Guidelines für die Visualisierung von Daten [Dataviz]

Prinzipien der Datenvisualisierung

- Bisher hatten wir uns hauptsächlich auf die technische Seite der Visualisierung mit `ggplot2` konzentriert
- Hier sollen allgemeine Prinzipien und Guidelines zur Datenvisualisierung vorgestellt werden
- Diese Vorlesungseinheit orientiert sich an einem Vortrag von Karl Broman mit dem Titel: "Creating effective figures and tables", Vorlesungsfolien von Peter Aldhous Introduction to Data Visualization course und dem Buch Introduction to Data Science (Kapitel 10)

Prinzipien der Datenvisualisierung

Aufbau der Vorlesungseinheit:

- Beispiele unvorteilhafter Grafiken aufzeigen
- Vorschläge machen, wie diese verbessert werden können
- Allgemeine Prinzipien zur Visualisierung aus diesen Beispielen ableiten

Sie sollten bei der Visualisierung von Daten immer ihr Ziel im Auge behalten:

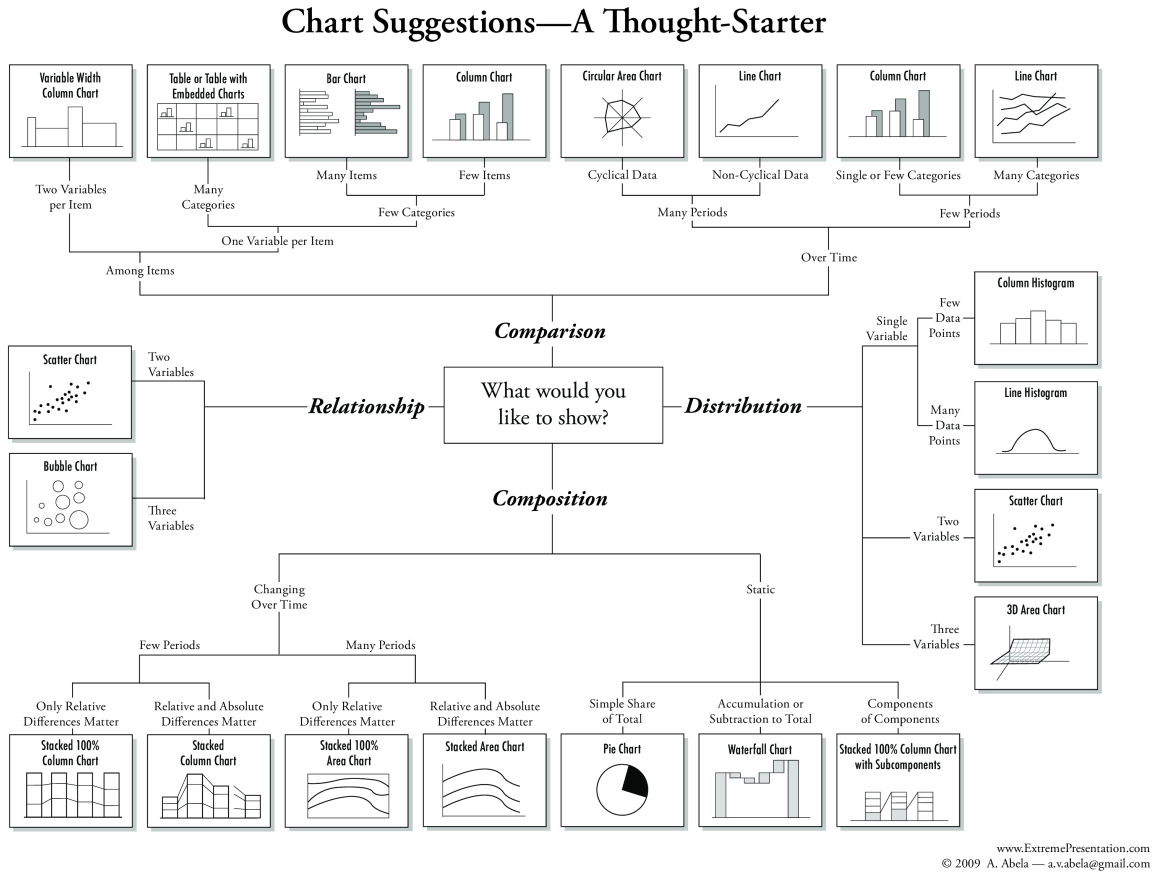
- Explorative Grafiken nur für Sie selbst können niedrigeren Standards genügen
- Grafiken in einer Präsentation oder Ausarbeitung sind für den Zuhörer/Leser und müssen diesen überzeugen

Prinzipien der Datenvisualisierung

Wir verwenden diese Pakete:

```
library(tidyverse)
library(readxl)
library(gridExtra)
library(ggthemes)
library(gganimate)
library(pander)
```

Möglichkeiten der Visualisierung



Quelle: https://extremepresentation.typepad.com/blog/2006/09/choosing_a_good.html

Möglichkeiten der Visualisierung

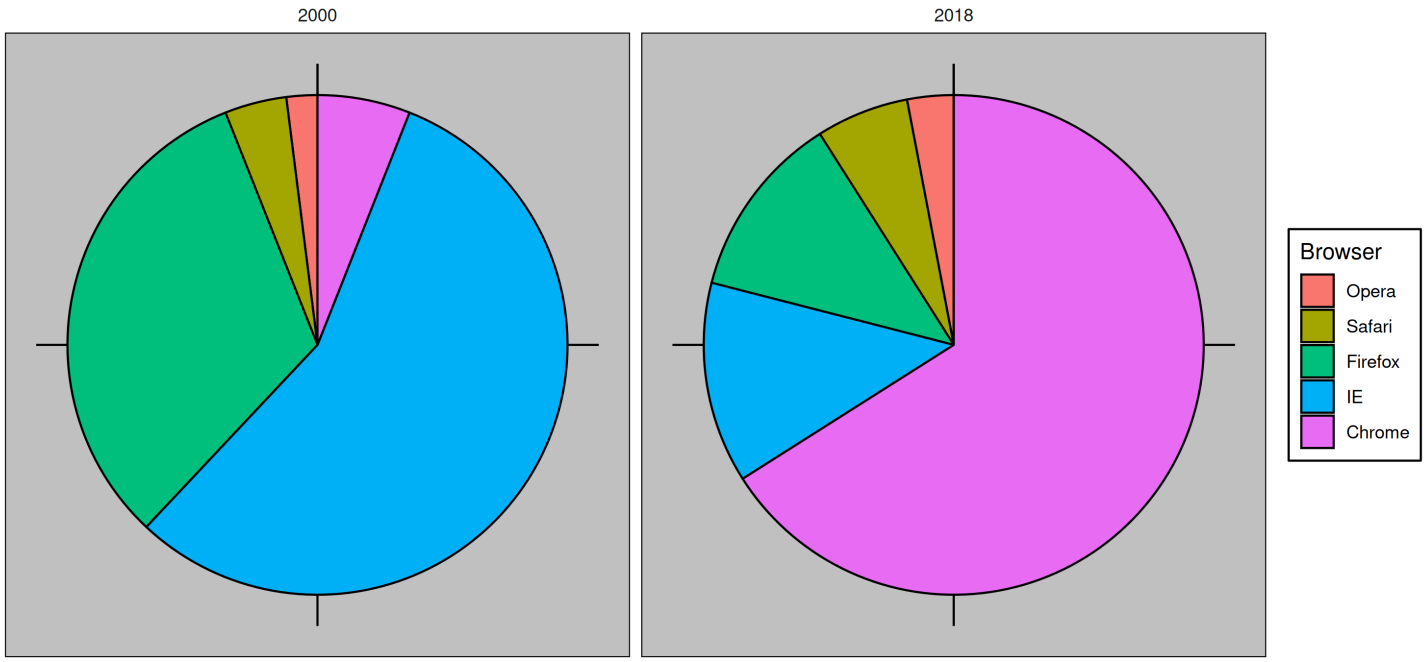
Neben den verschiedenen Typen von Grafiken stehen uns viele weitere Möglichkeiten zur Verfügung wie wir Daten visualisieren können. Bspw. durch Wahl der:

- Position
- Ausrichtung
- Winkel
- Fläche
- Helligkeit
- Farbgebung

Möglichkeiten der Visualisierung

Beispiel: Nutzung von Internet-Browsern für Januar 2000 und Januar 2018

- Daten zur Nutzung von Internet-Browsers können Sie unter StatCounter einsehen
- In vielen Präsentation wird für diese Fragestellung ein Kuchendiagramm verwendet:



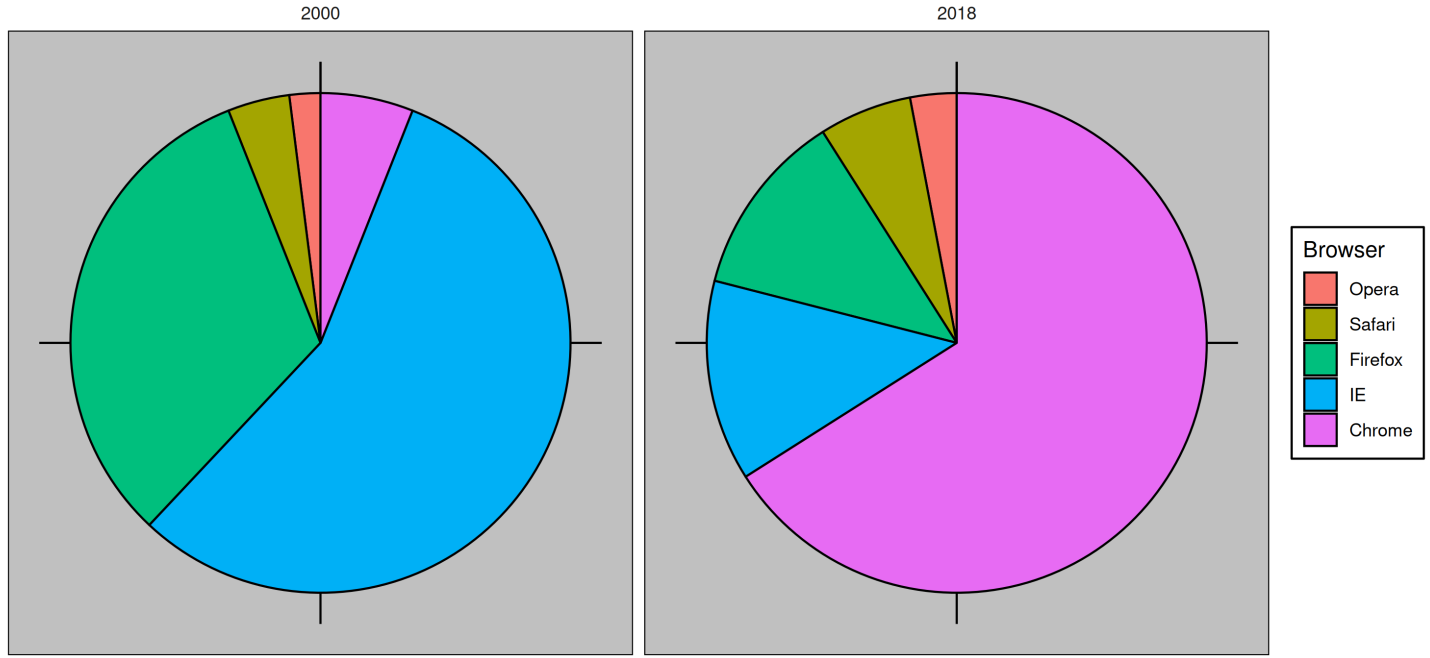
Möglichkeiten der Visualisierung

- Die **Fläche und der Winkel** im Kuchendiagramm werden verwendet um den Anteil jedes Browsers am Gesamtmarkt aufzuzeigen
- Diese Darstellungsweise ist **suboptimal**, da Menschen nicht gut darin sind Winkel abzuschätzen und sogar noch schlechter bei Flächen

Kuchendiagramme

Beantworten Sie sich beispielsweise folgende Fragen:

> Um wie viel verändert sich der Marktanteil jedes Browsers von 2000 zu 2018?



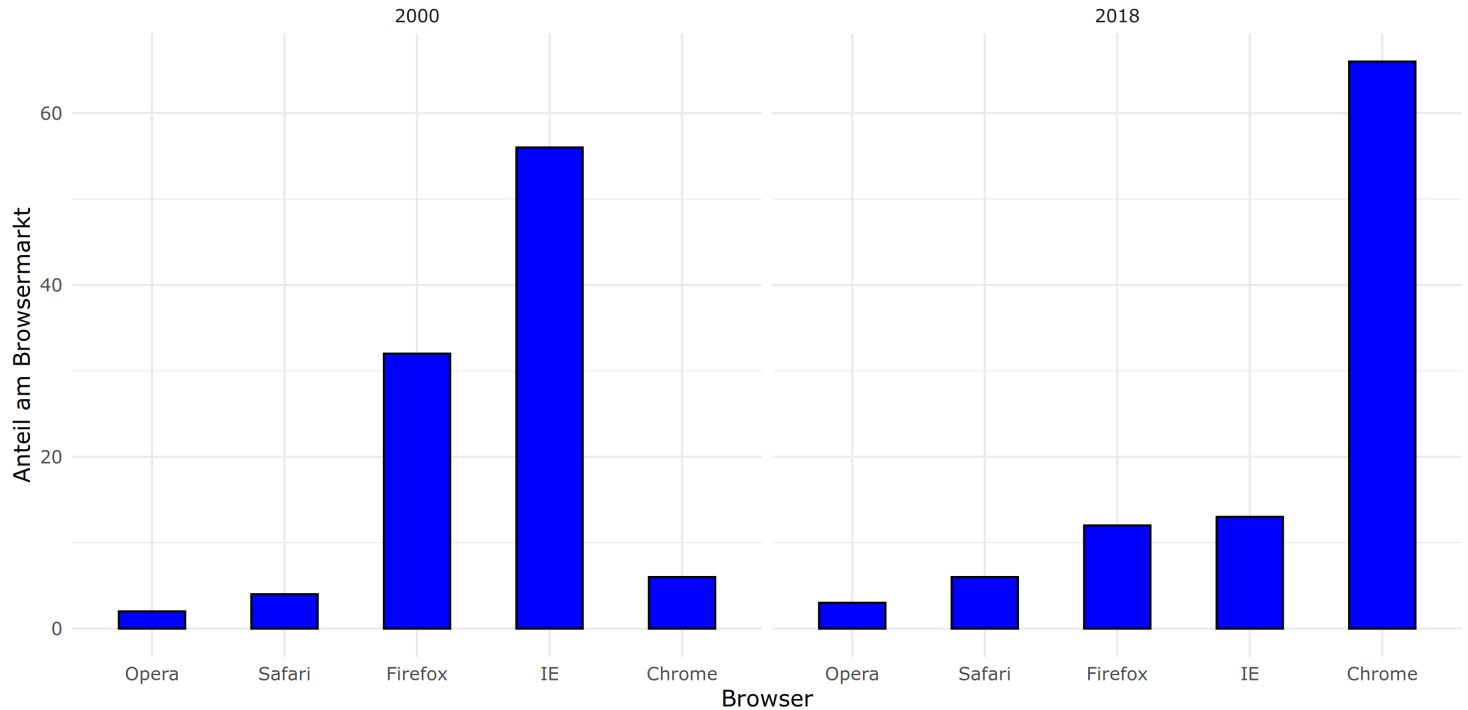
Kuchendiagramme

- ✚ Es wäre in diesem Fall einfacher und verständlicher die nackten Zahlen zu präsentieren:

Browser	2000	2018
Opera	2	3
Safari	4	6
Firefox	32	12
IE	56	13
Chrome	6	66

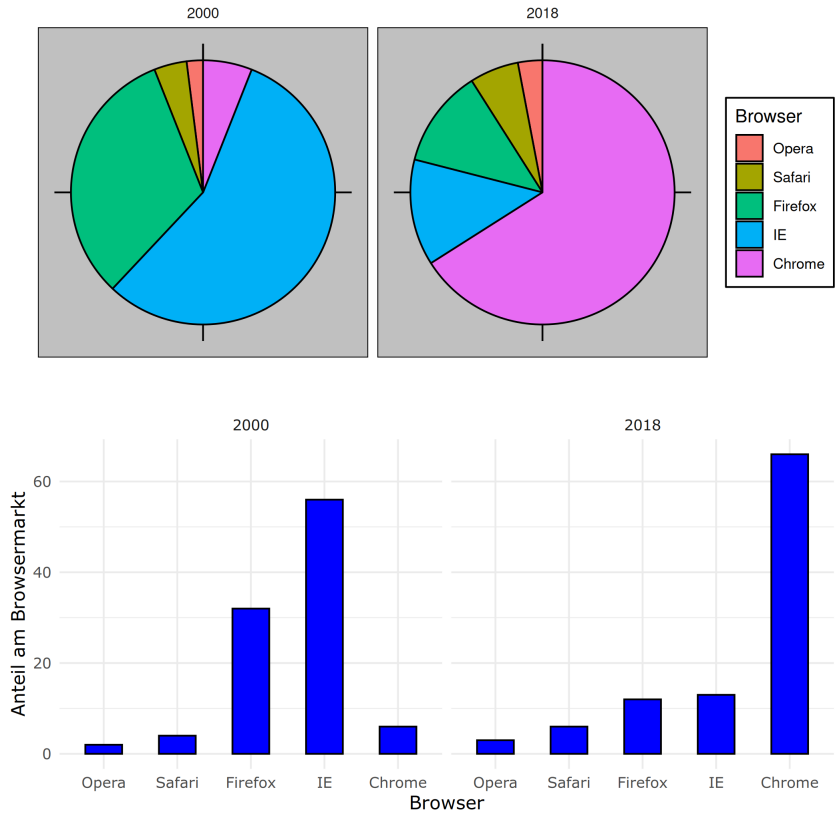
Balkendiagramm

- Mit einem Balkendiagrammen kommen Sie der menschlichen Wahrnehmung besser entgegen
- Das menschliche Gehirn ist deutlich besser im Abschätzen von Längen als von Winkeln
- Durch Hilfslinien (hier für jede 10%) können Sie den Leser zusätzlich beim schnellen Verständnis unterstützen:



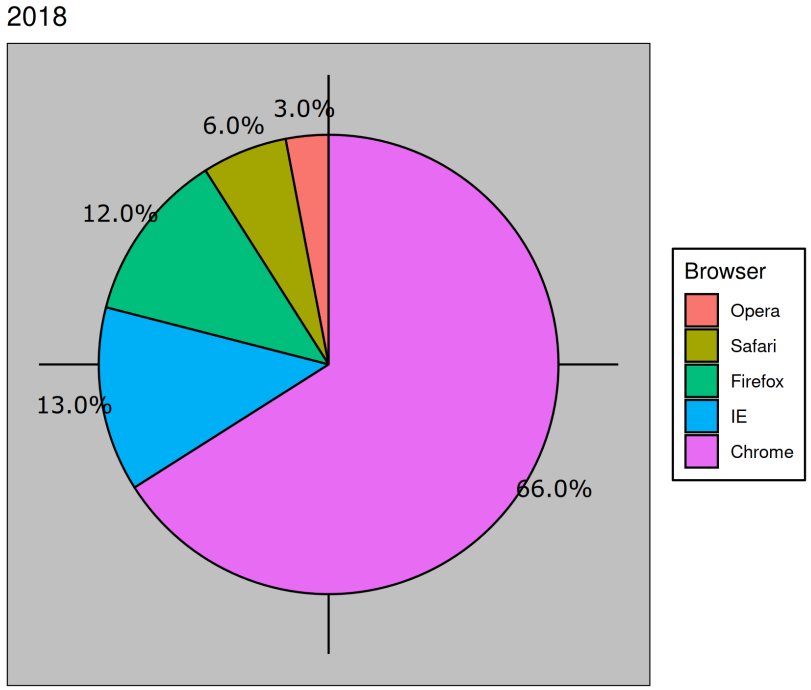
Balkendiagramm

Ein Vergleich zwischen Kuchen und Balkendiagramm:



Kuchendiagramm vs. Balkendiagramm

- Durch das Balkendiagramm können Sie prozentuale Unterschiede direkt ablesen
- Sollten Sie dennoch ein Kuchendiagramm nutzen, dann sollten Sie die prozentualen Anteile in das Diagramm mit aufnehmen um Abschätzungen der Fläche oder Winkel zu umgehen:



Balkendiagramme

> Es gilt: Balkendiagrammen **immer** bei Null beginnen

Balkendiagramme

Es gilt: Balkendiagrammen **immer** bei Null beginnen

- ✚ Durch die Verwendung eines Balkendiagramms wird automatisch impliziert, dass die Länge des Balkens proportional zur gezeigten Stückzahl ist

Balkendiagramme

> Es gilt: Balkendiagrammen **immer** bei Null beginnen

- ✚ Durch die Verwendung eines Balkendiagramms wird automatisch impliziert, dass die Länge des Balkens proportional zur gezeigten Stückzahl ist
- ✚ Wenn Sie die Null nicht in ihr Balkendiagramm aufnehmen können kleine Differenzen viel größer erscheinen, als sie eigentlich sind

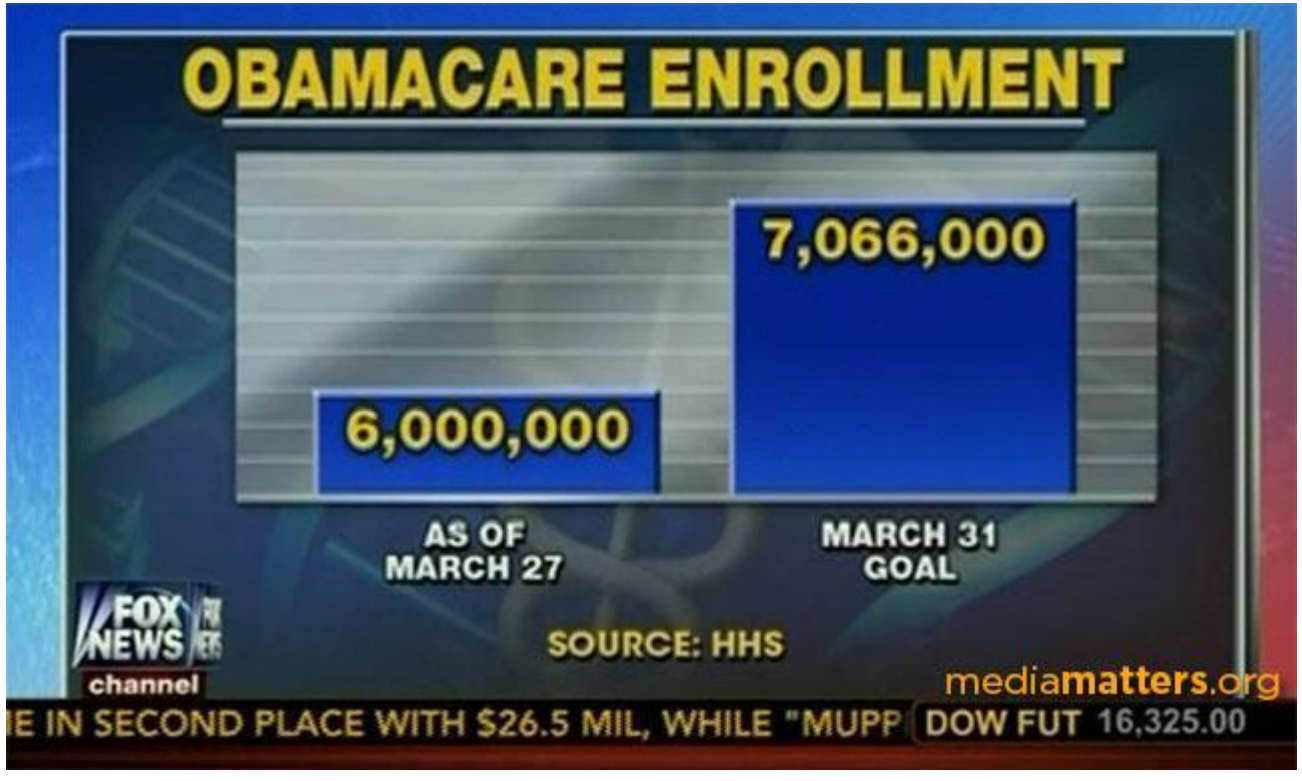
Balkendiagramme

Es gilt: Balkendiagrammen **immer** bei Null beginnen

- Durch die Verwendung eines Balkendiagramms wird automatisch impliziert, dass die Länge des Balkens proportional zur gezeigten Stückzahl ist
- Wenn Sie die Null nicht in ihr Balkendiagramm aufnehmen können kleine Differenzen viel größer erscheinen, als sie eigentlich sind
- Dies wird oft in der Politik oder den Medien verwendet um sich besser darzustellen

Balkendiagramme

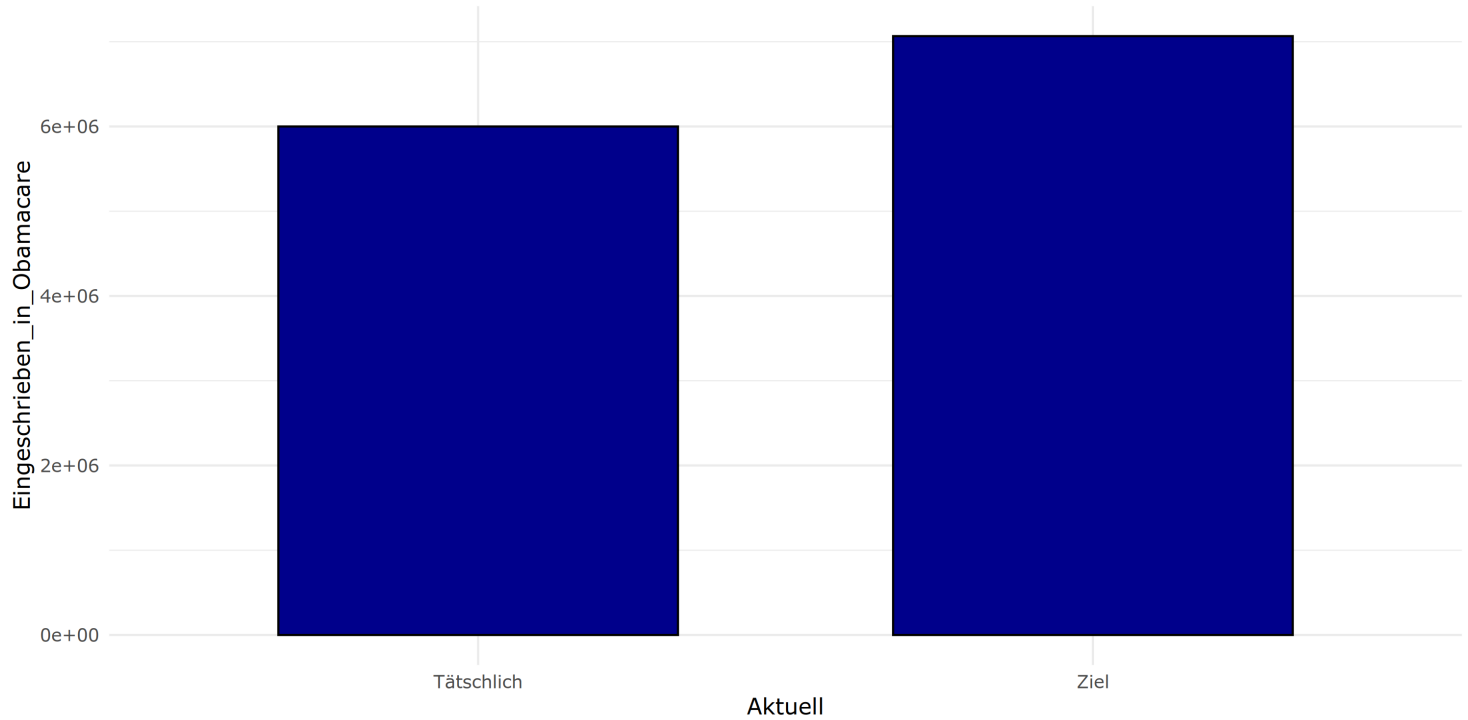
Hier ein Beispiel aus den FOX News:



Quelle: Fox News; https://flowingdata.com/2015/08/11/real-chart-rules-to-follow/

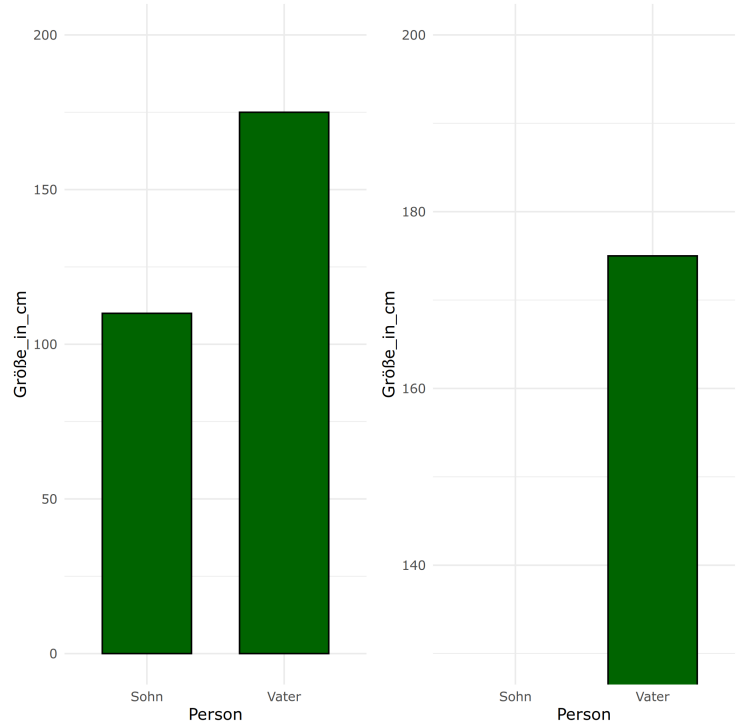
Balkendiagramme

- In der Grafik erscheint es so als ob Obamacare zum Ziele hatte drei mal so viele Personen zu versichern als aktuell versichert sind
- Jedoch ist das Ziel Ende März nur um 17,8% höher als die tatsächlich eingeschriebene Zahl an Versicherten
- Eine Grafik, welche bei Null beginnt macht den Unterschied deutlich:



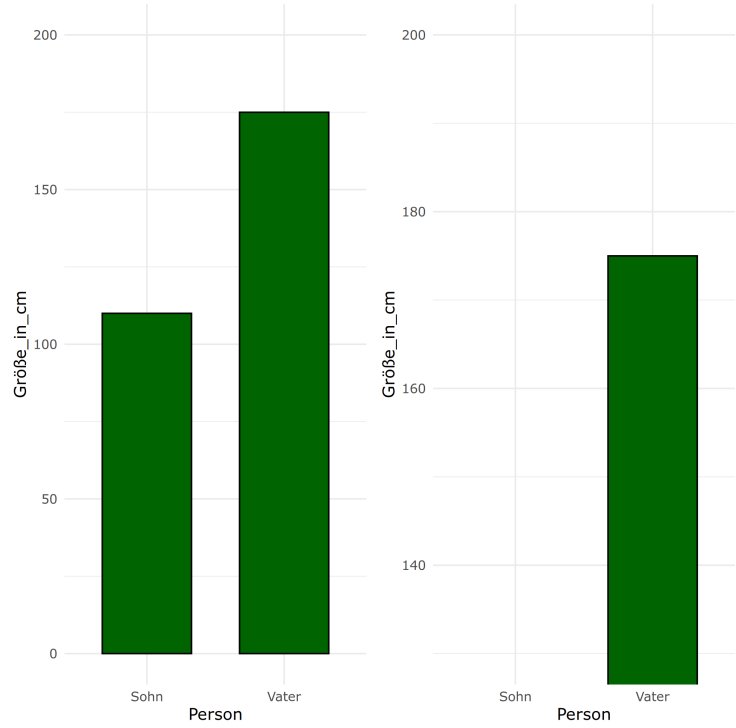
Balkendiagramme

> Angenommen Sie vergleichen die Körpergrößen von Vater (175 cm) und Sohn (110 cm) auf gleiche Weise



Balkendiagramme

> Angenommen Sie vergleichen die Körpergrößen von Vater (175 cm) und Sohn (110 cm) auf gleiche Weise



Hier wird deutlich, warum es wichtig ist ein Balkendiagramm bei Null zu beginnen

Schaubilder immer bei Null beginnen?

Nein!

Schaubilder immer bei Null beginnen?

Nein!

Wenn Sie die Position im Schaubild und nicht die Länge (von Balken) verwenden, dann ist es **nicht nötig** das Schaubild bei Null beginnen zu lassen

- Liniendiagramme
- Punktediagramme
- Bubble Grafiken
- ...

müssen alle **nicht** bei Null beginnen.

Zeigen Sie ihre Daten

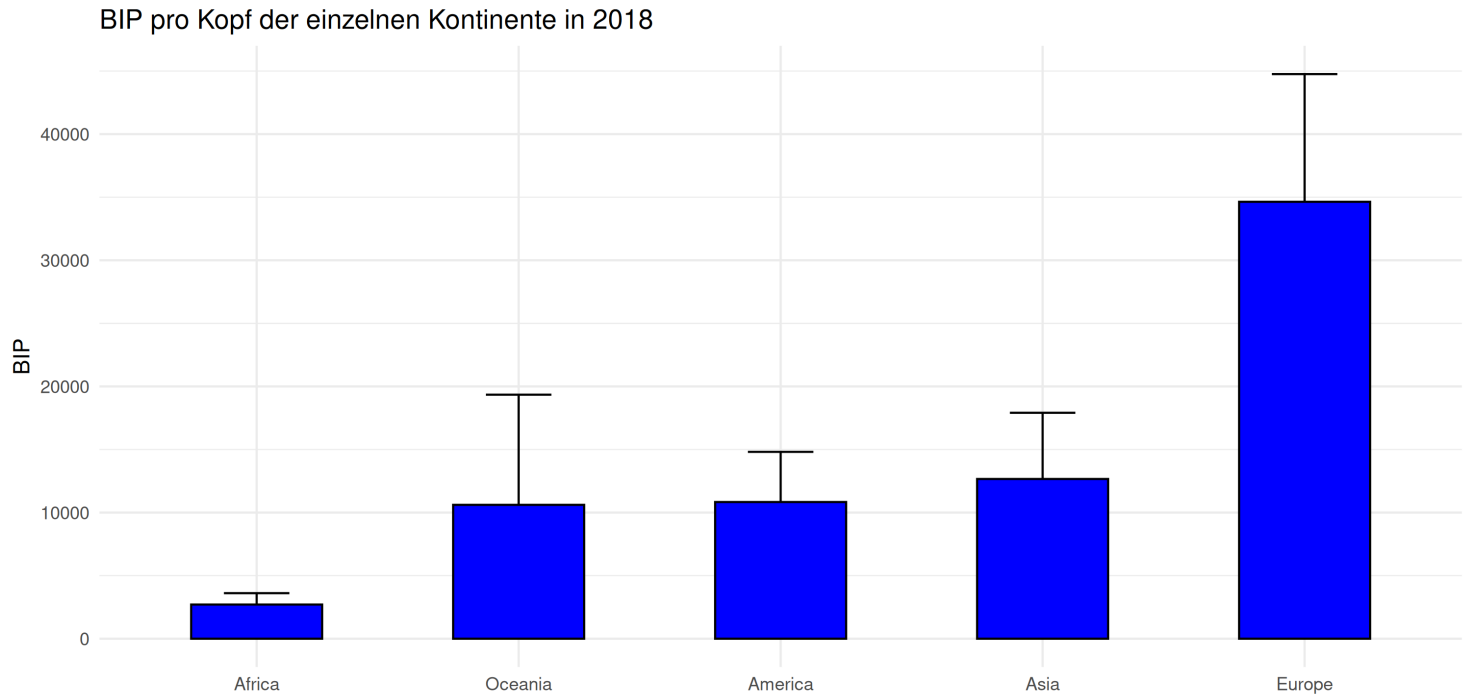
Gegeben Sie wollen das BIP aller Kontinente in 2018 einander gegenüberstellen

- Standardmäßig wird hierfür eine Grafik gezeigt, welche den Mittelwert als Balken mit den dazugehörigen Standardfehlern zeigt

Zeigen Sie ihre Daten

Gegeben Sie wollen das BIP aller Kontinente in 2018 einander gegenüberstellen

- Standardmäßig wird hierfür eine Grafik gezeigt, welche den Mittelwert als Balken mit den dazugehörigen Standardfehlern zeigt



Zeigen Sie ihre Daten

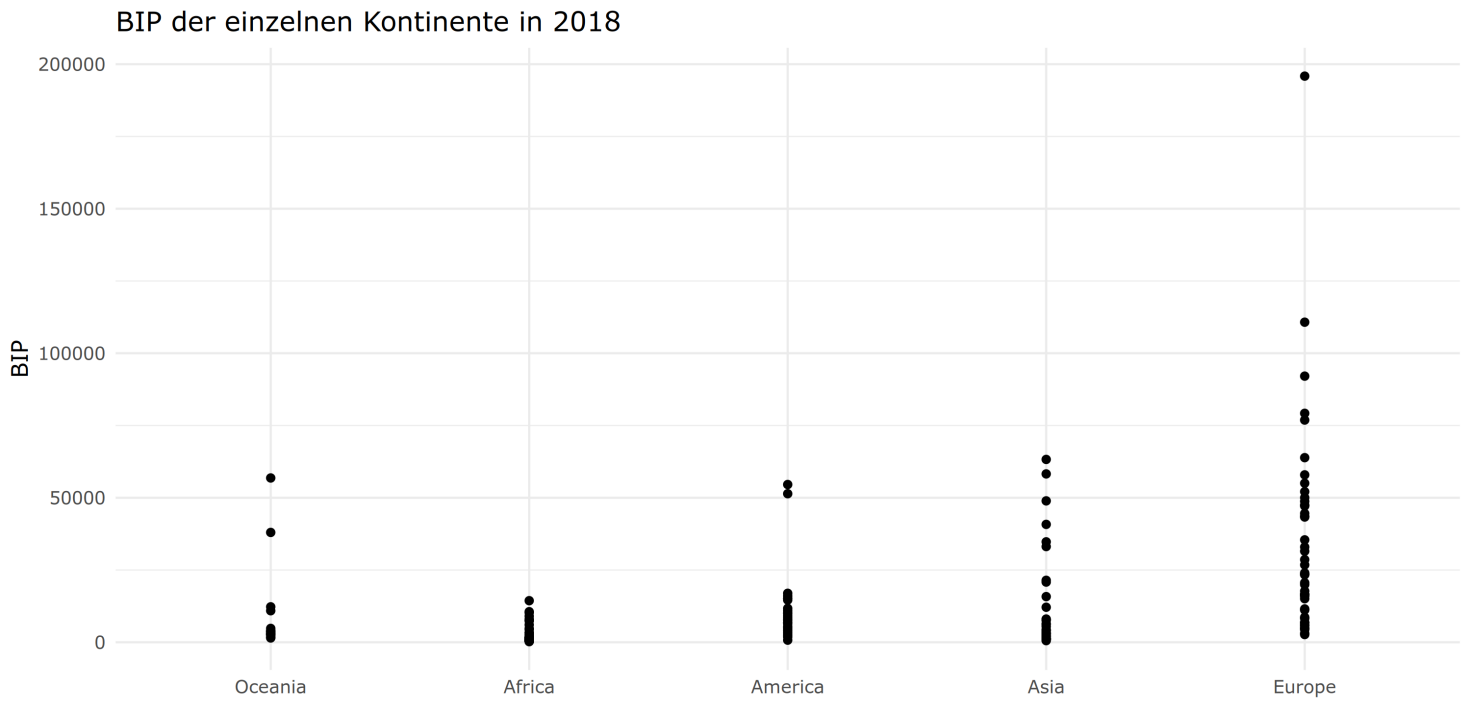
- Sie können allgemeine Trends in ihren Daten aufzeigen, z.B. ist das durchschnittliche BIP in Afrika am geringsten und Europa am höchsten
- Allerdings können Sie nichts zur Variabilität innerhalb der Kontinente sagen
  - Die Verteilung des BIPs ist hier schwer zu beurteilen
- Zwei einfache Fragen, welche mit dem Balkendiagramm nicht beantwortet werden können:
  - Haben alle Länder in Afrika ein geringeres BIP als in Ozeanien?
  - Wie ist die Verteilung auf den jeweiligen Kontinenten?

Zeigen Sie ihre Daten

- Sie können allgemeine Trends in ihren Daten aufzeigen, z.B. ist das durchschnittliche BIP in Afrika am geringsten und Europa am höchsten
- Allerdings können Sie nichts zur Variabilität innerhalb der Kontinente sagen
  - Die Verteilung des BIPs ist hier schwer zu beurteilen
- Zwei einfache Fragen, welche mit dem Balkendiagramm nicht beantwortet werden können:
  - Haben alle Länder in Afrika ein geringeres BIP als in Ozeanien?
  - Wie ist die Verteilung auf den jeweiligen Kontinenten?

> Diesen Fragen können Sie sich nähern indem Sie alle Datenpunkte zeigen

Zeigen Sie ihre Daten



Zeigen Sie ihre Daten

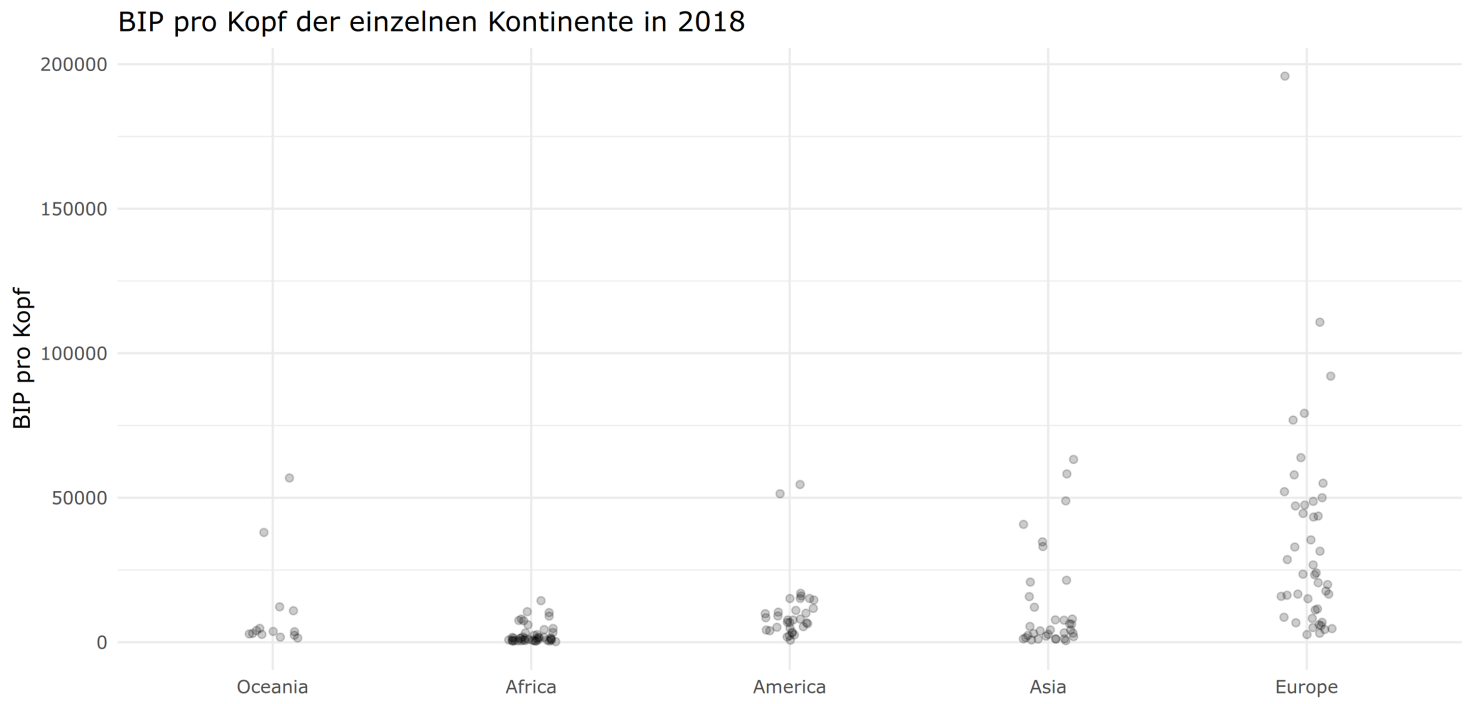
- Hier bekommen Sie eine Idee davon, in welchem Bereich das BIP für die einzelnen Kontinente liegen
- Diese Grafik ist immer noch problematisch, da Sie nicht das BIP eines jeden Landes sehen können; viele Datenpunkte liegen übereinander

Möglichkeiten um die Punktewolke etwas zu entzerren:

- Mit *jitter* können Sie jeden Datenpunkt zufällig um einen kleinen Bereich horizontal verschieben
  - In unserem Beispiel können wir dadurch die übereinander liegenden Punkte etwas voneinander abgrenzen
- Mit *alpha blending* können Sie die Datenpunkte transparent machen
  - Wenn mehrere Punkte aufeinander fallen erscheint dieser Bereich dunkler

Durch *jitter* und *alpha blending* bekommen Sie ein besseres Gefühl für die Verteilung der Daten

Zeigen Sie ihre Daten



Zeigen Sie ihre Daten

- Nun sehen Sie, dass die meisten Länder Afrikas deutlich ärmer sind als die Europas, aber das viele Länder Americas ein ähnliches BIP aufweisen wie die Asiens
- Weiterhin ist die Verteilung der Länder pro Kontinent recht ähnlich, mit Ausnahme von Europa

Zeigen Sie ihre Daten

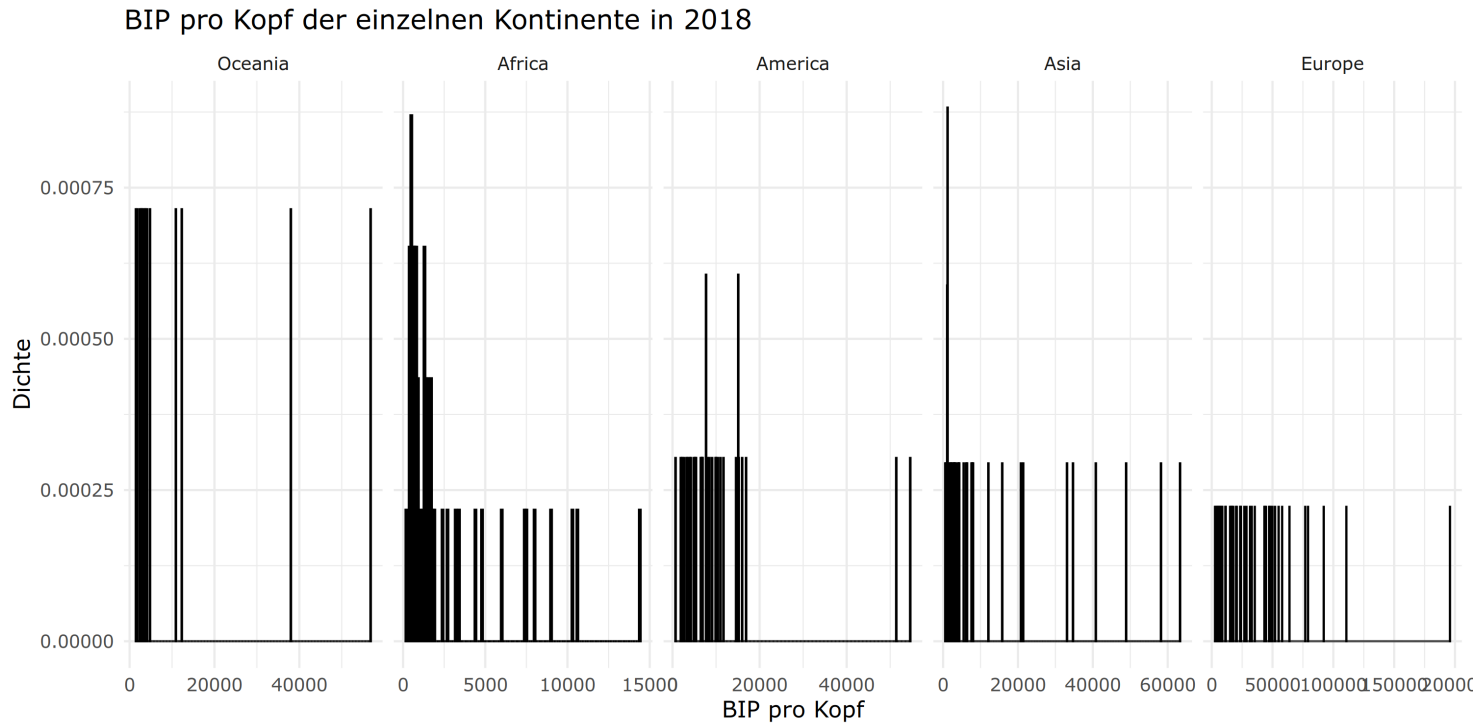
- Nun sehen Sie, dass die meisten Länder Afrikas deutlich ärmer sind als die Europas, aber das viele Länder Americas ein ähnliches BIP aufweisen wie die Asiens
- Weiterhin ist die Verteilung der Länder pro Kontinent recht ähnlich, mit Ausnahme von Europa

> **Überlgegung:** Ist es hier eventuell sinnvoller die komplette Verteilung anstatt einzelne Datenpunkte zu zeigen?

Zeigen Sie ihre Daten

- Nun sehen Sie, dass die meisten Länder Afrikas deutlich ärmer sind als die Europas, aber das viele Länder Americas ein ähnliches BIP aufweisen wie die Asiens
- Weiterhin ist die Verteilung der Länder pro Kontinent recht ähnlich, mit Ausnahme von Europa

Überlgegung: Ist es hier eventuell sinnvoller die komplette Verteilung anstatt einzelne Datenpunkte zu zeigen?



Verwenden Sie einheitliche Achsgrößen

- **Problem:** Hier wird nicht auf den ersten Blick ersichtlich, dass afrikanische Länder zum Großteil ein geringeres BIP haben als europäische Länder
- Dies liegt daran, dass die x-Achse in beiden Schaubildern unterschiedliche Zahlenbereiche umfasst

Daher ist es wichtig zu beachten:

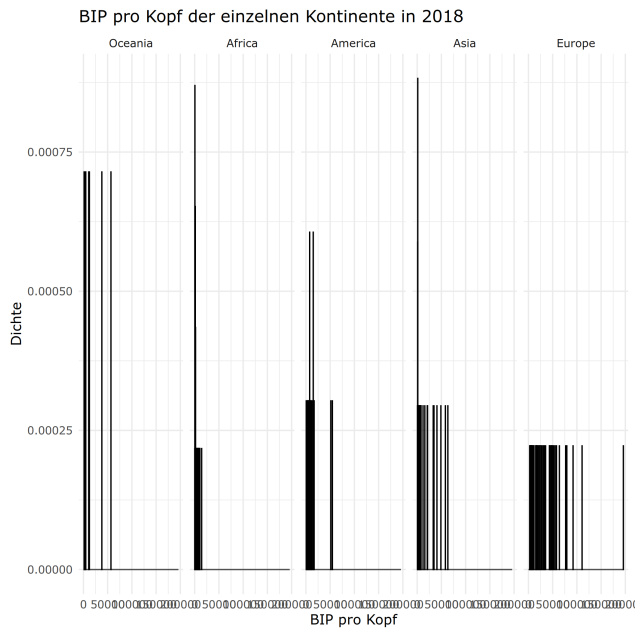
> **Verwenden Sie immer einheitliche Achsgrößen** wenn Sie Daten über mehrere Grafiken vergleichen

Verwenden Sie einheitliche Achsgrößen

- **Problem:** Hier wird nicht auf den ersten Blick ersichtlich, dass afrikanische Länder zum Großteil ein geringeres BIP haben als europäische Länder
- Dies liegt daran, dass die x-Achse in beiden Schaubildern unterschiedliche Zahlenbereiche umfasst

Daher ist es wichtig zu beachten:

> **Verwenden Sie immer einheitliche Achsgrößen** wenn Sie Daten über mehrere Grafiken vergleichen



Verwenden Sie einheitliche Achsgrößen

- Wenn Sie ihren Fokus auf vertikale Änderungen legen wollen, dann ordnen Sie ihre Grafiken horizontal zueinander an
- Wenn Sie ihren Fokus auf horizontale Änderungen legen wollen, dann ordnen Sie ihre Grafiken vertikal zueinander an
- In unserem Beispiel wollen Sie herausfinden, ob Kontinente und deren Länder sich im Hinblick auf deren BIP voneinander unterscheiden. Somit interessieren uns Veränderungen auf der x-Achse, d.h. horizontale Veränderungen

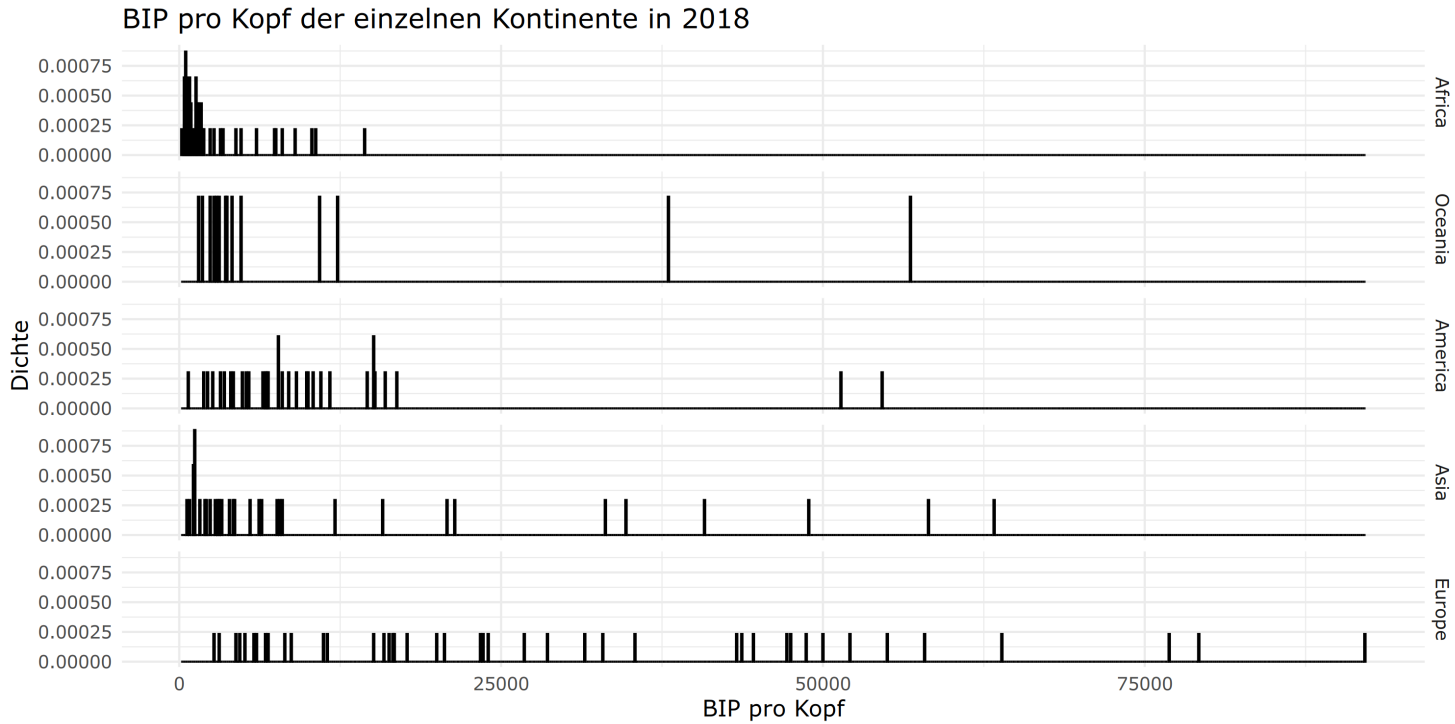
Verwenden Sie einheitliche Achsgrößen

- Wenn Sie ihren Fokus auf vertikale Änderungen legen wollen, dann ordnen Sie ihre Grafiken horizontal zueinander an
- Wenn Sie ihren Fokus auf horizontale Änderungen legen wollen, dann ordnen Sie ihre Grafiken vertikal zueinander an
- In unserem Beispiel wollen Sie herausfinden, ob Kontinente und deren Länder sich im Hinblick auf deren BIP voneinander unterscheiden. Somit interessieren uns Veränderungen auf der x-Achse, d.h. horizontale Veränderungen

> Ordnen Sie die Grafiken vertikal an

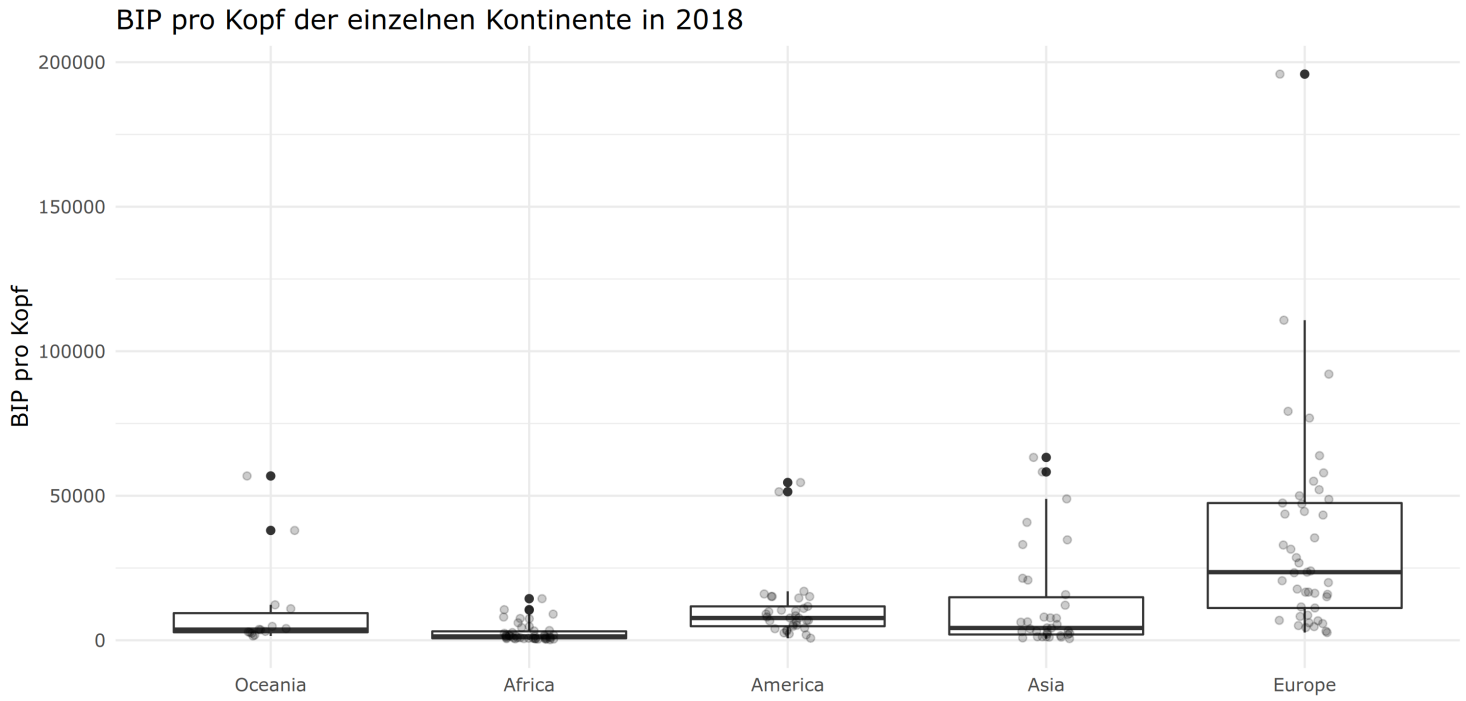
Verwenden Sie einheitliche Achsgrößen

Hier sehen Sie die unterschiede innerhalb der einzelnen Gruppen sehr schnell (wir klammern die Außreiser bei Europa aus und betrachten nur die Länder < $100 000 BIP pro Kopf:



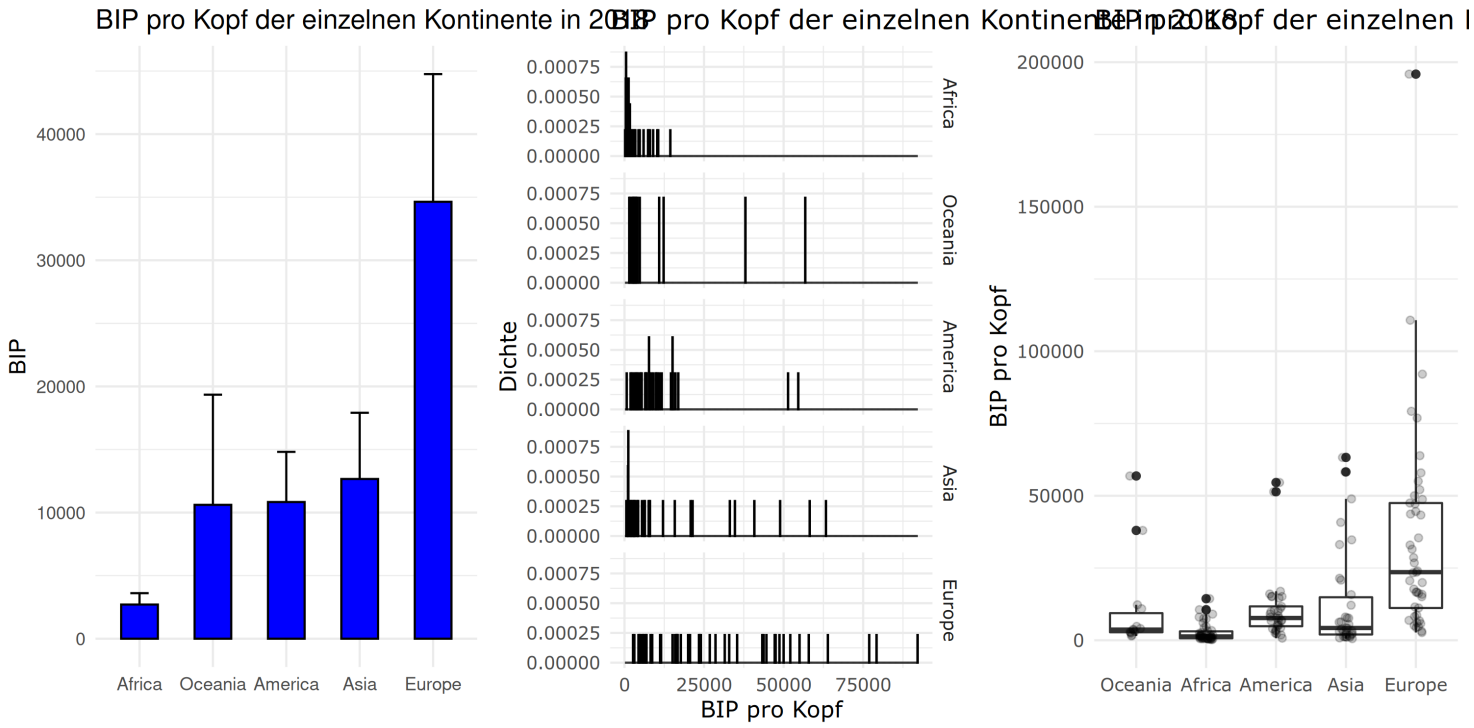
Verwenden Sie einheitliche Achsgrößen

- Eine weitere Möglichkeit die Information über das BIP kompakt darzustellen bieten Boxplots.
- Hier können Sie auch alle Datenpunkte zeigen um die Informationsdichte der Boxplots zu erhöhen:



Verwenden Sie einheitliche Achsgrößen

- Ein Vergleich unserer Darstellungsmethoden:

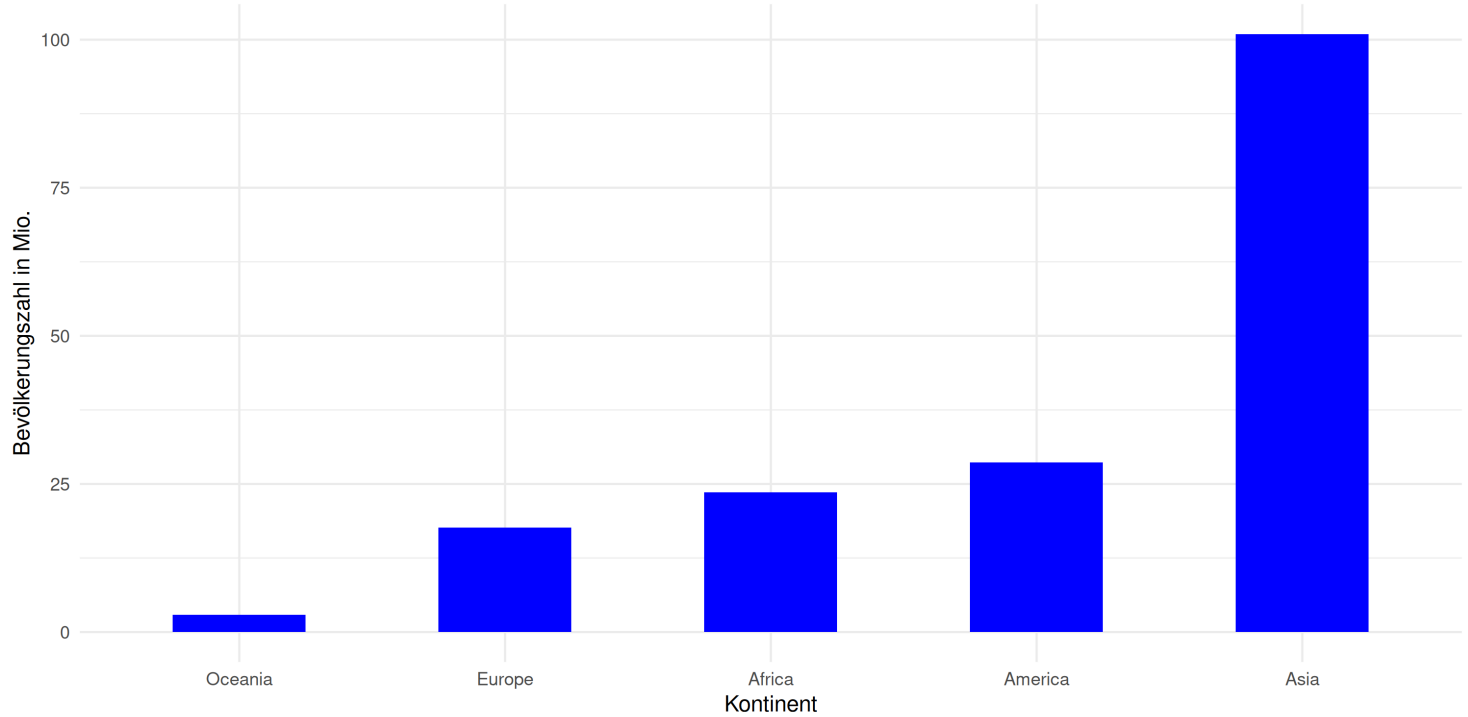


Transformieren Sie ihre Daten, falls nötig

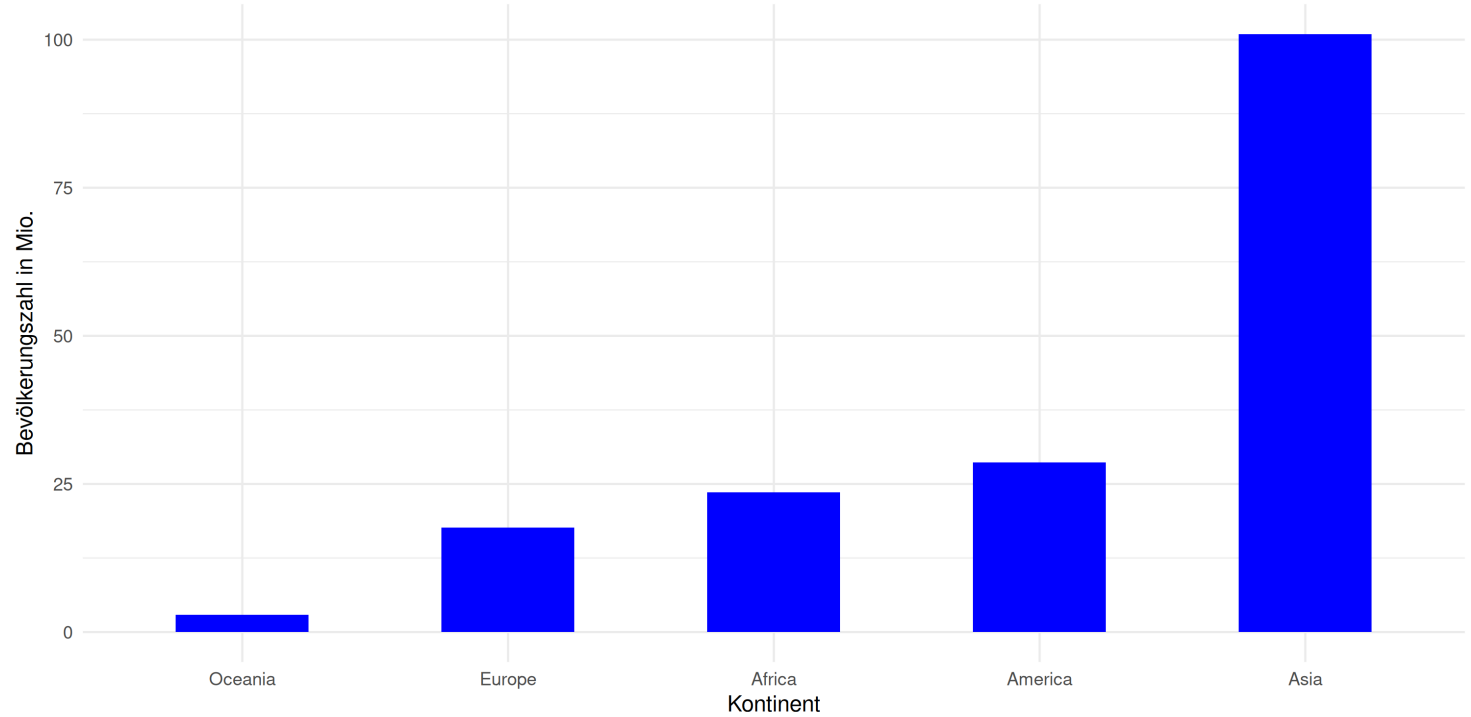
- Für Daten bei denen ein Wert deutlich über allen anderen Werten liegt sollten Sie sich Gedanken darüber machen, ob es nicht besser ist die Daten zu logarithmieren und dadurch vergleichbarer zu machen

> Visualisieren Sie die Bevölkerungszahl pro Kontinent (in 2018)

Transformieren Sie ihre Daten, falls nötig



Transformieren Sie ihre Daten, falls nötig

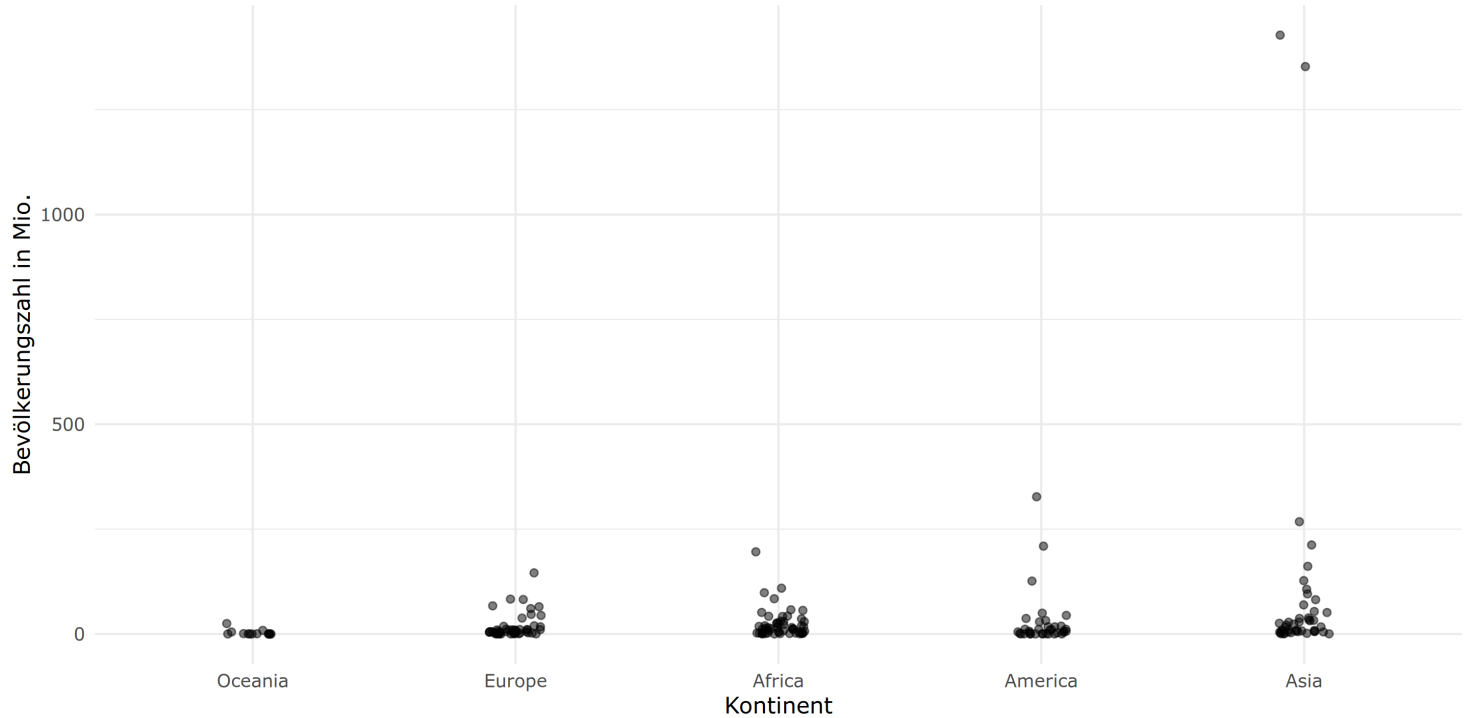


Schlussfolgerung: In asiatischen Länder wohnen im Durchschnitt deutlich mehr Menschen als in andere Teile der Welt

Transformieren Sie ihre Daten, falls nötig

Wenn Sie jedoch alle Datenpunkte darstellen, dann ergibt sich ein anderes Bild:

- Es gibt zwei Ausreiser in unseren Daten, d.h. zwei besonders bevölkerungsreiche Länder, speziell auf dem asiatischen Kontinent: vermutlich Indien und China



Transformieren Sie ihre Daten, falls nötig

> Logarithmieren Sie die Daten!

Transformieren Sie ihre Daten, falls nötig

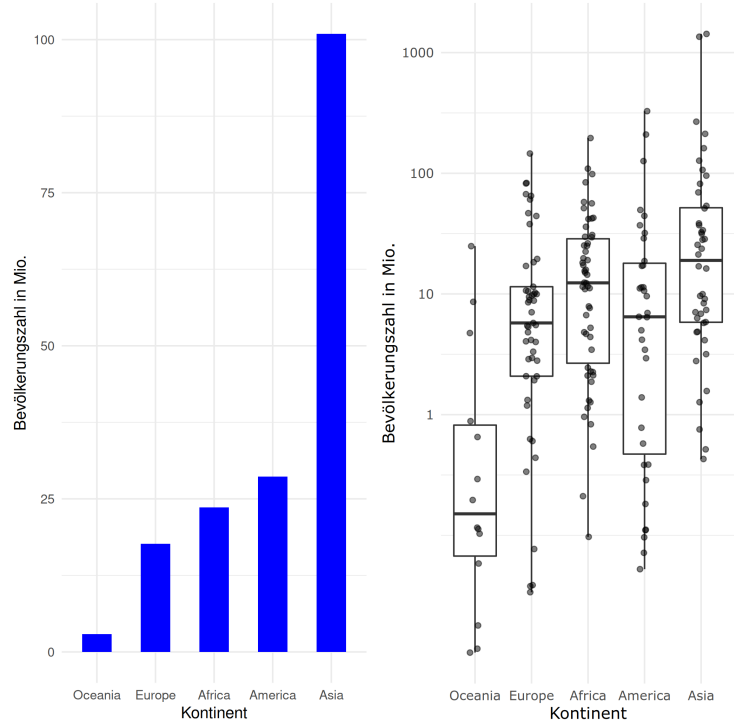
Logarithmieren Sie die Daten!

- Durch die Logarithmierung erhalten Sie eine viel bessere Einschätzung der tatsächlichen Bevölkerungszahlen auf den einzelnen Kontinenten
- Ein Vergleich der Grafiken macht dies besonders deutlich

Transformieren Sie ihre Daten, falls nötig

Logarithmieren Sie die Daten!

- ✚ Durch die Logarithmierung erhalten Sie eine viel bessere Einschätzung der tatsächlichen Bevölkerungszahlen auf den einzelnen Kontinenten
- ✚ Ein Vergleich der Grafiken macht dies besonders deutlich



Transformieren Sie ihre Daten, falls nötig

- Länder in Amerika haben **im Median** eine größere Bevölkerung als in Afrika, jedoch sind asiatische Länder am bevölkerungsreichsten

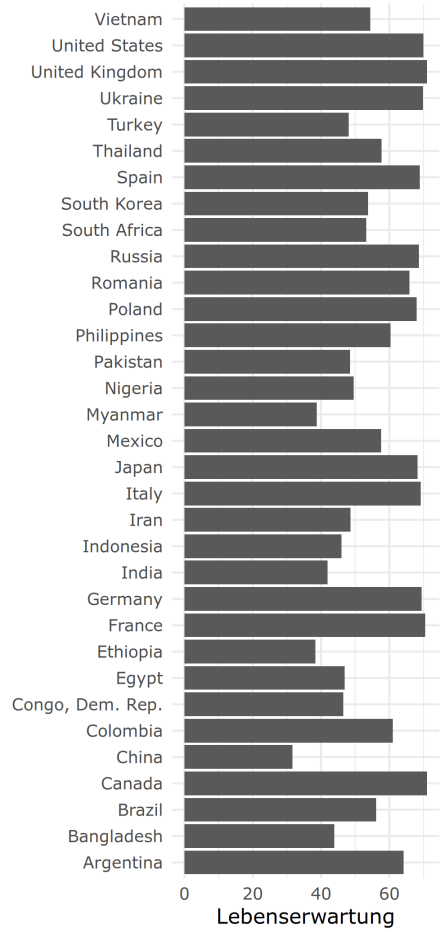
Die zwei Ausreiser China und Indien verzerren die erste Grafik!

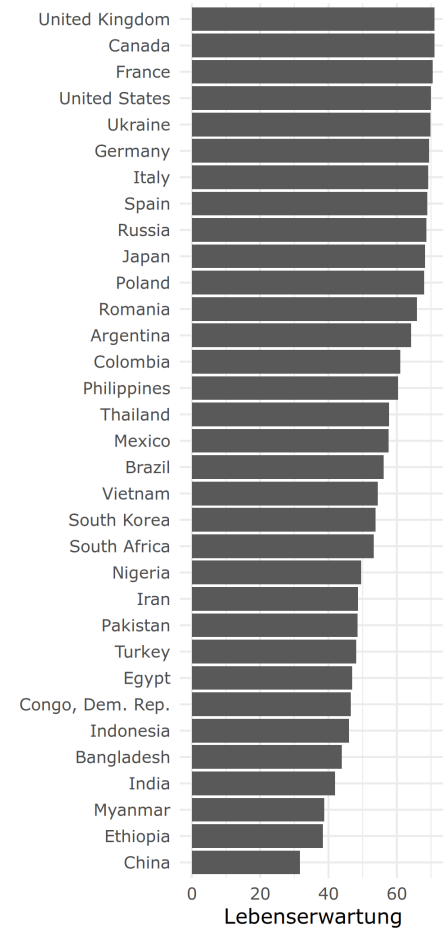
Sortierung des Outputs

- Wenn Sie in `ggplot` ein Balkendiagramm erstellen und mit Kategorien (z.B. Ländern) oder Faktorvariablen (z.B. Anzahl der Geschwister) arbeiten, dann sortiert `ggplot` immer nach diesen Kategorien oder Faktorvariablen
- Oft ist dies nicht gewünscht und Sie sollten nach geeigneteren Variablen für die Sortierung suchen
- Sie können die Funktion `reorder` verwenden um eine Sortierung nach ihren Wünschen zu erreichen

> Zeigen Sie die Lebenserwartung im Jahr 1960 für alle Länder mit mehr als 15 Mio. Einwohnern auf

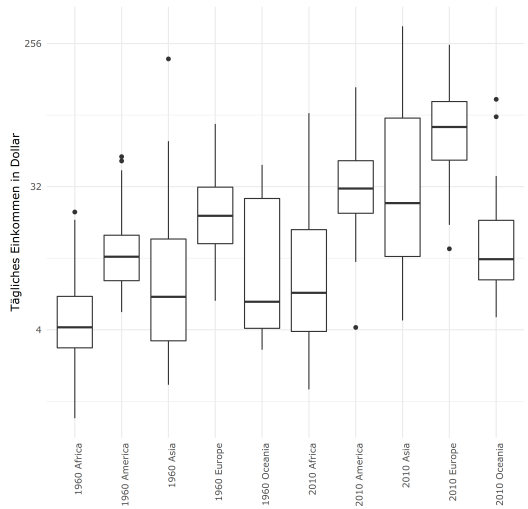
Sortierung des Outputs





Zeigen Sie Daten die zusammen gehören auch zusammen

- Gegeben Sie wollen das durchschnittliche tägliche Einkommen (Aufs Jahr mit 365 Tagen gerechnet) auf jedem Kontinent in 1960 dem Einkommen in 2010 gegenüberstellen
- `ggplot` sortiert hier per Default alphabetisch, wodurch alle Werte von 1960 *vor* allen Werten von 2010 kommen
  - Diese Darstellung erschwert den Vergleich zwischen den Gruppen

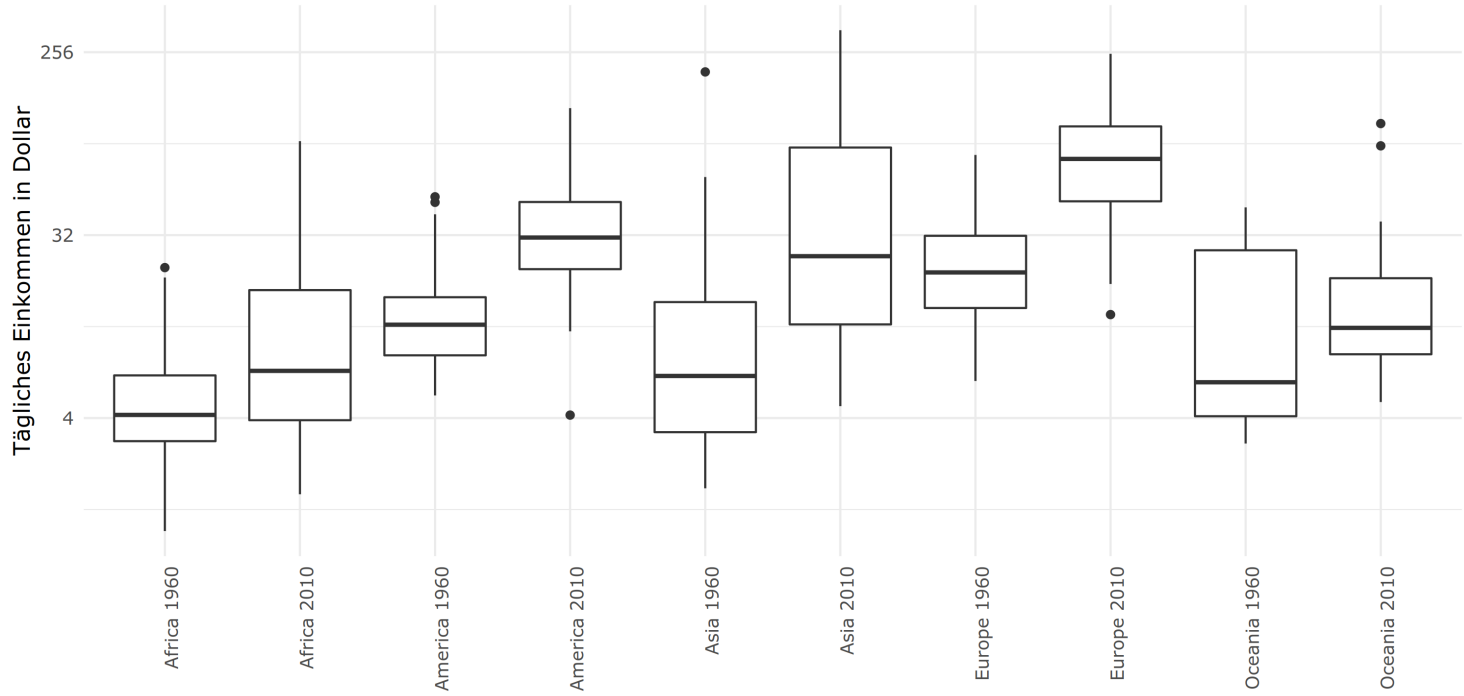


Zeigen Sie Daten die zusammen gehören auch zusammen

- ✚ Der Vergleich wird einfacher, wenn die relevanten Informationen nebeneinander gezeigt werden:

Zeigen Sie Daten die zusammen gehören auch zusammen

- Der Vergleich wird einfacher, wenn die relevanten Informationen nebeneinander gezeigt werden:

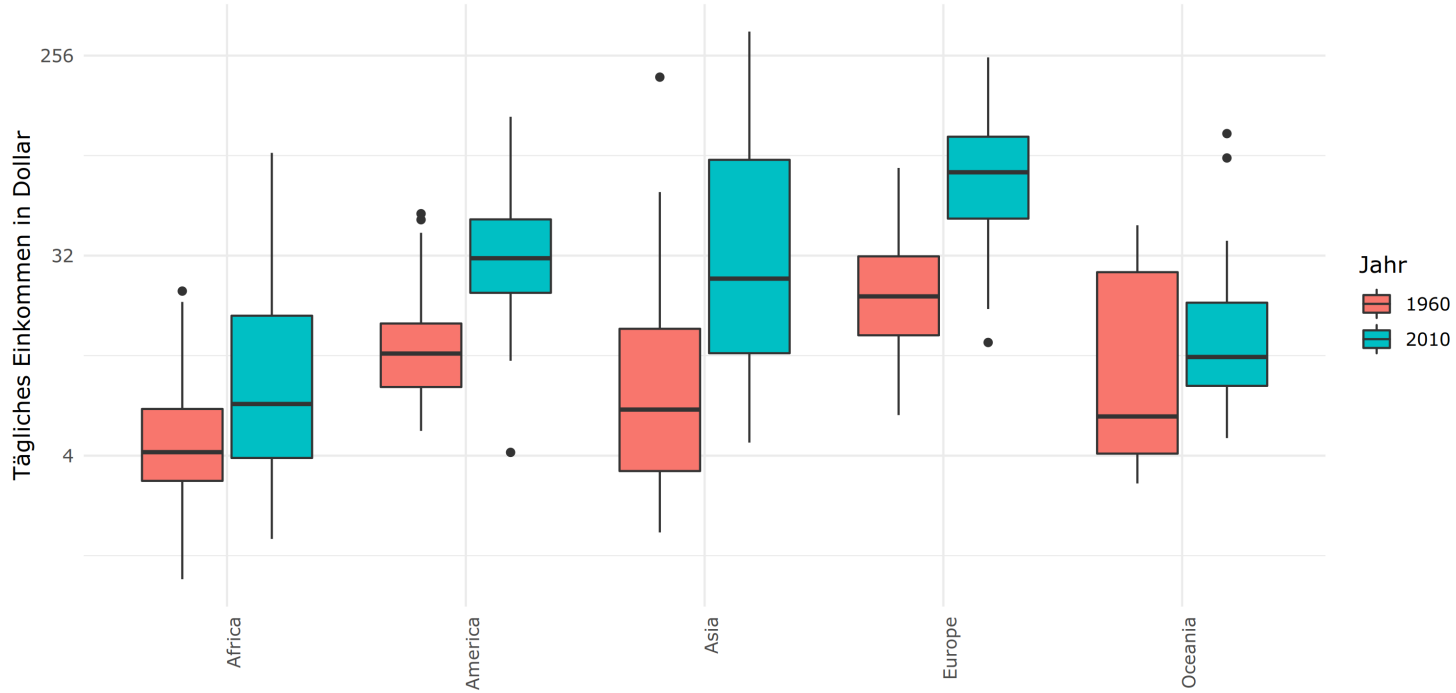


Die Verwendung von Farbe

- Der Vergleich wird noch einfacher, wenn wir verschiedene Farben für die einzelnen Jahre verwenden:

Die Verwendung von Farbe

- Der Vergleich wird noch einfacher, wenn wir verschiedene Farben für die einzelnen Jahre verwenden:

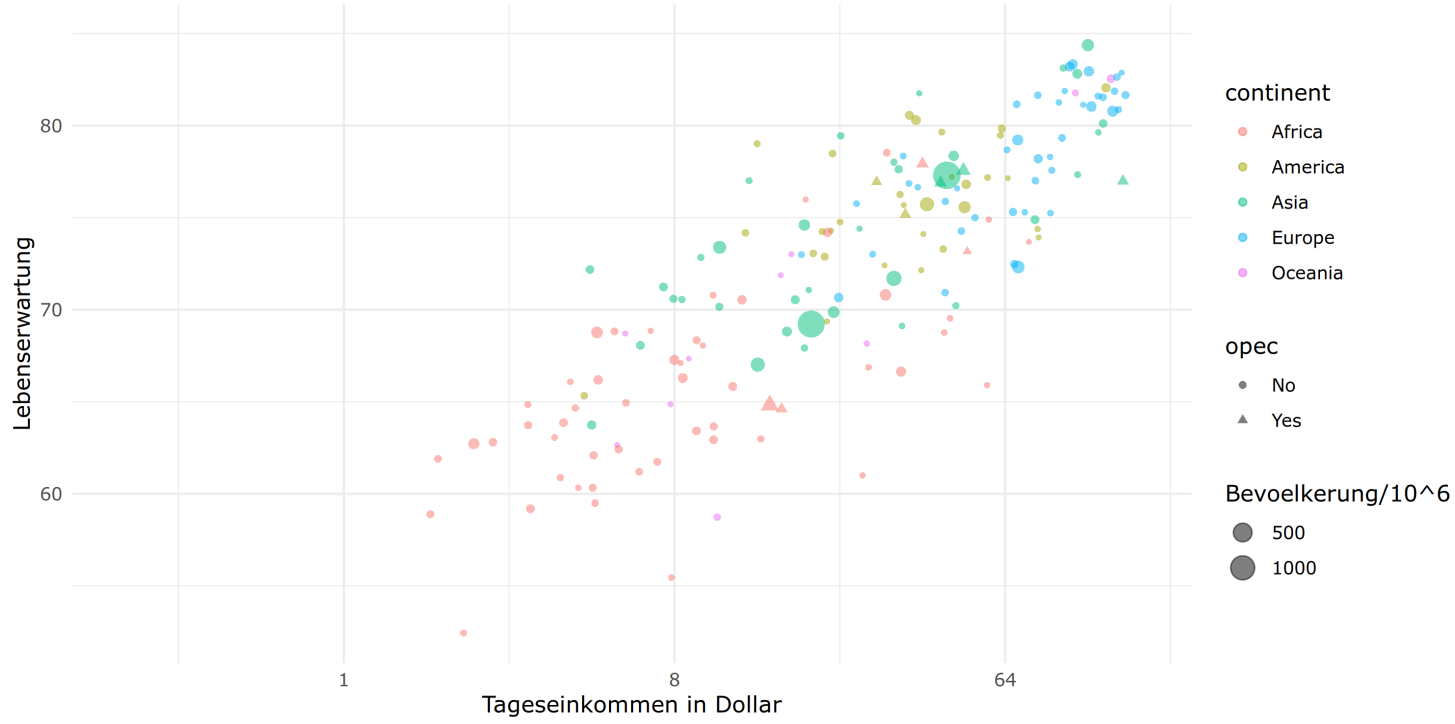


Eine dritte Variable aufnehmen

- Wenn Sie in ihrer Grafik mehr Informationen, d.h. eine dritte Variable, aufnehmen möchten können Sie dies bspw. folgendermaßen erreichen.
- Hier wird die Region, Bevoelkerung und OPEC Mitgliedschaft mehrerer Länder aus `gapminder` veranschaulicht

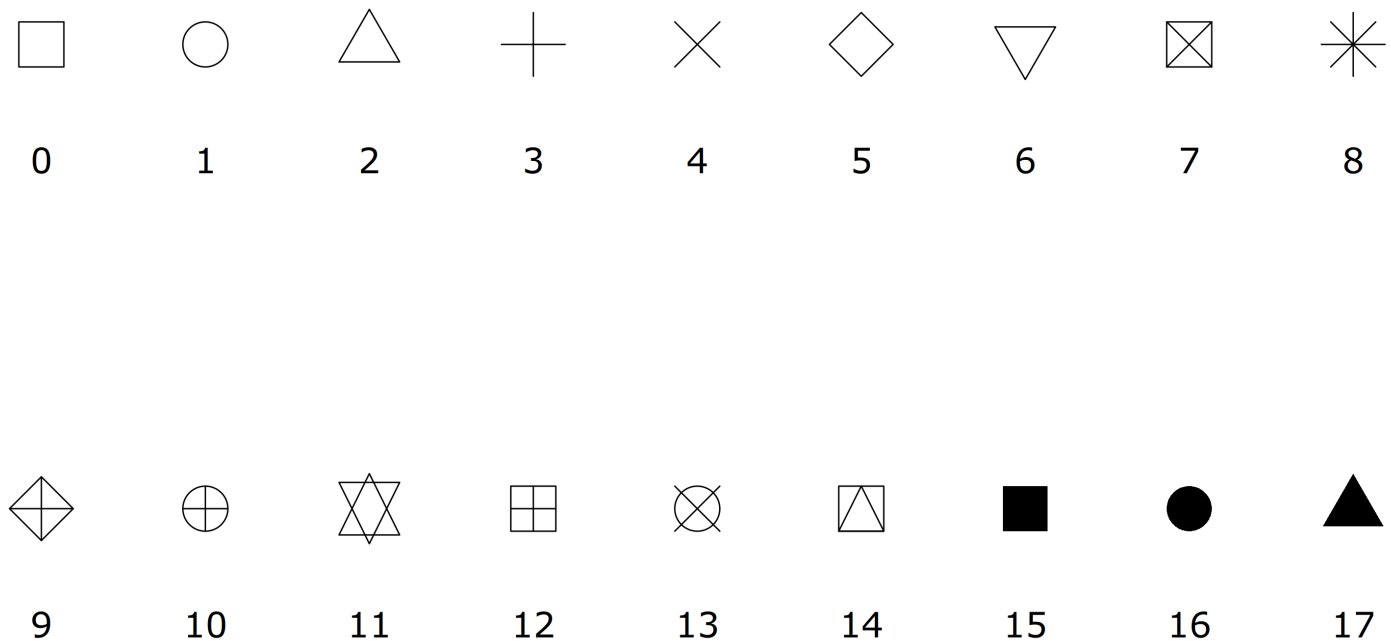
Eine dritte Variable aufnehmen

- Wenn Sie in ihrer Grafik mehr Informationen, d.h. eine dritte Variable, aufnehmen möchten können Sie dies bspw. folgendermaßen erreichen.
- Hier wird die Region, Bevoelkerung und OPEC Mitgliedschaft mehrerer Länder aus `gapminder` veranschaulicht



Eine dritte Variable aufnehmen

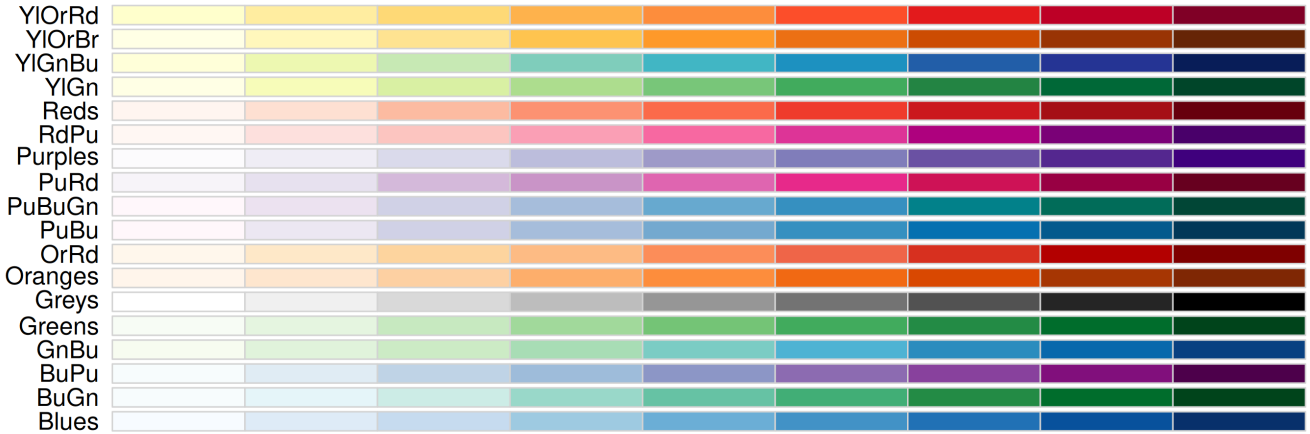
- Faktorvariablen können durch den Farbton und die Form dargestellt werden
- Sie können diese mit dem Paramter `shape` anpassen
- Hier die Formen, welche in R verfügbar sind:



Farbpaletten

Die Farbverläufe welche in R vorhanden sind:

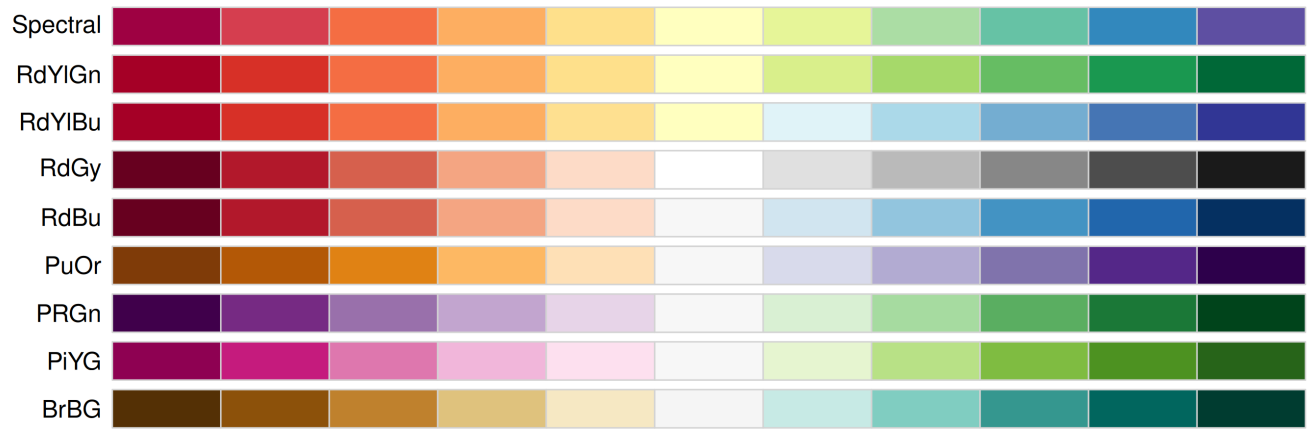
```
library(RColorBrewer)
display.brewer.all(type="seq")
```



Farbpaletten

- Divergente Farbverläufe werden verwendet, wenn Farben weg von einem Zentrum definiert werden sollen
- Dieser Farbverlauf legt auf beide Enden der Spanne gleichen Wert

```
library(RColorBrewer)
display.brewer.all(type="div")
```



Karten

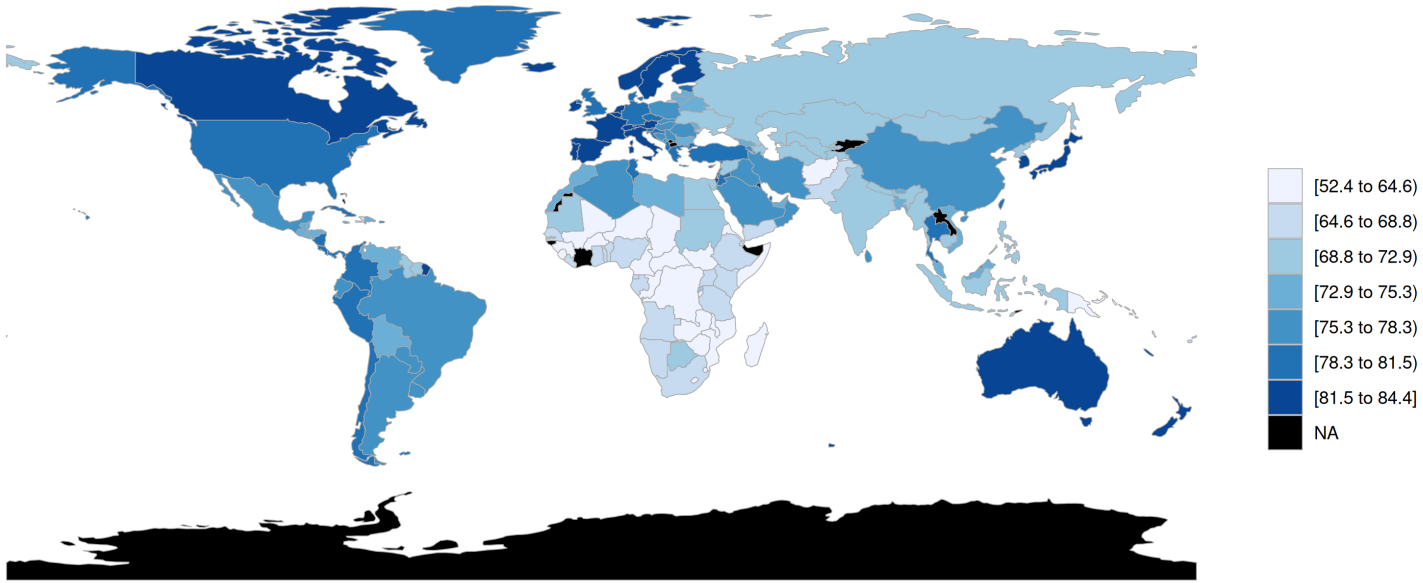
Eine weitere Möglichkeit Daten zu visualisieren, welche geografische Informationen beinhalten ist über Karten.

Wenn Sie sich die Lebenserwartung in 2018 für unterschiedliche Länder anschauen möchten, dann können Sie dies auch auf einer Karte effektiv darstellen:

Karten

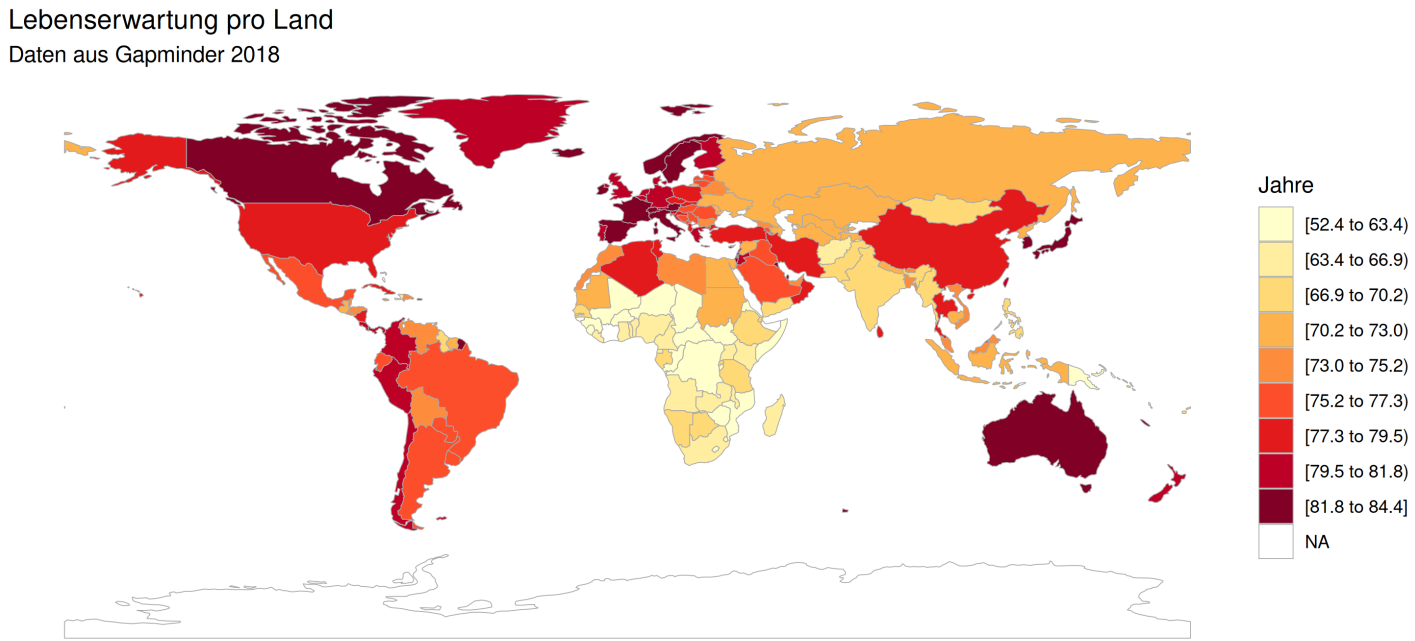
Eine weitere Möglichkeit Daten zu visualisieren, welche geografische Informationen beinhalten ist über Karten.

Wenn Sie sich die Lebenserwartung in 2018 für unterschiedliche Länder anschauen möchten, dann können Sie dies auch auf einer Karte effektiv darstellen:



Karten

Um die Aussage der Karte zu erhöhen können Sie eine andere Farbpalette wählen und die Grafik noch entsprechend beschriften:

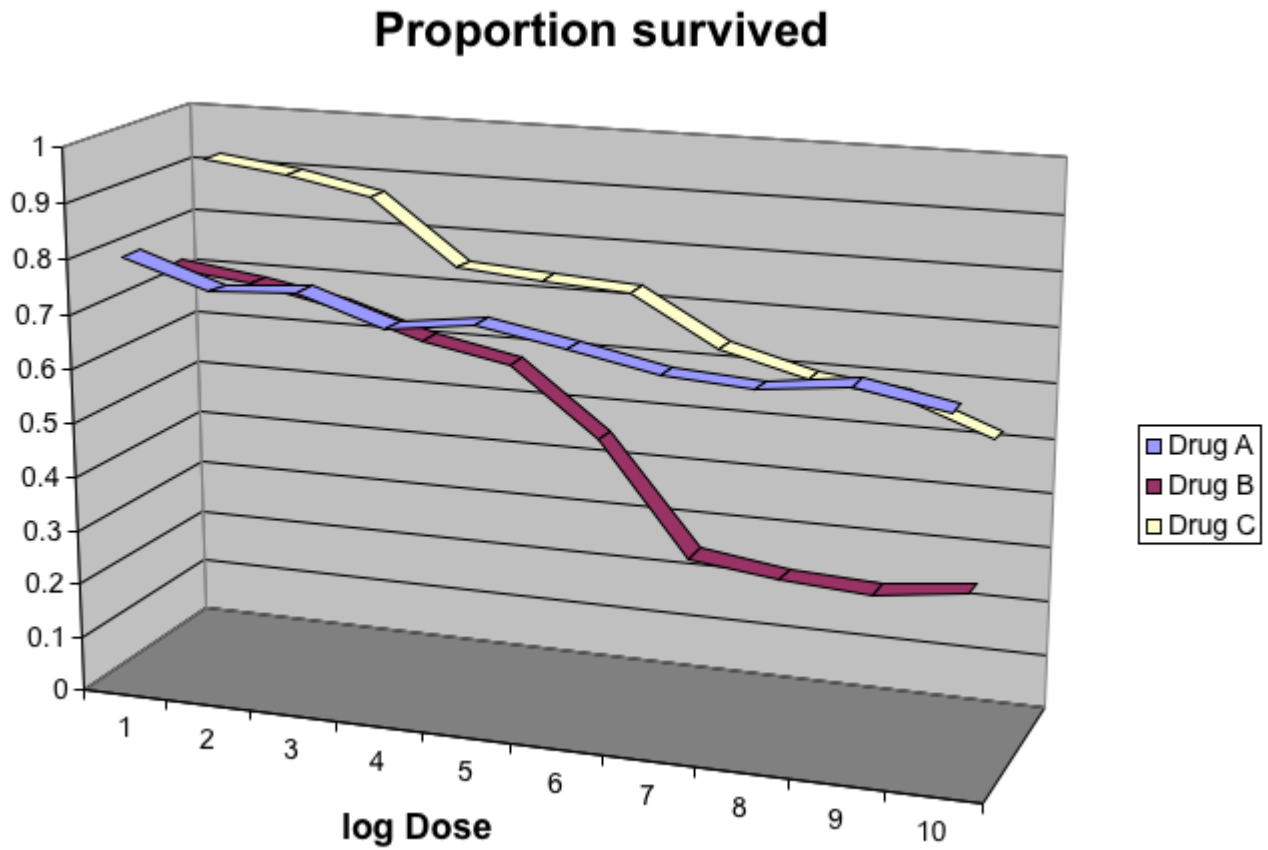


Quelle: https://www.gapminder.org,

Code angelehnt an: https://rkabacoff.github.io/datavis/GeoMaps.html

Abschreckende Beispiele

Vermeiden Sie pseudo 3D-Grafiken

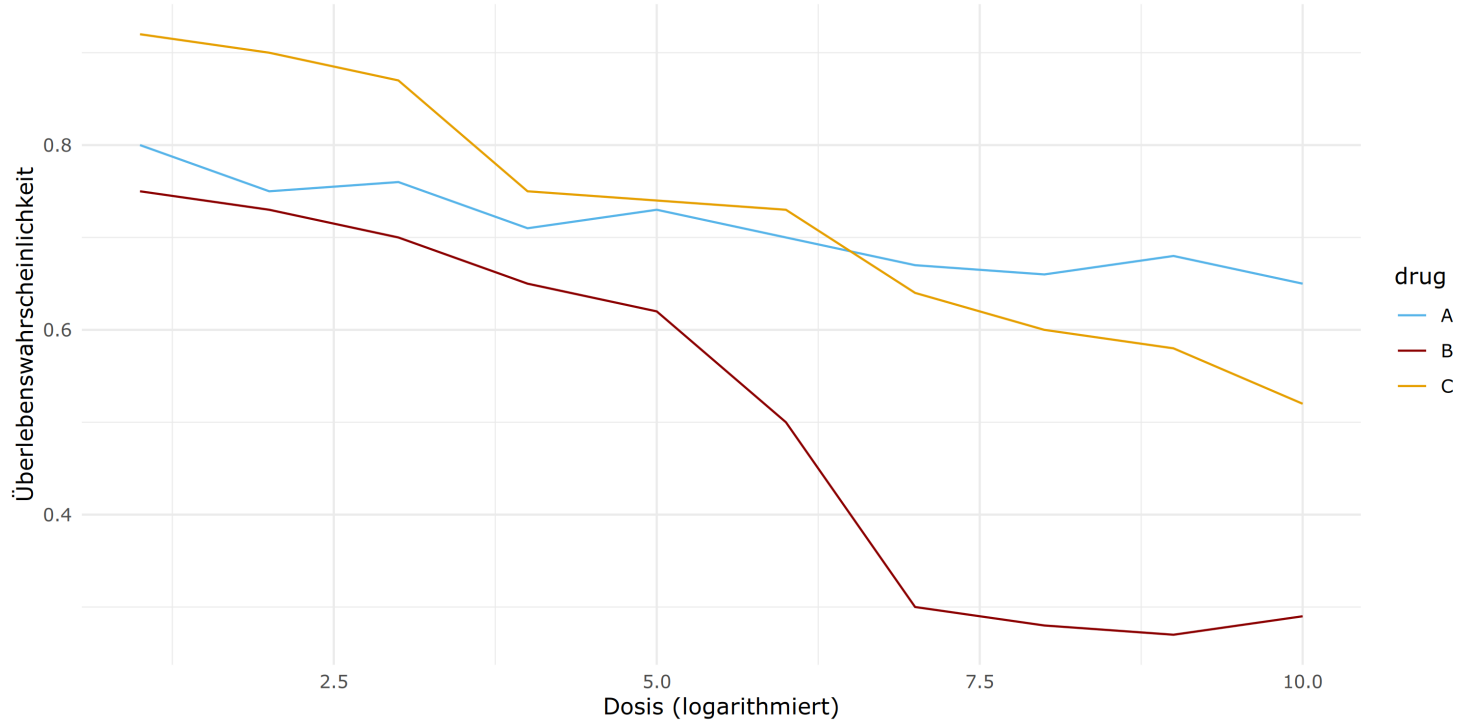


Vermeiden Sie pseudo 3D-Grafiken

- Die pseude 3D-Grafik auf der vorherigen Folie trägt drei unterschiedliche Dimensionen ab und wurde so in einem wissenschaftlichen Artikel veröffentlicht.
- Dargestellt wird hier: Die Höhe der verabreichten Dosis eines Medikaments, die Art des Medikaments und die Überlebenswahrscheinlichkeit der Patienten
- Die Grafik versucht einen Eindruck von Dreidimensionalität zu simulieren, doch dies will nicht so recht funktionieren!

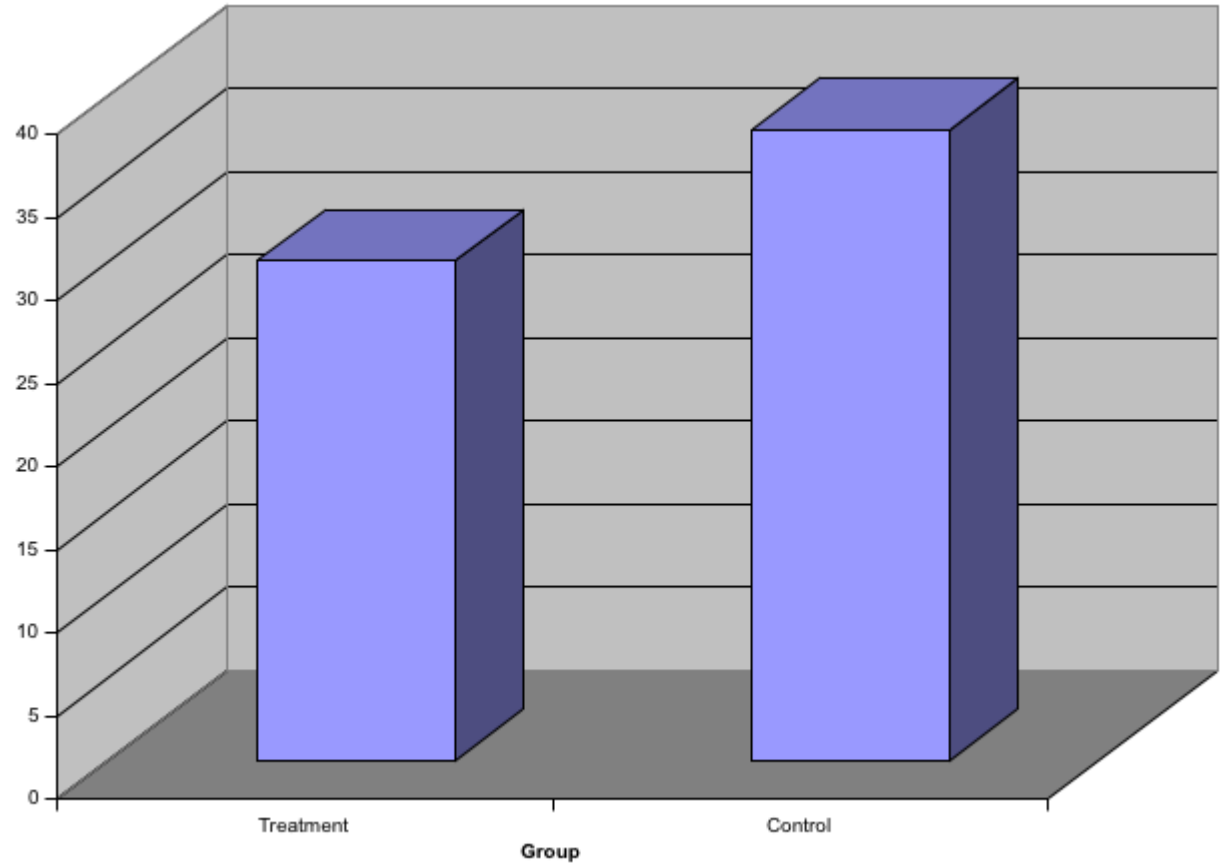
> Wo treffen sich die violette und rote Linie?

Wenn Sie die pseudo 3D-Grafik etwas umgestalten ist die Frage leichter zu beantworten:



Vermeiden Sie pseudo 3D-Grafiken

Grafiken bei welchen nicht ersichtlich ist, warum überhaupt auf 3D gesetzt wird, vewirren den Leser unnötig und sollten grundsätzlich vermieden werden:



Tabellen: Vermeiden Sie zu viele Nachkommastellen

Zu viele Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.435616
Germany	1980	76.673973
Germany	1990	85.717808
Germany	2000	100.72603
Germany	2010	110.76438

Tabellen: Vermeiden Sie zu viele Nachkommastellen

Zu viele Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.435616
Germany	1980	76.673973
Germany	1990	85.717808
Germany	2000	100.72603
Germany	2010	110.76438

Passende Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.44
Germany	1980	76.67
Germany	1990	85.72
Germany	2000	100.73
Germany	2010	110.76

Tabellen: Vermeiden Sie zu viele Nachkommastellen

Zu viele Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.435616
Germany	1980	76.673973
Germany	1990	85.717808
Germany	2000	100.72603
Germany	2010	110.76438

Passende Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.44
Germany	1980	76.67
Germany	1990	85.72
Germany	2000	100.73
Germany	2010	110.76

- ✚ R gibt ihnen standardmäßig immer sieben Nachkommastellen aus (falls diese vorhanden sind)
- ✚ Diese Genauigkeit ist unnötig und verunstaltet ihre Tabellen. Dadurch *erschweren* Sie es dem Leser die wichtigsten Informationen schnell aufzunehmen

Tabellen: Vermeiden Sie zu viele Nachkommastellen

Zu viele Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.435616
Germany	1980	76.673973
Germany	1990	85.717808
Germany	2000	100.72603
Germany	2010	110.76438

Passende Nachkommastellen:

country	Jahr	einkommen
Germany	1970	59.44
Germany	1980	76.67
Germany	1990	85.72
Germany	2000	100.73
Germany	2010	110.76

- R gibt ihnen standardmäßig immer sieben Nachkommastellen aus (falls diese vorhanden sind)
- Diese Genauigkeit ist unnötig und verunstaltet ihre Tabellen. Dadurch *erschweren* Sie es dem Leser die wichtigsten Informationen schnell aufzunehmen
- Die Tabelle trägt das Tageseinkommen in Deutschland zu verschiedenen Zeitpunkten ab
- Wir können keine 0,005 Cent verdienen, deshalb ist es hier nicht wichtig und verwirrt den Leser nur
- Zwei Nachkommastellen sind hier mehr als genug und zeigt gut auf, dass das Einkommensniveau steigt

Zielgruppe definieren

Sie können Grafiken für

- ✚ sich selbst erstellen um ein Gefühl für die Daten zu bekommen (explorativ)
- ✚ Experten erstellen, um ihre Analyse/Ergebnisse zu verdeutlichen
- ✚ eine größere Gruppe erstellen, um einen allgemeinen Sachverhalt darzustellen

> Überlegen Sie wer ihre Zielgruppe und designen Sie ihre Grafik so, dass sie von dieser Zielgruppe verstanden wird!