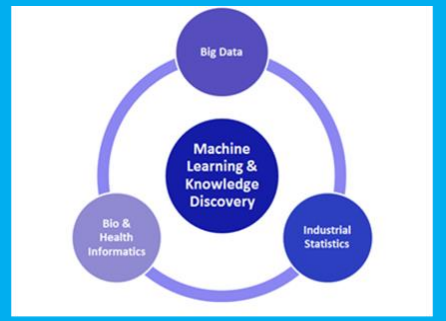
521156S
TOWARDS DATA MINING

MATKALLA TIEDONLOUHINTAAN

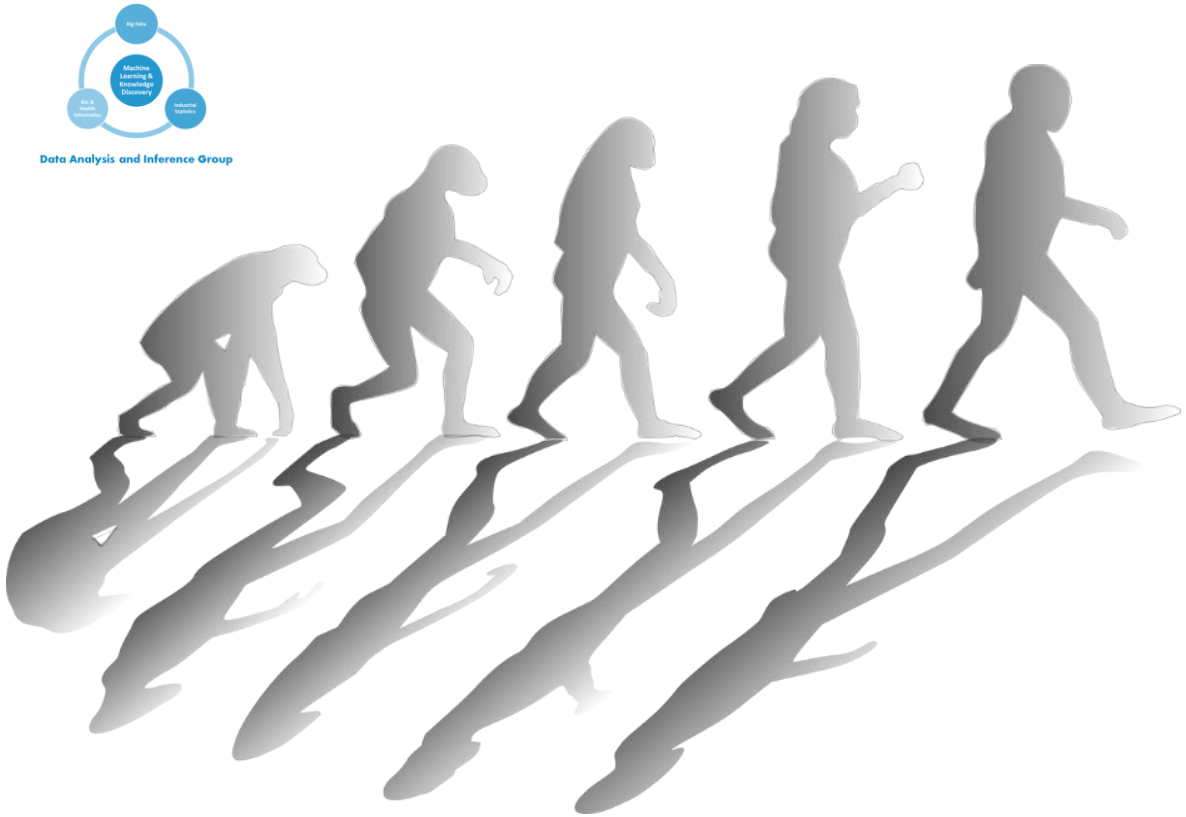
UNIVERSITY OF OULU

BISG Oulu



Data Analysis and Inference Group

Transformations, normalization, data reduction and distributions

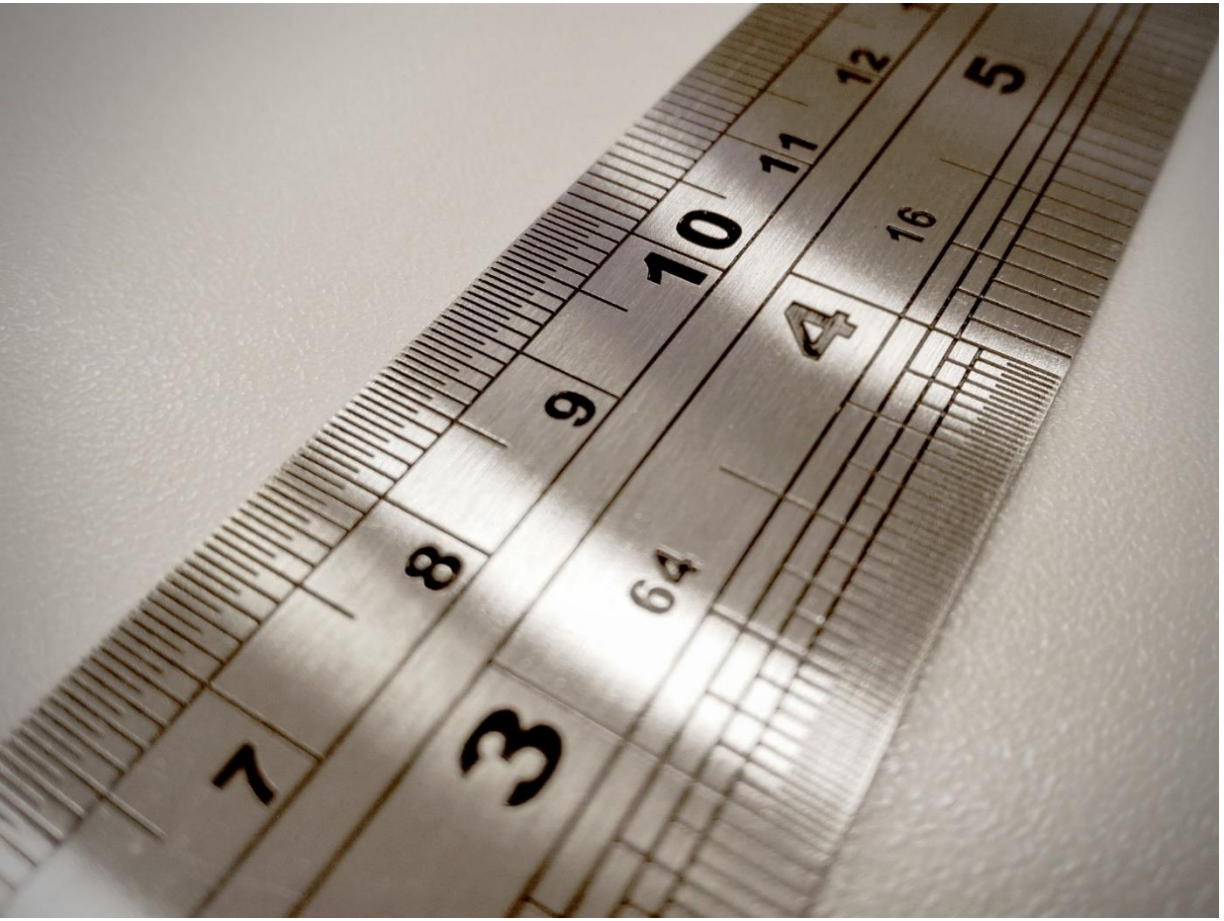


Why continue data pre-processing further after outlier detection and noise removal?

- It improves the performance of the following models
- it can be mandatory for some methods
- It can speed up the learning
- Helps the models to learn the association between inputs and outputs
- To simplify the statistical analyses and interpreting of the results

Domain knowledge is important, when choosing pre-processing methods to highlight underlying features in the data!

Normalization, scaling and standardization



What is normalization / scaling?

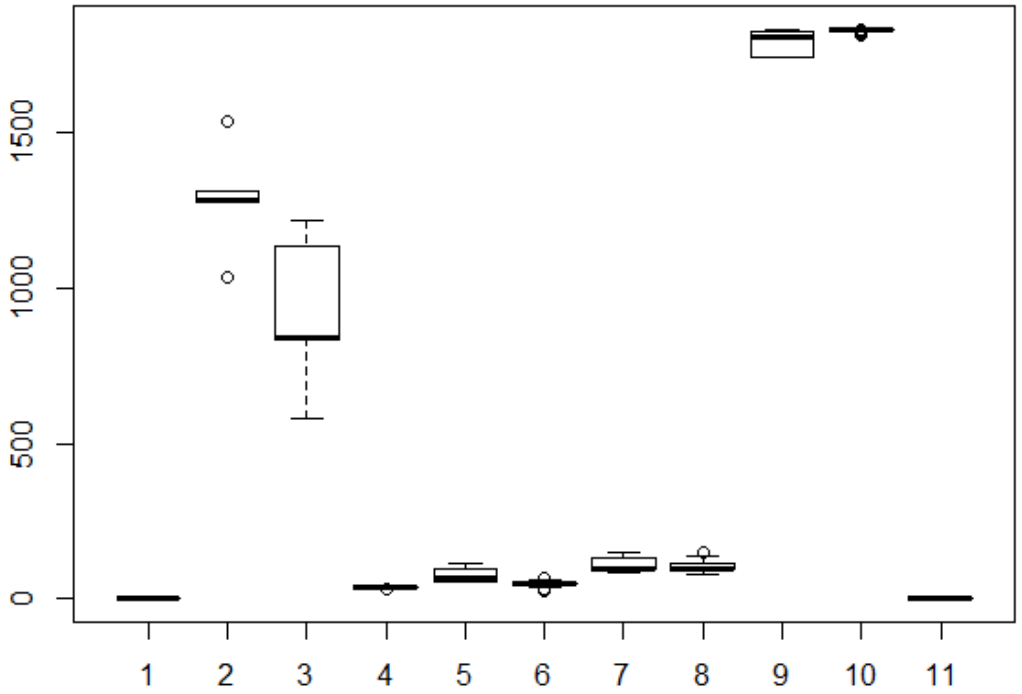
- Normalization means adjusting values measured on different scales to a notionally common scale.
- You do not have to utilize the same method for all the variables, just remember which one you selected for each variable.

Why normalization?

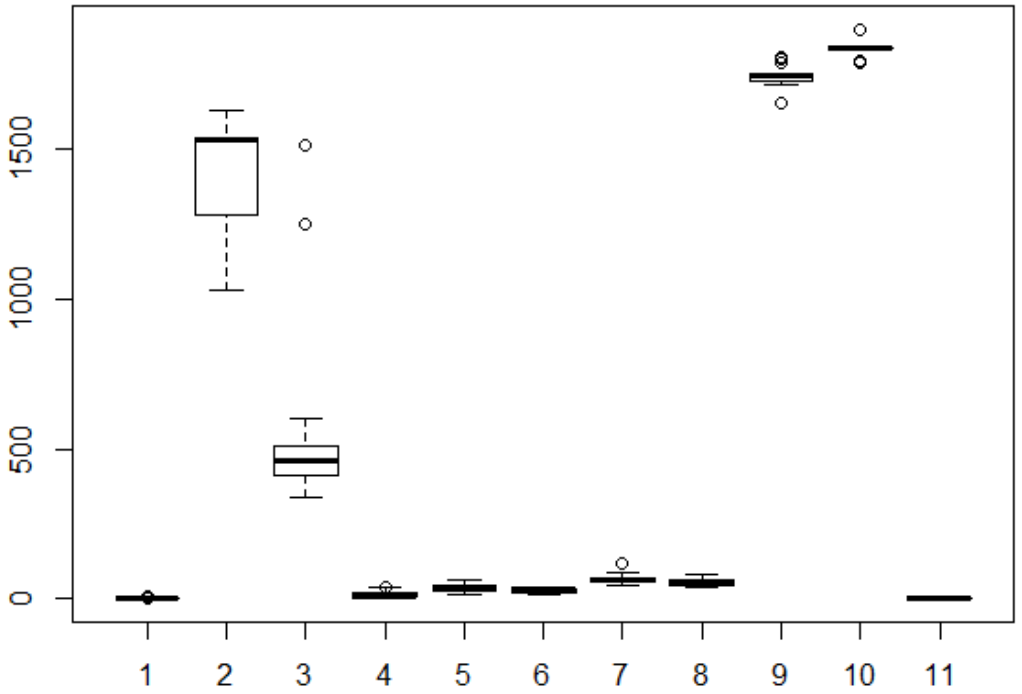
- **Speeds-up some learning techniques**
- **The learning technique might require normalization**
- **Helps to prevent attributes with large ranges to outweight ones with small ranges, for example:**
  - Variable1 [10 000 - 70 000]
  - Variable2 [2 - 90]
  - Variable3 [2 - 5]
  - Variable4 [0.002-0.07]
- **Note that you may need to store your normalization parameters!**
  - Rescaling back to original scale in reports
  - For scaling new data later

Is there any difference between good products and bad products?

Process parameters are visualized with boxplots in their original scales.



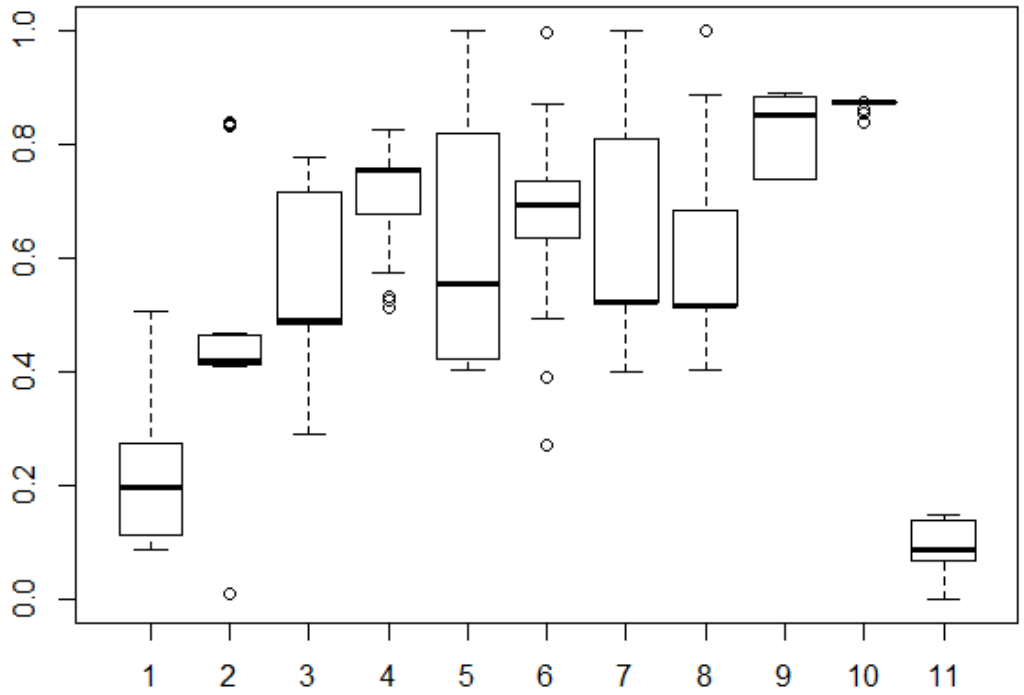
GOOD



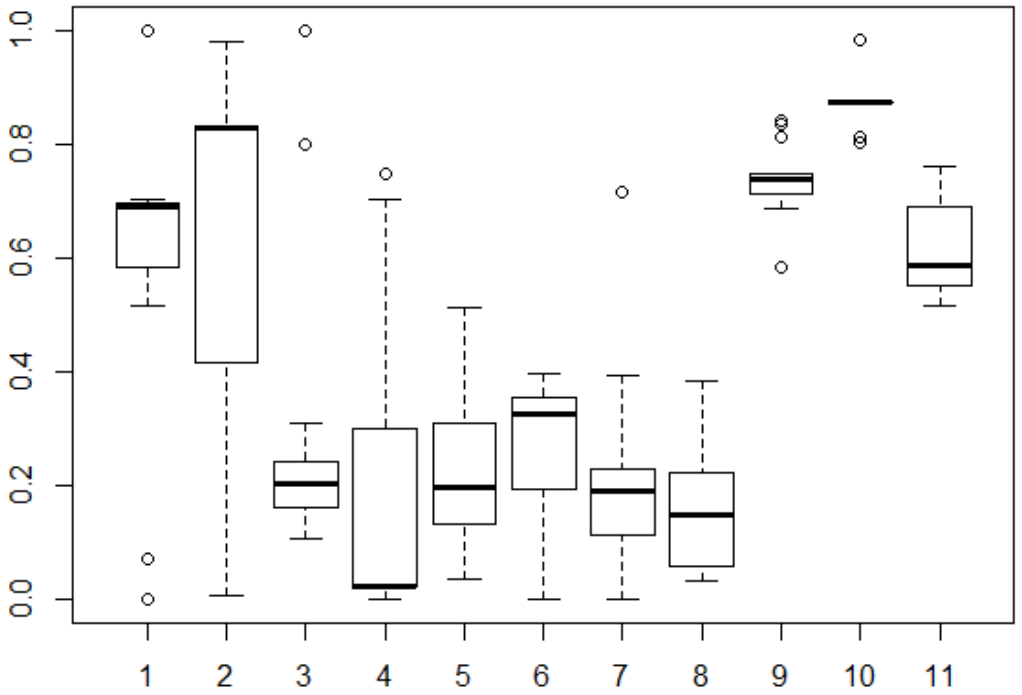
BAD

Is there any difference between good products and bad products?

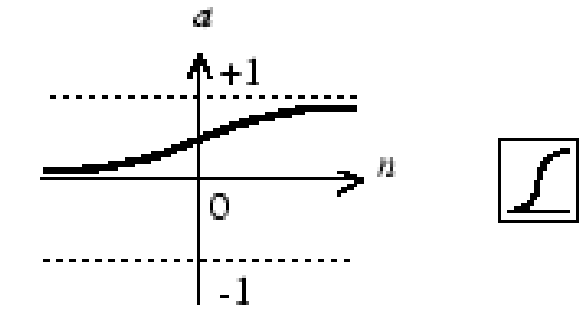
Process parameters are visualized with boxplots with normalized scales.



GOOD

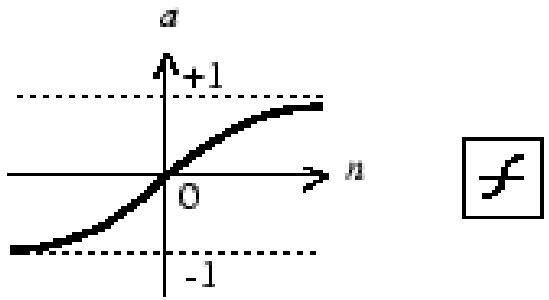


BAD



$a = logsig(n)$

Log-Sigmoid Transfer Function



$a = tansig(n)$

Tan-Sigmoid Transfer Function

Min-Max Normalization

- Range [0,1]

$$v' = \frac{v - v_{min}}{v_{max} - v_{min}}$$

- Any arbitrary range [a,b]

$$v' = \frac{(v - v_{min})(b - a)}{v_{max} - v_{min}} + a$$

Standardization or z-score standardization

- Transform the data to center by removing the mean value of each variable
- Then scale it by dividing non-constant variables by their standard deviation

$$v' = \frac{v - \mathbf{mean}(v)}{\mathbf{sd}(v)}$$

- Suitable especially, if the min and max values of the population are unknown
- Can be used with data containing outliers

Decimal scaling

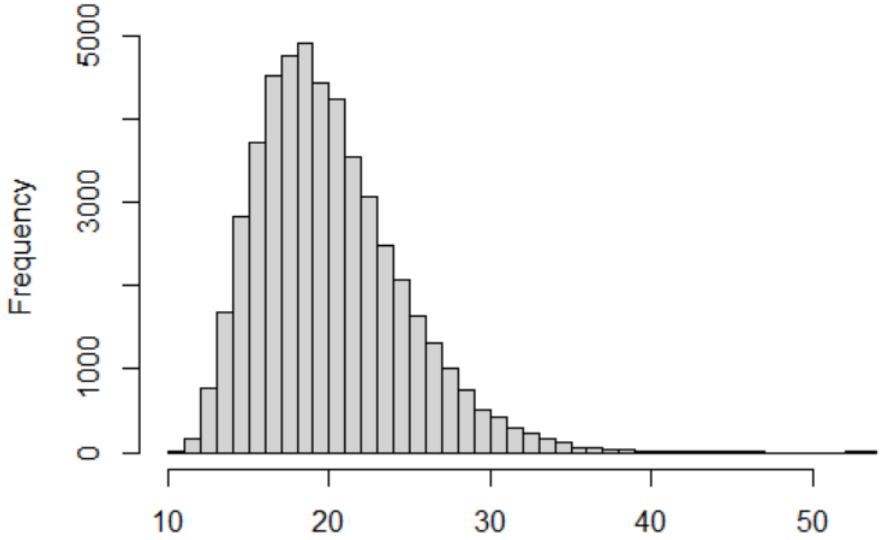
- Scaling in terms of decimals, simply by moving the decimal point of the values for each variable.
- The movement of decimal points depends on the max value of the variable.

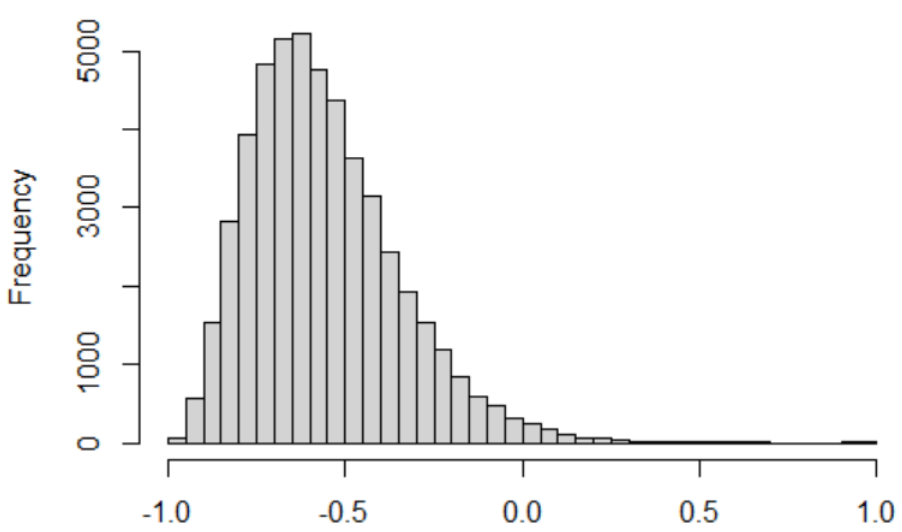
$$v' = v/10^j,$$

where *j* is the smallest integer

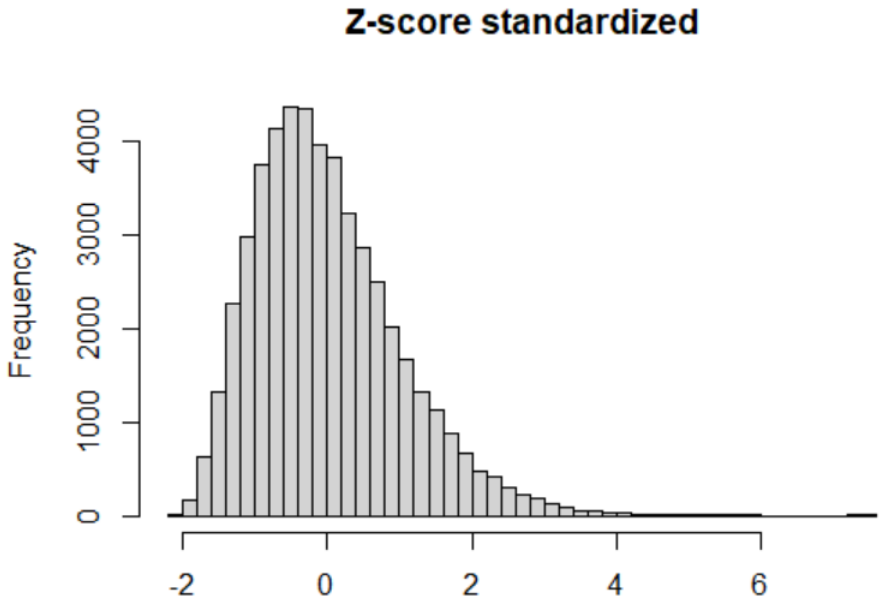
such that $\max(|v'|) < 1$.

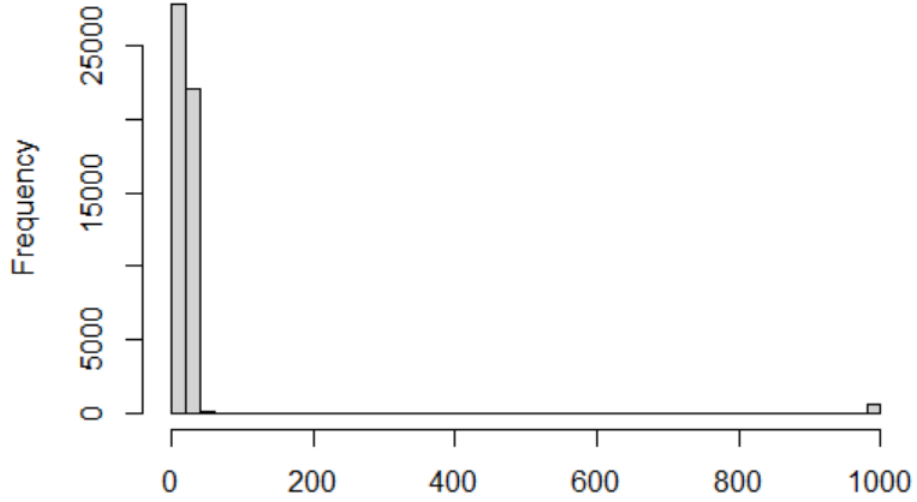
For example:
[2,6,18,79] -> *j* = 2
[30 231, 4 008, 299] -> *j* = 5

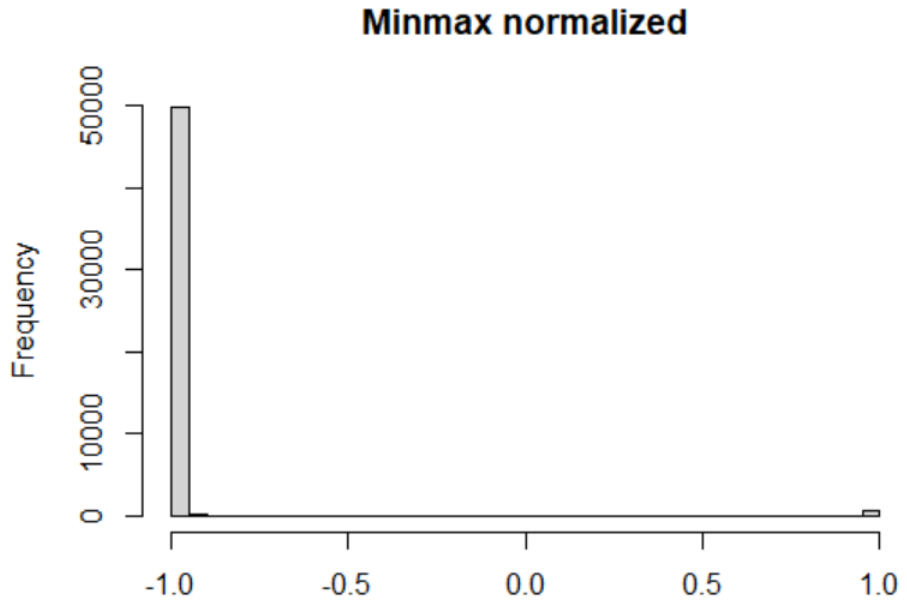




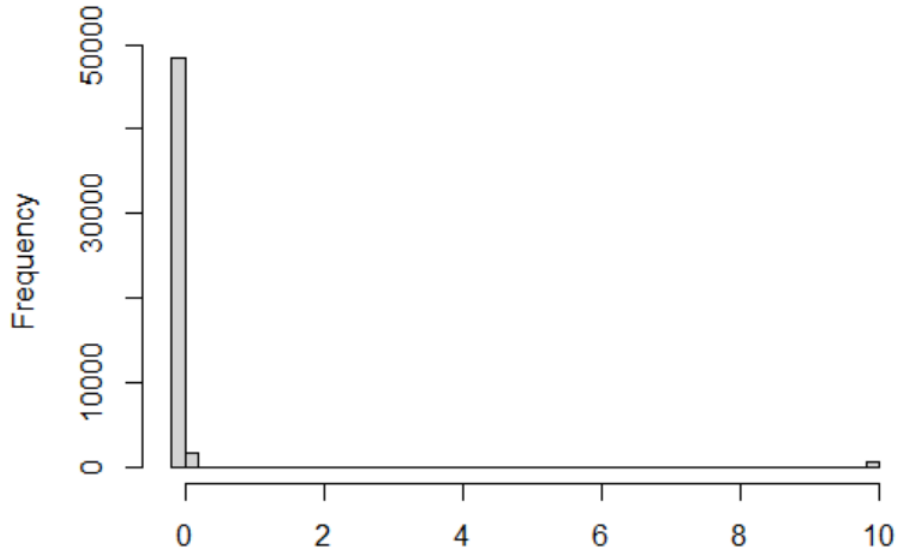
- Normalization methods do not affect the data distribution
- What happens to outliers?
- Z-score standardization can be utilized for outlier detection, but the data should be Gaussian.



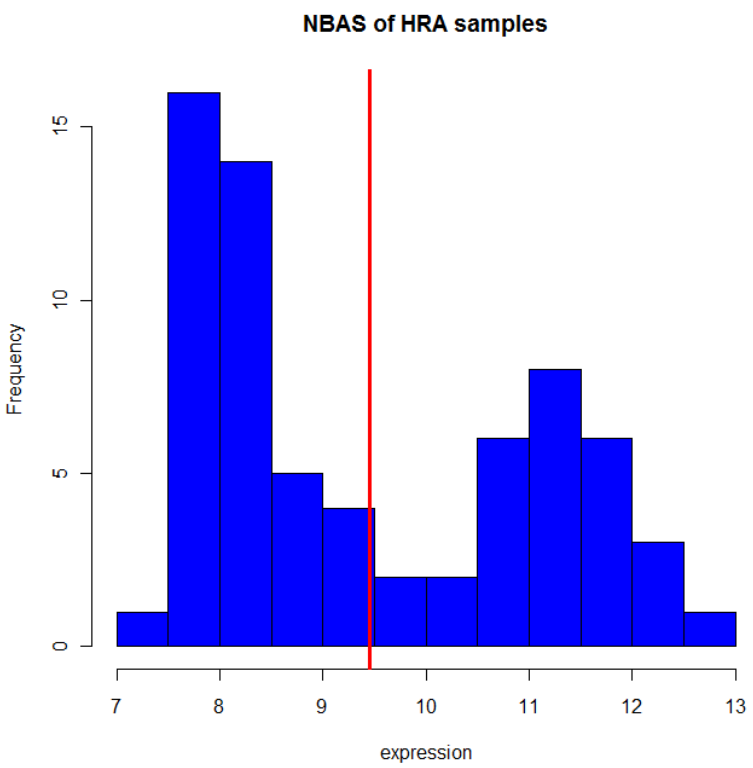




- If the data has clear measurement errors, the visual inspection helps to find them
- If outliers are left in the data set, the variablility in the correct data will become meaningless
- With N(0,1) distribution the |value|>3 is often used as cut-off, but in practise that is often too low, you could use 6-7 instead.

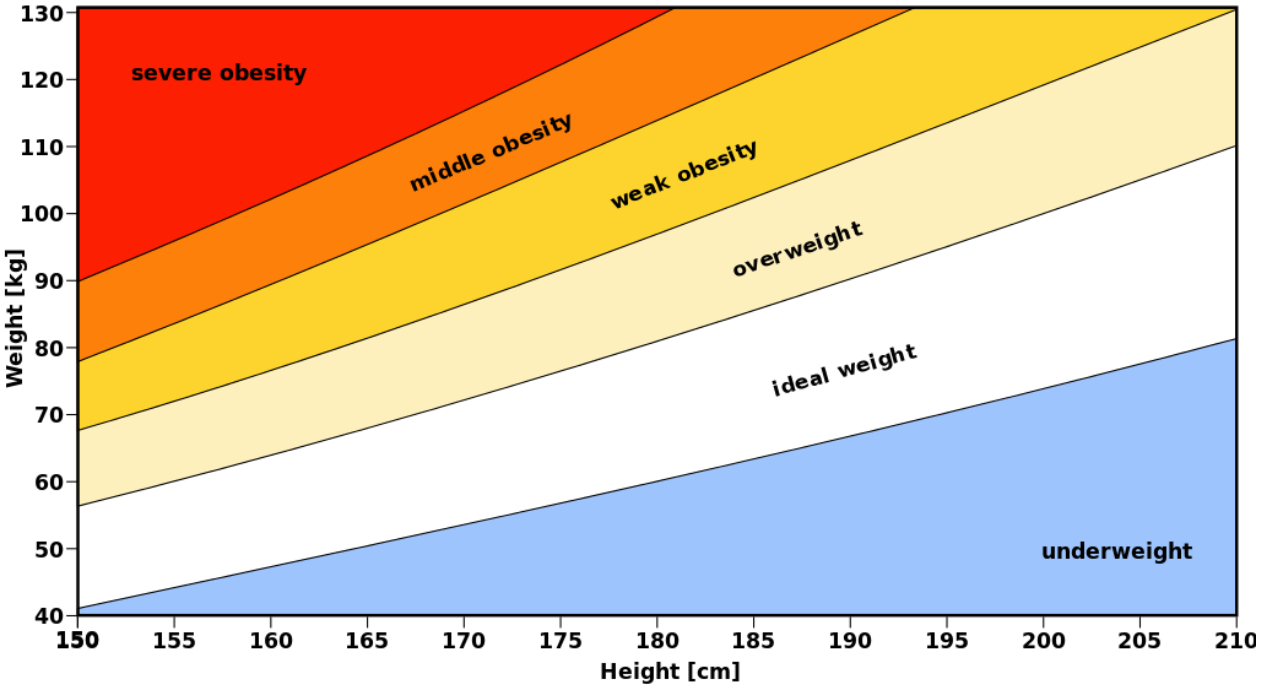


Classification of the continuous variables, factors

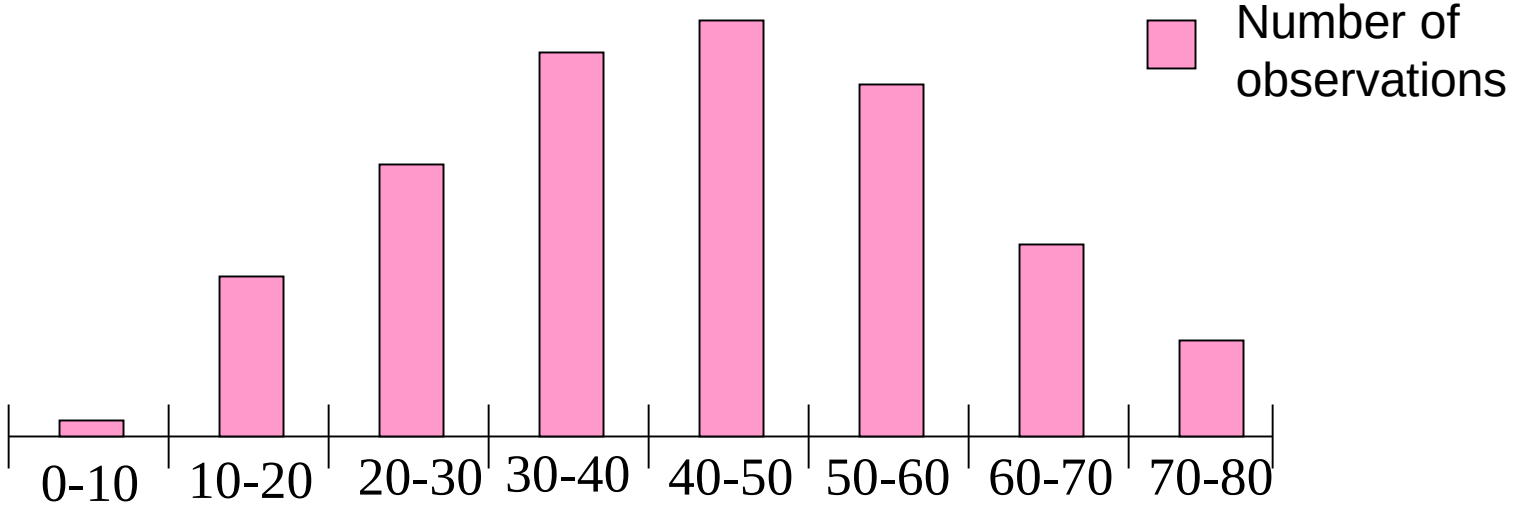


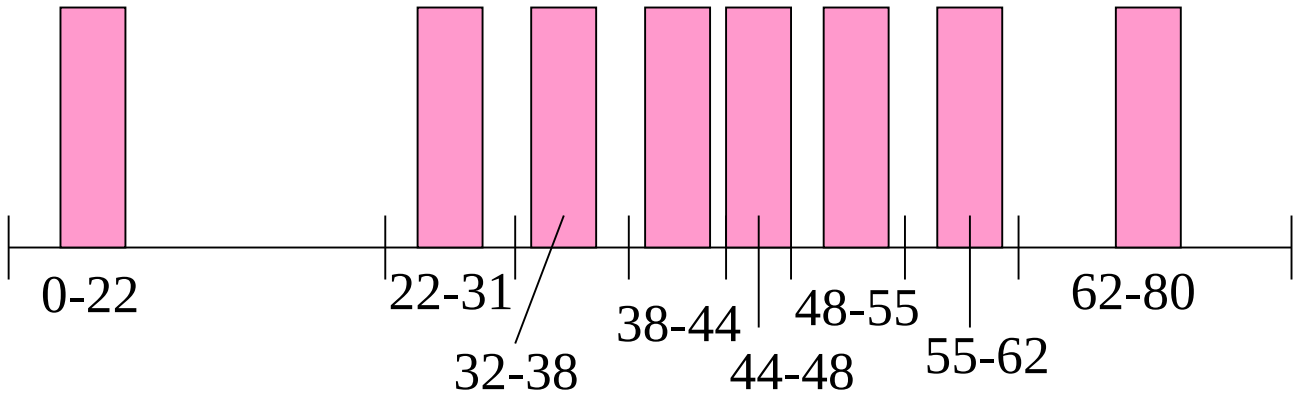
Why discretization of a continuous variables is used?

- When researchers believe there are distinct group of observations
- When researchers are interested in group differences rather than individual differences
- Utilizing a median split to create equal groups (very popular in social sciences)
- To simplify the statistical analyses and interpreting of the results
- When selected method requires categories
- Concept hierarchies can be useful
  - Instead of GPS coordinates it may be more convenient to use Street, District, City, Country…



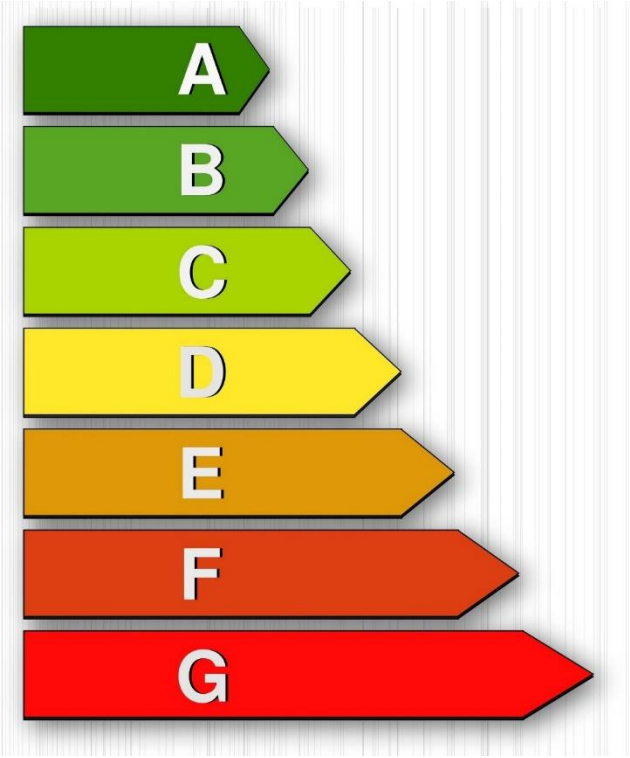
- **Sometimes quantitative scale reflects meaningful qualitative differences**
  - Physical measurements may have levels for different treatments: low, elevated, needs medical care
  - “The amount of cigarettes consumed in a week”
- **Quantitative variable may have a natural, meaningful cut points**
  - Age may have a legislative boundary: underage, adult
- **When there are only few values for a continuous variable**
- **When the relationship is not linear, but you cannot fit a nonlinear regression model.**


Number of observations
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80


0-22
22-31
32-38
38-44
44-48
48-55
55-62
62-80

Categorizing, discretizing or dichotomizing continuous variables

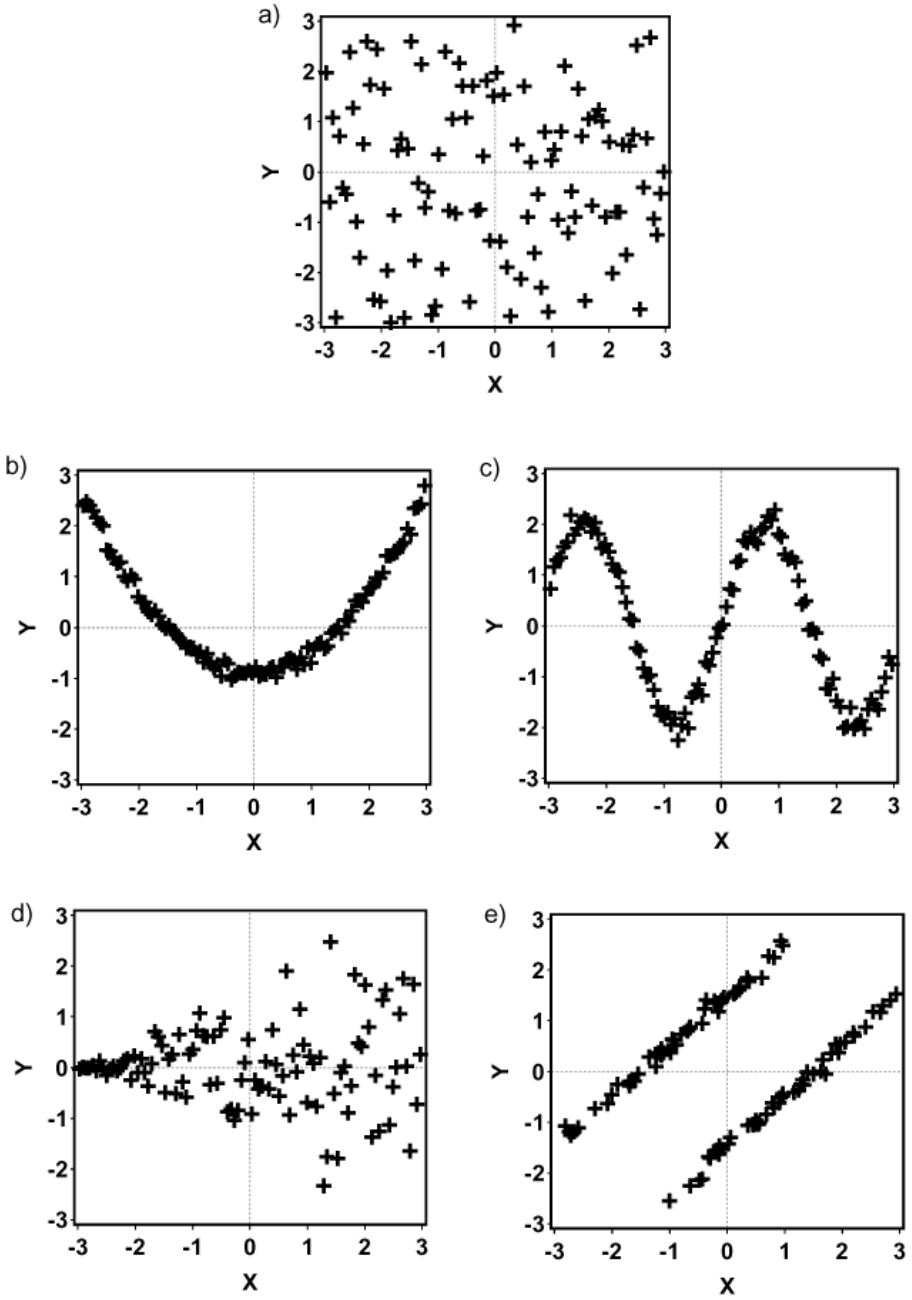
- **Some methods accept only categorical variables**
- **May simplify analysis and interpretation**
  - It is easy to understand what was done and what the results are
  - 2-class examples: normal/abnormal, risky/safe, treat/do not treat…
- **But simplicity may be gained at a high cost**
  - Instead of solving problems, new ones are created
- **Generally, categorizing is not recommendable**
  - Avoid it and plan your data mining project differently
  - If you use it, make sure that your selections are justified



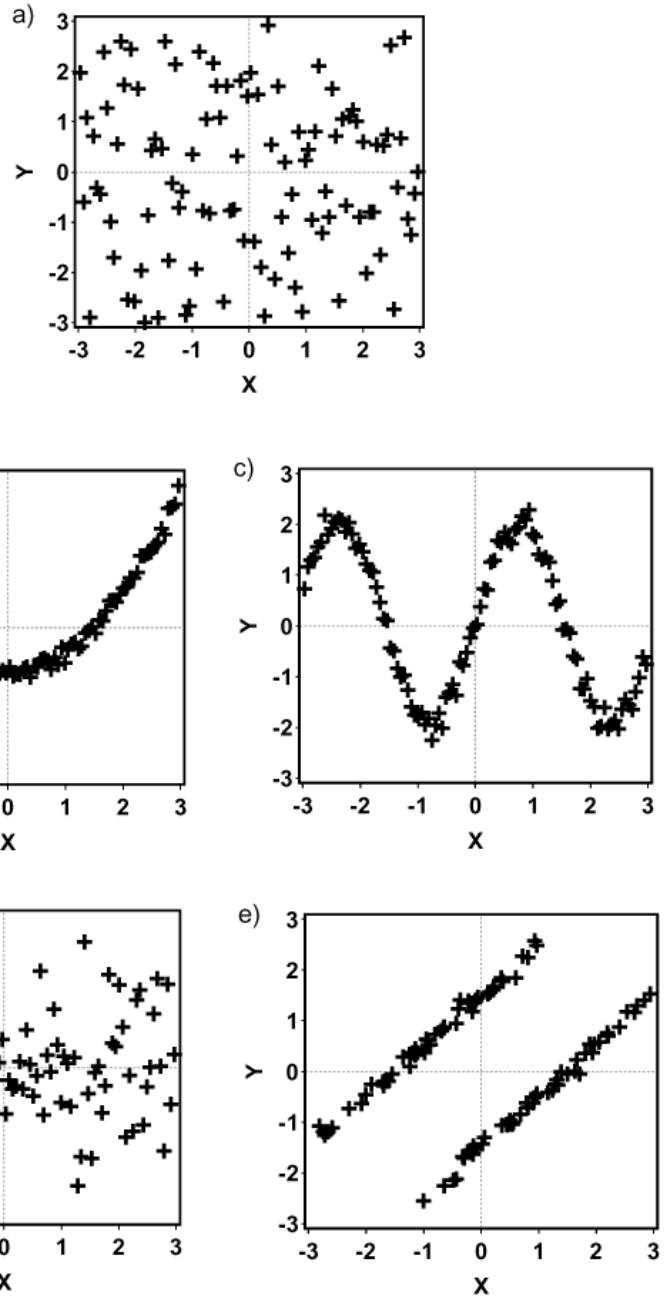
Problems that may occur when discretizing continuous variables:

- **Loss of information**
  - The loss is smaller with multiple categories but severe with only two categories
- **Categorization does not make use of within-category information**
  - Anyone above or below the cut-point (with two-class situation often the median) is treated differently: normal/high BMI
  - The risk of misclassification because of measurement error when it is high
- **Comparison with other studies that used different cut-points becomes impossible (reporting of the results!)**
- **By categorizing one continuous variable can artificially make another variable appear associated with the outcome**

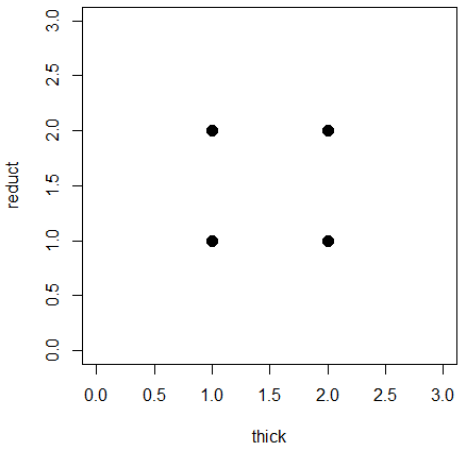
An example of information loss caused by dichotomization



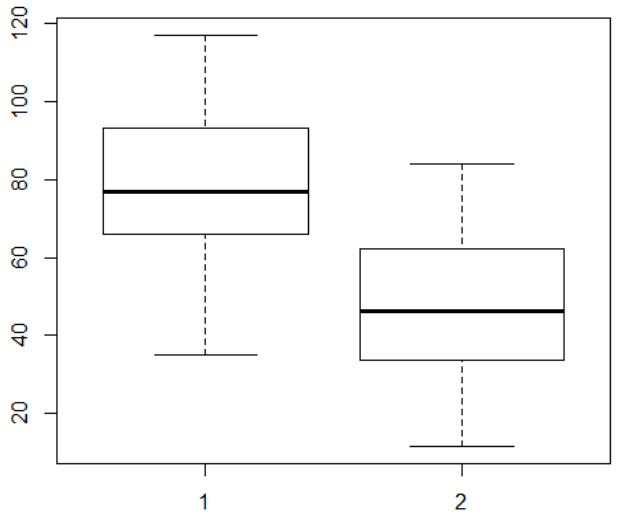
An example of information loss caused by dichotomization



		Y positive?	
		Yes	No
X positive?	Yes	25	25
	No	25	25



		Reduct
Thick	65	33
	20	80



Which data analysis methods work with categorical data?

- Correlations are not meaningful
- Scatter plots do not work

– **Frequency tables and correspondence tests**

– **Visualization of categorical data against continuous variables**

- Box plots for each category

– **Statistical tests**

It is important to understand, what method you can and cannot do!

What to do with the categorical variables?

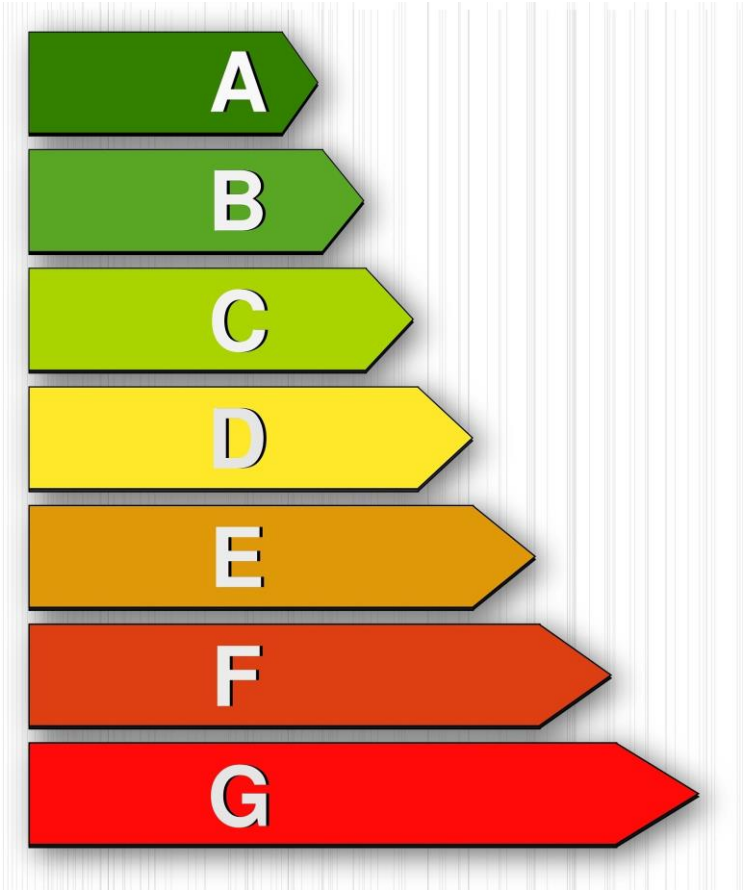
- **It depends on the method you plan to use in data mining / modelling**
  - Some methods do not allow categorical data (regression)
- **Choose method that allow the variable type “factor”**
  - Make sure that you have defined your variables correctly
- **Dummy-coding**
  - If you have *K* categories, you need *K*-1 dummy variables

Variable values: red, blue, green, yellow
Dummy_red: =1 if Variable=red, else =0
Dummy_blue: = 1 if Variable=blue, else =0
Dummy_green: = 1 if Variable=green, else =0

Data reduction

Data reduction brings efficiency:

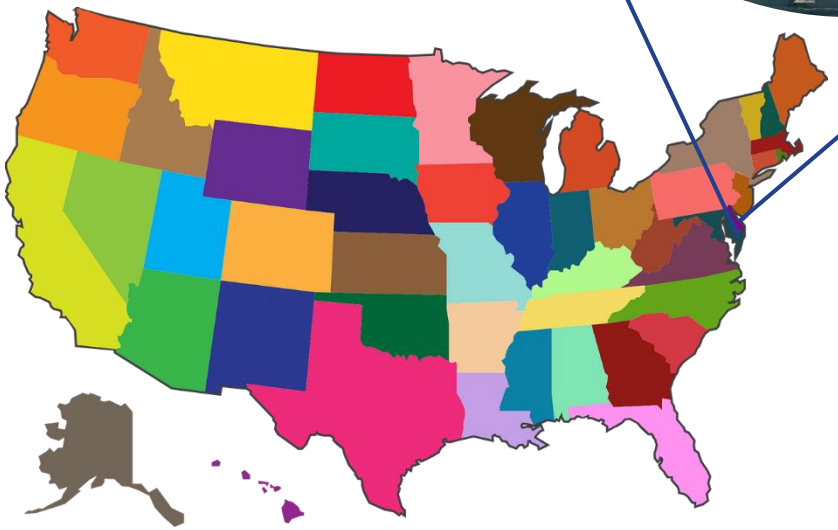
- As data warehouse may store terabytes of data, complex data analysis/mining may take a very long time to run on the complete data set
- Obtains a reduced representation of the data set that is much smaller in volume, but yet, produces the same (or almost the same) analytical results



DATA REDUCTION

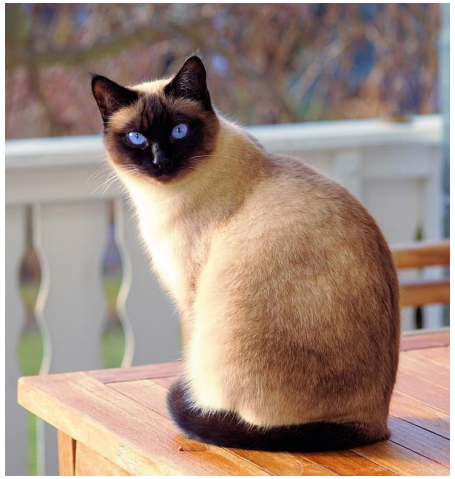
- **Reducing the number of variables**
  - Aggregation Q1,Q2,Q3,Q4 -> year
  - Removing irrelevant attributes
  - Principle component analysis
- **Reducing the number of variable values**
  - Binning (histograms): reducing the number of variable values by grouping them into intervals
  - Clustering: grouping values in clusters
  - Aggregation or generalization
- **Reducing the number of observations**
  - Sampling

Data aggregation and generalization



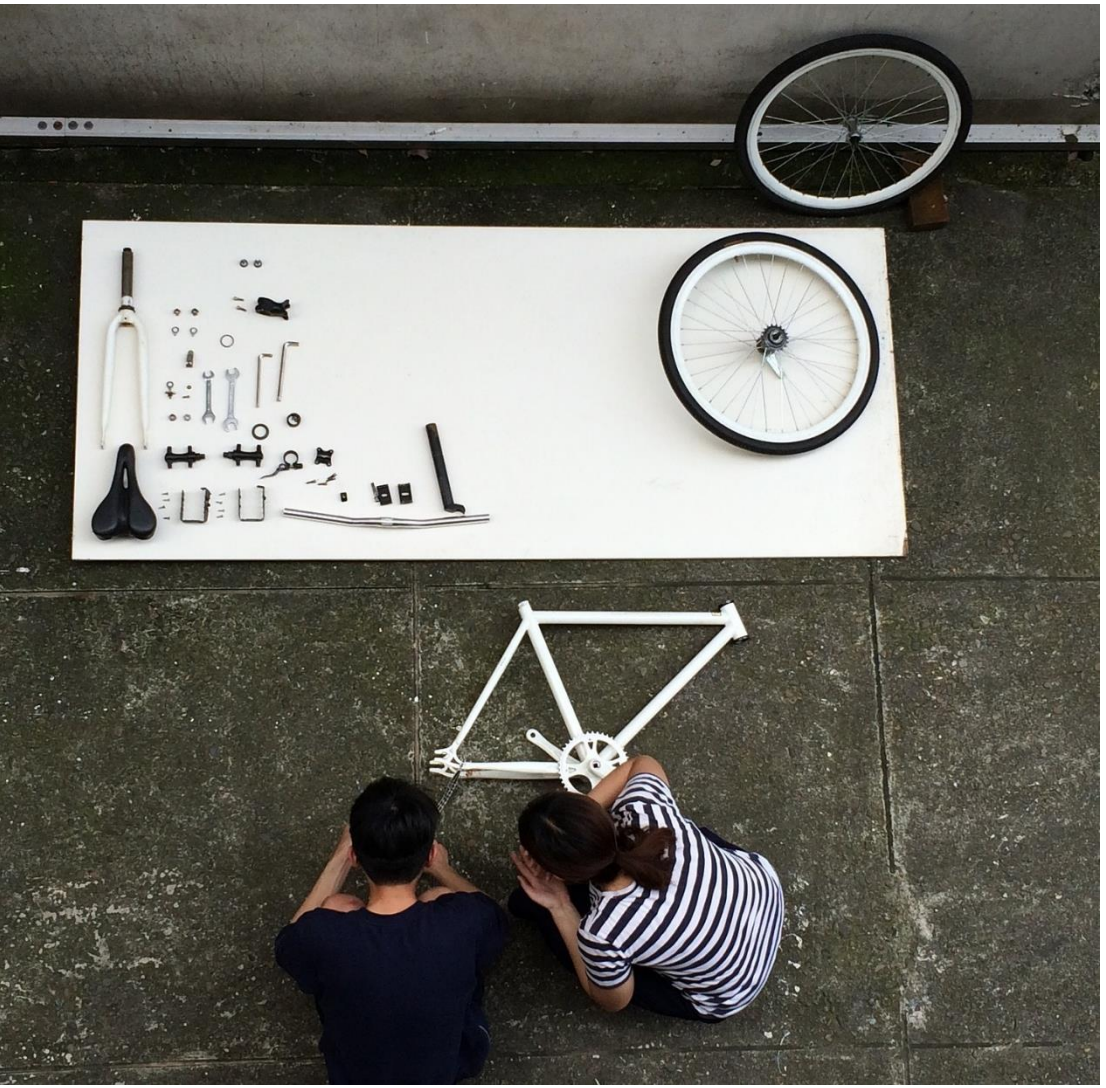
- **Information is gathered and expressed in a summary form**
- **Is based on a certain group or class**
  - Average age of residents of each district in the City
  - The average income within groups based on the education
  - District / city / country instead of accurate address

DATA GENERALIZATION



Images from pixabay

Variable construction

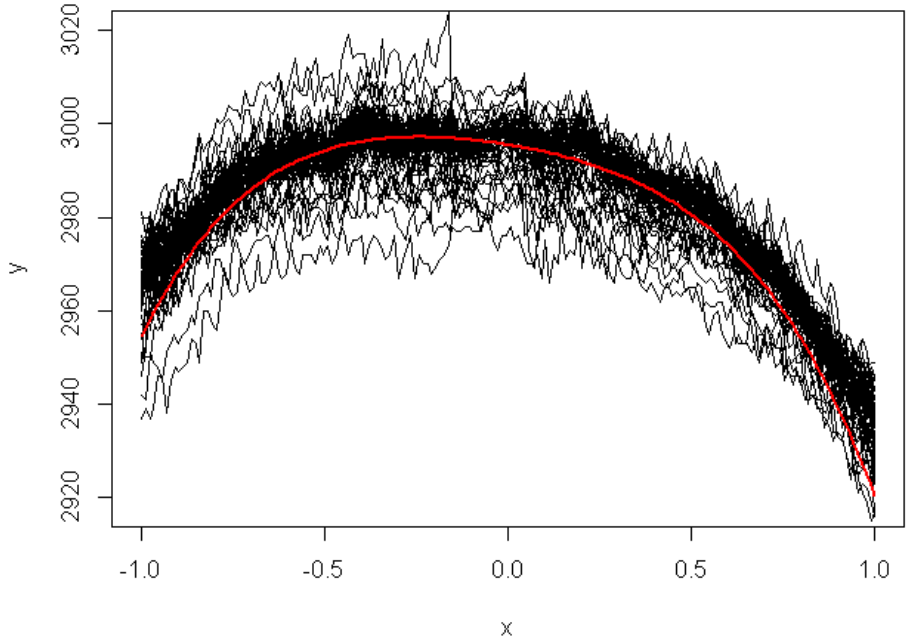


- **By using domain knowledge, combine or split existing variables into new one that has higher predictive power or is more meaningful**
  - Example for combining:

$$\textbf{specific force} = \frac{\textbf{force}}{\textbf{width}} \textbf{ or reduction} = \frac{\textbf{thickness1}}{\textbf{thickness2}}$$

  - Example for splitting:

exact date to **workdays** and **weekends**



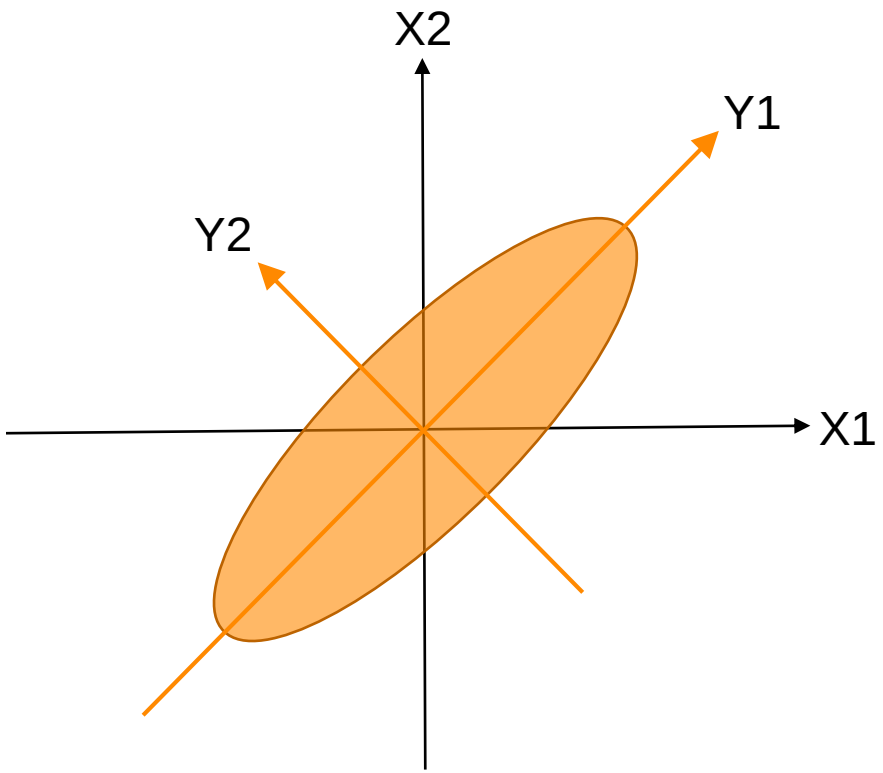
$$y = (29955 - 126x - 234x^2 - 45x^3 - 347x^4)/10$$

$T_0 = 29955$ $T_2 = 234$ $T_4 = 347$
$T_1 = 126$ $T_3 = 45$

Numerosity reduction

- **Technique of choosing smaller forms of data representation to reduce the volume of data**
- **Parametric**
  - Model (Regression, log-linear models etc) is used to estimate the data
  - The model parameters are stored instead of the actual data
- **Nonparametric**
  - Nonparametric methods are used for storing reduced representations of the data
  - Histograms, clustering, sampling

X1, X2: original axes
Y1, Y2: principal components



1. Order principal components by significance.
2. Eliminate weaker ones.

Principal component analysis (PCA)

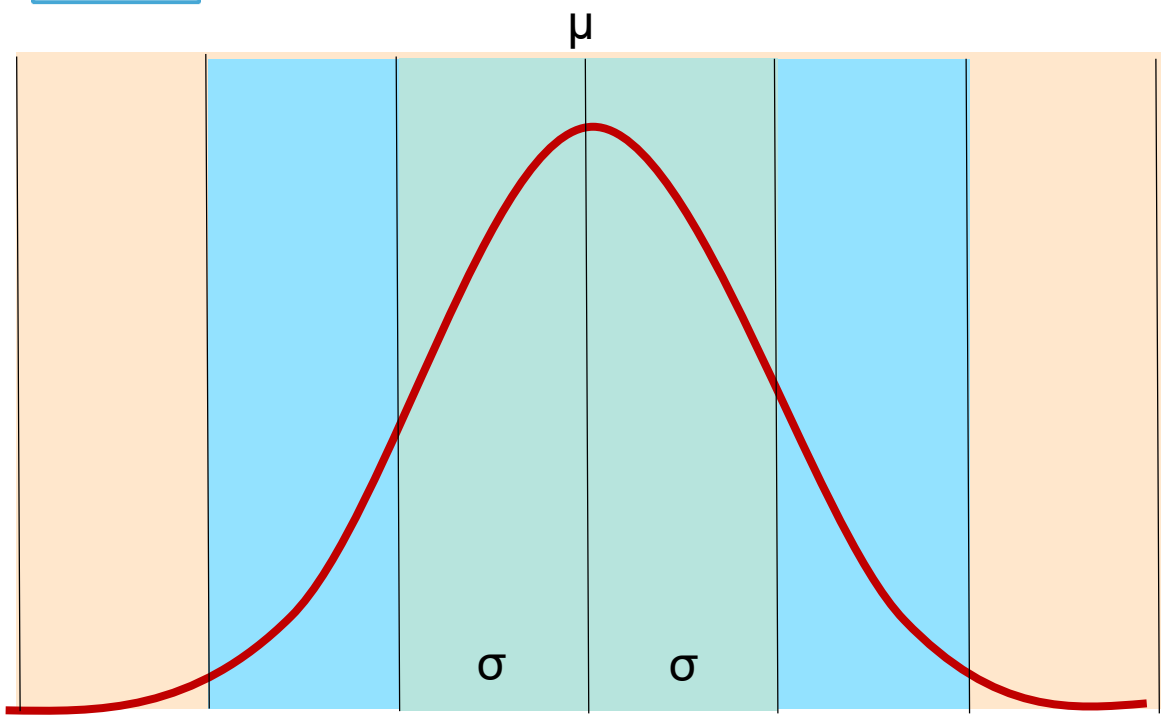
- Given *N* data vectors from *k*-dimensions, find *c* <= *k* orthogonal vectors that can be best used to represent data
- The original data set is reduced to one consisting of *N* data vectors on *c* principal components (reduced dimensions)
- Each data vector is a linear combination of the *c* principal component vectors
- Works for numeric data only
- Used when the number of dimensions is large
- Can help, when the model fails because of the multicollinearity of the variables
- The interpretation of the results gets more complicated
- An alternative: nonlinear principal component analysis (NLPCA)

Discrete Wavelet Transform

- **Linear signal processing technique**
- **Often used with very high dimensionality data**
- **Useful especially, when the number of variables is large and the number of observations is small**
- **Can be done to one sample at a time**
  - It is not necessary to put all samples together
- **For dimension reduction, the wavelet coefficients are selected according to their absolute values**
  - In PCA, the magnitude of the variance of each component is used for selection

Transformations to normality

Why normality matters?



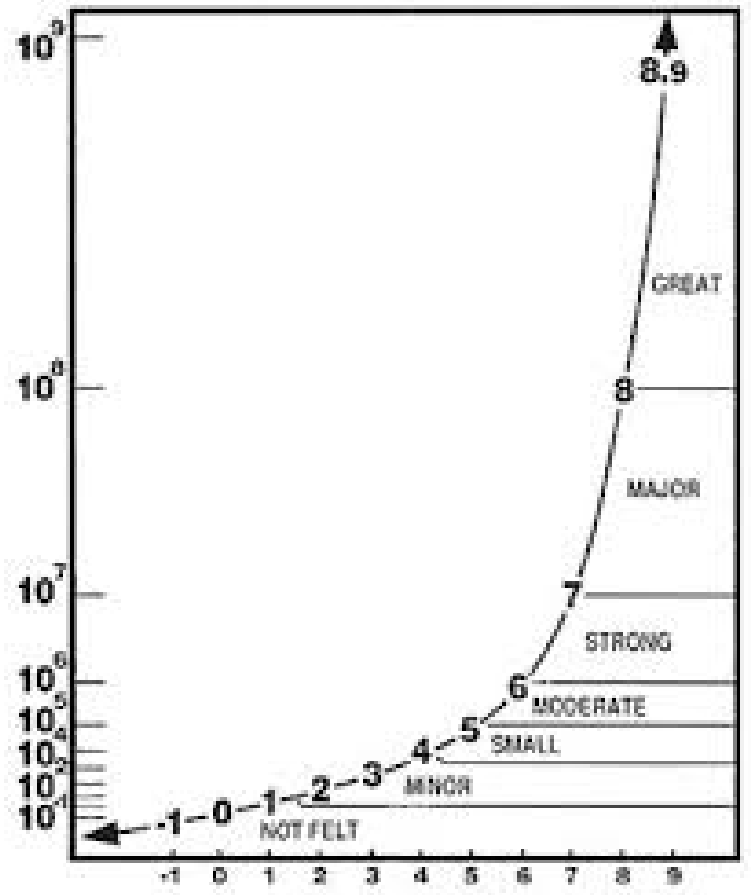
Normality may be required:

- **Many modelling methods and statistical tests have a normality assumption**

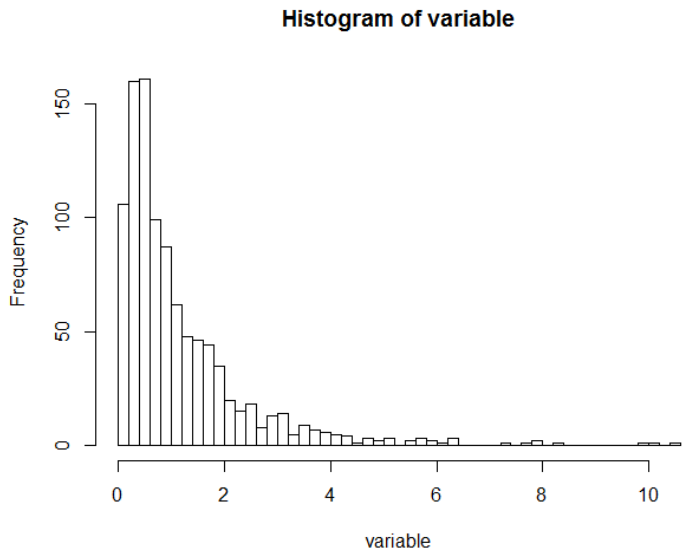
$$y = f(x) + \varepsilon, \varepsilon \sim N(0, \sigma_n^2)$$

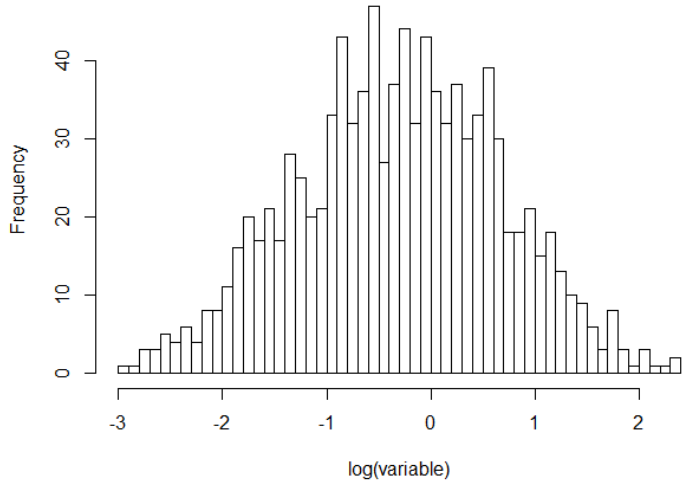
- **In real life, the data does not always follow Gaussian distribution**
- **Tests for normality**

Solutions for non-normality



- You can choose method that allows other distributions or use non-parametric methods
- You can use transformations to achieve normality
- Transformations may lead to more convenient way of representing information as well





Box-Cox transformation family

- **Box-Cox transformations are a family of useful transformations to achieve normality**

$$T(Y) = \frac{Y^{\lambda}-1}{\lambda}$$

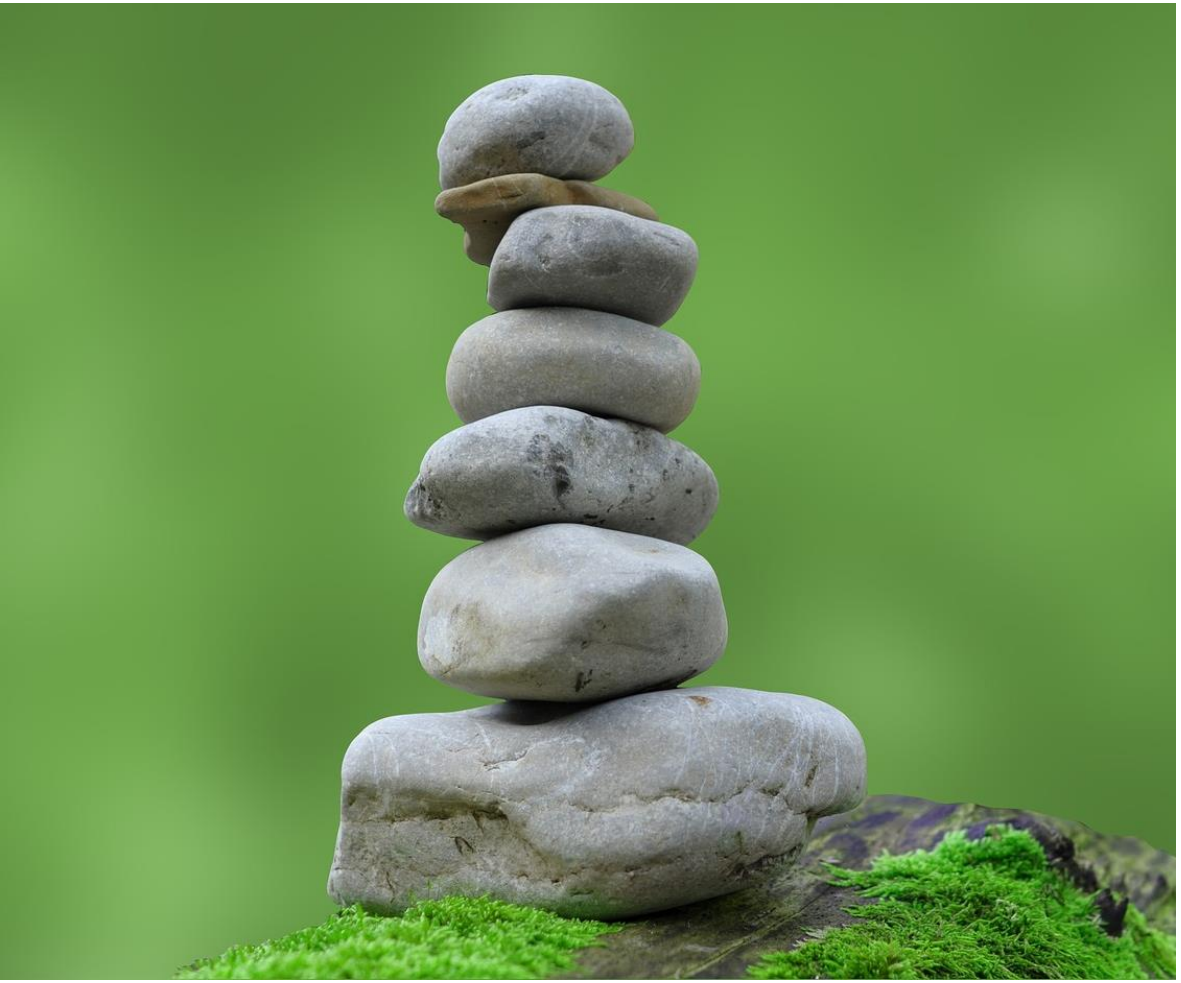
λ	-2	-1	-0.5	0	0.5	1	2
Transformation equation	$1/Y^2$	$1/Y$	$1/\sqrt{Y}$	$\ln Y$	$\sqrt{Y}$	Y	Y^2

- **B-C transformation is not a guarantee for normality**
- **B-C transformation works only for positive values**
  - By adding a suitable constant to a variable containing negative values, the data becomes positive

Summary

- What is normalization and why we need it
- Three methods
  - Min-max normalization
  - Standardization or z-score standardization
  - Decimal scaling
- What is classification
- How to classify
  - Median split
  - Distinctive groups
  - Equal lengths
  - Equal group sizes
- The advantages and disadvantages
- How to work with categorical variables
- Why data reduction is needed
- What is the goal of data reduction
- You can reduce
  - Number of variables
  - Number of variable values
  - Number of observations
- Methods
  - Data aggregation and generalization
  - Variable construction
  - Numerosity reduction
  - Principal component analysis (PCA)
  - Discrete wavelet transfrom
- Check your methods for assumptions
- Non-parametric methods
- Transformations for normality

Think further



Lecture 7 introduced 4 topics: data normalization, data classification, data reduction and transformations to normality.

- **Which of these methods was the most interesting to you?**
- **Which of these techniques you think you'll might need later?**

About the final exam

REGISTER TO THE EXAM AT https://exam.oulu.fi !

THE EXAM WILL BE AVAILABLE IN NOVEMBER 2020-JANUARY 2021.

More information about e-exam and exam visit in participating universities:

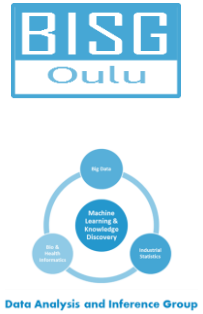
https://www.oulu.fi/forstudents/e-exam

EXAM:

- **You should be able to answer the exam questions with the provided materials, not forgetting listening to the lectures.**
- **The exams will be quite practical. Some essays about topics, maybe we will ask you to explain some concepts. Also, we may provide some data examples and you need to analyse them.**

Some examples of question types in the exam:

- **Explain terms briefly (1 p each)**
  - Open data
  - Data pollution
- **Small essays (2-3 p)**
  - Explain what can bias results when collecting data from humans.
- **Full essays (6 p)**
  - Introduce the characteristics that define data quality with examples?
- **Practical questions (2-4 p)**
  - Explain the intent of SQL queries and write down what they return

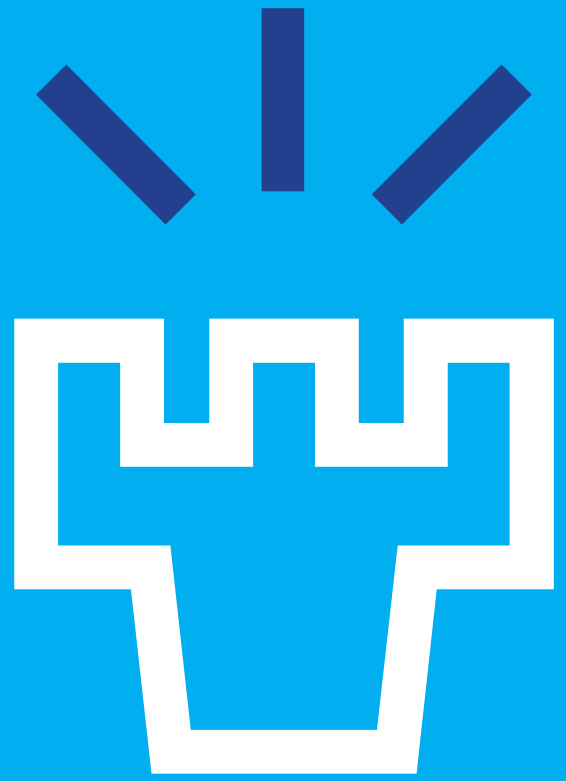


You can give your feedback for this course at

https://palaute.oulu.fi/

The feedback will be used for course development!

Thank you!



UNIVERSITY
OF OULU