

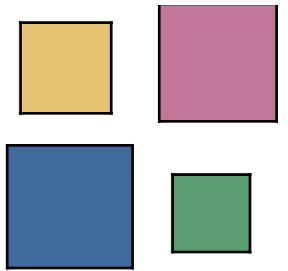
Datenanalyse

- Die Datenanalyse ist ein weitaus breiteres Feld als die Statistik
- Enthalten die Daten keine stochastische Komponente, so sind die klassischen statistischen Methoden nicht sehr nützlich

Beispiele:

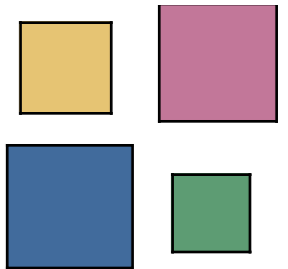
- Titanic Daten: “Stichprobe”?, Vollerhebung?, stat. Modell?
- Tour de France Daten: Reproduzierbarkeit?, Modelle?
- In allen Fällen ist dann eine graphische, deskriptive Betrachtung sehr hilfreich.
- Mögliches Modell:





Data Mining und KDD

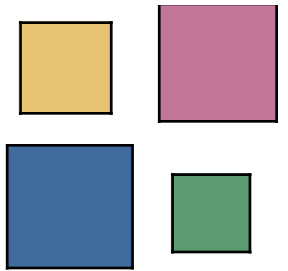
- **KDD** = Knowledge Discovery in Databases
- **Data Mining:**
 - Viele Definitionen, wenig Formalisierung
 - Meist große Datensätze
 - Keine a-priori Hypothese
 - Methoden aus der Informatik (oftmals in Unkenntnis der Statistik)
 - “Probleme an die sich klassisch ausgebildete Statistiker nicht herantrauen”
- KDD und Data Mining sind inzwischen Synonyme
- **Gemeinsamkeit zur EDA:**
 - Explorative Arbeitsweise
 - Mehr Graphik
- **Aber:**
 - die EDA basiert wesentlich auf den Grundlagen der Statistik



Schritte in der Datenanalyse

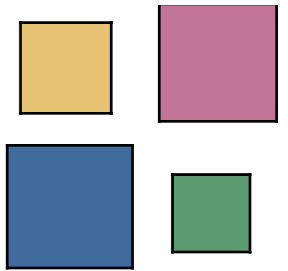
- **Beschaffung der Hintergrundinformation**
Grundlegendes Verständnis der Zusammenhänge die betrachtet werden sollen
- **Planung der Studie**
Welche Daten müssen wie erhoben werden? (Abhängig von den Fragestellungen)
- **Erhebung der Daten**
Tatsächliche Beschaffung der Daten
- **Überprüfung der Datenqualität**
Wie sehen ggf. Messfehler und/oder fehlende Werte aus? Korrekturen?
- **Bestimmung der Ziele der Analyse**
Was ist Ziel der Analyse? Welche Fragen sollten beantwortet werden?
- **Auswahl der Methoden**
Welche Methoden sollten zur Beantwortung benutzt werden?

...



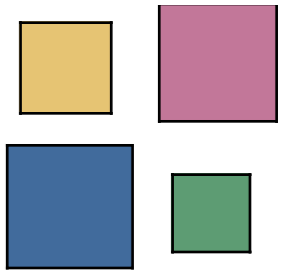
Schritte in der Datenanalyse

- **Exploration der Daten**
Umfassende Betrachtung der Daten mit im wesentlichen Mitteln aus der Statistik
 - Univariate Graphiken und Summaries
 - Untersuchung der Interaktionen
 - bivariat
 - multivariat
 - ggf. Tests und Modelle zur Überprüfung und Formulierung der Zusammenhänge (alternativ: Resampling Methoden)
- **Überprüfung der initialen Fragen**
Können die ursprünglichen Fragen beantwortet werden? Reichen Daten und Methoden dafür aus? Sind die Fragen noch relevant? Ergeben sich neue Fragen?
- **Erstellung des Modells**
Welches Model/Test ist des/der “Beste” um mein Problem zu lösen?
- **Überprüfung des Ergebnisses bzgl. Plausibilität und Relevanz**



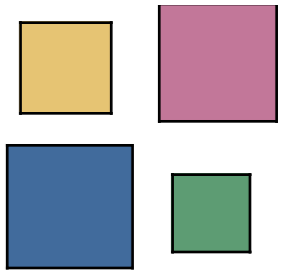
Datenanalyse: Bemerkungen

- Auch wenn die Schritte linear notiert sind, so ist der Ablauf meist iterativ.
- Im Kern steht die Exploration der Daten
- Im Gegensatz zur konfirmatorischen Statistik sind die Schritte innerhalb der Analyse nicht fest geplant.
- **Wichtig:**
Auswahl der relevanten Inhalte für den Report der Analyse
 - nicht alle Schritte der Exploration sind wichtig
 - **aber**, auch wichtige Negativ-Ergebnisse sollten erwähnt werden
- Sprache im Report
 - “So umgangssprachlich wie möglich, und so technisch wie nötig”
oder
“Man soll die Dinge so einfach erklären wie möglich, aber auch nicht einfacher”



Report-Writing

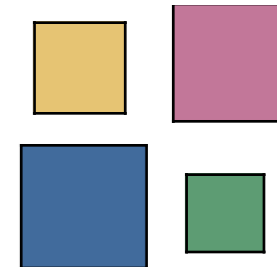
- Was gehört in einen Report:
 - Problembeschreibung
 - Beschreibung der Daten
 - Quellen
 - Variablen/Fälle
 - Ziel der Analyse
 - Beschreibung der nötigen Methoden
 - wichtige Ergebnisse der allgemeinen Analyse der Daten
 - Summaries und Graphiken,
 - uni-, bi- und multivariat
 - Beantwortung und Überprüfung der ursprünglichen Fragestellung (beinhaltet meist “den Test” bzw. “das Model”)
 - Zusammenfassung (zusätzliche Erkenntnisse, Relevanz)
 - Benutzte Quellen



Was man beim Report beachten sollte

- Aussagen im Text sollten sich immer auf entsprechende Tabellen oder Graphiken beziehen
- Diese Tabellen und Graphiken sollten möglichst nah zu ihrer Referenz platziert sein.
- Die Größe von Gruppen sollte immer in absoluten **und** relativen Zahlen angegeben werden.
- Summaries und Modelle sollten immer in sinnvollen Größen dargestellt werden.
(je nach Problem kann 1€ oder 1.000.000€ sinnvoll sein)
- Wenn man Daten in Gruppen oder Hierarchien zusammenfasst, sollte man sich möglichst an vorhandene Standards halten
(z.B. Bundesländer besser ost/west als nord/süd)

...



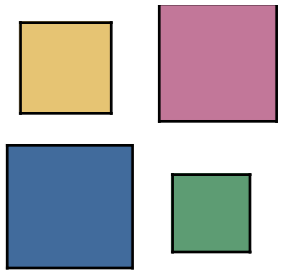
Reports (cont.)

- Bei komplexeren Tabellen ist das Layout sehr entscheidend für ihre Lesbarkeit:
 - zu vergleichende Werte sollten nah beieinander gehalten werden
 - Zeilen und Spalten sollten sinnvoll nach Kategorien und Werten sortiert werden
 - Hierarchien auf Zeilen und/oder Spalten sparen Platz

<i>Detergent usage</i>			<i>Watersoftness</i>		
<i>M-User?</i>	<i>Temperature</i>	<i>Preference</i>	<i>Hard</i>	<i>Medium</i>	<i>Soft</i>
no	low	X	68	66	63
		M	42	50	53
	high	X	42	33	29
		M	30	23	27
yes	low	X	37	47	57
		M	52	55	49
	high	X	24	23	19
		M	43	47	29

vs.

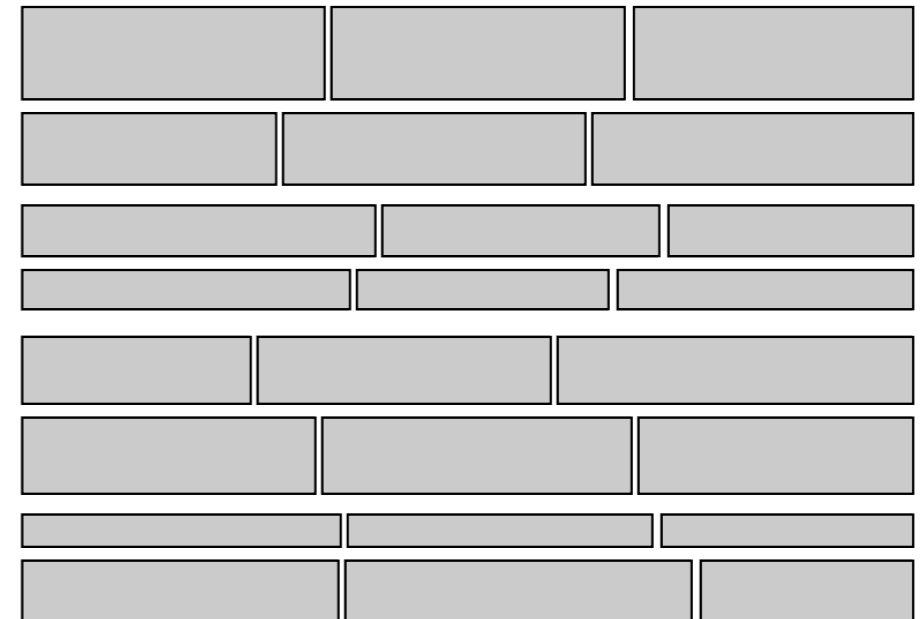
<i>Muser</i>	<i>Tempe</i>	<i>WaterSof.</i>	<i>Pref.</i>	<i>Number</i>
No	Low	Hard	X	68
No	Low	Hard	M	42
No	Low	Medium	X	66
No	Low	Medium	M	50
No	Low	Soft	X	63
No	Low	Soft	M	53
No	High	Hard	X	42
No	High	Hard	M	30
No	High	Medium	X	33
No	High	Medium	M	23
No	High	Soft	X	29
No	High	Soft	M	27
Yes	Low	Hard	X	37
Yes	Low	Hard	M	52
Yes	Low	Medium	X	47
Yes	Low	Medium	M	55
Yes	Low	Soft	X	57
Yes	Low	Soft	M	49
Yes	High	Hard	X	24
Yes	High	Hard	M	43
Yes	High	Medium	X	23
Yes	High	Medium	M	47
Yes	High	Soft	X	19
Yes	High	Soft	M	29



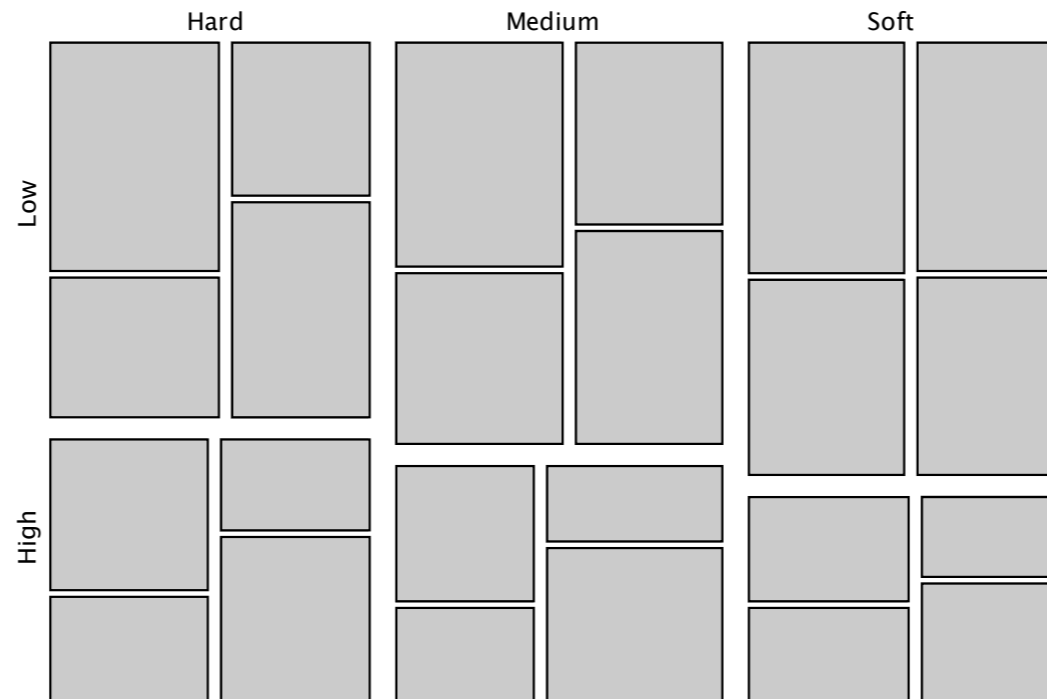
Reports (cont.)

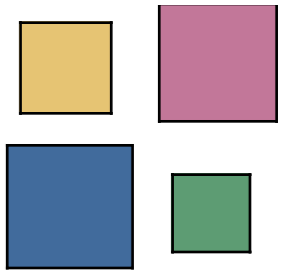
- Tabellen kommen schnell an ihr Ende bzgl. der Interpretation:

<i>Detergent usage</i>			<i>Watersoftness</i>		
<i>M-User?</i>	<i>Temperature</i>	<i>Preference</i>	<i>Hard</i>	<i>Medium</i>	<i>Soft</i>
no	low	X	68	66	63
		M	42	50	53
	high	X	42	33	29
		M	30	23	27
yes	low	X	37	47	57
		M	52	55	49
	high	X	24	23	19
		M	43	47	29



besser ⇒

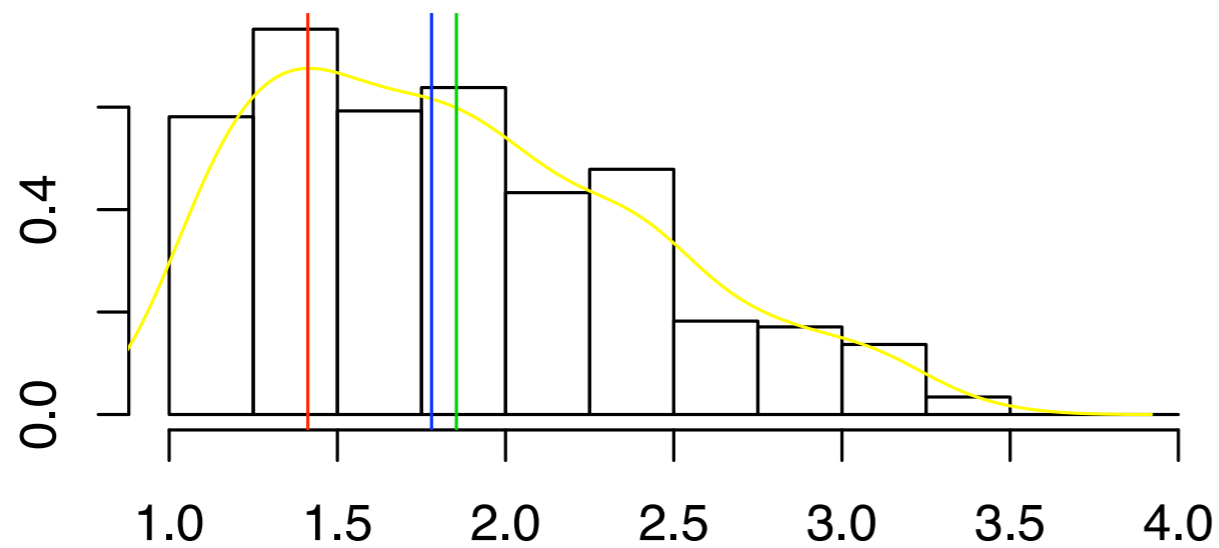




Reports (cont.)

- Übliche Probleme bei Gruppenvergleichen

- Robustheit von Lagemaßen



- Verteilung einer Untergruppe

