TP4_Web scraping

Svetlana Zhuk

2024-03-24

Introduction :

The issue of immigration is always of paramount importance in the politics of Quebec, as it is directly related to the protection of French identity in Quebec, the sole French-speaking province in Canada. Since the 1991 Canada–Quebec Accord, Quebec has gained more autonomy over the selection and integration of its immigrants Statistics Canada. And now we are spectators of the continuation of this debate according to which Quebec seeks to fully control the immigration power, arguing that the province has exceeded its capacity to welcome new arrivals, but that the federal government refuses to give full autonomy to the Quebec CBC Canada. Since the Coalition Avenir Québec (CAQ) party came to power, Premier François Legault has opposed an increase in immigration despite the federal's plan, arguing contrary to the federal government's plan, stating, "We're different than the rest of North America. And it's important to protect French in the future to ensure that newcomers speak French"Toronto.CityNews.

This raises the question about the etnic composition of immigrants arriving in Quebec. We propose to conduct a description analysis of etnic composition of immigrants from the beginning of 2001 to the pre-pandemic year of 2019, focusing on the research question:

- **What are the top countries of origin for immigrants to Quebec between 2001 and 2019?**

Subsequently, we will explore the relationship between political trends and immigration during the same period:

- **How does the political orientation of Quebec's governing parties (Parti Libéral du Québec, Parti Québécois, Coalition Avenir Québec) affect the trends in immigration from the top source countries to Quebec over time?**

Thus, understanding the comprehensive picture of ethnic immigration trends to Quebec and their relation to political trends will enrich our comprehension of multiculturalism and diversity in Quebec. This analysis will provide a foundation for further discussions on how this diversity can be harmonized with Quebecois identity.

Données et méthodes :

To address our research question, we employed a web scraping technique to access data on immigration to Quebec for the period from 2001 to 2019. Our primary datasets were derived from two Wikipedia pages. Our motivation for web scraping data from Wikipedia stems from its status as a verified source that cites its references. The organized presentation of data tables on Wikipedia facilitates the scraping process, making it accessible, easy to replicate, and verifiable.

After web scraping and cleaning the data, we produced four key datasets:

- **top_countries_by_year**
- **pm_quebec_clean_test**
- **dat_imm_parti_final**
- **top5_countries_party_2001_2019_final**

The dataset named **top_countries_by_year** and **imm_quebec_2001_2019_clean**, which details immigration from various countries to Quebec, was sourced from Immigration au Québec. The Wikipedia page references data from the Ministère des Relations avec les citoyens et de l'Immigration, specifically: 'Immigrants admis au Québec selon les 15 principaux pays de naissance.

The dataset **pm_quebec_clean_test**, which provides the year and the party that was in power, was sourced from the web page Premier ministre du Québec.

By merging these datasets into **dat_imm_parti_final** and **top5_countries_party_2001_2019_final**, we were able to produce visualizations that trace immigration trends, thereby facilitating an answer to our research question.

Dataset_1 : top_countries_by_year

The dataset **top_countries_by_year** was created by filtering the top 5 countries from **imm_quebec_2001_2019_clean**. This latter dataset was generated through web scraping and consolidating data 'Immigrants admis au Québec selon les 15 principaux pays de naissance.'

After consolidating the information from all years from 2001 to 2019 into the single dataset **imm_quebec_2001_2019_clean**, we focused on converting the variables *Nombre_immigrants* and *Proportion_immigrants* to numeric types. The primary challenge in transforming these data was removing spaces in the numbers and replacing commas with periods.

The key variables crucial for further analysis are:

- pays: categorical variable

- année: 2001 to 2019
- nombre_immigrants: continuous variable
- For a step-by-step description, check here: Dataset: top_countries_by_year
- For a glimpse of the data, check here: Glimpse: imm_quebec_2001_2019_clean
- For filtering the top 5 countries, check here: Nettoyage & Glimpse: top_countries_by_year

Dataset_2 : pm_quebec_clean_test

This dataset delineates the parties and their respective years in government, derived from the web-scraped dataset **pm_quebec**. A primary challenge involved separating the information within the 'Nom' column into three distinct columns: 'Name', 'DateOfBirth', and 'Party'. Additionally, we split the data in the 'Législatures et mandats' column into two columns: 'start_year' and 'end_year'. Given that we scraped information on all Premiers of Quebec since 1867, we refined our dataset to include only the years from 2001 onwards, inclusive of 2001 itself. Furthermore, we reclassified the categories of the 'Party' variable for clarity.

The key variables crucial for further analysis are:

- Party: PQ, PLQ, CAQ
- year: 2001 to 2019
- For a more detailed, step-by-step cleaning process, check here: Dataset: pm_quebec_clean_test.
- For a glimpse of the final dataset, check here: Glimpse: pm_quebec_clean_test.

Dataset_3 : dat_imm_parti_final

This dataset results from merging two datasets: **imm_quebec_total** and **pm_quebec_clean_test**, using the 'year' variable as the key for merging. It was crucial that the 'year' variable had the same name and was of the numeric type in both datasets.

imm_quebec_total is derived from the dataset imm_quebec_2001_2019_clean. We renamed the variable 'année' to 'year' and created a new column to capture the total number of immigrant arrivals from 15 countries. This modification facilitates further visualization of immigration trends in conjunction with the political context.

After merging the two datasets, the number of observations doubled. We used the distinct() function to filter for unique values.

The key variables crucial for further analysis are:

- year: 2001 to 2019

- nombre_imm_total: continuous variable
- party: PQ, PLQ, CAQ
- For a more detailed, step-by-step cleaning process, check here: Dataset 3: dat_imm_parti_final
- For a glimpse of the final dataset, see here: Glimpse: dat_imm_parti_final

Dataset_4 : top5_countries_party_2001_2019_final

For the final visualization of the ethnic composition from the top 5 countries of origin for immigrants to Quebec, and the political context under which immigration occurred, we utilized this dataset.

We merged the two datasets, **top5_2001_2019** and **pm_quebec_clean_test**, by the 'year' column. To address the duplication of observations resulting from the merge, we used the distinct() function to retain only unique combinations of variables. Subsequently, we removed any rows that were unnecessary.

The key variables crucial for further analysis:

- pays: a categorical variable representing countries.
- year: 2001 : 2019.
- nombre_immigrants: a continuous variable.
- Party: categorical variable PQ, PLQ, CAQ.
- For a more detailed, step-by-step cleaning process, check here: Dataset: top5_countries_party_2001_2019
- For a glimpse of the final dataset, check here: Glimpse: top5_countries_party_2001_2019_final

Les resultats :

Fig. 1 :

The first table highlights a central aspect of our analysis: the trends in immigration from the top 5 countries to Quebec from 2001 to 2019. We note a continuous flow of immigrants from countries such as Algeria and France, both considered Francophone, throughout the entire period, and, notably, from China as well. The graph depicts immigration arrivals from Romania during the period of 2002 to 2005, from Iran from 2014 to 2016, from Syria from 2015 to 2019, and from India from 2017 to 2019.

The graph serves as a foundation for further analysis, allowing us to concentrate on specific time periods for certain countries to better understand their immigration trends. For instance, the notable increase in immigration from Syria from 2015 to 2019 is directly related to the Syrian Civil War.

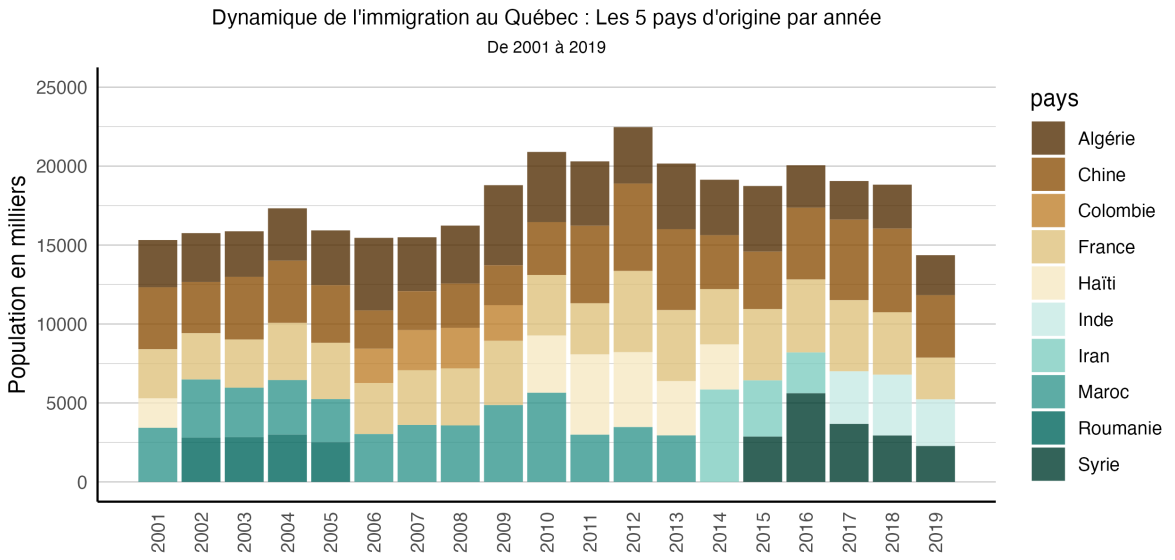


Figure 1: fig. 1

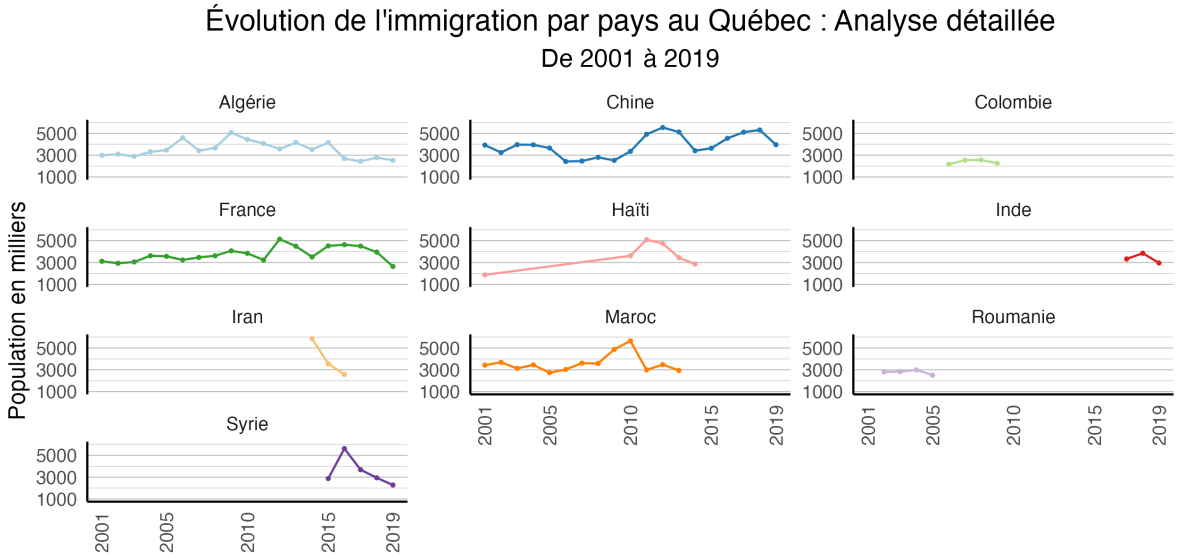


Figure 2: fig. 2

Fig. 2 :

In Figure 2, we see a detailed analysis of the previous graph showcasing the top 5 countries per year, but with specific trends for each country during the period from 2001 to 2019. The dots on the graph represent the number of immigrants arriving each year. Thus, it consistently shows that the top constant sources of immigration to Quebec are Algeria, China, and France. Haiti experienced a noticeable spike in immigration from 2010 to 2014, likely corresponding to the earthquake.

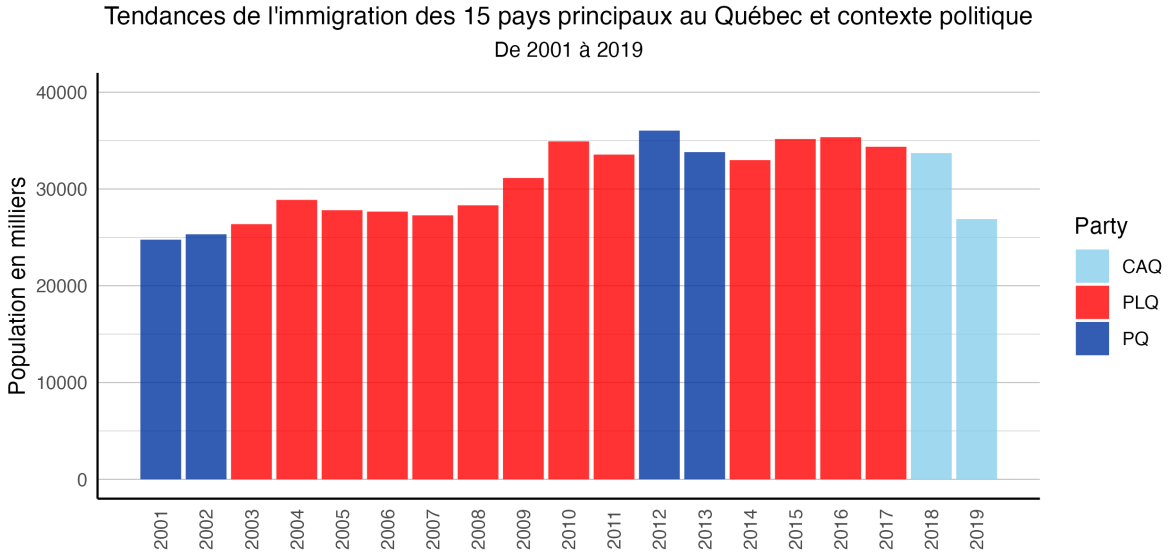


Figure 3: fig.3

Fig 3 :

We proceed to analyze our second area of interest: the immigration trends and political context. The graph consistently illustrates the total number of immigrants arriving from the key 15 countries. We observe an increase in immigration during the period when the PLQ was in power, and a corresponding decrease in immigration trends following the CAQ's rise to power. However, it's important to note that this data does not represent a complete observation of overall immigration trends but rather highlights information for the top 15 countries as identified by the Ministry of Immigration.

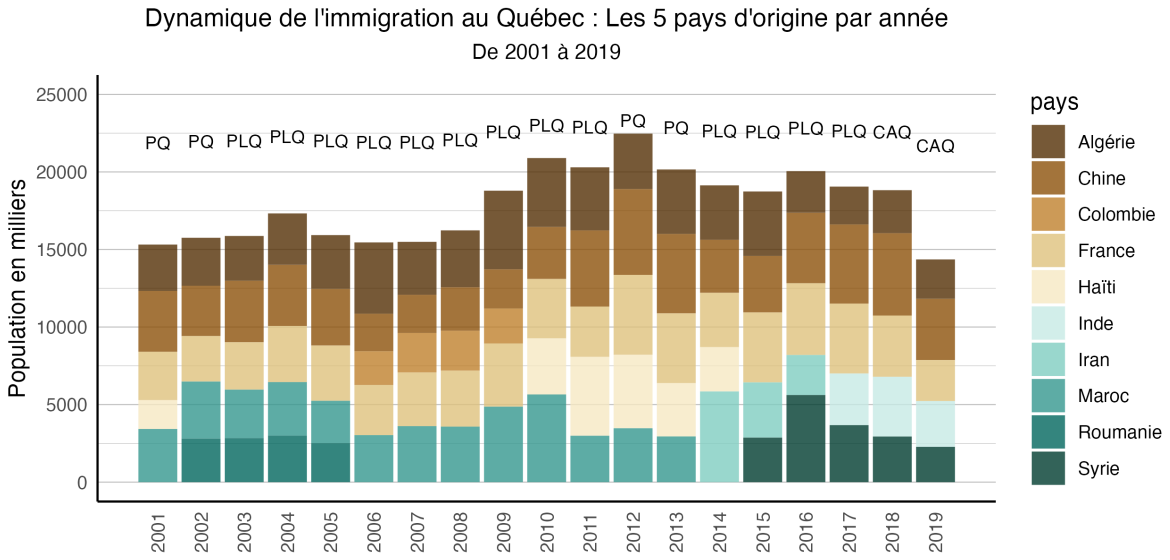


Fig 4 :

Figure 4 continues the analysis presented in the first figure. It depicts the ethnic composition of immigrants from the top 5 countries arriving in Quebec and the political context, including which party was in power.

Although a decrease in immigration is observed after the CAQ came to power—a move aligned with their electoral campaign promises to protect the French language and reduce the number of immigrants—it is not conclusive from the ethnic composition data of 2019 that priority was exclusively given to countries considered Francophone.

While the party in power can influence immigration policies, the graph suggests that the ethnic composition of immigrants is more significantly affected by external factors, such as civil wars or earthquakes.

The visualization does not demonstrate a clear correlation between the party in power and the ethnic composition of immigrant arrivals.

Conclusion :

In conclusion, motivated by recent news of the Quebec government's demand for full control over immigration, we aimed to conduct a detailed analysis of the ethnic composition of immigrants arriving from the top 5 countries to Quebec. Our goal was to identify the key source countries and determine whether these sources of immigration changed depending on the party in power. We discovered that the immigration flow from three countries—Algeria, France, and China—remained constant for the period from 2001 to 2019. However, the visualization did

not reveal any specific trends regarding the political context's influence on the ethnic composition of immigrant arrivals. Nonetheless, we observed a tendency for the number of immigrants to decrease after the CAQ came to power.

Since our primary method of data collection is web scraping, discussing the ethical aspects is necessary. Wikipedia is open source and is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. The license provides a human-readable summary, stating:

'You are free: - To Share — copy and redistribute the material in any medium or format - To Adapt — remix, transform, and build upon the material for any purpose, even commercially.'

In conducting our web scraping and data analysis, we share the datasets we've produced to ensure the transparency and reproducibility of our analysis, with proper citation included. According to this license, no rights are violated. Hooray!

Annexe : code de nettoyage

```
# libraries
library(tidyverse)
suppressMessages(library(tidyverse))
library(lubridate)
library(rvest)
```

Dataset : top_countries_by_year

```
#télécharger d'abord toutes les tables que nous voulions web scraping

imm_quebec <- read_html("https://fr.wikipedia.org/wiki/Immigration_au_Québec")

imm_quebec_1 <- imm_quebec |>
  html_elements("table") |>
  pluck(1) |>
  html_table(fill = T)

imm_quebec_2 <- imm_quebec |>
  html_elements("table") |>
  pluck(2) |>
  html_table(fill = T)

imm_quebec_3 <- imm_quebec |>
```

```
  html_elements("table") |>
  pluck(3) |>
  html_table(fill = T)

imm_quebec_4 <- imm_quebec |>
  html_elements("table") |>
  pluck(4) |>
  html_table(fill = T)
```

```
# Nettoyage 2001_2005
imm_quebec_1 <- imm_quebec_1 |> slice(-c(18, 19, 1, 2))

imm_quebec_1_clean2001 <- imm_quebec_1 |>
  select(X2, X3, X4) |> mutate(year = 2001) |>
  rename(Pays = X2,
         Nombre_immigrants = X3,
         Proportion_immigrants = X4)

imm_quebec_1_clean2002 <- imm_quebec_1 |>
  select(X5, X6, X7) |> mutate(year = 2002) |>
  rename(Pays = X5,
         Nombre_immigrants = X6,
         Proportion_immigrants = X7)

imm_quebec_1_clean2003 <- imm_quebec_1 |>
  select(X8, X9, X10) |>  mutate(year = 2003) |>
    rename(Pays = X8,
         Nombre_immigrants = X9,
         Proportion_immigrants = X10)

imm_quebec_1_clean2004 <- imm_quebec_1 |>
  select(X11, X12, X13) |> mutate(year = 2004) |>
  rename(Pays = X11,
         Nombre_immigrants = X12,
         Proportion_immigrants = X13)

imm_quebec_1_clean2005 <- imm_quebec_1 |>
  select(X14, X15, X16) |> mutate(year = 2005) |>
  rename(Pays = X14,
         Nombre_immigrants = X15,
         Proportion_immigrants = X16)
```

```
Data_combined_2001_2005 <- bind_rows(imm_quebec_1_clean2001,
                                     imm_quebec_1_clean2002,
                                     imm_quebec_1_clean2003,
                                     imm_quebec_1_clean2004,
                                     imm_quebec_1_clean2005)

Data_combined_2001_2005 <- Data_combined_2001_2005 |>
  select(Pays, year, Nombre_immigrants, Proportion_immigrants)
```

```
#Nettoyage 2006_2010

imm_quebec_2 <- imm_quebec_2 |> slice(-c(18, 19, 1, 2))

imm_quebec_2_clean2006 <- imm_quebec_2 |>
  select(X2, X3, X4) |> mutate(year = 2006) |>
  rename(Pays = X2,
         Nombre_immigrants = X3,
         Proportion_immigrants = X4)

imm_quebec_2_clean2007 <- imm_quebec_2 |>
  select(X5, X6, X7) |> mutate(year = 2007) |>
  rename(Pays = X5,
         Nombre_immigrants = X6,
         Proportion_immigrants = X7)

imm_quebec_2_clean2008 <- imm_quebec_2 |>
  select(X8, X9, X10) |>  mutate(year = 2008) |>
    rename(Pays = X8,
         Nombre_immigrants = X9,
         Proportion_immigrants = X10)


imm_quebec_2_clean2009 <- imm_quebec_2 |>
  select(X11, X12, X13) |> mutate(year = 2009) |>
  rename(Pays = X11,
         Nombre_immigrants = X12,
         Proportion_immigrants = X13)

imm_quebec_2_clean2010 <- imm_quebec_2 |>
  select(X14, X15, X16) |> mutate(year = 2010) |>
```

```
  rename(Pays = X14,
         Nombre_immigrants = X15,
         Proportion_immigrants = X16)

Data_combined_2006_2010 <- bind_rows(imm_quebec_2_clean2006,
                                     imm_quebec_2_clean2007,
                                     imm_quebec_2_clean2008,
                                     imm_quebec_2_clean2009,
                                     imm_quebec_2_clean2010)

Data_combined_2006_2010 <- Data_combined_2006_2010 |>
  select(Pays, year, Nombre_immigrants, Proportion_immigrants)
```

```
#Nettoyage 2011_2015

imm_quebec_3 <- imm_quebec_3 |> slice(-c(18, 19, 1, 2))

imm_quebec_3_clean2011 <- imm_quebec_3 |>
  select(X2, X3, X4) |> mutate(year = 2011) |>
  rename(Pays = X2,
         Nombre_immigrants = X3,
         Proportion_immigrants = X4)

imm_quebec_3_clean2012 <- imm_quebec_3 |>
  select(X5, X6, X7) |> mutate(year = 2012) |>
  rename(Pays = X5,
         Nombre_immigrants = X6,
         Proportion_immigrants = X7)

imm_quebec_3_clean2013 <- imm_quebec_3 |>
  select(X8, X9, X10) |>  mutate(year = 2013) |>
    rename(Pays = X8,
         Nombre_immigrants = X9,
         Proportion_immigrants = X10)

imm_quebec_3_clean2014 <- imm_quebec_3 |>
  select(X11, X12, X13) |> mutate(year = 2014) |>
  rename(Pays = X11,
         Nombre_immigrants = X12,
         Proportion_immigrants = X13)
```

```r
imm_quebec_3_clean2015 <- imm_quebec_3 |>
  select(X14, X15, X16) |> mutate(year = 2015) |>
  rename(Pays = X14,
         Nombre_immigrants = X15,
         Proportion_immigrants = X16)

Data_combined_2011_2015 <- bind_rows(imm_quebec_3_clean2011,
                                     imm_quebec_3_clean2012,
                                     imm_quebec_3_clean2013,
                                     imm_quebec_3_clean2014,
                                     imm_quebec_3_clean2015)

Data_combined_2011_2015 <- Data_combined_2011_2015 |>
  select(Pays, year, Nombre_immigrants, Proportion_immigrants)
```

```r
#Nettoyage 2016_2019

imm_quebec_4 <- imm_quebec_4 |> slice(-c(18, 19, 1, 2))

imm_quebec_4_clean2016 <- imm_quebec_4 |>
  select(X2, X3, X4) |> mutate(year = 2016) |>
  rename(Pays = X2,
         Nombre_immigrants = X3,
         Proportion_immigrants = X4)

imm_quebec_4_clean2017 <- imm_quebec_4 |>
  select(X5, X6, X7) |> mutate(year = 2017) |>
  rename(Pays = X5,
         Nombre_immigrants = X6,
         Proportion_immigrants = X7)

imm_quebec_4_clean2018 <- imm_quebec_4 |>
  select(X8, X9, X10) |>  mutate(year = 2018) |>
    rename(Pays = X8,
         Nombre_immigrants = X9,
         Proportion_immigrants = X10)

imm_quebec_4_clean2019 <- imm_quebec_4 |>
  select(X11, X12, X13) |> mutate(year = 2019) |>
  rename(Pays = X11,
         Nombre_immigrants = X12,
```

```
         Proportion_immigrants = X13)

Data_combined_2016_2019 <- bind_rows(imm_quebec_4_clean2016,
                                     imm_quebec_4_clean2017,
                                     imm_quebec_4_clean2018,
                                     imm_quebec_4_clean2019)


Data_combined_2016_2019 <- Data_combined_2016_2019 |>
  select(Pays, year, Nombre_immigrants, Proportion_immigrants)
```

```
#la fusion des données

imm_quebec_2001_2019 <- bind_rows(Data_combined_2001_2005,
                                  Data_combined_2006_2010,
                                  Data_combined_2011_2015,
                                  Data_combined_2016_2019)


#transformer la variable "nombre_immigrant" en numérique

imm_quebec_2001_2019_clean <- imm_quebec_2001_2019 |>
  rename(année = year) |>
  arrange(Pays, desc(année)) |>
  mutate(across(c(Nombre_immigrants, Proportion_immigrants),
                ~ as.numeric(
                  str_replace_all(
                    str_squish(.x),
                    " ", "") |>
                  str_replace_all(",", "."))))

class(imm_quebec_2001_2019_clean$Nombre_immigrants)
```

```
[1] "numeric"
```

```
#change les noms de colonnes en minuscules

colnames(imm_quebec_2001_2019_clean) <- tolower(colnames(imm_quebec_2001_2019_clean))
```

Glimpse : imm_quebec_2001_2019_clean

```
imm_quebec_2001_2019_clean |> glimpse()
```

```
Rows: 285
Columns: 4
$ pays                  <chr> "Algérie", "Algérie", "Algérie", "Algérie", "Alg~
$ année                 <dbl> 2019, 2018, 2017, 2016, 2015, 2014, 2013, 2012, ~
$ nombre_immigrants     <dbl> 2531, 2772, 2438, 2681, 4155, 3513, 4155, 3572, ~
$ proportion_immigrants <dbl> 6.2, 5.4, 4.7, 5.0, 5.5, 7.0, 8.0, 6.5, 7.9, 8.2~
```

Nettoyage & Glimpse : top_countries_by_year

```
#choisir les 5 principaux pays d'origine des immigrants
top_countries_by_year <- imm_quebec_2001_2019_clean |>
  arrange(année, desc(nombre_immigrants)) |>
  group_by(année) |>
  slice_max(order_by = nombre_immigrants, n = 5)

summary(top_countries_by_year$nombre_immigrants)
```

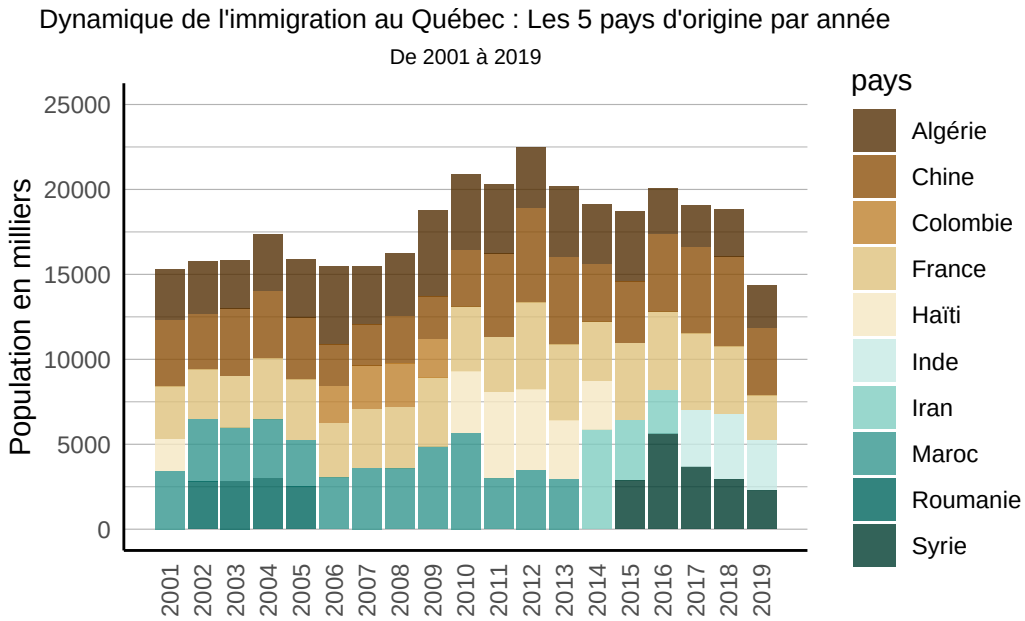
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1864    2932    3465    3581    4018    5853
```

```
glimpse(top_countries_by_year)
```

```
Rows: 95
Columns: 4
Groups: année [19]
$ pays                  <chr> "Chine", "Maroc", "France", "Algérie", "Haïti", ~
$ année                 <dbl> 2001, 2001, 2001, 2001, 2001, 2002, 2002, 2002, ~
$ nombre_immigrants     <dbl> 3924, 3427, 3112, 2991, 1864, 3686, 3245, 3093, ~
$ proportion_immigrants <dbl> 10.5, 9.1, 8.3, 8.0, 5.0, 9.8, 8.6, 8.2, 7.8, 7.~
```

Les résultats :

```
Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
i Please use the `linewidth` argument instead.
```




```
# saving the graph

save(graph_1, file = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_1.Rda")

ggsave(graph_1,
       filename = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_1.png",
       dpi = 320,
       bg = "white",
       units = "cm",
       height = 10,
       width = 20)
```

```
save(graph_2, file = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_2.Rda")

ggsave(graph_2,
       filename = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_2.png",
       dpi = 320,
```

```
        bg = "white",
        units = "cm",
        height = 10,
        width = 20)
```

Dataset : pm_quebec_clean_test

```
# données de webscraping Web provenant de Wikipédia

pm_quebec <- read_html("https://fr.wikipedia.org/wiki/Premier_ministre_du_Québec") |>
  html_elements("table") |>
  pluck(2) |>
  html_table(fill = T)

# Nettoyage des données: suppression des colonnes inutiles

pm_quebec <- pm_quebec[, -c(1, 2, 3, 5)]


# diviser les info de la variable "Nom" en colonnes, "Name", "DateOfBirth", "Party"
#used chat GPT

pm_quebec <- pm_quebec |>
  separate(Nom, into = c("Name", "DateOfBirth", "Party"),
           sep = "\\(|\\)", remove = TRUE, convert = TRUE, extra = "merge") |>  mutate(
    Name = str_trim(Name),
    DateOfBirth = str_extract(DateOfBirth, "\\d+\\s+\\w+\\s+\\d+"),
    Party = str_extract(Party, "[A-Za-zé]+")) |>
  select(-c(DateOfBirth, Événements))

# séparant la colonne législatures et mandats "

colnames(pm_quebec)
```

```
[1] "Name"                          "Party"
[3] "Législature(s) et mandat(s)" "Circonscription"
```

```
# en séparant les infos dans la colonne "législatures et mandats" par les deux: "start_yea
#used chat GPT
```

```r
pm_quebec_clean <- pm_quebec %>%
  mutate(
    years = str_match(`Législature(s) et mandat(s)`, "^(\\d{4})-?\\s*(\\d{4})?"),
    start_year = as.numeric(years[, 2]),
    end_year = as.numeric(years[, 3])
  ) %>%
  select(-c("years", "Législature(s) et mandat(s)"))

colnames(pm_quebec_clean)
```

```
[1] "Name"           "Party"          "Circonscription" "start_year"
[5] "end_year"
```

```r
# filtrer la colonne "start_year" pour l'année >= 2001

pm_quebec_clean <- pm_quebec_clean |> filter(start_year >= 2001) |>
  select("Name", "Party", "start_year", "end_year", "Circonscription")

# transformer l'observation NA de la variable "end_year" en 2019

pm_quebec_clean <- pm_quebec_clean |>
  mutate(end_year = if_else(is.na(end_year), 2019, end_year))

# Création d'une liste de séquences d'années de 2001 à 2019

pm_quebec_clean_test <- pm_quebec_clean |>
  rowwise() |>
  mutate(year = list(seq(start_year, end_year))) |>
  unnest(cols = c(year)) |>
  select(Name, Party, year, Circonscription)

# Renommer les catégories de la variable "Party"

pm_quebec_clean_test <- pm_quebec_clean_test |>
  mutate(Party = case_when(
    Party == "Parti" ~ "PQ",
    Party == "Libéral" ~ "PLQ",
    Party == "Coalition" ~ "CAQ",
    TRUE ~ as.character(Party)
  )) |> mutate(year = as.numeric(year))
```

Glimpse : pm_quebec_clean_test

```
glimpse(pm_quebec_clean_test)
```

```
Rows: 45
Columns: 4
$ Name            <chr> "Bernard Landry", "Bernard Landry", "Bernard Landry", ~
$ Party           <chr> "PQ", "PQ", "PQ", "PLQ", "PLQ", "PLQ", "PLQ", "PLQ", "~
$ year            <dbl> 2001, 2002, 2003, 2003, 2004, 2005, 2006, 2007, 2008, ~
$ Circonscription <chr> "Verchères, Montérégie", "Verchères, Montérégie", "Ver~
```

Dataset 3 : dat_imm_parti_final

```
## Unificateur : données des partis + données des immigrants


# calcul du total pour les 15 pays et rename de la colonne année

imm_quebec_total <- imm_quebec_2001_2019_clean |> select(nombre_immigrants, année) |>
  group_by(année) |> summarise(nombre_imm_total = sum(nombre_immigrants)) |>
  rename(year = année)

# essayons maintenant de combiner

dat_imm_parti <- left_join(imm_quebec_total, pm_quebec_clean_test, by = "year")

# since we doubled the rows somehow, for unique combinations we use distinct() and then ex


dat_imm_parti_demo <- dat_imm_parti |> distinct(year, nombre_imm_total, Name, Party, Circo

dat_imm_parti_final <- dat_imm_parti_demo[-c(3, 13, 16, 21), ]
```

Glimpse : dat_imm_parti_final

```
glimpse(dat_imm_parti_final)
```

```
Rows: 19
Columns: 5
$ year             <dbl> 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009,~
$ nombre_imm_total <dbl> 24764, 25321, 26370, 28874, 27808, 27662, 27279, 2831~
$ Name             <chr> "Bernard Landry", "Bernard Landry", "Jean Charest", "~
$ Party            <chr> "PQ", "PQ", "PLQ", "PLQ", "PLQ", "PLQ", "PLQ", "PLQ",~
$ Circonscription  <chr> "Verchères, Montérégie", "Verchères, Montérégie", "Sh~
```

```r
# graph_3

partie_couleurs <- c("PQ" = "#0033A0",
                     "CAQ" = "#87CEEB",
                     "PLQ" = "#FF0000")


graph_3 <- ggplot(data = dat_imm_parti_final, aes(x = year,
                                        y = nombre_imm_total,
                                        fill = Party)) +
  geom_bar(stat = "identity",
           alpha = 0.8) +
  scale_x_continuous(breaks = 2001:2019) +
  scale_y_continuous(limits = c(0, 40000),
                     breaks = seq(0, 40000, by = 10000)) +
  scale_fill_manual(values = partie_couleurs) +
  labs(title = "Tendances de l'immigration des 15 pays principaux au Québec et contexte po
       subtitle = "De 2001 à 2019",
       y = "Population en milliers",
       x = "") +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(colour = "grey70", size = .2),
        panel.grid.minor.y = element_line(colour = "grey70", size = .1),
        axis.line = element_line(colour = "black"),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        plot.title = element_text(size = 12,
                                  hjust = 0.5),
        plot.subtitle = element_text(size = 10,
                                  hjust = 0.5),
        text = element_text(face = "plain")
        )
```

```r
save(graph_3, file = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_3.Rda")

ggsave(graph_3,
       filename = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_3.png",
       dpi = 320,
       bg = "white",
       units = "cm",
       height = 10,
       width = 20)
```

Dataset : top5_countries_party_2001_2019_final

```r
# nettoyage des données pour fusionner les 5 principaux pays et partis politiques:

top5_2001_2019 <- top_countries_by_year |>  rename(year = année)

#merging

top5_countries_party_2001_2019 <- left_join(top5_2001_2019, pm_quebec_clean_test, by = "ye
```

```
Warning in left_join(top5_2001_2019, pm_quebec_clean_test, by = "year"): Detected an unexpec
i Row 11 of `x` matches multiple rows in `y`.
i Row 1 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```r
# On supprime les données en double formées lors de la combinaison de deux bases de donnée

colnames(top5_countries_party_2001_2019)
```

```
[1] "pays"                 "year"                 "nombre_immigrants"
[4] "proportion_immigrants" "Name"                 "Party"
[7] "Circonscription"
```

```r
top5_countries_party_2001_2019 <- top5_countries_party_2001_2019|>
  distinct(pays, year, nombre_immigrants, Party)

top5_countries_party_2001_2019_final <- top5_countries_party_2001_2019[-c(11, 13, 15, 17,
```

Glimpse : top5_countries_party_2001_2019_final

```
glimpse(top5_countries_party_2001_2019_final)
```

```
Rows: 95
Columns: 4
Groups: year [19]
$ pays              <chr> "Chine", "Maroc", "France", "Algérie", "Haïti", "Mar~
$ year              <dbl> 2001, 2001, 2001, 2001, 2001, 2002, 2002, 2002, 2002~
$ nombre_immigrants <dbl> 3924, 3427, 3112, 2991, 1864, 3686, 3245, 3093, 2926~
$ Party             <chr> "PQ", "PQ", "PQ", "PQ", "PQ", "PQ", "PQ", "PQ", "PQ"~
```

```
#graph 4

library(RColorBrewer)

graph_4 <- ggplot(data = top5_countries_party_2001_2019_final, aes(x = year,
                                                    y = nombre_immigrants,
                                                    fill = pays)) +
  geom_bar(stat = "identity",
           position = "stack",
           alpha = 0.8) +
  geom_text(
    aes(label = Party, group = year),
    stat = "summary",
    fun.y = max,
    vjust = -0.5,
    color = "black",
    size = 3,
    nudge_y = 18000) +
  scale_x_continuous(breaks = 2001:2019) +
  scale_y_continuous(limits = c(0, 25000),
                     breaks = seq(0, 25000, by = 5000)) +
  scale_fill_brewer(palette = "BrBG") +
  labs(title = "Dynamique de l'immigration au Québec : Les 5 pays d'origine par année",
       subtitle = "De 2001 à 2019",
       y = "Population en milliers",
       x = "") +
  theme_minimal() +
  theme(panel.grid.major.x = element_blank(),
```

```
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(colour = "grey70", size = .2),
        panel.grid.minor.y = element_line(colour = "grey70", size = .1),
        axis.line = element_line(colour = "black"),
        axis.text.x = element_text(angle = 90, vjust = 0.5),
        plot.title = element_text(size = 12,
                                  hjust = 0.5),
        plot.subtitle = element_text(size = 10,
                                  hjust = 0.5),
        text = element_text(face = "plain")
        )
```

```
Warning in geom_text(aes(label = Party, group = year), stat = "summary", :
Ignoring unknown parameters: `fun.y`
```

```
#

save(graph_4, file = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_4.Rda")

ggsave(graph_4,
       filename = "~/Dropbox/fas_1001_Zhuk/_tp/_tp4/fig/graph_4.png",
       dpi = 320,
       bg = "white",
       units = "cm",
       height = 10,
       width = 20)
```

```
No summary function supplied, defaulting to `mean_se()`
```