

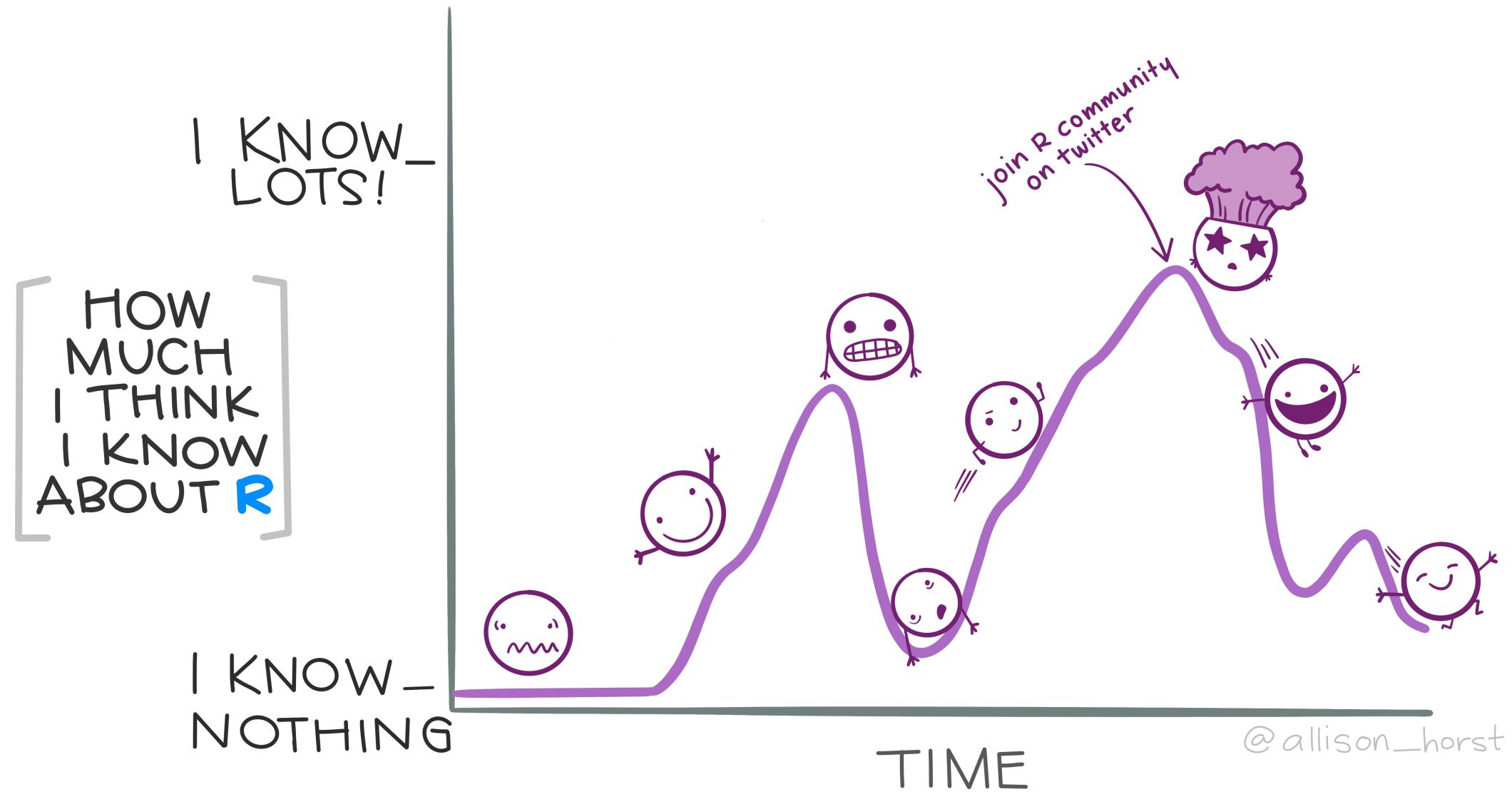
# Data Visualisation & Data Organisation in Spreadsheets

CVEN 5837 - Summer 2022

Lars Schöbitz

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

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



# Solving coding problems

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

# 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

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

# Give credit

from [r cookbook](#), where bp is your ggplot:

528 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

ew 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



1



but when I do something like this `bp + theme(legend.position`

# Give credit

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

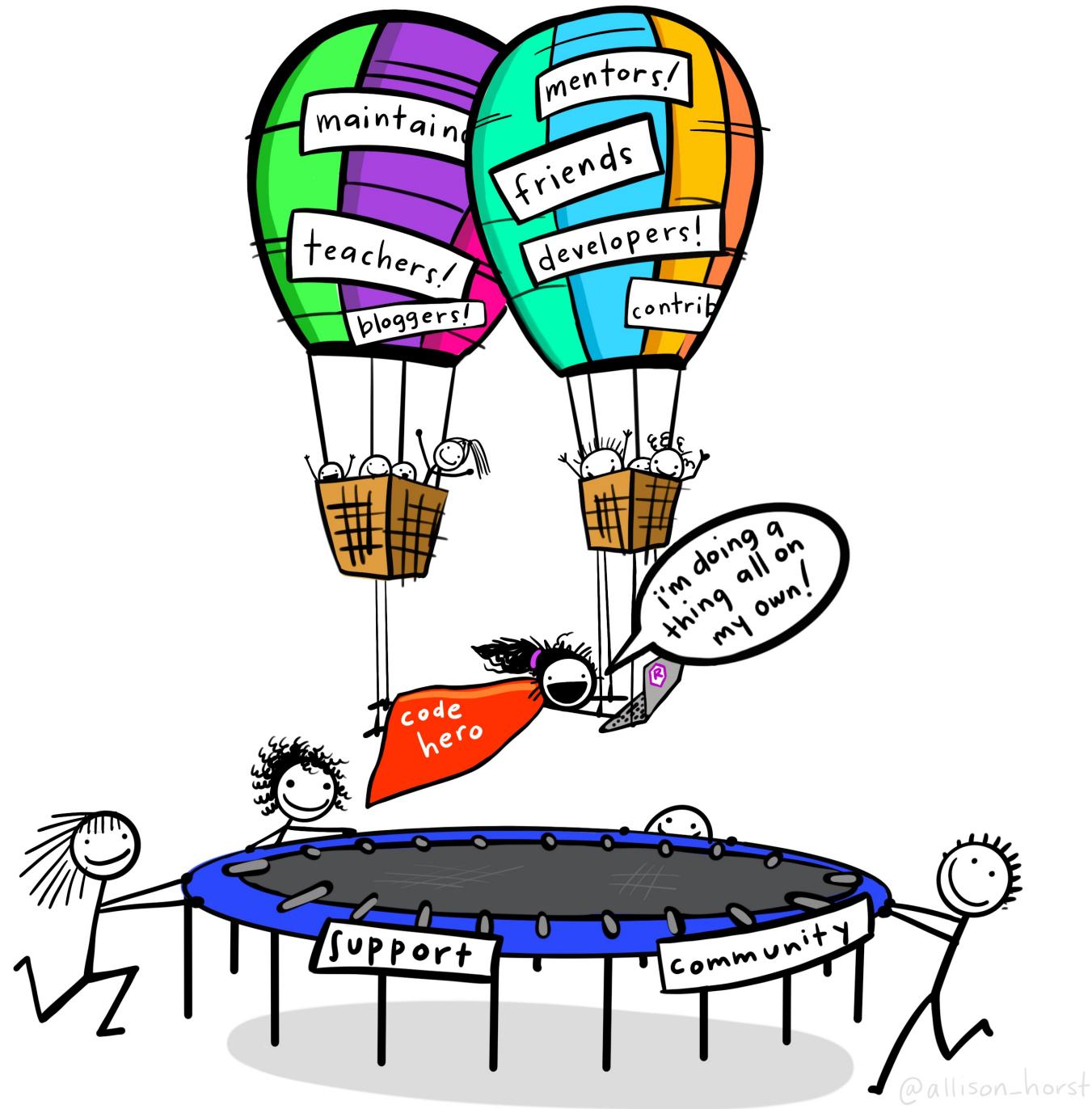
# Other sources for help

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



# Minimal reproducible example (reprex)

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



Artwork by @allison\_horst

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

# Learning Objectives (for this week)

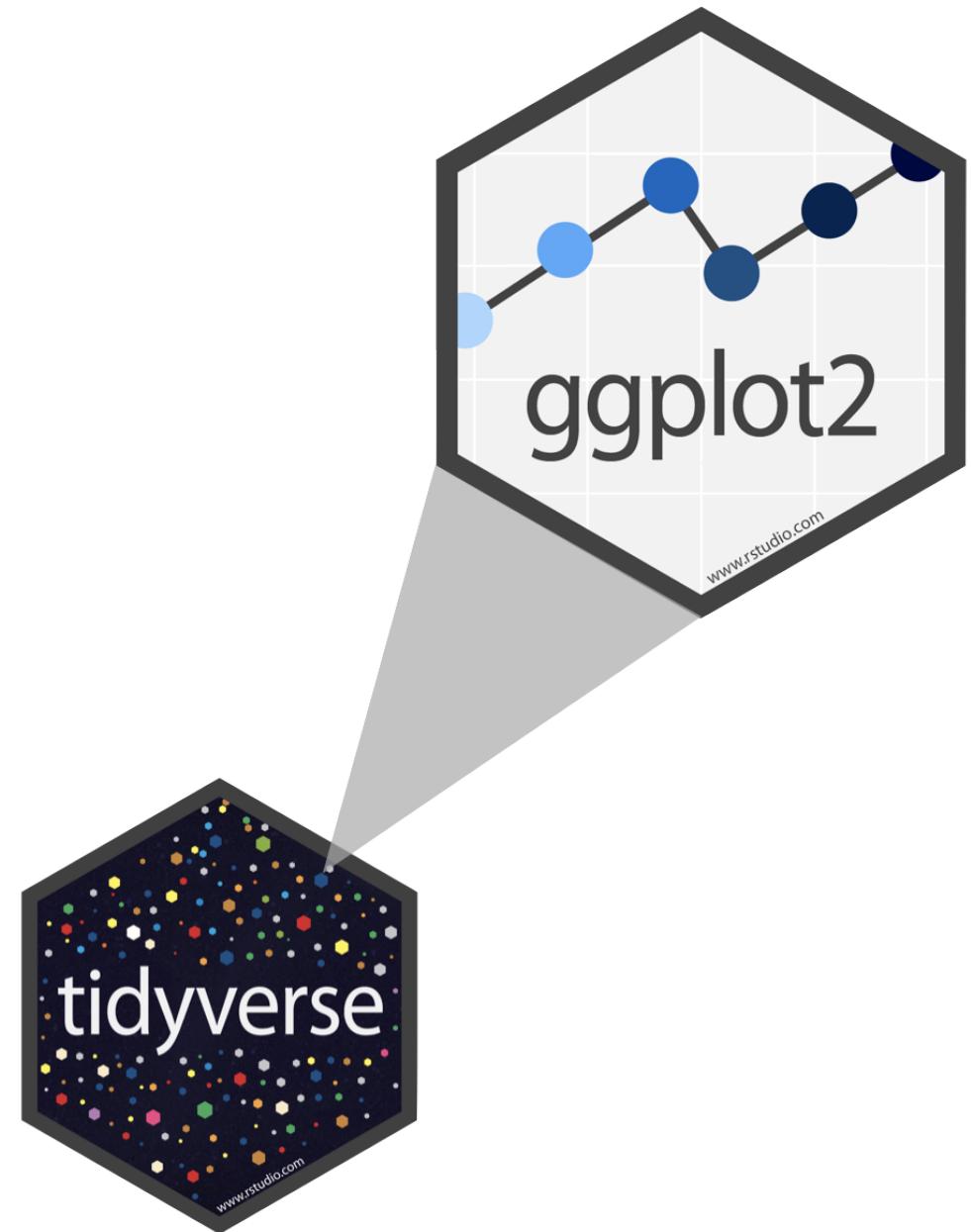
1. Learners can describe the four main aesthetic mappings that can be used to visualise data using the ggplot2 R Package
2. Learners can control the colour scaling applied to a plot using colour as an aesthetic mapping
3. Learners can compare three different geoms and their use case
4. Learners can apply 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`

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

- **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

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

# Code structure

```
1 ggplot()
```

# Code structure

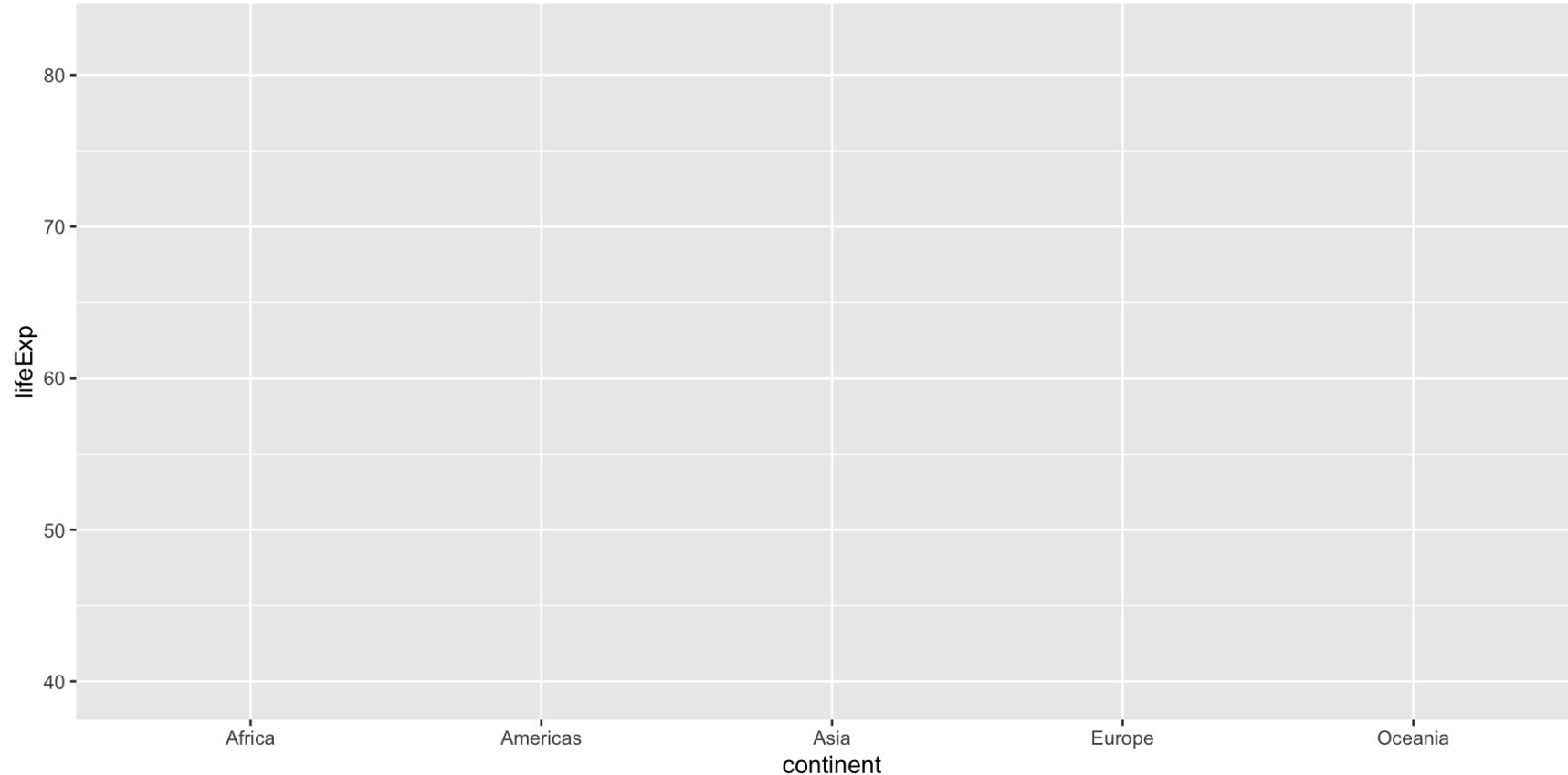
```
1 ggplot(data = gapminder_yr_2007)
```

# Code structure

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

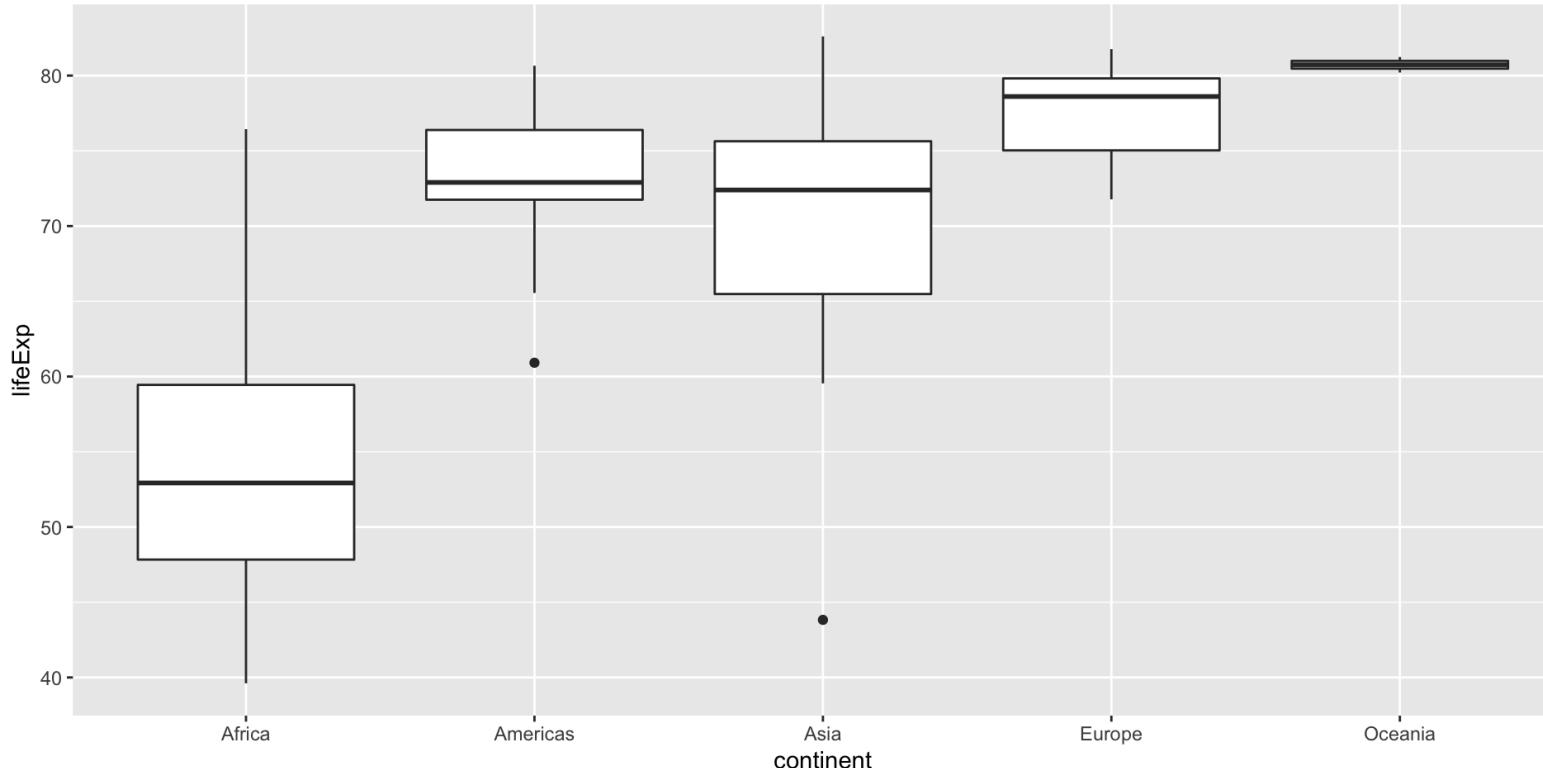
# Code structure

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



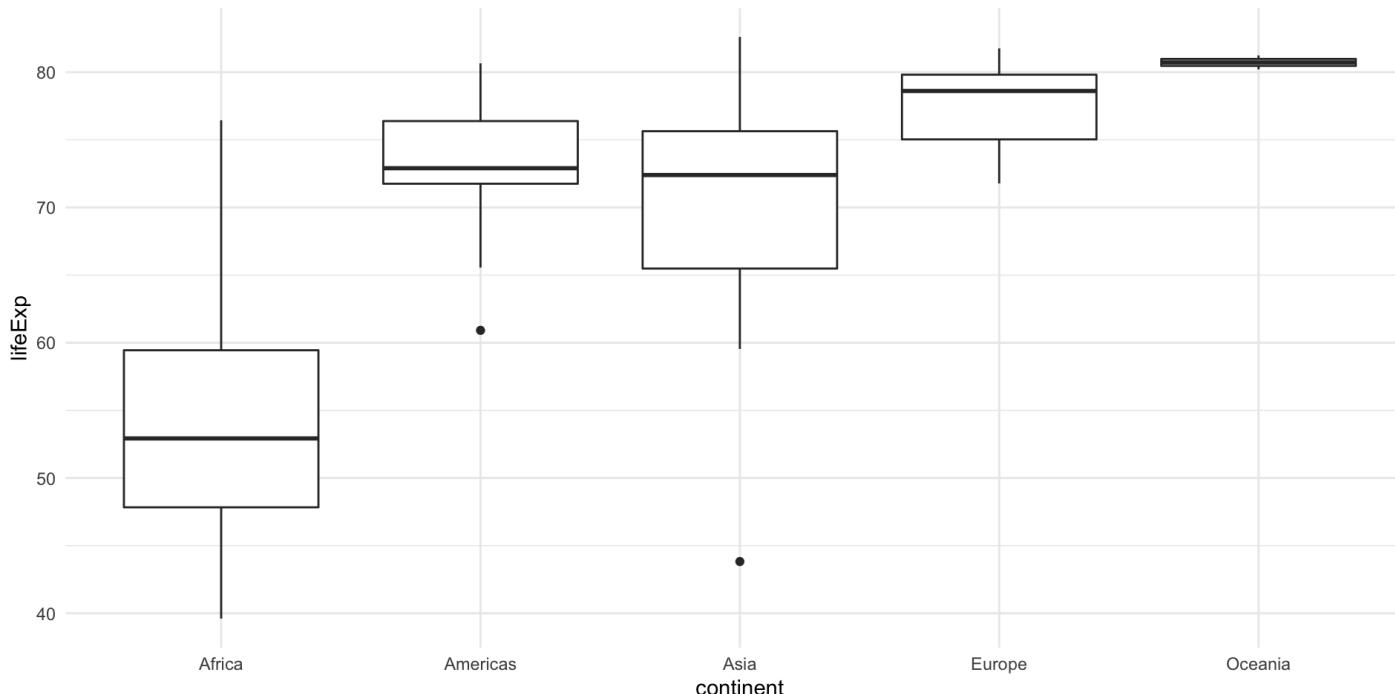
# Code structure

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



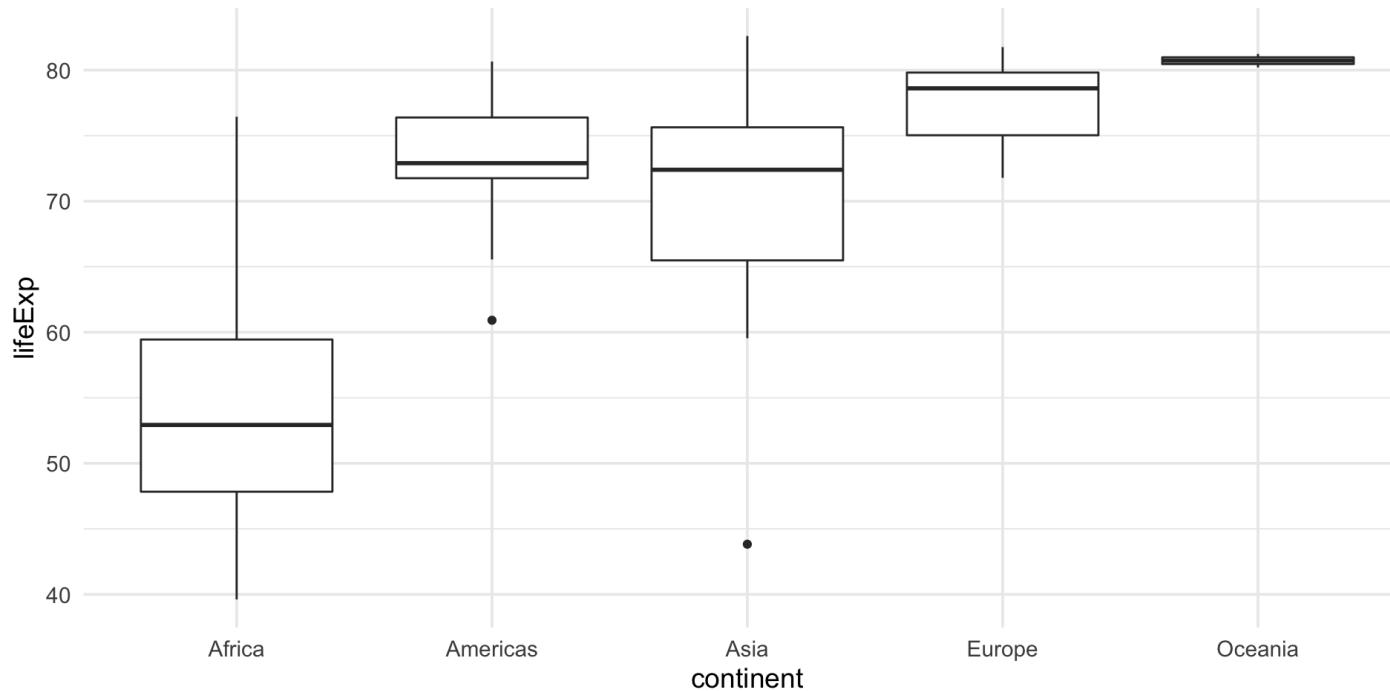
# Code structure

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



# Code structure

```
1 ggplot(data = gapminder_yr_2007,  
2         mapping = aes(x = continent,  
3                           y = lifeExp)) +  
4         geom_boxplot() +  
5         theme_minimal(base_size = 14)
```



# Live Coding Exercise: Reproduce this plot

```
{background-image="img/lec-  
02/plot-waste-generation-gdp.png"  
background-size="contain"} ^
```

# live-02a-data-visualiation - RStudio Cloud Setup

1. Head over to [rstudio.cloud](https://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

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

# Break



10 : 00

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

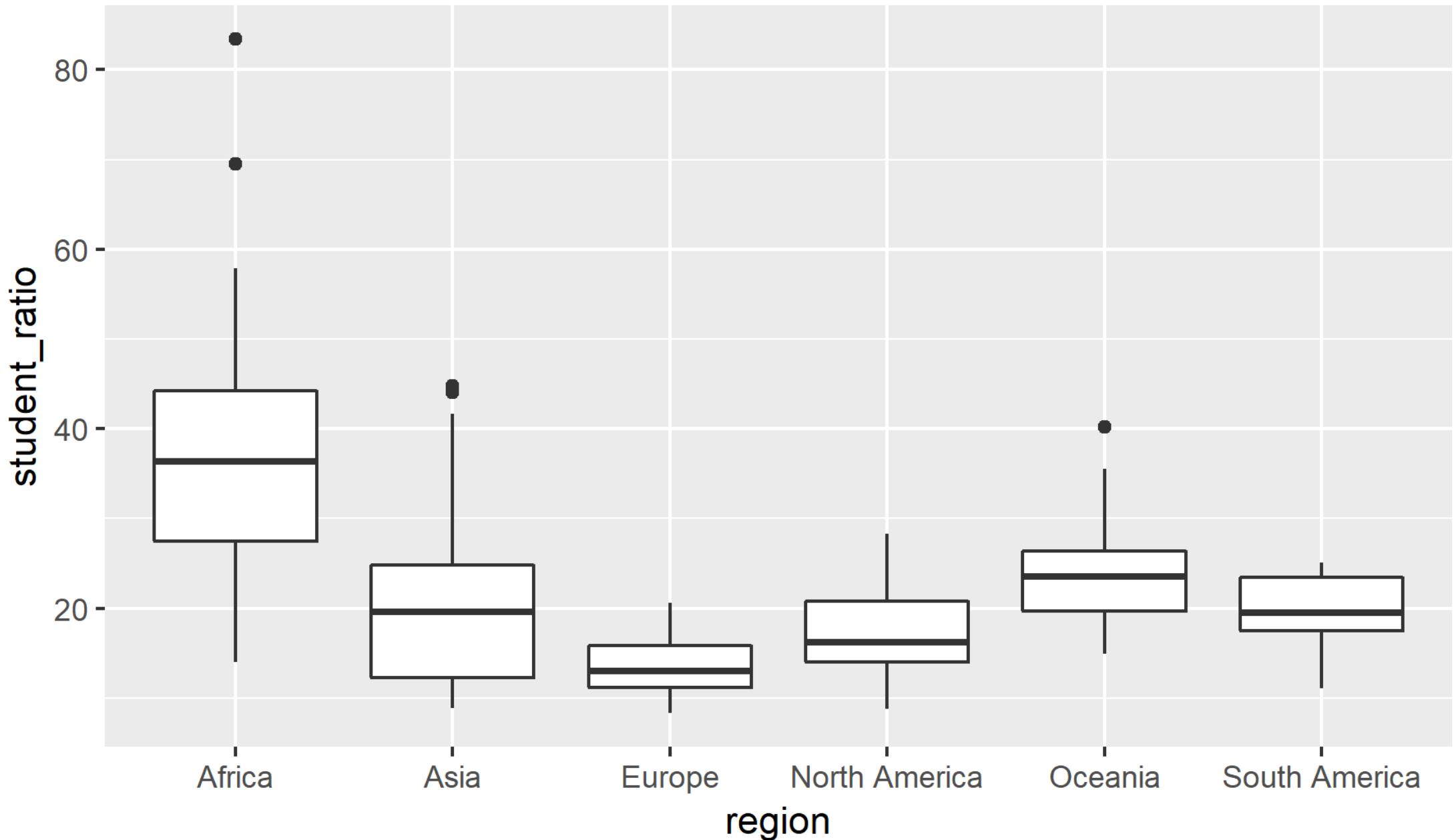
Photo by [Blake Wisz](#)

# Visualising numerical data

# Types of variables

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

# The Evolution of a ggplot



# data-to-viz.com

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

# Data Collection Tools

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

# Data Collection Tools

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

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

# Survey tools

Commonly used in the Global Engineering and Development sector

- Kobo Toolbox
- mWater
- OpenDataKit

# Data Organisation in Spreadsheets



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

# 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](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

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

# Why?

Because it will make your life easier!

Latest	Open access	Most read	Most cited
<a href="#">The ASA Statement on <i>p</i>-Values: Context, Process, and Purpose &gt;</a>			588993 Views
Ronald L. Wasserstein et al.			
Editorial   Published online: 9 Jun 2016			
			✓
<a href="#">Moving to a World Beyond “<i>p</i> &lt; 0.05” &gt;</a>			302054 Views
Ronald L. Wasserstein et al.			
Editorial   Published online: 20 Mar 2019			
			⌚
<a href="#">Data Organization in Spreadsheets &gt;</a>			272345 Views
Karl W. Broman et al.			
Article   Published online: 24 Apr 2018			
			⌚
<a href="#">Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don’t Expect Replication &gt;</a>			48111 Views
Valentin Amrhein et al.			
Article   Published online: 20 Mar 2019			
			⌚

# License? CCO (!)

☰ 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](https://kbroman.org/Talk_DataOrg/dataorg.pdf)

PDF of slides with notes: [https://kbroman.org/Talk\\_DataOrg/dataorg\\_notes.pdf](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.



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

# Pair Programming Exercise

# 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 - RStudio Cloud Setup

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

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

# 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

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

# Homework due dates

- All material on [course website](#)
- Homework assignment due: Friday, 15th July
- Learning reflection due: Monday, 18th July

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

Thanks! 🌻

Slides created via revealjs and Quarto:

<https://quarto.org/docs/presentations/revealjs/> Access slides as PDF on GitHub

All material is licensed under [Creative Commons Attribution Share Alike 4.0 International.](#)