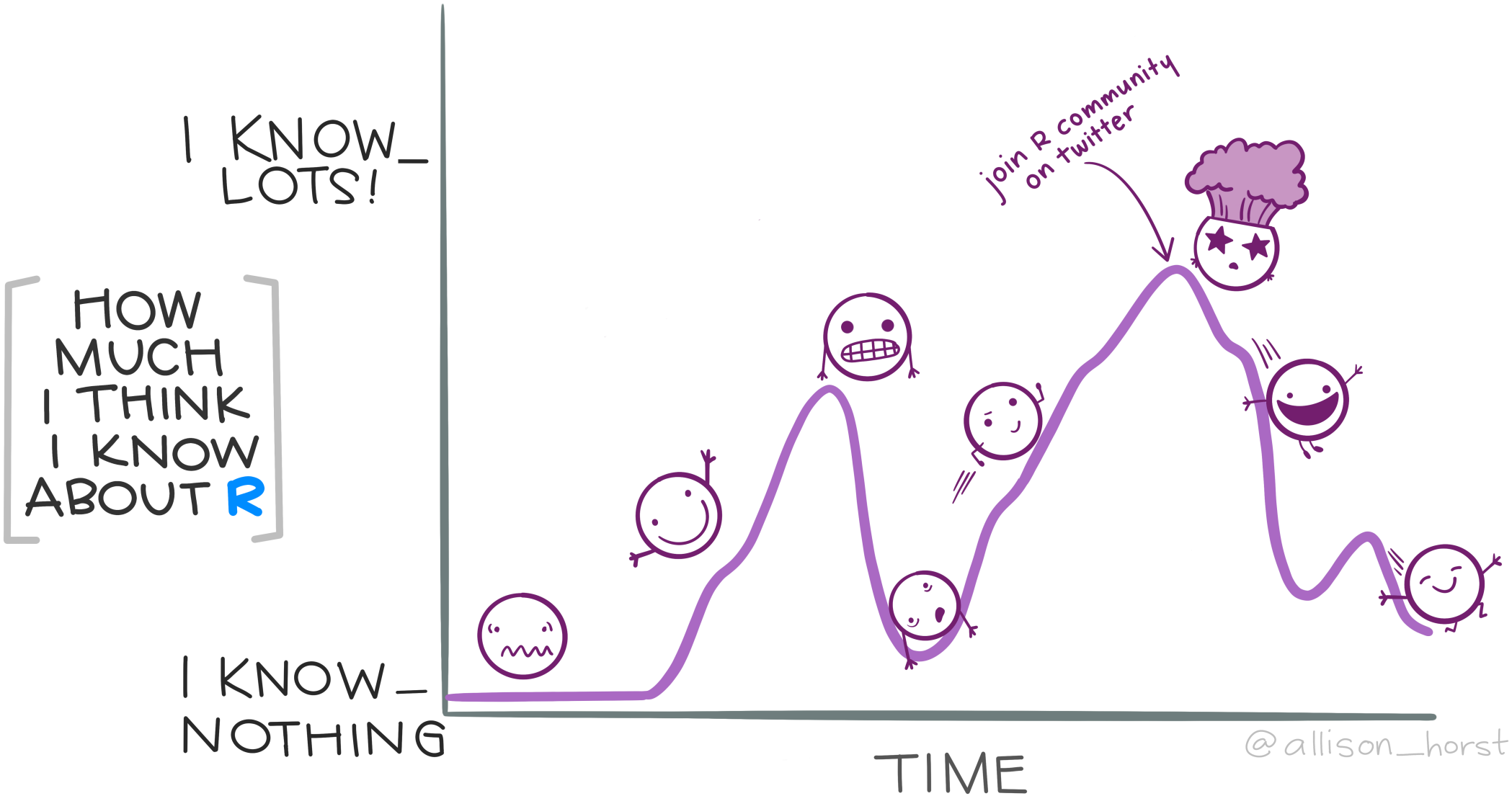
Exploratory data analysis using visualization & Data organization in spreadsheets

CVEN 5999 - Summer 2025

Lars Schöbitz

cven5999-ss25.github.io/website/

Mortenson Center in Global Engineering
UNIVERSITY OF COLORADO BOULDER


I KNOW LOTS!
HOW MUCH I THINK I KNOW ABOUT R
I KNOW NOTHING
join R community on twitter
TIME
@allison_horst

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

Tipps for AI tools

- 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

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

Other sources for help

- Posit Community Forum: https://forum.posit.co/
- Documentation websites: https://dplyr.tidyverse.org/
- bsky community: https://bsky.app/hashtag/RStats

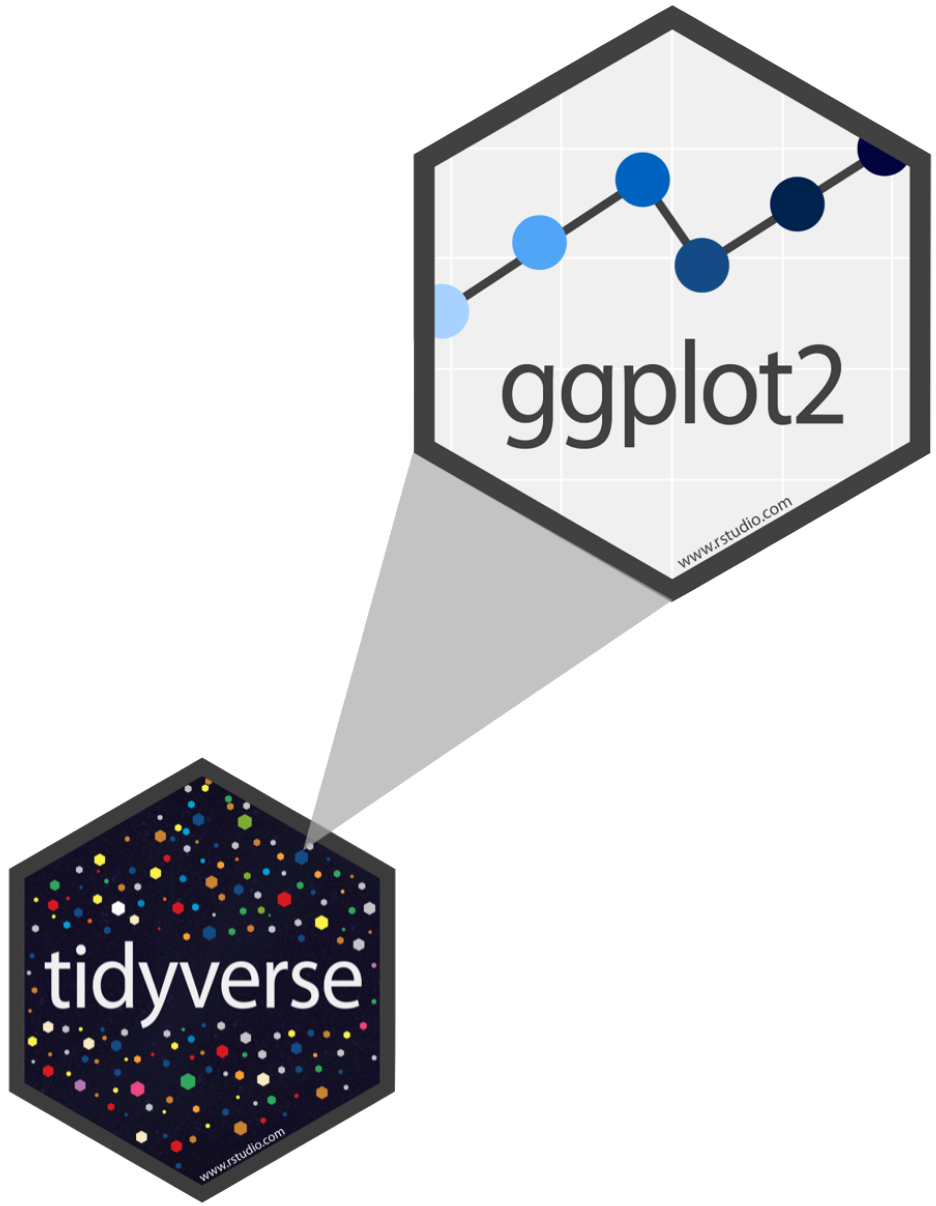
Learning Objectives (for this week)

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

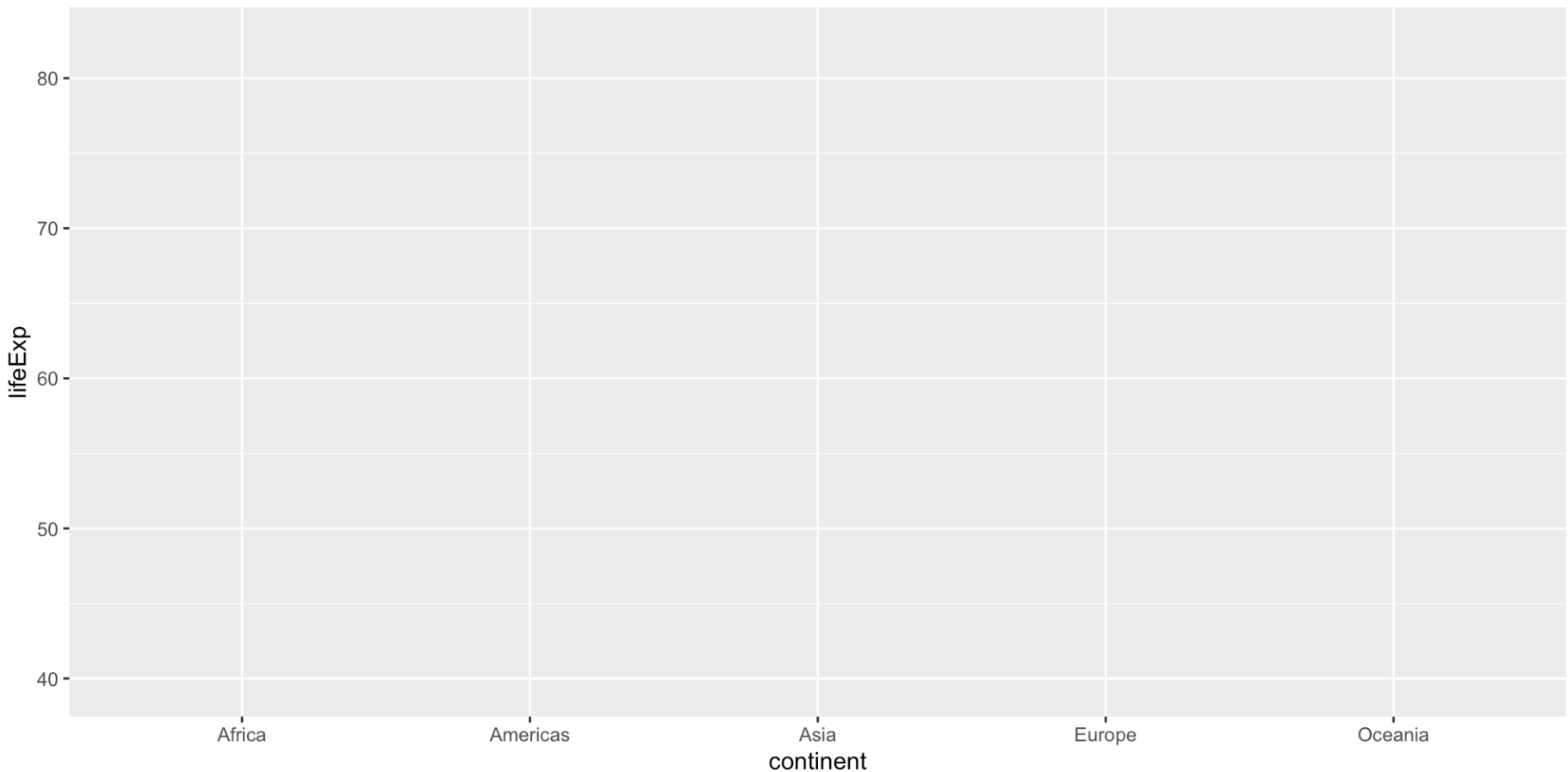
```
ggplot(data = gapminder_yr_2007)
```

Code structure

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

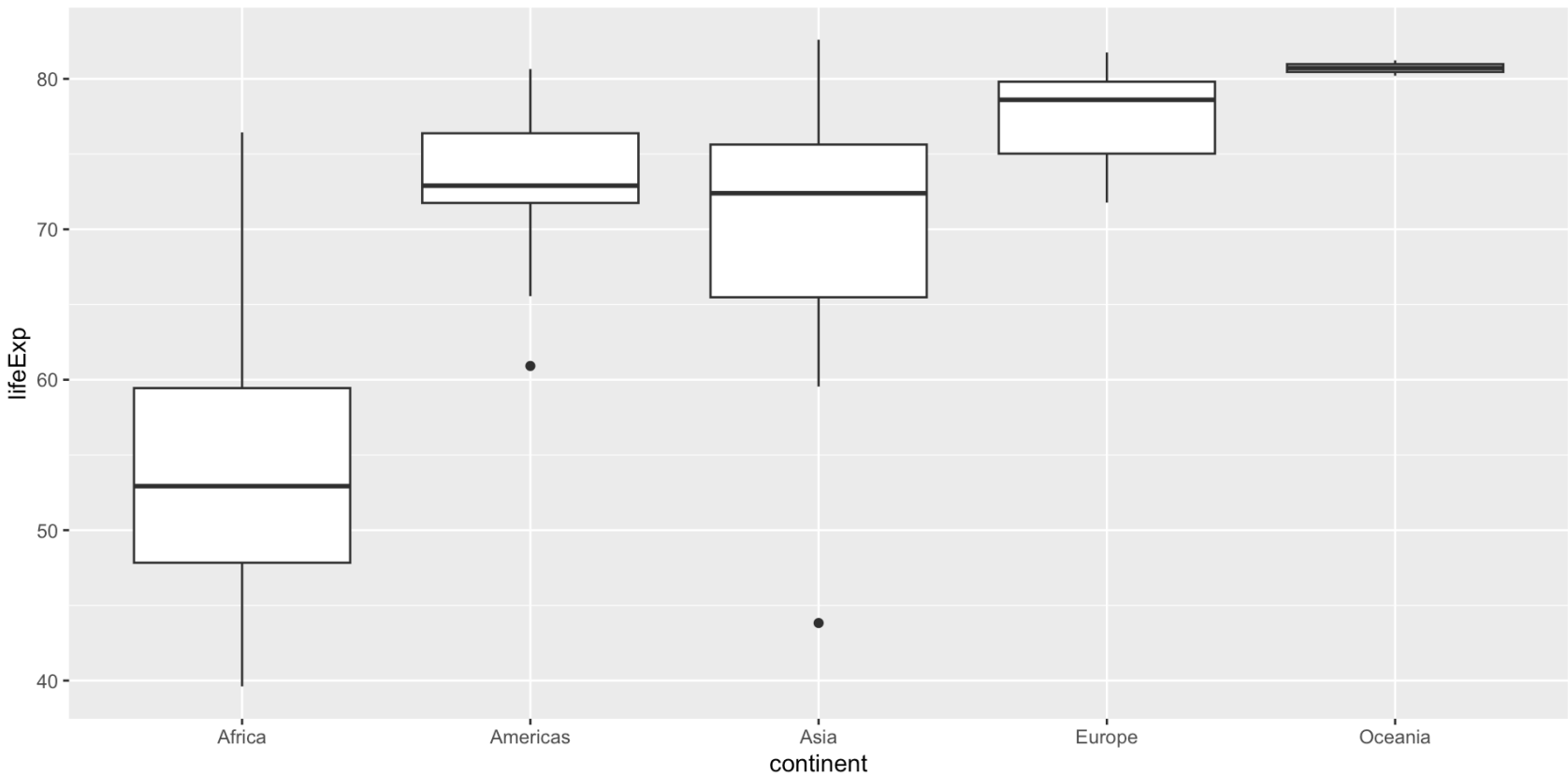
Code structure

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



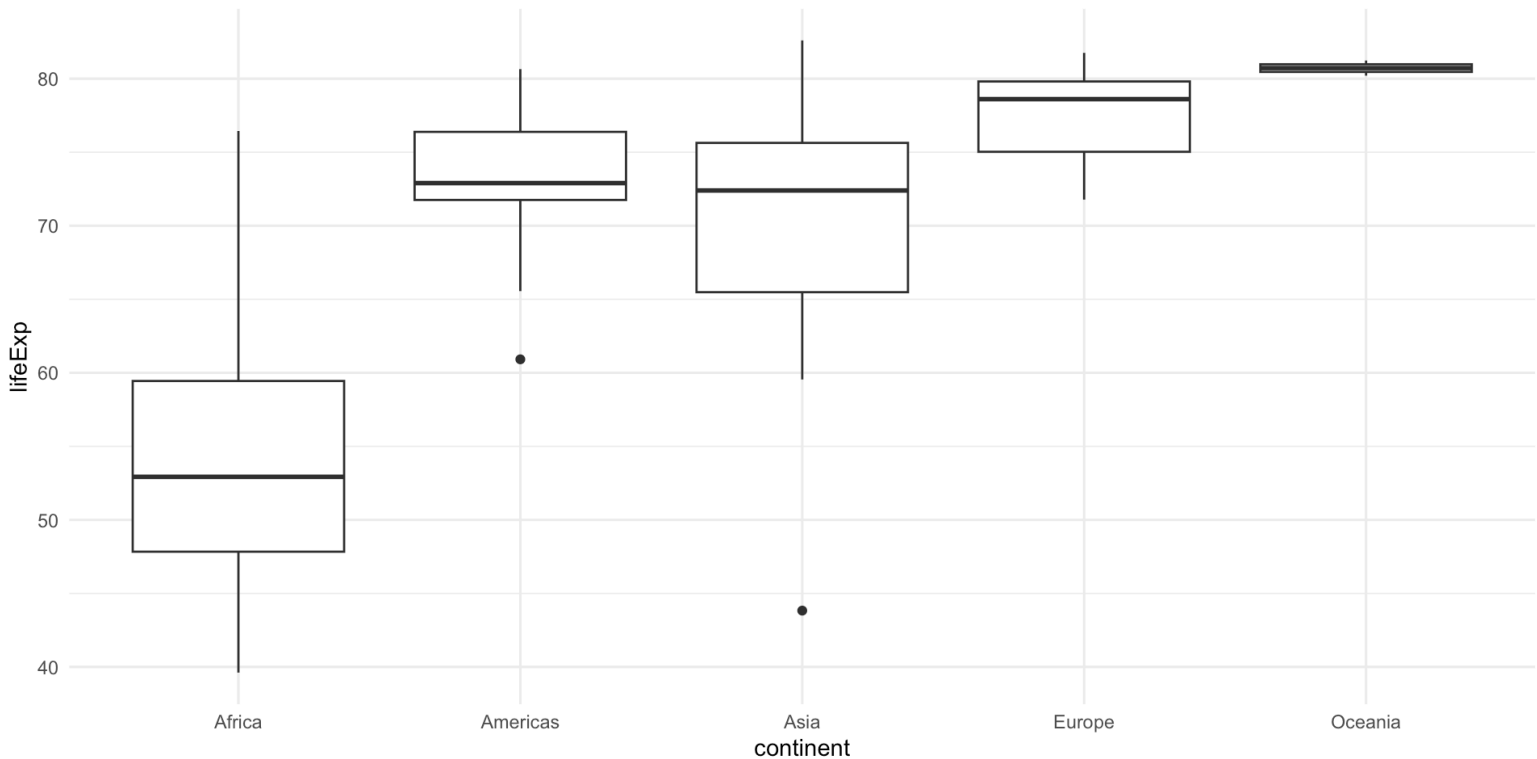
Code structure

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



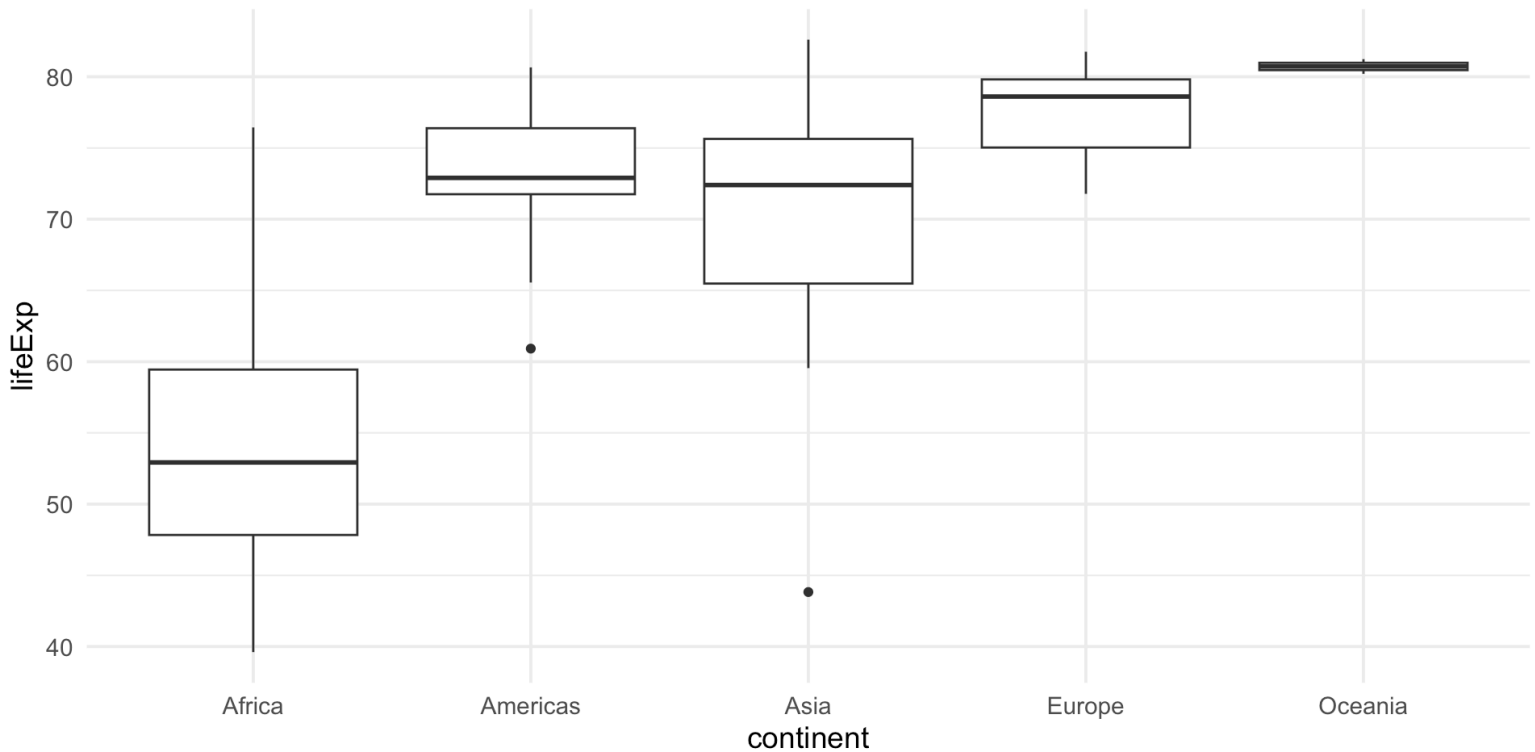
Code structure

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



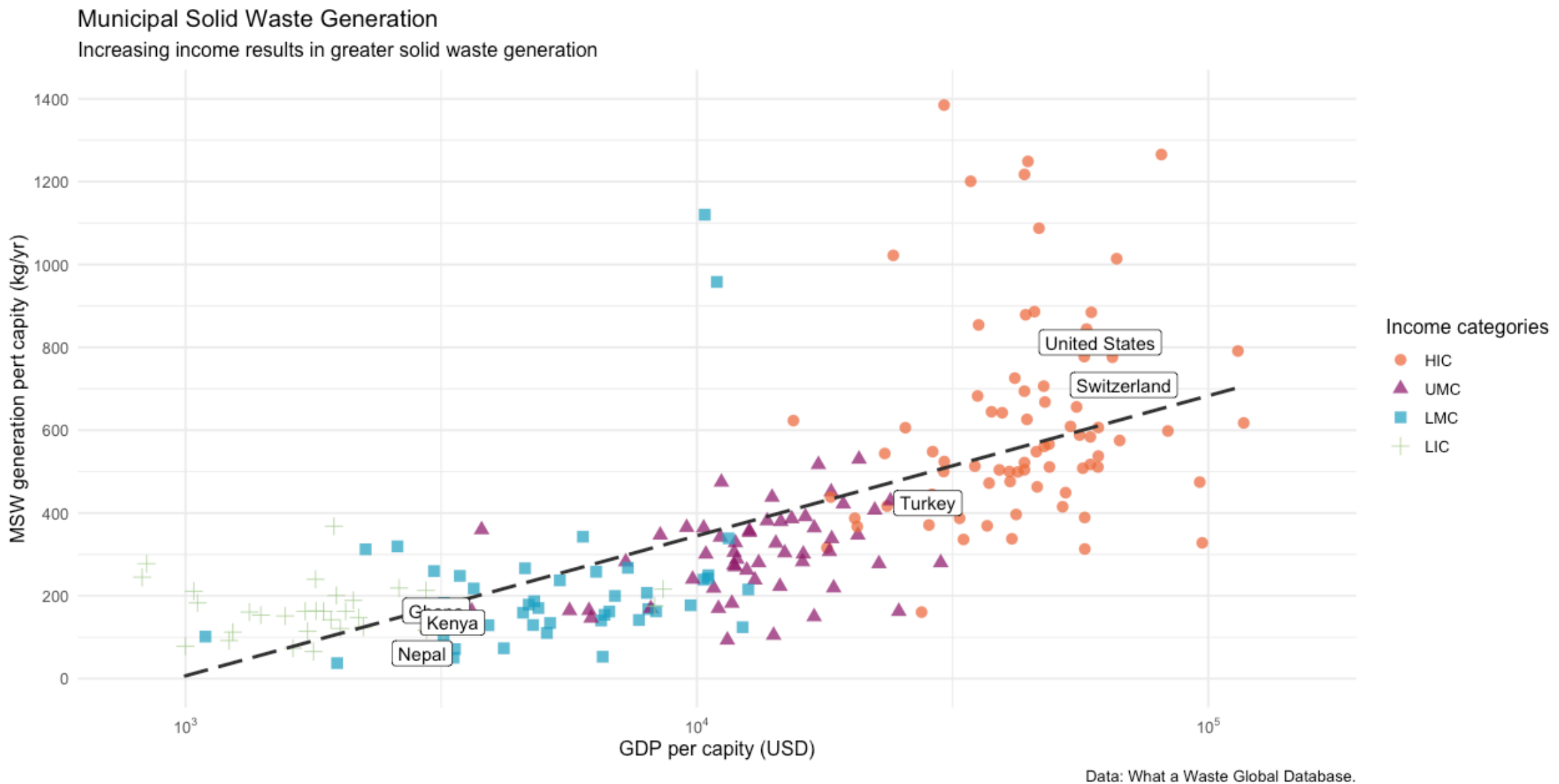
Code structure

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



Live Coding Exercise: Reproduce this plot

cven5999-ss25.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-data-visualiation

1. Head over to posit.cloud
2. Open the workspace for the course cven5999-ss25
3. Open “Projects”
4. Open the “course-materials” project
5. Follow along with me

Break

10:00

Photo by Blake Wisz

cven5999-ss25.github.io/website/

Visualising numerical data

Types of variables

numerical

discrete variables

- non-negative
- whole numbers
- e.g. number of students, roll of a dice

continuous variables

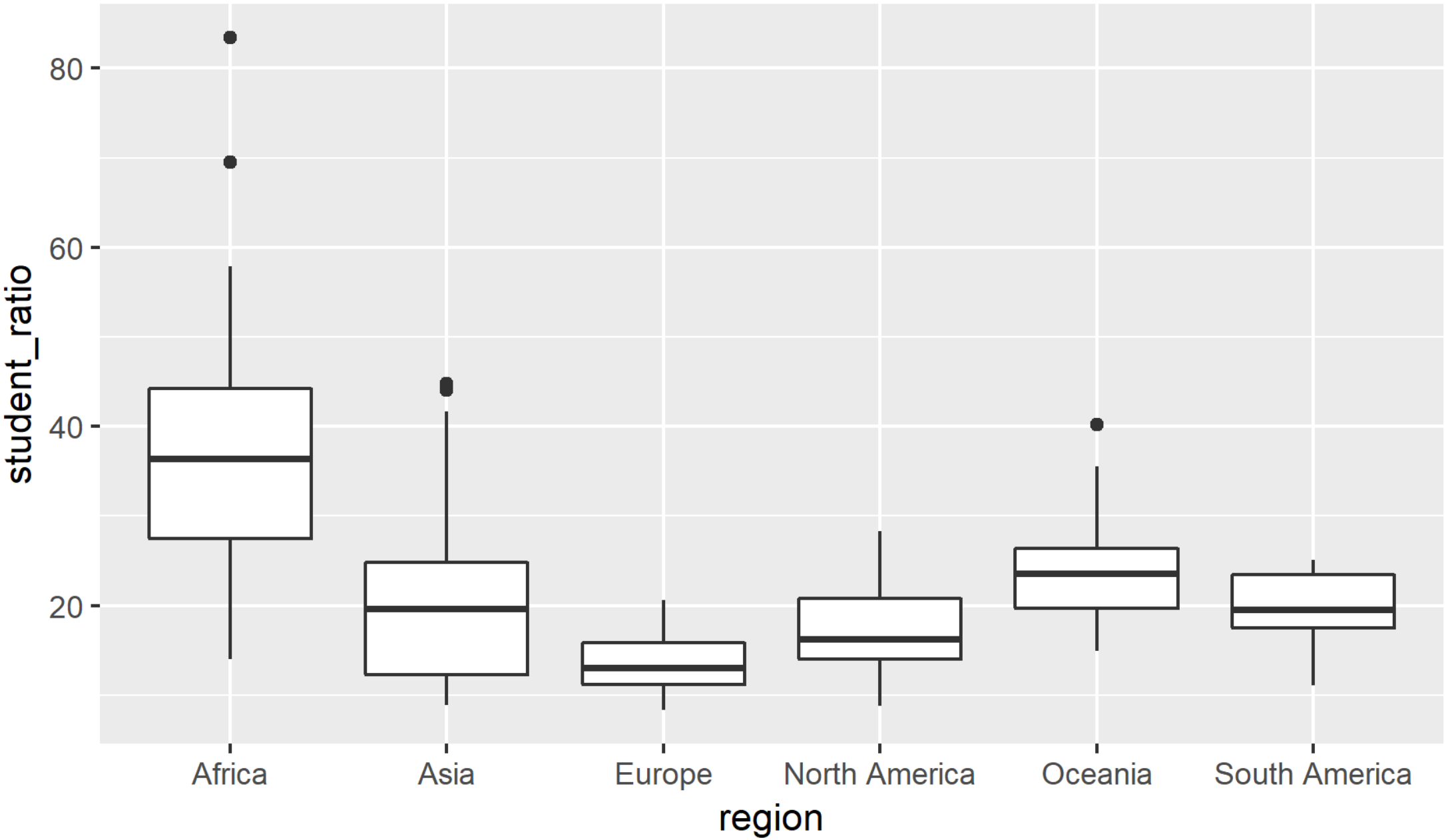
- infinite number of values
- also dates and times
- e.g. length, weight, size

non-numerical

categorical variables

- finite number of values
- distinct groups (e.g. EU countries, continents)
- **ordinal** if levels have natural ordering (e.g. week days, school grades)

The Evolution of a ggplot



cven5999-ss25.github.io/website/

Data: UNESCO Institute for Statistics
Visualization by Cédric Scherer

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

Why?

Because it will make your life easier!

via GIPHY

Editorial

The ASA Statement on *p*-Values: Context, Process, and Purpose ›

Ronald L. Wasserstein & Nicole A. Lazar

Published online: 9 Jun 2016 (Vol.70, No.2, 2016)

808032 Views

4465 CrossRef citations

2'259 Altmetric

Editorial

Moving to a World Beyond "$p < 0.05$" ›

Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar

Published online: 20 Mar 2019 (Vol.73, No.sup1, 2019)

439201 Views

2031 CrossRef citations

1'393 Altmetric

Article

Data Organization in Spreadsheets ›

Karl W. Broman & Kara H. Woo

Published online: 24 Apr 2018 (Vol.72, No.1, 2018)

374144 Views

78 CrossRef citations

1'628 Altmetric

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

Homework week 3

cven5999-ss25.github.io/website/

Identify a dataset for the capstone project

- A dataset from your own research
- A dataset from your work
- A dataset that you find interesting and is available as open data

Homework due dates

- All material on course website
- Homework assignment & learning reflection due: **2025-06-20**

Thanks! 

cven5999-ss25.github.io/website/

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/

Access slides as PDF on GitHub