

## 4 Daten verbildlichen

### 4.1 Lernsteuerung

---

#### 4.1.1 Standort im Lernpfad

Abb. [Abbildung 2](#) zeigt den Standort dieses Kapitels im Lernpfad und gibt damit einen Überblick über das Thema dieses Kapitels im Kontext aller Kapitel.

#### 4.1.2 Lernziele

- Sie können erläutern, wann und wozu das Visualisieren statistischer Inhalte sinnvoll ist.
- Sie kennen typische Arte von Datendiagrammen.
- Sie können typische Datendiagramme mit R visualisieren.
- Sie können zentrale Ergebnisse aus Datendiagrammen herauslesen.

#### 4.1.3 Benötigte R-Pakete

```
library(tidyverse)
library(easystats)
library(DataExplorer) # nicht vergessen zu installieren
```

#### 4.1.4 Wozu das alles?



[Quelle: GIPHY](#)



Wir müssen die Galaxis verteidigen, Kermit

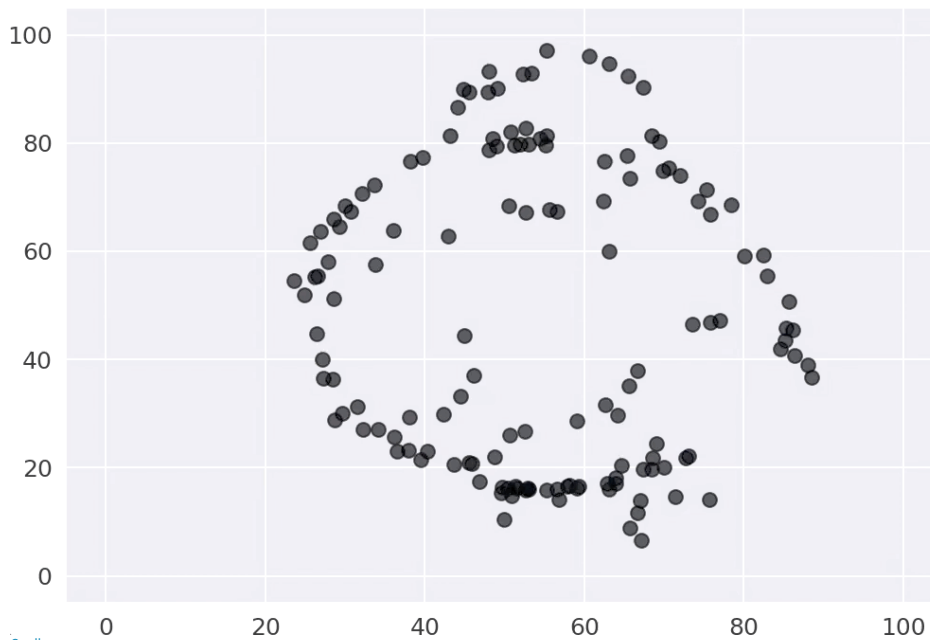


Schlock

### 4.2 Ein Dino sagt mehr als 1000 Worte

---

Es heißt, ein Bild sage mehr als 1000 Worte. Schon richtig, aber ein Dinosaurier sagt auch mehr als 1000 Worte, s. **?fig-dino1**.



X Mean: 54.2  
 Y Mean: 47.8  
 X SD : 16.76  
 Y SD : 26.93  
 Corr. : -0.060

In **fig-dino1** sieht man zwei verschiedene “Bilder”, also Datensätze: einmal einen Dino und einmal einen Kreis. Obwohl die Bilder grundverschiedene sind, sind die zentralen statistischen Kennwerte (praktisch) identisch.

Die Idee stammt von Anscombe (1973). [Hier](#) ist ein weiteres Beispiel (von Anscombe), das zeigt, dass Bilder mehr zeigen als typische Statistiken es vermögen.

Unter visueller Cortex ist sehr leistungsfähig. Wir können ohne Mühe eine große Anzahl an Informationen aufnehmen und parallel verarbeiten. Aus diesem Grund sind Datendiagramme eine effektive und einfache Art, aus Daten Erkenntnisse zu ziehen.

#### Tipp

Nutzen Sie Datendiagramme umfassend; sie sind einfach zu verstehen und doch sehr mächtig.

### 4.2.1 Datendiagramm

Ein *Datendiagramm* (kurz: Diagramm) ist ein Diagramm, das Daten und Statistiken zeigt, mit dem Zweck, Erkenntnisse daraus zu ziehen.

**Beispiel 4.1 (Aus der Forschung: Ein aufwändiges (und ansprechendes) Datendiagramm)** [Hier](#) finden Sie ein Beispiel für ein Datendiagramm, das mit R erzeugt wurde ([Scherer u. a. 2019](#)).

### 4.2.2 Ein Bild hat nicht so viele Dimensionen

[Abbildung 4.1](#) zeigt ein Bild mit mehreren Variablen. Wie man (nicht) sieht, wird es langsam unübersichtlich. Offenbar kann man in einem Bild nicht beliebig viele Variablen reinquetschen. Die “Dimensionalität” eines Diagramms hat ihre Grenzen, vielleicht bei 4-6 Variablen.

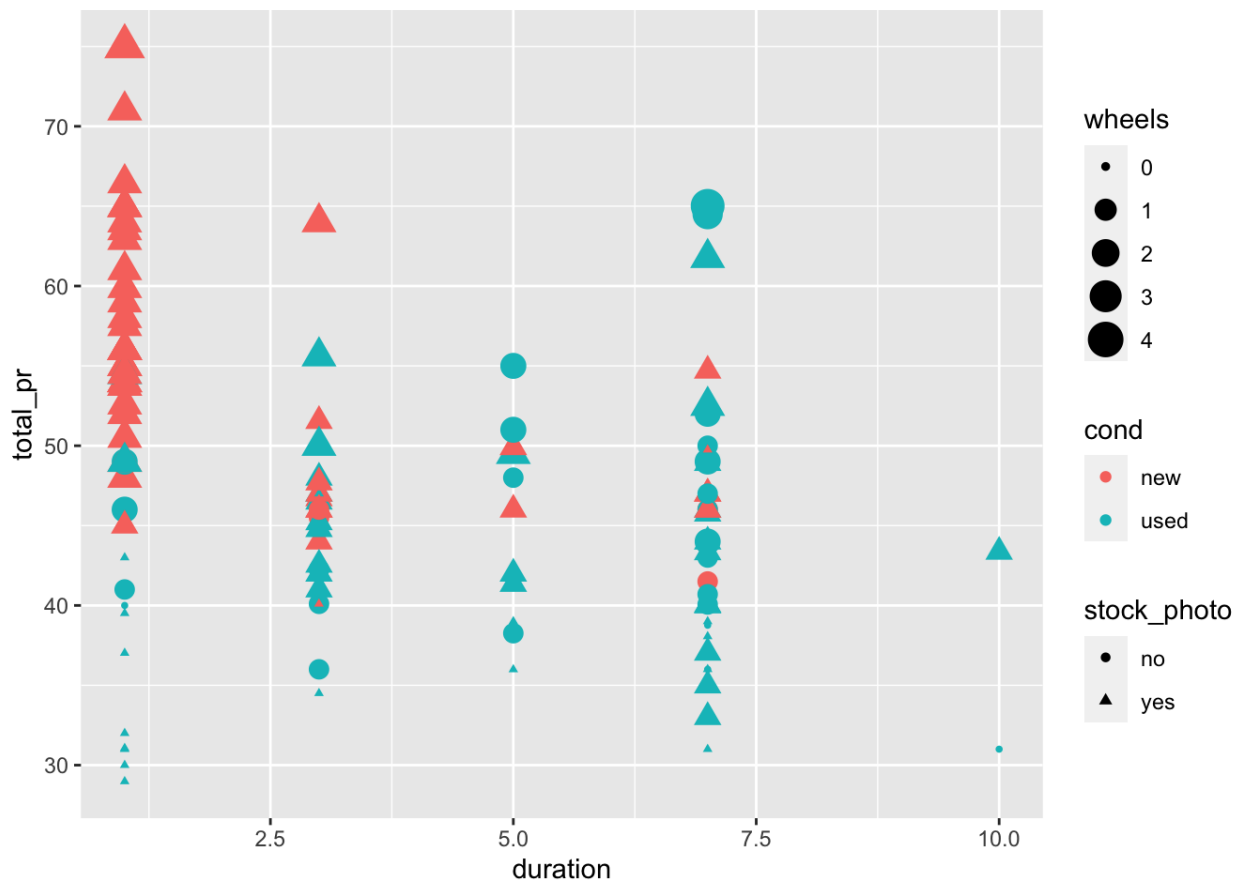


Abbildung 4.1: Ein Diagramm kann nur eine begrenzte Anzahl von Variablen zeigen

Möchten wir den Zusammenhang von vielen Variablen, z.B. mehr als 5, verstehen, kommen wir mit Bildern nicht weiter. Dann brauchen wir andere Werkzeuge: statistics to the rescue.

#### **Hinweis**

Bei klaren Zusammenhängen und wenig Variablen braucht man keine (aufwändige) Statistik. Ein Bild (Datendiagramm) ist dann (oft) ausreichend. Man könnte sagen, dass es Statistik nur deshalb gibt, weil unser Auge mit mehr als ca. 4-6 Variablen nicht gleichzeitig umgehen kann.

#### **Übungsaufgabe 4.1** Wie viele Variablen sind in [Abbildung 4.1](#) dargestellt?<sup>1</sup>

Eine weitere Möglichkeit, mehr Variablen in einem Diagramm unterzubringen, ist die "Flatlands" zu verlassen, also von 2D auf 3D zu wechseln, s. [Abbildung 4.2](#).

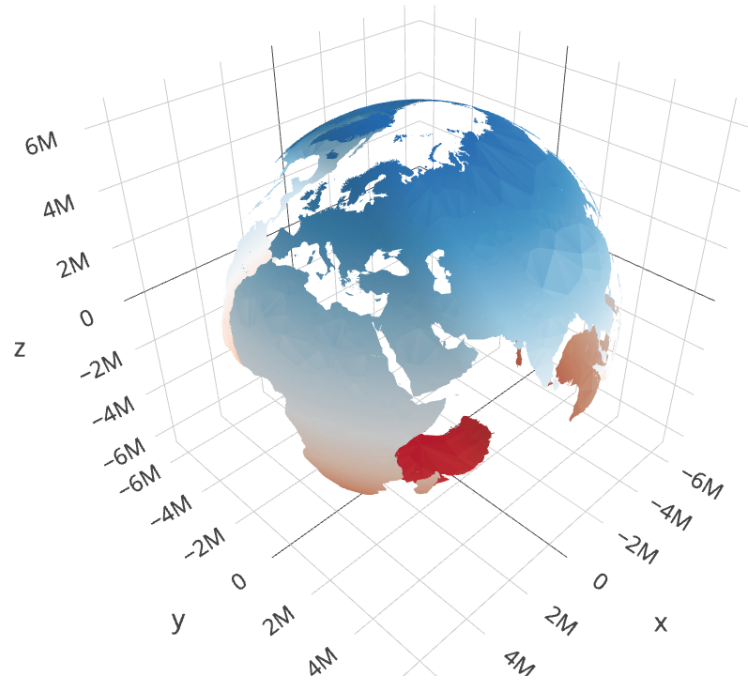


Abbildung 4.2: Eine 3D-Karte der Erde

[Quelle](#)

Etwas weniger spektakulär, aber näher an der Datenanalyse ist [Abbildung 4.3](#).

Abbildung 4.3: 3D-Punktediagramm zum Datensatz mariokart

Leider ist [Abbildung 4.3](#) nicht sehr aufschlussreich.

Daraus kann man zweierlei lernen:

- Nicht jedes Datendiagramm (ist auf Anhieb) informativ.
- Die Daten müssen ggf. erst umgeformt werden.

**Übungsaufgabe 4.2** Es gibt einen Extremwert im Diagramm. Finden Sie ihn?

## 4.3 Nomenklatur von Datendiagrammen

[Tabelle 4.1](#) zeigt eine - sehr kurze Nomenklatur - an Datendiagrammen.<sup>2</sup>

Tabelle 4.1:  
Ein (sehr kurze) Nomenklatur von Datendiagrammen

Erkenntnisziel <chr>	qualitativ <chr>	quantitativ <chr>
Verteilung	Balkendiagramm	Histogramm und Dichtediagramm
Zusammenhang	gefülltes Balkendiagramm	Streudiagramm
Unterschied	gefülltes Balkendiagramm	Boxplot

3 rows

### **Hinweis**

Wir arbeiten hier mit dem Datensatz `mariokart`. Hilfe bzw. ein Data-Dictionary (Codebook) finden Sie [hier](#).

## 4.4 Verteilungen verbildlichen

### 4.4.1 Verteilung: nominale Variable

**Definition 4.1 (Verteilung)** Eine (Häufigkeits-)Verteilung einer Variablen  $X$  schlüsselt auf, wie häufig jede Ausprägung von  $X$  ist.□

**Beispiel 4.2** [Tabelle 4.2](#) zeigt die Häufigkeitsverteilung von `cond` aus dem Datensatz `mariokart`. Die Variable hat 5 Ausprägungen; z.b. kommt die Ausprägung `new` 59 mal vor.□

Tabelle 4.2:  
Häufigkeitsverteilung von `cond` aus dem Datensatz `mariokart`

<code>cond</code> <fct>	<code>n</code> <int>
new	59
used	84

2 rows

Zugegeben, das Datendiagramm von `cond` ist nicht so aufregend, s. [Abbildung 4.4](#). Wie man sieht, besteht so ein Diagramm als *Balken*, daher heißt es *Balkendiagramm*. Man kann so ein Diagramm um 90° drehen, keine Ausrichtung ist unbedingt besser als die andere.

**Definition 4.2 (Balkendiagramm)** Ein Balkendiagramm eignet sich, um Häufigkeiten darzustellen

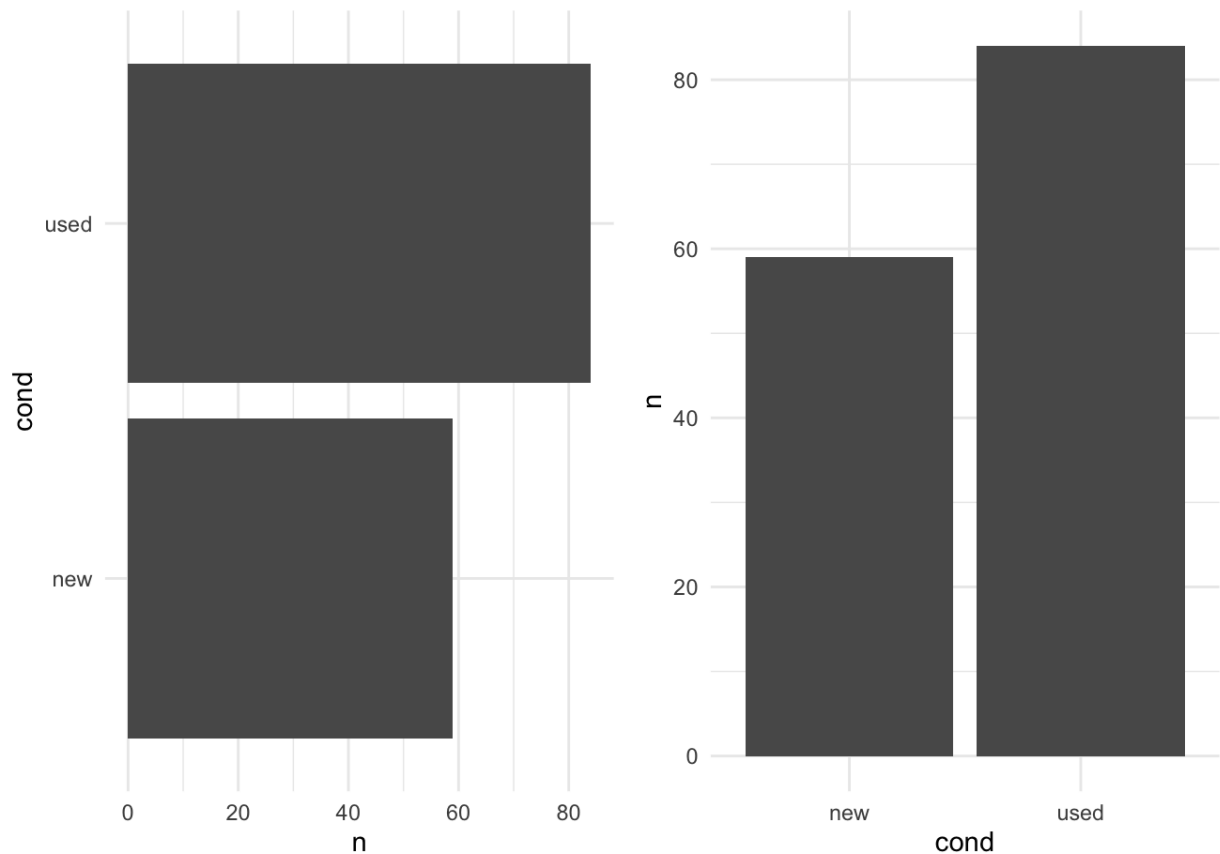


Abbildung 4.4: Häufigkeitsverteilung der Variable `cond`

Es gibt viele Methoden, sich mit R ein Balkendiagramm ausgeben zu lassen. Eine einfache, komfortable ist die mit dem Paket `DataExplorer`, s. [Abbildung 4.5](#).

Zuerst importieren wir die Daten und starten das R-Paket `DataExplorer`, s. [Listing 4.1](#).

Listing 4.1: Mariokart-Daten importieren von einer Webseite

```
mariokart <- read.csv("https://vincentarelbundock.github.io/Rdatasets/csv/openintro/mariokart")
```

Außerdem nicht vergessen, das Paket `DataExplorer` zu starten, s. [Listing 4.2](#).<sup>3</sup> In diesem Paket “wohnen” die Befehle, die wir zum Erstellen der Datendiagramme nutzen werden.

Listing 4.2: Wir starten das R-Paket `DataExplorer`

```
library(DataExplorer)
```

Listing 4.3: Syntax zur Erstellung eines Histogramms

```
mariokart %>%
  select(total_pr) %>%
  filter(total_pr < 100) %>% # ohne Extremwerte
  plot_histogram()
```

Listing 4.4: Syntax zur Erstellung eines Balkendiagramms

```
mariokart %>%
  select(cond) %>%
```

```
plot_bar()
```

Die Syntax ist in [Listing 4.4](#) abgedruckt. Übersetzen wir die Syntax ins Deutsche:

Nimm den Datensatz `mariokart` \*und dann\*  
wähle die Spalte `cond` \*und dann\*  
zeichne ein Balkendiagramm.

**Übungsaufgabe 4.3 (Spalten wählen für das Balkendiagramm)** Hätten wir andere Spalten ausgewählt, so würde das Balkendiagramm die Verteilung jener Variablen zeigen. Ja, Sie können auch mehrere Variablen auf einmal auswählen. Probieren Sie das doch mal aus!

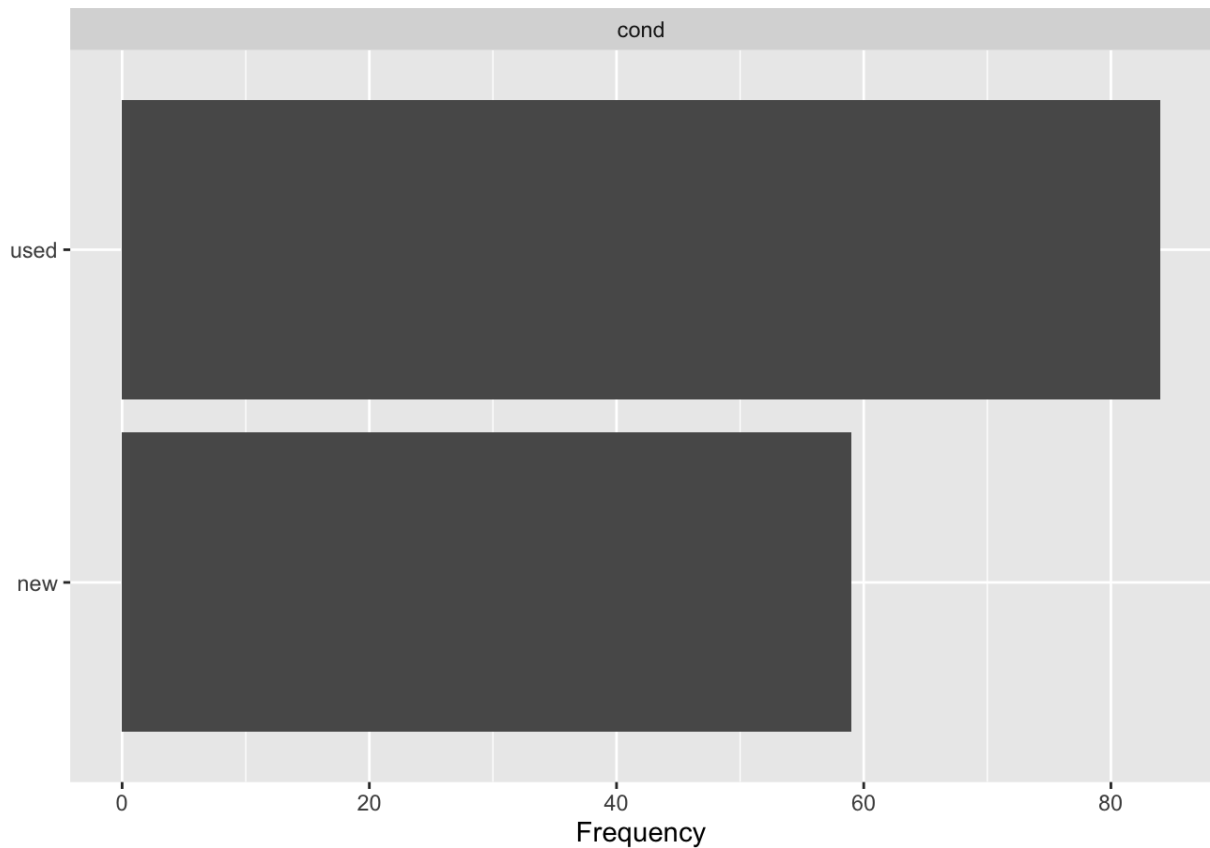


Abbildung 4.5: Balkendiagramm mit dem R-Paket `DataExplorer`

## 4.4.2 Verteilung: quantitative Variable

### 4.4.2.1 Histogramm

Bei einer quantitativen Variablen mit vielen Ausprägungen wäre ein Balkendiagramm nicht so aussagekräftig, s. [Abbildung 4.6](#). Es gibt einfach zu viele Ausprägungen.

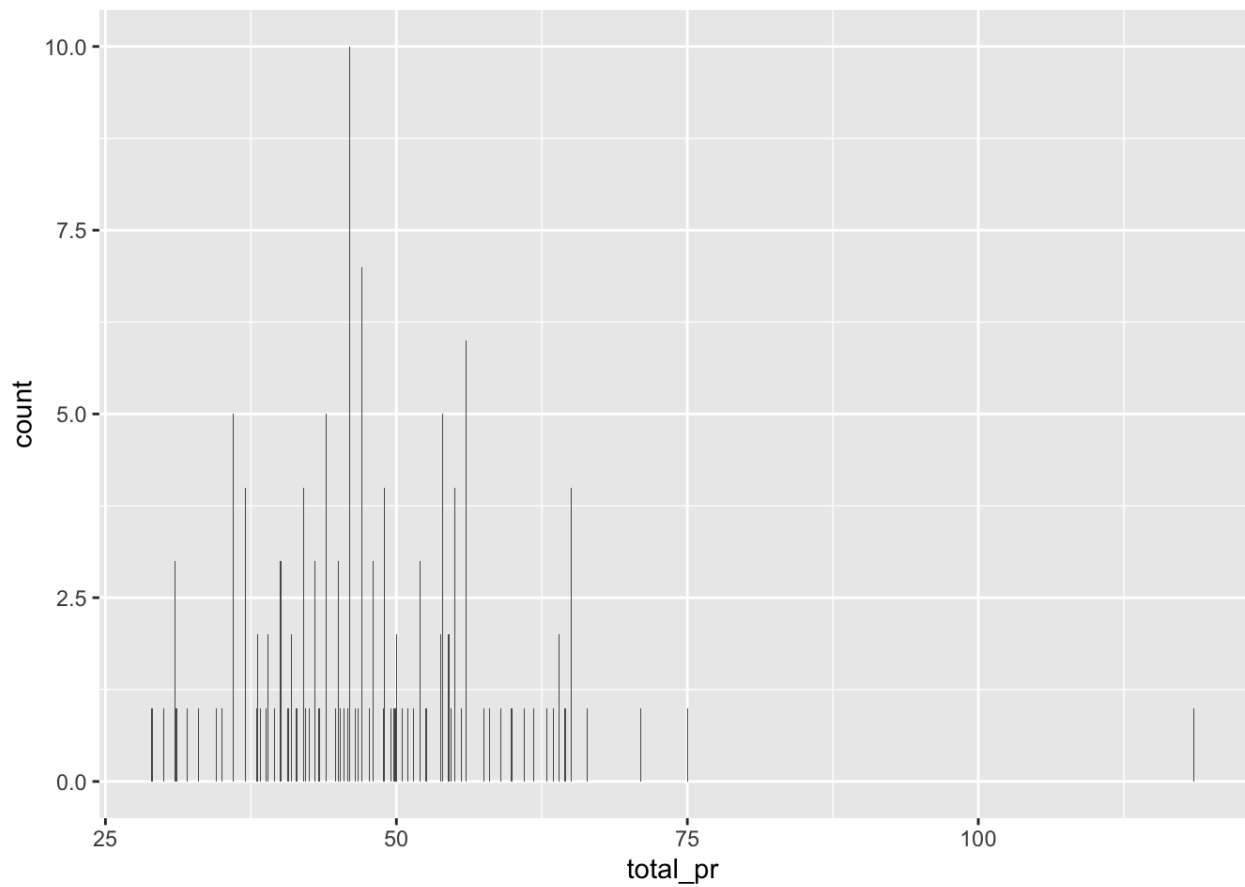


Abbildung 4.6: Balkendiagramm für `total_pr`

Die Lösung: Wir reduzieren die Anzahl der Ausprägungen, indem wir auf ganze Dollar runden. Oder, um noch weniger Ausprägungen zu bekommen, können wir einfach Gruppen definieren, z.B.

- Gruppe 1: 0-5 Dollar
- Gruppe 2: 6-10 Dollar
- Gruppe 2: 11-15 Dollar ...

In [Abbildung 4.7](#) sind z.B. die Ausprägungen des Verkaufspreis (`total_pr`) in in Gruppen der Breite von 5 Dollar aufgeteilt worden. Zusätzlich sind noch die einzelnen Werte als schwarze Punkte gezeigt.



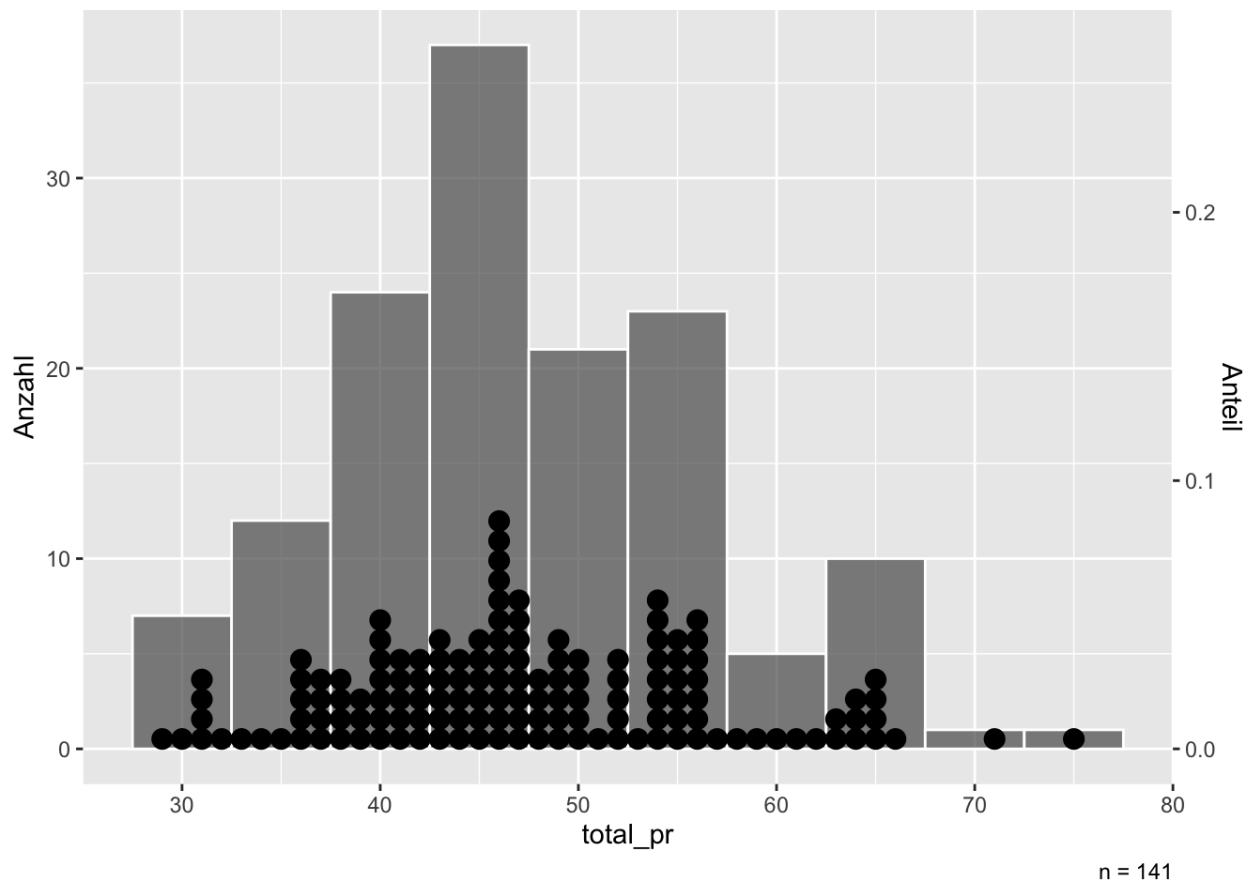
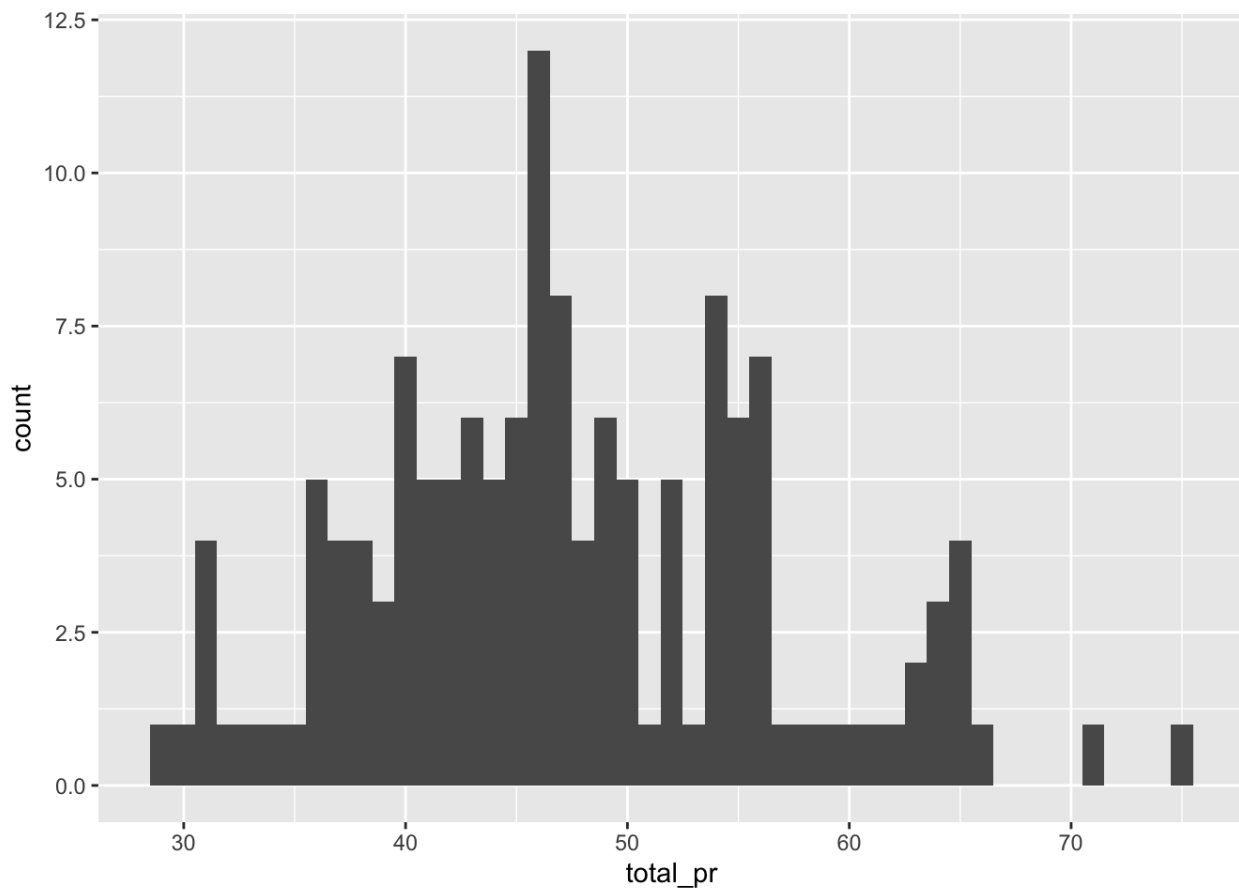


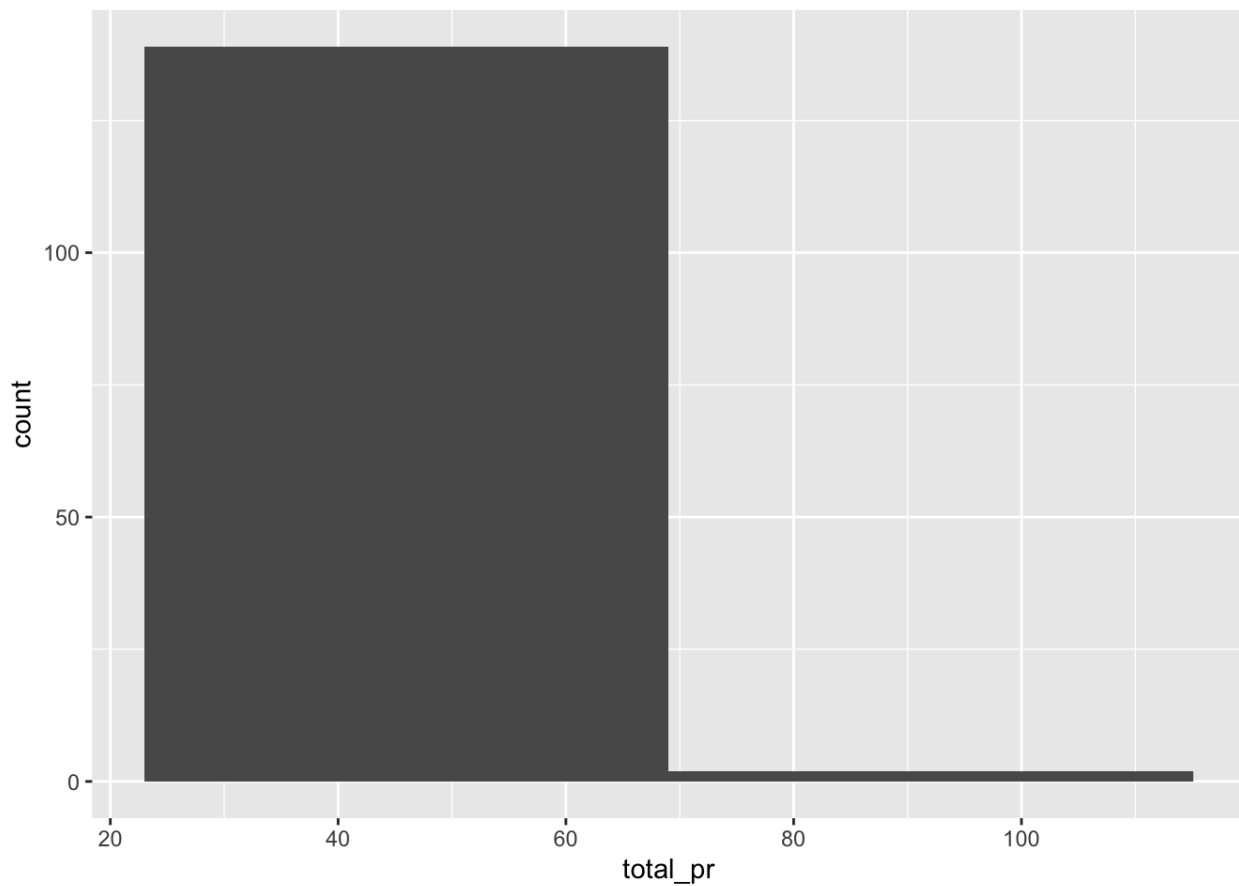
Abbildung 4.7: Balkendiagramm für `total_pr`

**Definition 4.3 (Histogramm)** Ein Histogramm ist ein Diagramm zur Darstellung der Häufigkeitsverteilung einer quantitativen Variablen. Die Daten werden in Gruppen (Klassen) eingeteilt, die dann durch einen Balken dargestellt sind. Die Höhe der Balken zeigt die Häufigkeit der Daten in dieser Gruppe/in diesem Balken<sup>4</sup>.

Es gibt keine klare Regel, wie viele Balken in einem Histogramm stehen sollten. Nur: Es sollten nicht sehr viele und nicht sehr wenig sein, s. [Abbildung 4.8](#) links bzw. [Abbildung 4.8](#), rechts.



(a) Zu viele Gruppen (Balken)



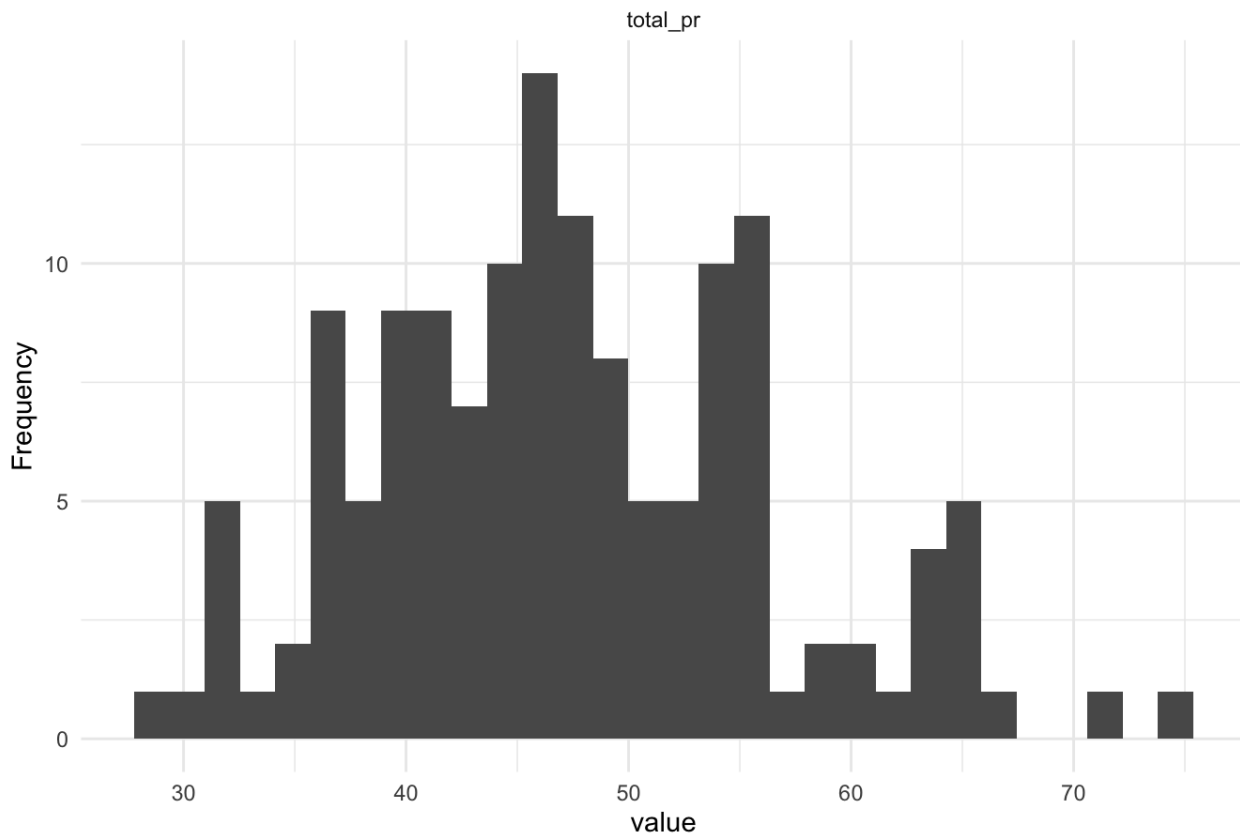
(b) Zu wenige Gruppen (Balken)

Abbildung 4.8: Nicht zu wenig und nicht zu viele Balken im Balkendiagramm

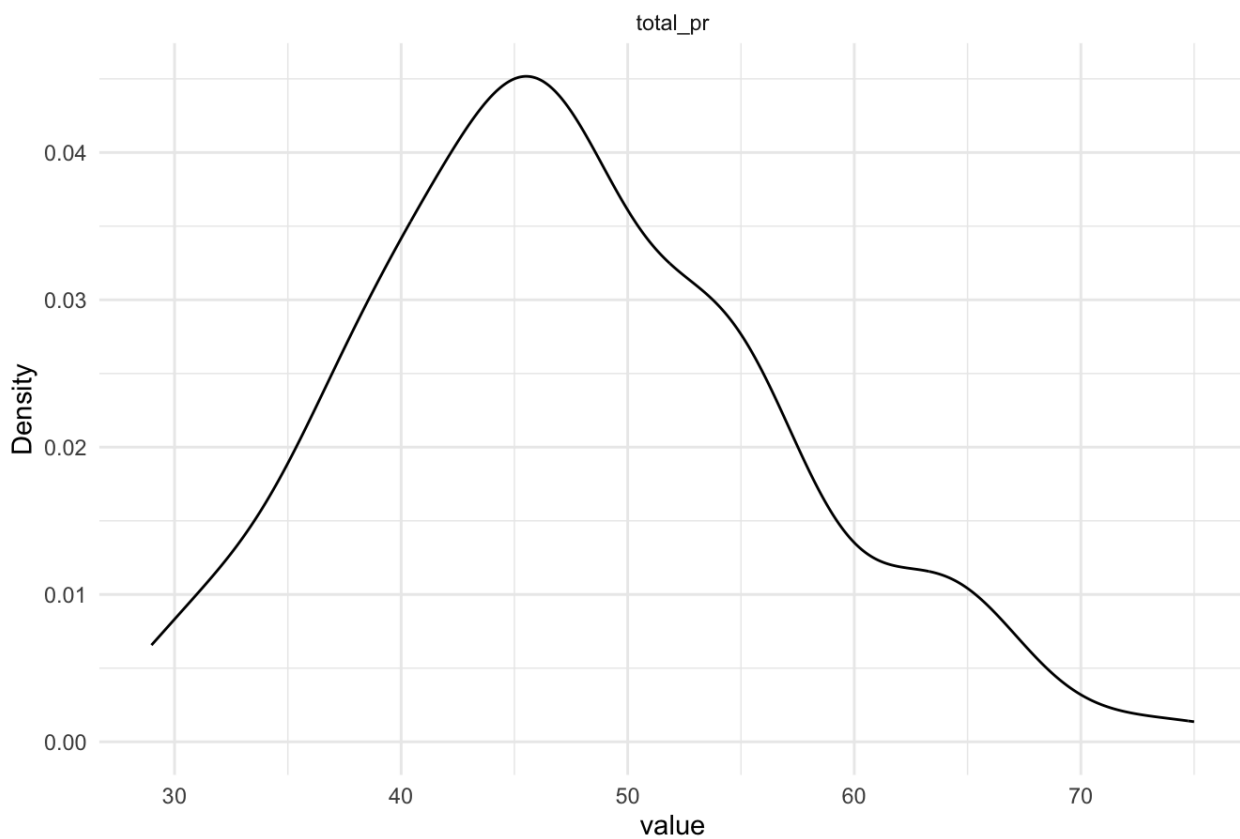
Zur Erstellung eines Histogramms können Sie die Syntax [Listing 4.5](#) nützen, vgl. [Abbildung 4.9](#), links.

Listing 4.5: Syntax zur Erstellung eines Histogramms

```
mariokart %>%  
  select(total_pr) %>%  
  filter(total_pr < 100) %>% # ohne Extremwerte  
  plot_histogram()
```



(a) Histogramm



(b) Dichtediagramm

Abbildung 4.9: Eine stetige Verteilung verbildlichen

### 4.4.2.2 Dichtediagramm

[Abbildung 4.10](#) fügt zu [Abbildung 4.7](#) ein *Dichtediagramm* hinzu (rote Linie). Ein Dichtediagramm ähnelt einem “glattgeschmirgeltem” Histogramm.

**Definition 4.4 (Dichtediagramm)** Ein Dichtediagramm visualisiert die Verteilung einer stetigen Variablen. Im Gegensatz zum Histogramm wird der Verlauf der Kurve geglättet, so kann Rauschen besser ausgeblendet werden.<sup>5</sup>

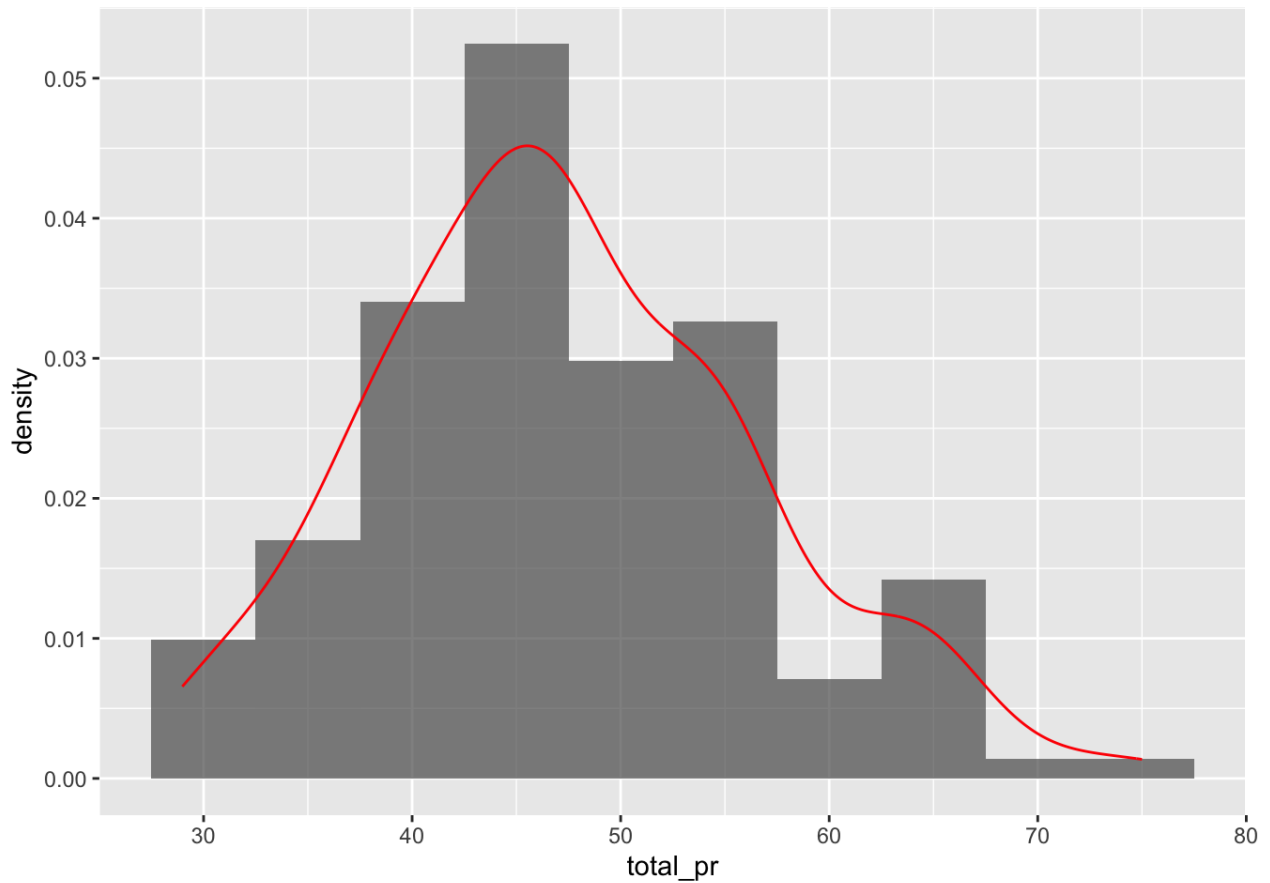
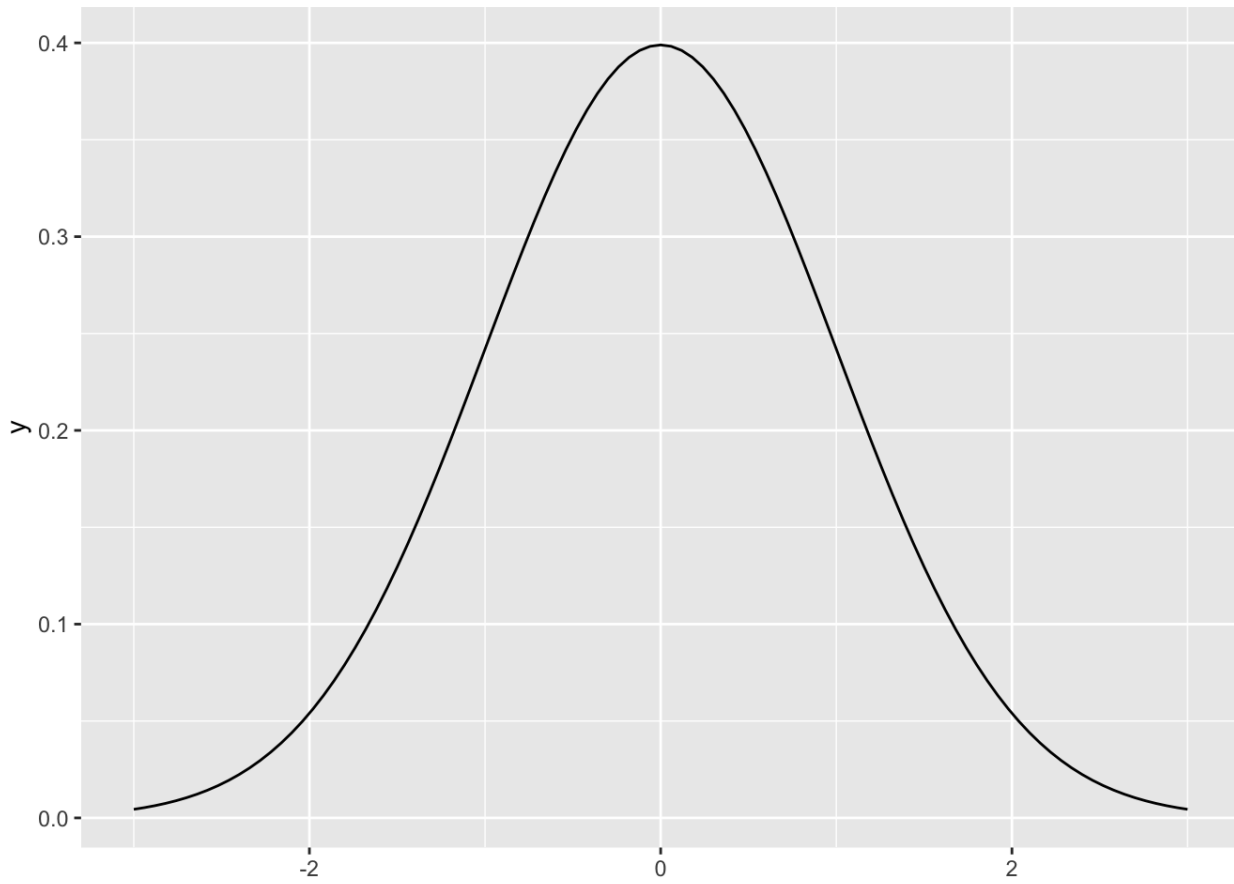


Abbildung 4.10: Balkendiagramm für `total_pr`

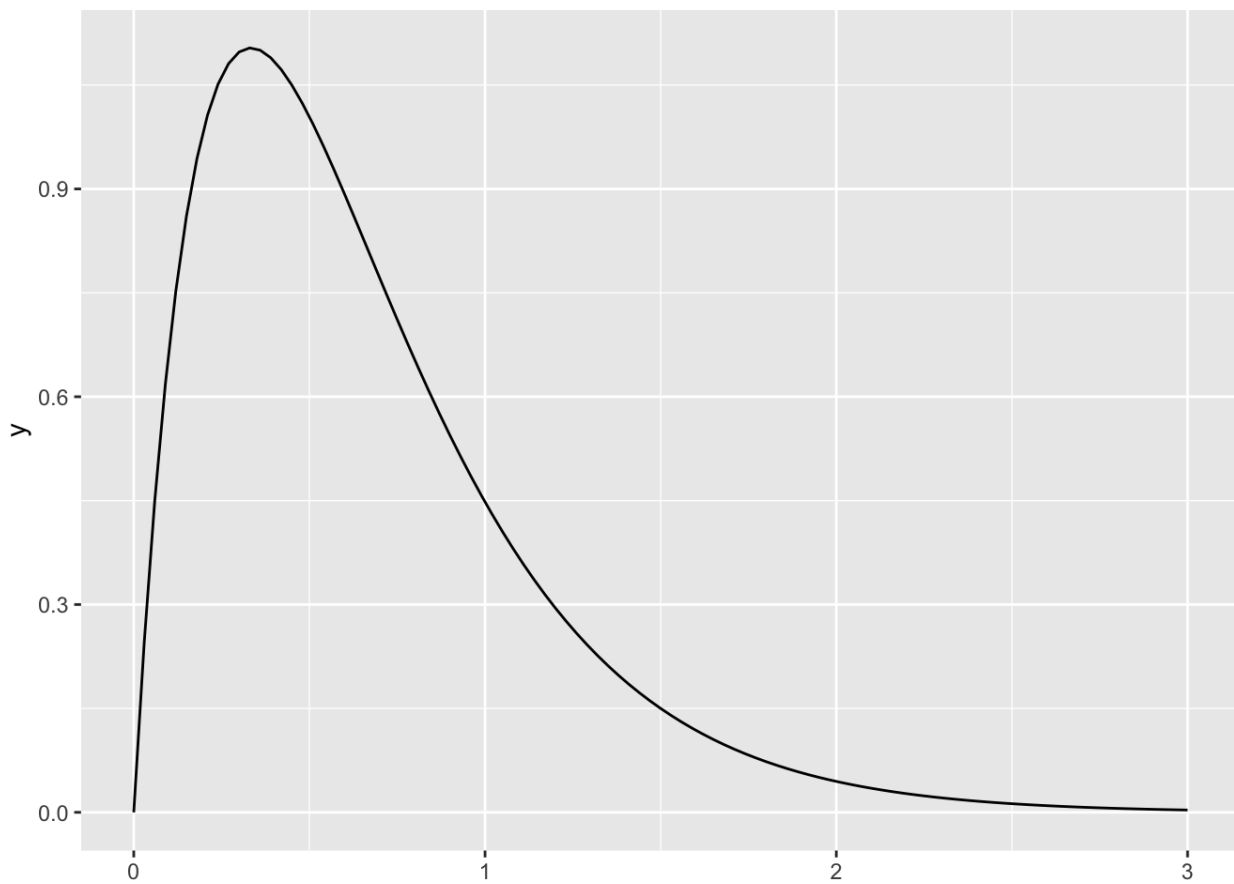
**Übungsaufgabe 4.4** Erstellen Sie das Diagramm [Abbildung 4.9](#), rechtes Teildiagramm!<sup>6</sup>□

### 4.4.2.3 Eigenschaften von Verteilungen

(Diagramme von) Verteilungen können symmetrisch oder schief (nicht symmetrisch) sein, s. [Abbildung 4.11](#).



(a) Symmetrisch (Normal)



(b) Schief

Abbildung 4.11: Symmetrische vs. schiefe Verteilung, verbildlicht

[Abbildung 4.12](#) zeigt verschiedene Formen von Verteilungen.

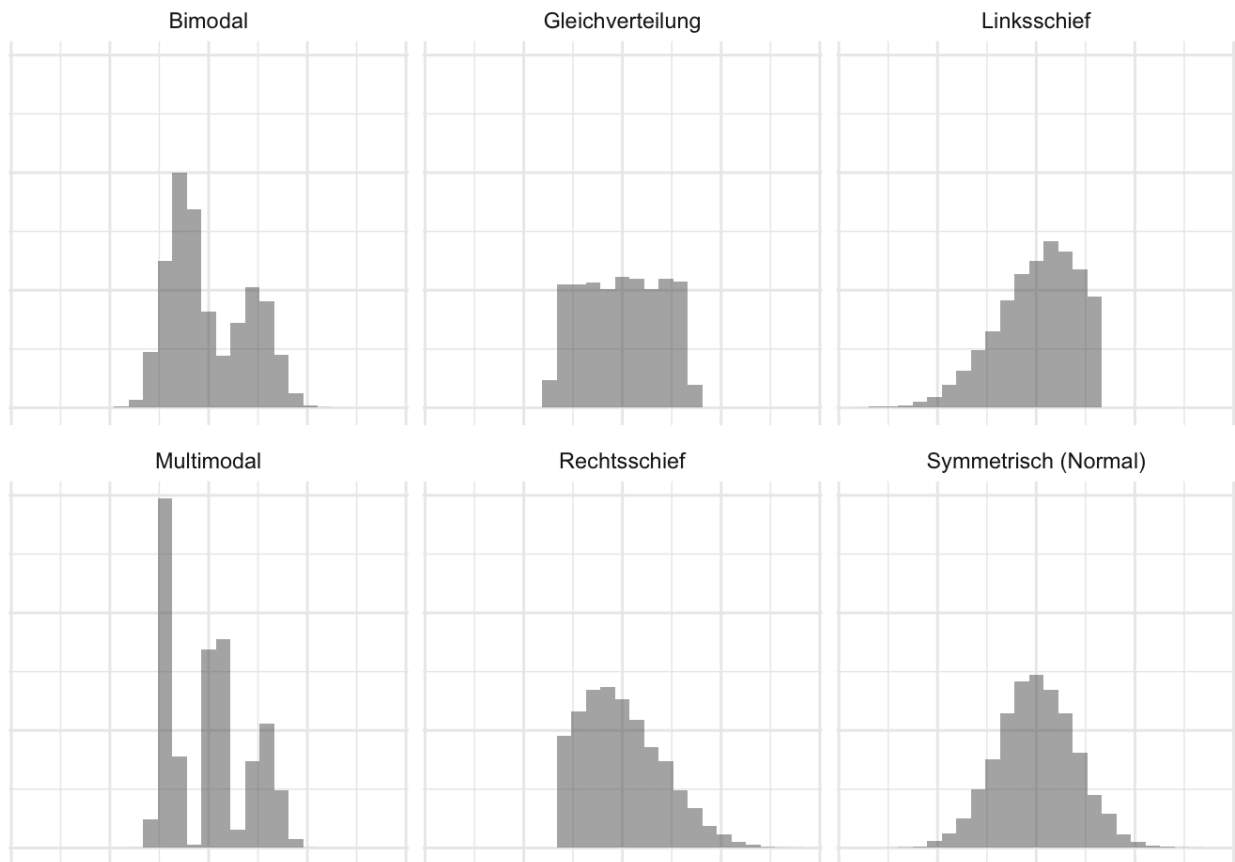


Abbildung 4.12: Verschiedene Verteilungsformen

[Quelle: ifes/FOM Hochschule](#)

### 4.4.3 Normalverteilung

Eine Normalverteilung sehen Sie in [Abbildung 4.11](#), links. Sie hat u.a. folgende Eigenschaften:

- symmetrisch
- glockenförmig
- stetig
- eingipflig (unimodal)
- Mittelwert, Median und Modus sind identisch

**Beispiel 4.3** Beispiele für normalverteilte Variablen sind Körpergröße von Männern oder Frauen, IQ-Werte, Prüfungsergebnisse, Messfehler, Lebensdauer von Glühbirnen, Gewichte von Brotlaiben, Milchproduktion von Kühen, Brustumfang schottischer Soldaten ([Lyon 2014](#)).□

Die Normalverteilung ist von hoher Bedeutung, da diese Verteilung unter (recht häufigen) Bedingungen zwangsläufig ergeben muss. Wenn sich eine Variable als Summe mehrerer, unabhängiger, etwa gleich starker Summanden, dann kann man erwarten, dass sich diese Variable normalverteilt. Dieses Phänomen kann man gut anhand des [Galton-Bretts](#) veranschaulichen.

## Galtonboard / Galtonbrett Simulation (or Bean machine or quincunx or Galton b...



### 4.5 Zusammenhänge verbildlichen

---

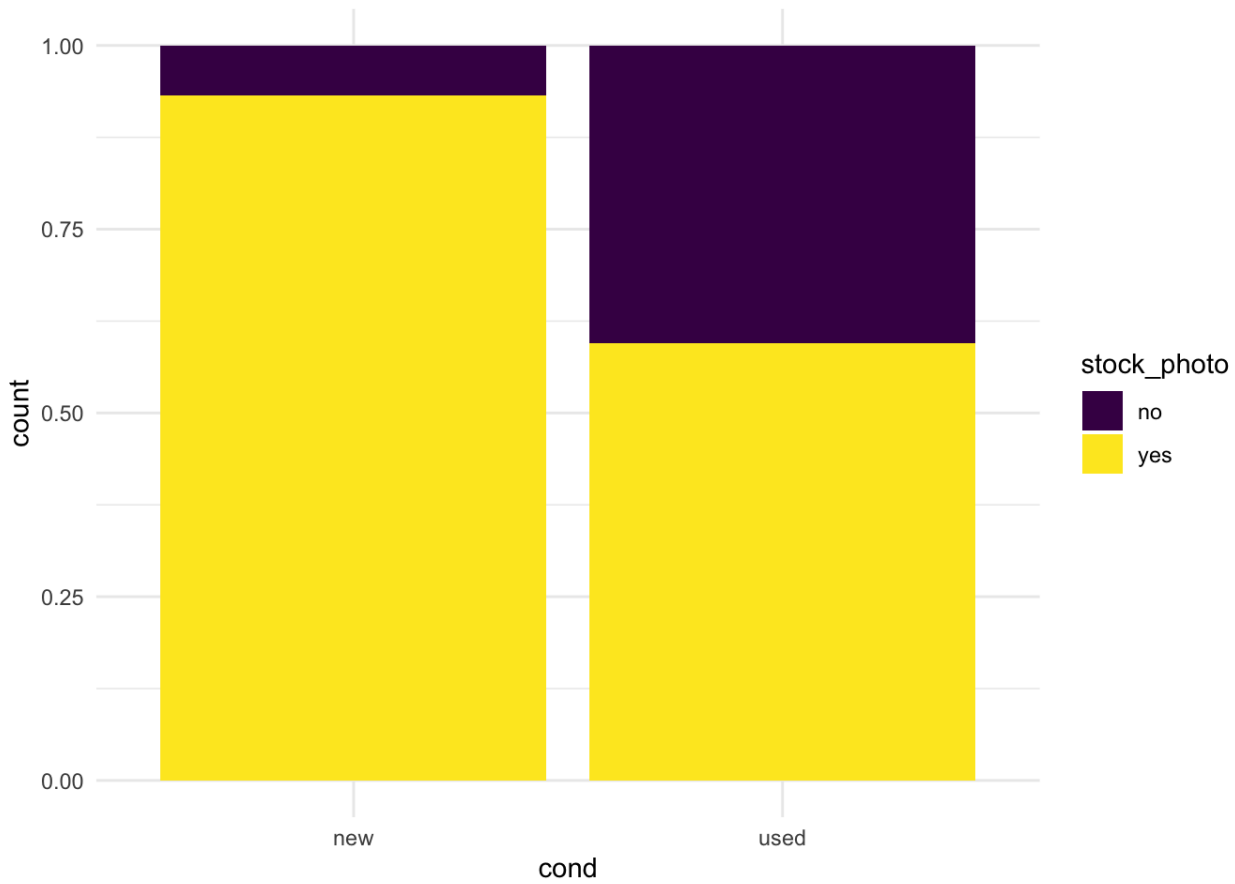
#### 4.5.1 Zusammenhang: nominale Variablen

##### **Beispiel 4.4 (Beispiele für Zusammenhänge bei nominalen Variablen)**

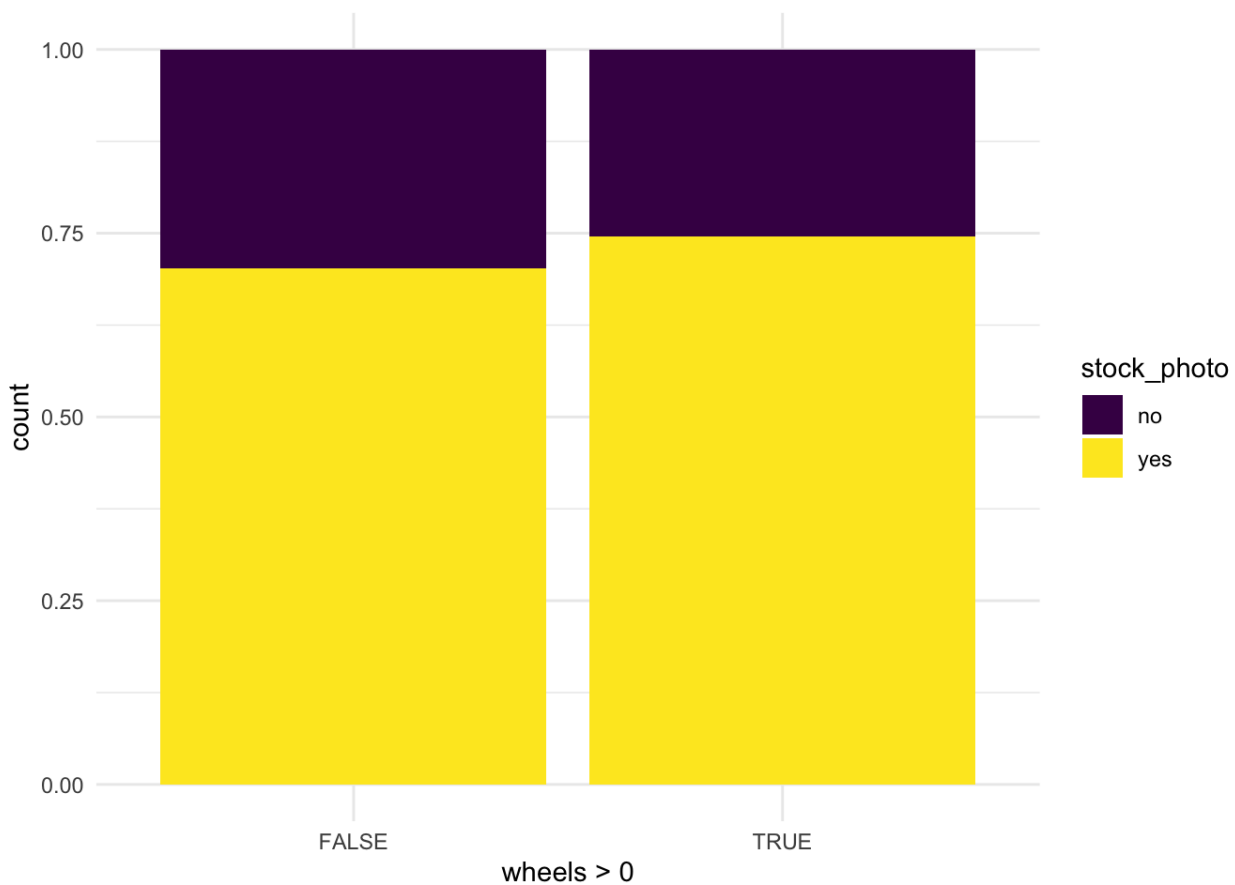
- Hängt Berufserfolg (Führungskraft ja/nein) mit dem Geschlecht zusammen?
- Hängt der Beruf des Vaters mit dem Schulabschluss des Kindes (Abitur, Realschule, Mittelschule) zusammen?
- Gibt es einen Zusammenhang zwischen Automarke und politische Präferenz einer Partei? ☐

Sagen wir, Sie arbeiten immer noch beim Online-Auktionshaus und Sie fragen sich, ob ein Produktfoto wohl primär bei neuwertigen Produkten beiliegt, aber nicht bei gebrauchten? Dazu betrachten Sie wieder die [mariokart](#)-Daten, s. [Abbildung 4.13](#).





(a) Es findet sich ein Zusammenhang von Foto und Zustand in den Daten



(b) Es findet sich (fast) kein Zusammenhang von `wheel` und Foto in den Daten

Abbildung 4.13: Zusammenhang zwischen nominalskalierten Variablen verbildlichen

Tatsächlich: Es findet sich ein Zusammenhang zwischen der Tatsache, ob dem versteigerten Produkt ein Foto bei lag und ob es neuwertig oder gebraucht war ([Abbildung 4.13](#), links). Bei neuen Spielen war fast immer (ca. 90%) ein Foto dabei. Bei gebrauchten Spiel immerhin bei gut der Hälfte der Fälle.

Anders sieht es aus für die Frage, ob ein (oder mehrere) Lenkräder dem Spiel beilagen (oder nicht) in Zusammenhang mit der Fotofrage. Hier gab es fast keinen Unterschied zwischen neuen und alten Spielen, was die Frage nach "Foto des Produkts dabei" betraf ([Abbildung 4.13](#), rechts), der Anteil betrug jeweils ca. 70%.

Anders gesagt: Unterscheiden sich die "Füllhöhe" in den Diagrammen, so gibt es einen Unterschied hinsichtlich "Foto ist dabei" zwischen den beiden Gruppen (linker vs. rechter Balken). Unterscheiden sich die Anteile in den Gruppen (neuwertige vs. gebrauchte Spiele), so spielt z.B. die Variable "Foto dabei" offenbar eine Rolle. Dann hängen Neuwertigkeit und "Foto dabei" also zusammen!

So können Sie sich in R ein gefülltes Balkendiagramm ausgeben lassen, s. [Abbildung 4.14](#).

```
mariokart %>%
  select(cond, stock_photo) %>%
  plot_bar(by = "cond")
```

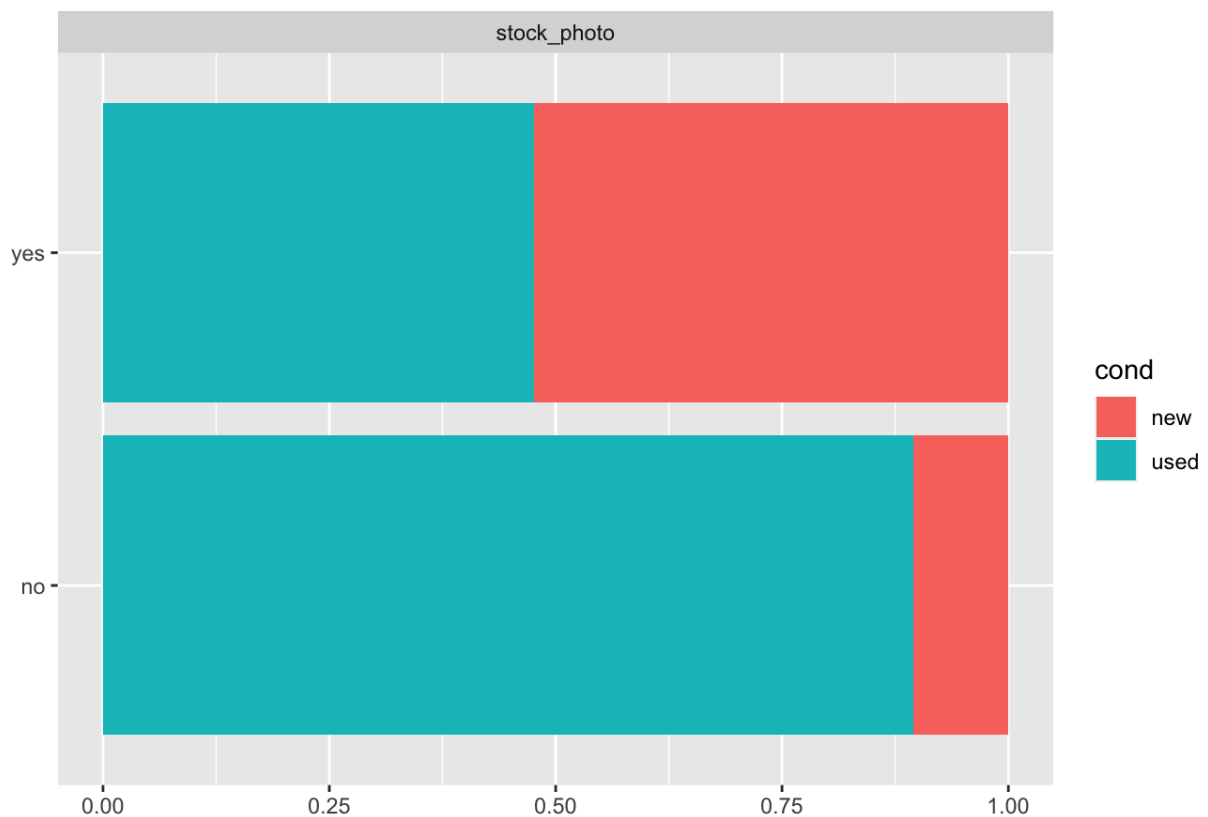


Abbildung 4.14: Ein gefülltes Balkendiagramm zur Untersuchung eines Zusammenhangs zwischen nominalskalierten Variablen

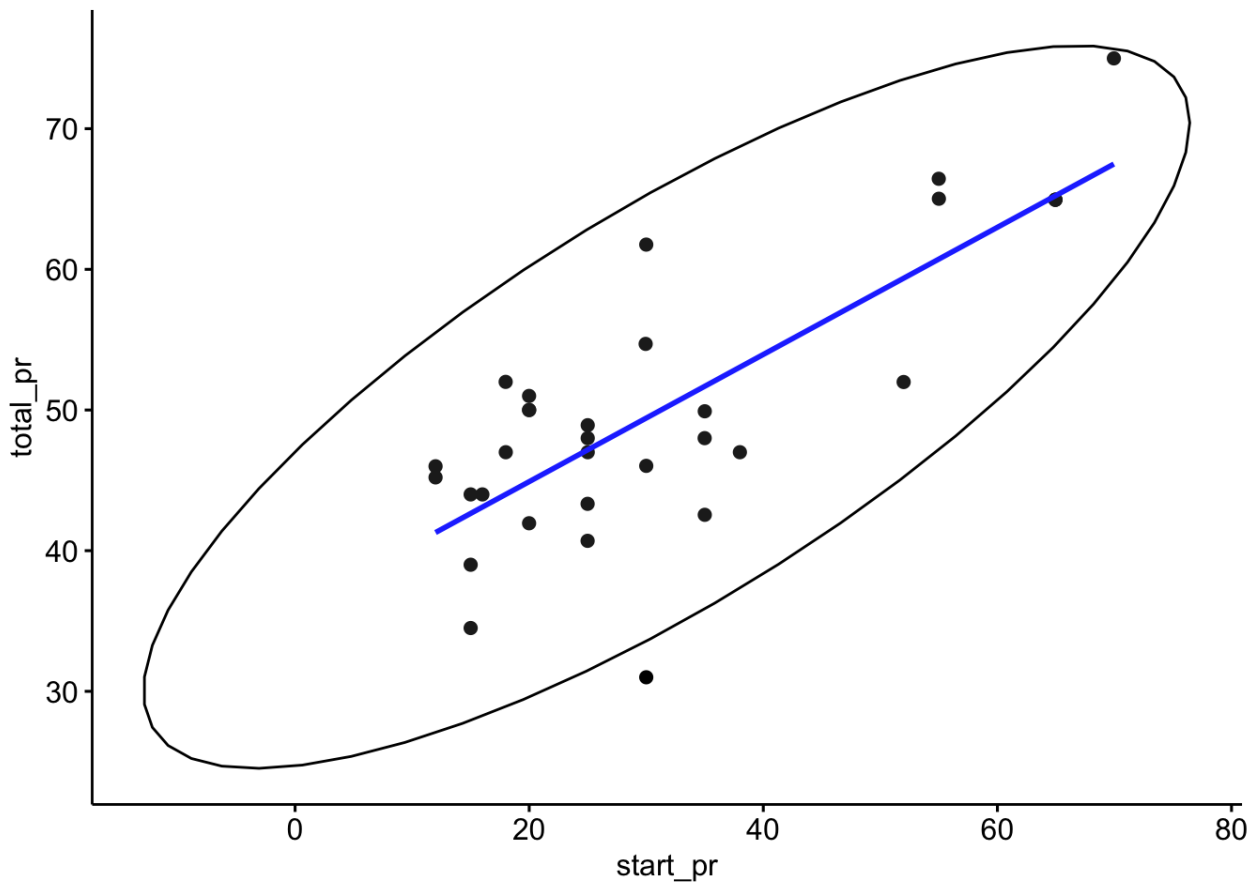
#### **Hinweis**

Gefüllte Balkendiagramme eignen sich zur Analyse eines Zusammenhangs zwischen nominalskalierten Variablen. Allerdings sollte eine der beiden Variablen nur zwei Ausprägungen aufweisen, sonst sind die Zusammenhänge nicht mehr so gut zu erkennen. □

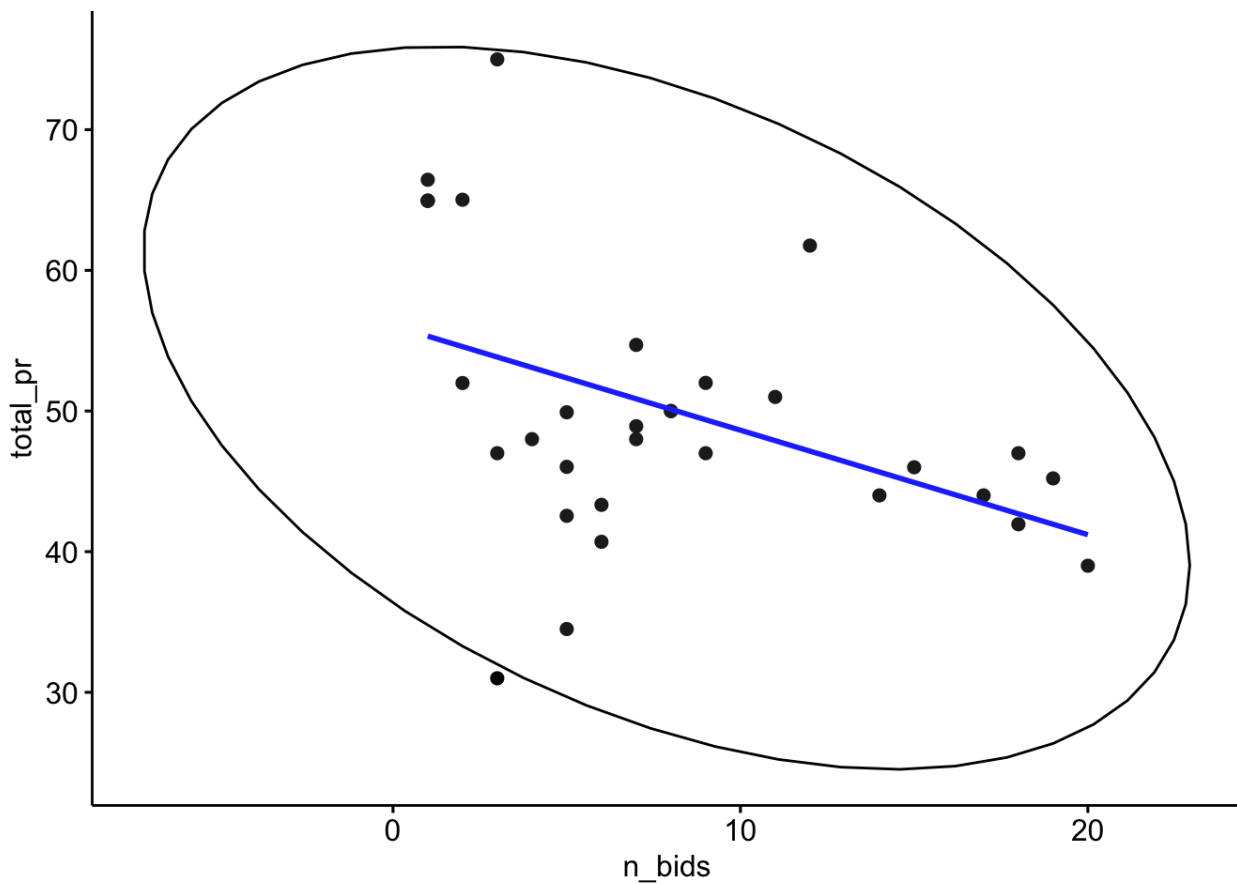
## 4.5.2 Zusammenhang: metrisch

Den (etwaigen) Zusammenhang zweier metrischer Variablen kann man mit einem *Streudiagramm* visualisieren (engl. scatterplot). [Abbildung 4.15](#) links untersucht den Zusammenhang des Einstiegspreises (X-Achse) und Abschlusspreises (Y-Achse) von Geboten bei Versteigerungen des Computerspiels Mariokart. In dem Diagramm ist eine Ellipse ergänzt, um die Art des Zusammenhangs besser zu verdeutlichen: Es handelt sich um einen *gleichsinnigen* (positiven) Zusammenhang: Je höher der Startpreis, desto *höher* der Abschlusspreis, zumindest tendenziell. Außerdem ist eine "Trendgerade" (Regressionsgerade) in blau eingezeichnet. Diese Gerade liegt "mittig" in den Daten (wir definieren dies später genauer). Diese Trendgerade gibt Aufschluss über "typische" Werte: Welcher Y-Wert ist "typisch" für einen bestimmten X-Wert?

[Abbildung 4.15](#) rechts untersucht den Zusammenhang zwischen Anzahl der Gebote (X-Achse) und Abschlusspreises (Y-Achse). Es handelt sich um einen negativen Zusammenhang: Je mehr Gebote, desto *geringer* der Abschlusspreis.



(a) positiver, mittelstarker Zusammenhang



(b) negativer, eher schwacher Zusammenhang

Abbildung 4.15: Streudiagramm zur Darstellung eines Zusammenhangs zweier metrischer Variablen

[Abbildung 4.16](#) bietet einen Überblick über verschiedene Beispiele von Richtung und Stärke von Zusammenhängen.

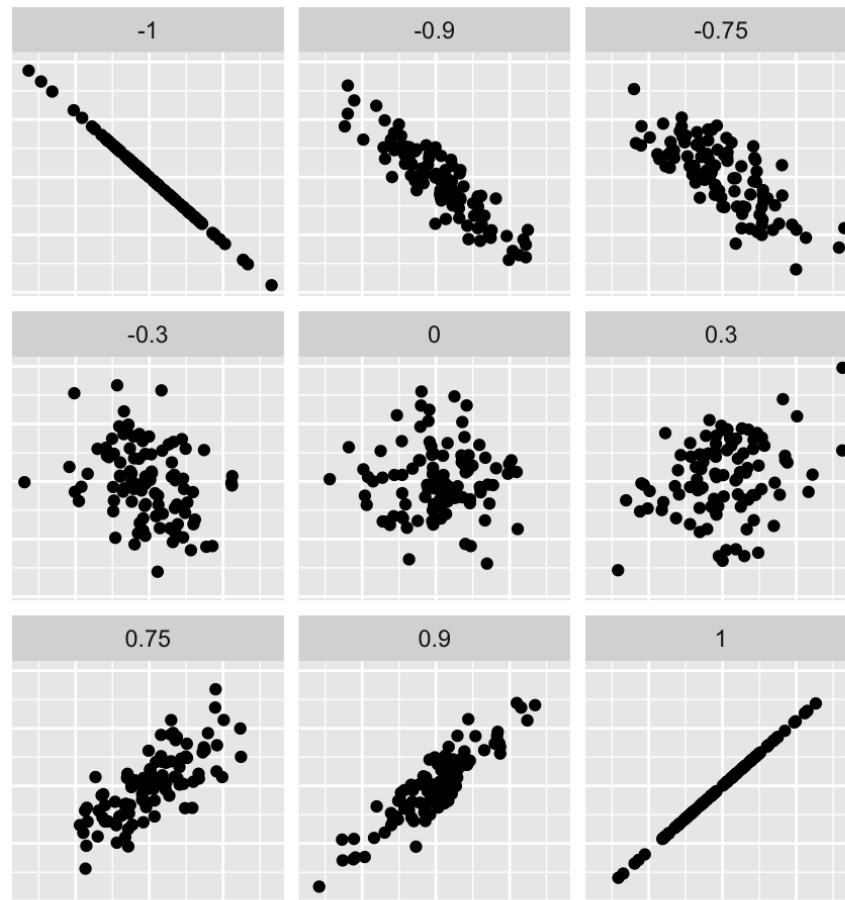


Abbildung 4.16: Überblick über verschiedene Beispiele von Streudiagrammen

Quelle: Aufbauend auf FOM/ifes, Norman Markgraf

In [Abbildung 4.16](#) ist für jedes Teildiagramm eine Zahl angegeben: der *Korrelationskoeffizient*. Diese Statistik quantifiziert Richtung und Stärke des Zusammenhangs (mehr dazu in Kap. XXX). Ein positives Vorzeichen steht für einen positiven Zusammenhang, ein negatives Vorzeichen für einen negativen Zusammenhang. Der (Absolut-)Wert gibt die Stärke des linearen Zusammenhangs an ([Cohen 1992](#)):

- $\pm 0$ : Kein Zusammenhang
- $\pm 0.1$ : schwacher Zusammenhang
- $\pm 0.3$ : mittlerer Zusammenhang
- $\pm 0.5$ : starker Zusammenhang
- $\pm 1$ : perfekter Zusammenhang

[Abbildung 4.17](#) hat die gleiche Aussage, ist aber plakativer, indem *Stärke* (schwach, stark) und *Richtung* (positiv, negativ) gegenübergestellt sind.

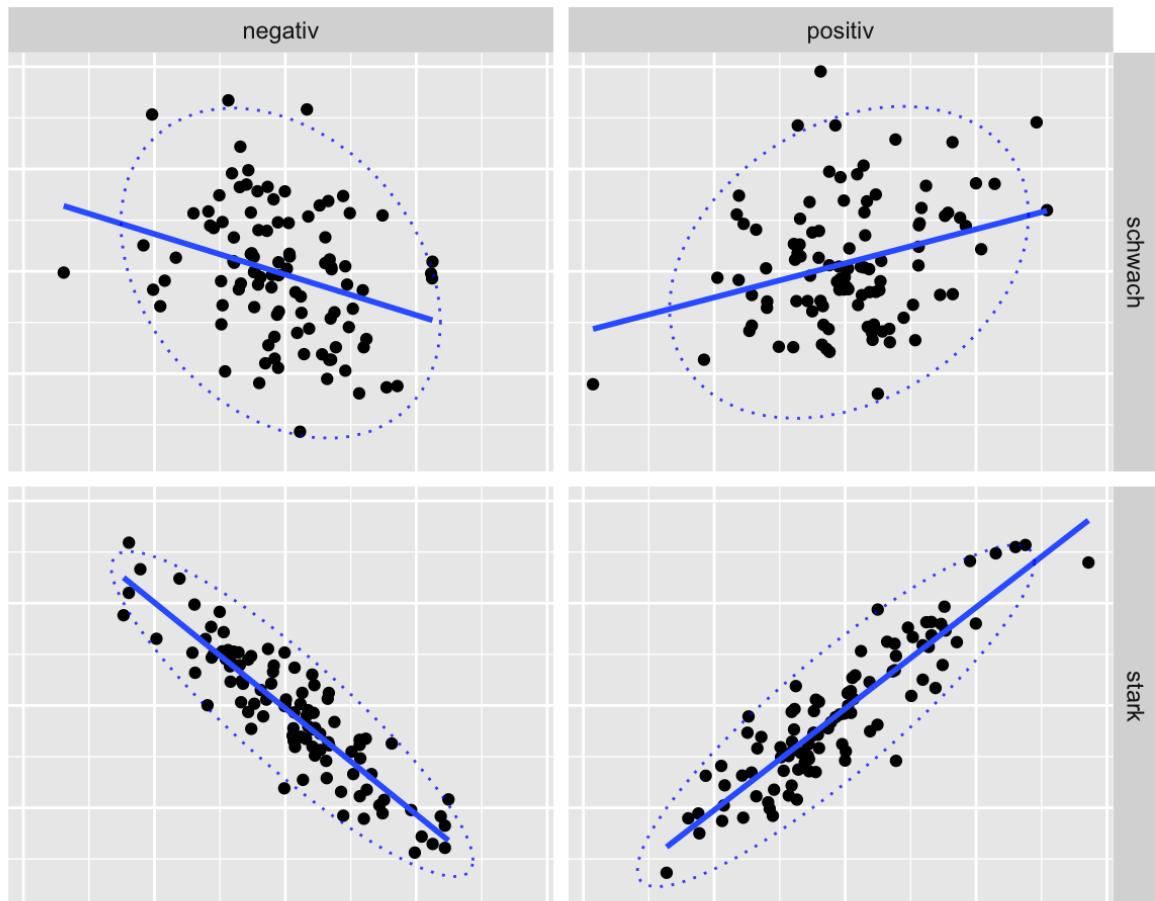


Abbildung 4.17: Überblick über starke vs. schwache bzw. positive vs. negative Zusammenhänge

Man sieht in [Abbildung 4.16](#) und [Abbildung 4.17](#), dass ein *negativer* Korrelationskoeffizient mit einer *absinkenden* Trendgerade (blaue Linie) einhergeht. Umgekehrt geht ein *positiver* Trend mit einer *ansteigenden* Trendgerade einher. Zweitens erkennt man, dass *starke* Zusammenhänge mit einer *schmalen* Ellipse einhergehen und *schwache* Zusammenhänge mit einer *breiten* Ellipse einhergehen.

**Definition 4.5 (Richtig und Stärke eines Zusammenhang)** Gleichsinnige (positive) Zusammenhänge erkennt man an aufsteigenden Trendgeraden; gegensinnigen (negative) Zusammenhänge an absteigenden Trendgeraden:

- + :
- - :

Starke Zusammenhänge erkennt man an schmalen Ellipsen ("Baguette"); schwache Zusammenhänge an breiten Ellipsen ("Torte"):

- schwach:
- stark:



#### Vorsicht

Ein Zusammenhang der Art "je mehr X, desto mehr Y" ist *linear*. Lineare Zusammenhänge erkennt man im Diagramm an einer Geraden bzw. inwieweit sich die Punkte an einer Geraden "anschmiegen". Natürlich könnte man auch nicht-lineare Zusammenhänge untersuchen, aber der Einfachheit halber begnügen wir uns mit linearen. □

**Beispiel 4.5** Sie arbeiten nach wie vor bei einem Online-Auktionshaus, und manchmal gehört Datenanalyse zu Ihren Aufgaben. Daher interessiert Sie, ob welche Variablen mit dem Abschlusspreis (`total_pr`) im Datensatz `mariokart`

zusammenhängen. Sie verbildlichen die Daten mit R, und zwar nutzen Sie das Paket `DataExplorer`. Starten Sie dieses Paket, s. [Listing 4.2](#). Außerdem müssen wir noch die Daten importieren, falls noch nicht getan, s. [Listing 4.1](#).

So, jetzt kann die eigentliche Arbeit losgehen. Da Sie sich nur auf metrische Variablen konzentrieren wollen, wählen Sie (mit `select`) nur diese Variablen aus. Dann weisen Sie R an, einen Scatterplot zu malen (`plot_scatterplot`) und zwar jeweils den Zusammenhang einer der gewählten Variablen mit dem Abschlusspreis (`total_pr`), da das die Variable ist, die Sie primär interessiert. Das Ergebnis sieht man in [Abbildung 4.18](#).

```
mariokart %>%
  select(duration, n_bids, start_pr, ship_pr, total_pr, seller_rate, wheels) %>%
  plot_scatterplot(by = "total_pr")
```

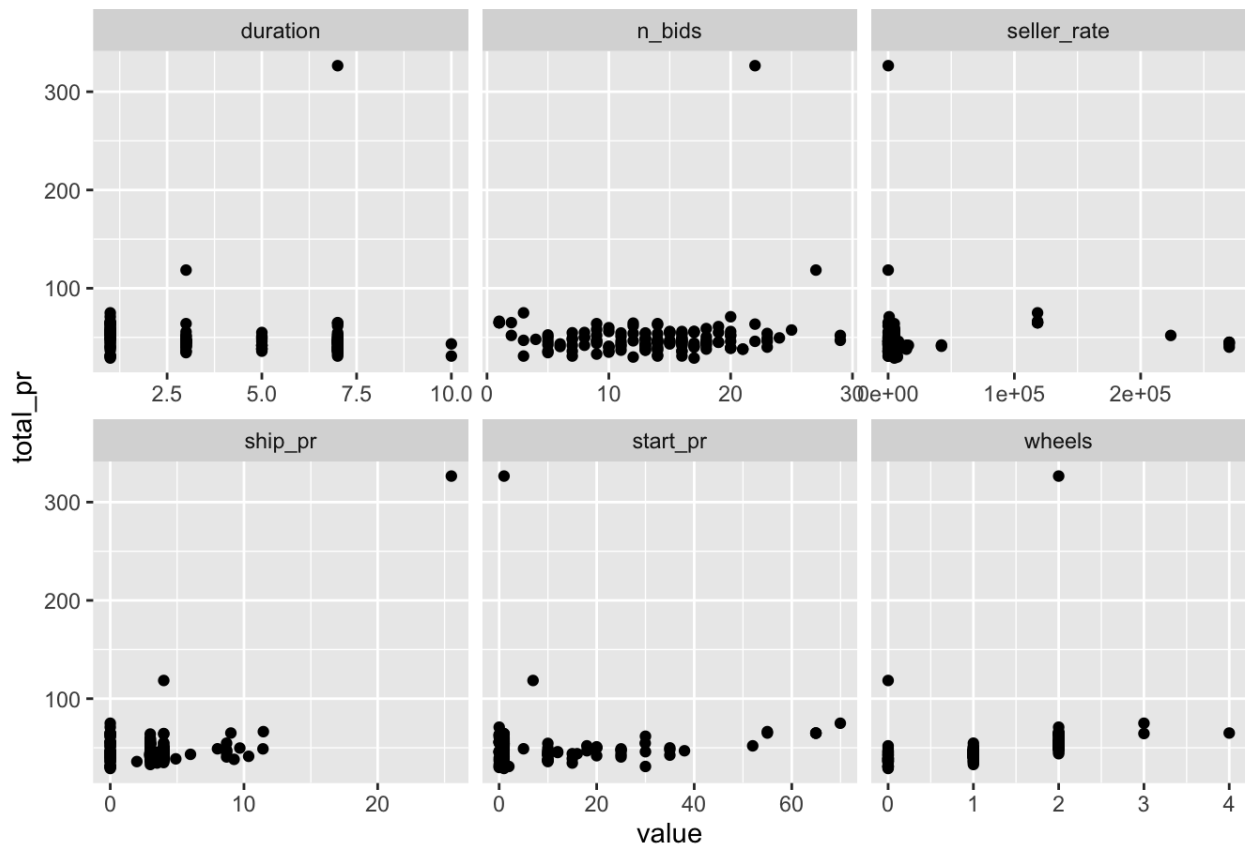


Abbildung 4.18: Der Zusammenhang metrischer Variablen mit Abschlusspreis

Aha... Was sagt uns das Bild? Hm. Es scheint einige Extremwerte zu geben, die dafür sorgen, dass der Rest der Daten recht zusammengequetscht auf dem Bild erscheint. Vielleicht sollten Sie solche Extremwerte lieber entfernen? Sie entscheiden sich, nur Verkäufe mit einem Abschlusspreis von weniger als 150 Dollar anzuschauen (`total_pr < 100`). Das Ergebnis ist in [Abbildung 4.19](#) zu sehen.

```
mariokart2 <-
  mariokart %>%
  filter(total_pr < 100)

mariokart2 %>%
  select(duration, n_bids, start_pr, ship_pr, total_pr, seller_rate, wheels) %>%
  plot_scatterplot(by = "total_pr")
```

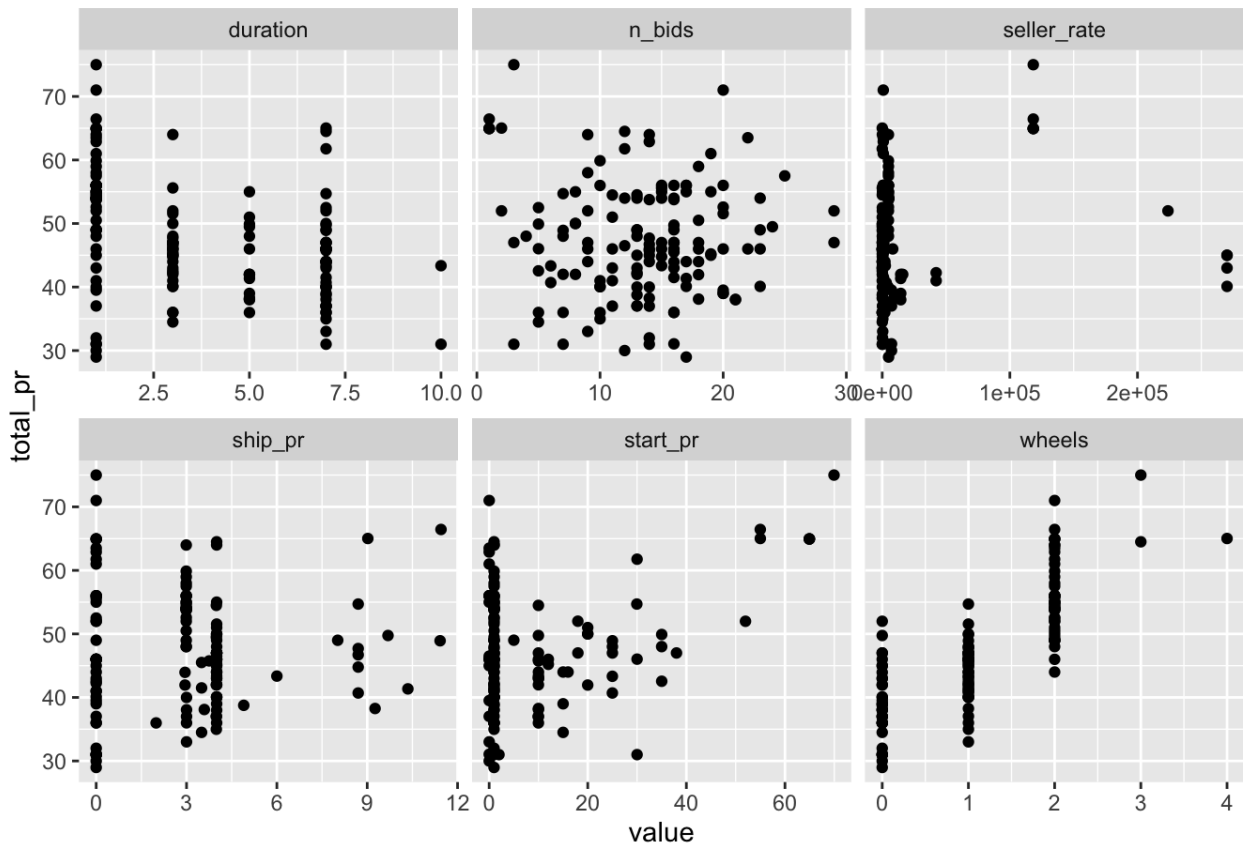


Abbildung 4.19: Der Zusammenhang metrischer Variablen mit Abschlusspreis

Ohne Extremwerte schält sich ein deutlicheres Bild ([Abbildung 4.19](#)) hervor: Startpreis (`start_pr`) und Anzahl der Räder (`wheels`) scheinen am stärksten mit dem Abschlusspreis zusammenzuhängen.

Das Argument `by = "total_pr"` bei `plot_scatterplot` weist R an, als Y-Variable stets `total_pr` zu verwenden. Alle übrigen Variablen kommen jeweils einmal als X-Variable vor.□

## 4.6 Unterschiede verbildlichen

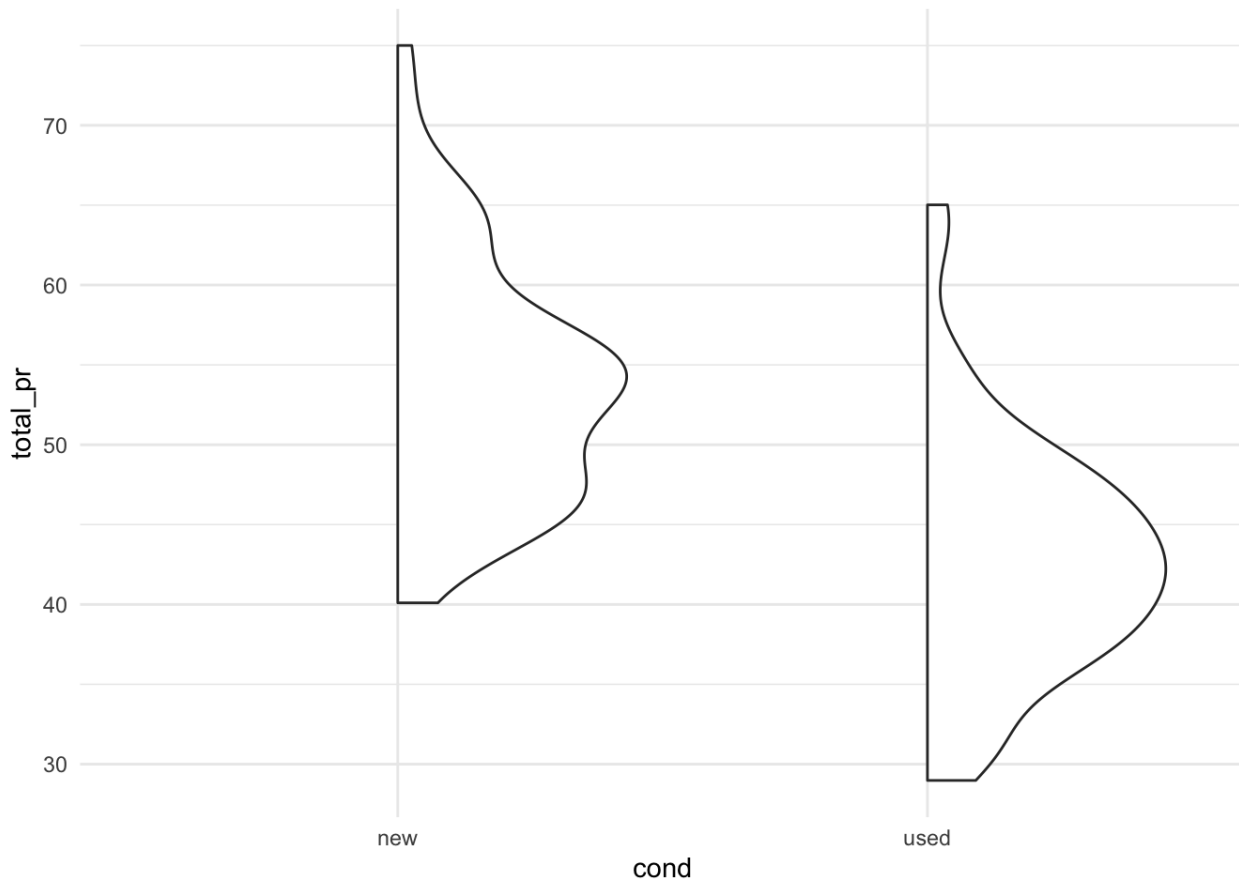
### 4.6.1 Unterschied: nominale Variablen

Gute Nachrichten: Für nominale Variablen bieten sich Balkendiagramme sowohl zur Darstellung von Zusammenhängen als auch von Unterschieden an. Genau genommen zeigt ja [Abbildung 4.13](#) (links) den *Unterschied* zwischen neuen und gebrauchten Spielen hinsichtlich der Frage, ob Photos beiliegen. Und wie man in [Abbildung 4.13](#) sieht, ist der Anteil der Spiele mit Foto bei den neuen Spielen höher als bei gebrauchten Spielen.

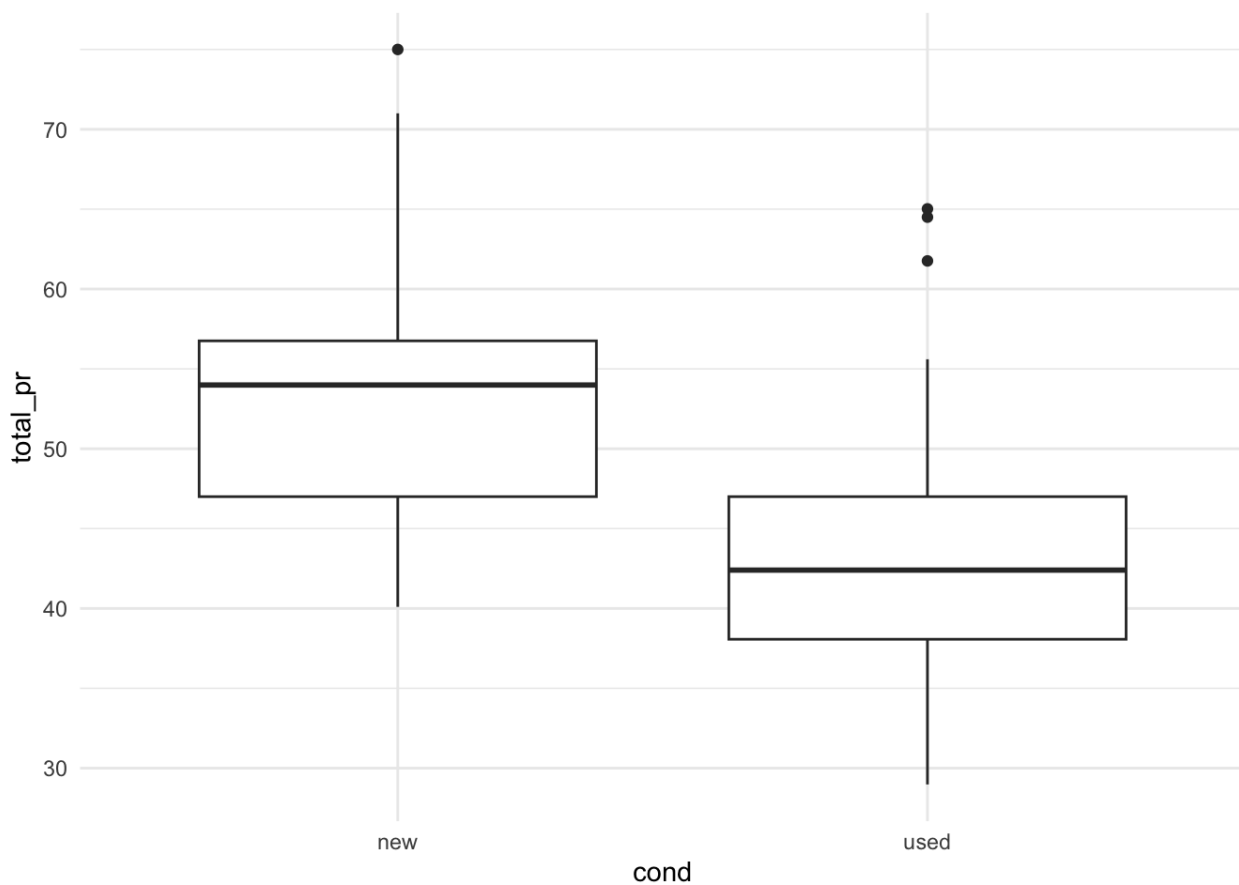
### 4.6.2 Unterschied: quantitative Variablen

Eine typische Analysefrage ist, ob sich zwei Gruppen hinsichtlich einer metrischen Zielvariablen deutlich unterscheiden. Genauer gesagt untersucht man z.B. oft, ob sich die Mittelwerte der beiden Gruppen zwischen der Zielvariablen deutlich unterscheiden. Das hört sich abstrakt an? Am besten wir schauen uns einige Beispiele an, s. [Abbildung 4.20](#).





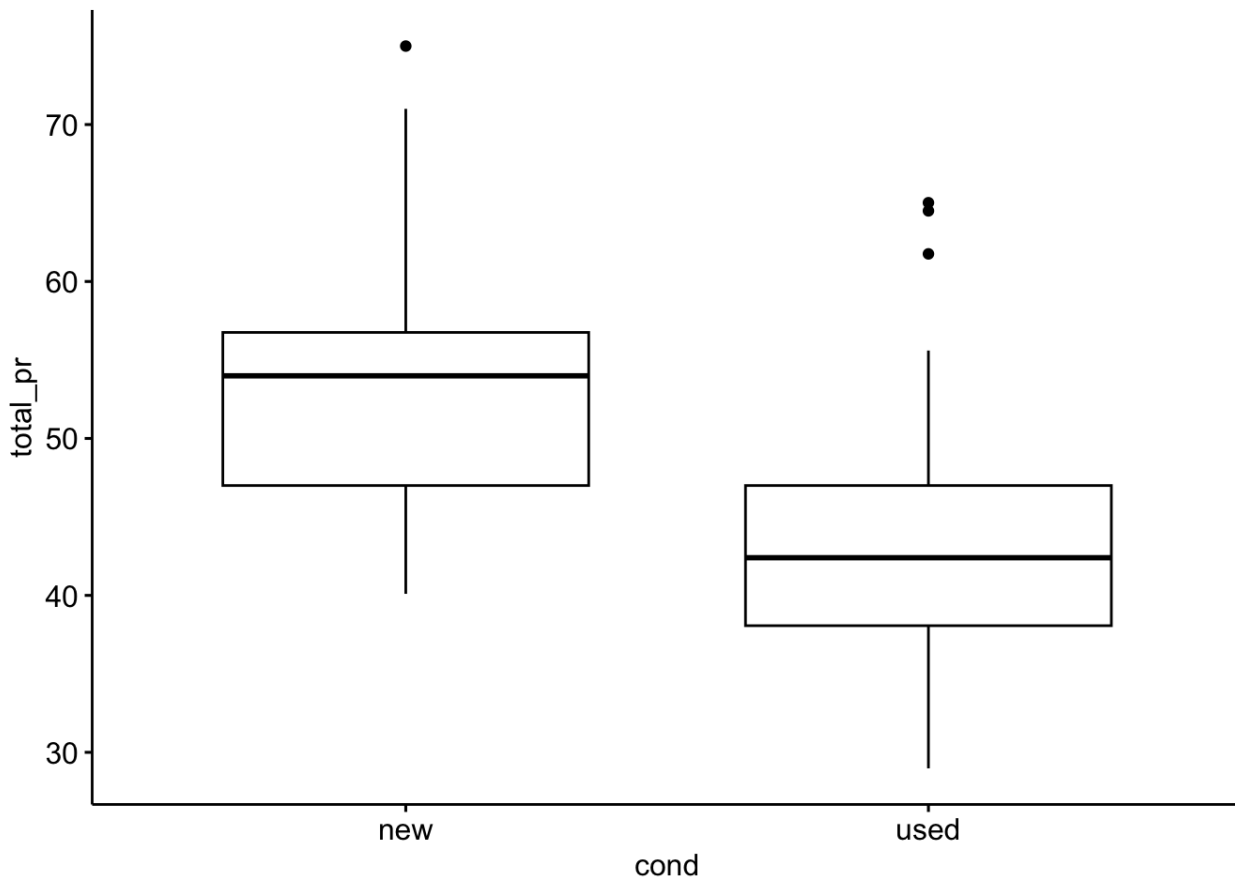
(a) Histogramm pro Gruppe



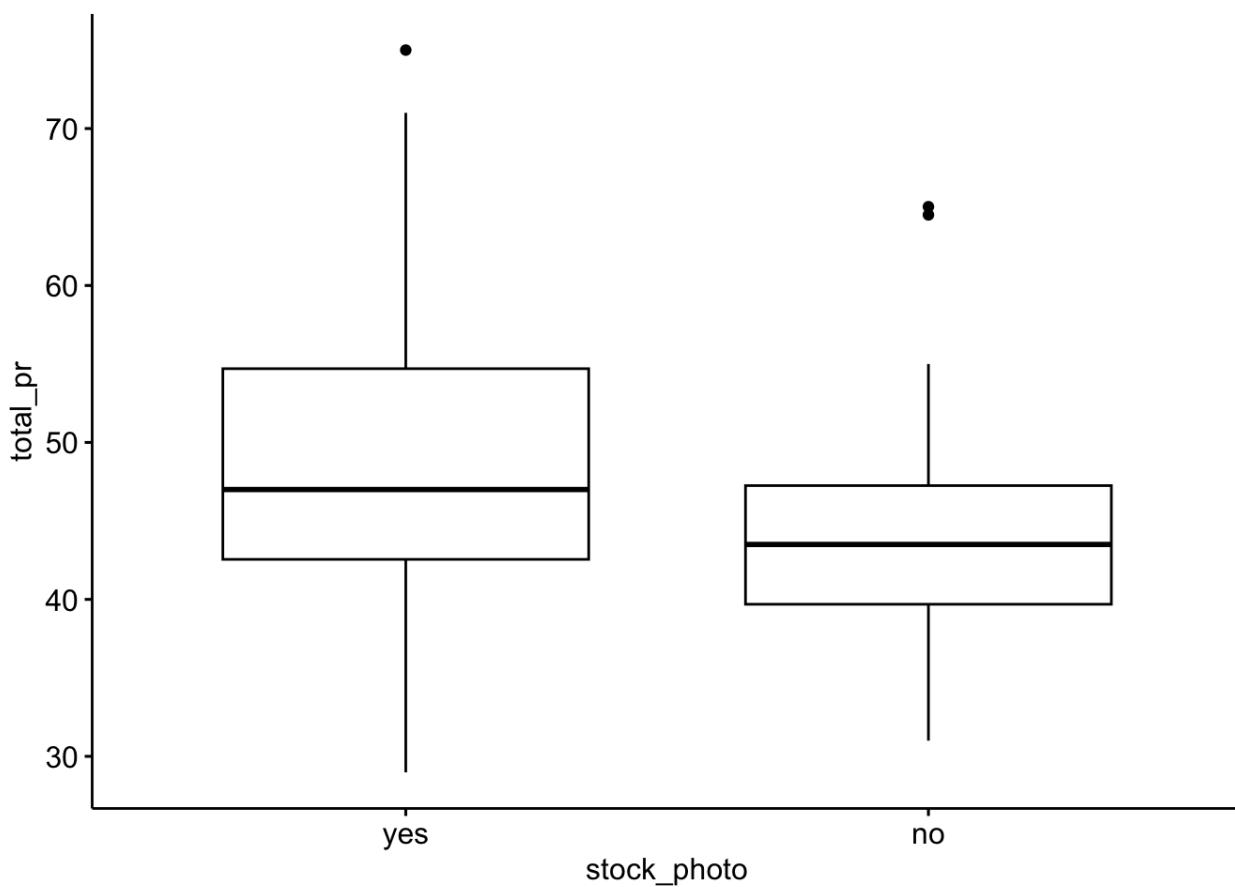
(b) Boxplot pro Gruppe

Abbildung 4.20: Unterschiede zwischen zwei Gruppen: Metrische Y-Variable, nominale X-Variable

Das linke Teildiagramm von [Abbildung 4.20](#) zeigt das Histogramm von `total_pr`, getrennt für neue und gebrauchte Spiele, vgl. [Abbildung 4.9](#). Das rechte Teildiagramm zeigt die gleichen Verteilungen, aber mit einer vereinfachten, groberen Darstellung, den *Boxplot*.



(a) Y: Abschlusspreis, X: Zustand



(b) Y: Abschlusspreis, X: Photo dabei?

Abbildung 4.21: Unterschiede zwischen zwei Gruppen: Metrische Y-Variable, nominale X-Variable

Das linke Teildiagramm von [Abbildung 4.21](#) zeigt den Unterschied in den Verteilungen von `total_pr`, einmal für die neuen Computerspiele (`cond == new`) und einmal für gebrauchte Spiele (`cond == used`).

**Definition 4.6 (Boxplot)** Der Boxplot ist eine Vereinfachung bzw. eine Zusammenfassung eines Histogramms.<sup>7</sup> Damit stellt der Boxplot auch eine Verteilung (einer metrischen Variablen) dar. □

In [Abbildung 4.22](#) sieht man die “Übersetzung” von Histogramm (oben) zu einem Boxplot (unten).

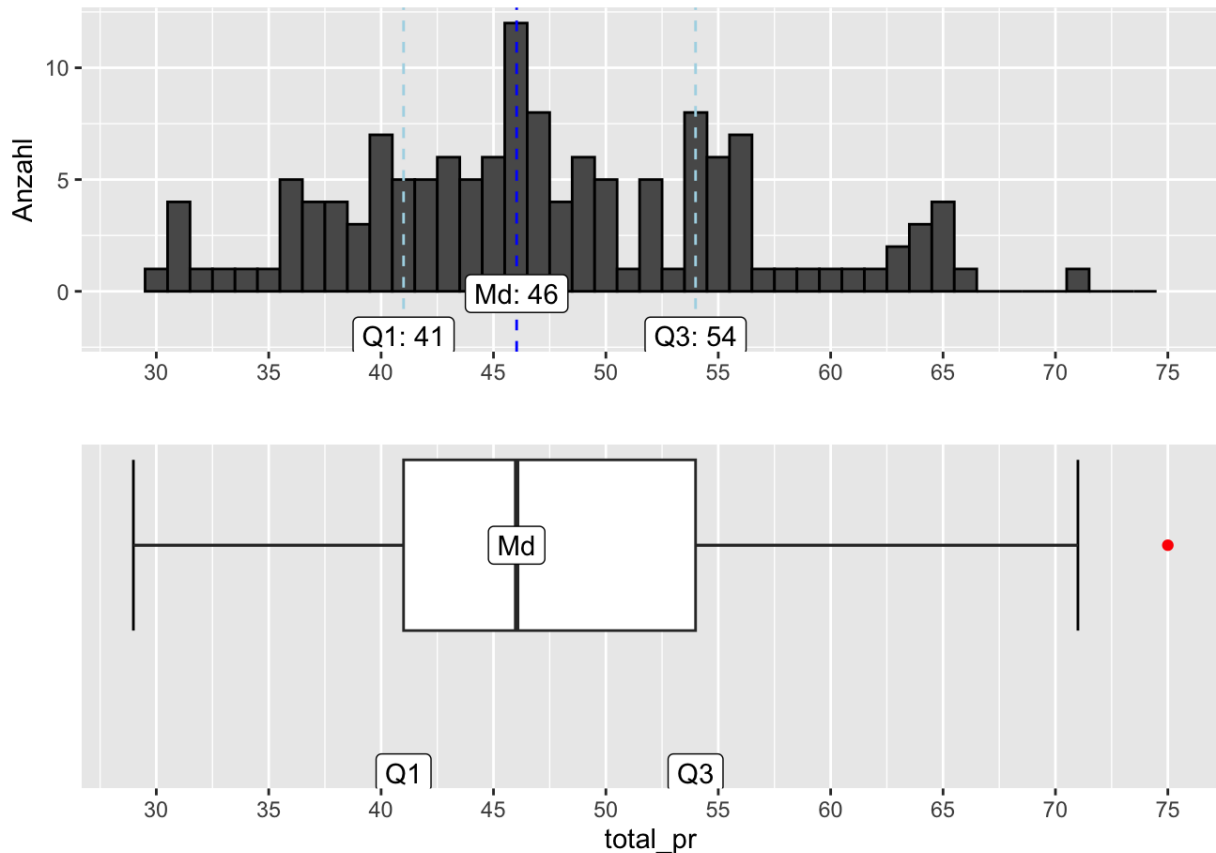


Abbildung 4.22: Übersetzung eines Histogramms zu einem Boxplot

Schauen wir uns die “Anatomie” des Boxplots näher an:

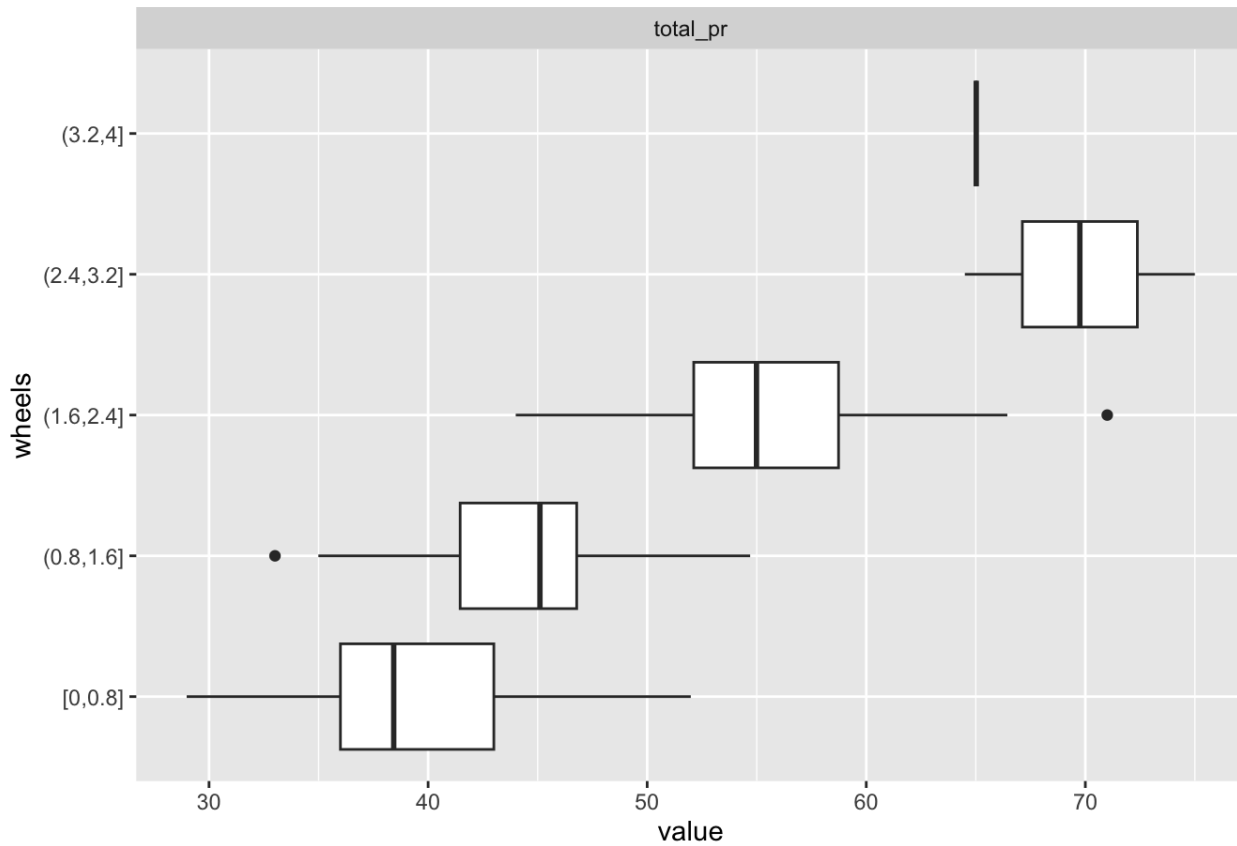
1. Der *dicke Strich* in der Box zeigt den Median der Verteilung
2. Die *Enden der Box* zeigen das 1. Quartil bzw. das 3. Quartil. Damit zeigt die Breite der Box die Streuung der Verteilung an, genauer gesagt die Streuung der inneren 50% der Beobachtungen. Je breiter die Box, desto größer die Streuung. Die Breite der Box nennt man auch den *Interquartilsabstand (IQR)*, da er die Strecke zwischen den Quartilen bemisst.
3. Die “*Antennen*” des Boxplots zeigen die Streuung in den kleinsten 25% der Werte (linke Antenne) bzw. die Streuung der größten 25% der Werte (rechte Antennen). Je länger die Antenne, desto größer die Streuung.
4. Falls es aber *Extremwerte* gibt, so sollten die lieber einzeln, separat, außerhalb der Antennen gezeigt werden. Daher ist die Antennenlänge auf die 1,5-fache Länge der Box beschränkt. Werte die außerhalb dieses Bereichs liegen (also mehr als das 1,5-fache der Boxlänge von Q3 entfernt sind) werden mittels eines Punktes dargestellt.
5. Liegt der Median-Strich in der Mitte der Box, so ist die Verteilung *symmetrisch* (bezogen auf die inneren 50% der Werte), liegt der Median-Strich nicht in der Mitte der Box, so ist die Verteilung nicht symmetrisch (*schief*). Gleiches gilt für die Antennenlängen: Sind die Antennen gleich lang, so ist der äußere Teil der Verteilung symmetrisch, andernfalls schief.

**Beispiel 4.6** In einer vorherigen Analyse haben Sie den Zusammenhang von Abschlusspreis und der Anzahl der Lenkräder untersucht. Jetzt möchten Sie eine sehr ähnliche Fragestellung betrachten: Wie *unterscheiden* sich die Verkaufspreise je nach Anzahl der beigelegten Lenkräder? Flink erstellen Sie dazu folgendes Diagramm,

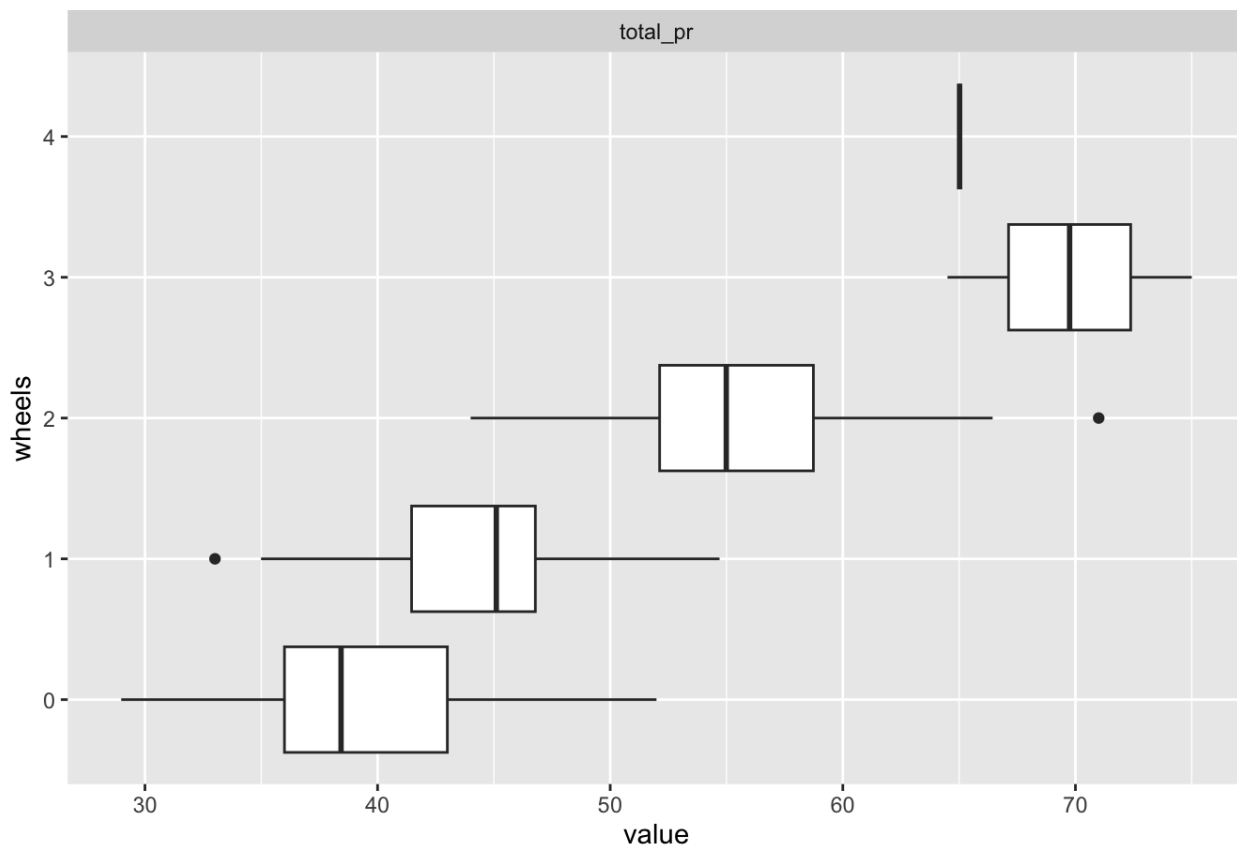
[Abbildung 4.23](#), links. Es zeigt die Verteilung des Abschlusspreises, aufgebrochen nach Anzahl Lenkräder ( `by = "wheels"` ).

Aber ganz glücklich sind Sie mit dem Diagramm nicht: R hat die Variable `wheels` komisch aufgeteilt. Es wäre eigentlich ganz einfach, wenn R die Gruppen `0`, `1`, `2`, `3` und `4` aufteilen würde. Aber schaut man sich die Y-Achse (im linken Teildiagramm von [Abbildung 4.23](#)) an, so erkennt man, dass R `wheels` als stetige Zahl betrachtet und nicht in ganze Zahlen gruppiert.<sup>8</sup> Aber wir möchten jeden einzelnen Wert von `wheels` (0, 1, 2, 3, 4) als *Gruppe* verstehen. Mit anderen Worten, wir möchten `wheels` als nominale Variable definieren. Das kann man mit dem Befehl `factor(wheels)` erreichen (verpackt in `mutate` ), s. [Abbildung 4.23](#), rechts.

```
mariokart2 %>%  
  select(total_pr, wheels) %>%  
  plot_boxplot(by = "wheels")  
  
mariokart2 %>%  
  select(total_pr, wheels) %>%  
  mutate(wheels = factor(wheels)) %>%  
  plot_boxplot(by = "wheels")
```



(a) wheels als metrische Variable



(b) wheels als nominale Variable

Abbildung 4.23: Abschlusspreis nach Anzahl von beigelegten Lenkrädern

Sie schließen aus dem Bild, dass Lenkräder und Preis (positiv) zusammenhängen. Allerdings scheint es wenig Daten für `wheels == 4` zu geben. Das prüfen Sie nach:

```
mariokart2 %>%  
  count(wheels)
```

	wheels	n
	<int>	<int>
	0	36
	1	52
	2	50
	3	2
	4	1

5 rows

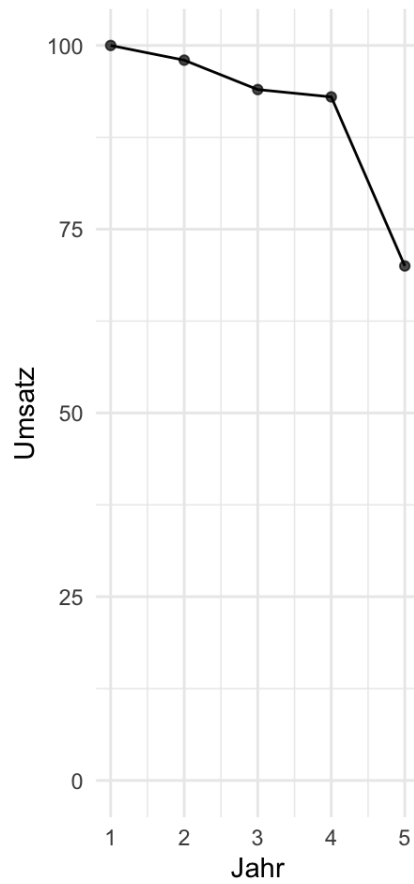
Tatsächlich gibt es (in `mariokart2`) auch für 3 Lenkräder schon wenig Daten, so dass wir die Belastbarkeit dieses Ergebnisses skeptisch betrachten sollten.□

## 4.7 So lügt man mit Statistik

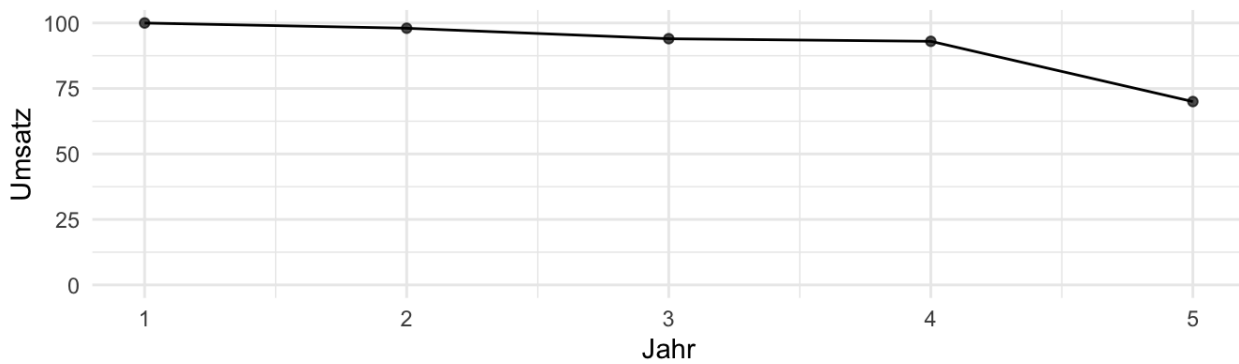
Diagramme werden häufig eingesetzt, um die Wahrheit “aufzuhübschen”.

### 4.7.1 Achsen manipulieren

Achsen zu stauchen ist ein einfacher Trick, s. [Abbildung 4.24](#).



(a) Oh nein, dramatischer Einbruch des Umsatzes!

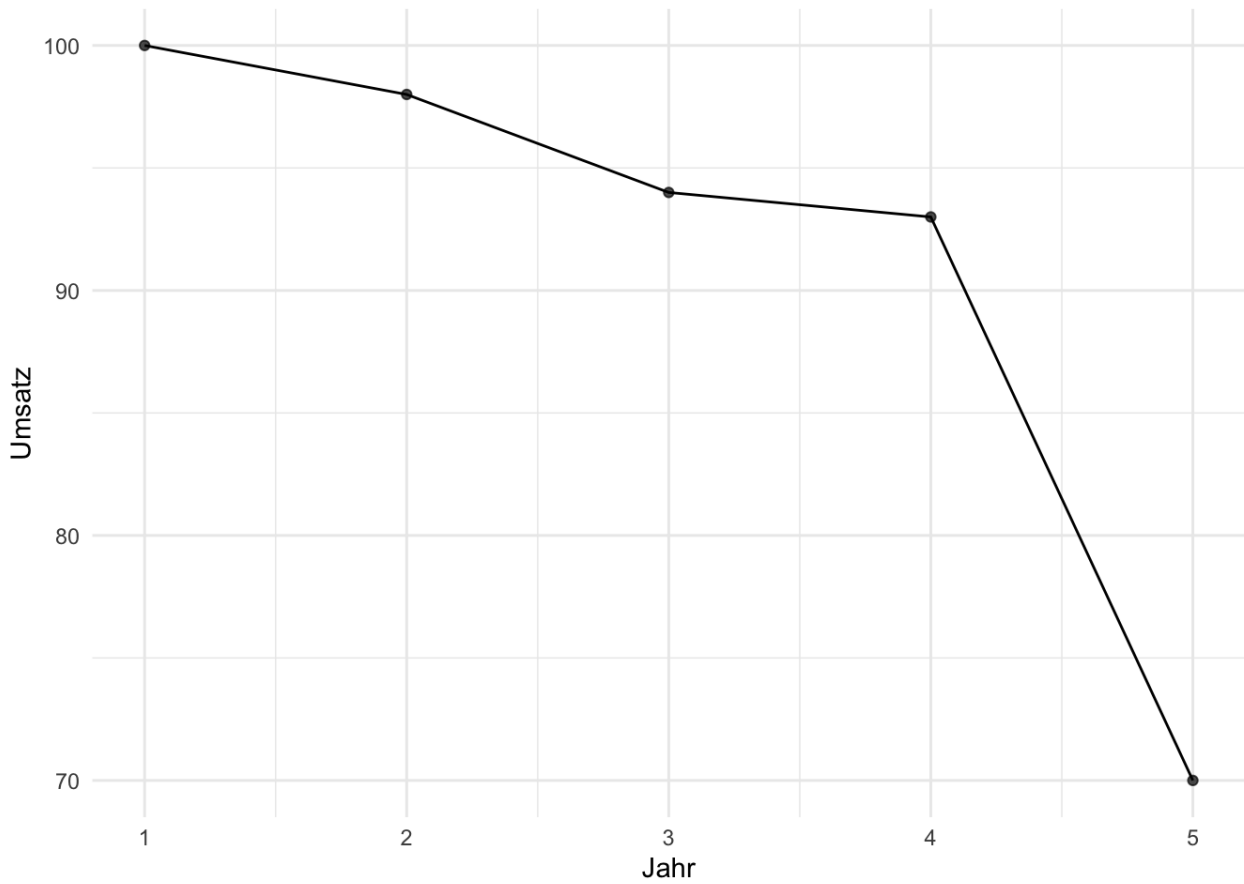


(b) Kaum der Rede wert, ist nur ein bisschen Schwankung!

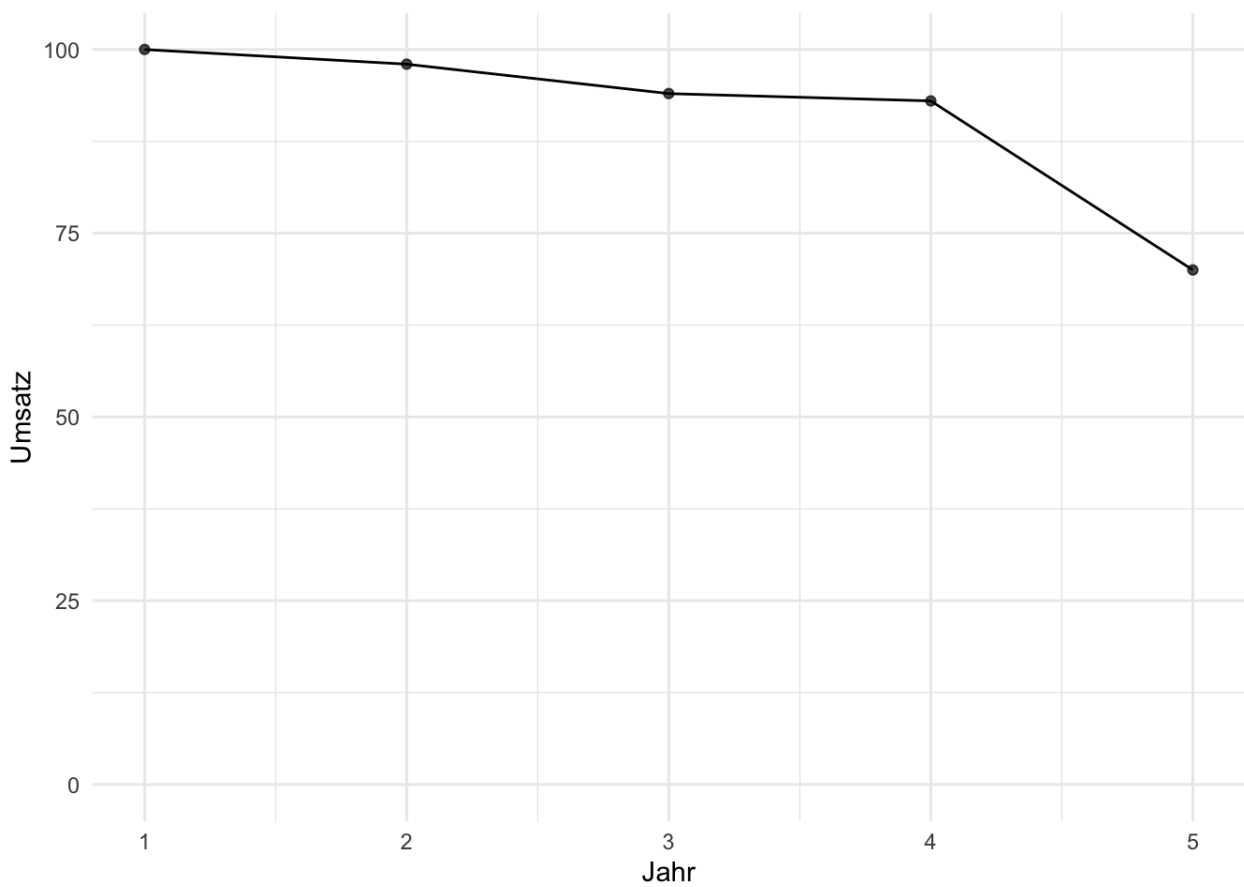
Abbildung 4.24: Stauchen der Y-Achse, um mit Statistik zu lügen



Natürlich kann man auch durch “Abschneiden” der Y-Achse einen eindrucksvollen Effekt erzielen, s. [Abbildung 4.25](#).



*(a) Oh nein, dramatischer Einbruch des Umsatzes!*



*(b) Kaum der Rede wert, ist nur ein bisschen Schwankung!*

Abbildung 4.25: Abschneiden der Y-Achse, um mit Statistik zu lügen

## 4.7.2 Scheinkorrelation

Messerli (2012) berichtet von einem Zusammenhang von Schokoladenkonsum und Anzahl von Nobelpreisen (Beobachtungseinheit: Länder), s. [Abbildung 4.26](#). Das ist doch ganz klar: Schoki füttern macht schlau und Nobelpreise! (?)

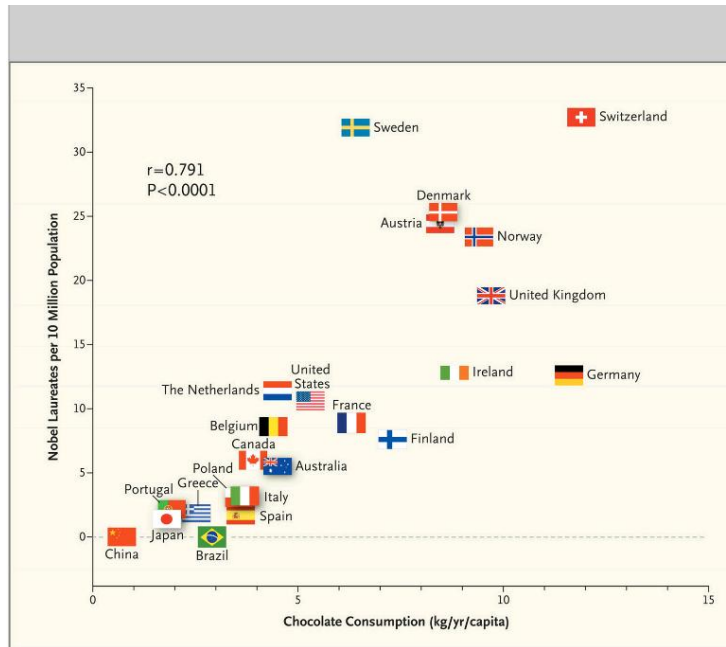


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Abbildung 4.26: Schokoladenkonsum und Nobelpreise

Leider ist hier von einer *Scheinkorrelation* auszugehen: Auch wenn die beiden Variablen *Schokoladenkonsum* und *Nobelpreise* zusammenhängen, heißt das *nicht*, dass die Variable die Ursache und die andere die Wirkung sein muss. So könnte auch eine Drittvariable im Hintergrund die gleichzeitige Ursache von Schokoladenkonsum und Nobelpreise sein, etwa der *allgemeine Entwicklungsstand* des Landes: In höher entwickelten Ländern wird mehr Schokolade konsumiert und es werden mehr Nobelpreise gewonnen im Vergleich zu Ländern mit geringerem Entwicklungsstand.

## 4.8 Praxisbezug

Ein, wie ich finde schlagendes Beispiel zur Stärke von Datendiagrammen ist [Abbildung 4.27](#). Das Diagramm zeigt die Häufigkeit von Masern, vor und nach der Einführung der Impfung. Die Daten und die Idee zur Visualisierung gehen auf Panhuis u. a. (2013) zurück. Das Diagramm und weitere finden sich in ähnlicher Form im [Wall Street Journal](#).

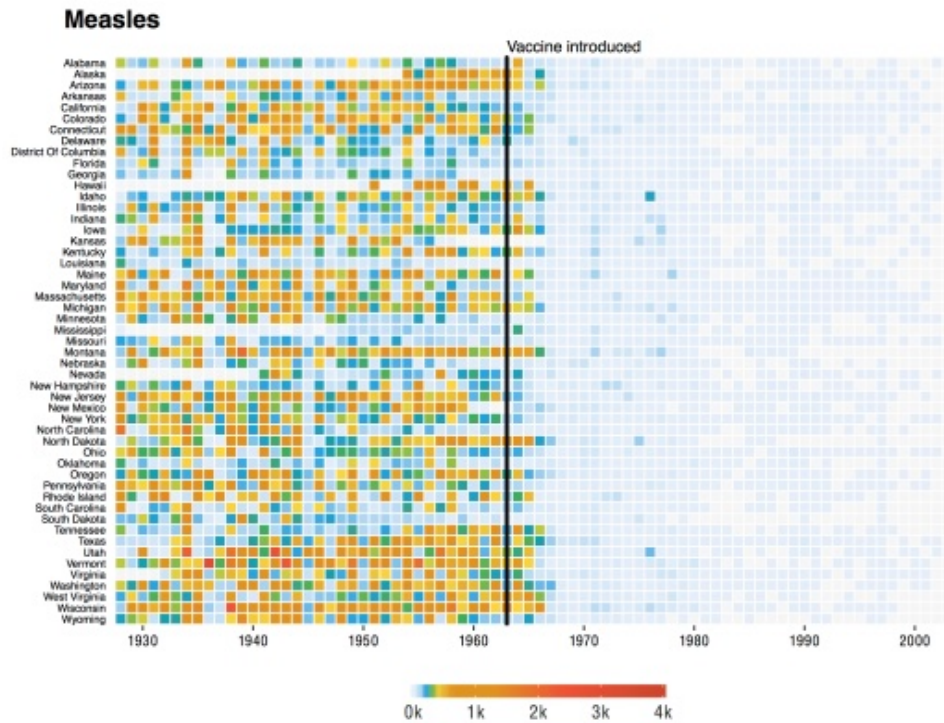


Abbildung 4.27: Häufigkeit von Masern und Impfung in den USA, Lizenz: MIT

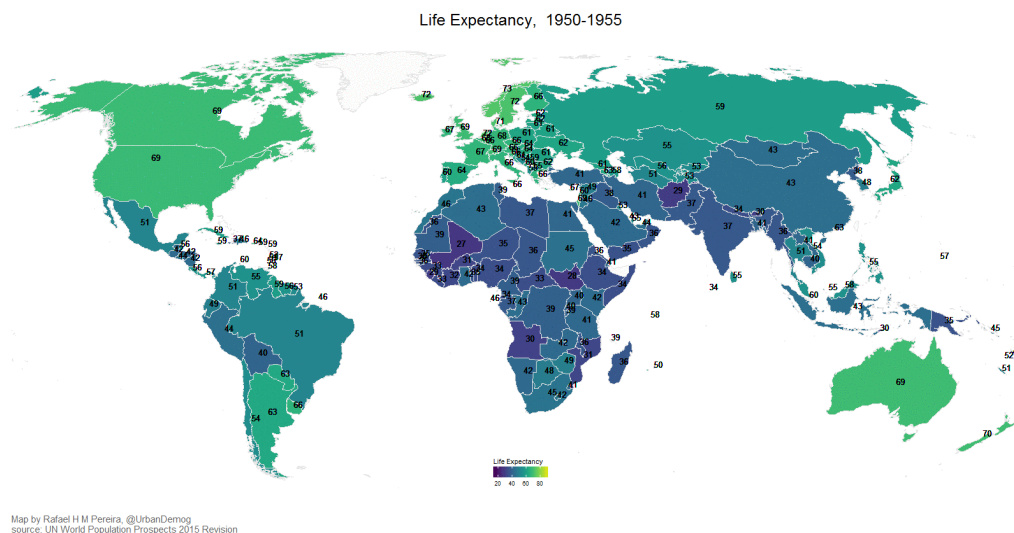
[Quellcode](#) [Datenquelle](#)

In der “freien Wildbahn” findet man häufig sog. “Tortendiagramme”. Zwar sind sie beliebt, doch ist [von ihrer Verwendung abzuraten](#); vgl. auch [hier](#).

## 4.9 Vertiefung

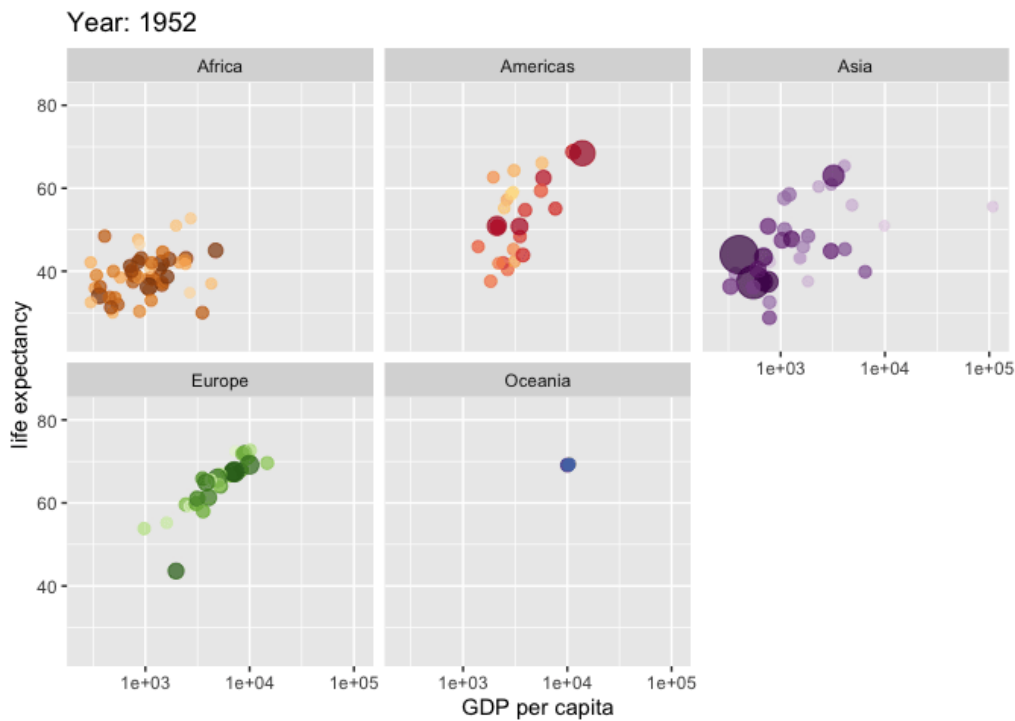
Mehr zu [DataExplorer](#) finden Sie [hier](#).

Eine weitere nützliche Art von Visualisierung sind Karten und Animationen. So zeigt z.B. **?fig-le-world** die Veränderung der Lebenserwartung (in Jahren) über die letzten Dekaden.



Der Quellcode der Animation ist [hier](#) zu finden.

In einigen Situation können Animationen zweckdienlich sein. Außerdem sind sie mitunter nett anzuschauen, s. **?fig-anim1**.



Natürlich sind der Fantasie keine Grenzen beim Visualisieren gesetzt, so ist etwa [diese Animationen](#) ziemlich atemberaubend.

Einen Überblick über verschiedene Typen an Diagrammen, sogar in Form einer systematischen Nomenklatur findet sich bei [data-to-vis](#).

Ein Teil der Diagramm dieses Kapitels wurden mit dem R-Paket [ggpubr](#) erstellt. Mit diesem Paket lassen sich einfach ansprechende Datendiagramme erstellen, so lautet die etwa die Syntax von [Abbildung 4.21](#) wie folgt.

```
library(ggpubr) # einmalig instalieren nicht vergessen
ggboxplot(mariokart2, x = "cond", y = "total_pr")
```

Möchte man Mittelwerte vergleichen, so sind Boxplots nicht ideal, da diese ja nicht den Mittelwert, sondern den Median herausstellen. Eine Abhilfe (also eine Darstellung des Mittelwerts) schafft man (z.B.) mit [ggpubr](#), s. [Abbildung 4.28](#).

```
ggviolin(mariokart2, x = "cond", y = "total_pr",
         add = "mean_sd")
```

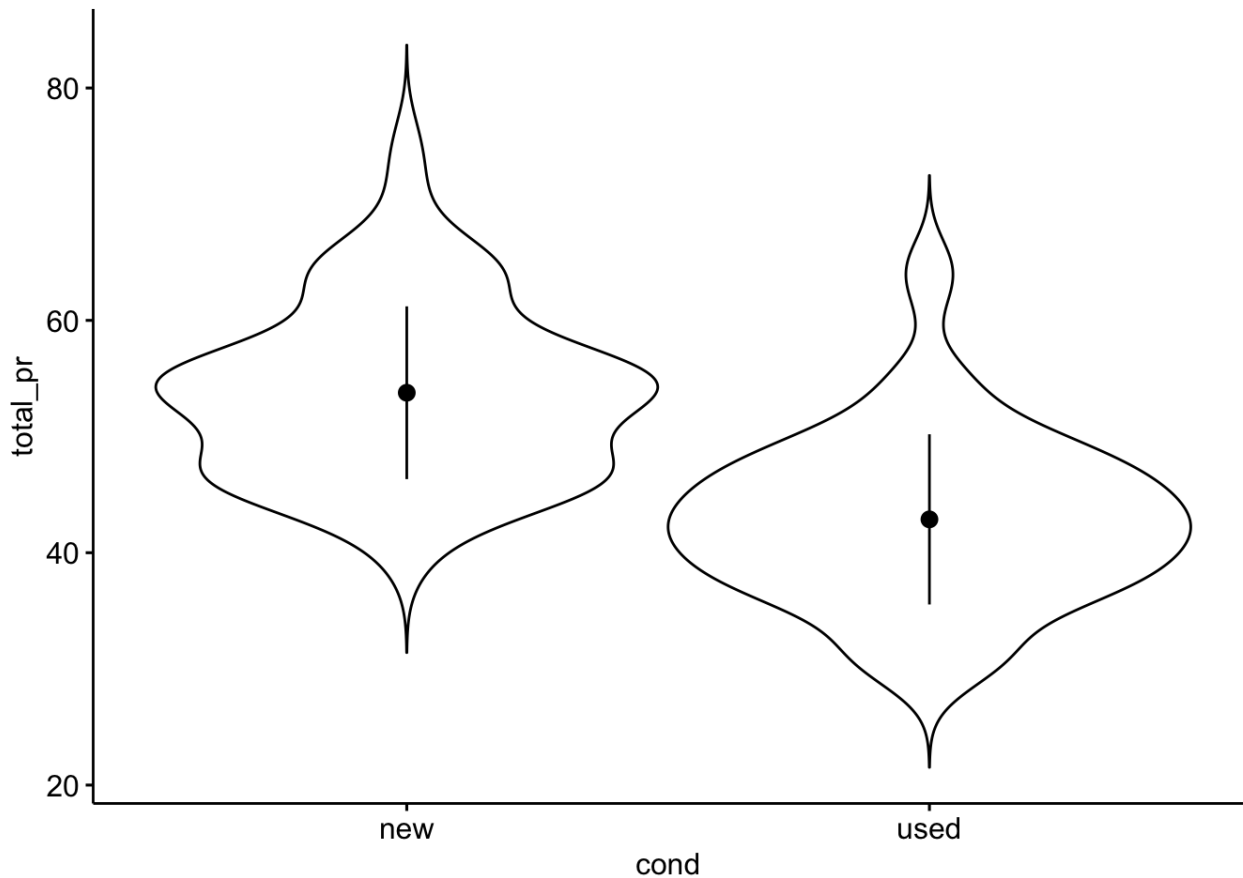


Abbildung 4.28: Vergleich der Verteilungen zweier Gruppen mit Mittelwert und Standardabweichung pro Gruppe hervorgehoben

Ein “Violinenplot” hat die gleiche Aussage wie ein Dichtediagramm: Je breiter die “Violine”, desto mehr Beobachtungen gibt es an dieser Stelle. Weitere Varianten zum Violinenplot mit `ggpubr` finden sich [hier](#).

Sowohl `ggpubr` als auch `DataExplorer` (und viele andere R-Pakete) bauen auf dem R-Paket `ggplot2` auf. `ggplot2` ist eines der am weitesten ausgearbeiteten Softwarepakete zur Erstellung von Datendiagrammen. Das Buch zur Software (vom Autor von `ggplot2`) ist empfehlenswert (Wickham 2009). Eine neue, gute Einführung in Datenvisualisierung findet sich bei Wilke (2019). Beide Bücher sind kostenfrei online lesbar.

Übrigens sind Modelle - und Diagramme sind Modelle - immer eine Vereinfachung, lassen also Informationen weg. Manchmal auch wichtige Informationen. [Dieses Beispiel](#) zeigt, wie etwa Histogramme wichtige Informationen unter den Tisch fallen lassen.



Ich würde gerne mal Beispiele von *schlechten* Datendiagrammen sehen.



Auf der Seite von [Flowingdata](#) findet sich eine nette Liste mit schlechten Datendiagrammen.

## 4.10 Aufgaben








1. [boxhist](#)
2. [max-corr1](#)
3. [max-corr2](#)
4. [Histogramm-in-Boxplot](#)
5. [Diamonds-Histogramm-Vergleich2](#)
6. [Boxplot-Aussagen](#)

- 7. [boxplots-de1a](#)
- 8. [movies-vis1](#)
- 9. [movies-vis2](#)

## 4.11 Literatur

---

- Anscombe, Francis J. 1973. „Graphs in statistical analysis“. *The American Statistician* 27 (1): 17–21.
- Cohen, J. 1992. „A power primer“. *Psychological Bulletin* 112 (1): 155–59.
- Lyon, Aidan. 2014. „Why Are Normal Distributions Normal?“ *The British Journal for the Philosophy of Science* 65 (3): 621–49. <https://doi.org/10.1093/bjps/axs046>.
- Messerli, Franz H. 2012. „Chocolate Consumption, Cognitive Function, and Nobel Laureates“. *New England Journal of Medicine* 367 (16): 1562–64. <https://doi.org/10.1056/NEJMon1211064>.
- Panhuys, Willem G. van, John Grefenstette, Su Yon Jung, Nian Shong Chok, Anne Cross, Heather Eng, Bruce Y. Lee, u. a. 2013. „Contagious Diseases in the United States from 1888 to the Present“. *New England Journal of Medicine* 369 (22): 2152–58. <https://doi.org/10.1056/NEJMms1215400>.
- Scherer, Cédric, Viktoriia Radchuk, Christoph Staubach, Sophie Müller, Niels Blaum, Hans-Hermann Thulke, und Stephanie Kramer-Schadt. 2019. „Seasonal Host Life-history Processes Fuel Disease Dynamics at Different Spatial Scales“. Herausgegeben von Ann Tate. *Journal of Animal Ecology* 88 (11): 1812–24. <https://doi.org/10.1111/1365-2656.13070>.
- Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Use R! New York: Springer. <https://doi.org/10.1007/978-0-387-98141-3>.
- Wilke, C. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. First edition. Sebastopol, CA: O'Reilly Media. <https://clauswilke.com/dataviz/>.

- 
1. 5 
  2. Weitere Nomenklaturen sind möglich, aber wir halten hier die Sache einfach. 
  3. Natürlich müssen Sie das Paket einmalig installiert haben, bevor Sie es starten können. 
  4. bei konstanter Balkenbreite 
  5. Mit *Dichte* ist die Anzahl der Beobachtungen pro Einheit der Variablen auf der X-Achse gemeint. 
  6. Grob gesagt: `mariokart %>% plot_density()` 
  7. Ob der Boxplot horizontal oder vertikal steht, ist Ihrem Geschmack überlassen. 
  8. Vielleicht so, dass in jeder Gruppe gleich viele Wert sind? 