Which Technologies are Most Frequently Used by Data Scientists?

Paula Pereira
University of Minho, Portugal
a77672@alunos.uminho.pt

João Paulo Fernandes
LIACC & DEI-FEUP, Portugal
jpaulo@fe.up.pt

Jácome Cunha
DEI-FEUP & HASLab/INESC TEC, Portugal
jacome@fe.up.pt

***Abstract*—Data collection is pervasively bound to our digital lifestyle. A recent study reports that the growth of the data created and replicated in 2020 was even higher than in the previous years to an astonishing global amount of 64.2 zettabytes of data. There are numerous companies whose services/products rely heavily on data analysis, and mining the produced data has already revealed great value for businesses in different sectors. In order to be able to support the professionals that do this job, typically known as data scientists, we first need to characterize them. To contribute towards this characterization, we conducted a public survey and in this work we present the results about a particular aspects of their life: the *tools* they use and need.**

***Index Terms*—data science, survey, empirical evidence**

I. Introduction

Every day huge amounts of data are created and mined [1], [2], [3]. It is estimated that every minute 5.7 million searches are conducted on Google and there are 575 thousand posts on Twitter. Companies from all sectors have realized the value of their data and are using it to gain competitive advantage [2], [4], [5], [**?**], [6]. This exploitation of the collected data makes it one of the most valuable resources to organizations: in 2017 *The Economist* stated that oil was no longer the world's most valuable resource, losing that position to data [3] and that *data scientist* was the *sexiest job of the 21st century* [4].

To fully understand how we can assist data science workers being more productive, we need to understand who they are, how they work, what are the skills they hold/lack, and which tools they use/need. The main goal of this preliminary work is to clarify one of these aspects, namely:

$\mathcal{RQ}$ Which technologies are most frequently used?

To accomplish this goal we conducted a public survey distributed worldwide whose results we now present.

II. Survey design and analysis methodology

The survey was divided in six sections, and followed the subsequent structure: 1) Academic background; 2) Professional situation; 3) Self-evaluation of strengths on several tasks; 4) Work characterization (problems, time spent coding); 5) Technologies used; 6) Demographic questions.

The survey was built using Google Forms and distributed online via several forums, namely Stack Overflow, Kaggle, Reddit, Facebook, and LinkedIn, but also through email people and communities known to be working in data science. The survey was open from April to the end of 2020.

This work is financed by National Funds through the Portuguese agency Fundação para a Ciência e a Tecnologia within project LA/P/0063/2020.

978-1-6654-4214-5/22/$31.00 ©2022 IEEE

III. Survey Results

We wanted to gather a large number of responses from a diverse group of people in terms of geography, gender, and age. To capture this information, the survey included a section with demographic questions asking participants their gender, their age and their country. This data is shown in Table I.

TABLE I
Country and gender demographics.

Country	Female	Male	Total
Australia	1	1	2
Brazil	0	4	4
Canada	1	7	8
Germany	0	6	6
India	0	6	6
Netherlands	1	1	2
Norway	2	1	3
Portugal	12	32	44
Spain	1	2	3
Switzerland	1	1	2
Thailand	1	1	2
UK	0	7	7
USA	3	9	12
Other	6	9	15
Total	**29**	**87**	**116**

A. Technology

We analyze these results under the perspective of the academic background of the participants since there are some difference for professionals with and without a computer science (CS) background.

a) IDEs or Editors: Figure 1 shows the choices of the participants regarding *IDEs or editors*. The most used IDE by people with a CS background is *IPython/Jupyter* (21,93%), which is quite similar to the usage of people without a CS background (20,69%), making *IPython/Jupyter* the second option in this group. The most used editor by people without a CS background is *RStudio*, but its use by people with a CS background is much lower (8,56%). In the group of people with CS background, the second most used text editor is *PyCharm* (14,97%), which reveals a preference for Python editors. In addition to *IDEs* and *text editors*, 16,58% of people with a CS background indicated using another option (e.g. *IntelliJ* or *Matlab*). For people without a CS background, this percentage is less than half (8,05%).

b) Programming, Scripting or Markup: The 3 programming languages (PLs) most used by people with or without a CS background coincide: *Python*, *SQL* and *R* (see Figure 2 in Appendix). The biggest difference concerns *R*, which was indicated by 10,98% (respectively, 21,11%) of the people with (respectively, without) a CS background. *SQL* was the second most indicated PL. In addition to *Scala*, there were six more

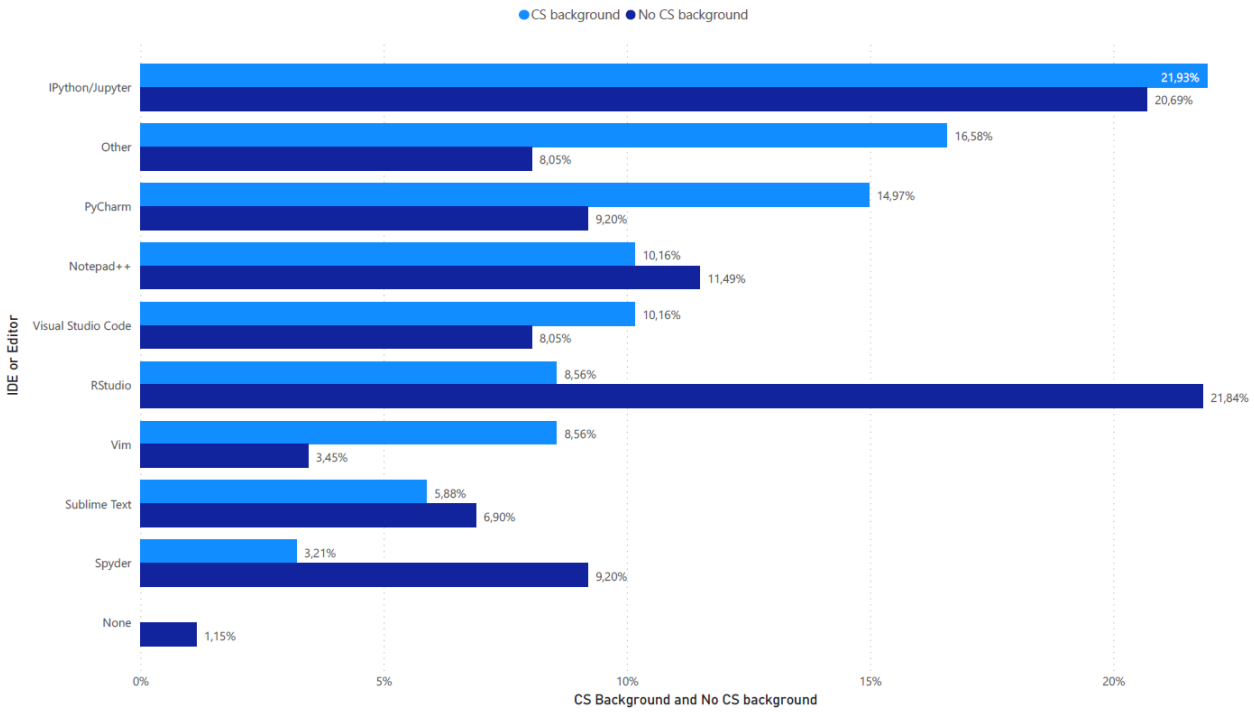

Fig. 1. IDE or Editor by background.

PLs that were only mentioned by people with CS background (*C#, Go, DAX, Julia, USQL*, and *Visual Basics*), and two PLs that were only mentioned by people without CS background (*SAS* and *Spark*). Overall, we note that Python and SQL are the two most used PLs by data science professionals and that people with a CS background use a greater diversity of PLs than people without a background in CS.

c) Machine Learning Frameworks/Libraries/Tools: As can be seen in Figure 3, concerning the participants with a CS background, the most indicated options were the library *scikit-learn* (21,46%), the open-source platform *Tensorflow* (13,24%), and the artificial neural networks library *Keras* (10,05%). In the group of people without CS background, *scikit-learn* (19,59%) and *Tensorflow* (11,34%) appear again as the first and second most indicated options, and in third place comes the machine learning framework *Torch/PyTorch* (10,31%). All four share the fact that they allow the application of machine learning techniques using *Python*, which reinforces the preference for *Python*. Some options were only mentioned by participants in one of the groups, with 10 (respectively, 3) being mentioned only by people with (respectively, without) a CS background. Finally, we also note that the percentage of participants without a CS background who indicated not using any type of tool is 6,19%, while in the group of people with a CS background this percentage is only 0,91%.

d) Statistics Packages/Tools: Figure 4 shows the responses we obtained. It is clear that *spreadsheet editors* are the preferred tools for statistical analysis amongst both professionals with (39,33%) and without CS background (33,82%). *Tableau* is also widely used. On the other hand, *SPSS* and *SAS*, despite sharing several features with *Tableau*, are mainly used only by professionals without CS background. Finally, it is relevant to notice that 14,61% and 7,35% of professionals with and without CS background, respectively, indicated not using any *statistics packages/tools* during their work, which amounts to 11,46% of the respondents.

e) Data visualization Libraries/Tools: Figure 5 shows the choices of the participants. Once again, there is a clear predominance of *Python-based* libraries. The most used libraries are *Matplotlib, Seaborn*, and *ggplot2*. Tools that provide an interactive way to create/manage visualizations, such as *Power BI* or *Tableau*, are also widely used. With this information we can also infer that almost no data science professional can conduct their work without a visualization tool.

B. Answering the RQ

There seems to be a slight difference in the choices between those with a CS background and those without. Professionals with a CS background have a strong preference for Python, which is reflected in their preferred IDE as well as their choices in machine learning and data visualization technologies. Aside from Python, R appears to be a common choice among professionals without a CS background. Finally there seems to be a consensus among professionals in choosing spreadsheet editors for the statistical analysis of data.

IV. Conclusions and Future Work

Our goal was to know the technologies used by data science professionals. We conducted an online worldwide survey, having obtained 116 valid answers. The results show that Python and R are the most popular tools. However, spreadsheets are also quite popular. Moreover, there seems to be some different between data scientists with and without a CS background.

APPENDIX

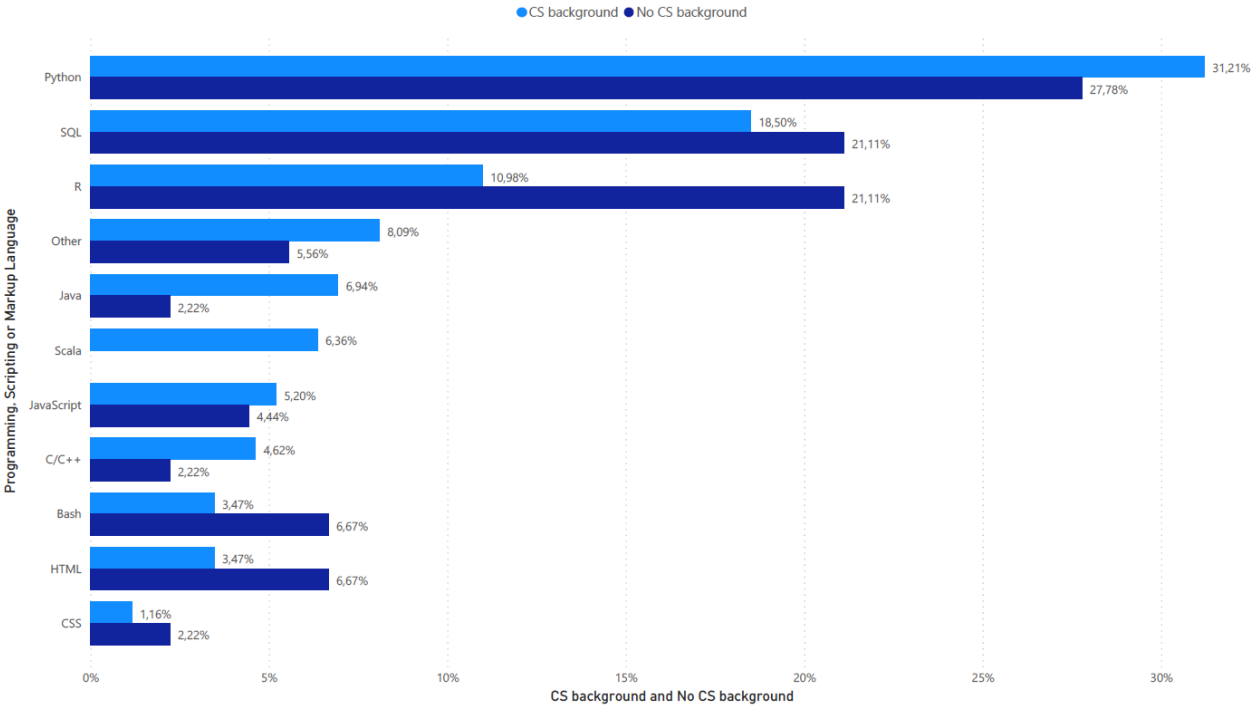

Fig. 2. Programming, Scripting or Markup Language by background.

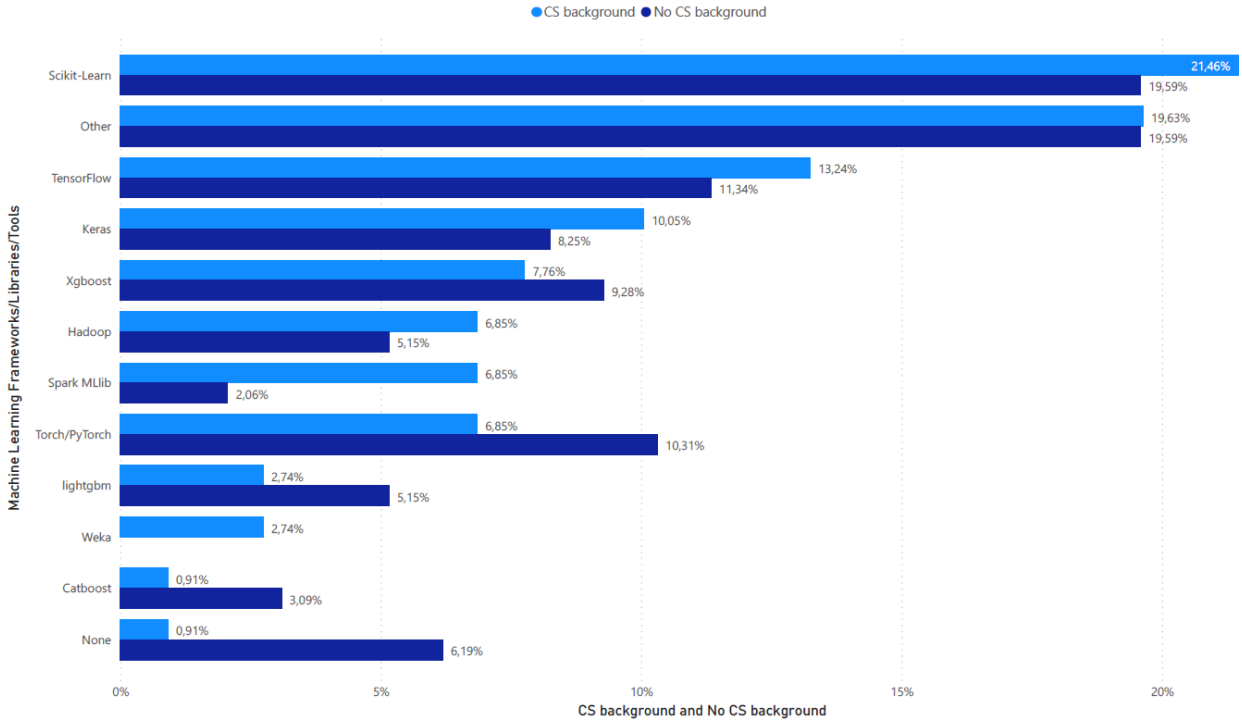

Fig. 3. Machine Learning Frameworks/Libraries/Tools by background.

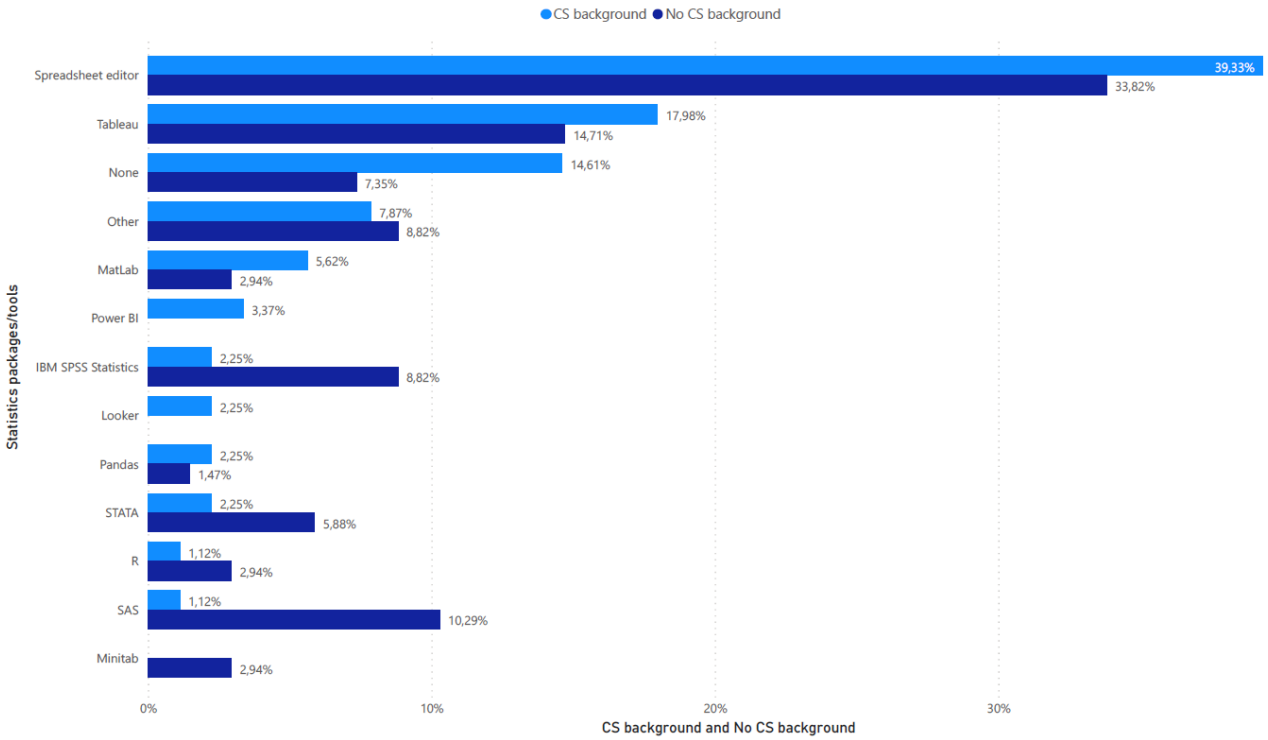

Fig. 4. Statistics packages/tools by background.

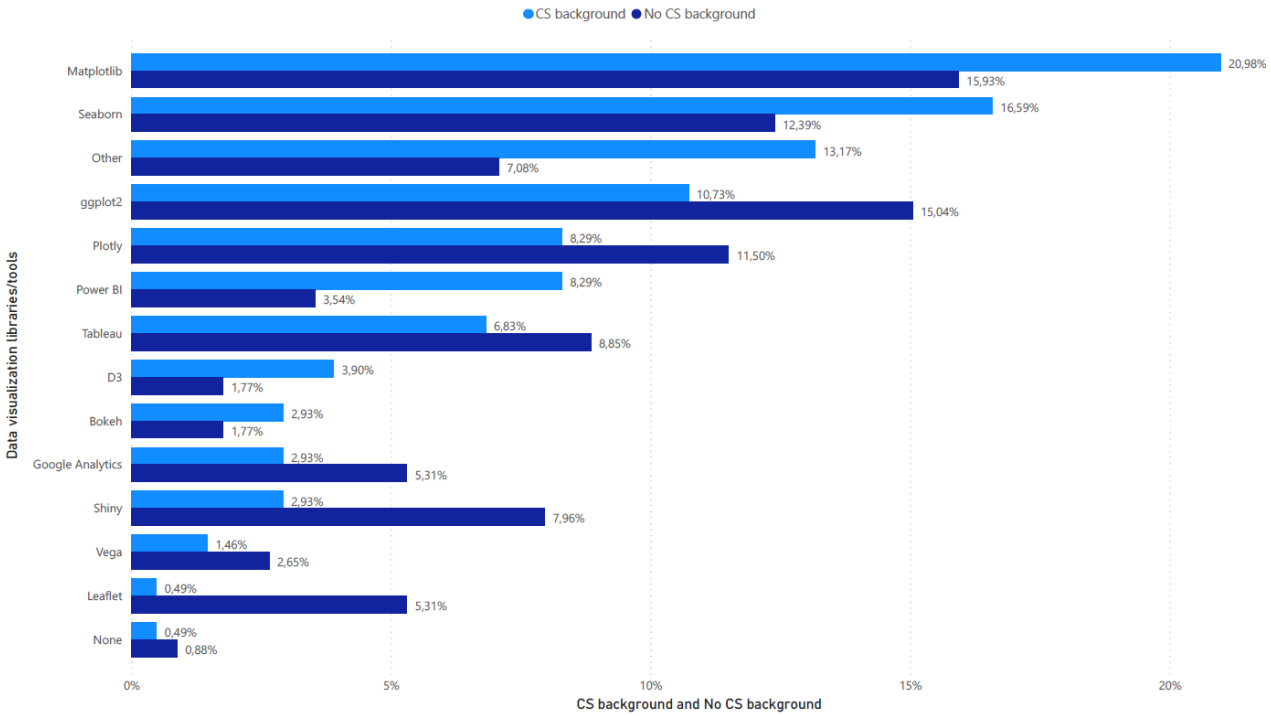

Fig. 5. Data visualization libraries/tools by background.

References

[1] A. Holst, "Data created worldwide 2010-2025 — Statista," 2019. [Online]. Available: https://www.statista.com/statistics/871513/worldwide-data-created/

[2] M. Kubina, M. Varmus, and I. Kubinova, "Use of Big Data for Competitive Advantage of Company," *Procedia Economics and Finance*, vol. 26, pp. 561–565, 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2212567115009557

[3] D. Parkins, "Regulating the internet giants: The world's most valuable resource is no longer oil, but data," *Economist (United Kingdom)*, vol. 413, no. 9035, 2017. [Online]. Available: https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

[4] T. H. Davenport and D. J. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Business Review*, vol. 90, no. 10, p. 5, 2012.

[5] F. H. Grupe and M. M. Owrang, "Data base mining: Discovering new knowledge and competitive advantage," *Information Systems Management*, vol. 12, no. 4, pp. 26–31, 1995.

[6] P. Pereira, J. Cunha, and J. P. Fernandes, "On Understanding Data Scientists," in *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 8 2020, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9127269/