Analyse und Visualisierung von Zeitreihen-Daten in Python

Hmetrica

ENABLE AI
MASTER YOUR DATA

Analyse und Visualisierung von Zeitreihen-Daten in Python

Tag 1

ENABLE AI
MASTER YOUR DATA
Hmetrica

Erwartungen

- Kurze Vorstellung
- Hintergrund, Vorerfahrungen Data Science, Programming
- Wie arbeite ich mit Zeitreihen?
- Was würde ich gern mitnehmen aus dem Kurs?
- Welcher Themenbereich interessiert mich am meisten?

Hmetrica GmbH

www.hmetrica.com |
info@hmetrica.com

Data Science Beratung

- Statistik, Machine Learning und Künstliche Intelligenz.
- Schwerpunkte: Datenvisualisierung, Machine Learning, Zeitreihenanalyse

ENABLE AI
MASTER YOUR DATA
Hmetrica

Inhalte – Was haben wir vor?

Tag 1	Tag 2
1. Einführung in Zeitreihendaten in Python 2. Zeitreihen und ihre Merkmale visualisieren 3. Zeitreihen vorhersagen (Statistik I): Exponentielle Glättung und Holt-Winters 4. Zeitreihen vorhersagen (Statistik II): ARIMA-Modelle	5. Einblick in andere Zeitreihenmodelle 6. Machine Learning für Zeitreihen: Überblick, Vorbereitung und Klassifikation 7. Machine Learning für Zeitreihen: Clustering 8. Deep Learning für Zeitreihen (Einblick)

Fragen?

Einführung in Zeitreihendaten in Python

Session 1 (Montag 09:15 – 10:45)

ENABLE AI
MASTER YOUR DATA

Hmetrica

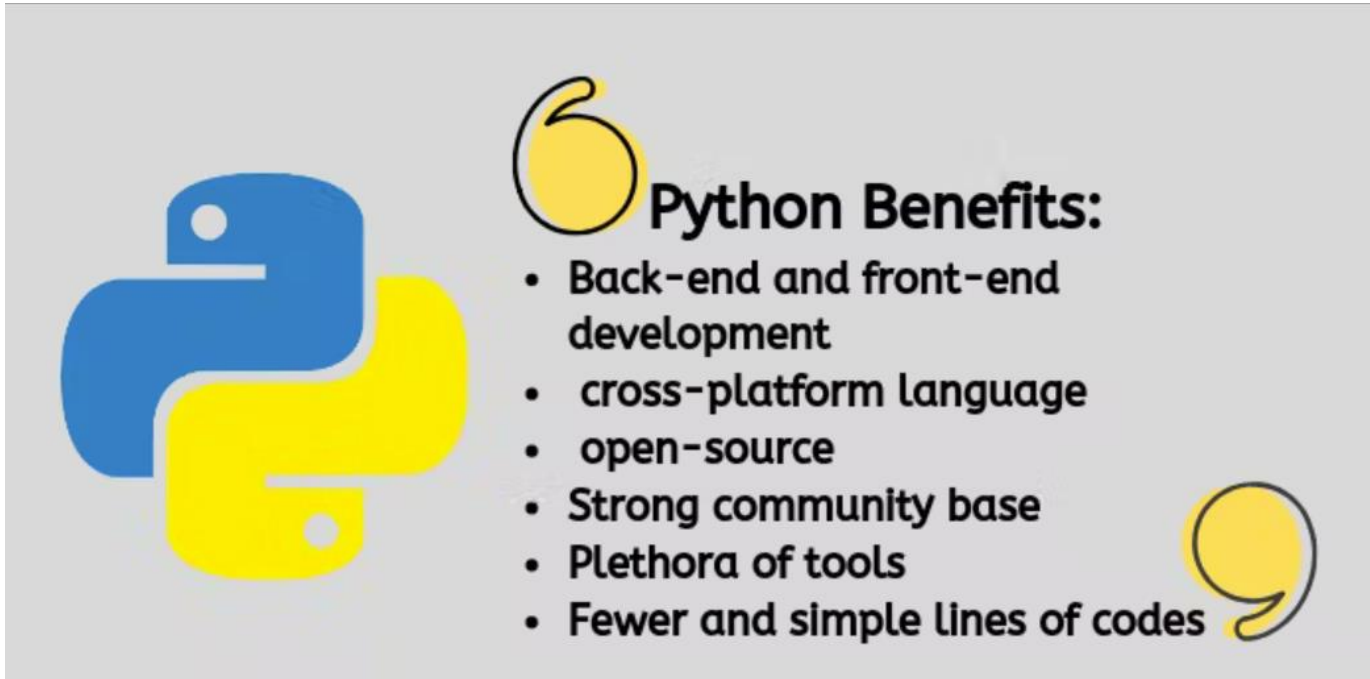
Was haben wir vor?

1. Einführung in Zeitreihendaten in Python
   1.1 Einführung in Python: pandas, matplotlib
   1.2 Einführung in Zeitreihendaten: Definitionen, einfache Merkmale

Einführung in Python

Python ist eine allgegenwärtige Skriptsprache, benutzt für:

- Webentwicklung,
- App-Entwicklung,
- wissenschaftliche und numerische Bereiche,
- Geschäftsanwendungen,
- GUI-Design,
- Automatisierung,
- künstliche Intelligenz,
- maschinelles Lernen
- und vieles mehr…



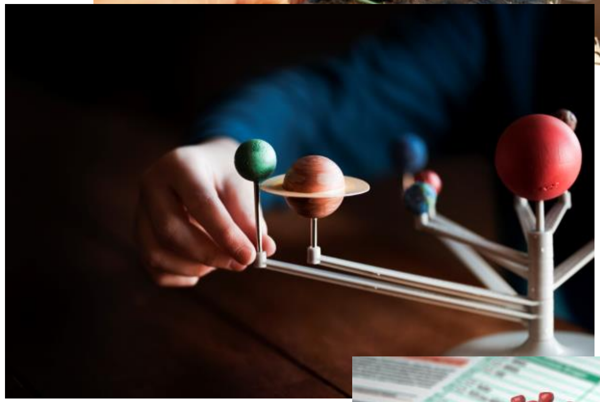
https://rotechnica.com/what-is-python-used-for/, https://hackernoon.com/what-is-python-used-for-an-exclusive-answer-1z2xo3xtf,

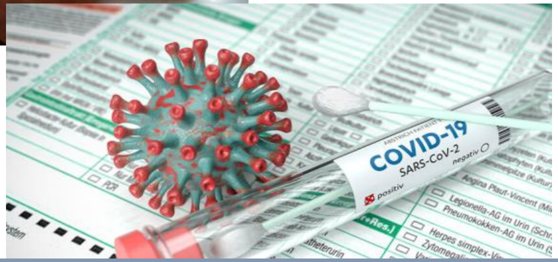
Ab ins Jupyter Notebook

Einführung in Zeitreihendaten

ENABLE AI
MASTER YOUR DATA
Hmetrica

Beispiele für Zeitreihenanwendungen

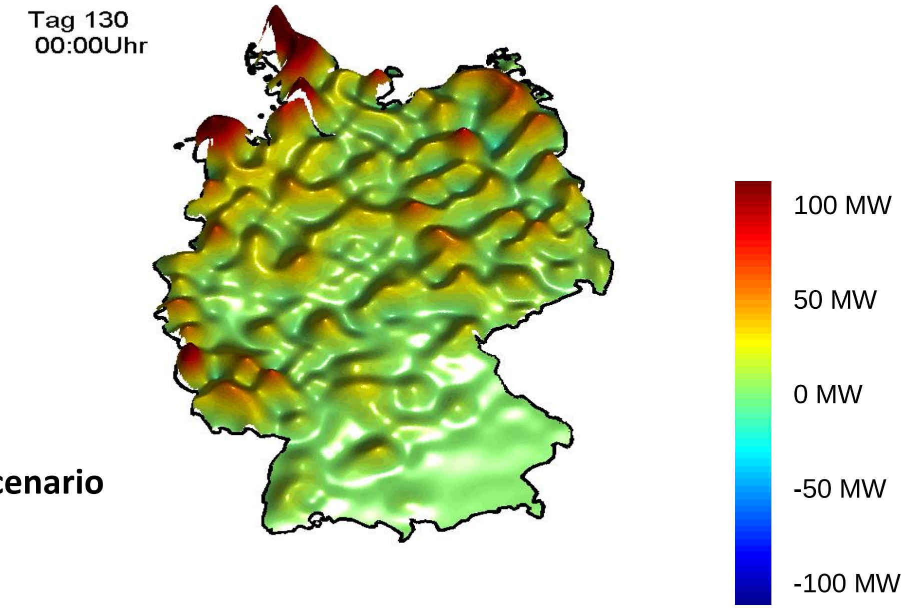




https://www.analyticssteps.com/blogs/introduction-time-series-analysis-time-series-forecasting-machine-learning-methods-models

Bsp: Energiewende

(PV) + (Wind) – (Consumption) = (Residual Load) = f(t)



Most likely scenario for 2025

© IFHT

Image source:earth.com; treehugger.com; Siemens, RWTH, Metzger, Huber, et al., 2021;

Was sind Zeitreihen?

- Umgangssprachlich: Daten, die sich auf aufeinanderfolgende Zeitpunkte (oder Zeiträume) beziehen
- Statistik betrachtet: Zeitreihe als “Realisation” eines “stochastischen Prozesses”

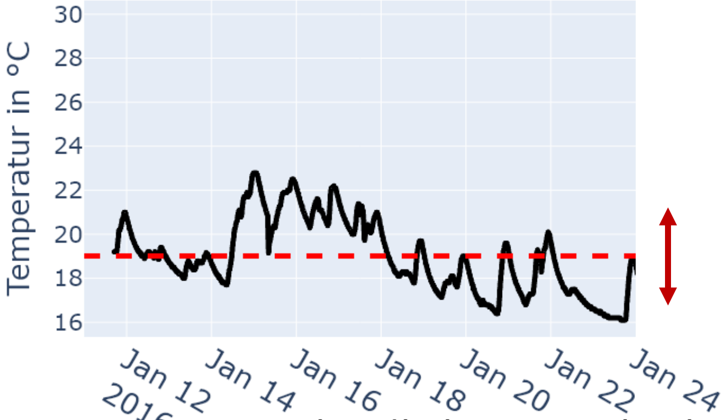
Stochastischer Prozess $\{X_t\}_{t \epsilon T}$ ist eine Folge von Zufallsvariablen (Funktionen $X_t: \Omega \to \mathbb{R}$), die einen Index t aus einer Indexmenge T (meist Zeit gemessen als $\mathbb{N}_0$ oder $\mathbb{R}_+$) haben

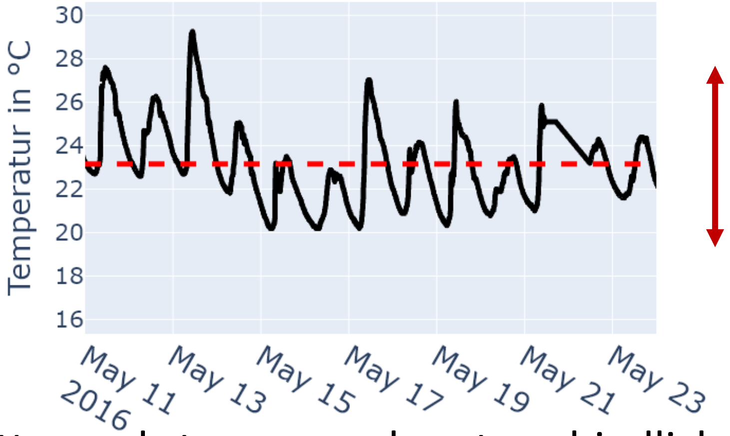
Zeitreihen beschreiben

Beispiel

Sie beschreiben die Temperatur in einem Gebäude im Januar und Mai

Wie unterscheiden sich diese beiden Zeitreihen?





Sie haben einen unterschiedlichen Durchschnitt – und streuen auch unterschiedlich darum herum

Stochastischer Prozess $\boldsymbol{X_1, X_2, X_3, \ldots}$	**Daten** $\boldsymbol{x_1, x_2, x_3, \ldots, x_t}$
Erwartungswert $\mu_t = \mathbb{E}[X_t]$	**Mittelwert** $\bar{x}_t = \hat{\mu}_t = \frac{1}{t}\sum_{i=1}^{t} x_i$
Varianz $\mathrm{Var}(X_t) = \mathbb{E}[(X_t - \mu_t)^2] = \gamma_t(0)$	**Stichprobenvarianz** $\hat{\gamma}_t(0) = \frac{1}{t}\sum_{i=1}^{t}(x_i - \bar{x})^2$
Autokovarianz $\mathrm{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)] = \gamma_t(h)$	**Stichprobenautokovarianz** $\hat{\gamma}_t(h) = \frac{1}{t}\sum_{i=1}^{t-\lvert h\rvert}\left(x_{i+\lvert h\rvert} - \bar{x}\right)(x_i - \bar{x})$

Hängt von der Zeit ab!

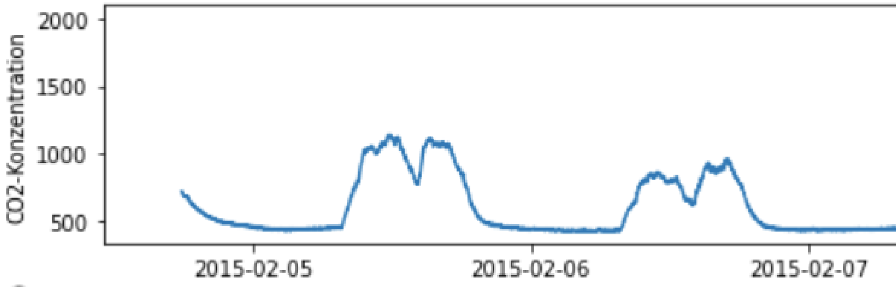
Warum sind Zeitreihen besonders?

Beispiel

(A) Ich messe heute Mittag die CO^2-Konzentration in 72 Büros in München **(keine Zeitreihe)**

(B) Ich messe 72 Stunden lang die CO^2-Konzentration in meinem München **(Zeitreihe)**

- Bei **Zeitreihen**:
- Beobachtung zu Zeitpunkt t: *könnte etwas zu tun haben* mit Beobachtungen zu anderem Zeitpunkt $t+h$
- Deshalb: **Kovarianz** der Zeitreihe mit sich selbst (zeitverzögert) wichtig!



Stochastischer Prozess $X_1, X_2, X_3, \ldots, X_n$	**Daten** $x_1, x_2, x_3, \ldots, x_n$
Kovarianz $\mathrm{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)] = \gamma_t(h)$	Stichprobenkovarianz $\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-\lvert h\rvert}\left(x_{t+\lvert h\rvert} - \bar{x}\right)(x_t - \bar{x})$

Die **Autokorrelationsfunktion ACF** (engl. Auto Correlation Function) ist definiert als

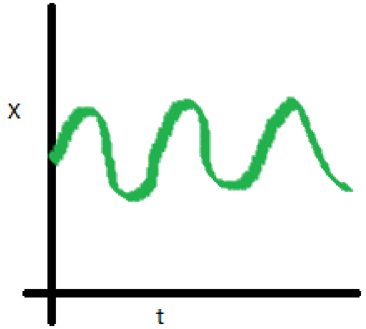
$$\rho_t(h) = \frac{\gamma_t(h)}{\gamma_t(0)}$$

Zeitreihen modellieren

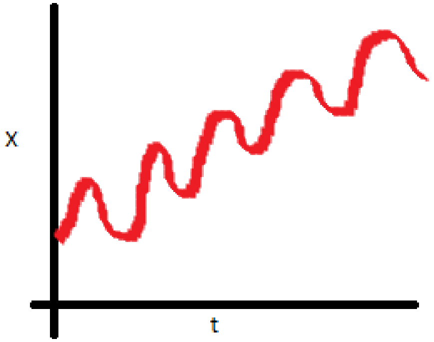
Eine Zeitreihe heißt **stationär**, wenn

$$\mu_t = \mu \qquad \text{und} \qquad \gamma_t(h) = \gamma(h)$$

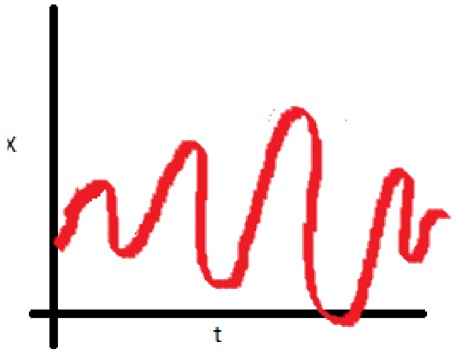
- Mittelwertfunktion und die Autokovarianzfunktion (und somit auch die Varianzfunktion) nicht abhängig von t sind.



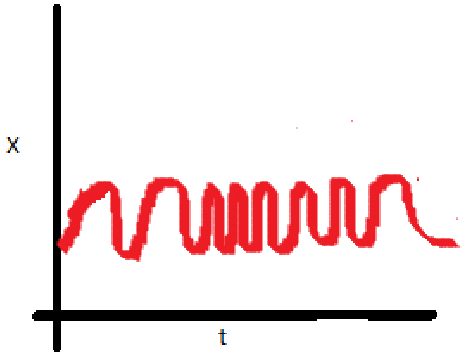
Stationär



Nicht stationär



Nicht stationär



Nicht stationär

Zeitreihen modellieren

Eine Zeitreihe heißt **stationär**, wenn

$$\mu_t = \mu \wedge \gamma_t(h) = \gamma(h)$$

- Mittelwertfunktion und die Autokovarianzfunktion (und somit auch die Varianzfunktion) nicht abhängig von t sind

- Wichtiger **Baustein**, um Zeitreihen zu modellieren:

Eine Zeitreihe $\{X_t\}_{t \epsilon T}$ heißt **weißes Rauschen**, wenn alle X_t unabhängig identisch verteilt und

$$\mu_t = \mu \wedge \gamma_t(h) = \begin{cases} \sigma^2, h = 0 \\ 0, h \neq 0 \end{cases}$$

- Wir schreiben $\epsilon_t \sim IID(0, \sigma^2)$
- Sonderfall: Gaußsches weißes Rauschen $\epsilon_t \sim N(0, \sigma^2)$

Fragen?

Ab ins Jupyter Notebook

ENABLE AI
MASTER YOUR DATA
Hmetrica