Data Visualisation & Data Organisation in Spreadsheets

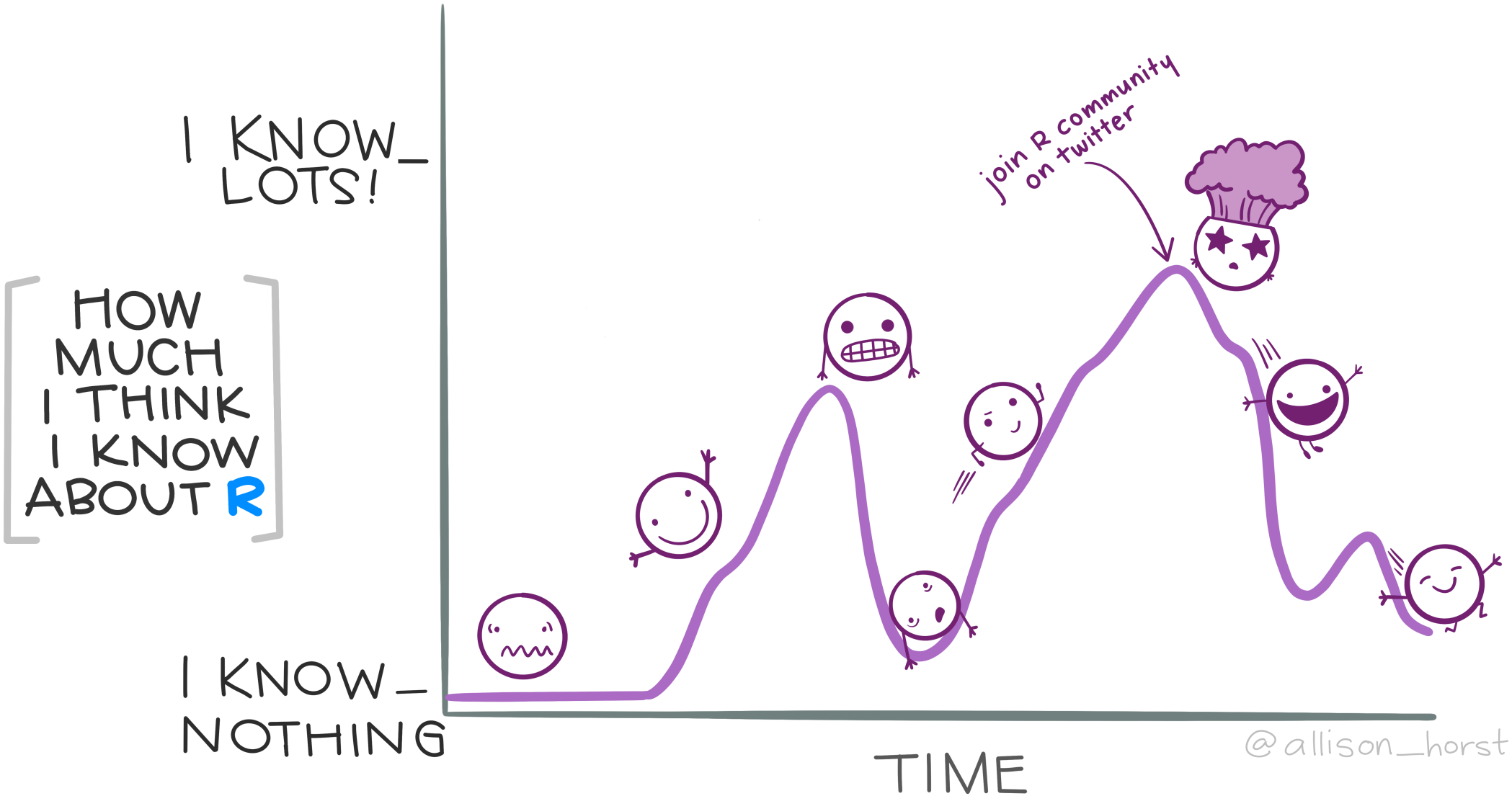
CVEN 5837 - Summer 2022

Lars Schöbitz

https://cven5837-ss22.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER

https://cven5837-ss22.github.io/website/



Artwork by @allison_horst

https://cven5837-ss22.github.io/website/

Solving coding problems

Tipps for search engines

- Use actionable verbs that describe what you want to do
- Be specific
- Add R to the search query
- Add the name of the R package name to the search query
- Scroll through the top 5 results (don’t just pick the first)

Example: “How to remove a legend from a plot in R ggplot2”

Stack Overflow

What is it?

- The biggest support network for (coding) problems
- Can be intimidating at first
- Up-vote system

Workflow

- First, briefly read the question that was posted
- Then, read the answer marked as “correct”
- Then, read one or two more answers with high votes
- Then, check out the “Linked” posts
- Always give credit for the solution

Give credit

528

from r cookbook, where bp is your ggplot:

Remove legend for a particular aesthetic (fill):

```
bp + guides(fill="none")
```

It can also be done when specifying the scale:

```
bp + scale_fill_discrete(guide="none")
```

This removes all legends:

```
bp + theme(legend.position="none")
```

Share Edit Follow Flag

edited Dec 2, 2021 at 7:07

Andrew Morris

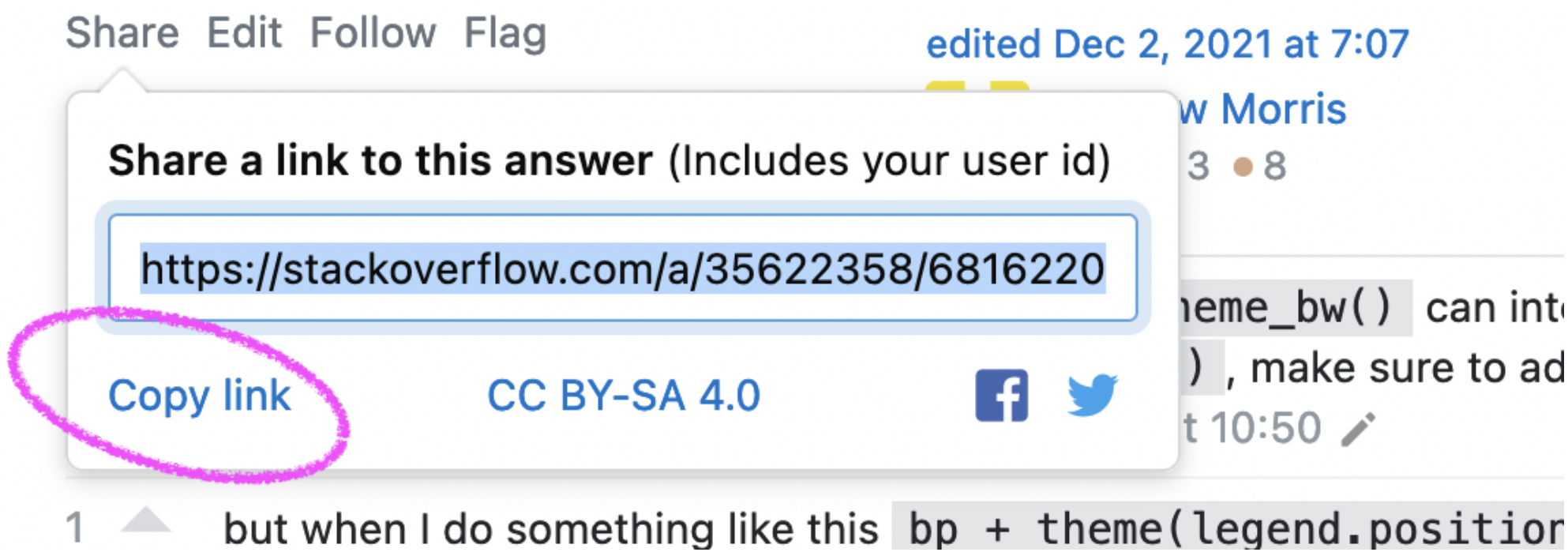
408 ● 3 ● 8

answered Feb 25, 2016 at 8:48

user3490026

5,388 ● 1 ● 9 ● 4

Give credit


Share Edit Follow Flag
edited Dec 2, 2021 at 7:07
w Morris
3 8
Share a link to this answer (Includes your user id)
https://stackoverflow.com/a/35622358/6816220
Copy link
CC BY-SA 4.0
eme_bw() can int
) , make sure to ad
t 10:50
1
but when I do something like this bp + theme(legend.positior

Give credit

```
ggplot(data = global_waste_data_kg_year,
       mapping = aes(x = income_id,
                     y = capita_kg_year,
                     color = income_id)) +
  ## Remove legend ref: https://stackoverflow.com/a/35622358/6816220
  theme(legend.position = "none")
```

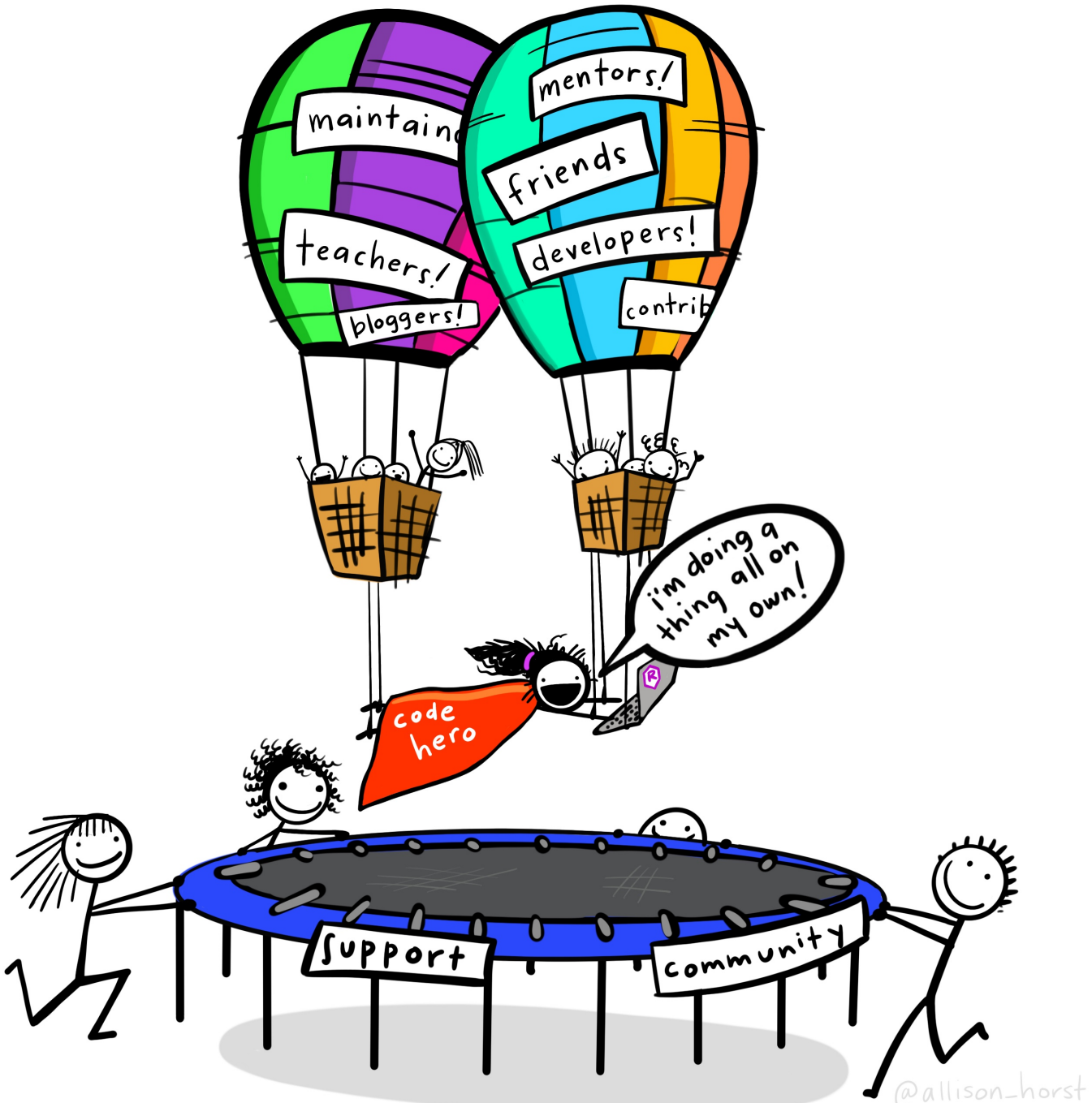
Other sources for help

- Our Google Space for the course
- RStudio Community Forum: https://community.rstudio.com/
- Documentation websites: https://dplyr.tidyverse.org/
- Twitter community: #rstats

Artwork by @allison_horst

Minimal reproducible example (reprex)

- Needed when asking questions online
- Good support information: https://www.tidyverse.org/help/#reprex



https://cven5837-ss22.github.io/website/

Artwork by @allison_horst

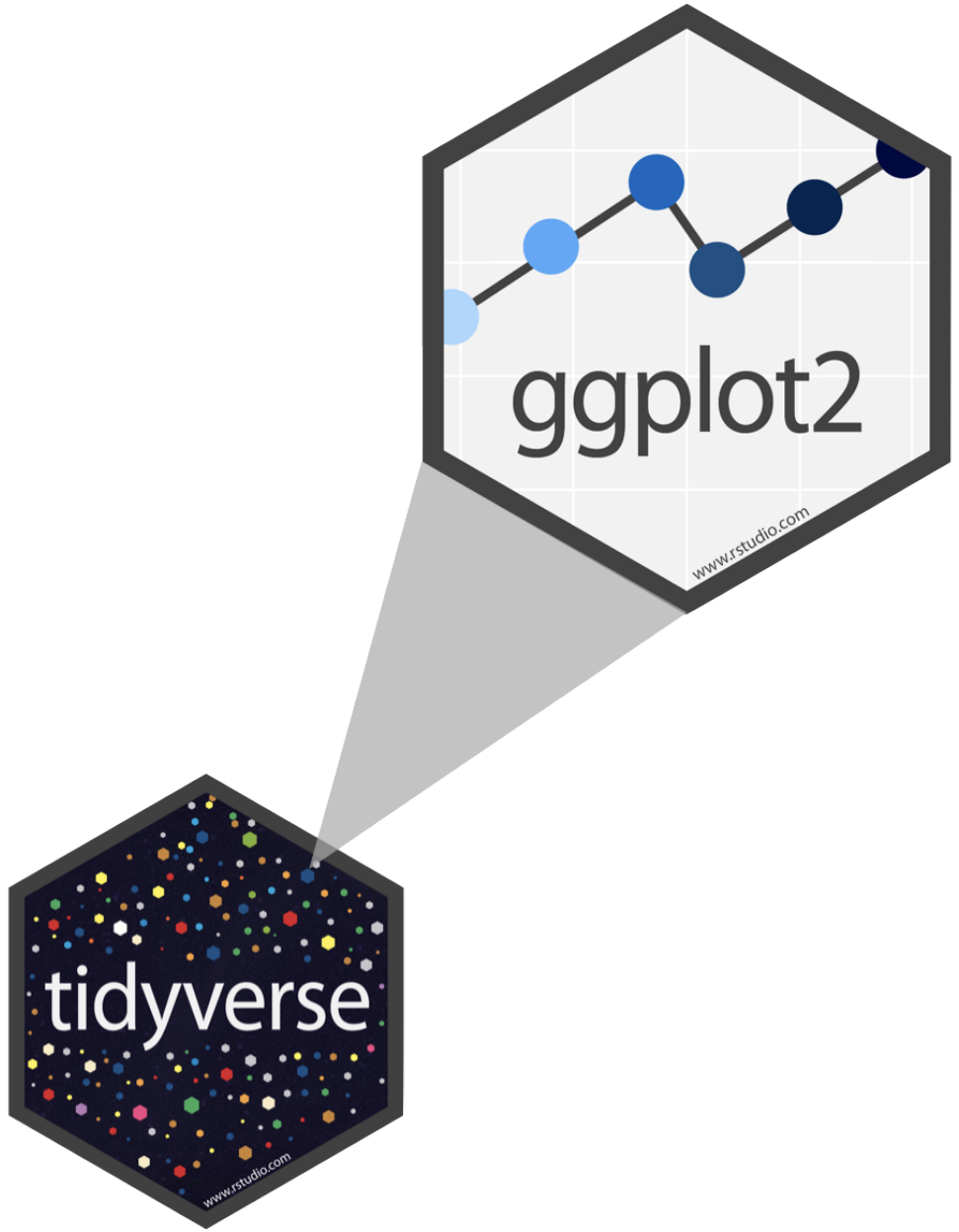
Learning Objectives (for this week)

1. Learners can describe the four main aesthetic mappings that can be used to visualise data using the ggplot2 R Package
2. Learners can control the colour scaling applied to a plot using colour as an aesthetic mapping
3. Learners can compare three different geoms and their use case
4. Learners can apply a theme to control font types and sizes within a plot
5. Learners can apply 12 principles for data organisation in spreadsheets in the layout of a collected dataset

Exploratory Data Analysis with `ggplot2`

R Package `ggplot2`

- **ggplot2** is tidyverse's data visualization package
- `gg` in `ggplot2` stands for Grammar of Graphics
- Inspired by the book **Grammar of Graphics** by Leland Wilkinson
- **Documentation:** https://ggplot2.tidyverse.org/
- **Book**: https://ggplot2-book.org



Code structure

- `ggplot()` is the main function in ggplot2
- Plots are constructed in layers
- Structure of the code for plots can be summarized as

```
ggplot(data = [dataset],
       mapping = aes(x = [x-variable],
                     y = [y-variable])) +
   geom_xxx() +
   other options
```

Code structure

```
1  ggplot()
```

Code structure

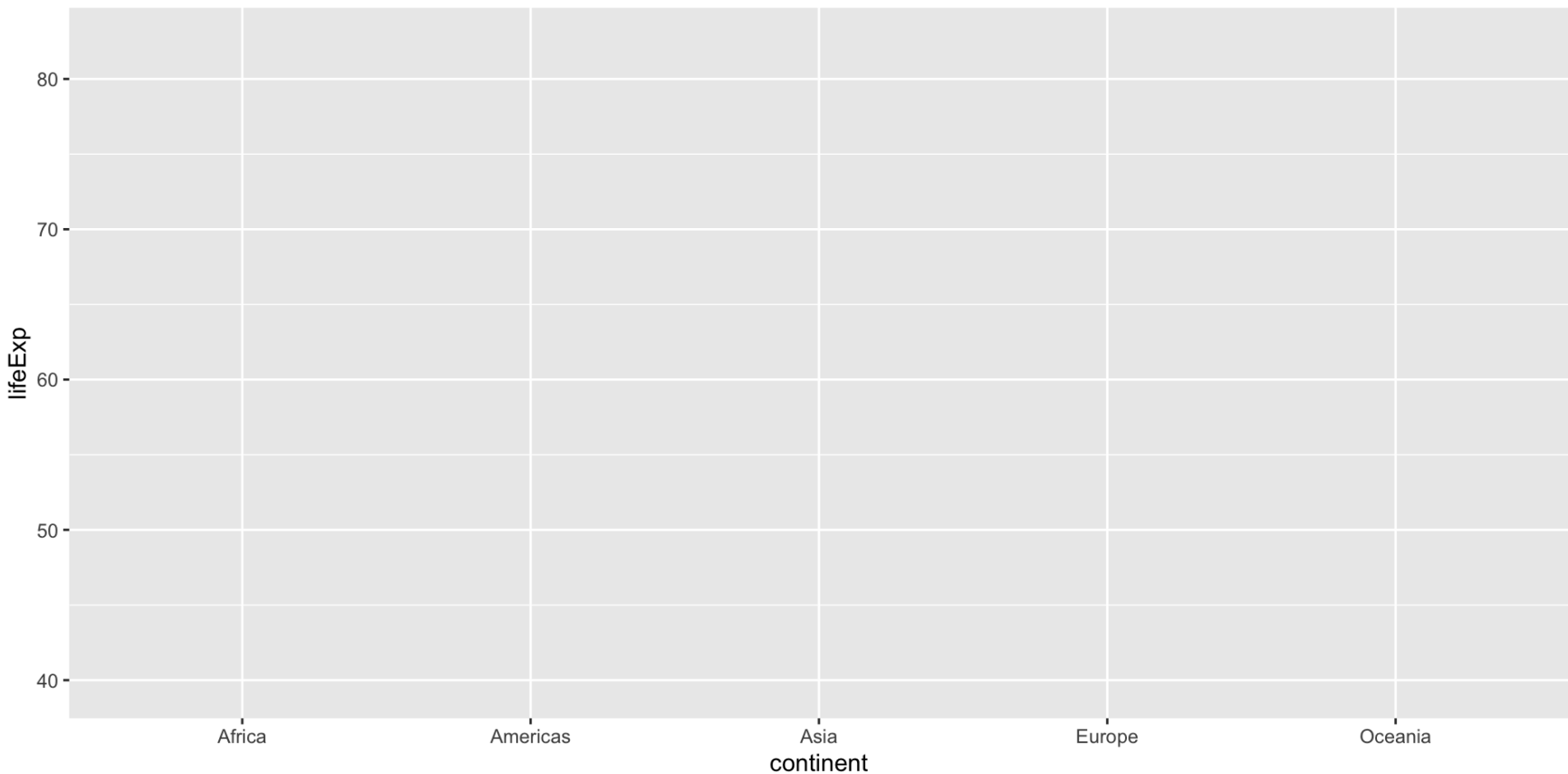
```
1  ggplot(data = gapminder_yr_2007)
```

Code structure

```
1  ggplot(data = gapminder_yr_2007,
2         mapping = aes())
```

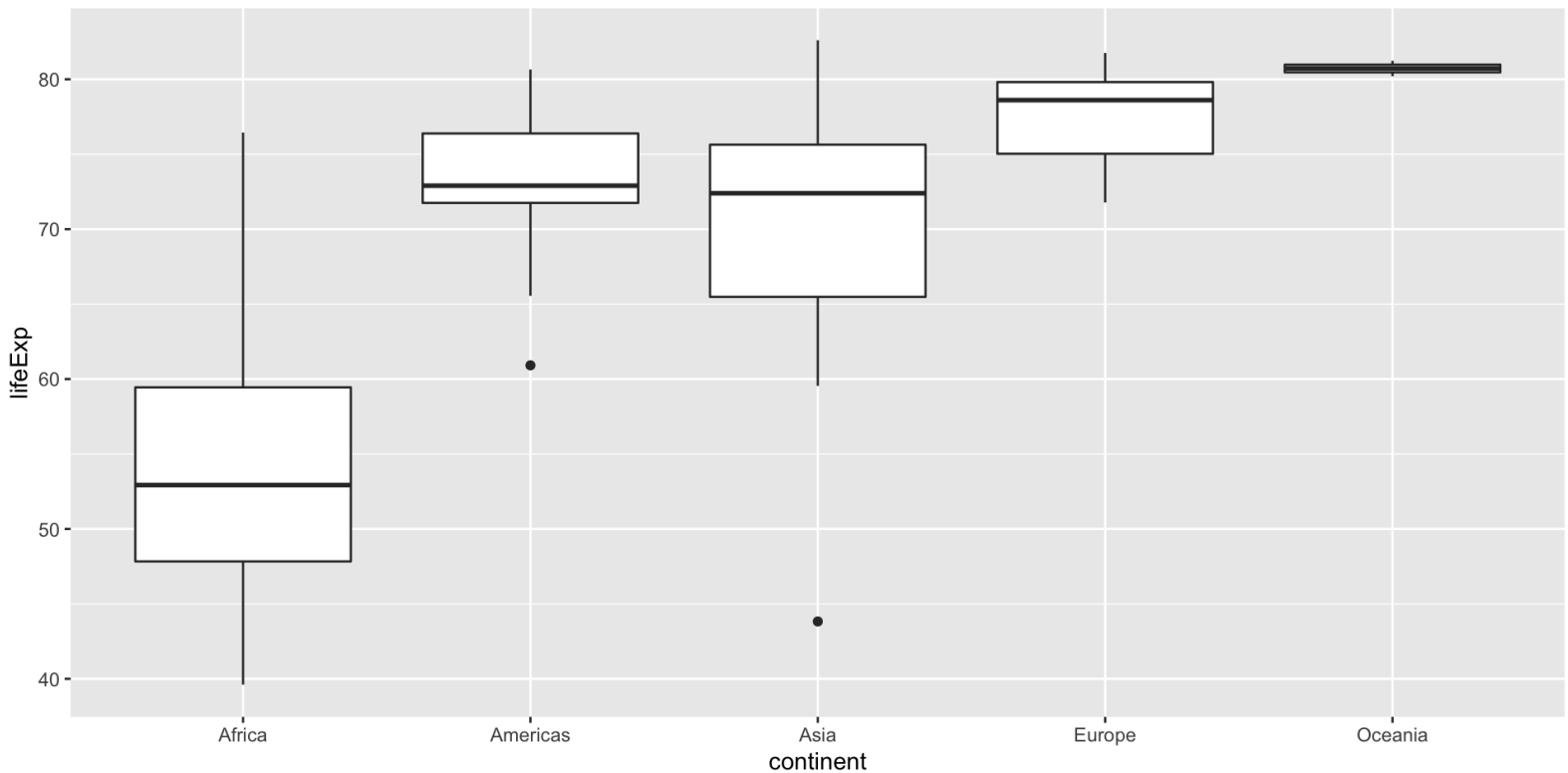
Code structure

```
1 ggplot(data = gapminder_yr_2007,
2        mapping = aes(x = continent,
3                      y = lifeExp))
```



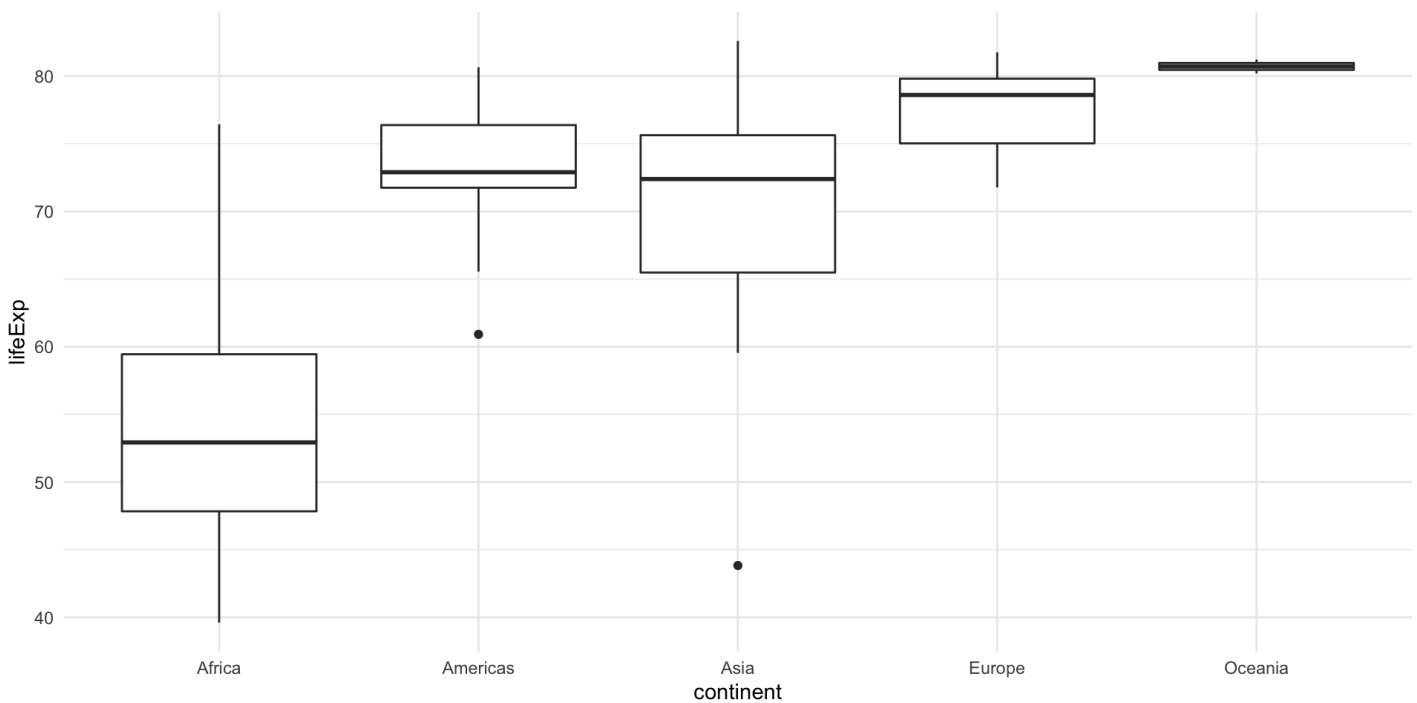
Code structure

```
1  ggplot(data = gapminder_yr_2007,
2         mapping = aes(x = continent,
3                       y = lifeExp)) +
4    geom_boxplot()
```



Code structure

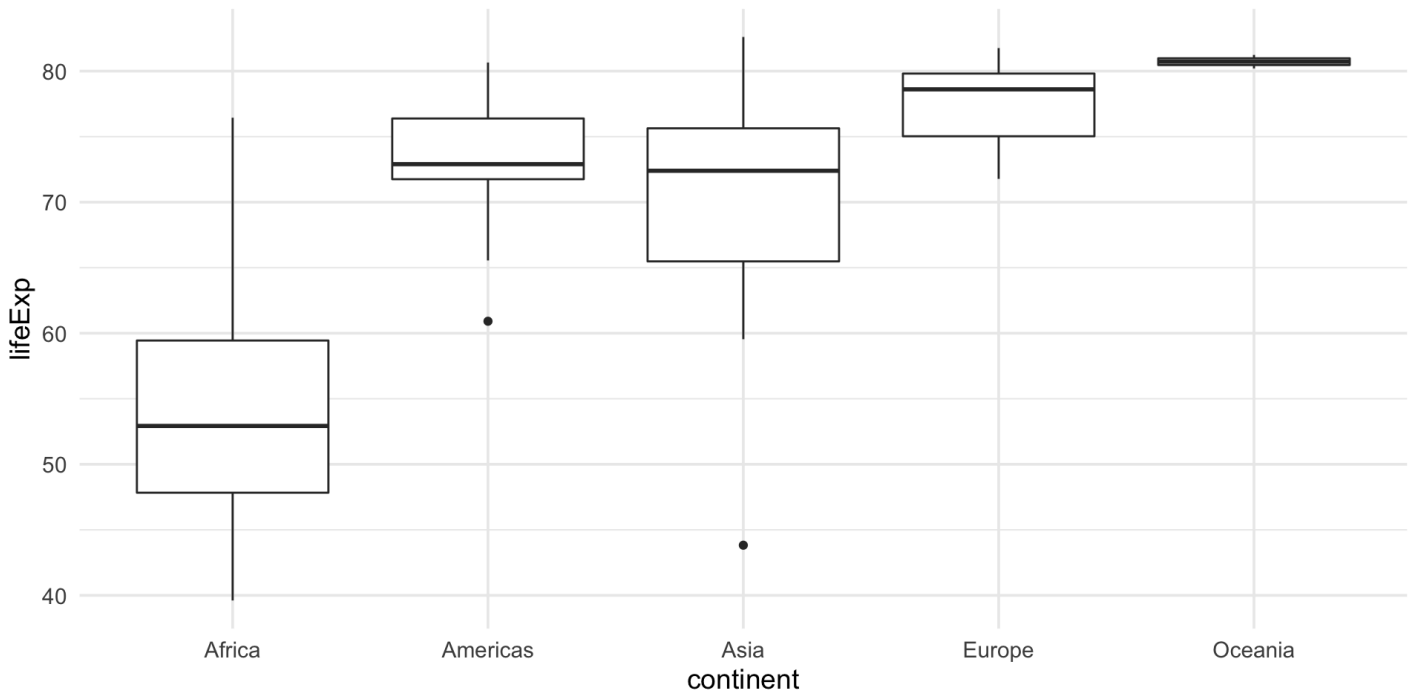
```
ggplot(data = gapminder_yr_2007,
       mapping = aes(x = continent,
                     y = lifeExp)) +
  geom_boxplot() +
  theme_minimal()
```



https://cven5837-ss22.github.io/website/

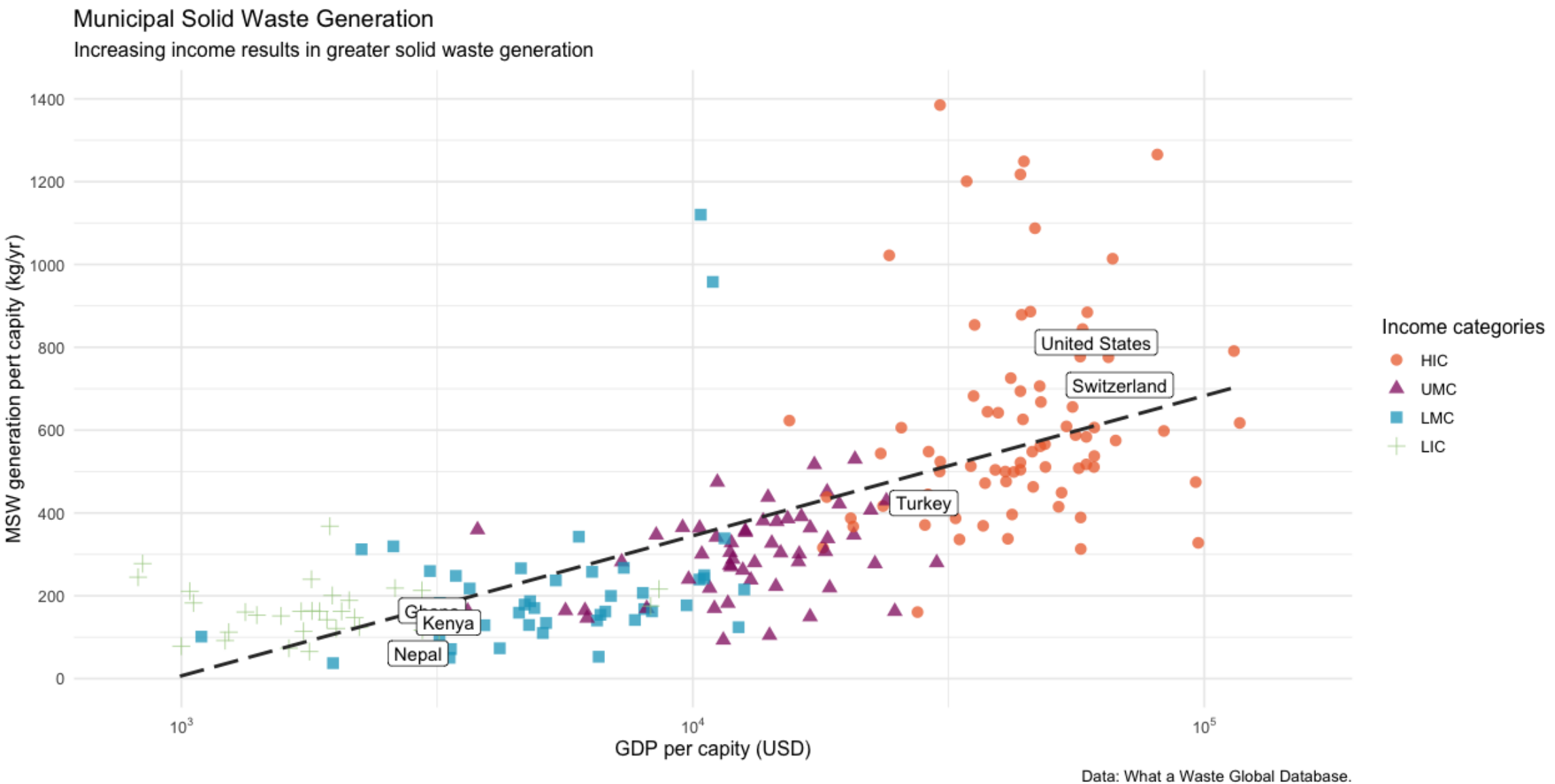
Code structure

```
ggplot(data = gapminder_yr_2007,
       mapping = aes(x = continent,
                     y = lifeExp)) +
  geom_boxplot() +
  theme_minimal(base_size = 14)
```



Live Coding Exercise: Reproduce this plot

https://cven5837-ss22.github.io/website/


Municipal Solid Waste Generation
Increasing income results in greater solid waste generation
MSW generation pert capity (kg/yr)
1400
1200
1000
800
600
400
200
0
10^3
10^4
10^5
GDP per capity (USD)
United States
Switzerland
Turkey
Kenya
Nepal
Income categories
HIC
UMC
LMC
LIC
Data: What a Waste Global Database.

live-02a-data-visualiation

1. Head over to rstudio.cloud
2. Open the workspace for the course (cven5837-ss22)
3. Open “Projects”
4. Open the “course-materials” project
5. Follow along with me

Break

10:00

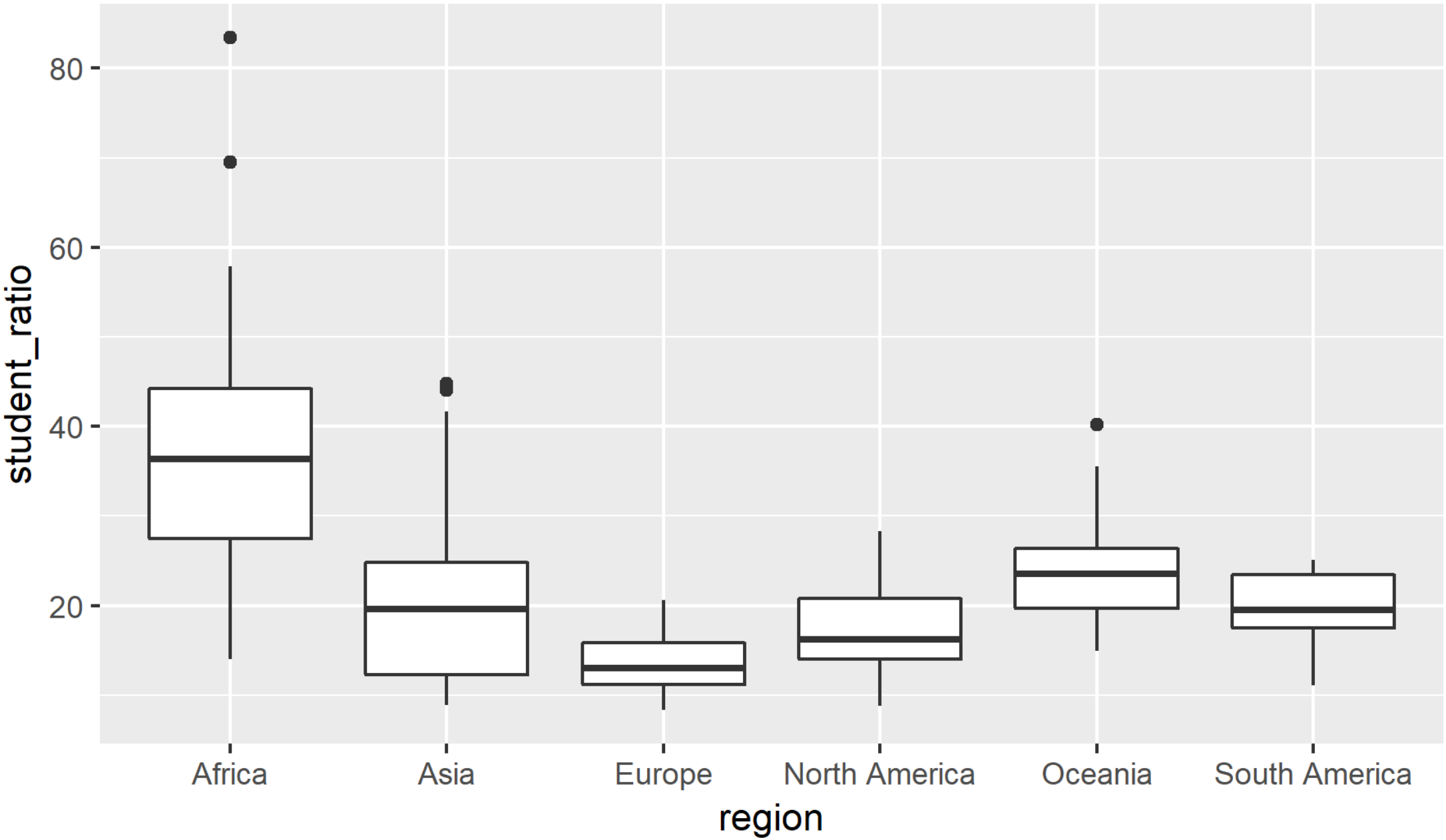
https://cven5837-ss22.github.io/website/

Photo by Blake Wisz

Visualising numerical data

Types of variables

The Evolution of a ggplot



Data: UNESCO Institute for Statistics
Visualization by Cédric Scherer

Tutorial: https://www.cedricscherer.com/2019/05/17/the-evolution-of-a-ggplot-ep.-1/

data-to-viz.com

https://cven5837-ss22.github.io/website/

Data Collection Tools

Data Collection Tools

- Questionnaires for survey based data
- Spreadsheets for manual experimental/observational data
- Sensors for automated near real-time data

Survey tools

Commonly used in the Global Engineering and Development sector

- KOBO Toolbox
- mWater
- OpenDataKit

Data Organisation in Spreadsheets

Listen

Article

Data Organization in Spreadsheets

Karl W. Broman & Kara H. Woo

Pages 2-10 | Received 01 Jun 2017, Accepted author version posted online: 29 Sep 2017, Published online:24 Apr 2018

Download citation https://doi.org/10.1080/00031305.2017.1375989 Check for updates

Jorunal: The American Statistician, Screenshot taken on 2022-03-23

Data Organisation in Spreadsheets

Read the paper (it's part of your homework), but you can also:

- Go through the annotated slides: https://kbroman.org/Talk_DataOrg/dataorg_notes.pdf
- Watch Karl Broman give the talk (02:36 to 45:00): https://youtu.be/t74E0a90gkA?t=156
- Read the content on a website: https://kbroman.org/dataorg/

But, especially apply it to your data

https://cven5837-ss22.github.io/website/

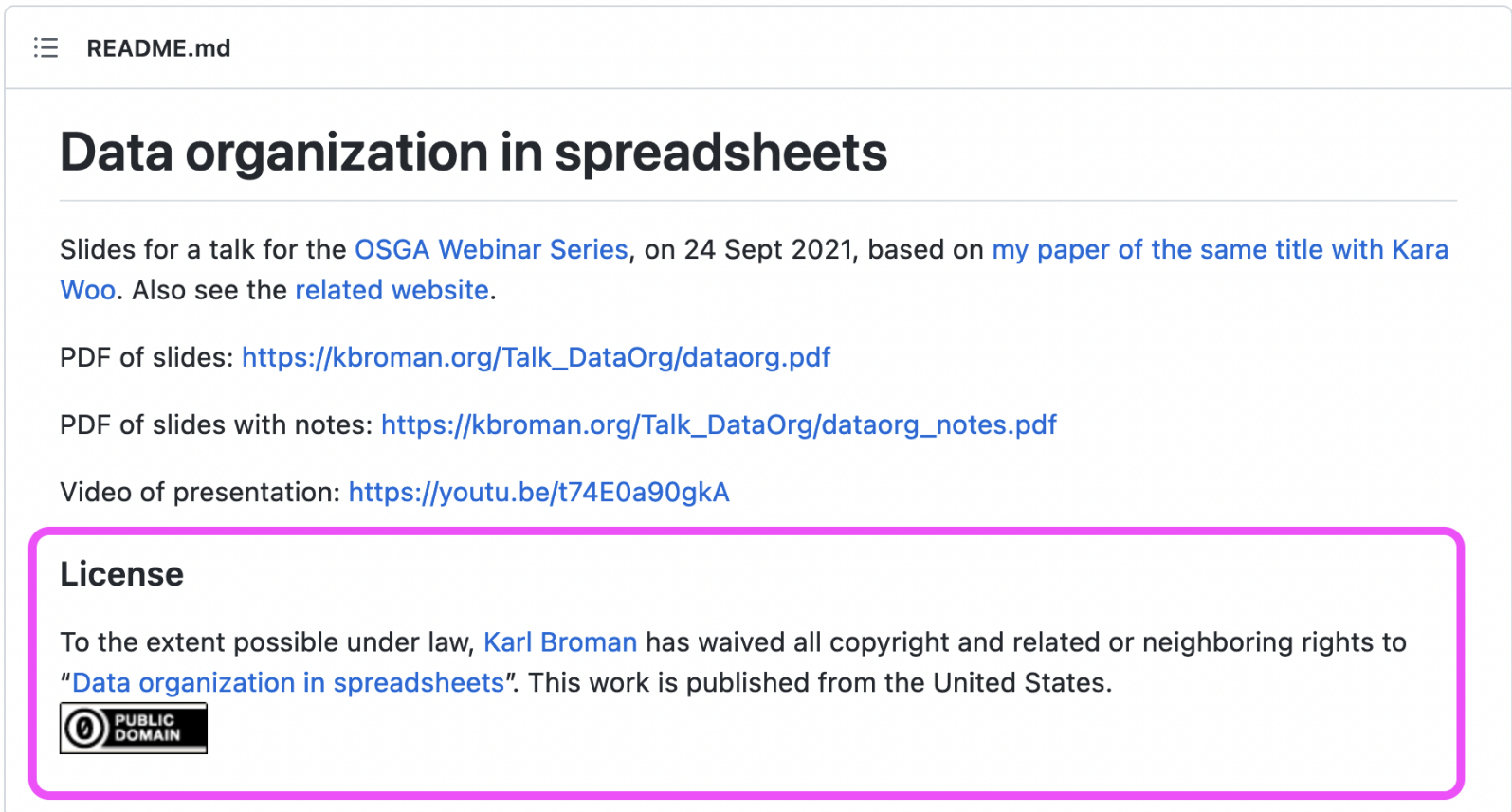
via GIPHY

Why?

Because it will make your life easier!

Jorunal: The American Statistician, [Screenshot taken on 2022-07-11]

License? CC0 (!)



README.md

Data organization in spreadsheets

Slides for a talk for the OSGA Webinar Series, on 24 Sept 2021, based on my paper of the same title with Kara Woo. Also see the related website.

PDF of slides: https://kbroman.org/Talk_DataOrg/dataorg.pdf

PDF of slides with notes: https://kbroman.org/Talk_DataOrg/dataorg_notes.pdf

Video of presentation: https://youtu.be/t74E0a90gkA

License

To the extent possible under law, Karl Broman has waived all copyright and related or neighboring rights to "Data organization in spreadsheets". This work is published from the United States.

PUBLIC DOMAIN

(https://github.com/kbroman/Talk_DataOrg) .footnote[Screenshot taken on 2022-03-23]

Karl Broman, GitHub repo broman/Talk_DataOrg, Screenshot taken on 2022-03-23

Pair Programming Exercise

https://cven5837-ss22.github.io/website/

Pair Programming Exercises

- Two learners work together in a break out session
- One person (the driver) shares the screen and does the typing
- The other person (the navigator) offers comments and suggestions
- Roles get switched

hw-02a-data-visualiation

1. Head over to rstudio.cloud
2. Open the workspace for the course (cven5837-ss22)
3. Open “Projects”
4. Open the “course-materials” project

Homework week 2

https://cven5837-ss22.github.io/website/

Bring your own data

- Generate data doing a short survey or observational study
- Find a data online that interests you
- Use a dataset that you already have available

Homework due dates

- All material on course website
- Homework assignment due: Friday, 15th July
- Learning reflection due: Monday, 18th July

Thanks! 

https://cven5837-ss22.github.io/website/

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/ Access slides as PDF on GitHub