FAS1002 – Initiation à la programmation en sciences sociales

Travail Final - Automne 2022

Samuel Guay

Date limite pour la remise du travail: 20 décembre 2022 - 23:59 - Anywhere On Earth (AOE)
Disponibilité: Prise de rendez-vous au rencontre.samuelguay.ca
Questions, suggestions et commentaires généraux: Discussion sur Ed

---

Introduction

Pour le travail final, votre mission principale est de **produire un rapport analytique de qualité professionnelle consultable à partir du Web.** Pour ce faire, vous devrez **manipuler**, **analyser** et **mettre en relation des données** en provenance de diverses bases de données précieusement choisies juste pour vous si vous n'avez pas. Vous devrez également appliquer les principes de la programmation littéraire, c'est-à-dire d'écrire de combiner des phrases explicatives entre vos blocs de code. Tout au long de la session, vous avez appris à utiliser divers outils qui vous permettront de réaliser votre rapport en suivant les bonnes méthodes de travail. Vous pouvez penser à R, RStudio, tous les packages que vous voulez, Git, GitHub, etc. Je vous répète qu'il est important de comprendre le code que vous utilisez, car le danger de *complexifier (over-engineer)* votre code vous guette sans même que vous ne le réalisiez! Tout au long du travail, vous devrez mettre à profit l'expertise que vous détenez en appliquant les concepts vus au courant de la session.

Évidemment, vos documents devront être reproductibles puisqu'ils seront la fondation même de votre site web qui sera hébergé sur GitHub. Un exemple d'un répertoire fonctionnel et disponible sur GitHub a été créé pour vous avec des indications des modifications possibles. Au cours de la session, nous avons vu qu'il était relativement facile de construire un site web sans connaissance en HTML, CSS ou Javascript.

Les données que j'ai choisies touchent un sujet d'actualité et chaque personne est libre de mettre en relation les variables qui l'intéressent, à quelques contraintes près qui vous sont décrites dans la section Consignes. J'ai choisi ces données, car 1- plusieurs analyses sont disponibles sur Internet, pouvant vous inspirer; et 2- en fonction de votre motivation et de votre rapport, cela pourrait devenir votre premier item à montrer dans votre portfolio en science des données! **Si vous voulez travailler sur vos propres données, veuillez m'écrire afin que nous les regardions ensemble** pour m'assurer qu'elles correspondent à un standard minimum de correction.

Base de données

Différentes sources de bases de données ont été choisies afin d'évaluer vos habiletés à manipuler, nettoyer et *tidy* des données, sans grande surprise.

La première source des données est Our World in Data qui offre librement "*Research and data to make progress against the world's largest problems.*" Leur site offre également de magnifiques visualisations interactives qui pourraient vous inspirer. Les deux ensembles de données sélectionnés pour le travail sont celui de **CO_2 and Greenhouse Gas Emissions** et celui **Energy** qui sont fréquemment mis à jour. De plus, leurs données sont disponibles directement sur GitHub: **Répertoire - CO_2 and Greenhouse Gas Emissions** et celui **Répertoire - Energy**. Leurs répertoires sont extrêmement bien construits, mentionnant ce que contient chaque fichier et ce que représente chaque variable. Il y a des données pour pratiquement tous les pays s'étendant sur plusieurs années. Prenez des notes sur comment ils ont écrit leur README.md ;).

La deuxième source des données est Gapminder, un organisme à but non lucratif indépendant qui lutte contre les idées fausses touchant le monde. Plus précisément, leur mission est de *"fight devastating ignorance with a fact-based worldview everyone can understand."* Les données librement offertes par Gapminder touchent une variété de sujets comme l'économie, l'environnement, la santé, les populations, l'éducation et beaucoup plus encore, similairement à Our World in Data! Dans le cadre du travail, vous aurez à travailler avec le jeu de données **Life Expectancy at Birth** et son **guide disponible dans Google Sheets**.

Si jamais vous regardez la liste des données disponibles sur Gapminder et que vous aimeriez utiliser d'autres indicateurs (indice de démocratie, corruption, etc.) à la place de ceux mentionnés précédemment, simplement me faire part de vous plan et je vous dirai si cela est trop ambitieux ou non. C'est rarement trop ambitieux ;). Enfin, si les données du Registre des entreprises du Québec vous intéressent vraiment, vous pourriez aussi les utiliser. Je ne veux pas restreindre vos ambitions!

Dans le cadre du travail, vous devrez d'abord incorporer au moins deux jeux de données ensemble.

Consignes générales

- [ ] D'abord, vous devez installer Quarto si ce n'est pas déjà fait. À ce stade, vous devriez être en mesure de l'installer par vous-mêmes.
    - Si vous rencontrez un bogue durant l'installation et que vos recherches sur le Web demeurent infructueuses, ouvrez un issue dans votre répertoire sur GitHub en me notifiant (@SamGuay).

- [ ] Un répertoire avec un exemple a été créé sur mesure pour vous et il se trouve au https://github.com/FAS1002/FAS1002_projet-final. Pour voir ce à quoi ressemble le rapport, vous pouvez regarder le site web associé au répertoire juste ici.
    - Les instructions pour pouvoir utiliser votre propre version de ce répertoire sont incluses dans le README.md. Il s'agit d'une façon typique d'inclure des instructions.
    - Vous **devez utiliser un *fork*** de ce répertoire pour effectuer votre travail et communiquer avec GitHub.
    - Tel que mentionné dans le README.md, presque *tous les fichiers seront amenés à être modifiés; [...] faudra créer quelques dossiers supplémentaires au besoin*, ainsi, vous devrez structurer votre travail afin d'être en mesure d'utiliser vos objets dans les différents fichiers .qmd, soit en mettant votre code commun dans des scripts .R ou toutes autres stratégies (par ex., sauvegarder des données nettoyées temporaires, etc.).

- [ ] Vous devrez travailler avec Git tout au long de votre projet en vous assurant de *commit* stratégiquement l'avancement de vos travaux. De plus, cela vous servira de sauvegarde lorsque vous faites un *push* sur GitHub.
    - Dans un travail d'une telle envergure, que vous devrez faire au minimum une dizaine de commits.
        - Petite aide pour la *rédaction* de messages pertinents pour vos commits.
    - Pensez à faire des commits de façon stratégique en fonction des différentes tâches à accomplir.
        - Exemple: Modifier les paramètres du rapport dans `_quarto.yml` vs ajouter du code dans un document d'analyse vs de nettoyage, etc.
    - Si vous voulez approfondir vos habiletés avec Git, vous pouvez vous amuser à créer différentes branches selon les fonctions que vous voulez ajouter et faire des *Pull requests* par la suite. Pour vous aider: Guide GitHub flow. À ce stade, je n'enlèverai pas de points si vous ne le faites pas, mais je pourrais très bien en ajouter si j'observe le processus de création de *branches*, pull requests et *merge*.

- [ ] Vous devrez remettre un répertoire Git hébergé sur GitHub avec un rapport fonctionnel dans n'importe quel fureteur sur Internet qui sera accessible par un lien similaire à ***https://votre-username.github.io/nom-de-votre-projet***. Vous pouvez changer le nom que je lui ai donné (c-à-d., *FAS1002_projet-final*) par un autre si vous jugez pertinent de le changer.

- [ ] Vous devrez démontrer vos connaissances à utiliser du Markdown et des citations dans les sections de texte de votre rapport de manière adéquate:
    - Pensez à vos titres de différentes grosseurs pour la logique de votre rapport, mots mis en évidence, listes à puces ou numéros, liens, images, etc.
    - Pour les citations, il vous faudra les citer adéquatement en ajoutant les citations pertinentes dans un ou des fichiers *`.bib`*. Les citations devront être ajoutées fichier de type bibtex (`.bibtext`) ou BibLaTeX (`.bib`). Pour le moment, vous sert d'exemple. Pour générer vos citations au format BibLaTeX, vous pouvez utilisez Zotero qui est gratuite et open source ou les trouver directement sur Internet.
        - Je vous ai fourni en exemple la référence pour un des ensembles de données de Our World in Data dans le fichier `references.bib`.

        - La documentation de Quarto à propos de comment citer est très bien construite: Authoring > Citations & Footnotes.
        - N'oubliez surtout pas de citer d'où proviennent toutes les données et les principaux packages utilisés!
            - Voir cet excellent article de ROpenSci au besoin.

- [ ] Votre fichier README.md, qui se trouve déjà à la racine de votre répertoire, devra décrire davantage ce qui se trouve dans votre répertoire au lieu de la description générique qui s'y trouve.
    - Conseil: Ce fichier risque fort probablement d'être le dernier que vous voudrez modifier, après avoir conçu votre page d'accueil du rapport.

- [ ] Ajustez la licence selon vos besoins, la licence actuelle qui s'applique est une licence Creative Commons CC BY SA.

- [ ] Si vous avez des questions, vous devrez ouvrir un ou des issues sur votre répertoire et m'identifier (@SamGuay) pour que je puisse vous répondre.
    - En procédant ainsi, il se pourrait même que je vous fasse une *pull request* qui vous aidera à avancer!

Consignes spécifiques

- [ ] Pour l'**importation** des données, vous aurez à les télécharger de plusieurs sources (GitHub, Google Sheet ou Excel). Il y a plusieurs stratégies possibles, à vous de jouer!
    - Cependant, vous devrez faire plus que de seulement importer les données. Puisque tout peut disparaître sur Internet ou être mis à jour sans préavis, vous devrez également télécharger les données. Par contre, il n'est pas optimal de procéder aux téléchargements à chaque fois que nous roulons le code. Les données brutes devront être téléchargées dans le dossier `data/raw/` en respectant les conditions:
        - Ainsi, pour les données qui proviennent de Our World in Data, vous devrez développer du code pour télécharger et sauvegarder les fichiers avec la date à laquelle le téléchargement a lieu. La fréquence du **téléchargement** devrait être **quotidienne** puisque les données sont mises à jour **fréquemment**. Pour vous simplifier la tâche, pensez à programmer également la suppression de l'ancien fichier une fois que le nouveau est téléchargé. En d'autres mots, ces données ne devraient être téléchargées qu'une seule fois par jour lorsque votre rapport est produit.
        - Pour les **données des autres sources**, le principe est le même, mais la fréquence du **téléchargement** devra être **mensuelle**, donc à chaque mois seulement.
        - ex: `data/raw/owid-energy-codebook_<yyyy>-<mm>-<jj>.csv`
    - Assurez-vous de lire les données dans un tibble et de transformer les types de variables afin qu'ils soient adéquats (par ex., transformer en facteurs les variables nominales avec les bons niveaux des facteurs, les dates en date, etc.).
    - Vous pouvez faire du nettoyage de données lors de l'importation ou plus tard de la processus, libre à vous selon vos préférences et votre projet.

- [ ] Pour la **manipulation** des données, vous devrez faire les procédures nécessaires sur vos tibbles pour les combiner, les séparer ou créer des sous-ensembles selon vos besoins.
    - Plusieurs packages que nous avons vus au cours de la session vous seront très certainement utiles. De plus, je ne vous limite à aucun ensemble de packages, vous pouvez utiliser tout ce que vous voulez.
    - Afin de tester vos habiletés, vous devrez un des ensembles de données d'Our World in Data vous aurez à créer une nouvelle variable qui catégorisera les différents pays en fonction des continents l'ensemble des données qui sera les continents qui permettra de regrouper les différents pays en fonction de cette nouvelle variables et d'observer des statistiques. À vous de choisir si vous voulez séparer l'Amérique en sous-groupes.

- [ ] Pour l'**exploration** et l'**analyse** des données, vous aurez à créer au moins une page descriptive des données:

- Vous pouvez, voire devriez, inclure plus d'éléments que ceux demandés ci-bas afin de produire un rapport plus exhaustif:
- Présenter un ou des beaux tableaux avec plusieurs variables d'intérêts en montrant les statistiques descriptives classiques (moyennes, écart-types, données manquantes, etc.). À vous de déterminer comment il est mieux de présenter les données globales.
    - Conseil: vous avez vu ou lu comment calculer plusieurs statistiques en même temps.
    - Présenter les mêmes tableaux en regroupant les données par continents.
- Une variable d'intérêt que vous devrez calculer est le temps écoulé en jours entre la production (*rendering*) de votre rapport et la *première année* que des données ont été récoltées pour chacun des pays qui se trouvent dans le jeu de données de Life Expectancy at Birth.
- Vous pouvez également inclure des graphiques descriptifs représentant la distribution des données, etc.
- Conseil: Il s'agit d'une bonne page pour rendre accessible le téléchargement des données.

☐ Pour l'**analyse** des données de vos données, vous devrez en effectuer deux et décrire les motivations de ces analyses et vos résultats. Les analyses ne sont pas obligées d'être hyper sophistiquées, mais le processus et l'interprétation des résultats doivent se retrouver dans votre document. Il y a énormément de possibilités dans toutes les variables que vous avez. Comme il ne s'agit pas d'un cours de statistiques, la correction sera moins sévère à ce niveau. Tâchez tout de même de démontrer une rigueur peu importe votre niveau de connaissances! Tentez d'expliquer vos résultats en décrivant ce qui peut les influencer, les forces et faiblesses, etc.

- Conseil: Partagez avec moi (et les autres si vous voulez) le plan de votre analyse dès que vous aurez une idée, ça évitera des surprises.
- Conseil 2: Vous pouvez *brainstormer* à plusieurs, vivement le travail collaboratif qui permet de faire profiter l'expertise respective de chacun au groupe!

☐ Pour la **visualisation** des données, il y a des centaines de possibilités qui s'offrent à vous.

- Cette partie est normalement la plus excitante pour les gens qui consultent votre rapport. Vous pouvez montrer des graphiques classiques, des cartes, etc. Vos graphiques peuvent être statiques ou dynamiques. Profitez de cette occasion pour générer de multiples graphiques sur une ou plusieurs variables afin de toucher concrètement au monde de la visualisation. Même s'ils ne sont pas parfaits, le but est que vous touchiez à différents types. Le seul ennemi est le thème de base de ggplot2, s.v.p., pas de fond gris atroce dans vos graphiques!

☐ Quant à l'**exportation** des données:

- Les gens devraient être en mesure de télécharger votre ou vos ensembles de données traitées (lire ici *tidy)* à partir de votre site web. Ainsi, il faudra exporter vos données dans un format standard et ouvert et les rendre disponibles sur le site de votre rapport. Vous avez appris les chemins relatifs pour une raison.
    - Ces données devraient être sauvegardées dans `data/processed` par exemple, tout comme vos données dites intermédiaires.
    - Vous pourriez également utiliser DT avec l'extension Buttons qui offre la possibilité de télécharger les données dans le format désiré (.csv, .xlsx, ...).
- Si vous avez plusieurs fichiers, je garderai l'oeil attentif afin de vérifier si vous arrivez à les combiner dans un dossier compressé! Encore une fois, ça n'enlèvera pas de points si vous ne le faites pas, mais je pourrais en ajouter si oui.

☐ Vous devez créer au moins une fonction dans votre travail et l'utiliser. À vous de choisir le but de la fonction. Assurez-vous de bien décrire en commentaire à quoi elle sert. Cette fonction devra se trouver dans un fichier qui se nomme `helper.R` dans le dossier `R` (`R/helper.R`). Vous devrez donc l'importer dans l'environnement au moment opportun..

☐ Je m'attends à ce que le site soit à l'image de votre personnalité, c'est-à-dire au minimum avec des couleurs différentes du thème original.

Conseils

- Si vous sentez que vous vous enfoncez dans un trou noir depuis des heures, n'hésitez surtout pas à me contacter, surtout pour des trucs qui ne concernent pas R directement.

- N'hésitez surtout pas à modifier les couleurs, les écritures, etc. Il est très facile de lui donner un look un plus *professionnel* en changeant la couleur du menu et du bas pour une couleur statique.

- Tous vos fichiers utilisés devront être répertoriés sur GitHub, incluant les données. Ainsi, il se peut que vos scripts ne soient pas tous visibles sur la face publique de votre rapport (votre site web), mais je dois avoir accès pour voir ce que vous avez fait!
    - Si vous modifiez les données, la trace des modifications doit être dans un script ou votre rapport.