Université Cergy Paris

Diplôme universitaire :
Data Analyst

UE 4 :
Report writing

January 6, 2023

Haury Fabien

Study of sociodemographic and socioprofessional influences on the number of programming languages used on a daily basis by Data Science practitioners

Haury Fabien

January 6, 2023

Contents

List of Figures

List of Tables

Part 1

Introduction

1.1 Context

The notion of Data Science was first put forward in 1974. However, it is from 2003 onwards that the emergence of specialized journals in this field has led to the emergence of Data Science as we define it today [7]. Since then, Data Science has become a distinct field of study whose evolution is constant. As a study done by the site KDnuggets (2019) points out [6], the tools used are also constantly evolving, or new tools created as needed.

1.2 Missing knowledge

A majority of the studies whose main subject is the tools used in data science focus only on the tools themselves, for example the evolution of the different programming languages used, etc. Rarely does a study focus on these tools and their links with sociodemographic (gender, etc.) or socioprofessional (salary, etc.) factors. Rarely does a study focus on these tools and their links with sociodemographic (gender, etc.) or socioprofessional (salary, etc.) factors.

Answering these questions would allow us to see the evolution of the tools from a new perspective. It would also allow us to establish a global cartography and thus to use the right variables for future studies, to detect new trends, etc.

1.3 Question

What are the sociodemographic or socioprofessional factors that influence the number of programming languages used in Data Science?

1.4 Methods used

The study will follow an exploratory logic. To answer these questions, the data will be transformed to be exploitable, using statistical tools (e.g. chi2, anova, etc.). The data are of qualitative and quantitative type.

Part 2

Methodology

2.1 Origin of the data

Every year since 2017, kaggle.com has offered a questionnaire to its users about the world of Data Sciences and Machine Learning. The datasets mobilized for this study are from the questionnaires offered in 2018, 2019, 2020 and 2021 [1] [2] [3] [4]. This data has been published under a CC 2.0 license [5].

2.2 Transformations performed on the data

A merge of the four datasets is performed. In order to keep a distinction between these four data sets, the year of each questionnaire is introduced by a new variable named *Annee*. The variable Q1 represents the age of the respondents and is normalized across the four years to obtain consistent age groups across the four years.

The variable Q2 corresponds to the gender of the respondents. People who are non-binary, do not prefer to answer, or prefer to write themselves will be recoded as "other" due to their small sample size.

The variable Q3 corresponds to the country of the respondents. There are 69 different countries, which will be grouped by their respective continents in a new variable named *Continent*, i.e. seven continents in total: North America, South America, Europe, Asia, Africa, Oceania and Middle East. This transformation is intended to clarify and account for differences in sample size among these countries. Some countries will also be renamed for clarity and readability, and the Q3 variable will be retained.

The variable Q4 corresponds to the respondents' level of education. Those who indicated that they did not prefer to answer, who had no education beyond high school, or who had completed a bachelor's degree or equivalent without validating that degree will be grouped in a new category named "Other" for greater clarity.

The variable Q5 corresponds to the sector of activity of the respondents. Some answers will be grouped together to ensure clarity. The grouping will be done in such a way as to preserve the field of activity as much as possible (i.e. engineers grouped together, etc.). The result of these groupings will be a new variable named *Role* and the original column will be kept.

Variable Q6 is the annual salary of respondents. Values are standardized to ensure consistent salary ranges across the four years. Respondents who do not wish to report their earnings will be recoded as Unknown.

The variables for the commonly used programming languages are recoded to 0 or 1, then a sum on each line is performed to get the total number of programming languages used by the respondent. This new information is kept in a new variable called *Language*.

A search for outliers is performed, followed by their elimination, in order to avoid any interference with the statistical tools.

2.3 Tools mobilized

The study will be done using the R language under RStudio. List of libraries used :

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

A use of Libre Office will be made during the study.
Functions have been created.

Part 3

Results

3.1 Introduction of the results

A normality test shows us that the variable *Language* does not follow a normal distribution. On average, respondents use 2.2 programming languages ± 1.5, are male, aged 18-29, with a master's degree, work in data science, reside in Asia, and earn an annual salary between zero and 10K dollars.

3.2 Sankey diagram

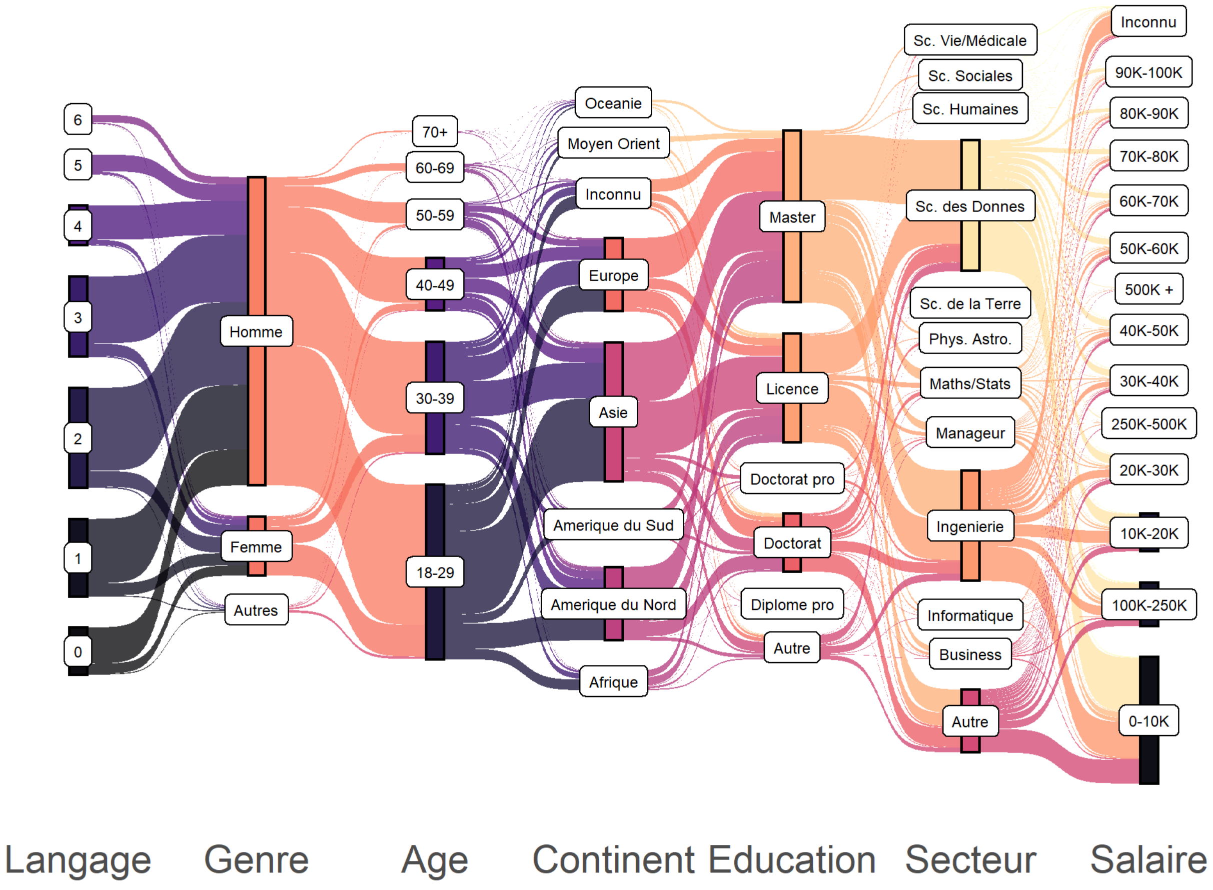

Figure 1: Sankey diagram for all the variables used

Figure 1 shows a Sankey diagram of the set of mobilized variables. Sankey's diagrams are a type of flow diagram in which the width of the arrows is proportional to the flow. We can see that the use of one, two or three programming languages is in the majority with 20.8%, 26.8% and 21.6% respectively. Men represent the majority of respondents with 82.6 %, 18-29 year olds represent with 46.7% the largest group and 30-39 year olds with 29.9% the second-largest group, which shows the youthfulness of the data professions. More than a third of respondents reside in Asia (37.2%). About 46% have earned a master's degree. Just over a third work in data science (35.1%). And one-third (33.9%) earn between \$0 and \$10,000 per year. Figure 10 in Appendix A represents an alluvial diagram, showing the path of each unique value of *Language* through the set of independent variables.

3.3 Normality test of the language variable by QQ plot

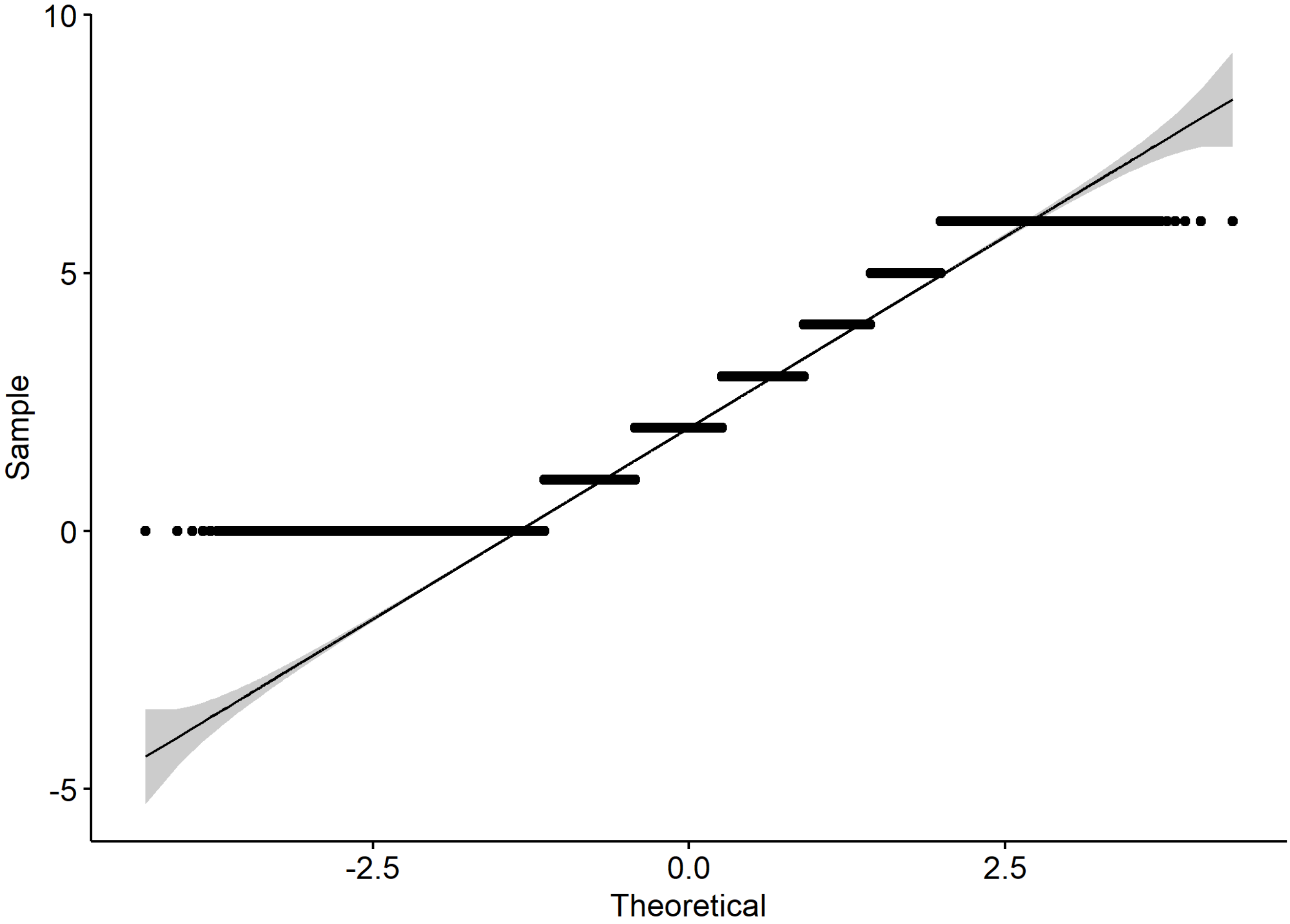

Figure 2: QQ plot of the language variable to test its normality

Figure 2 represents a QQ plot of the variable "Language" in order to test its normality. We find the seven categories of the variable "Language". It is decomposed in three parts. The first part, from quantile -4 to quantile -1, shows a spread and a distance of the points from the theoretical line, which means a distribution strongly biased towards the left. The second part going from the -1 quantile to the 2 quantile shows a set of points in the vicinity of the theoretical line, but without being on it, indicating that these points are not normally distributed. Finally, the third part going from quantile 2 to quantile 4 shows a spread and a distance of the points from the theoretical line, indicating a strongly right-biased distribution.

3.4 Number of programming languages used daily by respondents

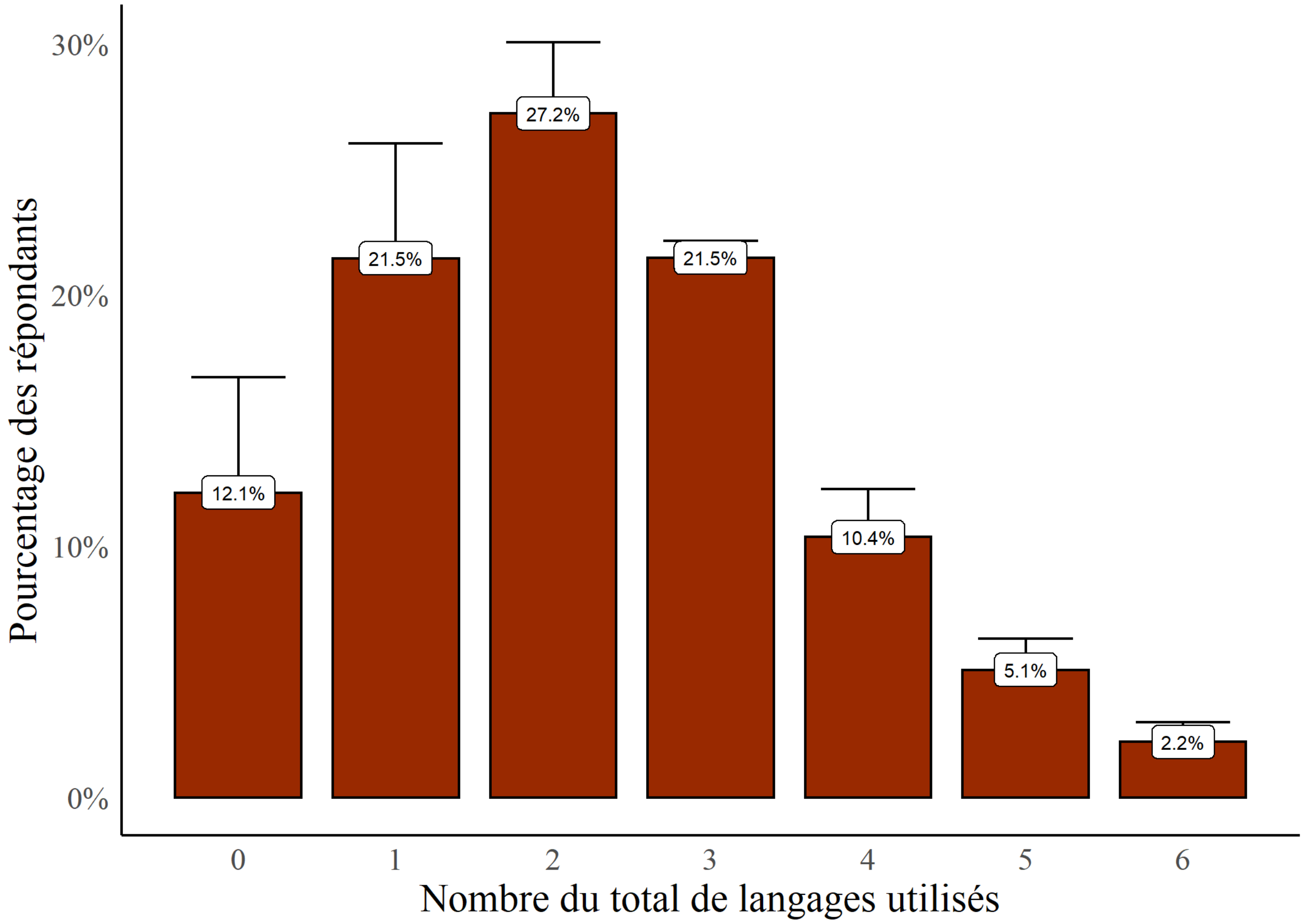

Figure 3: Distribution with standard deviation of the total number of programming languages used on a daily basis by respondents for all years 2018-2021

Figure 3 represents the distribution of the number of programming languages used for all years from 2018 to 2021, with their respective standard deviations. The number of respondents who use zero programming languages on a daily basis represents 12,1% ±4,7%. The number of respondents who use one programming language on a daily basis represents 21,5% ±4,5%. The number of respondents who use two programming languages on a daily basis represents about a quarter of the total with 27,2% ±2,8%. The number of respondents who use three programming languages on a daily basis represents 21,5% ±0,7%. The number of respondents who use four programming languages on a daily basis is 10,4% ±1,9%. The number of respondents who use five programming languages on a daily basis is 5,1% ±1,2%. The number of respondents who use six programming languages on a daily basis is 2,2% ±0,8%.

3.5 Age of respondents

3.5.1 Age distribution of respondents

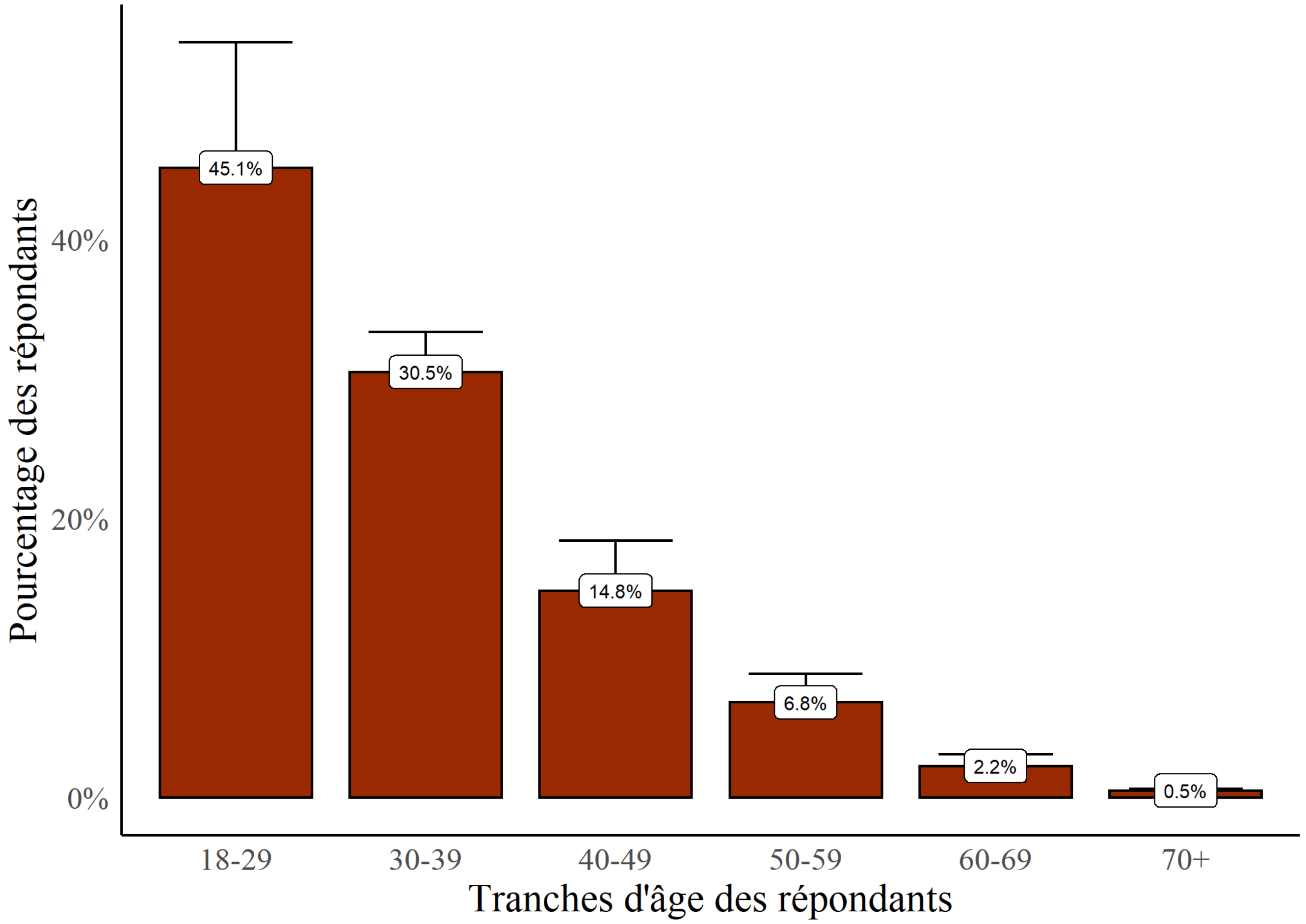

Figure 4: Distribution with standard deviation of respondent age ranges for all years 2018 to 2021

Figure 4 depicts the distribution of respondent age ranges for all years from 2018 to 2021, with their respective standard deviations. The number of respondents between 18 and 29 years of age represents 45.1% ±9%. The number of respondents between 30 and 39 years of age represents 30.5% ±2.9%. The number of respondents between 40 and 49 years of age represents 14.8% ±3.8%. The number of respondents between 50 and 59 years old represents 46.8% ±2.1%. The number of respondents between 60 and 69 years old represents 2,2% ±0,8%. The number of respondents aged 70 years or older represents 0,5% ±0,02%.

3.5.2 Contingency table

Table 1 represents the proportions of respondents' age groups by number of programming languages used daily.

The 18-29 and 30-39 year olds account for over three quarters of the respondents, 26,922 (46,7%) and 17,249 (29,9%) respectively. The 40-49 year olds represent a total of 8,200 respondents (14,2%). The number of respondents aged 50–59 years totaled 3,739 (6,5%). Those aged 60–69 years represent 1,221 (2,1%). Finally, respondents aged 70 and over numbered 274 (0,5%).

Table 1: Table of percentages of respondents' age groups by number of programming languages used, 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 Gender of respondents

3.6.1 Gender distribution of respondents

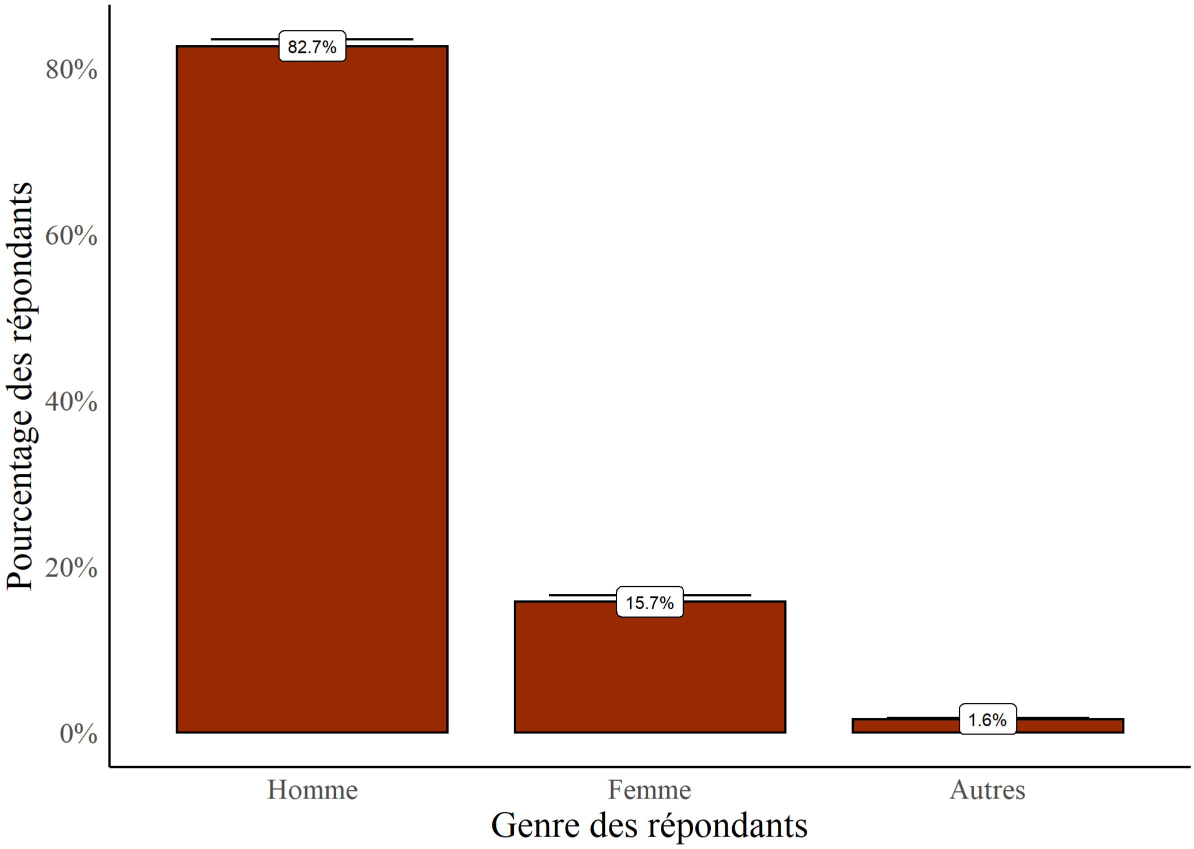

Figure 5: Distribution with standard deviation of gender of respondents for all years 2018-2021

Figure 5 represents the gender distribution of respondents for all years from 2018 to 2021, with their respective standard deviations. The percentage of respondents reporting as male is 82,7% ±0,8%. The percentage of respondents identifying as female is 15,7% ±0,7%. The percentage of respondents reporting as other is 1,6% ±0,05%.

3.6.2 Contingency table

Table 2: Table of percentages of respondents' gender by number of programming languages used, 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

Table 2 represents the proportions of respondents' gender by number of programming languages used daily. A majority of respondents identified as male with 47,571 (82,6%) responses. The number of people who identified themselves as female was 9,117 (15,8%). Finally, the number of respondents identifying as other is 917 (1,6%).

3.7 Education level of respondents

3.7.1 Distribution of respondents' education level

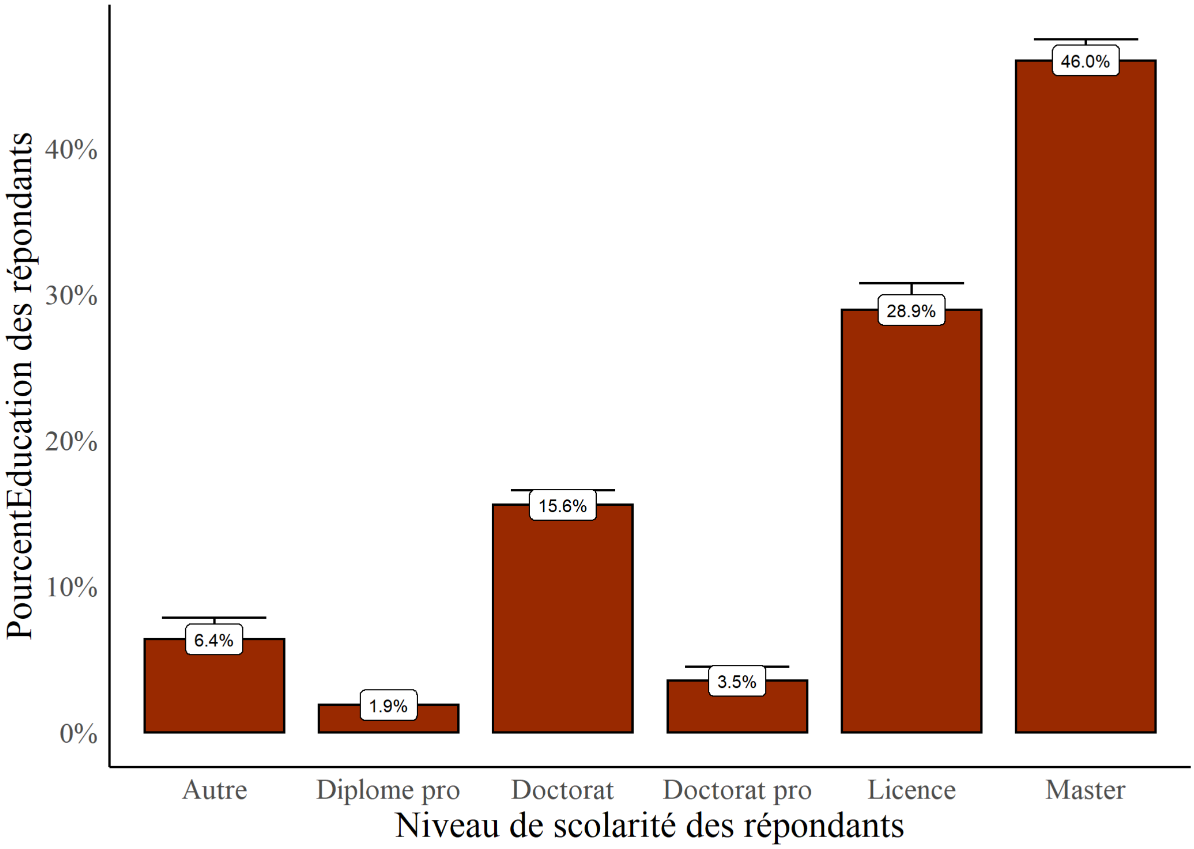

Figure 6: Distribution with standard deviation of respondents' educational attainment for all years 2018 to 2021

Figure 6 depicts the distribution of respondents' educational attainment for all of 2018 through 2021, with their respective standard deviations. The percentage of respondents reporting completion of a different degree is 6,4% ± 1,5 %. The percentage of respondents reporting completion of a professional degree is 1,5%. The percentage of respondents reporting having earned a doctorate degree is 15,6% ± 1 %. The percentage of respondents reporting having earned a professional doctorate is 3,5% ± 0,9 %. The percentage of respondents reporting having obtained a Bachelor's degree was 28,9% ± 1,8 %. The percentage of respondents reporting having obtained a Master's degree other than a Bachelor's degree was 46% ± 1,9 %.

3.7.2 Contingency table

Table 3: Table of percentages of respondents' education level by number of programming languages used, 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

Table 3 represents the proportions of respondents' education level by number of programming languages used daily.
Table-3: Educational Levels of Respondents Nearly half of the respondents have a master's degree, 26,551 (46,1%), followed by a bachelor's degree with 16,791 (29,1 %). 8,915 (15.5%) of the respondents have a PhD and 1,408 (2.4%) have a professional doctorate. 283 (0.5%) of them have a professional degree and finally, 3,657 (46.3%) did not wish to answer.

3.8 Respondents' sector of activity

3.8.1 Distribution of respondents' sectors of activity

Figure 7 represents the distribution of respondents' industries for all of 2018 through 2021, along with their respective standard deviations. The percentage of respondents reporting working in a different sector is 18,4% ±4,9%. The percentage of respondents reporting working in the business sector is 8,0%. The percentage of respondents reporting that they work in the IT sector is 4,4%. The percentage of respondents reporting that they work in the engineering sector is 26% ±20,5% The percentage of respondents reporting working in the managerial sector is 5,9% ±0,9%. The percentage of respondents who reported working in the Mathematics or Statistics sector was 4,8% ±11%. The percentage of respondents reporting working in Physics or Astronomy is 5,5%. The percentage of respondents reporting working in the Earth Sciences sector is 11,1% ±1,2%. The percentage of respondents reporting that they work in the Data Science sector is 53%. The percentage of respondents reporting working in the Humanities is 1,2%. The percentage of respondents reporting working in the Social Sciences is 2,4%. The percentage of respondents who reported working in the Life Sciences or Medical field was 53,9%.

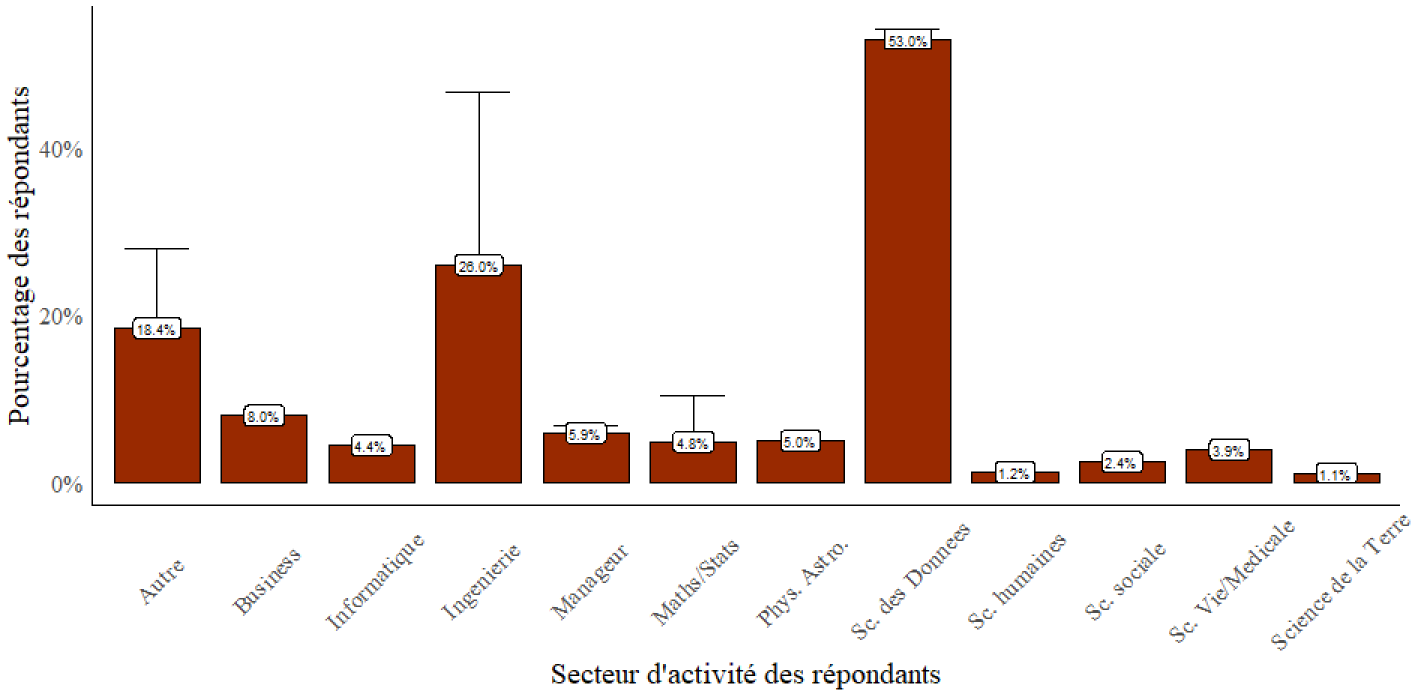

Figure 7: Distribution with standard deviation of respondents' industry for all years 2018 to 2021

3.8.2 Contingency table

Table 4 represents the proportions of the respondents' industry by the number of programming languages used daily. The variable *Number of Programming Languages* represents the number of programming languages used on a daily basis. The variable *Role* corresponds to the respondents' professional domain. We can see three major roles that account for over four-fifths of the total with 46,900 (81.4%) responses. The first is data science with 20,223 (35.1%) respondents reporting from this occupational area. The second is the engineering field with a total of 16,996 (29.5%) respondents reporting. The third is the one named other with 9,681 (16.8%) respondents of the total. Humanities, social, life, medical and earth sciences represent 1,697 (2.9%)) of the total respondents. Mathematics, statistics, physics, and astronomy accounted for 4,305 (7.5%) of the total. Business and managerial roles represent 3,844 (6.7%) of the total. Computer science roles account for 859 (1.5%) of the total.

Table 4: Table of percentages of respondents' sectors of activity according to the number of programming languages used, 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 of residence of respondents

3.9.1 Distribution of respondents' continents of residence

Figure 8 represents the distribution of respondents' continents of residence for all of 2018 to 2021, with their respective standard deviations. The percentage of respondents reporting residence in Africa is 5,2% ±2,3%. The percentage of respondents reporting residency in North America is 19,2% ±4,2%. The percentage of respondents reporting residence in South America was 6,5% ±1%. The percentage of respondents reporting residence in Asia was 37,3% ±2,9%. The percentage of respondents reporting residence in Europe was 19,4% ±2,6%. The percentage of respondents whose continent of residence is unknown is 7,5% ±0,9%. The percentage of respondents reporting residence in the Middle East was 3,3% 0,8%. The percentage of respondents reporting residence in Oceania is 1,6% ±0,04%.

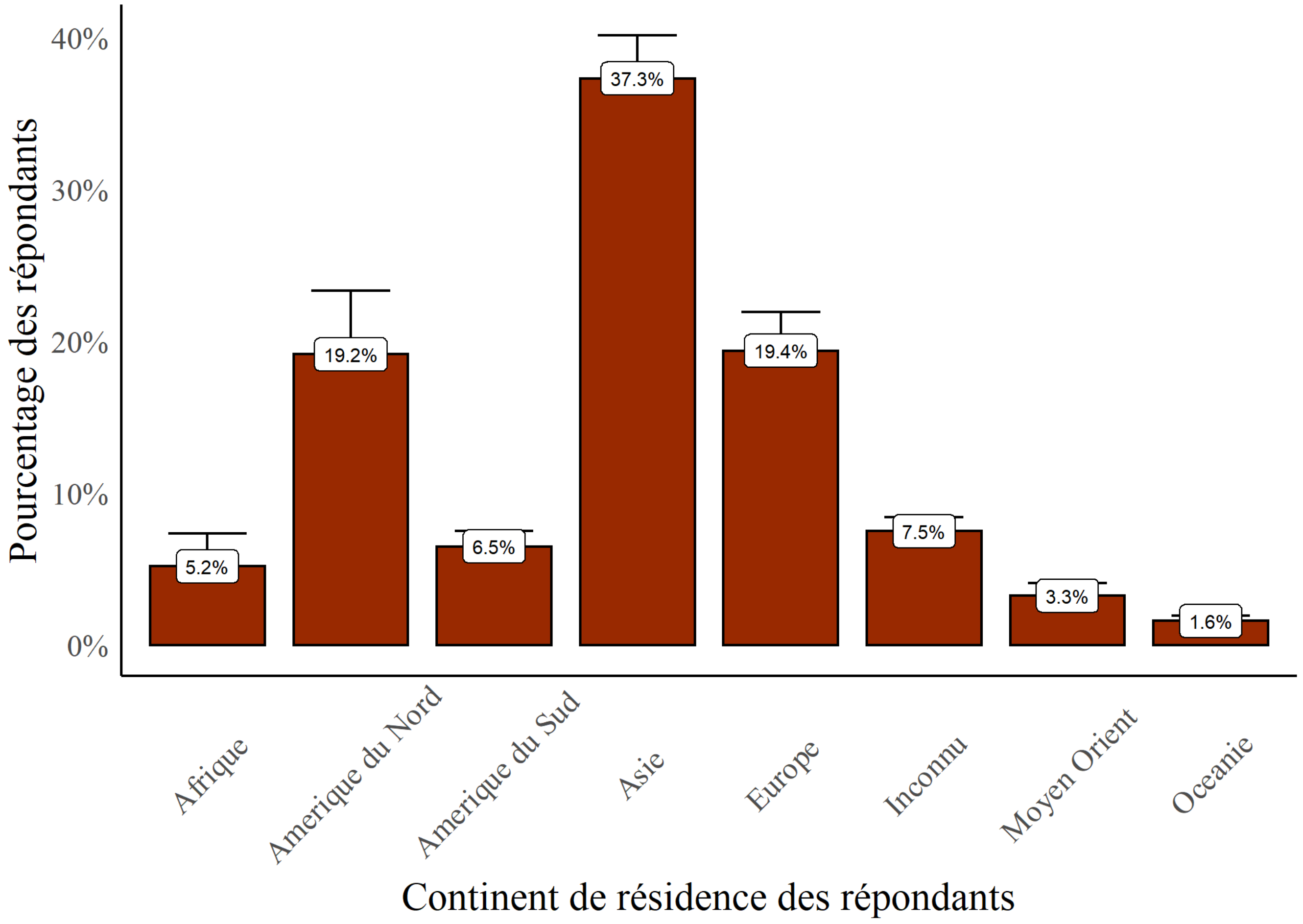

Figure 8: Distribution with standard deviation of respondents' continents of residence for all years 2018-2021

3.9.2 Contingency table

Table 5: Table of percentages of respondents' continents of residence according to the number of programming languages used, 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

Table 5 represents the proportions of the respondents' continent of residence by number of programming languages used daily.

Asia accounts for more than a third of the total with 21,430 (37,2%) respondents. North America and Europe with 11,319 (19.6%) and 11,306 (19,6%) responses respectively represent the next two poles. The number of people who did not wish to respond was 4,252 (7,4%) South America, Africa, Middle East and Oceania with 3,641 (6,3%), 2,884 ((5%), 1,843 (3,2%)and 927 (1,6%) responses respectively, are the minority continents.

3.10 Annual salary of respondents

3.10.1 Distribution of respondents' annual salaries

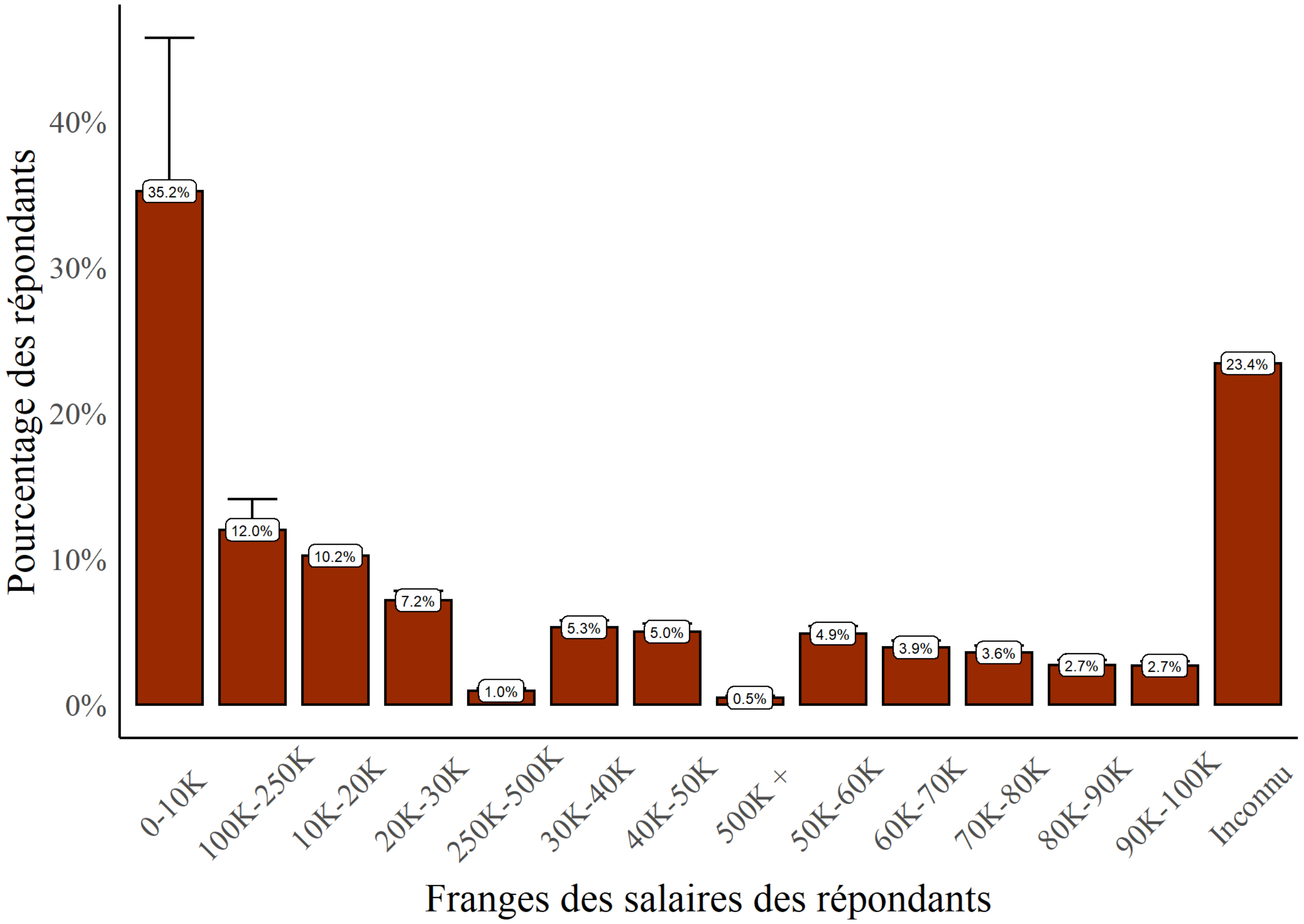

Figure 9: Distribution with standard deviation of respondents' annual salaries for all years 2018-2021

Figure 8 depicts the distribution of respondents' annual salaries for all of 2018 through 2021, with their respective standard deviations. The percentage of respondents reporting getting between zero and 10K dollars is 35,2% ±10,5%. The percentage of respondents reporting that they would get between $100k and $250K is 12% ±2,1%. The percentage of respondents who reported earning between $10K and $20K was 10,2% ±0.6%. The percentage of respondents who reported getting between $20K and $30K was 7,2% ±0,7%. The percentage of respondents reporting getting between $250K and $500K is 1% ± 0,02%. The percentage of respondents who reported obtaining between $30K and $40K was 5,3% ±0,45%. The percentage of respondents who reported getting between $40K and $50K was 5% ±0,6%. The percentage of respondents who reported getting $500K or more was 0,5% ±0,02%. The percentage of respondents who reported getting between $50K and $60K was 4,9% ±0,5%. The percentage of respondents who reported getting between $60K and $70K was 3,9% ±0,5%. The percentage of respondents who reported obtaining between $70K and $80K was 3,6% ±0,6%. The percentage of respondents who reported getting between $80K and $90K was 2,7% ±0,4% The percentage of respondents reporting getting between $90K and $100K is 2,7% ±0,4%. The percentage of respondents whose salary is unknown is 23,4%.

3.10.2 Contingency table

Table 6 represents the proportions of respondents' salaries by the number of programming languages used daily.
Table 6 The 0-10K fringe represents the largest group with 19,521 (33,9%). The 100k-250k fringe represents the second-largest group with 6,728 (11,7%) of the total respondents. The 10K-20K fringe represents the third group with 5,831 (10,1 %) respondents. The 20K-30K and unknown bangs represent with 4,119 (7,2%) and 4,555 (7,9%) respectively the next group. The 30K-40K, 40K-50K, and 50K-60K bangs represent with 3,066 (5,3 %), 2,858 (5%), and 2,783 (4,8%) respectively the next group. The 60K-70K and 70K-80K bangs represent with 2,235 (3,9%) and 2,025 (3,5%)respectively the next group. The 80K-90K and 90K-100K bangs represent with 1,160 (2,7%) and 1,539 (2,7%) respectively the next group. The 250K-500K fringe represents 539 (0,9%) of the total respondents. Finally, with 261 (0,5%)of the total respondents, the 500K+ fringe.

Table 6: Table of percentages of respondents' salary bangs by number of programming languages used, 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 Poisson regression

Table 7: Poisson's regression between the number of programming languages used in daily life and age, gender, continent of residence, education, annual salary and industry. Estimated effects associated with the modality "Gender Other, aged between 18 and 29, degree Other, continent Africa, industry Other and salary between zero and 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 ***

Table 7 represents a Poisson regression between the number of programming languages used in daily life and age, gender, continent of residence, education, annual salary, and industry without interaction. With an estimate of the effects associated with the modality "Gender Other, aged between 18 and 29, degree Other, continent Africa, sector Other and salary between zero and 10K dollars".

The Intercept modality: "Genre Autre, âgé entre 18 et 29 ans, diplôme Autre, continent Afrique, secteur Autre et salaire entre zéro et 10K dollars" has a coefficient of 0.52, p = <0.001.
The variable *Age* is broken down into three parts. The first, comprising the 40-49 (coefficient: 0.03, p <0.01) and 70+ (coefficient: -0.16, p <0.01) age groups, is highly significant. The second, comprising the age groups 30-39 (coefficient: -0.01, p = 0.041), 50-59 (coefficient: 0.03, p = 0.020) is weakly significant. The third, including the age range 60-69 (coefficient: -0.01, p = 0.520) is not statistically significant.
The variable *Genre* consists of females being very strongly significant (coefficient: -0.14, p <0.001) and males being weakly significant (coefficient: -0.06, p = 0.012).
The variable *Éducation* is very strongly significant as a whole with pro degree (coefficient: 0.22, p <0.001), doctorate (coefficient: 0.14, p <0.01), pro doctorate (coefficient: 0.10, p <0.01), bachelor's degree (coefficient: 0.06, p <0.01) and master's degree (coefficient: 0.10, p <0.01).
The variable *Continent* is broken down into three parts. The first, including South America (coefficient: 0.13, p <0.01) is very strongly significant. The second, including North America (coefficient: 0.13, p = 0.005) and Europe (coefficient: 0.04, p = 0.009) is highly significant. The third, including Asia (coefficient: 0.00, p = 0.622), Unknown (coefficient: 0.02, p = 0.197), Middle East (coefficient: 0.01, p = 0.666) and Oceania (coefficient: 0.02, p = 0.458) is not statistically significant.

The variable *Secteur* is decomposed into two parts. The first, with business (coefficient: 0.13, p <0.01), computer science (coefficient: 0.28, p <0.01), engineering (coefficient: 0.30, p <0.01), math/stats (coefficient: 0.20, p <0.01), phys/astro (coefficient: 0.21, p <0.01), data science (coefficient: 0.16, p <0.01), social science (coefficient: 0.17, p <0.01), life/medical science (coefficient: 0.17, p <0.01), and earth science (coefficient: 0.18, p <0.01) is highly significant. The second, with manager (coefficient: 0.04, p = 0.012) and humanities (coefficient: 0.10, p = 0.018) is weakly significant.

The variable *Salaire* is decomposed into three parts. The first, with 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), and unknown (coefficient: -0.06, p <0.01) is highly significant. The second, with the 10K-20k bracket (coefficient: 0.02, p = 0.045) is weakly significant. The third, with the 500K+ bracket (coefficient: -0.02, p = 0.678) is not statistically significant.

Part 4

Discussion

4.1 Structure of the discussion

We will start by discussing the different results by specifying the path of reflection. Then, we will talk about the limits of the work done. Finally, we will talk about the perspectives.

4.2 Discussion of the different results

4.2.1 Distribution of the total number of programming languages used daily by respondents

There is great heterogeneity in the number of programming languages used. Many use only one to three programming languages (70,2%). This can be explained by differences in training, work needs and required results

4.2.2 Study of the relationship between the number of programming languages and the age of the respondents

We can see two distinct patterns. The first is simply due to the fact that Data Science is a recent science and therefore will be largely represented by a young population. The second is the number of programming languages used, this one starts with a small number and a young age range. As the age increases, the number of programming languages used also increases, which is simply due to the experience gained, as well as the need to learn new programming languages in order to use new technologies.

4.2.3 Study of the relationship between the number of programming languages and the gender of the respondents

We can see that gender inequalities are well present, more than 82% are men, almost 16% are women and almost 2% declare themselves as others. Women are over-represented for zero programming languages and under-represented for four or more programming languages, and vice versa for men. This can be explained by the different trainings, higher job positions for men, less important career evolution for women.

4.2.4 Study of the relationship between the number of programming languages and the respondents' level of education

We can see that a high level of study does not mean the use of a large number of programming languages. We can see that, whatever the level of study, the number of programming languages used is between one and three. The school training courses focus essentially on one or two programming languages in general, which can explain the number of programming languages used and the fact that the professions generally use only one or two programming languages.

4.2.5 Study of the relationship between the number of programming languages and the professional sector of the respondents

It is obvious that professions such as those related to Data Science use a large number of different programming languages, if only for data analysis, which uses either Python, R, SQL, etc. However, the share of the number of programming languages equal to zero is surprisingly high across all industries. This can be explained by the fact that we are talking about sectors of activity and not professions, and therefore each sector includes a range of professions, some of which do not require any programming language.

4.2.6 Study of the relationship between the number of programming languages and the respondents' continent of residence

We can see different patterns emerging. The first one is the one where the respondents use mostly one or two programming languages in this order, with Africa 26,6% and 25,2% respectively. The second is where respondents predominantly use two or one programming language in that order, with Asia 26,8% and 22,7%, the Middle East 27,7% and 23,14=%, and Unknown 23,6% and 23,5%, respectively. The third is the one where the respondents use two or three programming languages, with North America 26,4% and 24,5%, South America 27,2% and 23,2%, Europe 28,3% and Oceania 26,1% and 22,8%, respectively. We can see that Africa is the only continent to have a single programming language in the first place. This can be explained in several ways: education is focused on a single programming language, professional needs only require a single programming language.

4.2.7 Study of the relationship between the number of programming languages and the salary of respondents

We can see that the salary bangs are in opposition, with the 0-10K fringe the most represented and the 100k-200K fringe the next most (not counting the "unknown" fringe). The 20-30K, 30K-40K, 40K-50K bangs are overall identically distributed. The 0-10K fringe can be explained in several ways, a problem in the data collected where the distinction between work done for free and paid work is not made, so we end up with a mixture of both. The first country in terms of respondents is India, with an average salary of about $750, which will increase the number of people with low salaries. The 100-200K range can be explained in the same way, but this time taking into account that the USA is the second most represented country. They have an average salary between $95,000 and $190,000. Finally, the 20-30K, 30K-40K and 40K-50K bands correspond to the average salaries of the various European countries.

4.2.8 Poisson egression

We can see that all the values of the variable *Education* are statistically significant, which is normal, because the programming languages and their number are directly linked to the future work sector of the students, and thus the number taught will be the one required for their future professional life. We can see that the values of the variable *Secteur* are mostly significant, which is normal, because each sector has its own needs and these are often multiple (i.e. Data Science with Python, R, SQL, etc.). We can see that the values of the variable *Salaire* are mostly significant, which is normal, because a higher salary means a mastery of several programming languages. We can see that the values of the variable *Age* are moderately significant overall. We can see that the continent of residence of the respondents is not statistically significant overall, which is normal, because the programming languages are not continent specific, but on the contrary universal and the number used for each job is the same (i.e. Data Science with Python, R, SQL, etc.).

4.3 Limits of the work

The social questions are too few and generic. Having more questions would allow the model to be refined even more finely. The first limitation is the place where the answers were collected, kaggle.com. This site is dedicated to data science, so respondents are likely to use one or more programming languages. The second limitation is the questionnaires themselves, which are formatted differently each year, resulting in a standardization of responses between each mandatory year and thus a potential loss of quality of results.

4.4 Perspectives

It would be necessary to make a questionnaire with more items on social topics and thus have more independent variables to study. This questionnaire should not be done only on kaggle.com, but in a much broader way to reduce the influence due to the fact that kaggle.com is a site dedicated to Data Science. The work done is only on a simple relationship between the dependent variable and one independent variable at a time. The research should be done by taking into account the interactions between independent variables to improve the results.

Bibliography

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

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

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

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

[5] *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.

[6] *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.

[7] *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..

Appendix A

Alluvial diagram

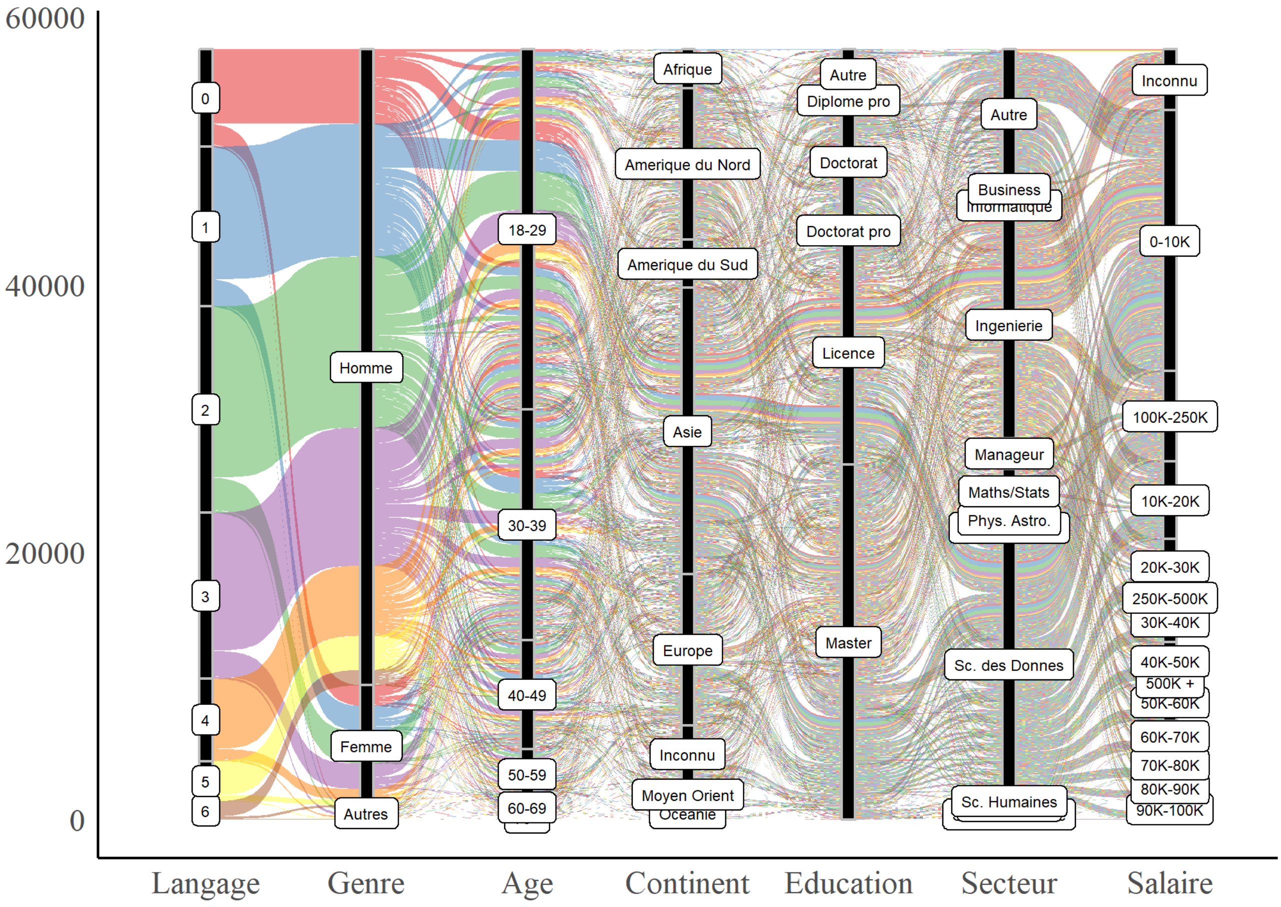

Figure 10: Alluvial diagram, total number of languages used through variables