Labo 6.1: Statistiques descriptives univariées avec BlueSky et RStudio

Visseho Adjiwanou, PhD.

12 February 2023

Plan

1. Comprendre les deux environnements
2. Processus d'analyse des données
3. Base de données
4. Fréquences, pourcentages
5. Mesures des paramètres de tendance centrale
6. Mesure des paramètres de dispersion/variation
7. Visualisation
8.

Comprendre les deux environnements

1.1. IDE versus GUI

- IDE (integrated development environment), en français, **environnement de développement intégré** est essentiellement un éditeur de langage de programmation.
  - Exemple: RStudio
- Avantages
  - Replicabilité
- Inconvénients
  - Écrire les codes

1.2. RStudio

RStudio offre aux utilisateurs :

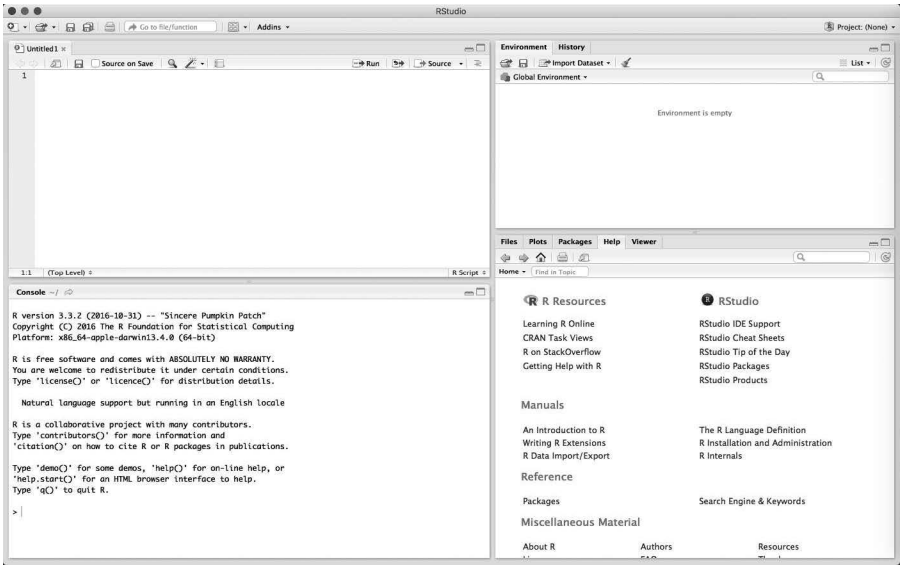
1. Un éditeur de texte pour écrire des programmes

- Console –> Calcul interactif, vérification rapide de commandes
- Script –> Écrire de programme à éxécuter
- RMarkDown –> Écriture de programme + Texte, moyen de replicabilité

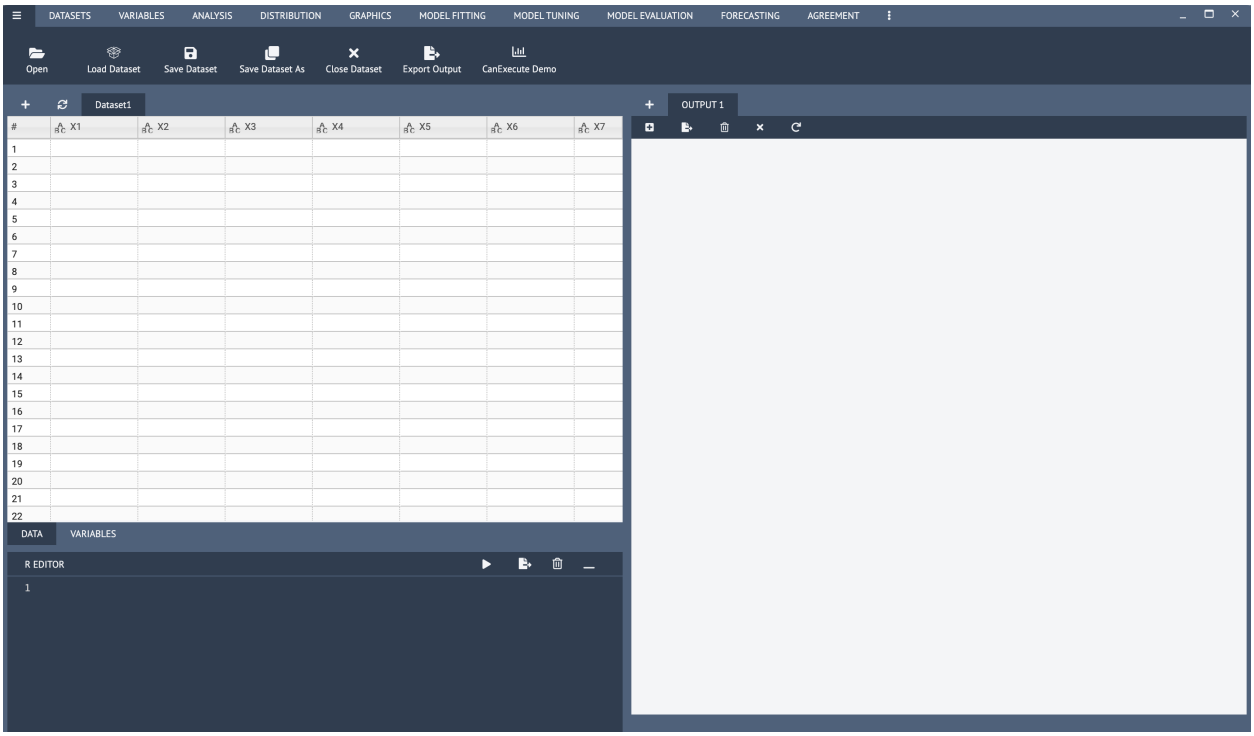
2. Un visualiseur de graphiques qui affiche les graphiques que nous créons,
3. La console R où les programmes sont exécutés,
4. Une aide section, et

5. De nombreuses autres fonctionnalités.

La figure 1.1 montre une capture d'écran de RStudio.



1.3.Présentation de BlueSky Statistics (BSS)



Les principaux menus et leur utilisation

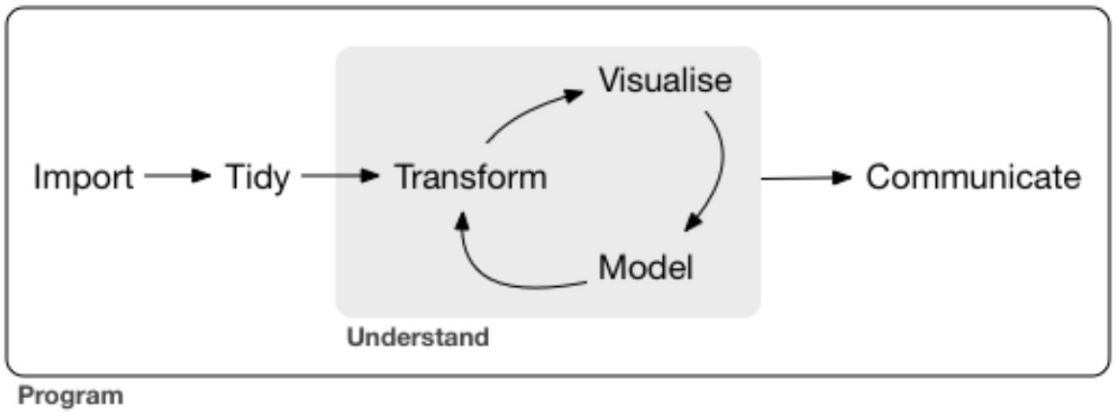
1. **Menu d'ouverture de base de données**
2. "Datasets": Traitement sur les bases de données
3. "Variables": traitement sur les variables

- Créer et recoder des variables
- Traiter les données manquantes ...

4. "Analysis":
   - *Summary* pour les fréquences, les paramètres de tendance centrale et de dispersion
   - Statistiques bivariées
5. "Graphics": pour faire les visualisations
6. "Model Fitting" pour les analyses de régression
7.

Processus d'analyse des données

- Comme dit plus haut, Tidyverse va nous servir à faire tout ce travail.
- Comme toujours, imitez au maximum ce que je fais



2.1. Processus d'analyse des données

- Résumons ce processus:
  1. Où sont les données? Vous devez les importer (**read**) pour les analyser. La manière dont vous allez les importer dépend du type de fichier.
  2. Est-ce que vous avez besoin de l'ensemble des variables du fichier de données? pas nécessairement. Vous devez sélectionner (**select**) celles qui vous intéresse
  3. Est-ce que vous travaillez sur l'ensemble de l'échantillon ou uniquement sur les femmes? Vous devez les filtrer (**filter**)
  4. Devez-vous utiliser les groupes d'âges ou les âges réels? Vous devez créer de nouvelles variables (**mutate**)
  5. Que faites-vous des individus qui n'ont pas répondu à certaines questions? leur attribuer une valeur (**impute**) ou les enlever (**na.rm pour remove na**)
  6. Que savons-nous sur les variables? Vous devez produire des statistiques descriptives (**summarize**)
- Les gras dans le diapositif précédent indique le langage que le logiciel comprend pour faire les étapes décrites plus haut
- Il comprend que l'Anglais. Chaque fois que vous voulez faire quelque chose, chercher le mot en anglais
- Il respecte une certaine manière de **parler**. Il va utiliser des symbole pour se simplifier la vie comme celui-ci par exemple %>%

2.2. Processus d'analyse des données

Chaque élément est associé à un `package` donné.

1. Importer (**readr**)
2. Préparation des données (data wrangling)
   - Arranger (**tidyr**)
   - Transformer (**dplyr**)
3. Analyse des donnés
   - Visualisation (**ggplot2**)
   - Modélisation
4. Communication (**rmarkdown**: ceci n'est pas un package de tidyverse)

PS. Intéressant sur data wrangling https://www.lemagit.fr/conseil/Quest-ce-que-le-Data-Wrangling

- Les autres packages de tidyverse
  - **sringr** : pour travailler avec les données caractères
  - **forcat** : pour travailler avec les facteurs : http://perso.ens-lyon.fr/lise.vaudor/manipulation-de-facteurs-avec-forcats/
  - **purrr** : pour travailler avec les fonctions
  - **tibble** : transformer les données en tribble.

La documentation est éparse sur chacun de ces packages.

3. Ouverture de la base des données

3.1. Introduction

Utiliser les données de l'enquête canadienne sur le revenu de 2015 pour répondre aux questions suivantes (https://search1.odesi.ca/#/). On se limitera à une base de données réduites de cette base. La base de données se nomme cis_short_5percent.csv est un échantillon de 5% pris dans la base de données originale.

Les variables retenues dans cette base sont:

- PERSONID: Person identifier (identifiant)
- AGEGP: Person's age group as of December 31 of reference year (groupe d'âge)
- SEX: Sex (Sexe du répondant)
- MARST: Marital status (Statut matrimonial)
- PROV: Province
- CFATINC: CF - After-tax income (Revenu familial après taxe)

Ce sont les variables originales de la base de données. A partir de ces variables, j'en ai créé 5 autres qui sont: - Province (créé à partir de PROV) - region (créé à partir de PROV) - sexe (créé à partir SEX) - statut_mat (créé à partir de MARST) - niveau_educ

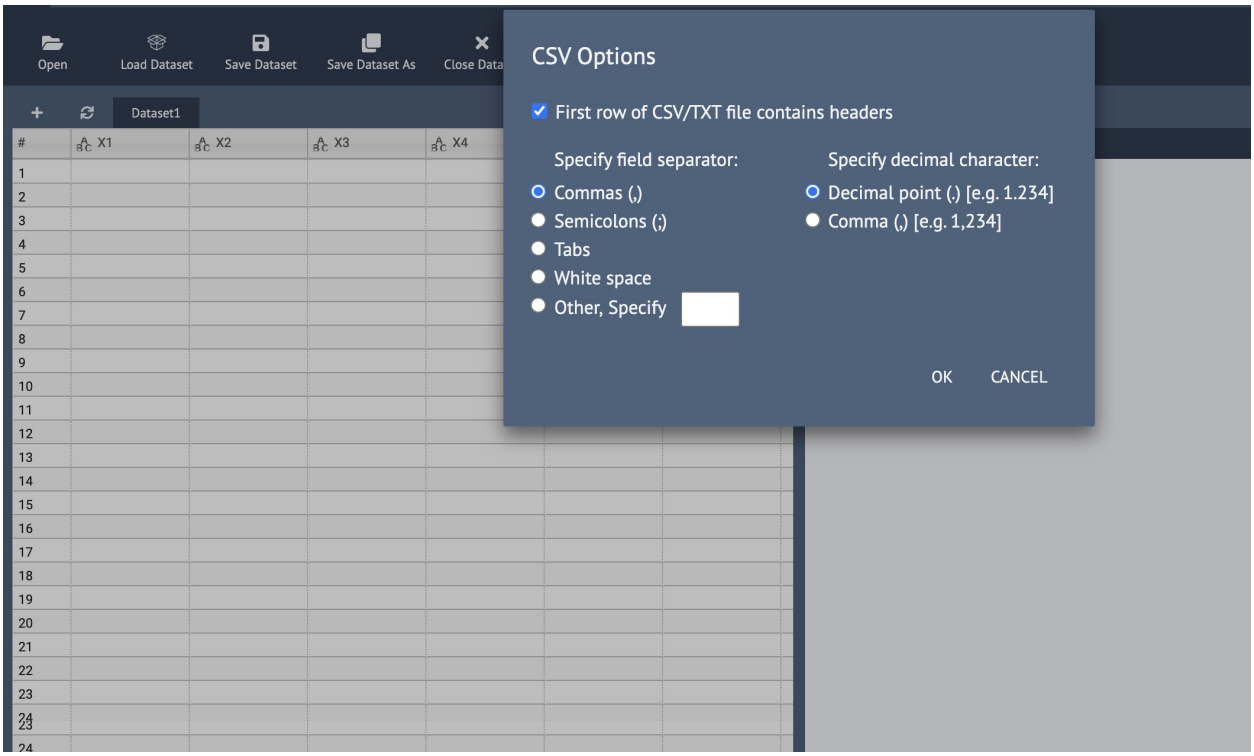
Pour comprendre les variables de cette base de données, vous pouvez consulter le fichier pdf **CIS-72M0003-E-2015_F1.pdf**. C'est ce qu'on appelle le dictionnaire de la base de données ou un Codebook. Il décrit l'ensemble des variables et présente les fréquences.

3.2. Question centrale de recherche

- Est-ce qu'il existe une inégalité de revenu entre les provinces du Canada?
- Est-ce qu'il existe une inegalité de revenu entre les hommes et les femmes au Canada?
- Quel est l'ampleur du niveau d'inégalité de revenu selon l'éducation?

3.3. Ouvrir la base de donnée

- **Avec BlueSky**



Vous devez laisser ces options comme tel.

- **Avec RSudio**

```
rm(list = ls())


library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(summarytools)
```

```
##
## Attaching package: 'summarytools'
##
## The following object is masked from 'package:tibble':
##
##     view
```

```
cis_2015 <- read_csv("cis_short_5percent_2015.csv")
```

```
## Rows: 3005 Columns: 26
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): CFID, EFID, Province, region, sexe, statut_mat, niveau_educ
## dbl (19): X___1, PUMFID, PERSONID, AGEGP, SEX, MARST, PROV, IMMSTP, YRIMMGP,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

4. Fréquences, pourcentages

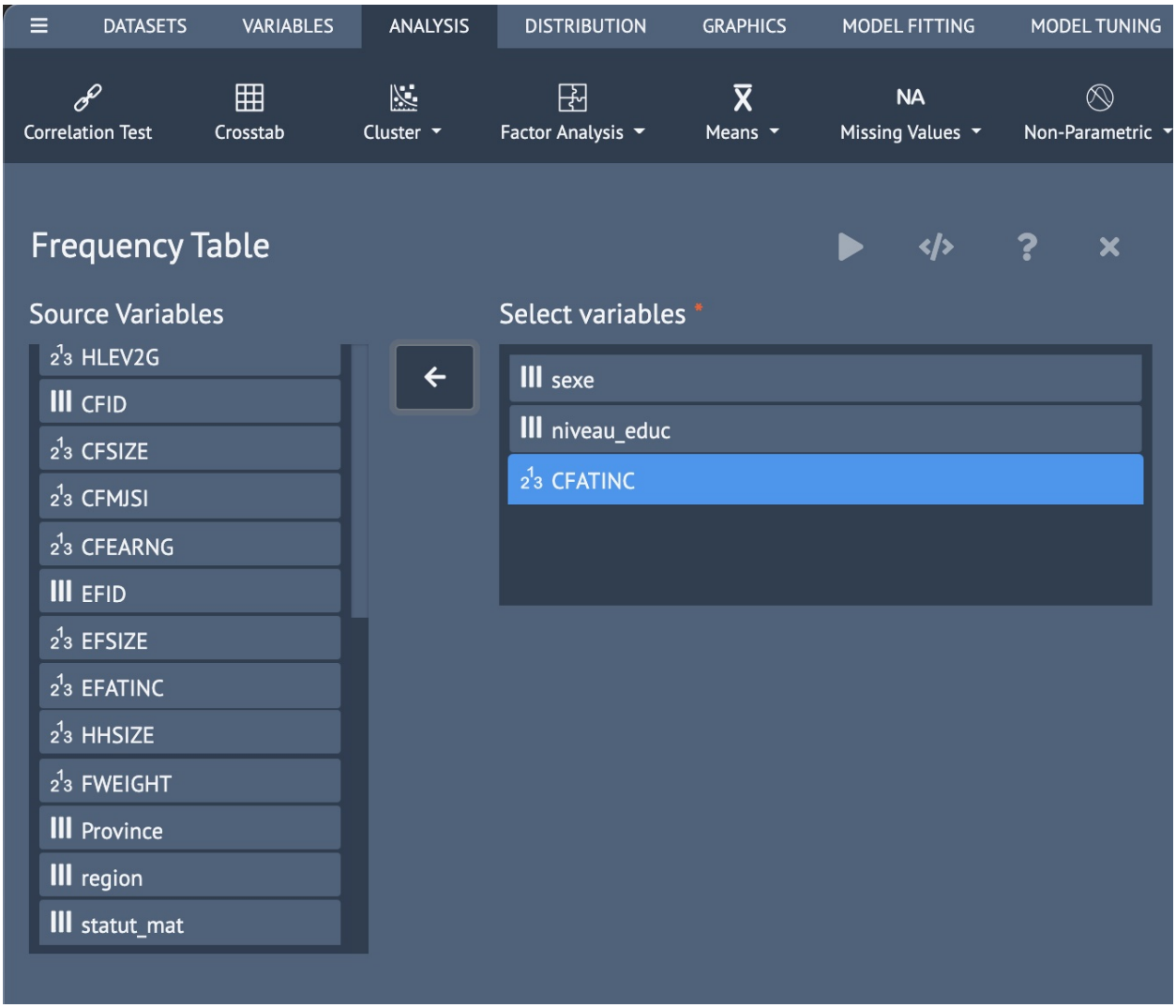
On va travailler avec les trois niveaux de mesures des variables.

- Nominale : Sexe (sexe)
- Ordinateur : Niveau d'éducation (niveau_educ)
- Intervalle/Ratio: Revenu (CFATINC)

4.1. Tableau de fréquences

- Avec BlueSky

ANALYSIS –> Summary –> Frequencies



- **Avec RStudio**

```
freq(cis_2015$sexe)
```

```
## Frequencies
## cis_2015$sexe
## Type: Character
##
##               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------ --------- -------------- --------- --------------
##       Femme   1541     51.28          51.28     51.28          51.28
##       Homme   1464     48.72         100.00     48.72         100.00
##        <NA>      0                               0.00         100.00
##       Total   3005    100.00         100.00    100.00         100.00
```

```
freq(cis_2015$niveau_educ)
```

```
## Frequencies
## cis_2015$niveau_educ
## Type: Character
##
```

```
##                                                                          Freq   % Valid   % Valid Cu
## ------------------------------------------------------------------------ ------ --------- -----------
##                  Certificat ou diplome postsecondaire non universitaire    778     25.89          25
##       Diplome d'etudes secondaires ou etudes postsecondaires partielles    640     21.30          47
##                                  Diplome ou certificat universitaire    557     18.54          65
##                               Moins que le diplome d'etudes secondaires    435     14.48          80
##                                                          Non applicable    567     18.87          99
##                                                             Non reponse     28      0.93         100
##                                                                    <NA>      0
##                                                                   Total   3005    100.00         100
```

```
#freq(cis_2015$CFATINC)
```

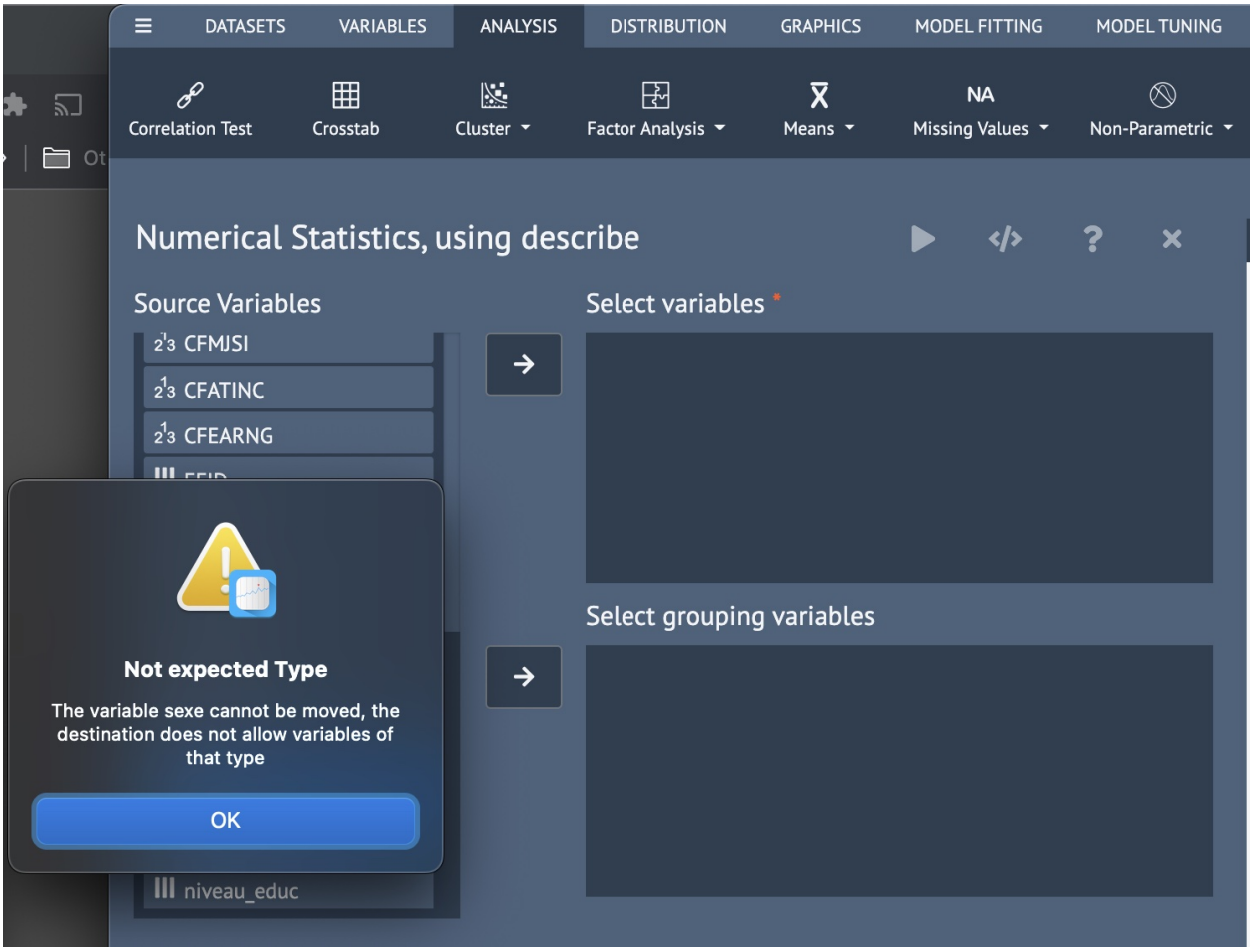
5. Mesures des paramètres de tendance centrale

5.1. Moyenne

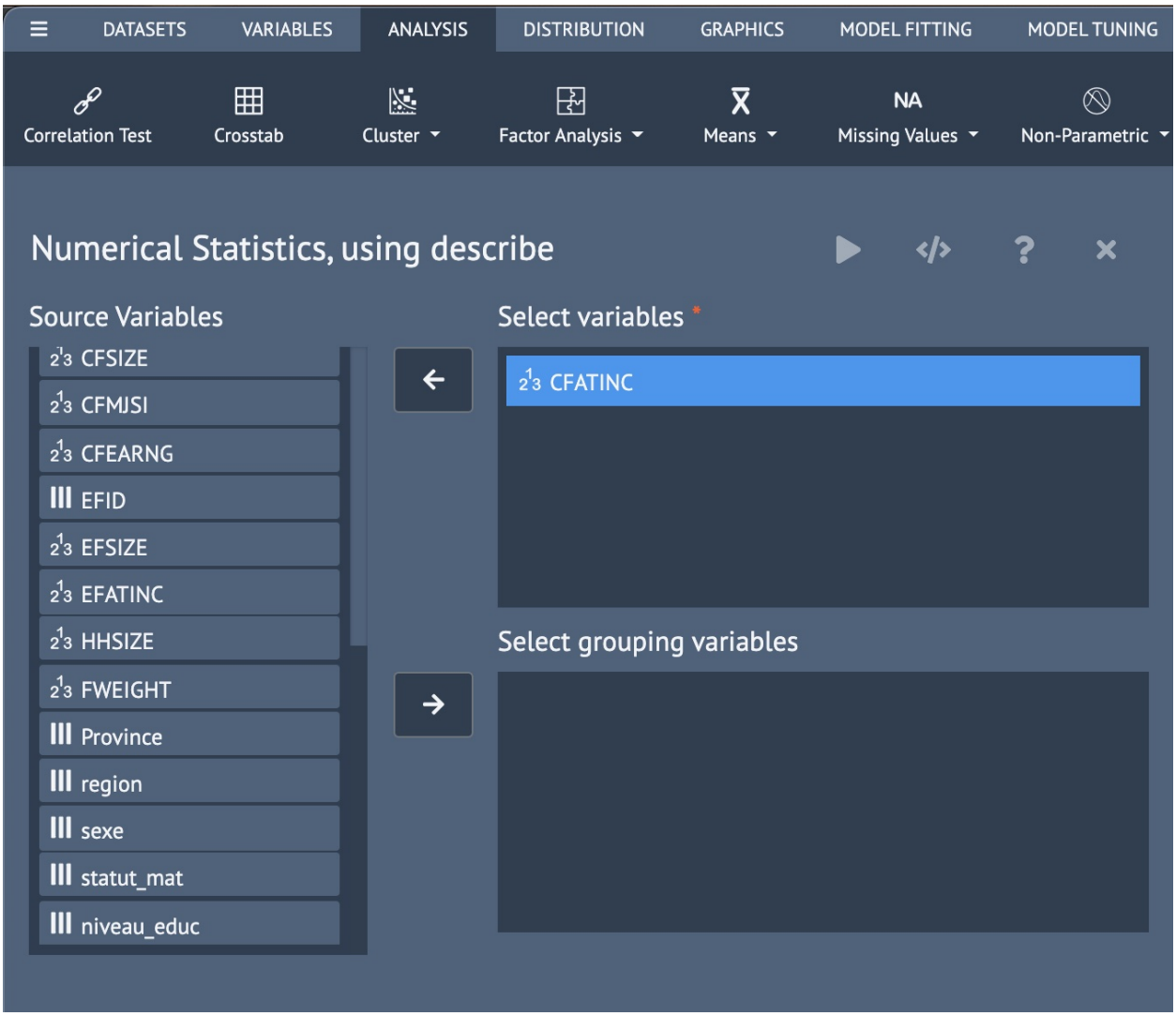
-Avec BlueSky

ANALYSIS –> Summary –> Describe

- Variable nominale ou ordinale



- Variables intervalle ou ratio



- **Avec RStudio**

Deux manières de faire:

1. Pas la bonne manière

```
mean(cis_2015$sexe)
```

```
## Warning in mean.default(cis_2015$sexe): argument is not numeric or logical:
## returning NA
```

```
## [1] NA
```

```
mean(cis_2015$SEX)
```

```
## [1] 1.512812
```

2. Une manière plus intéressante:

```
moyenne <-
  cis_2015 %>%
  summarise(moyenne = mean(CFATINC, na.rm = TRUE))
```

```
mean_income_province <-
  cis_2015 %>%
  group_by(Province) %>%
  summarise(mean(CFATINC))

mean_income_province
```

```
## # A tibble: 10 x 2
##    Province                `mean(CFATINC)`
##    <chr>                             <dbl>
##  1 Alberta                         107739.
##  2 Colombie Britanique              81451.
##  3 Ile du prince Edouard            61071.
##  4 Manitoba                         78080.
##  5 Nouveau-Brunswick                71494.
##  6 Nouvelle-<U+00C9>cosse           69294.
##  7 Ontario                          88137.
##  8 Quebec                           69958.
##  9 Saskatchewan                     85011.
## 10 Terre-Neuve-et-Labrador          76880.
```

On peut aussi calculer de cette manière les autres paramètres de tendance centrale comme la médiane ...

5.2. Revenu médian par province

```
md_income_province <-
  cis_2015 %>%
  group_by(Province) %>%
  summarise(median(CFATINC))

md_income_province
```

```
## # A tibble: 10 x 2
##    Province                `median(CFATINC)`
##    <chr>                               <dbl>
##  1 Alberta                             86805
##  2 Colombie Britanique                 71475
##  3 Ile du prince Edouard               58110
##  4 Manitoba                            69250
##  5 Nouveau-Brunswick                   64065
##  6 Nouvelle-<U+00C9>cosse              59490
##  7 Ontario                             76880
##  8 Quebec                              61650
##  9 Saskatchewan                        80540
## 10 Terre-Neuve-et-Labrador             62505
```

- A vous: calculer la médiane

6. Mesure des paramètres de dispersion/variation

6.1. Variance

- Avec BlueSky
- Avec RStudio

```
variance <-
  cis_2015 %>%
  summarise(variance = var(CFATINC, na.rm = TRUE))
```

```
variation_income <-
  cis_2015 %>%
  group_by(Province) %>%
  summarise(var(CFATINC, na.rm = TRUE))

variation_income
```

```
## # A tibble: 10 x 2
##    Province                `var(CFATINC, na.rm = TRUE)`
##    <chr>                                          <dbl>
##  1 Alberta                                   9237441980.
##  2 Colombie Britanique                       2931761450.
##  3 Ile du prince Edouard                     1278804681.
##  4 Manitoba                                  2928720726.
##  5 Nouveau-Brunswick                         2564733549.
##  6 Nouvelle-<U+00C9>cosse                    2176614009.
##  7 Ontario                                   3761201784.
##  8 Quebec                                    2096440971.
##  9 Saskatchewan                              2735434984.
## 10 Terre-Neuve-et-Labrador                   3114211663.
```

On voit que de manière générale, la variance des revenus est aussi très grands à Alberta. On voit que les niveaux de la variance sont très grands et ne sont pas de la même unité que le revenu. C'est pourquoi, on va lui préférer l'écart-type.

```
ecart_income <-
  cis_2015 %>%
  group_by(Province) %>%
  summarise(sd(CFATINC, na.rm = TRUE))

ecart_income
```

```
## # A tibble: 10 x 2
##    Province                `sd(CFATINC, na.rm = TRUE)`
##    <chr>                                         <dbl>
##  1 Alberta                                      96112.
##  2 Colombie Britanique                          54146.
##  3 Ile du prince Edouard                        35760.
##  4 Manitoba                                     54118.
##  5 Nouveau-Brunswick                            50643.
##  6 Nouvelle-<U+00C9>cosse                       46654.
##  7 Ontario                                      61329.
##  8 Quebec                                       45787.
##  9 Saskatchewan                                 52301.
```

```
## 10 Terre-Neuve-et-Labrador                    55805.
```

Vous pouvez alors aisément calculer les autres paramètres de variation.

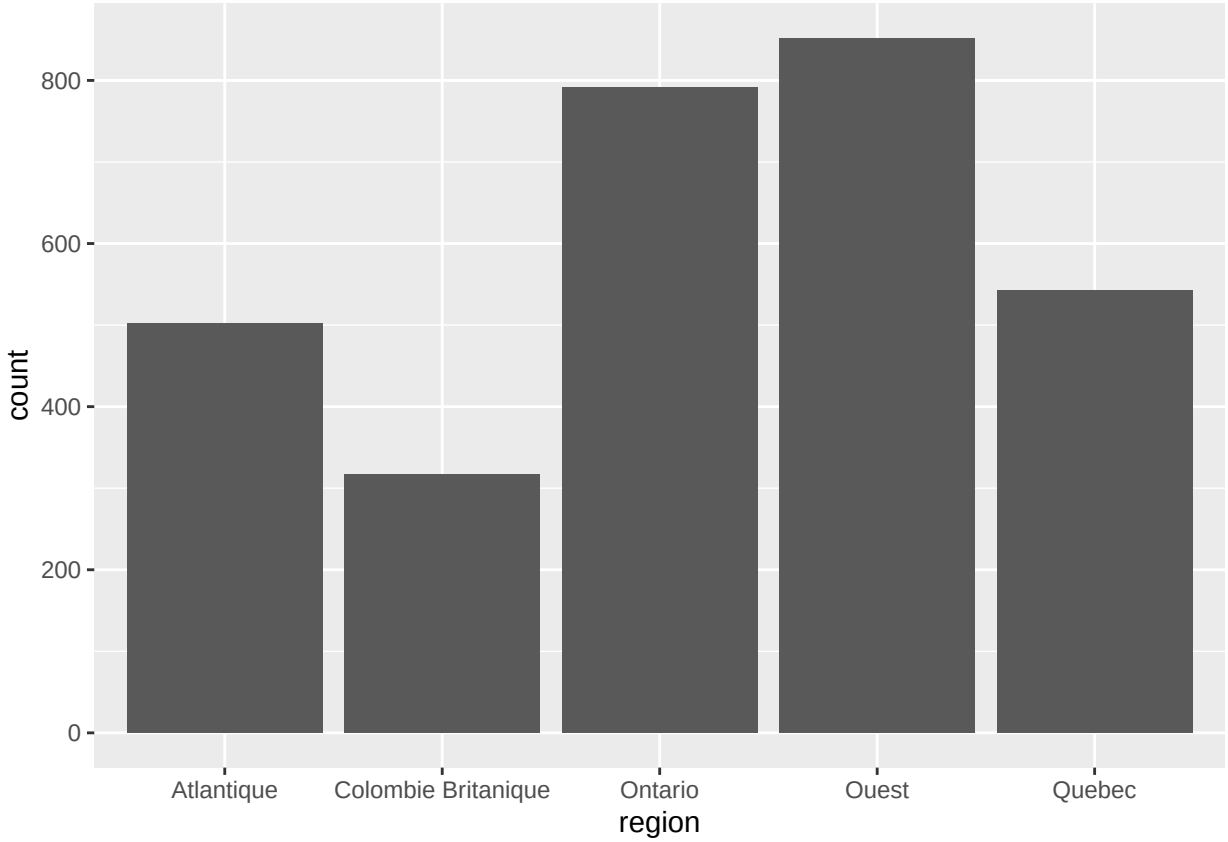
On ne peut pas comparer directement les revenus des gens de Québec avec Ontario, il faut pour cela standardisés les revenus avant de les comparer. Mais, avant de faire cela, visualisons la distribution du revenu. La visualisation est un bon moyen de se faire une première idée de la nature des données.

7. Visualisation

7.1. Variables nominales et ordinales

7.1.1. Diagramme de barre

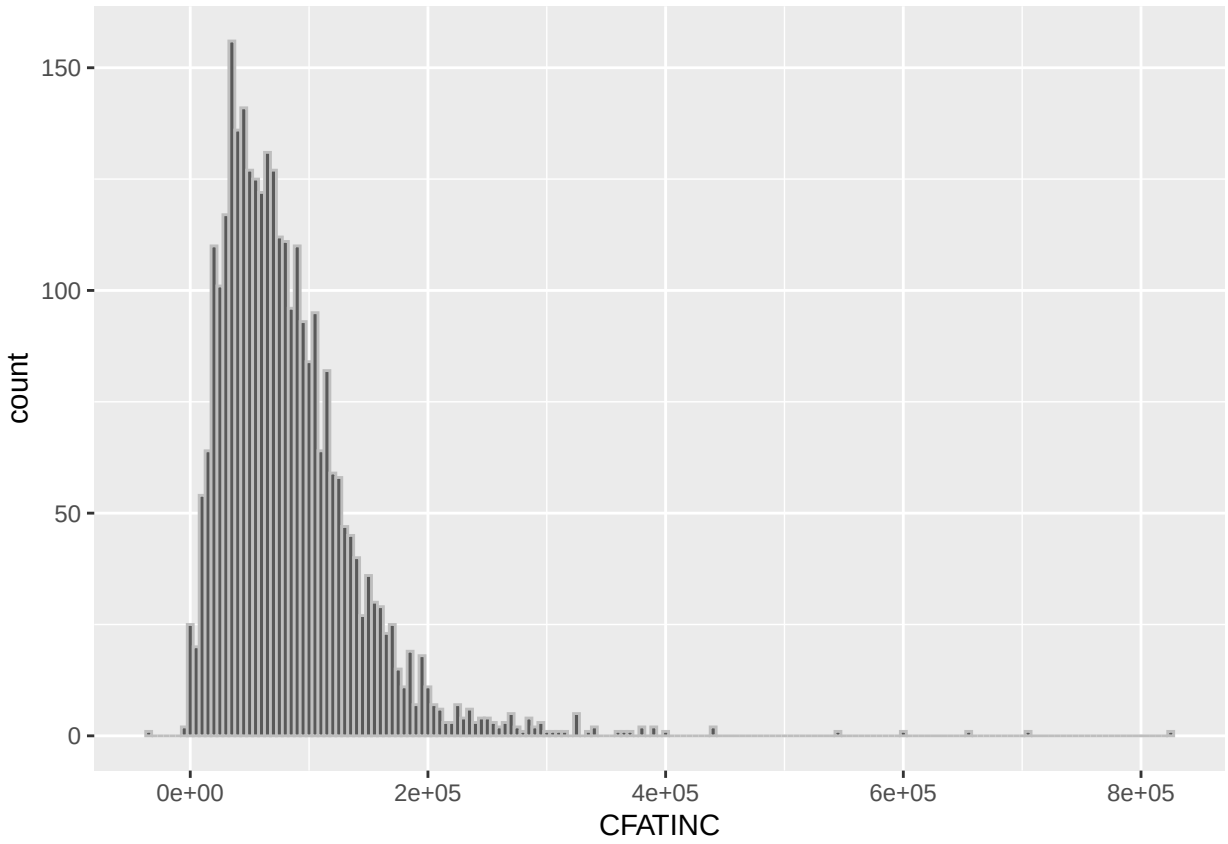
```
ggplot(cis_2015) +
  geom_bar(aes(x = region))
```



7.2. Variable de ratio et d'intervalle

7.2.1. Histogramme

```
ggplot(cis_2015) +
  geom_histogram(aes(x = CFATINC), binwidth = 5000, color = "gray")
```



Mais, on peut présenter la distribution pour chaque province. Comment pensez-vous qu'on puisse le faire?

7.2.2. Boxplot

https://statistique-et-logiciel-r.com/comment-detecter-les-outliers-avec-r/

```
ggplot(cis_2015) +
  geom_boxplot(aes(y = CFATINC))
```

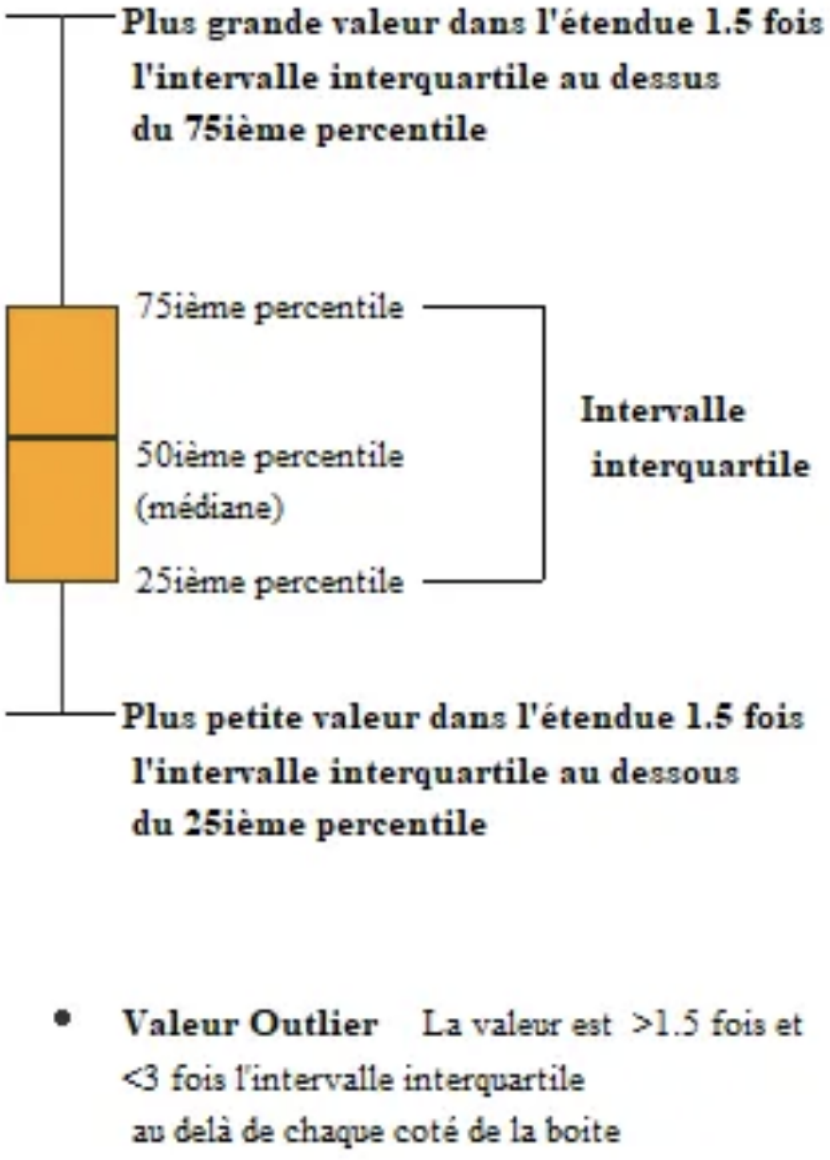
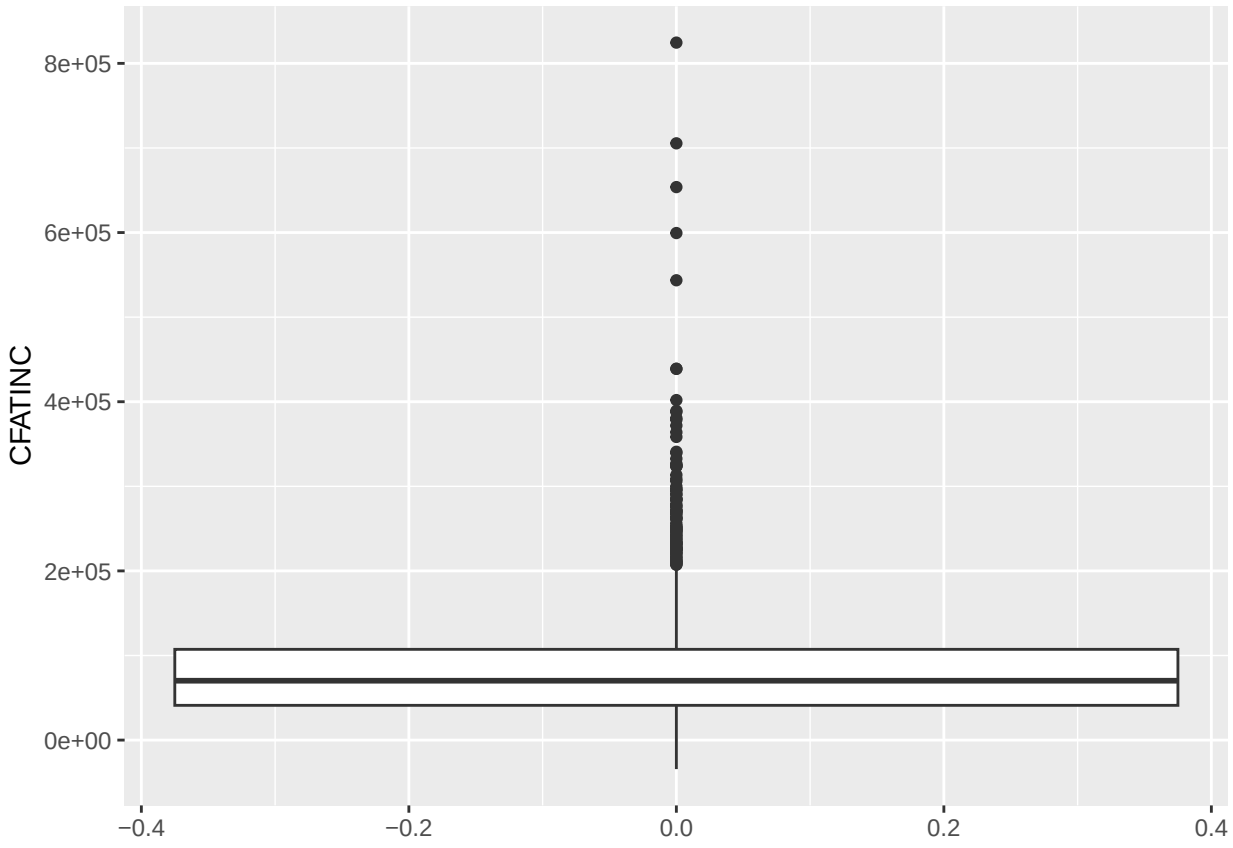
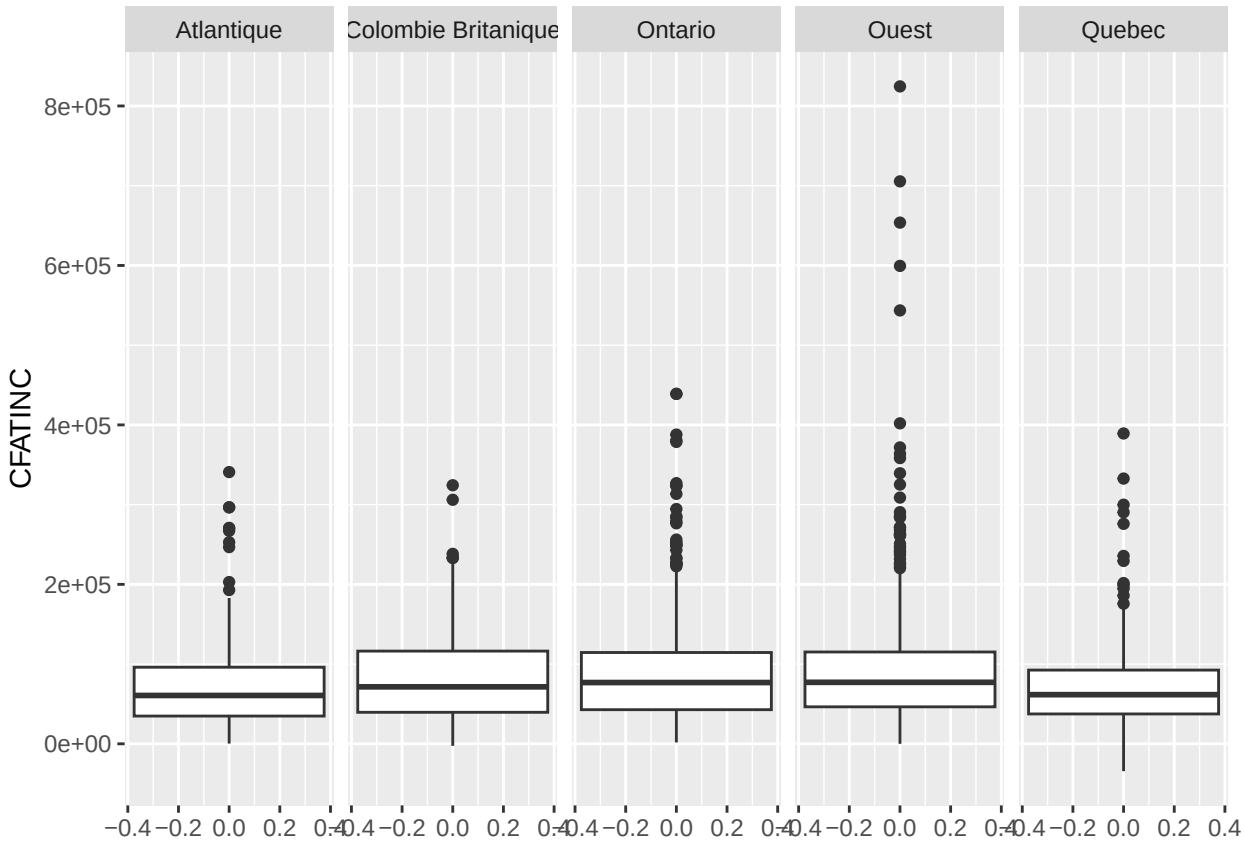


Figure 1: Interprétation boxplot



Ce graphique permet de visualiser les données abberantes ou les outliers. Un **outlier**, ou **donnée aberrante** est "une valeur ou une observation qui est « distante » des autres observations effectuées sur le même phénomène, c'est-à-dire qu'elle contraste grandement avec les valeurs « normalement » mesurées. Une donnée aberrante peut être due à la variabilité inhérente au phénomène observé ou bien elle peut aussi indiquer une erreur expérimentale. Les dernières sont parfois exclues de la série de données". Mais avant cela, voyons comment se présentent les distributions selon les régions du Canada.

```
ggplot(cis_2015) +
  geom_boxplot(aes(y = CFATINC)) +
  facet_grid(~ region)
```



8 Exercices - Extension

Base de données

1. Sélectionner les données de Québec
2. Sélectionner les données de Ontario
3. Fusionner à nouveau ces deux données pour créer la base quebec_ontario.csv
4. Calculer le revenu moyen par province et mettre le dans la base revenu_moyen_ensemble

   Datasets ==> Aggregate

Tableau de fréquence

1. calculer pour chaque province la fréquence des variables

- sexe
- état matrimonial
- niveau d'éducation
- revenu (CFATINC)

2. Montrer un peu la syntaxe des résultats

Graphiques

1. Montrer les graphiques qu'on peut créer avec ces 4 variables

- diagramme de barre
- diagramme circulaire
- carte (revenu moyen)
  - https://www.mapchart.net/

Paramètres de tendance centrale

1. Calculer pour les variables suivantes si c'est possible les paramètres de tendance centrale (moyenne, mode, médiane, premier quartile, troisimène quartile, les déciles)

- sexe
- état matrimonial
- niveau d'éducation
- revenu (CFATINC)

2. Faites les mêmes choses pour chaque province (Québec et Ontario)
3. Calculer l'indice de Palma pour le Québec et l'Ontario

Graphiques

4. Représenter graphiquement le diagramme de quartile pour la variable revenu selon la province.
5. Interpréter les résultats

Ce que je n'ai pas montré

Créer une variable

- Avec BlueSky
- Avec RStudio

```
cis_2015 <-
  cis_2015 %>%
  mutate(Sexe = case_when(
    SEX == 1 ~ "Homme",
    SEX == 2 ~ "Femme"
  ))
```