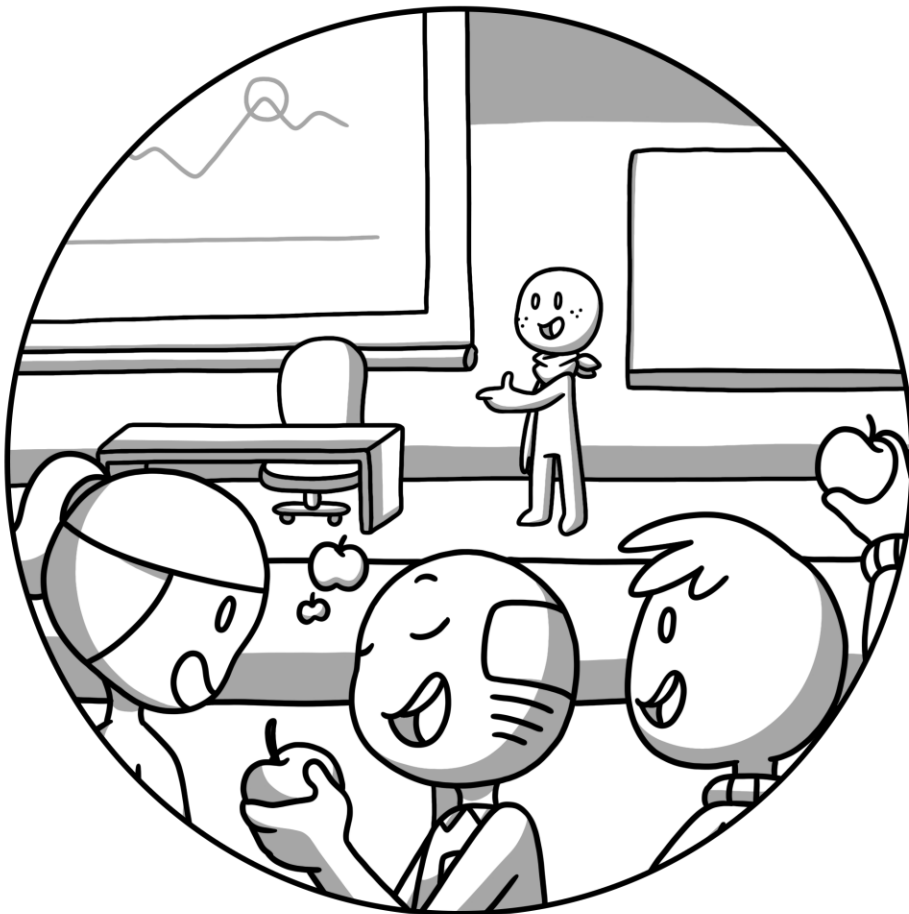


Lagemaße



COMIXPLAIN

Dieser Comic wurde im Zuge des Forschungsprojekts Comixplain, gefördert von der Fachhochschule St. Pölten im Rahmen des Innovation Call 2022, erstellt.

Projektteam:

Victor-Adriel De-Jesus-Oliveira
Hsiang-Yun Wu
Christina Stoiber
Magdalena Boucher
Alena Ertl

Kontakt:

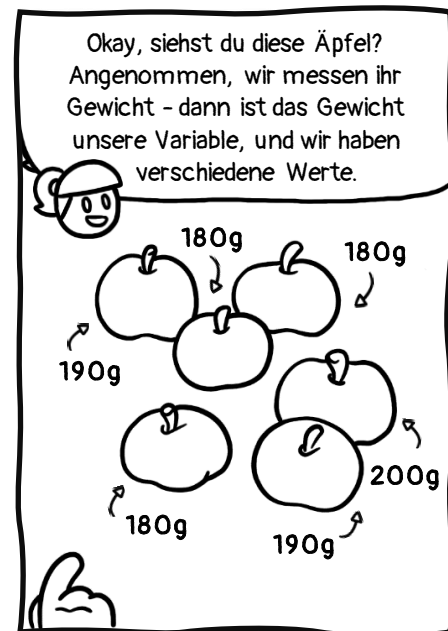
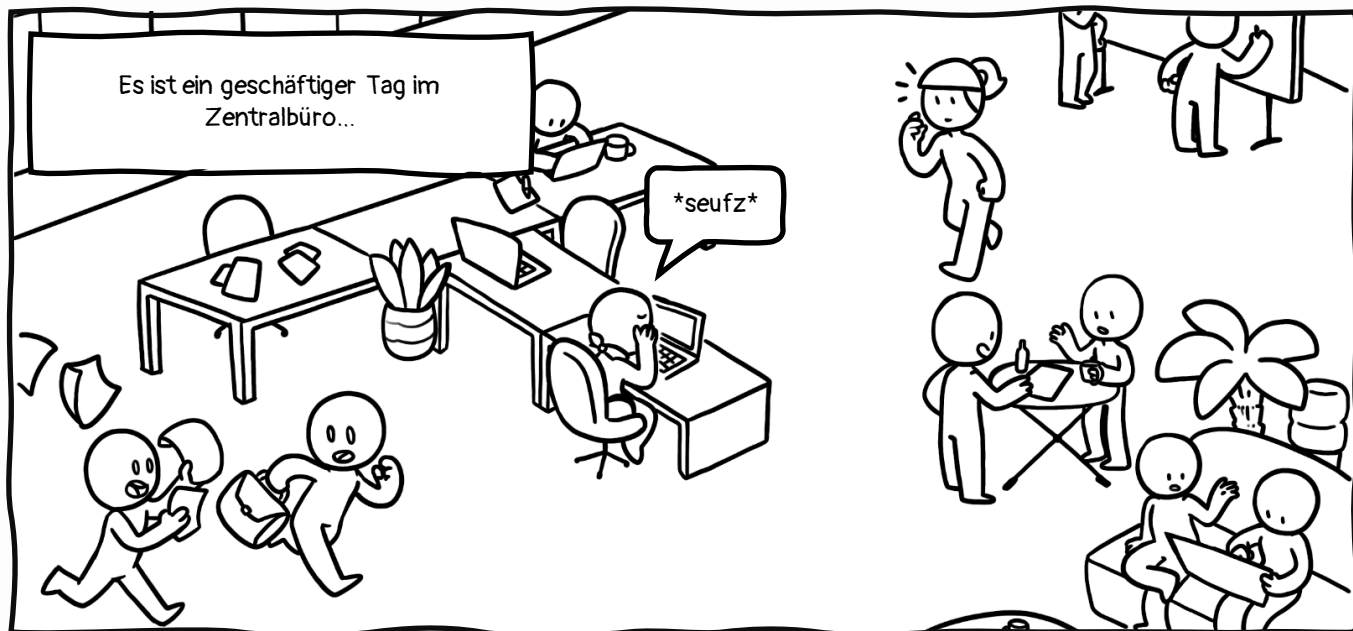
victor.oliveira@fhstp.ac.at

Illustrationen:

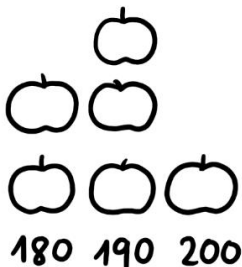
Magdalena Boucher & Alena Ertl



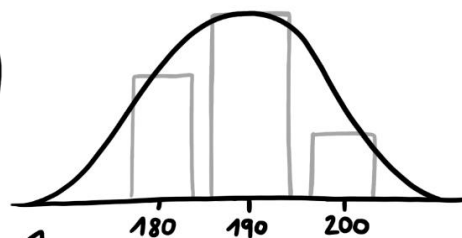
<https://fhstp.github.io/comixplain>



Der beste Weg, eine Variable zu beschreiben, ist zu zeigen, welche Werte darin vorkommen, und wie oft. Das nennt man die **VERTEILUNG** der Werte einer Variable.



Wenn wir das Gewicht unserer Äpfel visualisieren, würde die Verteilung so aussehen, weil alle Äpfel im Korb ungefähr gleich viel wiegen.

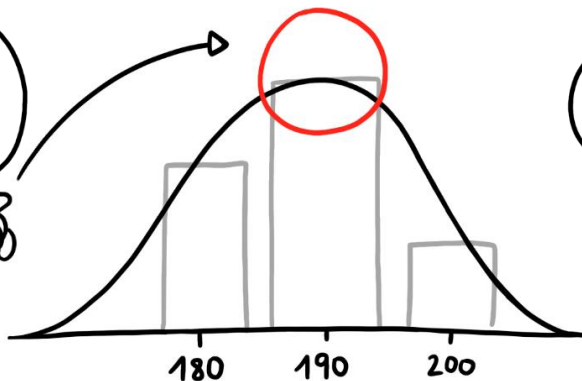


Weight of apples

Der Graph hat irgendwie die Form einer Glocke...



Ja. Wir können so schätzen, dass das Gewicht der meisten Äpfel im Korb um die Spitze der Glocke herum liegt.

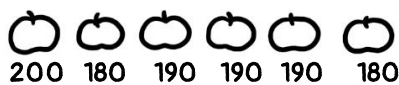


Ah, okay. Aber... wie würde ich das berechnen?

Dieser Mittelpunkt ist ein gutes Maß für die Verteilung unserer Daten - deshalb nennt man einen solchen Wert auch **LAGEMAß**.

Es gibt verschiedene Arten von Lagemaßen. Der **MITTELWERT** ist das bekannteste Maß, und er ist auch leicht zu berechnen.

Hier sind unsere sechs Äpfel:



Um den Mittelwert zu berechnen, addieren wir einfach alle Werte...

$$200 + 180 + 190 + 190 + 190 + 180$$

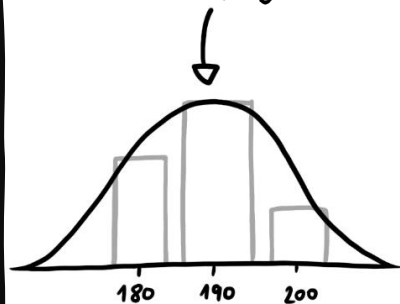
...und dividieren dann durch die Anzahl unserer Äpfel...

$$(200 + 180 + 190 + 190 + 190 + 180)$$

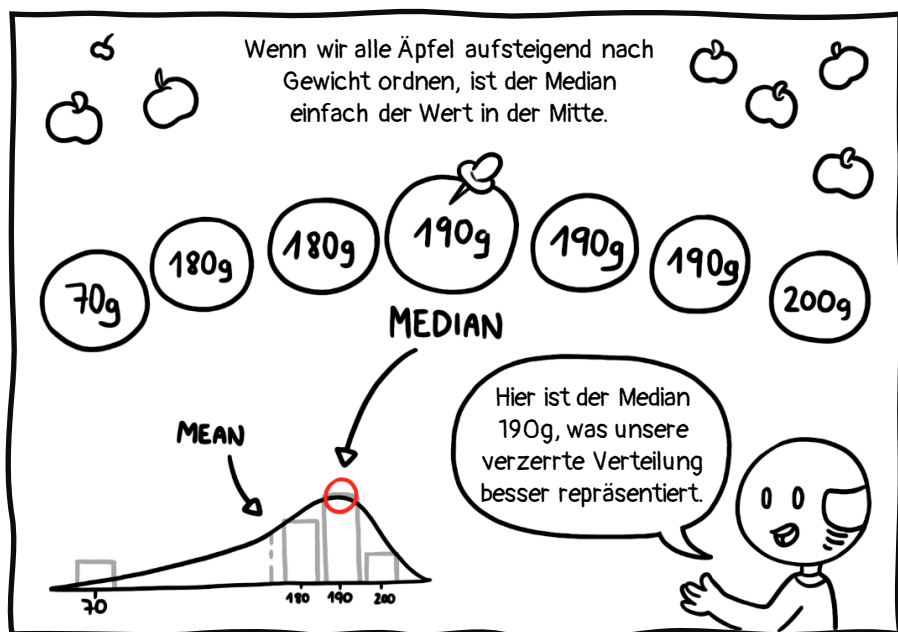
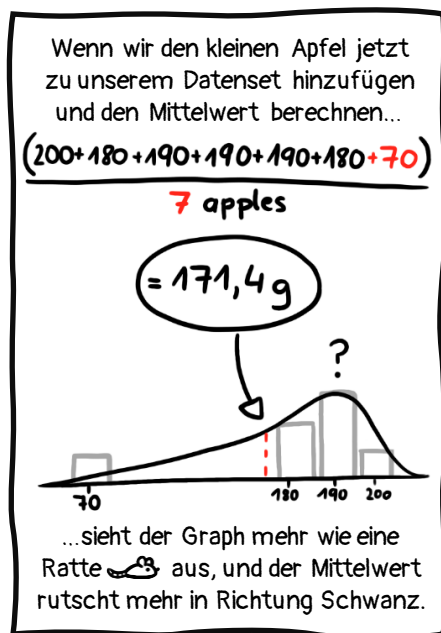
6

Das ergibt einen Mittelwert von:

188,3g



Oh, das ist wirklich fast die Spitze unserer Glockenkurve!



In diesem Fall hat unser Graph auf einmal zwei Hügel:



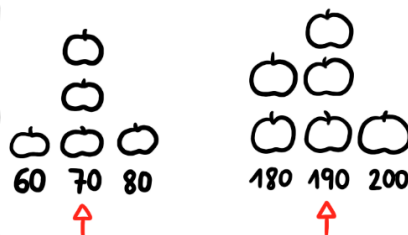
Zwei Ratten
die sich küssen!

Ähh, klar...

weight of tiny & big apples

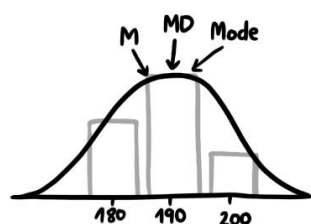
Naja, jedenfalls gibt es noch ein
Lagemaß namens **MODUS**, das
wir verwenden können, wenn
unsere Wertverteilung mehrere
Hügel hat.

Der Modus beschreibt die Werte,
die in einem Datenset am
häufigsten auftreten.



In diesem Fall haben wir
mehrere Modi, aber es kann
auch Datensets mit nur einem,
oder sogar gar keinem geben.

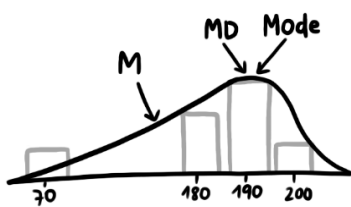
Wir können Mittelwert, Median und Modus für ganz verschiedene Stichproben von Äpfeln verwenden - aber es wird oft vorkommen, dass eines der Maße die Daten besser beschreibt als ein anderes.



180, 180, 190, 190, 190, 200

$M = 188,3$
 $MD = 190$
 $\text{Modus} = 190$

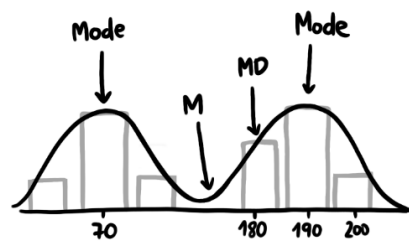
Gute Maße



70, 180, 181, 190, 191, 191, 200

$M = 171,8$
 $MD = 190$
 $\text{Modus} = 191$

Gute Maße

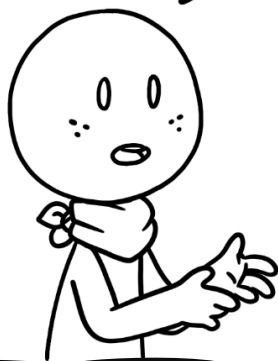


60, 70, 70, 70, 80, 180, 180, 190,
190, 190, 200

$M = 134,5$
 $MD = 180$
 $\text{Modus} = 70 \text{ \& } 190$

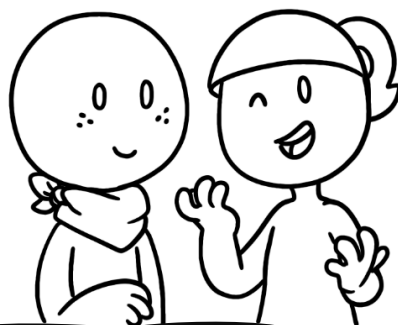
Gutes Maß

Okay, danke... Ich habe viel
dazugelernt. Jetzt muss ich das
nur noch auf meine eigenen Daten
übertragen. Sie sind aus einer App,
die Herzfrequenz misst.



Nutzer ID	Herz-Frequenz (bpm)	Nutzungszeit	Bewertung
1	45	13:00	1
2	50	9:00	5
3	55	10:00	3
4	57	9:00	4
5	63	14:00	5
6	70	15:00	5
7	65	16:00	4
8	75	15:00	2

Das sollte machbar sein - schau
dir deine Daten an und folge
denselben Schritten, die wir
gerade mit den Äpfeln gemacht
haben. Du kannst die nächste
Seite für Notizen verwenden.





Bevor du umblätterst: Versuche, den Mittelwert, Median und Modus für jede Variable in der Tabelle zu berechnen. Entscheide, welches Maß am besten für welche Variable geeignet ist. Du kannst auf dieser Seite Notizen machen!



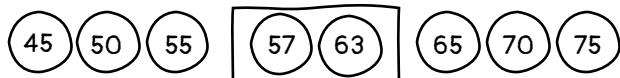
Du kannst die Lösungen auf der nächsten Seite zum Vergleich anschauen und auf dieser Seite weitere Notizen machen.

HERZFREQUENZ

Berechnung des MITTEL WERTS:

$$\frac{45+50+55+57+63+70+65+75}{8 \text{ users}} = \frac{480}{8} = 60 \text{ bpm}$$

Berechnung des MEDIANS:



Wenn es zwei mittlere Werte gibt, dann berechnet sich der Median aus dem Mittelwert der beiden Werte:

$$(57+63)/2 = 60 \text{ bpm}$$

Berechnung des MODUS:

45, 50, 55, 57, 63, 70, 65, 75

Jeder Wert kommt nur ein Mal vor - es gibt also keinen Modus!

Wenn die Verteilung der Werte symmetrisch, ohne Verzerrungen, ist, dann sind Mittelwert und Median gleich.



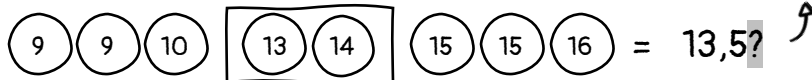
HÄUFIGSTENUTZUNGSZEIT

Berechnung des MITTEL WERTS:

$$\frac{9+9+10+13+14+15+15+16}{8 \text{ Nutzer*innen}} = \frac{101}{8} = 12,6?$$

Nutzungszeit ist kein quantitativer Wert - daher macht die Berechnung von Mittelwert und Median keinen Sinn!

Berechnung des MEDIANS:



Berechnung des MODUS:

9:00, 10:00, 13:00, 14:00, 15:00, 16:00
2x 1x 1x 1x 2x 1x = 2 Modi: 9:00 & 15:00

Der Modus ist nicht nur für multimodale Verteilungen geeignet, sondern auch für ordinale und kategoriale Daten.



STERNEBEWERTUNG

Berechnung des MITTEL WERTS:

$$\frac{1+2+3+4+4+5+5+5}{8 \text{ users}} = \frac{29}{8} = 3,6 \text{ stars}$$

Berechnung des MEDIANS:



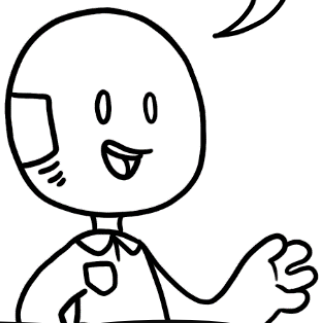
Berechnung des MODUS:

1 2 3 4 5
1x 1x 1x 2x 3x = 5 stars

Für Datensätze mit einer verzerrten Verteilung ist der Median ein besseres Lagemaß.

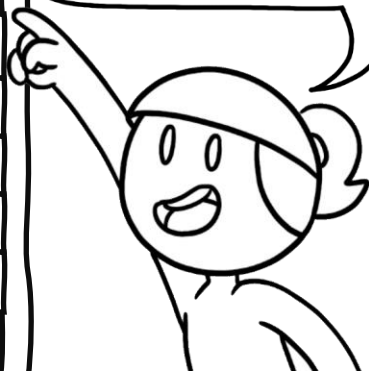


Mit Programmiersprachen wie **R** kannst du Lagemaße für Attribute in großen Datensatz berechnen. Mit libraries für R, wie etwa **tidyverse**, kannst du die Verteilung der Daten schnell visualisieren.



	model	year	hwy
1	jetta	1999	44
2	corolla	2008	37
3	civic	2008	36
4	civic	2008	36
5	corolla	1999	35
6	altima	2008	32
7	sonata	2008	31
+ 227 weitere Einträge			

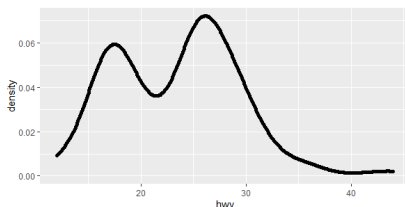
In tidyverse hast du Zugriff auf Datensätze wie **mpg** mit Kraftstoffverbrauchsdaten. Darin finden sich 11 Attribute, wie Automodell, Herstellungsjahr und Kraftstoffverbrauch pro Meile auf der Autobahn (hwy).



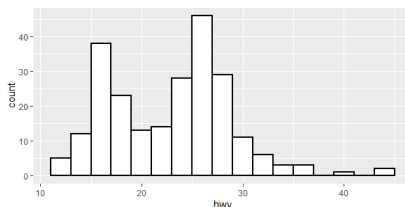
Die Wertverteilung beim Kraftstoffverbrauch (hwy) kannst du mit **ggplot**, das in tidyverse enthalten ist, in Form eines Histogramms, einer Dichtekurve, oder beidem visualisieren.



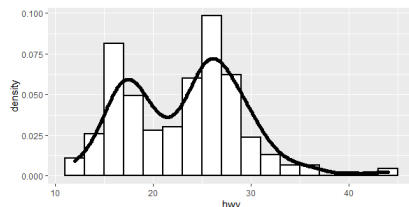
```
library(tidyverse)
plot <- ggplot(mpg, aes(x=hwy))
plot +
  geom_density()
```



```
plot +
  geom_histogram(
    colour="black",
    fill="white" )
```



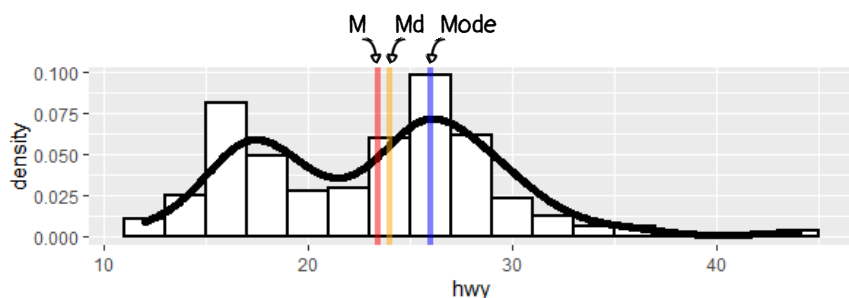
```
plot +
  geom_histogram(aes(y=..density..),
    colour="black",
    fill="white" ) +
  geom_density()
```



mean(mpg\$hwy) = 23,4

median(mpg\$hwy) = 24

library(**modeest**)
mlv(mpg\$hwy) = 26



R hat native Funktionen, um Mittelwert und Median zu berechnen. Für den Modus kannst du dir deine eigene Funktion schreiben, oder die Most Likely Values (mlv) aus der Library modeest verwenden.



Quellen:

Downey, A. (2014). Think stats: exploratory data analysis. O'Reilly Media, Inc.

Field, A. (2022). An adventure in statistics: The reality enigma. Sage.