
# Data analysis

Courses: Industrial automation, Automatizace pro průmyslovou praxi

CTU, FS, U12110

Matouš Cejnek


# Contents

- Introduction
- Basic flow
- Types of data
- Practical tools and examples


# Introduction

# Data analysis

- Data analysis involves inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.
- Basic flow:
  **Pre-processing → Core Analysis → Post-processing**

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

# Related tools, methods and fields

- Machine learning
- Statistics
- Data mining
- Natural language processing
- …


# Basic flow

# Data pre-processing

- **Handling Missing Data**
- **Data Cleaning** - removing outliers, errors
- **Data Transformation** - changing format, shape
- **Normalization** - data range adjustment
- **Feature Engineering**

**Pre-processing** → Core Analysis → Post-processing

# Core analysis

- Application of various **statistical methods, machine learning algorithms, or other analytical techniques** to extract meaningful insights and draw conclusions from the dataset.

Pre-processing → **Core Analysis** → Post-processing

# Core analysis

Most common:

- Regression
- Classification
- Clustering

Pre-processing → **Core Analysis** → Post-processing

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

# Post-processing

- Interpretation of results
- Visualization
- Validation (cross-validation)
- Documentation

Pre-processing → Core Analysis → **Post-processing**


# Types of data

# Types of data

- Qualitative (Categorical) Data
- Quantitative (Numerical) Data
- Mixed (Categorical and Numerical) Data

# Qualitative (Categorical) Data

Qualitative data represents categories or labels and is non-numeric. It describes qualities or characteristics and cannot be measured in numerical terms.

**Nominal Data**: (e.g., colors, types of fruits).

**Ordinal Data**: (e.g., education levels).

# Quantitative (Numerical) Data

Quantitative data consists of measurable quantities and is expressed in numerical terms. It represents quantities or amounts and can be further categorized as discrete or continuous.

**Discrete Data** (e.g., number of employees, customer purchases).

**Continuous Data** (e.g., height, temperature).

# Mixed (Categorical and Numerical) Data

Definition: Some datasets may include both qualitative and quantitative variables, leading to mixed data types.

Example: A survey dataset with both categorical variables (gender, education level) and numerical variables (income, age).


# Practical tools and examples

# Common metrics for numerical data

We need to calculate different metrics as a step for various other tasks (filtration, normalization, cleaning, modelling, …)

**Number of observations** (length of data)

**Min, max,** and **range** (max - min) of data

**Sum of data** (total value)

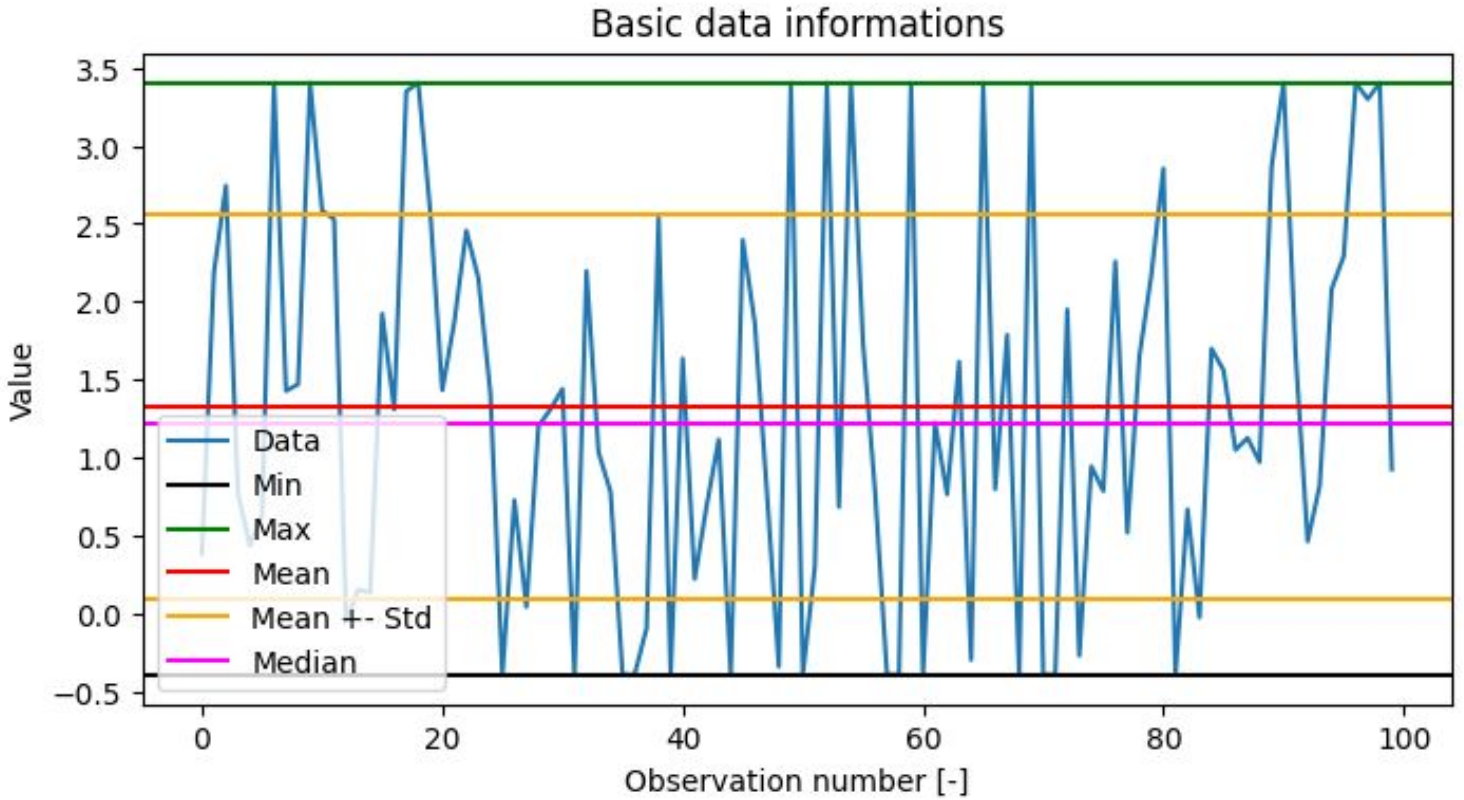
# Common metrics for numerical data

**Mean value** = sum / number of observations

**Median value** (most common value, the middle value of ordered data)

**Variance** - how the data are spread from mean

**Standard deviation** - square root of variance


Basic data informations
Value
Observation number [-]
Data
Min
Max
Mean
Mean +- Std
Median
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0
20
40
60
80
100

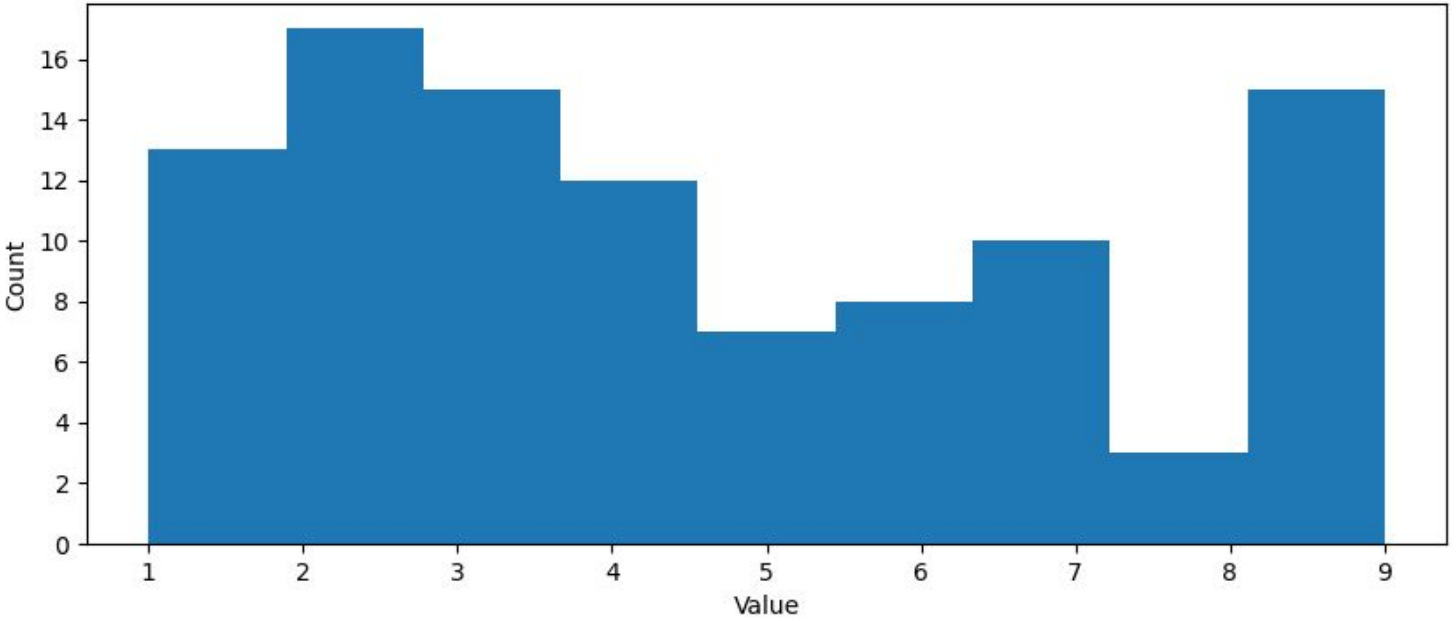
CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

# Histogram

- A histogram is a visual representation of the distribution of a dataset, displaying the frequency or probability of different ranges or bins of numerical data.
- The horizontal axis (x-axis) represents the numerical data, divided into intervals or bins.
- The vertical axis (y-axis) represents the frequency or probability of occurrences within each bin.

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

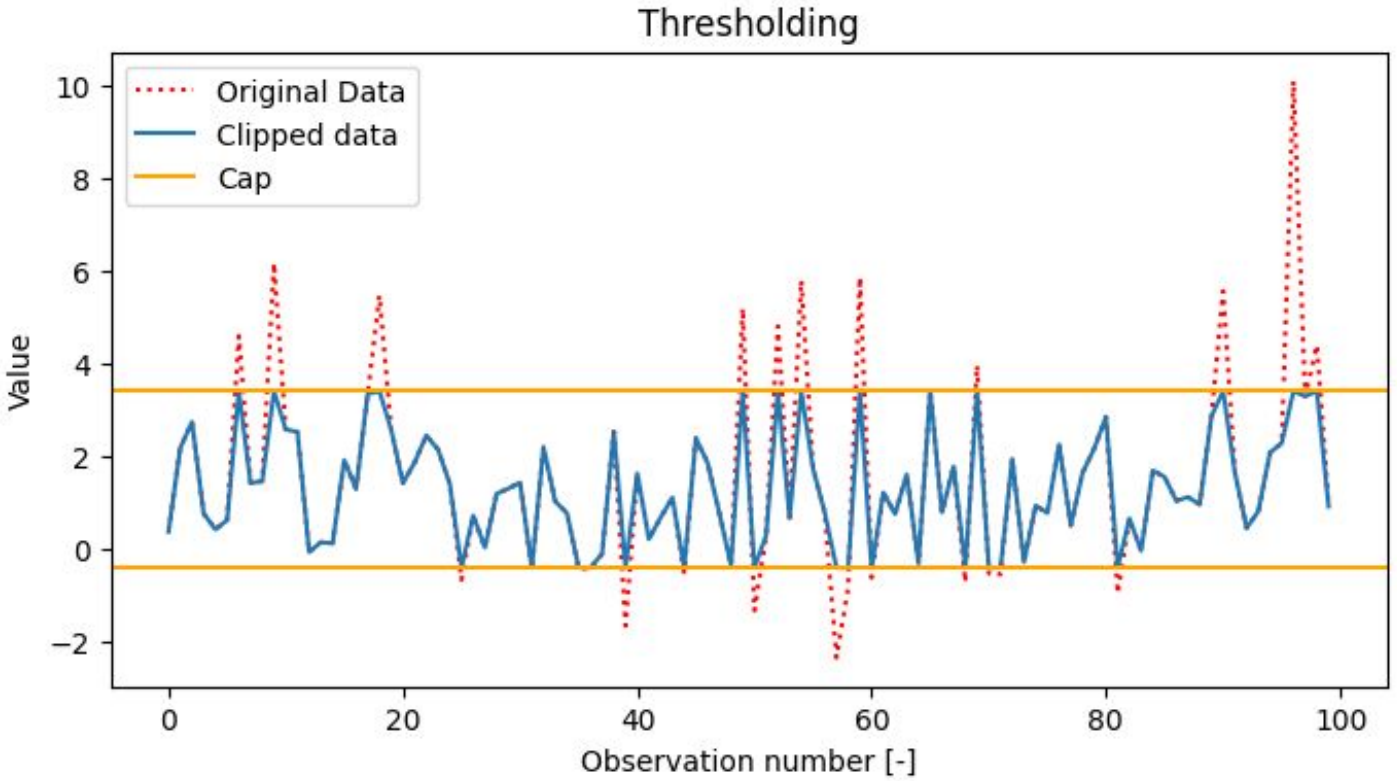
# Histogram

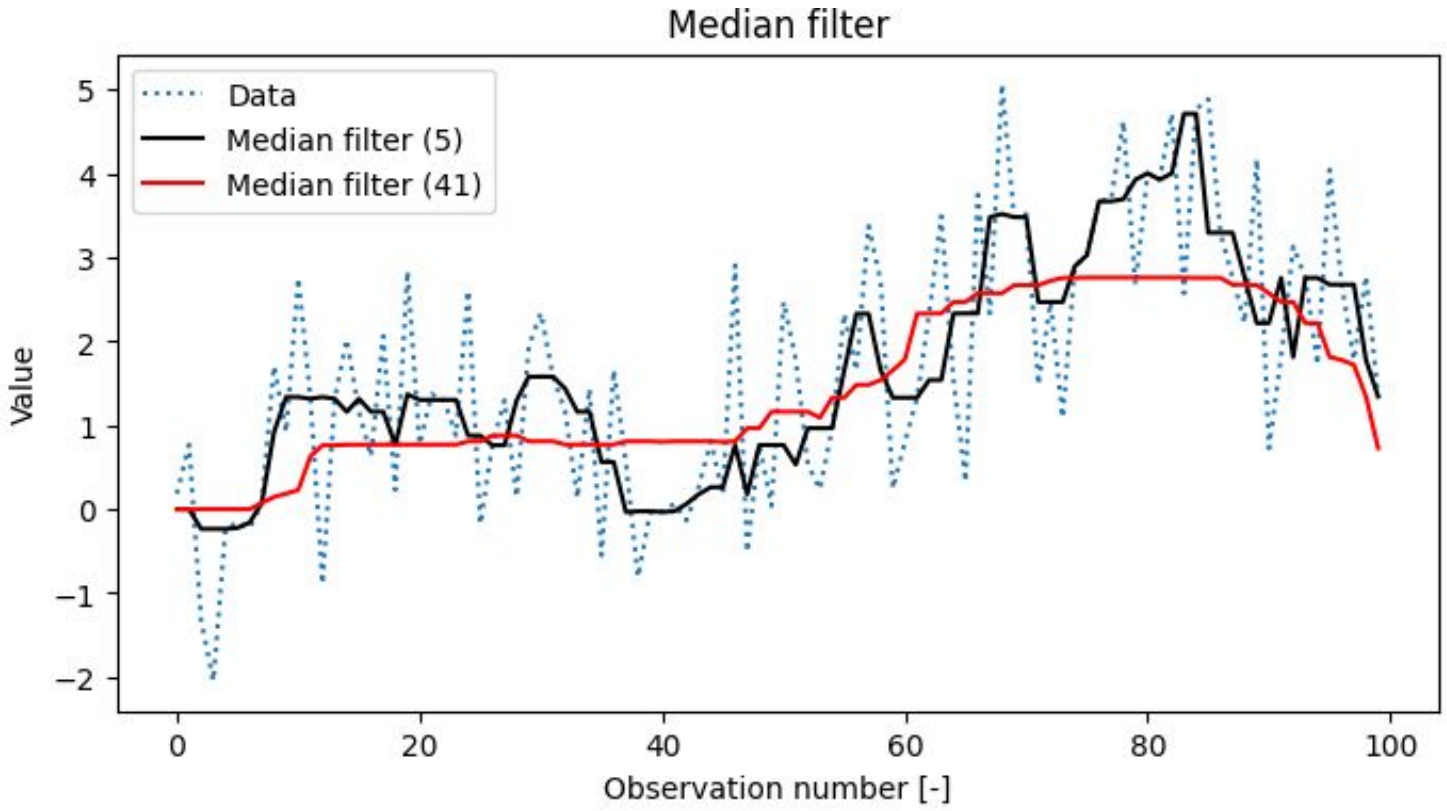


# Basic cleaning / filtering methods

**Thresholding -** removing/capping values on below or above a threshold (good for removing outliers)

**Mean / median filter** - replacing values with mean or median of defined surrounding (e.g. last few samples)

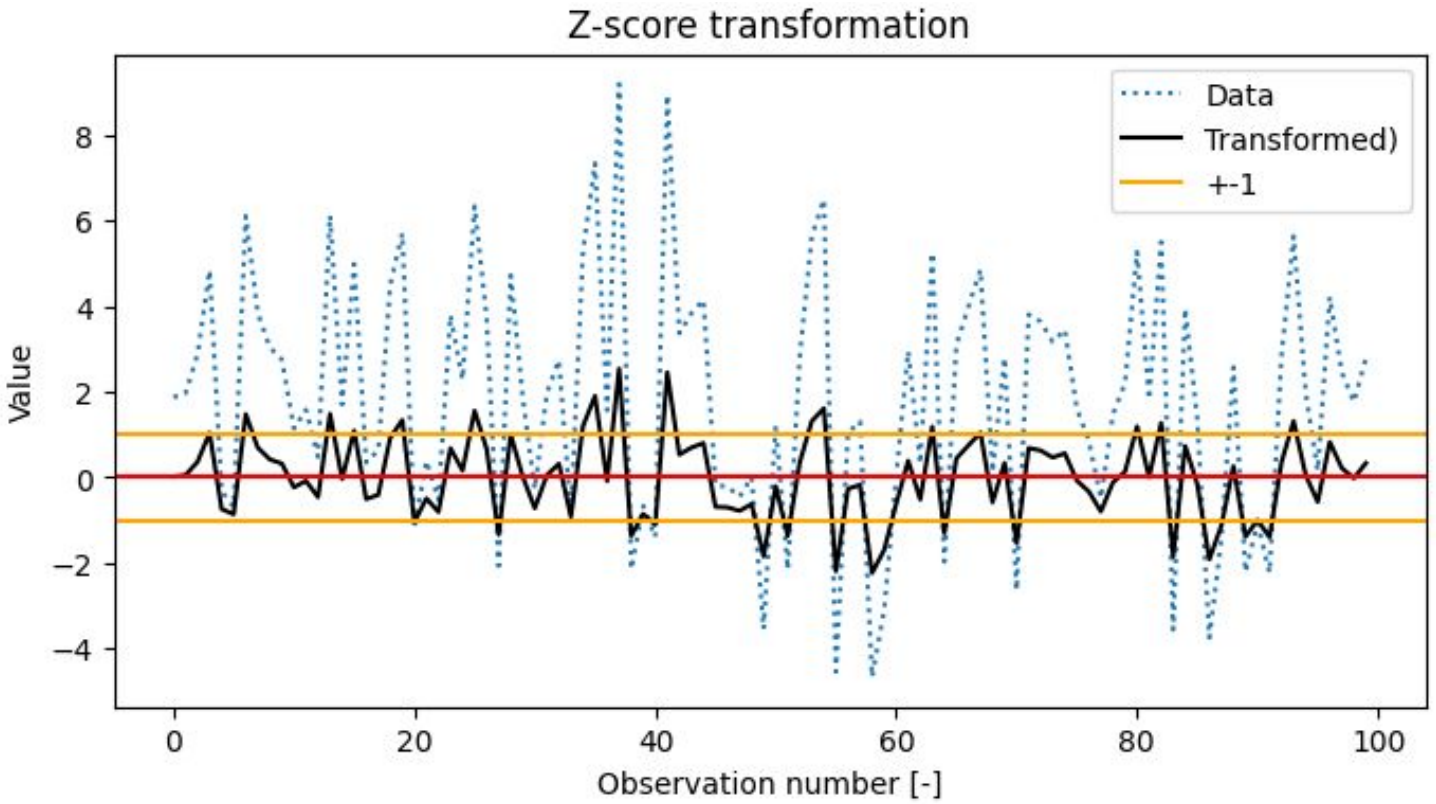

Thresholding
Original Data
Clipped data
Cap
Value
Observation number [-]
10
8
6
4
2
0
−2
0
20
40
60
80
100


Median filter
Data
Median filter (5)
Median filter (41)
Value
Observation number [-]
5
4
3
2
1
0
−1
−2
0
20
40
60
80
100

# Z-score (standard score)

Used for normalization of data into standard range.

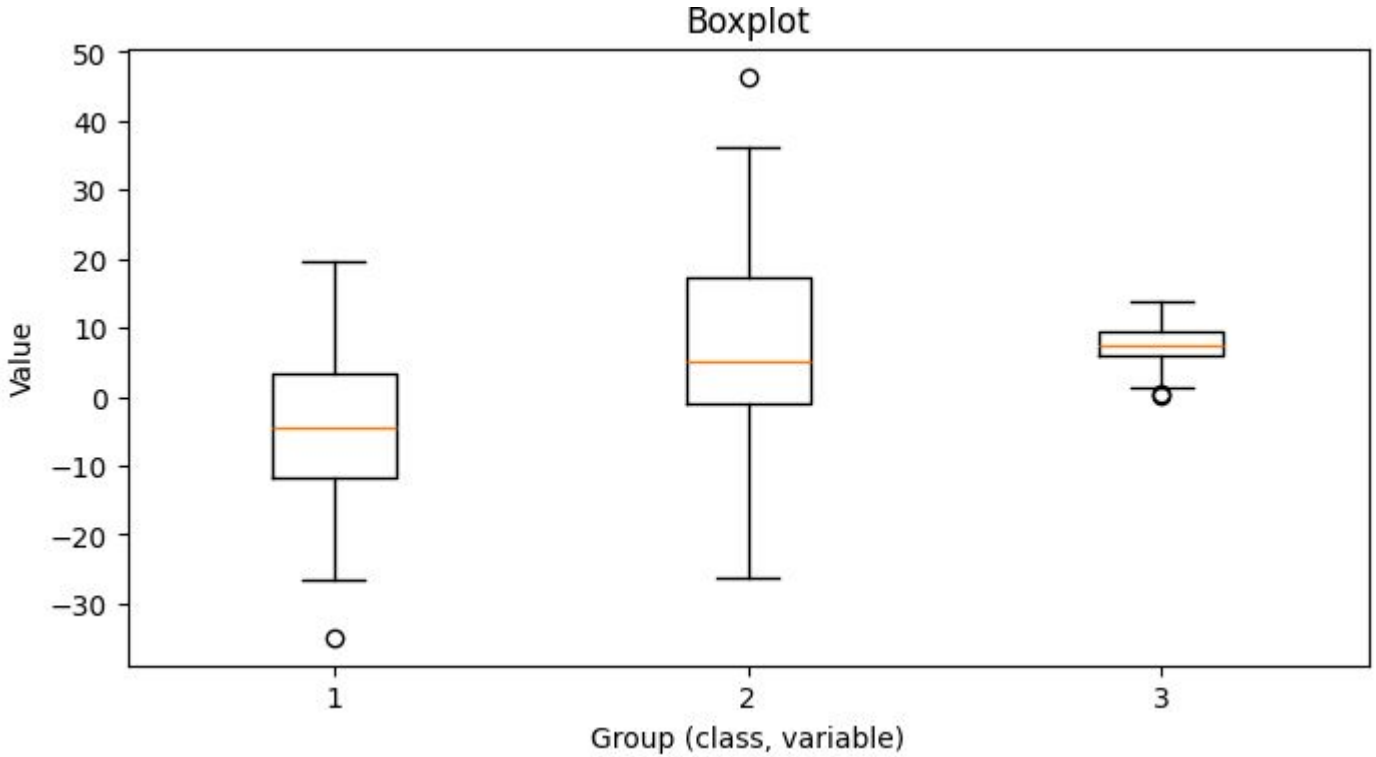
1. Remove mean value
2. Divide by standard deviation


Z-score transformation
Data
Transformed)
+-1
Value
Observation number [-]
8
6
4
2
0
−2
−4
0
20
40
60
80
100


# Boxplot

Boxplot is used to visualize multiple variables in a way that displays (for each):

- Median value
- Quartiles
- Outliers


Boxplot
50
40
30
20
10
0
−10
−20
−30
Value
1
2
3
Group (class, variable)

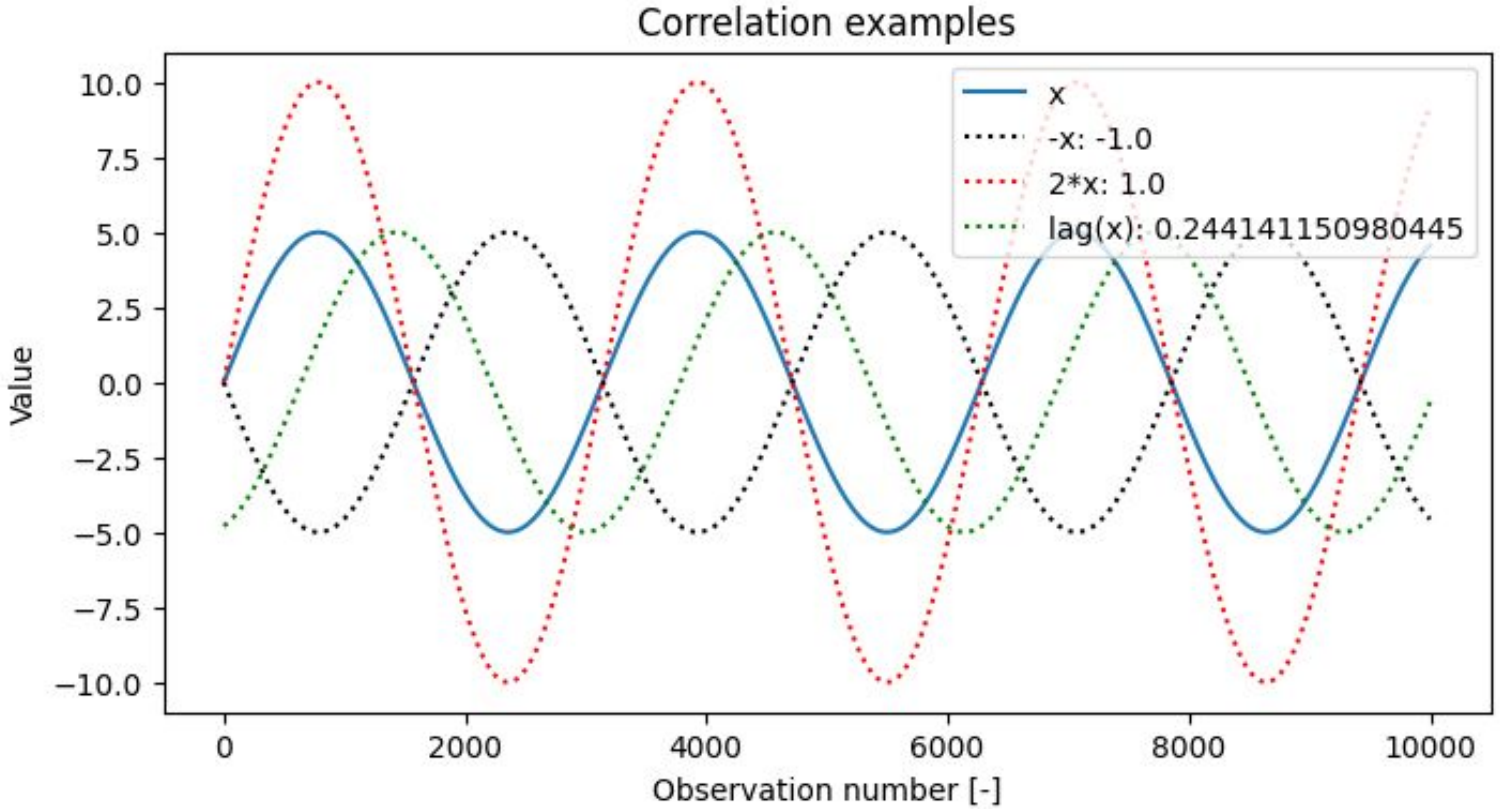
# Correlation coefficient

The correlation coefficient quantifies the strength and direction of a linear relationship between two numerical variables:

1: Perfect Positive Linear Relationship

0: No Linear Relationship

-1: Perfect Negative Linear Relationship

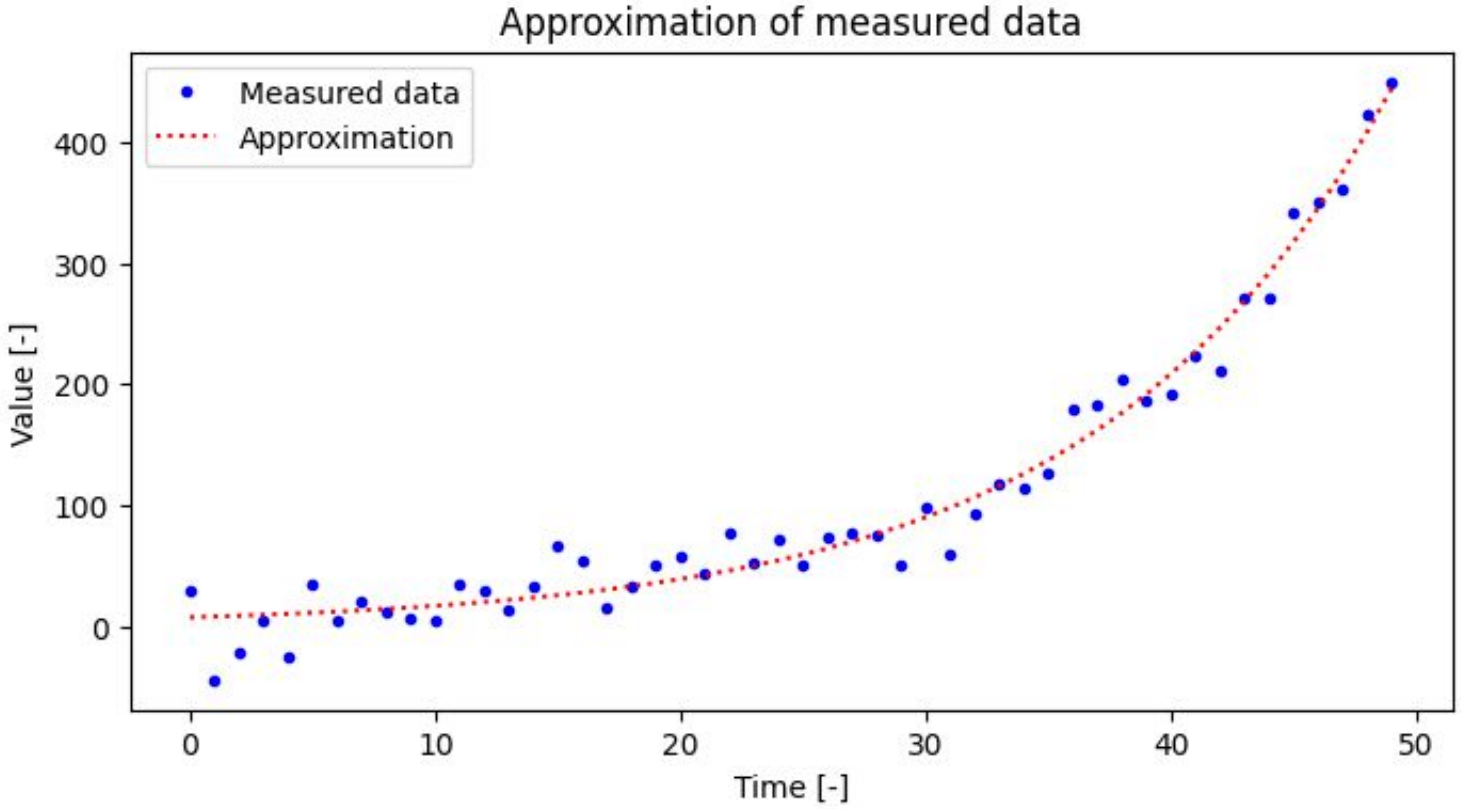

Correlation examples
x
-x: -1.0
2*x: 1.0
lag(x): 0.244141150980445
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
Value
0
2000
4000
6000
8000
10000
Observation number [-]

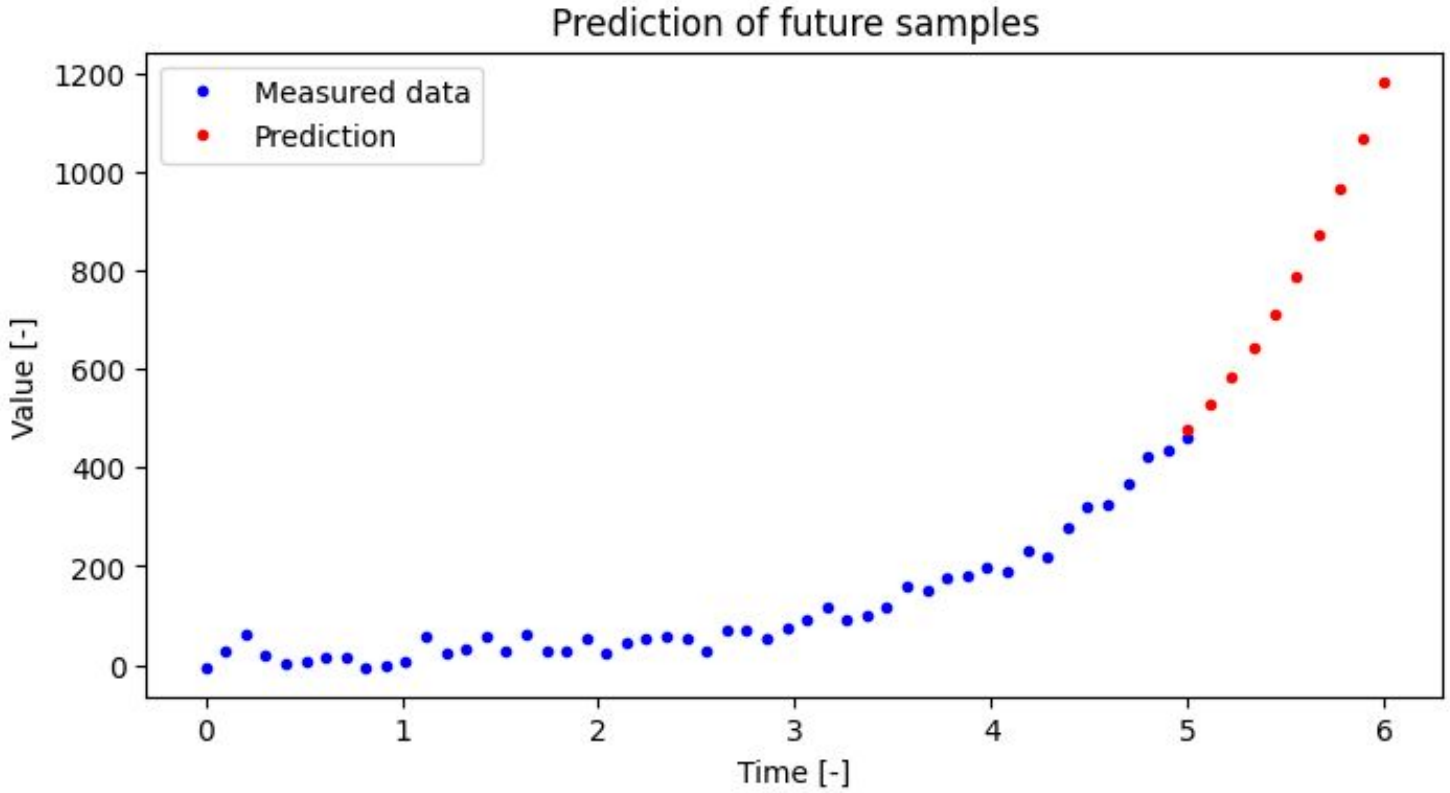
# Regression analysis

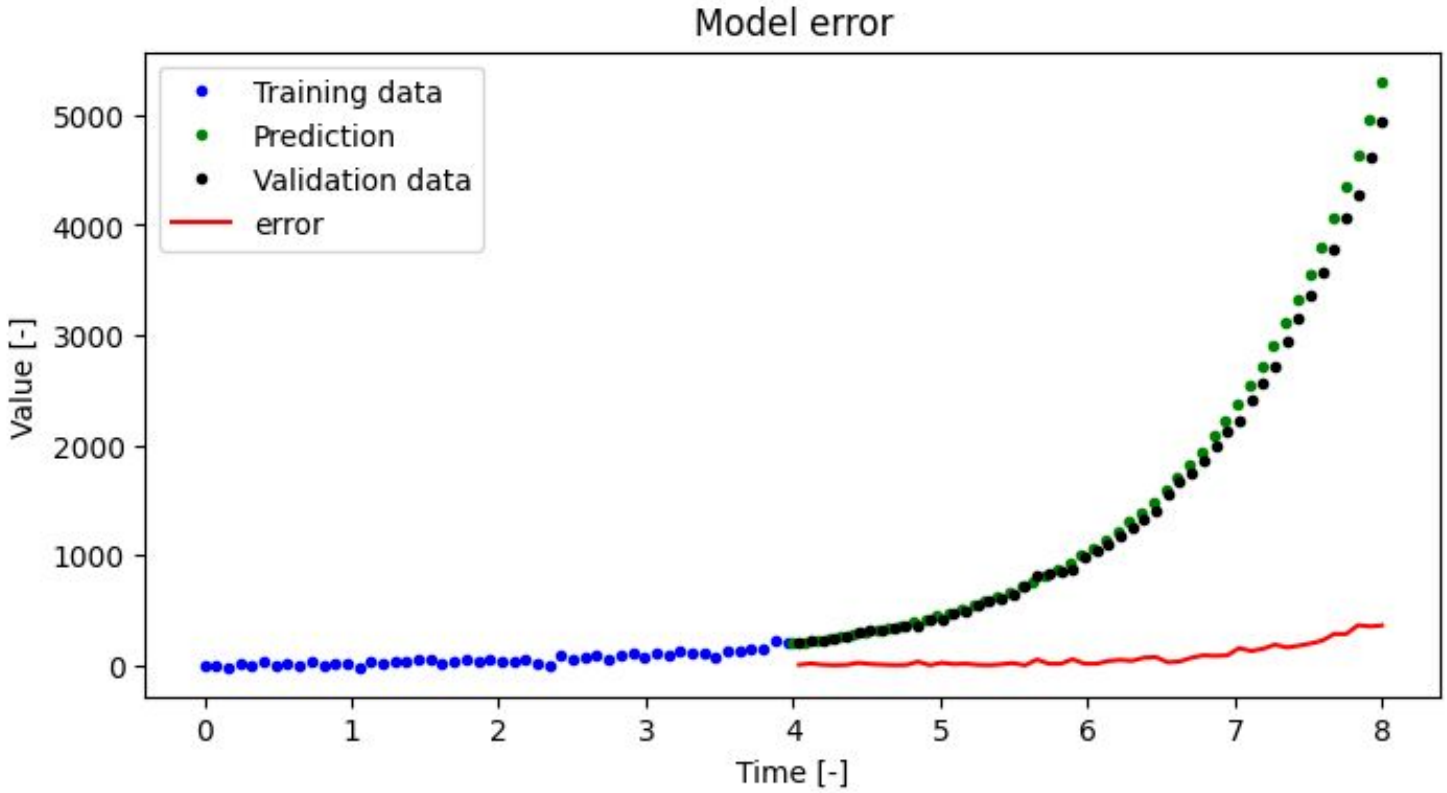
Regression analysis is a statistical method used to model and examine the relationship between a dependent variable and one or more independent variables.

**For example:**

What is the relation between measured value and time?


Approximation of measured data
Measured data
Approximation
Value [-]
Time [-]
400
300
200
100
0
0
10
20
30
40
50


Prediction of future samples
Measured data
Prediction
Value [-]
Time [-]
0
200
400
600
800
1000
1200
0
1
2
3
4
5
6


Model error
Training data
Prediction
Validation data
error
Value [-]
Time [-]
0
1000
2000
3000
4000
5000
0
1
2
3
4
5
6
7
8