

## **Seance 3: Introduction à R - Fin**

**Bases de données en sociologie et introduction à Tidyverse**

Visseho Adjewanou, PhD.

Département de Sociologie - UQAM

30 September 2020

# Plan de présentation

- ① Sources de données en sciences sociales
- ② Analyse univariée sur variables qualitatives
- ③ Exemples
- ④ Introduction à Tidyverse
- ⑤ Introduction à Summarytools
- ⑥ Introduction au Labo

# Sources de données en sciences sociales

# Sources de données en sciences sociales

- ① Données que vous collectez vous-mêmes
- ② Données qui existent déjà

# Collecter vos propres données

## ① Avantages

- Vous collectez ce qui vous intéresse si vous devez faire une collecte formelle
- Peut aussi recourir à collecter les données des médias et réseaux sociaux

# Collecter vos propres données

## ② Inconvénients

- Peut demander beaucoup de temps de préparation
- Peut demander de la programmation
- Coûteux
- Disponibilités de multiples données qui existent déjà, pourquoi ne pas utiliser une de ses données?

# Collecter vos propres données

- Exemple : Collecter les données twitter sur le premier ministre Trudeau
- Collecter des informations sur les étudiants de l'UQAM sur leur perception sur l'immigration

# Utiliser les données qui existent déjà

## ① Sur les pays en développement

- Enquêtes démographique et de santé
- <https://dhsprogram.com/data/>

# Utiliser les données qui existent déjà

## ② Sur le Canada

- Recensements
- Enquêtes sociales générales
- Pleins d'autres
- Sondage d'opinions
  - <https://www.queensu.ca/cora/our-data/data-holdings>

# Utiliser les données qui existent déjà

## ③ Sur les USA

- <http://www.pewresearch.org/>

# Et une bonne nouvelle pour vous... .

- Il existe une base de données sur tout

<https://blog.google/products/search/discovering-millions-datasets-web/>

## Description statistique univariée sur les variables qualitatives

# Description statistique des variables qualitatives

Soit une série de valeurs qualitative: H, F, F, F, H, F, H, F, F, F, F, H, H, F, H, H, F

- donner les effectifs de chaque modalité
- donner les proportions (= fréquences) de chaque modalité par rapport au total
- combiner si besoin les proportions, notamment des proportions cumulées pour des variables ordinaires)

# Description statistique des variables qualitatives

- Une telle information peut être présentée dans un tableau dit de distribution:

Occurrence de la variable (X)	H	F	Total
Effectifs	7	10	17
Fréquence	7/17	10/17	17/17

# Description statistique des variables qualitatives

- Notation générale (avec plus de modalités)

La variable  $X$  prend les valeurs  $x_1, x_2, \dots, x_p$ ,  $n$  valeurs avec  $p$  occurrences différentes:

Occurrence de $X$	$x_1$	$x_2$	$\dots$	$x_i$	$\dots$	$x_p$	total
Effectifs	$n_1$	$n_2$	$\dots$	$n_i$	$\dots$	$n_p$	$n$
Fréquence	$f_1$	$f_2$	$\dots$	$f_i$	$\dots$	$f_p$	1

# Description statistique des variables qualitatives

- Nombre total d'observation

$$n = \sum_{i=1}^p n_i$$

- Fréquence relative

$$f_i = \frac{n_i}{n}$$

- Somme des fréquences

$$\sum_{i=1}^p f_i = 1$$

# Exemples

# Exemple 1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

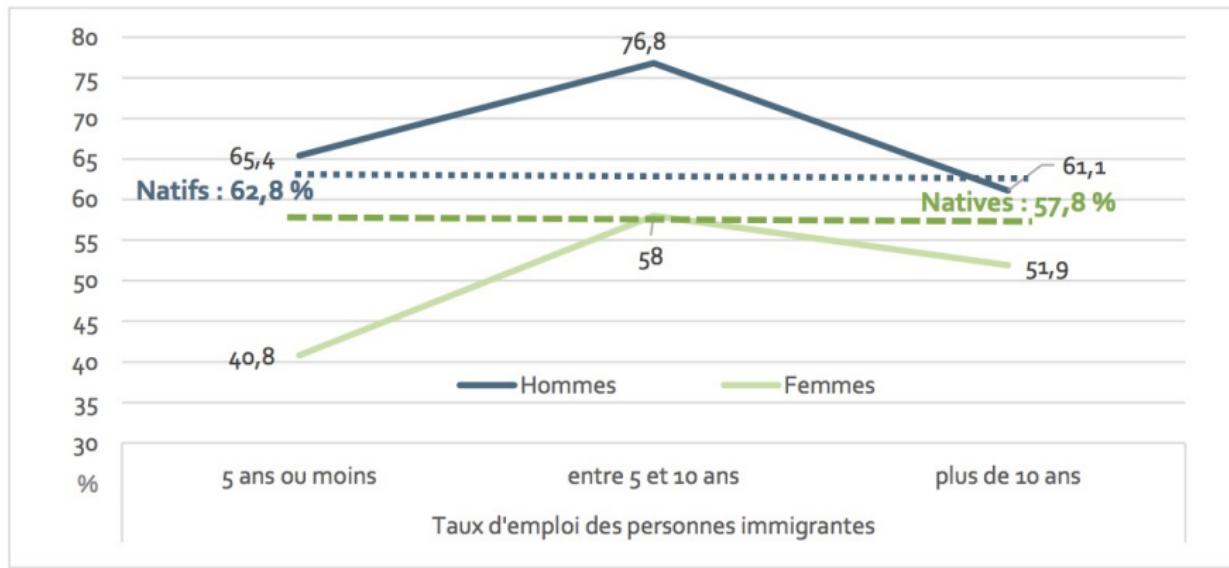
## Objectifs de l'étude:

- ① Décrire la participation des minorités ethnoculturelles dans 7 dimensions
  - Dimension 1: Économique
  - Dimension 2: Communautaire
  - Dimension 3: Culturelle
  - Dimension 4: Linguistique
  - Dimension 5: Citoyenne
  - Dimension 6: Identitaire
- ② Comparer la participation des minorités ethnoculturelles avec celle de la population majoritaire

# Exemple 1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

- Décrivez ce graphique

GRAPHIQUE 4 : TAUX D'EMPLOI SELON LA DURÉE DE RÉSIDENCE PAR SEXE, 2015



Source : Enquête sur la population active, 2015

# Exemple 1: Mesurer la participation des Québécoises et Québécois des minorités ethnoculturelles

“Les immigrants masculins participent au marché du travail avec un taux d’emploi dépassant celui des hommes natifs. Pour les femmes immigrantes, le taux d’emploi dépasse très légèrement celui des femmes natives chez celles résidant depuis 5 à 10 ans au Québec, mais demeure inférieur avant et après.” Laur, P. 19) \*

[http://www.midi.gouv.qc.ca/publications/fr/recherches-statistiques/RAP\\_Mesure\\_participation\\_2016.pdf](http://www.midi.gouv.qc.ca/publications/fr/recherches-statistiques/RAP_Mesure_participation_2016.pdf)

## Exemple 2: Relation entre niveau de scolarisation et attitude face à la violence

# Fréquences des femmes selon le niveau d'éducation

- Décrivez ce tableau

Frequencies

exemple\$education

Type: Factor

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Primaire	7572	18.64	18.64	18.64	18.64
Sans education	25999	64.02	82.66	64.02	82.66
Secondaire	6596	16.24	98.90	16.24	98.90
Université et plus	445	1.10	100.00	1.10	100.00
<NA>	0			0.00	100.00
Total	40612	100.00	100.00	100.00	100.00

# Fréquences des femmes selon l'acceptation de la violence

- Décrivez ce tableau

Frequencies

exemple\$attitude\_violence

Type: Factor

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
Non	28860	72.61	72.61	71.06	71.06
Oui	10889	27.39	100.00	26.81	97.88
<NA>	863			2.12	100.00
Total	40612	100.00	100.00	100.00	100.00

# Relation entre les deux variables

- Décrivez ce tableau

```
Cross-Tabulation, Row Proportions
education * attitude_violence
Data Frame: exemple
```

	attitude_violence	Non	Oui	<NA>	Total
education					
Primaire	5622 (74.2%)	1819 (24.0%)	131 (1.7%)	7572 (100.0%)	
Sans education	17093 (65.7%)	8297 (31.9%)	609 (2.3%)	25999 (100.0%)	
Secondaire	5712 (86.6%)	765 (11.6%)	119 (1.8%)	6596 (100.0%)	
Université et plus	433 (97.3%)	8 ( 1.8%)	4 (0.9%)	445 (100.0%)	
Total	28860 (71.1%)	10889 (26.8%)	863 (2.1%)	40612 (100.0%)	

# Défis à relever pour produire ces résultats

- Les données ne viennent pas sous une forme “clean”
- Les données sont “messy”, c'est à dire il y a beaucoup d'impuretés



# Comment ça se fait?

- Votre rôle va consister à les :
  - nettoyer
  - recoder
  - créer de nouvelles variables
- Afin de produire les résultats escomptés
- Pour nous aider à faire cela, nous allons nous servir d'un ensemble d'outils (encore appelé **packages**)
- Nous utiliserons principalement deux packages:
  - Tidyverse (qui en elle-même est un ensemble de package) et
  - Summarytools (pour faire des tableaux)

# Introduction à Tidyverse

# Processus d'analyse des données

- Tidyverse comprend un ensemble de packages qui suivent la même philosophie dont le but est de vous aider à répondre à chaque étape de votre 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**)

# Processus d'analyse des données

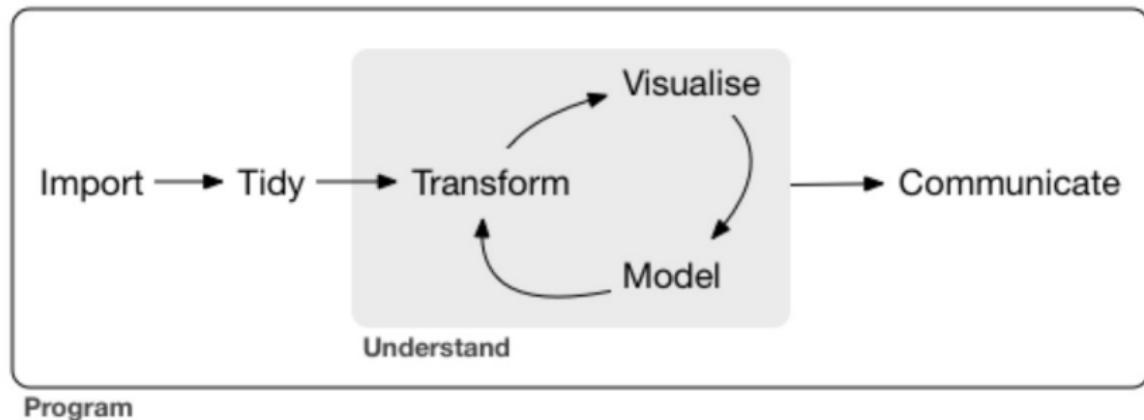
- Les gras dans le diapositif précédent indique le langage que le logiciel comprend pour faire les étapes décrites plus haut
- Il comprend que l'Anglais. Chaque fois que vous voulez faire quelque chose, chercher le mot en anglais
- Il respecte une certaine manière de **parler**. Il va utiliser des symbole pour ce simplifier la vie comme celui-ci par exemple **%>%**

# Packages de Tidyverse

```
#install.packages("tidyverse")
library(tidyverse)
```

# Processus d'analyse des données

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



# Processus d'analyse des données

Chaque élément est associé à un package donné.

- ① Importer (**readr**)
- ② Préparation des données (data wrangling)
  - Arranger (**tidyverse**)
  - Transformer (**dplyr**)
- ③ Analyse des données
  - 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>

# Processus d'analyse des données

- Les autres packages de tidyverse
  - **sringr** : pour travailler avec les données caractères
  - **forcats** : 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 éparses sur chacun de ces packages.

# Processus d'analyse des données

Et finalement deux autres packages que nous utiliserons:

- **haven**, **rio** ou **foreign** pour télécharger des données d'autres formats (sav, dta...)
- **Summarytools** pour les tableaux de fréquences et les tableaux croisés

# Informations sur les packages

- Les packages R sont une collection de fonctions R, de code conforme et d'exemples de données.
- Par défaut, R installe un ensemble de packages lors de l'installation.
- D'autres packages sont ajoutés plus tard, lorsqu'ils sont nécessaires à des fins spécifiques: c'est le cas de Tidyverse et de Summarytools
- Il existe un package pour presque tout
- Une manière de commencer par travailler facilement avec les nouveaux packages, c'est d'utiliser leur feuille de résumé s'il en existe.
- Ce lien vous renvoie à ces résumés :  
<https://rstudio.com/resources/cheatsheets/>

# Introduction à Summarytools

# Différences entre summarytools et tidyverse

- Les deux sont des packages qui s'attaquent à différents problèmes
- Summarytools va être utiliser pour présenter :
  - les statistiques descriptives sur les variables qualitatives (**freq**)
  - présenter les tableaux croisés liant deux variables descriptives (**ctable**)
- Comme pour tout package, vous devez en priorité l'installer avant son utilisation (**install.packages("nom du package")**) Cette installation se fait pour une fois de bon.
- Mais, au début de chaque utilisation, vous devez le chargez (**library(nom du package)**)
- Faites attention aux **guillemets** entre installer et charger un package: pour réussir à écrire les codes correctement, vous devez ouvrir grands vos yeux.

# Introduction au labo

# Base de données

- Nous allons utiliser les données de l'enquête “Socio-Cultural Survey” de 1996

“Les enquêtes socio-culturelles font partie d'une grande série d'études internationales comparatives des valeurs fondamentales. Des enquêtes parallèles sont réalisées chaque année dans plusieurs pays européens et aux États-Unis. CROP Inc. a commencé cette série au Canada en 1983. Ces enquêtes portent sur un large éventail d'attitudes de base - sociales, culturelles, économiques et politiques.”

# Travail à faire avant le cours

- Lisez cette note de cours et envoyez-moi vos questions
- Lisez la documentation sur l'étude **cora-crsc1996-E-1996.pdf**, disponible sur Moodle
- Choisissez :
  - Une variable dépendante de votre choix qui vous intéresse
  - Quatre variables indépendantes dont une de type qualitative **nominale**, une de type qualitative **ordonnée**, une de type quantitative **ratio** et une de type quantitative **intervalle** en lien avec votre variable dépendante
  - Votre échantillon, c'est-à-dire le groupe cible qui vous concerne
- Décrivez/Spéculer sur le lien que vous entrevoyez entre chacune de vos variables indépendantes et votre variable dépendante. Si vous faites recours à vos théories sociologiques, ce serait encore parfait.