Ansichtssache

Welches Diagramm passt zu meinen Daten?

Ein Diagramm ist schnell gebaut – wenn man erst mal die richtige Darstellungsform gefunden und die Daten passend umgewandelt hat. Unser kleiner Leitfaden hilft bei der Orientierung.

Von Andrea Trinkwalder

Ein Diagramm, in dem man von oben auf Balken schaut, lässt Werte klein erscheinen, von unten wirken sie dagegen groß: Wie man Daten aufbereitet, welchen Diagrammtyp man wählt, sogar die Farben und Formen beeinflussen, wie ein Betrachter die zugrunde liegenden Daten interpretiert. Bevor man eine Grafik baut, muss man sich eingehend mit dem Rohmaterial beschäftigen – und gegebenenfalls auch die Datenquellen kritisch hinterfragen: Welche Erkenntnisse kann man aus den Daten gewinnen? Welche davon sind wichtig und mit welcher Art der Darstellung kann man diese eindrucksvoll vermitteln?

Diagramm-Typen

Die wohl universellste Darstellungsform sind Säulen und Balken, die sich für die unterschiedlichsten Datentypen eignen. **Säulendiagramme** wirken am besten, wenn man nur eine Handvoll Kategorien oder überschaubare Zeitreihen visualisieren möchte, etwa die jährlichen Verkaufszahlen von Tablets, Smartphones, Notebooks und PCs während der letzten fünf Jahre. Kommen zu viele Kategorien ins Spiel, rücken die Säulen immer enger zusammen. Dann fehlt Platz für die Beschriftung – und eine vertikale Ausrichtung erschwert das Lesen.

Hier schlägt die Stunde des **Balkendiagramms**, das leichter zu beschriften ist, da selbst längerer Text links oder rechts neben den Balken genügend Platz findet. Diese Anordnung unterstützt zudem die natürliche Leserichtung. Setzt man gestalterische Mittel gekonnt ein, wirken auch Balken und Säulen alles andere als langweilig: Man denke etwa an die einprägsame Form der Bevölkerungspyramide.

Säulen und Balken gibt es auch in **gestapelter Form**, die häufig dann sinnvoll ist, wenn noch eine weitere Dimension hinzukommt. Jedes Jahr im oben genannten Beispiel lässt sich auch als Stapel anstatt als Säulengruppe darstellen – wodurch die schrumpfenden PC-Anteile und der Trend zu Smartphones und Tablets deutlicher werden.

Kreisdiagramme haben sich etabliert, um den Anteil jeder Kategorie am Ganzen zu zeigen. Je mehr Kategorien ins Spiel kommen und je kleiner die Segmente werden, umso schwieriger wird die Interpretation. Bei mehr als fünf Kategorien ist das Balkendiagramm die bessere Wahl. **Donuts** wiederum sind leichter zu erfassen als Torten, weil ein Vergleich der außen liegenden Bögen leichter fällt als zwischen den teils schmalen kompletten Kuchenstücken. Zudem lässt sich im Inneren platzsparend die Überschrift oder eine markante Aussage unterbringen.

Für zeitliche Verläufe mit vielen Messwerten gibt es **Linien- und Flächendiagramme**, wobei die x-Achse üblicherweise als Zeitleiste fungiert.

Ob es einen Zusammenhang zwischen zwei numerischen Variablen gibt, sie also korrelieren, illustriert das **Streudiagramm**, etwa zwischen Alter und Gewicht oder Lebenserwartung und durchschnittlichem Einkommen in den Ländern der Welt. Mit **Blasen** lässt sich eine weitere numerische Größe – beispielsweise die Einwohnerzahl des jeweiligen Landes – über die Kreisfläche visualisieren, eine kategorische Unterscheidung (etwa nach Kontinent) kann man über deren Färbung anzeigen.

Kunst-Graph

Der **Streamgraph** ist der ästhetisch ansprechendere Bruder von gestapelten Säulen- und Flächendiagrammen. Er vermittelt über einen längeren Zeitraum auftretende Trends und Muster eindrücklich. Fürs exakte Ablesen von Werten und zu viele Kategorien eignet sich die Flussdarstellung hingegen nicht.

Hierarchisch geordnete Daten lassen sich als **Baum**, **Sunburst** oder **Baumkarte** visualisieren. Die beiden letzten reflektieren durch ihre Flächen auch Mengenverhältnisse, etwa in Tierpopulationen. Die kompakten, rechteckigen Baumkarten

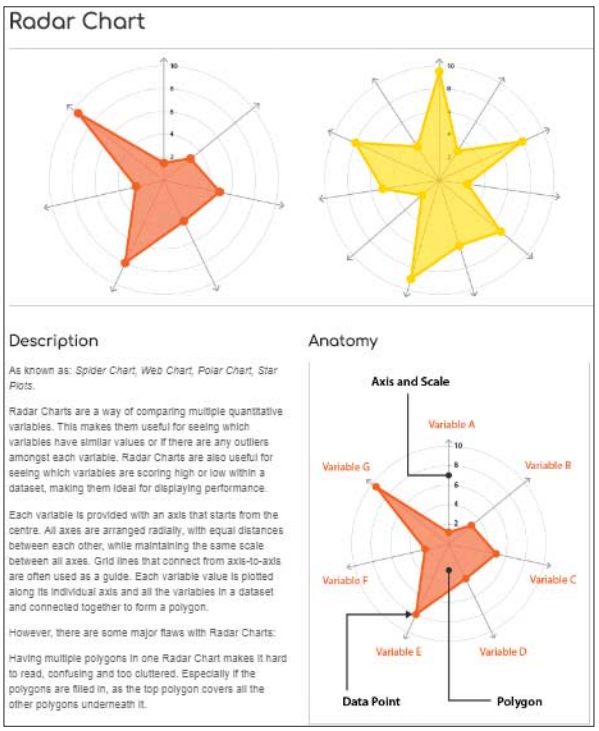


Bild: datavizcatalogue.com

Der Dataviz Catalogue erklärt jeden erdenklichen Diagrammtyp detailliert – auch sortiert nach Anwendungsgebiet.

Wide Format

Proband	Gewicht1	Gewicht2	Gewicht3
Martina	60.4	58.3	62.0
Michael	80.5	75.9	77.2
Thomas	95.3	92.0	94.1

Long Format

Proband	Zeitpunkt	Gewicht
Martina	1	60.4
Martina	2	58.3
Martina	3	62.0
Michael	1	80.5
Michael	2	75.9
Michael	3	77.2
Thomas	1	95.3
Thomas	2	92.0
Thomas	3	94.1

Im Long Format entspricht jeder Datenpunkt einer Zeile, das Wide Format fällt kompakter aus.

vermitteln die Proportionen besser, während der raumgreifende Sunburst die Hierarchien besser wiederspiegelt.

Alluvial- oder **Sankey-Diagramme** kommen eigentlich aus der technisch-naturwissenschaftlichen Ecke und werden genutzt, um Bewegungen innerhalb eines abgeschlossenen Systems darzustellen, etwa Energie- und Geldflüsse. Je breiter der Fluss, umso größer die Menge, die sich von A nach B bewegt. In der politischen Berichterstattung verdeutlichen sie Wählerwanderungen.

Mit einem sehr eindrucksvollen Sankey-Diagramm hat der Bauingenieur Charles Joseph Minard 1869 die Verluste der französischen Armee während Napoleons Russlandfeldzug verdeutlicht.

Choroplethen- und **Symbolkarten** vermitteln rasch einen Eindruck von den Zuständen auf der Welt, etwa wenn es um die Wahlgewinner in den einzelnen Bundesländern oder die Verfügbarkeit von Breitbandanschlüssen geht. Wichtig ist, die Daten zuvor zu normalisieren und nicht immer wieder gefärbte Karten als Darstellungsform für Daten mit geografischem Bezug zu benutzen: Exakte Größenverhältnisse lassen sich mit Säulen und Balken besser vermitteln – oder einer Kombination aus Karten und Balken.

Einen Überblick sämtlicher Diagrammtypen und ihres Verwendungszwecks gibt der Data Visualization Catalogue, siehe ct.de/y1vf. Das Chart-Kompendium lässt sich nach Namen und gewünschter Funktion sortieren.

Datenformate

Wie man Daten professionell aggregiert, aufbereitet und analysiert, füllt ganze Bücher – das lässt sich in diesem Rahmen nicht erschöpfend erklären. Für erste Experimente und zum Vergleichen der Apps präpariert man am besten eine Tabelle mit wenigen Spalten und Messwerten, die sich zu einem Balken- oder Tortendiagramm verarbeiten lässt. Oder man verwendet einfach die Beispieldatensätze, die nahezu jeder App-Anbieter zur Verfügung stellt.

Der kleinste gemeinsame Nenner beim Datenimport sind kommaseparierte Textdateien (.csv), die meisten Anwendungen akzeptieren auch Excel-Tabellen im Format .xls oder .xslx oder Google-Tabellen. Einige zapfen Datenbanken an oder analysieren Online-Aktivitäten über das API diverser Dienste wie Twitter und Facebook in Echtzeit. Als Konvention hat sich durchgesetzt, die erste Zeile als Benennung der Variablen zu interpretieren und die Zeilen darunter als Messwerte.

Beim Einlesen kommaseparierter Files sind manche Apps arg wählerisch. Manche erwarten tatsächlich ein Komma und akzeptieren kein Semikolon als Trennzeichen – oder umgekehrt. Das lässt sich per Suchen-Ersetzen leicht beheben (sofern sich diese „reservierten" Zeichen nicht auch in den Messwerten verstecken). Dann muss man eventuell zu schlau gewählten Suchstrings greifen.

Eine weitere kleine Hürde ist, dass zwei grundsätzlich unterschiedliche Darstellungsformen für Datensätze existieren: das Wide Format und das Long Format. Ersteres ist von Menschen leichter zu verstehen, wird aber von manchen statistischen Verfahren nicht unterstützt. Beim Long Format besitzt jeder Messwert eine eigene Zeile, das Wide Format verteilt Messwerte auf mehrere Spalten.

Beispiel: Bei drei Probanden wird zu drei Zeitpunkten das Gewicht gemessen. Das Wide Format besteht aus den Spalten Proband-Gewicht1-Gewicht2-Gewicht3, das Long Format baut sich nach dem Schema Proband-Zeitpunkt-Gewicht auf. Das erste kommt pro Proband mit einer Zeile aus, bei Letzterem belegt jeder Proband drei Zeilen.

Charticulator, RawGraphs und Tableau erwarten das Long Format, die anderen im Artikel auf Seite 72 gezeigten Apps das Wide Format; die beiden Letztgenannten können das Wide Format praktischerweise selbst konvertieren. Auf dem Desktop gelingt die Transformation mit Statistikprogrammen oder Pivot-Tabellen in einer Tabellenkalkulation.

JSON und GeoJSON

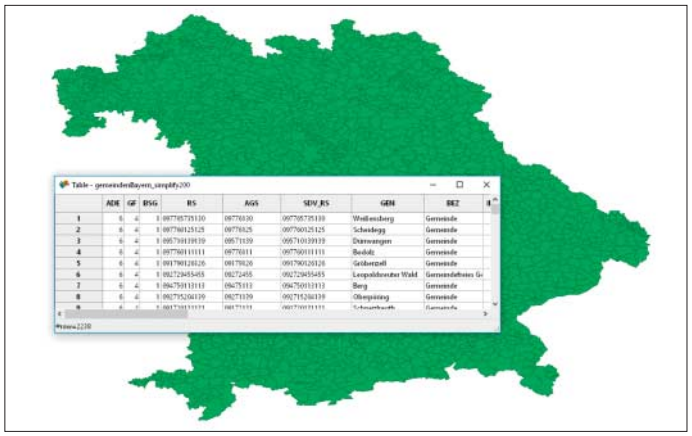
JSON (Javascript Object Notation) ist ein standardisiertes Format, das vor allem hierarchische Daten besser speichert als CSV – und auch komplette Visualisierungsprojekte wie etwa die von RawGraphs und Highcharts.

Auf geografische Daten ist GeoJSON spezialisiert, mit dem sich Regionen als Polygone beschreiben lassen. Die Open-Source-Software GeoDa hat keine eigenen Karten hinterlegt, weshalb sie GeoJSON als Input benötigt. Datawrapper und Tableau bringen eigenes Kartenmaterial mit, weshalb hier die Angabe von Ländernamen oder Geocodes – DE, AT und so weiter – genügt, um die Regionen anhand der Messwerte einzufärben oder mit Symbolen zu versehen.

Eine sehr reichhaltige **Quelle für hiesige Geodaten** bis hinunter auf Kreisebene ist das Open Data Lab. Hier lassen sich benötigte Flächen interaktiv auswählen und als GeoJSON-Datei herunterladen. **Weltweites Kartenmaterial** steht auf den OpenStreetMap-Seiten zum Download.

(atr@ct.de) c't

Datenquellen und Utilities: ***ct.de/y1vf***



Bei GeoDa muss man das Kartenmaterial importieren. Das kann man sich bei Open Streetmap und Open Data Lab besorgen.