

Datenaufbereitung und Data Cleaning mit Python

Datenqualität erkennen, verbessern und reproduzierbar sicherstellen

01 INTRO

Reale Daten sind selten unmittelbar analysierbar. Fehlende Werte, unterschiedliche Formate, Dubletten, ungültige Werte, widersprüchliche Kodierungen und fachliche Inkonsistenzen gehören zum Alltag der Datenanalyse. Ein erheblicher Teil professioneller Data-Analytics-Arbeit besteht deshalb aus der systematischen Aufbereitung und Validierung von Daten.

Der Kurs konzentriert sich auf diesen entscheidenden Arbeitsschritt. Mit Python und Pandas werden Datenbestände untersucht, Qualitätsprobleme identifiziert, Bereinigungsregeln entwickelt und reproduzierbare Cleaning-Prozesse aufgebaut. Dabei geht es nicht nur um technische Korrekturen, sondern auch um die Frage, wann Daten überhaupt verändert werden dürfen und wie Änderungen nachvollziehbar bleiben.

DAUER

2 Tage

FORMAT

Präsenz · Live Online · Inhouse

VORKENNTNISSE

Python-Grundkenntnisse und grundlegender Umgang mit Pandas

02 TRAININGSPROGRAMM

Inhalte

Datenqualität in der Praxis

- typische Qualitätsprobleme
- technische und fachliche Datenqualität
- Vollständigkeit
- Konsistenz
- Validität
- Eindeutigkeit
- Aktualität und Plausibilität

Daten systematisch profilieren

- Struktur eines Datensatzes untersuchen
- Datentypen
- Wertebereiche
- Häufigkeiten
- Kardinalitäten
- fehlende Werte
- auffällige Verteilungen
- Data Profiling als Ausgangspunkt

Fehlende Werte

- Arten fehlender Informationen
- fehlende Werte erkennen
- technische Repräsentationen
- Löschen oder Ersetzen
- Imputation
- fachliche Auswirkungen
- Missingness dokumentieren

Datentypen und Formate bereinigen

- fehlerhafte Datentypen
- numerische Werte
- Datumsformate
- boolesche Werte
- Kategorien
- Zeichensätze
- Dezimal- und Tausendertrennzeichen
- robuste Konvertierung

Textdaten normalisieren

- Whitespace
- Groß-/Kleinschreibung
- Sonderzeichen
- Schreibvarianten
- reguläre Ausdrücke
- Extraktion aus Textfeldern
- standardisierte Repräsentationen

Dubletten und Identitäten

- vollständige Dubletten
- fachliche Dubletten
- Schlüssel
- zusammengesetzte Identitäten
- unterschiedliche Schreibweisen
- Regeln für Deduplizierung

Ungültige Werte und Ausreißer

- Wertebereiche
- Plausibilitätsregeln
- statistische Ausreißer
- fachliche Ausreißer
- Fehler versus legitime Extremwerte
- Korrektur, Markierung oder Ausschluss

Inkonsistente Datenbestände

- widersprüchliche Kategorien
- Einheiten
- Kodierungen
- Referenzdaten
- Abhängigkeiten zwischen Feldern
- Cross-Field-Validation

Daten aus mehreren Quellen harmonisieren

- unterschiedliche Schemata
- Spaltenzuordnung
- Schlüssel
- Einheiten und Formate
- Mapping-Tabellen
- Referenzdaten
- Konflikte beim Zusammenführen

Validierungsregeln entwickeln

- technische Constraints
- fachliche Regeln
- Assertions
- erwartete Wertebereiche
- Schemaprüfungen
- Fehlerberichte
- wiederverwendbare Validatoren

Reproduzierbare Cleaning-Pipelines

- Cleaning-Schritte strukturieren
- Funktionen statt Notebook-Handarbeit
- Reihenfolge von Transformationen
- Idempotenz
- Logging
- Nachvollziehbarkeit
- Originaldaten erhalten
- Tests für Datenbereinigung

Praktisches Cleaning-Projekt

- problematischen Datensatz profilieren
- Qualitätsprobleme klassifizieren
- Bereinigungsstrategie entwickeln
- Regeln implementieren
- Ergebnis validieren
- Änderungen dokumentieren

03 DURCHFÜHRUNG

Rahmen und Organisation

Dauer	2 Tage
Durchführung	Präsenz oder Live Online, vorzugsweise als Inhouse-Schulung
Schwerpunkt	Inhouse-Schulungen für Unternehmen und Teams
Sprache	Deutsch, englische Fachbegriffe und Dokumentation
Unterlagen	Kursunterlagen, Beispiele und Übungsaufgaben
Technik	Eigener Rechner mit Python, Jupyter und den gängigen Analysebibliotheken

Inhouse-Schulungen

Eigene problematische Datenbestände können als Grundlage für die Übungen verwendet werden.

04 ZIELGRUPPE

Für wen ist der Kurs gedacht?

Data Analysts, Entwickler und Fachanwender, die reale Datenbestände für Analysen, Reporting oder weitere Verarbeitung zuverlässig aufbereiten müssen.

Voraussetzungen

Grundkenntnisse in Python und ein grundlegender Umgang mit Pandas.

05 LERNZIELE

Was Sie nach dem Kurs können

Nach dem Kurs können die Teilnehmer Datenqualitätsprobleme systematisch erkennen, geeignete Bereinigungsstrategien auswählen, Cleaning-Regeln mit Python implementieren und reproduzierbare sowie nachvollziehbare Aufbereitungsprozesse entwickeln.

06 FAQ

Häufige Fragen

Ist Data Cleaning nicht bereits Bestandteil der Datenanalyse mit Pandas?

Ja, als notwendiger Bestandteil einer Datenanalyse. Dieser Kurs behandelt Datenqualität und Datenbereinigung jedoch wesentlich tiefer und als eigenständigen Engineering-Prozess.

Werden Ausreißer automatisch entfernt?

Nein. Ein ungewöhnlicher Wert ist nicht automatisch falsch. Der Kurs behandelt ausdrücklich die Unterscheidung zwischen statistischer Auffälligkeit und fachlichem Datenfehler.

Geht es nur um Pandas?

Pandas ist das zentrale Werkzeug. Die vermittelten Prinzipien zu Datenqualität, Validierung und reproduzierbarer Bereinigung sind jedoch allgemeiner.

UC IT Service · Ulrich Cuber · kontakt@uc-it.de
<https://www.uc-it.de/coaching/data-cleaning/>