

# Data Science: Der nächste Schritt nach Data Analytics

by Woche 1

---

Nachdem wir uns in "Data Analyst mit Python" mit den Grundlagen der Datenverarbeitung, Manipulation und Visualisierung beschäftigt haben, machen wir in diesem Kurs schlichtweg da weiter wo wir aufgehört haben. Wir vertiefen unser Wissen und lernen, wie wir fortgeschrittene statistische Methoden und Machine Learning-Techniken anwenden können, um komplexe Datenprobleme zu lösen. Ebenso werden wir uns mit fortgeschrittenen Visualisierungstechniken beschäftigen, um unsere Analysen zu veranschaulichen und zu interpretieren. Natürlich wird man auch nach diesem Kurs nicht alles wissen, was es über Data Science zu wissen gibt, aber das Ziel ist immer ein Gefühl dafür zu bekommen was möglich ist und zumindest den Ansatz zu kennen um sich in neue Themen einzuarbeiten.

## Von Data Analytics zu Data Science: Der Unterschied

Während Data Analytics sich hauptsächlich auf die **Beschreibung und Aufbereitung** von Daten konzentriert (deskriptive und diagnostische Analyse), geht Data Science einen wichtigen Schritt weiter: Es umfasst auch **prädiktive und präskriptive Analysen**. Mit anderen Worten:

- **Data Analytics** beantwortet die Frage: "Was ist passiert und warum?"
- **Data Science** fügt hinzu: "Was wird passieren und was sollten wir tun?"

Diese Erweiterung des Fokus erfordert zusätzliche Methoden und Techniken, die über die grundlegende Datenanalyse hinausgehen.

 Verweis Data Analyst Workshop

Die Diskussion über die nicht ganz einfache Abgrenzung der verschiedenen Bereiche hatten wir bereits im Kapitel 1.2 Data Analytics.

## Die Säulen der Data Science und angrenzende Gebiete

Oft wird geschrieben, dass Data Science auf mehreren Säulen ruht. Eine Art diese zu unterteilen ist: Statistik, Machine Learning, Data Engineering und Fachwissen. Letzteres wird in diesem Kurs natürlich nicht behandelt, da es voll davon abhängt an welcher Position ihr arbeitet. Die anderen drei werden wir aber genauer betrachten.

### 1. Fortgeschrittene Statistik

Im ersten Teil haben wir uns mit deskriptiver Statistik beschäftigt - den grundlegenden Methoden zur Beschreibung von Daten. Nun erweitern wir unser statistisches Toolkit um:

- **Inferenzstatistik:** Treffen von Schlussfolgerungen über Populationen basierend auf Stichproben
- **Hypothesentests:** Formale Methoden zur Überprüfung von Annahmen über Daten
- **Regression und Korrelationsanalysen:** Untersuchung von Beziehungen zwischen Variablen
- **Zeitreihenanalyse:** Analyse von zeitabhängigen Daten

- **Multivariate Statistik:** Analyse von mehreren Variablen gleichzeitig

## 2. Machine Learning

Machine Learning ist das Herzstück moderner Data Science. Es ermöglicht Computern, aus Daten zu lernen und Vorhersagen zu treffen, ohne explizit programmiert zu werden. Wir werden uns mit folgenden Bereichen befassen:

- **Supervised Learning** (Überwachtes Lernen):
  - Klassifikationsalgorithmen
  - Regressionsalgorithmen
  - Ensemble-Methoden
- **Unsupervised Learning** (Unüberwachtes Lernen):
  - Clustering
  - Dimensionsreduktion
  - Anomalieerkennung
- **Model Evaluation** (Modellbewertung):
  - Kreuzvalidierung
  - Hyperparameter-Tuning
  - Metriken zur Modellbewertung

## 3. Data Engineering Grundlagen

Als Data Scientist benötigt man auch grundlegende Kenntnisse in Data Engineering, um effektiv mit Daten arbeiten zu können. Die Grenzen zwischen Data Science und Data Engineering sind in der Praxis oft fließend, und viele Aufgaben erfordern Kenntnisse aus beiden Bereichen:

- **Datenbankanbindung:** Verbindung zu SQL- und NoSQL-Datenbanken
- **API-Integration:** Nutzung von Web-APIs zur Datengewinnung
- **ETL-Prozesse:** Extraktion, Transformation und Laden von Daten
- **Big Data-Konzepte:** Grundlagen der Arbeit mit sehr großen Datensätzen

## 4. Fortgeschrittene Datenvisualisierung

Aufbauend auf den Grundlagen aus Teil 1 werden wir komplexere Visualisierungstechniken kennenlernen. Im Gegensatz zu den anderen Themenbereichen, tauchen die Kapitel zur Visualisierung nicht gebündelt auf, sondern sind über den Kursinhalt verteilt.

Diese integrierte Herangehensweise sorgt nicht nur für Abwechslung im Kurs, sondern vermittelt auch ein besseres Verständnis dafür, welche Visualisierungstechniken für welche Arten von Analysen am besten geeignet sind.

All dies wird außerdem regelmäßig mit Best Practice Empfehlungen aus der Praxis ergänzt..