



Promo avril 2025  
Data Analytics

# Projet: Data Job

## Rendu Final

Nelson RAMIREZ



# TABLE DES MATIÈRES

<b>TABLE DES MATIÈRES</b>	<b>2</b>
<b>Partie 1: Introduction au projet</b>	<b>3</b>
Contexte d'insertion du projet dans votre métier ou intérêts professionnels	3
Point de vue technique	3
Point de vue économique	4
Objectif principal	4
Mon niveau d'expertise	4
Échanges avec des experts métier	4
Références à des projets similaires	5
Cadre	5
Pertinence	5
Pré-processing et feature engineering	6
Visualisations et Statistiques	6
<b>Partie 2 : Visualisation Avancée</b>	<b>16</b>
Visualisation Power BI	16
Classification du problème	16
Process ETL	17
Modélisation	17
Création de graphiques	18
Conception et optimisation du tableau de bord	18
Page d'introduction	18
Vision globale	19
Analyse comparative des 4 métiers cibles	24
CV par métier (avec sélection dynamique)	28
Outils vs Besoins	29
Recommandation métier	31
Promo avril 2025	30



## Introduction au projet

### Contexte d'insertion du projet dans votre métier ou intérêts professionnels

Ayant un parcours professionnel dans le monde de l'informatique et du développement de logiciel, exploiter les données associées aux métiers de la data me permet d'avoir une visualisation globale sur le marché professionnel dans lequel je vais développer ma carrière prochainement.

### Point de vue technique

Le projet repose sur l'analyse d'un jeu de données issu de l'enquête Kaggle (2020). Ce dernier recense les pratiques, les outils et les rôles occupés par des professionnels ou non du domaine de la data à travers le monde. Il s'agit de données hétérogènes, structurées en colonnes correspondant aux réponses à un questionnaire, certaines en format multi-colonnes (choix multiples).

Le projet mobilise plusieurs compétences liées au métier de data analyst :

- Nettoyage des données : Suppression des NaN, doublons et autres erreurs.
- Structuration des données, choix des informations, mise en forme pour l'analyse, en l'occurrence concaténation des données à choix multiples.
- Création de visuels à l'aide d'outils comme Seaborn ou Matplotlib, Power BI.
- Interprétation des données et mise en place d'un système de recommandation.

### Point de vue économique

Le monde de la data est aujourd'hui omniprésent et connaît une forte croissance. Le fait de mieux comprendre les différents métiers dans ce domaine peut aider :

- Les entreprises à recruter plus efficacement et facilement.
- Les apprenants à identifier les métiers en lien avec leur compétences et appétences.
- Les entreprises à développer en interne l'apprentissage des outils de data science en lien avec leurs besoins.



## Objectif principal

L'objectif de cette enquête est de dresser un état des lieux de l'industrie de la data science en collectant des données sur les profils, les pratiques, les outils et les technologies utilisés, et de comprendre les relations entre les métiers de la data et les outils/compétences utilisés. Les participants ont répondu à un éventail de questions portant sur des informations les concernant, leur expérience en programmation, les langages et bibliothèques qu'ils utilisent, leurs activités professionnelles et d'autres encore.

## Mon niveau d'expertise

En tant qu'ingénieur informatique, le développement de ce projet me permet d'élargir mes compétences vers le domaine de la data, d'explorer, de découvrir de nouveaux outils et méthodologies, mais également de mettre en évidence et d'exploiter mes compétences actuelles pour un bon déroulement de la formation.

## Cadre

Le projet Data Job résulte d'une enquête réalisée par Kaggle, une plateforme de référence dans le domaine de la data science. Afin de réaliser ce projet, un questionnaire de plus de 40 questions a été déposé sur la plateforme et a rassemblé plus de 20 000 participants à travers le monde, allant d'étudiants à des professionnels confirmés.

**Source :** Données issues du sondage 2020 de Kaggle. Ce dataset est librement accessible.

**Volumétrie :** Environ 20 000 réponses, réparties sur plus de 350 colonnes (questions et sous-questions). Les données couvrent un large éventail de thématiques : le poste actuel du répondant, les outils et langages qu'il utilise, les environnements de développement, les tâches réalisées, ainsi que d'autres aspects comme le niveau d'expérience, les formations ou encore les salaires.

## Pertinence

Étant donné notre objectif de cartographier les profils métiers de la data en fonction des compétences, outils et langages utilisés, les variables les plus pertinentes pour notre analyse concernent les intitulés de poste, les langages de programmation utilisés, les outils et technologies utilisés, les tâches principales effectuées, et les compétences avancées spécifiques (ancienneté, expérience).

Les variables sélectionnées incluent notamment :

- Q1, Q2, Q3 : données démographiques
- Q5 : poste occupé
- Q7, Q8, Q9, Q10 : langages, IDE, notebooks
- Q14, Q16, Q17, Q19 : bibliothèques, frameworks, algorithmes ML, outils NLP
- Q23 : activités principales
- Q26A, Q29A, Q30, Q31A, Q32, Q38 : outils cloud, big data, BI, outils d'analyse



Ces variables nous permettent d'analyser à la fois les outils en usage réel, les tâches effectuées et la perception selon les profils.

Cela nous permettra de dégager des profils types et d'établir des liens entre les outils et les métiers.

Certaines limitations existent, notamment en raison de la nature déclarative des réponses, ce qui peut introduire des biais ou des valeurs manquantes. De plus, certaines réponses peuvent ne pas être représentatives de l'ensemble de l'industrie. Le sondage étant volontaire, un biais d'autosélection est possible (sur-représentation des autodidactes, influence de la communauté Kaggle).

#### Limitations

- Particularités : Format de certaines réponses à choix multiples (Q7, Q14...) sur plusieurs colonnes.
- Beaucoup de valeurs manquantes et de réponses incomplètes.
- Déséquilibre dans le nombre de répondants selon les métiers (Statistician, DBA) trop rares pour analyse isolée.
- Les intitulés des colonnes sont complexes à exploiter sans documentation.

## Pré-processing et feature engineering

Un nettoyage et une transformation des données ont été réalisés pour rendre celles-ci exploitables, cela a impliqué :

- Le traitement des valeurs manquantes : La totalité des colonnes étant des variables catégorielles, le choix qui a été fait à été de remplacer leurs valeurs nulles en fonction de ce que chaque colonne représente pour notre analyse.
  - La valeur la plus fréquente (mode) pour la colonne Q6 (Années d'expérience en code) et Q8 (Langage recommandé aux débutants), étant donné que le taux de NaNs n'est pas important et donc le risque de biais est bas.
  - La modalité "I prefer not to answer" pour la question Q4, et "No answer" pour le reste des questions afin de ne pas biaiser nos statistiques et ne pas perdre les données associées aux gens qui n'ont pas répondu à ces questions spécifiques.
  - *Drop* pour toutes les lignes correspondant aux personnes qui n'ont pas répondu à la question Q5 (Poste actuel) étant donné que notre analyse est centrée sur les personnes actuellement travaillant (y compris les étudiants) dans le monde de la data.



- La réorganisation et regroupement de certaines colonnes des questions multi-réponses en une seule colonne contenant la liste des valeurs sélectionnées pour cette question.  
Par exemple : Réunion des colonnes multiples comme Q7 (Langages de programmation): de *Q7\_Part\_1* à *Q7\_Part\_n* fusionnées en une seule colonne Q7 en gardant comme valeur de la ligne une liste du format ['Python', 'SQL', 'autres'].
- La suppression ou exclusion des colonnes non pertinentes, étant associées à des questions du formulaire ne nous apportent pas de données importantes pour notre cadrage et analyse.
- Filtrage des réponses "étudiant" ou "sans emploi" pour certaines analyses ciblées.

## Visualisations et Statistiques

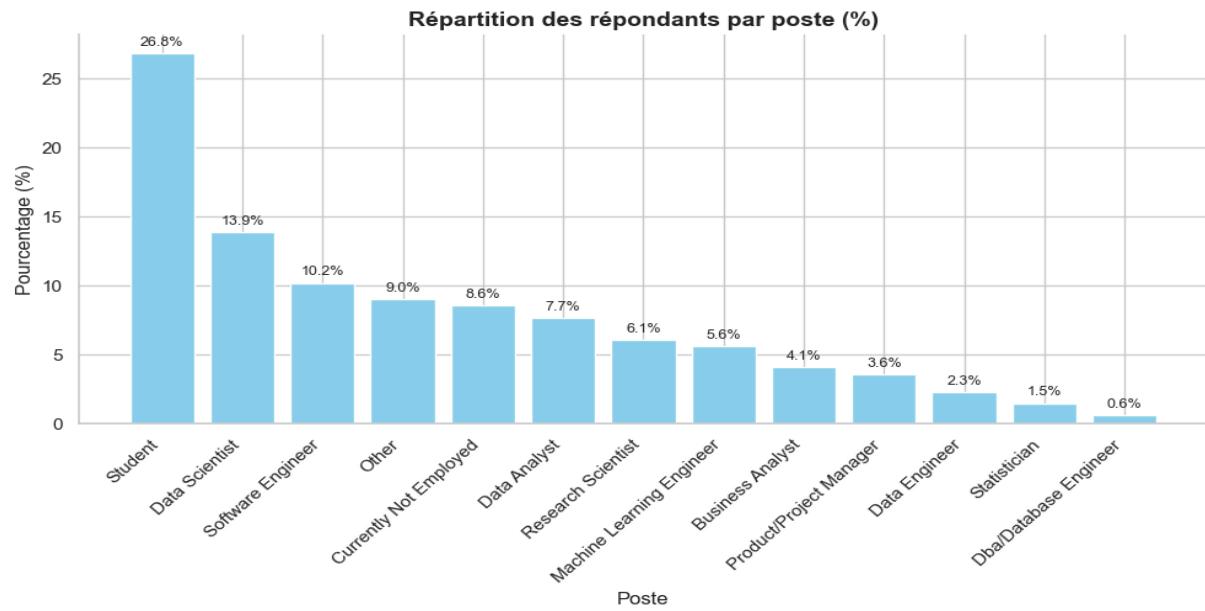
Pour explorer les données issues de l'enquête Kaggle 2020, nous avons réalisé plusieurs visualisations visant à mettre en lumière les grandes tendances du domaine de la Data Science.

Nous avons d'abord réalisé une série de visualisations globales afin de mieux comprendre la population des répondants.

En termes de distribution des données, il est intéressant d'analyser tout d'abord la répartition des postes actuels afin de présenter la composition globale de l'échantillon selon les métiers.



### Répartition des répondants par poste :



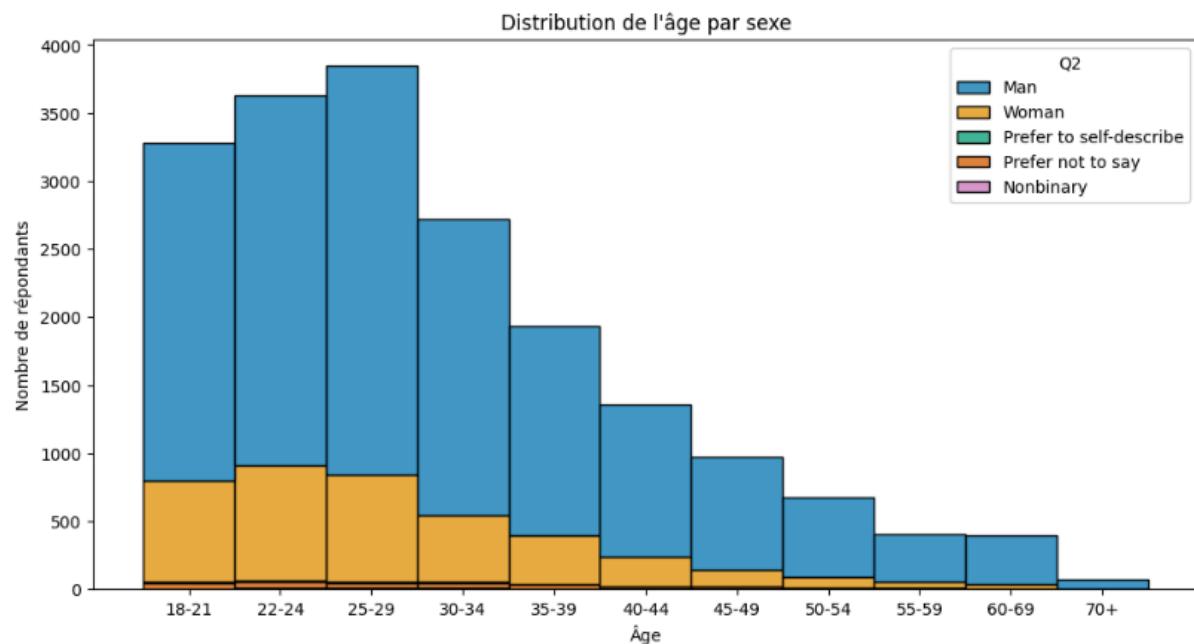
Ce graphique montre la répartition des différents métiers dans le domaine de la Data, tout en illustrant leur fréquence relative sur le marché. Il nous donne un premier aperçu de notre base de données. On y observe notamment une forte représentation des étudiants, environ un quart de l'échantillon, ce qui peut expliquer la proportion importante de jeunes dans les tranches d'âge analysées.

Les Data Scientists (13.3%) et les Software Engineers (10.2%) sont également bien représentés témoignant de l'importance de ces rôles dans le monde de la Data. Les Data Analysts (7.2%) et Data Engineers (2.7%) sont présents en proportion plus modestes. Enfin on remarque que 9% des répondants ne précise pas leur poste, ce qui peut introduire un biais, raison pour laquelle nous avons supprimé ces lignes pour l'analyse plus fine (Pré-processing).

Pour aller plus loin dans la compréhension de notre échantillon, il est pertinent d'examiner la répartition de l'âge selon le genre. Cette analyse permettra d'affiner notre lecture des profils interrogés et de mieux cerner les dynamiques démographiques au sein de la communauté data.



## Âge et genre des répondants :



Ce graphique met en évidence une forte représentation des jeunes dans le domaine de la data : près de la moitié des répondants ont moins de 30 ans, avec un pic dans les tranches 22-24 et 25-29 ans. Cela reflète un secteur dynamique, en pleine croissance, attirant majoritairement des jeunes professionnels en début ou milieu de carrière.

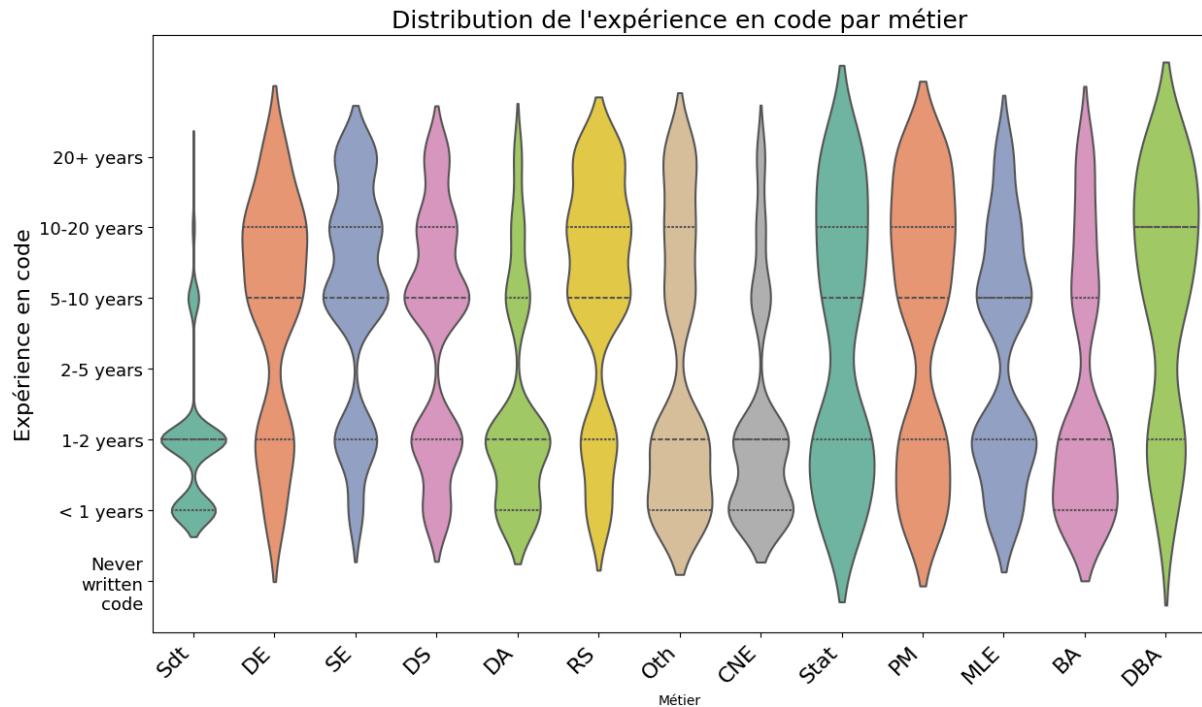
On observe également une nette prédominance des hommes, environ 80% de l'échantillon. Les femmes sont sous-représentées, ne constituant qu'un cinquième des répondants, un déséquilibre encore très marqué. Cette disparité de genre est visible dès les premières tranches d'âge, ce qui souligne la persistance d'un écart d'accès ou d'attractivité du secteur selon le genre.

Après avoir constaté une population majoritairement jeune et masculine, pour mieux cerner la maturité professionnelle de ces profils, intéressons-nous maintenant à leur niveau d'expertise par métier.



### Niveau d'expérience :

Cette analyse permet de mieux comprendre comment l'expertise évolue en fonction des rôles et d'identifier les profils junior, intermédiaire ou senior dans le domaine de la data. Le graphique suivant illustre la répartition du niveau d'expérience par métier.



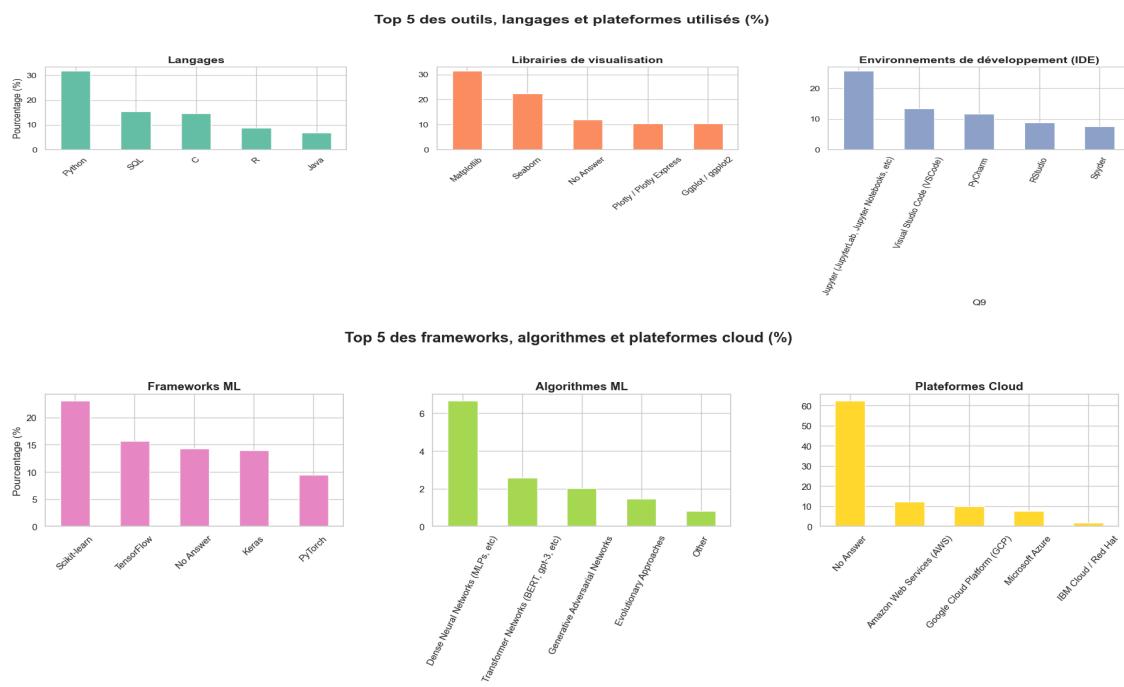
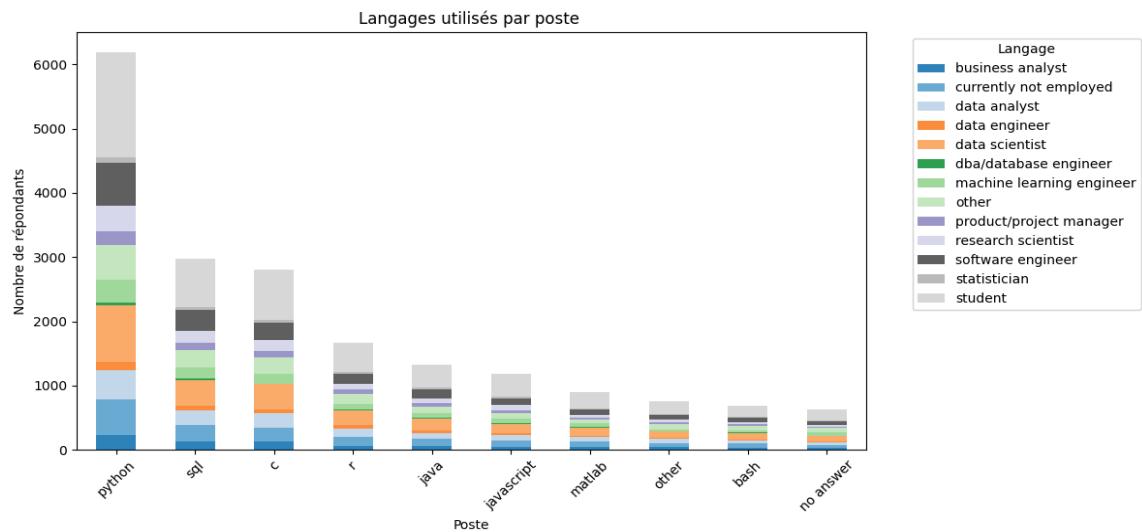
On observe que les métiers de Research Scientist (RS), Database Engineer (DBA) et Software Engineer (SE) requièrent généralement plus d'expérience, reflétant une dimension technique plus poussée et des profils plus expérimentés. En revanche, les postes de Data Analyst (DA) et Business Analyst (BA) semblent s'adresser à des profils avec moins d'ancienneté. D'autres métiers présentent quant à eux une répartition plus homogène en termes d'expérience.

Ainsi, le niveau d'expérience varie selon les rôles occupés, ce qui impacte les compétences techniques requises. Entrons à présent dans le détail des langages de programmation utilisés par les professionnels.



## Langages globalement utilisés :

Les outils sont au cœur des métiers de la data et varient en fonction des rôles occupés. Les graphiques suivants illustrent la popularité des principaux langages utilisés par les professionnels du secteur :



Sans surprise, le langage Python s'impose comme l'outil principal utilisé par la majorité des répondants, suivi de près par SQL et C.

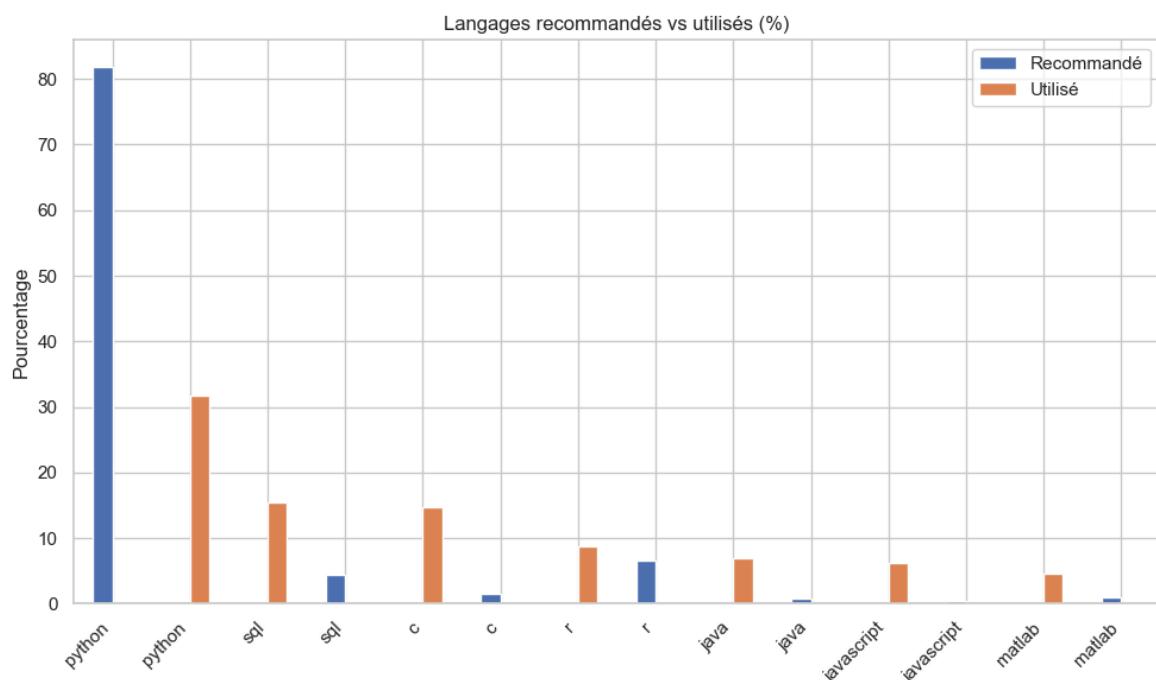
Ces graphiques permettent de comprendre les compétences techniques requises selon les différents métiers de la Data. En illustrant quels outils de programmation (Python, R, SQL, etc.) sont utilisés par chaque profil, ils peuvent aider à mieux cibler les formations ou les



recommandations de compétences à développer pour chaque type de poste. Par ailleurs, cette analyse permet aussi de faire le lien entre les outils et les tâches spécifiques associées à chaque fonction.

Ces langages représentent les compétences techniques de base dans le secteur. Il est aussi intéressant de comparer cela avec les recommandations faites aux débutants pour mieux comprendre les écarts entre apprentissage et pratiques.

### Langages recommandés pour débutants vs utilisés :

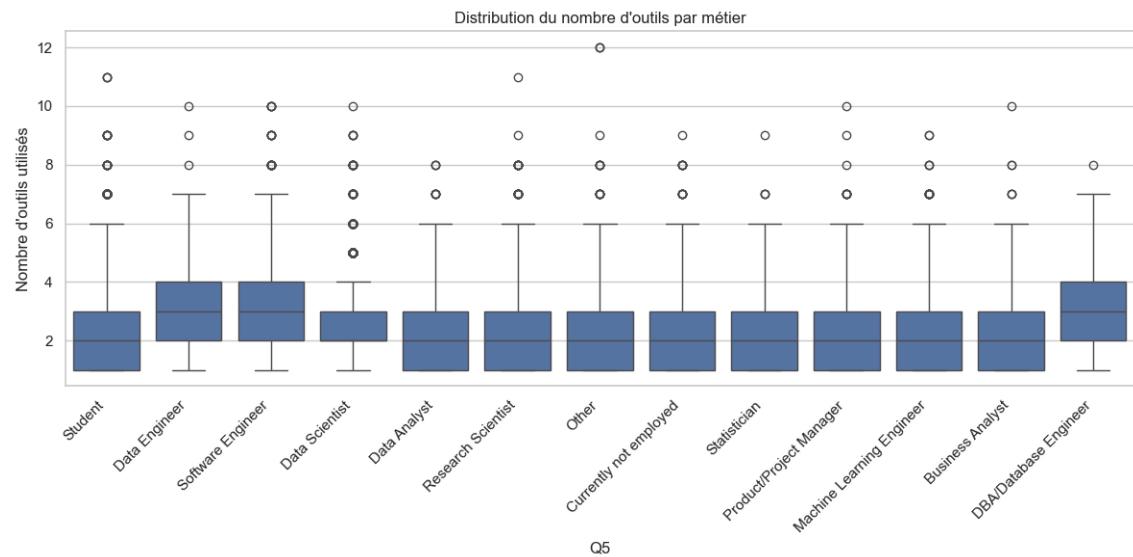


Les langages recommandés aux débutants ne reflètent pas toujours ceux réellement utilisés dans les métiers de la data. Python et SQL forment un socle commun, à la fois conseillés et pratiqués. En revanche, des langages comme R, Matlab ou C, bien que recommandés, restent minoritaires en entreprise. À l'inverse, la présence de Java et JavaScript dans les usages suggère une spécialisation vers le big data ou la visualisation web. Ce décalage entre apprentissage et pratique pourrait refléter soit un retard pédagogique, soit l'émergence de nouveaux besoins métier.

Ce contraste met en évidence des dynamiques d'apprentissage et d'évolution professionnelle. Au-delà des langues, la maîtrise d'un ensemble d'outils est essentielle. Regardons donc la diversité des outils utilisés par métier.



### Distribution du nombre d'outils par métier :

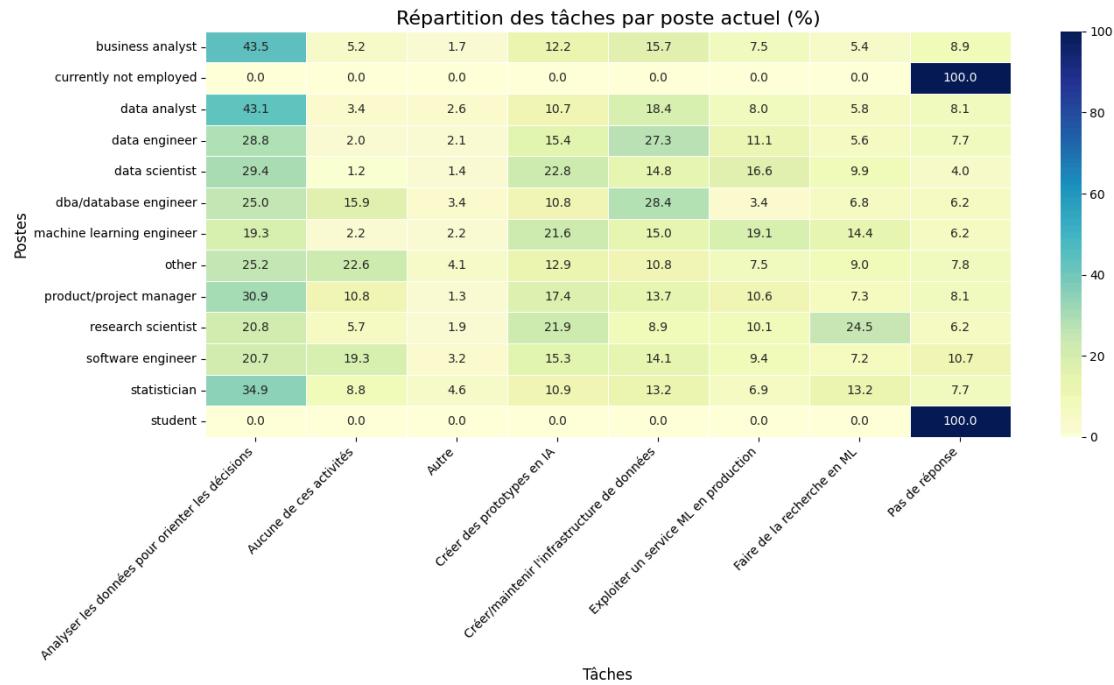


Le boxplot montre la répartition du nombre d'outils utilisés selon les métiers. La médiane est proche de la moyenne, ce qui indique peu d'asymétrie dans la distribution.

Certains métiers, comme Data Scientists, ont une plus grande variabilité et présentent des valeurs atypiques, suggérant que certains profils utilisent beaucoup plus d'outils que la moyenne. Ces écarts peuvent refléter des spécialisations ou cas particuliers que nous allons étudier dans le graphique suivant.



## Tâches réalisées :

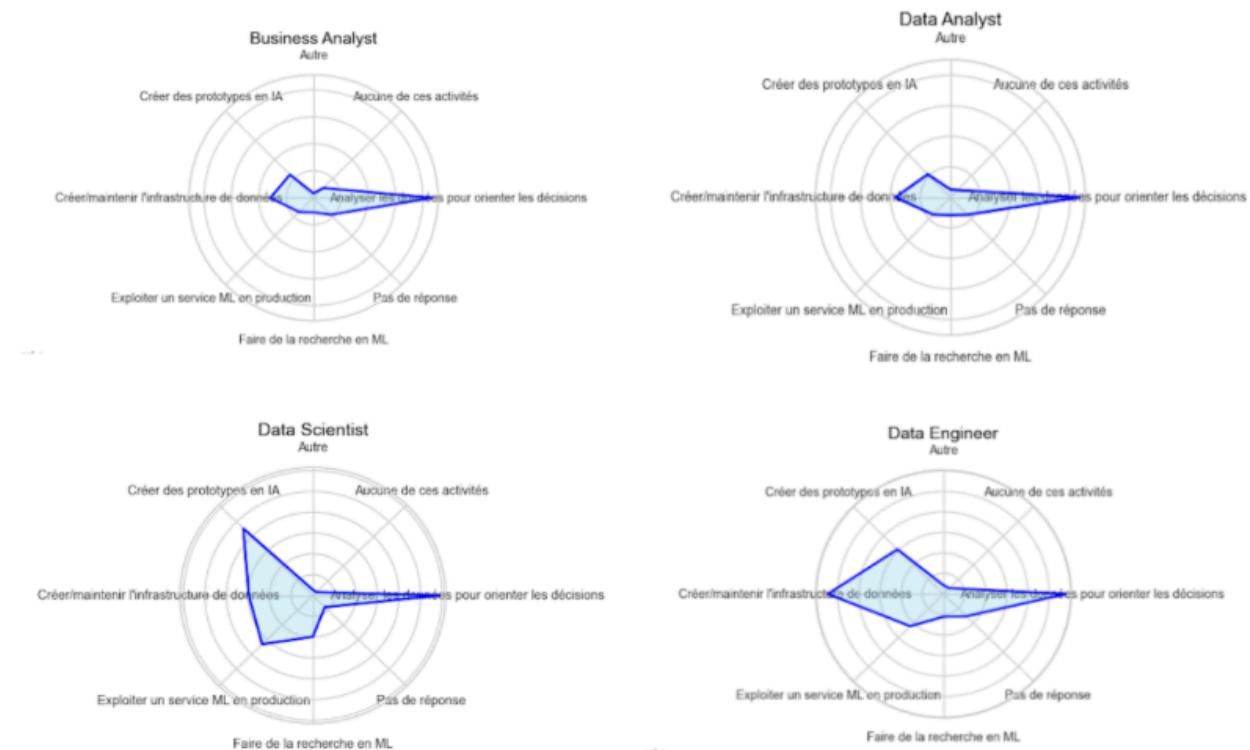


Nous pouvons observer la prédominance des tâches d'analyse des données pour orienter les décisions dans la quasi majorité des postes. Nous pouvons identifier les spécificités de chaque poste. Les tâches auront un lien direct avec les outils ou langages utilisés.

Les tâches réalisées, ainsi que le niveau d'expérience et la diversité des outils maîtrisés sont autant d'éléments analysés pour dégager des profils types. Pour mieux synthétiser les profils, nous allons détailler les caractéristiques clés de quatre métiers cibles.



### Vision par Métier Cible :



Ce graphique radar permet de visualiser la répartition des tâches principales selon quatre métiers ciblés : Data Scientist, Data Engineer, Data Analyst et Business Analyst.

On observe des profils bien distincts :

- Les Data Analysts et Business Analysts se concentrent avant tout sur l'analyse de données et la visualisation, avec des tâches orientées vers la communication des résultats et le support à la décision. Leur rôle demande une compréhension fine, mais moins d'expertise en codage ou en machine learning.
- Les Data Engineers sont chargés de la mise en place de l'infrastructure, la gestion des bases de données et des flux de données à grande échelle. Ils utilisent des outils robustes de traitement de données et interviennent en amont du pipeline data.
- Les Data Scientists , quant à eux, sont très impliqués dans la modélisation statistique, le prototypage de solutions IA, et parfois même le déploiement en production. Leur rôle exige un bon équilibre entre connaissances statistiques et compétences en programmation.  
Un Data Scientist est une personne qui utilise des algorithmes de machine learning pour créer des modèles à partir des données, afin d'aider l'entreprise à gagner en efficacité.



Cependant, il est aussi souvent attendu qu'un Data Scientist assume des fonctions proches de celles d'un Data Analyst.

- La frontière entre Business Analyst et Data Analyst est aujourd'hui tellement floue qu'ils sont parfois considérés comme équivalents.

Dans les deux cas, leurs analyses et leurs rapports servent à aider la direction à prendre des décisions et à définir des objectifs.

**Cette première phase exploratoire nous a permis de mieux comprendre les spécificités des postes data en matière d'outils, de langages et de tâches et les recoupements entre les rôles , offrant un premier aperçu des différents profils**

**Elle sert de socle solide pour la deuxième phase, où nous chercherons à représenter ces résultats à travers des visualisations et des tableaux de bord interactifs, afin d'aider à l'orientation, au ciblage et à la communication efficace de ces insights.**



## Partie 2 : Visualisation Avancée

### Visualisation Power BI

Après avoir identifié les spécificités propres à chaque métier, nous allons à présent concevoir un Dashboard de visualisation à l'aide de Power BI. Cela nous permettra de dégager des insights pertinents et de formuler des recommandations adaptées à chaque profil-type.

#### Classification du problème

Ce projet vise à transformer la richesse des données issues du sondage Kaggle 2020 en visualisations claires, interactives et utiles pour différents acteurs :

- Les apprenants ou candidats à une reconversion, qui souhaitent identifier les compétences les plus valorisées dans les métiers de la data.
- Les centres de formation, qui désirent aligner leurs cursus sur les besoins réels du marché.
- Les recruteurs et décideurs, qui cherchent à mieux comprendre les profils types.

Notre objectif est de concevoir des visualisations permettant de mieux comprendre les profils professionnels et les compétences clés dans les métiers de la data, de comparer en profondeur les quatre métiers cibles (Data Analyst, Data Scientist, Business Analyst, Data Engineer), d'identifier les écarts entre étudiants et professionnels (en matière d'outils et de pratiques), et de proposer des profils types (CV interactifs) ainsi qu'une aide à l'orientation adaptée aussi bien à un usage individuel que collectif.

Pour cela, nous allons créer des visualisations afin de :

- Présenter une vue globale du dataset pour contextualiser les répondants.
- Comparer en profondeur quatre métiers cibles (Data Analyst, Data Scientist, Business Analyst, Data Engineer) sur plusieurs dimensions (outils, langages, tâches, salaires...).
- Explorer dynamiquement les compétences métiers sous forme de "CV type" interactif.
- Mettre en évidence les différences entre les professionnels et les étudiants en termes de pratiques et d'aspirations.
- Proposer un système d'orientation vers un métier cible selon les compétences ou préférences.



## Process ETL

Dans un premier temps, nous avons importé le fichier ***df\_resume.csv***, qui a été préalablement traité et exporté à l'aide de Python.

Les transformations apportées par Python consistaient à nettoyer et préparer ce jeu de données en vue de sa migration vers Power BI. Lors de cette étape de nettoyage, il a été nécessaire de supprimer les espaces superflus et les caractères indésirables, notamment lors du fractionnement des colonnes.

Une fois importé dans PowerBI, via Power Query, nous avons modifié le type de données de quelques colonnes. Ensuite nous avons regroupé les questions/colonnes à choix unique dans une table, et les questions à choix multiples (comme les langages, outils ou IDE utilisés) dans une autre, via un fractionnement par délimiteur suivi d'une transformation en lignes distinctes (unpivot).

## Modélisation

Pour la modélisation, nous avons opté pour une architecture en étoile simple : deux tables, une table de dimension appelée “Respondents” et une table de faits “Questions”.

La table principale, *Respondents*, contient les informations sociodémographiques des participants (âge, genre, pays), ainsi que des données sur leur situation professionnelle (poste actuel, taille de l'entreprise, niveau d'études, expérience, salaire, etc.). Elle constitue une **table de dimension** servant de point d'entrée pour filtrer les analyses selon les profils des répondants.

La deuxième table, nommée “Questions”, adopte une structure normalisée avec trois colonnes :

- **ID\_Respondent** : identifiant unique du répondant.
- **Question** : libellé de la question ou identifiant.
- **Answer** : réponse correspondante.

Cette table agit comme une **table de faits**, stockant l'ensemble des réponses du questionnaire sous un format exploitable pour des visualisations dynamiques. Elle permet une grande flexibilité analytique : on peut regrouper les réponses par type de question, croiser les outils utilisés avec les métiers ou encore analyser la fréquence d'utilisation des technologies selon les niveaux d'expérience.

À ces deux tables principales s'ajoute une table spécifique (*RecommendationList*) restructurée sous forme de liste pour la dernière partie du rapport, la conclusion, où nous proposons une estimation du métier correspondant le mieux aux choix de l'utilisateur. Cela permet d'alimenter directement les listes qui servent à choisir des outils et des compétences spécifiques dans cette dernière page, qui sera détaillée plus tard.



Cette approche modulaire et relationnelle assure une visualisation fluide et interactive des résultats, tout en gardant une structure de données claire et scalable. Elle facilite également l'enrichissement du modèle en cas d'intégration de nouvelles données.

## Création de graphiques

Nous avons ensuite défini des critères pour guider la conception de nos visualisations

- Lisibilité et simplicité : en privilégiant des visuels clairs tels que des KPI, des histogrammes, des matrices, des treemaps et des graphiques en radar.
- Interactivité : en intégrant des filtres croisés (slicer par métier, etc) et des fonctionnalités de “drill-down” pour explorer les données à différents niveaux de détail.
- Storytelling : chaque page du dashboard suit une logique narrative ou comparative afin de guider l'utilisateur dans la compréhension des métiers.
- Pertinence métier : les visuels sont conçus pour répondre à des besoins concrets de compréhension des compétences, des écarts et des attentes du marché.

## Conception et optimisation du tableau de bord

Le tableau de bord a été construit sur Power BI et se compose de neuf pages : une page d'introduction, trois pages “Vision Globale”, deux pages “Analyse Comparative”, une page “CV métier”, une page “Outils vs besoins” et enfin une page “Conclusion”.

### Page d'introduction

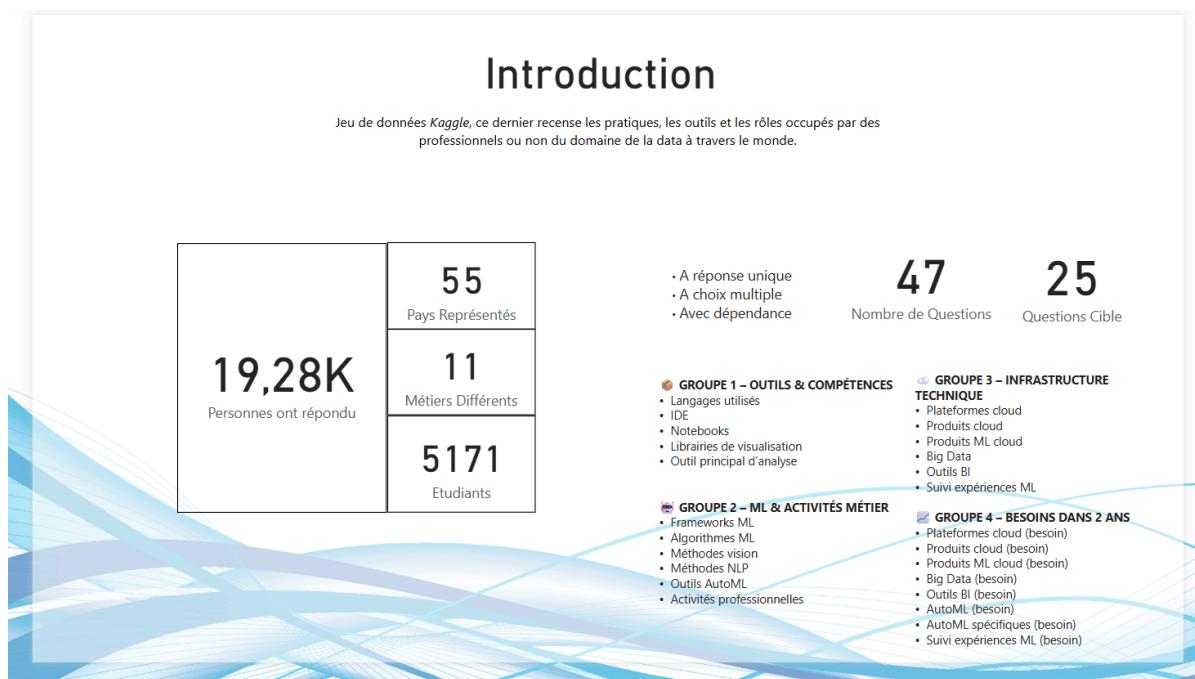
Dans un premier temps, nous avons créé une page d'introduction destinée à présenter notre projet, le jeu de données utilisé, ainsi que le cadre de notre analyse. Cette page comprend plusieurs indicateurs clés (KPI) et un aperçu de la liste des questions :

- Une courte présentation du dataset.
- Une présentation synthétique des données : nombre de répondants, pays représentés, nombre de métiers, nombre d'étudiants, nombre de questions.
- Une liste des questions cibles analysées.



## Introduction

Jeu de données Kaggle, ce dernier recense les pratiques, les outils et les rôles occupés par des professionnels ou non du domaine de la data à travers le monde.



## Vision globale

Ces 3 pages offrent une vue d'ensemble du dataset et permettent de poser le contexte avant d'entrer dans les analyses ciblées.

Les pages comportent les indicateurs suivants :

### Page 1

- **La distribution par genres** : On remarque que la majorité des répondants sont des hommes.
- **La distribution par métier** : Elle montre une forte présence d'étudiants dans le dataset, ce qui nous a confirmé la pertinence d'inclure une comparaison entre profils étudiants et professionnels.
- **Les répondants par pays** : Nous remarquons sur cette carte que nous avons des répondants du monde entier, en Europe, en Amérique, en Asie..
- **Le top 7 des pays représentés** : Les pays les plus représentés sont notamment l'Inde et les Etats-Unis. Cela a une influence notable sur les niveaux de salaires et les pratiques professionnelles observées, et doit être pris en compte dans l'interprétation des résultats globaux.



## Page 2

Sur la gauche de cette page un filtre par métier permet à l'utilisateur d'isoler à tout moment un ou plusieurs métiers et permet une analyse dynamique des différents graphiques. Par défaut, l'ensemble des métiers sont sélectionnés afin de conserver une vue globale.

- **Un graphique en barres de l'expérience en code par niveau d'étude :** En croisant le niveau d'études et l'expérience en code, on constate que les profils juniors sont nombreux, mais déjà bien formés techniquement.
- **Salaire par genre :** On observe que près de 70 % des répondants déclarent un salaire situé dans la tranche la plus basse (0-12 K€ par an). Cette forte concentration s'explique notamment par la surreprésentation de certains pays comme l'Inde ou des régions d'Amérique du Sud, où les niveaux de rémunération sont globalement plus faibles que dans les pays occidentaux. À l'inverse, environ 8% des répondants perçoivent un salaire supérieur à 100 K€ par an. Cette proportion plus réduite reflète la présence de profils qualifiés exerçant dans des pays à plus forte rémunération, comme les Etats-unis, également bien représentés dans notre échantillon. Cela illustre les écarts géographiques importants en matière de rémunération.
- **Distribution des genres par âge :** La majorité des répondants appartiennent aux tranches d'âge les plus jeunes, notamment entre 18 et 34 ans. Deux interprétations principales peuvent être avancées :
  - D'une part, les métiers de la data sont relativement récents, en constante évolution et attirent particulièrement les jeunes générations, souvent issues de formations techniques récentes.
  - D'autre part, notre échantillon pourrait être biaisé par le canal de diffusion du questionnaire, notamment la plateforme Kaggle, très prisée par les étudiants et jeunes professionnels.



## Page 3

Sur cette page, nous avons regroupé l'ensemble des outils et pratiques déclarés par les répondants, en les organisant selon 4 grands groupes thématiques :

- Outils et langages
- Machine Learning et activités
- Infrastructure technique
- Besoins à deux ans

Grâce aux boutons en haut à gauche, nous pouvons filtrer dynamiquement chaque catégorie pour explorer les résultats sous forme de Treemaps interactifs.

**Outils et langages** : On observe une forte dominance de Python et SQL dans les langages utilisés, ce qui confirme leur rôle de socle commun dans la data. Du côté des IDE, Jupyter Notebook et VS Code sont largement en tête, et les bibliothèques de visualisation les plus utilisées sont Matplotlib et Seaborn.

**Machine Learning et activités** : On trouve des méthodes plus avancées comme l'AutoML, les méthodes de vision par ordinateur (Computer Vision), et celles liées au traitement du langage naturel (NLP). Les treemaps de cette page montrent que certaines techniques comme l'automatisation du choix de modèle ou du réglage des hyperparamètres en AutoML, ou encore la classification d'images et les modèles transformeurs en NLP, commencent à émerger dans les pratiques mais restent encore relativement peu représentées parmi l'ensemble des répondants.

**Infrastructure technique** : Du côté des bases de données et technologies Big Data, MySQL et PostgreSQL ressortent nettement. En BI, Tableau et Microsoft Power BI dominent largement. Pour le cloud, Amazon Web Services (AWS) et Google Cloud sont les plus utilisés, tandis qu'Azure est un peu moins utilisé.

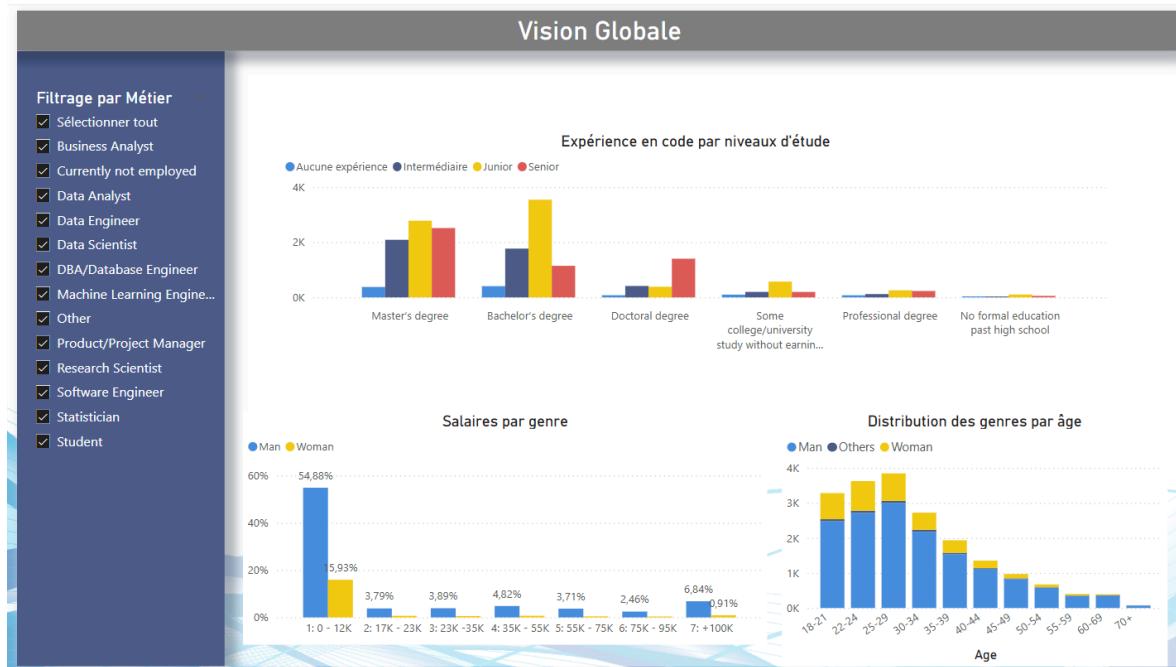
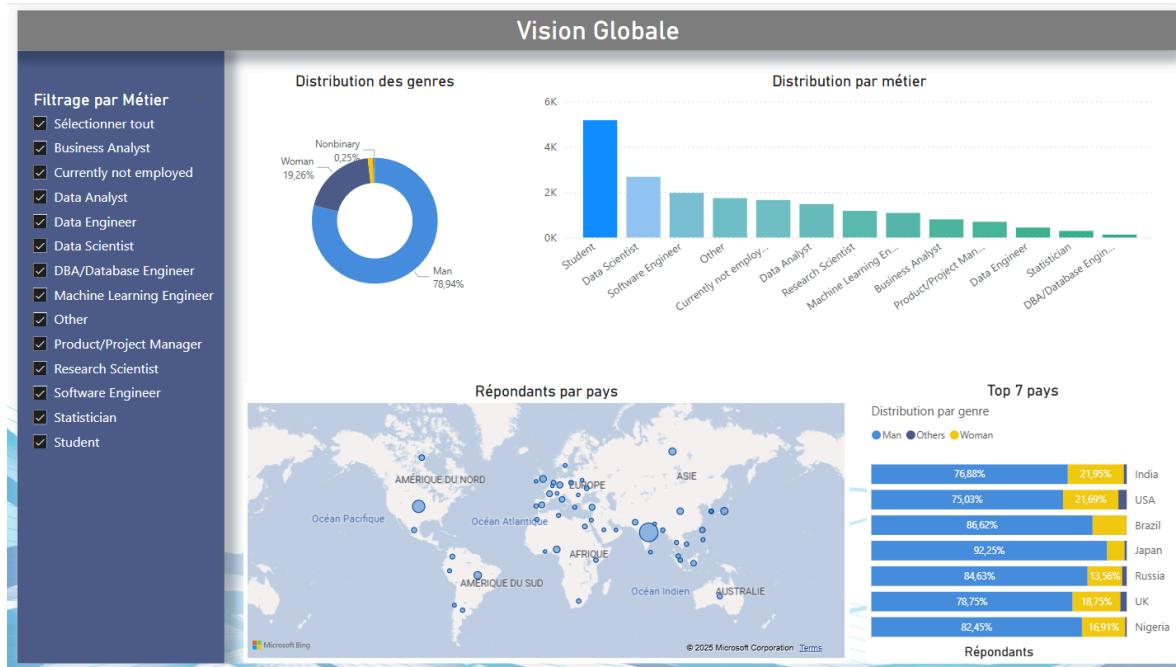
**Besoins à deux ans** : Lorsqu'on interroge les répondants sur leurs besoins futurs, les réponses confirment globalement les tendances actuelles : En BI, Power BI et Tableau restent les outils les plus souhaités. En Big Data, MySQL reste très cité, mais on voit apparaître MongoDB, ce qui traduit une volonté d'ouverture vers des solutions NoSQL.

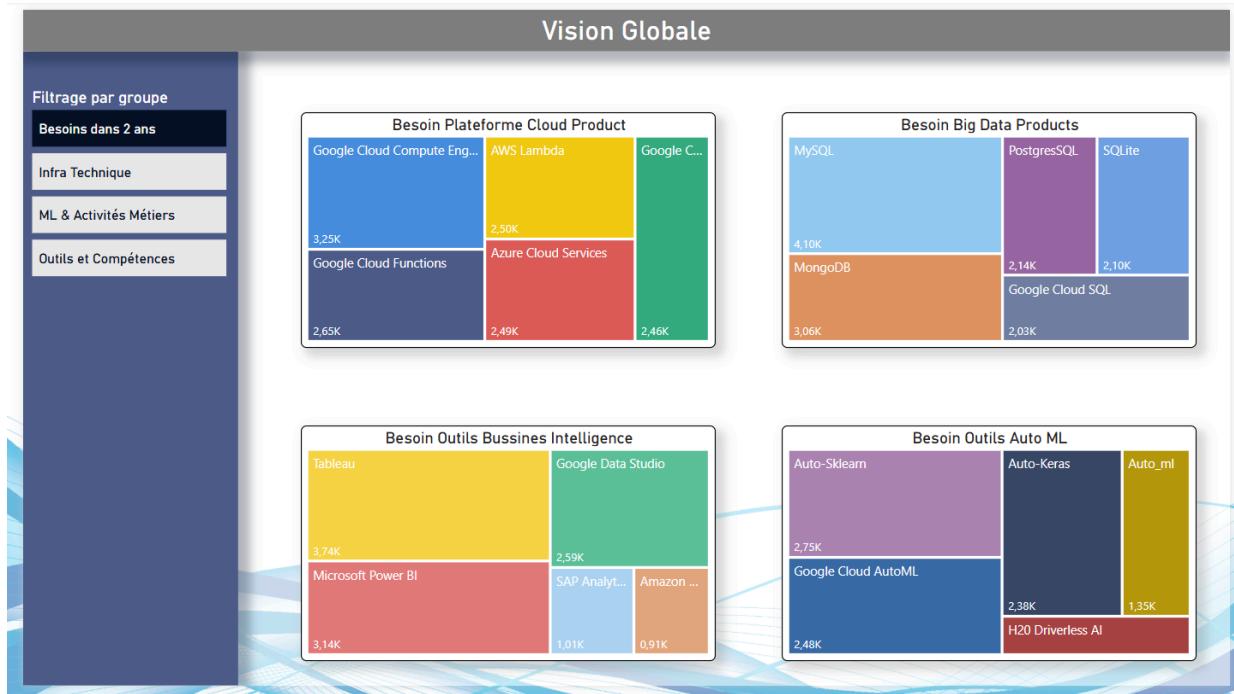
Ces résultats globaux donnent une première vision des outils dominants dans l'ensemble des métiers de la data.

Dans les pages suivantes, nous allons voir comment ces outils se répartissent et se spécialisent en fonction des quatre rôles ciblés : Data Scientist, Analyst, Engineer et Business Analyst.



## Promo avril 2025 Data Analytics





### Préambule aux pages qui suivent :

Avant de présenter les pages qui suivent, il convient dans un premier temps d'expliquer notre choix des 4 métiers cibles. Comme évoqué précédemment, nous porterons notre analyse sur les métiers suivants : **Business Analyst**, **Data Analyst**, **Data Engineer** et enfin **Data Scientist**.

Nous avons fait ce choix de métier pour plusieurs raisons :

- Ces métiers sont parmi les plus fréquents dans l'univers de la data et également dans notre base de données.
- Ces métiers représentent bien la chaîne de valeur complète, à la fois en termes de pertinence métier et de complémentarité :
  - Business Analyst : Compréhension métier, besoin business, KPI.
  - Data Analyst : Exploration, visualisation et reportings.
  - Data Engineer : préparation, ingestion et qualité des données.
  - Data Scientist : modélisation, prédiction, machine learning.
- Simplicité de compréhension, compromis entre spécialisation et diversité :
  - Deux profils orientés analyse et métier : Business Analyst et Data Analyst.
  - Deux profils orientés technique et modélisation : Data Engineer et Data Scientist.



## Analyse comparative des 4 métiers cibles

### Page 4

Sur la gauche de cette page, nous avons 4 boutons, un bouton pour chaque métier, permettant à l'utilisateur d'analyser les données d'un métier ou d'en exclure un. Afin de faciliter la comparaison entre les métiers, les graphiques, à l'exception du salaire, seront présentés en proportion interne à chaque métier.

Ainsi, cette page vise à comparer directement les quatre métiers sur plusieurs dimensions clés :

- **Histogramme empilé du salaire moyen par poste :** On remarque que les salaires moyens varient fortement entre les métiers :
  - Les Data Scientists et Data Engineers affichent les rémunérations les plus élevées, ce qui peut s'expliquer par le haut niveau de compétences techniques sur ces postes. D'autres facteurs peuvent également entrer en jeu, comme le pays, la taille de l'entreprise, la demande pour un poste.
  - Les Data Analysts sont nettement moins rémunérés, ce qui peut refléter des exigences techniques moins élevées.
- **Histogramme groupé de la distribution des métiers par âge :** Ce graphique met clairement en évidence les tendances d'âge selon les métiers. On observe que les Data Analysts sont majoritairement représentés dans les tranches d'âge les plus jeunes, tandis que les Data Engineers appartiennent plus souvent à des tranches d'âges plus élevées. De manière générale, on constate une tendance à la jeunesse dans les métiers de la data, ce qui peut s'expliquer par le fait que ces professions sont relativement récentes et attirent de nouvelles générations de diplômés.  
Le fait que les data analysts soient généralement plus jeunes que les autres profils de la data pourrait témoigner d'un cheminement professionnel où les compétences se spécialisent progressivement, menant certains vers des postes plus techniques comme celui de data scientist ou de data engineer.
- **Histogramme groupé de la présence des métiers par taille d'entreprise :** Ce graphique met en évidence une forte présence des Data Analysts dans les petites entreprises, comparativement aux autres métiers de la data. À l'inverse, les Data Engineers sont davantage représentés dans les grandes entreprises. On remarque également que les entreprises de taille intermédiaire semblent moins recruter ce type de profils, toutes spécialités confondues. On peut expliquer cela par plusieurs facteurs :
  - Les petites structures, souvent moins équipées techniquement, recherchent des profils polyvalents comme les Data Analysts, capables de produire rapidement des analyses exploitables pour le pilotage opérationnel.



- Les grandes entreprises, en revanche, disposent généralement de systèmes d'informations complexes nécessitant des compétences techniques pointues, ce qui explique le recours plus fréquent aux Data Engineers.
  - Les entreprises intermédiaires peuvent se situer dans un entre-deux : pas encore assez matures sur le plan technologique pour structurer des équipes data complètes, mais plus exigeantes qu'une startup. Ce positionnement peut freiner l'intégration en raison de contraintes budgétaires ou organisationnelles.
- **Histogramme groupé de l'expérience en code des métiers :** On constate que les Data Engineers disposent généralement de davantage d'expérience en programmation que les autres métiers de la data tandis que les Data Analysts en ont souvent moins.
- On observe que plus de 50 % des Data Scientists ou Engineers ont entre trois et dix ans d'expérience en codage. À l'inverse, la majorité des Data Analysts se situent dans une tranche de un à cinq ans et cette expérience est encore plus faible chez les business analysts, dont plus de la moitié déclare entre zéro et deux ans d'expérience.
- Cet écart s'explique par la nature même des postes, les Data Engineers ou Scientist travaillent sur des problématiques techniques comme la gestion des pipelines de données, le déploiement de modèles ou l'implémentation d'algorithmes complexes, ce qui nécessite une solide maîtrise en code. A l'inverse ,les Data Analysts sont davantage tournés vers l'analyse métier et la restitution des résultats à travers des visualisations, où l'expertise en programmation est moins centrale.

## Page 5

Cette page propose une lecture comparative des outils mobilisés selon les quatre métiers ciblés : Data Scientist, Data Engineer, Data Analyst et Business Analyst.

- **Langages de programmation :**

On observe une base commune entre les métiers avec Python et SQL présents partout, notamment chez les Data Scientists (Python dominant), Data Engineers et Data Analysts.

Cependant, on note quelques distinctions : R est plus utilisé par les Data Analysts, Data Scientist et Business Analyst. Les Data Engineers mobilisent davantage Java, Bash, ce qui témoigne de leur proximité avec l'infrastructure logicielle et les systèmes.

Les socles communs Python et SQL confirment une base universelle dans la data, mais les langages secondaires permettent de distinguer les spécialités techniques.



- **Frameworks de Machine Learning :**

Les Data Scientists utilisent largement des frameworks avancés comme Scikit-learn, TensorFlow, XGBoost et Keras, montrant leur rôle dans la modélisation. Les Data Engineers les utilisent également, mais dans une moindre mesure. Les Data Analysts et Business Analysts les utilisent peu, ce qui est cohérent avec leur rôle moins axé sur la modélisation.

Les frameworks de ML sont un marqueur fort de spécialisation : seuls les métiers centrés sur la modélisation (Data Scientist, parfois Data Engineer) les mobilisent.

- **Outils BI (Business Intelligence) :**

Les trois outils BI majeurs (Tableau, Power BI, Google Data Studio) sont présents dans les quatre métiers. Même les Data Scientists et Data Engineers, censés être plus “techniques”, les utilisent de manière significative.

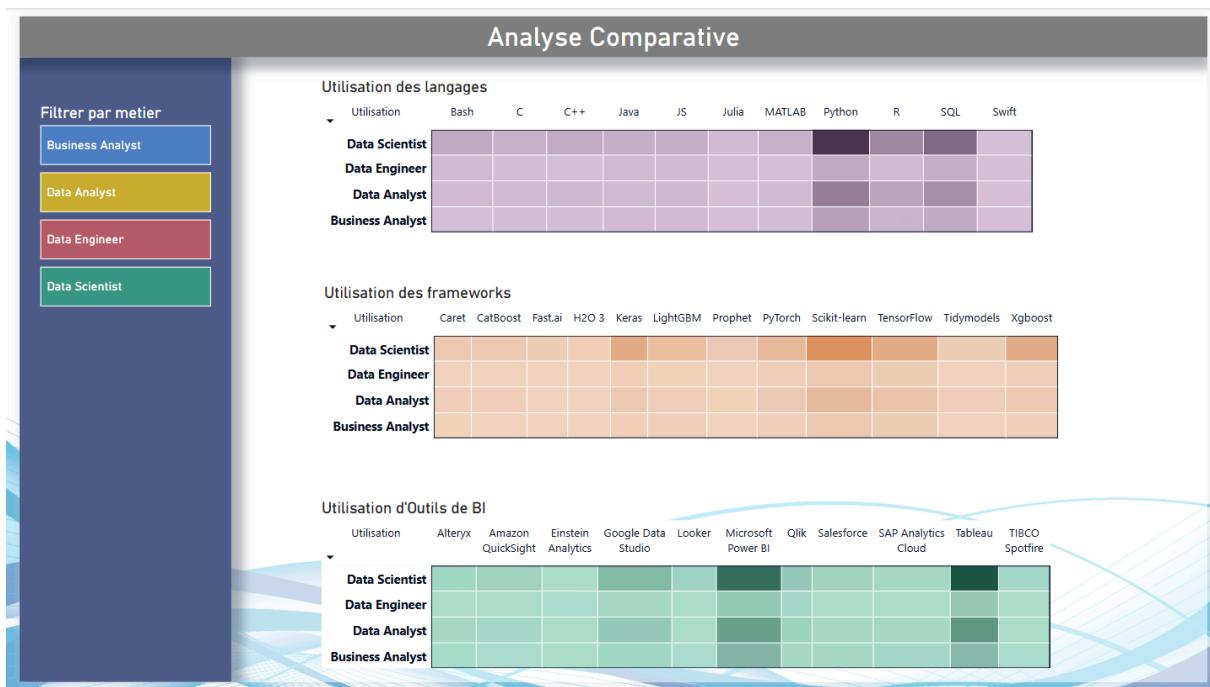
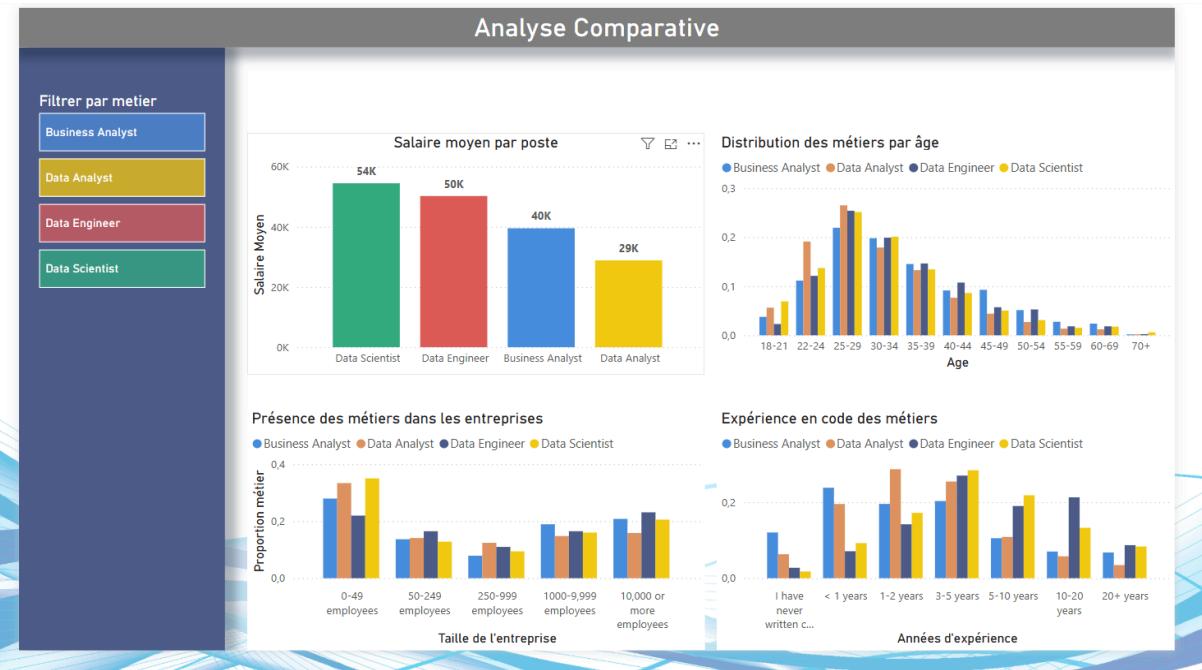
Cela traduit soit une réalité hybride des pratiques, soit une proximité croissante entre métiers, notamment dans des contextes où chacun doit pouvoir communiquer ses résultats de manière visuelle. On peut donc en déduire que ces outils, s’ils ne diffèrent pas les métiers en termes d’exclusivité, jouent un rôle transversal dans les équipes data.

Ces résultats illustrent la complémentarité des métiers : tous partagent un socle technique, mais se différencient par l’usage et la finalité des outils. Cela montre que la nature des tâches (modélisation, exploration, reporting, automatisation...) guide les choix d’outils plus que le métier lui-même.

Ce constat est à la base de notre idée de “CV de compétences” interactif par métier dans le tableau de bord.



## Promo avril 2025 Data Analytics

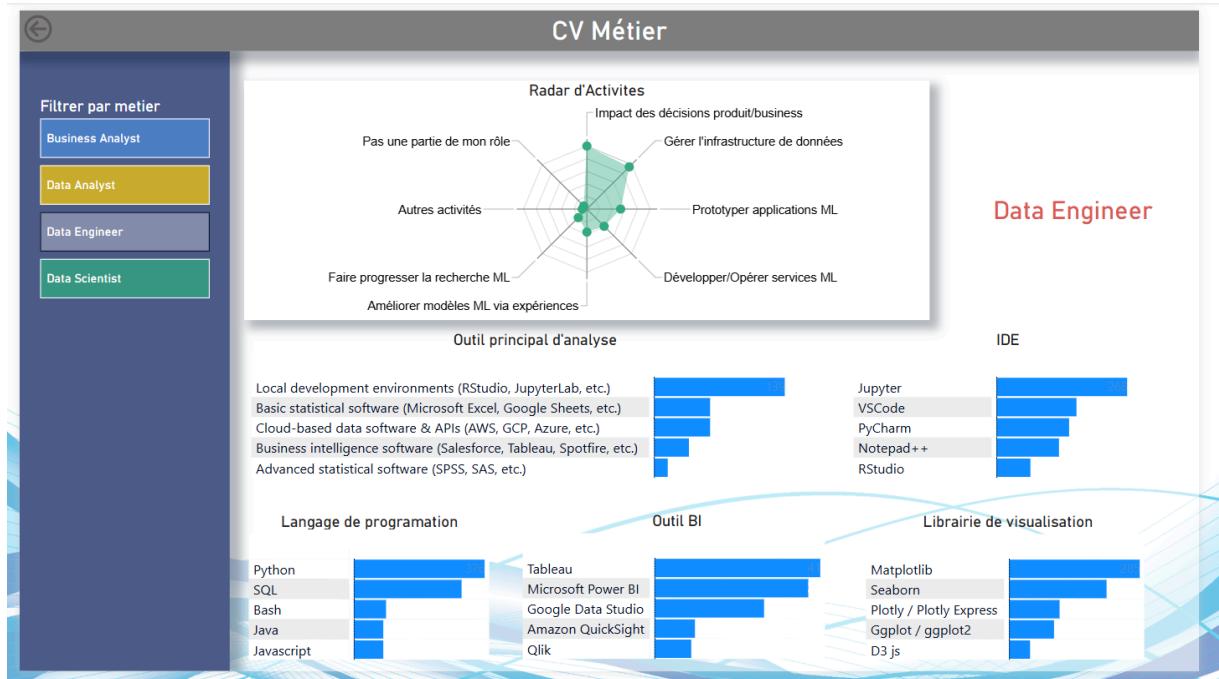




## CV par métier (avec sélection dynamique)

Cette page présente “un profil type” ou un CV de compétences pour chaque métier cible. Elle est interactive grâce à un filtre par métier.

- **Le radar des activités** montre la répartition des activités principales selon le métier sélectionné. Il permet de bien visualiser la spécialisation et la complémentarité de chaque métier.
- Les quatre métiers étudiés partagent un socle commun de langages et outils, avec cependant des spécificités propres à chacun :
  - **Langages de programmation :**  
Tous utilisent principalement Python et SQL. Le bash est plus fréquent chez les Data Engineers, tandis que le langage R est privilégié par les autres.
  - **Outils de Business Intelligence (BI) :**  
Des outils comme Power BI et Tableau sont très présents dans les quatre profils. Toutefois, ils sont priorisés différemment selon les métiers, ils sont davantage au cœur des pratiques des Business Analysts et Data Analysts,
  - **Librairies de visualisation :**  
Les quatre métiers utilisent des librairies communes telles que Matplotlib et Seaborn. Les analystes priviléguent ggplot2 (notamment avec R), tandis que les Data Scientists et Data Engineers utilisent souvent Plotly pour des visualisations.
  - **Environnements de développement (IDE) :**  
Jupyter Notebook et Visual Studio Code sont largement utilisés dans tous les profils. Les Data Scientists et Data Engineers préfèrent souvent PyCharm pour le développement Python, alors que les analystes priviléguent RStudio.
  - **Outils principaux d'analyse :**  
Les analystes s'appuient prioritairement sur des outils de statistique de base comme Excel, des environnements locaux comme RStudio, ainsi que des solutions BI (ex. Salesforce, Tableau). En revanche, les Data Scientists et les Data Engineers combinent usage d'environnements locaux, d'outils statistiques, puis de plateformes cloud et d'API (ex. AWS, Azure).



## Outils vs Besoins

Cette page du rapport compare les outils actuellement utilisés avec ceux jugés importants pour l'avenir, dans quatre domaines clés : Business Intelligence, Machine Learning, Big Data et automatisation du machine learning. Il convient de préciser que la hauteur des barres reflète le nombre de répondants, et que les deux séries ne sont pas directement comparables en termes absolus. Néanmoins, les écarts entre les barres permettent de dégager des tendances claires sur les évolutions attendues dans l'écosystème des outils data.

Dans la catégorie BI, des outils comme Power Bi et Tableau conservent une forte présence, et les autres outils ne montrent pas de changement de tendance significatif. En Machine Learning, les produits cloud (Google AI, Azure ML, Amazon SageMaker) affichent une hausse marquée des attentes, ce qui suggère une transition vers des solutions plus intégrées. Pour le Big Data, les outils comme PostgreSQL ou SQL Server, moins utilisés aujourd'hui, apparaissent comme des technologies à suivre selon les répondants. Enfin, l'automatisation du machine learning connaît une forte poussée d'intérêt futur, en particulier pour les plateformes de Google et Microsoft.

Si nous comparons seulement les besoins des étudiants:

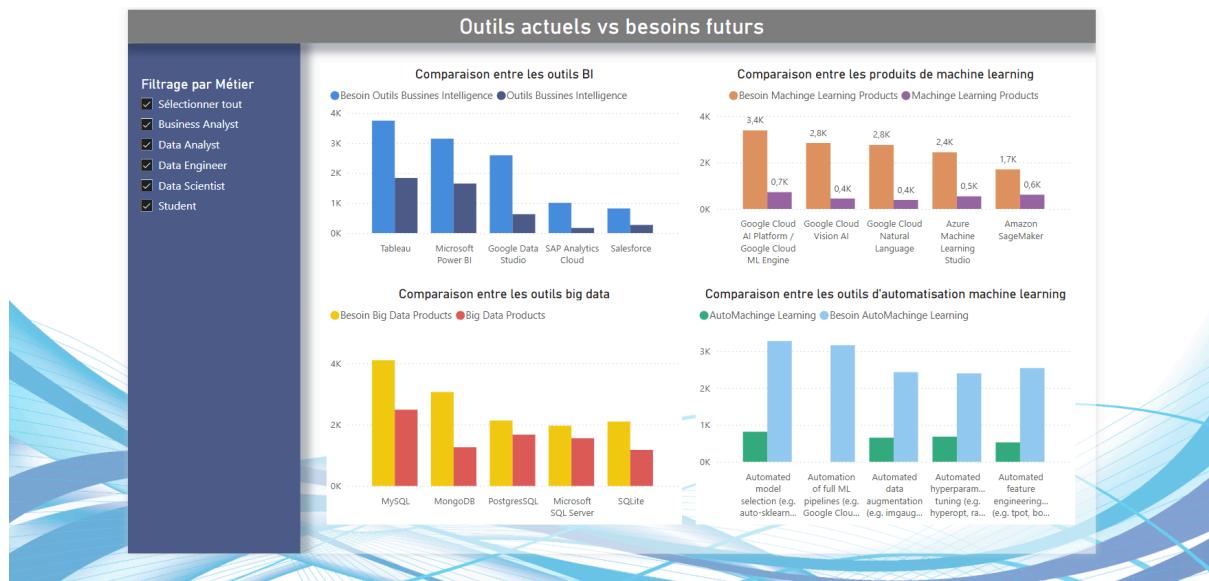
Les étudiants déclarent vouloir utiliser des outils plus avancés, parfois surreprésentés dans leurs attentes. En pratique, les professionnels utilisent surtout des outils robustes, simples et bien implantés.



Certains outils avancés comme AutoML sont très peu présents dans les entreprises, bien que fortement cités dans les aspirations étudiantes. Ce décalage peut refléter une surestimation de certains outils par les étudiants, ou un besoin d'ajustement des formations.

Les aspirations des étudiants ne seraient-elles pas les prémisses d'une transformation profonde du marché du travail Data ?

Ce tableau illustre donc des tendances nettes vers le cloud, l'automatisation et une diversification des outils, signes d'un secteur en mutation rapide.



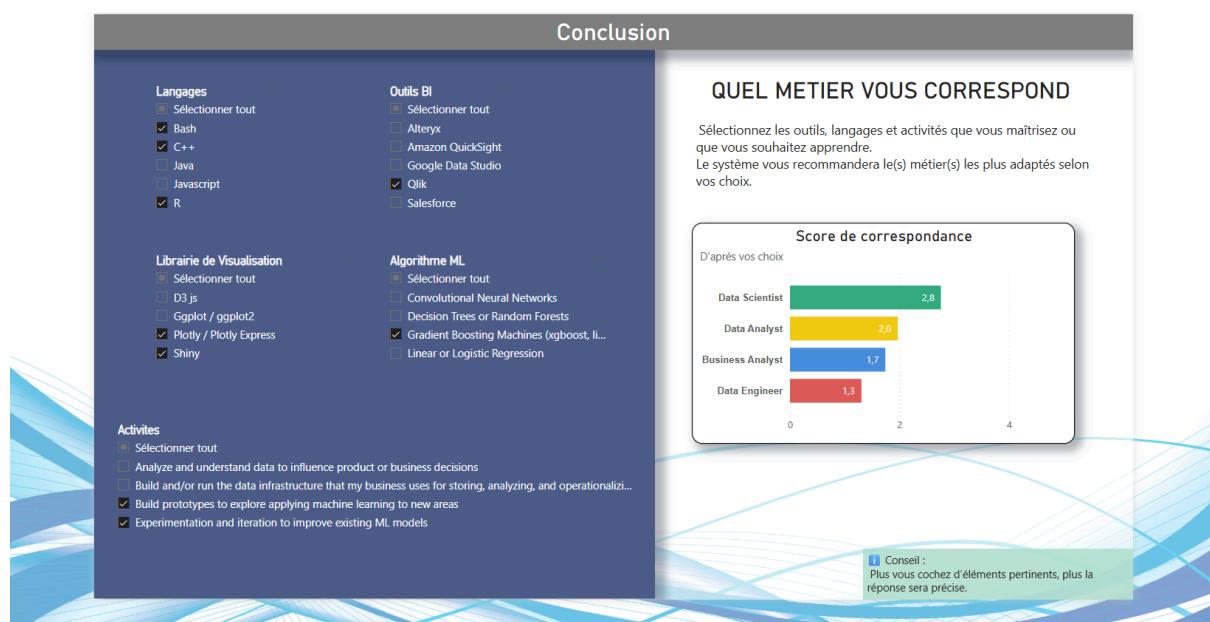


## Recommandation métier

Cette page propose un métier en fonction des outils, langages et activités sélectionnés par un utilisateur.

1. L'utilisateur sélectionne ses préférences via des slicers interactifs.
2. Un score de correspondance est calculé en fonction de la proximité avec les profils-types.
3. Le score est pondéré selon l'ordre de priorité des éléments.

En termes d'optimisation, le tableau a été fait afin d'être le plus simple possible à lire, le plus interactif. Il permet aux différents acteurs de naviguer aisément à travers les différents métiers grâce à des slicers synchronisés (métier, outils...).



## Interprétation des résultats

Ce projet nous a permis d'explorer en profondeur les métiers de la data science à travers le sondage Kaggle 2020. En centrant notre analyse sur quatre métiers cibles, **Data Analyst**, **Data Scientist**, **Data Engineer** et **Business Analyst**, nous avons pu identifier, comparer et visualiser les compétences, outils et trajectoires caractéristiques de chacun.

Le travail a combiné des étapes rigoureuses de prétraitement des données, de visualisations avancées et un effort de structuration métier. Nous avons mis en lumière :

- Les disparités de salaires et d'expériences entre les postes.
- Les socles de compétences communs à tous les métiers (Python, SQL).
- Les spécialisations propres à chaque rôle (outils BI, frameworks, tâches).
- Les radars d'activités permettent une lecture claire des points forts et des dominantes de chaque métier.
- Les écarts entre les attentes et les pratiques réelles.



Ce constat a justifié la création des pages CV-type et de recommandation de métier. La réalisation d'un simulateur de recommandation interactif marque l'aboutissement du projet. Il permet de transformer une exploration descriptive en outil d'aide

Bien que cette description mette en lumière des différences claires, il est important de noter que ces frontières restent souvent floues en pratique.

- Recoulements fonctionnels :

Par exemple, un Data Scientist peut tout à fait utiliser Excel ou des outils BI pour communiquer ses résultats à des équipes métier, tandis qu'un data analyste peut exploiter des solutions cloud pour traiter des volumes importants.

- Influence du contexte métier :

Le choix des outils dépend aussi de l'environnement de travail, de la taille de l'entreprise, ou de la culture technique. Dans une startup, les rôles sont souvent plus transverses, tandis que dans une grande entreprise, les tâches peuvent être plus spécialisées.

### Conclusion :

Le tableau de bord Power BI nous a permis de mieux comprendre les compétences et les trajectoires des professionnels de la data en 2020. Il offre une vision claire et interactive des profils types selon les métiers, et pose les bases d'un outil de recommandation qui pourrait guider aussi bien les étudiants que les recruteurs.

En ce sens, le projet répond pleinement à notre problématique initiale : rendre lisible un jeu de données complexe, en extraire des tendances utiles pour mieux comprendre et mieux éclairer les différents acteurs sur les profils types et les différents parcours dans la data.

Il est également important de noter que les données utilisées datent de 2020. Si elles offrent un aperçu des métiers de la data à cette période, certaines tendances, outils ou pratiques ont depuis évolué. Ces résultats doivent donc être considérés comme une base informative, à comparer, actualiser et compléter avec des données plus récentes.

L'analyse montre enfin que les outils seuls ne suffisent pas à définir un métier. Ce sont surtout leurs usages concrets et les tâches associées qui permettent de distinguer les profils.

En combinant analyse exploratoire, structuration métier et recommandation, ce projet démontre comment la data peut éclairer les métiers de la data eux-mêmes.