

Allgemeine Übersicht

Statistik in a nutshell

Was ist das Ziel hier?

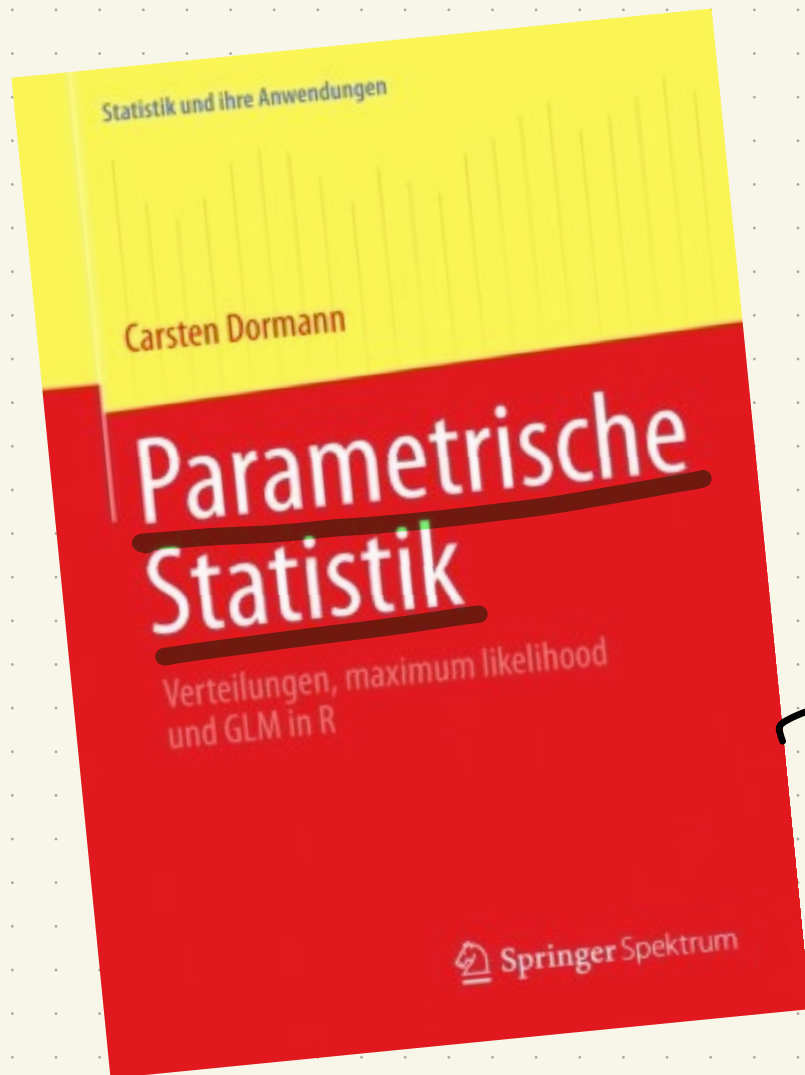
Wir wollen modellieren.

Wir wollen Daten D
modellieren

Wir wollen Parameter
von Verteilungen schätzen.

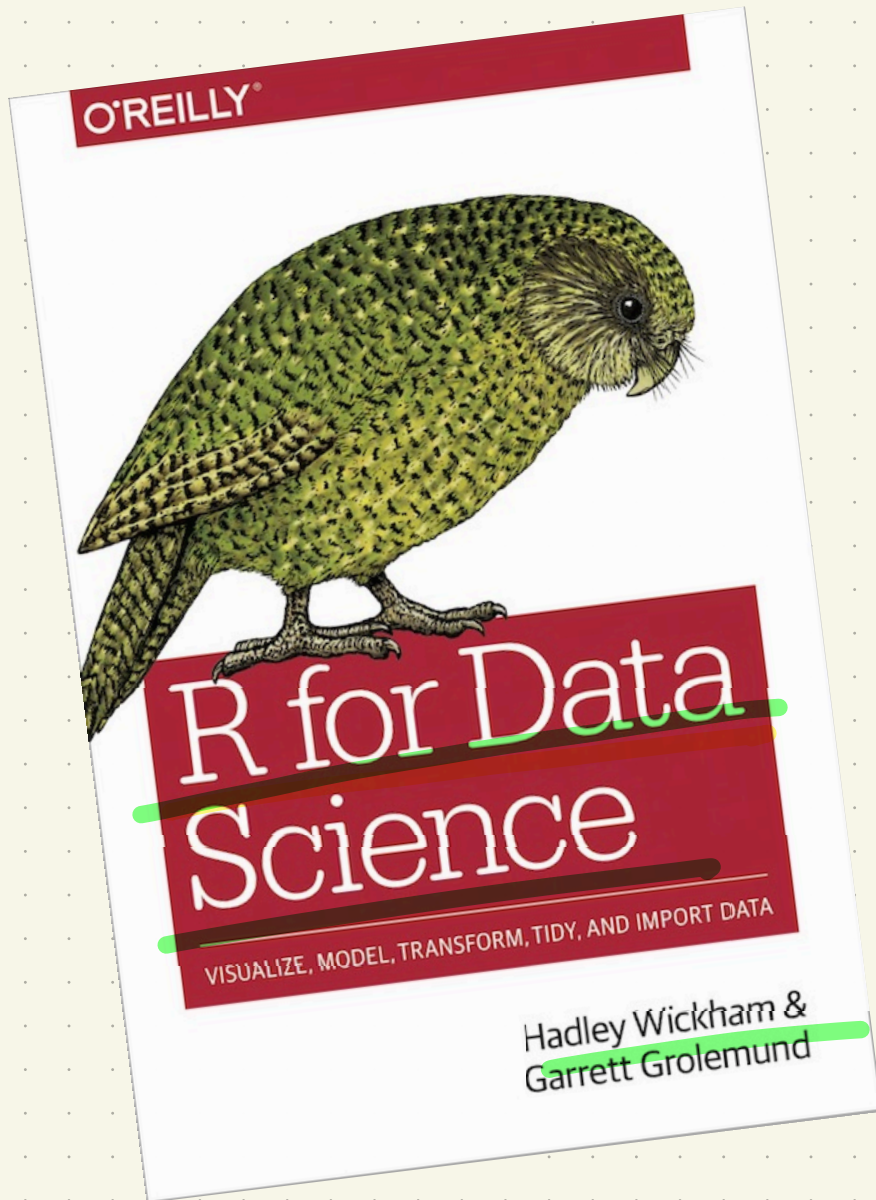
$$D \rightarrow Y \sim X$$

Eine
Spalte



Es geht um die Anwendung!

Unsere Grammatik
ist Tidyverse



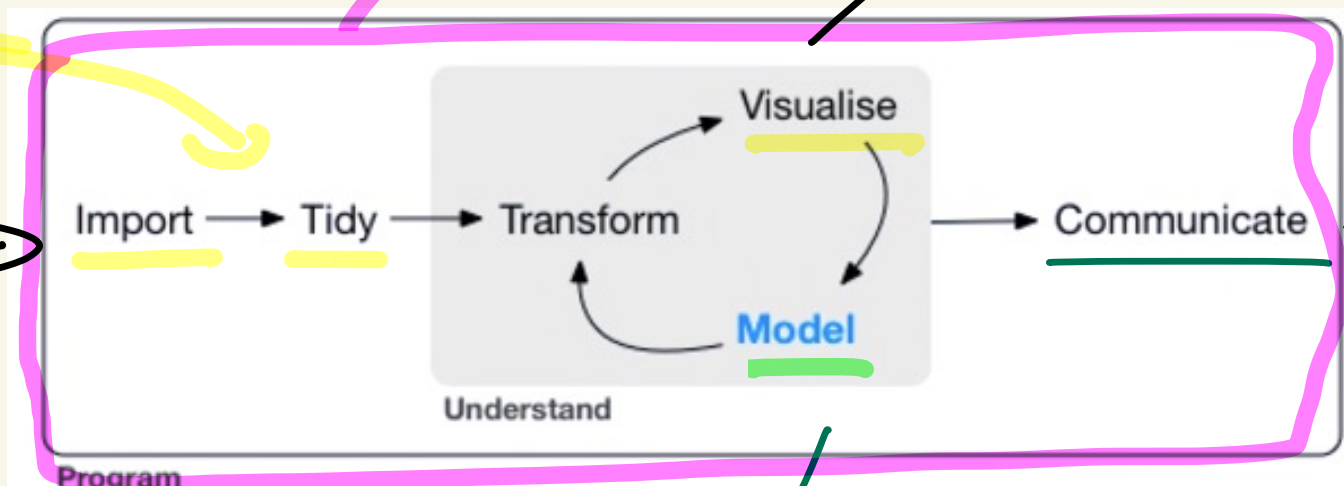


Daten I

Excel

R

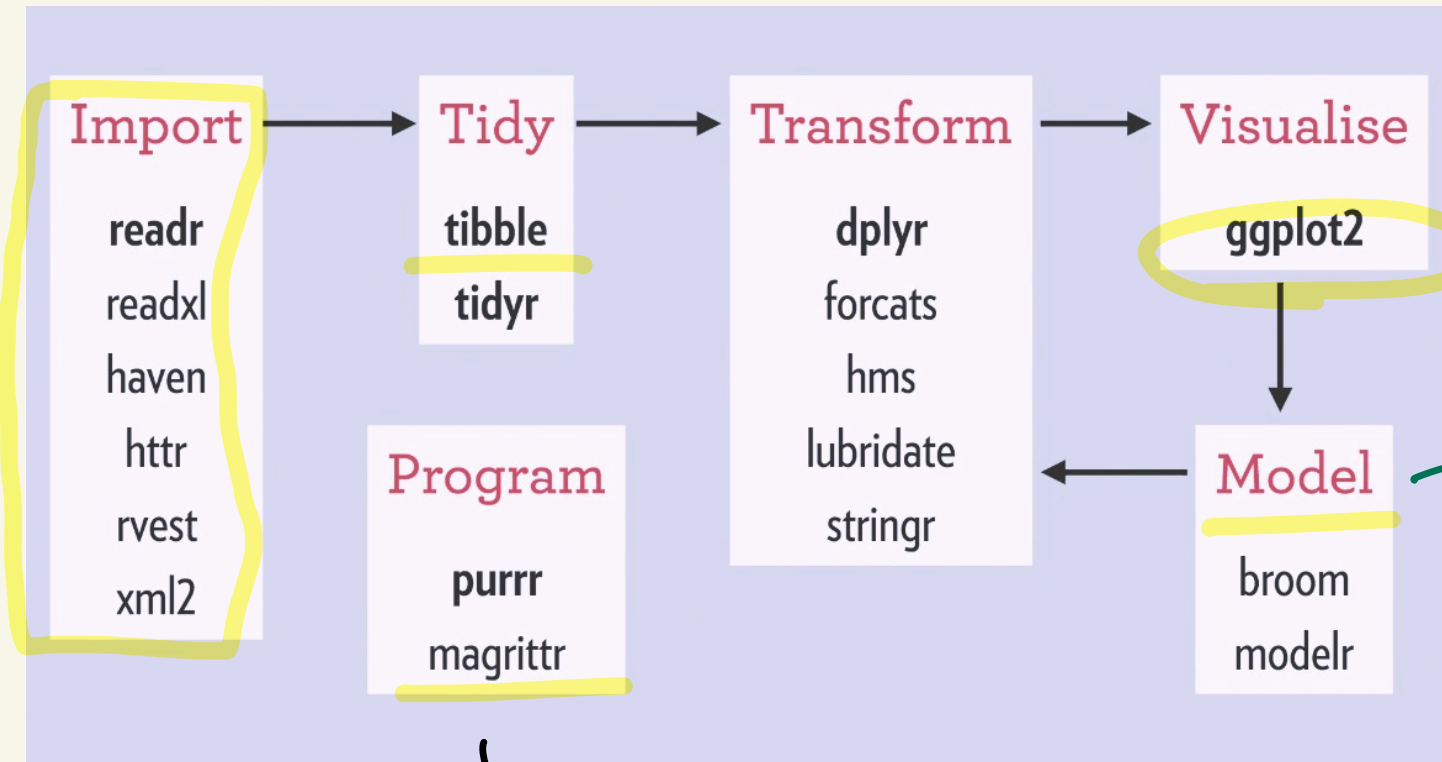
Explorative Daten-
analyse



Word

$Y \sim X$

tibble lagern wir Daten in R



→ Pipes %>%

"In einem
Modell kann
es nur
ein Y geben"

←
eine Spalte

Outcome
Response
Endpoint

1 Spalte ~ anderen Spalten
 ≥ 1



Kausale

Prädiktive

Machine Learning
KI

Variable V
Riskfactor
Covariate
X

Using the *performance* R package



Model 1

`glm(response ~ V1*V2)`

$Y \sim X_1 + X_2$

`compare_performance()`

`check_model()`

- ✓ Linearity
- ✓ Homogeneity
- ✓ Heteroscedasticity
- ✓ Collinearity
- ✓ Outliers
- ✓ ...

`model_performance()`

- ✓ AIC
- ✓ BIC
- ✓ ICC
- ✓ R2
- ✓ RMSE
- ✓ ...

Model	AIC	BIC	R2
Model 1	301	318	0.14
Model 2	284	297	0.12

`test_performance()`
Model 1 > Model 2 ($p < .042$)

Model 2

`glm(response ~ V1)`

$Y \sim X_1$

`check_model()`

- ✓ Linearity
- ✓ Homogeneity
- ✓ Heteroscedasticity
- ✓ Collinearity
- ✓ Outliers
- ✓ ...

`model_performance()`

- ✓ AIC
- ✓ BIC
- ✓ ICC
- ✓ R2
- ✓ RMSE
- ✓ ...

Grundgesamtheit



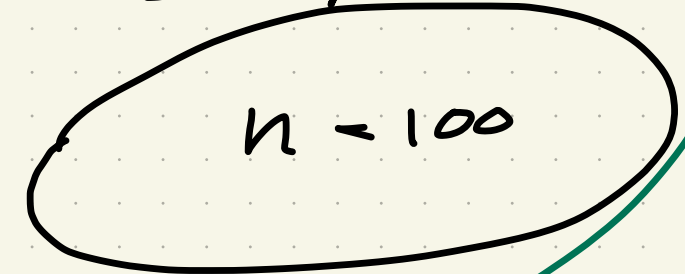
Randomisieren

Y? X?

1			
2			
3			
100			

Stichprobe

D



Rückschluß

in der Bewertung
des Anwenders...

Modell

$$Y \sim X$$