Données

Constitution et nettoyage

Véronique Tremblay

Retour sur les étapes d'un projet d'analyse de données

1. Définition des objectifs
2. Données
   - Inventaire et qualité
   - Constitution et nettoyage
   - Exploration et traitement préliminaire
3. Élaboration et validation des modèles
4. Mise en oeuvre
5. Suivi de la performance et amélioration

Extraction des données

Extraire les données (suite)

Librairies pour l'importation de données en R

Format	Extension	Librairie
Texte	.txt, .csv	readr
Excel	.xlsx	readxl
SAS	.sas7bdat	haven
SPSS	.sav, .zsav	haven
JSON	.json	jsonlite
Shapefile	.shp	shapefile
Raster	.grd (et autres)	raster

Nettoyage

Jeu de données propre

- Une et une seule ligne par observation

- Une colonne par variable

Les types de variables

«In computer science, the pedagogical focus is on preventing the computer from doing stupid things with data.
In statistics, it's about preventing you from doing stupid things with data.»

Cassie Kozyrkov

Chief Decision Scientist at Google, Inc.

Le fomat des types de variables

Type	Format en R
Quantitatives	`numeric` ou `double`
Catégorielles	`factor`
Ordinales	`ordered factor`
Identifiants	`character`
Dates	`POSIXlt` ou `POSIXct`

Jeu de donnée propre, la suite

- Retirer les doublons
  Truc: fonction `discinct`

- Uniformiser les modalités
  Truc: librairie `stringr` et regex

- Vérifier le format des valeurs spéciales (`NA`)
  Truc: fonction `na_if`

Résumé

- Extraire

- Nettoyer