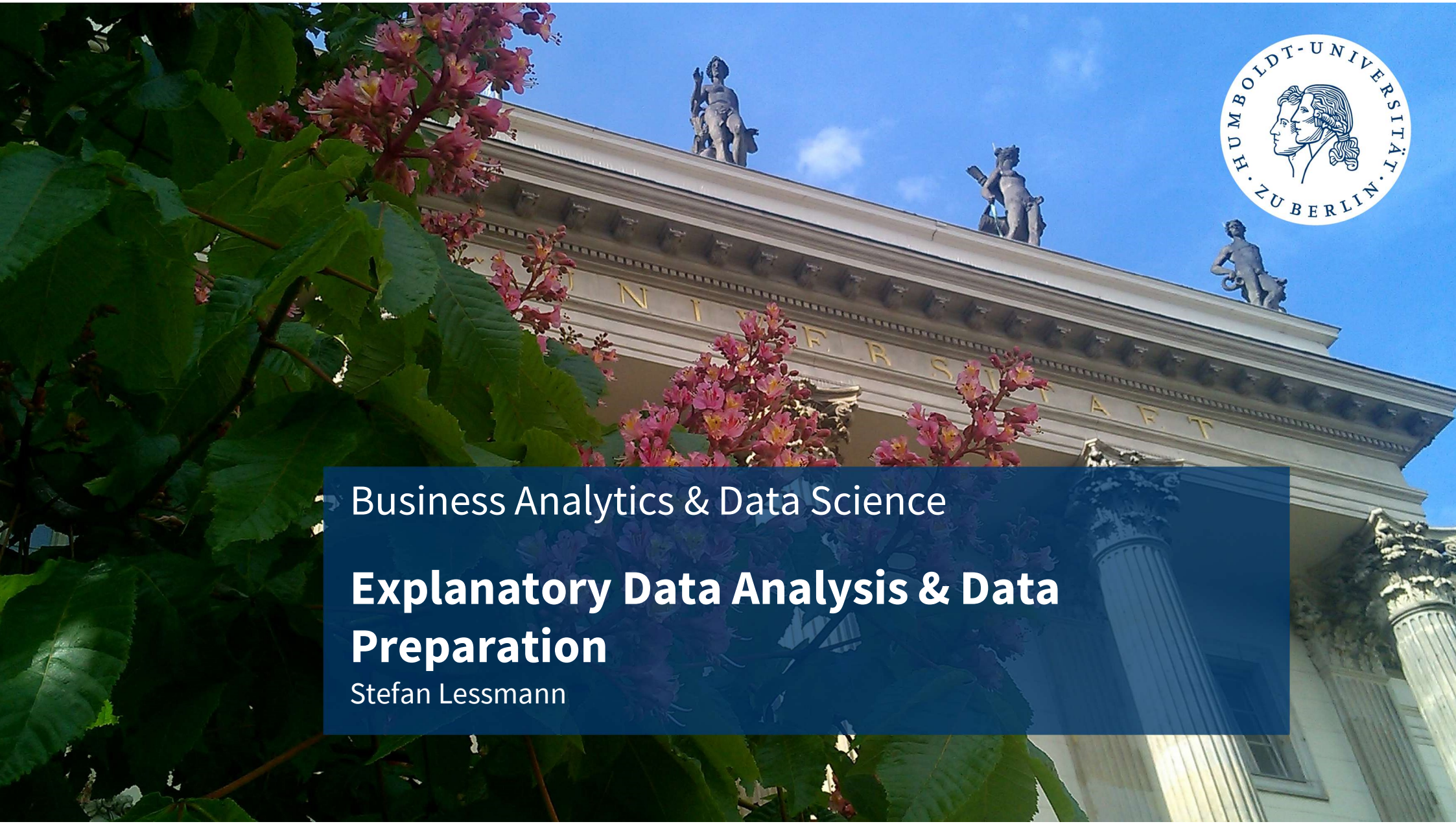


Business Analytics & Data Science

Explanatory Data Analysis & Data Preparation

Stefan Lessmann

Agenda

Introduction
Motivation, drivers, and types of data

Explanatory Data Analysis
Scope, motivation, popular visualizations

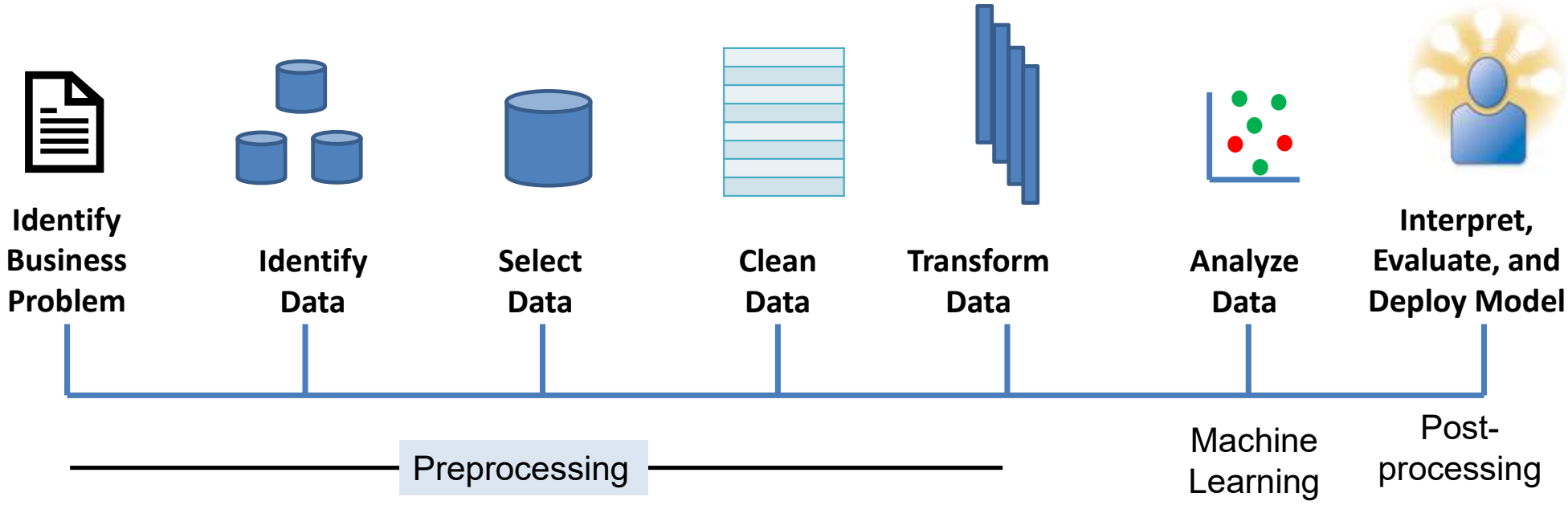
Data preparation
Process model, cleaning strategies, handling continuous & categorical variables

Summary

Introduction

Motivation, drivers, and types of data

Recap: Data Analytics Process Model



Recap: Data Structure for Business Analytics

Target (…)

Features (aka variables, covariates, characteristics, predictors, attributes)

Default (e.g., 90 days late)	Bureau score	Collateral	Debt/Income	Years at address	Age	…
No	650	Yes	20%	2	<21	…
Yes	280	No	43%	0	21-29	…
No	750	Yes	27%	8	30-39	…
No	600	Yes	18%	4	40-50	…
No	575	No	33%	12	>50	…
No	715	Yes	24%	1	21-29	…
Yes	580	No	18%	6	40-50	…
No	410	Yes	29%	4	21-29	…
Yes	800	Yes	14%	10	40-50	…

Motivation and Drivers

- **Dirty, noisy data**
  - Age = -2003
- **Data integration and data merging problems**
  - Amounts in euro versus amounts in dollar
- **Inconsistent data**
  - Value '0' means actual zero or missing value
- **Incomplete data**
  - Income = ?
- **Duplicate data**
  - Salary versus professional Income
- **Garbage in, garbage out (GIGO)**
- **Very time consuming (80% rule)**

Explanatory Data Analysis

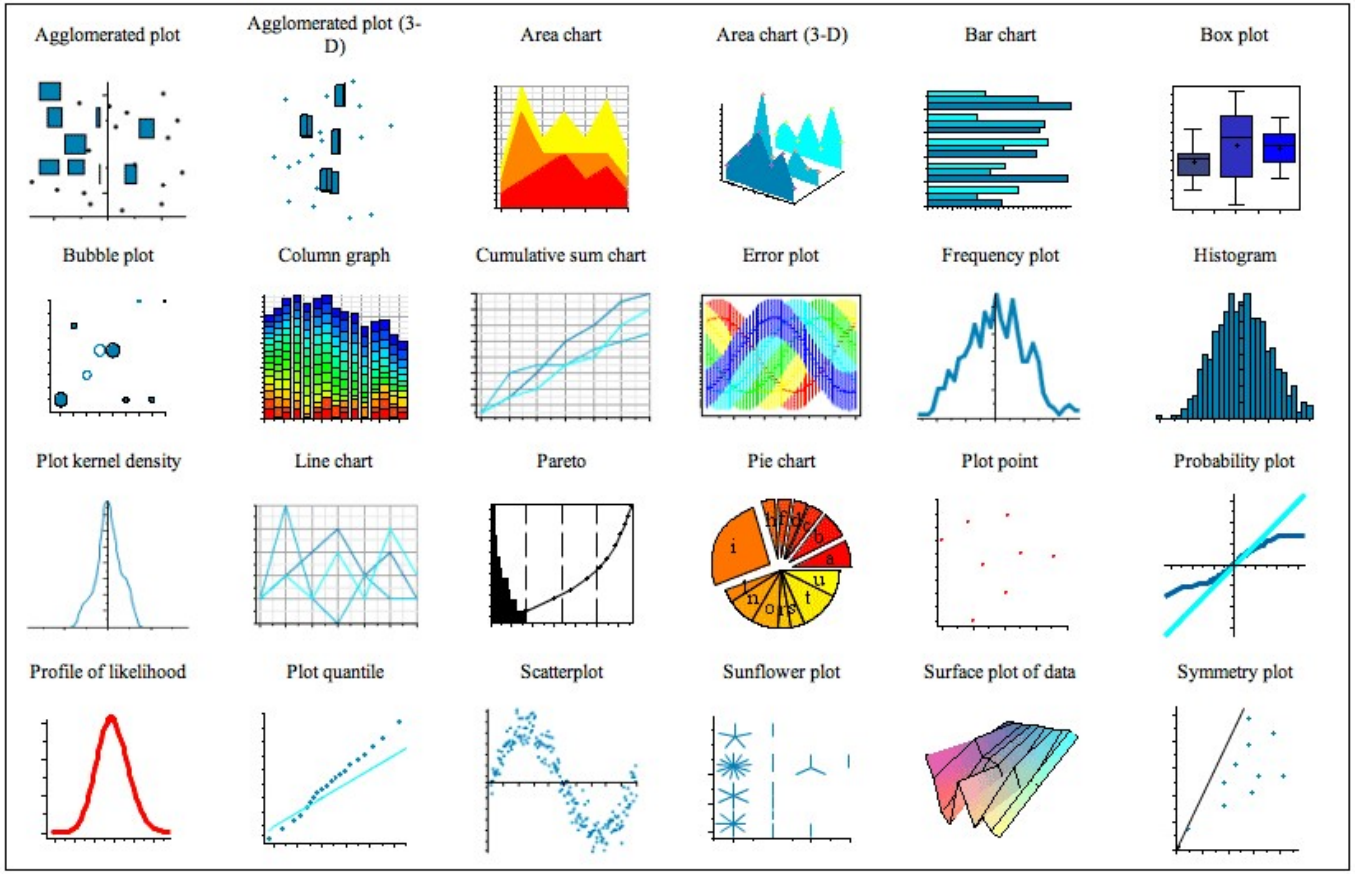
Scope, motivation, popular visualizations

Explanatory Data Analysis (EDA)

“A first look at the data”

EDA methods play a key role in data preparation. The modern term visual analytics also displays some overlap with what is traditionally referred to as EDA in that similar techniques are employed.

However, VA would typically make use of more sophisticated, possibly interactive means of visualizations.



Explanatory Data Analysis

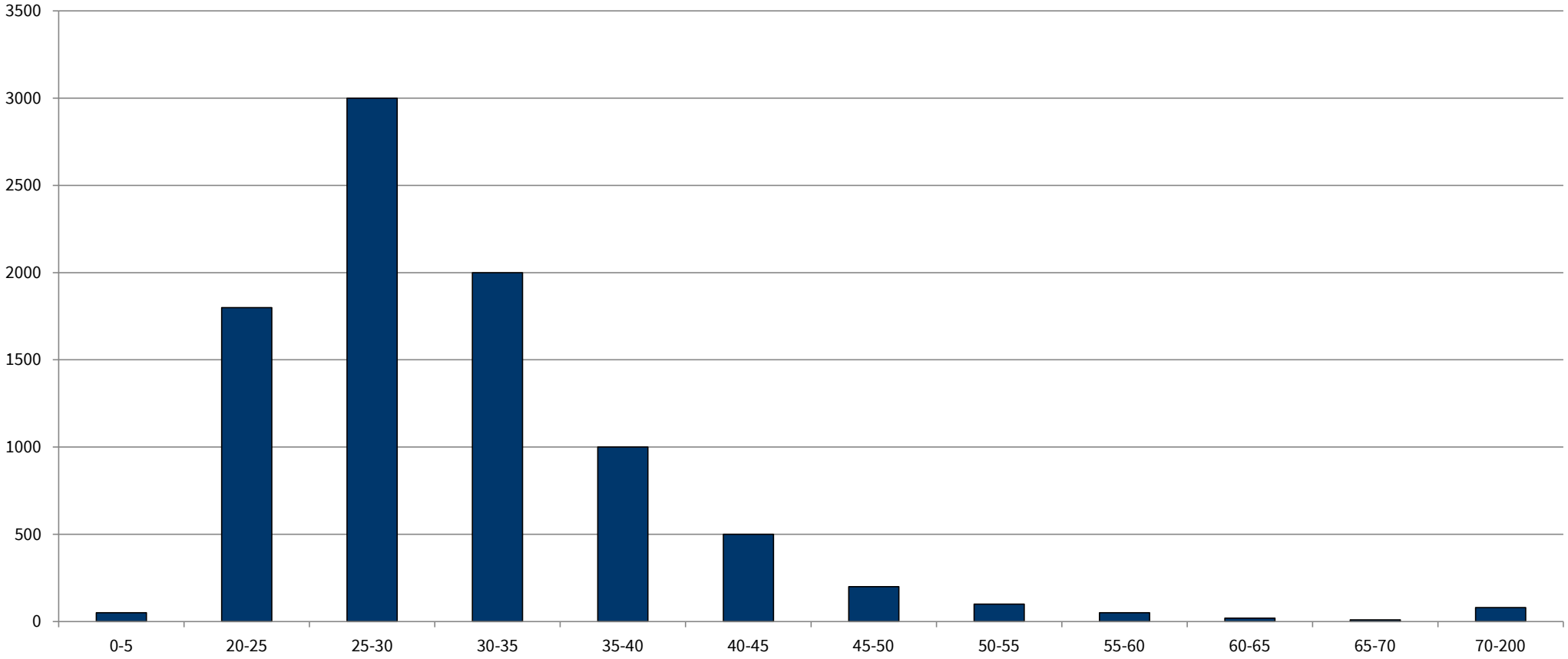
"A first look at the data"

- **Goes back to Tuckey's pioneering work in 1960s**
- **Prepare formulation of hypotheses**
- **Prepare model selection (e.g., check assumptions)**
- **Find relationships (e.g., dependency, co-occurrence, mistakes)**

	Univariate	Multivariate
Graphical		
Non-graphical		

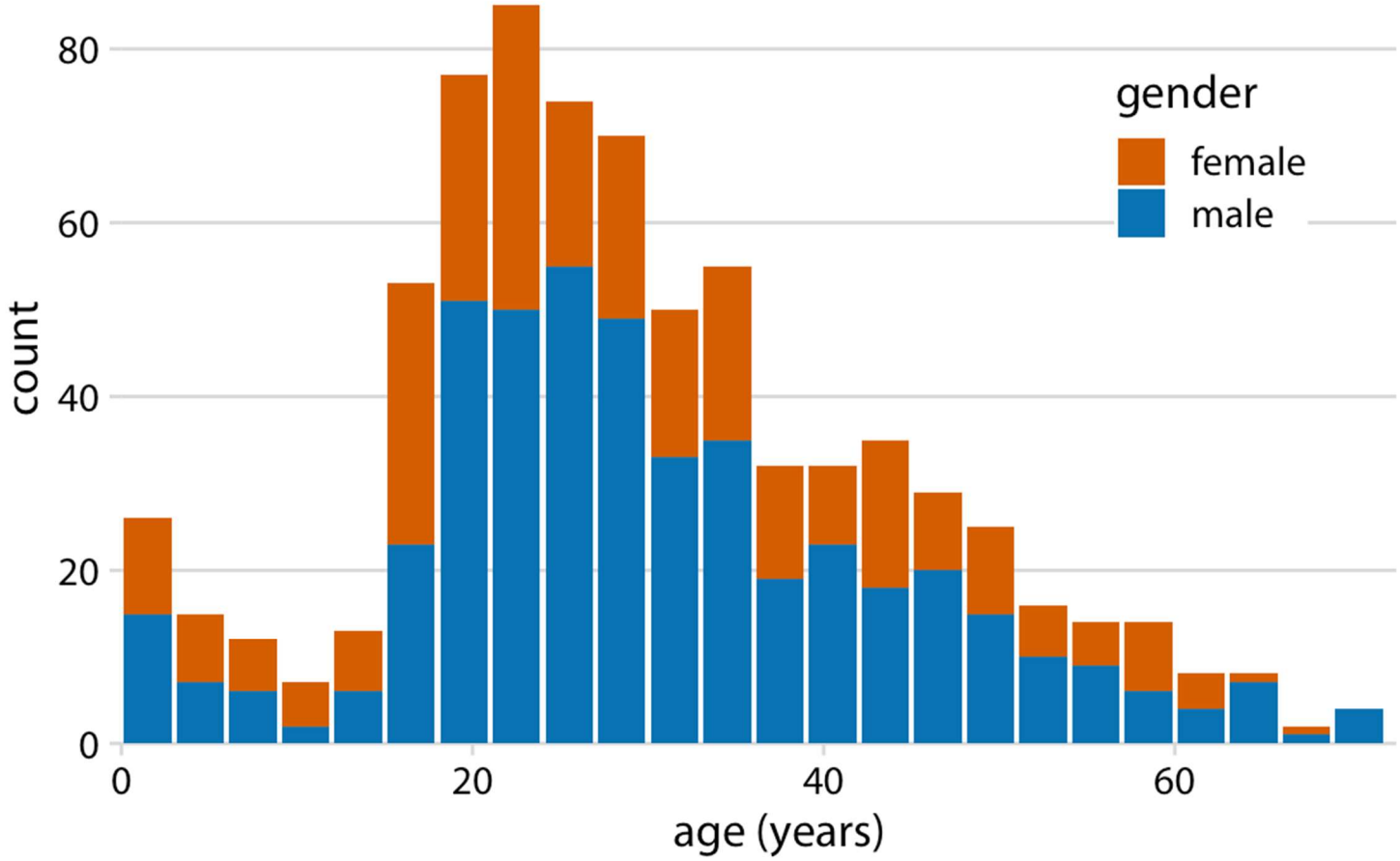
Visual Data Exploration

Distribution of a categorical feature AGE using count plot



Visual Data Exploration

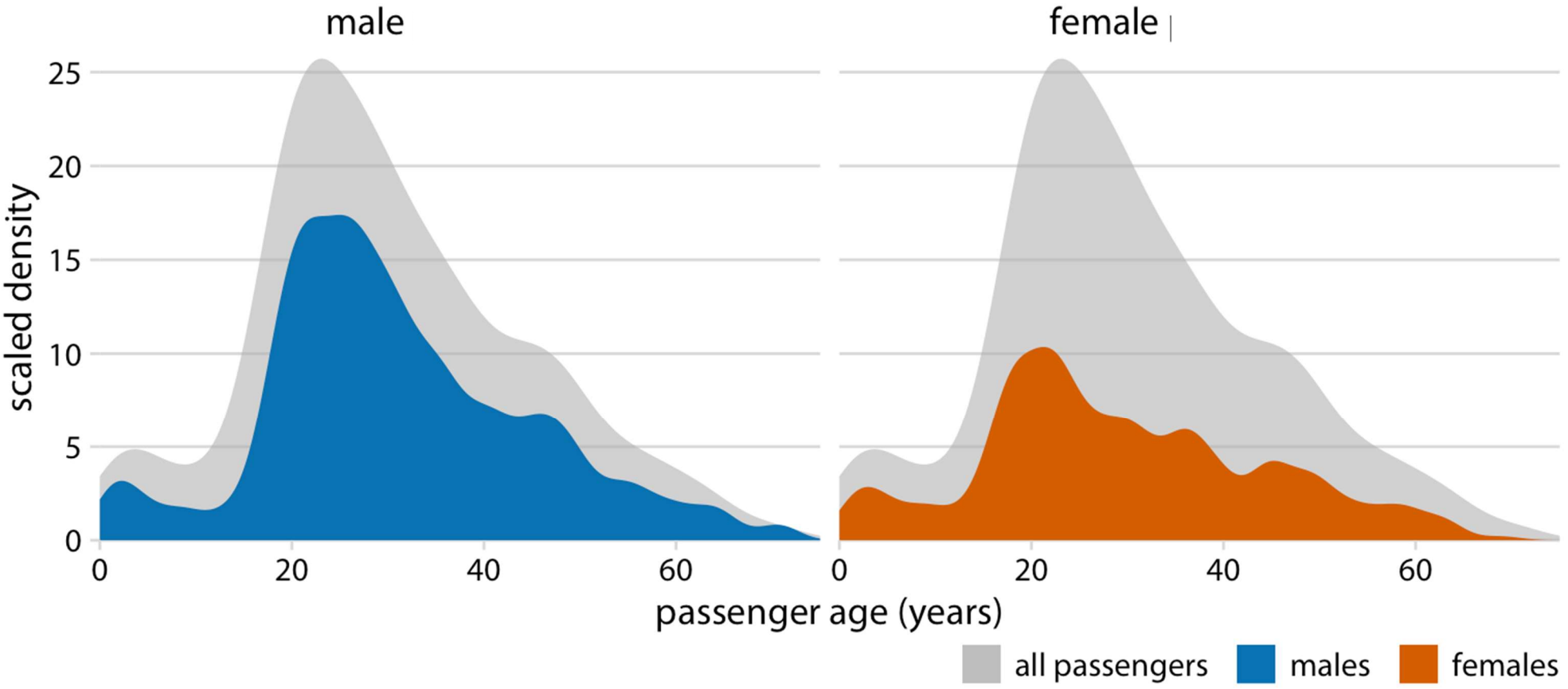
Stacked histogram to visualize multiple distributions



Wilke (2019)

Visual Data Exploration

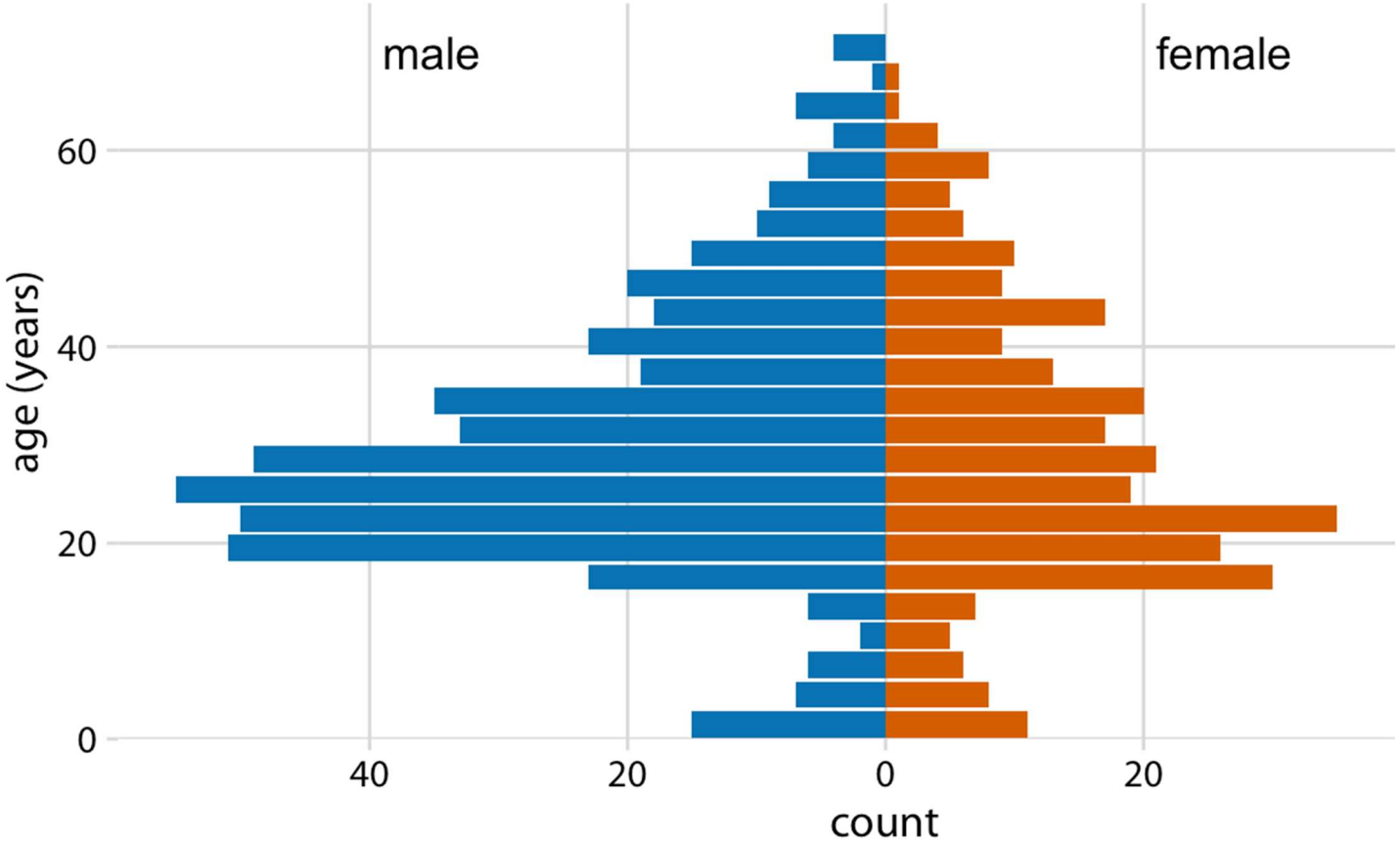
Better ways to depict the same data?!



Wilke (2019)

Visual Data Exploration

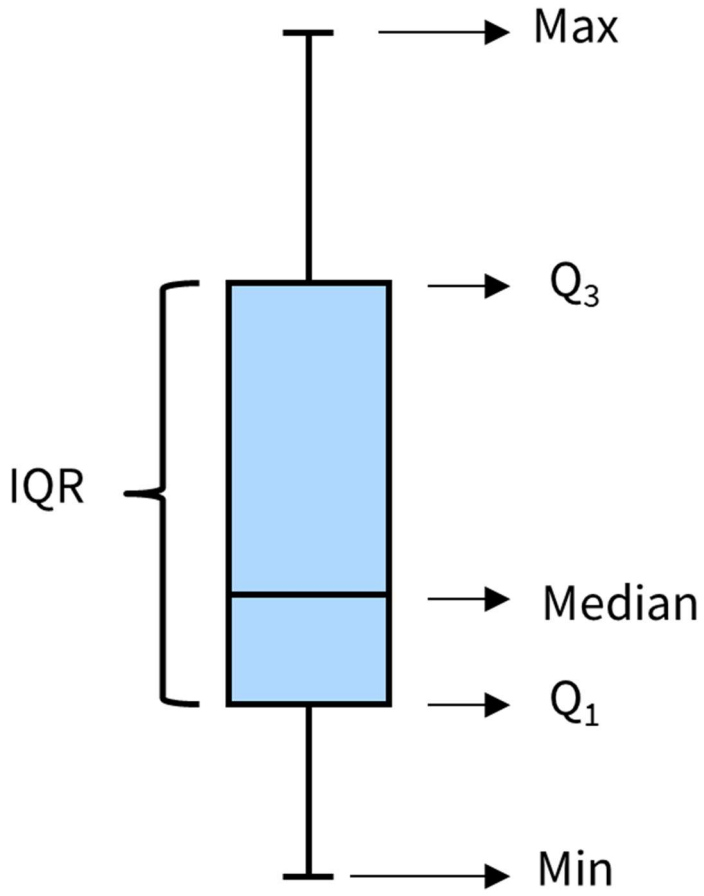
Better ways to depict the same data ?!



Wilke (2019)

Visual Data Exploration
Box plot

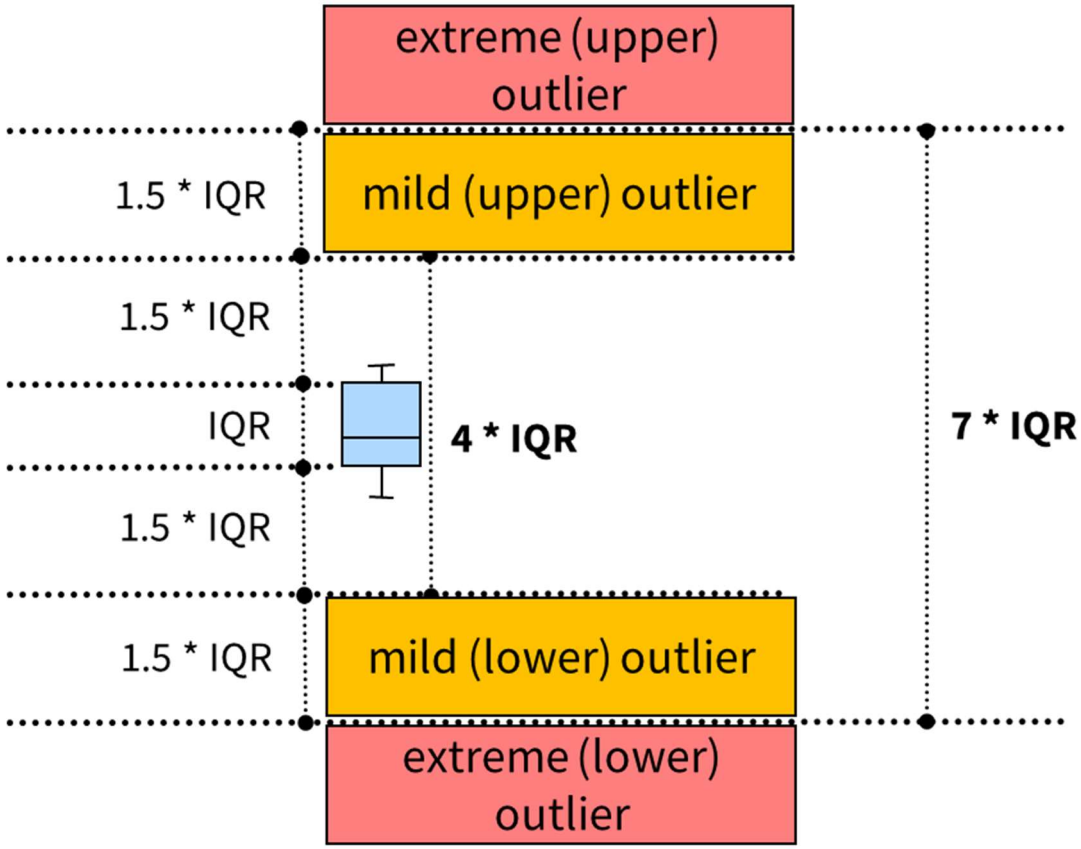
- **Visual representation of five numbers**
  - Median M $P(X \leq M)=0.50$
  - First Quartile Q_1 $P(X \leq Q_1)=0.25$
  - Third Quartile Q_3 $P(X \leq Q_3)=0.75$
  - Minimum
  - Maximum
  - Inter-Quartile-Range (IQR)=Q_3-Q_1
    - Covers the middle fifty percent of the data
    - Simple, robust measure of spread
- **Note that some variants of the box plot exists**



Visual Data Exploration

Box-plot-based outlier rule of Tuckey 1977

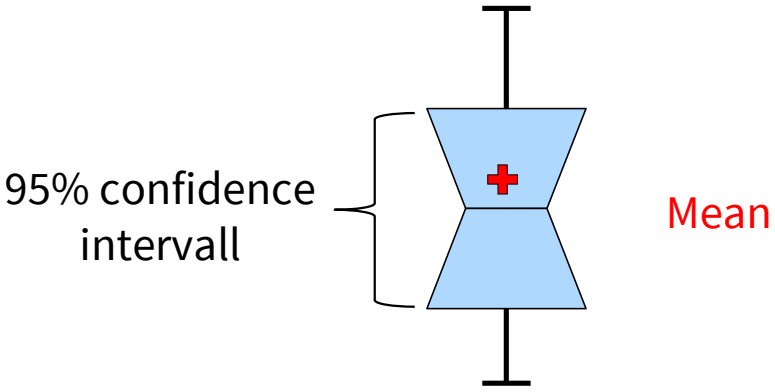
- **Definition of outliers based on distance to first or third quartile and inter-quartile range**
  - Upper outliner
  - Lower outlier
- **Mild outliers:** values between 1.5 and 3 IQR away from Q1 / Q3
- **Extreme outliers:** values with more than 3 IQR away from Q1 / Q3

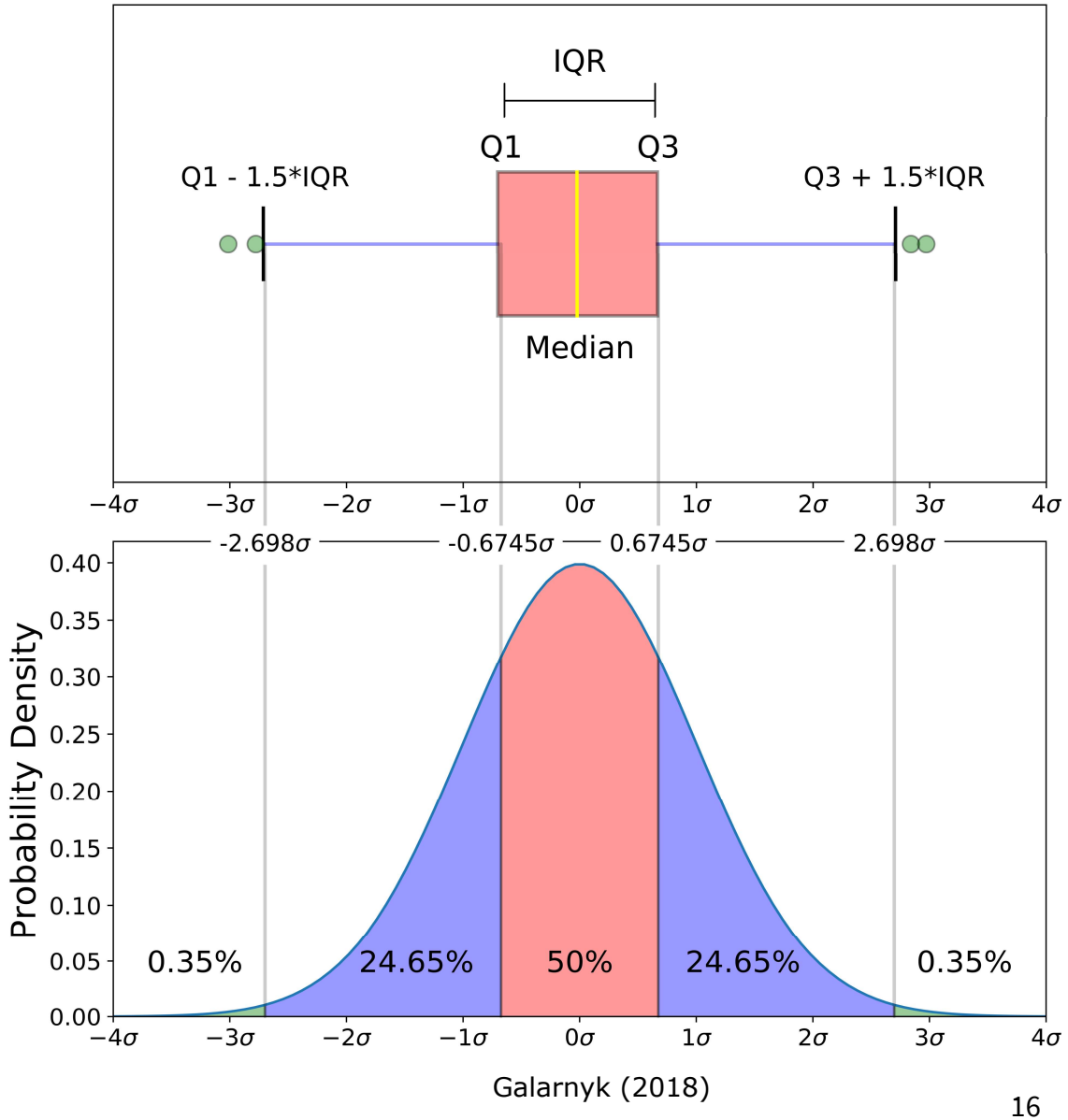


Visual Data Exploration

Box plot revisited

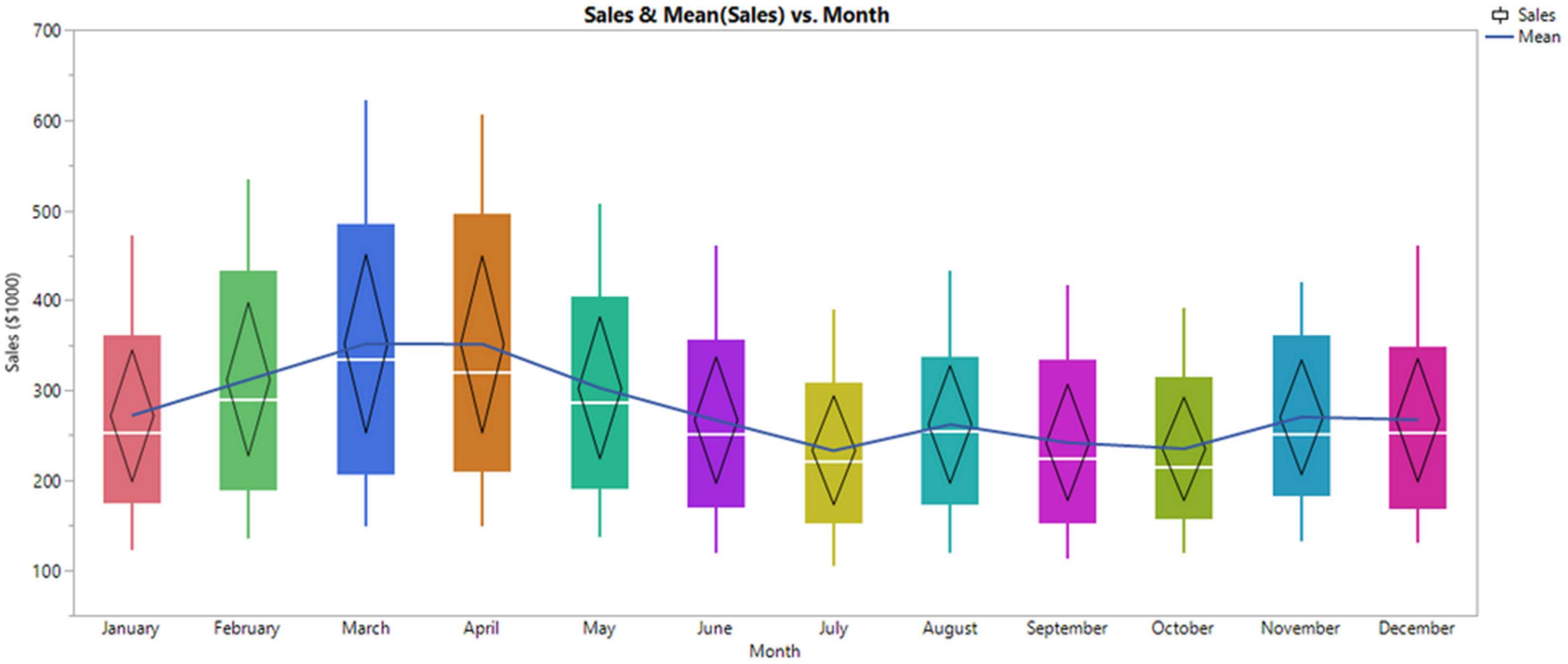
- **Upper panel depicts more common form of the box plot where whiskers indicate the boundaries of (soft) upper and lower outliers**
- **Then outliers enter the plot as dots**
- **Plot can also incorporate notches and/or depict the mean**





Box Plot

Multiple variables or repeated measures of the same variable

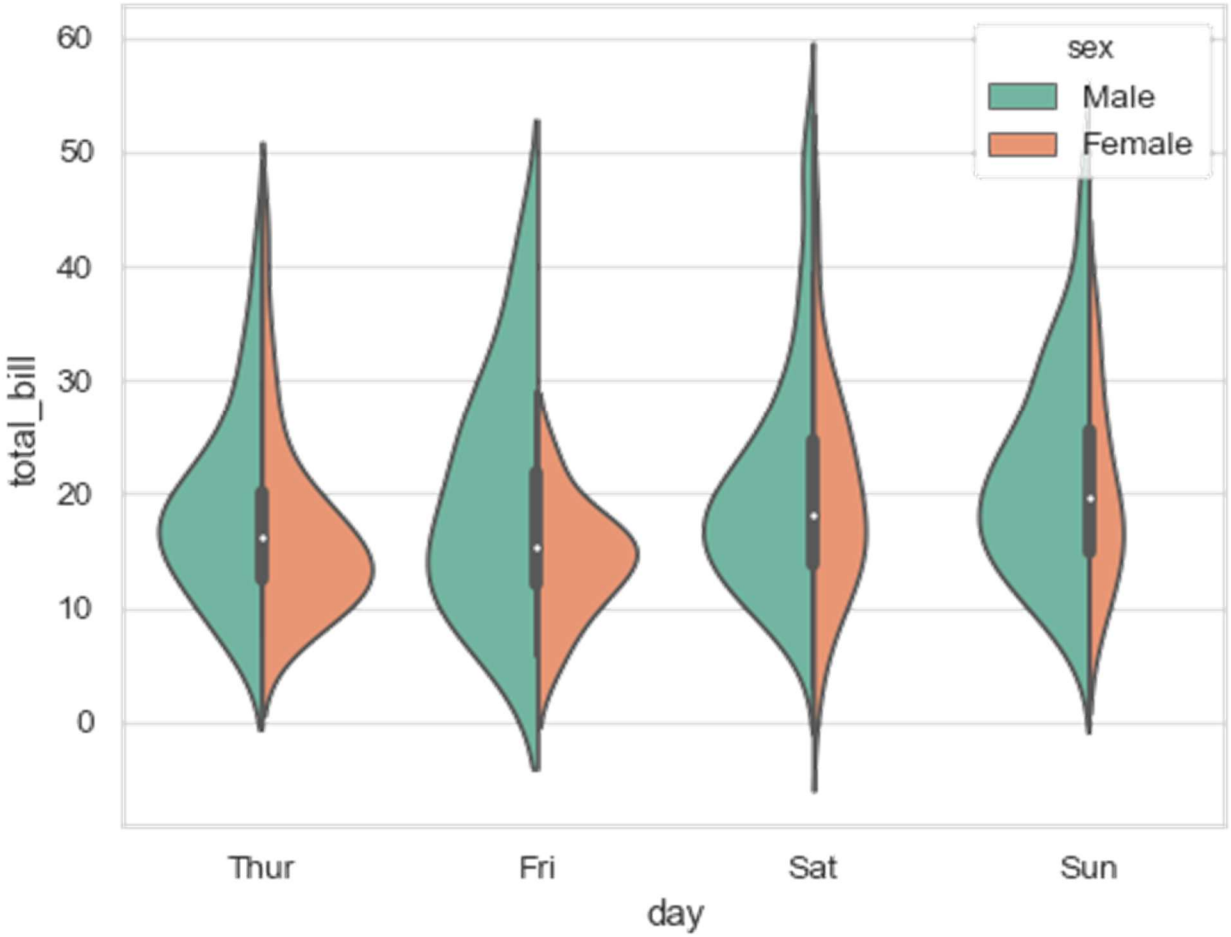


https://community.jmp.com/t5/Discussions/Show-means-on-box-plot-in-graph-builder/td-p/67947

Violin Plot

Box plot with added kernel density plot

- **Common box plot elements**
  - Median
  - Interquartile range
  - Possibly sample points
- **Probability density**
  - Typically smoothed
  - Useful for multi-modal distributions
  - Symmetric around the “box plot”
  - Stratified by categories, in which case the width can indicate prevalence

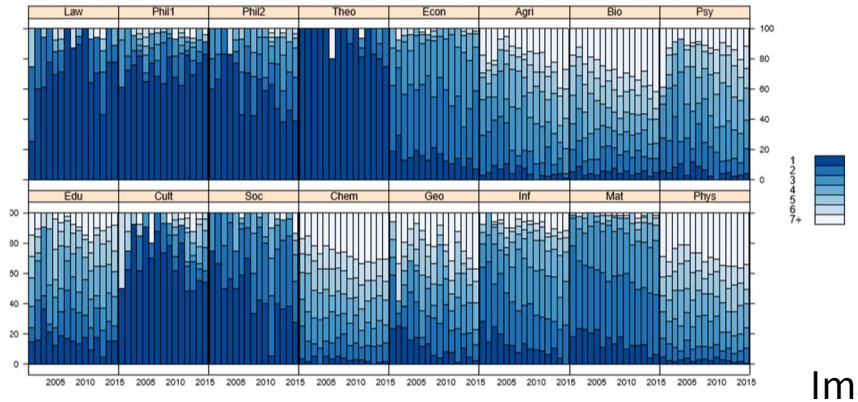


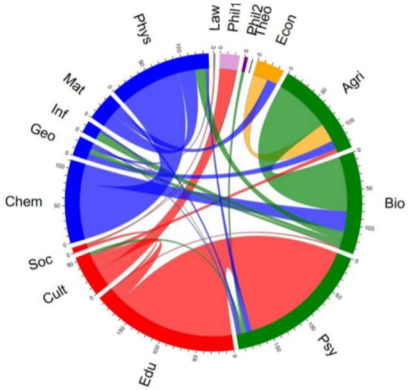
https://seaborn.pydata.org/generated/seaborn.violinplot.html

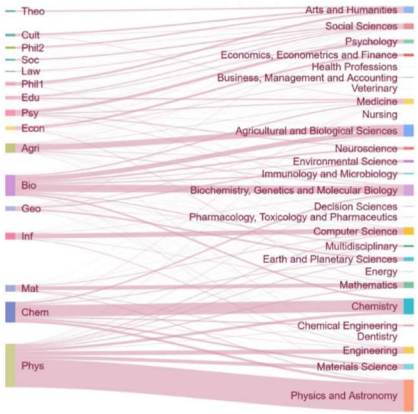
Explanatory Data Analysis – Lessons learnt

The single first task when obtaining a new data set

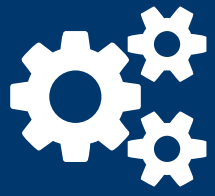
- **Multiple perspectives:** uni- vs. multivariate, (non-)graphical, …
- **No silver bullet or reference model**
  - Creative process
  - Requires understanding of the domain and statistics
- **Many *non-standard* ways to depict complex data**
  - Relatively easy to construct using contemporary software
  - Ggplot2, Matplotlib, Pandas, Seaborn,…







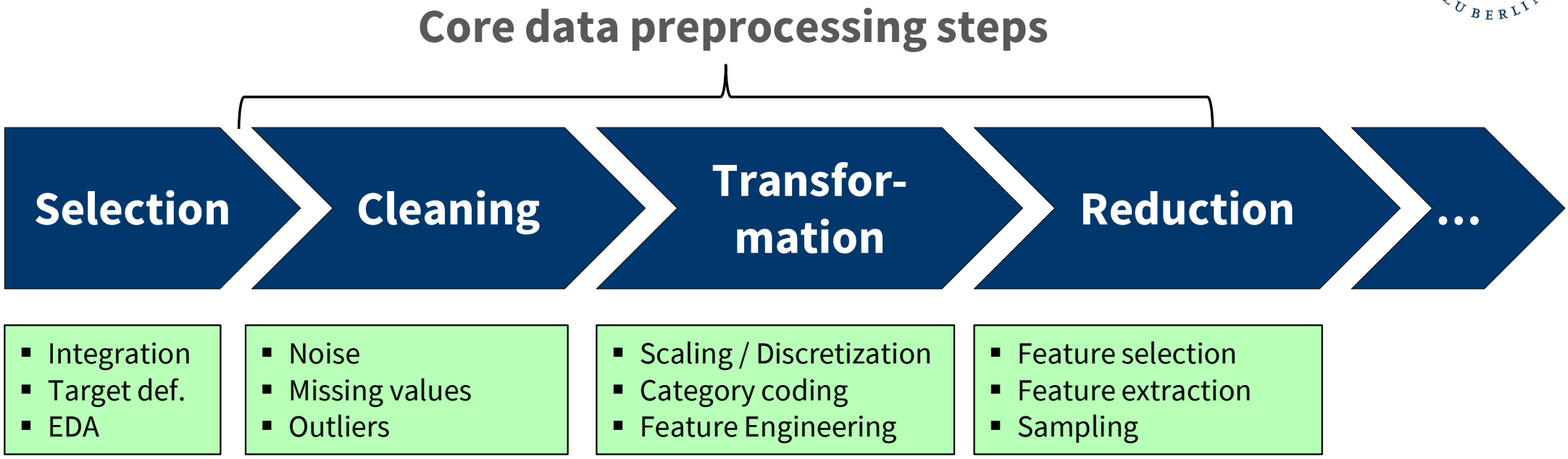
Images from: Zharova et al. (2017)



Data Preparation

Process model, cleaning strategies, handling continuous and categorical variables

Data Preparation Process



In practice, data preprocessing activities do not follow a strictly sequential order. Operations in different core stages are interrelated, so that decisions in a later stage may affect the suitability of preprocessing operations in an earlier stage. Therefore, data preprocessing is best thought of as an iterative approach, which traverses the above steps in multiple cycles.

Structured Tabular Data

Types of Variables

- **Continuous**
  - Synonyms: numeric, real
- **Categorical**
  - Synonyms: discrete, category, non-numeric, factors (-> *R* terminology)
  - Admissible values are called levels
  - Three types
    - Binary: just two levels
    - Nominal: no ordering between levels
    - Ordinal: implicit ordering between levels

Numeric	Categorical			Other	
	Binary	Nominal	Ordinal		

- **Preprocessing operation vary across different types of variables**

Data Cleaning

- **Noise**
  - Different viewpoints
    - Umbrella term for various data problems
    - Measurement inaccuracies
  - Application specific concepts
    - White noise in time series analysis
    - Label noise in classification
  - Actual data errors
- **Missing values**
  - Attribute value is not available
  - For example, customer did not give her/his date of birth
- **Outliers**
  - Attribute value appears extreme
  - Detection versus treatment

Missing value

Default	DATE OF BIRTH	SALARY in K$	...
NO	16.03.1973	$75,00	...
NO	09.12.1984	$65,00	...
NO	03.05.1961	$125,00	...
NO	17.02.1979	$55,00	...
NO	08.08.1988	$9,250,00	...
NO	?	$60,00	...
NO	24.09.1976	$83,00	...
YES	13.06.1998	$15,00	...
YES	09.04.1789	$45,00	...
YES	17.11.1979	$111,00	...

Actual data error

Outlier?

Data Cleaning Strategies

High-level overview

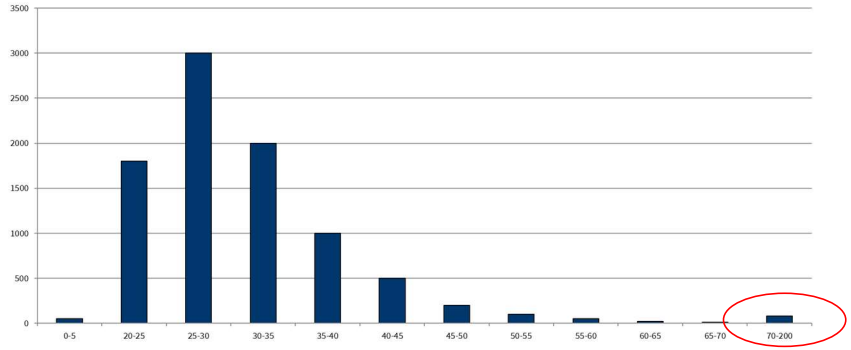
- **Errors: correct if feasible; else treat as missing value**
- **Missing values**
  - Keep
    - Fact that a variable is missing can be important information
    - Encode variable in a special way (e.g., as separate category)
  - Delete
    - When number of missing values is excessive
    - Horizontally versus vertically missing values
  - Replace using some imputation procedure
- **Outliers**
  - Detection using graphical / statistical approaches or clustering
  - Treatment options
    - Keep as is if outlier is valid
    - Treat as missing value if outlier is invalid
    - Truncate (if expected impact on the analysis appears excessive)

Data Cleaning Strategies
Missing Values

- **Replacement**
  - Continuous attributes: mean/median replacement
  - Nominal attributes: mode replacement (= most frequent category)
- **Imputation:** estimate replacement value with a “mini-model” using other covariates
  - Tree-based algorithms
  - Nearest neighbour approaches
  - Markov Chain Monte Carlo methods
- **Stratification by grouping variable**
  - For example in classification
  - Compute replacement value among observations of the same class
- **Consider adding a dummy variable to flag missing value correction**
- **Is data missing at random?**

Data Cleaning Strategies
Outliers

- **Valid unusual observations vs. data entry errors**
- **Difficult to handle**
  - Univariate versus multivariate
  - Detection versus treatment
- **Visual detection using histogram of box plot**
- **Numeric detection using *z*-scores or IQR**
  - How many std. deviations observation is away from the mean
  - μ is the mean of variable x_i and σ its standard deviation
  - Outliers are variables having $|z_i| > 3$ (or 2.5)



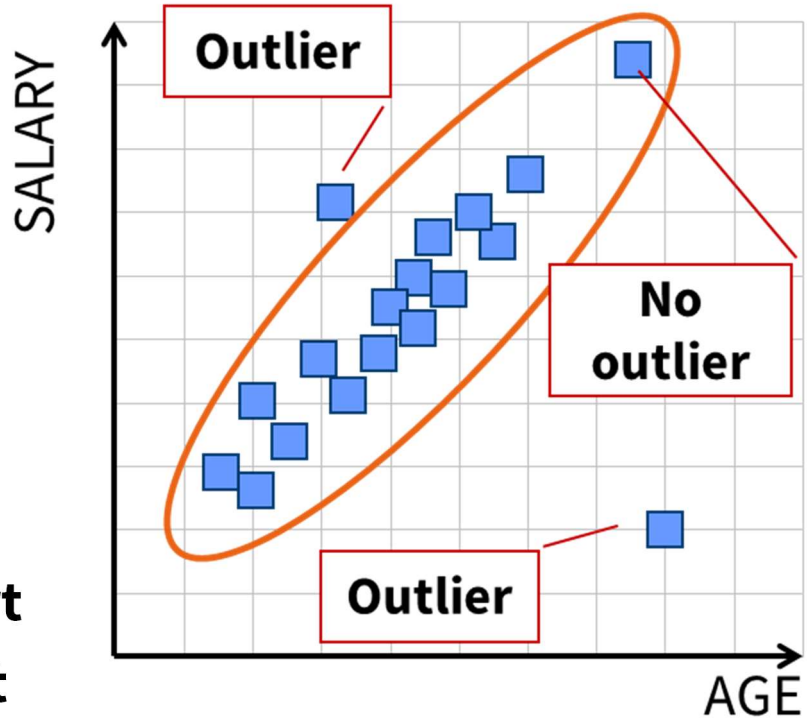
$$z_i = \frac{x_i - \mu}{\sigma}$$

Data Cleaning Strategies
Multivariate outliers

- **Previous approaches focus at one variable at a time**
  - Better to consider the data set as a whole
  - Multivariate perspective
- **Mahalanobis distance**

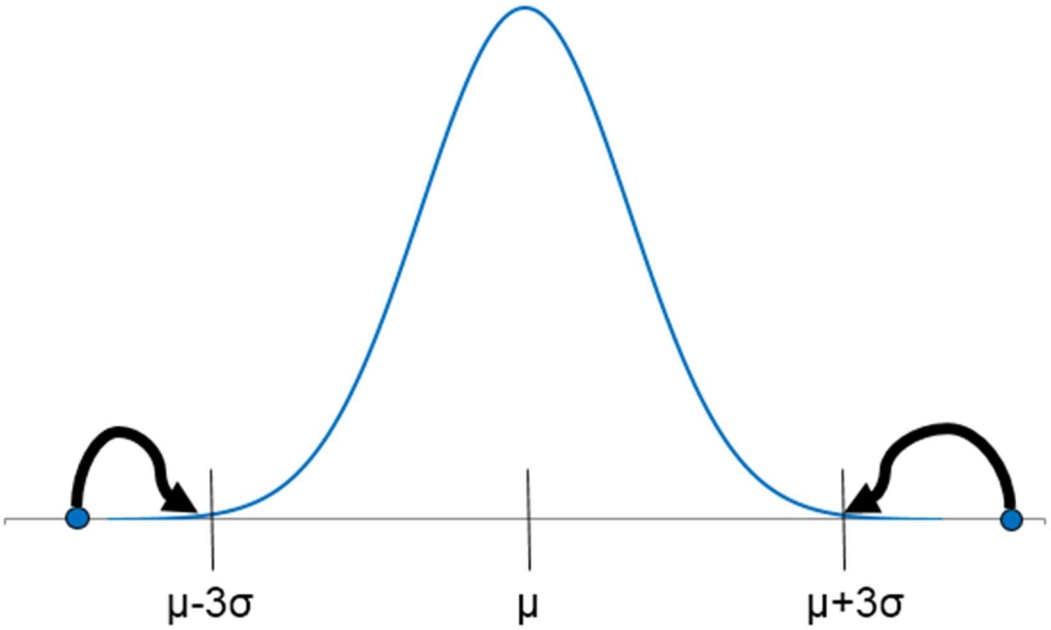
$$D^2 = (\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

  - μ is the vector of means
  - Σ is the covariance matrix
  - Different notion of distance compared to (standard) Euclidian distance
- **To identify outliers, calculate D^2 for every x_i and sort**
- **K-Means and other clustering methods also support the identification of multivariate outliers**



Data Cleaning Strategies
Outlier treatment

- **Invalid outliers**
  - E.g., age=300 years
  - Treat as missing value (keep, delete, replace)
- **Valid outliers**
  - Keep as is
  - Truncation based on z-scores:
    - Replace values having z > 3 by m + 3 s
    - Replace values having z < -3 by m - 3 s
  - Truncation based on IQR
    - More robust than z-scores
    - Winsorizing
  - Other forms of truncation (e.g., using sigmoid)

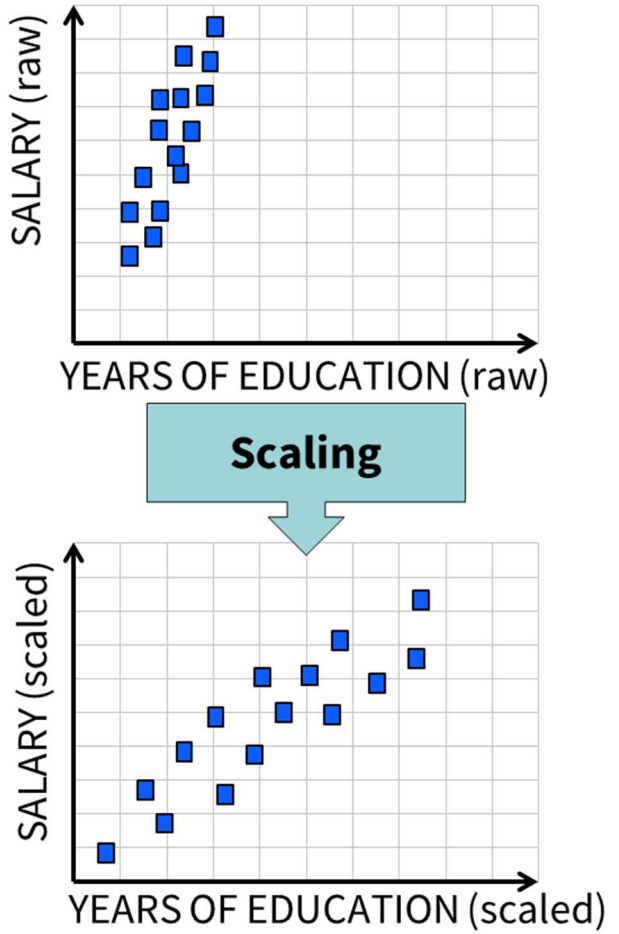


Preprocessing of Continuous Variables

Scaling to ensure comparable values ranges across variables

- **Motivation**
  - Many statistical methods calculate distances
  - Contribution of one variable depends on its variability relative to other variables
- **Different value ranges across variables**
  - Distort distance computations
  - One variable may dominate another
  - Adversely affect statistical method
- **Scaling approaches**
  - Z-transformation (see outlier identification above)
  - Min/max scaling

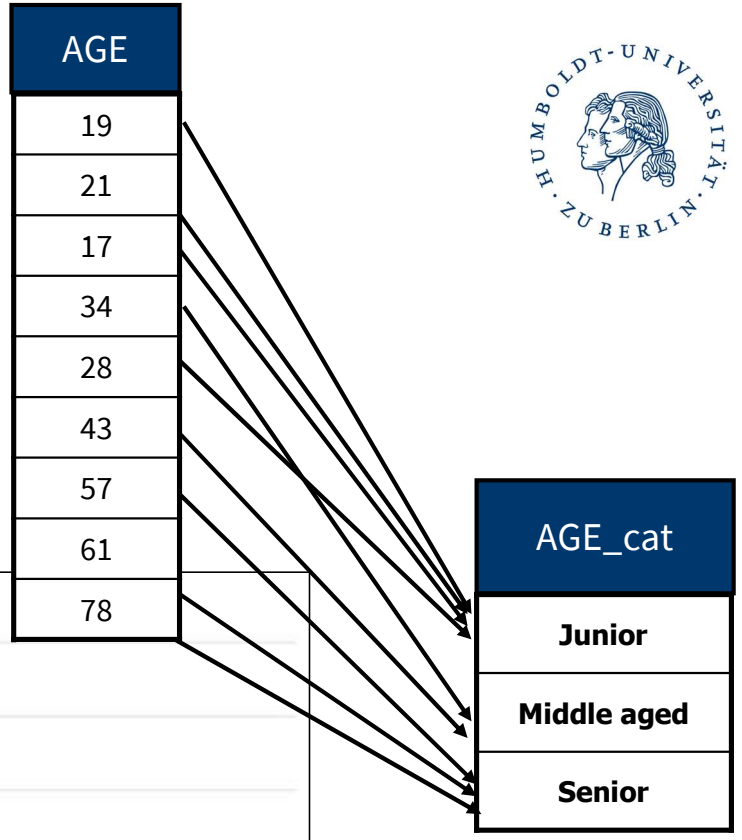
$$x_n = \frac{x_o - \min(x_o)}{max(x_o) - \min(x_o)} \cdot (max_n - min_n) + min_n$$

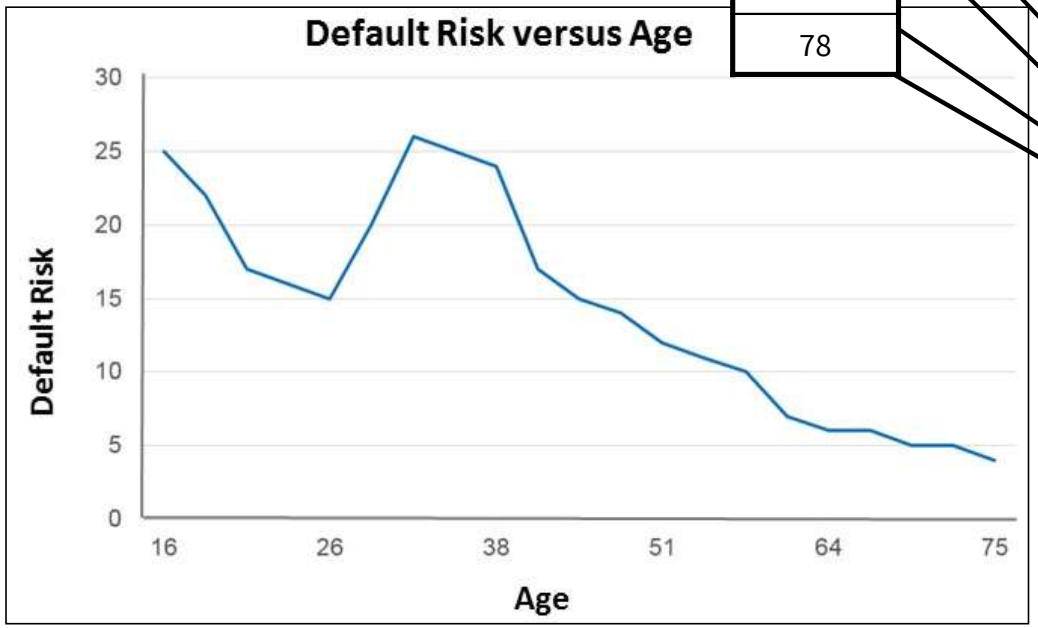


Preprocessing of Continuous Variables

Discretization (aka binning, coarse classification)

- **Motivation**
  - Avoid negative impact of outliers
  - Increase comprehensibility
  - Capture non-linear effects
- **Disadvantage**
  - Loss of information
  - Additional pre-processing for handling the new categorical feature

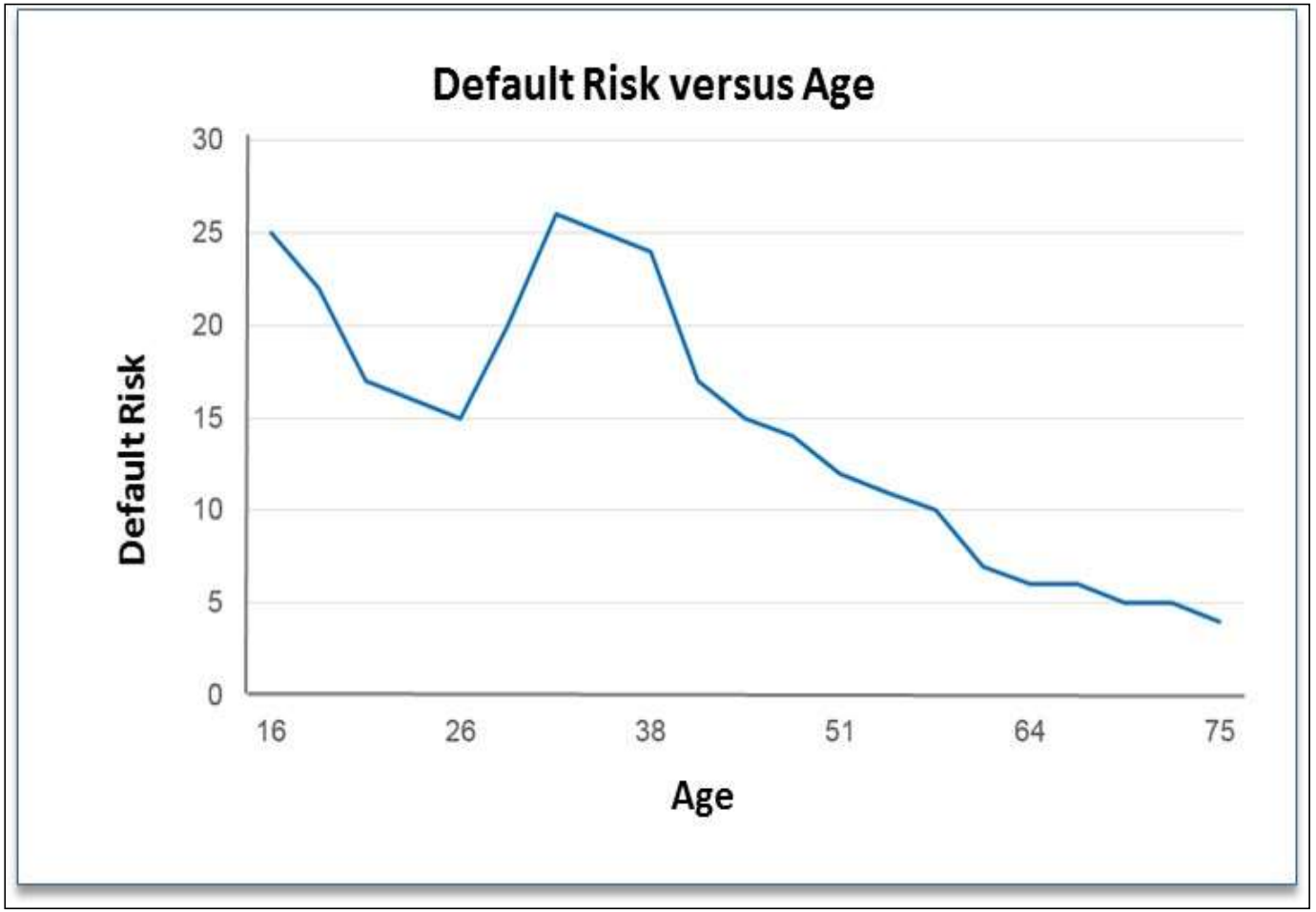




[Thomas et al., 2000]

Preprocessing of Continuous Variables

Discretization to capture non-linear effects in a linear model



[Thomas 2000]

Preprocessing of Continuous Variables

Unsupervised discretization

- **Unsupervised approaches**
  - Equal interval binning
  - Equal frequency binning (histogram equalization)
- **Example: variable SALARY**
  - Analyst decides on bin width / no. of bins
  - Equal interval binning with bin width = 500
    - Bin 1, [1000, 1500[: 1000, 1200, 1300, 1400
    - Bin 2, [1500, 2000[: 1800, 2000
  - Equal frequency binning with two bins
    - Bin 1: 1000, 1200, 1300
    - Bin 2: 1400, 1800, 2000

SALARY
1000
1200
1300
2000
1800
1400

Preprocessing of Continuous Variables

Supervised discretization

- **Use target variable to discretize**
  - Form groups to maximize predictive value
  - Often done via tree-based algorithms
- **Splitting criteria for tree growing**

$$IG(N) = I(N) - p_{N_1} I(N_1) - p_{N_2} I(N_2)$$

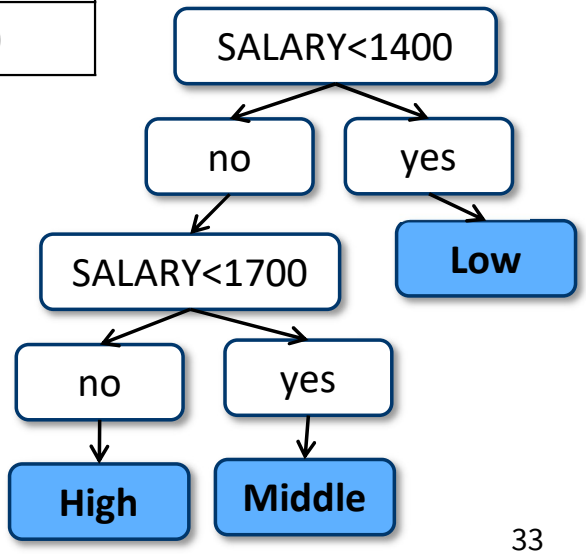
with

$$I_{entropy}(N) = -\sum_j \left(p(c_j|N) \cdot \log_2 \left(p(c_j|N)\right)\right)$$

$$I_{Gini}(N) = 1 - \sum_j p(y_j|N)^2$$

- **Control group number through meta-parameter tree depth**

GOOD/ BAD	SALARY
No	1000
Yes	1200
Yes	1300
No	2000
No	1800
No	1400

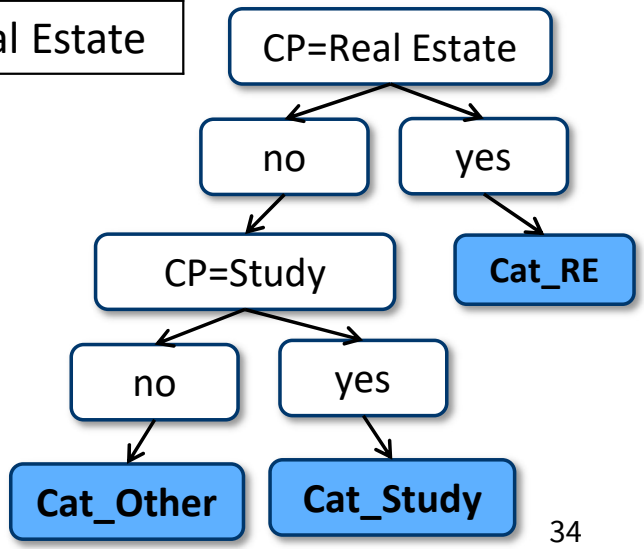


Preprocessing of Continuous Variables

Supervised discretization (cont.)

- **Note how trees also facilitate re-grouping a categorical variable**
  - CREDIT PURPOSE (CP) has four levels in the original data set
    - Real estate (RE), car, holiday, and study
  - After performing two splits using a tree, the number of levels reduces to three
    - Cat_RE, Cat_Study, and Cat_Other
    - Where prefix Cat_ simply stands for category
  - Such regrouping is also called coarse classification (see Appendix for example using Chi^2 method)
- **Reducing number of category levels is often useful**

GOOD/ BAD	CREDIT PURPOSE
No	Real Estate
Yes	Car
Yes	Holiday
No	Study
No	Real Estate
No	Real Estate



Preprocessing of Categorical Variables

Category encoding

- **Credit scoring example**
  - Variable (credit) PURPOSE
  - How to incorporate into empirical model?
- **Code as number**
  - Car=1, house=2, travel=3, study=4
  - $y = \beta_0 + \beta_1 \mathrm{AGE} + \beta_2 \mathrm{PURPOSE}$
- **Problems?**

ID	G/B	AGE	PURPOSE
C1	G	44	car
C2	B	29	house
C3	B	58	travel
C4	G	26	car
C5	G	30	study
C6	G	32	house
…	…	…	…

Preprocessing of Categorical Variables

Category encoding

- **Credit scoring example**
  - Variable (credit) PURPOSE
  - How to incorporate into empirical model?
- **Code as number**
  - Car=1, house=2, travel=3, study=4
  - $y = \beta_0 + \beta_1 \mathrm{AGE} + \beta_2 \mathrm{PURPOSE}$
- **Problems?**
  - Introduces artificial ordering
  - Adversely affects learning
  - Distance argument

    Keeping everything else constant, applicants who apply for a car loan are more similar to applicants who apply for a mortgage than to applicants who apply for study financing.
  - **Never code a nominal variable as a number**

ID	G/B	AGE	PURPOSE
C1	G	44	car
C2	B	29	house
C3	B	58	travel
C4	G	26	car
C5	G	30	study
C6	G	32	house
…	…	…	…

	Car	House	Travel	Study
Car	**0**	**1**	**4**	**9**
House	**1**	**0**	**1**	**4**
Travel	**4**	**1**	**0**	**1**
Study	**9**	**4**	**1**	**0**

Pairwise Euclidian distances after numbering

Preprocessing of Categorical Variables

Category encoding using dummy variables

- **Replace variable with N-1 binary (dummy) variables**
  - N = level of categories
  - N-1 to avoid linear dependency
- **Regression** $y = \beta_0 + \beta_1 \mathrm{AGE} + \beta_2 \mathrm{P_{CAR}} + \beta_3 \mathrm{P_{HOUSE}} + \beta_4 \mathrm{P_{TRAVEL}}$
- **Problems**

Original variable	New dummy variables			Reference level: study
	P_{CAR}	P_{HOUSE}	P_{TRAVEL}	
PURPOSE=car	1	0	0	0
PURPOSE=house	0	1	0	0
PURPOSE=travel	0	0	1	0
PURPOSE=study	0	0	0	1

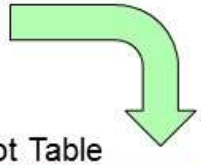
Preprocessing of Categorical Variables

Practical recommendation: reduce the number of category levels

- **Feasibility of dummy coding**
- **Also benefits WOE coding**
- **Heuristic approach using pivot table method**
  - Create a pivot table of the category vs. target (e.g., G/B status) and compute the odds.
  - Group variable values having similar odds ratio
- **Further information**
  - See appendix for more formal approach using the Chi^2 statistic
  - See chapter on feature engineering for WOE-coding

Data Table

Customer ID	Age	Purpose	...	Bad/Good
C1	44	car	...	Good
C2	20	cash	...	Good
C3	58	travel	...	Bad
C4	26	car	...	Good
C5	30	study	...	Bad
C6	32	house	...	Good
...	...	...	...	...

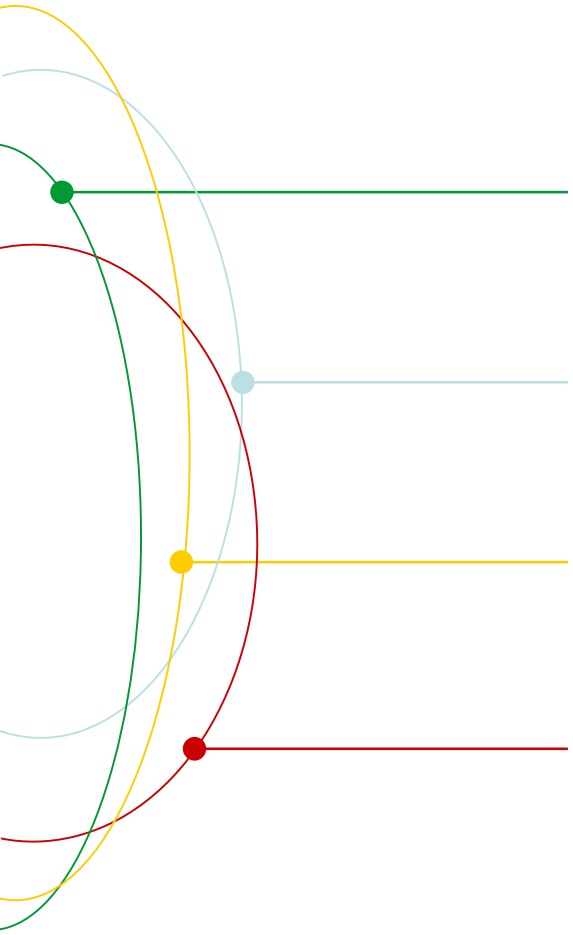


Pivot Table

	Car	Cash	Travel	Study	House	...
Good	1000	2000	3000	100	5000	...
Bad	500	100	200	80	800	...
Odds	2	20	15	1,25	6,25	...

Summary

Summary



Learning goals

- Scope and need of data preparation
- Selected preprocessing activities

Findings

- Visualizations for uni- and multivariate EDA
- Missing value replacement & imputation
- Outlier identification and treatment
- Scaling and discretization of numeric variables
- Dummy coding of discrete variables

What next

- Evaluation of predictive models
- Performance indicators & experimental designs

Thank you for your attention!

Stefan Lessmann

Chair of Information Systems
School of Business and Economics
Humboldt-University of Berlin, Germany

Tel. +49.30.2093.5742
Fax. +49.30.2093.5741

stefan.lessmann@hu-berlin.de
http://bit.ly/hu-wi

www.hu-berlin.de

Photo: Heike Zappe

Appendix

Chi^2 approach toward coarse classification

Chi^2 approach toward coarse classification

- The term coarse classification refers to the discretization of a continuous variable or, alternatively, to the task of re-grouping a variable that is already a category
- The following example considers the latter case. It starts from a categorical variable, which captures information on the housing conditions of credit applicants, and ask the questions how the number of category levels can be reduced
- Reducing the levels of a categorical variable can, for example, be useful in regression modeling when using dummy codes. Fewer category levels mean less (new) dummy variables
- The point of the example is to demonstrate the Chi2 approach toward coarse classification
- After working through the example, think about the similarities between the Chi2 method and a decision tree
- A popular tree growing algorithm, CHAID, actually operates on the basis of the Chi2 method.
- Once you see the connections between the Chi2 method and tree-based algorithms, you can immediately generalize the tree-based discretization example in the main part of the lecture to the task of coarse classification. That is, you should understand that decision trees can also be useful to re-group categorical variables and/or merge levels of a categorical variable in an informed manner.

Appendix: Coarse Classification

- **Consider the following example (Thomas et al., 2002)**
  - Categorical variable HOUSING
  - How to reduce the number of levels?

Attribute HOUSING	Owner	Rent Unfurnished	Rent Furnished	With parents	Other	No answer	Total
Goods	6000	1600	350	950	90	10	**9000**
Bads	300	400	140	100	50	10	**1000**
G/B odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	**9:1**

Appendix: Coarse Classification

- **Consider the following example (Thomas et al., 2002)**
  - Categorical variable HOUSING
  - How to reduce the number of levels?

Attribute HOUSING	Owner	Rent Unfurnished	Rent Furnished	With parents	Other	No answer	Total
Goods	6000	1600	350	950	90	10	**9000**
Bads	300	400	140	100	50	10	**1000**
G/B odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	**9:1**

- **Suppose we want three levels. Which option is better?**
  - Option 1: “owners”, “renters”, and “others”
  - Option 2: “owners”, “with parents”, and “others”

Appendix: Coarse Classification

The Chi² method

- **Assume housing does not affect class membership**
  - Statistical independence of G/B status and HOUSING
  - Independence frequencies per housing type for G and B should be the same as in the population
  - Example
    - Owner & Good
    - 6300 * 9000/10000 = 5670
- **Chi-square distance**
  - Sum of squared cell-wise difference between the tables
  - Large values cast doubt on independence assumption

Empirical frequencies option 1:

HOUSING	Owner	Renters	Others	Total
Goods	6000	1950	1050	9000
Bads	300	540	160	1000
Total	6300	2490	1210	10000

Independence frequencies option 1:

HOUSING	Owner	Renters	Others	Total
Goods	5670	2241	1089	9000
Bads	630	249	121	1000
Total	6300	2490	1210	10000

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(1950-2241)^2}{2241} + \frac{(540-249)^2}{249} + \frac{(1050-1089)^2}{1089} + \frac{(160-121)^2}{121} = 583$$

Appendix: Coarse Classification

The Chi² method (cont.)

- **Empirical frequencies for option 2 (check for yourself):**

$$\chi^2 = \frac{(6000-5670)^2}{5670} + \frac{(300-630)^2}{630} + \frac{(950-945)^2}{945} + \frac{(100-105)^2}{105} + \frac{(2050-2385)^2}{2385} + \frac{(600-265)^2}{265} = 662$$

- **The higher the test statistic, the better the split**
  - Formally, compare with chi-square distribution with k-1 degrees of freedom for *k* classes of the characteristic
  - Not needed to answer the focal question
- **Since $\chi^2_{\text{Option2}} > \chi^2_{\text{Option1}}$ option 2 gives the better split**
  - Three categories should be “owners”, “with parents”, “others”
  - Stronger relationship with Good/Bad status