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

La constitution de la base de données

Extraire les données

Les données peuvent être emmagasinées de différentes façons:

- Fichiers plats
- BD Relationnelle
- Entrepôts pour données massives (type Hadoop)
- ...

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
Page web	.html	xml, httr, RCurl

Formater

Jeu de données structuré et 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 vaviables

Type	Format en R
Quantitatives	`numeric` ou `double`
Catégorielles	`factor`
Ordinales	`ordored factor`
Identifiants	`character`
Valeurs manquantes	`NA`
Dates	`POSIXlt` ou `POSIXct`

Nettoyer

- Retirer les doublons
  Truc: fonction `discinct`

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

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

Pivoter

Format Long

Client	Âge	Produit	Montant
A	32	Auto	350
A	32	Habitation	1200
B	55	Auto	200
C	55	Habitation	900

Format Large

Client	Âge	Auto	Habitation
A	32	350	1200
B	55	200	
C	55		900

Pivot: long vers large

Client	Âge	Produit	Montant
A	32	Auto	350
A	32	Habitation	1200
B	55	Auto	200
C	55	Habitation	900

```
pivot_wider(don,
            names_from = Produit,
            values_from = Montant) %>%
  kable()
```

Pivot: large vers long

Client	Âge	Auto	Habitation
A	32	350	1200
B	55	200	
C	55		900

```
pivot_longer(large,
             cols = c('Auto', 'Habitation'),
            names_to = "Produit",
            values_to = "Montant",
            values_drop_na = TRUE)
```

Les jointures

Auto

Client	Âge	Montant
A	32	350
B	55	200

Habitation

Client	Âge	Montant
A	32	1200
C	55	900

Jointure externe complète (`full_join`)

Avant la jointure

Client	Âge	Montant
A	32	350
B	55	200

Client	Âge	Montant
A	32	1200
C	55	900

Jointure à complète

Client	Âge	Montant.Auto	Montant.Habitation
A	32	350	1200
B	55	200	
C	55		900

Jointure à gauche (`left_join`)

Avant la jointure

Client	Âge	Montant
A	32	350
B	55	200

Client	Âge	Montant
A	32	1200
C	55	900

Jointure à gauche

Client	Âge	Montant.Auto	Montant.Habitation
A	32	350	1200
B	55	200	

Jointure à droite (`right_join`)

Avant la jointure

Client	Âge	Montant
A	32	350
B	55	200

Client	Âge	Montant
A	32	1200
C	55	900

Client	Âge	Montant.Auto	Montant.Habitation
A	32	350	1200
C	55		900

Jointure interne (`inner_join`)

Avant la jointure

Client	Âge	Montant
A	32	350
B	55	200

Client	Âge	Montant
A	32	1200
C	55	900

Jointure interne

Client	Âge	Montant.Auto	Montant.Habitation
A	32	350	1200

Jointure en R

```
joint <- inner_join(Auto, Habitation,
                    by = c('Client', 'Âge'),
                    suffix = c(".Auto", ".Habitation")
                    )
```

Client	Âge	Montant.Auto	Montant.Habitation
A	32	350	1200

Documenter

- Une ligne par variable
- Le type de variable (continu, catégorielle ...)
- La source
- Les modalités
- Les unités
- Une note explicative et des éléments descriptifs

Documentez dans le script!

Résumé

Constitution et nettoyage de la base de données

- Extraire
- Formater
- Nettoyer
- Pivoter
- Joindre
- Documenter