# Données

## Exploration et traitement préliminaire

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

# Exploration des données

# Pourquoi explorer les données?

- Vérifier si on a tout le nécessaire
- Poursuivre le nettoyage
- Choisir un modèle approprié
- Prévenir des problèmes

# Qu'est-ce qu'on cherche?

- Les modalités rares
- Les modalités trop nombreuses
- L'asymétrie
- Le débalancement des classes
- Les valeurs extrêmes ou aberrantes
- Les variables fortement corrélées
- Les valeurs manquantes

# Statistiques descriptives unidimensionnelles

- Numérique:
  Moyenne, minimum et maximum
  Tableau de fréquences (sommes)
- Graphique
  Histogrammes
  Diagrammes à boîte (box-plot)

# Statistiques descriptives bidimensionnelles

- Numériques:
  Tableaux croisés
  Corrélation
- Graphique:
  Diagrammes de dispersion
  Analyse des correspondances

# Statistiques descriptives multidimensionnelles

- Numérique
  Distance de Mahalanobis
  ACP
- Graphique
  ACP, ACM

# Transformer les variables

# Transformations simples

- $\log(X)$

- Standardisation

# Transformations simples

Regrouper des modalités

- Voiture: sous-compacte, compacte, intermédiaire, grande et très grande
- Entreprises :« SIC » à un niveau supérieur
- Codes postaux en RTA

# Combinaison de variables

- Date du premier achat - date de naissance
- Date d'aujourd'hui - date du dernier achat
- Somme de tous les achats dans les 12 derniers mois
- Prix payé l'an passé - prix payé cette année

# Réduction de la dimensionnalité

- Sélection de variables
- ACP
- ACM

# Résumé

- Exploration

- Transformation