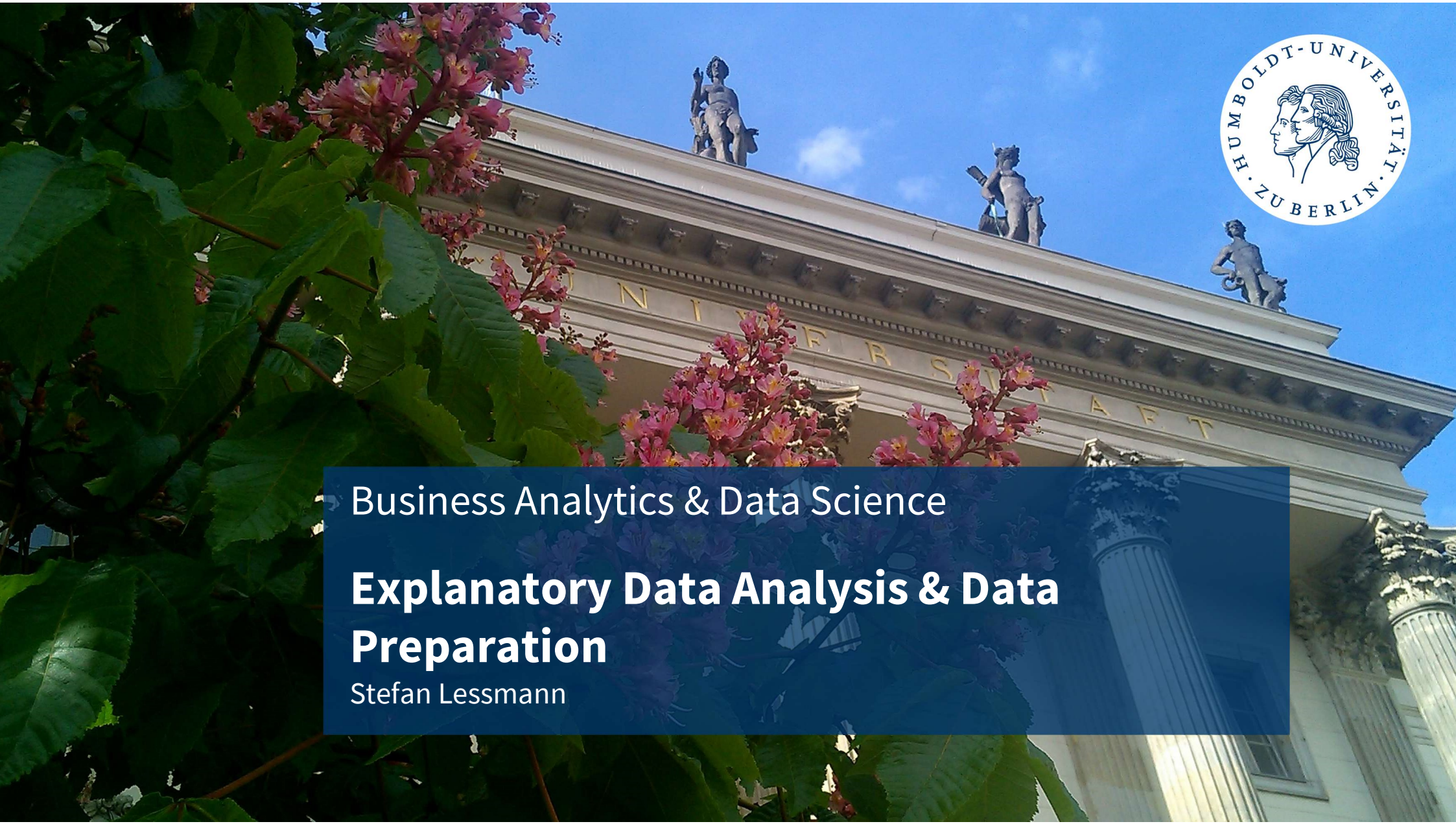


Business Analytics & Data Science

Explanatory Data Analysis & Data Preparation

Stefan Lessmann

Agenda

Introduction
Preliminaries, recap, and types of data

Explanatory Data Analysis
Scope, motivation, popular approaches

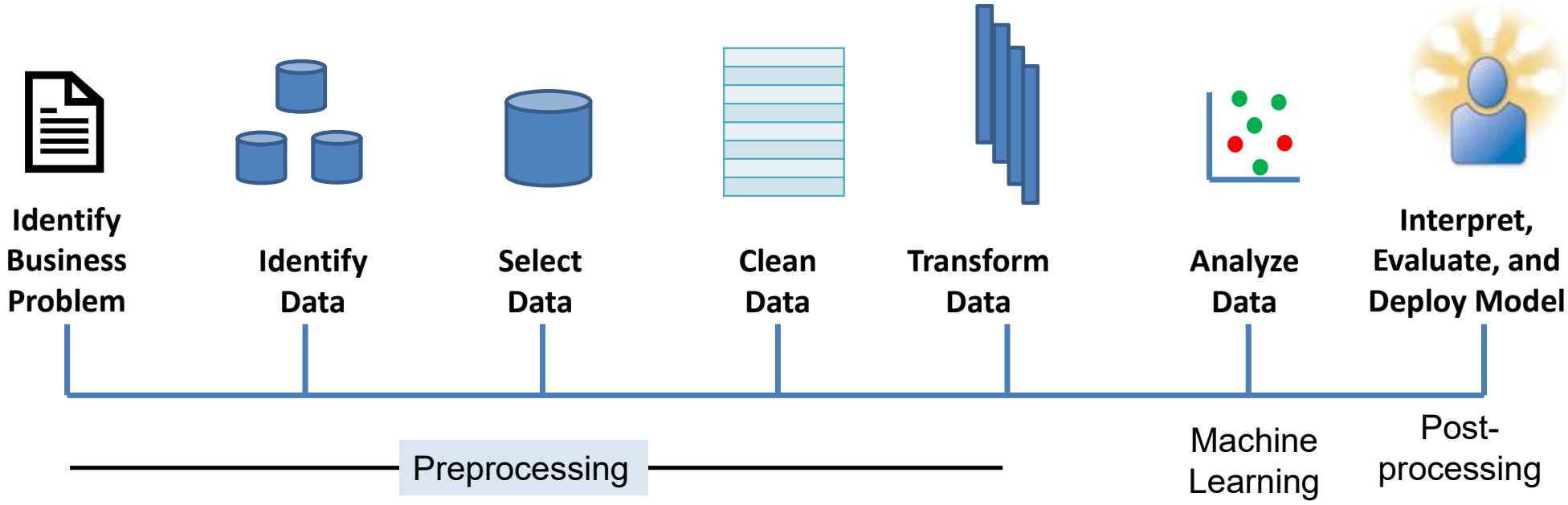
Data preparation
Motivation & need, process, cleaning & preparation strategies for continuous and categorical variables

Summary

Introduction

Preliminaries, recap, and types of data

Recap: Data Analytics Process Model



Recap: Machine Learning Lingo

Some terminology and ML jargon

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

Target, outcome, label, response (variable), dependent (variable)

Observations, cases, examples, data items, subjects $\mathcal{D} = \{Y_i, \boldsymbol{X}_i\}_{i=1}^{n}$

BUREAU SCORE	COLLATERAL	DEBT/ INCOME	YEARS AT ADDRESS	AGE	…	**DEFAULT (E.G., 90 DAYS LATE)**
650	Yes	20%	2	<21	…	No
280	No	43%	0	21-29	…	Yes
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	…	No
580	No	18%	6	40-50	…	Yes
410	Yes	29%	4	21-29	…	No
800	Yes	14%	10	40-50	…	Yes

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_m) \in \mathbb{R}^m$$

$$Y \in \{0,1\}$$

Structured Tabular Data
Types of Variables

- **Continuous**
  - Synonyms: numeric(al)/real/metric variables
  - Domain: whole and fractional numbers
- **Categorical**
  - Synonyms: discrete, categories, non-numeric
  - 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	DateTime	…

- **EDA and preprocessing operations vary across different types of variables**

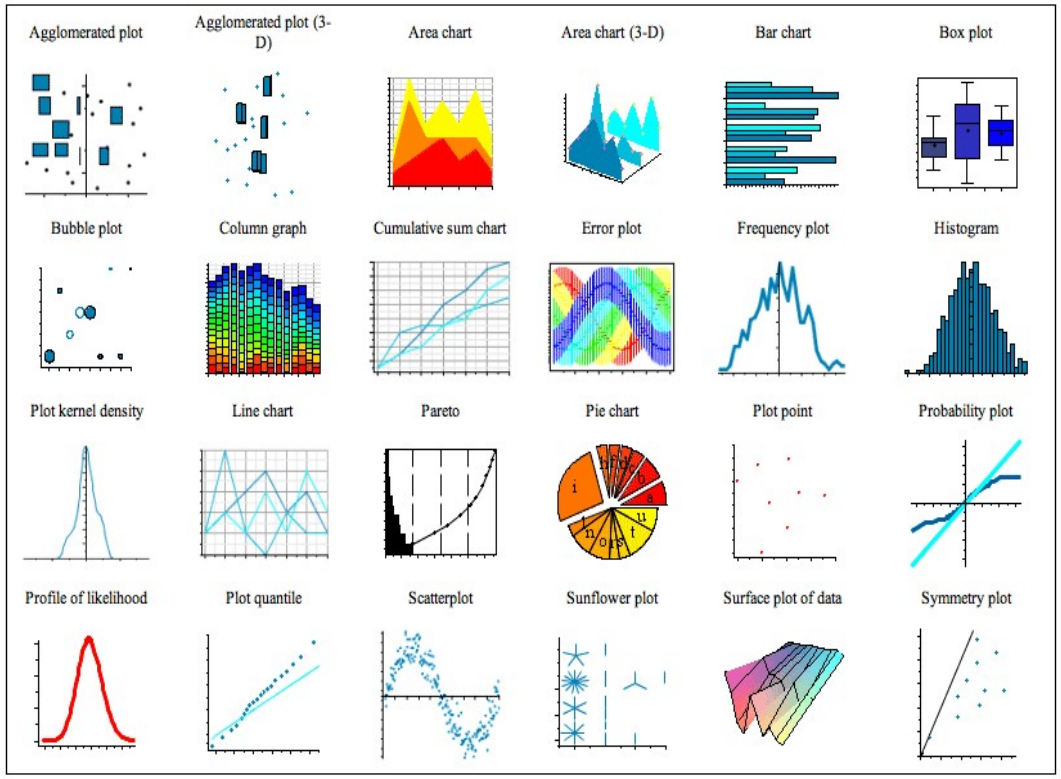
Explanatory Data Analysis

Scope, motivation, popular approaches

Explanatory Data Analysis (EDA)

"A first look at the data"

- **Motivation and use cases**
  - Inform analyst about the structure of a data set
  - Hint at patterns that warrant further investigation
  - Reveal data quality problems
  - Suggest suitable data preparation operations
- **Approaches**
  - Descriptive statistics
  - Extensive use of charts to explore data
  - Somewhat related to the concept of visual analytics (VA), which emphasizes sophisticated and possibly interactive means of visualization

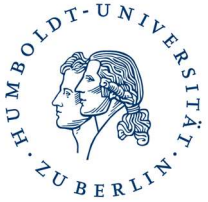


Given a Tabular Data Set, What Methods Would You Consider for Taking „A First Look at the Data “?

BUREAU SCORE	COLLATERAL	DEBT/INCOME	YEARS AT ADDRESS	AGE	…	**BAD RISK** (e.g., 90 days late)
650	Yes	20%	2	<21	…	**No (0)**
280	No	43%	0	21-29	…	**Yes (1)**
750	Yes	27%	8	30-39	…	**No (0)**
600	Yes	18%	4	40-50	…	**No (0)**
575	No	33%	12	>50	…	**No (0)**
715	Yes	24%	1	21-29	…	**No (0)**
580	No	18%	6	40-50	…	**Yes (1)**
410	Yes	29%	4	21-29	…	**No (0)**
800	Yes	14%	10	40-50	…	**Yes (1)**

Explanatory Data Analysis Landscape

Descriptive statistics and visualizations



- **Goes back to Tukey'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		

Descriptive Statistics

	LOAN	MORTDUE	VALUE	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
count	5960.000000	5442.000000	5848.000000	5445.000000	5252.000000	5380.000000	5652.000000	5450.000000	5738.000000	4693.000000
mean	18607.969799	73760.817200	101776.048741	8.922268	0.254570	0.449442	179.766275	1.186055	21.296096	33.779915
std	11207.480417	44457.609458	57385.775334	7.573982	0.846047	1.127266	85.810092	1.728675	10.138933	8.601746
min	1100.000000	2063.000000	8000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.524499
25%	11100.000000	46276.000000	66075.500000	3.000000	0.000000	0.000000	115.116702	0.000000	15.000000	29.140031
50%	16300.000000	65019.000000	89235.500000	7.000000	0.000000	0.000000	173.466667	1.000000	20.000000	34.818262
75%	23300.000000	91488.000000	119824.250000	13.000000	0.000000	0.000000	231.562278	2.000000	26.000000	39.003141
max	89900.000000	399550.000000	855909.000000	41.000000	10.000000	15.000000	1168.233561	17.000000	71.000000	203.312149

```
hmeq.describe()
```

Cross-Tables

Compute summary statistics across levels of the discrete target

```
features = ["LOAN", "DEBTINC"]
print(hmeq.groupby("BAD")[features].mean())
print(hmeq.groupby("BAD")[features].median())
print(hmeq.groupby("BAD")[features].quantile(q=0.95))
```

```
              LOAN    DEBTINC
BAD
False  19028.107315  33.253129
True   16922.119428  39.387645
         LOAN    DEBTINC
BAD
False  16900.0  34.541671
True   14900.0  38.079762
         LOAN    DEBTINC
BAD
False  39550.0  42.135485
True   41080.0  62.777490
```

Count occurences of category levels across levels of a discrete target

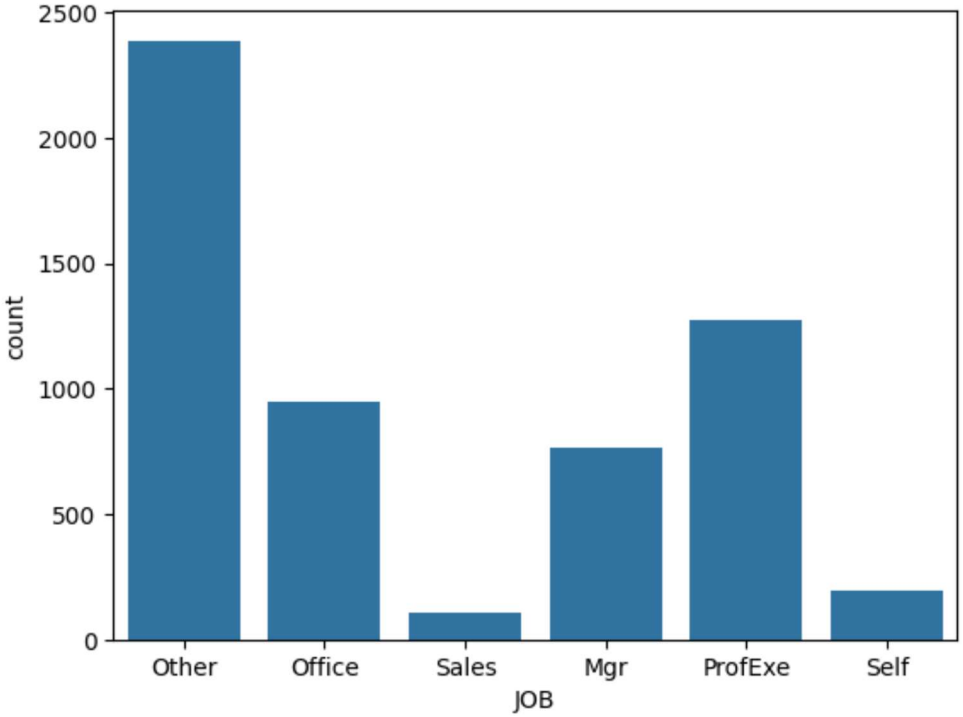
```
pd.crosstab(hmeq.BAD, hmeq.JOB, normalize=True)
```

JOB	Mgr	Office	Other	ProfExe	Sales	Self
BAD						
False	0.103503	0.144869	0.322830	0.187291	0.012498	0.023763
True	0.031509	0.022003	0.097518	0.037317	0.006689	0.010209

The Countplot

Visualizing the empirical distribution of a categorical feature

- **HMEQ credit risk data set**
- **Feature: JOB**
  - Client profession
  - Categorical feature with six levels
- **Count the number of observations with a given level (aka feature value)**
  - Can group by another feature
  - Illustrated for binary target BAD
- **Findings**
  - Truncated at zero
  - Roughly normal in a certain range
  - Long tail (right skewed distribution)



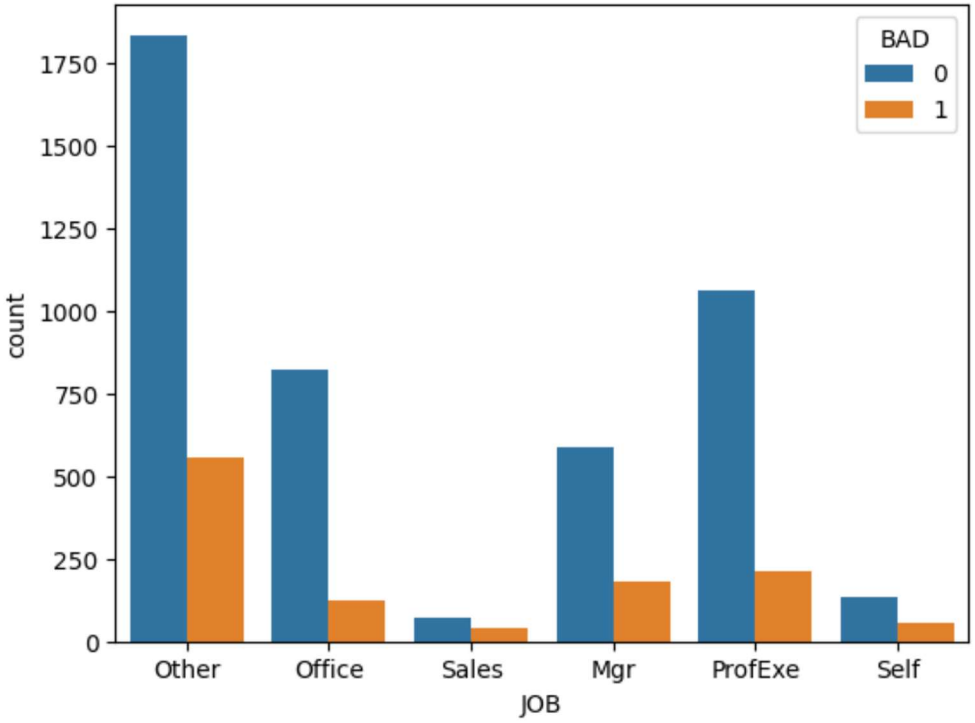

```
sns.countplot(x=hmeq.JOB)
plt.show()
```

The Countplot

Visualizing the empirical distribution of a categorical feature

Source Code

- **HMEQ credit risk data set**
- **Feature: JOB**
  - Client profession
  - Categorical feature with six levels
- **Count the number of observations with a given level (aka feature value)**
  - Can group by another feature
  - Illustrated for binary target BAD
- **Findings**
  - Truncated at zero
  - Roughly normal in a certain range
  - Long tail (right skewed distribution)



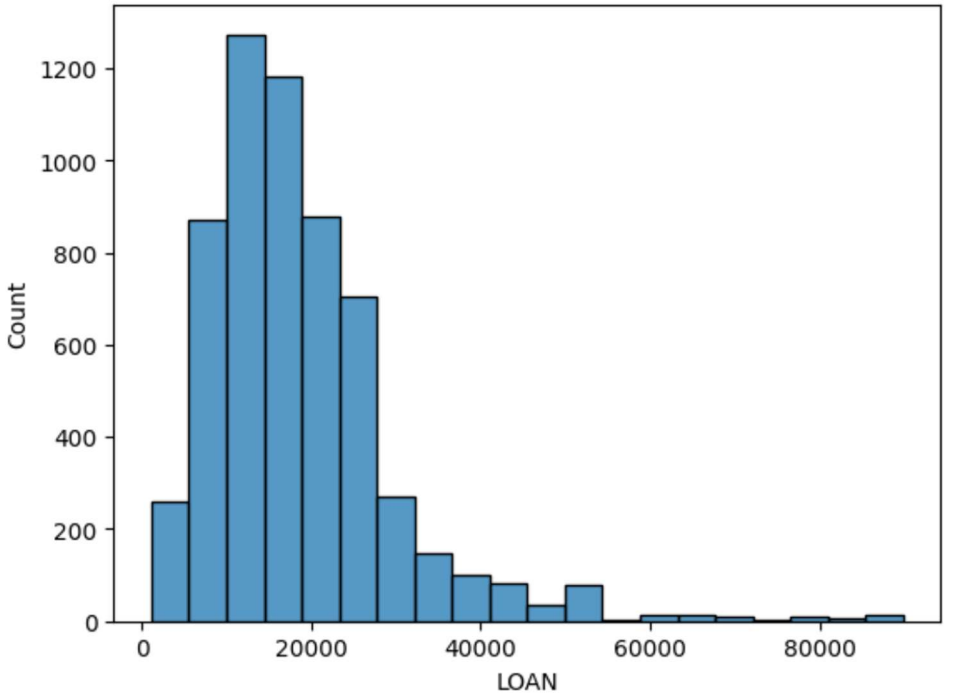

```
sns.countplot(data=hmeq, x='JOB', hue='BAD')
plt.show()
```

The Histogram

Visualizing the empirical distribution of a continuous feature

Source Code

- **HMEQ credit risk data set**
- **Feature: LOAN**
  - Amount of the loan request
  - Numerical feature of type int
- **Categorized into 20 bins**
- **Count observations in each bin**
- **Findings**
  - Truncated at zero
  - Roughly normal in a certain range
  - Long tail (right skewed distribution)



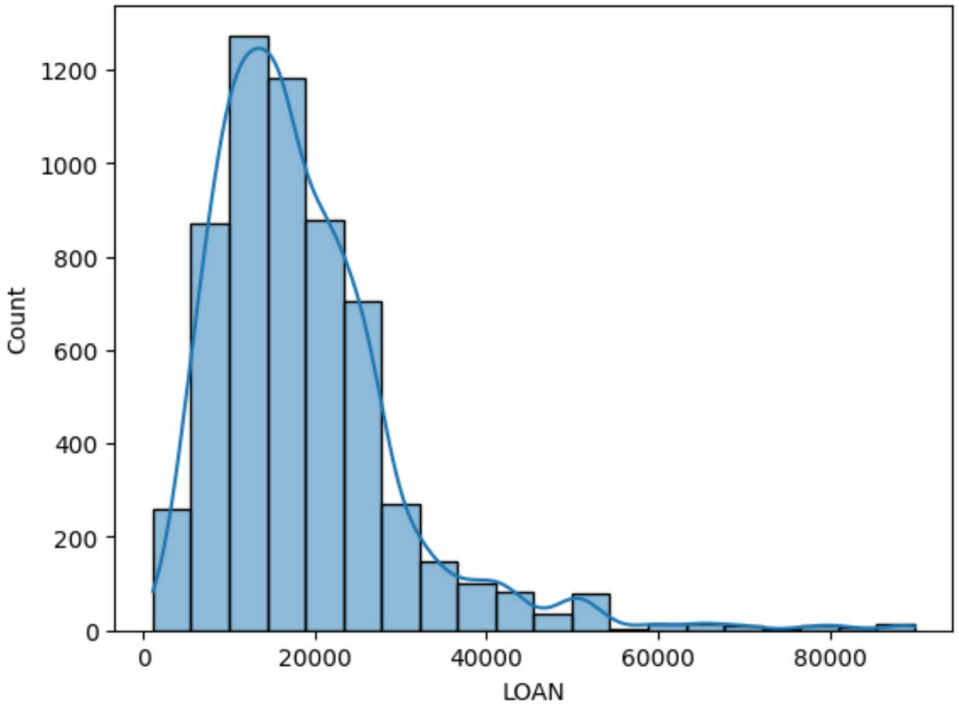

```
sns.histplot(hmeq.LOAN, bins=20)
plt.show()
```

The Histogram

Visualizing the empirical distribution of a continuous feature

Source Code

- **HMEQ credit risk data set**
- **Feature: LOAN**
  - Amount of the loan request
  - Numerical feature of type int
- **Categorized into 20 bins**
- **Count observations in each bin**
- **Add a density estimator**
- **Findings**
  - Truncated at zero
  - Roughly normal in a certain range
  - Long tail (right skewed distribution)

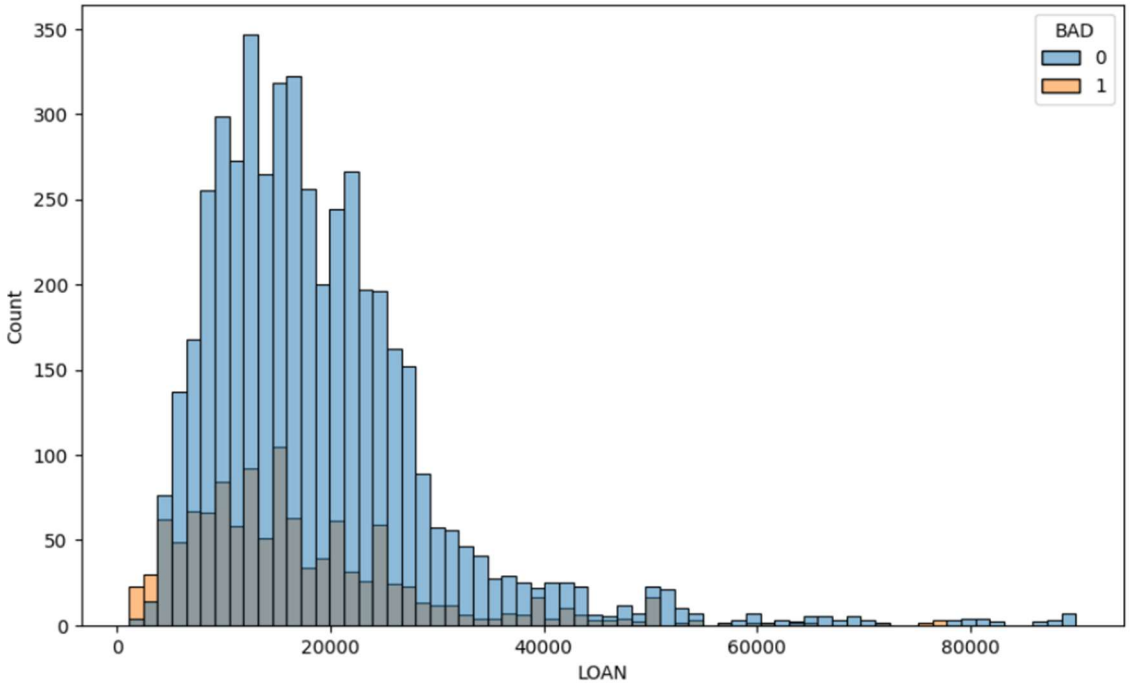



```
sns.histplot(hmeq.LOAN, bins=20, kde=True)
plt.show()
```

The Histogram

Visualizing the empirical distribution of a continuous feature

- **HMEQ credit risk data set**
- **Feature: LOAN**
  - Amount of the loan request
  - Numerical feature of type int
- **Categorized into 20 bins**
- **Count observations in each bin**
- **Group by the target variable**
- **Findings**
  - Check of feature predictiveness
  - Do we find evidence of the distribution of LOAN to differ between good vs. bad risks?

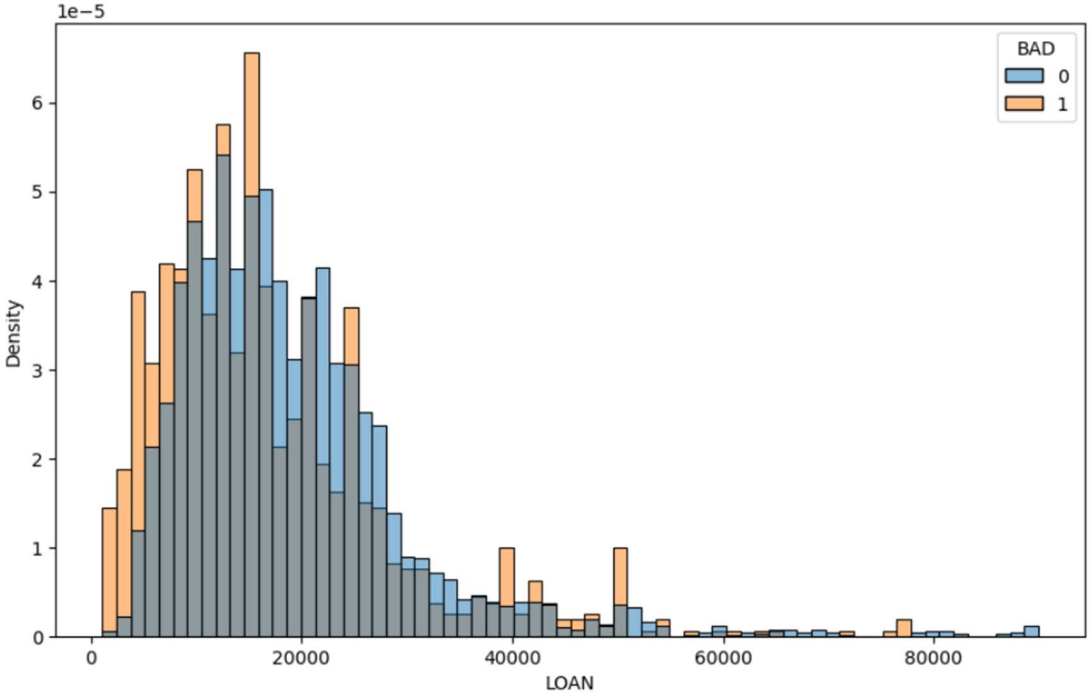



```
plt.figure(figsize=(10,6))
sns.histplot(data=hmeq, x='LOAN', hue='BAD')
plt.show()
```

The Histogram

Visualizing the empirical distribution of a continuous feature

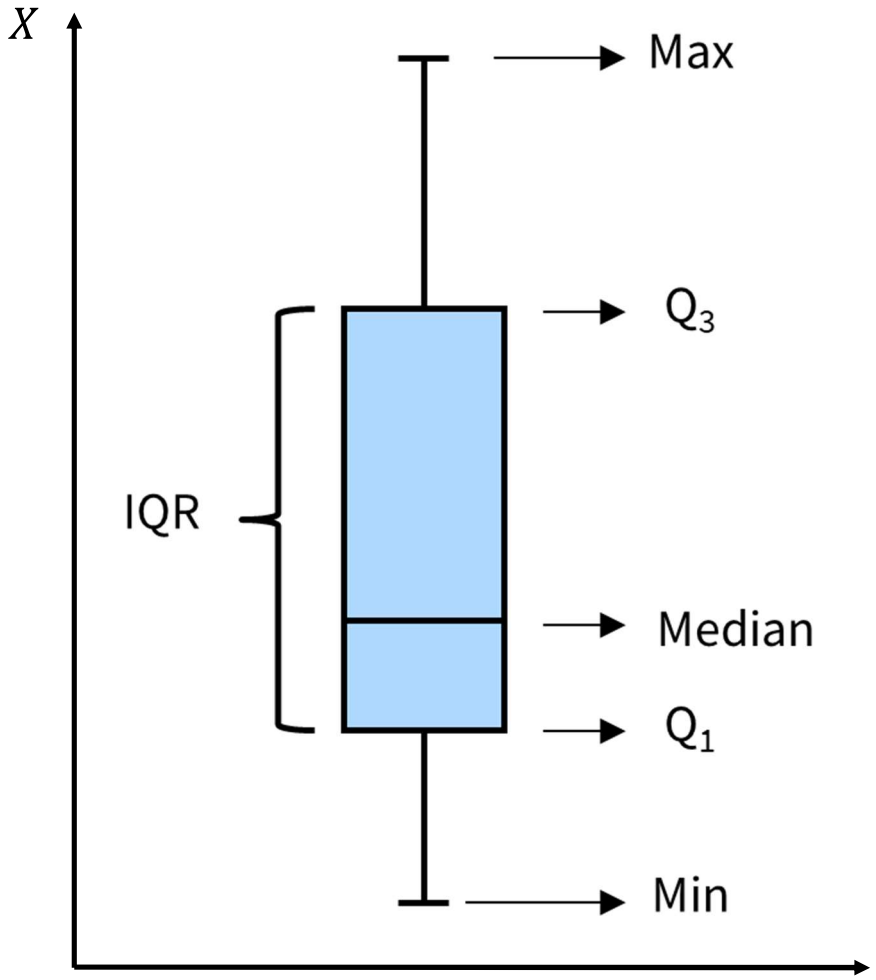
- **HMEQ credit risk data set**
- **Feature: LOAN**
  - Amount of the loan request
  - Numerical feature of type int
- **Categorized into 20 bins**
- **Count observations in each bin**
- **Group by the target variable**
- **Findings**
  - Check of feature predictiveness
  - Do we find evidence of the distribution of LOAN to differ between good vs. bad risks?
  - Easier to answer when normalizing the counts




```
plt.figure(figsize=(10,6))
sns.histplot(data=hmeq, x='LOAN', hue='BAD',
             stat='density', common_norm=False)
plt.show()
```

The Boxplot (aka Box Whisker Plot)

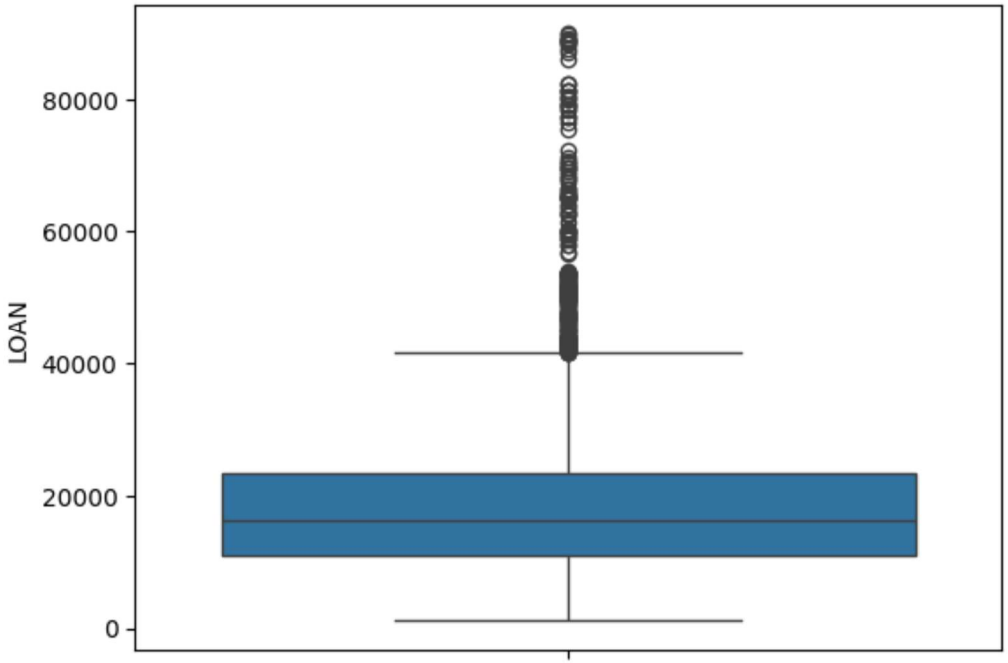
- **Five-number summary of a distribution**
  - Minimum, Median, Maximum
  - First Quartile: $Q_1 \equiv P(X \leq Q_1) = 0.25$
  - Third Quartile $Q_3 \equiv P(X \leq Q_3) = 0.75$
- **Box**
  - Drawn from $Q1$ to $Q3$
  - Line inside the box highlights median
- **Interquartile range (IQR) defined as $Q3 - Q1$**
  - Covers the middle fifty percent of the data
  - Simple, robust measure of spread
- **Whiskers**
  - Reveal information about boundaries of the distribution
  - Option 1: draw whiskers to depict Min & Max
  - Option 2: draw whiskers to depict IRQ



The Boxplot (aka Box Whisker Plot)

Outlier rule of Tukey 1977

- **Definition of outliers based on distance to first or third quartile and inter-quartile range**
  - Upper outliner
  - Lower outlier
- **Whiskers depict IQR boundaries**
  - Calculate threshold as IQR*1.5
  - Start drawing from Q1/Q3
  - Draw whisker down/up to the lowest/highest data point observed within IQR*1.5
- **Outliers**
  - Observed values outside this range
  - Depict these separately

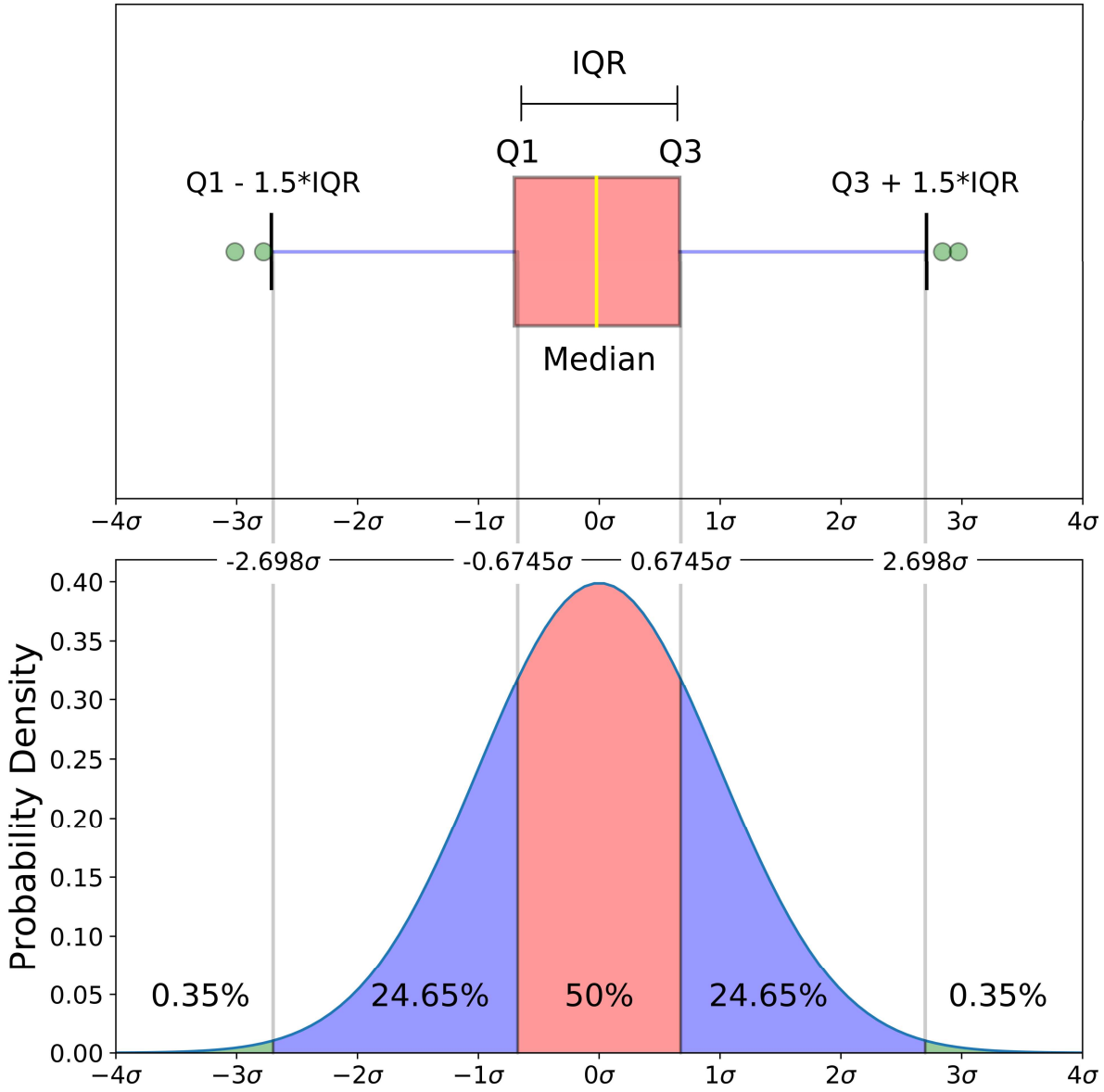



```
sns.boxplot(data=hmeq, y='LOAN')
plt.show()
```

The Boxplot and the Normal Distribution

While remaining a heuristic, aka rule of thumb, the connections between the boxplot and the normal distribution explain, to some extent, the definition of outliers as observations 1.5*IQR above/below Q3/Q1, respectively.

But note that the „boxplot-rule“ does not assume the underlying data to come from a normal distribution.

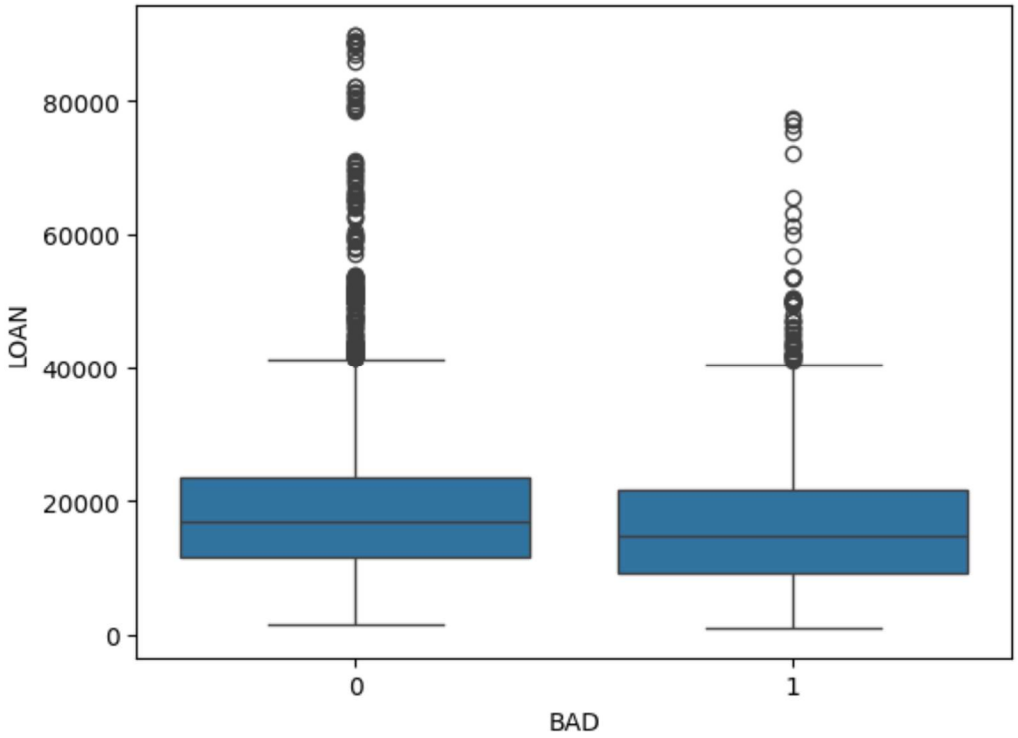


Source: Galarnyk (2018)

The Boxplot (aka Box Whisker Plot)

Examining feature predictiveness

- **Group data points using the target**
- **Compare feature values across summary statistics**
- **Differences discernible?**
  - Medians
  - First and/or third quartile
  - Inter-quartile range
  - Outliers

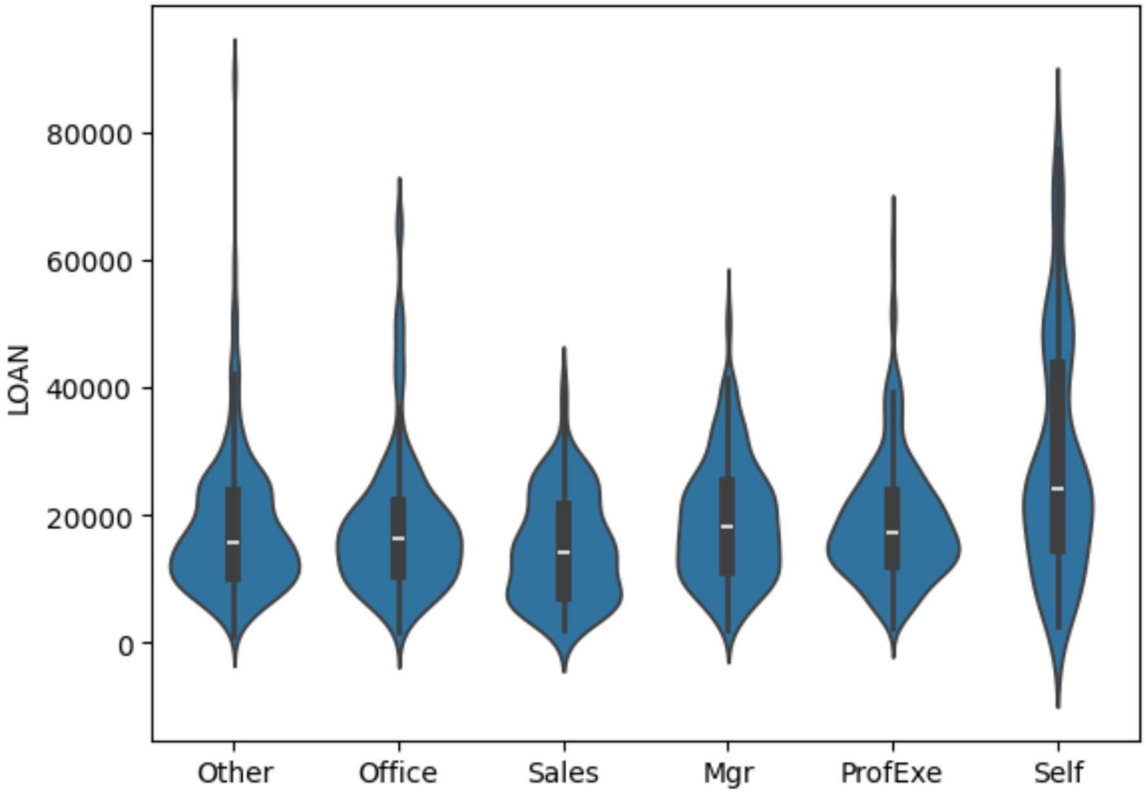



```
sns.boxplot(data=hmeq, y='LOAN', x='BAD')
plt.show()
```

Adding Complexity: The Violin Plot

Source Code

- **Distribution of numerical feature (here LOAN)**
- **Common box plot elements**
  - Median, Q1, Q3, IQR
  - Alternative options
    - Only quartiles, data points, …
    - See argument `inner`
- **Kernel estimator of the density**
  - Typically smoothed
  - Useful for multi-modal distributions
  - Symmetric around the “box plot”
- **Option to group by categorical feature (here JOB)**



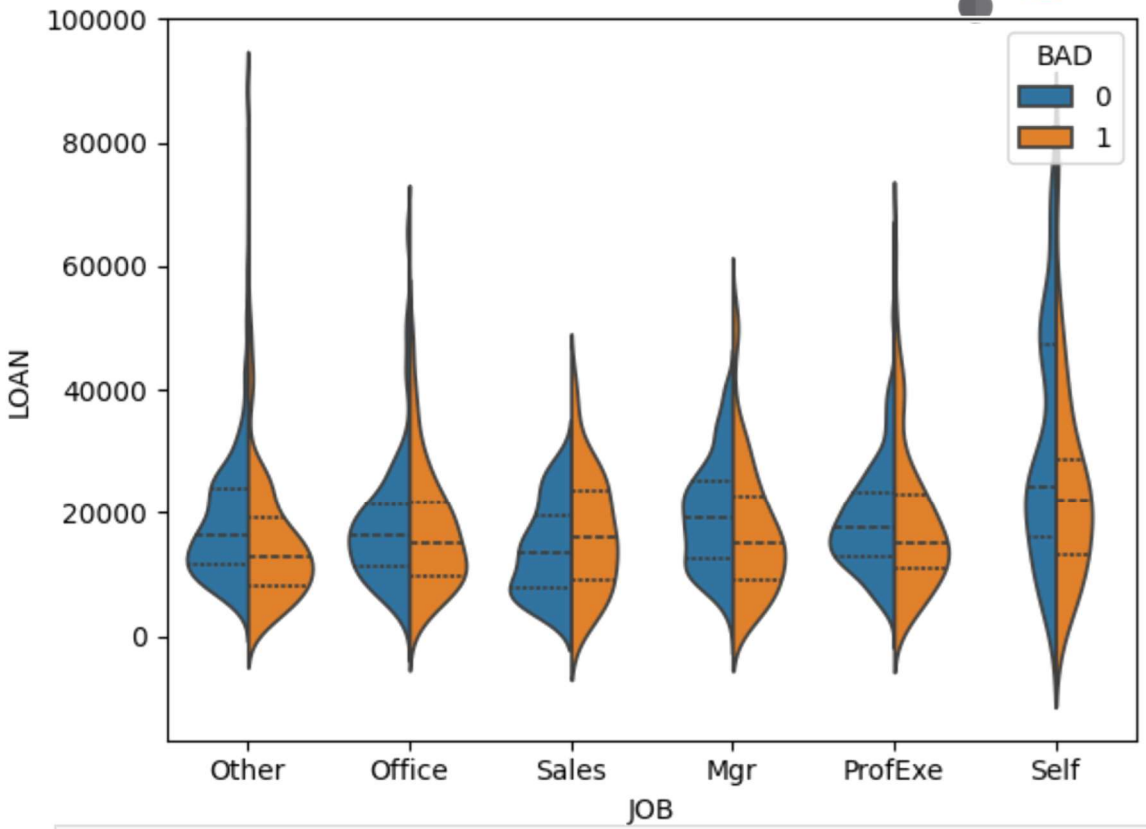

```
sns.violinplot(data=hmeq, x='JOB', y='LOAN',
               hue='BAD', split=True, inner='quart')
plt.show()
```

Adding Complexity: The Violin Plot

Two criteria for grouping

- **Distribution of numerical feature (here LOAN)**
- **Option to group by categorical feature (here JOB)**
- **Auxiliary grouping by, e.g., the target**
- **Densities estimated separately for each level of the target**
  - Reduced representation of boxplot elements
  - Dashed lines represent Q1, Q2, and Q3

Source Code



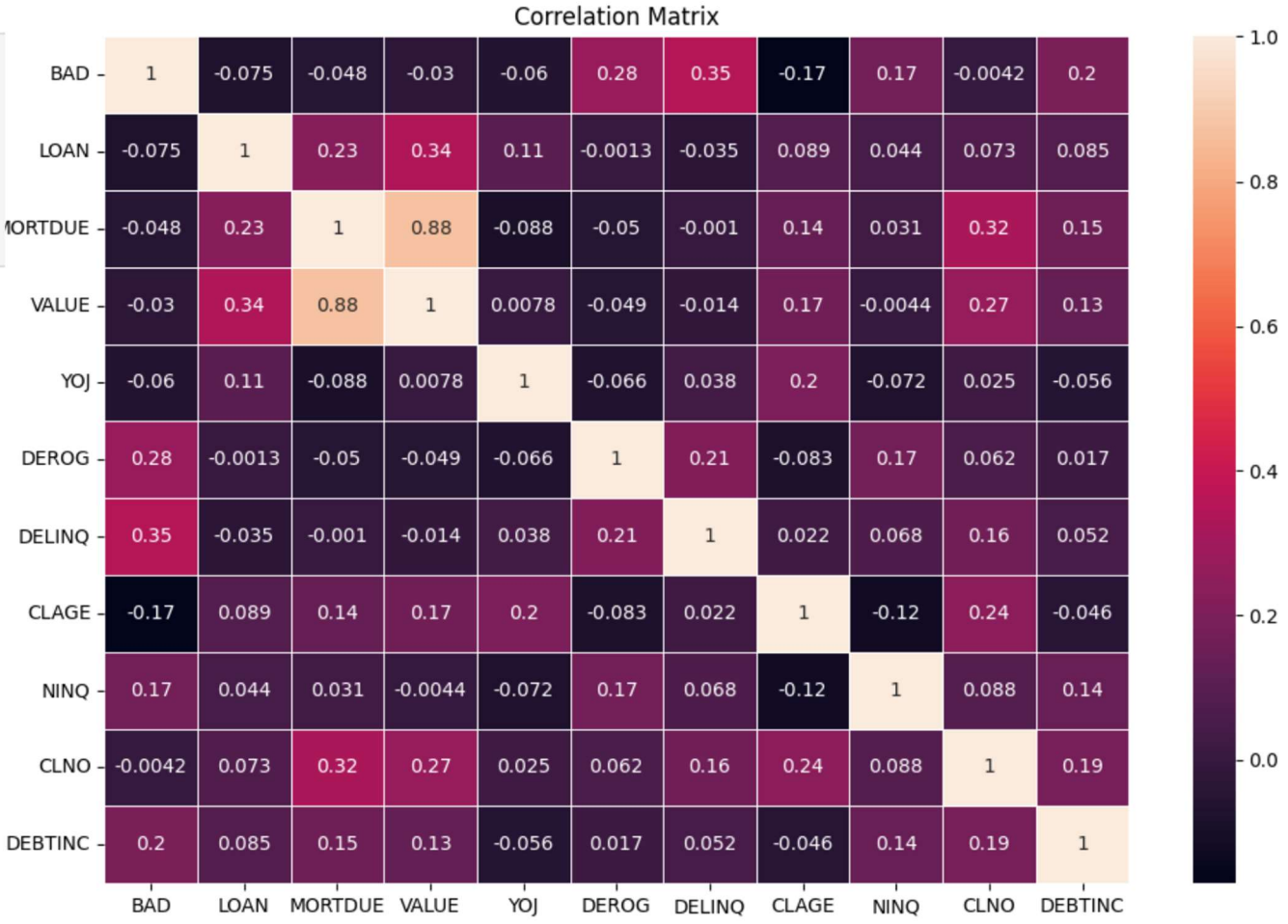

```
sns.violinplot(data=hmeq, x='JOB', y='LOAN',
               hue='BAD', split=True, inner='quart')
plt.show()
```

Correlation Matrix

All pairwise correlations of over a set of numerical features

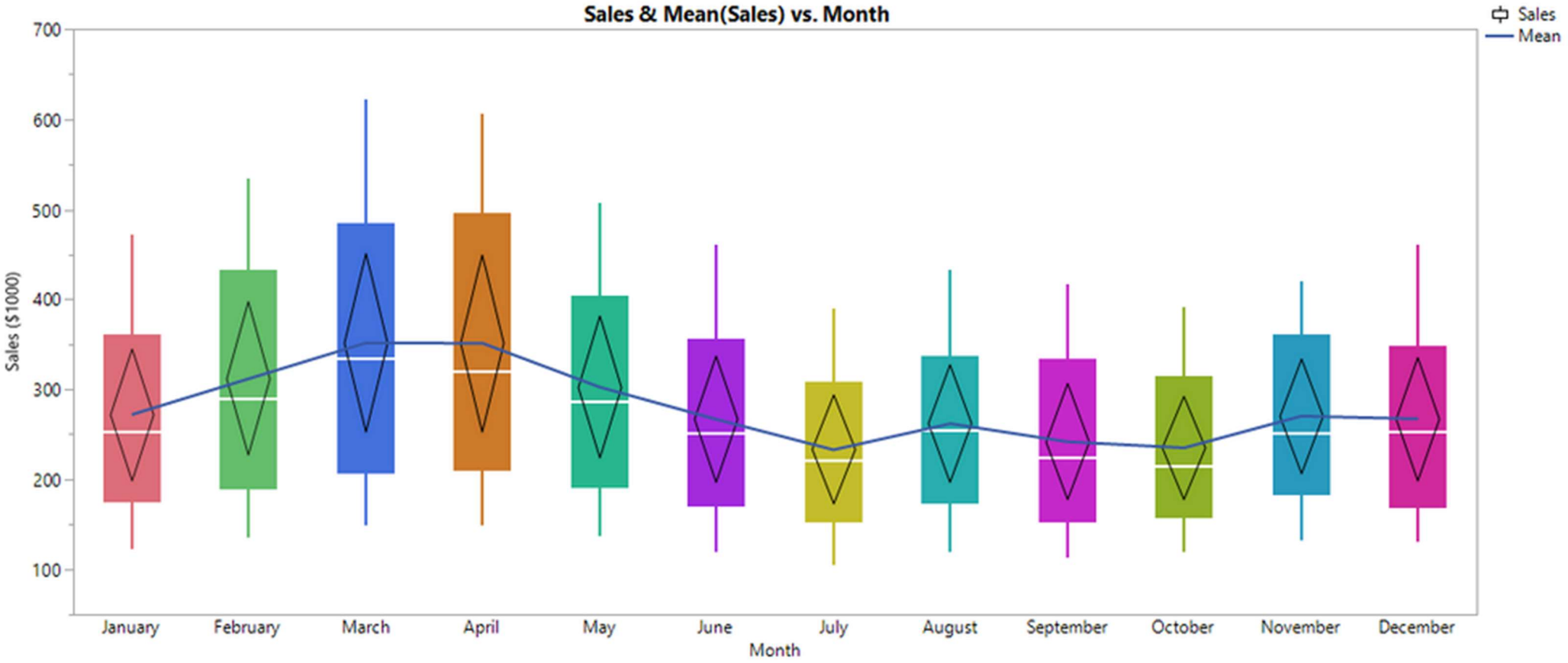
Source Code

```
corr= hmeq.select_dtypes(exclude='O').corr()
f,ax = plt.subplots(figsize=(12, 8))
sns.heatmap(corr, annot=True, linewidth=.5)
plt.title('Correlation Matrix')
plt.show()
```



Adding Complexity: Further Options

Multiple variables or repeated measures of the same variable

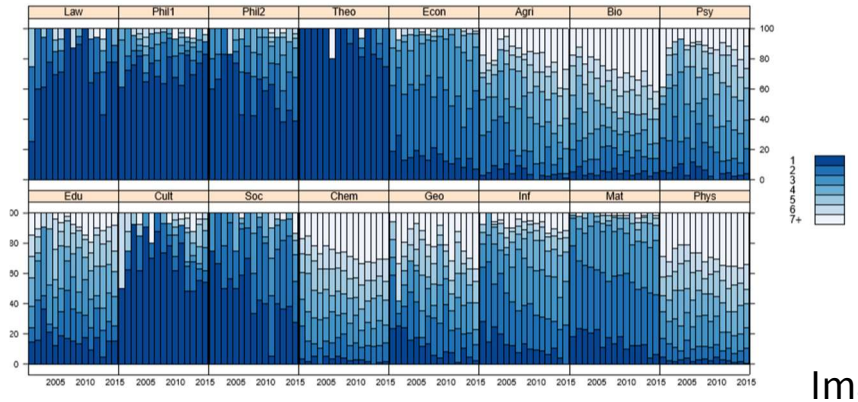


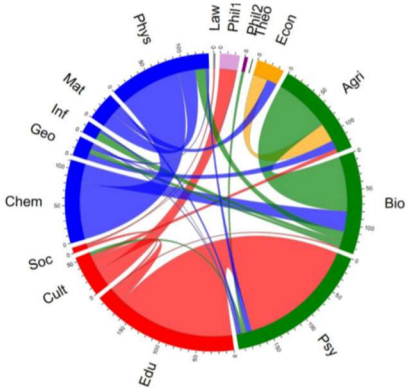
https://community.jmp.com/t5/Discussions/Show-means-on-box-plot-in-graph-builder/td-p/67947

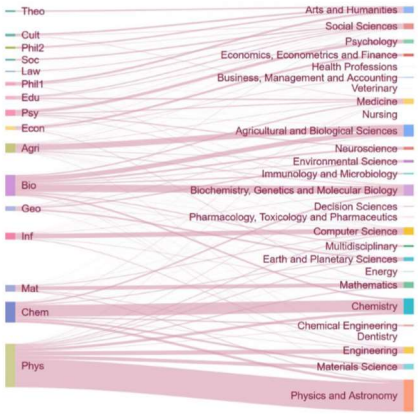
Explanatory Data Analysis – Lessons learnt

The single first task when obtaining a new data set

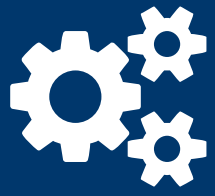
- **Multiple perspectives:** uni- vs. multivariate, (non-)graphical, …
- **No silver bullet or reference model**
  - Creative process
  - Requires understanding of the domain and statistics
- **Many *non-standard* ways to depict complex data**
  - Relatively easy to construct using contemporary software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







Images from: Zharova et al. (2017)



Data Preparation

Motivation & need, process, cleaning & preparation strategies for continuous and categorical variables

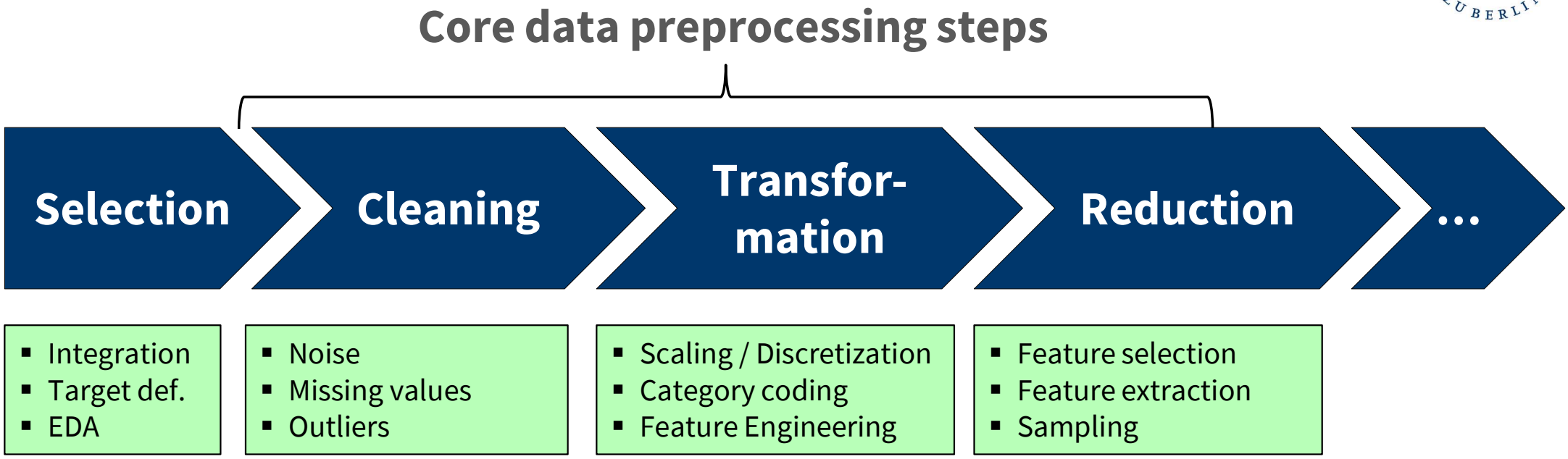
Background, Motivation, and Drivers

Real-world data needs much work before algorithms can process it

- **Typical data quality problems**
  - Data entry errors (e.g., AGE <0)
  - Missing values and outliers
  - Data integration and data merging problems
    - Often resulting from different systems using different keys
      - For ex. same client has a record in marketing and accounting database
    - Inconsistencies
      - Value '0' means actual zero or missing value
      - Amounts in different currencies
    - Duplicates (e.g., SALARY vs INCOME)
- **Making data preparation the most time consuming step in an ML process**

GIGO principle: Garbage-in-garbage-out

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.

Data Cleaning
Credit scoring example

- **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 ex., customer did not give their date of birth
- **Outliers**
  - Feature value that differs substantially from other values
  - Could be due to an error or a vali but extreme value

Missing value

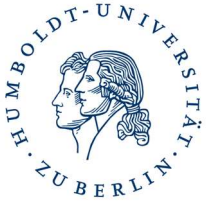
DEFAULT	DATE OF BIRTH	SALARY in K$	...
NO	16.03.1973	$75,00	...
NO	09.12.1984	$65,00	...
NO	03.05.1961	$125,00	...
NO	17.02.1979	$55,00	...
NO	08.08.1988	$9,250,00	...
NO	?	$60,00	...
NO	24.09.1976	$83,00	...
YES	13.06.1998	$15,00	...
YES	09.04.1789	$45,00	...
YES	17.11.1979	$111,00	...

Actual data error

Outlier?

Data Cleaning Strategies
Missing values and errors

Is the data
missing at random???



- **Errors: correct if feasible; else treat as missing value**
- **Missing values**
  - Keep: fact that a variable is missing can be important info
  - Delete
    - When number of missing values is excessive
    - Horizontally versus vertically missing values
- **Missing value treatment (if they are kept)**
  - Univariate replacement
    - Continuous attributes: use mean or median
    - Nominal attributes: use mode (most frequent category)
  - Multivariate imputation
    - Estimate replacement value using a statistical model
      - Derive this model from the other features
      - Feature with the missing values becomes the target
  - Consider adding a dummy variable to flag treated missing values

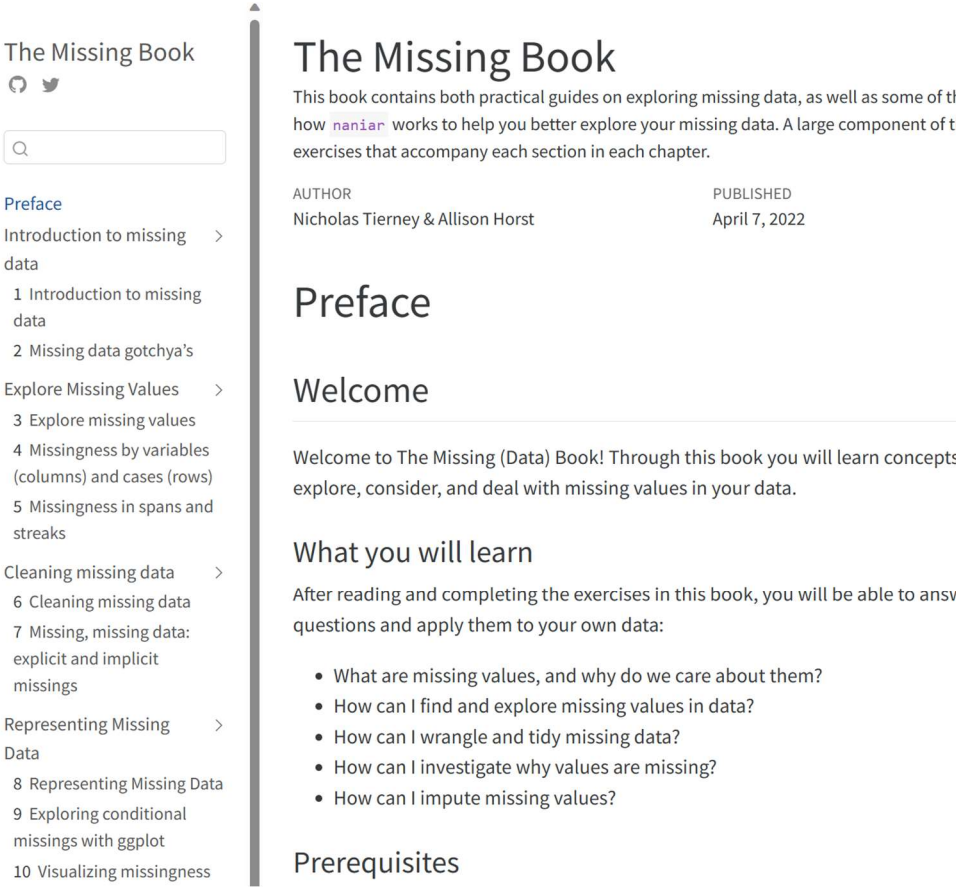
DEFAULT	DATE OF BIRTH (DoB)	DoB_IsMissing	...
NO	16.03.1973	0	...
NO	09.12.1984	0	...
NO	03.05.1961	0	...
NO	17.02.1979	0	...
NO	08.08.1988	0	...
NO	?	1	...
NO	24.09.1976	0	...
YES	13.06.1998	0	...
YES	~~09.04.1789~~	1	...
YES	17.11.1979	0	...

Still need to impute the original missing value.

Same for errors.

Data Cleaning Strategies

Missing values: if you want to learn more



https://tmb.njtierney.com/

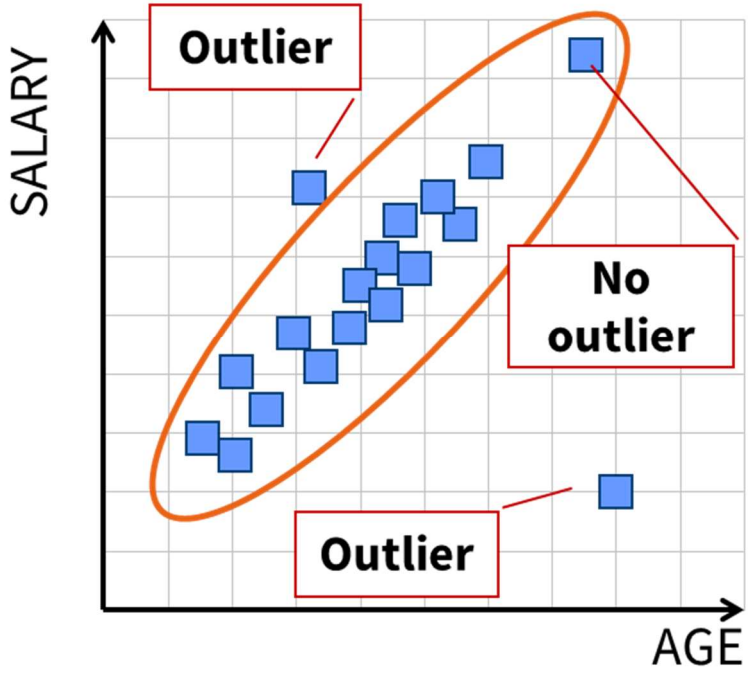
Data Cleaning Strategies
Outliers

- **Univariate versus multivariate outliers**
  - Do we consider only the feature itself to judge whether an observed feature value is 'unusual'
  - Or do we also account for the values of other features
- **Outlier detection versus outlier treatment**

- **Outliers**
  - Detect using graphs, statistics, or clustering
  - Treatment options
    - Keep as is if the outlier is valid
    - Treat as missing value if outlier is invalid
    - Truncate (if expected impact on the analysis appears excessive)

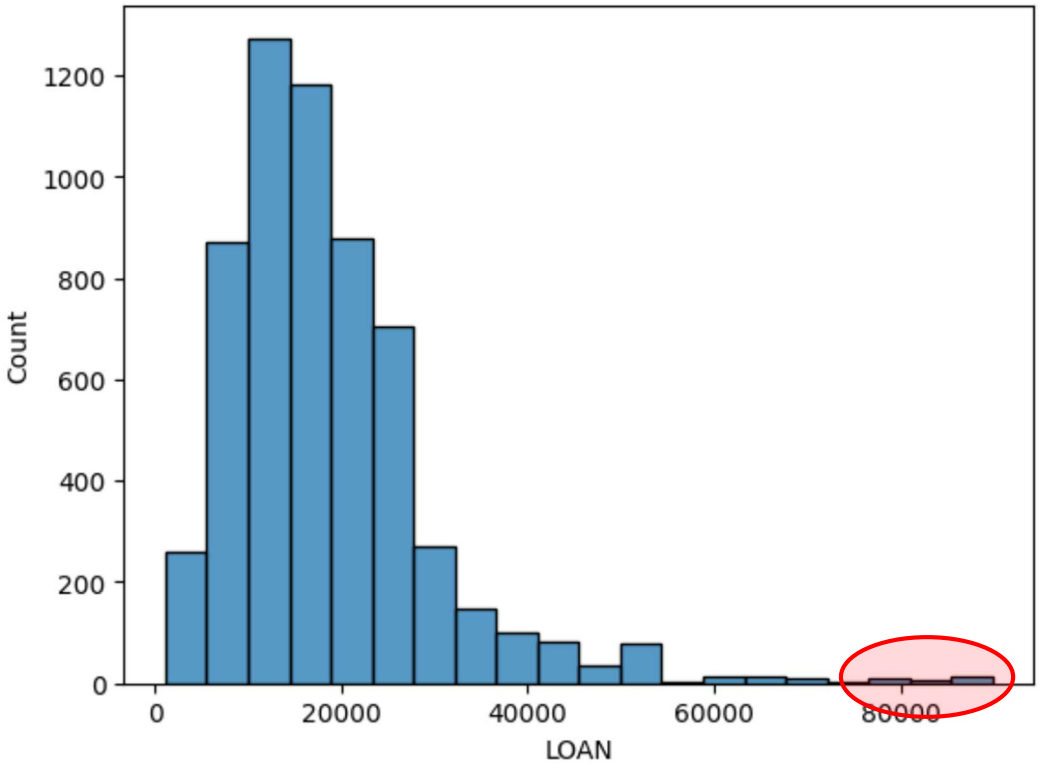
Outlier?

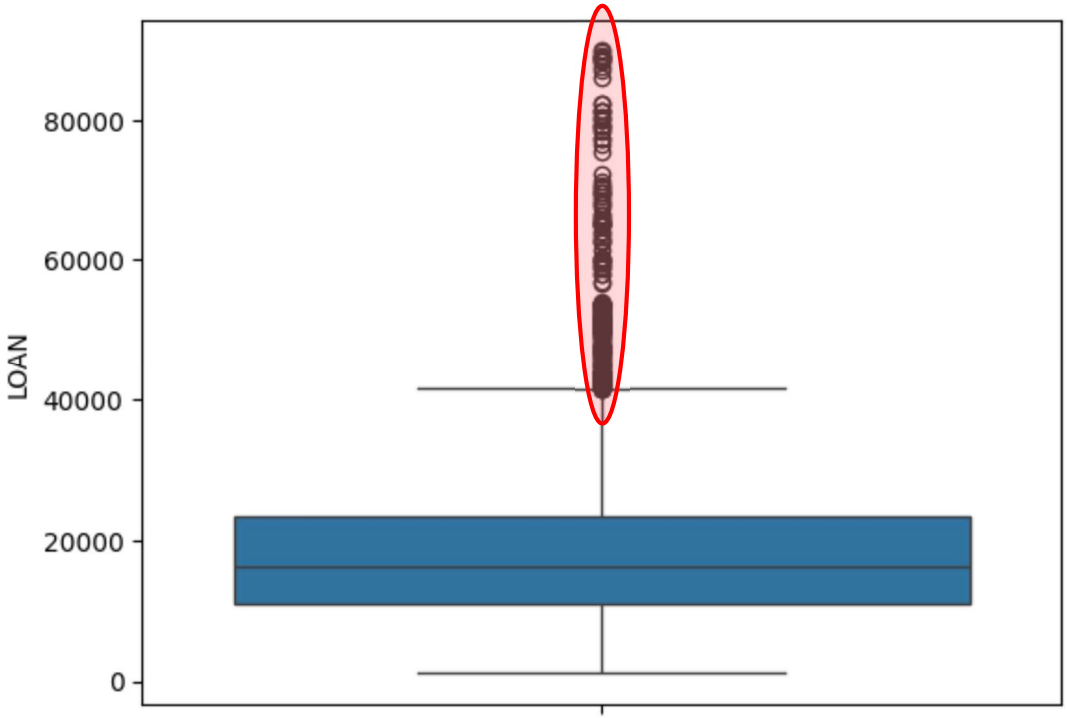
DEFAULT	DATE OF BIRTH	SALARY in K$	...
NO	16.03.1973	$75,00	...
NO	09.12.1984	$65,00	...
NO	03.05.1961	$125,00	...
NO	17.02.1979	$55,00	...
NO	08.08.1988	$9,250,00	...
NO	?	$60,00	...
NO	24.09.1976	$83,00	...
YES	13.06.1998	$15,00	...
YES	09.04.1789	$45,00	...
YES	17.11.1979	$111,00	...



Data Cleaning Strategies

Outlier detection using histogram or boxplot

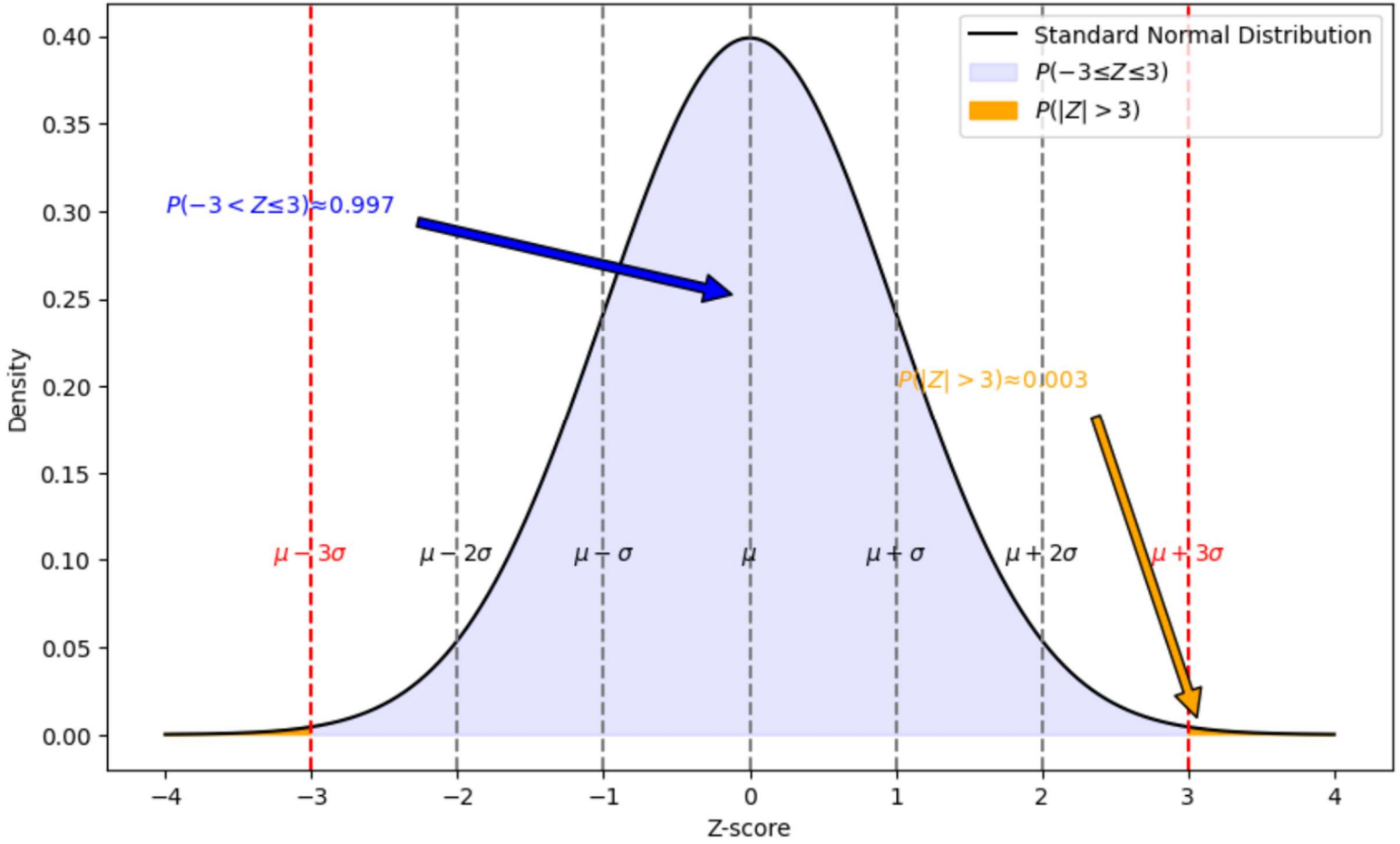




Data Cleaning Strategies

Outlier detection based on Z-Score

- **Standardize feature value**
  - Subtract from feature value X_i the mean of that feature, μ_X
  - Divide by the feature's standard deviation, σ_X
- **Rule of thumb**
  - Outliers have $|Z_i| > 3$
  - Assuming X is normally distributed, $\mathrm{P}(|Z_i| > 3) \approx 0.0027$

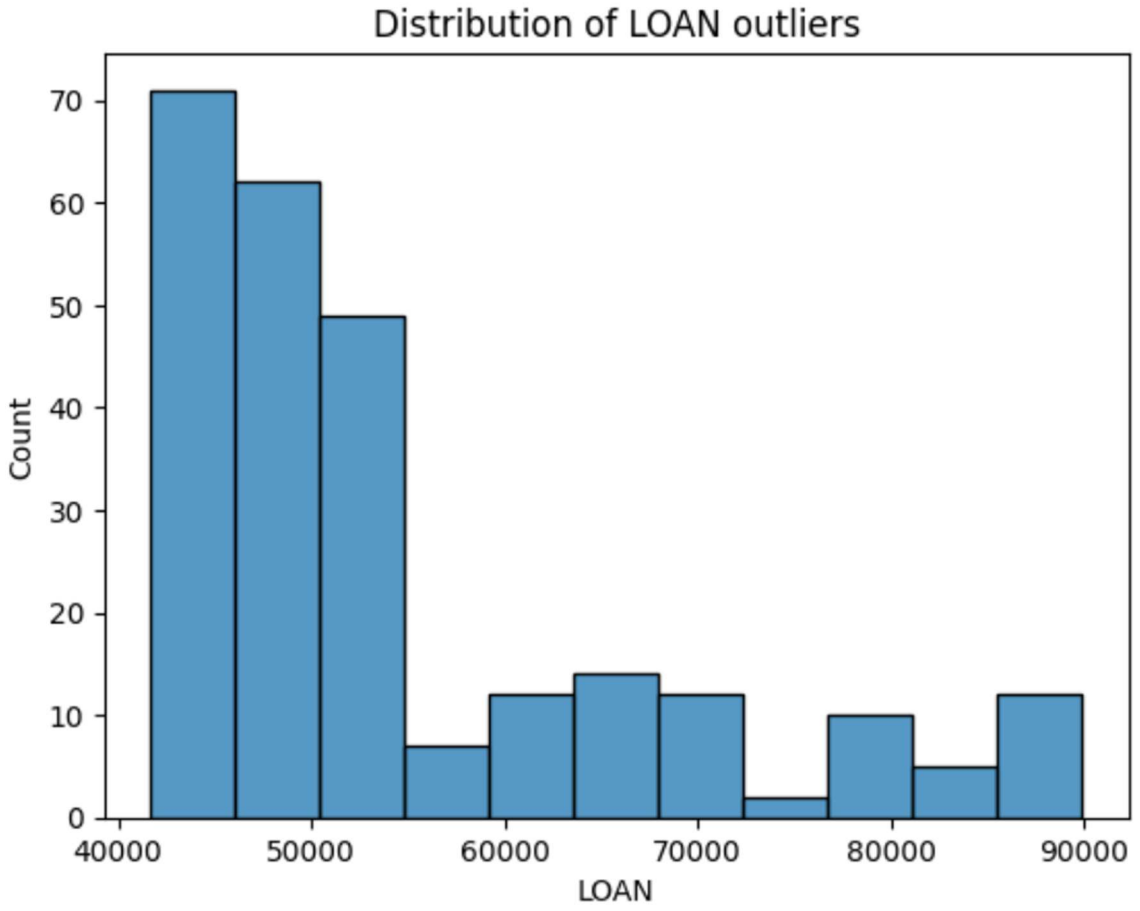


Data Cleaning Strategies

Outlier detection based on IQR

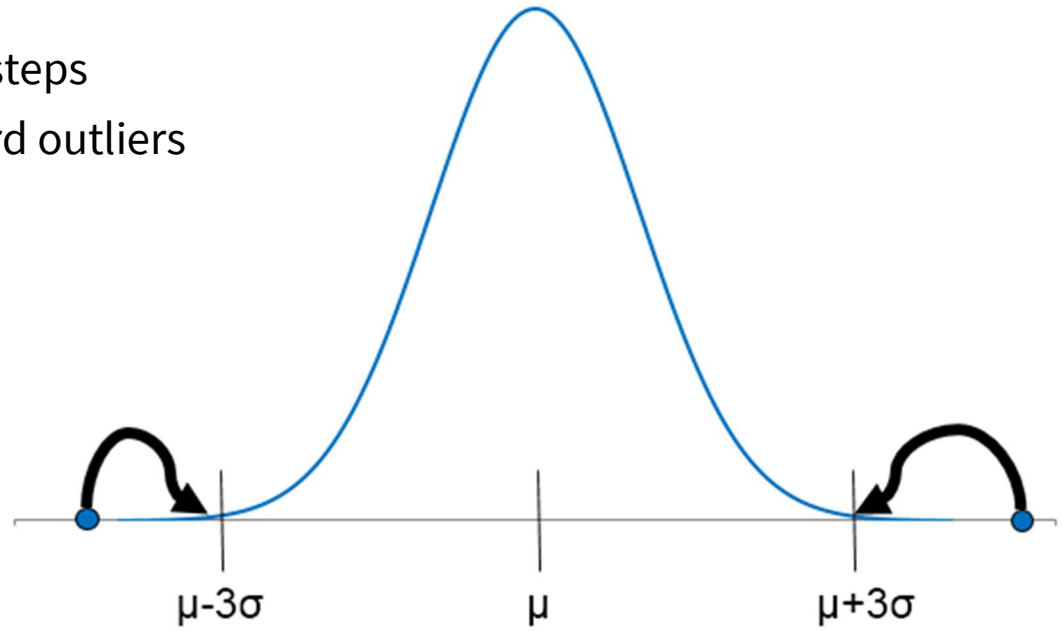
- **Same logic as Z-score approach**
- **Just compute threshold based on IQR**
  - Mild upper outlier: $X_i > Q3 + 1.5 * IQR$
  - Extreme upper outlier: $X_i > Q3 + 3 * IQR$
  - Analog for lower outliers

```
# Calculate IQR
X = hmeq['LOAN']  # select a feature
# Compute quantiles and IQR
q1 = X.quantile(0.25)
q3 = X.quantile(0.75)
iqr = q3 - q1
# Define upper/lower bound
upper = q3 + 1.5*iqr
lower = q1 - 1.5*iqr
# Find outliers
outliers = X[(X < lower) | (X > upper)]
```



Data Cleaning Strategies
Outlier treatment

- **Treat invalid outliers as missing value (keep, delete, replace)**
- **Keep values as they are**
  - Use robust data science methods in subsequent steps
  - For ex., decision trees (see later) are robust toward outliers
- **Truncate outliers**
  - Based on Z-scores:
    - Replace values having $Z > 3$ by $\mu + 3\sigma$
    - Replace values having $Z < -3$ by $\mu - 3\sigma$
  - Based on IQR
    - More robust than z-scores
    - Sometimes called winsorizing
- **Other forms of truncation (e.g., using sigmoid)**

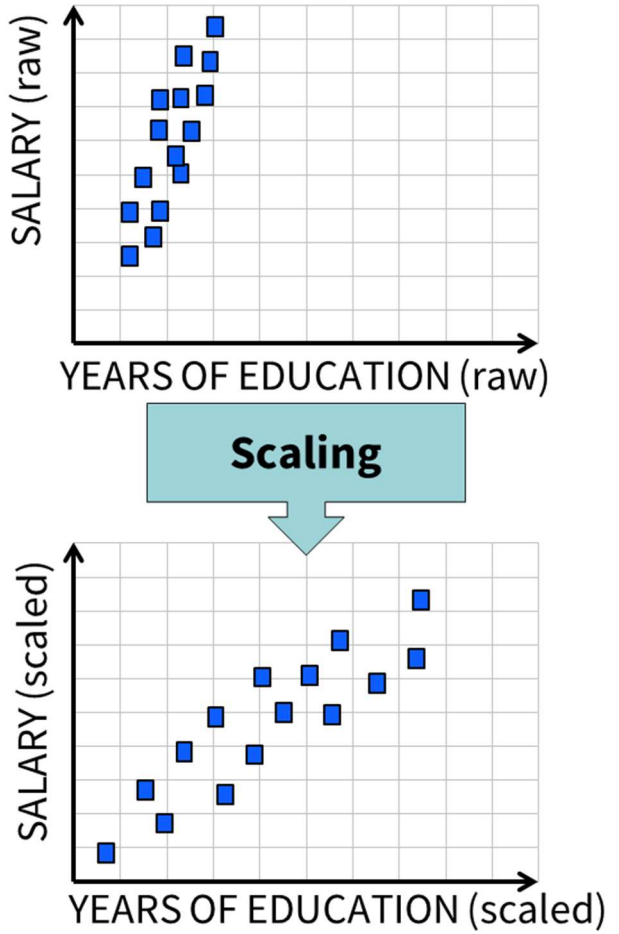


Preprocessing of Continuous Variables

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 methods
- **Scaling approaches**
  - Z-transformation (see above)
  - Min/max scaling

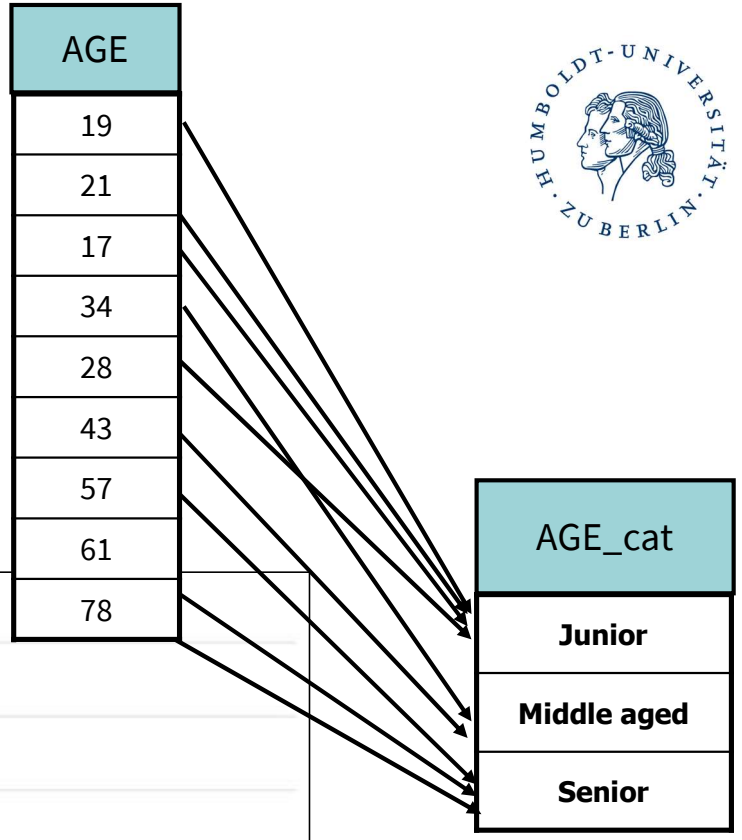
$$x_n = \frac{x_o - \min(x_o)}{max(x_o) - \min(x_o)} \cdot (max_n - min_n) + min_n$$

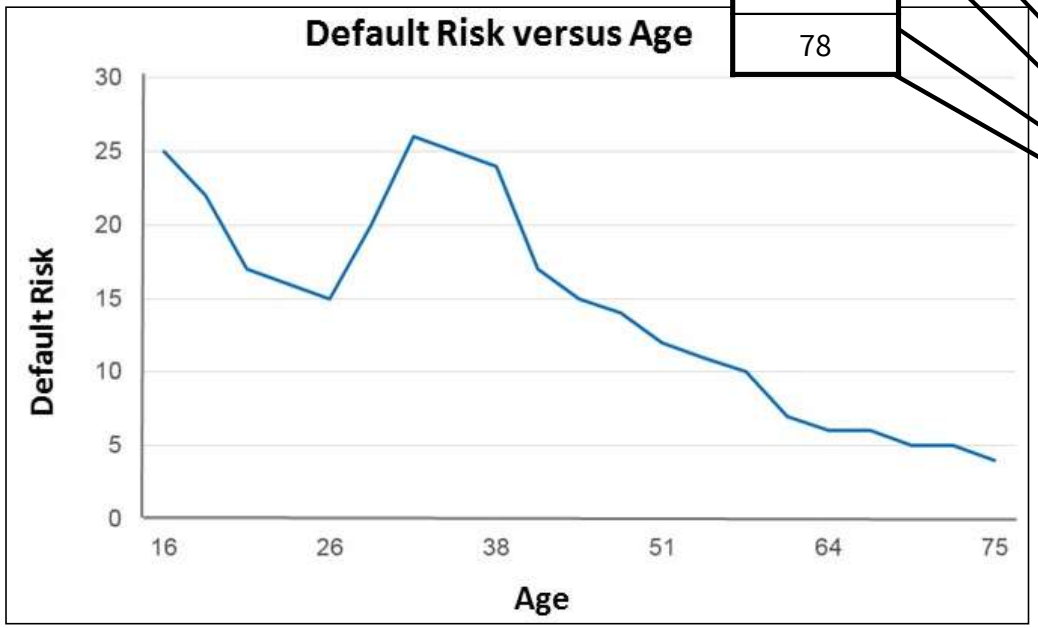


Preprocessing of Continuous Variables

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

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 Categorical Variables

Category encoding

- **Credit scoring example**
  - Variable (credit) PURPOSE
  - How to incorporate into an 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
1	G	44	car
2	B	29	House
3	B	58	travel
4	G	26	car
5	G	30	study
6	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
  - $y = \beta_0 + \beta_1 \text{AGE} + \beta_2 \text{PURPOSE}$
- **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
1	G	44	car
2	B	29	House
3	B	58	travel
4	G	26	car
5	G	30	study
6	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

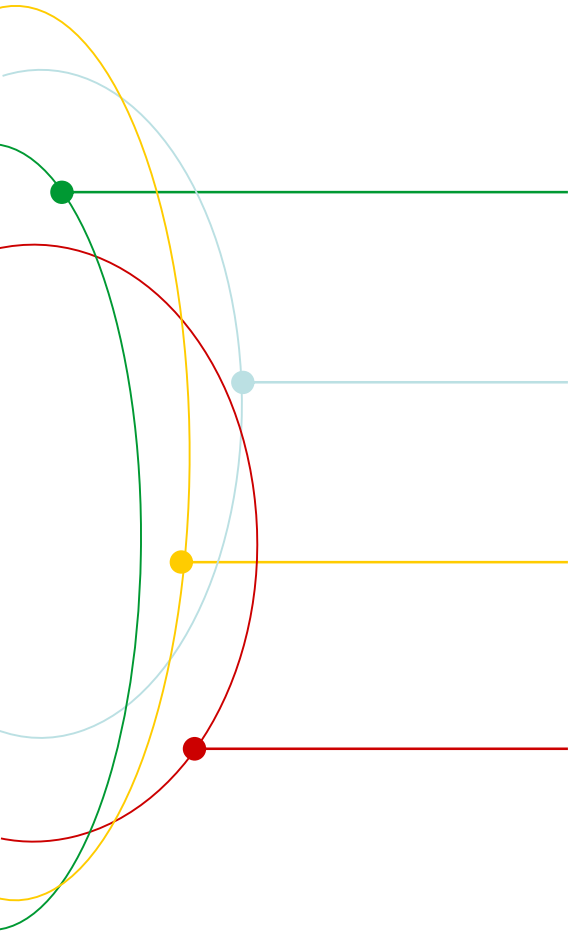
Category encoding using dummy variables

- **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

Summary

Summary



Learning goals

- Scope and need of EDA & 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

- Basic algorithms for supervised learning
- Python demo on EDA and data preparation

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