DataCamp

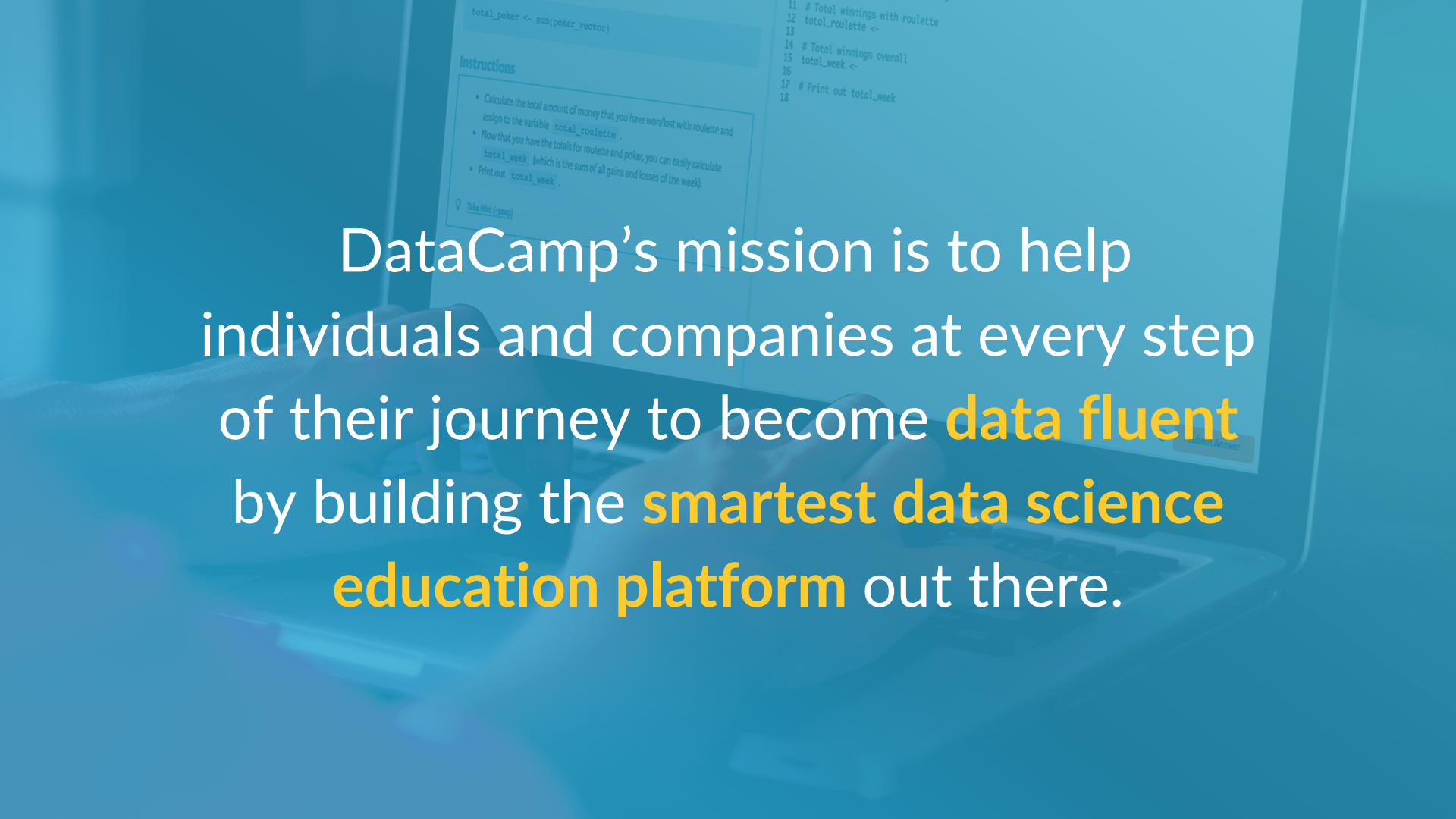
# Using Data to Drive Curriculum Development

Chester Ismay
**Senior Curriculum Lead**

2018-07-30

PDF of slides available at http://bit.ly/ismay-jsm

# What is DataCamp?




DataCamp's mission is to help individuals and companies at every step of their journey to become **data fluent** by building the **smartest data science education platform** out there.

R python SQL Spark git >_ Shell

1

**Learn**

Acquire new skills. Choose from over 100 intuitive Courses on R, Python, SQL, Git, Shell,...

2

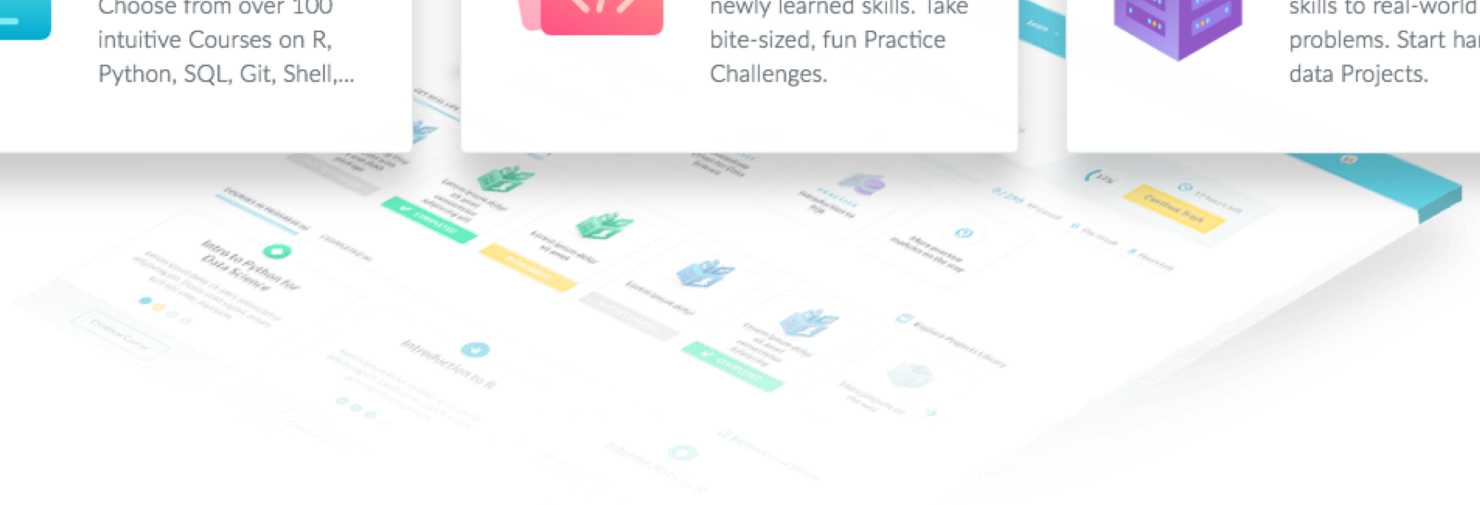
**Practice**

Sharpen and train your newly learned skills. Take bite-sized, fun Practice Challenges.

3

**Build**

Apply your data science skills to real-world problems. Start hands-on data Projects.



DataCamp

Course Outline

EXERCISE

## Visualizing the weights

Graphics are a much better way of visualizing data than just staring at the raw data frame! Therefore, in this activity, we will load the data visualization package `ggplot2` to create a histogram of the survey weights.

INSTRUCTIONS 100 XP

- Load the data visualization package `ggplot2` .
- From the pre-loaded data frame, `ce` , construct a histogram of the survey weights.
- Remember that the survey weights are stored in the column labeled `FINLWT21` .

Take Hint (-30 XP)

SCRIPT.R

```
1  # Load ggplot2
2  library(___)
3
4  # Construct a histogram of the weights
5  ggplot(data = ___, mapping = aes(x = ___)) +
6      geom____()
```

Run Code | Submit Answer

R CONSOLE | SLIDES

```
>
```

## Complete the script to produce the output shown

**OUTPUT**

```
-ducation
```

**SCRIPT**

```
z = "education"
print(z.          ("e", "-"))
```

Try Code (3)

1 change 2 swap 3 edit 4 replace 5 alter

Check

DataCamp

How it works | Report an Issue | BETA

PROJECT INSTRUCTIONS

## Task 1: Instructions

First, get the data into your workspace and summarize it by year. Do you notice a pattern over time?

- Read `datasets/breath_alcohol_ames.csv` into your workspace using the `read_csv()` function. Save it as `ba_data`.
- Count how many tests were administered in each `year` using the `count()` function, creating a new data set called `ba_year`.

## Good to know

If you have taken the Introduction to the Tidyverse, this project is for you! You may also find this data transformation cheat sheet and this `ggplot2` cheat sheet helpful. For even more, visit the `tidyverse` documentation.

Previous Task | Next Task

PROJECT: WHO IS DRUNK AND WHEN IN AMES, IOWA?

jupyter notebook (autosaved)

File Edit View Insert Cell Kernel Help | Not Trusted | R O

Markdown

### 1. Breath alcohol tests in Ames, Iowa, USA

Ames, Iowa, USA is the home of Iowa State University, a land grant university with over 36,000 students. By comparison, the city of Ames, Iowa, itself only has about 65,000 residents. As with any other college town, Ames has had its fair share of alcohol-related incidents. (For example, Google 'VEISHEA riots 2014'.) We will take a look at some breath alcohol test data from Ames that is published by the State of Iowa.


Check Project

1. How do we evaluate and improve our curriculum?
2. How do we decide what to teach next?

# How can we evaluate an online course?

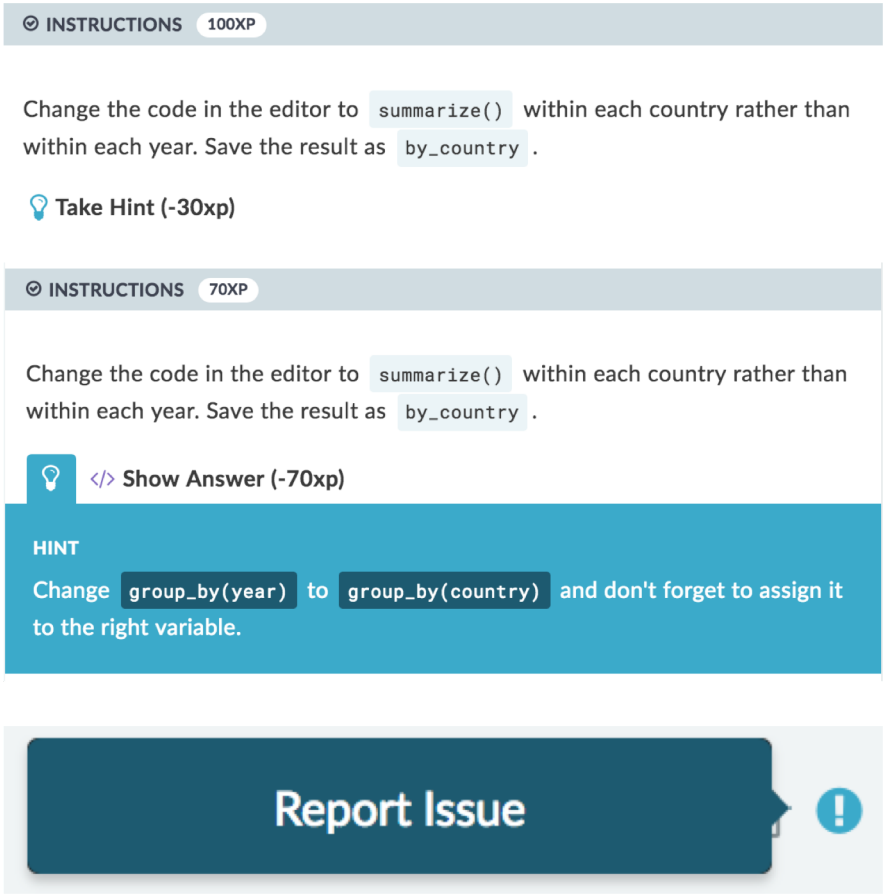
# Exercise quality metrics

DataCamp

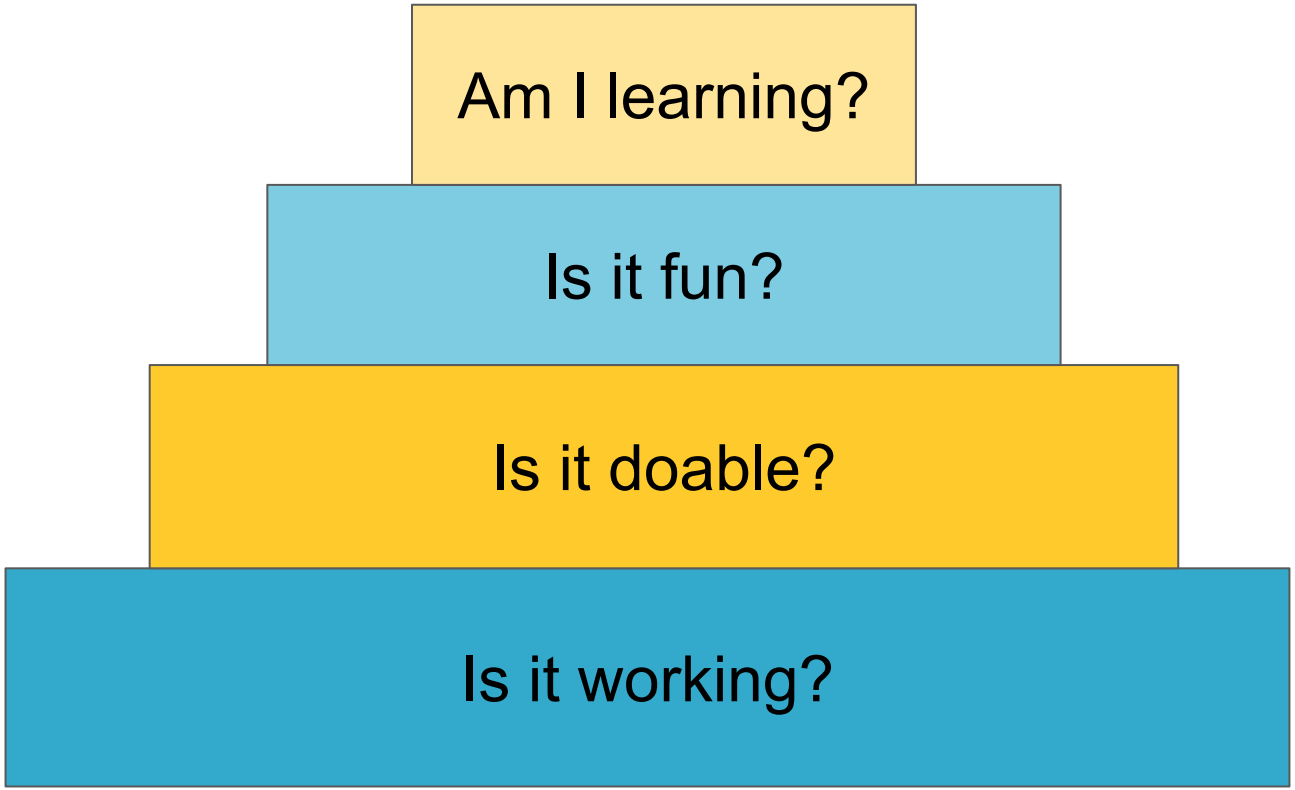
% asked hint

% asked solution

% reporting an issue



# Hierarchy of content quality




COMING SOON

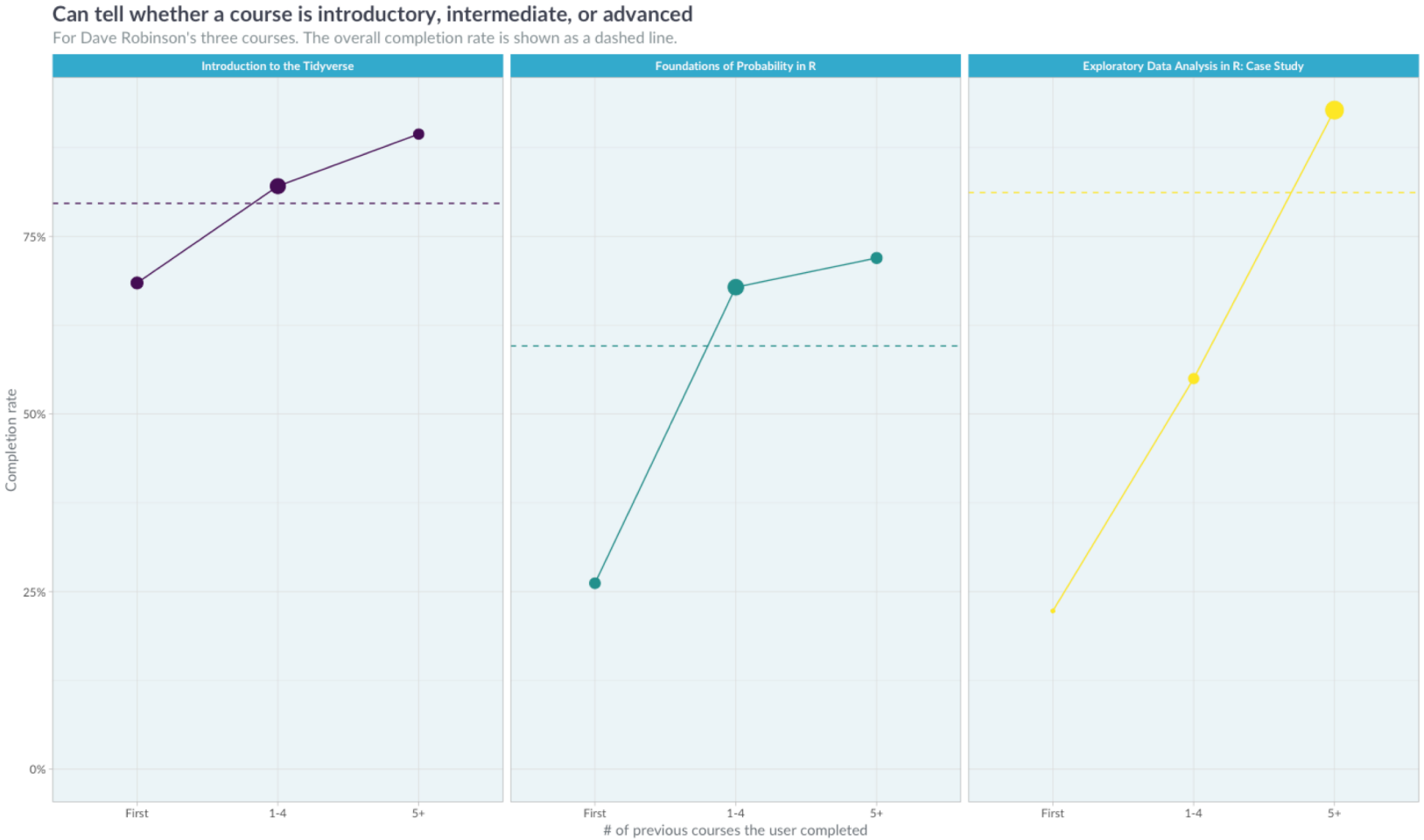
Course rating

% asking hint

% asking solution

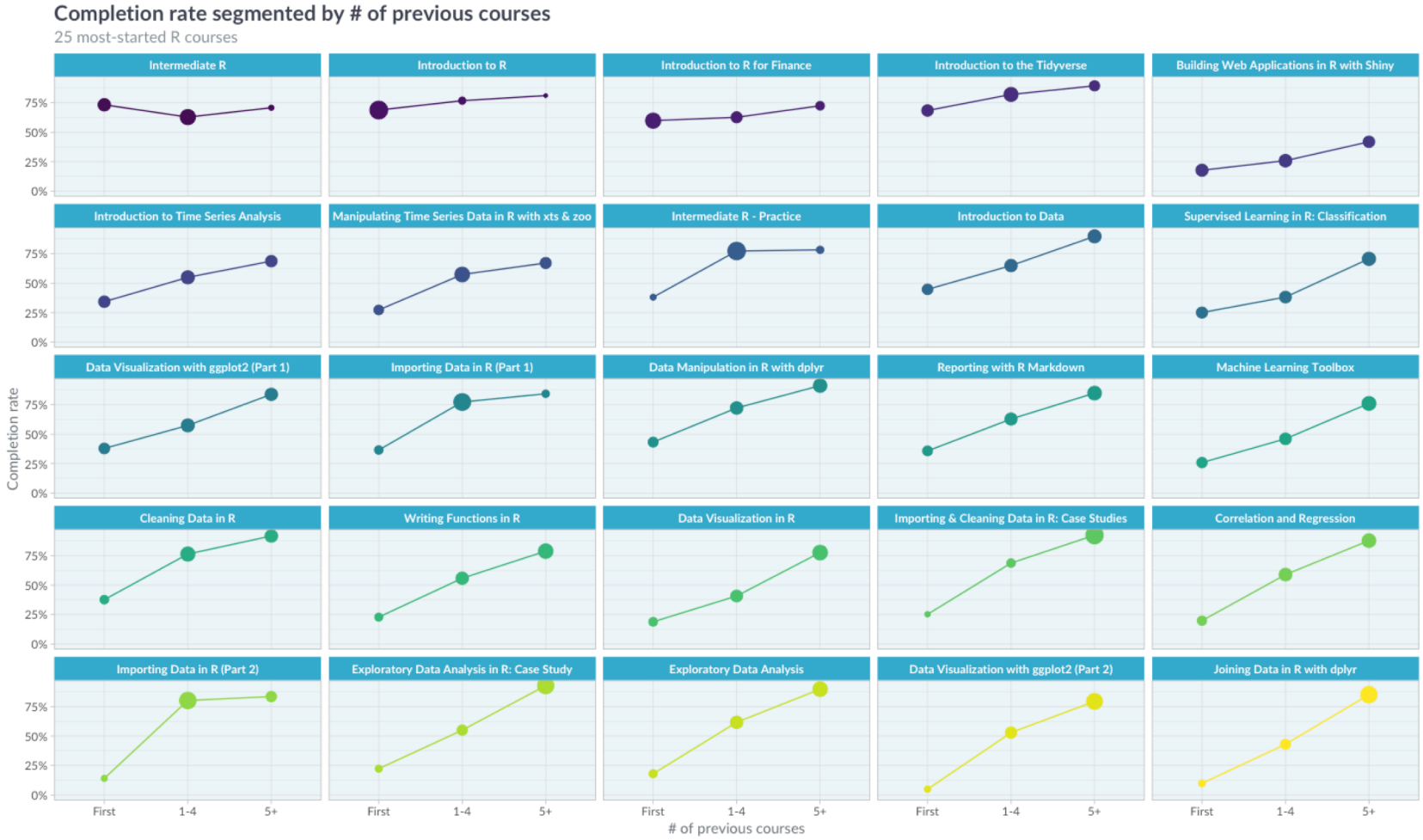
% reporting an issue

# Looking deeper into completion rate



# Segmenting users by experience shows the true completion rates

DataCamp



# What can we tell about the process of learning data science?

# We understand what in a course is challenging

DataCamp

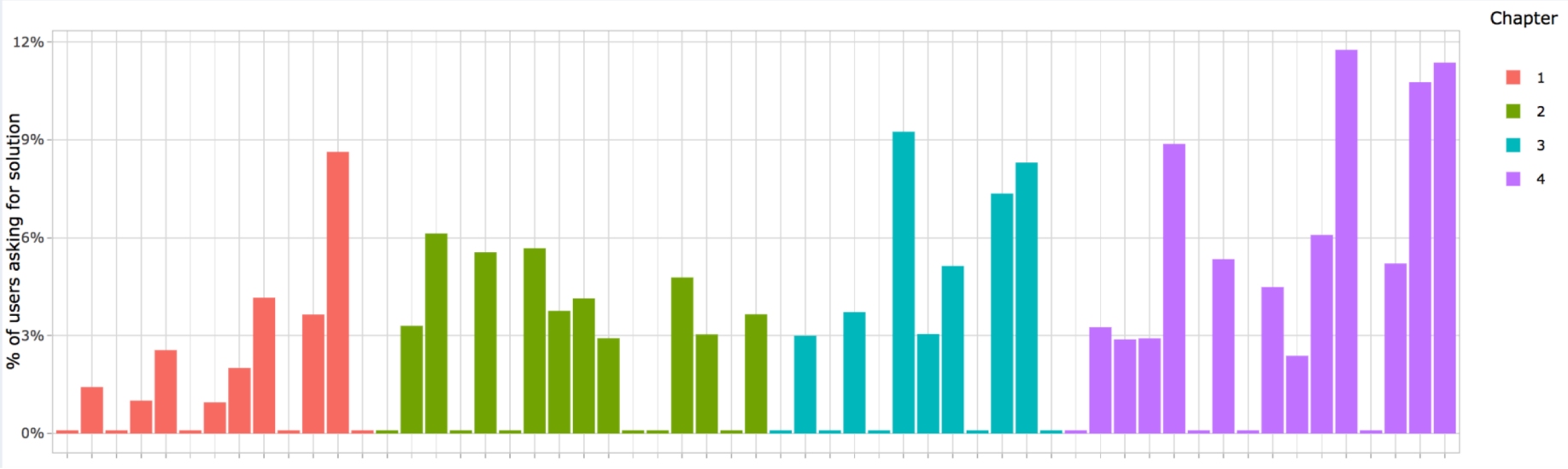
**Course:**

Intro to Python for Data Science

**Y-axis:**

% asked solution | % asked hint | Average rating | # of feedback messages | % of feedback messages

| course_id | course_title | technology | started | completed | avg_rating | nb_ratings | ex_used_hint | ex_used_solution | ex_avg_rating | ex_nb_ratings | completion |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 735 | Intro to Python for Data Science | Python | 7145 | 4571 | 4.72 | 4539 | 0.06 | 0.03 | 4.66 | 2744 | 0.64 |



# We understand what in a course is challenging

DataCamp

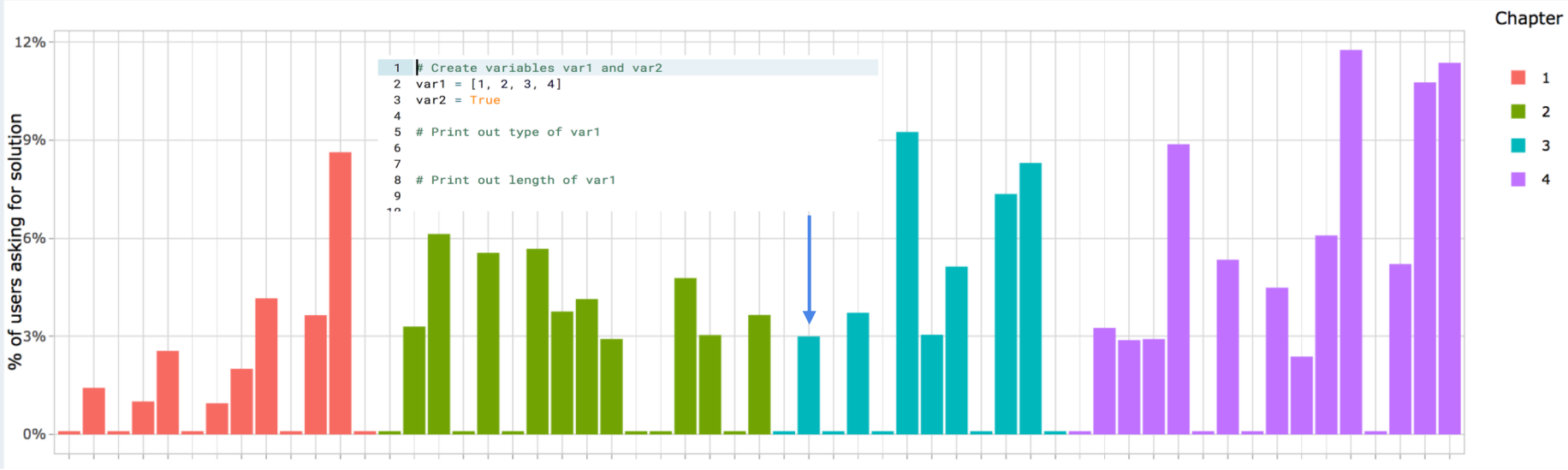
**Course:**

Intro to Python for Data Science

**Y-axis:**

% asked solution | % asked hint | Average rating | # of feedback messages | % of feedback messages

| course_id | course_title | technology | started | completed | avg_rating | nb_ratings | ex_used_hint | ex_used_solution | ex_avg_rating | ex_nb_ratings | completion |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 735 | Intro to Python for Data Science | Python | 7145 | 4571 | 4.72 | 4539 | 0.06 | 0.03 | 4.66 | 2744 | 0.64 |



# We understand what in a course is challenging

DataCamp

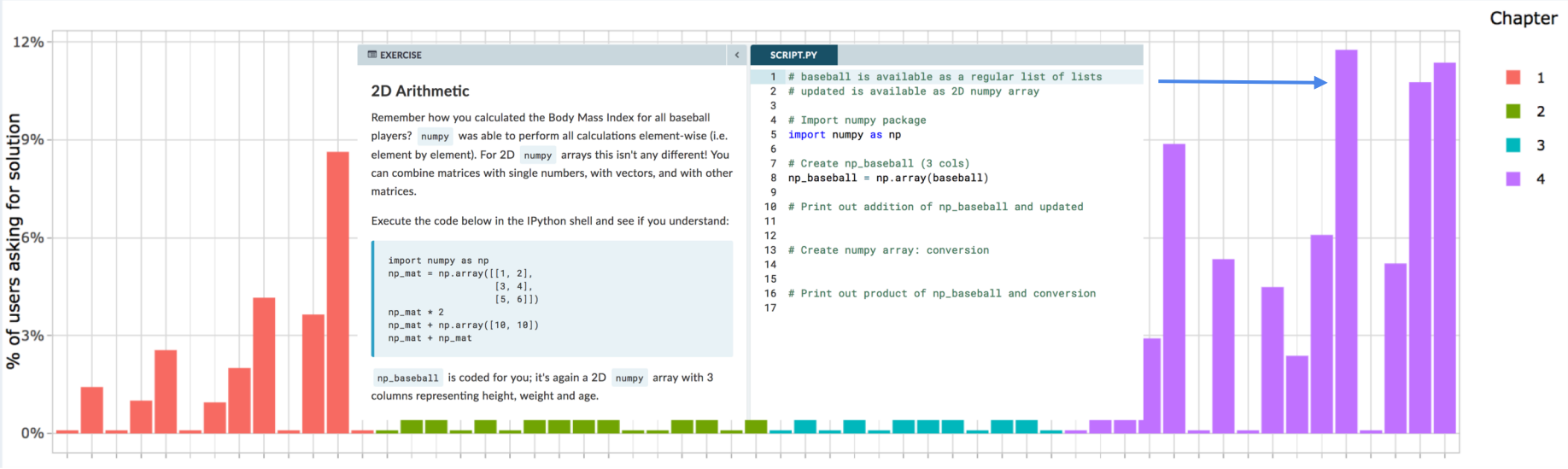
**Course:**

Intro to Python for Data Science

**Y-axis:**

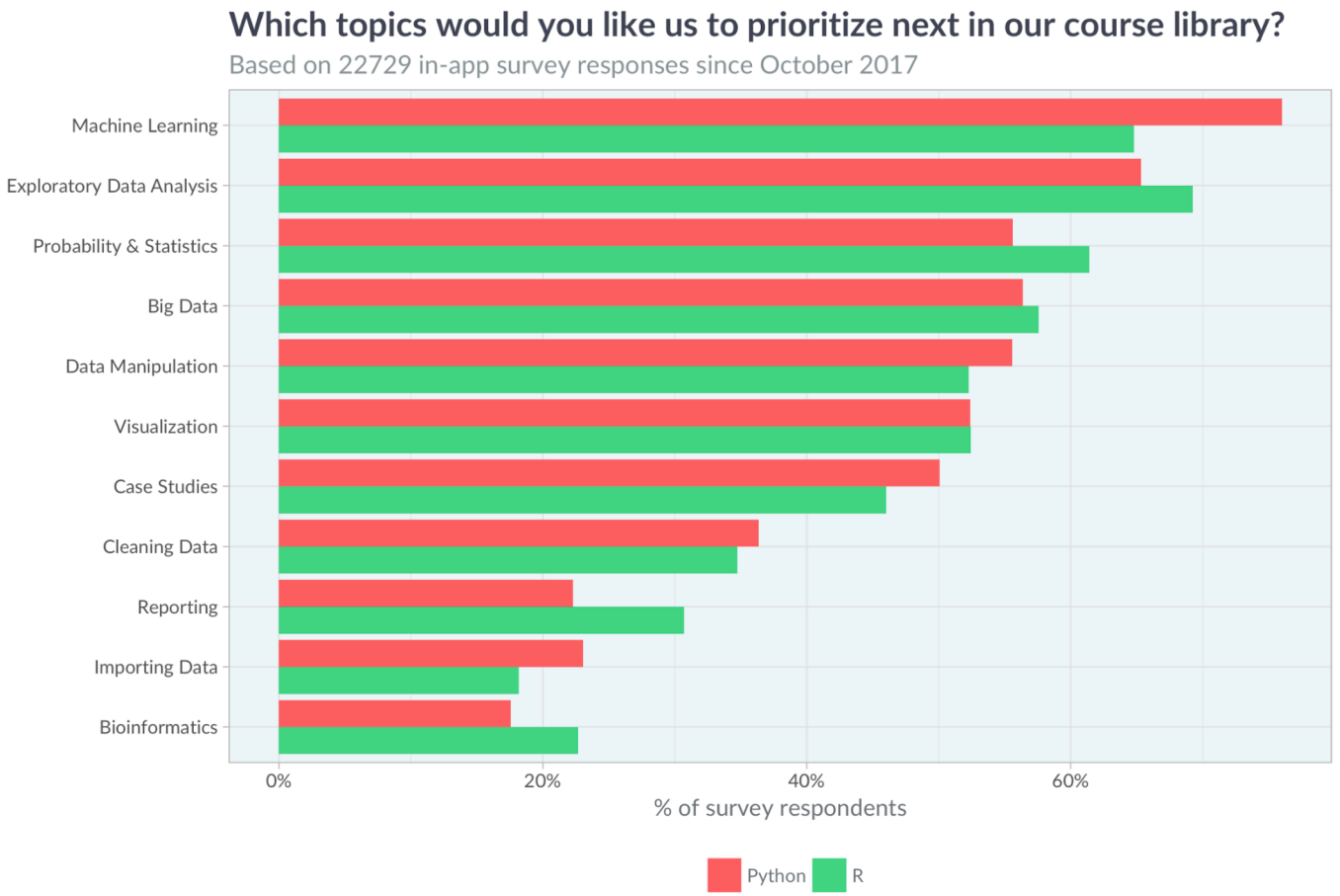
% asked solution | % asked hint | Average rating | # of feedback messages | % of feedback messages

| course_id | course_title | technology | started | completed | avg_rating | nb_ratings | ex_used_hint | ex_used_solution | ex_avg_rating | ex_nb_ratings | completion |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 735 | Intro to Python for Data Science | Python | 7145 | 4571 | 4.72 | 4539 | 0.06 | 0.03 | 4.66 | 2744 | 0.64 |



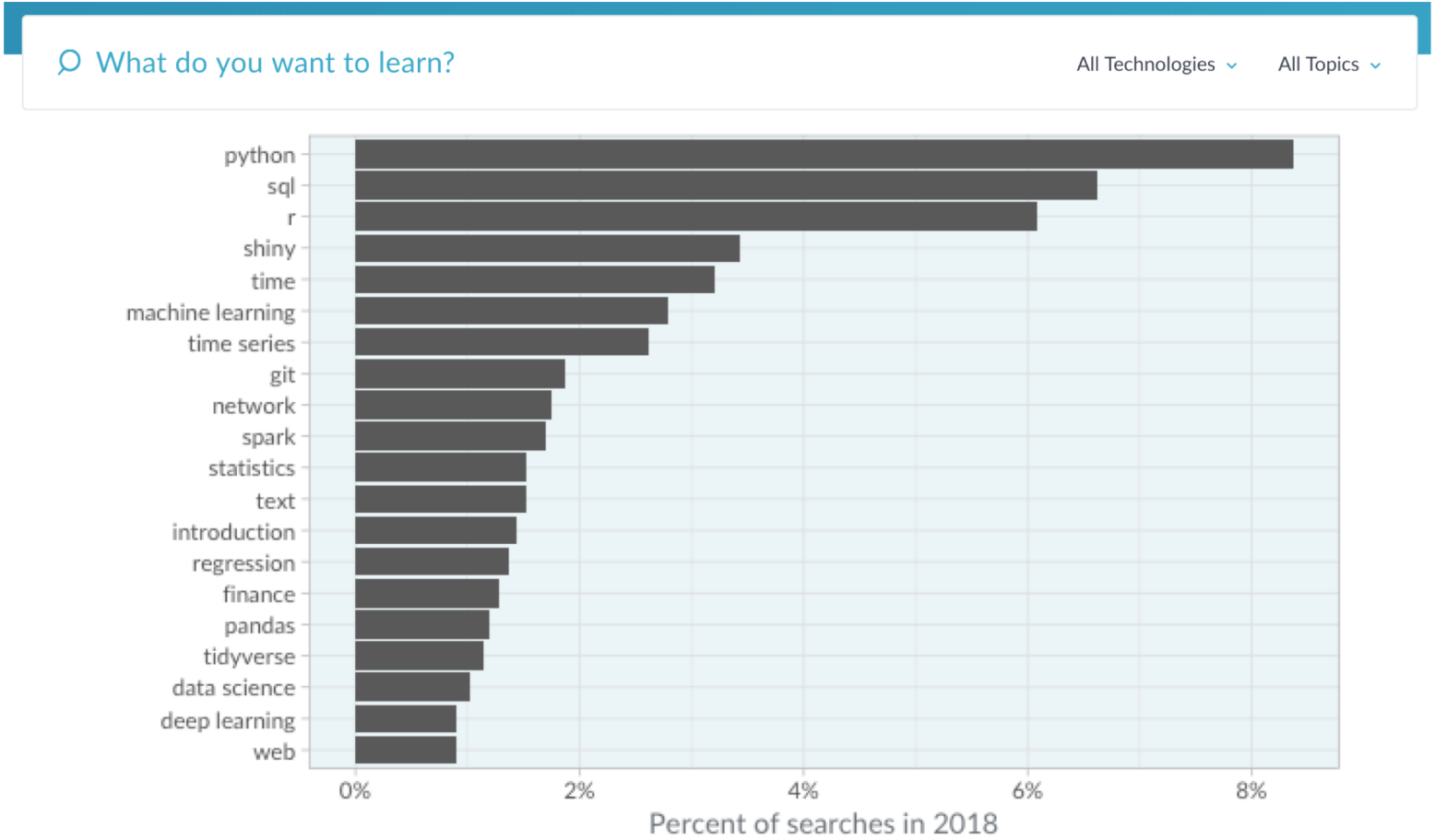
# Deciding what to teach

# DataCamp users want new courses in machine learning and EDA



# They search for time series, ML, networks, Spark, and text analysis

DataCamp



# Understand where data science is going

DataCamp

stackoverflow

## Side-by-side plots with ggplot2

231

115

I would like to place two plots side by side using the ggplot2 package, i.e. do the equivalent of `par(mfrow=c(1,2))` .

For example, I would like to have the following two plots show side-by-side with the same scale.

```
x <- rnorm(100)
eps <- rnorm(100,0,.2)
qplot(x,3*x+eps)
qplot(x,2*x+eps)
```

Do I need to put them in the same data.frame?

```
qplot(displ, hwy, data=mpg, facets = . ~ year) + geom_smooth()
```

r visualization ggplot2

share edit close flag protect

edited Dec 22 '16 at 5:41

jazzurro

15k ● 13 ● 38 ● 57

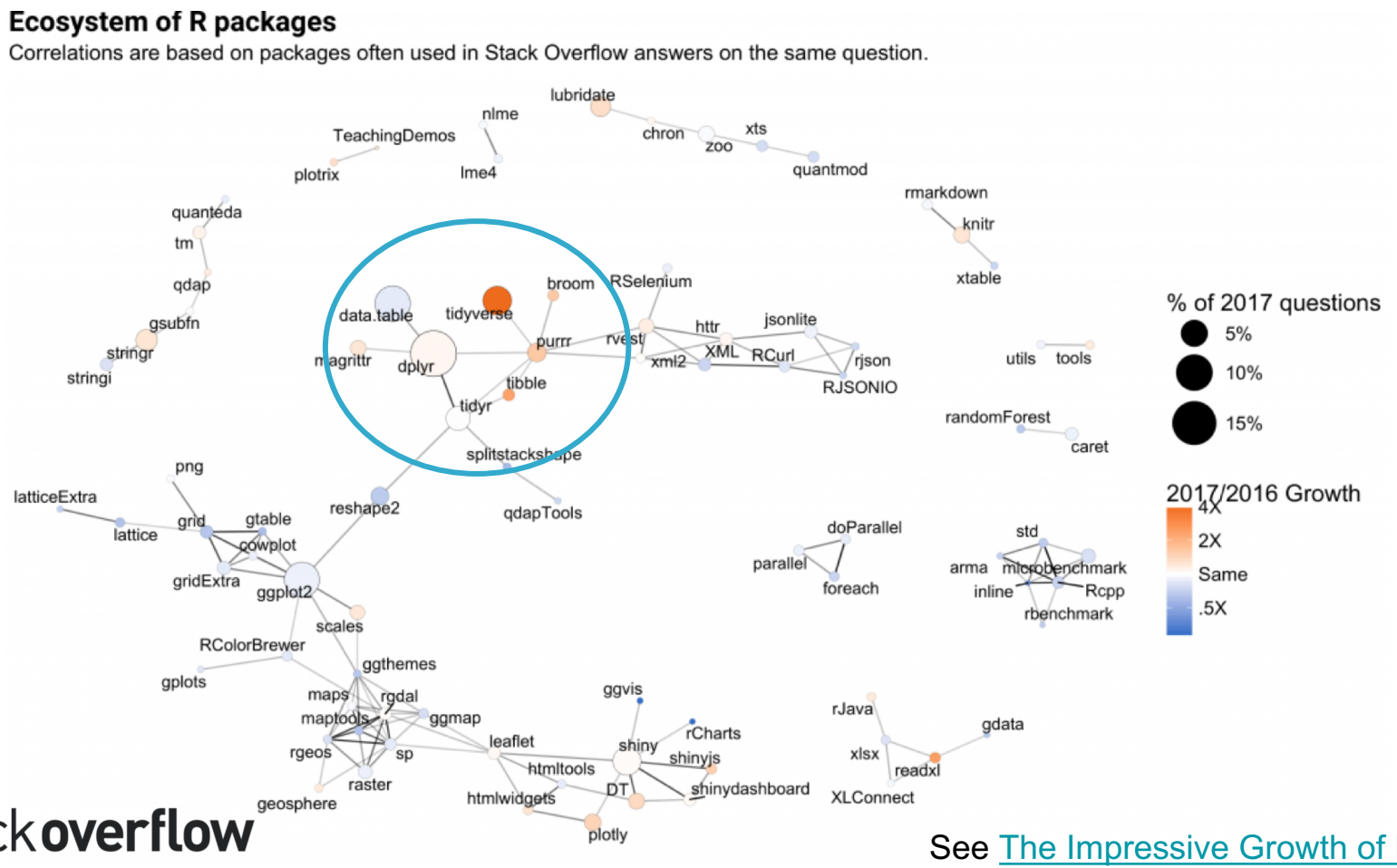
asked Aug 8 '09 at 18:16

Christopher DuBois

16.9k ● 20 ● 60 ● 90

# R package ecosystem: what is growing?

DataCamp




stackoverflow

See [The Impressive Growth of R]

# Development of a tidyverse curriculum

DataCamp

## 13 courses currently launched

### Sentiment Analysis in R: The Tidy Way

In this course, you will the learn principles of sentiment analysis from a tidy data perspective.

4 hours  Play preview

JULIA SILGE
Data Scientist at Stack Overflow

### Data Manipulation in R with dplyr

Master techniques for data manipulation using the select, mutate, filter, arrange, and summarise functions in dplyr.

4 hours

GARRETT GROLEMUND
Data Scientist at RStudio

### String Manipulation in R with stringr

Learn how to pull character strings apart, put them back together and use the stringr package.

4 hours  Play preview

CHARLOTTE WICKHAM
Assistant Professor at Oregon State University

### Communicating with Data in the Tidyverse

Leverage the power of tidyverse tools to create publication-quality graphics and custom-styled reports that communica...

4 hours

TIMO GROSSENBACHER
Data Journalist at SRF Data

## With many more to come! Including...

### Working with Data in the Tidyverse

Learn to work with data using tools from the tidyverse, and master the important skills of taming and tidying your data.

4 hours

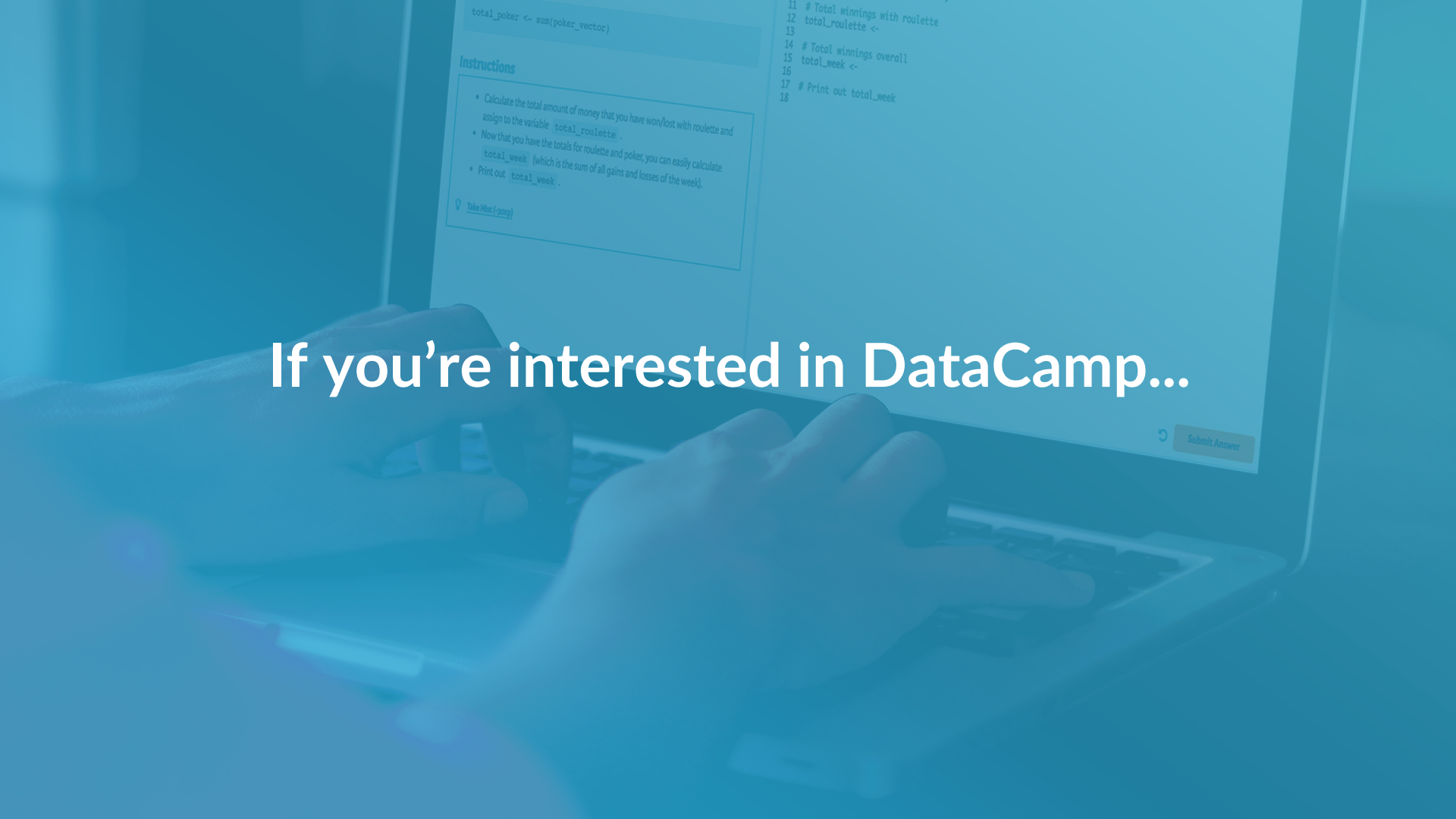
ALISON HILL
Professor and Data Scientist

### Modeling with Data in the Tidyverse

Explore Linear Regression in a tidy framework.

4 hours

ALBERT Y. KIM
Assistant Professor of Statistical & Data Sciences at Smith Co...




# If you're interested in DataCamp...

# Teach | www.datacamp.com/create

DataCamp

BAYESIAN MODELING WIT...

- 1. INTRODUCTION TO...
- 2. BAYESIAN MODELS ...
  - 1. The Normal-N...
  - 2. Normal-Norma...
  - 3. Sleep study data
  - 4. Insights from t...
  - 5. Simulating the ...
  - 6. Define, compil...
  - 7. Posterior insig...
  - 8. Markov chains
  - 9. Storing Markov...
  - 10. Markov chain...
  - 11. Markov chain...
  - 12. Markov chain...
  - 13. Multiple chains
  - 14. Naive standar...
  - 15. Reproducibility
  - Add Exercise
- 3. BAYESIAN INFEREN...
- 4. MULTIVARIATE & G...

Add Chapter

## Sleep study data

NORMAL EXERCISE | Reported Issues | Delete Exercise

### CONTEXT

Researchers enrolled 18 subjects in a sleep deprivation study. Their observed `sleep_study` data are loaded in the workspace. These data contain the `day_0` reaction times and `day_3` reaction times after 3 sleep deprived nights for each `subject`.

You will define and explore `diff_3`, the observed *difference* in reaction times for each subject. This will require the `mutate()` & `summarize()` functions. For example, the following would add variable `day_0_s`, `day_0` reaction times in *seconds*, to `sleep_study`:

```{r}
sleep_study <- sleep_study %>%
  mutate(day_0_s = day_0 * 0.001)
```

You can then `summarize()` the `day_0_s` values, here by their minimum & maximum:

```{r}
sleep_study  %>%
  summarize(min(day_0_s), max(day_0_s))
```

### INSTRUCTIONS

- Check out the first 6 rows of `sleep_study`.
- Define a new `sleep_study` variable `diff_3`, the `day_3` less the `day_0` reaction times.
- Use `ggplot()` with a `geom_histogram()` layer to construct a histogram of the `diff_3` data.
- `summarize()` the mean and standard deviation of the `diff_3` observations.

### HINT

- The first 6 rows are the `head()` of a data frame.
- Define `diff_3 = day_3 - day_0` in `mutate()`.

Back to Dashboard | BAYESIAN MODELING WITH RJAGS | MASTER | UNSAVED CHANGES | Save | Preview

DataCamp | Course Outline

### EXERCISE

**Sleep study data**

Researchers enrolled 18 subjects in a sleep deprivation study. Their observed `sleep_study` data are loaded in the workspace. These data contain the `day_0` reaction times and `day_3` reaction times after 3 sleep deprived nights for each `subject`.

You will define and explore `diff_3`, the observed *difference* in reaction times for each subject. This will require the `mutate()` & `summarize()` functions. For example, the following would add variable `day_0_s`, `day_0` reaction times in *seconds*, to `sleep_study`:

```
sleep_study <- sleep_study %>%
    mutate(day_0_s = day_0 * 0.001)
```

You can then `summarize()` the `day_0_s` values, here by their minimum & maximum:

```
sleep_study  %>%
    summarize(min(day_0_s), max(day_0_s
```

### INSTRUCTIONS 100 XP

- Check out the first 6 rows of `sleep_study`.
- Define a new `sleep_study` variable `diff_3`, the `day_3` less the `day_0` reaction times.
- Use `ggplot()` with a `geom_histogram()` layer to construct a histogram of the `diff_3` data.
- `summarize()` the mean and standard deviation of the `diff_3` observations.

Take Hint (-30 XP)

### SCRIPT.R

```
1  # Check out the first 6 rows of sleep_study
2
3
4  # Define diff_3
5
6
7
8  # Histogram of diff_3
9  ggplot(___, aes(x = ___)) +
10     geom_histogram(binwidth = 20, color = "white")
11
12 # Mean and standard deviation of diff_3
```

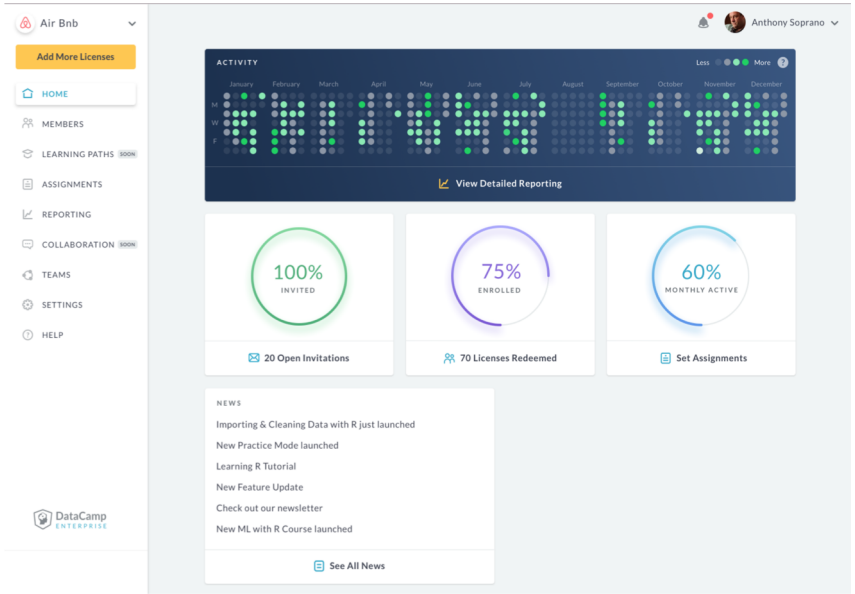
Run Code | Submit Answer

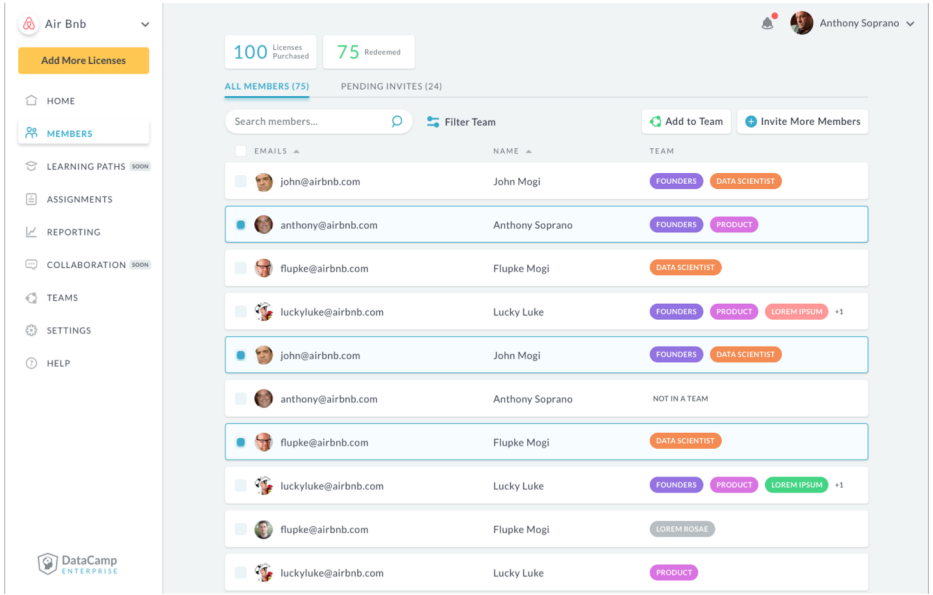
### R CONSOLE

>

# DataCamp for Business | www.datacamp.com/groups/business

DataCamp



# DataCamp for the Classroom | www.datacamp.com/groups/education

DataCamp

# Come work with us | www.datacamp.com/careers

DataCamp

Where would you like to work?

(•) All ( ) Belgium ( ) London ( ) New York City

Content

**Content Development Lead**

Leuven, Belgium — Apply

**Instructor Recruiter, Data Science**

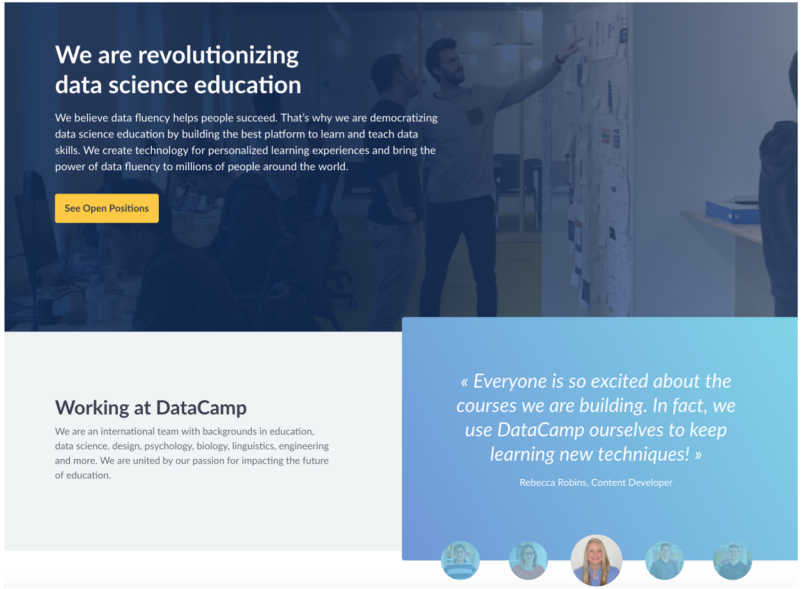
New York City — Apply

**Product Manager, Data Science Curriculum**

New York City — Apply

**Product Manager, Mobile Content**

Leuven, Belgium — Apply



# Thank you



-  @old_man_chester
-  @ismayc
-  chester@datacamp.com


Special thanks to Dave Robinson

- PDF of slides available at

  http://bit.ly/ismay-jsm