Analyse de données

Dr. Matthieu Cisel - Ingénieur Sciences Po

Septembre 2024

1 Objectifs

La manipulation et le prétraitement de données constituent une étape incontournable pour toute analyse. Il est fréquent que celles-ci représentent près de la moitié du travail de l'analyste. Dans la première partie de ce cours, nous revenons sur les manipulations basiques de jeux de données (import d'une base, premières visualisations, identification de données manquantes, de problèmes et d'outliers). Dans la seconde, nous insistons davantage sur la visualisation de données.

1.1 Cours

S'agissant des cours magistraux, nous traiterons les sujets suivants :

1. Typologie de données et typologie d'analyses
2. Les bonnes pratiques de programmation
3. Qualité des données, qualité des analyses
4. Visualisation de données : typologie de graphiques
5. Bonnes et mauvaises pratiques de conception de graphiques

Le choix du langage de programmation sera discuté en début d'U.E. 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, matplotlib et seaborn. Après avoir appliqué les premières lignes de commande sur deux jeux de données facile à s'approprier (age_gender et MCF), 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 Activités sur Datacamp

2.1 Cours

Vous devrez suivre les cours suivants sur Datacamp :

1. Data Manipulation with dplyr
2. Introduction to Data Visualization with ggplot2
3. Intermediate Data Visualization with ggplot2
4. Dealing With Missing Data in R

Le suivi des cours et l'obtention de leurs certificats est requis.

2.2 Evaluation obligatoire

Au terme de l'U.E., vous devrez soumettre les preuves (par capture d'écran) de l'atteinte d'un niveau au minimum intermédiaire avancé (plus proche d'expert que de débutant) pour l'assessment "Data Manipulation with R".

3 MOOC à suivre

En sus des cours Datacamp, vous aurez à suivre un MOOC sur les arguments fallacieux, conçu par CY Cergy Paris Université. Vous serez invité par l'enseignant sur la plateforme FUN MOOC, après vous y être créé un compte. Des examens de connaissance seront organisés en classe pour attester de votre maîtrise du contenu.

4 Premières manipulations d'un jeu de données

Commençons par un exercice simple. Nous vous fournissons un jeu de données (age, gender) présentant deux variables, l'âge et le genre. Ce jeu ne comporte pas de données manquantes. Vous exporterez le notebook Jupyter correspondant au format PDF (d'abord au format html, puis enregistrerez l'html au format PDF en passant par l'impression). L'utilisation de ChatGPT d'Open AI ou de Github Copilot est autorisée.

1. Importez ce jeu de données au format csv
2. Utilisez la fonction cut pour couper l'âge des individus par morceaux de cinq années (15 à 20, 20 à 25, etc.), puis affichez la distribution via matplotlib ou ggplot2.
3. Représentez la distribution des deux variables, l'âge d'une part, et le genre d'autre part, sur un même graphique, sous la forme de deux subplots

4. Ajoutez un titre (de votre choix) à la dernière figure

5. Représentez dans un même graphique la distribution de l'âge des individus en faisant une courbe pour les hommes, une courbe pour les femmes

6. Faites le même graphique en jouant sur la transparence des barres (paramètre alpha) pour mieux mieux visualiser l'intersection des deux distributions

7. Faites un barplot représentant sur une barre l'âge moyen des hommes, et sur l'autre l'âge moyen des femmes. Vous ajouterez une barre d'erreur correspondant à l'écart-type

5 Manipulations de données : maîtrise des bases

Vous devez démontrer votre capacité à mettre en œuvre rapidement pandas pour atteindre des objectifs spécifiques. Nous commençons par des exercices faciles, et passons progressivement à des tâches plus complexes.

5.1 Opérations les plus simples

1. Créez deux jeux de données simples df1, df2 (quelques lignes, quelques colonnes), similaires en termes de dimensions

2. Chaque colonne doit avoir un nom (comme a, b, c, d, etc.), les valeurs doivent être des nombres générés aléatoirement en utilisant numpy

3. Pour df1, changez la valeur de la première ligne, première colonne, en NaN (valeur manquante)

4. En utilisant fillna, changez cette valeur manquante en 0

5. Exportez df1 comme une base de données csv

6. Fusionnez ces jeux de données, une fois verticalement, une fois horizontalement, et affichez les résultats dans votre notebook

7. Rassemblez toutes les colonnes de df en une seule colonne en utilisant la fonction melt

8. Créez df3, un jeu de données modifié à partir de df1, où vous avez décalé les valeurs d'une colonne d'une ligne vers le bas

5.2 Une approche géographique des taux de complétion en France

Vous devez d'abord télécharger le jeu de données appelé "Formations engagées", depuis le dépôt de données ouvertes de la CDC.

Nous voulons voir quelle proportion de personnes inscrites terminent effectivement la formation, en fonction du "département" (localisation géographique).

Pour statut_dossier : Réalisation totale : a terminé complètement le cours. Réalisation partielle : n'a pas complètement terminé la formation. Annulation titulaire : l'inscrit a annulé la session de formation. Le centre de formation peut également annuler la session de formation pour une personne donnée avant même qu'elle ne commence.

1. Montrez l'en-tête des jeux de données intermédiaires chaque fois que possible, pour plus de clarté

2. N'oubliez pas de réinitialiser l'index lorsque c'est pertinent

5.2.1 Travaillons sur un jeu de données réel

1. Examinez le jeu de données à la recherche de doublons (il n'y en a pas, mais quand même).

2. Créez une nouvelle variable en multipliant nb_dossiers (nombre de fichiers) par prix_moyen (coût moyen), et évaluez dans quelle mesure cette nouvelle variable (dont vous choisirez judicieusement le nom) correspond à montant_engage. Utilisez l'opérateur == à un moment donné.

3. Remplacez toutes les occurrences du motif "Yoga" par "Cours de yoga"

4. Créez une table dans laquelle vous calculez la somme cumulée de montant_engage pour tous les départements, en utilisant groupby, puis filtrez cette table pour ne conserver que le département de Paris. La table intermédiaire contenant les sommes par département doit s'appeler dep_money. Utilisez la méthode reset_index() pour en faire un dataframe approprié, et renommez les colonnes de manière pertinente.

5. Effectuez un décompte des sessions de formation par région en 2022 (une ligne équivaut à une session de formation), en utilisant la commande size. Cela crée le dataframe train_reg. Trouvez un moyen de calculer le montant total d'argent dépensé par région.

6. Téléchargez la taille de la population de la région (au 1er janvier 2023) à partir de cette adresse. Fusionnez les deux jeux de données (train_reg et régions) pour créer un jeu de données intermédiaire où le nombre de lignes correspond au nombre de régions distinctes. La variable dans le jeu de données principal s'appelle region_lieu_formation. Précisez le type de fusion que vous avez effectué et expliquez pourquoi vous l'avez utilisé. Quelles régions avez-vous perdues / conservées dans le processus ? En vous basant sur ce nouveau jeu de données, calculez le montant d'argent dépensé par habitant pour chaque région.

7. Revenez au jeu de données d'origine "Fondations engagées". Supprimez les données où status_dossier n'est pas "Clos" (au fait, pourquoi est-ce utile si nous voulons calculer un taux de réalisation ?). Utilisez str.contains.

8. Développez une stratégie pour obtenir le pourcentage de "Réalisation totale" (par rapport aux 4 types de "dossiers" clos), en fonction du département. Fournissez une justification.

6 Analyse du jeu de données de thèses

Maintenant que vous avez rédigé vos premières lignes de code, 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).

Les institutions d'enseignement supérieur et de recherche sont devenues réticentes à payer des abonnements coûteux aux éditeurs qui bénéficient indirectement de l'argent des contribuables qui subventionne les travaux de recherche, et les organisations de financement poussent depuis des années en faveur de l'accès ouvert. Par exemple, le Plan S, initialement lancé par l'Europe mais rejoint rapidement par la Chine, vise à terme à contraindre les travaux de recherche financés par des fonds publics à être publiés dans des revues en accès ouvert. Alors que les débats sur l'accès ouvert se sont principalement concentrés sur les articles de recherche, certains chercheurs ont souligné l'importance d'étendre la question aux manuscrits de thèses de doctorat. En moyenne, de tels manuscrits ne rapportent aucun bénéfice économique aux éditeurs scientifiques. Cependant, ils sont souvent financés par l'argent des contribuables, et certains auteurs argumentent donc qu'ils devraient être accessibles au grand public. Alors que des sites Web privés comme ProQuest ont gagné du terrain aux États-Unis, des dépôts en ligne publics se sont développés au fil des années dans d'autres pays.

La France a lancé theses.fr et son archive associée TEL. La recherche basée sur les métadonnées dans les archives ou les dépôts en ligne est fortement contrainte par la nature et la qualité des informations collectées sur les manuscrits. Dans cette unité d'apprentissage, vous devez d'abord scruter le site Web pour obtenir une quantité significative d'entrées (au moins 10 000) afin de construire le premier jet de l'ensemble de données. Dans une deuxième étape, vous disposez de l'ensemble de données des thèses de doctorat, qui comprend environ 480 000 entrées d'étudiants ayant soutenu leur thèse en France de 1985 à 2020. Vous devrez identifier les problèmes pertinents lors de la phase d'exploration de l'ensemble de données. Nous vous fournissons un nombre limité d'instructions ; votre objectif est de fournir des graphiques et des statistiques afin d'expliquer quels sont les problèmes (données manquantes ou peu fiables).

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 sera produit 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 - il peut aussi être trouvé en ligne.

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 et résultats anormaux

V. Résultats préliminaires

6.1 Identifier des données manquantes

Vous aurez notamment besoin de dplyr et visdat (fonction vis-miss) pour R. En premier lieu, chargez le jeu de données PhD mis à disposition dans l'espace Teams.

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.

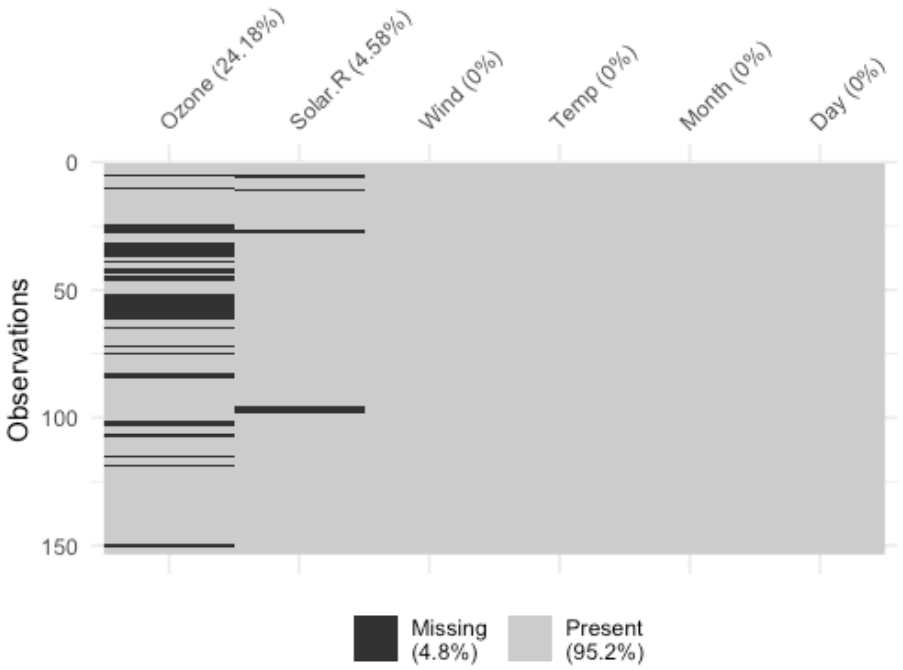

Figure 1: Exemple de visualisation de données manquantes

Faites une première heatmap avec la librairie missingno pour représenter les coocurrences de données manquantes.

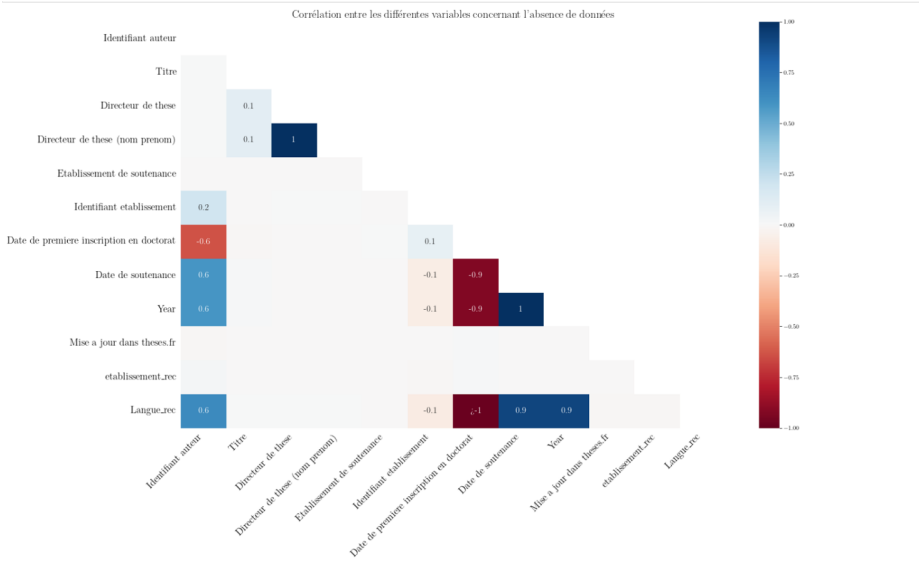

Figure 2: Visualisation de patterns dans les données manquantes

Faites une seconde heatmap via des manipulations avec pandas (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.

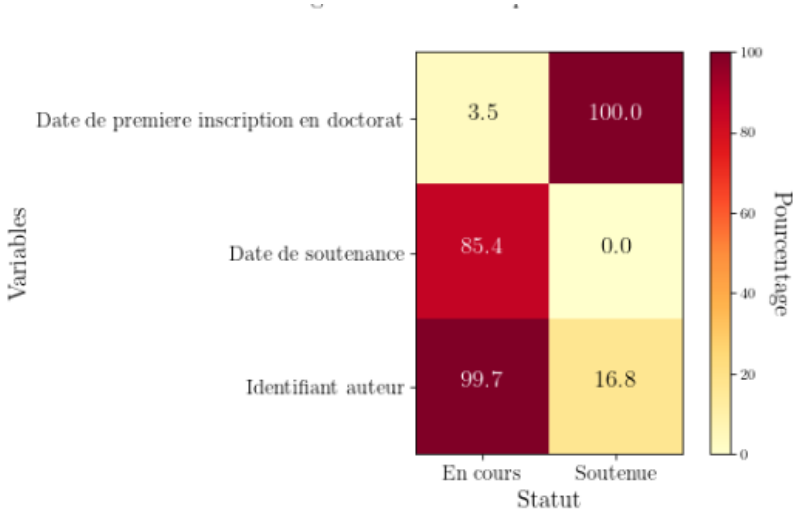

Figure 3: Visualisation de patterns dans les données manquantes - zoom modalités

Observez-vous des régularités dans le caractère manquant des données ? Tâchez de les visualiser. 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 ? Faites un dernier travail sur les patterns dans les données manquantes en utilisant le package UpSet. Vous interpréterez vos résultats.

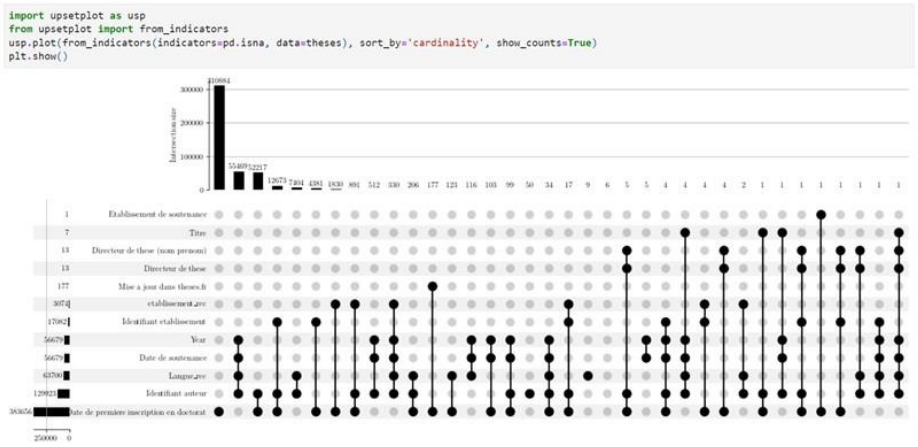

Figure 4: Exemple de visualisation de données manquantes

6.2 Détection d'un problème dans les données

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 mobilisera les logiques de filtre et d'aggrégation (group.by), et ce que l'on nomme un facet-wrap. 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 intermédiaire qui vous permettra de calculer des moyennes et des écarts-type.

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.

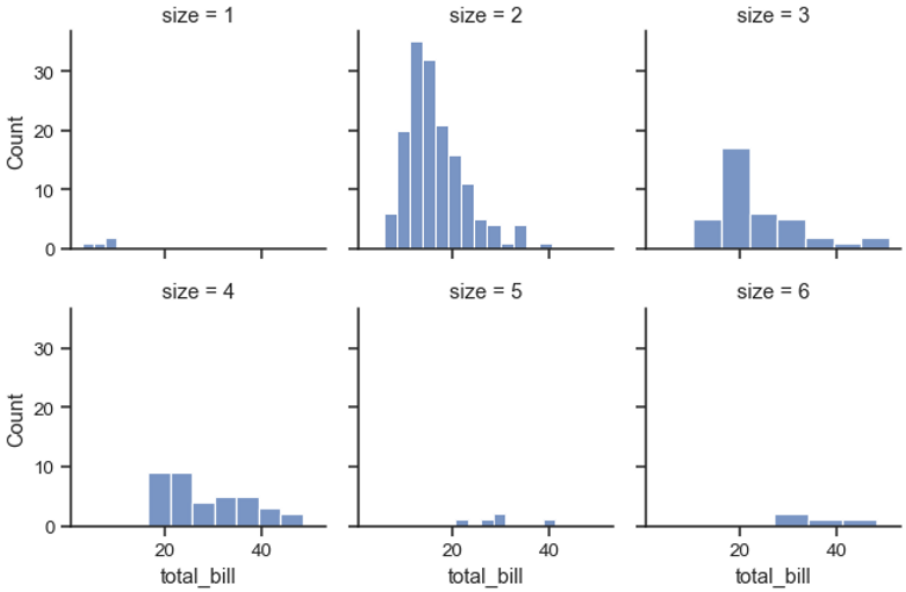

Figure 5: Exemple de graphe en "facets"

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

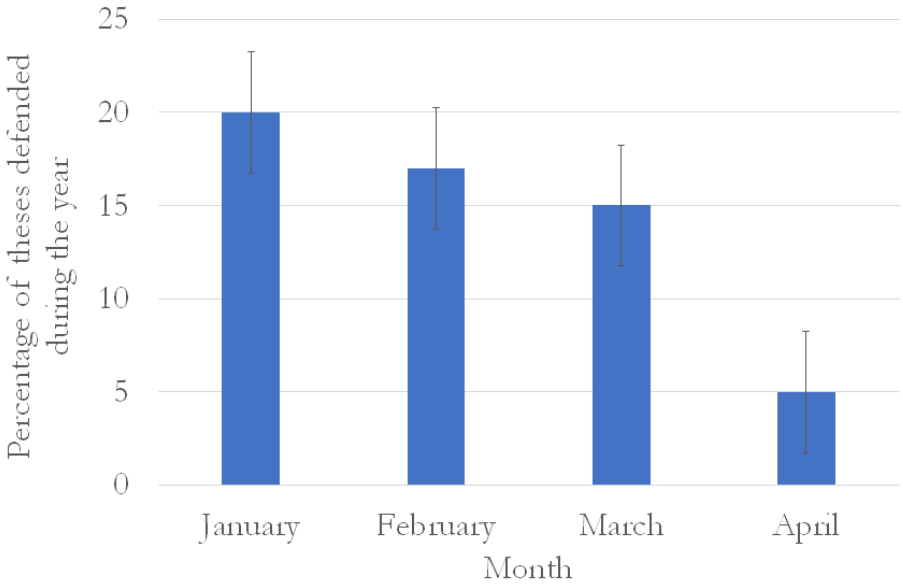

Figure 6: Proportion des thèses soutenues au fil des mois

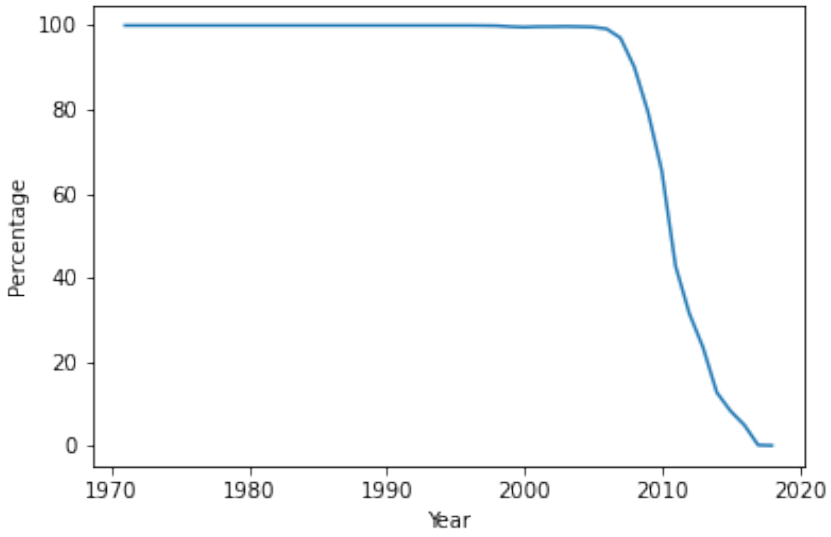

Figure 7: Proportion des thèses soutenues au premier janvier

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.

6.3 Détection d'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 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.

6.4 Obtention de résultats préliminaires

Une variable est dédiée à la langue d'écriture. Vous recoderez (en anglais : to recode) l'ensemble des différents niveaux en quatre niveaux (Français, Anglais, Bilingue - uniquement pour Anglais-Français ou Français-Anglais, soit enfr et fren, et Autre). Appelez cette nouvelle variable language.rec. Traitez la date de la soutenance comme une date, et montrez à travers un plot matplotlib ou ggplot2 (ou pourquoi pas, seaborn), comment le choix de la langue d'écriture a évolué au cours des deux dernières décennies.

Réalisez en plus 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 ...).

Proposez dans un PDF un plan détaillé de type IMRAD permettant de représenter un sélection des résultats de votre choix, en explicitant la ou les questions de recherche ainsi que la ou les hypothèses. Il faut au minimum 3 sous-parties pour l'introduction, la méthodologie, et pour les résultats.

7 Option bonus (+2) pour les avancés : scraping

Nous allons maintenant faire comme si nous n'avions pas eu accès au jeu de données, et qu'il avait fallu le constituer en navigant sur le site theses.fr. Après un cours magistral dédié à la question, et le suivi éventuel de Web Scraping with Python sur Datacamp, utilisez BeautifulSoup, Scrapy ou Selenium pour collecter toutes les informations pertinentes pour 1000 thèses à partir du site theses.fr. Vous devez créer une base de données au format CSV que vous appellerez "votre nom PhD1000.csv".

8 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).