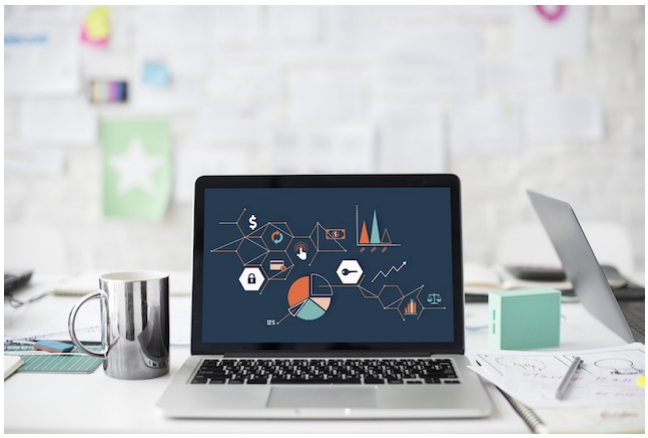
Montrez que vous êtes un scientique de données

En un mot



Choisissez un jeu de données, n’importe quel jeu de données…

…et faire quelque chose avec. C’est votre projet final en un mot. Plus de détails ci-dessous.

C’est long, mais veuillez lire au complet

Le projet final pour ce cours consistera en une analyse sur un ensemble de données de votre choix. L’ensemble de données peut déjà exister, ou vous pouvez collecter vos propres données à l’aide d’une enquête ou en menant une expérience. Vous pouvez choisir les données en fonction de vos intérêts ou en fonction du travail dans d’autres cours ou projets de recherche. L’objectif de ce projet est que vous démontriez votre maîtrise des techniques que nous avons couvertes dans ce cours (et au-delà, si vous le souhaitez) et que vous les appliquiez à un nouvel ensemble de données de manière significative.

Le but n’est pas de faire une analyse exhaustive des données, c’est-à-dire de ne pas calculer toutes les statistiques et procédures que vous avez apprises pour chaque variable, mais plutôt de me faire savoir que vous êtes capable de poser des questions significatives et d’y répondre avec les résultats de l’analyse des données, que vous êtes compétent dans l’utilisation de R, et que vous êtes compétent dans l’interprétation et la présentation des résultats. Concentrez-vous sur les méthodes qui vous aident à commencer à répondre à vos questions de recherche. Vous n’êtes pas obligé d’appliquer toutes les procédures statistiques que nous avons apprises (et vous pouvez utiliser des techniques que nous n’avons pas officiellement couvertes en classe, si vous vous sentez aventureux). Critiquez également vos propres méthodes et faites des suggestions pour améliorer votre analyse. Les questions relatives à la fiabilité et à la validité de vos données et à la pertinence de l’analyse statistique doivent être abordées ici.

Le projet est très ouvert. Vous devez créer une sorte de visualisation convaincante de ces données dans R. Il n’y a pas de limite sur les outils ou les packages que vous pouvez utiliser, mais il est nécessaire de s’en tenir aux packages que nous avons appris en classe (`tidyverse`). Vous n’avez pas besoin de visualiser toutes les données à la fois. Une seule visualisation de haute qualité recevra une note beaucoup plus élevée qu’un grand nombre de visualisations de mauvaise qualité. Faites également attention à votre présentation. La netteté, la cohérence et la clarté compteront. Toutes les analyses doivent être effectuées dans RStudio, en utilisant R.

Données

Pour que vous ayez les meilleures chances de succès avec ce projet, il est important que vous choisissiez un ensemble de données gérable. Cela signifie que les données doivent être facilement accessibles et suffisamment volumineuses pour que de multiples relations puissent être explorées. En tant que tel, votre ensemble de données doit avoir au moins 50 observations et entre 10 et 20 variables (des exceptions peuvent être faites mais vous devez d'abord me parler). Les variables de l'ensemble de données doivent inclure des variables catégorielles, des variables numériques discrètes et des variables numériques continues.

Si vous utilisez un ensemble de données qui vient dans un format que nous n'avons pas rencontré en classe, assurez-vous que vous pouvez le charger dans R car cela peut être délicat selon la source. Si vous rencontrez des difficultés, demandez de l'aide avant qu'il ne soit trop tard.

Remarque sur la réutilisation des ensembles de données de la classe : Ne réutilisez pas les ensembles de données utilisés dans les exemples, les devoirs ou les laboratoires de la classe.

Vous trouverez ci-dessous une liste de référentiels de données qu'il pourrait être intéressant de parcourir. Vous n'êtes pas limité à ces ressources et, en fait, vous êtes encouragé à vous aventurer au-delà. Mais vous pourriez y trouver quelque chose d'intéressant :

- TidyTuesday (https://github.com/rfordatascience/tidytuesday)
- NHS Scotland Open Data (https://www.opendata.nhs.scot/) - Edinburgh Open Data (https://edinburghopendata.info/)
- Open access to Scotland's official statistics (https://statistics.gov.scot/home)
- Bikeshare data portal (https://www.bikeshare.com/data/) - UK Gov Data (https://data.gov.uk/)
- Kaggle datasets (https://www.kaggle.com/datasets) - OpenIntro datasets (http://openintrostat.github.io/openintro/)
- Awesome public datasets (https://github.com/awesomedata/awesome-public-datasets)
- Youth Risk Behavior Surveillance System (YRBSS) (https://chronicdata.cdc.gov/Youth-Risk-Behaviors/DASH-Youth-Risk-Behavior-Surveillance-System-YRBSS/q6p7-56au)
- PRISM Data Archive Project (https://www.icpsr.umich.edu/icpsrweb/content/ICPSR/fenway.html)
- Harvard Dataverse (https://dataverse.harvard.edu/)
- Google Dataset Search (https://datasetsearch.research.google.com/)
- Canadien govt open data (https://open.canada.ca/en/open-data)
- Github Public Dataset (https://github.com/awesomedata/awesome-public-datasets)
- Québec données (https://www.donneesquebec.ca/)

Livrables

- Proposition : due le 10 novembre 2023 à 23h59
- Présentation - due le 15 décembre 2023 à 23h59
- Résumé exécutif - dû le 15 décembre 2023 à 23h59

Proposition

Il s'agit d'une ébauche de la section d'introduction de votre projet ainsi que d'un plan d'analyse des données et de votre jeu de données.

- **Section 1 - Introduction :** L'introduction doit présenter votre

  question de recherche et vos données (d'où elles proviennent, comment elles ont été collectées,

  quels sont les sujets, quelles sont les variables, etc.).

- **Section 2 - Données :** Placez vos données dans le dossier `/data` et ajoutez les dimensions et la source de vos données au README de ce dossier. Executez ensuite la sortie du README Executez ensuite la sortie de `glimpse()` ou `skim()` de votre bloc de donnée pour avoir un appercu.

- **Section 3 - Plan d'analyse des données :**

  - Les variables de résultat (réponse, Y) et prédictives (explicatives, X) que vous utiliserez pour répondre à votre question.
  - Les groupes de comparaison que vous utiliserez, le cas échéant.
  - Analyse de données exploratoire très préliminaire, comprenant des statistiques et des visualisations récapitulatives, ainsi que des explications sur la manière dont elles vous aident à en savoir plus sur vos données. (Vous pouvez en ajouter plus tard au fur et à mesure que vous travaillez sur votre projet.)
  - La ou les méthodes qui, selon vous, seront utiles pour répondre à vos questions. (Vous pouvez les mettre à jour ultérieurement au fur et à mesure que vous travaillez sur votre projet.)
  - Quels résultats de ces méthodes statistiques spécifiques sont nécessaires pour vous aider à répondre à votre question?

Chaque section ne doit pas dépasser 1 page (hors figures). Vous pouvez vérifier un aperçu avant impression pour confirmer la longueur.

Le système de notation pour la proposition de projet est le suivant.

Total	**10 points**
Données	3 points
Proposition	5 points
Workflow, organisation, qualité du code	1 point
Travail d'équipe	1 point

Présentation

5 minutes maximum, et chaque membre de l'équipe doit dire quelque chose de substantiel. Vous devrez préparer un enregistrement vidéo de 5min maximum avec votre présentation et la soumettre.

Préparez un diaporama en utilisant le modèle de votre repo Ce modèle utilise un package appelé `xaringan` et vous permet de créer des diapositives de présentation à l'aide de la syntaxe R Markdown. Il n'y a pas de limite au nombre de diapositives que vous pouvez utiliser, juste une limite de temps (5 minutes au total). Chaque membre de l'équipe devrait avoir la chance de prendre la parole pendant la présentation. Votre présentation ne doit pas simplement être un compte-rendu de tout ce que vous avez essayé ("puis nous avons fait ceci, puis nous avons fait cela, etc."), mais elle doit plutôt indiquer les choix que vous avez faits, pourquoi et ce que vous avez trouvé.

Avant de finaliser votre présentation, assurez-vous que vos blocs sont désactivés avec `echo = FALSE`.

A la fin du semestre, vous regarderez les présentations d'autres équipes et fournirez des commentaires sous la forme d'évaluations par les pairs.

Le barème de notation de la présentation est le suivant :

Total	**50 pts**
Gestion du temps : L'équipe s'est-elle bien répartie le temps ou a-t-elle été interrompue au fil du temps ?	4 pts
Contenu : la question de recherche est-elle bien conçue et les données utilisées sont-elles pertinentes pour la question de recherche ?	5 pts
Professionnalisme : dans quelle mesure l'équipe s'est-elle bien présentée ? La présentation semble-t-elle bien pratiquée ? Est-ce que tout le monde a eu la chance de dire quelque chose de significatif sur le projet ?	5 pts
Travail d'équipe : L'équipe a-t-elle présenté une histoire unifiée, ou est-ce que cela ressemblait à des morceaux de travail indépendantes assemblées ?	6 pts
Contenu : L'équipe a-t-elle utilisé avec précision les procédures statistiques appropriées et les interprétations des résultats ?	10 pts
Créativité et pensée critique : Le projet est-il mûrement réfléchi ? Les limites sont-elles soigneusement prises en compte ? Semble-t-il que du temps et des efforts ont été consacrés à la planification et à la mise en œuvre du projet ?	10 pts
Diapositives : les diapositives sont-elles bien organisées, lisibles, pas pleines de texte, comportant des figures avec des étiquettes lisibles, des légendes, etc. ?	10 pts

Résumé

En plus de vos diapositives de présentation, je souhaite que vous fournissiez un bref résumé de votre projet dans le fichier README de votre repo Github

Ce résumé devrait fournir des informations sur l'ensemble de données que vous utilisez, vos questions de recherche, votre méthodologie et vos résultats.

Le résumé vaut 15 points et sera évalué selon qu'il suit les directives et qu'il est concis mais suffisamment détaillé.

Organisation des dépôts

Les dossiers et fichiers suivants dans votre dépôt Github de projet :

- `presentation.Rmd` + `presentation.html` : vos diapositives de présentation
- `README.Rmd` + `README.md` : Votre rédaction
- `/data` : Votre jeu de données au format CSV ou RDS et votre dictionnaire de données
- `/proposition` : Votre proposition de projet

Le style et le format comptent pour ce devoir, alors prenez le temps de vous assurer que tout semble bon et que vos données et votre code sont correctement formatés.

Des astuces

- Vous travaillez maintenant dans le même référentiel que vos coéquipiers, donc des conflits de fusion se produiront, des problèmes surgiront, et c'est très bien. Engagez-vous et poussez souvent, et posez des questions lorsque vous êtes bloqué.
- Passez en revue les directives de notation ci-dessous et posez des questions si l'une des attentes n'est pas claire.
- Assurez-vous que chaque membre de l'équipe contribue, à la fois en termes de qualité et de quantité de contribution (j'examinerai les commits des différents membres de l'équipe).
- Prévoyez du temps pour travailler ensemble et séparément (physiquement).
- Lorsque vous avez terminé, passez en revue les documents sur GitHub pour vous assurer que vous êtes satisfait de l'état final de votre travail. Alors allez vous reposer !
- Code : Dans votre présentation votre code doit être caché (`echo = FALSE`) afin que votre document soit propre et facile à lire. Cependant, votre document doit inclure tout votre code de sorte que si je reconstitue votre fichier R Markdown, je devrais pouvoir obtenir les résultats que vous avez présentés.
- Exception : si vous souhaitez mettre en évidence quelque chose de spécifique à propos d'un morceau de code, vous pouvez afficher cette partie.
- Travail d'équipe : Vous devez effectuer ce travail en équipe. Tous les membres de l'équipe doivent contribuer de manière égale à l'achèvement de cette mission et des évaluations d'équipe seront données à sa fin - toute personne jugée comme n'ayant pas suffisamment contribué au produit final verra sa note pénalisée. Bien que différents membres d'équipes puissent avoir des antécédents et des capacités différents, il est de la responsabilité de chaque membre de l'équipe de comprendre comment et pourquoi tout le code et les approches de la mission fonctionnent.

Notation

totale	**100 points**
Proposition	10 points
Présentation	50 points
Sommaire exécutif	15 points
Reproductibilité et organisation	10 points
Évaluation par les pairs de l'équipe	10 points
Évaluation des camarades de classe	5 points

Critère

Votre projet sera évalué sur les critères suivants :

- Contenu - Quelle est la qualité de la question recherche et la pertinence des données par rapport à cette question?
- Exactitude - Les procédures statistiques sont-elles exécutées et expliquées correctement ?

- Rédaction et présentation - Quelle est la qualité de la présentation statistique, de la rédaction et des explications ?
- Créativité et Pensée Critique - Le projet est-il mûrement réfléchi ? Les limites sont-elles soigneusement prises en compte ? Semble-t-il que du temps et des efforts ont été consacrés à la planification et à la mise en œuvre du projet ?

Une répartition générale de la notation est la suivante :

- 90%-100% - Effort exceptionnel. L'élève comprend comment appliquer tous les concepts statistiques, peut mettre les résultats dans un argument convaincant, peut identifier les faiblesses de l'argument et peut clairement communiquer les résultats aux autres.
- 80%-89% - Bon effort. L'élève comprend la plupart des concepts, élabore une argumentation adéquate, identifie certaines faiblesses de son argumentation et communique clairement la plupart des résultats aux autres.
- 70%-79% - Effort de passe. L'élève a une mauvaise compréhension des concepts dans plusieurs domaines, a de la difficulté à rassembler les résultats dans un argument convaincant et la communication des résultats n'est parfois pas claire.
- 60%-69% - Effort difficile. L'élève fait des efforts, mais a une mauvaise compréhension de nombreux concepts et est incapable de mettre en place un argument convaincant. La communication des résultats n'est pas claire.
- En dessous de 60 % - L'élève ne fait pas un effort suffisant.

Évaluation par les pairs de l'équipe

Je vous demanderai de remplir un sondage où vous évaluerez la contribution et le travail d'équipe de chaque membre de l'équipe sur 10 points. Vous déclarerez en outre un pourcentage de contribution pour chaque membre de l'équipe. Remplir le sondage est une condition préalable pour obtenir sa propre note sur l'évaluation des membres de l'équipe. Si vous suggérez qu'une personne a effectué moins de 20 % du travail, veuillez fournir une explication. Si un individu obtient un score moyen par ses pairs indiquant qu'il a effectué moins de 10% du travail, cette personne recevra la moitié de la note du reste du groupe.