# Projet de programmation statistique avec R

December 2023

## Objectifs

L'objectif de ce projet est d'appliquer les connaissances acquises pendant le cours de programmation statistique avec R. Chaque groupe doit s'approprier sa base de données, en faire des résumés statistiques et réaliser différents graphiques d'analyse descriptive. Les graphiques et les statistiques diffèrent en fonction des sujets, il faut choisir lesquels sont pertinents à réaliser.

## Déroulement

Le projet a lieu sur 3 séances : le 06/12, le 13/12 et le 19/12. **Le projet est à rendre le 19/12 à 14h au plus tard.** Chaque groupe passera à l'oral lors de la séance du 20/12 matin. Les modalités de l'oral restent à préciser, mais il consistera principalement en des questions sur votre code et vos choix d'analyse.

## Cahier des charges

Les projets doivent contenir, **au minimum**:

- l'import des données en R,
- la description des données : nombre de lignes, de colonnes, de valeurs manquantes,
- des résumés statistiques pertinents par variables : moyennes/médianes/écarts type, effectifs,
- des résumés statistiques croisés entre au moins deux variables,
- des résumés statistiques par populations/groupes,
- des graphiques descriptifs par variables,
- des graphiques contenant la description de plusieurs variables (options *facet* ou *facetgrid*)
- une fonction ou une boucle
- un élément de R avancé : une fonction complexe (par exemple une fonction permettant de réaliser des graphiques), une carte, une interface RShiny...

Les résumés statistiques et la manipulation de données sont à faire, dans la mesure du possible, avec la librairie *dplyr* vue en cours. Les graphiques sont à réaliser avec la librairie *ggplot2*. La notation sera adaptée au sujet : plus un sujet est facile à prendre en main, plus il faudra produire un rapport avancé. Le rendu est à effectuer en **RMardown** (ou en RShiny si vous décidez de faire une application). **Un soin particulier doit être apporté à la rédaction, la présentation et l'interprétation qui seront pris en compte dans la notation.**

## Sujets

L'ensemble des données est sur le Moodle du cours. La plupart des sujets ont déjà été étudiés et des graphiques sont disponibles, vous pouvez vous en inspirer. N'hésitez pas à faire quelques recherches pour contextualiser votre projet, sans y consacrer tout votre temps.

**Dinosaures : espèces, masse, époque**

La description des données est disponible sur ce lien. En complément, les subdivisions du Jurassique sont disponibles sur Wikipedia. Les deux onglets principaux du jeux de données sont les onglets *Full data* et *Mass estimates*. Les données manquantes peuvent être supprimées. Il n'est pas nécessaire de lire l'article scientifique, mais vous pouvez le parcourir pour vous inspirer des figures réalisées. Il n'est également pas nécessaire de décrire toutes les variables, vous pouvez faire une sélection.

**Automobile : caractéristiques du parc du Michigan**
La description des données est disponible sur ce lien.

**Echecs : ouvertures de parties de haut niveau**
La description des données est disponible sur ce lien.

**Ecologie : jour du dépassement**
La description des données est disponible sur ce lien. Un dictionnaire des termes principaux est accessible à ce lien. Les données exportées sur Moodle sont les données d'empreinte écologique et de biocapacité en gha par personne.

**Fortnite : performances et consommation de marijuana**
La description des données est disponible sur ce lien.

**Ligue 1 : résultats de 1999 à 2019**
La description des données est disponible sur ce lien

**Musique : chansons de 1950 à 2019**
La description des données est disponible sur ce lien

**NASA : analyse des météores et boules de feu**
La description des données est disponible sur ce lien

**Netflix : films et séries**
La description des données est disponible sur ce lien