Übung 10

Roman Kessler und Valentina Cisternas Seeger

03/12/2019

Übung 10

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. Döhler, B. Nedic (FBMN, h_da)

Name:

Punkte:

Aufgabe 22. Auf der Moodle-Seite zur Vorlesung finden Sie 4 verschiedene zweidimensionale Datensätze.

a) Ermitteln Sie in jedem Fall $\overline{x}, \overline{y}, s_x, s_y$ und r_{xy}. Was fällt Ihnen auf?

b) Ermitteln Sie in allen 4 Fällen das Regressionsmodel und die Regressionsgerade. Benutzen Sie hierfür den `lm`-Befehl.

c) Schreiben Sie eine Funktion `plot.regression`. Als Argumente sollen dieser Funktion übergeben werden:

   - `daten`: Ein Datensatz
   - `model`: Das zugehörige `lm`-Objekt
   - `header`: Eine Überschrift
   - `x.lim,y.lim`: Wertebereich der x- und y-Achse

   Als Output soll die Funktion einen Scatterplot der Daten (auf den übergebenen Wertebereichenen) mit zugehöriger Regressionsgerade sowie eines Titels liefern.

d) Plotten Sie die 4 Scatterplots inklusive ihrer Regressionsgeraden in einem gemeinsamen Grafikpanel. Benutzen Sie die Funktion `plot.regression`.

e) Führen Sie für die 4 Datensätze jeweils eine Residuenanalyse durch und interpretieren Sie die Ergebnisse.

Interpretieren Sie die Ergebnisse.

Lösung

Anmerkungen/Korrektur

Aufgabe 23. Arbeiten Sie weiter an dem Datensatz `UmfrageBis2019.csv` (s. Aufgabe 21).

a) Untersuchen Sie, ob und wenn ja, welcher Zusammenhang zwischen den Merkmalen 'Fussballfan' und 'Geschlecht' besteht. Erzeugen Sie hierzu eine Vierfeldertafel und einen Mosaicplot mit geeigneten R-Befehlen.

b) Untersuchen Sie, die gleiche Fragestellung wie in a) bei den Merkmalen 'Musikalitaet' und 'Geschlecht'. Was unterscheidet das Merkmal 'Musikalitaet' vom Merkmal 'Fussballfan'? Wie sollte man das in der Analyse berücksichtigen?

c) Untersuchen Sie, die gleiche Fragestellung wie in a) bei den Merkmalen 'Musikalitaet' und 'Haarfarbe'.

d) In welchen der 3 obigen Situation scheint der Mosaicplot am ehesten auf Unabhängigkeit der Merkmale hinzudeuten? Begründen Sie Ihre Antwort!

Lösung

Aufgabe 1 a)

```
reg1 <- read.csv("~/Desktop/Uni/M. Sc. Data Science/1.Semester /03_EDA/Reg1.csv",header=TRUE
x1<-reg1$x1
y1<-reg1$y1

mean(x1)
#> [1] 9
mean(y1)
#> [1] 7.500909
sd(x1)
#> [1] 3.316625
sd(y1)
#> [1] 2.031568
cor.test(x1,y1)
#>
#>  Pearson's product-moment correlation
#>
#> data:  x1 and y1
#> t = 4.2415, df = 9, p-value = 0.00217
```

```
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.4243912 0.9506933
#> sample estimates:
#>       cor
#> 0.8164205
```

```
reg2 <- read.csv("~/Desktop/Uni/M. Sc. Data Science/1.Semester /03_EDA/Reg2.csv",header=TRUE
x2<-reg2$x2
y2<-reg2$y2

mean(x2)
#> [1] 9
mean(y2)
#> [1] 7.500909
sd(x2)
#> [1] 3.316625
sd(y2)
#> [1] 2.031657
cor.test(x2,y2)
#>
#>  Pearson's product-moment correlation
#>
#> data:  x2 and y2
#> t = 4.2386, df = 9, p-value = 0.002179
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.4239389 0.9506402
#> sample estimates:
#>       cor
#> 0.8162365
```

```
reg3 <- read.csv("~/Desktop/Uni/M. Sc. Data Science/1.Semester /03_EDA/Reg3.csv",header=TRUE
x3<-reg3$x3
y3<-reg3$y3

mean(x3)
#> [1] 9
mean(y3)
#> [1] 7.5
sd(x3)
#> [1] 3.316625
sd(y3)
#> [1] 2.030424
cor.test(x3,y3)
```

```
#>
#>  Pearson's product-moment correlation
#>
#> data:  x3 and y3
#> t = 4.2394, df = 9, p-value = 0.002176
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.4240623 0.9506547
#> sample estimates:
#>       cor
#> 0.8162867
```

```
reg4 <- read.csv("~/Desktop/Uni/M. Sc. Data Science/1.Semester /03_EDA/Reg4.csv",header=TRUE
x4<-reg4$x4
y4<-reg4$y4
mean(x4)
#> [1] 9
mean(y4)
#> [1] 7.500909
sd(x4)
#> [1] 3.316625
sd(y4)
#> [1] 2.030579
cor.test(x4,y4)
#>
#>  Pearson's product-moment correlation
#>
#> data:  x4 and y4
#> t = 4.243, df = 9, p-value = 0.002165
#> alternative hypothesis: true correlation is not equal to 0
#> 95 percent confidence interval:
#>  0.4246394 0.9507224
#> sample estimates:
#>       cor
#> 0.8165214
```

Alle Datensätze haben den gleichen Mittelwert für x und bis auf Reg3 den gleichen Mittwelwert für y. Abweichung von 0,000909 bei Reg3. Die Standardabweichung ist bei allen Datensätzen für x gleich und für y ungefähr gleich. Auch die Korrelation ist bei allen ungefähr gleich. Insgesamt sind die Werte für alle Datensätze, auf zwei Kommastellen gerundet, gleich.

Aufgabe 1[b]:

```r
linearMod1 <- lm(y1 ~ x1, data=reg1)
print(linearMod1)
#>
#> Call:
#> lm(formula = y1 ~ x1, data = reg1)
#>
#> Coefficients:
#> (Intercept)           x1
#>      3.0001       0.5001
```

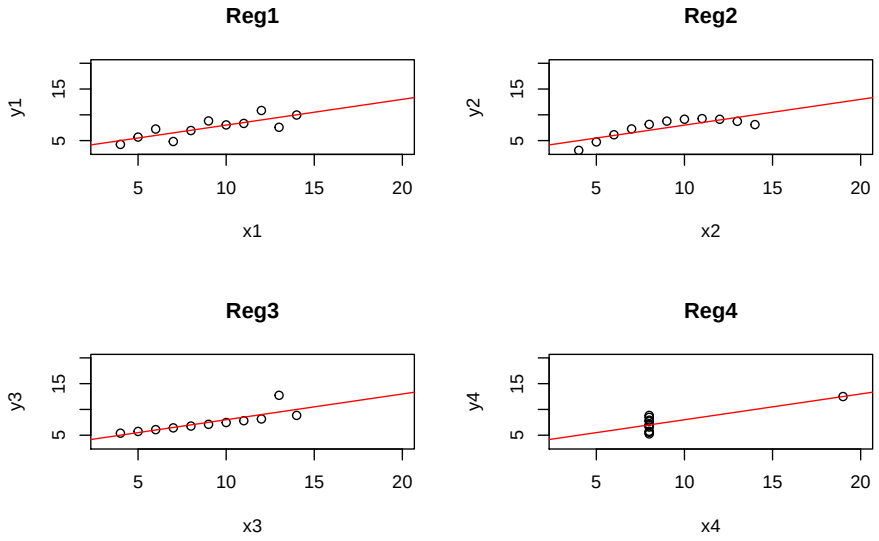
```r
linearMod2 <- lm(y2 ~ x2, data=reg2)
print(linearMod2)
#>
#> Call:
#> lm(formula = y2 ~ x2, data = reg2)
#>
#> Coefficients:
#> (Intercept)           x2
#>       3.001        0.500
```

```r
linearMod3 <- lm(y3 ~ x3, data=reg3)
print(linearMod3)
#>
#> Call:
#> lm(formula = y3 ~ x3, data = reg3)
#>
#> Coefficients:
#> (Intercept)           x3
#>      3.0025       0.4997
```

```r
linearMod4 <- lm(y4 ~ x4, data=reg4)
print(linearMod4)
#>
#> Call:
#> lm(formula = y4 ~ x4, data = reg4)
#>
#> Coefficients:
#> (Intercept)           x4
#>      3.0017       0.4999
```

```r
a=c(3,20)
b=c(3,20)
par(mfrow=c(2,2))
plot(x1,y1,main = "Reg1",xlim=a,ylim=b)
```

```
abline(linearMod1,col="red")
plot(x2,y2,main = "Reg2",xlim=a,ylim=b)
abline(linearMod2,col="red",)
plot(x3,y3,main = "Reg3",xlim=a,ylim=b)
abline(linearMod3,col="red")
plot(x4,y4,main="Reg4",xlim=a,ylim=b)
abline(linearMod4,col="red")
```



Die Regressionsgerade ist ungefähr gleich bei allen Datensätzen. Dennoch sind die Stichproben sehr unterschiedlich strukuriert. Bei Reg 1 bilden die Punkte eine Art Welle. Bei reg 2 eine Parabel, bei Reg3 eine Linea (bis auf ein Ausreißer) und bei Reg4 gibt es zu einem x-Wert mehrere y-Werte und sonst nur einen anderen x-Wert dazu.

Aufgabe 1[c]:

plot.regression<-function(daten,model,header,x.lim,y.lim){ x<-daten[[1]] y<-daten[[2]]

#Regressionsplot und scatterplot plot(x,y,main=header,xlim=x.lim,ylim=y.lim,xlab="X-Achse",ylab="Y-Achse") abline(model,col="red") }

Aufgabe 1[d]:

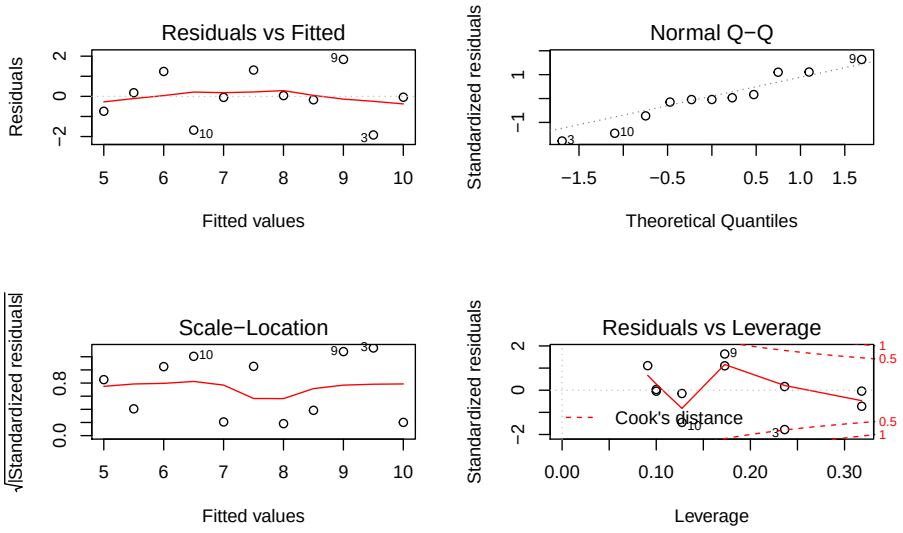
```
a=c(1,10)
b=c(2,10)
par(mfrow=c(2,2))
plot.regression(reg1,linearMod1,"Reg1",a,b)
```

```
#> Error in plot.regression(reg1, linearMod1, "Reg1", a, b): konnte Funktion "plot.regressio
plot.regression(reg2,linearMod2,"Reg2",a,b)
#> Error in plot.regression(reg2, linearMod2, "Reg2", a, b): konnte Funktion "plot.regressio
```

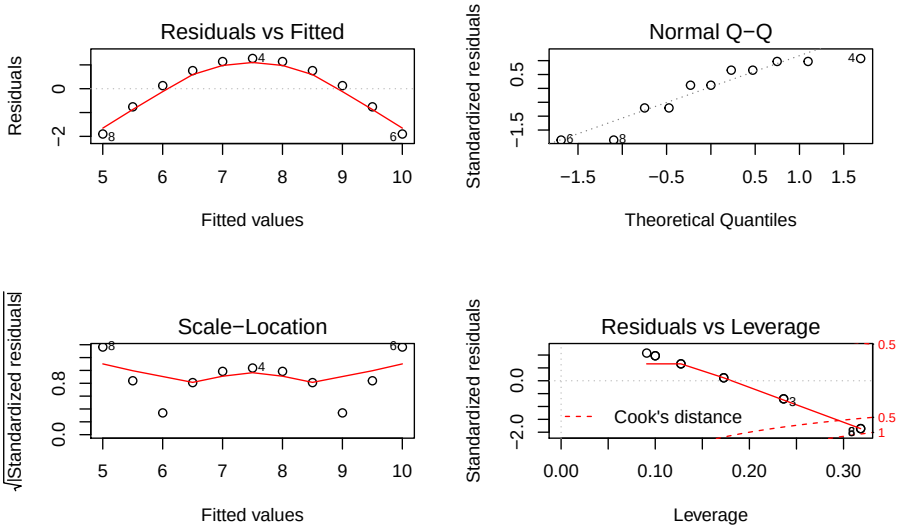
```
plot.regression(reg3,linearMod3,"Reg3",a,b)
#> Error in plot.regression(reg3, linearMod3, "Reg3", a, b): konnte Funktion "plot.regressio
plot.regression(reg4,linearMod4,"Reg4",a,b)
#> Error in plot.regression(reg4, linearMod4, "Reg4", a, b): konnte Funktion "plot.regressio
```

Aufgabe 1[e]

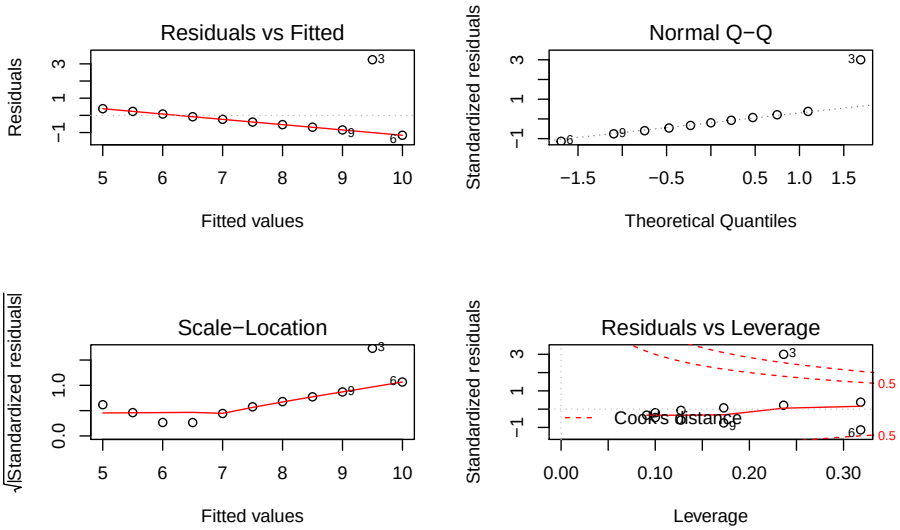
```
par(mfrow=c(2,2))
plot(linearMod1)
```




```
plot(linearMod2)
```

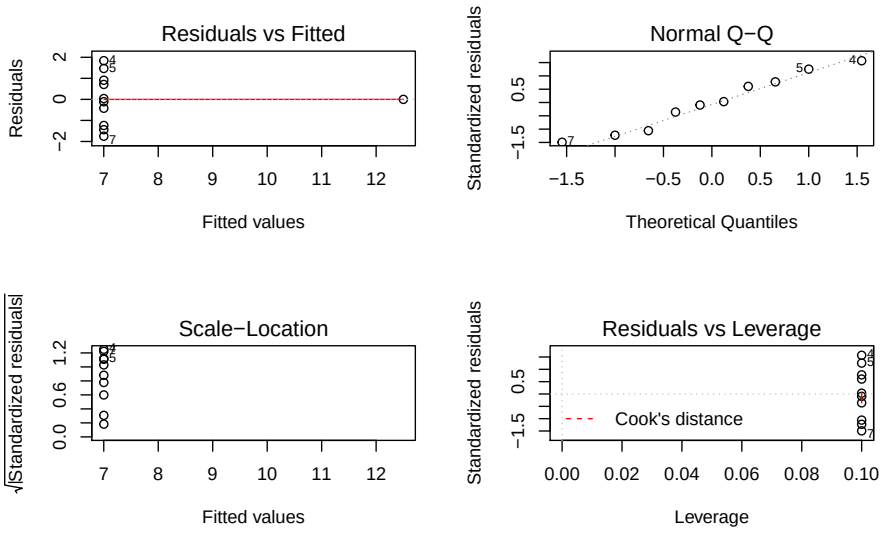



```
plot(linearMod3)
```




```
plot(linearMod4)
#> Warning: not plotting observations with leverage one:
#>   8
```

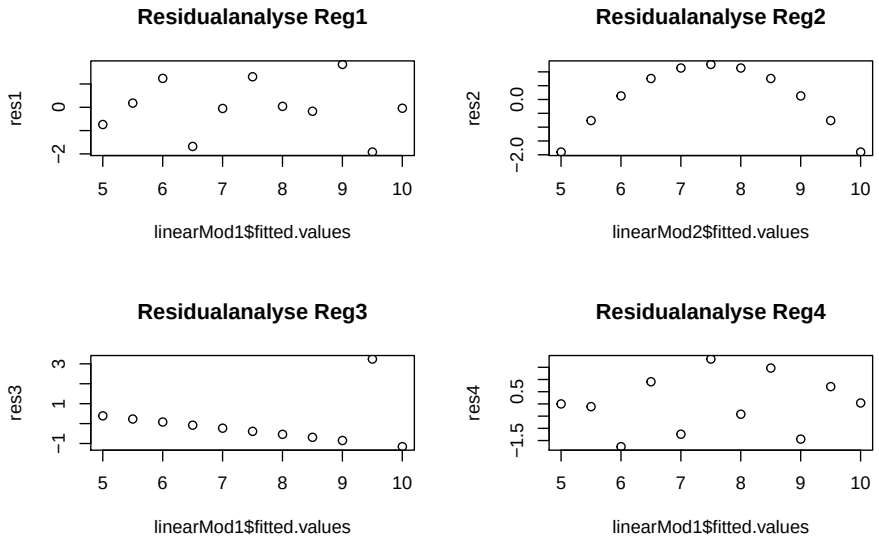
```
#> Warning: not plotting observations with leverage one:
#>   8
```




```
res1<-linearMod1$residuals
res2<-linearMod2$residuals
res3<-linearMod3$residuals
res4<-linearMod4$residuals



par(mfrow=c(2,2))
plot(linearMod1$fitted.values,res1,main="Residualanalyse Reg1")
plot(linearMod2$fitted.values,res2,main="Residualanalyse Reg2")
plot(linearMod1$fitted.values,res3,main="Residualanalyse Reg3")
plot(linearMod1$fitted.values,res4,main="Residualanalyse Reg4")
```



Gesamtfazit:

Wie der folgenden Grafiken zu entnehmen ist, unterstellt das Regressionsmodell stets einen linearen Zusammenhang, wie er in Beispiel (a) vorliegt. Selbst wenn ein nichtlinearer Zusammenhang wie im Beispiel (b) existiert, legt das Regressionsverfahren eine Gerade durch die umgekehrte Parabel. Daher wissen wir nicht, worauf eine schlechte Modellanpassung zurückzuführen ist. Besteht wirklich kein Zusammenhang zwischen beiden Merkmalen oder ist dieser Zusammenhang nur nicht linear ? Die Beispiele (c) und (d) veranschaulichen den Einfluß von Ausreißern auf die Lage der Regressionsgeraden. Im Beispiel (c) zieht der Außreiser die Regressionsgerade nach oben, im zweiten Beispiel konstituiert der Ausreißer einen Scheinzusammenhang. Ohne ihn läge im Beispiel (d) kein Zusammenhang vor, da alle Beobachtungen über identische X-Werte verügen und somit keine Varianz aufweisen.

Aufgabe 2[a]:

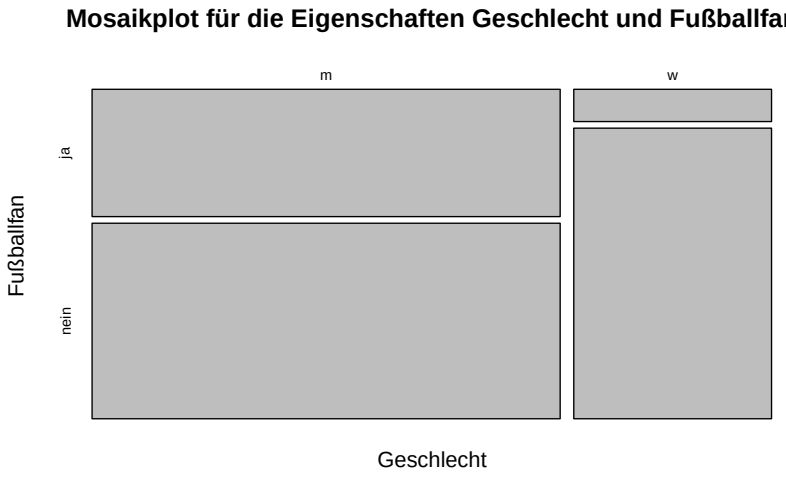
```
data <- read.csv("~/Desktop/Uni/M. Sc. Data Science/1.Semester /03_EDA/UmfrageBis2019.csv",h
G<-data$Geschlecht
F<-data$Fussballfan
table(G,F)
#>    F
#> G   ja nein
#>   m 28   43
#>   w  3   27
chisq.test(table(G,F))
#>
#>  Pearson's Chi-squared test with Yates' continuity correction
```

```
#>
#> data:  table(G, F)
#> X-squared = 7.2624, df = 1, p-value = 0.007041
```

Der p-Wert 0.007041 ist kleiner als 0.05. Somit haben wir nachgewiesen, dass zwischen Geschlecht und Fussbalfan ein statistisch signifikanter Zusammenhang besteht.

Nun kommen wir zum Mosaikplot

```
mosaicplot(table(G,F),xlab="Geschlecht",ylab="Fußballfan",main="Mosaikplot für die Eigenscha
```



Über die Vierfeldertafel erkennt man sofort, dass kaum eine Frau Fußballfan ist, während viele Männer Fußballfans sind. Fußballfans sind bei 28 Männer und 3 Frauen. Von denjenigen, die keine Fußballfans sind, sind es widerum mehr Männer, die keine Fußballfans sind als Frauen.

Im Mosaikplot wird erkenntlich, das ein wesentlich größerer Anteil von Männern Fußballfan sind als Frauen. Es gibt kaum Frauen, die Fußballfans sind. Dies erkennt man durch die verschiedenen Längen der Männlich/Weiblichen Boxen, die Fußballfans sind.Im Gegensatz zum Viertafelplot, wird jetzt klar, das prozentual gesehen mehr Frauen keine Fußballfans sind als Männer. Dies liegt daran, das das Mosaikplot noch die Geschlechtergewichtung visualisiert. Demensprechend ist die Länge der Box der Frauen die keine Fußballfans sind größer als die der Männer.

Daraus kann man schließen, dass es einen Zusammenhang zwischen den Eigenschaften Geschlecht und Fußballfann schließen. Welcher Zusammenhang?

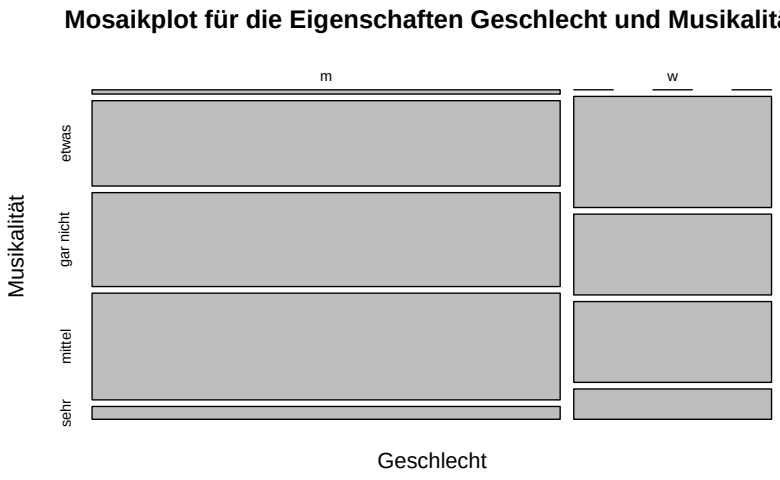
Aufgabe 2[b]:

```
M<-data$Musikalitaet
table(G,M)
#>    M
#> G      etwas gar nicht mittel sehr
#>   m  1    20        22     25    3
#>   w  0    11         8      8    3
```

Nun haben wir nicht jeweils zwei Kategorien. Musikalität besitzt fünf Kategorien. Diese determinieren nicht, ob jemand musikalisch ist oder nicht. Stattdessen wird gerankt, wie musikalisch man ist. Dies sollte man in der Analyse berücksichtigen. Dementsprechend ist die Vierfeldertafel wörtlich genommen keine Viefeldertafel. Würde man die Eigenschaften aus a und b vergleichen wollen, um die jeweiligen Zusammenhänge zu vergleichen, müsste man die kategorien Musikalität auch in zwei Kategorien einteilen.

Aus der Tafel erkennt man das gleich viele Männer wie Frauen sehr musikalisch sind.Bei allen anderen Kategorien sind wesentlich mehr Männer als Frauen vertreten. Auffällig ist auch, dass fast doppelt so viele Männer etwas musikalisch sind wie Frauen. Weierhin sind jeweils ungefähr 3-fach soviele Männer gar nicht bis mittelmäßig musikalisch.

```
mosaicplot(table(G,M),xlab="Geschlecht",ylab="Musikalität",main="Mosaikplot für die Eigensch
```



Im Mosaikplot werden die Häufigkeiten der Musikalitäten nach der Häufigkeit des weiblichen bzw. männlichen Geschlechts gewichtet. Es wird klas das

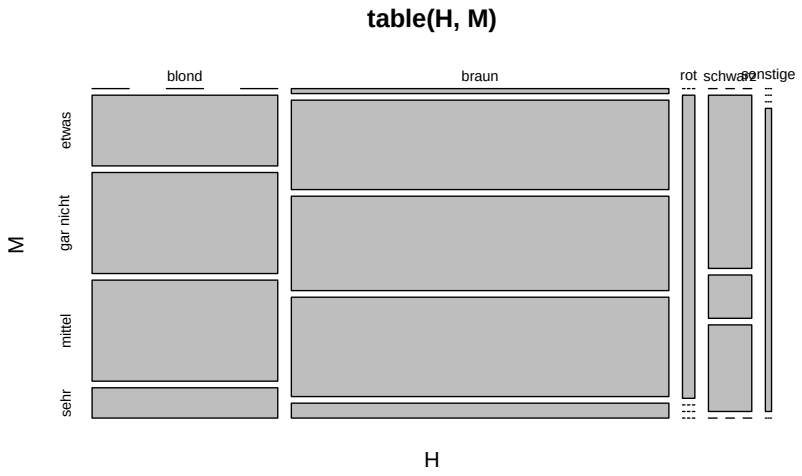
prozentual gesehen mehr Frauen sehr musikalisch sind als Männer.(In der Tafel waren es 3:3!) Die extremen Unterschiede in den Häufigkeiten bei gar nicht und mittel musikalisch fallen in dem Mosaikplot gar nicht mehr so auf. Zwar sind verhältnissmässig mehr Männer gar nicht bis mittelmässig musikalisch, aber der Unterschied ist nicht so groß. Zudem fällt auf, das im Verhältnis zu Männer und Frauen wesentlich mehr Frauen etwas musikalisch sind als Männer.(Vergleich zu Vierfeldertafel=20:11). Insgesamt kann man definitiv einen Zusammenhang erkennen, das mehr Frauen musikalisch sind und Männer eher unmusikalisch bzw. ein bisschen musikalisch.

Aufgabe 2[c]:

```
H<-data$Haarfarbe
table(H,M)
#>           M
#> H             etwas gar nicht mittel sehr
#>   blond     0     7        10     10    3
#>   braun     1    18        19     20    3
#>   rot       0     2         0      0    0
#>   schwarz   0     4         1      2    0
#>   sonstige  0     0         0      1    0
```

Nun haben wir jeweils fünf Kategorien zu einer Eigenschaft zugeordnet. Hauptsächlich bei bei Blonden und Braunhaarigen treten Ereignisse zu den Kategorien auf. Es lässt sich kein Zusammenhang erkennen zwischen Haarfarbe und Musikalität.

```
mosaicplot(table(H,M))
```



Auch im Mosaikplot wird erkenntlich, dass es keinen Zusammenhang zwischen den beiden Eigenschaften gibt. Zum Beispiel sind Blonde im Mosaikplot verhältnismässig "sehr musikalischer" als Braunhaarige. Dennoch sind Blonde verhältnissmäßig eher gar nicht musikalisch als musikalisch verglichen mit Braunhaarigen. Daher schließen wir auf keinen Zusammenhang zwischen den Eigenschaften.

Aufgabe 2[d]: Am ehesten lässt der letzte Mosaikplot auf Unabhängigkeit hindeuten. Es lässt sich keine Struktur erkennen, welche Haarfarbe zu welcher Musikalität beiträgt.Zum Beispiel sind Blonde im Mosaikplot verhältnismässig musikalischer als Braunhaarige. Dennoch sind Blonde verhältnissmäßig eher gar nicht musikalisch als musikalisch verglichen mit Braunhaarigen. Daher schließen wir auf keinen Zusammenhang zwischen den Eigenschaften.

Anmerkungen/Korrektur
