

# EINFÜHRUNG IN DIE DATENANALYSE MIT R - EINFACHE GRAFIKEN

Jan-Philipp Kolb

09 Mai, 2019

# EIN PLOT SAGT MEHR ALS 1000 WORTE

- ▶ Grafisch gestützte Datenanalyse ist toll
- ▶ Gute Plots können zu einem besseren Verständnis beitragen
- ▶ Einen Plot zu generieren geht schnell
- ▶ Einen guten Plot zu machen kann sehr lange dauern
- ▶ Mit R Plots zu generieren macht Spaß
- ▶ Mit R erstellte Plots haben hohe Qualität
- ▶ Fast jeder Plottyp wird von R unterstützt
- ▶ R kennt eine große Menge an Exportformaten für Grafiken

# PLOT IST NICHT GLEICH PLOT

- ▶ Bereits das base Package bringt eine große Menge von Plot Funktionen mit
- ▶ Das lattice Packet erweitert dessen Funktionalität
- ▶ Eine weit über diese Einführung hinausgehende Übersicht findet sich in Murrell, P (2006): R Graphics.

# CRAN TASK VIEWS

- ▶ Zu einigen Themen sind alle Möglichkeiten in R zusammengestellt. (Übersicht der Task Views)
- ▶ Zur Zeit gibt es 35 Task Views
- ▶ Alle Pakete eines Task Views können mit folgendem Befehl installiert werden:

```
install.packages("ctv")  
library("ctv")  
install.views("Bayesian")
```

## CRAN Task Views

[Bayesian](#)

Bayesian Inference

[ChemPhys](#)

Chemometrics and Computational Physics

[ClinicalTrials](#)

Clinical Trial Design, Monitoring, and Analysis

[Cluster](#)

Cluster Analysis & Finite Mixture Models

[DifferentialEquations](#)

Differential Equations

[Distributions](#)

Probability Distributions

[Econometrics](#)

Econometrics

[Environmetrics](#)

Analysis of Ecological and Environmental Data

# TASK VIEW ZU THEMA GRAPHIKEN

CRAN Task View: Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization

**Maintainer:** Nicholas Lewin-Koh

**Contact:** nikko at haimail.net

**Version:** 2015-01-07

**URL:** <https://CRAN.R-project.org/view=Graphics>

R is rich with facilities for creating and developing interesting graphics. Base R contains functionality for many plot types including coplots, mosaic plots, biplots, and the list goes on. There are devices such as postscript, png, jpeg and pdf for outputting graphics as well as device drivers for all platforms running R. [lattice](#) and grid are supplied with R's recommended packages and are included in every binary distribution. [lattice](#) is an R implementation of William Cleveland's trellis graphics, while grid defines a much more flexible graphics environment than the base R graphics.

# GRAFIKEN FÜR BEDINGTE, BI- UND MULTIVARIATE VERTEILUNGEN - SCATTERPLOTS

- ▶ Funktion `plot()` ist eine generische Funktion - bspw. kann ein einfacher Scatterplot erstellt werden
- ▶ Für einen solchen muss `plot()` ein `x` und ein `y` Beobachtungsvektor übergeben werden
- ▶ Um die Farbe der Plot-Symbole anzupassen gibt es die Option `col` (Farbe als character oder numerisch)
- ▶ Die Plot-Symbole selbst können mit `pch` (plotting character) angepasst werden (character oder numerisch)
- ▶ Die Achsenbeschriftungen (labels) werden mit `xlab` und `ylab` definiert

# BEISPIEL - IRIS DATENSATZ

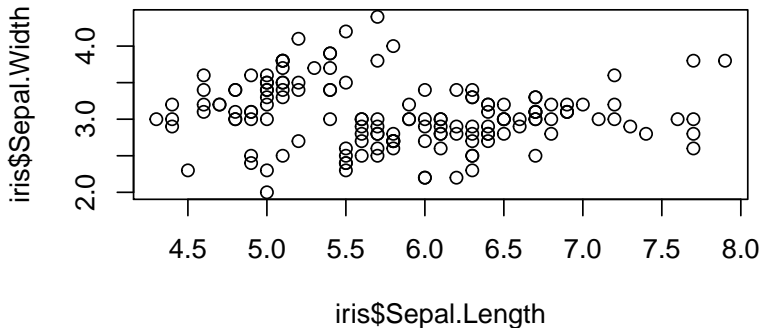
```
data(iris)
```

```
head(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
## 4	4.6	3.1	1.5	0.2	setosa
## 5	5.0	3.6	1.4	0.2	setosa
## 6	5.4	3.9	1.7	0.4	setosa

## EIN ERSTER SCATTERPLOT

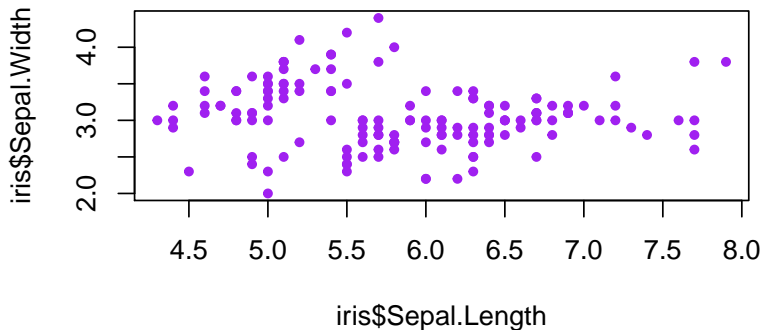
```
plot(iris$Sepal.Length,iris$Sepal.Width)
```





## ANDERE FARBE UND PUNKTTYP

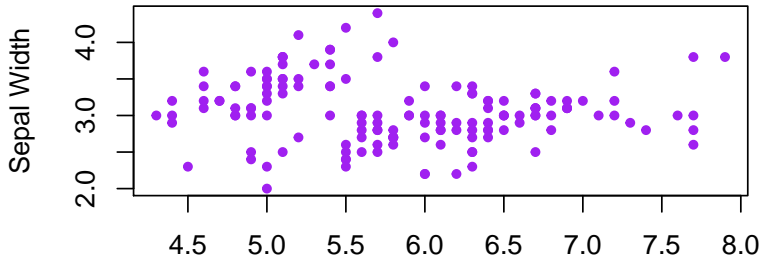
```
plot(iris$Sepal.Length,iris$Sepal.Width,pch=20,  
     col="purple")
```



## BESCHRIFTUNG HINZUFÜGEN

```
plot(iris$Sepal.Length,iris$Sepal.Width,pch=20,  
     col="purple",main="Iris dataset",  
     xlab="Sepal Length",ylab="Sepal Width")
```

**Iris dataset**



# DATENSATZ

```
library(mlmRev)
data(Chem97)
```

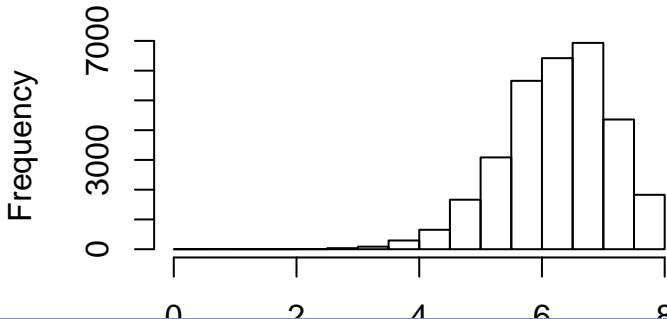
- ▶ [lea] Local Education Authority - a factor
- ▶ [school] School identifier - a factor
- ▶ [student] Student identifier - a factor
- ▶ [score] Point score on A-level Chemistry in 1997
- ▶ [gender] Student's gender
- ▶ [age] Age in month, centred at 222 months or 18.5 years
- ▶ [gcsecore] Average GCSE score of individual.
- ▶ [gcsecnt] Average GCSE score of individual, centered at mean.

# HISTOGRAMM - DIE FUNKTION HIST()

Wir erstellen ein Histogramm der Variable gcscscore:

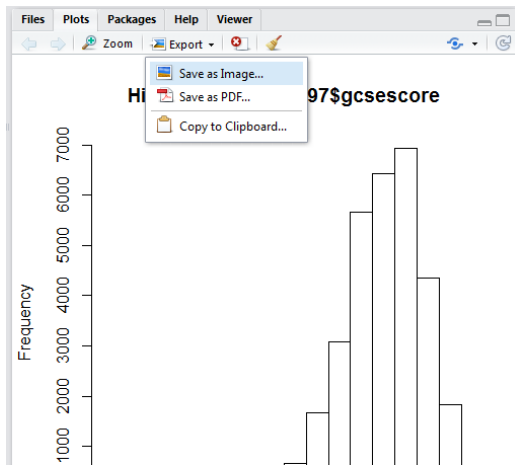
```
hist(Chem97$gcscscore)
```

## Histogram of Chem97\$gcscscore



# GRAPHIK SPEICHERN

- ▶ Mit dem button Export in Rstudio kann man die Grafik speichern.



# BEFEHL UM GRAPHIK ZU SPEICHERN

- Alternativ auch bspw. mit den Befehlen `png`, `pdf` oder `jpeg`

```
png("Histogramm.png")  
hist(Chem97$gcsescore)  
dev.off()
```

# HISTOGRAMME

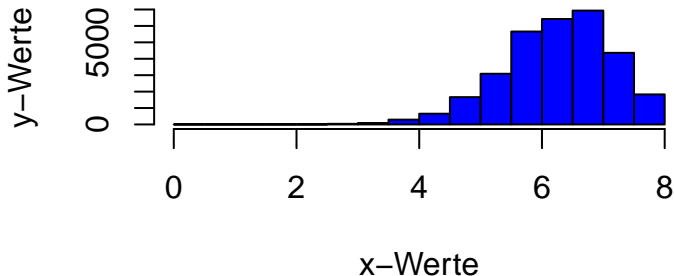
- ▶ Die Funktion `hist()` plottet ein Histogramm der Daten
- ▶ Der Funktion muss mindestens ein Beobachtungsvektor übergeben werden
- ▶ `hist()` hat noch sehr viel mehr Argumente, die alle (sinnvolle) default values haben

	Argument	Bedeutung	Beispiel
main	Überschrift		main='Hallo Welt'
xlab	x-Achsenbeschriftung		xlab='x-Werte'
ylab	y-Achsenbeschriftung		ylab='y-Werte'
col	Farbe		col='blue'

# HISTOGRAMM

```
hist(Chem97$gcsescore,col="blue",  
     main="Hallo Welt",ylab="y-Werte", xlab="x-Werte")
```

## Hallo Welt





# BARPLOT

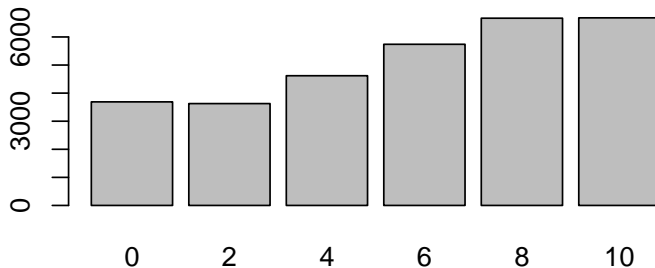
- ▶ Die Funktion `barplot()` erzeugt aus einer Häufigkeitstabelle einen Barplot
- ▶ Ist das übergebene Tabellen-Objekt zweidimensional wird ein bedingter Barplot erstellt

```
tabScore <- table(Chem97$score)
```

```
barplot(tabScore)
```

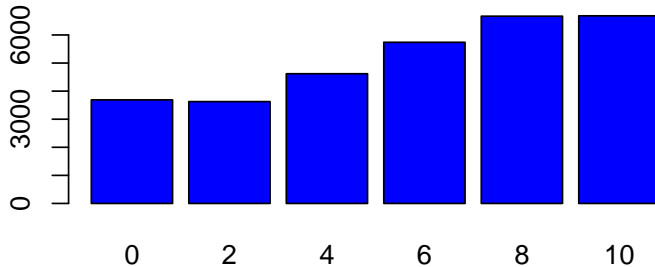
# BARPLOTS UND BARCHARTS

```
barplot(tabScore)
```



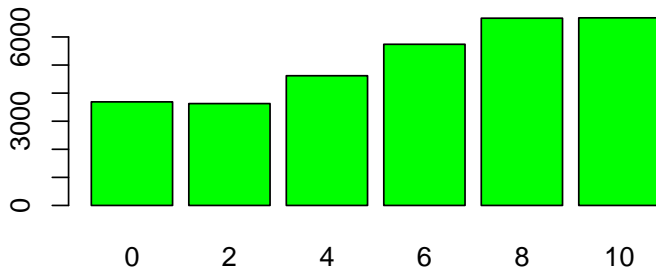
## MEHR FARBEN:

```
barplot(tabScore,col=rgb(0,0,1))
```



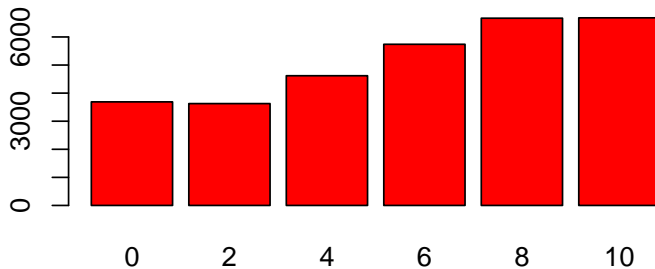
## GRÜNE FARBE

```
barplot(tabScore,col=rgb(0,1,0))
```



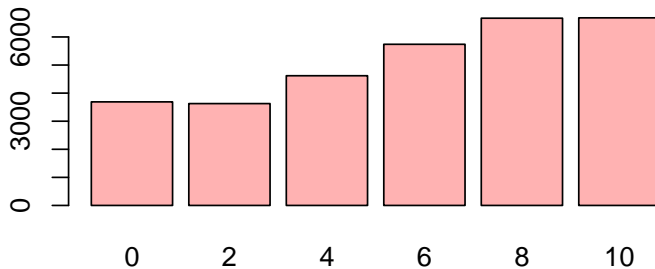
# ROTE FARBE

```
barplot(tabScore,col=rgb(1,0,0))
```



# TRANSPARENT

```
barplot(tabScore,col=rgb(1,0,0,.3))
```



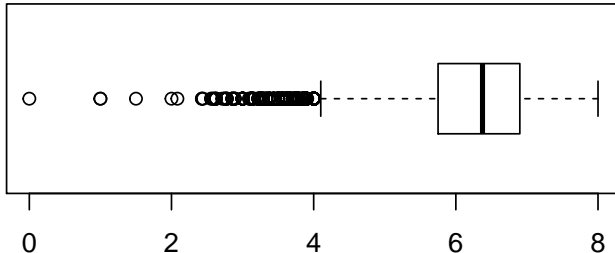
# BOXPLOT

- ▶ Einen einfachen Boxplot erstellt man mit `boxplot()`
- ▶ Auch `boxplot()` muss mindestens ein Beobachtungsvektor übergeben werden

```
?boxplot
```

# HORIZONTALER BOXPLOT

```
boxplot(Chem97$gcsescore,  
horizontal=TRUE)
```



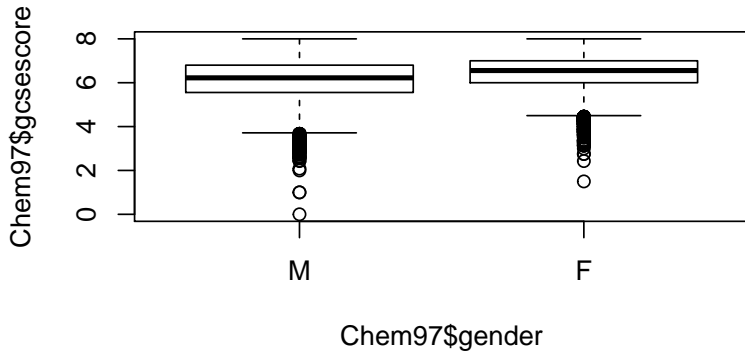


# GRUPPIERTE BOXPLOTS

- ▶ Ein sehr einfacher Weg, einen ersten Eindruck über bedingte Verteilungen zu bekommen ist über sog. Gruppierte notched Boxplots
- ▶ Dazu muss der Funktion `boxplot()` ein sog. Formel-Objekt übergeben werden
- ▶ Die bedingende Variable steht dabei auf der rechten Seite einer Tilde

## BEISPIEL GRUPIERTER BOXPLOT

```
boxplot(Chem97$gcsescore~Chem97$gender)
```



# ALTERNATIVEN ZU BOXPLOT

## VIOPLLOT

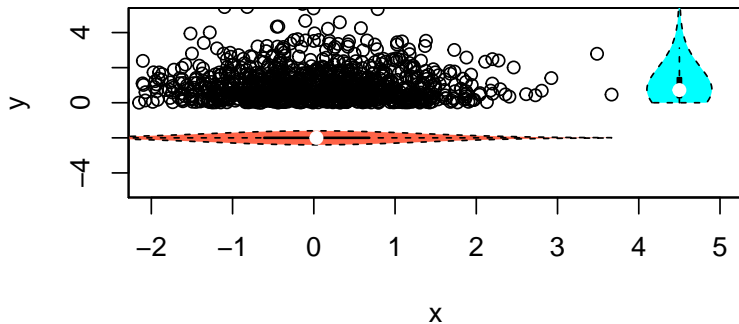
- ▶ Baut auf Boxplot auf
- ▶ Zusätzlich Informationen über Dichte der Daten
- ▶ Dichte wird über Kernel Methode berechnet.
- ▶ weißer Punkt - Median
- ▶ Je weiter die Ausdehnung, desto größer ist die Dichte an dieser Stelle.

```
# Beispieldaten erzeugen  
x <- rnorm(1000)  
y <- rexp(1000,1)
```

# BEISPIEL VIOPLOT

```
library(vioplot)
plot(x, y, xlim=c(-2,5), ylim=c(-5,5))
vioplot(x, col="tomato", horizontal=TRUE, at=-2,
        add=TRUE,lty=2, rectCol="gray")
vioplot(y, col="cyan", horizontal=FALSE, at=4.5,
        add=TRUE,lty=2)
```

# BEISPIEL VIOPLLOT



# ALTERNATIVEN ZUM BOXPLOT

```
library(beanplot)
par(mfrow = c(1,2))
boxplot(count~spray,data=InsectSprays,col="blue")
beanplot(count~spray,data=InsectSprays,col="orange")
```

