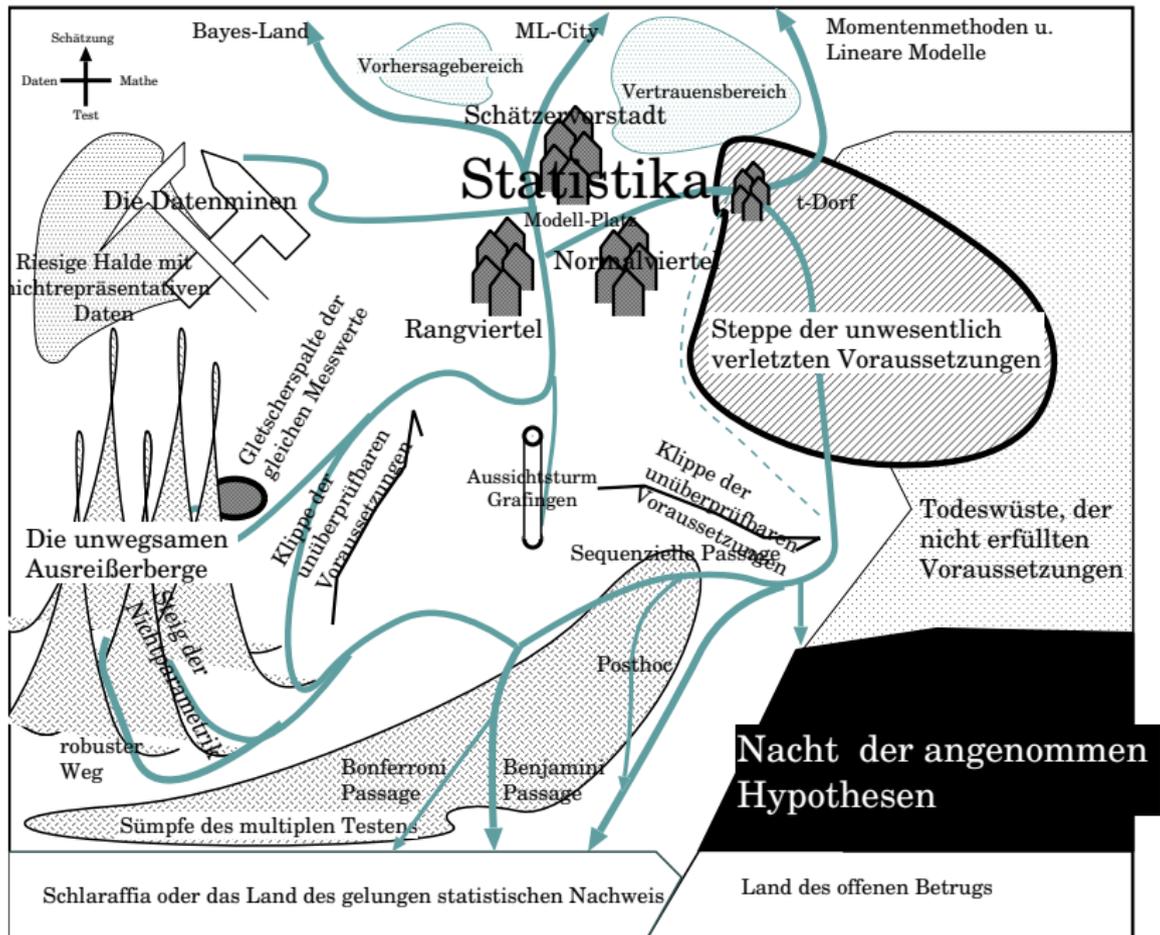


# Datenanalyse und Statistik

## Vorlesung 3 (Graphik für diskrete Skalen und mehr als eine Variable)

K.Gerald van den Boogaart  
<http://www.stat.boogaart.de>

21. Oktober 2019



# Einteilung der Graphiken und Parameter

		Erste Variable	
		diskret	stetig
zweite Variable	keine		x
	diskret		
	stetig		

- ▶ \*stetige Daten
- ▶ diskrete Daten
- ▶ stetig–stetig
- ▶ diskret–diskret
- ▶ diskret–stetig

# Einteilung der Graphiken und Parameter

		Erste Variable	
		diskret	stetig
zweite Variable	keine	X	x
	diskret		
	stetig		

- ▶ \*stetige Daten
- ▶ diskrete Daten als nächstes
- ▶ stetig–stetig
- ▶ diskret–diskret
- ▶ diskret–stetig

# Graphiken und Kenngrößen für diskrete Skalen

- ▶ Kenngrößen
- ▶ Balkendiagramme (praktisch und funktional)
- ▶ Kuchendiagramme (Vorsicht Zucker)
- ▶ Tortendiagramm (Vorsicht Cholesterolin und Zucker)

# Datensatz

```
> margin <- function(x,...) apply(x, pmatch(c(...), names(dimnames(x))), sum)
> data(Titanic)
> ftable(Titanic, col.vars=c("Class", "Survived"))
```

		Class		1st		2nd		3rd		Crew	
		Survived		No	Yes	No	Yes	No	Yes	No	Yes
Sex	Age										
	Male	Child		0	5	0	11	35	13	0	0
	Adult	118	57	154	14	387	75	670	192		
Female	Child	0	1	0	13	17	14	0	0		
	Adult	4	140	13	80	89	76	3	20		

# Kenngrossen

## Anteile:

```
> margin(Titanic, "Survived")/sum(Titanic)
```

No	Yes
0.676965	0.323035

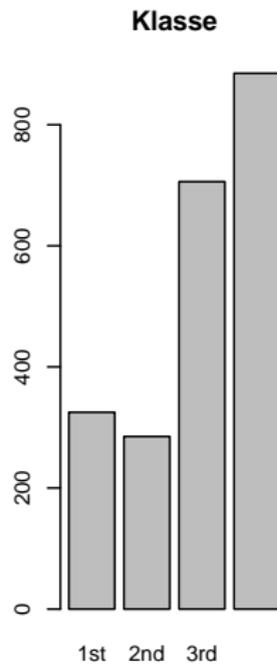
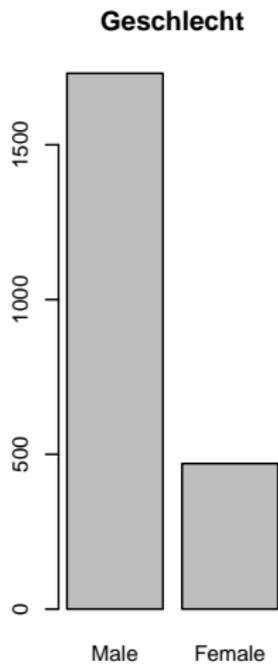
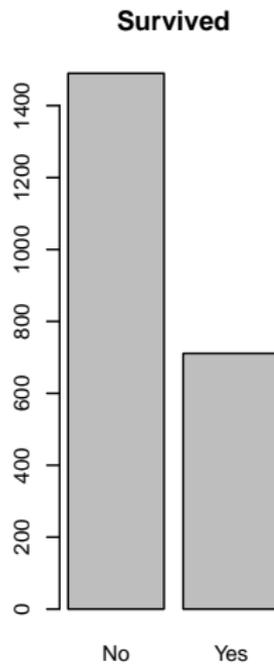
```
> margin(Titanic, "Sex")/sum(Titanic)
```

Male	Female
0.7864607	0.2135393

```
> margin(Titanic, "Class")/sum(Titanic)
```

1st	2nd	3rd	Crew
0.1476602	0.1294866	0.3207633	0.4020900

# Balkendiagramm

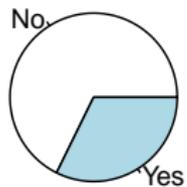


# Balkendiagramm

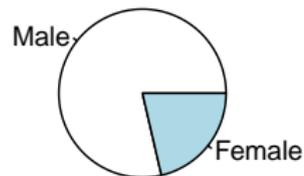
- ▶ Häufigkeiten werden als Flächen dargestellt.
- ▶ Häufigkeiten werden als Höhen dargestellt.
- ▶ Was sind die Unterschiede zum Histogramm?
- ▶ Was muß man bei ordinalen Daten beachten?

# Kuchendiagramme

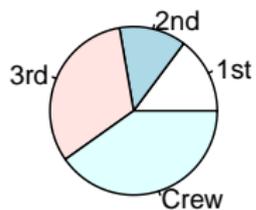
## Survived



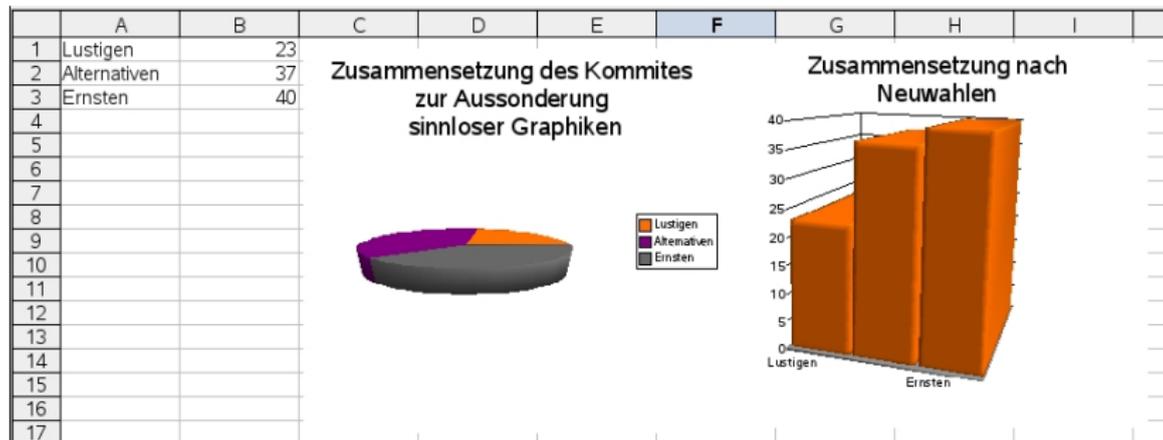
## Geschlecht



## Klasse



# Torte oder Diät



# Lesbare diskrete Graphiken

- ▶ Balkendiagramme

# Einteilung der Graphiken

		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret		
	stetig		

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ stetig–stetig
- ▶ diskret–diskret
- ▶ diskret–stetig

# Einteilung der Graphiken

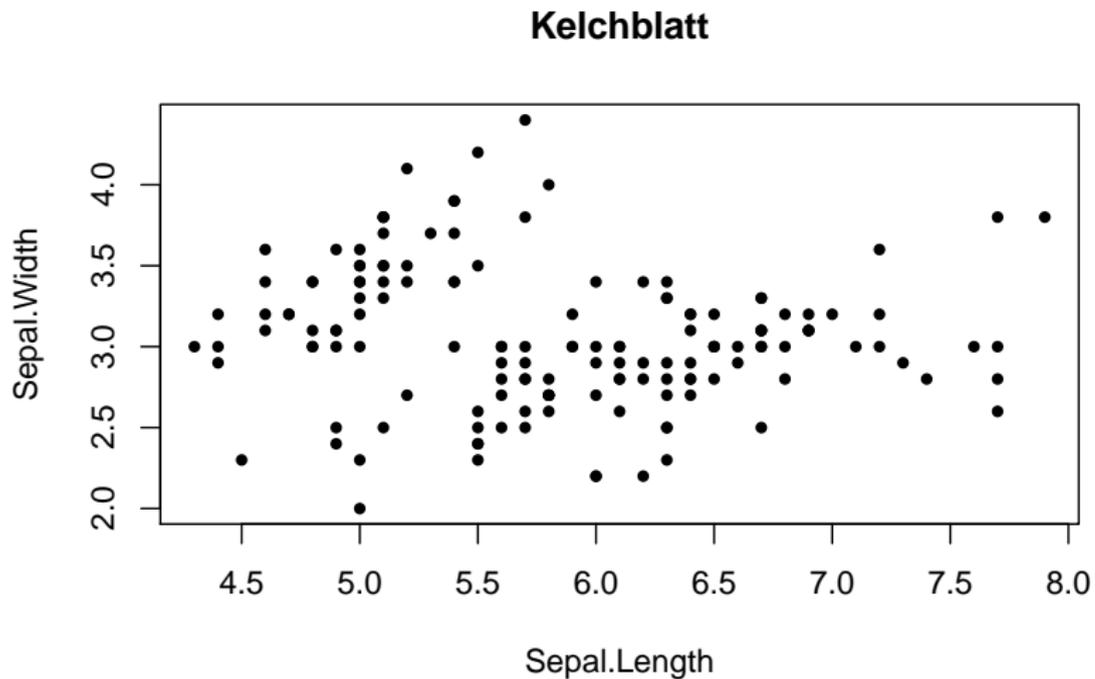
		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret		
	stetig		X

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ stetig–stetig **als nächstes**
- ▶ diskret–diskret
- ▶ diskret–stetig

# Stetig–Stetig

- ▶ Streudiagramm
- ▶ Kenngrößen für stetige Abhängigkeit
- ▶ QQ-plot
- ▶ Streudiagrammmatrix

# Streudiagramm



# Streudiagramm

- ▶ Überlagerung bei Bindungen
- ▶ Verzerrung durch Ausreißer
- ▶ Probleme bei extremer Schiefe
- ▶ Nicht: Kenngrößen, nahe Ausreißer

## (Pearson) Korrelation

$$\hat{\text{côr}}(X, Y) = \frac{\hat{\text{côv}}(X, Y)}{\sqrt{\hat{\text{vâr}}(X)\hat{\text{vâr}}(Y)}}$$

$$\hat{\text{vâr}}(X) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\hat{\text{vâr}}(Y) = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

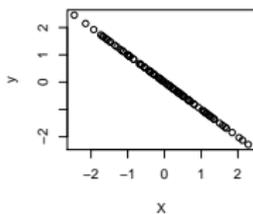
$$\hat{\text{côv}}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

# Theoretischen Interpretation

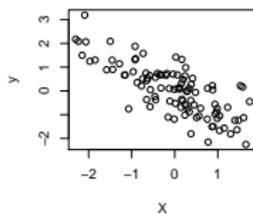
- ▶  $-1 \leq \text{cor}(X, Y) \leq 1$
- ▶ stochastisch unabhängig  
 $\Rightarrow \text{cov}(X, Y) = 0 \Rightarrow \text{cor}(X, Y) \neq 0 \nRightarrow \text{cor}(X, Y) = 0$
- ▶  $\text{cor}(X, Y) = 1 \Leftrightarrow X \propto Y$
- ▶  $\text{cor}(X, Y) = -1 \Leftrightarrow X \propto -Y$

# (Pearson) Korrelation

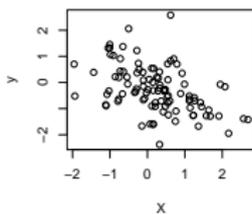
$\text{cor}(X,Y) = -1$



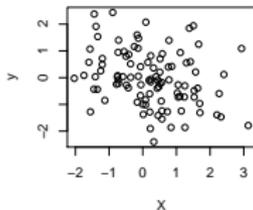
$\text{cor}(X,Y) = -0.75$



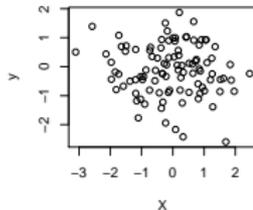
$\text{cor}(X,Y) = -0.5$



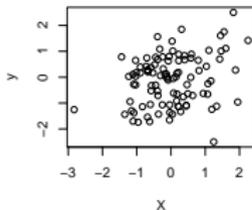
$\text{cor}(X,Y) = -0.25$



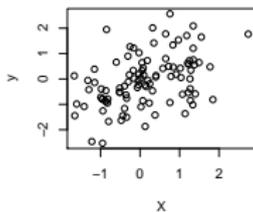
$\text{cor}(X,Y) = 0$



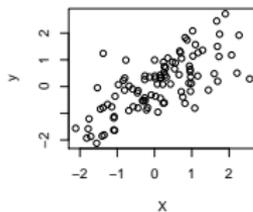
$\text{cor}(X,Y) = 0.25$



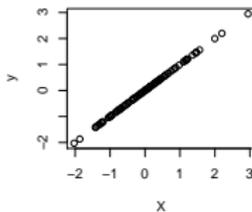
$\text{cor}(X,Y) = 0.5$



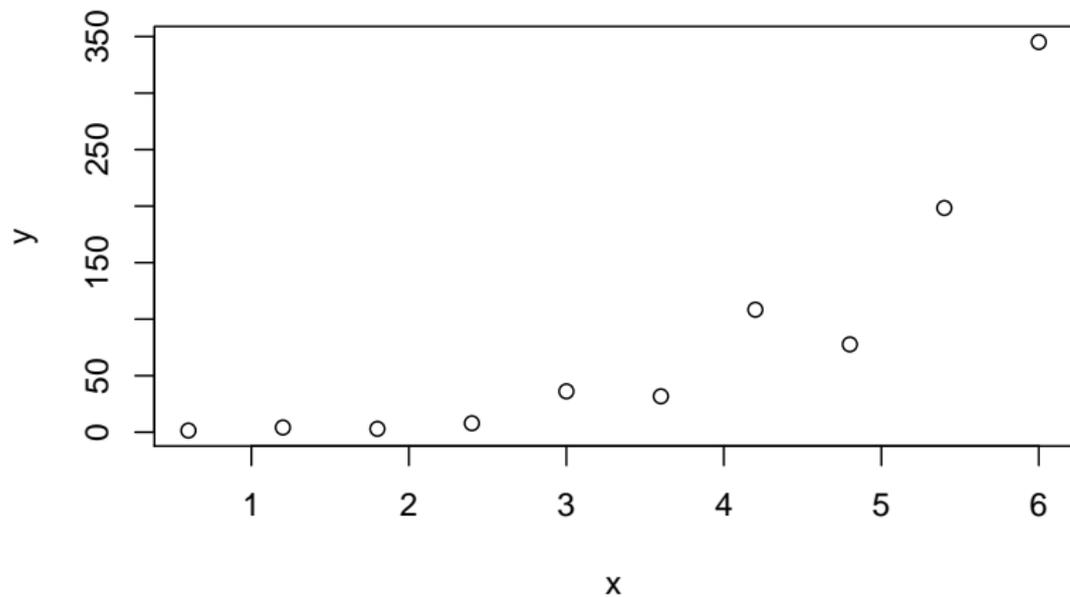
$\text{cor}(X,Y) = 0.75$



$\text{cor}(X,Y) = 1$



# Motivation für Rangkorrelation



# Rangziffern

$r_{\uparrow i}$  = Rang der i-ten Beobachtung

```
> x
```

```
[1] 1.8 1.2 0.6 2.4 3.0 3.6 4.2 4.8 6.0 5.4
```

```
> rank(x)
```

```
[1] 3 2 1 4 5 6 7 8 10 9
```

```
> y
```

```
[1] 2.963372 4.191878 1.482011 7.934833 36.176051  
[6] 31.753800 108.401706 77.721528 345.201323 198.373120
```

```
> rank(y)
```

```
[1] 2 3 1 4 6 5 8 7 10 9
```

# Rangverfahren

- ▶ **Idee:** Ersetze Daten durch ihren Rang

# Rangverfahren

- ▶ **Idee:** Ersetze Daten durch ihren Rang
- ▶ **Vorteil:** Die Auswertbarkeit ist unabhängig von der Verteilung.

# Rangverfahren

- ▶ **Idee:** Ersetze Daten durch ihren Rang
- ▶ **Vorteil:** Die Auswertbarkeit ist unabhängig von der Verteilung.
- ▶ **Nachteil 1:** Man verliert Information.

# Rangverfahren

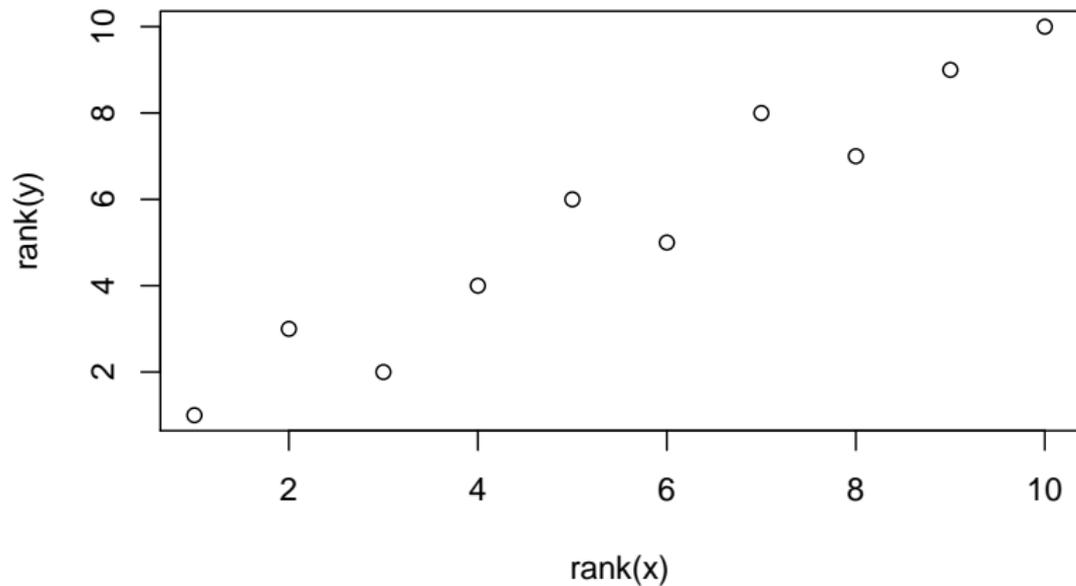
- ▶ **Idee:** Ersetze Daten durch ihren Rang
- ▶ **Vorteil:** Die Auswertbarkeit ist unabhängig von der Verteilung.
- ▶ **Nachteil 1:** Man verliert Information.
- ▶ **Nachteil 2:** Interpretation schwieriger.

# Rangverfahren

- ▶ **Idee:** Ersetze Daten durch ihren Rang
- ▶ **Vorteil:** Die Auswertbarkeit ist unabhängig von der Verteilung.
- ▶ **Nachteil 1:** Man verliert Information.
- ▶ **Nachteil 2:** Interpretation schwieriger.
- ▶ **Problem:** Rangziffernbestimmung bei Bindungen problematisch

# Spearman Korrelation

```
> plot(rank(x),rank(y))
```



# Spearman Korrelation

```
> cor(x,y)
```

```
[1] 0.8411657
```

```
> cor(rank(x),rank(y))
```

```
[1] 0.9636364
```

```
> cor(x,y,method="spearman")
```

```
[1] 0.9636364
```

```
> cor(x,y,method="pearson")
```

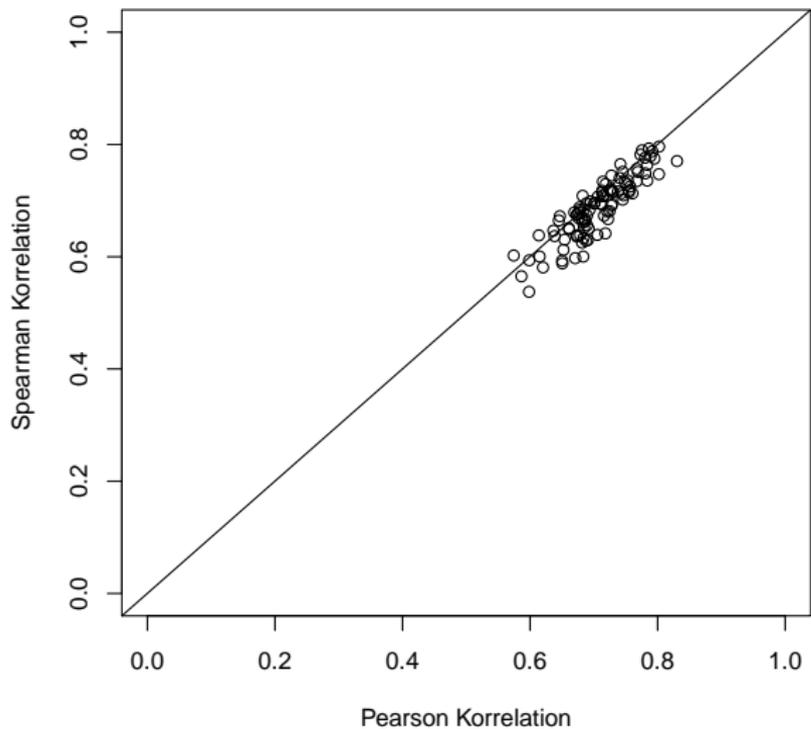
```
[1] 0.8411657
```

# Vergleich

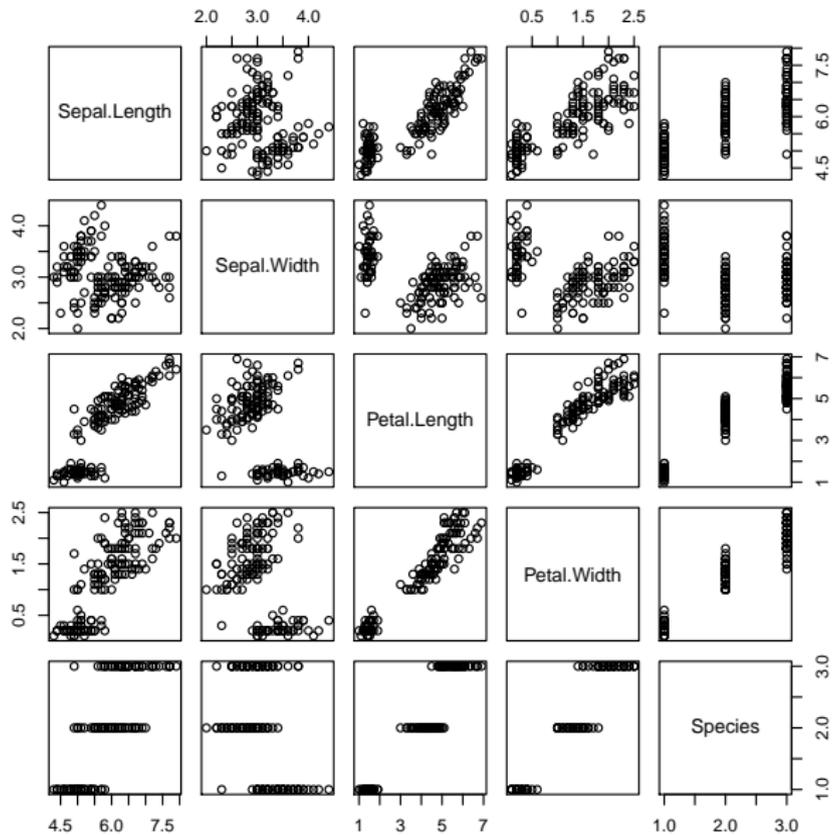
- ▶ Pearson Korrelation quantifiziert lineare Abhängigkeit
- ▶ Spearman Korrelation quantifiziert monotone Abhängigkeit

# Vergleich

100 Datensätze a 100 Beobachtungen mit  $\text{cor}=0.707$



# Streudiagrammmatrix



# Einteilung der Graphiken

		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret		
			x

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ \*stetig–stetig
- ▶ diskret–diskret
- ▶ diskret–stetig

# Einteilung der Graphiken

		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret	X	
	stetig		x

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ \*stetig–stetig
- ▶ diskret–diskret **als nächstes**
- ▶ diskret–stetig

# diskret–diskret

- ▶ gestapelte Balkendiagramme
- ▶ parallele Balkendiagramme
- ▶ Mosaikplots

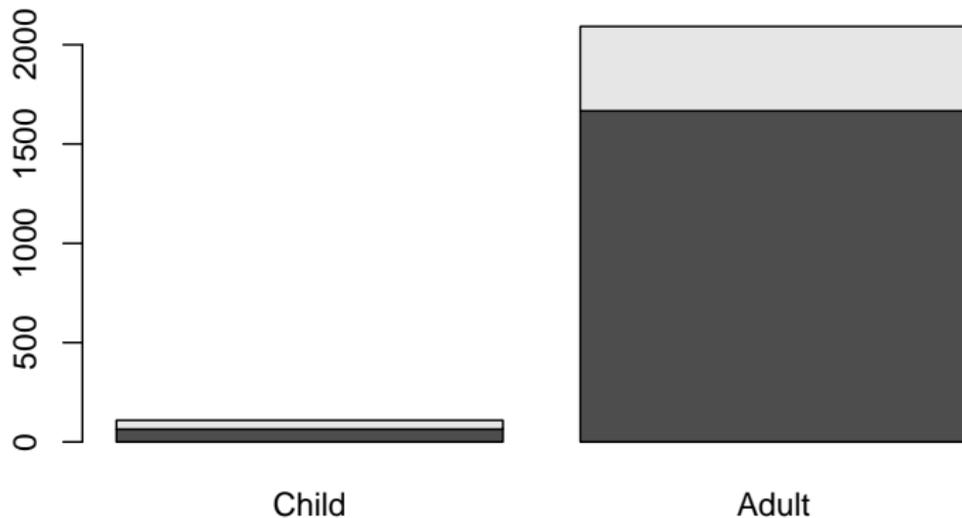
# Passagiere der Titanic

```
> data(Titanic)
> X <- apply(Titanic, c(2,3), sum)
> X
```

Sex	Age	
	Child	Adult
Male	64	1667
Female	45	425

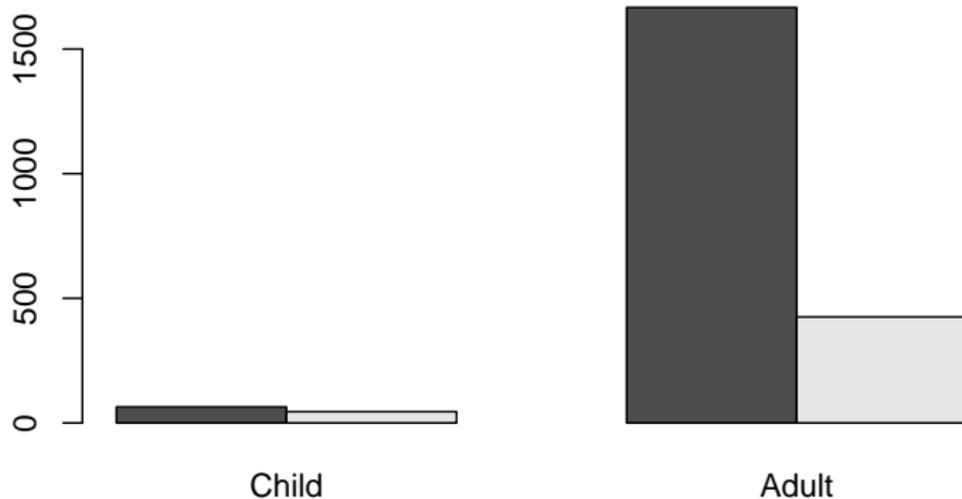
# gestapelte Balkendiagramme

**Gestapeltes Balkendiagramm**



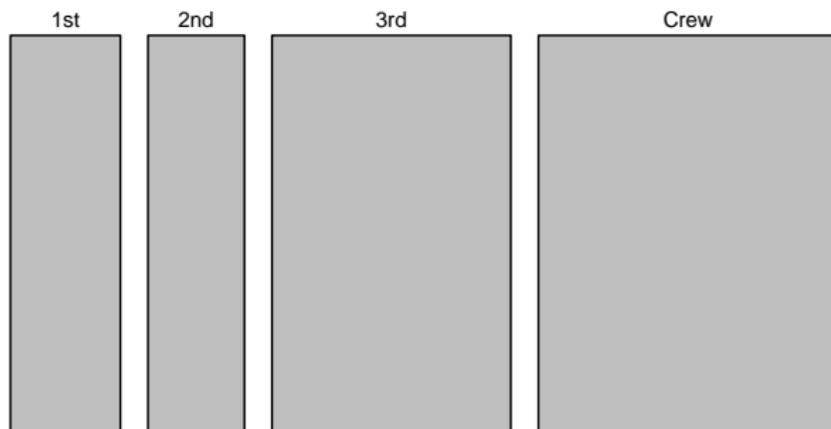
# parallele Balkendiagramme

**Paralleles Balkendiagramm**

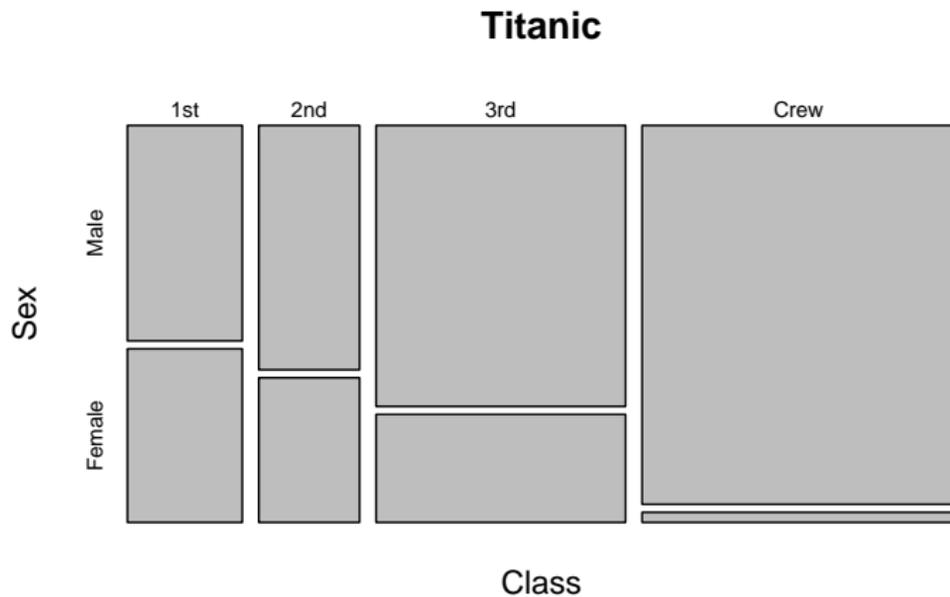


# Mosaikplot (1 Variable)

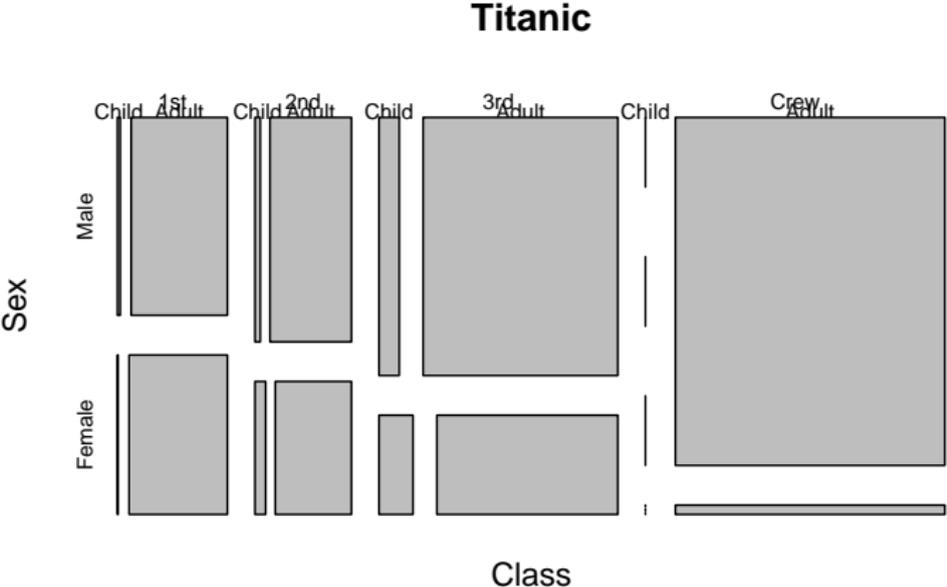
## Titanic



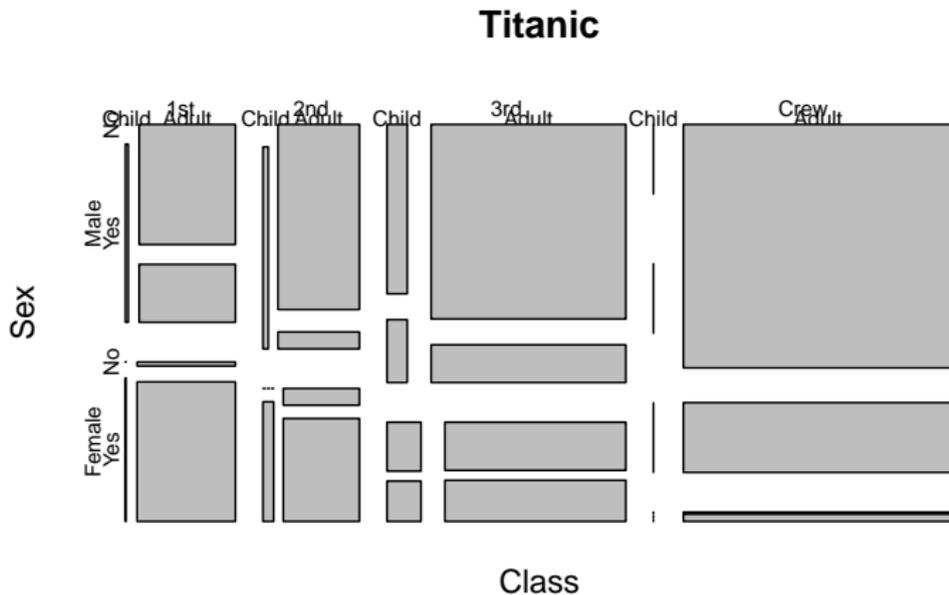
# Mosaikplot (2 Variablen)



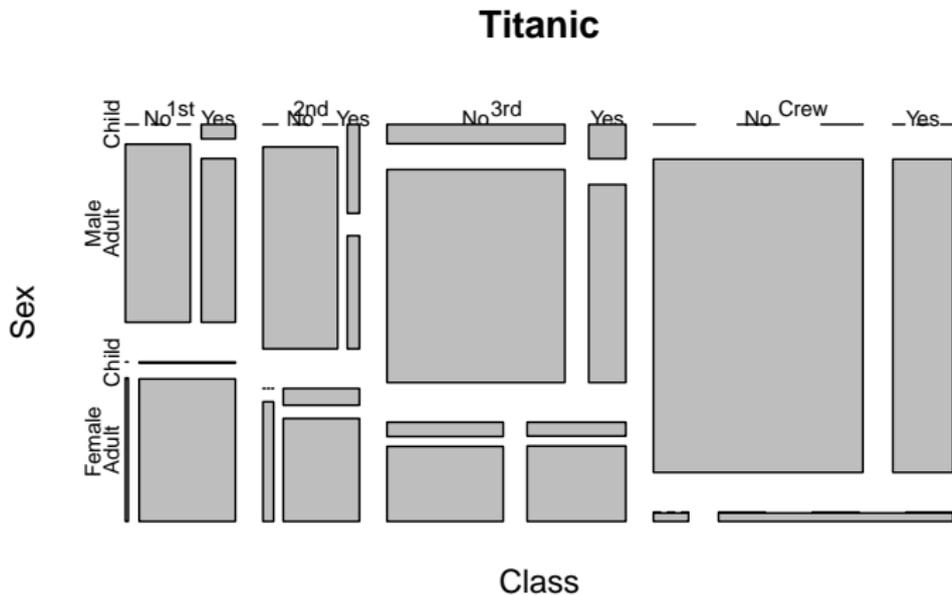
# Mosaikplot (3 Variablen)



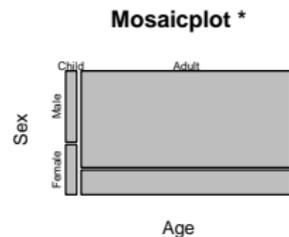
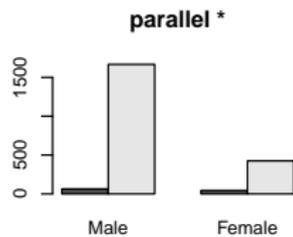
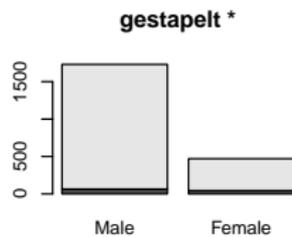
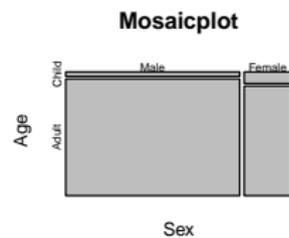
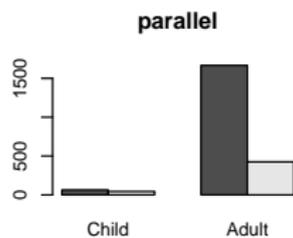
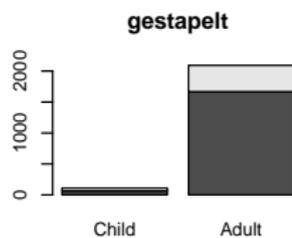
# Mosaikplot (4 Variablen)



# Mosaikplot (Andere Reihenfolge)



# Vergleich



# Wer kann was?

- ▶ Überblick: stapeln
- ▶ Vergleich von Teilgruppen: parallel
- ▶ Bedingte Wahrscheinlichkeiten: Mosaik

# Einteilung der Graphiken

		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret	x	
	stetig		x

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ \*stetig–stetig
- ▶ \*diskret–diskret
- ▶ diskret–stetig

# Einteilung der Graphiken

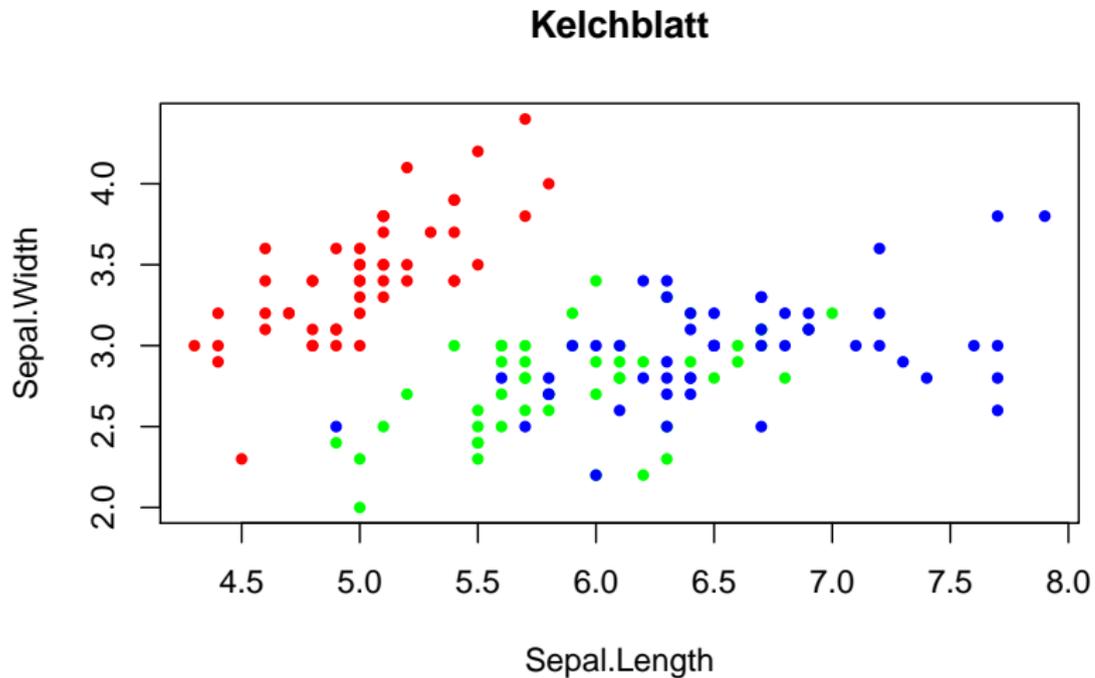
		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret	x	X
	stetig	X	x

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ \*stetig–stetig
- ▶ \*diskret–diskret
- ▶ diskret–stetig als nächstes

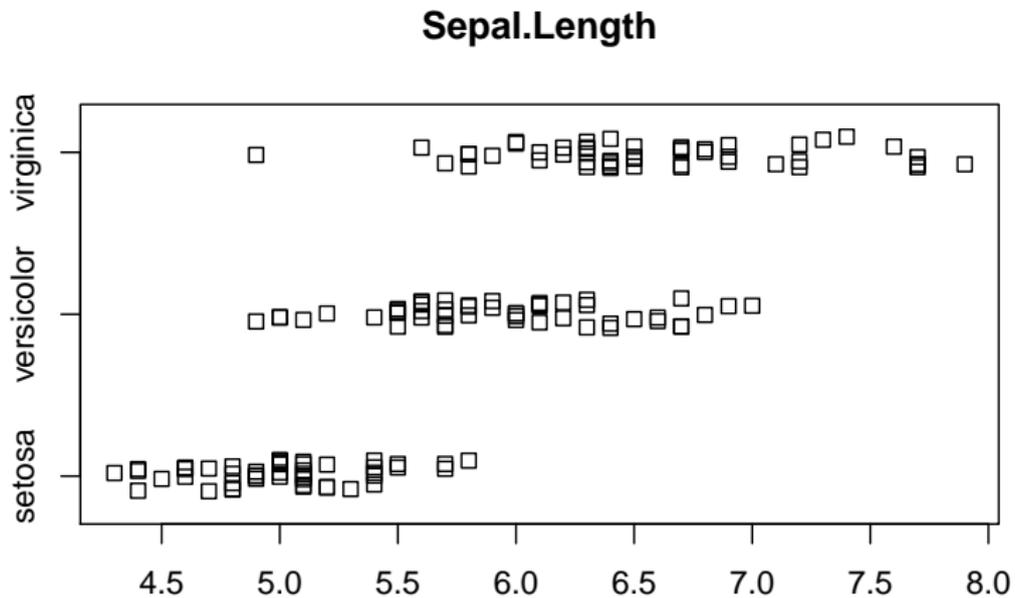
# diskret–stetig

- ▶ Farben und Symbole
- ▶ parallele Punktdiagramme
- ▶ parallele Boxplots
- ▶ gekerbte Boxplots

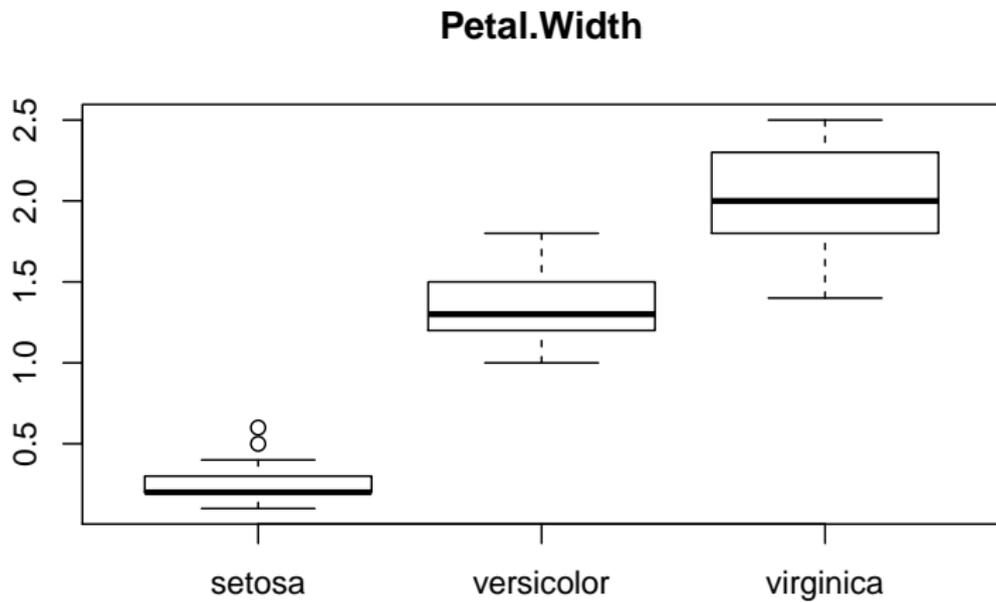
# Farben und Symbole



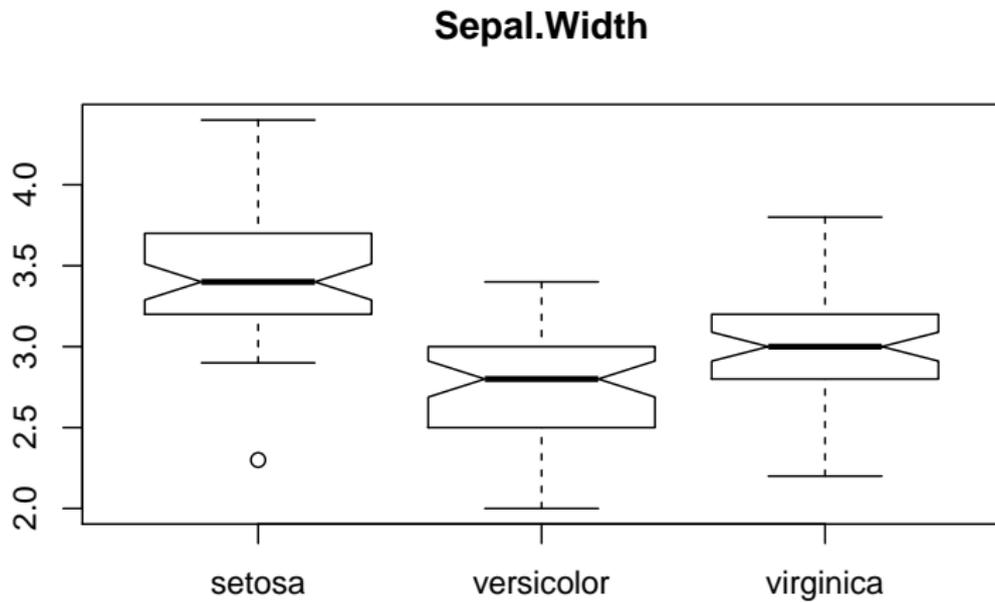
# parallele Punktdiagramme



# parallele Boxplot



# Boxplot (gekerbt)



# Interpretation

- ▶ Sind die Mediane gleich so überlagern sich die Kerben mit einer Wahrscheinlichkeit von 95%.
- ▶ Überlagern sich die Kerben nicht, so ist das ein Hinweis auf verschiedene Mediane.

# Einteilung der Graphiken

		Erste Variable	
		diskret	stetig
zweite Variable	keine	x	x
	diskret	x	x
	stetig	s.o.	x

- ▶ \*stetige Daten
- ▶ \*diskrete Daten
- ▶ \*stetig–stetig
- ▶ \*diskret–diskret
- ▶ \*diskret–stetig

# Symbolik

- ▶ Kategorien  $\sim$  Farben, Formen, Position
- ▶ Reelle Zahlen  $\sim$  Position
- ▶ Positive Zahlen  $\sim$  Position, Fläche, log Positionen
- ▶ Anzahlen, Wahrscheinlichkeiten  $\sim$  Flächen, Höhen
- ▶ Dichten  $\sim$  Höhe

# Zweck der Graphik

- ▶ Wie sind die Daten?
- ▶ Gibt es Ausreißer und Verteilungsbesonderheiten?
- ▶ Welche Zusammenhänge kann man erkennen/vermuten?
- ▶ Können wir unsere Vermutungen graphisch bestätigen?
- ▶ Wie geht es weiter?

# Fragen an die Graphiken

- ▶ Ist etwas ungewöhnlich? Warum?
- ▶ Wie sind die Daten verteilt?
- ▶ Gibt es Ausreißer oder Bindungen?
- ▶ Wird der optische Eindruck durch Besonderheiten verfälscht (z.B. Bindungen, zu kleine Balken, Überlagerung)
- ▶ Welche Abhängigkeiten sind erkennbar?
- ▶ Sind die Abhängigkeiten stark oder schwach, linear oder nichtlinear, zunehmen oder abnehmend?
- ▶ Entsprechend die Beobachtungen dem, was man inhaltlich erwarten würde?
- ▶ Was fällt sonst auf?

# Masszahlen

Masszahlen werden verwendet um bestimmte Aspekte der Verteilung zusammenfassend darzustellen.

- ▶ Lage
- ▶ Streuung
- ▶ Form
- ▶ Zusammenhang
- ▶ Anteil
- ▶ fehlt noch: diskret-diskret, diskret-stetig (später  $R^2$ )