Manipulation et prétraitement de données

Matthieu Cisel - D.U. Data Analyst

Mars 2022

1 Objectifs

La manipulation et le prétraitement de données n'est sans doute pas la partie la plus fascinante de la Data Science, mais il s'agit d'une étape incontournable pour toute analyse de données. Il est fréquent que celle-ci représente près de la moitié du travail de l'analyste. Dans la première partie ce cours, nous revenons sur les manipulations de base de jeux de données (import d'une base, premières visualisations, identification de données manquantes, de problèmes et d'outliers).

Pour les utilisateurs de R, nous nous focaliserons sur des packages comme dplyr or tidyr. Pour les utilisateurs de Python, nous nous focaliserons sur des librairies comme pandas et matplotlib. Après avoir appliqué les premières lignes de commande sur un jeu de données facile à s'approprier, nous nous concentrerons sur un jeu de données portant sur les soutenances de thèse, qui est utilisé d'une part au sein des cours donnés dans l'établissement, et qui fait d'autre part l'objet de travaux de recherche.

2 Premières manipulation d'un jeu de données

Nous vous fournissons un jeu de données (age, gender) présentant deux variables, l'âge et le genre, et qui ne comporte pas de données manquantes. Vous devez importer le jeu de données et représenter la distribution des deux variables. Exportez le notebook Jupyter correspondant au format PDF (d'abord au format html, puis enregistrez l'html au format PDF en passant par l'impression).

3 Analyse d'un jeu de données réel

Maintenant que vous avez effectué vos premières manipulations de jeux de données, nous allons prendre en main un premier jeu de données un peu complexe, portant sur les soutenances de thèse en France. Vous allez devoir le nettoyer, étudier la question des données manquantes, et identifier des problèmes associés au jeu de données. Vous devez produire un notebook Jupyter (en PDF) qui comporte de manière exhaustive toutes les opérations que vous réaliserez (dans

l'ordre des consignes). En parallèle, vous produirez un rapport (au format PDF) ne comportant qu'une sélection de votre travail. Ce rapport doit être structuré, avec titres et sous-titres, et peut être produit à partir de Word, on préférablement en Latex, à partir d'overleaf (un template est fourni, à copier-coller dans la partie gauche d'overleaf, pour accélérer la prise en main du logiciel).

Dans le rapport, les figures doivent avoir une légende, et être numérotées. Il est obligatoire de faire référence à chaque figure dans le corps du texte (exemple : dans la Figure 1, nous voyons que ...). Le rapport inclura toutes les figures que vous aurez produites, et devra suivre la structure suivante :

I. Présentation des données

II. Données manquantes

III. Principaux problèmes détectés

IV. Outliers

V. Résultats préliminaires

3.1 Données manquantes

La détection de données manquantes mobilisera les techniques des cours de Datacamp Dealing With Missing Data (R ou Python). Vous aurez notamment besoin ici des packages pandas et missingno (fonction matrix) pour Python, et dplyr et visdat (fonction vis-miss) pour R. En premier lieu, chargez le jeu de données PhD-v1 mis à disposition dans l'espace Teams. Affichez le nombre de lignes. Êtes-vous certain.e que toutes les données sont bien chargées ? Montrez que vous pouvez retirer ce jeu de données de la mémoire, puis chargez et analysez à partir de maintenant le jeu PhD-v2.

Montrez que vous êtes capable de visionner les premières lignes du jeu de données avec la fonction head. Faites un "summary" des différentes variables et cherchez à identifier la nature des différentes variables que vous allez manipuler, sur la base des noms attribués à ces variables, et d'échanges avec l'enseignant. Dans votre rapport, vous devrez rapporter ce bref travail de compréhension des principales variables de votre jeu de données, sans souci d'exhaustivité.

Dans un second temps, réalisez un graphique pour représenter la répartition des données manquantes au sein du jeu de données. Le résultat de votre travail doit ressembler à la Figure 1.

Observez-vous des régularités dans le caractère manquant des données. Il existe par exemple un lien entre la date de soutenance de la thèse et la date de lancement de la thèse. Comment pourrait-on expliquer ce pattern ?

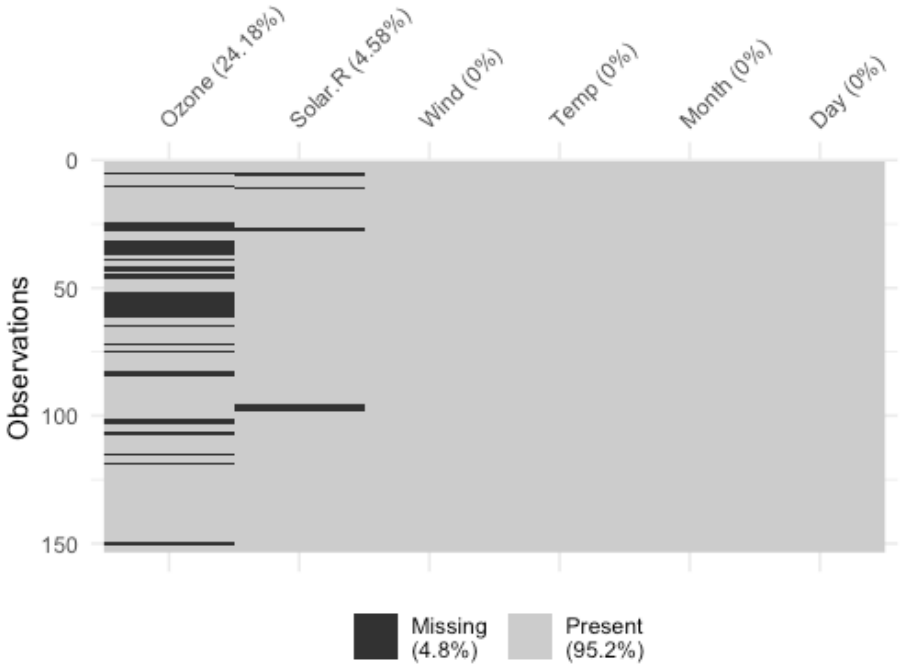

Figure 1: Exemple de visualisation de données manquantes

3.2 Détection d'un problème

Représentez la distribution du mois de soutenance pour l'intégralité du jeu de données, sur la période 1984-2018 (Pourquoi le choix de s'arrêter en 2018 ?). Vous devez trouver les lignes de commandes pour que les mois soient ordonnés dans le bon ordre (janvier à gauche, décembre à droite). Il s'agit ici d'une simple distribution. Pour ce faire, vous devrez convertir la variable associée à la date dans un format qui pourra facilement être traité. Sur R, le package lubridate et la fonction month pourront vous être utiles. Vous devez utiliser la commande qui permet d'extraire automatiquement le mois à partir de la date. Comment interprétez-vous le résultat relatif aux soutenances du mois de janvier ?

La prochaine mission, sur R, mobilisera des fonctions comme filter, group-by ou count du package dplyr, et facet-wrap. La fonction Facetgrid de seaborn peut être utile sur Python. Vous devez représentez en premier lieu la distribution du mois de soutenance pour chaque année, de 2005 à 2018. Pour réaliser ce travail, vous devrez suivre une logique de group.by sur deux variables (le mois et l'année), puis un count. Alors seulement vous aurez la base de données intermédiaires qui vous permettra de calculer des moyennes et des écarts-type.

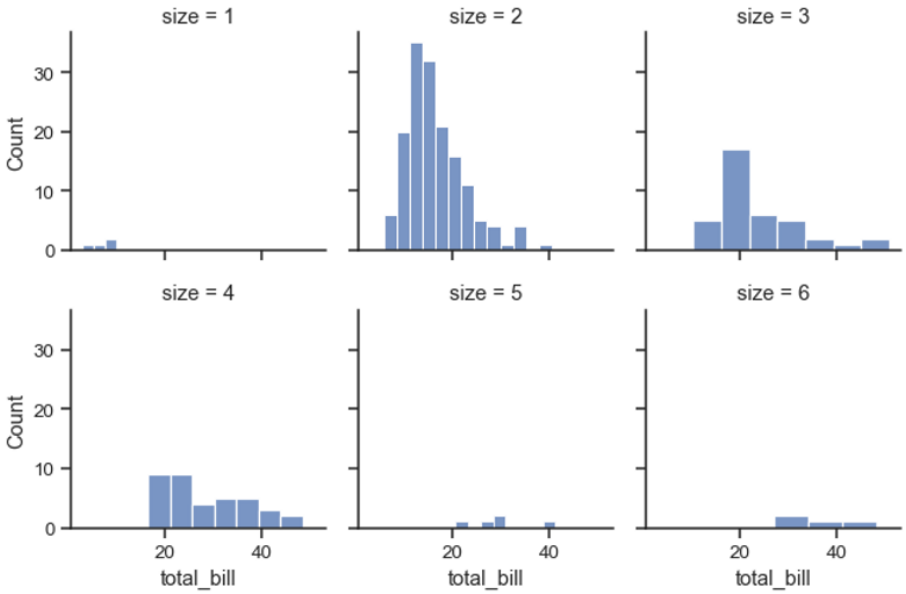

Figure 2: Exemple de graphe en "facets"

Trouvez ensuite une solution pour que seule la proportion des soutenances d'un mois donné soit représentée, pour une année donnée. Il faut donc pour commencer, un graphique par année, soit quatorze graphiques. La Figure 2 vous donne une indication du type de graphe attendu.

Dans l'étape suivante, compilez toutes les années pour ne produire qu'un seul et unique graphique, avec une erreur-type, qui ressemblera à ce que l'on observe en Figure 3. Comment la proportion des soutenances au premier janvier a-t-elle évolué au fil des ans ? Réalisez un graphique analogue à la Figure 4. Comment interprétez-vous cette évolution ? Refaites le graphe analogue à la Figure 3 mais cette fois en enlevant toutes les thèses où la date de soutenance est le premier janvier. Quel est le mois de soutenance préféré ?

Nous allons ensuite nous intéresser à la question des homonymes chez les noms d'auteurs. Focalisez-vous sur le cas de Cécile Martin. Réalisez une enquête pour essayer de comprendre les résultats que vous obtenez. Proposez dans le rapport diverses interprétations des résultats obtenus, que vous représenterez sous la forme d'un tableau.

3.3 Outliers

Nous allons traiter la question des outliers sous l'angle des directeurs et directrices de thèse (supervisors en anglais). Créez un nouveau jeu de données à partir du jeu de données qui vous est fourni, mais en vous focalisant cette fois

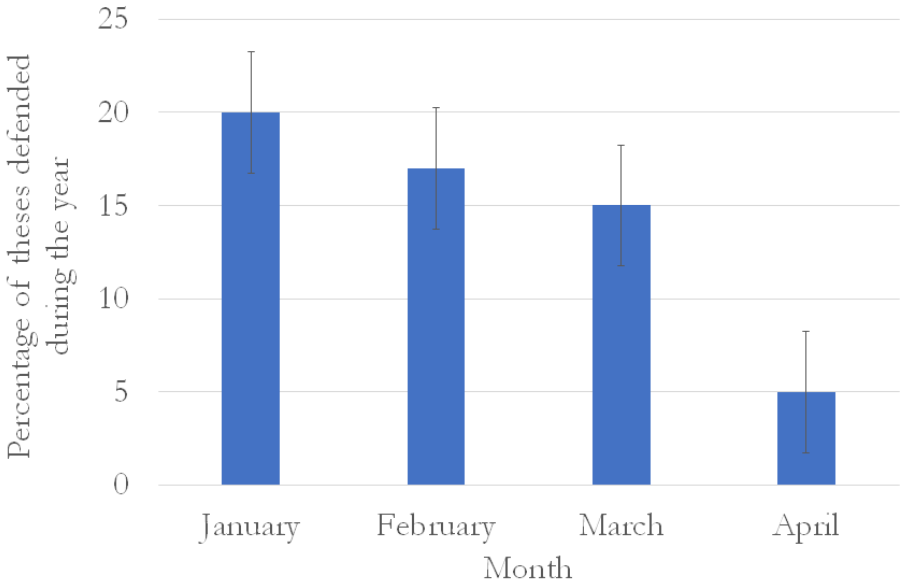

Figure 3: Proportion des thèses soutenues au fil des mois

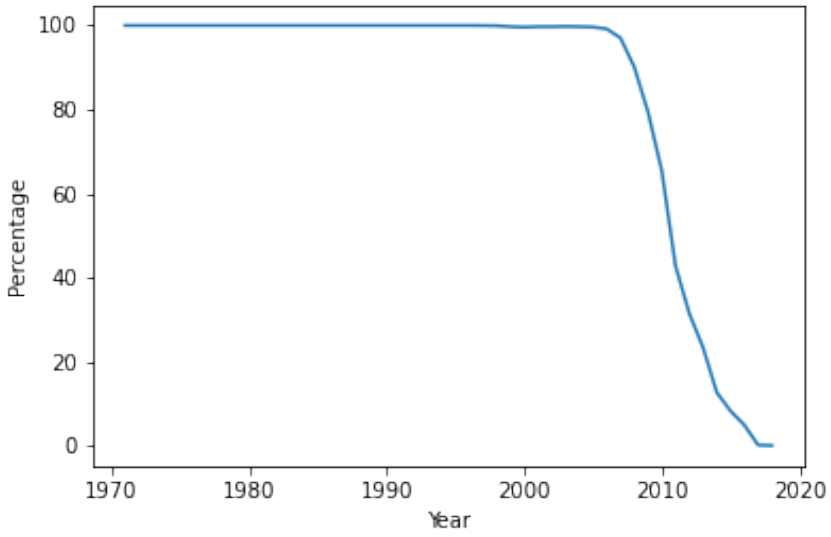

Figure 4: Proportion des thèses soutenues au premier janvier

sur la question des directeurs. Il vous faut une ligne par directeur/directrice, vous devez conserver également l'information suivante : nom et prénom, et créer une nouvelle variable : le nombre de thèses encadrées sur la période considérée (1984-2018).

Identifiez les individus ayant encadré un nombre relativement anormal de thèses, et enquêtez de la manière que vous souhaitez pour déterminer s'il s'agit d'outliers ou d'erreurs dans les données. Précisez dans le rapport la manière dont vous avez procédé pour arriver à vos conclusions, présentez éventuellement des tableaux pour illustrer des cas typiques, selon la logique que vous avez adoptée pour Cécile Martin.

3.4 Obtention de résultats préliminaires

Dans cette partie, le défi consiste à recoder une variable, dont nous allons suivre ensuite la distribution au fil des ans. Recodez la langue d'écriture de la thèse en utilisant quatre niveaux (levels en anglais) : Anglais, Français, Bilingue (enfr et fren), et Autres. Vous devez aussi montrer votre capacité à changer le nom de la variable (Langue de la thèse est trop long). Montrez avec le graphique approprié et sur une période pertinente comment le choix de la langue d'écriture a évolué au fil des ans. Avec R, vous devrez à nouveau utiliser ggplot2, et matplotlib avec Python. Réalisez une description précise des résultats que vous avez obtenus, en rapportant des chiffres dans le corps du texte. Proposez-en une première interprétation. Utilisez une des références fournies dans ce polycopié, et faites référence à la publication en suivant la norme APA7 (exemple : Comme le souligne Martin (2015), la publication de thèses en ligne ...).

3.5 Travail en bonus

Pour ceux qui veulent approfondir le travail réalisé dans le cadre de cette UE en lieu et place de l'initiation à SQL, nous vous proposons de développer trois axes d'analyse pour le jeu de données : les données manquantes, l'évolution des rapports homme/femme au sein des disciplines universitaires, et les techniques de scraping pour conclure. Il vous suffit pour cela de mettre à jour le notebook et le rapport avec les suggestions d'analyse que nous allons vous proposer, et de le resoumettre au niveau du devoir final noté.

Concernant les données manquantes, nous vous proposons de faire une heat map (la couleur dépendant du pourcentage de données manquantes), en choisissant « statut » comme variable en « abscisse » : vous contrastez ainsi les niveaux « enCours » et « soutenue ». A vous de choisir en ordonnée un sous-ensemble de variables qui vous semblerait pertinent par rapport aux analyses. Si vous voulez utiliser R en bonus, vous pouvez utiliser le package UpSetR (ou son équivalent Python s'il existe).

Concernant l'évolution des genres au sein des disciplines, l'exercice consiste à trouver les genres des auteurs avec la librairie Python gender guesser. Une fois que vous avez réussi à créer une nouvelle variable genre, penchez-vous sur la variable « discipline ». Que constatez-vous en termes de nombre de niveaux / modalités pour cette colonne dans les versions 1 et 2 des jeux de données fournies ? Faites un décompte du nombre de modalités et mettez en évidence les disciplines qui ressortent le plus souvent. Le croisement genre / discipline semble compromis. Chargez le jeu de données V3 et faites des « aera charts » pour décrire l'évolution des genres selon les nouvelles disciplines recodées. Reproduisez l'analyse en regardant cette fois l'évolution des langues d'écriture en fonction des disciplines.

Dans une dernière partie de ce travail, nous vous proposons de scraper le site theses.fr pour produire de novo la base de données que vous avez mobilisées pendant l'UE. Suivez le cours sur le scraping donné dans Datacamp et utilisez la technique de votre choix (Selenium ou, plus simplement, Beautifulsoup), pour extraire le maximum de données possibles pour un échantillon d'une cinquantaine de thèses. Il faudra donc en amont trouver une technique pour produire une base de données d'URL pour les 50 thèses concernées. En bonus final (non obligatoire), trouvez une méthode pour extraire les PDFs des manuscrits correspondants lorsque c'est possible.

4 Références

Copeland, S., Penman, A., Milne, R. (2005). Electronic theses : The turning point. Program, 39(3), 185 197. https://doi.org/10.1108/00330330510610546

ElSabry, E. (2017). Who needs access to research? Exploring the societal impact of open access. Revue Française Des Sciences de l'information et de La Communication, 11, Article 11. https://doi.org/10.4000/rfsic.3271

Harnad, S. (2011). Open Access to Research. Changing Researcher Behavior Through University and Funder Mandates. JeDEM - EJournal of EDemocracy and Open Government, 3(1), 33 41. https://doi.org/10.29379/jedem.v3i1.54

Martin, I. (2015). Le signalement des thèses de doctorat. I2D - Information, donnees documents, Volume 52(1), 46 47.

Moxley, J. M. (2001). American universities should require electronic theses and dissertations. Educause Quarterly, (3), 61.

Moyle, M. (2008). Improving access to European e-theses: the DART-Europe Programme. Liber Quarterly, 18(3-4).

Park, E. G., Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. The Electronic Library, 29(3), 394 407. https://doi.org/10.1108/02640471111141124

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., Haustein, S. (2018). The state of OA : A large-scale analysis of the prevalence and impact of Open Access articles. PeerJ, 6, e4375. https://doi.org/10.7717/peerj.4375

Rabesandratana, T. (2019). The world debates open-access mandates. Science, 363(6422), 11–12. https://doi.org/10.1126/science.363.6422.11

Stanton, K. V., Liew, C. L. (2011). Open Access Theses in Institutional Repositories : An Exploratory Study of the Perceptions of Doctoral Students. Information Research: An International Electronic Journal, 16(4).