Introduction

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

datacamp

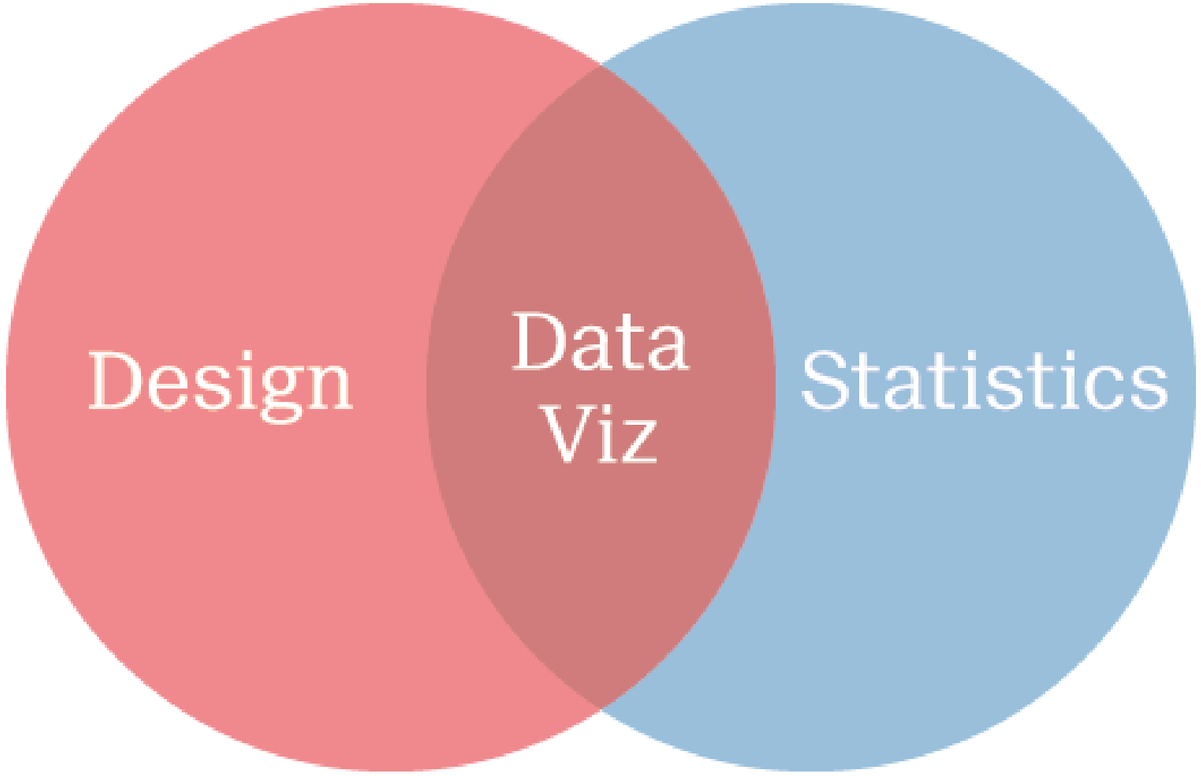
Your instructor - Rick Scavetta

- e-mail: office@scavetta.academy

- Twitter: @Rick_Scavetta

Data visualization & data science

- A core skill in Data Science.



Exploratory versus explanatory

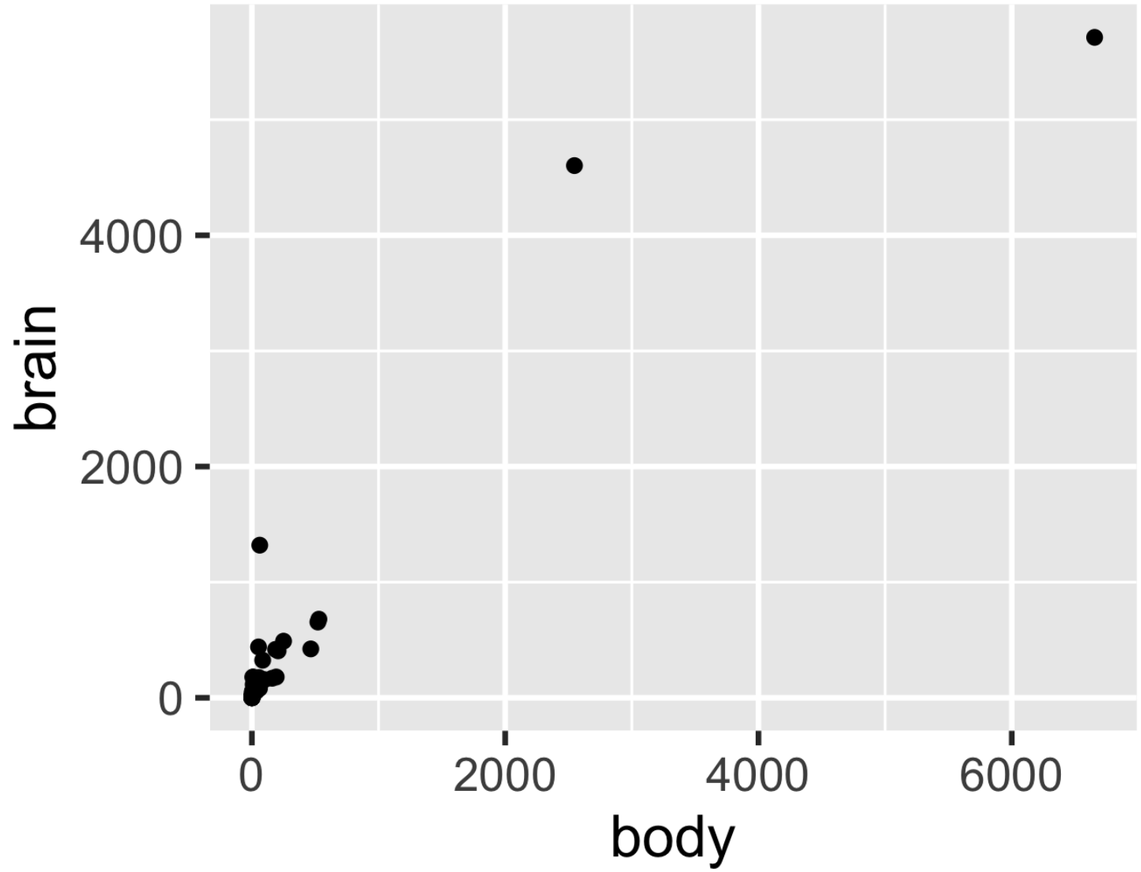
MASS::mammals

```
MASS::mammals
```

```
                          body   brain
Arctic fox               3.385   44.50
Owl monkey               0.480   15.50
Mountain beaver          1.350    8.10
Cow                    465.000  423.00
Grey wolf               36.330  119.50
Goat                    27.660  115.00
Roe deer                14.830   98.20
...
Pig                    192.000  180.00
Echidna                  3.000   25.00
Brazilian tapir        160.000  169.00
Tenrec                   0.900    2.60
Phalanger                1.620   11.40
Tree shrew               0.104    2.50
Red fox                  4.235   50.40
```

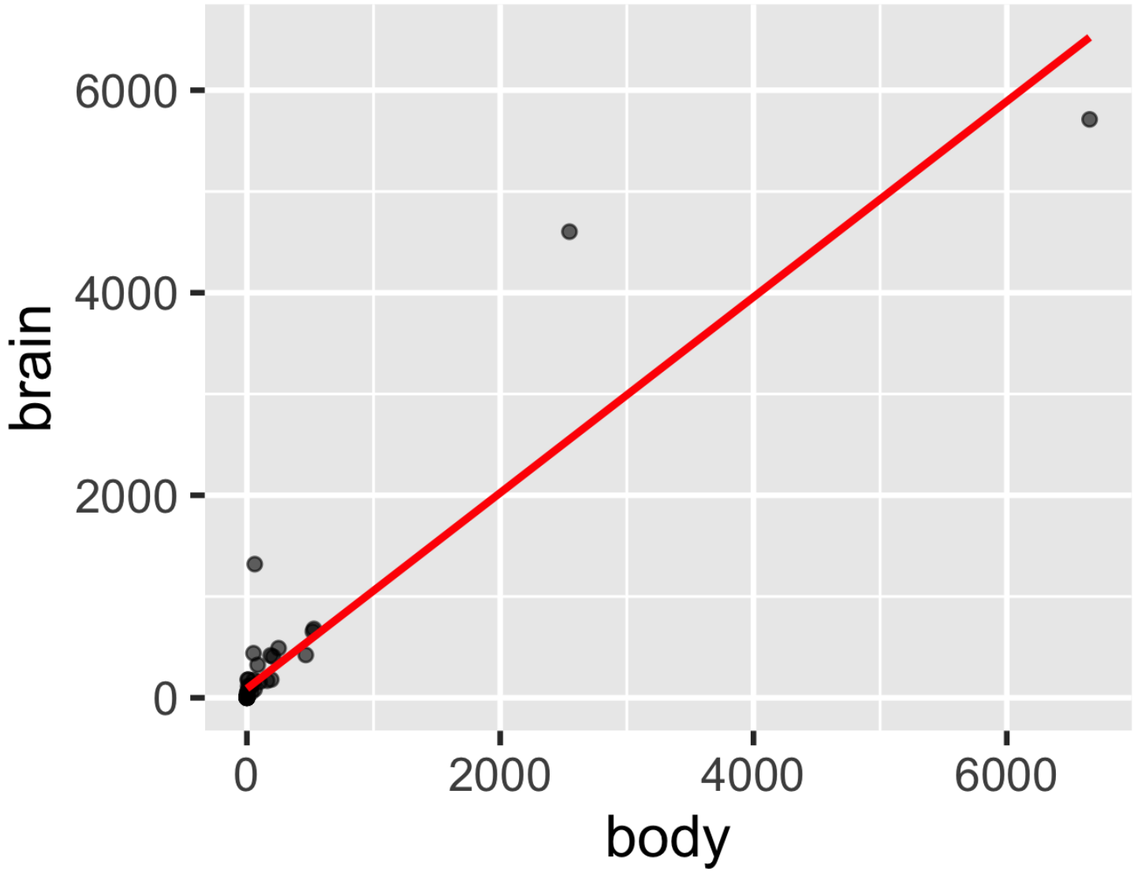
A scatter plot

```
ggplot(mammals, aes(x = body, y = brain)) +
  geom_point()
```



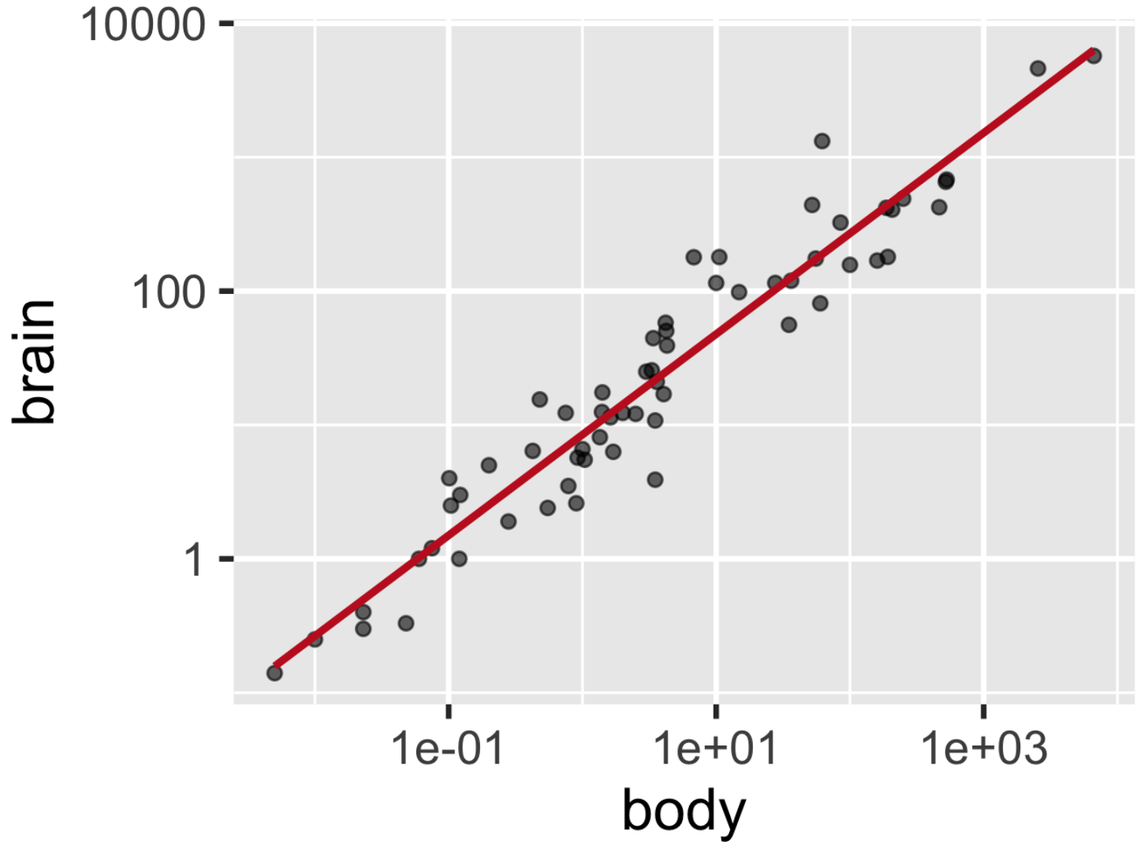
Explore with a linear model

```
ggplot(mammals, aes(x = body, y = brain)) +
  geom_point(alpha = 0.6) +
  stat_smooth(
    method = "lm",
    color = "red",
    se = FALSE
  )
```

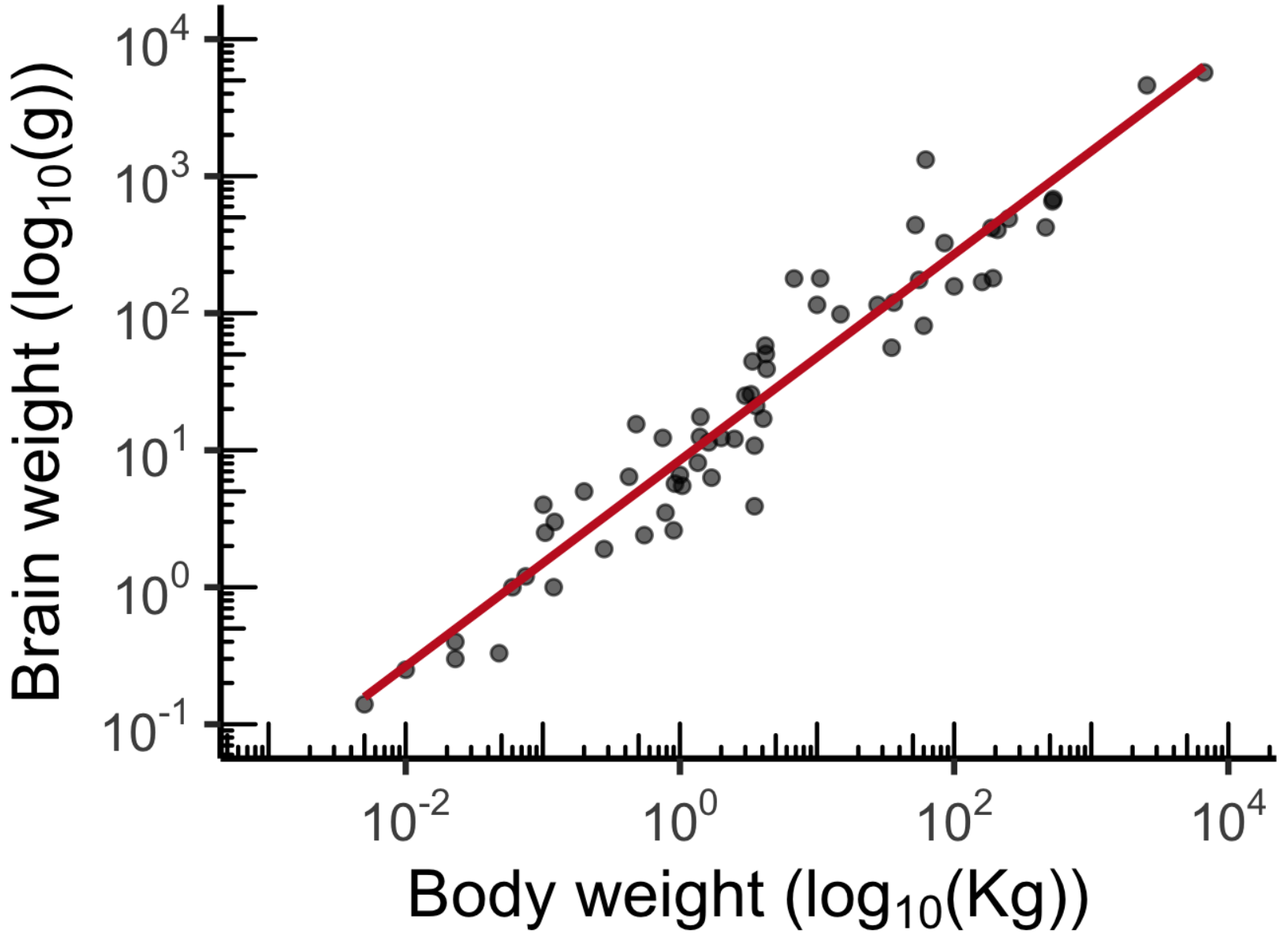


Explore: fine-tuning

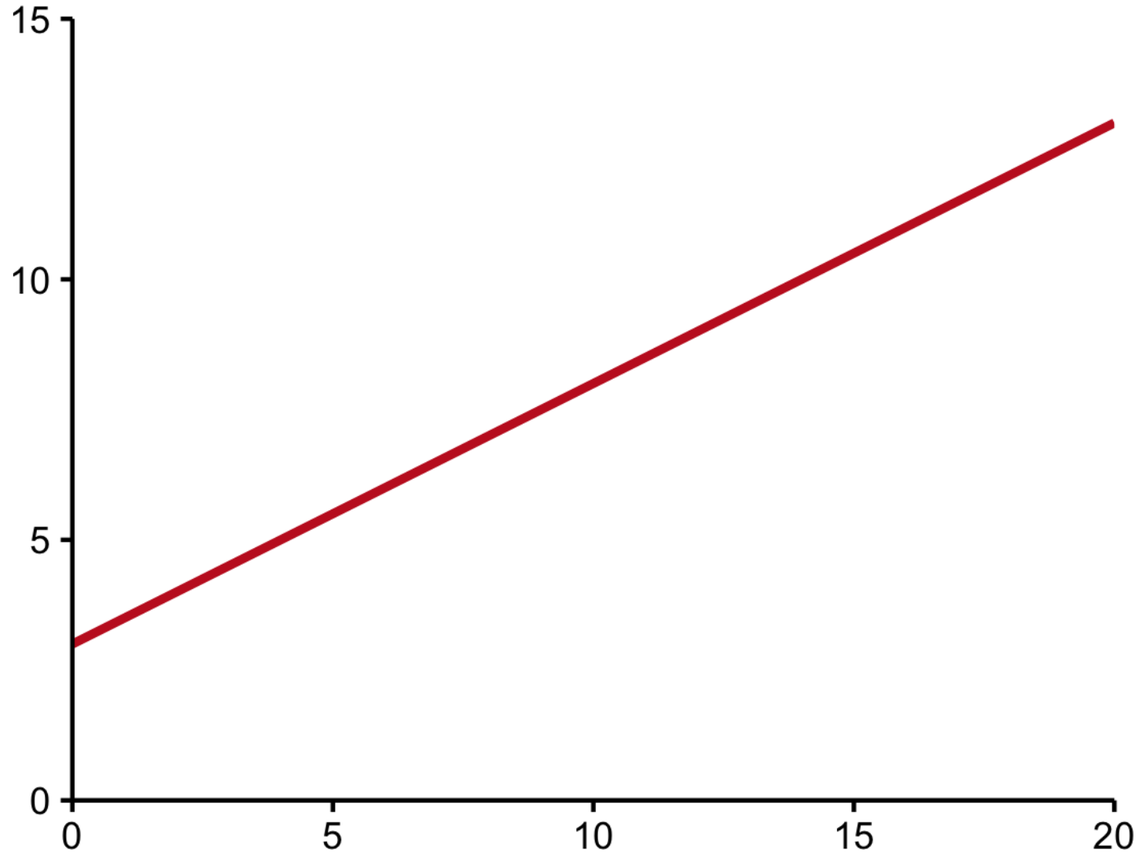
```
ggplot(mammals, aes(x = body, y = brain)) +
  geom_point(alpha = 0.6) +
  coord_fixed() +
  scale_x_log10() +
  scale_y_log10() +
  stat_smooth(
    method = "lm",
    color = "#C42126",
    se = FALSE,
    size = 1
  )
```



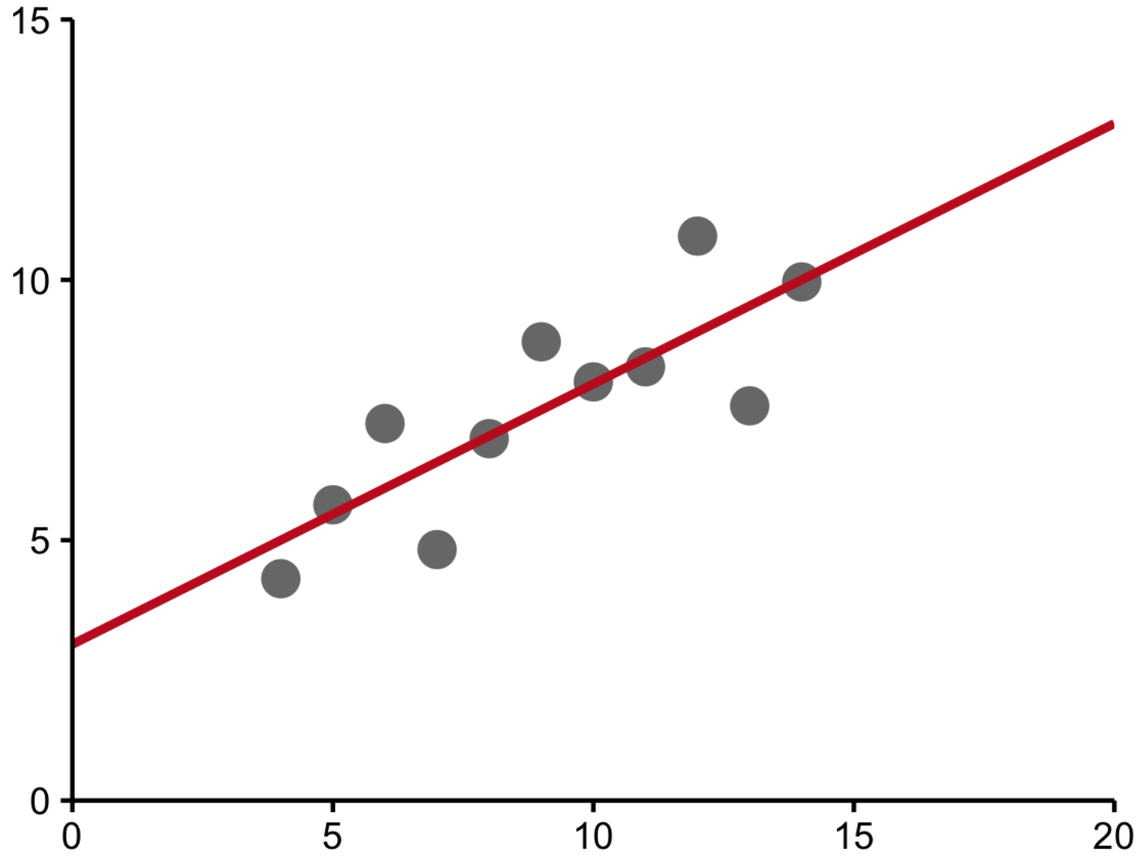
Publication-ready plot



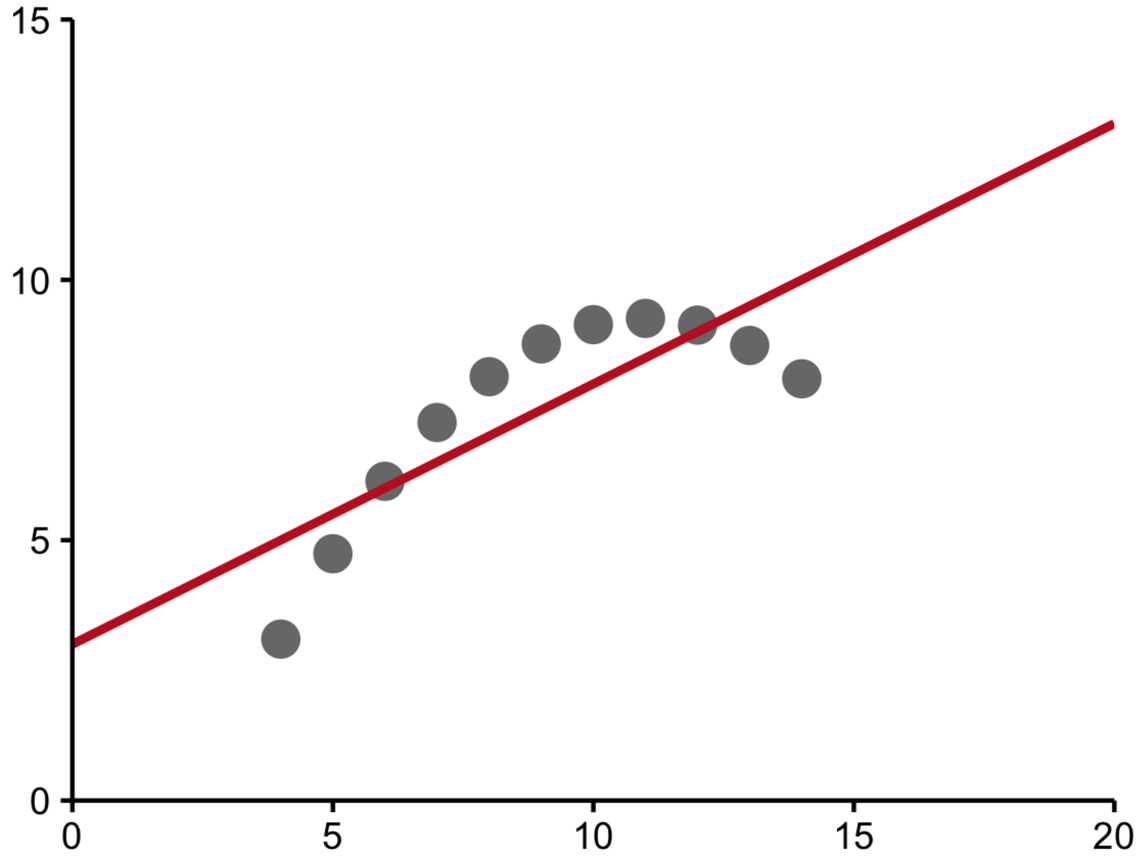
Anscombe's plots



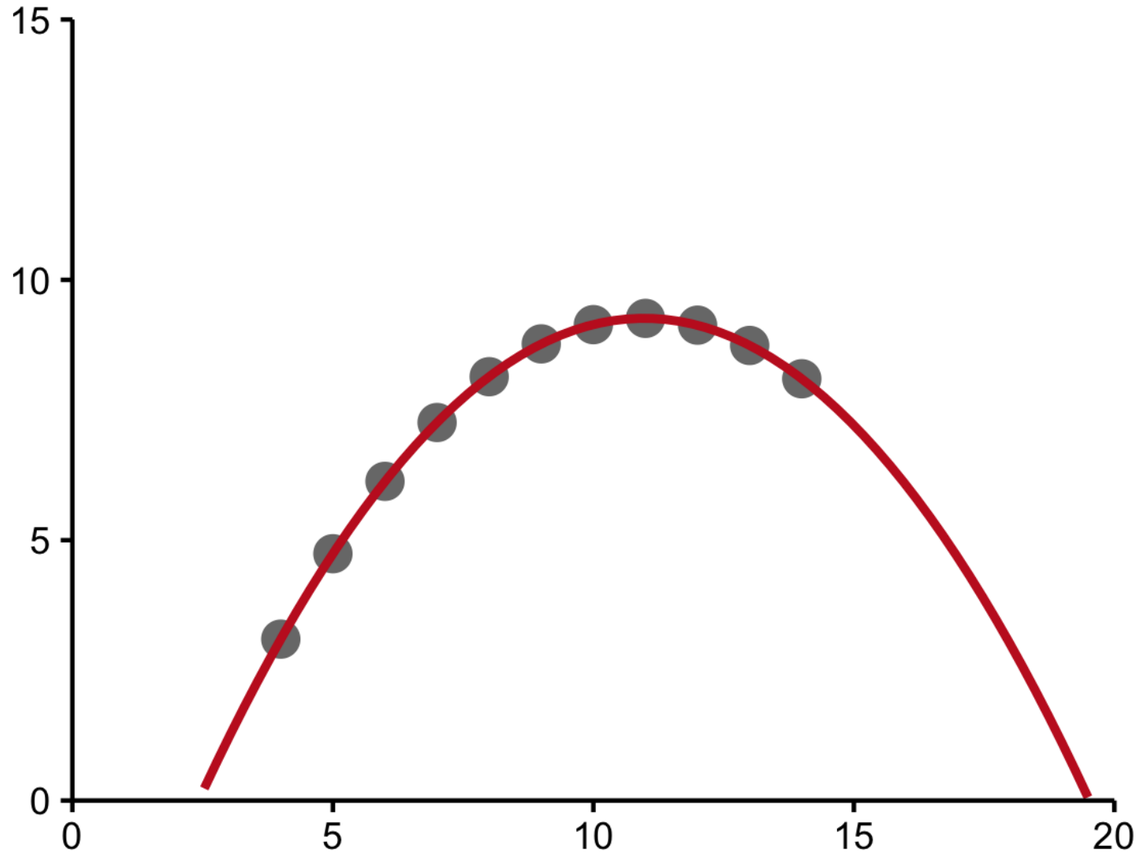
Anscombe's plots



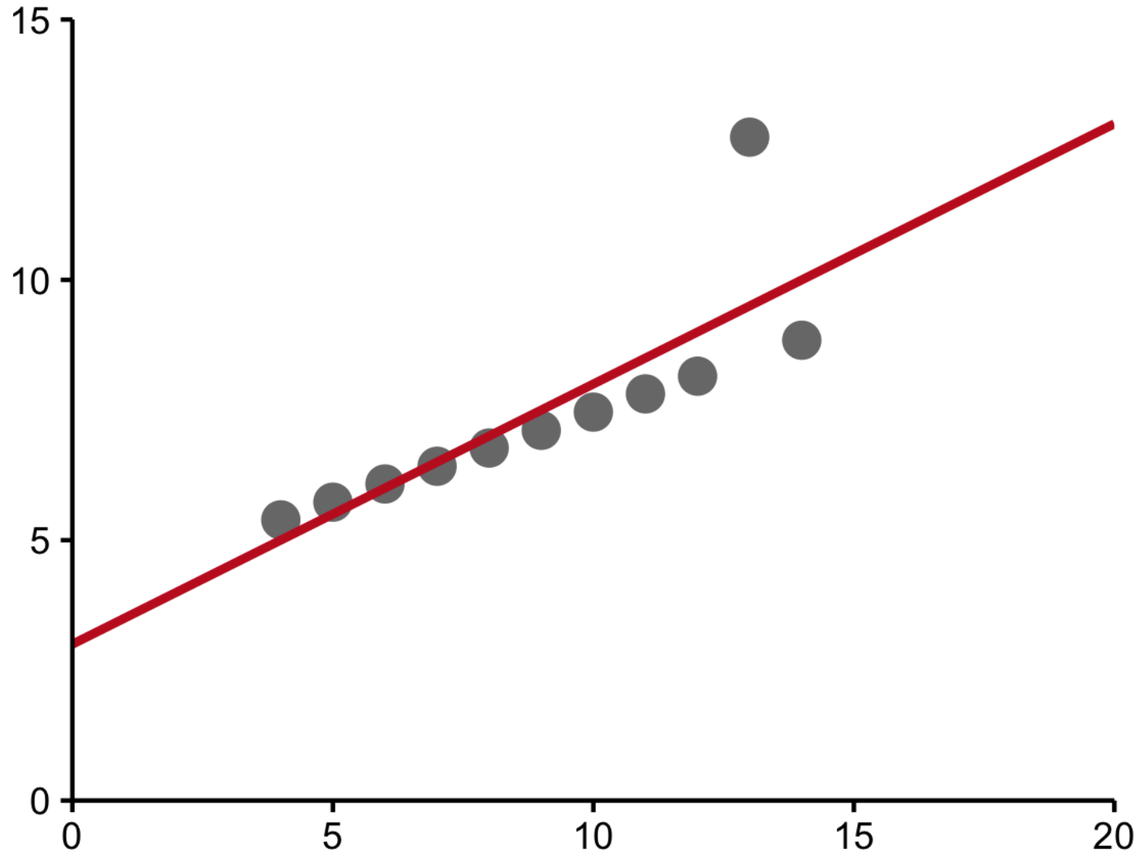
Anscombe's plots



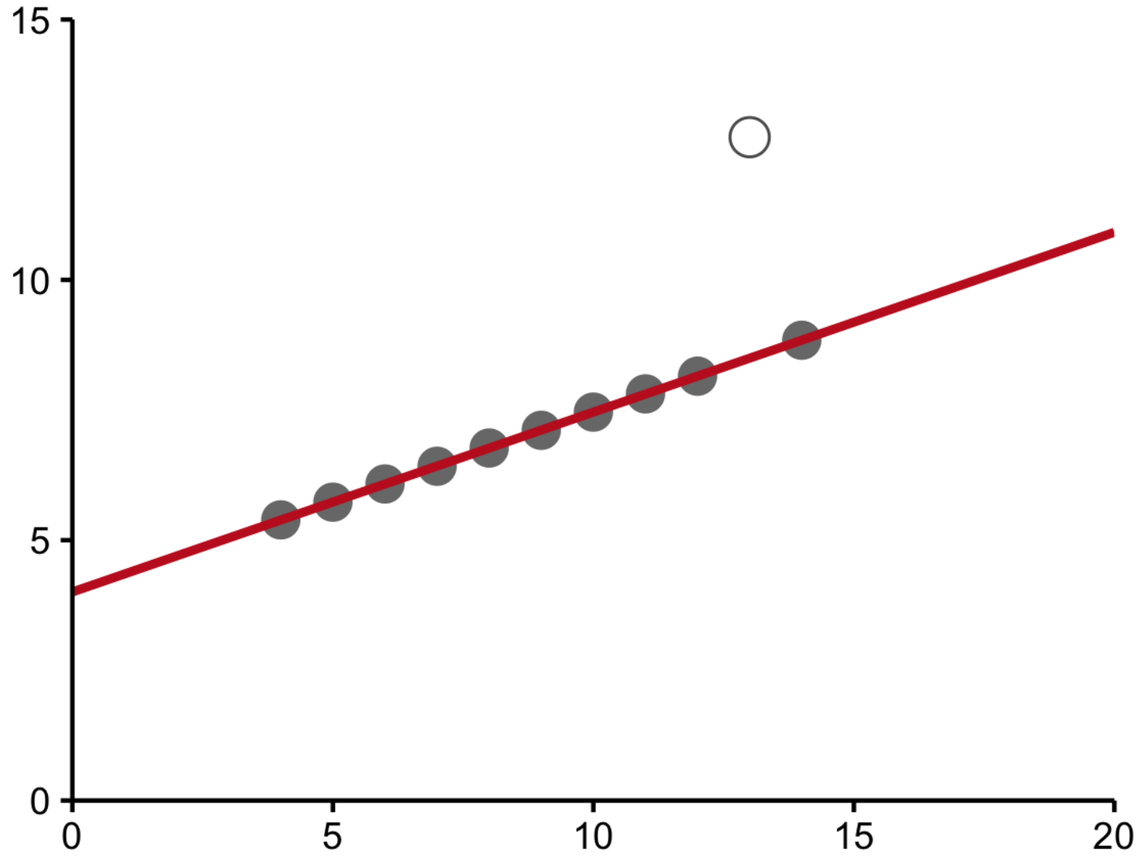
Anscombe's plots



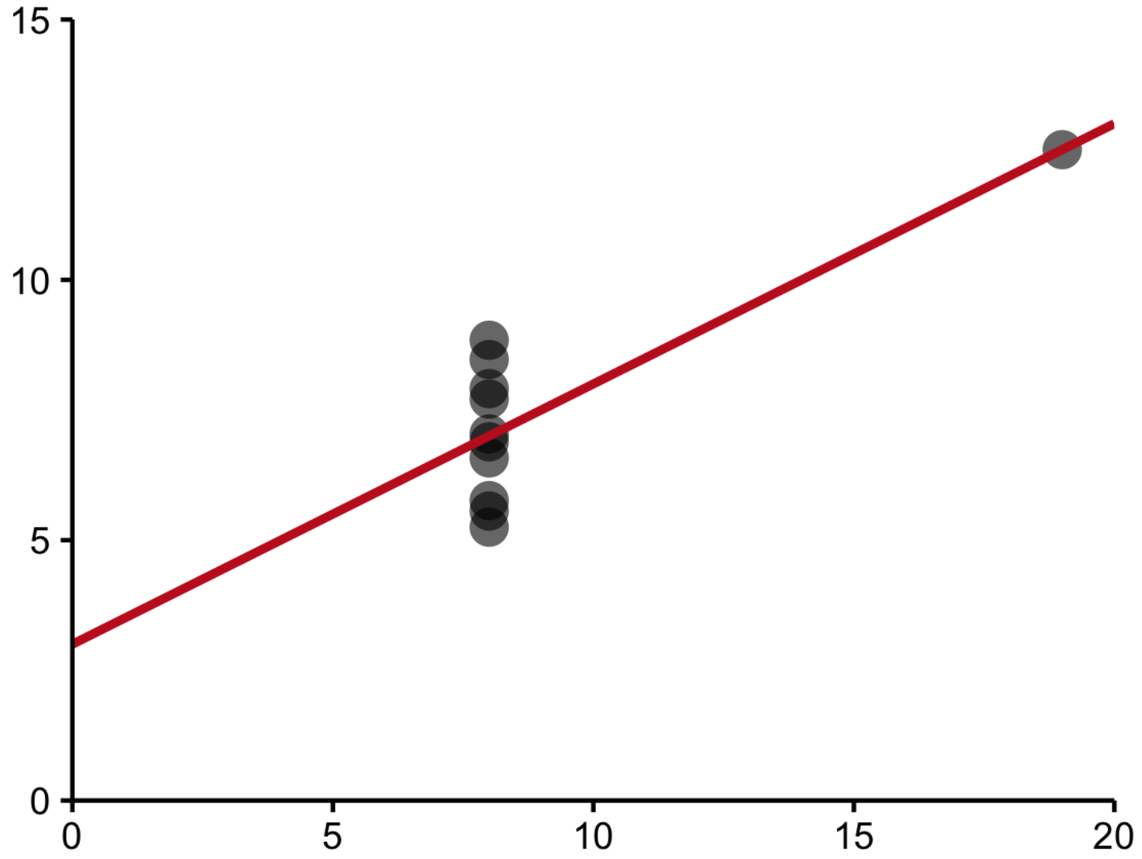
Anscombe's plots



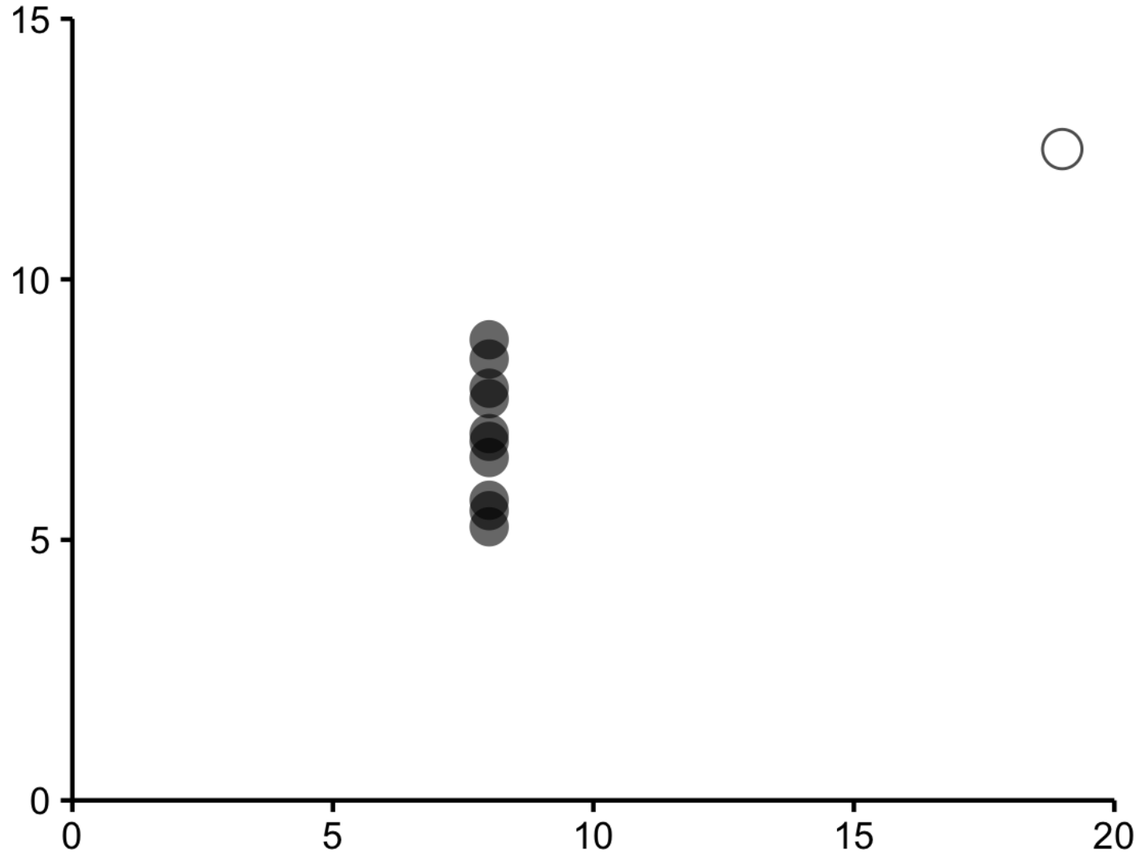
Anscombe's plots



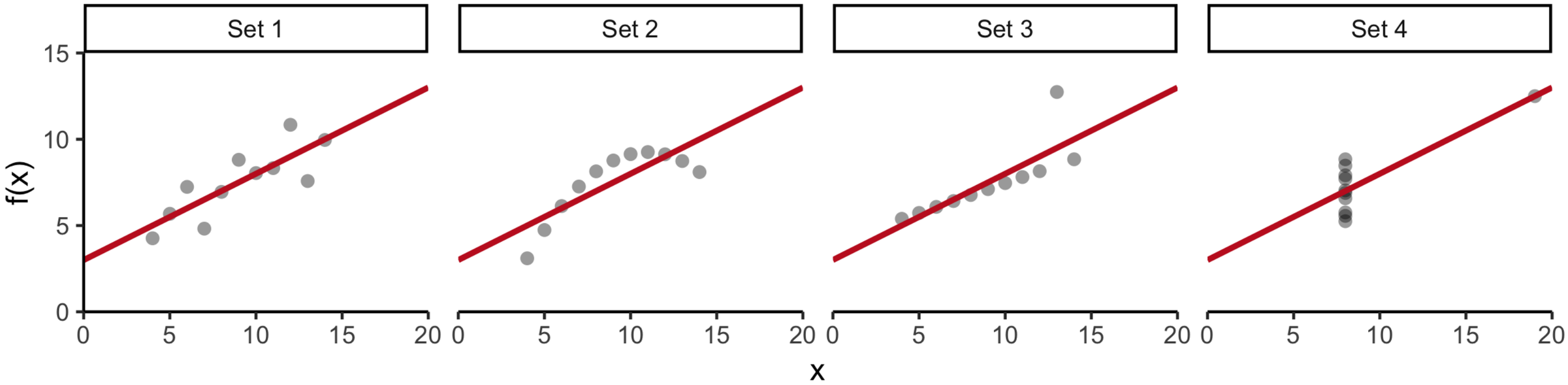
Anscombe's plots



Anscombe's plots



Anscombe's plots



Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2

The grammar of graphics

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

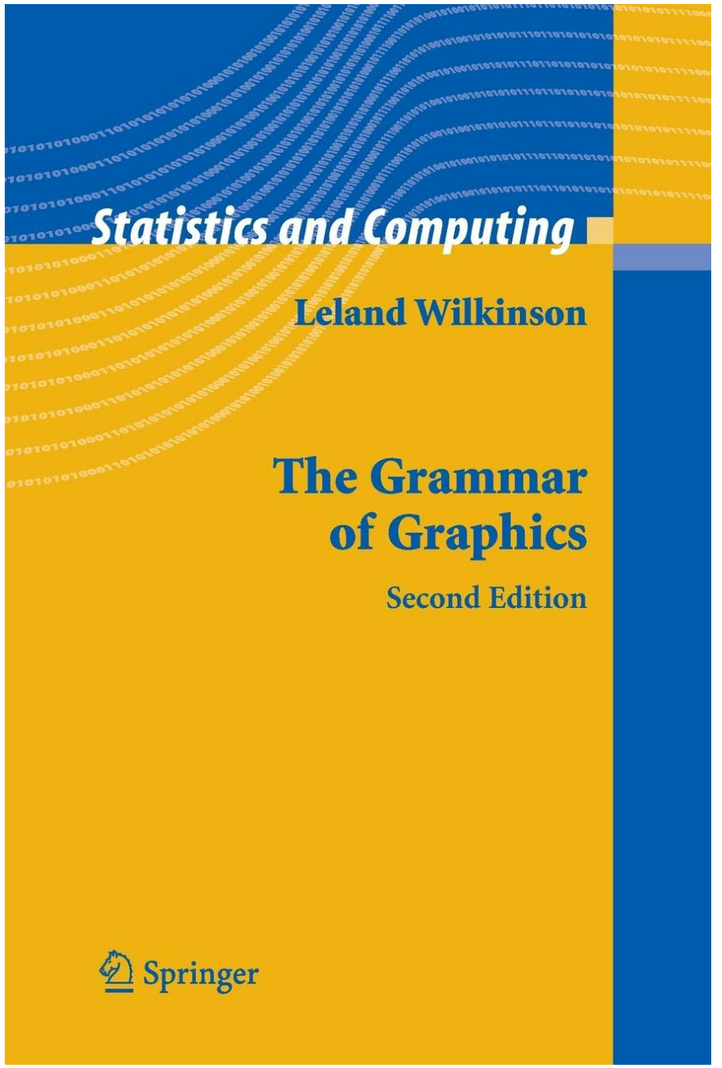
The quick brown fox jumps over the lazy dog

The quick brown fox jumps over the lazy dog

Article	*The*	*A*	*The*
Adjective	*quick brown*	*rabid red*	
Noun	*fox*	*fox*	*Hunter*
Verb	*jumps*	*bit*	*shot*
Preposition	*over*		
Article	*the*	*the*	*the*
Adjective	*lazy*	*friendly*	*rabid red*
Noun	*dog.*	*dog.*	*fox.*

Grammar of graphics

- Plotting framework
- Leland Wilkinson, Grammar of Graphics, 1999
- 2 principles
  - Graphics = distinct layers of grammatical elements
  - Meaningful plots through aesthetic mappings



The three essential grammatical elements

Element	Description
Data	The data-set being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.

Course 1: core competency

Element	Description
Data	The data-set being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.
Themes	All non-data ink.

The seven grammatical elements

Element	Description
Data	The data-set being plotted.
Aesthetics	The scales onto which we *map* our data.
Geometries	The visual elements used for our data.
Themes	All non-data ink.
Statistics	Representations of our data to aid understanding.
Coordinates	The space on which the data will be plotted.
Facets	Plotting small multiples.

Jargon for each element

Data	*{variables of interest}*				
Aesthetics	*x-axis* *y-axis*	*colour* *fill*	*size* *labels*	*alpha* *shape*	*line width* *line type*
Geometries	*point*	*line*	*histogram*	*bar*	*boxplot*
Themes	*non-data ink*				
Statistics	*binning*	*smoothing*	*descriptive*	*inferential*	
Coordinates	*cartesian*	*fixed*	*polar*	*limits*	
Facets	*columns*	*rows*			

Course 2: Tools for EDA

- Remaining 3 layers
- Best practices for Data Viz

Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2

ggplot2 layers

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2

Rick Scavetta
Founder, Scavetta Academy

datacamp

ggplot2 package

- The grammar of graphics implemented in R
- Two key concepts:
  1. Layer grammatical elements
  2. Aesthetic mappings

Data

Iris dataset

Setosa

Versicolor

Virginica

[1] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179–188. [2] Anderson, Edgar (1935). The irises of the Gaspe Peninsula, Bulletin of the American Iris Society, 59, 2–5.

Iris dataset

```
iris
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
...
50           5.0         3.3          1.4         0.2     setosa
51           7.0         3.2          4.7         1.4 versicolor
52           6.4         3.2          4.5         1.5 versicolor
53           6.9         3.1          4.9         1.5 versicolor
...
100          5.7         2.8          4.1         1.3 versicolor
101          6.3         3.3          6.0         2.5  virginica
102          5.8         2.7          5.1         1.9  virginica
103          7.1         3.0          5.9         2.1  virginica
...
150          5.9         3.0          5.1         1.8  virginica
```

Aesthetics

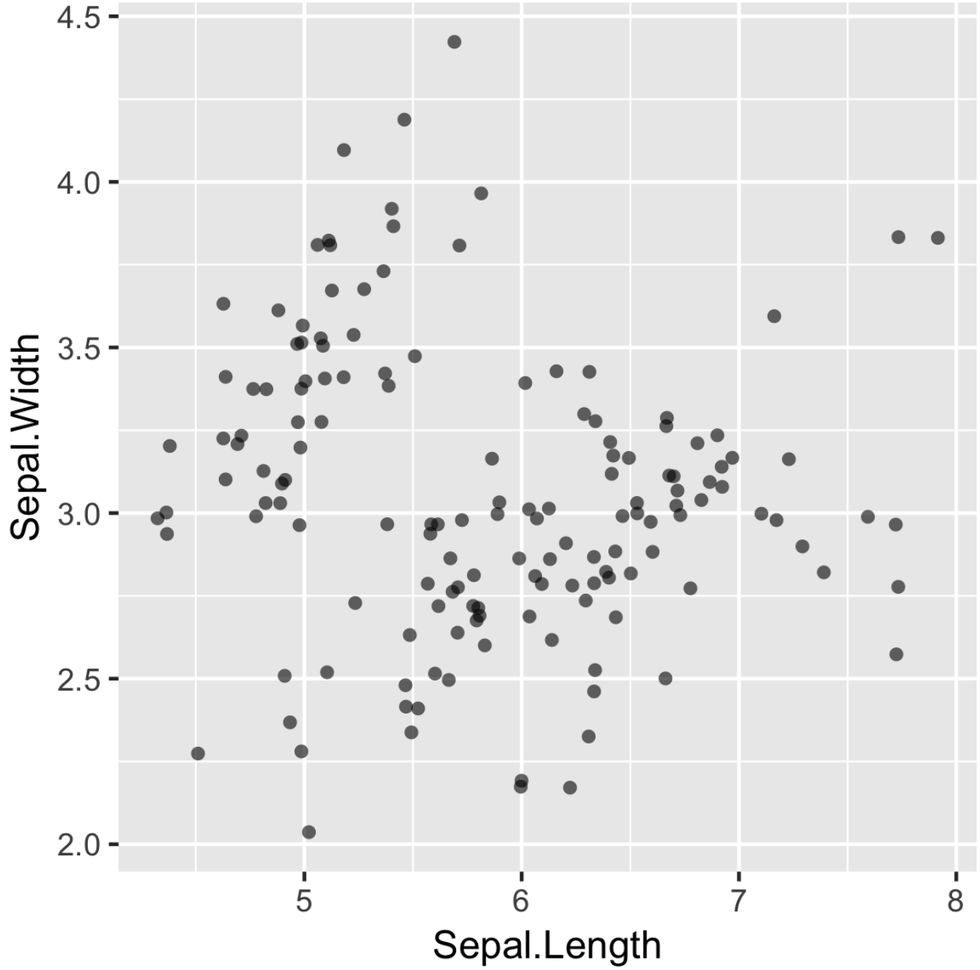
Iris aesthetics

Species	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
	X	Y		

Geometries

Iris geometries

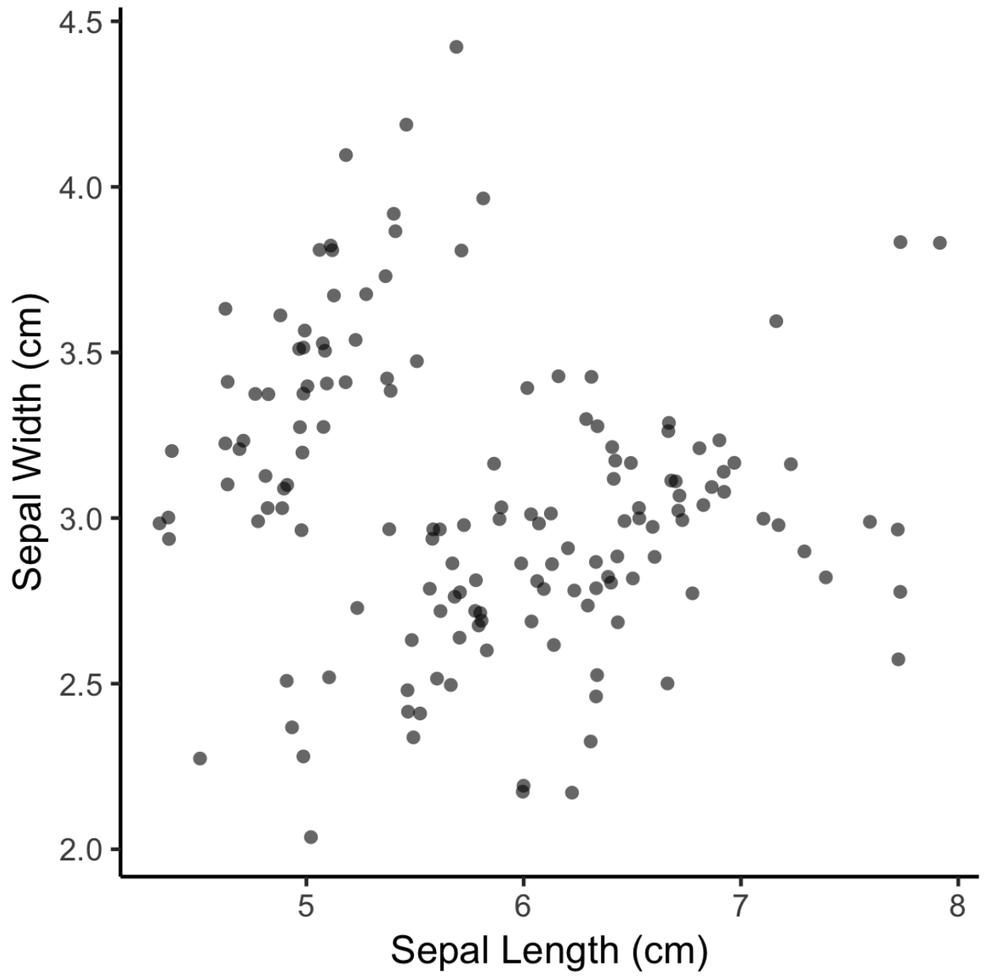
```
g <- ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width)) +
  geom_jitter()
g
```



Themes

Iris themes

```
g <- g +
  labs(x = "Sepal Length (cm)", y = "Sepal Width (cm)") +
  theme_classic()
g
```



Let's practice!

INTRODUCTION TO DATA VISUALIZATION WITH GGPLOT2