Insee
Mesurer pour comprendre

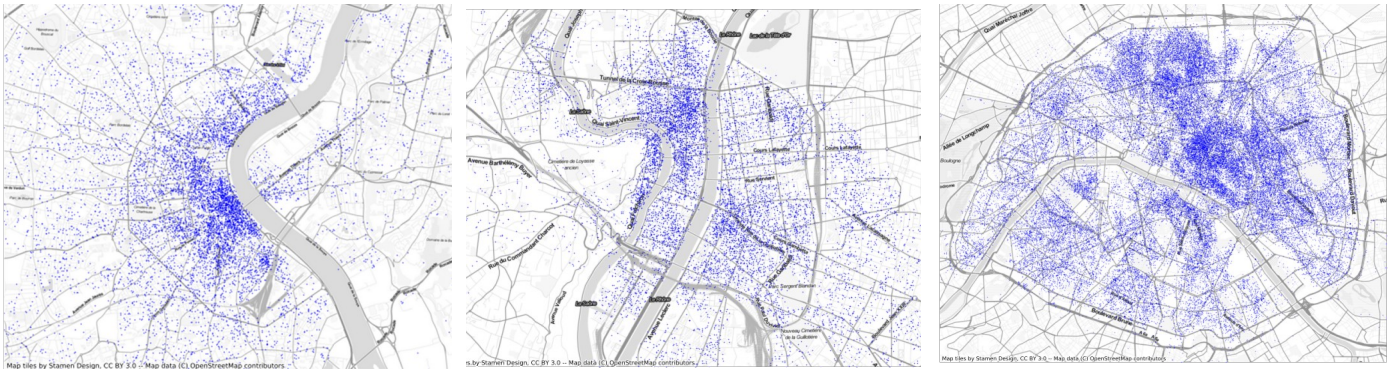
Big Data et Statistique Publique

Lettre d'information
N°11 – Octobre 2021

Entrez dans la Da.tascie.nce ! Retour sur le premier Funathon du Service statistique public

Des données comme s'il en pleuvait, riches, hétérogènes, imparfaites, prêtes à dévoiler leurs mystères. Des méthodes. Des langages. Des codes source. Des *packages*. Des mots qui claquent : *datascience*, *dataviz*, *machine learning*. Et, face à cela, des femmes et des hommes, statisticiens publics de leur état, prêts à gravir des montagnes. Mais comment faire ? Tous seuls, bardés d'enthousiasme, de *MOOC* et d'*outils open source* ? Que nenni : la *datascience* sera collective ou ne sera pas. L'échange de bonnes pratiques devient donc vital. Il restait à inventer des façons efficaces de travailler ensemble, de partager : bienvenue au Funathon.

Pascal Rivière, chef de l'Inspection générale, Insee



Source : Inside Airbnb

Datascience, science des données, de quoi parle-t'on ?

La science des données (*datascience*) est un domaine interdisciplinaire qui utilise des méthodes scientifiques et des algorithmes pour extraire des connaissances à partir de données que celles-ci soient « structurées » (comme des enquêtes, des formulaires) ou « non structurées » (comme des textes, des images, ou des informations provenant de capteurs). La science des données couvre la collecte, la structuration si besoin, la manipulation, les traitements des données, leur analyse, leur modélisation qui peut relever de la statistique ou être issue de l'apprentissage automatique (*machine learning*). Elle couvre aussi la visualisation des résultats obtenus (*dataviz*). Une définition qui entre en forte résonance avec les cœurs de métier de la statistique publique !

L'essor actuel de la science des données rime avec la révolution numérique, les nouvelles données, le développement des algorithmes et des traitements adaptés aux textes, aux images, l'ouverture de certains jeux de données accessibles librement sur le web, et plus généralement la valorisation de la donnée. Il rime aussi avec le développement des outils collaboratifs informatiques (*versioning* avec Git) et l'ouverture des dépôts de codes (Gitlab/Github). En effet, tout cela s'entremêle : la mise en commun des codes, de packages, des bonnes pratiques, les jeux de données ouverts sur lesquels pratiquer, tester pour améliorer, ont pour conséquence que les outils de sciences de données se développent et évoluent vite. C'est par la pratique que l'on en gagne connaissance et maîtrise.

Funathon, une pratique collective non compétitive de la science des données au sein du SSP

Ces habitudes d'échanges dans la communauté de science des données fournissent le matériel pour s'initier, se former et pratiquer. Les *hackathons*, les *challenges*, souvent compétitifs sur des problèmes ciblés y sont courants pour stimuler l'innovation. C'est dans ce contexte que l'idée du « Funathon » est apparue. Son objectif est simple : offrir à tous les agents du service statistique public et de ses partenaires proches, un moment de pratique collective où chacun peut se confronter, s'entraîner, apprendre ou de se perfectionner en *datascience* sur un cas d'usage, dans un esprit collaboratif et non compétitif. Le Funathon vise tous les agents – du curieux débutant, à l'expert en science des données, en passant par celles et ceux qui ont déjà reçu une formation mais qui n'ont pas l'occasion de pratiquer, partant du principe que si un agent n'est pas confronté régulièrement à un cas d'utilisation de ses nouvelles connaissances, celles-ci risquent de se perdre progressivement. Comme dit le dicton, « c'est en forgeant qu'on devient forgeron ». Le Funathon complète ainsi l'éventail des formations aux formats divers offertes aux agents du service statistique public.

Un événement en distanciel, incitation à utiliser les outils collaboratifs

Concrètement, les 21 et 22 juin derniers, cet événement de formation a réuni 150 participants, inscrits au préalable auprès de leur responsable de formation et souvent regroupés en équipes. L'événement s'est tenu en distanciel, même si certaines équipes ont pu se retrouver dans un même lieu. Les participants ont travaillé sur la plateforme Datalab/SSP Cloud, ce qui leur a aussi permis de mobiliser les outils de travail collaboratif (Gitlab et Minio). Le SSP Lab et la Division Innovation et Instruction Technique ont assuré l'assistance tout au long de l'événement pour débloquer, ou aider à réfléchir aux différentes problématiques.

Des données open data géographiques et textuelles…

Les participants ont travaillé sur la base d'un jeu de données Inside Airbnb, site proposant une extraction des logements Airbnb sur quelques villes (ici, Bordeaux, Lyon, Paris). Guidés par des

propositions de sujets, de la documentation mise à disposition ainsi que des *notebooks* en R et python préparés par le SSP Lab, Ils ont pu mettre en œuvre des analyses spatiales sur la répartition des logements et leurs caractéristiques, réaliser des traitements du langage sur les commentaires des clients ou les descriptions des logements, prédire les prix des logements Airbnb et (data)visualiser les résultats de ces analyses.
Le jeu de données mis à disposition était riche. Il contenait des informations sur la description des logements *via* les champs de description ainsi que leur description textuelle, leur localisation, les commentaires laissés par les clients ou encore le calendrier de location. Les participants étaient incités aussi à mobiliser d'autres données descriptives des villes : données carroyées relatives aux conditions de vie, aux logements, à la démographie ou encore des données plus exotiques comme celles localisant les terrasses éphémères lesquelles commençaient à s'étendre dans ces trois villes en ces premiers jours d'été.

… pour pratiquer diverses techniques de sciences des données

Où sont localisés les logements Airbnb ? À quoi ressemblent-ils ? Comment les clients évaluent-ils les logements ? Le profil des loueurs influence-t-il le potentiel locatif du logement ? À quel prix louer son logement ? Quel est le potentiel touristique des logements Airbnb dans les villes ? L'activité touristique liée ? Le profil des clients Airbnb ? En pratique, huit sujets ont été proposés aux équipes, chaque équipe pouvant aussi décider de travailler sur un autre sujet de son choix. Ces sujets ont été construits pour pouvoir aborder un large éventail de techniques de sciences des données : statistiques descriptives, traitement du langage naturel (*natural langage processing)*, analyse de sentiment, analyse de l'image, *deep learning,* prédiction des prix, *machine learning* supervisé ou non, modèles hédoniques, économétrie, statistiques spatiales, jointures spatiales ou entre sources, tableaux de bord, cartes, *dataviz*. Et ceux-ci étaient accompagnés de bouts de code en R comme en Python pour entrer dans le sujet.
Beaucoup de cartes, d'interfaces R-shiny ont été produites à l'issue du Funathon. Plusieurs équipes ont cherché à modéliser les prix, parfois en tenant compte de la dimension spatiale ou en allant jusqu'à construire un simulateur « à quel prix louer mon logement ? ». Certains ont étudié la concurrence fonctionnelle entre la sphère résidentielle et touristique. D'autres ont construit un tableau de bord de l'activité touristique. D'autres encore ont développé un pipeline complet d'accueil et de traitement. Il y a même eu des travaux sur un générateur automatique de texte pour décrire un logement. Comme les données étaient à la fois géographiques et textuelles on pouvait imbriquer les techniques. Certains ont ainsi cartographié les logements selon les langues utilisées dans les commentaires ; d'autres selon les monuments et autres points d'intérêt mentionnés dans le descriptif.

RETEX. Pourquoi es-tu venu ? Qu'est-ce que ton équipe et toi avez réalisé ? Qu'en as-tu tiré ?

Anh Van Lu, service connaissance et développement durable, **DREAL Grand-Est**, **venu** au sein d'une équipe de cinq agents de la DREAL pour pouvoir se sensibiliser collectivement aux innovations offertes par la ***datascience*** à travers des projets concrets de traitement de la donnée. « Notre équipe a réalisé un projet complet de traitement des données allant de la gestion à la valorisation à travers des cartes interactives générées automatiquement, en passant par des traitements des données textuelles contenues dans les annonces Airbnb ». Concrètement un générateur de cartes cartographiant les logements Airbnb dont l'annonce comporte un mot donné – métro par exemple. Van **en retire une démystification** de certains anglicismes de la ***datascience*** - *cloud computing*, *webscraping*, *natural language processing*, *machine learning*… et cerne leur intérêt et la diversité des usages. « Ce Funathon a été aussi l'occasion de se rendre compte de toute la complexité de ces usages, à la hauteur des possibilités qu'offre la *datascience* ».

Pierre Girard, division Services à la Direction des statistiques d'entreprise de l'Insee, **venu pour les données.** « Les données sur les réservations d'Inside Airbnb, constituent un gisement de statistiques très intéressantes pour compléter les connaissances sur l'hébergement touristique marchand et sa conjoncture. À côté, l'enquête de fréquentation touristique se concentre sur la fréquentation des hôtels, campings, résidences de tourisme, villages vacances et auberges de jeunesse. Or en 2018 déjà, 15% des nuitées touristiques dans l'hébergement marchand concernaient des logements offerts par des particuliers via les plateformes internet ». **Venu aussi pour la méthode de travail. «** Le confinement a été l'occasion d'une mise à niveau salutaire en R, comme de nombreux collègues. Le Funathon nous a permis de nous replonger en équipe dans l'environnement R, forts de cette remise à niveau, sur des jeux de données volumineux. Nous avons travaillé sur les réservations aboutissant à des **commentaires** laissés par les locataires, preuve tangible que la réservation a été honorée. Nous avons développé des indicateurs conjoncturels selon le profil des loueurs, du simple particulier louant occasionnellement son logement, au véritable professionnel de l'hébergement. Nous avons tiré profit des résultats du Funathon pour commenter l'actualité et décidé d'intégrer une exploitation basée sur cette nouvelle source dans la Note de conjoncture d'octobre. L'objet consiste en une comparaison entre grandes places internationales du tourisme dont Paris ». Pierre **retire du Funathon** l'intime conviction que « nous avons progressé collectivement, durablement, dans la technique de manipulation, contrôle et fabrication de données, mais aussi dans nos façons de faire. Les modes d'organisation du travail autour des nouveaux outils et sources sont très efficaces : partage du code à plusieurs avec Git, par petites touches itératives, nombreux allers-retours lors de la fabrication de statistiques sur des volumes importants de données. Nous cherchons maintenant à aller plus loin dans l'exploitation de ce type de sources complémentaires aux statistiques usuelles sur le tourisme et vérifier leur qualité ». Et, un groupe de travail transversal a été lancé sur le *scraping* des plateformes (Airbnb, Booking) suite au Funathon.

Bilan et suites

Dans l'ensemble, ce premier événement semble avoir été un succès, 90% des participants ayant répondu à l'enquête de satisfaction à la fin du Funathon envisagent de s'inscrire à une prochaine occurrence. Des marges d'amélioration ont aussi été pointées : besoin de parcours plus fléchés pour les débutants, besoins de formation ou de pratique à l'utilisation de Git, parfois à R ou python. Une nouvelle édition est prévue au printemps prochain. Le format pourra évoluer un peu - pour améliorer la prise en main pour les débutants, pour mettre en place différentes sessions de micro formations pendant l'événement. Les principes devraient cependant rester les mêmes. D'ici là, n'hésitez pas à contacter l'équipe du SSP Lab, ou à rejoindre le canal Slack de l'événement pour en discuter. Vous pouvez également consulter l'ensemble du matériel mis à disposition sur le repo associé.

Ressources/Actualités / Brèves

utilitR, documentation R destinée à tous sans pré-requis de niveau
Autoformations déployées sur le datalab-SSP Cloud et plus
Matériel de formation sur le site intranet du SSP Lab
Prochaines sessions de formation en *datascience*, e-formations par l'Insee ou l'IGPDE listées ici (rubrique informatique)
Spyrales : Soutien python et R entre agents de l'Etat
Symposium international de 2021 de Statistique Canada « Adopter la science des données en statistique officielle pour répondre aux besoins émergents de la société » les vendredis du 15 octobre au 5 novembre
Merci à tous les participants et à ceux qui ont rendu possible le Funathon !

Cette lettre est une occasion d'informer largement et d'échanger.
N'hésitez pas à nous transmettre vos réactions et suggestions d'articles à innovation@insee.fr.
Les archives de cette lettre sont disponibles ici et là.
Demande d'inscription individuelle à la lettre : dg75-l001@insee.fr