Licence de sociologie – Semestre 5
Année universitaire 2024-2025
Mathieu FERRY
mathieu.ferry@uvsq.fr

Consignes pour le projet collectif

Le devoir maison consiste à trouver une base de données en ligne, de poser une mini question de recherche à partir de ces données (à quel problème souhaitez-vous répondre ?), de recoder les données suivant les besoins de l'analyse, de proposer quelques statistiques descriptives univariées (moyenne, médiane, quantiles, tri à plat), bivariées (corrélation, nuage de points, corrélogramme, tri croisé, test du khi-2, V de Cramer…), puis de réaliser l'un des modèles d'analyse factorielle (Analyse en Composantes Principales, Analyse des Correspondances ou Analyse des Correspondances Multiples) sur cette même base et enfin de rédiger une analyse des résultats obtenus.

Trouver une base de données

De nombreux sites offrent des bases de données en accès libre. Ils contiennent des listes de bases sur divers sujets et il vous suffit d'en trouver une qui suscite votre intérêt. Notez que ces bases n'ont pas à avoir un intérêt sociologique. Il s'agit simplement de poser une question et y répondre en exploitant ces données. Voici une liste non-exhaustive de sites où vous pouvez trouver des bases de données :

- Données publiques françaises : https://www.data.gouv.fr/fr/datasets/
- Données publiques US : https://catalog.data.gov/dataset
- Ville de Paris : https://opendata.paris.fr/pages/home/
- SNCF : https://data.sncf.com/explore/?sort=modified
- Le site de l'INSEE : https://insee.fr/fr/statistiques?debut=0
- Les résultats des élections législatives/présidentielles françaises par commune (numérisation d'archives) : https://unehistoireduconflitpolitique.fr/telecharger.html

D'autres données sont disponibles. Si vous voulez vous intéresser à un autre pays, une autre ville ou organisation, en tapant son nom et Open Data, vous tomberez certainement sur des bases accessibles librement.

Les bases que vous trouverez pourront prendre plusieurs formes. Néanmoins, elles doivent généralement avoir la structure d'un tableau avec des individus-statistiques en ligne (comme des personnes, des organisations ou des pays) et des variables en colonne. Ce croisement constitue le fondement d'une base de données. Toutes les données présentes dans la base n'ont pas à vous intéresser, vous pourrez n'en sélectionner qu'une partie. Vous pouvez vous intéresser à des variables quantitatives ou à des variables qualitatives : les différentes méthodes mobilisées ce semestre vous permettront de mobiliser les unes aussi bien que les autres.

Enfin, les bases sont idéalement dans un format .csv. Mais ce n'est pas toujours le cas. Les packages foreign ou haven permettent généralement de lire les formats différents. Si vous ne

parvenez pas à lire votre base de données, n'hésitez pas à me poser la question à la fin d'un cours ou par mail.

Rédiger un script pour exploiter la base de données

Les scripts que nous utilisons à chaque séance sont autant de sources d'aide et d'inspiration pour votre devoir.

La rubrique Help ou Aide sur le panel de droite contient de nombreuses informations sur les fonctions que vous voulez utiliser. De plus, si vous rencontrez une erreur, vous pouvez souvent copier coller dans un moteur de recherche l'erreur qui apparait dans votre console et accéder à des forums qui cherchent où des gens se sont posés la même question que vous et l'ont résolue. Enfin, si vous rencontrez une difficulté vous pouvez toujours me poser la question, par mail ou à la fin d'un cours.

Lorsque vous souhaitez me poser une question par écrit : précisez bien l'erreur que vous rencontrez et quelle ligne de code l'a produite !

Le script R (structuré et commenté comme il se doit) devra accompagner le rendu.

Rédiger le devoir

Le devoir final devra faire au maximum 5 pages (figures et tableaux inclus, il est possible d'inclure une annexe en plus). Cette partie sera évidemment entièrement et clairement **rédigée**. Elle devra contenir les éléments suivants :

- une introduction avec une mini problématique
- une présentation des données et de comment elles vous permettent de répondre à la problématique
- une partie de statistique descriptive univariée / bivariée : les analyses statistiques doivent être commentées.
- une partie de statistique multidimensionnelle (analyse géométrique des données)
    - une justification du type d'analyse factorielle que vous utilisez
    - une justification du nombre d'axes retenus
    - une interprétation des résultats
- une conclusion récapitulative
- votre devoir doit contenir les graphiques de l'analyse factorielle (représentant les axes retenus) et peut contenir tout autre tableau ou graphique d'intérêt.

Le rendu final doit contenir : la base de données utilisée, le script R, un fichier PDF avec le devoir final. Envoyez-les par email à mon adresse (mathieu.ferry@uvsq.fr).

Pensez à soigner votre présentation.

*Date de rendu :au plus tard le **6 janvier à 9h**.*