?bung 12

Explorative Datenanalyse und Visualisierung

Wintersemester 2019

S. D?hler (FBMN, h_da)

Name:

Aufgabe 27. Arbeiten Sie das Kapitel 22 in *R for Data Science* durch.

Aufgabe 28. Analysieren Sie die bereinigten Daten `UmfrageBis2019.csv` (s. Aufgabe 21) mit `ggplot2` (arbeiten Sie wieder mit den neuen Spaltennamen...).

a) Erzeugen Sie ein Histogramm der Variablen `Groesse`

   i) F?r die Gesamtpopulation

   ii) Getrennt nach Geschlechtern (arbeiten Sie mit facets)

   Was f?llt Ihnen bei der Default-Klasseneinteilung auf? Verwenden Sie zus?tzlich die Aufteilung nach der Diaconis-Friedman-Methode.

b) Stellen Sie Kerndichtesch?tzer der Variablen `Groesse` dar.

   – Getrennt nach Geschlechtern, jedoch in einer gemeinsamen Grafik (mit verschiedenen Farben und Schraffierungen)

   – F?gen Sie einen rug-Plot mit entsprechenden Farben hinzu.

c) Stellen Sie die empirische Dichtefunktion der Variablen `Groesse` dar.

   – Getrennt nach Geschlechtern, jedoch in einer gemeinsamen Grafik (mit verschiedenen Farben und Schraffierungen)

   – F?gen Sie einen rug-Plot mit entsprechenden Farben hinzu.

d) Erzeugen Sie Box- und Violin-Plots der Variablen `Groesse` getrennt nach Geschlechtern, jedoch in jeweils einer gemeinsamen Grafik.

e) Erzeugen Sie einen Scatterplot der Variablen `Schuhgroesse` (y-Achse) und `Groesse` (x-Achse).

   – F?rben Sie die Datenpunkte nach Geschlecht

   – Passen Sie pro Geschlecht jeweils eine lineare Regression an und stellen Sie die resultierenden Regressionsgeraden mit der passenden Farbe zusammen mit den Daten dar.

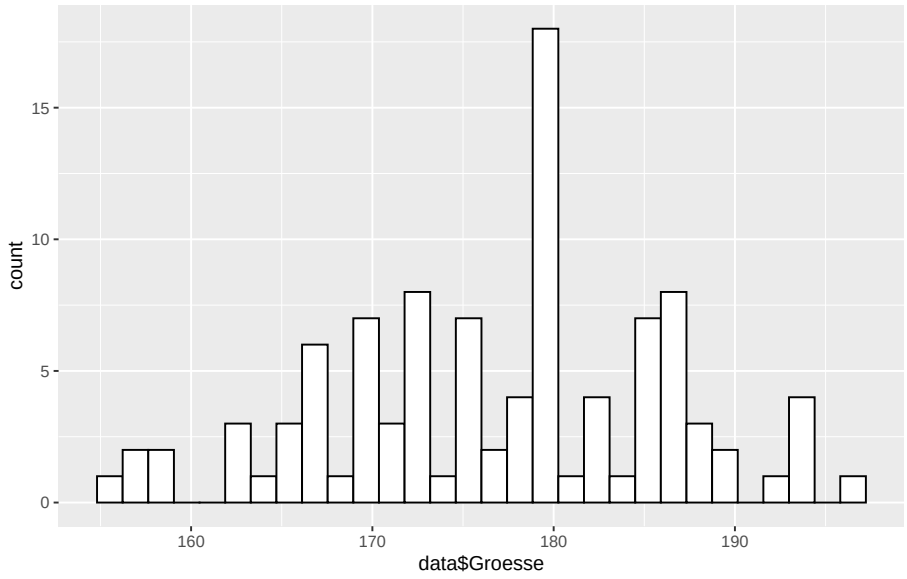
L?sung

Zun?cht laden wir die Daten runter.

```
data <-read.csv("C:/Users/Roman/Dropbox/hda/Explorative_Datenanalyse/uebungen/uebung12/Umfra
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"
```

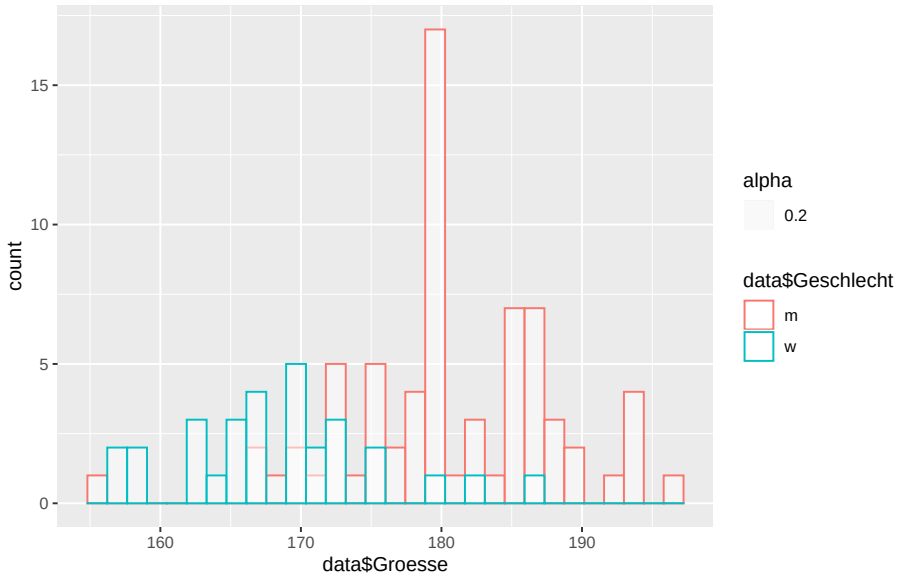
[a)]

```
library(ggplot2)
ggplot(data=data, aes(data$Groesse)) + geom_histogram(colour="black",fill="white")
```



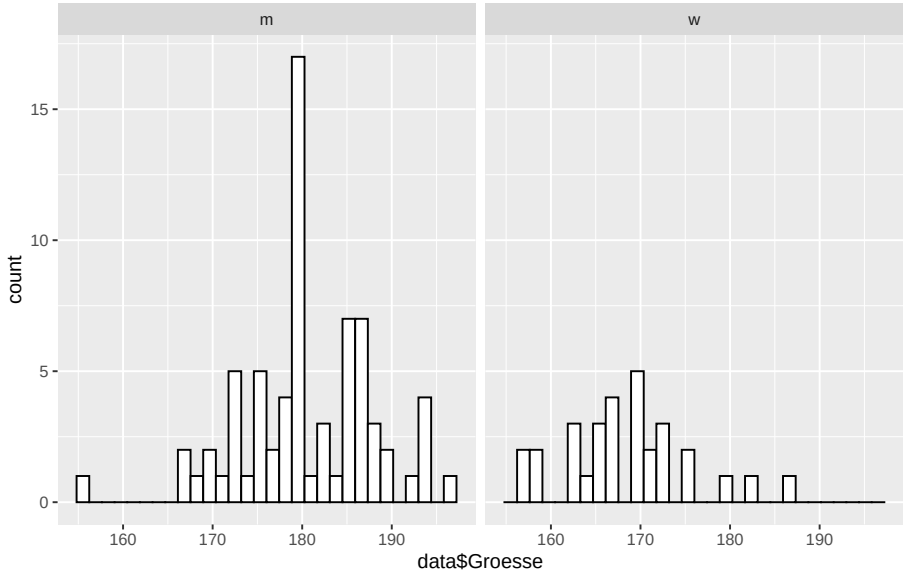
In einer Grafik nach Geschlecht aufgeteilt:

```
ggplot(data, aes(x=data$Groesse, color=data$Geschlecht,alpha=.2)) +
  geom_histogram(fill="white", position="identity")
```



In zwei Grafiken nach Geschlecht aufgeteilt unter Verwendung von Facets:

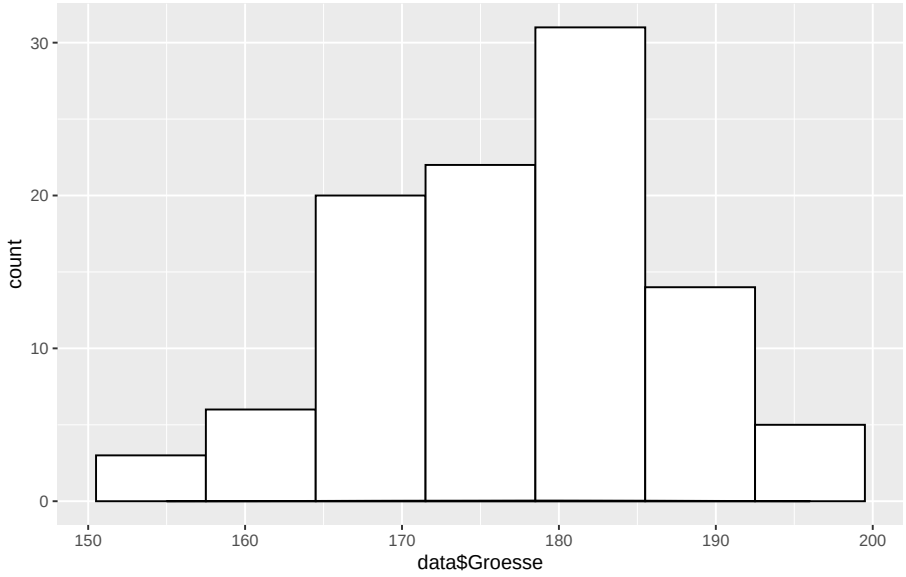
```
ggplot(data, aes(x=data$Groesse))+
  geom_histogram(color="black", fill="white")+
  facet_grid(.~data$Geschlecht)
```



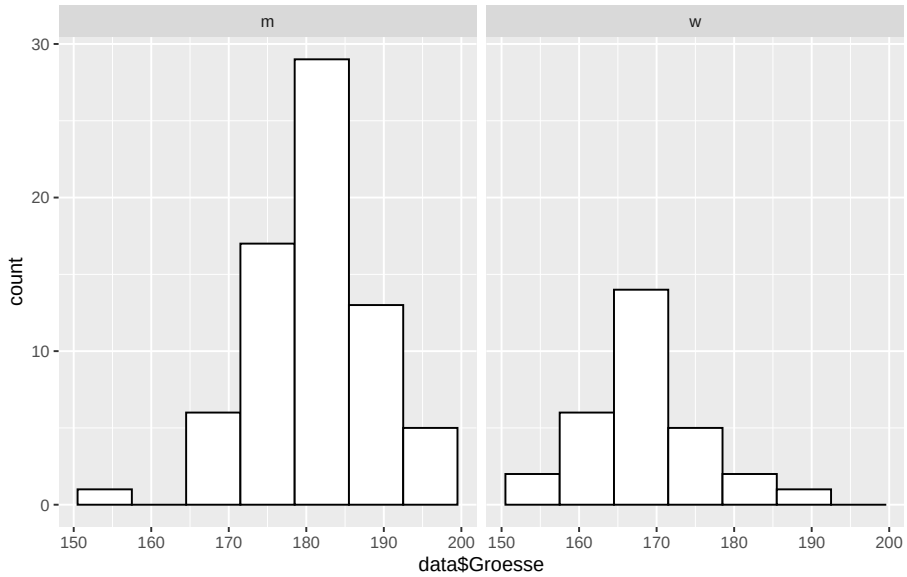

```
#qplot(data$Groesse, geom="histogram",xlab="Gr??e",xlim=c(150,200),binwidth=1,col=I("blue"),
#qplot(data$Groesse, geom="histogram",xlab="Gr??e",xlim=c(150,200),binwidth=1,col=I("blue"),
```

Die Bandbreite ist sehr unvorteilhaft gew?hlt, da es zwischen jeden zehn Einheitn 7 Bins gibt. Wir werden mit der Freedman-Diacons Methode, die Bandbreite optimieren.

```
nclass.FD(data$Groesse)
#> [1] 7
ggplot(data=data, aes(data$Groesse)) + geom_histogram(colour="black", fill="white",binwidth
```

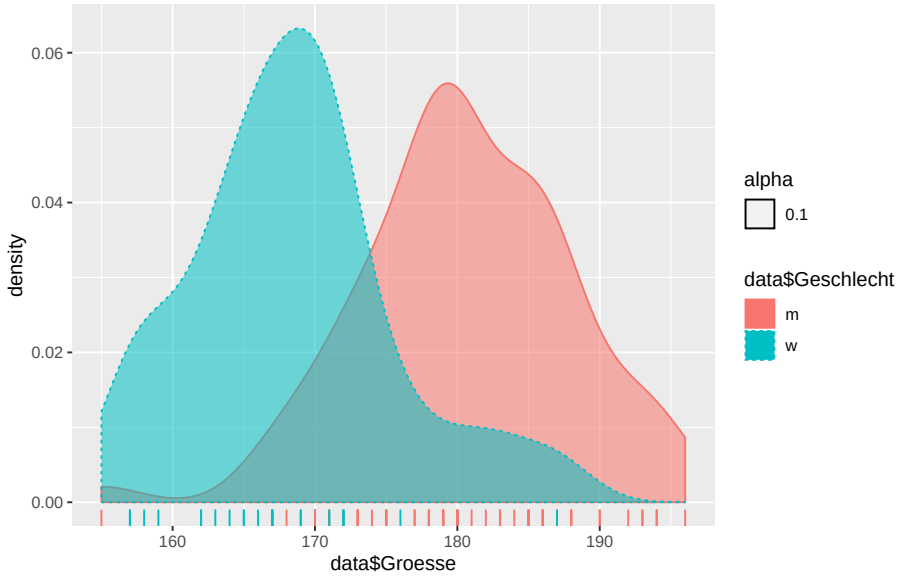



```
G<-ggplot(data, aes(x=data$Groesse))
  G+geom_histogram(color="black", fill="white",binwidth = 7)+
  facet_grid(.~data$Geschlecht)
```



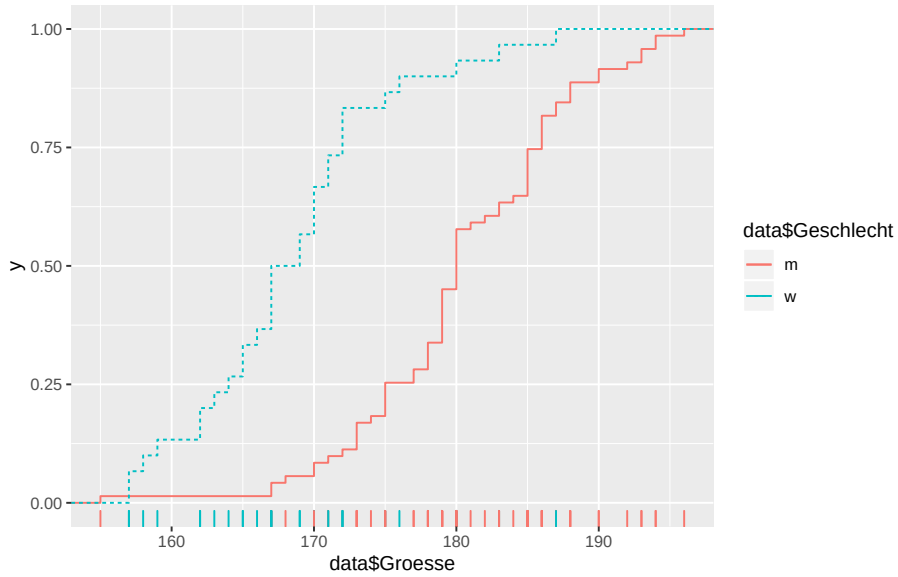
[b)] Nun wollen wir die Kerndichtesch?tzer der Variablen “Gr??e” darstellen.

```
G+geom_density(aes(col=data$Geschlecht,fill=data$Geschlecht,alpha=.1,linetype=data$Geschlech
```



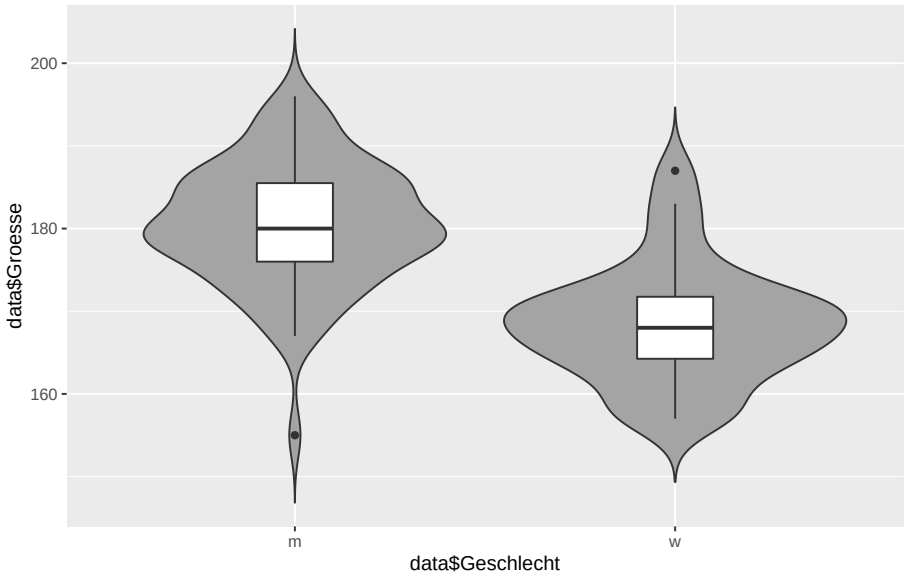
[c)]

```
ggplot(data,aes(x=data$Groesse))+stat_ecdf(aes(col=data$Geschlecht,linetype=data$Geschlecht)
```



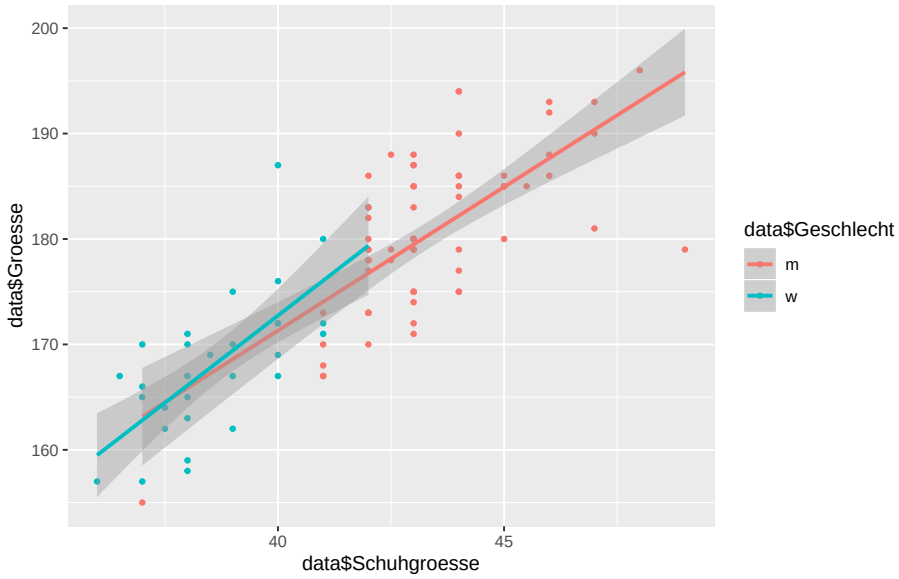
[d)]

```
ggplot(data,aes(x=data$Geschlecht,y=data$Groesse))+geom_violin(trim=FALSE,fill='#A4A4A4',)+g
```



[e)]

```r
ggplot(data, aes(x=data$Schuhgroesse, y=data$Groesse,col=data$Geschlecht)) +
  geom_point(size=1)+geom_smooth(method=lm)
```



Anmerkungen/Korrektur
