Statistik für Digital Humanities

Visualisierung

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

27. April 2020

[Letzte Aktualisierung: 26/04/2020, 16:59]

Überblick

1. Grundlagen statistischer Visualisierung
2. Typische Diagrammarten
3. Lies, damned Lies and Statistics

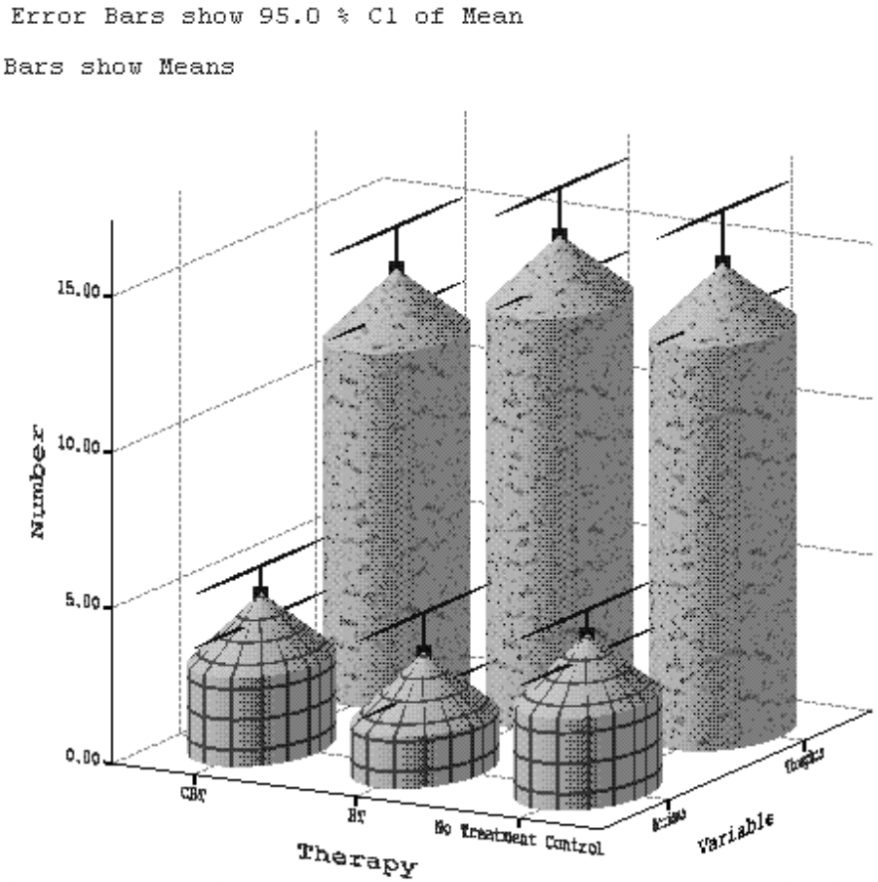
(Das letzte Kapitel ist nicht aus Fields Lehrbuch)

Grundsätze guter Visualisierung

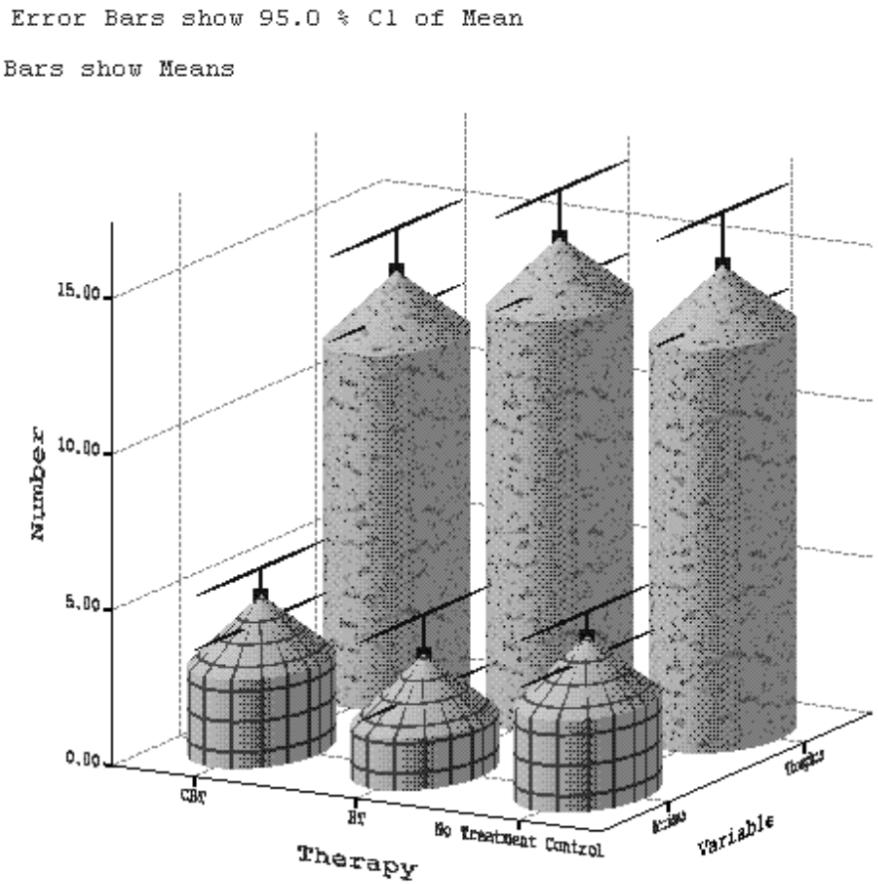
Nach Tufte E. R. (2001): *The visual display of quantitative information*

- Daten zeigen
- Leser_Innen dazu bringen, über die Daten nachzudenken (statt darüber, wie rot sie sind)
- **Graphmüll** (Chartjunk) vermeiden
- Verzerrung vermeiden
- **Minimum Ink**: Viel Information mit wenig Tinte zeigen
- Kohärenz herstellen, vor allem bei großen Datenmengen
- Vergleichbarkeit ermutigen
- Informationen offenbaren, Schlüsse zulassen

Graphmüll

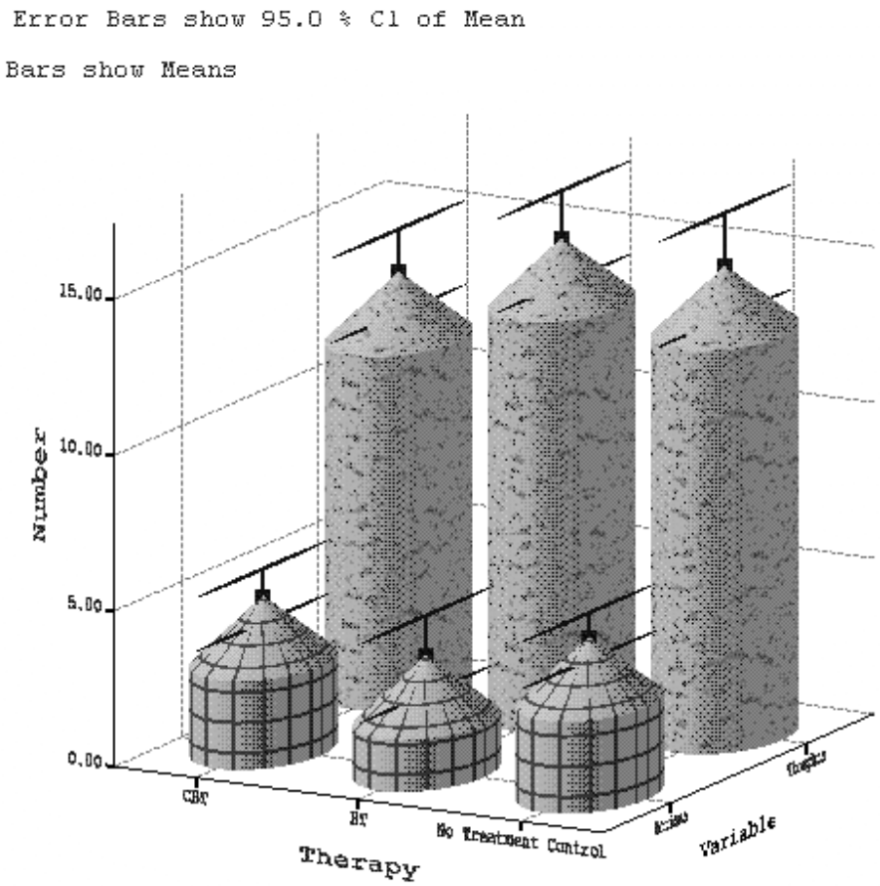

Error Bars show 95.0 % CI of Mean
Bars show Means
15.00
10.00
5.00
0.00
Number
CBT
BT
No Treatment Control
Therapy
Variable

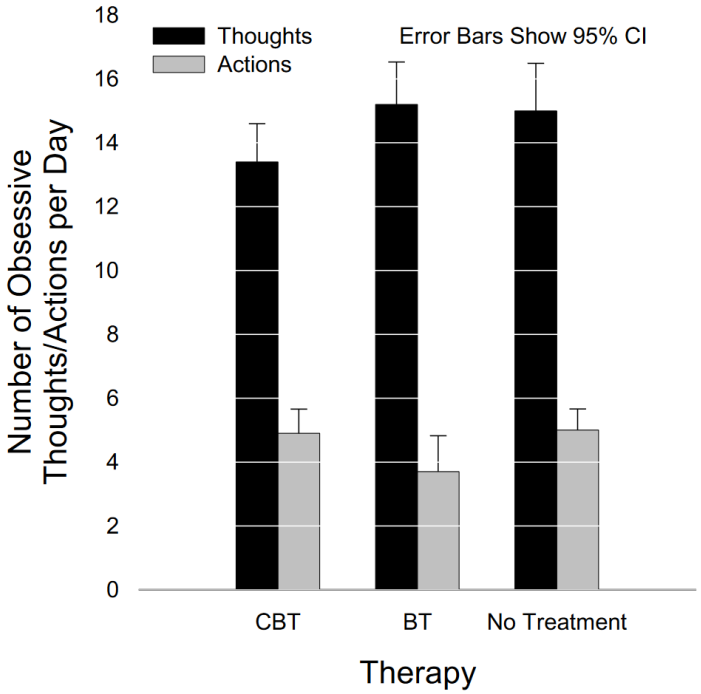
Graphmüll



- Nie 3D-Effekt bei 2D-Daten verwenden
- Informationslose Muster/Texturen
- Zylindrische Balken verzerren die Daten
- Schlecht benannte Achse *Numbers*

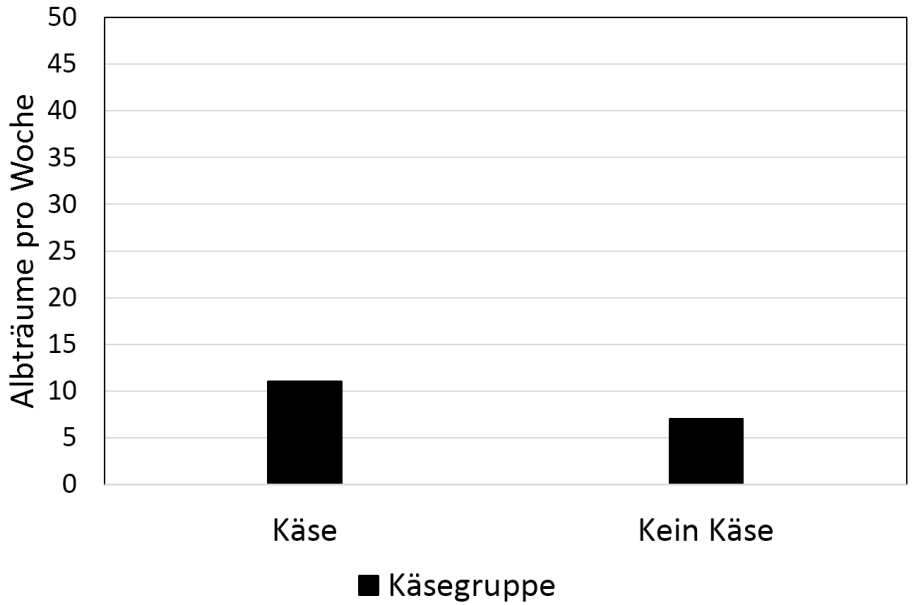
Graphmüll





- 2D
- Keine unnötigen Ablenkungen
- Minimum Ink
- aussagekräftige Achsenbezeichnung

Suggestion durch Formatierung



Grundlagen statistischer Visualisierung

- Vertikale Achse = y-Achse
- Horizontale Achse = x-Achse

Grundlagen statistischer Visualisierung

- Vertikale Achse = y-Achse
- Horizontale Achse = x-Achse

Grundregeln:

- Keine nicht vorhandenen Eindrücke suggerieren
- Keine wichtigen Effekte verbergen
- Minimum Ink
- Kein Graphmüll (Chartjunk)

Ausflug ggplot2

- ggplot2 = Visualisierungspaket für R
- Codesnippets im folgenden zur praktischen Veranschaulichung
- Installation

```
install.packages("ggplot2")
library(ggplot2)
```

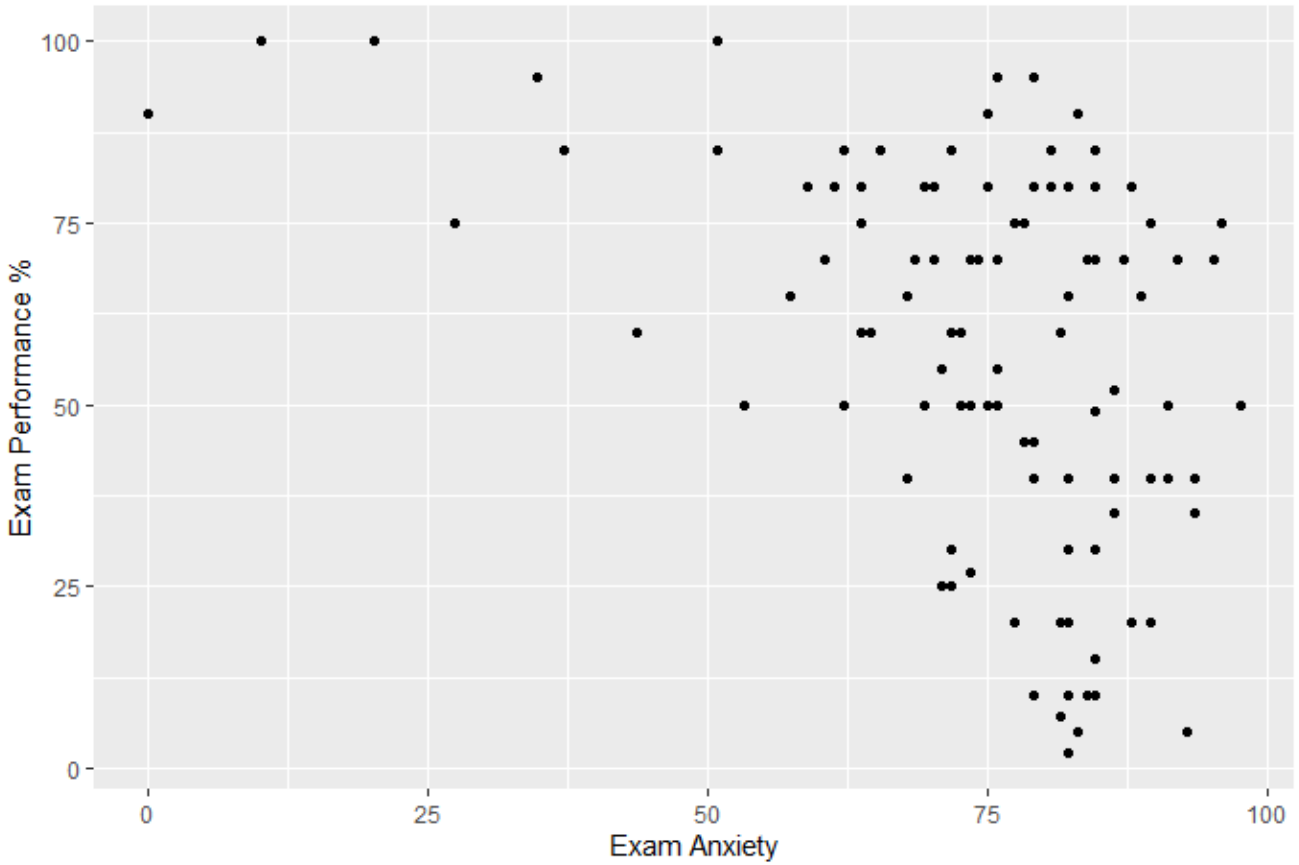
- Siehe Begleitlektüre für Tutorial und Datensätze sowie einen fertigen Start-Workspace für RStudio
- Folgendes Skript sollte einen Plot erzeugen

```
library(ggplot2)
data<-read.delim("FacebookNarcissism.dat", header=TRUE)
graph<-ggplot(data, aes(NPQC_R_Total, Rating))
graph + geom_point()
```

Überblick

1. Grundlagen statistischer Visualisierung
2. Typische Diagrammarten
   - Scatterplot
   - Histogramm
   - Line Charts
   - Boxplot
   - Density Plot
   - Bar Charts
     - 1 Unabhängige Variable
     - Bar Chart mit Konfidenzintervall
     - Mehrere Unabhängige Variablen
     - Facettierung
3. Lies, damned Lies and Statistics

Scatterplot

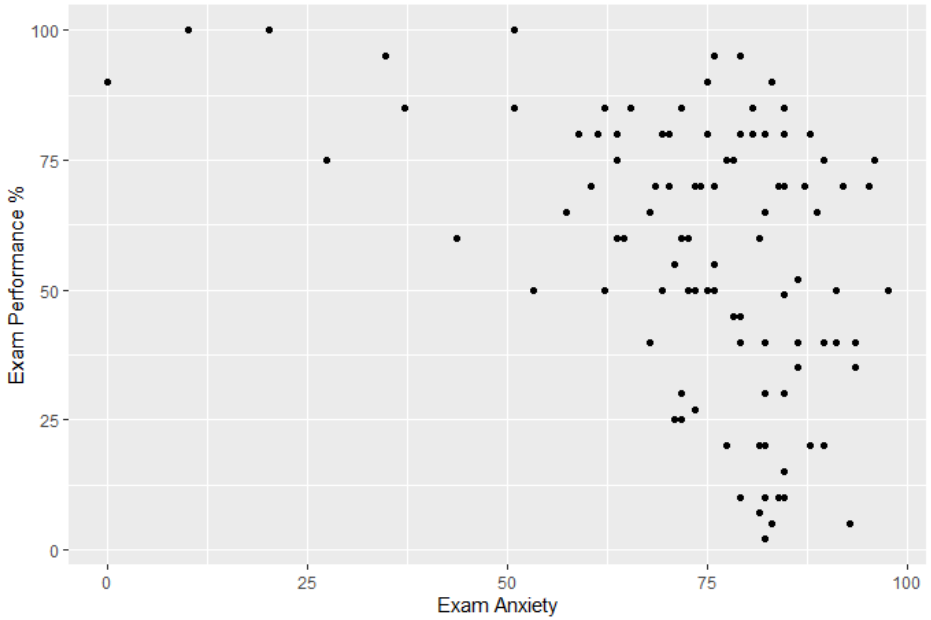



```
data<-read.delim("Exam Anxiety.dat", header=TRUE)
graph<-ggplot(data, aes(Anxiety, Exam))
graph + geom_point() + labs(x = "Exam Anxiety", y = "Exam Performance %")
```

Scatterplot

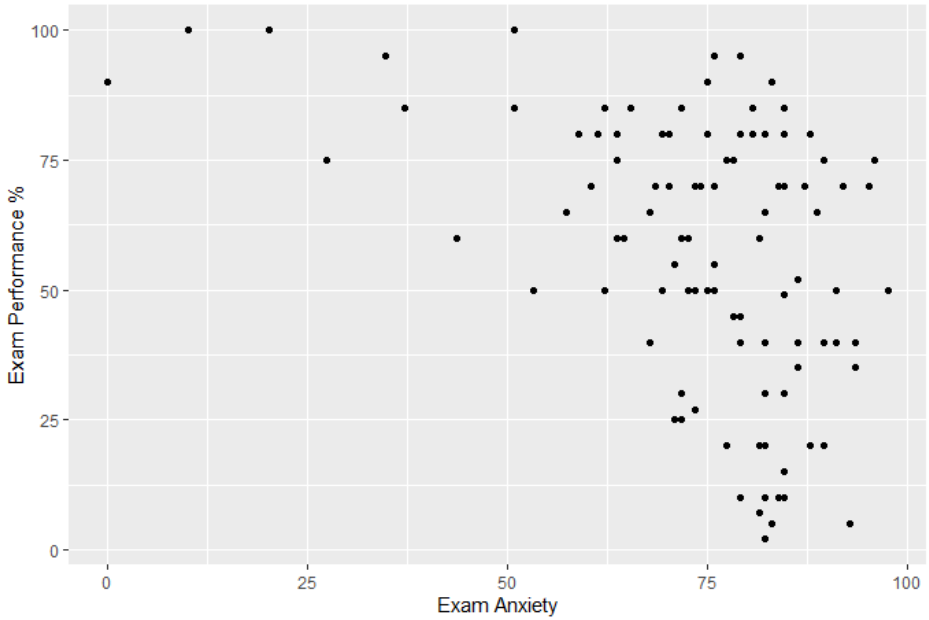
- 2D Daten
  - Mehr Dimensionen möglich (Farbe, Form,...) $\rightarrow$ Gruppierter Scatterplot
- Beziehungen zwischen 2 Variablen
- Extremwerte (Outlier) identifizierbar

Scatterplot



Interpretation:

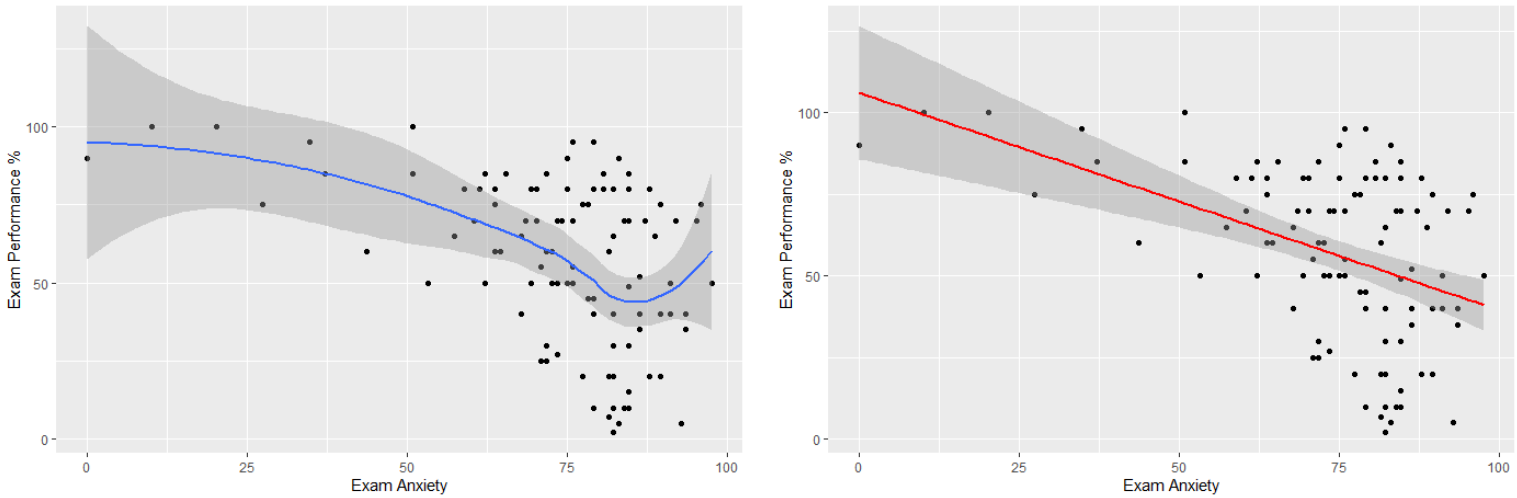
Scatterplot



Interpretation:

- Großteil hat Prüfungsangst
- Keine offensichtlichen Outlier
- Trend guter Noten bei geringer Prüfungsangst
- Trend variable Noten bei hoher Prüfungsangst
  - $\rightarrow$ Weniger Sorge erhöht Notenschnitt?

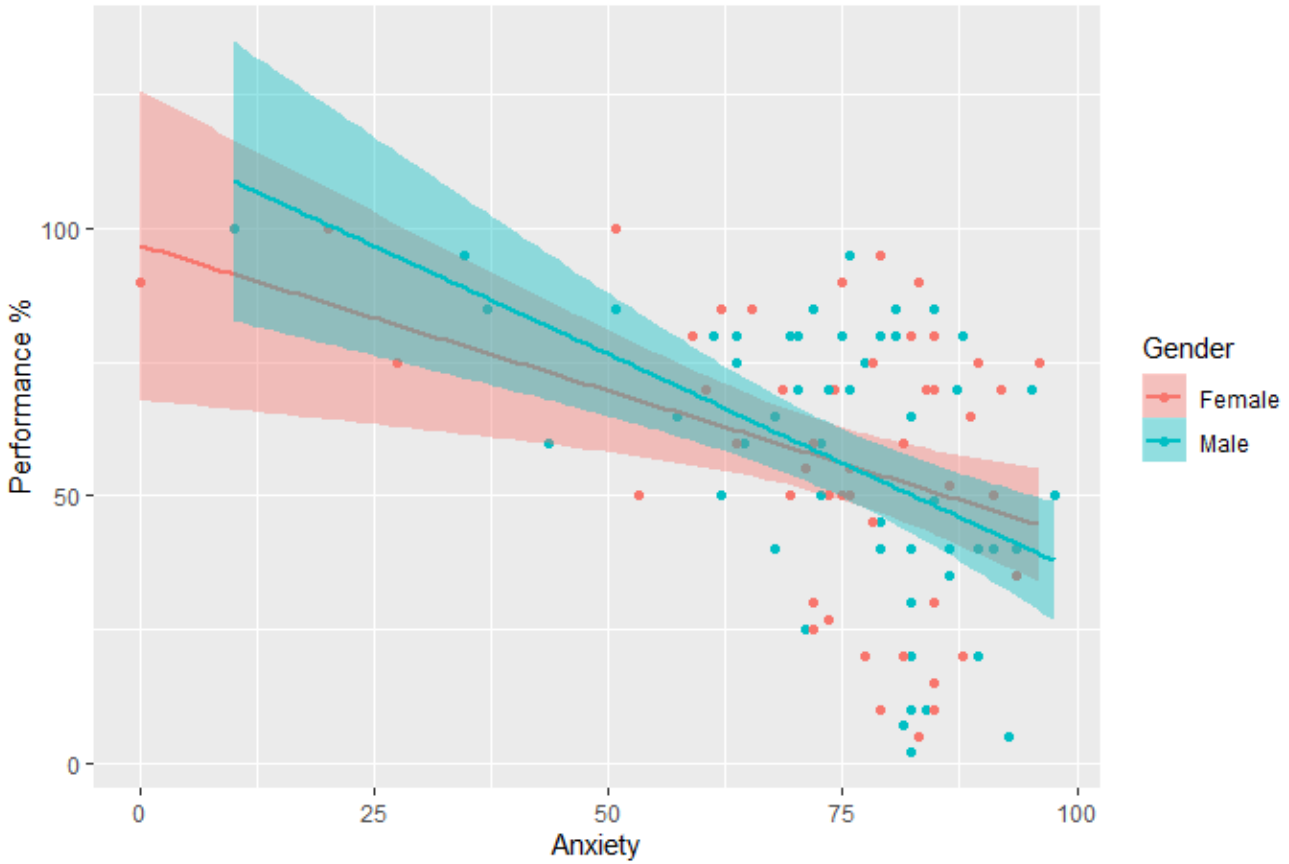
Scatterplot Regressionsgerade



geom_smooth()
Grauer Bereich = 95% Konfidenzintervall

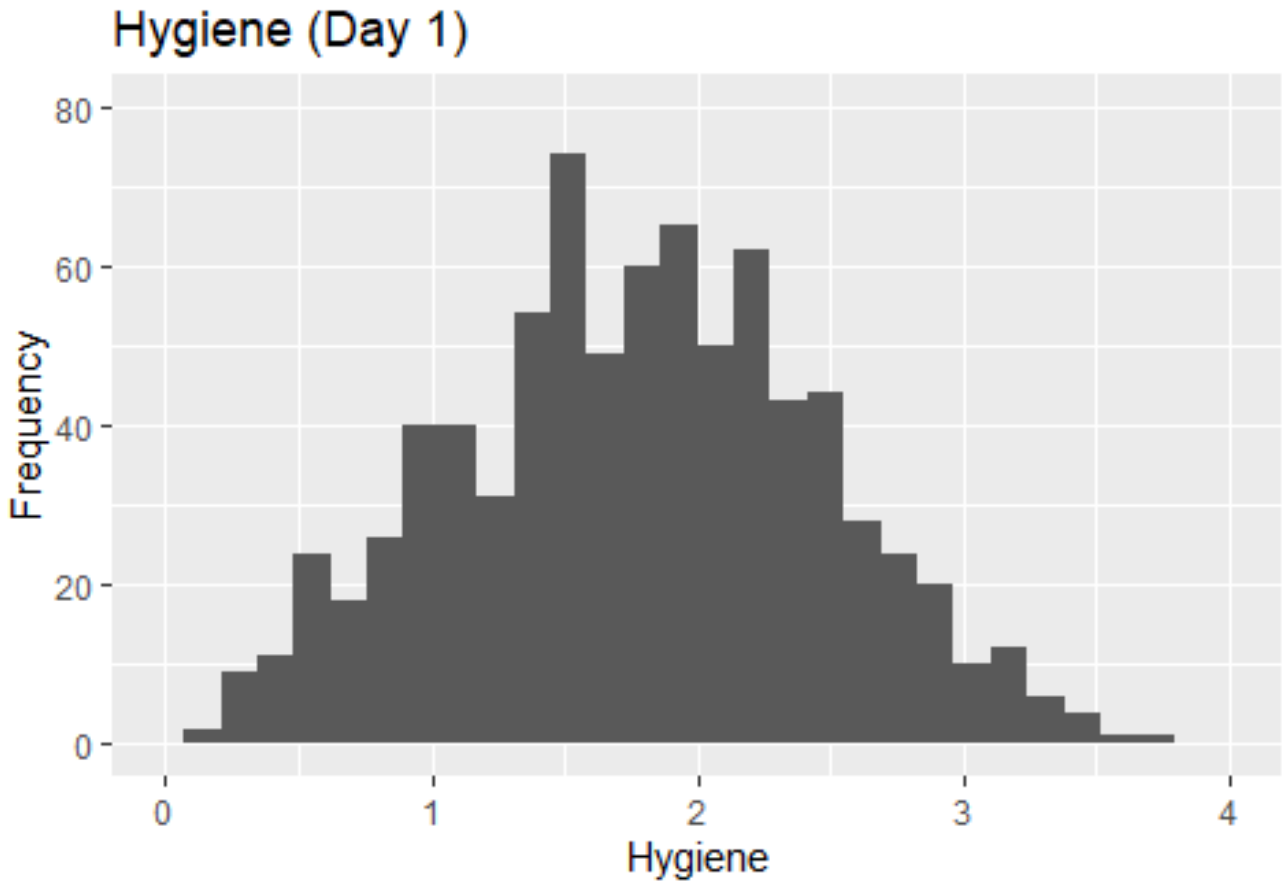
```
data<-read.delim("Exam Anxiety.dat", header=TRUE)
graph<-ggplot(data, aes(Anxiety, Exam))
graph + geom_point() + geom_smooth(method="lm") +
  labs(x = "Exam Anxiety", y = "Exam Performance %")
```

Gruppierter Scatterplot



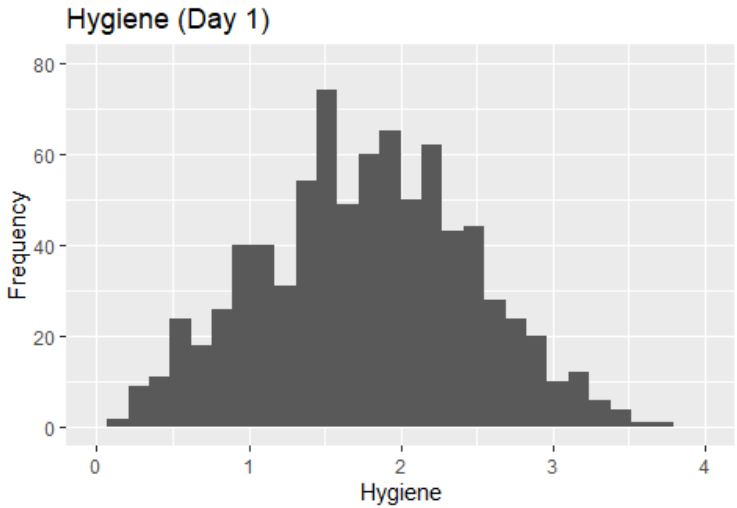

```
data<-read.delim("Exam Anxiety.dat", header=TRUE)
graph<-ggplot(data, aes(Anxiety, Exam, color=Gender))
graph + geom_point() + geom_smooth(method="lm", aes(fill=Gender)) +
  labs(x = "Anxiety", y = "Performance %")
```

Histogramm (Häufigkeitsverteilung)




```
festivaldata<-read.delim("DownloadFestival.dat", header=TRUE)
festivalhistogram<-ggplot(festivaldata, aes(day1))
festivalhistogram + ggtitle ("Hygiene (Day 1)") + xlim(0,4) + geom_histogram() +
  labs(legend.position = "none", x = "Hygiene", y = "Frequency")
```

Histogramm (Häufigkeitsverteilung)

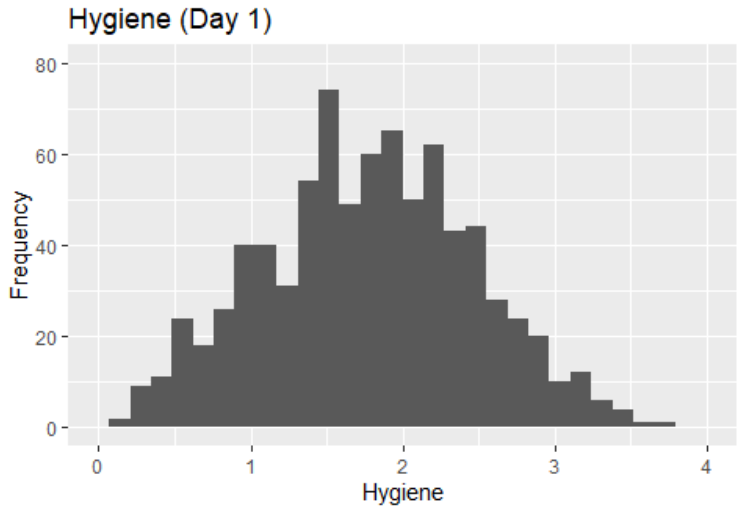


0 = "You smell like a corpse that's been left to rot in a skunk's arse" (Andy Field)

4 = "You smell of sweet roses on a fresh spring day" (Andy Field)

Interpretation:

Histogramm (Häufigkeitsverteilung)



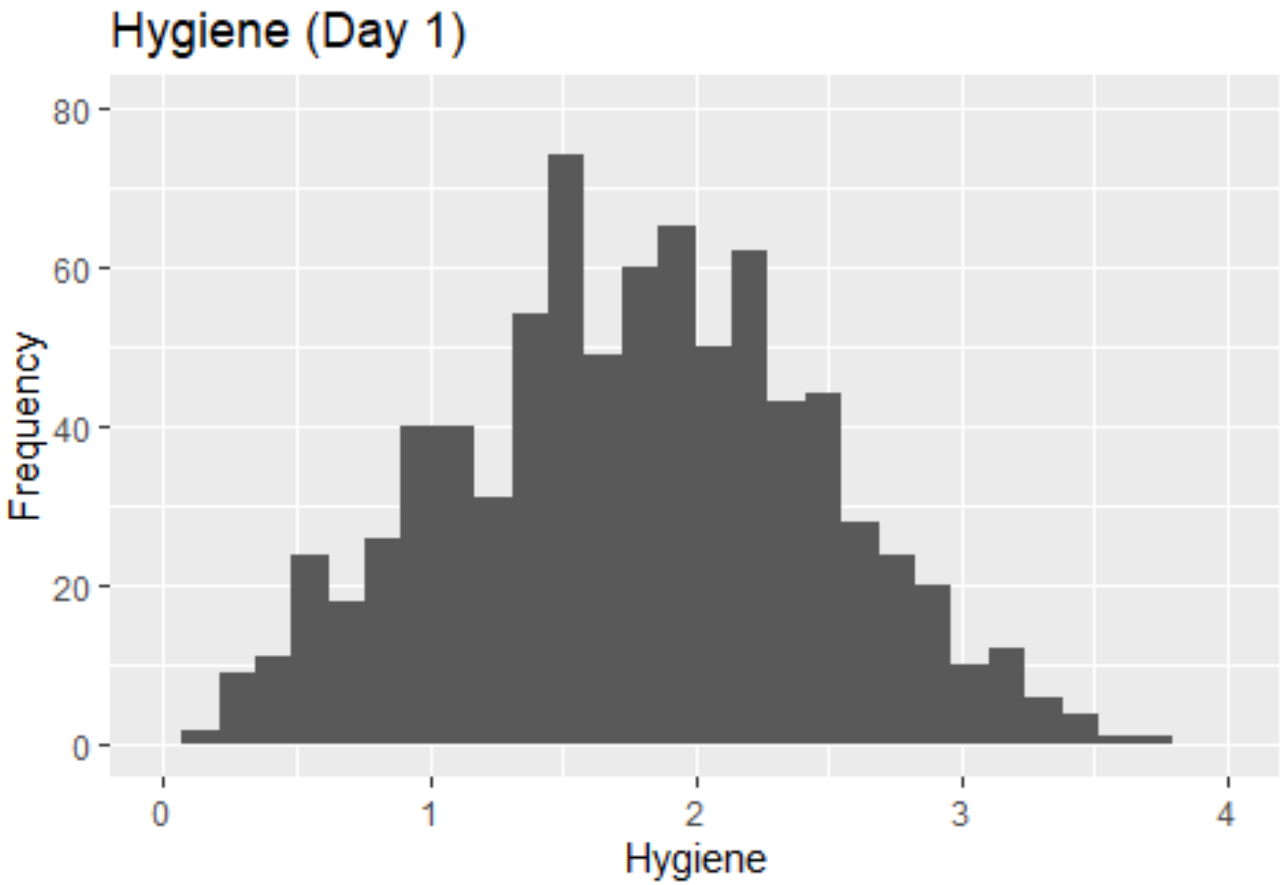
0 = "You smell like a corpse that's been left to rot in a skunk's arse" (Andy Field)

4 = "You smell of sweet roses on a fresh spring day" (Andy Field)

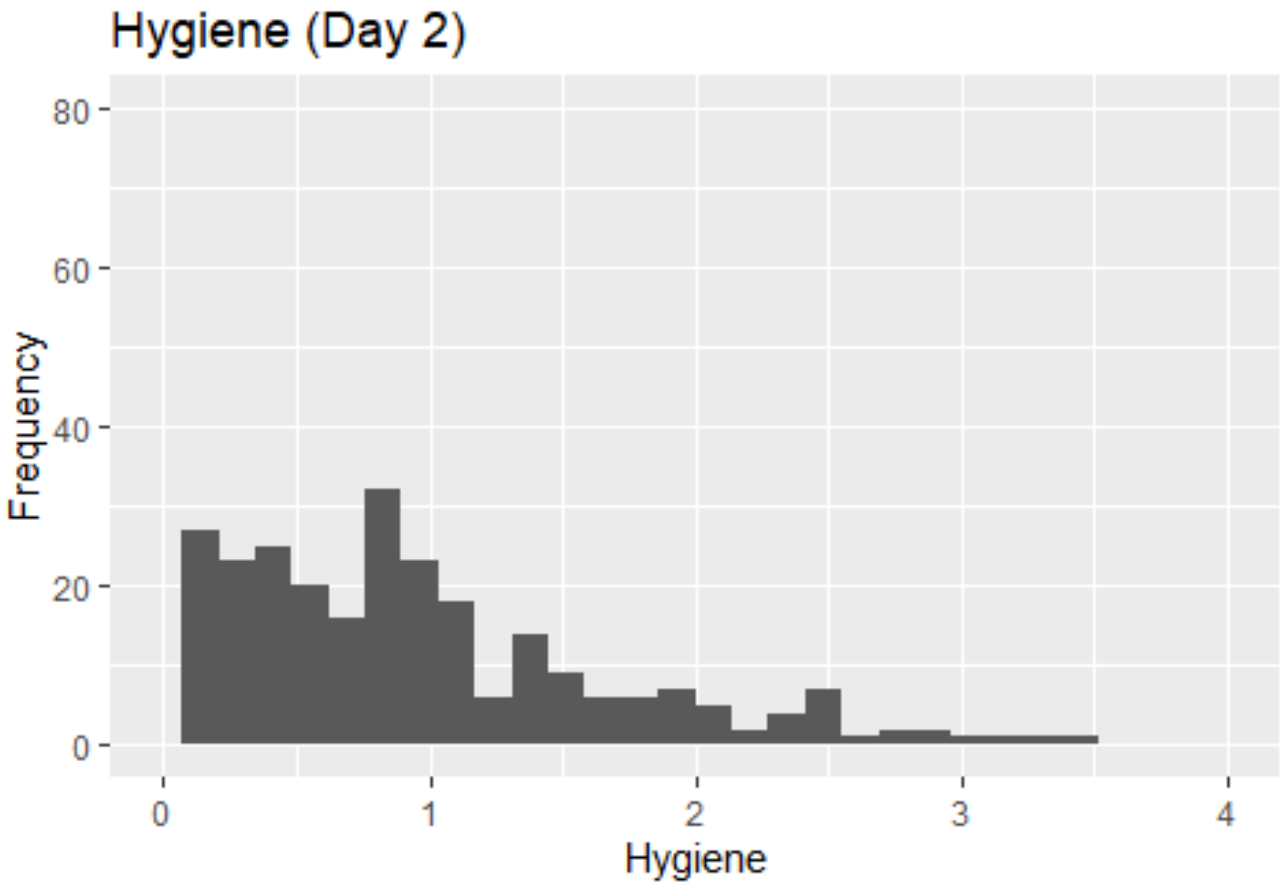
Interpretation:

- Hygiene etwa normalverteilt

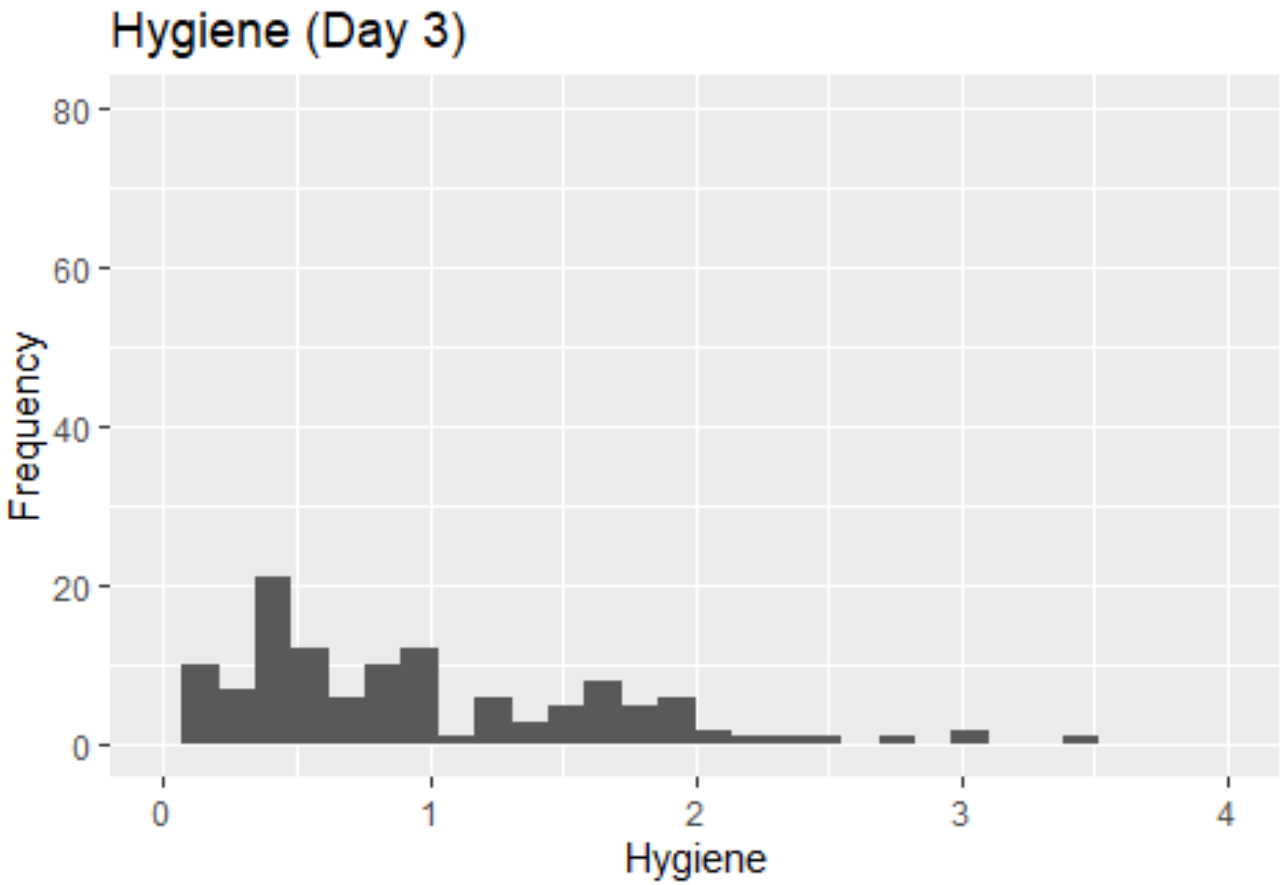
Histogram Festival Tag 1



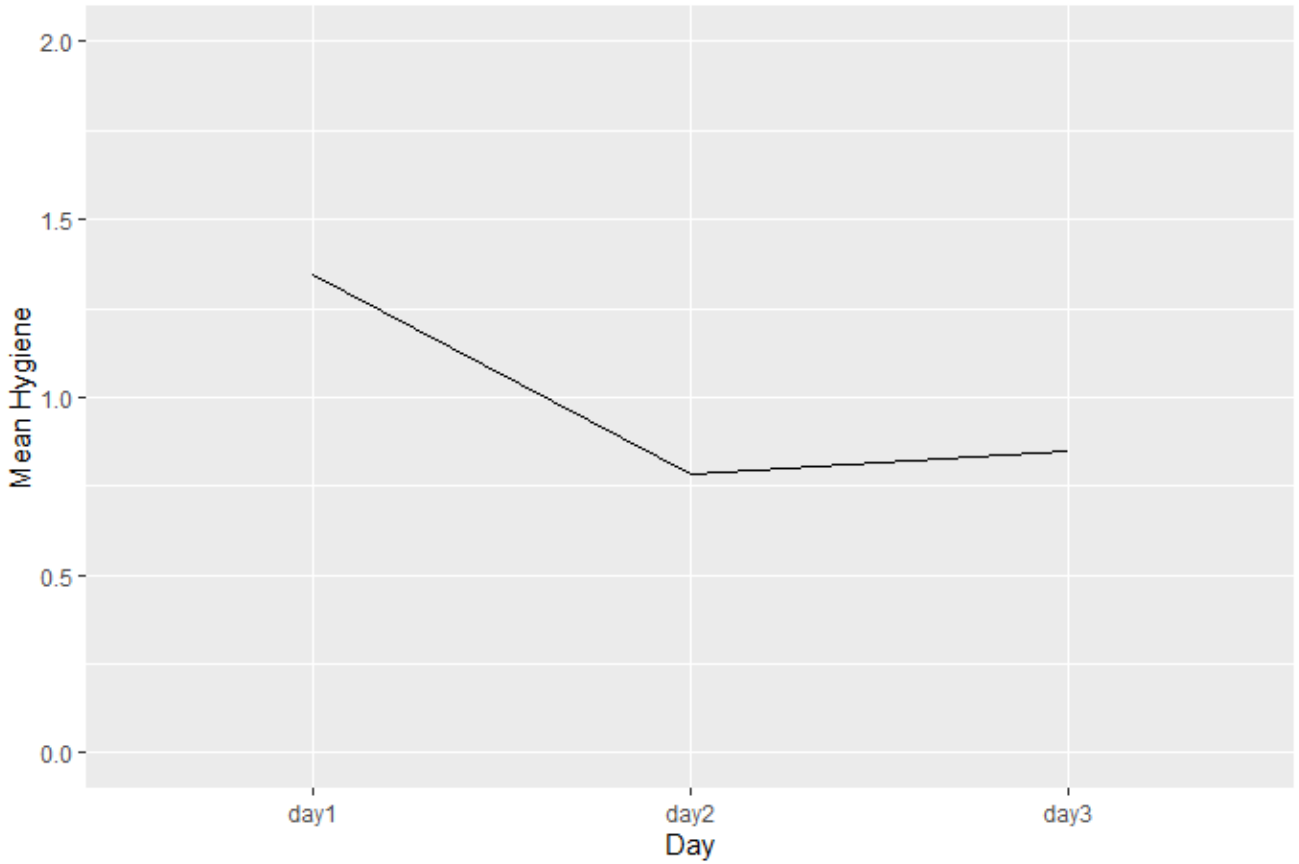
Histogram Festival Tag 2



Histogram Festival Tag 3

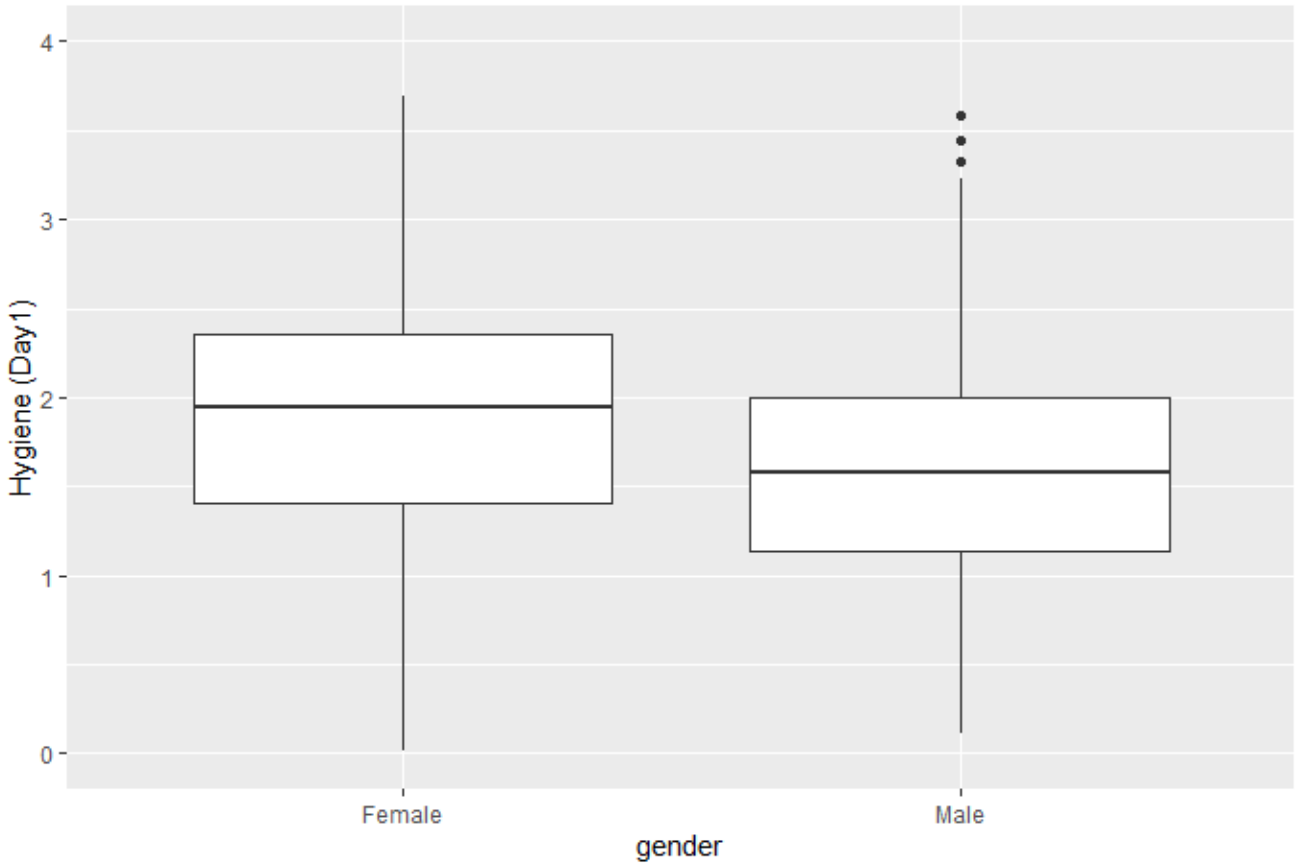


Line Charts




```
festivalstack<-read.delim("festival_stack.dat", header=TRUE)
bar <- ggplot(festivalstack, aes(day, hygiene))
bar + stat_summary(fun.y =mean, geom="line", aes(group=1)) +
  labs(y="Mean Hygiene", x = "Day") + ylim(0,2)
```

Boxplot

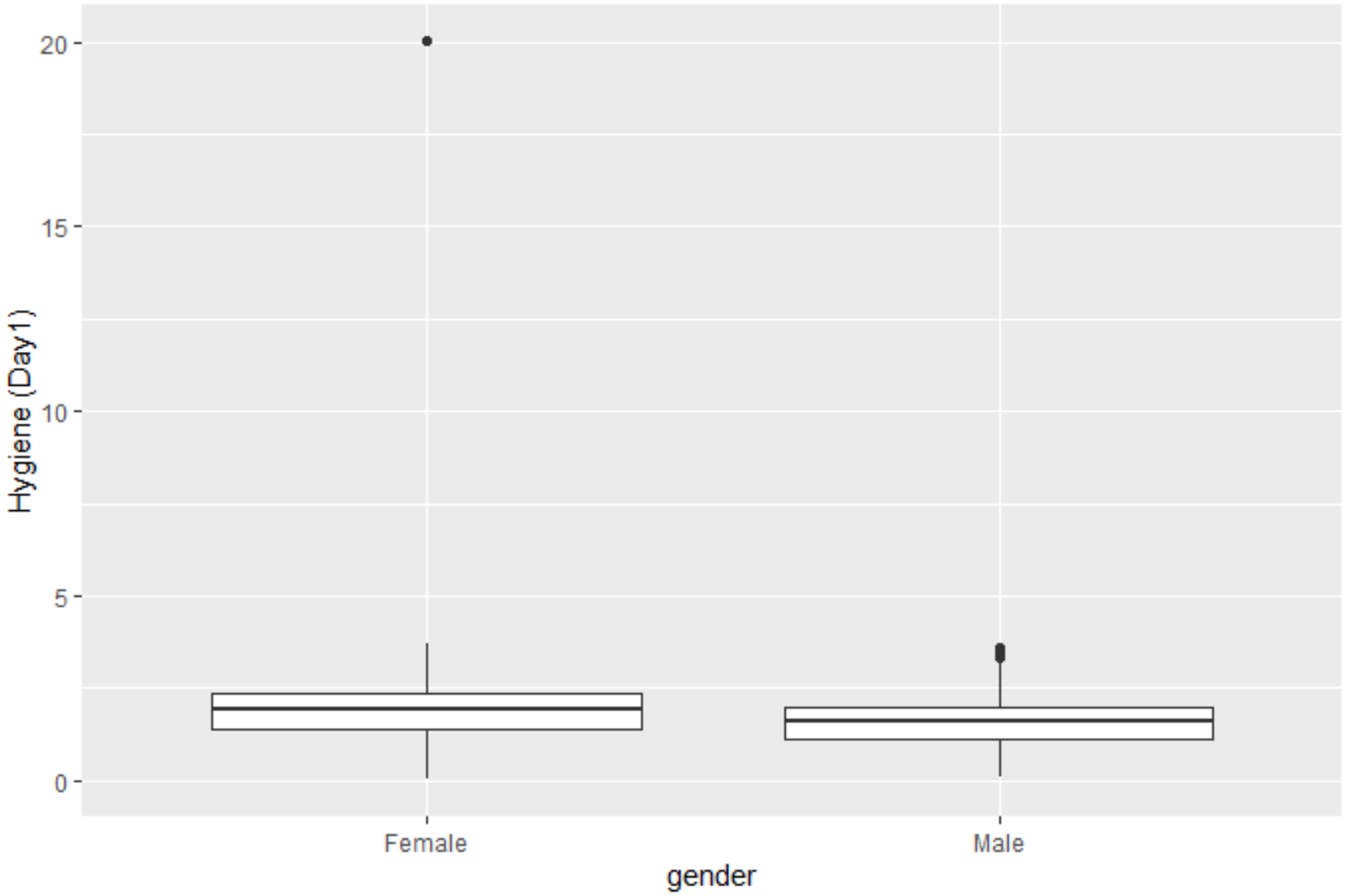



```
festivaldata<-read.delim("DownloadFestival.dat", header=TRUE)
festivalboxplot<-ggplot(festivaldata, aes(gender, day1))
festivalboxplot + geom_boxplot()+labs(y="Hygiene (Day1)") + ylim(0,4)
```

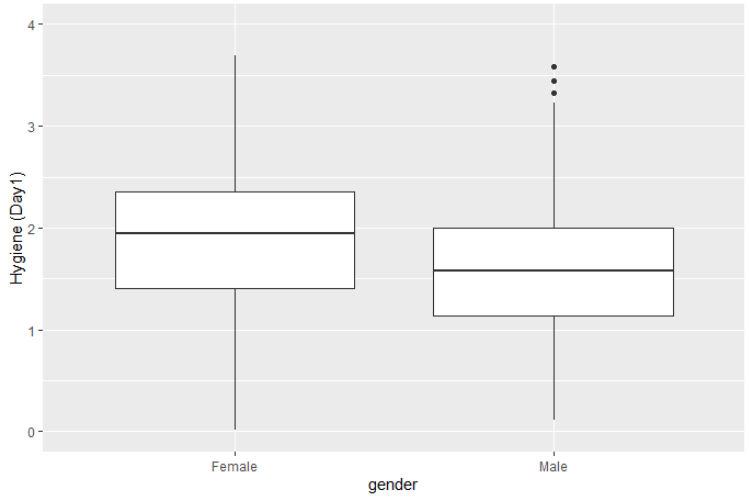
Boxplot

- Box-Whisker-Diagramm
- Mittelpunkt = Median
- Box = Interquartilsabstand (IR) (50% aller Werte um Median)
- Oberer / Unterer Whisker = Maximaler/Minimaler Wert im Bereich $1.5 * IR$ in jede Richtung von der Box ausgehend
- Werte außerhalb der Whisker als Sternchen (Outlier)
- Workaround wenn Gruppierung nicht gewünscht
  aes(x = factor(0), day1), also für einzelnen Boxplot
  - `https://stackoverflow.com/questions/15027659/how-do-you-draw-a-boxplot-without-specifying-x-axis`

Boxplot und Outlier



Boxplot

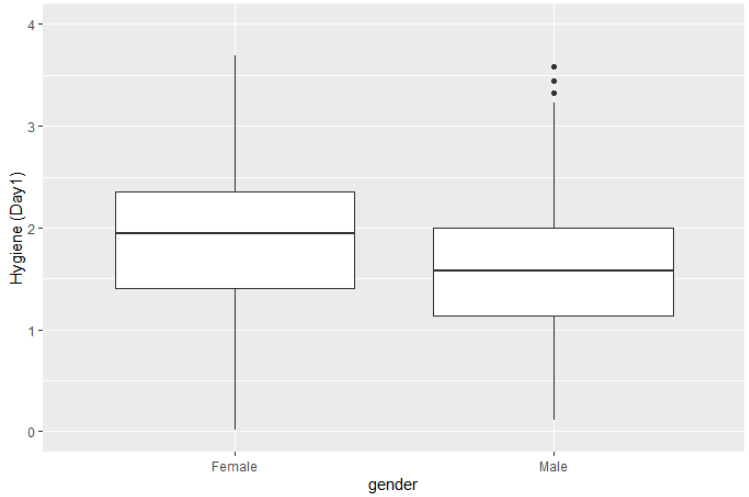


0 = "You smell like a corpse that's been left to rot in a skunk's arse" (Andy Field)

4 = "You smell of sweet roses on a fresh spring day" (Andy Field)

Interpretation:

Boxplot



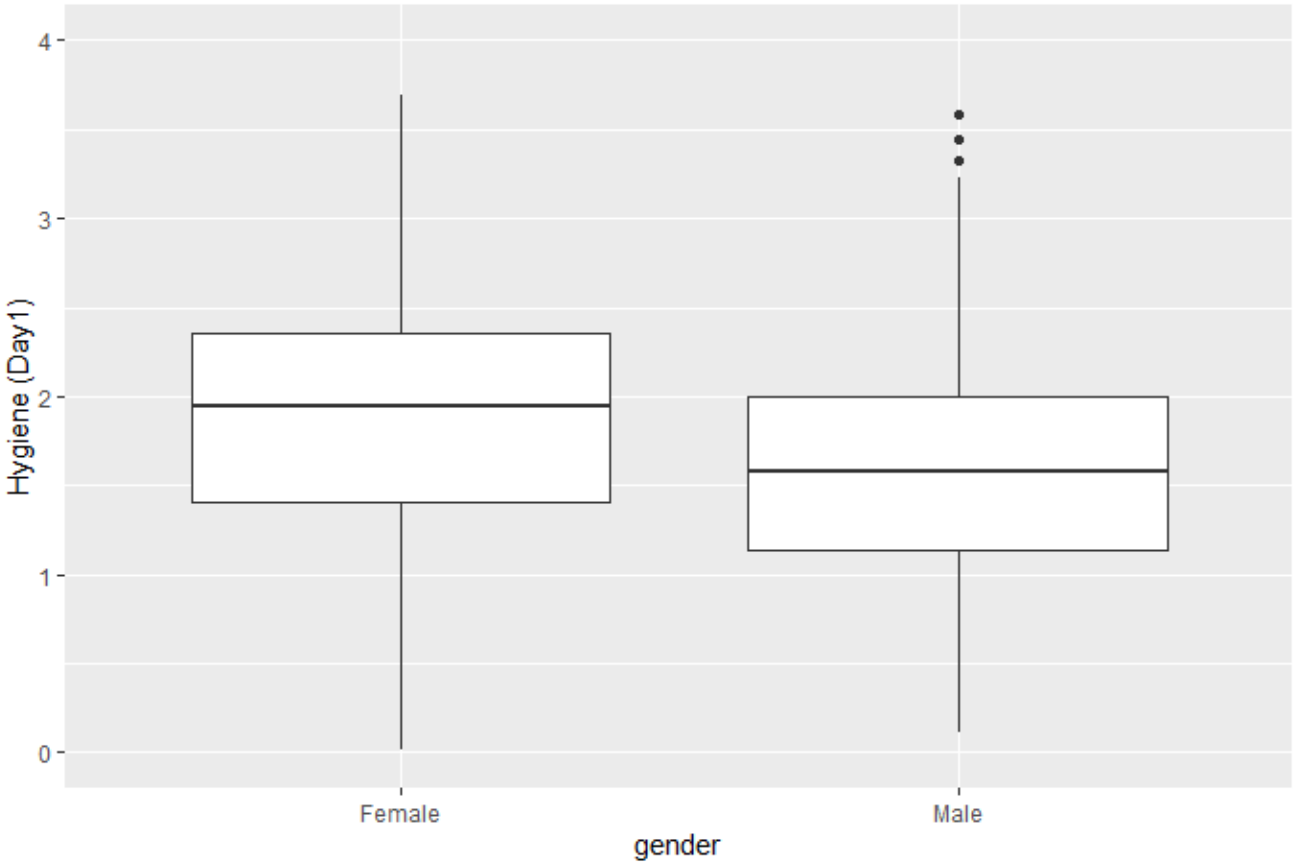
0 = "You smell like a corpse that's been left to rot in a skunk's arse" (Andy Field)

4 = "You smell of sweet roses on a fresh spring day" (Andy Field)

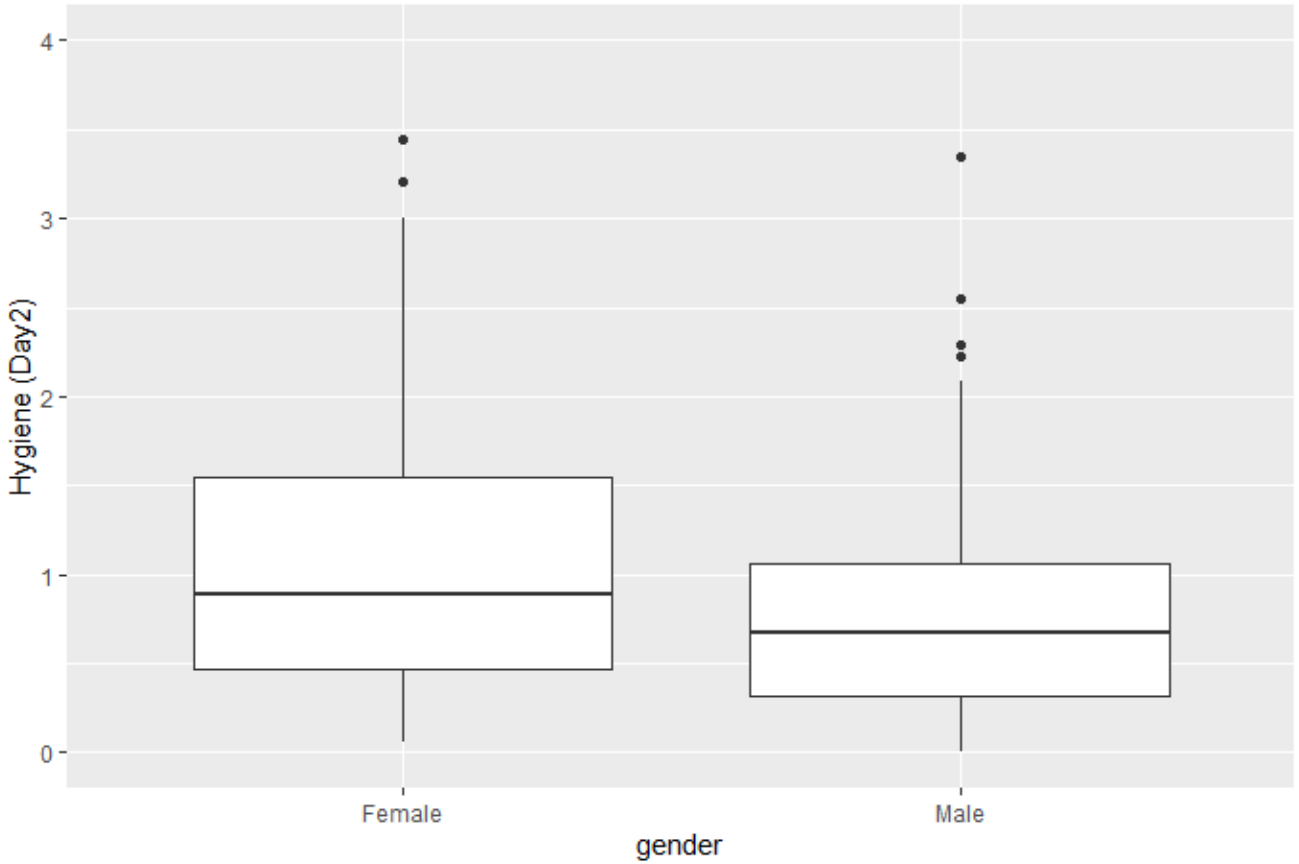
Interpretation:

- Corpsigkeit etwa gleich
- Blumigkeit bei Females höher
- Blumigste Females entsprechen Outliern bei Males
- IR etwa gleich groß, Verteilung pro Gruppe ähnlich
- Spannweite bei Females größer
- Box bei Females höher, insgesamt Females also blumiger

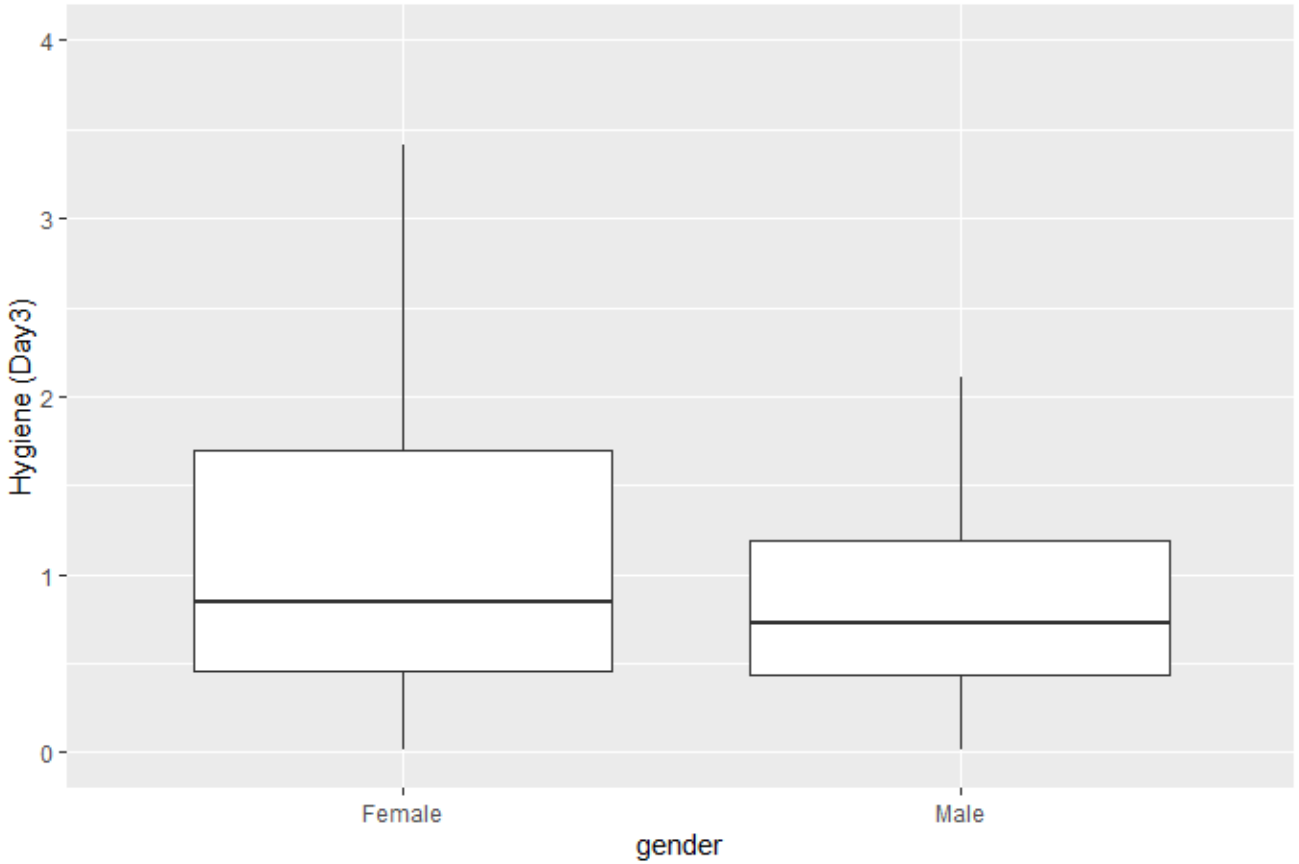
Boxplot Festival Tag 1



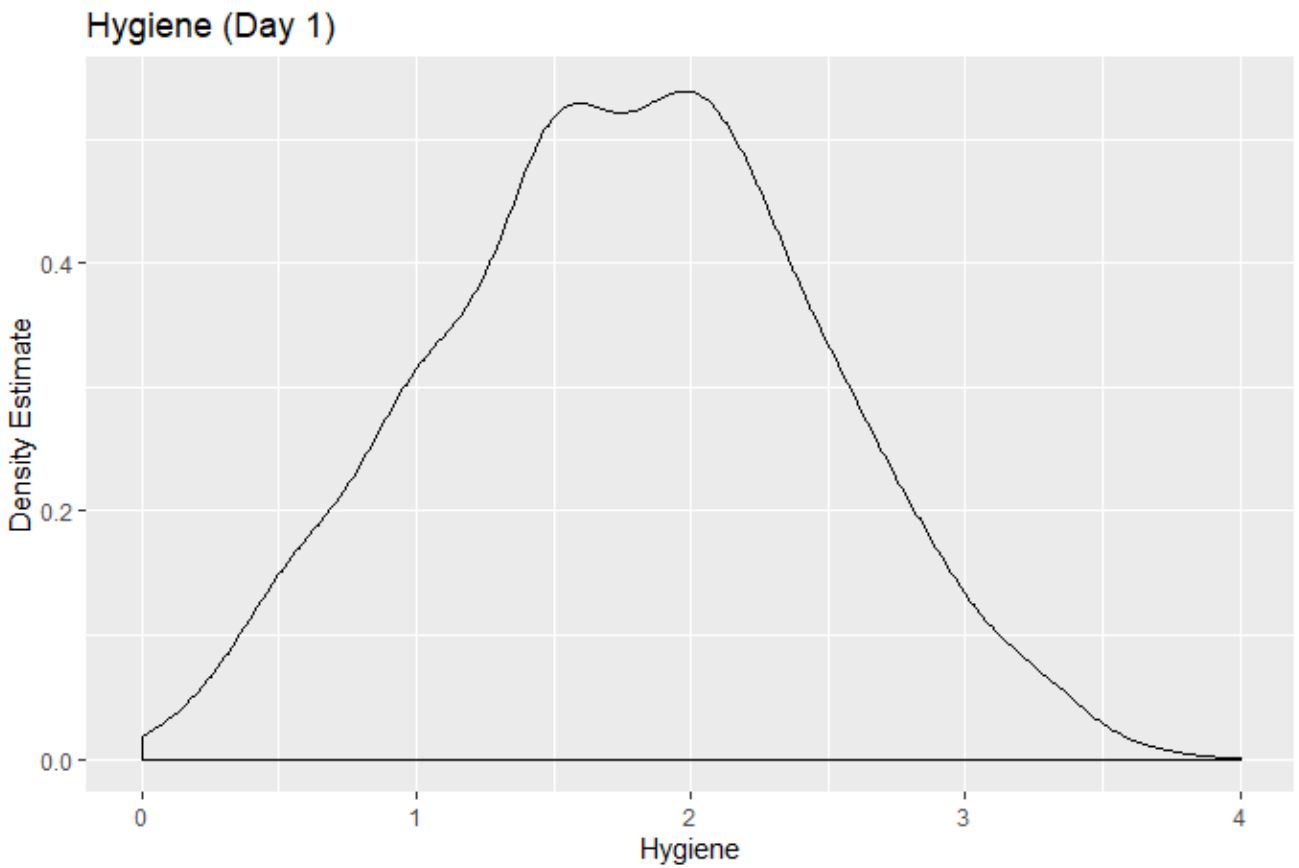
Boxplot Festival Tag 2



Boxplot Festival Tag 3



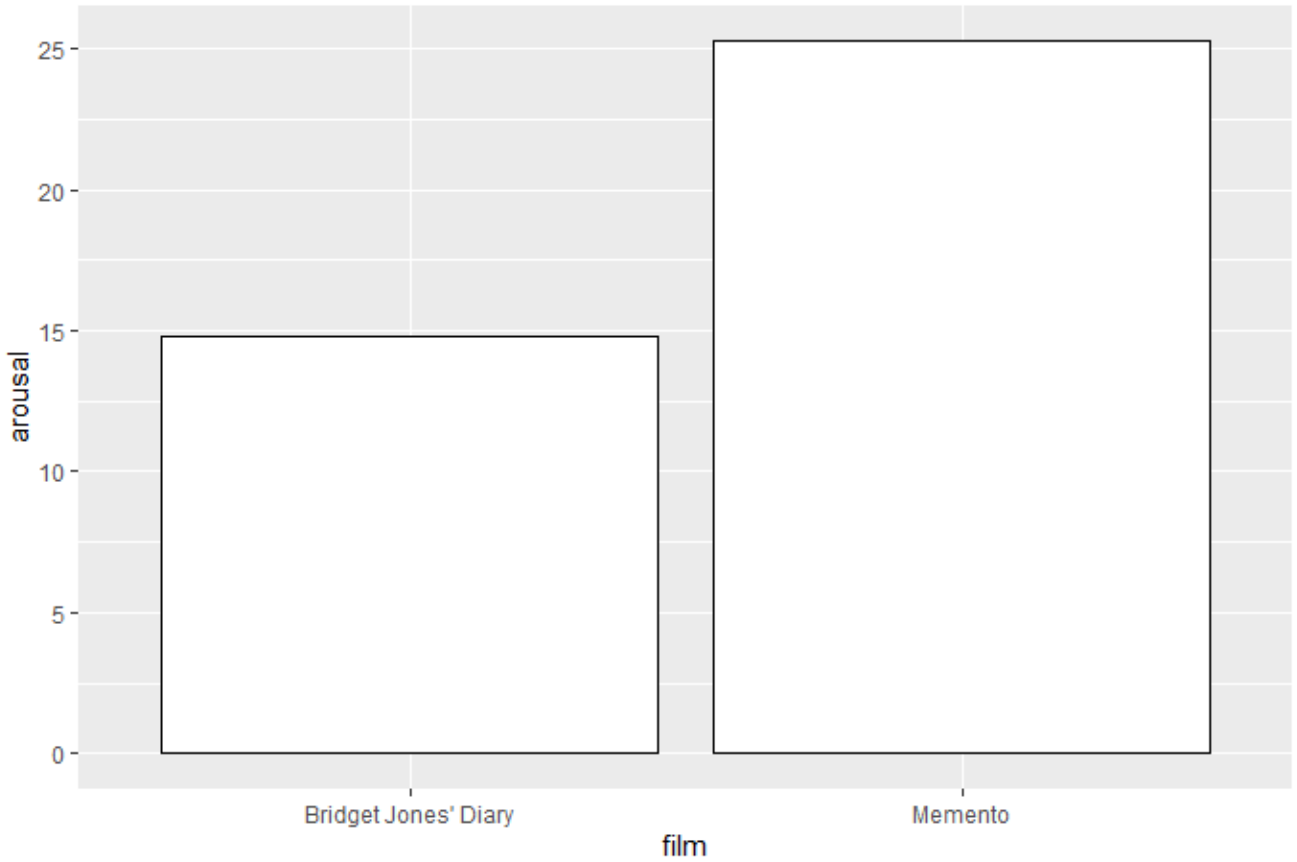
Density Plot



"Smoothe" Häufigkeitsverteilung

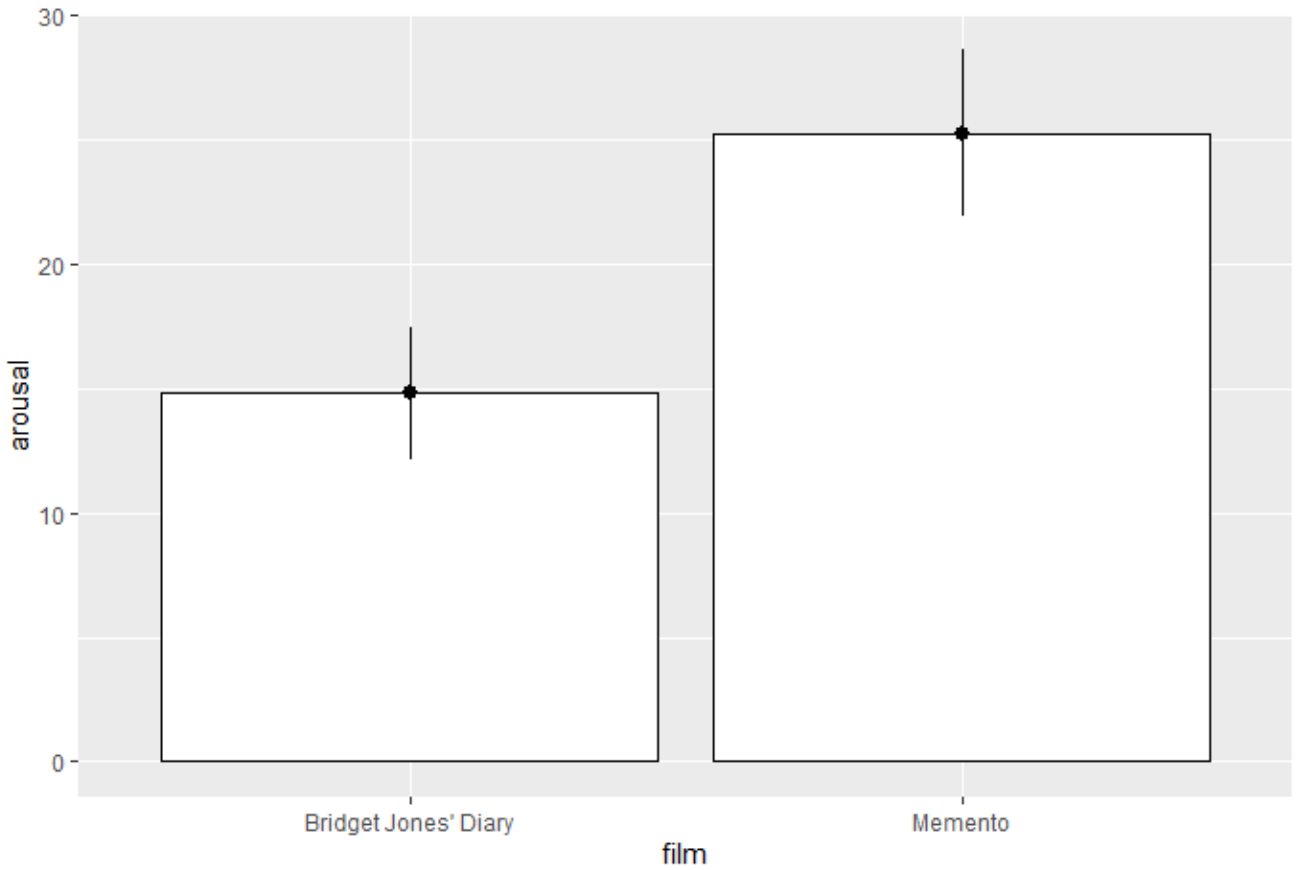
```
festivaldata<-read.delim("DownloadFestival.dat", header=TRUE)
festivalhistogram<-ggplot(festivaldata, aes(day1))
festivalhistogram + ggtitle ("Hygiene (Day 1)") + xlim(0,4) + geom_density() +
  labs(legend.position = "none", x = "Hygiene", y = "Density Estimate")
```

Bar Charts / 1 Unabhängige Variable



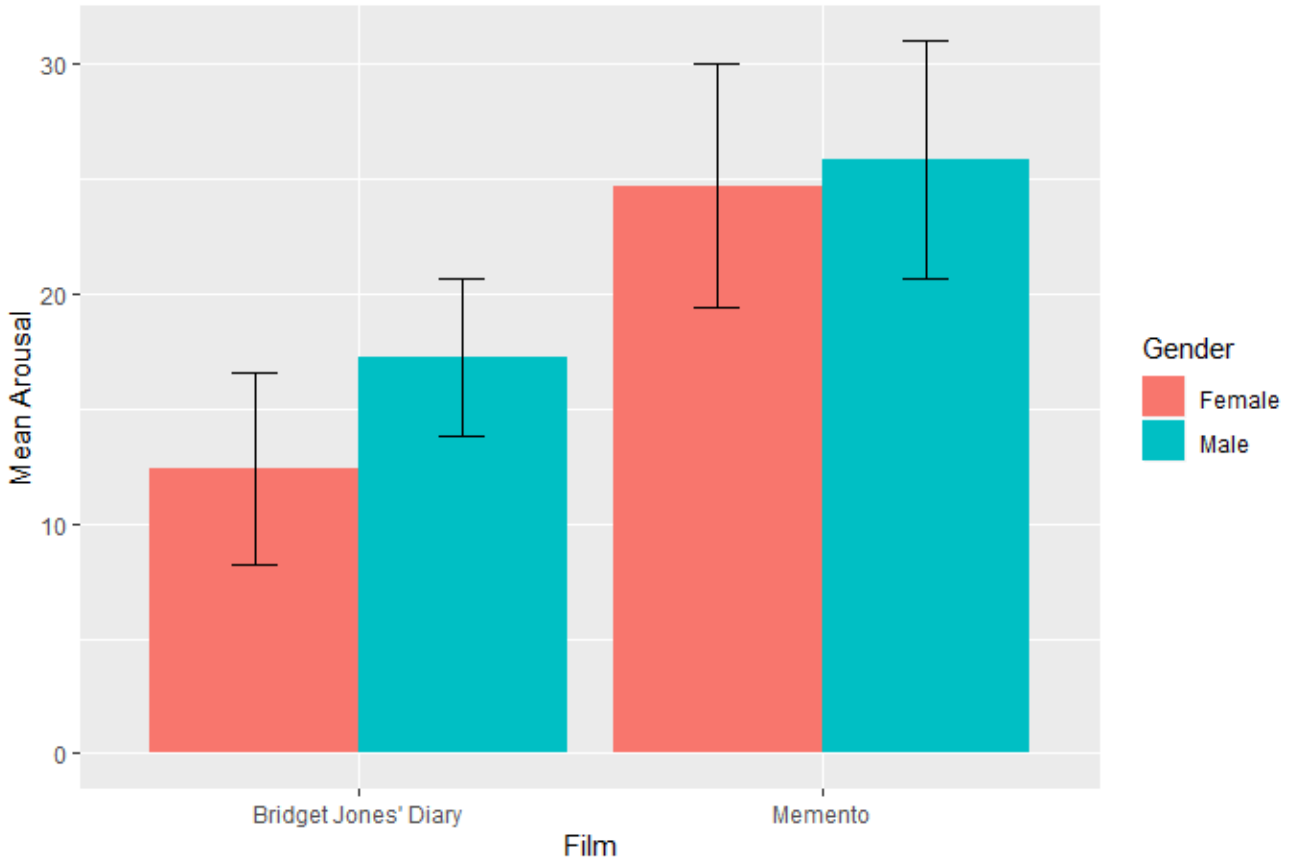

```
chickFlickdata<-read.delim("ChickFlick.dat", header=TRUE)
bar <- ggplot(chickFlickdata, aes(film,arousal))
bar + stat_summary(fun.y =mean, geom="bar", fill="White", color="Black")
```

Bar Chart mit Konfidenzintervall (95%)



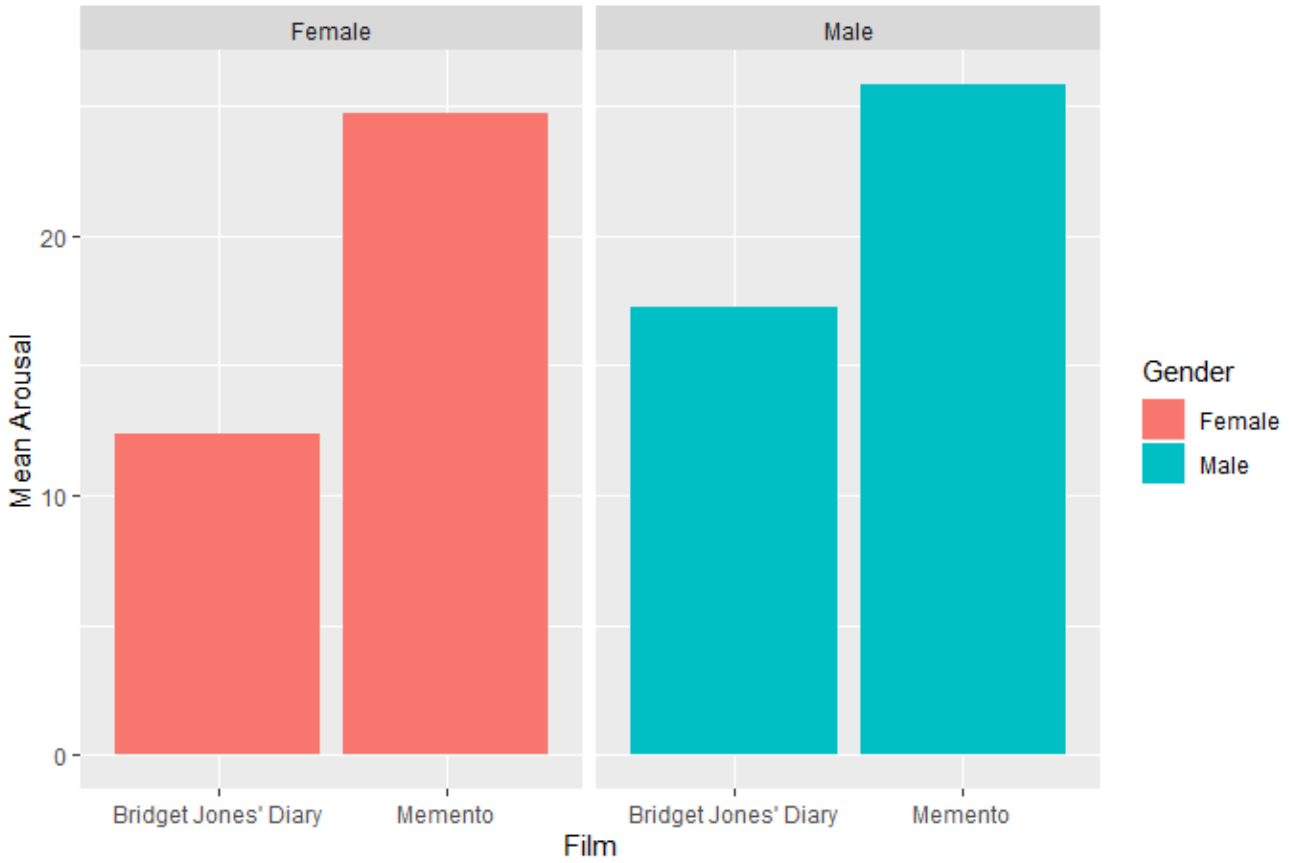

```
chickFlickdata<-read.delim("ChickFlick.dat", header=TRUE)
bar <- ggplot(chickFlickdata, aes(film,arousal))
bar + stat_summary(fun.y =mean, geom="bar", fill="White", color="Black") +
  stat_summary(fun.data = mean_cl_normal, geom="pointrange")
```

Mehrere Unabhängige Variablen + Error Bar



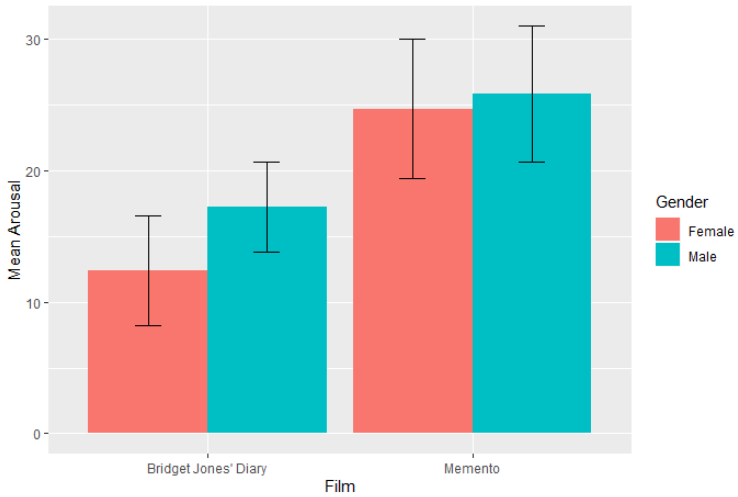

```
chickFlickdata<-read.delim("ChickFlick.dat", header=TRUE)
bar <- ggplot(chickFlickdata, aes(film,arousal, fill=gender))
bar + stat_summary(fun.y =mean, geom="bar", position="dodge") +
  stat_summary(fun.data = mean_cl_normal, geom="errorbar",
  position = position_dodge(width=0.90), width=0.2) +
  labs (x = "Film", y = "Mean Arousal", fill="Gender")
```

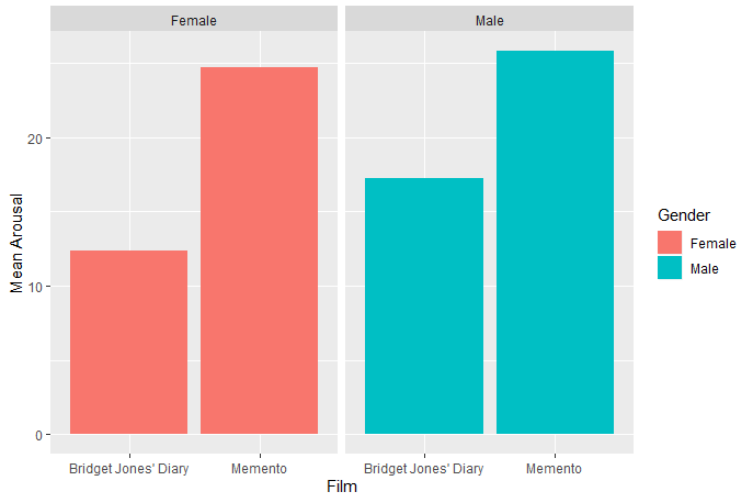
Mehrere Unabhängige Variablen + Facette




```
chickFlickdata<-read.delim("ChickFlick.dat", header=TRUE)
bar <- ggplot(chickFlickdata, aes(film,arousal, fill=gender))
bar + stat_summary(fun.y = mean, geom = "bar") +
  facet_wrap (~ gender) +
  labs (x = "Film", y = "Mean Arousal", fill="Gender")
```

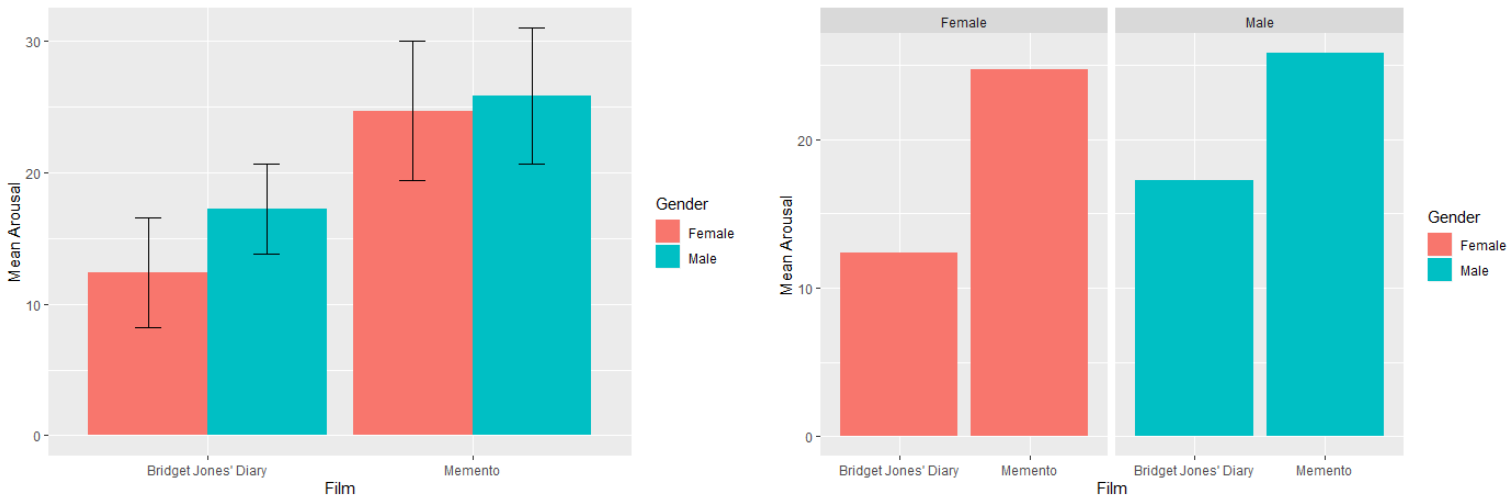
Bar Charts





Interpretation:

Bar Charts



Interpretation:

- Arousal pro Film zwischen Females und Males etwa gleich
- Bridget Jones finden beide langweiliger als Memento

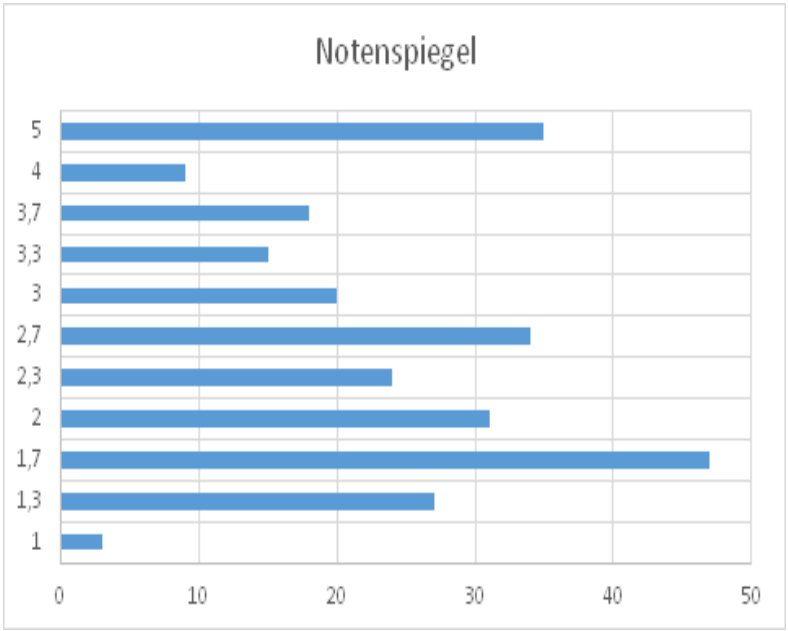
Überblick

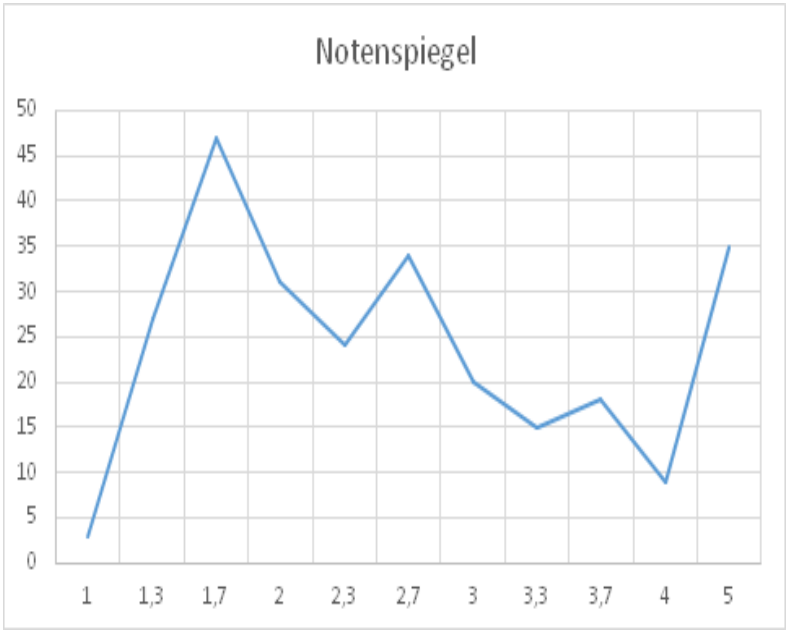
1. Grundlagen statistischer Visualisierung
2. Typische Diagrammarten
   - Scatterplot
   - Histogramm
   - Line Charts
   - Boxplot
   - Density Plot
   - Bar Charts
     - 1 Unabhängige Variable
     - Bar Chart mit Konfidenzintervall
     - Mehrere Unabhängige Variablen
     - Facettierung
3. Lies, damned Lies and Statistics

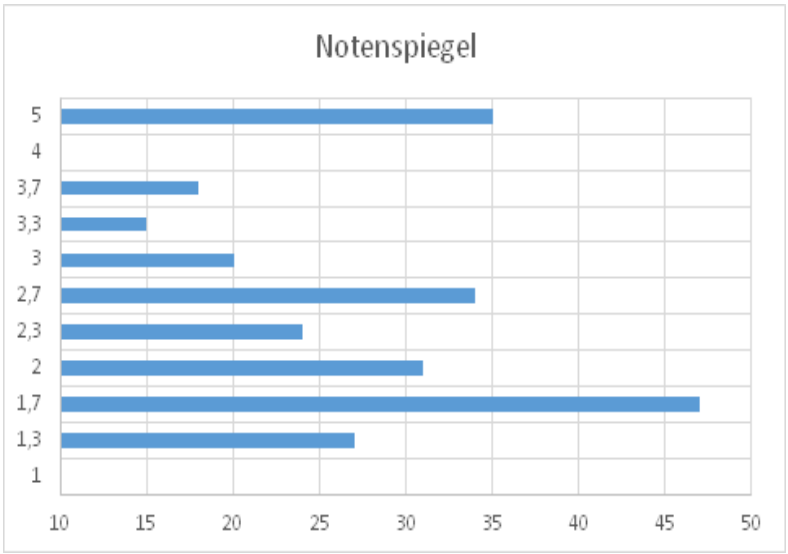
Lies, damned Lies and Statistics

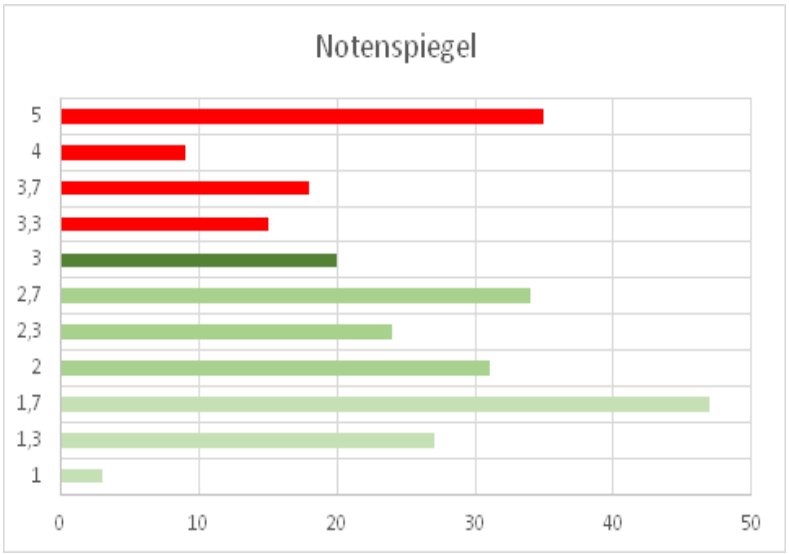
- Gefundene Beispiele gerne an mich schicken
- p-Hacking: Daten anpassen um p hochzutreiben
- Kummulation
- Relative vs. Absolute Zahlen
- Too Much Information (TMI)
- Diagramme & Skalierung
  - Fehlender Nullpunkt
  - Logarithmische vs. Lineare Skalierung
  - Suggerierte Zusammenhänge zwischen Nominaldaten (bspw durch Reihenfolge oder Linien)
  - Zoom
- Suggestive Farbcodierung
  - Künstliche irreführende Abgrenzung durch scharfe Farbtongrenzen
  - kulturelle Interpretation von Farben (rot = negativ)
  - http://colorbrewer2.org/ kann helfen bei Farbwahl

Irreführende Darstellung (links oben ist neutral)

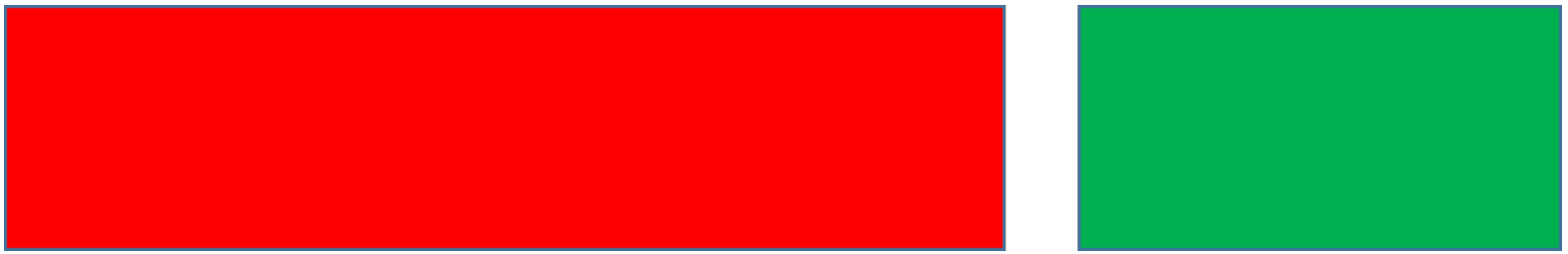






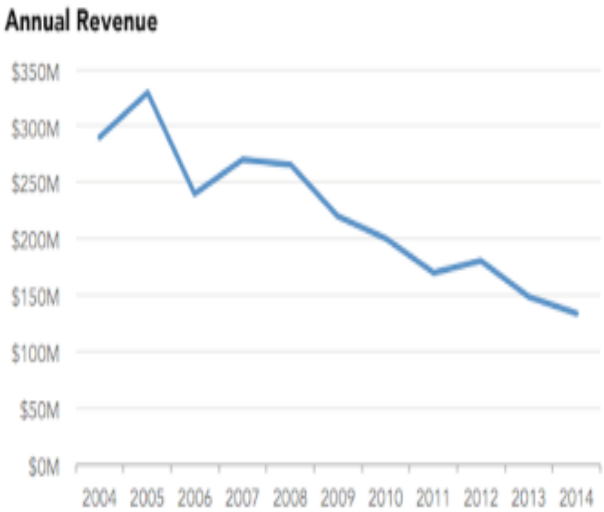


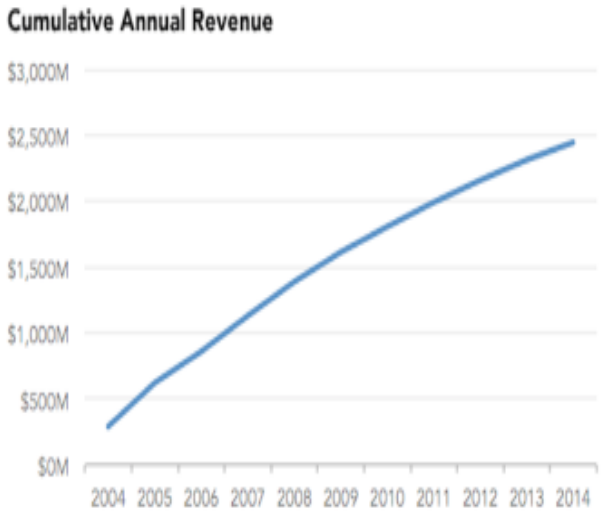
Suggestive Farbcodierung

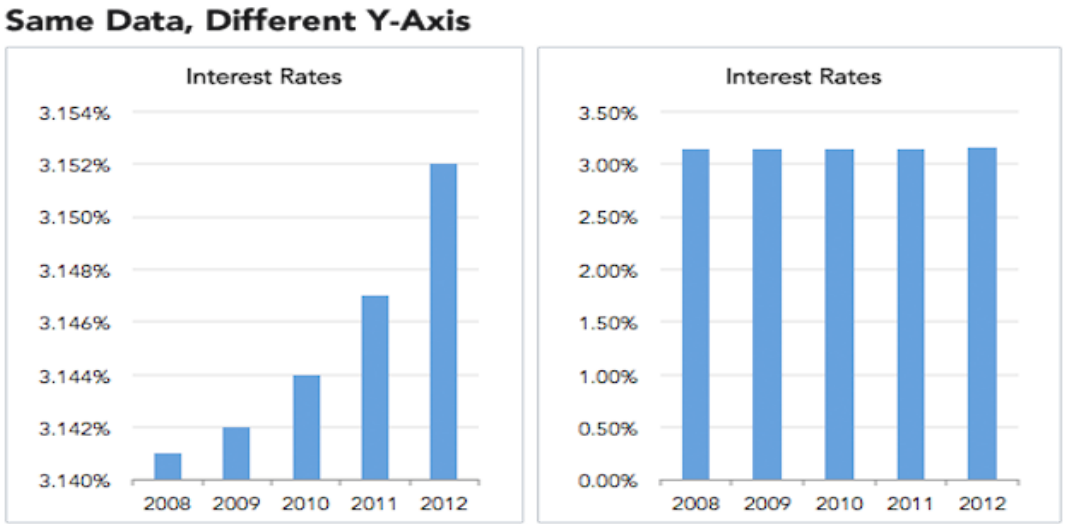


Suggestive Farbcodierung

Irreführende Darstellungen



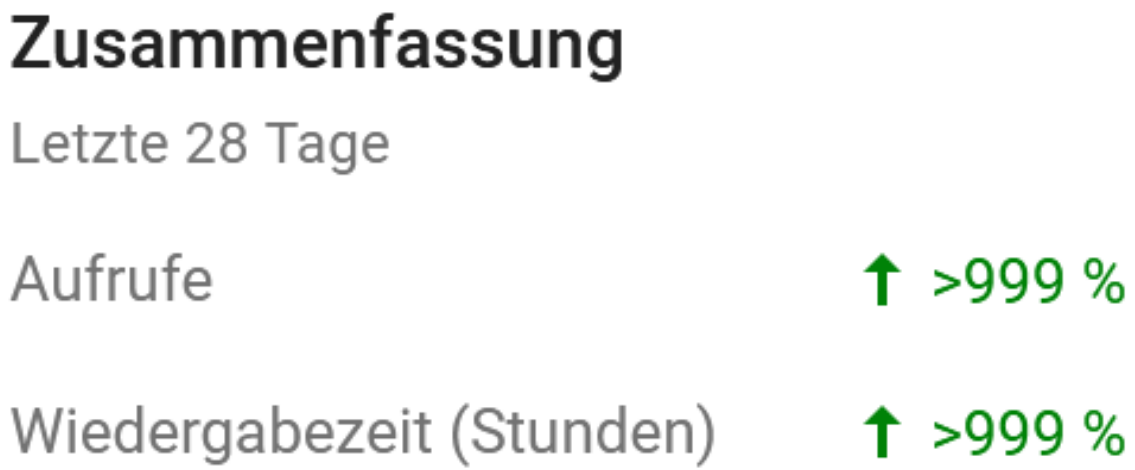




https://www.mapbusinessonline.com/Whitepaper.aspx/Avoid-Data-Visualization-Misinformation

Relative vs Absolute Zahlen

Youtube Kanal der Pandemie-Vorlesung

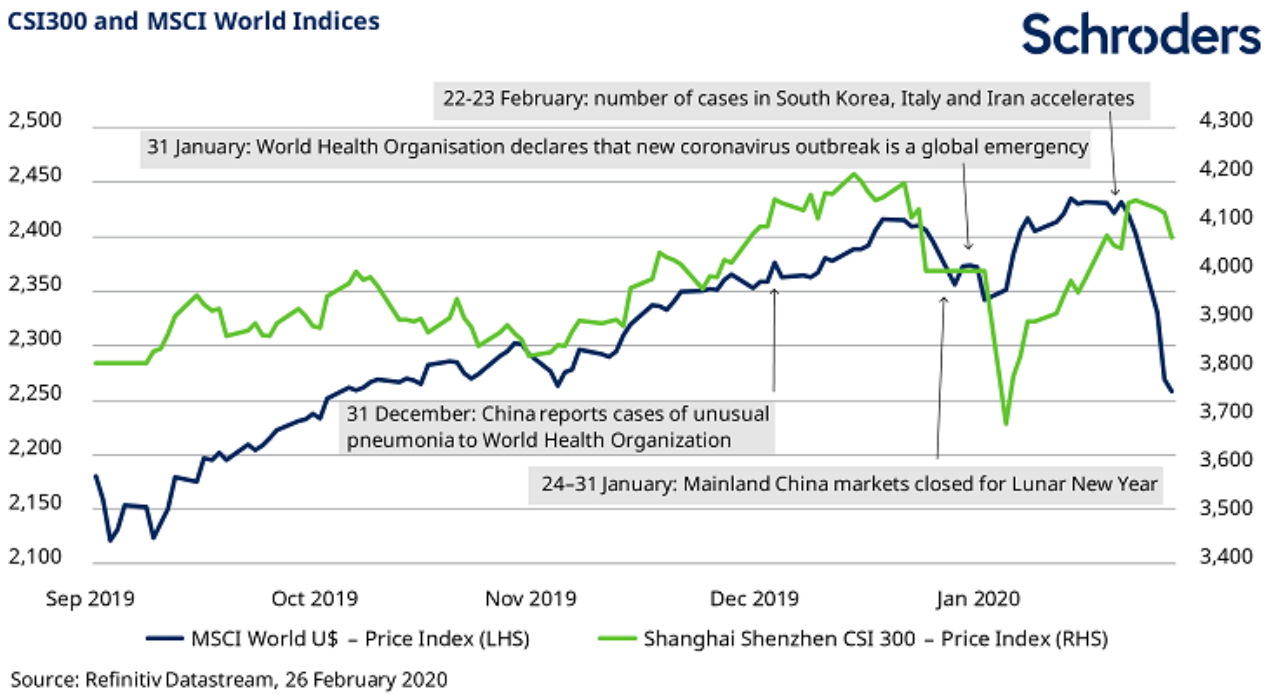


Relative vs. Absolute Zahlen

Youtube Kanal der Pandemie-Vorlesung

- Im Zweifel lieber mehr Diagramme als unklar zu interpretierende Vermischungen

Beispiel 2 verschiedene Skalierungen in 1 Diagramm:

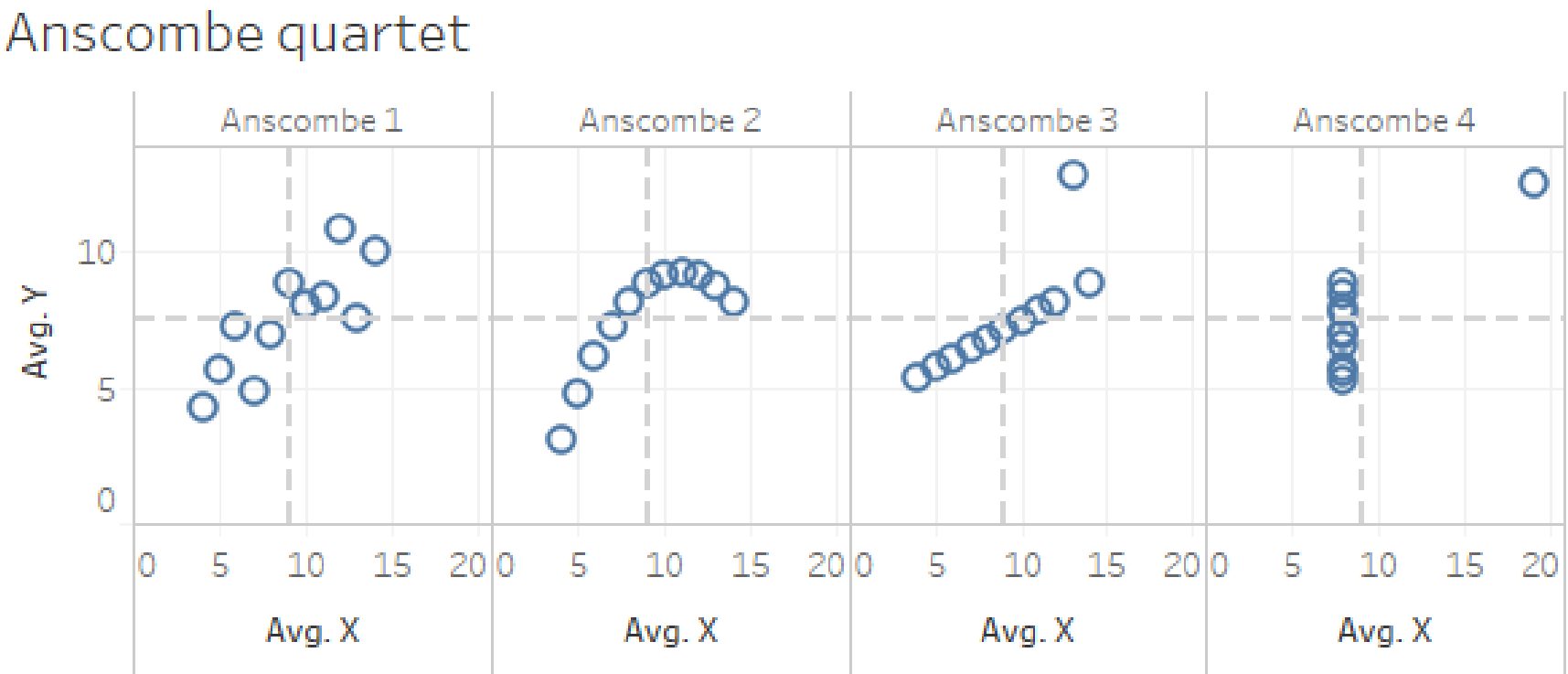


https://www.schroders.com/de/de/institutionelle/insights/maerkte/coronavirus-die-folgen-fuer-die-maerkte-in-sieben-charts/

Fox News

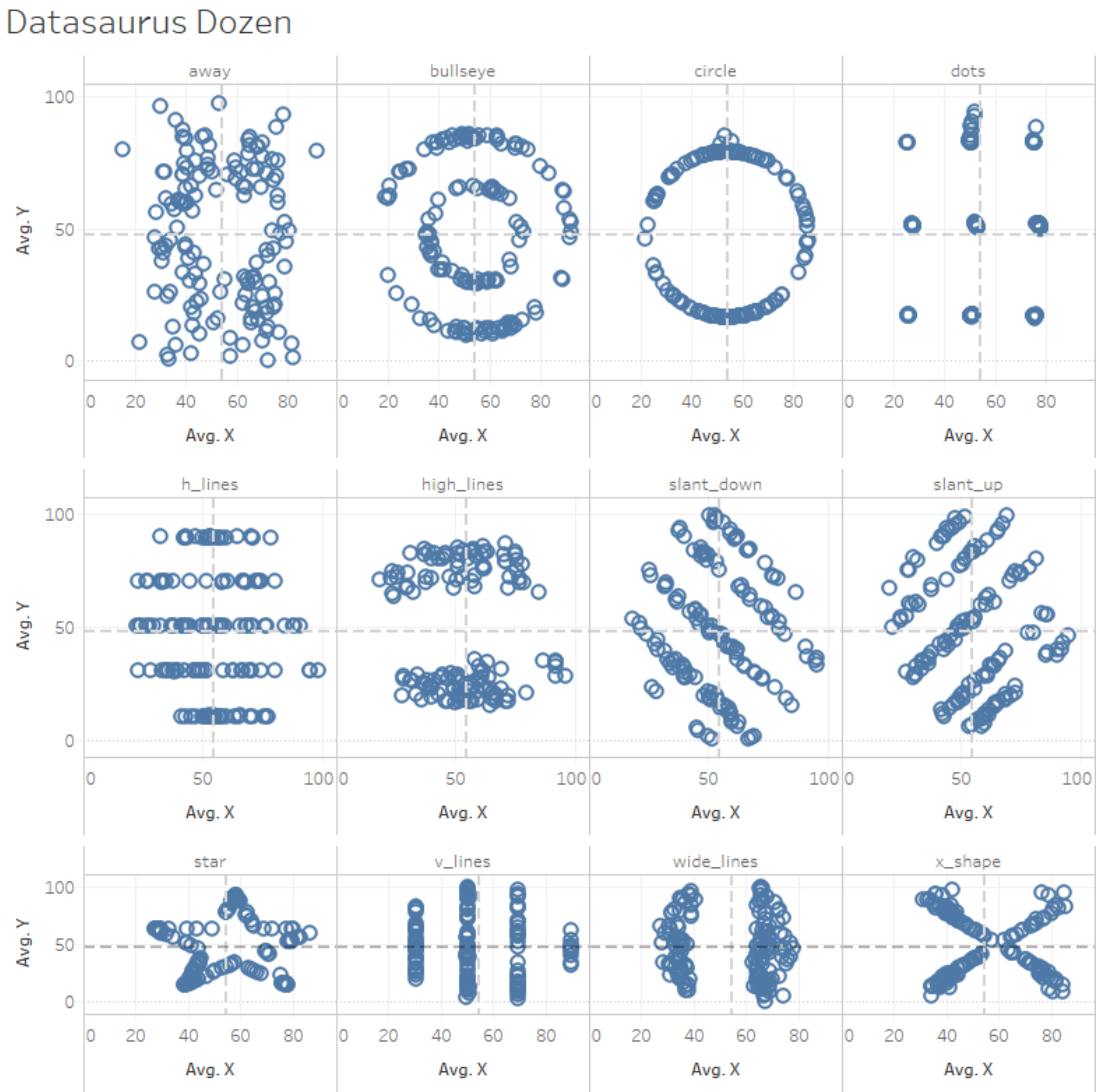
https://twitter.com/Carnage4Life/status/1246579721585868800/photo/1

Anscombes Quartett



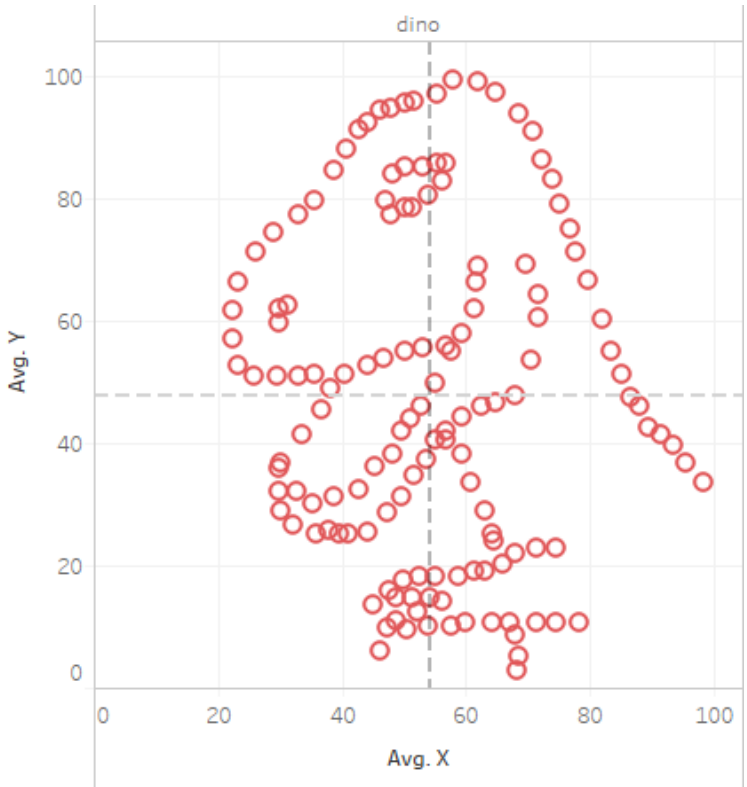
Anscombe, F. J. (1973). Graphs in Statistical Analysis. American Statistician. 27 (1): 17–21. doi:10.1080/00031305.1973.10478966. JSTOR 2682899.

Datasaurus Dozen



Justin Matejka and George Fitzmaurice (1973). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.

Datasaurus Dozen



Justin Matejka and George Fitzmaurice (1973). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.

Zusammenfassung

- Grundlagen statistischer Visualisierung
  - Verzerrung und Chartjunk vermeiden, Minimum Ink, Aussagekraft, ...
- Verschiedene Diagramme
  - Scatterplots, Histogram, Line Charts, Boxplots, Density Plots, Bar Charts, ...
- Irreführung oder fehlerhafte Darstellung erkennen
  - p-Hacking, Kummulation, Diagramme & Skalierung, Suggestive Farbcodierung, TMI, ...

Aktuelle Beispiele

- Ted Underwood (2016): *The Life Cycles of Genres*
  - Untersuchung zu Stabilität und Trennschärfe von Genrebezeichnungen, konkret *Detective Story*, *Science Fiction* und *Gothic*
- Ted Underwood, David Bamman, and Sabrina Lee (2018): *The Transformation of Gender in English-Language Fiction*
  - Vergleich von Genderzuordnung und -anteil zwischen 18. und 20. Jhd bezogen auf Rollen / Verhalten der Charaktere, Sprache und Autorenschaft
  - Stabilität Trennschärfe des Begriffs *gender* (Klassifikation, Vorhersage auf Basis der Sprache)
- RWI–Leibniz-Institut für Wirtschaftsforschung: *Unstatistik des Monats*
  - Regelmäßig erweiterte Sammlung handwerklicher Fehler und irreführender Rechnungen in öffentlichen Medienberichten