Aufgabe 4.7: Histogramme, Säulen- und Streudiagramme mit R (5 Punkte) (STO-deskrSta13 | 1)

(i) Zeichnen Sie Histogramme für die Merkmale `mpg, cyl, hp, wt` der R-internen Datentabelle *mtcars*.

(ii) Finden Sie heraus, für welche der Merkmale `mpg, cyl, hp, wt` ein Säulendiagramm sinnvoll ist.

(iii) Zeichnen Sie für die in (ii) ermittelten Merkmale Säulendiagramme.

(iv) Zeichnen Sie jeweils ein Streudiagramm für die Merkmalspaare `mpg` $\sim$ `hp, mpg` $\sim$ `wt` und interpretieren Sie die Diagramme. Berechnen Sie die zugehörigen Pearson-Korrelationskoeffizienten.

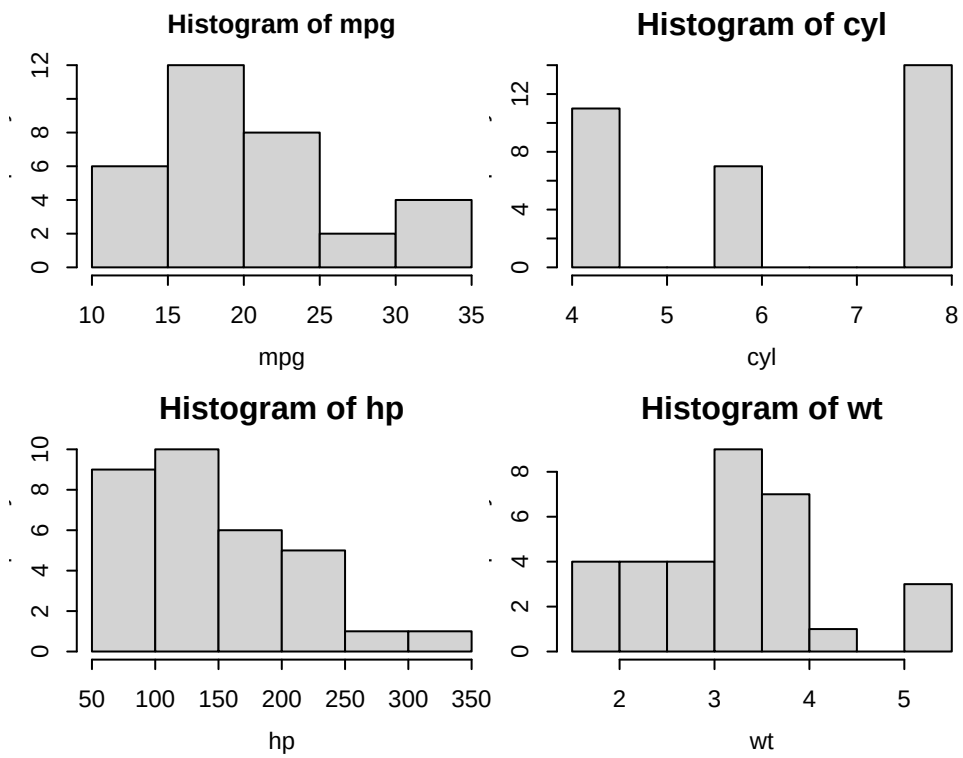
Lösung:

```
# (i)
attach(mtcars)
# Alternative: mtcars$...
# draw four plots in one window
par(mfrow=c(2,2), mar=c(5, 4, 4, 2)-2 + 0.1)
hist(mpg, cex.main = 1)
```

```
hist(cyl)
```

```
hist(hp)
```

```
hist(wt)
```




```
# alternativ mit ggplot
# (1 Punkt)
#
# (ii)
table(mpg)

mpg
10.4 13.3 14.3 14.7   15 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.2 19.7   21
   2    1    1    1    1    2    1    1    1    1    1    1    1    2    1    2
21.4 21.5 22.8 24.4   26 27.3 30.4 32.4 33.9
   2    1    2    1    1    1    2    1    1
```

```r
table(cyl)
```

```
cyl
 4  6  8
11  7 14
```

```r
table(hp)
```

```
hp
 52  62  65  66  91  93  95  97 105 109 110 113 123 150 175 180 205 215 230 245
  1   1   1   2   1   1   1   1   1   1   3   1   2   2   3   3   1   1   1   2
264 335
  1   1
```

```r
table(wt)
```

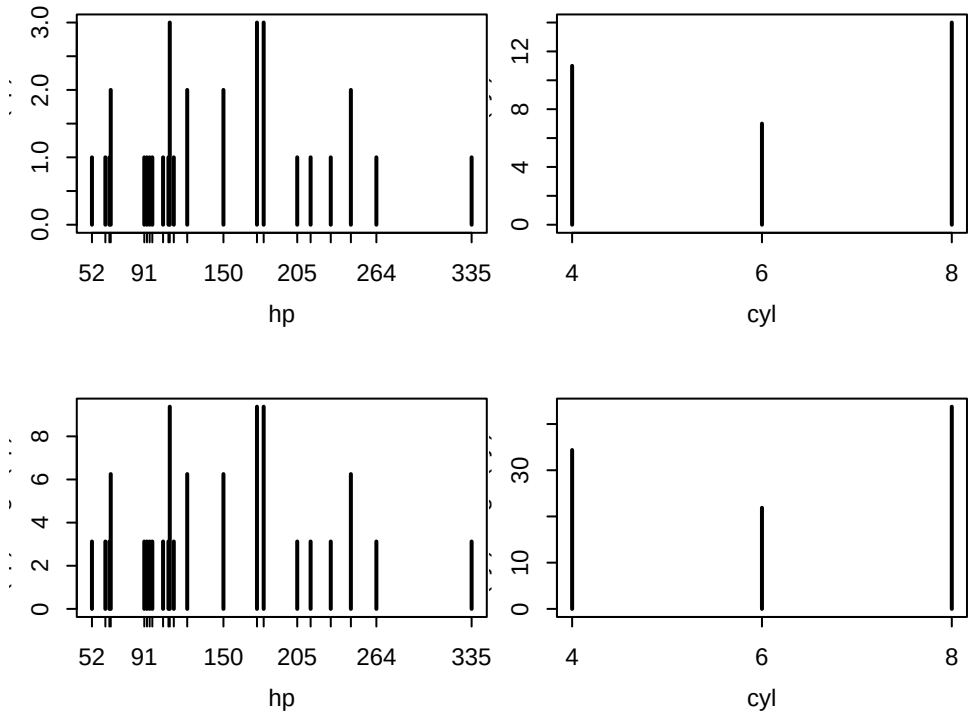
```
wt
1.513 1.615 1.835 1.935  2.14   2.2  2.32 2.465  2.62  2.77  2.78 2.875  3.15
    1     1     1     1     1     1     1     1     1     1     1     1     1
 3.17  3.19 3.215 3.435  3.44  3.46  3.52  3.57  3.73  3.78  3.84 3.845  4.07
    1     1     1     1     3     1     1     2     1     1     1     1     1
 5.25 5.345 5.424
    1     1     1
```

```r
# Säulendiagramm macht für hp und cyl Sinn.
# (1 Punkt)
#
# (iii)
plot(table(hp))
```

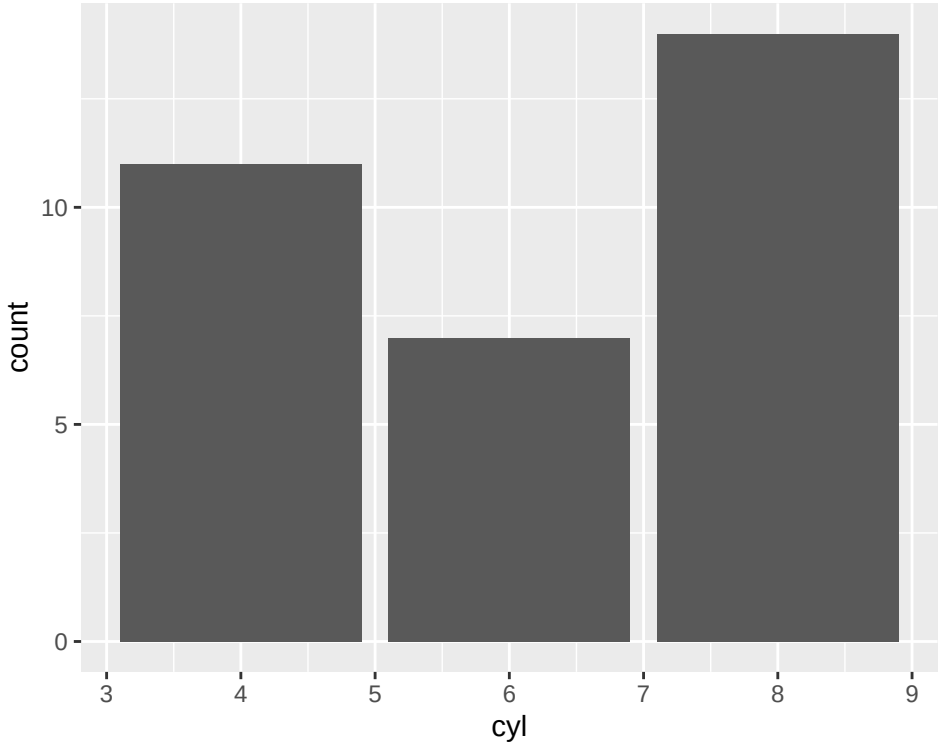
```r
plot(table(cyl))
```

```r
# mit relativen Häufigkeiten
plot(table(hp)/length(hp)*100)
```

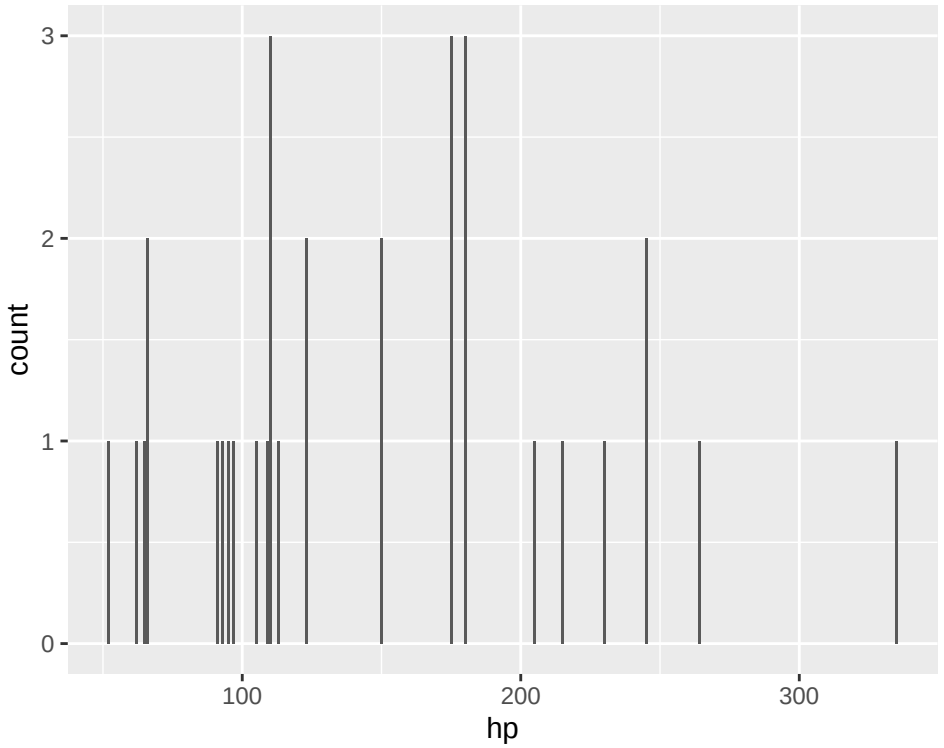
```r
plot(table(cyl)/length(cyl)*100)
```



```r
# alternativ mit tidyverse
library(tidyverse)
mtcars %>% ggplot() + geom_bar( aes( x = cyl) )
```

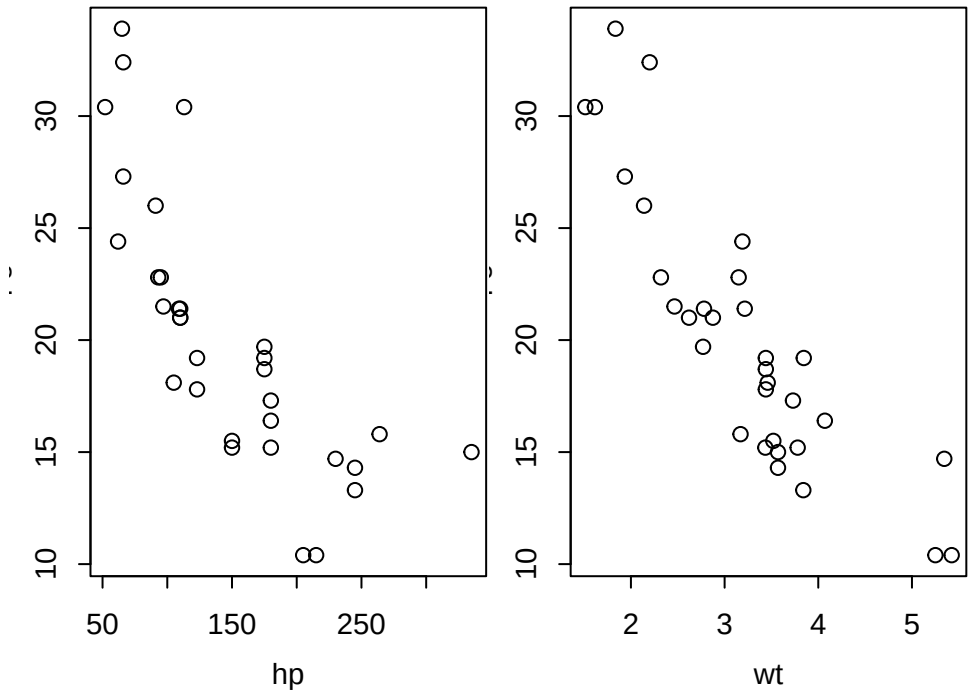



```r
mtcars %>% ggplot() + geom_bar( aes( x = hp) )
```




```r
#
# (1 Punkt)
# (iv)
par(mfrow=c(1,2))
plot(hp,mpg) # erst x, dann y Koordinate
```

```
plot(wt,mpg)
```



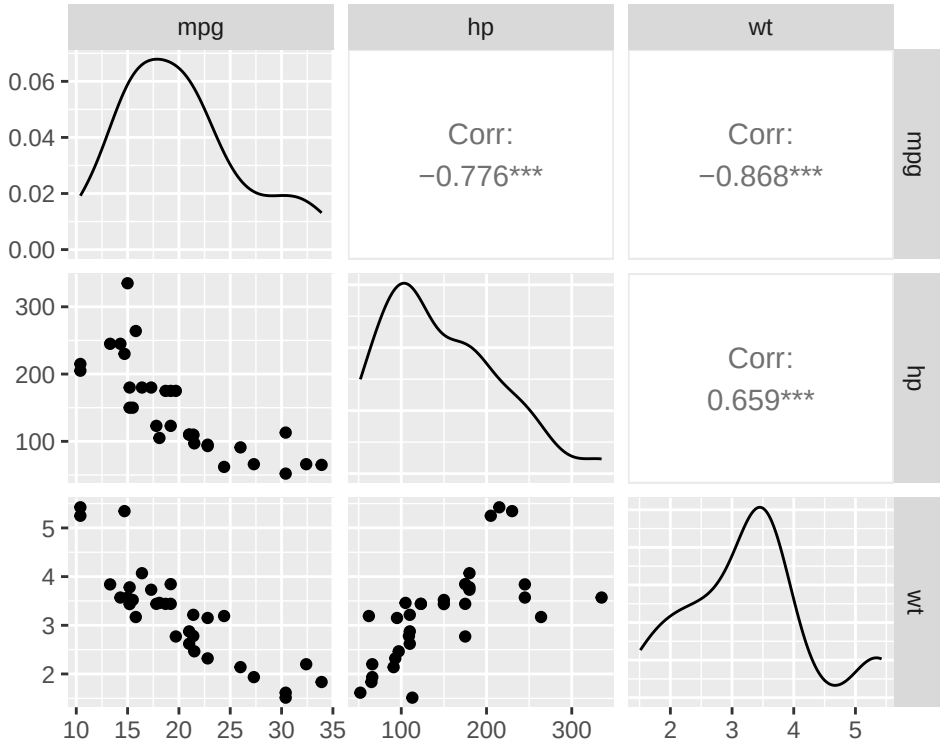

```
cor(hp,mpg)

[1] -0.776168

cor(wt,mpg)

[1] -0.867659

# alternative mit GGally
library(GGally)
ggpairs( mtcars %>% select(mpg, hp, wt))
```



(2 Punkte)

Aufgabe 4.8: Streuungszerlegung (keine Punkte)

Gegeben seien $m \in \mathbb{N}$ (Teil-)Stichproben (man sagt auch Schichten) $x_1, \ldots, x_m$ reeller Zahlen mit den Umfängen $n_1, \ldots, n_m, n := \sum_{i=1}^m n_i$, arithmetischen Mitteln $\bar{x}_1, \ldots, \bar{x}_m$ und Stichproben-Varianzen $s_1^2, \ldots, s_m^2$. Zeigen Sie, dass für das arithmetische Mittel $\bar{z}$ und die Stichproben-Varianz s_z^2 der gemeinsamen Stichprobe $z = (z_1, \ldots, z_n)$, die aus allen Stichprobenwerten der m Teilstichproben besteht, folgende Formeln gelten.

(i) $\bar{z} = \frac{1}{n} \sum_{i=1}^m n_i \bar{x}_i$ (gewichtetes Mittel)

(ii) $s_z^2 = \frac{1}{n-1} \sum_{i=1}^m (n_i - 1)\, s_i^2 + \frac{1}{n-1} \sum_{i=1}^m n_i \, (\bar{x}_i - \bar{z})^2$ (Streuungszerlegung)

(iii) Die Streuungszerlegung, vgl. (ii)), wird oft in der Form:

Gesamtstreuung = Streuung in den Schichten + Streuung zwischen den Schichten

ausgedrückt. Interpretieren Sie diese Aussage, z.B. grafisch, anhand der Formel in (ii).

Hinweis: Ohne Einschränkung der Allgemeinheit kann man davon ausgehen, dass die ersten n_1 Stichprobenwerte $z_1, \ldots, z_{n_1}$ der gemeinsamen Stichprobe z die Werte der Stichprobe x_1 sind, die folgenden n_2 Werte $z_{n_1+1}, \ldots, z_{n_1+n_2}$ die Werte der Stichprobe x_2 sind usw. bis zu den n_m Werten $z_{n_1+\ldots+n_{m-1}+1}, \ldots, z_n$, die die Stichprobenwerte von x_m sind.

Lösung:

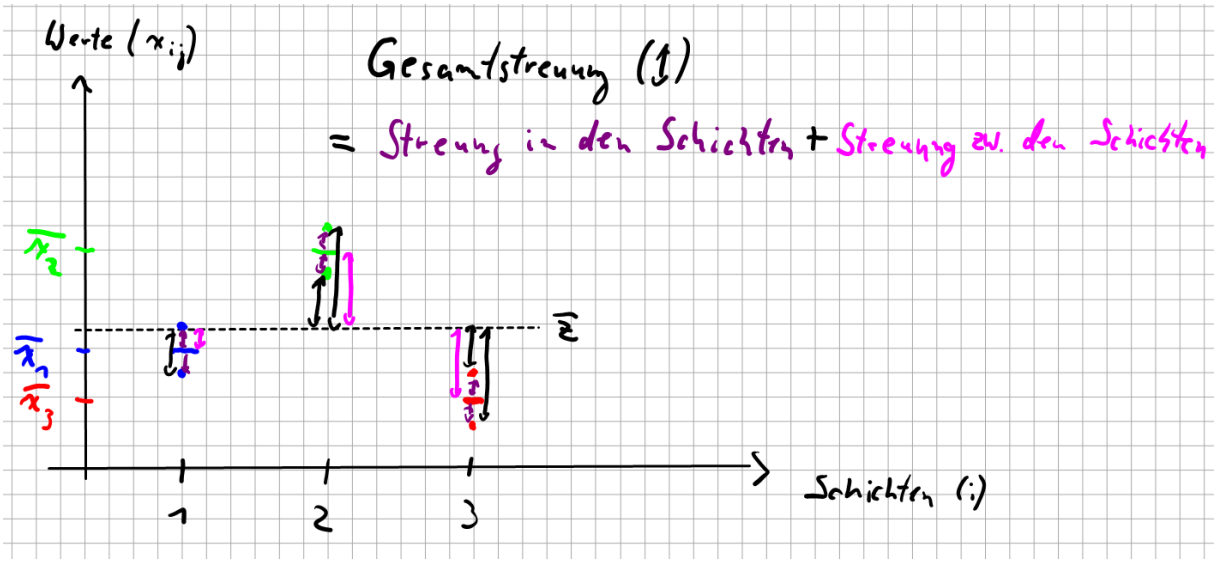
(i) $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^m \sum_{j=1}^{n_i} x_i j = \frac{1}{n} \sum_{i=1}^m n_i \bar{x}_i$ #

(ii)

$$\begin{aligned}
s_z^2 &= \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n-1} \left(\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{z})^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i + \bar{x}_i - \bar{z})^2 \right) \\
&= \frac{1}{n-1} \left(\sum_{i=1}^m \left(\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + 2 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{z}) + \sum_{j=1}^{n_i} (\bar{x}_i - \bar{z})^2 \right) \right) \\
&= \frac{1}{n-1} \sum_{i=1}^m (n_i - 1)\, s_i^2 + \frac{1}{n-1} \sum_{i=1}^m n_i \, (\bar{x}_i - \bar{z})^2
\end{aligned}$$

#

(iii)



Aufgabe 4.9: Korrelationen (R) (keine Punkte)

(i) Lesen Sie die Datei „AbsolventenDat.csv“ wie in Aufgabe 2.4 in den tibble `dat.Absolventen` ein.

(ii) Wandlen Sie die beiden Merkmale

Kontakt der Studierenden untereinander und *Umgang der Lehrenden mit den Studierenden*

in den Datentyp `factor` um und kodieren Sie die Stufen der beiden Merkmale mit dem Befehl `levels` in Schulnoten (1 bis 5) um.

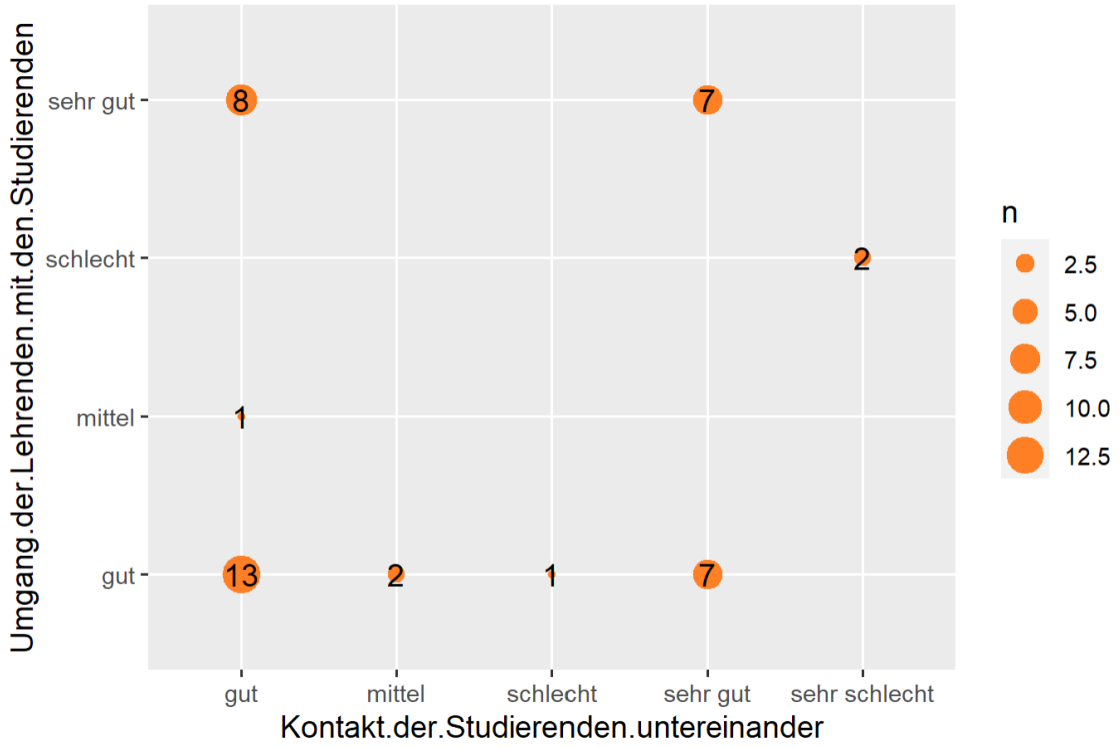
(iii) Berechnen Sie mit dem Befehl `table` eine Kontingenztabelle der beiden oben genannten Merkmale.

(iv) Zeichen Sie einen Häufigkeitsplot der beiden Merkmale.

Hinweis: Verwenden Sie den `geom_count()`-Befehl.

(v) Schreiben Sie zusätzlich die absolute Häufigkeit in die Kreise des Häufigkeitsplots.

Beispielsweise könnte die Grafik so aussehen:



Hinweis: Sehen Sie sich die Internetseite:

`www.r-graph-gallery.com/5-correlation-of-discrete-variables`

an. Auf der Seite `www.r-graph-gallery.com` finden Sie eine Menge schöner Grafiken und den zugehörigen R-code. Außerdem ist `stackoverflow.com` immer eine Suche wert.

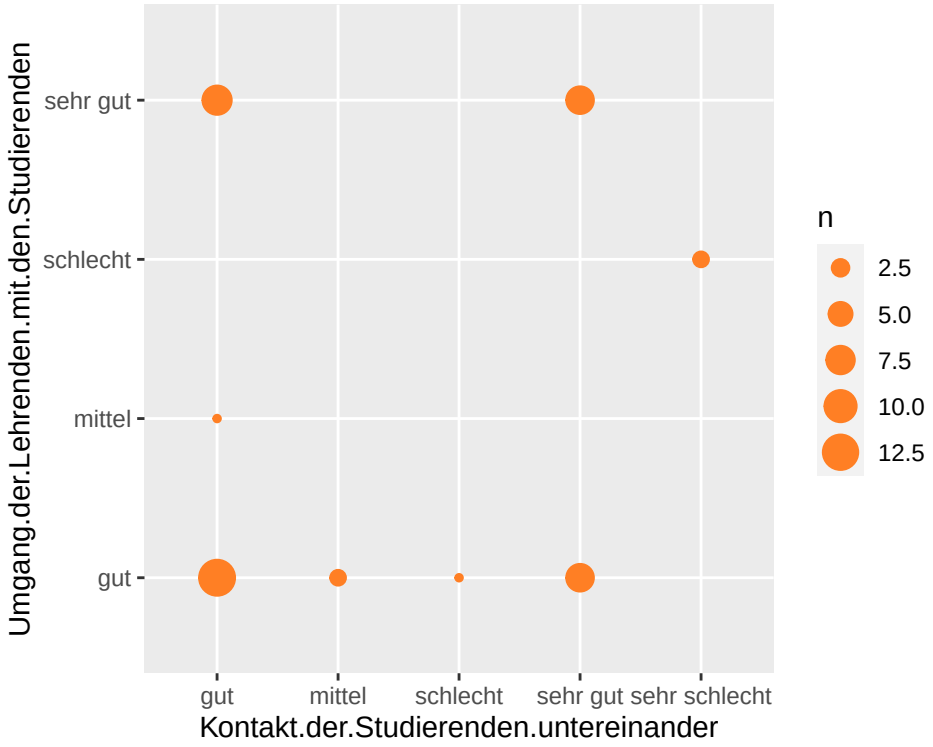
Lösung:

```
#
# (i)
dat.Absolventen <- read.csv2("AbsolventenDat.csv",na.strings = c("-","k.A."))
# (ii)
dat.Absolventen |> mutate(
  Kontakt.der.Studierenden.untereinander =
    factor( Kontakt.der.Studierenden.untereinander,
            levels = list( "1"="sehr gut", "2"="gut", "3"="mittel",
                           "4"="schlecht", "5"="sehr schlecht" ) ),
            Umgang.der.Lehrenden.mit.den.Studierenden =
    factor( Umgang.der.Lehrenden.mit.den.Studierenden,
                    levels = list( "1"="sehr gut", "2"="gut", "3"="mittel",
                                   "4"="schlecht", "5"="sehr schlecht" ) ) )
```

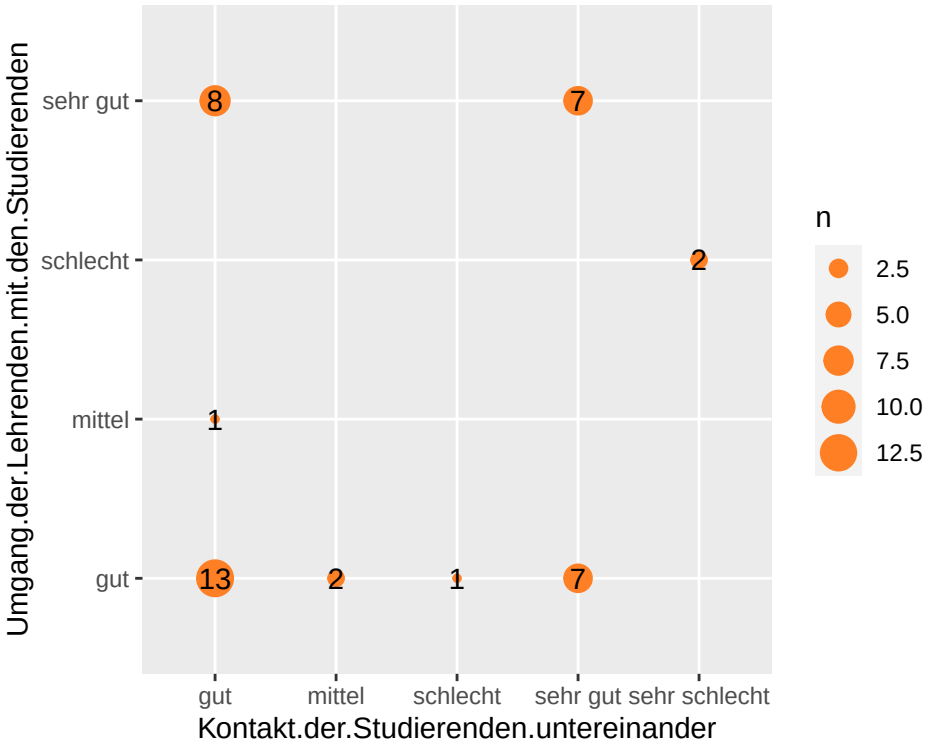
```r
#
# (iii)
table( dat.Absolventen$Kontakt.der.Studierenden.untereinander,
       dat.Absolventen$Umgang.der.Lehrenden.mit.den.Studierenden )
```

```
                gut mittel schlecht sehr gut
  gut            13      1        0        8
  mittel          2      0        0        0
  schlecht        1      0        0        0
  sehr gut        7      0        0        7
  sehr schlecht   0      0        2        0
```

```r
#
# (iv)
dat.Absolventen |> ggplot() +
  geom_count(mapping = aes(x = Kontakt.der.Studierenden.untereinander,
                           y = Umgang.der.Lehrenden.mit.den.Studierenden),
             color="chocolate1")
```




```r
# (v) [Lösung aus Stackoverflow]
p <- dat.Absolventen |> ggplot(mapping = aes(x = Kontakt.der.Studierenden.untereinander,
                           y = Umgang.der.Lehrenden.mit.den.Studierenden)) +
  geom_count(color="chocolate1")
p + geom_text(data = ggplot_build(p)$data[[1]], aes(x, y, label = n), color = "black")
```


Umgang.der.Lehrenden.mit.den.Studierenden
sehr gut
schlecht
mittel
gut
8
7
2
1
13
2
1
7
gut
mittel
schlecht
sehr gut
sehr schlecht
Kontakt.der.Studierenden.untereinander
n
2.5
5.0
7.5
10.0
12.5

Aufgabe 4.10: Korrelationen 2 (keine Punkte)

Sei $(x_i, y_i), i = 1, \ldots, n$, eine bivariate Stichprobe zweier metrischer Merkmale mit $s_x > 0, s_y > 0$ und Pearson-Korrelationskoeffizient $r_{x,y}$, wobei $x = (x_1, \ldots, x_n)^T$ und $y = (y_1, \ldots, y_n)^T$. Zeigen Sie, dass die folgenden Aussagen gelten.

(i) $-1 \leq r_{x,y} \leq 1$

(ii) $\forall i = 1, \ldots, n : y_i = a + bx_i, a, b \in \mathbb{R}, b \neq 0 \Rightarrow r_{x,y} = \text{sgn}(b)$, wobei $\text{sgn} : \mathbb{R} \to \{-1{,}0{,}1\}$ die Signum- oder Vorzeichenfunktion mit der Abbildungsvorschrift

$$\text{sgn(x)} = \begin{cases} -1 & \text{für } x < 0 \\ 0 & \text{für } x = 0 \\ 1 & \text{für } x > 0 \end{cases}$$

bezeichnet.

(iii) $r_{x,y} = \pm 1 \Rightarrow \exists a, b \in \mathbb{R}, \text{sgn}(b) = \text{sgn}\,(r_{x,y}) : y_i = a + bx_i, \forall i = 1, \ldots, n.$

Hinweis: Verwenden Sie die Ungleichung von Cauchy-Schwarz (Schwarzsche Ungleichung): Für alle $u, v \in \mathbb{R}^n, n \in \mathbb{N}$, gilt, dass

$$|u \cdot v| \leq \|u\|\|v\|.$$

Das Gleichheitszeichen gilt genau dann, falls u und v linear abhängig sind. Der Malpunkt steht dabei für das Skalarprodukt.

Lösung:

(i) Da

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

folgt nach der Cauchy-Schwarz-Ungleichung (CSU) sofort die Behauptung. #

(ii) Es gelte $\forall i = 1, \ldots, n : y_i = a + bx_i, a, b \in \mathbb{R}, b \neq 0$ Damit gilt auch für die Vektoren: $y = a + bx$. Also sind x und y linear abhängige Vektoren (siehe LA-Vorlesung, S. 110). Damit folgt mit der CSU: $|x \cdot y| = \|x\| \cdot \|y\|$ und damit $|r_{x,y}| = 1$. Da

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})b(x_i - \bar{x}) = b\|x - \bar{x}\|^2 (*)$$

folgt die Behauptung.

(iii) Sei $r_{x,y} = \pm 1$. Dann gilt:

$$|(x - \bar{x}) \cdot (y - \bar{y})| = \|x - \bar{x}\| \cdot \|y - \bar{y}\|$$

($\bar{x}$ und $\bar{y}$ decken wir uns hierbei als Vektoren mit n gleichen Elementen). Nach der CSU sind daher $x - \bar{x}$ und $y - \bar{y}$ linear abhängig. Folglich existiert nach der LA-Vorlesung ein $b \in \mathbb{R}$ mit: $y - \bar{y} = b(x - \bar{x})$ und damit $y = bx + a$ mit $a := \bar{y} - b\bar{x}$.

Die Tatsache, dass b und $r_{x,y}$ das gleiche Vorzeichen haben müssen, folgt aus (*).