Departamento de Ingeniería Informática

Exploratory Data Analysis

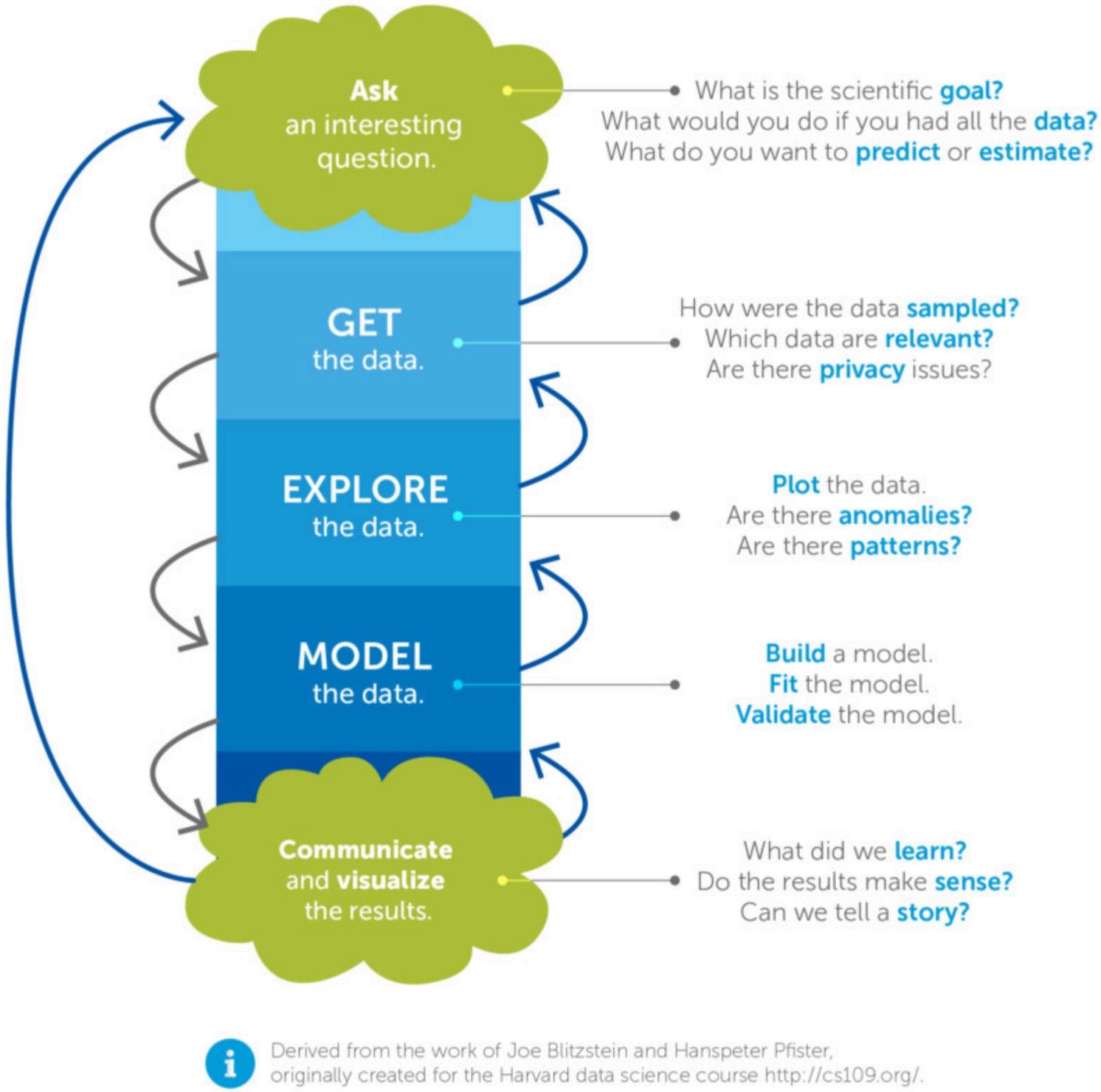
Guillermo Cabrera-Vives
guillecabrera@udec.cl

Based on work by Weiwei Pen, Edward Tufte

Announcements

- Lectures on Thursday 24/8 and 31/8 postponed.
- Next Tuesday 5/9: paper presentation
  - 15 minutes for presenting + 5 minutes for questions
  - Read paper, present, prepare small quiz (5 mins), and grade.
  - Grades: 50% presentation, 50% quizzes
  - Next week: Monserrat & Mabel

The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

https://www.class-central.com/report/best-intro-data-science-courses/

Data

- Get Data
- Explore data
  - Descriptive statistics
  - Data Visualization

What is data?

- “A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data is multiple such measurements.”

- Provocative claim: everything is (can be) data!

Where does data come from?

- **Internal sources:** already collected by or is part of the overall data collection of you organization.
    - For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data.
- **Existing External Sources:** available in ready to read format from an outside source for free or for a fee.
    - For example: public government databases, stock market data, Yelp reviews
- **External Sources Requiring Collection Efforts:** available from external source but acquisition requires special processing.
    - For example: data appearing only in print form, or data on websites

How to get data generated, published or hosted online

- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use.
  - For example: Google Map API, Facebook API, Twitter API
- **RSS (Rich Site Summary):** summarizes frequently updated online content in standard format. Free to read if the site has one.
  - For example: news-related sites, blogs
- **Web scraping:** using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file.

Web scraping

- **Why do it?** Older government or smaller news sites might not have APIs for accessing data, or publish RSS feeds or have databases for download. You don't want to pay to use the API or the database.
- **How do you it?** Scrapping libraries for pulling data out of HTML and XML files. (e.g. BeautifulSoup)
- **Should you do it?**
  - **You just want to explore:** Are you violating their terms of service? Privacy concerns for website and their clients?
  - **You want to publish your analysis or product:** Do they have an API or fee that you're bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

What does data look like?

- What kind of values are in your data (data types)?
- Simple or atomic:
  - **Numeric:** integers, floats
  - **Boolean:** binary or true false values
  - **Strings:** sequence of symbols

What does data look like?

- Compound, composed of a bunch of atomic types:
  - **Date and time:** compound value with a specific structure
  - **Lists:** a list is a sequence of values
  - **Dictionaries:** A dictionary is a collection of key-value pairs, a pair of values x : y where x is usually a string called key representing the “name” of the value, and y is a value of any type.
    - **Example:** Student record
    - First: Weiwei
    - Last: Pan
    - Classes: [CS109A, STAT121A, AC209A]

What does data look like?

- How is your data represented and stored (data format)?
- **Tabular Data:** a dataset that is a two-dimensional table, where each row typically represents a single data record, and each column represents one type of measurement (csv, tsp, xlsx etc.).
- **Structured Data:** each data record is presented in a form of a, possibly complex and multi-tiered, dictionary (json, xml etc.)
- **Semistructured Data:** not all records are represented by the same set of keys or some data records are not represented using the key-value pair structure.

Tabular data

- In tabular data, we expect each record or observation to represent a set of measurements of a single object or event.

	Hight	Radius	Do I Like It?
Cylinder # 1	10	5	Yes
Cylinder # 2	3	7.5	No

- Each type of measurement is called a variable or an attribute of the data (e.g. Height, Radius and “Do I Like It?” are variables or attributes). The number of attributes is called the dimension of the data.
- We expect each table to contain a set of records or observations of the same kind of object or event (e.g. our table above contains observations of cylinders).

Tabular data

- You'll see later that it's important to distinguish between classes of variables or attributes based on the type of values they can take on.
- **Quantitative variable:** is numerical and can be
  - **discrete:** a finite number of values are possible in any bounded interval
    - **For example:** "Number of siblings" is a discrete variable.
  - **Continuous:** an infinite number of values are possible in any bounded interval
    - **For example:** "Temperature" is a continuous variable
- **Categorical variable:** no inherent order among the values
  - **For example:** "What kind of pet you have" is a categorical variable

Is the data any good?

- Common issues with data:
  - **Missing values:** how do we fill in?
  - **Wrong values:** how can we detect and correct?
  - **Messy format**
  - **Not usable:** the data cannot answer the question posed

Messy data

- The following is a table accounting for produce deliveries over a weekend.
- What are the variables in this dataset?
- What object or event are we measuring?

	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Messy data

- We're measuring individual deliveries; the variables are Time, Day, Number of Produce.
- Problem: each column header represents a single value rather than a variable. Row headers are "hiding" the Day variable. The values of the variable, "Number of Produce", is not recorded in a single column.

	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Messy data

- We need to reorganize the information to make explicit the event we're observing and the variables associated to this event.

Delivery	Time	Day	No. of Produce
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	90
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45

Messy data

- What object or event are we measuring?
- What are the variables in this dataset?

Delivery	Amount
On Sunday	
10:30	43
12:30	12
12:35	30
On Monday	
11:30	29
11:57	87
11.59	63
On Tuesday	
11:33	19
11:15	27
12.59	54

Messy data

- We're measuring individual deliveries; the variables are Time, Day, Number of Produce:

Days	times	Amount
Sunday	10:30	43
Sunday	12:30	12
Sunday	12:35	30
Monday	11:30	29
Monday	11:57	87
Monday	11.59	63
Tuesday	11:33	19
Tuesday	11:15	27
Tuesday	12.59	54

Messy data

- Common causes of messiness are:
  - Column headers are values, not variable names
  - Variables are stored in both rows and columns
  - Multiple variables are stored in one column
  - Multiple types of experimental units stored in same table
- In general, we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation.

Exploring Data

Describing Data

- Given some large dataset, we'd like to compute a few quantities that intuitively summarizes the data.

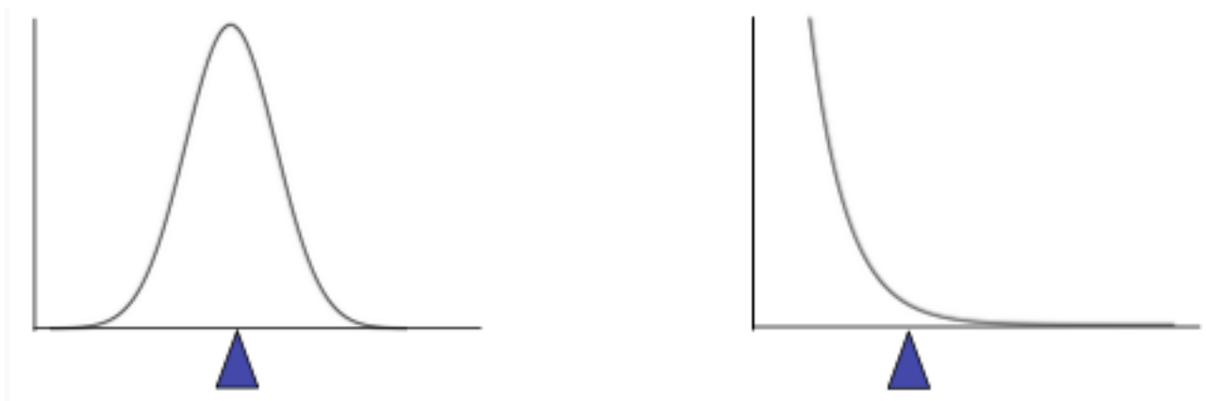
To begin with we'd like to know

- what are typical values for our variables or attributes?
- how representative are these typical values?

Centrality

- The **mean** of a set of *n* samples of a variable is defined by

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



- The mean describes what a “typical” sample value looks like, or where is the “center” of the distribution of the data.

Centrality

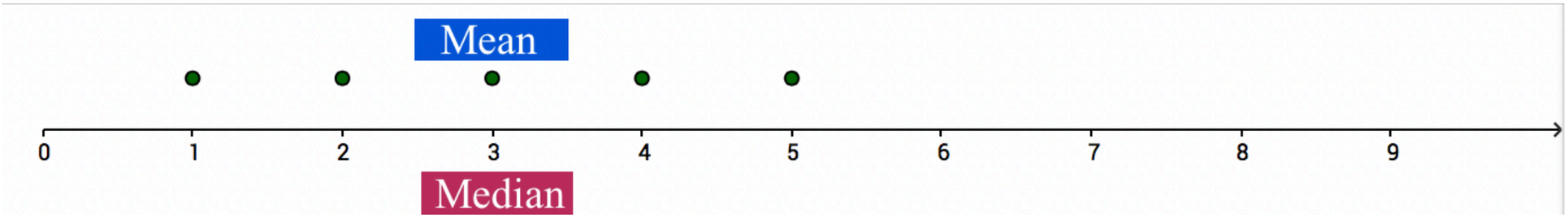
- The median of a set of *n* number of samples, ordered by value, of a variable is is defined by

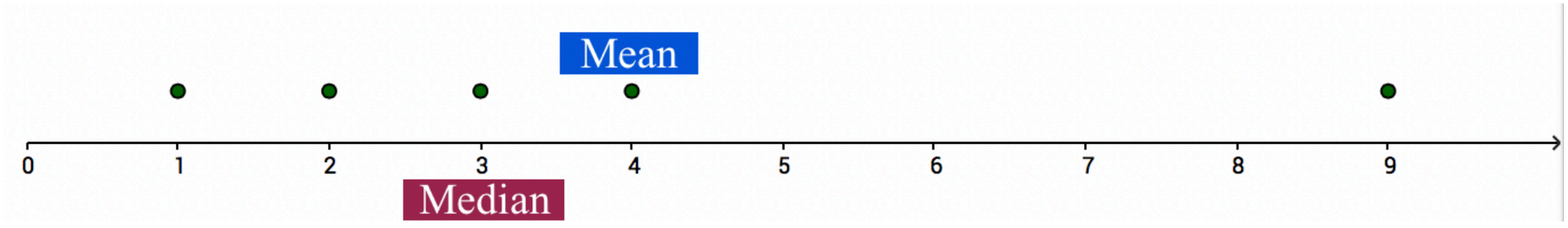
$$\text{Median} = \begin{cases} x_{\lfloor n/2 \rfloor + 1}, & \text{if } n \text{ is odd} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & \text{if } n \text{ is even} \end{cases}$$

- Example:
- Ages: 17, 19, 21, 22, 23, 23, 23, 38
- Median = (22+23) / 2 = 22.5
- The median describes what a “typical” sample looks like, or where is the “center” of the distribution of the samples.

Centrality

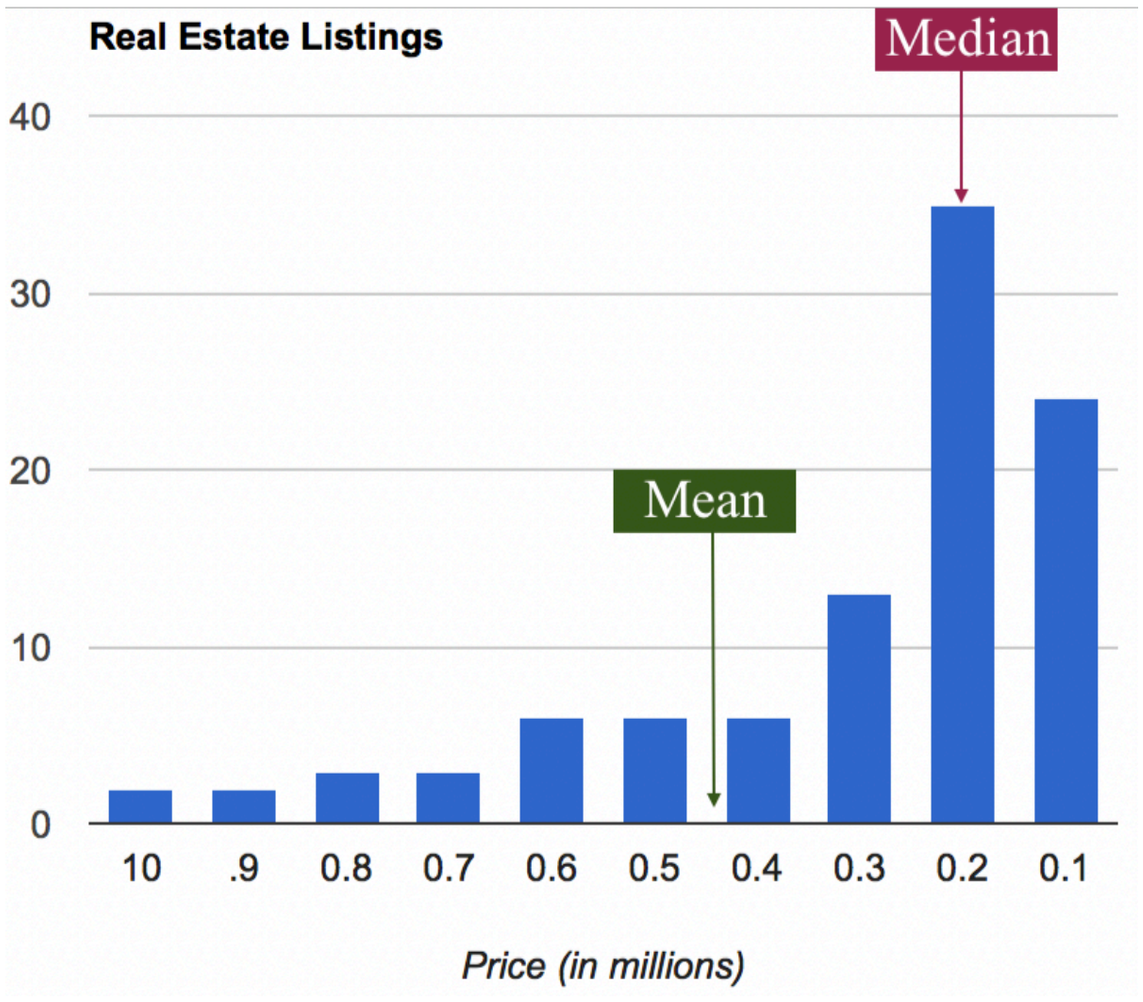
- The mean is sensitive to outliers





Centrality

- The mean is sensitive to skewness (asymmetry) of distributions.



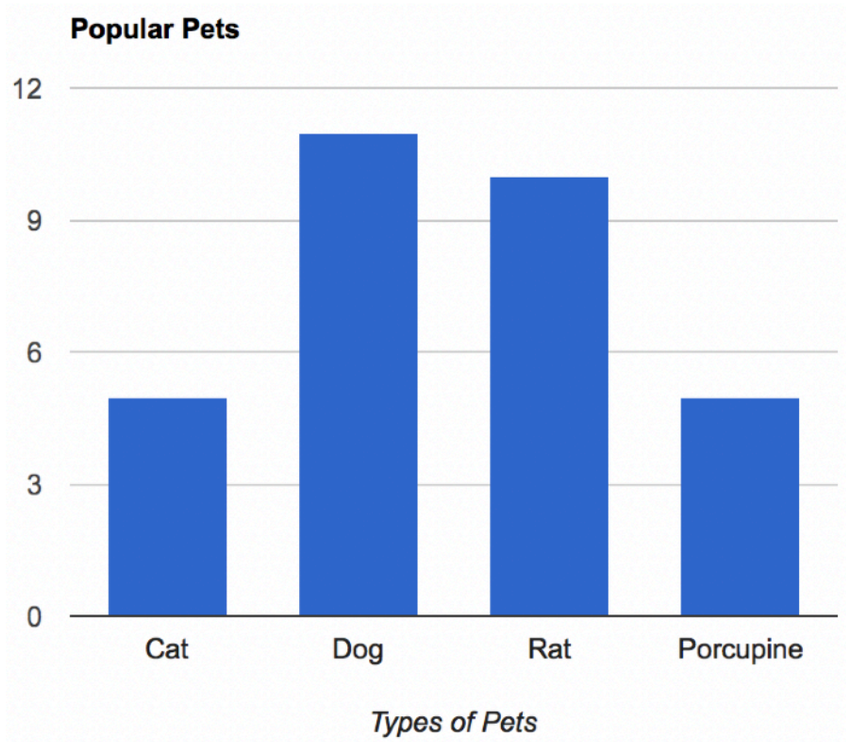
Centrality

- How hard (in terms of algorithmic complexity) is it to calculate
  - the mean: at most O(n)
  - the median: at leat O(n log n)

- Note: Practicality of implementation has to be considered!

Centrality

- For samples of categorical variables, neither mean or median make sense.



- The **mode** (value that appears most often) might be a better way to find the most “representative” value.

Spread

- The spread of samples measures how well the mean or median describes the sample set.
- One way to measuring spread of a set of samples is via the range.
  - Range = Maximum Value – Minimum Value

Spread

- The (sample) variance measures how much on average the sample values “deviates” from the mean

$$s^2 = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|^2}{n-1}$$

- **Note**: the term |xi – x| measure the amount by which xi deviates from the mean x. Squaring these deviation means that s2 is sensitive to extreme values (outliers).
- **Note**: s2 doesn’t have the same units as xi! What does a variance of 1, 008 mean? Or 0.0001?

Spread

- The (sample) standard deviation is the square root of the variance

$$s = \sqrt{\frac{\sum_{i=1}^{n} |x_i - \bar{x}|^2}{n-1}}$$

- Note: the standard deviation has the same units as x!

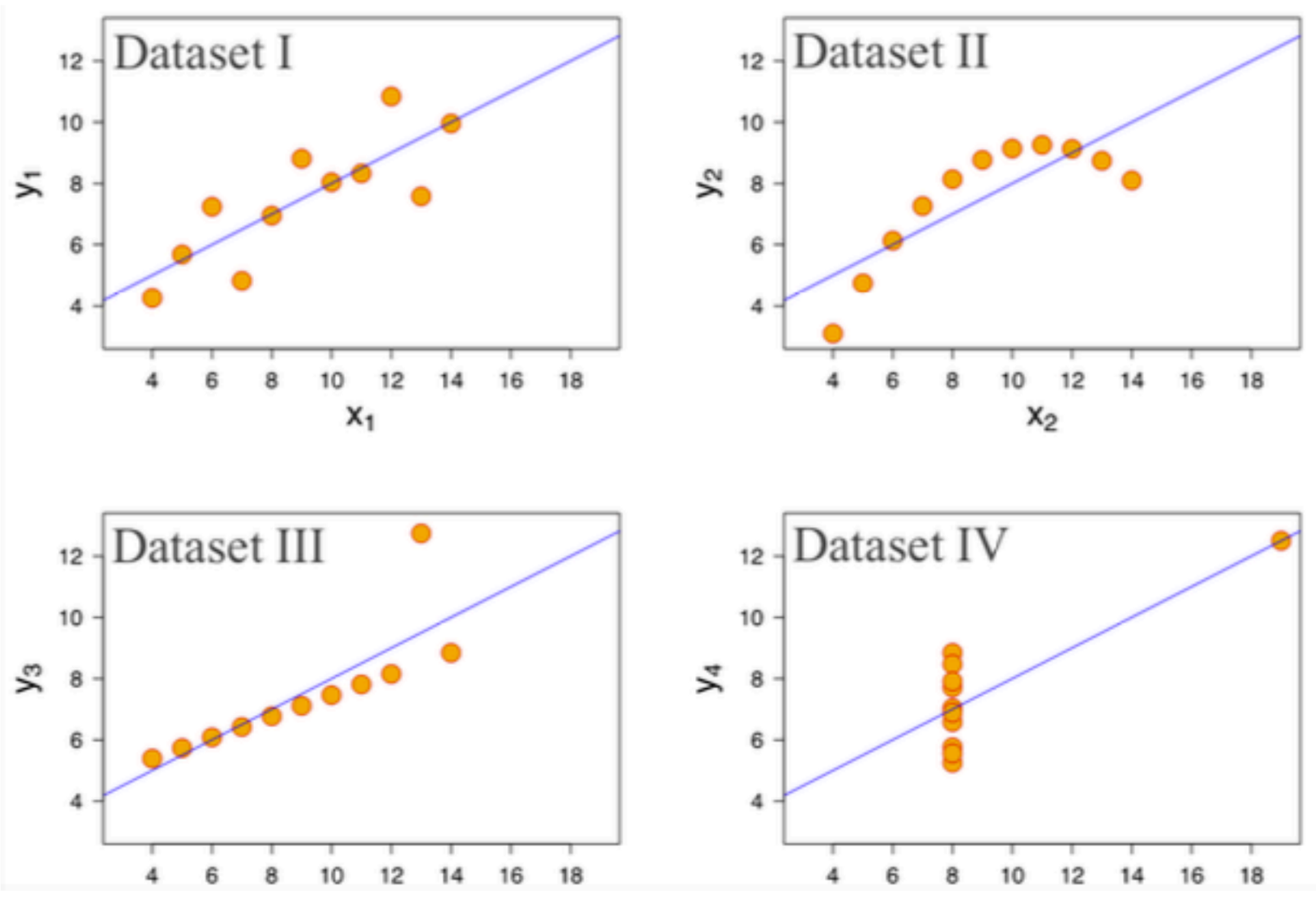
Data Visualization

- The following data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.

	Dataset I		Dataset II		Dataset III		Dataset IV	
	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	9.13	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
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03

Data Visualization

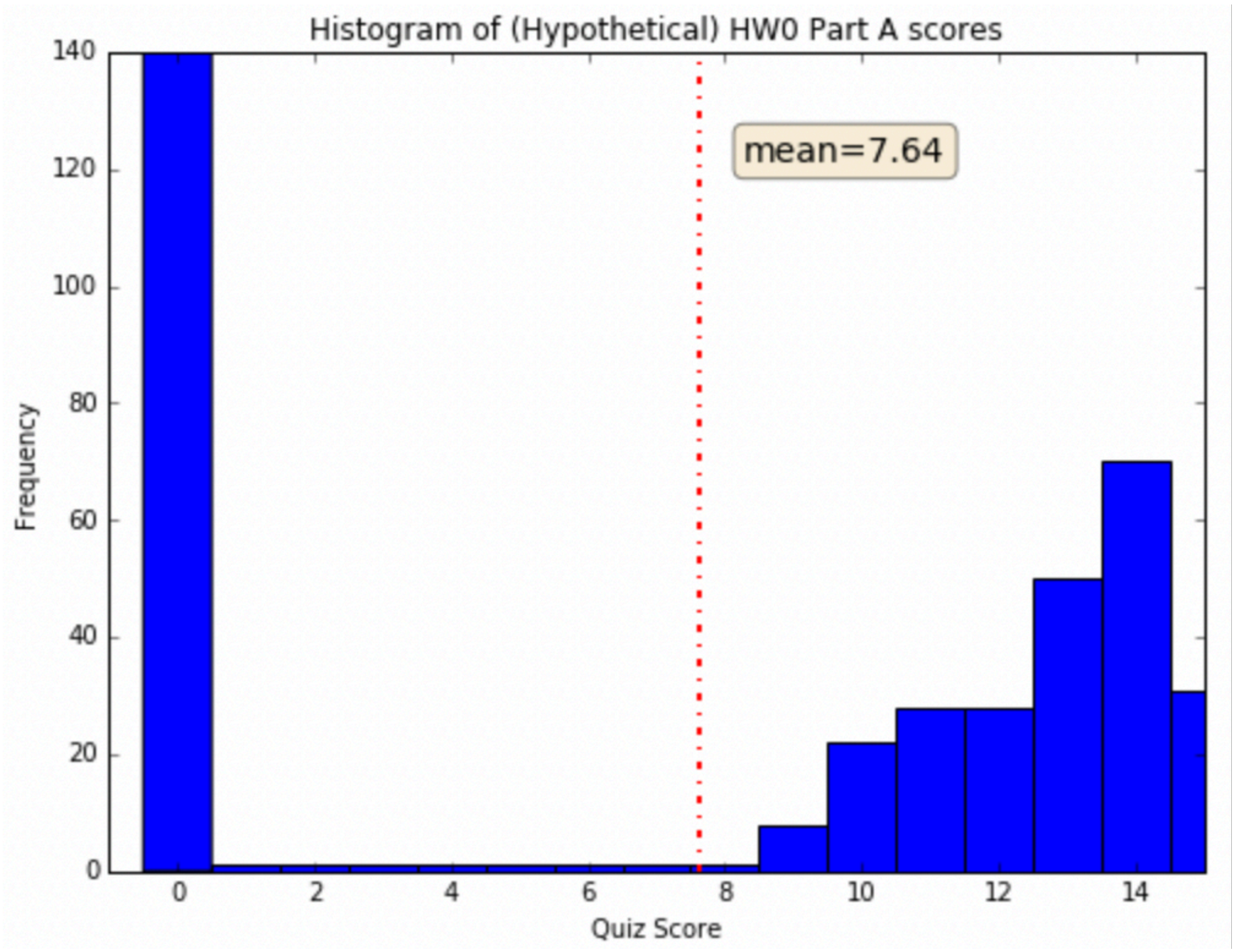
- The following data sets comprise the Anscombe's Quartet; all four sets of data have identical simple summary statistics.



Data Visualization

- If I tell you that the average score for Homework 0 Part A is: 7.64/15. What does that suggest?

Data Visualization



Data visualization

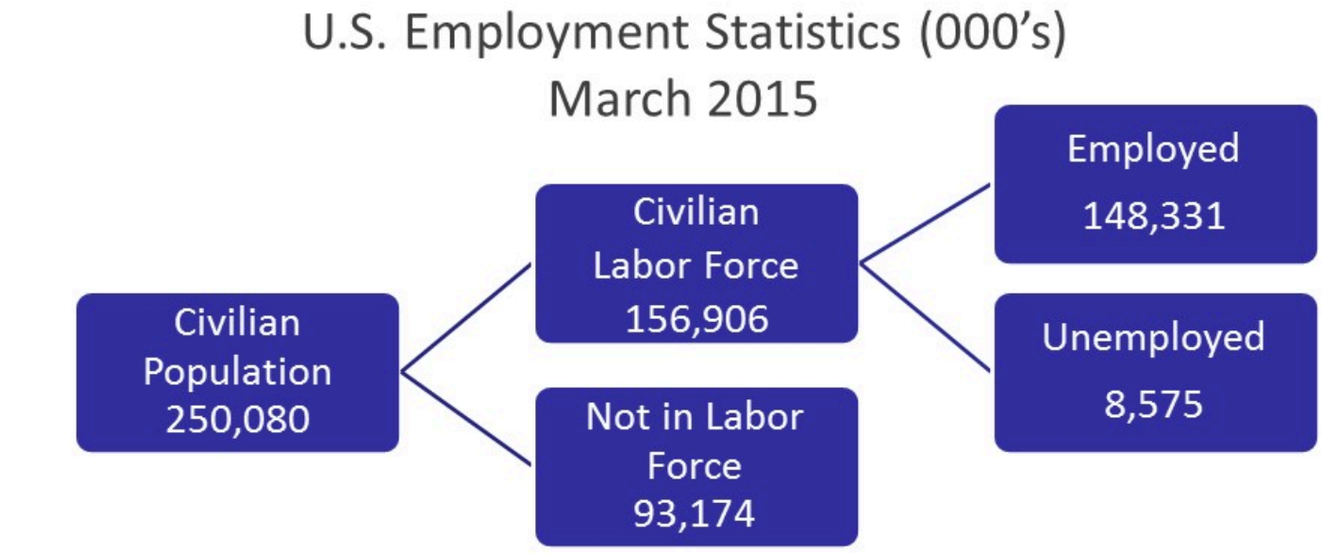
- Good for:
  - **Analyze:**
    - Identify hidden patterns and trends
    - Help formulate/test hypotheses
    - Help determine the next step in analysis/modeling
  - **Communicate:**
    - Present information and ideas succinctly
    - Provide evidence and support
    - Influence and persuade

Tufte's Criteria for Good Visual Information Representation

- Graphical Excellence
  - "the greatest number of ideas, in the shortest time, using the least amount of ink, in the smallest space."

Tufte's Criteria for Good Visual Information Representation

- The graphic below, relating to US employment statistics in March 2015, offers many ideas in a very small space and is easy to digest.



Unemployment Rate = Unemployed / Civilian Labor Force = 8,575 / 156,906 = 5.5%

Labor Force Participation Rate = Labor Force / Civilian Population = 156,906 / 250,080 = 62.7%

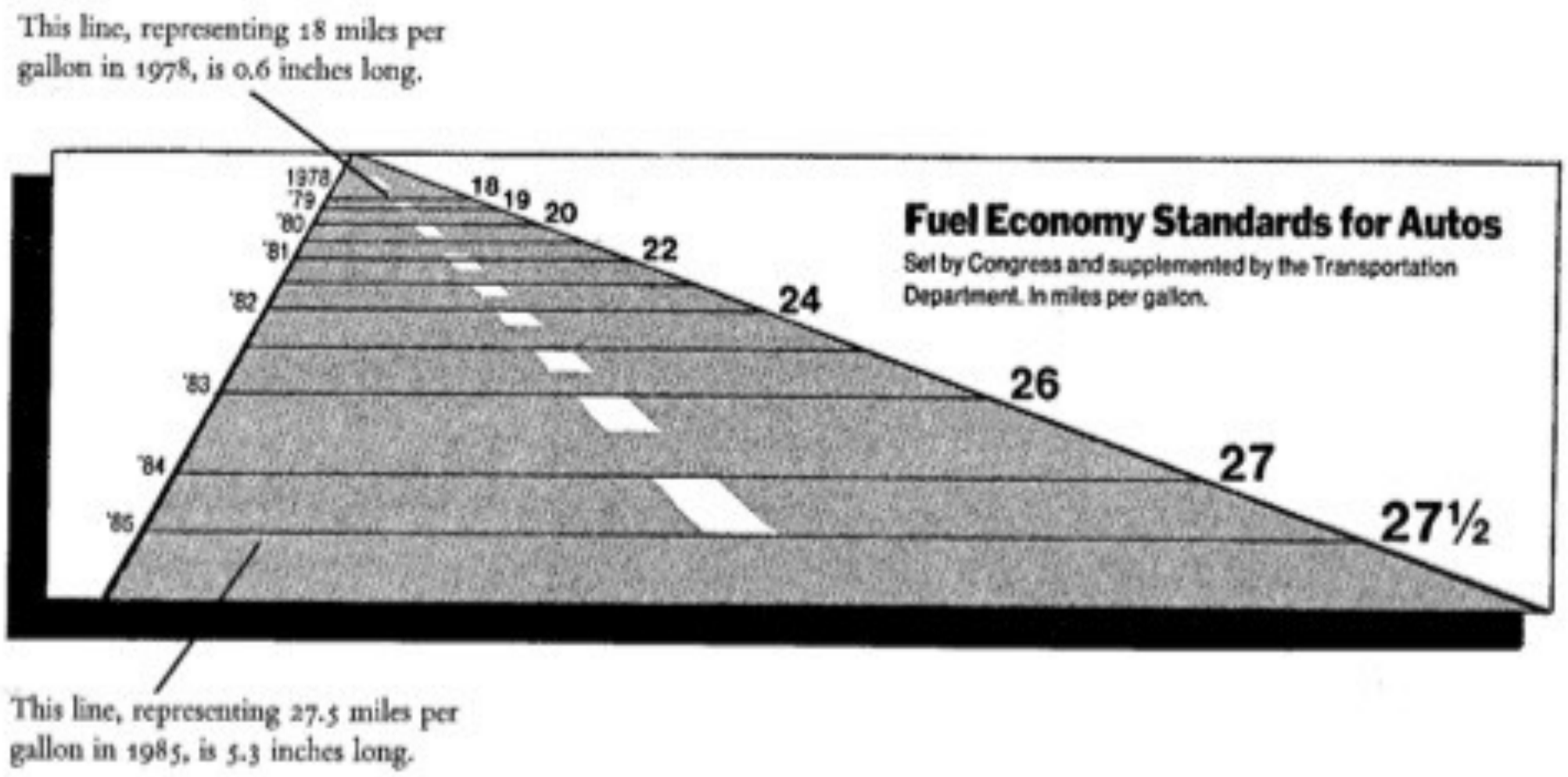
Employment to Population Ratio = Employed / Civilian Population = 148,331 / 250,080 = 59.3%

Source Data: FRED Database

https://www.interaction-design.org/literature/article/guidelines-for-good-visual-information-representations

Tufte's Criteria for Good Visual Information Representation

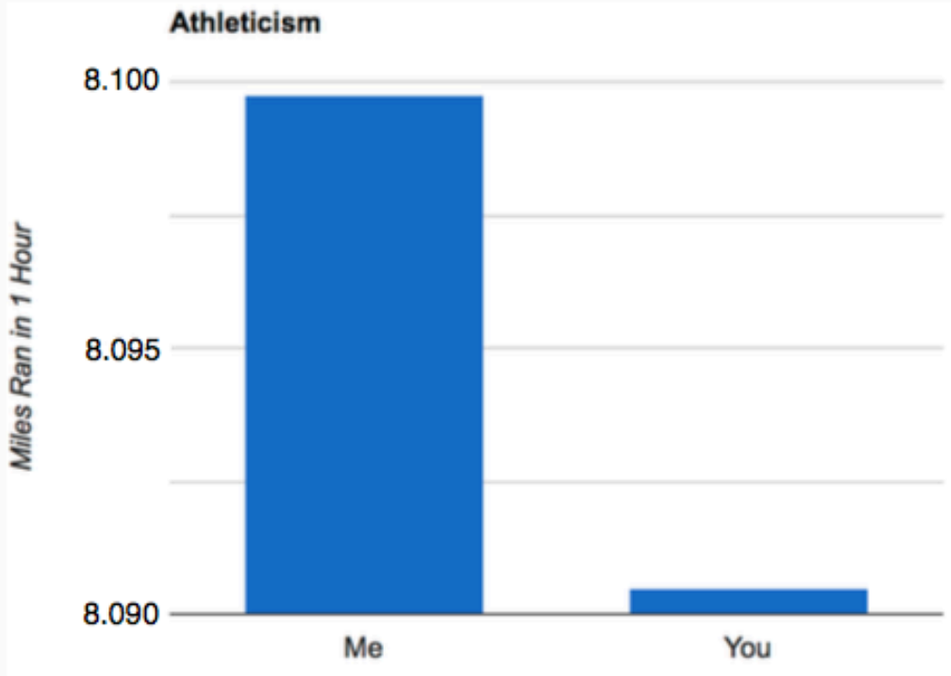
- **Visual Integrity:** Numerical scales should be properly proportionate (and not fudged to exaggerate the fall or rise of a curve at a particular point, for example).



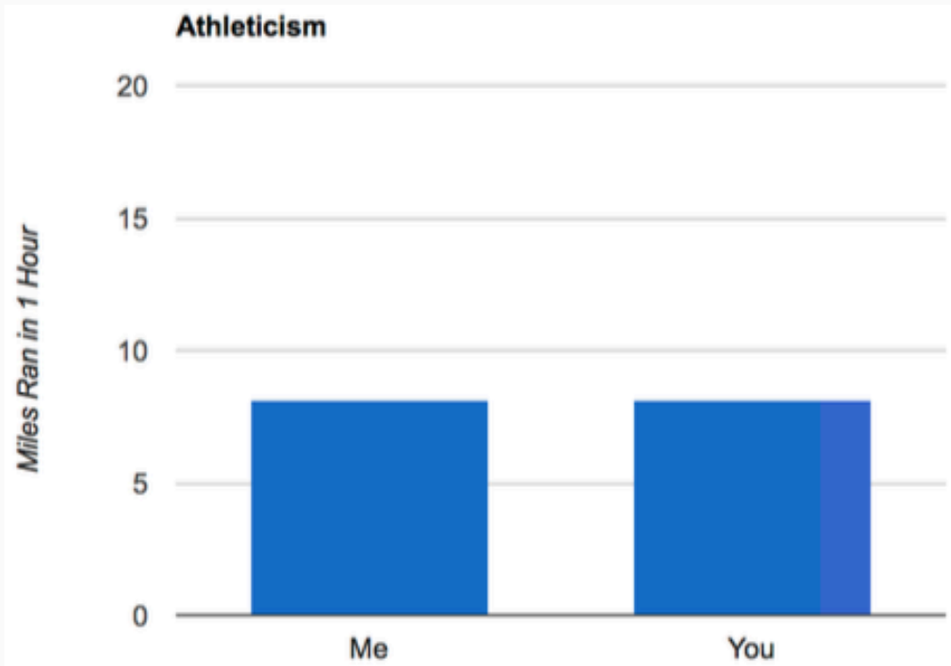
Tufte's Criteria for Good Visual Information Representation

- **Visual Integrity:** Numerical scales should be properly proportionate (and not fudged to exaggerate the fall or rise of a curve at a particular point, for example).

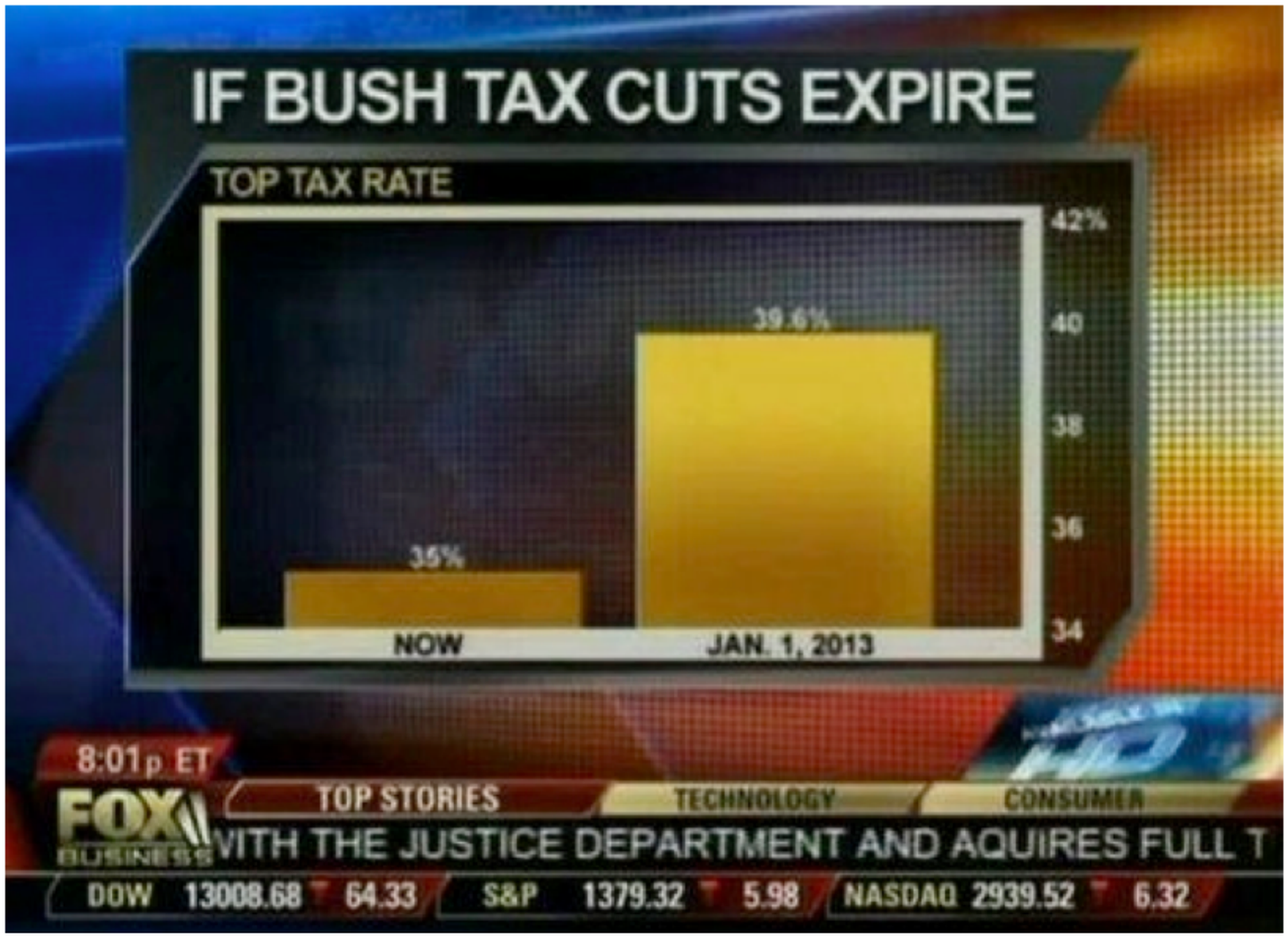
Bad

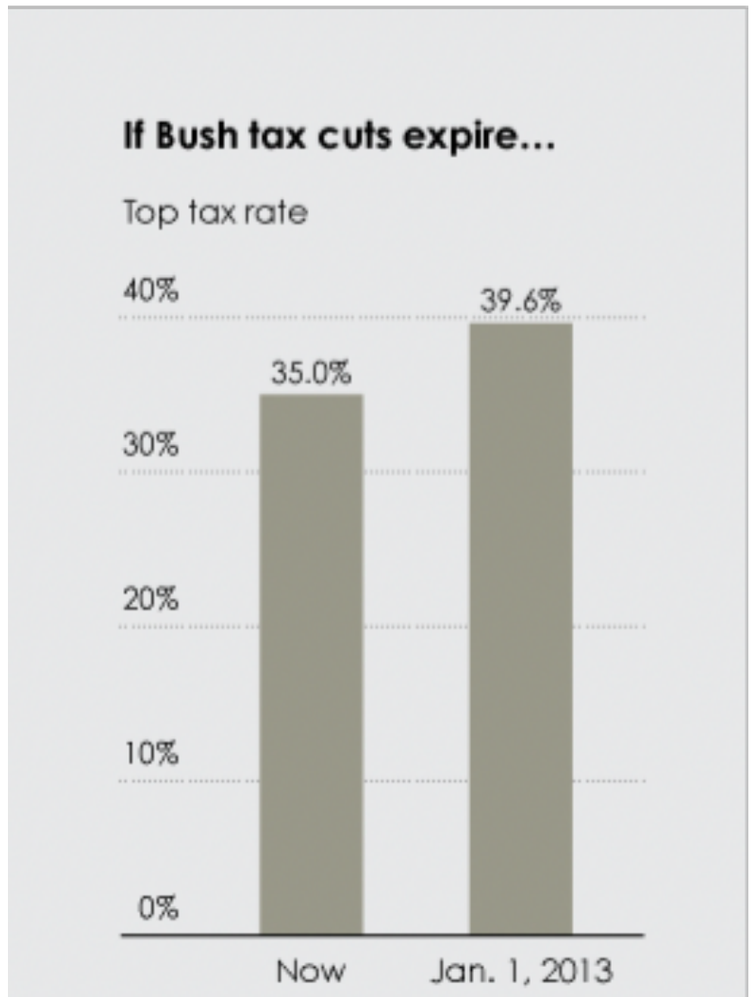


Better



Tufte's Criteria for Good Visual Information Representation



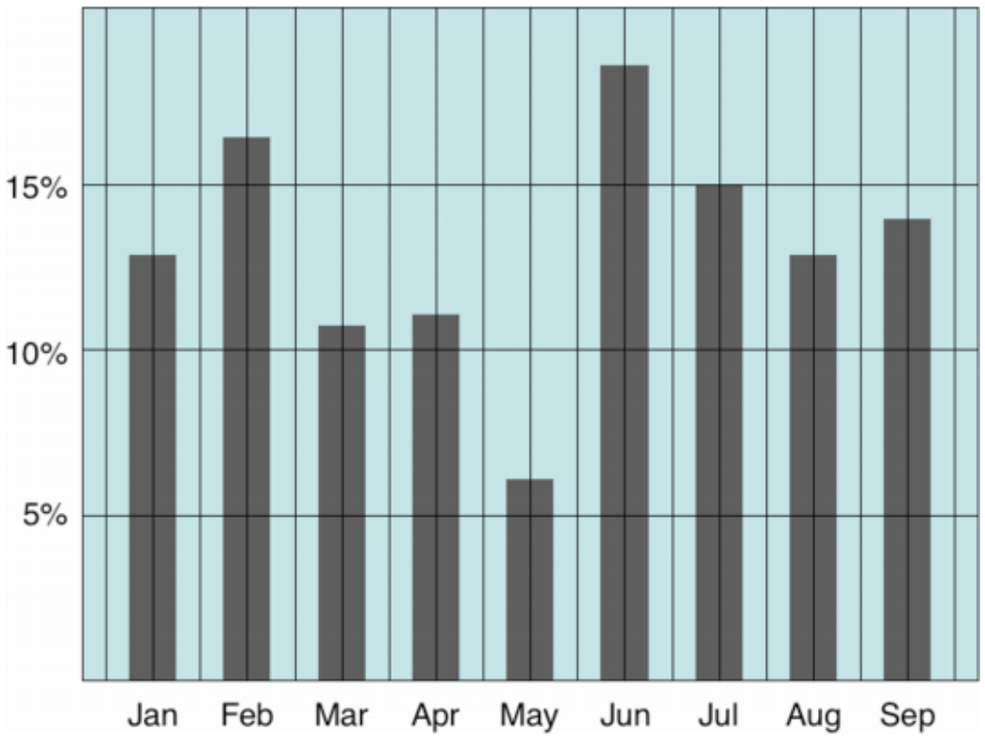


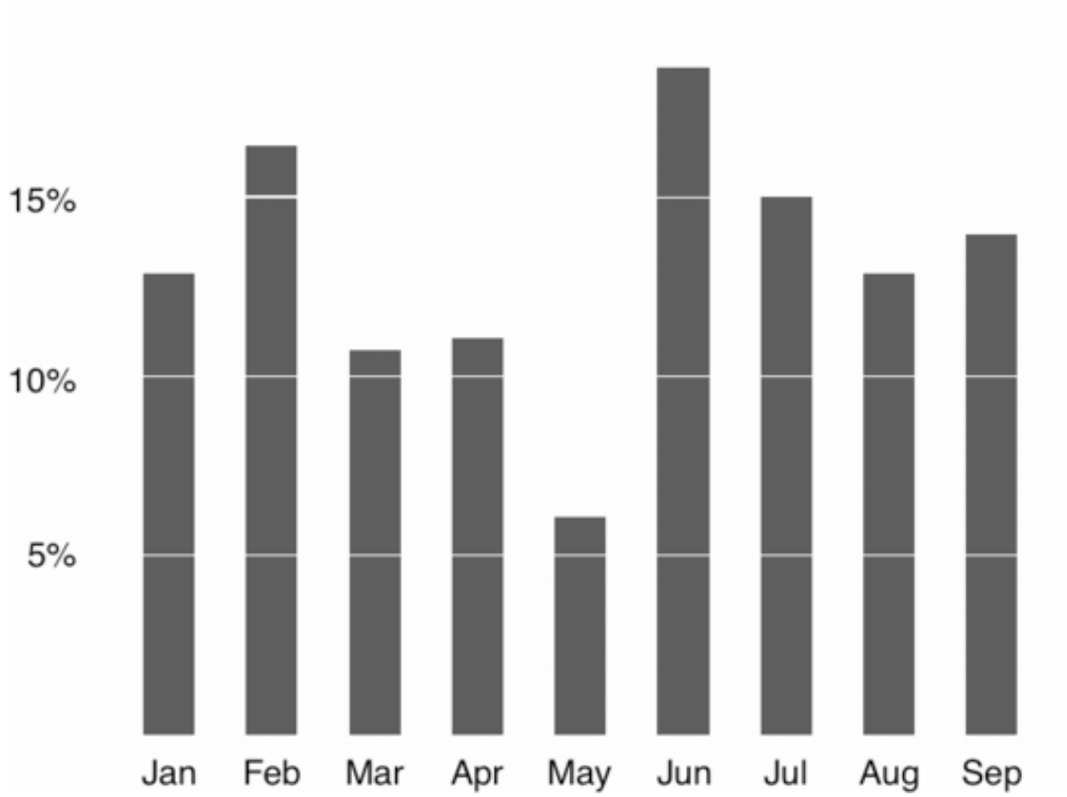
Scale Distortions

Tufte's Criteria for Good Visual Information Representation

- Maximizing the Data-Ink Ratio

$$\text{Data-Ink Ratio} = \frac{\text{Data ink}}{\text{Total ink used in graphic}}$$

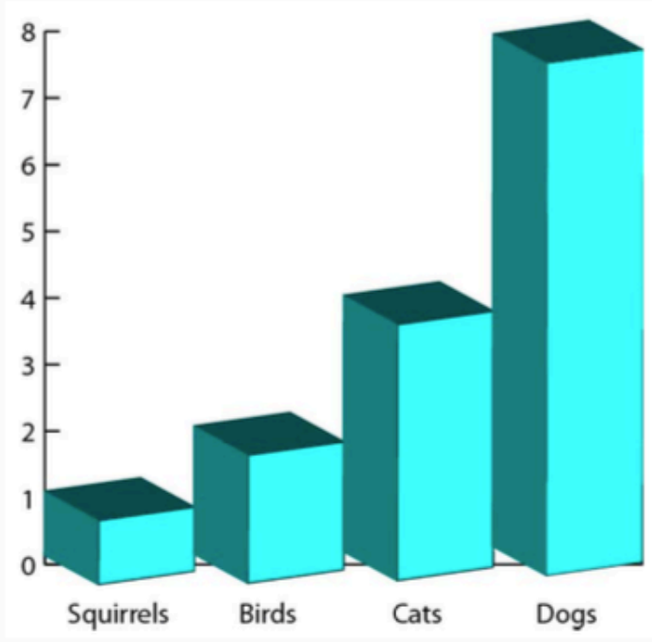




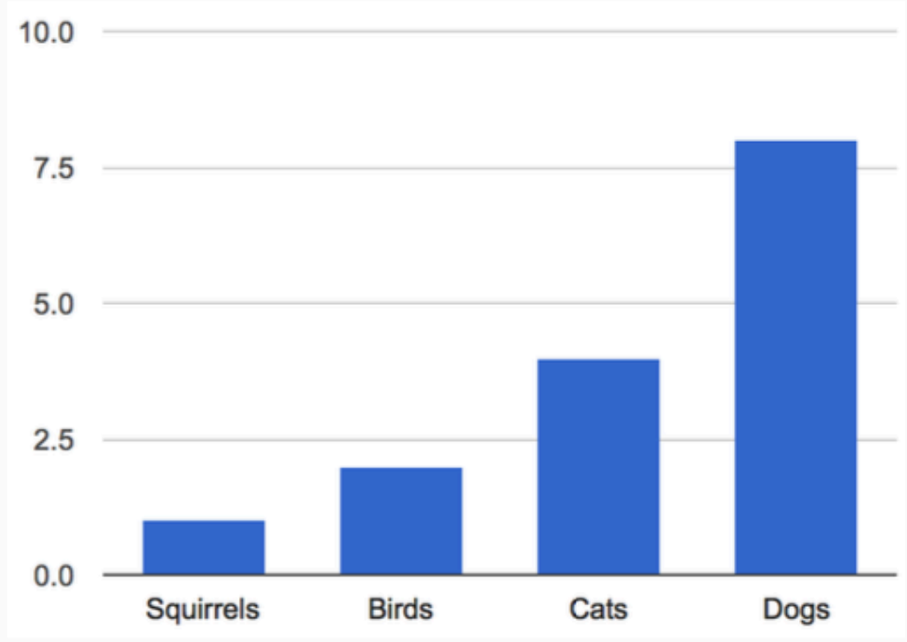
Tufte's Criteria for Good Visual Information Representation

- Minimize chart-junk: show data variation, not design variation.
- The number of visual parameters should not exceed the dimension of the data!

Bad



Better



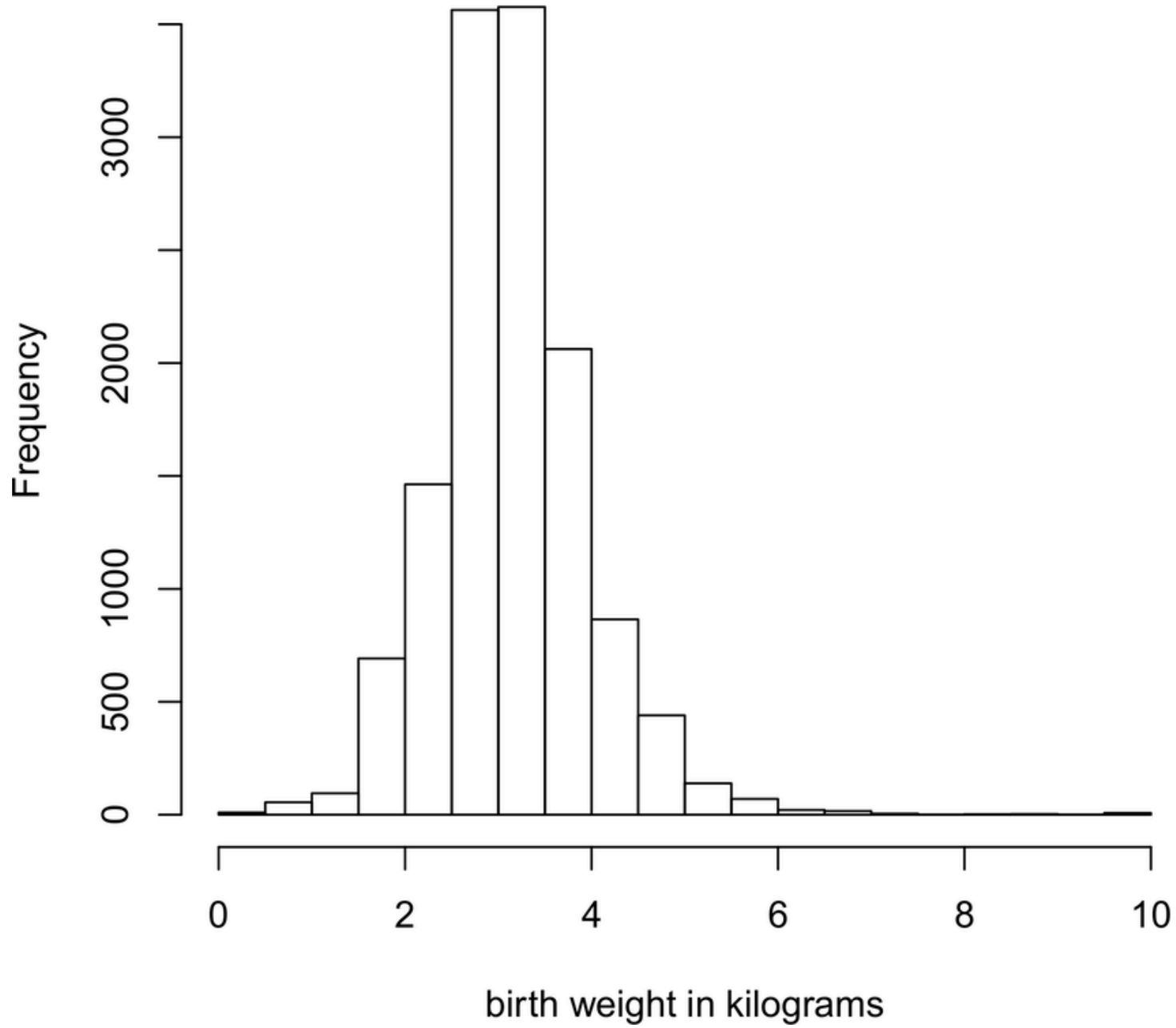
Tufte's Criteria for Good Visual Information Representation

- Basic data visualization guidelines from Edward Tufte:
  - Maximize data to ink ratio: show the data
  - Don't lie with scale: minimize size of effect in graph (Lie Factor) size of effect in data
  - Minimize chart-junk: show data variation, not design variation
  - Clear, detailed and thorough labeling (including important events)

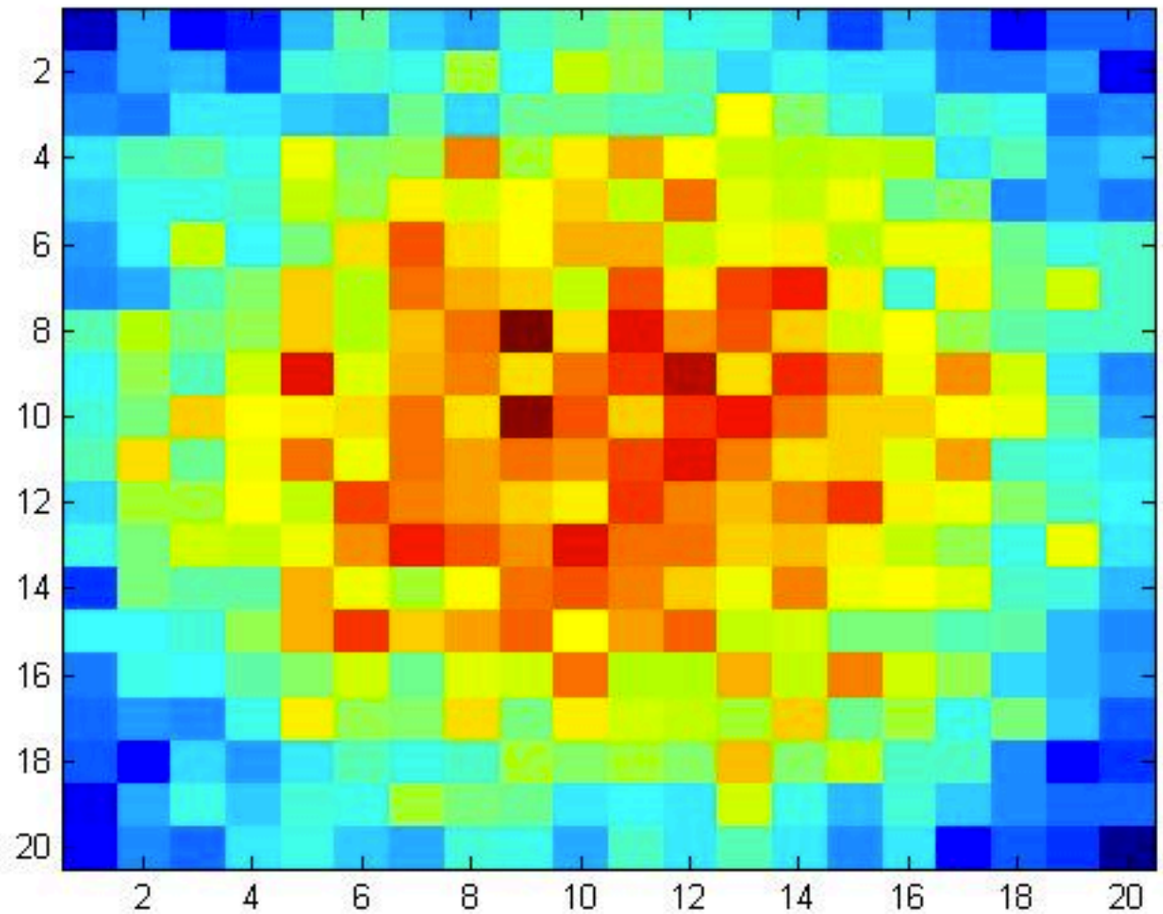
Types of visualization

- **Distributions**
- **Relationship**
- **Composition**
- **Comparison**

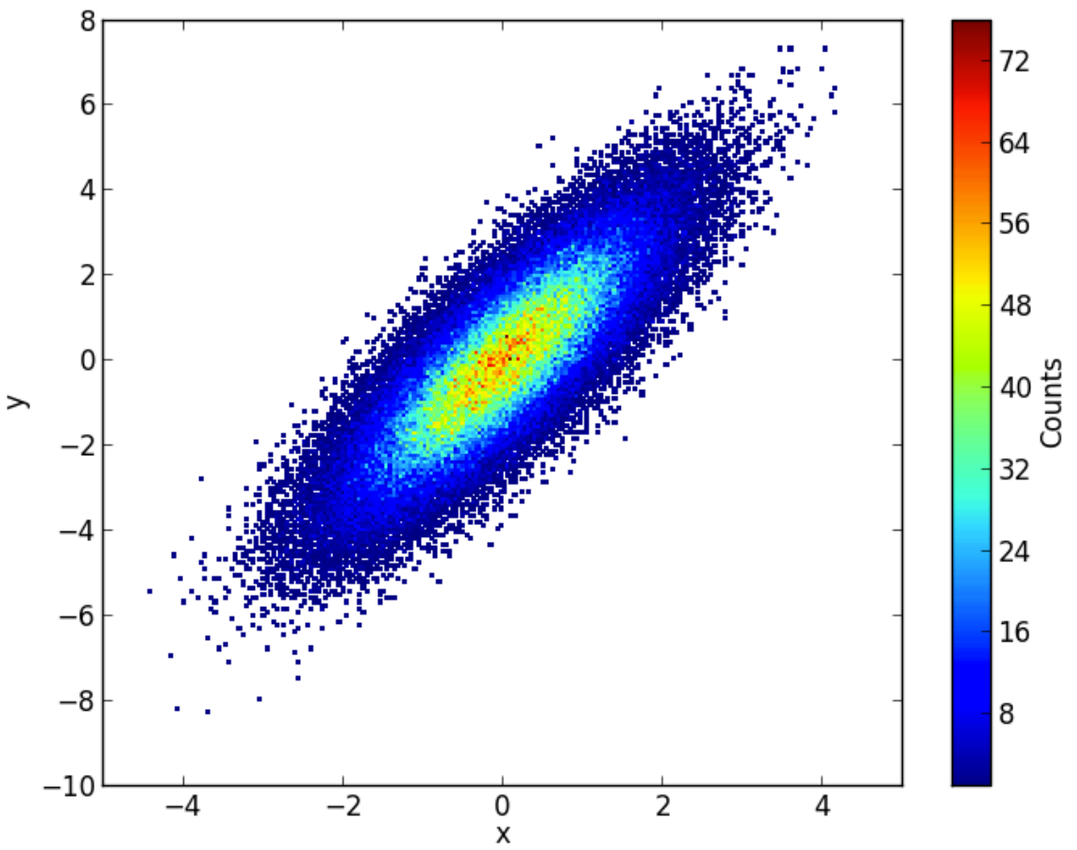
Distributions: histogram



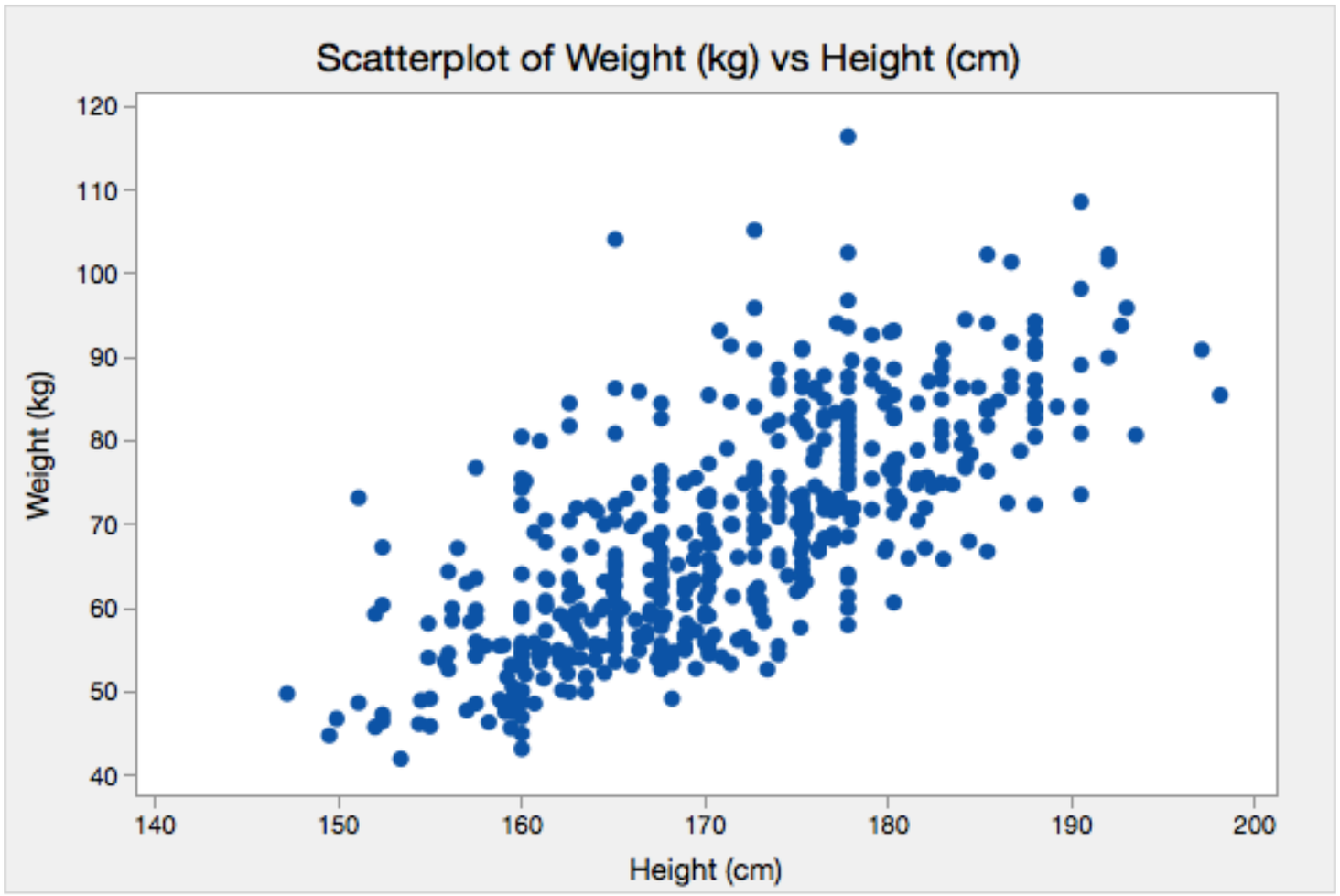
Distributions: 2D histograms



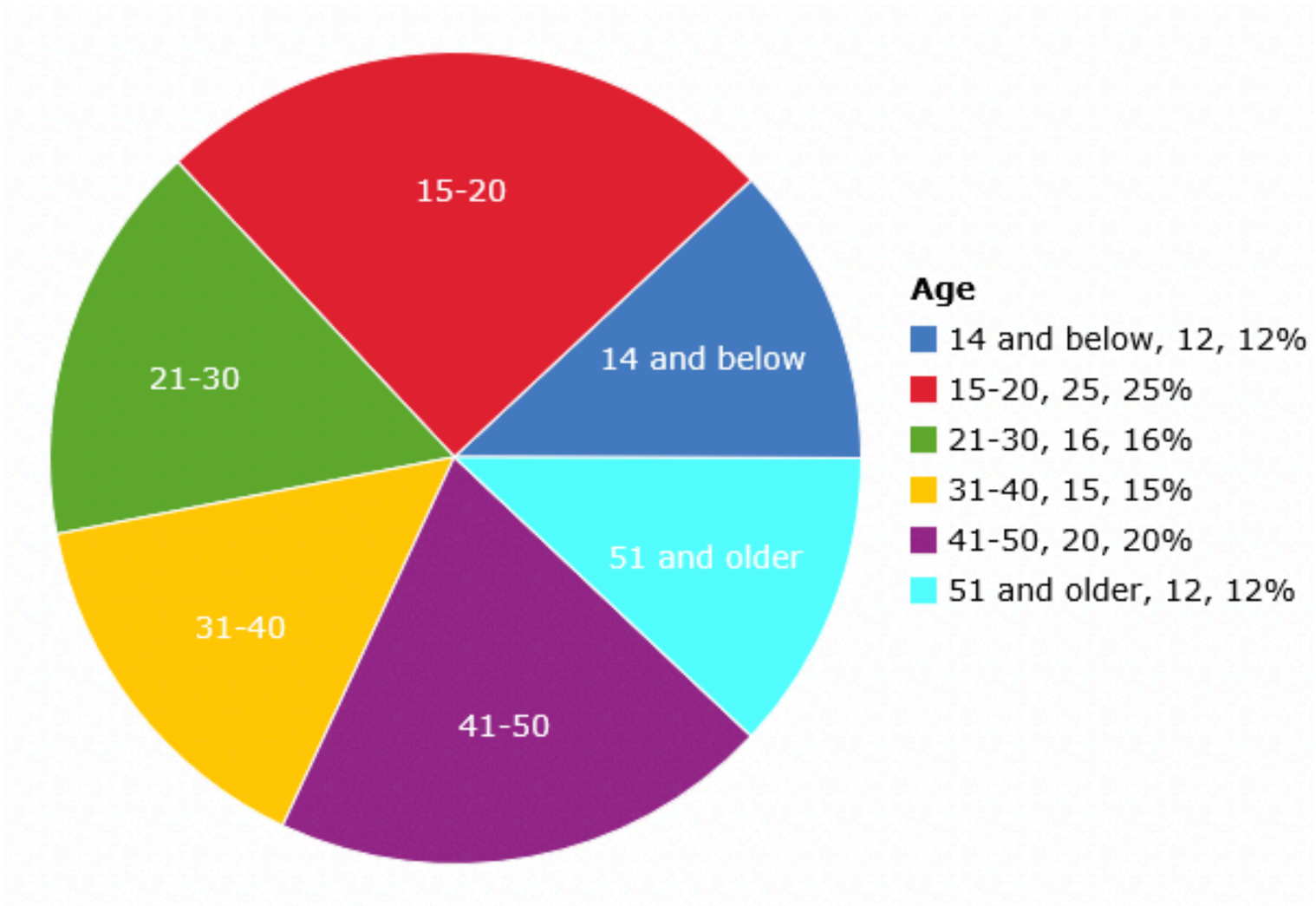
Distribution: Scatter plot



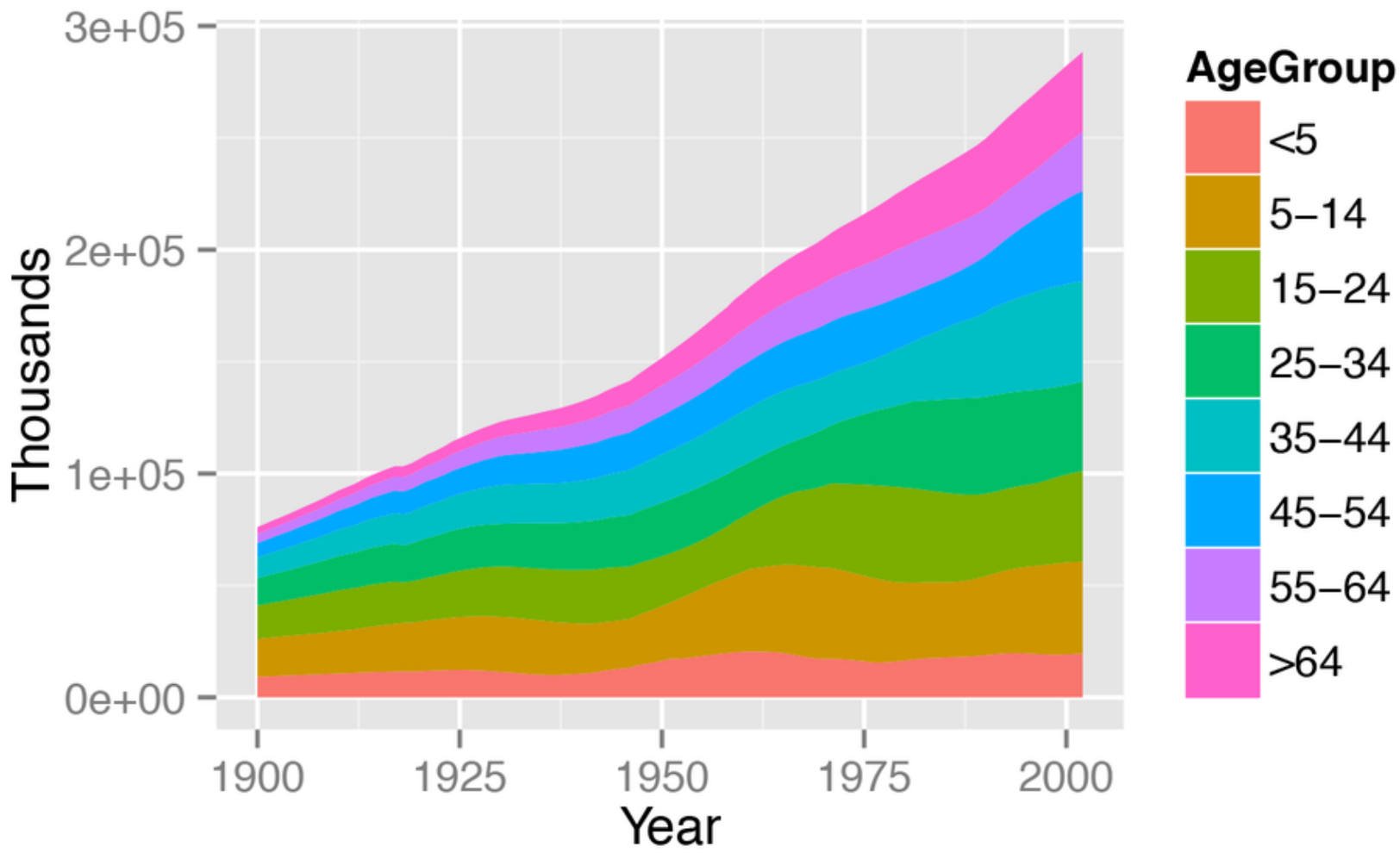
Relationship: scatter plot



Composition: pie chart

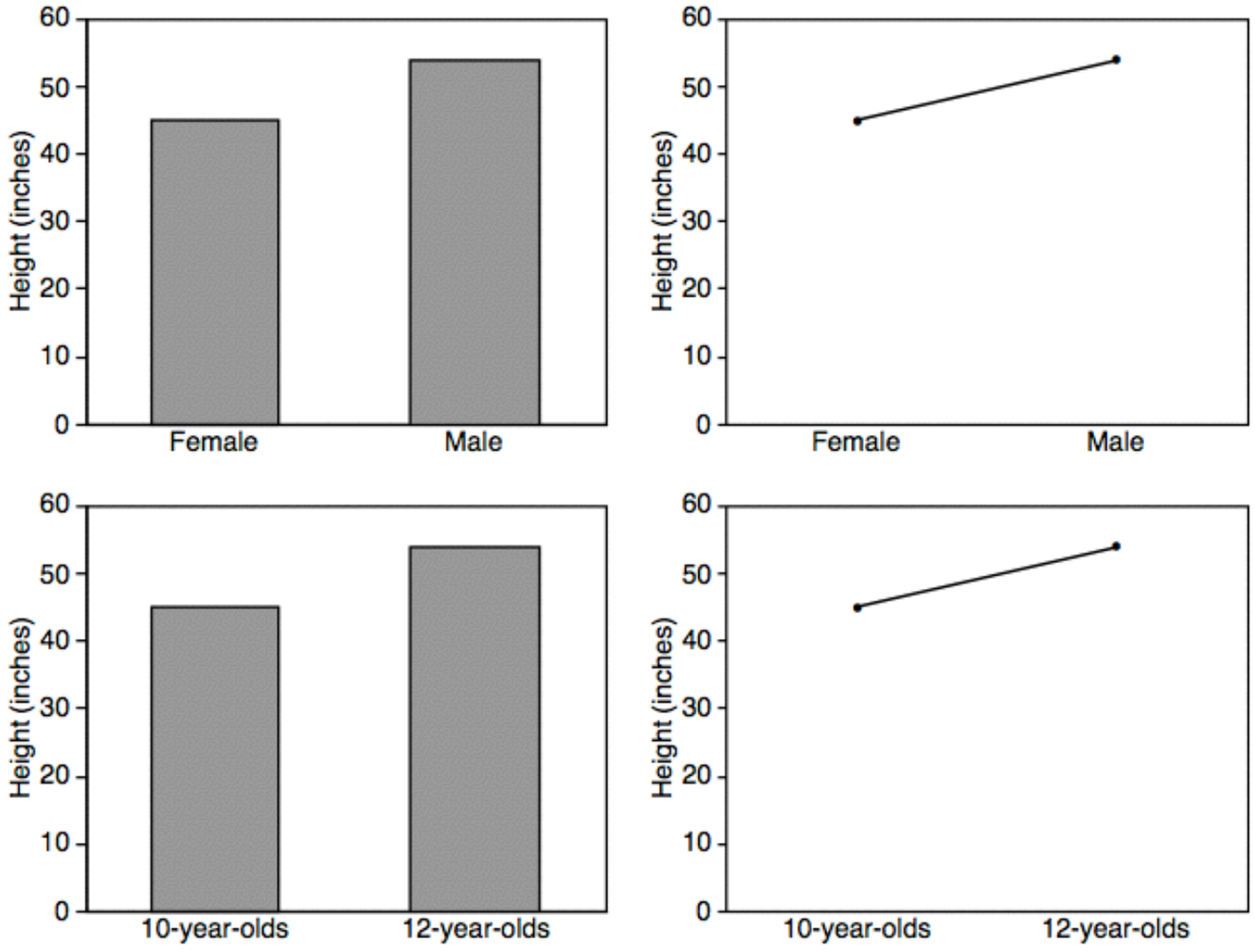


Composition: stacked area graph

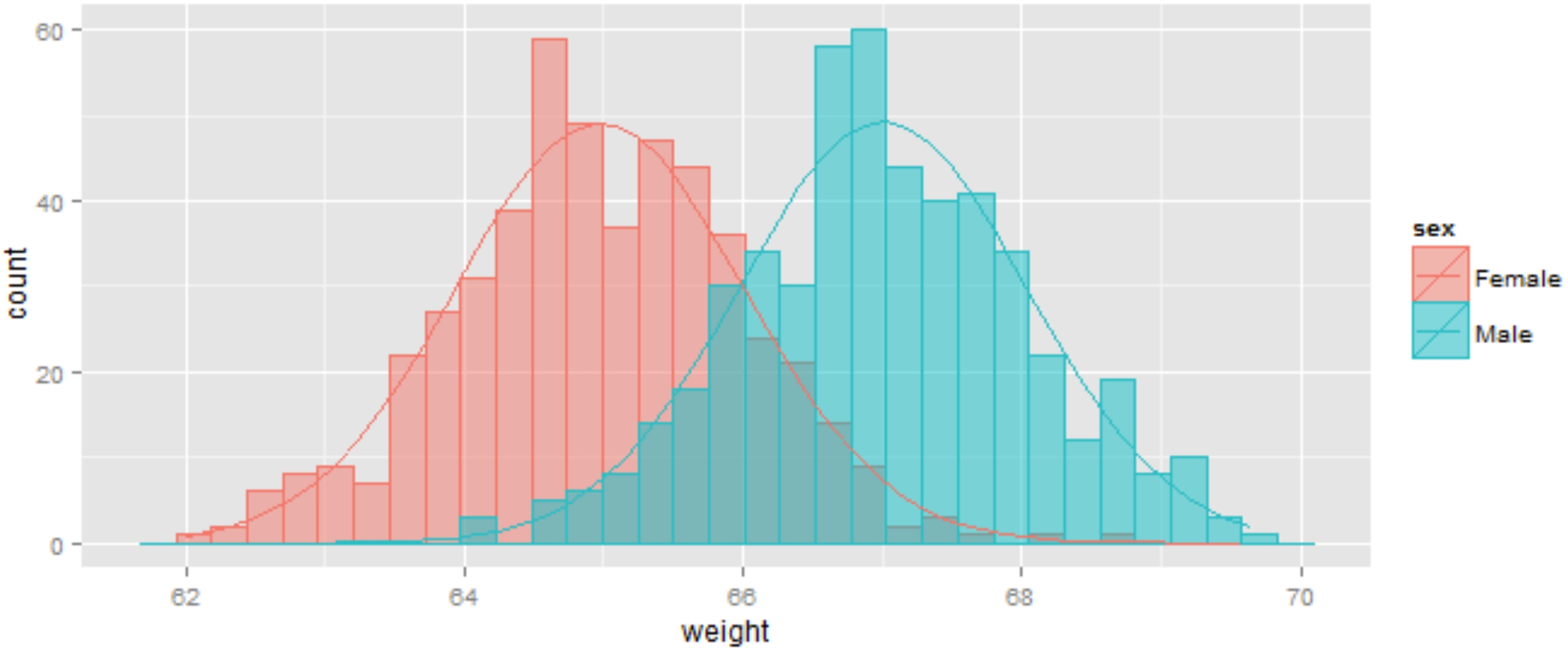


Comparison: bar chart

Comparison: bar chart vs lines



Comparisons: multiple distributions



Trends


Apple Inc. (AAPL) - NasdaqGS
Add to Portfolio
Like 6k
601.10 15.53(2.65%) 4:00PM EDT | After Hours: 604.60 3.50 (0.58%) 7:15PM EDT - Nasdaq Real Time Price
Enter name(s) or symbol(s)
GET CHART
COMPARE
EVENTS
TECHNICAL INDICATORS
CHART SETTINGS
RESET
Feb 10, 2012 : AAPL 493.42
600
550
500
450
400
350
© 2012 Yahoo! Inc.
2011
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
2012
Feb
Mar
Volume 22,523,900
40M
20M
1D
5D
1M
YTD
3M
6M
1Y
2Y
5Y
Max
FROM: Mar 18 2011
TO: Mar 16 2012
1984
1989
1994
1999
2004
2009
Basic Chart | Full Screen | Print | Share | Send Feedback

Next week

- For Tuesday:: 10 Minutes to Pandas
  - http://pandas.pydata.org/pandas-docs/stable/10min.html
- For Thursday:
  - Bishop:
    - 1.4 The curse of Dimensionality
    - 12.1 Principal Component Analysis
  - Non-negative matrix factorization

Departamento de Ingeniería Informática

Thank you

Guillermo Cabrera-Vives
guillecabrera@udec.cl