Shifting Sands of Canada

Analyzing Changing Interprovincial Migration Trends

2023-11-29

---

1 TEAM MEMBERS: Group SAAS

1.1 *Huynh Hiep Tran (Alex)* - T00728369

1.2 *Shivani Tyagi* - T00727866

1.3 *Sayantika Saha* - T00731231

1.4 *Mohd Asaf Shaikh* – T00728877

---

2 TABLE OF CONTENT

---

3 PROJECT OBJECTIVE

This project focuses on investigating **interprovincial migration in Canada from 1971 to 2022**, with a specific emphasis on migration patterns on both a yearly and quarterly basis. We aim to explore whether people prefer to migrate during specific times, the provinces that are less favored and those that attract the most migrants, and whether migration trends have changed over the decades. This research will provide valuable insights about the movement of people within the country and can become a rich resource for studying Canada's population mobility, with applications in economics, sociology, policy development, and urban planning.

4 ABOUT DATA SET

This comprehensive dataset provides valuable information on the migration patterns of individuals and families across Canada from 1971 to 2022. It specifically focuses on interprovincial migration, tracking the movement of people between different provinces and territories. This dataset is a crucial resource for understanding the dynamics of population movement within Canada over several decades.

Data Fields - The dataset typically includes the following key data fields:

1. *Year:* The time period indicating the specific year in which the migration occurred.
2. *Quarter:* The time period indicating the specific quarter in which the migration occurred.
3. *Province or Territory of Origin:* The province or territory from which migrants originated. This includes data on the number of individuals or families moving out of each province.
4. *Province or Territory of Destination:* The province or territory to which migrants are relocating. This includes data on the number of individuals or families moving into each province.
5. *Number of Migrants:* The total count of individuals or families moving from the destination to the specified quarter.

5 LOADING DATA SET

```
# Reading data file

migration_data <- read.csv("C:/Users/shiva/Documents/ADSC1910/ProjectWork/interprovincial_migration.csv
                           header=TRUE)
head(migration_data)
```

```
##   Year Quarter Origin N.L. P.E.I. N.S. N.B. Que.  Ont. Man. Sask. Alta. B.C.
## 1 1971       3   N.L.    0     39  378  279  218  1732   40    20    87  111
## 2 1971       3 P.E.I.   35      0  326  256   77   563   34    24   102   52
## 3 1971       3   N.S.  596    283    0 1272  590  3754  214    90   518  871
## 4 1971       3   N.B.  432    260 1199    0  942  2785  124    66   328  384
## 5 1971       3   Que.  373     92  651 1346    0 11692  535   123   821 1365
## 6 1971       3   Ont. 2436    658 3942 2967 7014     0 2958  1053  4444 6287
##   Y.T. N.W.T. Nvt. Total
## 1    4      6    0  2914
## 2    5      7    0  1481
## 3   19     30    0  8237
## 4    7     11    0  6538
## 5   21     31    0 17050
## 6  153    238    0 32150
```

5.1 Displaying Column Names

```
# Display variable names
cat("Variable Names:\n")
```

```
## Variable Names:
```

```
print(names(migration_data))
```

```
##  [1] "Year"    "Quarter" "Origin"  "N.L."    "P.E.I."  "N.S."    "N.B."
##  [8] "Que."    "Ont."    "Man."    "Sask."   "Alta."   "B.C."    "Y.T."
## [15] "N.W.T."  "Nvt."    "Total"
```

5.2 Checking Dimensions of the dataset

```
# Display dimensions of the dataset
cat("\nDimensions of the Dataset :\n")
```

```
##
## Dimensions of the Dataset :
```

```
print(dim(migration_data))
```

```
## [1] 2585   17
```

5.3 Exploring Data Summary

```
# Display summary statistics for numeric variables
cat("\nSummary Statistics of the Dataset :\n")
```

```
##
## Summary Statistics of the Dataset :
```

```
print(summary(migration_data))
```

```
##       Year         Quarter         Origin               N.L.
##  Min.   :1971   Min.   :1.000   Length:2585        Min.   :   0.0
##  1st Qu.:1984   1st Qu.:2.000   Class :character   1st Qu.:  15.0
##  Median :1997   Median :3.000   Mode  :character   Median :  52.0
##  Mean   :1997   Mean   :2.503                      Mean   : 168.2
##  3rd Qu.:2010   3rd Qu.:3.000                      3rd Qu.: 170.0
##  Max.   :2022   Max.   :4.000                      Max.   :2515.0
##      P.E.I.             N.S.             N.B.             Que.
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Min.   :   0.0
##  1st Qu.:   3.00   1st Qu.:  31.0   1st Qu.:  14.0   1st Qu.:  21.0
##  Median :  26.00   Median : 166.0   Median : 106.0   Median :  97.0
##  Mean   :  64.04   Mean   : 358.4   Mean   : 262.1   Mean   : 486.8
##  3rd Qu.:  88.00   3rd Qu.: 458.0   3rd Qu.: 354.0   3rd Qu.: 406.0
##  Max.   :1711.00   Max.   :6790.0   Max.   :4937.0   Max.   :8655.0
##       Ont.            Man.            Sask.            Alta.
##  Min.   :    0   Min.   :   0.0   Min.   :   0.0   Min.   :    0
##  1st Qu.:  104   1st Qu.:  18.0   1st Qu.:  15.0   1st Qu.:  147
##  Median :  876   Median :  76.0   Median :  54.0   Median :  519
##  Mean   : 1514   Mean   : 322.2   Mean   : 355.3   Mean   : 1404
##  3rd Qu.: 2107   3rd Qu.: 539.0   3rd Qu.: 505.0   3rd Qu.: 1861
##  Max.   :17432   Max.   :3250.0   Max.   :4129.0   Max.   :18017
##       B.C.             Y.T.            N.W.T.            Nvt.
##  Min.   :    0   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.:   99   1st Qu.:  0.00   1st Qu.:  7.00   1st Qu.:  0.00
##  Median :  305   Median : 12.00   Median : 28.00   Median :  0.00
##  Mean   : 1207   Mean   : 37.09   Mean   : 56.14   Mean   : 11.72
##  3rd Qu.: 1113   3rd Qu.: 44.00   3rd Qu.: 73.00   3rd Qu.: 16.00
##  Max.   :12505   Max.   :381.00   Max.   :551.00   Max.   :295.00
##      Total
##  Min.   :   80
##  1st Qu.:  975
##  Median : 4091
##  Mean   : 6246
##  3rd Qu.: 9682
##  Max.   :49032
```

```
# We can also use the 'describe' function from the 'psych' package for a more detailed summary
library(psych)
describe(migration_data)
```

```
##          vars    n    mean      sd median trimmed     mad  min   max range  skew
## Year        1 2585 1997.11   14.73   1997 1997.20   19.27 1971  2022    51 -0.05
## Quarter     2 2585    2.50    1.12      3    2.50    1.48    1     4     3 -0.01
## Origin*     3 2585    6.97    3.80      7    6.96    4.45    1    13    12  0.02
## N.L.        4 2585  168.16  300.89     52   95.18   71.16    0  2515  2515  3.46
## P.E.I.      5 2585   64.04  103.43     26   42.96   38.55    0  1711  1711  4.41
## N.S.        6 2585  358.36  544.71    166  243.93  232.77    0  6790  6790  3.62
## N.B.        7 2585  262.08  402.47    106  178.89  152.71    0  4937  4937  3.39
## Que.        8 2585  486.77 1100.27     97  201.28  143.81    0  8655  8655  3.84
## Ont.        9 2585 1513.65 1942.40    876 1130.05 1197.94    0 17432 17432  2.29
## Man.       10 2585  322.22  477.18     76  223.39  112.68    0  3250  3250  2.22
## Sask.      11 2585  355.31  596.80     54  211.40   75.61    0  4129  4129  2.28
## Alta.      12 2585 1403.62 1983.20    519  977.51  713.13    0 18017 18017  2.32
## B.C.       13 2585 1206.69 2015.03    305  708.27  422.54    0 12505 12505  2.38
## Y.T.       14 2585   37.09   60.37     12   22.37   17.79    0   381   381  2.48
## N.W.T.     15 2585   56.14   75.62     28   40.28   37.06    0   551   551  2.46
## Nvt.       16 2585   11.72   22.26      0    6.77    0.00    0   295   295  4.10
## Total      17 2585 6245.84 6517.49   4091 5164.14 5002.29   80 49032 48952  1.69
##          kurtosis     se
## Year        -1.18   0.29
## Quarter     -1.36   0.02
## Origin*     -1.26   0.07
## N.L.        15.32   5.92
## P.E.I.      38.02   2.03
## N.S.        20.09  10.71
## N.B.        18.69   7.92
## Que.        16.32  21.64
## Ont.         7.35  38.20
## Man.         6.30   9.39
## Sask.        5.39  11.74
## Alta.        7.12  39.01
## B.C.         5.58  39.63
## Y.T.         6.53   1.19
## N.W.T.       7.47   1.49
## Nvt.        29.16   0.44
## Total        3.72 128.19
```

```
# Display the first few rows of the dataset
print(tail(migration_data))
```

```
##      Year Quarter Origin N.L. P.E.I. N.S. N.B. Que. Ont. Man. Sask. Alta. B.C.
## 2580 2022       3  Sask.    8     15  135    8  104 1426  297     0  2608 1052
## 2581 2022       3  Alta.  370     98  682  406  975 3645  659  1572     0 4968
## 2582 2022       3   B.C.  123     78  742  373 1062 4204  537   951 10928    0
## 2583 2022       3   Y.T.    8      0   17    5    6   55    0    41   183   69
## 2584 2022       3 N.W.T.   15      4   21   71   42  154   18    10   146   70
## 2585 2022       3   Nvt.    5     15   31    5    4  198    0    21    84    0
##      Y.T. N.W.T. Nvt. Total
## 2580   14      3   15  5685
## 2581   65     65   26 13531
## 2582   50     57    0 19105
## 2583    0     30    0   414
## 2584   24      0   13   588
## 2585   13     11    0   387
```

```
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.1.1 --
```

```
## v broom        1.0.5     v recipes      1.0.8
## v dials        1.2.0     v rsample      1.2.0
## v dplyr        1.1.3     v tibble       3.2.1
## v ggplot2      3.4.4     v tidyr        1.3.0
## v infer        1.0.5     v tune         1.1.2
## v modeldata    1.2.0     v workflows    1.1.3
## v parsnip      1.1.1     v workflowsets 1.0.1
## v purrr        1.0.2     v yardstick    1.2.0
```

```
## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x ggplot2::%+%()   masks psych::%+%()
## x ggplot2::alpha() masks scales::alpha(), psych::alpha()
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Search for functions across packages at https://www.tidymodels.org/find/
```

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
glimpse(migration_data)
```

```
## Rows: 2,585
## Columns: 17
## $ Year    <int> 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971, 1971, 19~
## $ Quarter <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4,~
## $ Origin  <chr> "N.L.", "P.E.I.", "N.S.", "N.B.", "Que.", "Ont.", "Man.", "Sas~
## $ N.L.    <int> 0, 35, 596, 432, 373, 2436, 107, 43, 188, 150, 7, 10, 0, 29, 4~
## $ P.E.I.  <int> 39, 0, 283, 260, 92, 658, 105, 13, 80, 41, 2, 3, 50, 0, 210, 2~
## $ N.S.    <int> 378, 326, 0, 1199, 651, 3942, 178, 279, 431, 463, 0, 0, 575, 2~
## $ N.B.    <int> 279, 256, 1272, 0, 1346, 2967, 231, 127, 243, 298, 5, 8, 399, ~
## $ Que.    <int> 218, 77, 590, 942, 0, 7014, 526, 152, 478, 576, 22, 34, 432, 8~
## $ Ont.    <int> 1732, 563, 3754, 2785, 11692, 0, 3923, 1605, 3675, 4171, 49, 7~
## $ Man.    <int> 40, 34, 214, 124, 535, 2958, 0, 2039, 1410, 1361, 31, 48, 53, ~
## $ Sask.   <int> 20, 24, 90, 66, 123, 1053, 1692, 0, 2406, 1262, 20, 31, 26, 18~
## $ Alta.   <int> 87, 102, 518, 328, 821, 4444, 2562, 5586, 0, 6362, 242, 378, 1~
```

```
## $ B.C.    <int> 111, 52, 871, 384, 1365, 6287, 2556, 3025, 8816, 0, 241, 377, ~
## $ Y.T.    <int> 4, 5, 19, 7, 21, 153, 50, 94, 347, 260, 0, 0, 5, 3, 13, 5, 18,~
## $ N.W.T.  <int> 6, 7, 30, 11, 31, 238, 78, 147, 544, 410, 0, 0, 7, 4, 20, 8, 3~
## $ Nvt.    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ Total   <int> 2914, 1481, 8237, 6538, 17050, 32150, 12008, 13110, 18618, 153~
```

```r
# Check for missing values in the migration_data data frame
missing_values <- sapply(migration_data, function(x) sum(is.na(x)))

# Print the number of missing values for each column
print(missing_values)
```

```
##    Year Quarter  Origin    N.L.  P.E.I.    N.S.    N.B.    Que.    Ont.    Man.
##       0       0       0       0       0       0       0       0       0       0
##   Sask.   Alta.    B.C.    Y.T.  N.W.T.    Nvt.   Total
##       0       0       0       0       0       0       0
```

---

6 DATA PRE-PROCESSING

6.1 CREATE A NEW COLUMN BASED ON "Origin" Column ; a new column will be created with the full names of the Canadian provinces and territories using a lookup table.

```r
# Load the dplyr package if you haven't already
#library(dplyr)

# Create a data frame with the abbreviations and full names
state_data <- data.frame(
  Origin = c("N.L.", "P.E.I.", "N.S.", "N.B.", "Que.", "Ont.", "Man.", "Sask.", "Alta.", "B.C.", "Y.T."
  ProvinceNames = c("Newfoundland and Labrador", "Prince Edward Island", "Nova Scotia", "New Brunswick"
)
state_data
```

```
##    Origin             ProvinceNames
## 1    N.L. Newfoundland and Labrador
## 2  P.E.I.      Prince Edward Island
## 3    N.S.               Nova Scotia
## 4    N.B.             New Brunswick
## 5    Que.                    Quebec
## 6    Ont.                   Ontario
## 7    Man.                  Manitoba
## 8   Sask.              Saskatchewan
## 9   Alta.                   Alberta
## 10   B.C.          British Columbia
## 11   Y.T.                     Yukon
## 12 N.W.T.     Northwest Territories
## 13   Nvt.                   Nunavut
```

```r
# Use mutate to add the Full_Name column to the migration dataset
result_data <- migration_data %>%
  right_join(state_data, by = c("Origin" = "Origin"))%>%
  select(Year, Quarter, Origin, ProvinceNames, everything())
```

```r
result_data <- result_data %>%
  rename(
    `Newfoundland and Labrador` = "N.L.",
    `Prince Edward Island` = "P.E.I.",
    `Nova Scotia` = "N.S.",
    `New Brunswick` = "N.B.",
    `Quebec` = "Que.",
    `Ontario` = "Ont.",
    `Manitoba` = "Man.",
    `Saskatchewan` = "Sask.",
    `Alberta` = "Alta.",
    `British Columbia` = "B.C.",
    `Yukon` = "Y.T.",
    `Northwest Territories` = "N.W.T.",
    `Nunavut` = "Nvt."
  )

head(result_data, 5)
```

```
##   Year Quarter Origin             ProvinceNames Newfoundland and Labrador
## 1 1971       3   N.L. Newfoundland and Labrador                         0
## 2 1971       3 P.E.I.      Prince Edward Island                        35
## 3 1971       3   N.S.               Nova Scotia                       596
## 4 1971       3   N.B.             New Brunswick                       432
## 5 1971       3   Que.                    Quebec                       373
##   Prince Edward Island Nova Scotia New Brunswick Quebec Ontario Manitoba
## 1                   39         378           279    218    1732       40
## 2                    0         326           256     77     563       34
## 3                  283           0          1272    590    3754      214
## 4                  260        1199             0    942    2785      124
## 5                   92         651          1346      0   11692      535
##   Saskatchewan Alberta British Columbia Yukon Northwest Territories Nunavut
## 1           20      87               111     4                     6       0
## 2           24     102                52     5                     7       0
## 3           90     518               871    19                    30       0
## 4           66     328               384     7                    11       0
## 5          123     821              1365    21                    31       0
##   Total
## 1  2914
## 2  1481
## 3  8237
## 4  6538
## 5 17050
```

7 TREND ANALYSIS

7.1 Trends in Migration Over the years

```r
# Total migration over years
total_migration <- aggregate(. ~ Year, data = result_data[, -c(3:4)], FUN = sum, na.rm = TRUE)

# Plotting total migration over years
ggplotly(ggplot(total_migration, aes(x = Year, y = Total)) +
```

```
  geom_line() +
  labs(title = "Total Migration Over Years",
       x = "Year",
       y = "Total Migration"))
```

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

OBSERVATIONS : Immigrants trend across Canadian provinces doesn't follow a patterned trend.Although, the recent years have a range of migration count peaks (280k~330k) that can be seen in the plot above.

8 Quarterly migration trends by province

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(plotly)


# Filtering numeric columns from the original dataset
numeric_cols <- result_data %>%
  select(-c(Year, Quarter, ProvinceNames, Total)) %>%
  sapply(is.numeric)

# Extracting the names of numeric columns
numeric_col_names <- names(numeric_cols)[numeric_cols]

# Selecting the numeric columns and necessary identifier columns for melting
selected_cols <- c("Year", "Quarter", "ProvinceNames", "Total",numeric_col_names)
selected_data <- result_data[, selected_cols]

# Melting the selected numeric columns
melted_data <- selected_data %>%
  pivot_longer(cols = -c(Year, Quarter, ProvinceNames, Total), names_to = "Province", values_to = "Migra

# Plotting quarterly migration trends by province for numeric columns
plot_years <- ggplot(melted_data, aes(x = Quarter, y = MigrationCount, color = Province)) +
  geom_line(size = 1) +
  facet_wrap(~Province, scales = "free_y") +
  labs(title = "Quarterly Migration Trends Across Provinces",
       x = "Quarters",
       y = "Migration Count",
       color = "Province") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        axis.title = element_text(size = 12),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8))
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
# Interactive plot
interactive_plot <- ggplotly(plot_years, tooltip = c("x", "y", "color")) %>%
  layout(title = list(text = "Quarterly Migration Trends Across Provinces", size = 16),
         xaxis = list(title = "Quarters", tickfont = list(size = 10)),
         yaxis = list(title = "Migration Count", tickfont = list(size = 10)),
         legend = list(font = list(size = 8)))

interactive_plot
```

OBSERVATIONS : Not much of a pattern can be seen except for the fact that majority of the provinces have high migration phase going on across Quarter 2 & 3.

8.1 Migration Across Candidan Province

```r
library(openintro)
```

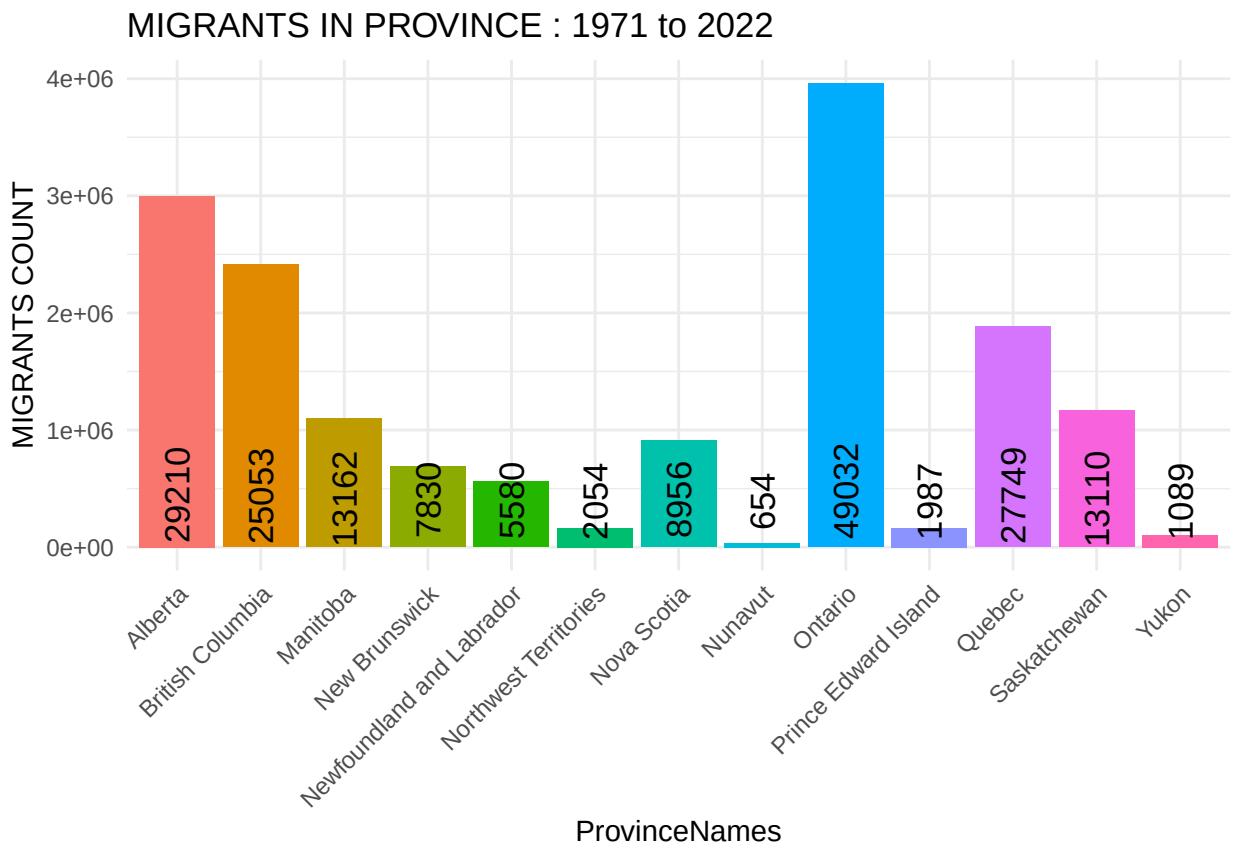
```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
##
## Attaching package: 'openintro'
## The following object is masked from 'package:modeldata':
##
##     ames
```

```r
library(ggplot2)
library(plotly)


bar_chart <-  ggplot(result_data, aes(x = ProvinceNames, y = Total, fill = ProvinceNames)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(
    title = "MIGRANTS IN PROVINCE : 1971 to 2022",
    y = "MIGRANTS COUNT"
  ) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


# Get the maximum count for each province
max_counts <- result_data %>%
  group_by(ProvinceNames) %>%
  summarise(Max_Count = max(Total))

# Add maximum count labels directly to the bars
bar_chart +
  geom_text(data = max_counts, aes(label = Max_Count, y = Max_Count+205000), size = 4.5,
            position = position_stack(vjust = 1.9), angle = 90,
            hjust = 0.5, vjust = 0.5,
            family = "sans") + theme(legend.position = "none")
```



8.2 Migration Across Candidan Province Across Three Year Ranges

```
# Dividing years into three equal ranges
range_1 <- result_data %>% filter(Year >= 1971 & Year <= 1991)
range_2 <- result_data %>% filter(Year >= 1992 & Year <= 2006)
range_3 <- result_data %>% filter(Year >= 2007 & Year <= 2022)

# Creating plots for each range
plot_range_1 <- ggplot(range_1, aes(x = ProvinceNames, y = Total, fill = ProvinceNames)) +
  geom_bar(stat = "identity") +
  labs(title = "1971-1991", x = "Province", y = "Count") +
  theme_minimal()+
  xlab("") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  geom_text(data = range_1 %>%
  group_by(ProvinceNames) %>%
  summarise(Max_Count = max(Total)), aes(label = Max_Count, y = Max_Count+205000), size = 3.5,
            position = position_stack(vjust = 1.9), angle = 90,
            hjust = 0.5, vjust = 0.5,
            family = "sans")


plot_range_2 <- ggplot(range_2, aes(x = ProvinceNames, y = Total, fill = ProvinceNames)) +
  geom_bar(stat = "identity") +
```

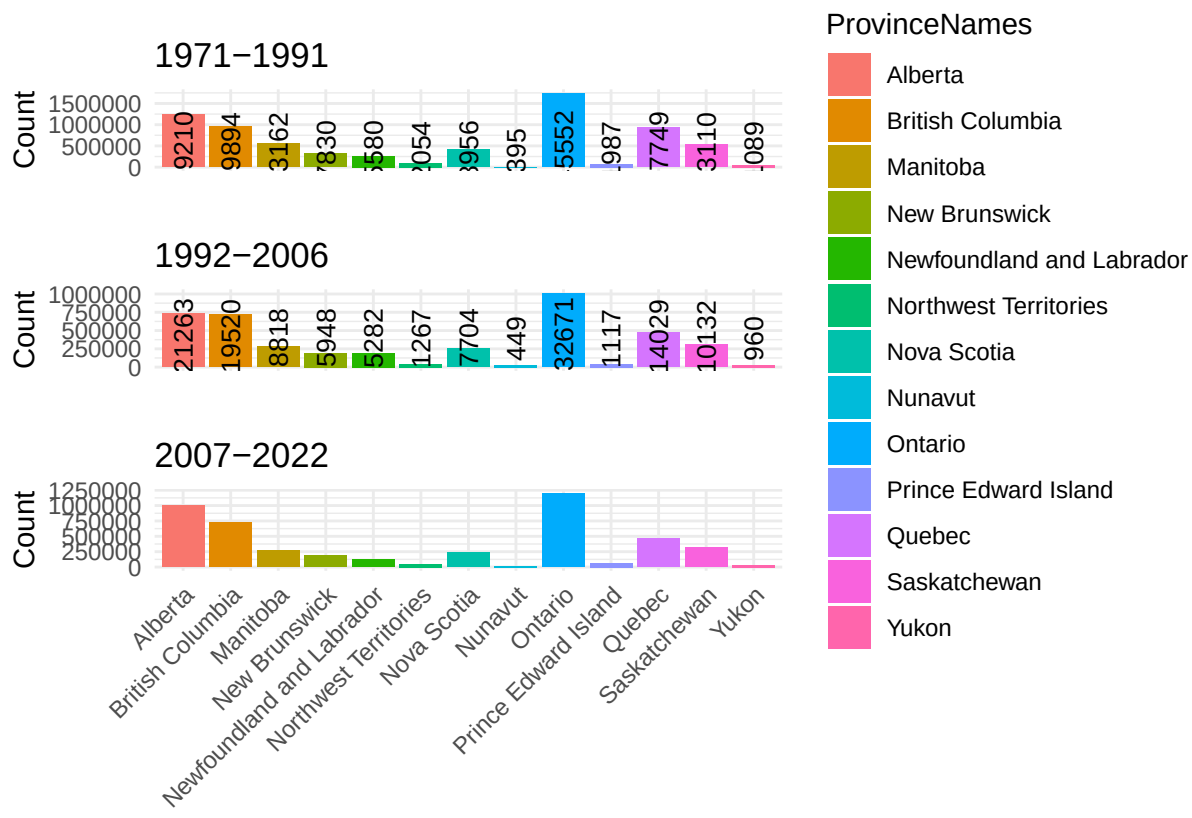
```
  labs(title = "1992-2006", x = "Province", y = "Count") +
  theme_minimal()+
  xlab("") +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank()) +
  geom_text(data = range_2 %>%
  group_by(ProvinceNames) %>%
  summarise(Max_Count = max(Total)), aes(label = Max_Count, y = Max_Count+205000), size = 3.5,
            position = position_stack(vjust = 1.9), angle = 90,
            hjust = 0.5, vjust = 0.5,
            family = "sans")


plot_range_3 <- ggplot(range_3, aes(x = ProvinceNames, y = Total, fill = ProvinceNames)) +
  geom_bar(stat = "identity") +
  labs(title = "2007-2022", x = "Province", y = "Count") +
  theme_minimal()+
  xlab("") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
  geom_text(data = range_3 %>%
  group_by(ProvinceNames) %>%
  summarise(Max_Count = max(Total)), aes(label = Max_Count, y = Max_Count+205000), size = 3.5,
            position = position_stack(vjust = 1.9), angle = 90,
            hjust = 0.5, vjust = 0.5,
            family = "sans")
```

```
## mapping: y = ~Max_Count + 205000, label = ~Max_Count
## geom_text: parse = FALSE, check_overlap = FALSE, na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_stack
```

```
library(plotly)
library(patchwork)

combined_plot <- plot_range_1+ theme(legend.position = "none") + plot_range_2 + plot_range_3 + theme(leg
combined_plot
```



OBSERVATIONS : "Ontario" is the province with highest migration count and "Nunavut" have the least migration count.

8.3 Migration Trends for Higher and Lower Migration Province Over the Quater.

```
# Finding the province with the highest migration and its corresponding quarter
highest_province <- melted_data %>%
  group_by(Province) %>%
  summarise(total_migration = sum(MigrationCount)) %>%
  arrange(desc(total_migration)) %>%
  slice(1)

highest_province_name <- highest_province$Province

highest_province_quarters <- melted_data %>%
  filter(Province == highest_province_name) %>%
  group_by(Quarter) %>%
  summarise(total_migration = sum(MigrationCount)) %>%
  arrange(desc(total_migration))

highest_province_quarters$Quarter <- factor(highest_province_quarters$Quarter, levels = unique(highest_

# Finding the province with the lowest migration and its corresponding quarter
lowest_province <- melted_data %>%
  group_by(Province) %>%
```

```
  summarise(total_migration = sum(MigrationCount)) %>%
  arrange(total_migration) %>%
  slice(1)

lowest_province_name <- lowest_province$Province

lowest_province_quarters <- melted_data %>%
  filter(Province == lowest_province_name) %>%
  group_by(Quarter) %>%
  summarise(total_migration = sum(MigrationCount)) %>%
  arrange(total_migration)

lowest_province_quarters$Quarter <- factor(lowest_province_quarters$Quarter, levels = unique(lowest_prov

highest_province_quarters$proportion <- highest_province_quarters$total_migration / sum(highest_provinc

custom_colors <- c("#17becf", "#1f77b4", "#bcbd22","#9467bd", "#e377c2", "#7f7f7f")


plot_ly(data = highest_province_quarters, labels = ~Quarter, values = ~proportion, type = 'pie', marker
  layout(title = paste("Migration Trends - Highest Province:", highest_province_name),
         showlegend = TRUE)
```

OBSERVATIONS : Ontario have the highest trends of migrations in the Quarter 2 & 3. This was concluded by the quarterly trends as well.

```
lowest_province_quarters$proportion <- lowest_province_quarters$total_migration / sum(lowest_province_q

custom_colors <- c("#17becf", "#1f77b4", "#bcbd22","#9467bd", "#e377c2", "#7f7f7f")

# Create the pie chart with custom colors
plot_ly(data = lowest_province_quarters, labels = ~Quarter, values = ~proportion, type = 'pie',
        marker = list(colors = custom_colors)) %>%
  layout(title = paste("Migration Trends - Lowest Province:", lowest_province_name),
         showlegend = TRUE)
```

OBSERVATIONS : Nunavut have the highest trends of migrations in the Quarter 2 & 1.

---

9 HYPOTHESIS TESTING

HYPOTHESIS 1 : People tend to migrate more frequently either yearly or quarterly between Canadian provinces.

Null Hypothesis (H0): There is no significant association between the year/quarter and the frequency of migration between Canadian provinces.

Alternate Hypothesis (H1): There is a significant association between the year/quarter and the frequency of migration between Canadian provinces.

9.1 To test this hypothesis, are using a Chi-Square (^{2}) test

We will organize the data into a contingency table, where rows represent one variable (e.g., year/quarter) and columns represent the other variable (e.g., provinces), with cell values being the frequencies of migration between provinces in each specific year/quarter.

```r
# Selecting the relevant columns for the contingency table
data_for_test <- result_data[, c("Year", "Quarter", "Newfoundland and Labrador", "Prince Edward Island"

# Creating a contingency table
contingency_table <- table(data_for_test$Year, data_for_test$Quarter)

# Performing Chi-Square test
chi_square_result <- chisq.test(contingency_table)

#Test result
print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 36.932, df = 153, p-value = 1
```

OBSERVATIONS :

1. *Chi-squared Value:* The calculated chi-squared value is 36.932.
2. *Degrees of Freedom (df):* The degrees of freedom are 153.
3. *P-Value:* The obtained p-value is approximately 1.

- Chi-squared Value Interpretation: The chi-squared value of 36.932 indicates the magnitude of the difference between the observed and expected frequencies within the contingency table.
- Degrees of Freedom: With 153 degrees of freedom, this test has considered a significant number of categories and observations.
- P-Value Interpretation: The p-value of approximately 1 suggests that **there's insufficient evidence to reject the null hypothesis**. A p-value of 1 indicates very high probability under the null hypothesis. It implies that there is no significant association between the years/quarters and the migration counts among Canadian provinces based on the provided data.

In summary, based on this test's results, there doesn't appear to be a significant relationship between the timing (years/quarters) and the migration counts among Canadian provinces.

HYPOTHESIS 2 : Certain provinces consistently attract more migrants and can be recognized as the most favorable destinations both annually and quarterly. HYPOTHESIS 3 : Certain provinces consistently attract less migrants and can be recognized as the least favorable destinations both annually and quarterly.

Null Hypothesis (H0): There is no significant difference in the average number of migrants attracted by different provinces both annually and quarterly.

Alternate Hypothesis (H1): Certain provinces consistently attract a significantly different number of migrants, establishing themselves as the more or less favorable destinations both annually and quarterly.

9.2 To test this hypothesis,we are using a t-test

```r
library(dplyr)

# Finding provinces with the highest total migration
highest_total <- result_data %>%
  arrange(desc(Total)) %>%
```

```
  distinct(ProvinceNames, .keep_all = TRUE) %>%
  head(2) %>%
  select(ProvinceNames, Total)

# Finding provinces with the lowest total migration
lowest_total <- result_data %>%
  arrange(Total) %>%
  distinct(ProvinceNames, .keep_all = TRUE) %>%
  head(2) %>%
  select(ProvinceNames, Total)

# Extracting province names into lists
highest_province_list <- as.list(highest_total$ProvinceNames)
lowest_province_list <- as.list(lowest_total$ProvinceNames)

cat("Provinces with the highest total migration:\n")
```

```
## Provinces with the highest total migration:
```

```
print(highest_province_list)
```

```
## [[1]]
## [1] "Ontario"
##
## [[2]]
## [1] "Alberta"
```

```
cat("\nProvinces with the lowest total migration:\n")
```

```
##
## Provinces with the lowest total migration:
```

```
print(lowest_province_list)
```

```
## [[1]]
## [1] "Nunavut"
##
## [[2]]
## [1] "Yukon"
```

```
# Selecting columns of provinces for comparison
provinces <- c('Ontario', 'Alberta' , 'Nunavut', 'Yukon')

# Creating an empty list to store t-test results
t_test_results <- list()

# Looping through combinations of provinces for t-tests
for (i in 1:(length(provinces) - 1)) {
  for (j in (i + 1):length(provinces)) {
    # Selecting migration data for two provinces
    province1 <- result_data[[provinces[i]]]
    province2 <- result_data[[provinces[j]]]

    # Performing a t-test between the two provinces
    t_test_result <- t.test(province1, province2)
```

```
    # Storing t-test result in the list
    comparison <- paste(provinces[i], "-", provinces[j])
    t_test_results[[comparison]] <- t_test_result
  }
}

# Printing the results of all t-tests
for (comparison in names(t_test_results)) {
  cat(comparison, ":\n")
  print(t_test_results[[comparison]])
  cat("\n")
  cat("--------------------------------------------\n")
}
```

```
## Ontario - Alberta :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = 2.0153, df = 5165.8, p-value = 0.04393
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    2.993842 217.068054
## sample estimates:
## mean of x mean of y
##  1513.649  1403.618
##
##
## --------------------------------------------
## Ontario - Nunavut :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = 39.311, df = 2584.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1427.012 1576.849
## sample estimates:
##  mean of x  mean of y
## 1513.64913   11.71876
##
##
## --------------------------------------------
## Ontario - Yukon :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = 38.631, df = 2589, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1401.606 1551.505
## sample estimates:
```

```
##  mean of x  mean of y
## 1513.64913   37.09362
##
##
## ------------------------------------------
## Alberta - Nunavut :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = 35.682, df = 2584.7, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1315.407 1468.391
## sample estimates:
##  mean of x  mean of y
## 1403.61818   11.71876
##
##
## ------------------------------------------
## Alberta - Yukon :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = 35.017, df = 2588.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1290.002 1443.047
## sample estimates:
##  mean of x  mean of y
## 1403.61818   37.09362
##
##
## ------------------------------------------
## Nunavut - Yukon :
##
##  Welch Two Sample t-test
##
## data:  province1 and province2
## t = -20.052, df = 3274, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -27.85606 -22.89365
## sample estimates:
## mean of x mean of y
##  11.71876  37.09362
##
##
## ------------------------------------------
```

OBSERVATIONS :

1. *Ontario vs. Alberta:* There's a statistically significant difference in the average number of migrants attracted by Ontario and Alberta, with Ontario having a slightly higher average. The p-value of

0.04393 suggests a moderate level of confidence in rejecting the null hypothesis, indicating a potential difference in migration numbers between these provinces.

2. *Ontario vs. Nunavut / Yukon:* Comparing Ontario with Nunavut and Yukon shows a substantial difference in migration numbers. e p-values are extremely low (p-value < 2.2e-16), indicating an exceedingly high level of confidence in rejecting the null hypothesis. Ontario attracts significantly more migrants compared to Nunavut and Yukon.
3. *Alberta vs. Nunavut / Yukon:* Similar to Ontario, Alberta attracts significantly more migrants compared to Nunavut and Yukon. The p-values are very low, indicating strong evidence against the null hypothesis.
4. *Nunavut vs. Yukon:* There's a significant difference in migration numbers between Nunavut and Yukon. The p-value is extremely low, indicating a clear distinction in migration patterns between these territories.

These results support the alternate hypothesis, suggesting that certain provinces consistently attract significantly different numbers of migrants. Ontario and Alberta emerge as leading destinations, drawing considerably higher migration numbers compared to Nunavut and Yukon. Additionally, Nunavut and Yukon showcase notably lower migration numbers compared to the provinces, indicating a distinct migration pattern between territories and provinces.

HYPOTHESIS 4 : There is a changing trend in interprovincial migration patterns over the years, reflecting evolving factors that influence people's decisions to relocate within Canada.

Null Hypothesis (H0): There is no difference in the distribution of interprovincial migration patterns over the years.

Alternate Hypothesis (H1): There exists a difference in the distribution of interprovincial migration patterns over the years.

9.3 To test this hypothesis,we are using a Kolmogorov-Smirnov test (KS test)

```
# Aggregate Migration Count by Year in each range
sum_range_1 <- aggregate(Total ~ Year, data = range_1, FUN = sum)
sum_range_2 <- aggregate(Total ~ Year, data = range_2, FUN = sum)
sum_range_3 <- aggregate(Total ~ Year, data = range_3, FUN = sum)
# Performing KS tests between aggregated migration counts
ks_test_range1_range2 <- ks.test(sum_range_1$Total, sum_range_2$Total)
ks_test_range1_range3 <- ks.test(sum_range_1$Total, sum_range_3$Total)
ks_test_range2_range3 <- ks.test(sum_range_2$Total, sum_range_3$Total)
# Printing the test results
print(ks_test_range1_range2)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  sum_range_1$Total and sum_range_2$Total
## D = 0.7619, p-value = 1.628e-05
## alternative hypothesis: two-sided
```

```
cat("---------------------------------")
```

```
## ---------------------------------
```

```
print(ks_test_range1_range3)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  sum_range_1$Total and sum_range_3$Total
## D = 0.5744, p-value = 0.002707
## alternative hypothesis: two-sided
```

```
cat("---------------------------------")
```

```
## ---------------------------------
```

```
print(ks_test_range2_range3)
```

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  sum_range_2$Total and sum_range_3$Total
## D = 0.3625, p-value = 0.1879
## alternative hypothesis: two-sided
```

OBSERVATIONS :

The Kolmogorov-Smirnov (KS) test results suggest interesting findings regarding the distribution of inter-provincial migration patterns across the specified time ranges.

Range 1 vs. Range 2: The KS test between the migration counts of Range 1 (1971-1991) and Range 2 (1992-2006) indicates a significant difference in their distributions (p-value = 1.628e-05). This suggests that there's a notable change or shift in migration patterns between these time periods.

Range 1 vs. Range 3: The KS test between the migration counts of Range 1 (1971-1991) and Range 3 (2007-2022) also reveals a significant difference in their distributions (p-value = 0.002707). This further emphasizes a substantial shift or alteration in migration trends over time.

Range 2 vs. Range 3: However, the KS test between the migration counts of Range 2 (1992-2006) and Range 3 (2007-2022) does not show a significant difference (p-value = 0.1879). This suggests that the migration patterns between these periods might be relatively similar.

- Given these results, the evidence supports rejecting the null hypothesis for comparisons between Range 1 vs. Range 2 and Range 1 vs. Range 3. This implies that there are notable differences in migration distributions between these time frames.
- However, for Range 2 vs. Range 3, the p-value is not significant, suggesting that the distributions might not differ significantly between these periods.

Thus, there is support for the alternative hypothesis (H1) that indicates differences in migration distributions between certain periods, signifying changing trends in interprovincial migration patterns over the years.

---

10 CONCLUSION

The project on analyzing changing interprovincial migration trends in Canada from 1971 to 2022 has yielded valuable insights through various statistical tests, shedding light on different aspects of migration patterns within the country.

Chi-Square Test:

The Chi-Square test outcomes suggest that there isn't a significant relationship between the timing (years/quarters) and migration counts among Canadian provinces. This implies that migration trends might not be notably influenced by specific temporal patterns on a yearly or quarterly basis.

T-Test Findings:

The T-test results strongly support the alternate hypothesis, highlighting that specific provinces consistently attract significantly higher numbers of migrants. Ontario and Alberta emerge as primary destinations, drawing notably higher migration figures compared to Nunavut and Yukon. This emphasizes distinct migration patterns between territories and provinces, with the latter attracting considerably more migrants.

KS Test Insights:

The Kolmogorov-Smirnov (KS) test results provide evidence supporting the idea of changing trends in interprovincial migration patterns over time. Notably, there are significant differences in migration distributions between certain periods, indicating evolving migration trends across the years studied.

In conclusion, these comprehensive statistical analyses offer nuanced insights into Canada's interprovincial migration dynamics. They uncover the influence of specific provinces in attracting migrants, highlight the absence of clear temporal migration patterns, and confirm the presence of changing trends in migration distributions over the studied decades. This research contributes significantly to understanding population mobility within Canada, with implications across diverse fields such as economics, sociology, policy development, and urban planning.

---

11 REFERNCES

1. DATASET Source : Estimates of interprovincial migrants by province or territory of origin and destination, quarterly and yearly in Canada from 1971 to 2022. (https://catalogue.data.gov.bc.ca/dataset/inter-provincial-and-international-migration/resource/f6171cc3-3845-40dd-9855-d87e8f524064/)
2. ADSC1910_01 - Applied Data Science/ Lecture Notes
3. Hypothesis Testing : https://www.r-bloggers.com/2022/12/hypothesis-testing-in-r/
4. Interactive Plots in R : https://r-graph-gallery.com/interactive-charts.html