Regression and inference

January 22, 2020

PMAP 8521: Program Evaluation for Public Service
Andrew Young School of Policy Studies
Spring 2020

Fill out your reading report on iCollege!

Plan for today

The magic of `ggplot()`

Drawing lines

Lines, Greek, and regression

Multiple regression

The magic of `ggplot()`

Your turn (#1)

Type this code into the empty chunk under “Your Turn 1” and run it

Pay attention to spelling, capitalization, and parentheses!

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

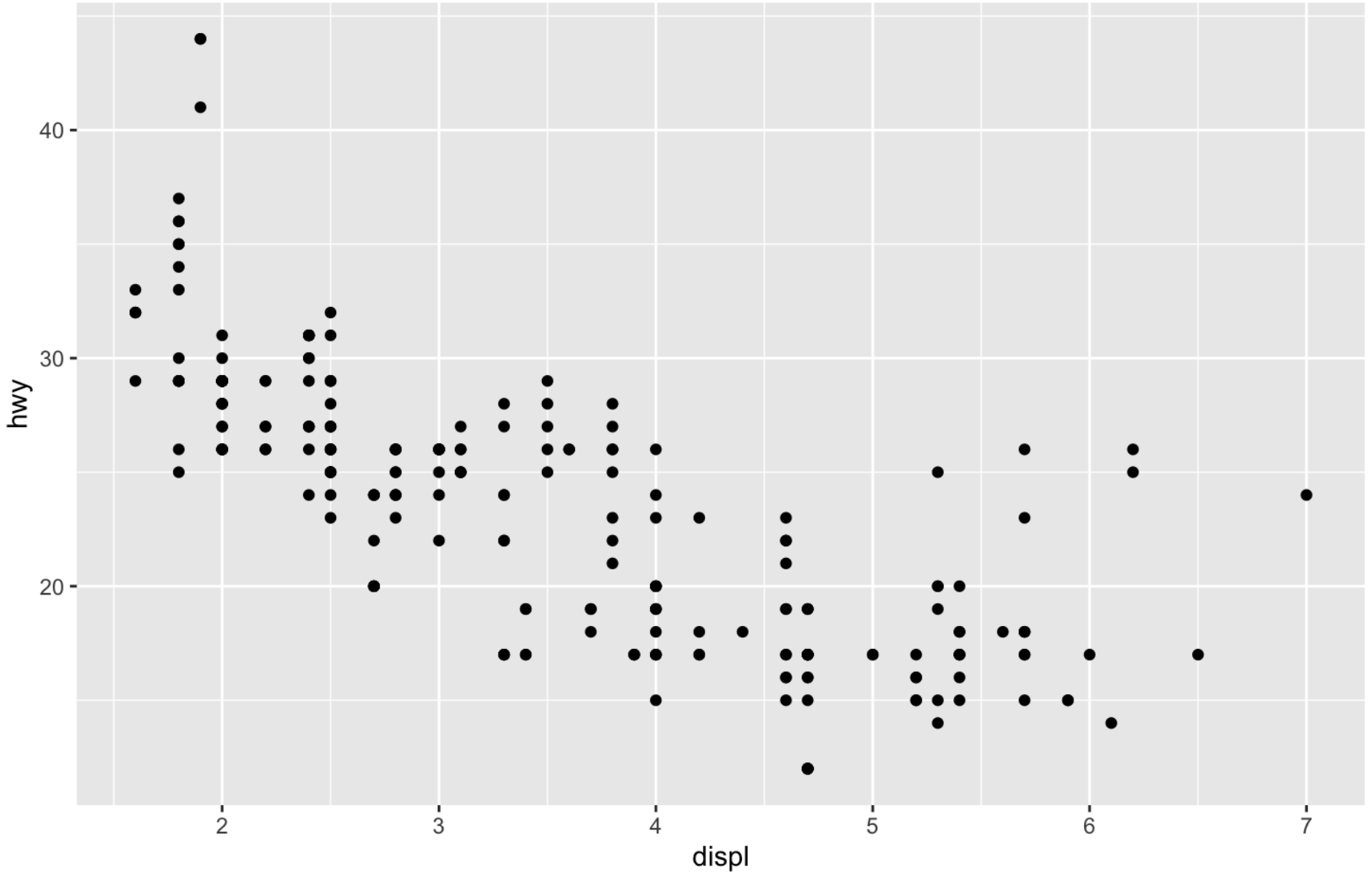
Your turn (#1)

Type this code into the empty chunk under “Your Turn 1” and run it

Pay attention to spelling, capitalization, and parentheses!

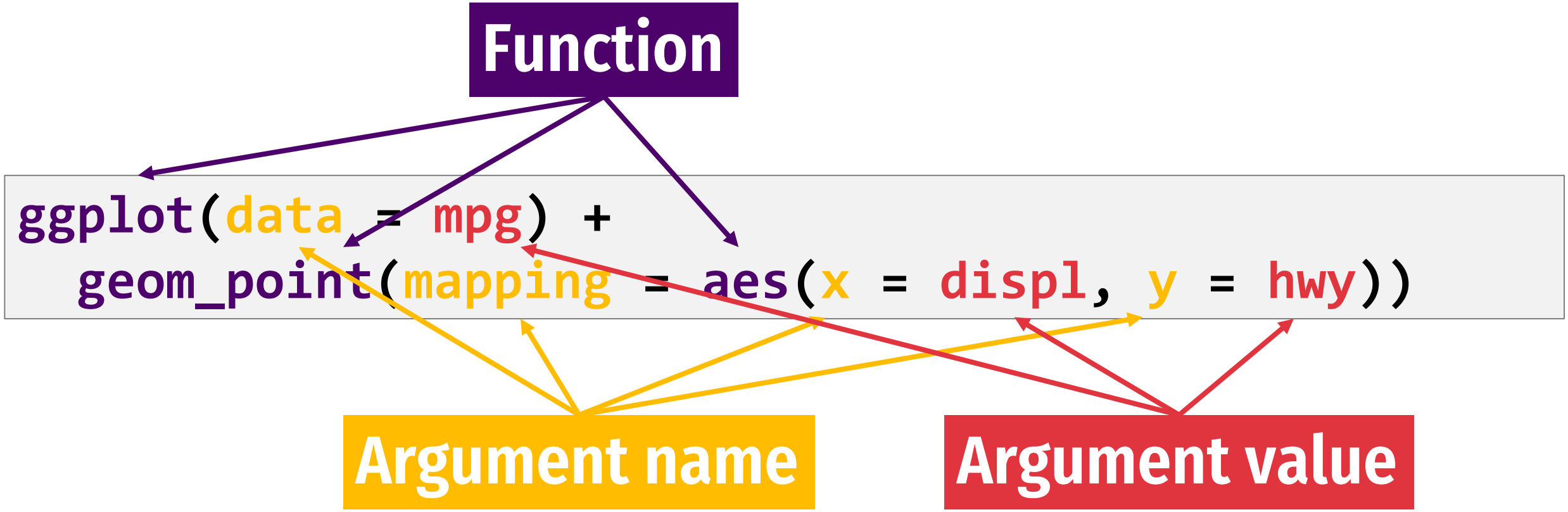
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

02:00




```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
```

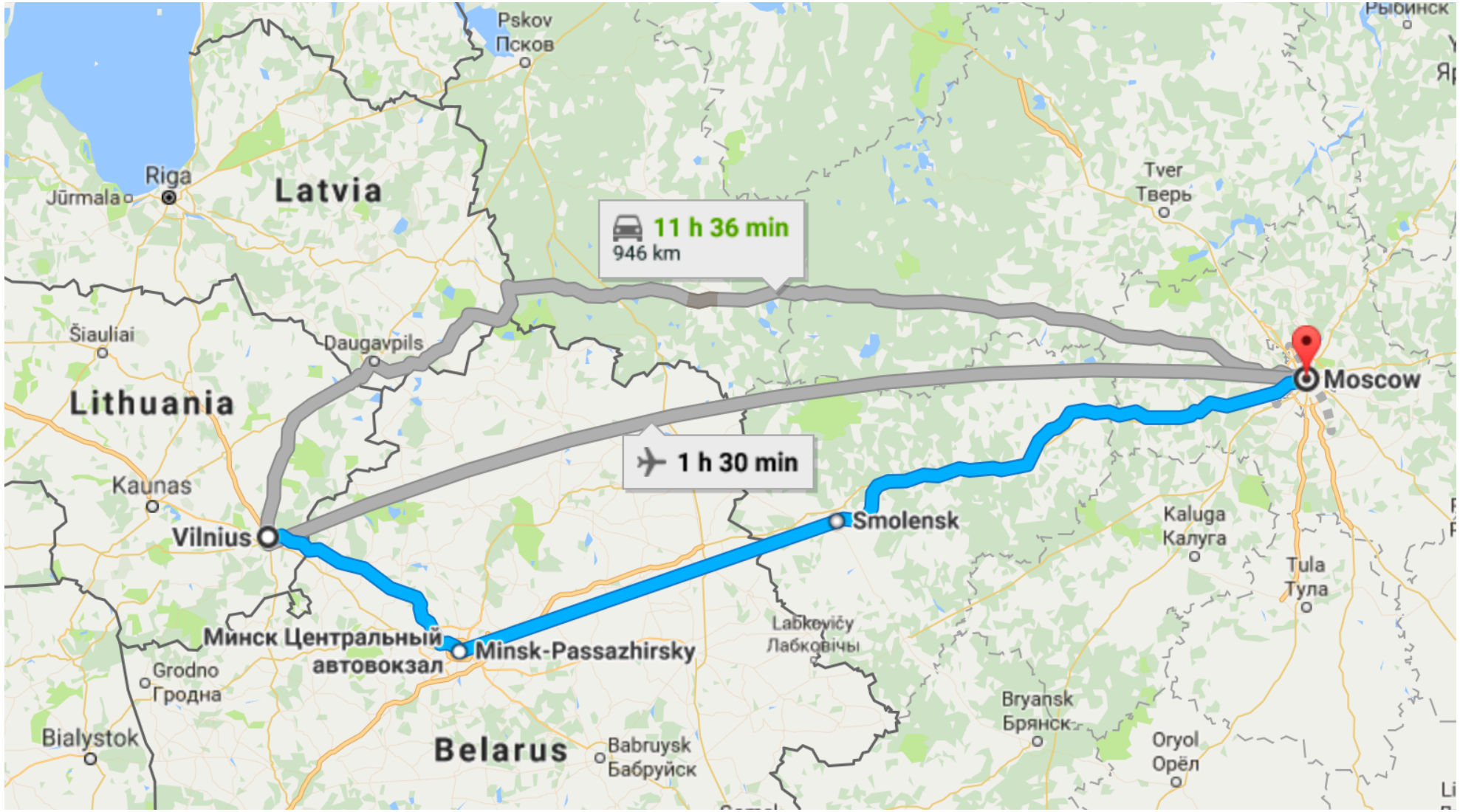

Function
ggplot(data = mpg) +
geom_point(mapping = aes(x = displ, y = hwy))
Argument name
Argument value

ggplot template

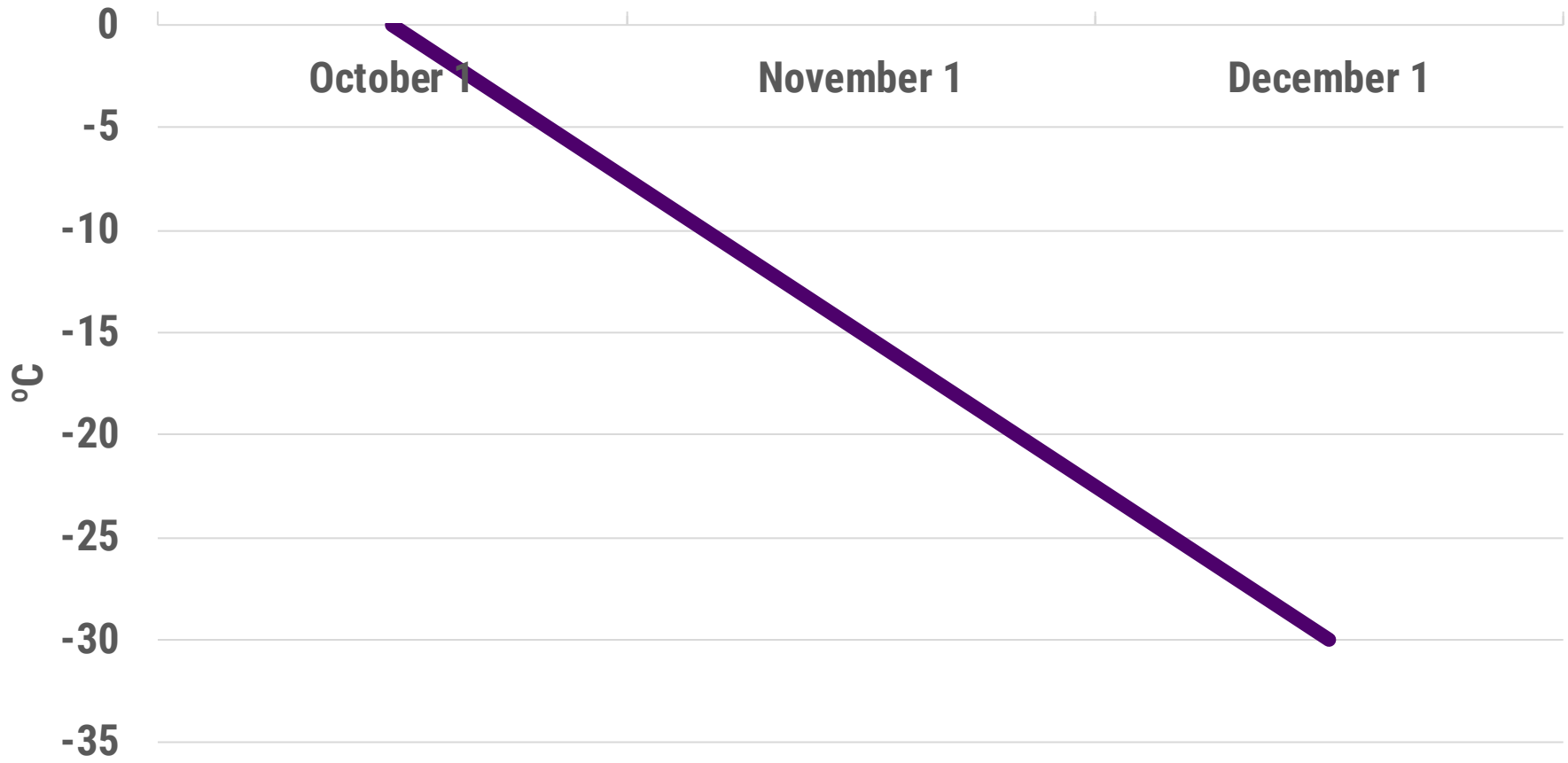
```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

```
  geom_point(mapping = aes(x = displ, y = hwy))
```

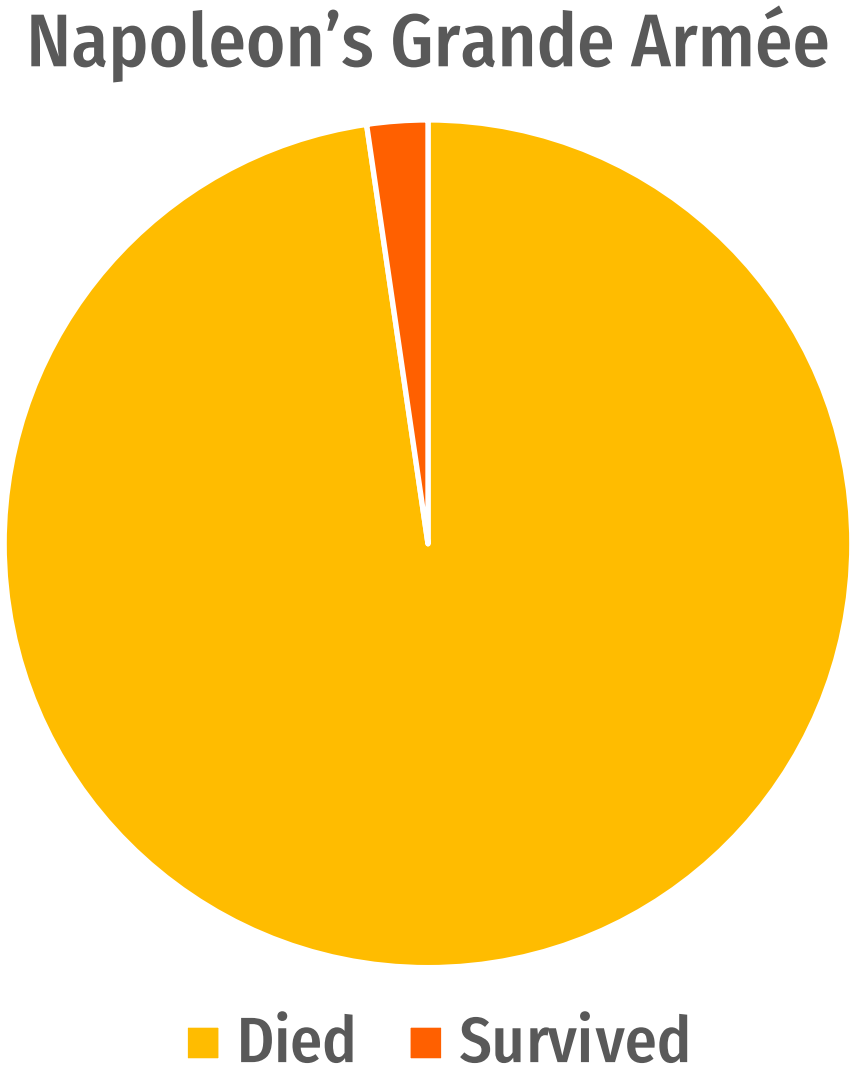
Long distance!



Very cold!



Lots of people died!



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. ___ Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow en ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

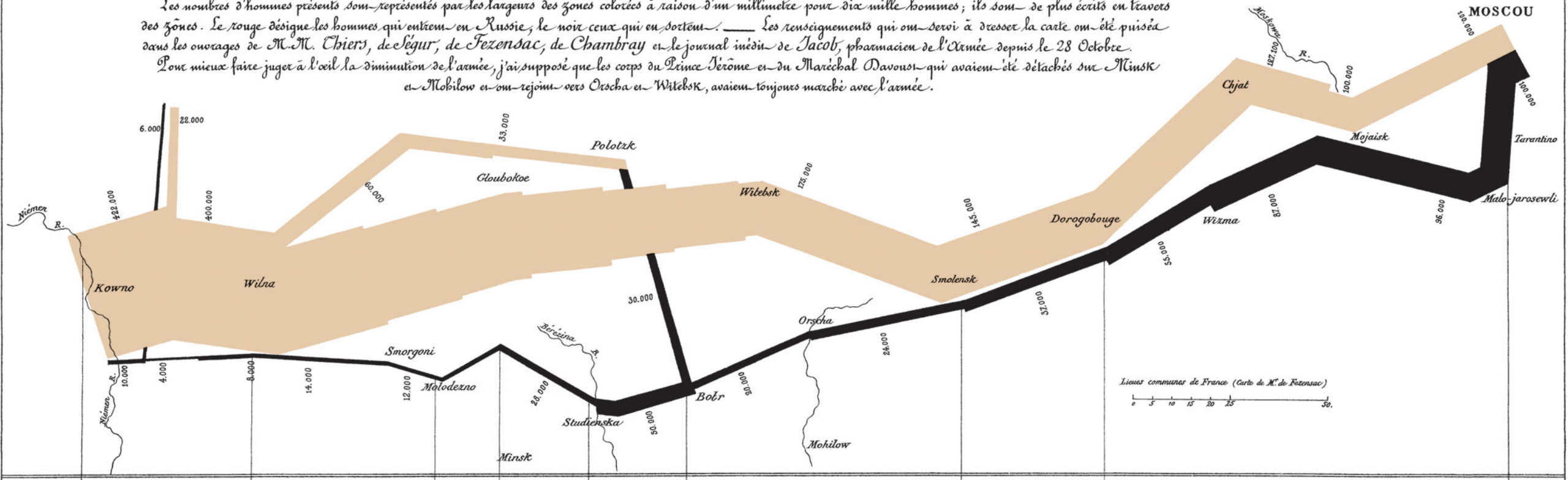
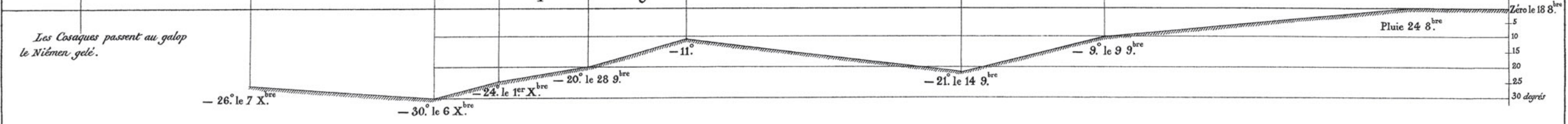


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

Imp. Lith. Regnier et Dourdet.

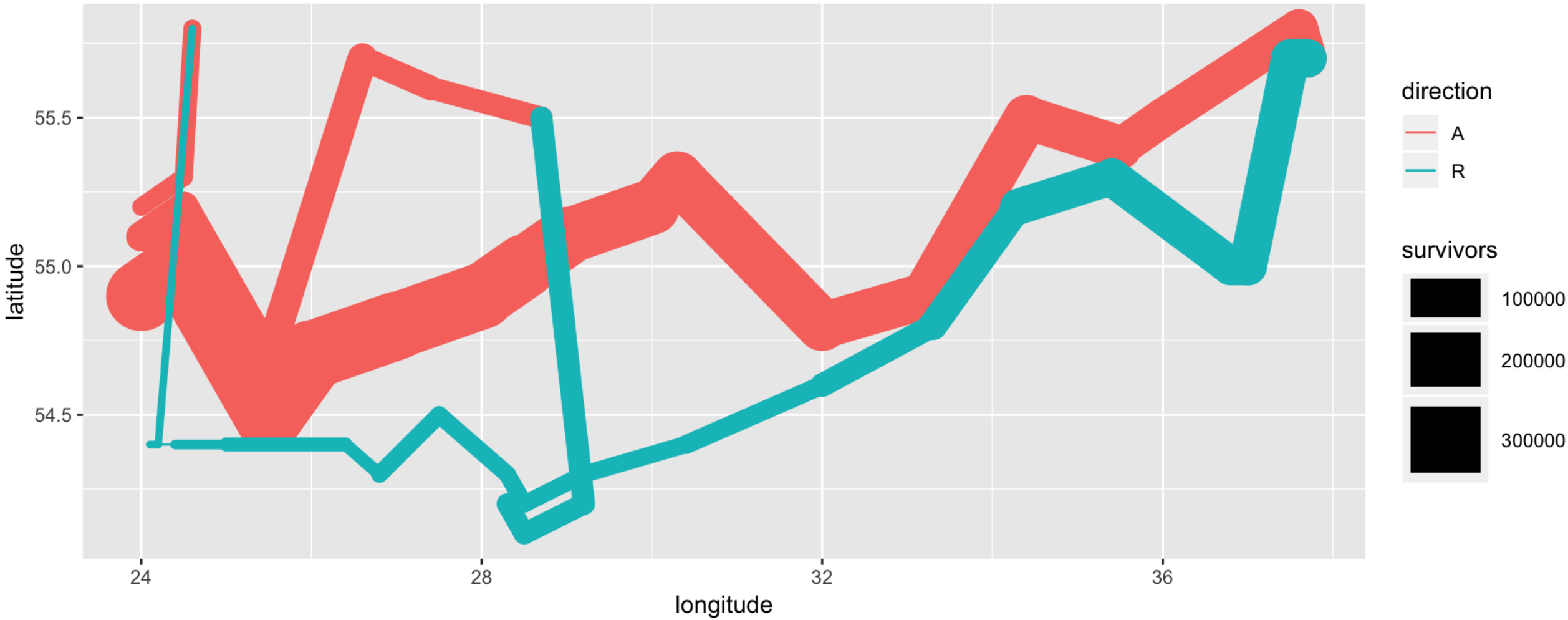
Aesthetics and data

Data	Aesthetic	Graphic
Longitude	x	point
Latitude	y	point
Army size	size	path
Army direction	color	path
Date	x	line and text
Temperature	y	line and text

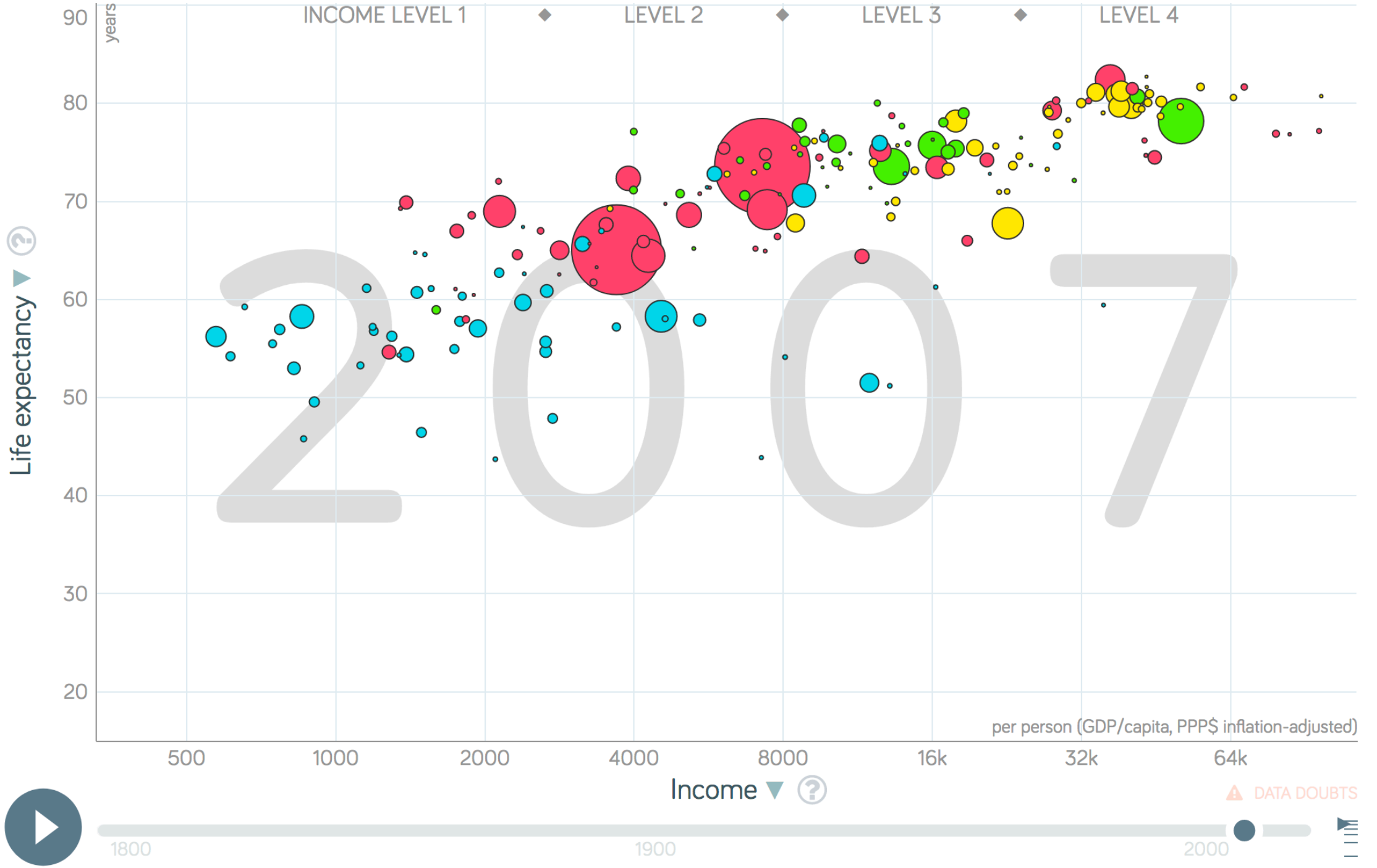
Aesthetics and data

data	aes()	geom_
Longitude	x	point
Latitude	y	point
Army size	size	path
Army direction	color	path
Date	x	line and text
Temperature	y	line and text

```
ggplot(data = troops,
       mapping = aes(x = longitude,
                     y = latitude,
                     color = direction,
                     size = survivors)) +
  geom_path()
```


latitude
55.5
55.0
54.5
24
28
32
36
longitude
direction
A
R
survivors
100000
200000
300000


India
Thailand
400
4000
10000


INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
years
90
80
70
60
50
40
30
20
Life expectancy
2007
500
1000
2000
4000
8000
16k
32k
64k
per person (GDP/capita, PPP$ inflation-adjusted)
Income
DATA DOUBTS
1800
1900
2000

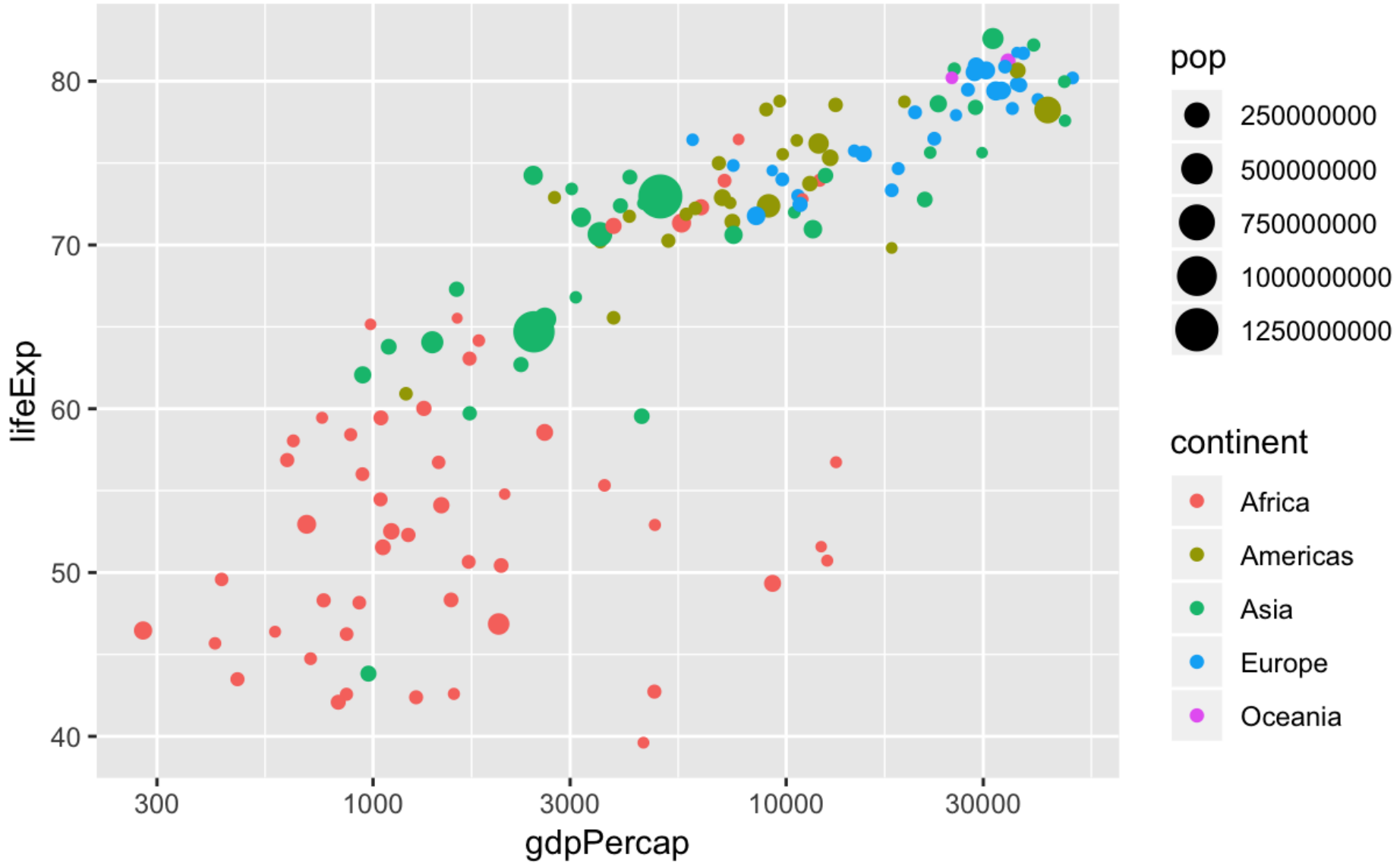
Aesthetics and data

data	aes()	geom_
Wealth (GDP/capita)	x	point
Health (Life expectancy)	y	point
Continent	color	point
Population	size	point

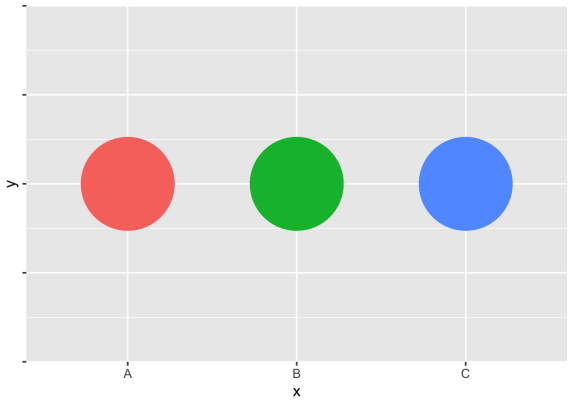
Health and wealth

```
ggplot(data = gapminder_2007,
       mapping = aes(x = gdpPercap,
                     y = lifeExp,
                     color = continent,
                     size = pop)) +
  geom_point() +
  scale_x_log10()
```

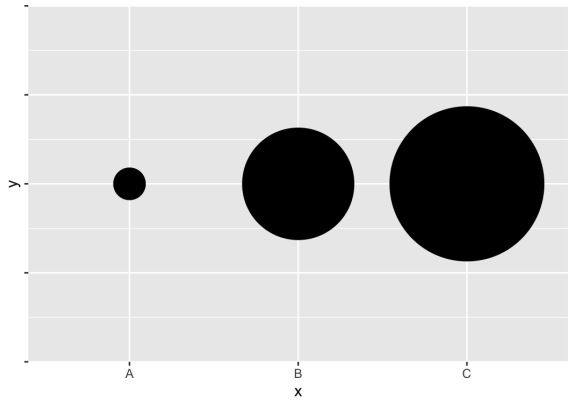
Health and wealth



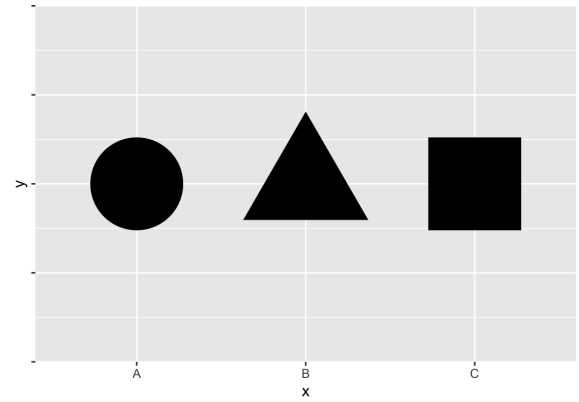
Possible aesthetics



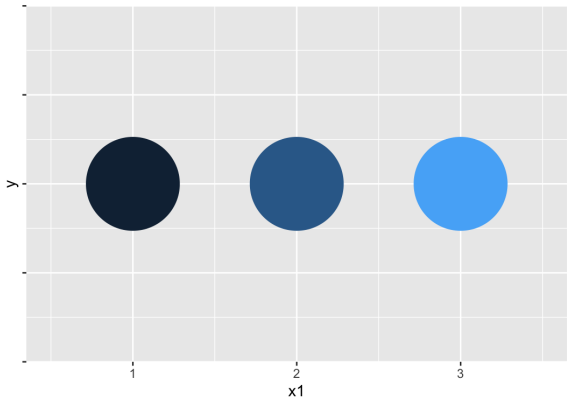
color



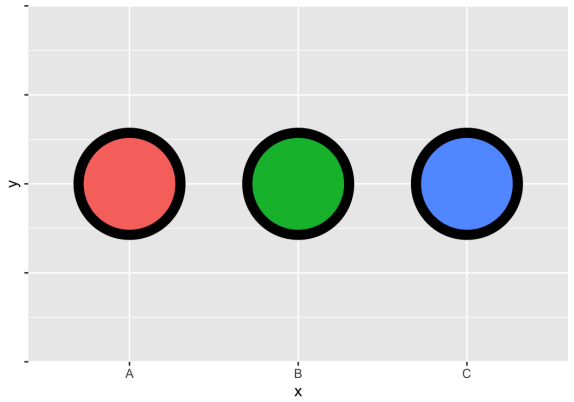
size



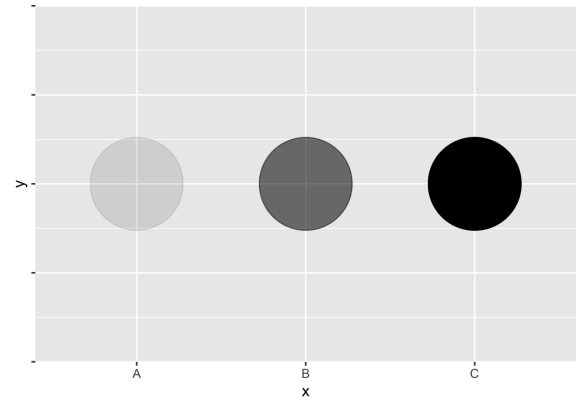
shape



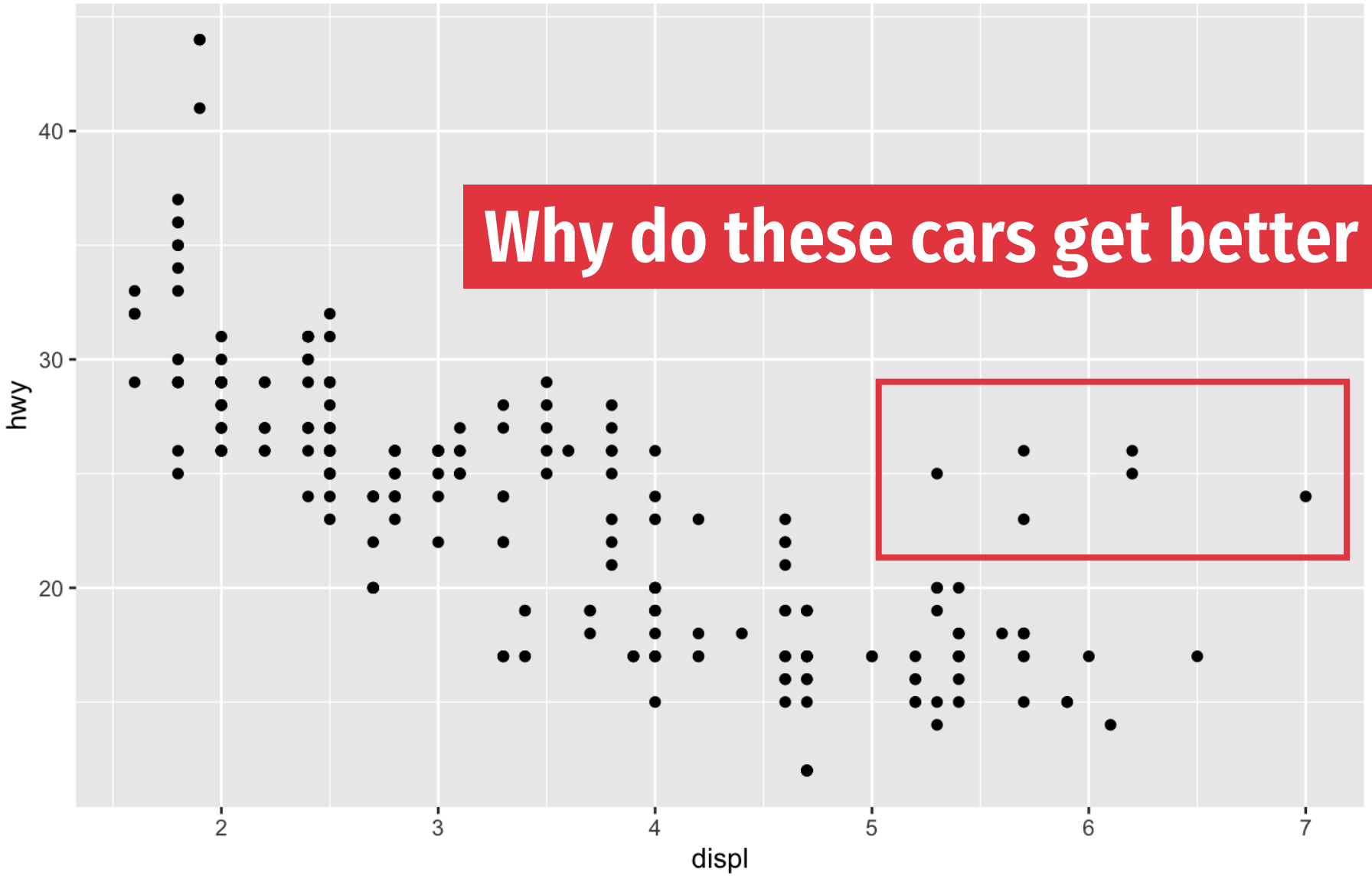
color



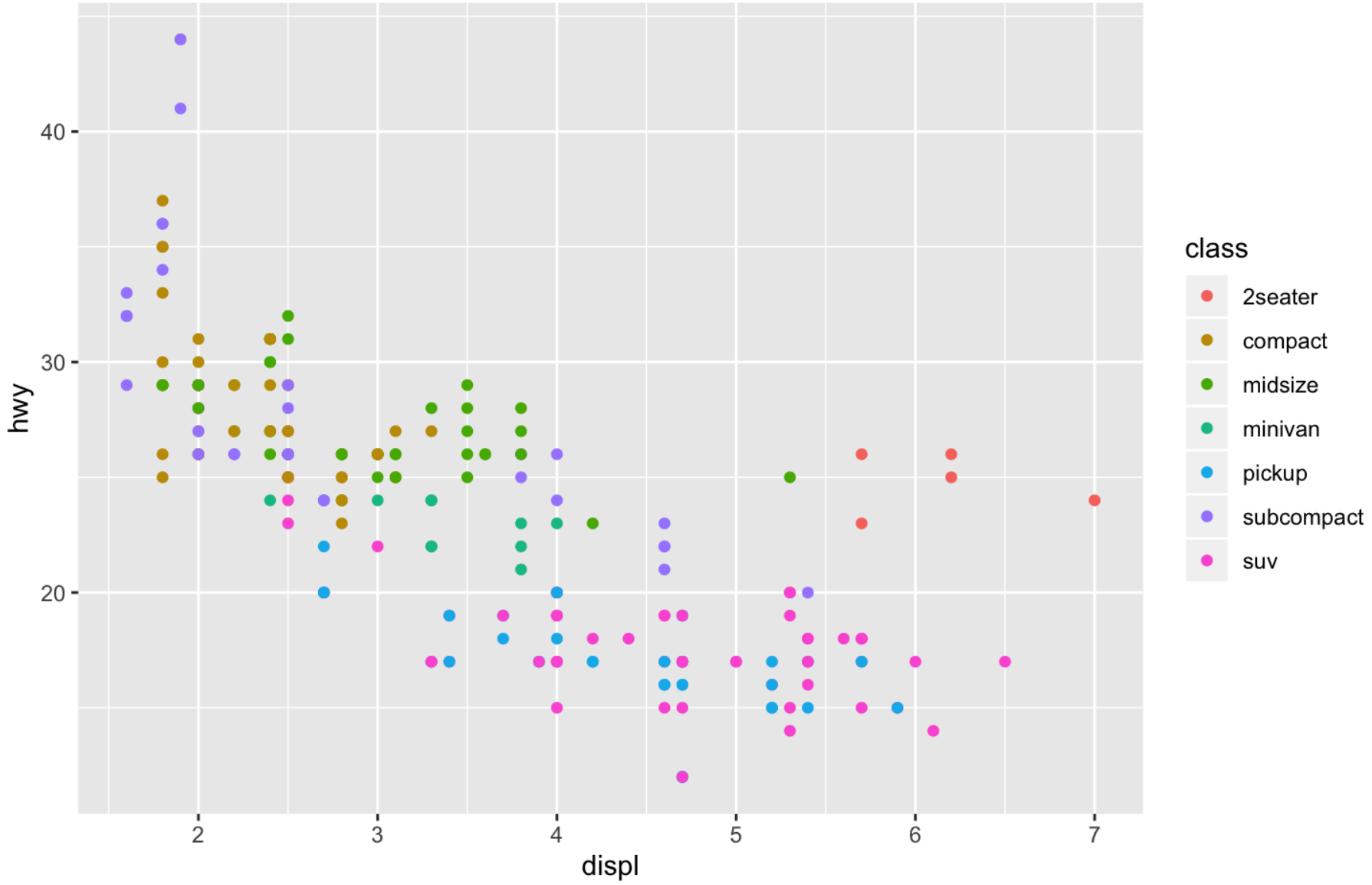
fill



alpha


Why do these cars get better mileage?
hwy
40
30
20
displ
2
3
4
5
6
7

```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, color = class))
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, size = class))
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, shape = class))
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, alpha = class))
```




```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = class))
```

Your turn (#2)

Add color, size, alpha, and shape aesthetics to your graph. Experiment!

Look at the data to see what columns you can use

Do different things happen when you map aesthetics to discrete and continuous variables?

What happens when you use more than one aesthetic?

Your turn (#2)

Add color, size, alpha, and shape aesthetics to your graph. Experiment!

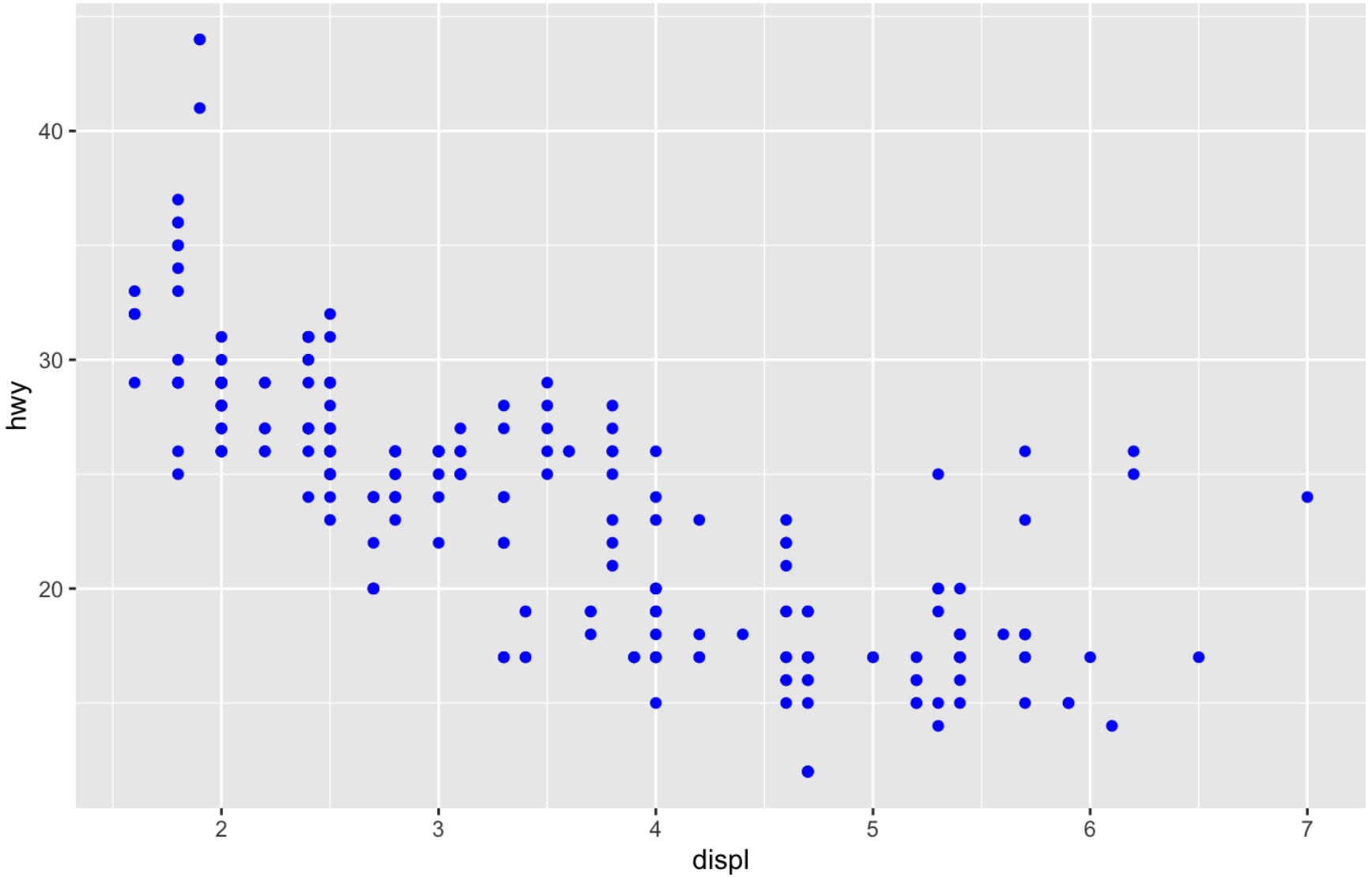
Look at the data to see what columns you can use

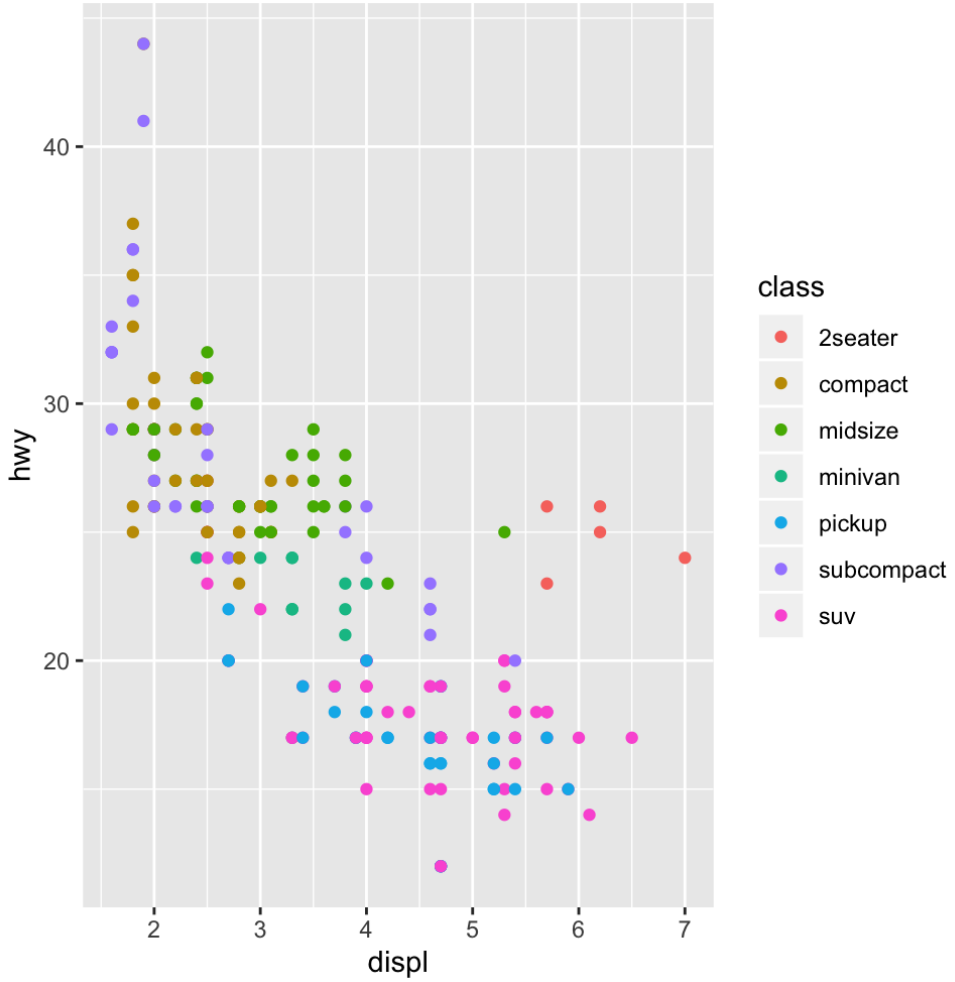
Do different things happen when you map aesthetics to discrete and continuous variables?

What happens when you use more than one aesthetic?

05:00

How would you make this plot?

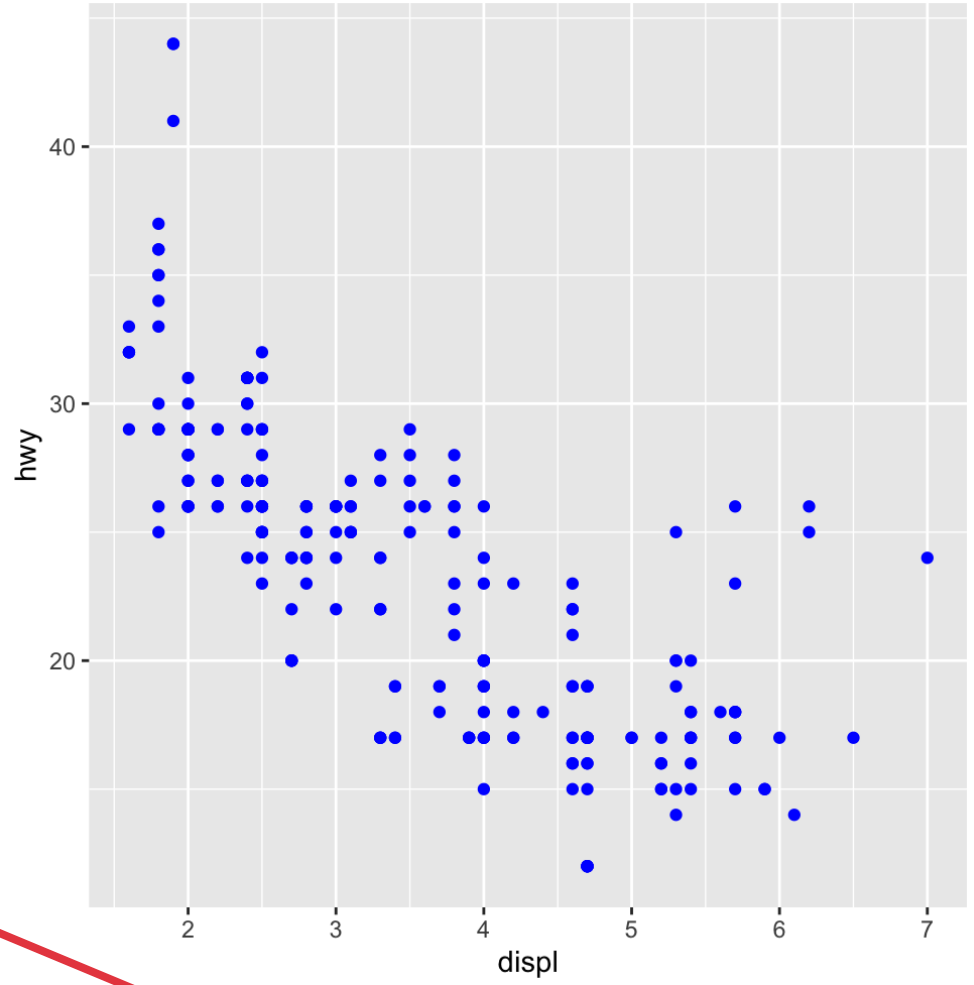




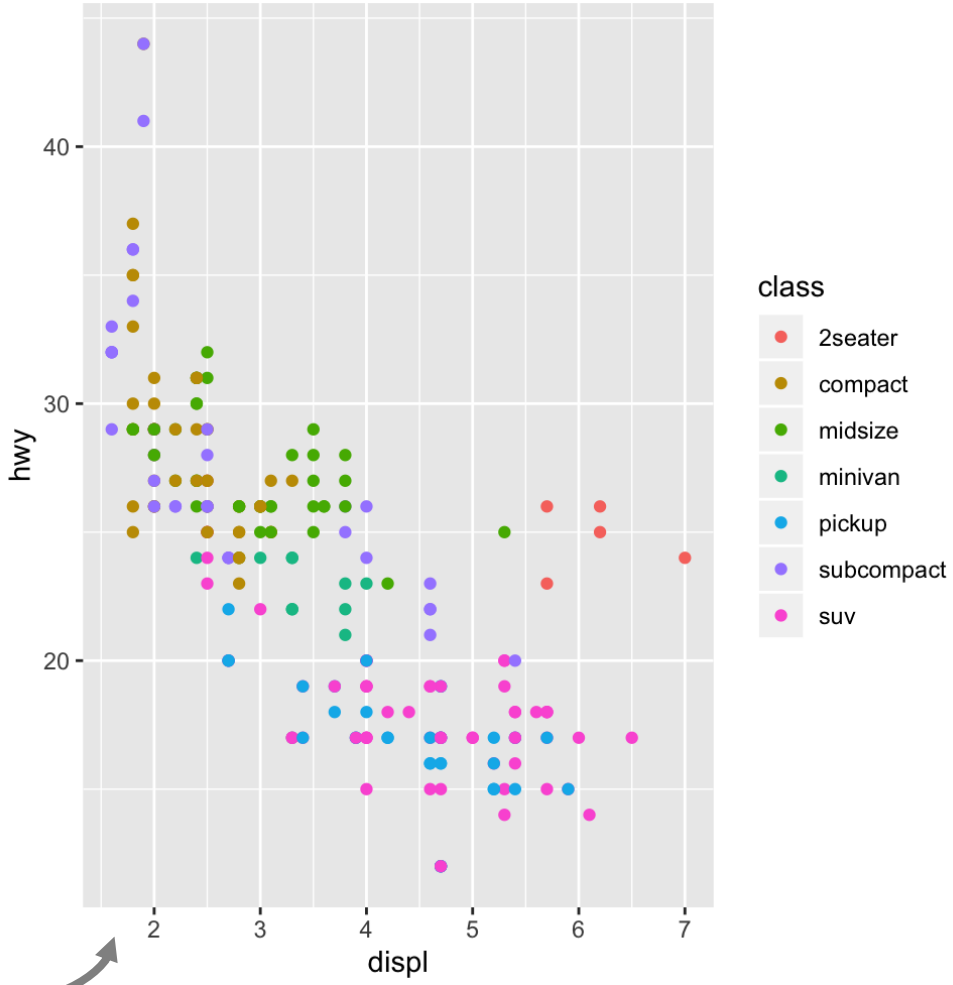
Inside of aes(): sets an aesthetic to a variable

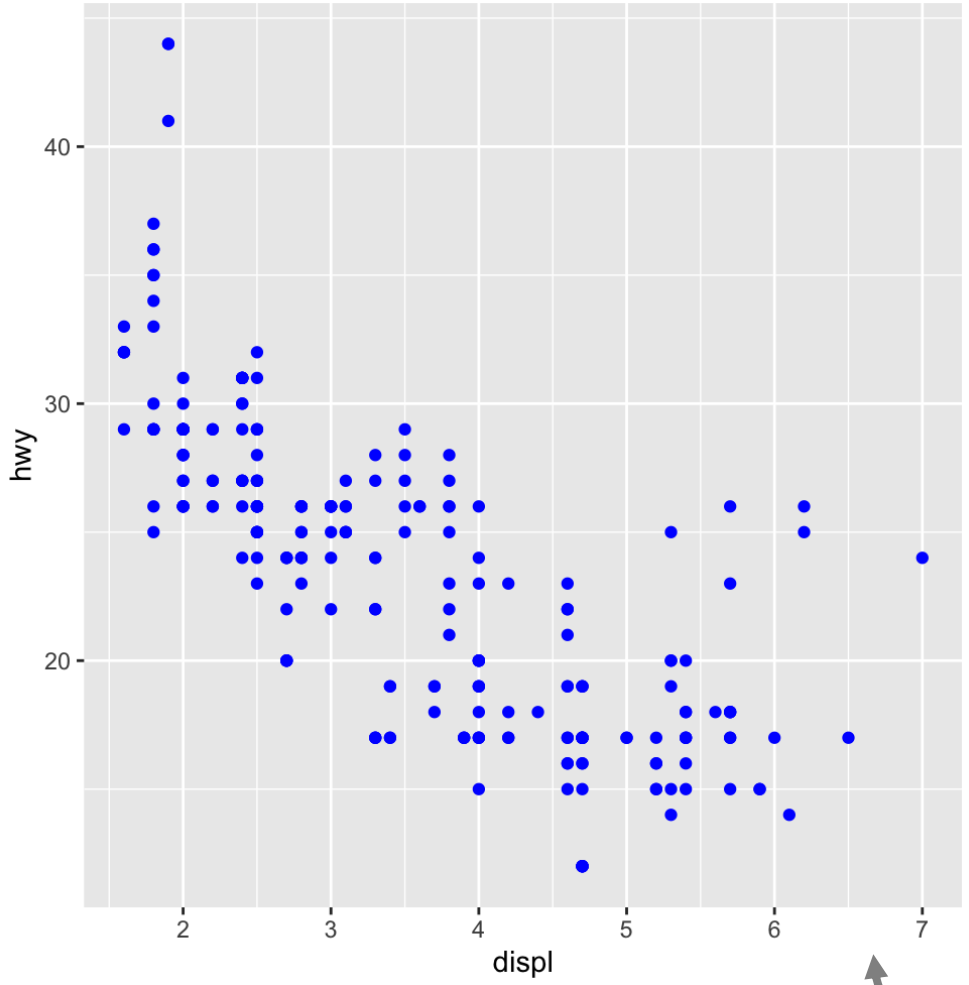
```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, color = class))
```

Outside of aes(): sets an aesthetic to a value



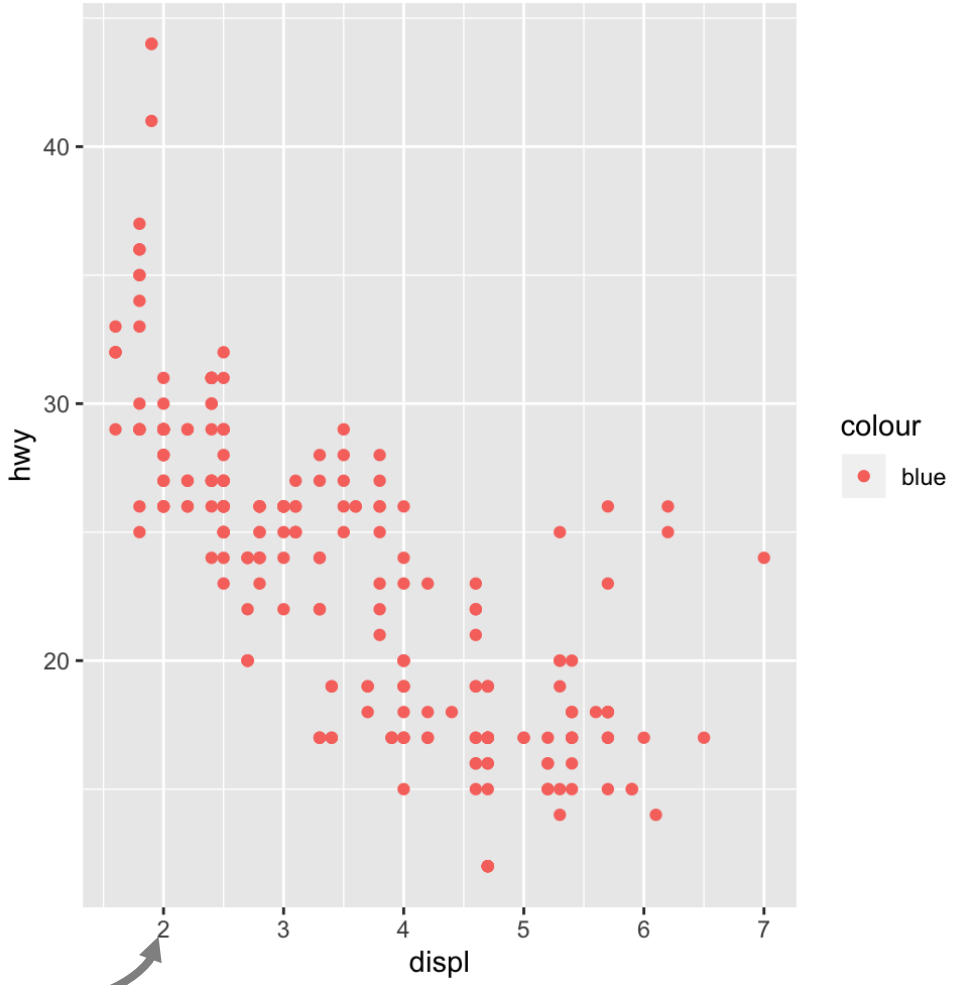

```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy), color = "blue")
```

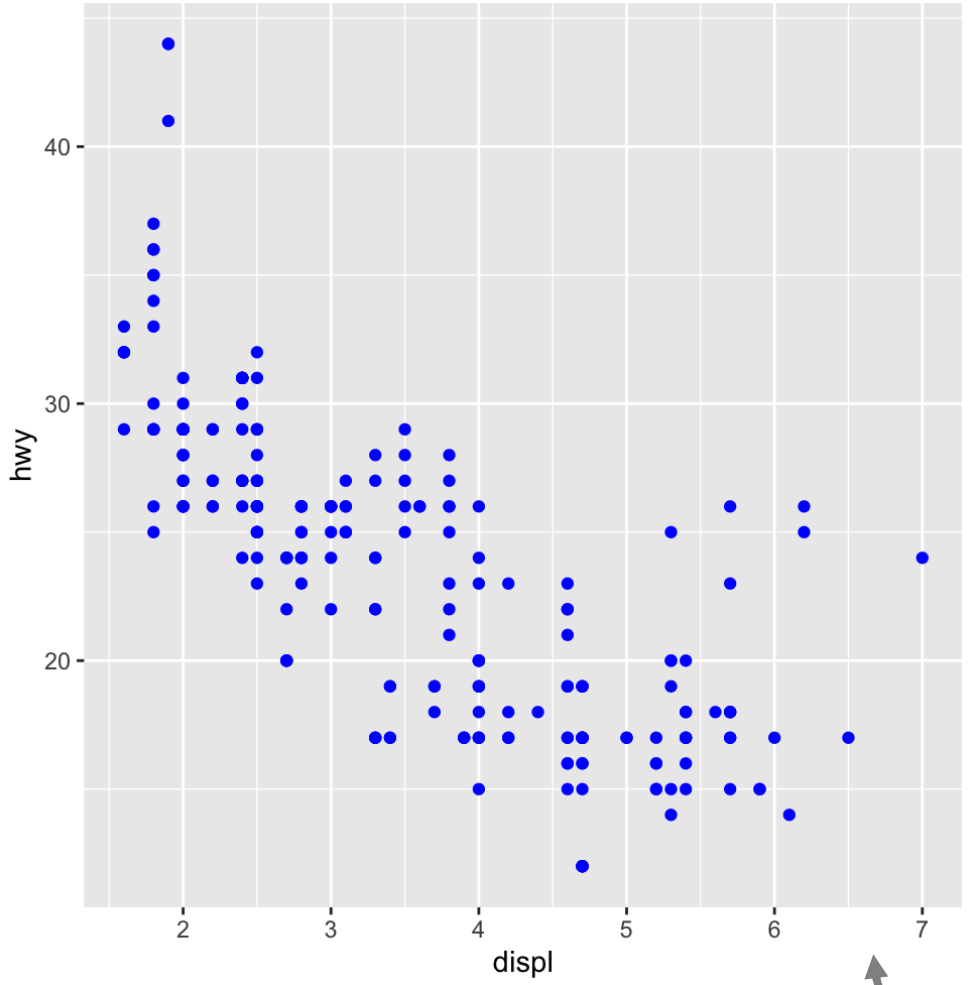





```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, color = class))
```

```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy), color = "blue")
```

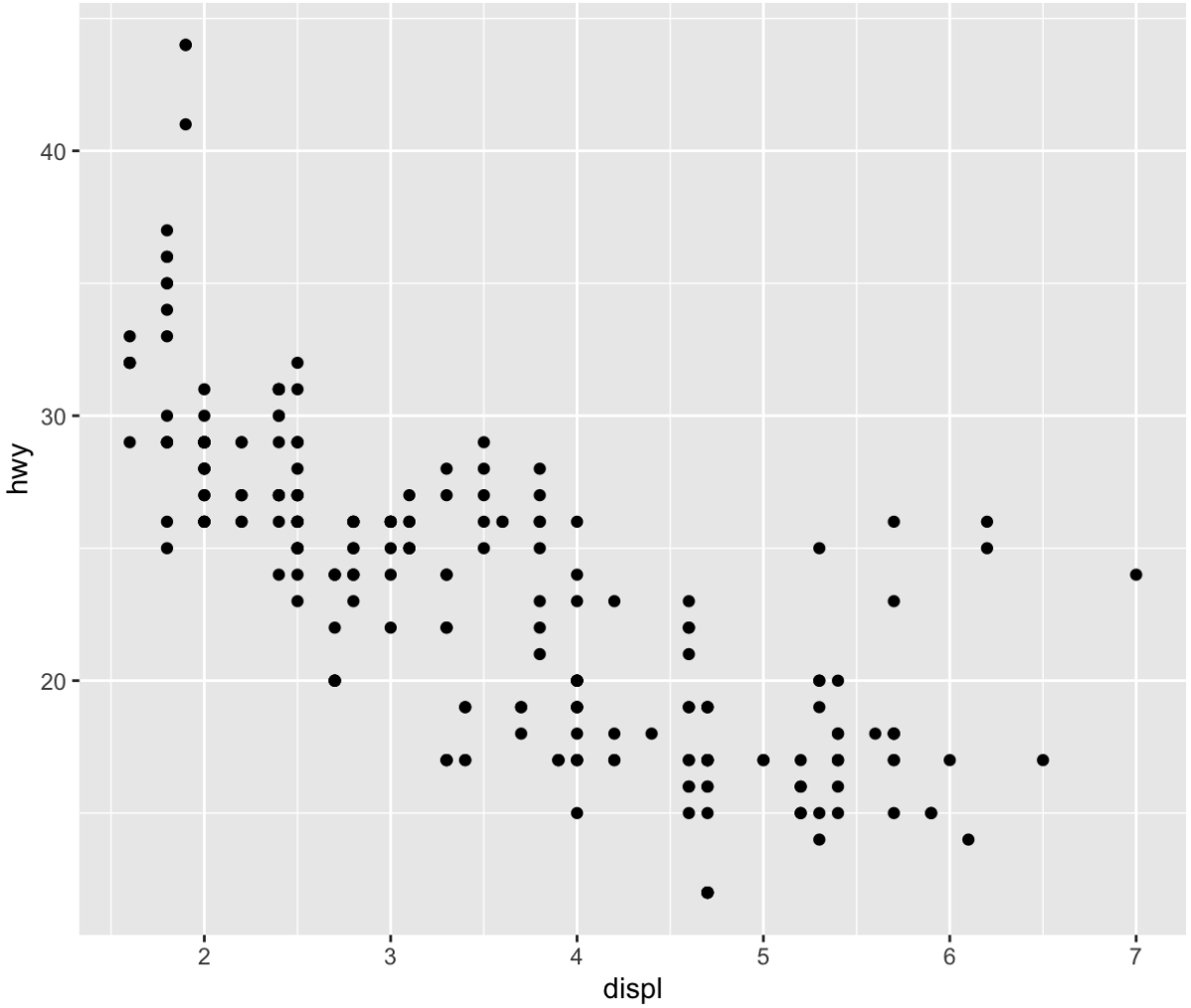


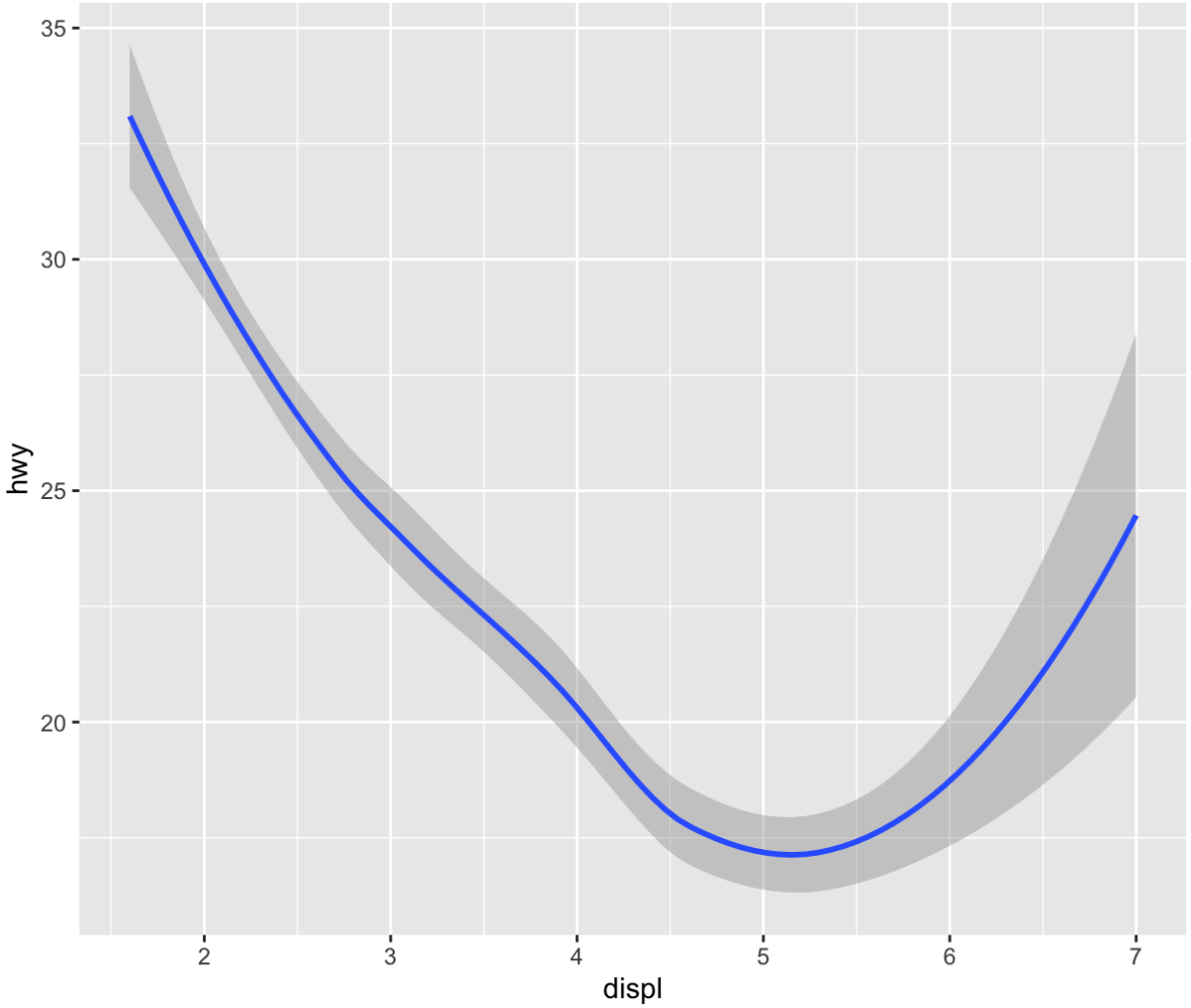



```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy, color = "blue"))
```

```
ggplot(mpg) + geom_point(aes(x = displ, y = hwy), color = "blue")
```

What's the same? What's different?

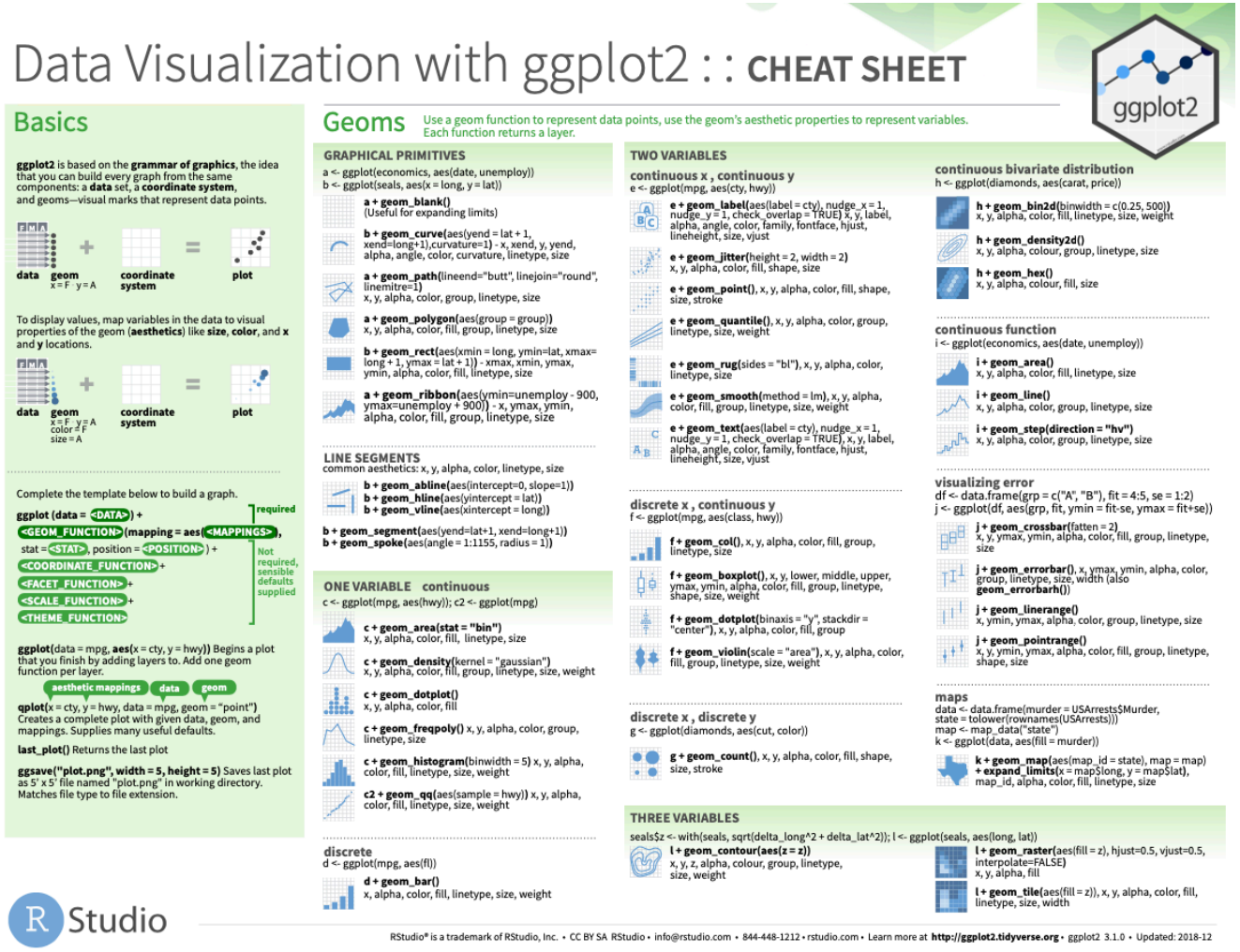




geoms

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

geom_*() functions



Data Visualization with ggplot2 : : CHEAT SHEET

ggplot2

Basics

ggplot2 is based on the grammar of graphics, the idea that you can build every graph from the same components: a data set, a coordinate system, and geoms—visual marks that represent data points.

data + geom (x = F, y = A) = coordinate system + plot

To display values, map variables in the data to visual properties of the geom (aesthetics) like size, color, and x and y locations.

data + geom (x = F, y = A, color = F, size = A) = coordinate system + plot

Complete the template below to build a graph.

required:
ggplot (data = <DATA>) +
<GEOM_FUNCTION> (mapping = aes(<MAPPINGS>),

Not required, sensible defaults supplied:
stat = <STAT>, position = <POSITION>) +
<COORDINATE_FUNCTION> +
<FACET_FUNCTION> +
<SCALE_FUNCTION> +
<THEME_FUNCTION>

ggplot(data = mpg, aes(x = cty, y = hwy)) Begins a plot that you finish by adding layers to. Add one geom function per layer.

aesthetic mappings | data | geom

qplot(x = cty, y = hwy, data = mpg, geom = "point") Creates a complete plot with given data, geom, and mappings. Supplies many useful defaults.

last_plot() Returns the last plot

ggsave("plot.png", width = 5, height = 5) Saves last plot as 5' x 5' file named "plot.png" in working directory. Matches file type to file extension.

Geoms

Use a geom function to represent data points, use the geom's aesthetic properties to represent variables. Each function returns a layer.

GRAPHICAL PRIMITIVES

a <- ggplot(economics, aes(date, unemploy))
b <- ggplot(seals, aes(x = long, y = lat))

- a + geom_blank() (Useful for expanding limits)
- b + geom_curve(aes(yend = lat + 1, xend=long+1),curvature=1) - x, xend, y, yend, alpha, angle, color, curvature, linetype, size
- a + geom_path(lineend="butt", linejoin="round", linemitre=1) x, y, alpha, color, group, linetype, size
- a + geom_polygon(aes(group = group)) x, y, alpha, color, fill, group, linetype, size
- b + geom_rect(aes(xmin = long, ymin=lat, xmax= long + 1, ymax = lat + 1)) - xmax, xmin, ymax, ymin, alpha, color, fill, linetype, size
- a + geom_ribbon(aes(ymin=unemploy - 900, ymax=unemploy + 900)) - x, ymax, ymin, alpha, color, fill, group, linetype, size

LINE SEGMENTS

common aesthetics: x, y, alpha, color, linetype, size

- b + geom_abline(aes(intercept=0, slope=1))
- b + geom_hline(aes(yintercept = lat))
- b + geom_vline(aes(xintercept = long))
- b + geom_segment(aes(yend=lat+1, xend=long+1))
- b + geom_spoke(aes(angle = 1:1155, radius = 1))

ONE VARIABLE continuous

c <- ggplot(mpg, aes(hwy)); c2 <- ggplot(mpg)

- c + geom_area(stat = "bin") x, y, alpha, color, fill, linetype, size
- c + geom_density(kernel = "gaussian") x, y, alpha, color, fill, group, linetype, size, weight
- c + geom_dotplot() x, y, alpha, color, fill
- c + geom_freqpoly() x, y, alpha, color, group, linetype, size
- c + geom_histogram(binwidth = 5) x, y, alpha, color, fill, linetype, size, weight
- c2 + geom_qq(aes(sample = hwy)) x, y, alpha, color, fill, linetype, size, weight

discrete

d <- ggplot(mpg, aes(fl))

- d + geom_bar() x, alpha, color, fill, linetype, size, weight

TWO VARIABLES

continuous x , continuous y

e <- ggplot(mpg, aes(cty, hwy))

- e + geom_label(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE) x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust
- e + geom_jitter(height = 2, width = 2) x, y, alpha, color, fill, shape, size
- e + geom_point(), x, y, alpha, color, fill, shape, size, stroke
- e + geom_quantile(), x, y, alpha, color, group, linetype, size, weight
- e + geom_rug(sides = "bl"), x, y, alpha, color, linetype, size
- e + geom_smooth(method = lm), x, y, alpha, color, fill, group, linetype, size, weight
- e + geom_text(aes(label = cty), nudge_x = 1, nudge_y = 1, check_overlap = TRUE), x, y, label, alpha, angle, color, family, fontface, hjust, lineheight, size, vjust

discrete x , continuous y

f <- ggplot(mpg, aes(class, hwy))

- f + geom_col(), x, y, alpha, color, fill, group, linetype, size
- f + geom_boxplot(), x, y, lower, middle, upper, ymax, ymin, alpha, color, fill, group, linetype, shape, size, weight
- f + geom_dotplot(binaxis = "y", stackdir = "center"), x, y, alpha, color, fill, group
- f + geom_violin(scale = "area"), x, y, alpha, color, fill, group, linetype, size, weight

discrete x , discrete y

g <- ggplot(diamonds, aes(cut, color))

- g + geom_count(), x, y, alpha, color, fill, shape, size, stroke

continuous bivariate distribution

h <- ggplot(diamonds, aes(carat, price))

- h + geom_bin2d(binwidth = c(0.25, 500)) x, y, alpha, color, fill, linetype, size, weight
- h + geom_density2d() x, y, alpha, colour, group, linetype, size
- h + geom_hex() x, y, alpha, colour, fill, size

continuous function

i <- ggplot(economics, aes(date, unemploy))

- i + geom_area() x, y, alpha, color, fill, linetype, size
- i + geom_line() x, y, alpha, color, group, linetype, size
- i + geom_step(direction = "hv") x, y, alpha, color, group, linetype, size

visualizing error

df <- data.frame(grp = c("A", "B"), fit = 4:5, se = 1:2)
j <- ggplot(df, aes(grp, fit, ymin = fit-se, ymax = fit+se))

- j + geom_crossbar(fatten = 2) x, y, ymax, ymin, alpha, color, fill, group, linetype, size
- j + geom_errorbar(), x, ymax, ymin, alpha, color, group, linetype, size, width (also geom_errorbarh())
- j + geom_linerange() x, ymin, ymax, alpha, color, group, linetype, size
- j + geom_pointrange() x, y, ymin, ymax, alpha, color, fill, group, linetype, shape, size

maps

data <- data.frame(murder = USArrests$Murder, state = tolower(rownames(USArrests)))
map <- map_data("state")
k <- ggplot(data, aes(fill = murder))

- k + geom_map(aes(map_id = state), map = map) + expand_limits(x = map$long, y = map$lat), map_id, alpha, color, fill, linetype, size

THREE VARIABLES

seals$z <- with(seals, sqrt(delta_long^2 + delta_lat^2)); l <- ggplot(seals, aes(long, lat))

- l + geom_contour(aes(z = z)) x, y, z, alpha, colour, group, linetype, size, weight
- l + geom_raster(aes(fill = z), hjust=0.5, vjust=0.5, interpolate=FALSE) x, y, alpha, fill
- l + geom_tile(aes(fill = z)), x, y, alpha, color, fill, linetype, size, width

RStudio® is a trademark of RStudio, Inc. • CC BY SA RStudio • info@rstudio.com • 844-448-1212 • rstudio.com • Learn more at http://ggplot2.tidyverse.org • ggplot2 3.1.0 • Updated: 2018-12

---

ggplot2 part of the tidyverse 3.2.1 — Reference

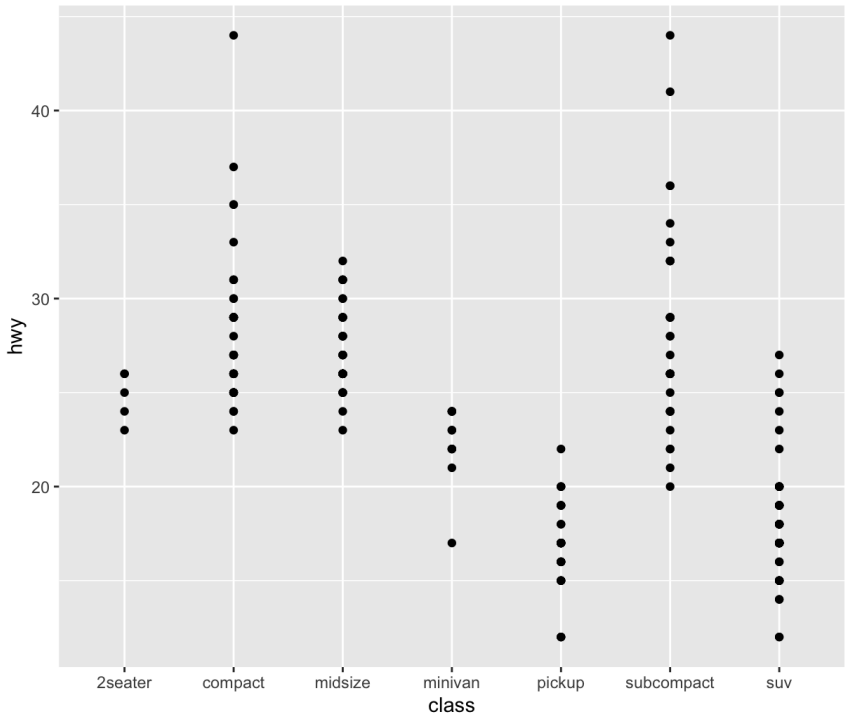
Layer: geoms

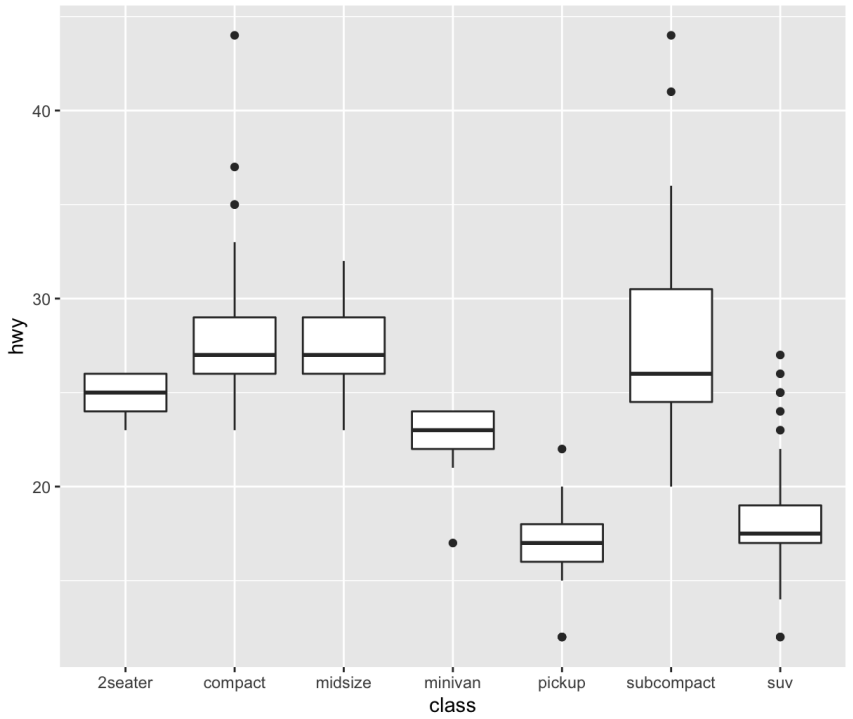
A layer combines data, aesthetic mapping, a geom (geometric object), a stat (statistical transformation), and a position adjustment. Typically, you will create layers using a `geom_` function, overriding the default position and stat if needed.

`geom_abline()` `geom_hline()` `geom_vline()`	Reference lines: horizontal, vertical, and diagonal
`geom_bar()` `geom_col()` `stat_count()`	Bar charts
`geom_bin2d()` `stat_bin_2d()`	Heatmap of 2d bin counts
`geom_blank()`	Draw nothing
`geom_boxplot()` `stat_boxplot()`	A box and whiskers plot (in the style of Tukey)
`geom_contour()` `stat_contour()`	2d contours of a 3d surface
`geom_count()` `stat_sum()`	Count overlapping points
`geom_density()` `stat_density()`	Smoothed density estimates
`geom_density_2d()` `stat_density_2d()`	Contours of a 2d density estimate
`geom_dotplot()`	Dot plot
`geom_errorbarh()`	Horizontal error bars
`geom_hex()` `stat_bin_hex()`	Hexagonal heatmap of 2d bin counts
`geom_freqpoly()` `geom_histogram()` `stat_bin()`	Histograms and frequency polygons

Your turn (#3)

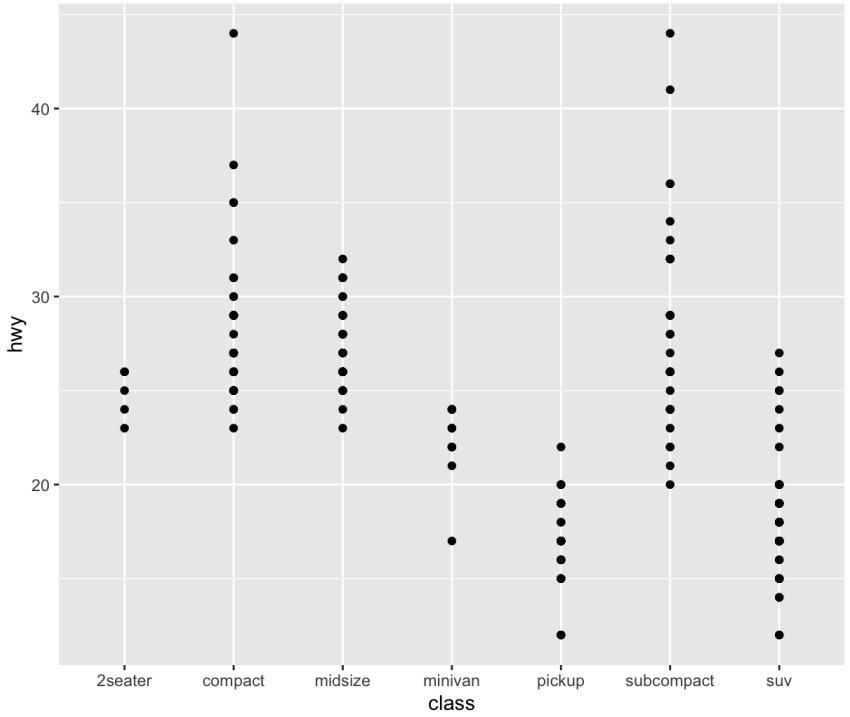
With a partner, decide how to replace this scatterplot with boxplots. Use the cheatsheet.

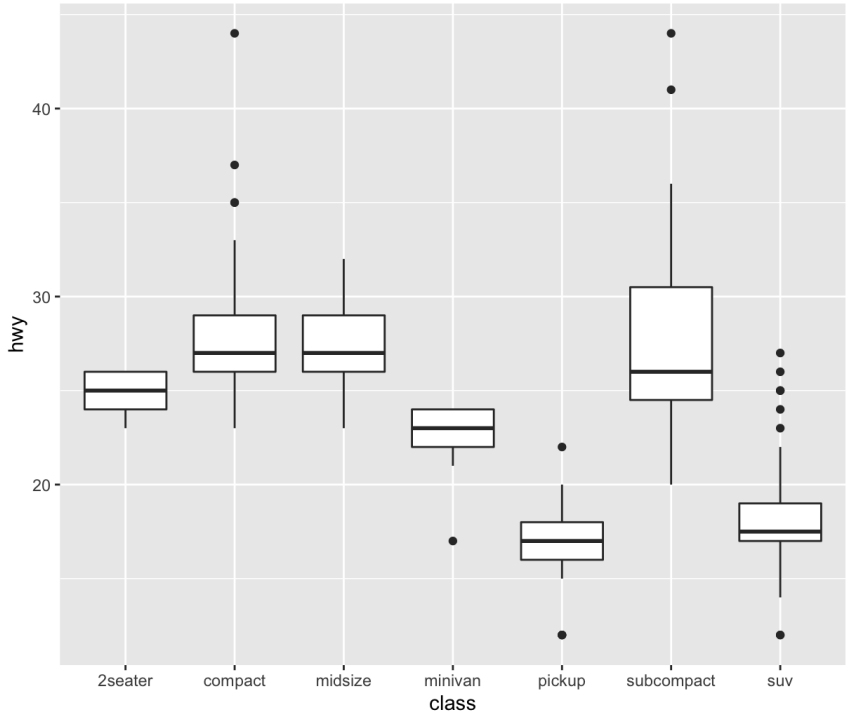




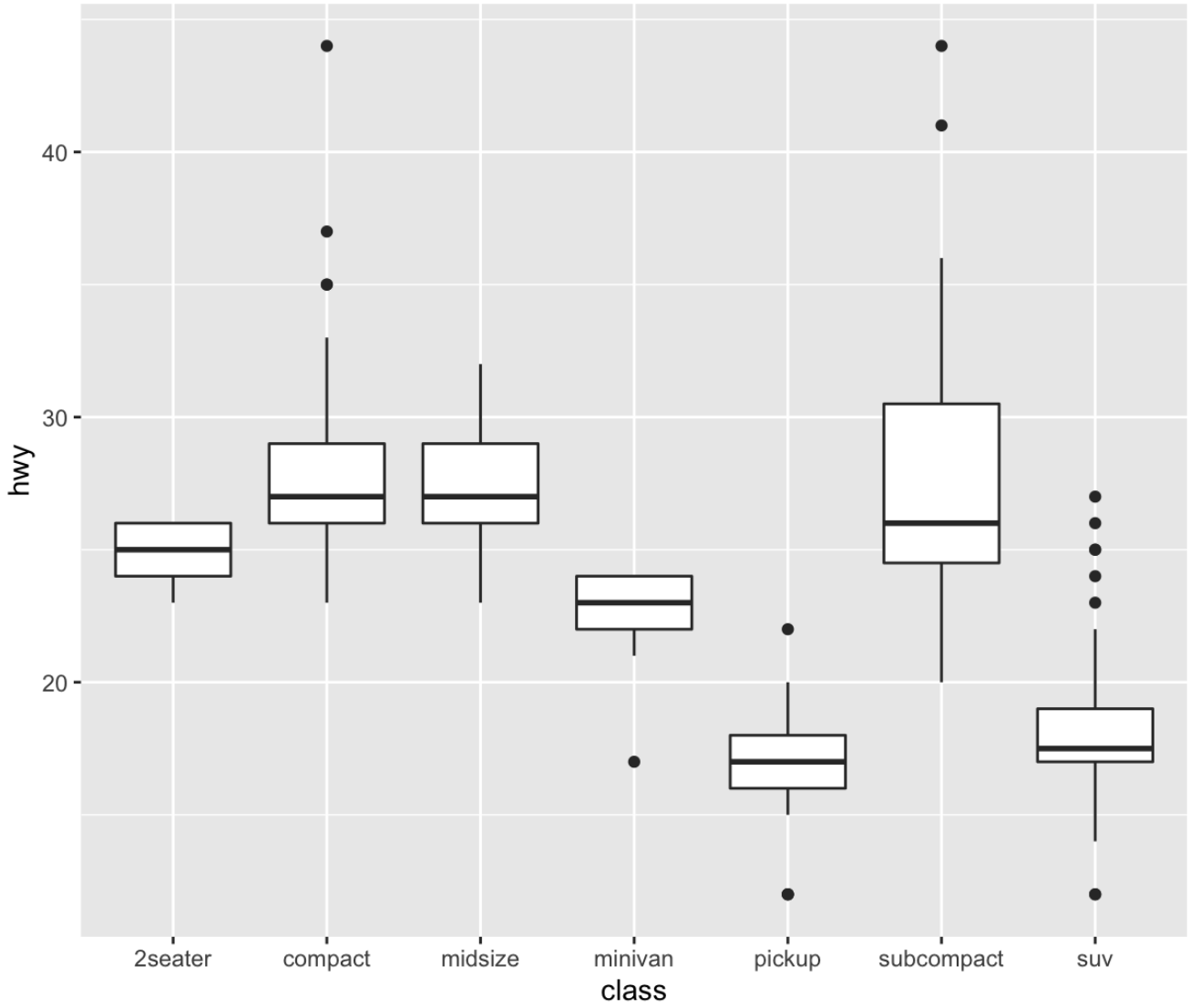
Your turn (#3)

With a partner, decide how to replace this scatterplot with boxplots. Use the cheatsheet.





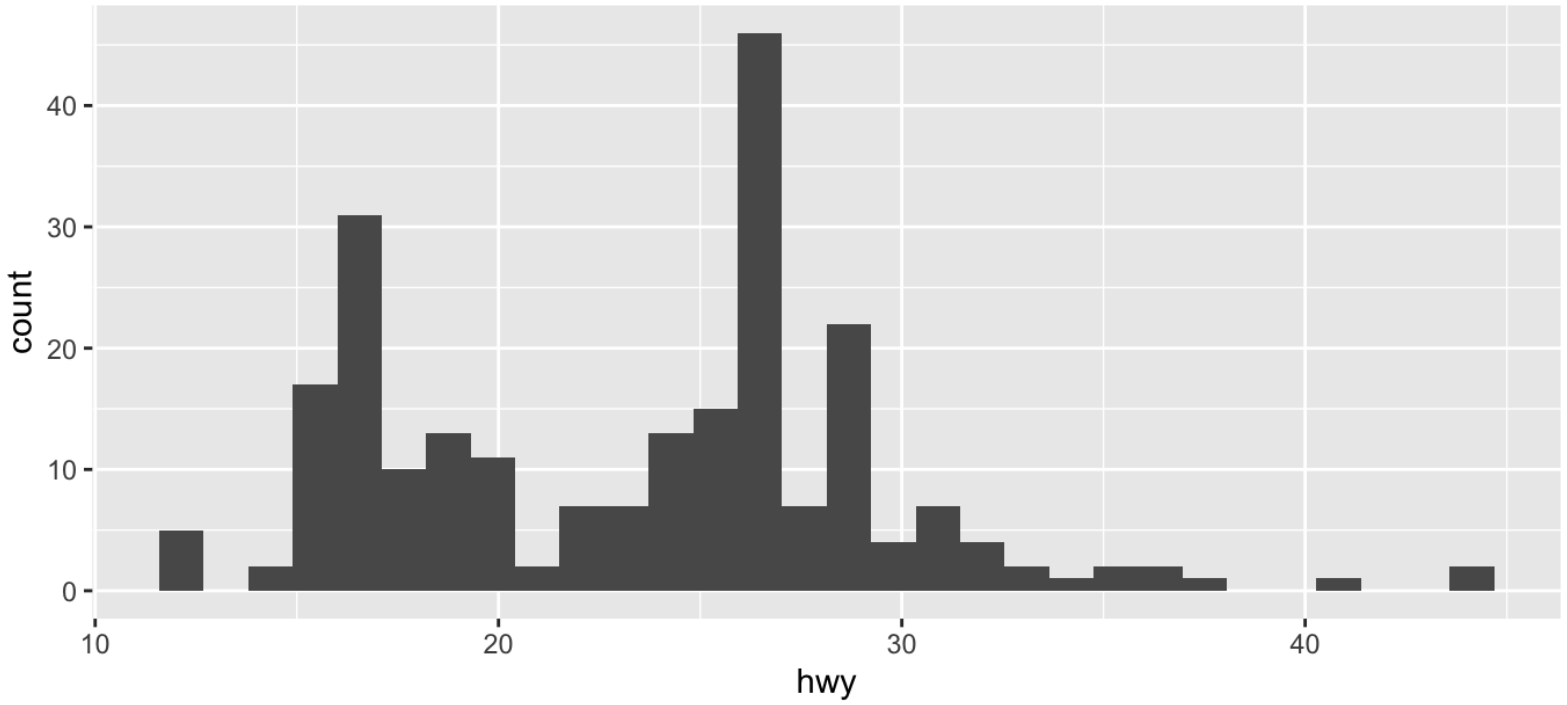
02:00




```
ggplot(data = mpg) +
  geom_boxplot(mapping = aes(x = displ, y = hwy))
```

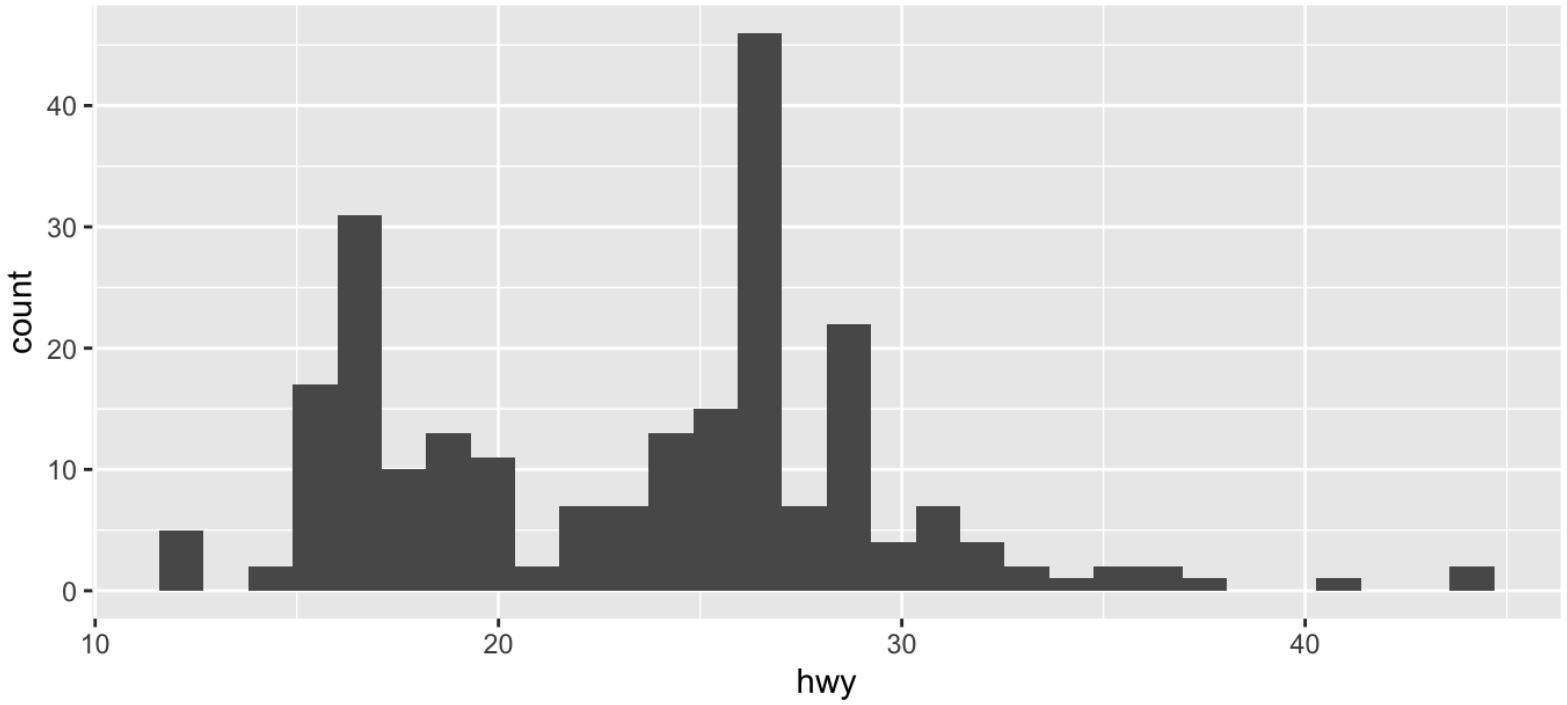
Your turn (#4)

With a partner, make the histogram of hwy. Use the cheatsheet. Hint: don't supply a y variable.

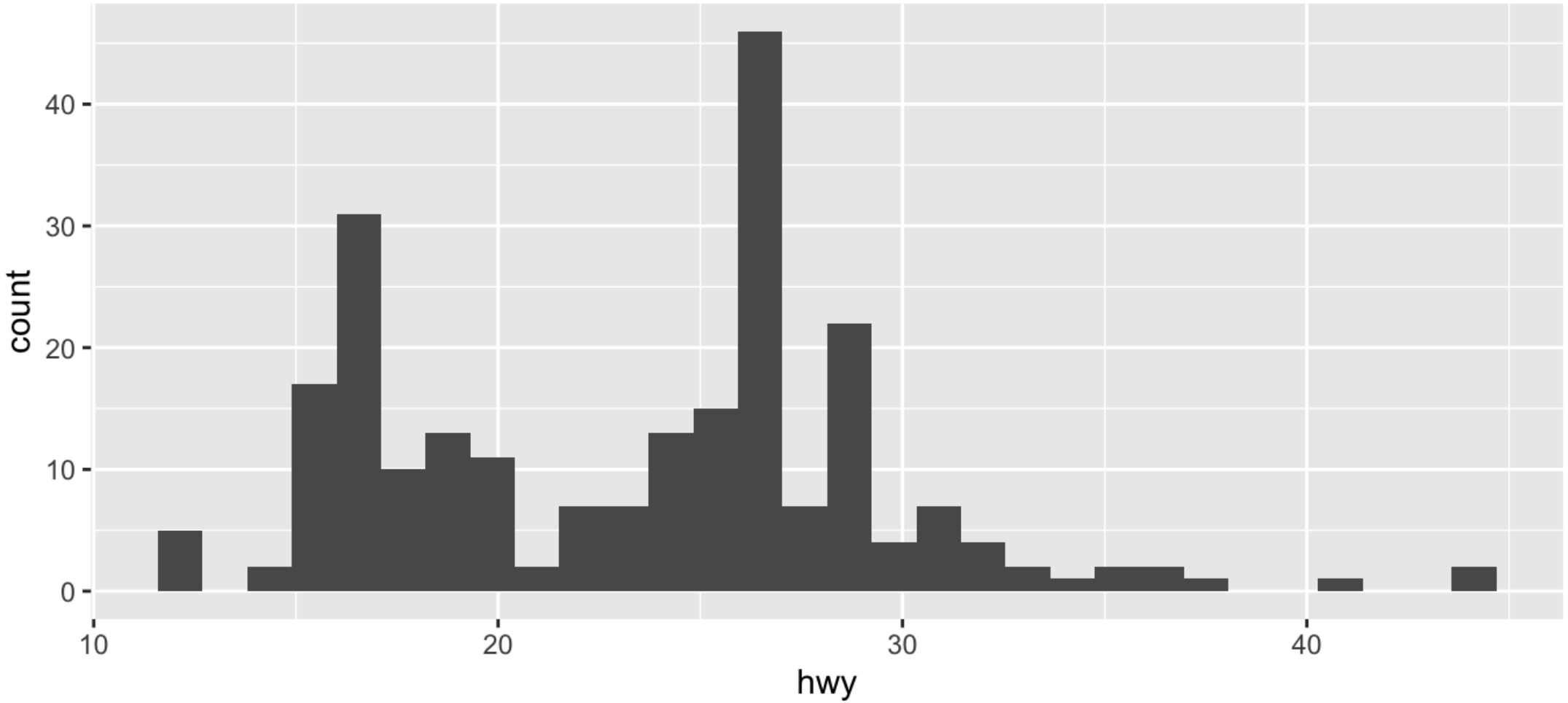


Your turn (#4)

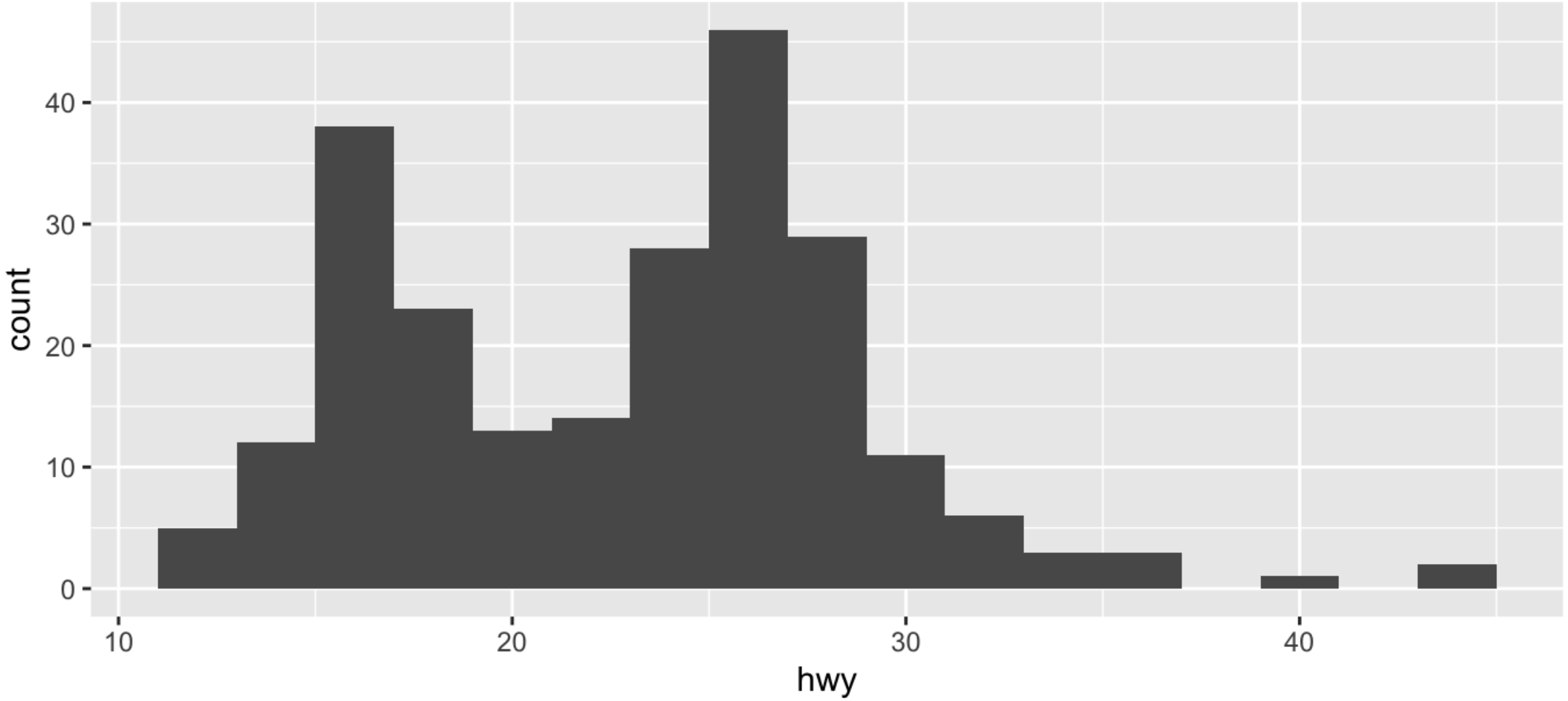
With a partner, make the histogram of hwy. Use the cheatsheet. Hint: don't supply a y variable.



02:00

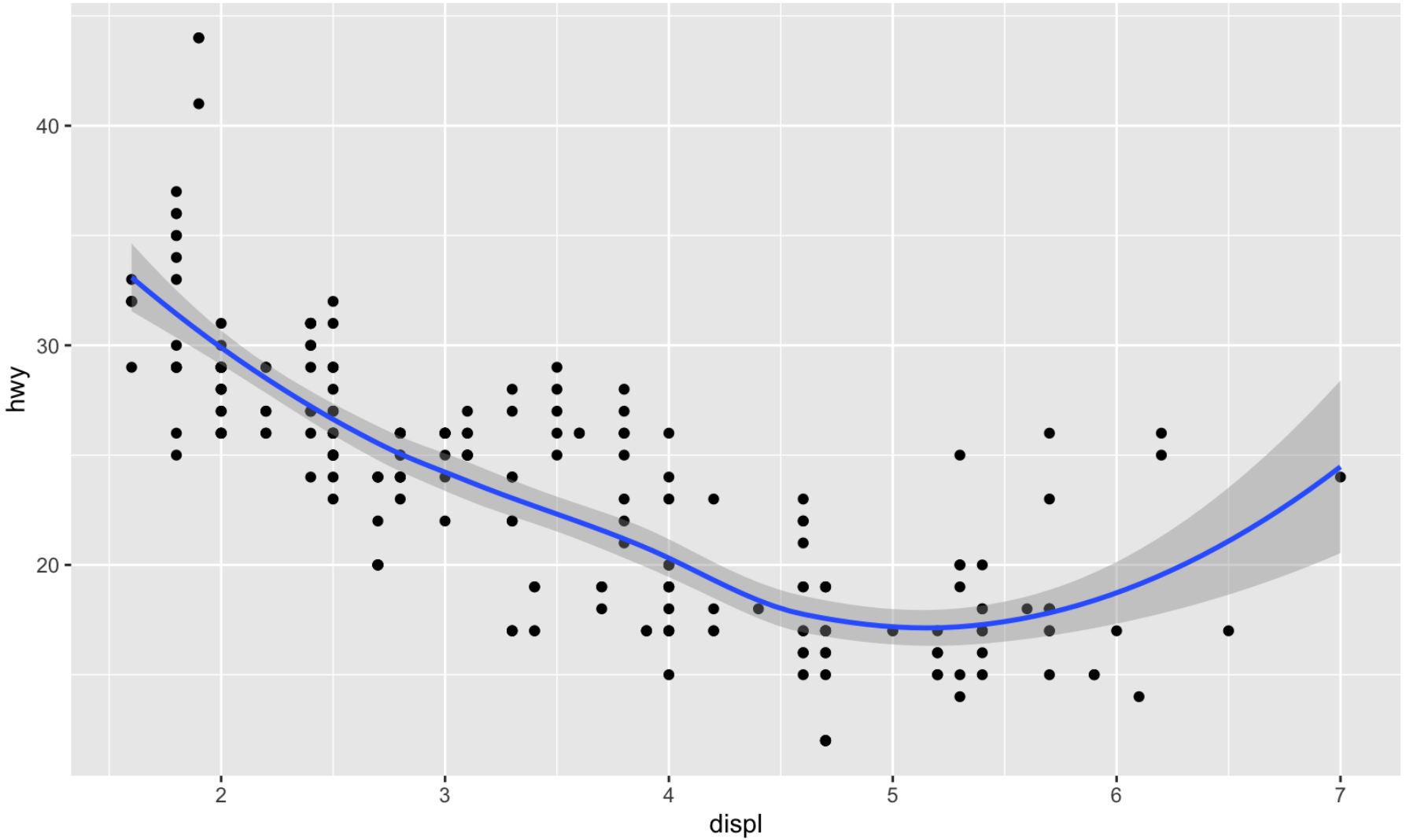



```
ggplot(data = mpg) +
  geom_histogram(mapping = aes(x = hwy))
```




```
ggplot(data = mpg) +
  geom_histogram(mapping = aes(x = hwy), binwidth = 2)
```

Complex graphs!



Your turn (#5)

With a partner, predict what this code will do. Then run it.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

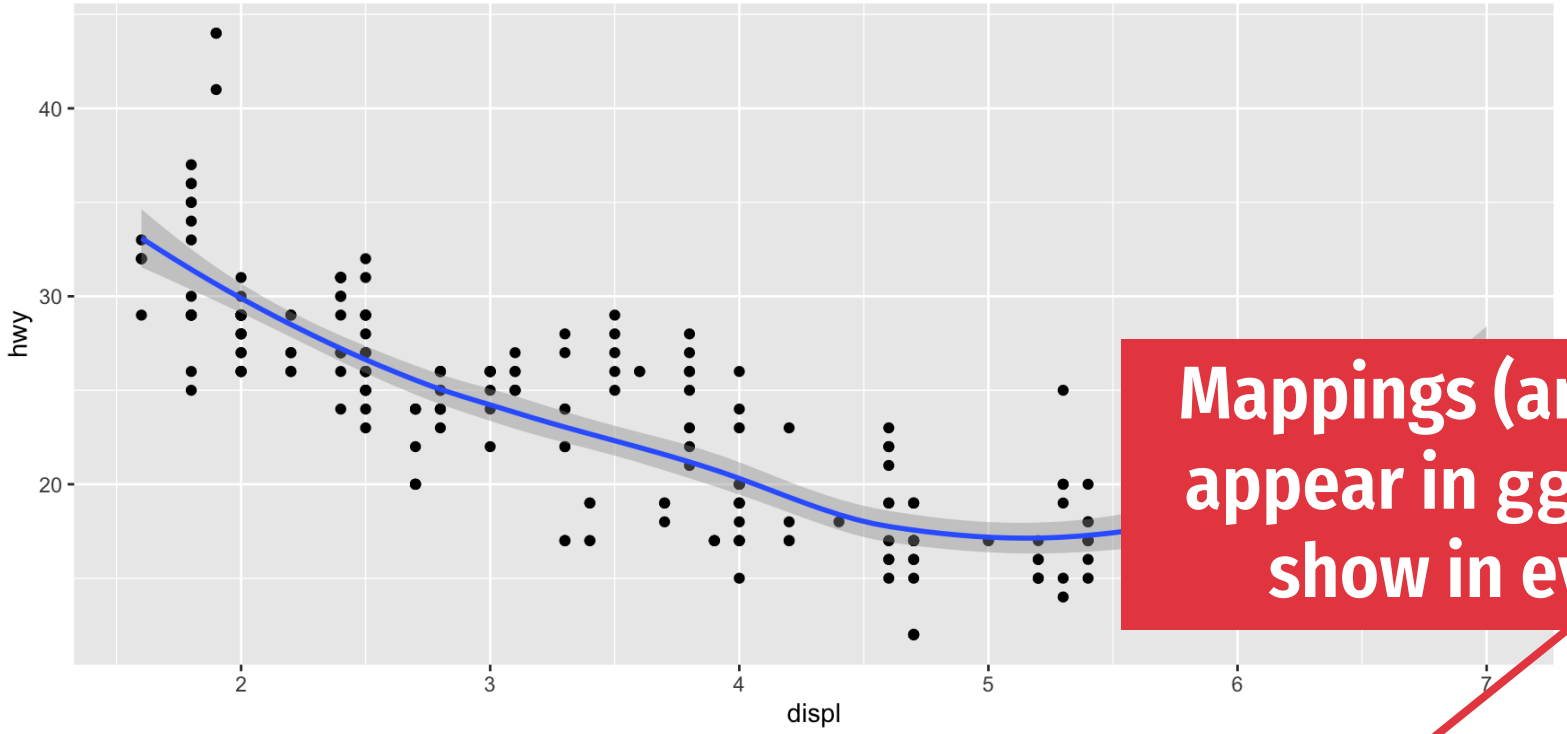
Your turn (#5)

With a partner, predict what this code will do.
Then run it.

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  geom_smooth(mapping = aes(x = displ, y = hwy))
```

02:00

Global vs. local

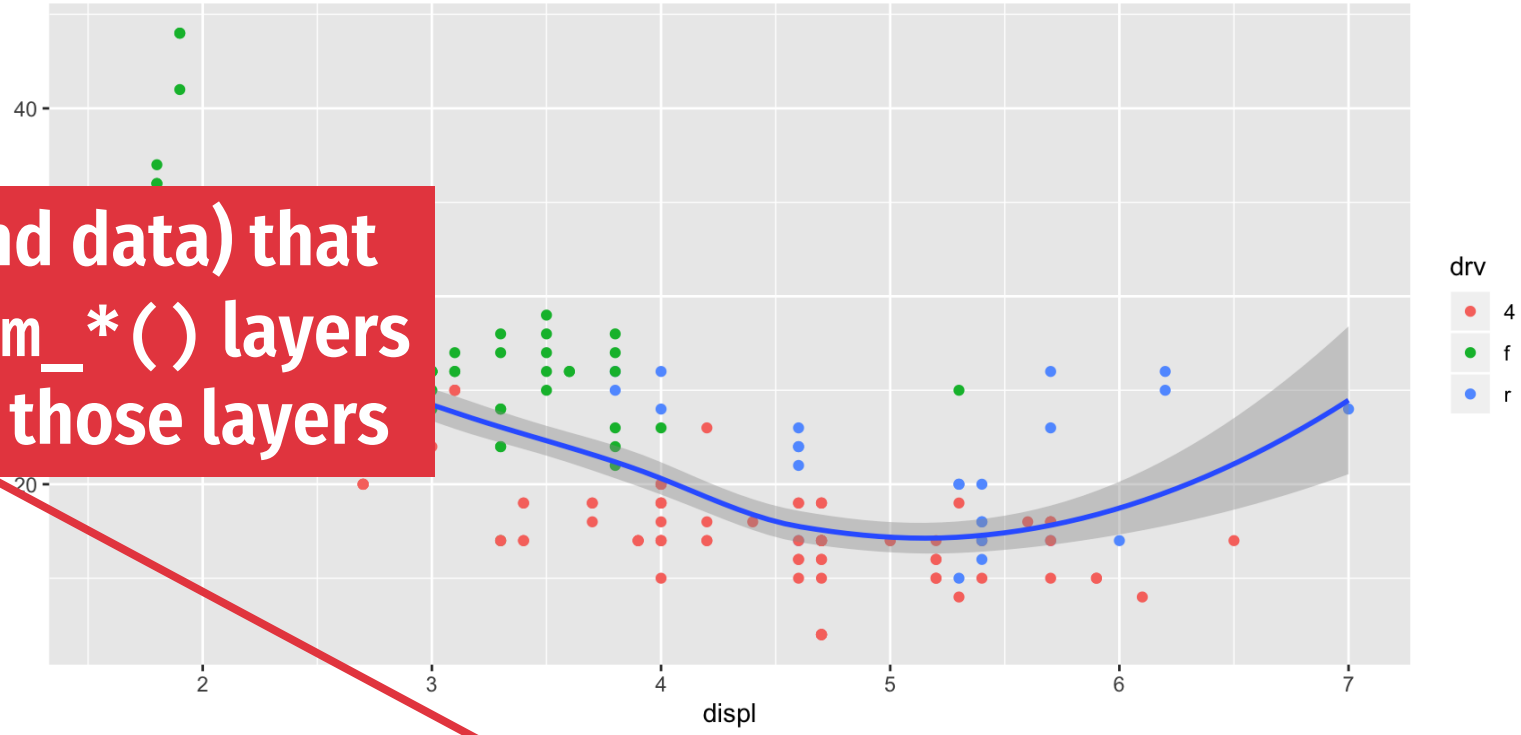


Mappings (and data) that appear in `ggplot()` will show in every layer

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()
```

Global vs. local

Mappings (and data) that appear in `geom_*()` layers only apply to those layers




```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point(mapping = aes(color = drv)) +
  geom_smooth()
```

Drawing lines

Essential parts of regression

Y	~	X (or lots of Xs)
Outcome variable		Explanatory variable
Response variable		Predictor variable
Dependent variable		Independent variable
Thing you want to explain or predict		Thing you use to explain changes in Y

Identify variables

A study examines the effect of smoking on lung cancer

You want to see if students taking more AP classes in high school improves their college grades

Researchers predict genocides by looking at negative media coverage, revolutions in neighboring countries, and economic growth

Netflix uses your past viewing history, the day of the week, and the time of the day to guess which show you want to watch next

02:00

Two purposes of regression

Prediction	Explanation
Forecast the future	**Explain effect of X on Y**
Focus is on Y	**Focus is on X**
Netflix trying to guess your next show	Netflix looking at the effect of time of day on show selection
Predicting who will escape poverty	Looking at the effect of food stamps on poverty reduction

How?

Plot X and Y

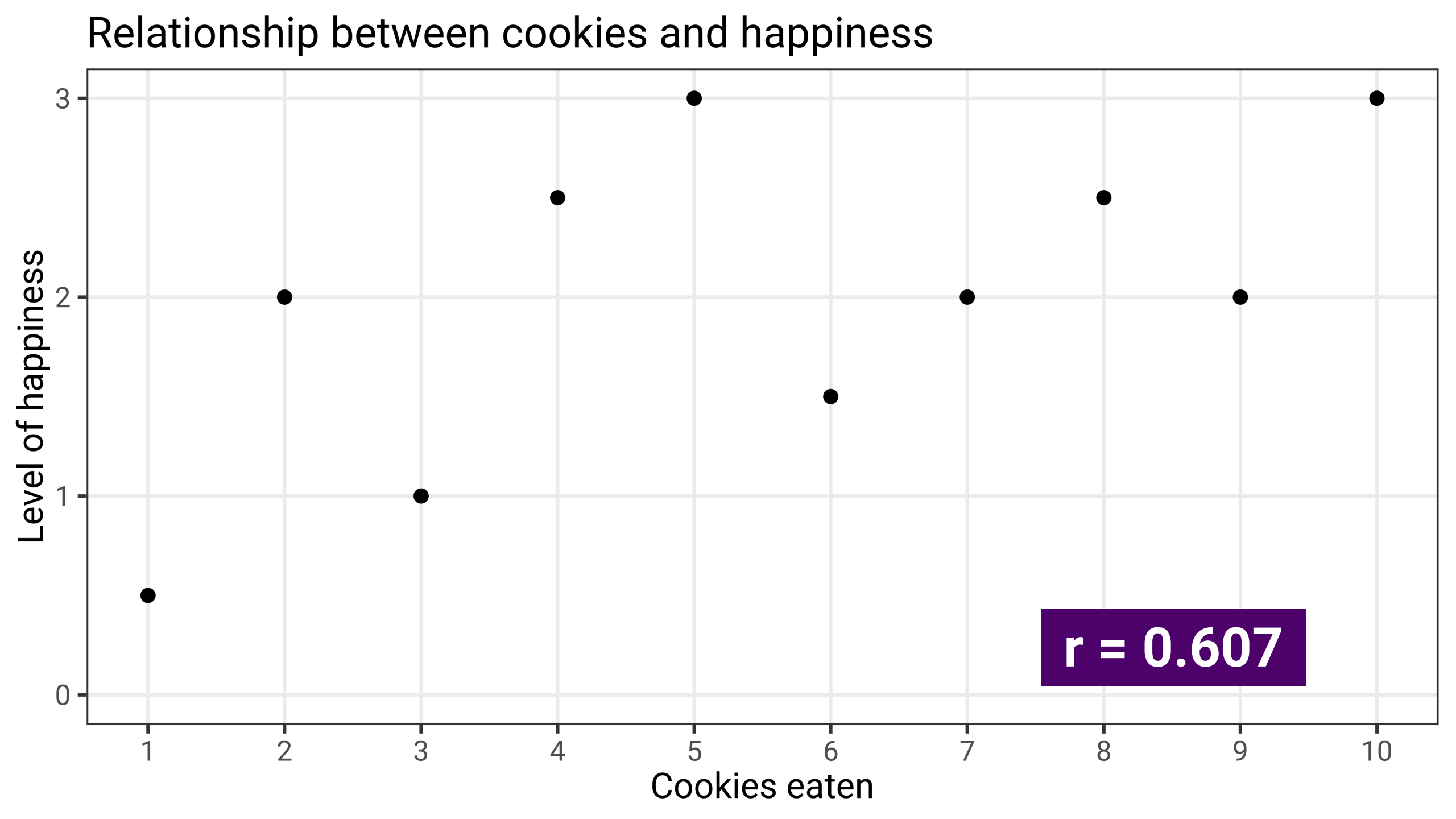
Draw a line that approximates the relationship

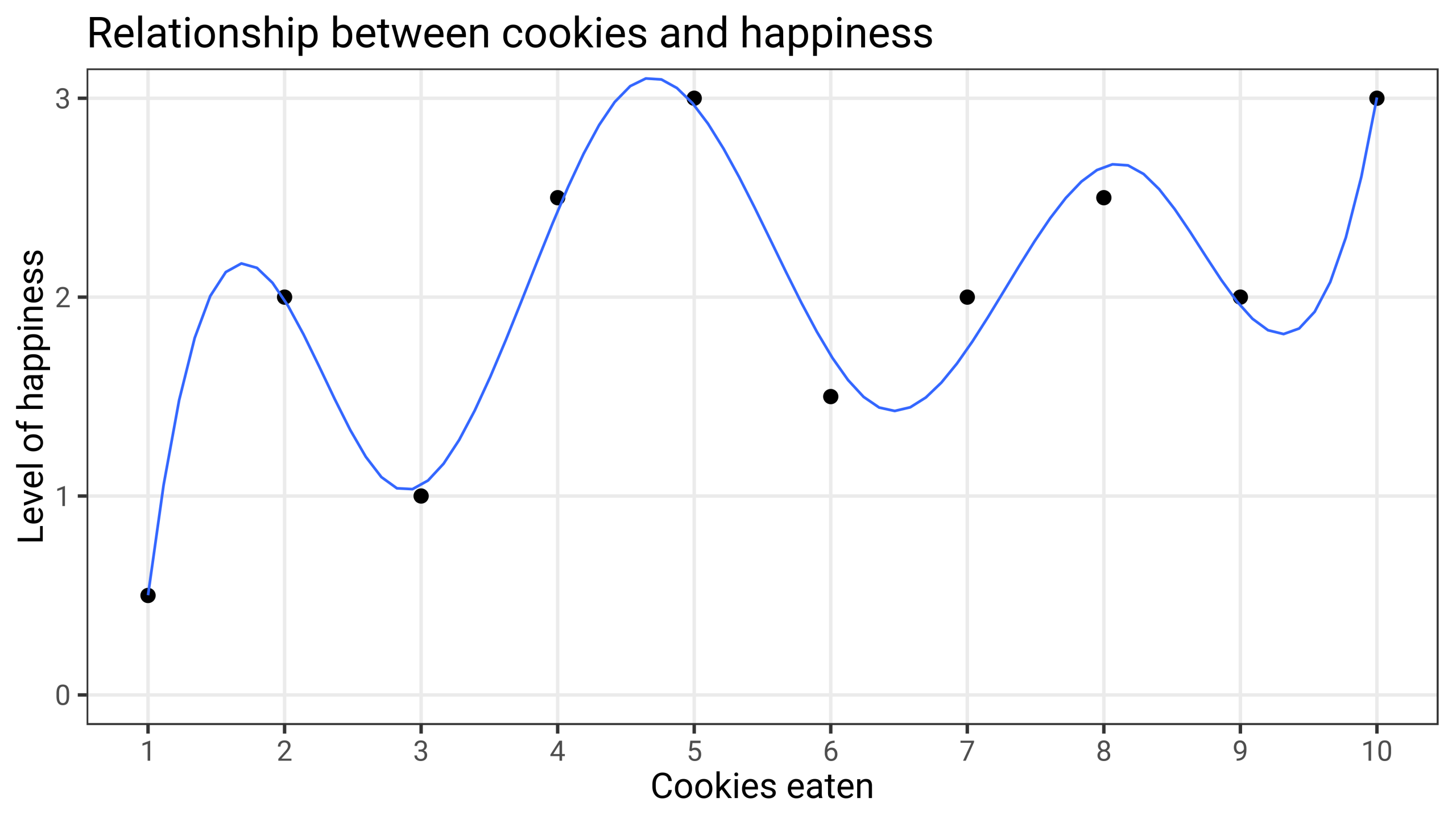
Find mathy parts of the line

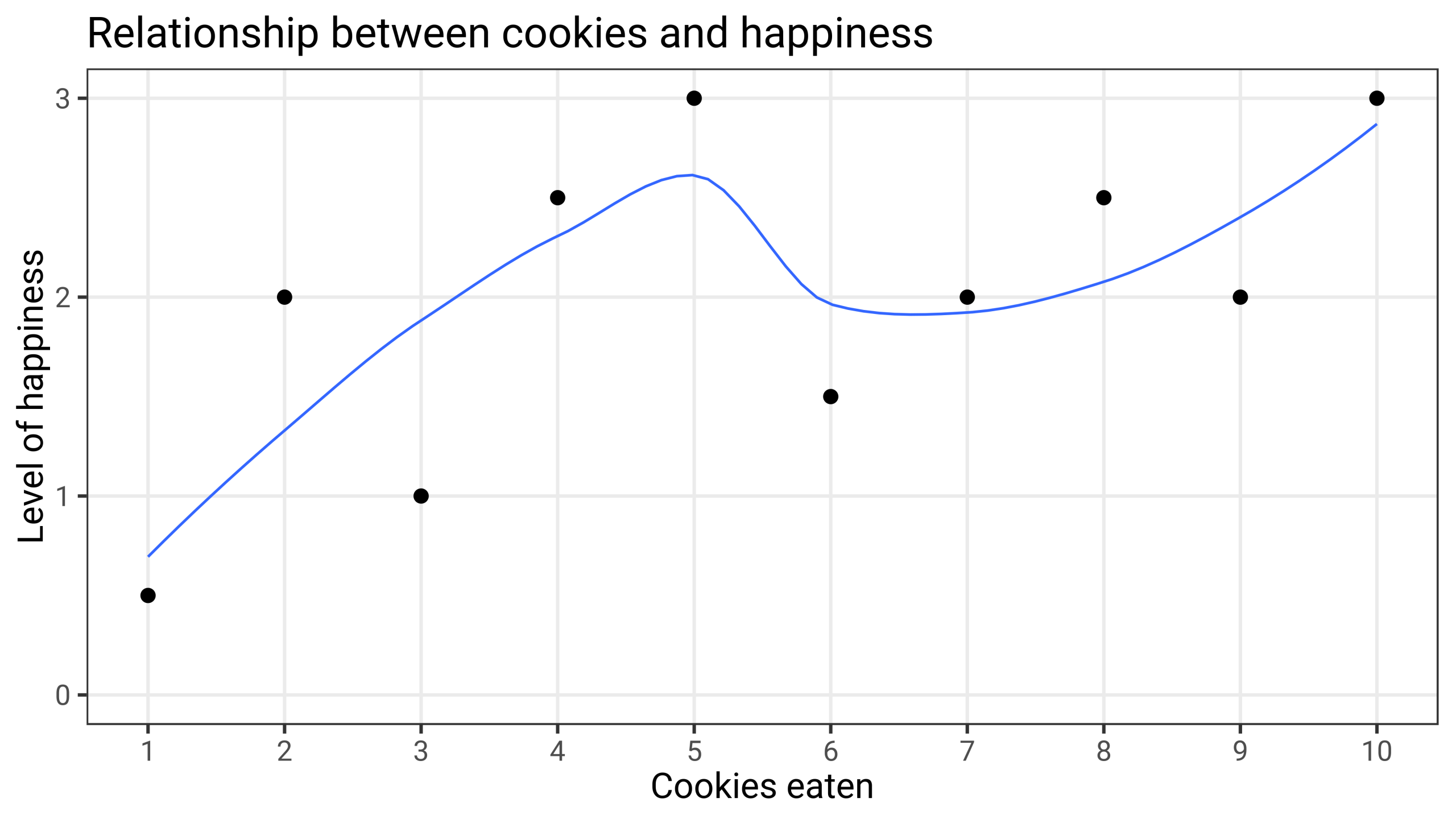
Interpret the math

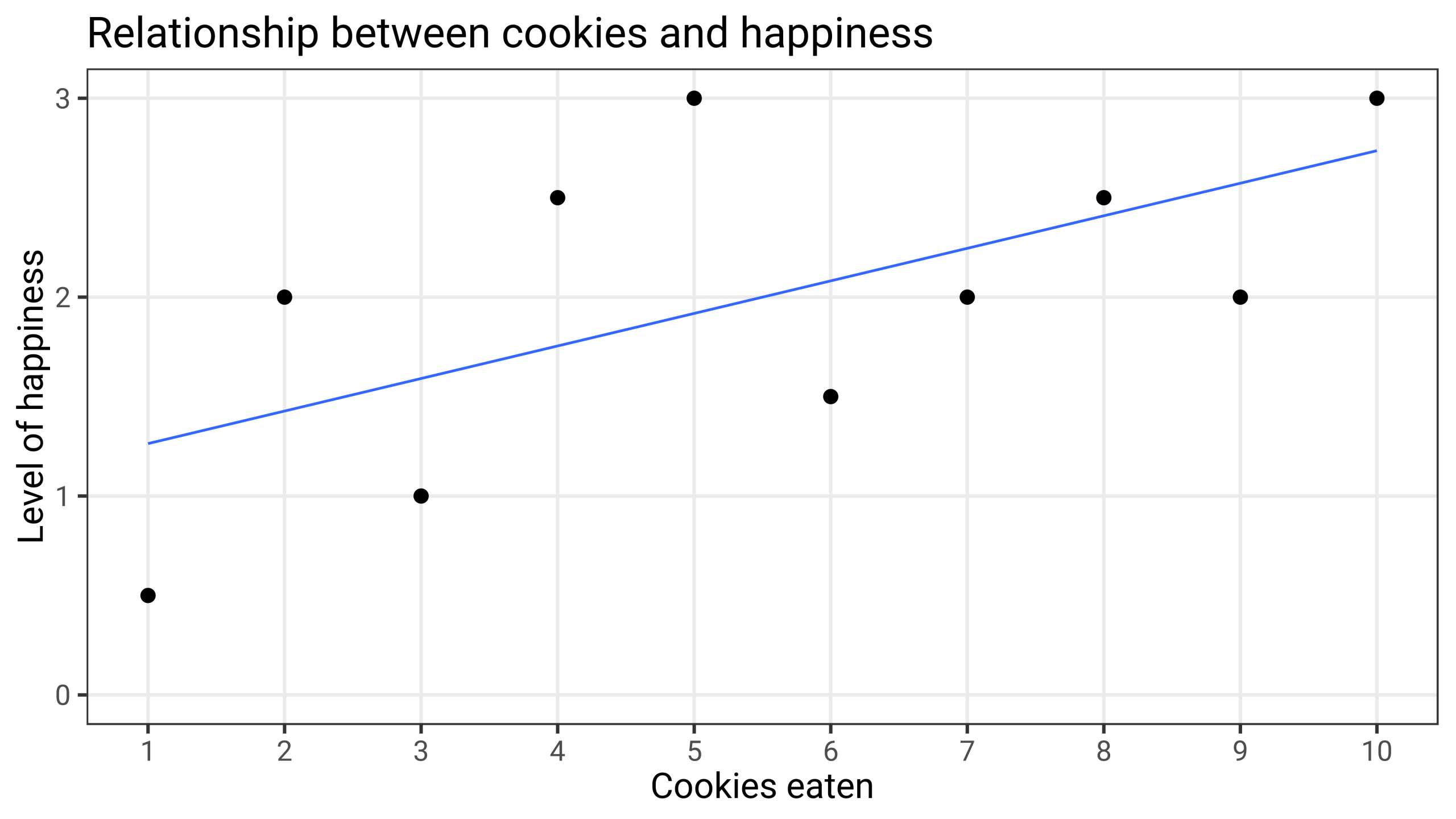
Cookies and happiness

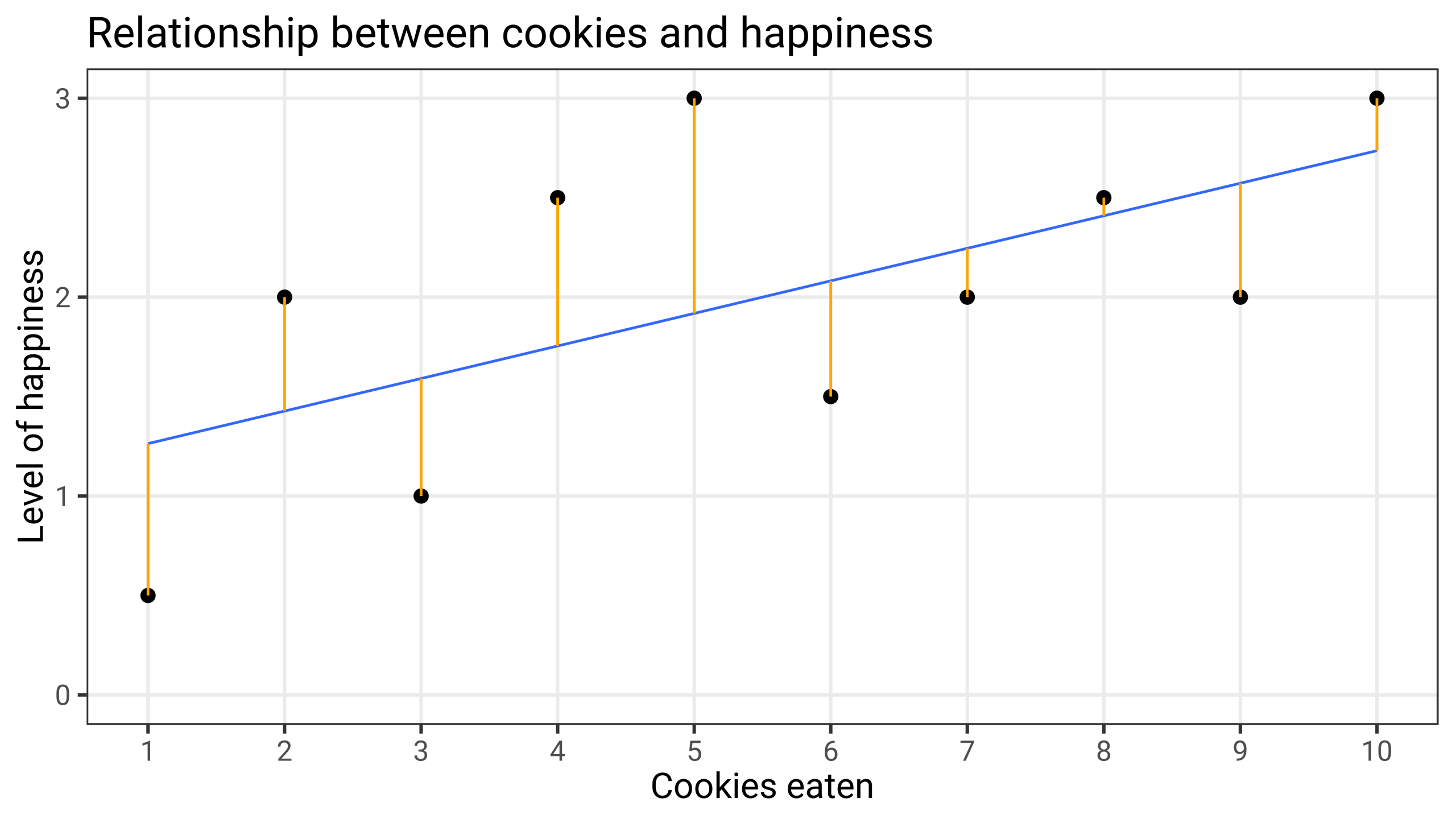
	happiness	cookies
1	0.5	1
2	2.0	2
3	1.0	3
4	2.5	4
5	3.0	5
6	1.5	6
7	2.0	7
8	2.5	8
9	2.0	9
10	3.0	10

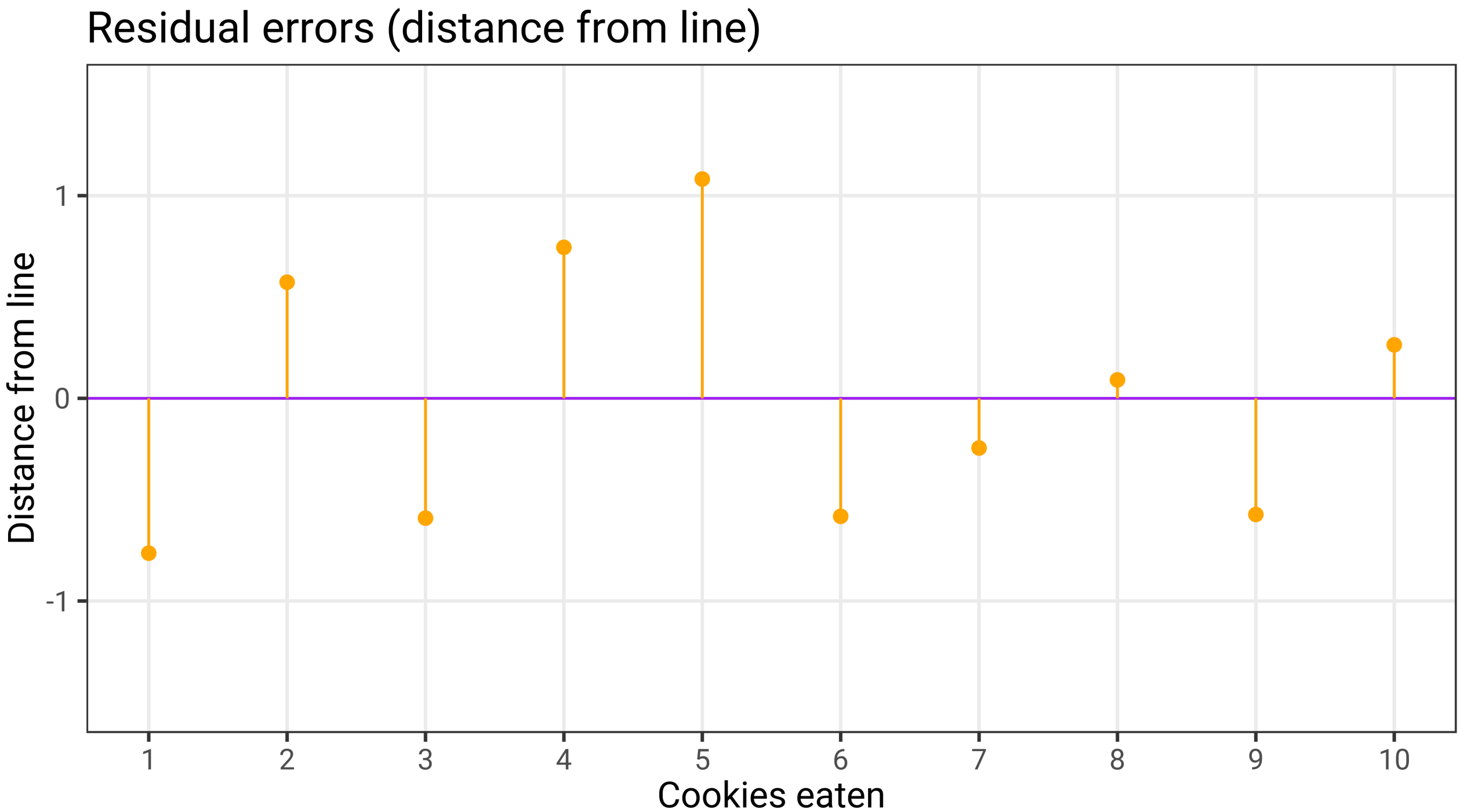

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10
r = 0.607

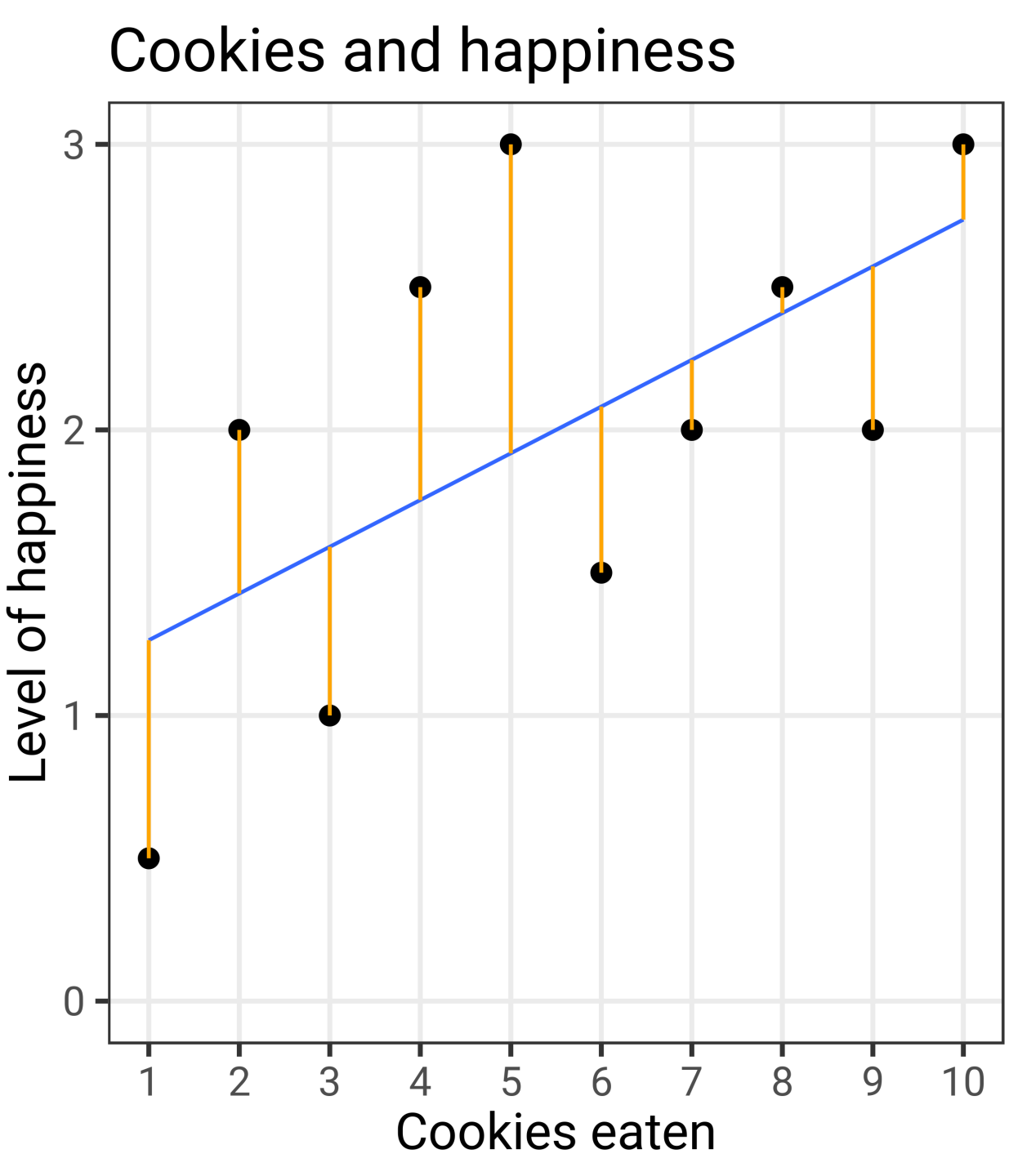

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10

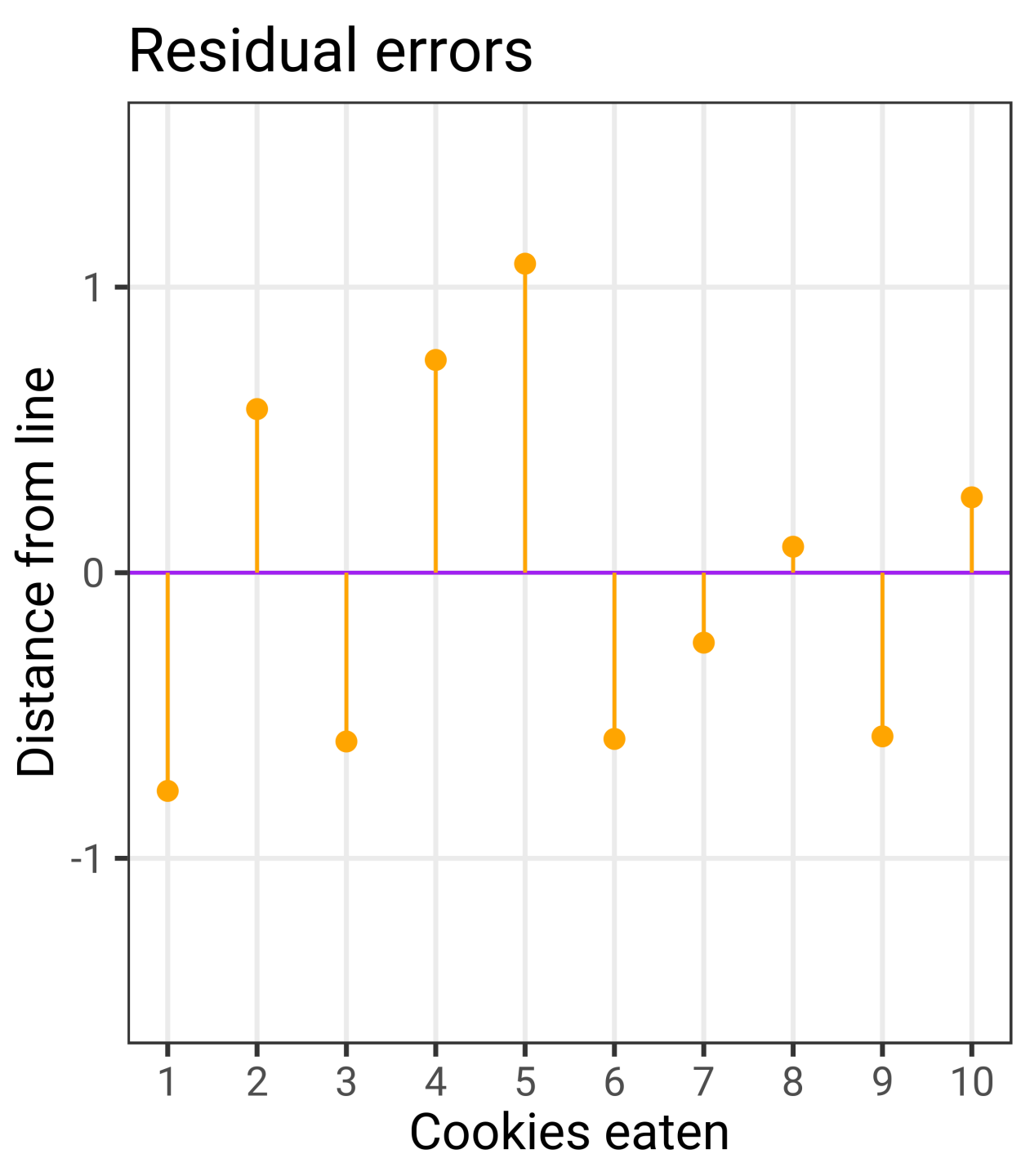

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10

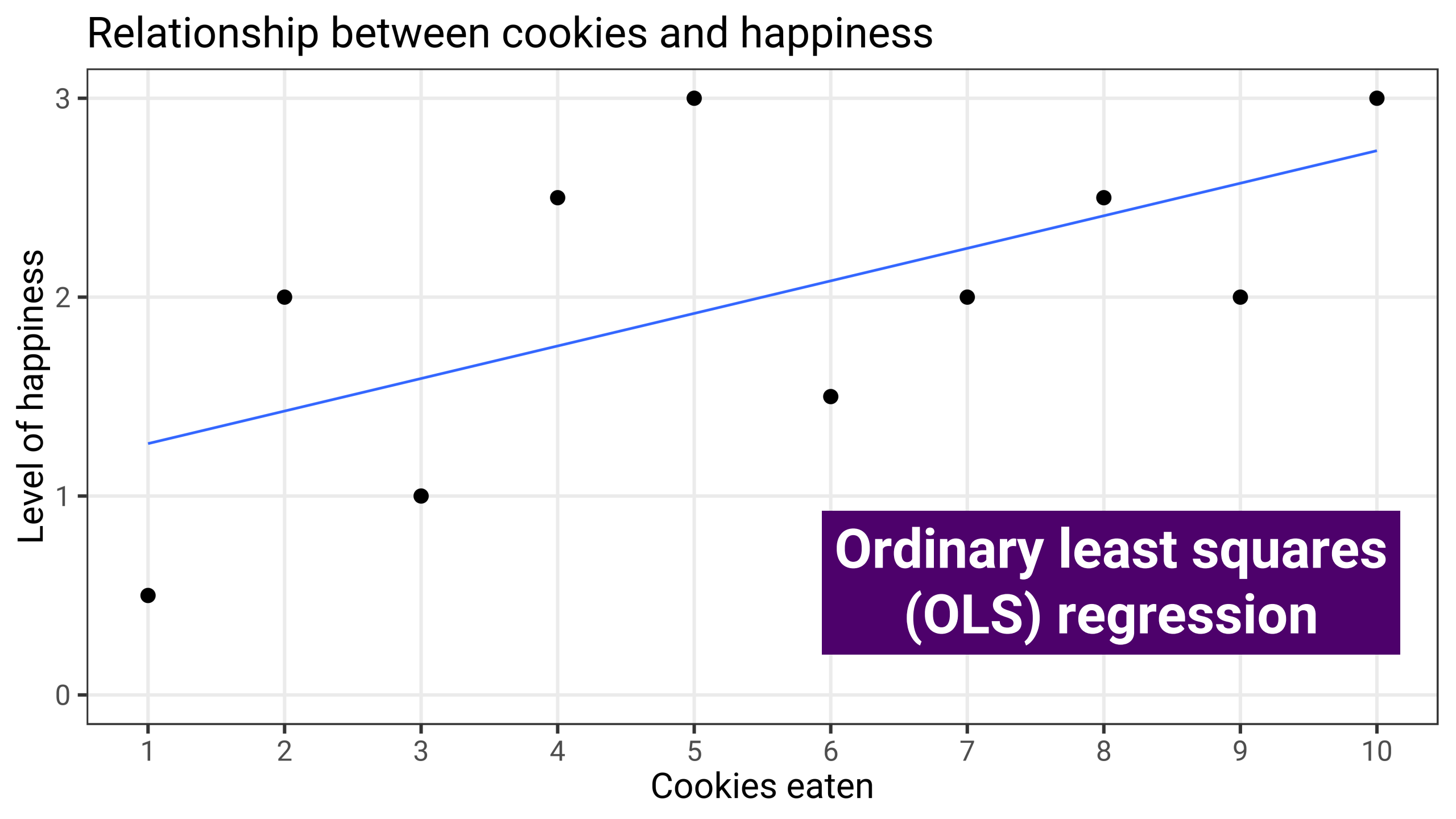

Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10


Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10


Residual errors (distance from line)
Distance from line
1
0
-1
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Cookies and happiness
Level of happiness
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Residual errors
Distance from line
1
0
-1
1
2
3
4
5
6
7
8
9
10
Cookies eaten


Relationship between cookies and happiness
Level of happiness
Cookies eaten
0
1
2
3
1
2
3
4
5
6
7
8
9
10
Ordinary least squares (OLS) regression

Lines, Greek, and regression

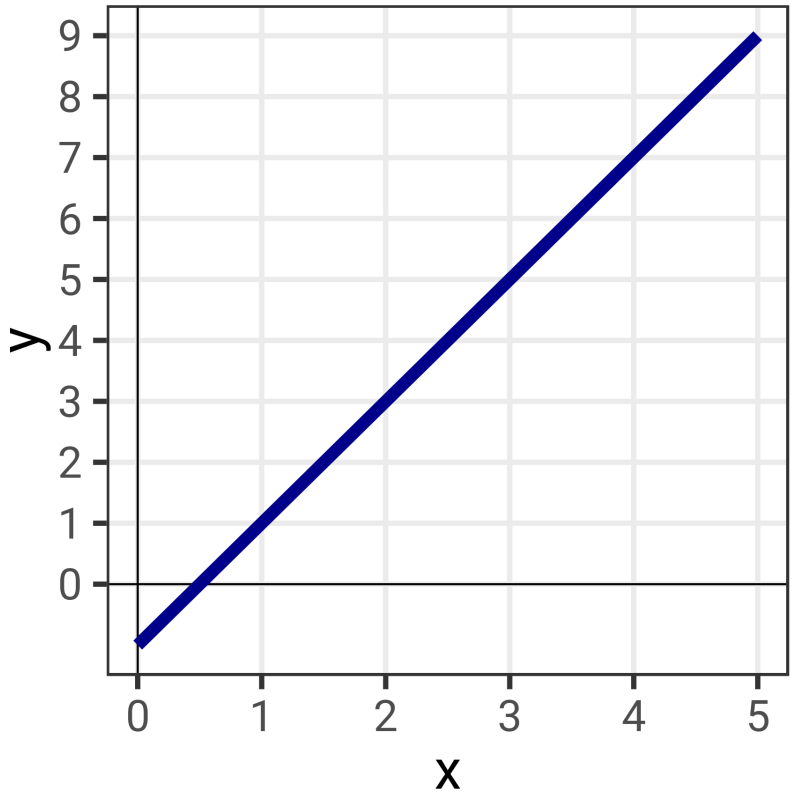
Drawing lines with math

$$y = mx + b$$

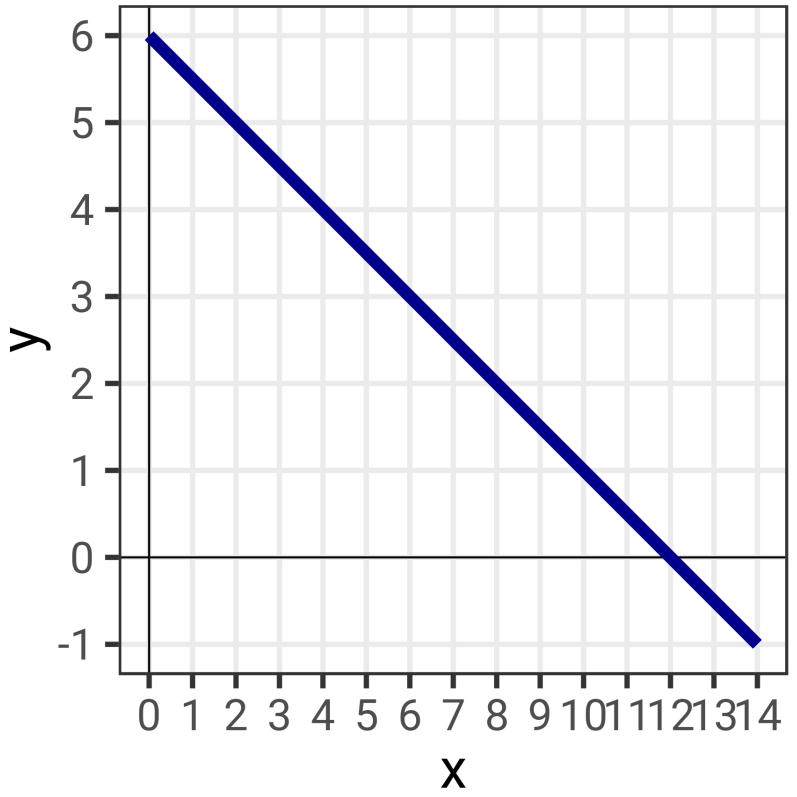
y	A number	
x	A number	
m	Slope	$\frac{rise}{run}$
b	y intercept	

Slopes and intercepts

$$y = 2x - 1$$



$$y = -0.5x + 6$$



Graph these

$y = 5x + 2$

$y = 1 - x$

$y = -2x + 11$

$y = 6 - 2x$

$y = -1 + 0.33x$

$y = 0.75x - 3$

02:00

Drawing lines with stats

$$\hat{y} = \beta_0 + \beta_1 x_1 + \varepsilon$$

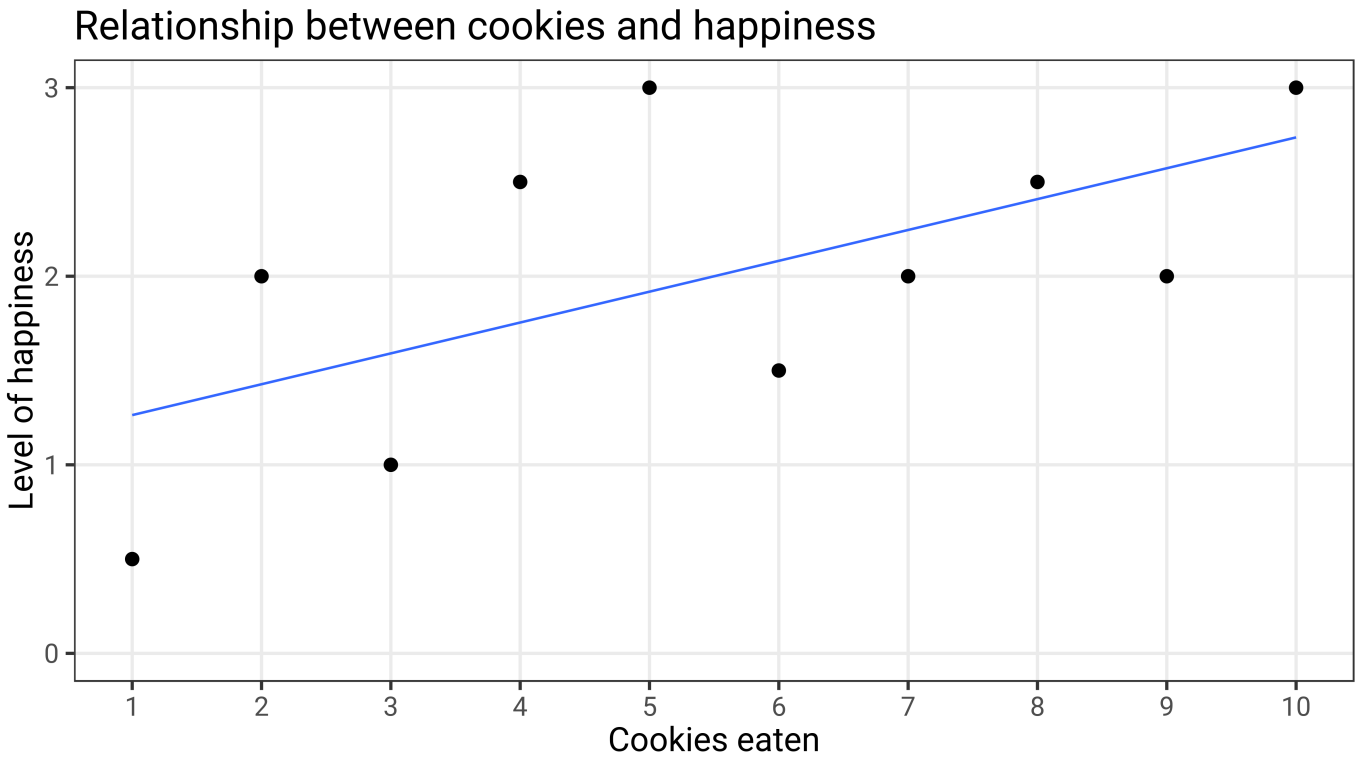
$y = mx + b$

y	$\hat{y}$	Outcome variable
x	x_1	Explanatory variable
m	β_1	Slope
b	β_0 (α)	y-intercept
	ε	Error (residuals)

Modeling cookies and happiness

$$\hat{y} = \beta_0 + \beta_1 x_1 + \epsilon$$

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{cookies} + \epsilon$$



OLS in R

```
name_of_model <- lm(<Y> ~ <X>, data = <DATA>)

summary(name_of_model)
tidy(name_of_model)
glance(name_of_model)
```

Modeling cookies and happiness

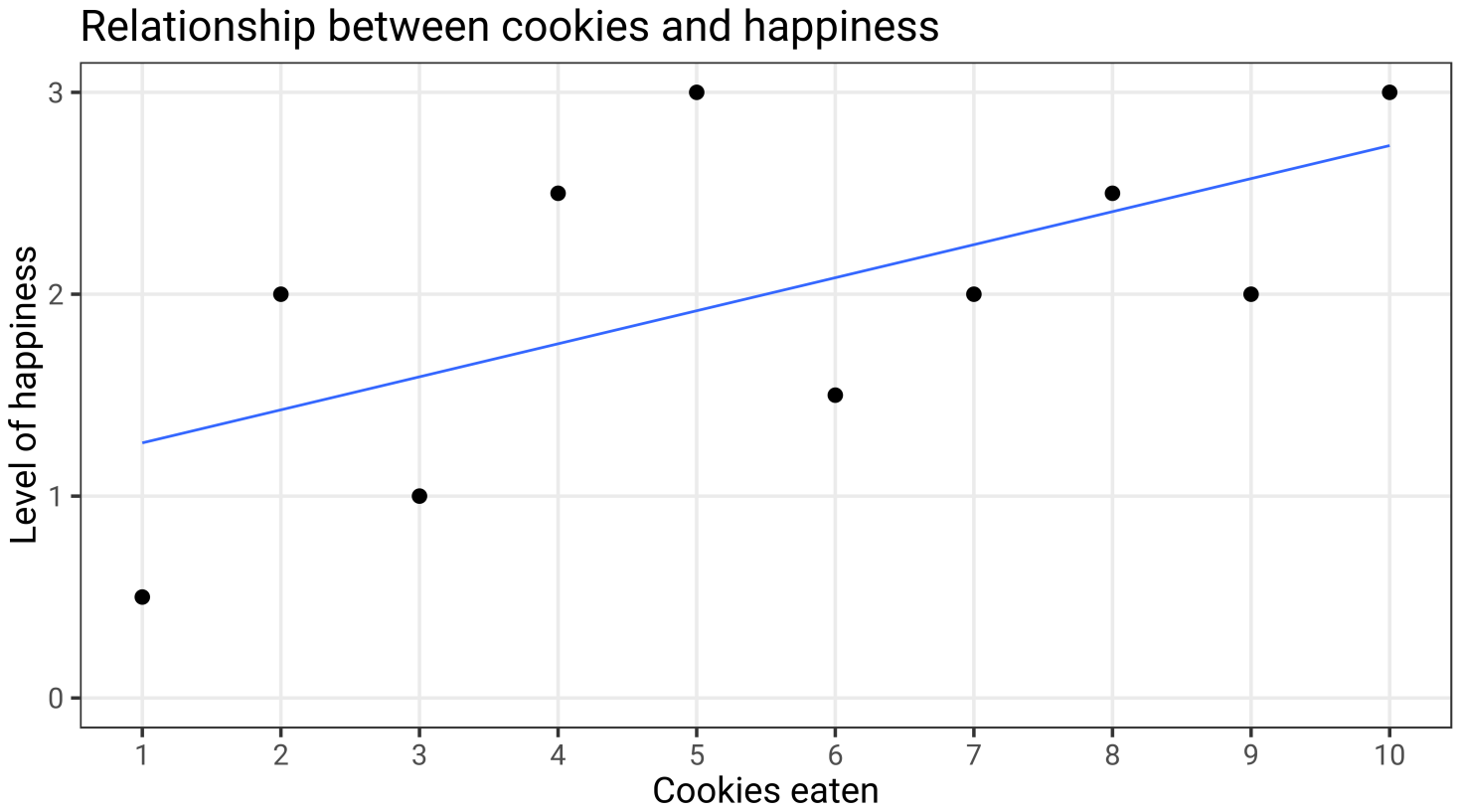
```r
cookies_model <- lm(happiness ~ cookies,
                    data = cookies_data)

tidy(cookies_model, conf.int = TRUE)
```

```
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept    1.1       0.47       2.34   0.047    0.016    2.18
2 cookies      0.164     0.076      2.16   0.063   -0.011    0.338
```

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{cookies} + \epsilon$$

$$\hat{\text{happiness}} = 1.1 + (0.164 \times \text{cookies}) + \epsilon$$



term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	1.1	0.47	2.339	0.047	0.016	2.184
cookies	0.164	0.076	2.159	0.063	-0.011	0.338

Template for continuous variables

A one unit increase in X is *associated* with a β_1 increase (or decrease) in Y, on average

$$\hat{\text{happiness}} = 1.1 + (0.164 \times \text{cookies}) + \epsilon$$

Econometric equations

$$ln\, Y_i = \alpha + \beta P_i + \gamma A_i + \delta_1 SAT_i + \delta_2 PI_i + \varepsilon_i$$

$\boldsymbol{Y}$: Income

$\boldsymbol{\alpha}$: Intercept

$\boldsymbol{\beta}$: Treatment

$\boldsymbol{P}$: Private school

$\boldsymbol{\gamma}$: Identification

$\boldsymbol{A}$: Group A

$\boldsymbol{\delta_1}$ and $\boldsymbol{\delta_2}$: Coefficients for control variables

$\boldsymbol{PI}$: Parental income

$\boldsymbol{SAT}$: SAT score

$\boldsymbol{\varepsilon}$: Error

These are all the same!

$$ln\, Y_i = \alpha + \beta P_i + \gamma A_i + \delta_1 SAT_i + \delta_2 PI_i + \varepsilon_i$$

$$ln\, Y = \beta_0 + \beta_1 P + \beta_2 A + \beta_3 SAT + \beta_4 PI + \varepsilon$$

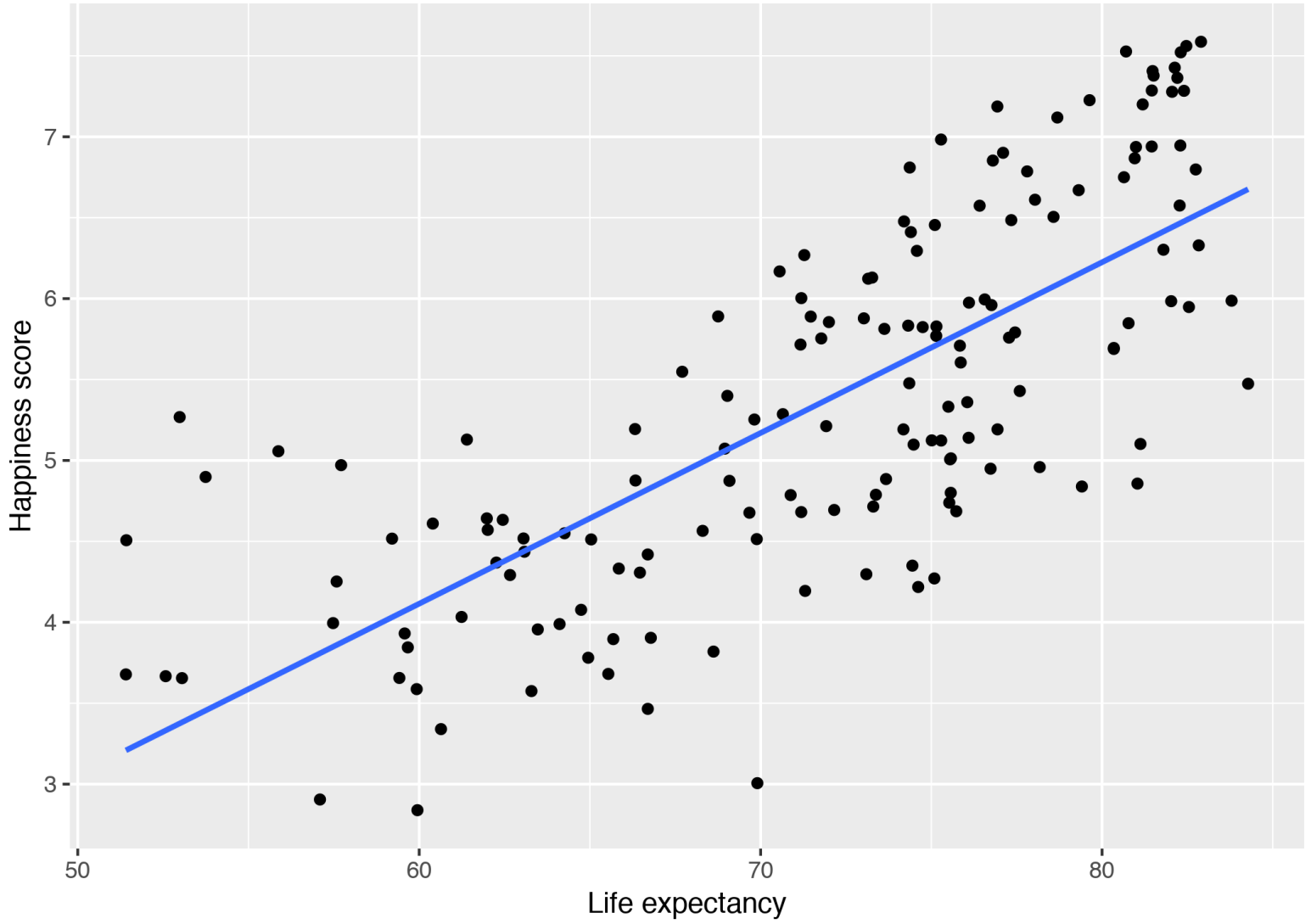
$$ln\, Y = \alpha + \beta_1 P + \beta_2 A + \beta_3 SAT + \beta_4 PI + \varepsilon$$

$$ln\, \text{Income} = \alpha + \beta_1 \text{Private} + \beta_2 \text{Group A} + \beta_3 \text{SAT score} + \beta_4 \text{Parental income} + \varepsilon$$

```
lm(log(income) ~ private + group_a + sat +
    parental_income, data = income_data)
```

Multiple regression

World happiness



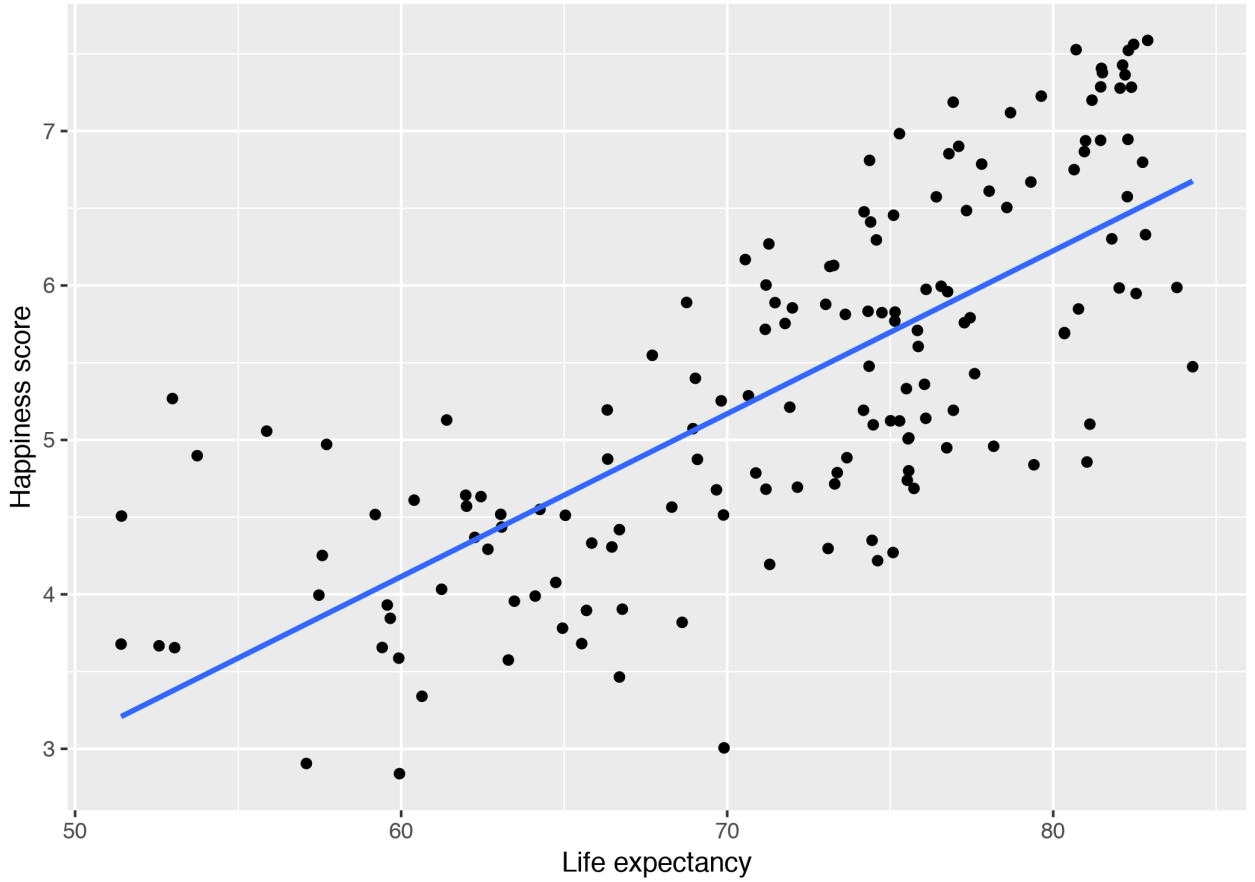
```r
model1 <- lm(happiness_score ~ life_expectancy,
             data = world_happiness)
tidy(model1)
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-2.215	0.556	-3.983	0	-3.313	-1.116
life_expectancy	0.105	0.008	13.73	0	0.09	0.121

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$

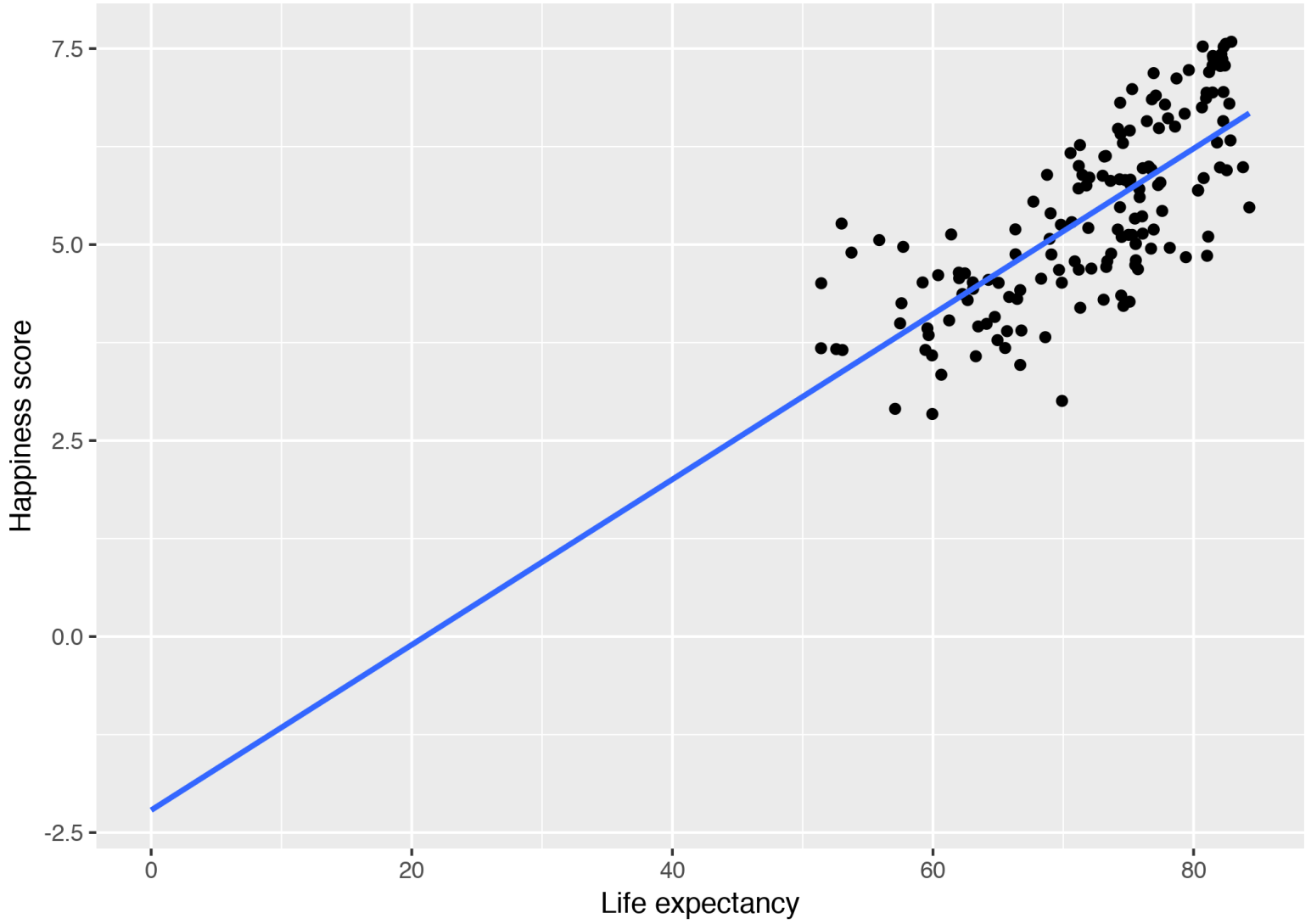
$$\hat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

World happiness



$$\hat{\text{happiness}} = -2.215 + (0.105 \times \text{life expectancy}) + \epsilon$$

World happiness



Variable types

Numeric variables

(Continuous)

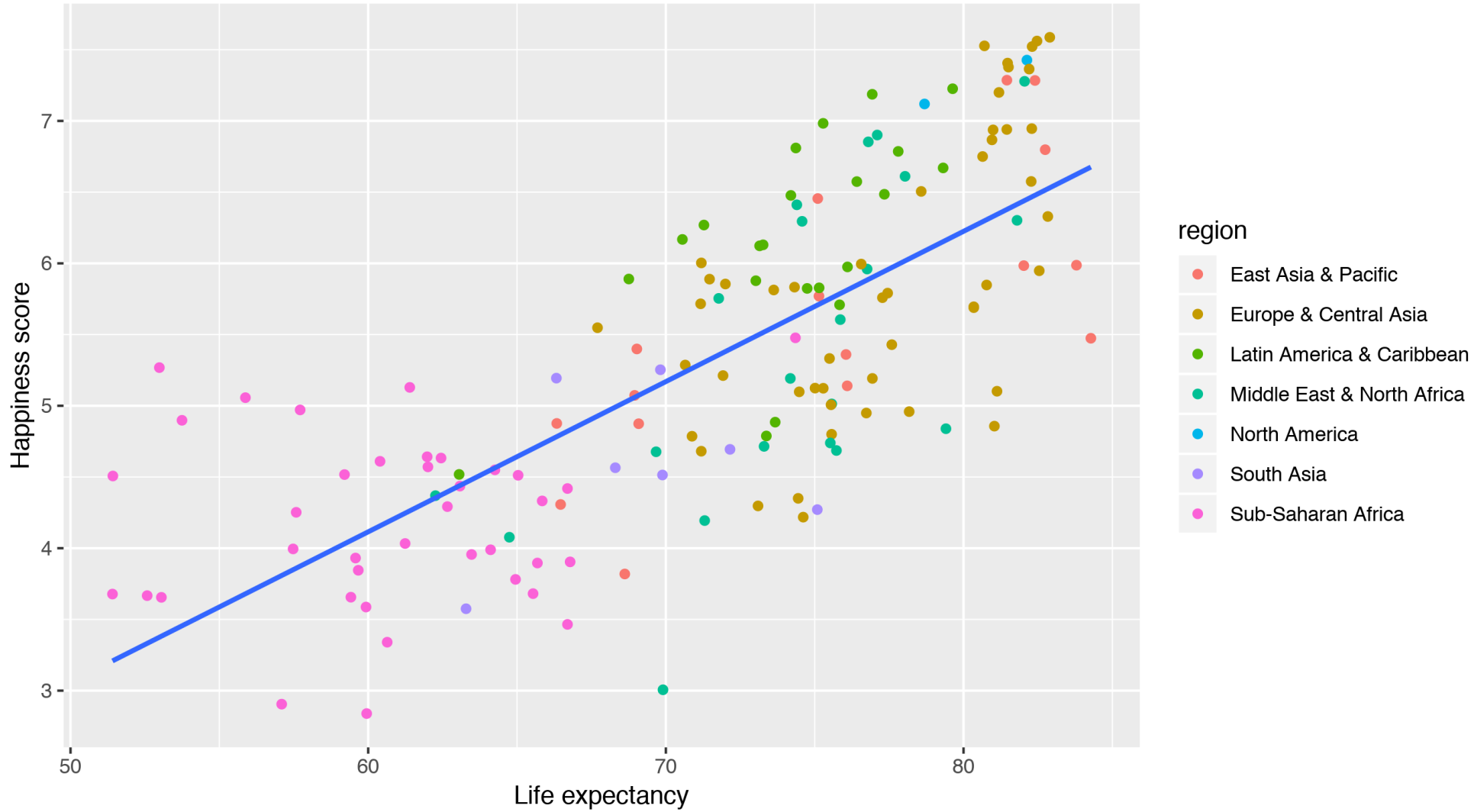
Numbers

Categorical variables

(Factors)

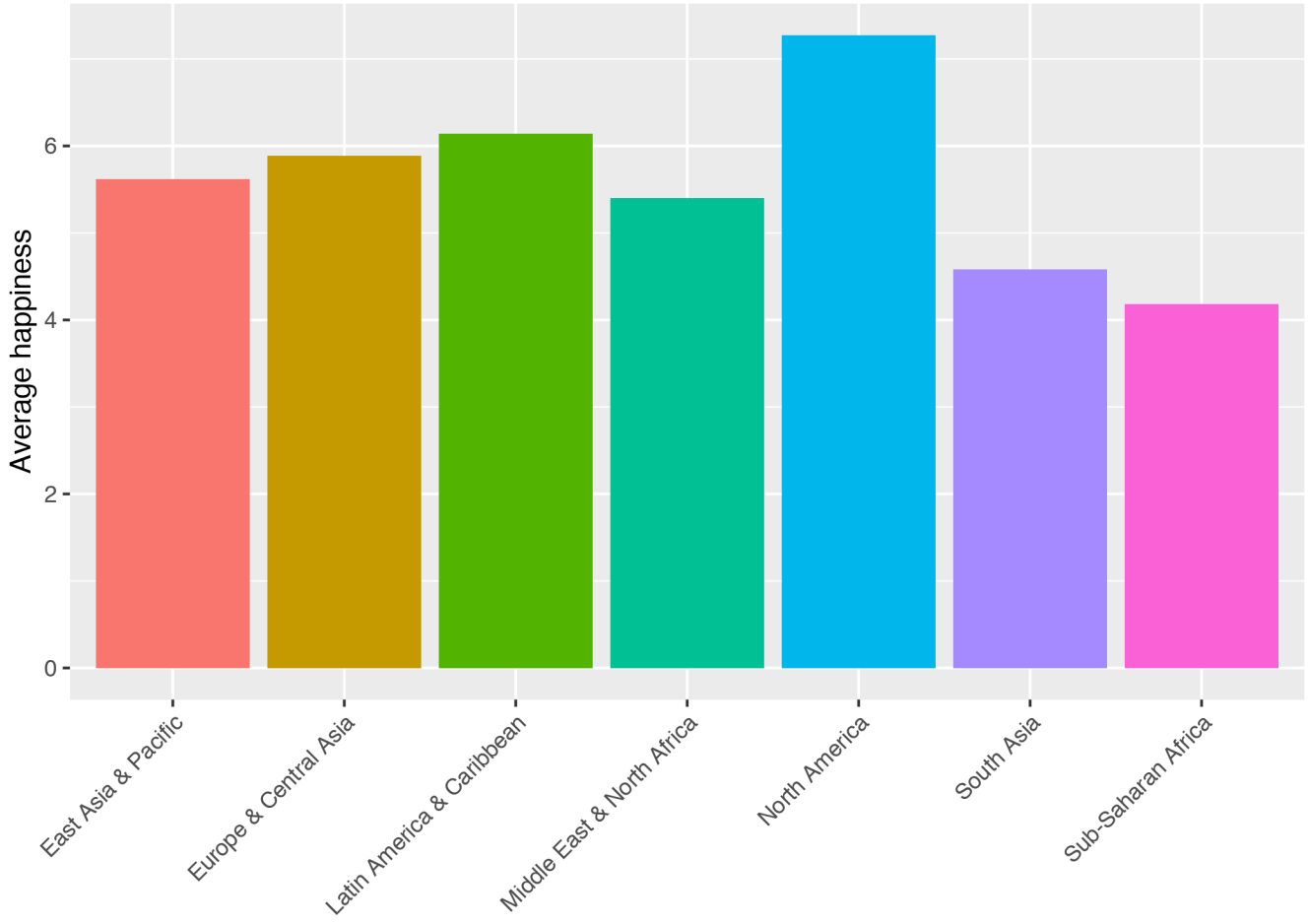
Not numbers

Life expectancy isn't the full story



Regional differences

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181



```r
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{Europe} + \beta_2 \text{Latin America} +$$
$$\beta_3 \text{MENA} + \beta_4 \text{North America} +$$
$$\beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon$$

```
model2 <- lm(happiness_score ~ region, data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	5.618	0.217	25.84	0
regionEurope & Central Asia	0.271	0.25	1.084	0.28
regionLatin America & Caribbean	0.527	0.286	1.844	0.067
regionMiddle East & North Africa	-0.214	0.289	-0.742	0.459
regionNorth America	1.655	0.652	2.538	0.012
regionSouth Asia	-1.037	0.394	-2.631	0.009
regionSub-Saharan Africa	-1.437	0.259	-5.544	0

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

Happiness in East Asia

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 0) + (0.527 \times 0) + (-0.214 \times 0) + (1.655 \times 0) + (-1.037 \times 0) + (-1.437 \times 0) + \epsilon$$

$$\hat{\text{happiness}} = 5.618$$

Happiness in Europe

$$\hat{\text{happiness}} = 5.618 + (0.271 \times \text{Europe}) + (0.527 \times \text{Latin America}) + (-0.214 \times \text{MENA}) + (1.655 \times \text{North America}) + (-1.037 \times \text{South Asia}) + (-1.437 \times \text{Sub-Saharan Africa}) + \epsilon$$

$$\hat{\text{happiness}} = 5.618 + (0.271 \times 1) + (0.527 \times 0) + (-0.214 \times 0) + (1.655 \times 0) + (-1.037 \times 0) + (-1.437 \times 0) + \epsilon$$

$$\begin{aligned}\hat{\text{happiness}} &= 5.618 + (0.271 \times 1) \\ &= 5.889\end{aligned}$$

Regression coefficients

term	estimate
intercept	5.618
regionEurope & Central Asia	0.271
regionLatin America & Caribbean	0.527
regionMiddle East & North Africa	-0.214
regionNorth America	1.655
regionSouth Asia	-1.037
regionSub-Saharan Africa	-1.437

Averages

region	avg
East Asia & Pacific	5.618
Europe & Central Asia	5.889
Latin America & Caribbean	6.145
Middle East & North Africa	5.404
North America	7.273
South Asia	4.581
Sub-Saharan Africa	4.181

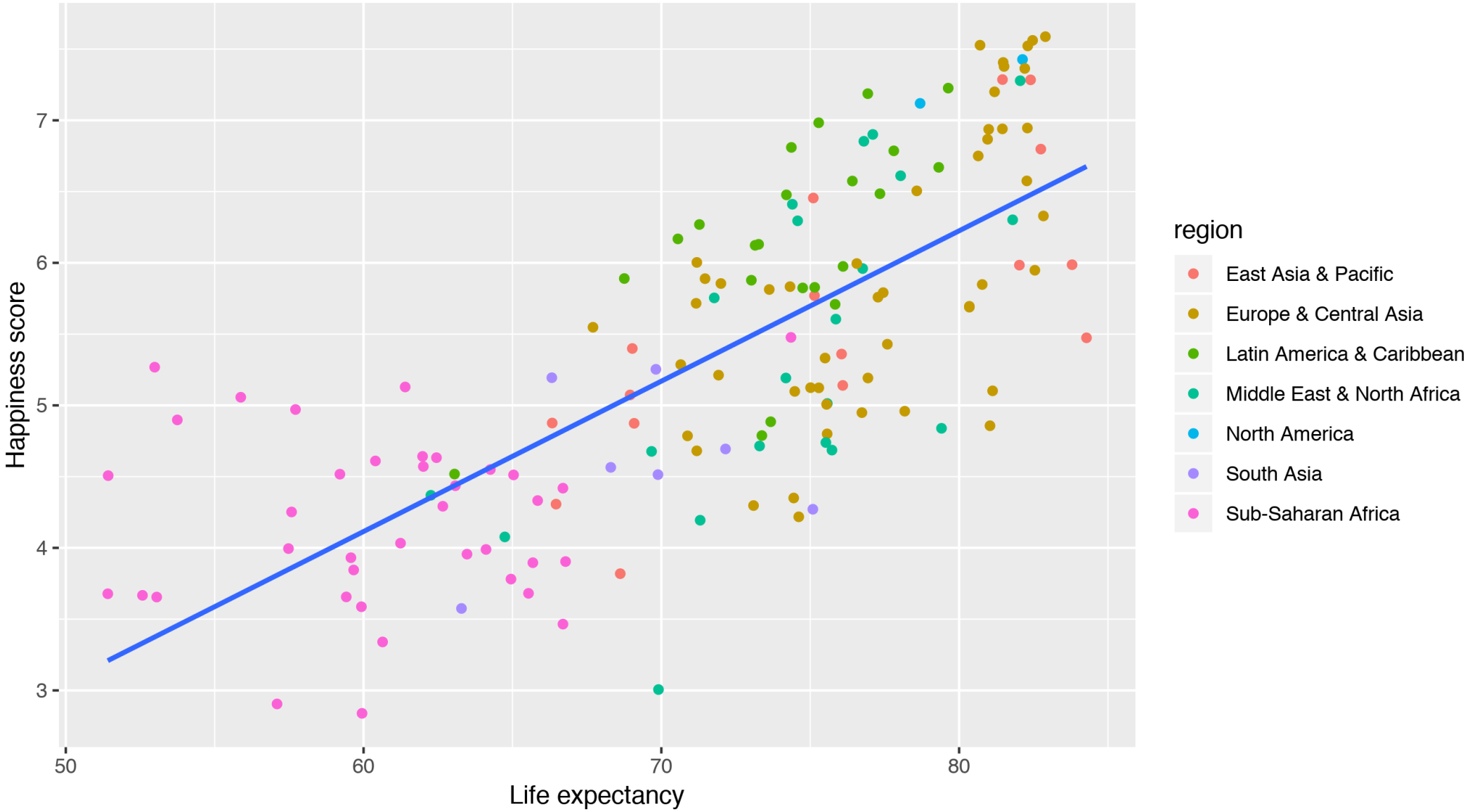
Template for indicator variables

On average, y is β_n units larger (or smaller) in x_n, compared to x_0

On average, national happiness is 1.65 points higher in North America than in East Asia

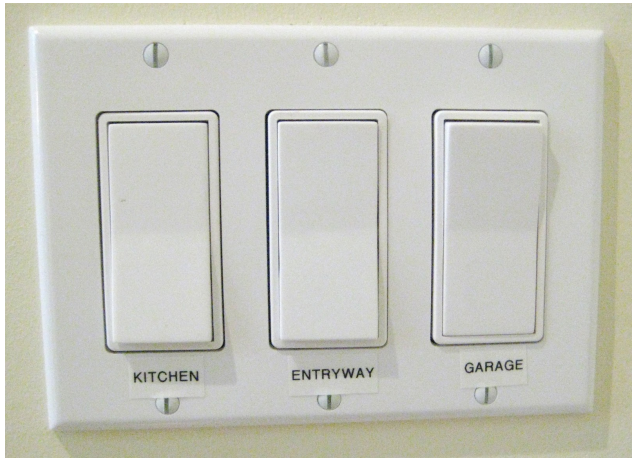
On average, compared to East Asia, national happiness is 1.44 points lower in Sub Saharan Africa

Getting closer



Sliders and switches

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$



$$\begin{aligned}\hat{\text{happiness}} = \beta_0 &+ \beta_1 \text{Europe} + \beta_2 \text{Latin America} + \\ &\beta_3 \text{MENA} + \beta_4 \text{North America} + \\ &\beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon\end{aligned}$$

All at once!

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

```r
model_life_school <- lm(happiness_score ~ life_expectancy +
                          school_enrollment,
                        data = world_happiness)
```

term	estimate	std_error	statistic	p_value	lower_ci
intercept	-2.111	0.835	-2.529	0.013	-3.767
life_expectancy	0.101	0.014	7.447	0	0.074
school_enrollment	0.003	0.01	0.331	0.741	-0.016

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} + \epsilon$$

$$\hat{\text{happiness}} = -2.11 + (0.101 \times \text{life expectancy}) + (0.003 \times \text{school enrollment}) + \epsilon$$

Filtering out variation

Each x in the model explains some portion of the variation in y

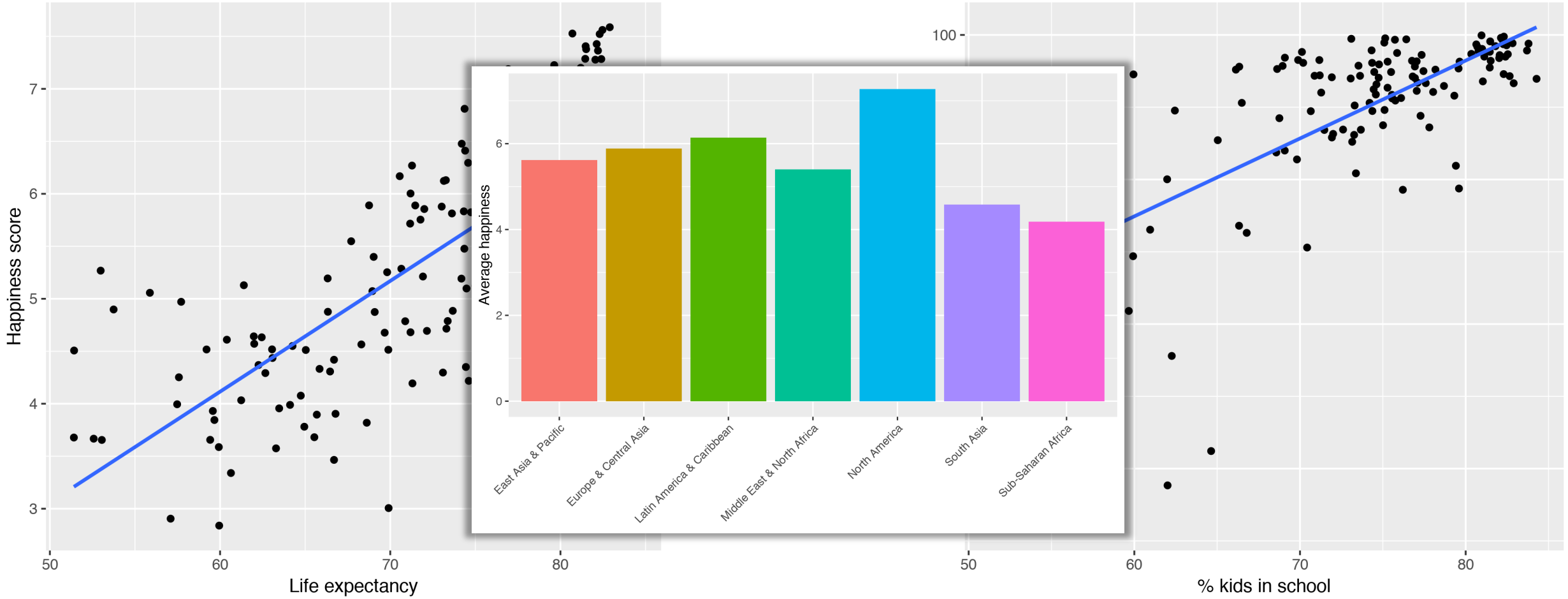
Interpretation is a little trickier, since you can only ever move **one** switch or slider (or variable)

Template for multiple regression

Taking all other variables in the model into account, a one unit increase in x_n is associated with a β_n increase (or decrease) in y, on average

Controlling for school enrollment, a 1 year increase in life expectancy is associated with a 0.1 point increase in national happiness, on average

Happiness ~ Life + School + Region



```r
model_life_school_region <-
  lm(happiness_score ~ life_expectancy + school_enrollment + region,
     data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	-2.821	1.355	-2.083	0.04
life_expectancy	0.102	0.017	5.894	0
school_enrollment	0.008	0.01	0.785	0.435
regionEurope & Central Asia	0.031	0.255	0.123	0.902
regionLatin America & Caribbean	0.732	0.294	2.489	0.015
regionMiddle East & North Africa	0.189	0.317	0.597	0.552
regionNorth America	1.114	0.581	1.917	0.058
regionSouth Asia	-0.249	0.45	-0.553	0.582
regionSub-Saharan Africa	0.326	0.407	0.802	0.425

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$