Seance 3: Introduction à R - Fin

Bases de données en sociologie et introduction à Tidyverse

Visseho Adjiwanou, PhD.

Département de Sociologie - UQAM

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/

Exemples

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

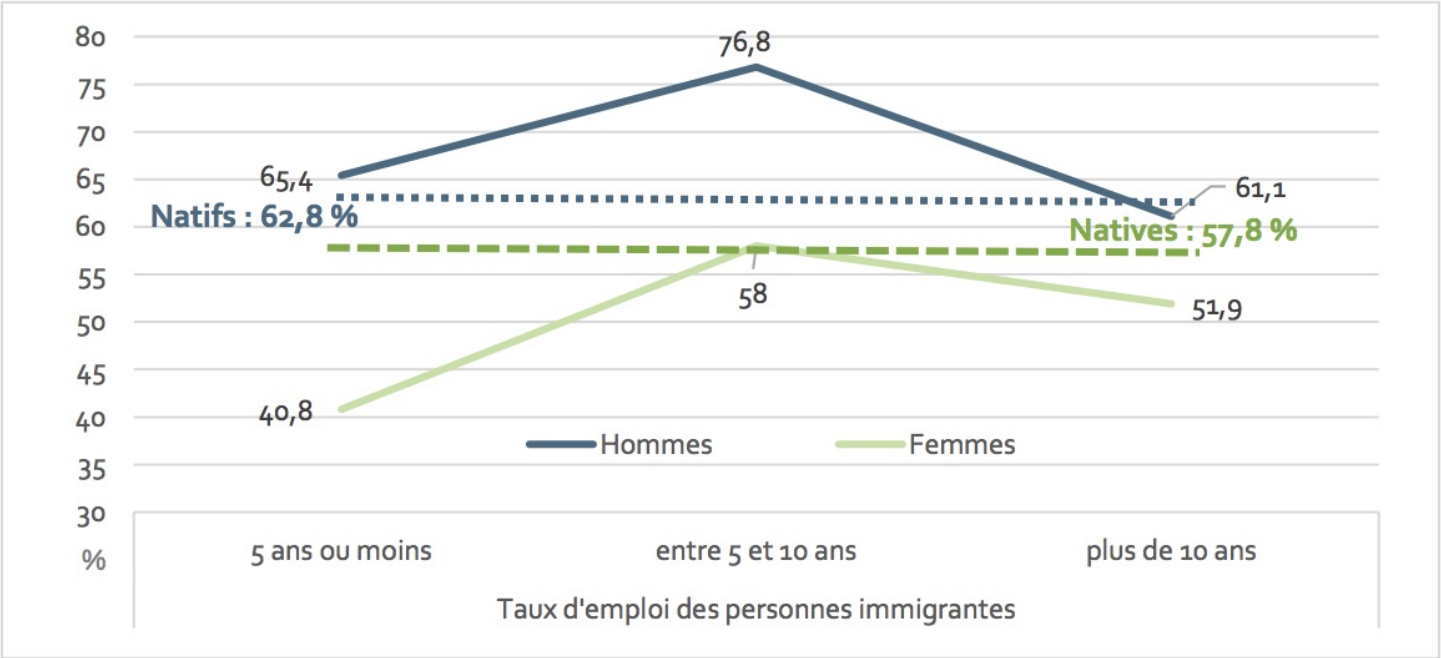
Objectifs de l'étude:

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

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

Fréquences des femmes selon le niveau d'éducation

- Décrivez ce tableau

Fréquences des femmes selon l'acceptation de la violence

- Décrivez ce tableau

Relation entre les deux variables

- Décrivez ce tableau

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 (**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

Processus d'analyse des données

- 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.

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 (**libreary(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.