DA503 Applied Statistics

Lecture 01

Introduction

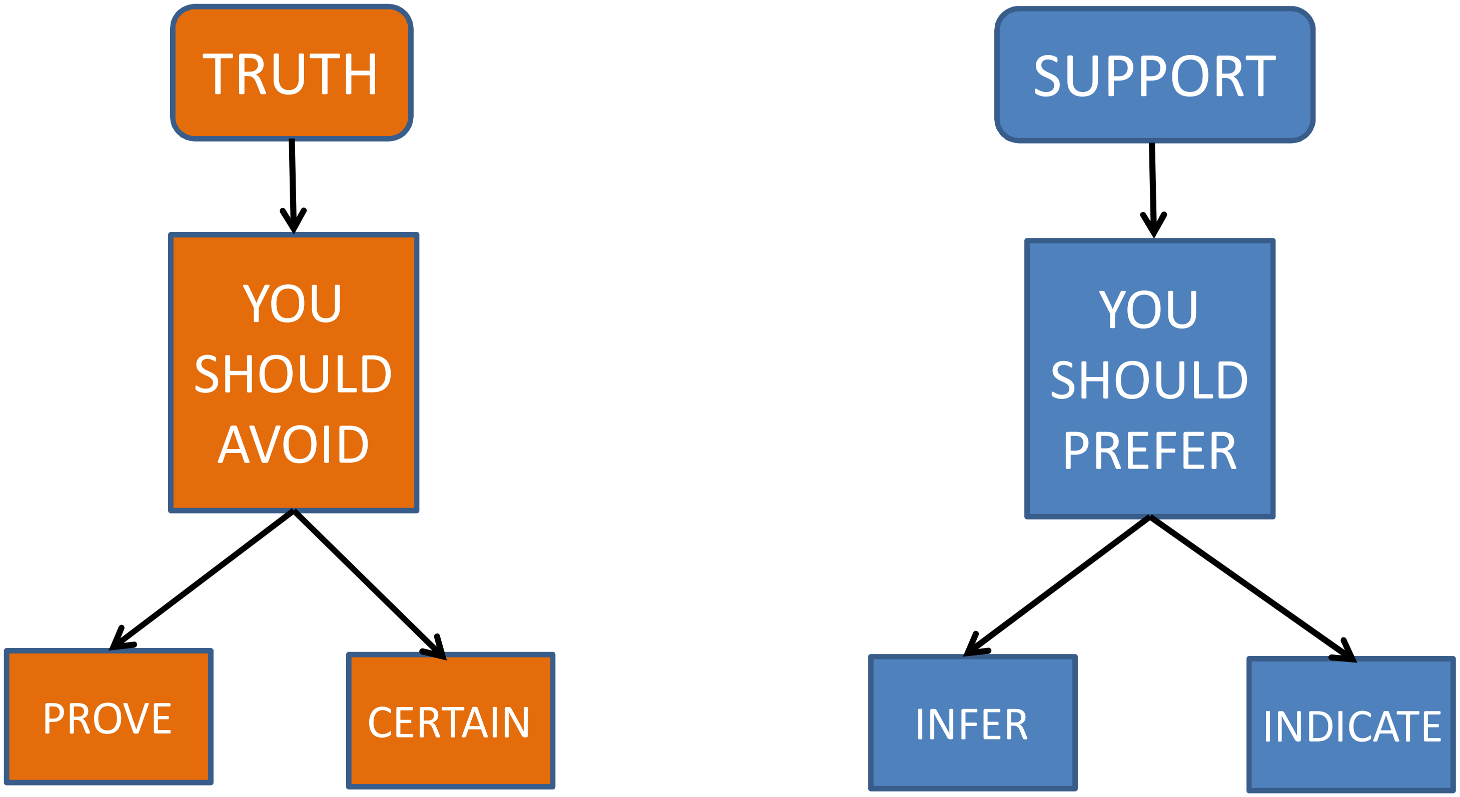
Fall 2021 – H . Sait Ölmez

Course Contents

- **Introduction**
  - General concepts in Statistics
  - Design, experimental setup and data collection
  - Preliminary data analysis
- **Descriptive Statistics**
  - Frequency distributions and histograms
  - Location and central tendency
- **A Primer on Probability**
  - Basic rules of probability
  - Conditional probability and independence
  - Probability distributions
- **Inferential Statistics**
  - Point estimation
  - Interval estimation
  - Hypothesis testing
  - Computational approaches in Inferential Statistics
  - ANOVA
  - Simple/Multiple Linear Regression

Wording matters

- Statistics is never 100% certain; but it states its limitations explicitly



TRUTH

YOU SHOULD AVOID

PROVE

CERTAIN

SUPPORT

YOU SHOULD PREFER

INFER

INDICATE

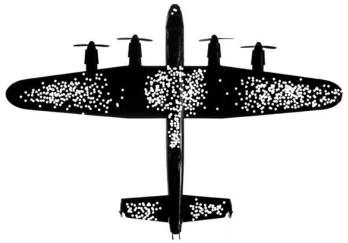
What is Statistics?

- **Statistics** (/stəˈtɪstɪks/): The discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- We use Statistics to
  - separate signal from noise
  - summarize and understand data
  - infer from a sample to a population
  - make a decision in the face of uncertainty
- When do you not need statistics?
  - When you have the data for the whole population
  - When there is no variability

Understanding data

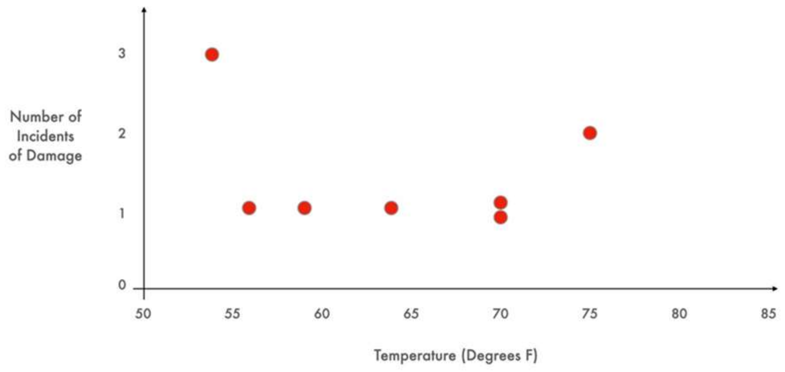
- Aaron Levenstein: *"… What it reveals is suggestive but what it conceals is vital"*
- Survivor bias by Abraham Wald

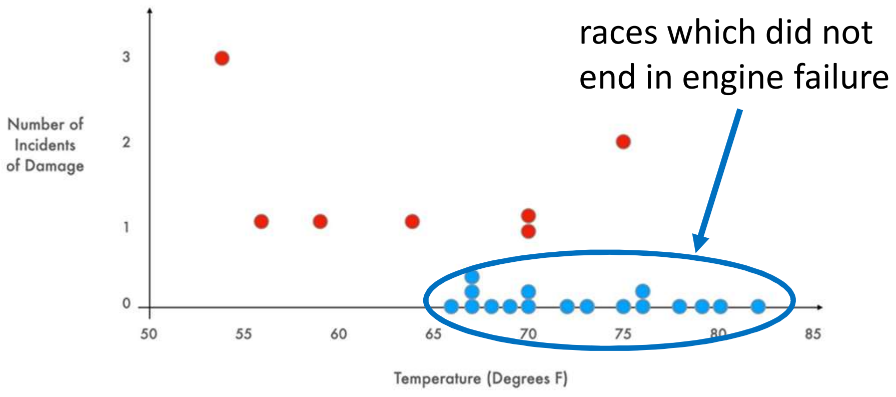
Planes that were able to come back from an airstrike



How could we make these planes stronger?

- Here is the number of engine failures in race cars as a function of ambient temperature. If the temperature for the next day is forecasted to be 45 ^{o}F, would you go for the race?

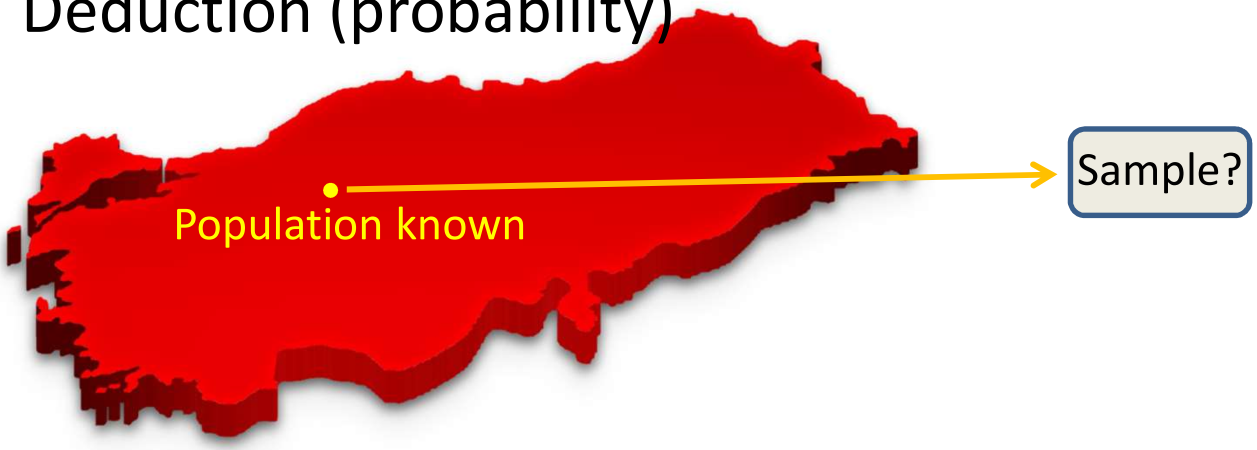




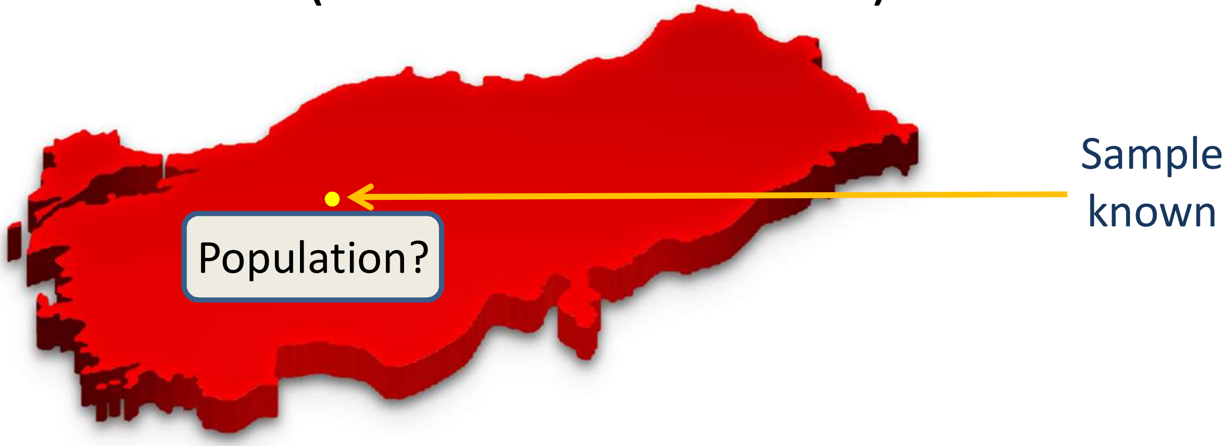
races which did not end in engine failure

Deduction & Induction in Statistics

Deduction (probability)



Induction (statistical inference)



A probabilist asks the probability of drawing a red ball given the proportions in the whole jar (population). A statistician infers the proportion of the red balls by sampling from the jar (population).

Image source: mesmes.deviantart.com

Phases of Statistics

- Three major phases:
  1. **Prelude to Data Analysis**
     - Investigation, design, data collection & exploratory analysis
     - “To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.” R. A. Fisher
  2. **Descriptive Statistics**
     - Understand data (numerical/graphical)
  3. **Inferential Statistics**
     - Make inferences about the population using samples randomly selected from the population.

Prelude to Data Analysis

- **Experiment design and setup**
  - What data do we need and how do we collect it?
- **Data integration and cleansing**
  - Consolidate and clean data, and make it ready for analysis
- **Data screening and exploration**
  - Get a feel for what you have before the analysis

Prelude to Data Analysis

- **Experiment design and setup**
  1. Ask the right questions and create a use-case
  2. Carefully design your questionnaire (for the right & relevant data)
  3. Create your sample (watch out for hidden bias)
  4. Work with the right sample size
     - Is sample size large enough to observe an effect of desired magnitude?
       - Variance of the parameter under investigation?
       - Magnitude of the expected effect in comparison to the standard deviation of the parameter?
  5. Collect data (interview, online surveys, etc.)

Experimental setup

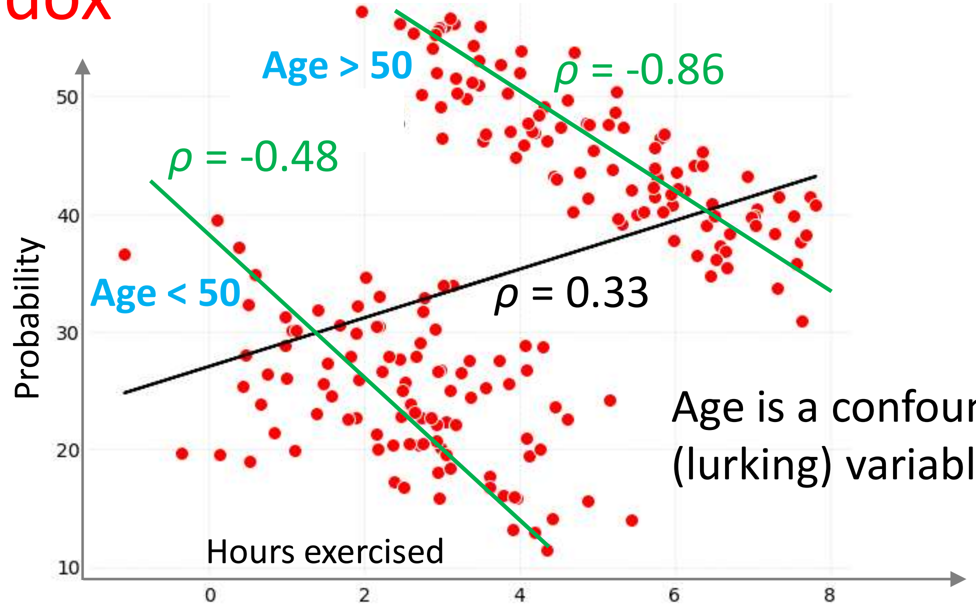
- **Experiment design and setup (cont'd)**
  1. Know your operating conditions
  2. Optimize for the right thing!
     - A/B test gone wrong for the New Coke

Experimental setup (cont’d)

- **Experiment design and setup** (cont'd)
  - Do you know what data do you need?

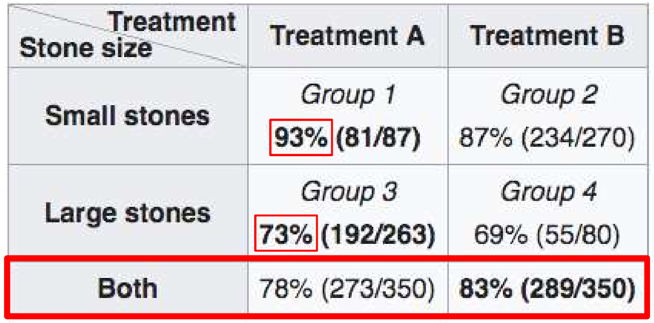
Simpson's Paradox

Hours of exercise per week versus the probability of risk for developing a disease for 2 sets of patients:



Age is a confounding (lurking) variable here

Effectiveness of 2 kidney stone treatments:



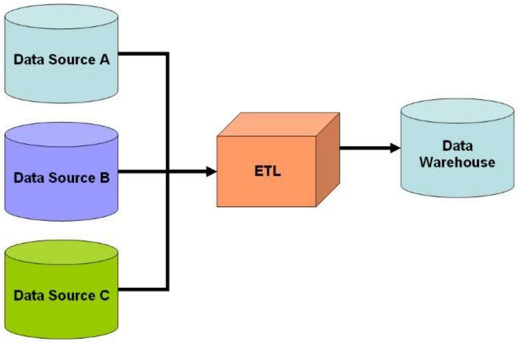
Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* 93% (81/87)	*Group 2* 87% (234/270)
Large stones	*Group 3* 73% (192/263)	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

Stone size is a confounding variable here

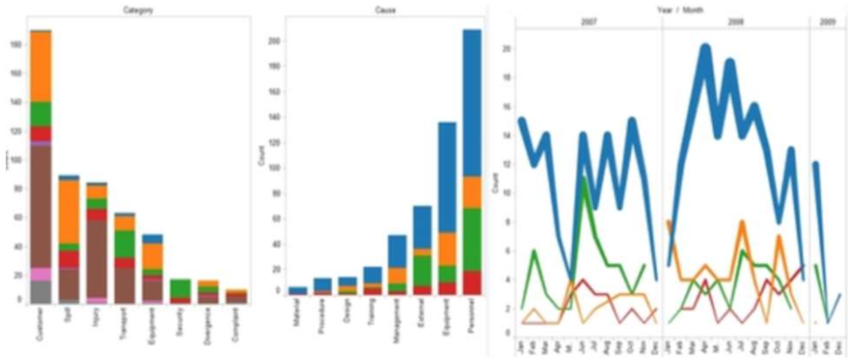
➔ Treatment B is better!

Data Integration and Cleansing

- **Integrating data**
  – Combining data from multiple sources into a coherent store

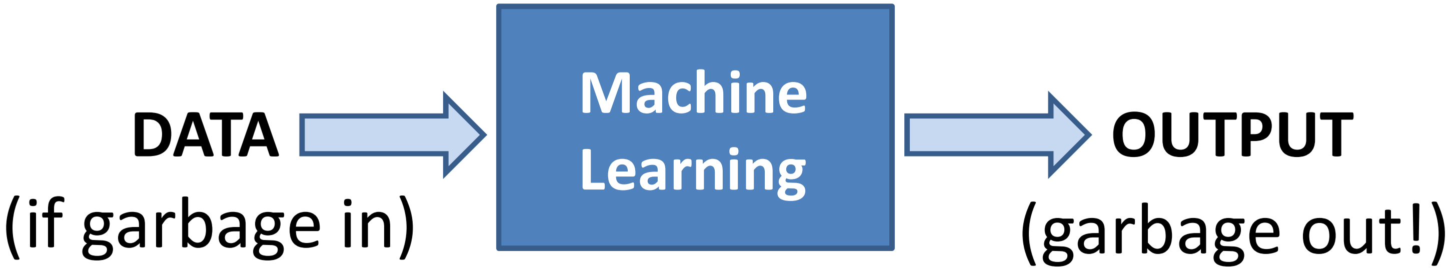


- **Cleaning and exploring the data**
  – Data cleaning (Missing values, outliers, noisy data)
  – Data Preparation (Variable transformation, dimension reduction, feature engineering, etc.)
  – Data screening (visualization and exploration)



Data Quality

- Data quality: Why is it so important?
  - Incomplete/inconsistent/noisy data
- For the problem we're trying to solve, the data used has to be **accurate, consistent** and **relevant** to the problem at hand.
- Data quality/integrity is, and will always be a critical part of data management. No matter what technologies are in play, if the data is bad, then the information coming out cannot be trusted.



DATA

(if garbage in)

Machine Learning

OUTPUT

(garbage out!)

Data cleansing

- A major part of any data analytics project
  - More than 70-80% of a data analytics project is spent on getting the data ready for analysis
- Data quality issues
  - Missing, incomplete or duplicate values
  - Inconsistency in data type or data format
  - Erroneous data
    - Typographical errors in categorical values
    - Numerical values way out of range
  - Outliers
- Usually more data pre-processing tasks:
  - Data aggregation, data conversion, data normalization, dimension reduction, etc.

Data cleansing (cont'd)

- Before you go any further, check:

Any missing data?

Data format?

Data length?

Data type?

Any duplicates?

Data range?

Confusing column names?

Data Screening

- How to screen (look at) data?
  - Inspect raw data
  - Summary statistics
    - Mean, median, mode, max, min, range, variance (standard deviation) etc.
  - Visualize
    - Visualize what?
      - Examples across all features (rarely)
      - Features across all examples (a lot more common)

Exploratory Data Analysis (EDA)

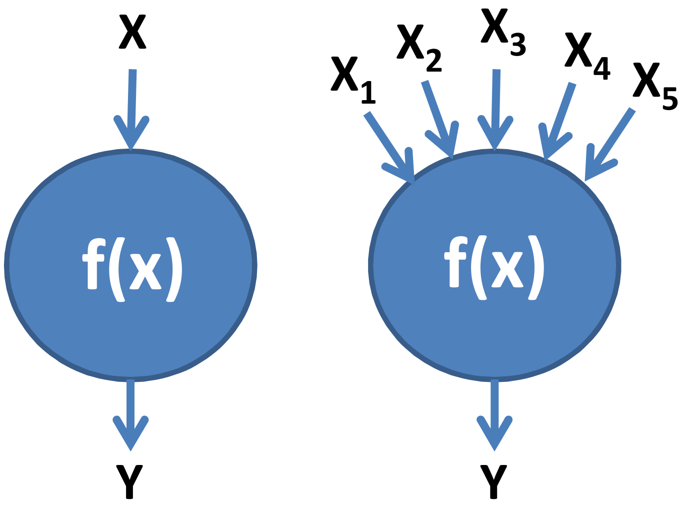
Exploratory Data Analysis (EDA)

- Why EDA?
  - Understand the behavior of your numbers
  - Detect errors early in the analysis
  - Find violations of statistical assumptions and assess assumptions for confirmatory analysis
  - What does the distribution look like? Symmetric, too tall and narrow, too short and wide spread, right- or left-skewed etc? Is the normality assumption violated? These are important as most of our analyses will assume a reference distribution to infer conclusions about the population parameters
  - Anomalous patterns? **Outliers**?

EDA – cont'd

- EDA provides hints on relations among the variables and might reveal patterns in the data, thus helping us generate hypotheses
- EDA serves as a sanity check before we dive into the mechanics of statistical learning

- Statistical learning process:

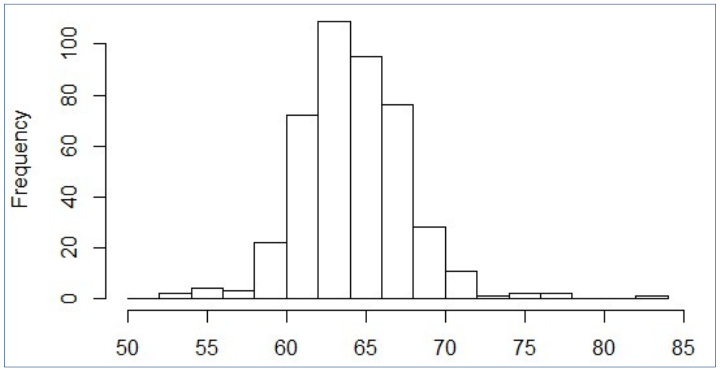


X_i : input, covariate, explanatory variable, independent variable, predictor, feature, attribute

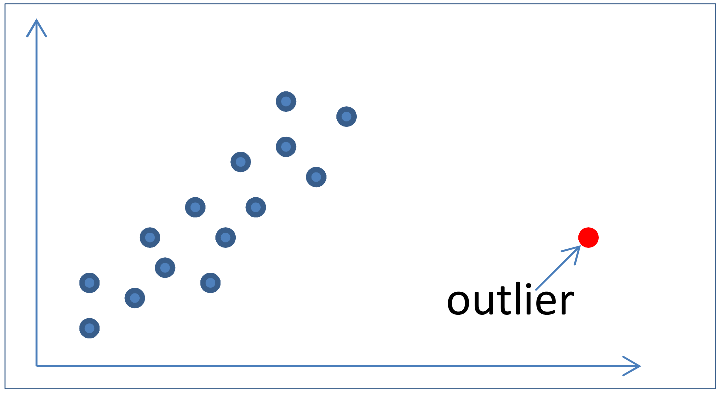
Y : output, dependent variable, target variable, response variable

Importance of visualizing data

- Helps with getting to know your data (eye test)
  - Simple visualization tools (graphs/plots) are very useful
  - Does the data make sense?
  - **Nominal attributes**: Histograms (distribution consistent with experience?)



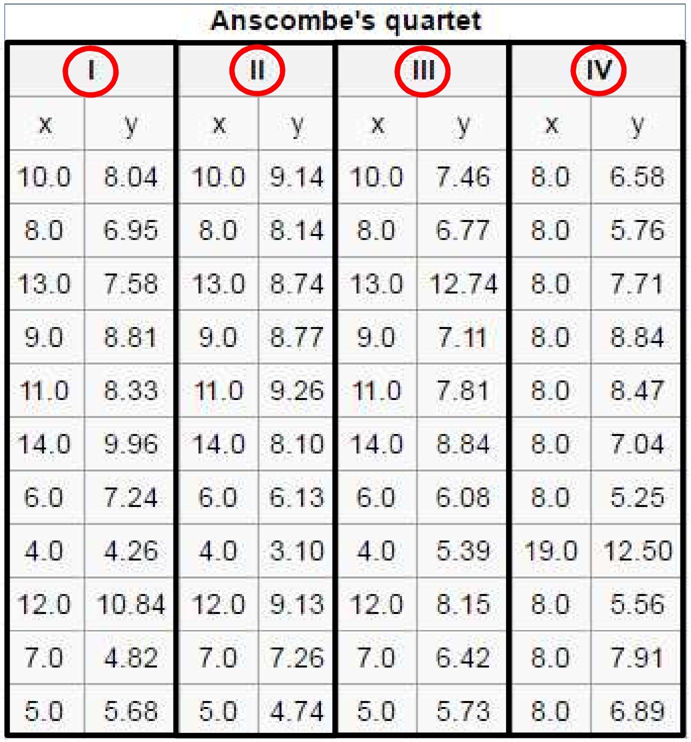
  - **Numeric attributes**: Graphs (any obvious outliers?)



  - 2D and 3D plots may show dependencies
  - Need to consult domain experts

Importance of visualizing data – cont'd

Dangers of summary statistics!



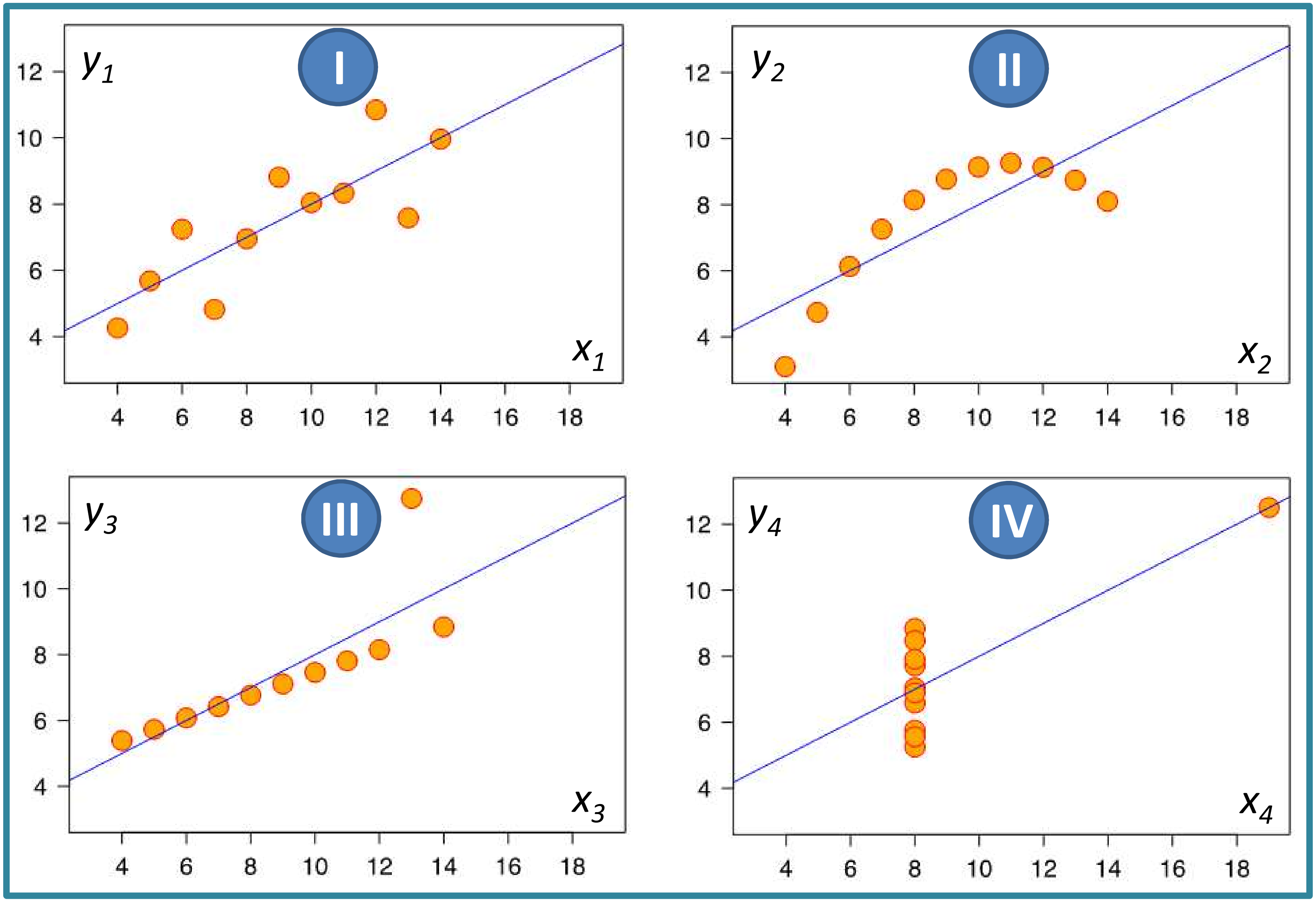
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of *x* in each case	9
Sample variance of *x*	11
Mean of *y*	7.50
Sample variance of *y*	4.122 or 4.127
Correlation between *x* and *y*	0.816
Linear regression line	*y* = 3.00 + 0.500*x*

Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet (Francis Anscombe, British statistician)

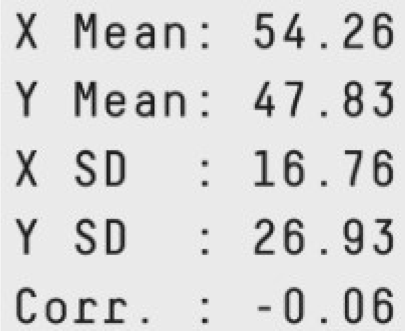
Importance of visualizing data – cont'd



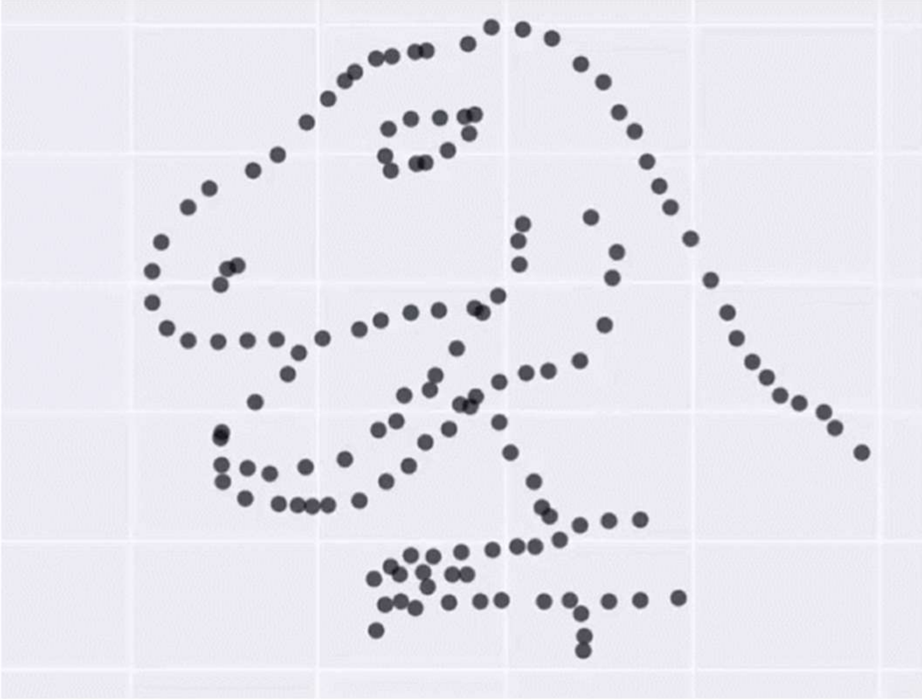
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Importance of visualizing data – cont'd

- **Same Stats, Different Graphs:**

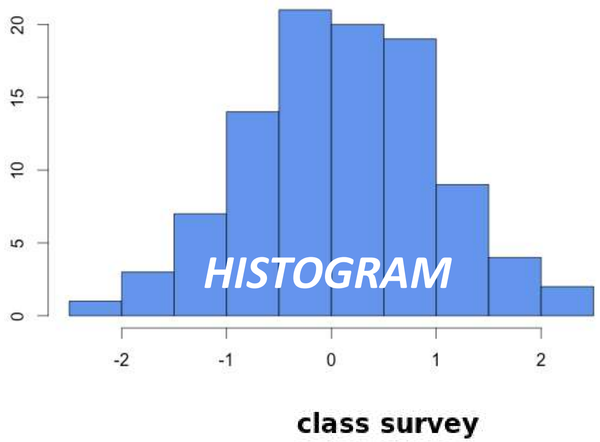


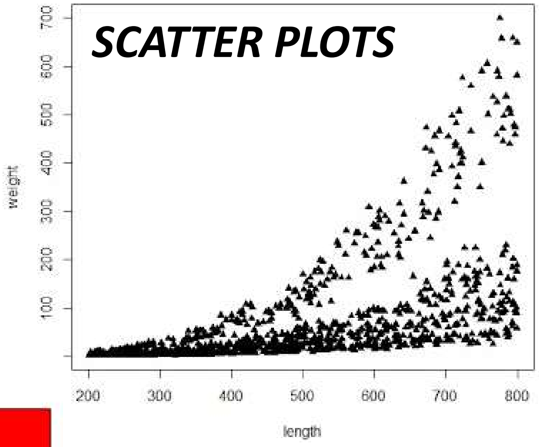
```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

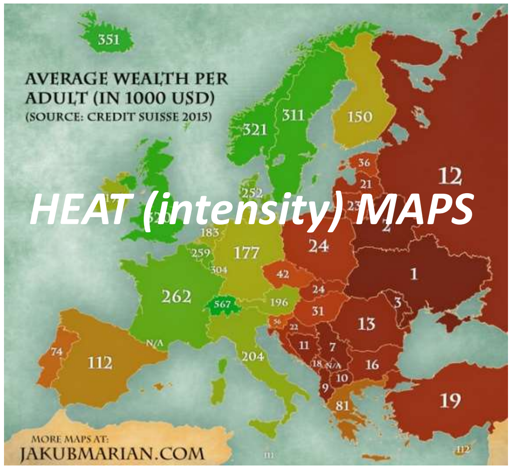


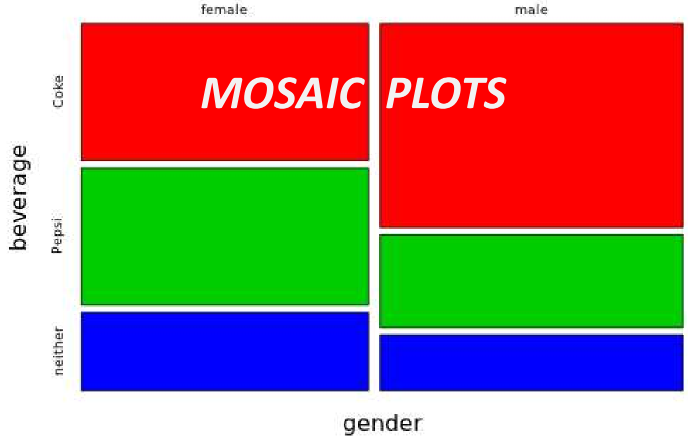
Ref: https://www.autodeskresearch.com/publications/samestats

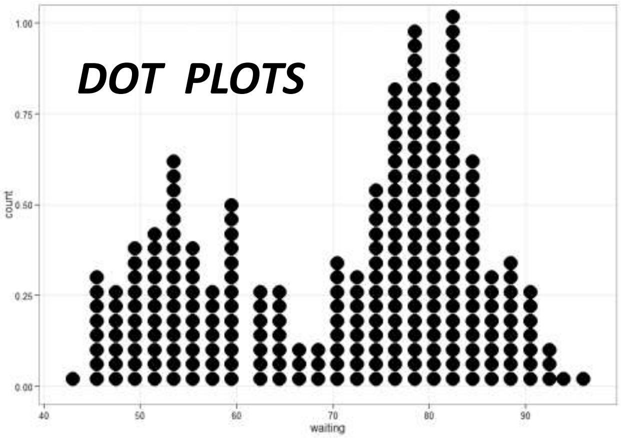
Visualization techniques

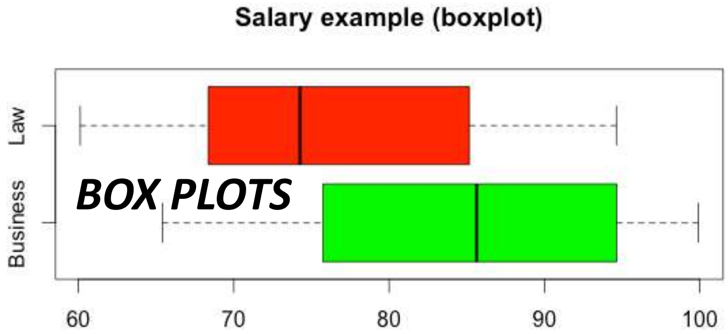


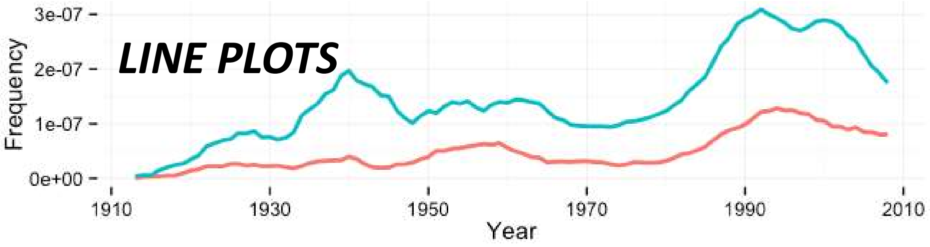






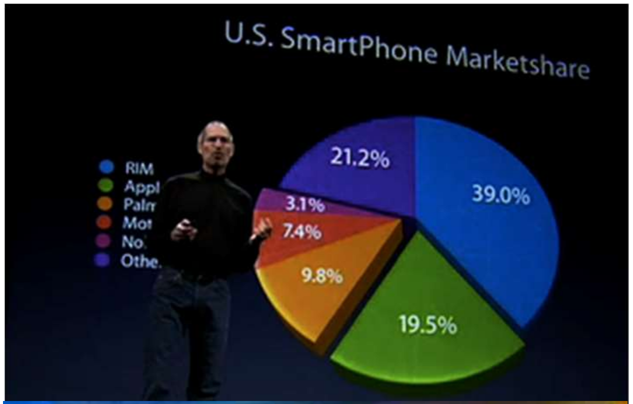


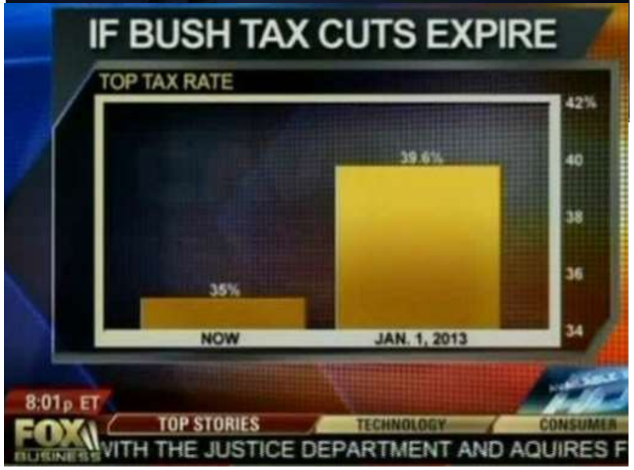




Visualization guidelines

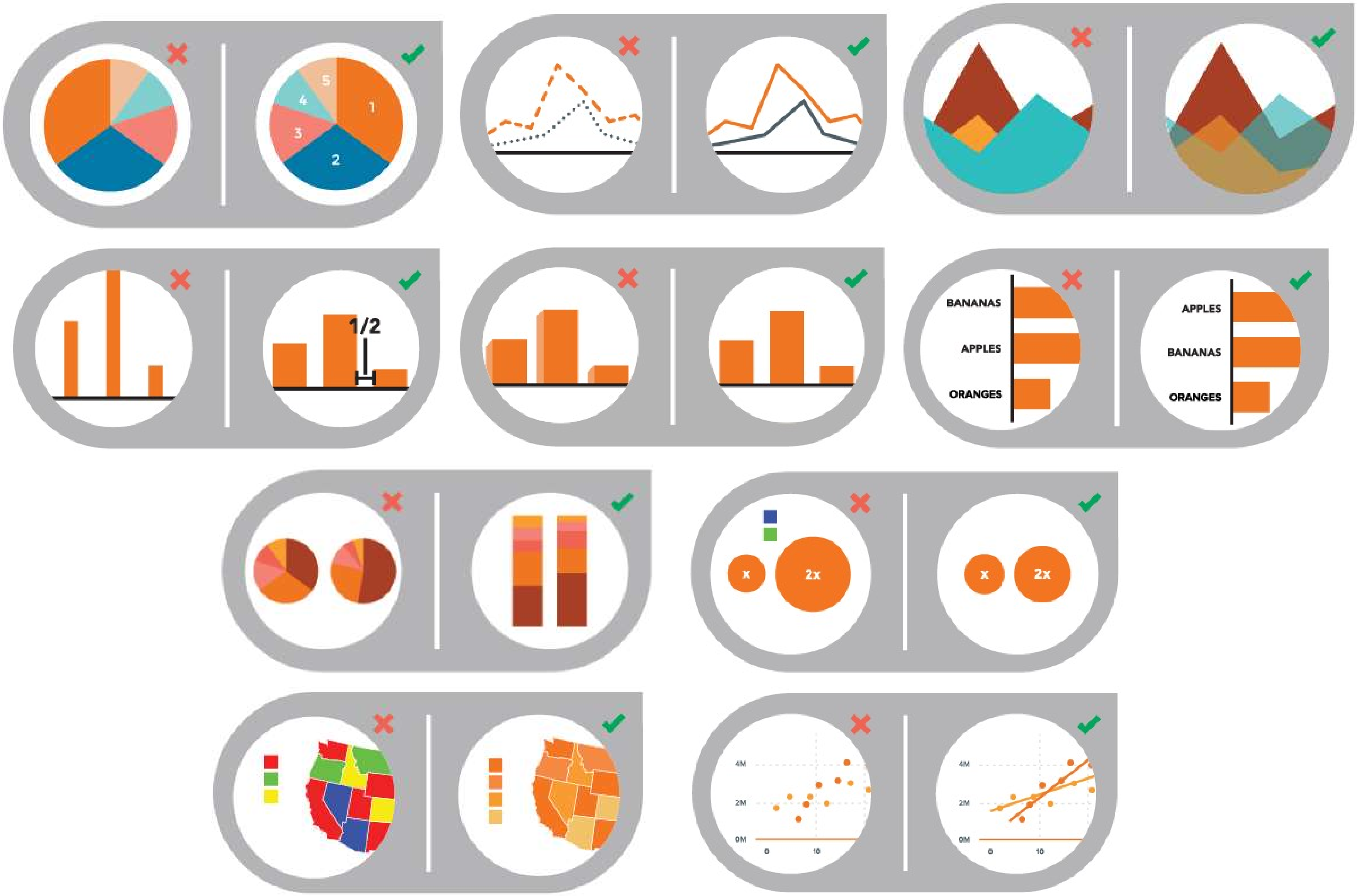
- Stick with “better graphics”:
  - **Know your audience**
  - **Identify your message**
  - Captions are not optional
  - Do not trust the defaults
  - Use color effectively
  - **Don’t mislead the reader:**
  - Avoid “chartjunk”
  - **Use the right tool**
  - Message and readability trump aesthetics
  - Adapt the figure to support your medium





Ten Simple Rules for Better Figures: N.P. Rougier , M. Droettboom, P.E. Bourne

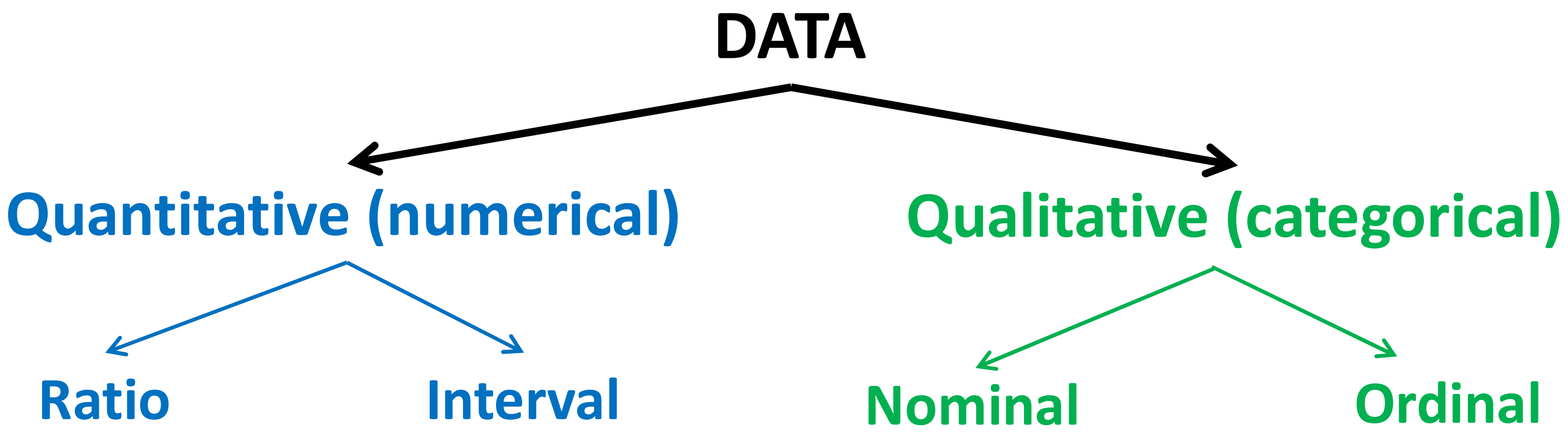
Common visualization mistakes



Source: http://blog.hubspot.com/marketing/data-visualization-mistakes

Data types

- **Quantitative vs Qualitative data**



DATA

Quantitative (numerical)

Qualitative (categorical)

Ratio

Interval

Nominal

Ordinal

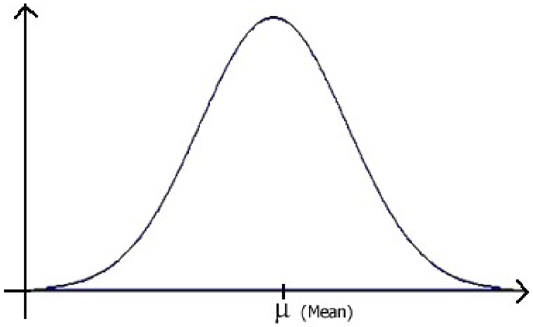
- The difference between the two can be established by asking the following 3 questions:
  1. Ordered: Can the data be ordered meaningfully?
  2. Equidistant: Is the difference between adjacent data points or categories consistent?
  3. Meaningful zero: Does the scale of measurements include a unique, non-arbitrary “zero” value?

Data types (cont'd)

- **Ratio scale:**
  - Interval variables with the added condition that zero of the measurement indicates that there is none of that variable. Has a true zero point.
  - Ex: weight, height, etc.
- **Interval scale:**
  - Has a fixed size of difference between data points with a no true zero point
  - Ex: Temperature (0 ^{o}C doesn't mean that there is no temperature)
- **Nominal scale:**
  - Categories with no inherent order between
  - circle-ellipse-square, eye color
  - A common and special case: Binary scale (1/0, True/False, Male-Female)
- **Ordinal scale:**
  - Categories that can be logically arranged in a meaningful order (but no distance)
  - Ex: low-medium-high, cold-cool-warm-hot, good-better-best

Major branches of Statistics

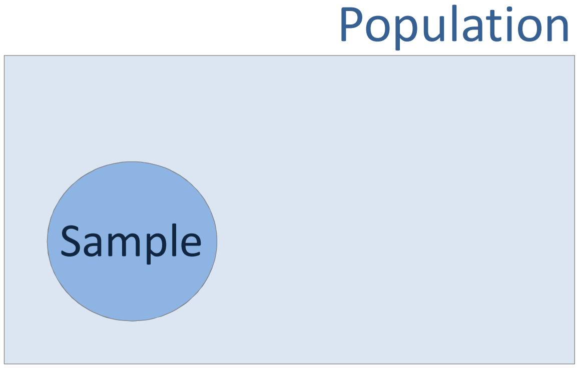
- **Descriptive Statistics**
  - Organizes, describes and summarizes charactersitics of data.
  - Includes construction of graphs, charts, tables and the calculation of various numeric measures such as mean, median, standard deviation, percentiles, etc.
  - It doesn't involve generalizing beyond the data at hand.



- Example:
  - Given the number of hits for a web site for the whole year, find out the average number of hits per week and state how much variation from the average exists.

Major branches of Statistics

- **Inferential Statistics**
  - Concerns with drawing conclusions or predictions about a **population** from the analysis of a random **sample** drawn from that population.
  - It includes methods like:
    - Point & interval estimation
    - Hypothesis testing
    - Regression
    - Classification



- Example:
  - Testing the efficacy of a new medicine on a random sample of patients for curing a disease.