Exploratory Data Analysis

INFO 370

Learning Objectives

Discuss the **purpose** of exploratory data analysis

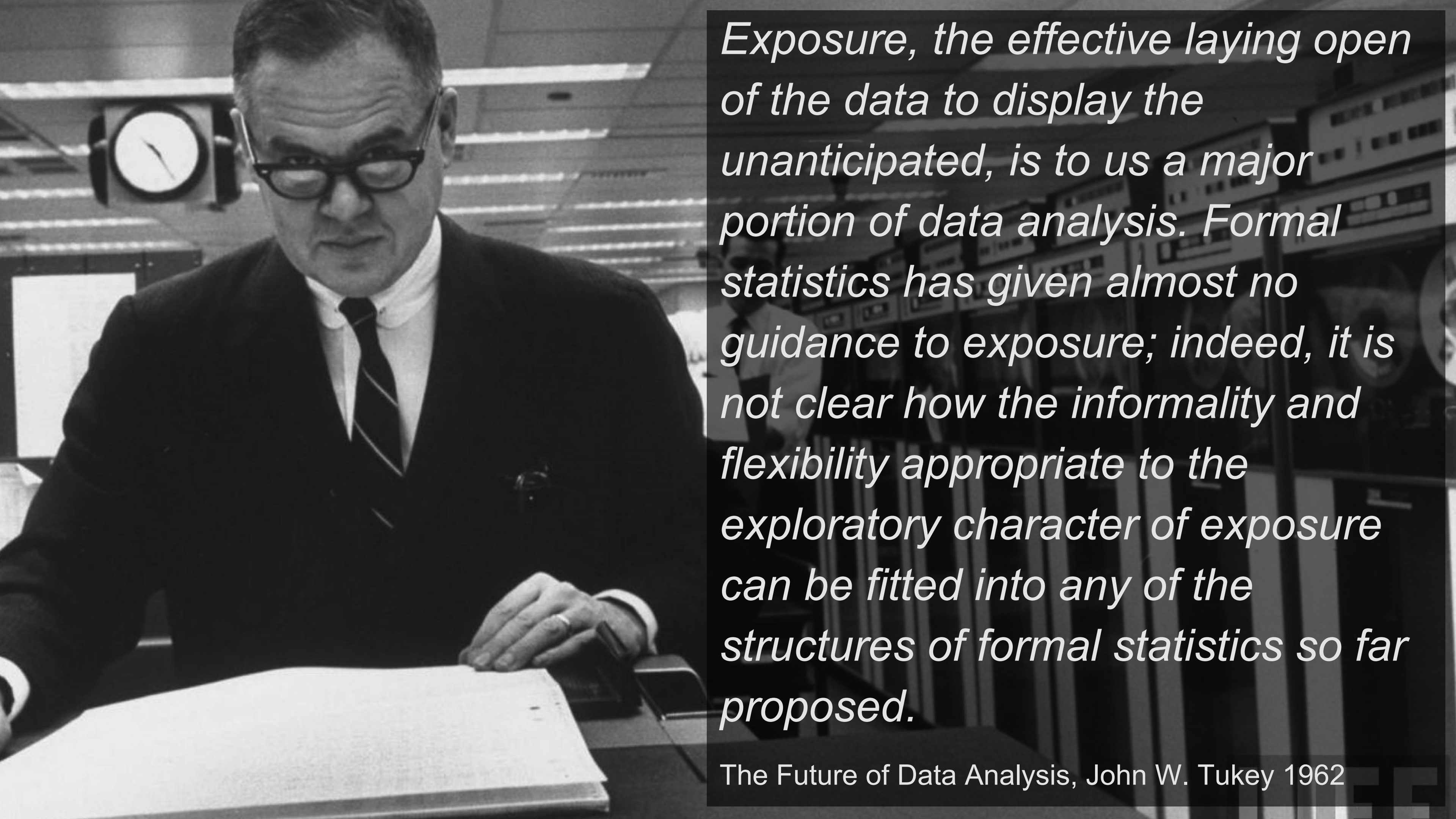
Develop a **set of questions** to ask of our datasets

Discuss **effectiveness** and **expressiveness** in visual layouts

Introduce the **Pandas** Python library for 2D data structures

Begin EDA by asking, how to health risks vary across the globe?

Exploratory Data Analysis



Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

The Future of Data Analysis, John W. Tukey 1962

X	Y		X	Y		X	Y		X	Y
10	8.04		10	9.14		10	7.46		8	6.58
8	6.95		8	8.14		8	6.77		8	5.76
13	7.58		13	8.74		13	12.74		8	7.71
9	8.81		9	8.77		9	7.11		8	8.84
11	8.33		11	9.26		11	7.81		8	8.47
14	9.96		14	8.1		14	8.84		8	7.04
6	7.24		6	6.13		6	6.08		8	5.25
4	4.26		4	3.1		4	5.39		19	12.5
12	10.84		12	9.11		12	8.15		8	5.56
7	4.82		7	7.26		7	6.42		8	7.91
5	5.68		5	4.74		5	5.73		8	6.89

Mean x: 9.0

Mean y: 7.5

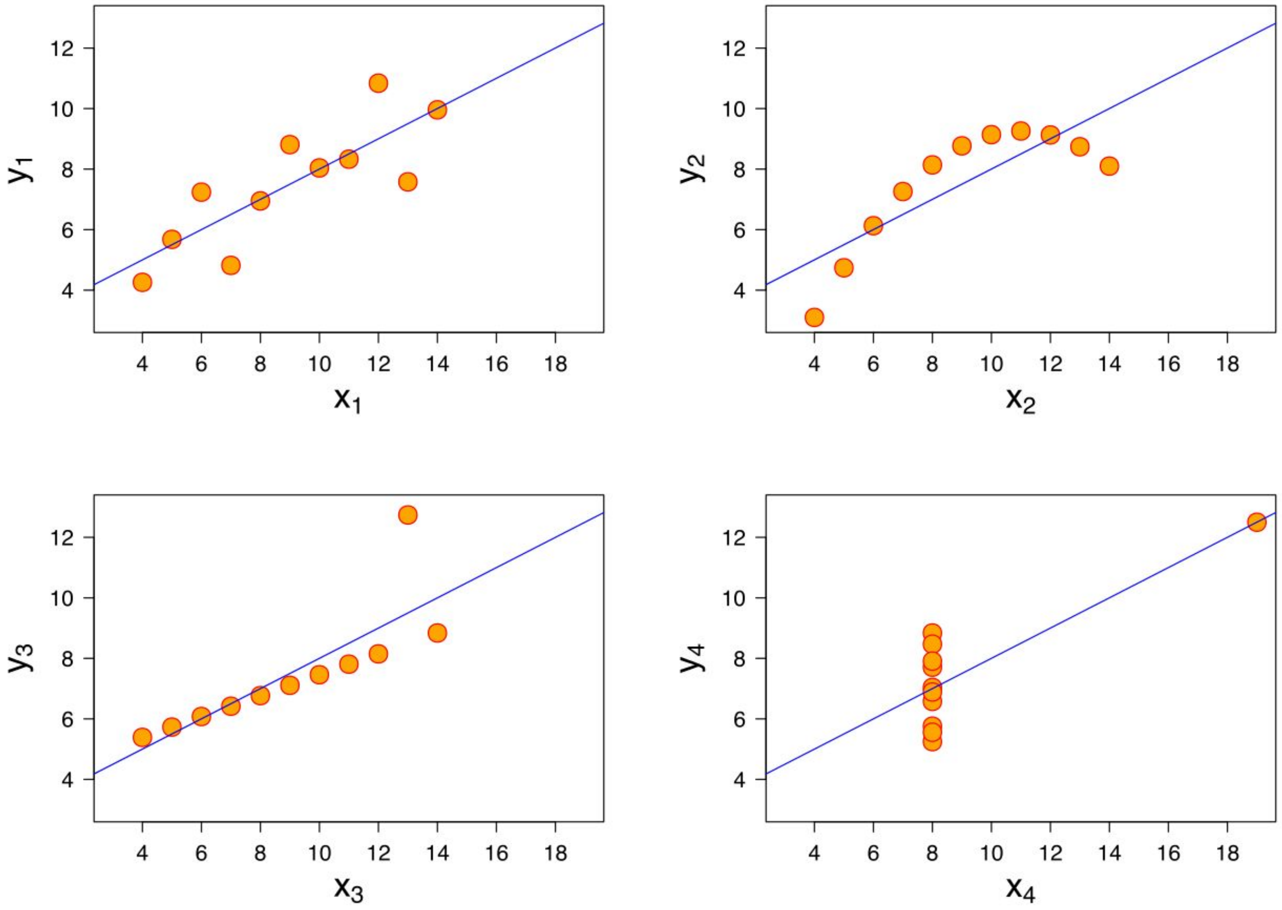
SD x: 3.317

SD y: 2.03

$y = 3 + .5x$

$R^2 = 0.67$

Even understanding small datasets is difficult...



But visualization can help.

Exploratory Data Analysis Purpose

Understand the structure of your data

Discover any data quirks (missingness, NA values, impossible values)

Test prior assumptions and assess data quality

Identify pertinent research questions

EDA Questions

Given a dataset for analysis, what questions do you need to ask about it?

EDA Questions

Data Structure

- # of rows/columns, variable names, data type for each variable

Univariate analysis

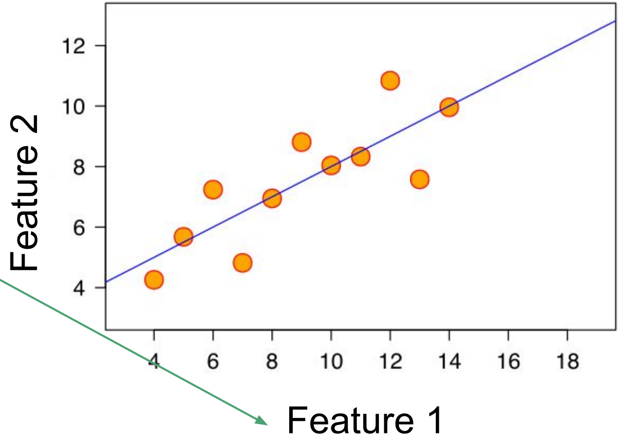
- Range/summary stats (min, max, mean, etc.), distribution, missingness

Multivariate

- Univariate distribution consistency across groupings
- Correlations between variables

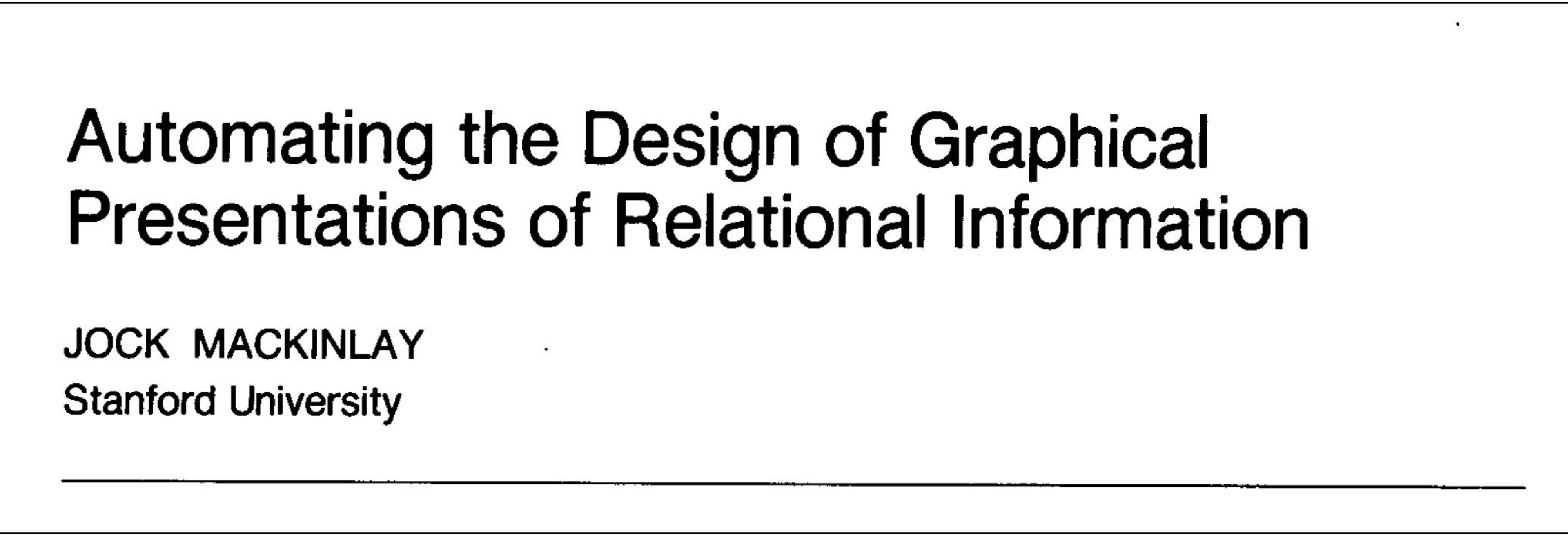
What is the process for providing visual answers to these questions (EDA)?

Feature 1	Feature 2
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68



Map from data features to visual features

Effectiveness and Expressiveness



Automating the Design of Graphical Presentations of Relational Information

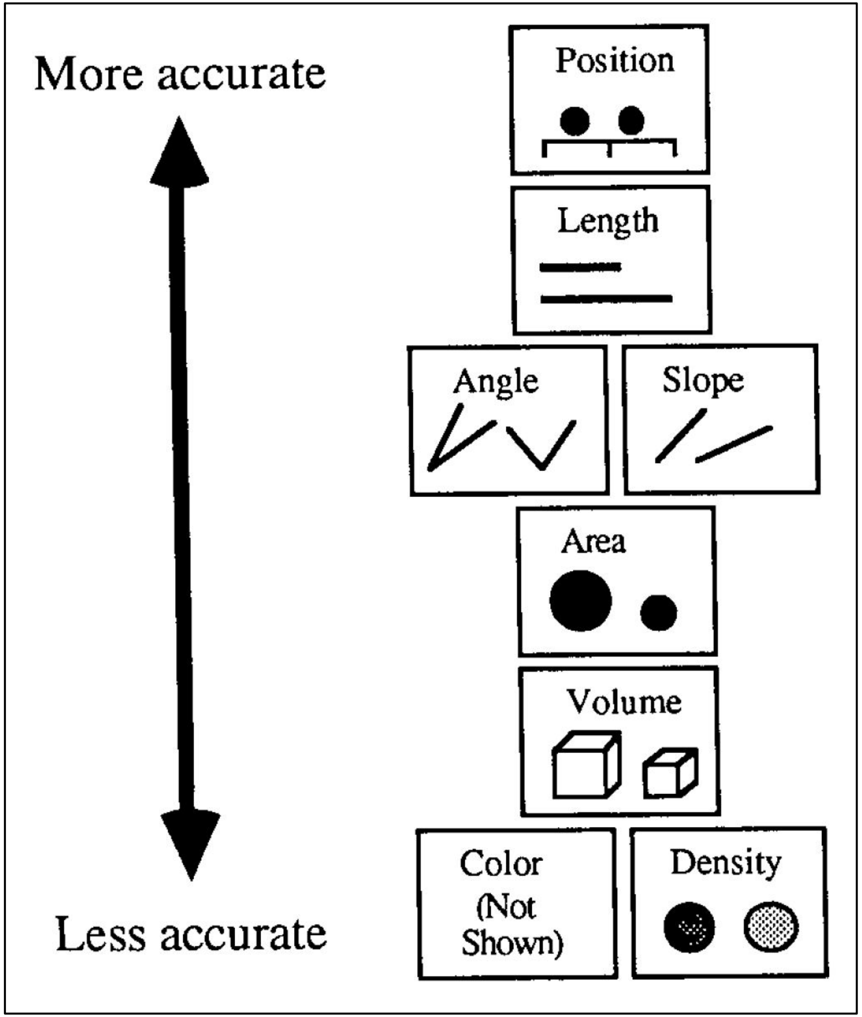
JOCK MACKINLAY
Stanford University

Link

“The graphic design issues are codified as ***expressiveness*** *and* ***effectiveness*** *criteria for graphical languages.* ***Expressiveness*** *criteria determine whether a graphical language can express the desired information.* ***Effectiveness*** *criteria determine whether a graphical language exploits the capabilities of the output medium and the human visual system.”*

- Mackinlay ‘86

Effectiveness

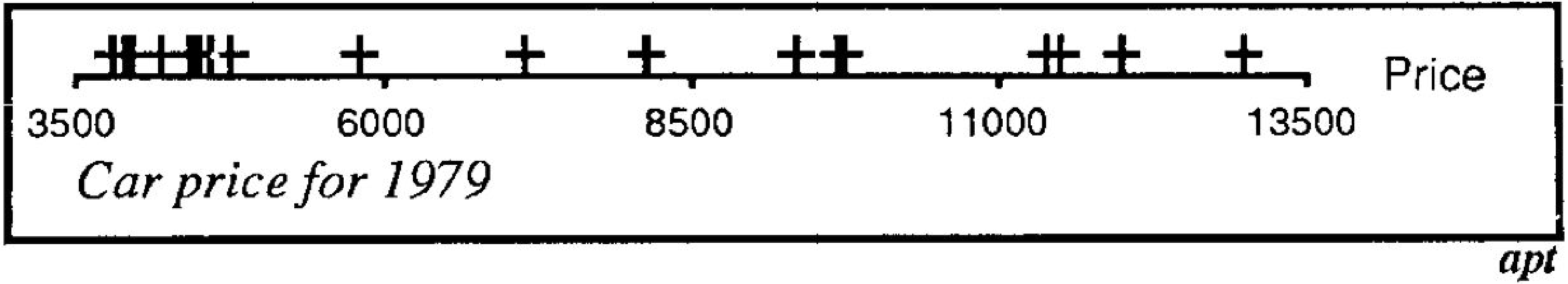


Expressiveness

"A set of facts {data} is expressible in a language {chart-type} if it contains a sentence {instance} that:

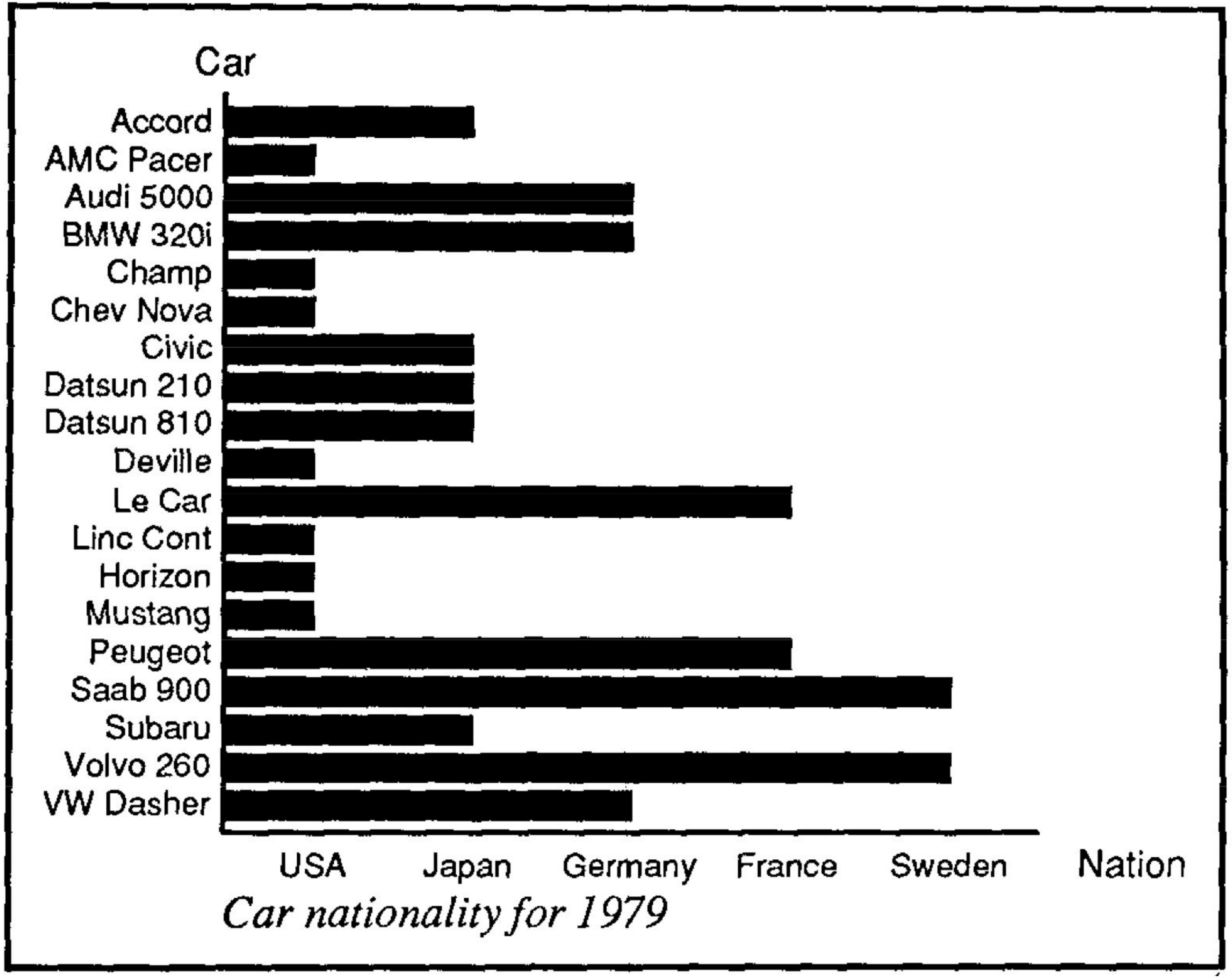
(1) encodes all the facts in the set,
(2) encodes only the facts in the set"

- Mackinlay '86 {added}

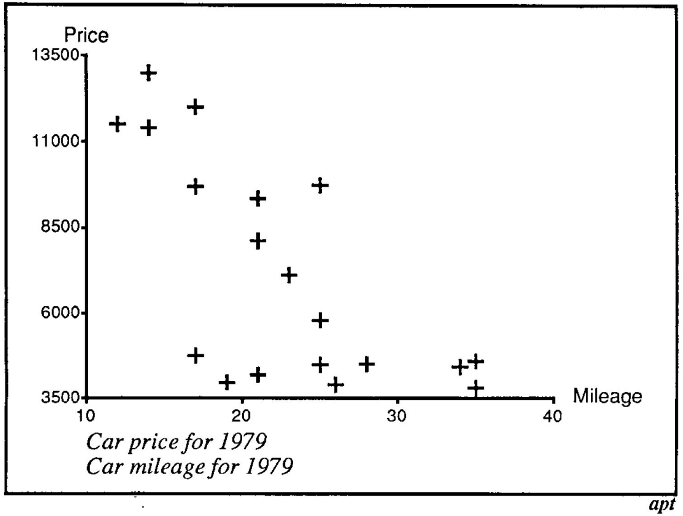


Unable to express all facts in the set
(fails first criterion)

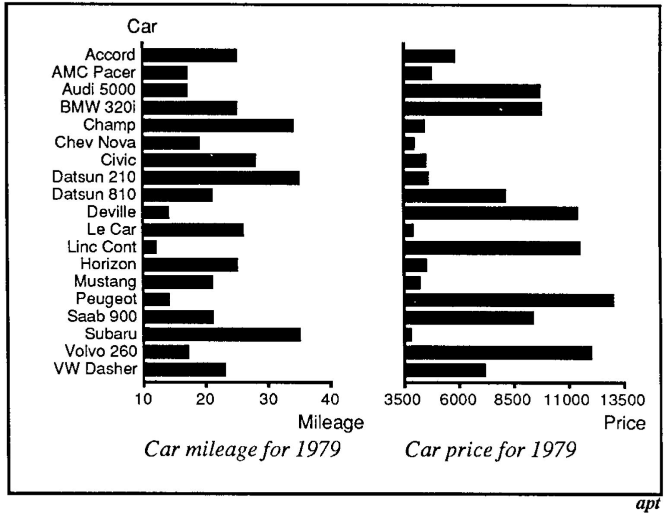
Expresses information not inherent in the dataset (fails second criterion)



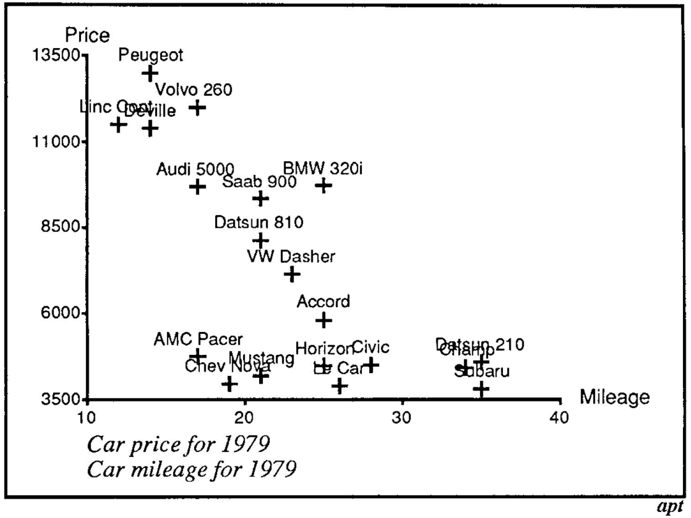
What are the trade-offs between effectiveness and expressiveness in these layouts?



Car price for 1979
Car mileage for 1979



Car mileage for 1979 Car price for 1979



Car price for 1979
Car mileage for 1979

Health Burden

Risk Dataset

Investigating the health burden of 5 risks:

- Smoking
- Low physical activity
- High red-meat consumption
- Drug use
- Alcohol use

Burden as measured by **death rates** (deaths per 100K people)

Data is broken down for each **country** by **age** and **sex** notebook-set-2

Upcoming...

Notebook set 2 due **Friday night**

Reading 2 (probability and statistics) due **next Tuesday** before class

This week: Developing metrics + R review