Université Cergy Paris

Diplôme universitaire : Data Analyst

UE 4 : Rédaction de rapport

7 septembre 2022

Haury Fabien

Étude des influences sociodémographiques et socioprofessionnelles sur le nombre de langages de programmation utilisés au quotidien par des personnes pratiquant la Data Science

Haury Fabien

7 septembre 2022

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction

1.1 Contexte

La notion de Data Science a été avancée dès 1974. Cependant, c'est à partir de 2003 que l'émergence de revues spécialisées dans ce domaine fait apparaître la Data Science telle que nous la définissons aujourd'hui [1]. Depuis lors, la Data Science est devenue un domaine d'étude distinct dont l'évolution est constante. Comme le souligne une étude faite par le site KDnuggets(2019) [2], les outils employés sont également en constante évolution, ou de nouveaux outils créés si besoin.

1.2 Connaissances manquantes

Une majorité des études ayant pour principal sujet d'études les outils utilisés dans les sciences des données se focalise uniquement sur les outils eux-mêmes, par exemple l'évolution des différents langages de programmation utilisés, etc. Rarement, une étude se concentre sur ces outils et leurs liens avec des facteurs sociodémographiques (genre, etc.) ou socioprofessionnels (salaire, etc.).

Répondre à ces questions nous permettrait de voir l'évolution des outils sous une nouvelle perspective. Cela permettrait aussi d'établir une cartographie globale et donc d'utiliser les bonnes variables pour des futures études, de détecter de nouvelles tendances, etc.

1.3 Question

Quels sont les facteurs sociodémographiques ou socioprofessionnels qui influent sur le nombre de langages de programmation utilisés dans la Data Science ?

1.4 Méthodes utilisées

L'étude se fera en suivant une logique exploratrice. Pour répondre à ces interrogations, les données seront transformées pour être exploitables, une utilisation d'outil statistique sera faite (e.g chi2, anova, etc.). Les données sont de type qualitative et quantitative.

Chapitre 2

Méthodologie

2.1 Origine des données

Chaque année depuis 2017, kaggle.com propose un questionnaire à ses utilisateurs sur le monde des Data Sciences et du Machine Learning. Les jeux de données mobilisées pour cette étude sont issus des questionnaires proposés en 2018, 2019, 2020 et 2021 [3] [4] [5] [6]. Ces données ont été publiées sous licence CC 2.0 [7].

2.2 Transformations réalisées sur les données

Une fusion des quatre jeux de données est effectuée. Afin de pouvoir garder une distinction entre ces quatre jeux de données, l'année de chaque questionnaire est introduite par une nouvelle variable nommée *Annee*.

La variable Q1 représente l'âge des répondants et est normalisée entre les quatre années pour obtenir des groupes d'âge cohérents entre les quatre années.

La variable Q2 correspond au genre des répondants. Les personnes étant non-binaires, ne préférant pas répondre ou bien préférant s'autodécrire seront recodés comme étant *Autres* dû au fait de leur taille d'échantillon trop petit.

La variable Q3 correspond au pays des personnes interrogées. Il y a 69 pays différents, ceux-ci seront regroupés par leurs continents respectifs dans une nouvelle variable nommée *Continent*, soit sept continents au total : Amérique du Nord, Amérique du Sud, Europe, Asie, Afrique, Océanie et Moyen-Orient. Cette transformation a pour but de clarifier et de tenir compte des différences de taille de l'échantillon entre ces pays. Certains pays seront également renommés à des fins de clarté et de lisibilité et la variable Q3 sera conservée.Les répondants ne souhaitant pas déclarer leur pays seront recodés comme étant *Inconnu*

La variable Q4 correspond au niveau d'éducation des répondants.Les personnes ayant indiqué qu'elles ne préfèrent pas répondre, n'ayant aucune formation au-delà du lycée ou bien ayant suivi une licence ou équivalente sans valider ce diplôme seront regroupées dans une nouvelle catégorie nommée *Autre* pour plus de clarté.

La variable Q5 correspond au secteur d'activité des personnes sondées. Certaines réponses seront groupées pour en assurer la clarté. Le regroupement sera fait de façon à conserver au maximum le domaine d'activité (i.e ingénieurs regroupés ensemble, etc.). Le résultat de ces regroupements se fera dans une nouvelle variable nommée *Role* et la colonne originale sera conservée.

La variable Q6 correspond au salaire annuel des personnes interrogées. Les valeurs sont uniformisées afin de garantir des échelles salariales uniformes entre les quatre années. Les répondants ne souhaitant pas déclarer leurs revenus seront recodés comme étant *Inconnu*.

Les variables portant sur les langages de programmation couramment utilisés sont recodées en 0 ou 1, ensuite une somme sur chaque ligne est effectuée afin d'avoir le nombre total de langages de programmation utilisé par répondant. Cette nouvelle information est gardée dans une nouvelle variable appelée *Langage*.

Une recherche des valeurs aberrantes est effectuée, suivie de leurs éliminations, afin d'éviter toute inter-

férence avec les outils statistiques.

2.3 Outils mobilisés

L'étude sera faite en utilisant le langage R sous RStudio. Liste des librairies utilisées :

- plyr
- tidyverse
- scales
- ggthemes
- finalfit
- summarytools
- crosstable
- flextable
- vcd
- ggsankey
- ggalluvial
- ggpubr
- nortest
- xtable
- clipr

Une utilisation de Libre Office sera faite pendant l'étude.
Des fonctions ont été créées.

Chapitre 3

Résultats

3.1 Introduction des résultats

Un test de normalité nous montre que la variable *Langage* ne suit pas une distribution normale. En moyenne, les répondants utilisent 2,2 langages de programmation ± 1,5, sont des hommes âgés de 18 à 29 ans ayant un master, travaillant dans le secteur des sciences des données, résidant en Asie et percevant un salaire annuel compris entre zéro et 10K dollars.

3.2 Diagramme de Sankey

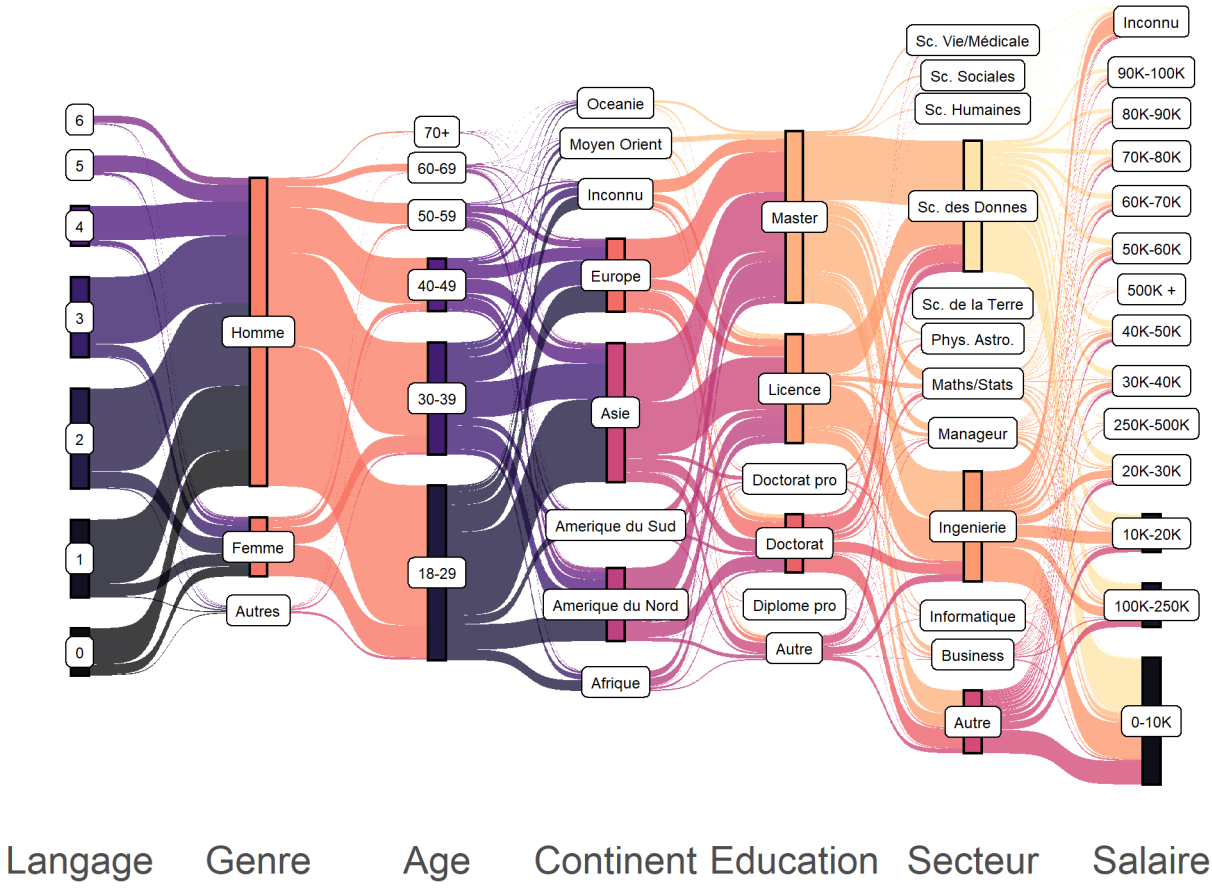

FIGURE 1 – Diagramme de Sankey pour l'ensemble des variables mobilisées

La Figure 1 représente un diagramme de Sankey de l'ensemble des variables mobilisées. Les diagrammes de Sankey sont un type de diagramme de flux dans lequel la largeur des flèches est proportionnelle au débit. Nous pouvons voir que l'utilisation d'un, deux ou trois langages de programmation est majoritaire avec 20,8 %, 26,8 % et 21,6 % respectivement. Les hommes représentent la majorité des répondants avec 82,6 %, Les 18-29 ans représentent avec 46,7 % le plus gros groupe et les 30-39 ans avec 29,9 % le second groupe, ce qui montre bien la jeunesse des métiers des données. Plus d'un tiers des répondants résident en Asie (37,2 %). Environ 46 % ont obtenu un master. Un peu plus d'un tiers travaille dans la Data Science (35,1 %). Et un tiers (33,9 %) gagne entre 0 et 10 000 dollars par an. La Figure 10 en annexe A représente un diagramme alluvial, montrant le cheminement de chaque valeur unique de *Langage* à travers l'ensemble des variables indépendantes.

3.3 Test de normalité de la variable langage par QQ plot

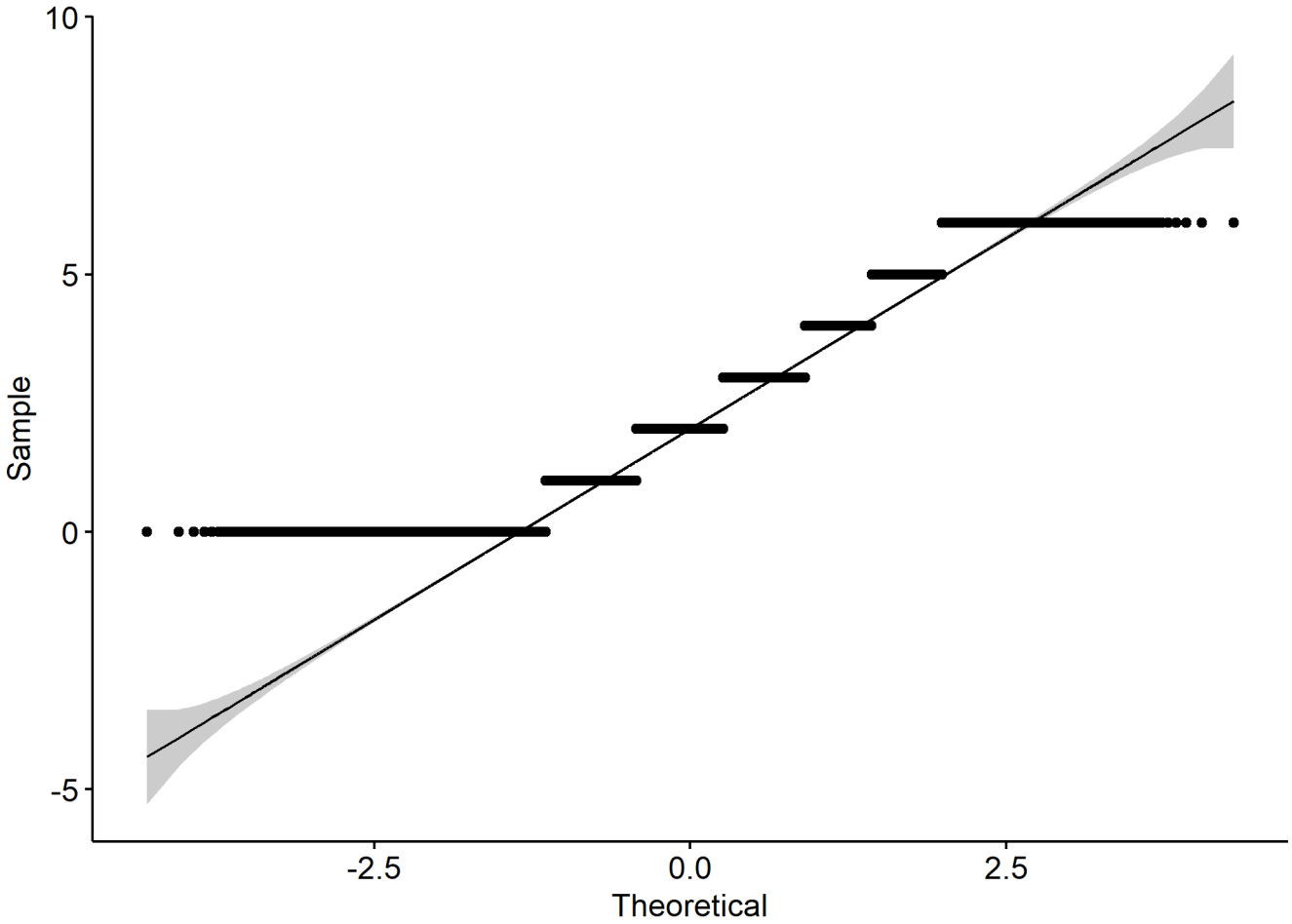

FIGURE 2 – QQ plot de la variable langage pour tester sa normalité

La Figure 2 représente un QQ plot de la variable *Langage* afin de tester sa normalité. Nous retrouvons les sept catégories de la variable *Langage*. Il se décompose en trois parties. La première allant du quantile -4 au quantile -1 montre un étalement et un éloignement des points par rapport à la droite théorique, cela signifie une distribution fortement biaisée vers la gauche. La deuxième partie partant du quantile -1 au quantile 2 montre un ensemble de points au voisinage de la droite théorique, mais sans être sur celle-ci, indiquant que ces points ne sont pas normalement distribués. Enfin, la troisième partie allant du quantile 2 au quantile 4 montre un étalement et un éloignement des points par rapport à la droite théorique, indiquant une distribution fortement biaisée vers la droite.

3.4 Nombre de langages de programmation utilisés au quotidien par les répondants

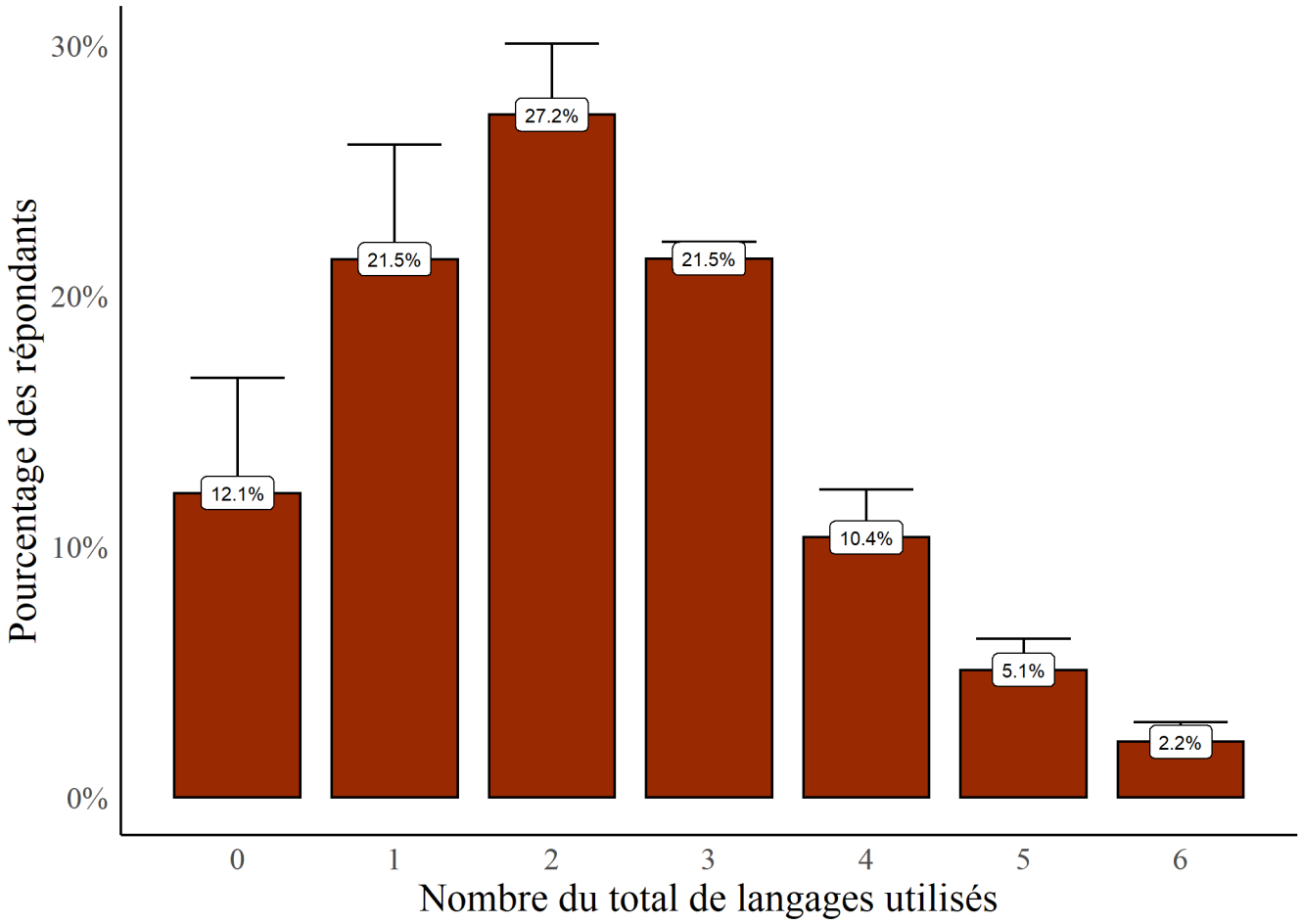

FIGURE 3 – Distribution avec écart-type du nombre total de langages de programmation utilisés au quotidien par les répondants pour l'ensemble des années 2018 à 2021

La Figure 3 représente la distribution du nombre de langages de programmation utilisés pour l'ensemble des années allant de 2018 à 2021, avec leurs écart-types respectifs. Le nombre de répondants qui utilisent zéro langage de programmation au quotidien représente 12,1 % ± 4,7 %. Le nombre de répondants qui utilisent un langage de programmation au quotidien représente 21,5 % ± 4,5 %. Le nombre de répondants qui utilisent deux langages de programmation au quotidien représente environ un quart du total des répondants avec 27,2 % ± 2,8 %. Le nombre de répondants qui utilisent trois langages de programmation au quotidien représente 21,5 % ± 0,7 %. Le nombre de répondants qui utilisent quatre langages de programmation au quotidien représente 10,4 % ± 1,9 %. Le nombre de répondants qui utilisent cinq langages de programmation au quotidien représente 5,1 % ± 1,2 %. Le nombre de répondants qui utilisent six langages de programmation au quotidien représente 2,2 % ± 0,8 %.

3.5 Âge des répondants

3.5.1 Distribution des tranches d'âge des répondants

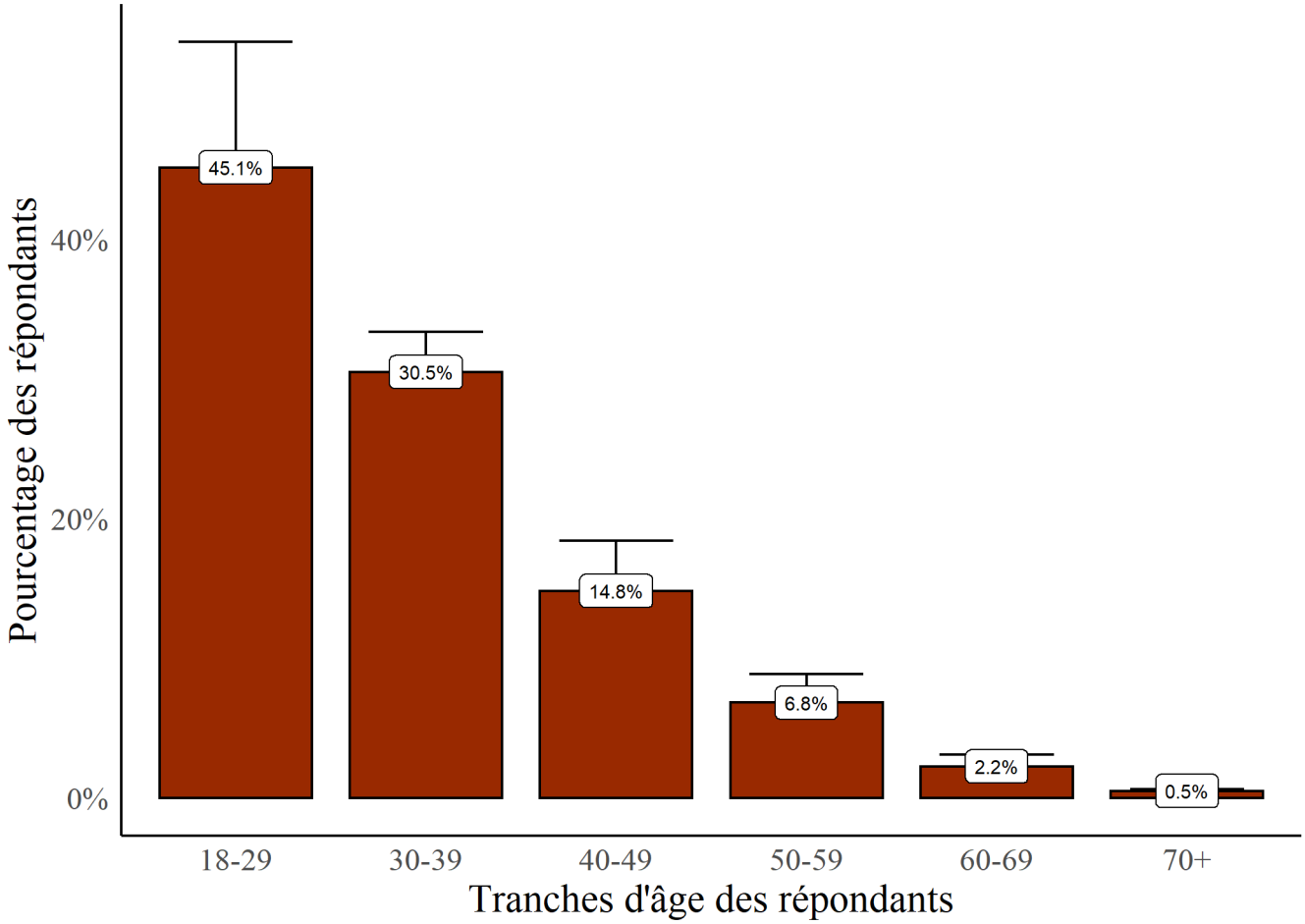

FIGURE 4 – Distribution avec écart-type des tranches d'âge des répondants pour l'ensemble des années 2018 à 2021

La Figure 4 représente la distribution des tranches d'âge des répondants pour l'ensemble des années allant de 2018 à 2021, avec leurs écart-types respectifs. Le nombre de répondants entre 18 et 29 ans représente 45,1 % ± 9 %. Le nombre de répondants entre 30 et 39 ans représente 30,5 % ± 2,9 %. Le nombre de répondants entre 40 et 49 ans représente 14,8 % ± 3,8%. Le nombre de répondants entre 50 et 59 ans représente 46,8 % ± 2,1 %. Le nombre de répondants entre 60 et 69 ans représente 2,2 % ± 0,8 %. Le nombre de répondants ayant 70 ans ou plus ans représente 0,5 % ± 0,02 %.

3.5.2 Table de contingence

La Table 1 représente les proportions des tranches d'âges des répondants par nombre de langages de programmation utilisés quotidiennement.

Les 18-29 ans et les 30-39 ans comptent pour plus des trois quarts des répondants, soit 26 922 (46,7 %) et 17 249 (29,9 %) respectivement. Les 40-49 ans représentent un total de 8 200 répondants (14,2 %). Le nombre de répondants âgés de 50 à 59 ans s'élève à 3 739 (6,5 %). Les personnes de 60 à 69 ans représentent 1 221 (2,1 %). Enfin, les répondants de 70 ans et plus sont au nombre de 274 (0,5 %).

TABLE 1 – Table des pourcentages des tranches d'âges des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Age	18-29	13,8	20,4	26,9	21,1	10,4	5,1	2,3	26 922
	30-39	11,0	21,2	28,1	22,6	10,7	4,4	1,9	17 249
	40-49	1147	20,6	25,7	21,7	11,4	6,4	2,8	8 200
	50-59	12,67	20,8	24,0	20,5	11,2	7,4	3,4	3 739
	60-69	13,2	23,3	22,2	21,5	11,2	5,8	2,8	1 221
	70+	21,5	27,0	20,1	13,8	10,6	4,4	2,6	274

3.6 Genre des répondants

3.6.1 Distribution du genre des répondants

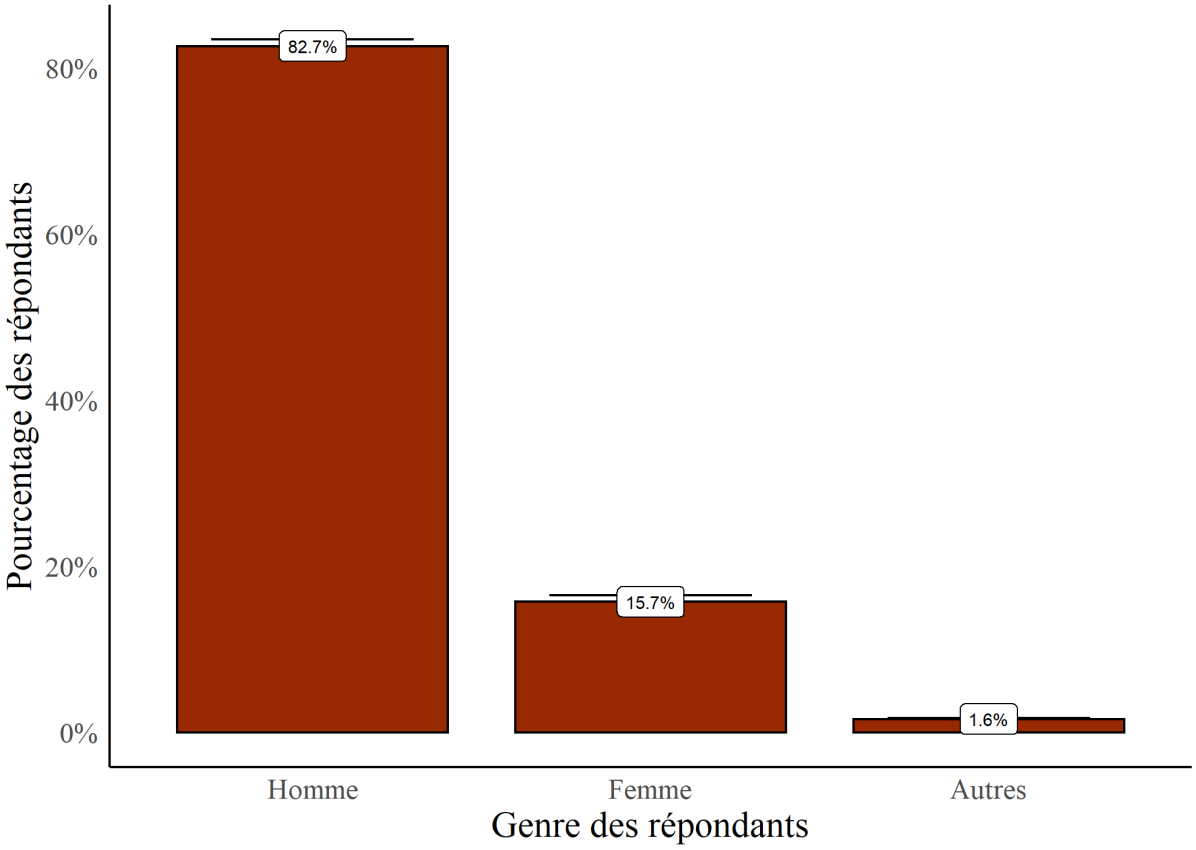

FIGURE 5 – Distribution avec écart-type du genre des répondants pour l'ensemble des années 2018 à 2021

La Figure 5 représente la distribution des genres des répondants pour l'ensemble des années allant de 2018 à 2021, avec leurs écart-types respectifs. Le pourcentage de répondants se déclarant comme homme est 82,7 % ± 0,8 %. Le pourcentage de répondants se déclarant comme femme est 15,7 % ± 0,7 %. Le pourcentage de répondants se déclarant comme autre est 1,6 % ± 0,05 %.

3.6.2 Table de contingence

TABLE 2 – Table des pourcentages du genre des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Genre	Autres	15,8	17,7	22,4	20,5	13,1	7,1	3,5	917
	Femme	17,1	20,6	26,5	20,9	8,6	4,4	1,9	9 117
	Homme	11,7	20,9	26,9	21,7	11,0	5,4	2,4	47 571

La Table 2 représente les proportions du genre des répondants par nombre de langages de programmation utilisés quotidiennement.

Une majorité des répondants s'identifient comme étant des hommes avec 47 571 (82,6 %) réponses. Le nombre de personnes qui se sont identifiées comme étant des femmes est de 9 117 (15,8 %). Enfin, le nombre de répondants s'identifiant comme autres est de 917 (1,6 %).

3.7 Niveau de scolarité des répondants

3.7.1 Distribution du niveau de scolarité des répondants

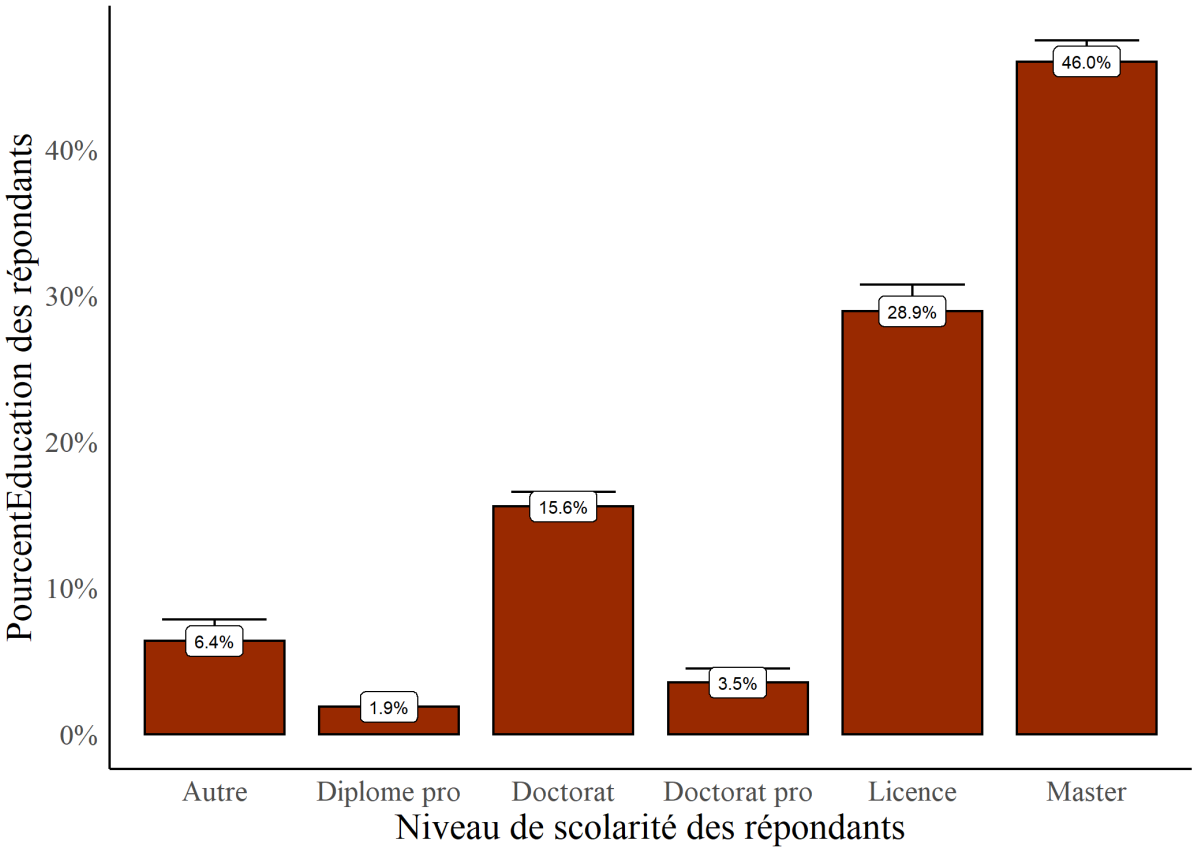

FIGURE 6 – Distribution avec écart-type du niveau de scolarité des répondants pour l'ensemble des années 2018 à 2021

La Figure 6 représente la distribution des niveaux de scolarité des répondants pour l'ensemble des années 2018 à 2021, avec leurs écart-types respectifs. Le pourcentage de répondants déclarant avoir obtenu un diplôme autre est de 6,4 % ± 1,5 %. Le pourcentage de répondants déclarant avoir obtenu un diplôme professionnel est de 1,5 %. Le pourcentage de répondants déclarant avoir obtenu un doctorat est de 15,6 % ± 1 %. Le pourcentage de répondants déclarant avoir obtenu un doctorat professionnel est de 3,5 % ± 0,9 %. Le pourcentage de répondants déclarant avoir obtenu une licence est de 28,9 % ± 1,8 %. Le pourcentage de répondants déclarant avoir obtenu un master autre est de 46 % ± 1,9 %.

3.7.2 Table de contingence

TABLE 3 – Table des pourcentages du niveau de scolarité des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Education	Autre	19,2	21,9	23,3	19,6	8,8	4,9	2,4	3657
	Diplôme pro	6,4	31,4	20,5	19,4	14,1	4,2	3,9	283
	Doctorat	9,2	22,8	26,9	21,7	11,4	5,4	2,5	8 915
	Doctorat pro	15,4	20,0	22,9	20,7	11,4	7,0	2,6	1 408
	Licence	14,2	20,6	26,9	20,6	10,3	5,2	2,2	16 791
	Master	11,8	20,0	27,4	22,5	10,9	5,2	2,3	26 551

La Table 3 représente les proportions du niveau de scolarité des répondants par nombre de langages de programmation utilisés quotidiennement.

Près de la moitié des répondants possède un master, 26 551 (46,1 %), suivi de la licence avec 16 791 (29,1 %). 8 915 (15,5 %) des répondants possèdent un doctorat et 1 408 (2,4 %) possède un doctorat professionnel. 283 (0,5 %) d'entre eux possèdent un diplôme professionnel et enfin, 3 657 (46,3 %) n'ont pas souhaité répondre.

3.8 Secteur d'activité des répondants

3.8.1 Distribution des secteurs d'activité des répondants

La Figure 7 représente la distribution des secteurs d'activité des répondants pour l'ensemble des années 2018 à 2021, avec leurs écart-types respectif. Le pourcentage de répondants déclarant travailler dans un secteur autre est de 18,4 % ± 4,9 %. Le pourcentage de répondants déclarant travailler dans le secteur business est de 8,0 %. Le pourcentage de répondants déclarant travailler dans le secteur de l'informatique est de 4,4 %. Le pourcentage de répondants déclarant travailler dans le secteur de l'ingénierie est de 26 % ± 20,5 %. Le pourcentage de répondants déclarant travailler dans le secteur managérial est de 5,9 % ± 0,9 %. Le pourcentage de répondants déclarant travailler dans le secteur des Mathématiques ou Statistiques est de 4,8 % ± 11 %. Le pourcentage de répondants déclarant travailler dans le secteur de la Physique ou Astronomie est de 5,5 %. Le pourcentage de répondants déclarant travailler dans le secteur des Sciences de la Terre est de 11,1 % ± 1,2 %. Le pourcentage de répondants déclarant travailler dans le secteur des Sciences des Données est de 53 %. Le pourcentage de répondants déclarant travailler dans le secteur des Sciences Humaines est de 1,2 %. Le pourcentage de répondants déclarant travailler dans le secteur des Sciences Sociales est de 2,4 %. Le pourcentage de répondants déclarant travailler dans le secteur des Sciences de la vie ou médicale est de 53,9 %.

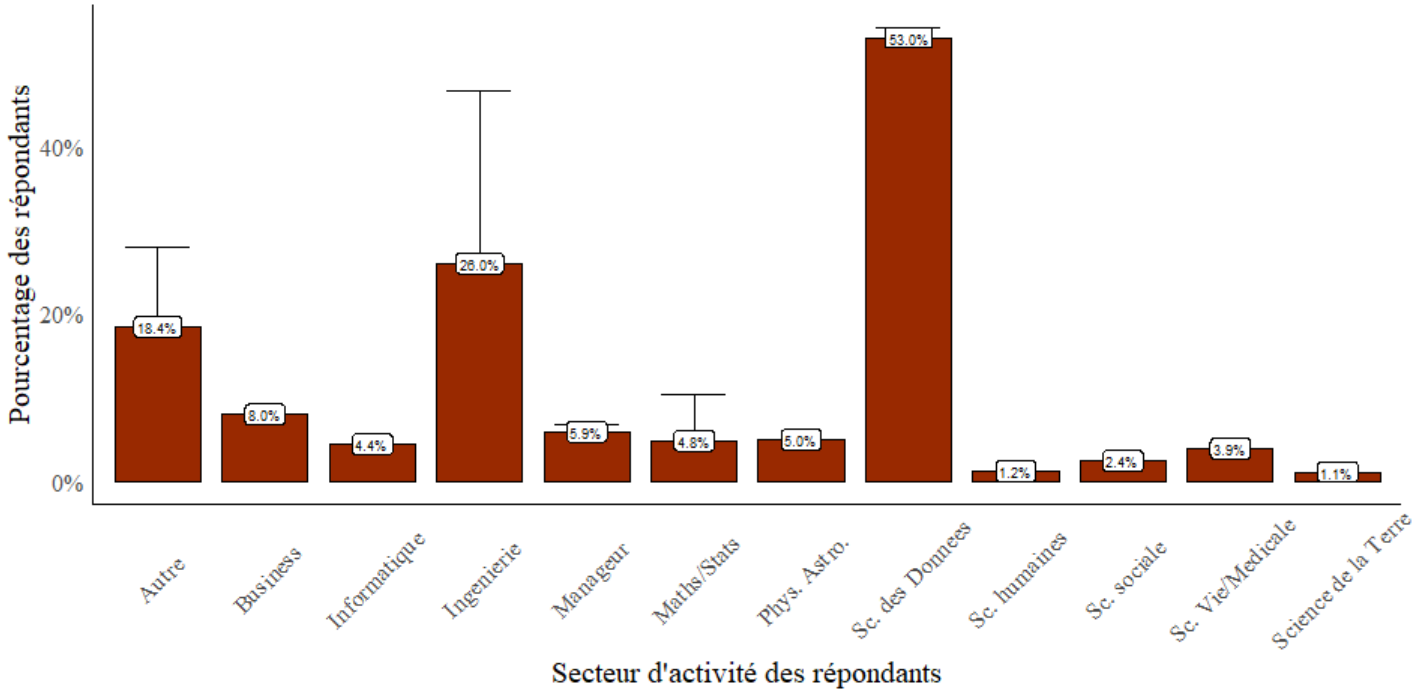

FIGURE 7 – Distribution avec écart-type du secteur d'activité des répondants pour l'ensemble des années 2018 à 2021

3.8.2 Table de contingence

La Table 4 représente les proportions du secteur d'activité des répondants par le nombre de langages de programmation utilisés quotidiennement. La variable *Nombre de langages de programmation* représente le nombre de langages de programmation utilisés au quotidien. La variable *Rôle* correspond au domaine professionnel des répondants.

Nous pouvons voir trois grands rôles qui comptent pour plus des quatre cinquièmes du total avec 46 900 (81.4 %) réponses. Le premier est celui des sciences de la donnée avec 20 223 (35,1 %) répondants se déclarant de ce domaine professionnel. Le second est celui du domaine de l'ingénierie avec un total de 16 996 (29,5 %) personnes déclarées. Le troisième est celui nommé autre avec 9 681 (16,8 %) répondants du total. Les sciences humaines, sociales, de la vie, médicales et de la terre représente 1 697 (2,9 %) du total des répondants. Les mathématiques, les statistiques, la physique et l'astronomie représente 4 305 (7,5 %) du total. Les rôles liés au business et aux manageurs représente 3 844 (6,7 %) du total. Les rôles liés à l'informatique représentent 859 (1,5 %) du total.

TABLE 4 – Table des pourcentages des secteurs d'activités des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Secteur d'activité	Autre	16,2	28,6	25,2	16,7	7,7	3,7	1,8	9 681
	Business	16,6	17,5	26,2	23,7	9,5	4,9	1,7	1 551
	Informatique	17,2	15,2	22,1	21,4	12,8	6,3	4,9	859
	Ingenierie	14,5	14,6	23,0	22,0	14,2	8,0	3,7	16 996
	Manageur	15,9	25,3	24,3	19,3	8,5	5,1	1,6	2 293
	Maths/Stats	14,3	16,6	25,9	23,4	12,6	5,0	2,1	3 338
	Phys. Astro.	10,3	17,4	26,7	24,5	12,7	5,0	3,4	967
	Sc. des Donnees	8,1	23,1	31,5	23,1	9,0	3,7	1,5	20 223
	Sc. humaines	16,0	18,5	25,2	22,3	13,4	3,8	0,8	238
	Sc. sociale	12,8	20,2	24,4	24,6	11,5	4,4	2,1	476
	Sc. Vie/Medicale	13,6	20,9	24,1	22,7	10,2	5,8	2,7	763
	Science de la Terre	12,7	19,5	28,1	19,5	12,7	5,4	1,8	220

3.9 Continent de résidence des répondants

3.9.1 Distribution des continents de résidence des répondants

La Figure 8 représente la distribution des continents de résidence des répondants pour l'ensemble des années 2018 à 2021, avec leurs écart-types respectif. Le pourcentage de répondants déclarant résider en Afrique est de 5,2 % ± 2,3 %. Le pourcentage de répondants déclarant résider en Amérique du Nord est de 19,2 % ±4,2 %. Le pourcentage de répondants déclarant résider en Amérique du Sud est de 6,5 % ± 1 %. Le pourcentage de répondants déclarant résider en Asie est de 37,3 % ± 2,9 %. Le pourcentage de répondants déclarant résider en Europe est de 19,4 % ± 2,6 %. Le pourcentage de répondants dont le continent de résidence est inconnu est de 7,5 % ± 0,9 %. Le pourcentage de répondants déclarant résider en Moyen-Orient est de 3,3 % ± 0,8 %. Le pourcentage de répondants déclarant résider en Océanie est de 1,6 % ± 0,04 %.

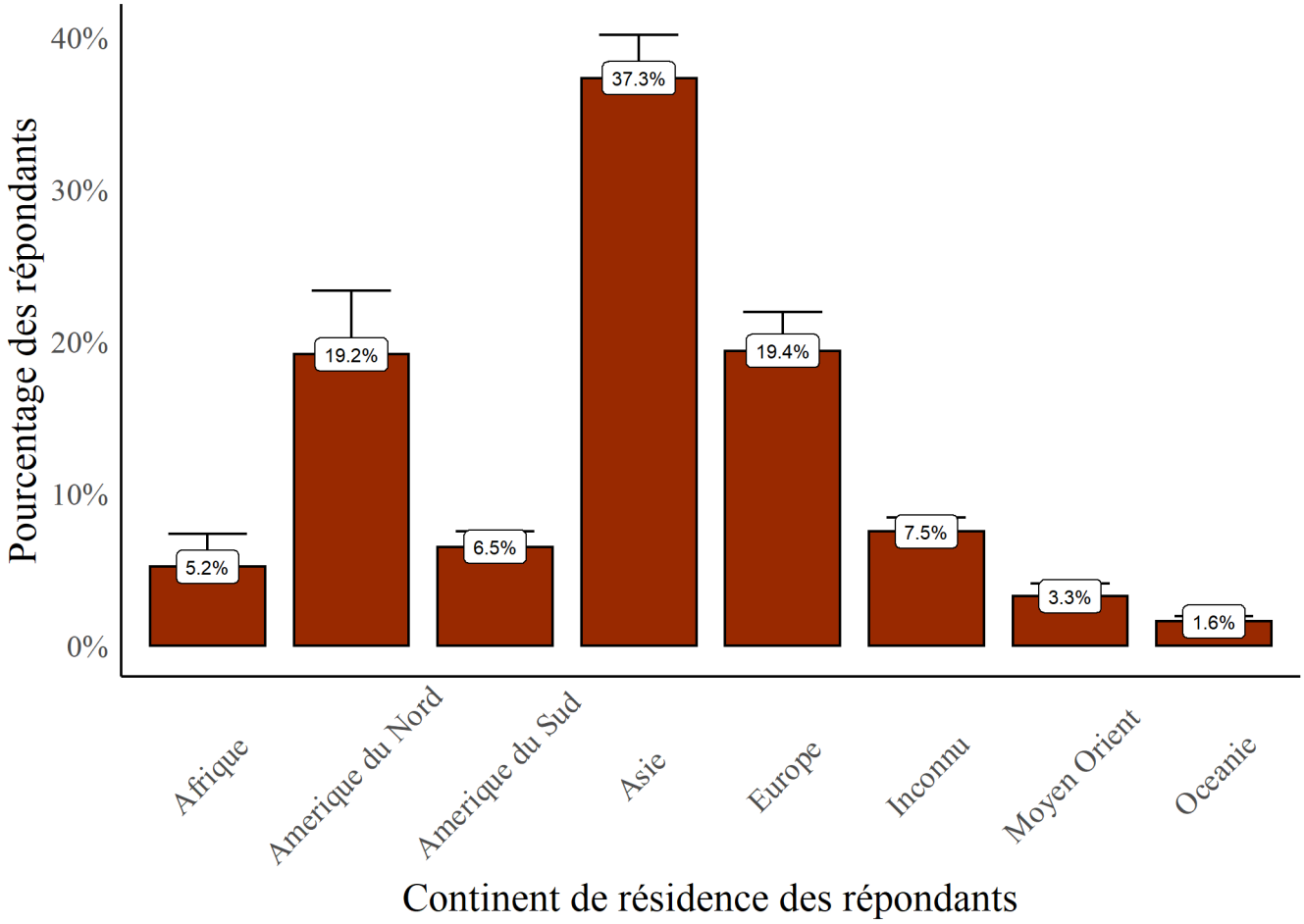

FIGURE 8 – Distribution avec écart-type des continents de résidence des répondants pour l'ensemble des années 2018 à 2021

3.9.2 Table de contingence

TABLE 5 – Table des pourcentages des continents de résidence des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Continent	Afrique	15,0	26,6	25,2	17,3	8,5	4,4	2,9	2 884
	Amérique du Nord	12,2	16,4	26,4	24,5	12,4	5,8	2,3	11 319
	Amérique du Sud	10,4	17,3	27,2	23,2	12,4	6,7	2,8	3 641
	Asie	13,5	22,7	26,8	20,4	9,5	4,7	2,2	21 430
	Europe	10,4	19,9	28,3	22,7	11,3	5,3	2,2	11 309
	Inconnu	14,9	23,5	23,6	18,8	10,1	6,0	3,0	4 252
	Moyen-Orient	13,3	23,1	27,7	18,9	9,7	5,0	2,3	1 843
	Océanie	12,7	18,1	26,1	22,8	13,8	4,4	2,0	927

La Table 5 représente les proportions du continent de résidence des répondants par nombre de langages de programmation utilisés quotidiennement.

L'Asie compte pour plus du tiers du total avec 21 430 (37,2 %) répondants. L'Amérique du Nord et l'Europe avec 11 319 (19,6 %) et 11 306 (19,6 %) réponses respectivement représentent les deux prochains pôles. Le nombre de personnes n'ayant pas souhaité répondre est de 4 252 (7,4 %). L'Amérique du Sud, l'Afrique, Moyen-Orient et l'Océanie avec 3 641 (6,3 %), 2 884 (5 %), 1 843 (3,2 %) et 927 (1,6 %) réponses respectivement, sont les continents minoritairement représentés.

3.10 Salaire annuel des répondants

3.10.1 Distribution des salaires annuels des répondants

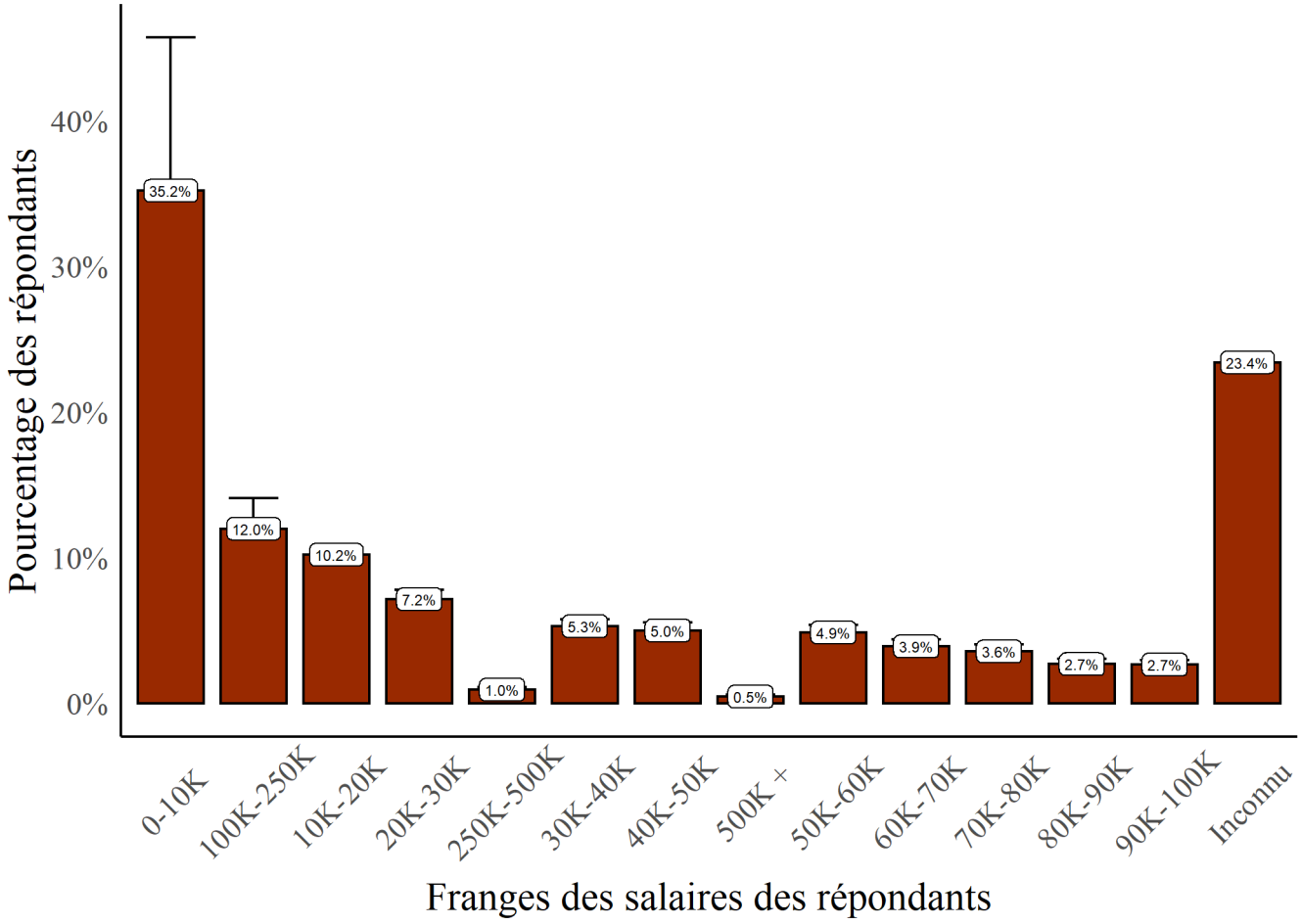

FIGURE 9 – Distribution avec écart-type des salaires annuels des répondants pour l'ensemble des années 2018 à 2021

La Figure 8 représente la distribution des salaires annuels des répondants pour l'ensemble des années 2018 à 2021, avec leurs écart-types respectif. Le pourcentage de répondants déclarant obtenir entre zéro et 10K dollars est de 35,2 % ± 10,5 %. Le pourcentage de répondants déclarant obtenir entre 100k et 250K dollars est de 12 % ± 2,1 %. Le pourcentage de répondants déclarant obtenir entre 10K et 20K dollars est de 10,2 % ± 0.6 %. Le pourcentage de répondants déclarant obtenir entre 20K et 30K dollars est de 7,2 % ± 0,7 %. Le pourcentage de répondants déclarant obtenir entre 250K et 500K dollars est de 1 % ± 0,02 %. Le pourcentage de répondants déclarant obtenir entre 30K et 40K dollars est de 5,3 % ± 0,45 %. Le pourcentage de répondants déclarant obtenir entre 40K et 50K dollars est de 5 % ± 0,6 %. Le pourcentage de répondants déclarant obtenir 500K dollars ou plus est de 0,5 % ± 0,02 %. Le pourcentage de répondants déclarant obtenir entre 50K et 60K dollars est de 4,9 % ± 0,5 %. Le pourcentage de répondants déclarant

obtenir entre 60K et 70K dollars est de 3,9 % ± 0,5 %. Le pourcentage de répondants déclarant obtenir entre 70K et 80K dollars est de 3,6 % ± 0,6 %. Le pourcentage de répondants déclarant obtenir entre 80K et 90K dollars est de 2,7 % ± 0,4 % Le pourcentage de répondants déclarant obtenir entre 90K et 100K dollars est de 2,7 % ± 0,4 %. Le pourcentage de répondants dont le salaire est inconnu est de 23,4 %.

3.10.2 Table de contingence

La Table 6 représente les proportions du salaire des répondants par le nombre de langages de programmation utilisés quotidiennement.

La frange 0-10K représente le plus gros groupe avec 19 521 (33,9 %). La frange 100k-250k représente le second groupe avec 6 728 (11,7 %) du total des répondants. La frange 10K-20K représente le troisième groupe avec 5 831 (10,1 %) répondants. Les franges 20K-30K et inconnu représentent avec 4 119 (7,2 %) et 4 555 (7,9 %) respectivement le groupe suivant. Les franges 30K-40K, 40K-50K et 50K-60K représentent avec 3 066 (5,3 %), 2 858 (5 %) et 2 783 (4,8 %) respectivement le groupe suivant. Les franges 60K-70K et 70K-80K représentent avec 2 235 (3,9 %) et 2 025 (3,5 %) respectivement le groupe suivant. Les franges 80K-90K et 90K-100K représentent avec 1 160 (2,7 %) et 1 539 (2,7 %) respectivement le groupe suivant. La frange 250K-500K représente 539 (0,9 %) du total des répondants. Enfin, avec 261 (0,5 %) du total des répondants, la frange 500K+.

TABLE 6 – Table des pourcentages des franges des salaires des répondants en fonction du nombre de langages de programmation utilisés, N = 57 605

		Nombre de langages de programmation							N_{ligne}
		0	1	2	3	4	5	6	
Salaire	0-10K	14,9	23,9	25,7	19,1	9,3	4,8	2,3	19 521
	100K-250K	8,1	17,0	27,2	25,6	13,1	6,3	2,7	6 728
	10K-20K	11,5	22,4	28,1	21,6	9,8	4,6	2,0	5 831
	20K-30K	10,8	19,6	29,1	22,9	10,5	5,0	2,1	4 119
	250K-500K	11,1	16,0	24,3	23,0	13,5	8,7	3,3	539
	30K-40K	9,8	21,2	29,7	21,8	10,6	4,6	2,3	3 066
	40K-50K	8,5	22,7	28,8	21,8	10,3	5,5	2,4	2 858
	500K +	23,0	21,1	14,6	21,8	11,1	4,6	3,8	261
	50K-60K	9,3	20,4	28,3	24,0	11,8	4,3	1,9	2 783
	60K-70K	9,5	20,7	28,1	22,0	12,5	5,3	2,0	2 235
	70K-80K	8,5	20,8	28,9	22,9	11,4	5,7	1,8	2 025
	80K-90K	8,9	18	28,3	25,5	11,6	5,6	2,1	1 545
	90K-100K	8,5	16,6	27,9	25,1	13,7	5,9	2,2	1 539
	Inconnu	24,5	13,7	21,1	19,6	11,2	6,6	3,3	4 555

3.11 Régression de Poisson

TABLE 7 – Régression de Poisson entre le nombre de langages de programmation utilisés au quotidien et l'âge, le genre, le continent de résidence, le niveau de scolarité, le salaire annuel et secteur d'activité. Estimation des effets associés à la modalité "Genre Autre, âgé entre 18 et 29 ans, diplôme Autre, continent Afrique, secteur Autre et salaire entre zéro et 10K dollars"

Variable		Coefficient	Erreur standard	z	P(>|z|)
	Intercept	0,52	0,03	18,29	<0,001 ***
Age :					
	30-39	-0,01	0,01	-2,04	0,041 *
	40-49	0,03	0,01	3,52	<0,001 ***
	50-59	0,03	0,01	2,33	0,020 *
	60-69	-0,01	0,02	-0,64	0,520
	70+	-0,16	0,04	-3,59	<0,001 ***
Genre :					
	Femme	-0,14	0,02	-5,94	<0,001 ***
	Homme	-0,06	0,02	-2,51	0,012 *
Éducation :					
	Diplôme pro	0,22	0,04	5,45	<0,001 ***
	Doctorat	0,14	0,01	9,86	<0,001 ***
	Doctorat pro	0,10	0,02	4,71	<0,001 ***
	Licence	0,06	0,01	5,03	<0,001 ***
	Master	0,10	0,01	8,06	<0,001 ***
Continent :					
	Amérique du Nord	0,04	0,02	2,78	0,005 **
	Amérique du Sud	0,13	0,02	7,81	<0,001 ***
	Asie	0,00	0,01	0,49	0,622
	Europe	0,04	0,02	2,59	0,009 **
	Inconnu	0,02	0,02	1,29	0,197
	Moyen-Orient	0,01	0,02	0,43	0,666
	Océanie	0,02	0,03	0,74	0,458
Secteur :					
	Business	0,13	0,02	6,79	<0,001 ***
	Informatique	0,28	0,02	11,73	<0,001 ***
	Ingénierie	0,30	0,01	32,05	<0,001 ***
	Manageur	0,04	0,02	2,51	0,012 *
	Maths/Stats	0,20	0,01	14,46	<0,001 ***
	Phys. Astro.	0,21	0,02	9,80	<0,001 ***

		Coefficient	Erreur standard	z	P(>|z|)
	Sc. des Donnees	0,16	0,01	17,66	<0,001 ***
	Sc. humaines	0,10	0,05	2,36	0,018 *
	Sc. sociale	0,17	0,03	5,21	<0,001 ***
	Sc. Vie/Médicale	0,17	0,03	6,84	<0,001 ***
	Science de la Terre	0,18	0,05	4,02	<0,001 ***
Salaire :					
	100K-250K	0,15	0,01	13,11	<0,001 ***
	10K-20K	0,02	0,01	2,01	0,045 *
	20K-30K	0,05	0,01	4,62	<0,001 ***
	250K-500K	0,17	0,03	5,92	<0,001 ***
	30K-40K	0,05	0,01	3,90	<0,001 ***
	40K-50K	0,07	0,01	5,40	<0,001 ***
	500K +	-0,02	0,04	-0,41	0,678
	50K-60K	0,07	0,01	5,21	<0,001 ***
	60K-70K	0,08	0,02	5,45	<0,001 ***
	70K-80K	0,09	0,02	5,50	<0,001 ***
	80K-90K	0,11	0,02	6,27	<0,001 ***
	90K-100K	0,13	0,02	7,52	<0,001 ***
	Inconnu	-0,06	0,01	-5,43	<0,001 ***

La Table 7 représente une régression de Poisson entre le nombre de langages de programmation utilisés au quotidien et l'âge, le genre, le continent de résidence, le niveau de scolarité, le salaire annuel et secteur d'activité sans interaction. Avec une estimation des effets associés à la modalité "Genre Autre, âgé entre 18 et 29 ans, diplôme Autre, continent Afrique, secteur Autre et salaire entre zéro et 10K dollars".

La modalité Intercept : "Genre Autre, âgé entre 18 et 29 ans, diplôme Autre, continent Afrique, secteur Autre et salaire entre zéro et 10K dollars" a un coefficient de 0.52, p = <0.001.
La variable *Age* se décompose en trois parties. La première, comprenant les tranches d'âges 40-49 (coefficient : 0,03, p <0,01) et 70+ (coefficient : -0,16, p <0,01) est très fortement significatives. La seconde, comprenant les tranches d'âges 30-39 (coefficient : -0,01, p = 0,041), 50-59 (coefficient : 0,03, p = 0,020) est faiblement significatives. La troisième, comprenant la tranche d'âge 60-69 (coefficient : -0,01, p = 0,520) n'est pas statistiquement significative. La variable *Genre* se compose des femmes qui sont très fortement significatives (coefficient : -0,14, p <0,001) et des hommes faiblement significatifs (coefficient : -0,06, p = 0,012). La variable *Éducation* est très fortement significative dans son ensemble avec diplôme pro (coefficient : 0,22, p <0,001), doctorat (coefficient : 0,14, p <0,01), doctorat pro (coefficient : 0,10, p <0,01), licence (coefficient : 0,06, p <0,01) et master (coefficient : 0,10, p <0,01). La variable *Continent* se décompose en trois parties. La première, comprenant l'Amérique du Sud (coefficient : 0,13, p <0,01) est très fortement significative. La seconde, comprenant l'Amérique du Nord (coefficient : 0,13, p = 0,005) et l'Europe (coefficient : 0,04, p = 0,009) est fortement significatives. La troisième, comprenant l'Asie (coefficient : 0,00, p = 0,622), Inconnu (coefficient : 0,02, p = 0,197), Moyen-Orient (coefficient : 0,01, p = 0,666) et l'Océanie (coefficient : 0,02, p = 0,458) n'est pas statistiquement significative. La variable *Secteur* se décompose en deux parties. La première,

avec business (coefficient : 0,13, p <0,01), informatique (coefficient : 0,28, p <0,01), ingénierie (coefficient : 0,30, p <0,01), maths/stats (coefficient : 0,20, p <0,01), phys/astro (coefficient : 0,21, p <0,01), sciences des données (coefficient : 0,16, p <0,01), sciences sociales (coefficient : 0,17, p <0,01), sciences de la vie/médicale (coefficient : 0,17, p <0,01) et science de la terre (coefficient : 0,18, p <0,01) est très fortement significatives. La seconde, avec manageur (coefficient : 0,04, p = 0,012) et sciences humaines (coefficient : 0,10, p = 0,018) est faiblement significative. La variable *Salaire* se décompose en trois parties. La première, avec 100K-250K (coefficient : 0,15, p <0,01), 20K-30K (coefficient : 0,06, p <0,01), 250K-500K (coefficient : 0,17, p <0,01), 30K-40K (coefficient : 0,06, p <0,01), 40K-50K (coefficient : 0,07, p <0,01), 50k-60K (coefficient : 0,07, p <0,01), 60K-70K (coefficient : 0,08, p <0,01), 70K-80K (coefficient : 0,09, p <0,01), 80K-90K (coefficient : 0,11, p <0,01), 90K-100K (coefficient : 0,13, p <0,01) et inconnu (coefficient : -0,06, p <0,01) est très fortement significatives. La seconde, avec la tranche 10K-20k (coefficient : 0,02, p = 0,045) est faiblement significative. La troisième, avec la tranche 500K+ (coefficient : -0,02, p = 0,678) n'est pas statistiquement significative.

Chapitre 4

Discussion

4.1 Structure de la discussion

Nous commencerons par discuter des différents résultats en précisant le cheminement de réflexion. Ensuite, nous parlerons des limites du travail effectué. Pour finir, nous parlerons des perspectives.

4.2 Discussions des différents résultats

4.2.1 Distribution du nombre total de langages de programmation utilisés au quotidien par les répondants

Il existe une grande hétérogénéité dans le nombre de langages de programmations utilisés. Beaucoup n'utilisent qu'un à trois langages de programmation (70,2 %). Cela peut s'expliquer par des différences dans les formations suivies, des besoins au travail et de résultats demandés.

4.2.2 Étude du lien entre le nombre de langages de programmation et l'âge des répondants

Nous pouvons voir deux schémas distincts. Le premier est simplement dû au fait que la Data Science est une science récente et donc va grandement être représentée par une population jeune. Le second point est le nombre de langages de programmation utilisés, celui-ci commence avec un nombre restreint et une jeune tranche d'âge. Au fur et à mesure que l'âge augmente, le nombre de langages de programmation utilisés augmente aussi, ce qui s'explique simplement par l'expérience acquise, ainsi que la nécessité d'apprendre des nouveaux langages de programmation pour pouvoir utiliser de nouvelles technologies.

4.2.3 Étude du lien entre le nombre de langages de programmation et le genre des répondants

Nous pouvons voir que les inégalités du genre sont bien présentes, plus de 82 % sont des hommes, près de 16 % des femmes et près de 2 % se déclare comme autres. Les femmes sont sur-représentées pour zéro langage de programmation et sous-représentées pour quatre et plus langages de programmation et inversement pour les hommes. Cela peut s'expliquer par les différentes formations, des postes de travail plus élevé pour les hommes, une évolution de carrière moins importantes femmes.

4.2.4 Étude du lien entre le nombre de langages de programmation et le niveau de scolarité des répondants

Nous pouvons voir qu'un haut degré d'étude ne signifie pas une utilisation d'un grand nombre de langages de programmation. Nous pouvons voir, que quelque soit le niveau d'étude, le nombre de langages de programmation utilisés se situe entre un et trois. Les formations scolaires se focalisent essentiellement sur un ou deux langages de programmation en général, peuvent expliquer le nombre de langages de programmation utilisés et le fait que les métiers ne mobilise en général qu'un ou deux langages de programmation.

4.2.5 Étude du lien entre le nombre de langages de programmation et le secteur d'activité professionnelle des répondants

Il est évident que des métiers tels que ceux liés aux Data Science utilisent un grand nombre de langages de programmation différents, ne serait-ce que pour les analyses de données qui utilisent soit Python, R, SQL, etc. Cependant, la part du nombre de langages de programmation égal à zéro est surprenante, et ce, pour l'ensemble des secteurs. Cela peut s'expliquer par le fait que nous parlons ici de secteur d'activité et non de métier, et donc chaque secteur regroupe un éventail de métiers dont certains ne nécessitent aucun langage de programmation.

4.2.6 Étude du lien entre le nombre de langages de programmation et le continent de résidence des répondants

Nous pouvons voir différents motifs apparaître. Le premier est celui où les répondants utilisent majoritairement un ou deux langages de programmation dans cet ordre, avec l'Afrique 26,6 % et 25,2 % respectivement. Le second est celui où les répondants utilisent majoritairement deux ou un seul langage de programmation dans cet ordre, avec l'Asie 26,8 % et 22,7 %, le Moyen-Orient 27,7 % et 23,1 % et Inconnu 23,6 % et 23,5 %, respectivement. Le troisième est celui où les répondants utilisent majoritairement deux ou trois langages de programmation, avec l'Amérique du Nord 26,4 % et 24,5 %, l'Amérique du Sud 27,2 % 23,2 %, l'Europe 28,3 % 22,7 % et l'Océanie 26,1 % et 22,8 %, respectivement. Nous pouvons constater que l'Afrique est le seul continent à avoir un seul langage de programmation en première place. Cela peut s'expliquer de plusieurs manières, l'enseignement se concentre sur un seul langage de programmation, les besoins professionnels ne nécessitent qu'un seul langage de programmation.

4.2.7 Étude du lien entre le nombre de langages de programmation et le salaire des répondants

Nous pouvons voir que des franges salariales sont en opposition, avec la frange 0-10K la plus représentée et la frange 100k-200K la suivante (en ne comptant pas la frange "inconnu"). Les franges 20-30K, 30K-40K, 40K-50K sont globalement distribuées de façon identique. La frange 0-10K s'explique de plusieurs façons, un problème dans les données récoltées ou la distinction entre un travail fait de façon gratuite et un travail rémunéré n'est pas faite, ainsi on se retrouve avec un mélange des deux. Le premier pays en terme de répondants est l'Inde à un salaire moyen de d'environ 750 $, ce qui donc va augmenter le nombre de personnes ayant un faible salaire. La frange 100-200K peut s'expliquer de la même façon, mais cette fois-ci en prenant en compte que les USA sont le deuxième pays le plus représenté. Ceux-ci ont un salaire moyen compris entre 95 000 $ et 190 000 $. Enfin, les franges 20-30K, 30K-40K, 40K-50K vont correspondre aux salaires moyens des différents pays européens.

4.2.8 Régression de Poisson

Nous pouvons voir que l'ensemble des valeurs de la variable *Education* sont statistiquement significatifs, ce qui est normal, car les langages de programmation et leur nombre sont directement liés au futur secteur de travail des étudiants, et donc le nombre enseigné sera celui demandé pour leur future vie professionnelle. Nous pouvons voir que les valeurs de la variable *Secteur* sont majoritairement significatives, ce qui est normal, car chaque secteur possède ses propres besoins et ceux-ci sont souvent multiples (i.e Science des données avec Python, R, SQL, etc.). Nous pouvons voir que les valeurs de la variable *Salaire* sont majoritairement significatives, ce qui est normal, car un salaire plus élevé signifie une maitrise de plusieurs langages de programmation. Nous pouvons voir que les valeurs de la variable *Age* sont moyennement significatives dans l'ensemble. Nous pouvons voir que le continent de résidence des répondants n'est dans l'ensemble pas statistiquement significatif, ce qui est normal, car les langages de programmation ne sont pas spécifiques à un continent, mais au contraire universel et le nombre utilisé pour chaque métier et le même (i.e Science des données avec Python, R, SQL, etc.).

4.3 Limites du travail

Les questions d'ordre social sont trop peu nombreuses et génériques. Avoir plus de questions permettrait d'affiner le modèle encore plus finement.

La première limite est le lieu de récolte des réponses, kaggle.com. Ce site étant un dédié à la Data Science, les répondants ont une forte chance d'utiliser un ou plusieurs langages de programmation. La deuxième limite sont les questionnaires eux-mêmes, ceux-ci sont formatés de façon différente chaque année, ce qui entraine une normalisation des réponses entre chaque année obligatoire et donc une perte potentielle de la qualité des résultats.

4.4 Perspectives

Il serait nécessaire de faire un questionnaire avec plus d'items portant sur des sujets sociaux et ainsi avoir plus de variables indépendantes à étudier. Ce questionnaire ne devrait pas être effectué uniquement sur kaggle.com, mais de façon beaucoup plus large pour ainsi réduire l'influence due au fait que kaggle.com soit un site dédié à la Data Science. Le travail effectué porte uniquement sur une relation simple entre la variable dépendante et une seule variable indépendante à la fois. Il faudrait effectuer les recherches en prenant en compte cette fois-ci les interactions entre variables indépendantes pour améliorer les résultats.

Total mots : 4916

Bibliographie

[1] *Définition et histoire de la data science en 5 dates clés*. 2020.
URL : https://datascientest.com/definition-et-histoire-de-la-data-science-en-5-dates-cles-les-dessous-dune-ascension-fulgurante#:~:text=La%20data%20science%20est%20mise,au%20centre%20de%20la%20conversation..

[2] *Python leads the 11 top Data Science, Machine Learning platforms : Trends and Analysis*. 2019.
URL : https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html.

[3] Kaggle. *2018 Kaggle Machine Learning & amp ; Data Science Survey*. Nov. 2018.
URL : https://www.kaggle.com/datasets/kaggle/kaggle-survey-2018.

[4] Kaggle. *2019 kaggle machine learning & amp ; data science survey*. 2019.
URL : https://www.kaggle.com/competitions/kaggle-survey-2019/overview.

[5] Kaggle. *2020 kaggle machine learning & amp ; data science survey*$_2$020. 2020.
URL : https://www.kaggle.com/competitions/kaggle-survey-2020/overview.

[6] Kaggle. *2021 kaggle machine learning & amp ; data science survey*. 2021.
URL : https://www.kaggle.com/competitions/kaggle-survey-2021/overview.

[7] *Creative Commons License Deed. Creative Commons - Attribution 2.0 Générique - CC BY 2.0. (n.d.)*. 2022. URL : https://creativecommons.org/licenses/by/2.0/deed.fr.

Annexe A

Diagramme alluvial

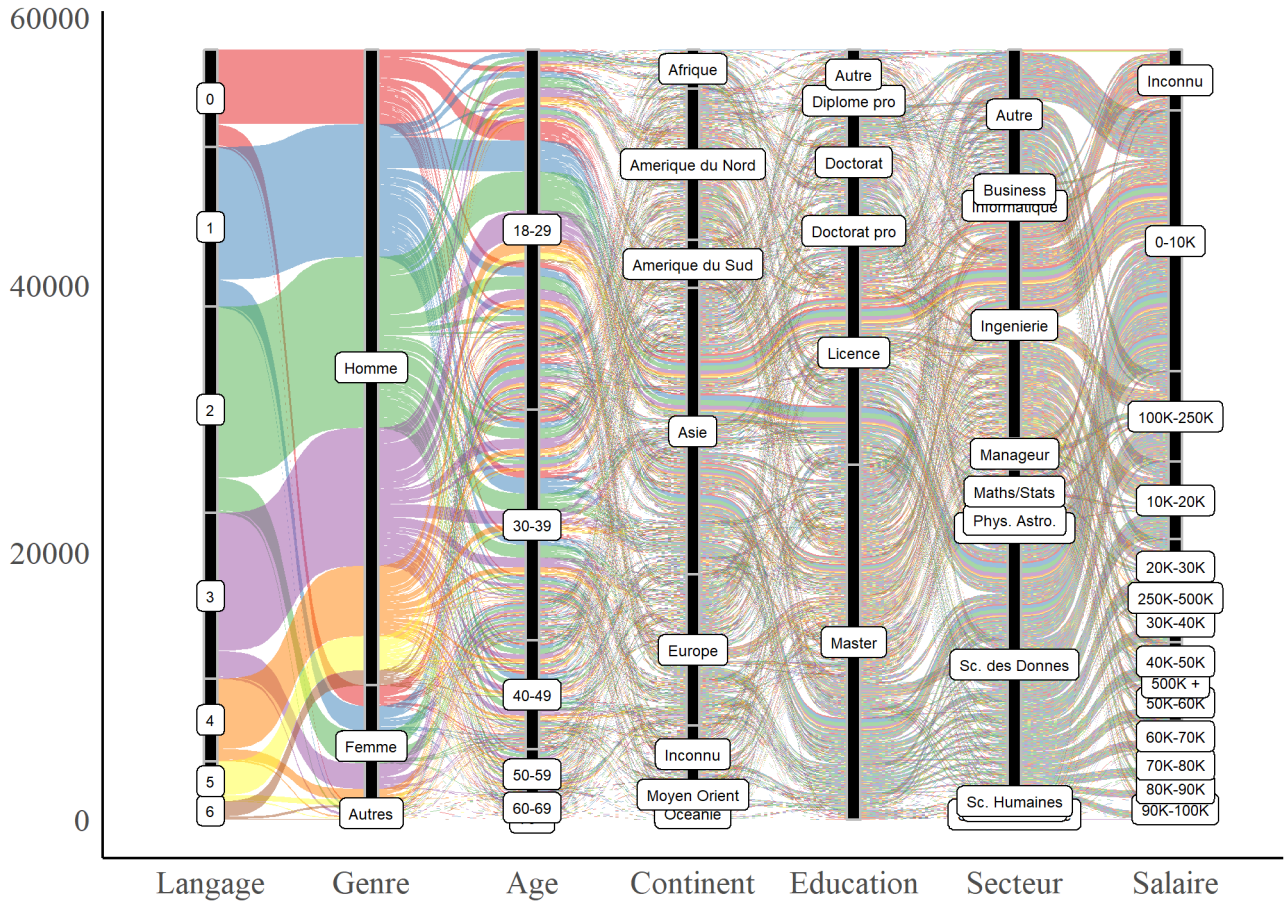

FIGURE 10 – Diagramme alluvial, nombre total de langages utilisés au travers des variables