



DataCamp

Using Data to Drive Curriculum Development

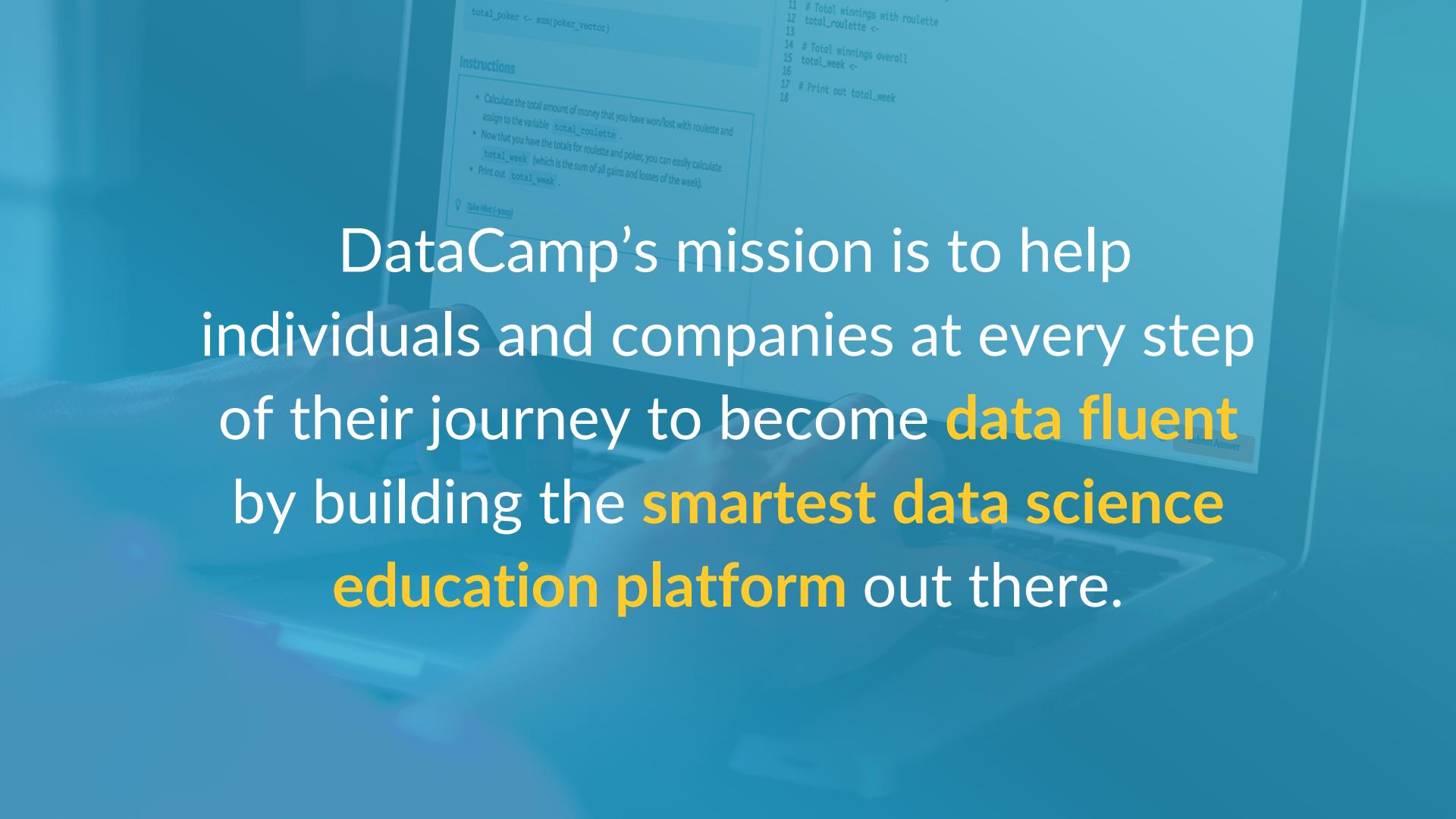
Chester Ismay
Senior Curriculum Lead

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.

```
total_poker <- sum(poker_vector)
11 # Total winnings with roulette
12 total_roulette <-
13
14 # Total winnings overall
15 total_week <-
16
17 # Print out total_week
18
```



pythonTM

SQL

SPARK

git

Shell

1



Learn

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

2



Practice

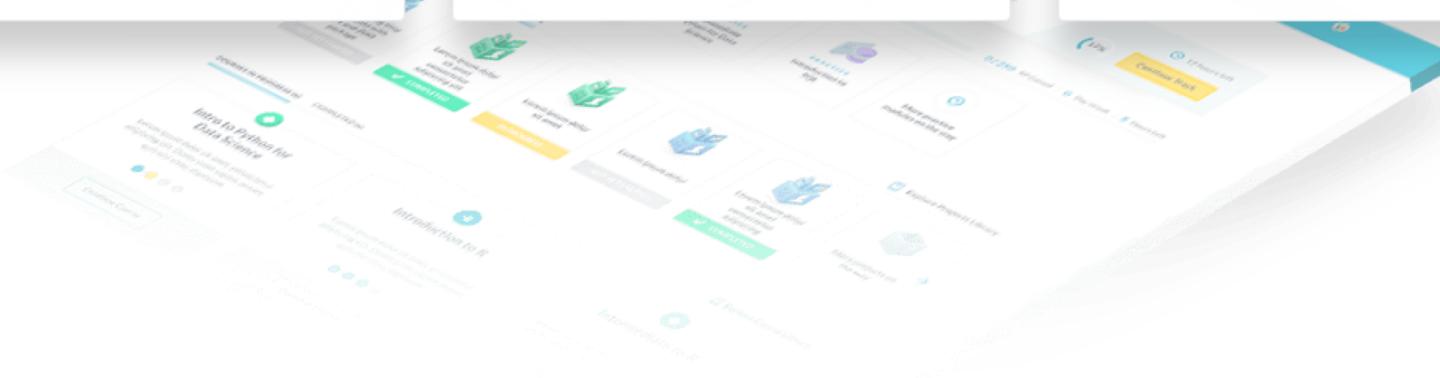
Sharpen and train your
newly learned skills. Take
bite-sized, fun Practice
Challenges.

3



Build

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





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)

Course Outline

SCRIPT.R

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



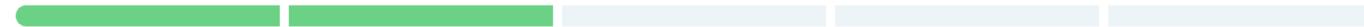
Run Code

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



PROJECT INSTRUCTIONS

1

Task 1: Instructions

2

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

3

> Read

```
datasets/breath_alcohol_ames.csv
```

4

into your workspace using the

```
read_csv()
```

 function. Save it as
`ba_data`.

5

> Count how many tests were administered in each `year` using the `count()` function, creating a new data set called `ba_year`.

6

7

Good to know

8

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](#)[How it works](#) [Report an Issue](#)

BETA

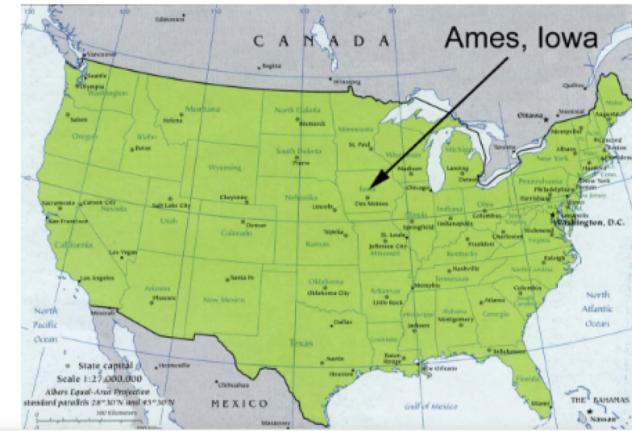
PROJECT: WHO IS DRUNK AND WHEN IN AMES, IOWA?

 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?

INSTRUCTIONS 100XP

% asked hint

Change the code in the editor to `summarize()` within each country rather than within each year. Save the result as `by_country`.

 Take Hint (-30xp)

INSTRUCTIONS 70XP

Change the code in the editor to `summarize()` within each country rather than within each year. Save the result as `by_country`.

 `</>` Show Answer (-70xp)

HINT

Change `group_by(year)` to `group_by(country)` and don't forget to assign it to the right variable.

% asked solution

% reporting an issue

Report Issue 

Am I learning?

Is it fun?

Is it doable?

Is it working?

COMING SOON

Course rating

% asking hint

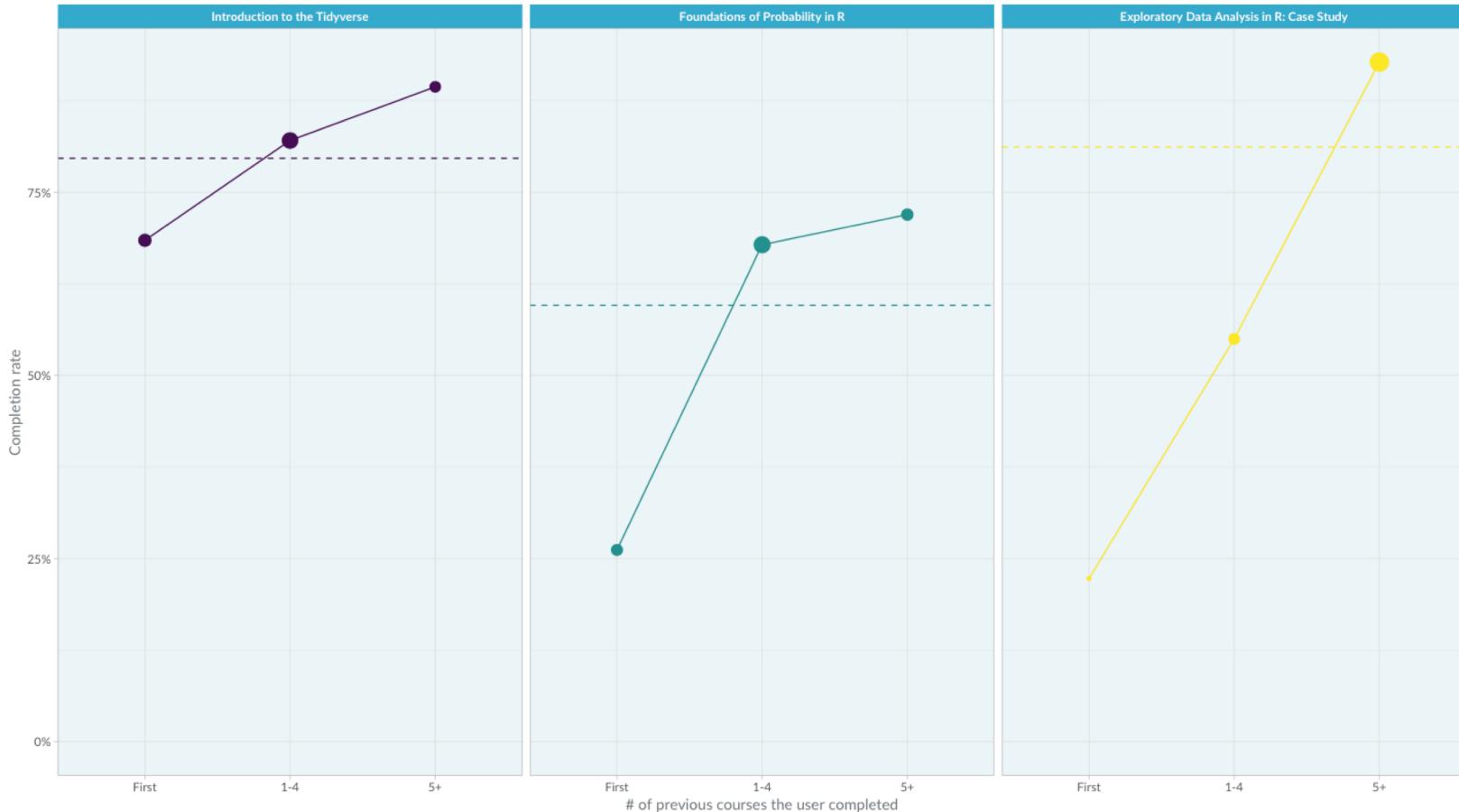
% asking solution

% reporting an issue

Looking deeper into completion rate

Can tell whether a course is introductory, intermediate, or advanced

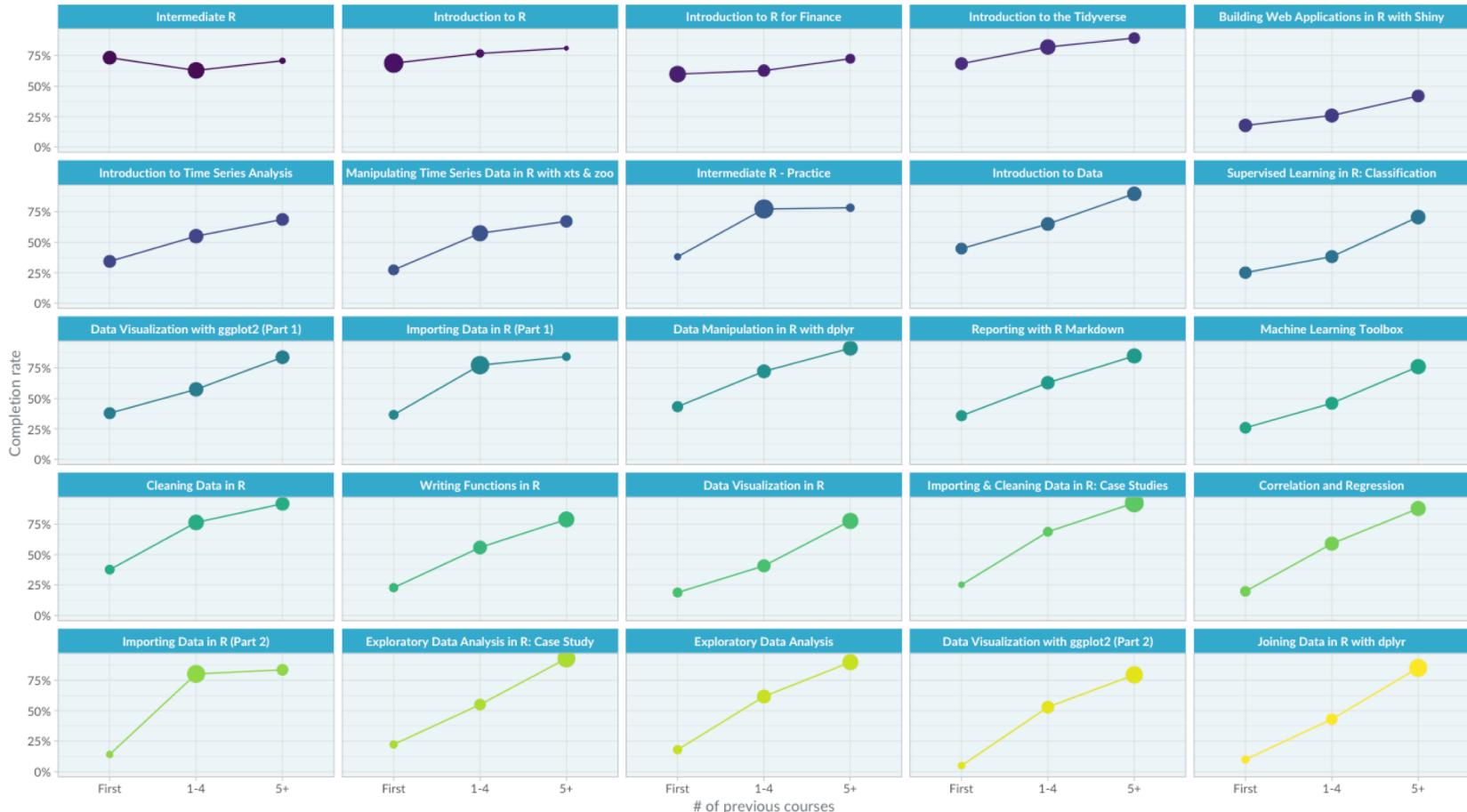
For Dave Robinson's three courses. The overall completion rate is shown as a dashed line.



Segmenting users by experience shows the true completion rates

Completion rate segmented by # of previous courses

25 most-started R courses



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

We understand what in a course is challenging

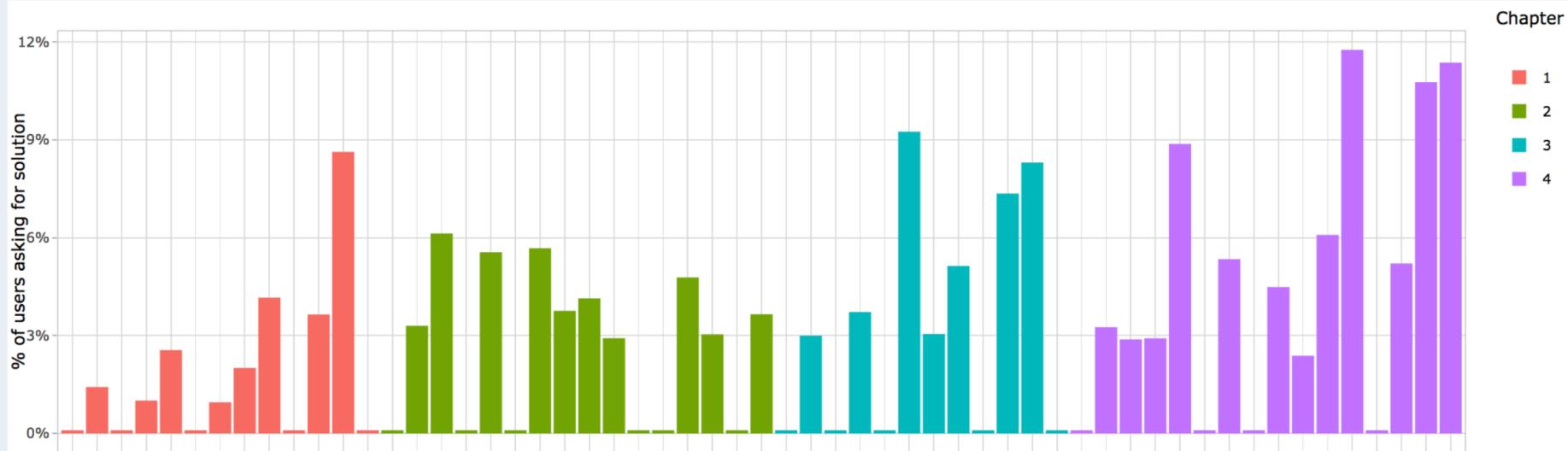
Course:

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

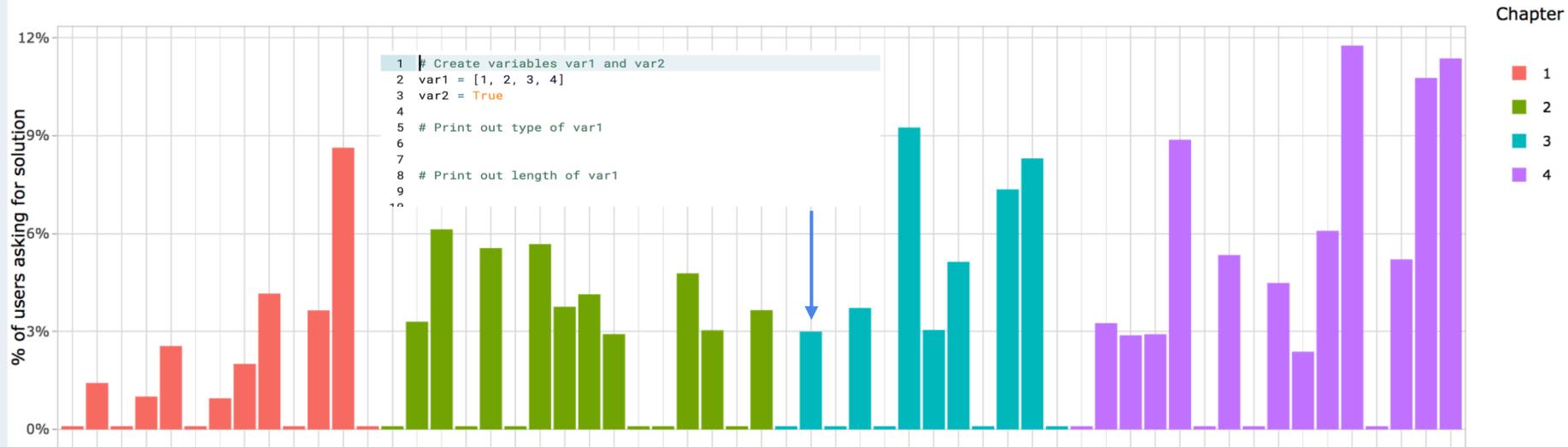
Course:

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

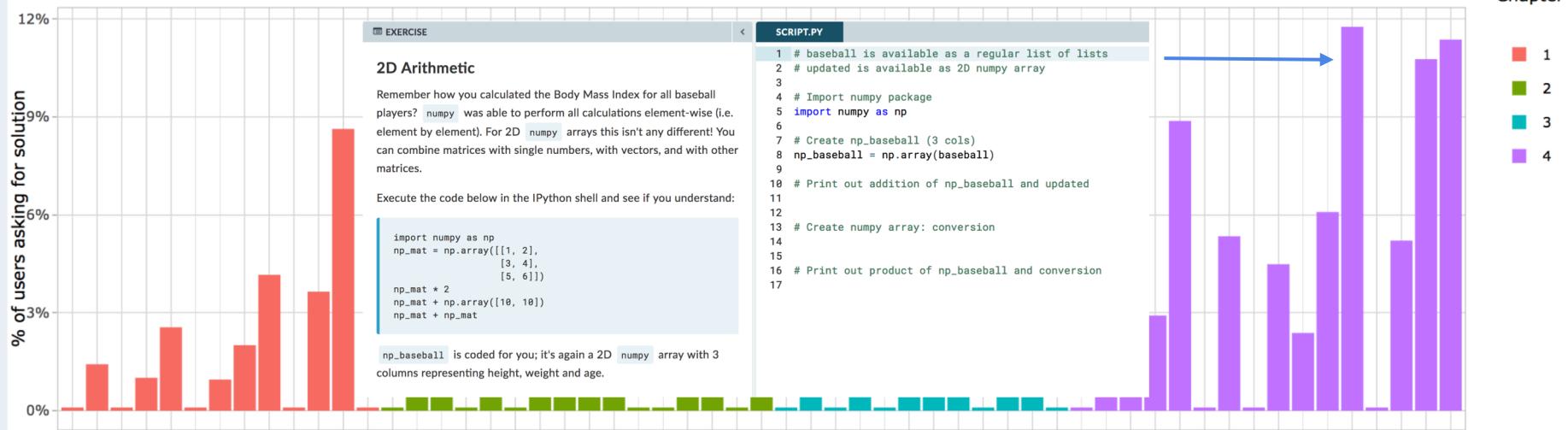
Course:

Intro to Python for Data Science

Y-axis:

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

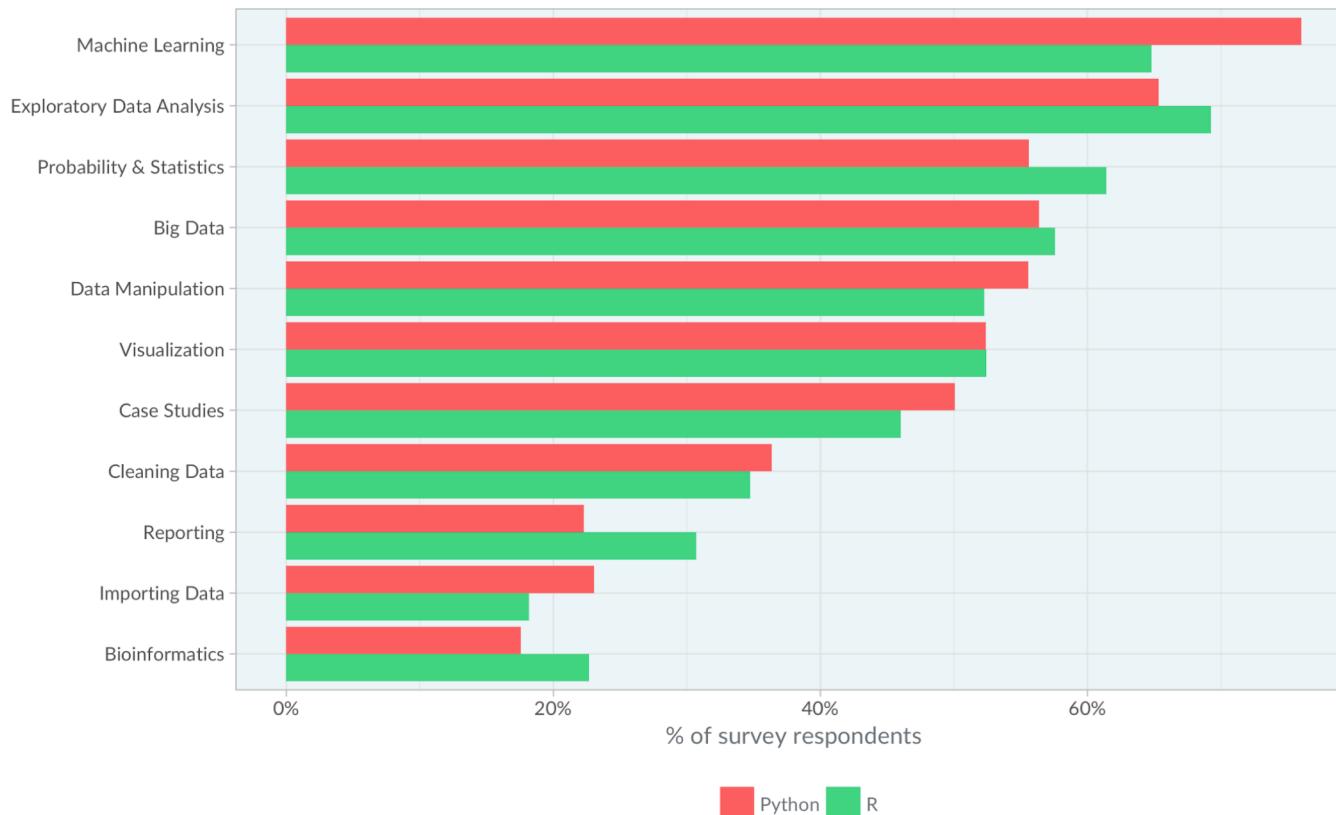
course_id	course_title	technology	started	completed	avg_rating	nb_ratings	ex_used_hint	ex_used_solution	ex_avg_rating	ex_nb_ratings	completion
735	Intro to Python for Data Science	Python	7145	4571	4.72	4539	0.06	0.03	4.66	2744	0.64



Deciding **what** to teach

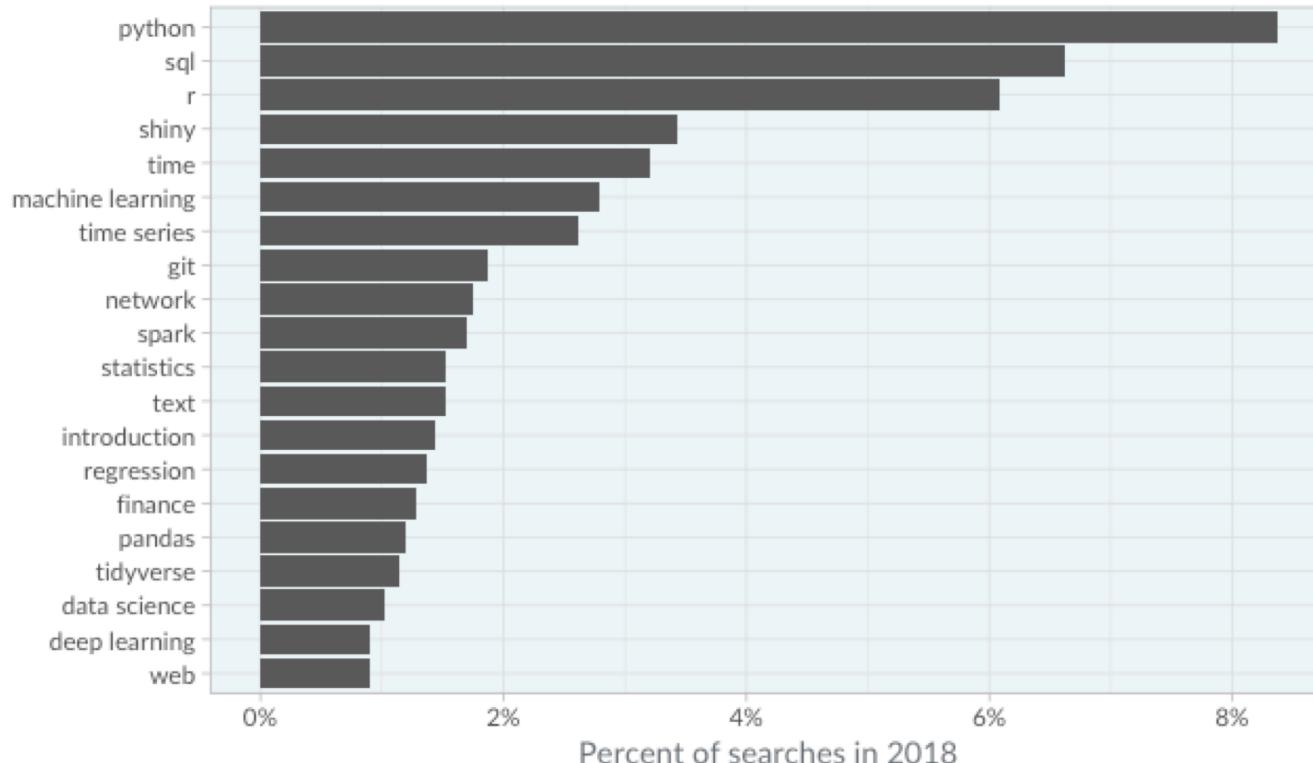
Which topics would you like us to prioritize next in our course library?

Based on 22729 in-app survey responses since October 2017



What do you want to learn?

All Technologies ▾ All Topics ▾





Side-by-side plots with ggplot2



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))`.

231

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)
```

115

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



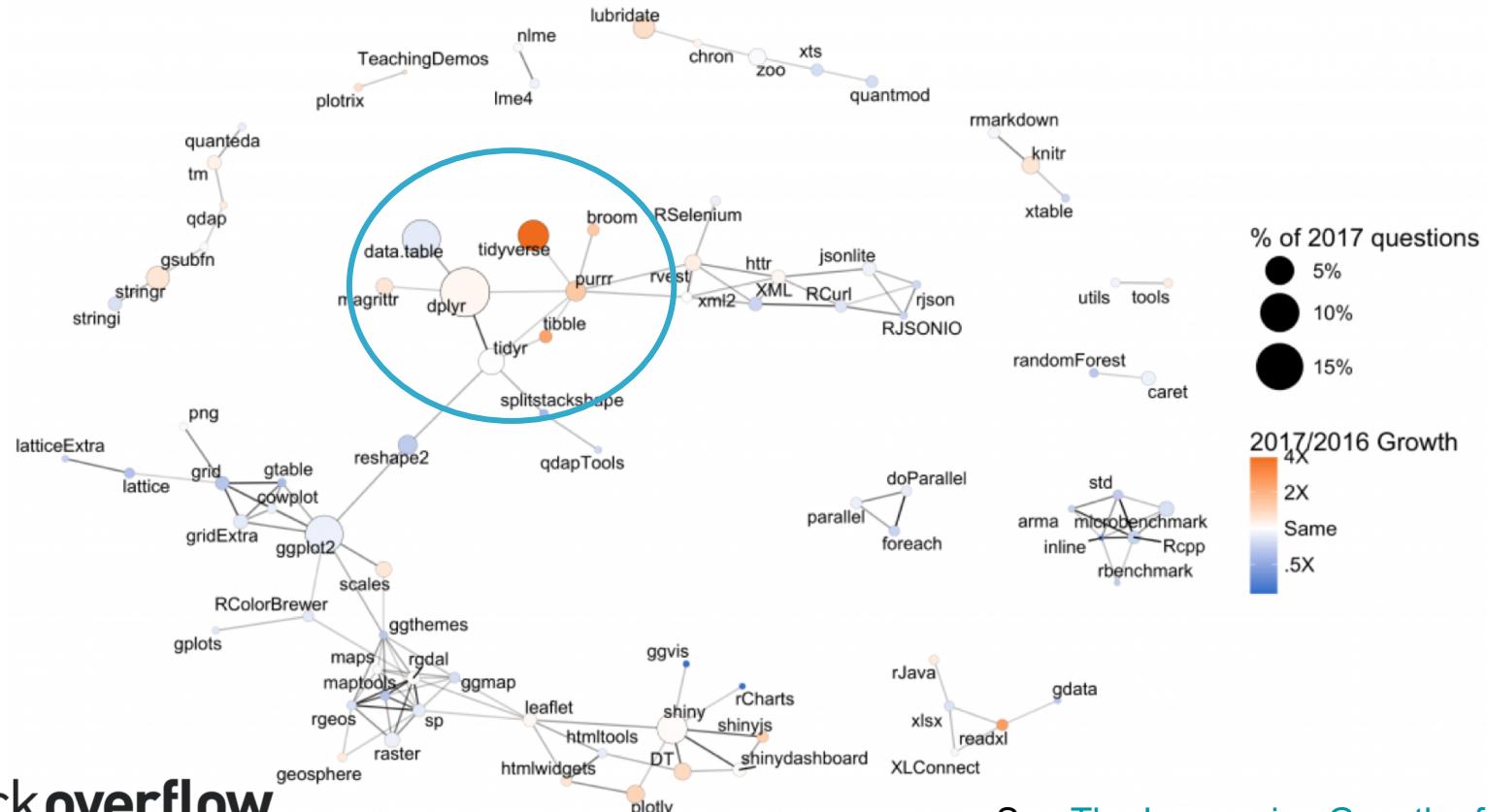
Christopher DuBois

16.9k 20 60 90

R package ecosystem: what is growing?

Ecosystem of R packages

Correlations are based on packages often used in Stack Overflow answers on the same question.



See The Impressive Growth of R

Development of a tidyverse curriculum

13 courses currently launched



Sentiment Analysis in R: The Tidy Way

In this course, you will 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 communicate...

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

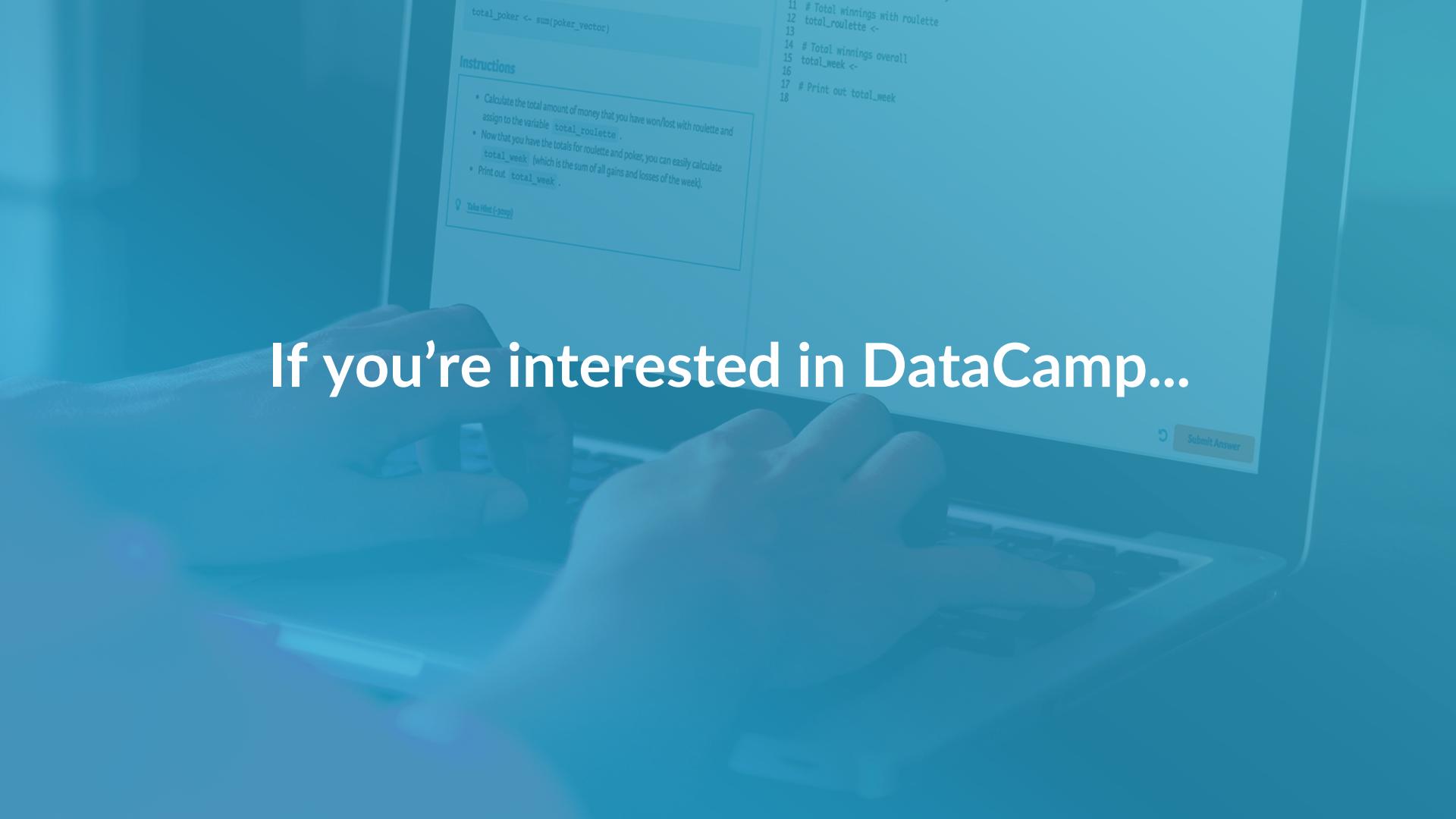
Explore Linear Regression in a tidy framework.

4 hours



ALBERT Y. KIM

Assistant Professor of Statistical & Data Sciences at Smith Co...



If you're interested in DataCamp...

```
total_poker <- sum(poker_vector)

Instructions
  • Calculate the total amount of money that you have won/lost with roulette and
    assign to the variable total_roulette.
  • Now that you have the totals for roulette and poker, you can easily calculate
    total_week (which is the sum of all gains and losses of the week).
  • Print out total_week.
```

Take Hint (-20%)

```
11 # Total winnings with roulette
12 total_roulette <-
13
14 # Total winnings overall
15 total_week <-
16
17 # Print out total_week
18
```

BAYESIAN MODELING WITH RJAGS

- > 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()'.
- 'summarize()' the 'mean(diff_3)' and 'sd(diff_3)'.

DataCamp

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.

R CONSOLE

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

> |

Air Bnb

Add More Licenses

HOME

MEMBERS

LEARNING PATHS 900K

ASSIGNMENTS

REPORTING

COLLABORATION 900K

TEAMS

SETTINGS

HELP

Activity

January February March April May June July August September October November December

Anthony Soprano

View Detailed Reporting

100% INVITED

75% ENROLLED

60% MONTHLY ACTIVE

20 Open Invitations

70 Licenses Redeemed

Set Assignments

NEWS

Importing & Cleaning Data with R just launched

New Practice Mode launched

Learning R Tutorial

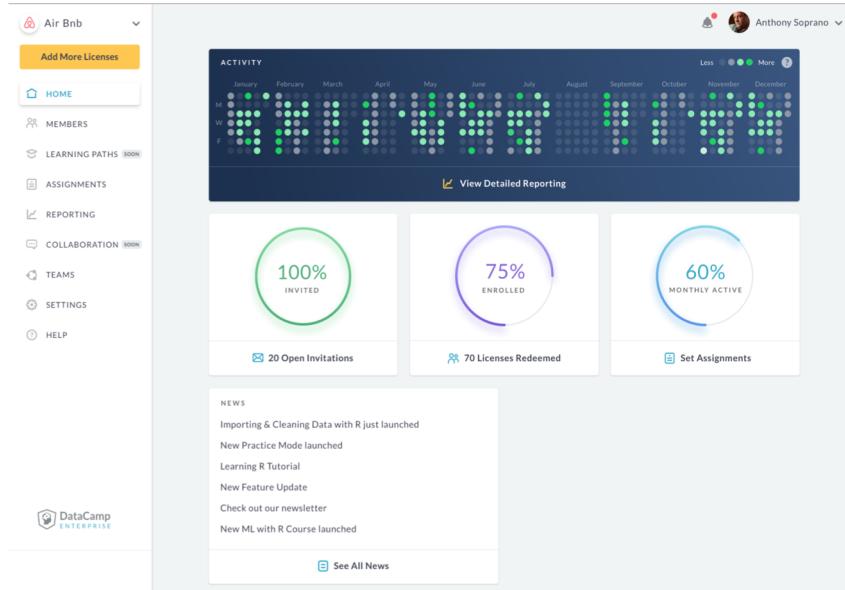
New Feature Update

Check out our newsletter

New ML with R Course launched

See All News

DataCamp ENTERPRISE



Air Bnb

Add More Licenses

HOME

MEMBERS

LEARNING PATHS 900K

ASSIGNMENTS

REPORTING

COLLABORATION 900K

TEAMS

SETTINGS

HELP

100 Licenses Purchased

75 Redeemed

ALL MEMBERS (75)

PENDING INVITES (24)

Search members...

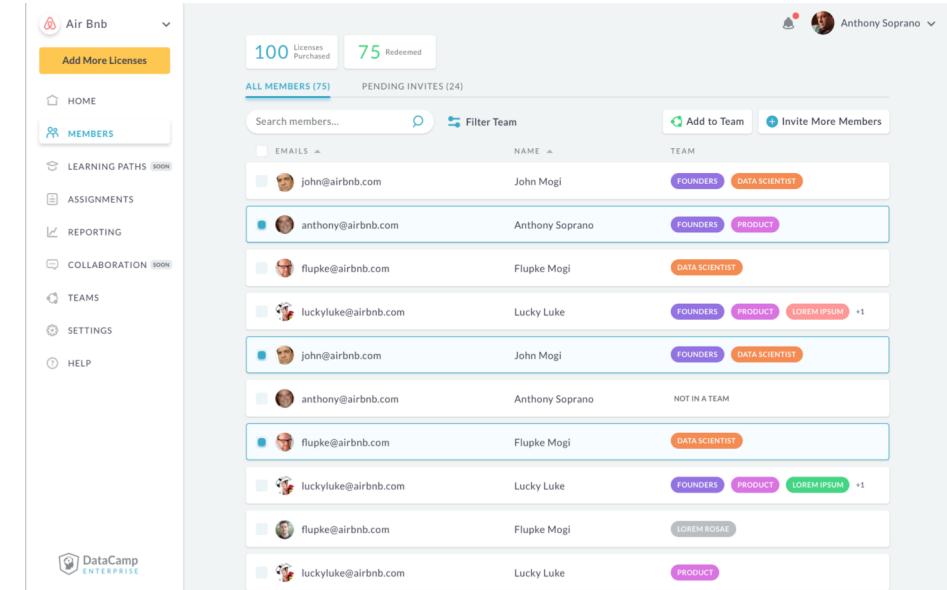
Filter Team

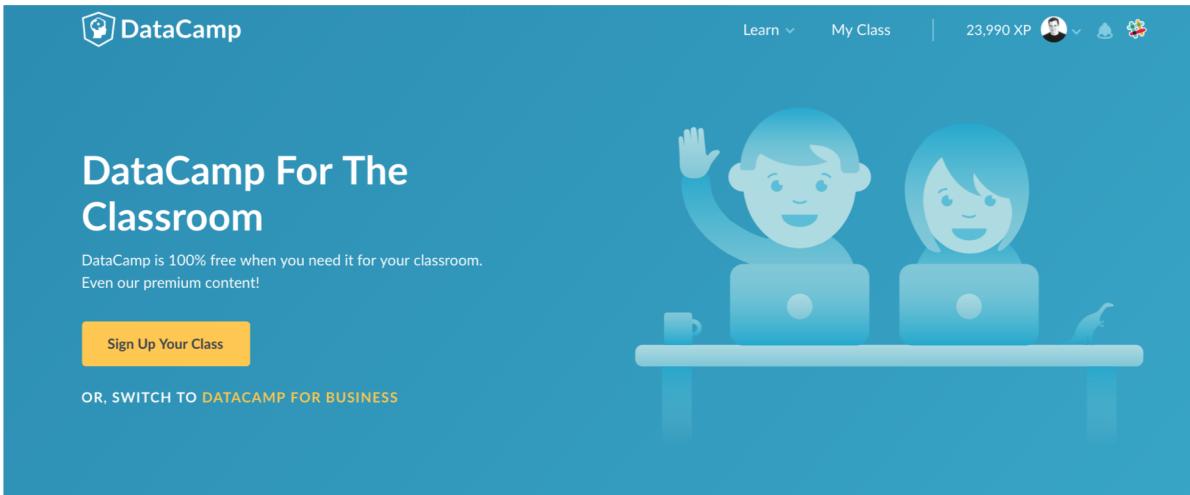
Add to Team

Invite More Members

| NAME | TEAM | FOUNDERS | DATA SCIENTIST |
|-----------------|-----------------|----------------|----------------|
| John Mogi | John Mogi | FOUNDERS | DATA SCIENTIST |
| Anthony Soprano | Anthony Soprano | FOUNDERS | PRODUCT |
| Flupke Mogi | Flupke Mogi | DATA SCIENTIST | |
| Lucky Luke | Lucky Luke | FOUNDERS | PRODUCT |
| John Mogi | John Mogi | FOUNDERS | DATA SCIENTIST |
| Anthony Soprano | NOT IN A TEAM | | |
| Flupke Mogi | Flupke Mogi | DATA SCIENTIST | |
| Lucky Luke | Lucky Luke | FOUNDERS | PRODUCT |
| Flupke Mogi | Flupke Mogi | LOREM IPSUM | +1 |
| Lucky Luke | Lucky Luke | FOUNDERS | PRODUCT |
| Lucky Luke | Lucky Luke | LOREM IPSUM | +1 |

DataCamp ENTERPRISE





The screenshot shows the DataCamp website's 'For The Classroom' section. At the top, there's a navigation bar with the DataCamp logo, 'Learn', 'My Class', and user stats (23,990 XP). Below the header, the main title 'DataCamp For The Classroom' is displayed in large white font. A subtext below it says 'DataCamp is 100% free when you need it for your classroom. Even our premium content!' followed by a yellow 'Sign Up Your Class' button. To the right, there's a cartoon illustration of two students at a desk with laptops, one raising their hand. Below the main title, there's a link 'OR, SWITCH TO DATACAMP FOR BUSINESS'.

Students Learning with DataCamp

Instructors across the world use DataCamp to teach large groups of students, develop their own bootcamps, help researchers strengthen their data science skills, and much more.



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

We are revolutionizing data science education

We believe data fluency helps people succeed. That's why we are democratizing data science education by building the best platform to learn and teach data skills. We create technology for personalized learning experiences and bring the power of data fluency to millions of people around the world.

See Open Positions



Working at DataCamp

We are an international team with backgrounds in education, data science, design, psychology, biology, linguistics, engineering and more. We are united by our passion for impacting the future of education.

« Everyone is so excited about the courses we are building. In fact, we use DataCamp ourselves to keep learning new techniques! »

Rebecca Robins, Content Developer





-  [@old_man_chester](https://twitter.com/old_man_chester)
-  [@ismayc](https://github.com/ismayc)
-  chester@datacamp.com

Special thanks to Dave Robinson

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