Rédaction d'un rapport d'analyse

Matthieu Cisel

Juin 2022

1 Introduction

Nous allons réaliser ici une première analyse de jeux de données où les résultats ne sont pas parfaitement connus à l'avance, et votre mission consiste à obtenir des résultats nouveaux, à les organiser, les décrire, les interpréter. Vous pouvez partir soit de votre jeu de données - auquel cas vous disposerez d'une moindre guidance - soit du jeu de données portant sur les thèses (la version 3, mobilisée dans l'UE sur la visualisation de données). Dans ce dernier cas, votre mission consiste en premier lieu à trouver un angle d'attaque, et une thématique clairement définie. Nous initions la rédaction du rapport jusqu'en juillet, puis vous avez plusieurs semaines pour le rédiger, et le soumettre avant la fin août.

2 Attentes

Vous devez avoir suivi un cours Datacamp sur la rédaction de fonction, et en montrer le certificat. Ce peut être soit en Python, soit en R. Vous n'êtes pas obligé(e) de créer une fonction dans votre notebook dans le cadre de votre analyse, mais vous gagneriez 2 points si vous le fait (notifiez-le dans le rapport dans ce cas).

Dans le notebook, commentez votre code. Vous devez montrer à un moment que vous avez la capacité de transférer une table produite dans la console sur un tableur comme Excel pour travailler sur son esthétique, **en passant par le presse-papier**. Par exemple, la fonction writeClipboard, sur R (ou ses variantes) suffit à réaliser cette étape. Vous devez également montrer votre capacité à exporter une table (également de votre choix) en csv depuis la console (une nouvelle table intermédiaire, issue d'un merge par exemple). Vous ajouterez dans les rendus une version Excel (ou libre Office) de la table (de votre choix) produite après passage par le presse-papier (vous la nommerez table.clipboard), et une table produite par export csv (vous la nommerez table.exportcsv).

Votre rapport devra remplir les conditions suivantes

1. La structure du rapport est une IMRAD (Introduction - Méthodologie - Résultats - Discussion). Un canevas (template) sur la structure du rapport vous est fourni, suivez-le.

2. Il devra comporter au moins deux figures sur lesquels vous aurez travaillé l'esthétique. Une description précise des résultats afférents est attendue.

3. Il devra comporter au moins deux tests statistiques présentés dans les cours précédents (liés au chi2, à l'ANOVA, à une régression logistique). Les tests statistiques devront être décrits finement.

4. Il fera au moins 2000 mots. Vous pouvez monter jusqu'à 5000.

5. Il doit être réalisé en Latex.

6. Vous ne pouvez pas reprendre des graphiques créés lors des UE précédentes.

Vous n'êtes pas obligé de réaliser un travail de nettoyage des données conséquent (visualisation de données manquantes, outliers, etc.) si vous partez dès le début d'un jeu de données propre, mais cela sera considéré comme un plus, pris en compte dans la notation.

Un plan d'analyse listant les grandes orientations que vous souhaitez suivre devra être envoyé à l'instructeur avant la dernière semaine de juillet.

3 Jeux de données

3.1 Apportez votre propre jeu de données

Assurez-vous auprès de votre employeur, ou de manière générale, d'avoir le droit de communiquer les résultats de votre analyse de données auprès de votre instructeur. Votre jeu de données devra être suffisamment riche pour pouvoir réaliser des tests statistiques. Vous pouvez également trouver un jeu de données en ligne (sur un site comme Kaggle par exemple, ou d'open data d'un gouvernement quelconque).

3.2 Rejoignez des recherches en cours sur les thèses

3.2.1 Mobilisation de références et discipline académique

Nous avons téléchargé et extrait les références de dizaines de milliers de manuscrits. Il s'agit de distinguer comment les disciplines se différencient en termes de nombre de références par manuscrit, de faire une étude sur les données manquantes, et d'aborder la question du type de référence mobilisée (livre, article scientifique, article dans une conférence, etc.).

Les trois premières colonnes permettent d'identifier les textes tandis que les deux dernières sont les informations qui vont être étudiées. La variable **section**

permet de déterminer à quel domaine scientifique chaque bibliographie correspond. La variable **numéro TEL** permet d'identifier la thèse dont sont issues les références. La variable **numéro citation** permet de savoir de quelle citation il s'agit au sein d'un texte. La variable **type** donne l'information concernant le type de document c'est-à-dire s'il s'agit d'un article de journal, d'un livre ou d'un chapitre d'un livre. La variable **titre** donne le titre de la citation.

3.2.2 La question du genre dans la supervision de thèse

Nous avons mis à disposition un jeu de données sur les supervisions de thèse. Vous pouvez vous pencher, en termes de variables indépendantes, sur la discipline d'appartenance, le genre de l'étudiant.e, de son encadrant.e. la période de soutenance (en années), la date de soutenance, etc. Vous pouvez vous centrer sur les étudiant.e.s ou les encadrant.e.s, voire sur le jury, ou même sur les interactions étudiant.e.s / encadrant.e.s.

Pour la supervision, la variable **Year** correspond à l'année durant laquelle la thèse a été présentée. **Genre** correspond au genre du doctorant tandis que **Genre.1** et **Genre.2** correspond au genre des différents encadrants.

3.2.3 Carte blanche sur le jeu de données theses.fr

Ce sujet vise à vous donner le maximum d'autonomie dans le choix des analyses. Il existe de nombreux croisements de variables que vous pouvez explorer, de la géographie au choix de la langue d'écriture en passant par les disciplines académiques.

3.3 Jeu de données artificiel didactisé

Vous avez enfin l'option de travailler sur un jeu de données didactisé, que nous avons conçu spécialement pour un cours visant à développer l'autonomie dans l'analyse de données. Ce jeu de données porte sur des profils Tinder artificiels, c'est-à-dire que nous avons créé le jeu de données de bout en bout, cachant des relations entre variables (nombre de photos postées, score et popularité, etc.). L'exercice consiste à retrouver et décrire le plus grand nombre possible de patterns dans les données, ainsi que les quelques incohérences que nous avons intentionnellement introduites. Si vous faites le choix de travailler en équipe, vous devrez prouver, via votre notebook, que vous n'avez pas simplement pris les figures/résultats d'un coéquipier (ces figures doivent être présentes dans votre notebook mais pas dans le sien).

1. userid : id de l'utilisateur
2. date.crea : date de création du compte
3. score : score associé au profile (reflétant le succès sur l'applications)

4. n.matches : Nombre total de matchs depuis la création du compte
5. n.photos : Nombre de photos sur le profil
6. last.up.photo : date de la dernière mise à jour de la photo de profil
7. last.pr.update : date de la dernière mise à jour du texte du profil
8. last.connex : date de la dernière connexion
9. genre. O pour homme, 1 pour femme. 2 pour autre
10. sent.ana : analyse du sentiment du texte du profil
11. length.prof : Nombre de mots dans le profil
12. voyage : Mot-clé voyage trouvé dans le profil
13. laugh : Mot-clé rire trouvé dans le profil
14. photo.keke : Photo de profil prise en maillot de bain, dans un ascenseur, ou avec des lunettes de soleil.
15. photo.beach : une des photos de profil est prise à la plage

4 Modalités de collaboration éventuelle

Travailler en équipe est recommandé dans cette UE. Néanmoins, les rapports doivent être dans une large mesure individuels. Il n'est pas autorisé d'avoir des figures en commun, ou de présenter des figures qui auront été produites par d'autres. En d'autres termes, vous pouvez travailler sur la même thématique (par exemple, les conditions d'encadrement), mais vous devez avoir des angles d'attaque suffisamment différents pour que les résultats soient distincts. Seule pourra être mutualisée une partie de la méthodologie.