CUNY School of Professional Studies

Data 607, Spring 2023, Project 3

Team Members:
Warner Alexis
John Droescher
Shaya Engelman
Heleine Fouda
Fomba Kassoh

Question:
We were presented with the question:

Which are the most valued data science skills?

Data Sources:
Our initial plan was to draft two surveys: the first would be an open-ended survey with responses that would be aggregated and used to create a larger, closed-ended survey. However, time was too short for this project to collect design and collect sufficient responses to two surveys. While this original plan was being explored, in parallel we moved forward with the second plan.

Our second plan involved directly scraping data from job-posting sites such as Indeed and LinkedIn. While we did make some progress on this plan, we ran into some technical issues with the scraper being detected and blocked by the posting sites. Due to these technical difficulties and the short time window for the project, this approach was abandoned before retrieving any useful data.

A third option for data gathering was identified as looking at previously performed studies – a meta-study approach. Some progress was made in this area in that we were able to find multiple studies with relevant results. Unfortunately, however, were not able to obtain the raw data involved in these studies. The aggregate results of these studies that were included with the study report (Appendix B) were referred to when designing our survey responses.

Ultimately, we went with a single survey with 17 response options, chosen by our group, intended to provide a wide range of identified skill sets. Our survey (Appendix B) was designed to allow respondents to choose and rank five (5) of the available responses. Additionally, we requested information on each respondent's position in the data science community as well as years of experience.

Data Source Caveats:

1) Our dataset, as of 28 Oct 2023, consists of 41 responses. While this does allow us to have some statistical insights, it is a non-significant sample for making more than generalized analysis to support funding for further research and/or indicate general trends.
2) The data collected is not truly randomized as the results were obtained from our personal and professional networks. While some of us have more extensive networks, as humans we tend to connect with those who have a similar mindset. This is likely to result in biases in our dataset.
3) During our analysis, we identified that in designing the survey we failed to restrict respondents to only five (5) responses. As such, some of our respondents assigned values to more than only

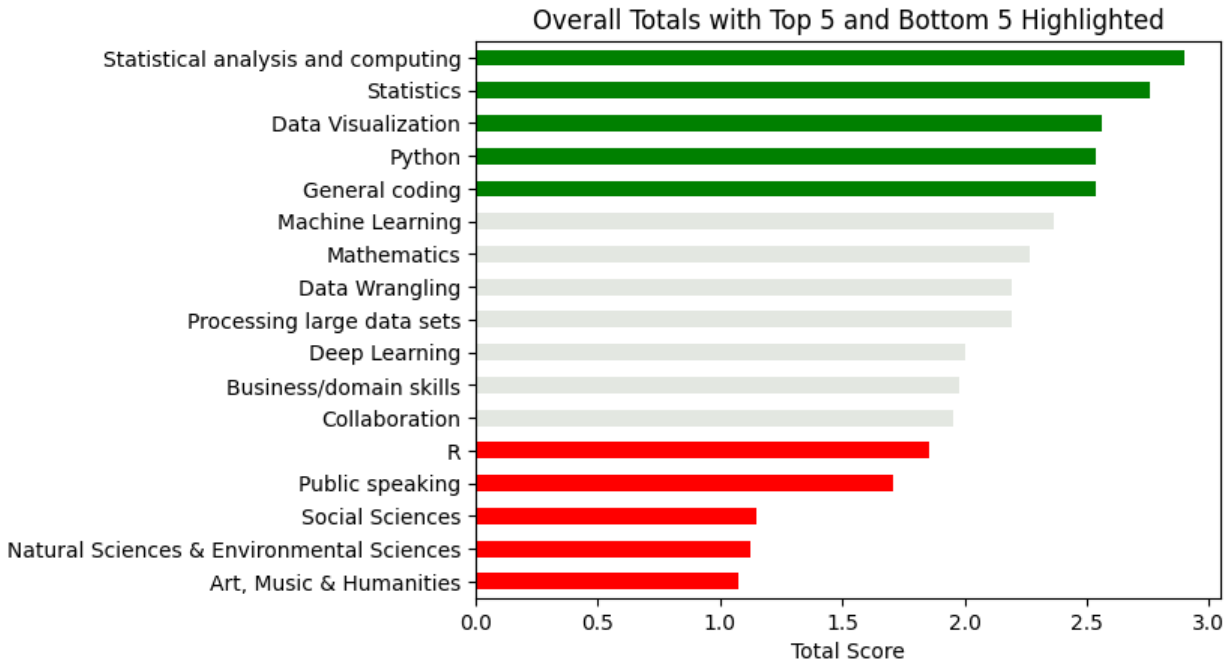
five responses. As such, our data will have some skew that is difficult to remove as we cannot assume how to re-rank their responses.

Data Preparation:
Our data preparation involved three (3) stages: data ingestion into Pandas, conversion from categorical encoding to numerical valuation, and grouping based on remaining categorical features.

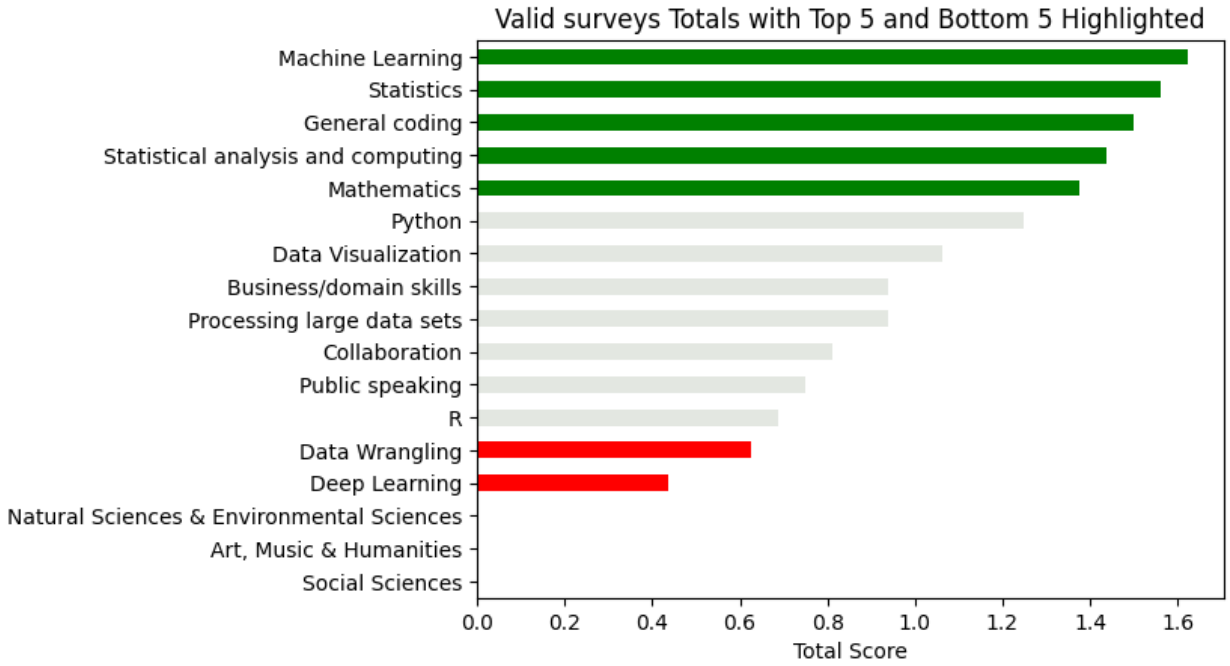
Because our survey was created using Google Forms, we were able to directly load the dataset into Python Pandas using built-in Pandas functions. After loading the data, we renamed some columns to make it easier to work with and converted categorical responses into numerical values, 1-5. As part of our preparation, we calculated the mean value for each response to help analyze different subsets of various sizes.

Next, we grouped the data based on different demographic features we had included in our survey. These included length of experience and type of experience. We also added a label to any invalid response to be able to exclude those results from our analyses. An invalid response was defined as one where a respondent selected other than the requested five responses or the score was greater than 15 points.

Data Analysis:
Once we had our data prepared, we created a bar plot of the different skills respondents said they value. The green bars in the below plot represent the skills that received the highest mean rating of importance. The red bars represent the five skills ranked the least important using the same metric. Initial analysis of this chart suggests the most highly ranked skills were "hard" skills directly associated with data science while "soft" skills like public speaking and collaboration tended to be ranked with much lower importance.

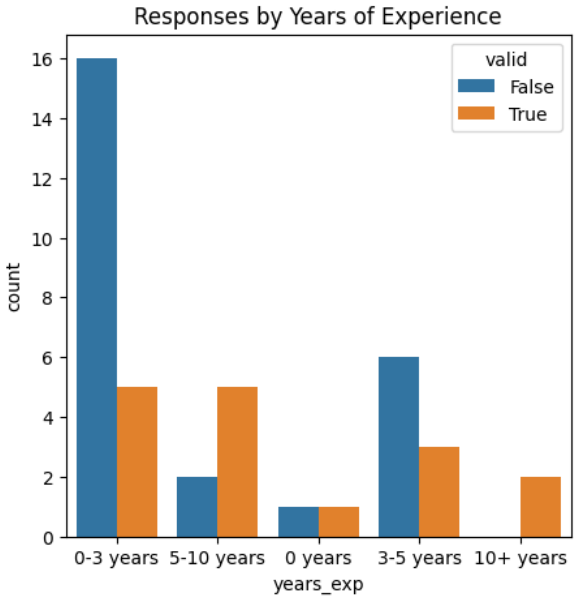


After we sent out our survey, we realized we had allowed respondents to select more than the five responses we intended the survey to allow. This led to skewed results. We recreated our first ranking chart as above but including only survey responses considered valid.

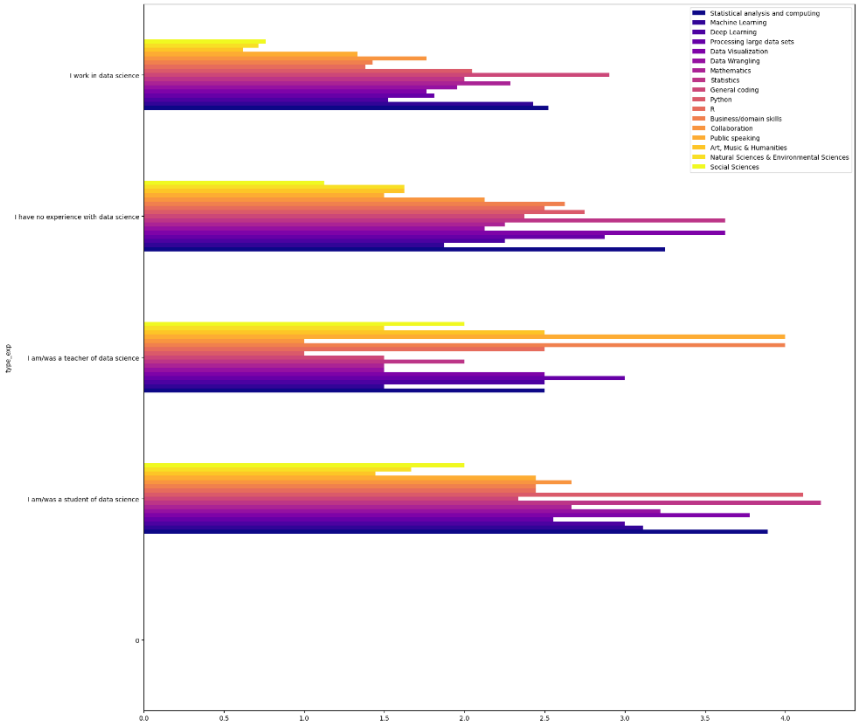


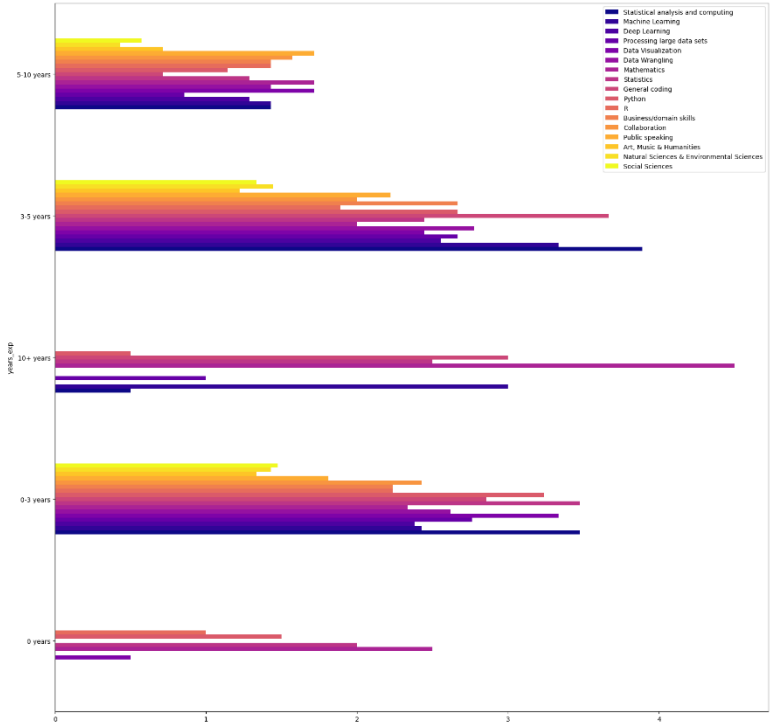
In the new plot of valid responses we find the social science skills were selected only when more than five skills were selected by the respondent. Other than this change, however, the results remained mostly unchanged – "hard" skills tended to be ranked most valuable and "soft" skills tended to be ranked less valuable. The most significant change was machine learning jumping from just outside the top five into the top position.

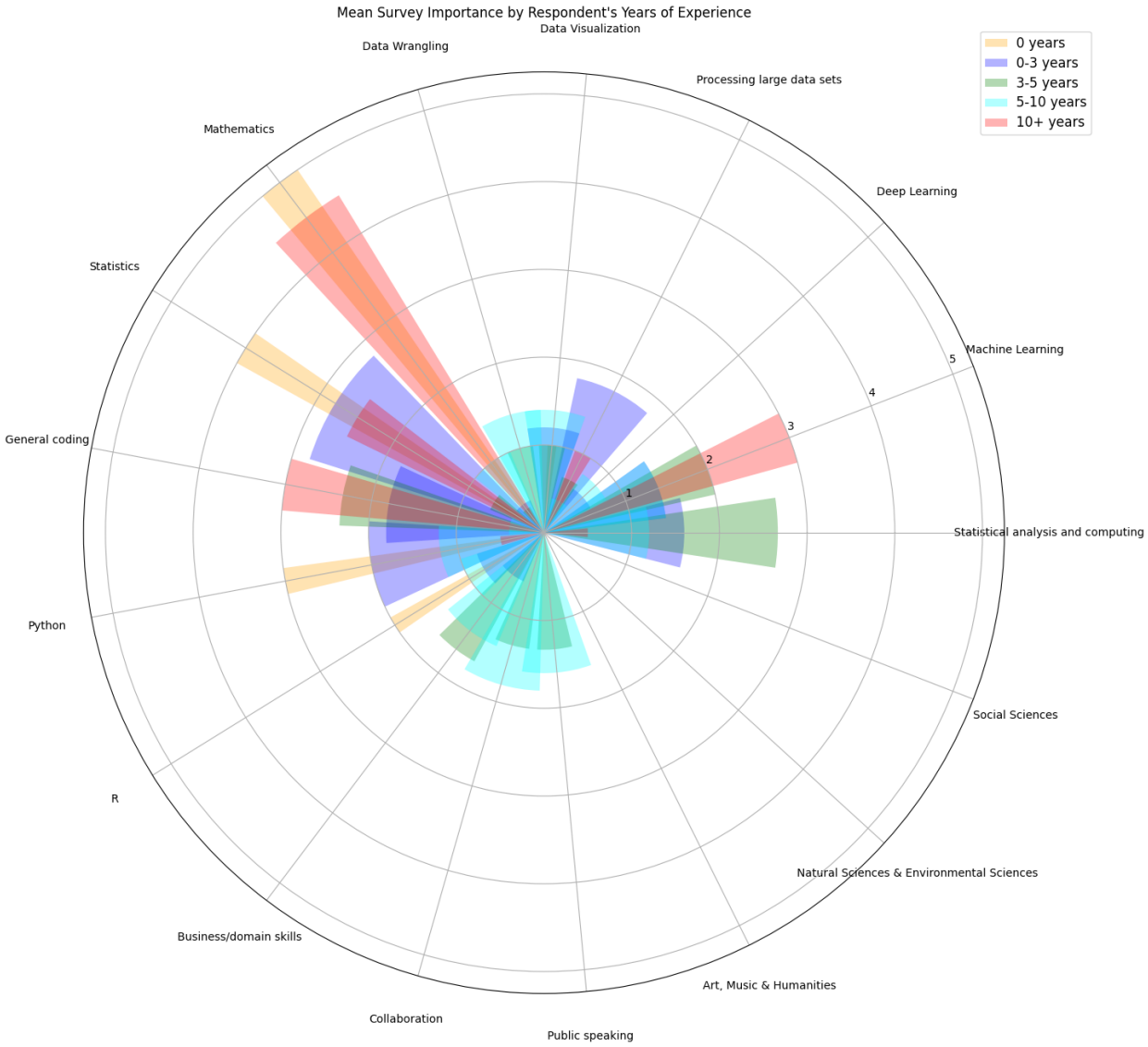
We were interested in how these results were reflected across our demographic categories. The next step was to group your responses by years of experience and valid / invalid. We find that, in general, the younger data scientists tended to leave more invalid responses to the survey.



We then looked at our demographic categories, length of experience and type of experience with data science, plotting the grouped data means. These plots clearly demonstrate how the answers vary by the demographic groupings. This suggests a disagreement between the skills valued by more experienced and less experienced data scientists. Similarly, there is a difference in the distribution of answers based on the type of experience respondents have.







Finally, we combined all the above into a polarized bar chart. From this chart we were clearly able to distinguish between the different years of experience and the mean skill ratings. The difference between the skills most highly valued by highly experienced data scientists and those with significantly less experience is quite clear.

Conclusions:

There is a disconnect between the critical skills identified by students and entry level data scientists versus the critical skills identified by senior data scientists.

Appendix A

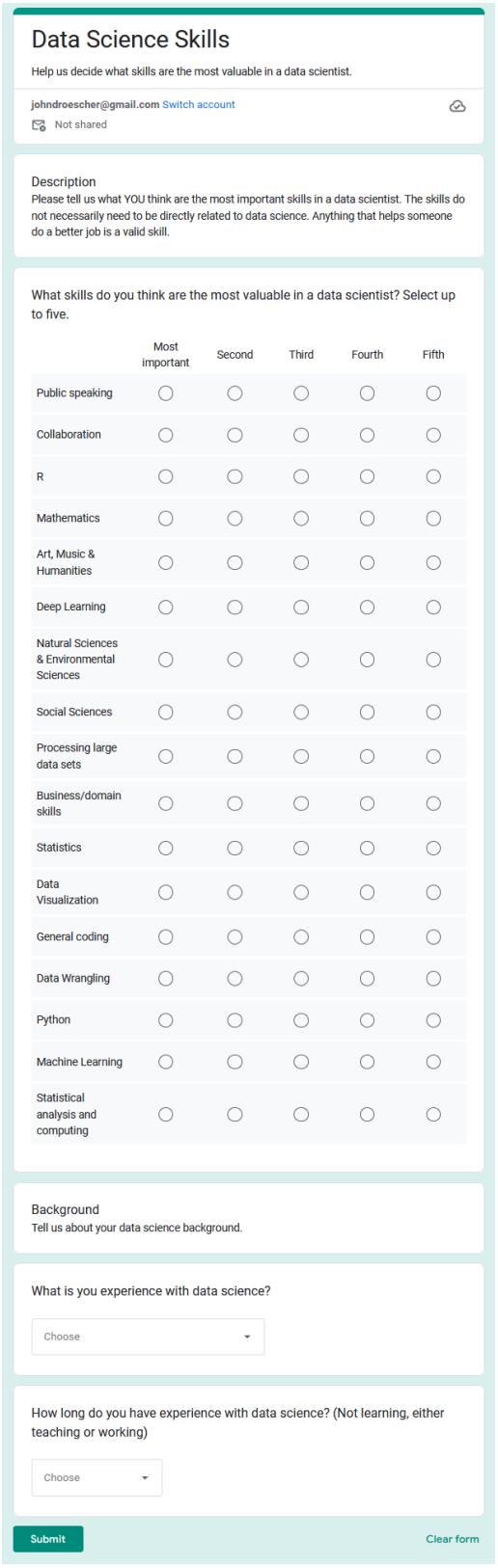
Wilbur W Stanton & Angela A Stanton (2020). Helping Business Students Acquire the Skills Needed for a Career in Analytics: A Comprehensive Industry Assessment of Entry-Level Requirements. Decision Sciences Journal of Innovative Education. Volume 18, Issue1. Retrieved from: https://onlinelibrary-wiley-com.ez.lib.jjay.cuny.edu/share/N9Y2QSCNFDAFZ3J6YUCB?target=10.1111/dsji.12199

Table 1: Top 20 Credential requirements for an entry-level position.

Data Science Position requirements	%	Data Analytics Position requirements	%	Business Analytics Position requirements	%
Prior experience	80.7%	Prior experience	76.1%	Prior experience	78.4%
Degree in computer science	38.2%	Degree in business	44.3%	Degree in business	67.1%
Degree in management	33.5%	Degree in management	42.3%	Degree in management	49.3%
Degree in engineering	31.3%	Degree in engineering	26.4%	Bachelor's degree	27.1%
Degree in business	29.2%	5+ years of experience	25.7%	5+ years of experience	23.8%
Bachelor's degree	25.9%	Bachelor's degree	25.6%	Degree in engineering	20.5%
5+ years of experience	20.7%	Degree in computer science	23.3%	Degree in computer science	19.9%
Degree in information systems	13.7%	Certifications	15.9%	Degree in marketing	19.4%
Master's degree	11.1%	Degree in marketing	14.3%	Degree in business intelligence	15.3%
Degree in information science	10.3%	Degree in statistics	12.9%	Certifications	14.9%
Degree in IT	10.0%	Degree in information Systems	12.2%	Degree in finance	13.9%
Degree in mathematics	9.8%	Master's degree	11.5%	Degree in information systems	13.1%
Degree in statistics	9.0%	Degree in mathematics	10.1%	Quantitative degree	13.0%
Degree in decision science	6.8%	Degree in IT	10.0%	Degree in statistics	12.2%
Quantitative degree	6.4%	Degree in business intelligence	8.8%	Degree in IT	11.2%
Degree in MIS	6.0%	Quantitative degree	8.8%	Master's degree	9.7%
1-3 Years of experience	5.8%	Degree in finance	8.4%	Microsoft certifications	8.6%
Degree in marketing	4.7%	1-3 Years of experience	7.8%	Degree in operations mgnt	8.5%
Degree in business intelligence	4.6%	Degree in operations mgnt	7.3%	Degree in MIS	7.5%
Degree in operations mgnt	3.7%	Degree in economics	6.4%	Degree in accounting	7.5%

Appendix B

https://docs.google.com/forms/d/e/1FAIpQLSfJyN2MdpCHNpXymslx1MlkiKmzbpKnrCoZDOH9lXRtzY8UgQ/viewform



Data Science Skills

Help us decide what skills are the most valuable in a data scientist.

johndroescher@gmail.com Switch account

Not shared

Description

Please tell us what YOU think are the most important skills in a data scientist. The skills do not necessarily need to be directly related to data science. Anything that helps someone do a better job is a valid skill.

What skills do you think are the most valuable in a data scientist? Select up to five.

	Most important	Second	Third	Fourth	Fifth
Public speaking	○	○	○	○	○
Collaboration	○	○	○	○	○
R	○	○	○	○	○
Mathematics	○	○	○	○	○
Art, Music & Humanities	○	○	○	○	○
Deep Learning	○	○	○	○	○
Natural Sciences & Environmental Sciences	○	○	○	○	○
Social Sciences	○	○	○	○	○
Processing large data sets	○	○	○	○	○
Business/domain skills	○	○	○	○	○
Statistics	○	○	○	○	○
Data Visualization	○	○	○	○	○
General coding	○	○	○	○	○
Data Wrangling	○	○	○	○	○
Python	○	○	○	○	○
Machine Learning	○	○	○	○	○
Statistical analysis and computing	○	○	○	○	○

Background

Tell us about your data science background.

What is you experience with data science?

Choose

How long do you have experience with data science? (Not learning, either teaching or working)

Choose

Submit

Clear form