STAT 215A Fall 2019
Week 3

Tiffany Tang

9/13/19

Lab 1 Check-In

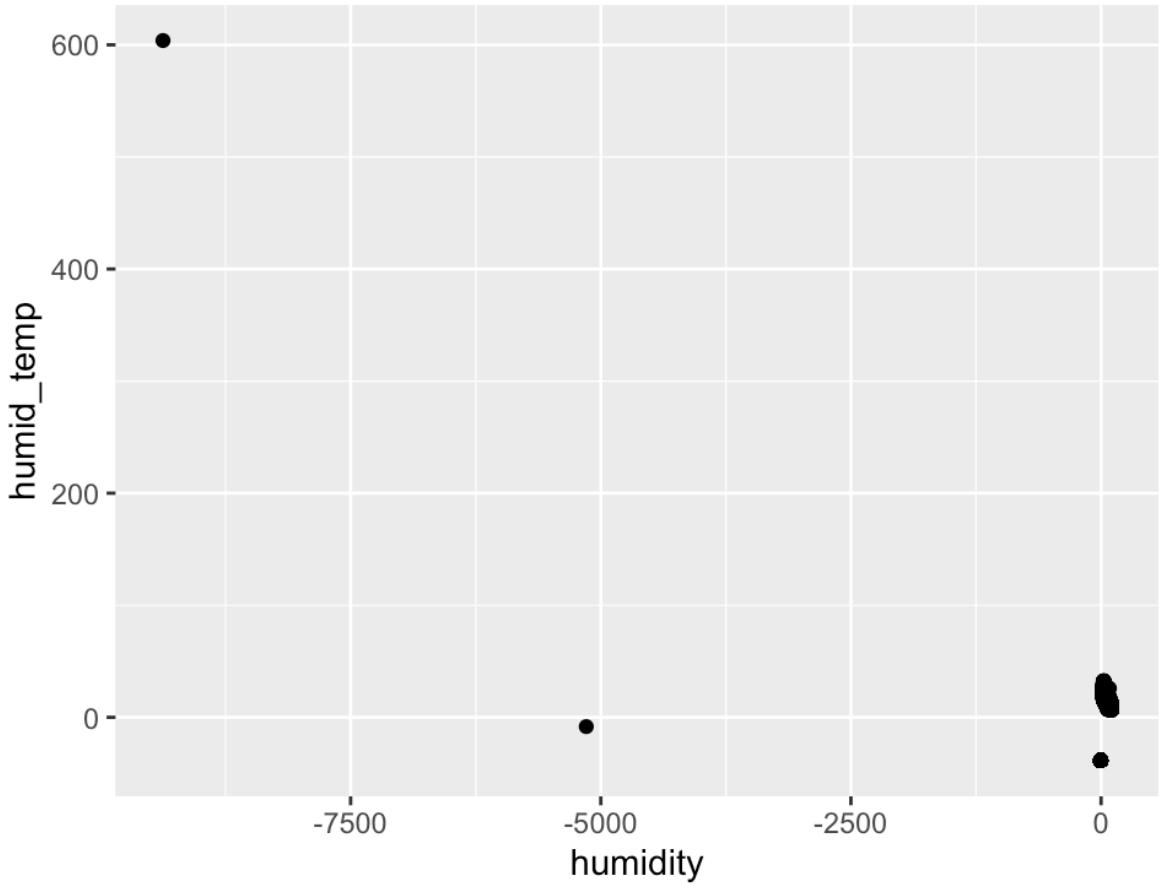
- How is it going?
- Having fun?
- Challenges?
- Questions?
- Remember it's due Thursday, Sept 19 at 11:59pm!!
- Berkeley SCF Resources: https://github.com/berkeley-scf

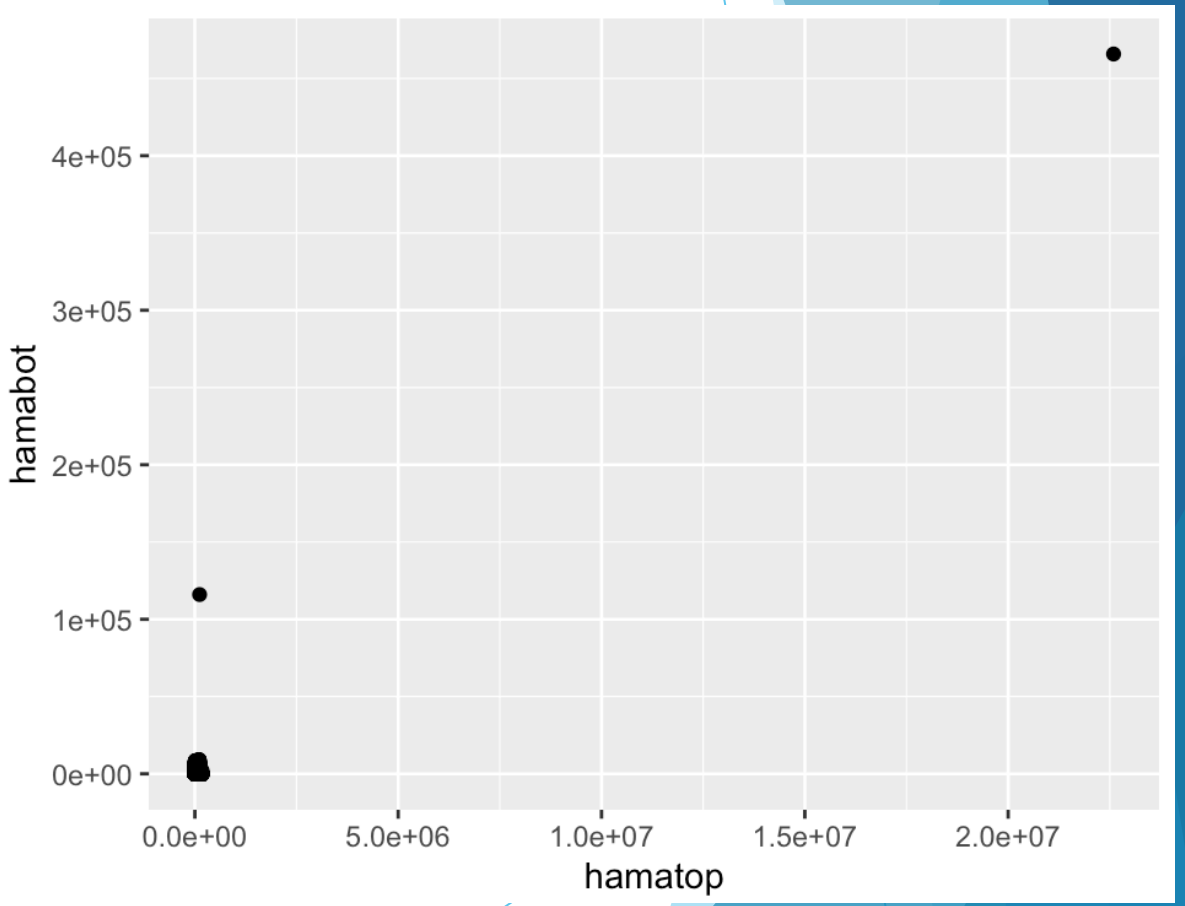
Lab 1 Check-In and Clarifications

Some thoughts if you're stuck

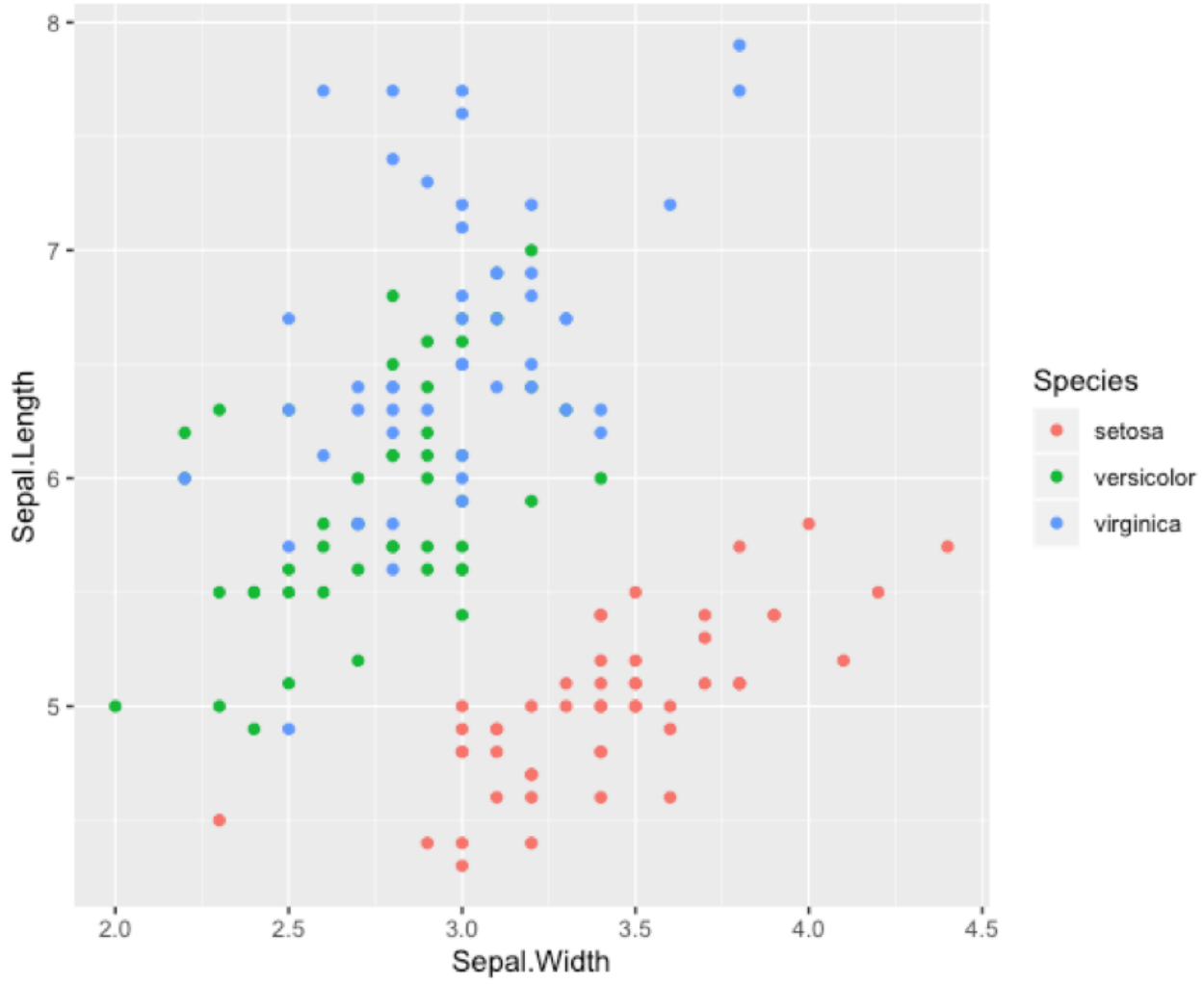
- Use your domain knowledge and curiosity to come up with questions you may want to answer
- Look at smaller parts of the data
  - Zoom in on a specific day or time of day

Lab 1 Check-In and Clarifications



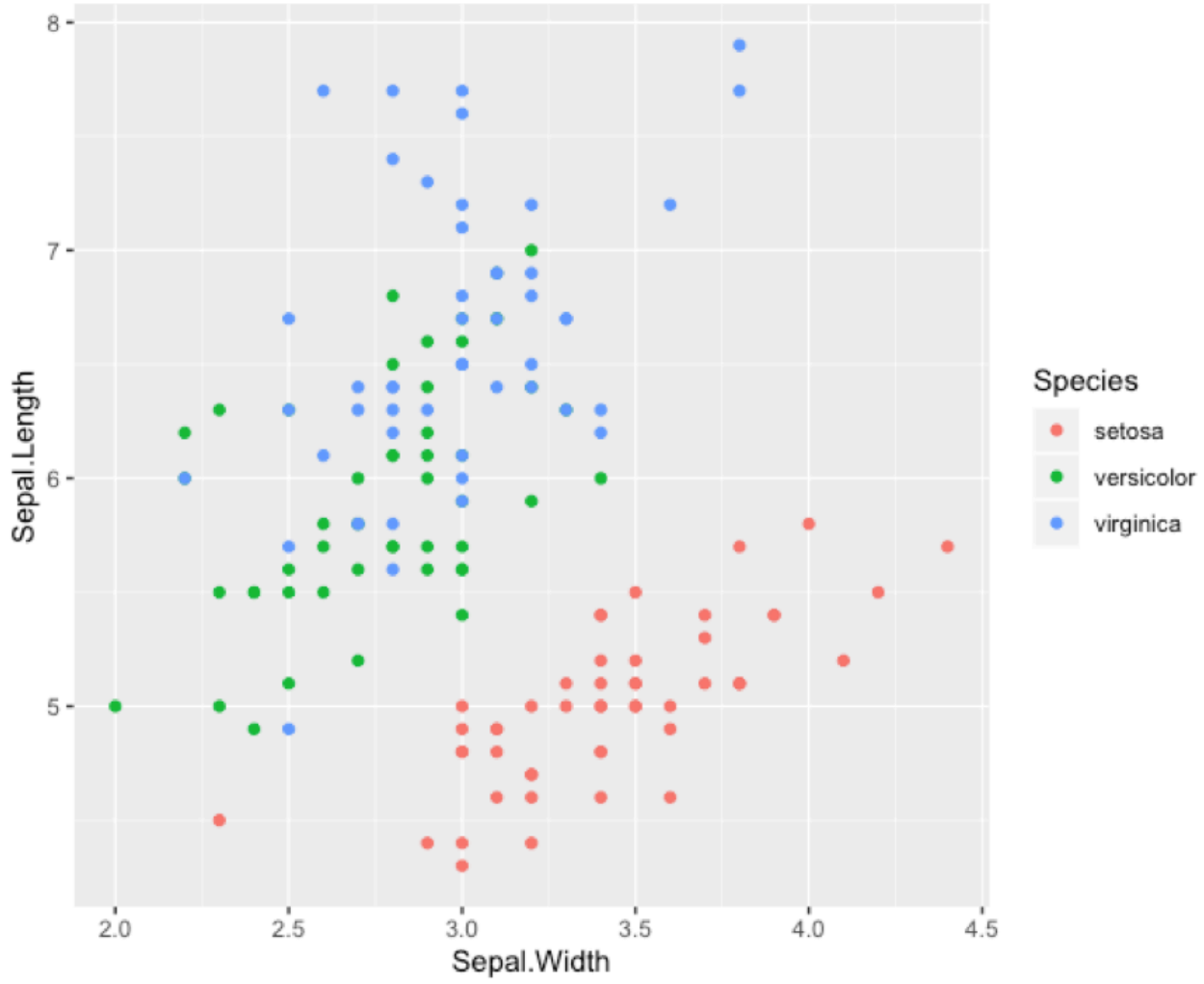


Motivation for Today



(Selfish) Motivation for Today

As your GSI, it can become monotonous to look at 100+ plots with the same gridded gray ggplot background and the same default ggplot color scheme… please don't make me go through that



We’re going to fix this problem

- Built-in and custom `ggplot` themes
- Color schemes
- Heatmaps with `superheat`
- `GGally` pair plots
- Ridge Density Plots
- Interactive plots
- Last part of class: PCA

Quick improvements to the classic ggplot theme

- Recall in the gapminder lab last week, we had defined this `theme_nice` in utils.R

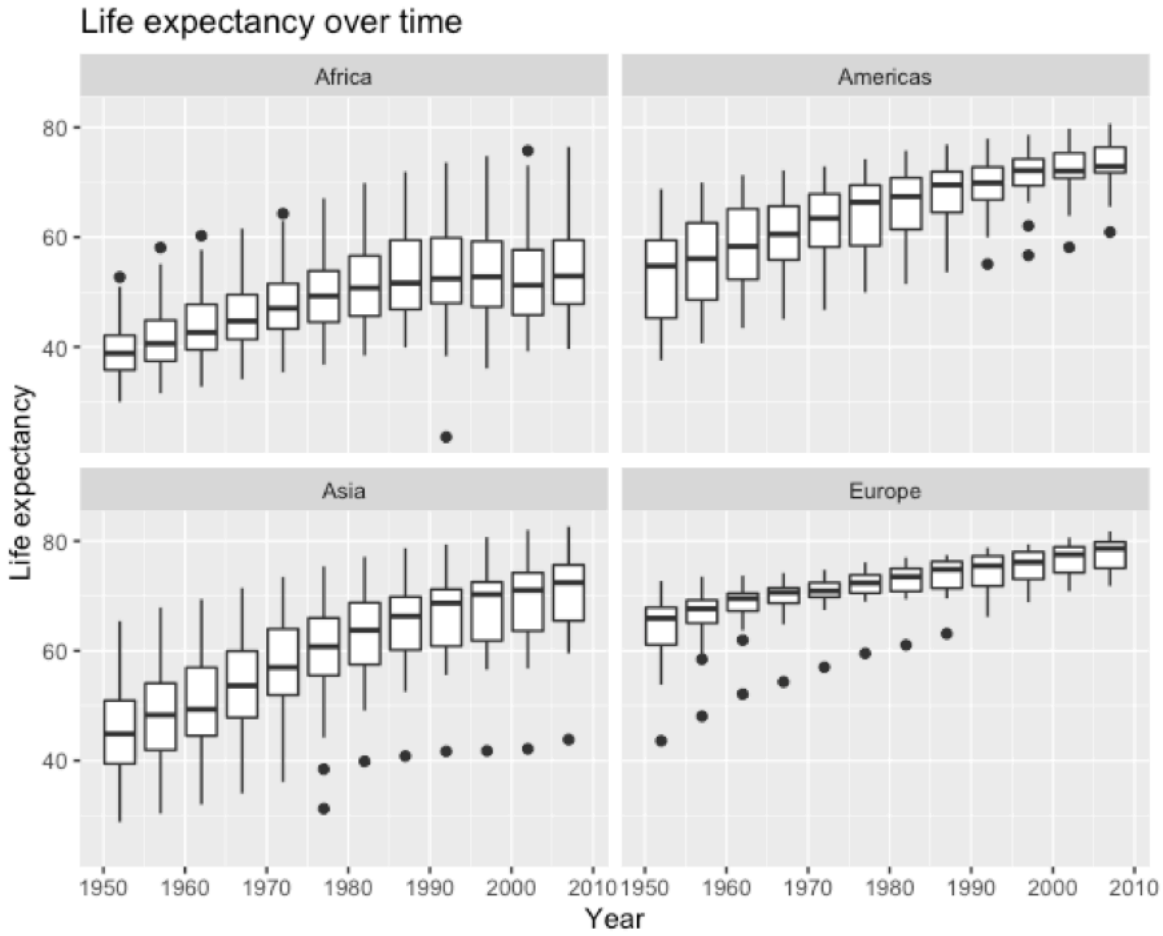
```
> theme_nice <- theme_classic() + theme(axis.line.y = element_blank())
```

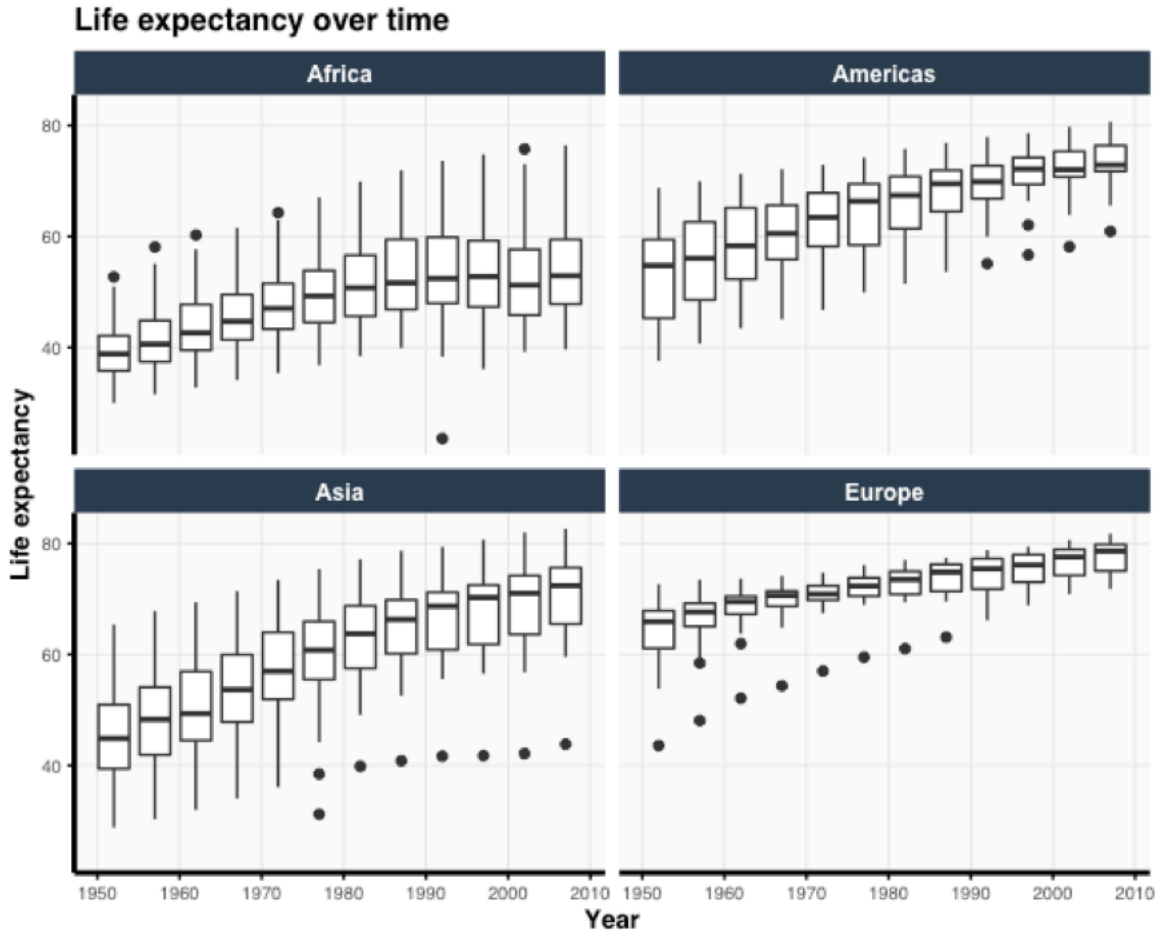
- Then to use this modified theme, we simply ran something like

```
> ggplot(gapminder %>% filter(continent != "Oceania")) +
+   facet_wrap(~continent) +
+   geom_boxplot(aes(x = year, y = life_exp, group = year), fill = "grey90") +
+   theme_nice
```

- Built-in ggplot themes: https://ggplot2.tidyverse.org/reference/ggtheme.html
- Or simply google “custom ggplot themes”

Custom ggplot themes with theme()

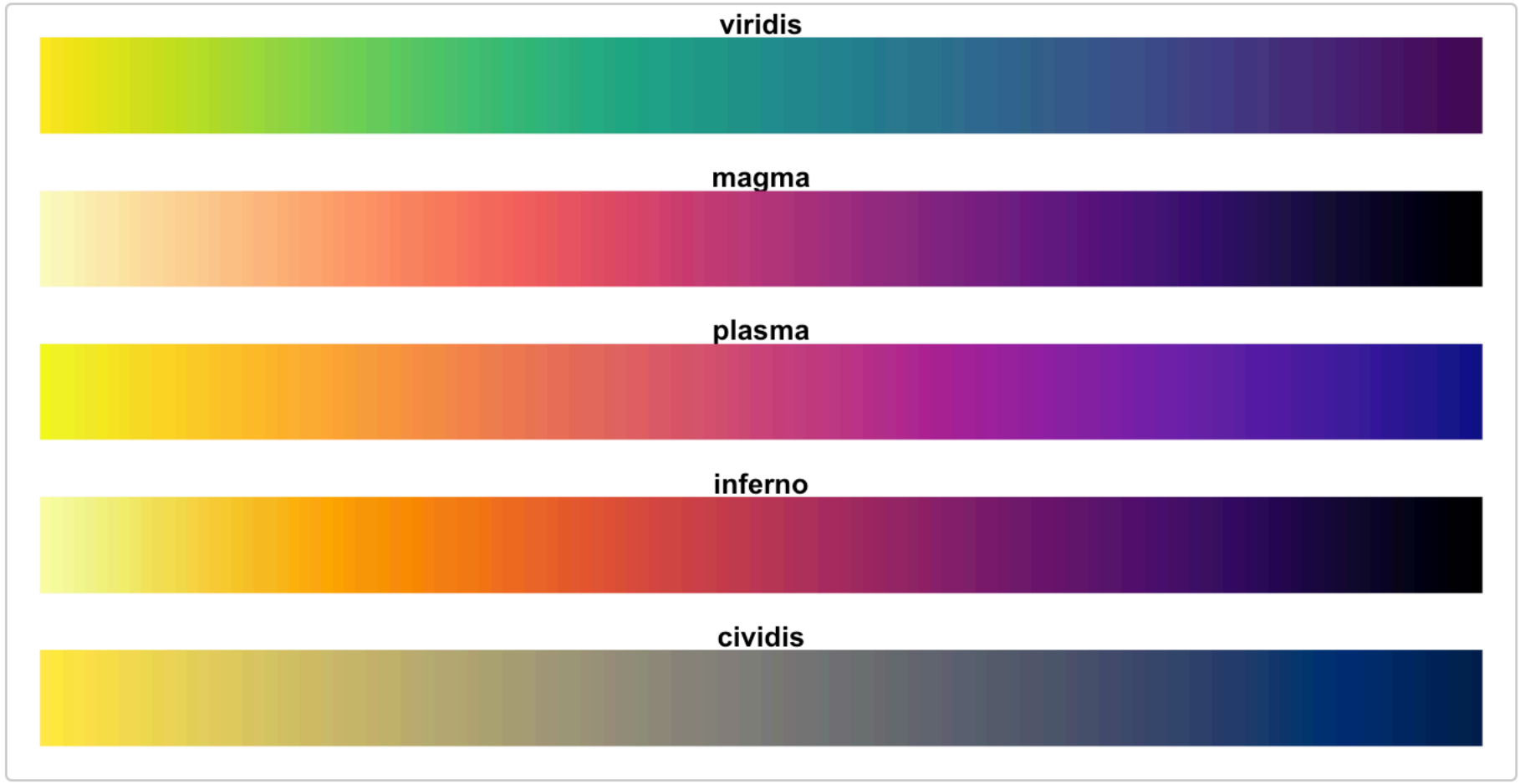




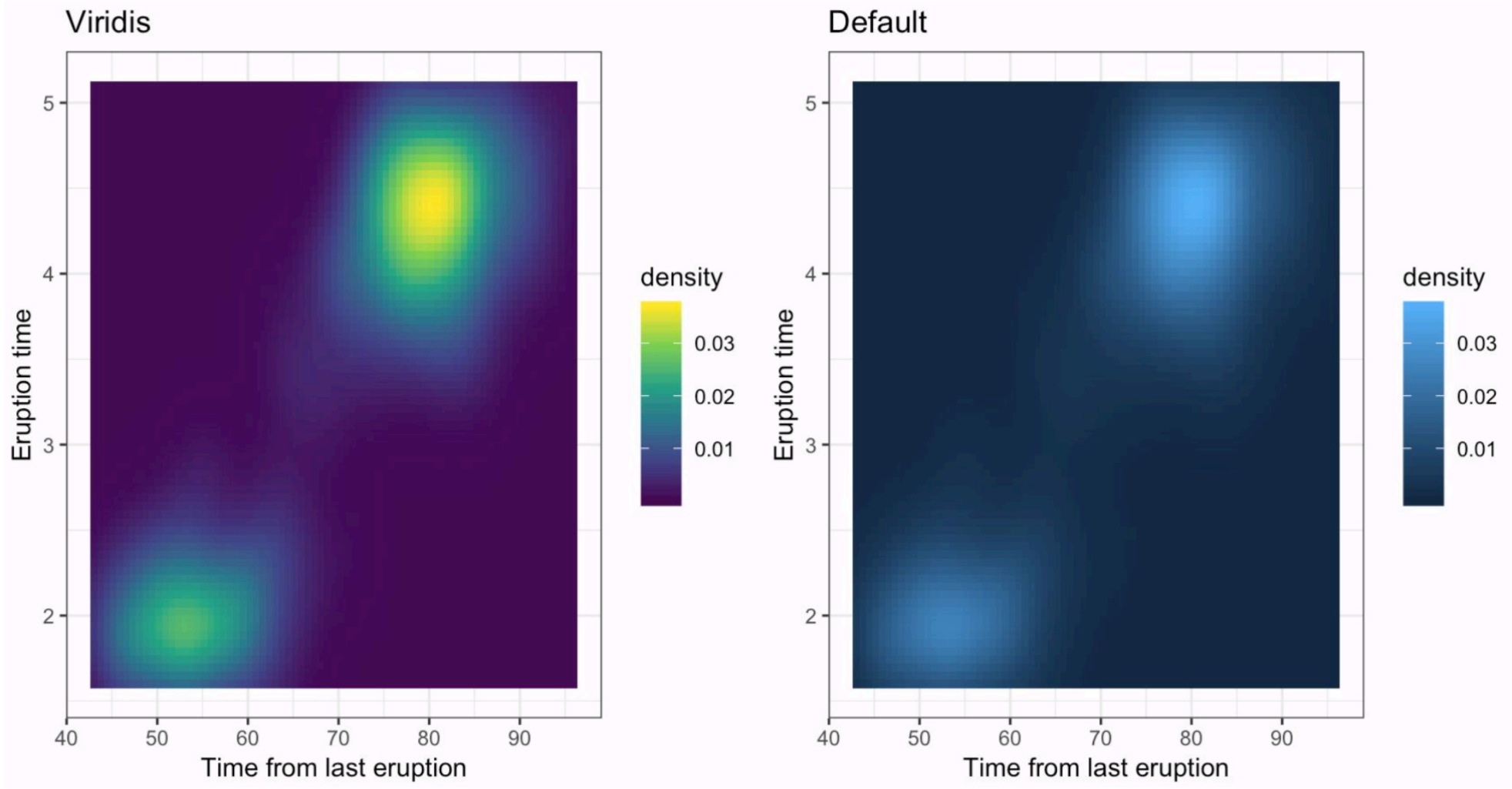
Color Schemes

- Default color scheme in base R or ggplot is not always the best choice
- Think about what you are trying to convey in the plot
- Color choices can affect the way we perceive the plot
- Two helpful websites
  - https://coolors.co/app
  - http://colorbrewer2.org/

Viridis Color Scheme



Viridis Color Scheme



Viridis Color Scheme

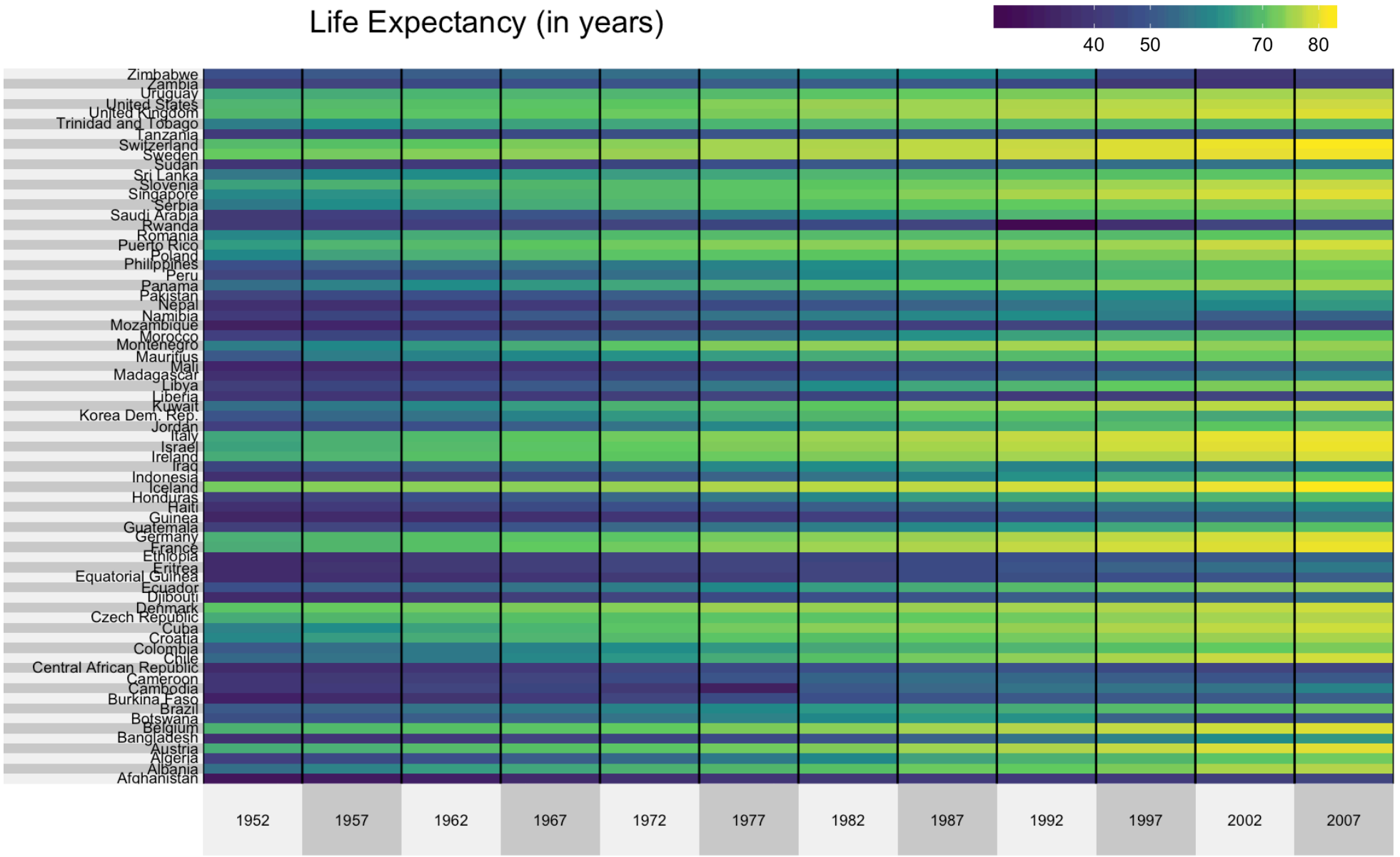
- Makes pretty plots!
- Perceptually uniform colors (meaning changes in the data should be accurately decoded by our brains)
  - Another colormap with this quality is **RColorBrewer**
- Perceived by most common forms of color blindness

Beyond the world of ggplot…

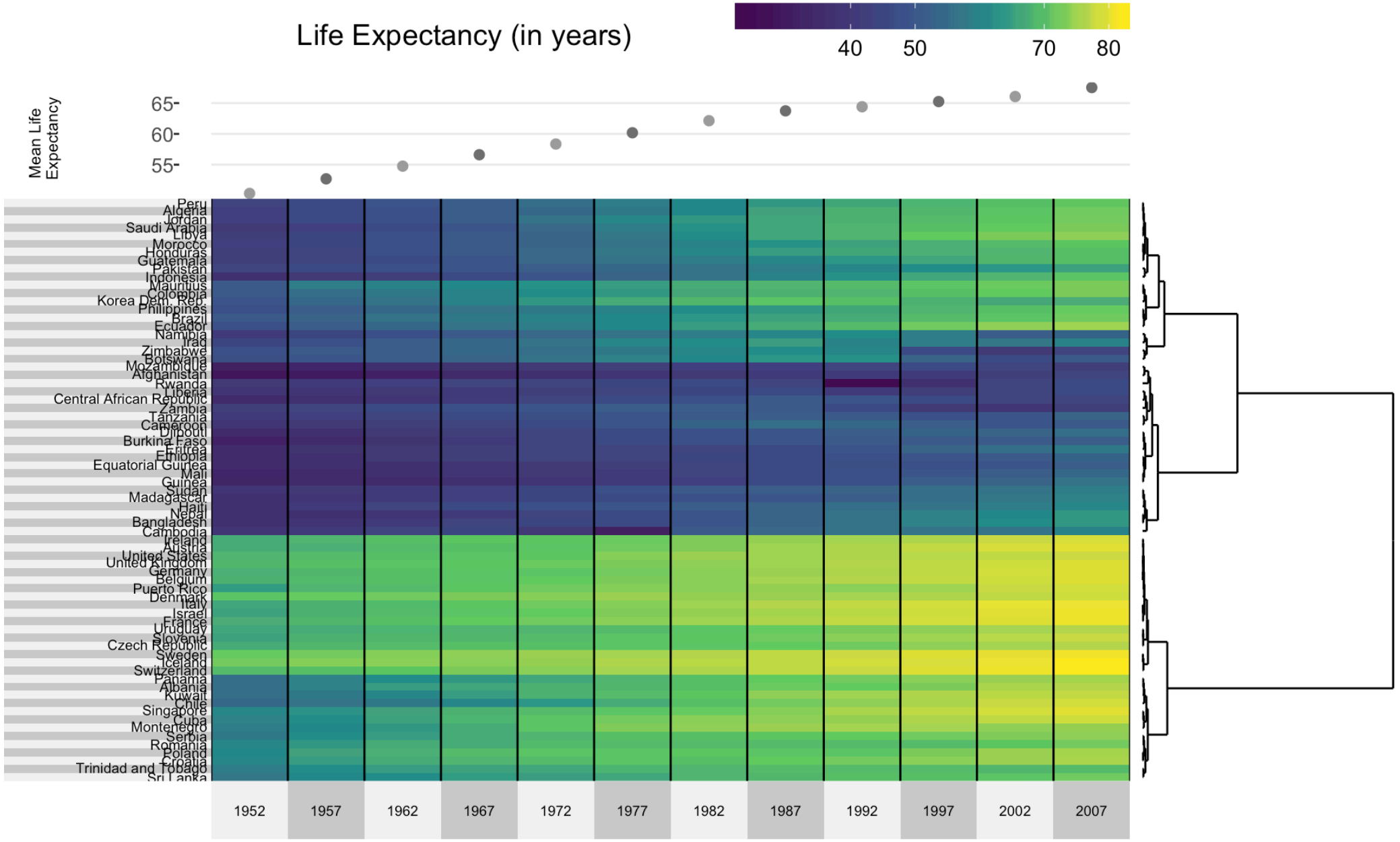
Heatmaps with superheat

```
install.packages("devtools")
devtools::install_github("rlbarter/superheat")
library(superheat)
```

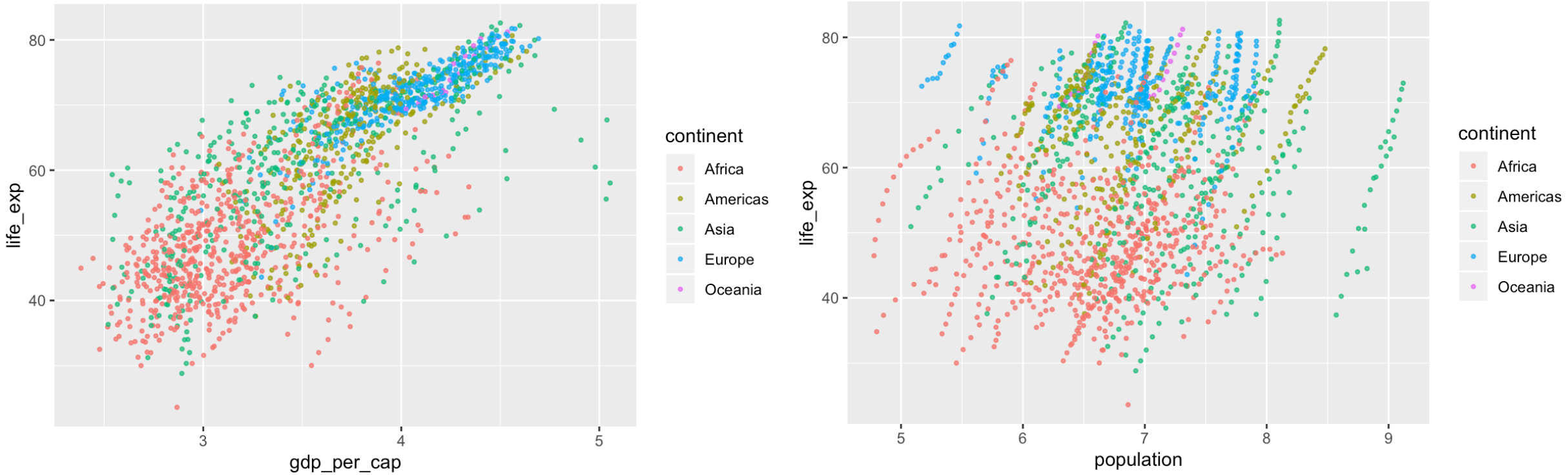
Heatmaps with superheat



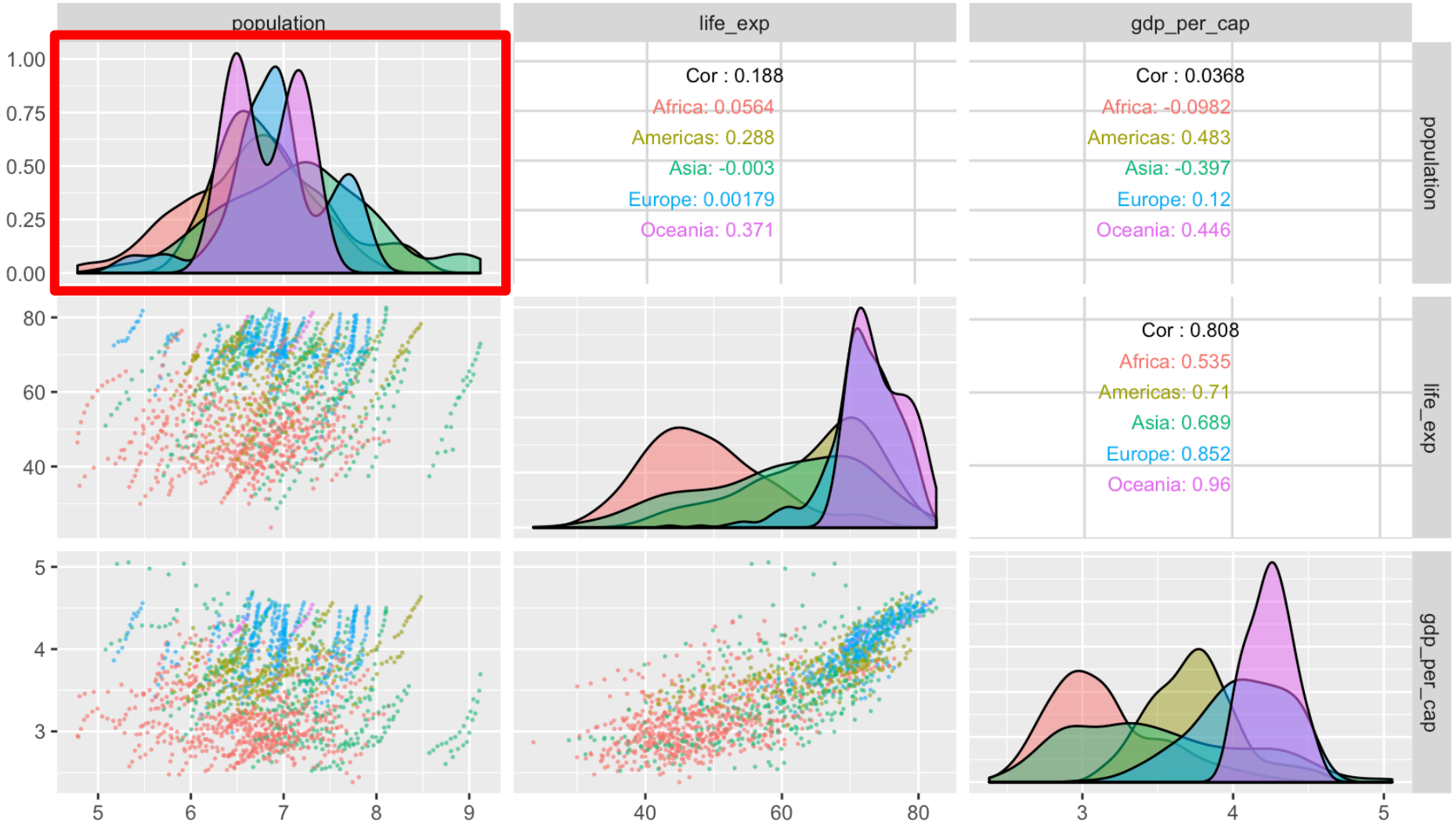
Heatmaps+ with superheat



Pair Plots

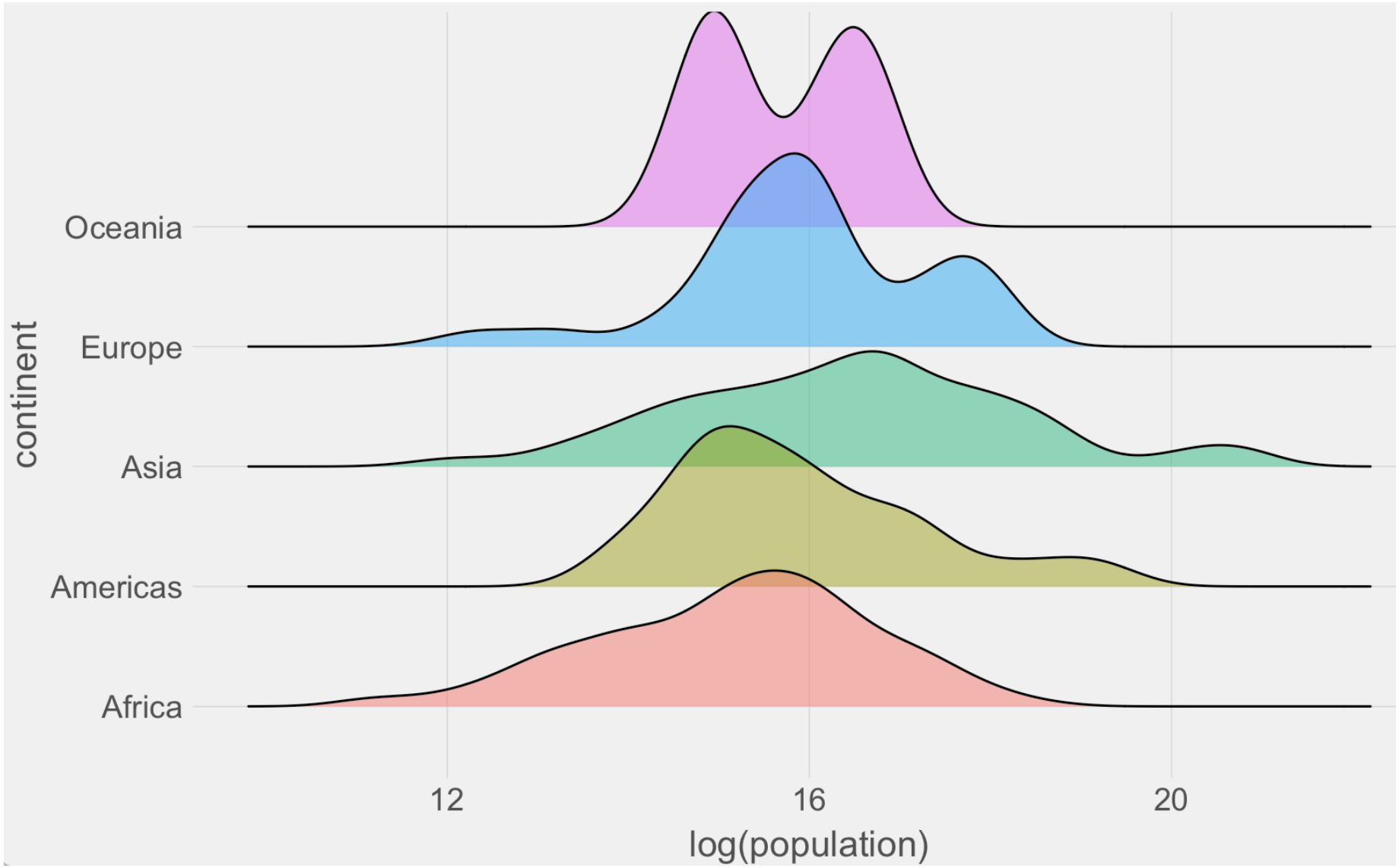

life_exp
80
60
40
3
4
5
gdp_per_cap
continent
Africa
Americas
Asia
Europe
Oceania
life_exp
80
60
40
5
6
7
8
9
population
continent
Africa
Americas
Asia
Europe
Oceania

Pair Plots with GGally::ggpairs



- A word of caution: be wary of over-plotting; consider subsampling points, limiting the number of variables in pair plot, etc.

ggridges: another way of viewing multiple density plots

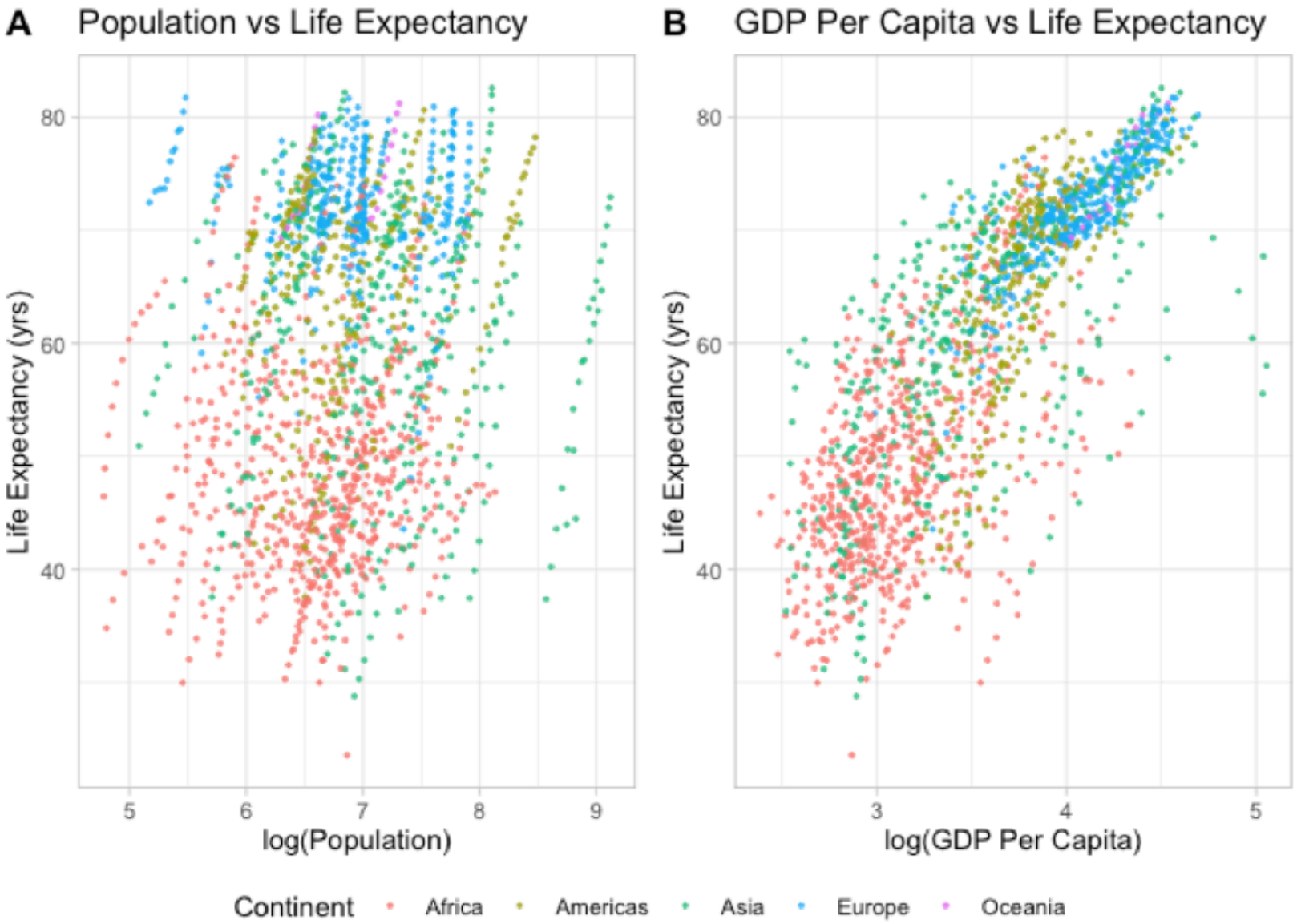


Creating sub-plots

- Two useful functions:
  - gridExtra::grid.arrange()
  - ggpubr::ggarrange
- Can easily set a common legend and subplot labels with ggarrange()
- grid.arrange() is better for fancier “non-matrix” arrangements

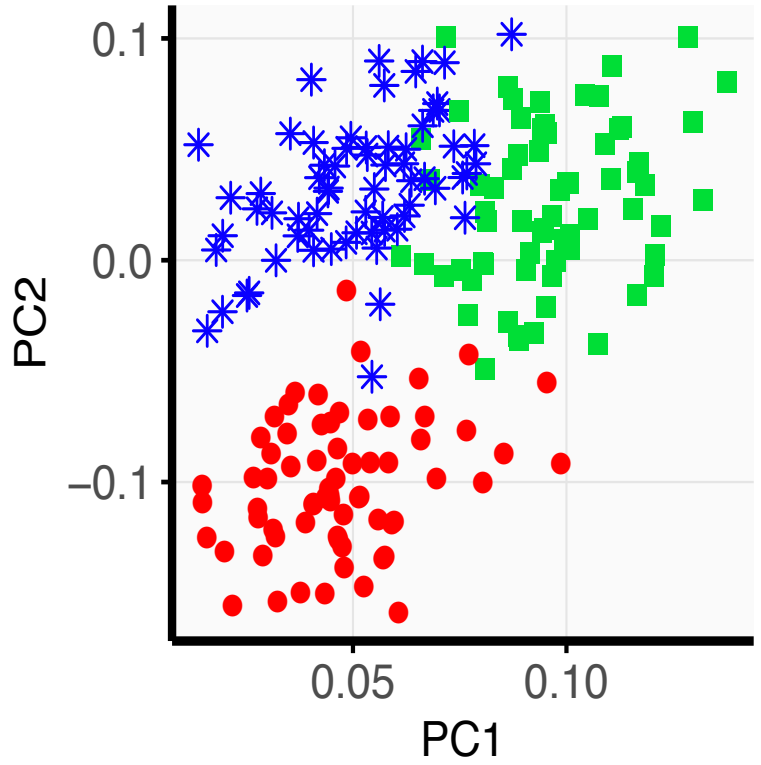
gridExtra::grid.arrange

ggpubr::ggarrange



Interactive Plots

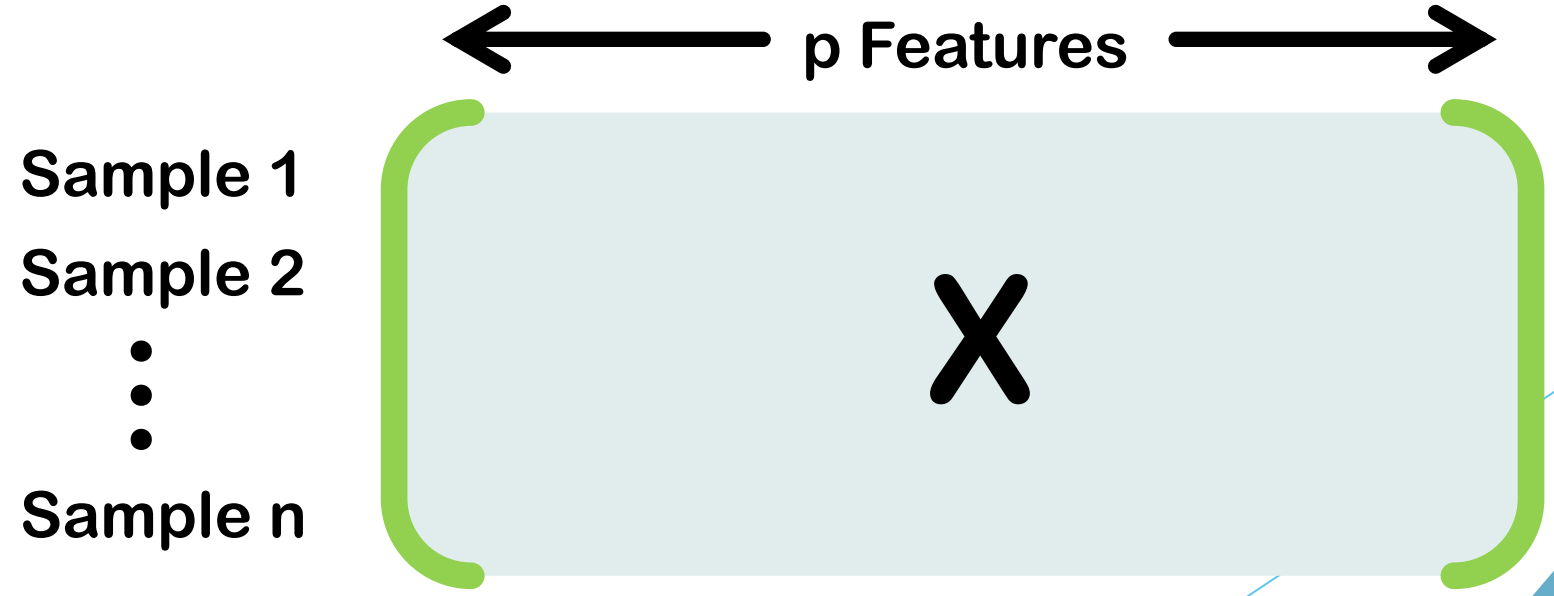
- Shiny: https://shiny.rstudio.com/gallery
  - Tutorial: https://shiny.rstudio.com/tutorial/
- Plotly: https://plot.ly/r/
- Crosstalk: https://rstudio.github.io/crosstalk/using.html
- Highcharter: http://jkunst.com/highcharter/hchart.html



Principal Components Analysis

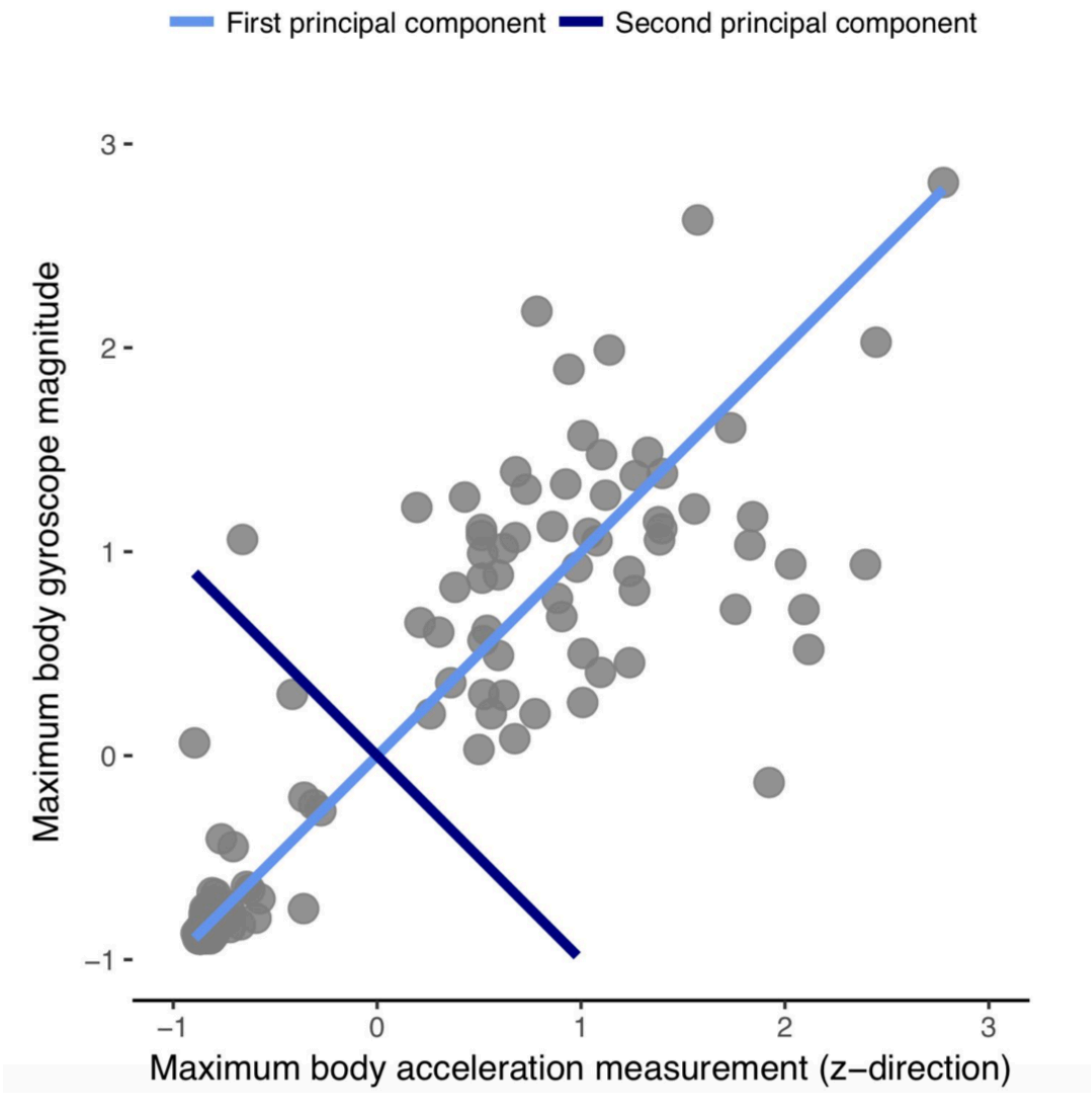
PCA: Motivation

- Exploratory Data Analysis and Visualization
- Dimension Reduction
  - Data compression
  - Denoising
- Pattern recognition
  - Clustering



PCA: Intuition

- Want to find a new (“better”) coordinate system
- Idea: find orthogonal directions (i.e., PCs) which maximize the variance in the data
  - Why? The hope is to find a lower-dimensional hyperplane (or representation) that retains most of the information in the data



PCA: Intuition

- Want to find a new (“better”) coordinate system
- Idea: find orthogonal directions (i.e., PCs) which maximize the variance in the data
  - Why? The hope is to find a lower-dimensional hyperplane (or representation) that retains most of the information in the data
- Equivalently, can view PCA as minimizing the projected (perpendicular) distance between the data points and the PC direction