Quelques actualités

- **Point open data ministères** : https://www.etalab.gouv.fr/lengagement-des-ministeres-sur-louverture-des-codes-sources-et-lutilisation-de-logiciels-libres-retour-sur-les-feuilles-de-route-publiees-en-septembre-2021/ : Le 27 septembre dernier, 14 ministères ont publié leurs feuilles de route sur la politique de la donnée, des algorithmes et des codes sources.

- **Bilan 2021datagouv** : https://www.data.gouv.fr/fr/posts/retour-sur-les-activites-de-data-gouv-fr-en-2021/ Point actualités diverses (analyse données, logiciels, gouvernance, etc.)

Quelques actualités

- **Gouvernance (données personnelles)** :https://datascientest.com/vie-privee-quelles-alternatives-a-google-analytics: quelles alternatives à *google analytics*, liste de recommandations de la CNIL : https://www.cnil.fr/fr/cookies-solutions-pour-les-outils-de-mesure-daudience

- **Gouvernance (souveraineté numérique)** : https://www.economie.gouv.fr/conference-pfue-souverainete-numerique : Les mois de décembre et janvier dernier ont vu l'adoption de deux textes européens permettant une meilleure régulation de la donnée: Le *Digital Market Act +Le Digital Service Act*

Quelques actualités

- **Analyse données (visualisation, analyse textuelle)** : Mesurer la réaction publique et médiatique à la sortie du nouveau rapport du GIEC : https://dataforgood.fr/blog/giec , analyse textuelle de twitter : analyse fréquence mots, analyse de sentiments, analyse des sujets de discussion (bertopic), etc

- **Analyse données (création jeux de données)** : https://datascientest.com/avec-scale-synthetic-les-donnees-de-synthese-vont-ameliorer-le-machine-learning: fort enjeu autour des données de synthèse pour partager les données entre organisations, des entreprises qui sortent leur solution propre

Packages R / Modules Python liés

- Python

o **Sklearn** :sur l'imputation univariée: https://scikit-learn.org/stable/modules/impute.html (multivariée est encore au stade expérimental, inspiré de mice dans R), KNNImputer pour l'imputation plus proche voisins, Imputer dans sklearn.preprocessing

o **statsmodel** :
mice, avec predictive mean matching

- R

o **Mice** :
Missing at random hypothesis : la probabilité qu'une valeur est manquante ne dépend que d'observables et peut être prédite en les utilisant https://cran.r-project.org/web/packages/mice/index.html, fondé sur l'article de 2011 Van Buuren and Groothuis-Oudshoorn, dernière *release* en novembre 2021 : https://www.researchgate.net/publication/44203418_MICE_Multivariate_Imputation_by_Chained_Equations_in_R: idée lorsqu'on a plusieurs variables manquantes et qu'on veut utiliser la distribution jointe de toutes les variables pour l'imputation

Packages R / Modules Python liés

- **VIM**
package sur R est aussi très utile pour faire de la visualisation de valeurs manquantes (voir exemple ici : https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17). Contient fonction kNN pour faire de l'imputation par plus proche voisins. Contient différents autres types d'imputations aussi

- **missForest** : implémentation de randomForest. Contient la fonction prodNA permettant de générer des valeurs manquantes de manière aléatoire pour pouvoir ensuite tester
ces différents algos.

- **Hmisc** :impute() : imputation par rapport moyenne, médiane, max. aregImpute() :
imputation en appliquant régression, *predictive mean matching*, etc.