# PROJET EN SCIENCE DE DONNEES

# TITRRE : ETUDE DES TENDANCES DE L’IMMIGRATION DU CANADA ENTRE JANVIER 2015 et JUIN 2024

AUTEUR : MONTHE MBOUOWO CYRILLE

UNIVERSITE TELUQ, Trimestre Hiver 2025

PROFFESSEUR RESPONSABLE : Richard Hotte

ENCADREUR RESPONSIBLE : Fatima Bensalma

***Lien github :*** *https://github.com/moncyr/Tendances\_Immigration\_Canada\_2015-2024*

Table des matières

[PROJET EN SCIENCE DE DONNEES 1](#_Toc199633206)

[TITRRE : ETUDE DES TENDANCES DE L’IMMIGRATION DU CANADA ENTRE JANVIER 2015 et JUIN 2024 1](#_Toc199633207)

[1. La problématique. 3](#_Toc199633208)

[2. Les enjeux 3](#_Toc199633209)

[3. Le contexte 3](#_Toc199633210)

[4. Objectifs du projet et hypothèses. 3](#_Toc199633211)

[5. La Collecte des données 4](#_Toc199633212)

[6. Description, exploration et présentation des données 4](#_Toc199633213)

[7. Nettoyage, préparation et reformatage des données. 7](#_Toc199633214)

[8. Filtrage des données 8](#_Toc199633215)

[1. Filtrage par province sur les données (filtrage sur les catégories) 8](#_Toc199633216)

[2. *Filtrage sur les variables* 9](#_Toc199633217)

[9. Visualisation et recherche de valeur extrêmes 12](#_Toc199633218)

[1. Visualisation des données totale de l’immigration du Canada par trimestres 12](#_Toc199633219)

[2. Construction graphique par années 13](#_Toc199633220)

[10. Analyse statistique et vérification des hypothèses d’étude. 14](#_Toc199633221)

[- Les données de l’immigration sur le Canada tendent à suivre une régression linéaire, qui montre que les chiffres sont en augmentation selon une tendance linéaire. 14](#_Toc199633222)

[- Les trimestres sont constants. 14](#_Toc199633223)

[1. Transformation des données 15](#_Toc199633224)

[2. Histogrammes montrant les répartitions par années et par trimestre. 16](#_Toc199633225)

[a. Repartion des chiffres par trimestres 16](#_Toc199633226)

[b. Repartion des moyennes par trimestre 17](#_Toc199633227)

[c. Répartition des chiffres par années 18](#_Toc199633228)

[3. Le test ANOVA (analyse de la variance) 19](#_Toc199633229)

[*Hypothèse de référence H0*: tous les trimestres ont la même moyenne. 19](#_Toc199633230)

[*Hypothèse alternative Ha*: au moins un des trimestres diffère des autres. 19](#_Toc199633231)

[*Degré de signification* : 5% 19](#_Toc199633232)

[Test statistique : ANOVA 19](#_Toc199633233)

[4. Vérification des suppositions pour le test ANOVA 19](#_Toc199633234)

[a. Independence des observations : 19](#_Toc199633235)

[b. L’homogénéité de la variance : 20](#_Toc199633236)

[c. Vérification de la normalité des résidus*:* 20](#_Toc199633237)

[5. Analyse et interpretation des résultats ANOVA 21](#_Toc199633238)

[6. Test de régression linéaire 21](#_Toc199633239)

[a. Transformons les données pour trouver la régression 21](#_Toc199633240)

[7. Intervalles de confiance 21](#_Toc199633241)

[8. Vérification des suppositions pour le test de régression linéaire. 23](#_Toc199633242)

[a. La normalité des résidus 23](#_Toc199633243)

[b. L’homogénéité de la variance 24](#_Toc199633244)

[9. Recherche de la régression linéaire en prenant pour variable explicatives l’année et la démographie du pays 25](#_Toc199633245)

[a. Collecte et traitement des données de démographie du pays entre 2015 et 2024 25](#_Toc199633246)

[b. Données sur l’employabilité au Canada entre 2015 et 2024 26](#_Toc199633247)

[c. Combinaison des données d’immigration, de démographie et d’employabilité. 27](#_Toc199633248)

[d. La régression multiple avec comme variables explicatives Années et démographie. 28](#_Toc199633249)

[e. Interprétation et analyse du coefficient de détermination (R2) 29](#_Toc199633250)

[11. Communication des résultats 29](#_Toc199633251)

[1. L’égalité des chiffres pour tous les trimestres. 29](#_Toc199633252)

[2. La régression linéaire pour les années entre 2015 et 2024 31](#_Toc199633253)

[12. Recommandations et améliorations. 32](#_Toc199633254)

[1. Les limites 32](#_Toc199633255)

[2. Les données aberrantes et valeurs extrêmes. 32](#_Toc199633256)

# La problématique.

La tendance de l’immigration au Canda selon les provinces entre les années 2015 et 2024.

# Les enjeux

La Canada fait face à une demande croissante dans ses besoins en immigration, tant de l’extérieure que de l’intérieur, ceci dans le but d’assoir une croissance économique importante. Il est clair que l’immigration est devenue un atout incontournable pour le pays. Les besoins sont variés et nombreux (secteur de la sante, de l’éducation etc..). Le manque de main d’œuvre touche tous les secteurs : l’industrie, la sante, l’éducation sont les principaux secteurs touchés par la pénurie de personnel.

# Le contexte

Bien que l’immigration soit un des atouts majeurs pour assurer sa croissance économique, le pays doit faire face à une croissance démographique du a l’apport de la mouvance démographique.  Dans le contexte actuel, le pays se trouve dans une crise jamais enregistrée comme au par avant. Il s’agit de la crise du logement. S’intéresser aux tendances démographiques est très essentiel pour le pays. Cela permettra aux régions d’ajuster leurs chiffres en matière d’immigration en vue de résorber la crise du logement. Cela passe aussi par une prévision d’investissement dans le secteur de la construction conséquemment aux besoins occasionnés par les flux migratoires. Connaitre les chiffres en immigration permet aussi aux provinces d’ajuster leurs politiques dans le secteur, de planifier leur développement en matière des besoins sociaux comme l’éducation, la sante et l’employabilité.

# Objectifs du projet et hypothèses.

Au regard des données sur l’immigration du Canada entre 2015 et 2024, nous nous proposons d’étudier les tendances sur l’immigration. Des hypothèses sur les tendances seront formulées et tout au long de notre travail il sera question de les vérifier en utilisant les procéder et les étapes clés de l’analyse des données.

1. Doit-on considérer que les chiffres de l’immigration sont en hausse dans cette période ?
2. Les données de l’immigrations sont-elles les même pour tous les trimestres ?
3. Existe-il une modèle de régression linéaire dans les données l’immigration ?

# La Collecte des données

Il s’agit des données sur l’immigration du canada, pris sur [www.kaggle.com](http://www.kaggle.com) . Les données seront importées sur la plateforme RStudio qui est notre plateforme de travail. Le lien vers les données est <https://www.kaggle.com/datasets/zongaobian/canada-permanent-residents> , le nom du fichier est La « canada\_immigrant\_2024\_q2.csv(217.51 kB) »

> # Commande pour importer les donnees

> immigration<-read.csv("canada\_immigrant\_2024\_q2(1).csv", sep=",", header = TRUE)

# Description, exploration et présentation des données

> str(immigration)

> # aperçu général des données. On peut clairement y voir les différentes variables, leur type et leurs valeurs. Les types des sont essentiellement des chaines de caractères (chr).

> # Nous avons 163 variables (colonnes) et 371 observations (lignes)

> dim(immigration) [1] 371 163

> # Nom de la première variable :

> colnames(immigration)[1]

[1] "Canada...Admissions.of.Permanent.Residents.by.Province.Territory.of.Intended.Destination.and.Immigration.Category..January.2015...June.2024.Province.Territory.and.Immigration.Category"

> # La première variable décrit les différentes catégories d’immigration sur chaque observation :

> # La commande suivante nous donne les différentes valeurs de cette variable :

> unique(immigration [,1])

[1] "Agri-Food Pilot"

[2] "Atlantic Immigration Pilot Programs"

[3] "Atlantic Immigration Programs"

[4] "Canadian Experience"

[5] "Caregiver"

[6] "Skilled Trade"

[7] "Skilled Worker"

[8] "Worker Program"

[9] "Self-Employed"

[10] "Start-up Business"

[11] "Business"

[12] "Provincial Nominee Program"

[13] "Temporary Resident to Permanent Resident Pathway"

[14] "Economic - Total"

[15] "Sponsored Children"

[16] "Sponsored Extended Family Member"

[17] "Sponsored Parent or Grandparent"

[18] "Sponsored Spouse or Partner"

[19] "Sponsored Family"

[20] "Sponsored Family - Total"

[21] "Blended Sponsorship Refugee"

[22] "Government-Assisted Refugee"

[23] "Privately Sponsored Refugee"

[24] "Resettled Refugee"

[25] "Protected Person in Canada"

[26] "Resettled Refugee & Protected Person in Canada - Total"

[27] "All Other Immigration - Total"

[28] "Newfoundland and Labrador - Total"

[29] "Prince Edward Island - Total"

[30] "Federal Economic Mobility Pathways Pilot"

[31] "Nova Scotia - Total"

[32] "New Brunswick - Total"

[33] "Entrepreneur"

[34] "Investor"

[35] "Quebec - Total"

[36] "Rural and Northern Immigration"

[37] "Ontario - Total"

[38] "Manitoba - Total"

[39] "Saskatchewan - Total"

[40] "Alberta - Total"

[41] "British Columbia - Total"

[42] "Yukon - Total"

[43] "Northwest Territories - Total"

[44] "Nunavut - Total"

[45] "Province/Territory not stated - Total"

[46] "Total"

Il s’agit de toutes les catégories qui se retrouves dans notre table de données. Pour notre étude nous nous intéresserons uniquement aux observations sur les provinces et la dernière ligne qui est le total pour tout le pays.

> # La commande suivante nous montre les autres variables :

> head(colnames(immigration)[-1], 20)

[1] "X2015.Q1.Jan" "X2015.Q1.Feb" "X2015.Q1.Mar" "X2015.Q1.Total"

[5] "X2015.Q2.Apr" "X2015.Q2.May" "X2015.Q2.Jun" "X2015.Q2.Total"

[9] "X2015.Q3.Jul" "X2015.Q3.Aug" "X2015.Q3.Sep" "X2015.Q3.Total"

[13] "X2015.Q4.Oct" "X2015.Q4.Nov" "X2015.Q4.Dec" "X2015.Q4.Total"

[17] "X2015.Total" "X2016.Q1.Jan" "X2016.Q1.Feb" "X2016.Q1.Mar"

Comme variables nous avons trois grandes formes :

* Les mois de chaque année (Jan, Feb, Mar …) et selon le trimestre (Q1, Q2, Q3, Q4) : Ex : "X2023.Q1.Jan"
* Le total par trimestre : Ex : "X2022.Q3.Total"
* Le total de chaque année : Ex : "X2021.Total"

Exploration des valeurs des variable pour vérifier la forme : je choisis une colonne au hasard et j’explore ses 20 premières et les 20 dernières valeurs :

> head(immigration[,2], 20)

[1] "0" "0" "0" "--" "0" "0" "--" "5" "0"

[10] "0" "0" "25" "25" "0" "0" "35" "--" "0"

> tail(immigration[,2], 20)

[1] "0" "0" "--" "0" "0" "--" "0" "0" "0"

[10] "0" "0" "0" "0" "0" "0" "0" "0" "--"

[19] "5" "12,910"

On peut clairement constater l’existence de valeur insignifiantes non définis : "--", il faudra donc changer ses valeurs pendant la phase de nettoyage pour leur donner une valeur significative comme 0 par exemple. Le caractère ‘ , ’ sera aussi à enlever pour faciliter la conversion en type numérique. *Ex : "12,910" sera changé en “12910"*

# Nettoyage, préparation et reformatage des données.

> # Nous commençons par changer le nom de la première variable qui semble trop long :

> colnames(immigration)[1]<-"Categories d'Immigration"

> # Nous remplaçons les valeurs non définies par 0 :

> immigration[immigration=='--']=0

> head(immigration[,2],20)

[1] "0" "0" "0" "0" "0" "0" "0" "5" "0" "0" "0" "25" "25" "0" "0" "35"

[17] "0" "0" "0" "15"

Remplaçons le caractère ‘ , ’ par un caractère vide ‘ ‘, pour faciliter la conversion et la lecture en décimal.

# remplacer la virgule ',' dans les valeurs numériques

> immigration<-data.frame(lapply(immigration, function(x) gsub(',','',x)))

> tail(immigration[,2], 20)

[1] "0" "0" "0" "0" "0" "0" "0" "0" "0" "0"

[11] "0" "0" "0" "0" "0" "0" "0" "0" "5" "12910"

> # Conversion des valeurs en type décimaux

> immigration<-data.frame(immigration[,1],lapply(immigration[,c(2:length(immigration))], as.numeric))

> str(immigration)

'data.frame': 371 obs. of 163 variables:

$ immigration...1.: chr "Agri-Food Pilot" "Atlantic Immigration Pilot Programs" "Atlantic Immigration Programs" "Canadian Experience" ...

$ X2015.Q1.Jan : num 0 0 0 0 0 0 0 5 0 0 ...

$ X2015.Q1.Feb : num 0 0 0 0 0 0 15 15 0 0 ...

$ X2015.Q1.Mar : num 0 0 0 0 0 0 10 10 0 0 ...

# Filtrage des données

## Filtrage par province sur les données (filtrage sur les catégories)

# fonction qui ressort les données d'une province.

# la fonction de nom 'province' prend en paramètre le nom de la province ou une partie du nom.

province<-function(prov){

index<-grep(prov,immigration[,1],ignore.case = TRUE)

return(immigration[index,])

}

**Exemple :**

Données totales sur la province du Nunavut :

> # données de l'immigration du nunavut

> immigration\_Nunavut<-province('nunavut')

> str(immigration\_Nunavut)

'data.frame': 1 obs. of 163 variables:

$ immigration...1.: chr "Nunavut - Total"

$ X2015.Q1.Jan : num 0

$ X2015.Q1.Feb : num 5

$ X2015.Q1.Mar : num 5

$ X2015.Q1.Total : num 10

$ X2015.Q2.Apr : num 0

$ X2015.Q2.May : num 0

Données totales pour la province de l’Ontario

> # Donnees de l'immigration de l'Ontario

> immigration\_Ontario<-province('ontario')

> str(immigration\_Ontario)

'data.frame': 1 obs. of 163 variables:

$ immigration...1.: chr "Ontario - Total"

$ X2015.Q1.Jan : num 5320

$ X2015.Q1.Feb : num 6590

$ X2015.Q1.Mar : num 7860

$ X2015.Q1.Total : num 19770

$ X2015.Q2.Apr : num 7610

$ X2015.Q2.May : num 9510

On détermine les données totales pour toute l’étendue du Canada

> #immigration de tout le pays, le total

> immigration\_Canada<-immigration[nrow(immigration),]

> str(immigration\_Canada)

'data.frame': 1 obs. of 163 variables:

$ immigration...1.: chr "Total"

$ X2015.Q1.Jan : num 12910

$ X2015.Q1.Feb : num 16440

$ X2015.Q1.Mar : num 21770

$ X2015.Q1.Total : num 51125

$ X2015.Q2.Apr : num 21165

$ X2015.Q2.May : num 23895

$ X2015.Q2.Jun : num 27025

$ X2015.Q2.Total : num 72090

$ X2015.Q3.Jul : num 27770

## ***Filtrage sur les variables***

On sélectionne les données totales pour chaque trimestre (Q1, Q2, Q3, Q4) selon l’année

# fonction pour avoir les données par trimestres qui prend deux paramètres : le trimestre et le dataframe.

# Q est le trimestre (Q1, Q2, Q3, Q4) et data est le dataframe qu'on filtre

trimestre<-function(Q,data){

index<-grep(Q,colnames(data),ignore.case = TRUE)

return(data[,index])

}

On calcule les index correspondants aux colonnes qui affichent les totaux :

# index pour conserver uniquement les colonnes des totaux dans le dataframe

index\_total<-grep('total', colnames(immigration), ignore.case = TRUE)

On va conserver uniquement les colonnes qui affichent les totaux pour les trimestres et les totaux pour chaque année en utilisant la variable objet index\_total calculée précédemment.

> # immigartion Canada total, on a conservé uniquement les totaux

> immigration\_Canada\_total<-immigration\_Canada[,index\_total]

> names(immigration\_Canada\_total)

[1] "X2015.Q1.Total" "X2015.Q2.Total" "X2015.Q3.Total" "X2015.Q4.Total"

[5] "X2015.Total" "X2016.Q1.Total" "X2016.Q2.Total" "X2016.Q3.Total"

[9] "X2016.Q4.Total" "X2016.Total" "X2017.Q1.Total" "X2017.Q2.Total"

**Exemple :**

Données du premier trimestres Q1 pour l’immigration totale du Canada :

> canada\_Q1<-trimestre('Q1',immigration\_Canada\_total)

> str(canada\_Q1)

'data.frame': 1 obs. of 10 variables:

$ X2015.Q1.Total: num 51125

$ X2016.Q1.Total: num 86235

$ X2017.Q1.Total: num 72835

$ X2018.Q1.Total: num 80005

$ X2019.Q1.Total: num 65955

$ X2020.Q1.Total: num 69140

$ X2021.Q1.Total: num 70495

$ X2022.Q1.Total: num 113805

$ X2023.Q1.Total: num 145495

$ X2024.Q1.Total: num 121730

On fait de même pour avoir les données autres trimestres Q2, Q3 et Q4

#Q2

canada\_Q2<-trimestre('Q2',immigration\_Canada\_total)

#Q3

canada\_Q3<-trimestre('Q3',immigration\_Canada\_total)

#Q4

canada\_Q4<-trimestre('Q4',immigration\_Canada\_total)

Si on désire observer seulement l’immigration pour tous les trimestres

> # conserver les index des colonnes trimestre (Q1, Q2, Q3, Q4) dans immigration\_Canada\_total

>index\_trimestre<-grep('Q',names(immigration\_Canada\_total), ignore.case = TRUE)

> #immigration totale par trimestre

> immigration\_trimeestre<-immigration\_Canada\_total[,index\_trimestre]

> names(immigration\_trimeestre)

[1] "X2015.Q1.Total" "X2015.Q2.Total" "X2015.Q3.Total" "X2015.Q4.Total"

[5] "X2016.Q1.Total" "X2016.Q2.Total" "X2016.Q3.Total" "X2016.Q4.Total"

[9] "X2017.Q1.Total" "X2017.Q2.Total" "X2017.Q3.Total" "X2017.Q4.Total"

[13] "X2018.Q1.Total" "X2018.Q2.Total" "X2018.Q3.Total" "X2018.Q4.Total"

On désire maintenant observer seulement l’immigration totale pour chaque année :

> # immigration pour chaque annees

> immigration\_annees<-immigration\_Canada\_total[,-index\_trimestre]

> str(immigration\_annees)

'data.frame': 1 obs. of 10 variables:

$ X2015.Total: num 271840

$ X2016.Total: num 296375

$ X2017.Total: num 286535

$ X2018.Total: num 321055

$ X2019.Total: num 341175

$ X2020.Total: num 184600

$ X2021.Total: num 406050

$ X2022.Total: num 437625

$ X2023.Total: num 471815

$ X2024.Total: num 255705

# Visualisation et recherche de valeur extrêmes

## Visualisation des données totale de l’immigration du Canada par trimestres

# les graphiques

# graphiques des données par trimestre

# graphique avec 'xaxt='n': desactiver l'axe des x pour le personnaliser,

# las=2, ecriture sur l'axe des x en verticale

plot(0,0, main="Nombre d'immigrant par trimestre",

xlab=" ", ylab="Nombre d'immigrants",

xlim=c(1,ncol(immigration\_trimestre)),

ylim=c(min(immigration\_trimestre[1,]),

max(immigration\_trimestre[1,])), type='n', xaxt='n' )

#type='n' pour ne rien afficher, type='b' ligne et point, type='h' lignse horizontales

axis(side=1, at=c(1:ncol(immigration\_trimestre)),labels =names(immigration\_trimestre), las=2 )

points(x=1:ncol(immigration\_trimestre),

y=immigration\_trimestre[1,], type = 'b', lwd=2 )

Une image contenant texte, Police, ligne, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 1 : Graphique de l'immigration par trimestre

## Construction graphique par années

# graphique des valeurs totales selon les annees

plot(0,0, type='n', main="Nombre d'immigrant par an",

xlab='', ylab="Nombre d'immigrant",xlim=c(1,ncol(immigration\_annees)),

ylim=c(min(immigration\_annees[1,]), max(immigration\_annees[1,])), xaxt='n')

axis(side=1, at=c(1:ncol(immigration\_annees)), labels=2015:2024, las=2)

points(x=1:ncol(immigration\_annees), y=immigration\_annees[1,], type='b',

lwd=2)

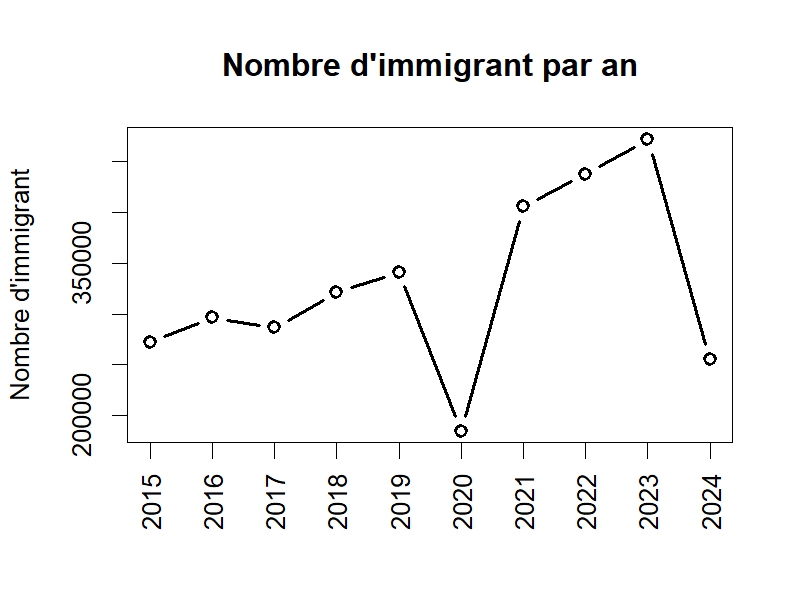


Figure 2 : Graphiques de l'immigration par années

**Interprétation** : Nous remarquons immédiatement les valeurs extrêmes dans les années 2020 et 2024, qui s’écartent de la distribution. Ces valeurs pourraient ne pas être considérées dans la recherche de la régression.

# Analyse statistique et vérification des hypothèses d’étude.

Selon les observations des graphiques que nous avons obtenus, nous pouvons émettre les hypothèses d’analyse suivantes. Ces hypothèses semblent encore plus prononcées lorsqu’on retire les valeurs aberrantes :

### Les données de l’immigration sur le Canada tendent à suivre une régression linéaire, qui montre que les chiffres sont en augmentation selon une tendance linéaire.

### Les trimestres sont constants.

Pour la vérification de nos hypothèses, il convient d’effectuer des tests statistiques. Le premier sera un test ANOVA et le second un test de régression linéaire.

Mais avant nous allons effectuer une mise en forme de nos données afin d’y effectuer nos tests statistiques. Nous allons constituer un dataframe avec les valeurs de l’immigration pour différentes catégories de trimestre. Les trimestres (Q1, Q2, Q3, Q4) constituent ici les groupes indépendants d’observation.

## Transformation des données

# reconstitution d'une table de données pour le test ANOVA

# data.frame avec les groupes par trimestre pour l'analyse l'Anova

# variable categorique 'Trimestre'

Trimestre<-c(rep('Q1',length(canada\_Q1)),rep('Q2',length(canada\_Q2)),

rep('Q3',length(canada\_Q3)),rep('Q4',length(canada\_Q4)))

# variables nombre d'immigrants

Nombre<-unlist(c(canada\_Q1[1,],canada\_Q2[1,],canada\_Q3[1,],canada\_Q4[1,]), use.names = FALSE))

> str(Nombre)

# data frame canada\_Trimestre qui regroupe les valeurs trimestre

canada\_trimestre<-data.frame(Nombre,Trimestre)

canada\_trimestre

# transformation en variable categorique

>canada\_trimestre$Trimestre<-as.factor(canada\_trimestre$Trimestre)

>str(canada\_trimestre)

> Nombre Trimestre

1 51125 Q1

2 86235 Q1

3 72835 Q1

4 80005 Q1

5 65955 Q1

6 69140 Q1

7 70495 Q1

8 113805 Q1

9 145495 Q1

10 121730 Q1

11 72090 Q2

12 88295 Q2

13 78015 Q2

14 87675 Q2

15 94280 Q2

16 34075 Q2

17 74440 Q2

18 118320 Q2

19 117925 Q2

20 133970 Q2

21 78890 Q3

22 68230 Q3

23 70135 Q3

24 82290 Q3

25 103715 Q3

26 40120 Q3

27 122915 Q3

28 122205 Q3

29 107990 Q3

30 69735 Q4

31 53610 Q4

32 65550 Q4

33 71080 Q4

34 77230 Q4

35 41265 Q4

36 138200 Q4

37 83295 Q4

38 00405 Q4

## Histogrammes montrant les répartitions par années et par trimestre.

### Repartion des chiffres par trimestres

# somme totale par trimestre

sum\_trimestre<-tapply(canada\_trimestre$Nombre,

INDEX = canada\_trimestre$Trimestre, FUN =sum)

# changement d'échelle pour la représentation sur l'histogramme (1/1000)

sum\_trimestre<-sum\_trimestre/1000

barplot(sum\_trimestre,xlab='Trimestres' , ylab = 'Nombres en millier (x1000)',

main='Repartition par trimestre', col = 'red', xpd = TRUE,

ylim=c(0, max(sum\_trimestre) +200))

# Nous obtenons les agrégations suivantes sur la répartition par trimestre

> sum\_trimestre

Q1 Q2 Q3 Q4

876.820 899.085 796.490 700.370

Une image contenant texte, capture d’écran, Police, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 3 : Histogramme montrant la répartition par trimestre en millier.

### Repartion des moyennes par trimestre

L’histogramme des valeurs moyennes selon les trimestre nous donne idee sur l’hypothese de l’egalite des groupes :

# repartition des voleurs moyennes selon les trimestres

barplot(moy\_Q, main='Repartion des moyennes par trimestre',

xlab='Trimestres', ylab='Valeurs moyennes',

ylim=c(0,max(moy\_Q)), col='blue')

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Figure 4 : Répartition des moyennes des trimestres

### Répartition des chiffres par années

# Histogramme de la répartition selon années

sum\_annee<-unlist(immigration\_annees[1,],use.names = FALSE)

# Changement d'echelle (1/1000)

sum\_annee<-sum\_annee/1000

#on nomme les valeurs

names(sum\_annee)<-2015:2024

sum\_annee

barplot(sum\_annee, main="Repartition par annees",

col='pink',xlab='Annees',

ylab='Nombre en millier (x1000)',

ylim=c(0, max(sum\_annee)+100))

# Total par annee

> sum\_annee

2015 2016 2017 2018 2019 2020 2021 2022

271.840 296.375 286.535 321.055 341.175 184.600 406.050 437.625

2023 2024

471.815 255.705

Une image contenant texte, capture d’écran, Police, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 5 : Répartition par années en millier

## Le test ANOVA (analyse de la variance)

### *Hypothèse de référence H0*: tous les trimestres ont la même moyenne.

### *Hypothèse alternative Ha*: au moins un des trimestres diffère des autres.

### *Degré de signification* : 5%

### Test statistique : ANOVA

# test ANOVA

aov<-aov(Nombre~Trimestre,data=canada\_trimestre)

summary(aov)

Df Sum Sq Mean Sq F value Pr(>F)

Trimestre 3 8.390e+08 279677560 0.342 0.795

Residuals 34 2.781e+10 817952228

## Vérification des suppositions pour le test ANOVA

### Independence des observations :

Les observations sont indépendantes. Il s’agit des données pour chaque trimestre.

### L’homogénéité de la variance :

Le graphe des résidus en fonction des valeurs prédites montre une distribution assez homogène des points.

# graphe des résidus en fonction des valeurs prédites

# vérification de l’égalité de la variance

plot(residuals(aov), fitted(aov), main='Résidus en fonction des valeurs predites',

xlab = 'Résidus', ylab = 'Valeurs prédites')

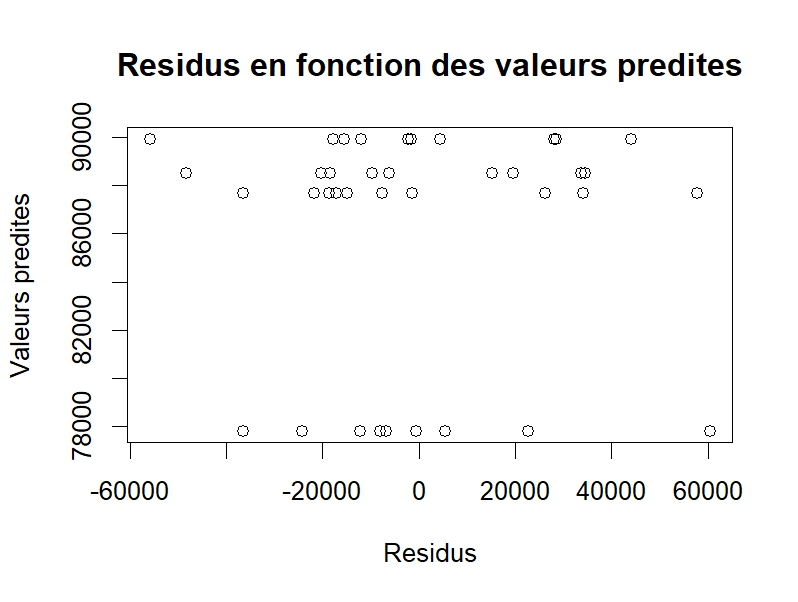


Figure 5: Graphe des Residus en fonction des valeurs predites

### Vérification de la normalité des résidus*:*

Le graphe confirme que les résidus suivent bien une distribution normale

# vérification de la normalité des résidus

# graphe des résidus

qqnorm(residuals(aov), main="Quantiles-Quantiles",

xlab='Quantiles Théoriques', ylab='Quantiles observe')

qqline(residuals(aov))

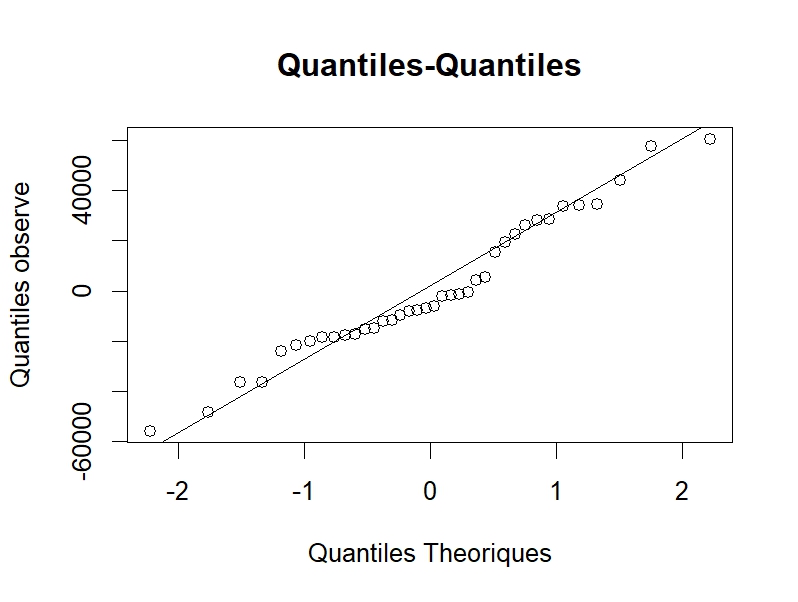


Figure 4 : Graphe quartiles-quartiles

## Analyse et interpretation des résultats ANOVA

Les groupes sont identiques avec une probabilité de 0.795. On ne peut pas rejeter l’hypothèse de référence. Il y a égalité des moyennes entre les trimestres.

## Test de régression linéaire

1. Transformons les données pour trouver la régression :

# transformation des donnees

Annees<-c(2015:2024)#la colonne Annees

canada\_annees<-data.frame(Annees, Nombre=unlist(immigration\_annees[1,],use.names = FALSE) )

> str(canada\_annees)

'data.frame': 8 obs. of 2 variables:

$ Annees: int 2015 2016 2017 2018 2019 2021 2022 2023

$ Nombre: num 271840 296375 286535 321055 341175 ...

## Intervalles de confiance

# Calcule de l'erreur type : SSE

# SEE\_Q1

Q1<-canada\_trimestre[canada\_trimestre$Trimestre=='Q1',]

SSE\_Q1<-sum((Q1$Nombre-mean(Q1$Nombre))^2)

# SEE\_Q2

Q2<-canada\_trimestre[canada\_trimestre$Trimestre=='Q2',]

SSE\_Q2<-sum((Q2$Nombre-mean(Q2$Nombre))^2)

# SEE\_Q3

Q3<-canada\_trimestre[canada\_trimestre$Trimestre=='Q3',]

SSE\_Q3<-sum((Q3$Nombre-mean(Q3$Nombre))^2)

# SEE\_Q4

Q4<-canada\_trimestre[canada\_trimestre$Trimestre=='Q4',]

SSE\_Q4<-sum((Q4$Nombre-mean(Q4$Nombre))^2)

# SSE pour l'ensemble des trimestres

SSE<- SSE\_Q1+SSE\_Q2+SSE\_Q3+SSE\_Q4

# degre de liberte : df=k(n-1)

# k: le nombre de groupe=4

# n le nombre de le nomdre d'observation par groupe

df<-34

# l'erreur type SE:

SE<-sqrt(SSE/df)

SE

# repartition des voleurs moyennes selon les trimestres

barplot(moy\_Q, main='Repartion des moyennes par trimestre',

xlab='Trimestres', ylab='Valeurs moyennes',

ylim=c(0,max(moy\_Q)), col='blue')

# construction de l'intervalle de confiance

#limites inferieures de l'intervalle de confiance

lim.inf<-moy\_Q - SE

#limites superieurs de l'intervalle de confiance

lim.sup<-moy\_Q + SE

# graphique des intervalles

plot(x=0, y=0, type ='n',ylim=c(min(lim.inf), max(lim.sup)),

main='Intervalles de confiances', xlab='Trimestres',

xlim=c(0,5), ylab='Valeurs moyennes', xaxt='n' )

#definition de l'axe des abscisses

axis(side=1, at=1:4, labels = names(moy\_Q))

# les moyennes

points(x=1:4, y=moy\_Q, col='blue')

# intervalle

arrows(x0 =1:4,y0 = lim.inf, x1 = 1:4, y1 = lim.sup,

col='red', code=3, angle=90 )

Une image contenant texte, diagramme, ligne, Rectangle

Le contenu généré par l’IA peut être incorrect.

Figure 5: Intervalles de confiance

## Vérification des suppositions pour le test de régression linéaire.

### La normalité des résidus

# vérification des suppositions

# vérification de la normalité des résidus

qqnorm(résiduals(Régression), main='Quartiles-Quartiles',

xlab = 'Quartile Théorique',

ylab = 'Quartile Observe')

qqline(residuals((Régression)))

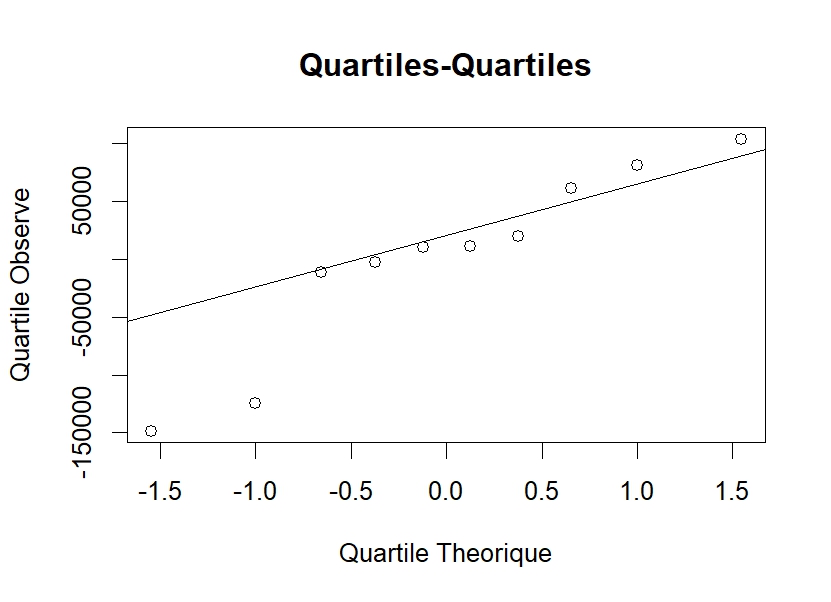


Figure 6: Quartiles observes en fonction des Quartiles theoriques

### L’homogénéité de la variance

#verification de l'homogénéité de la variance

plot(fitted(Regression),residuals(Regression), main='Residus en fonction des valeurs predites',

xlab='Residus',

ylab='Valeurs Predites')

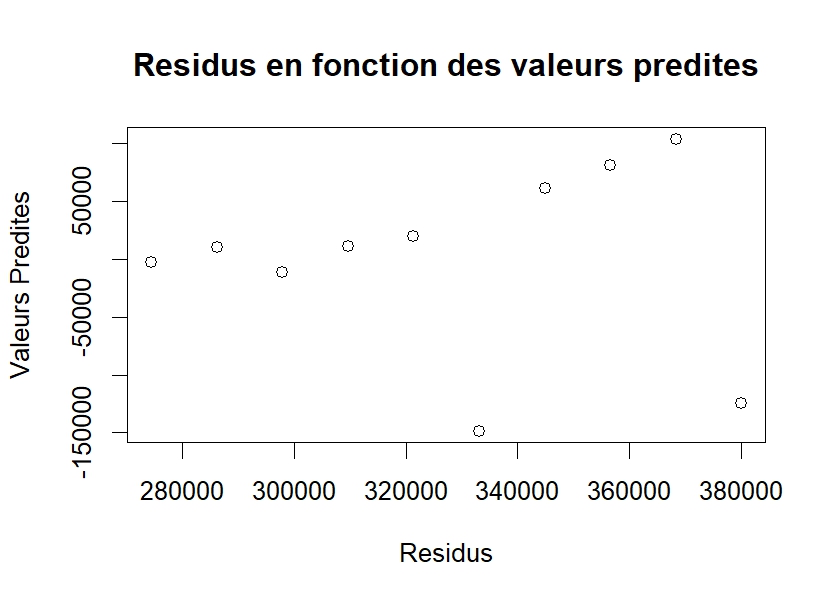


Figure7: Graphe des residus en fonction des valeurs predites

## Recherche de la régression linéaire en prenant pour variable explicatives l’année et la démographie du pays

### Collecte et traitement des données de démographie du pays entre 2015 et 2024

Données démographiques sur le pays entre 2015 et 2024 téléchargées sur le site de statique Canada : <https://www150.statcan.gc.ca/t1/tbl1/fr/tv.action?pid=1710000901&cubeTimeFrame.startMonth=01&cubeTimeFrame.startYear=2015&cubeTimeFrame.endMonth=10&cubeTimeFrame.endYear=2024&referencePeriods=20150101%2C20241001>

# Traitement des données de demographie entre 2015 et 2024

demographie<-read.csv2('Demographie\_Canada.csv', sep=';', header = TRUE)

# remplacer les espaces ' ' dans les valeurs numeriques

demographie<-data.frame(lapply(demographie, function(x) gsub(' ','',x)))

# convertir les valeurs en type numerique

demographie<-data.frame(demographie[,1],lapply(demographie[,c(2:length(demographie))], as.numeric))

# creer un dataframe qui regroupe les donnees par annees

somme\_annee<-function(data){

annees<-2014:2024

# Initialiser un data.frame vide avec les mêmes lignes que data

demographie\_annees <- data.frame(matrix(nrow = nrow(data), ncol = 0))

for ( annee in annees) {

# on determine les trimestres pour chaque annees

trim<-grep(annee,names(demographie),ignore.case = TRUE)

# on additionne pour avoir la somme pour l'annee

if (length(trim) == 4) {

# Calculer la somme des 4 trimestres

demographie\_annees[[as.character(annee)]] <- rowSums(data[,

trim], na.rm = TRUE)

}

# la somme de toutes les provinces

demographie\_somme<-colSums(demographie\_annees)

}

return(demographie\_somme)

}

demographie\_somme<-somme\_annee(demographie)

> head(demographie\_somme)

2015 2016 2017 2018 2019 2020

285411732 288420230 291954718 296074710 300441712 303982386

### Données sur l’employabilité au Canada entre 2015 et 2024

Source : <https://www.kaggle.com/datasets/pienik/unemployment-in-canada-by-province-1976-present>. Les données sont reparties sur tous les mois allant de 1976 à 2023. Nous allons donc uniformiser toutes nos données précédentes jusqu’à 2023.

# données de la population active

population\_active<-read.csv('Unemployment\_Canada\_1976\_present.csv',sep=',',header = TRUE)

# conversion des types

population\_active$Employment<-as.numeric(population\_active$Employment)

# création de la colonne years pour faciliter le filtrage

years <- substr(as.character(population\_active$REF\_DATE), 1, 4)

# ajout de la colonne Annees dans le dataframe population\_active

population\_active$Annes<-years

# coversion en type facteur

population\_active$Annes<-as.factor(population\_active$Annes)

# population active moyenne par annees

employment<-tapply(population\_active$Employment,INDEX=population\_active$Annes,FUN =mean)

# population active de 2015 a 2024

employment\_canada<-employment[names(employment)>='2015' & names(employment)<='2024']

employment\_canada

2015 2016 2017 2018 2019 2020 2021 2022

2061432 2075042 2118477 2156642 2199248 2081160 2182816 2265091

2023

2301906

### Combinaison des données d’immigration, de démographie et d’employabilité.

canada\_demo\_annees<-data.frame(canada\_annees, demographie=demographie\_annees )

> head(canada\_demo\_annees)

Annees Nombre demographie

2015 2015 271840 285411732

2016 2016 296375 288420230

2017 2017 286535 291954718

2018 2018 321055 296074710

2019 2019 341175 300441712

2020 2020 184600 303982386

# on coupe les données jusqu'a 2023 pour uniformiser le dataframe

canada\_demo\_annees\_2023<-canada\_demo\_annees[1: nrow(canada\_demo\_annees)-1,]

# on ajoute les donnees employment au dataframe canada\_demo\_annees

canada\_demo\_annees\_emp <-data.frame(canada\_demo\_annees\_2023,employabilite=employment\_canada)

>canada\_demo\_annees\_emp

Annees Nombre demographie employabilite

2015 2015 271840 285411732 2061432

2016 2016 296375 288420230 2075042

2017 2017 286535 291954718 2118477

2018 2018 321055 296074710 2156642

2019 2019 341175 300441712 2199248

2020 2020 184600 303982386 2081160

2021 2021 406050 305771888 2182816

2022 2022 437625 310927782 2265091

2023 2023 471815 319748258 2301906

### La régression multiple avec comme variables explicatives Années et démographie.

Regression\_Emp<-lm(Nombre~Annees+demographie+employabilite,data=canada\_demo\_annees\_emp)

> summary(Regression\_Emp)

Call:

lm(formula = Nombre ~ Annees + demographie + employabilite, data = canada\_demo\_annees\_emp)

Residuals:

2015 2016 2017 2018 2019 2020 2021 2022 2023

22832 36186 -21421 -23994 -45369 -33863 51208 -6819 21239

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.423e+07 7.296e+07 -0.332 0.7533

Annees 1.166e+04 3.749e+04 0.311 0.7684

demographie -6.203e-03 9.790e-03 -0.634 0.5541

employabilite 1.336e+00 3.578e-01 3.734 0.0135 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42870 on 5 degrees of freedom

Multiple R-squared: **0.8586**, Adjusted R-squared: 0.7737

F-statistic: 10.12 on 3 and 5 DF, p-value: 0.01453

### Interprétation et analyse du coefficient de détermination (R2)

Le coefficient de détermination (R2) ici est de 86% avec un p-value de 1.45%, ce qui indique une valeur significative. C’est un indicateur d’un bon ajustement, en outre au moins une des variables explicatives, explique bien la variable réponse.

Les coefficients de régression :

Intercept = -24 230 000 avec p-value=75%, valeur non significative. *Aucun effet statistique sur la variable reponse.*

Variable explicative = Années 11 660 : p-value=77%, valeur non significative. *Aucun effet statistique sur la variable reponse.*

Variable explicative démographie = -0.006203 : p-value=55%, valeur non significative. *Aucun effet statistique sur la variable reponse.*

Variable explicative employabilité = 1.336 : p-value=1.35%, valeur significative. Statistiquement significatif pour expliquer la variable reponse.

# Communication des résultats

## L’égalité des chiffres pour tous les trimestres.

L’analyse de l’ANOVA nous atteste que les trimestres sont égaux à un degré de signification 5% avec une probabilité de P= 0,795. Donc au regard de nos supposition de départ on peut considérer les trimestres sont restés constants et donc conclure que l’immigration suit une tendance régulière que l’on peut estimer par une régression sur les années. Les résultats du test de régression permettront de confirmer cette tendance régulière.

Une image contenant texte, capture d’écran, Police, diagramme

Le contenu généré par l’IA peut être incorrect.

Figure 10: Repartition par trimestres

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Figure 2 : Répartition des moyennes par trimestres

Une image contenant texte, diagramme, ligne, Rectangle

Le contenu généré par l’IA peut être incorrect.

Figure 3: Intervalles de confiance

L’erreur type entre les groupes est d’environs SE = 28874.

## La régression linéaire pour les années entre 2015 et 2024

Le coefficient de détermination obtenu par le test de régression, est de 86%, ce qui indique environ 86% de la variance la réponse est expliquée par les variables explicatives (années, démographie et employabilité), avec un **p-value = 0.0145** : Le **modèle global est significatif**, ce qui signifie que **au moins une** des variables explicatives explique significativement. Il s’agit d’un très bon ajustement.

# Recommandations et améliorations.

## Les limites

Le travail ci-dessus pourrait être amélioré en l’étendant à l’étude de chaque province. L’étude précédente couvre toute l’étendue du territoire du Canada. Il est connu que chacune des provinces a ses propres défis et besoins en immigration selon son contexte socio-économique et culturel. Les provinces comme le Québec ont des pouvoirs plus larges en immigration, donc une étude à l’échelle nationale ne renseigne pas d’avantage plus sur les besoins de cette province. En outre plusieurs autres variables pourraient expliquer la régression, comme le PIB par habitant. On pourrait améliorer le travail en incluant plusieurs autres variables explicatives pour étudier la variable à prédire.

## Les données aberrantes et valeurs extrêmes.

Les valeurs extrêmes principalement observées en 2020 et en 2024 peuvent s’expliquer par les effets de la pandémie mondiale de la COVID19 de 2020 qui a forcé le gouvernement fédéral à fermer les frontières pendant cette période, et le manque de données sur les trimestres 3 et 4 de l’année 2024 dans notre fichier. Ces deux éléments pourraient justifier à suffisance les valeurs extrêmes.