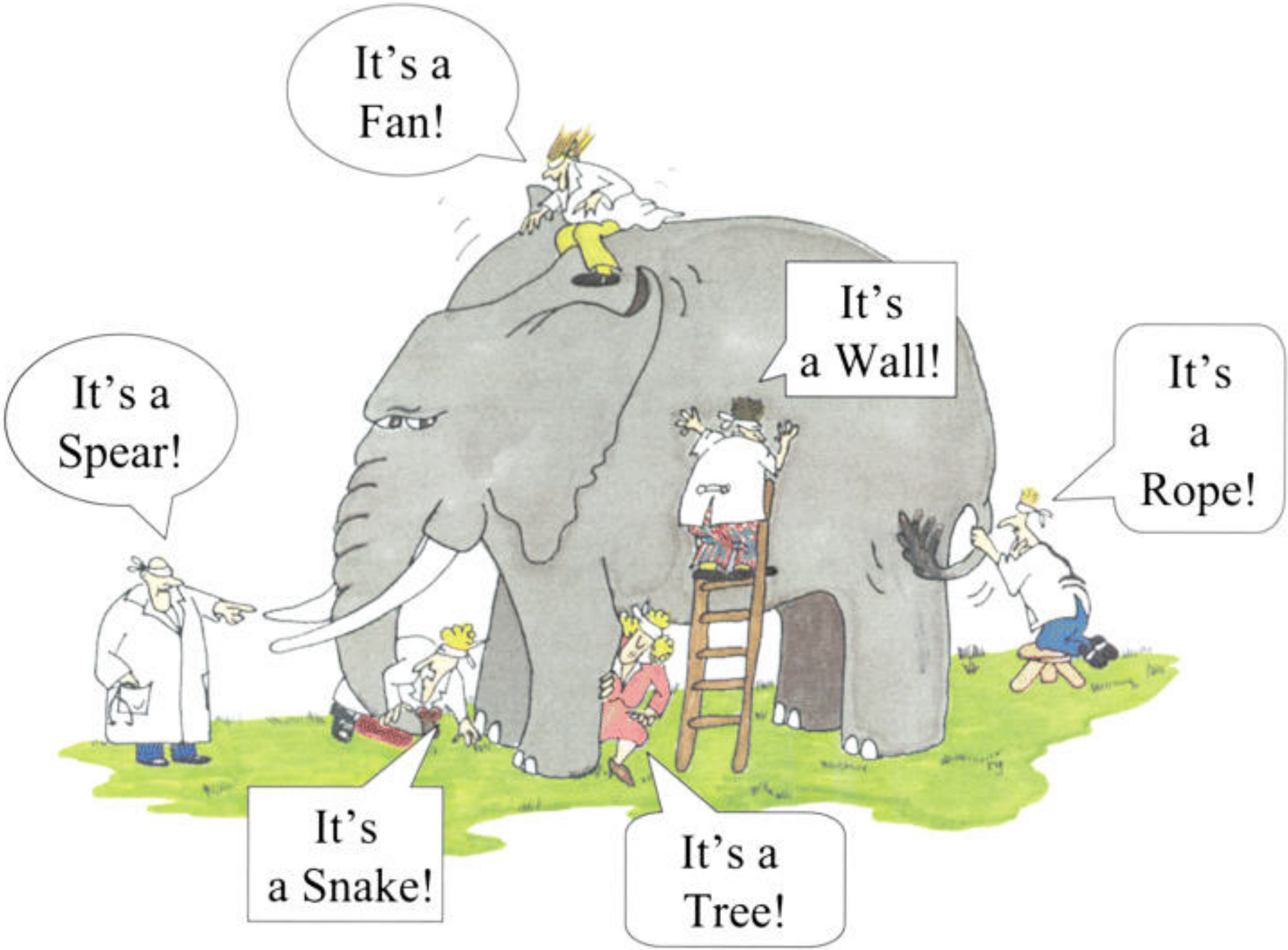
Practice and Applications of Data Management

CMPSCI 345

Lecture: Data Visualization


It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!

Data visualization

- The creation and study of the visual representation of data.
- “The goal of visualization is to aid our understanding of data by leveraging the human visual system’s highly tuned ability to see patterns, spot trends, and identify outliers.”

Understanding data

- It’s hard to interpret raw data consisting of more than a few numbers.
- One approach: summarize with statistics

Set 1

x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Set 2

x	y
10.0	9.14
8.0	8.14
13.0	8.74
9.0	8.77
11.0	9.26
14.0	8.10
6.0	6.13
4.0	3.10
12.0	9.13
7.0	7.26
5.0	4.74

Set 3

x	y
10.0	7.46
8.0	6.77
13.0	12.74
9.0	7.11
11.0	7.81
14.0	8.84
6.0	6.08
4.0	5.39
12.0	8.15
7.0	6.42
5.0	5.73

Set 4

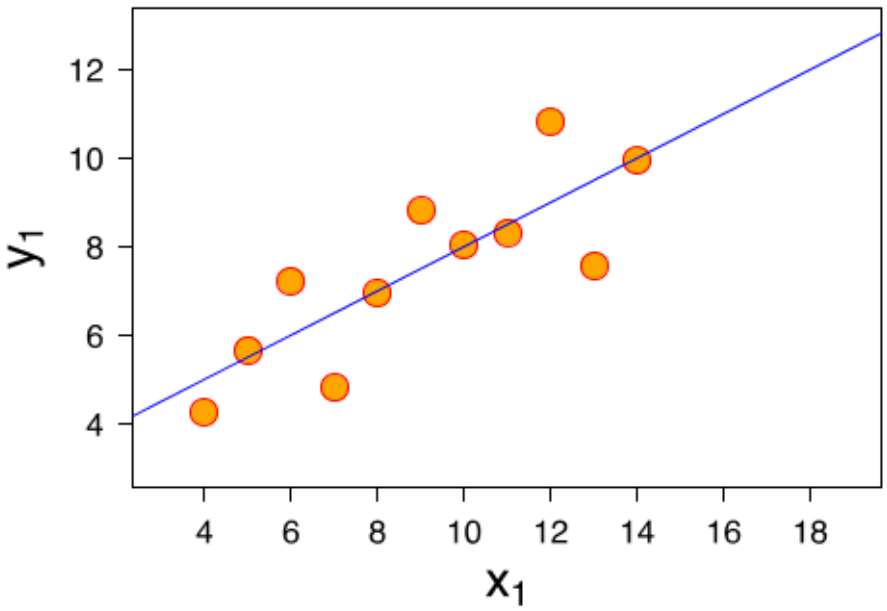
x	y
8.0	6.58
8.0	5.76
8.0	7.71
8.0	8.84
8.0	8.47
8.0	7.04
8.0	5.25
19.0	12.50
8.0	5.56
8.0	7.91
8.0	6.89

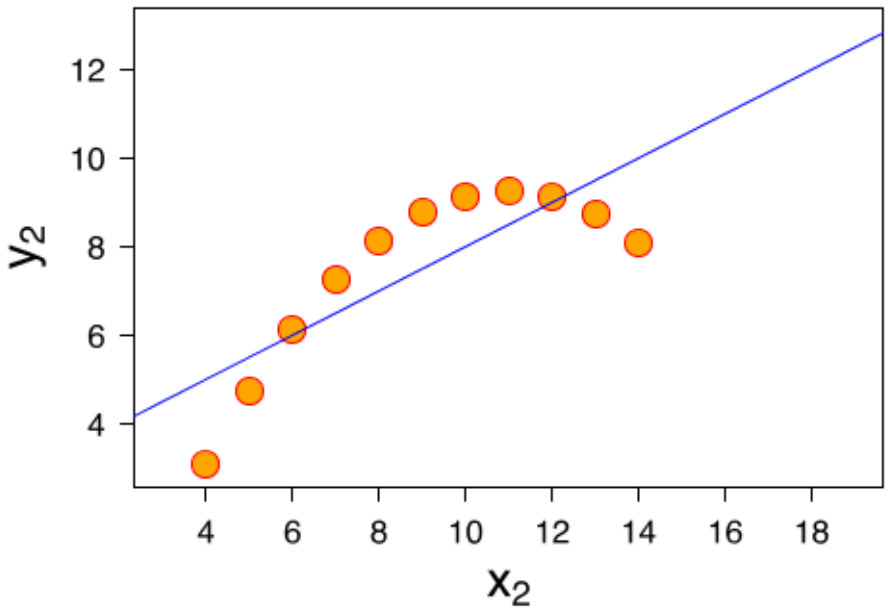
Understanding data

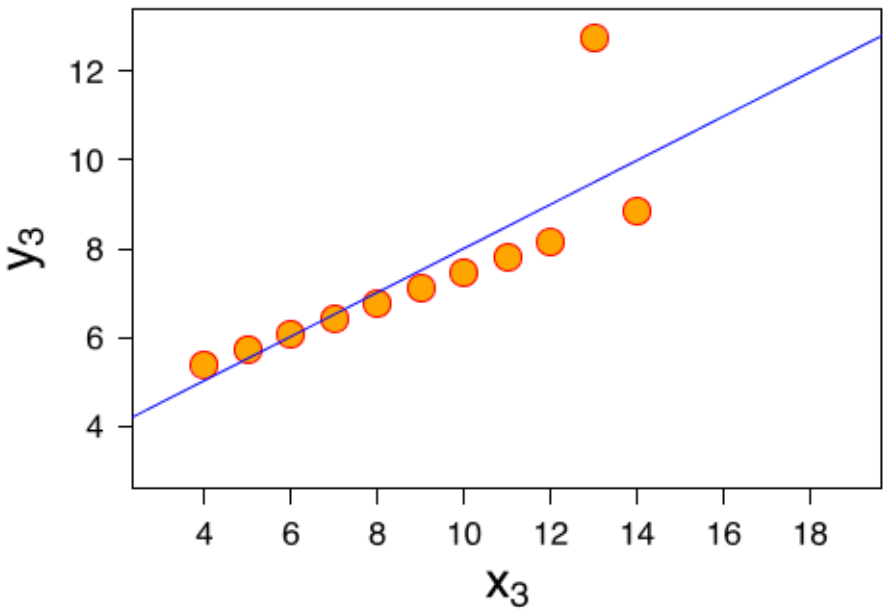
- It's hard to interpret raw data consisting of more than a few numbers.
- One approach: summarize with statistics
- Example: each of the four datasets, they have identical summary statistics:
  - mean of x: 9
  - mean of y: 7.50
  - correlation(x,y): .816
  - regression line: y = 3.00 + .5x

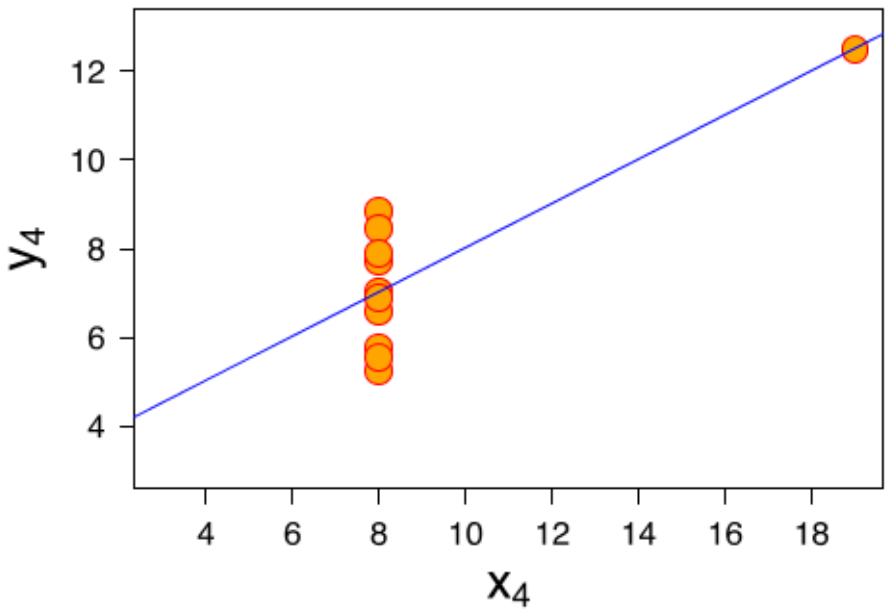
x	y
10.0	8.04
8.0	6.95
13.0	7.58
9.0	8.81
11.0	8.33
14.0	9.96
6.0	7.24
4.0	4.26
12.0	10.84
7.0	4.82
5.0	5.68

Anscombe's quartet







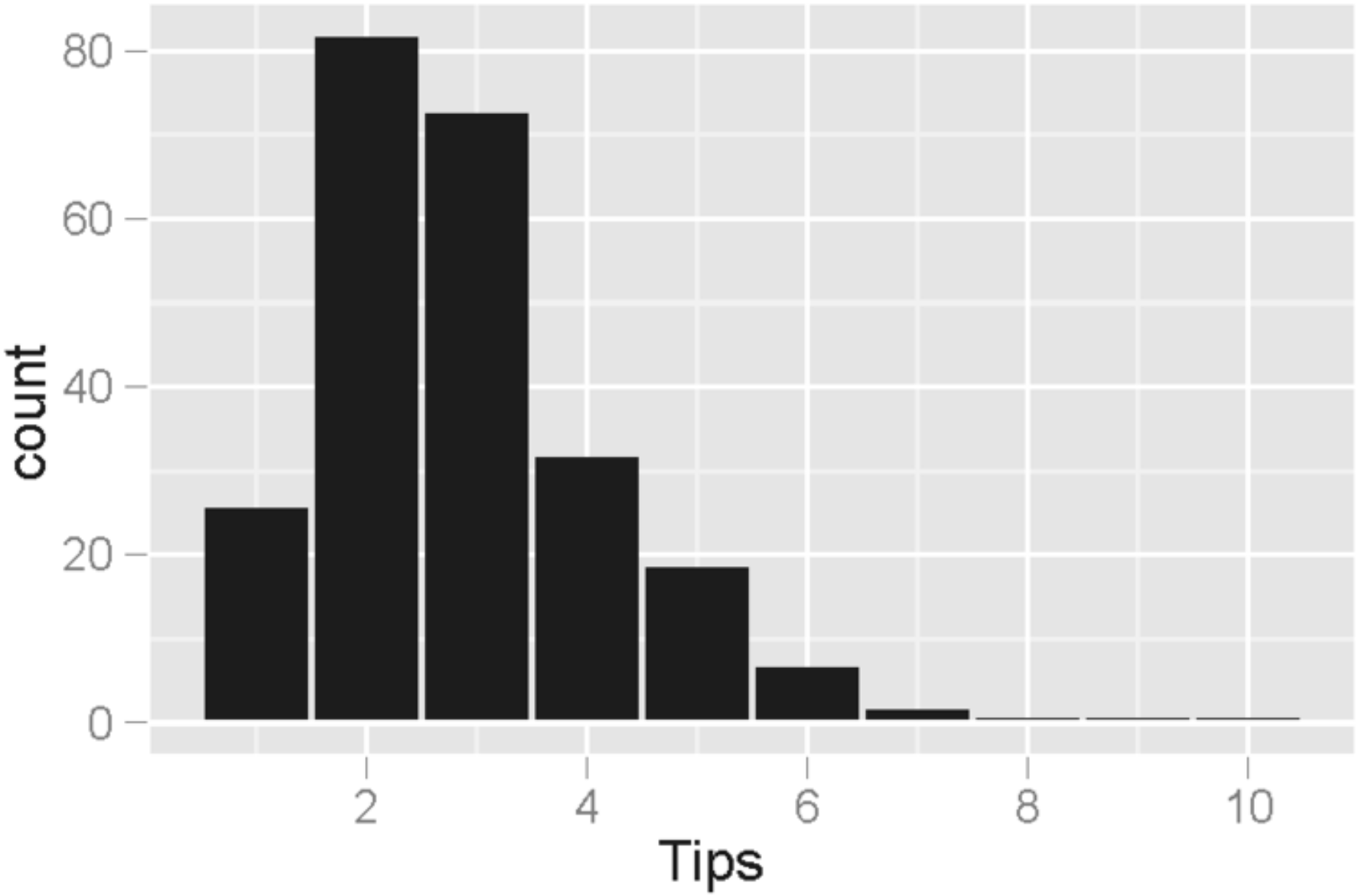


Encoding data in graphics

- There are many ways to encode information in graphical properties:
  - **spatial position**
  - angle
  - one-dimensional length
  - two-dimensional area
  - three-dimensional volume
  - color saturation

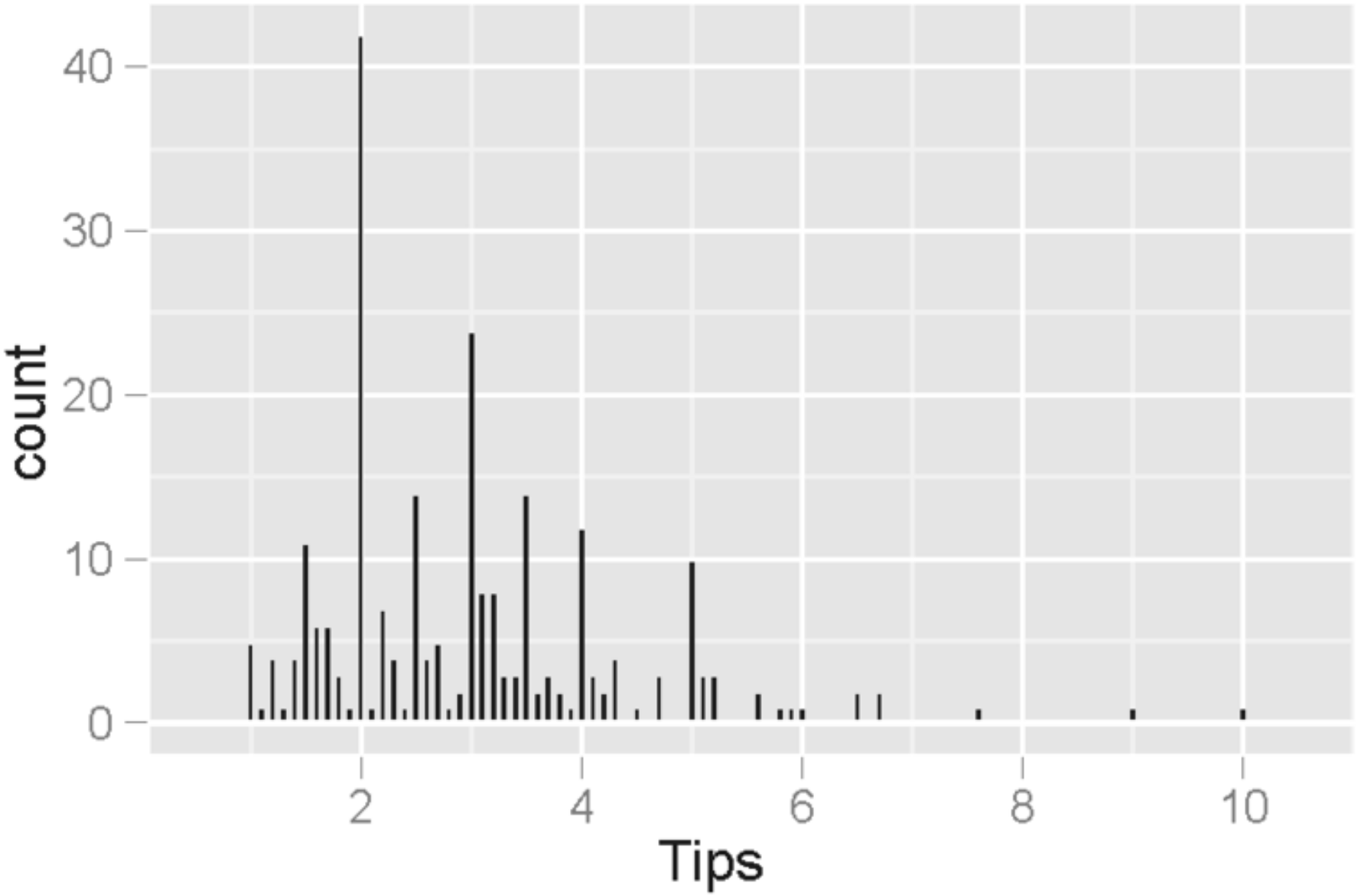
What to visualize?

source: Wikipedia



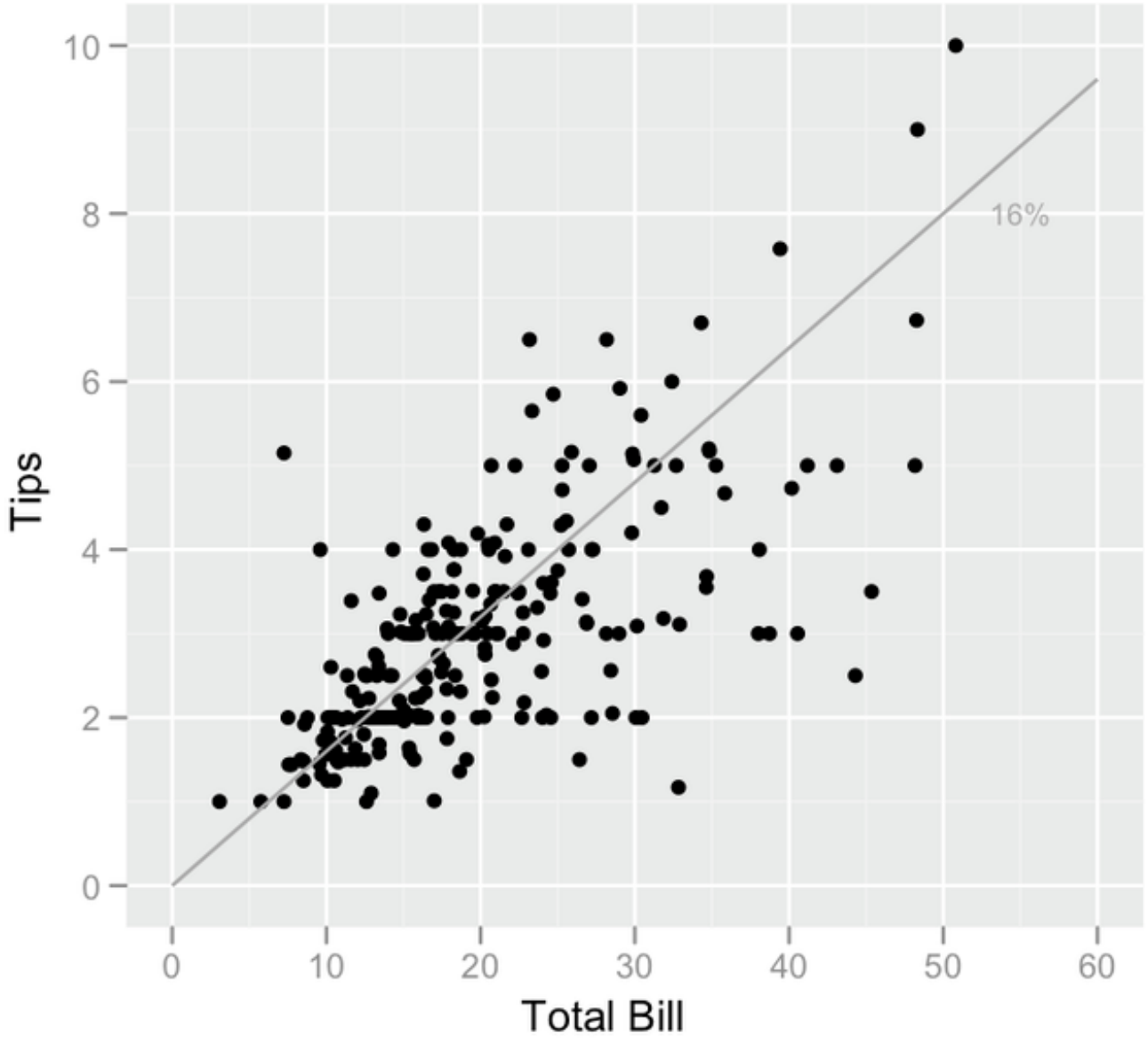
What to visualize?

source: Wikipedia



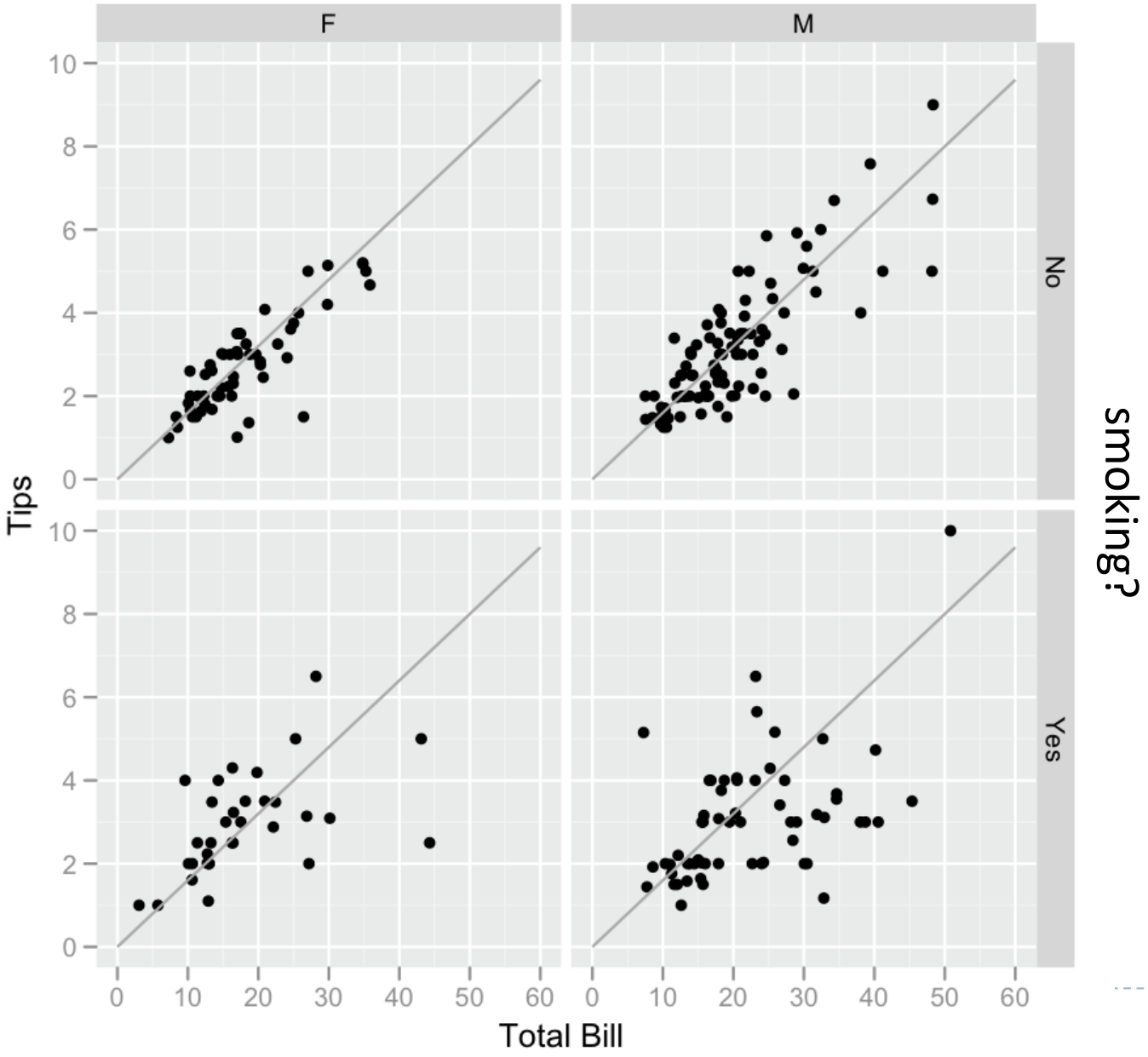
What to visualize?

source: Wikipedia

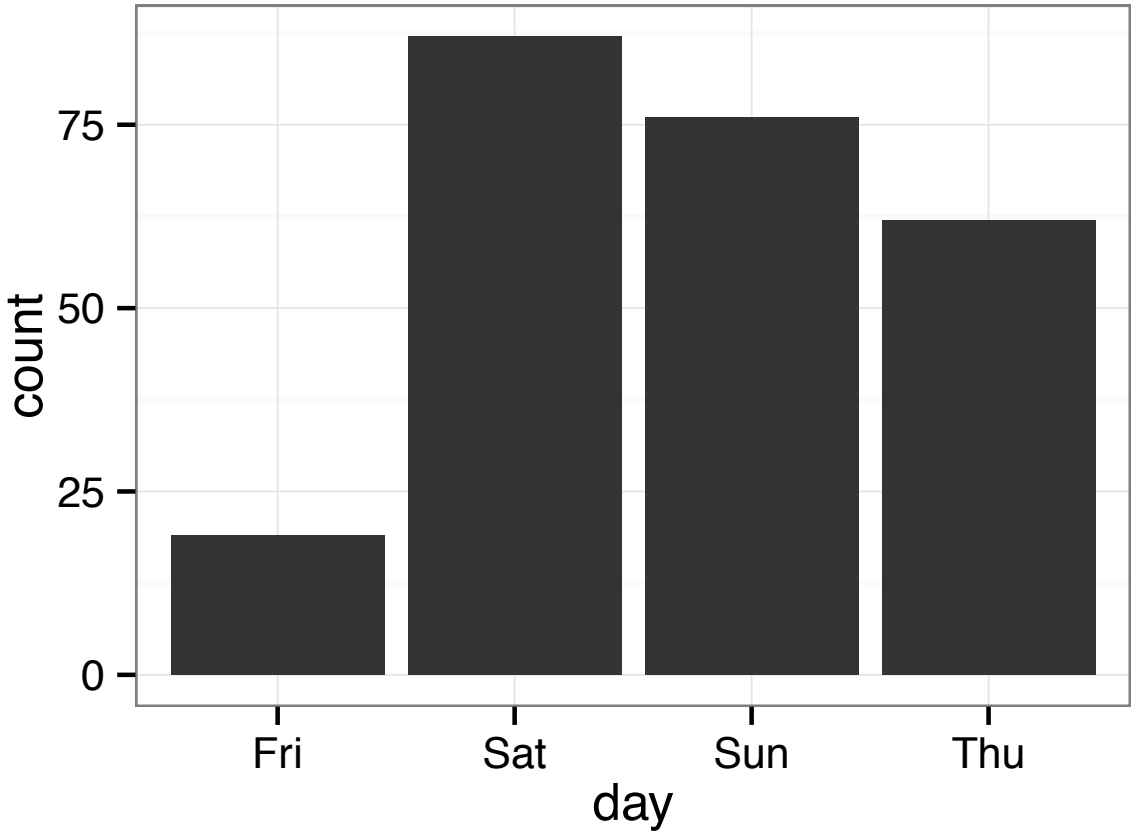


What to visualize?

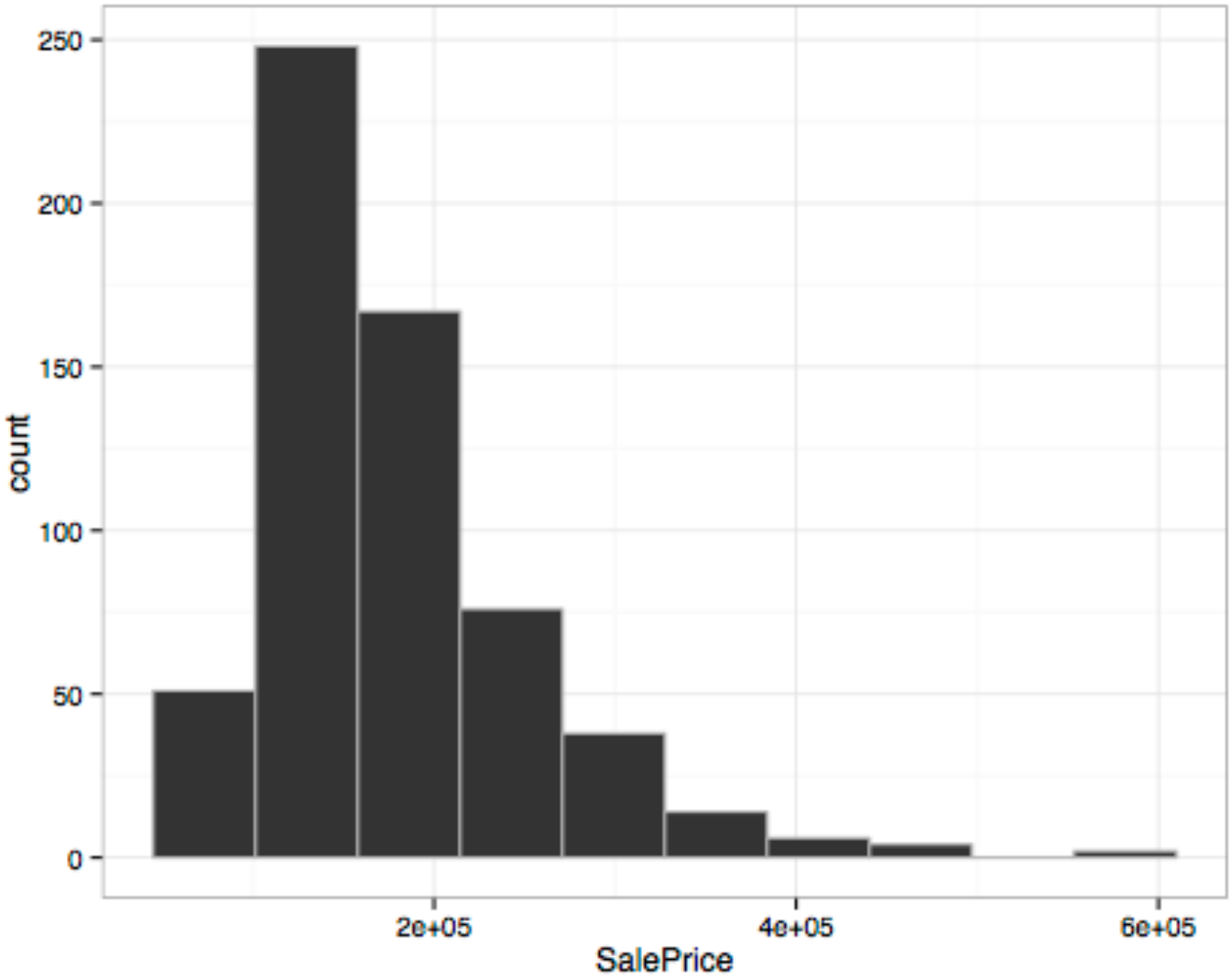
source: Wikipedia



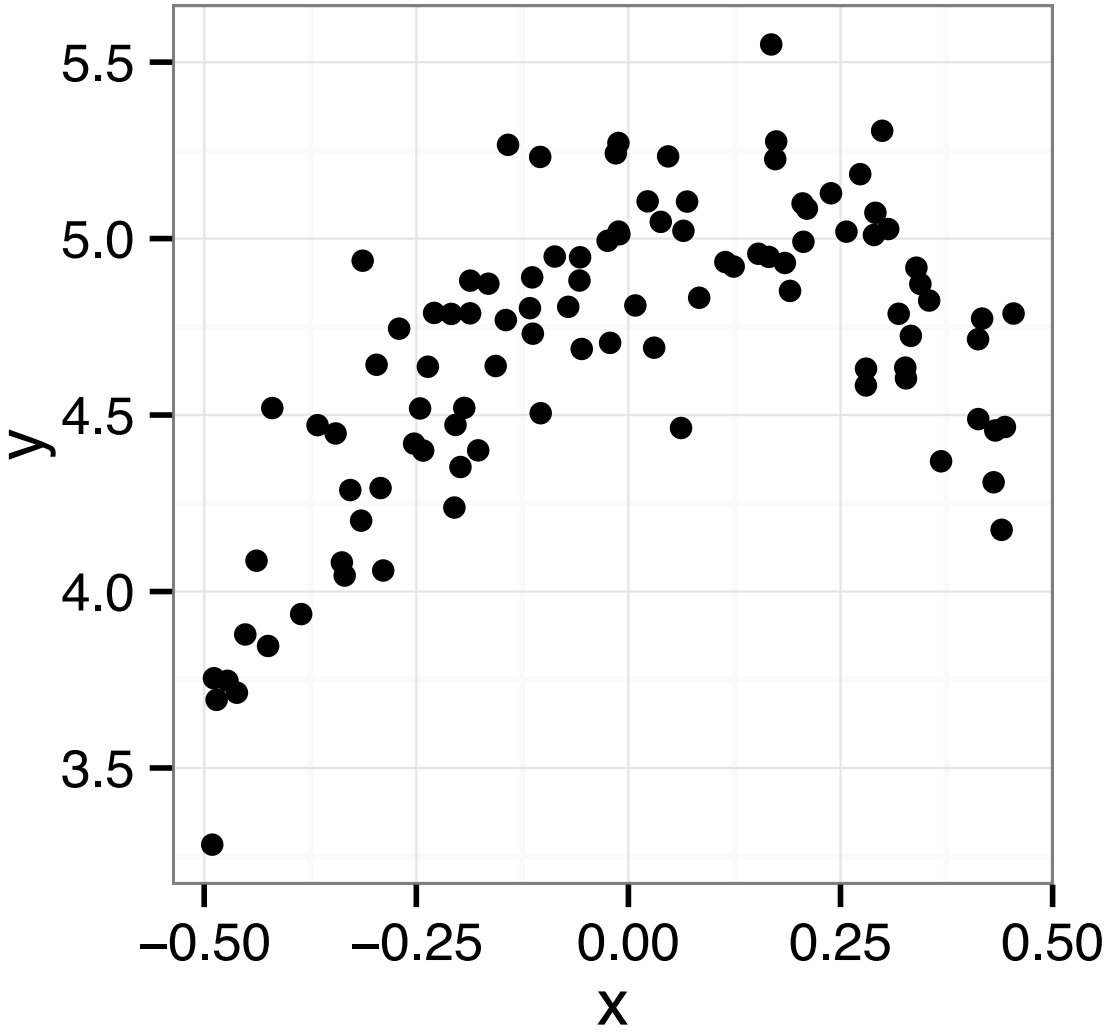
What data do we need for this plot?



What data for this plot?



What data for this plot?

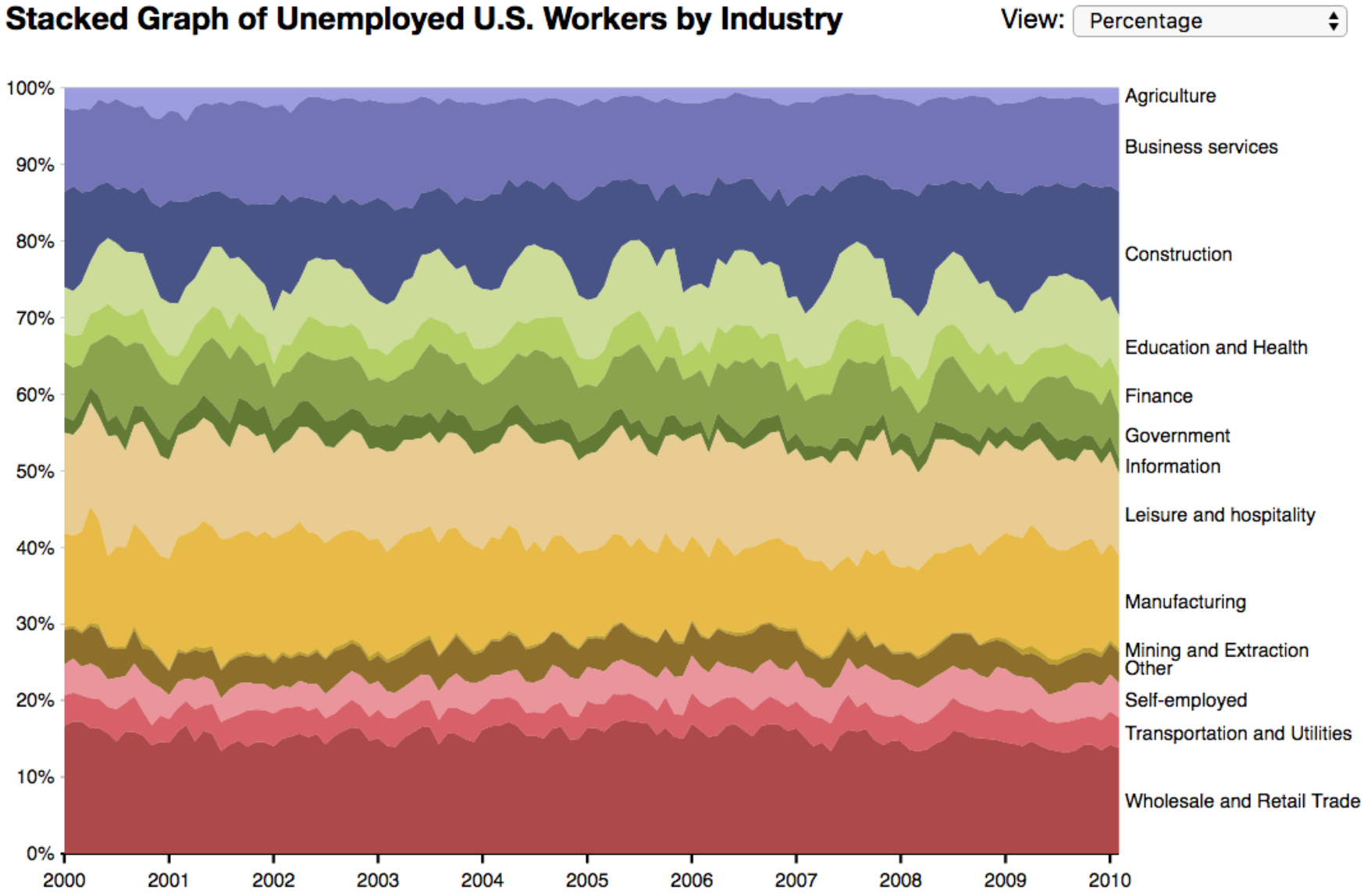


What data underlies this plot?

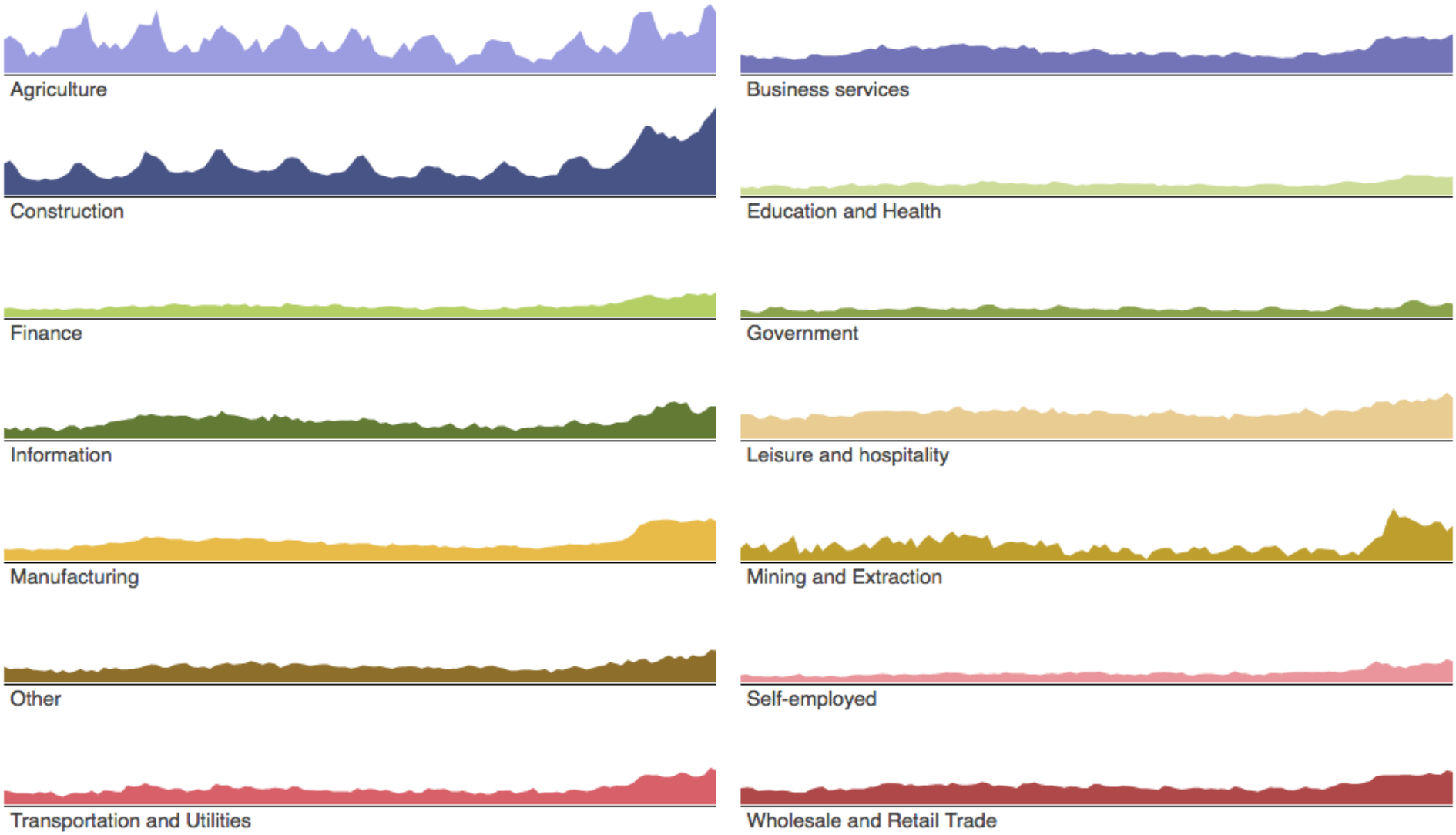
Types of visualizations

- The examples that follow are taken from this article:
- "A tour through the visualization zoo" by Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky.
  - http://goo.gl/MxQU3E

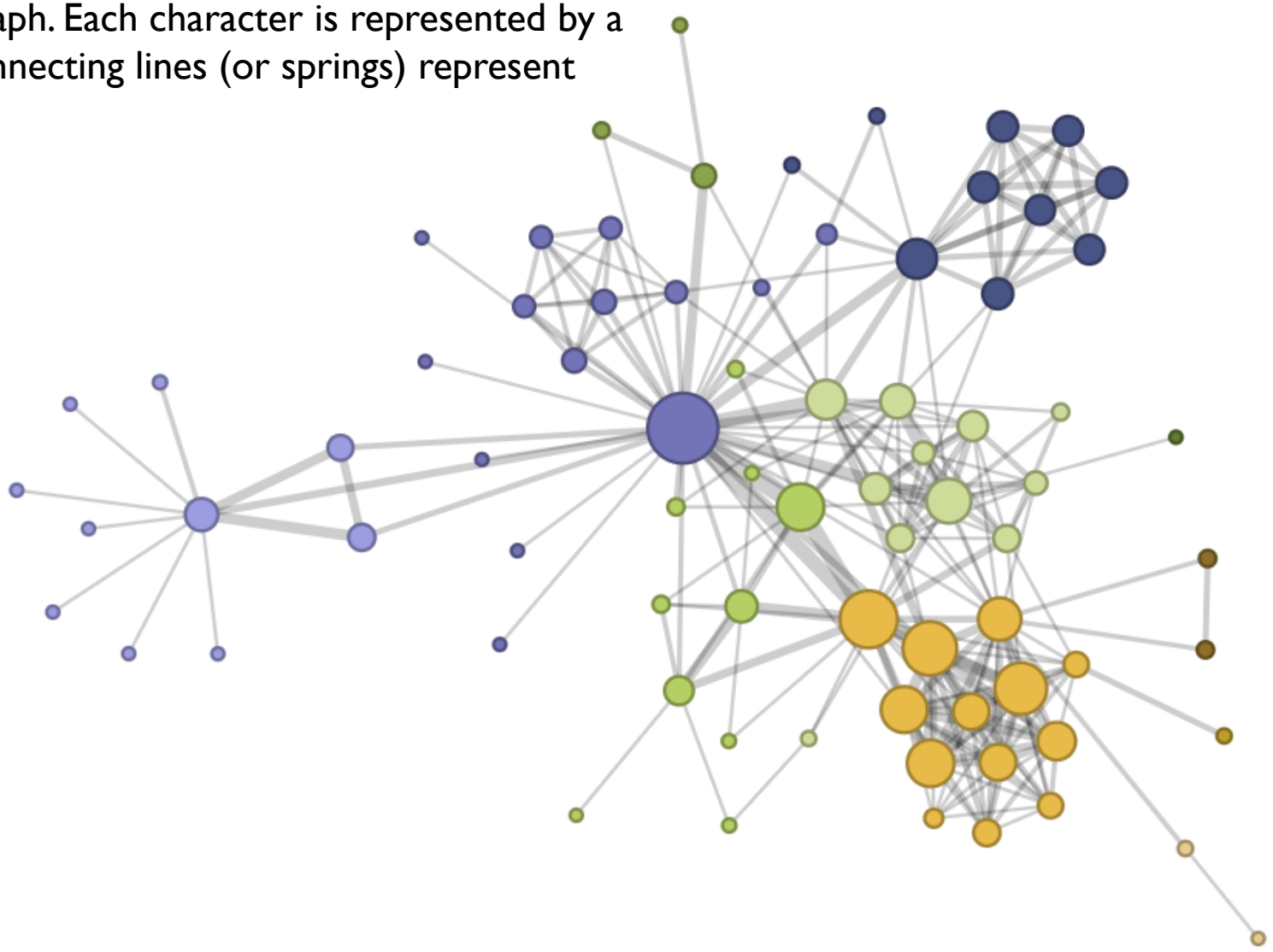
  (http://queue.acm.org/detail.cfm?id=1805128)


Stacked Graph of Unemployed U.S. Workers by Industry
View:
Percentage
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
Agriculture
Business services
Construction
Education and Health
Finance
Government
Information
Leisure and hospitality
Manufacturing
Mining and Extraction
Other
Self-employed
Transportation and Utilities
Wholesale and Retail Trade

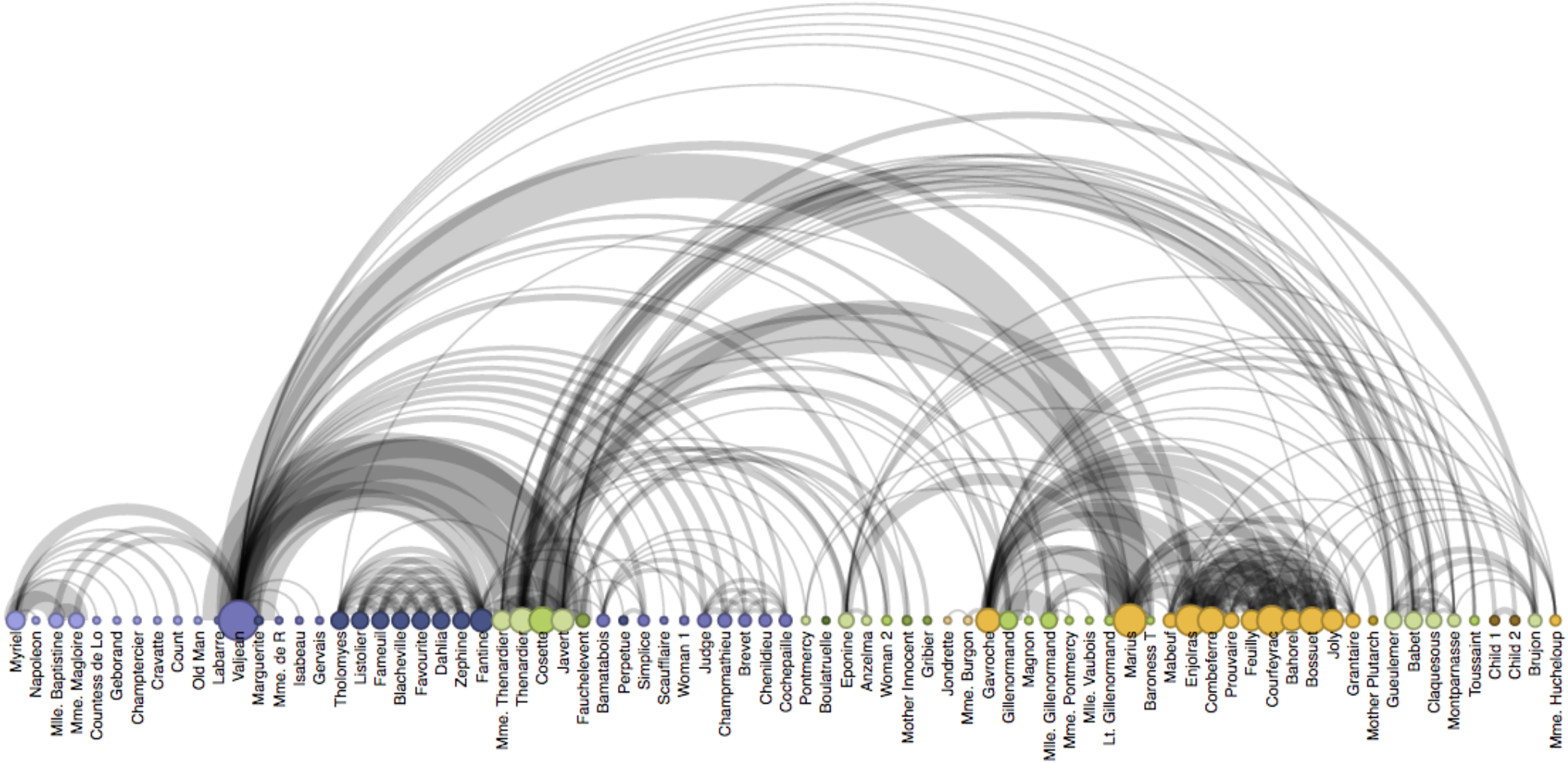
Unemployment Rate of U.S. Workers by Industry, 2000-2010



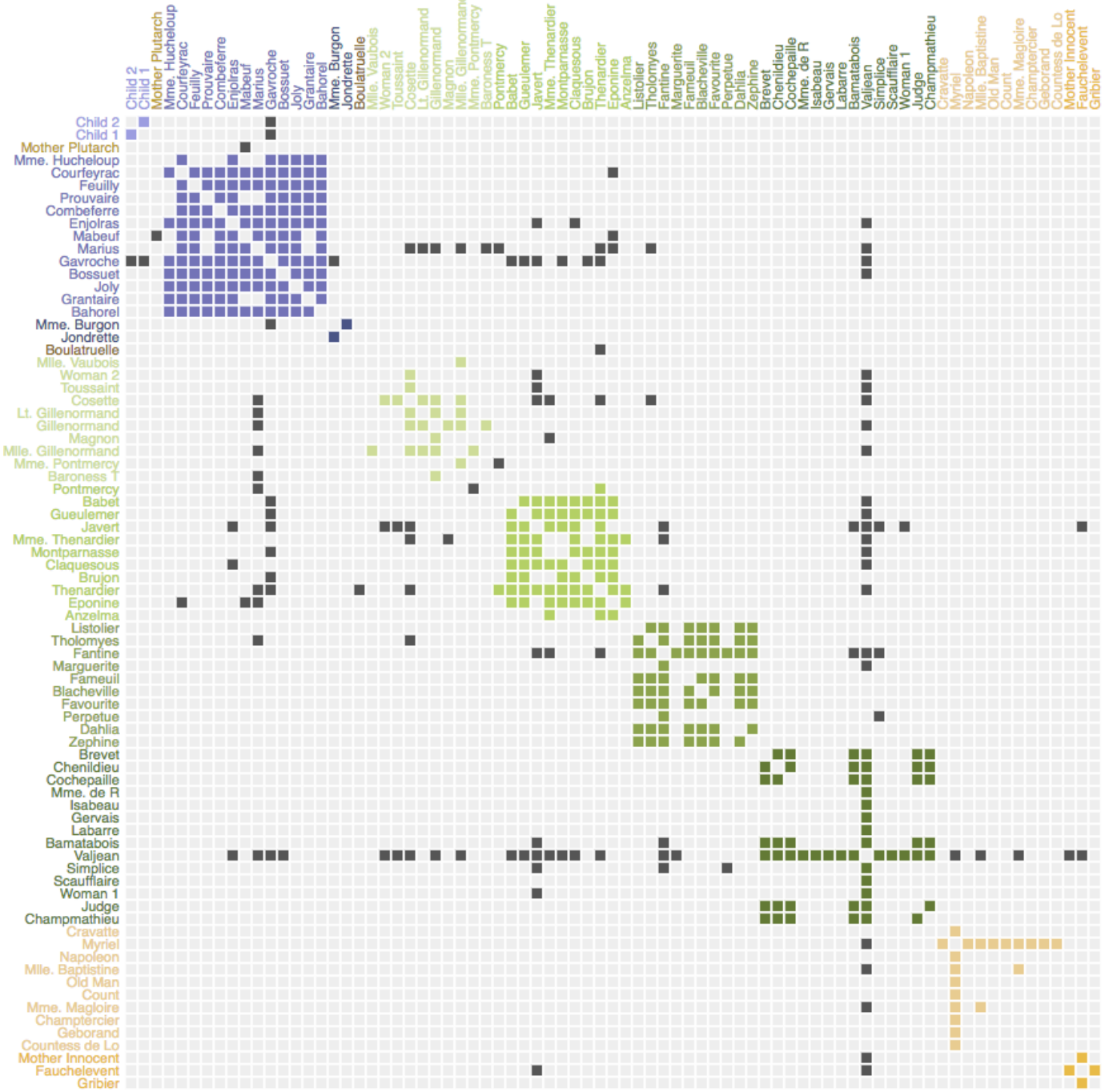
Les Misérables character interaction presented as a force directed graph. Each character is represented by a circle and the connecting lines (or springs) represent interaction.



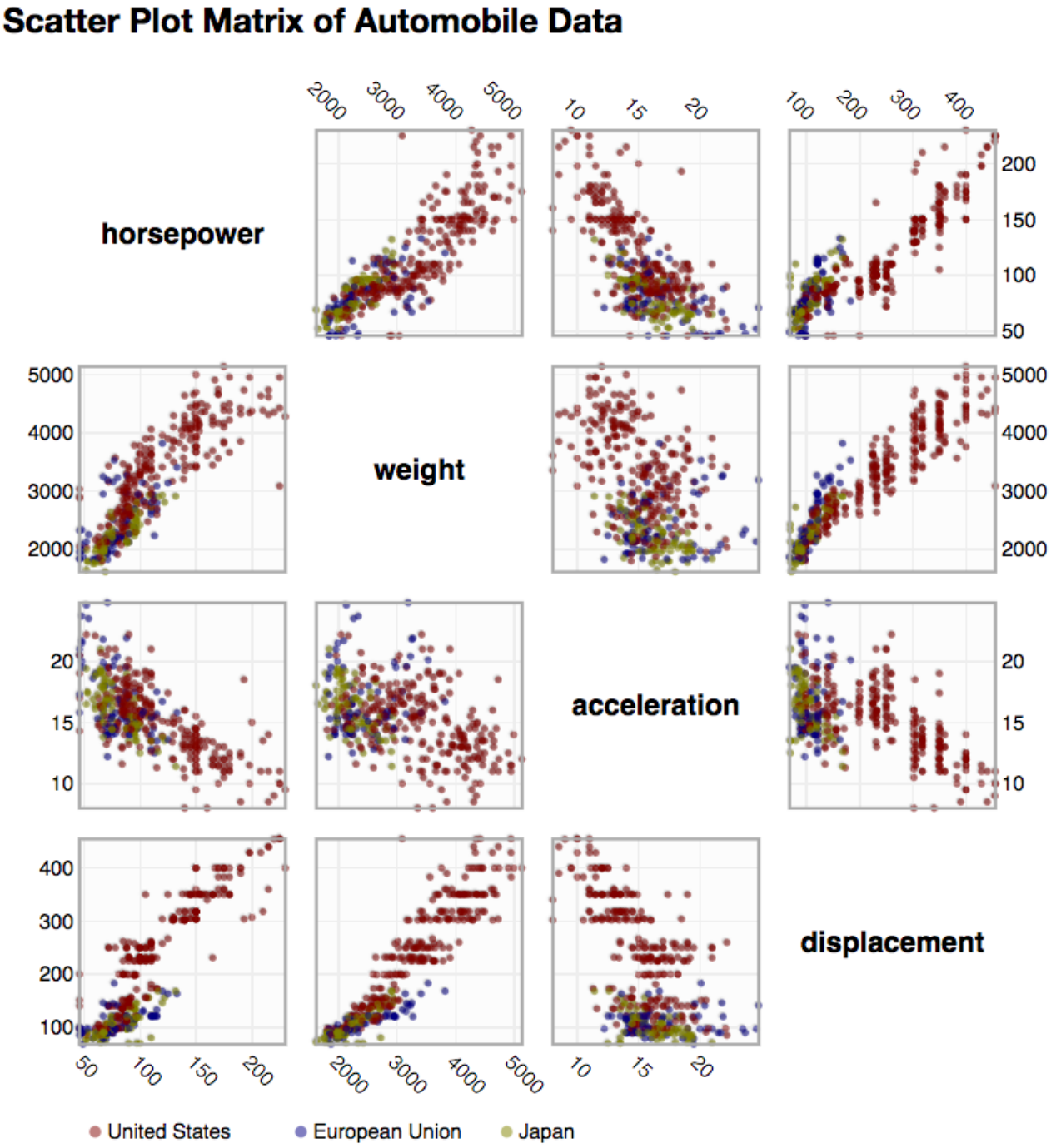
Les Misérables character interaction. Each character is represented by a circle and the connecting arc represents co-occurrence in a chapter. The character's size indicates the number of appearances they have over the entire work.



Les Misérables characters presented in an interaction matrix. Each character is represented by a row and a column in the matrix. An entry in the matrix is colored if it's row and column characters interact.



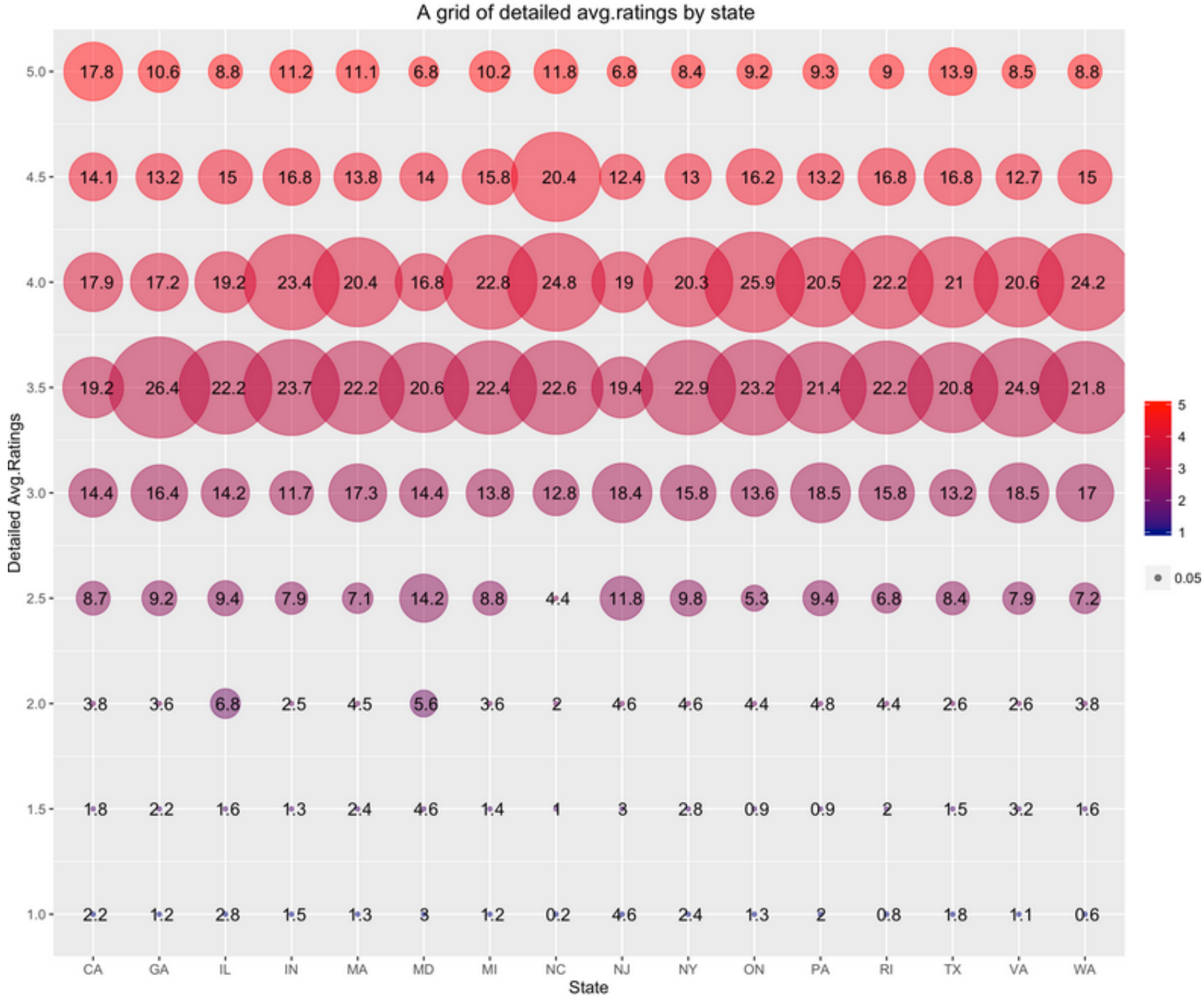
Four dimensions of a database of cars plotted in a scatter plot matrix, with different colors to indicate the country of origin. Each pair of variables is represented in two (transposed) plots.



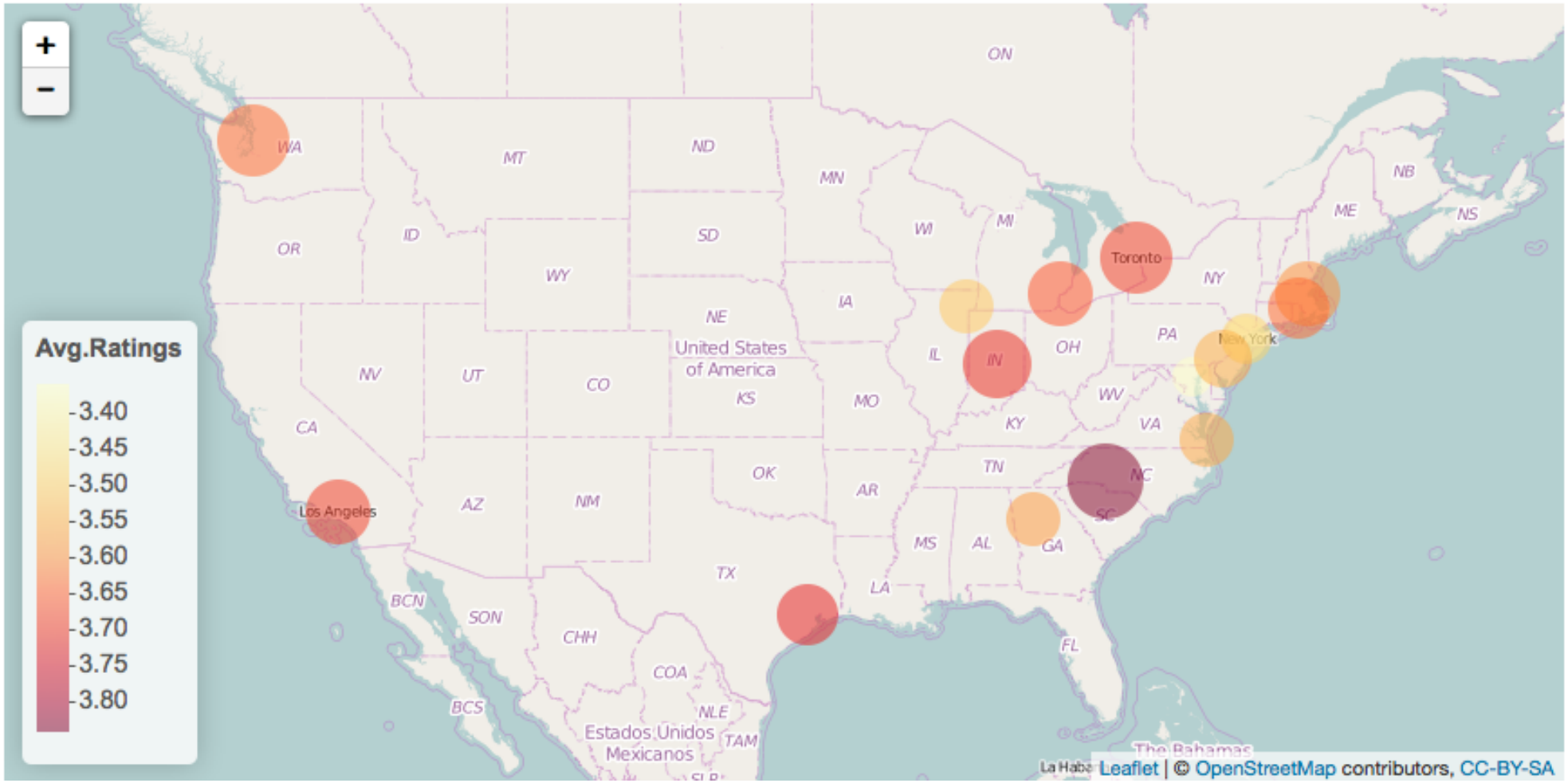

yelp

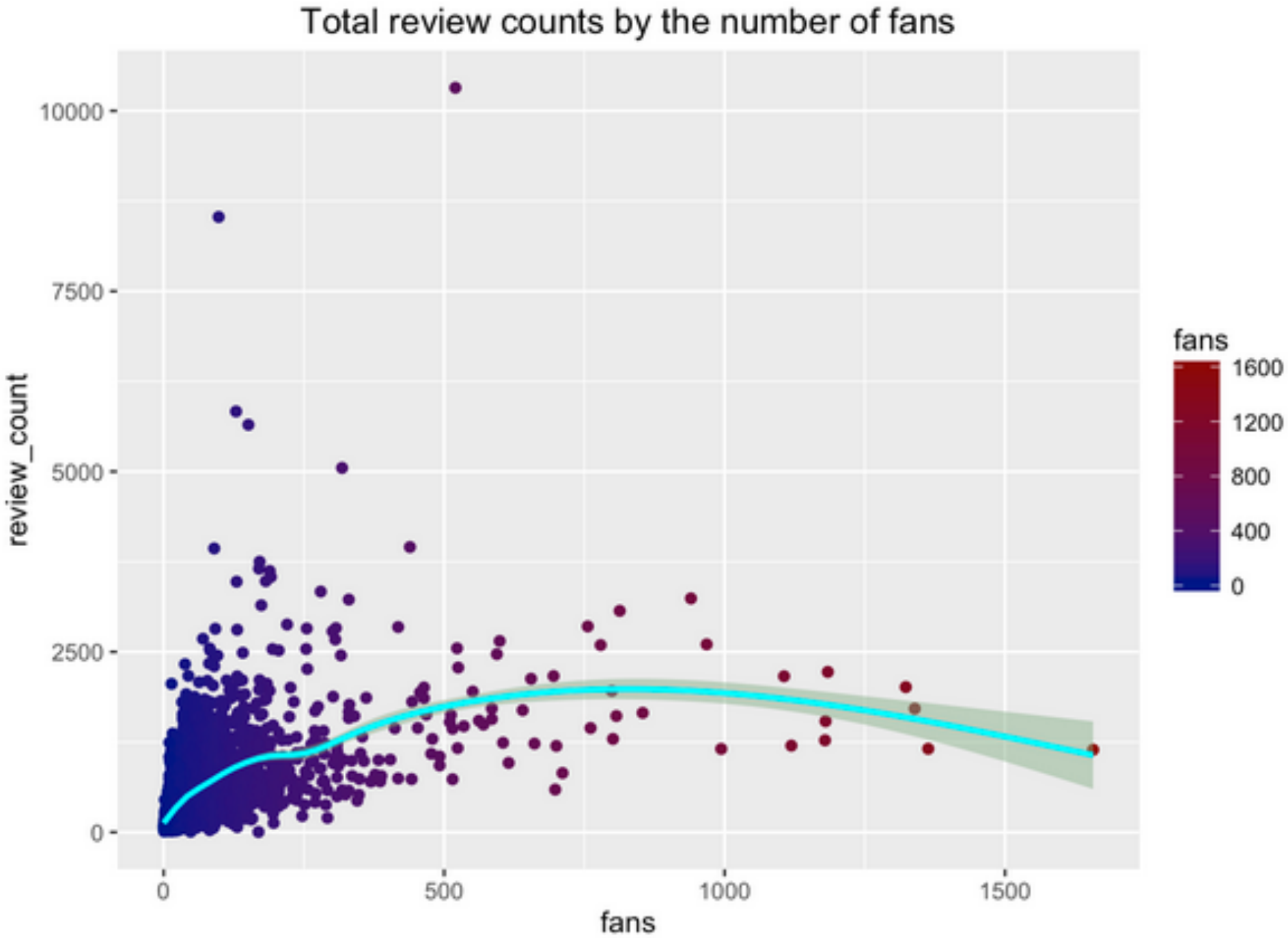
Topics suggested by Yelp

- **Cultural Trends**: By adding a diverse set of cities, we want participants to compare and contrast what makes a particular city different. For example, are people in international cities less concerned about driving in to a business, indicated by their lack of mention about parking? What cuisines are Yelpers raving about in these different countries? Do Americans tend to eat out late compared to the Germans and English? In which countries are Yelpers sticklers for service quality? In international cities such as Montreal, are French speakers reviewing places differently than English speakers?
- **Location Mining and Urban Planning**: How much of a business' success is really just location, location, location? Do you see reviewers' behavior change when they travel?
- **Seasonal Trends**: What about seasonal effects: Are HVAC contractors being reviewed just at onset of winter, and manicure salons at onset of summer? Are there more reviews for sports bars on major game days and if so, could you predict that?
- **Infer Categories**: Do you see any non-intuitive correlations between business categories e.g., how many karaoke bars also offer Korean food, and vice versa? What businesses deserve their own subcategory (i.e., Szechuan or Hunan versus just "Chinese restaurants"), and can you learn this from the review text?
- **Changepoints and Events**: Can you detect when things change suddenly (i.e. a business coming under new management)? Can you see when a city starts going nuts over cronuts?
- **Social Graph Mining**: Can you figure out who the trend setters are and who found the best waffle joint before waffles were cool? How much influence does my social circle have on my business choices and my ratings?



source: http://blog.nycdatascience.com/r/project-1-exploratory-visualizations-of-yelp-academic-dataset-draft/


Avg.Ratings
3.40
3.45
3.50
3.55
3.60
3.65
3.70
3.75
3.80
Toronto
Los Angeles
New York
United States of America
Estados Unidos Mexicanos
The Bahamas
Leaflet | © OpenStreetMap contributors, CC-BY-SA


Total review counts by the number of fans
review_count
10000
7500
5000
2500
0
0
500
1000
1500
fans
fans
1600
1200
800
400
0

Warm-start bias in reviews

- “Yelp ratings are often viewed as a reputation metric for local businesses. In this paper we study how Yelp ratings evolve over time. Our main finding is that on average the first ratings that businesses receive overestimate their eventual reputation.”
  - In particular, the first review that a business receives in our dataset averages 4.1 stars, while the 20th review averages just 3.69 stars.