Séance 2.1: Transformation et exploration des données

Visseho Adjiwanou, PhD.

SICSS-Montréal

08 June 2021

Plan de présentation

1. Manipuler et transformer les données avec base R
2. Manipuler et transformer les données avec Tidyverse
3. Exploration des données

Introduction

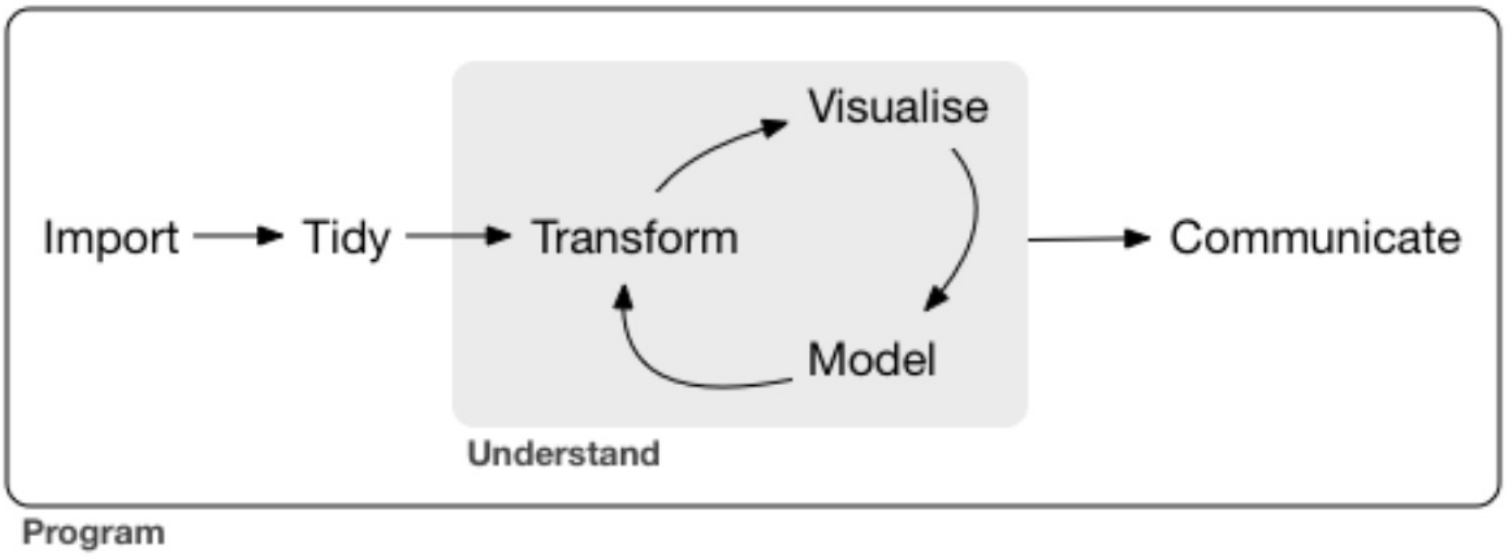

Figure 1

1. Manipuler et transformer les données avec base R

Opération sur les fichiers de données

```
UNpop_URL <- "https://raw.githubusercontent.com/kosukeimai/
UNpop <- read.csv(UNpop_URL)
```

Opération sur les fichiers de données

```
head(UNpop)
```

```
##   year world.pop
## 1 1950   2525779
## 2 1960   3026003
## 3 1970   3691173
## 4 1980   4449049
## 5 1990   5320817
## 6 2000   6127700
```

Fichier/base de données - Opération sur les fichiers de données

```
#Kenya <- read.csv("/Users/visseho/Documents/Documents - M

#head(Kenya)
```

Fichier/base de données - Opérations sur les bases de données

```
class(UNpop)
```

```
## [1] "data.frame"
```

```
names(UNpop)
```

```
## [1] "year"      "world.pop"
```

```
nrow(UNpop)
```

```
## [1] 7
```

```
ncol(UNpop)
```

Fichier/base de données - Opérations sur les bases de données

```
dim(UNpop)
```

```
## [1] 7 2
```

```
length(UNpop)
```

```
## [1] 2
```

Fichier/base de données - Opérations sur les bases de données

```
summary(UNpop)
```

```
##       year         world.pop
##  Min.   :1950   Min.   :2525779
##  1st Qu.:1965   1st Qu.:3358588
##  Median :1980   Median :4449049
##  Mean   :1980   Mean   :4579529
##  3rd Qu.:1995   3rd Qu.:5724258
##  Max.   :2010   Max.   :6916183
```

Fichier/base de données - Opérations sur les bases de données

- L'opérateur **$** est un moyen d'accéder à une variable individuelle à partir d'un objet fichier de données.
- Il renvoie un vecteur contenant la variable spécifiée.

```
UNpop[c(1, 2, 3), ]
```

```
##   year world.pop
## 1 1950   2525779
## 2 1960   3026003
## 3 1970   3691173
```

Fichier/base de données - Opérations sur les bases de données

```r
# Selectionner une variable
UNpop[, "world.pop"]
```

```
## [1] 2525779 3026003 3691173 4449049 5320817 6127700 6916
```

```r
UNpop$world.pop
```

```
## [1] 2525779 3026003 3691173 4449049 5320817 6127700 6916
```

```r
UNpop[["world.pop"]]
```

```
## [1] 2525779 3026003 3691173 4449049 5320817 6127700 6916
```

- select(UNpop, world.pop) marche aussi mais, select vient d'un

Fichier/base de données - Opérations sur les bases de données

```
UNpop[1:3, "year"]
```

```
## [1] 1950 1960 1970
```

- Que fait la commande?
- select(slice(UNpop, 1:3), year) marche aussi mais. select vient d'un autre package

Fichier/base de données - Opérations sur les bases de données

- Sélectionner les observations impaires

```
UNpop$world.pop[seq(from = 1, to = nrow(UNpop), by = 2)]
```

```
## [1] 2525779 3691173 5320817 6916183
```

Fichier/base de données - création de nouvelles variables

- Quand vous créez une nouvelle variable, il est important de la créer dans la même base de données.
- Exemple: Calculer le taux de croissance

```
UNpop$taux <- UNpop$world.pop / UNpop$world.pop[1]
head(UNpop)
```

```
##   year world.pop     taux
## 1 1950   2525779 1.000000
## 2 1960   3026003 1.198047
## 3 1970   3691173 1.461400
## 4 1980   4449049 1.761456
## 5 1990   5320817 2.106604
## 6 2000   6127700 2.426063
```

Fichier/base de données - Statistique

- Quand vous désirez calculer une statistique sur une variable, il faut créer un objet différent.
- Exemple : calculer la population mondiale totale

```
pop_totale <- sum(UNpop$world.pop)
pop_totale
```

```
## [1] 32056704
```

Fichier/base de données - Statistique

- Exemple : calculer la population mondiale moyenne

```
pop_moyenne <- pop_totale / 6
pop_moyenne
```

```
## [1] 5342784
```

```
pop_moyenne1 <- mean(UNpop$world.pop)
pop_moyenne1
```

```
## [1] 4579529
```

2. Manipuler et transformer les données avec 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.

Processus d'analyse des données

- Résumons ce processus:

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.

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.

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

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

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**)

Processus d'analyse des données

- Résumons ce processus:

Processus d'analyse des données

- Résumons ce processus:

- 4 Devez-vous utiliser les groupes d'âges ou les âges réels? Vous devez créer de nouvelles variables (**mutate**)

Processus d'analyse des données

- Résumons ce processus:

- 4 Devez-vous utiliser les groupes d'âges ou les âges réels? Vous devez créer de nouvelles variables (**mutate**)

Processus d'analyse des données

- Résumons ce processus:

- 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**)

Processus d'analyse des données

- Résumons ce processus:

- 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**)

Processus d'analyse des données

- Résumons ce processus:

- 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 se simplifier la vie comme celui-ci par exemple **%>%**

Packages de Tidyverse

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

```
## -- Attaching packages ---------------------------------

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6

## Warning: package 'tibble' was built under R version 3.6.

## Warning: package 'tidyr' was built under R version 3.6.2
```

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

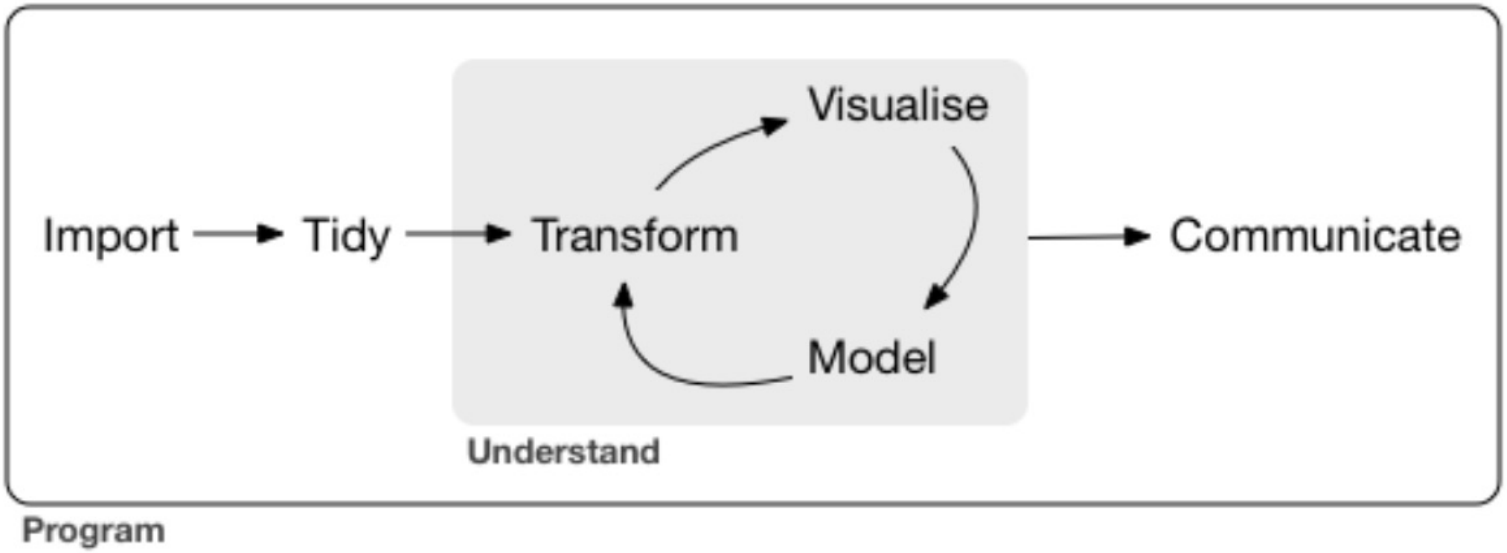

Figure 2

Processus d'analyse des données

Chaque élément est associé à un `package` donné.

1. Importer (**readr**)
2. Transformation 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)

Processus d'analyse des données

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.

Différences dans les codes

	Base r	Tidyverse
	. (ex. read.csv)	_ (ex. read_csv)
pipes	()	%>%
Creer variable	()	mutate
Position	mean()	summarise
	median()	
Dispersion	var()	summarise
	sd()	
Analyse/groupe		group_by
graphique	hist	ggplot

1. Sélection des variables

Voir Séance 2.2

2. Sélection des observations

Séance 2.2

3. Créations de nouvelles variables

- Variables quantitatives
- Variables factorielles

Séance 2.2

3. Exploration des données

1. Variables factorielles

- Tableau de fréquences

Voir Séance 2.3

2. Variables quantitatives

- Paramètres de tendances centrales
- Paramètres de dispersion

Voir Séance 2.3

3. Relations entre deux variables

Voir Séance 2.3