Utrecht, April 30th 2019

Business Intelligence

Lecture 02 - Descriptive Analystics Part A
Nature of Data, Statistical Modeling, and Visualization

Georg Krempl

Algorithmic Data Analysis
Information and Computing Sciences
Utrecht University, The Netherlands

With particular thanks to

- Armel Lefebvre (tutor, A.E.J.Lefebvre@uu.nl)
- Koen Niemeijer (student teaching assistant)
- Jordan v. Dijk (student teaching assistant)

Outline: Today's Lecture in the Context of Business Intelligence

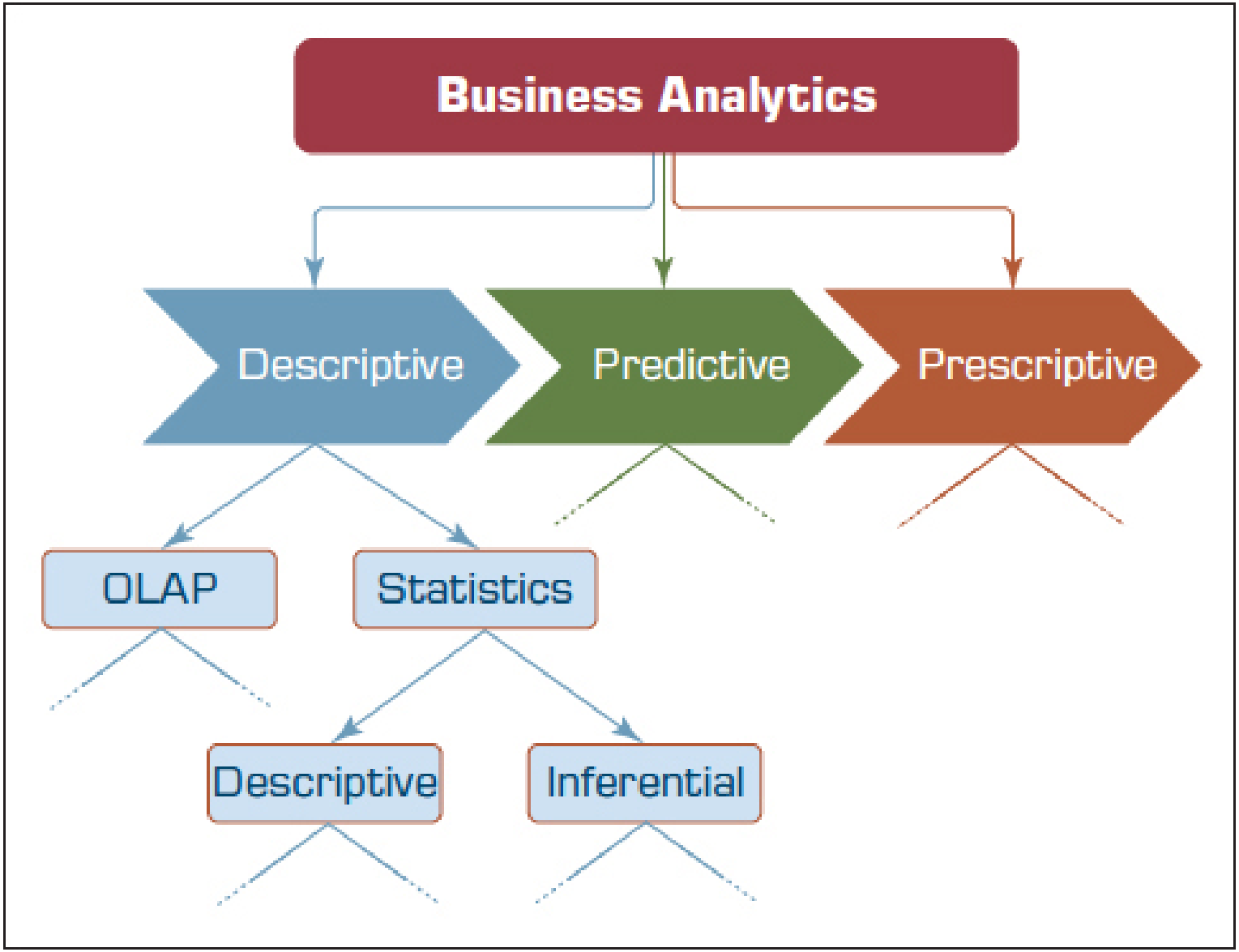


Figure: Context and Types of Descriptive Analytics within Business Intelligence (Source: [Sharda et al., 2018, page 101])

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Descriptive Analytics Part A: Nature of Data, Statistical Modeling, and Visualization

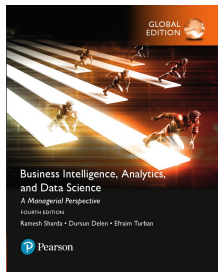


Figure: Textbook [Sharda et al., 2018]
Sharda, Delen, Turban & King (2018). Business Intelligence, Analytics & Data Science: A Managerial Perspective 4th Global Edition, Pearson. ISBN-13: 9781292220567

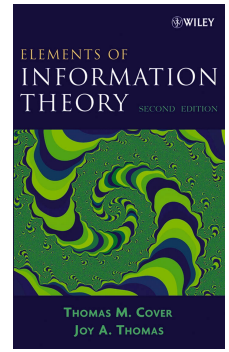


Figure: Textbook [Cover and Thomas, 2006]
Cover, Thomas (2006). Elements of Information Theory. 2nd Edition, Wiley Pearson. ISBN-13: 3 978-0-471-24195-9

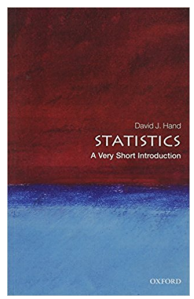


Figure: Textbook [Hand, 2008]
Hand (2008). Statistics: A very short introduction. Oxford University Press. 978-0-19-923356-4

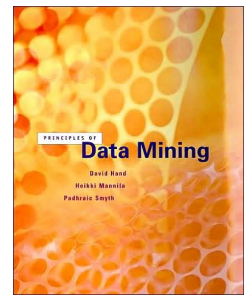


Figure: Textbook [Hand et al., 2001]
Hand, Mannila, Smyth (2001). Principles of Data Mining. The MIT Press . ISBN 978-0262082907

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data as a Source for Knowledge

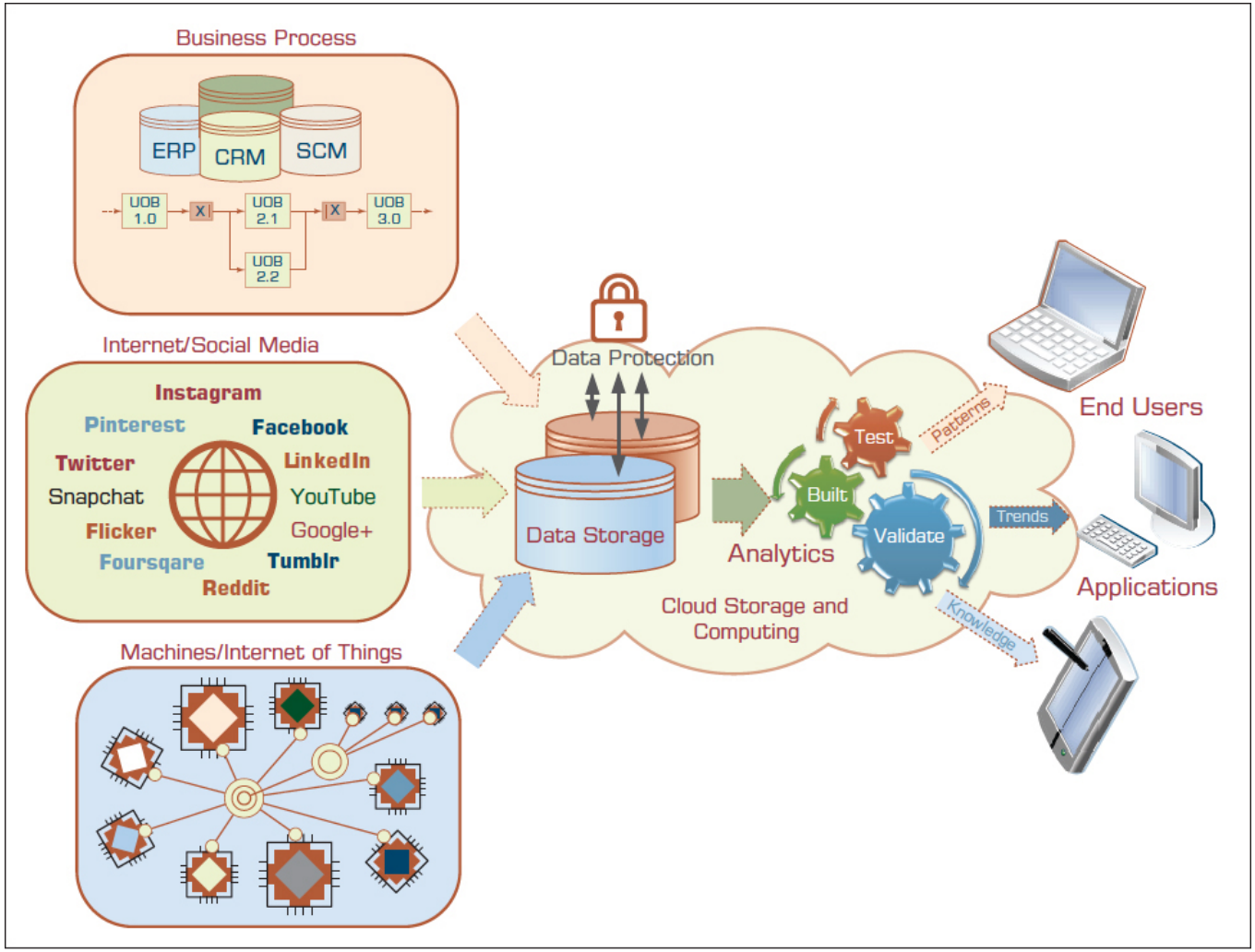


Figure: Data to Knowledge (Source: [Sharda et al., 2018, page 84])

The Nature of Data

Definition of Data

- Data (latin sg. Datum, pl. Data)
- "A collection of facts, obtained e.g., through observations, experiences, experiments, transactions." [Sharda et al., 2018]
- Lowest level of abstraction.
  The source for deriving information, insight, and knowledge
- As such, the quality of the analytics' output critically depends on the quality of the data going into it
  "Garbage in, garbage out"!
- Aim: Making data "analytics ready"
  Understanding characteristics of data and subsequent analytics
  Selecting and performing appropriate preprocessing

The Nature of Data: Taxonomies of Data

Appearance

- as numbers
- as categories
- as words, text, ...
- as images, video stream, ...
- ...

Sources

- Databases
  E.g. normalized OLTP data, or non-normalized OLAP data from a data warehouse, or from legacy databases
- external data
  E.g., web data provided by APIs, or in flat files like CSV-files, sensor data
- ...

The Nature of Data: Taxonomies of Data (2)

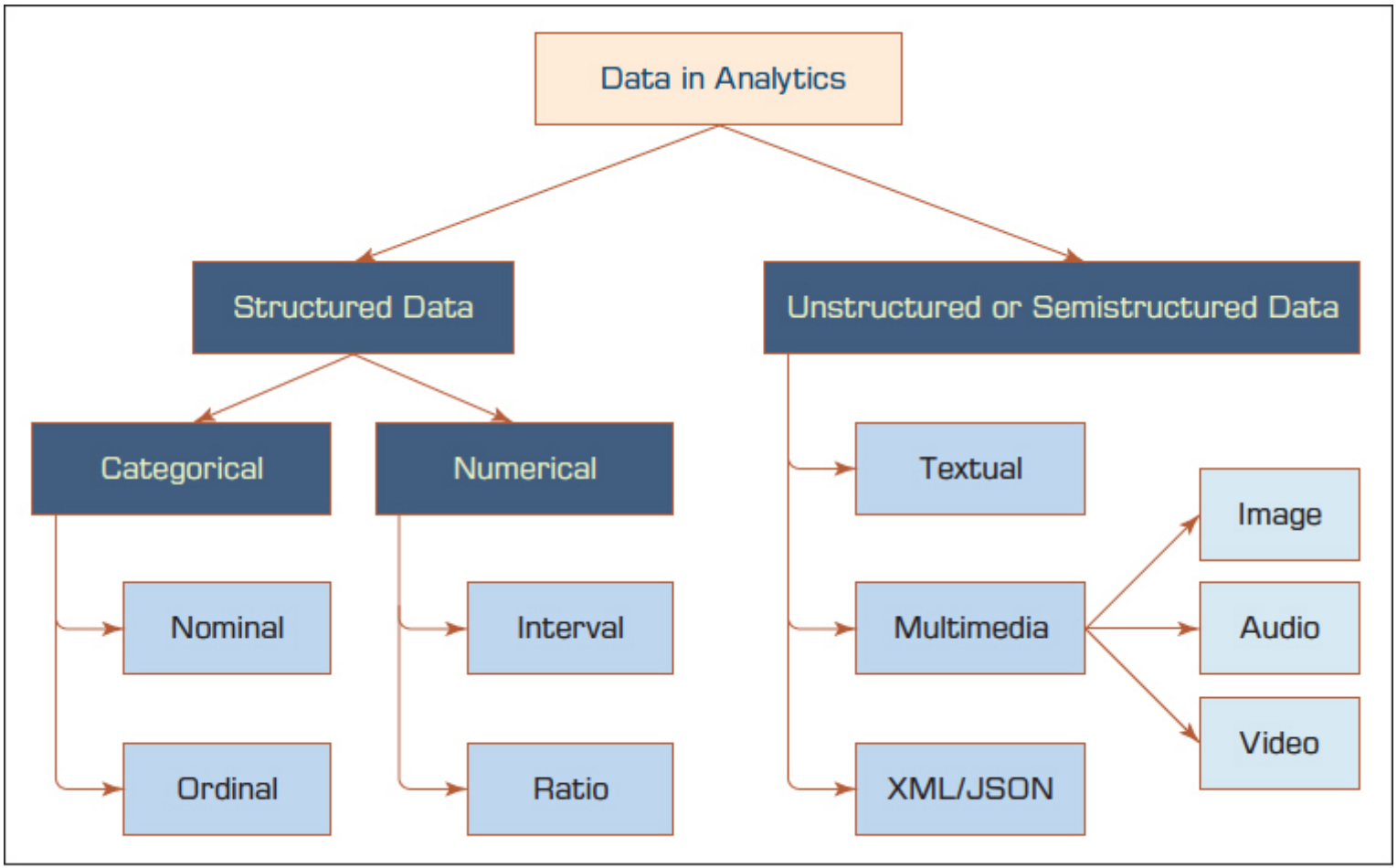


Figure: Taxonomy based on structure and levels of measurements (Source: [Sharda et al., 2018, page 87])

The Nature of Data: Taxonomies of Data (2)

Level of Structure

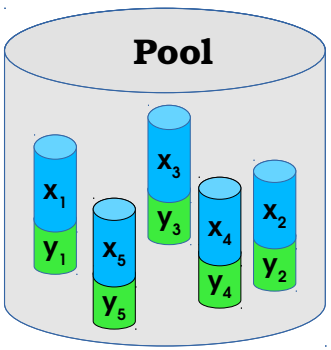
- Structured
  Being in a specific format that allows processing/interpretation by a computer
- Semi-structured (e.g., XML) or unstructured (e.g., text)
  Requires specialized analytics techniques (e.g. text mining) or transformation into structured data before analysing

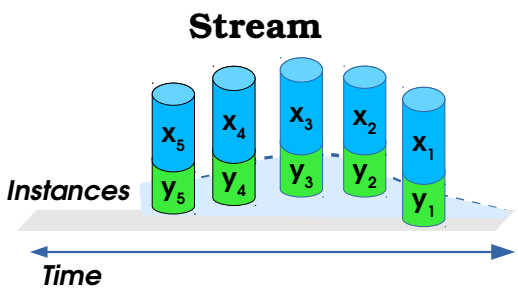
Level of Measurements

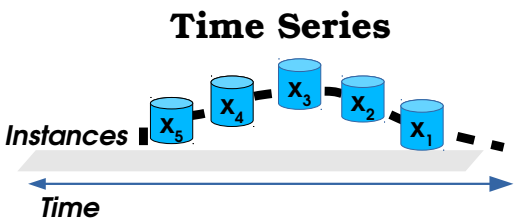
- categorical
  - nominal
    E.g., gender, nationality, product category
    - mode
  - ordinal: the order or rank of the data is meaningful
    E.g., grades from "very good" to "fail"
    - also median
- numerical
  - interval: interval between observations is meaningful (fixed unit of measure) E.g., temperature in Celcius
    - + (arithmetic) mean
      + standard deviation
  - ratio: the ratio of two values is meaningful
    E.g., length, income
    - + geometric mean
      + harmonic mean
      + variation coeff.

The Nature of Data: Taxonomy by Arrival / Availability of Data

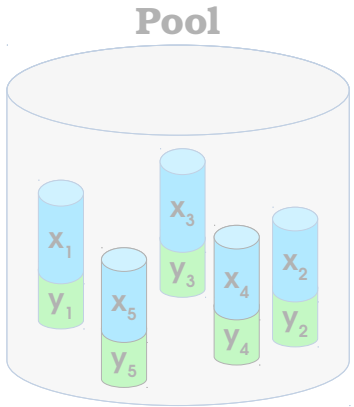
- Pool/Batch
  all datapoints are available at once
- Streaming
  Datapoints arrive *sequentially*
  - in chunks
  - one-by-one
- Time Series
  - Time-stamped data
  - Might become available sequentially or all at once
- Other types of sequential data . . .

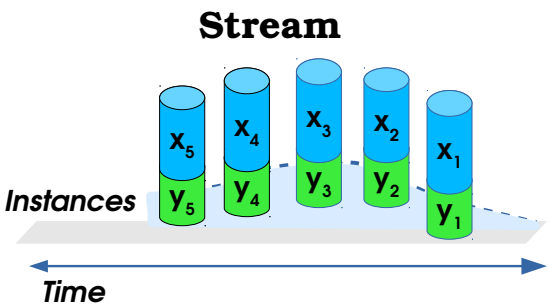


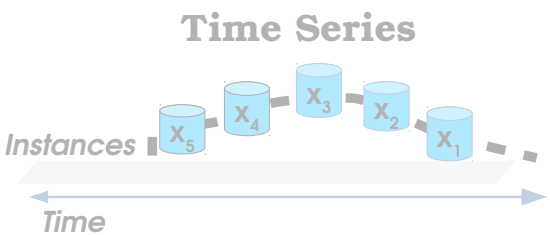




The Nature of Data: Arrival/Timing Scenarios







Data Stream

- Instances arrive sequentially
- Few instances available at beginning
- Possibly infinite number of instances
- Non-stationary distributions (drift)
- "Big Data" is often streaming data

General Challenges

- Limited computational resources
  - Online processing, limited storage capacity
  - Chunk-based vs. instance-wise processing
- Adaptation to change

The Nature of Data: Stationary vs. Non-Stationary Distributions

- Why should we bother about (non-)stationarity?
- What is the "correct" distribution of these points?
- The distribution varies over time (non-stationary distribution)!
- Better modelled considering time, e.g., as non-stationary data stream
- Chunk-based (several instances arrive at once) vs. instance-based (one-by-one)

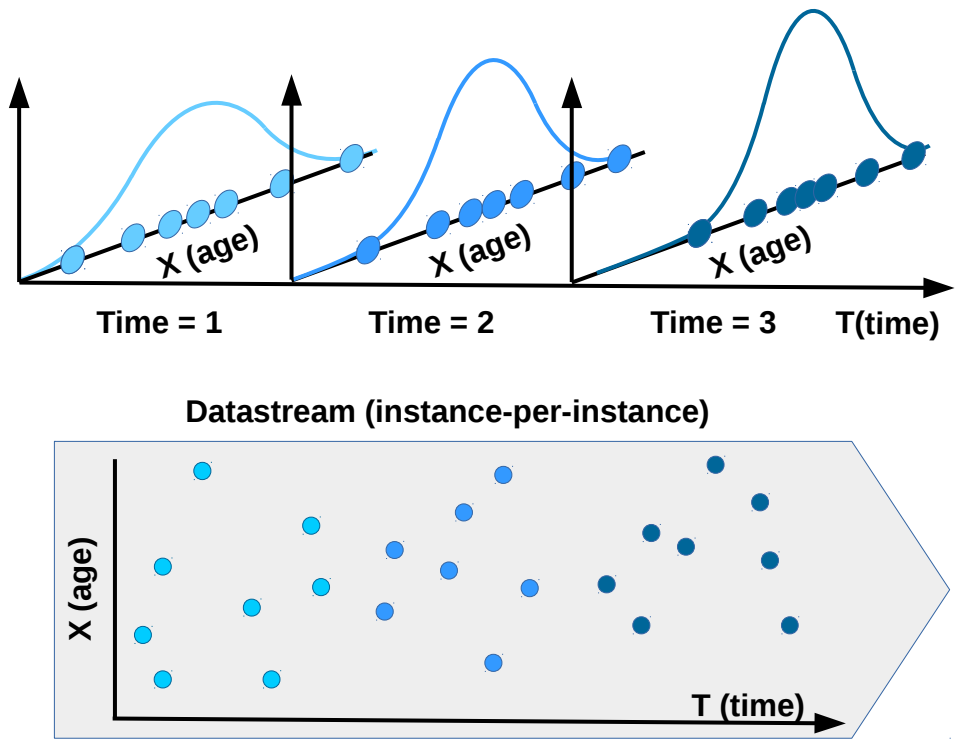


Figure: Datapoints with/without temporal information

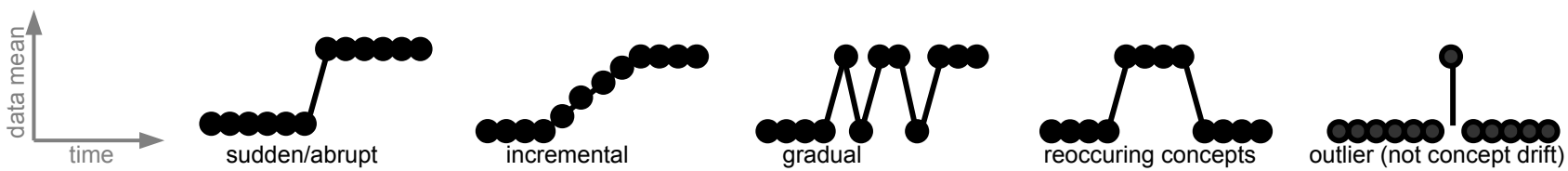
The Nature of Data: Types of Drift

Phenomenon known as

- *Concept drift*, e.g. in [Schlimmer and Granger, 1986]
- *Population drift*, e.g. in [Kelly et al., 1999]
- Related to *dataset shift* [Quionero-Candela et al., 2009]
- Broader Terms: *non-stationary*, *evolving* or *changing* data

Categorizing Drift[1]

- By the affected distributions: $P(X, Y)$, $P(X)$, $P(Y)$, $P(Y|X)$, $P(X|Y)$
- Smoothness of concept transition: sudden shift vs. gradual drift
- Singular or recurring contexts: recurring: "obsolete" data & models gain relevance again
- Systematic or unsystematic: E.g. distributions change according to patterns
- Real or virtual: real: changes affects the decision boundary



[2]See e.g., [Zliobait, 2009, Hofer and Krempl, 2013, Krempl et al., 2014, Webb et al., 2017].

[2]Illustration from [Zliobaite et al., 2016]

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Quality and Integrity

Objective & Approach

- make data analytics ready, i.e.
  organize it in a format that fits the needs of the subsequent analytics approaches
- Example: prepare for a particular classifier, e.g., logistic regression:
  flat file, transformation to numerical attributes, scaling, missing values imputation, etc.

Question

- What are quality or integrity problems?
- Why do they arise?
- Why are they problematic?

Data Quality and Usability Metrics[3]

- Data source reliability
  Do we have the right confidence and believe in this data source?
  Prefer original source rather than 2nd hand data
- Data content accuracy
  Correctness and appropriateness of data for objective of our analysis
  E.g., right data (e.g., scoring: customer residence/work address),
  Intention by original data provider correctly reflected?
  E.g., Is a zero a missing value or a value of zero, e.g. in income?
- Data accessibility
  Are the data accessible when needed?
  The type of data used during development of a prediction model should be available when deploying the model
  E.g., 3rd-party data (e.g., credit bureau/Facebook)

[3] See [Sharda et al., 2018, page 85–86]

Data Quality and Usability Metrics[4] (2)

- Data security and data privacy
  Ensure that solely authorized people have read/write access to the data
  Ensure no data are lost due to, e.g., system failures or attacks
- Data richness / comprehensiveness
  Are all required data elements (variables) included in the data set?
  E.g., potentially confounding variables
- Data consistency
  Are the data accurately collected and combined/merged?
  E.g., when merging data (attribute values) about the same subject from different sources, this data needs to be consistently assigned to the same subject
- Data currency/data timeliness
  Are the data as recent/new as required?
  Best practice is to enter an observation as soon as it is made,
  in order to avoid incorrect remembering/encoding later,

[4] See [Sharda et al., 2018, page 85–86]

Data Quality and Usability Metrics[5] (3)

- Data granularity
  Are the variables and data values defined in the adequate (e.g., lowest) level of detail for their intended use?
  E.g., customer income encoded as numerical value (vs. discretization)
- Data validity
  Are the actual and expected data values aligned?
  E.g., age wrongly defined as range between 0 and 99
- Data relevancy
  Are the variables included in the data set relevant for objective of our analysis?

  Note: Inclusion of variables of lesser relevance depends on the analysis/algorithm
  E.g., some similarity/distance calculations might be difficult with large numbers of variables (see "curse of dimensionality")
  E.g., when segmenting customers into groups using euclidean distance measure, the number of variables should be small (e.g., us Principal Component Analysis)

[5] See [Sharda et al., 2018, page 85–86]

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Preprocessing: Application Case 2.1

Questions

- What is student attrition, why is it an important problem in higher education?
- What were the traditional methods to deal with the attrition problem?
- List and discuss the data-related challenges within context of this case study.
- What was the proposed solution? And, what were the results?

Data Preprocessing: Application Case 2.1 (2)

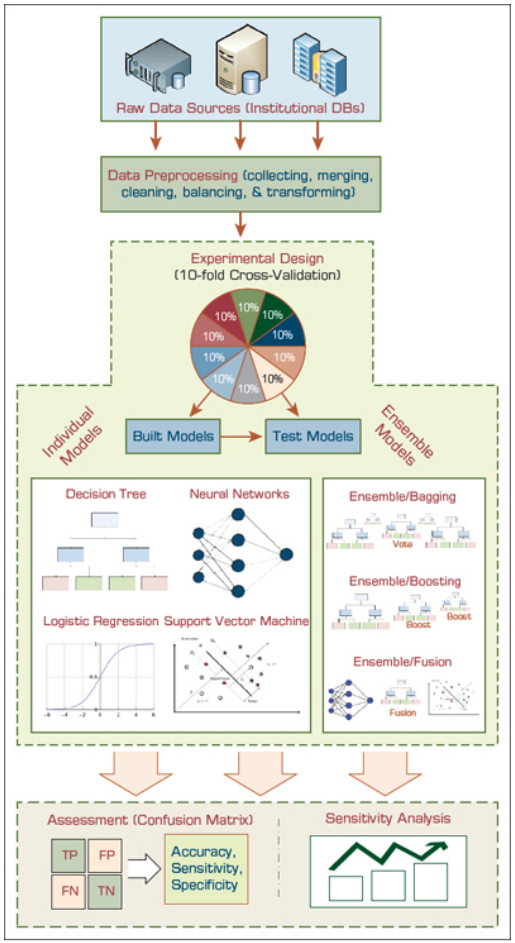


Figure: Preprocessing Steps (Source: [Sharda et al., 2018, p. 95])

- Improve student retention, understand reasons behind attrition, identify students at risk, provide individual support
- Traditional qualitative research fails to predict/suport it on an individual level
- Many variables, some related (e.g. hours earned/registered, years since enrollment)
- Majority of students continues, minority quits "Naive" classification rule is highly accurate, but not usable ("all continue")

Class Imbalance: Use Re-Sampling

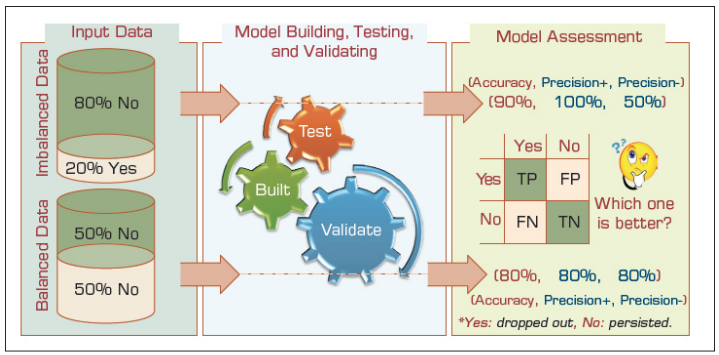


Figure: Class imbalance (Source: [Sharda et al., 2018, p. 97])

Data Preprocessing: Application Case 2.1 (3)

TABLE 2.2 Prediction Results for the Original/Unbalanced Dataset

	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	1494	384	1518	304	1478	255	1438	376
Yes	1596	11142	1572	11222	1612	11271	1652	11150
SUM	3090	11526	3090	11526	3090	11526	3090	11526
Per-Class Accuracy	48.35%	96.67%	49.13%	97.36%	47.83%	97.79%	46.54%	96.74%
Overall Accuracy	86.45%		87.16%		87.23%		86.12%	

TABLE 2.3 Prediction Results for the Balanced Data Set

Confusion Matrix	ANN(MLP)		DT(C5)		SVM		LR	
	No	Yes	No	Yes	No	Yes	No	Yes
No	2309	464	2311	417	2313	386	2125	626
Yes	781	2626	779	2673	777	2704	965	2464
SUM	3090	3090	3090	3090	3090	3090	3090	3090
Per-class Accuracy	74.72%	84.98%	74.79%	86.50%	74.85%	87.51%	68.77%	79.74%
Overall Accuracy	79.85%		80.65%		81.18%		74.26%	

TABLE 2.4 Prediction Results for the Three Ensemble Models

	Boosting (Boosted Trees)		Bagging (Random Forest)		Information Fusion (Weighted Average)	
	No	Yes	No	Yes	No	Yes
No	2242	375	2327	362	2335	351
Yes	848	2715	763	2728	755	2739
SUM	3090	3090	3090	3090	3090	3090
Per-Class Accuracy	72.56%	87.86%	75.31%	88.28%	75.57%	88.64%
Overall Accuracy	80.21%		81.80%		82.10%	

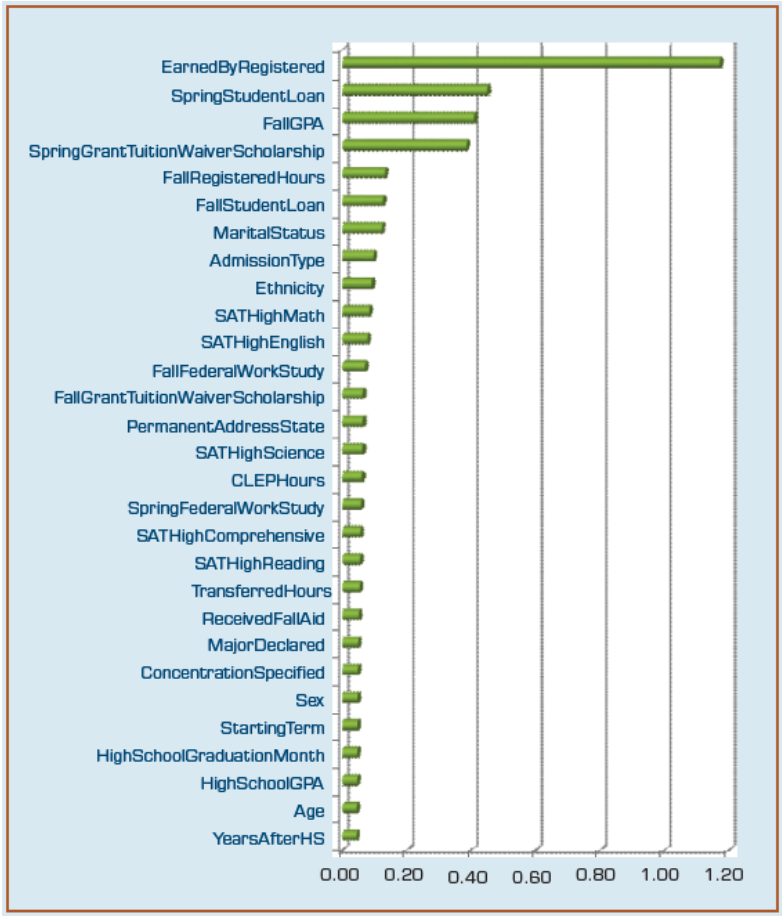


Figure: Results of the Student Attrition Analytics Study (Source: [Sharda et al., 2018, pages 97–99])

Data Preprocessing

Motivation

- Why do we need preprocessing? What is this?
- Real-World data is often dirty, misaligned, overly complex/irrelevant, inacurate
- Remember: garbage in-garbage out
- Improve quality of data, make it ready for analytics!

Preprocessing Steps: From **raw** to **well-formed** data

- Data consolidation
- Data cleaning
- Data transformation
- Data reduction

Example

- select the right attributes (variables) from data sources,
- clean, e.g. zeros that encode missing values, impute missing values
- normalize data, e.g., income vs age for euclidean distance
- reduce dimensions, e.g. PCA

Data Preprocessing: Application Case 2.1 (4)

Question

- What are the limits of the previous analysis in the application case 2.1?

(Short) Answer

Be careful with re-balancing, as some aspects of the learned model (e.g., class prior) are misleading.

Data Preprocessing Steps: Details

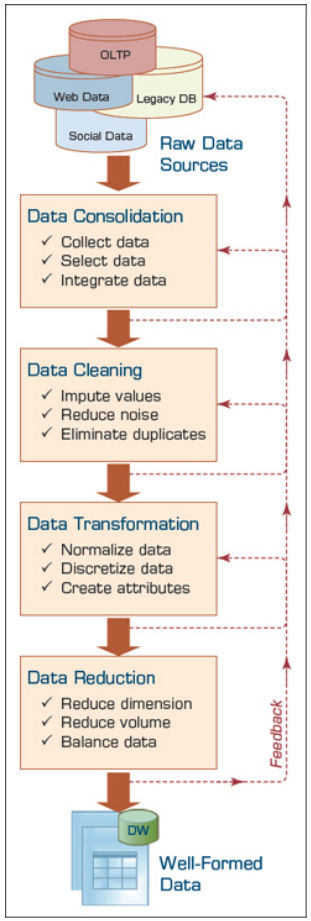


Figure: Data Preprocessing Steps (Source: [Sharda et al., 2018, page 92])

1. Data Consolitation
   - Collect, select, integrate data
   - select relevant data, filter out irrelevant;
   - integrate data from multiple sources
2. Data Cleaning
   - Impute (missing) values - missing at random?
   - Detect outliers, reduce noise
   - Eliminate duplicates
3. Data Transformation
   - Adjust data type according to capabilities/requirements of subsequent analytics tools, e.g., classifiers
   - Normalize data, e.g. range of attributes (e.g. for euclidean distance)
   - Discretize data
   - Create attributes, e.g., ratio between income and age
4. Data Reduction
   - Aim is to make data volume more manageable
   - Dimension reduction, e.g. PCA
   - Volume reduction, e.g., subsample
   - Balancing of data, e.g. in case of few positives

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Statistics: Random Variables [6]

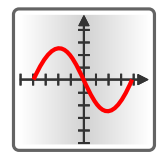
Examples

Throwing (fair) dices.

- Discrete outcomes (domain $\mathcal{X} = \{1, 2, \ldots, 6\}$),
- with probabilities $\Pr(X = x_i) = \frac{1}{6}$
  $\{\Pr(x_1), \Pr(x_2), \ldots, \Pr(x_6)\}$ is the *probability mass function*
- When orderable values, $\Pr(X \leq x) = F(x)$
  is the *cumulative distribution function*

Return of an investment.

- Continuous outcomes (domain $\mathcal{X} = \mathbb{R}$),
- with *cumulative distribution function* $F(x)$ as above,
- with probability of $\Pr(X \in [a, b]) = F(b) - F(a)$
- with *probability density function* $f(x)$ expressing the probability, that the value will be in an infinitesimal interval around x



[5] Images: Nuvola icon theme KDE 3.x. Source: `commons.wikimedia.org`

[6] See [Hand et al., 2001, Appendix A.1: Review of Univariate Random Variables].

Statistics: Random Variables

Definition

- A random "variable" X who's values are outcomes of a random phenomenon
- Actually, a *function* $X : \mathcal{X} \to \mathbf{E}$
- **Domain**: the set of all possible values (outcomes) the variable can take
  - e.g., the discrete domain $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ for the results of throwing dices,
  - or the continuous domain $\mathcal{X} = \mathcal{R}$ for the return of an investment,
- Notation:
  - Capital letters for Variables (e.g. X),
  - lowercase letters (e.g. x) for values
- Different types of domains and levels of measurement (see slide 10)
- Probability of a value:
  - Frequentist: Outcome
    How often did the values occur in an experiment?
  - Bayesian: Belief, i.e. a measure of plausibility of propositions
    How often should one expect a value as an outcome of an experiment?

Statistics: Distribution of Random Variables

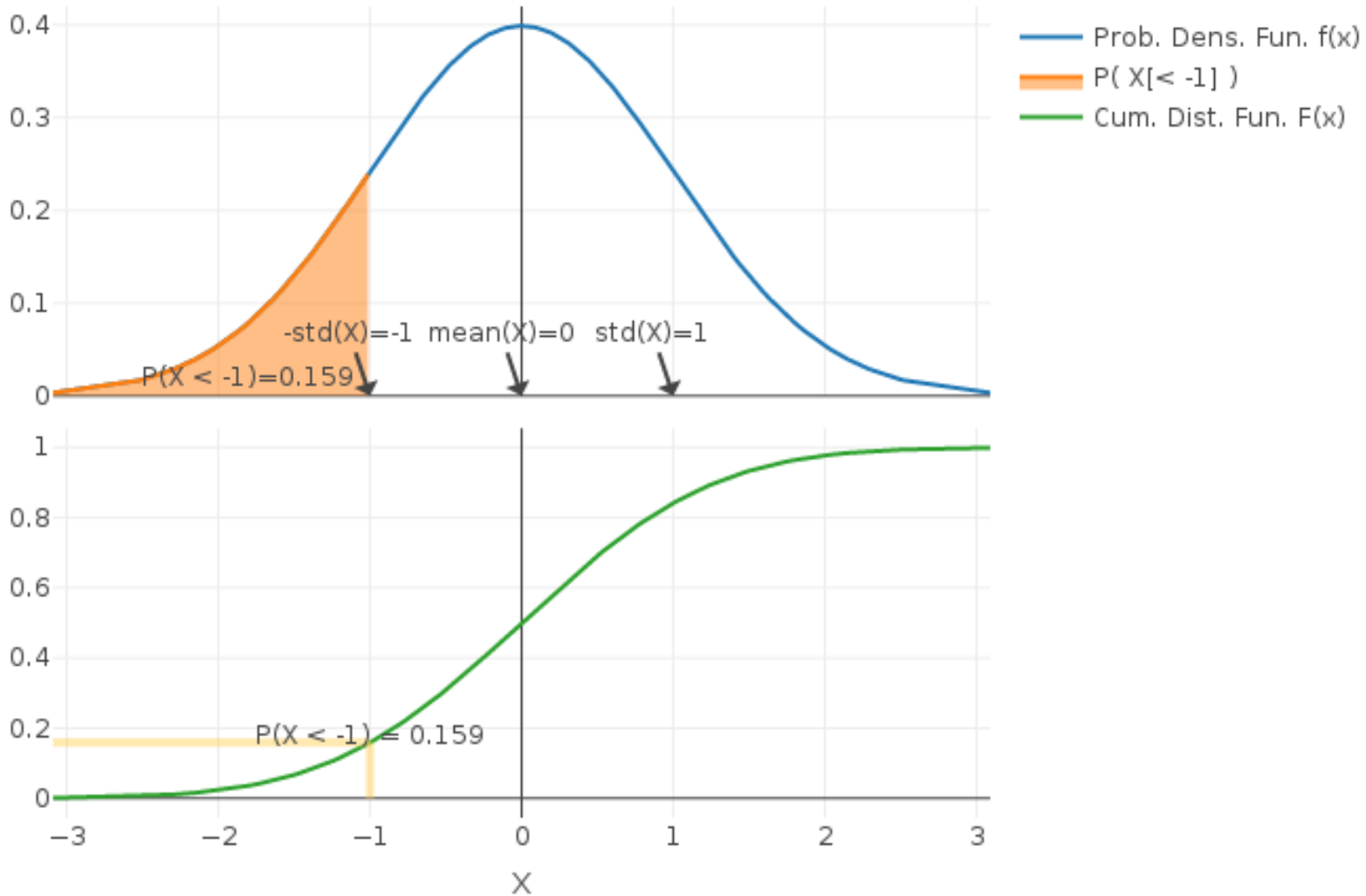


Figure: Probability Density Function (PDF) $f(x)$, Cumulative Distribution Function (CDF) $F(x)$ of a Gaussian distributed variable X (standard normal with *mean* $= 0$, *std* $= 1$)

Statistics: Visualization with a Histogram

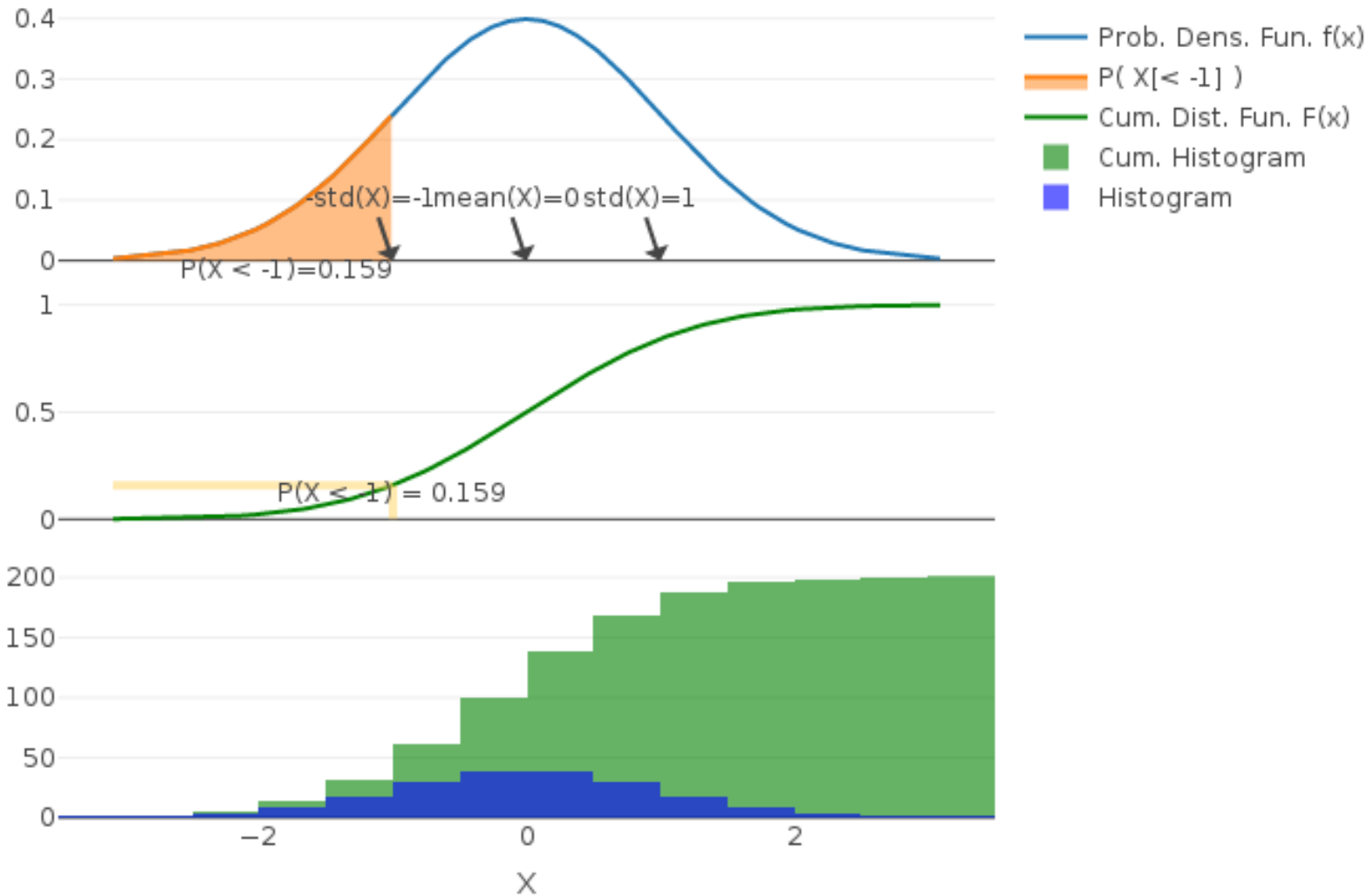


Figure: PDF, CDF, Histogram and Cumulative Histogram of a Gaussian distributed variable X (standard normal with $mean = 0$, $std = 1$)

Statistics: Visualization with a Box Plot

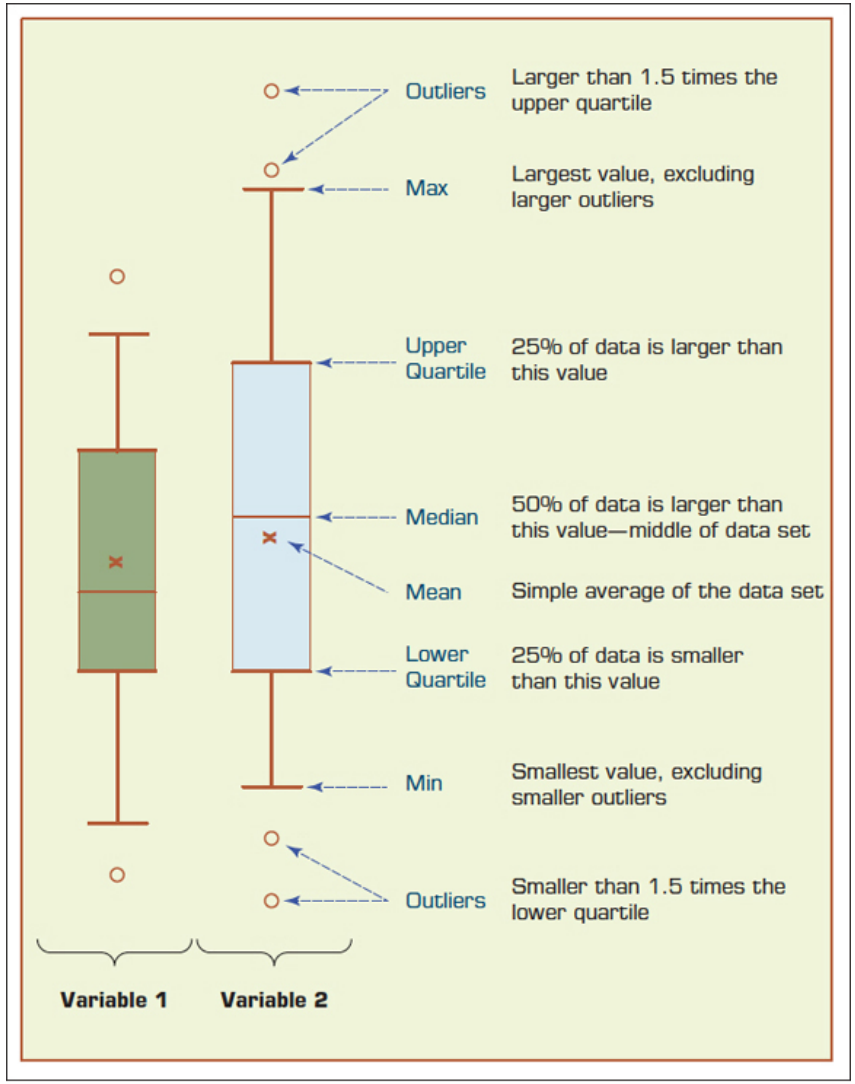


Figure: Box-Plot (Source: [Sharda et al., 2018, page 105])

Statistics: Visualizing Univariate Data

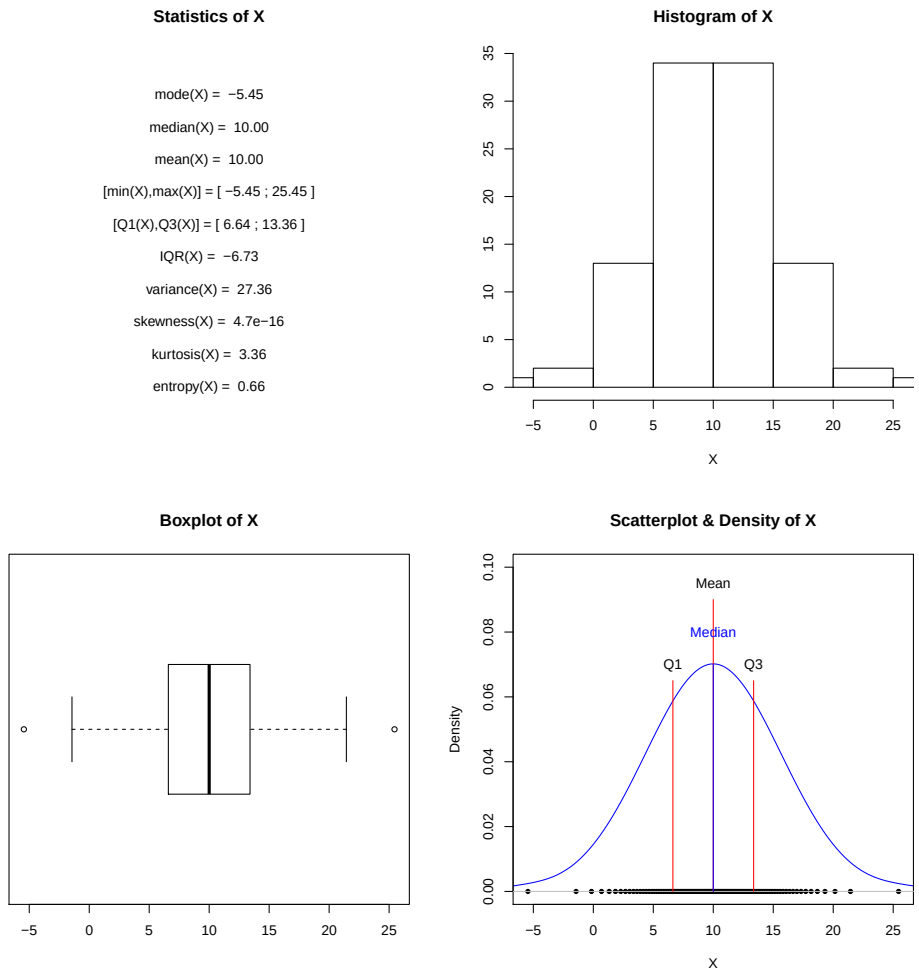


Figure: Statistics & Visualizations of a Gaussian Normal Distributed Variable

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Subfields of Statistics:

Descriptive Statistics: describing the data sample (as it is)

- Used in descriptive analytics!

Inferential Statistics: generalizing, drawing conclusions about the population

- Used in predictive analytics!

Caveat

- Our objective is to *describe* data & facts
- That is, summarizing by focusing on the important aspects
- Involves a loss of (hopefully irrelevant) information
- Always ask what the summary is able to tell and what not!
  Summaries and largely compressed information (e.g., in KPIs) might be misleading! You are supposed to be the one to identify potential issues!

Outline of Statistical Modelling

- Describing univariate data: measuring central tendency
- Describing univariate data: measuring dispersion
- Describing univariate data: measuring skewness & kurtosis
- Describing (uni-/bi-)variate data: measuring information content
- Describing bi-/multivariate data
- Visualization

Describing Univariate Data: Measuring Central Tendency

Main Objective

- Location of the "gravity" centre of the observations

Mode

- mode: nominal scale
- most frequent value
- in real-valued attributes a distinct mode can help detect spurious values (e.g., a zero encoding a missing value)

Median

- at least ordinal scale
- 50% of observations smaller, 50% greater

Describing Univariate Data: Measuring Central Tendency (2)

(Arithmetic) Mean

- at least interval scale

$$mean(x) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Note: susceptible to outliers!

(Geometric) Mean

- at least ratio scale

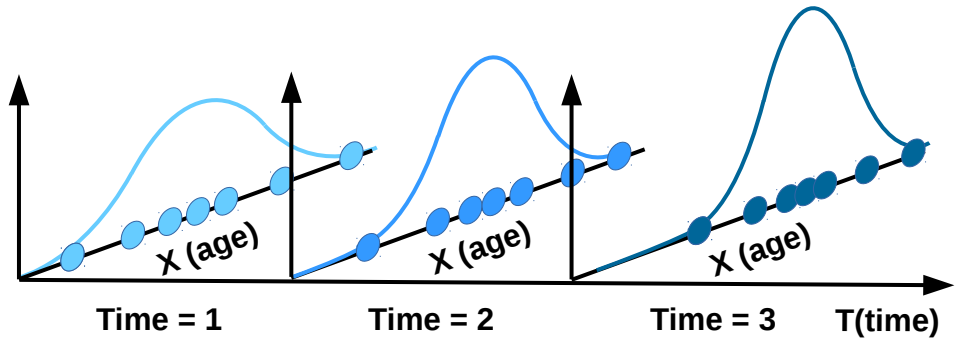
$$geometricmean(x) = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

Describing Univariate Data: Measuring Central Tendency (3)

A Note on Evolving/Streaming Data

- Streaming data: what to do?
- Computational Efficiency:
  Calculate incrementally:
  Store partial terms
  E.g., partial sums

  Caveat: Overflows!
- Non-Stationarity:
  - Forgetting:
    Drop old instance,
    annulate their contribution
  - Windowing:
    Use only subsample
    with most recent instances
- Further reading:
  [Meng, 2015,
  Babcock et al., 2003]



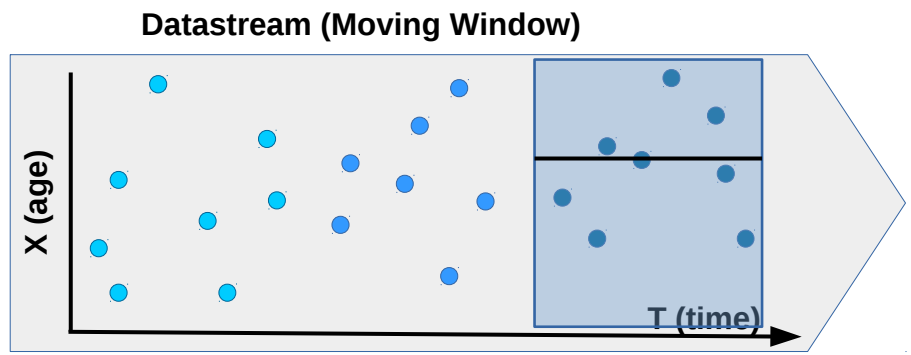


Figure: Prossing Streaming Data: Incrementally or with Moving Windows

Describing Univariate Data: Measuring Dispersion

Main Objective

- Degree of variation in the observations
- Range that these observations span
- Note: influence of noise on interquartile range and standard deviation

Range

- Maximum - Minimum Value

$$range = \max_{x \in X} x - \min_{x \in X} x$$

Quartiles and Interquartile Range

- 1st Quartile: Q1: 25% observations smaller, 75% greater
- 3rd Quartile: Q3: 75% observations smaller, 25% greater
- Interquartile Range: Distance between 1st and 3rd Quartiles
  Interquartile Range = Q3-Q1
- Related: Percentiles
- More robust (ignorant) against outliers than, e.g., variance

Describing Univariate Data: Measuring Dispersion (2)

Variance

- Measures squared differences from mean:

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

$(1-n)$ unbiased estimator of sample variance of iid observations
(n) second moment around mean

Mean Absolute Deviation

- Measures absolute values of differences from mean:

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

- Compared to standard deviation, lesser penalization of larger deviations

Describing Univariate Data: Measuring Dispersion (3)

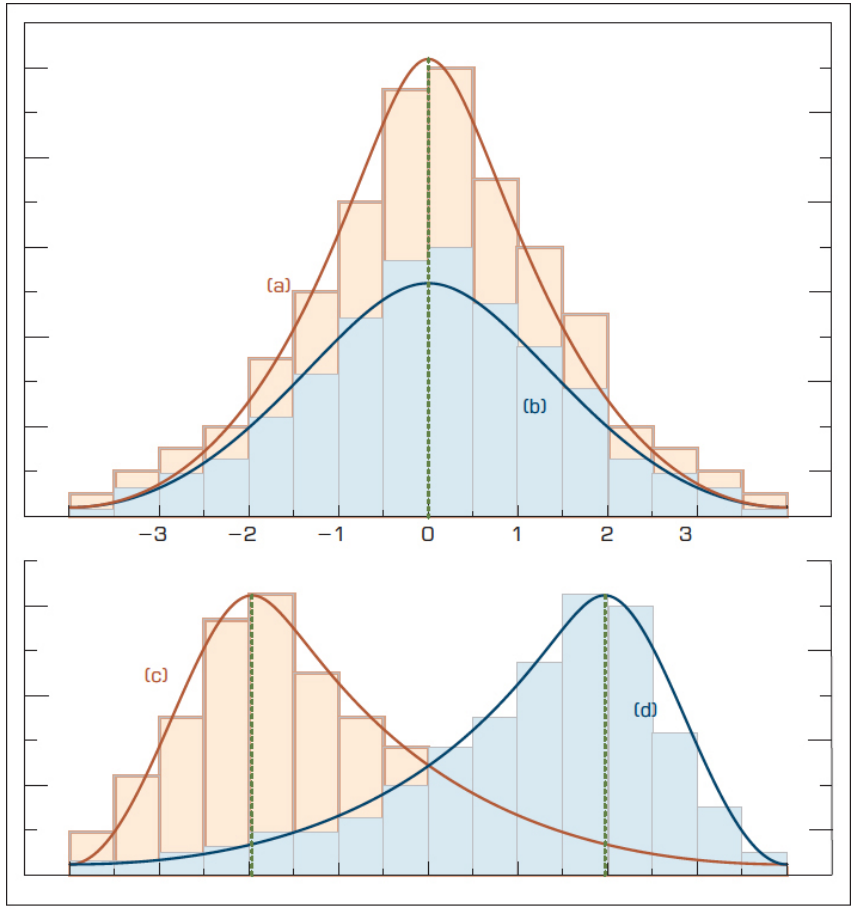


Figure: Dispersion and Shape Properties (Source: [Sharda et al., 2018, page 107])

Describing Univariate Data: Measuring Skweness and Kurtosis

Main Objective

- Describe shape of distribution to a higher degree
- Example: Symmetry

(Sample) Skewness[7]

- Describes symmetry of the data

$$skewness(X) = \frac{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}} \cdot \frac{n^2}{(n-1)(n-2)}$$

- Positive: tail on the right side, swayed left
  mean greater than median, e.g., point (c) in Fig. 18
- Negative: tail on the left side, swayed right, e.g., mean smaller than median, point (d) in Fig. 18
- **Note: the discussion of skewness is wrong in the book (see page 106)!**

[7] There are various existing definitions (see e.g., [Joanes and Gill, 1998])

Describing Univariate Data: Measuring Skweness and Kurtosis (2)

(Excess) Kurtosis[8]

- Describes the peak/tall/skinny nature of the data

$$kurtosis(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4} - 3$$

- A normal distribution has a kurtosis of 3

Further Generalizations: Moments

$$moment_p(X) = \mathbb{E}\left[(X - \bar{x})^p\right] = \int_{-\infty}^{\infty} (X - \bar{x})^p P(x) dx$$

[8] This is actual the formula for **excess kurtosis**, where the kurtosis (see below) is adjusted such that it is zero for the Gaussian normal distribution.

$$kurtosis(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4}$$

Main Objective

- Measure the uncertainty (or information content)

Entropy of a single variable

- Entropy of a discrete random variable X

$$H(X) = -\sum_{x \in \mathcal{X}} \mathcal{P}r(x) \cdot \log(\mathcal{P}r(x))$$

  Convention: $0 \log 0 = 0$

- The higher the entropy, the more informative (& uncertain) is the result of X.
- E.g., if logarithm base 2 is used, the entropy of a fair coin toss is 1 bit.
- See [Cover and Thomas, 2006, chapter 2].

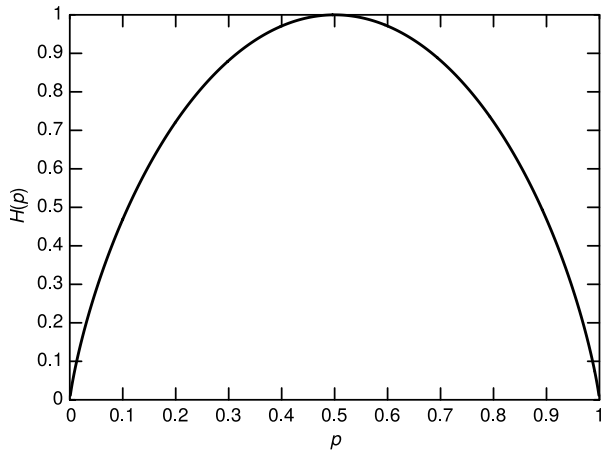


FIGURE 2.1. $H(p)$ vs. p.

Figure: Entropy of a binary, Bernoulli-distributed random variable against its success probability p. Source: [Cover and Thomas, 2006, p. 16]

Describing Similarity of Two Distributions: Measuring Divergence

Relative Entropy or Kullback-Leibler Divergence

- Describes a relationship between two distributions
- Measures the divergence between the two distributions
- Definition for discrete random variables p,q [9]:

$$D(p\|q) = \sum_{x\in\mathcal{X}} p(x) \cdot \log \frac{p(x)}{q(x)} = \mathrm{E}_p \log \frac{p(X)}{q(X)}$$

Conventions:

- $0 \log \frac{0}{0} = 0$
- $0 \log \frac{0}{q} = 0$ (for $q > 0$)
- $p \log \frac{p}{0} = \infty$ (for $p > 0$).

- Characteristics:
  - Given a true distribution p and an approximation q thereof, the relative entropy or Kullback-Leibler divergence $D(p\|q)$ measures the divergence of q from the actual distribution p.
  - It is non-negative
  - Zero if and only if $p = q$
  - Not symmetric, does not satisfy triangly inequality (thus no true distance measure)
  - $D(p\|q) = \infty$, if there is an $x \in \mathcal{X}$ such that $p(x) > 0$ and $q(x) = 0$

[9] See [Cover and Thomas, 2006, chapter 2.3, equation 2.26].

Describing Univariate Data: Gaussian

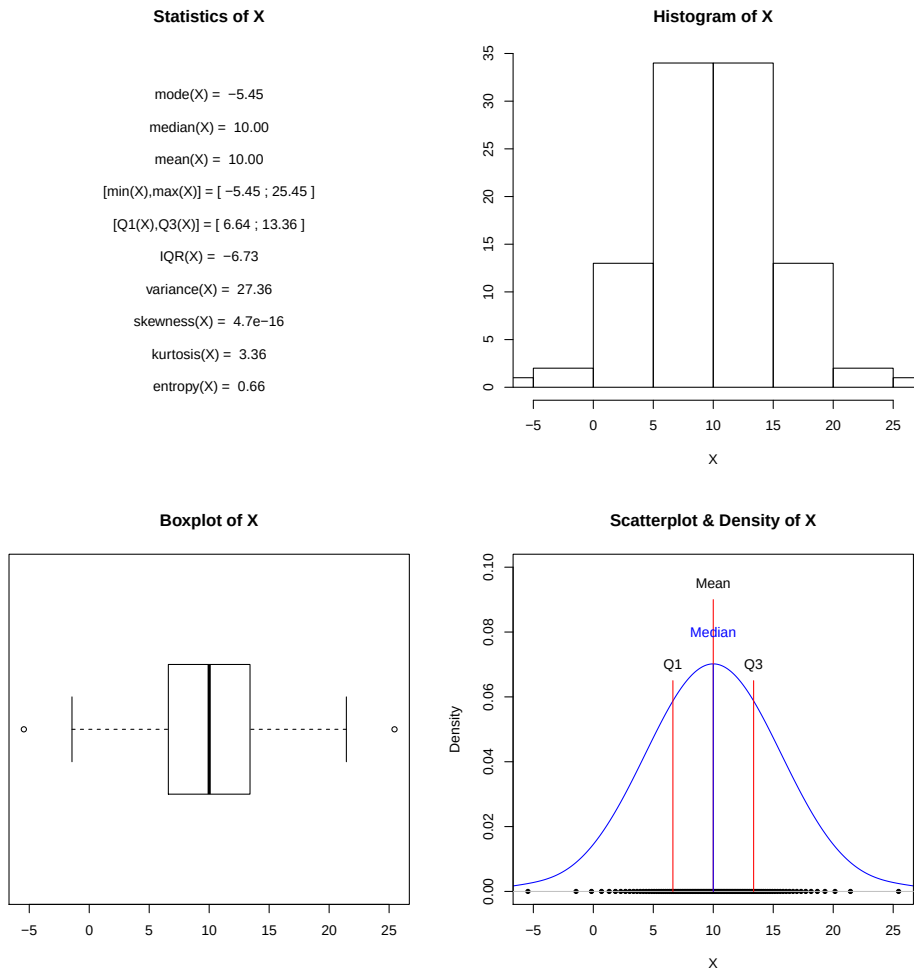


Figure: Statistics of a Gaussian Normal Distributed Variable

Describing Univariate Data: Peaked

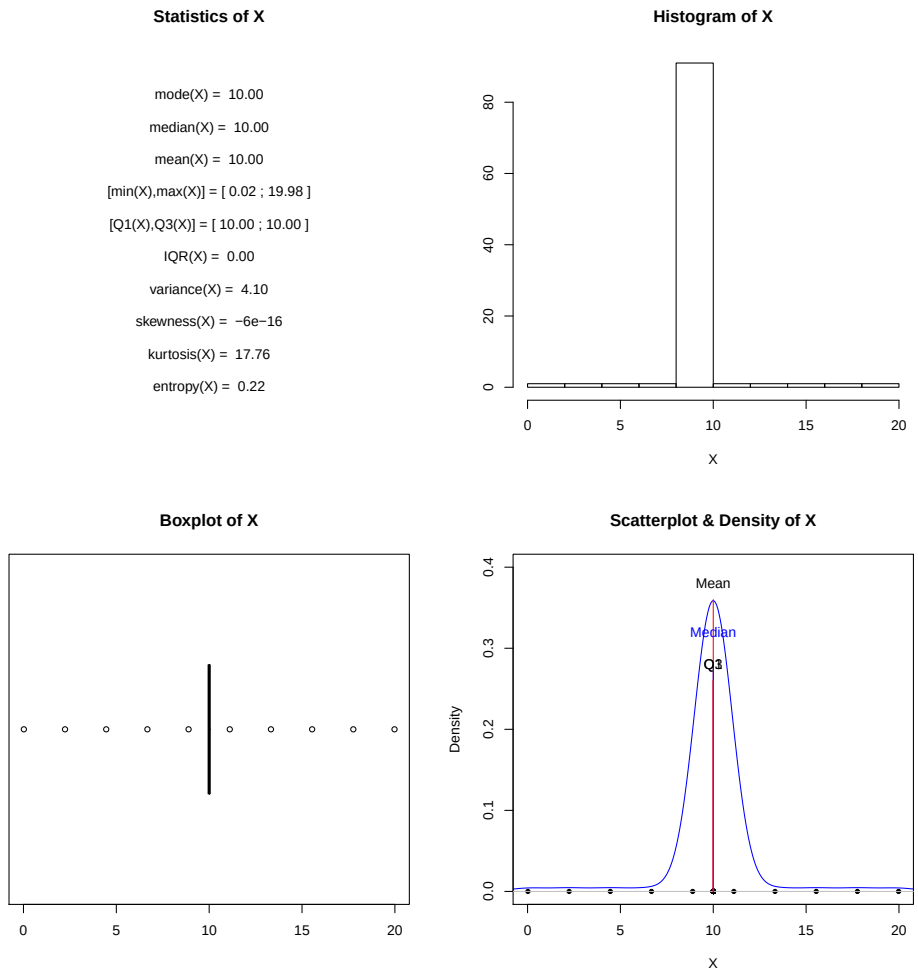


Figure: Statistics of an Variable with Distinct Mode

Describing Univariate Data: Bimodal / Gaussian Mixture

Boxplot is the same while the scatterplot clearly shows the difference with the unimodal/gaussian distribution

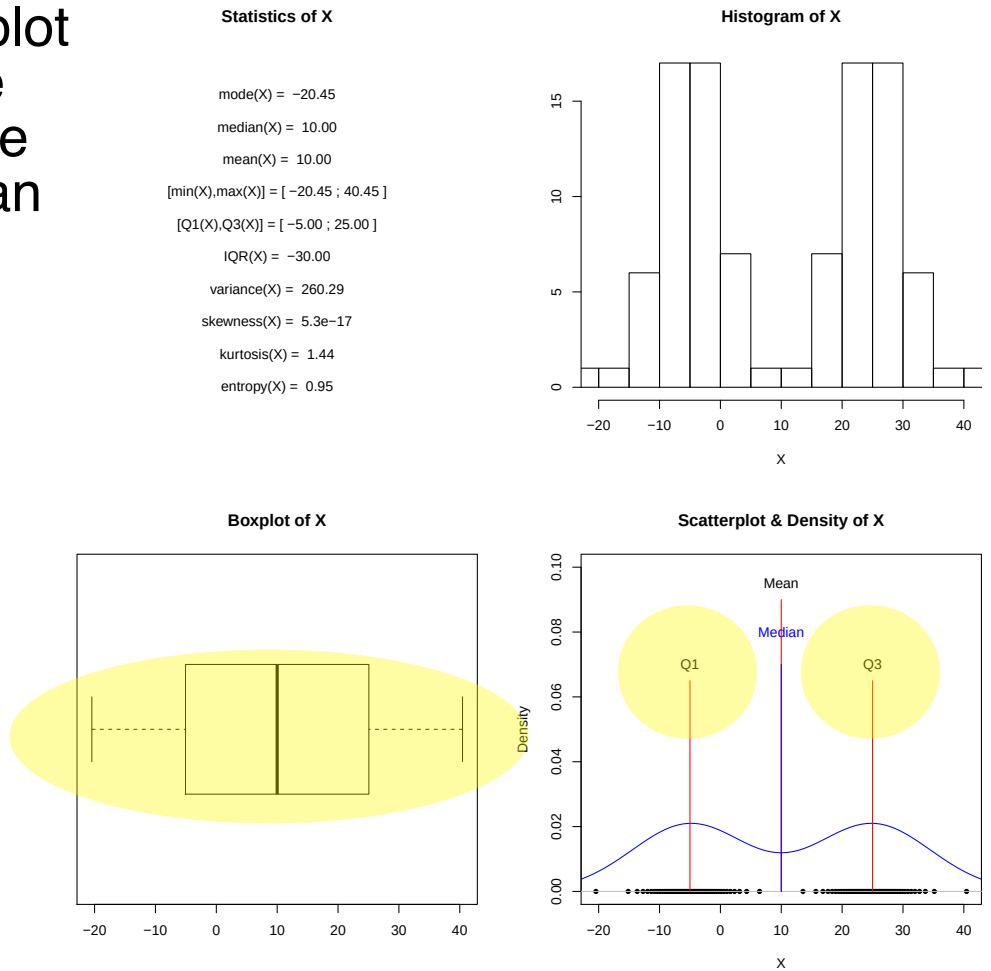


Figure: Statistics of a Bimodal Distribution (Gaussian Mixture)

Describing Univariate Data: Uniform

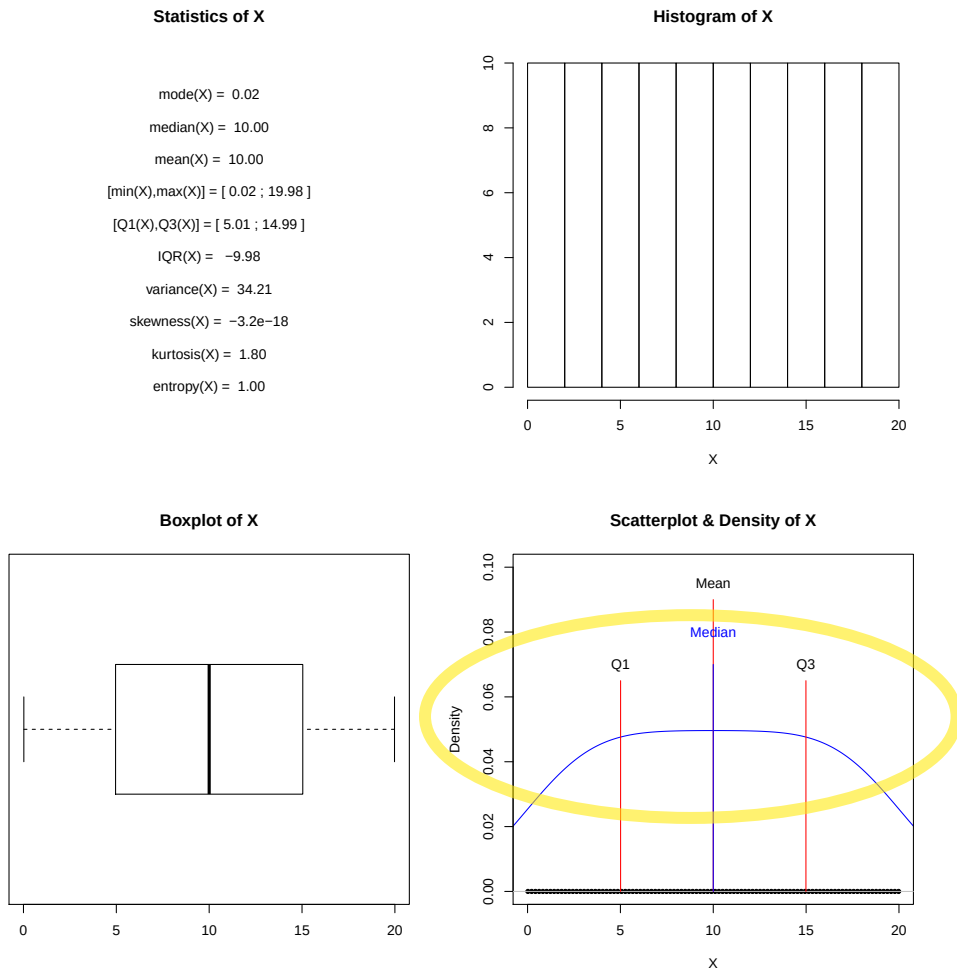


Figure: Statistics of a Uniformly Distributed Variable

Describing Univariate Data: Gaussian with Outlier

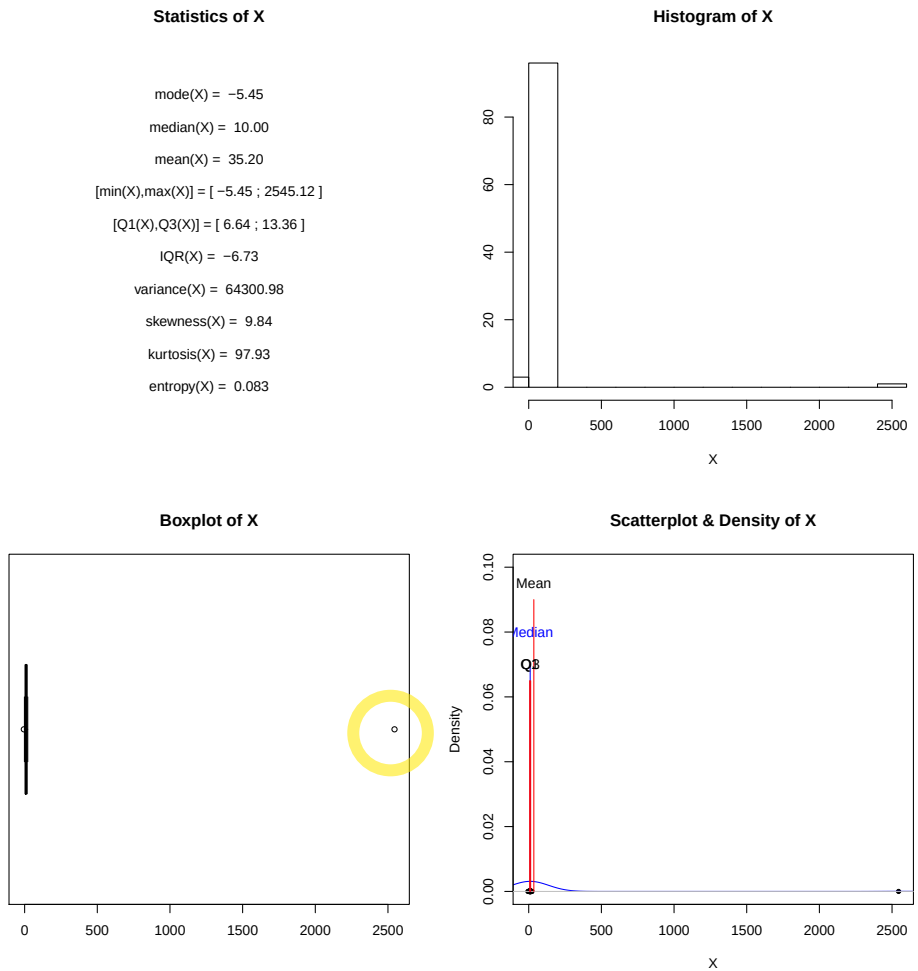


Figure: Statistics of a Gaussian Distributed Variable with Outlier

Describing Univariate Data: Lognormal

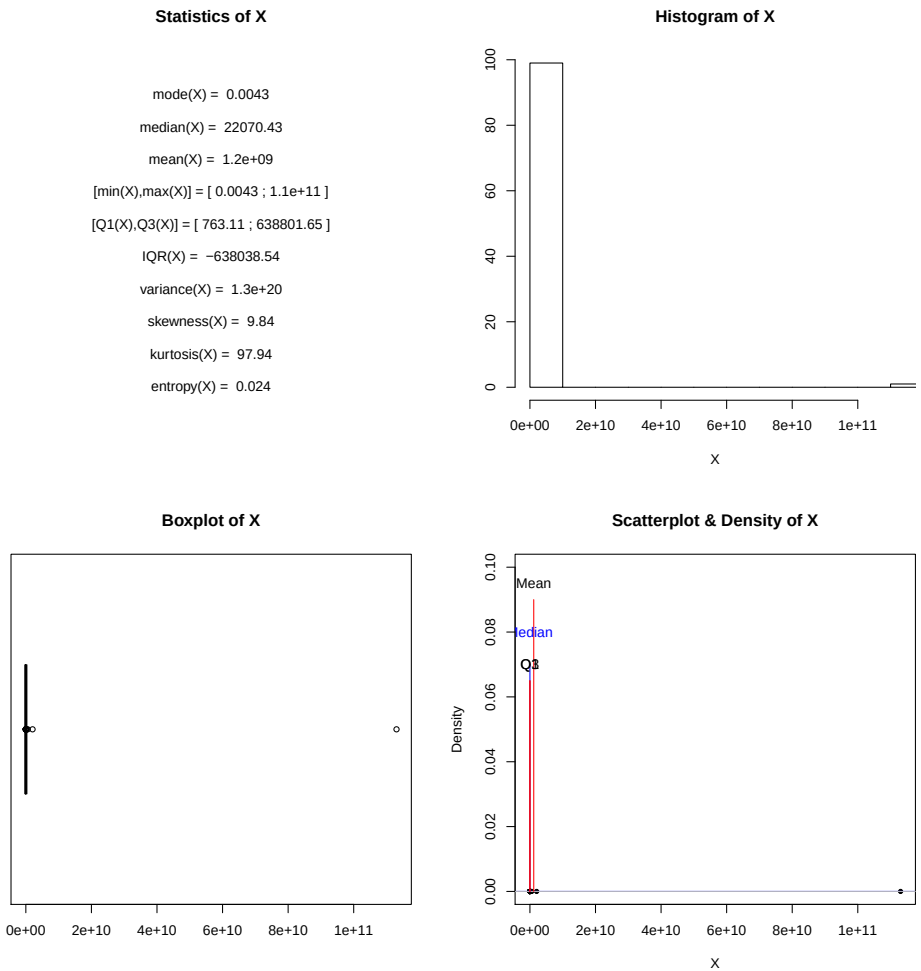


Figure: Statistics of a Gaussian Log-Normal Distributed Variable

Describing Univariate Data: Exponential

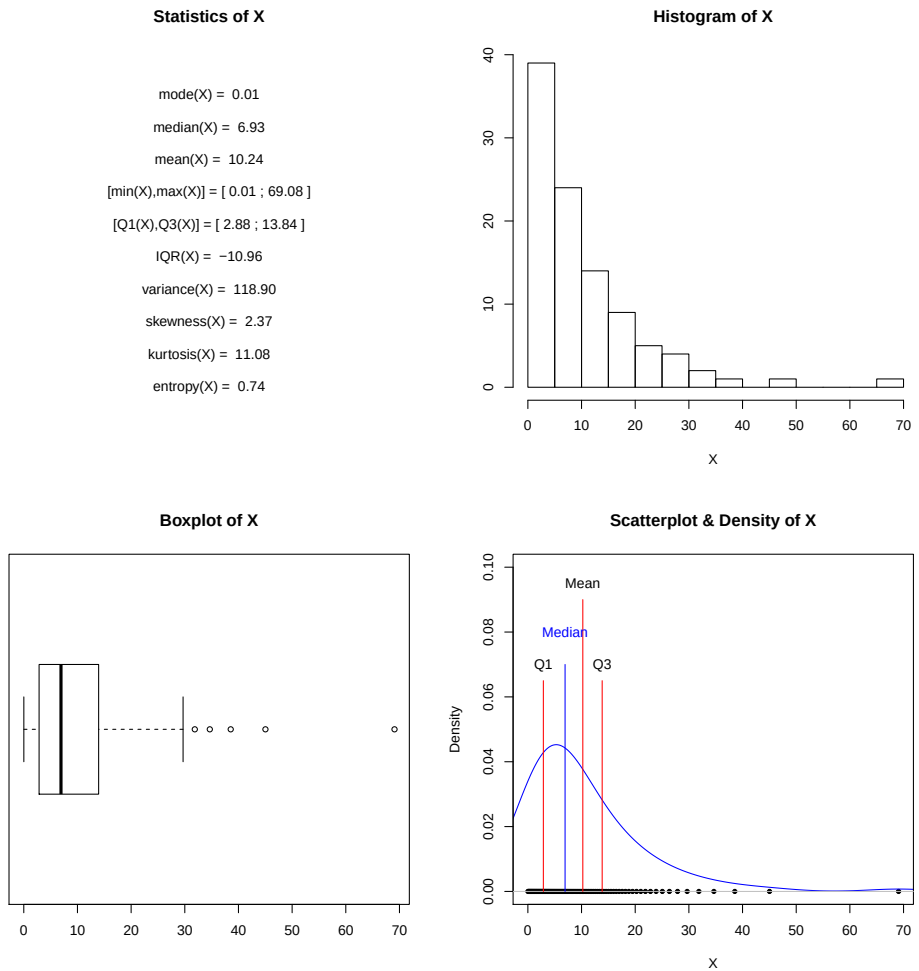


Figure: Statistics of an Exponentially Distributed Variable

Describing Multivariate Data: Correlation

Main Objective

- Describe the relationship between two variables

Covariance

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{N - 1}$$

Pearson's linear correlation coefficient

$$corr(x, y) = \frac{cov(x, y)}{s_x * s_y}$$

- Discussion: Correlation vs. Causality?

Spearman's rank correlation coefficient

- Sort the values and assign ranks
- Compute the correlation coefficient on the ranks, instead of the original values

Describing Multivariate Data: Covariance

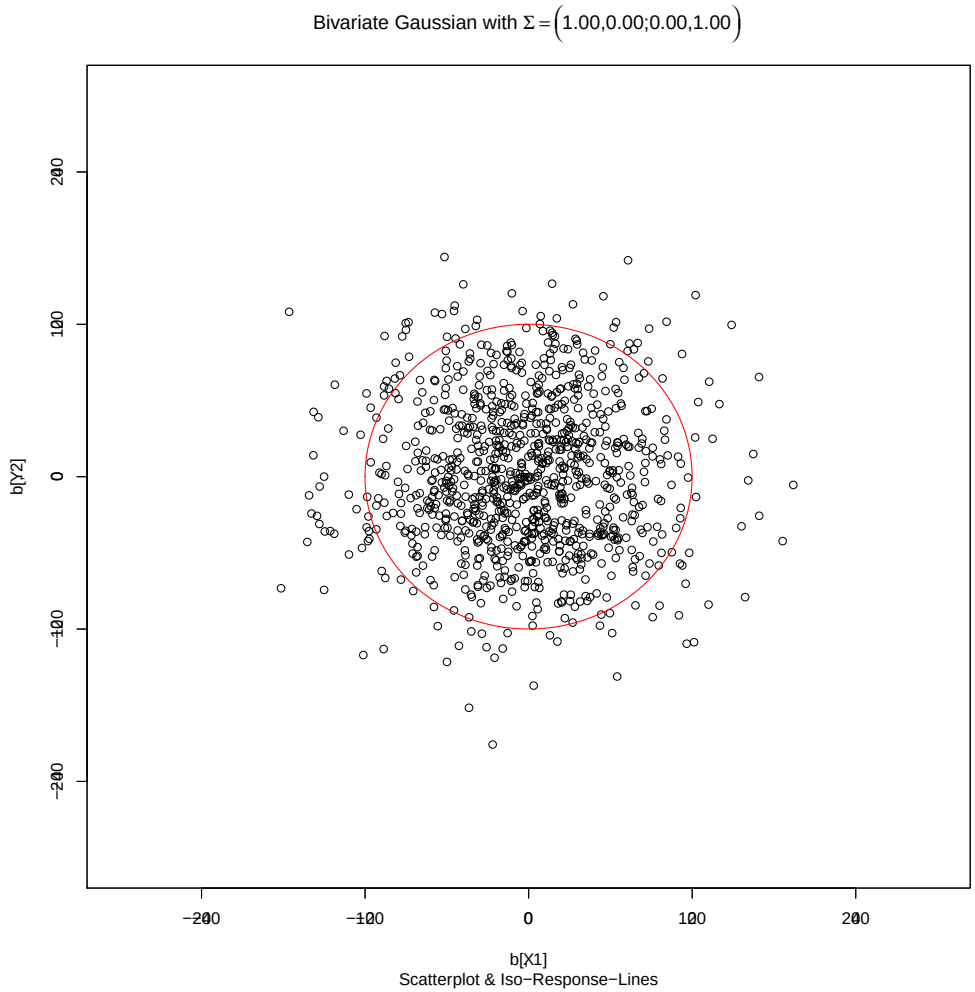


Figure: Scatterplot and Iso-Response-Line of an Bivariate Gaussian with Covariancematrix $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and Meanvector $\mu = (0, 0)$

Describing Multivariate Data: Covariance

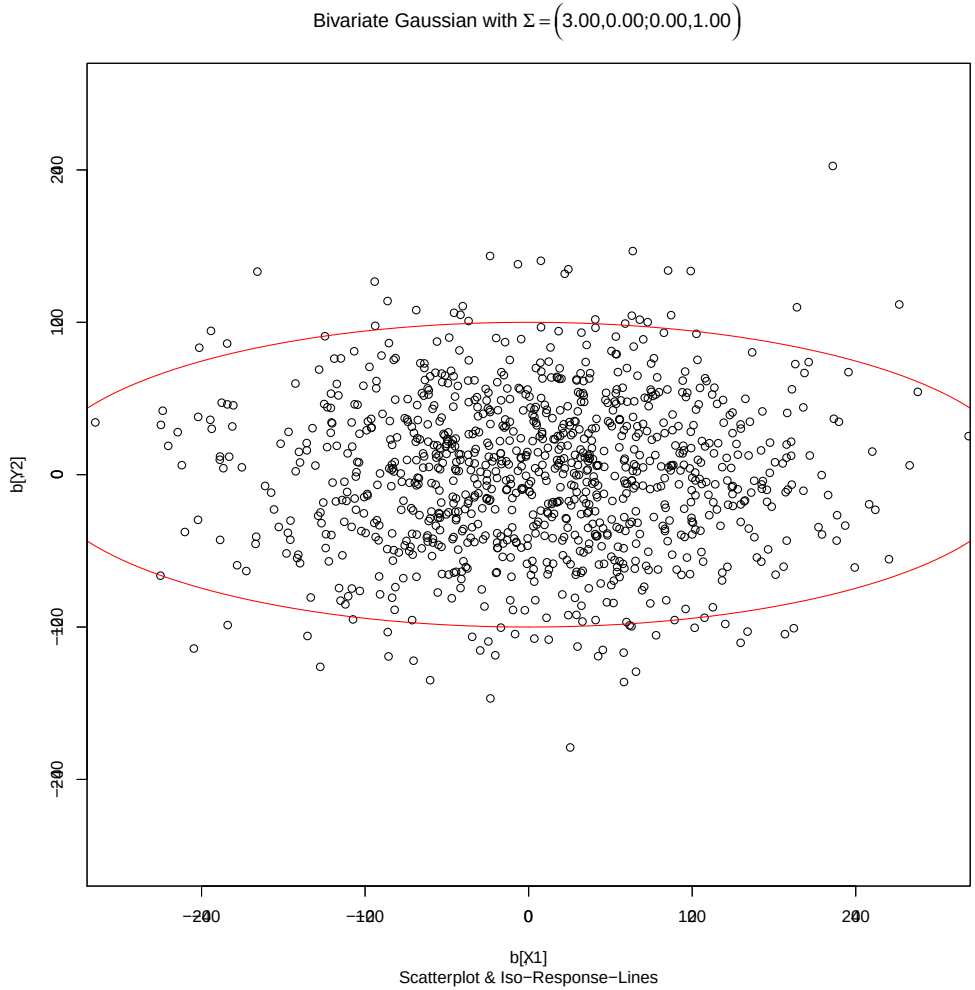


Figure: Scatterplot and Iso-Response-Line of an Bivariate Gaussian with Covariancematrix $\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$ and Meanvector $\mu = (0, 0)$

Describing Multivariate Data: Covariance

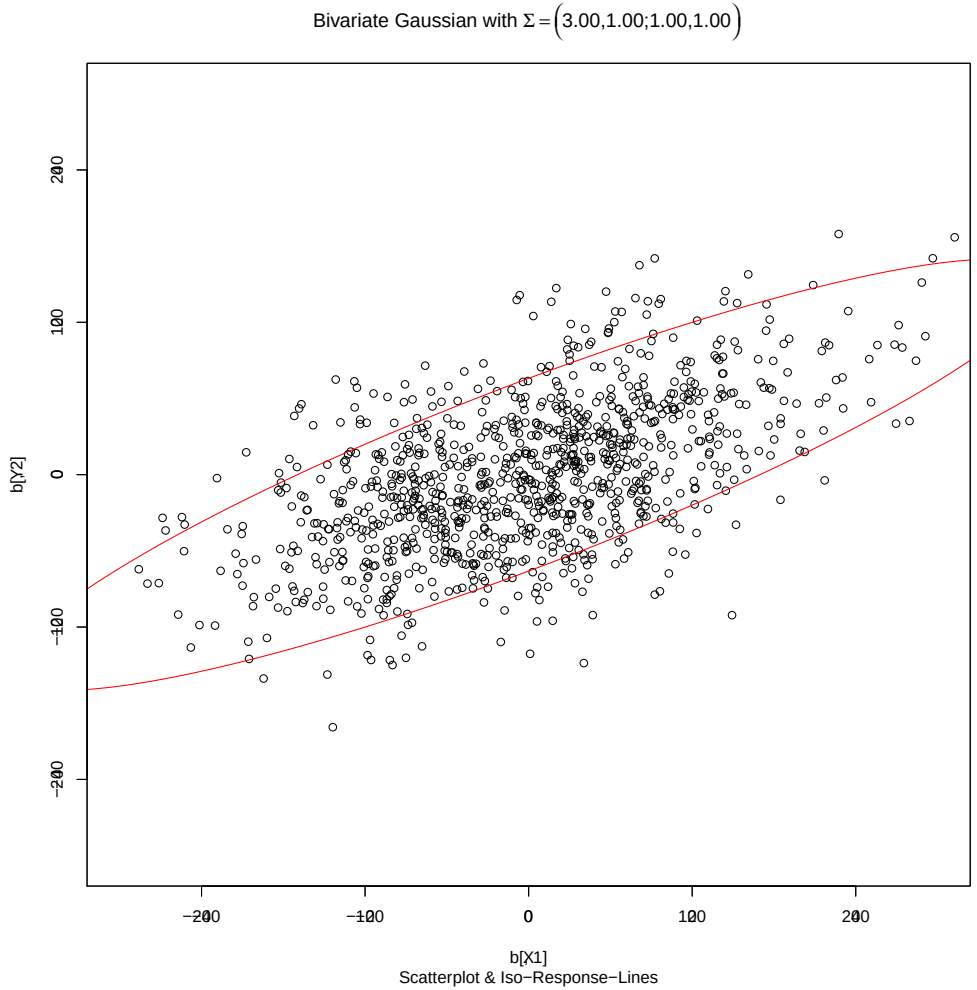
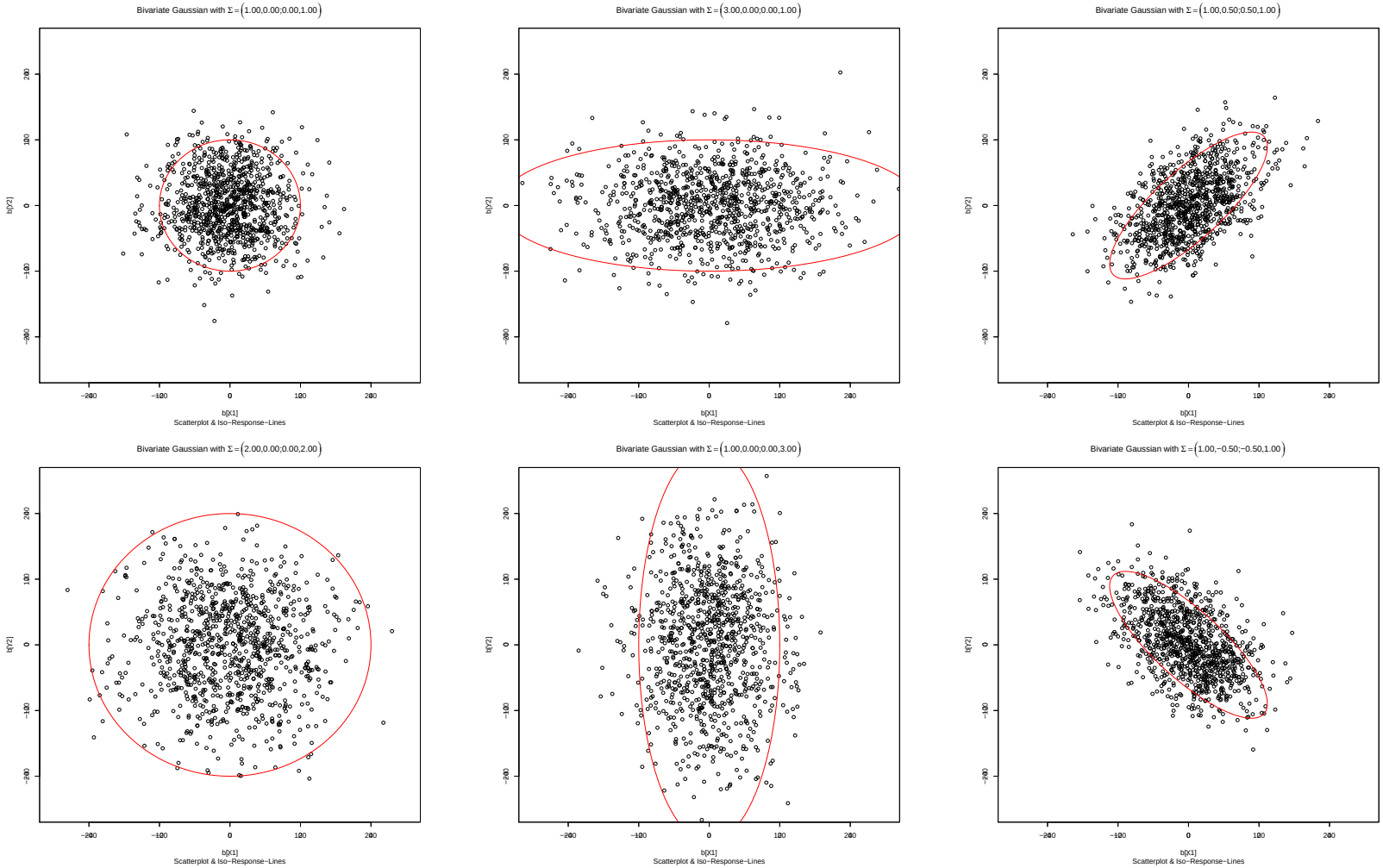


Figure: Scatterplot and Iso-Response-Line of an Bivariate Gaussian with Covariancematrix $\Sigma = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$ and Meanvector $\mu = (0, 0)$

Describing Multivariate Data: Covariance



(a) Identical Variance (b) Varying Variance (c) Varying Covariance

Figure: Scatterplots and Iso-Response-Lines of an Bivariate Gaussians with Meanvector $\mu = (0, 0)$ and varying Covariancematrices Σ.

Describing Multivariate Data: (In)Dependence of Variables

Independence

- Two random variables A and B are *indepedent*, if and only if

$$\mathcal{P}r(A, B) = \mathcal{P}r(A) \cdot \mathcal{P}r(B)$$

  for all values of A and B.
- In this case, the realisations of the two variables are not related
- Knowing the realisation of A does not help in predicting B.

Conditional Independence

- A random variable A is *conditionally independent* of B given C, if for all its values holds that

$$\mathcal{P}r(A, B|C) = \mathcal{P}r(A|C) \cdot \mathcal{P}r(B|C) \qquad \text{or, likewise}$$
$$\mathcal{P}r(A|B, C) = \mathcal{P}r(A|C)$$

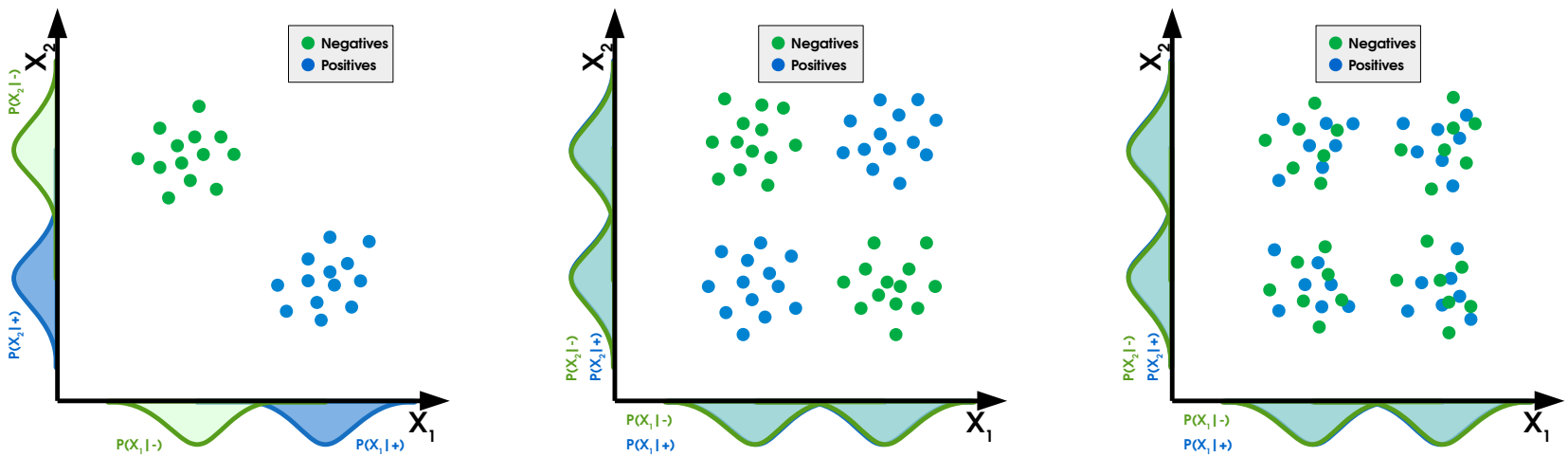
- Conditional independence does *not imply* unconditional independence!
- Example: (done on whiteboard)

Conditional Independence[10]

Two variables X_i and X_j are conditionally independent given a third variable Y, if for all values of X_i, X_j, Y holds

$$\mathcal{P}r(X_i, X_j|Y) = \mathcal{P}r(X_i|Y) \cdot \mathcal{P}r(X_j|Y)$$

$$\mathcal{P}r(X_i|X_j, Y) = \mathcal{P}r(X_i|Y)$$

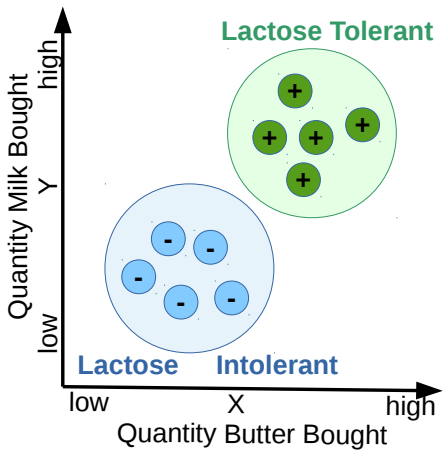


(a) (b) (c)

Figure: In scatterplots (a,c), X_1, X_2 are cond. indep. given Y. In (b), they are not.

[10] See [Hand et al., 2001, Section 4.3.1].

Describing Multivariate Data: (In)Dependence of Variables [11]



Independence Example

- Sales records of products butter X and milk Y
- Third variable lactose tolerant Z (class label)
- Without knowing class Z of customer, X and Y are dependent
- Given a value of Z, they are independent (in this illustration at least . . .)

[11] Example inspired by [Hand et al., 2001].

Illustration

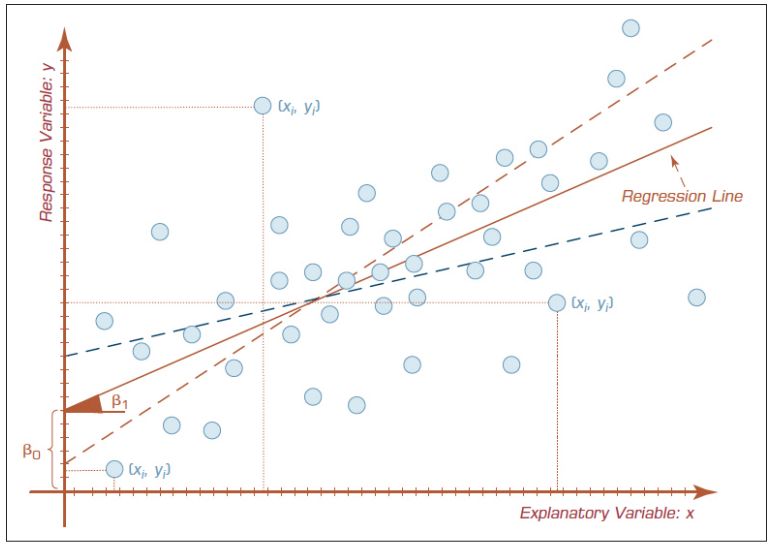


Figure: Linear Regression Line in a Scatterplot (Source: [Sharda et al., 2018, page 113])

- $f : X \to Y$
- X is the explanatory or input variable
- Y is the dependent variable, or output/response

Describing Multivariate Data: Regression (2)

Types of Regression

- simple regression: one explanatory variable and one response
- multiple regression: several explanatory variables
- multivariate regression: several dependent variables

Different Purposes

- descriptive/explanatory: how much influence does X have on Y?
- predictive/forecasting: given a value of X, what is the value of Y?

Note:

- used in descriptive analytics
- **and** in predictive analytics

Describing Multivariate Data: Regression (3)

Simple Linear Regression

- linear (or generalized to polynomes of higher order)

$$y = \beta_0 + \beta_1 \cdot x$$

- β_0 intercept
- β_1 slope
- ϵ error term

Multiple linear regression

- More than one explanatory variables

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_p \cdot x_p$$

- y response
- $x_1, x_2, \cdots, x_p$ the p different explanatory variables
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ the beta-coefficients

Describing Multivariate Data: Regression (4)

Different optimization functions

Different performance measures are frequently used, e.g.,

- ordinary least squares: minimize sum of squared errors
- ...

Assumptions

- linearity
- independence
- normality of errors
- constant variance
- multicollinearity

Descriptive Statistics: Multivariate Examples

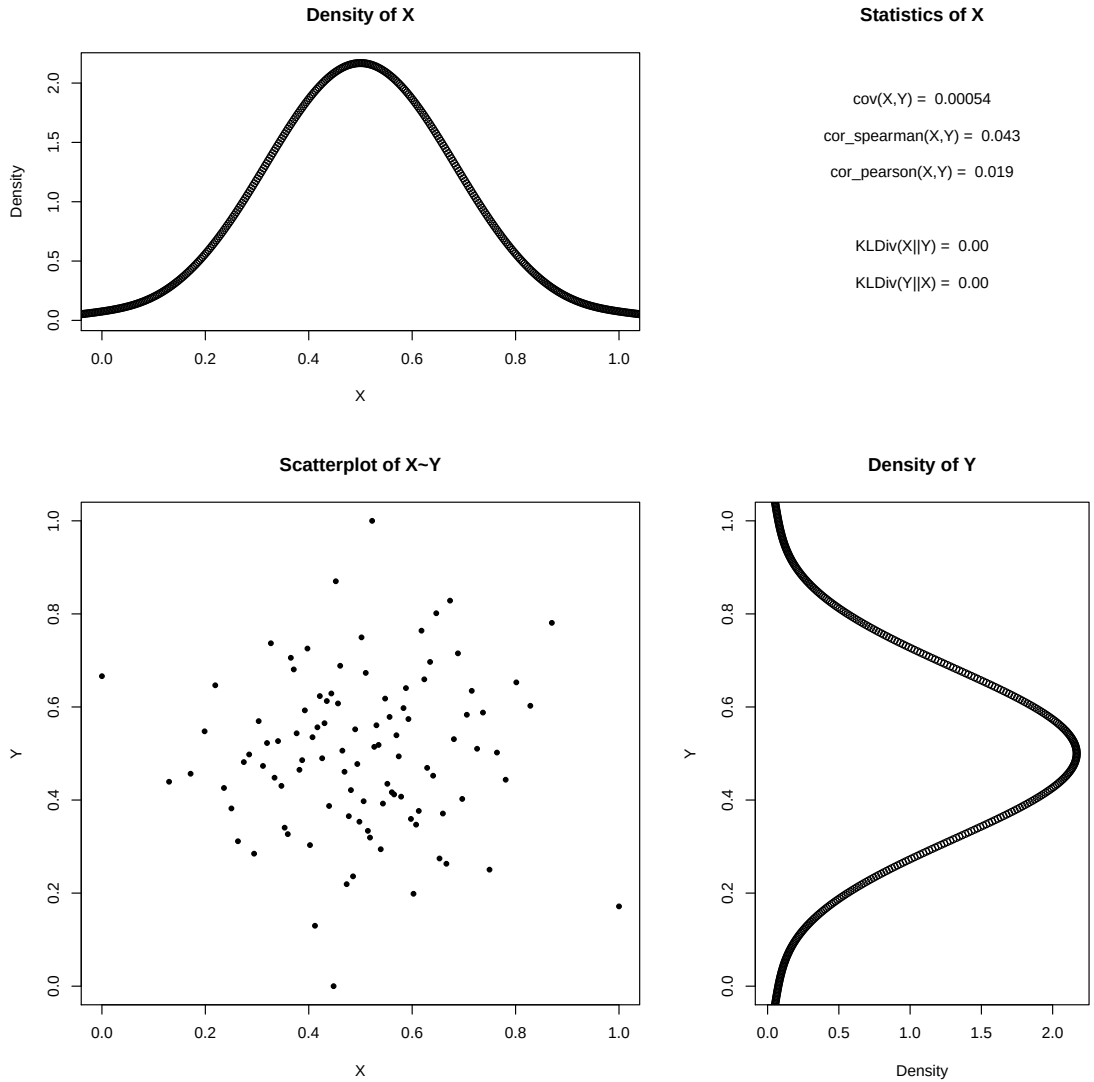


Figure: Statistics of Independent Variables

Descriptive Statistics: Multivariate Examples

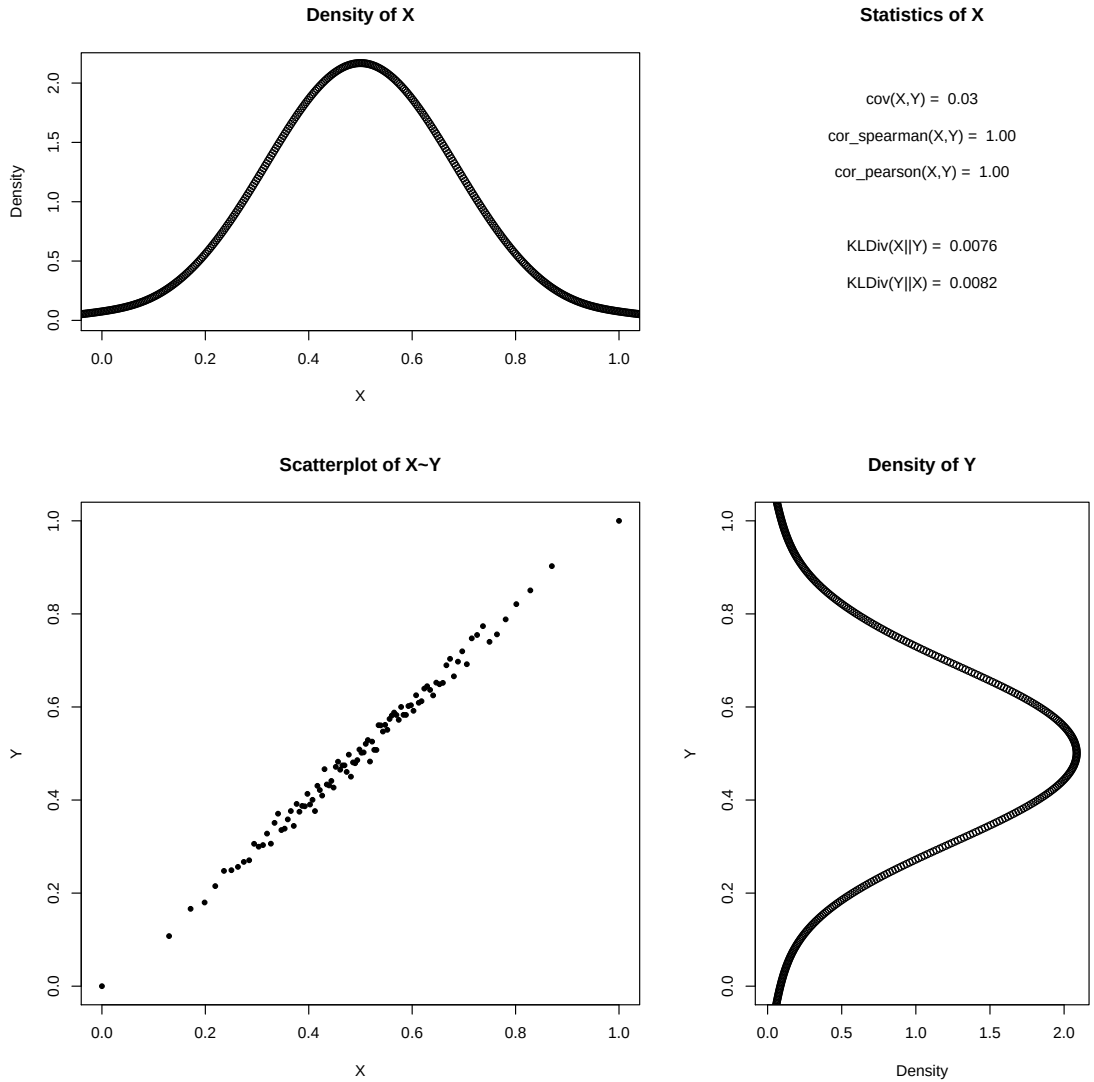


Figure: Statistics of Linearly Dependent Variables

Descriptive Statistics: Multivariate Examples

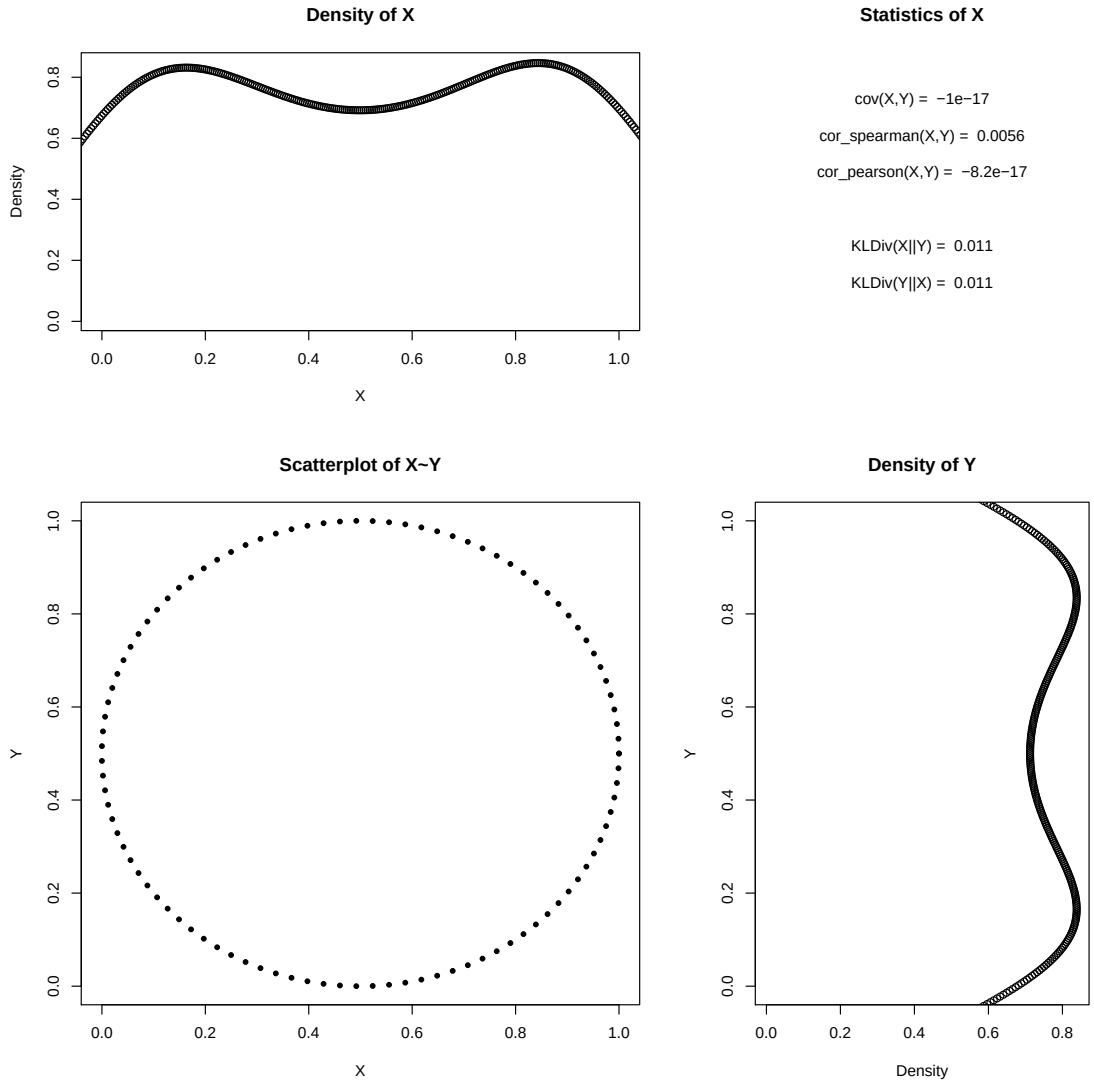


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Multivariate Examples

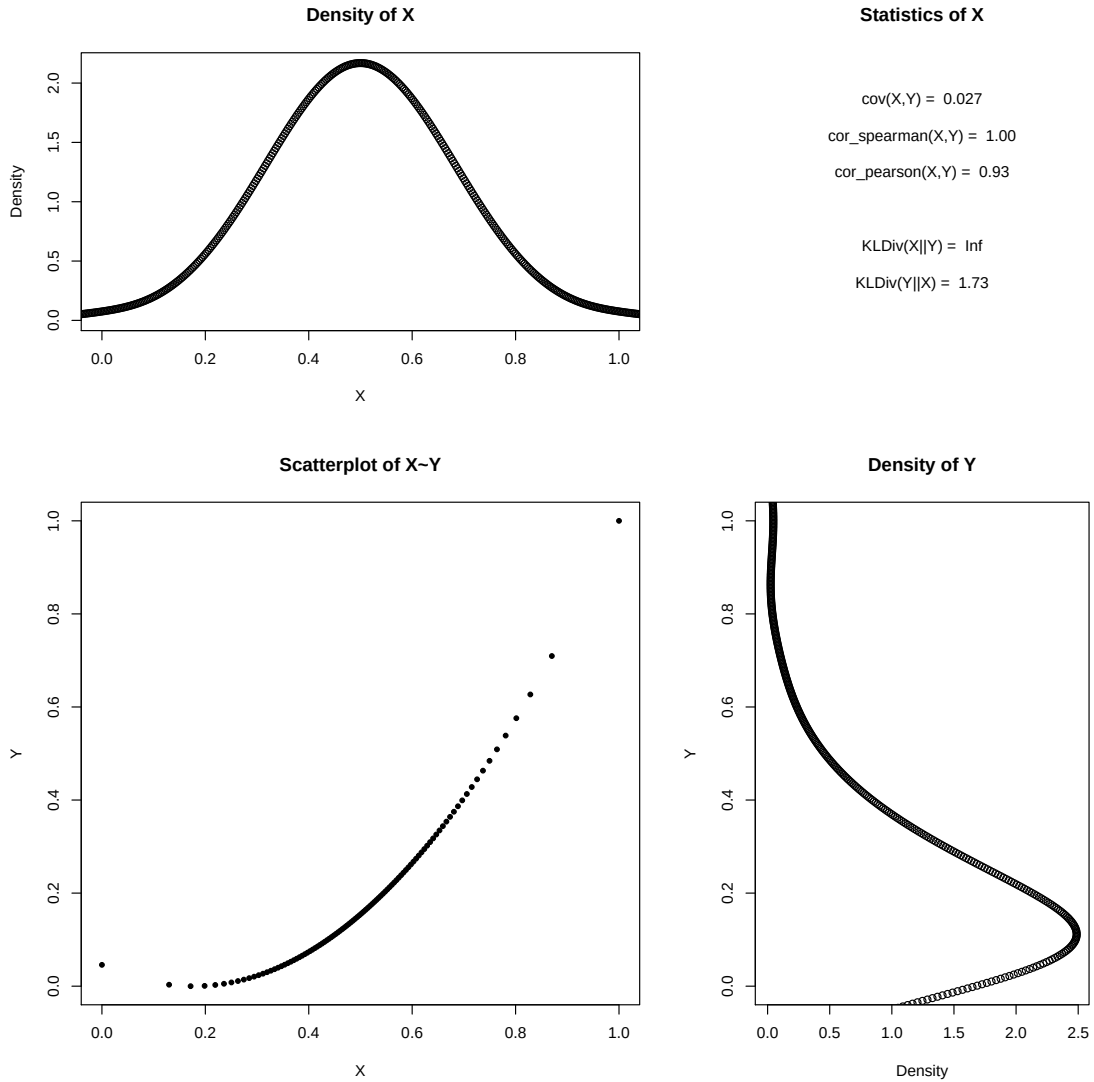


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Multivariate Examples

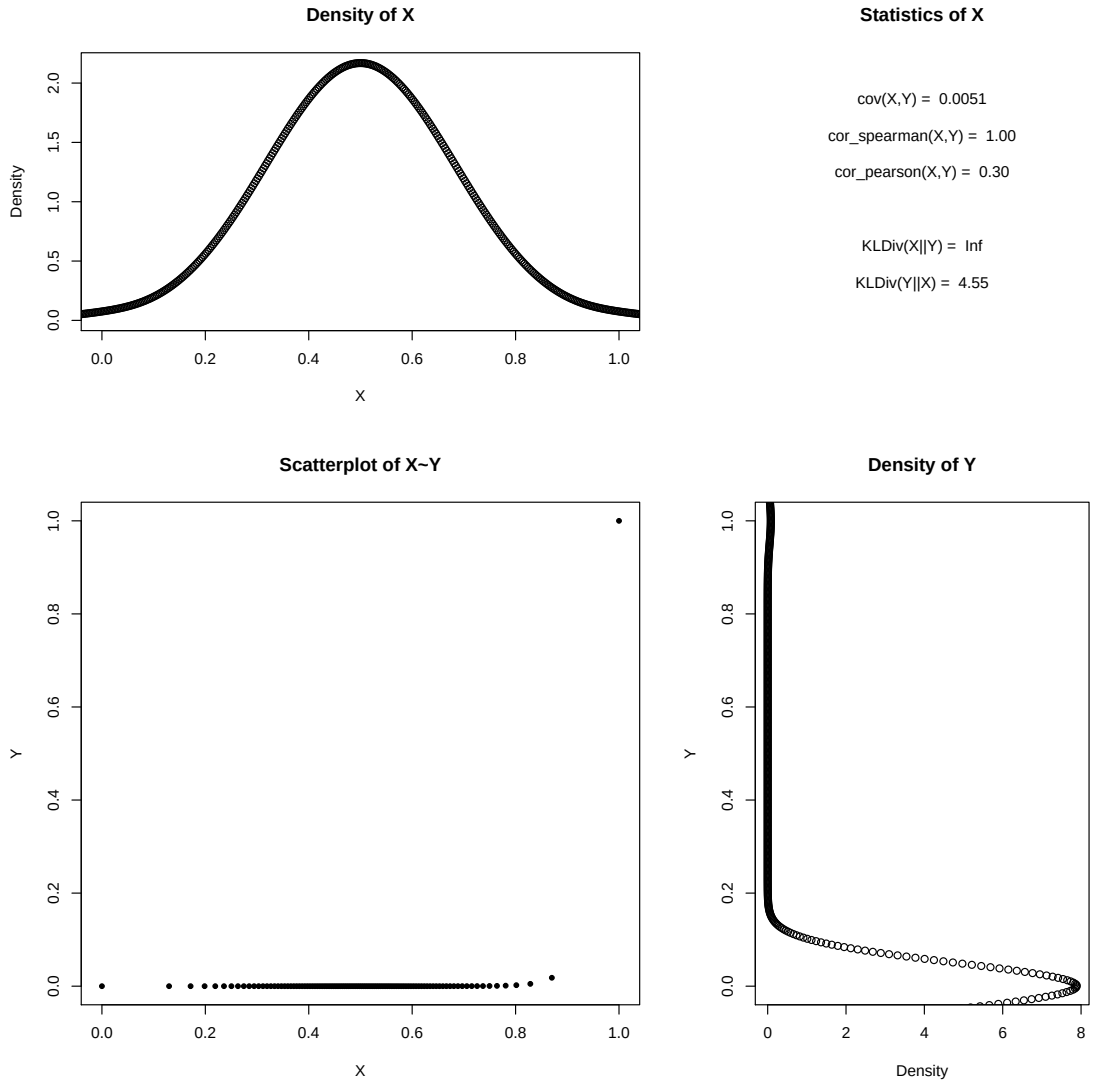


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Multivariate Examples

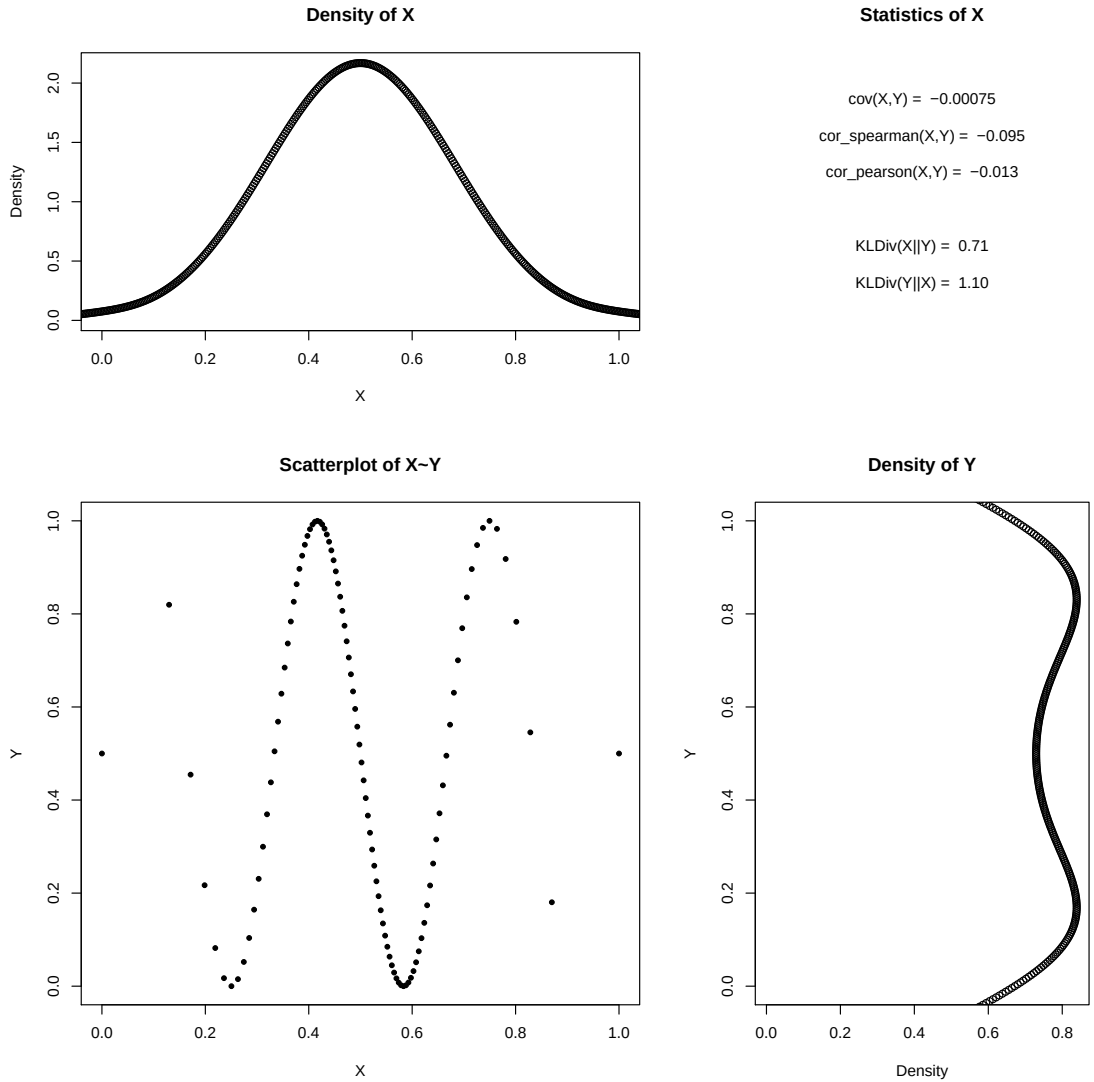


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Multivariate Examples

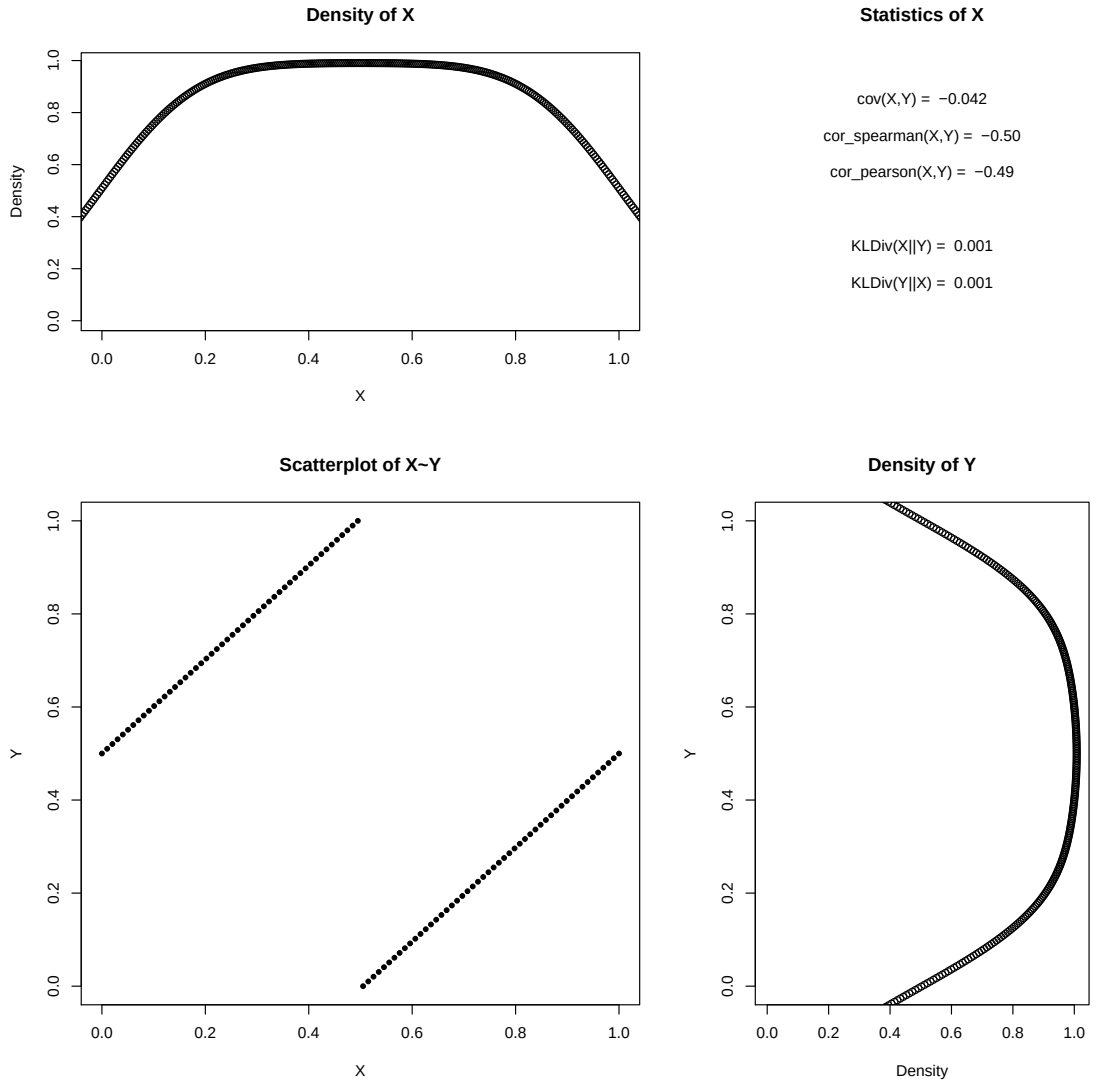


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Multivariate Examples

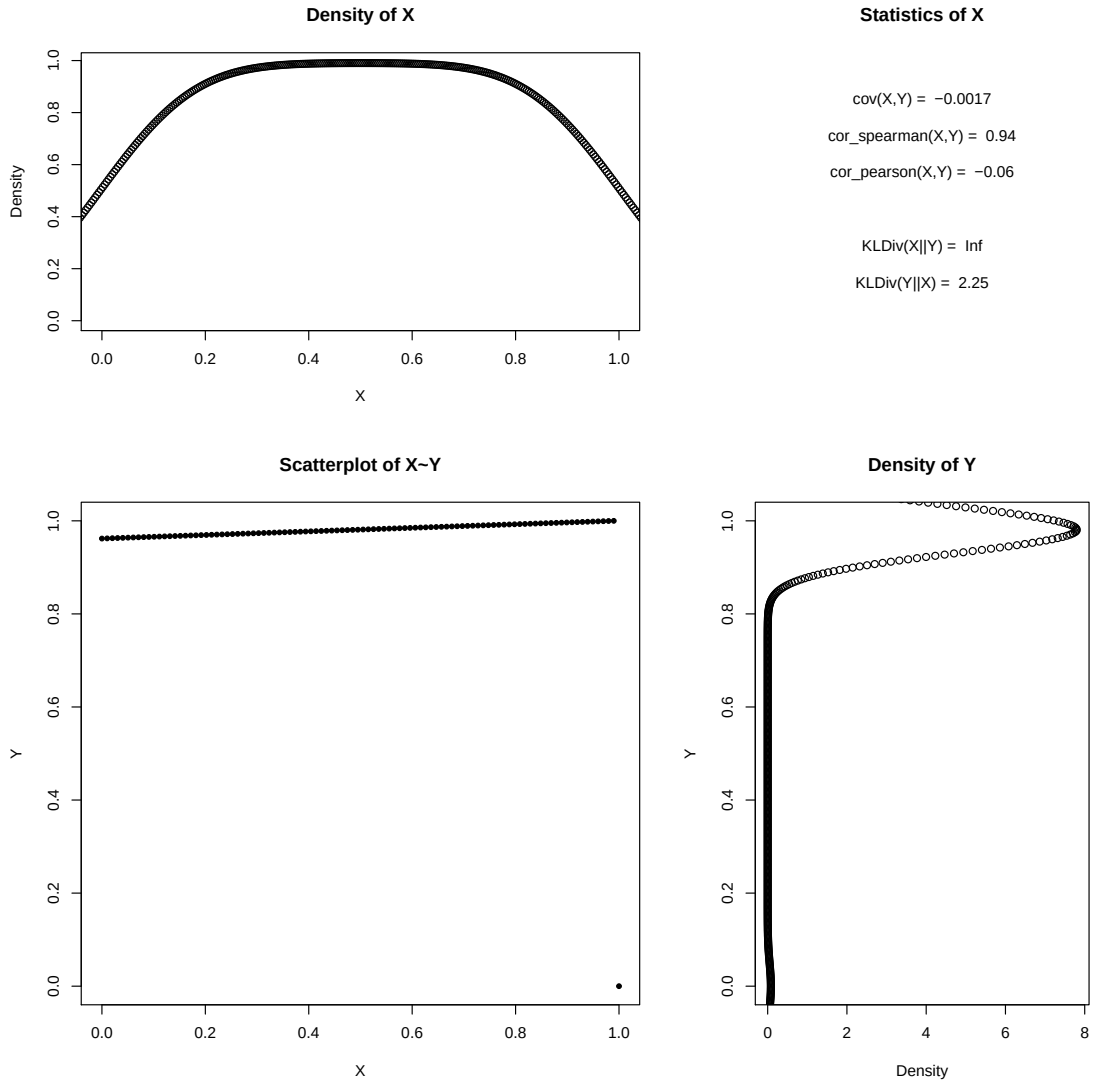


Figure: Statistics of Dependent but **not** Linearly Correlated Variables

Descriptive Statistics: Epilogue

Point Estimates

- The above are so-called *point* estimates
- They aim to summarize the data into a single value
- They do not tell how *well* this estimate describes the data

- Extensions are *(confidence) interval* estimations and *statistical testing*
- Consider, e.g., the sample size (number of observations)

- Not in the scope of this lecture, but rather subject of *Inferential Statistics* (recommended when focusing on predictive analytics)

Statistics: Distribution of Random Variables

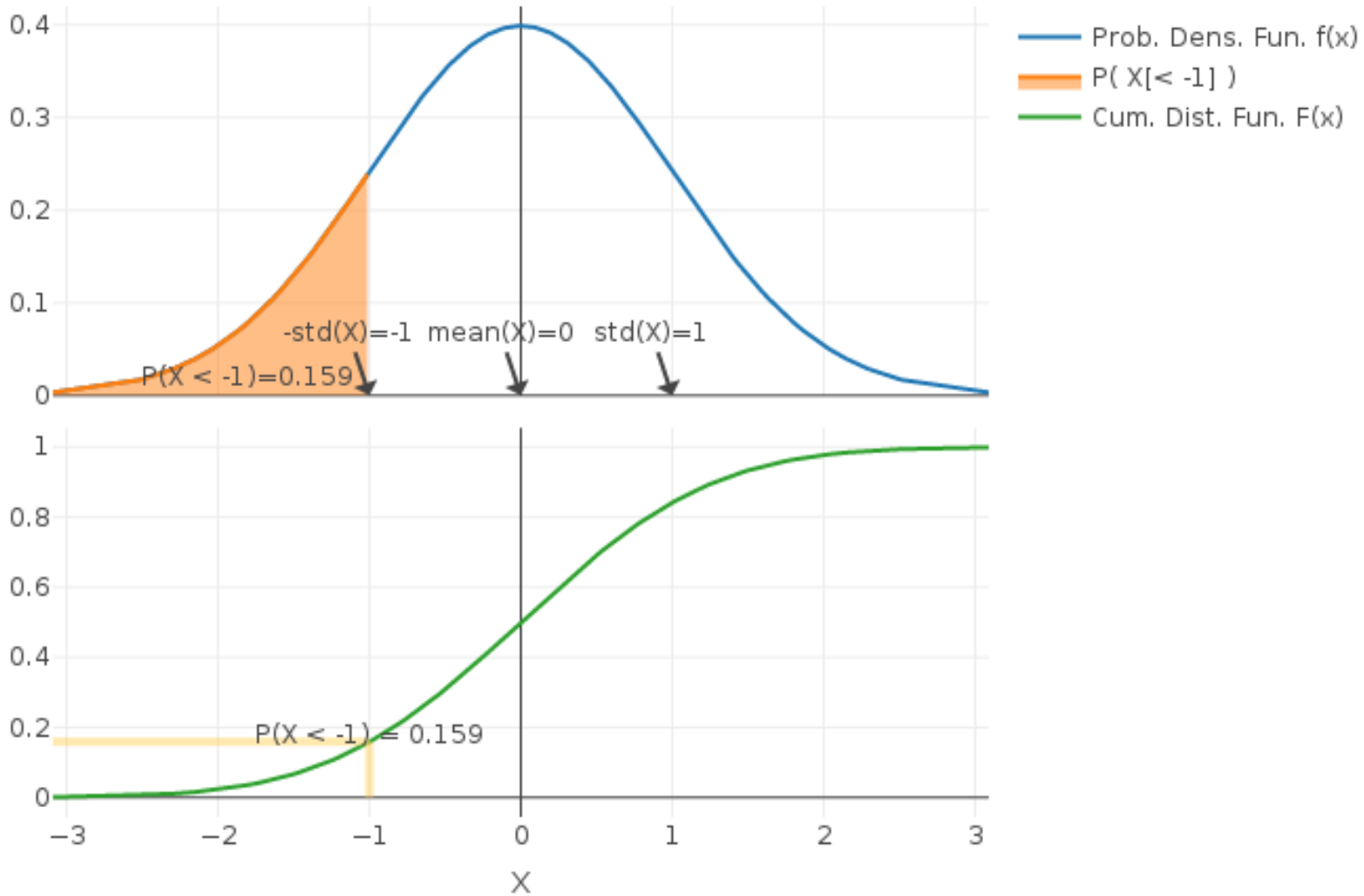


Figure: Probability Density Function (PDF) $f(x)$, Cumulative Distribution Function (CDF) $F(x)$ of a Gaussian distributed variable X (standard normal with *mean* $= 0$, *std* $= 1$)

Statistical Testing and Confidence Interval: Example

- Given a sample S[12], and assuming data is normal distributed, aim is to assess whether the population mean is greater than 40.
- Interested in a 95%-Confidence interval
- Alternative hypothesis $H_1 : \mu > 40$ (corresponding null hyp.: $H_0 : \mu \leq 40$)
- Performing a one-sample t-test with significance level $\alpha = 5\%$ (α probability of wrongly rejecting the null hypothesis)
- Null hypothesis not rejected (p-value $0.5 \not< 0.05$, the predefined sig. level α), thus our alternative hypothesis is *not* confirmed
- 95%-Confidence interval is $[39.281; \infty[$

```
> S=c(37,38 ,39 ,39,39 ,40 ,40 ,40 ,41 ,41,41 ,41 ,41,43)
> t.test(x=S,mu=40)
        One Sample t-test
data:  S
t = 0, df = 13, p-value = 1
alternative hypothesis: true mean is not equal to 40
95 percent confidence interval:
39.12289 40.87711
sample estimates:
mean of x
      40
```

[12] For simplicity, the sample here is very small (in reality, a bigger sample size would be needed).

Outline and Summary[2]

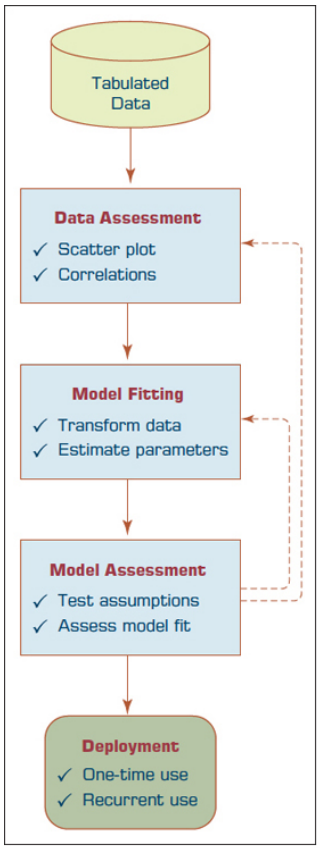
- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

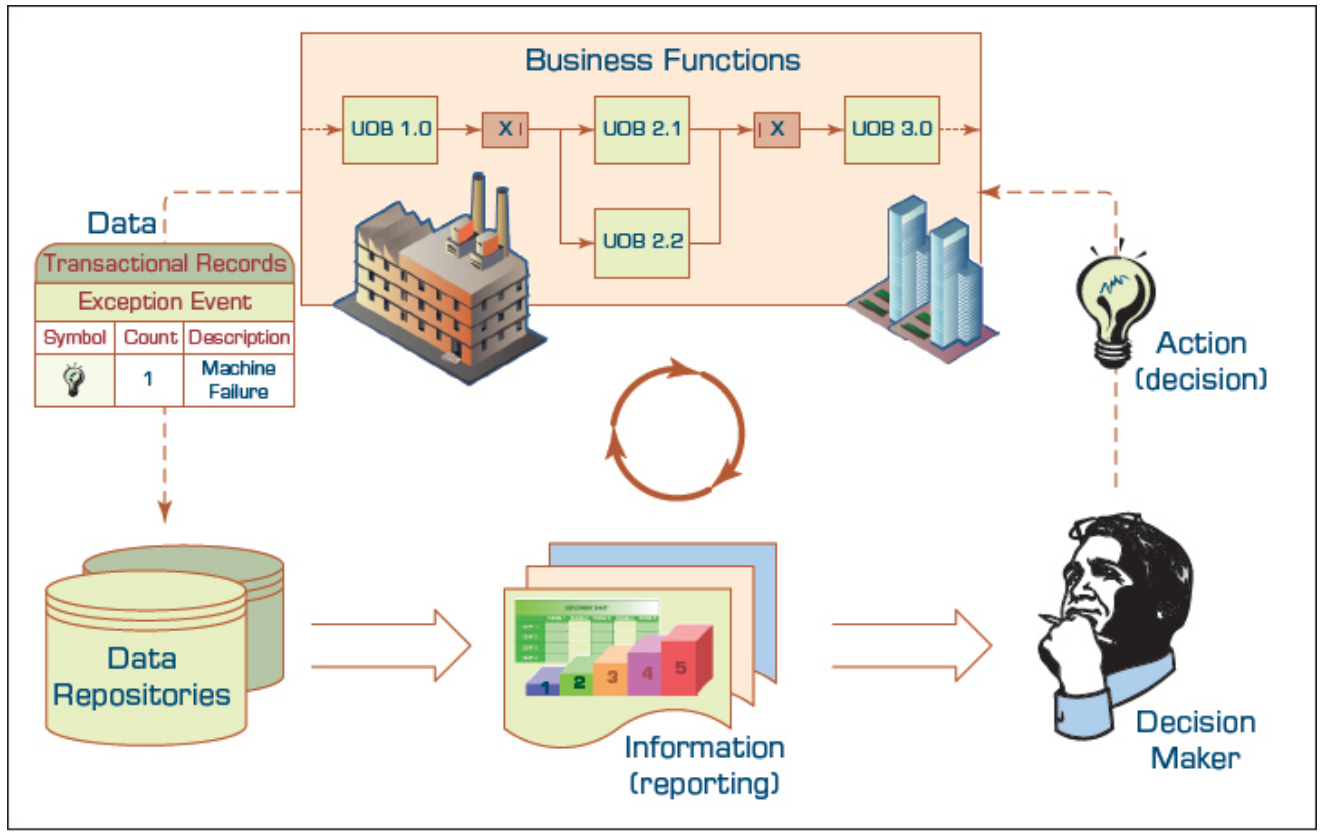
▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

From Data to Reporting: Context within Business Reporting





Copyright © 2018, 2014, 2011, 2008 by Pearson Education, Inc.

Figure: The Context of Business Reporting (Source: [Sharda et al., 2018, page 115,125])

Business Report

- A Business Report conveys business related information
- Aim: supporting & improving managerial decisions
- Integrates data from inside & outside (e.g., ETL)
- Format: text, tables, visualization
- Distribution: print, digital (interactive!)

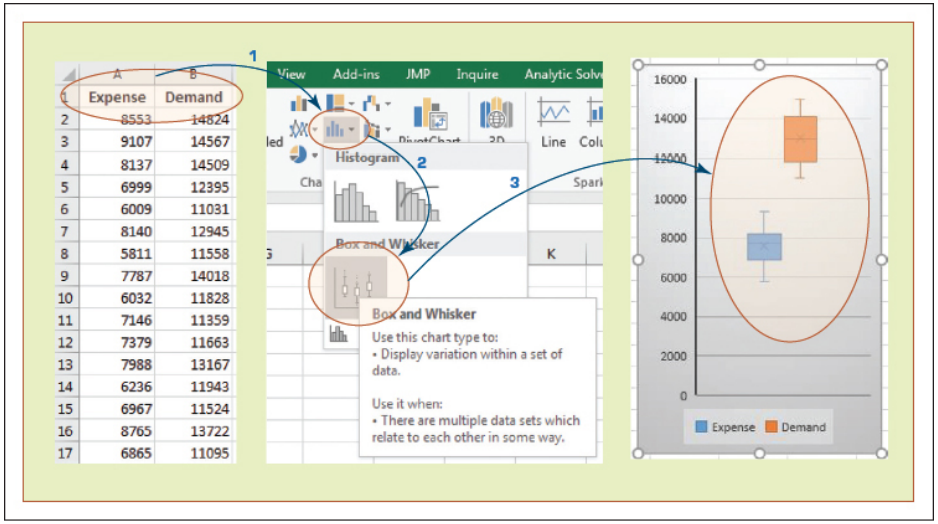


Figure: Business Report
(Source: [Sharda et al., 2018, page 110])

Business Reporting: Types of Reports[13]

Metric Management Reports

- Help manage business performance through metrics (SLAs for externals; KPIs for internals)
- Can be used as part of Six Sigma and/or TQM

Dashboard–Type Reports

- Graphical presentation of several performance indicators in a single page using dials/gauges

Balanced Scorecard–Type Reports

- Include financial, customer, business process, and learning & growth indicators

[13] Source: [Sharda et al., 2018].

From Data to Business Reports/Dashboards

What do we need to know to get from data to a "good" dashboard?

- Understanding what defines a business reporting instrument like a dashboard
- Selecting/Integrating, (pre)processing and analysing the data: Data Preprocessing
- Deriving the right information: Descriptive Statistics
- Selecting the right (visualization) tools: Graph/Table/...
- Visual design of the report/dashboard: Storytelling

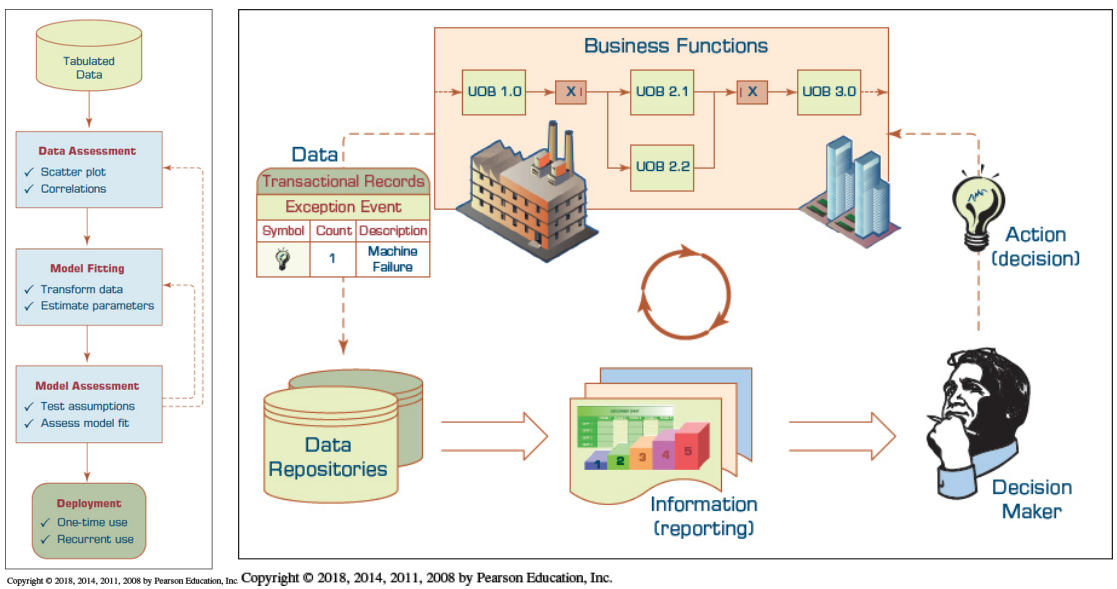


Figure: The Context of Business Reporting (Source: [Sharda et al., 2018, page 115,125])

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Data Visualization[14]

Definition

The use of visual representations to explore, make sense of, and communicate data.

- Data visualization vs. Information visualization
- Information: aggregation, summarization, and contextualization of data
- Related to information graphics, scientific visualization, and statistical graphics
- Often includes charts, graphs, illustrations,. . .

[14] Definition and slide based on [Sharda et al., 2018].

Data Visualization: Historical Examples

- Visualization of information is very old (beyond 2nd century A.D., according to [Sharda et al., 2018, p.127])
- "Modern" visualization started approx. 200 years ago
- Probably the first "modern" chart, created by William Playfair in 1801
- He supposedly created the first line and pie charts

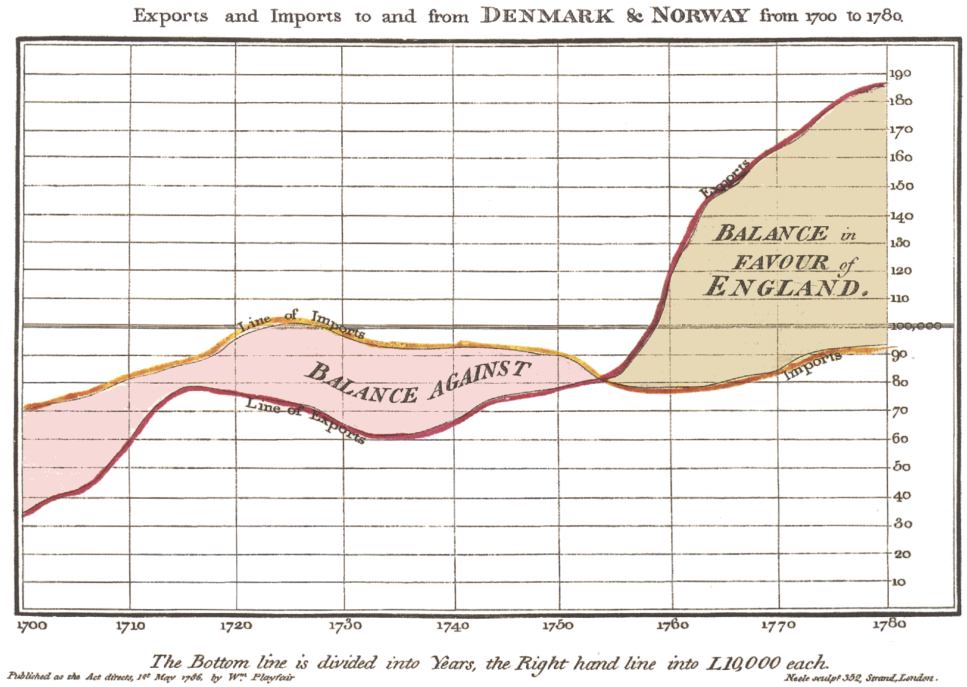


Figure: Time Series Line Chart created by William Playfair in 1801 (Source: `commons.wikimedia.org`)

Data Visualization: Historical Examples (2)

- One of the most popular early visualizations (Charles Joseph Minard, 1869)
- Visualization itself has recently been recognized and developed as discipline
- There is a move towards visualization
- Current Developments: Interactiveness and Storytelling with Data Visualization: [Segel and Heer, 2010]

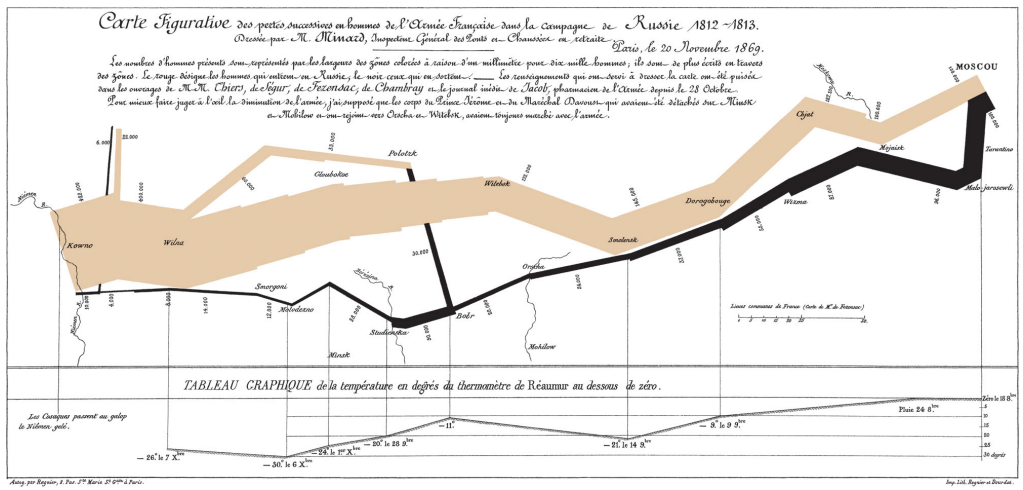


Figure: Decimation of Napoleaon's Army during the 1812 Russian Campaign by Charles Joseph Minard, 1869. Source: `commons.wikimedia.org`

From Data Visualization to Visual Analytics

Origin

- Information visualization and predictive analytics
- Strong tendency towards visual analytics in BI!

Information Visualization

- descriptive, retrospective
- what is happening / has happened?

Predictive Analytics

- predictive, forward-looking
- what will happen and why?

From Data to Business Reports/Dashboards

What do we need to know to get from data to a "good" dashboard?

- Understanding what defines a business reporting instrument like a dashboard
- Selecting/Integrating, (pre)processing and analysing the data: Data Preprocessing
- Deriving the right information: Descriptive Statistics
- Selecting the right (visualization) tools: Graph/Table/...
- Visual design of the report/dashboard: Storytelling

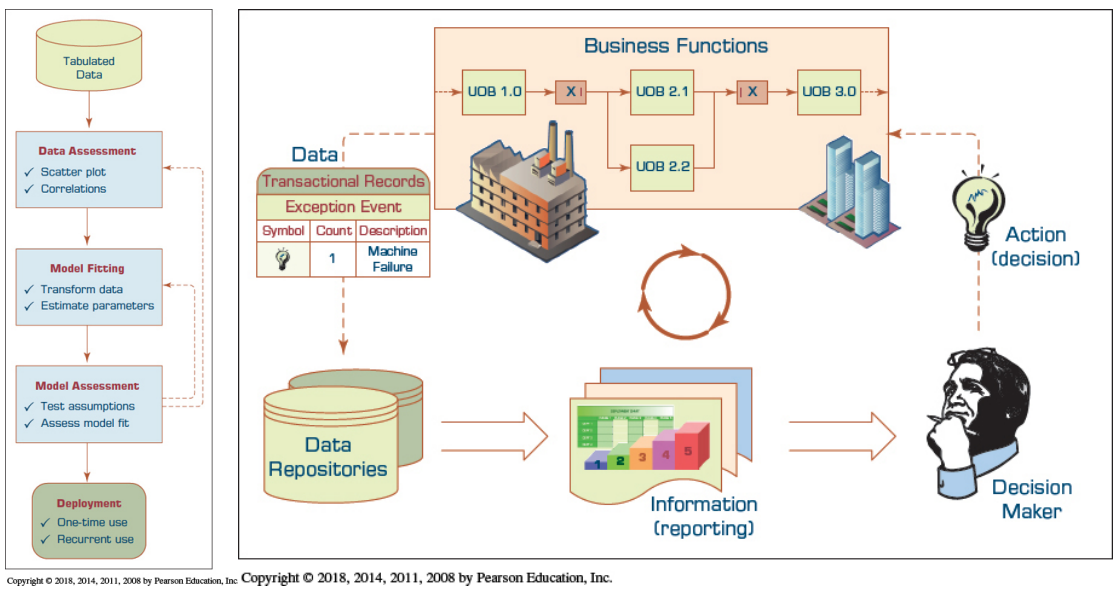


Figure: The Context of Business Reporting (Source: [Sharda et al., 2018, page 115,125])

Performance Dashboards

Performance Dashboard

- Provides visual display of important information
- Consolidates and arranges information on a single screen
- Common in BI platforms and BPM software suites

Aims

- Allow fast capturing of the information
- Allow drill down and exploration by user

Source: dundas.com.

Figure: Exemplary Information Dashboard (Source: [Sharda et al., 2018, page 143])

Performance Dashboards

Design Challenges

- fit all the information on one screen
- clearly and without distraction
- easily to "grasp" / "digest"
- logically consistent and intuitive, inviting/guiding user to "explore"

Compare with Visual Storytelling (next slides)

Layers of Information[15]

- Monitoring: Graphical, abstracted data that allows monitoring
- Analysis: Summarized dimensional data that allows analysing
- Management: Detailed operational data that allows decision making

[15] See [Eckerson, 2010].

Performance Dashboards (2)

Desirable Dashboard Qualities

- highlight data and exceptions that require action
- intuitive use, transparent functioning
  (do not assume user are reading the docu . . .)
- combine and integrate data from variety of systems
  into a single, summarized, unified view
- allow drill-down or drill-htrough to underlying data sources or reports (analysis!)
- dynamic, real-world view with timely data
- easy implementation, deployment, maintenance

Performance Dashboards (3)

Best Practices[16]

- Benchmark KPIs with industry standards
  i.e., which metrics are used by others / leading players?
- Wrap the metrics with contextual metadata
  e.g., where did the data come from, % of missing value, time horizon, ...
- Validate the design by a usability specialist
  i.e., dashboard needs to be (perceived as) user-friendly, otherwise it is not used
- Prioritize and rank alerts and exceptions
  i.e., (important) information should quickly find the user
- Enrich dashboard with business-user comments
  e.g., present other user's comments on the context
- Present information in three different levels
  i.e., the visual dashboard level, static report level, self-service cube level
- Pick the right visual constructs
  i.e., formulate design principles, document them, select instruments (tables, line charts, scatterplots, ...) accordingly
- Provide for guided analytics
  i.e. ensure usability by users of different expertise (guide, enable learning)

[16] Source: [Sharda et al., 2018].

From Data to Business Reports/Dashboards

What do we need to know to get from data to a "good" dashboard?

- Understanding what defines a business reporting instrument like a dashboard
- Selecting/Integrating, (pre)processing and analysing the data: Data Preprocessing
- Deriving the right information: Descriptive Statistics
- Selecting the right (visualization) tools: Graph/Table/...
- Visual design of the report/dashboard: Storytelling

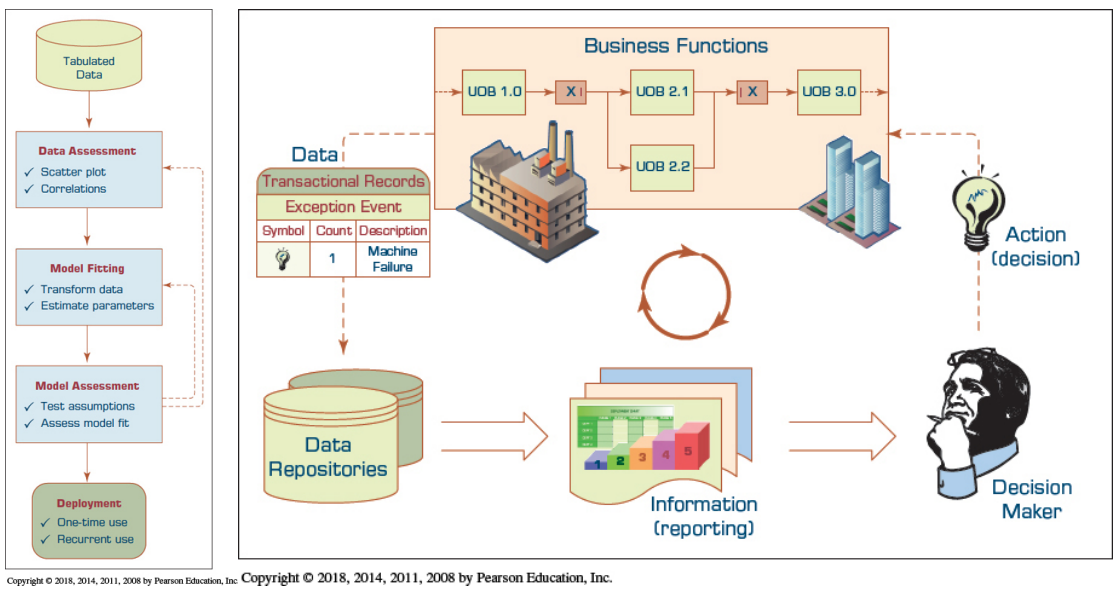


Figure: The Context of Business Reporting (Source: [Sharda et al., 2018, page 115,125])

Data Visualization: Selecting the Right Graph

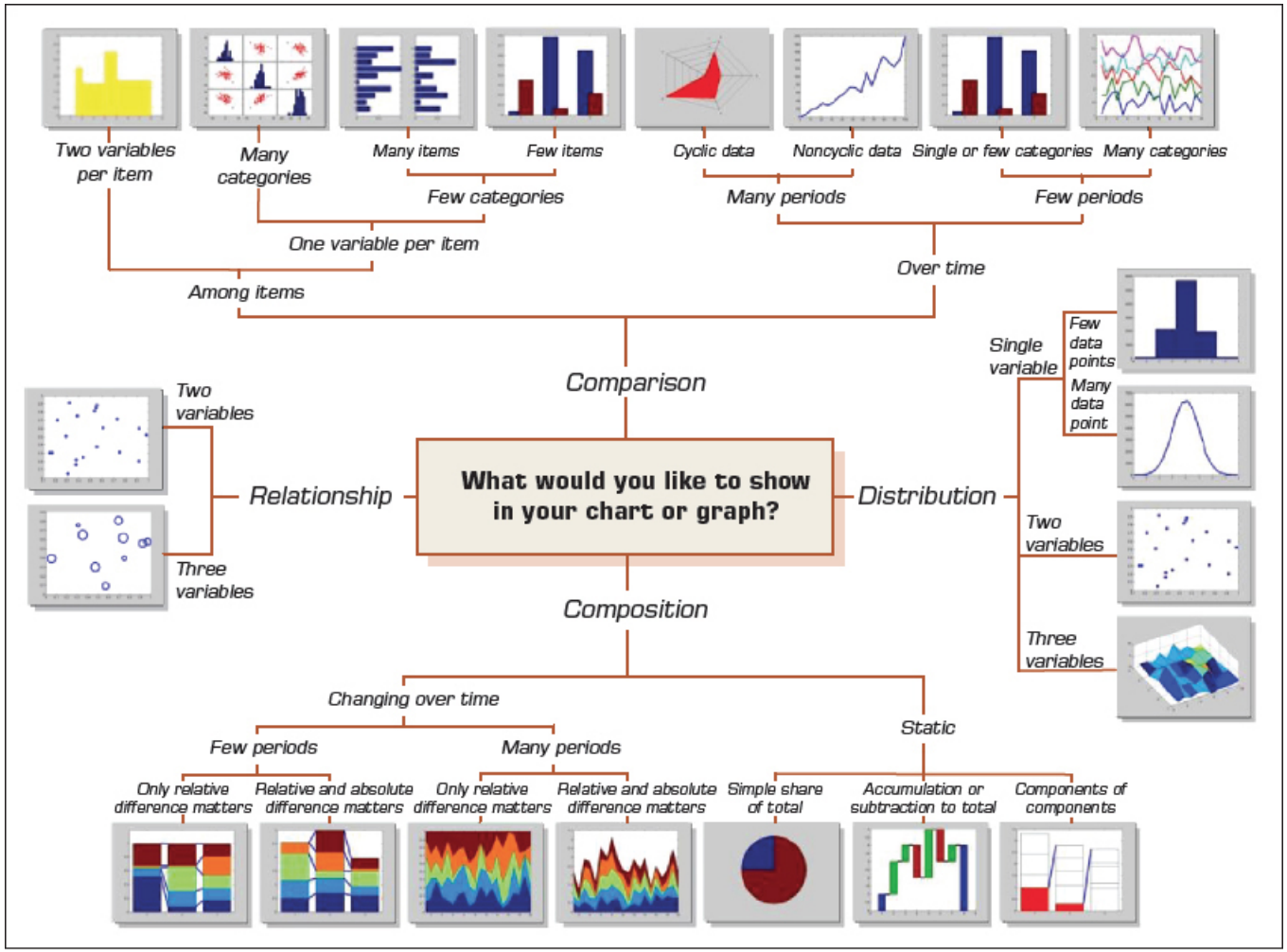


Figure: Taxonomy of Graphs (Source: [Sharda et al., 2018, page 135])

Data Visualization: Selecting the Right Graph (2)

Chart Galleries

- Many websites provide galleries of chart(types)
- R Shiny Gallery: `http://shiny.rstudio.com/gallery/`
- GoogleViz-based Gallery: `https://developers.google.com/chart/interactive/docs/gallery` (usable with R Shiny)
- [Sharda et al., 2018, p.132–135] have a short description of the most common chart types
- Of course, many more ...

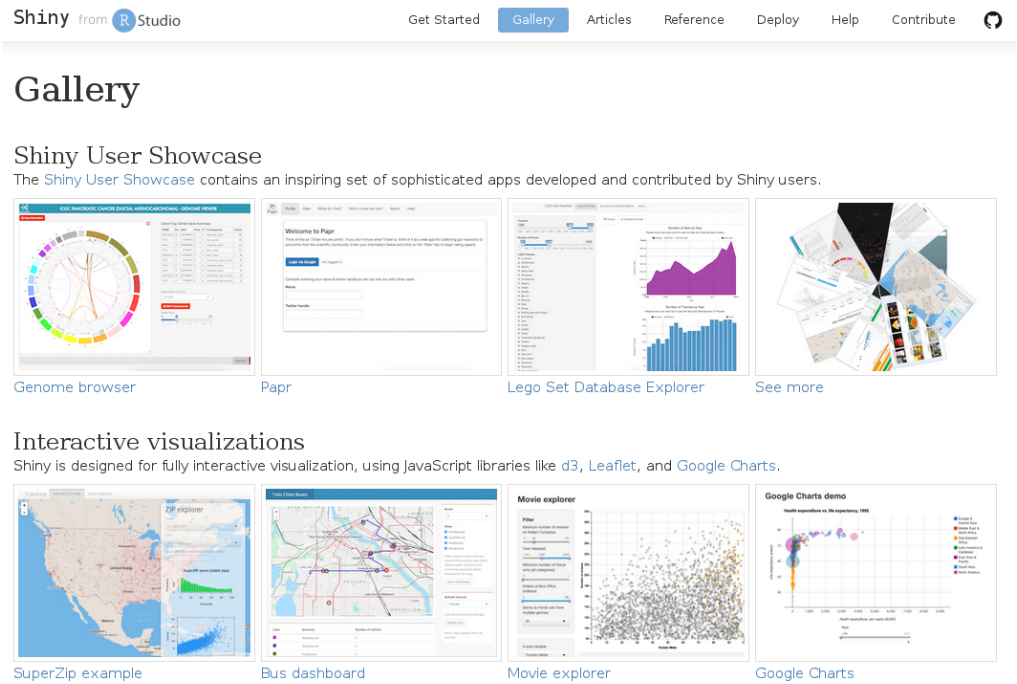


Figure: R Shiny Gallery, 2018-05-02

Data Visualization: Comparison Analysis

Comparison Analysis

- **Bar charts** are well-suited for comparison
- E.g., sorting items (bars) by ranks

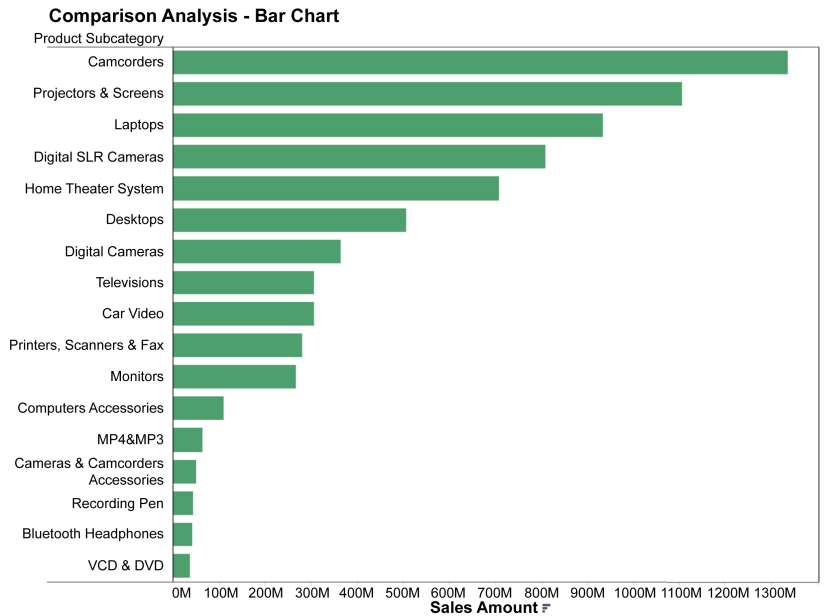


Figure: Example of a Comparison Analysis with a Bar Chart(Source: [Sherman, 2015, page 352])

Time Series/Trend Analysis

- **Point graphs** with time on horizontal axis (if interpolation between time points is *not* meaningful)
- **Line graphs** with time on horizontal axis (if interpolation between time points is meaningful)
- Line graph allows and suggests exploring trends

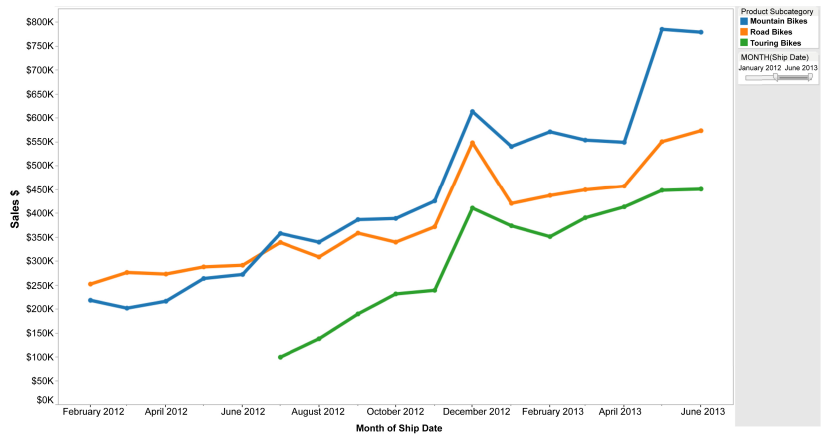


Figure: Example of a Time Analysis with a Line Graphs (Source: [Sherman, 2015, page 353])

Contribution Analysis

- Aim is to represent shares (e.g., percentages)
- **Pie chart** is one possibility only suited for few segments (slices)
  Avoid 3D (biased perception!), better 2D!
- **Heat map** for many segments
- Line graph allows and suggests exploring trends

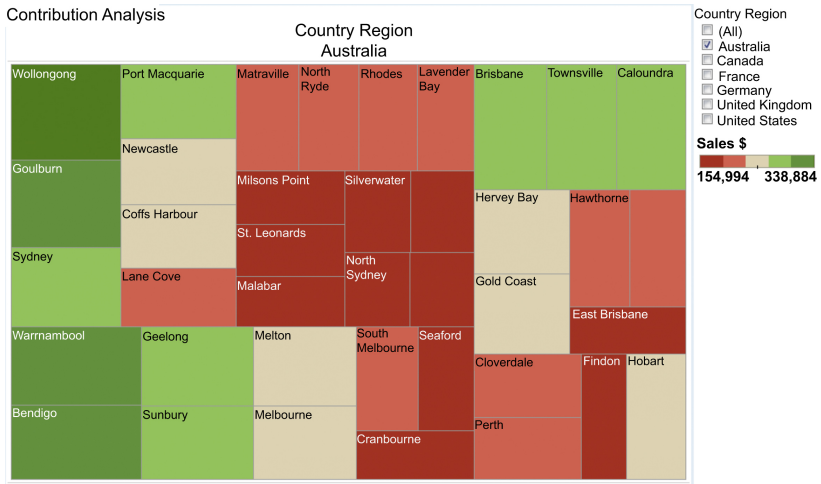


Figure: Example of a Contribution Analysis with a Heat Map (Source: [Sherman, 2015, page 354])

Data Visualization: Correlation Analysis

Correlation Analysis

- Aim is to represent relationship (e.g., correlation)
- **Map** is recommended
- Possible to plot in indicators for correlation or regression analysis

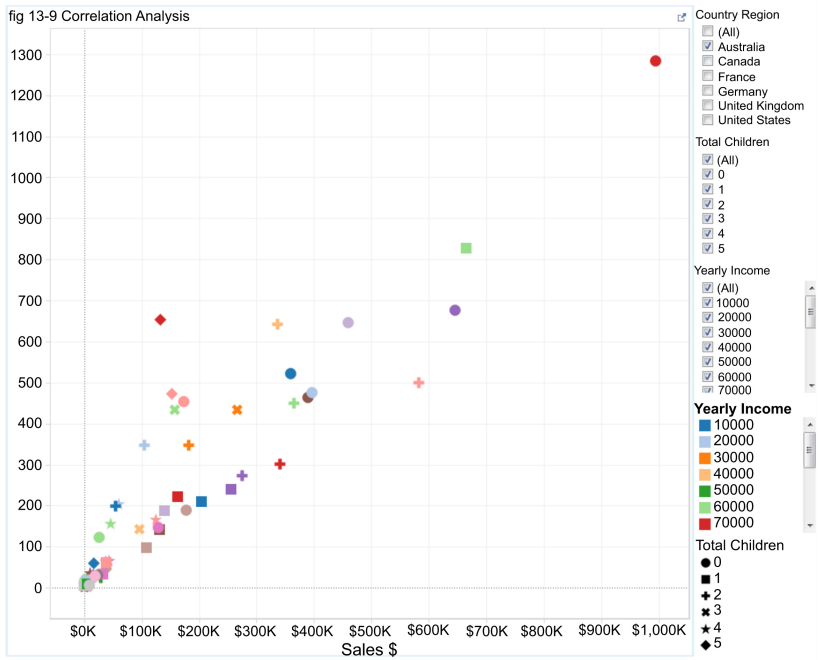


Figure: Example of a Correlation Analysis with a Scatterplot (Source: [Sherman, 2015, page 356])

From Data to Business Reports/Dashboards

What do we need to know to get from data to a "good" dashboard?

- Understanding what defines a business reporting instrument like a dashboard
- Selecting/Integrating, (pre)processing and analysing the data: Data Preprocessing
- Deriving the right information: Descriptive Statistics
- Selecting the right (visualization) tools: Graph/Table/...
- Visual design of the report/dashboard: Storytelling

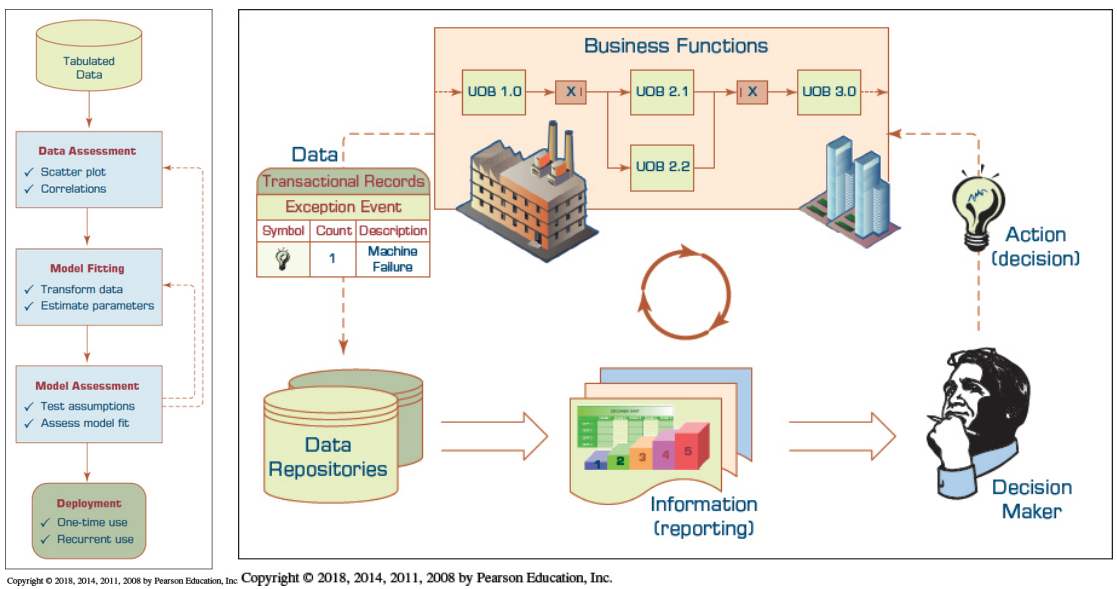


Figure: The Context of Business Reporting (Source: [Sharda et al., 2018, page 115,125])

Storytelling with Data Visualization

What is Storytelling?

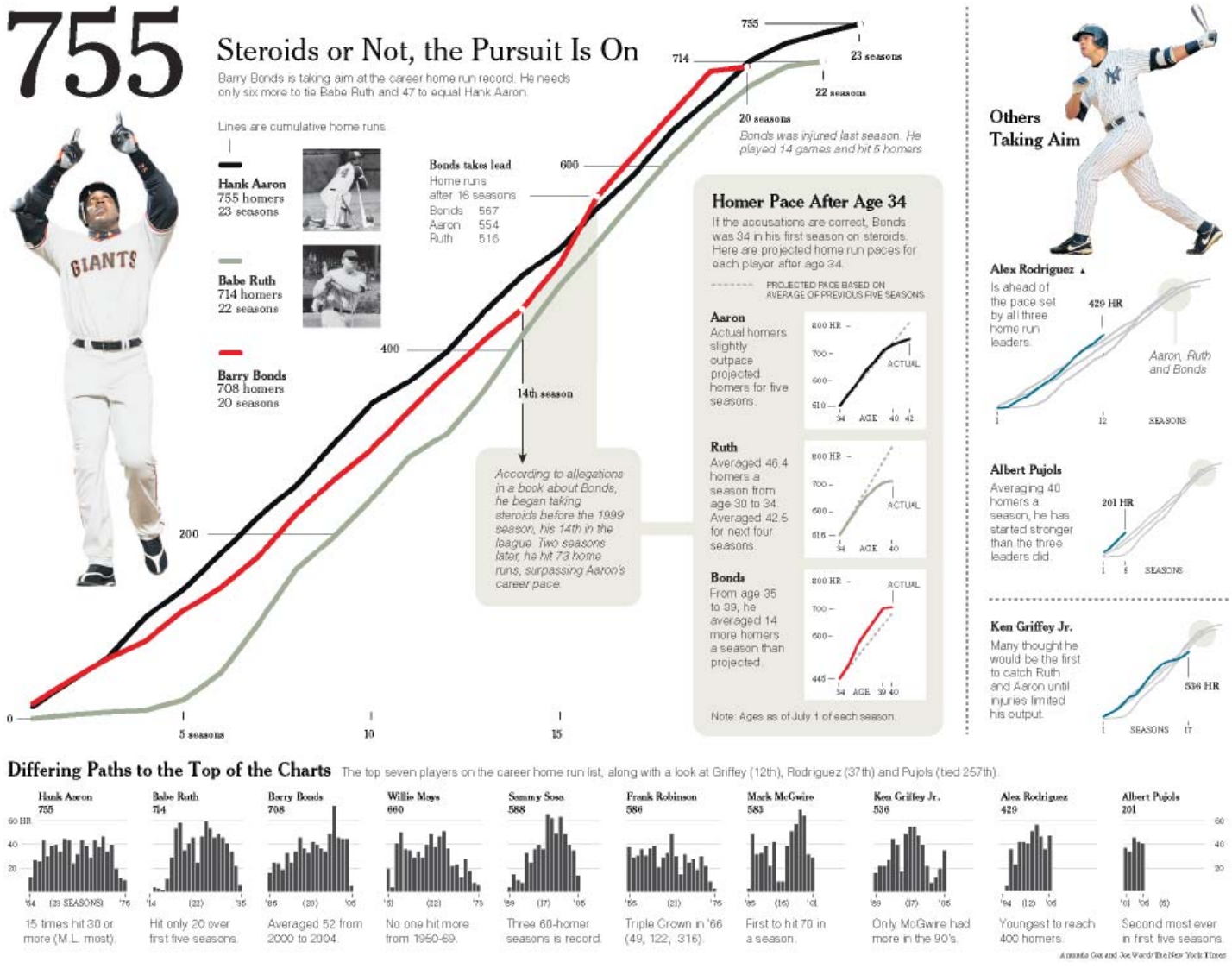


Figure: Example of Storytelling with Data Vis. (Source: [Segel and Heer, 2010, p. 1141])

Storytelling with Data Visualization: Example (2)

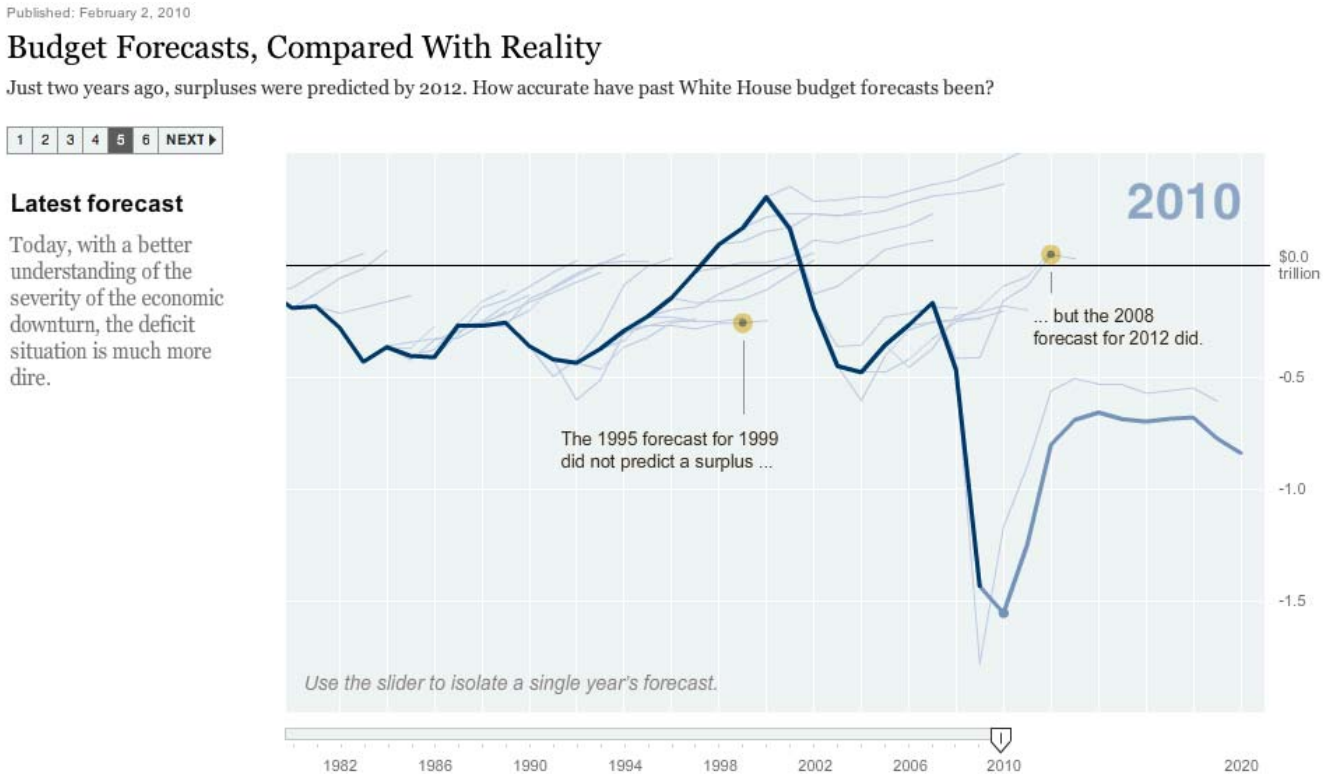


Figure: Example of Storytelling with Data Visualization (Source: [Segel and Heer, 2010, p. 1141])

Storytelling with Data Visualization: Dashboard Example

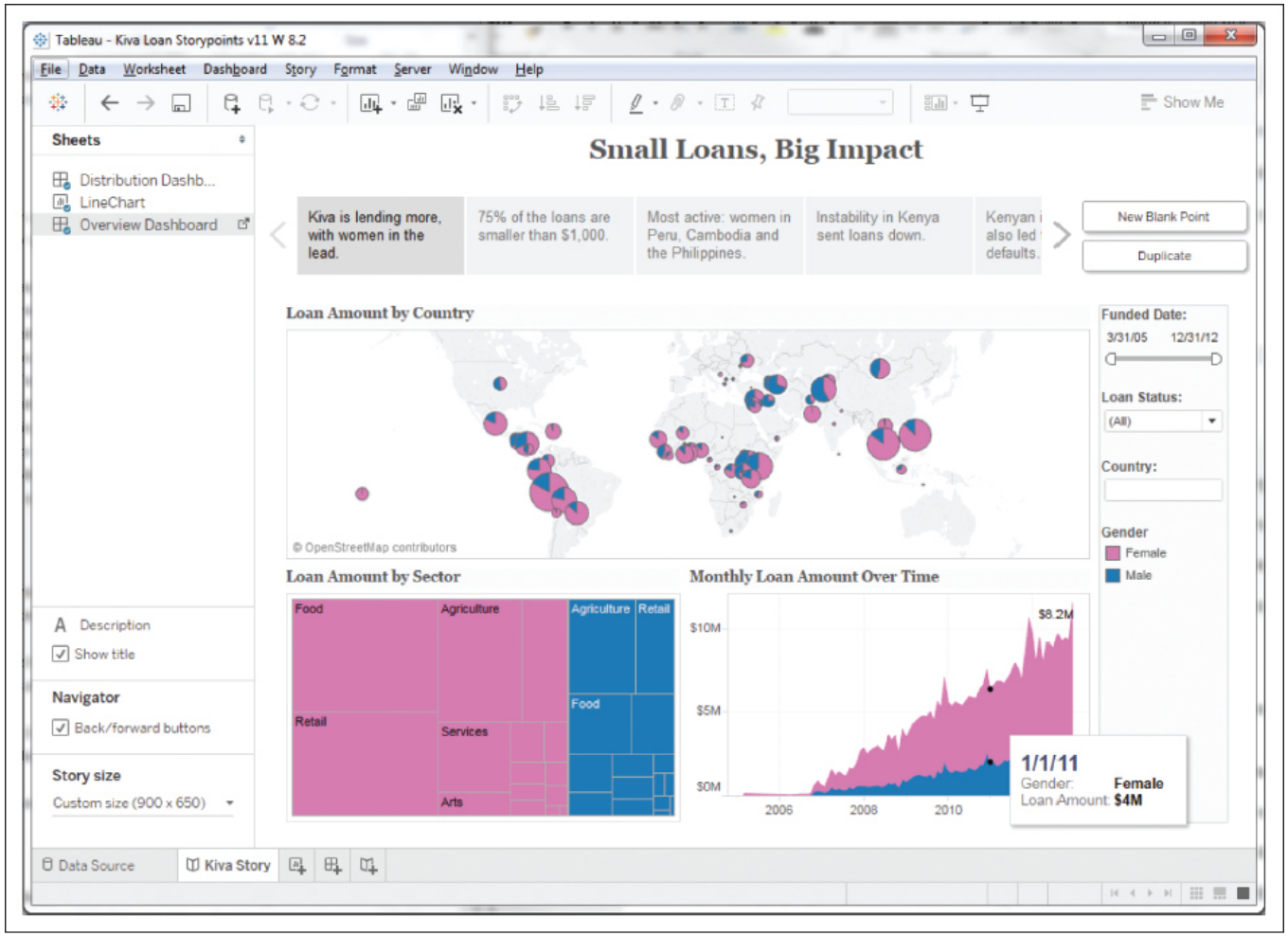
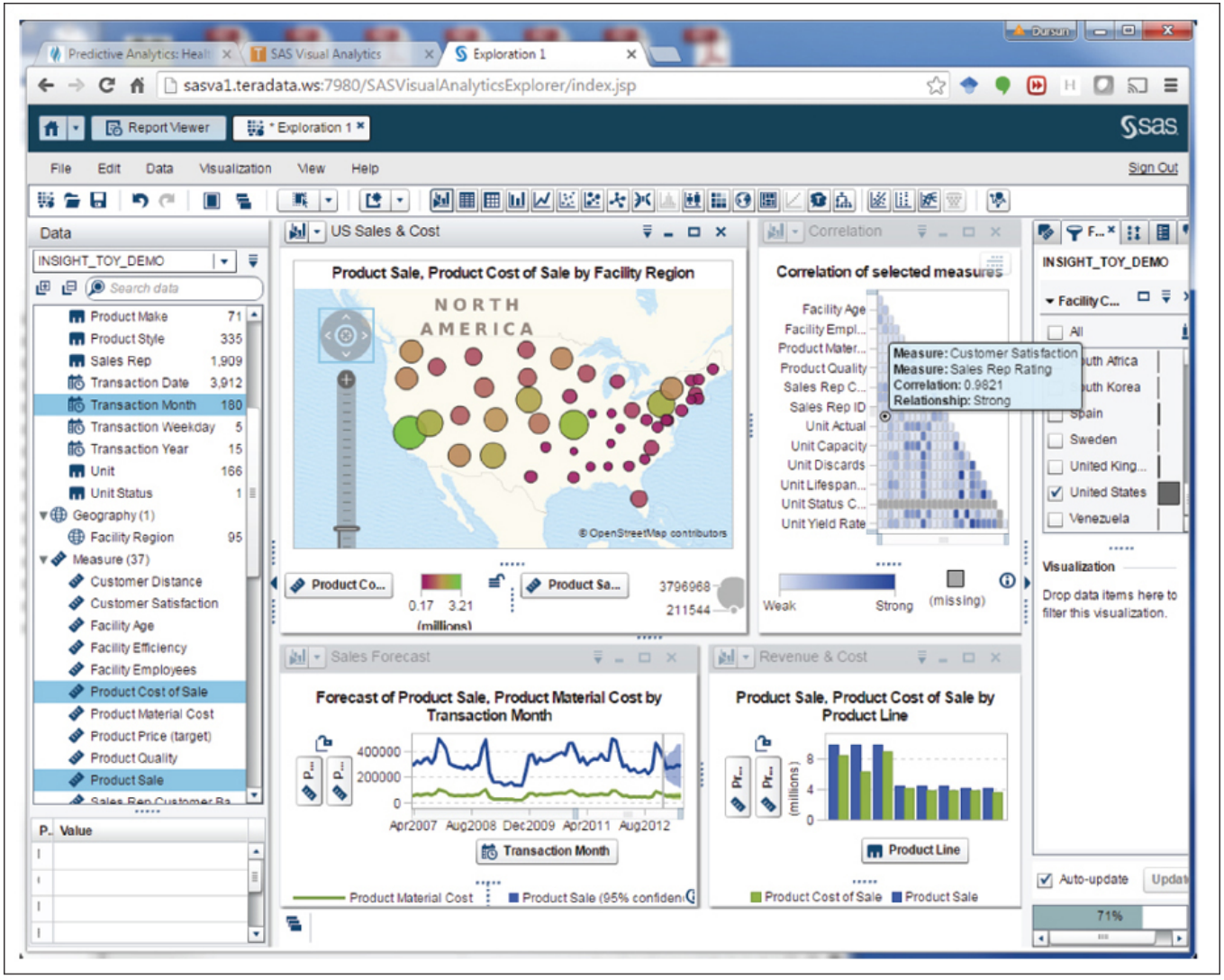


Figure: Example of Storytelling with Data Visualization (Source: [Sharda et al., 2018, page 140])

Storytelling with Data Visualization: Report Example



Source: SAS.com.

Figure: Example of Storytelling with Data Visualization (Source: [Sharda et al., 2018, page 142])

Narrative Visualization: Storytelling Principles[17]

Author-Driven

- linear ordering of scenes
- heavy messaging
- no interactivity

Reader-Driven

- no prescribed ordering
- no messaging
- free interactivity

[17] See [Segel and Heer, 2010].

Narrative Visualization: Genres Overview

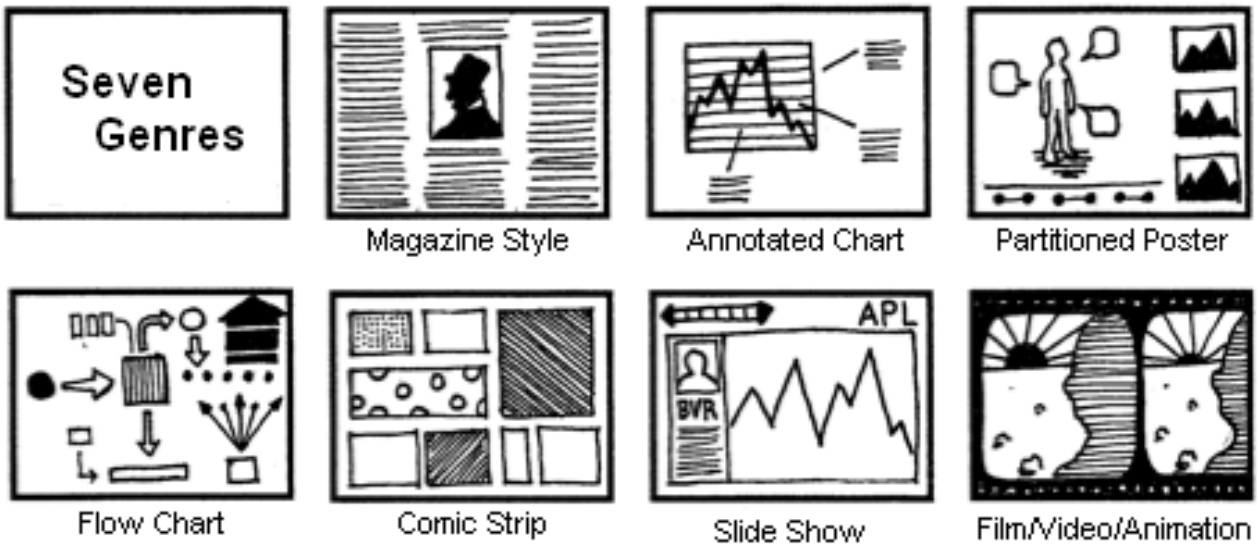
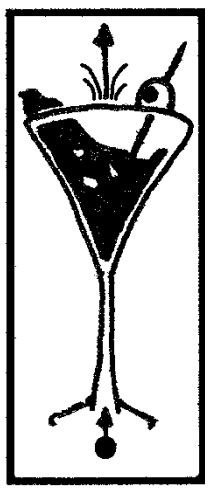


Figure: Genres of Narrative Visualization (Source: [Segel and Heer, 2010, page 1145])

Narrative Visualization: Storytelling Schemas[18] (1)

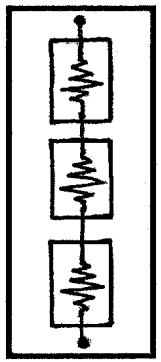


Martini Glass Structure

- Provide a clear starting point to focus the user's attention
- Examples: a "catchy" illustration, headline, ...
- From that point on, ensure a logical flow
- Sequentially direct attention by *vectorial reference*
- Example: arrows, presentation in a row/column, animation
- Eventually open the flow up for exploration
- User might want to explore different parts, depending on their information needs
- Example: Details-on-demand (interactive)

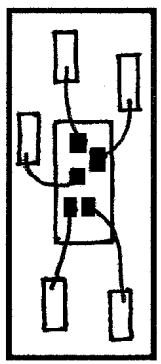
[18] Source: [Segel and Heer, 2010, page 1146].

Narrative Visualization: Storytelling Schemas (2) [19]



Interactive Slideshow

- Follows slide-show-format, but allows for interaction early on
- Mix of author-driven and reader-driven
- Allows the user to stay on one "slide" longer, if needed
- Example: time slider, with animations between



Drill-Down-Story

- Start with a general theme showing the big picture
- Allow the user to drill down on different parts of the picture
- Example: A map providing -upon interaction- more detailed information in each location

[19] See [Segel and Heer, 2010].

Narrative Visualization: Hints & Caveats [20]

Tacit Tutorial

- Stimulate and enable self-exploration and learning-by-doing
- Indicate interaction possibilities, explain them on the way
- Example: animated "ghost-guided" interaction to create figures
  indicate interactive components (buttons, ...) by pointer symbols

Outliers among Visual Features Attract Attention

- Examples: differences in color, shape, orientation
- For consistency, align visualization of similar content
- For highlighting outliers, use distinct visual features

Mind Cultural Differences and Technical Difficulties!

- reading orientation (left-right/right-left, bottom-up/bottom-down)
- color and intensity coding
- avoid red/green, avoid yellow/light green for presentations (beamer!)
- ...

[20] See [Segel and Heer, 2010].

Outline and Summary[2]

- The Nature of Data
  See [Sharda et al., 2018, chapter 2.1], [Zliobaite et al., 2016]
- Data Quality and Integrity
  See [Sharda et al., 2018, chapter 2.2–2.3]
- Data Preprocessing
  See [Sharda et al., 2018, chapter 2.4 & Application Case 2.1]
- Statistics Repetition[1]
  See [Hand et al., 2001, Appendix A.1]
- Statistical Modelling
  See [Sharda et al., 2018, chapter 2.5–2.7],
  [Cover and Thomas, 2006, chapter 2], [Hand, 2008], and others
- Business Reporting
  See [Sharda et al., 2018, chapter 2.7]
- Data Visualization
  See [Sharda et al., 2018, chapter 2.8–2.11] and [Segel and Heer, 2010]

▸ Start ▸ Appendix

[1] **The Freudenthal Institute provides a self-learning tool for statistics. If you are interested, contact Marcela Ruiz** `m.ruiz@uu.nl`

[2] Note: This lecture integrates knowledge from several further sources (cited where used, except if my own).

Outlook: Descriptive Analytics Part B

Lecture Next Tuesday

- **09:15 – 10:45**
- Business Intelligence and Data Warehousing: [Sharda et al., 2018, chapter 3]

Preparation

In preparation, please read

- An Overview of Business Intelligence Technology: [Chaudhuri et al., 2011]
- Organizational factors in data warehouse architecture selection: [Ariyachandra and Watson, 2010]

Any More Questions?

Thank you!

Appendix

Bibliography I


Ariyachandra, T. and Watson, H. (2010).
Key organizational factors in data warehouse architecture selection.
Decision Support Systems, 49:200212.


Babcock, B., Datar, M., Motwani, R., and O'Callaghan, L. (2003).
Maintaining variance and k-medians over data stream windows.
In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, page 234243. ACM.


Chaudhuri, S., Dayal, U., and Narasayya, V. (2011).
An overview of business intelligence technology.
Communications of the ACM, 54(8):8898.


Cover, T. M. and Thomas, J. A. (2006).
Elements of Information Theory.
Wiley-Interscience, 2 edition.


Eckerson, W. W. (2010).
Performance Dashboards: Measuring, Monitoring, and Managing Your Business.
Wiley, 2 edition.


Hand, D. J. (2008).
Statistics: A very short introduction.
Oxford University Press.

Bibliography II


Hand, D. J., Mannila, H., and Smyth, P. (2001).
Principles of Data Mining.
Adaptive Computation and Machine Learning. The MIT Press.


Hofer, V. and Krempl, G. (2013).
Drift mining in data: A framework for addressing drift in classification.
Computational Statistics and Data Analysis, 57(1):377391.


Joanes, D. N. and Gill, C. A. (1998).
Comparing measures of sample skewness and kurtosis.
Journal of the Royal Statistical Society (Series D): The Statistician, 47:183189.


Kelly, M. G., Hand, D. J., and Adams, N. M. (1999).
The impact of changing populations on classifier performance.
In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, page 367371.


Krempl, G., Zliobait, I., Brzeziski, D., Hüllermeier, E., Last, M., Lemaire, V., Noack, T., Shaker, A., Sievi, S., Spiliopoulou, M., and Stefanowski, J. (2014).
Open challenges for data stream mining research.
SIGKDD Explorations, 16(1):110.
Special Issue on Big Data.

Bibliography III


Meng, X. (2015).
Simpler online updates for arbitrary-order central moments.
arXiv preprint arXiv:1510.04923.


Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D., editors (2009).
Dataset Shift in Machine Learning.
MIT Press.


Schlimmer, J. C. and Granger, R. H. (1986).
Beyond incremental processing: Tracking concept drift.
In *AAAI*, page 502507.


Segel, E. and Heer, J. (2010).
Narrative visualization: Telling stories with data.
IEEE Transactions on Visualization and Computer Graphics, 16(6):11391148.


Sharda, R., Delen, D., and Turban, E. (2018).
Business Intelligence, Analytics, and Data Science: A Managerial Perspective.
Pearson, 4 edition.


Sherman, R. (2015).
Business Intelligence Guidebook: From Data Integration to Analytics.
Morgan Kaufmann.

Bibliography IV


Webb, G., Lee, L. K., Goethals, B., and Petitjean, F. (2017).
Understanding concept drift.
arXiv preprint, 1704.00362v1.


Winston, W. L. (1997).
Operations Research: Applications and Algorithms.
Wadsworth Publishing Company, 3rd edition edition.


Zliobaite, I., Pechenizkiy, M., and Gama, J. (2016).
An overview of concept drift applications.
In Japkowicz, N. and Stefanowski, J., editors, *Big Data Analysis: New Algorithms for a New Society*, page 91114. Springer, Cham.


Zliobait, I. (2009).
Learning under concept drift: an overview.
Technical report, Vilnius University.