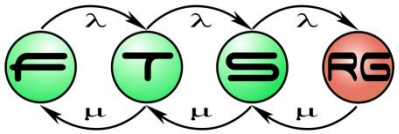
Visuelle Datenanalyse

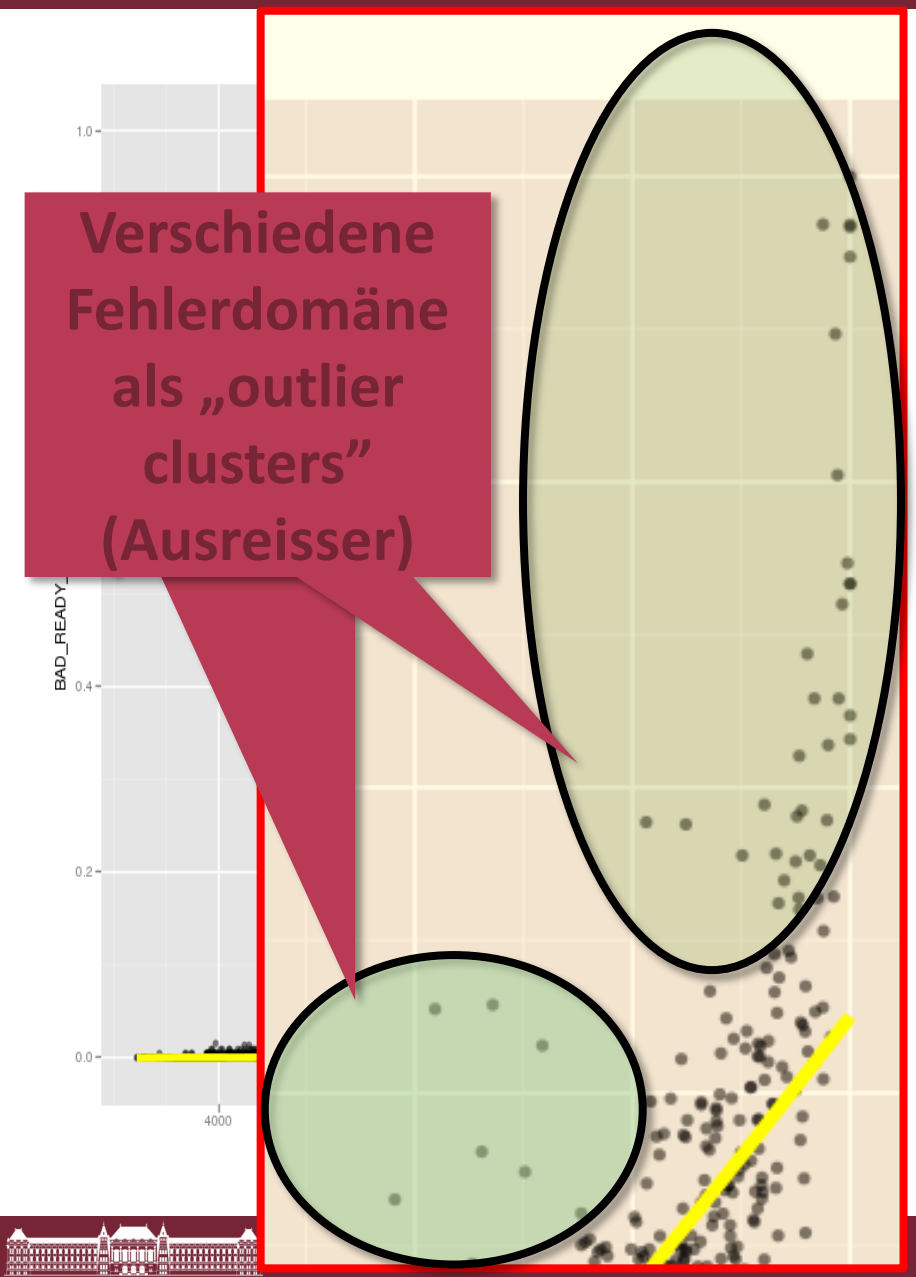
Salánki Ágnes, Dr. Guta Gábor

Dr. Pataricza András

Budapest University of Technology and Economics
Fault Tolerant Systems Research Group



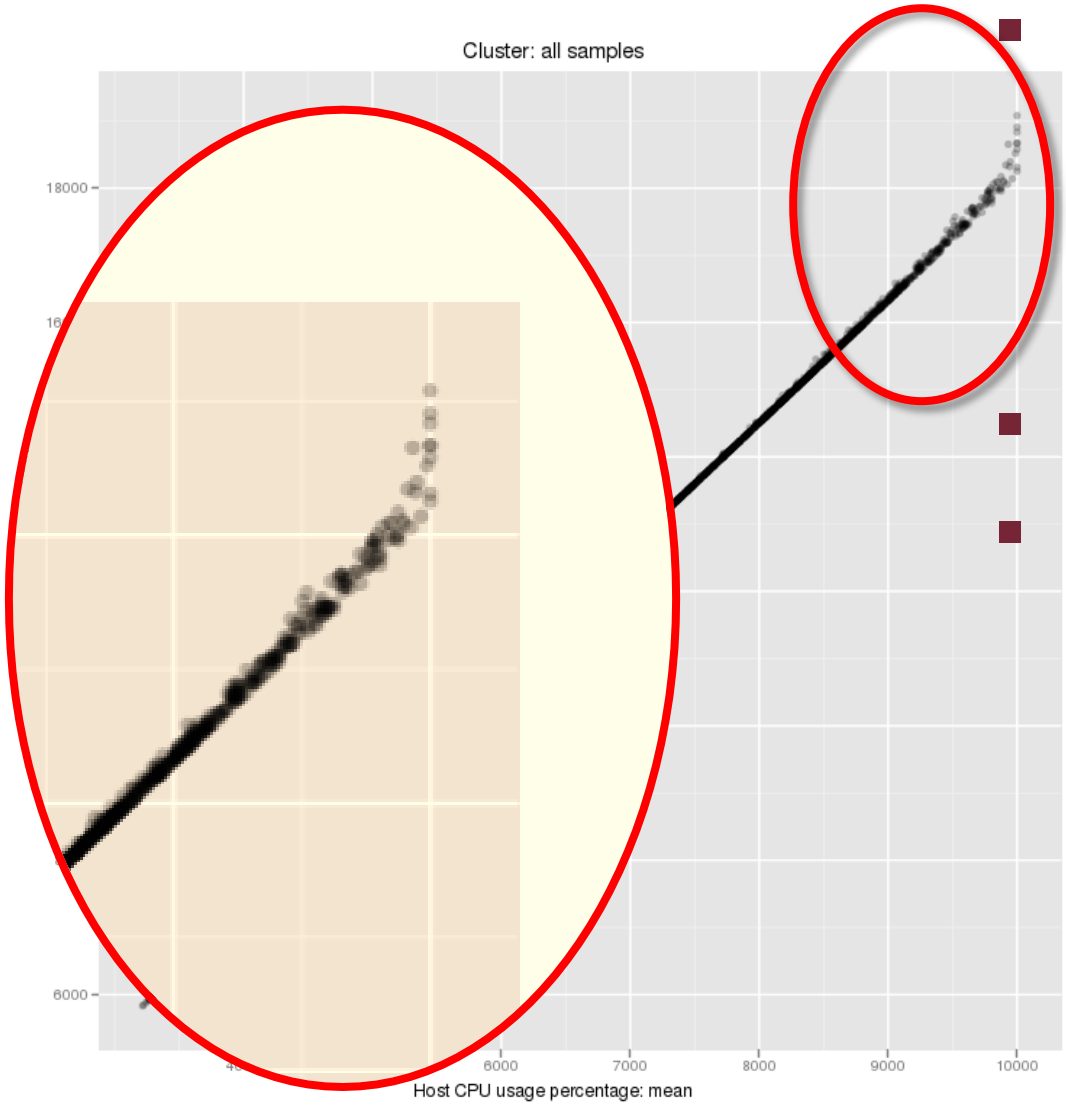
Beispiel: host CPU usage



- Last am Host-CPU
- Anteil der „bad vCPU ready”

- Mit etwas Reserve kann die Last besser bedient werden
- Scheduling hilft

Beispiel: Eine seltsame Beobachtung

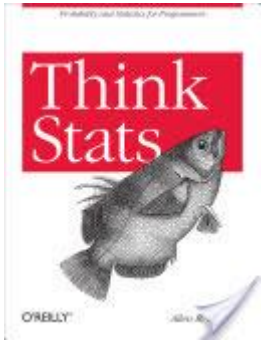


- *Verbrauchte Rechen-leistung = CPU Auslastung* × CPU Taktfreq. (const.)
- Sollte proportional sein
- Korrelationskoeffizient: 0.99998477434137

Wiederholung: Statistik (Grundlagen)

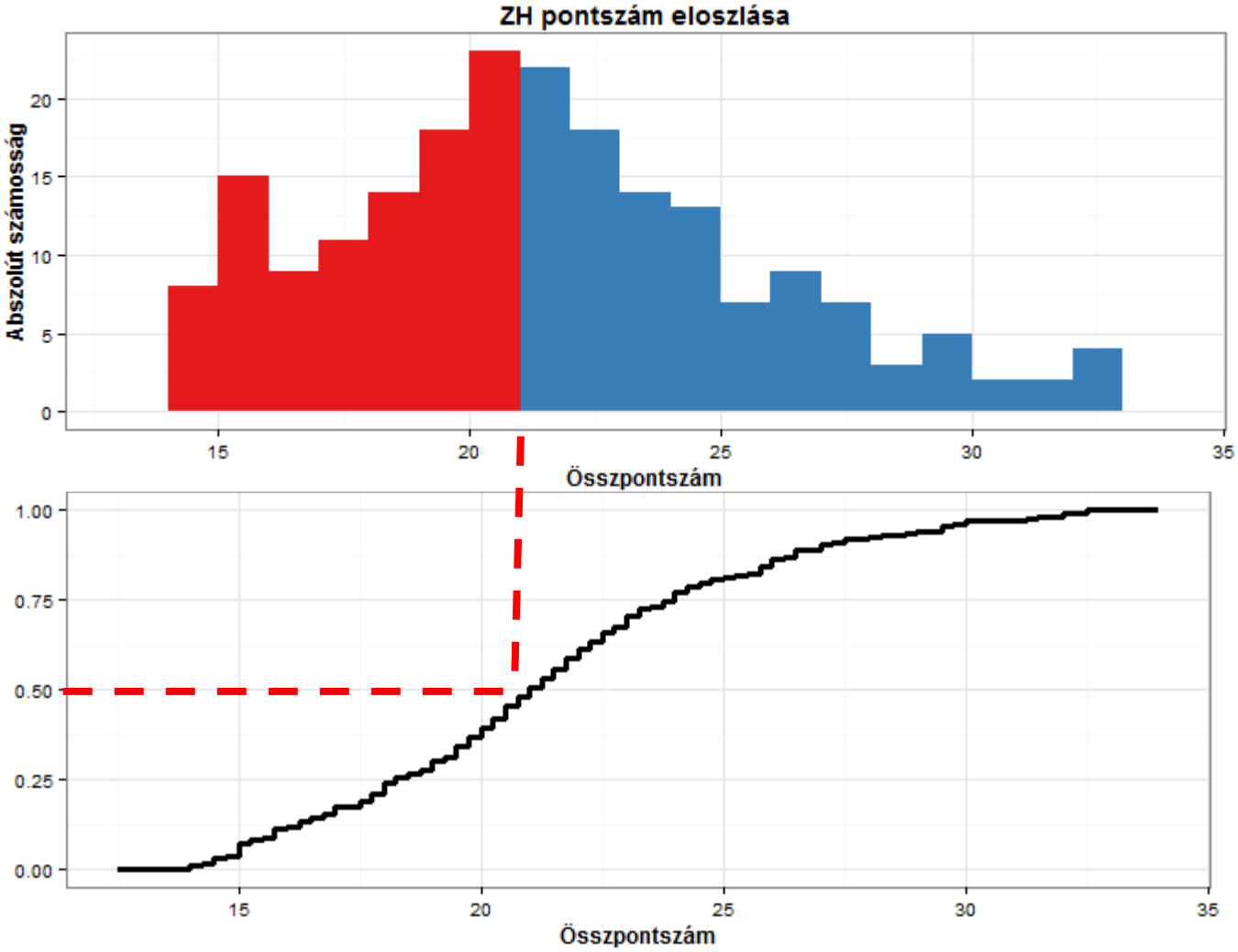
- Min: der kleinste Wert
- Max: der größte Wert
- Der Durchschnitt: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$
- Die Varianz (Streuung): $\frac{\sum_{i=1}^{n}(x_i - mean)^2}{n}$

- Think Stats: Probability and Statistics for Programmers



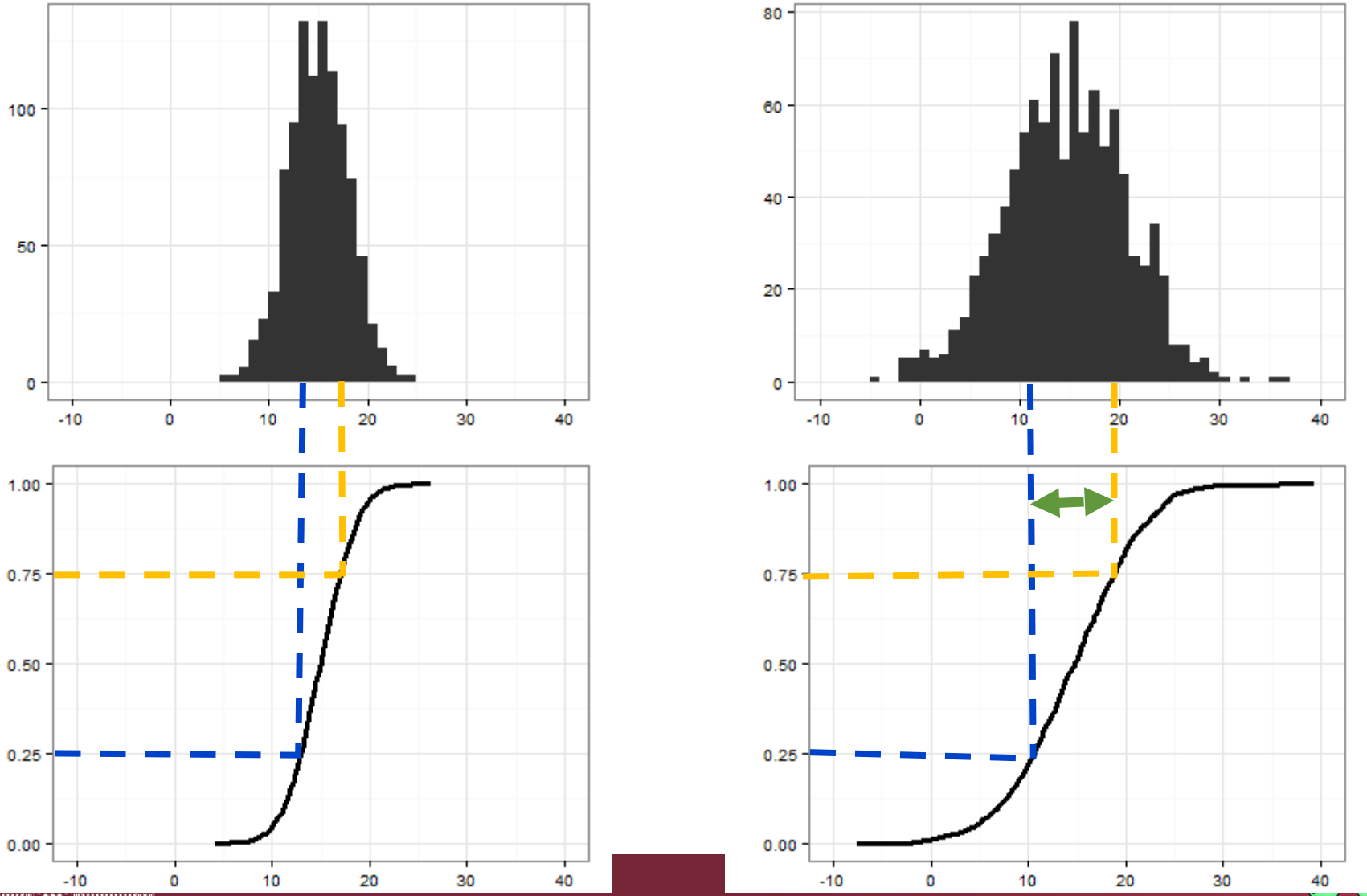
Einfache statistische Beschreibung

- Wo ist „die Mitte” der Werte?



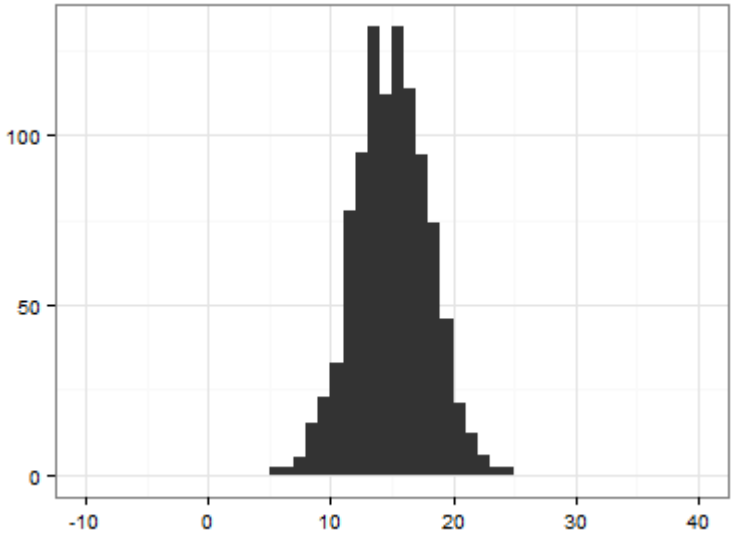
Einfache statistische Beschreibung

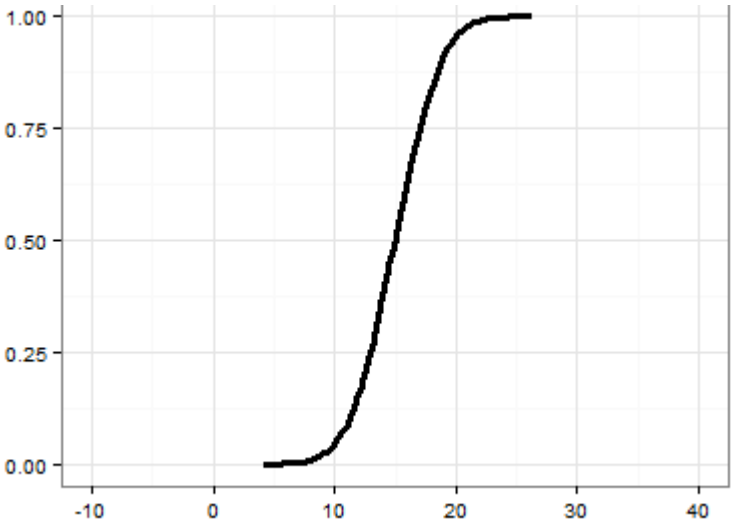
- Wie weit sind die Werte „gestreut”?

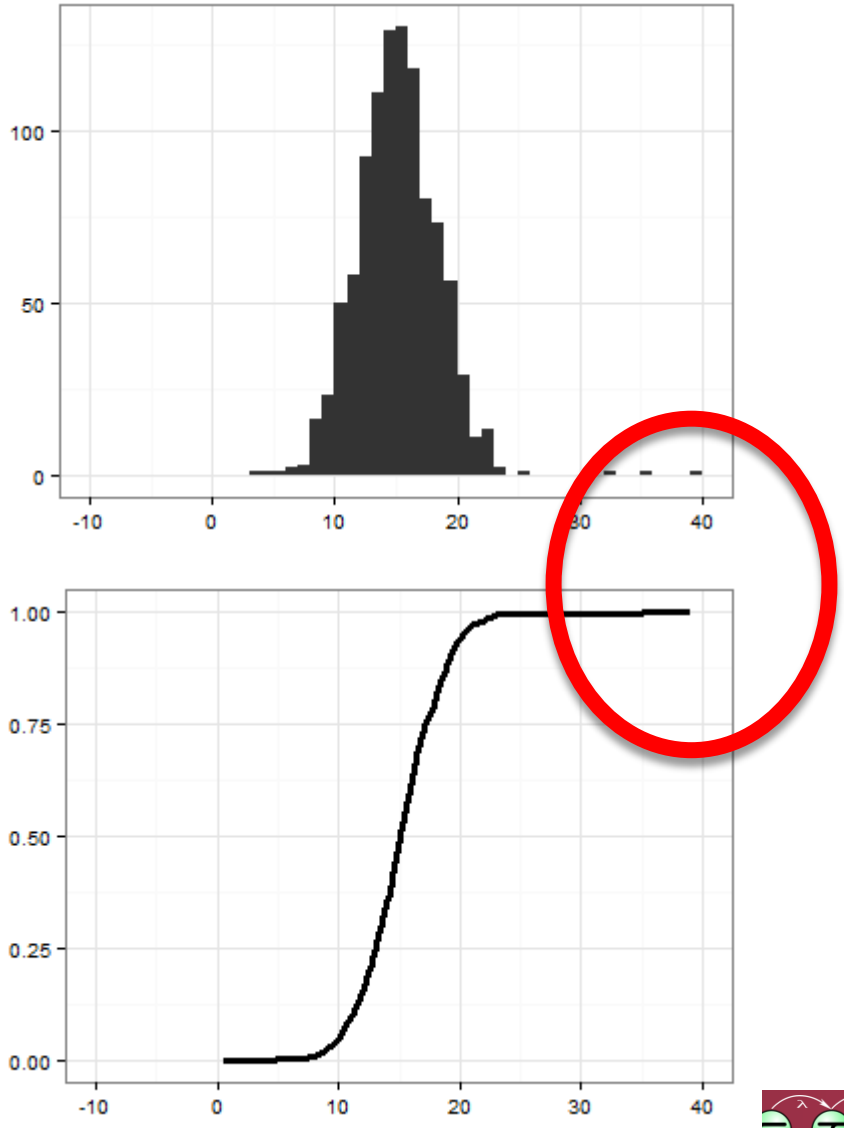


Einfache statistische Beschreibung

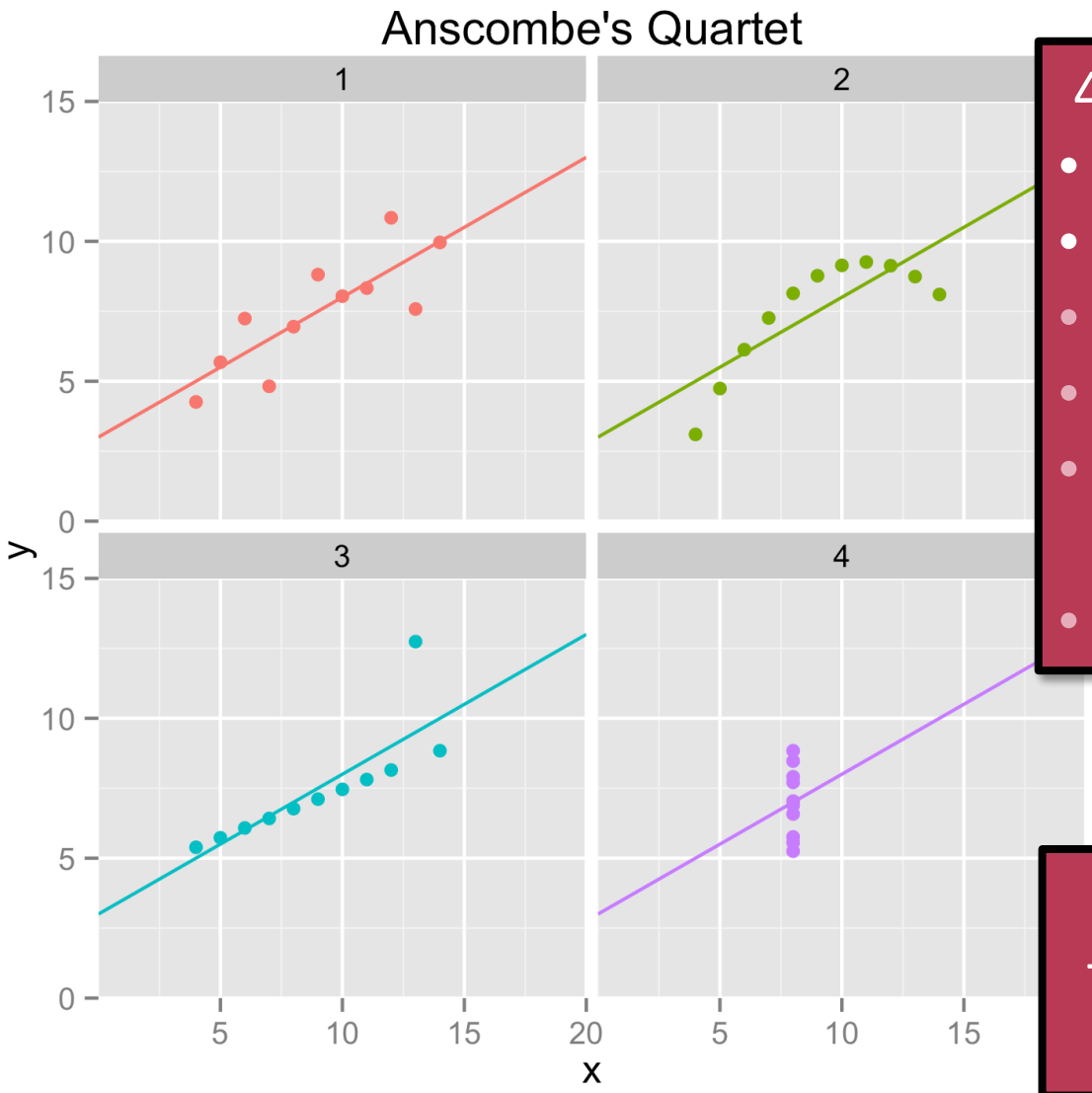
- Gibt es Ausreisser?







Überprüfung der Berechnungen



4 Datensätze mit gleichen:

- Durchschnitt von x (9)
- Streuung von x (11)
- Durchschnitt von y (7,5)
- Streuung von y (4,12)
- Korrelation der beiden (0,816)
- Regressionsgerade

Vermeidung der fehlerhaften Annahmen… und Intuition

Inhalt

Visualisieren – Warum?

Visualisieren – Was?

Visualisieren – Wie?

Inhalt

Visualisieren – Warum?

Visualisieren – Was?

Visualisieren – Wie?

Visualisieren im Alltag

Analoge Anzeige

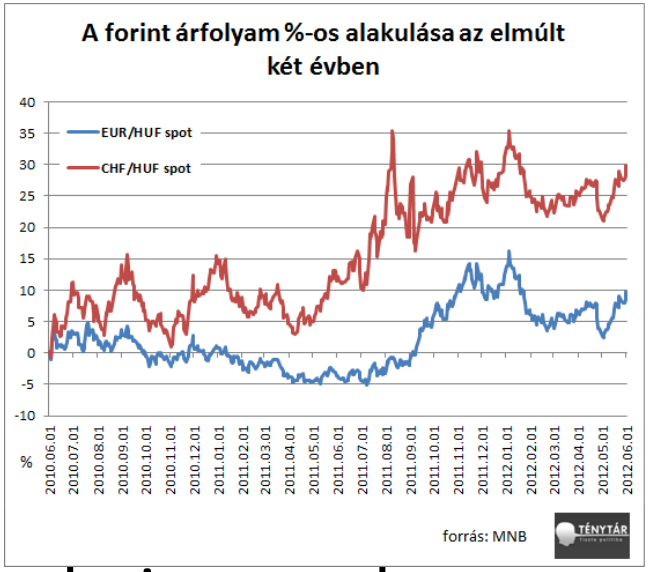
Digitale Anzeige

Analog + Koordinat

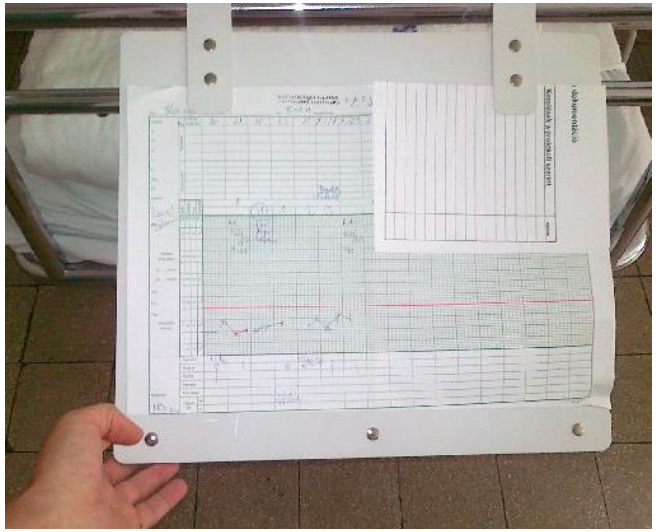
·ide Anzeige

Visualisieren im Alltag

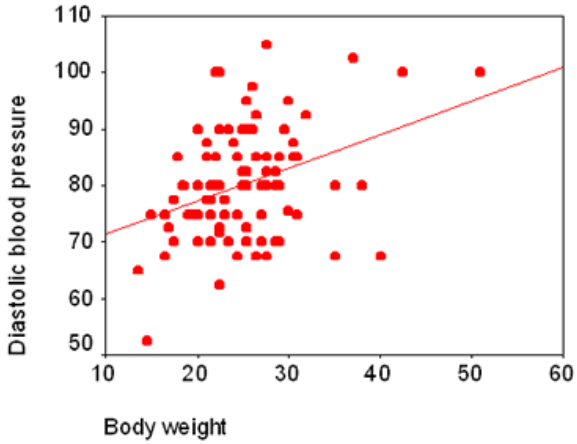
Trendanalyse und Vorhersage



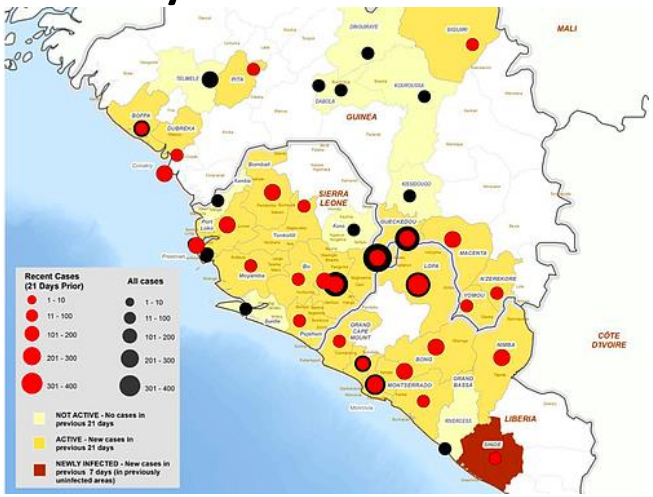
Zeitreihenanalyse



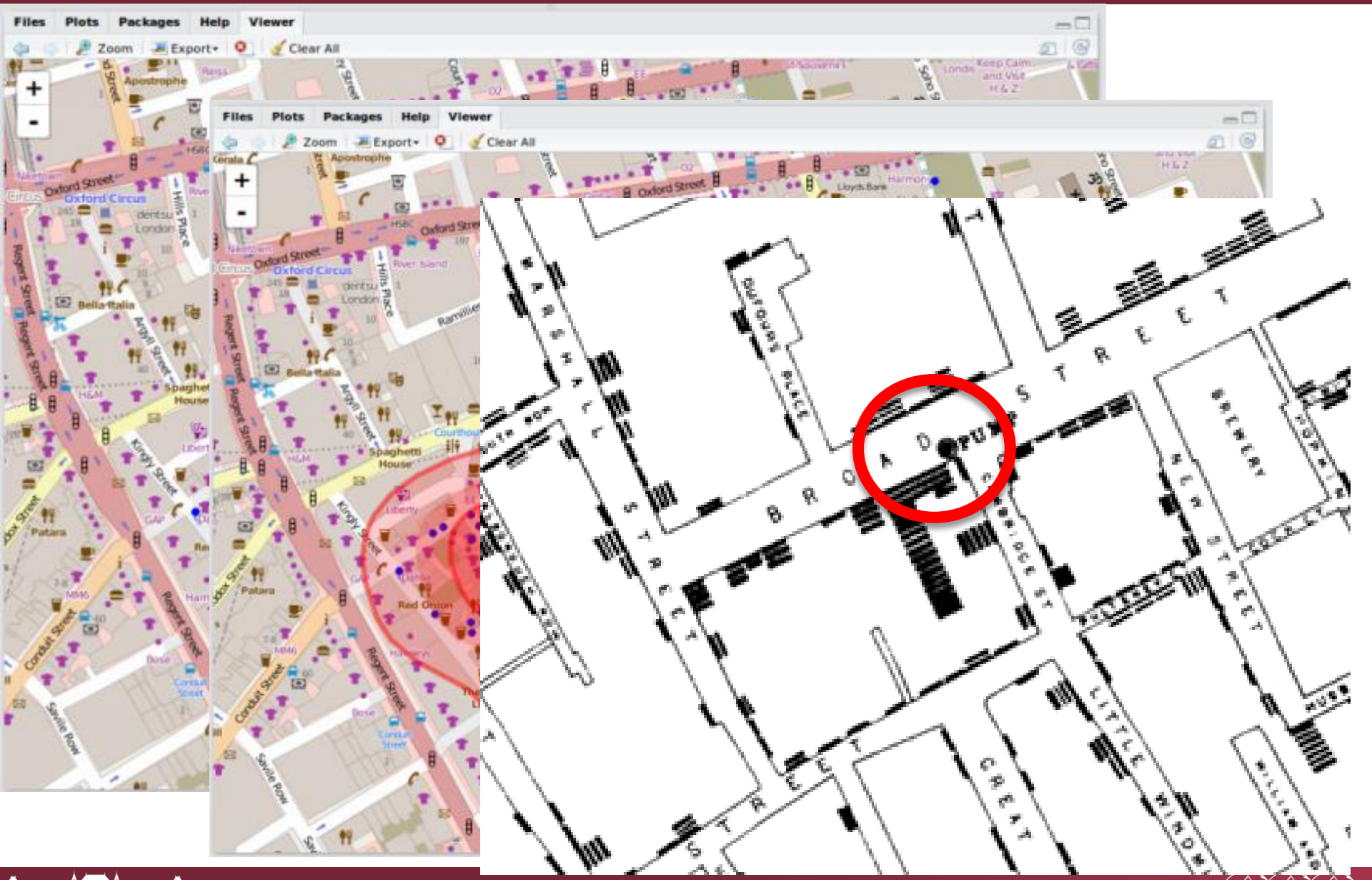
Korrelationsanalyse



Analyse der räumlichen Vert.



Erschließen von Zusammenhängen



Alle Augen auf die Daten!

„Massiv-Parallele” Bearbeitung

- 120.000.000 Sensoren
- 10^{10} B[illegible]

5. Visuelle Auswahl und Manipulation
4. Interpretation, Korrelation mit anderen Modellen, Auswertung

Inhalt

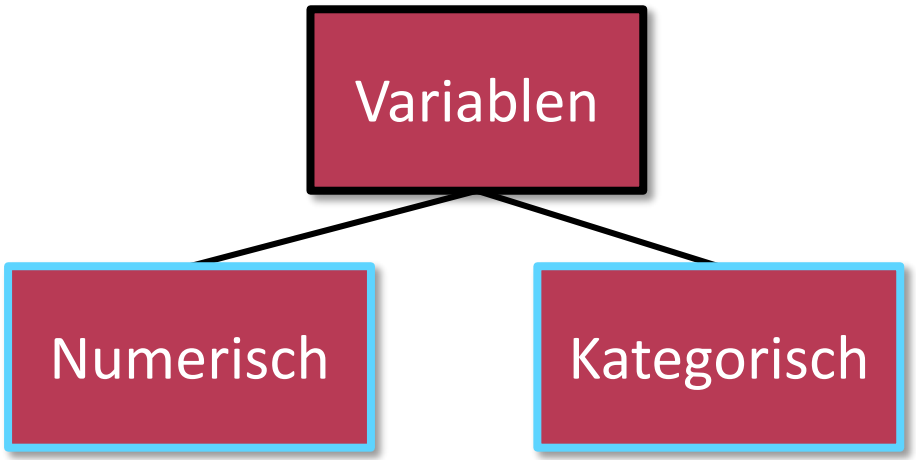
Visualisieren – Warum?

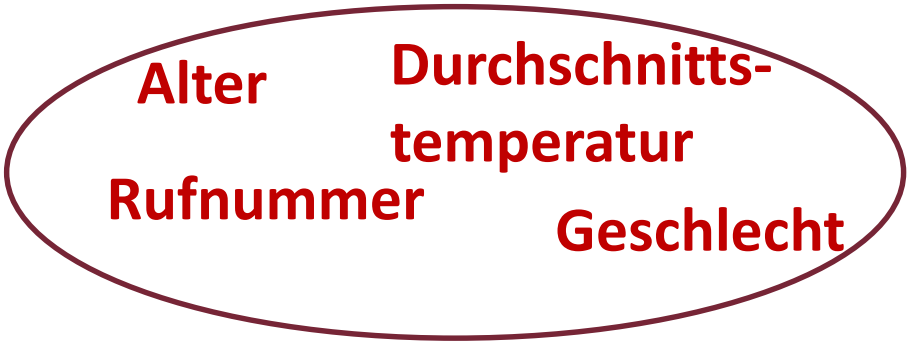
Visualisieren – Was?

Visualisieren – Wie?

Numerische und kategorische Variablen

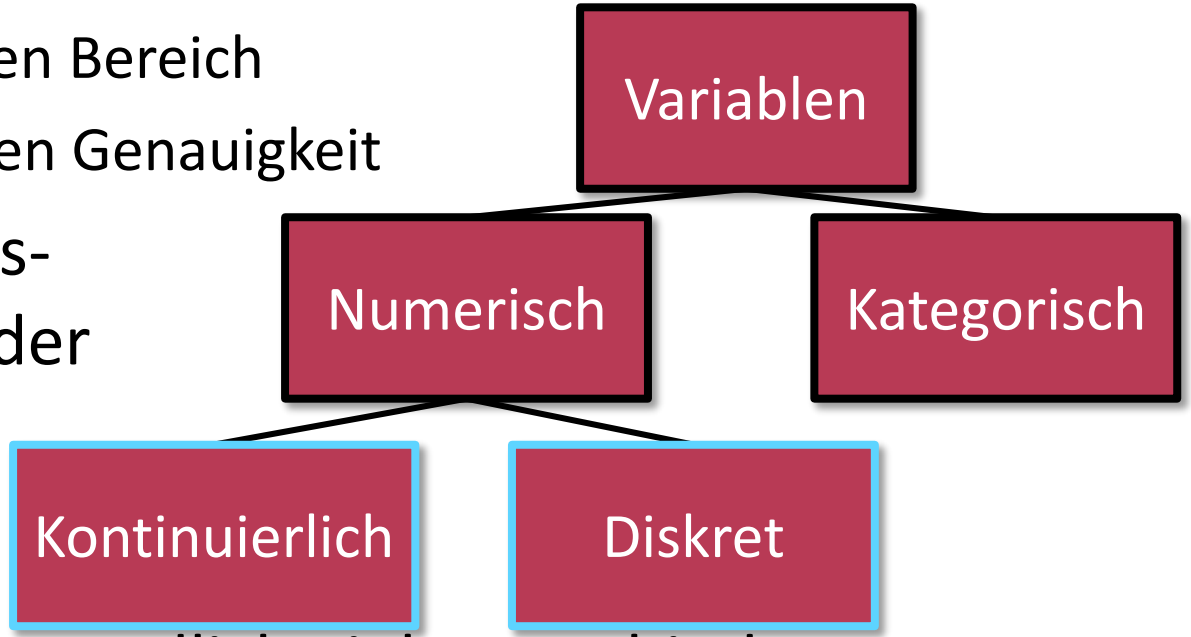
- Numerisch (numerical)
  - Die arithmetische Grundoperationen sind sinnvoll definierbar
- Kategorisch (categorical)
  - Die Kategorien dienen die Unterscheidung, die Operationen sind sinnlos





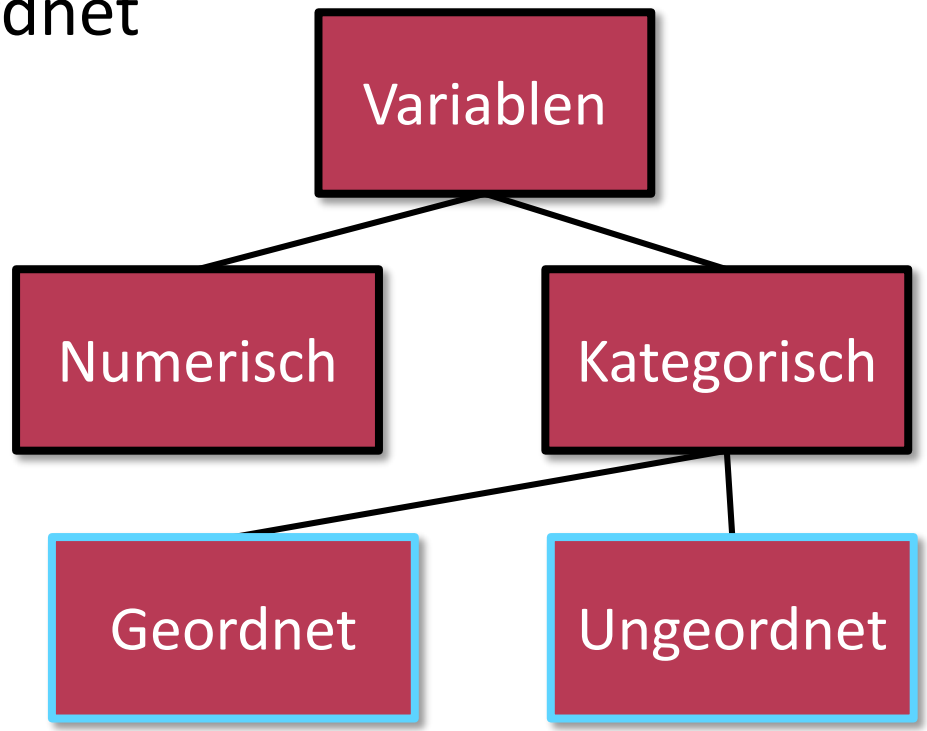
Numerische Variablen

- Kontinuierlich
  - Gemessen – kann beliebige Werte aufnehmen
    - in dem gegebenen Bereich
    - bei der gegebenen Genauigkeit
  - z.B. Durchschnitts-Klausurergebnis der Anwesenden
- Diskret
  - Gezählt – kann nur endlich viele verschieden Werte aufnehmen
  - z.B. Anzahl der Anwesenden



Kategorische Variablen

- Geordnet (Kategorisch → Ordinal)
  - Der Wertebereich ist geordnet
  - Vollständig geordnet?
- Ungeordnet (regulär)



9. Ajánlanád-e a tárgyat másoknak?

- Mindenkit rábeszélnék
- Nyugodtan ajánlanám
- Esetleg ajánlanám
- Inkább lebeszélném róla
- Feltétlenül lebeszélném
- Nem kívánok válaszolni

Inhalt

Visualisieren – Warum?

Visualisieren – Was?

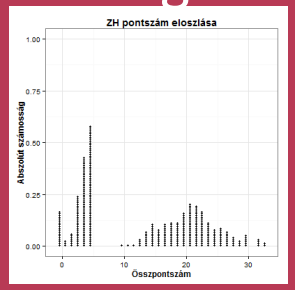
Visualisieren – Wie?

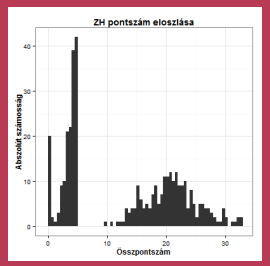
Verteilung einer Variable

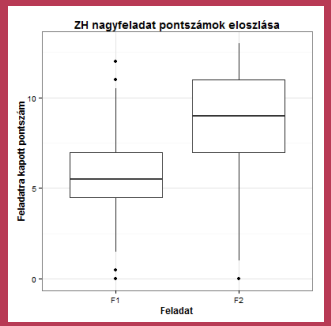
Variablen

Numerisch

Kl-Ergebnis: [13, 15, 2, …]

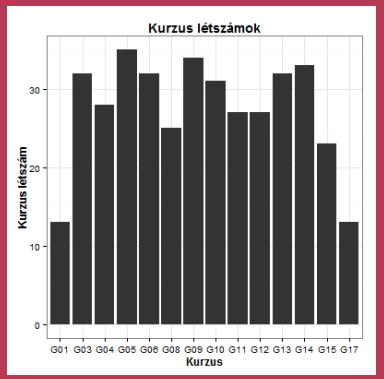






Kategorisch

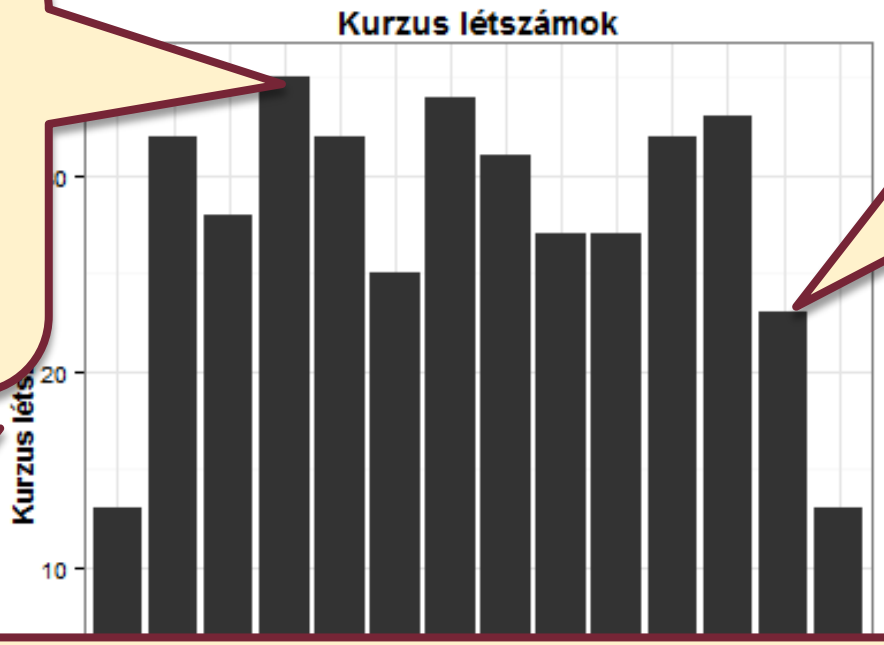
Kurse: [G01, G03, G15, G17, …]



Säulendiagramm/Balkendiagramm

- Eingabevariable: Kurs-Kode
- Frage: wie viele haben sich angemeldet?

Gibt es populäre Zeitpunkte/ Übungsleiter?



Säulenhöhe: Häufigkeit des gegebenen Wertes

absolute Häufigkeit!

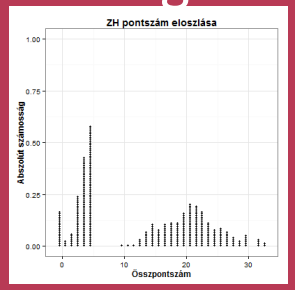
Entwurfsentscheidung: Aufteilung des Definitionbereiches
z.B.: Dienstag-Donnerstag-Freitag?

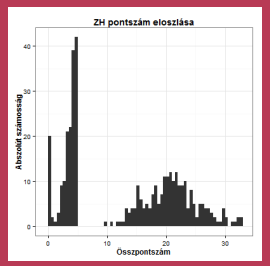
Verteilung einer Variable

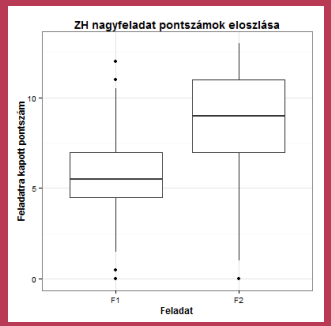
Variablen

Numerisch

Kl-Ergebnis: [13, 15, 2, …]

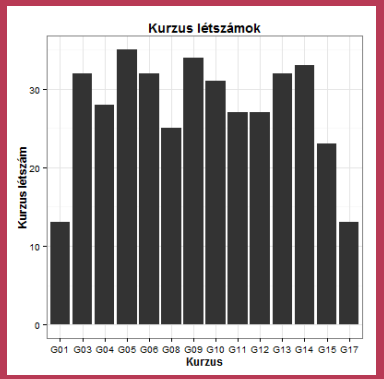






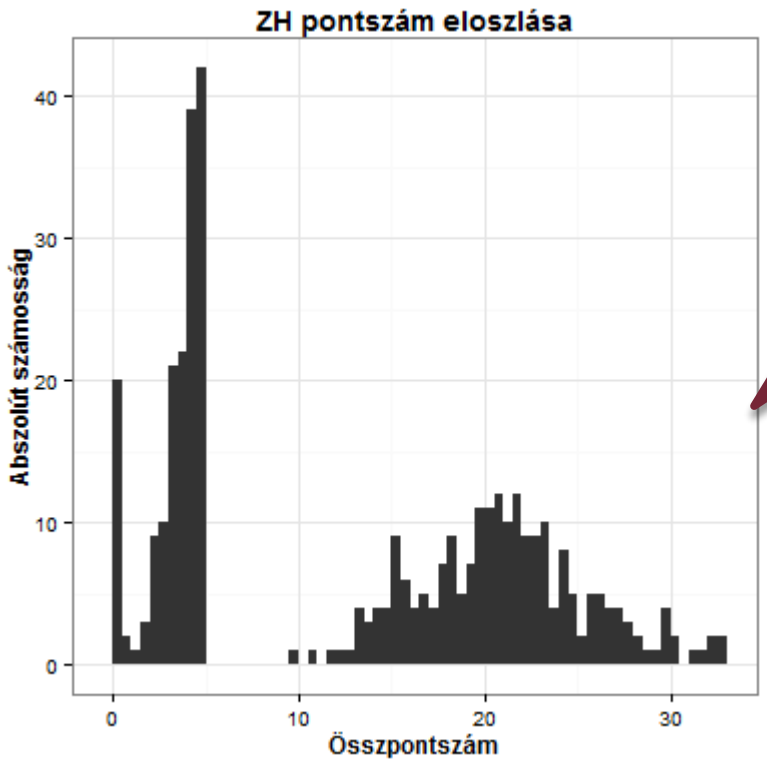
Kategorisch

Kurse: [G01, G03, G15, G17, …]



Histogramm

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren?



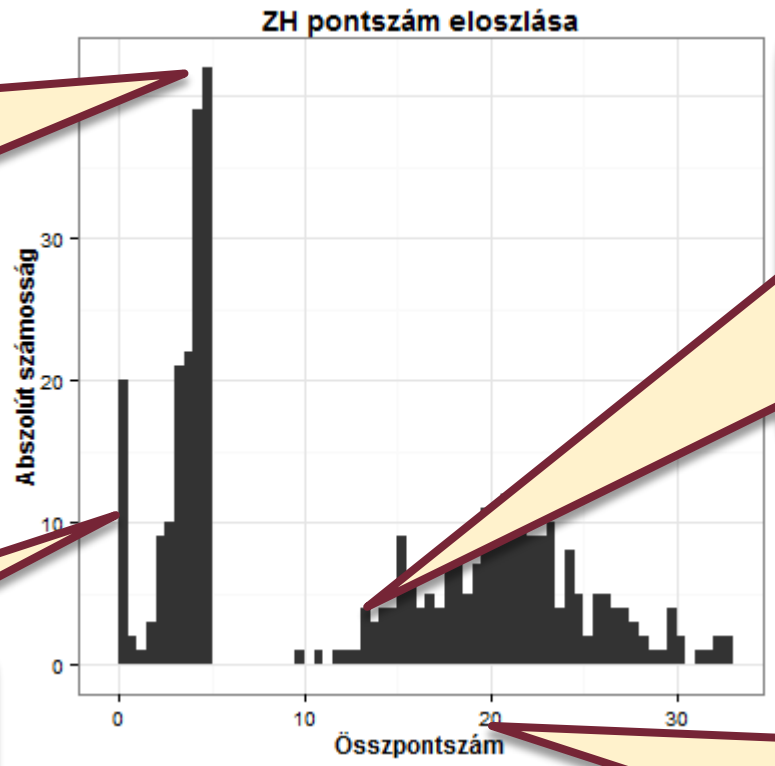
Säulenhöhe: Häufigkeit des gegebenen **Intervalls**

absolute Häufigkeit!

Entwurfsentscheidung: Bestimmung der Intervalllänge
Z.B.: 1-Punkte-Auflösung vs. 0,5-Punkte-Auflösung?

Histogramm

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren?



Viele haben das Test nur fast geschafft

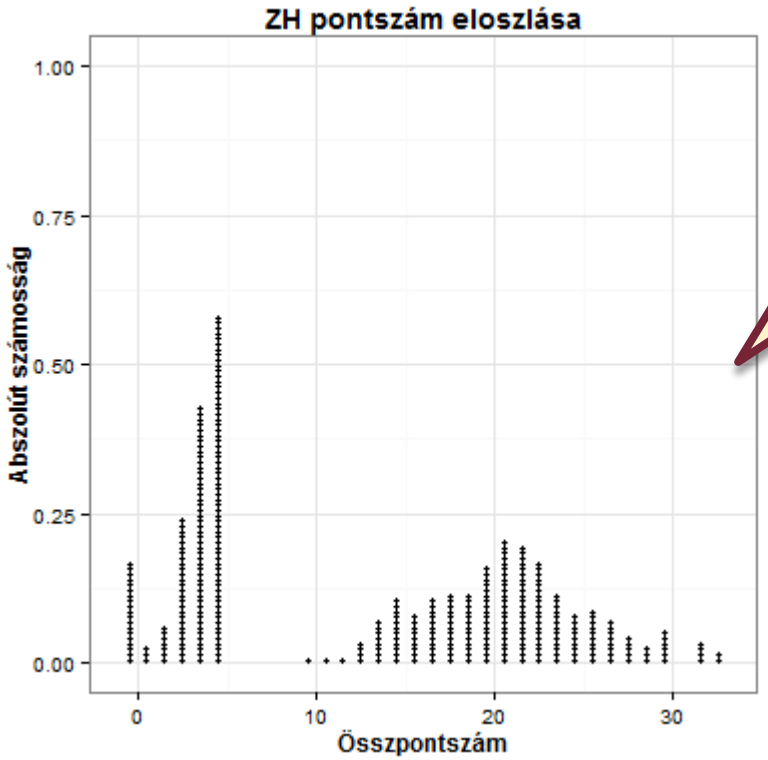
Die meisten von denen, die das Test schon bestanden haben, haben die Klausur geschafft.

Die gar nicht erschienen

Der Durchschnitt/ Median war so um 20 Punkte

Relative Häufigkeit

- Eingabevariable: Klausurergebnis
  - Frage: was für Ergebnisse sind geboren?



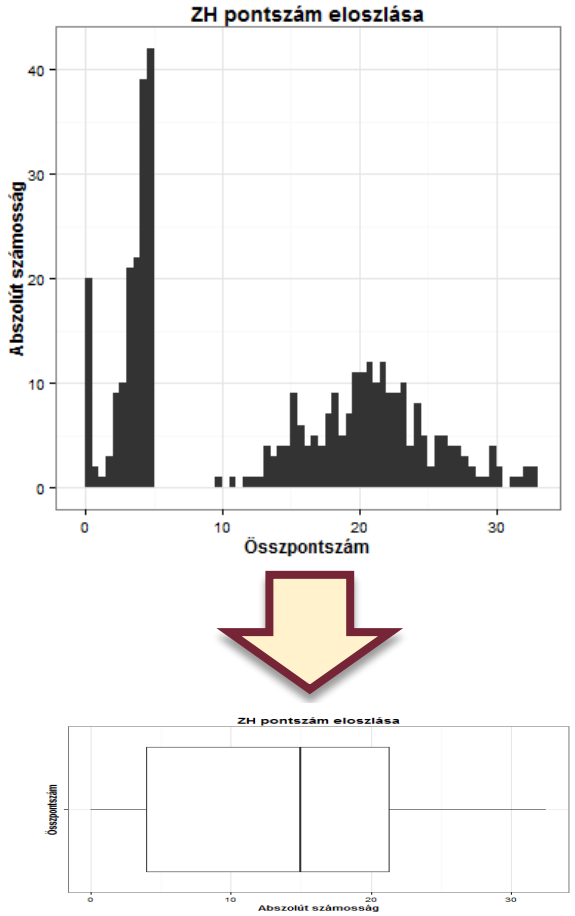
Genau wie beim Histogramm
(Histogramm mit relativer Frequenz)

Die **relative** Häufigkeit!

Kastengraphik/Boxplot

- Eingabevariable: Klausurergebnis
- Frage: was für Ergebnisse sind geboren **so ungefähr**?

Eine Art Abstraktion:
nehmen wir Intervalle,
die einzelnen Werte sind
uninteressant



Beschreibung (kontinuierlicher) Beobachtungen

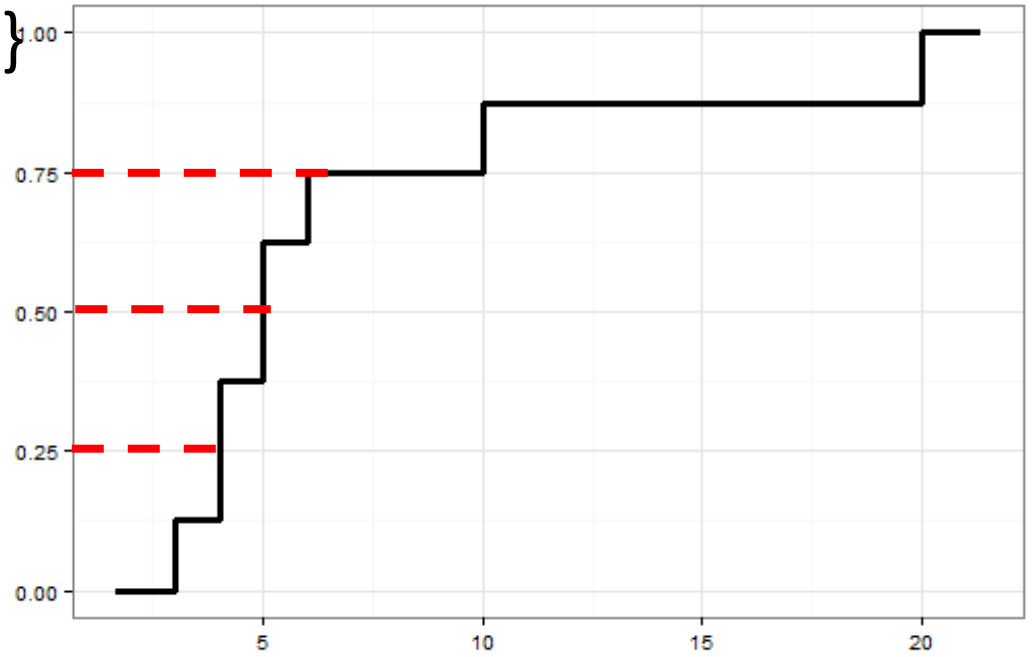
- Beschreibung der „Mitte”
  1. Durchschnitt (Mittelwert) – arithmetisches Mittel
  2. **Median** (Zentralwert) – mittleres Element (geordnet)
  3. Modalwert – häufigstes Element
  - Beispiel: {3, 4, 4, 5, 5, 6, 10, 20}
    - Durchschnitt: ~ 7.125
    - Median: 5
    - Modalwert: 4 und 5

- Beschreibung der „Ausdehnung”
  - Perzentile (Frequenzen für kategorische Typen)

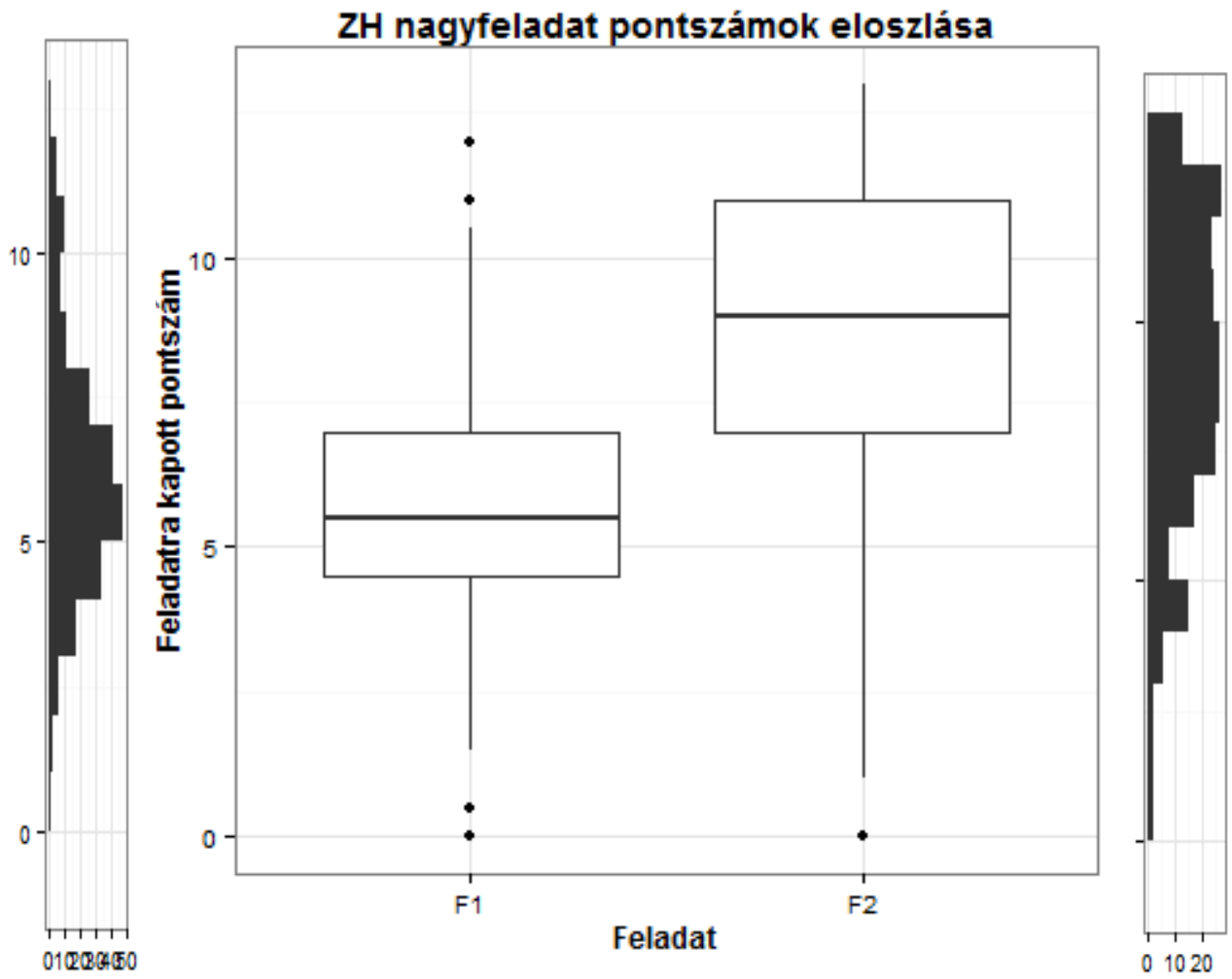
Perzentile

- Perzentil
  - $n\%$ der Werte sind kleiner als das n-te Perzentil
  - {3, 4, 4, 5, 5, 6, 10, 20}
    - 50. Perzentil: 5
    - 25. Perzentil: 4
    - 75. Perzentil: 10
- Quartil
  - Q1: 25. Perzentil
  - Q3: 75. Perzentil
  - **Q2: Median**

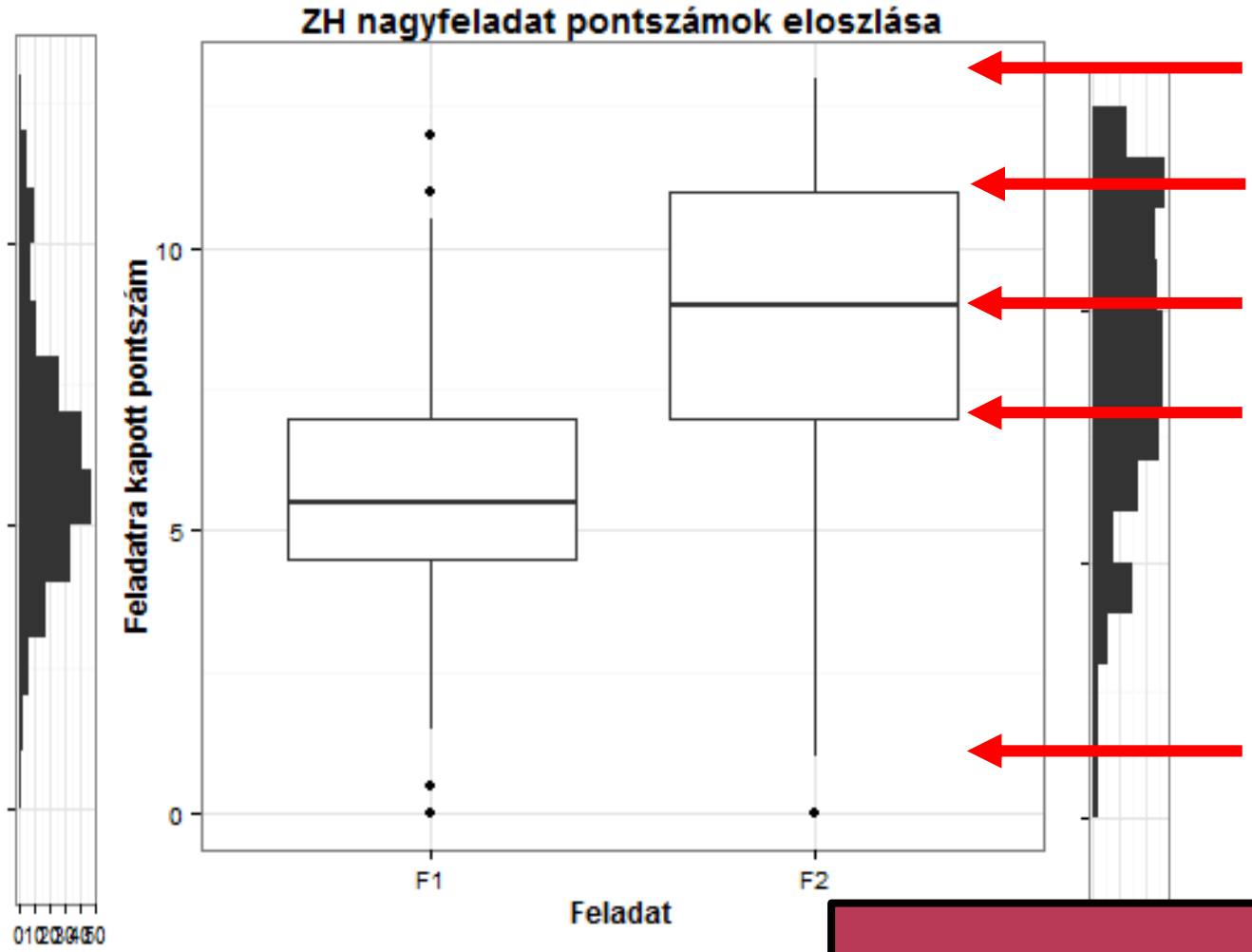
Frequenz: $n\%$ der Werte fallen in die gegebene Kategorie(n)



Kastengraphik (Box and whisker plot)



Kastengraphik (Box and whisker plot)



Obere Ausreissergrenze

Q3

Median

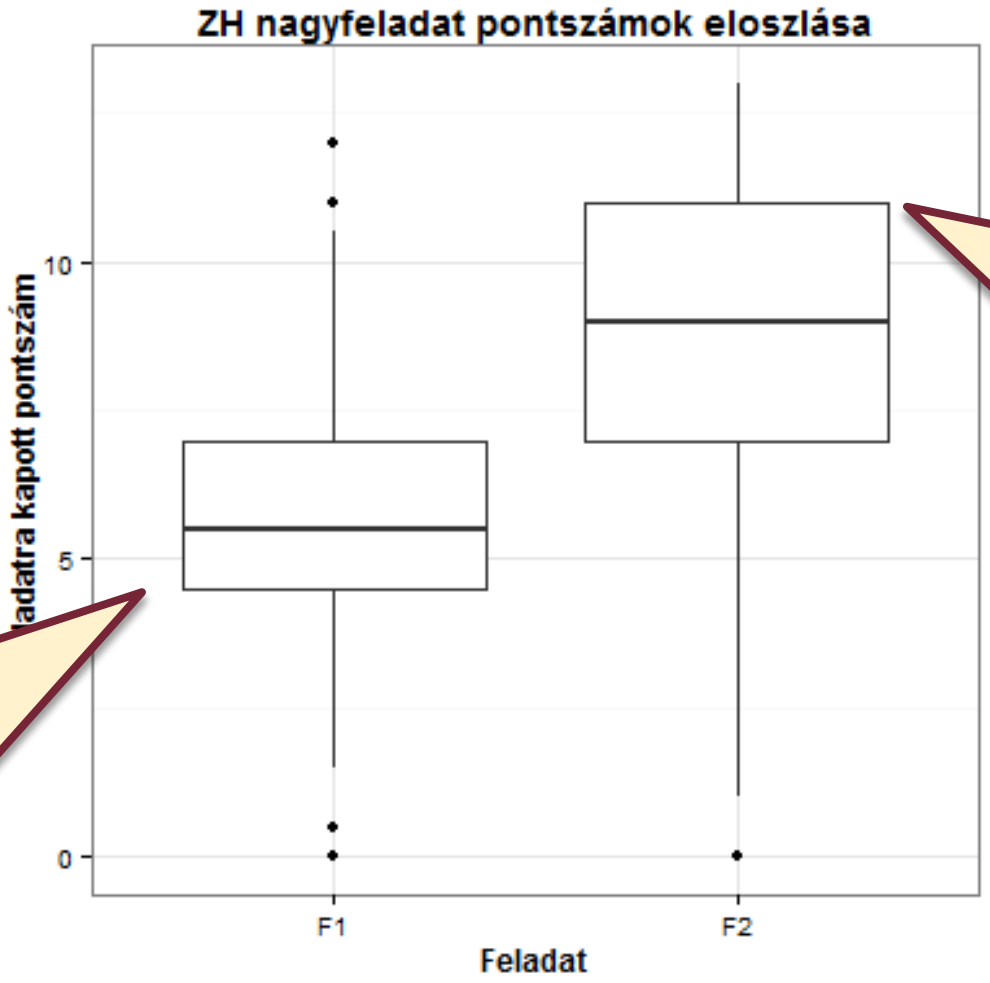
Q1

Untere Ausreissergrenze

Wie geht es in Excel?

(www.youtube.com/watch?v=ucWmfmXb1kk)

Kastengraphik (Box and whisker plot)

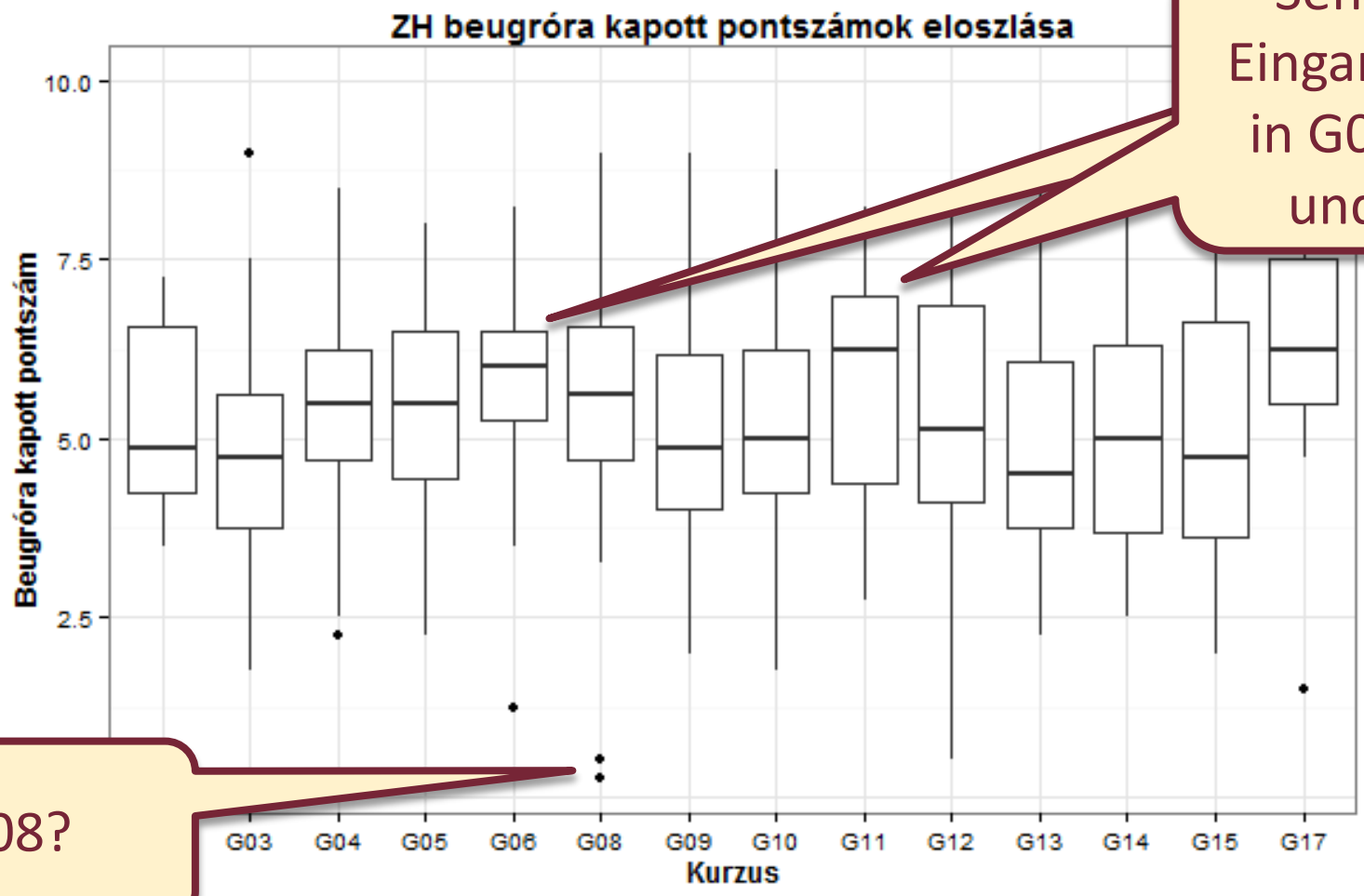


Aufgabe 2 lieferte (im Allg.) bessere Ergebnisse als Aufgabe 1.

50% der Ergebnisse der Aufgabe 1 liegen zwischen 4.5 und 7.5.

Kastengraphik (Box and whisker plot)

- Was waren die Ergebnisse per Übungsgruppe?

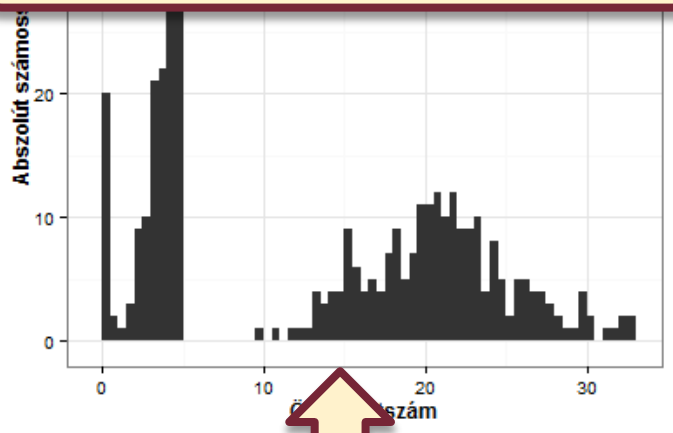


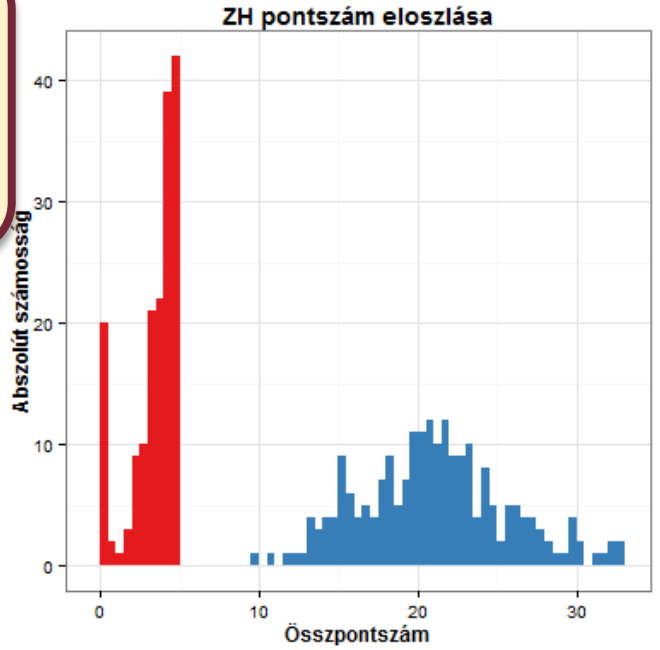
Sehr gute Eingangstests in G06, G11 und G17

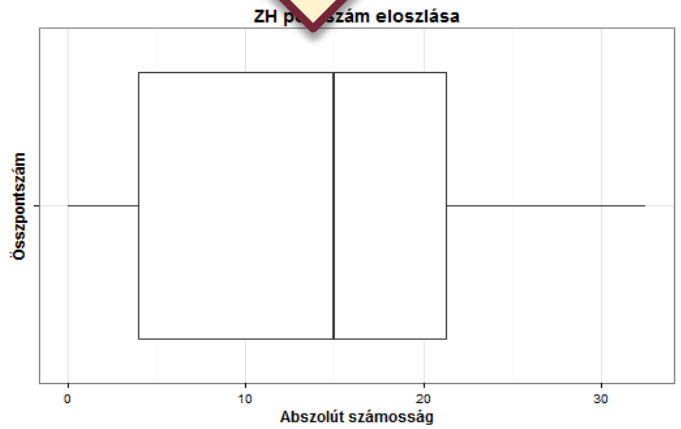
G08?

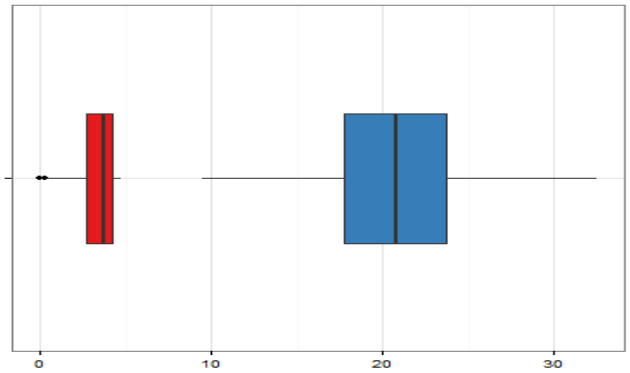
Kastengraphik (Box and whisker plot)

Abstraktion: Mit Kastengraphik können wir wichtige Informationen verlieren.



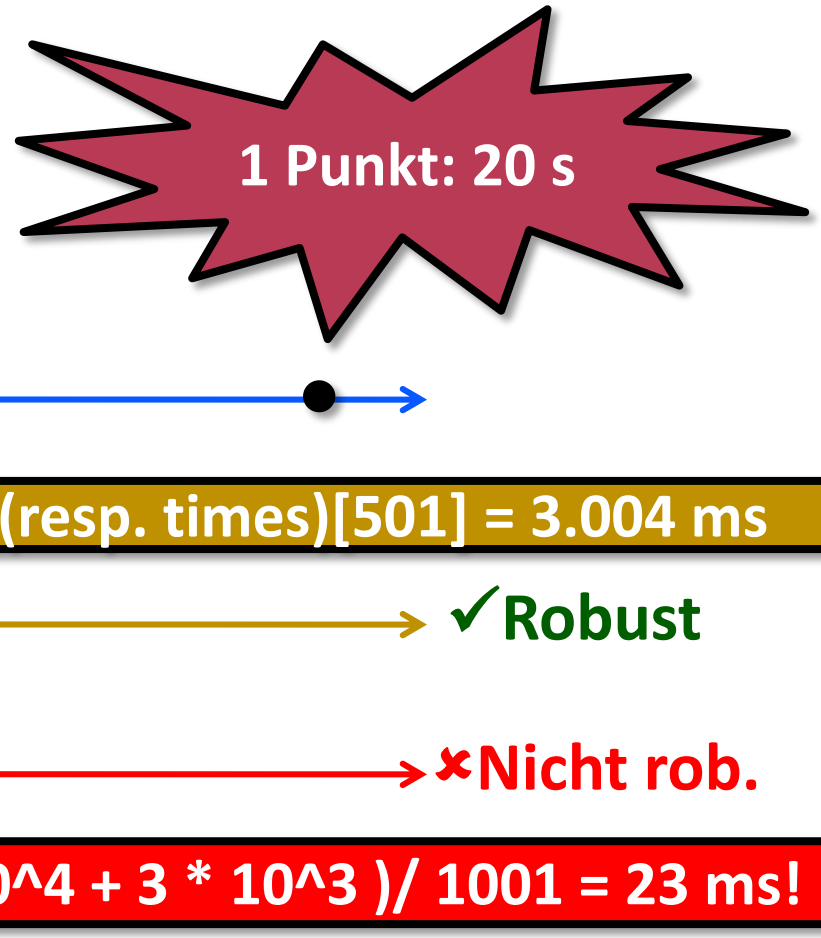




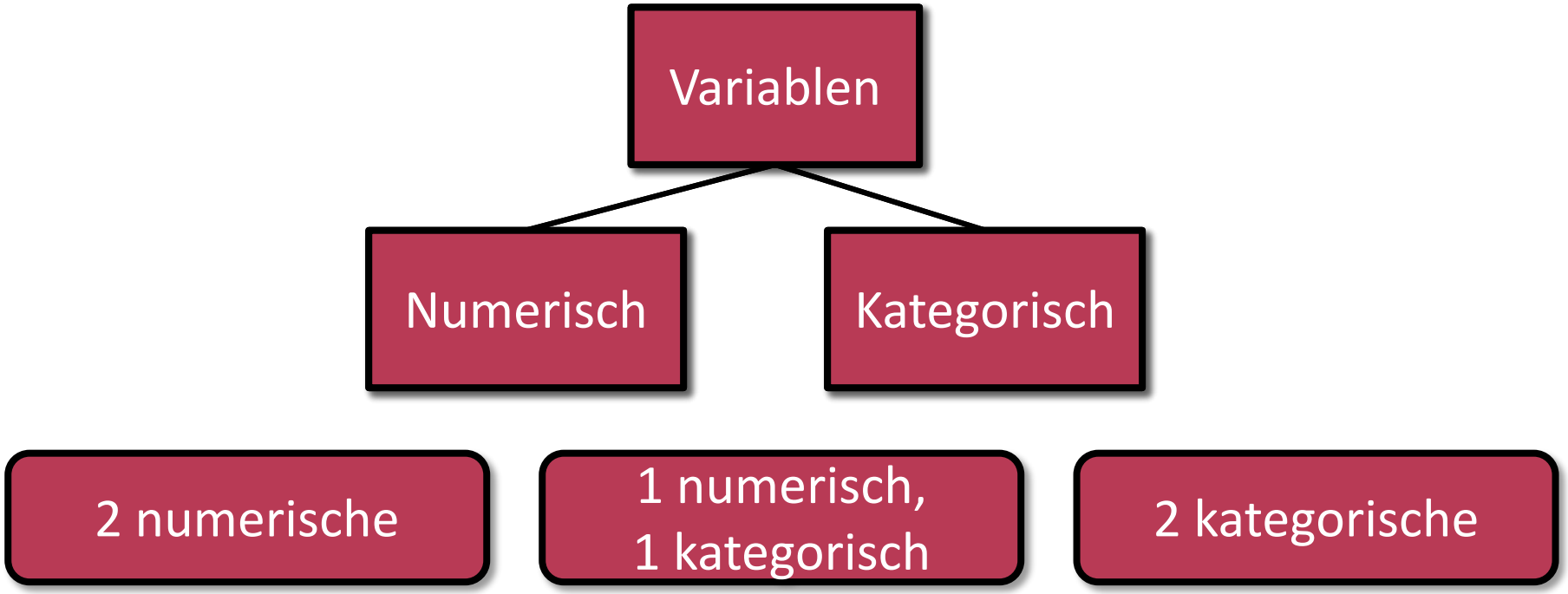


Zentralwert anstatt Mittelwert – Warum?

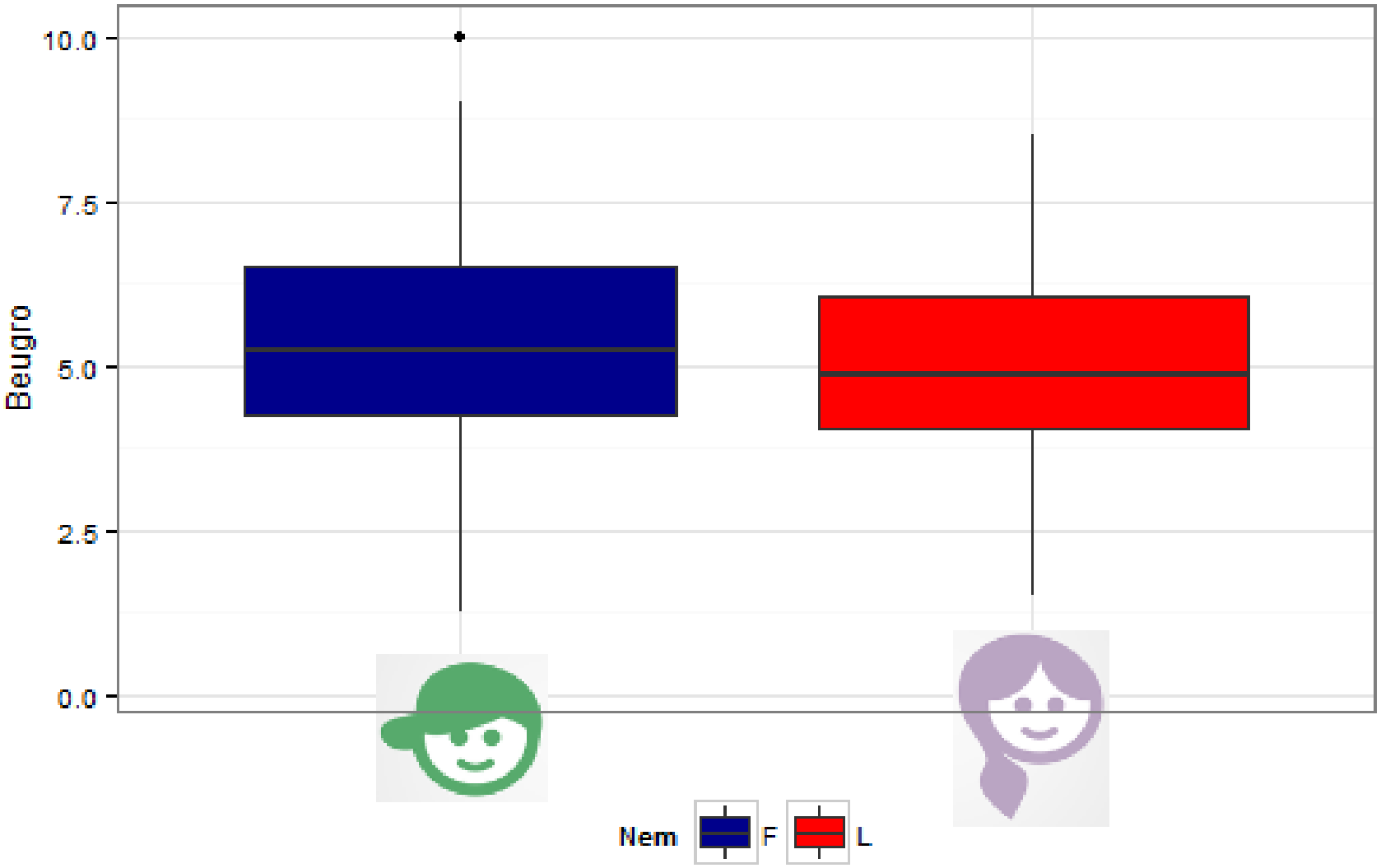
- Grundmenge: 1000 Punkte ~ U(1, 5) mit gleichmässiger Verteilung
  - *Mittelwert = Zentralwert = 3 ms*



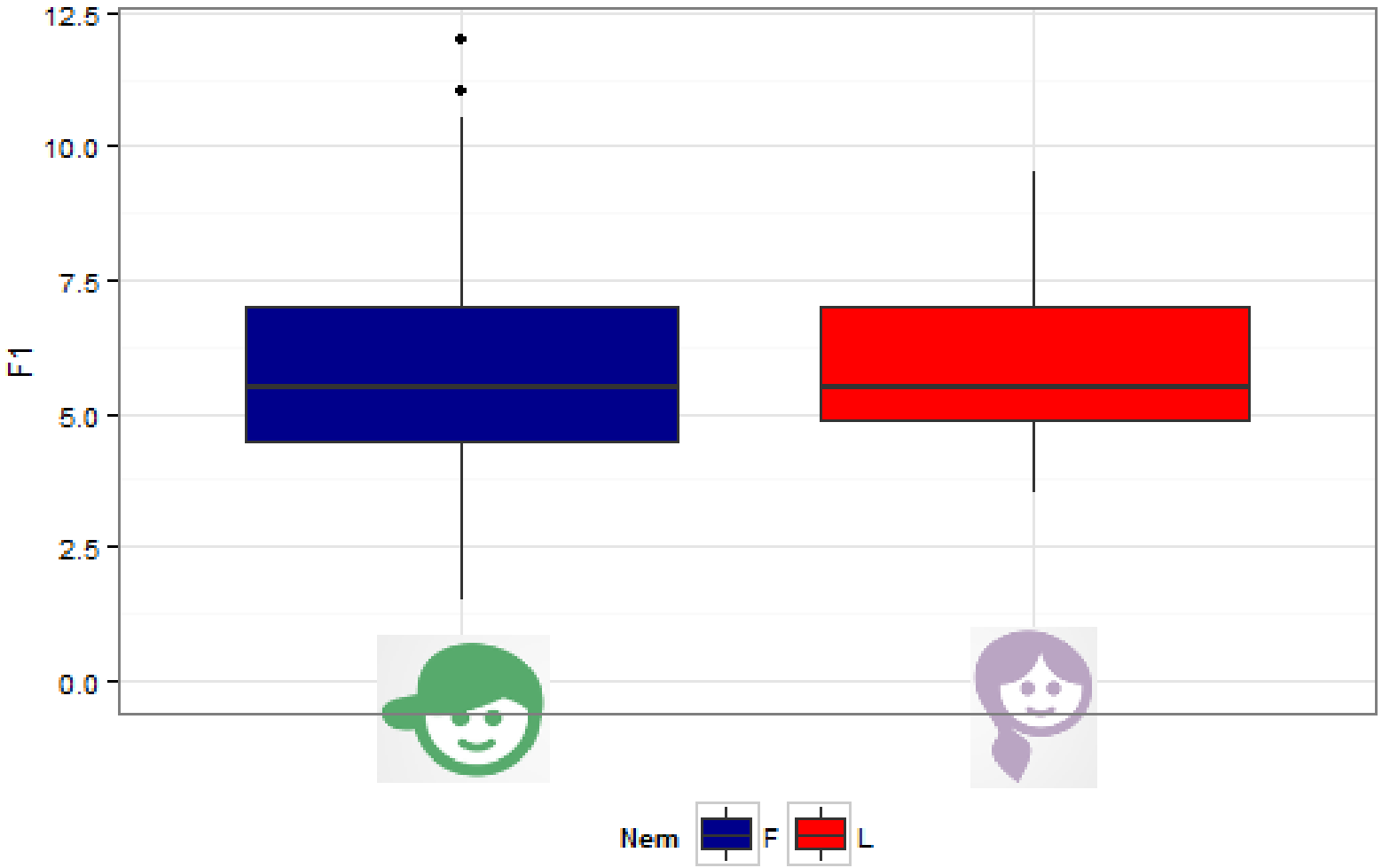
Zusammenhänge zweier Variablen



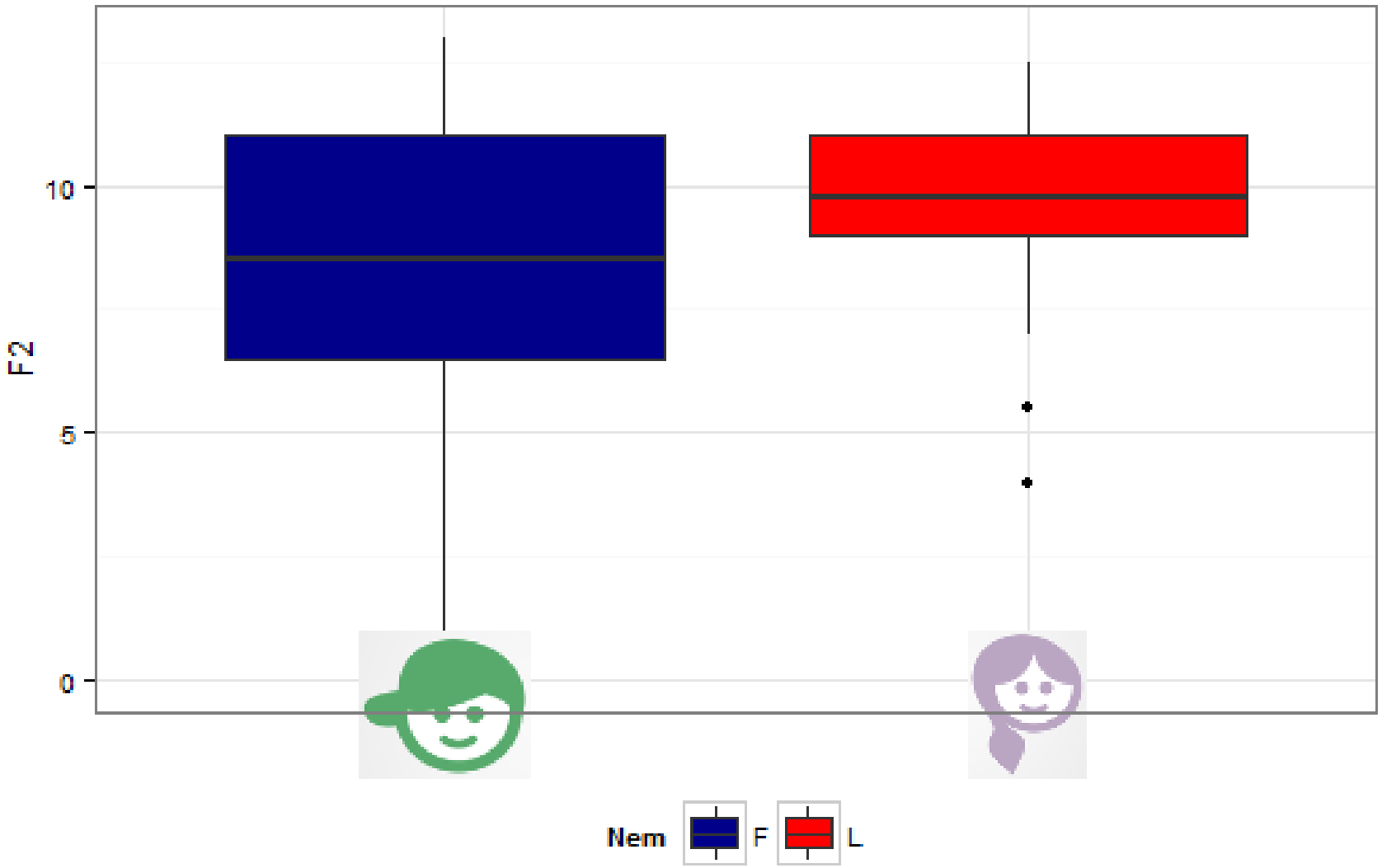
Numerisch, per Kategorie



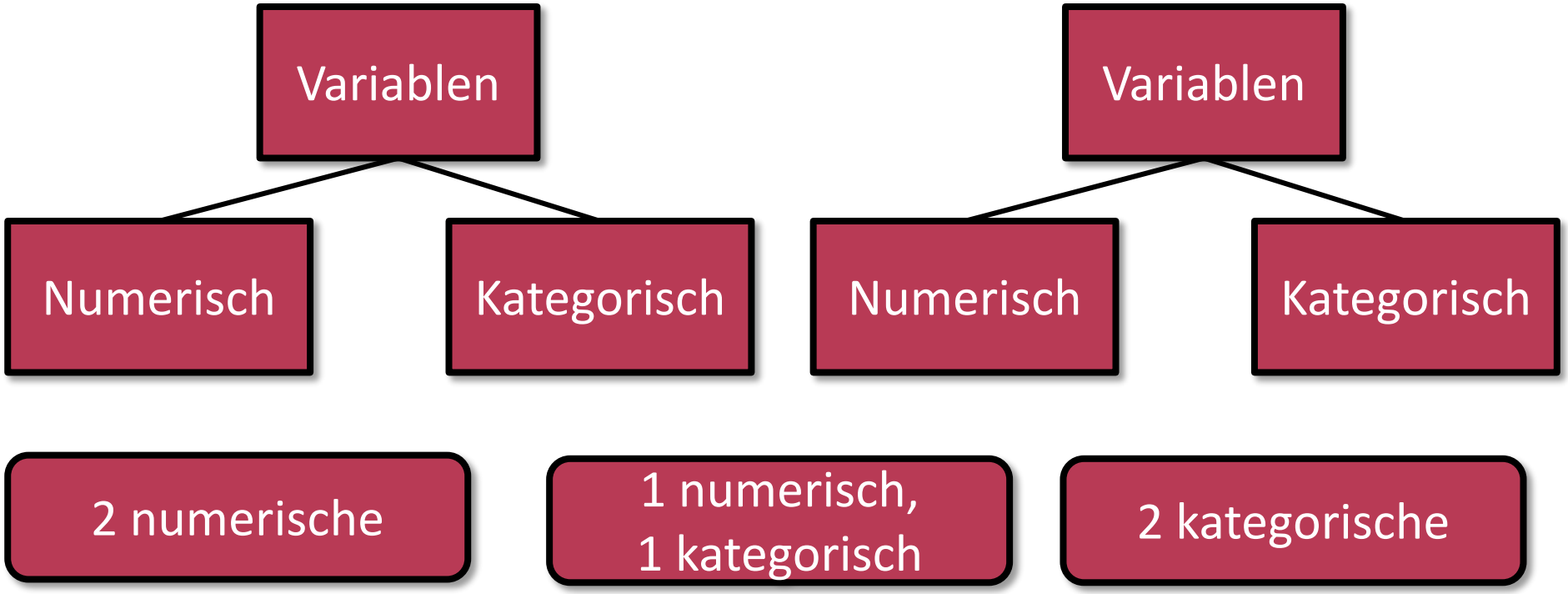
Numerisch, per Kategorie



Numerisch, per Kategorie



Zusammenhänge zweier Variablen

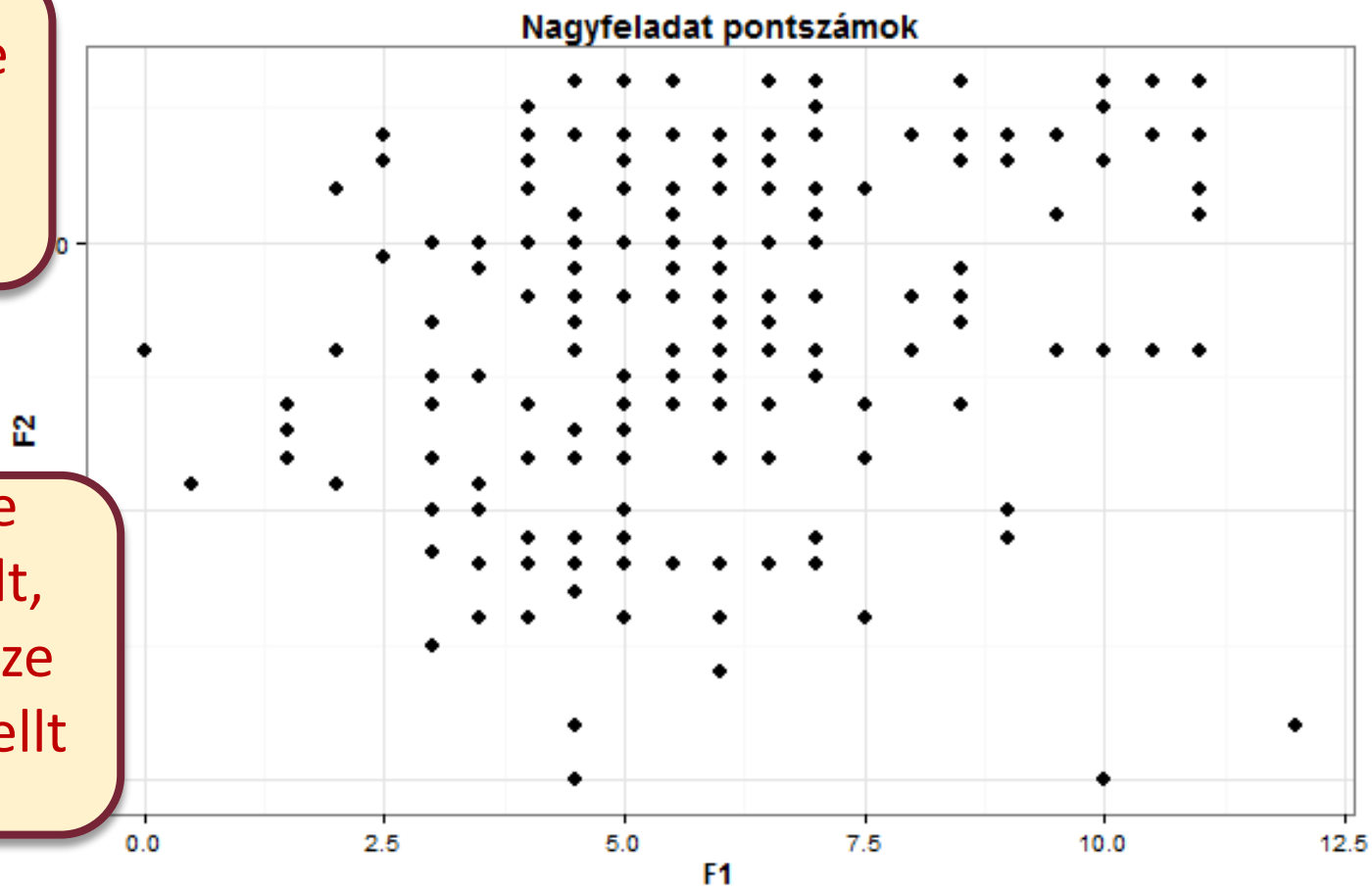


Streudiagramme (scatterplot)

- Eingabevariablen: Ergebnisse der Großaufgaben
- Frage: Was isr ihr Verhältnis?

Ergebnispaare sind visualisiert

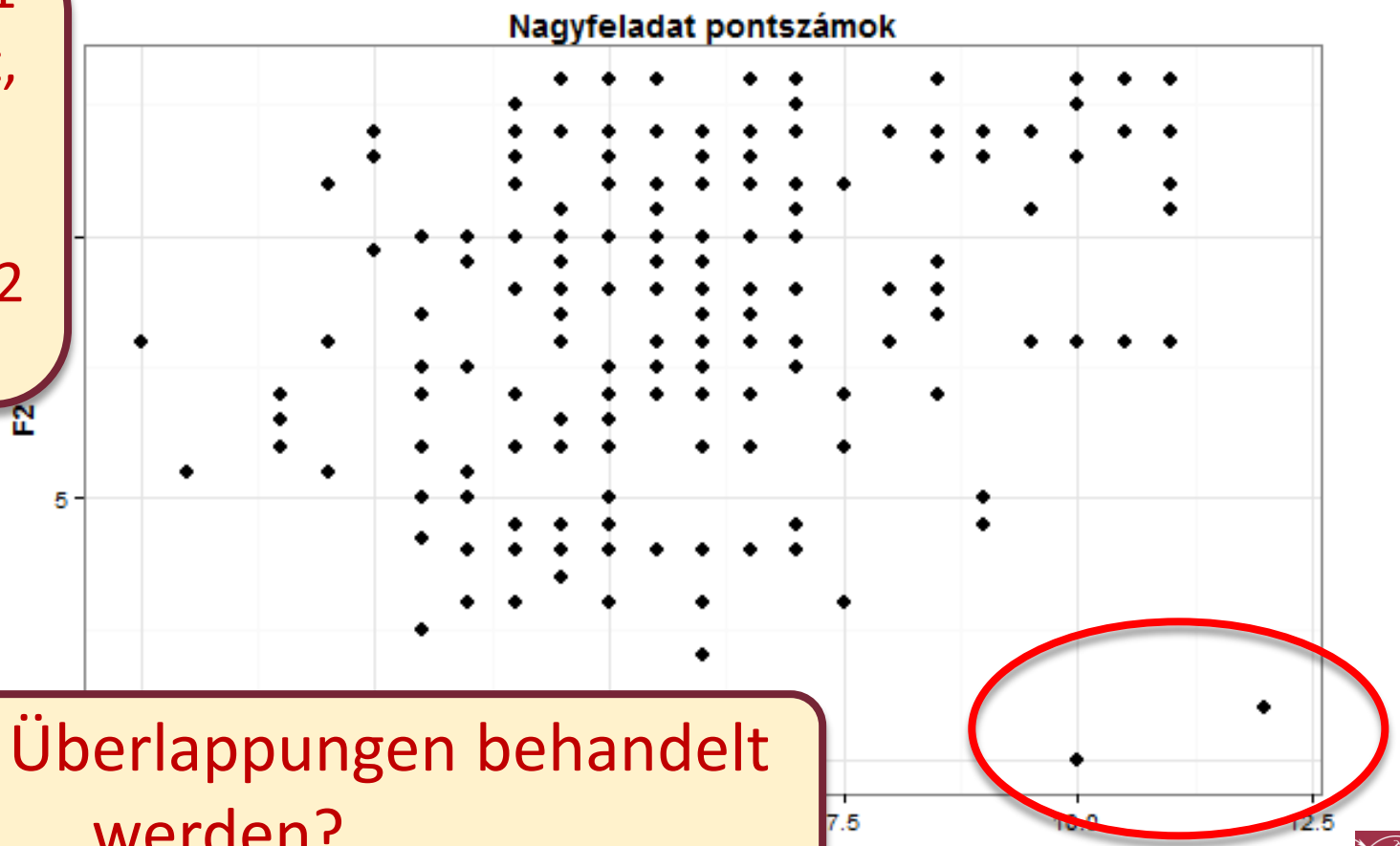
Wo das eine Ergebnis fehlt, kann das ganze nicht dargestellt werden. ☹



Streudiagramme (scatterplot)

- Eingabevariablen: Ergebnisse der Großaufgaben
- Frage: Was isr ihr Verhältnis?

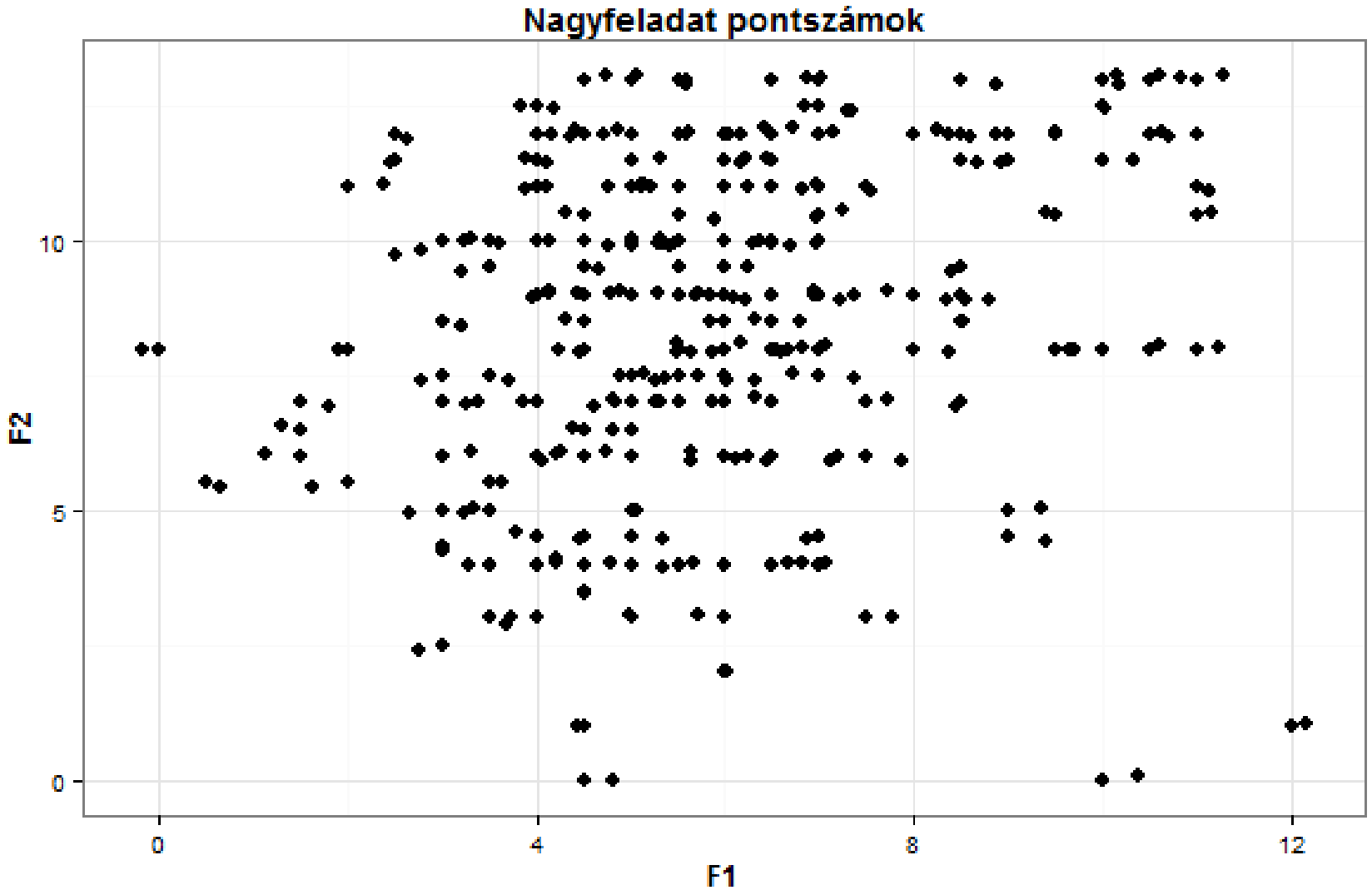
Wer Aufgabe1 gut gelöst hat, hat nicht unbedingt auch Aufgabe2 gut gelöst.



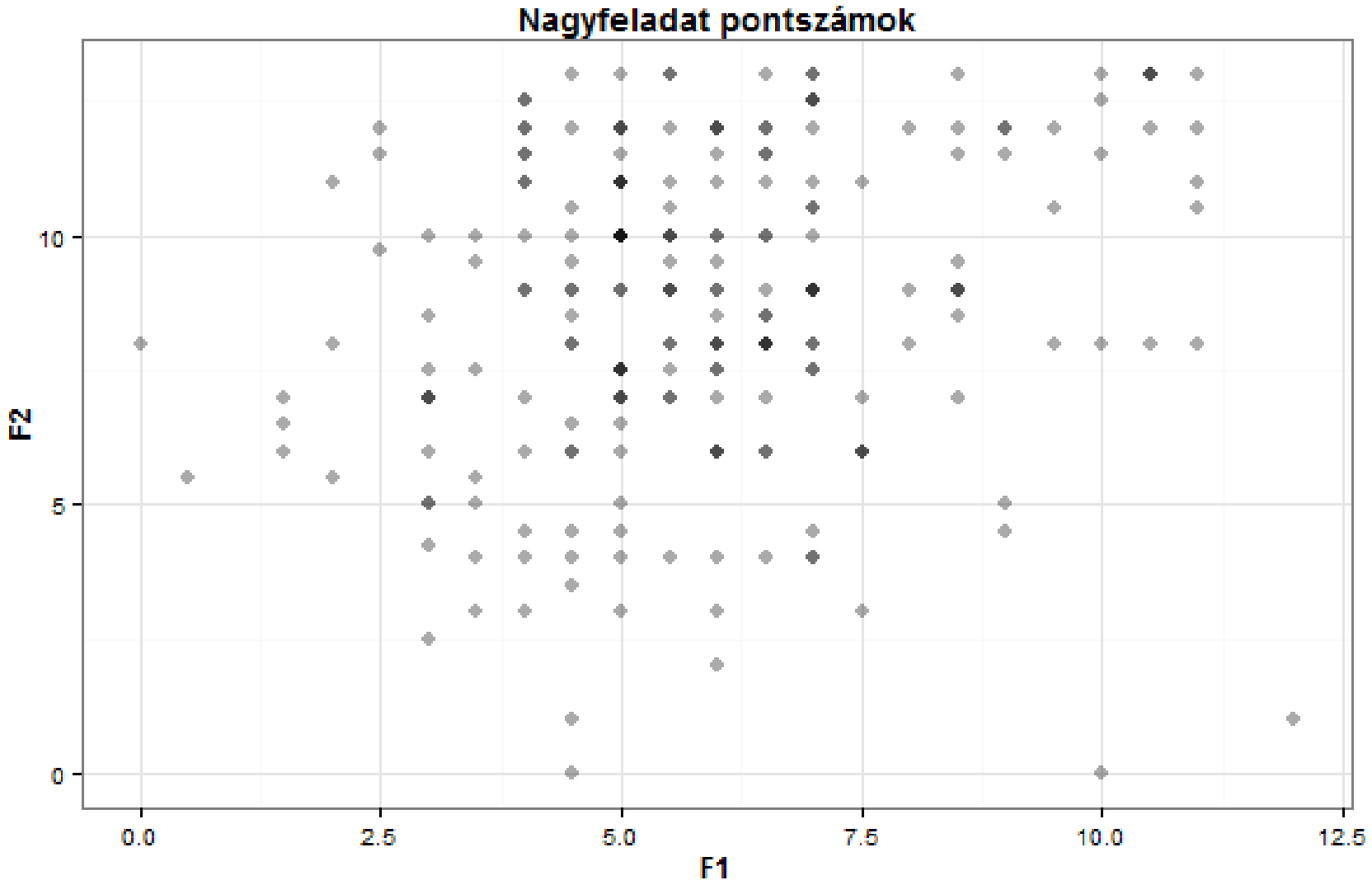
Wie sollten Überlappungen behandelt werden?

Overplotting

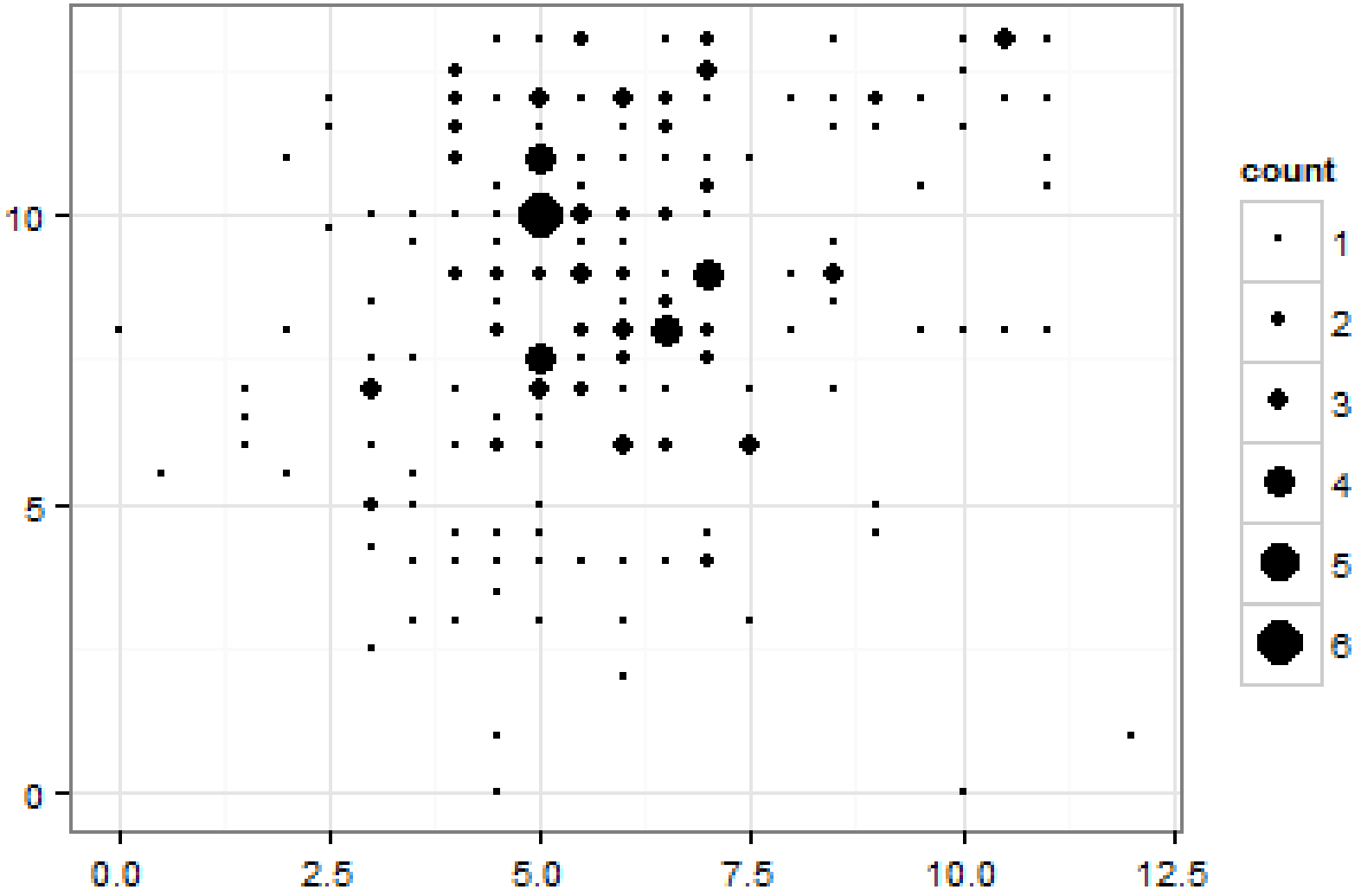
Overplotting – Lösung 1: jitter



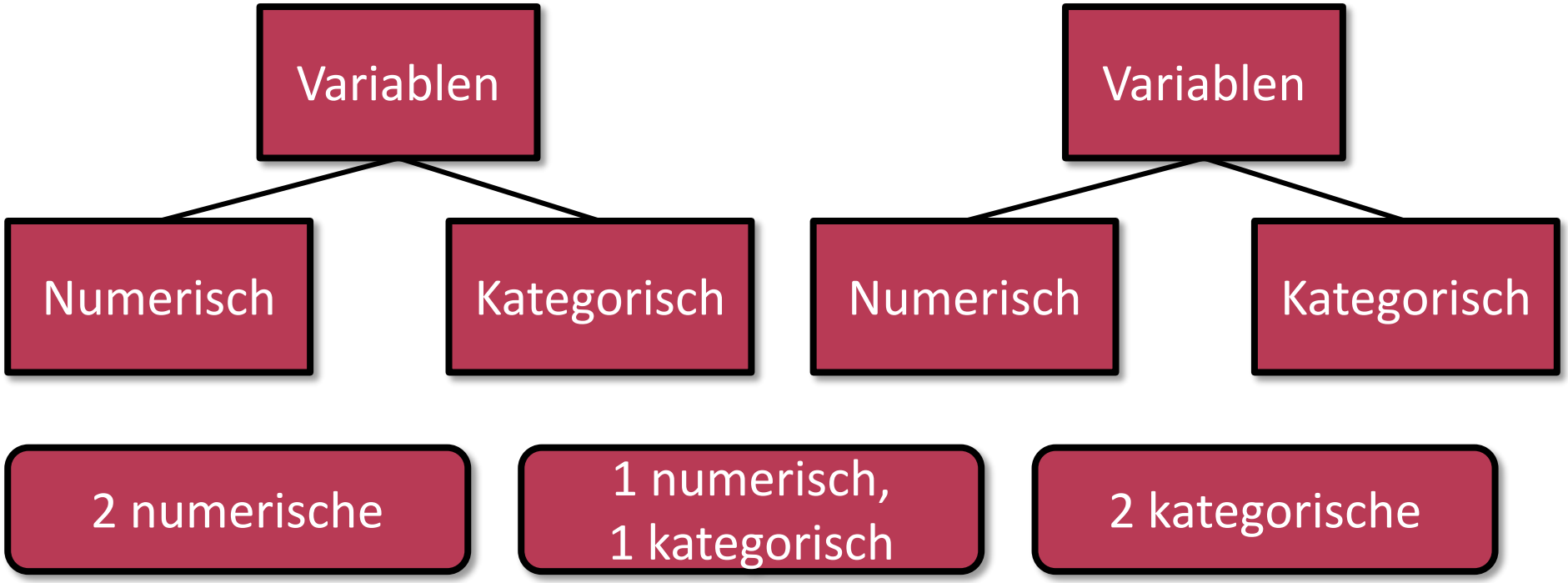
Overplotting – Lösung 2: Durchsichtigkeit



Overplotting – Lösung 3: Grösse



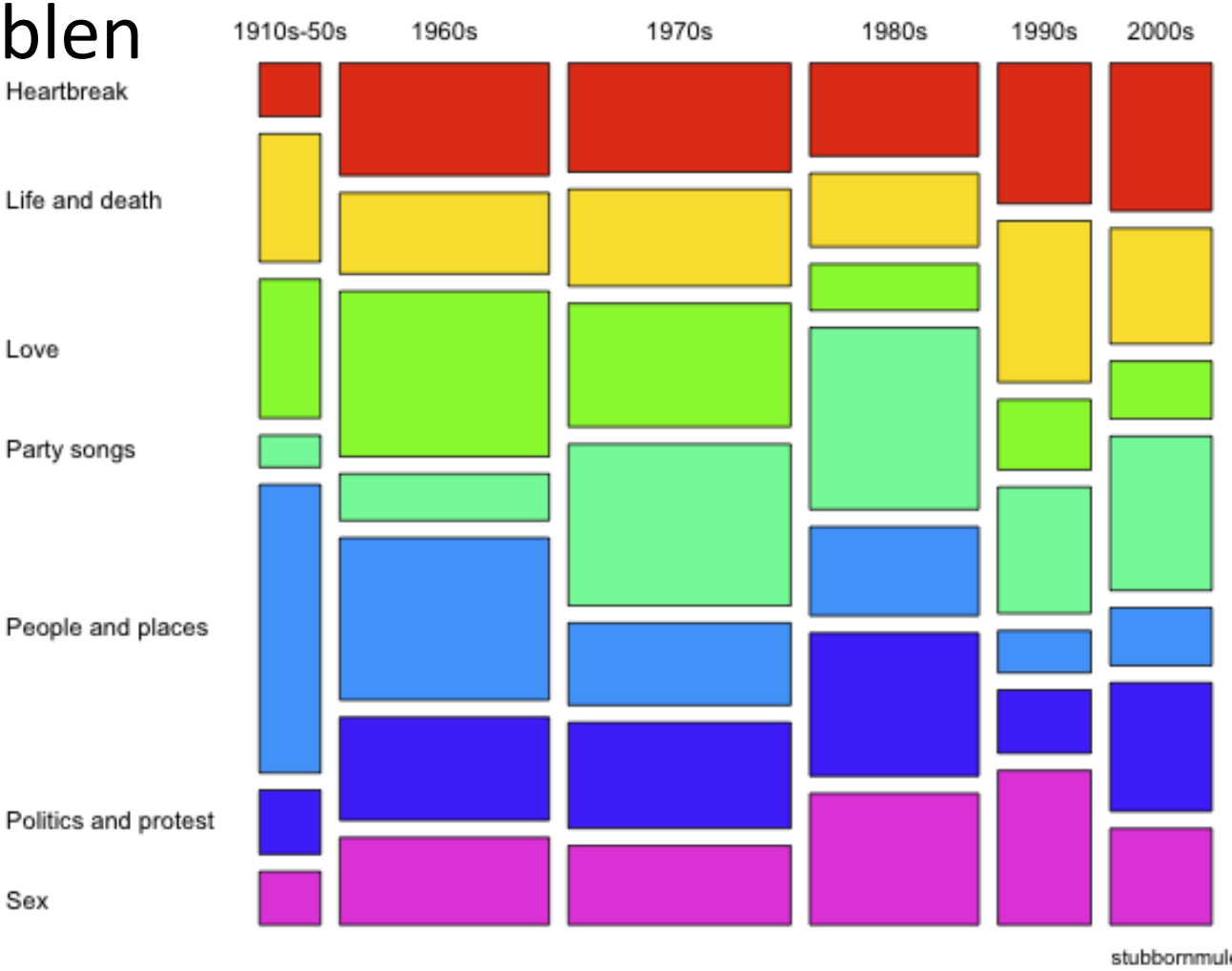
Zusammenhänge zweier Variablen



Wird jetzt nicht behandelt.

Mosaikplot

- Zusammenhänge zweier oder mehr kategorischen Variablen



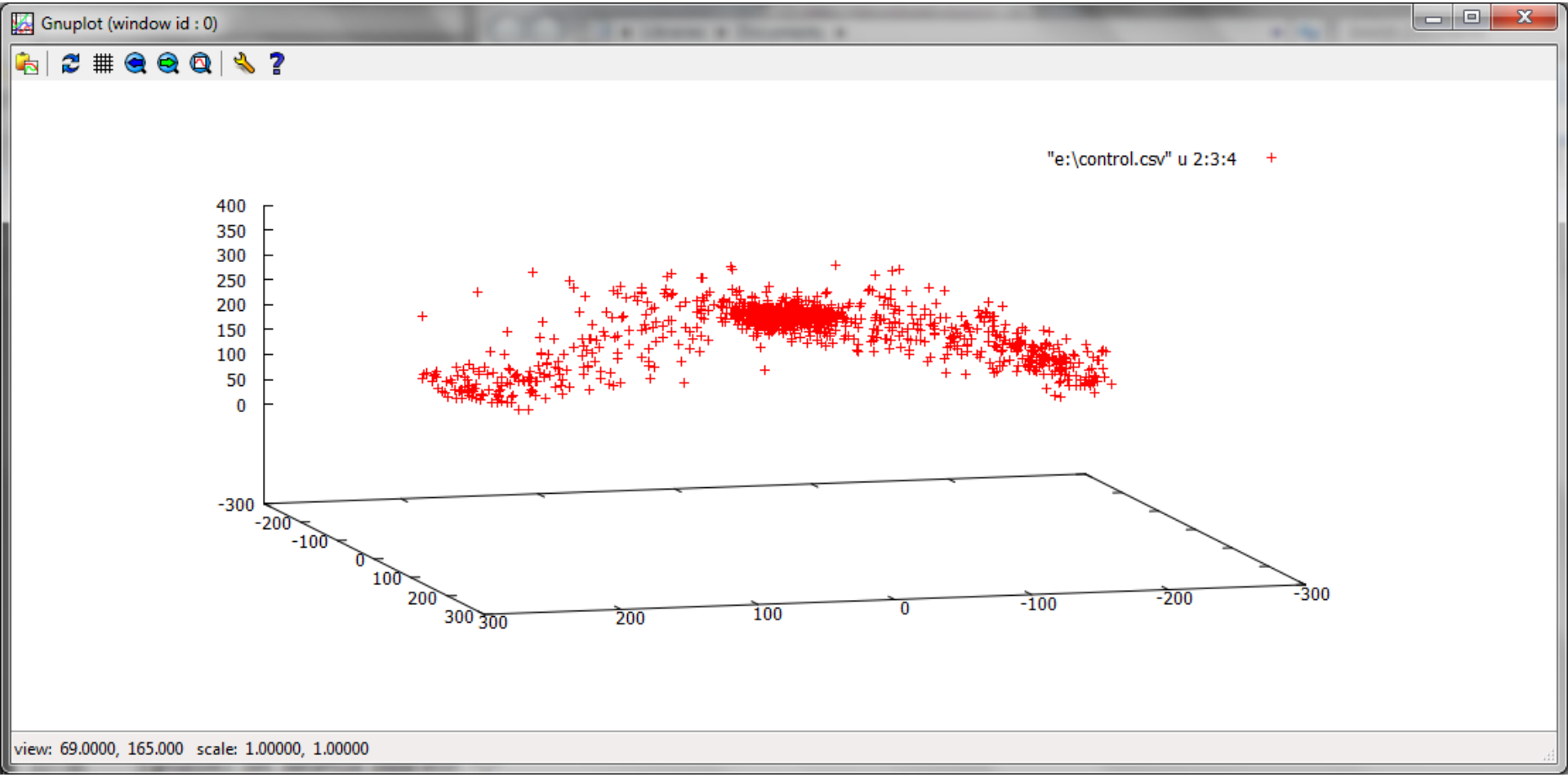
Guardian's list of "1000 songs to hear before you die"

MEHRERE VARIABLEN

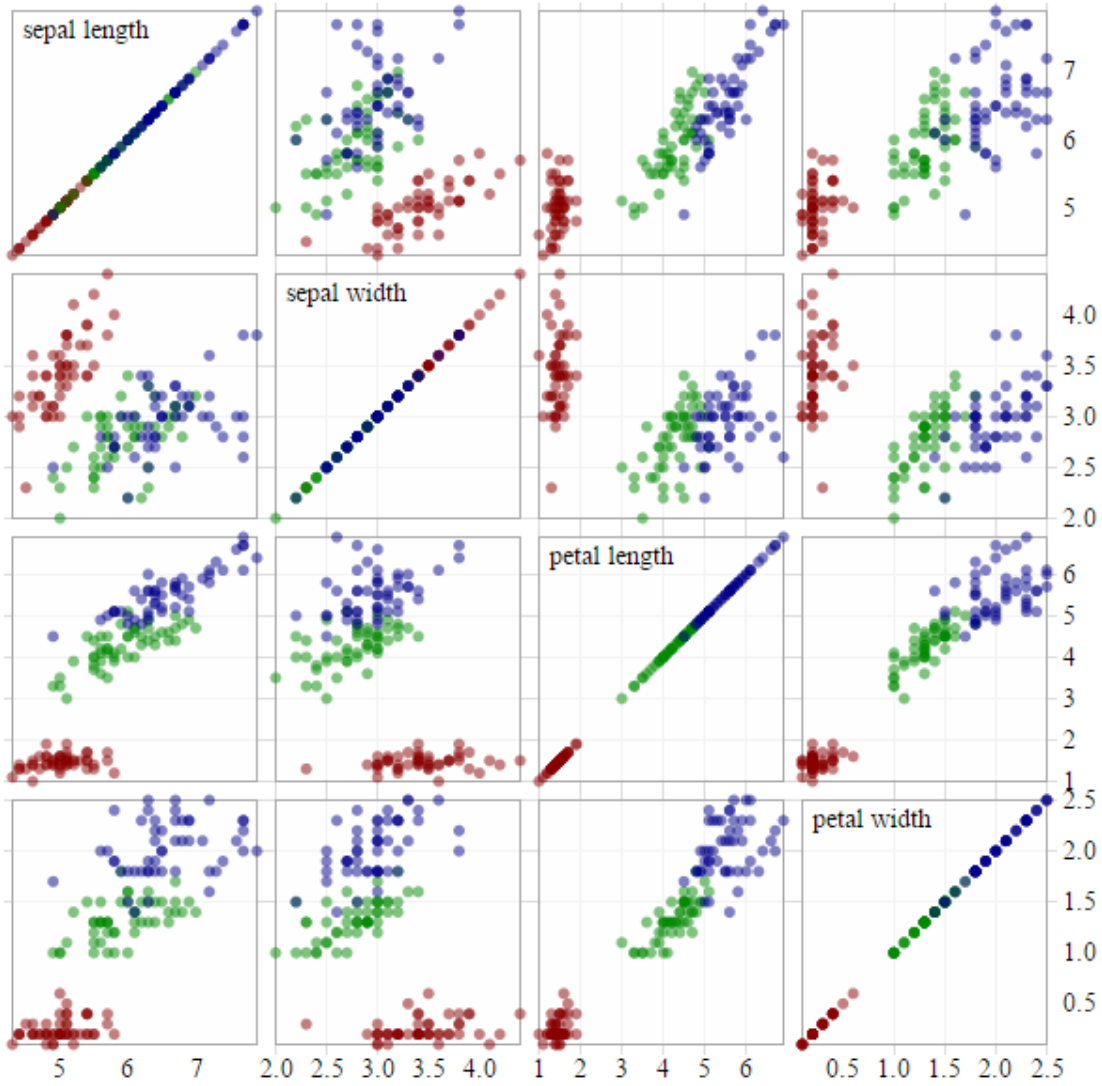
>=3 Variablen

- Wir ändern die Attribute der graphischen Objekte
  - Farbe
  - Grösse
  - Textur
  - Stelle – es kann auch trivial sein, aber bei „treemaps” hat Stelle eine konkrete Bedeutung
- Z.B. heatmap, bubble chart, treemap

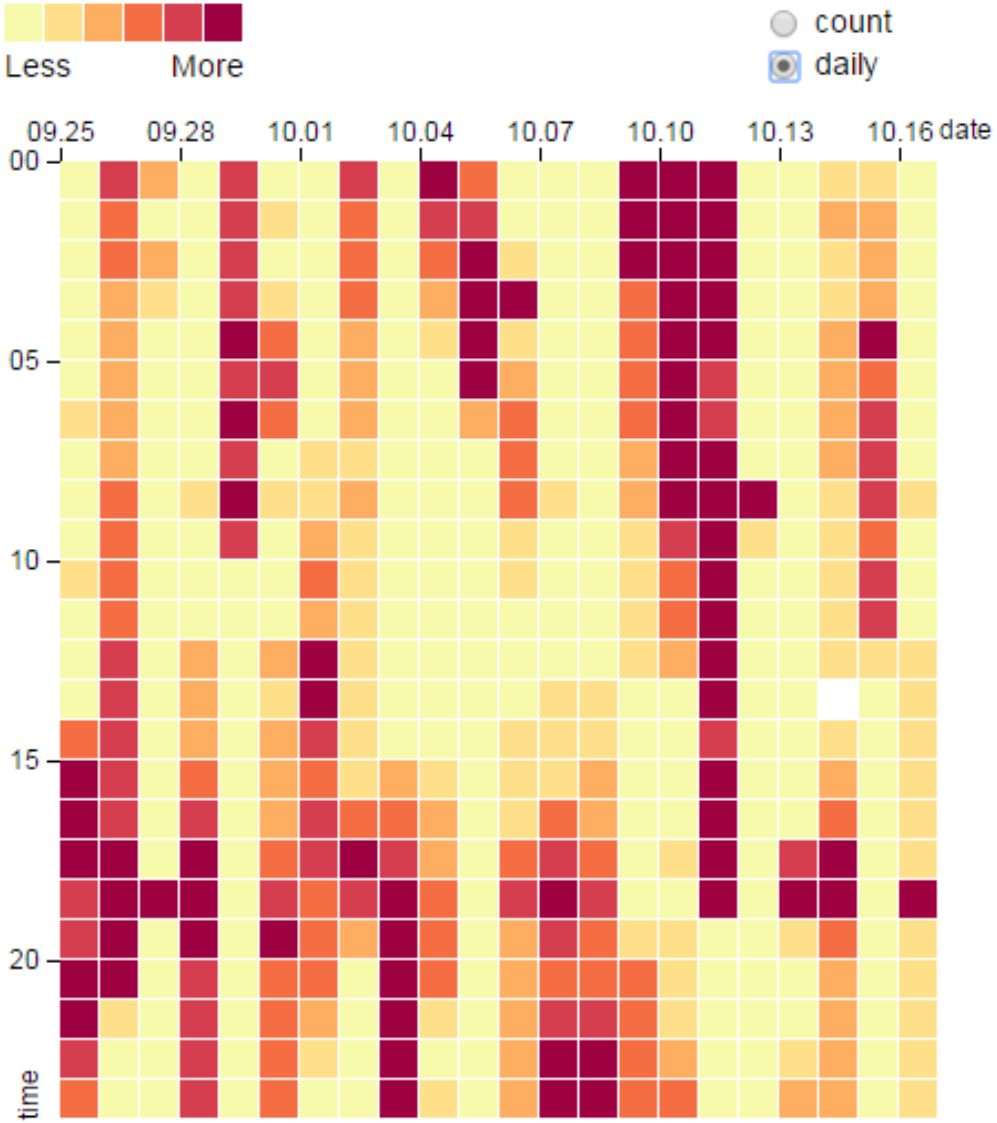
3D Plot


Gnuplot (window id : 0)
"e:\control.csv" u 2:3:4
400
350
300
250
200
150
100
50
0
-300
-200
-100
0
100
200
300
300
200
100
0
-100
-200
-300
view: 69.0000, 165.000 scale: 1.00000, 1.00000

Scatterplot Matrix

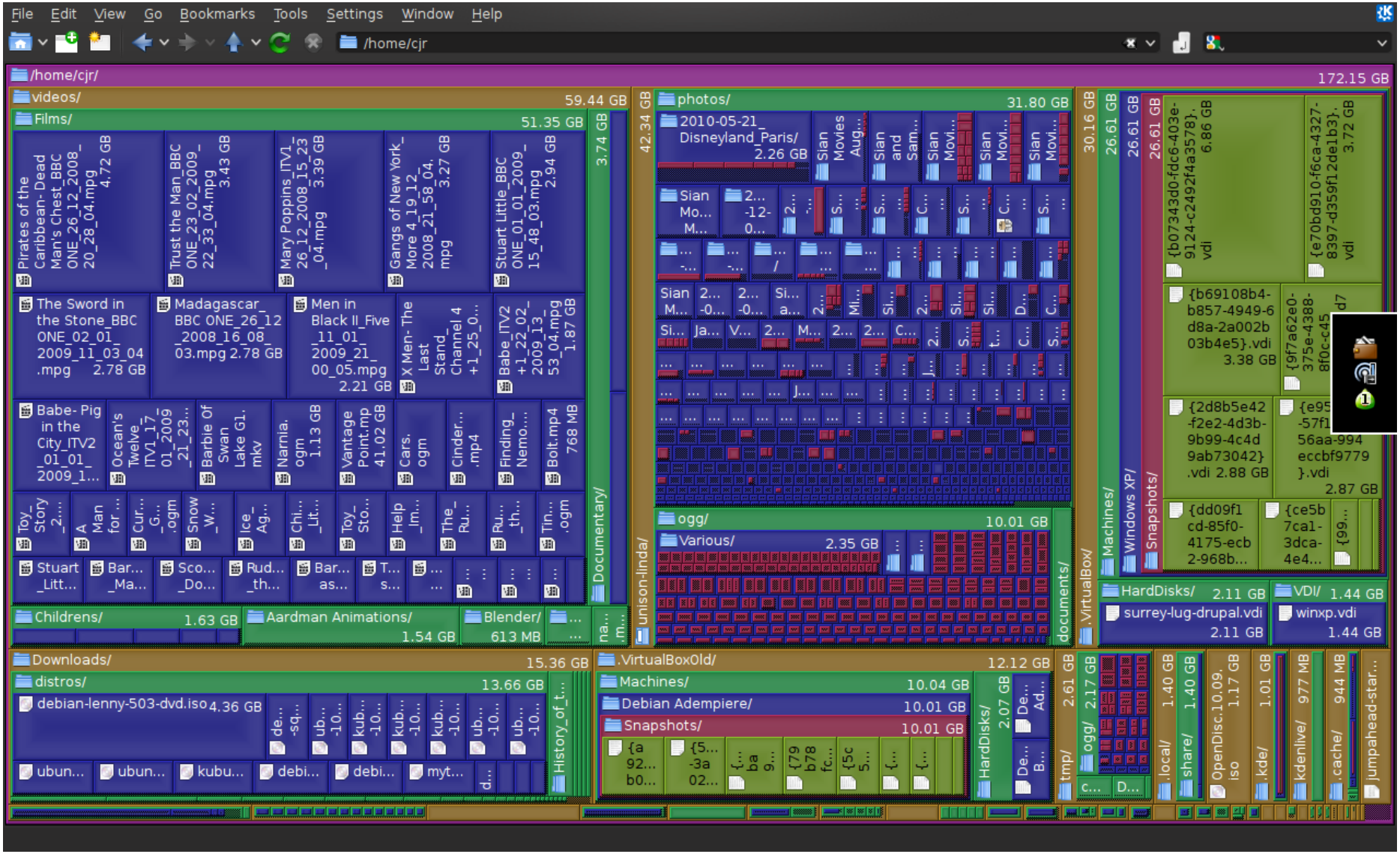


Heat Map


Less
More
count
daily
09.25
09.28
10.01
10.04
10.07
10.10
10.13
10.16
date
00
05
10
15
20
time

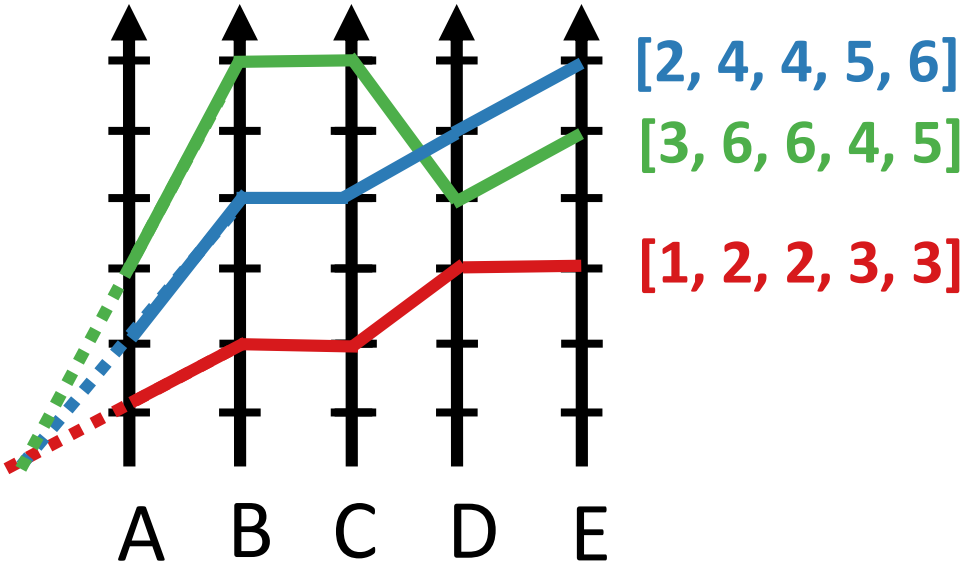
Bubble chart: Durchschnittsalter nach Regionen

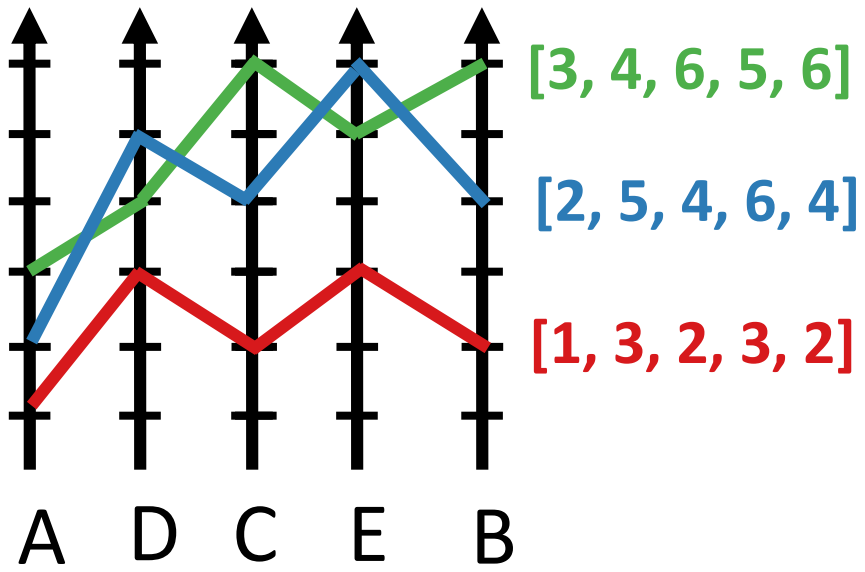
Treemap: Dateisystem



Parallele Koordinaten

- Multi-dimensionale Visualisation
- Kompakt, skalierbar
- Reihenfolge der Achsen?





Radar chart: Erweiterung der parallelen Koord.