



ADAM SPANNBAUER

Python Developers Survey 2019

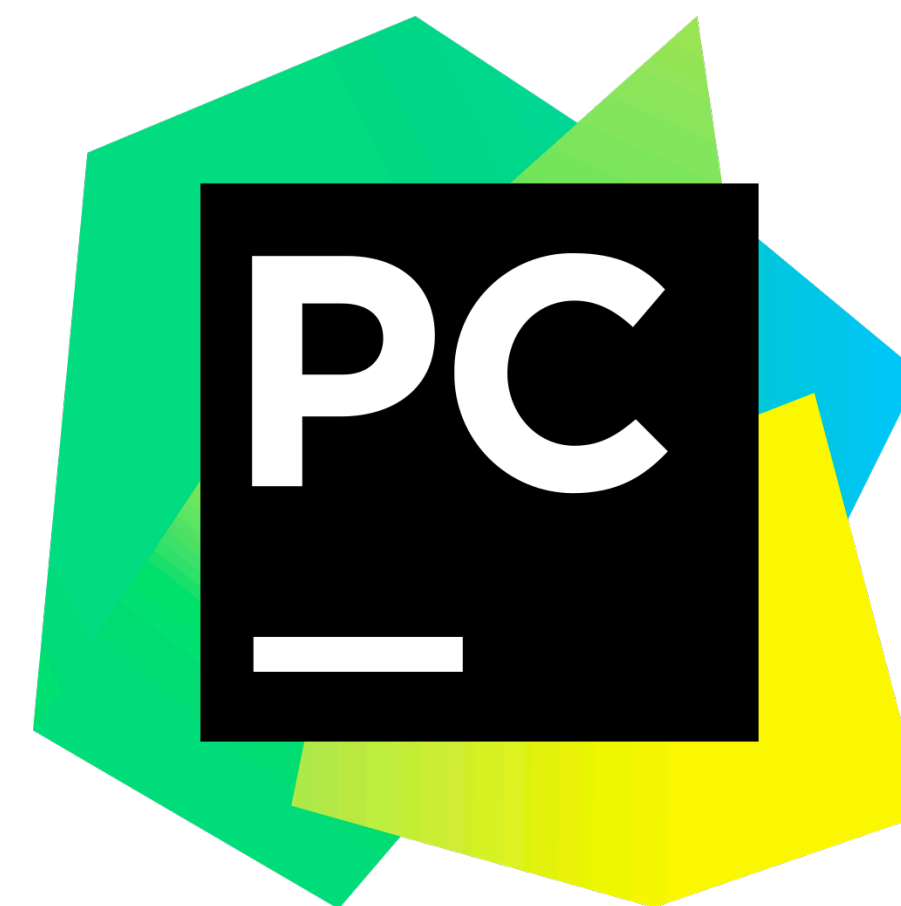
Cluster Analysis

AGENDA

- ▶ Introduce survey questions / data
- ▶ Describe the clustering process
- ▶ Interpret cluster membership
- ▶ Present limitations / next steps of analysis

ABOUT THE SURVEY

- ▶ Survey performed by JetBrains in 2017, 2018, & 2019
- ▶ PyCharm (a JetBrains product) is a popular Python IDE
- ▶ 47,308 respondents in 2019 (no requirement to answer all questions)



PyCharm

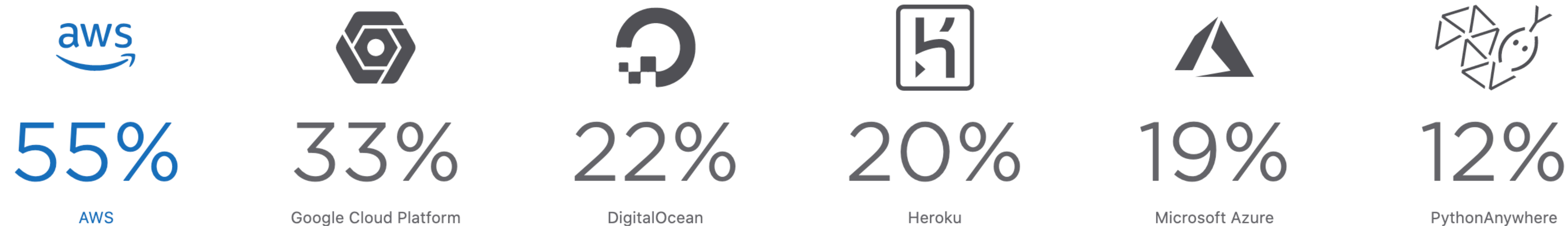
ABOUT THE SURVEY

- ▶ Includes questions on:
 - ▶ Python usage / programming experience
 - ▶ Company / team size
 - ▶ Area of work based on:
 - ▶ What tools do you use?
 - ▶ What's your role?

ABOUT THE SURVEY

- ▶ JetBrains has published some great exploratory analysis of the data:
<https://www.jetbrains.com/lp/python-developers-survey-2019/>

Top Cloud platforms > 100%

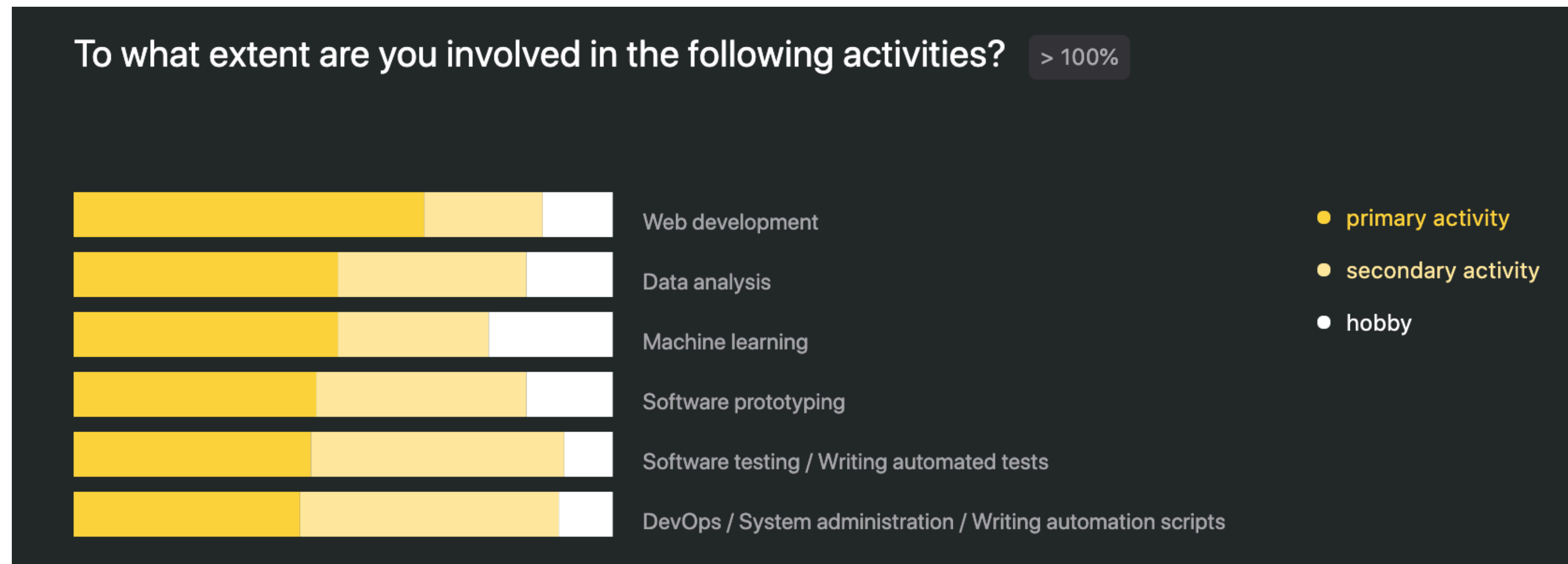


Google Cloud Platform usage has grown 2 percentage points and Microsoft Azure 4 percentage points since last year, while DigitalOcean and Heroku have decreased their shares.

In addition to the popular cloud platforms listed above, OpenStack got 6%, Linode 5%, OpenShift 3%, and Rackspace 1%. Other cloud platforms that were not listed in the options were also identified by 8% of respondents. The most common write-ins were OVH, Vultr, Hetzner and Scaleway.

ABOUT THE SURVEY

- ▶ JetBrains has published some great exploratory analysis of the data:
<https://www.jetbrains.com/lp/python-developers-survey-2019/>



From JetBrains analysis. Example of a "what's your role?" question. Results have been cropped to top 6 bars from original viz.

ABOUT THE SURVEY


- ▶ JetBrains has published some great exploratory analysis of the data:
<https://www.jetbrains.com/lp/python-developers-survey-2019/>
- ▶ Want to analyze the data yourself? It is free to download from the survey results link above.

ABOUT THE CLUSTERING PROCESS

- ▶ Only ordinal type questions were used for clustering
 - ▶ Examples:
 - ▶ How many years of professional coding experience do you have?
 - ▶ How often do you use Version Control Systems?
 - ▶ How involved are you with Web development?

ABOUT THE CLUSTERING PROCESS

- ▶ Only **ordinal** type questions were used for clustering
 - ▶ Examples:
 - ▶ How many years of professional coding experience do you have?
 - ▶ How often do you use Version Control Systems?
 - ▶ How involved are you with Web development?


- 
1. "Less than 1 year"
 2. "1-2 years"
 3. "3-5 years"
 4. "6-10 years"
 5. "11+ years"

ABOUT THE CLUSTERING PROCESS

- ▶ Only **ordinal** type questions were used for clustering


- ▶ Examples:

- ▶ How many years of professional coding experience do you have?
- ▶ How often do you use Version Control Systems?
- ▶ How involved are you with Web development?

- 
1. "Never or Almost never"
 2. "From time to time"
 3. "Often"

ABOUT THE CLUSTERING PROCESS

- ▶ Only **ordinal** type questions were used for clustering
 - ▶ Examples:
 - ▶ How many years of professional coding experience do you have?
 - ▶ How often do you use Version Control Systems?
 - ▶ How involved are you with Web development?

- 
1. "Hobby"
 2. "Secondary Activity"
 3. "Primary Activity"

ABOUT THE CLUSTERING PROCESS

- ▶ Only ordinal type questions were used for clustering
 - ▶ Examples:
 - ⦿ How many years of professional coding experience do you have?
 - ❖ How often do you use Version Control Systems?
 - ◆ How involved are you with Web development?

3 main categories of questions used:

- ⦿ General
- ❖ "How often do you..." (tools)
- ◆ "How involved are you with..." (role)

ABOUT THE CLUSTERING PROCESS

- ▶ Number of records in clustering process: 10,169
 - ▶ ~21% of original 47,308
- ▶ This was based on lack of responses to the ordinal questions being used*

*A lack of response is tricky to define on 2 of the question categories being used; see next slide

ABOUT THE CLUSTERING PROCESS

- ▶ The tools / roles questions were presented as below.
- ▶ A lack of response might be interpreted as missing or a lack of involvement

How involved are you with:

	Hobby	Secondary Activity	Primary Activity
Web Development			
Machine Learning			
Software prototyping			

ABOUT THE CLUSTERING PROCESS

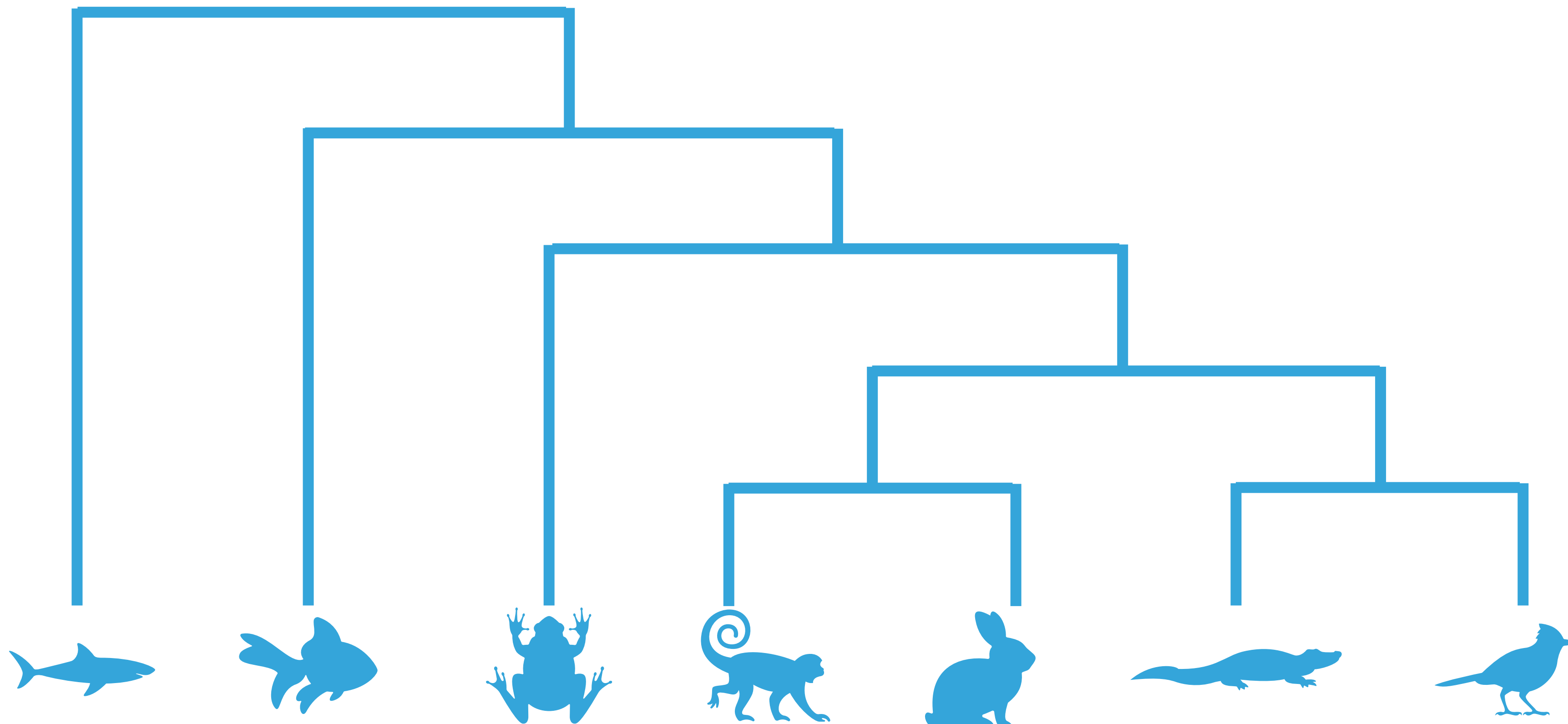
- ▶ The tools / roles questions were presented as below.
- ▶ A lack of response might be interpreted as missing or a lack of involvement.
- ▶ This analysis treated missing-ness on these questions as a lack of involvement

How involved are you with:

	Hobby	Secondary Activity	Primary Activity
Web Development			
Machine Learning			
Software prototyping			

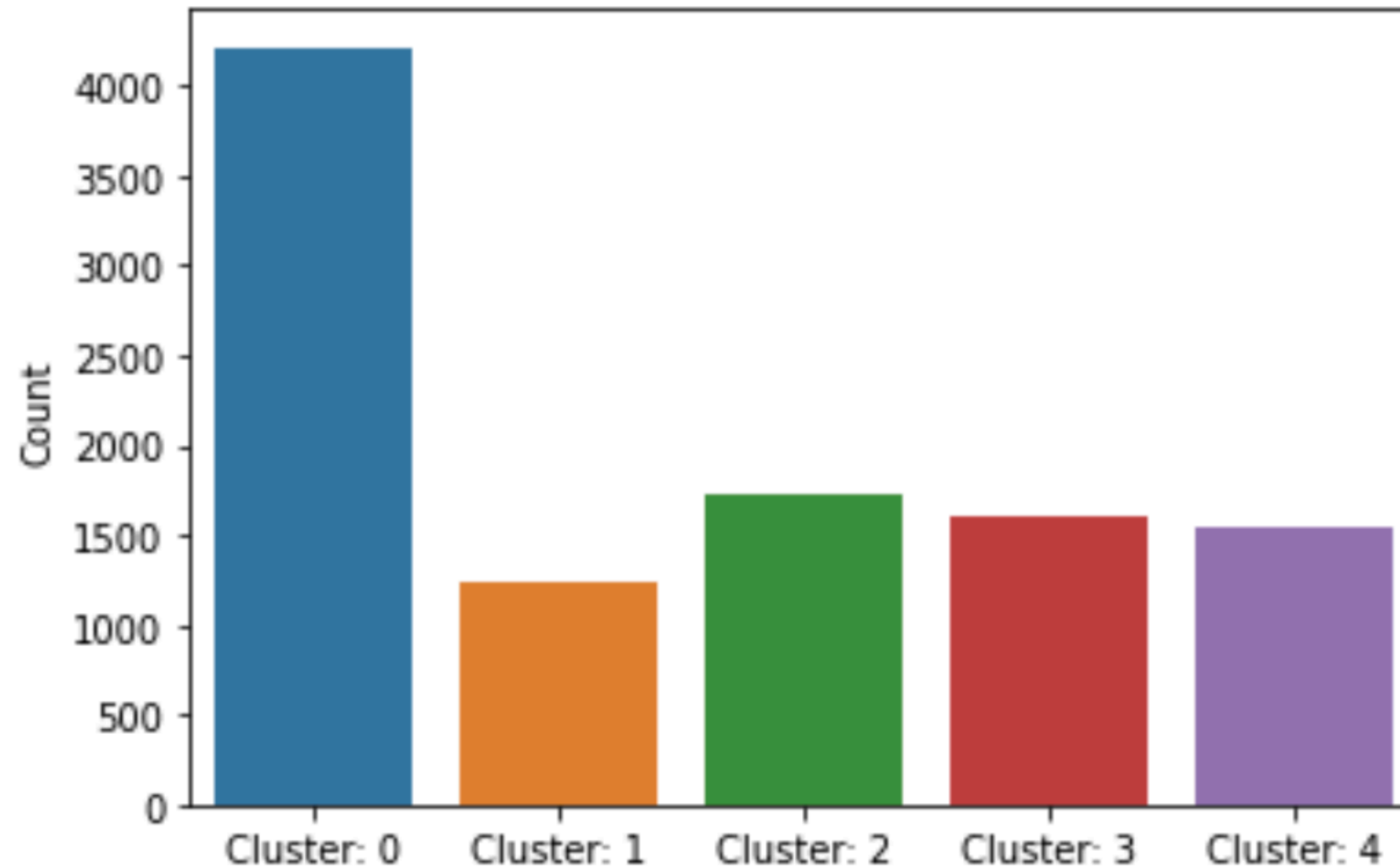
ABOUT THE CLUSTERING PROCESS

- Hierarchical clustering was used to cluster similar observations into 5 groups



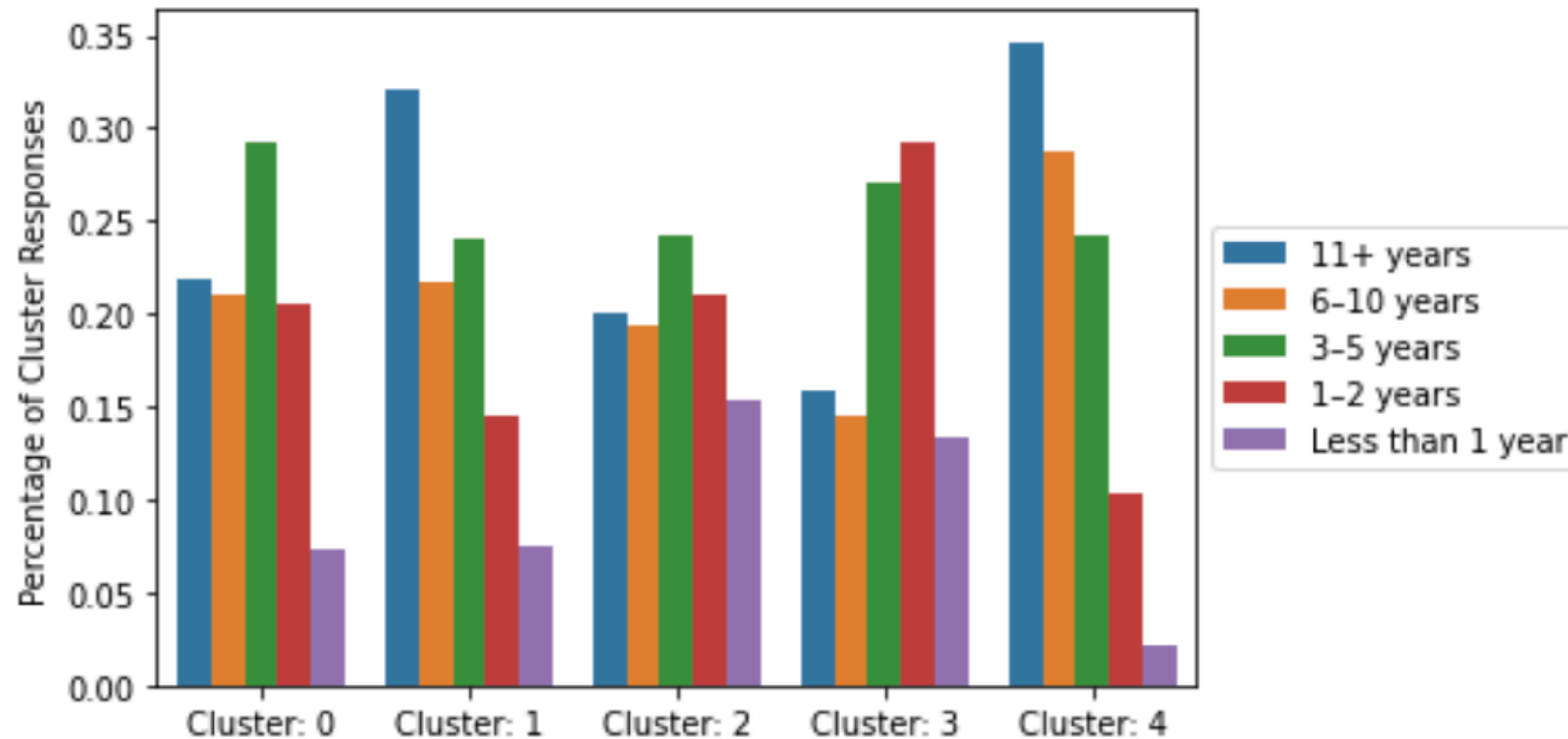
ABOUT THE CLUSTERS

► Cluster size



ABOUT THE CLUSTERS

- ▶ How many years of professional coding experience do you have?



ABOUT THE CLUSTERS

- ▶ How many years of professional coding experience do you have?



Clusters in order of average experience (most to least):

- ▶ Cluster 4
- ▶ Cluster 1
- ▶ Cluster 0
- ▶ Cluster 2
- ▶ Cluster 3

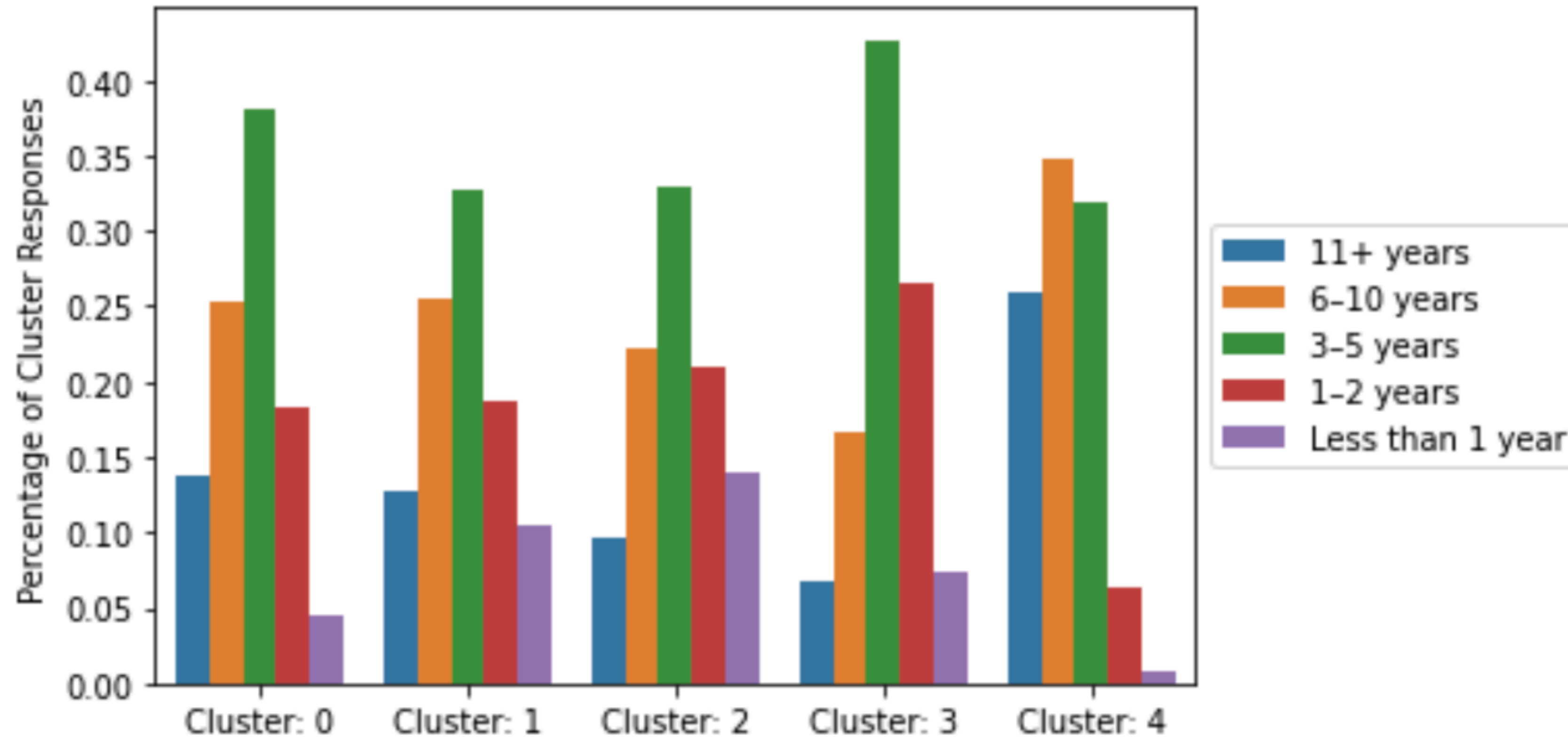
ABOUT THE CLUSTERS

► How many years of professional coding experience do you have?



ABOUT THE CLUSTERS

- ▶ How long have you been programming in Python?



ABOUT THE CLUSTERS

► How long have you been programming in Python?



Clusters in order of average experience (most to least):

- Cluster 4
- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3

ABOUT THE CLUSTERS

- ▶ How long have you been programming in Python?

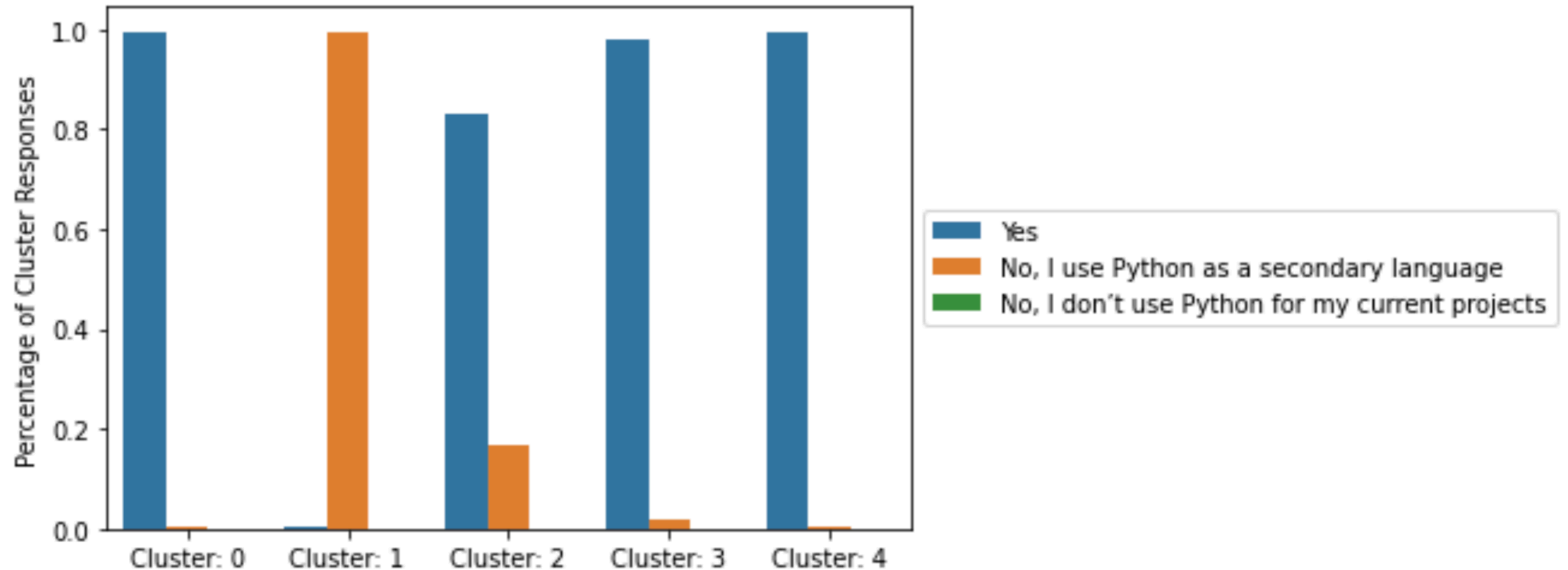


Clusters in order of average experience (most to least):

- ▶ Cluster 4
- ▶ Cluster 0
- ▶ Cluster 1
- ▶ Cluster 2
- ▶ Cluster 3

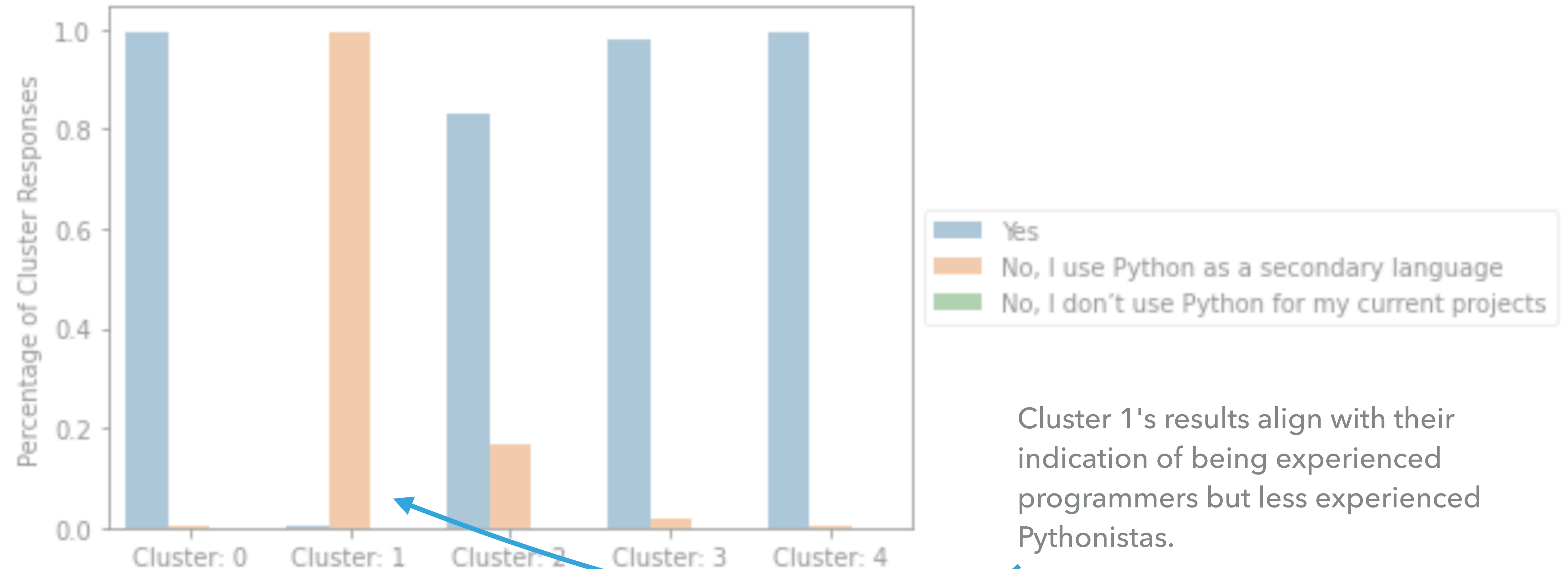
ABOUT THE CLUSTERS

- ▶ Is Python the main language you use for your current projects?



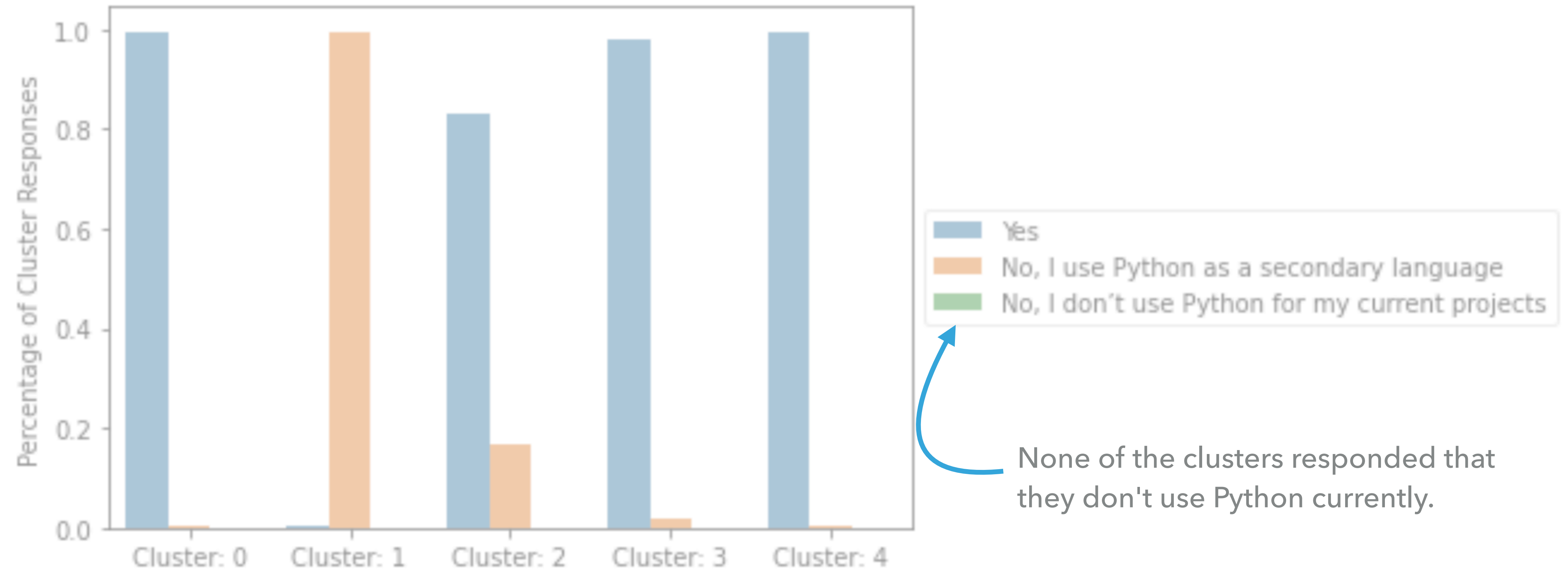
ABOUT THE CLUSTERS

- ▶ Is Python the main language you use for your current projects?



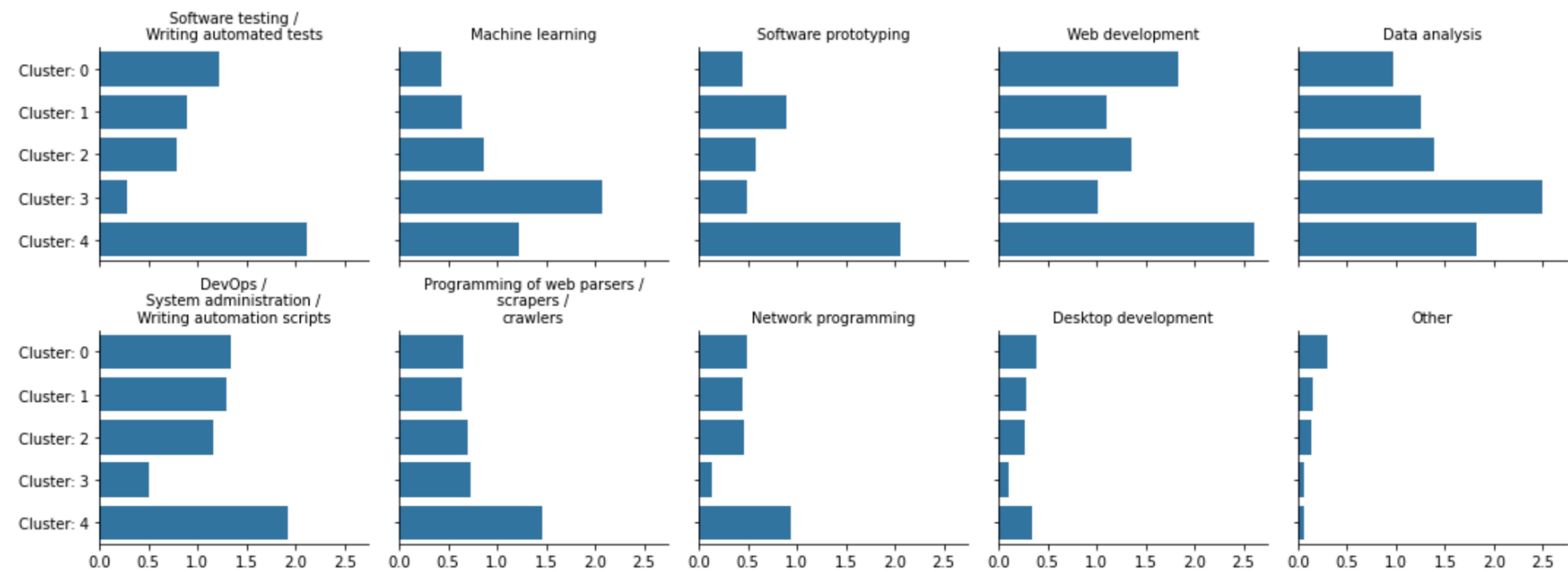
ABOUT THE CLUSTERS

- ▶ Is Python the main language you use for your current projects?



ABOUT THE CLUSTERS

► How involved are you with... ?

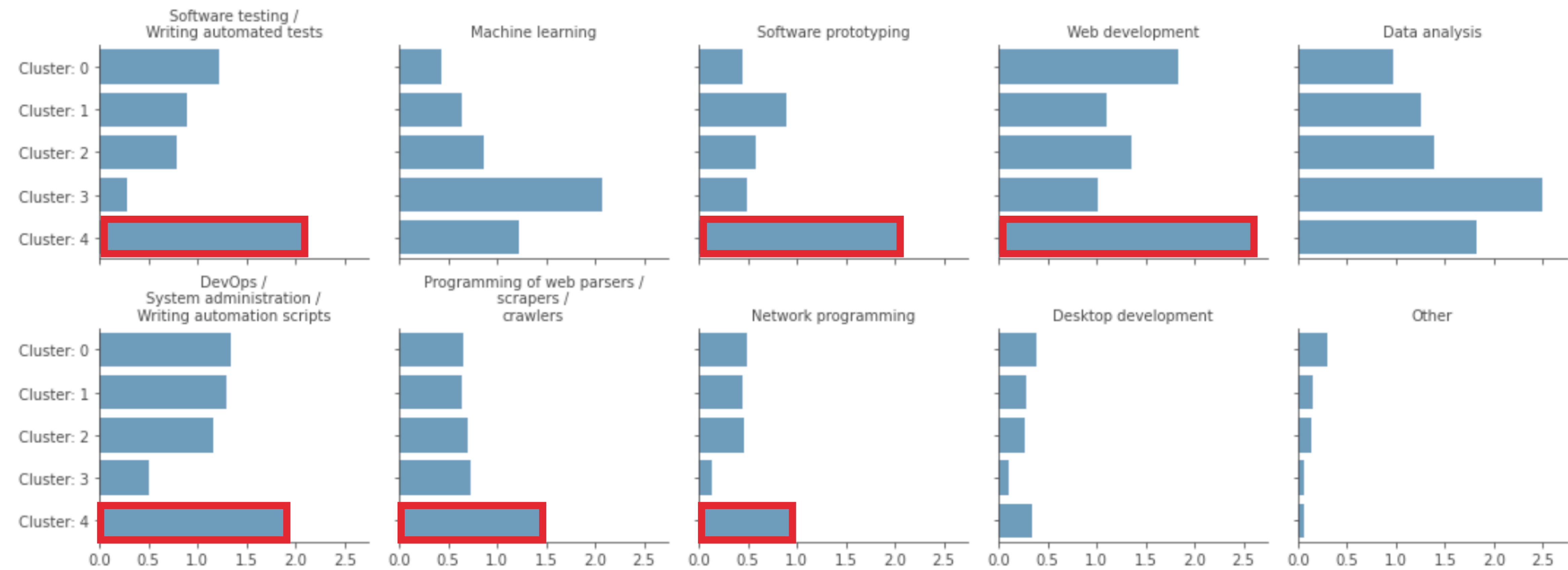


There were other 'involved' questions, but these were the top 10 for differentiating cluster members.

ABOUT THE CLUSTERS

Our most experienced cluster is the most involved with varying techniques

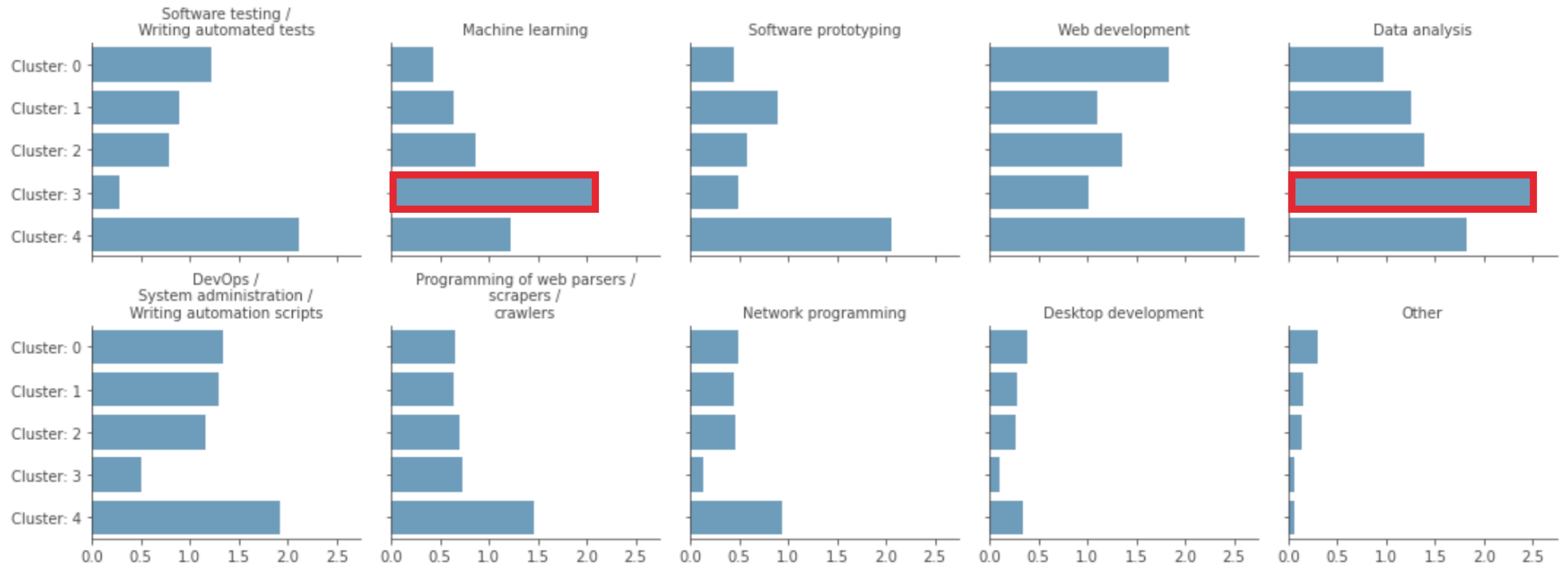
► How involved are you with... ?



ABOUT THE CLUSTERS

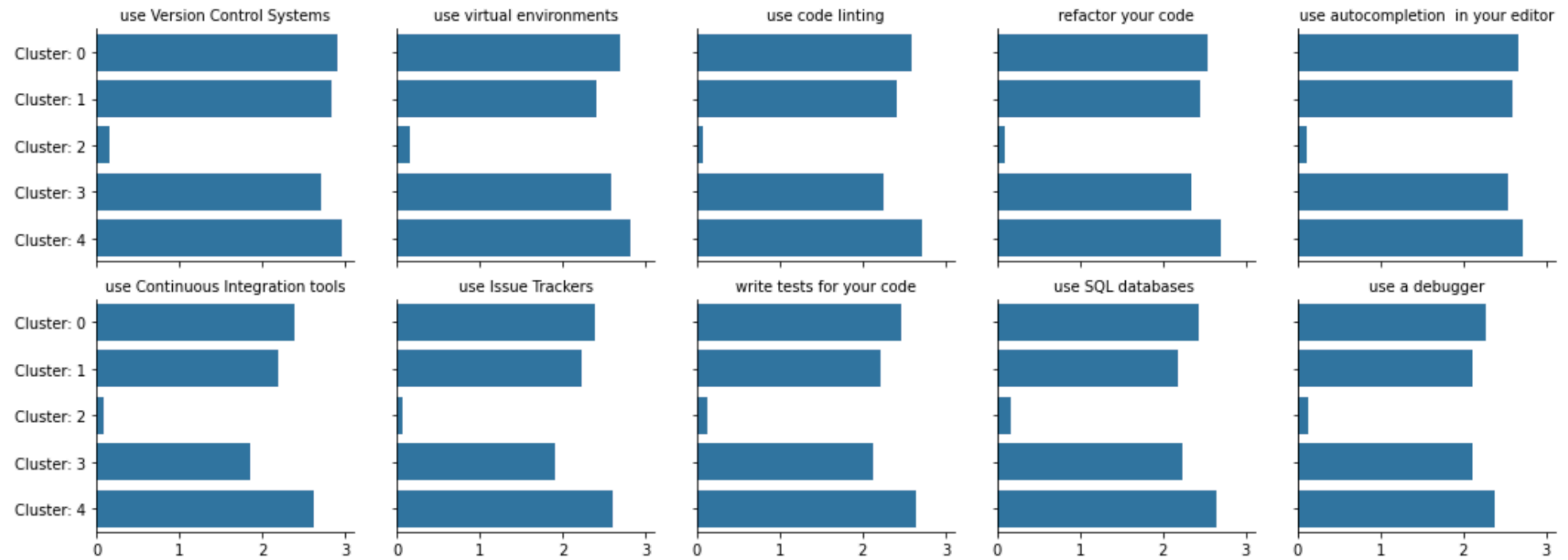
Our least experienced cluster is the most involved with more "Data Science" techniques.

► How involved are you with... ?



ABOUT THE CLUSTERS

► How often do you... ?

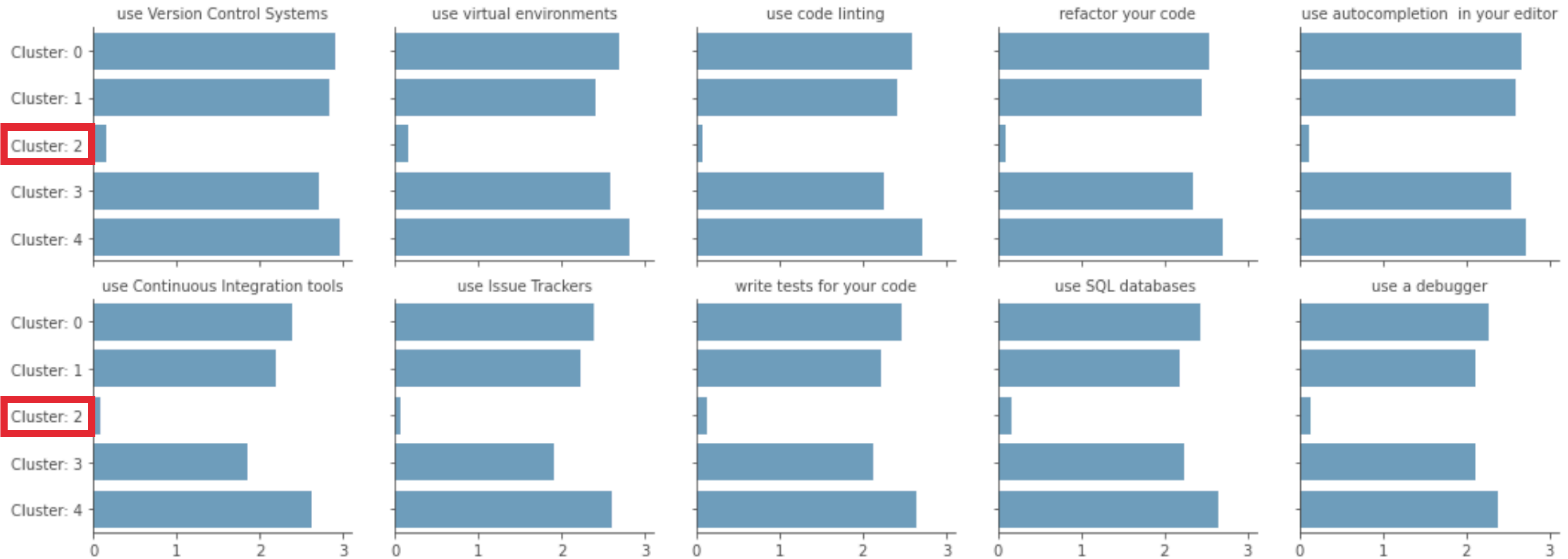


There were other 'how often' questions, but these were the top 10 for differentiating cluster members.

ABOUT THE CLUSTERS

Cluster 2 was by far the least likely to mark that they did any of these activities 🤔

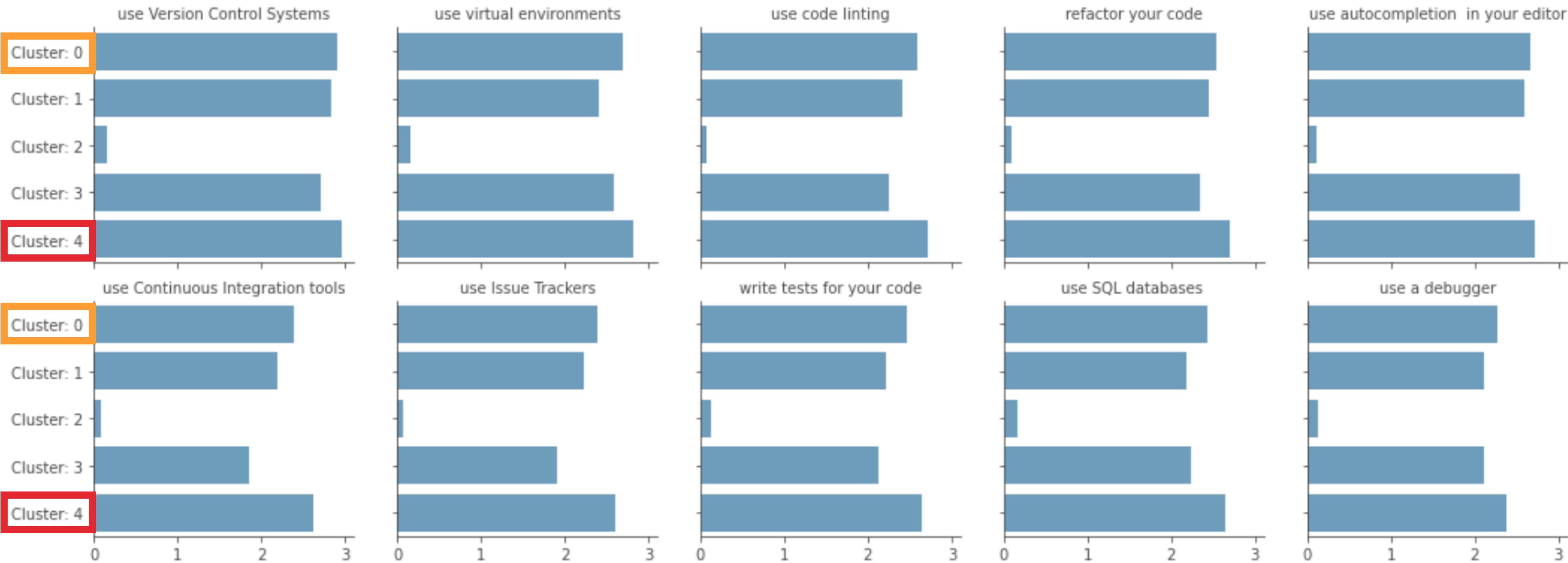
► How often do you... ?



ABOUT THE CLUSTERS

Cluster 4 (most experienced programmers) does all of these activities the most. Cluster 0 is right behind.

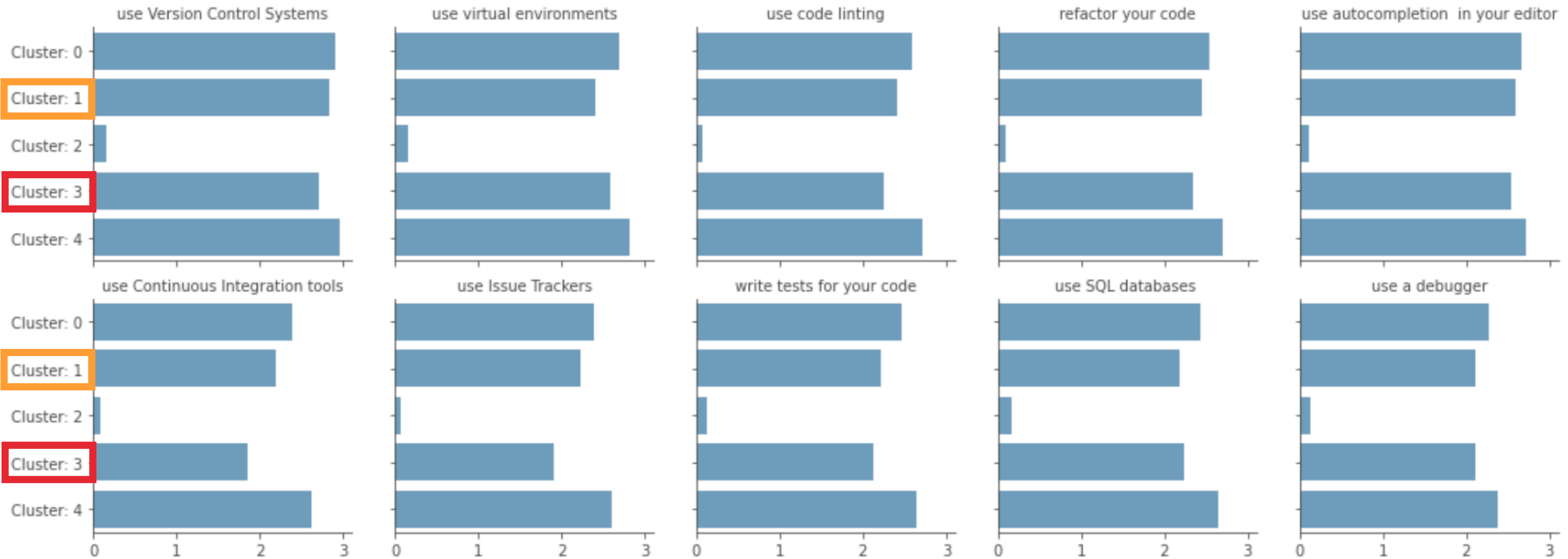
► How often do you... ?



ABOUT THE CLUSTERS

► How often do you... ?

Cluster 1 (non-Pythonistas) & Cluster 3 (less experienced programmers) are hard to differentiate based on these responses.



ABOUT THE CLUSTERS – SUMMARY STATEMENTS

▶ Cluster 1

- ▶ Python is not their first choice for programming. They are experienced programmers, and they use varying technologies and techniques

▶ Cluster 2

- ▶ Might not be a cluster of interest. Membership seems mostly based on not answering the "How often do you... ?" series of questions.

ABOUT THE CLUSTERS – SUMMARY STATEMENTS

▶ Cluster 3

- ▶ Less experienced programmers with a focus on data science style tasks. Is their lack of programming experience due to a lack of experience in industry or due to a heavier focus on the math/stats of data science tasks?

▶ Cluster 4

- ▶ Most experienced Programmers and most experienced Pythonistas. They are more classical programmers than they are data scientists.

ABOUT THE CLUSTERS – SUMMARY STATEMENTS

- ▶ Cluster 0

- ▶ Appears to be more of a 'catch-all' cluster for the remaining participants. Future work might include a more granular cluster analysis.

LIMITATIONS

- ▶ This analysis ignored many of the survey participants and many of the questions
 - ▶ These decisions were made due to a short project timeline
- ▶ Treating a lack of response as an indication of not using a technology / technique should be revisited
- ▶ There is survey data available for 2017, 2018, & 2019. Only 2019 was considered. Could previous results be leveraged in analysis?

(POTENTIAL) **ACTIONABLE NEXT STEPS**

- ▶ If these clusters were to be used to segment PyCharm customers. The main groups might be:
 1. Less experienced programmers, interested in more Data Science focused features (Cluster 3)
 2. More experienced programmers, interested in more traditional Software Engineering features (Clusters 1 & 4)
- ▶ Analyzing / polling these groups might lead to valuable market analysis of these segments