Multivariate Data &
Tables and Graphs

CS 4460 – Intro. to Information Visualization
Aug. 28, 2017
John Stasko

Learning Objectives

• Explain different types of data models
• Describe different variable types (categories)
• Define metadata
• Know when to use a table versus a graph
• Explain marks and mark properties
• Identify effective techniques for low-dimensional (<=3) data
• Given raw data, be able to analyze, model, and transform into tabular data
Data

- Data is taken from and/or representing some phenomena from the world
- Data models something of interest to us
- Data comes in many different forms
  - Typically, not in the way you want it

- What is available to me (in the raw)?

Example

- Cars
  - make
  - model
  - year
  - miles per gallon
  - cost
  - number of cylinders
  - weights
  - ...

Fall 2017  CS 4460
Example

- Web pages

Data Models

- Often characterize data through three components
  - Objects
    - Items of interest
      - (students, courses, terms, ...)
  - Attributes
    - Characteristics or properties of data
      - (name, age, GPA, number, date, ...)
  - Relations
    - How two or more objects relate
      - (student takes course, course during term, ...)

Fall 2017
CS 4460
6
Data Tables

- We take raw data and transform it into a model/form that is more workable
- Main idea:
  - Individual items are called *cases*
  - Cases have *variables* (attributes)

Statistical Model

- Independent and Dependent variables
- Dimensions
  - Discrete, categorical info
- Measures
  - Continuous, quantitative info
## Data Table Format

<table>
<thead>
<tr>
<th>Variable</th>
<th>Case₁</th>
<th>Case₂</th>
<th>Case₃</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable₁</td>
<td>Value₁₁</td>
<td>Value₂₁</td>
<td>Value₃₁</td>
<td></td>
</tr>
<tr>
<td>Variable₂</td>
<td>Value₁₂</td>
<td>Value₂₂</td>
<td>Value₃₂</td>
<td></td>
</tr>
<tr>
<td>Variable₃</td>
<td>Value₁₃</td>
<td>Value₂₃</td>
<td>Value₃₃</td>
<td></td>
</tr>
</tbody>
</table>

Think of as a function

\[ f(\text{case₁}) = \langle \text{Val}_{11}, \text{Val}_{12}, \ldots \rangle \]

## Example

<table>
<thead>
<tr>
<th>Mary</th>
<th>Jim</th>
<th>Sally</th>
<th>Mitch</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSN</td>
<td>145</td>
<td>294</td>
<td>563</td>
<td>823</td>
</tr>
<tr>
<td>Age</td>
<td>23</td>
<td>17</td>
<td>47</td>
<td>29</td>
</tr>
<tr>
<td>Hair</td>
<td>brown</td>
<td>black</td>
<td>blonde</td>
<td>red</td>
</tr>
<tr>
<td>GPA</td>
<td>2.9</td>
<td>3.7</td>
<td>3.4</td>
<td>2.1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

People in class
Or

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Age</th>
<th>Hair</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>145</td>
<td>23</td>
<td>brown</td>
<td>2.9</td>
</tr>
<tr>
<td>Jim</td>
<td>294</td>
<td>17</td>
<td>black</td>
<td>3.7</td>
</tr>
<tr>
<td>Sally</td>
<td>563</td>
<td>47</td>
<td>blonde</td>
<td>3.4</td>
</tr>
<tr>
<td>Mitch</td>
<td>823</td>
<td>29</td>
<td>red</td>
<td>2.1</td>
</tr>
</tbody>
</table>

People in class

Example

Baseball statistics
Wide vs. Long Data

Wide  

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>32</td>
<td>128</td>
</tr>
<tr>
<td>Alice</td>
<td>24</td>
<td>80</td>
</tr>
<tr>
<td>Steve</td>
<td>64</td>
<td>95</td>
</tr>
</tbody>
</table>

Each attribute gets a column

Long (Narrow)

<table>
<thead>
<tr>
<th>Person</th>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>Age</td>
<td>32</td>
</tr>
<tr>
<td>Bob</td>
<td>Weight</td>
<td>128</td>
</tr>
<tr>
<td>Alice</td>
<td>Age</td>
<td>24</td>
</tr>
<tr>
<td>Alice</td>
<td>Weight</td>
<td>86</td>
</tr>
<tr>
<td>Steve</td>
<td>Age</td>
<td>64</td>
</tr>
<tr>
<td>Steve</td>
<td>Weight</td>
<td>95</td>
</tr>
</tbody>
</table>

For each data case, there is an attribute-value pair

https://en.wikipedia.org/wiki/Wide_and_narrow_data

Variable Types

- Three main types of variables
  - N-Nominal (equal or not equal to other values)
    - Example: gender
  - O-Ordinal (obeys < relation, ordered set)
    - Example: fr, so, jr, sr
  - Q-Quantitative (can do math on them)
    - Example: age
Metadata

- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn’t enough
  - Example: if variable1 is “l”, then variable3 can only be 3, 7 or 16

Data Cleaning

- Data may be missing/corrupted
  - Remove?
  - Modify?
- You may want to adjust values
  - Use inverse
  - Map nominal to ordinal/quantitative
  - Normalize values
    Scale between 0 and 1
Nice Interactive Tool

https://www.trifacta.com/start-wrangling/

Administratia

- Sign up for Piazza
- Class slides: external & internal
- Office hours coming
  - John S.
  - John T.
  - Ayshwarya
  - Ayan
  - Bethany
Surveys

• Who hasn’t completed one?

How Many Variables?

• Data sets of dimensions 1, 2, 3 are common
• Number of variables per class
  – 1 - Univariate data
  – 2 - Bivariate data
  – 3 - Trivariate data
  – >3 - Hypervariate data
Representation

• What are two main ways of presenting multivariate data sets?
  – Directly (textually) → Tables
  – Symbolically (pictures) → Graphs

• When use which?

Strengths?

S. Few
Show Me the Numbers

• Use tables when
  – The document will be used to look up individual values
  – The document will be used to compare individual values
  – Precise values are required
  – The quantitative info to be communicated involves more than one unit of measure

• Use graphs when
  – The message is contained in the shape of the values
  – The document will be used to reveal relationships among values
Effective Table Design

- See *Show Me the Numbers*
  - Next examples taken from there
- Proper and effective use of layout, typography, shading, etc. can go a long way
- (Tables may be underused)

Example

2003 Q1-to-Date Regional Sales
March 15, 2003

<table>
<thead>
<tr>
<th>Region</th>
<th>Sales (U.S.$)</th>
<th>Percent of Total Sales</th>
<th>Current Qtr Plan %</th>
<th>Qtr End Projected Sales (U.S.$)</th>
<th>Qtr End Projected Percent of Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Americas</td>
<td>469,384</td>
<td>60%</td>
<td>85%</td>
<td>586,730</td>
<td>107%</td>
</tr>
<tr>
<td>Europe</td>
<td>273,854</td>
<td>35%</td>
<td>91%</td>
<td>353,272</td>
<td>118%</td>
</tr>
<tr>
<td>Asia</td>
<td>34,847</td>
<td>5%</td>
<td>50%</td>
<td>43,210</td>
<td>62%</td>
</tr>
</tbody>
</table>

Note: To date, 83% of the quarter has elapsed.
Graphs

- Visual structures composed of
  - Spatial substrate
  - Marks
  - Graphical properties of marks
Space

- Visually dominant
- Often put axes on space to assist
- Use techniques of composition, alignment, folding, recursion, overloading to
  1) increase use of space
  2) do data encodings

Marks

- Things that occur in space
  - Points
  - Lines
  - Areas
  - Volumes
Graphical Properties

- Size, shape, color, orientation...

<table>
<thead>
<tr>
<th>Expressing extent</th>
<th>Spatial properties</th>
<th>Object properties</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Position</td>
<td>Grayscale</td>
</tr>
<tr>
<td></td>
<td>Size</td>
<td></td>
</tr>
<tr>
<td>Differentiating marks</td>
<td>Orientation</td>
<td>Color Shape Texture</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Back to Data

- What were the different types of data sets?
- Number of variables per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data
Univariate Data

Representations

What Goes Where?

- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another
**Alternative View**

- We may think of graph as representing independent (data case) and dependent (value) variables
- Guideline:
  - Independent vs. dependent variables
    - Put independent on x-axis
    - See resultant dependent variables along y-axis

**Bivariate Data**

- Representations

![Scatter plot example](scatter_plot_example.png)

Scatter plot is common

Two variables, want to see relationship

Is there a linear, curved or random pattern?
Trivariate Data

- Representations

3D scatter plot is possible

```
price

horsepower

mileage
```

Alternative Representation

Still use 2D but have mark property represent third variable
**Alternative Representation**

Represent each variable in its own explicit way.

---

**Hypervariate Data**

- Ahhh, the tough one
- Number of well-known visualization techniques exist for data sets of 1-3 dimensions
  - line graphs, bar graphs, scatter plots
  - We see a 3-D world (4-D with time)
- What about data sets with more than 3 variables?
  - Often the interesting, challenging ones
Multiple Views

Give each variable its own display

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

Scatterplot Matrix

Represent each possible pair of variables in their own 2-D scatterplot

Useful for what?
Misses what?
Thoughts

- Liked the "living more in the present"
- Note each data case is not a simple event occurrence
  - Typically some attribute as well
    - Which animal did you see?
    - How did you feel?
    - What did you eat?
- What would you log?
More to Come...

- Subsequent days will explore other general techniques for handling hypervariate data

Advice

- Take DB & IR courses
  - Learn about query languages, relational data models, data cubes, data warehouses, ...
Learning Objectives

• Explain different types of data models
• Describe different variable types (categories)
• Define metadata
• Know when to use a table versus a graph
• Explain marks and mark properties
• Identify effective techniques for low-dimensional (<=3) data
• Given raw data, be able to analyze, model, and transform into tabular data

HW 1

• Data analysis without vis

• Due Friday
Upcoming

- Statistical Charts & Graphs
  - Prep: Few article, pp. 1-20

- Lab: HTML, CSS, DOM

Sources Used

- Few book
- CMS book
- Referenced articles
- Marti Hearst SIMS 247 lectures
- Kosslyn '89 article
- A. Marcus, Graphic Design for Electronic Documents and User Interfaces
- W. Cleveland, The Elements of Graphing Data