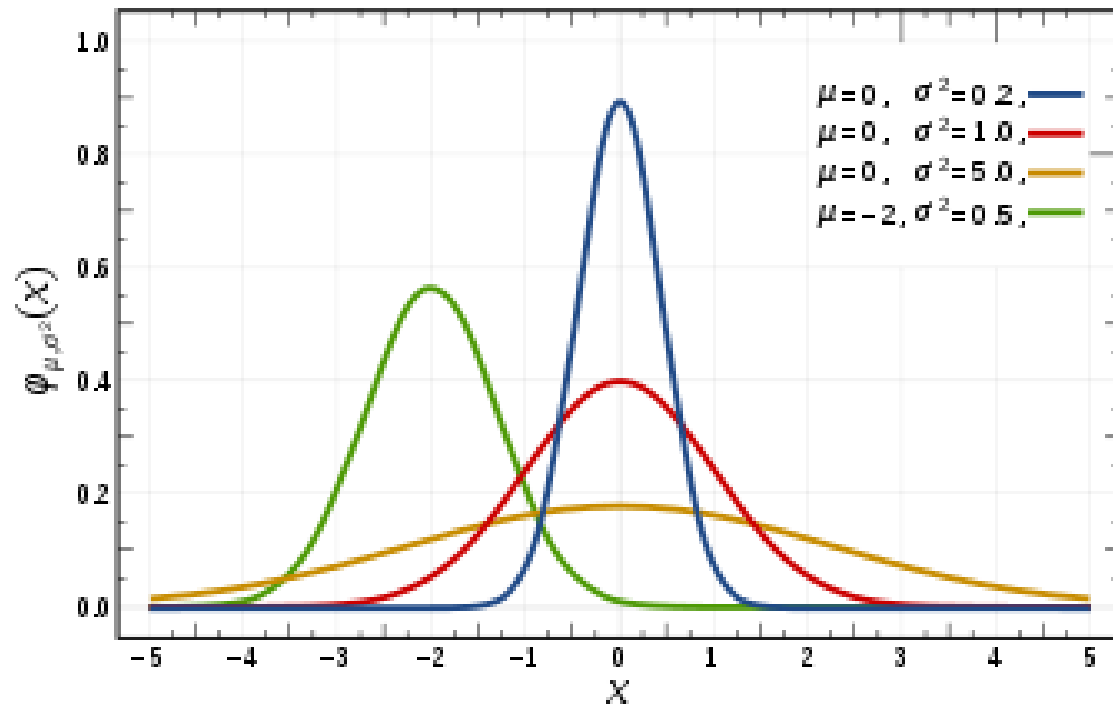


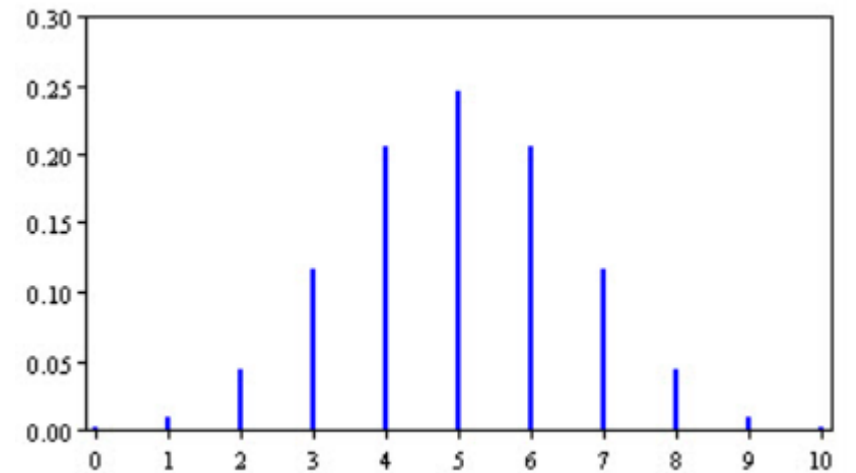
Data analyse og præprocessering

Sandsynlighedsregning, statistik og data manipulation

Sandsynlighedsfordelinger

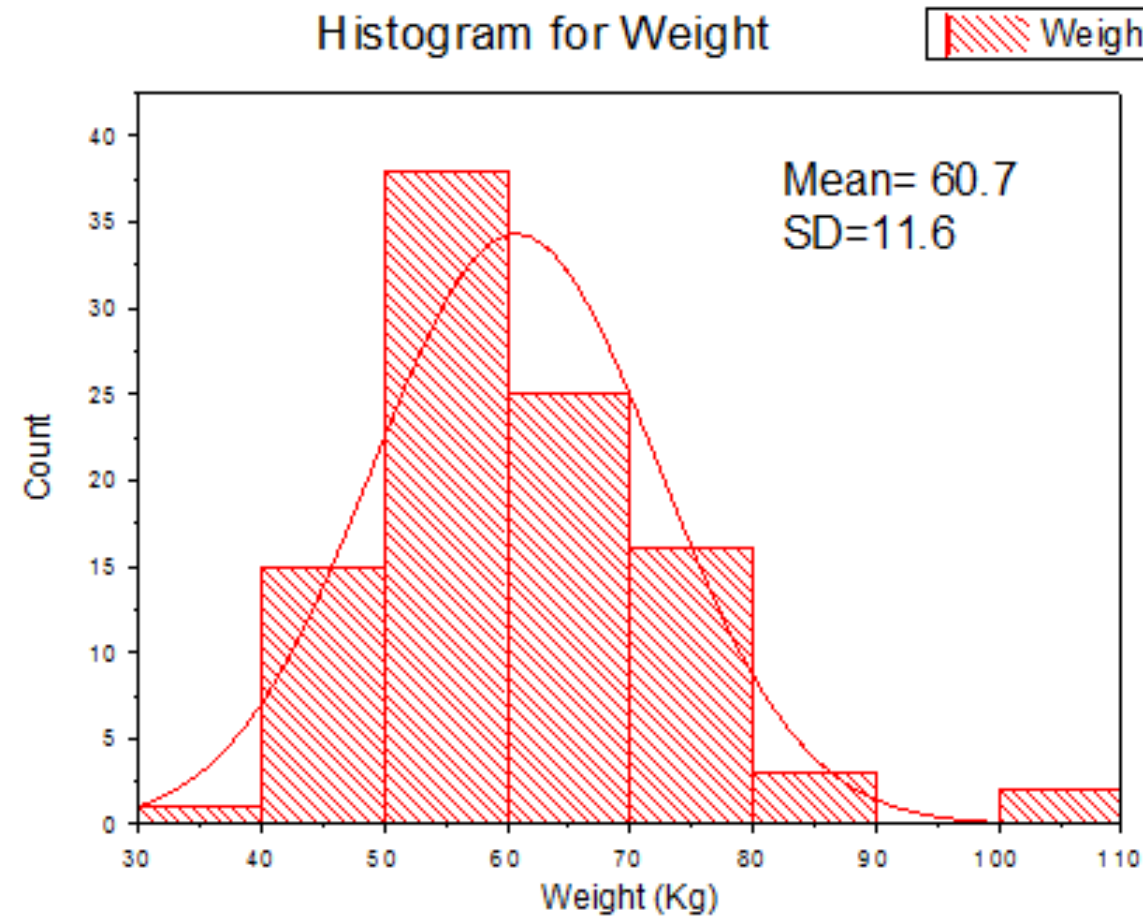


Kontinuert variabel

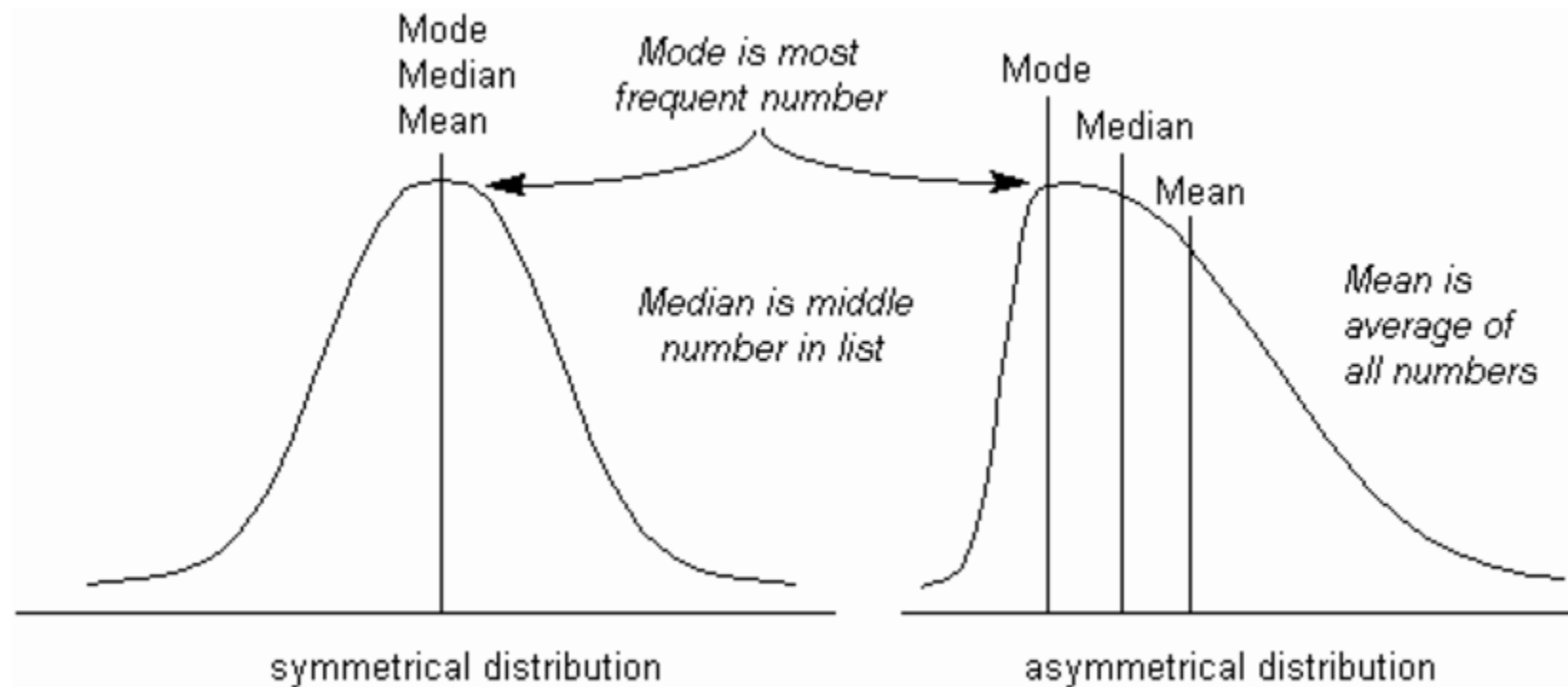


Diskret variabel

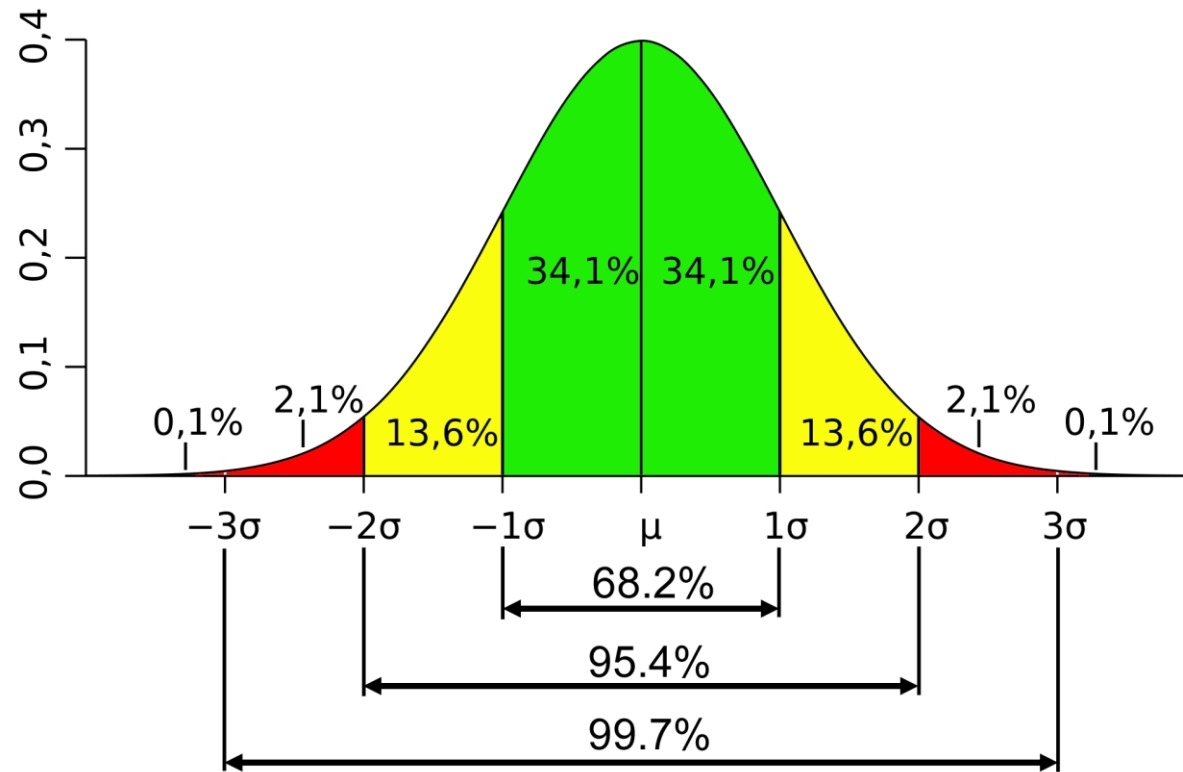
Kontinuerte data – histogram eller funktions- approximation



Mean, median og mode



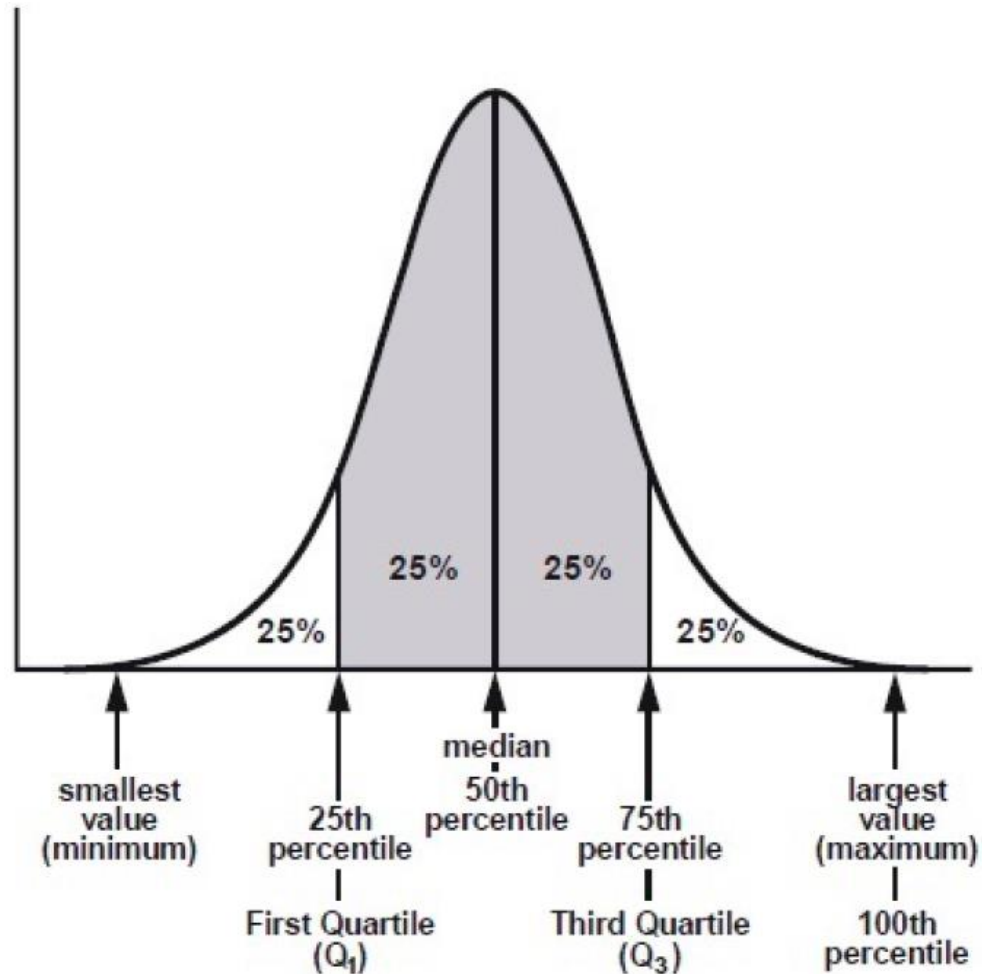
Standard afvigelse og varians



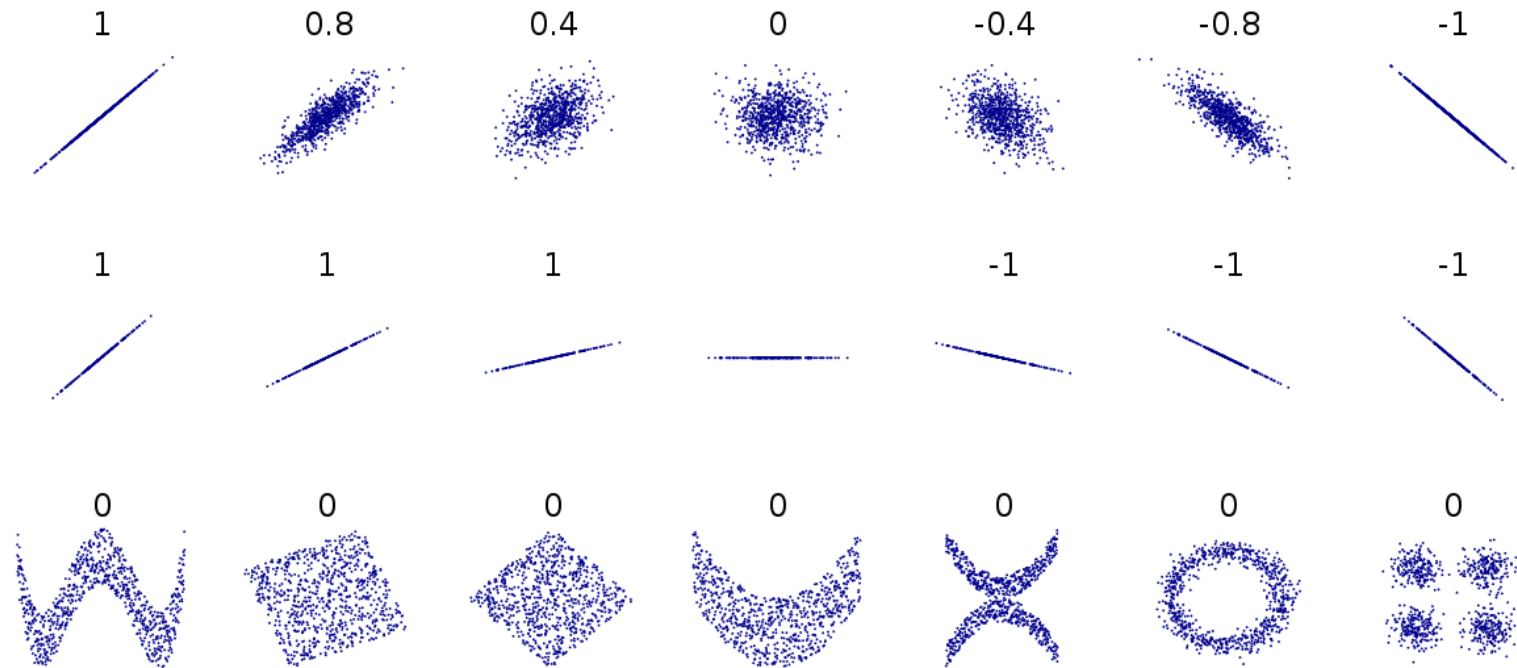
$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$

KUN FOR NORMALFORDELING (= GAUSSISK FORDELING) !

Percentiler

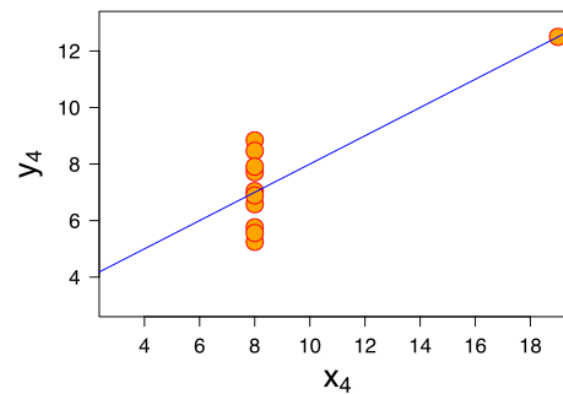
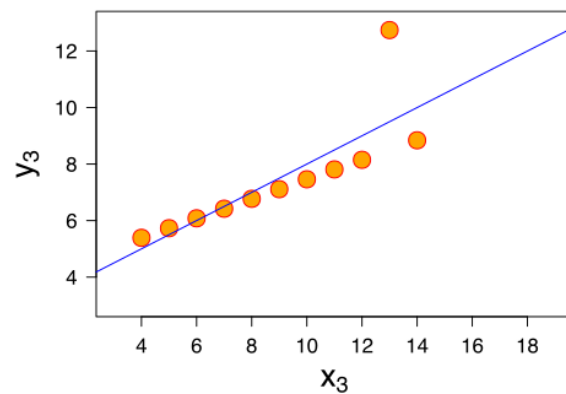
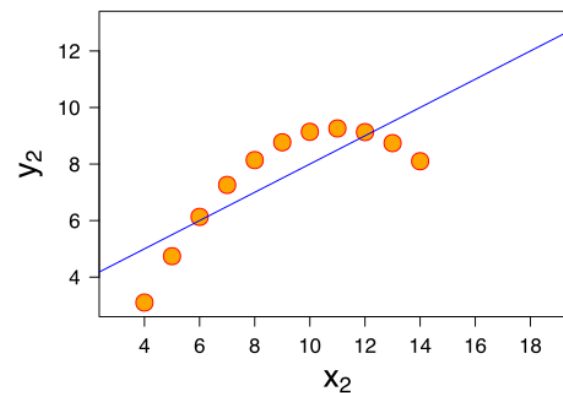
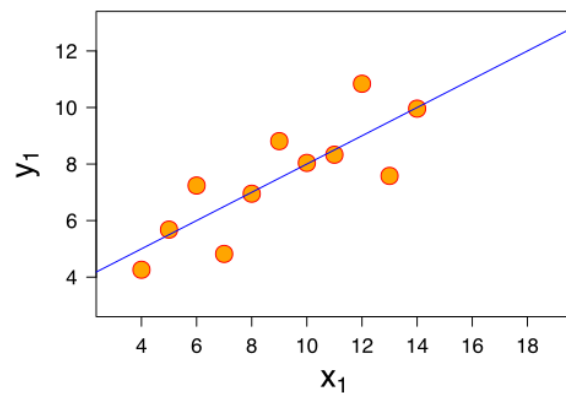


Analyse af sammenhæng mellem to variable - Korrelationskoefficient (Pearsons)

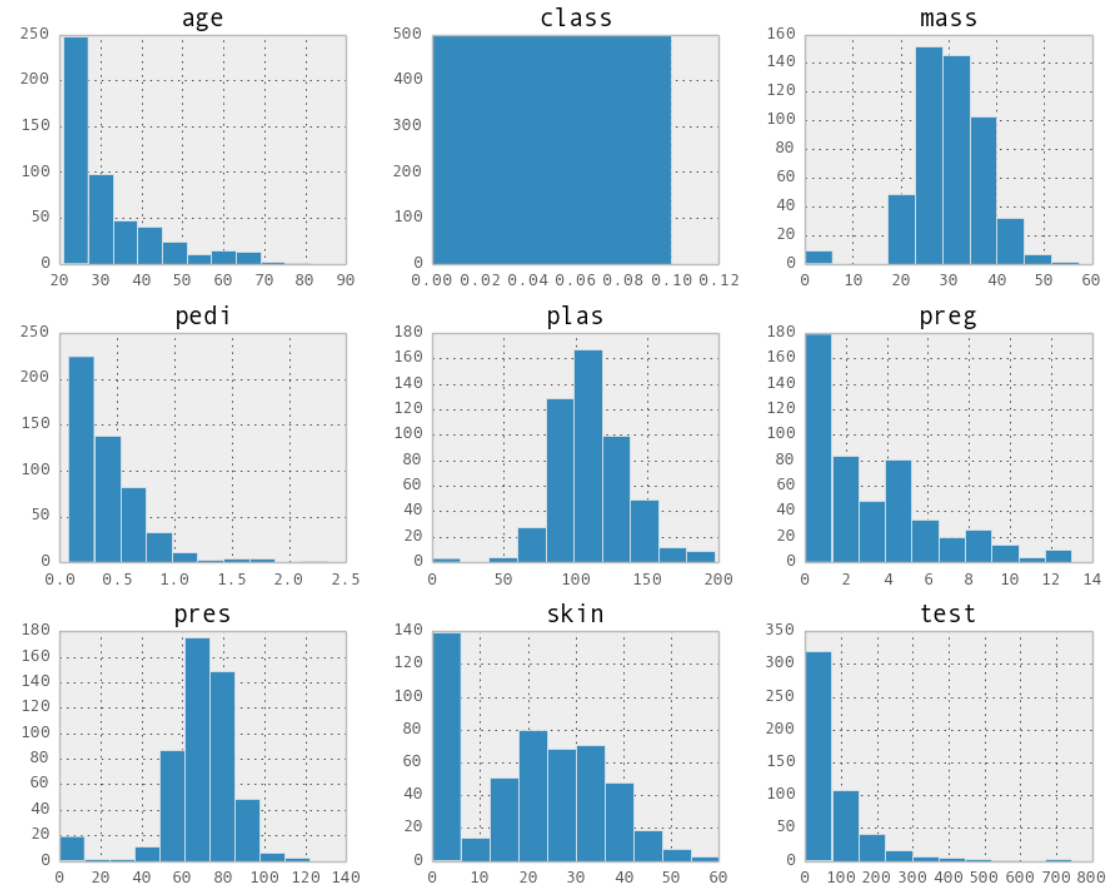


$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

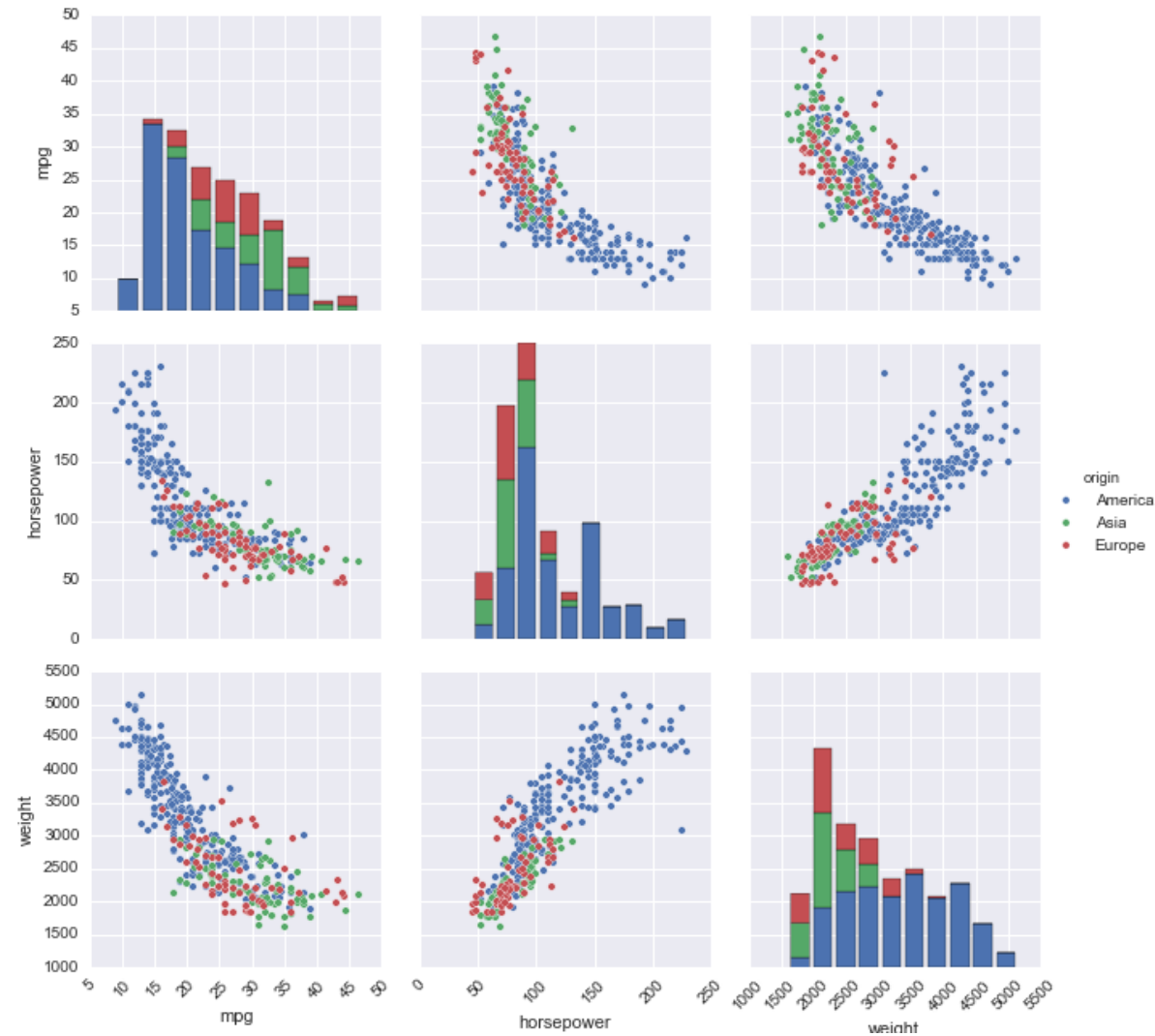
Anscombe's Quartet



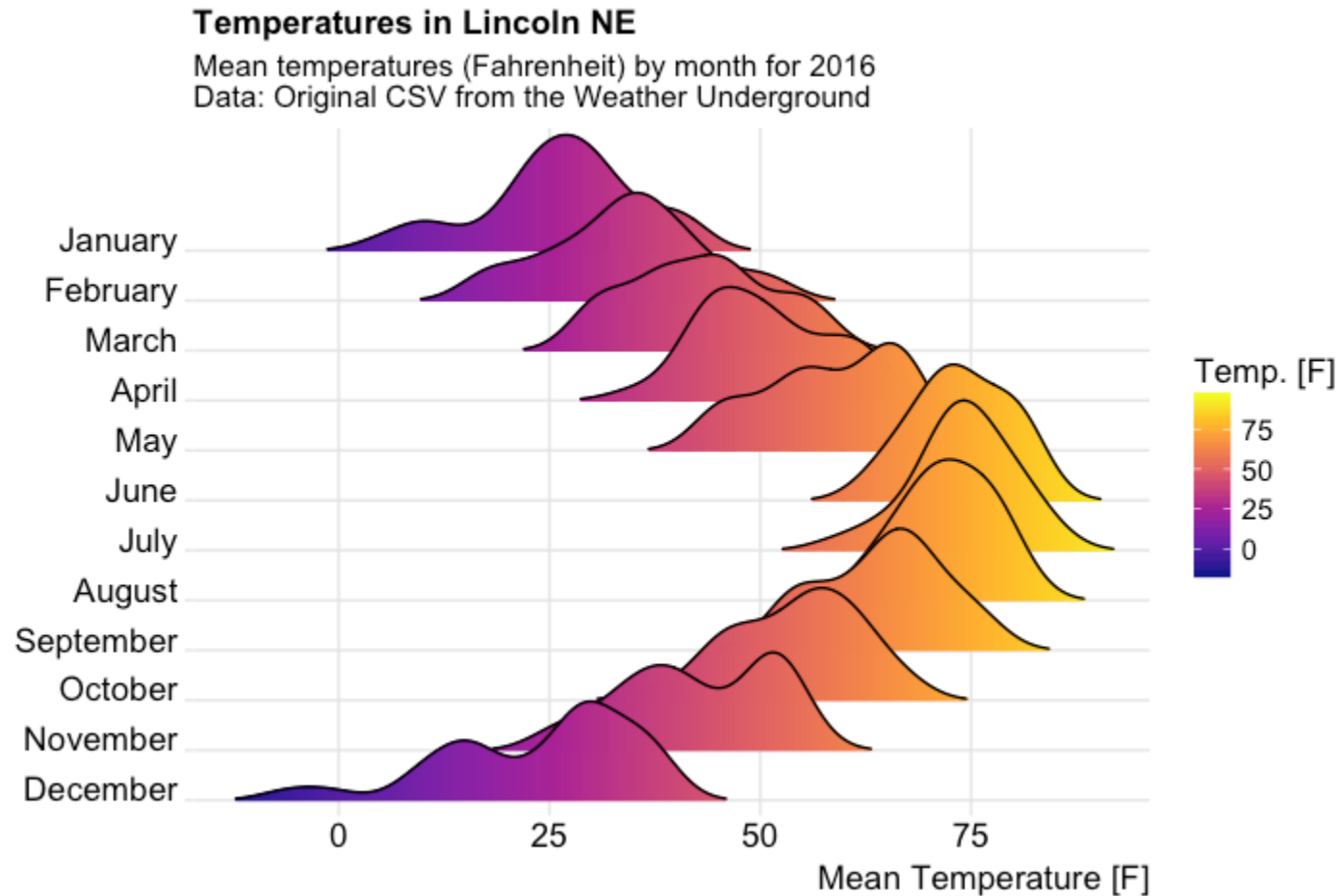
Eksempel : Analyse med histogram



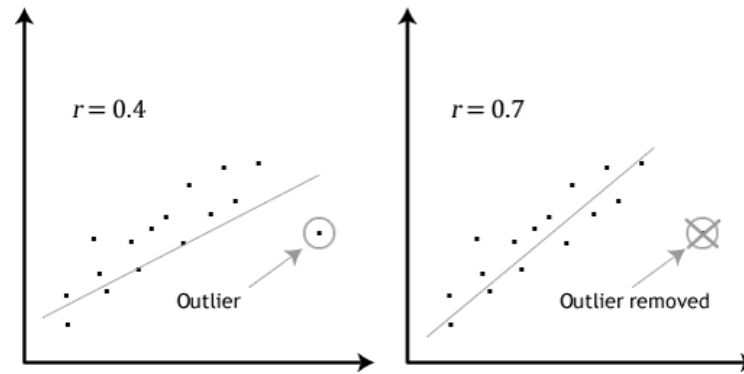
Eksempel : Analyse af klassifikationsproblem med scatterplot



Seaborn (kun hvis du har oceaner af tid 😊)

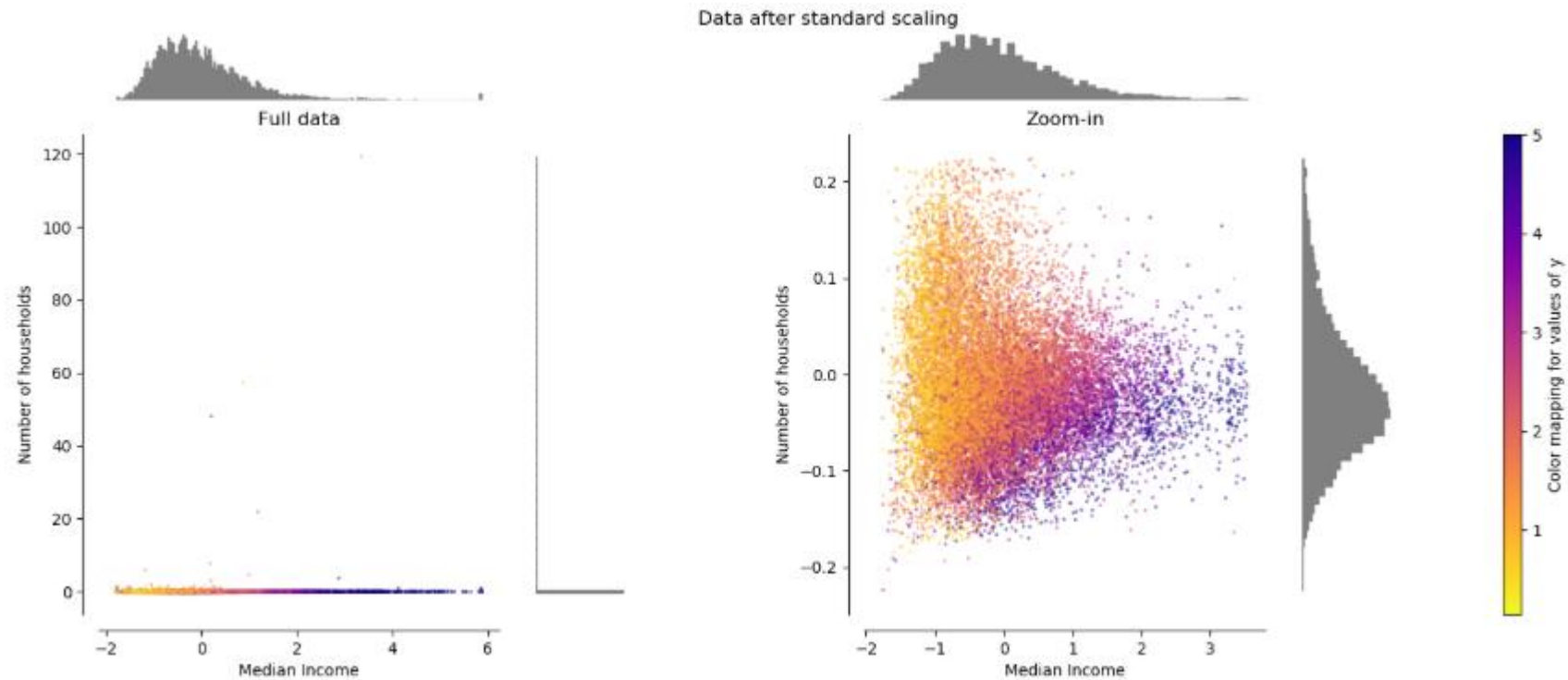


Data cleaning – missing values and outliers



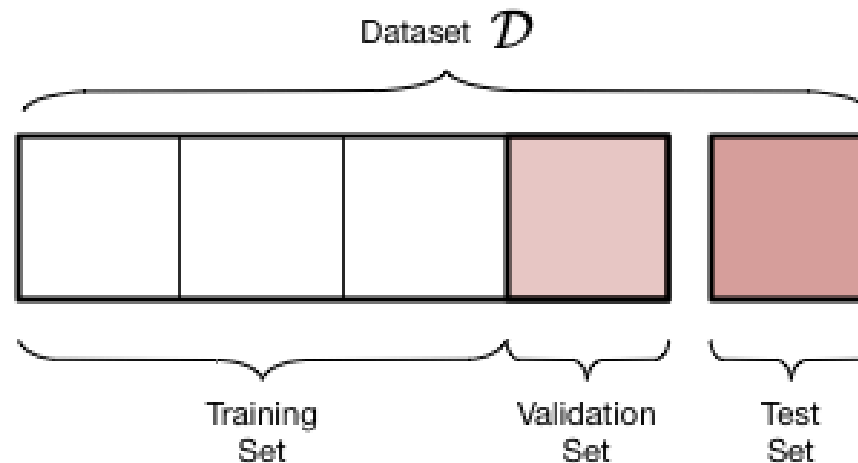
Respondent	Variables			
	A	B	C	D
1	1	2	3	4
2	1	2	3	4
3	4	3	2	1
4	4	3	2	1
5	1	2		1
6		2	2	1
7	1	2	2	
8	1		2	1

Standardisation og normalisation



Train-, test-, og validation-sets

- Fokus : Mindst mulig generalisationsfejl
- Vigtig pointe – algoritmen lærer kun ud fra de data som trænes med !
- Pas på afhængigheder mellem datasæt



Cross-validation split

