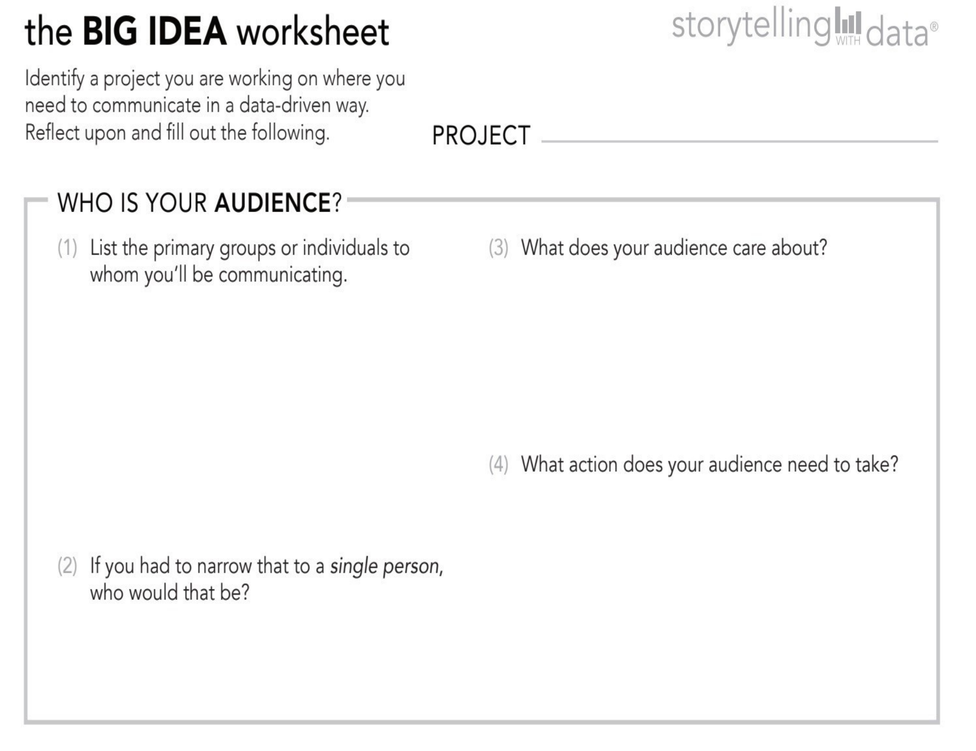
GG501

2. Tabular data visualization

Questions for storytelling with data



the **BIG IDEA** worksheet

storytelling with data®

Identify a project you are working on where you need to communicate in a data-driven way. Reflect upon and fill out the following.

PROJECT ______________________

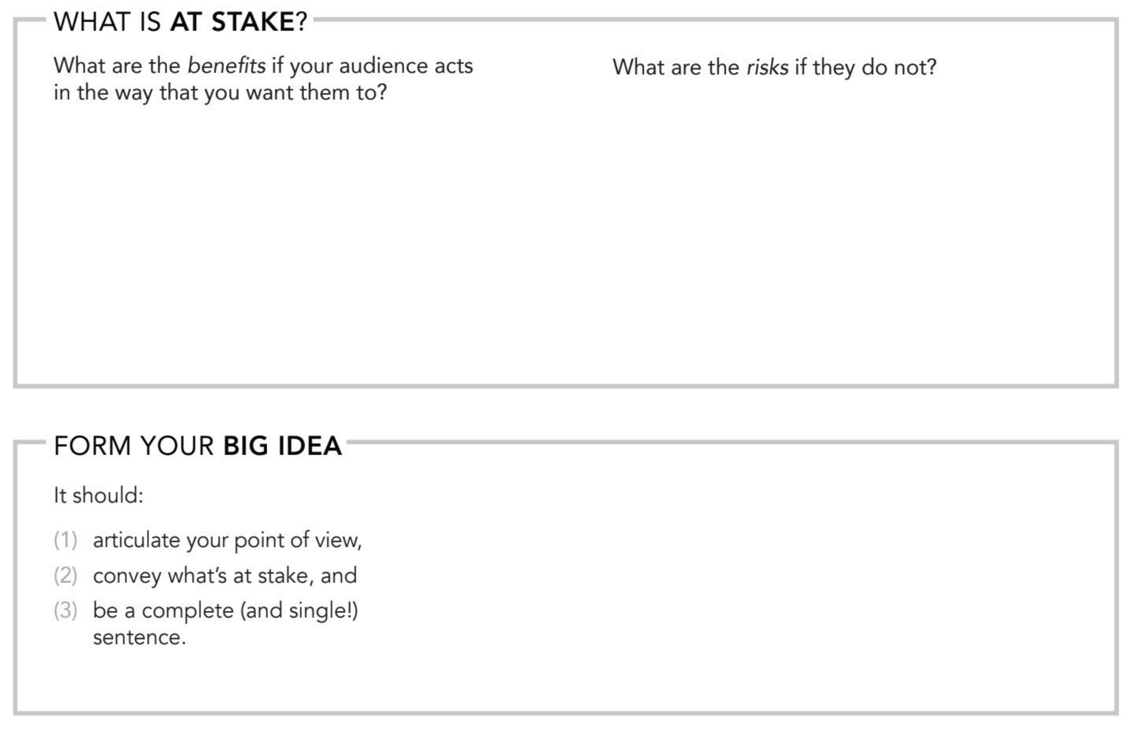
WHO IS YOUR **AUDIENCE**?

(1) List the primary groups or individuals to whom you'll be communicating.

(2) If you had to narrow that to a *single person*, who would that be?

(3) What does your audience care about?

(4) What action does your audience need to take?



WHAT IS **AT STAKE**?

What are the *benefits* if your audience acts in the way that you want them to?

What are the *risks* if they do not?

FORM YOUR **BIG IDEA**

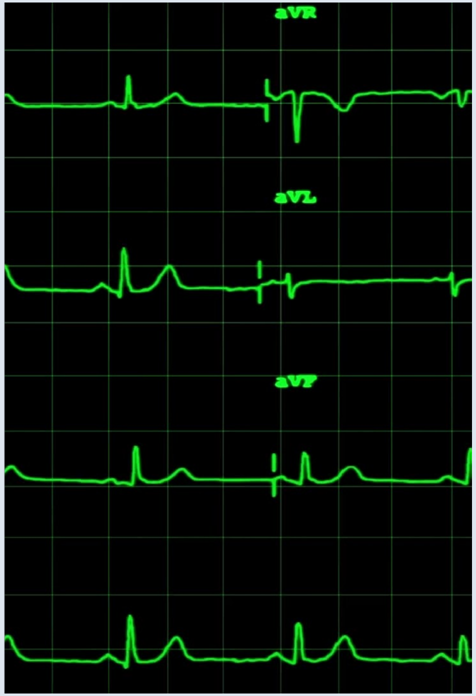
It should:

(1) articulate your point of view,

(2) convey what's at stake, and

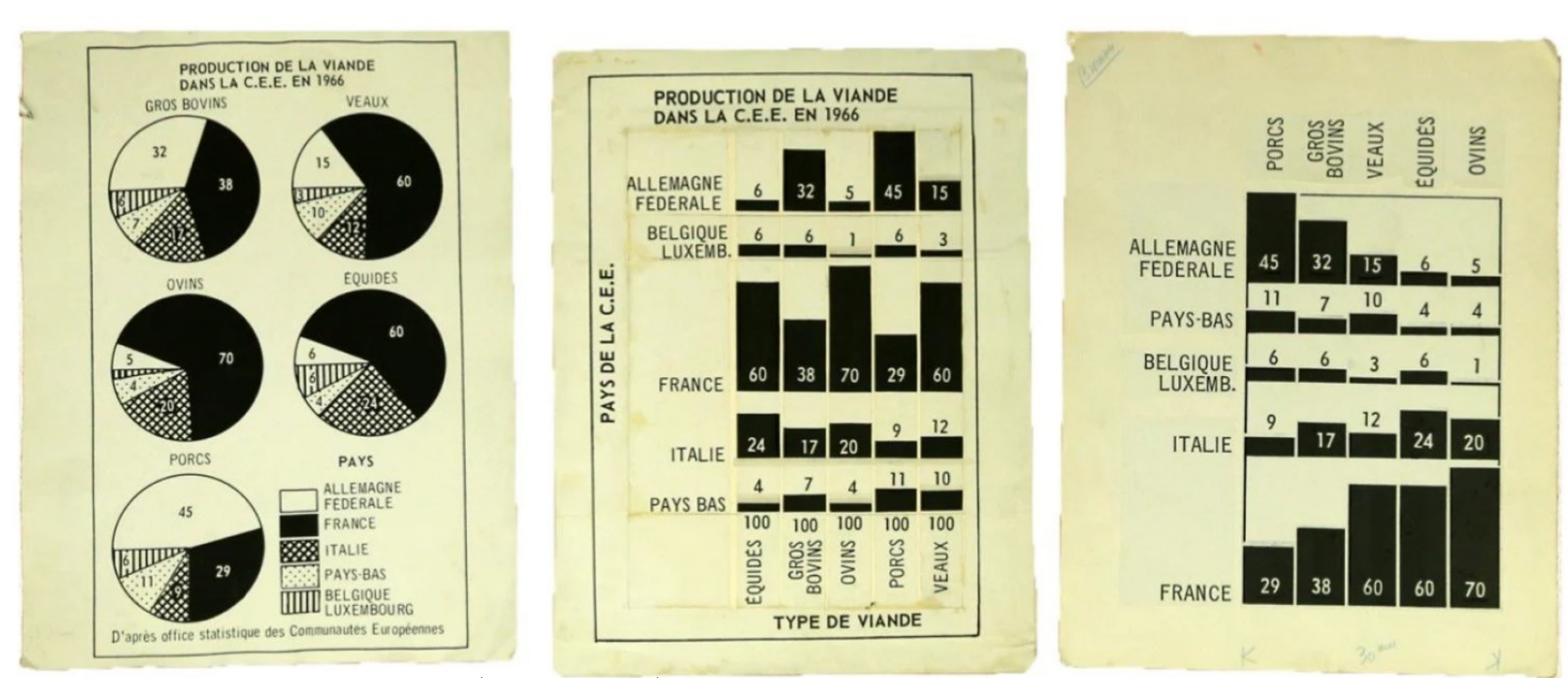
(3) be a complete (and single!) sentence.

Recognizing effective visualization



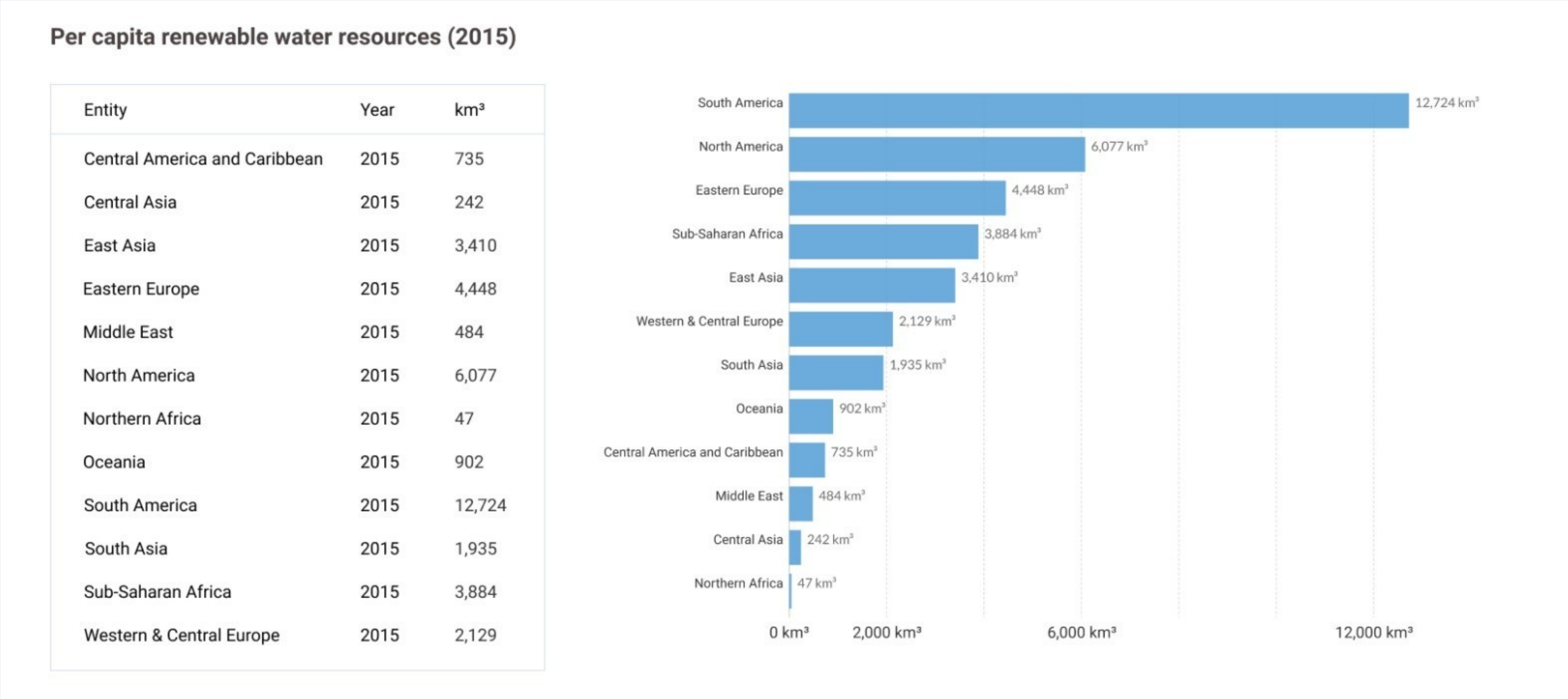
- Does the visualization tell a clear story
  - or the best story possible given the data at hand
- Is there too much or too little detail?
- Is natural ordering or hierarchy exploited properly and according to convention?
  - e.g., time typically increasing on the X-axis
- static vs. dynamic – incorporating interactivity can enhance or degrade the experience – depends on the context

Visualizations can always be improved



Drafts for the book La Graphique (Bertin, 1977), Courtesy of EHESS/AN ref. 20010291/36. - Tetiana Donska

Tabular vs Visual



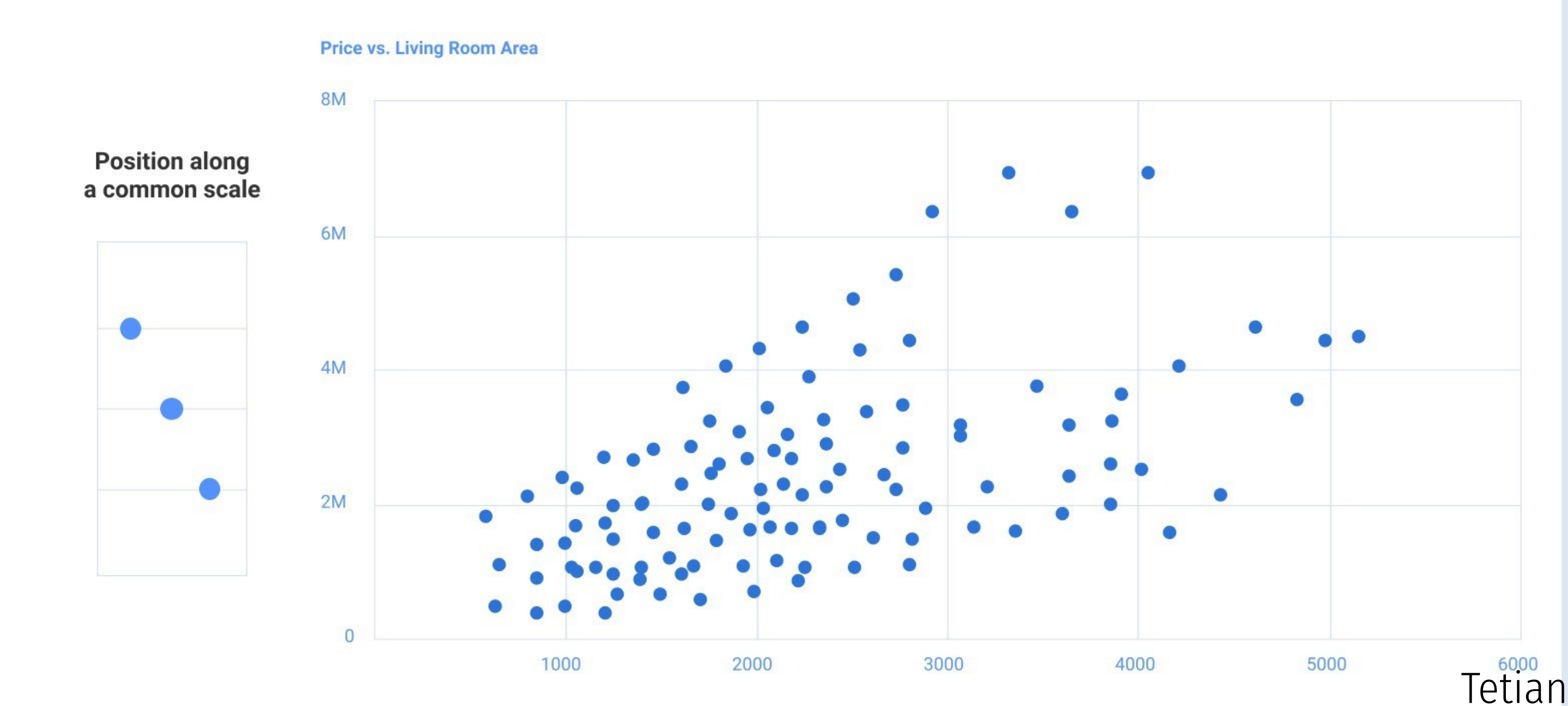
Per capita renewable water resources (2015)

Entity	Year	km³
Central America and Caribbean	2015	735
Central Asia	2015	242
East Asia	2015	3,410
Eastern Europe	2015	4,448
Middle East	2015	484
North America	2015	6,077
Northern Africa	2015	47
Oceania	2015	902
South America	2015	12,724
South Asia	2015	1,935
Sub-Saharan Africa	2015	3,884
Western & Central Europe	2015	2,129

Visual elements & human perception

1. Position along a common scale

   position is the easiest feature to recognize & evaluate with regard to elements in space

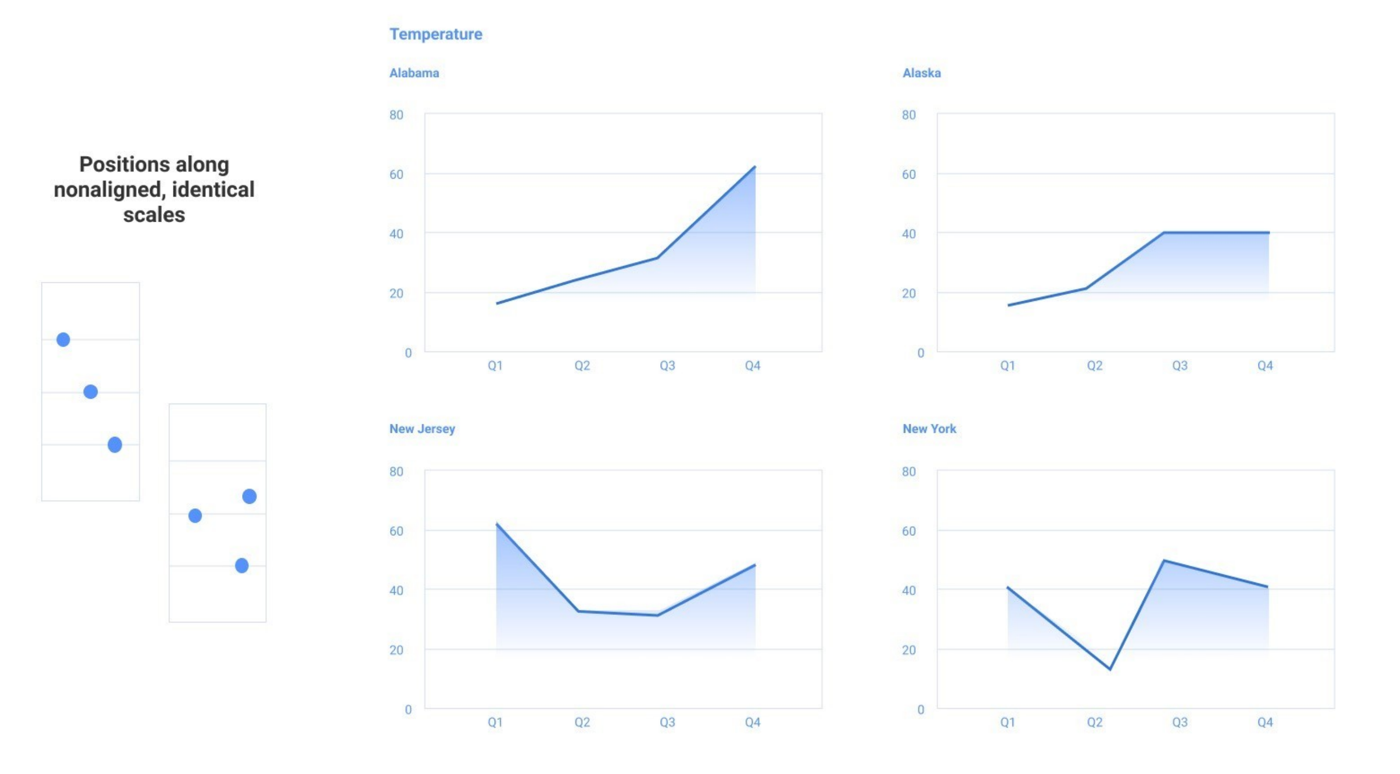


Tetiana Donska

Visual elements & human perception

2. Position along non-aligned by identical scales

often more effective to break into multiple subplots

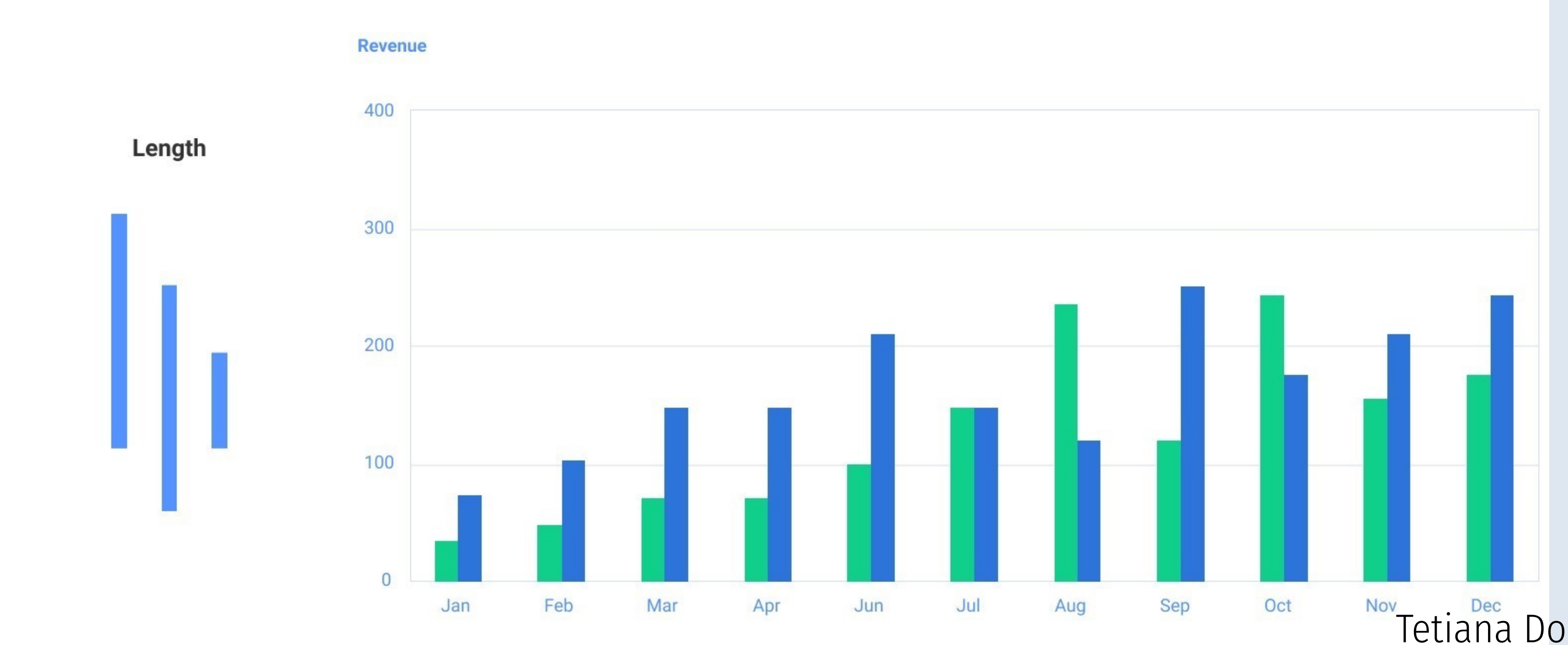


Tetiana Donska

Visual elements & human perception

3. Length

human brain easily recognizes proportions and evaluated length, even if objects are not aligned

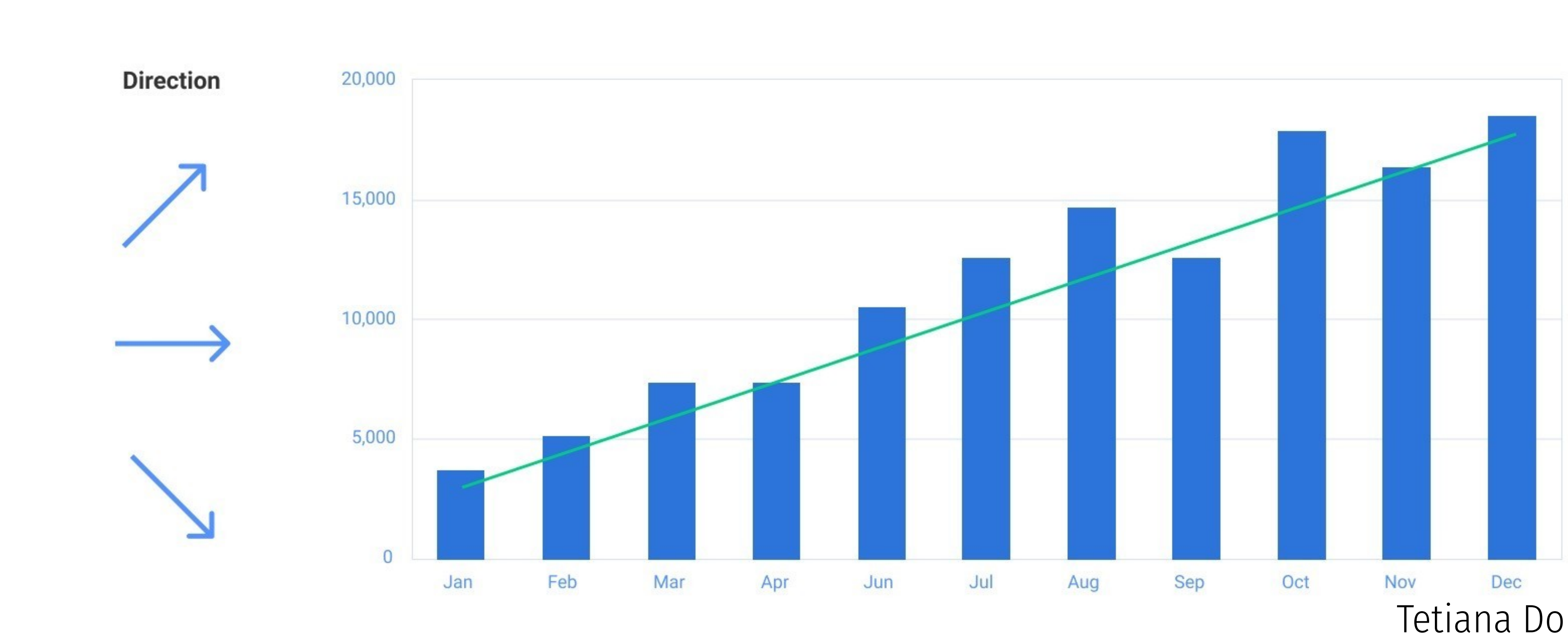


Tetiana Donska

Visual elements & human perception

4. Direction

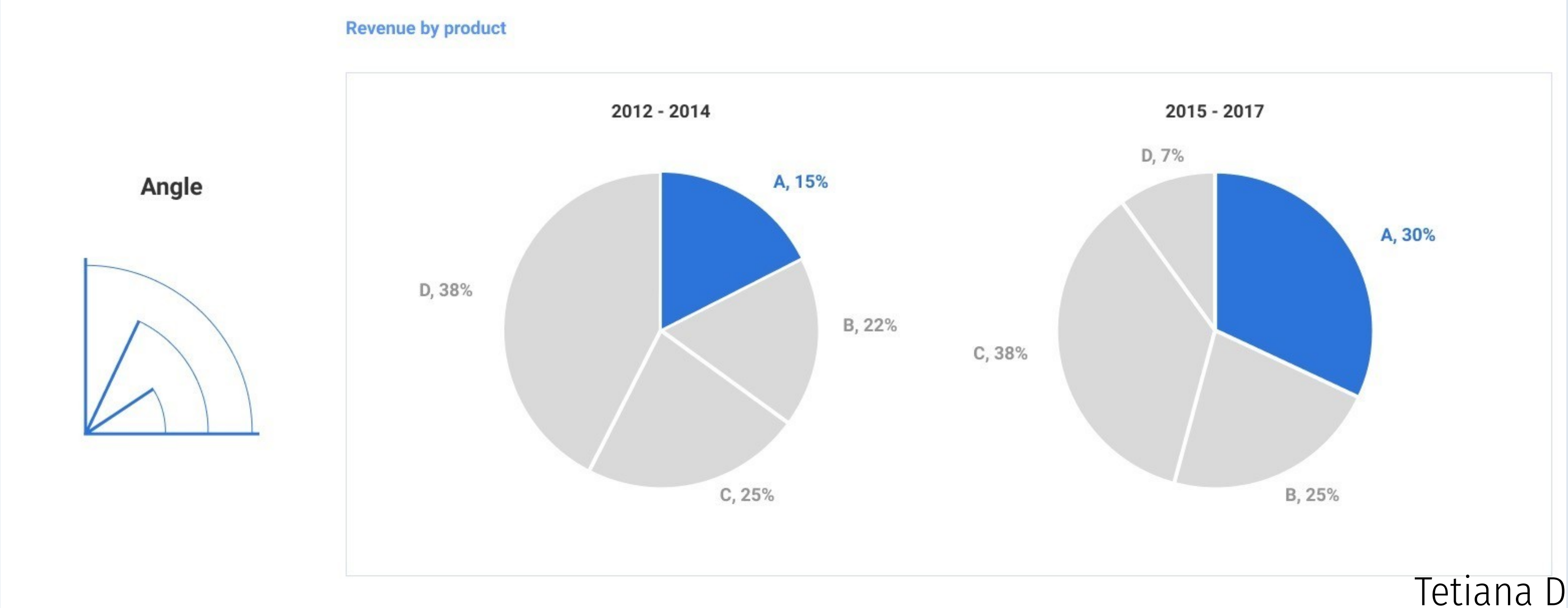
easily recognized by the human eye –line and trend charts to present data that changes over time



Tetiana Donska

Visual elements & human perception

5. Angle
   - angles are harder to evaluate than length or position – generally bar charts should be used first
   - if the number of categories is small pie charts can sometimes be effective

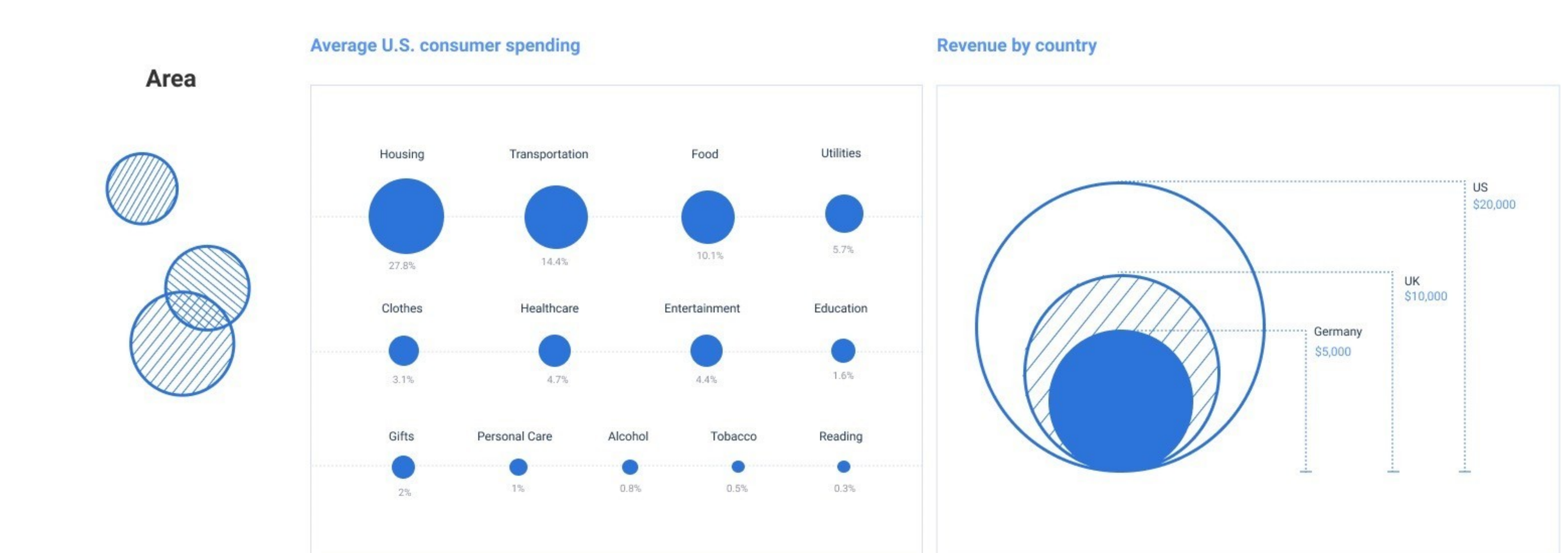


Visual elements & human perception

6. Area

relative magnitude of areas is harder to compare versus the length of lines

however area often useful on maps as proportional symbols

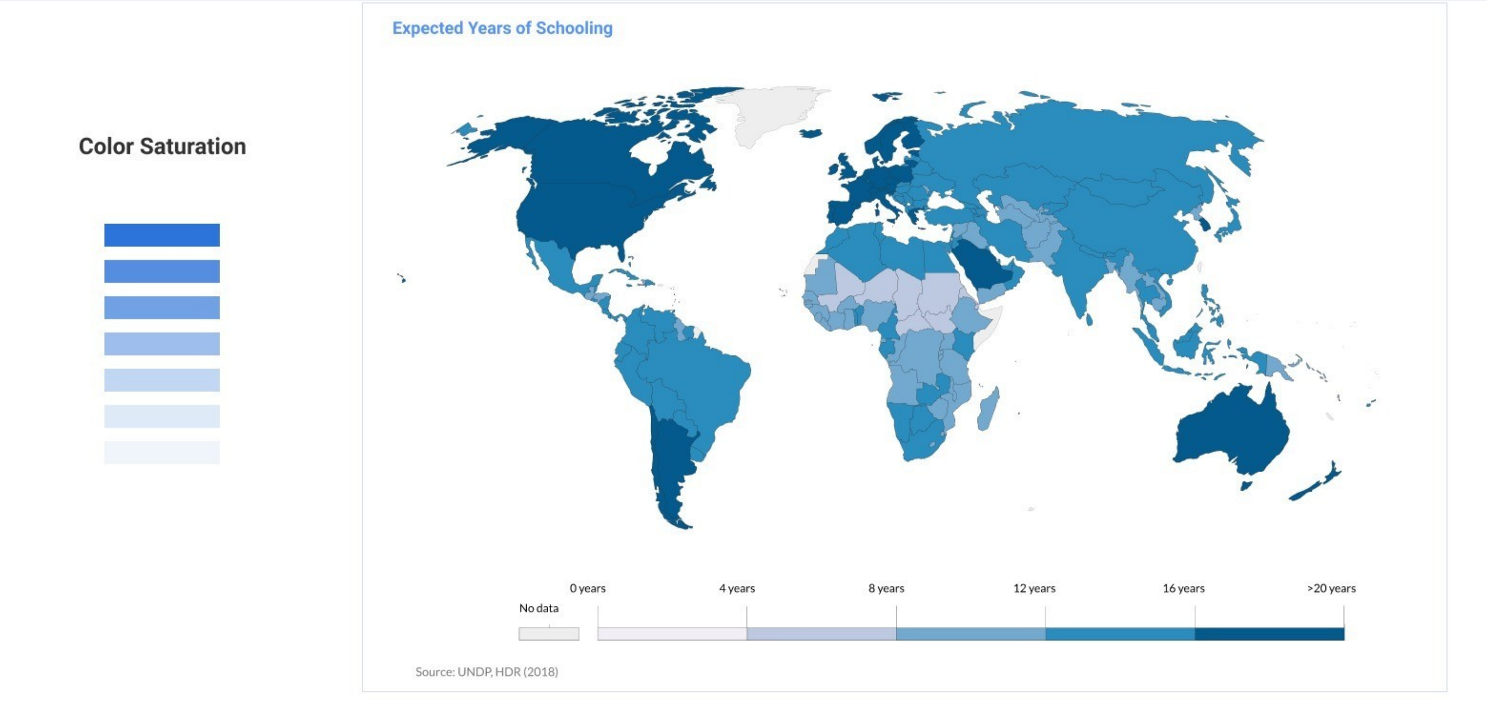


Tetiana Donska

Visual elements & human perception

7. Colour intensity

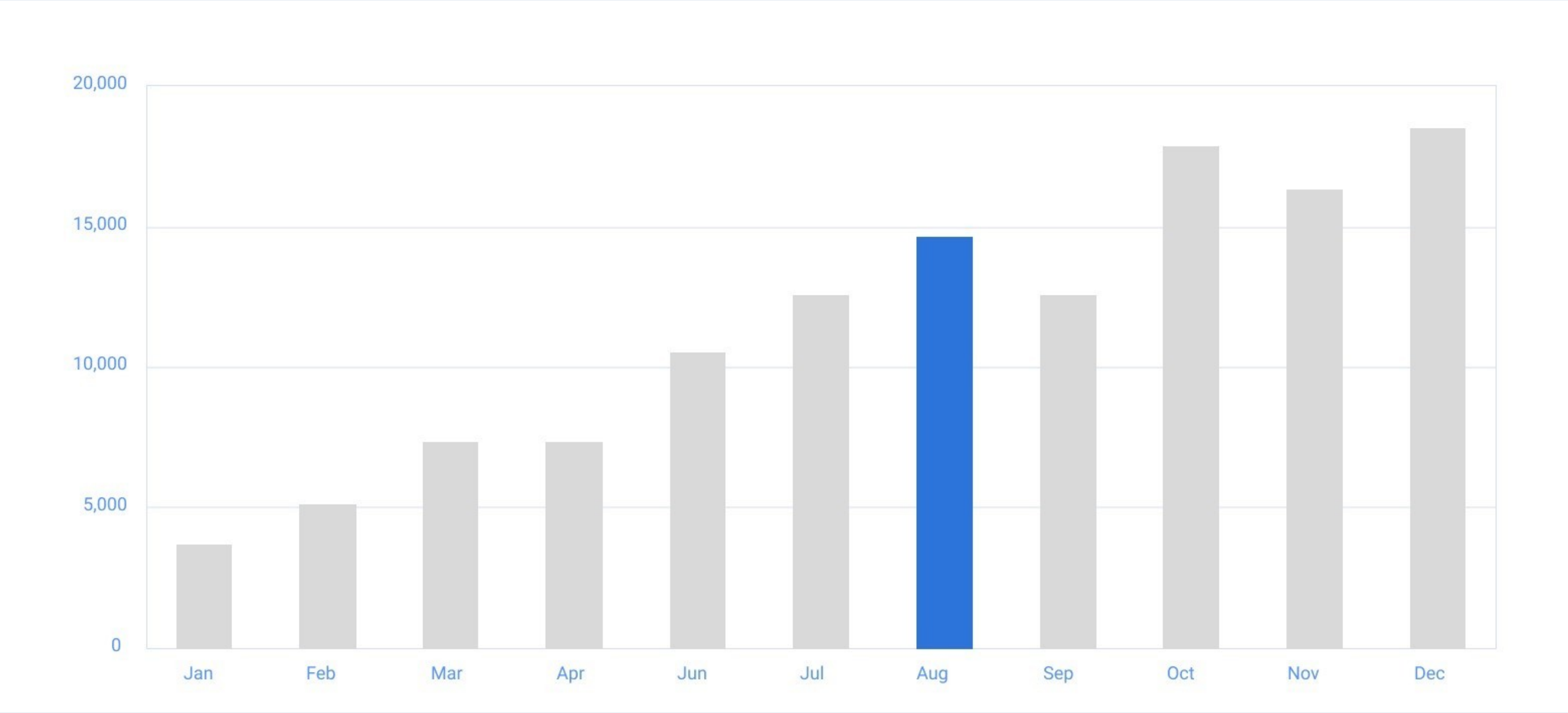
increasing intensities of colour can be perceived intuitively as numbers of increasing value but it’s hard to evaluate the results precisely.



Source: UNDP, HDR (2018)

Visual elements & human perception

- Colour can be used to highlight and help tell a story
- Choice palette is critical, use integrated palettes (e.g., ggplot2, colorbrewer2.org for mapping pallets)



Learning objectives

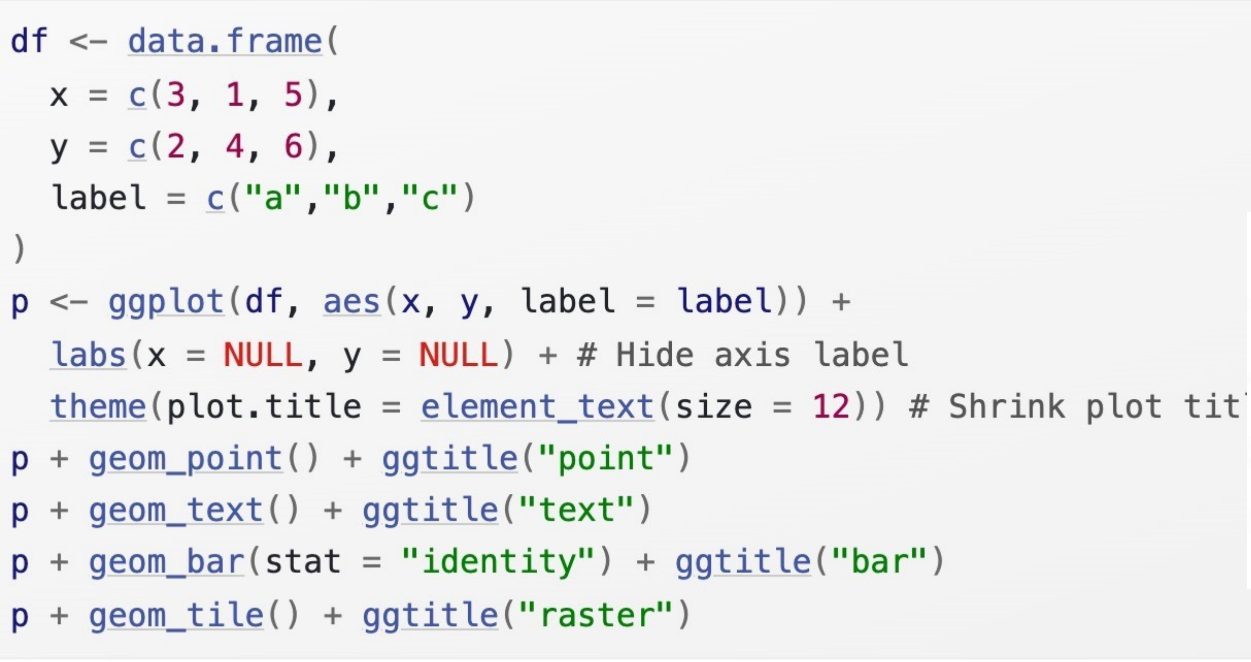
- Do basic plotting of tabular data within R/R-Studio
- Describe when to use scatterplots, line plots, bar plots, and histograms
- Understand key aspects of ggplot objects
  - Aesthetic mappings
  - Geometric objects
  - Scales
- Generate data visualization outputs at appropriate resolutions and in appropriate file formats

Grammar or Graphics

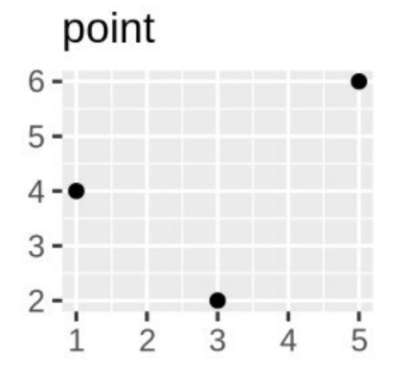
- All plots are composed of the **data**, the **information you want to visualise**, and a **mapping**, the *description of how the data's variables are mapped to aesthetic attributes*
  - layer – collection of geometric elements ('geom') and statistical transformations ('stat')
  - scales – map values in data space to values in plotting space
  - coord – how data coordinates are mapped to the plane of the graphic
  - facet – how to break up data/plots into multiple sub-plots
  - theme – controls finer points of display such as font size, background colours, etc. Themes can be groups of default values for these into cohesive sets of values

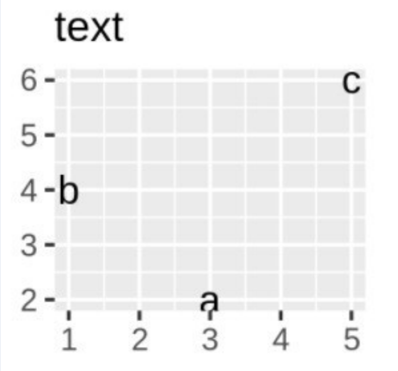
ggplot2 basics

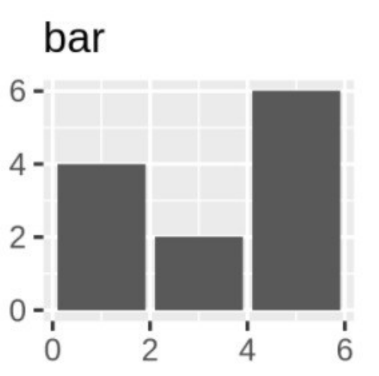
- Individual geoms – obs/geom

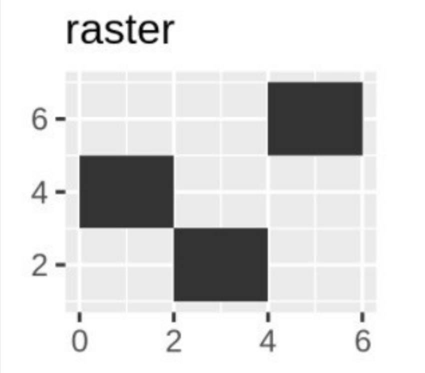


```r
df <- data.frame(
  x = c(3, 1, 5),
  y = c(2, 4, 6),
  label = c("a","b","c")
)
p <- ggplot(df, aes(x, y, label = label)) +
  labs(x = NULL, y = NULL) + # Hide axis label
  theme(plot.title = element_text(size = 12)) # Shrink plot tit
p + geom_point() + ggtitle("point")
p + geom_text() + ggtitle("text")
p + geom_bar(stat = "identity") + ggtitle("bar")
p + geom_tile() + ggtitle("raster")
```

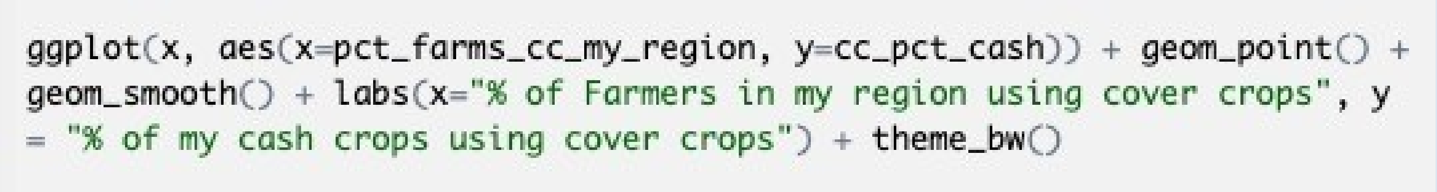




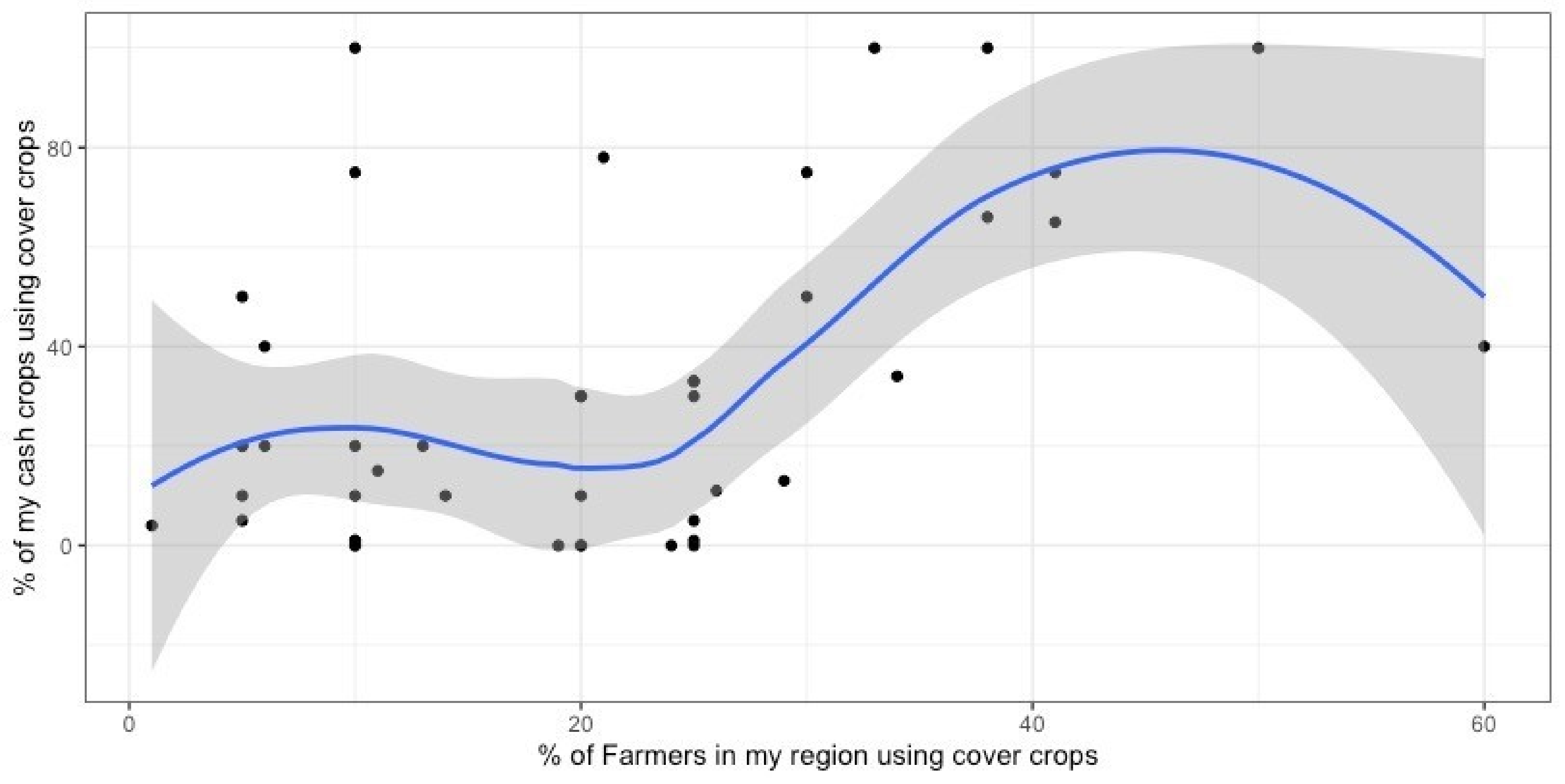




ggplot2 applied



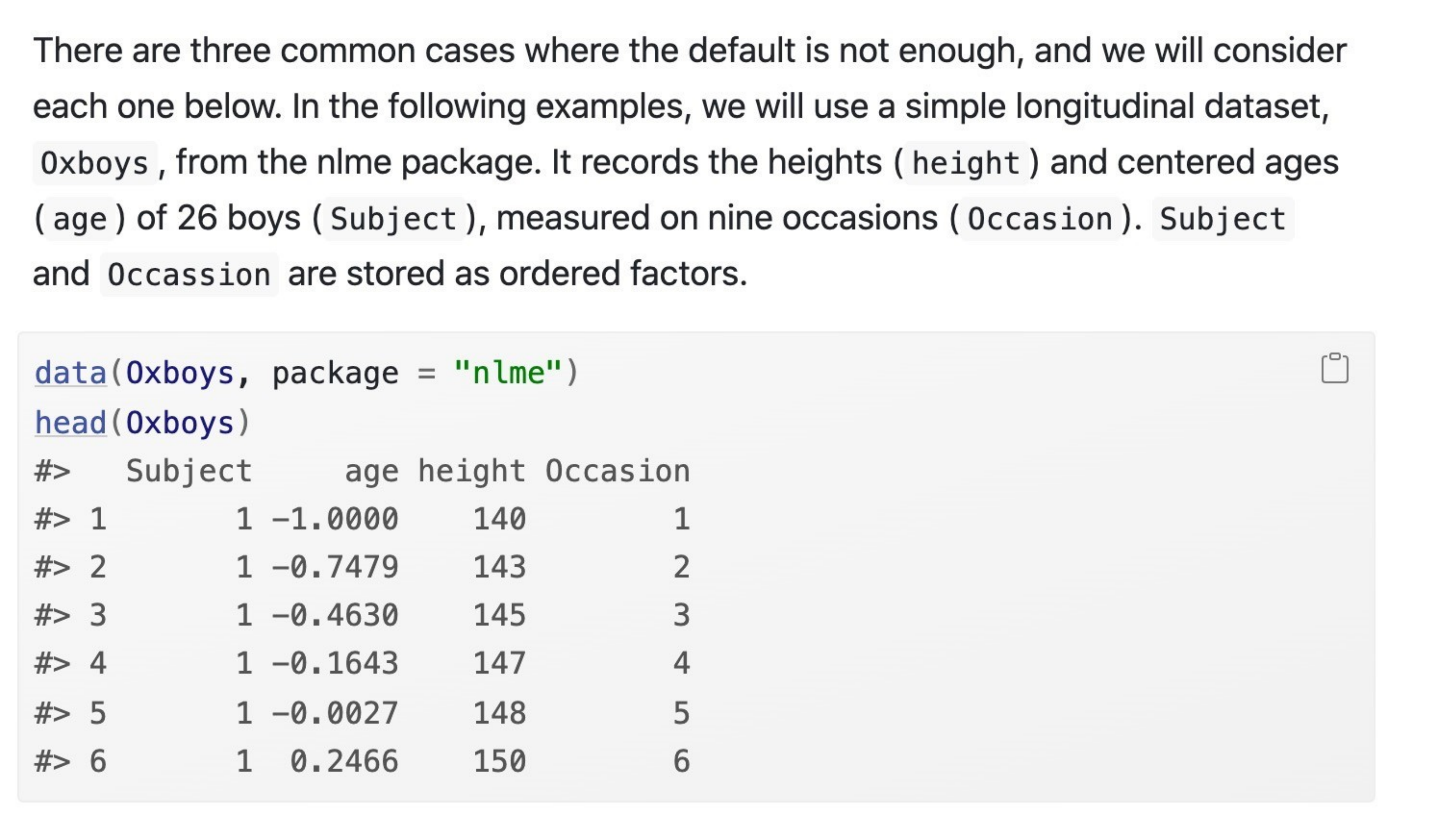
```
ggplot(x, aes(x=pct_farms_cc_my_region, y=cc_pct_cash)) + geom_point() +
geom_smooth() + labs(x="% of Farmers in my region using cover crops", y
= "% of my cash crops using cover crops") + theme_bw()
```



ggplot2 basics

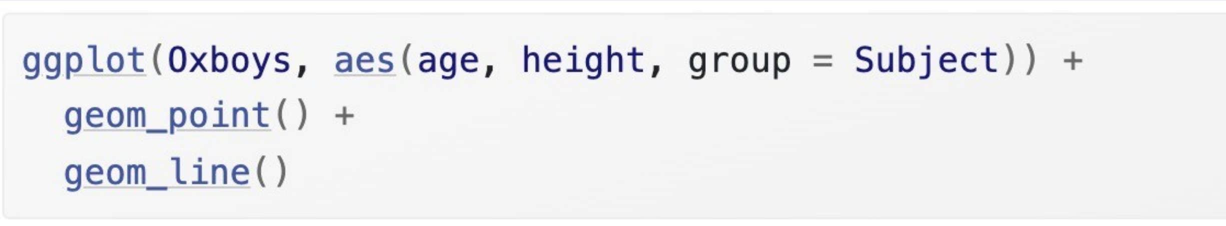
- Collective geoms – multiple observations per geometric element on the graph
  - statistical summaries – such as a boxplot or bar graph
- Grouping structure needed to associate individual

  observations (i.e., rows in a data frame) to geometric elements
  - default is by each level or value of a discrete variable

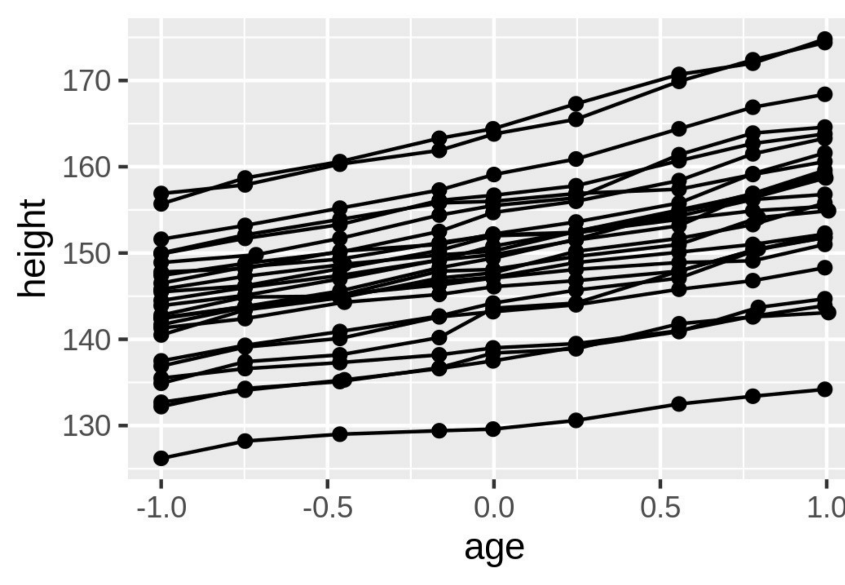


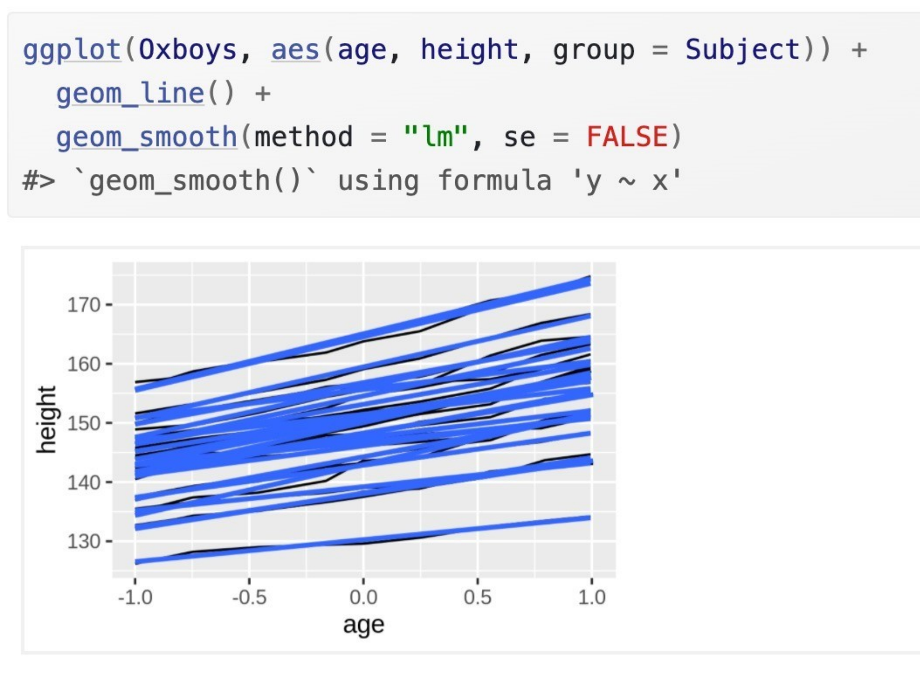
There are three common cases where the default is not enough, and we will consider each one below. In the following examples, we will use a simple longitudinal dataset, `Oxboys`, from the nlme package. It records the heights (`height`) and centered ages (`age`) of 26 boys (`Subject`), measured on nine occasions (`Occasion`). `Subject` and `Occassion` are stored as ordered factors.

```
data(Oxboys, package = "nlme")
head(Oxboys)
#>   Subject     age height Occasion
#> 1       1 -1.0000    140        1
#> 2       1 -0.7479    143        2
#> 3       1 -0.4630    145        3
#> 4       1 -0.1643    147        4
#> 5       1 -0.0027    148        5
#> 6       1  0.2466    150        6
```

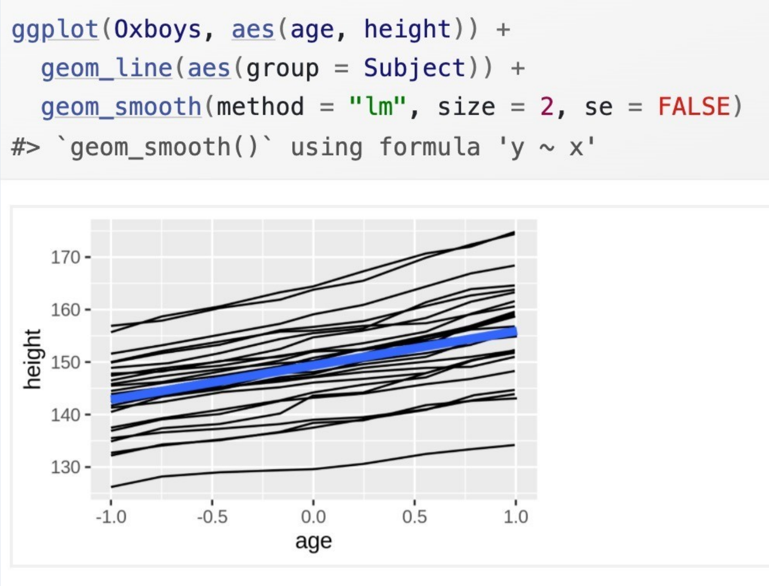


```r
ggplot(Oxboys, aes(age, height, group = Subject)) +
  geom_point() +
  geom_line()
```





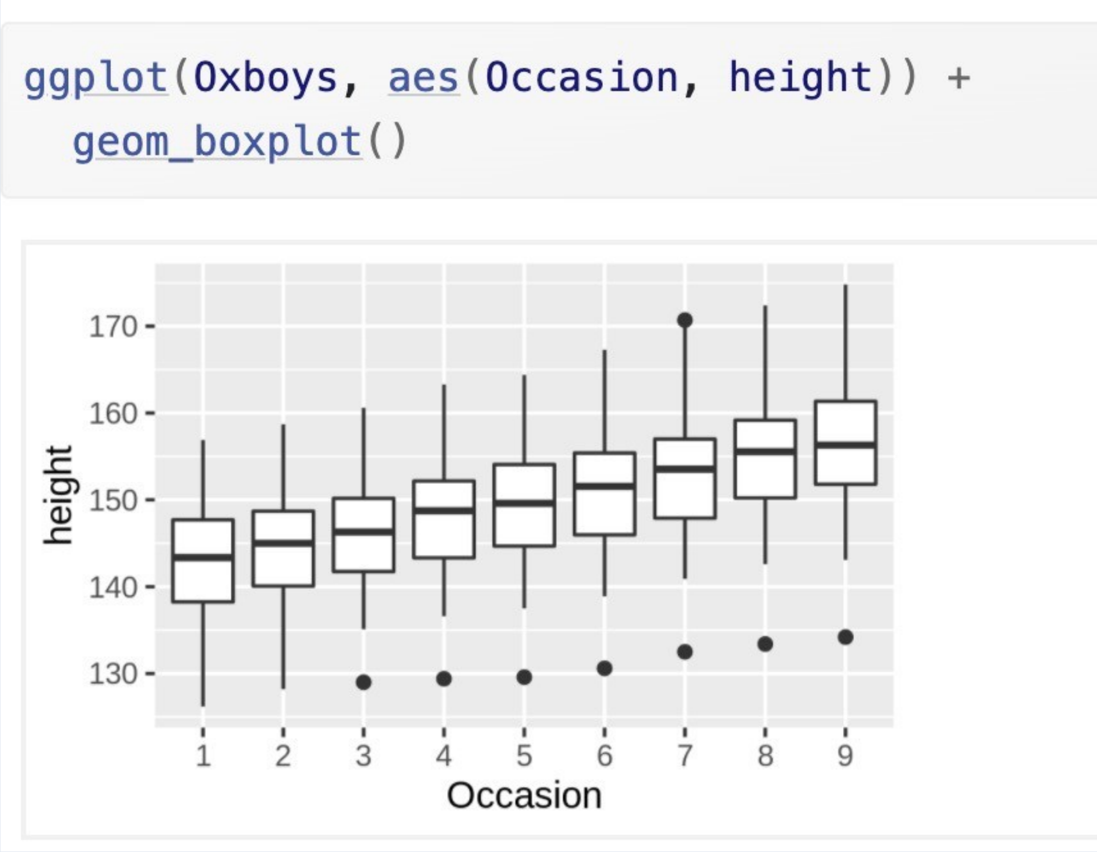
```r
ggplot(Oxboys, aes(age, height, group = Subject)) +
  geom_line() +
  geom_smooth(method = "lm", se = FALSE)
#> `geom_smooth()` using formula 'y ~ x'
```



```r
ggplot(Oxboys, aes(age, height)) +
  geom_line(aes(group = Subject)) +
  geom_smooth(method = "lm", size = 2, se = FALSE)
#> `geom_smooth()` using formula 'y ~ x'
```

ggplot2 basics

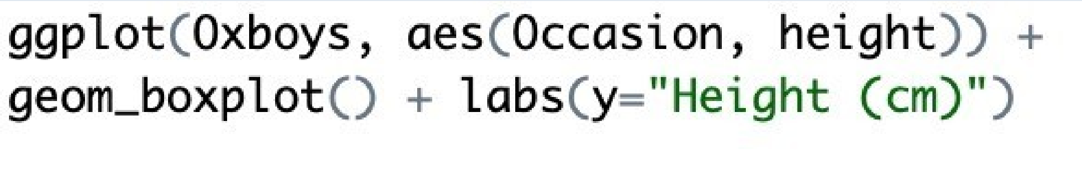
- What is wrong with the labelling in this basic boxplot?
- What could we do to correct it?



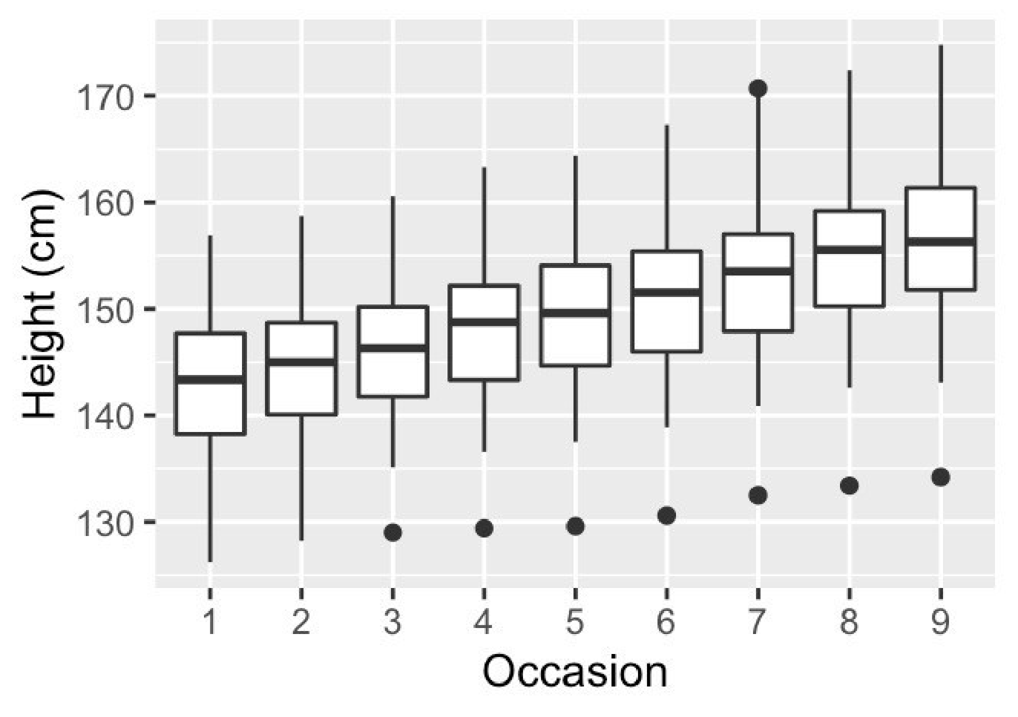
```
ggplot(Oxboys, aes(Occasion, height)) +
  geom_boxplot()
```

ggplot2 basics

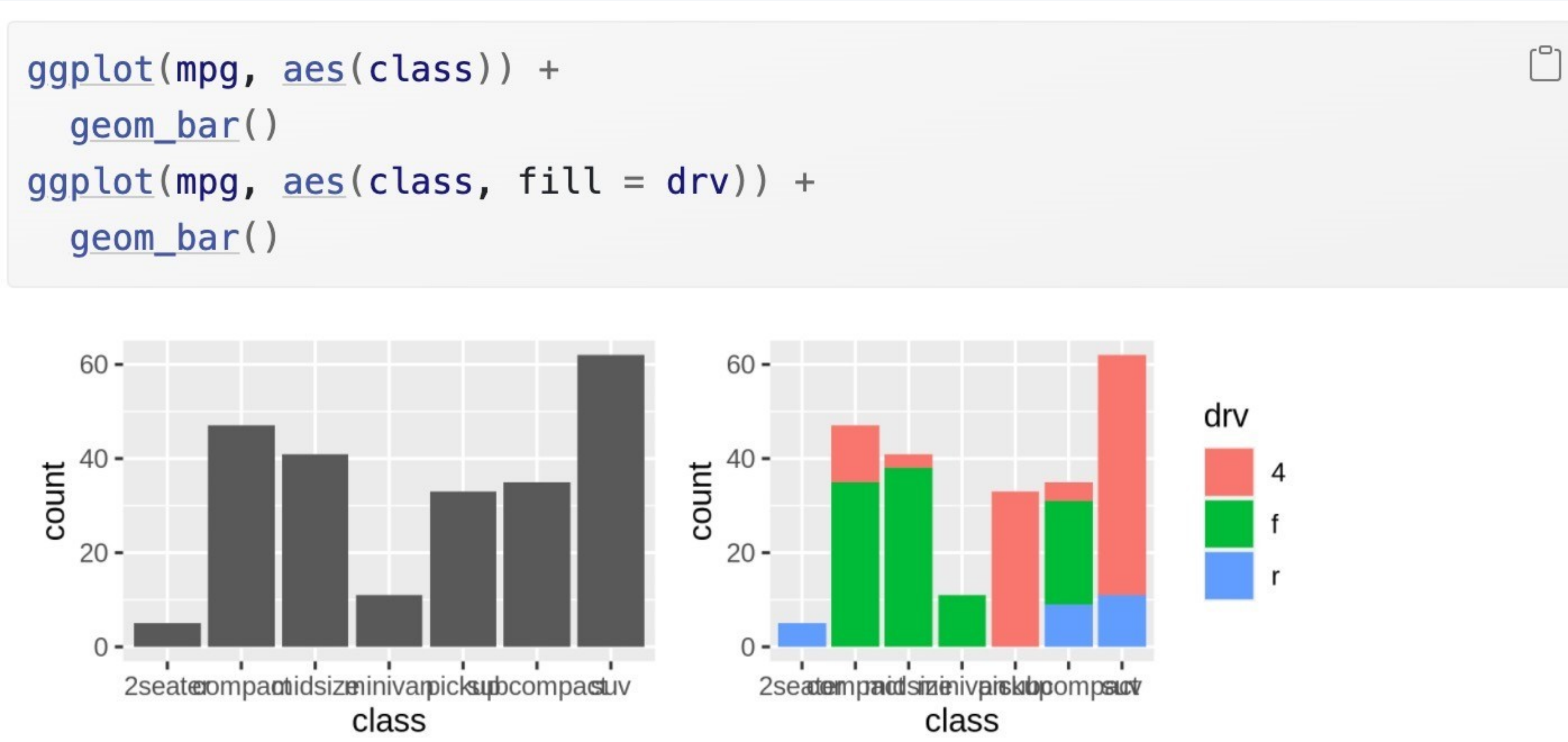
- What is wrong with the labelling in this basic boxplot?
- What could we do to correct it?



```
ggplot(Oxboys, aes(Occasion, height)) +
geom_boxplot() + labs(y="Height (cm)")
```



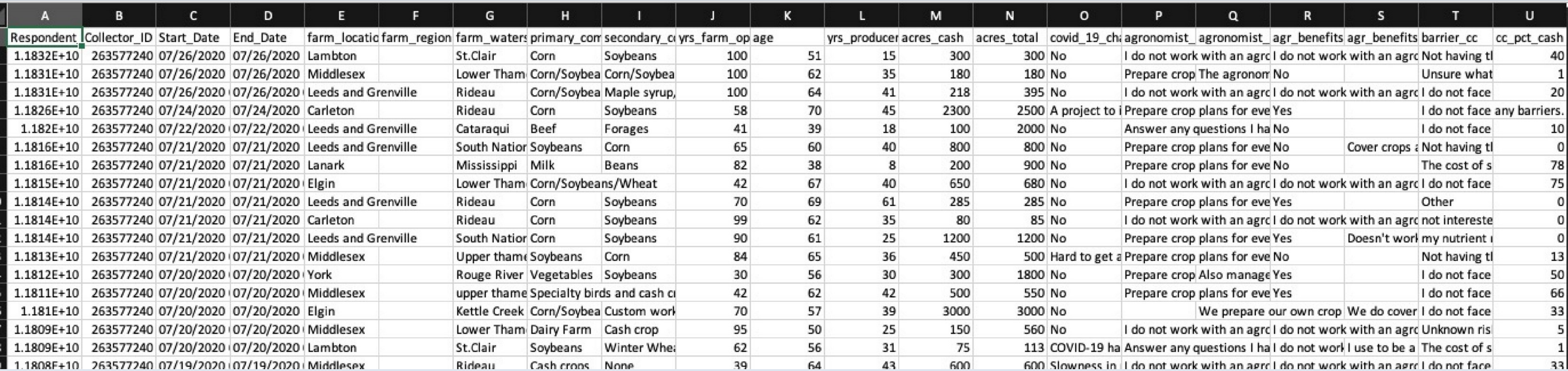
Aesthetics & collective geoms



```
ggplot(mpg, aes(class)) +
  geom_bar()
ggplot(mpg, aes(class, fill = drv)) +
  geom_bar()
```

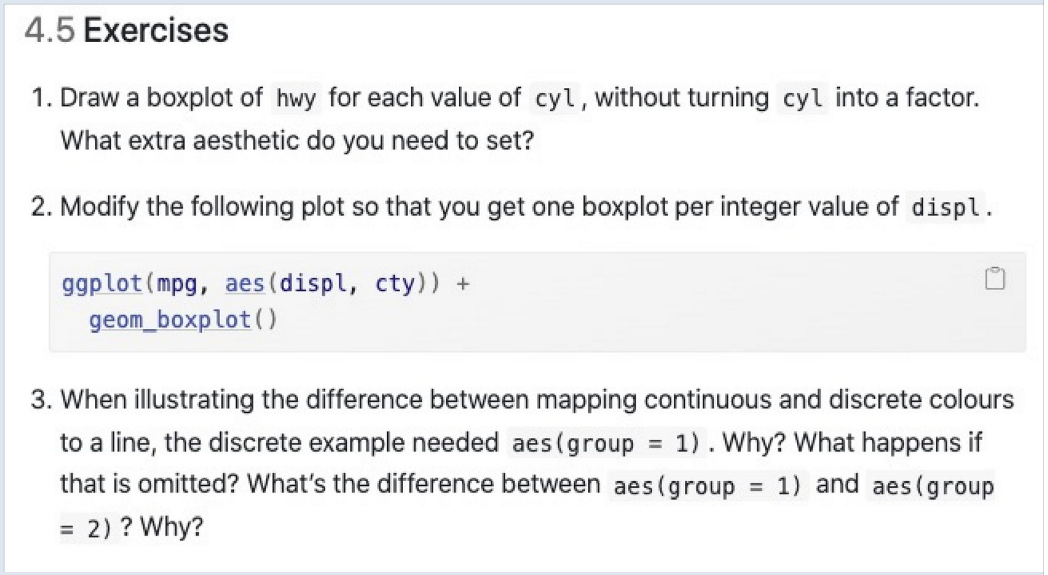
ggplot2 applied

- Farmer survey data on environmental / agricultural practices



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Respondent_	Collector_ID	Start_Date	End_Date	farm_locatio	farm_region	farm_waters	primary_com	secondary_c	yrs_farm_op	age	yrs_producer	acres_cash	acres_total	covid_19_cha	agronomist_	agronomist_	agr_benefits	agr_benefits	barrier_cc	cc_pct_cash
1.1832E+10	263577240	07/26/2020	07/26/2020	Lambton		St.Clair	Corn	Soybeans	100	51	15	300	300	No	I do not work with an agrc		I do not work with an agrc		Not having tl	40
1.1831E+10	263577240	07/26/2020	07/26/2020	Middlesex		Lower Tham	Corn/Soybea	Corn/Soybea	100	62	35	180	180	No	Prepare crop	The agronom	No		Unsure what	1
1.1831E+10	263577240	07/26/2020	07/26/2020	Leeds and Grenville		Rideau	Corn/Soybea	Maple syrup,	100	64	41	218	395	No	I do not work with an agrc		I do not work with an agrc		I do not face	20
1.1826E+10	263577240	07/24/2020	07/24/2020	Carleton		Rideau	Corn	Soybeans	58	70	45	2300	2500	A project to i	Prepare crop plans for eve		Yes		I do not face any barriers.	
1.182E+10	263577240	07/22/2020	07/22/2020	Leeds and Grenville		Cataraqui	Beef	Forages	41	39	18	100	2000	No	Answer any questions I ha		No		I do not face	10
1.1816E+10	263577240	07/21/2020	07/21/2020	Leeds and Grenville		South Natior	Soybeans	Corn	65	60	40	800	800	No	Prepare crop plans for eve		No	Cover crops a	Not having tl	0
1.1816E+10	263577240	07/21/2020	07/21/2020	Lanark		Mississippi	Milk	Beans	82	38	8	200	900	No	Prepare crop plans for eve		No		The cost of s	78
1.1815E+10	263577240	07/21/2020	07/21/2020	Elgin		Lower Tham	Corn/Soybeans/Wheat		42	67	40	650	680	No	I do not work with an agrc		I do not work with an agrc		I do not face	75
1.1814E+10	263577240	07/21/2020	07/21/2020	Leeds and Grenville		Rideau	Corn	Soybeans	70	69	61	285	285	No	Prepare crop plans for eve		Yes		Other	0
1.1814E+10	263577240	07/21/2020	07/21/2020	Carleton		Rideau	Corn	Soybeans	99	62	35	80	85	No	I do not work with an agrc		I do not work with an agrc		not intereste	0
1.1814E+10	263577240	07/21/2020	07/21/2020	Leeds and Grenville		South Natior	Corn	Soybeans	90	61	25	1200	1200	No	Prepare crop plans for eve		Yes	Doesn't work	my nutrient i	0
1.1813E+10	263577240	07/21/2020	07/21/2020	Middlesex		Upper thame	Soybeans	Corn	84	65	36	450	500	Hard to get a	Prepare crop plans for eve		No		Not having tl	13
1.1812E+10	263577240	07/20/2020	07/20/2020	York		Rouge River	Vegetables	Soybeans	30	56	30	300	1800	No	Prepare crop	Also manage	Yes		I do not face	50
1.1811E+10	263577240	07/20/2020	07/20/2020	Middlesex		upper thame	Specialty birds and cash cı		42	62	42	500	550	No	Prepare crop plans for eve		Yes		I do not face	66
1.181E+10	263577240	07/20/2020	07/20/2020	Elgin		Kettle Creek	Corn/Soybea	Custom work	70	57	39	3000	3000	No		We prepare our own crop		We do cover	I do not face	33
1.1809E+10	263577240	07/20/2020	07/20/2020	Middlesex		Lower Tham	Dairy Farm	Cash crop	95	50	25	150	560	No	I do not work with an agrc		I do not work with an agrc		Unknown ris	5
1.1809E+10	263577240	07/20/2020	07/20/2020	Lambton		St.Clair	Soybeans	Winter Whea	62	56	31	75	113	COVID-19 ha	Answer any questions I ha		I do not work	I use to be a	The cost of s	1
1.1808E+10	263577240	07/19/2020	07/19/2020	Middlesex		Rideau	Cash crops	None	39	64	43	600	600	Slowness in	I do not work with an agrc		I do not work with an agrc		I do not face	33

Take the next 20 minutes to work on these questions then we will report back on the answers



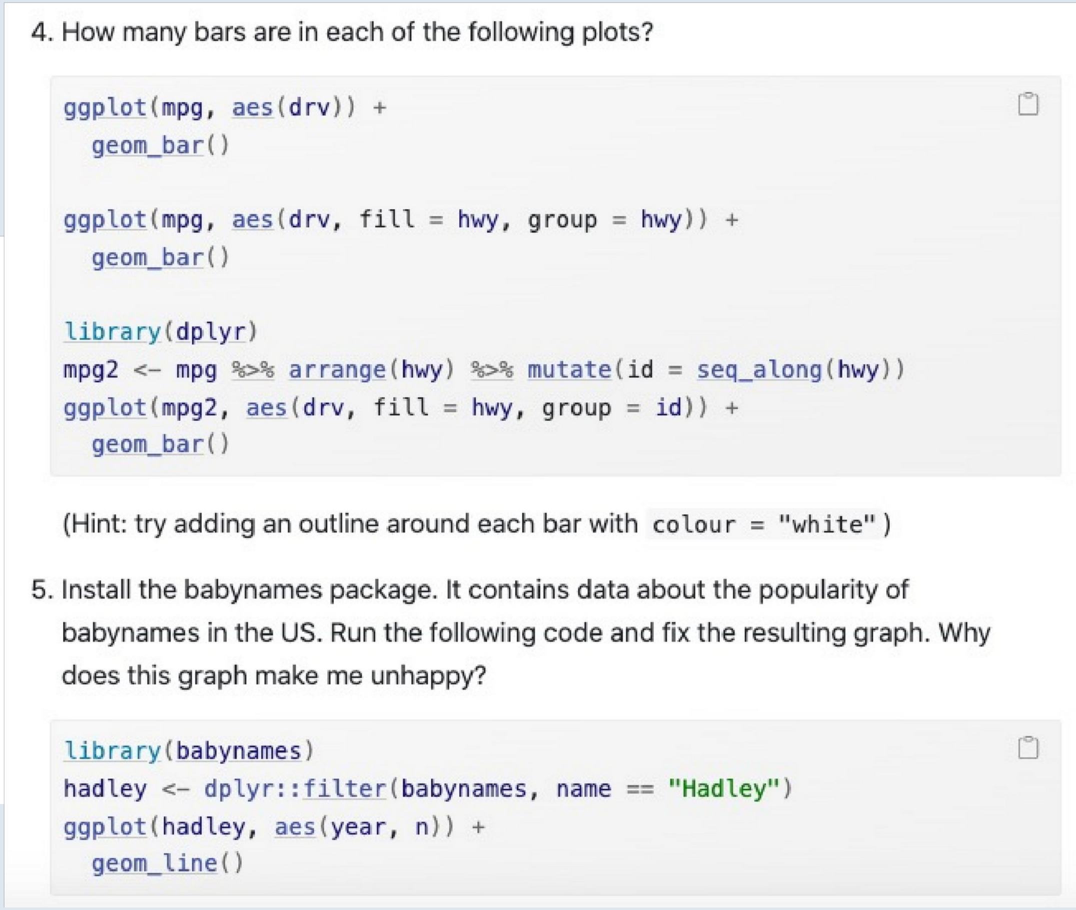
4.5 Exercises

1. Draw a boxplot of `hwy` for each value of `cyl`, without turning `cyl` into a factor. What extra aesthetic do you need to set?

2. Modify the following plot so that you get one boxplot per integer value of `displ`.

```
ggplot(mpg, aes(displ, cty)) +
  geom_boxplot()
```

3. When illustrating the difference between mapping continuous and discrete colours to a line, the discrete example needed `aes(group = 1)`. Why? What happens if that is omitted? What's the difference between `aes(group = 1)` and `aes(group = 2)`? Why?



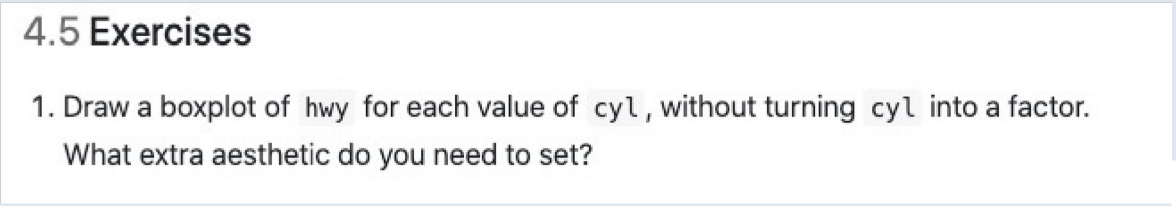
4. How many bars are in each of the following plots?

```
ggplot(mpg, aes(drv)) +
  geom_bar()

ggplot(mpg, aes(drv, fill = hwy, group = hwy)) +
  geom_bar()

library(dplyr)
mpg2 <- mpg %>% arrange(hwy) %>% mutate(id = seq_along(hwy))
ggplot(mpg2, aes(drv, fill = hwy, group = id)) +
  geom_bar()
```

(Hint: try adding an outline around each bar with `colour = "white"`)

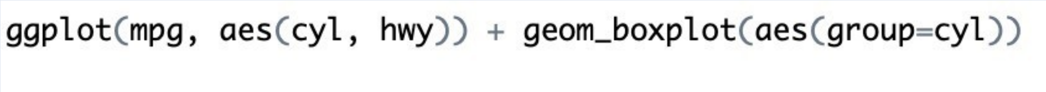
5. Install the babynames package. It contains data about the popularity of babynames in the US. Run the following code and fix the resulting graph. Why does this graph make me unhappy?

```
library(babynames)
hadley <- dplyr::filter(babynames, name == "Hadley")
ggplot(hadley, aes(year, n)) +
  geom_line()
```

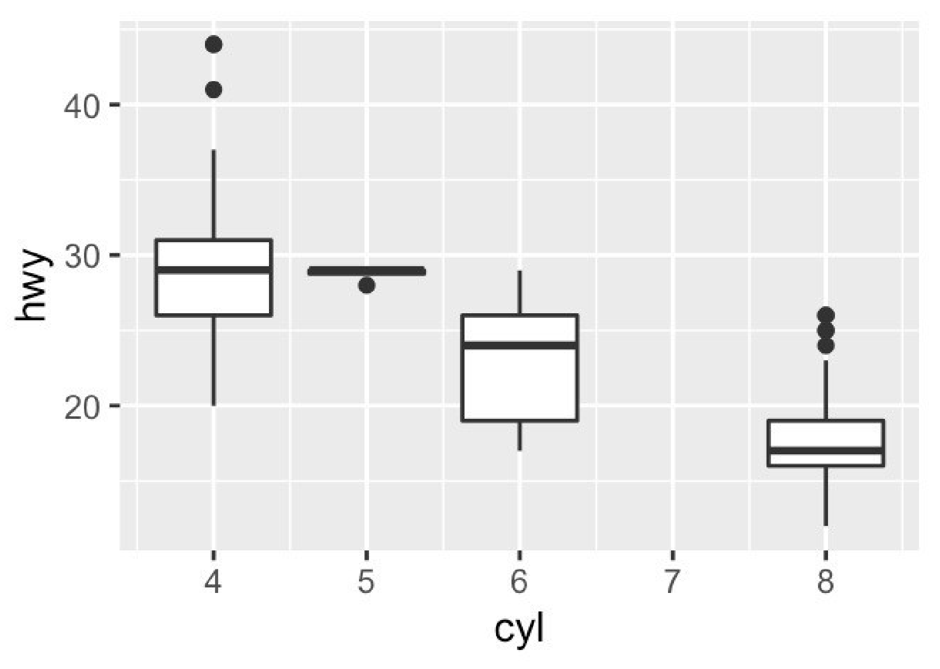


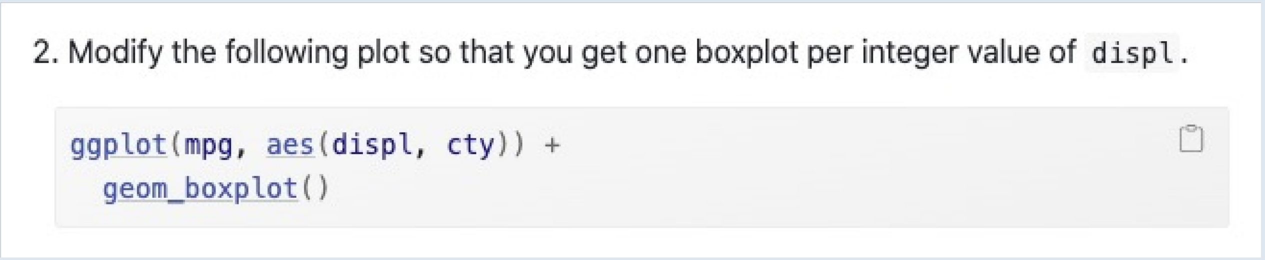
4.5 Exercises

1. Draw a boxplot of `hwy` for each value of `cyl`, without turning `cyl` into a factor. What extra aesthetic do you need to set?



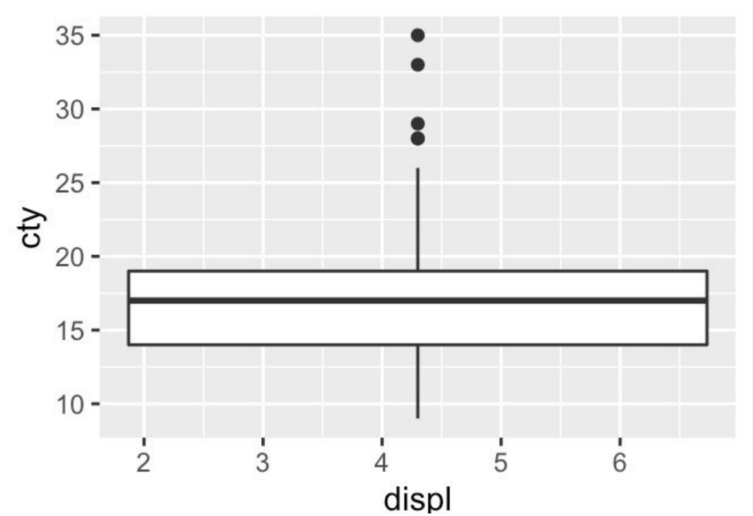
```
ggplot(mpg, aes(cyl, hwy)) + geom_boxplot(aes(group=cyl))
```



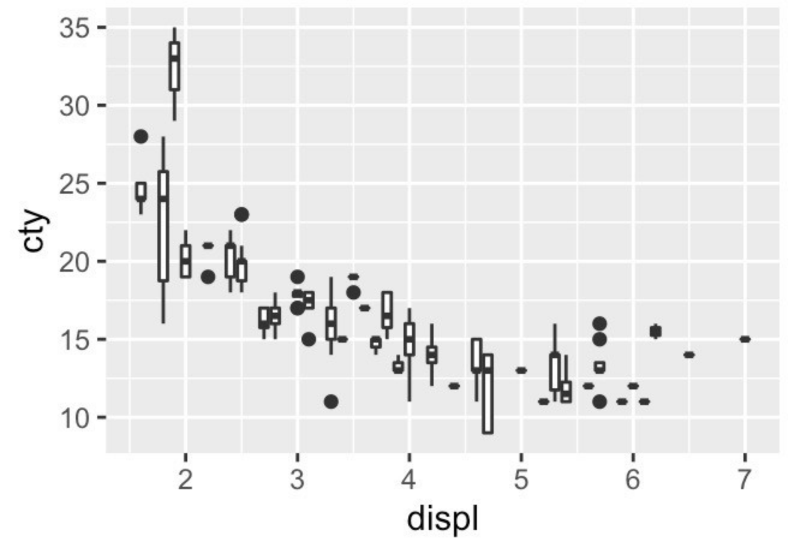


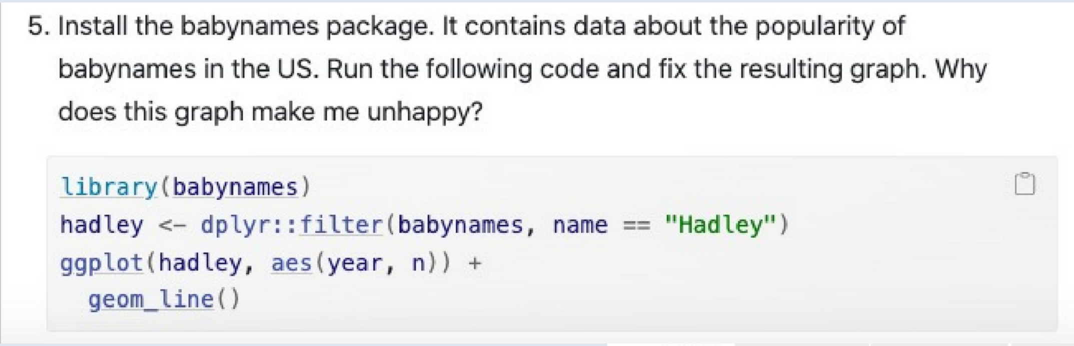
2. Modify the following plot so that you get one boxplot per integer value of `displ`.

```
ggplot(mpg, aes(displ, cty)) +
  geom_boxplot()
```



```
ggplot(mpg, aes(displ, cty, group=displ)) +
  geom_boxplot()
```

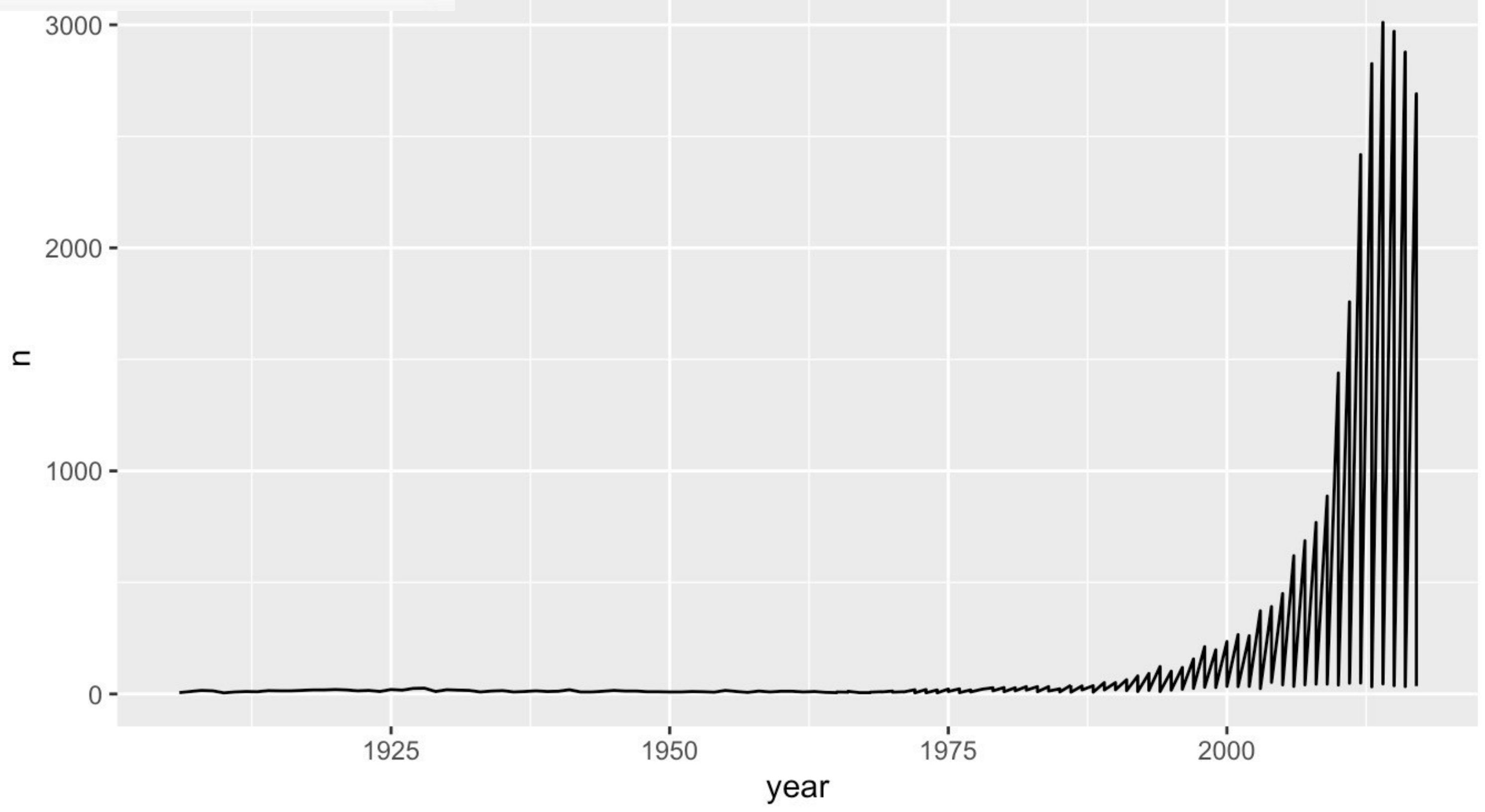


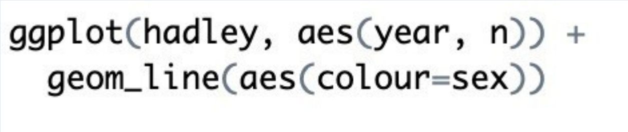


5. Install the babynames package. It contains data about the popularity of babynames in the US. Run the following code and fix the resulting graph. Why does this graph make me unhappy?

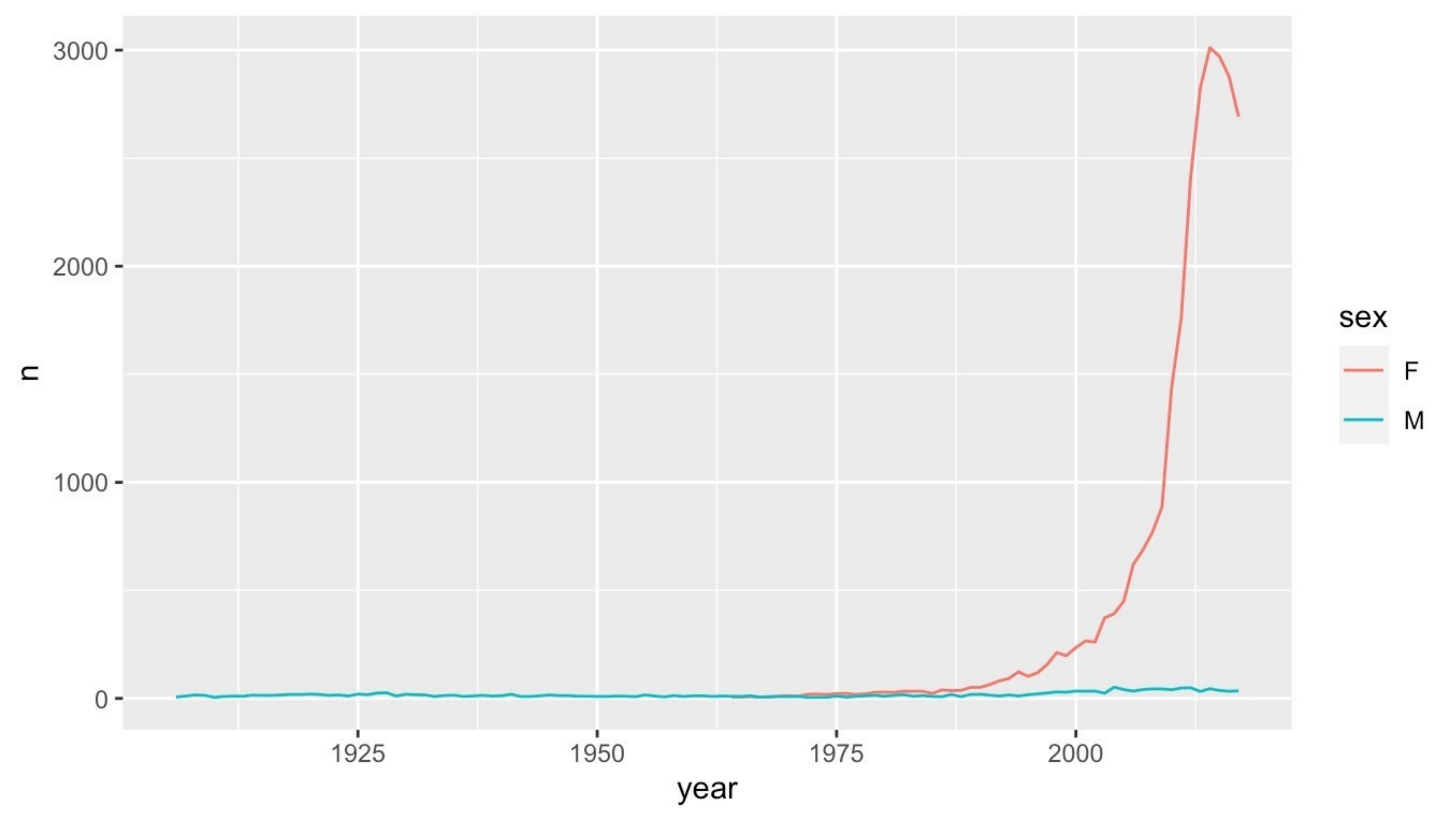
```r
library(babynames)
hadley <- dplyr::filter(babynames, name == "Hadley")
ggplot(hadley, aes(year, n)) +
  geom_line()
```

What is wrong?
How to fix it?

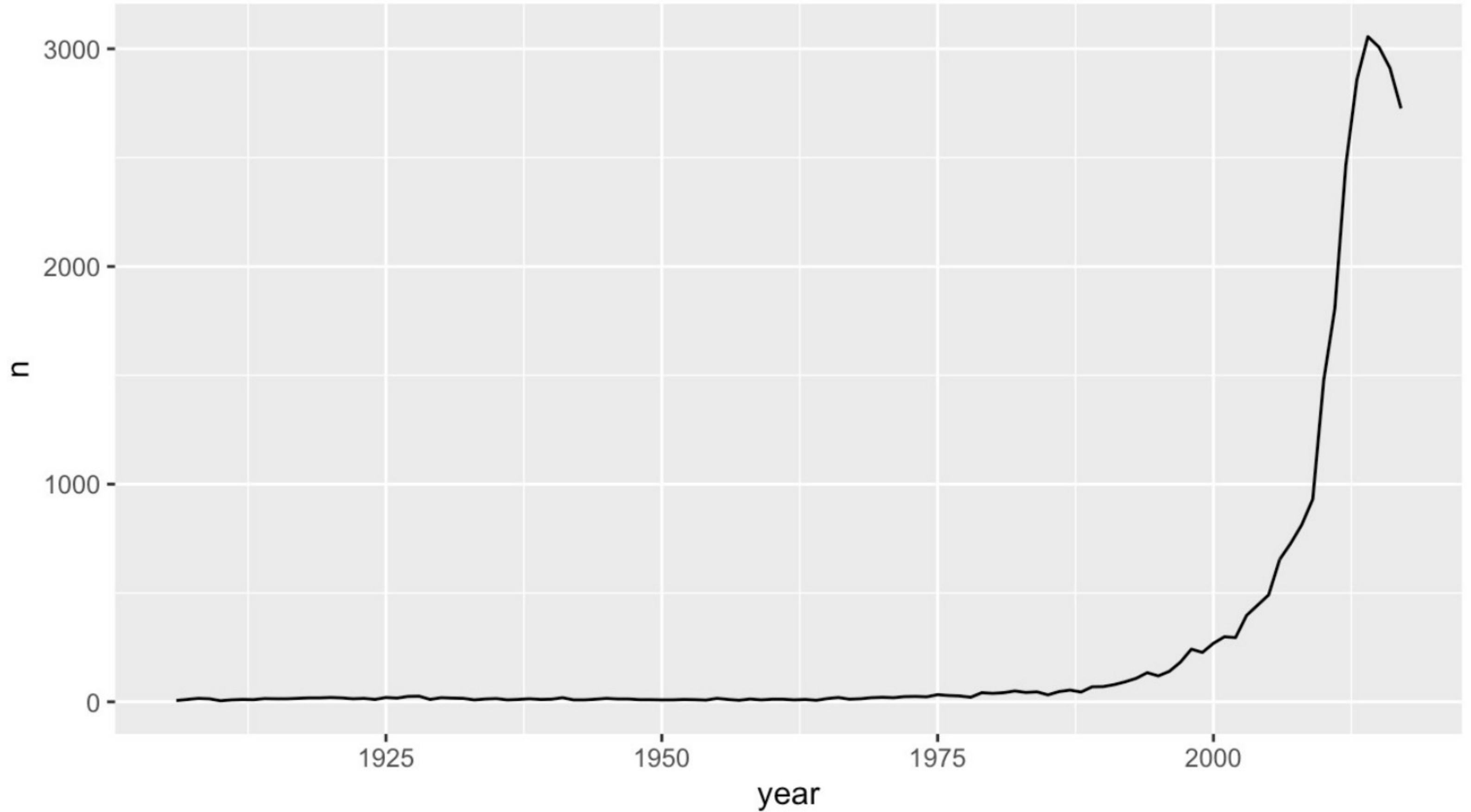




```
ggplot(hadley, aes(year, n)) +
  geom_line(aes(colour=sex))
```

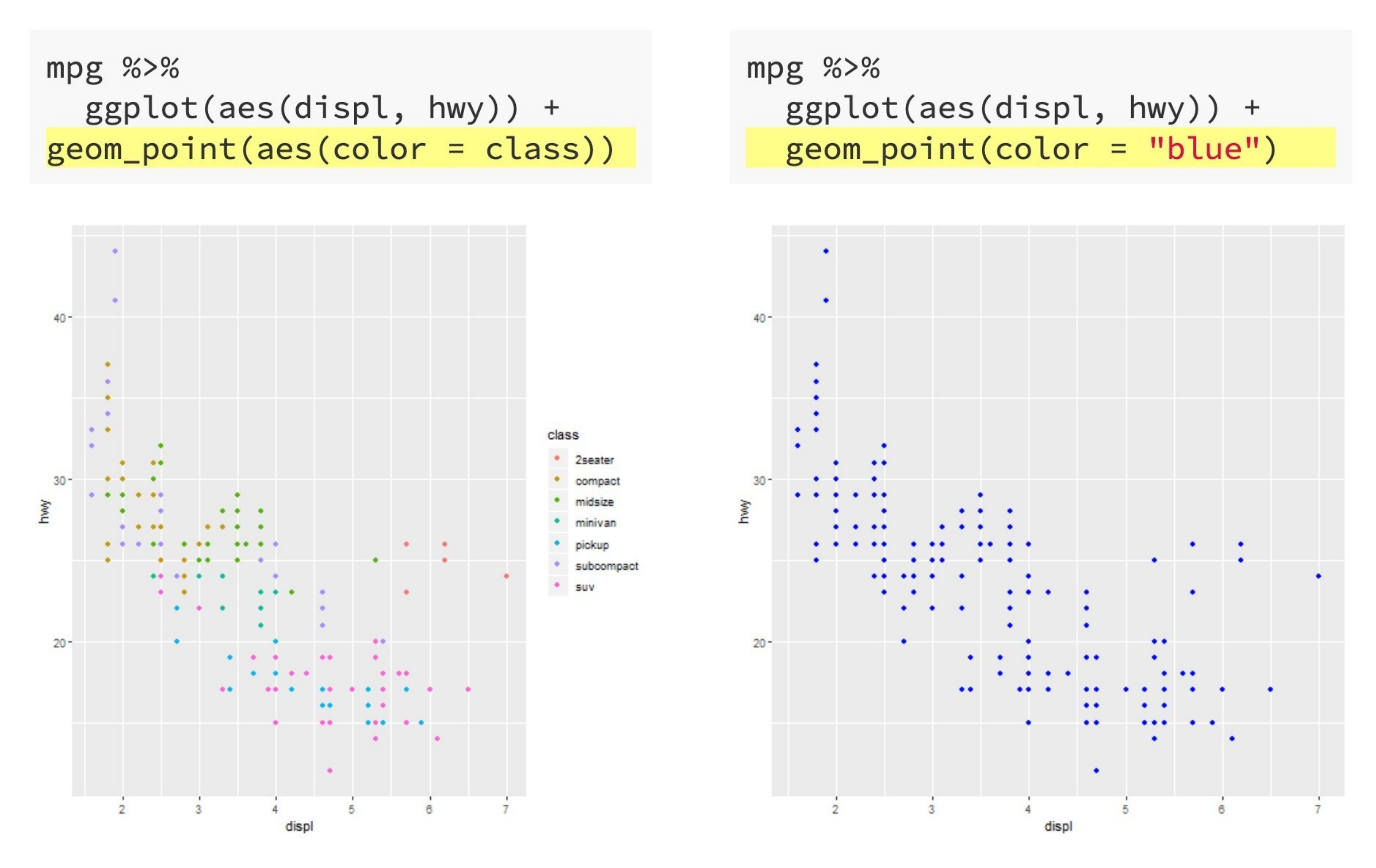


```r
hadley <- filter(babynames, name == "Hadley")  %>% group_by(year) %>% summarise(n=sum(n))

ggplot(hadley, aes(year, n)) +
  geom_line()
```



- Break
- Case Study
- Explore https://rstudio.cloud/learn/primers/

Mapping vs Setting



```
mpg %>%
  ggplot(aes(displ, hwy)) +
geom_point(aes(color = class))
```

```
mpg %>%
  ggplot(aes(displ, hwy)) +
  geom_point(color = "blue")
```

John Little 2020

Graphic outputs

- Vector graphics:
  - .pdf (for publication)
  - .svg (for editing in Illustrator or Inkscape) -> export to PDF
  - Get aspect ratio and relative font size right
- Bitmap graphics:
  - .png (*lossless* compression) for charts and text
  - .jpg (*lossy*, quality 90+) for photos or complex illustrations with tonal gradients
- Minimum DPI for printing of 240, 300-600 preferred
- Minimum DPI of 150 for displaying on screen
- Need to get width and height exactly right since resizing involves interpolation

Summary

- Start simple and build up complexity in your learning
- Use the grammar of graphics features
  - layers, facets, etc.
- Look at the raw data and double check your plots and transformations are doing what you think they are
- Use external resources – there are countless resources free and online for learning ggplot2 visualization