Übung 12

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. Döhler (FBMN, h_da)

Name: Valentina Cisternas Seeger, Roman Kessler

Aufgabe 29. Arbeiten Sie weiter an dem Datensatz `UmfrageBis2019.csv` (s. Aufgabe 21).

a) Erzeugen Sie eine Scatterplotmatrix des gesamten Datensatzes (entfernen Sie ggf. irrelevante Merkmale). Experimentieren Sie auch mit der Funktion 'gpairs' aus dem gleichnamigen R-Paket.

b) Stellen Sie nun den 3-dimensionalen Datensatz der Merkmale Schuhgroesse, Groesse und Anzahl Schuhe als 3-dimensionalen scatterplot dar (Sie können dazu das Paket 'scatterplot3d' verwenden). Welche Einstellung der Parameter liefert eine aufschlussreiche Darstellung? Interpretieren Sie die Daten.

c) Visualisieren Sie die Correlationsmatrix durch ein Correlogramm der Daten aus b) z.B. mithilfe der Pakete 'corrplot' und 'psych'. Interpetieren Sie die Ergebnisse.

d) Stellen Sie die Daten aus b) durch einen parallelen Koordinaten-Plot dar. Interpretieren Sie diesen.

e) Betrachten Sie nun die gemeinsame Verteilung von 'Groesse' und 'Schuhgroesse'. Sie sollen einen 2-dimensionalen Kerndichteschätzer der Daten visualisieren. Ein solcher ist beispielsweise im Paket 'MASS' implementiert.

   i) Stellen Sie einen Höhenlinienplot des Schätzers zusammen mit den Originaldaten dar.

   ii) Stellen Sie einen Graphen des Schätzers dar.

   iii) Stellen Sie eine heatmap des Schätzers dar.

   Interpretieren Sie die Ergebnisse.

Denken Sie bitte wie immer daran, die Parameter der Grafiken sorfältig zu wählen sowie sie sinnvoll zu beschriften (Titel, Achsen, etc)!

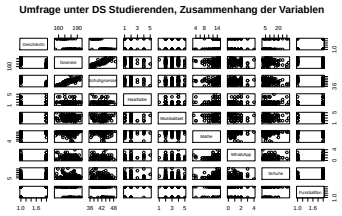
Aufgabe a)

```
df <- read.csv("C:/Users/Roman/Dropbox/hda/Explorative_Datenanalyse/uebungen/uebung13/Umfrag
df$Teilnehmer <- NULL
names(df)[names(df) == "Letzte.Schulnote.in.Mathematik"] <- "Mathe"
```

```r
names(df)[names(df) == "Stunden.am.Tag.in.WhatsApp"] <- "WhatsApp"
names(df)[names(df) == "Anzahl.Paar.Schuhe.im.Schrank"] <- "Schuhe"
summary(df)
#>  Geschlecht    Groesse       Schuhgroesse      Haarfarbe     Musikalitaet
#>  m:71       Min.   :155.0   Min.   :36.00   blond   :30             : 1
#>  w:30       1st Qu.:170.0   1st Qu.:40.00   braun   :61   etwas     :31
#>             Median :178.0   Median :42.50   rot     : 2   gar nicht:30
#>             Mean   :176.9   Mean   :41.99   schwarz : 7   mittel    :33
#>             3rd Qu.:185.0   3rd Qu.:44.00   sonstige: 1   sehr      : 6
#>             Max.   :196.0   Max.   :49.00
#>
#>      Mathe          WhatsApp         Schuhe       Fussballfan
#>  Min.   : 3.00   Min.   :0.000   Min.   : 2.000   ja  :31
#>  1st Qu.:12.00   1st Qu.:0.500   1st Qu.: 5.000   nein:70
#>  Median :13.00   Median :1.000   Median : 7.000
#>  Mean   :12.51   Mean   :1.055   Mean   : 9.337
#>  3rd Qu.:14.00   3rd Qu.:1.500   3rd Qu.:12.000
#>  Max.   :15.00   Max.   :4.000   Max.   :32.000
#>  NA's   :1
```

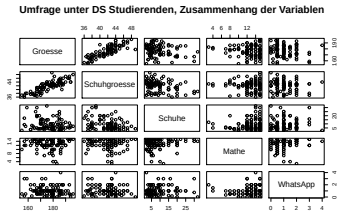
Scatterplotmatrix des gesamten Datensatzes:

```r
plot(df, main="Umfrage unter DS Studierenden, Zusammenhang der Variablen")
```



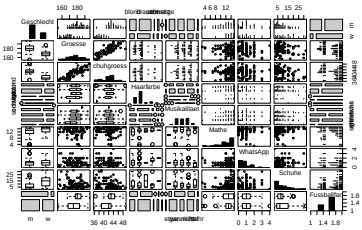
Scatterplotmatrix des bereinigten Datensatzes. Wir entfernen die kategoriellen Variablen.

```r
df2 <- subset(df, select = c("Groesse","Schuhgroesse","Schuhe","Mathe","WhatsApp"))
plot(df2, main="Umfrage unter DS Studierenden, Zusammenhang der Variablen")
```



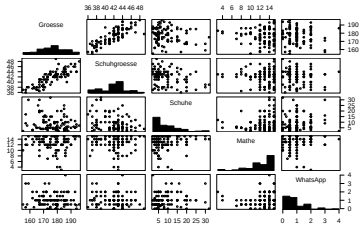
Experimentieren mit *gpairs*:

```
library(gpairs)
gpairs(df, gap = 0.1)
```



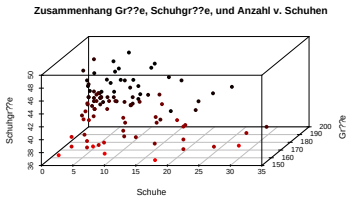
Mit dem gek?rzten Datensatz:

```
gpairs(df2, gap = 0.1)
```



Aufgabe b: 3D Zusammenhang zwischen Schuhgroesse, Groesse und Anzahl der Schuhe

```
library(scatterplot3d)
scatterplot3d(x = df$Schuhe,
              y = df$Groesse,
              z = df$Schuhgroesse,
              main = "Zusammenhang Gr??e, Schuhgr??e, und Anzahl v. Schuhen",
              zlab = "Schuhgr??e",
              ylab = "Gr??e",
              xlab = "Schuhe",
              highlight.3d = TRUE,
              scale.y = 0.9,
              angle = 60,
              pch = 16)
```

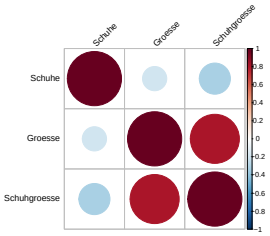


Wenn man den 3D-Scatterplot dreht und wendet (und aus den Ergebnissen der paarweisen Scatterplots aus der letzten Aufgabe), sieht man, dass die Merkmale Gr??e und Schuhgr??e einen gleichsinnigen Zusammenhang haben. In der hier gew?hlten Ansicht wird es durch den Parameter "highlight.3d" deutlich, welcher die z-Dimension (hier das Merkmal Gr??e) farbig einf?rbt. Rote Datenpunkte bedeuten eine eher niedrige Merkmalsauspr?gung, schwarze Datenpunkte eine hohe Merkmalsauspr?gung (= gro?e K?rpergr??e). Die roten Datenpunkte verteilen sich eher im unteren/vorderen Teil der Graphik, und die schwarzen Datenpunke eher im hinteren/oberen Teil der Graphik. Dies illustriert den gleichsinnigen Zusammenhang der beiden Merkmale. Interessanterweise ist hier kein Zusammenhang zwischen Schuhgr??e und Anzahl von Schuhen zu erkennen.

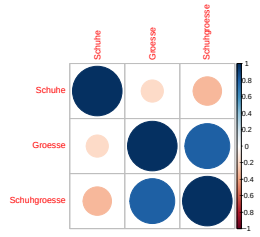
Aufgabe c)

'corrplot' und 'psych'

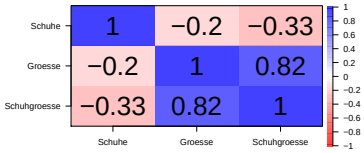
```
library(corrplot)
library(psych)
#> Warning: package 'psych' was built under R version 3.6.2
source("http://www.sthda.com/upload/rquery_cormat.r")
# we use this function for our correlations

# aus dem Paket "rquery_cormat"
corri <- rquery.cormat(subset(df, select = c("Groesse","Schuhgroesse","Schuhe")), type = "fu
```




```
# aus dem Paket "corrplot"
corrplot(corri$r)
```




```
# aus dem Paket "psych"
corPlot(corri$r)
```




```
# und der r-Wert dazu:
print(corri$r)
#>               Schuhe Groesse Schuhgroesse
#> Schuhe          1.00   -0.20        -0.33
#> Groesse        -0.20    1.00         0.82
#> Schuhgroesse   -0.33    0.82         1.00

# und der p-Wert dazu:
print(corri$p)
#>                Schuhe Groesse Schuhgroesse
#> Schuhe       0.00000 4.8e-02      6.7e-04
#> Groesse      0.04800 0.0e+00      3.2e-25
#> Schuhgroesse 0.00067 3.2e-25      0.0e+00
```

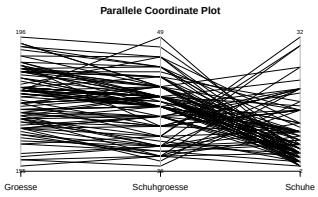
Man erkennt, dass die Merkmale Gr??e und Schuhgr??e stark positiv miteinander korrelieren (r = 0.82), w?hrend die Merkmale “Anzahl der Schuhe” mit den beiden Merkmalen Groesse (r = -0.2) und Schuhgroesse (r = -0.33) jeweils negativ korrelieren. Wenn wir auf die Signifikanzen (p-Werte) schauen, sehen wir dass alle 3 Zusammenh?nge auf einem alpha-Level von 0.05 (ohne Korrektur f?r multiple Vergleiche) signifikant sind. Der gleichsinnige Zusammenhang Gr??e/Schuhgr??e ist sehr stark signifikant (p sehr klein), w?hrend der gegensinnige Zusammenhang Schuhgroesse/Anzahl Schuhe immernoch recht signifikant ist, w?hrend der Zusammenhang Groesse/Anzahl Schuhe eine Korrektur f?r Multiple Vergleiche wahrscheinlich nicht ?berleben w?rde (p=0.048).

Zur Interpretation: Teilnehmer mit hoher K?rpergr??e zeigen auch eine hohe Schuhgr??e (positive Korrelation). Dieser Zusammenhang erscheint logisch, denn die Gr??e von K?rperteilen korreliert mit der K?rpergr??e (Quelle wird nachgereicht). Die negative Korrelation zwischen den Merkmalen Schuhgr??e und Anzahl der Schuhe k?nnte man ?ber gesellschaftliche Unterschiede zwischen M?nnern und Frauen interpretieren: Frauen haben meist mehr Schuhe, und kleinere F??e. Den Zusammenhang k?nnte man zum Beispiel mittels mutlipler Regression ermitteln: Man k?nnte ermitteln, ob das Geschlecht oder die Schuhgr??e st?rker pr?diktiv f?r die Anzahl der Schuhe ist.

Aufgabe d Parallel Coordinates Plot

```
library(MASS)
parcoord(subset(df, select = c("Groesse","Schuhgroesse","Schuhe")),
```

```
          var.label = TRUE, lty=1, lwd=0.8,
          main= "Parallele Coordinate Plot")
```

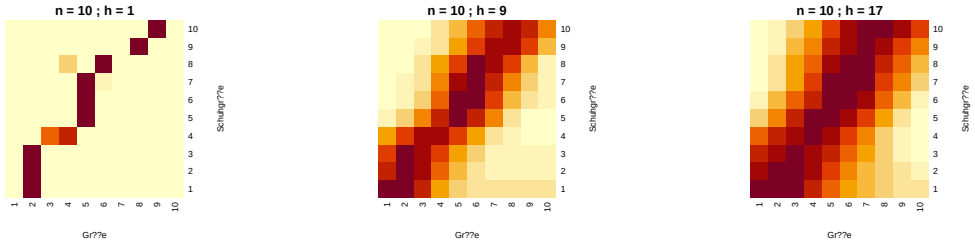


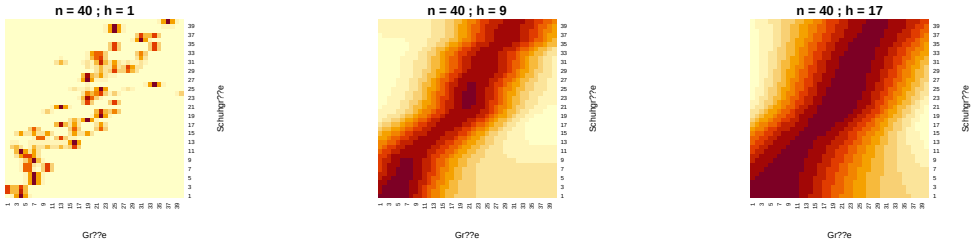
Der Parallele Coordinate Plot zeigt die Merkmalsauspr?gung (Y-Achse) der 3 Merkmale (X-Achse), wobei f?r jedes Merkmal die Y-Achse durch die Spannweite der Daten skaliert ist. An den zumeist horizontalen Verbindungen zwischen Groesse und Schuhgroesse sieht man auch den gleichsinnigen Zusammenhang der Daten: Teilnehmer mit einer gro?en Gr??e haben tendentiell auch eine gro?e Schuhgr??e, und vice versa. Zwischen Schuhgr??e und Anzahl der Schuhen sieht es schon anders aus. Hier erkennt man einen leichten gegensinnigen Zusammenhang dadurch, dass viele Verbindungen eine gr??ere (positive oder negative) Steigung haben und sich ?berkreuzen. Teilnehmer mit einer kleinen Schuhgr??e tendieren zu einer h?heren Anzahl von Schuhen, und umgekehrt.

Aufgabe e Wir bearbeiten zun?chst den Aufgabenteil (iii), um einen geeigneten Sch?tzer zu finden, und anschlie?end die Teile (i) und (ii).

iii

```
library(MASS)

for (n in c(10,40)){
  for (h in seq(from= 1, to = 20, by = 8)){
    kdes <- kde2d(x = df$Groesse, y = df$Schuhgroesse, n = n, h = h)
    heatmap(kdes$z,Colv = NA, Rowv = NA,
            main = paste0(c("n =",n,"; h =",h), collapse = " "),
            #main = expression("n =" ~ n ~ "; h = " ~ h),
            xlab = "Gr??e", ylab = "Schuhgr??e")
  }
}
```

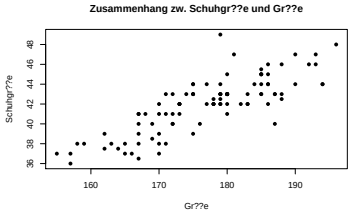




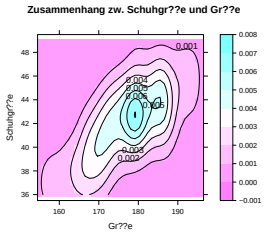
Wir entscheiden uns f?r die folgenden Aufgaben f?r einen Sch?tzer mit den Parametern n = 40 (Einheiten pro Achse), und h (Bandbreite) = 8.

i

```
n <- 40
h <- 9
kdes <- kde2d(x = df$Groesse, y = df$Schuhgroesse, n = n, h = h)

plot(x=df$Groesse, y = df$Schuhgroesse, pch = 16,
             xlab = "Gr??e",
             ylab = "Schuhgr??e",
             main = "Zusammenhang zw. Schuhgr??e und Gr??e")
```

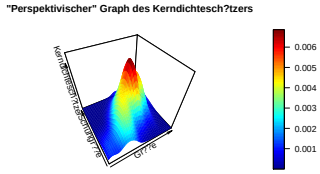



```
contourplot(kdes$z,
             pretty = TRUE, region = TRUE,
             row.values = kdes$x,
             column.values = kdes$y,
             aspect = "square",
             xlab = "Gr??e",
             ylab = "Schuhgr??e",
             main = "Zusammenhang zw. Schuhgr??e und Gr??e",
           )
```



ii

```
library(plot3D)

persp3D(z = kdes$z, theta = -30, phi = 45,
        main = "\"Perspektivischer\" Graph des Kerndichtesch?tzers",
        xlab = "Gr??e",
        ylab = "Schuhgr??e",
        zlab = "Kerndichtesch?tzer")
```



Anmerkungen/Korrektur
