

Meetup 3: Visualization

Quality Features

George Hagstrom

Week Summary

- This week we will read about more quality principles used to structure graphs
- There is lots of “smart” sounding advice, but is it good and how to do it?
- Discussion on different quality principles and begin Data Story 2
- Planning to make a coding vignette on [plotnine](#)

Data Story 2



The Fed has a dual mandate of price stability and maximum employment.

St. Louis FED

- Data Story 2 is about the Federal Reserve
- Emphasis is on improving figure basics

Principle of Uniformity of Graphics

- Represent data directly proportional to ink/graphics
 - Lie Factor = (size of effect shown in graphic) / (size of effect in data)
 - Proportional Ink
- Clear, detailed, and thorough labeling
 - Write explanations on the graphic.
 - Label important events
- Show data variation, not design variation

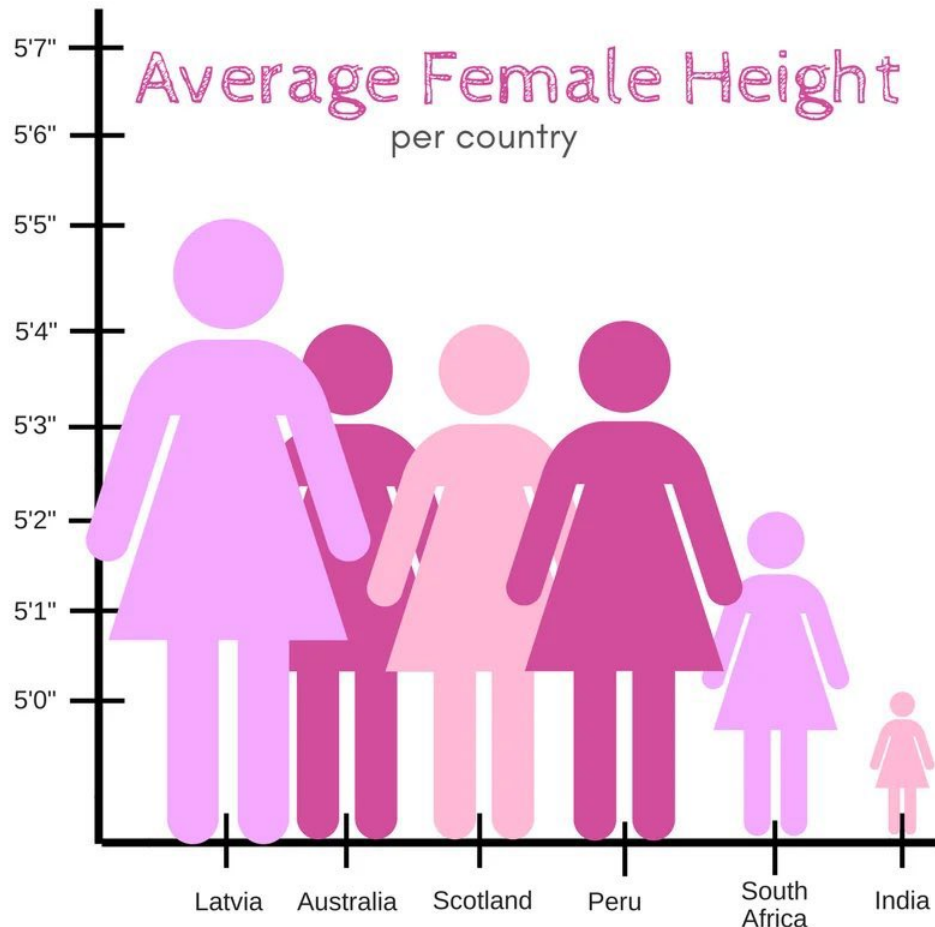
Principal of Proportional Ink

“The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.”

(1983, p.56) Edward Tufte, The Visual Display of Quantitative Information

- This principle avoids many misleading visualizations

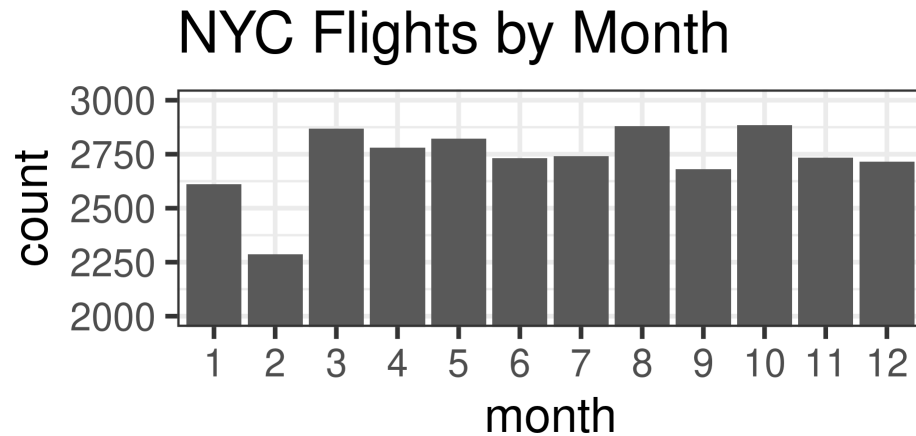
Example



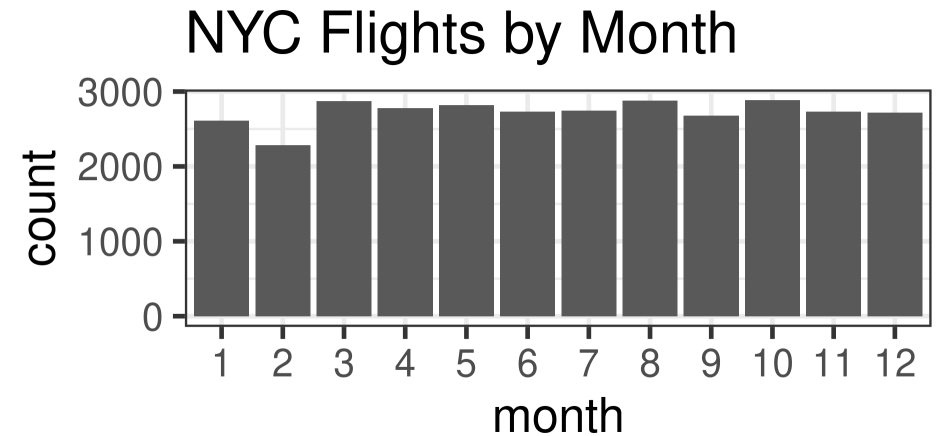
- Scale doesn't start at 0
- Width expands with height
- Danger in bar plots, filled line plots, plots with elements that scale with data

Bar Plots Should Start at 0

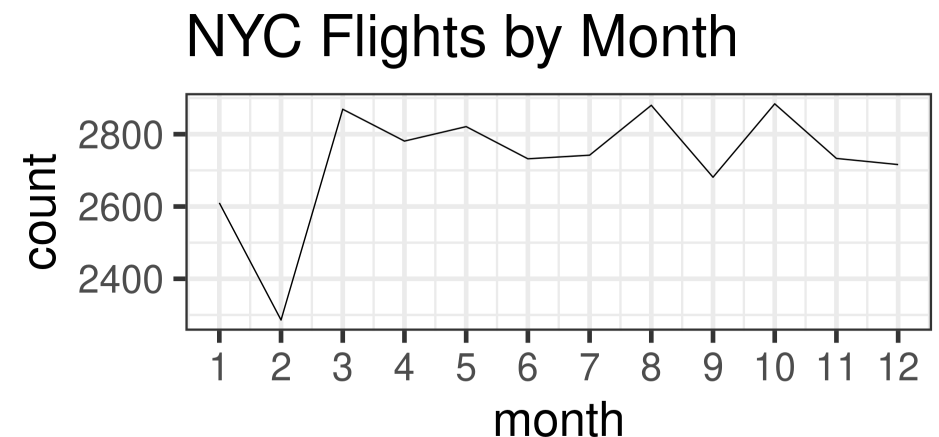
Wrong



Right



Or make them line charts:

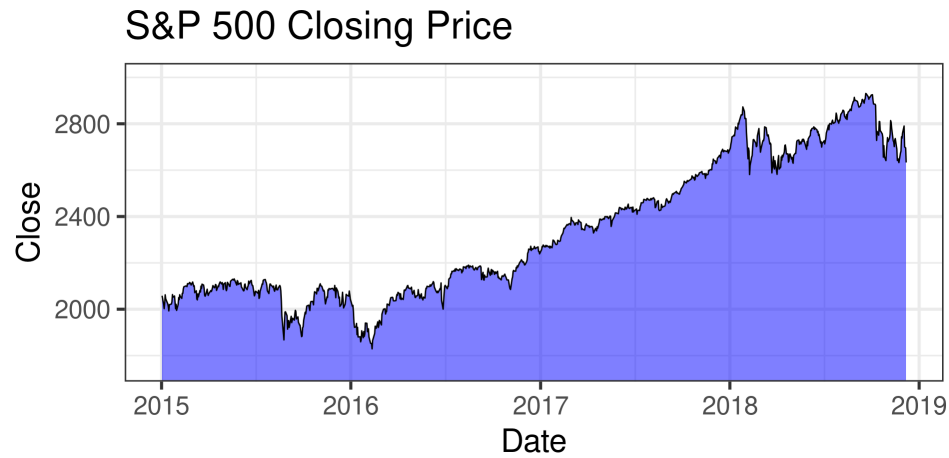


Bar Plots at 0: How?

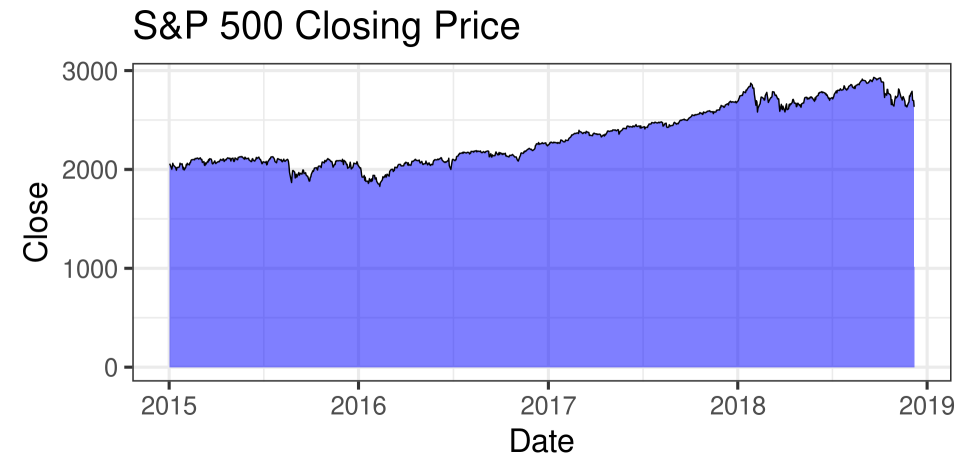
```
1 nycflights |> mutate(month = as.factor(month)) |> ggplot(aes(x=month)) +  
2   geom_bar() +  
3   theme_bw(base_size = 36) +  
4   labs(title = "NYC Flights by Month") +  
5   coord_cartesian(ylim = c(0, 3000))  
6  
7 nycflights |> group_by(month) |>  
8   summarise(count = n()) |> ggplot(aes(x=month,y=count)) +  
9   geom_line() +  
10  theme_bw(base_size = 36) +  
11  labs(title = "NYC Flights by Month") +  
12  scale_x_continuous(breaks = seq(1,12,by=1))
```


Filled Plots Should Start at 0

Wrong



Right



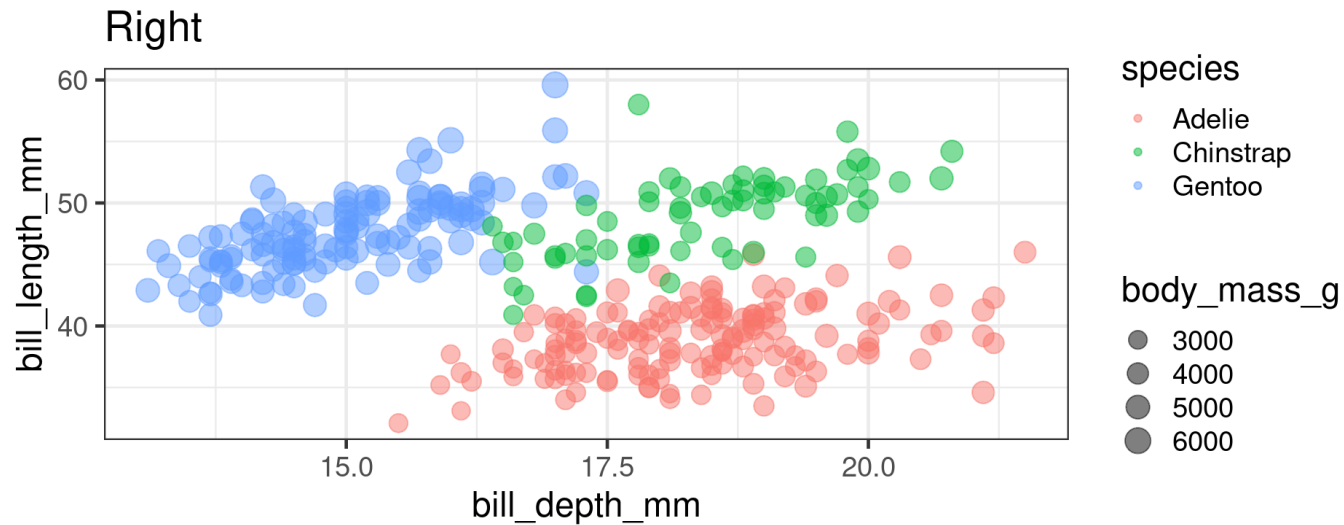
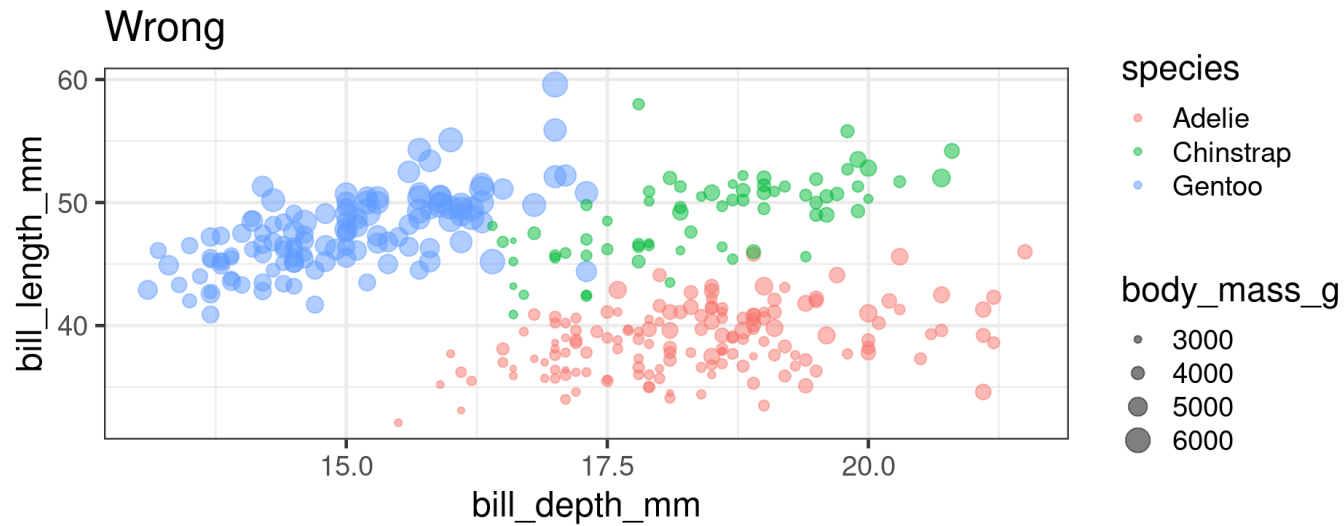
Or transform data to represent a net change:



Filled Plots: How?

```
1 sp500_1950_2018 |> mutate(Date = as.Date(Date)) |>
2   filter(Date > as.Date("2015-01-01")) |> ggplot(aes(x=Date,y=Close)) +
3   geom_area(fill = "blue",alpha = 0.5) +
4   geom_line() +
5   theme_bw(base_size = 24) +
6   labs(title = "S&P 500 Closing Price") +
7   coord_cartesian(ylim = c(0, 3000))
8
9 Initial_Date = sp500_1950_2018 |> filter(Date == "2015-01-02")
10 Initial_Open = Initial_Date$Open[1]
11
12 sp500_1950_2018 |> mutate(Date = as.Date(Date)) |>
13   filter(Date > as.Date("2015-01-03")) |>
14   mutate(Price_Change = Close - Initial_Open) |> ggplot(aes(x=Date,y=Price_
```

Scale Area not Radius



Scale Area: How?

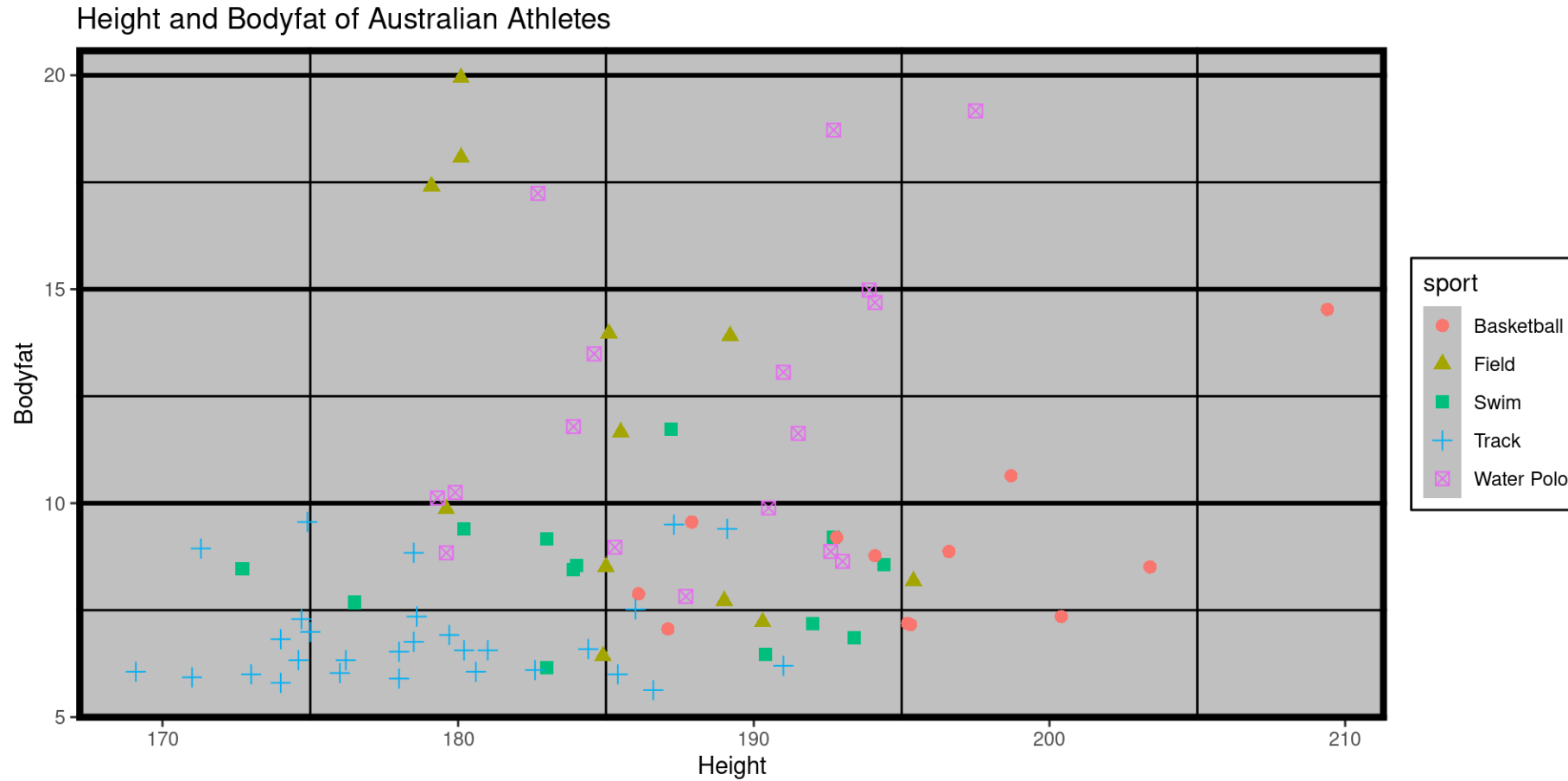
```
1 plotRight = penguins |> ggplot(aes(x=bill_depth_mm,  
2                                   y=bill_length_mm,  
3                                   size = body_mass_g,  
4                                   color=species)) +  
5   geom_point(alpha = 0.5) +  
6   labs(title = "Right") +  
7   scale_size_area() +  
8   theme_bw(base_size = 18)
```

Principles of Graphical Excellence

Graphical Excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space (Tufte)

- Two types of visual elements in figures
 - Data
 - Other Stuff
- Maximize Data-Ink Ratio

Example



- This figure has lots of chart-junk

How so ugly?

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5) +  
3   theme_excel(base_size = 24) +  
4   theme(  
5     panel.grid.major = element_line(color = "black", size = 1),  
6     panel.grid.minor = element_line(color = "black", size = 0.5),  
7     panel.border = element_rect(color = "black", size = 3)) +  
8   labs(title = "Height and Bodyfat of Australian Athletes") +  
9   xlab("Height") +  
10  ylab("Bodyfat")
```

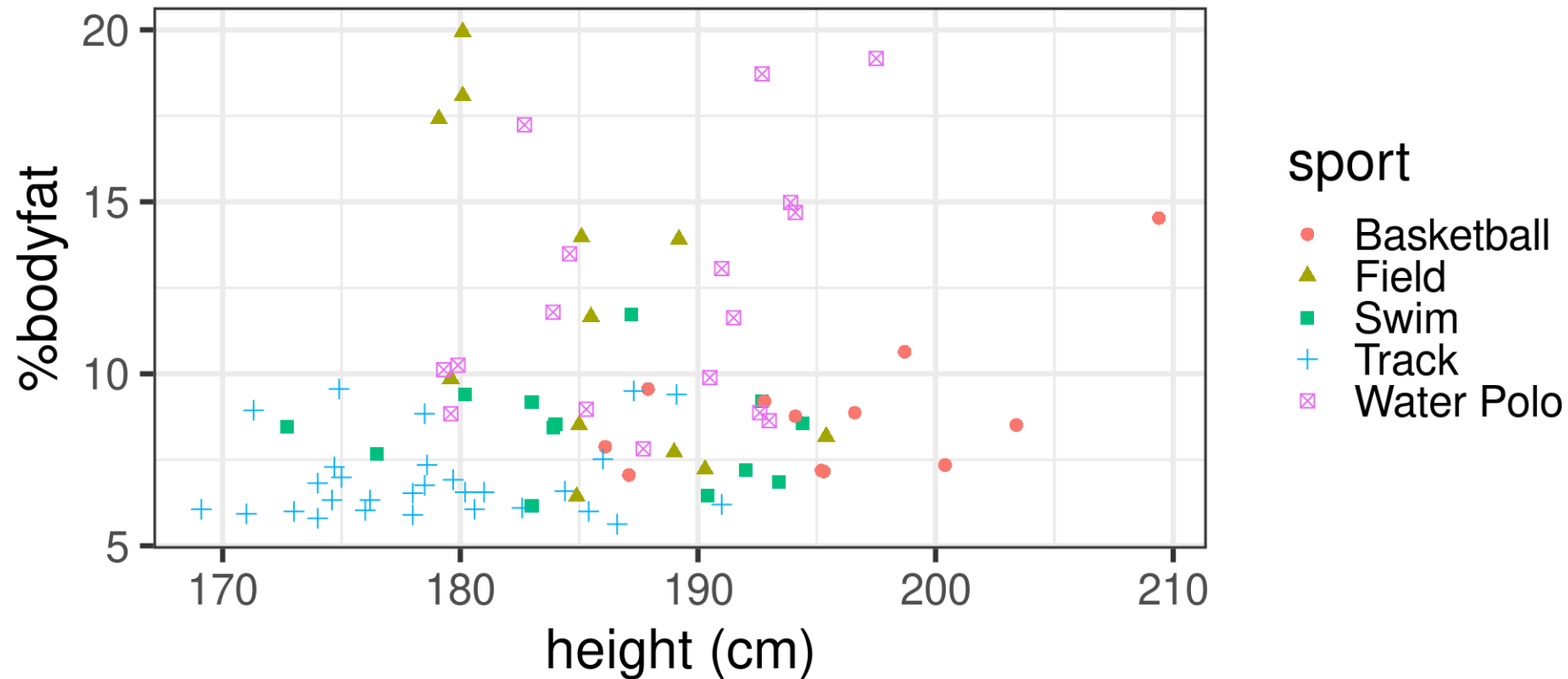
- `theme_excel` lots of poor decisions there
- `panel.grid.major`
- `panel.grid.minor`
- `panel.border`

Improve It

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5) +  
3   theme_bw(base_size = 24) +  
4   labs(title = "Height and Bodyfat of Australian Athletes") +  
5   xlab("height (cm)") +  
6   ylab("%bodyfat")
```


Improve It

Height and Bodyfat of Australian Athletes



- Removed the intentionally bad items

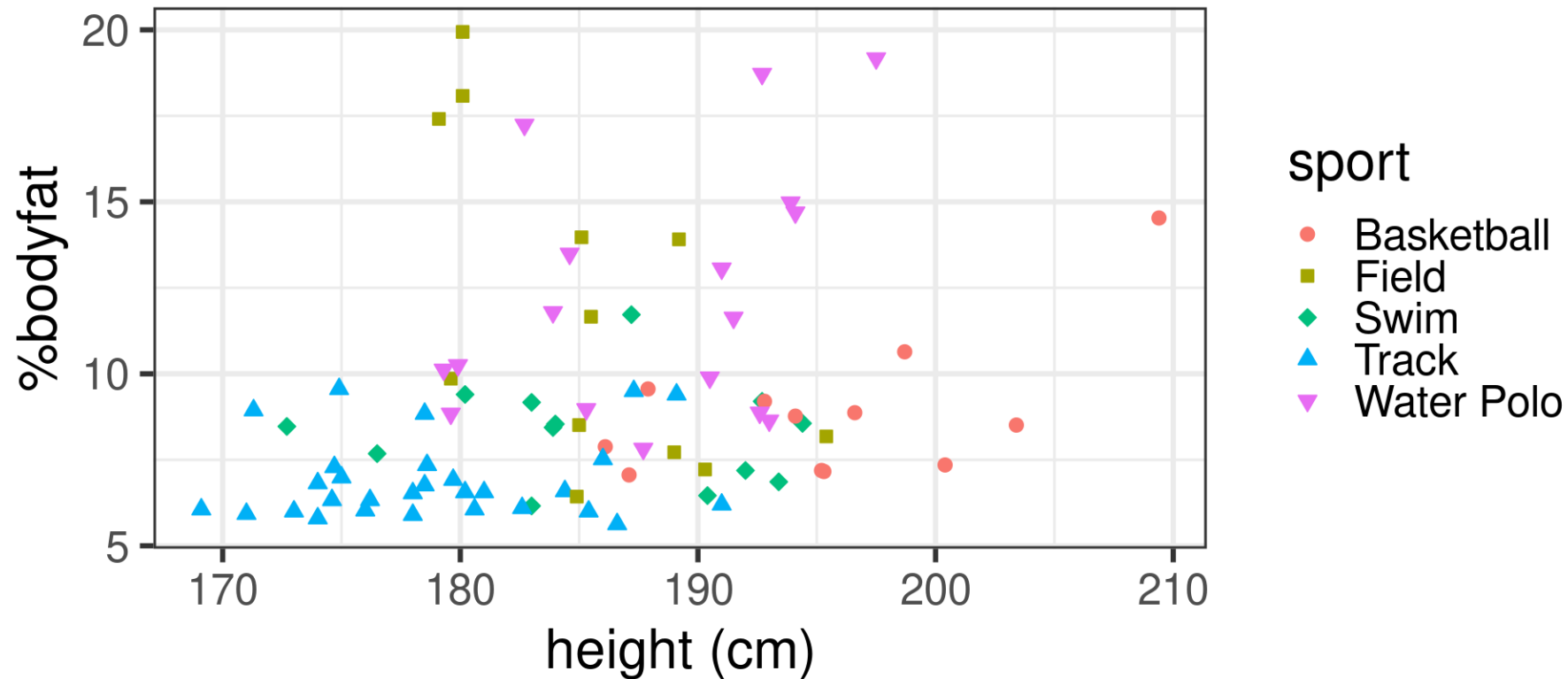
Improve It

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5) +  
3   theme_bw(base_size = 24) +  
4   scale_shape_manual(values = 21:25) +  
5   labs(title = "Height and Bodyfat of Australian Athletes") +  
6   xlab("height (cm)") +  
7   ylab("%bodyfat")
```

- Make the symbols solid

Improve It

Height and Bodyfat of Australian Athletes



- Make the symbols solid

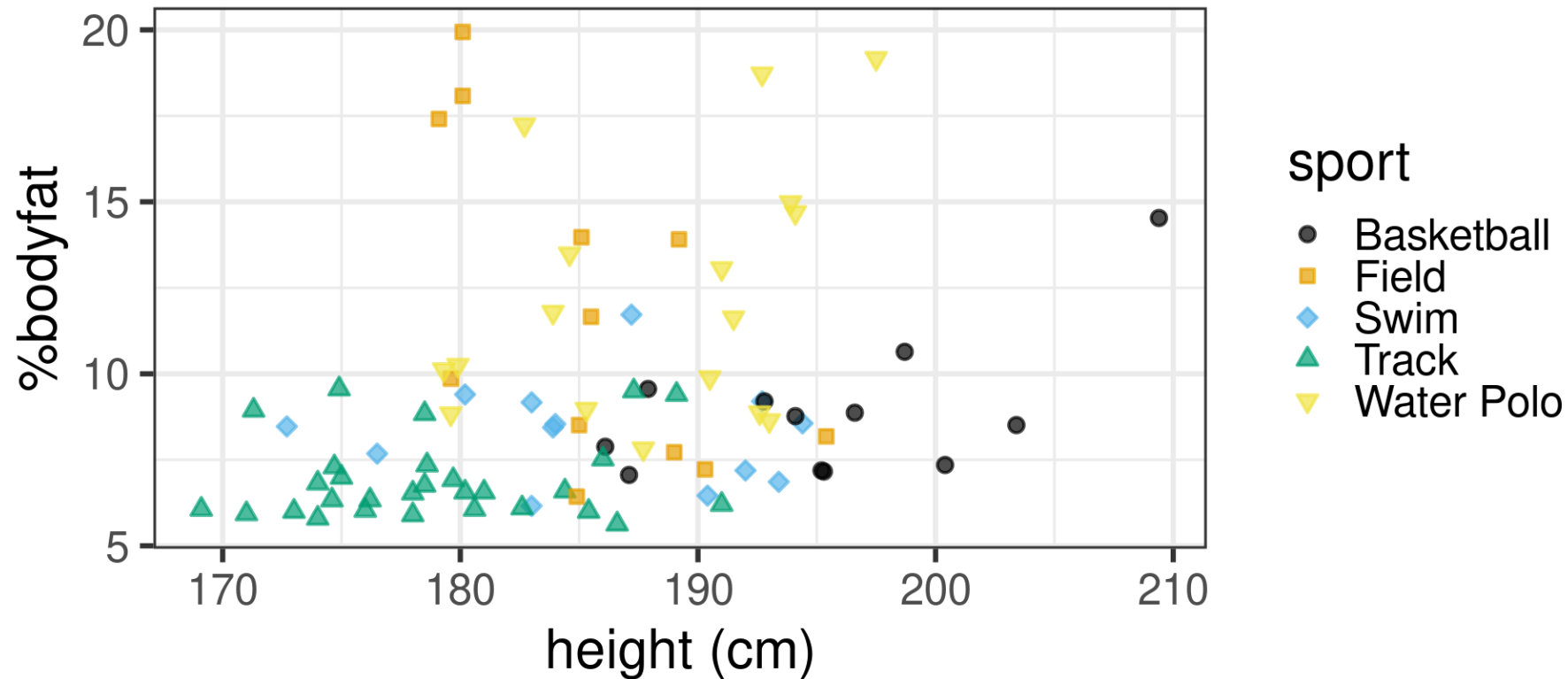
Improve It

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5,alpha=0.7,stroke=1) +  
3   theme_bw(base_size = 24) +  
4   scale_shape_manual(values = 21:25) +  
5   scale_color_colorblind() +  
6   scale_fill_colorblind() +  
7   labs(title = "Height and Bodyfat of Australian Athletes") +  
8   xlab("height (cm)") +  
9   ylab("%bodyfat")
```

- Switch to colorblind friendly palette

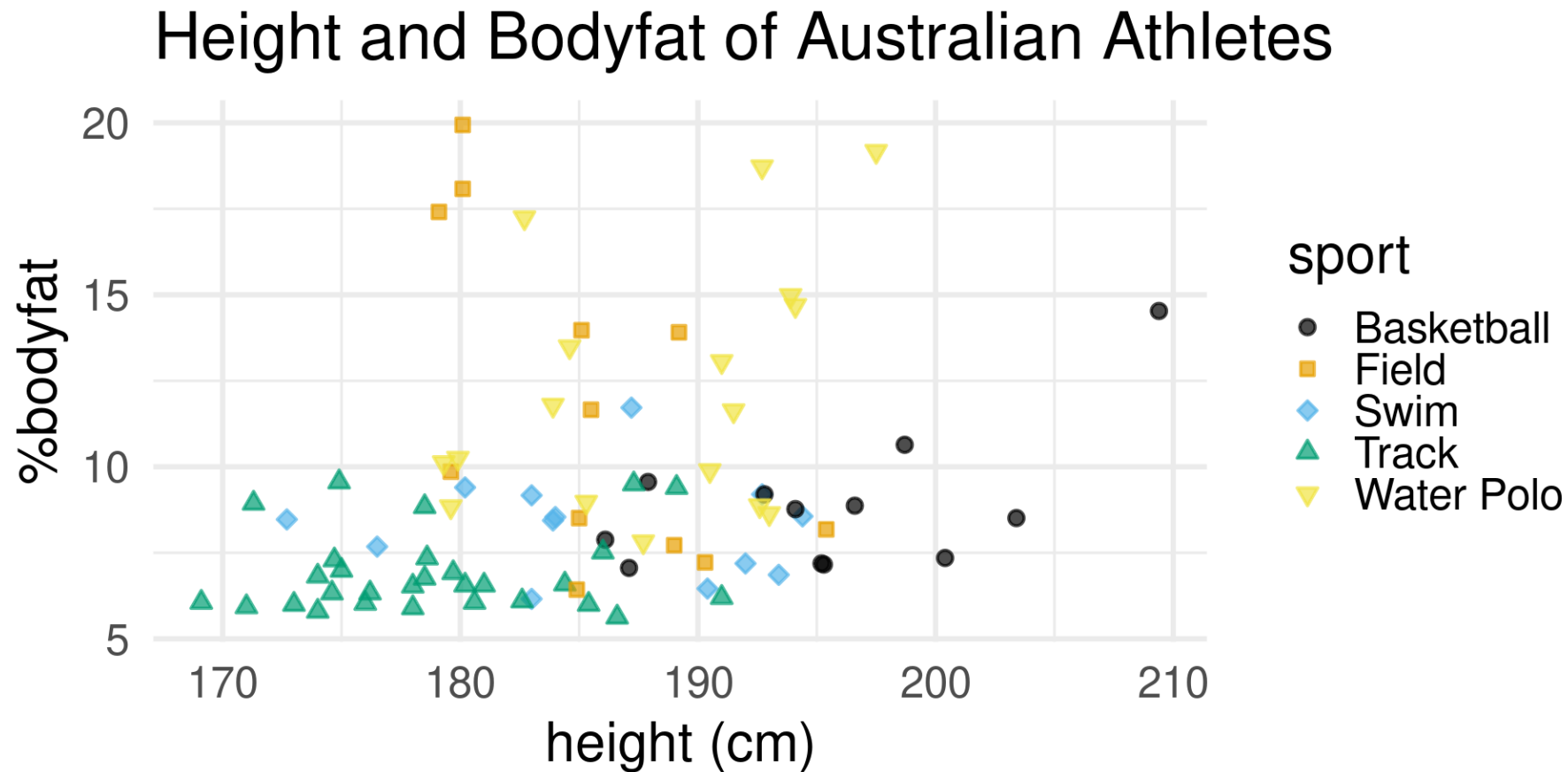
Improve It

Height and Bodyfat of Australian Athletes



- Switch to colorblind friendly palette

Improve It



- Try a theme without background frame

Improve It

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5,alpha=0.7,stroke=1.) +  
3   theme_minimal(base_size = 24) +  
4   scale_shape_manual(values = 21:25) +  
5   scale_color_colorblind() +  
6   scale_fill_colorblind() +  
7   labs(title = "Height and Bodyfat of Australian Athletes") +  
8   xlab("height (cm)") +  
9   ylab("%bodyfat")
```

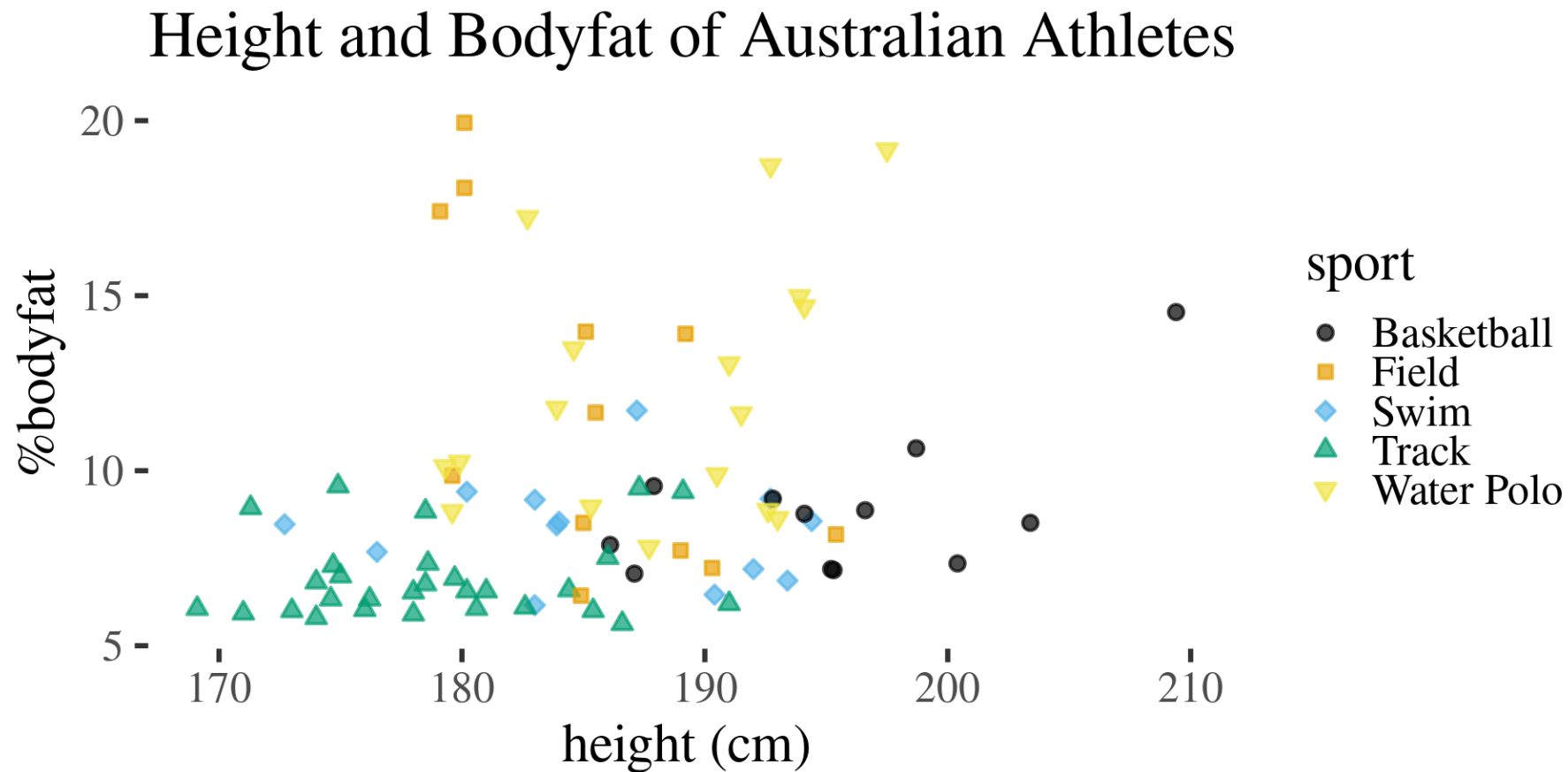
- Try a theme without background frame

Improve It?

```
1 male_Aus |> ggplot(aes(x=ht,y=pcBfat,color=sport,fill=sport,shape=sport)) +  
2   geom_point(size=2.5,alpha=0.7,stroke=1.) +  
3   theme_tufte(base_size = 24) +  
4   scale_shape_manual(values = 21:25) +  
5   scale_color_colorblind() +  
6   scale_fill_colorblind() +  
7   labs(title = "Height and Bodyfat of Australian Athletes") +  
8   xlab("height (cm)") +  
9   ylab("%bodyfat")
```

- Tufte theme gets rid of everything

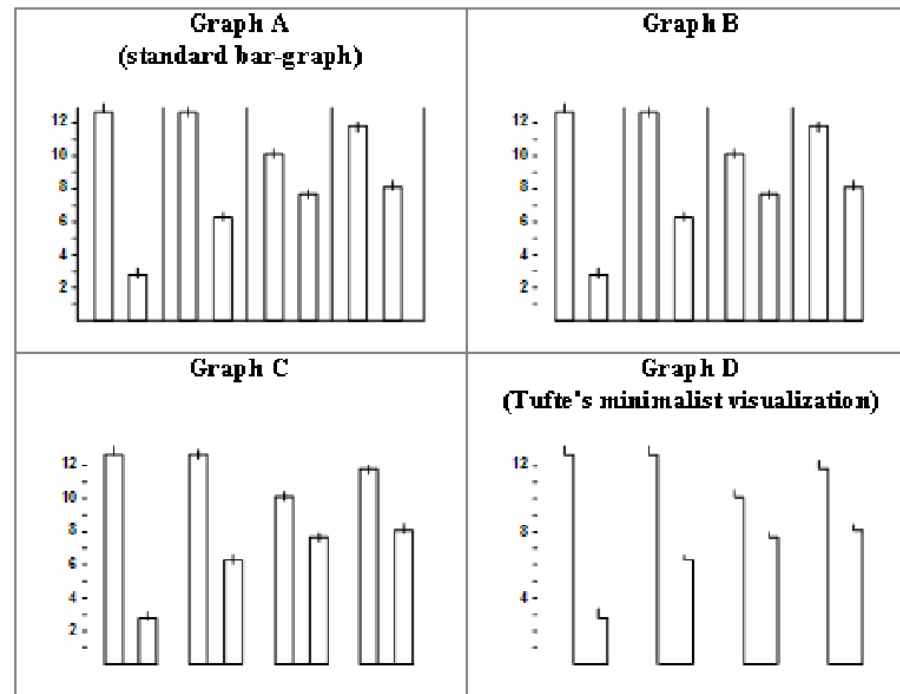
Improve It?



- Tufte theme gets rid of everything

How far is too far?

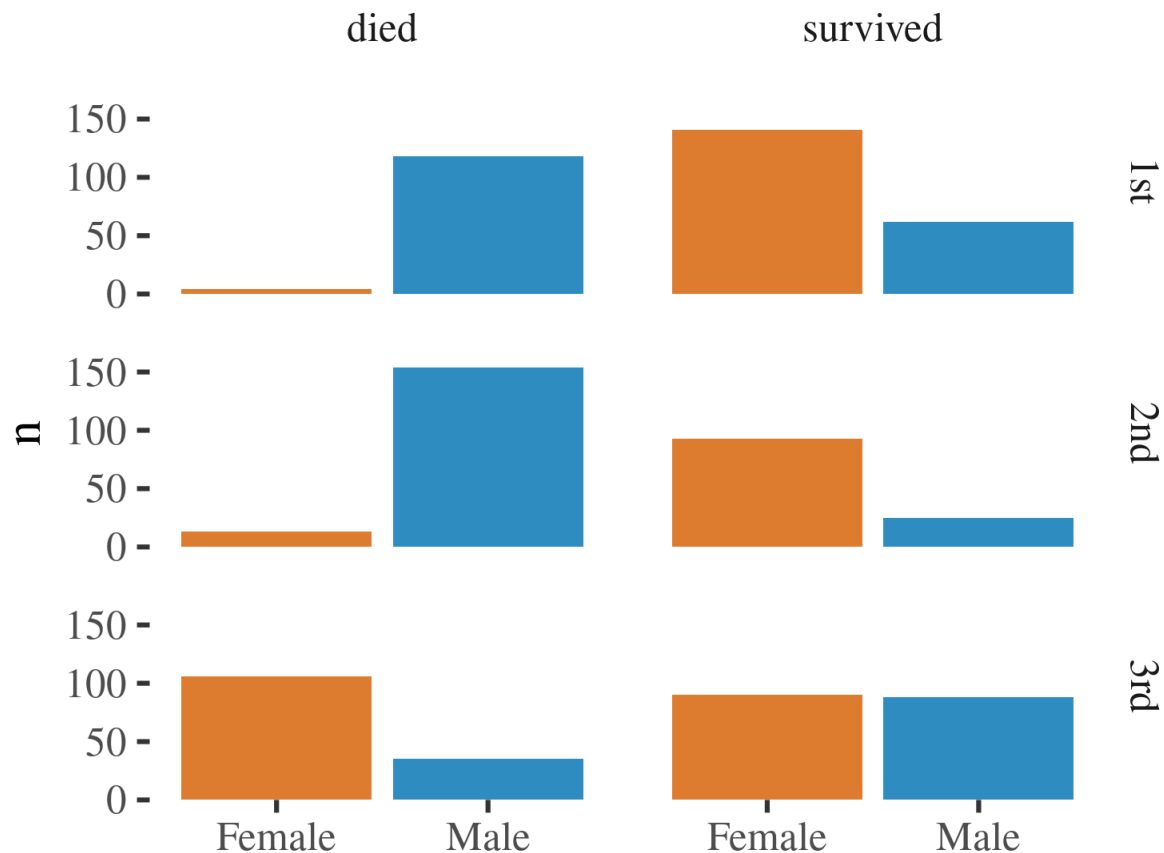
- I think it is possible to be too minimal
- Research supports this idea:



Indmar 2007

Maximize Data:Ink Within Reason

- When figures are too minimalistic elements appear to “float”



Maximize Data:Ink Within Reason

- When figures are too minimalistic elements appear to “float”

```
1 data("Titanic")
2 titanic = as_tibble(Titanic)
3 titanic |>
4   filter(Class != "Crew") |>
5   mutate(Survived =
6     if_else(Survived == "No", "died","survived")) |>
7   ggplot(aes(Sex,n,fill = Sex)) +
8   geom_col() +
9   facet_grid(Class ~ Survived,scales = "free_x") +
10  scale_x_discrete(name = NULL) +
11  scale_y_continuous(limits = c(0, 195), expand = c(0, 0)) +
12  scale_fill_manual(values = c("#D55E00D0", "#0072B2D0"), guide = "none")
13  theme_tufte(base_size = 24)
```

Maximize Data:Ink Within Reason

- Adding a Gray Background grounds figures



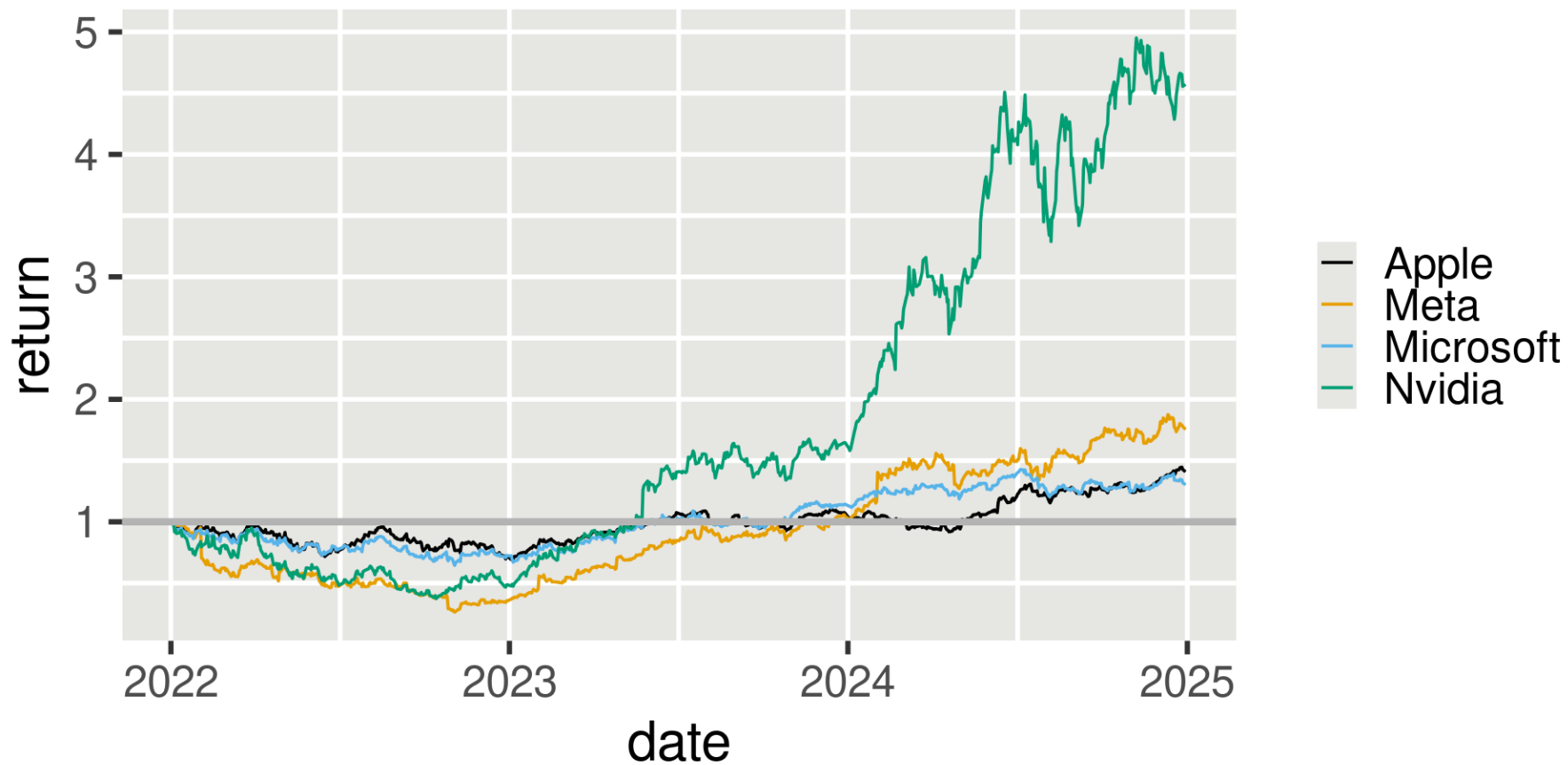
Maximize Data:Ink Within Reason

- Adding a Gray Background grounds figures

```
1 p +
2     theme_gray(base_size = 24) +
3     theme(axis.line = element_blank(),
4           #axis.line.y = element_line(color = colors[2]),
5           axis.ticks = element_blank(),
6           axis.ticks.length = grid::unit(0, "pt"),
7           #axis.ticks.y = element_line(color = colors[2]),
8           axis.text.x = element_text(margin = margin(7, 0, 3.5, 0)),
9     strip.text = element_text(margin = margin(3.5, 3.5, 3.5, 3.5)),
10    strip.background = element_blank(),
11    panel.background = element_rect(fill = "#E6E6E3",
12                                     linewidth = 0),
13    panel.grid.major.x = element_blank(),
14    panel.grid.minor.y = element_line(color="white",linewidth = 0.3),
15    panel.grid.major.y = element_blank(),
16    panel.grid.minor.x = element_blank())
```

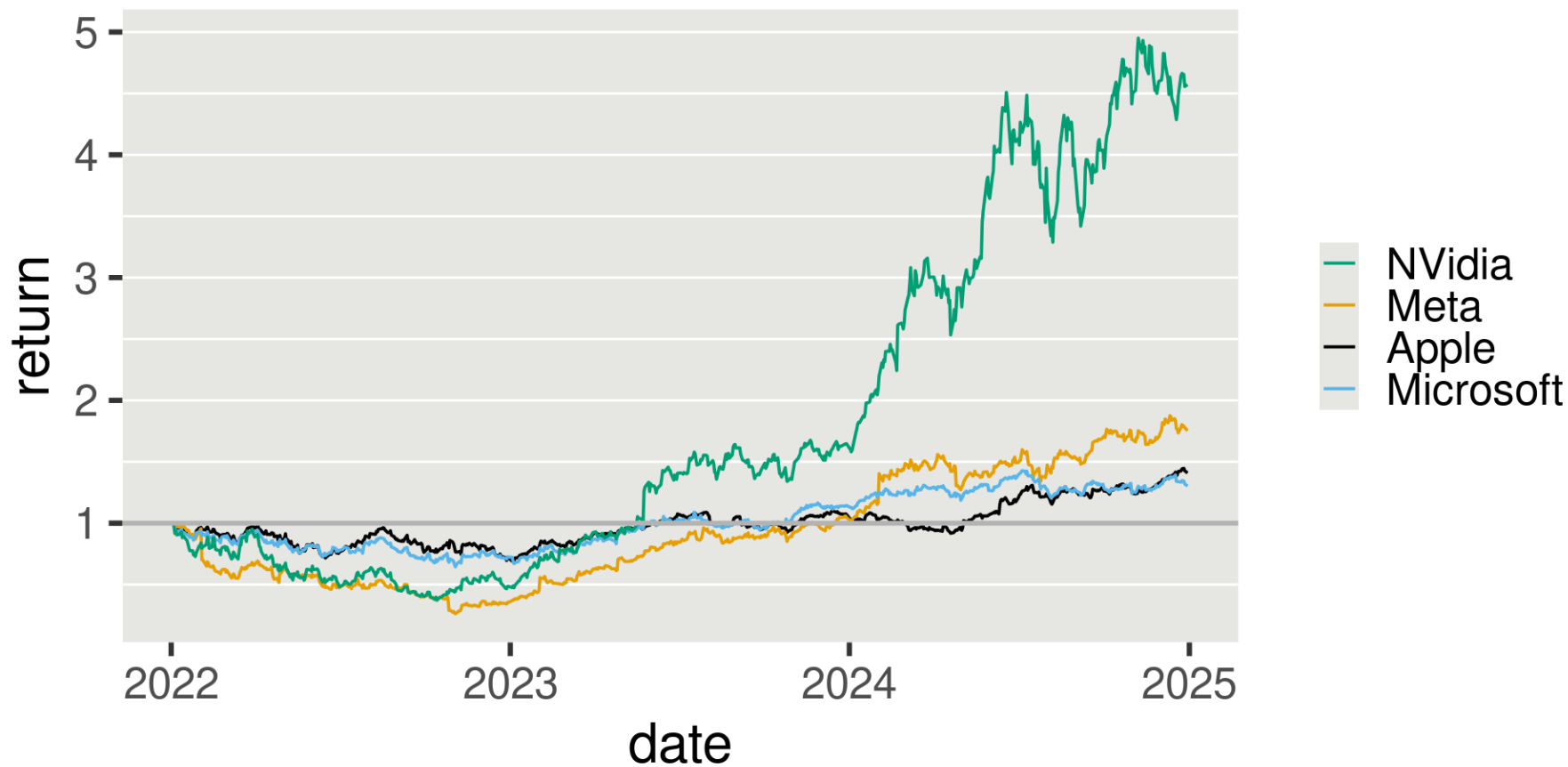
More Gridline Tips

- Consider gridlines perpendicular to axis of variation:



More Gridline Tips

- Consider gridlines perpendicular to axis of variation:



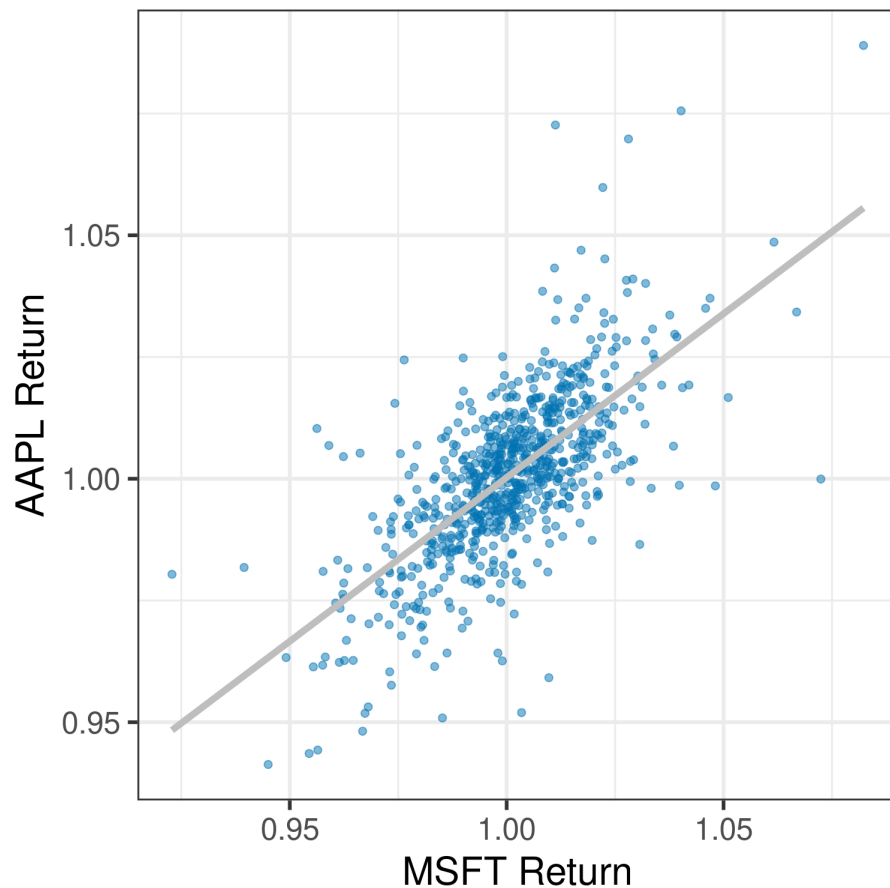
Gridlines: How?

- Consider gridlines perpendicular to axis of variation:

```
1 tech_stocks |>
2   ggplot(aes(x=date,y=return,group=Company,color=Company)) + geom_line(size
3   geom_hline(yintercept = 1, size = 1.0, color="grey70") +
4   theme_gray(base_size = 24) +
5   scale_color_manual(
6     values = c( "#009E73", "#E69F00", "#000000", "#56B4E"),
7     name = NULL,
8     breaks = c("NVDA", "META", "AAPL", "MSFT"),
9     labels = c("NVidia", "Meta", "Apple", "Microsoft")) +
10  theme(
11    strip.background = element_blank(),
12    panel.background = element_rect(fill = "#E6E6E3",
13    linewidth = 0),
14    panel.grid.minor.y = element_line(color="white",linewidth = 0.5),
15    panel.grid.major.y = element_line(color="white",linewidth = 0.5),
16    panel.grid.minor.x = element_blank(),
17    panel.grid.major.x = element_blank(),
18  )
```

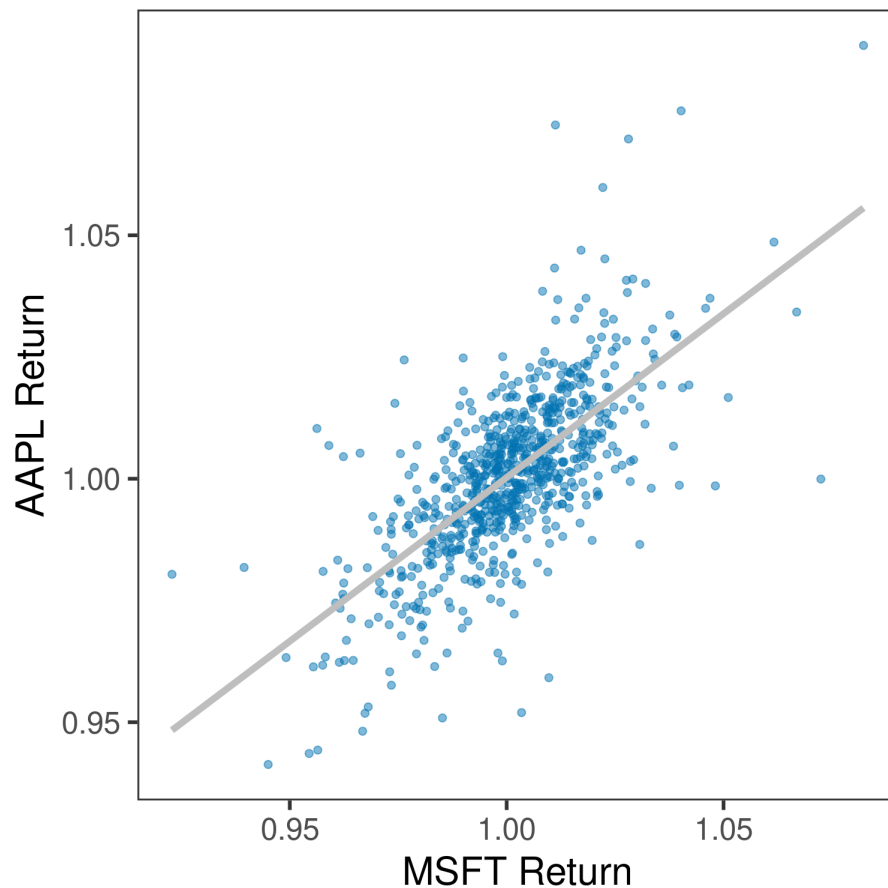
More Gridline Tips Paired Data

- For paired data, draw a regression line



More Gridline Tips Paired Data

- For paired data, but drop gridlines



More Gridline Tips Paired Data

- For paired data, but drop gridlines

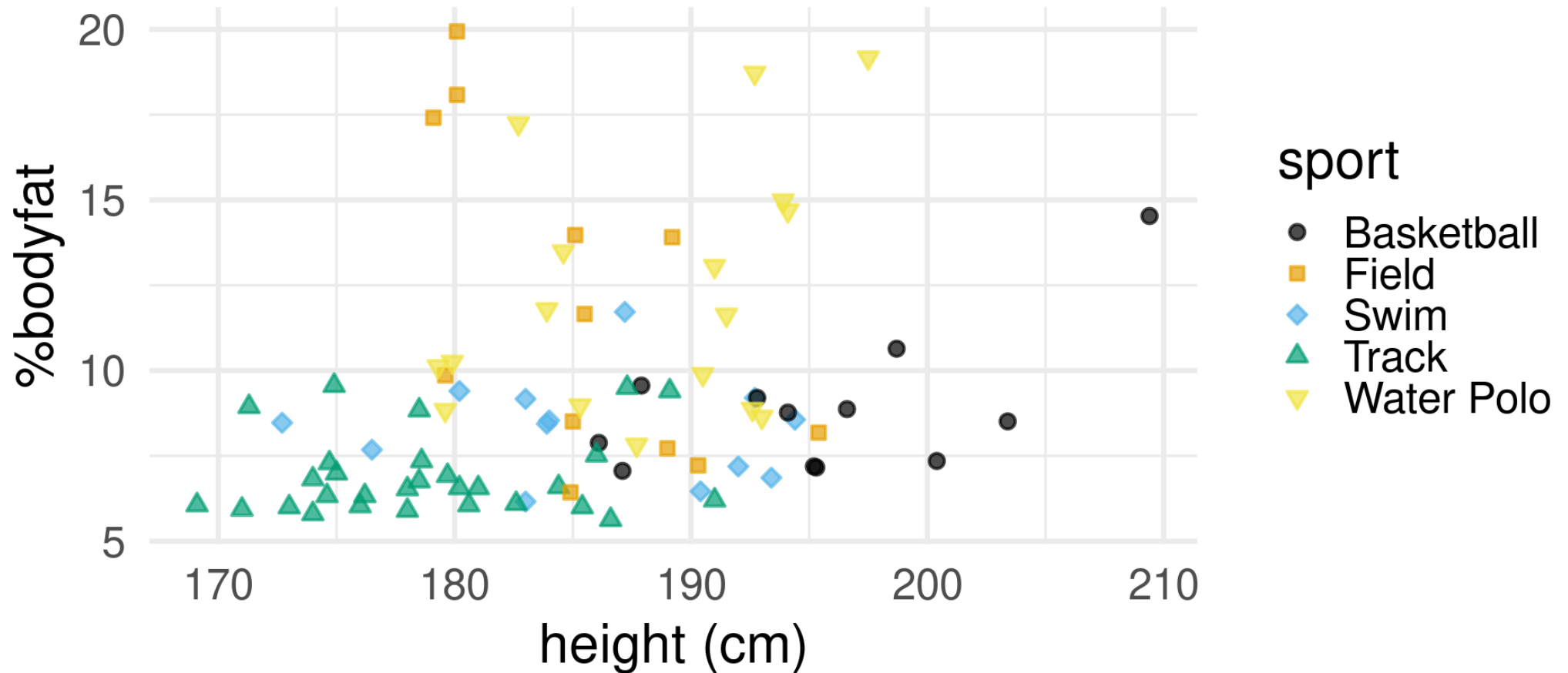
```
1 tech_stocks = tech_stocks |> group_by(Company) |> mutate(daily_return = ret
2 tech_stocks |> select(Company,date,daily_return) |> pivot_wider(names_from=
3   ggplot(aes(x=MSFT,y=AAPL)) +
4   geom_point(alpha = 0.5,fill = "#0072B2D0",color="#0072B2D0",size=2) +
5   stat_smooth(method="lm",color="gray",se = FALSE,linewidth=2) +
6   theme_bw(base_size = 24) +
7   xlab("MSFT Return") +
8   ylab("AAPL Return") +
9   theme(
10     panel.grid.minor.y = element_blank(),
11       panel.grid.major.y = element_blank(),
12       panel.grid.minor.x = element_blank(),
13       panel.grid.major.x = element_blank(),
14   )
```

Consider Larger Font Size

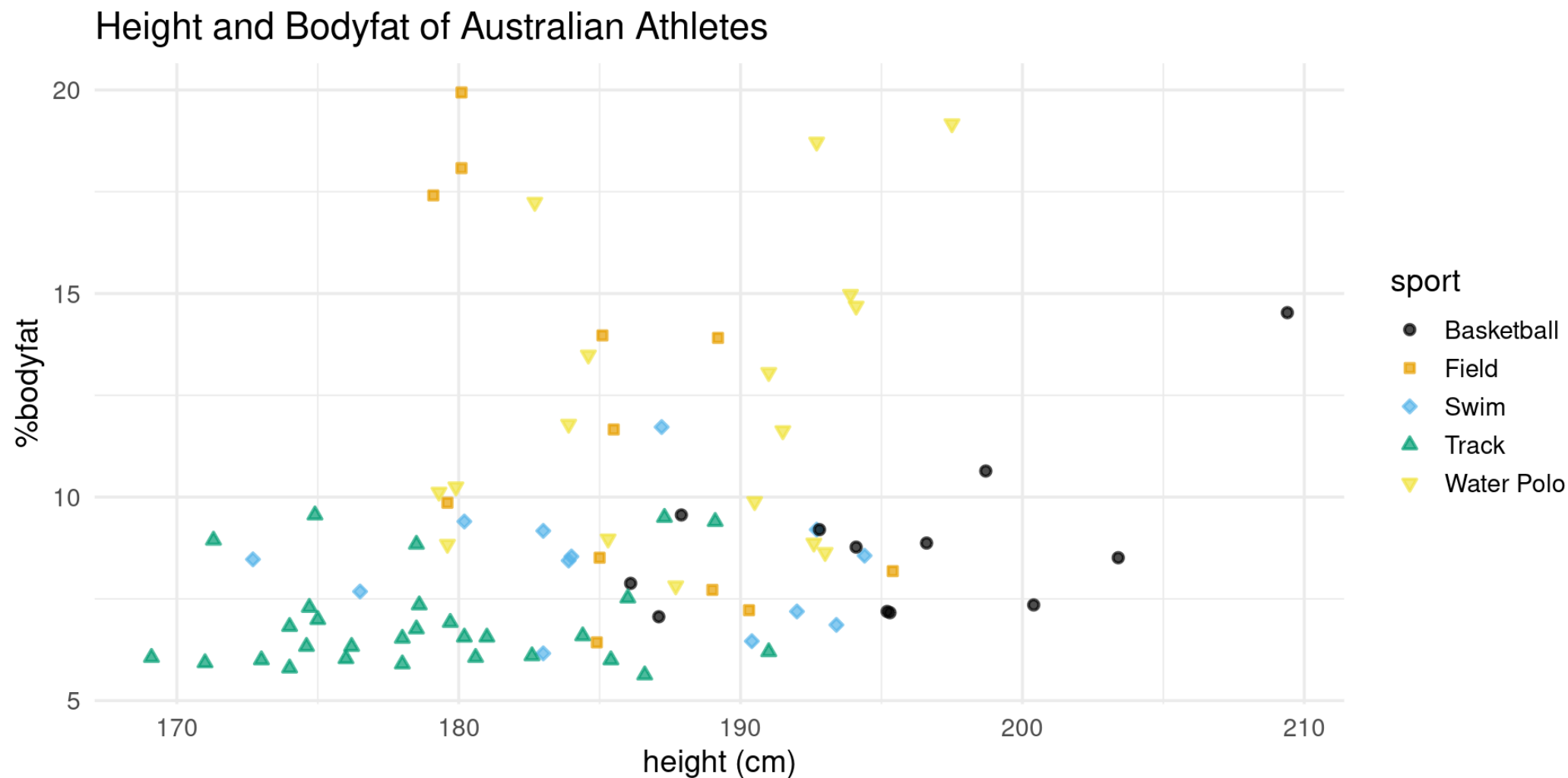
- Most common mistake in in person presentations
 - If you can, look at your slides from the back of the room
- Also an issue for figures in other formats
 - You see figure in full size on your screen
 - Publisher/Publishing software scales it down to fit it with your text

Good Font Size

Height and Bodyfat of Australian Athletes

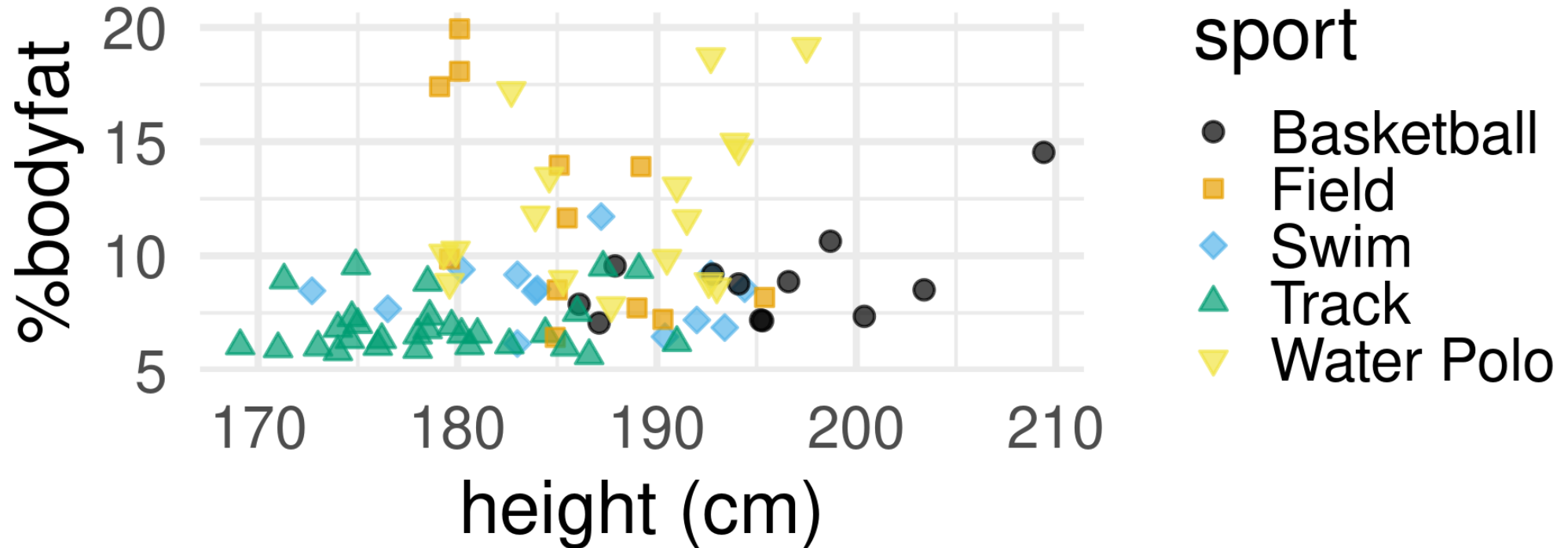


Small Font Size



Large Font Size

Height and Bodyfat of Australian Athletes



How to change size

ggplot2 based advice:

- `theme_x(base_size = ?)`
- `geom_point(size = ?)`
- `geom_line(linewidth=?)`
- `panel.grid.major.y =
element_rect(linewidth = ?)`
- `axis.text.x = element_text(size=?)`

Thanks!