Interpréter et explorer

Approfondissement en analyse de données avec R

Hubert Cadieux

Plan de l'atelier

9h à 12h

- Qui suis-je?
- R et RStudio

- Analyse textuelle
- PCA et clustering

12h à 15h

- Atelier libre avec des mentors de la CLESSN
- Exercices avec R

Qui suis-je?

- Maitrise en science politique
- Membre de la Chaire de leadership en enseignement des sciences sociales numériques (CLESSN)
- Membre du Centre d'analyse des politiques publiques (CAPP)
- Membre du Groupe de recherche en communication politique (GRCP)
- Membre du Centre d'étude sur la citoyenneté démocratique (CECD)

La CLESSN

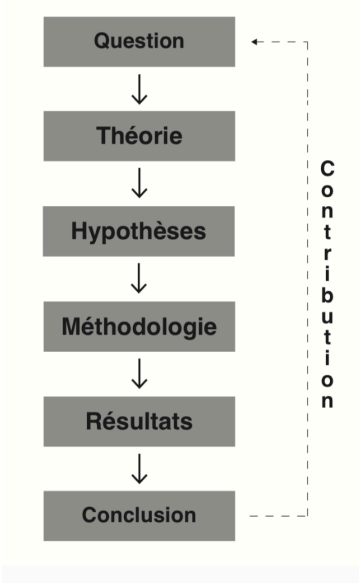
Chaire de leadership en enseignement des sciences sociales numériques (CLESSN)

- Titulaire: Yannick Dufresne
- Objectif: Développer des compétences en science des données pour les étudiants en sciences sociales à travers des projets de recherche et des formations

Pourquoi suivre cet atelier?

- Base solide pour la recherche académique

  Permet également de comprendre la pertinence de l'utilisation d'un logiciel de traitement de données et d'obtenir des outils pour dédramatiser la science des données



Pourquoi suivre cet atelier?

Comprendre le monde autour de nous à travers des données

Les données sont partout, mais pas nécessairement compréhensible dans leur forme « raw ». L'analyse de données permet de simplifier ces données pour comprendre et expliquer des phénomènes ou des comportements dans le monde qui nous entoure.

Pouvoir de prédiction: prédire comment certaines variables ou facteurs influencent d'autres variables ou événements.

Qu'est-ce que l'analyse de données?

> L'analyse de données vise à explorer, décrire et interpréter des informations recueillies sur les comportements humains, les relations sociales et les phénomènes sociétaux. **Ce processus peut inclure des données quantitatives (comme des données issus de sondages) et des données qualitatives (comme des entretiens ou des observations)**.

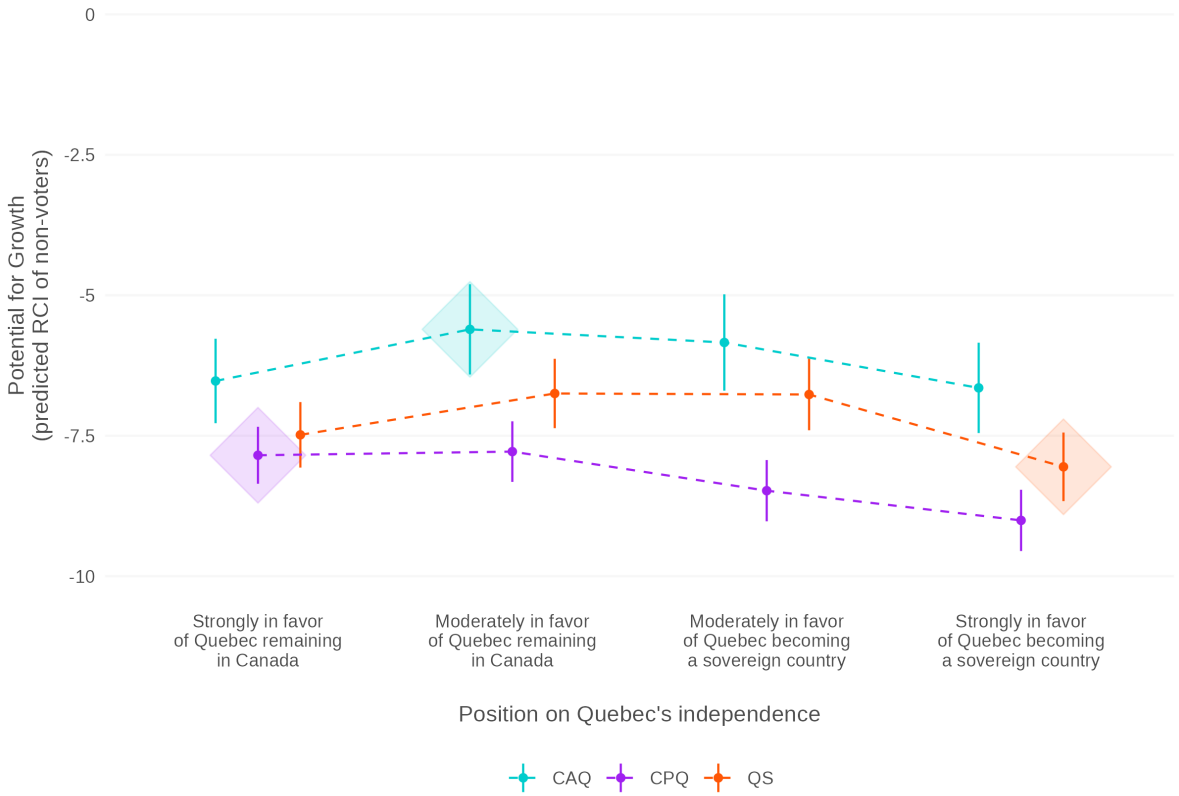
Principales étapes de l'analyse de données :

1. Collecte des données
2. Préparation des données
3. Analyse statistique

Les exercices de cet atelier

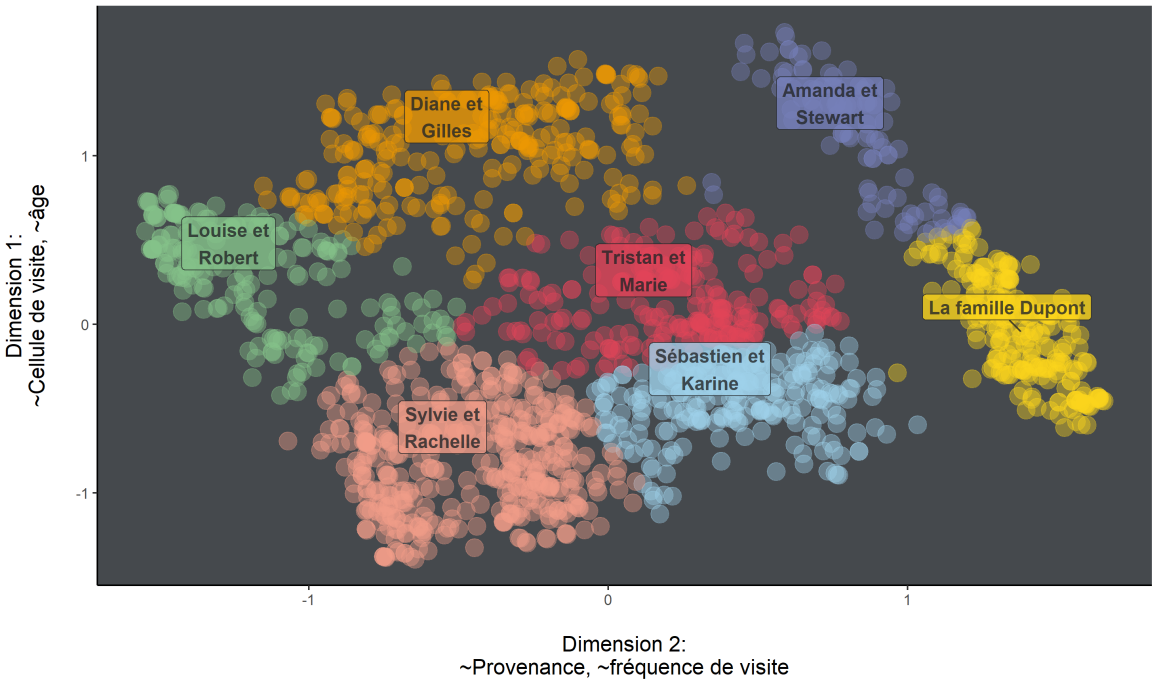
- Analyse textuelle
- PCA et clustering

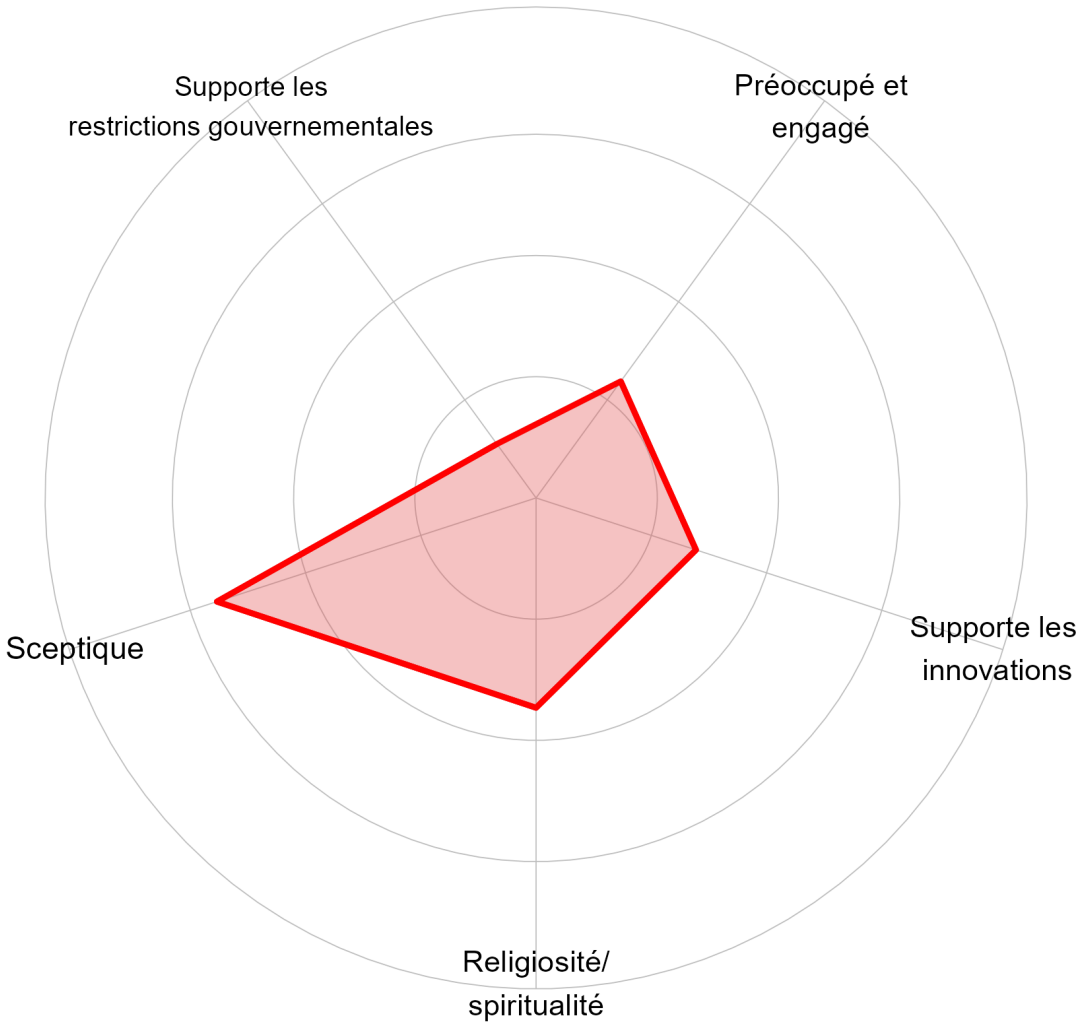
Voici quelques exemples de ce que vous pourrez faire avec R

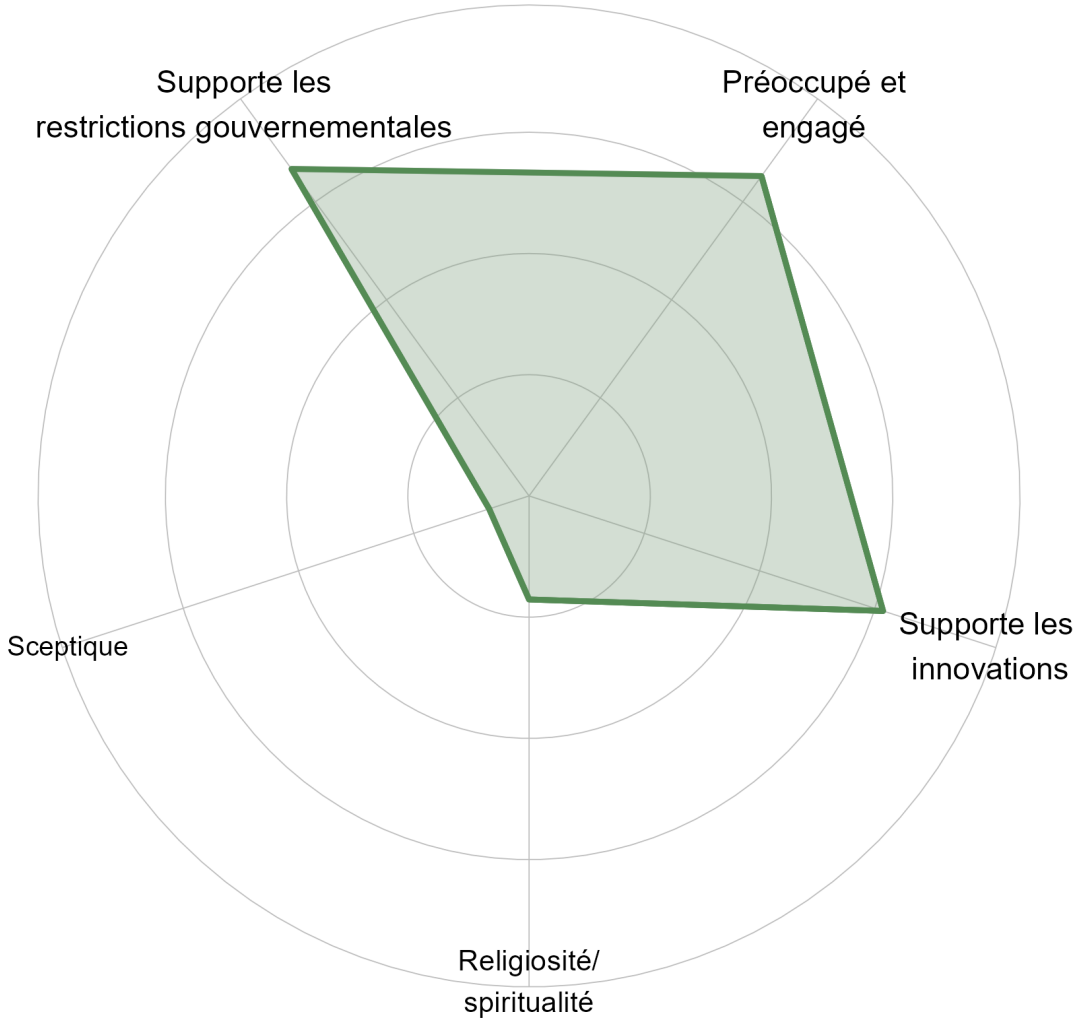


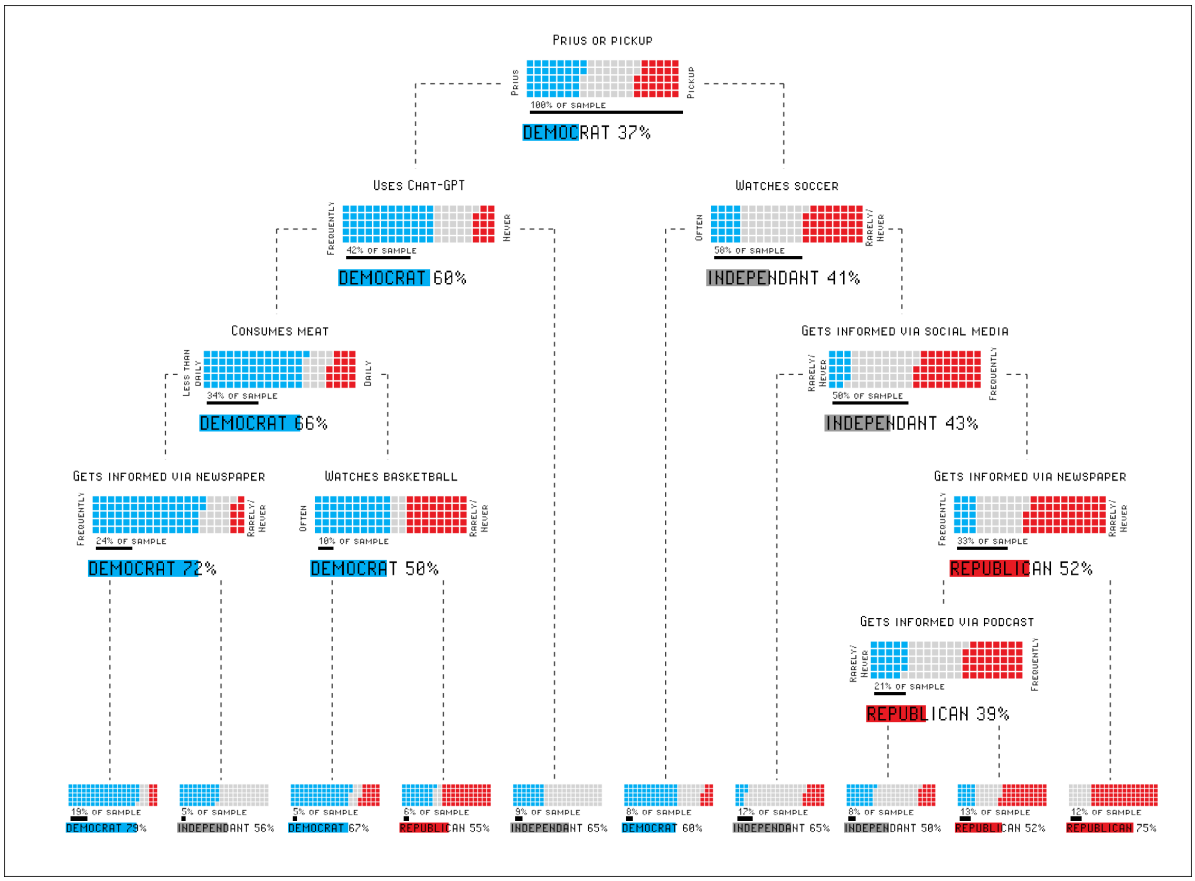
Diamonds indicate the parties' positions on the issue.
Vertical lines indicate the 95% confidence interval.

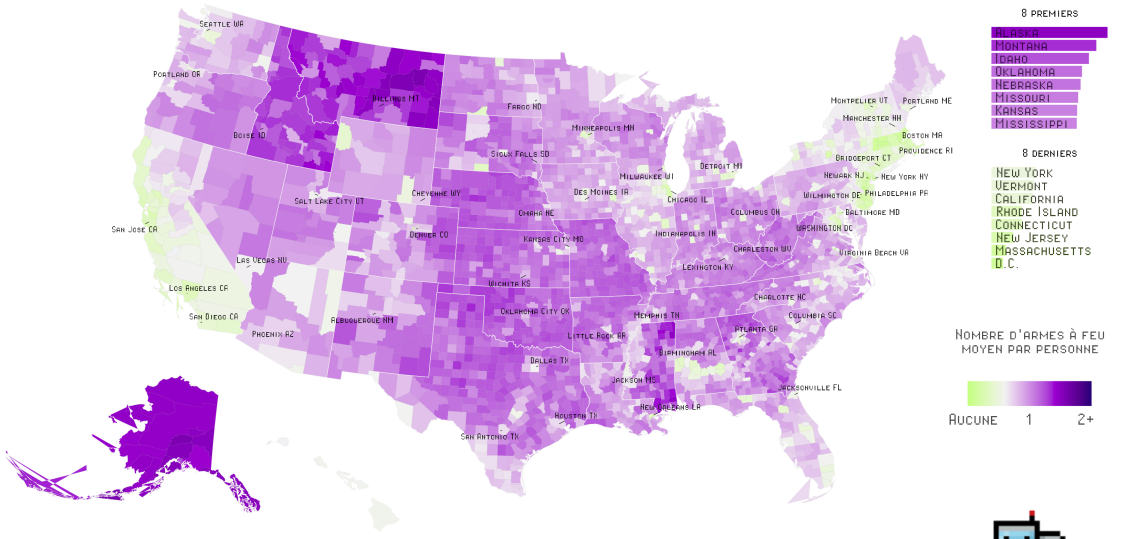
##


Diane et Gilles
Amanda et Stewart
Louise et Robert
Tristan et Marie
La famille Dupont
Sébastien et Karine
Sylvie et Rachelle
1
0
-1
-1
0
1
Dimension 1:
~Cellule de visite, ~âge
Dimension 2:
~Provenance, ~fréquence de visite


Supporte les
restrictions gouvernementales
Préoccupé et
engagé
Sceptique
Supporte les
innovations
Religiosité/
spiritualité


Supporte les
restrictions gouvernementales
Préoccupé et
engagé
Supporte les
innovations
Religiosité/
spiritualité
Sceptique

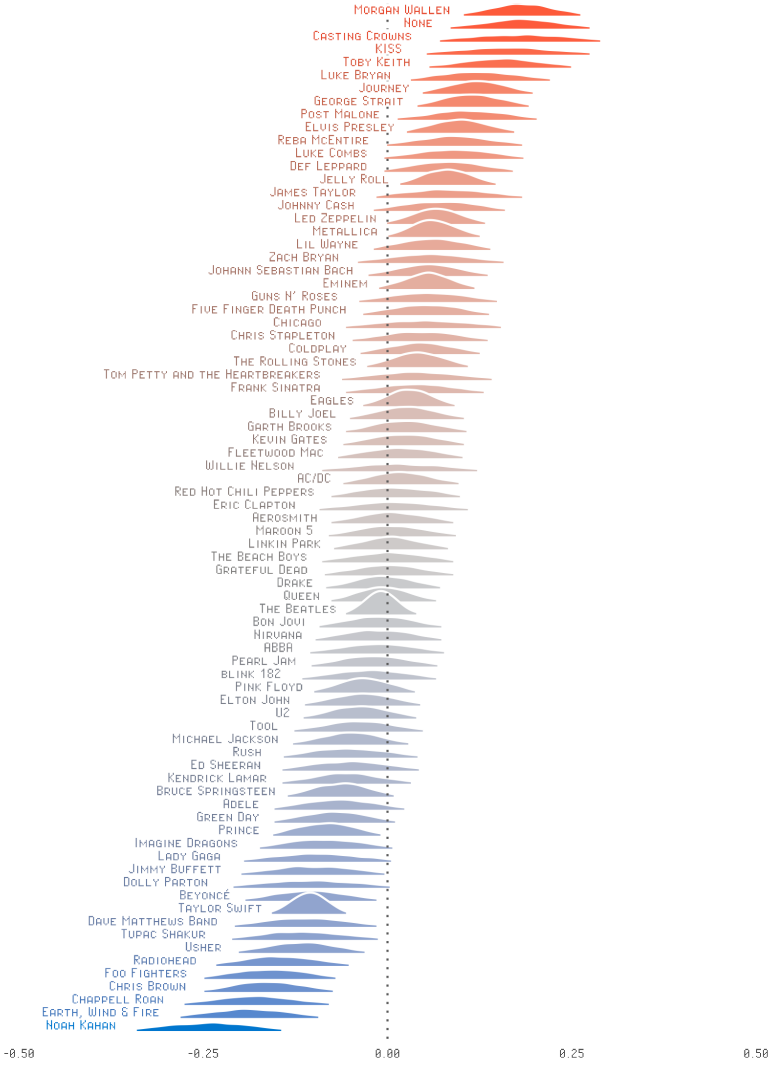

PRIUS OR PICKUP
PRIUS
PICKUP
100% OF SAMPLE
DEMOCRAT 37%
USES CHAT-GPT
FREQUENTLY
NEVER
42% OF SAMPLE
DEMOCRAT 60%
WATCHES SOCCER
OFTEN
RARELY/ NEVER
58% OF SAMPLE
INDEPENDANT 41%
CONSUMES MEAT
LESS THAN DAILY
DAILY
34% OF SAMPLE
DEMOCRAT 66%
GETS INFORMED VIA SOCIAL MEDIA
RARELY/ NEVER
FREQUENTLY
58% OF SAMPLE
INDEPENDANT 43%
GETS INFORMED VIA NEWSPAPER
FREQUENTLY
RARELY/ NEVER
24% OF SAMPLE
DEMOCRAT 72%
WATCHES BASKETBALL
OFTEN
RARELY/ NEVER
10% OF SAMPLE
DEMOCRAT 50%
GETS INFORMED VIA NEWSPAPER
FREQUENTLY
RARELY/ NEVER
33% OF SAMPLE
REPUBLICAN 52%
GETS INFORMED VIA PODCAST
RARELY/ NEVER
FREQUENTLY
21% OF SAMPLE
REPUBLICAN 39%
18% OF SAMPLE
DEMOCRAT 78%
5% OF SAMPLE
INDEPENDANT 56%
5% OF SAMPLE
DEMOCRAT 67%
6% OF SAMPLE
REPUBLICAN 55%
9% OF SAMPLE
INDEPENDANT 65%
8% OF SAMPLE
DEMOCRAT 60%
12% OF SAMPLE
INDEPENDANT 65%
8% OF SAMPLE
INDEPENDANT 58%
13% OF SAMPLE
REPUBLICAN 52%
12% OF SAMPLE
REPUBLICAN 75%



À partir des données de l'application Datagotchi (N = 14 329), le nombre d'armes à feu par personne a été prédit pour chaque comté en utilisant une régression multiniveau et une post-stratification (MRP). Le dégradé de couleur sur la carte reflète la prédiction du nombre d'armes d'un répondant agrégée pour chaque comté. Cette valeur est basée sur les prédictions du modèle de régression multiniveau.

DATAGOTCHI

L'effet de l'artiste préféré sur le vote aux élections américaines 2024



Intervalle de crédibilité du coefficient

Cet intervalle de crédibilité bayésien illustre la plage des valeurs probables du coefficient associé à chaque artiste calculée à partir de la distribution postérieure d'un modèle de régression bayésien. Le prior a été défini sur la base d'une enquête d'experts regroupant 12 étudiants et professeurs en science politique. Les coefficients sont mesurés sur l'échelle suivante : Strong Democrat, Soft Democrat, Independent, Soft Republican, Strong Republican, par incréments de 0.25. Un coefficient de 0.25 indique qu'un artiste progresse d'un niveau complet sur cette échelle, toutes choses étant égales par ailleurs, dans un modèle ajusté pour le statut socio-économique (SES) et certains indicateurs de style de vie.

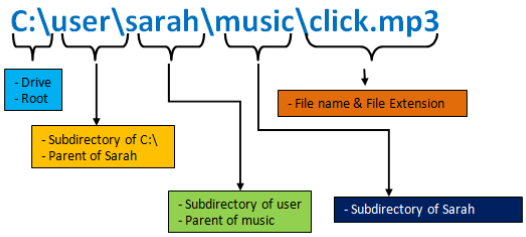
DATAGOTCHI

Les outils utilisés dans cet atelier

- R et RStudio
- C'est quoi la différence entre R et RStudio?
- R est le moteur, RStudio est l'interface
- Quarto
- GitHub

Concept important : Le chemin d'arborescence

- À tout moment vous devez savoir où vous êtes dans votre ordinateur pour pouvoir importer des données, exporter des graphiques ou mettre des fichiers en relation
- Votre R est toujours ouvert dans un dossier, et donc vous devez savoir où il est pour pouvoir importer des données
- La fonction `getwd()` dans R permet de savoir où vous êtes
- Un fichier quarto se réfère toujours à lui même



Concept important : Planifier avant de coder

La plus grosse erreur est de commencer à coder sans savoir ce que vous voulez faire

- Clarifer vos objectifs: Qu'est-ce que vous voulez faire?
  - Nettoyer des données?
  - Faire un graphique?

Les possibilités sont infinies, donc il est important de savoir où vous voulez aller

Concept important : Décomposer le problème

- Une fois que vous savez ce que vous voulez faire, il est important de décomposer le problème en petites étapes
- Un script R pour une seule tâche
  - Bien nommer vos scripts pour savoir ce qu'ils font
  - Exemples:
    - `nettoyage_donnees.R`
    - `graphique.R`
- Chaque script doit être clair et facile à comprendre
- Commenter votre code avec des `#`

Bonnes pratiques

- Nommer vos objets de façon explicite
- Commenter votre code
- Organiser votre code

Organiser votre répertoire et placer vos données

```
/MonProjet
  data/
      raw_data.csv
      cleaned_data.csv
  scripts/
      analyse.R
      visualization.R
  results/
      summary_statistics.csv
      plots/
          data_distribution.png
  docs/
      methodology.md
      references.bib
      project_presentation.qmd
  README.md
```

Apprendre plus de R

- swirl
- Datacamp
- R4DS (R for Data Science)
- Advanced R

Ressources

Quoi faire quand ça ne fonctionne pas?

- Google (Stack Overflow)
- ChatGPT / Claude
- Réessayer ChatGPT / Claude
- La documentation de R

Codons!