LECTURER: TAI LE QUY

DATA SCIENCE

TOPIC OUTLINE

UNIT 5

SELECTED MATHEMATICAL TECHNIQUES

On completion of this unit, you will have learned …

- how to apply principal component analysis to data.
- how to perform cluster analysis on a dataset.
- how to describe the linear regression model and compute its coefficients.
- how to describe the important features of time-series data.
- the popular models for forecasting future values in time-series data.
- the common approaches for dataset transformation.

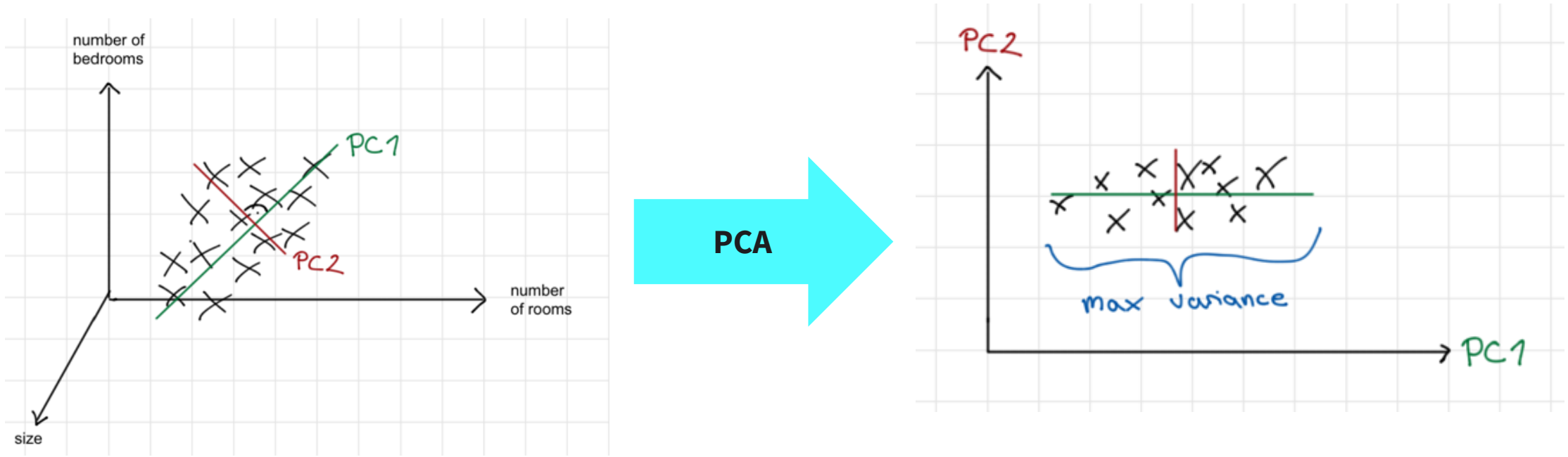
EXPLAIN SIMPLY

1. Explain when to use the Principal Component Analysis (PCA) in practice.
2. Describe the concept of linear regression models and its coefficients using your own words.
3. Identify when the use of clustering techniques is helpful for business.

PRINCIPAL COMPONENT ANALYSIS

Transform potentially correlated variables into fewer uncorrelated variables (principal components - PCs).

→ dimensionality reduction of the dataset while loosing only a small amount of information.



Source of the graphics: Drawings by Antonia Schulze, 2021.

Step (1): Get and subtract the mean

For an input dataset with N records (1, 2, …, N) and M variables (x_1, x_2, …, x_M)

- Calculate the mean

$$\overline{x}_i = \frac{1}{N} \sum_{k=1}^{N} x_{i_k}$$

$$i = 1, 2, ..., M$$

- Calculate new values of variables

$$x_i = x_i - \overline{x}_i$$

$$i = 1, 2, ..., M$$

Step (2): Calculate the covariance matrix

The covariance $C(x_i, x_j)$ is a measure of the changes in variable xi with respect to changes in variable x_j

$$C(x_i,x_j) = \frac{1}{N-1}\sum_{k=1}^{N}(x_i \cdot x_j)_k$$

$$C = \begin{bmatrix} C(x_1,x_1) & C(x_1,x_2) & C(x_1,x_3) & \cdots & C(x_1,x_M) \\ C(x_2,x_1) & C(x_2,x_2) & C(x_2,x_3) & \cdots & C(x_2,x_M) \\ C(x_3,x_1) & C(x_3,x_2) & C(x_3,x_3) & \cdots & C(x_3,x_M) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(x_M,x_1) & C(x_M,x_2) & C(x_M,x_3) & \cdots & C(x_M,x_M) \end{bmatrix}$$

The exact value of $C(x_i,x_j)$ is an indication of how strongly the two variables depend on each other. If the covariance is zero, the variables are uncorrelated.

Step (3): Calculate the eigenvalues and eigenvectors

The objective of PCA is to transform the calculated covariance matrix into an optimum form where all the variables are uncorrelated linearly to first order ($C(x_i, x_j) = 0, i \neq j$)

$$C = \begin{bmatrix} C(x_1,x_1) & C(x_1,x_2) & C(x_1,x_3) & \cdots & C(x_1,x_M) \\ C(x_2,x_1) & C(x_2,x_2) & C(x_2,x_3) & \cdots & C(x_2,x_M) \\ C(x_3,x_1) & C(x_3,x_2) & C(x_3,x_3) & \cdots & C(x_3,x_M) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C(x_M,x_1) & C(x_M,x_2) & C(x_M,x_3) & \cdots & C(x_M,x_M) \end{bmatrix} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_M \end{bmatrix} \qquad C - [\lambda_1 ... \lambda_M] \cdot I = 0$$

eigenvalues (λ), I: identity matrix (i.e., a matrix with “1” in its diagonal and “0” otherwise)

- Finding eigenvalues: to solve determinant $\det(C - \lambda \cdot I) = |C - \lambda \cdot I| = 0$
- Eigenvectors (principal components (PCs)): $C \cdot PC_i = (\lambda_i \cdot I) \cdot PC_i$

$$i = 1, 2, ..., M$$

- i^{th} PC is the solution of

$$[(C - \lambda_i \cdot I) \cdot PC_i = 0]$$

Step (4): Formulate the PCs

The PC (i.e., eigenvector) that corresponds to the highest eigenvalue is the first principal component of the dataset

- Percentage of how much variance (V) each PC represents

$$H_{PC_i} = \frac{\lambda_i}{\lambda_1 + \cdots + \lambda_M} \cdot 100\,\%$$

Step (5): Dimensionality reduction

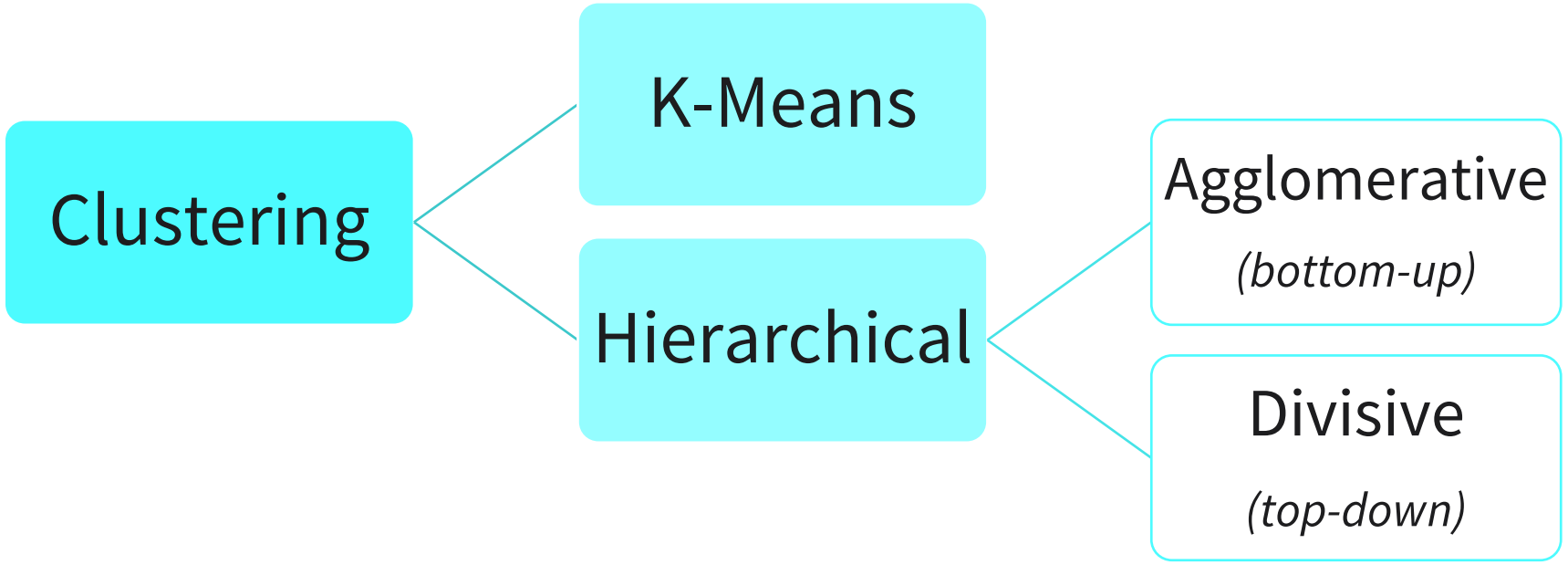
Ignore the PCs with less significance

Step (6): Reconstruct the dataset

$$[y]^T = [PC_1 \ldots PC_{M*}]^T \cdot [x]^T$$

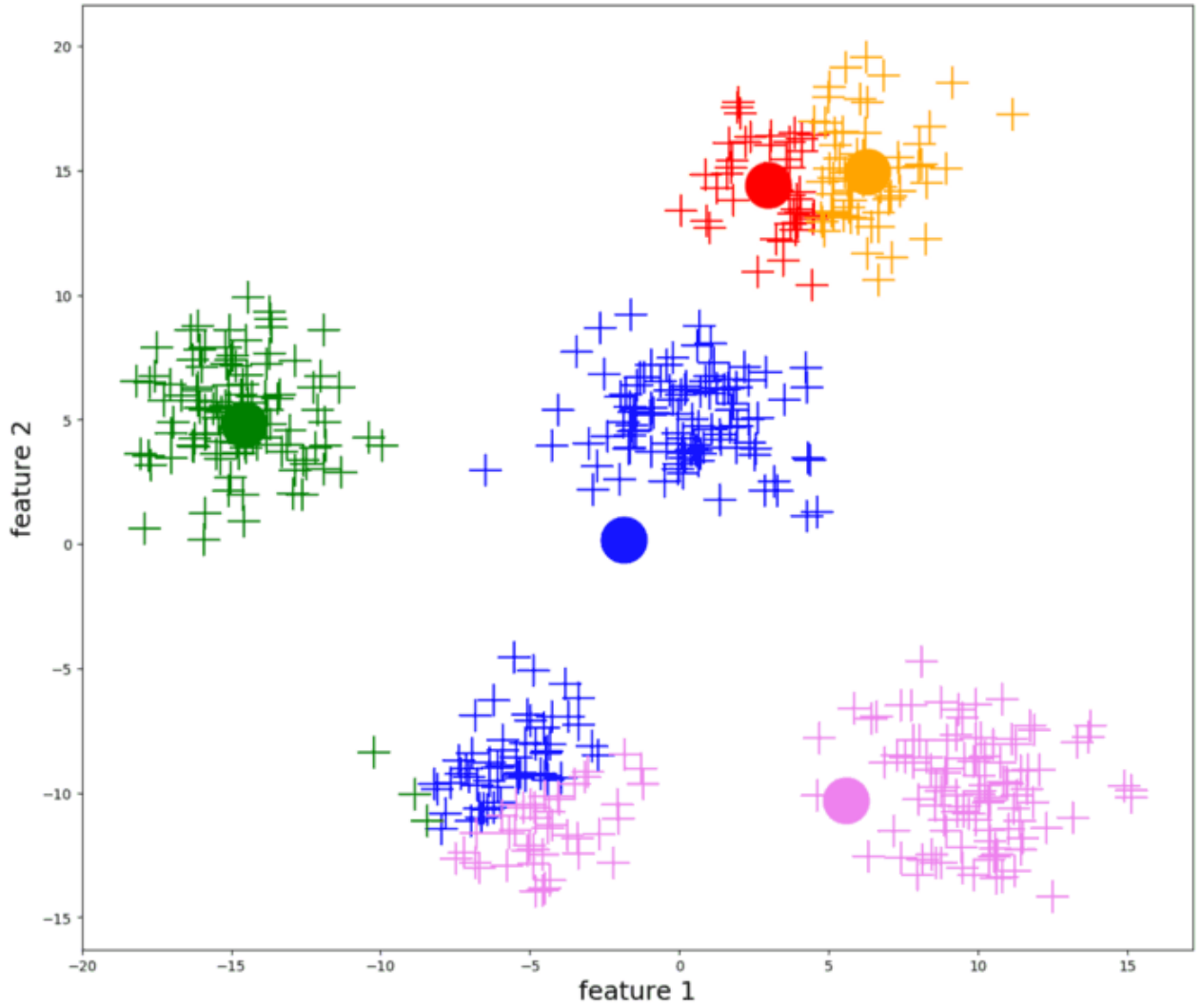
Grouping objects into unlabeled, meaningful clusters

- maximize similarity within a cluster (distance to centroids)
- maximize dissimilarity between clusters



K-MEANS CLUSTERING

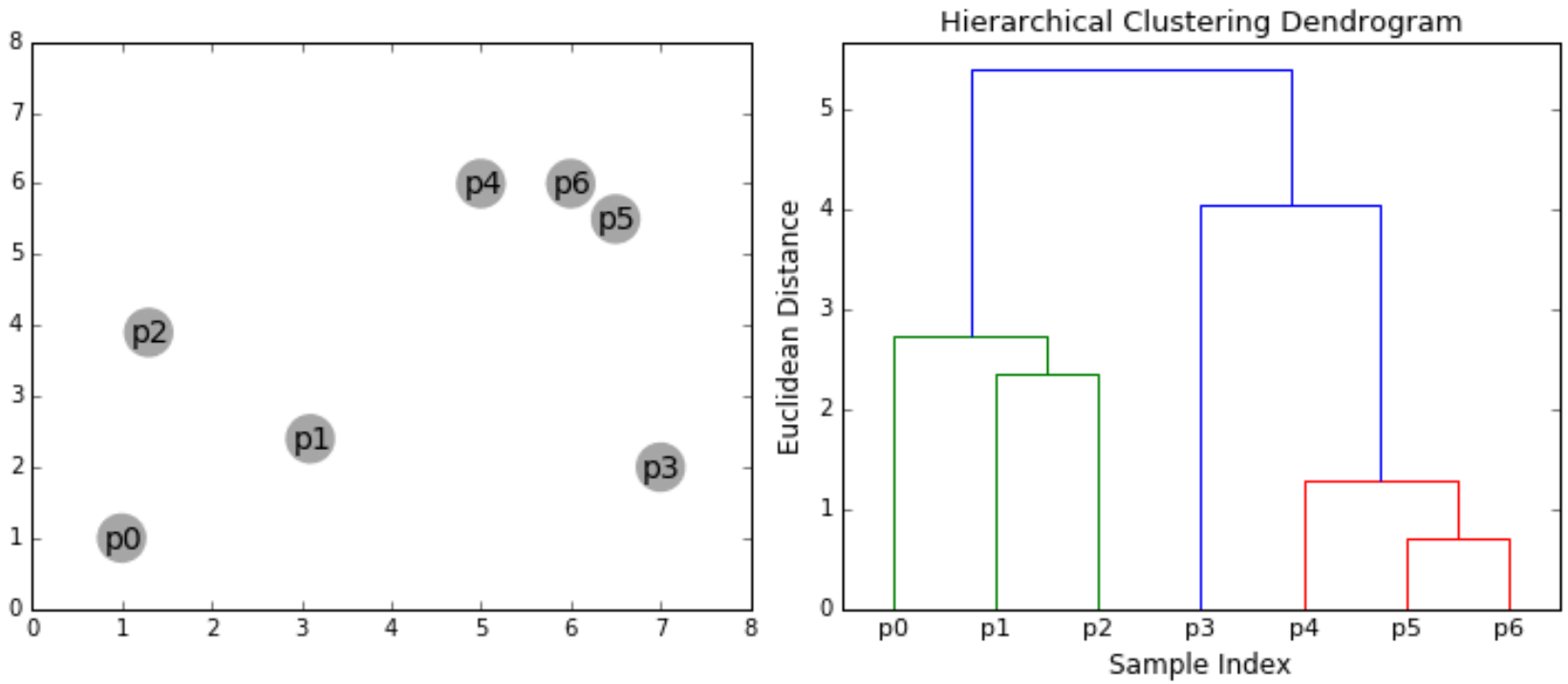
- select # of clusters (k)
- choose random centroids
- assign data points to clusters based on minimal distance to centroid
- calculate new centroid
- start over until no changes made to centroids



Source of the animation: Brilenkov, 2021.

HIERARCHICAL CLUSTERING

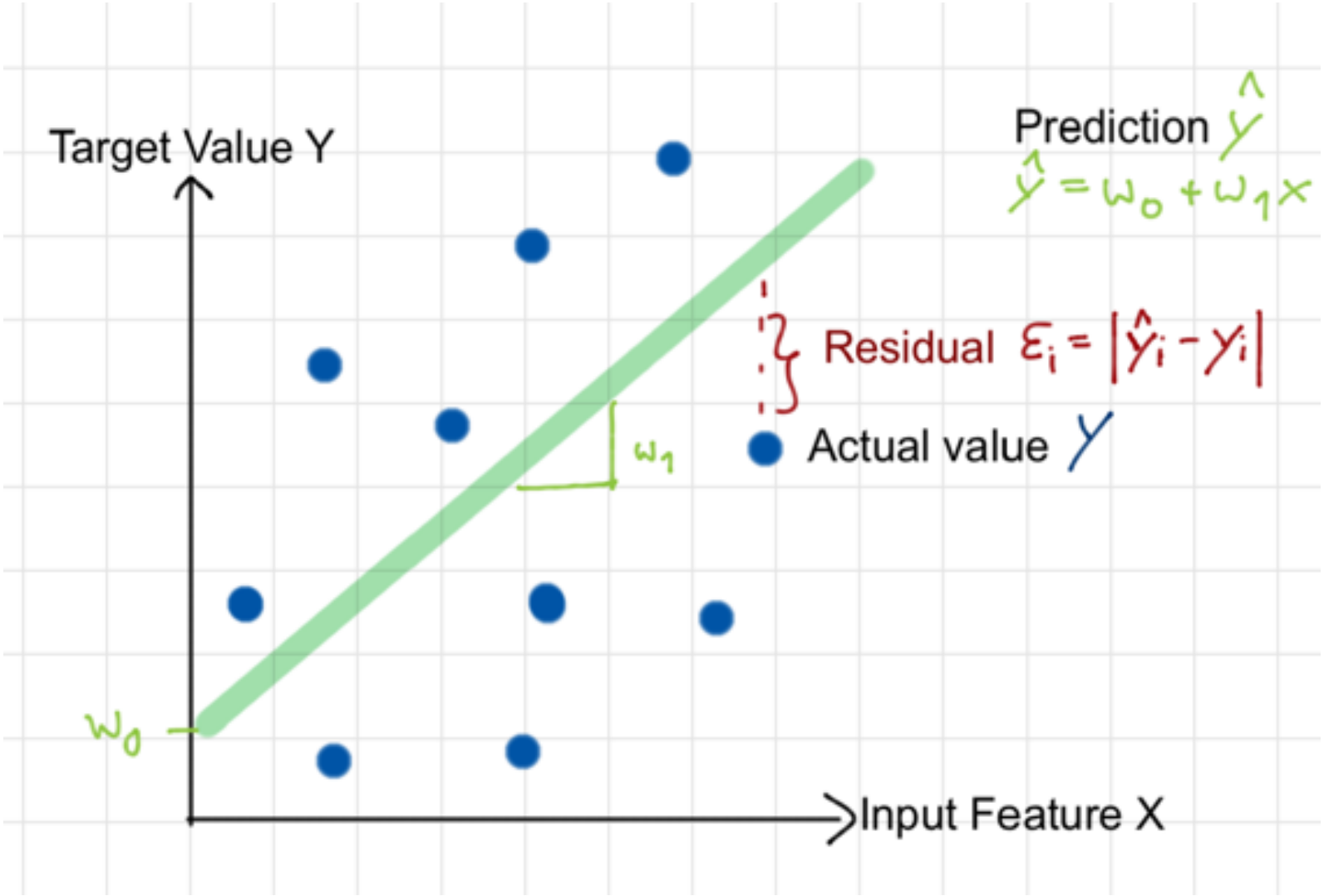
- assign each record to a unique cluster
- merge clusters with minimum distance
- repeat until only one cluster left



Source of the animation: Sheenan, 2017.

LINEAR REGRESSION

- predict value of dependent (*target*) variable given independent (*predictor*) variables
- assumption: linear relationship between variables



Source of the graphic: Drawing by Antonia Schulze, 2021.

Simple Linear Regression Model

$$y = w_0 + w_1 x$$

Calculate w_1:

$$w_1 = \frac{n\sum_{i=1}^{n} \hat{y}x - \sum_{i=1}^{n} \hat{y} \sum_{i=1}^{n} x}{-\left(\sum_{i=1}^{n} x\right)^2 + n\sum_{i=1}^{n} x^2}$$

Calculate w_0:

$$w_0 = \frac{1}{n}\sum_{i=1}^{n} \hat{y} - \frac{w_1}{n}\sum_{i=1}^{n} x$$

Predict:

$$\widehat{y} = w_0 + w_1 \widehat{x}$$

MULTIPLE LINEAR REGRESSION MODEL

A dataset with m variables [x_1, x_2, …, x_m] and n data records (1, 2, …, n)

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m$$

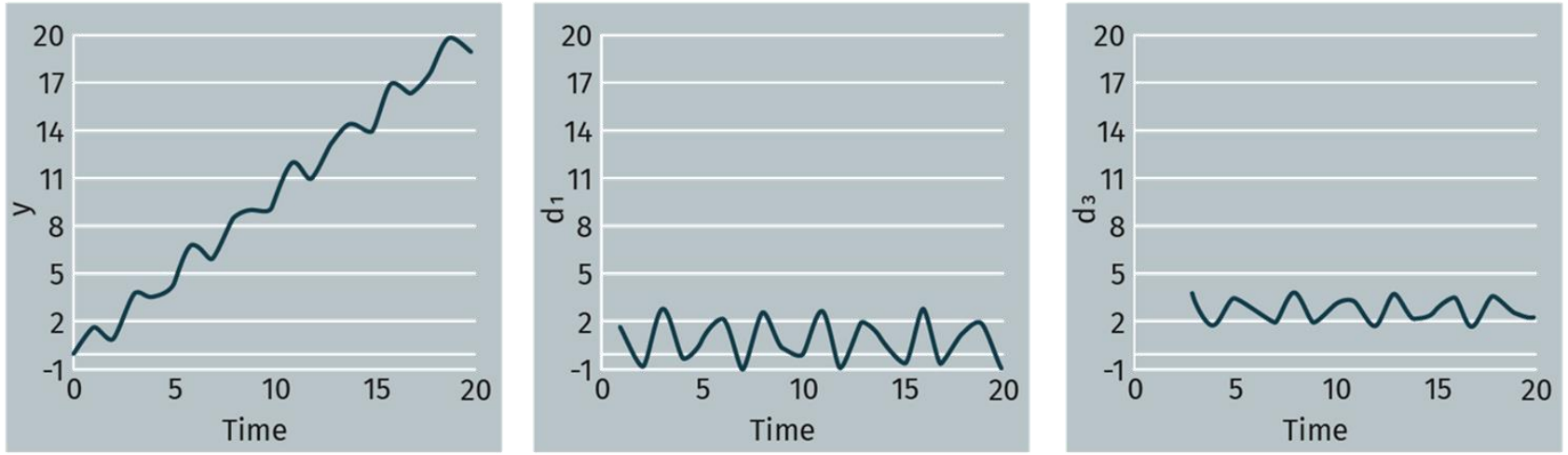
$\hat{y}$: predicted value, w_0 is the bias and (w_1, w_2, …, w_m) are the weights

TIME-SERIES FORECASTING

Source of image: Google finance

- Stationary time series = constant mean and standard deviation over time

t	y	d_1 $(y_t - y_{t-1})$	d_2 $(y_t - y_{t-2})$	d_3 $(y_t - y_{t-3})$
0	0.000			
1	1.650	1.650		
2	1.012	-0.638	1.012	
3	3.851	2.839	2.201	3.851
4	3.695	-0.156	2.683	2.045
5	4.612	0.917	0.761	3.600
6	6.894	2.282	3.199	3.043
7	6.029	-0.865	1.417	2.334
8	8.581	2.551	1.687	3.968
9	9.088	0.508	3.059	2.194
10	9.285	0.197	0.705	3.256



Source of the graphic: IU International University, Course Book DLMBDSA01.

- Stationary time series = constant mean and standard deviation over time
- Lag(n) = backshift of a time-series by n time steps
- Autocorrelation (ACF) = correlation between variable and previous lags
- Partial Autocorrelation (PACF) = autocorrelation between y_t and y_{t-k} that is not accounted for by the autocorrelations from the 1^{st} to the $(k-1)^{st}$ lags.

- Autocorrelation coefficient ACF(n) at lag(n)

$$\mathrm{ACF(n)} = \frac{\mathrm{C(y_t, y_t - n)}}{\sqrt{\mathrm{V(y_t)} \cdot \mathrm{V(y_t - n)}}}$$

where $C(y_t, y_{t-n})$ is the covariance coefficient between y_t and y_{t-n} and $V(y_t)$ ist the variance of y_t, and $V(y_t) = C(y_t, y_t)$.

- Partial autocorrelation function PACF(k) at lag(k)

$$\begin{pmatrix} \mathrm{ACF(0)} & \mathrm{ACF(1)} & \cdots & \mathrm{ACF(k-1)} \\ \mathrm{ACF(1)} & \mathrm{ACF(0)} & \cdots & \mathrm{ACF(k-2)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{ACF(k-1)} & \mathrm{ACF(k-2)} & \cdots & \mathrm{ACF(0)} \end{pmatrix} \begin{pmatrix} \mathrm{PACF_1} \\ \mathrm{PACF_2} \\ \vdots \\ \mathrm{PACF_k} \end{pmatrix} = \begin{pmatrix} \mathrm{ACF(1)} \\ \mathrm{ACF(2)} \\ \vdots \\ \mathrm{ACF(k)} \end{pmatrix}$$

TIME-SERIES FORECASTING

Autoregressive Model (AR)

models future values as a function of recent past **sequential values**

Moving Average Model (MA)

models future values as a function of recent past **sequential error terms**

Autoregressive Integrated Moving Average Model (ARIMA)

combination of **AR** & **MA** models with an **I**ntegration of differencing the time-series until stationarity reached

AUTOREGRESSIVE (AR) MODEL

AR(n) stands for an autoregressive model of order n, which indicates that n previous observations (i.e., **lag(n)**) are used in the prediction of the next observation

$$y_t = p_0 + p_1 y_{t-1} + p_2 y_{t-2} + \cdots + p_n y_{t-n} + \varepsilon_t$$

where $\{p_0, p_1, \ldots, p_n\}$ are the model coefficients and ε_t is a white noise term

The moving average model predicts future observations:

$$y_t = q_0 + q_1 \varepsilon_{t-1} + q_2 \varepsilon_{t-2} + \cdots + q_n \varepsilon_{t-n}$$

where ε_{t-n} are the noise error terms; $\{q_0, q_1, \ldots, q_n\}$ are the model coefficients.

AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) MODEL

- ARIMA models mix both the (AR) and (MA) models with integrated parameters in one model to obtain a better understanding of time-series data and/or to forecast future data points in the series
- ARIMA(p, d, q) with p: the number of (AR) terms, d is the degree of differencing, and q is the number of (MA) terms
- General form of the model:

$$\begin{aligned} y_t &= \text{costant} + \text{weighted sum of the previouspvalues ofy} \\ &\quad + \text{weighted sum of the previousqforecast errors} \\ y_t &= c + p_1 y_{t-1} + \cdots + p_n y_{t-n} + q_1 \varepsilon_{t-1} + \cdots + q_m \varepsilon_{t-n} \end{aligned}$$

- Set p and q based on the Autocorrelation (ACF) and Partial Autocorrelation (PACF) values

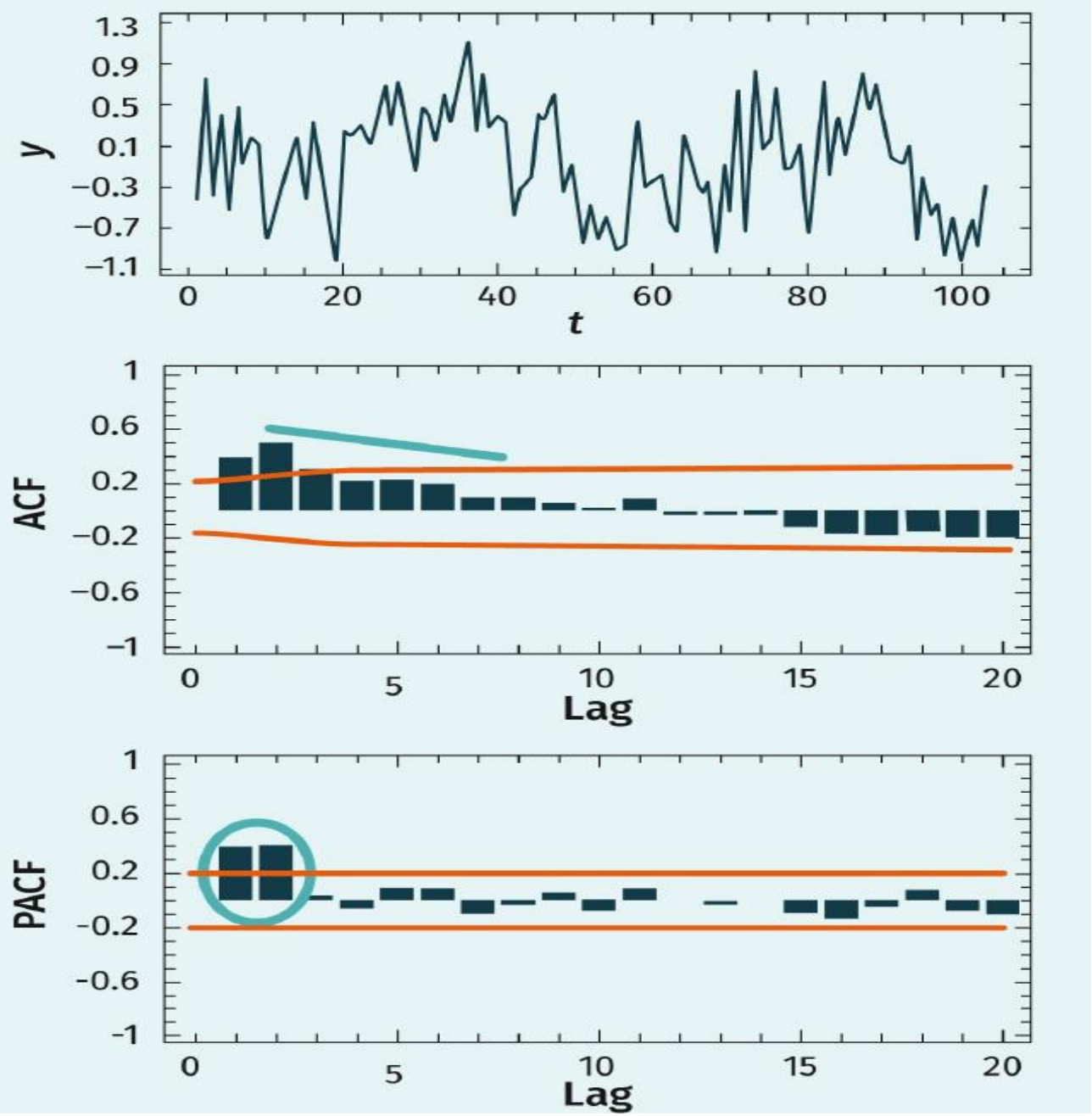
1. If the ACF plot cuts off sharply at lag(k) (i.e., if the autocorrelation is significantly different from zero at lag(k) and extremely low in significance at the next higher lag and the ones that follow), while there is a more gradual decay in the PACF plot (i.e., if the decay is significance beyond lag(k) is more gradual), then set $q = k$ and $p = 0$.

2. If the PACF plot cuts off sharply at lag(k) while there is a more gradual decay in the ACF plot beyond lag(k), then set $p = k$ and $q = 0$.

3. If there is a single spike at lag(1) in both the ACF and PACF plots, then set $p = 1$ and $q = 0$ if the spike is positive, and set $p = 0$ and $q = 1$ if it is negative.

EXAMPLE OF ARIMA(2,0,0)

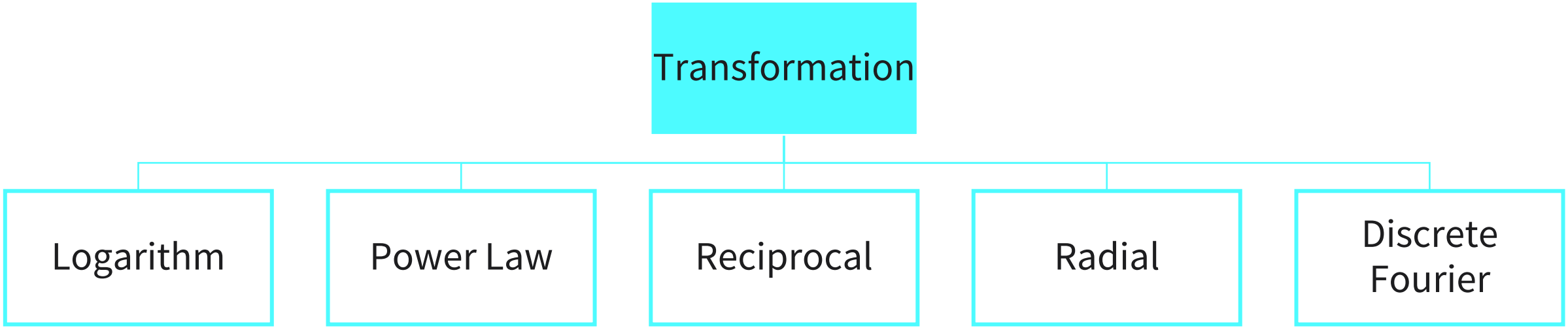
- The time-series looks stationary over time; therefore, there is no need for a differencing process.
- The PACF values cut off sharply at lag(2), and the ACF values gradually decay after lag(2).
- Set p = 2 and q = 0, resulting in the ARIMA(2, 0, 0) model



Source of image: Course book

TRANSFORMATION APPROACHES

Process of transforming variables to improve its interpretability



Log(x)
When linear regression model fails to hold the linearity between the input variables

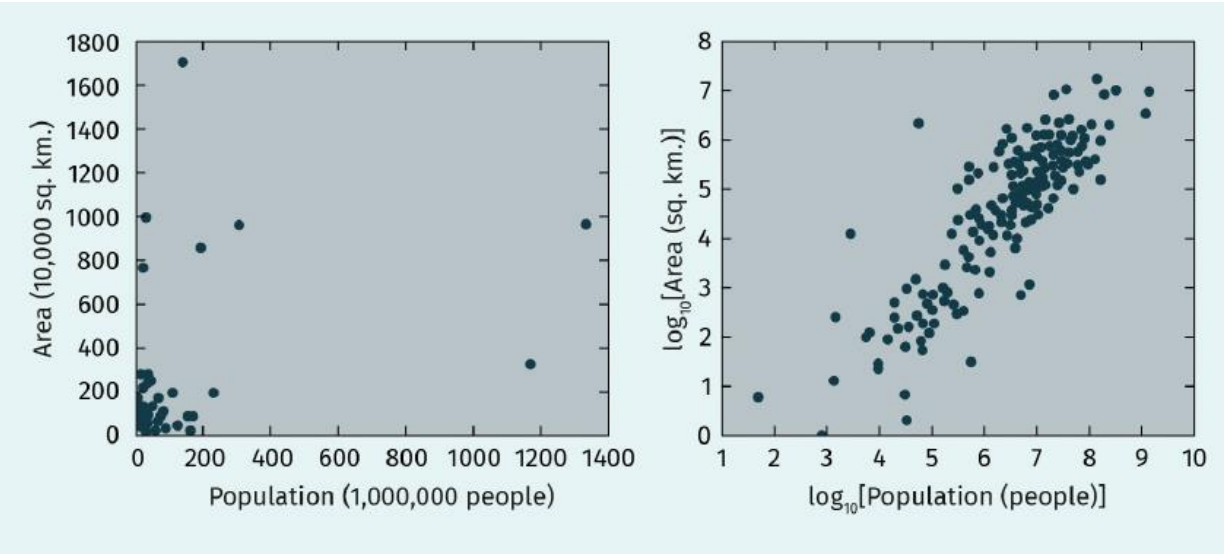
$X = x^{\gamma}$
Updated during the training phase of the model-building process until the highest performance accuracy

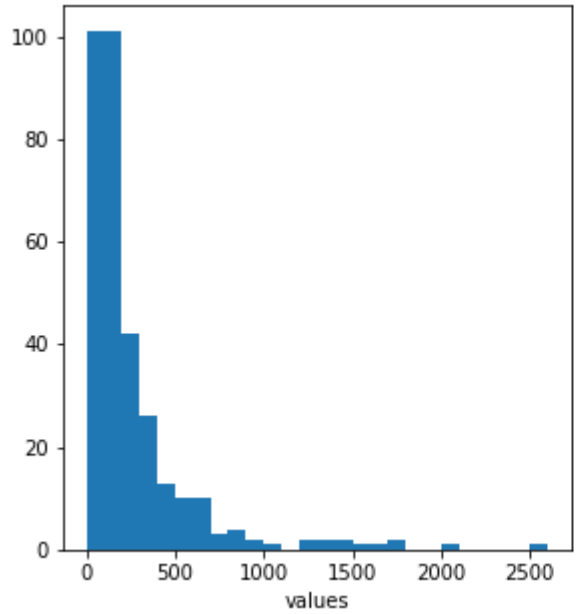
X = 1/x
Changes the shape of a variable to its inverse

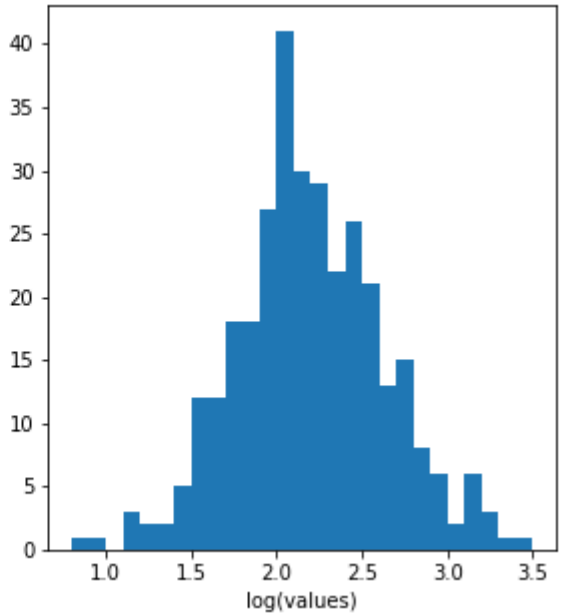
Combines two variables (x1) and (x2) and transforms them into the radial coordinates of radius $(r = \sqrt{x_1^2 + x_2^2})$ and angle $(\theta = \tan^{-1}\frac{x_1}{x_2})$

Transfer a variable of a dataset from its traditional domain (e.g., variable (x) versus time (t)) to its frequency domain

TRANSFORMATION APPROACHES - EXAMPLE







Source of image: course book (left) and https://julienbeaulieu.gitbook.io/wiki/sciences/programming/data-analysis/data-visualization/univariate-exploration-of-data/scales-and-transformations-log (right)

You have learned …

- how to apply principal component analysis to data.
- how to perform cluster analysis on a dataset.
- how to describe the linear regression model and compute its coefficients.
- how to describe the important features of time-series data.
- the popular models for forecasting future values in time-series data.
- the common approaches for dataset transformation.

SESSION 5

TRANSFER TASK

TRANSFER TASK 1

You are facing a big dataset and want to apply your previous knowledge of PCA to get a smaller, but still informative dataset. Your colleague, however, has some questions that you will need to answer. Prepare a role play.

Inspiration:

- Discuss: More data → more information?
- Analyze: advantages and disadvantages of using PCA

Working in groups

- Select your domain (e.g., healthcare, education, finance, etc.)
- Select a task (e.g., customer segmentation, customer demand prediction, diabetes prediction, etc.)
- Describe how PCA and/or clustering and/or time-series prediction techniques can be applied?
- Present your finding in 5 minutes

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. The transformation approach, which transfers data variables to their frequency domain, is called the…

   a) radial transformation.
   b) reciprocal transformation.
   c) Fourier transformation.
   d) logarithm transformation.

2. The auto-regressive model assumes a…

   a) linear function between the future output and past outputs.
   b) repeated pattern in the time-series data.
   c) constant output over time.
   d) sinusoidal wave that relates the outputs and the inputs.

3. The operation of sorting data variables according to their level of changeability along data records is part of…

a) regression modelling.
b) classification modelling.
c) clustering analysis.
d) principal component analysis.

LIST OF SOURCES

Brilenkov, R. (2021). *Understanding K-Means Clustering: Hands-on Visual Approach* [blog post]. Retrieved from: https://ai.plainenglish.io/understanding-k-means-clustering-hands-on-visual-approach-c2dc46f0ed18
Sheenan, D. (2017). *Clustering with Scikit with GIFs.* [blog post]. Retrieved from: https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/