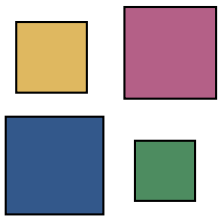


Prof. Antony Unwin, Alexander Pilhöfer
Lehrstuhl für Rechnerorientierte Statistik und Datenanalyse
Institut für Mathematik
Universität Augsburg
http://stats.math.uni-augsburg.de/

Statistik I
Übungsblatt 2

Abgabe: *Mittwoch 02. Mai 2012, bis spätestens 12.00 Uhr*; Briefkasten: Statistik I oder per email an die Übungsleiter

Die Aufgaben können auch in 2er-Gruppen bearbeitet und abgegeben werden!

1. (a) Laden Sie den Datensatz `Cars` in R. Welcher Anteil Fahrzeuge hat mehr als 300 PS? **(1P)**
   (b) Was ist der durchschnittliche Händlerpreis für diese Autos? **(1P)**
   (c) Was ist der Median der Leistung für diese Autos und welchem Quantil entspricht dies in der Gesamtstichprobe? **(1P)**
   (d) Berechnen Sie, wie groß der Prozentsatz der Werte einer Normalverteilung ist, die in einem Boxplot als Ausreißer bzw. krasse Ausreißer klassifiziert werden. **(1P)**

2. **Zehnkampf (5P)**

   Laden Sie den Datensatz `Zehnkampf2011` (von der Internetseite der Vorlesung) in `Mondrian`.

   (a) Erstellen Sie in MONDRIAN parallele Boxplots
   - der Punkte-Variablen
   - der Resultate

   Welche interaktiven Optionen bietet die Grafik?

   (b) Welche Optionen und Anordnungen würden Sie nutzen, um die Daten in den beiden Fällen zu interpretieren? Welche Schlüsse ziehen Sie?
   (c) Welche Disziplinen haben den größten Einfluss auf das Gesamtergebnis?

   Ein alternatives System schlägt vor, die rohen Daten zu standardisieren, um so die Punkte zu erhalten: $P = (X - \overline{X})/s_X$.

   (d) Berechnen Sie diese Variante in R und vergleichen Sie die beiden Herangehensweisen: führen sie zum selben Ergebnis? Falls nicht, wo liegen die Unterschiede?
   (e) Welches System würden Sie bevorzugen und warum?

3. **Cars I (5P)**

   Laden Sie den Datensatz `Cars` (von der Internetseite der Vorlesung) in R und `Mondrian`.

   (a) Erstellen Sie ein Histogramm der Variable "Dealercost". Variieren Sie die Binbreite und den Ankerpunkt. Welche Wahl würden Sie treffen? Welche Verteilung könnte auf diese Variable passen? Überprüfen Sie Ihre Vermutung in R mit einem einen geeigneten QQ-plot. Interpretieren Sie den Plot!
   (b) Erstellen Sie für die Variablen "Horsepower", "DealerCost" und "City Miles Per Gallon" paarweise Scatterplots (Streudiagramme). Gibt es Ausreißer? Vermuten Sie einen funktionalen Zusammenhang zwischen den Variablen?

(c) Erstellen Sie einen Barchart für die Variable "Number of Cylinders". Färben Sie im Barchart die einzelnen Zylindergruppen in verschiedenen Farben. Zeigen die Scatterplots ein konsistentes Muster? Welche Fälle würden Sie nun als Ausreißer bezeichnen?

(d) Löschen Sie die bisherige Färbung. Visualisieren Sie durch eine geeignete Färbung die proportionalen Verläufe der 4- und 8-Zylindergruppen in einem Spinogramm der Variablen "Highway Miles Per Gallon". Interpretieren Sie diesen Plot.

(e) Untersuchen Sie auf ähnliche Weise, für welche Fahrzeuge (Typ, Antrieb, Zylinder) die Preise am deutlichsten unter dem UVP-Preis liegen. (In Mondrian können Sie das Menü *Calc* verwenden).

4. **Cars II (5P)**

(a) Untersuchen Sie die Korrelationen zwischen den Variablenpaaren. Lassen Sie sich eine auf 2 Nachkommastellen gerundete Korrelationsmatrix ausgeben. Welche Variablen sind am stärksten korreliert?

(b) Erstellen Sie eine Scatterplotmatrix und ggf. geeignete einzelne Streudiagramme um die paarweisen Zusammenhänge genauer unter die Lupe zu nehmen. Sind die berechneten Korrelationen in allen Fällen aussagekräftig?

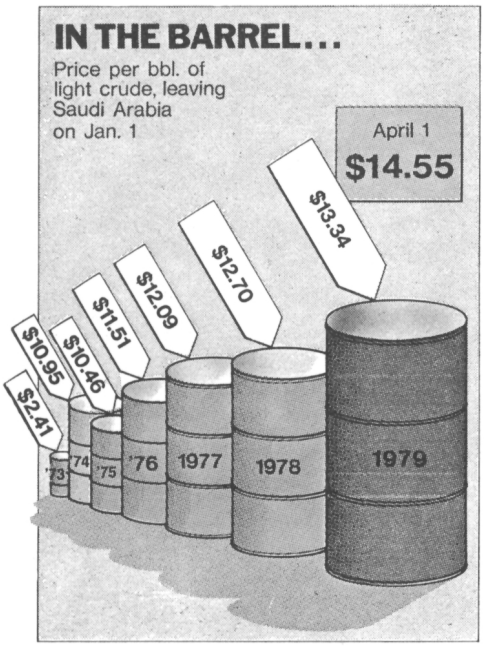
(c) In welchen Fällen würde durch Transformation einer der beteiligten Variablen eine höhere Korrelation erreicht werden? Berechnen Sie eine entsprechende Transformation und die resultierende Korrelation!

(d) Untersuchen Sie die paarweisen Zusammenhänge zwischen den kategorialen Variablen *Type, Drive* und *Number of Cylinders*. Sind die Variablen unabhängig? Wie visualisieren Sie dies und anhand welcher Größen treffen Sie Ihre Aussage?

5. **Barrel (5P)**

Analysieren Sie die nebenstehende Graphik.

- Was wird hier Ihrer Meinung nach dargestellt?
- Ist die Visualisierung der Zahlen gelungen? Schätzen Sie den (wahrgenommenen) Unterschied zwischen 1973 und 1979.
- Messen Sie die Höhen der Fässer. Berechnen Sie die mit Farbe ausgefüllten Frontflächen sowie die Volumina der Zylinder. Wozu sollten die Preise proportional sein?
- Was fällt Ihnen sonst noch auf?



Ölpreise im Zeitraum von 1973 bis 1979 (entnommen aus Tufte, S. 62).

Tipp für weitere Übungsblätter:
Installieren Sie in R das Paket *Rserve*, um in MONDRIAN mehr Optionen zur Verfügung zu haben:

```
install.packages("Rserve")
```

Nun sollten in Mondrian auch Dichteschätzer, CD-plots, Glättungen usw. zur Verfügung stehen.