DATA ANALYST

Von Daten zu Entscheidungen – praxisorientiert, interaktiv und anwendungsnah.

IHK Akademie München und Oberbayern

KENNENLERNEN

- Name
- Rolle / Firma / …
- Erfahrungen mit Datenanalyse
- Welche Tools hast du bereits verwendet für Datenanalyse und/oder Visualisierung?
- Gab es in deinem Alltag schon Mal eine Daten-Herausforderung?
- Was sind deine Erwartungen an den Workshop?

WORKSHOP – DATA ANALYST

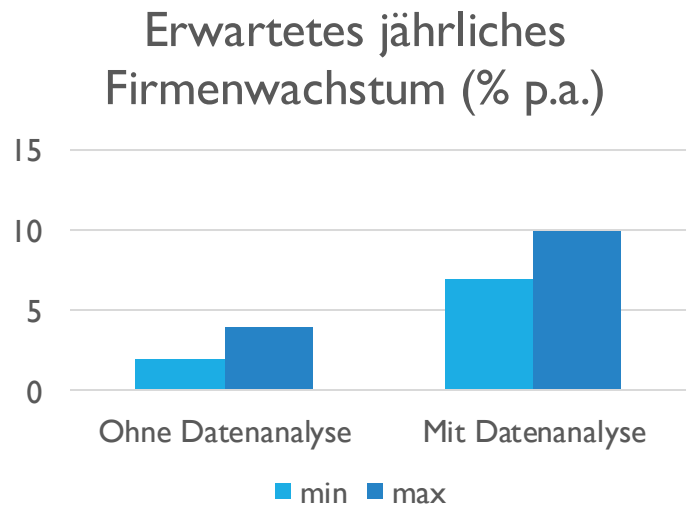
- **Trainer:** Dr. Denis Düsseldorf
- **Zeitraum:** 17.11. – 04.12.2025
  - 5 Module, 10 Tage
- **Format:** Online, Live
- **Unterlagen:** Digital verfügbar während der Veranstaltung
- **Zertifizierung:** Durch Abschlussprojekt mit anschließender Präsentation in Gruppen

WARUM EIGENTLICH DATA ANALYSIS?

- In der digitalen Welt generieren Unternehmen enorme Datenmengen
- Oft werden Daten bereits seit mehreren Jahrzehnten gesammelt
  - Vielen Unternehmen mangelte es bisher oft an Ressourcen um diese Daten auszuwerten
- Durch „Big Data“ und „Künstliche Intelligenz“ gewinnt die Datenanalyse stetig an Bedeutung im industriellen Umfeld

Potenziale einer guten Datenanalyse

- Optimierung von Geschäftsprozessen
- Datenbasierte Entscheidungen treffen:
  - Harte Fakten statt Bauchgefühl
- Wettbewerbsvorteile finden und nutzen

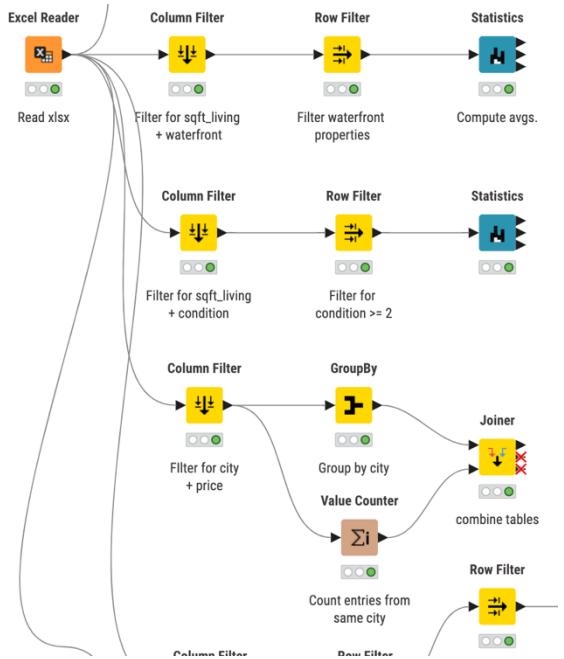


Basiert auf Quelle: Big data: Getting a better read on performance, McKinsey

MODUL 1: GRUNDLAGEN DER DATA ANALYTICS – DER ETL-PROZESS

- Einführung und Definition der Ziele des Workshops
- Verständnis des ETL-Prozesses (Extract, Transform, Load)
- Aufgaben und Funktionen der Data Analytics im Unternehmen
- Analyse, Organisation und Dokumentation von Datenprozessen
- Einführung in KNIME und erste ETL-Workflows
- Explorative Datenanalyse

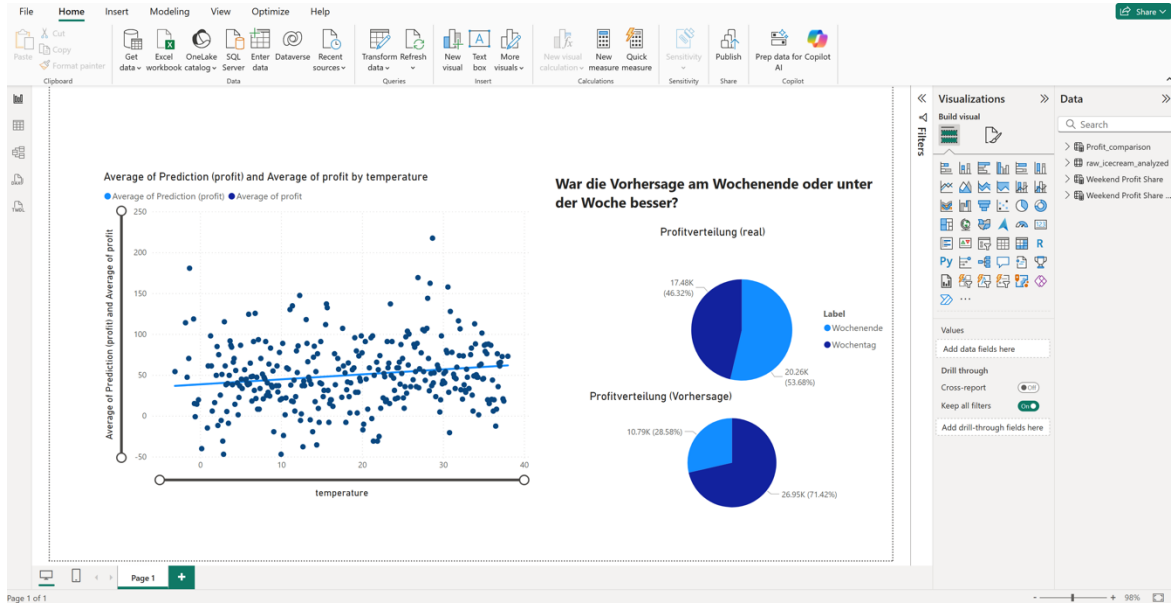
Tools: KNIME



MODUL 2: VISUELLE ANALYSE UND REPORTING – BI-TOOLS MIT MICROSOFT POWER BI

- Einführung in Microsoft Power BI
- Übernahme von ETL-Ergebnissen aus KNIME
- Feintuning der Datenanalyse-Ergebnisse
- Aufbau aussagekräftiger Visualisierungen
- Erstellung von interaktiven Dashboards

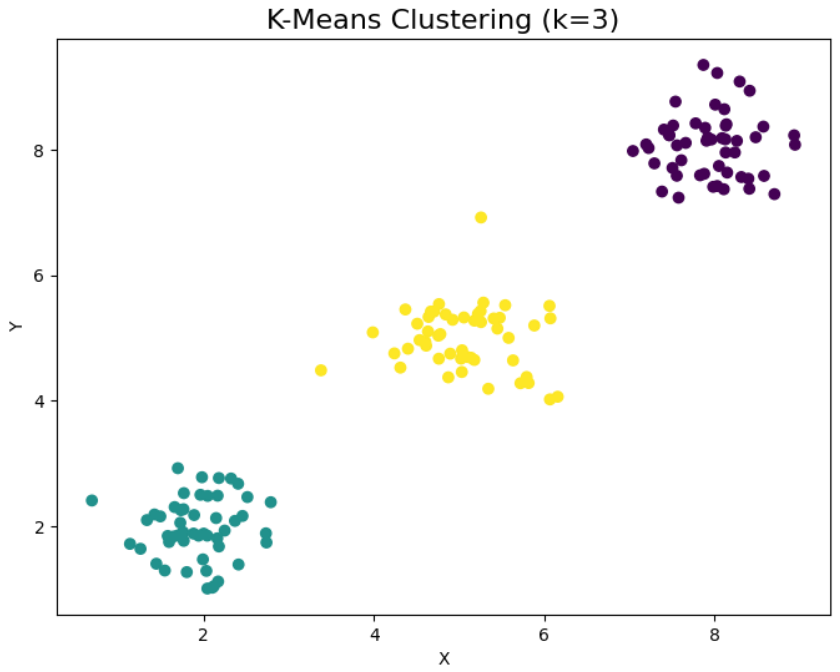
Tools: Microsoft Power BI



MODUL 3: DATA ANALYTICS FÜR FORTGESCHRITTENE – MACHINE LEARNING & WORKFLOW CONTROL

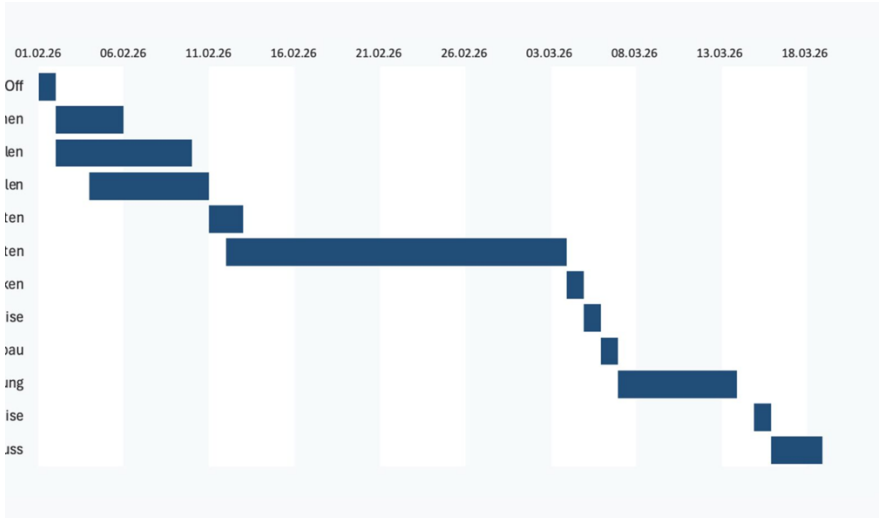
- Theoretische Grundlagen des Machine Learning
- Identifikation von ML-Einsatzpotenzialen
- Anwendung von ML-Algorithmen in KNIME
- Workflow Control und Automatisierung
- Datenbankanbindung in KNIME und Power BI

Tools: KNIME & Microsoft Power BI



MODUL 4: EIGENE DATENPROJEKTE – BEWERTEN, PLANEN, UMSETZEN

- Planung und Konzeption eigener Datenprojekte
- Definition und Kommunikation von Projektzielen
- Umsetzung datenbasierter Lösungen im Team
- Anwendung agiler Methoden zur Projektsteuerung
- Effizienzsteigerung durch iterative Vorgehensweisen



Tools: KNIME & Microsoft Power BI

MODUL 5: PRAXISTRAINING – PROJEKTARBEIT, BEWERTUNG & ABSCHLUSS

- Gruppenarbeit: Konzeption, Durchführung und Auswertung
- Trainer-Feedback zur Projektarbeit
- Vorbereitung der Abschlusspräsentation
- Bewertung der Projektarbeiten
- Abschlussprüfung (Präsentation + mündlich)

Praxistraining mit Trainer: 27.11.2025

Bewertung der Projektarbeit: 01.12.2025

Prüfung: 04.12.2025

ERWARTUNGEN UND REGELN

- Aktive Beteiligung
- Pünktlichkeit (auch bei der Abgabe des Abschlussprojektes)
- Anwesenheit
- Bitte während der Online-Meetings Kameras einschalten

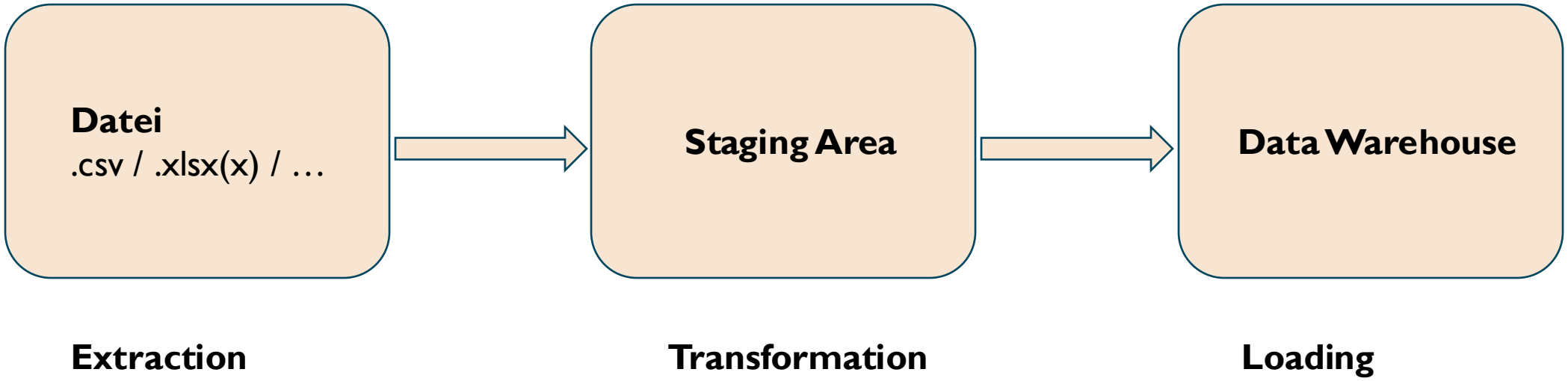
MODUL 1
AGENDA

- Verständnis des ETL-Prozesses
- Aufgaben und Funktion der Data Analytics
- Analyse, Organisation und Dokumentation von Datenprozessen
- **KNIME und ETL-Workflows**
- Explorative Datenanalyse (EDA)
- Ports - Arten des Datentransfers in KNIME

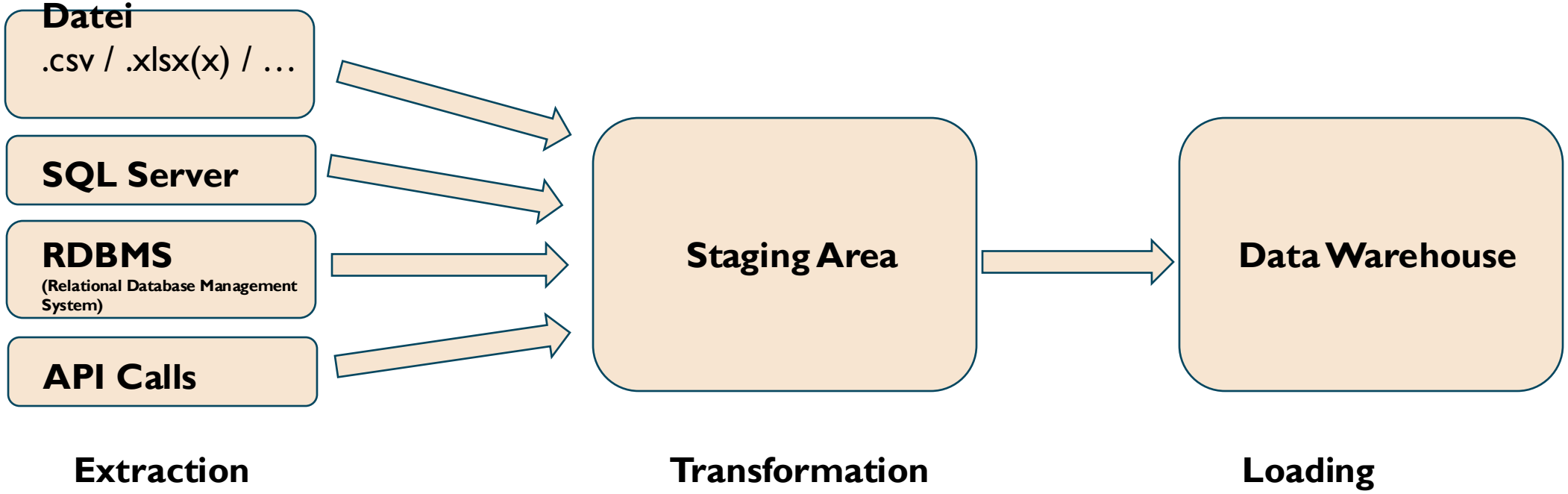
EXTRACT → TRANSFORM → LOAD

VERSTÄNDNIS DES ETL-PROZESSES

DER ETL-PROZESS (EINFACHSTER FALL)



DER ETL-PROZESS (GENERELL)



DER ETL-PROZESS (GENERELL)

Der ETL-Prozess ist wie Kochen

WARUM ETL? VORTEILE UND BEISPIELE

Vorteile

- Datenkonsistenz
- Effizienz
- Entscheidungsfindung

Herausforderungen

- Komplexität bei Big Data
- Performance bei großen Datenmengen
- Sicherstellung der Datenqualität

Anwendungsbeispiele

- Kundenanalyse im Marketing, bspw. um neue Rabattaktionen auf eine spezielle Kundengruppe zuzuschneiden (Verkaufsdaten, Social Media Daten)
- Reporting und Controlling in der Finanzbranche, bspw. um bestehende Investments neu zu bewerten und neue Möglichkeiten zu identifizieren

SCHRITT 1: **EXTRACTION**

1

- Daten aus einer / mehreren Quelle(n) müssen gesammelt werden
  - Aus Dateien
  - Über API Calls
  - Aus Datenbanken

Herausforderungen: Volumen, Vielfalt, Echtzeit vs. Batch Daten

Beispiel: Extraktion von Verkaufsdaten aus Customer Relationship Management (CRM) Systemen

Beispiel: Formatierung zweier Datenquellen weicht signifikant voneinander ab

```
order_id,customer_id,product,category,quantity,price
10001,501,"Bluetooth Speaker","Electronics",1,49.99,
10002,502,"Running Shoes","Sportswear",2,69.50,139.0
10003,503,"Coffee Mug","Home & Kitchen",3,8.90,26.70
10004,504,"Wireless Mouse","Electronics",1,24.99,24.
10005,505,"Yoga Mat","Sportswear",1,29.99,29.99,2025
```

```
order_id;order_part;customer_id;product;category;quantity;price;
10001;1;501;Bluetooth Speaker;Electronics;1;49,99;49,99;2025-10-
10001;2;501;Bluetooth Speaker;Electronics;1;49,99;49,99;2025-10-
10002;1;502;Running Shoes;Sportswear;1;69,50;69,50;2025-10-01;Pa
10002;2;502;Running Shoes;Sportswear;1;69,50;69,50;2025-10-01;Pa
10003;1;503;Coffee Mug;Home & Kitchen;2;8,90;17,80;2025-10-02;Cr
```

SCHRITT 2: **TRANSFORMATION**

2

Ziel: Daten konsistent und analysierbar machen und analysieren! Z.B. durch:

- Daten bereinigen
  - Duplikate entfernen, Fehler korrigieren, Einheiten angleichen
- Formatieren
  - Datumsformate angleichen
  - Währungsumrechnung
  - Adressstandardisierung
- Aggregieren
  - Summen bilden, Mittelwerte und andere Kennzahlen berechnen
  - Visualisierungen, Machine Learning
  - ..

SCHRITT 2: **TRANSFORMATION**

2

- Eine gute Datenqualität muss gewährleistet werden
- Ansonsten ist eine vernünftige Datenanalyse nicht möglich!

„Garbage in, garbage out“

George Fuechsel, 1965

WAS IST "DATENQUALITÄT"?

Datenqualität ist ein Maß für die Güte und Zuverlässigkeit von Daten. Sie beschreibt, wie geeignet Daten für Analyse- und Entscheidungszwecke sind

Typische Aspekte:

- **Vollständigkeit:** Keine fehlenden Werte
- **Korrektheit:** Daten stimmen mit der Realität überein
- **Konsistenz:** Gleiche Werte in allen Systemen und Formaten
- **Aktualität:** Daten sind auf dem neuesten Stand
- **Eindeutigkeit:** Keine Duplikate
- **Relevanz:** Daten passen zum Analysezweck
- **Validität:** Daten entsprechen definierten Formaten und Regeln

ANZUSTREBENDE DATENQUALITÄT

- Einheitliche Datenquellen mit klaren Standards
- Regelmäßige Datenprüfung und Bereinigung
- Automatisierte Datenvalidierung im ETL-Prozess
- Dokumentierte Datenherkunft (Data Lineage)
- Verantwortlichkeiten für Datenqualität festgelegt
- Aufbau einer „Single Source of Truth“
- Kontinuierliche Qualitätskontrolle durch Monitoring und KPIs

SCHRITT 3: **LOADING**

3

- Daten müssen in Zielsysteme geladen werden, bspw.
  - in eine neue Datei (nicht skalierbar, meist nur zu Testzwecken)
  - in eine Datenbank / RDBMS
  - in ein Data Warehouse
- Man unterscheidet zwischen **Full Load** und **Incremental Load**
  - **Full Load:** Die gesamten transformierten Daten werden in das Zielsystem geladen
  - **Incremental Load:** Nur Änderungen werden in bestehende Daten des Zielsystems eingepflegt

ARTEN VON ZIELSYSTEMEN

3

	Datei
Zweck	Archivieren, testen
Datenvolumen	Klein bis mittel
Aktualisierung	Batch oder on-demand
Abfragen	Dateioperationen

ARTEN VON ZIELSYSTEMEN

3

	Datei	Datenbank
Zweck	Archivieren, testen	Operative Verarbeitung
Datenvolumen	Klein bis mittel	Relativ klein bis mittel
Aktualisierung	Batch oder on-demand	Häufig oder sogar in Echtzeit
Abfragen	Dateioperationen	Viele kleine Transaktionen

ARTEN VON ZIELSYSTEMEN

3

	Datei	Datenbank	Data Warehouse
Zweck	Archivieren, testen	Operative Verarbeitung	Analytische Auswertung
Datenvolumen	Klein bis mittel	Relativ klein bis mittel	Sehr groß (historische Daten)
Aktualisierung	Batch oder on-demand	Häufig oder sogar in Echtzeit	Periodisch, z.B. täglich oder wöchentlich
Abfragen	Dateioperationen	Viele kleine Transaktionen	Komplexe Aggregationen und Analysen

ARTEN VON ZIELSYSTEMEN

3

	Datei	Datenbank	Data Warehouse	Data Lake
Zweck	Archivieren, testen	Operative Verarbeitung	Analytische Auswertung	Speicherung heterogener, großer Datenmengen
Datenvolumen	Klein bis mittel	Relativ klein bis mittel	Sehr groß (historische Daten)	Sehr groß, beliebige Datenmengen
Aktualisierung	Batch oder on-demand	Häufig oder sogar in Echtzeit	Periodisch, z.B. täglich oder wöchentlich	Batch oder Streaming
Abfragen	Dateioperationen	Viele kleine Transaktionen	Komplexe Aggregationen und Analysen	Komplexe Analysen, oft über externe Tools

AUFGABEN UND FUNKTION DER DATA ANALYTICS IN UNTERNEHMEN

WAS IST DATA ANALYTICS?

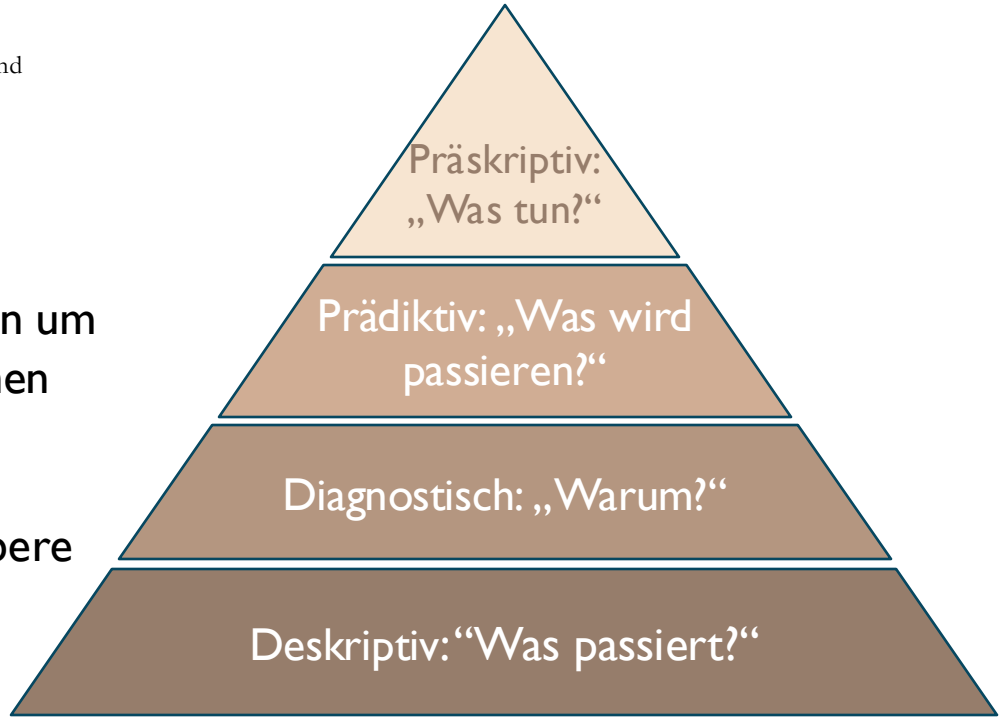
Data Analytics liefert Einblicke für bessere Entscheidungen, Prozessoptimierung und Trendvorhersagen.

Basierend auf Definitionen von Deloitte und Talend

Hauptbestandteil:

Strukturierte Analyse von Daten um (verborgene) Muster zu erkennen

Achtung: Hierfür werden saubere Daten benötigt (ETL)



KERNFUNKTIONEN

Anwendungsbereich	Branchenbeispiele
Trends erkennen und Vorhersagen treffen	**Modebranche –** Analyse von Social-Media-Trends zur Sortimentsplanung
Prozesse optimieren (bspw. für das Lieferkettenmanagement)	**Logistik –** Optimierung des Lieferkettenmanagements durch Echtzeit-Tracking und Bestandsanalyse
Kundenverhalten analysieren	**E-Commerce –** Analyse von Klick- und Kaufverhalten zur Personalisierung von Angeboten
Verkaufsprognosen für Handelsunternehmen erstellen	**Einzelhandel –** Prognosen basierend auf saisonalen Mustern und historischen Verkaufsdaten
Risiken minimieren (Fraud Detection)	**Finanzwesen –** Fraud Detection bei Kreditkartentransaktionen durch Anomalieerkennung

AUFGABEN EINES DATA ANALYSTS

- Daten sammeln
- Daten analysieren
- Daten (und Ergebnisse) visualisieren
- Reports erstellen
- Mit Stakeholdern kommunizieren

- Daten aus der echten Welt sind im Normalfall viel zu komplex um „per Hand“ analysiert zu werden. Stattdessen bedient sich der Data Analyst einer Reihe von Tools
  - Tabellenkalkulation, bspw. Microsoft Excel
  - SQL zur Kommunikation mit Datenbanken
  - Python-Kenntnisse
  - KNIME
  - Business Intelligence Tools, bspw. Microsoft PowerBI

AUFGABEN EINES DATA ANALYSTS

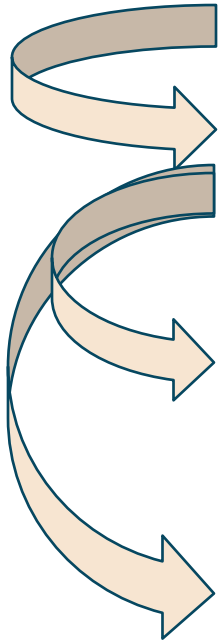
Modul 1 & 3
- Daten sammeln
- Daten analysieren

Modul 2
- Daten (und Ergebnisse) visualisieren

Modul 4
- Reports erstellen
- Mit Stakeholdern kommunizieren

- Daten aus der echten Welt sind im Normalfall viel zu komplex um „per Hand“ analysiert zu werden. Stattdessen bedient sich der Data Analyst einer Reihe von Tools
  - Tabellenkalkulation, bspw. Microsoft Excel
  - SQL zur Kommunikation mit Datenbanken
  - Python-Kenntnisse
  - KNIME
  - Business Intelligence Tools, bspw. Microsoft PowerBI

ROLLEN IM DATA ANALYTICS TEAM



- Der **Data Engineer** pflegt die technische Infrastruktur (Datenbanken, Pipelines)
- Der **Data Scientist** ist verantwortlich für die Erstellung von datenbasierten Modellen
- Der **Data Analyst** ist verantwortlich für die Analyse vorhandener Daten, erstellt Berichte und identifiziert Muster oder Trends zur Unterstützung von Entscheidungen
- Der **BI Analyst** arbeitet die Daten visuell ansprechend auf

Synergie: Gemeinsam bilden diese Rollen das Rückgrat datengetriebener Unternehmen: Der Engineer liefert die Datenbasis, der Scientist modelliert, der Analyst interpretiert und der BI-Analyst kommuniziert die Ergebnisse.

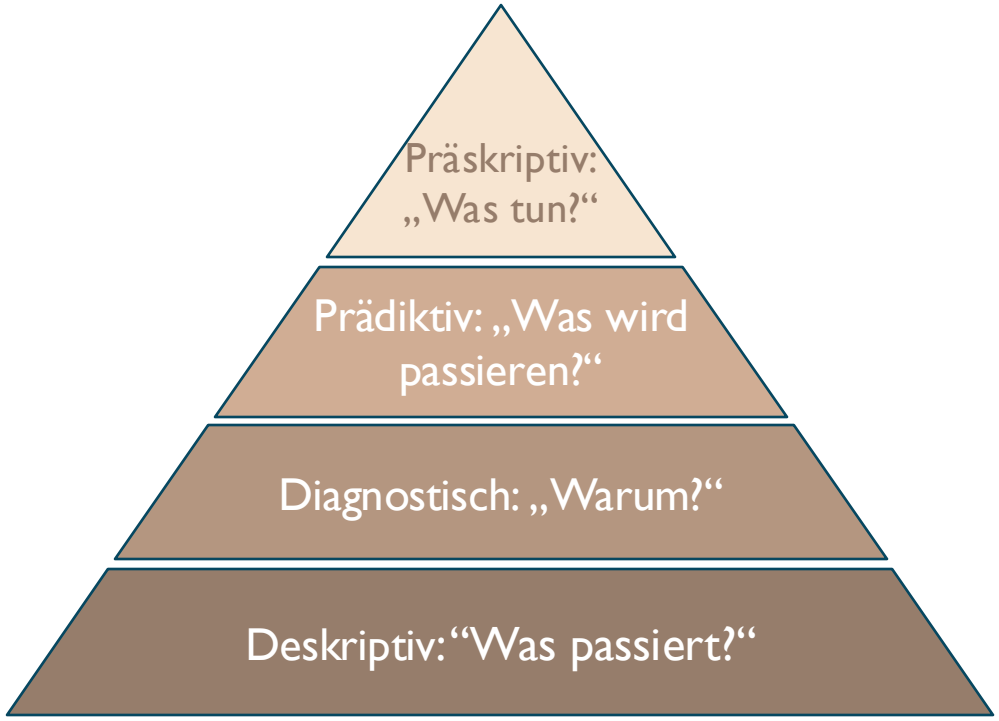
Zwischen dem Data Analyst und BI Analyst kann keine klare Grenze gezogen werden, da die Bereiche in vielen Unternehmen überlappen und beide am Ende der Datenkette stehen. Aus diesem Grund beschäftigen wir uns in diesem Workshop mit der Analyse und Visualisierung von Daten gleichermaßen!

FALLBEISPIELE IM DETAIL

NETFLIX

- Prädiktive Analyse und Machine Learning für personalisierte Empfehlungen
- Analyse von Sehgewohnheiten, Bewertungen, Abbruchraten

Content-Investitionen können gezielter gesteuert und die Kundenbindung deutlich erhöht werden

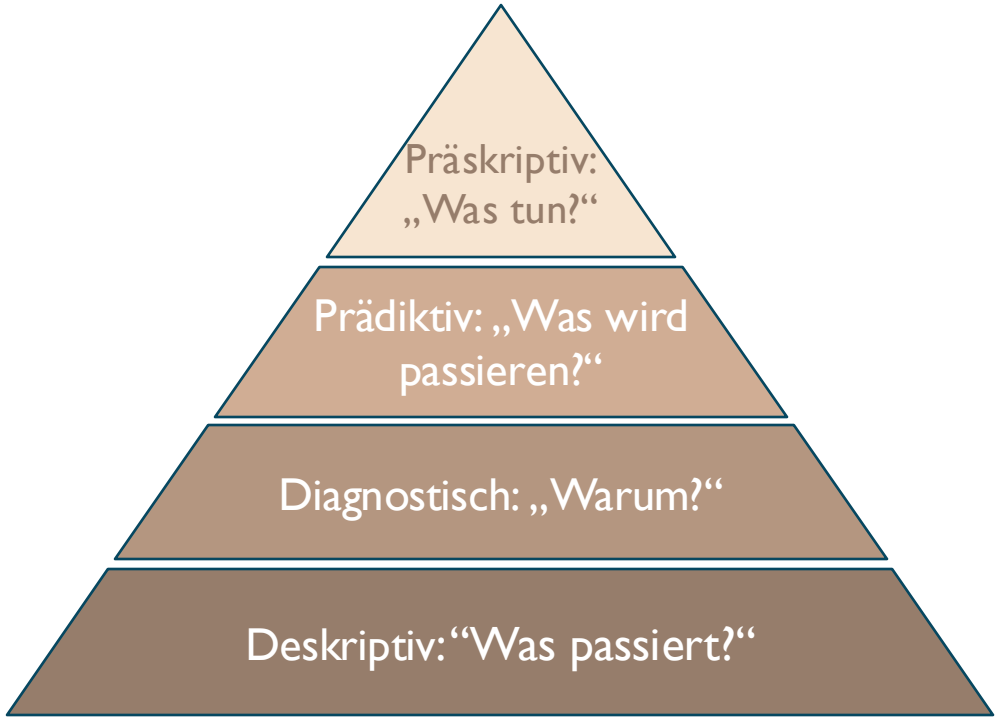


FALLBEISPIELE IM DETAIL

SIEMENS

- Data Analytics in der Maschinenproduktion
- Maschinendaten werden in Echtzeit gesammelt und ausgewertet um prädiktive Wartung möglich zu machen

Wartungsbedarf kann erkannt werden, bevor eine Maschine ausfällt. Das reduziert Ausfallzeiten und spart Geld.

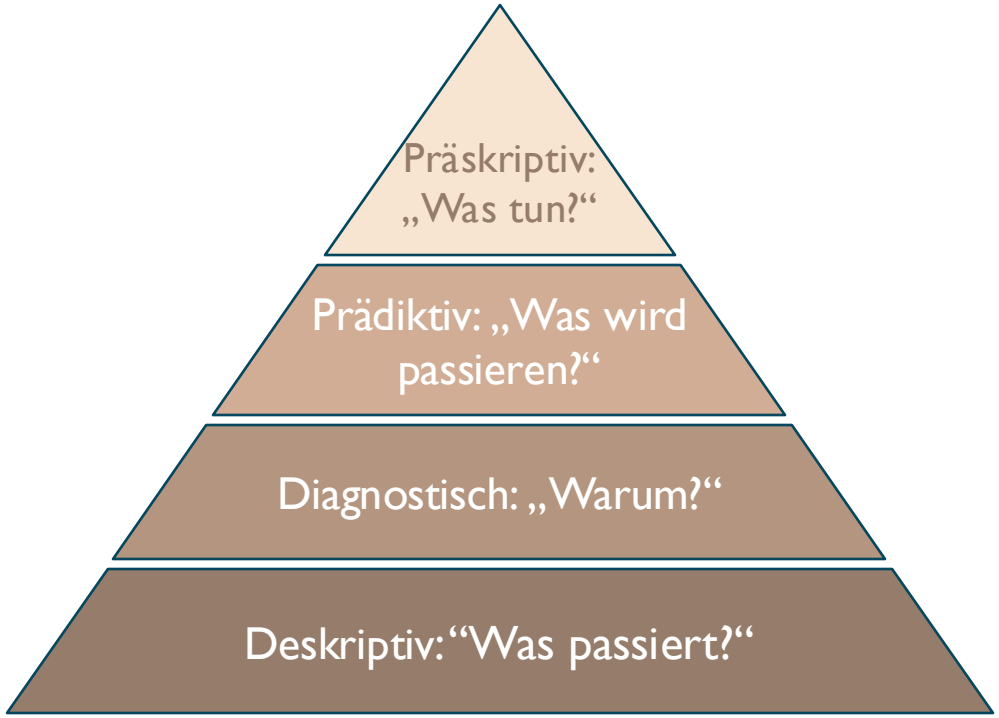


FALLBEISPIELE IM DETAIL

AMAZON

- Intensive Nutzung von Data Analytics in der Logistik und Lieferkette
- Echtzeitdaten zu Beständen, Nachfrage, Versandrouten
- Vorhersagen

Lieferzeiten können minimiert werden, Engpässe werden erkannt, Preise können dynamisch gestaltet werden.



Frage: Habt ihr andere Beispiele für die Anwendung von Datenanalyse in Unternehmen?

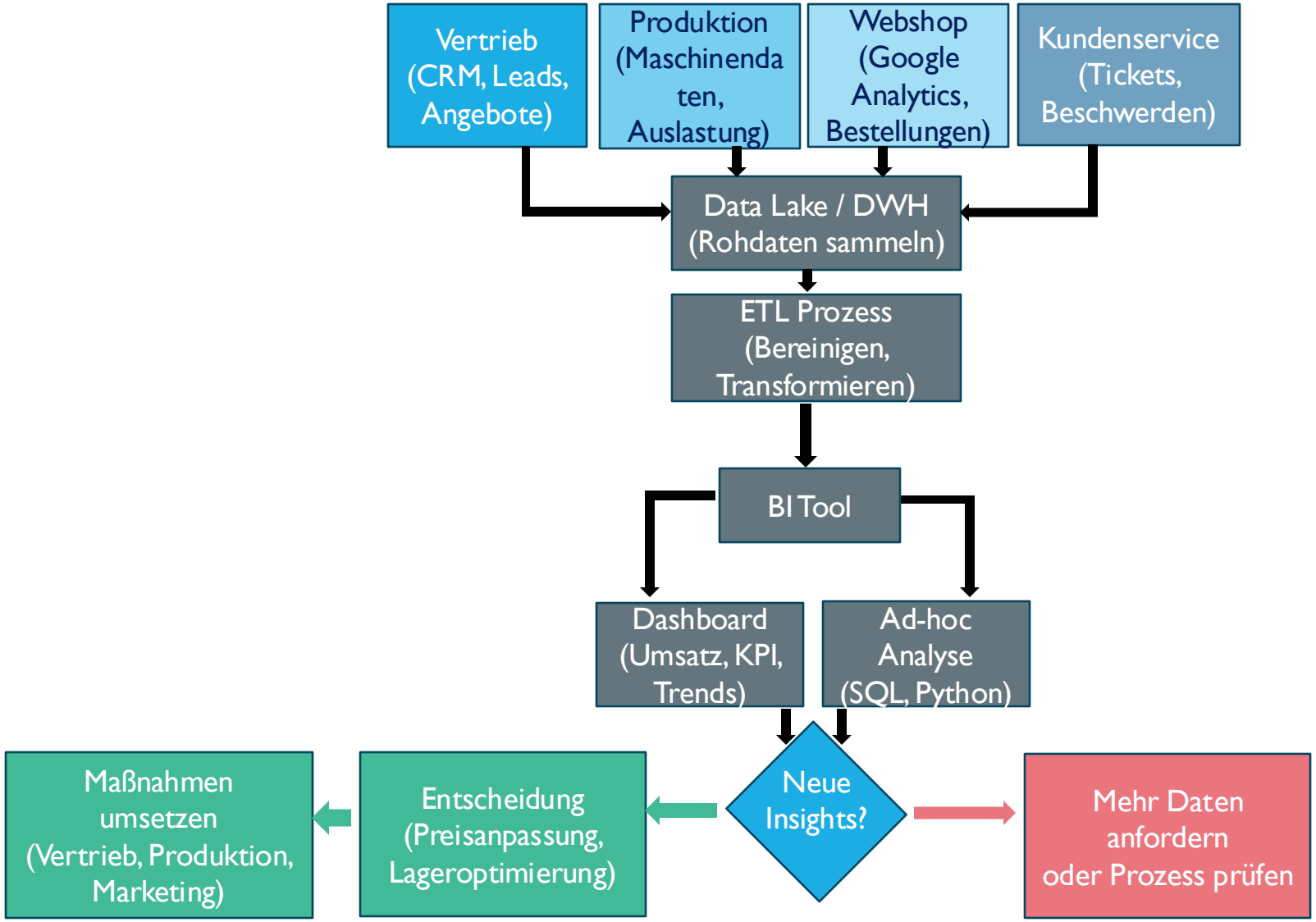
ANALYSE, ORGANISATION UND DOKUMENTATION VON DATENPROZESSEN

ANALYSE VON DATENPROZESSEN

- Die Analyse von Datenprozessen beginnt mit der **Identifikation zentraler Abläufe**
  - Von der Datenerfassung, über die Verarbeitung, bis hin zur Nutzung
- Im Anschluss können Datenflüsse visualisiert und Abhängigkeiten zwischen Systemen und Akteuren dargestellt werden
  - In diesem Schritt werden Schwachstellen (bspw. Bottlenecks, doppelte Datenerhaltung, Medienbrüche) erkannt und bewertet

Die Analysephase ist die Vorbereitung für den ETL-Prozess: Wer seine Datenflüsse versteht, kann ETL-Pipelines gezielter und effizienter gestalten!

BEISPIELHAFTE FLOWCHART



ORGANISATION VON DATENPROZESSEN

- Daten sollten kategorisiert sein
- Man unterscheidet zwischen strukturierten und unstrukturierten Daten

Strukturierte Daten sind direkt analysierbar und tabellarisch darstellbar, z.B.

- Kundendaten (Name, Kundennummer, Adresse, Umsatz)
- Verkaufszahlen (Produkt-ID, Datum, Menge, Preis)
- Lagerbestände (Artikelnummer, Standort, Stückzahl)
- Sensorwerte (Zeitstempel, Temperatur, Druck)

Unstrukturierte Daten sind nicht tabellarisch und schwerer zu analysieren, z.B.

- E-Mails / Chatverläufe
- Bilder / Videos / Audioaufnahmen (nicht nur Userbezogen, auch bspw. aus der Fertigung: Fotos aller produzierten Teile zur Fehlererkennung)
- Freitext-Kommentare im Kundenfeedback
- Dokumente mit gemischten Inhalten (PDF, Präsentationen)

ORGANISATION VON DATENPROZESSEN

- Neben der Kategorisierung von Daten ist ebenfalls die Speicherstrategie der Daten wichtig:
  - Speichern in der Cloud oder On-Premise (= lokal vor Ort)
- Es müssen Zugriffsregeln definiert werden
  - Wer darf was sehen?

- **Best Practices:** Erstellen eines Datenkatalogs, sowie Metadatenmanagement

ORGANISATION VON DATENPROZESSEN

DATENKATALOG

Ein Datenkatalog ist ein zentrales Verzeichnis, das Informationen über alle Datenbestände einer Organisation speichert.

- Hilft Mitarbeitern dabei, Daten schnell zu finden, zu verstehen und zu nutzen
- Unterstützt das Datenmanagement
- Verbessert Datenqualität
- Fördert die Zusammenarbeit
- Vereinfacht Compliance (Einhaltung von Richtlinien, Gesetzen, Vorschriften und internen Regeln)

METADATENMANAGEMENT

Das Metadatenmanagement bezieht sich auf die systematische Organisation, Verwaltung und Pflege von Metadaten, d. h. von Daten, die Informationen über andere Daten liefern.

- Enthält Infos über Struktur, Kontext (Definitionen, Datenbesitzer), Herkunft und Zweck und Lebenszyklus von Daten in einem Unternehmen
- Entscheidend für effektive Datenverwaltung
- Verbessert Zugänglichkeit, Genauigkeit, Konsistenz und Sicherheit von Daten

DOKUMENTATION VON DATENPROZESSEN

- Datenflüsse in Unternehmen werden meist nach einheitlichen Dokumentationsleitfäden erstellt
- Dort werden auch Schnittstellen und Verantwortlichkeiten definiert
- Der Dokumentationsleitfaden kann auch verlangen, dass zur visuellen Aufbereitung von Datenprozessen Flowcharts mit standardisierten Symbolen für die verschiedene Schritte verwendet werden müssen
  - Das nennt man BPMN-Diagramm (**B**usiness **P**rocess **M**odel and **N**otation)
- Hierdurch wird es für andere Personen derselben Organisation einfacher, die Visualisierung zu verstehen.

GUIDELINES FÜR ETL & DATENPROZESS-MANAGEMENT

- **Datenquellen zentralisieren:** Alle relevanten Daten an einem Ort sammeln
- **Datenqualität sichern:** Fehlende, doppelte oder fehlerhafte Daten erkennen und bereinigen
- **Klare Definitionen:** Einheitliche Begriffe und Standards für Daten festlegen
- **Aktualität gewährleisten:** Daten regelmäßig aktualisieren
- **Verantwortlichkeiten klären:** Data Owner für Pflege und Qualität bestimmen
- **Daten dokumentieren:** Prozesse, Quellen und Transformationen transparent festhalten
- **Struktur & Organisation:** Daten logisch aufbereiten, für einfache Analyse und Reporting
- **Zielgerichtete Analyse:** Vorher definieren, welche Fragen die Daten beantworten sollen
- **Visualisierung & Interpretation:** Daten verständlich aufbereiten, Entscheidungen unterstützen
- **Team-Schulungen:** Kontinuierliche Weiterbildung zu Tools, Datenschutz, etc.
- **Regelmäßige Audits:** Überprüfung von Datenqualität, Prozessen und Compliance-Standards

HERAUSFORDERUNGEN IM DATENPROZESS-MANAGEMENT

- **Wachsendes Datenvolumen:** Speicher- und Verarbeitungslösungen müssen skalierbar sein
- **Silodenken**: Fehlende bereichsübergreifende Kommunikation behindert Integration
- **Datenqualität & Standardisierung:** Fehlerhafte, unvollständige oder uneinheitliche Daten erschweren Analysen
- **Manuelle Prozesse:** Zeitaufwendig, fehleranfällig und schwer reproduzierbar
- **Verantwortlichkeiten & Governance:** Unklare Rollen und fehlende Datenpflege
- **Dokumentation & Nachvollziehbarkeit:** Fehlende Transparenz bei Datenherkunft und ETL-Schritten

VON DER THEORIE ZUR PRAXIS MIT KNIME

KEY-TAKEAWAYS

ETL ist das Rückgrat
Extract, Transform, Load schafft saubere Daten

Data Analytics treibt Entscheidungen
Von Deskriptiv bis Präskriptiv

Analyse + Dokumentation = Nachhaltigkeit
Strukturierte Prozesse = skalierbar

Die Theorie haben wir verstanden. Jetzt bauen wir echte ETL-Workflows

KNIME UND ETL-WORKFLOWS

- Was ist KNIME?
- Aufbau von KNIME
- Grundlegende Funktionalität
- Einfache Visualisierungen

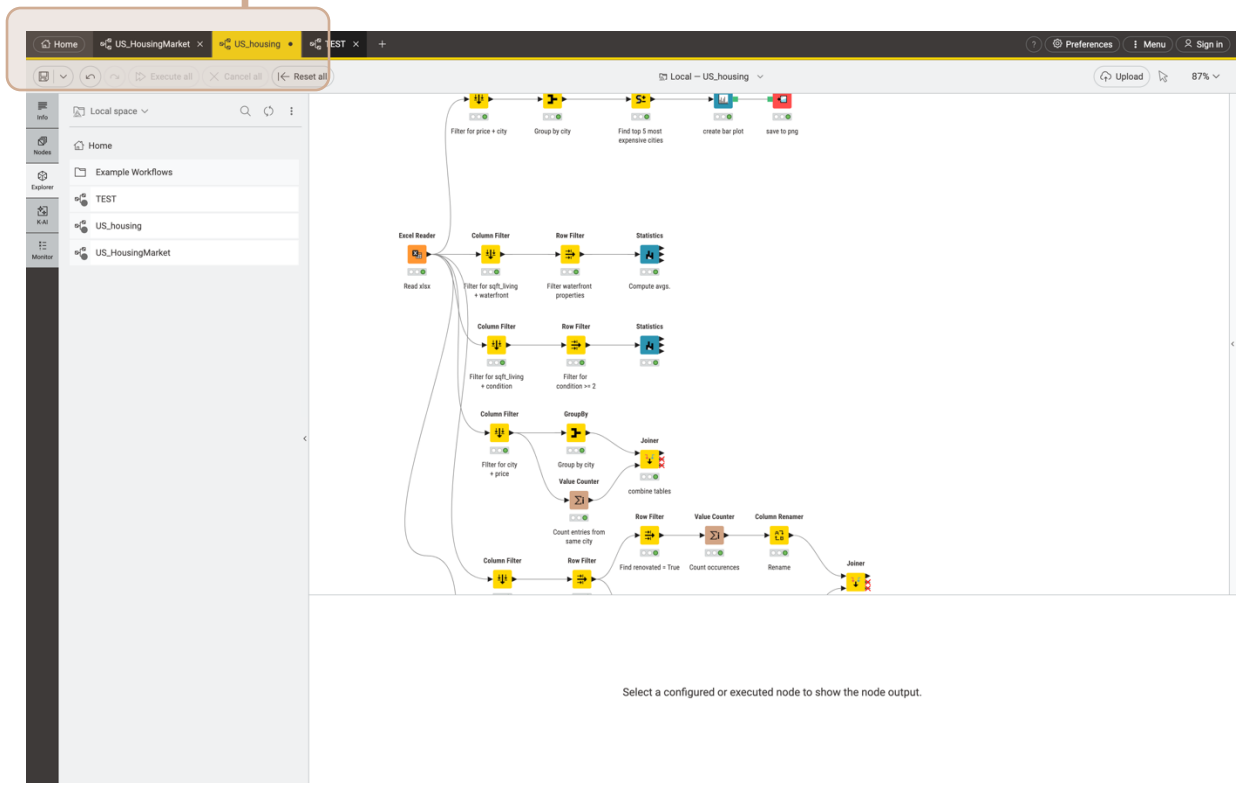
WAS IST KNIME?

- KNIME ist eine Open Source Analytics Plattform
- Es erlaubt die Datenanalyse und Datenaufbereitung aus verschiedenen Quellen
- ETL-Workflows können visuell per Drag-And-Drop erstellt werden
  - Programmieren ist also optional!
- KNIME ist erweiterbar für Machine Learning, Text Mining und Big Data
- KNIME ist plattformübergreifend: Identische Oberfläche auf Windows, Mac und Linux
- Es gibt zusätzliche Enterprise-Funktionen (optional)

AUFBAU VON KNIME

Anwendungstabs

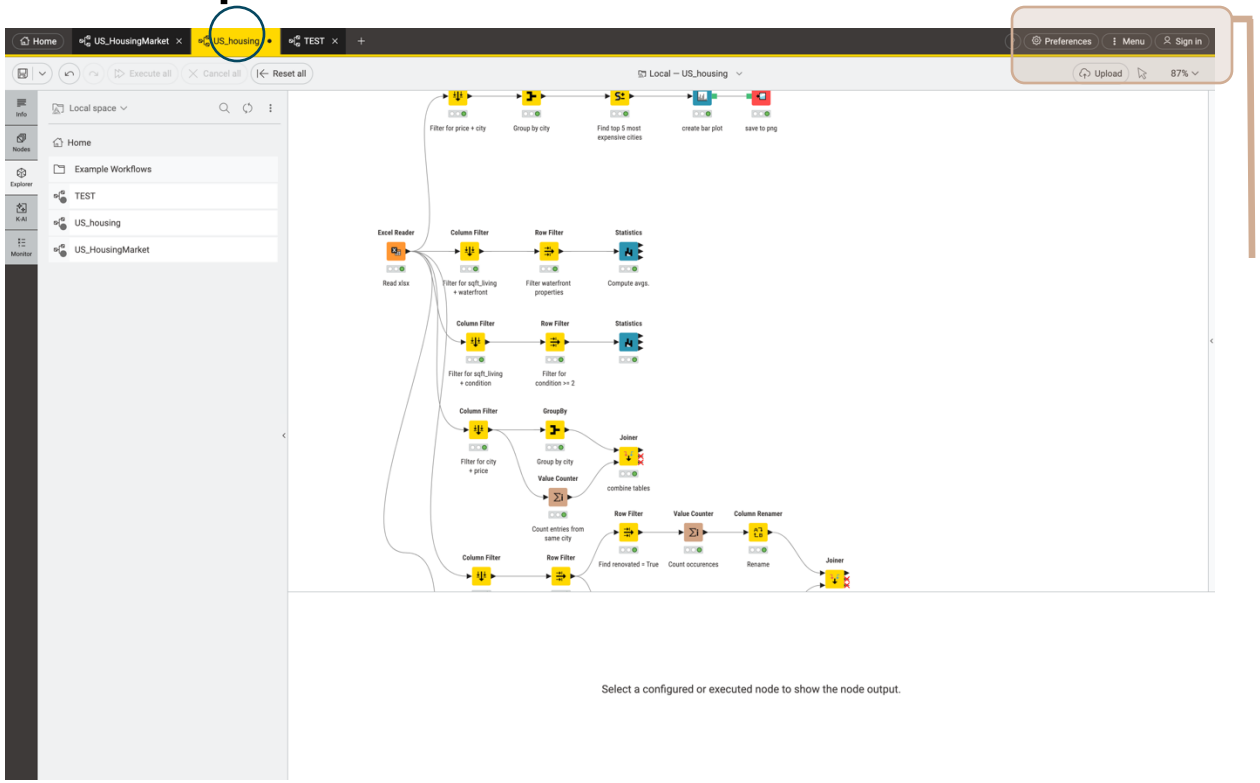
Zeigt den Startbildschirm und alle offenen Workflows



AUFBAU VON KNIME

Anwendungstabs

Zeigt den Startbildschirm und alle offenen Workflows

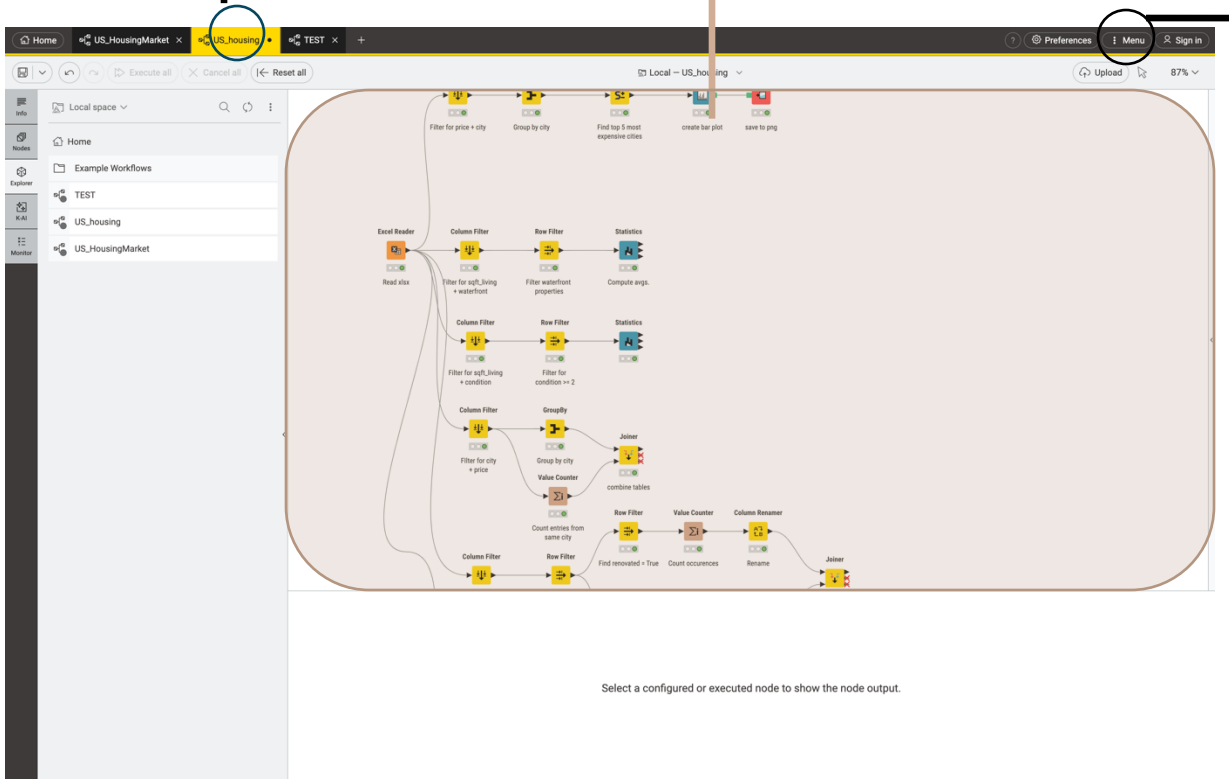


Menü für Hilfe, Einstellungen

Zugriff auf Zusatzmaterialien, Erweiterungen und Einstellungen für das Knotenverzeichnis

AUFBAU VON KNIME

Anwendungstabs
Zeigt den Startbildschirm und alle offenen Workflows



Workflow Editor
In diesem Bereich werden die Workflows erstellt und der ausgewählte Workflow angezeigt

Menü für Hilfe, Einstellungen
Zugriff auf Zusatzmaterialien, Erweiterungen und Einstellungen für das Knotenverzeichnis

AUFBAU VON KNIME

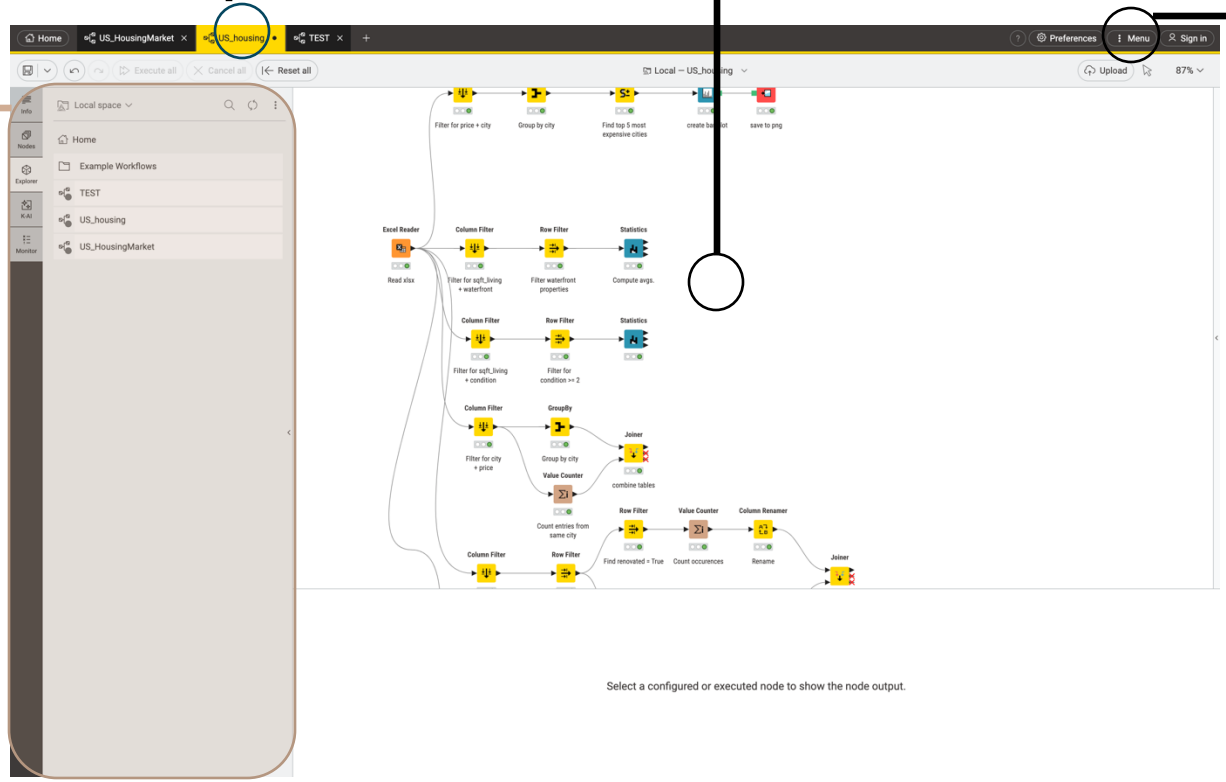
Anwendungstabs
Zeigt den Startbildschirm und alle offenen Workflows

Seitenpanel zur Navigation
Beschreibung: Zeigt eine Beschreibung des aktuellen Workflows

Knoten-Repository: Zeigt alle verfügbaren Knoten die für Workflows genutzt werden können

Space Explorer: Zur Navigation auf dem Computer oder in KNIME Hub Spaces, um zu eigenen Workflows, Komponenten und Dateien zu navigieren

Workflow Monitor: Hilft beim Debuggen von Workflows indem Warnungen / Fehler aller verwendeten Knoten angezeigt werden



Workflow Editor
In diesem Bereich werden die Workflows erstellt und der ausgewählte Workflow angezeigt

Menü für Hilfe, Einstellungen
Zugriff auf Zusatzmaterialien, Erweiterungen und Einstellungen für das Knotenverzeichnis

AUFBAU VON KNIME

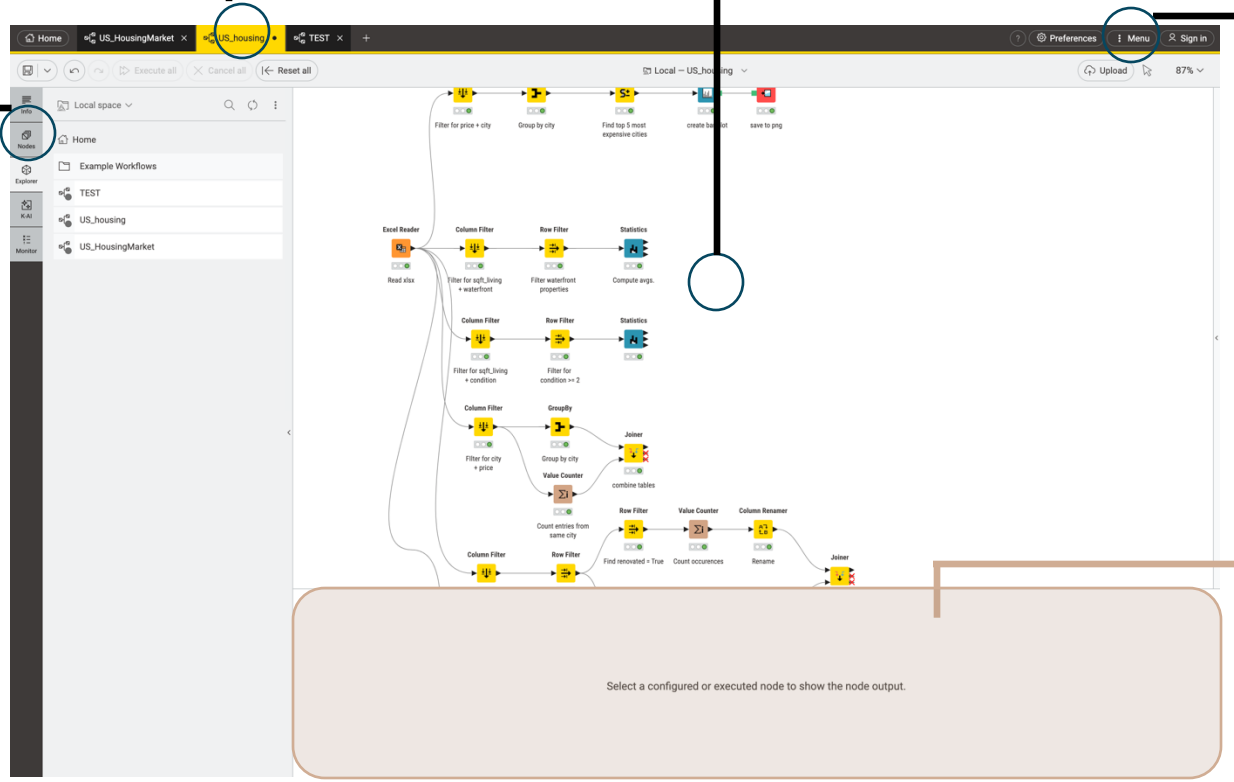
Anwendungstabs
Zeigt den Startbildschirm und alle offenen Workflows

Seitenpanel zur Navigation
Beschreibung: Zeigt eine Beschreibung des aktuellen Workflows

Knoten-Repository: Zeigt alle verfügbaren Knoten die für Workflows genutzt werden können

Space Explorer: Zur Navigation auf dem Computer oder in KNIME Hub Spaces, um zu eigenen Workflows, Komponenten und Dateien zu navigieren

Workflow Monitor: Hilft beim Debuggen von Workflows indem Warnungen / Fehler aller verwendeten Knoten angezeigt werden



Workflow Editor
In diesem Bereich werden die Workflows erstellt und der ausgewählte Workflow angezeigt

Menü für Hilfe, Einstellungen
Zugriff auf Zusatzmaterialien, Erweiterungen und Einstellungen für das Knotenverzeichnis

Knotenmonitor
Zeigt den Output eines ausgewählten Knotens und den Wert von Flussvariablen

AUFBAU VON KNIME

- Workflows bestehen aus Knoten (Node), die einzelne Schritte darstellen
- Jeder Knoten hat eine definierte Funktion: Daten einlesen, transformieren, analysieren und ausgeben (auch grafisch)
- Knoten sind per Drag-And-Drop verbunden, um den Datenfluss zu steuern
- Daten fließen von Knoten zu Knoten, Ergebnisse werden schrittweise erzeugt
- KNIME ist dadurch einfach erweiterbar durch Knoten für Machine Learning, Text Mining oder Big Data
- Fehlerbehandlung und Datenvorschau ist in jedem Knoten möglich!

KNOTEN IN KNIME

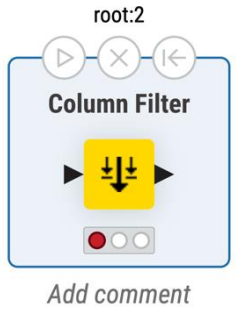
- KNIME bietet eine riesige Vielfalt an vorgefertigten Knoten, welche man in einen ETL-Workflow kombinieren kann
- Zusätzlich ist KNIME erweiterbar
- Im Folgenden schauen wir uns die wichtigsten Standardknoten für ETL-Prozesse an
- Datenbankanbindungen lernen wir in Modul 3 kennen

HINWEIS BEVOR ES LOS GEHT

- **„Den richtigen“ ETL-Workflow gibt es nicht!**
- Viele Wege führen zum Ziel – die Reihenfolge der Abarbeitung ist manchmal relevant, oft aber auch nicht!
  - **Solange die Logik stimmt**

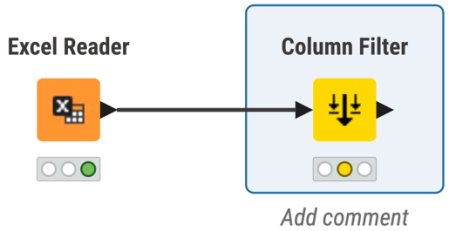
GENERELLE SCHRITTE ZUM ARBEITEN MIT KNOTEN

1. Knoten per Drag und Drop zum Workflow hinzufügen



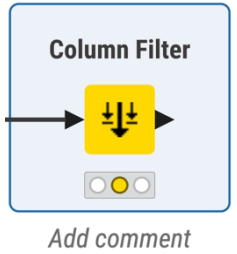
KNIME nutzt für die Knoten ein Ampelsystem. Anfangs rot, weil nicht konfiguriert!

2. Ggf. Knoten mit anderem Knoten verbinden



Daten-Ein und -Ausgang sind am Knoten markiert. Der Ausgang eines bestehenden Knotens kann mit dem Eingang des nächsten Knotens verbunden werden (Maus gedrückt halten). Das symbolisiert den Datenfluss

3. Ggf. noch (nach-)konfigurieren, je nach Knotenart über einfachen Klick oder Rechtsklick auf Knoten → Konfigurieren

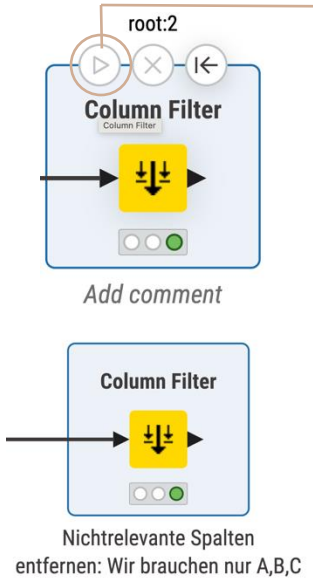


Im Pop-Up Dialog kann der Knoten noch (nach-)konfiguriert werden. Nach dem Konfigurieren Dialog schließen. Nach erfolgreicher Konfiguration: Ampel schaltet auf orange; Knoten ist bereit, wurde aber noch nicht ausgeführt.

GENERELLE SCHRITTE ZUM ARBEITEN MIT KNOTEN

4. Knoten kann ausgeführt werden

5. Dokumentation: Aussagekräftige kurze Annotation zum Knoten hinzufügen

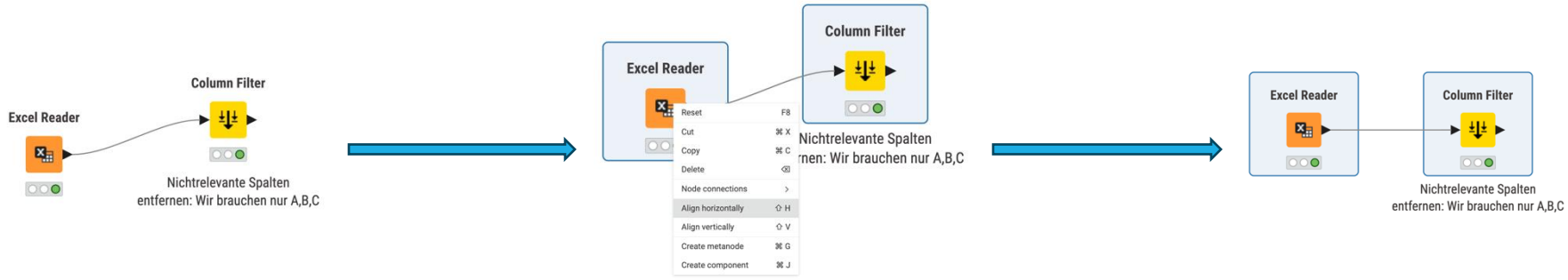


Nach dem ausführen des Knotens schaltet die Ampel (falls alles gut läuft) auf grün. Zusätzlich wird nun im Knotenmonitor eine Übersicht der Daten / produzierten Ergebnisse des Knotens angezeigt.

Ein Kommentar zum Knoten hilft, den ETL-Workflow besser verständlich zu machen. **Dieser Schritt ersetzt keine vollständige Dokumentation!**

WORKFLOW ANSPRECHEND GESTALTEN

- Jeder Knoten sollte kommentiert sein
- Knoten sollten sinnvoll angeordnet sein: Besonders ansprechend sieht es aus, wenn der Datenfluss gerade verläuft (horizontal oder vertikal)
- Das können wir einfach umsetzen: Auszurichtende Knoten markieren, Rechtsklick, „Align horizontally“ oder „Align vertically“



KNOTEN FÜR DIE WICHTIGSTEN GRUNDFUNKTIONEN

1. Excel Reader
2. CSV Reader
3. Missing Value (fehlende Daten behandeln)
4. Column Filter (Spaltenfilter)
5. Row Filter (Zeilenfilter)
6. Math Formula (einfache math. Formeln)
7. String Manipulation (einfache Textbearbeitung)
8. GroupBy (Datenaggregation)
9. Joiner (Zusammenführen von Tabellen)
10. Sorter (Daten Sortieren)
11. Value Counter
12. Statistics
13. Top k row filter (sortieren + trunkieren)
14. Excel Writer
15. CSV Writer
16. Bar Chart (Balkendiagramme)
17. Line Plot / Line Chart (Liniendiagramme)
18. Histogram (Häufigkeitsverteilungen)
19. Box Plot
20. Image Writer (Port)

KNOTEN: EXCEL READER

1

- **Liest xls / xlsx / xlsb und xlsm Dateien**
- Lokal, aus Mountpoint (KNIME Business Server), oder genereller Server
- Eine oder mehrere Dateien gleichzeitig
- Unterstützte Datentypen: string, number, boolean, date und time
- **Lädt nur Daten, keine Grafiken o.ä.**
- Spalten können direkt umbenannt werden
- Datentypen werden automatisch erkannt, können aber manuell geändert werden
- Zeilenzahl kann begrenzt werden

KNOTEN: CSV READER

2

- **Liest csv Dateien**
- Lokal, aus Mountpoint (KNIME Business Server), oder genereller Server
- Unterstützte Datentypen: string, number, boolean, date und time
- Ähnlich zum Excel Reader
- Dateiformat kann automatisch identifiziert oder angegeben werden (Dezimalzeichen, Trennzeichen, …)
- Spalten können direkt umbenannt werden
- Datentypen werden automatisch erkannt, können aber manuell geändert werden
- Zeilenzahl kann begrenzt werden

KNOTEN: MISSING VALUE

3

- **Identifiziert und behandelt fehlende Werte in einer Datentabelle**
- Fehlende Werte werden ersetzt
  - Entweder mit einem fixen Wert
  - Oder mit dem häufigsten Wert aus einer Spalte / dem Mittelwert / dem Maximum / Minimum aus der Spalte, …
- Man definiert die Regeln entweder für jeden Datentyp pauschal, oder für jede Spalte individuell

- Angenommen wir haben Sensordaten aus einer Produktion mit Zeitstempel. Falls ein Sensor temporär keine Werte geliefert hat, so fehlen die Messwerte zu den zugehörigen Zeitstempeln.
  - Der Missing Value Knoten kann genutzt werden, um diese Werte bspw. auf 0 zu setzen

Zentraler Knoten für die Datenbereinigung!

AUFGABE

Dateien:

www.bileso.de/material/IHK/house_prices.xlsx

- Bitte lade dir die Excel-Datei herunter.
- Stelle sicher, dass KNIME auf deinem Computer installiert ist.
- Erstelle einen neuen Workflow. Das Ziel dieses Workflows ist es, die Excel-Datei zu lesen
- Nach dem Ausführen können die Daten im Knotenmonitor angesehen werden

KNOTEN: COLUMN FILTER

4

- **Wird genutzt um Spalten aus gegebenen Daten herauszufiltern**
- Hierzu werden die Spalten in eine „Include“ und „Exclude“ Liste aufgeteilt

- Vorteil: Nach dem Filtern hat man nur Spalten, welche man auch tatsächlich haben will
- Spalten zu filtern ist mehr als ein nice-to-have:
  - **Falls bspw. der nächste Knoten Mittelwerte berechnen soll, so ist das herausfiltern von allen nicht-numerischen Spalten unbedingt notwendig (was ist der Mittelwert einer Textspalte?)**

KNOTEN: ROW FILTER

5

- **Wird genutzt um Zeilen aus gegebenen Daten herauszufiltern**
- Hierzu werden ein oder mehrere Kriterien festgelegt. Nur Zeilen, welche alle definierten Kriterien erfüllen, schaffen es durch den Row Filter

- Angenommen man hat einen Datensatz zu Immobilien aus Deutschland (mit Preisen):
  - Mit dem Row Filter können wir alle Immobilien einer Stadt herausfiltern
  - Oder alle die einen Verkaufspreis von mindestens 500.000€ haben
  - Oder beides (und mehr)

KNOTEN: MATH FORMULA

6

- **Wird genutzt um zeilenweise mathematische Formeln anzuwenden**
- Die Formel (=Expression) kann aus den Einträgen aller Spalten und zusätzlichen Funktionen wie Mittelwertberechnungen, Minima / Maxima etc. bestehen
- Das Ergebnis der Berechnung ist ein Wert pro Zeile, also wieder eine neue Spalte: Diese kann man entweder nutzen um eine bestehende Spalte zu überschreiben, oder man hängt die Spalte als zusätzliche Spalte an die Daten dran.

- Angenommen man hat einen Datensatz zu Immobilien aus Deutschland (mit Preisen):
  - Man kann mit dem Math Formula Knoten einfach herausfinden, welche Immobilien günstiger als der Mittelwert aller Hauspreise sind
  - Wenn die Preise in einer Spalte *price* stehen, dann ist die Formel

    COL_MEAN($price$) > $price$
  - Das zeilenweise Ergebnis: 0 falls ein Haus teurer als der Mittelwert ist, 1 falls es maximal den Mittelwert kostet.

AUFGABE

Dateien:

www.bileso.de/material/IHK/house_prices.xlsx

- Erweitere den Workflow aus der vorherigen Aufgabe:
  1. Filtere die Spalten „*waterfront*“ und „*price*“ aus der Datentabelle heraus
  2. Extrahiere aus dem Ergebnis aus 1. je eine Datentabelle mit allen Immobilien bei denen *waterfront* = 1 und *waterfront* = *0* ist.
  3. Berechne nun für beide erstellten Datentabellen aus 2. je eine neue Spalte, in der die Abweichung von *price* zum zugehörigen Mittelwert steht, d.h.
     1. die Abweichung jeder Uferpromenaden-Immobilie zum Mittelwert aller Preise von Uferpromenaden-Immobilien
     2. Und die Abweichung jeder Nicht-Uferpromenaden-Immobilie zum Mittelwert aller Preise von Nicht-Uferpromenaden-Immobilien.

KNOTEN: STRING MANIPULATION

7

- **Wird genutzt um zeilenweise eine einfache Textbearbeitung durchzuführen**
- Die Bearbeitungsvorschrift (=Expression) kann aus allen Spaltenwerten und Funktionen wie „upperCase“, „capitalize“ etc. bestehen
- Das Ergebnis kann wie bei den Math Formula Knoten entweder eine Spalte ersetzen, oder als neue Spalte hinzugefügt werden

- Angenommen man hat einen Datensatz mit Namen von Personen, die in einem Onlineshop eingekauft haben
  - Die Einträge in der Spalte „name“ sollten also mit Großbuchstaben beginnen („Denis Düsseldorf“ nicht „denis düsseldorf“)
  - Die Expression hierzu lautet
    capitalize($name$)

KNOTEN: STRING MANIPULATION

7

Wir können mit diesem Knoten auch in andere Datenformate konvertieren, was diesen Knoten sehr praktisch macht!

- Um eine Spalte mit dem Namen ‚income' in eine Gleitkommazahl zu konvertieren, können wir diesen Ausdruck hier verwenden:

  toDouble($income$)

Der String Manipulation Knoten ist ein zentraler Ausgangspunkt für das Bereinigen der Daten!

AUFGABE

Dateien:

www.bileso.de/material/IHK/ingame_purchasedata.csv

- Lade dir die CSV-Datei runter. Erstelle einen neuen Workflow in KNIME und lade die Daten des Ingame-Purchase Datensatzes.
  - Berechne anschließend zeilenweise für die gesamte Spalte ‚income' die Abweichung zum Mittelwert und speicher diese Spalte in ‚income_diffToExpectation".

Tipp: 1) Der Ausdruck zur Berechnung lautet
$income$ - COL_MEAN($income$)

2) **Denk' dran:** Diese Berechnung macht nur für numerische Datentypen Sinn (integer, float / double).

3) Achte auf das kleine Warndreieck innerhalb eines Knotens: Um die Informationen hierzu einzusehen, fahre mit der Maus über das Warndreieck (ohne Klick, nur hovern).

KNOTEN: GROUP BY

8

- **Wird genutzt um Datenaggregationen durchzuführen**
- Datensätze (Zeilen) werden basierend auf dem Wert aus einer / mehrerer Spalten in Gruppen eingeteilt
- Spalten, welche für diese Gruppierung nicht wichtig sind, werden gemittelt / gezählt / etc.
  - Das ist festzulegen!

- Angenommen man hat einen Datensatz zu Immobilien aus Deutschland (mit Preisen)
  - Um pro Stadt nun den Mittelwert zu berechnen, würde man mit einem Group By Knoten nach der Stadt gruppieren, und den Mittelwert der Preisspalte berechnen lassen

KNOTEN: JOINER

9

- **Wird genutzt um Datentabellen zu vereinen**
- Bspw. Wenn die Daten aus mehreren Quellen kommen, oder Ergebnisse von Zwischenberechnungen aus einem anderen Arm des ETL-Workflows einfließen sollen
- Hierzu muss festgelegt werden, welche Spalten zum vereinen beider Datentabellen übereinstimmen sollen (dies können mehrere sein)
- Man kann so auch nur Zeilen aus einer Tabelle identifizieren, die kein Pendant in der anderen Tabelle haben
- Angenommen man Sensordaten zweier Maschinen aus unterschiedlichen Dateien / Datenbanken (mit Zeitstempel)
  - Zum gleichen Zeitstempel gibt es also je Maschine einen (oder mehrere) Messwert(e)
  - Nach dem getrennten Einlesen der Daten könnte man mit dem Joiner Knoten über gleiche Zeitstempel die Daten zusammenführen

KNOTEN: SORTER

10

- **Wird genutzt Zeilen einer Datentabelle zu sortieren**
- Aufsteigend oder absteigend
- Kann für mehrere Spalten simultan genutzt werden

- Angenommen hat einen Datensatz mit Immobilien aus Deutschland (inkl. Baujahr).
- Mit dem Sorter Knoten kann man die Immobilien nach dem Baujahr aufsteigend sortieren.

KNOTEN: VALUE COUNTER

11

- **Zählt, wie oft jeder Wert aus einer Spalte in der Spalte vorkommt**

- Angenommen wir haben einen Datensatz mit Immobilien aus Deutschland (inkl. Stadt)
  - Der Value Counter Knoten kann genutzt werden, um für jede Stadt die Anzahl der Immobilien zu zählen.

AUFGABE

Dateien:

www.bileso.de/material/IHK/house_prices.xlsx

Zurück zum Workflow des Immobilienpreis-Datensatzes!

- Erweitere den Workflow zur Analyse der Immobilienpreise aus der vorherigen Aufgabe:
  1. Die Spalten “city“ und „price“ herausfiltern
  2. Den durchschnittlichen Preis aller Immobilien pro Stadt berechnen
  3. Zusätzlich soll die Anzahl der Immobilien pro Stadt gezählt werden
  4. Die Ergebnisse aus 2. und 3. sollen in eine gemeinsame Datentabelle gebracht werden

KNOTEN: STATISTICS

12

- **Berechnet die wichtigsten statistischen Kennzahlen einer Datentabelle**
- Berechnung erfolgt spaltenweise
- Ergebnis enthält
  - Minimum / Maximum
  - Mittelwert
  - Standardabweichung
  - Varianz
  - Anzahl fehlender Werte
  - …
- Dieser Knoten erlaubt es eine schnelle Übersicht über eine Datentabelle zu erhalten
- Das ganze funktioniert für alle Spalten simultan
- Angenommen wir haben einen Datensatz mit Immobilien aus Deutschland (inkl. Preis, Wohnfläche, …)
  - Eine Anwendung des Statistics Knoten zeigt eine detaillierte statistische Übersicht über die Preisspanne, Wohnfläche etc.

KNOTEN: TOP K ROW FILTER

13

- **Sortiert eine Datentabelle nach festgelegten Kriterien und schneidet die Ergebnistabelle zusätzlich nach einer festgelegten Anzahl an Zeilen ab**
- Entweder ist die Anzahl der Zeilen fix, oder man schneidet nach den ersten k unterschiedlichen Werten ab
- Es gibt verschiedene Anwendungsfälle für diesen Knoten
- Insbesondere wird er häufig vor dem Schreiben von Dateien oder dem Plotten von Daten benutzt

- Angenommen wir haben einen Datensatz mit Immobilien aus Deutschland (inkl. Preis und Stadt)
  - Der Top K Row Filter kann genutzt werden, um bspw. die teuersten 5 Immobilien aus dem Datensatz herauszuschneiden

KNOTEN: EXCEL WRITER

14

- **Schreibt eine Datentabelle in eine xls oder xlsx Datei**
- Kann auch genutzt werden um eine Seite zu einer bestehenden Excel Datei hinzuzufügen

- Dieser Knoten ist ein möglicher Endpunkt in einem ETL-Workflow
- Nachdem die Ergebnisse in eine Excel-Datei geschrieben wurden, kann man zur Visualisierung und Erstellung von Dashboards mit BI Tools (bspw. Microsoft PowerBI) übergehen

KNOTEN: CSV WRITER

15

- **Schreibt eine Datentabelle in eine csv Datei**
- Kann auch genutzt werden um Zeilen an eine bestehende csv Datei anzuhängen
- Einstellungen wie das Festlegen eines Trennzeichens können vorgenommen werden
- Dieser Knoten ist ein möglicher Endpunkt in einem ETL-Workflow
- Nachdem die Ergebnisse in eine CSV geschrieben wurden, kann man zur Visualisierung und Erstellung von Dashboards mit BI Tools (bspw. Microsoft PowerBI) übergehen

KNOTEN: BAR CHART

16

- **Erstellt Balkendiagramme zur Darstellung kategorischer Daten**
- Generiert den Plot nur zur Visualisierung innerhalb von KNIME, oder auch als PNG oder SVG Datei (als Image Port Objekt)
- Eine Achse = Kategorien, andere Achse = numerische Werte
- Unterstützt vertikale und horizontale Balken
- Kann vor dem Plotten die Daten auch aggregieren
- Gut geeignet für Vergleiche zwischen Gruppen oder summierte Werte

- **Wenn eine Datei geschrieben werden soll:** Dieser Knoten generiert die grafische Darstellung, schreibt diese jedoch NICHT selbst in eine Datei. Das geschieht im Anschluss mit einem Image Writer Knoten
- Angenommen wir haben einen Datensatz mit Immobilien aus Deutschland (inkl. Stadt und Preis):
  - Der Bar Chart Knoten kann genutzt werden um den durchschnittlichen Preis der Immobilien pro Stadt visuell als Balkendiagramm darzustellen.

KNOTEN: LINE PLOT

17

- **Erstellt Liniendiagramme zur Darstellung numerischer oder zeitlicher Daten**
- Generiert den Plot nur zur Visualisierung innerhalb von KNIME, oder auch als PNG oder SVG Datei (als Image Port Objekt)
- Farben / Marker / Linienarten können ausgewählt werden
- Kann auch mehrere Liniendiagramme gleichzeitig generieren
- Gut geeignet für Zeitreihenanalysen oder den Vergleich von Verläufen

- **Wenn eine Datei geschrieben werden soll:** Dieser Knoten generiert die grafische Darstellung, schreibt diese jedoch NICHT selbst in eine Datei. Das geschieht im Anschluss mit einem Image Writer Knoten
- Angenommen wir haben einen Datensatz mit numerischen Sensordaten über die Zeit:
  - Der Line Plot Knoten kann genutzt werden um die numerischen Sensordaten im Zeitverlauf darzustellen

KNOTEN: HISTOGRAM

18

- **Erstellt Histogramme zur Darstellung der Verteilung von Werten aus einer Spalte**
- Generiert den Plot nur zur Visualisierung innerhalb von KNIME, oder auch als PNG oder SVG Datei (als Image Port Objekt)
- Die Daten werden anhand der Werte der ausgewählten Spalte in eine festgelegte Anzahl von Gruppen eingeteilt. Die Anzahl der Datensätze pro Gruppe bestimmt dann wie hoch der jeweilige Balken im Histogramm ist.
- Diese Visualisierung hilft dabei zu verstehen, wie die Daten verteilt sind (Ausreißer)

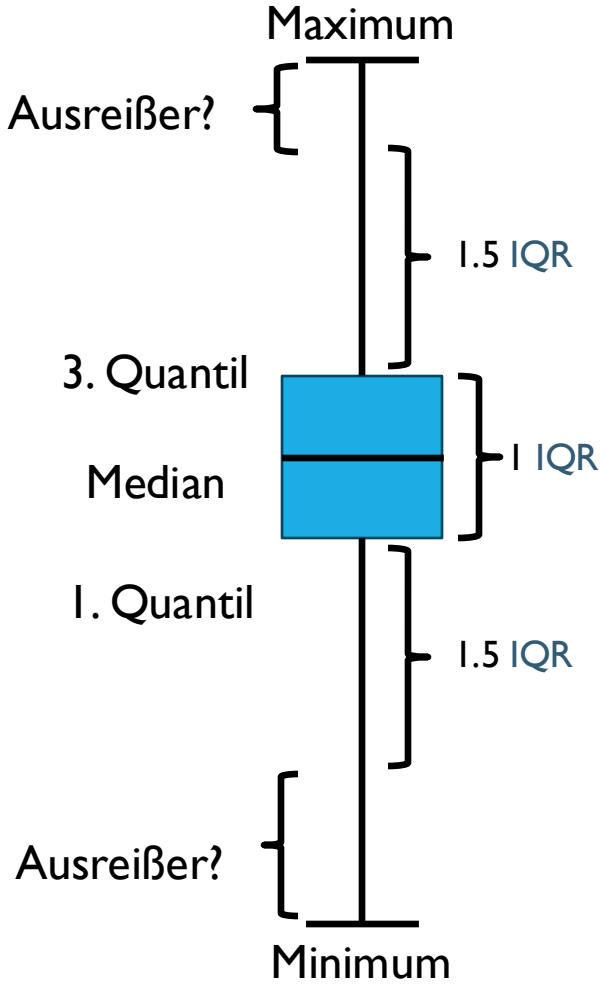
- **Wenn eine Datei geschrieben werden soll:** Dieser Knoten generiert die grafische Darstellung, schreibt diese jedoch NICHT selbst in eine Datei. Das geschieht im Anschluss mit einem Image Writer Knoten
- Angenommen wir haben einen Datensatz mit numerischen Sensordaten über die Zeit:
  - Der Histogram Knoten kann genutzt werden, um die Messwerte in 3 Kategorien einzuteilen (unterstes Drittel der Werte, mittleres Drittel, oberstes Drittel) und die Verteilung der Datensätze anschließend visuell als Histogramm darzustellen.

KNOTEN: BOX PLOT

19

- **Erstellt spaltenweise Boxplots zur Darstellung der Verteilung von Werten**
- Generiert den Plot nur zur Visualisierung innerhalb von KNIME, oder auch als PNG oder SVG Datei (als Image Port Objekt)
- Die Daten werden anhand der Werte der ausgewählten Spalte in eine festgelegte Anzahl von Gruppen eingeteilt. Die Anzahl der Datensätze pro Gruppe bestimmt dann wie hoch der jeweilige Balken im Histogramm ist.
- Diese Visualisierung hilft dabei zu verstehen, wie die Daten verteilt sind (Ausreißer)

- **Wenn eine Datei geschrieben werden soll:** Dieser Knoten generiert die grafische Darstellung, schreibt diese jedoch NICHT selbst in eine Datei. Das geschieht im Anschluss mit einem Image Writer Knoten
- Angenommen wir haben einen Datensatz mit numerischen Sensordaten über die Zeit:
  - Der Box Plot Knoten kann bei der Identifikation von Ausreißern, bspw. Messfehler, helfen.



IQR = Inter quartil range

KNOTEN: IMAGE WRITER (PORT)

20

- **Erstellt PNG / SVG Dateien aus Image Port Objekten**
- Image Port Objekte sind der Output von Knoten wie dem Bar Chart / Line Plot / Histogram, falls diese konfiguriert worden den Inhalt einer Bilddatei (PNG / SVG) zu generieren
- Der Image Writer bereitet die Daten also nicht selbst visuell auf, sondern schreibt nur den Output eines vorherigen Knotens in eine Bilddatei!

- Dieser Knoten ist ein möglicher Endpunkt in einem ETL-Workflow
- Angenommen wir haben ein Histogram erstellt. Der Image Writer kann den grafischen Output nun in eine Datei schreiben.

AUFGABE

Dateien:

www.bileso.de/material/IHK/house_prices.xlsx

- Erweitere den Immobilienpreis-Workflow aus der vorherigen Aufgabe:
  1. Erstelle ein Balkendiagramm **ALS PNG-DATEI**, dass folgende Infos zeigt:
     - Auf der x-Achse die Namen der 5 Städte mit den höchsten Durchschnittspreisen aller Immobilien mit einer „condition“ > 2
     - Auf der y-Achse die zugehörigen Durchschnittspreise Preise
  2. Berechne den Anteil der Renovierten Häuser für die Stadt „Bellevue“

AUFGABE

Dateien:

www.bileso.de/material/IHK/house_prices.xlsx

- Zusätzlich zu den bisherigen Betrachtungen: Welche weiteren Kennzahlen / Grafiken könnten interessant sein? **Beschränke dich hierbei auf die Knoten, die wir bereits kennengelernt haben!**

EXPLORATIVE DATENANALYSE (EDA)

EXPLORATIVE DATENANALYSE

Ziel: Muster, Anomalien & Hypothesen entdecken – vor dem Modellieren!

- Man möchte Daten verstehen bevor man sie erklärt
- Datenstruktur, Verteilung, Zusammenhänge & Probleme erkennen
- Offene (visuelle) Erkundung, keine Hypothesen! „Was sagen die Daten wirklich?“
- Vermeidung falscher Schlüsse durch unbekannte Datenprobleme
- Entdeckung unerwarteter Muster
- Grundlagen für Feature Engineering und Modellwahl

EDA - UNIVARIANTE ANALYSE

Ziel: Verstehen, wie einzelne Variablen verteilt sind

- Für numerische Daten:
  - Bspw. durch Histogramme
  - Boxplots helfen bei der Analyse von Ausreißern
- Für kategorische Daten
  - Balkendiagramme
  - Häufigkeitstabellen

EDA – BIVARIATE / MULTIVARIATE ANALYSE

Ziel: Zusammenhänge zwischen Variablen

- Für numerische Daten:
  - Scatter plot
  - Korrelationsmatrix
- Für kategorische Daten:
  - Gruppenweise Boxplots
  - Gestapelte Balkendiagramme

TYPISCHE ERKENNTNISSE DER EDA

- Starke Korrelation zwischen einem / mehreren Features und der Zielvariable
- Nicht-lineare Zusammenhänge erfordern ggf. Transformation
- Ausreißer: Korrekte (Mess-)Werte, oder sind das Fehler?
- Fehlende Werte in bestimmten Regionen: Systematisch?

EDA – BEST PRACTICES

- **Kurz: Immer visualisieren**
- Alles dokumentieren!
- Fragen stellen: „Warum ist das so?“
- Keine Voreiligen Schlüsse! Die EDA dient dazu, Hypothesen zu generieren!

EDA - ZUSAMMENFASSUNG

- **EDA ist das Fundament einer jeden Datenanalyse, und damit auch der erste Schritt in einem ETL-Workflow**
- Durch eine Kombination aus Statistik, Visualisierung und Neugier werden **Fragen** generiert, noch **keine Antworten**!
- Im nächsten Schritt können die Hypothesen dann durch vertiefende Analysen und Modellierungen überprüft / widerlegt werden!

AUFGABE*

Dateien:

www.bileso.de/material/IHK/netflix_titles.csv

Du bist Data Analyst bei Netflix. Deine Aufgabe ist es, den gegebenen Datensatz zu analysieren um länderspezifische Content-Trends, Vorlieben und Konsummuster zu erkennen. Ultimativ soll das dabei helfen, marktübergreifende Handlungsempfehlungen für höheres Engagement und mehr Views zu produzieren.

Generier hierfür einige erste Insights.

VERFÜGBARMACHEN VON WORKFLOWS

- KNIME Worflows werden im knwf (KNIME Workflow) / knwa (KNIME Archive) Format gespeichert
  - Jeder der KNIME installiert hat, kann diese Dateien dann öffnen
- KNIME Hub: Upload, veröffentlicht / privat
- Git / SVN: Versionskontrolle einer .knwf Datei
- KNIME Server: Nur bei Enterprise Lizenzen, inkl. Zugriffssteuerung
- WebPortal: Bereitstellung auf einem Server, interaktive Ausführung im Browser
- …

KNWF UND KNAR DATEIEN

Endung	Inhalt	Größe	Verwendung	Öffnen in KNIME	Mit hochgeladenen Dateien?	Flow-Variablen mit Wert?	Komprimiert?
.knwf	Nur Workflow Struktur: Knoten, Einstellungen, Namen von Flow-Variablen	Klein (oft < 1 MB)	Teilen der Logik, Versionskontrolle (Git), Nur Struktur	File → Import → KNIME Workflow	Nein, nur Pfad-Referenzen	Nur Name und Typ, keine Werte	Nein (reines JSON)
.knar	Kompletter Workflow inkl. Daten	Variable, aber größer (MB bis GB)	Reproduzierbare Analysen, Teilen mit Ergebnissen, Backup mit Daten, Tutorials	File → Import → KNIME Workflow (identisch zu knwf)	Ja – enthält lokale Dateien (z.B. CSV, Excel) aus File Reader	Ja – Flow Variablen sind mit aktuellen Werten enthalten	Ja (ZIP-basiert)

PORTS - ARTEN DES DATENTRANSFERS IN KNIME

ARTEN DES DATENTRANSFERS IN KNIME

- Bisher haben wir in KNIME Datentabellen von Knoten zu Knoten weitergereicht
- Das ist nicht alles: Neben Datentabellen können auch andere Objekte (bzw. Datenflüsse) verarbeitet werden
- Die Ein- und Ausgänge eines Knotens nennt man „Ports“
- Es gibt verschiedene Arten von Ports
  - **Ein Beispiel:** Der Image Writer Knoten erhält die grafische Aufbereitung als Input für seinen **Image Port**

- Welche Arten von Ports sind für uns am wichtigsten?

PORTS (ZUR DATENANALYSE)

- **File Reader / Writer Ports** sind für die Kommunikation mit Dateien
- **Database Ports** handhaben Datenbankverbindungen
- **Data Ports** sind die Standard-Ports, welche für tabellarische Daten genutzt werden
- **Flow Variable Ports** erlauben die Übergabe von Variablen (Automatisierung)
- **Image Ports** leiten Bilddaten weiter
- **Model Ports** verarbeiten trainierte Modelle, bspw. Regressionsmodelle
- …

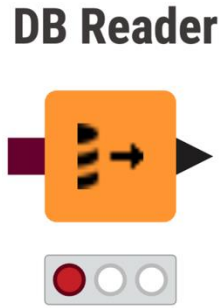
FILE READER / WRITER PORTS

- Erlauben die Kommunikation mit dem Dateisystem
- Werden genutzt um Daten zu lesen / schreiben
  - Sie sind also am Anfang bzw. Ende eines ETL-Prozesses

DATABASE PORTS

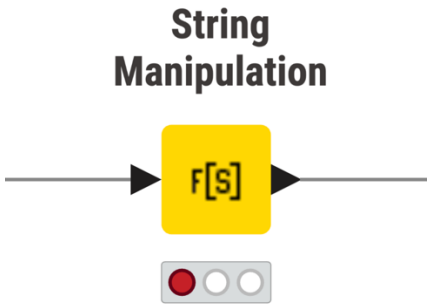
- Zur Verbindung mit Datenbanken

- Zum Empfangen / Senden von Datenbank-Daten



DATA PORTS

- Zur Weiterreichung tabellarischer Daten
- Standard Port-Typ in KNIME
- Haben wir bereits im Detail betrachtet!



FLOW VARIABLE PORTS

- Flussvariablen können an vielen Knoten definiert werden und auch an die folgenden Knoten (zusätzlich zu den regulären Daten) weitergereicht werden
- Flussvariablen sind wichtig für die Automatisierung von ETL-Workflows
- Schauen wir uns noch genau an
- **Diese Ports sind standardmäßig ausgeblendet und werden sichtbar, wenn man mit der Maus über die obere linke Ecke (Eingang), bzw. die obere rechte Ecke (Ausgang) von Knoten fährt**

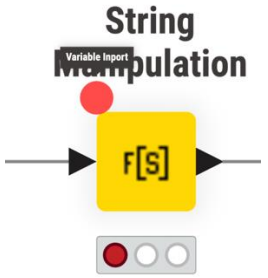
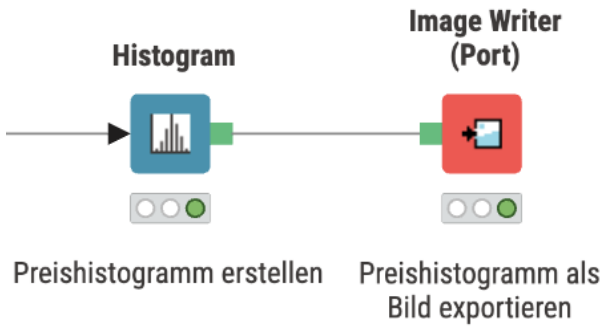


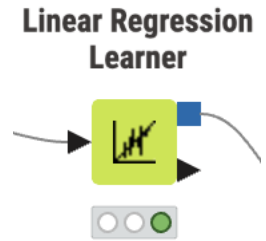
IMAGE PORTS

- Leiten Bilddaten weiter
- Ein Histogram Knoten kann bspw. die Ausgabe in eine Datei vorbereiten
  - Das bedeutet, dass er einen Image Port als Ausgang hat, welcher dann mit einem Image Writer Knoten verbunden werden kann
  - Der Image Writer akzeptiert an seinem Image Port (Eingang) die Bilddaten und schreibt diese dann in eine Datei



MODEL PORT

- Machine Learning Modelle können ebenfalls als Knoten genutzt werden
- Sie akzeptieren tabellarische Daten, auf denen das Modell dann trainiert wird
- Das trainierte Modell kann dann neben den produzierten Daten ebenfalls weitergeleitet werden
- Model Ports finden sich wieder in der oberen linken Ecke (Eingang) und oberen rechten Ecke (Ausgang) der jeweiligen Knoten
- Wir sehen diese Ports noch detailliert in Modul 3



AUFGABE

Dateien:

www.bileso.de/material/IHK/predictive_maintenance.csv

Der Datensatz zeigt Maschinendaten und Informationen zu Ausfällen aus einer Produktion. Dieser Datensatz könnte klassischerweise dafür genutzt werden, um mit Maschinellem Lernen Vorhersagen zu Ausfällen zu treffen.

Ohne Maschinelles Lernen: Was könnte man bereits jetzt untersuchen? Gibt es irgendwelche Muster? Du kannst den Binner Knoten verwenden, um Ergebnisse in verschiedene numerische Bereiche zu gruppieren.