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
  - ...

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, …)

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

Dimensions

	Case_1	Case_2	Case_3	…
Variable_1	Value_{11}	Value_{21}	Value_{31}	
Variable_2	Value_{12}	Value_{22}	Value_{32}	
Variable_3	Value_{13}	Value_{23}	Value_{33}	
…				

Think of as a function
$f(\text{case}_1) = <\text{Val}_{11}, \text{Val}_{12},\ldots>$

Example

	Mary	Jim	Sally	Mitch	…
SSN	145	294	563	823	
Age	23	17	47	29	
Hair	brown	black	blonde	red	
GPA	2.9	3.7	3.4	2.1	
…					

People in class

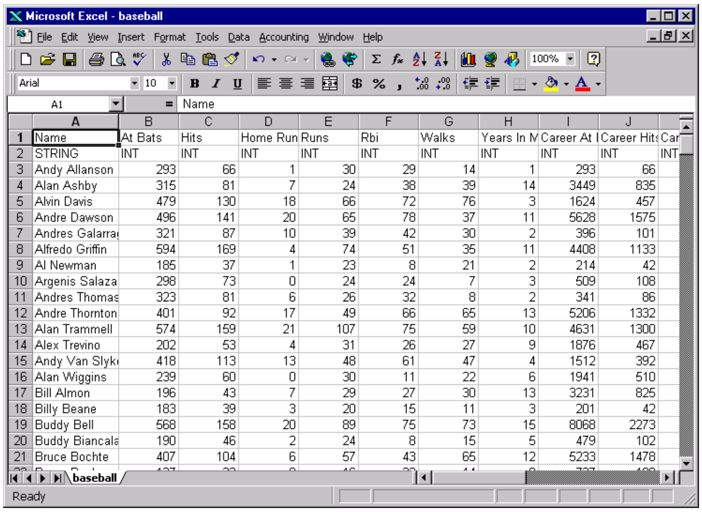
Or

	P1	P2	P3	P4	…
Name	Mary	Jim	Sally	Mitch	
SSN	145	294	563	823	
Age	23	17	47	29	
Hair	brown	black	blonde	red	
GPA	2.9	3.7	3.4	2.1	
…					

People in class

Example

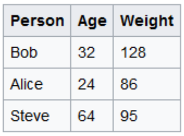
Baseball statistics


Microsoft Excel - baseball
File Edit View Insert Format Tools Data Accounting Window Help
100%
Arial
10
A1
=
Name
baseball
Ready

	A	B	C	D	E	F	G	H	I	J	
1	Name	At Bats	Hits	Home Run	Runs	Rbi	Walks	Years In M	Career At I	Career Hits	Car
2	STRING	INT	INT	INT	INT	INT	INT	INT	INT	INT	INT
3	Andy Allanson	293	66	1	30	29	14	1	293	66	
4	Alan Ashby	315	81	7	24	38	39	14	3449	835	
5	Alvin Davis	479	130	18	66	72	76	3	1624	457	
6	Andre Dawson	496	141	20	65	78	37	11	5628	1575	
7	Andres Galarra	321	87	10	39	42	30	2	396	101	
8	Alfredo Griffin	594	169	4	74	51	35	11	4408	1133	
9	Al Newman	185	37	1	23	8	21	2	214	42	
10	Argenis Salaza	298	73	0	24	24	7	3	509	108	
11	Andres Thomas	323	81	6	26	32	8	2	341	86	
12	Andre Thornton	401	92	17	49	66	65	13	5206	1332	
13	Alan Trammell	574	159	21	107	75	59	10	4631	1300	
14	Alex Trevino	202	53	4	31	26	27	9	1876	467	
15	Andy Van Slyk	418	113	13	48	61	47	4	1512	392	
16	Alan Wiggins	239	60	0	30	11	22	6	1941	510	
17	Bill Almon	196	43	7	29	27	30	13	3231	825	
18	Billy Beane	183	39	3	20	15	11	3	201	42	
19	Buddy Bell	568	158	20	89	75	73	15	8068	2273	
20	Buddy Biancala	190	46	2	24	8	15	5	479	102	
21	Bruce Bochte	407	104	6	57	43	65	12	5233	1478	

Wide vs. Long Data

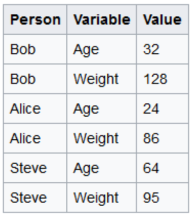
Wide



Person	Age	Weight
Bob	32	128
Alice	24	86
Steve	64	95

Each attribute gets a column

Long (Narrow)



Person	Variable	Value
Bob	Age	32
Bob	Weight	128
Alice	Age	24
Alice	Weight	86
Steve	Age	64
Steve	Weight	95

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

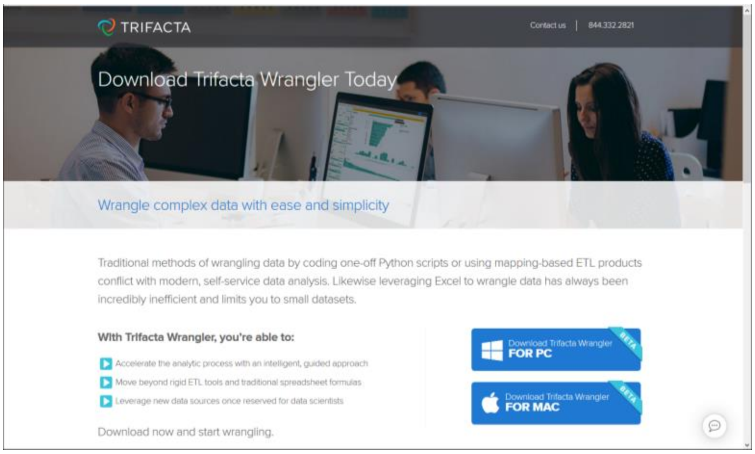
Metadata

- 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



Download Trifacta Wrangler Today

Wrangle complex data with ease and simplicity

Traditional methods of wrangling data by coding one-off Python scripts or using mapping-based ETL products conflict with modern, self-service data analysis. Likewise leveraging Excel to wrangle data has always been incredibly inefficient and limits you to small datasets.

With Trifacta Wrangler, you're able to:

- Accelerate the analytic process with an intelligent, guided approach
- Move beyond rigid ETL tools and traditional spreadsheet formulas
- Leverage new data sources once reserved for data scientists

Download now and start wrangling.

Download Trifacta Wrangler FOR PC
BETA
Download Trifacta Wrangler FOR MAC
BETA

`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?

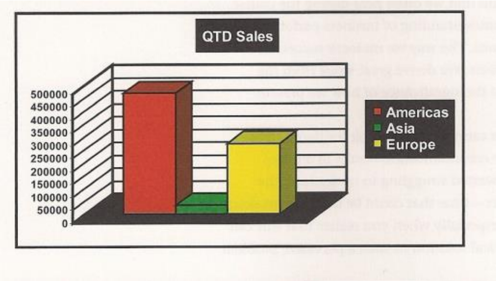
S. Few
Show Me the Numbers

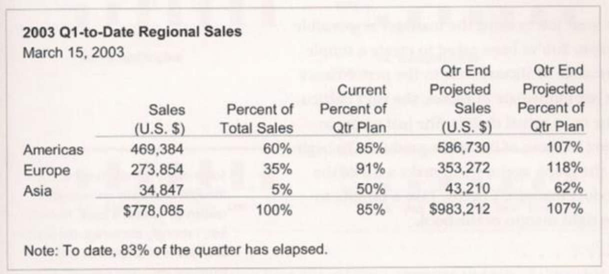
- Use tables when
  - The document will be used to look up individual values
  - The document will be used to compare individual values
  - Precise values are required
  - The quantitative info to be communicated involves more than one unit of measure
- Use graphs when
  - The message is contained in 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



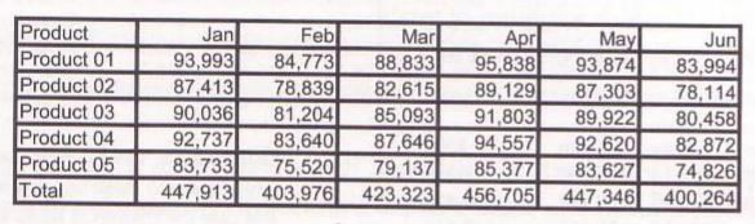


2003 Q1-to-Date Regional Sales
March 15, 2003

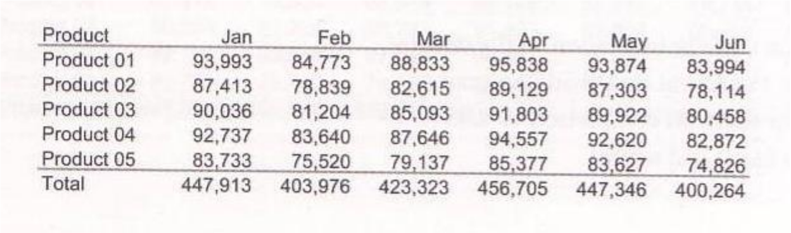
	Sales (U.S. $)	Percent of Total Sales	Current Percent of Qtr Plan	Qtr End Projected Sales (U.S. $)	Qtr End Projected Percent of Qtr Plan
Americas	469,384	60%	85%	586,730	107%
Europe	273,854	35%	91%	353,272	118%
Asia	34,847	5%	50%	43,210	62%
	$778,085	100%	85%	$983,212	107%

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

Example



Product	Jan	Feb	Mar	Apr	May	Jun
Product 01	93,993	84,773	88,833	95,838	93,874	83,994
Product 02	87,413	78,839	82,615	89,129	87,303	78,114
Product 03	90,036	81,204	85,093	91,803	89,922	80,458
Product 04	92,737	83,640	87,646	94,557	92,620	82,872
Product 05	83,733	75,520	79,137	85,377	83,627	74,826
Total	447,913	403,976	423,323	456,705	447,346	400,264

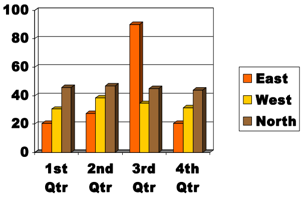


Product	Jan	Feb	Mar	Apr	May	Jun
Product 01	93,993	84,773	88,833	95,838	93,874	83,994
Product 02	87,413	78,839	82,615	89,129	87,303	78,114
Product 03	90,036	81,204	85,093	91,803	89,922	80,458
Product 04	92,737	83,640	87,646	94,557	92,620	82,872
Product 05	83,733	75,520	79,137	85,377	83,627	74,826
Total	447,913	403,976	423,323	456,705	447,346	400,264

Graphs

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

Yuk

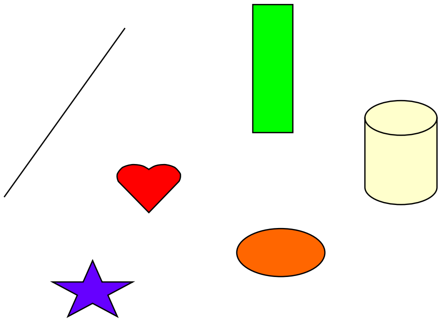


Space

- 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...

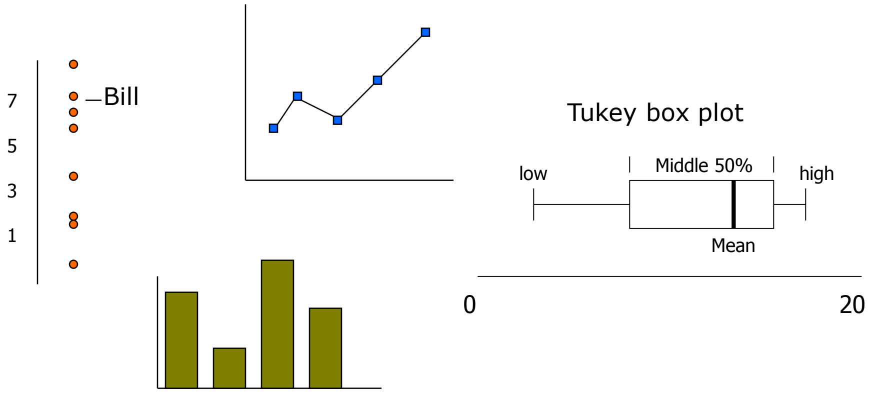
	Spatial properties	Object properties
Expressing extent	Position Size	Grayscale
Differentiating marks	Orientation	Color Shape Texture

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

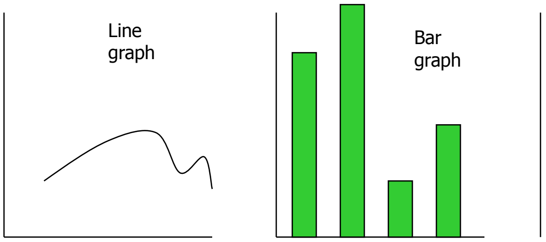
Univariate Data

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



Y-axis is quantitative variable

See changes over consecutive values

Y-axis is quantitative variable

Compare relative point values

Alternative View

- We may think of graph as representing independent (data case) and dependent (value) variables
- Guideline:
  - Independent vs. dependent variables

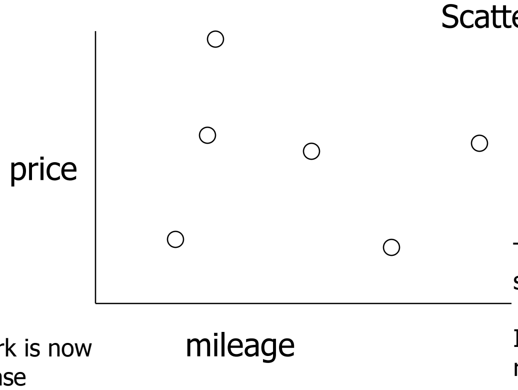
    Put independent on x-axis

    See resultant dependent variables along y-axis

Bivariate Data

- Representations

Scatter plot is common



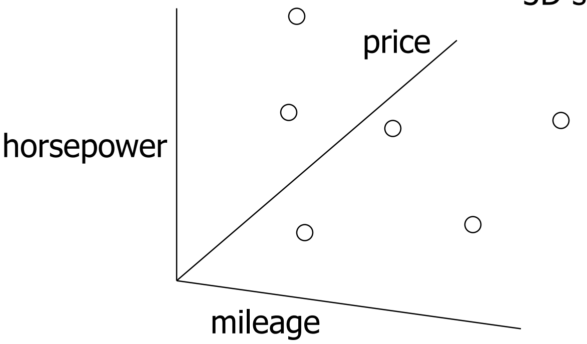
Two variables, want to see relationship

Each mark is now a data case

Is there a linear, curved or random pattern?

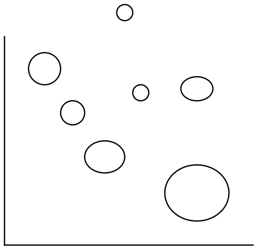
Trivariate Data

- Representations



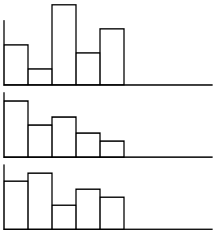
3D scatter plot is possible

Alternative Representation



Still use 2D but have mark property represent third variable

Alternative Representation



Represent each variable in its own explicit way

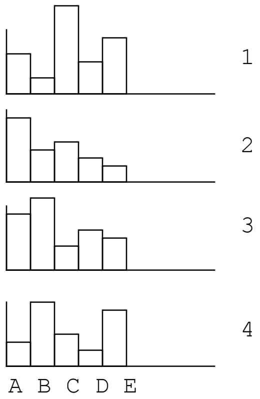
Hypervariate Data

- 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

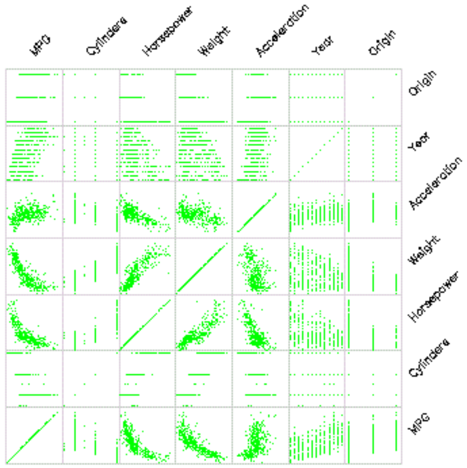
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5



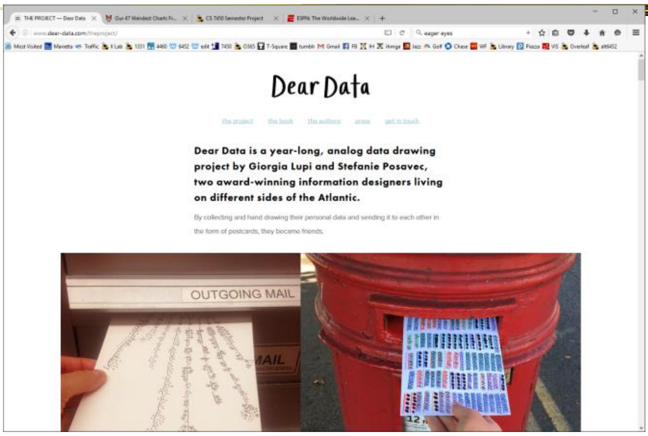
Scatterplot Matrix

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

Useful for what?
Misses what?



Dear Data



Dear Data

the project the book the authors press get in touch

Dear Data is a year-long, analog data drawing project by Giorgia Lupi and Stefanie Posavec, two award-winning information designers living on different sides of the Atlantic.

By collecting and hand drawing their personal data and sending it to each other in the form of postcards, they became friends.

```
http://www.dear-data.com
http://www.dear-data.com/all
```

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, datacubes, 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*