Übung 02 - Datenanalyse und Statistik WS 2015/2016

Aufgabe 1: Laden Sie die Datei "Uebung2.R" von der Website herunter, wo sich auch die Übungsblätter befinden. Öffnen Sie es in R und führen jede Zeile aus. Es gibt auch da 3 kleine Aufgaben zu lösen.

Aufgabe 2: Über *Old Faithful*
Laden und speichern Sie die Textdatei "OldFaithful". Lesen Sie den Datensatz anschließend mit folgendem Befehl in R ein:

```
setwd("...")  # ... steht fuers Arbeitsverzeichnis
( old <- read.table(file="OldFaithful.txt") )
```

Der Geysir Old Faithful im Yellowstone-Nationalpark ist bekannt für die vergleichsweise geringe Variationsbreite der Wartezeit, die zwischen zwei Wasser-Eruptionen vergeht. Der Ihnen vorliegende Datensatz enthält zwei Messreihen an Wartezeiten (in Minuten), die zu zwei verschiedenen Zeitabschnitten ermittelt wurden.

a. Welche Skala hat jede Variable?

b. Für welche Grundgesamtheit(en) könnten diese Daten representativ sein? Was könnte an diesem Datensatz besonders problematisch sein in Hinblick auf die Representativität?

c. Stellen Sie jede Variable mit den passenden Diagrammen dar. Gibt es potenziell problematische Werte? (z.B.: Nulls, Bindungen, Ausreißer, fehlende Werte)

d. Stellen Sie die Abhängigkeit zwischen beide Variablen mit den folgenden Befehlen:

```
boxplot(wtime~period, old)
spineplot(period~wtime, old)
```

Welche von den 2 Diagramme ist für diesen bestimmten Datensatz geeigneter?

e. Ist die Verteilung uni- oder bimodal? Für beide Perioden?

Aufgabe 3: Laden sie das Beispieldatensatz "iris" aus R (Befehl `data`).

a. Diskutieren Sie die Skala jeder Variable. Stellen sie jede Variable unabhängig von den anderen auf der beste Weise dar.

b. Characterisieren Sie die Abhängigkeit zwischen Breite und Länge von Kelchblat. Soll man die Spezies berücksichtigen? Wie kann man das bekommen?

c. Characterisieren Sie die Abhängigkeit zwischen Breite und Länge von den Blütenblätter. Soll man auch hier die Spezies berücksichtigen?

d. Stellen Sie ein Diagramm von den Kelchblattgrößen dar, wo Farbe die Spezies beschreibt, und die Größe der Symbole die Blattlänge darstellt. Fügen Sie eine Legende (Befehl `legend`) hinzu

e. Stellen Sie ein Diagramm von allen stetigen Variablen dar (Befehl `plot`), wo Farbe die Spezies beschreibt.

Graphische Darstellung Zusammenfassung von R-Funktionen, Prinzipien und Tricks zur effektive Darstellung von Daten:

Merkmal	diskrete Variable	stetige Variable
Streu und Verteilung	Balkendiagramm (`barplot`)	Punktdiagramm (`stripchart`) Kastendiagramm (`boxplot`) Histogramm (`hist`) emp. Verteilungsfunktion (`ecdf`)
2D-Abhängigkeit von		
diskrete Variable	gestapelte oder parallele Balkendiagramme Mosaik (`mosaicplot`)	parallele Kastendiagramme, Histogramme, usw. (eventl. Farbe nutzen)
stetige Variable	`spineplot`,`cdplot`	Streudiagramm (`plot`)
extra Abhängigkeit von		
diskrete Variable	hD-Mosaik; Nutzung von (kontrastreiche) Farben	parallele Diagramme Farbe & Symbole nutzen
stetige Variable	lattice plots; Nutzung von Farbskalen & Symbolgröße	