Projet de programmation statistique avec R

2024-2025

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 de travail : le 17/12, le 18/12 et le 07/01.
Le rendu final est attendu le 12/01 à 22h au plus tard.
Les présentations orales auront lieu le **15/01** après-midi. Les modalités précises de l'oral seront communiquées ultérieurement, mais celui-ci portera principalement sur vos choix méthodologiques, vos analyses et votre code R.

Cahier des charges

Votre projet devra inclure **au minimum** les éléments suivants :

1. Gestion des données

- Importation des données dans R.
- Description générale des données (par exemple : nombre de lignes, colonnes, et valeurs manquantes).

2. Analyse descriptive

- Résumés statistiques pertinents pour chaque variable : moyennes, médianes, écarts types, effectifs, etc.
- Résumés croisés entre au moins deux variables.
- Analyses par sous-groupes ou populations spécifiques.

3. Visualisations graphiques

- Graphiques descriptifs pour des variables individuelles.
- Graphiques combinant plusieurs variables (par exemple, avec les options `facet` ou `facet_grid` de `ggplot2`).

4. Programmation avancée

- Une fonction ou une boucle pour automatiser une tâche.
- Un élément de programmation avancé, tel que :
  - Une fonction complexe (par exemple une fonction à plusieurs paramètres ou une fonction permettant de générer des graphiques).
  - Une carte interactive.
  - Une interface construite avec `RShiny`.

5. Outils recommandés

- Manipulations de données avec la librairie `dplyr`.
- Visualisations réalisées avec `ggplot2`.

6. Structure et rendu

Le rendu devra être organisé sous la forme d'un projet R contenant :

- **Des fichiers de code R** (exemples : un fichier pour la gestion des données et un autre pour les fonctions).
- **Un fichier Rmardown et le rapport HTML correspondant** ou une application `RShiny`.
  - Le rapport `RMarkdown` devra intégrer les fichiers R via la commande `source()`.

Critères d'évaluation

La notation sera adaptée en fonction de la complexité de votre sujet. Les sujets simples nécessiteront des analyses plus approfondies et un rapport plus complet.

Un soin particulier sera apporté à l'évaluation de :

- L'organisation du code R
- La qualité et la pertinence des analyses statistiques.
- La cohérence et la lisibilité des graphiques.
- La clarté et la structuration du rapport ou de l'application.

Sujets

L'ensemble des données est sur le Moodle du cours. Certains 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, mais sans y passer trop de temps.

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

2. Cancer du sein : caractéristiques des tumeurs.
La description des données est disponible sur ce lien.

3. Cryptomonnaies : données de la bourse.
Le fichier `3_code_data_crypto.R` permet de télécharger les données dans R. Pour plus d'informations, allez voir l'aide du package quantmod ou les différentes utilisations de quantmod sur Internet.

4. NASA : chutes de météorites.
La description des données est disponible sur ce lien.

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

6. Transports : validations dans les gares d'Île-de-France.
La description des données est disponible sur ce lien.

7. Pokémons : caractéristiques de pokemons.
La description des données est disponible sur ce lien. De manière facultative, des données complémentaires sont disponibles sur ce lien et ce lien.

8. LoL : statistiques des joueurs du LoL World Championship 2024.

La description des données est disponible sur ce lien.

9. Réseaux sociaux : analyse de posts.
La description des données est disponible sur ce lien.

10. NBA : statistiques des joueurs.
La description des données est disponible sur ce lien (cliquer sur **Glossary**).

11. Ligue 1 : statistiques des joueurs.
Les données proviennent du site MPG Stats. Un glossaire est disponible sur ce lien.

12. Tennis : analyse des joueurs et/ou matchs de l'ATP Tour.
Les données proviennent de ce site. Sur Moodle, 3 fichiers de données sont disponibles :

- `12_atp_matches_2024.csv` : données sur les matchs
- `12_atp_players.csv` : données sur les joueurs
- `12_atp_ranking_current.csv` : classement des joueurs.

Vous pouvez choisir de travailler sur les matchs ou sur les joueurs (ou les deux si vous vous en sentez capable). Si vous décidez de vous focaliser sur les joueurs, il vous est vivement conseillé de joindre les fichiers `12_atp_players.csv` et `12_atp ranking_current.csv`. Le fichier `12_matches_data_dictionary.txt` disponible sur Moodle est un dictionnaire de données.

13. Écologie : Jour du dépassement
Les données proviennent de ce site. Un dictionnaire des termes principaux est accessible à ce lien. Les données exportées sur Moodle incluent l'empreinte écologique et la biocapacité en gha par personne, pour chacun des continents. Vous pouvez choisir de travailler sur un ou plusieurs continents.