

HEC MONTRÉAL

Analyse multidimensionnelle appliquée



Survol du cours

- **Objectif** : Comprendre et appliquer des techniques statistiques utiles à l'intelligence d'affaire

Techniques	No de séances (approximatif)
2. Analyse factorielle exploratoire	3
3. Analyse de regroupement	4-5
4. Sélection de variables et de modèles	5-7
5. Régression logistique	8-10
6. Analyse de survie	11-12
7. Données manquantes	13

Mise en contexte

DOMAINES D'APPLICATION

Marketing (intelligence d'affaires)

- Acquisition de clientèle
- Fidélisation de la clientèle
- Vente croisée (*cross-sell*)
- Vente incitative (*up-sell*)
- Rétention

Recherche marketing

- Analyse de données de sondage
- Segmentation

Gestion des risques

- Automatisation des décisions de crédit
- Prévion des pertes

Fraude

SECTEURS D'ACTIVITÉ

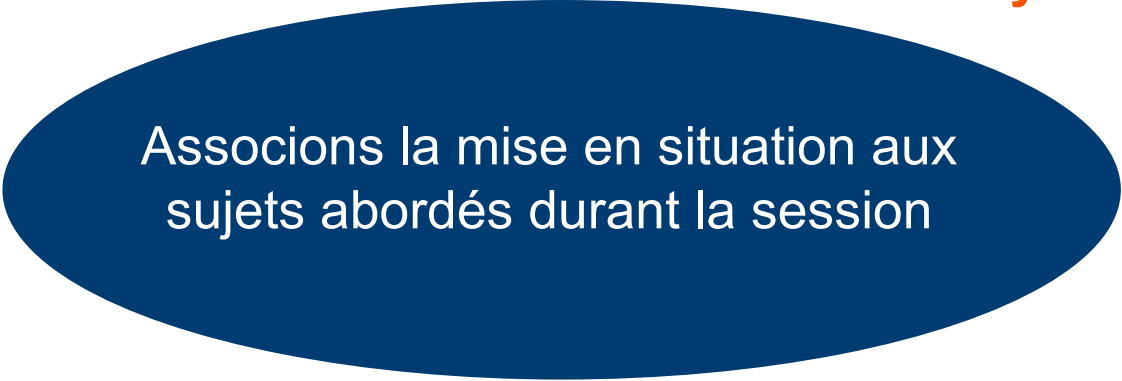
- Commerce au détail
- Télécommunication
- Firmes de sondage
- Institutions financières
- Assurances
- **Toute entreprise accumulant une grande quantité d'information (données)**

Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle



Associons la mise en situation aux
sujets abordés durant la session

Analyse de regroupement

Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Répondez aux questions suivantes sur une échelle de 1 à 5 où 1 veut dire pas important et 5 très important
Pour vous, à quel point est-ce important...

- 1) que le magasin offre de bons prix tous les jours?
- 2) que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)?
- 3) que le magasin offre des produits de qualité?
- 4) que les vendeurs connaissent bien les produits?
- 5) qu'il y ait des ventes spéciales régulièrement?
- 6) que les marques connues soient disponibles?
- 7) que le magasin ait sa propre carte de crédit?
- 8) que le service soit rapide?
- 9) qu'il y ait une vaste sélection de produits?
- 10) que le magasin accepte le paiement par carte de débit?
- 11) que le personnel soit courtois?
- 12) que le magasin ait en stock les produits annoncés?

Organisation du cours – sujets abordés

Répondez aux questions suivantes sur une échelle de 1 à 5 où 1 veut dire pas important et 5 très important
Pour vous, à quel point est-ce important...

- 1) que le magasin offre de bons prix tous les jours?
- 2) que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)?
- 3) que le magasin offre des produits de qualité?
- 4) que les vendeurs connaissent bien les produits?
- 5) qu'il y ait des ventes spéciales régulièrement?
- 6) que les marques connues soient disponibles?
- 7) que le magasin ait sa propre carte de crédit?
- 8) que le service soit rapide?
- 9) qu'il y ait une vaste sélection de produits?
- 10) que le magasin accepte le paiement par carte de débit?
- 11) que le personnel soit courtois?
- 12) que le magasin ait en stock les produits annoncés?

Ces trois questions indiquent l'importance accordée au service.

Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Une firme de recherche marketing voudrait regrouper les questions de son sondage qui sont très corrélées afin de construire des sous-échelles.

Analyse de regroupement

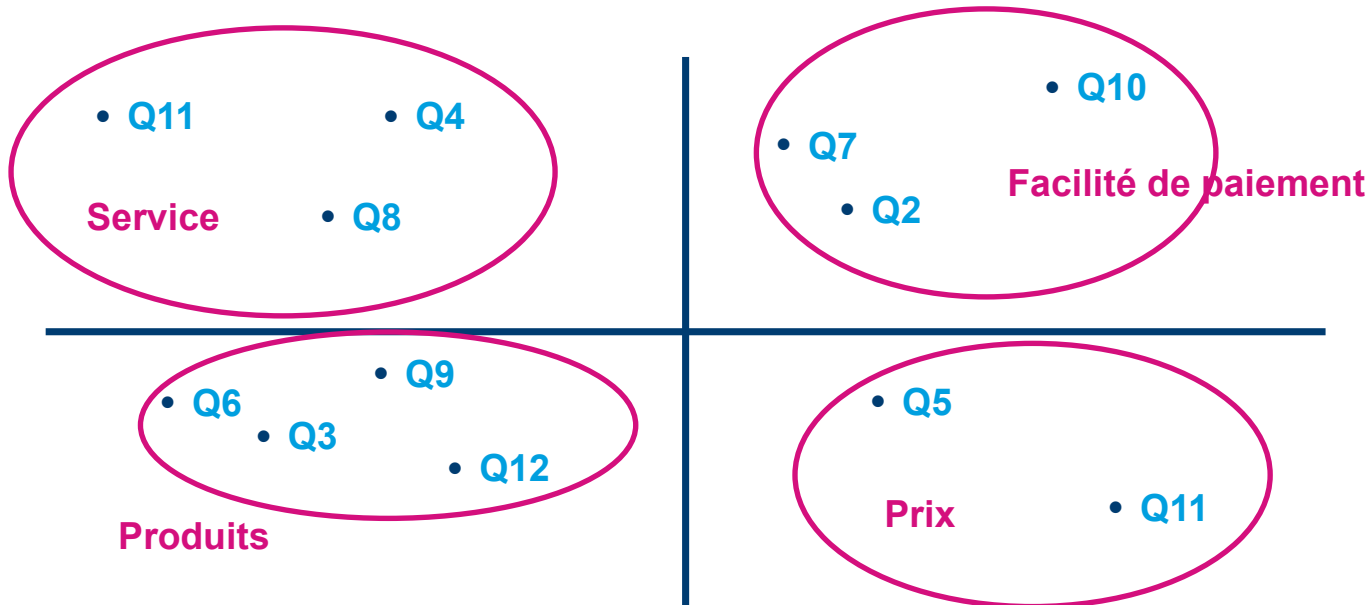
Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Analyse factorielle



Organisation du cours – sujets abordés

Analyse de regroupements

Segmentation de marché.

« ...définir des sous-groupes réunissant des consommateurs qui partagent les mêmes préférences ou qui réagissent de façon semblable à des variables de marketing ».

tiré de :

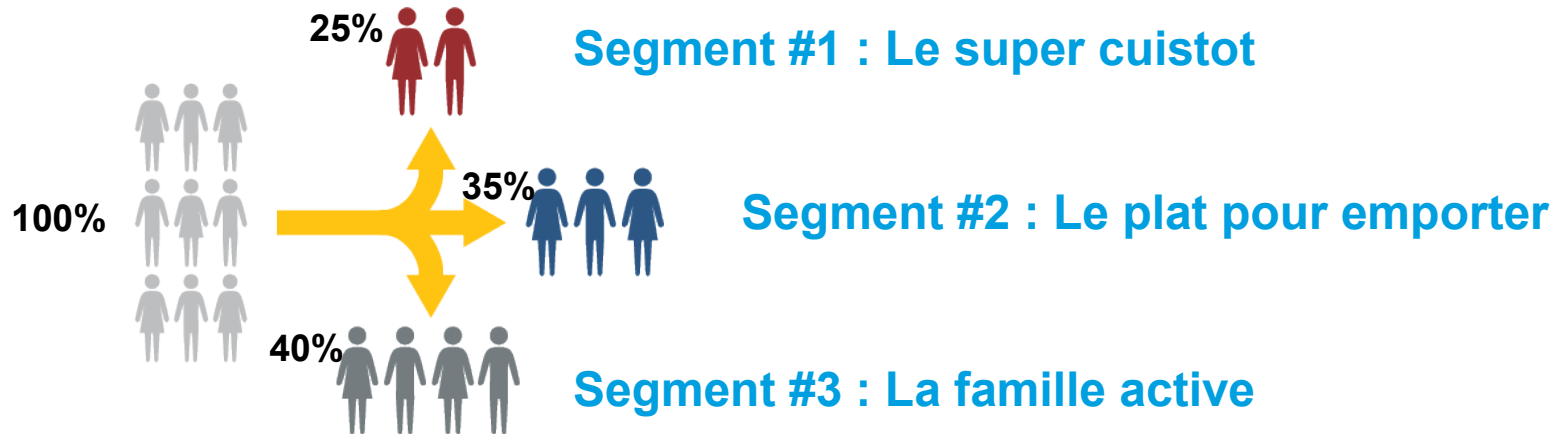
d'Astous, A. (2000). Le projet de recherche en marketing, 2/e édition. Chenelière/McGraw-Hill.



Organisation du cours – sujets abordés

Analyse de regroupement (suite)

Exemple : Suite à un sondage auprès de la population québécoise sur les habitudes alimentaires, une chaîne de magasins d'alimentation veut segmenter les consommateurs.



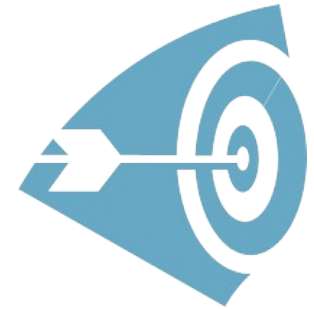
La chaîne de magasins pourrait se poser les questions suivantes :

- Est-ce que mes clients se distribuent comme la population québécoise?
- Dans quels segments ai-je les clients les plus profitables?

Organisation du cours – sujets abordés

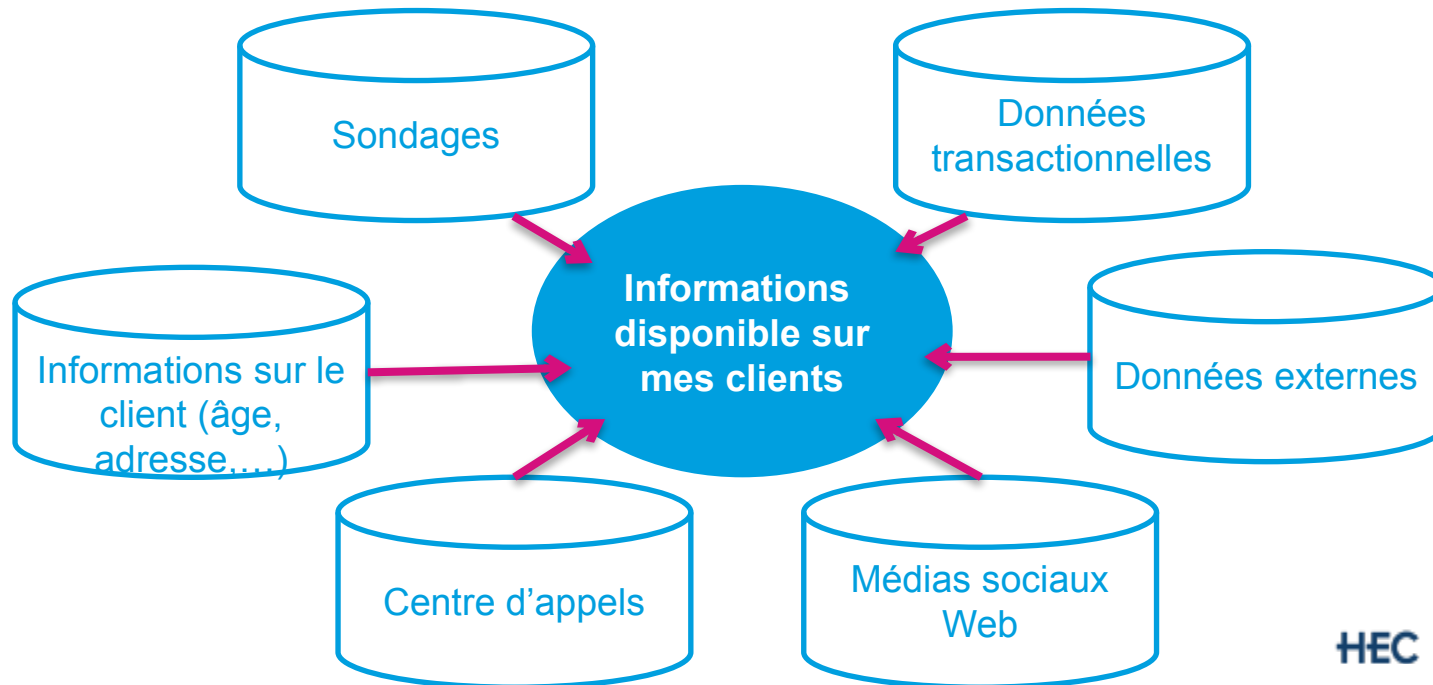
Développement d'un modèle de prévision

- Problématiques :
 - Détecter les faillites des clients (ou des entreprises)
 - Cibler les clients qui seront intéressés par une offre promotionnelle
 - Détecter les fraudes (par carte de crédit ou dans les rapports de revenus)
 - Prévoir d'avance si un client va nous quitter.
- Techniques :
 - régression linéaire ou logistique
 - réseaux de neurones
 - arbres de régression ou de classification
 - Etc...



Organisation du cours – sujets abordés

Plusieurs sources de données...



Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Vous voulez construire un modèle prédictif.
Pour chaque client de l'entreprise, vous avez
accès à plus de 200 variables. Quelles sont les
informations à considérer dans votre modèle?

Analyse de regroupement

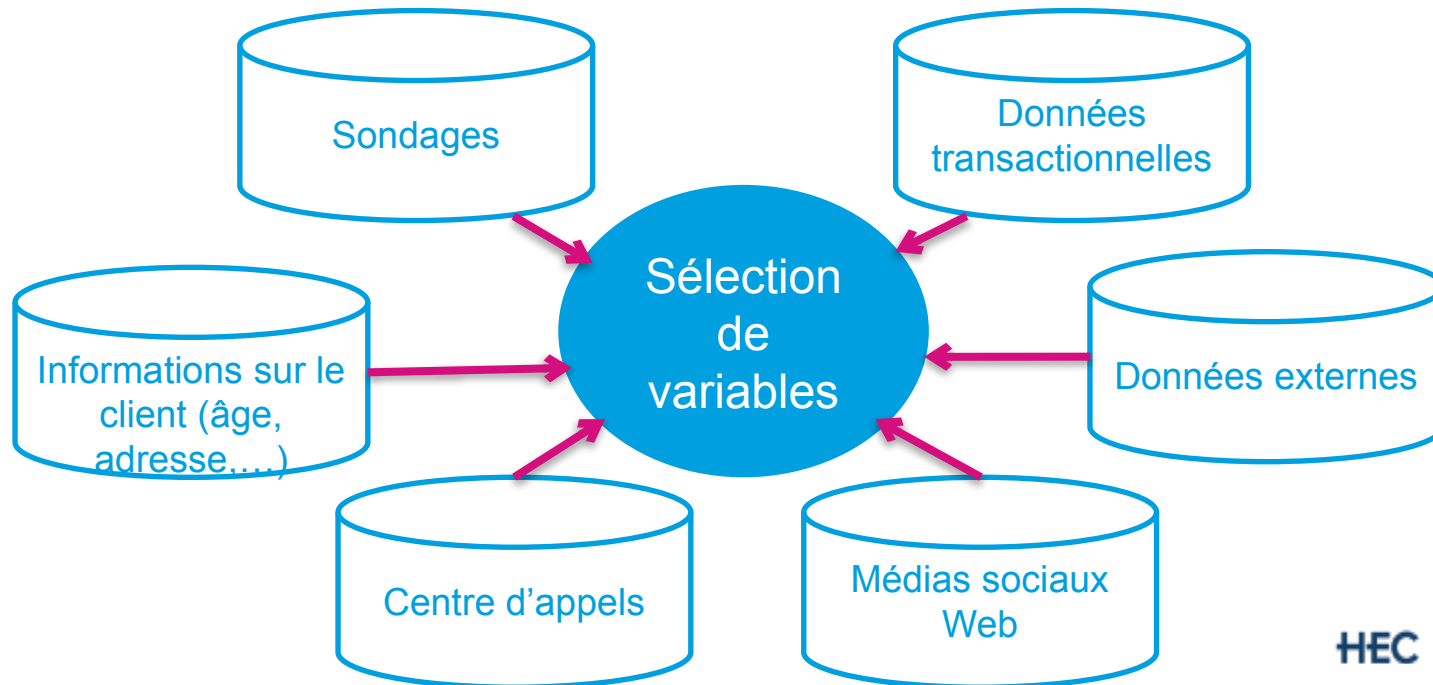
Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Sélection de variables et de modèles



Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Une banque veut construire un modèle de prévision permettant d'évaluer la probabilité pour un client de cesser le remboursement de son prêt.

Analyse de regroupement

Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Régression logistique

- La régression logistique est un modèle adéquat lorsque la variable à expliquer (cible) est binaire (0-1) :
 - Accepter / Refuser
 - Quitter / Rester
 - Faillite
- Les objectifs :
 - Explicatif
 - Prédictif

Organisation du cours – sujets abordés

Régression logistique

Marketing

- Une compagnie de télécommunication veut déterminer parmi les détenteurs du produit A, ceux qui sont les plus propices à acheter le produit B (modèle de vente croisée).
- Suite à une campagne marketing, une compagnie veut déterminer les caractéristiques qui distinguent les clients qui ont accepté l'offre de ceux qui l'ont refusée.

Gestion des risques

- Une banque veut déterminer si un client est à risque de cesser le remboursement de son prêt ou non.
- Une compagnie d'assurance veut estimer la probabilité pour ses clients de faire une réclamation dans la prochaine année.

Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Un organisme de charité aimerait classer ses donateurs dans des groupes selon leurs caractéristiques (âge, montant des dons, fréquence des dons, canal de communication,...)

Analyse de regroupement

Régression logistique

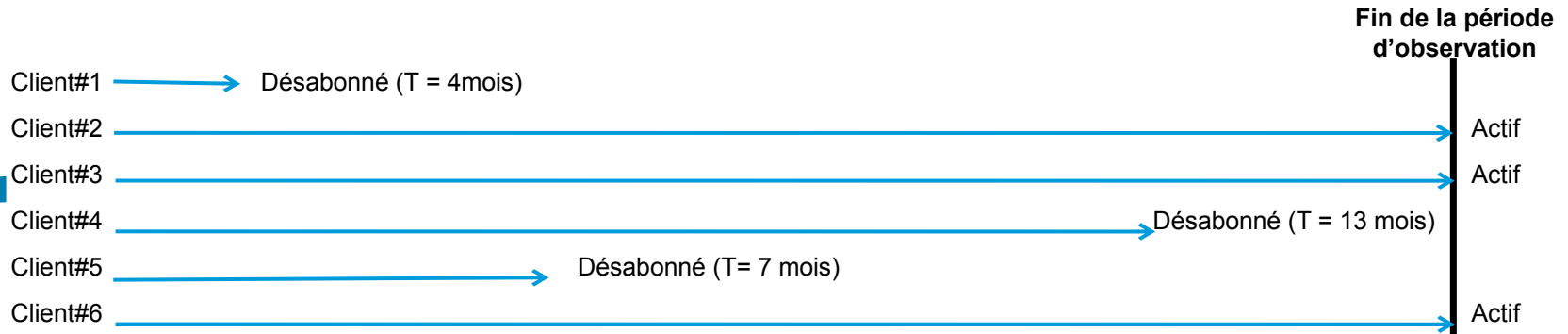
Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

On s'intéresse au temps avant qu'un événement survienne

- Exemple : une compagnie de télécommunication est intéressée au temps qu'un client demeure abonné au service de téléphonie mobile.



- Quelles sont les caractéristiques qui expliquent le mieux le temps avant le désabonnement?

Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Une entreprise de télécommunication veut évaluer la probabilité qu'un client quitte dans 1 mois, 2 mois, 3 mois, ..., 36 mois.

Analyse de regroupement

Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Analyse de survie

Marketing

- Temps avant le prochain achat d'un client.
- Temps avant qu'un client change de segment.

Gestion des risques

- Temps avant la faillite d'une entreprise (ou d'un particulier)

Ressources humaines

- Temps qu'un employé demeure au service de la compagnie.

Organisation du cours – sujets abordés

Sélection de variables et de modèles

Analyse de survie

Analyse factorielle

Une agence de voyage aimerait faire la promotion d'un forfait tout inclus dans le sud pour la famille. Pour bien cibler son offre, l'agence doit envoyer la promotion à ses clients qui ont des enfants. Toutefois, cette information n'est pas toujours disponible

Analyse de regroupement

Régression logistique

Données manquantes

HEC MONTRÉAL

Organisation du cours – sujets abordés

Données manquantes

Pourquoi avons-nous des données manquantes?

- Refus de répondre à une question d'un sondage.
- L'information sur les clients est incomplète.
- Le client ne détient pas d'information pour certaines caractéristiques (exemple : client qui n'a jamais eu de crédit).

Quel est l'impact des valeurs manquantes

- En régression (linéaire et logistique) et dans plusieurs autres techniques multivariées, si un sujet a au moins une valeur manquante parmi les variables utilisées dans le modèle, le sujet sera supprimé des analyses.
- Simplement ignorer les sujets avec des valeurs manquantes et faire l'analyse avec les autres sujets conduit généralement à des estimations biaisées et à de l'inférence invalide.

La solution étudiée...

- L'imputation multiple