

COURSE OVERVIEW

(Data Science Tools and Methods)

MARCUS BIRKENKRAHE

Created: 2021-08-19 Do 10:50

TABLE OF CONTENTS

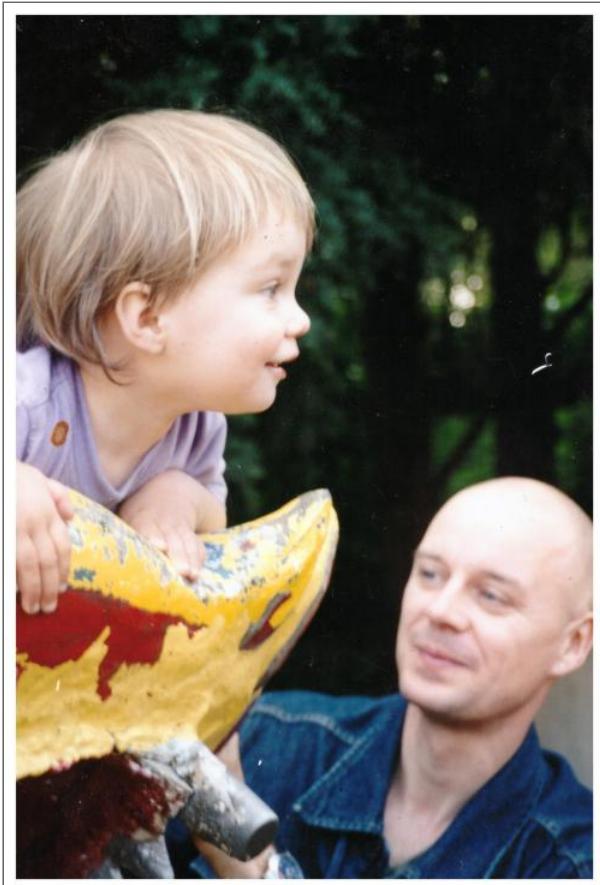
- What're you going to learn today?
- Who am I?
 - Science
 - Industry
 - Teaching
 - Pleasure
- What are your expectations?
- Which topics will we cover?
 - Introduction to data science
 - Introduction to R programming
 - Visualization using R
 - FasteR approach
 - Schedule (see Syllabus)
- How will we do it?
 - Classroom sessions
 - Lecture scripts with exercises (GitHub)
 - Reading suggestions
 - Video lectures (YouTube)
 - Online assignments (DataCamp)
 - Team EDA project
 - Agile project management
 - Tests and final exam
 - Podcasts and feeds
 - Summary of course activities
- What do you have to do to pass?

- DataCamp assignments (> 50%)
- Team project (> 50%)
 - What is a team project?
 - Do you have project examples?
 - Can I do a project as an absolute beginner?
- Final exam (> 50%)
- What's next?
 - In the course
 - Your challenges
- Any questions?

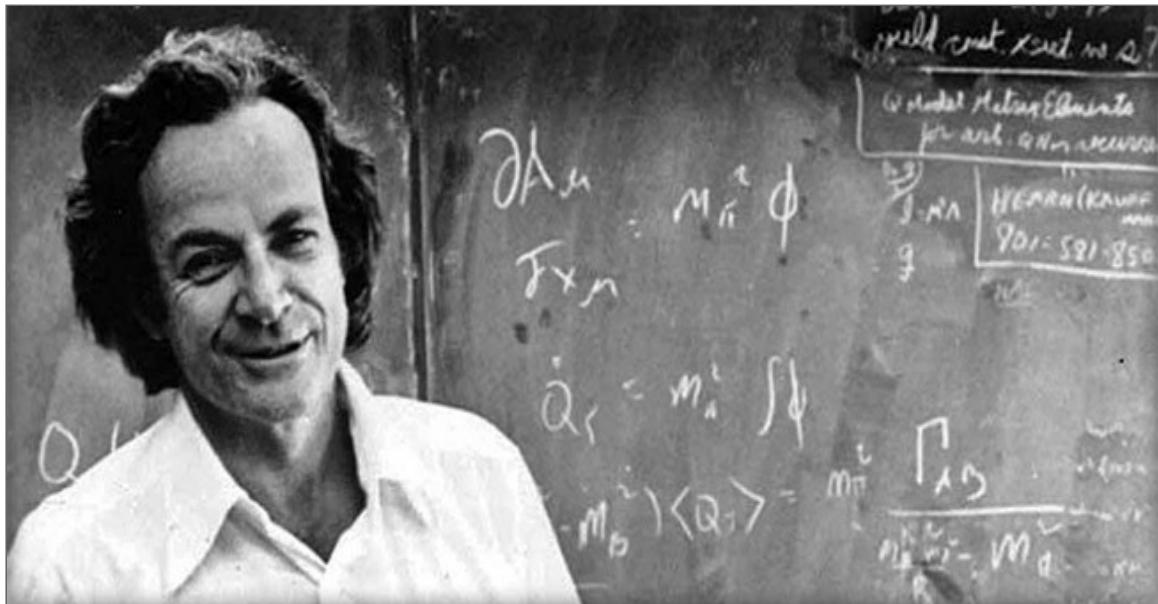
WHAT'RE YOU GOING TO LEARN TODAY?

- Who is your lecturer?
- Who are you and what do you want?
- Which topics will we cover?
- How will we do it?
- What do you have to do to pass?
- What's next?

WHO AM I?



SCIENCE



- Development of WWW
- PhD theoretical particle physics
- 60 research publications
- Assoc. Ed. Int. J. of Data Science
- Ed. Board Int. J. of Big Data Mgmt.
- Scientific member [d-cube@Berlin](#)

INDUSTRY



- Executive at Accenture & Shell
- Coach and consultant
- Certified psychotherapist
- Startup mentor

TEACHING



- Business informatics @HWR Berlin
- Visiting professor of data science @Lyon
- Adviser for CPU @LA
- Internship supervision

PLEASURE

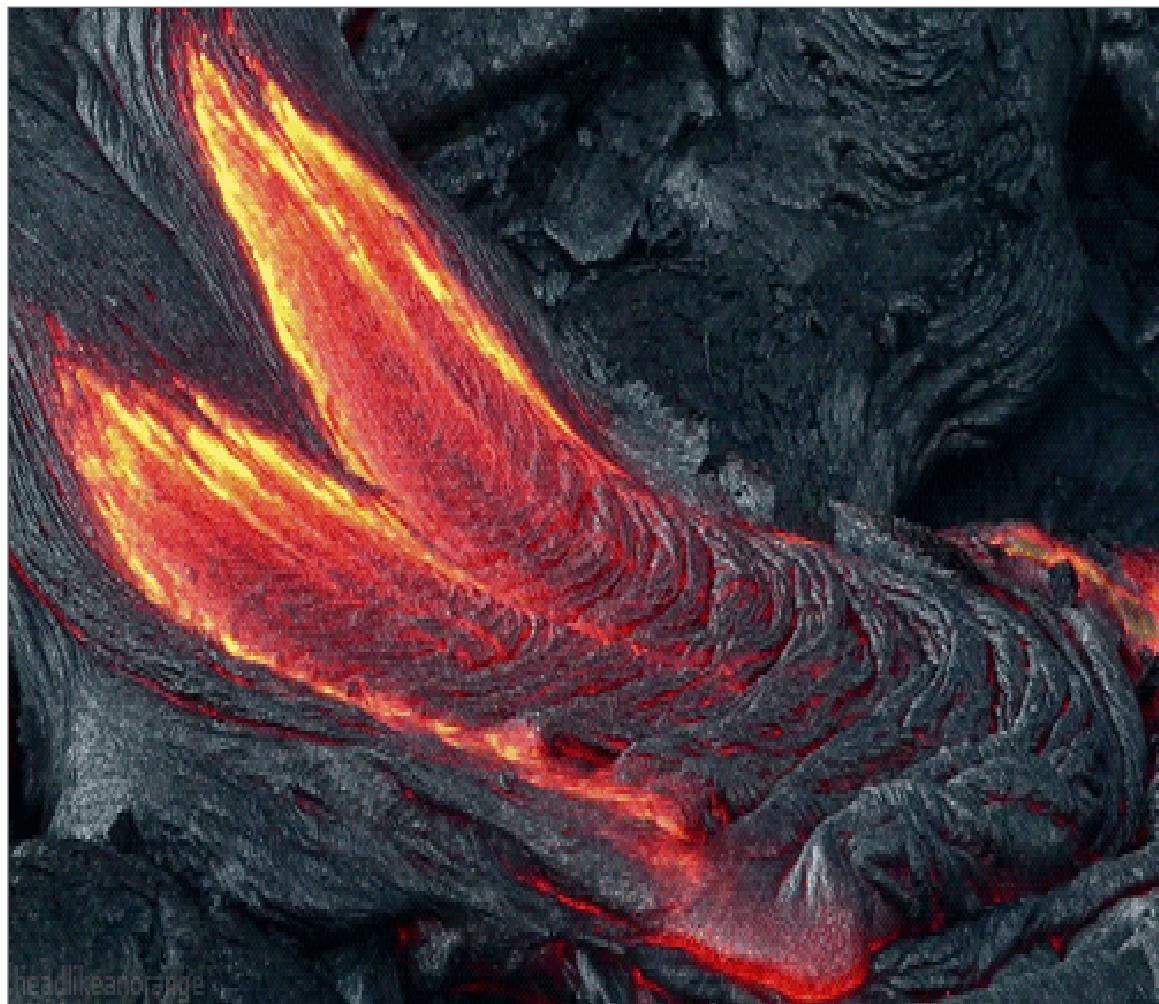


- Playing: **Assassin's Creed Valhalla** (2020)
- Reading: **Waugh, Sword of Honour** (1952-1961)
- Watching: **The Middle** (2009-2018)

WHAT ARE YOUR EXPECTATIONS?

- What do you want to learn here?
- What would you like to avoid?
- What did you take away from another course?
- What did you really not like in another course?

WHICH TOPICS WILL WE COVER?



INTRODUCTION TO DATA SCIENCE

6

Data visualization will go mainstream

In 2020, data visualizations helped us make sense of an increasingly complex world. Creating, critically understanding, and evaluating data visualizations will become a foundational skill for every citizen.

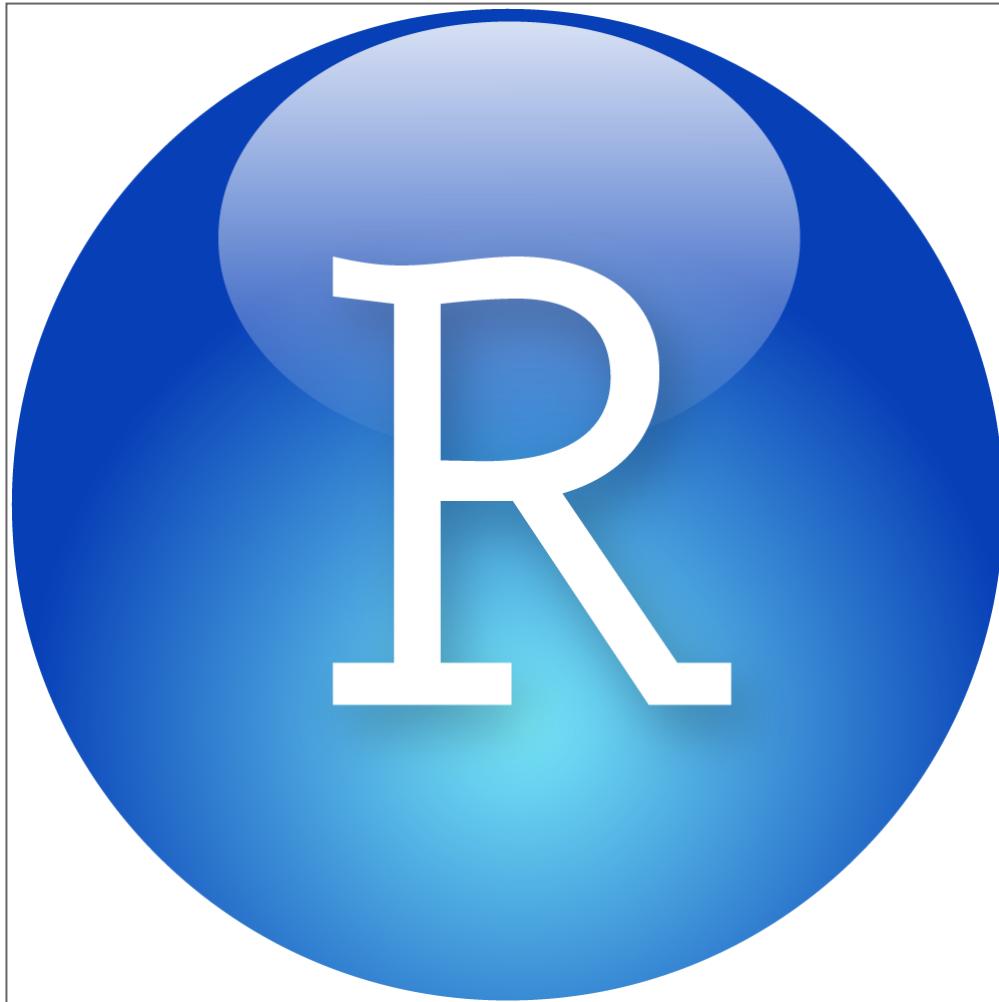
8

Data skills will cross over to every discipline

From primary to tertiary education, data literacy will become foundational for every discipline.

Source: [datacamp.com](https://www.datacamp.com)

INTRODUCTION TO R PROGRAMMING



Source: [RStudio](#)

VISUALIZATION USING R



Source: [Thomas Lin Pedersen](#)

FASTER APPROACH



- Focus on data exploration (EDA)
- Stay close to base R
- Use real data sets
- Compute interactively
- Prepare for DSC201 (ML)

SCHEDULE (SEE SYLLABUS)

No	Date	Lectures ¹	DataCamp ²	Tests/Quiz ³	fasterR ⁴
1	17-Aug	Overview			
2	19-Aug	On the R Shell			
3	24-Aug	Vectors in R			
4	26-Aug	Data frames in R			
5	31-Aug	Factors in R			
6	2-Sep	apply functions			
7	7-Sep	Cleaning data			
8	9-Sep	Lists in R			
9	14-Sep	Nile exploration			
10	16-Sep	Visualization			
11	21-Sep	Base R graphics			
12	23-Sep	Writing functions			
13	28-Sep	Iteration I			
14	30-Sep	Fibonacci series			
15	5-Oct	Literate Programming			
16	7-Oct	Conditions			
17	12-Oct	EDA example I			
18	14-Oct	Linear regression I			
19	19-Oct	Object-orientation			
20	21-Oct	EDA example II			
21	26-Oct	Packages			
22	28-Oct	Grammar of Graphics			
23	2-Nov	Functional Programming			
24	4-Nov	Text mining I			
25	9-Nov	Text mining II			
26	11-Nov	Linear regression II			
27	16-Nov	Dates and times			
28	18-Nov	Coding style			
29	23-Nov	Logistic regression			
30	25-Nov	Version control			
31	30-Nov	Iteration II			
32	2-Dec	Summary and outlook			
33	TBD				
VISUALIZATION					
APPLICATIONS					
				Final exam	

HOW WILL WE DO IT?



CLASSROOM SESSIONS



LECTURE SCRIPTS WITH EXERCISES ([GITHUB](#))

The screenshot shows a GitHub repository page for 'dsc101'. The URL bar at the top contains the URL <https://github.com/birkenkrahe/dsc101>. The page features a navigation bar with links for Pulls, Issues, Marketplace, Explore, and a user profile. Below the navigation bar, there are buttons for Raise issues, Watch, and Discuss. The 'Issues' button is highlighted with an orange box. The main content area displays a list of repository files and commits. A blue box highlights the '1_overview' folder. The 'Lectures' file is prominently displayed in a large blue box. Other files listed include LICENSE and README.md. On the right side, there are sections for About, Releases, and Packages, each with a 'Create a new release' or 'Publish your first package' link.

URL: https://github.com/birkenkrahe/dsc101

Birkenkrahe / dsc101

Raise issues **Watch** **Discuss**

Code **Issues** **Pulls** **Unwatch** 1 **Unstar** 0

Discussions

main Go to file Add file **Code** About

birkenkrahe Update README.md ... 15 minutes ago 30

1_overview Add files via upload Lectures 19 hours ago

LICENSE Initial commit 2 months ago

README.md Update README.md 15 minutes ago

README.md

dsc101

- Repository for DSC 101 - Data science methods and tools.
- Contains Emacs Org-mode files (rendered as Markdown), Wiki, Discussion forum.
- First offered: fall 2021.

About

Repository for DSC 101
- Data science methods and tools

Readme

GPL-3.0 License

Releases

No releases published [Create a new release](#)

Packages

No packages published [Publish your first package](#)

READING SUGGESTIONS



- Matloff: **fasteR: Fast Lane to Learning R!** (2021)
- Matloff: **The Art of R Programming** (2011)

VIDEO LECTURES ([YOUTUBE](#))



Vectors in R (part 1)

9 videos • 127 views • Last updated on Nov 23, 2020

Unlisted ▾

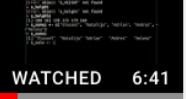
✖️ ➡️ ⋮

Everything is an object, and vectors are among the most important objects in R. In this video series, we cover creating, sorting, and measuring vectors. It may sound all a little technical, and it is, but it is an important building block towards our end game of data-driven storytelling. Once we understand objects and vectors, many other concepts will fall in our lap! - Note: try to complete the exercises at the end BEFORE looking at the solution video. The exercises are contained in the description of the solution video!

= SORT

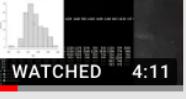
**1 Everything is an object**
Marcus Birkenrahe

**2 Assigning Objects**
Marcus Birkenrahe

**3 Who needs vectors anyway?**
Marcus Birkenrahe

**4 Creating vectors**
Marcus Birkenrahe

**5 Down the river Nile**
Marcus Birkenrahe

**6 Plotting histograms**
Marcus Birkenrahe

ONLINE ASSIGNMENTS (DATACAMP)

The screenshot shows the 'Team Assignments' page for the 'BSEL Berlin Analytics 21' team. A prominent orange callout box with the text 'Don't miss the deadline' points to the 'ACTIVE' tab in the filter bar. The page lists four assignments:

Title	Assignees	Status	Due By	C	A	CR	Details
Data Science for Everyone Introduction to Data Science Chapter	Team	Active	Mar 10, 10:00 CET	0	0	0%	View
Data Science for Everyone Data Collection and Storage Chapter	Team	Active	Mar 15, 10:00 CET	0	0	0%	View
Data Science for Everyone Preparation, Exploration, and Visualization Chapter	Team	Active	Mar 22, 10:00 CET	0	0	0%	View
Data Science for Everyone Experimentation and Prediction Chapter	Team	Active	Mar 29, 10:00 CEST	0	0	0%	View

- Register at DataCamp today!

TEAM EDA PROJECT

Kaggle Notebook (Left): Pima Indians Diabetes Database Analysis

Sample R project on Kaggle

corrrplot(correlat)

RStudio Session (Right):

```

is.edu/FasterR/data/Pima.csv",
header=TRUE)

27
28
29 The data set is in a CSV ("comma-separated-values") file. Here we
read it using the *read.csv* function. The
30 file header, if it exists, is the first
line in the file. If the file does not
have a header, you can set header to
31 *FALSE* and add one later using the names
function.
32
33 ##### Look at the *pima* dataset
34 Now, let's look at the data frame: in the
following code chunk, substitute the
correct function for *...* so that you
see the first few lines of the dataset
*pima*.
35 ````{r Print first few lines of pima}
36 head(pima)
37 ````
```

	pregnant	glucose	diastolic
1	6	148	72
2	1	85	66
3	8	183	64
4	1	89	66
5	0	137	40
6	5	116	74

6 rows | 4 of 9 columns

38 Execute the chunk to check that you did the right thing. The output will appear right below it. The data frame should have nine (numeric) variables and 768 observations.

34:91 Look at the "pima" dataset : R Markdown

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > OneDrive > R > workspace

Name Size

- .Rhistory 1.9 KB
- Pima_solutions.docx 188.6 KB
- Pima_solutions.Rmd 6 KB
- Pima_solutions.html 897.9 KB
- workspace.Rproj 205 B
- Pima_solutions.aux 662 B
- Pima_solutions.log 20.4 KB
- Pima_solutions.tex 15.6 KB
- notebook_2.knit.md 476 B
- notebook_2.nb.html 892.3 KB
- Pima_problems.Rmd 5.4 KB
- notebook_2.Rmd 426 B

Console

AGILE PROJECT MANAGEMENT



TESTS AND FINAL EXAM

DS101 Entry Quiz 📝

Challenge 🏆 Ends in 5 days

Start date: Feb 22 2021, 5:06 pm

End date: Mar 3 2021, 10:00 am

Hosted by birkenkrahe

Challenges are available throughout the course

Summary Players (10) Questions (20)

All (20) Difficult questions (6)

Question ▾ Type

Question	Type	Progress
1 Which of these are good problems for ...	Quiz	0%
2 Which of these are skills that data scie...	Quiz	0%
3 Which of these things have to do with...	Quiz	60%
4 Which part of the data science proces...	Poll	?
5 What is "R" (in data science)?	Quiz	0%
6 According to the TIOBE ranking, R is t...	True or false	80%

Quiz questions will be recycled in the final exam

PODCASTS AND FEEDS

Google Podcasts Search for podcasts

Build a Career in Data Science Sep 10, 2020

Chapter 1: What is Data Science?

48 min left

Podcasts on:
data science careers,
data-driven storytelling,
visualization,
applications

People also listened to

- Data Futurology - Leadership... Felipe Flores #144 Machine Learning: Getting th... We are joined by Alexey Grigorev for an episode that will be very useful for anyone wanting to identify the skills needed to g... 1 hr 5 min
- Data Stories Enrico Bertini and Moritz Stefaner 163 | svelte.js for web-based... 163 | svelte.js for web-based dataviz with Amelia Wattenberger 47 min

More episodes from Build a Career in Data Science

HWR Berlin

Feed Actions

- Analytics Vidhya https://www.analyticsvidhya.com/feed/ Learn everything about Analytics
- BBC Technology News http://newsrss.bbc.co.uk/rss BBC News - Technology
- R-Bloggers https://www.business-scien 'rss.xml' Blog feeds on technology, applications, current events
- CNET How-To http://feed.cnet.com/feed/how-to CNET editors and users share the top tech 'how to' tips and tricks with advice for getting the most out of all your gadgets.
- CNET Tech News http://feed.cnet.com/feed/news CNET news editors and reporters provide top technology news, with investigative reporting and in-depth coverage of tech issues and events.
- Data is beautiful https://www.reddit.com/r/dataisbeautiful/.rss A place to share and discuss visual representations of data: Graphs, charts, maps, etc.

Shared via GitHub / Schoology

SUMMARY OF COURSE ACTIVITIES

- Twice weekly classroom meetings
- Lecture scripts (GitHub)
- Reading assignments (Online)
- Video lectures (YouTube)
- **Online assignments** (DataCamp)
- **Team EDA projects** (Sprints)
- **Tests and final exam**
- Podcasts and feeds

WHAT DO YOU HAVE TO DO TO PASS?



DATA CAMP ASSIGNMENTS (> 50%)

1 Introduction to Data Science

100%

We'll start the course by defining what data science is. We'll cover the data science workflow and how data science is applied to real-world problems. We'll finish the chapter by learning about different roles within the data science field.

▶ What is data science?	Videos	✓ 50 xp
</> Customer segmentation workflow		✓ 100 xp
</> Building a customer service chatbot		✓ 100 xp
▶ Applications of data science		
</> Assigning data science project		
☰ Investment research		
▶ Data science roles and tools		✓ 50 xp
☰ Editing a job post	Mixed exercises	✓ 50 xp
</> Matching skills to jobs		✓ 100 xp
</> Classifying data tasks		✓ 100 xp

Videos

Ca. 30 min per chapter

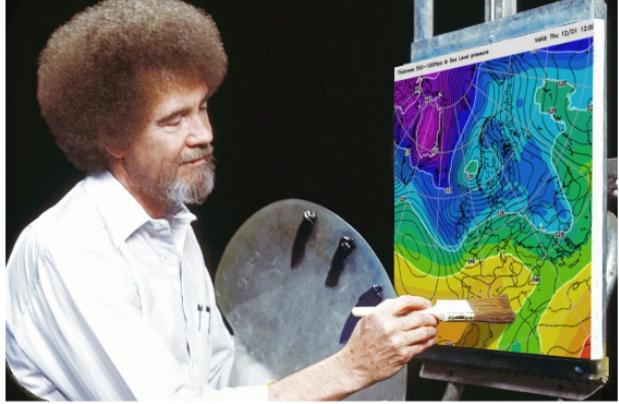
Mixed exercises

Complete at least 8 of 15 assignments

TEAM PROJECT (> 50%)

 **Election 2016 Trump-Clinton Spatial Visualization**
R notebook using data from [multiple data sources](#) · 1,275 views · 7mo ago
data visualization, exploratory data analysis, politics, +1 more

Kaggle sample R project



Version 10 of 10

[Notebook](#)

[Table Of Contents](#)

- [1. Packages](#)
- [2. How To Create A Map Using Ggplot2](#)
- [3. Election Data & First Map](#)
- [4. Trump Vs Clinton](#)
- [5. Statebins](#)
- [6. References](#)

[Input \(2\)](#)

[Output](#)

[Execution Info](#)

[Log](#)

[Comments \(12\)](#)

Table of Contents

- [1. Packages](#)
- [2. How to create a map using ggplot2](#)
- [3. Election Data & First Map](#)
- [4. Trump vs Clinton](#)
- [5. Statebins](#)
- [6. References](#)

Present on Nov 30 or Dec 2

WHAT IS A TEAM PROJECT?

- Description of the dataset
- Introduction of the problem statement
- Description of the methods used
- Visualization of the data (plots!)
- Analysis of the plots
- Limitations of own analysis
- References

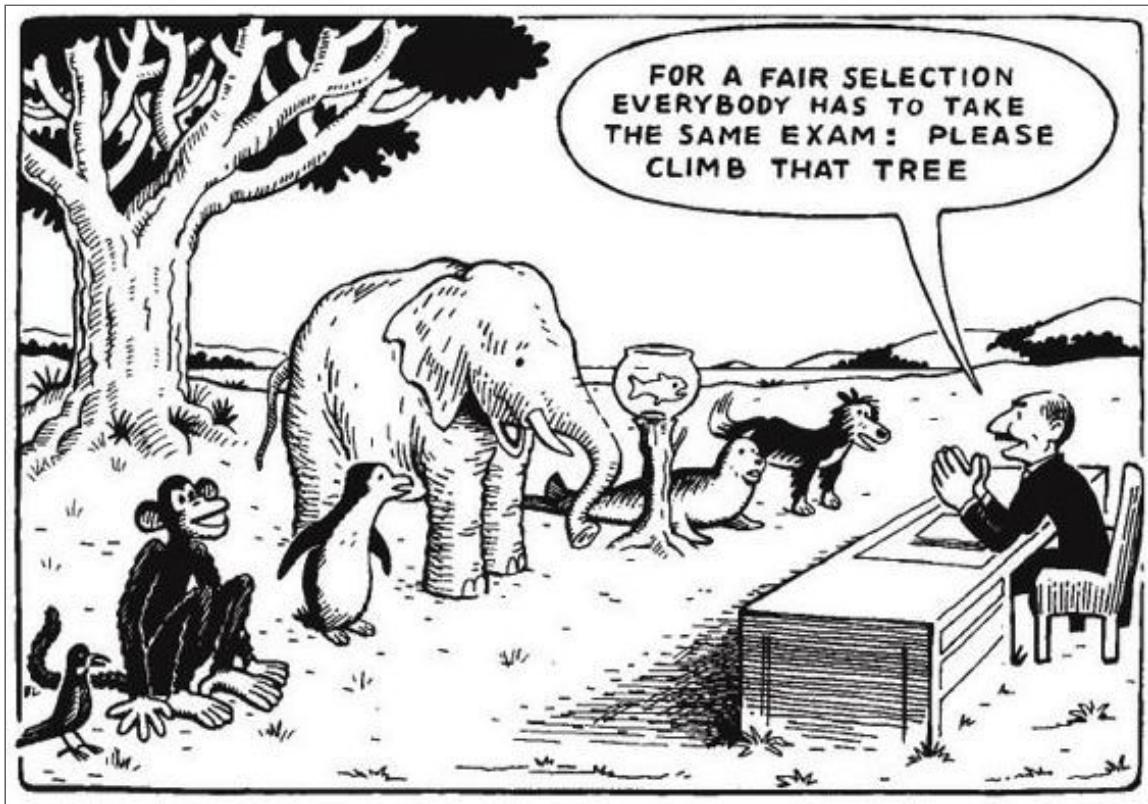
DO YOU HAVE PROJECT EXAMPLES?

- Examples on Kaggle ([example](#))
- Examples on data science blogs ([example](#))
- Translate from Python to R ([example](#))
- Extend someone else's EDA ([example](#))
- Document an R package ([example](#))
- Use your own data ([example](#))

CAN I DO A PROJECT AS AN ABSOLUTE BEGINNER?

- Keep It Simply Scientific (IMRaD)
- Look at examples (e.g. in my bookmarks)
- Create data set (e.g. your productivity)
- Researchers are beginners

FINAL EXAM (> 50%)



Final exam: date TBD

WHAT'S NEXT?



IN THE COURSE

- Intro to Data science (Lecture)
- Intro to DataCamp (Practice)
- Intro to GitHub (Productivity)
- Intro to R (Language)

YOUR CHALLENGES

What?	When?
Register at DataCamp	Today
Register at GitHub	Today
Complete test challenge	Aug 24
Complete DataCamp assignment	Aug 24
Set up team project (2-3 ppl)	Sep 2
Check FAQs x 2 in GitHub	n.d.
Ask questions (class/GitHub)	n.d.

**) do this every week until December*

ANY QUESTIONS?



A copy of this presentation is available.