

Übungsblatt 2 - Informationsvisualisierung und Visual Analytics

Nick Martin, Tim Lukas

Aufgabe 1: Daten- und Task-Analyse

What? (Datenanalyse)

- **Datentyp:** bei den Rohdaten handelt es sich um eine **Tabelle**.
- **Items (Zeilen):** Die Items sind die 6 **Personen** (A - F). Jede Person stellt eine Entität dar.
- **Attribute (Spalte):** Die Attribute sind die 5 **Programmiersprachen** (Java, Perl, JS, C#, Python). Sie beschreiben Eigenschaften der Items.
- **Werte (Zellen):** Die Werte in den Zellen repräsentieren die Selbsteinschätzung der Kenntnisse einer Person für eine bestimmte Programmiersprache, die Werte liegen auf einer Skala von 1-5.

Attribut Typen die Attribute der Tabelle entsprechen folgenden Klassifikationen:

- **Personen:** Personen sind kategorische Attribute. Sie haben keine implizite Ordnung und dienen in dieser Tabelle als eindeutiger Schlüssel, um auf die Werte eines Items zuzugreifen.
- **Programmiersprachen:** Auch die Programmiersprachen sind kategorische Attribute.
- **Werte der Selbsteinschätzung:** Die Selbsteinschätzungen sind ordinale Attribute. Es existiert eine klare Rangfolge (5 ist besser als 4), die Abstände zwischen den Werten sind jedoch weder quantifizierbar noch notwendigerweise gleichmäßig.

Why? (Task)

Die gegebene Aufgabe "*Wer ist der beste und wer der schlechteste Programmierer*" ist eine Domänenfrage. Sie ist:

- **Mehrdeutig:** Die Definition von "bester" ist nicht klar. Ist es die Person mit konstant hohen Werten oder eher die Person mit Spitzenwerten in wenigen Kategorien?
- **Unpräzise:** Es wird keine klare Metrik für den Vergleich vorgegeben.
- **Nicht direkt ausführbar:** Die Frage kann nicht ohne weitere Annahmen direkt an die Daten gestellt werden.

Die Frage könnte in folgende Datenfragen übersetzt werden:

"Welche Person hat die höchste Summe an Fähigkeitspunkten über alle fünf Programmiersprachen, und welche die niedrigste?"

Abstraktion der Aufgabe: Das übergeordnete Ziel ist der Vergleich der Gesamtkompetenzen aller Personen, um die Extremwerte zu identifizieren. Um dieses Ziel zu erreichen sind folgende Aktionen erforderlich:

1. **Zusammenfassen / Aggregieren:** Da jede Person durch mehrere Attribute (fünf Programmiersprachen) beschrieben wird, kann kein direkter Vergleich stattfinden. Zuerst müssen diese verschiedenen Werte für jede Person zu einem einzigen, repräsentativen Wert zusammengefasst werden.
 - **Beispiel:** Berechnung der Summe oder des Medians der Fähigkeitswerte pro Person.

2. **Ordnen:** Die aus der Aggregation resultierenden Gesamtwerte werden anschließend genutzt, um eine Rangliste aller Personen zu erstellen. Diese Aktion bringt die Items(Personen) in eine explizite Reihenfolge.
3. **Suchen/Identifizieren:** Im finalen Schritt müssen die beiden gesuchten Items aus der im vorherigen Schritt erstellten, geordneten Liste extrahiert werden: Die erste Person (der "beste") und die letzte Person (der "schlechteste").

Ergebnis der Task-Abstraktion: Die Aufgabe besteht darin, für eine Menge von Items (Personen) deren Attribute zu aggregieren, um eine Rangordnung zu erstellen, aus der die Extremwerte identifiziert werden können.

Aufgabe 2 - Datenattribute

Diskussion zwischen den in der Vorlesung vorgestellten und in der Publikation vorgeschlagenen Attribut-Typen

Der größte Unterschied zwischen der Aufgliederung von S. S. Stevens' Attribut-Typen und dem in der Vorlesung behandelten System liegt in ihrer jeweiligen Fokus. Stevens' unterscheidet zwischen Nominal-, Ordinal-, Intervall- und Verhältnisskalen. Seine Einteilung ist auf die statistische Analyse ausgelegt. Sie definiert, welche mathematischen Operationen für einen Datentyp zulässig sind, damit die Ergebnisse statistisch sinnvoll bleiben. In der Vorlesung haben wir die Attribut-Typen kategorial, geordnet und quantitativ behandelt. Diese sind dagegen besser für die grafische Darstellung von Daten geeignet. Dabei geht es vor allem darum, herauszufinden, wie die Daten am besten visualisiert werden können (z. B. durch Farbe, Form oder Position), damit schnell erkannt werden kann was die Daten aussagen.

Bei den Gemeinsamkeiten fällt auf, dass Stevens' Nominalskala direkt dem kategorialen Attributstyp entspricht, der in der Vorlesung behandelt wurde. Auch die Ordinalskala entspricht dem geordneten (ordered) Typ aus der Vorlesung.

Der größte Unterschied liegt bei den numerischen Daten. Stevens unterscheidet zwischen Intervallskalen, die keinen echten Nullpunkt haben (bspw. Celsius), und Verhältnisskalen mit einem absoluten Nullpunkt (z.B. Größe). Diese Trennung ist statistisch entscheidend, da Verhältnisse nur bei einem absoluten Nullpunkt sinnvoll sind. Das in der Vorlesung behandelte Modell fasst beide Typen von Stevens zu einem einzigen quantitativen Typ zusammen. Dies hat den Grund, dass für die grafische Umsetzung die Art des Nullpunkts meist keine Rolle spielt. Sowohl Intervall- als auch Verhältnisskalen werden typischerweise durch dieselben visuellen Kanäle, wie die Position auf einer Achse oder die Balkenlänge, abgebildet.

Abschließend lässt sich sagen, dass das in der Vorlesung behandelte System eine für die Informationsvisualisierung optimierte Version des Systems von Stevens ist. Es vereinfacht Stevens' strenge statistische Regeln zu einem Modell, das hilft, gute Design-Entscheidungen für die visuelle Darstellung von Daten zu treffen.