Übung 9

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. Döhler (FBMN, h_da)

Name: Roman Kessler, Valentina Cisternas Seeger

Aufgabe 21. *In der Datei* `UmfrageBis2019.csv` *finden Sie (bereinigte) Daten, die von den aktuellen und bisherigen BesucherInnen der EDA-Veranstaltung erhoben wurden.*

a) Laden Sie die Daten und lassen Sie sich die ersten Datensätze mit dem `head`*-Befehl anzeigen. Verschaffen Sie sich einen Überblick über die Struktur der Daten mit dem* `str`*-Befehl. Nennen Sie den Datensatz* `data`*.*

b) Einige Spaltennamen sind sehr lang und daher unpraktisch zu handhaben. Benennen Sie die Spalten , `Letzte. Schulnote....`*,* `Stunden.am.Tag...` *bzw.* `Anzahl.Paar.Schuhe...` *um in* `Mathe`*,* `WhatsApp`*,* `AnzSchuhe` *bzw. und lassen Sie sich den header des neuen datensatzes anzeigen.*

c) Erzeugen Sie 5-point summaries für alle Merkmale.

d) Was passiert, wenn man `plot(data)` *eingibt?*

e) Definieren Sie eine Funktion `plot.cont.data`*, die als Argument einen Datensatz* `dat` *eines stetigen Merkmals nimmt und einen grafischen Outpunkt in Form eines 2x2 Grafikpanels liefert, der folgende Grafiken enthält:*

- *Boxplot*
- *QQ-Plot*
- *Histogramm mit Dichtefunktion der angepassten Normalverteilung*
- *Empirische Verteilungsfunktion mit Verteilungsfunktion der angepassten Normalverteilung.*

f) Wenden Sie `plot.cont.data` *auf das Merkmal* `Groesse` *an und interpretieren Sie das Ergebnis.*

g) Definieren Sie eine Funktion `analyse.regression`*, die als Argument einen Teil-Datensatz* `dat` *des gesamten Datensatzes nimmt und eine Regression von* `Groesse` *auf* `Schuhgroesse` *durchführt. Als Ausgabe sollen das lm-Objekt und die ANOVA-Tabelle zurückgegeben werden.*

h) Wenden Sie `analyse.regression` *auf die gesamte Stichprobe, sowie jeweils auf die männlichen und weiblichen Teilnehmer an und interpretieren Sie*

die Ergebnisse. Um Teile der Gesamtstichprobe auszuwählen können Sie den `subset`*-Befehl verwenden.*

i) Analysieren Sie Körpergröße in Abhängigkeit von der Haarfarbe durch mehrere Boxplots in einer Grafik. Was erscheint besonders? Untersuchen Sie die Daten ggf. genauer, um den Grund zu erkennen.

Lösung

a. Import the data:

```
data = read.table(
  file ="C:/Users/Roman/Dropbox/hda/Explorative_Datenanalyse/uebungen/uebung9/UmfrageBis2019
  sep = ";", header = TRUE)
```

Nun zeigen wir die ersten Zeilen der Tabelle:

```
head(data)
#>   Teilnehmer Geschlecht Groesse Schuhgroesse Haarfarbe Musikalitaet
#> 1          1          m     184           44     blond         sehr
#> 2          2          w     163           38     braun       mittel
#> 3          3          m     175           44     braun       mittel
#> 4          4          m     177           44     braun       mittel
#> 5          5          m     180           43     blond       mittel
#> 6          6          m     180           43     braun    gar nicht
#>   Letzte.Schulnote.in.Mathematik Stunden.am.Tag.in.WhatsApp
#> 1                             12                      1.500
#> 2                             13                      0.167
#> 3                             13                      0.200
#> 4                              8                      0.500
#> 5                              7                      0.333
#> 6                             13                      2.000
#>   Anzahl.Paar.Schuhe.im.Schrank Fussballfan
#> 1                             9          ja
#> 2                             8        nein
#> 3                             5          ja
#> 4                             5          ja
#> 5                             6          ja
#> 6                             5          ja
```

Nun lassen wir uns ausgeben, welche Variablen wir haben, was für einen Typ sie haben, wieviele vorhanden sind, etc.

```
str(data)
#> 'data.frame':    101 obs. of  10 variables:
#>  $ Teilnehmer                   : int  1 2 3 4 5 6 7 8 9 10 ...
#>  $ Geschlecht                   : Factor w/ 2 levels "m","w": 1 2 1 1 1 1 1 1 1 1 ...
#>  $ Groesse                      : int  184 163 175 177 180 180 173 168 179 179 ...
```

```
#>  $ Schuhgroesse                  : num  44 38 44 44 43 43 42 41 42.5 42 ...
#>  $ Haarfarbe                     : Factor w/ 5 levels "blond","braun",..: 1 2 2 2 1 2 1 2
#>  $ Musikalitaet                  : Factor w/ 5 levels "","etwas","gar nicht",..: 5 4 4 4
#>  $ Letzte.Schulnote.in.Mathematik: int  12 13 13 8 7 13 11 13 9 10 ...
#>  $ Stunden.am.Tag.in.WhatsApp    : num  1.5 0.167 0.2 0.5 0.333 2 0.5 0.5 0.5 0 ...
#>  $ Anzahl.Paar.Schuhe.im.Schrank : int  9 8 5 5 6 5 7 4 4 3 ...
#>  $ Fussballfan                   : Factor w/ 2 levels "ja","nein": 1 2 1 1 1 1 2 2 1 1 ..
```

b) Umbenennung der Spalten.

```
names(data)[names(data) == "Letzte.Schulnote.in.Mathematik"] <- "Mathe"
names(data)[names(data) == "Stunden.am.Tag.in.WhatsApp"] <- "WhatsApp"
names(data)[names(data) == "Anzahl.Paar.Schuhe.im.Schrank"] <- "Schuhe"
head(data)
#>   Teilnehmer Geschlecht Groesse Schuhgroesse Haarfarbe Musikalitaet Mathe
#> 1          1          m     184           44     blond         sehr    12
#> 2          2          w     163           38     braun       mittel    13
#> 3          3          m     175           44     braun       mittel    13
#> 4          4          m     177           44     braun       mittel     8
#> 5          5          m     180           43     blond       mittel     7
#> 6          6          m     180           43     braun    gar nicht    13
#>   WhatsApp Schuhe Fussballfan
#> 1    1.500      9          ja
#> 2    0.167      8        nein
#> 3    0.200      5          ja
#> 4    0.500      5          ja
#> 5    0.333      6          ja
#> 6    2.000      5          ja
```

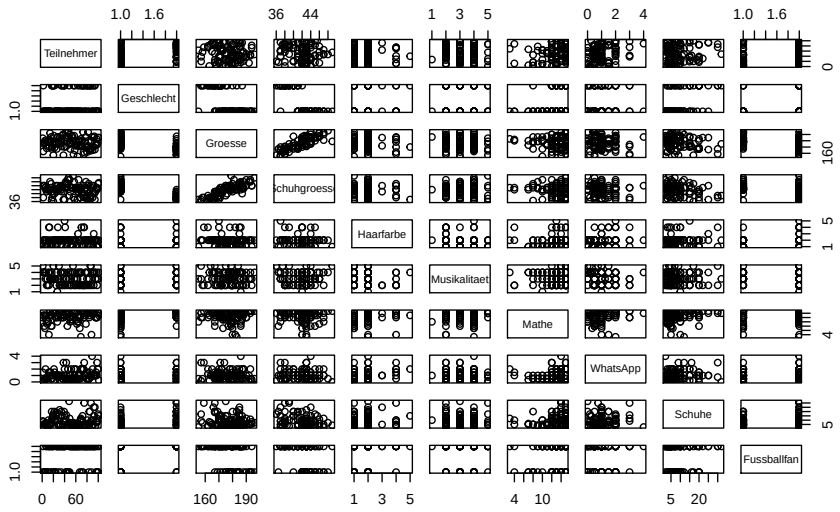
c) 5-Point-Summaries für alle Features.

```
summary(data)
#>    Teilnehmer  Geschlecht    Groesse       Schuhgroesse      Haarfarbe
#>  Min.   :  1   m:71       Min.   :155.0   Min.   :36.00   blond   :30
#>  1st Qu.: 26   w:30       1st Qu.:170.0   1st Qu.:40.00   braun   :61
#>  Median : 51              Median :178.0   Median :42.50   rot     : 2
#>  Mean   : 51              Mean   :176.9   Mean   :41.99   schwarz : 7
#>  3rd Qu.: 76              3rd Qu.:185.0   3rd Qu.:44.00   sonstige: 1
#>  Max.   :101              Max.   :196.0   Max.   :49.00
#>
#>     Musikalitaet     Mathe          WhatsApp         Schuhe
#>            : 1    Min.   : 3.00   Min.   :0.000   Min.   : 2.000
#>  etwas     :31    1st Qu.:12.00   1st Qu.:0.500   1st Qu.: 5.000
#>  gar nicht:30    Median :13.00   Median :1.000   Median : 7.000
#>  mittel    :33    Mean   :12.51   Mean   :1.055   Mean   : 9.337
#>  sehr      : 6    3rd Qu.:14.00   3rd Qu.:1.500   3rd Qu.:12.000
```

```
#>                  Max.   :15.00   Max.   :4.000   Max.   :32.000
#>                  NA's   :1
#>  Fussballfan
#>  ja  :31
#>  nein:70
#>
#>
#>
#>
#>
```

d) Was passiert bei plot(data) ?

```
plot(data)
```



Es erscheint eine Graphik, in welcher jede Variable gegen jede andere Variable in jeweils einem Streudiagramm geplottet wird. Der Plot scheint sehr hilfreich, um erste Einblicke über mögliche Zusammenhänge einzelner Variablen zu erhalten. Die Auftragung gegen sich selbst (auf der Diagonalen) wurde ausgespart, und stattdessen der Variablennamen als Label ausgegeben.

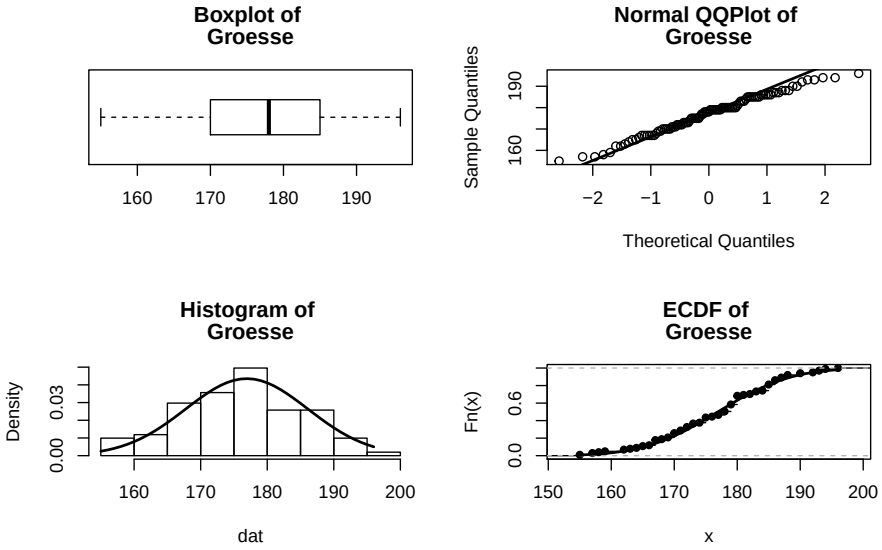
e) Definiere EDA Funktion

```
plot.cont.data <- function(dat, varName){
  # for dat a stetic variable must be chosen !

  # boxplot
  par(mfrow=c(2,2)) # create 2x2 subplots
```

```
  boxplot(dat, horizontal = TRUE,
          main = paste(c("Boxplot of ",varName)))

  # qqplot
  {qqnorm(dat,
          main = paste(c("Normal QQPlot of ",varName)))
    qqline(dat, lwd = 2)
  }

  # hist. with density
  {hist(dat, freq = FALSE,
        main = paste(c("Histogram of ",varName)))
    xs <- seq(min(dat),max(dat))
    ys <- dnorm(xs, mean = mean(dat), sd = sd(dat))
    lines(x = xs, y = ys, col = "black", lwd = 2)
  }

  # ecdf
  {plot(ecdf(dat),
       main = paste(c("ECDF of ",varName)))
    ys2 <- pnorm(xs, mean = mean(dat), sd = sd(dat))
    lines(x = xs, y = ys2, col = "black", lwd = 2)
  }

}
```

f) Anwenden auf Merkmal Groesse!

```
plot.cont.data(data$Groesse, "Groesse")
```



Der Boxplot zeigt bereits, dass sich die Körpergrößen grob von 150 bis 200 cm verteilen, wobei der Median bei etwa 178cm liegt, das untere Quartil bei etwa 170cm, und das obere Quartil bei etwa 185cm.

Der QQPlot gibt etwas Aufschluss darüber, ob die Daten normalverteilt sind. Wir sehen am oberen und unteren Ende, dass die Daten etwas von der Winkelhalbierenderen Geraden abweichen. Jedoch nur bei den Extremwerten, während in der Mitte die Werte halbwegs auf der Geraden liegen. Somit können wir von einer fast-Normalverteilung ausgehen.

Der Boxplot vorher hat auch schon erste Hinweise (wenn auch viel ungenauer) auf eine Normalverteilung gegeben. da er symmetrisch aussieht (die beiden mittleren Quartile als auch die beiden Whiskers sind etwa gleich groß).

Das Histogram ergänzt noch einmal die Aussagen des Boxplots. Wir sehen die Verteilung der Daten hier besser, und sehen auch, dass sie recht symmetrisch um den Erwartungswert verteilt sind. Die Eingezeichnete Dichtefunktion bestätigt noch einmal, dass die Daten halbwegs einer Normalverteilung folgen könnten.

Die Empirische Verteilungsfunktion mit der eingezeichneten Kumulativen Verteilungsfunktion einer theoretischen Normalverteilung mit mean = mean(Groesse) und sd = sd(Groesse) verstärkt die Vermutung nach einer Normalverteilung. Die Daten der Emp. Verteilungsfunktion liegen sehr gut auf der theoretischen Funktion.

g) definiere Regressionsfunktion

```
library(lme4) # import of linear model
analyse.regression <- function(dat){
```

```
  gr  <- dat$Groesse
  sgr <- dat$Schuhgroesse
  model <- lm(sgr ~ gr)
  return(list(model, anova(model)))
  # to return multiple objects, we need to use a list
}
```

h) Anwendung der Regression auf:

h.i) die ganze Stichprobe

```
analyse.regression(data)
#> [[1]]
#>
#> Call:
#> lm(formula = sgr ~ gr)
#>
#> Coefficients:
#> (Intercept)           gr
#>     -1.7228       0.2471
#>
#>
#> [[2]]
#> Analysis of Variance Table
#>
#> Response: sgr
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> gr         1 513.13  513.13  196.25 < 2.2e-16 ***
#> Residuals 99 258.86    2.61
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

zu den Modellparametern: Wir sehen zunächst einen positiven Zusammenhang (slope = 0.25). Das bedeutet, innerhalb unseres Wertebereiches (X-Achse) steigt mit jeder Einheit Grösse, die Schuhgrösse um 0.25 Einheiten. Der intercept liegt bei -1.73. Er ist in unserem Fall nicht interpretierbar (jemand mit der Größe Null, hat keine Schuhgröße, und wenn, dann wäre sie nicht negativ).

zu der ANOVA-Tabelle: Wir sehen die Ergebnisse des F-Tests. Die Wahrscheinlichkeit, dass die Nullhypothese verworfen werden kann, und somit dass ein linearer Zusammenhang zwischen Grösse und Schuhgrösse vorhanden ist, beträgt 1 - 2.2*10^-15 und ist somit sehr hoch. Das ist auch an dem hohen F-Wert zu sehen.

h.ii) männliche Teilnehmer

```
analyse.regression(
  data[which(data$Geschlecht == "m"),]
```

```
)
#> [[1]]
#>
#> Call:
#> lm(formula = sgr ~ gr)
#>
#> Coefficients:
#> (Intercept)           gr
#>     12.3557       0.1719
#>
#>
#> [[2]]
#> Analysis of Variance Table
#>
#> Response: sgr
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> gr         1 113.59 113.588  60.689 4.834e-11 ***
#> Residuals 69 129.14   1.872
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

zu den Modellparametern: Wir sehen zunächst einen positiven Zusammenhang (slope = 0.17). Das bedeutet, innerhalb unseres Wertebereiches (X-Achse) steigt mit jeder Einheit Grösse, die Schuhgrösse um 0.17 Einheiten (bei den Männern). Der intercept liegt bei 12.3557. Er ist in unserem Fall nicht interpretierbar.

zu der ANOVA-Tabelle: Wir sehen die Ergebnisse des F-Tests. Die Wahrscheinlichkeit, dass die Nullhypothese verworfen werden kann, und somit dass ein linearer Zusammenhang zwischen Grösse und Schuhgrösse vorhanden ist, beträgt 1 - 4.8*10^-11 und ist somit sehr hoch. Das ist auch an dem immernoch sehr hohen F-Wert zu sehen.

Der positive Zusammenhang zwischen Groesse und Schuhgroesse ist somit auch in der Subgruppe der Männer vorhanden.

h.iii) weibliche Teilnehmer

```
analyse.regression(data[which(data$Geschlecht == "w"),])
#> [[1]]
#>
#> Call:
#> lm(formula = sgr ~ gr)
#>
#> Coefficients:
#> (Intercept)           gr
#>     13.3217       0.1507
#>
```

```
#>
#> [[2]]
#> Analysis of Variance Table
#>
#> Response: sgr
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> gr         1 33.585  33.585  27.892 1.286e-05 ***
#> Residuals 28 33.715   1.204
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

zu den Modellparametern: Wir sehen zunächst einen positiven Zusammenhang (slope = 0.15). Das bedeutet, innerhalb unseres Wertebereiches (X-Achse) steigt mit jeder Einheit Grösse, die Schuhgrösse um 0.15 Einheiten (bei den Frauen). Der Slope ist nur minimal unterschiedlich zu dem der Männer. Der intercept liegt bei 13. Er ist in unserem Fall nicht interpretierbar.

Interessant ist, dass die Analysen für Männer und Frauen getrennt jeweils niedrigere Slopes, und höhere Intercepts liefern als in der Analyse der gesamten Gruppe.

zu der ANOVA-Tabelle: Wir sehen die Ergebnisse des F-Tests. Die Wahrscheinlichkeit, dass die Nullhypothese verworfen werden kann, und somit dass ein linearer Zusammenhang immernoch hoch. Das ist auch an dem immernoch hohen F-Wert zu sehen.

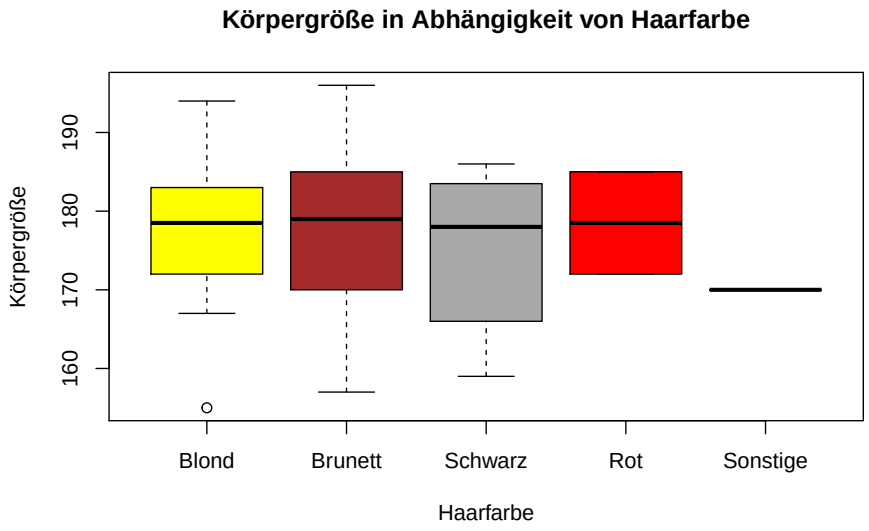
Trotz allem Sind F-Wert geringer als bei der gleichen Analyse der Männer Subgruppe, und bei weitem geringer als in der Analyse der Gesamtstichprobe. Das gilt umgekehrt natürlich auch für die p-Werte.

Der positive Zusammenhang zwischen Groesse und Schuhgroesse ist somit auch in der Subgruppe der Frauen vorhanden und signifikant.

i) Körpergröße in Abhängigkeit von Haarfarbe

```
# chose subset
groesse <- data.frame(row.names = list(c("Blond","Brunett","Schwarz","Rot","Sonstige")))

blonde   <- subset(data, Haarfarbe == "blond", select = Groesse)
brunette <- subset(data, Haarfarbe == "braun", select = Groesse)
schwarze <- subset(data, Haarfarbe == "schwarz", select = Groesse)
rote     <- subset(data, Haarfarbe == "rot", select = Groesse)
sonstige <- subset(data, Haarfarbe == "sonstige", select = Groesse)

# boxplots

#par(mfrow=c(1,5))
boxplot( c(blonde,brunette,schwarze,rote,sonstige),
```

```
        col = c("yellow","brown","darkgrey","red","green"),
        names = c("Blond","Brunett","Schwarz","Rot","Sonstige"),
        main = "Körpergröße in Abhängigkeit von Haarfarbe",
        ylab = "Körpergröße",
        xlab = "Haarfarbe"
        )
```



besonders erscheint: Die Mediane der 4 Großen Gruppen (im Folgenden vernachlässigen wir den einen Datenpunk von "Sonstige") liegen sehr ähnlich.

Das gilt auch für die Gruppe "Rothaarig", auch wenn dort nur 2 Datenpunkte vorliegen.

Die Gruppe "Sonstige Haarfarbe" hat nur einen Datenpunkt, deswegen interpretieren wir diese Gruppe erstmal nicht. Auffällig (positiv) ist, dass die wenigen Datenpunkte der Gruppen mit wenigen Datenpunkten ("Rothaarig" und "Sonstige Haarfarbe") halbwegs innerhalb der IQR der anderen Gruppen liegen, auf jeden Fall innerhalb der Daten-Ranges der anderen Gruppen.

Eine genauere Untersuchung sparen wir an dieser stelle aus, da wir keine besonderen Anzeichen für bedeutende Gruppenunterschiede sehen.

Anmerkungen/Korrektur
