Details zur Visualisierung von Daten

Details zur Grammer of Graphics

Andreas Reschreiter

24.11.2020

Prinzipien für Graphen - Grammer of Graphics (GG)

Die “Grammatik” für Grafiken formalisiert die Grafikdarstellung

- für eine gewisse Einheitlichkeit
- basiert auf wissenschaftlichen Analysen

Visual Cue

Kann ein

- Datenpunkt
- eine Farbe
- eine Linie
- ein Winkel
- etc.

sein. Zum Beispiel ein einzelner Punkt im Scatter Plot.

Types of Visual Cues nach Vorteilhaftigkeit

- **Position**: Punkte sind intutitiv bei numerischen Werten und zwei Dimensionen. Bei nur einer Dimension, liegen die Punkte auf einer Achse, je dichter die Werte beisamen sind, desto näher liegen die Punkte auf einer Achse zusammen.
- **Länge**: Höhe eines Balkens (nur die Höhe zählt, nicht die Balkenfläche).
- **Winkel**: Relativ gut erkennbar, ob ein Winkel parallel ist oder nicht. Die Winkelgröße ist sehr schlecht abschätzbar. Die länge ist viel besser erkennbar, z.B. eine unterschiedliche Balkenhöhe wird sofort erkannt. Unterschiedliche Winkel sind nur schwer erkennbar (ob größer oder kleiner) Menschen können Winkel schlecht abschätzen, daher sollten Zahlen nicht mit Winkel darstellen. Pie-Chart visualisieren Anteile, die Fläche ist proportional zum Anteil und der Winkel ist entscheidend für die Fläche. Pie Charts sind nur sehr schwer abschätzbar, daher wichtig die Werte anzugeben. In einem Balkendiagramn sind die unterschiedlichen Werte zueinander viel besser abschätzbar.

Types of Visual Cues nach Vorteilhaftigkeit

- **Richtung**: Spezialfall eines Winkels, z.B. eine Regressionslinie kann rauf oder runter gehen, kann man gut unterscheiden.
- **Form (Shape)**: Symbole wie Punkt, Stern etc. in einem Scatterplot.
- **Fläche**: Flächen sind sehr schwer abschätzbar, weil zwei Dimensionen und führt oft zu falschen Vorstellungen. Wenn die Größe des Symbols die Größe einer dritten Kenngroße angibt, z.B. bei mehr Umsatz ist der Punkt ein größerer Kreis in in einem Scatter Plot. Die Flächen sind schwer zu unterscheiden, um wieviel mal so groß die Flächen sind. Flächen zu vergleichen ist nicht einfach und funktioniert nur bei wenigen Datenpunkten (Flächen).
- **Volumen**: Noch schwieriger - 3D-Balkendiagrammme. Nur die Höhe ist relevant, die 3. Dimension lenkt nur ab, daher sind 3D-Balkendiagramme eher sinnlos, ablenkend und eventuell verwirrend.
- **Schattierung**: Zum Beispiel engere Strichlierung vs. weitere Strichlierung.

Types of Visual Cues nach Vorteilhaftigkeit

- **Farben**: Oft verwendet (sehr beliebt) haben aber viele Probleme. Problem bei Farben, dass extrem ablenken können. Gewisse Farben stechen (subjektiv) stark hervor, z.B. violett. Farben können sehr grell sein (außer schwarz). Wenn objektiv sein will, dann sollte (k)eine Farbe - für ein kategoriales Merkmal - sollte nicht hervorstechen. Wenn die Graustufen gleich sind, dann gleich hell/grell, dann haben auf einem schwarz/weiss Drucker alle Farben dasselbe Grau auf dem Ausdruck.
- `RColorBrewer` für eine gute Farbwahl - weil jede Abstufung um gleichen Faktor intensiver. Qualitative Paletten, sodass die Nachbarn sich möglichst stark unterscheiden. Standardfarben in `ggplot` ist gut gewählt, in Standard `R` sind die Farben schlecht gewählt.

Coordinate System

Punkt alleine ist sinnlos, muss verorten, brauche daher ein Koordinaten System.

Kontext und erweiterte Grafiken

- **Context** Titel & Achsenbeschreibung: Beschriftung nötig, weil ohne Kontext das Diagram nicht verständlich ist. Auch eine Achsenbeschriftung ist nötig.
- **Skala** Skala ist die Abbildung zwischen Visual Cue und Werten, z.B. ein Punkt im Scatter auf die x und y Achse. Die Art der Abbildung der Daten zum Diagram ist die Skala. Man muss die Skala wählen - welchen visuellen Anker verwende - die Zuordnung.
- **Layers** Man kann viele Layer übereinaqnder schichten. Zum Beispiel, wenn eine Regressionachse durch Punkte gelegt wird, dann ist die Regressionslinie ein (zusätzlicher) Layer.
- **Facets** Mit Facets kann man mehrere Gruppen unterscheiden. Zum Beispiel, der Umsatz nach verschiedenen Betriebsstätten. Die kategoriale Variable Betriebsstätte dient als Gruppierungsmerkmal. Die Daten für die Facets kommen normalerweise aus dem Datensatz.

Facets und Layer sind optional, die vorherigen sind immer da.

Zwei Grafiksysteme in R

Es gibt in R zwei Grafiksysteme:

Base Graphics

- **Low Level** Devices (Files mit Pixel, pdf, svg) -> Pixel -> kann nicht nachträglich verändern (daher sehr eingeschränkt in Möglichkeiten) -> es gibt keine Container oder Layers auf die zugreifen kann.
- **High Level** Die Punkte werden mit Funktionen erstellt, wie `plot`, `barplot()`, `hist()`, `segment`, `rect`.

Das ist sehr eingeschränkt, daher Erweiterung mit `grid`

Grid

- **Low Level** Mit containern.
- **High Level** Es gibt verschiedene Frameworks. Das `ggplot2` Framework implementiert Grammer for Grafiks (gg). Mit `ggplot2` vieles relativ elegant machbar. Nicht alles machen, aber vieles geht sehr gut.

Die zwei R-Grafikansätze kennen sich nicht, kann nicht mischen.