Dieses Aufgabenblatt soll Sie mit der Struktur und Besonderheiten longitudinaler Daten vertraut machen, sowie deren graphische Darstellungsmöglichkeiten in `R` vermitteln. Die zu bearbeitenden Aufgaben beziehen sich auf die Inhalte der ersten und zweiten Vorlesungsfolien.

*Hinweis:* Aufgrund der Fülle an Stoff ist die Übung so konzipiert, dass vorausgesetzt wird, dass zumindest versucht wurde die Aufgaben **im Voraus** zu bearbeiten.

**Aufgabe 1** *Korrelation longitudinaler Daten: Grundlagen und Notation*

Sei $\boldsymbol{\xi} \in \mathbb{R}^n$ ein Zufallsvektor mit $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$ und $\mathbf{V} = \text{Cov}(\boldsymbol{\xi})$ und sei $\mathbf{A}$ eine $m \times n$ - Matrix.

a) Zeigen Sie, dass $\text{Cov}(\mathbf{A}\boldsymbol{\xi}) = \mathbf{AVA}^\mathsf{T}$.

Sei nun eine Zufallsstichprobe der Länge $n$ für ein Individuum $i = 1$ gegeben mit Response $\mathbf{Y}_i = (Y_1, ..., Y_n)^\mathsf{T}$ und Kovariablen-Designmatrix

$$\mathbf{X}_i = \begin{pmatrix} 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in} \end{pmatrix}$$

mit $x_{ij} \in \mathbb{R},\ j = 1, ..., n$. Wir betrachten das lineare Modell

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\xi}_i$$

mit Koeffizientenvektor $\boldsymbol{\beta} \in \mathbb{R}^2$ und $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{in})^\mathsf{T}$ Zufallsvektor mit Kovarianzmatrix $V_i$. *Hinweis: In der Aufgabe gehen wir vereinfachend davon aus, dass nur Daten eines einzigen Individuums erhoben wurden. Der Individuen-Index 'i' wird zur Vorbereitung auf die Vorlesung dennoch eingeführt.*

b) Berechnen Sie $\text{Cov}(\hat{\boldsymbol{\beta}})$ für den einfachen kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$.

c) Seien $\xi_{i1}, ..., \xi_{i(n-1)}$ i.i.d. standardnormalverteilt und $\xi_{in} = \frac{1}{n-1}\sum_{j=1}^{n-1} \xi_{ij} \prod_{g=1}^{n-1} \text{sgn}(\xi_{ig})$. Zeigen Sie, dass $\text{Cov}(\xi_{ij}, \xi_{in}) = 0$ für alle $j = 1, ..., n-1$, falls $n \geq 4$. Was können Sie daraus über den Zusammenhang von Korrelation und stochastischer Abhängigkeit folgern?

d) Sei $\boldsymbol{\xi}_i$ multivariat normalverteilt und $\text{Cov}(\boldsymbol{\xi}_i)$ Diagonalmatrix. Zeigen Sie, dass $\xi_{i1}, ..., \xi_{in}$ unabhängig sind.

e) Sei nun $\boldsymbol{\xi}_i = \mathbf{X}_i\mathbf{b}_i$ für einen Zufallsvektor $\mathbf{b}_i \in \mathbb{R}^2$ mit $\mathbb{E}(\mathbf{b}_i) = \mathbf{0}$ und $\text{Cov}(\mathbf{b}_i) = \mathbf{D}$. Fertigen sie eine Skizze an, wie eine Realisation des Modells für $n = 5$ aussehen könnte und bestimmen Sie $\text{Cov}(\hat{\boldsymbol{\beta}})$. Nach wie vielen Messungen dürfte sich in diesem Modell die Schätzung (idealerweise) nicht mehr verändern? Inwiefern spiegelt ihr Ergebnis für die Kovarianz dies wieder?

**Aufgabe 2** *Visualisierung longitudinaler Daten*

In dieser Aufgabe beschäftigen wir uns mit dem Datensatz `rats`. Lesen Sie hierzu zunächst die Beschreibung des Datensatzes (auf der Homepage) durch.

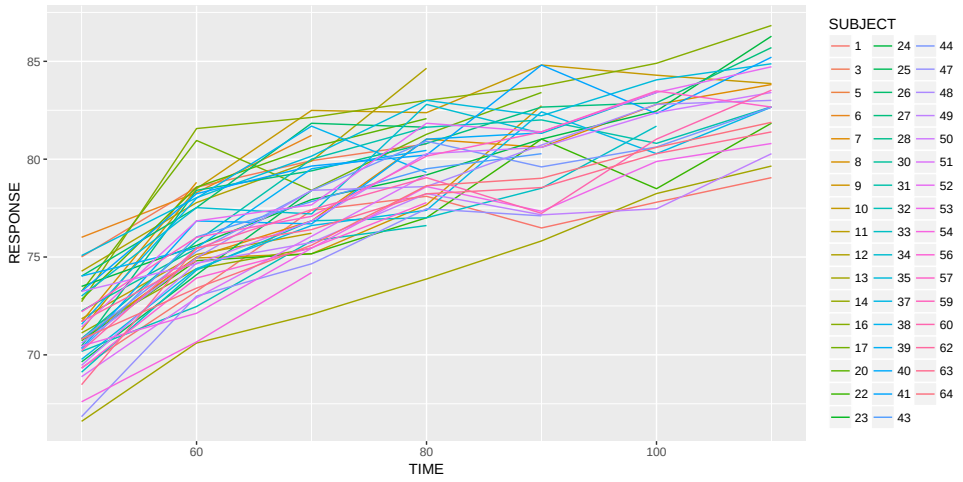
a) i) Laden Sie den Datensatz `rats` von der Homepage herunter und lesen Sie ihn mit `read.csv2` in `R` ein. Wandeln Sie die Variablen `GROUP` und `SUBJECT` in Faktorvariablen mit passenden Labels um.

ii) Wandeln Sie den Datensatz in 'Long-Format' um. D.h. formatieren Sie ihn so, dass sich im Datensatz eine Zeile pro Messung befindet und dass die Messzeitpunkte in einer Spalte `TIME` und die Response-Werte in einer Spalte `RESPONSE` stehen. Nennen Sie den umformatierten Datensatz `rats.long`.
*Hinweis: Sie können hierfür die Funktion `melt` im R-Paket `reshape2` verwenden.*

b) Im Folgenden wurde der Datensatz mit verschiedenen Zielsetzungen mit Hilfe des `R`-Pakets `ggplot2` visualisiert. Die gezeigten Plots weisen aber noch Probleme auf. Identifizieren Sie diese und sammeln Sie Verbesserungsideen. Setzen Sie jeweils eine Idee in `R` um.

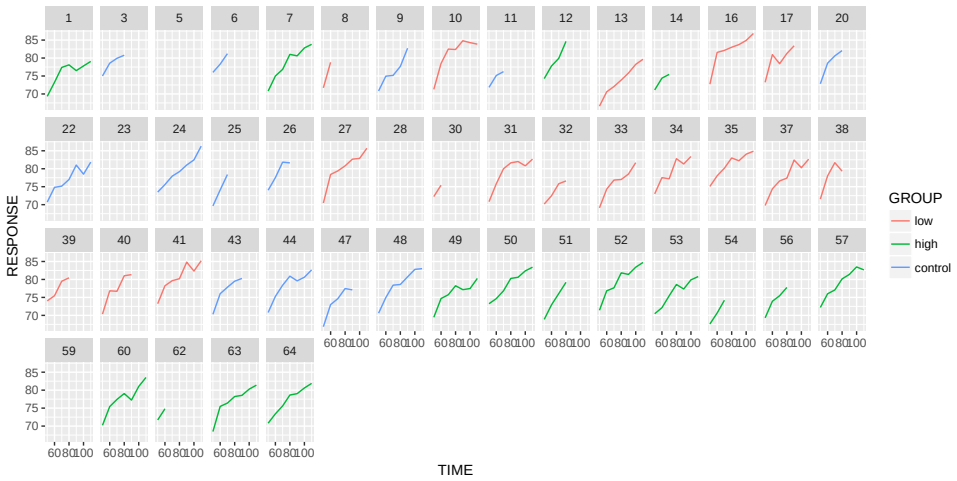
i) Zielsetzung: Sie wollen die Datenstruktur (bspw. einem Kollegen) anhand eines Plots erklären.

```
library(ggplot2)
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = SUBJECT)) +
  geom_line()
```



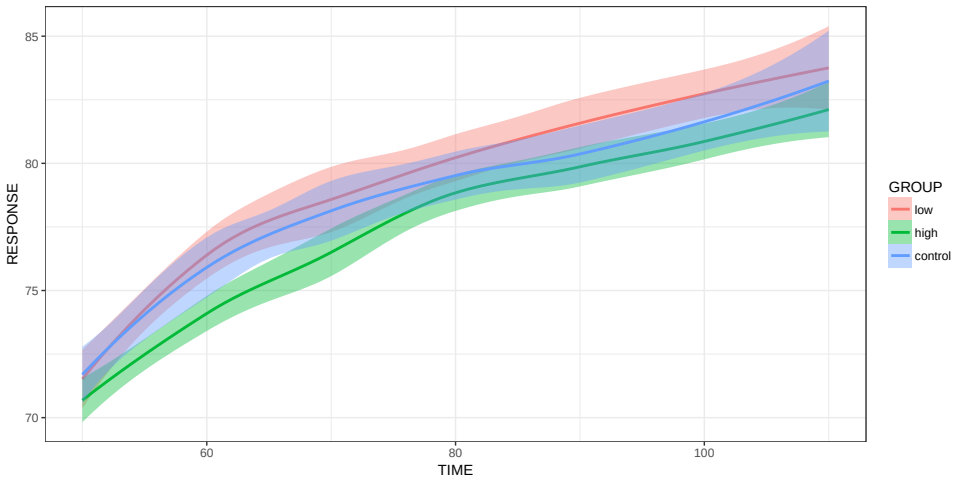
ii) Zielsetzung: Sie wollen Strukturen in den Daten mit einem der Biologen besprechen, der am Versuch beteiligt war und sich möglicherweise an Details einzelner Tiere erinnern kann.

```
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = GROUP)) +
      geom_line() + facet_wrap(~SUBJECT, ncol = 15)
```




iii) Zielsetzung: Sie wollen Gruppenunterschiede mit Hilfe von glatten Ausgleichskurven veranschaulichen.

```
ggplot(rats.long, aes(x = TIME, y = RESPONSE, col = GROUP, fill = GROUP)) +
  stat_smooth() + theme_bw()
```




c) Erstellen Sie einen Lasagna-Plot der Daten. Welche Vor- und Nachteile sehen Sie beim Lasagna-Plot gegenüber den üblichen Plots wie oben?