MACHINE LEARNING

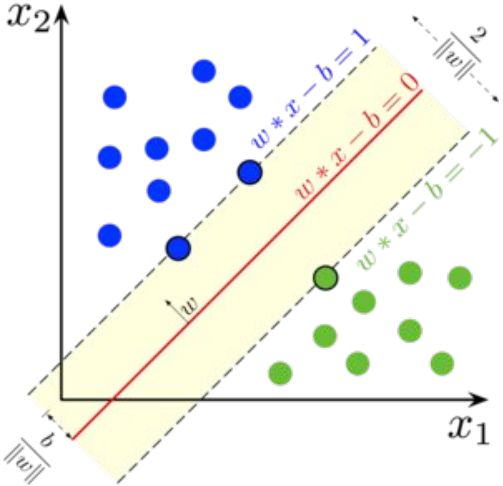
Module 1

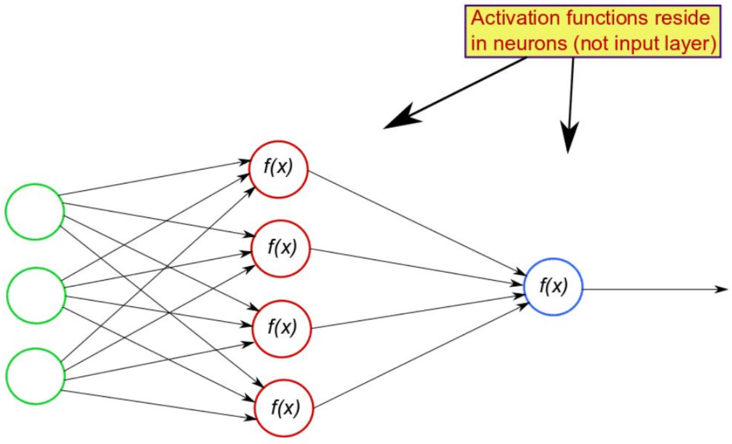
Exploratory Data Analysis

(*a gentle introduction*)

Universal Approximation Theorem - [Kurt Hornik & George Cybenko]

- Artificial Neural Networks:
  - non linearity
  - limits
  - continue



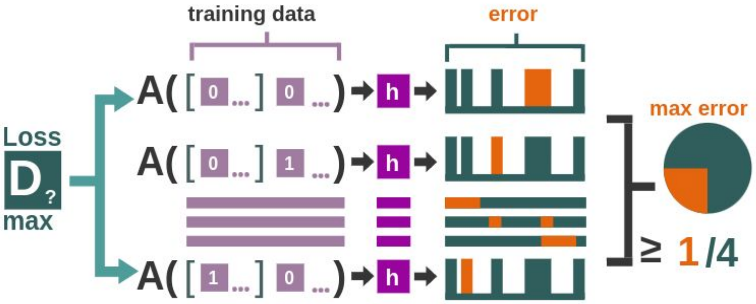


- Support Vector Machine
  - Gaussian Kernel
  -

- ... others (Decision Tree, Polynomials)

No Free Lunch Theorem - D. H. Wolpert & W. G. Macready

- The Core Concept
- Implications in ML
- Diversity of ML Problems
- Algorithm Selection
- Final Reflection



No Free Lunch Theorems for Optimization

David H. Wolpert
IBM Almaden Research Center
N5Na/D3
650 Harry Road
San Jose, CA 95120-6099

William G. Macready
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM, 87501

December 31, 1996

Abstract

A framework is developed to explore the connection between effective optimization algorithms and the problems they are solving. A number of "no free lunch" (NFL) theorems are presented that establish that for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class. These theorems result in a geometric interpretation of what it means for an algorithm to be well suited to an optimization problem. Applications of the NFL theorems to information theoretic aspects of optimization and benchmark measures of performance are also presented. Other issues addressed are time-varying optimization problems and *a priori* "head-to-head" minimax distinctions between optimization algorithms, distinctions that can obtain despite the NFL theorems' enforcing of a type of uniformity over all algorithms.

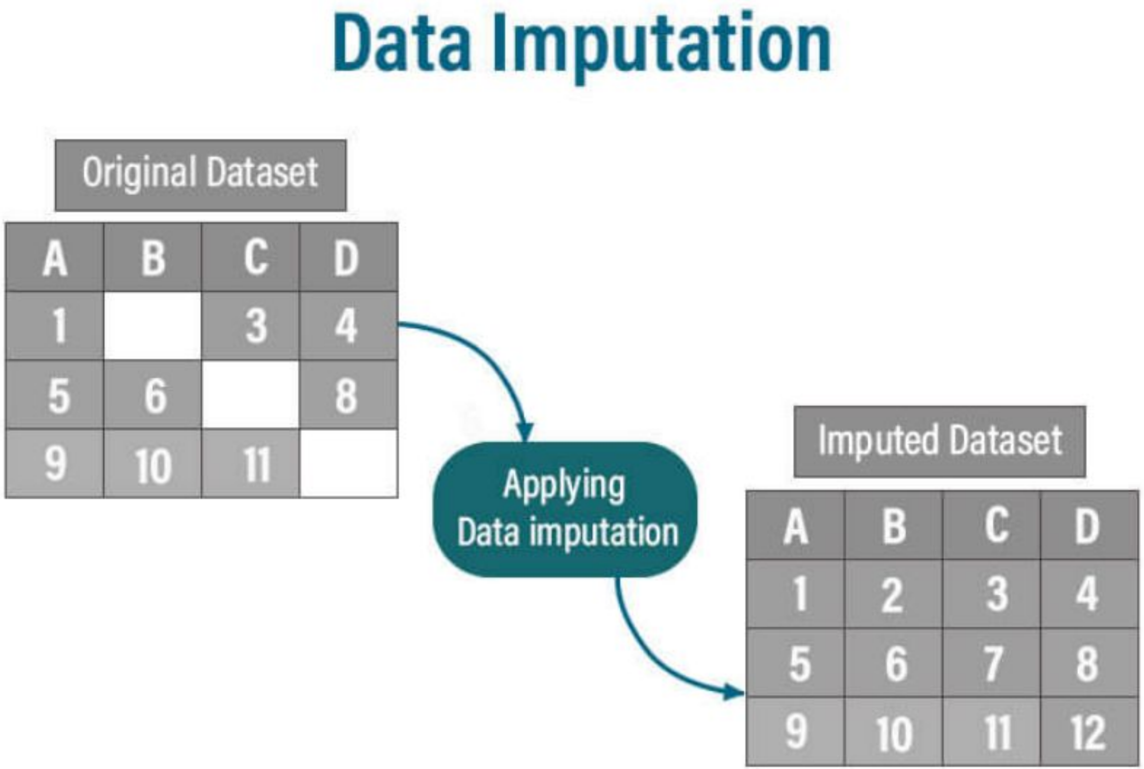
1 Introduction

The past few decades have seen increased interest in general-purpose "black-box" optimization algorithms that exploit little if any knowledge concerning the optimization problem on which they are run. In large part these algorithms have drawn inspiration from optimization processes that occur in nature. In particular, the two most popular black-box optimization strategies, evolutionary algorithms [FOW66, Hol93] and simulated annealing [KGV83] mimic processes in natural selection and statistical mechanics respectively.

1

Module 1: Imputation

- Missing Values
- Why it is important
    - data retention
    - Performance improving
    - data preparation
- Different Strategies
    - univariate, multivariate
    - statistical, MICE, ...



Module 1: Imputation

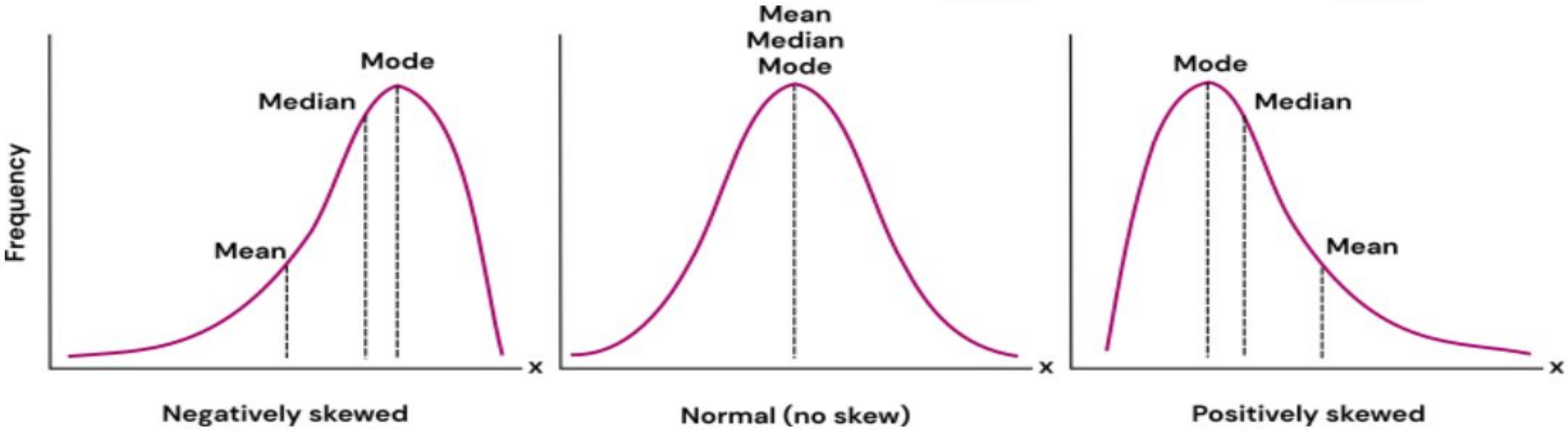
- mean/median/mode

Price
100
90
50
40
20
100
60
120
200

Mean = 86.66

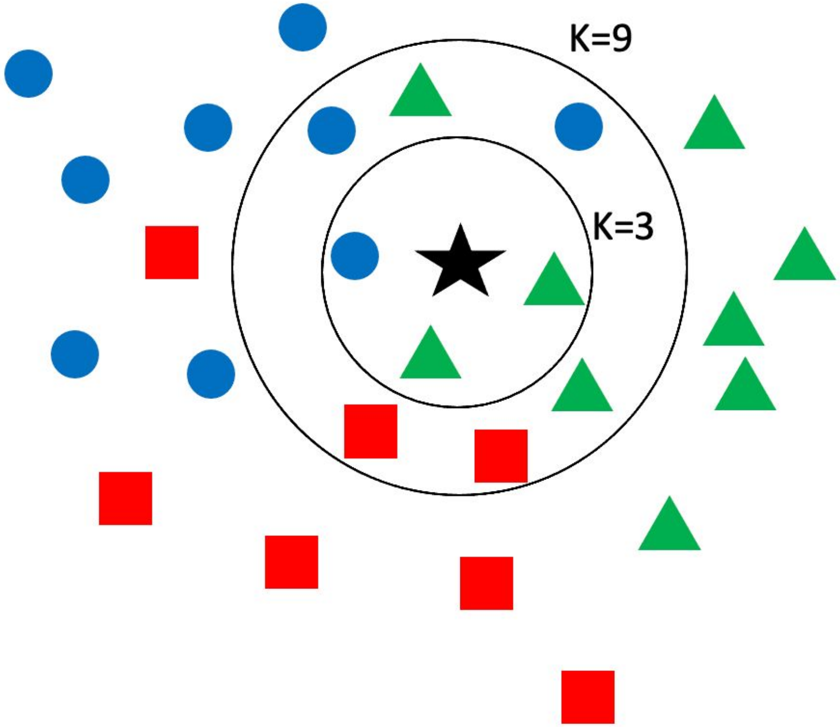
Median = 90

Price
100
90
50
40
20
100
86.66
60
120
86.66
200



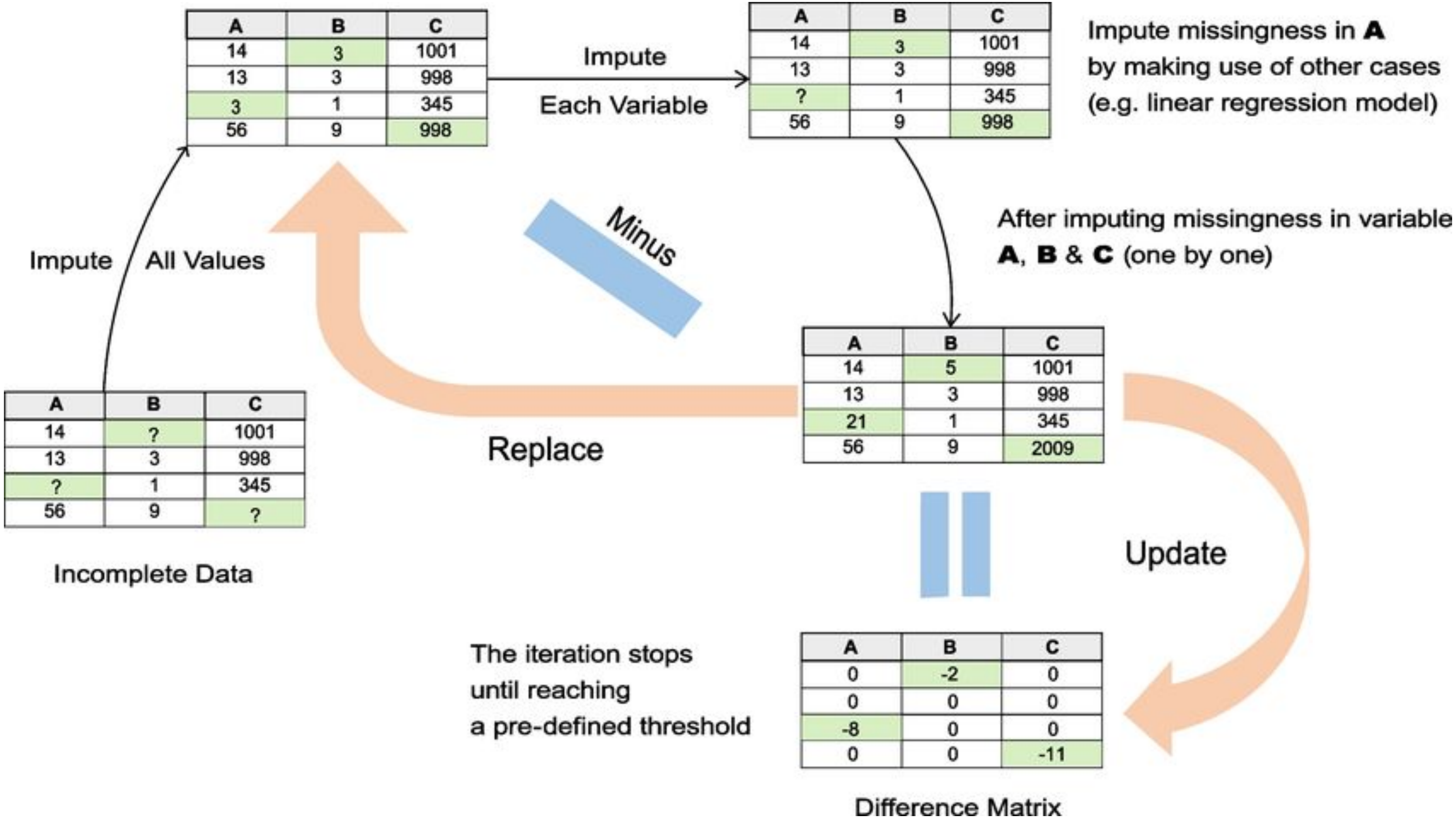
Module 1: Imputation

- Identification of Neighbours
- Selection of k-neighbours
- Missing values estimation
- Repetition for All Samples



Imputation

Multivariate Imputation by Chained Equations (MICE)



A	B	C
14	3	1001
13	3	998
3	1	345
56	9	998

A	B	C
14	3	1001
13	3	998
?	1	345
56	9	998

A	B	C
14	5	1001
13	3	998
21	1	345
56	9	2009

A	B	C
14	?	1001
13	3	998
?	1	345
56	9	?

A	B	C
0	-2	0
0	0	0
-8	0	0
0	0	-11

Imputation *SoftImpute* (Singular Value Decomposition)

- Matrix Decomposition
- Dimensional Reduction
- Regularization
- Imputation
- Iterate Process

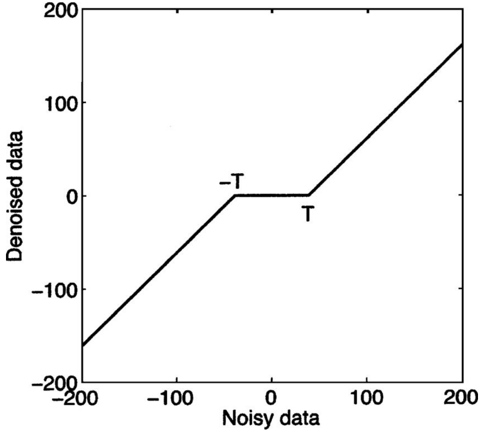
$$\mathbf{A} = \begin{bmatrix} 1 & -2-i & 5 \\ 1+i & i & 4-2i \end{bmatrix}$$

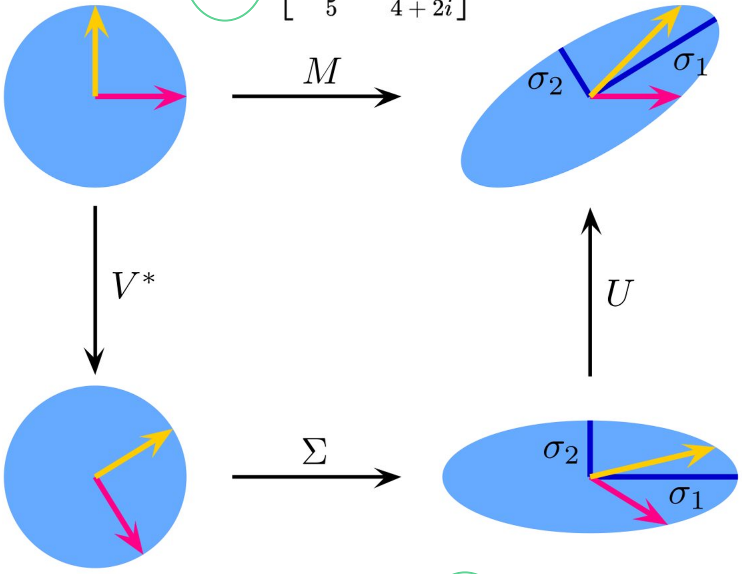
We first transpose the matrix:

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 1 & 1+i \\ -2-i & i \\ 5 & 4-2i \end{bmatrix}$$

Then we conjugate every entry of the matrix:

$$\mathbf{A}^{\mathrm{H}} = \begin{bmatrix} 1 & 1-i \\ -2+i & -i \\ 5 & 4+2i \end{bmatrix}$$

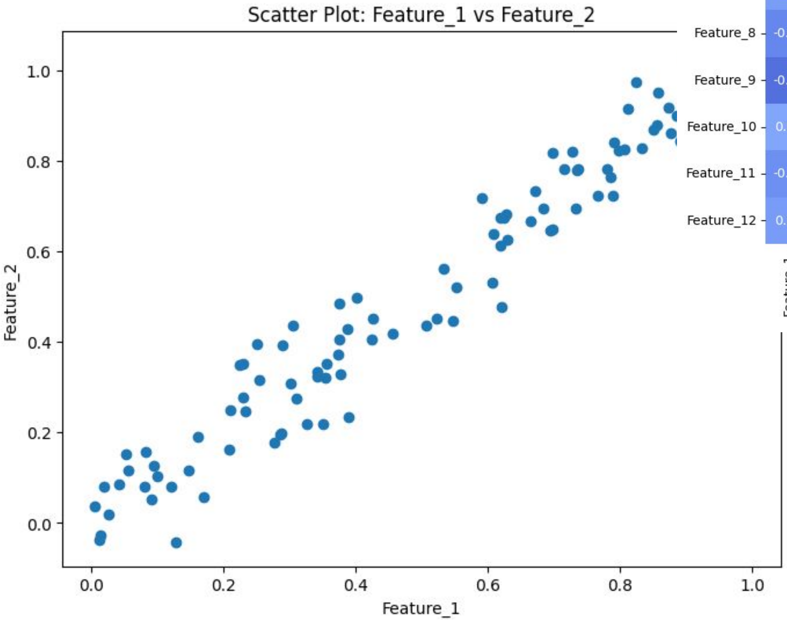


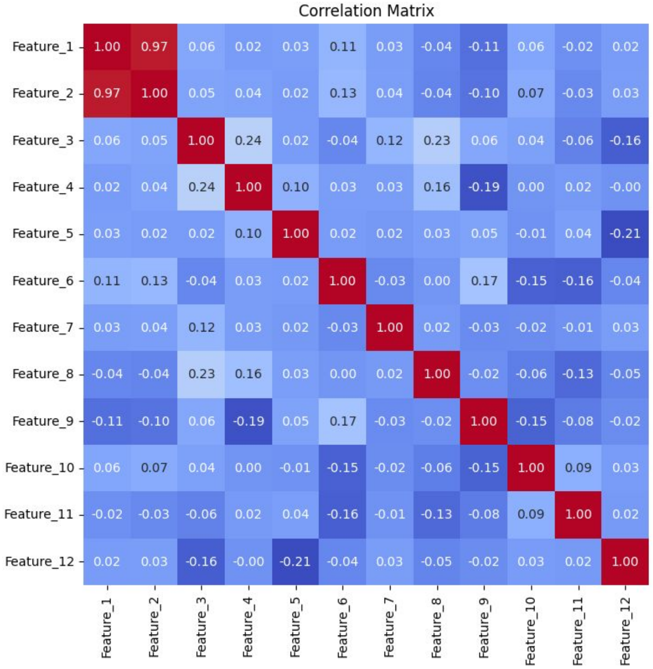


$$M = U \cdot \Sigma \cdot V^*$$

Module 1: Data Correlation

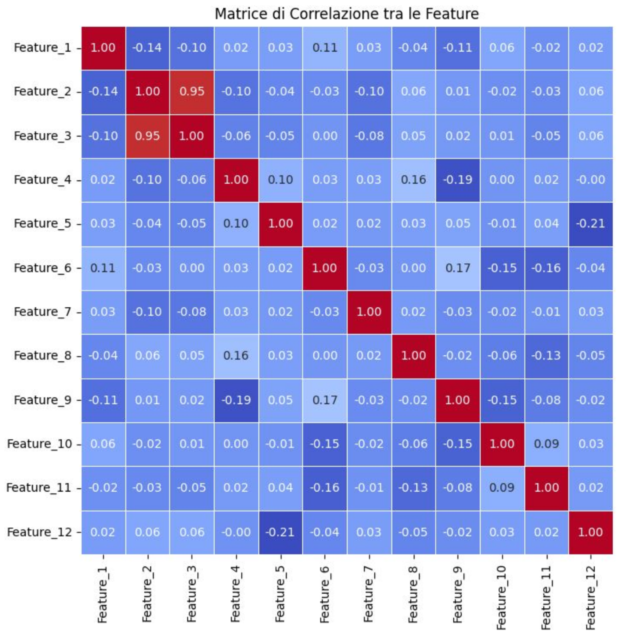
- ***Feature Selection***
- Multicollinearity Prevention
- Feature Engineering
- Noise Reduction

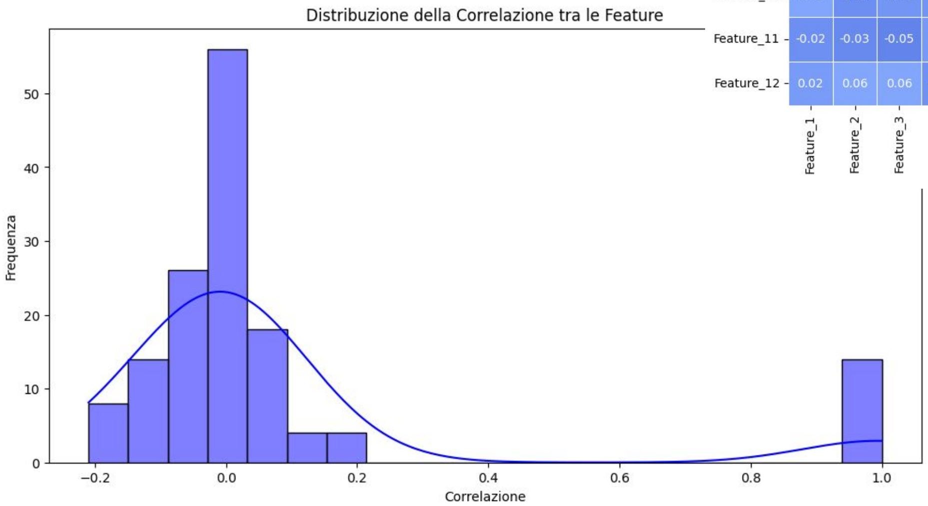




Module 1: Data Correlation

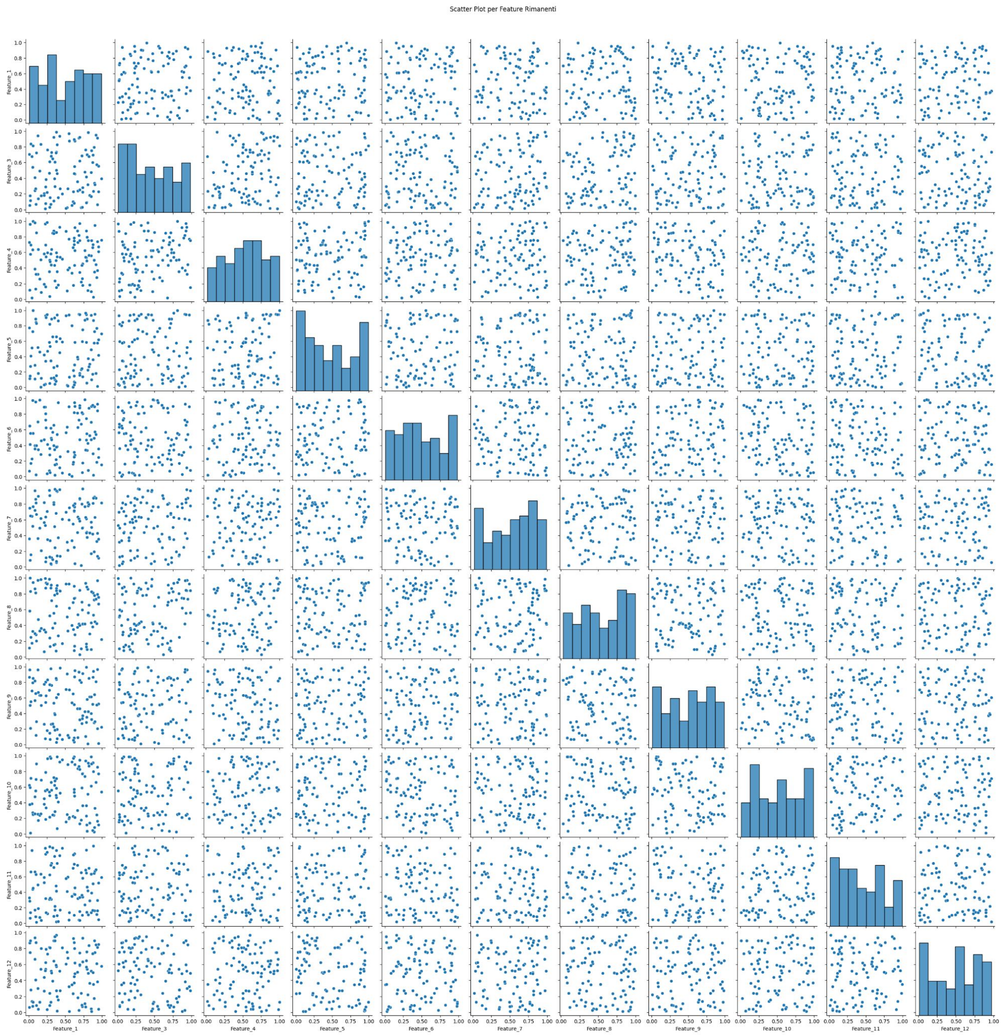
- Feature Selection
- ***Multicollinearity Prevention***
- Feature Engineering
- Noise Reduction





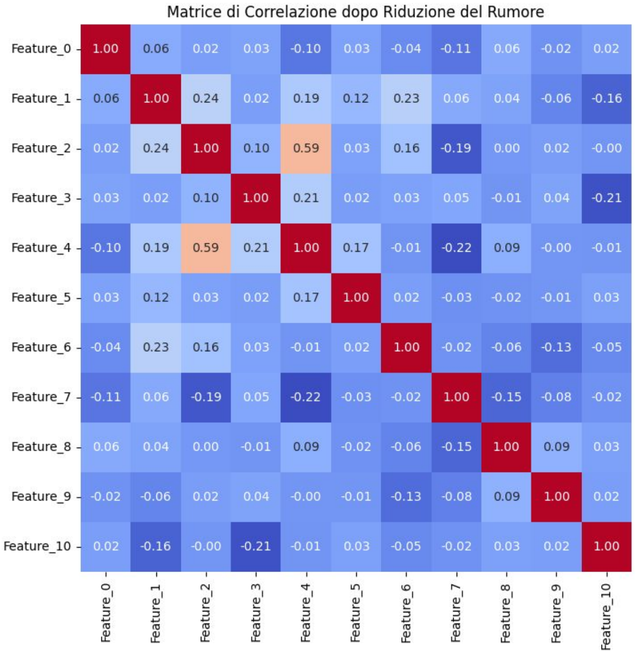
Module 1: Data Correlation

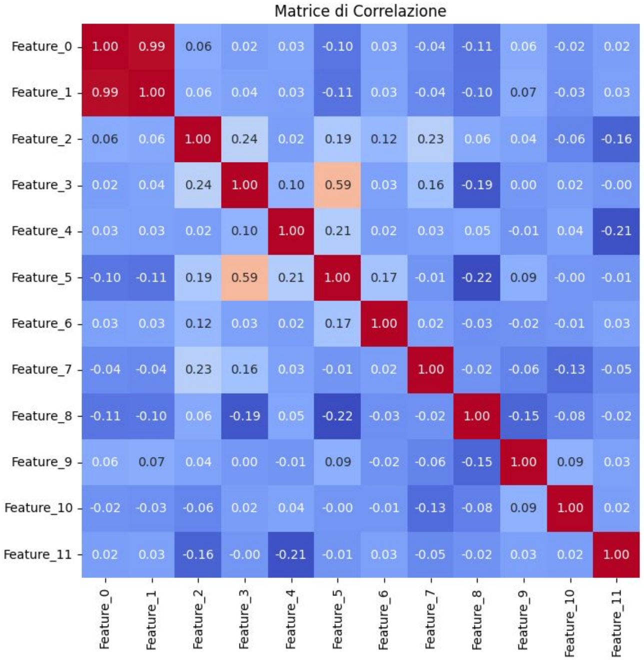
- Feature Selection
- Multicollinearity Prevention
- ***Feature Engineering***
- Noise Reduction



Module 1: Data Correlation

- Feature Selection
- Multicollinearity Prevention
- Feature Engineering
- ***Noise Reduction***





Module 1: Data Distribution

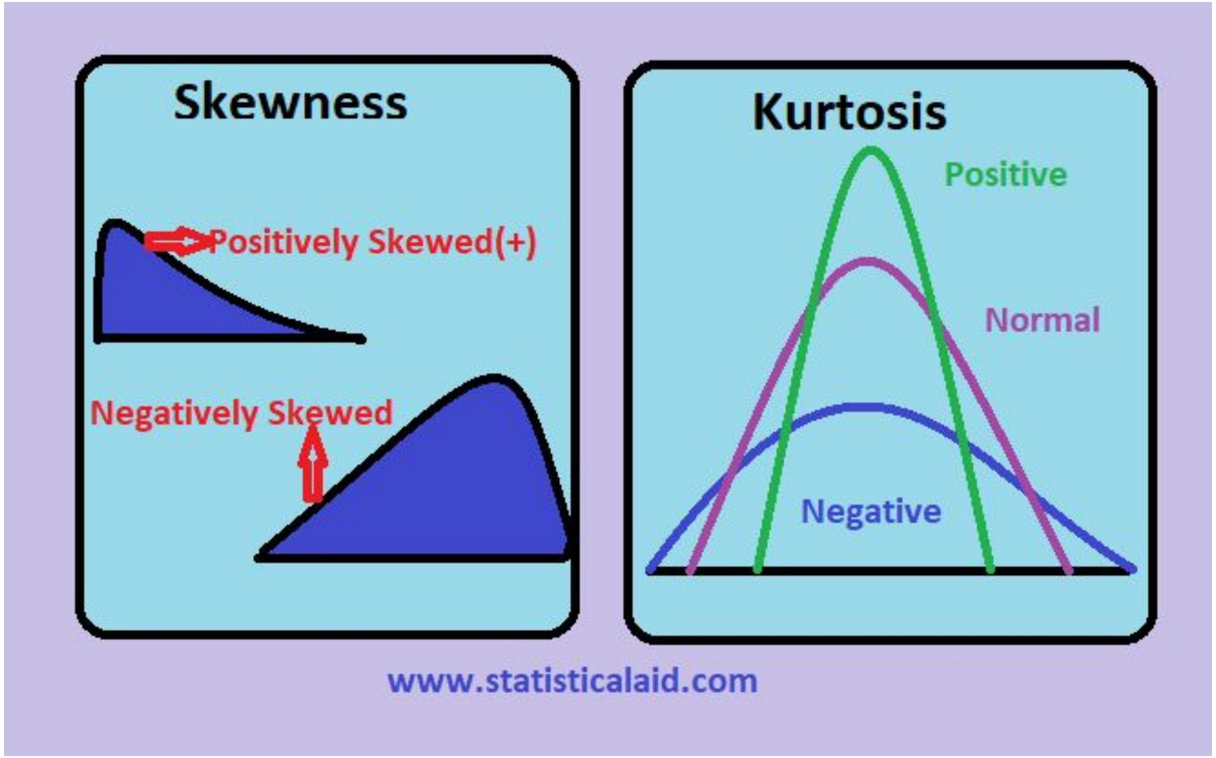
- Data Information
  - Normal Distribution
  - Skew Distribution
  - Kutosis

- ***Multicollinearity***

  Prevention

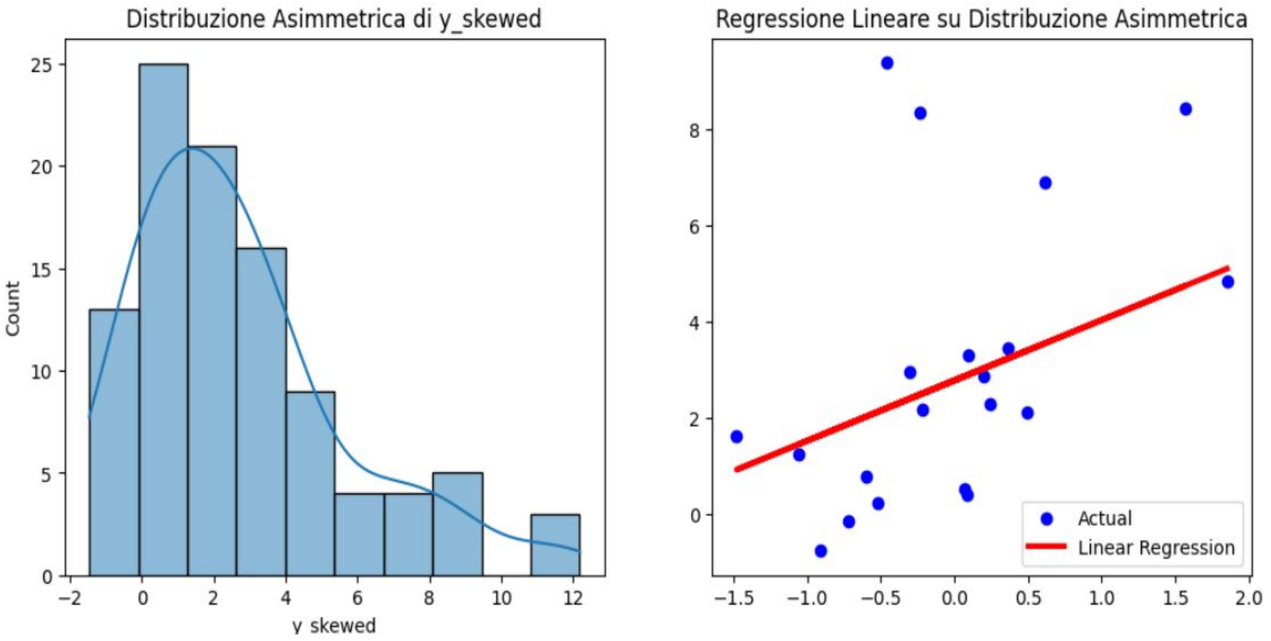
- Feature Engineering

- Noise Reduction



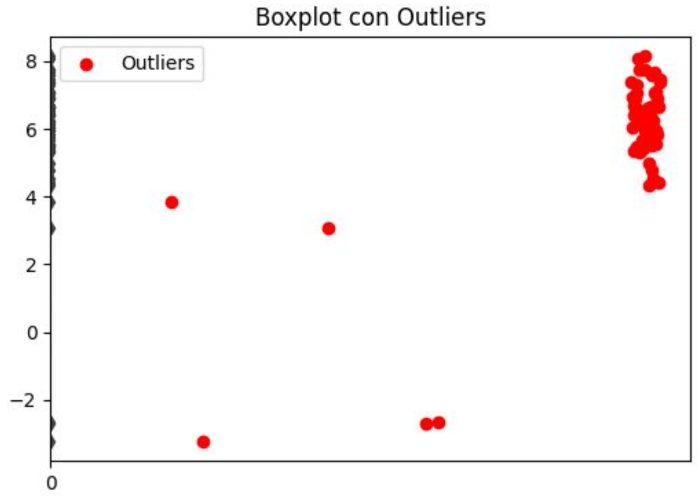
Module 1: Data Distribution

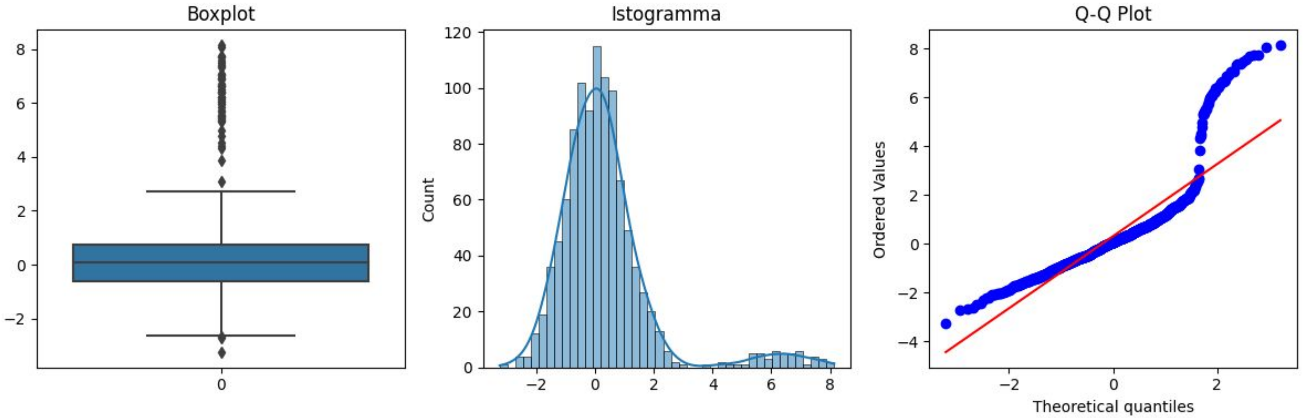
- Data Information
- **Model Choosing**
  - linear regression
- Outliers Detection
- Feature Selection



Module 1: Outliers

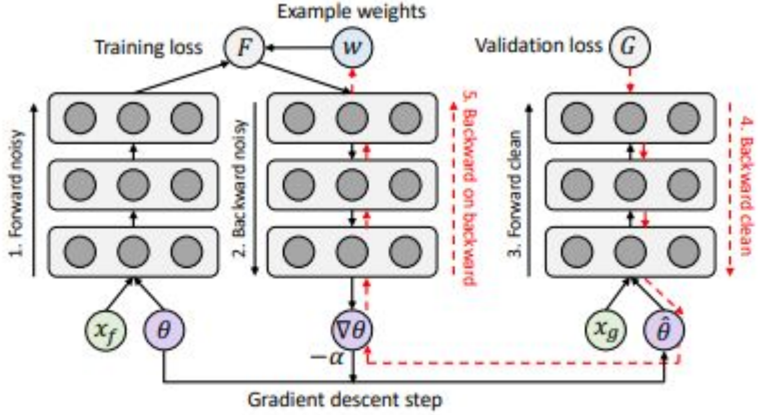
- Data Information
- Model Choosing
- **Outliers Detection**
- Feature Selection

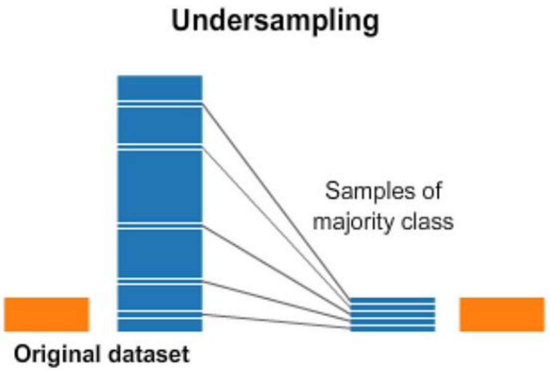


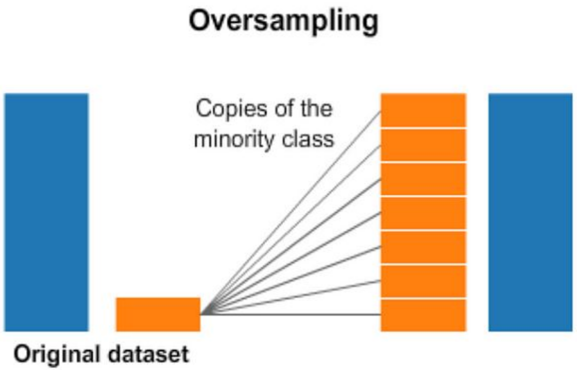


Module 1: Dataset Balancing

- Reweighting
- Resampling
  - Undersampling
  - oversampling

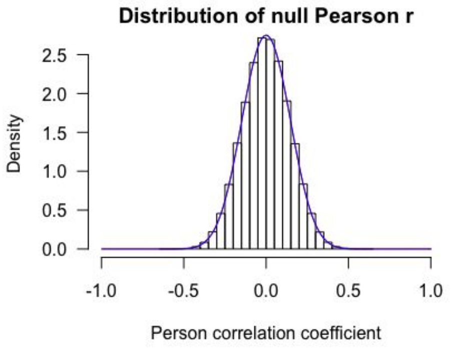




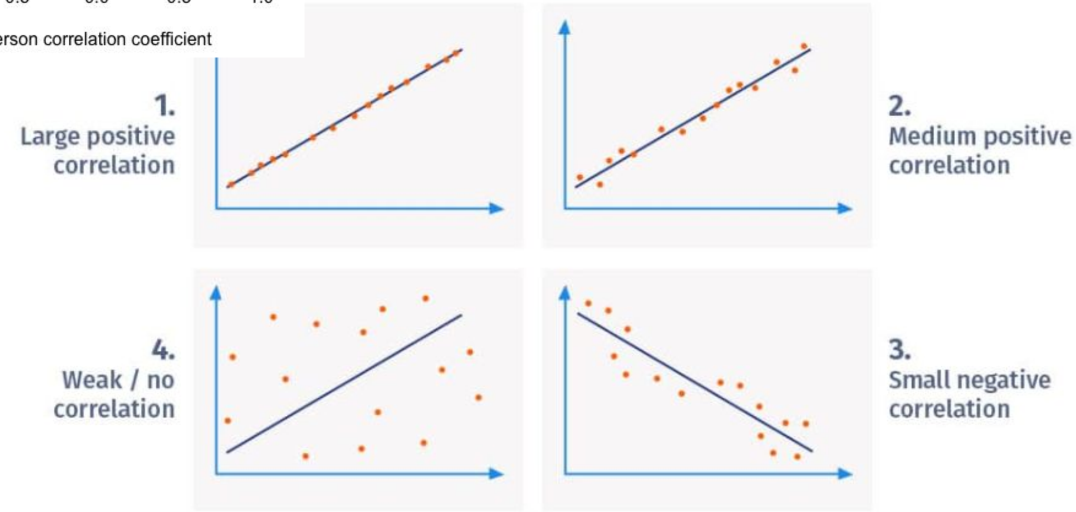


Module 1: … *again* Data Correlation

- **Pearson**
  - **linear correlation**
  - **Normal Distribution**
  - **Linear Scalability**
  - **Without Outliers**



$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$



- Spearman

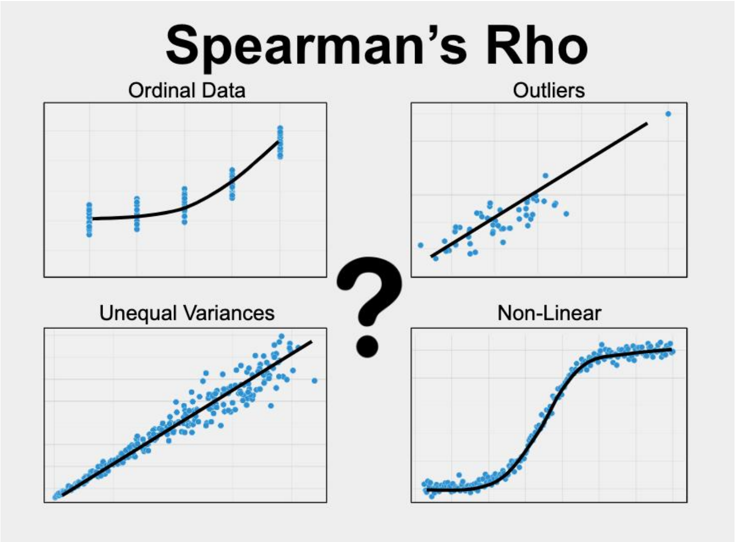
- Kendall

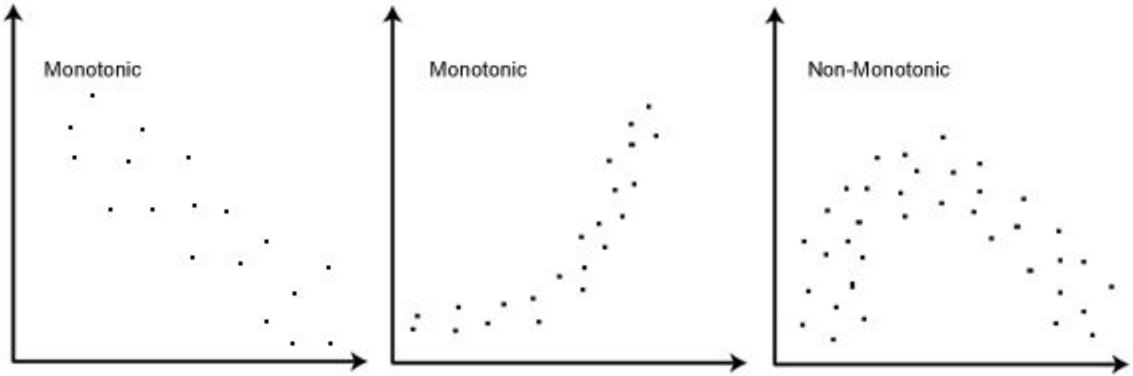
Module 1: ... *again* Data Correlation

- Pearson

- **Spearman**
  - **Ordinals**
  - **Non-linear**
  - **With Outliers**
  - **No Normal Distribution**
  - **Monotone Relation**

- Kendall





Module 1: ... *again* Data Correlation

- Pearson

- Spearman

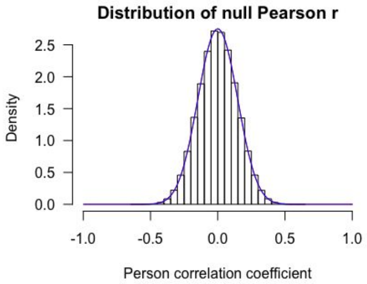
- **Kendall**
  - **Non Normal Distribution**
  - **Robustness to Outliers**
  - **Ordinal Representation**

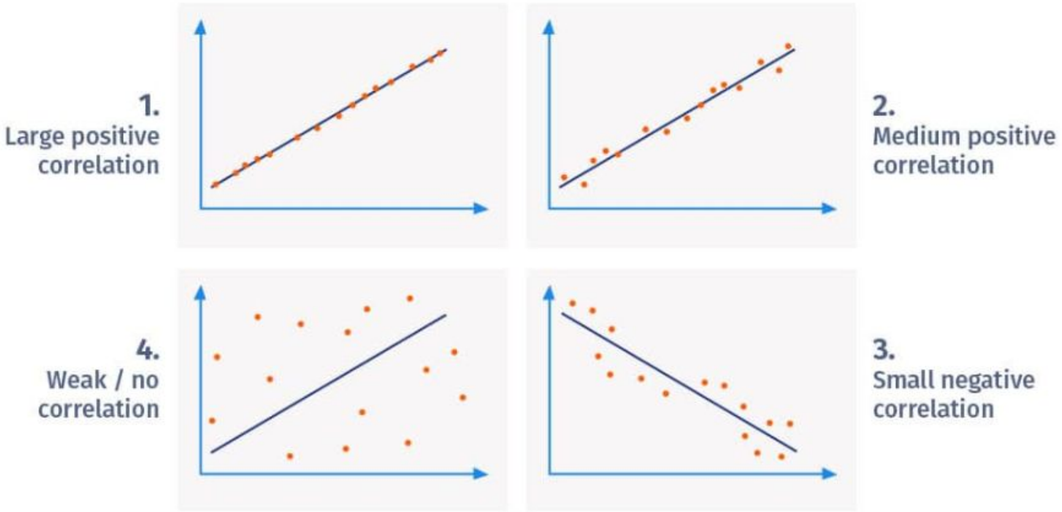
$$\tau = \frac{R}{\sqrt{\frac{(n^2 - n - K_x)(n^2 - n - K_y)}{4}}}$$

Module 1: … *again* Data Correlation

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

- **Pearson**
  - **linear correlation**
  - **Normal Distribution**
  - **Linear Scalability**
  - **Without Outliers**



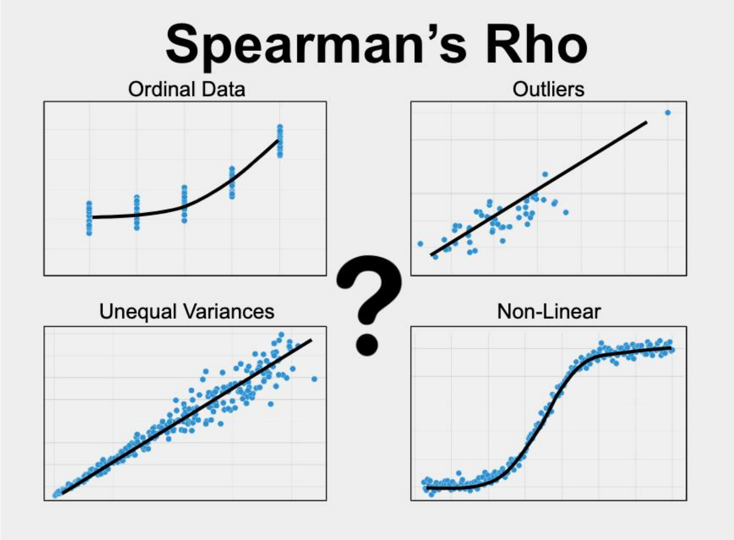


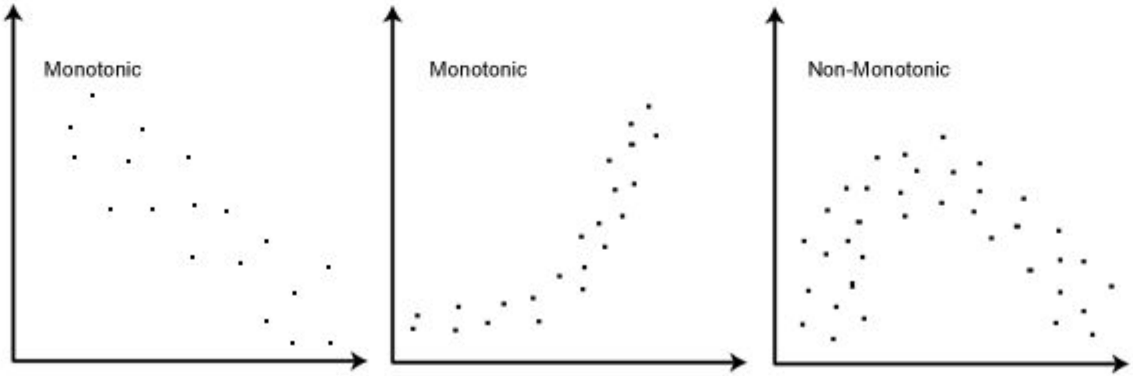
- Spearman

- Kendall

Module 1: ... *again* Data Correlation

- Pearson
- **Spearman**
  - **Ordinals**
  - **Non-linear**
  - **With Outliers**
  - **No Normal Distribution**
  - **Monotone Relation**
- Kendall





Module 1: … *again* Data Correlation

- Point-Biserial
- Cramer's V
- Partial Correlation
- Phi Correlation
- ……

Mean value of the persons who failed

Mean value of the persons who passed

$$r_{pb} = \frac{\bar{x}_2 - \bar{x}_1}{s_x} \cdot \sqrt{\frac{n_1 \cdot n_2}{n^2}}$$

Number of those who have falied

Number of people who have passed

Total number

$$\boxed{V = \sqrt{\frac{\Phi^2}{\min\{(r-1),(c-1)\}}}}$$

$$\rho_{XY|Z} = \frac{N \Sigma_{i=1}^{N} e_{X,i} e_{Y,i}}{\sqrt{N \Sigma_{i=1}^{N} e_{X,i}^2} \sqrt{N \Sigma_{i=1}^{N} e_{Y,i}^2}}$$

$\rho_{XY|Z}$: partial correlation of X and Y (controlled for Z)

N: number of observations

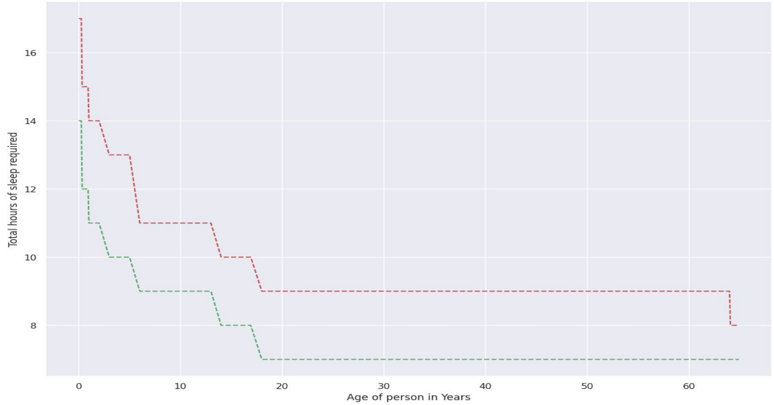
$e_{X,i}$: i-th residual of the linear model predicting X by Z

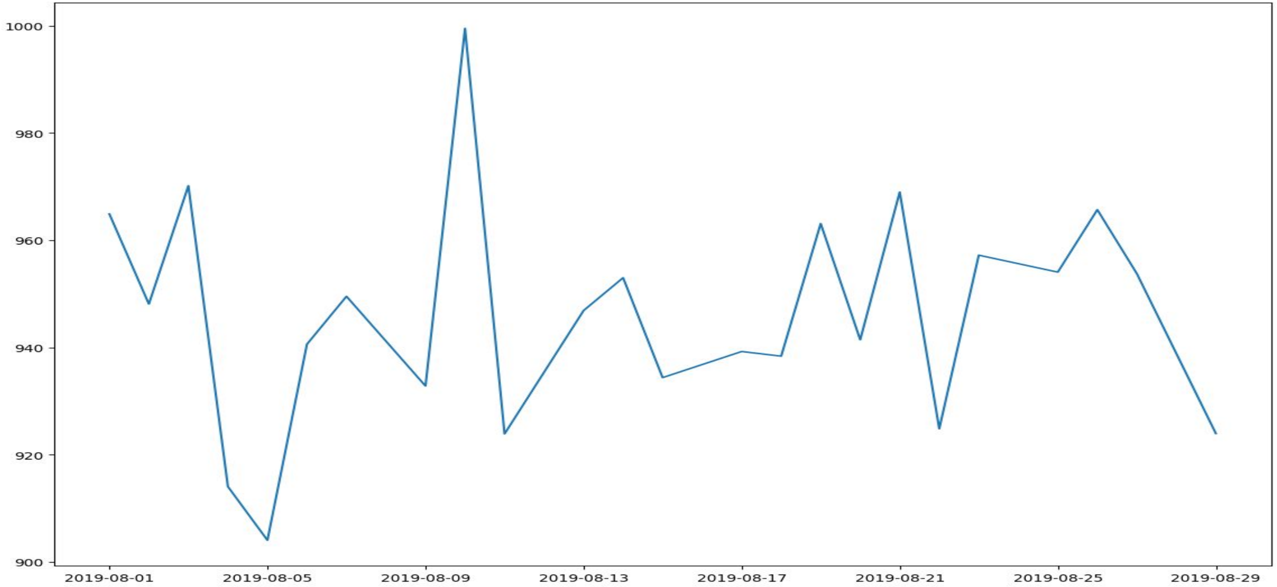
$e_{Y,i}$: i-th residual of the linear model predicting Y by Z

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Module 1: Descriptive Statistics

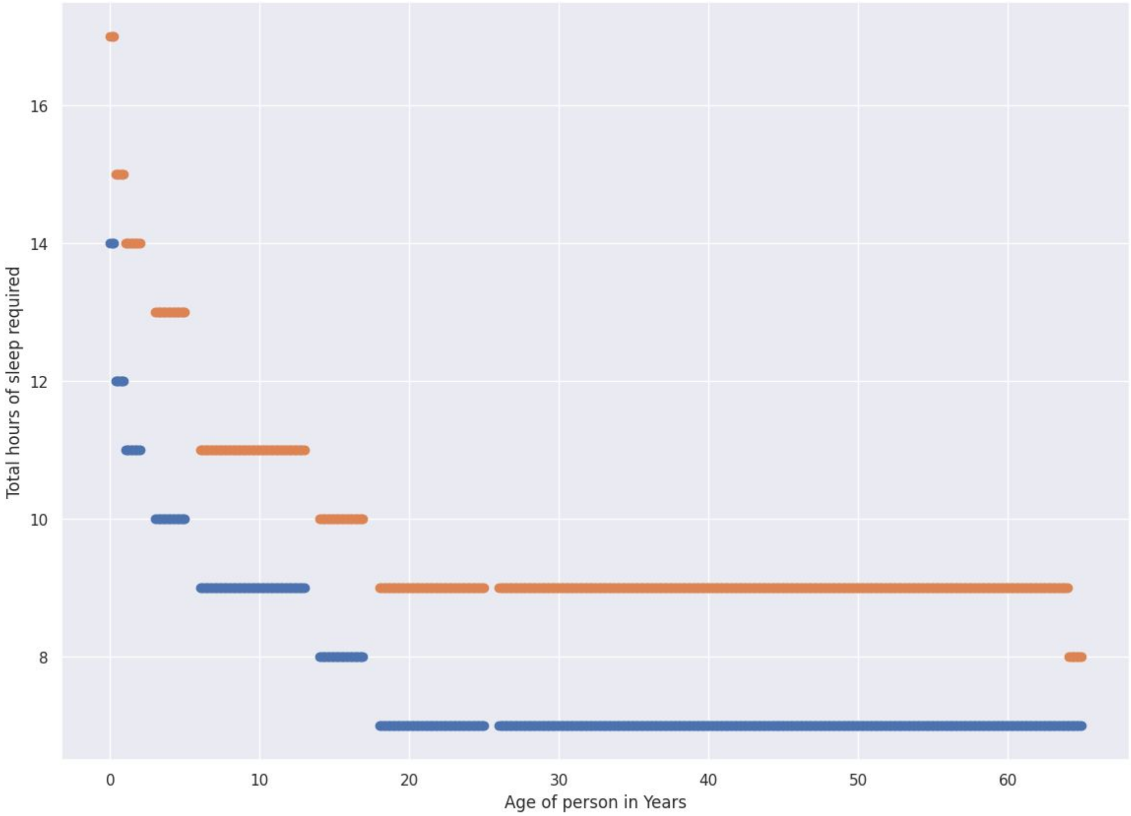
- **Line Chart & Plot**
- Scatter Plots
- Area & Stacked Plots
- Pie and Table Charts
- Polar & Lollipop Charts





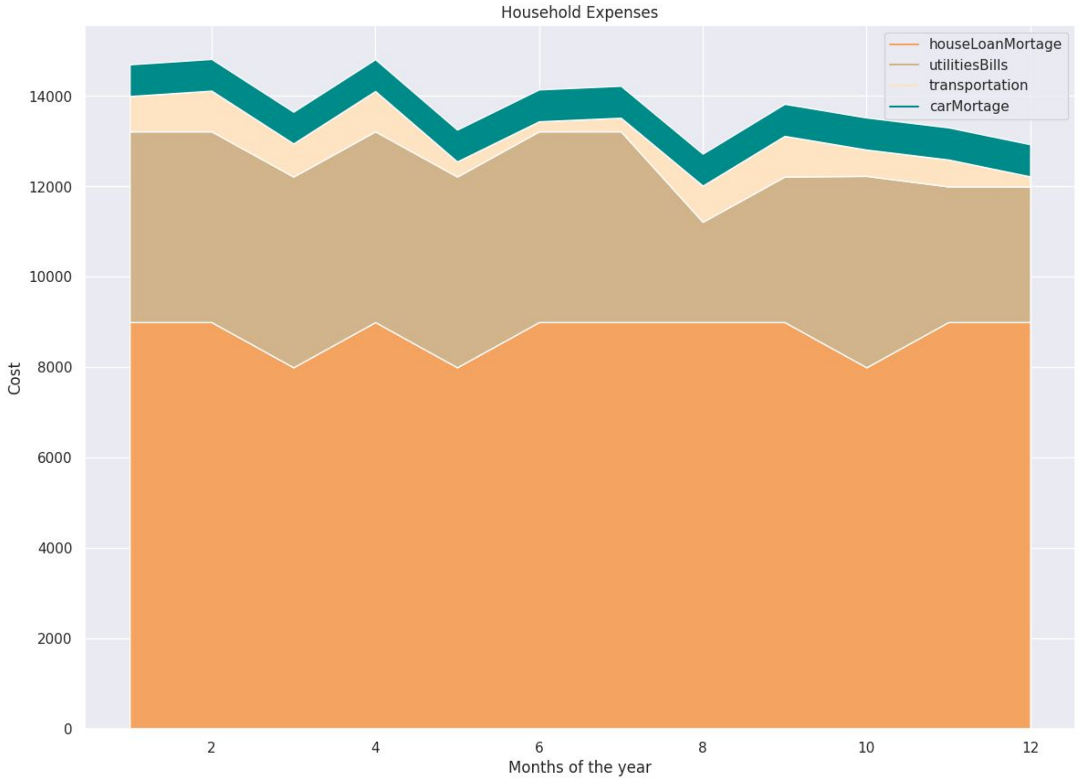
Module 1: Descriptive Statistics

- Line Chart
- **Scatter Plots**
- Area & Stacked Plots
- Pie and Table Charts
- Polar & Lollipop Charts



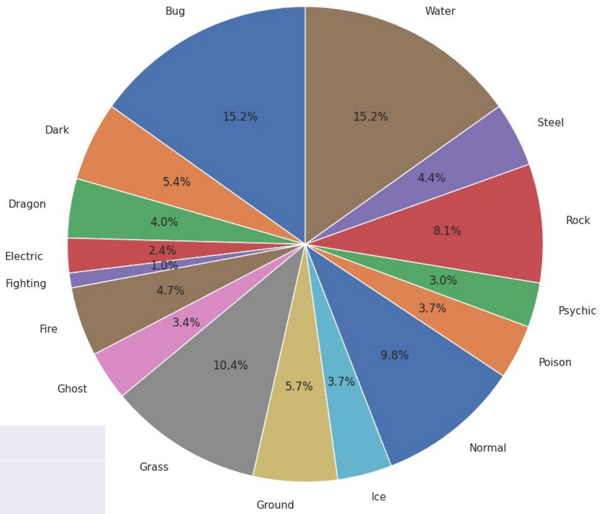
Module 1: Descriptive Statistics

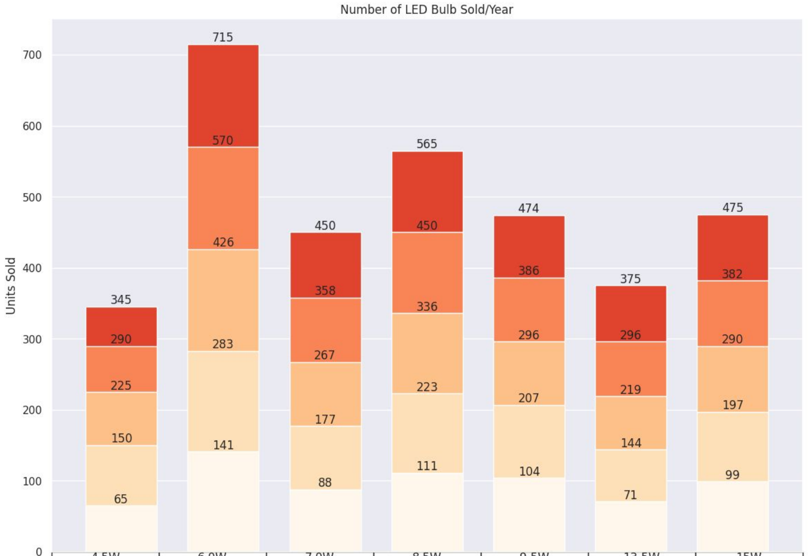
- Line Chart
- Scatter Plots
- **Area & Stacked Plots**
- Pie and Table Charts
- Polar & Lollipop Charts



Module 1: Descriptive Statistics

- Line Chart
- Scatter Plots
- Area & Stacked Plots
- **Pie and Table Charts**
- Polar & Lollipop Charts





	4.5W	6.0W	7.0W	8.5W	9.5W	13.5W	15W
2010	65.0	141.0	88.0	111.0	104.0	71.0	99.0
2011	150.0	283.0	177.0	223.0	207.0	144.0	197.0
2012	225.0	426.0	267.0	336.0	296.0	219.0	290.0
2013	290.0	570.0	358.0	450.0	386.0	296.0	382.0
2014	345.0	715.0	450.0	565.0	474.0	375.0	475.0

Module 1: Descriptive Statistics

- Line Chart
- Scatter Plots
- Area & Stacked Plots
- Pie and Table Charts
- **Polar & Lollipop Charts**