Exercices

Les exercices utilisent l'outil Gapminder [la version en ligne est disponible à l'adresse https://www.gapminder.org/tools/; la version hors ligne, à l'adresse https://www.gapminder.org/tools-offline/].

Module 1 - Principes fondamentaux de l'analyse des données

Prenez le temps d'explorer l'outil. Dans la version en ligne, le point de départ par défaut est un graphique à bulles montrant l'espérance de vie en 2020, ainsi que le revenue par personne, par pays (la taille des bulles étant associée à la population totale). Dans la version hors ligne, sélectionnez l'option "Bubbles".

1. Pouvez-vous identifier les catégories de variables disponibles, ainsi que certaines des variables? [Vous devrez peut-être fouiller un peu].
2. Pourquoi pensez-vous que Gapminder ait choisi l'espérance de vie et le revenu par personne comme variables par défaut?
3. Remplacez l'espérance de vie par le nombre de bébés par femme. Observez et discutez des changements par rapport au graphique par défaut.
4. Formulez quelques questions auxquelles vous pourriez répondre avec les données par défaut.
5. Formulez quelques questions auxquelles vous pourriez répondre en utilisant certaines des autres variables.
6. À quel moment du "flux de travail de la science des données" pensez-vous que des visualisations de cette nature pourraient être utiles?
7. Ces visualisations permettent-elles de bien comprendre le système étudié (la Terre géopolitique)?

Module 2 - Collecte et gestion des données

1. Quelles sont, selon vous, les sources de données de l'ensemble de données sous-jacent? [Vous devrez peut-être fouiller sur Internet pour répondre à cette question].
2. Toutes les variables et mesures sont-elles également dignes de confiance? Comment pouvez-vous le déterminer?
3. L'ensemble de données sous-jacent est-il structuré ou non structuré?
4. Fournissez un modèle de données ("data model") potentiel pour l'ensemble de données sous-jacent.

Module 3 - Traitement et nettoyage des données

1. Explorez l'ensemble de données avec les outils Gapminder dans leur configuration par défaut. Pensez-vous qu'il pourrait y avoir des problèmes avec les valeurs rapportées? Par exemple, sélectionnez la Suède et les États-Unis dans le menu de cases à cocher à droite et suivez leur parcours de 1799 à 2018/2020. À partir de quel moment les valeurs sont-elles raisonnables? À votre avis, que se passe-t-il au début de la série chronologique?
2. Suivez l'Érythrée pendant la même durée. Recherchez la date d'indépendance de ce pays (vis-à-vis de l'Éthiopie). A votre avis, que représentent les mesures antérieures à cette date?
3. Suivez l'Autriche pendant la même durée. Recherchez la chronologie historique des frontières du pays (Autriche-Hongrie, Anschluss, frontières modernes, etc.). Qu'est-ce que cela implique pour les mesures rapportées?
4. Suivez la Finlande pendant la même durée. Que se passe-t-il en 1809? Cela vous apprend-il quelque chose sur la façon dont les données sont codées dans l'ensemble de données?
5. Désélectionnez tous les pays et laissez la simulation se dérouler de 1799 à 2018/2020. Pouvez-vous identifier des cas où un grand sous-ensemble d'observations se comporte de manière inattendue? Si oui, pensez-vous que cela est dû à des problèmes de nettoyage/traitement des données?
6. Continuez à explorer l'ensemble de données. Vous pouvez modifier les variables affichées ou utiliser d'autres méthodes de visualisation. Globalement, pensez-vous que l'ensemble de données est fiables? L'utiliseriez-vous pour effectuer des analyses? Quelles sont ses forces et ses faiblesses?

Module 4 – Techniques de base d'analyse des données

1. Quels sont les types des 4 variables par défaut (espérance de vie, revenu, population, régions du monde)?
2. Jouez un peu avec les graphiques. Pouvez-vous trouver des paires de variables qui sont positivement corrélées? Négativement corrélées? Non corrélées?
3. Parmi les variables qui sont corrélées, certaines vous semblent-elles présenter une relation dépendante-indépendante? Comment pouvez-vous identifier de telles paires?
4. Pouvez-vous fournir une estimation visuelle de la moyenne, de la médiane, et de l'étendue de diverses variables numériques?
5. Pouvez-vous estimer à vue d'œil le mode des variables catégorielles?
6. Pouvez-vous identifier des moments spéciaux (points temporels particuliers) dans les données, où un changement à longue haleine se produit, par exemple?

Module 5 - Apprentissage statistique

1. Dans la configuration par défaut, nous pouvons identifier quelques règles d'association potentielles. En utilisant des estimations visuelles et approximatives, évaluez la performance des règles suivantes:
    a. Revenu > 8000 → Espérance de vie > 70
    b. Revenu < 8000 ET Espérance de vie < 70 → Région du monde = Afrique
2. Jouez avec divers graphiques et variables et identifiez/évaluez 5+ règles d'association supplémentaires.
3. Identifiez des groupes de pays "similaires" en 2018 [veillez à valider vos groupes à l'aide de divers graphiques].
4. Dans la configuration par défaut, suivez les trajectoires de la Finlande, de la Suède, de l'Islande, de la Norvège et du Danemark entre 1900 et 2018. Les pays semblent-ils suivre des trajectoires similaires? Y a-t-il des valeurs aberrantes ou des trajectoires anormales?
5. Répétez l'étape 4 pour le Brésil, le Paraguay, l'Uruguay, le Venezuela, la Colombie, le Pérou et l'Équateur.
6. D'après les résultats des étapes 4 et 5, pensez-vous que la trajectoire de l'Argentine ressemblerait davantage à celle des pays nordiques ou à celle des pays d'Amérique du Sud? Ou peut-être ni l'un ni l'autre? Votre réponse est-elle la même pour tous les horizons temporels?

Module 6 - Visualisation des données

1. Maintenant que vous avez eu un peu plus de temps pour jouer avec l'ensemble de données, revenons sur une question à laquelle vous avez répondu au module 1: à quel moment du flux de travail de la science des données pensez-vous que les visualisations de cette nature pourraient être utiles?
2. De quelle manière les observations peuvent-elles être anormales? Avez-vous trouvé de telles anomalies? Pouvez vous offrir des explications? [Repensez à l'exemple de l'Afrique du Sud discuté en classe].
3. Choisissez 2+ visualisations "définitives" (méthodes, variables, etc.) autres que la configuration par défaut. Quelles sont les conclusions importantes?
4. Comment décririez-vous les idées de l'étape 3 sans avoir recours à un vocabulaire visuel?
5. Pouvez-vous imaginer comment les données qui vous intéressent dans vos activités quotidiennes pourraient bénéficier du même traitement? Quelles situations pourriez-vous explorer dans un tel scénario? Comment cela aiderait-il votre équipe à mieux comprendre le système à l'étude?