3-6 months" ~ 3,
    readiness == "In more than 6 months" ~ 2,
    readiness == "When a vaccine will be ready or the virus will have disappeared" ~ 1,
    readiness == "I don’t foresee a time when I will be ready to do it again" ~ 0,
    readiness == "Not applicable for me" ~ -1,
    TRUE ~ NA_real_  # Handles any unexpected values
  ))
```

3.0.1 Stacked bar chart

distribution of readiness levels for each behavior across different age groups

```r
options(repr.plot.width = 12, repr.plot.height = 8)
```

```r
# Stacked bar chart showing readiness across different behaviors and age groups
ggplot(when_ready_long, aes(x = behavior, fill = readiness)) +
  geom_bar(position = "fill") +
  facet_wrap(~ age_group) +
  labs(title = "Readiness to Resume Activities by Age Group",
       x = "Behavior",
       y = "Proportion of Responses",
       fill = "Readiness Level") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "bottom",
        plot.title = element_text(hjust = 0.5))
```



3.1 Behavior Descriptions

Behavior Code	Description
`whenready_1`	Travel domestically by plane for business or essential purposes
`whenready_2`	Travel internationally by plane for business or essential purposes

Behavior Code	Description
whenready_3	Travel domestically by plane for leisure
whenready_4	Travel internationally by plane for leisure
whenready_5	Go shopping in-person at stores or malls
whenready_6	Return to the workplace
whenready_7	Go eat in restaurants
whenready_8	Go to bars, clubs, or crowded spaces
whenready_9	Attend large public events like a festival or outdoor gathering
whenready_10	Take public transit

```
# Save the plot
ggsave("readiness_by_age_group.png", width = 12, height = 8, dpi = 300)
```

The stacked bar chart provides a detailed look at readiness levels for various activities across different age groups. Older age groups (50-64 and 65+ years) have a higher proportion of responses in categories like "I don't foresee a time when I will be ready to do it again" (represented in pink) and "When a vaccine will be ready or the virus will have disappeared" (light blue), especially for high-risk activities such as international travel and attending large public events. In contrast, younger age groups (18-34 and 35-49 years) show more responses indicating they are already ready or will be ready within 1-6 months for various activities, demonstrating less concern or greater willingness to resume normal activities.

3.2 Heatmap of Average Readiness Scores

how different age groups feel about resuming various activities. The color intensity represents the average readiness score

```
#Calculate average readiness score for each behavior and age group
average_readiness <- when_ready_long %>%
  group_by(age_group, behavior) %>%
  summarize(avg_score = round(mean(readiness_numeric, na.rm = TRUE), 2), .groups = "drop")
```

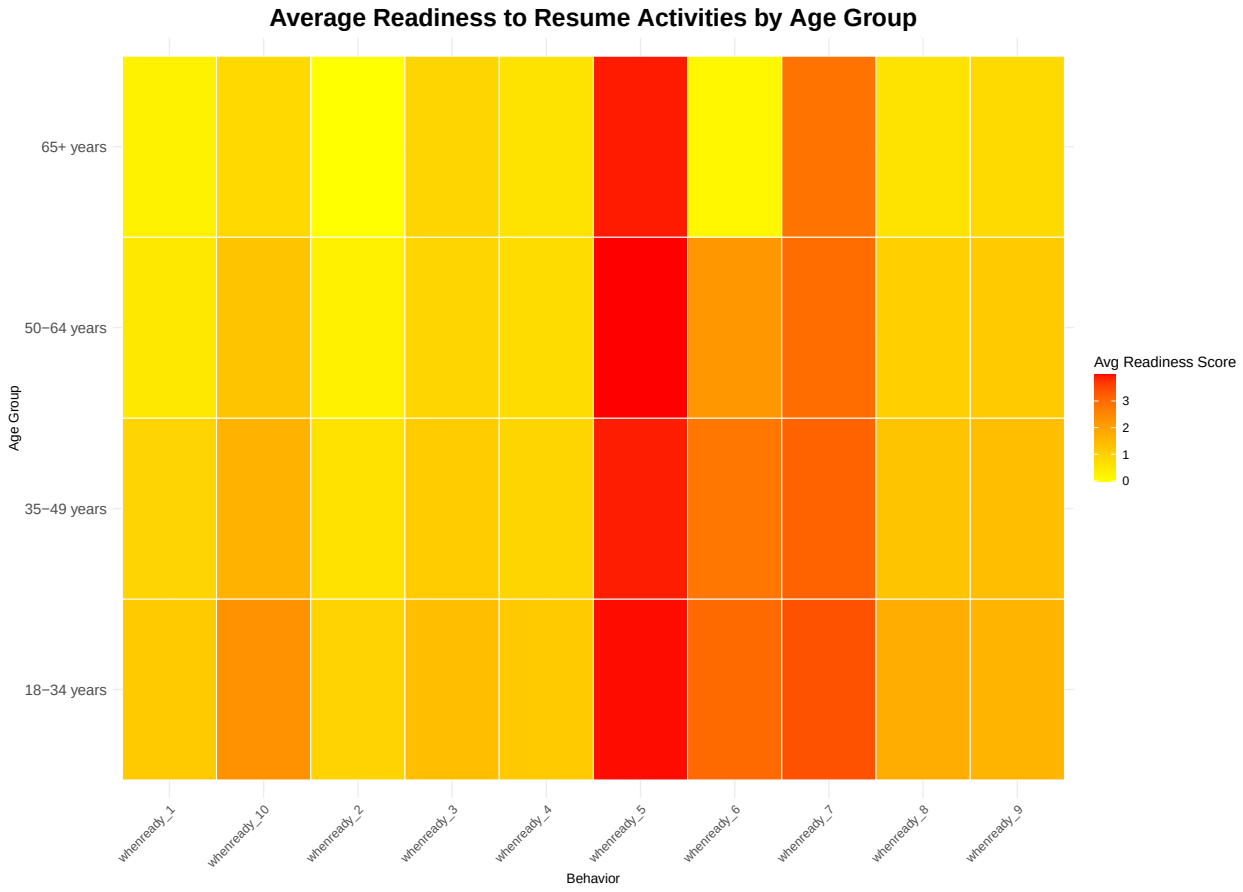
```
head(average_readiness)
```

```
## # A tibble: 6 x 3
##   age_group   behavior     avg_score
##   <chr>       <chr>            <dbl>
## 1 18-34 years whenready_1       1.12
## 2 18-34 years whenready_10      2.28
## 3 18-34 years whenready_2       0.94
## 4 18-34 years whenready_3       1.35
## 5 18-34 years whenready_4       1.11
## 6 18-34 years whenready_5       3.94
```

```
# create heatmap of average readiness scores
heatmap_plot <- ggplot(average_readiness, aes(x = behavior, y = age_group, fill = avg_score)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "yellow", high = "red", name = "Avg Readiness Score") +
  labs(
    title = "Average Readiness to Resume Activities by Age Group",
```

```
    x = "Behavior",
    y = "Age Group"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),  # Customize x-axis text
    axis.text.y = element_text(size = 12),                         # Customize y-axis text
    plot.title = element_text(size = 20, hjust = 0.5, face = "bold"),  # Adjust title size, centering,
    legend.text = element_text(size = 10),                         # Customize legend text size
    legend.title = element_text(size = 12)                         # Customize legend title size
  )
```

```
heatmap_plot
```



3.3 Readiness Level Descriptions

`readiness_numeric` Value	Readiness Description
5	I am already ready or doing it
4	In 1-3 months
3	In 3-6 months
2	In more than 6 months
1	When a vaccine will be ready or the virus will have disappeared

readiness_numeric Value	Readiness Description
0	I don't foresee a time when I will be ready to do it again
-1	Not applicable for me

```
ggsave("average_readiness_heatmap.png", plot = heatmap_plot, width = 10, height = 6, dpi = 300)
```

The heatmap illustrates the average readiness scores for different activities across various age groups. The darkest red color indicates the highest readiness to resume specific activities. Notably, younger age groups (18-34 and 35-49 years) show higher average readiness scores (dark red) for activities like whenready_5 (likely corresponding to "Go shopping in-person at stores or malls") and whenready_6 (which could be "Return to the workplace"), indicating they feel more ready to engage in these activities. In contrast, older age groups (50-64 and 65+ years) show lighter colors across many activities, suggesting a more cautious approach and lower readiness scores. This trend highlights that younger individuals are generally more eager or prepared to return to normal activities, whereas older adults exhibit more hesitation.

4 Bonus: Sources of Information

```
# Add a unique ID column to the dataset
data <- data %>%
  mutate(ID = row_number())
```

```
#Extract and Clean Familiarity Data
familiarity_data <- data %>%
  select(ID, starts_with("news_canada_")) %>%
  pivot_longer(
    cols = starts_with("news_canada_"),
    names_to = "familiarity_source",
    values_to = "source_name"
  ) %>%
  filter(!is.na(source_name) & source_name != "0" & source_name != "")
# Keep only recognized sources
```

```
head(familiarity_data)
```

```
## # A tibble: 6 x 3
##      ID familiarity_source source_name
##   <int> <chr>              <chr>
## 1     1 news_canada_9      Toronto Star
## 2     2 news_canada_27     The Wall Street Journal
## 3     3 news_canada_1      CTV News
## 4     3 news_canada_4      MacLeans
## 5     3 news_canada_9      Toronto Star
## 6     3 news_canada_24     The New York Times
```

```
#Extract and Clean Trust Data
trust_data <- data %>%
  select(ID, starts_with("news_trust_")) %>%
  pivot_longer(
```

```
    cols = starts_with("news_trust_"),
    names_to = "trust_source",
    values_to = "trust_value"
  ) %>%
  filter(!is.na(trust_value) & trust_value != "0" & trust_value != "")
# Keep only valid trust responses
```

```
head(trust_data, n=12)
```

```
## # A tibble: 12 x 3
##       ID trust_source  trust_value
##    <int> <chr>         <chr>
##  1     1 news_trust_8  Yes
##  2     2 news_trust_26 Yes
##  3     3 news_trust_1  No
##  4     3 news_trust_4  No
##  5     3 news_trust_8  No
##  6     3 news_trust_23 No
##  7     4 news_trust_7  Not sure
##  8     4 news_trust_10 Not sure
##  9     4 news_trust_11 Not sure
## 10     4 news_trust_26 Not sure
## 11     4 news_trust_27 Not sure
## 12     5 news_trust_1  Not sure
```

Since the *news_canada_...* columns and *news_trust_...* columns don't align directly, we need to create a mapping manually. This helps us to understand which *news_trust_...* column corresponds to each *news_canada_...* column. *news_canada_29...* was not included because there is no entry other than None, 0, and blank. *news_canada_5...* is not present in the dataframe

```
familiarity_columns <- colnames(data)[4:30]
trust_columns <- colnames(data)[32:60]
```

```
head(familiarity_columns, n=28)
```

```
##  [1] "news_canada_1"  "news_canada_2"  "news_canada_3"  "news_canada_4"
##  [5] "news_canada_6"  "news_canada_7"  "news_canada_8"  "news_canada_9"
##  [9] "news_canada_10" "news_canada_11" "news_canada_12" "news_canada_13"
## [13] "news_canada_14" "news_canada_15" "news_canada_16" "news_canada_17"
## [17] "news_canada_18" "news_canada_19" "news_canada_20" "news_canada_21"
## [21] "news_canada_22" "news_canada_23" "news_canada_24" "news_canada_25"
## [25] "news_canada_26" "news_canada_27" "news_canada_28"
```

```
head(trust_columns, n=28)
```

```
##  [1] "news_trust_1"  "news_trust_2"  "news_trust_3"  "news_trust_4"
##  [5] "news_trust_5"  "news_trust_6"  "news_trust_7"  "news_trust_8"
##  [9] "news_trust_9"  "news_trust_10" "news_trust_11" "news_trust_12"
## [13] "news_trust_13" "news_trust_14" "news_trust_15" "news_trust_16"
## [17] "news_trust_17" "news_trust_18" "news_trust_19" "news_trust_20"
## [21] "news_trust_21" "news_trust_22" "news_trust_23" "news_trust_24"
## [25] "news_trust_25" "news_trust_26" "news_trust_27" "news_trust_28"
```

```r
#Create a manual mapping based on the first row inspection
mapping <- data.frame(
  familiarity_source = familiarity_columns,
  trust_source = trust_columns[1:length(familiarity_columns)]
)
```

```r
print(mapping)
```

```
##    familiarity_source  trust_source
## 1       news_canada_1  news_trust_1
## 2       news_canada_2  news_trust_2
## 3       news_canada_3  news_trust_3
## 4       news_canada_4  news_trust_4
## 5       news_canada_6  news_trust_5
## 6       news_canada_7  news_trust_6
## 7       news_canada_8  news_trust_7
## 8       news_canada_9  news_trust_8
## 9      news_canada_10  news_trust_9
## 10     news_canada_11 news_trust_10
## 11     news_canada_12 news_trust_11
## 12     news_canada_13 news_trust_12
## 13     news_canada_14 news_trust_13
## 14     news_canada_15 news_trust_14
## 15     news_canada_16 news_trust_15
## 16     news_canada_17 news_trust_16
## 17     news_canada_18 news_trust_17
## 18     news_canada_19 news_trust_18
## 19     news_canada_20 news_trust_19
## 20     news_canada_21 news_trust_20
## 21     news_canada_22 news_trust_21
## 22     news_canada_23 news_trust_22
## 23     news_canada_24 news_trust_23
## 24     news_canada_25 news_trust_24
## 25     news_canada_26 news_trust_25
## 26     news_canada_27 news_trust_26
## 27     news_canada_28 news_trust_27
```

4.1 Trust Levels for Media Sources

```r
# Merge familiarity data with trust data using the ID column and the mapping
combined_data <- familiarity_data %>%
  inner_join(mapping, by = "familiarity_source") %>%
  inner_join(trust_data, by = c("ID", "trust_source"))
```

```r
head(combined_data, n= 22)
```

```
## # A tibble: 22 x 5
##       ID familiarity_source source_name            trust_source  trust_value
##    <int> <chr>              <chr>                  <chr>         <chr>
##  1     1 news_canada_9      Toronto Star           news_trust_8  Yes
```

```
##  2     2 news_canada_27     The Wall Street Journal news_trust_26 Yes
##  3     3 news_canada_1      CTV News                news_trust_1  No
##  4     3 news_canada_4      MacLeans                news_trust_4  No
##  5     3 news_canada_9      Toronto Star            news_trust_8  No
##  6     3 news_canada_24     The New York Times      news_trust_23 No
##  7     4 news_canada_8      The National Post       news_trust_7  Not sure
##  8     4 news_canada_11     The Vancouver Sun       news_trust_10 Not sure
##  9     4 news_canada_12     The Province            news_trust_11 Not sure
## 10     4 news_canada_27     The Wall Street Journal news_trust_26 Not sure
## # i 12 more rows
```

```r
# Summarize the data to count the number of "Yes", "No",
# and "Not sure" responses for each media source
summary_data <- combined_data %>%
  group_by(source_name) %>%
  summarize(
    familiarity_count = n(),  # Count of recognitions
    trust_yes_count = sum(trust_value == "Yes", na.rm = TRUE),
    trust_no_count = sum(trust_value == "No", na.rm = TRUE),
    trust_not_sure_count = sum(trust_value == "Not sure", na.rm = TRUE)
  ) %>%
  arrange(desc(familiarity_count))
```

```r
str(summary_data)
```

```
## tibble [27 x 5] (S3: tbl_df/tbl/data.frame)
##  $ source_name         : chr [1:27] "Toronto Star" "The CBC" "CTV News" "The Globe and Mail " ...
##  $ familiarity_count   : int [1:27] 687 668 650 561 482 380 333 304 302 296 ...
##  $ trust_yes_count     : int [1:27] 358 399 460 332 257 229 156 60 64 68 ...
##  $ trust_no_count      : int [1:27] 130 107 72 94 80 48 89 78 68 150 ...
##  $ trust_not_sure_count: int [1:27] 199 162 118 135 145 103 88 166 170 78 ...
```

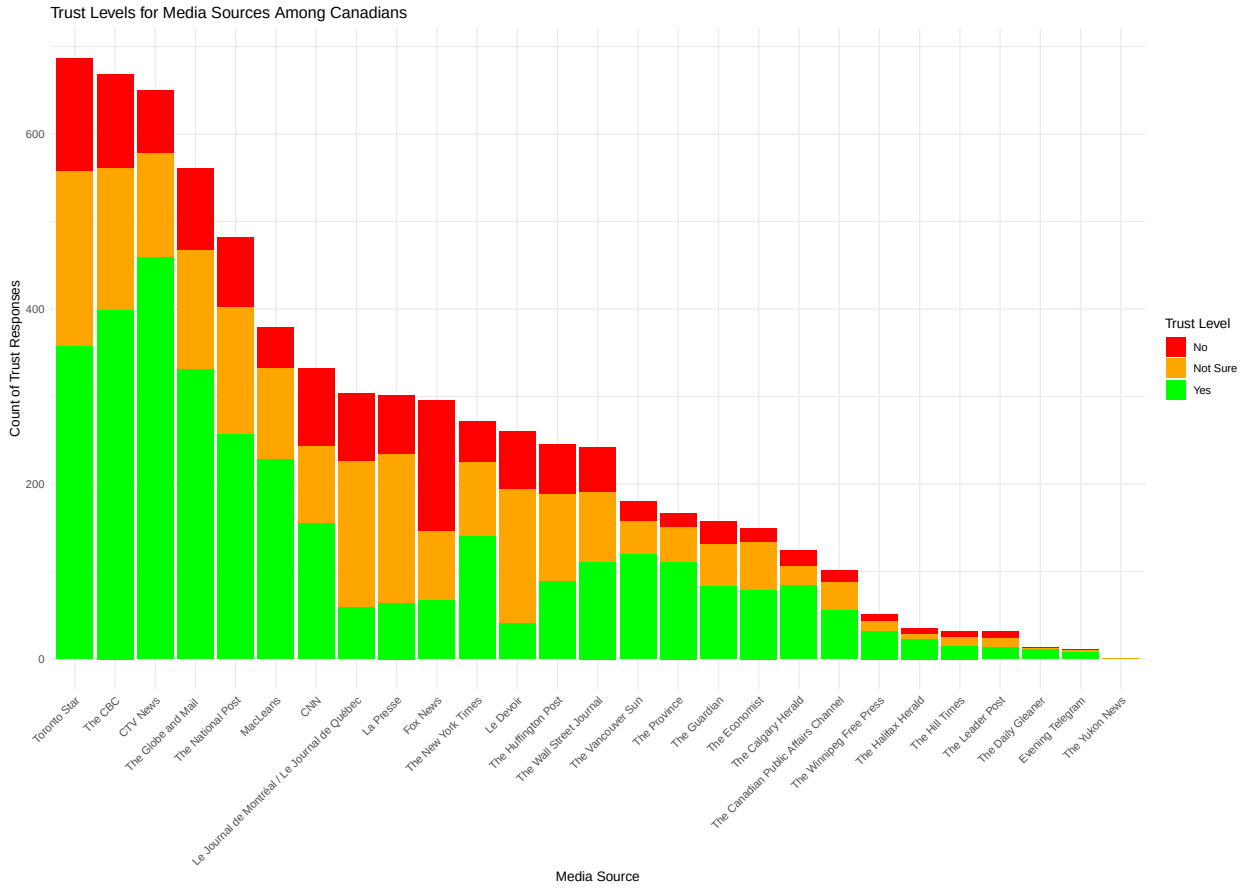
```r
summary_data
```

```
## # A tibble: 27 x 5
##    source_name                  familiarity_count trust_yes_count trust_no_count
##    <chr>                                    <int>           <int>          <int>
##  1 "Toronto Star"                             687             358            130
##  2 "The CBC"                                  668             399            107
##  3 "CTV News"                                 650             460             72
##  4 "The Globe and Mail "                      561             332             94
##  5 "The National Post"                        482             257             80
##  6 "MacLeans"                                 380             229             48
##  7 "CNN"                                      333             156             89
##  8 "Le Journal de Montréal / L~               304              60             78
##  9 "La Presse"                                302              64             68
## 10 "Fox News"                                 296              68            150
## # i 17 more rows
## # i 1 more variable: trust_not_sure_count <int>
```

```r
# Reshape data for plotting
trust_plot_data <- summary_data %>%
  pivot_longer(cols = starts_with("trust_"), names_to = "trust_level", values_to = "count")
```

4.1.1 Graph 1: Trust Level among Canadians

```r
# Plot the data
ggplot(trust_plot_data, aes(x = reorder(source_name,
                                        -familiarity_count), y = count, fill = trust_level)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(
    title = "Trust Levels for Media Sources Among Canadians",
    x = "Media Source",
    y = "Count of Trust Responses"
  ) +
  scale_fill_manual(values = c("trust_yes_count" = "green",
                               "trust_no_count" = "red", "trust_not_sure_count" = "orange"),
                    name = "Trust Level",
                    labels = c("No", "Not Sure", "Yes")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



4.2 Trust Level Media Sources by *Age*

```
birthyear_data <- data %>%
  select(ID, birthyear)
```

```
combined_with_birthyear <- combined_data %>%
  inner_join(birthyear_data, by = "ID")
```

```
head(combined_with_birthyear)
```

```
## # A tibble: 6 x 6
##      ID familiarity_source source_name        trust_source trust_value birthyear
##   <int> <chr>              <chr>              <chr>        <chr>           <dbl>
## 1     1 news_canada_9      Toronto Star       news_trust_8 Yes              1986
## 2     2 news_canada_27     The Wall Street J~ news_trust_~ Yes              1978
## 3     3 news_canada_1      CTV News           news_trust_1 No               1999
## 4     3 news_canada_4      MacLeans           news_trust_4 No               1999
## 5     3 news_canada_9      Toronto Star       news_trust_8 No               1999
## 6     3 news_canada_24     The New York Times news_trust_~ No               1999
```

```
age_summary <- combined_with_birthyear %>%
  group_by(birthyear, source_name) %>%
  summarize(
    familiarity_count = n(),
    trust_yes_count = sum(trust_value == "Yes", na.rm = TRUE),
    trust_no_count = sum(trust_value == "No", na.rm = TRUE),
    trust_not_sure_count = sum(trust_value == "Not sure", na.rm = TRUE),
    .groups = 'drop'  # This removes the grouping after summarizing
  ) %>%
  arrange(desc(familiarity_count))
```

```
head(age_summary)
```

```
## # A tibble: 6 x 6
##   birthyear source_name  familiarity_count trust_yes_count trust_no_count
##       <dbl> <chr>                    <int>           <int>          <int>
## 1      1986 Toronto Star                27              15              4
## 2      1963 The CBC                     24              14              4
## 3      1956 The CBC                     23              14              4
## 4      1989 Toronto Star                22              11              7
## 5      1954 The CBC                     20              11              7
## 6      1956 Toronto Star                19               9              3
## # i 1 more variable: trust_not_sure_count <int>
```

```
age_group_summary <- combined_with_birthyear %>%
  mutate(
    age = 2024 - birthyear,  # Calculate age assuming the current year is 2024
    age_group = case_when(
      age >= 18 & age <= 34 ~ "18-34 years",
      age >= 35 & age <= 49 ~ "35-49 years",
```

```
      age >= 50 & age <= 64 ~ "50-64 years",
      age >= 65 ~ "65+ years",
      TRUE ~ "Unknown"  # Handles missing or incorrect values
    )
  ) %>%
  group_by(age_group, source_name) %>%
  summarize(
    familiarity_count = n(),
    trust_yes_count = sum(trust_value == "Yes", na.rm = TRUE),
    trust_no_count = sum(trust_value == "No", na.rm = TRUE),
    trust_not_sure_count = sum(trust_value == "Not sure", na.rm = TRUE),
    .groups = 'drop'  # To remove grouping after summarizing
  )
```

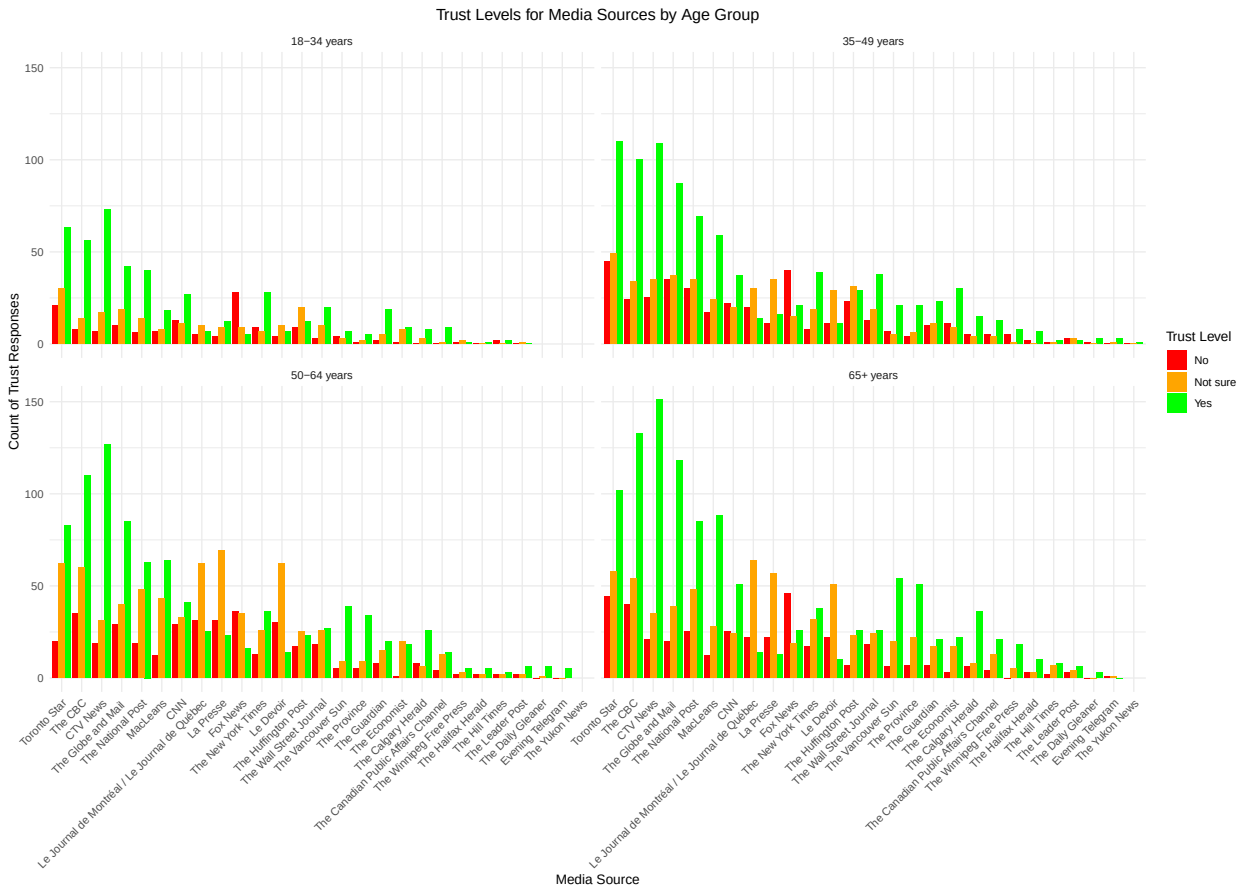
```
head(age_group_summary)
```

```
## # A tibble: 6 x 6
##   age_group   source_name       familiarity_count trust_yes_count trust_no_count
##   <chr>       <chr>                         <int>           <int>          <int>
## 1 18-34 years CNN                              51              27             13
## 2 18-34 years CTV News                         97              73              7
## 3 18-34 years Fox News                         42               5             28
## 4 18-34 years La Presse                        25              12              4
## 5 18-34 years Le Devoir                        21               7              4
## 6 18-34 years Le Journal de Mo~                22               7              5
## # i 1 more variable: trust_not_sure_count <int>
```

```
# Reshape data for plotting with grouped bars
trust_plot_data_age_group <- age_group_summary %>%
  pivot_longer(cols = starts_with("trust_"), names_to = "trust_level", values_to = "count")
```

4.2.1 Graph 2: Trust level by Age

```
# Plot grouped bar plot with age groups
ggplot(trust_plot_data_age_group, aes(x = reorder(source_name,
                                                  -familiarity_count), y = count, fill = trust_level)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ age_group) +  # Separate plots for each age group
  labs(
    title = "Trust Levels for Media Sources by Age Group",
    x = "Media Source",
    y = "Count of Trust Responses"
  ) +
  scale_fill_manual(values = c("trust_yes_count" = "green",
                               "trust_no_count" = "red", "trust_not_sure_count" = "orange"),
                    name = "Trust Level",
                    labels = c("No", "Not sure", "Yes")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5)
        )
```



Graph 2 shows trust levels for media sources by different age groups (18-34, 35-49, 50-64, and 65+ years). It highlights that older age groups (especially 50-64 and 65+ years) tend to trust traditional media outlets like The BBC and The New York Times more, as seen by the higher proportion of green bars. Younger groups (18-34 years) show more variability in their trust responses, with significant "Not sure" responses (in orange) across different media. Overall, trust in media appears to increase with age, with older generations showing more decisive "Yes" or "No" responses compared to younger audiences.

```
# Save the last displayed plot directly
ggsave(filename = "trust_levels_by_age_group.png", width = 10, height = 6, dpi = 300)
```

4.3 Trust Level for Media Source by *Gender*

```
gender_data <- data %>%
  select(ID, gender)
```

```
combined_with_gender <- combined_data %>%
  inner_join(gender_data, by = "ID")
```

```
head(combined_with_gender)
```

```
## # A tibble: 6 x 6
##      ID familiarity_source source_name          trust_source trust_value gender
```

```
##   <int> <chr>              <chr>                 <chr>        <chr>       <chr>
## 1     1 news_canada_9      Toronto Star          news_trust_8 Yes         Male
## 2     2 news_canada_27     The Wall Street Jour~ news_trust_~ Yes         Male
## 3     3 news_canada_1      CTV News              news_trust_1 No          Female
## 4     3 news_canada_4      MacLeans              news_trust_4 No          Female
## 5     3 news_canada_9      Toronto Star          news_trust_8 No          Female
## 6     3 news_canada_24     The New York Times    news_trust_~ No          Female
```

```
gender_summary <- combined_with_gender %>%
  group_by(gender, source_name) %>%
  summarize(
    familiarity_count = n(),
    trust_yes_count = sum(trust_value == "Yes", na.rm = TRUE),
    trust_no_count = sum(trust_value == "No", na.rm = TRUE),
    trust_not_sure_count = sum(trust_value == "Not sure", na.rm = TRUE),
    .groups = 'drop'  # To remove grouping after summarizing
  )
```

```
head(gender_summary)
```

```
## # A tibble: 6 x 6
##   gender source_name      familiarity_count trust_yes_count trust_no_count
##   <chr>  <chr>                        <int>           <int>          <int>
## 1 Female CNN                            167              78             40
## 2 Female CTV News                       324             226             35
## 3 Female Evening Telegram                 3               3              0
## 4 Female Fox News                       144              33             68
## 5 Female La Presse                      158              34             34
## 6 Female Le Devoir                      138              20             35
## # i 1 more variable: trust_not_sure_count <int>
```

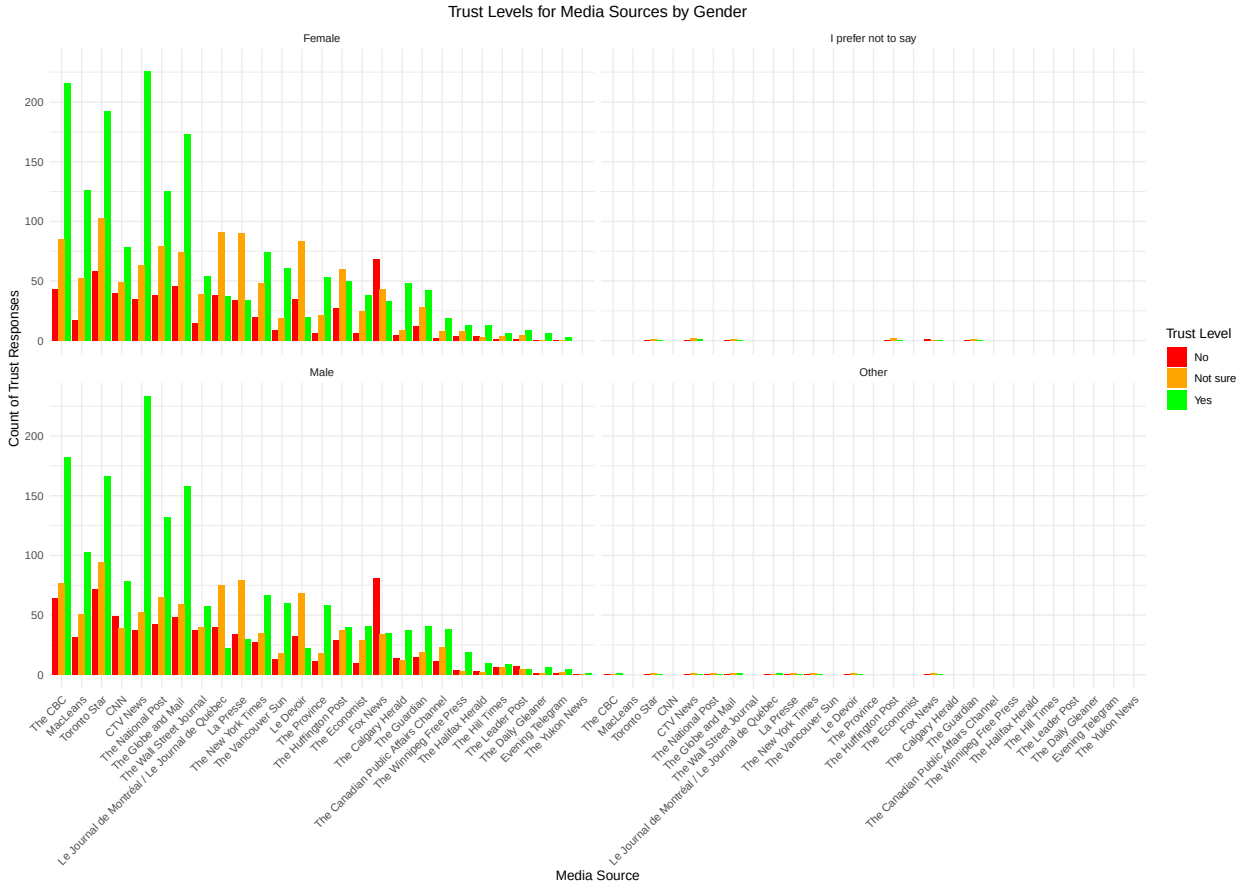
```
# Reshape data to a long format for plotting
trust_plot_data_gender <- gender_summary %>%
  pivot_longer(
    cols = starts_with("trust_"),
    names_to = "trust_level",
    values_to = "count"
  )
```

```
head(trust_plot_data_gender)
```

```
## # A tibble: 6 x 5
##   gender source_name familiarity_count trust_level          count
##   <chr>  <chr>                  <int> <chr>                <int>
## 1 Female CNN                      167 trust_yes_count         78
## 2 Female CNN                      167 trust_no_count          40
## 3 Female CNN                      167 trust_not_sure_count    49
## 4 Female CTV News                 324 trust_yes_count        226
## 5 Female CTV News                 324 trust_no_count          35
## 6 Female CTV News                 324 trust_not_sure_count    63
```

4.3.1 Graph3: Trust level by Gender

```
# Plot grouped bar plot with gender
ggplot(trust_plot_data_gender, aes(x = reorder(source_name,
                                               -familiarity_count), y = count, fill = trust_level)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ gender) +  # Separate plots for each gender
  labs(
    title = "Trust Levels for Media Sources by Gender",
    x = "Media Source",
    y = "Count of Trust Responses"
  ) +
  scale_fill_manual(values = c("trust_yes_count" = "green",
                               "trust_no_count" = "red", "trust_not_sure_count" = "orange"),
                    name = "Trust Level",
                    labels = c("No", "Not sure", "Yes")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5)
  )
```



Graph 3 illustrates the distribution of trust levels for various media sources, segmented by gender categories: Female, Male, Other, and "I prefer not to say." Across all gender groups, we observe a trend where certain media sources, such as The BBC and The Toronto Star, receive higher trust levels ("Yes" in green). Males

appear to express more trust overall, with higher green bars, compared to the other gender groups. In contrast, the "No" trust response (in red) is more prominent in media sources like CNN and The Toronto Sun, particularly in the "I prefer not to say" category, indicating a diverse perception of trust across gender groups.