

Online-Zertifikatslehrgang

Data Analyst IHK

Die neue Generation digitaler
IHK-Weiterbildungen

Modul 1 – GRUNDLAGEN DER DATA ANALYTICS – DER ETL-PROZESS

- **Data Analytics – Definition und Inhalt**
 - Was ist Data Analytics und womit beschäftigt es sich
 - **Daten**
 - Datenformen und -quellen
 - **Der Datenprozess**
 - Strukturiert von der Datenquelle zur Anwendung
 - **Datenimport**
 - Systematisches Extrahieren von Daten
 - **Datenqualität**
 - Merkmale der Datenqualität und Methoden zu ihrer Verbesserung
-
- **Die explorative Datenanalyse**
 - Daten zusammenfassen und verstehen
 - Ausreißer und Fehlende Werte entdecken und handhaben
 - **Daten bearbeiten und transformieren**
 - Typenkonvertierung, Wertetransformation
 - Tabellentransformation
 - Aggregationen
-
- **Datensicherheit und Datenschutz**
 - Inhalte, Abgrenzungen, wichtige Elemente
 - **Datenexport**
 - Daten sichern und Anwendungen zur Verfügung stellen
 - **Dokumentation und Organisation von Workflows**
 - Methoden zur Dokumentation und Workflowstrukturierung

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower-left foreground, with fingers slightly curled as if about to touch the globe. The overall aesthetic is high-tech and digital.

Data Analytics

Was ist Data Analytics?

Definition:

Data Analytics ist eine Methode, mit der Daten unter wirtschaftlicher oder wissenschaftlicher Zielsetzung untersucht werden, um Schlussfolgerungen zu ziehen, Zusammenhänge zu visualisieren und Handlungsempfehlungen auszusprechen.

Data Analytics in der Gegenwart

Innovative Technologien wie maschinelles Lernen und Deep Learning eröffnen durch die Nutzung von Algorithmen ganz neue Möglichkeiten.

Statt der Programmierung einzelner Entscheidungs- und Auswertungskriterien werden Anwendungen dank **Künstlicher Intelligenz** in die Lage versetzt:

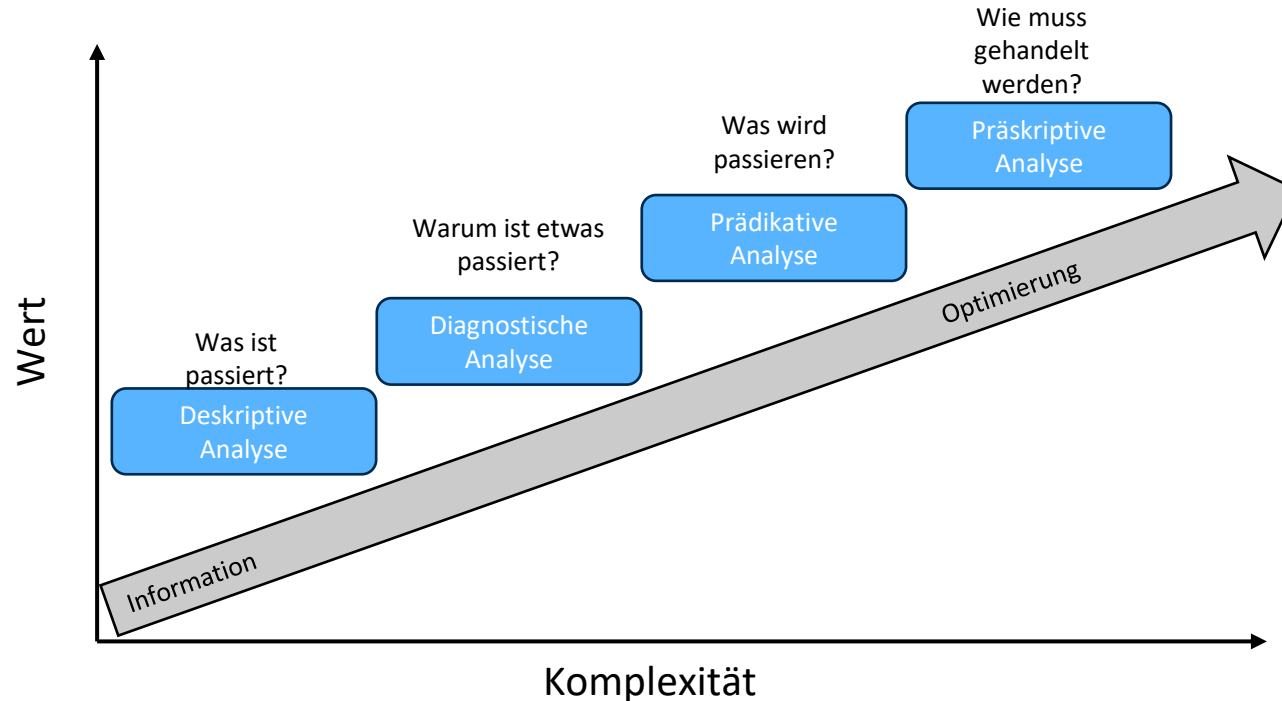
- selbstständig aus Daten zu lernen
- sich zu verbessern
- Entscheidungen zu treffen



**Maschinelles
Lernen**

Mit welchen 4 Zielen und
Fragestellungen beschäftigt sich die
Data Analytics?

Ziele und Fragestellungen der Data Analytics



Deskriptive Analyse: Was ist passiert?

Die **deskriptive Analyse** befasst sich mit der Betrachtung der Vergangenheit.

Anhand überschaubarer Tabellen, grafischer Darstellungen und zusammengeführter Kennzahlen (z. B. Durchschnitt, Toleranz etc.) beschreibt diese Analysen die Zusammenhänge der Daten und bislang unbekannte Strukturen und Informationen.

Diagnostische Analyse: Warum ist es passiert?

Die **diagnostische Analyse** ermöglicht es, Ursachen und Wechselwirkungen aufzudecken und zu erklären.

Diese Analyse ermöglicht einen tiefgehenden Einblick in bestimmte Probleme.

Achtung: Um Muster aufzudecken und Beziehungen der Daten zueinander analysieren zu können, bedarf es einer ausreichend detaillierten Datenbasis.

Prädikative Analyse: Was wird passieren?

Gegenstand der **prädikativen Analyse** ist der Blick in die Zukunft.

Sie nutzt die Erkenntnisse aus der deskriptiven und diagnostischen Analyse, um Abweichungen von Standardwerten vorzeitig zu erkennen und zukünftige Trends möglichst genau vorherzusagen.

Solche Prognosen erfordern mitunter den Einsatz hoch entwickelter Algorithmen und intelligenter Modelle.

Dennoch bleiben es „nur“ Schätzungen auf Basis statistischer Auswertung vergangener Daten. Die Genauigkeit der Modelle hängt immer von der Qualität der eingesetzten Daten ab.

Präskriptive Analyse: Welches Handeln ist erforderlich?

Die **präskriptive Analyse** erweitert die prädikative Analyse um ein Handlungselement und erfordert sowohl die Auswertung von historischen und gegenwärtigen Daten als auch die Integration von vorläufigen Analysen und Prognosen.

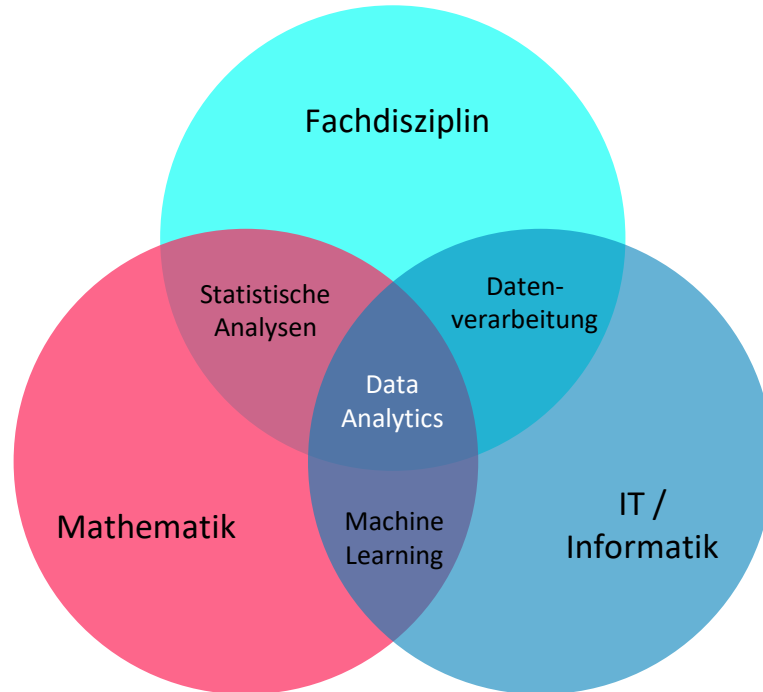
Gute Datenqualität und Flexibilität in der Modellentwicklung sind unerlässliche Elemente.

Es ist die wohl komplexeste Form der Datenanalyse und macht den Einsatz verschiedener Technologien (z. B. Simulationsmodelle, maschinelles Lernen und den Einsatz neuronaler Netze) erforderlich. Mehrwert und Aufwand dieser Methode sollten daher genau gegeneinander abgewogen werden.

1. Welche Disziplinen verbinden sich in der Data Analytics?

2. Welche Anforderungen erwarten einen Data Analyst und welche Fähigkeiten helfen diese zu erfüllen?

Data Analytics ist eine Schnittstelle dreier Disziplinen



Fachdisziplin

- Thematische Zusammenhänge
- Verständnis der Datenentstehung
- Interpretation der Ergebnisse

Mathematik

- Statistik
- Modelle und Algorithmen

IT / Informatik

- Programmierung
- Anwendung

Anforderungen an einen Data Analysten

- Die Verbindung von **Statistik, Fachexpertise und IT** macht es möglich, Themen zu analysieren und daraus Erkenntnisse zu gewinnen sowie Lösungen abzuleiten.
- Durch die Einbeziehung von Computertechnologie und hoch entwickelten Anwendungen ist es zudem möglich, große Datenmengen mit schnellen und effizienten Mitteln zu analysieren.
- Neben dem **fachlichen Know-How** sollte ein Data Analyst über **kommunikative Fähigkeiten** verfügen, um die neu gewonnenen Informationen überzeugend und kreativ in die verschiedenen Ebenen einer Organisation hineintragen zu können.

Fähigkeiten eines Data Analyst

- Verständnis für betriebswirtschaftliche Vorgänge zur Interpretation der Ergebnisse
- Verständnis für datengenerierende Prozesse
- Verständnis der Datenstrukturen, -banken und -modelle
- Kenntnisse zur Verknüpfung verschiedener Datenquellen, die Erstellung komplexer Abfragen und die Beherrschung sehr großer Datenmengen
- Statistische und analytische Fähigkeiten zur Ableitung von Vorhersagen über zukünftige Ereignisse
- Visualisierung von Ergebnissen
- Kommunikation komplexer Sachverhalte und Modelle
- Moderationstechniken
- Projektmanagementtechniken

Wo arbeiten Data Analysts?

- Financial Analyst – oft im Versicherungswesen anzutreffen
- Risk Analyst – häufig im Bankensektor und der Unternehmensberatung tätig
- Data Analyst BI – Expert:in für Unternehmensprozesse, in fast allen Branchen gefragt
- Customer Data Analyst – „Kundenverstehender:in“
- UX Data Analyst – „Userverstehender:in“
- Big Data Analyst – analysiert mit Hilfe von Algorithmen automatisch gigantische Datenmengen
- Clinical Data Analyst – unerlässlich für die Weiterentwicklungen in der E-Health-Branche

Data Analytics im Berufsalltag

Neben der Bearbeitung, Transformation, Analyse, Auswertung und Präsentation von Daten gibt es für den Data Analysten noch weitere Einsatzgebiete:

So coacht er z. B. Kollegen und Vorgesetzte im Umgang mit Daten und Analysetools, wartet und implementiert Datensysteme oder unterstützt bei der Qualitätskontrolle.

Wo Data Analysts zu finden sind:

- **Financial Analyst** – oft im Versicherungswesen anzutreffen
- **Risk Analyst** – häufig im Bankensektor und der Unternehmensberatung tätig
- **Data Analyst BI** – ein Experte für Unternehmensprozesse, der in fast allen Branchen gefragt ist
- **Customer Data Analyst** – der „Kundenversther“
- **UX Data Analyst** – der „Userversther“
- **Big Data Analyst** – analysiert mit Hilfe von Algorithmen automatisch gigantische Datenmengen
- **Clinical Data Analyst** – unerlässlich für die Weiterentwicklungen in der E-Health Branche
- **Weather Analyst** – erstellt Wetterprognosen anhand von Wetterdaten

Daten

Was sind Daten und welche
Bedeutung haben sie?

Was sind Daten?

Menschen und Maschinen produzieren ständig neue Daten. Sie umgeben uns überall und sind Anlass zu Diskussionen über Datensicherheit und Datenschutz. Was aber sind Daten? Der Begriff „Daten“ wird häufig mit Tabellen, Zahlen oder Werten verbunden. Es steckt jedoch mehr dahinter.

Was meistens erwartet wird:

Zahlen & Variablen

	A	D	E	F	G	H
1	row ID	PassagierID	Klasse	Alter	Geschwister	Preis
2	Row0	1	1	60	0	76,2917
3	Row1	4	1	37	1	52,5542
4	Row2	7	1	30	1	57,75
5	Row3	9	1	71	0	49,5042
6	Row4	10	1	48	1	76,7292
7	Row5	14	1	49	0	26
8	Row6	23	1	41	0	134,5
9	Row7	27	1	47	1	227,525
10	Row8	29	1	61	0	32,3208
11	Row9	45	1	58	0	153,4625
12	Row10	47	1	42	0	26,55
13	Row11	49	1	61	1	262,375
14	Row12	53	1	36	0	71
15	Row13	55	1	36	0	75,2417
16	Row14	60	1	47	1	52,5542
17	Row15	65	1	33	1	53,1

Text

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusions and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, and is used in different business, science, and social science domains. In today's business world, data analysis plays a role in making decisions more scientific and helping businesses operate more effectively.

Xia, B. S., & Gong, P. (2015). Review of business intelligence through data analysis. *Benchmarking*, 21(2), 300-311. doi:10.1108/BIJ-08-2012-0050

Bilder

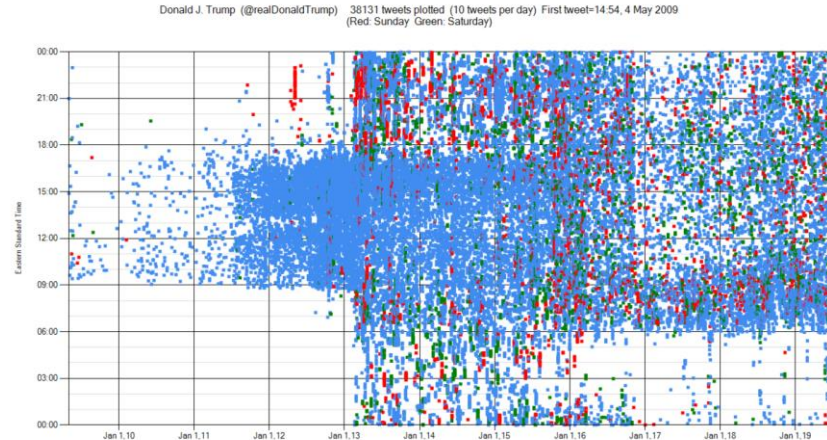


Aber auch das sind Daten:

Die Formen, in denen Daten auftreten, sind so vielseitig wie die Informationen, die aus ihnen gewonnen werden können. So können wir Daten aus Texten ziehen oder auf digitale Daten von z. B. Smart Devices wie einem Fitnesstracker zugreifen.

- Namen, Adressen
- Tweets
- Zeitungsartikel
- Logs (Zeitstempel)
- User Interaktionen
- Standort:
Mobile Geräte, Fahrzeuge
- Sensoren:
Erfassung von Messungen (Ströme, Lautstärke, Beschleunigung etc.)

Tweets Donald J. Trump (2009- 2019)



Bedeutung von Daten für die heutigen Unternehmen

Daten bieten Orientierung und Entscheidungsgrundlagen:

- Informationen über Kunden und ihr Verhalten (Kundenprofile, saisonales Kaufverhalten, etc.)
- Kontrolle und Transparenz in den eigenen Prozessen (Prozesszeiten, Lagerengpässe, Performancewerte, etc.)
- Informationen über Wettbewerber und Partner (Preise, Marktanteile, Lieferverhalten, etc.)
- Trends, Veränderungen, Ereignisse (Welche Wirkungen haben Ereignisse auf bestimmte Zielgruppen? Verändert sich ein Markttrend? ...)

Daten bringen zusätzlichen Nutzen

- Produkte und Dienstleistungen erfahren eine Aufwertung und können vielfältiger oder intensiver genutzt werden (App-Steuerung smarter Produkte, Mobile Buchung, etc.)
- Erweiterung oder Erneuerung des Geschäftsmodells (bessere Erreichbarkeit und Planung auf Datenbasis)
- Absicherung der eigenen Geschäftsprozesse durch kontinuierliche Risiko- und Chancenbewertungen.

1. Was ist der Unterschied zwischen strukturierten und unstrukturierten Daten?

2. In welchen Dimensionen unterscheiden sich Big Data?

Was charakterisiert Daten?

- Es gibt strukturierte und unstrukturierte Daten

Sed ut perspiciatis unde
omnis iste natus error
sit voluptatem
accusantium
doloremque laudantium,
totam rem aperiam,
eaque ipsa



Year Built	Bedrooms	Bathrooms	Sq. Ft.	Price
1901	3	1	1,800	\$200,000
1995	4	3	2,500	\$350,000
1980	2	1	1,300	\$150,000

- Nur strukturierte Daten können analysiert werden!



Year Built	Bedrooms	Bathrooms	Sq. Ft.	Price
1901	3	1	1,800	\$200,000
1995	4	3	2,500	\$350,000
1980	2	1	1,300	\$150,000

Was macht Daten zu Big Data?

Der Begriff „Big Data“ bezeichnet die große Menge an strukturierten und unstrukturierten Daten, die Unternehmen empfangen, weiterleiten oder für nachstehende Prozesse nutzen.

Big Data lässt sich in 5 Dimensionen beschreiben:

Menge
(Volume)

Unternehmen sammeln Daten aus einer Vielzahl von Quellen (z.B. intelligente Geräten (IoT), Produktionsanlagen, Social Media, etc.

Geschwindigkeit
(Velocity)

Moderne Arbeitsprozesse erfordern hohe Geschwindigkeit und zeitnahe Verarbeitung der Datenströme. Sensoren, RFID-Tags, etc. liefern riesige Datenmengen, die nahezu in Echtzeit bewältigt werden müssen.

Vielfalt
(Variety)

Daten werden in vielfältigen Formaten verarbeitet wie z.B. Texte, Videos, Audioquellen, Messprotokolle, etc.

Datenquellen
(Reach)

Nicht nur die Daten selbst werden immer vielfältiger, auch die Datenquellen nehmen zu (Datenbanken, Dateien, Sensoren, Clouds, etc.)

Komplexität
(Variability)

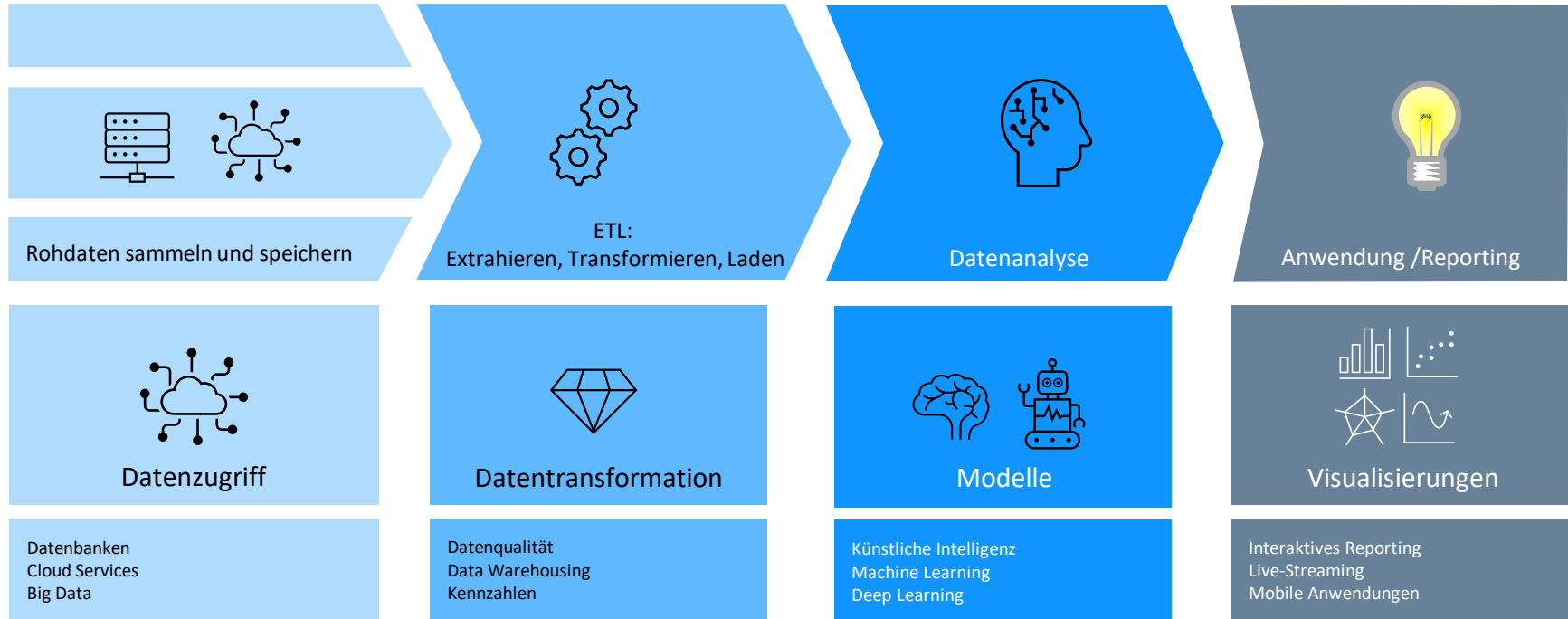
Künstliche Intelligenz, Data Mining, etc. verarbeiten zunehmend mehr Daten parallel in hoher Komplexität.

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower left, with fingers slightly curled as if about to touch the globe. The overall aesthetic is high-tech and digital.

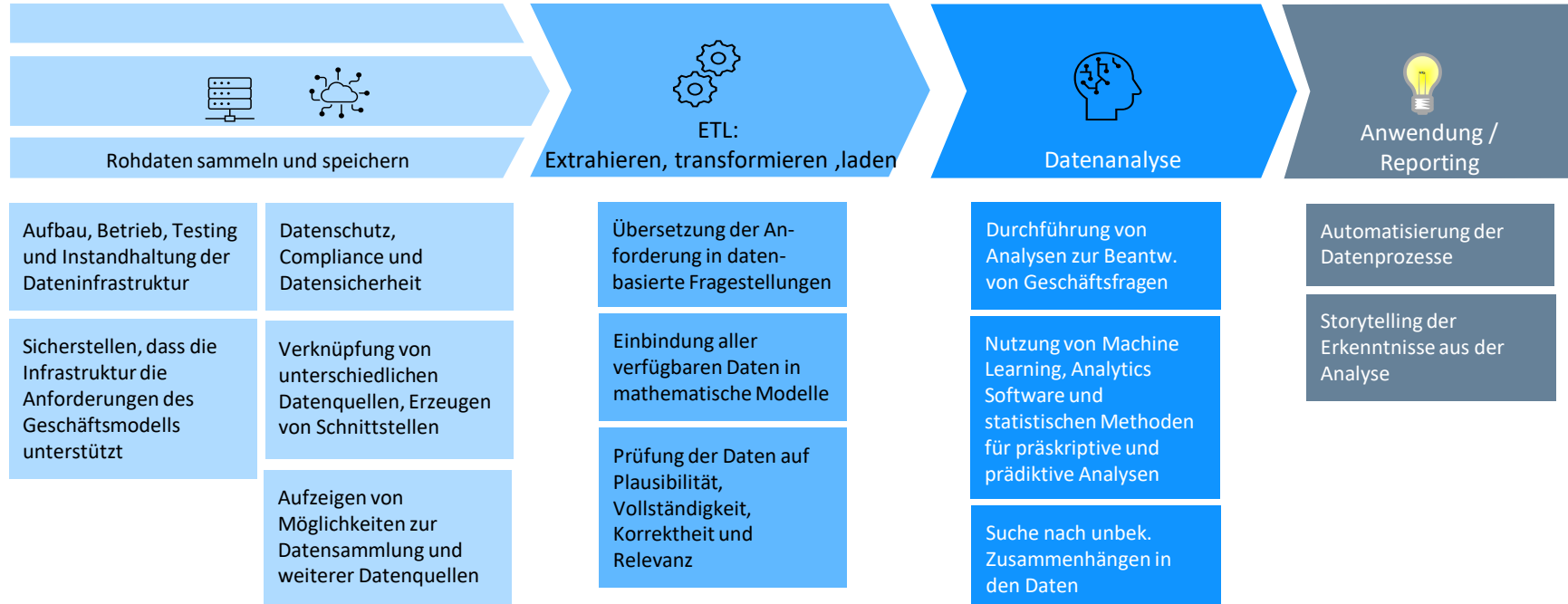
Der Datenprozess

1. In welche Teilschritte lässt sich der Data Analytics Prozess unterteilen?
2. Welche Aufgaben und Merkmale prägen die einzelnen Teilschritte?

Der Data Analytics Prozess – vom Dateninput bis zum Mehrwert



Aufgaben des Data Analyst im Datenprozess



The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower left, with fingers slightly curled as if about to touch the globe.

Datenimport

1. Wie werden Daten in relationalen Systemen strukturiert?

2. In welchen Formaten liegen Daten vor und was kennzeichnet diese Formate?

Relationale Tabellen

Für die Datenanalyse sind relationale Daten (im Zusammenhang stehende Daten) interessant und werden dementsprechend miteinander verknüpft:

Jede Zeile oder Reihe in einer Tabelle ist ein Datensatz. Jede Zeile besteht aus einer Reihe von Attributwerten (Attribute = Eigenschaften), den Spalten der Tabelle.

Die Verbindung wird in der Regel ein Verbindungselement , dem Index, wie die Reihen-ID, oder andere Verbinder (Kunden-ID, Datum, etc.) hergestellt.

Kunden - ID	Nachname	Vorname
001	Anders	Achim
002	Brücke	Beate
003	Chor	Christian
004	Denker	Dagmar
005	Eisen	Ernst
006	Funkel	Frauke

Datenformen

In Tabellen werden in der Regel Zeichen verarbeitet. Das bedeutet, wenn Daten in anderen Formaten wie Bilder, Videos oder Tonaufnahmen vorliegen, müssen diese in einen Zeichensatz umgewandelt werden.

Die gängigen 3 Hauptformate von Daten sind:

1. Ganze Zahlen (Integer, Long): -1; 0; 1; 2; 3; 4; ...
2. Gleitkommazahlen (Flow, Double): 0,34; 5,01; 250,34; ...
3. Zeichen (String, Varchar): AaBb-+*/!“ ...

I	Geschw...	S	Ticketnr	D	Preis
0			11813		76.292
1			11751		52.554
1			13236		57.75
0			PC 17609		49.504
1			PC 17572		76.729
0			19924		26

Je nach Anwendung werden noch weitere Formate als Zellen oder Spalteneigenschaft geführt (hier noch ein paar typische Vertreter):

4. Datum: 09.04.2021
5. Zeit: 09:32:00
6. Prozent: 73%
7. Währung: 257,53 €

Da diese Eigenschaften in der Regel anwendungsspezifisch verwaltet werden, werden sie bei Ex- und Import zwischen verschiedenen Systemen häufig nicht übertragen und müssen im Folgesystem entsprechend angepasst werden.

Sampling

Stichprobenentnahme

- Warum erfassen wir nicht einfach alle Daten?
- Habe ich die Zeit um zu warten bis alle Daten erfasst sind?
- Sind die damit verbundenen Kosten vertretbar?
- Kann ich mit solchen Datenmengen umgehen (Anforderungen an Infrastruktur, Big Data spezifische Lösungen)?

Wichtig: Bewusstsein entwickeln welche Daten sinnvoll sind

Sampling, wenn richtig angewendet, ist eine gute Methode für schnelle Auswertungen mit validen Ergebnissen

Sampling

Stichprobenentnahme

Der **Prozess der Stichprobenauswahl** kann die Daten verzerren. Dadurch sind die Informationen, die wir aus der Stichprobe gewinnen nicht auf die Grundgesamtheit übertragbar.

Beispiel: Verkehrsaufkommen ermitteln

- Zählen der vorbeifahrenden Fahrzeuge über einen definierten Zeitraum und hochrechnen der Anzahl auf den Rest des Jahres.
- Zähle ich einen Tag der Woche? Eine ganze Woche? Einen Monat lang?
- Wochentag oder Wochenende? Feiertage? Saison? Schulferien? Events? Wetter?

Datenimport in KNIME

Weitere Knoten zum Dateninput

Decompress Files



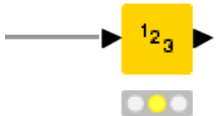
- Dekomprimiert Archive in einen definierten Zielordner(.zip, .rar, etc.)
 - Erstellt eine Liste der Dekomprimierten Ordner und Dateien
 - Funktioniert bisher nur mit englischem Zeichensatz
-

Table Creator



- Erstellt eine manuell definierte Tabelle
 - Spaltennamen und Datentypen können frei gewählt werden
 - Daten werden manuell in die einzelnen Zellen eingetragen
-

Counter Generation



- Erstellt eine kontinuierliche Zahlenreihe
- Beginn und Teilschritte sind frei wählbar

Übung Datenimport

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower left, with fingers slightly curled as if about to touch the globe.

Datenqualität

1. In welchen Kategorien lässt sich Datenqualität beurteilen?

2. Mit welchen Maßnahmen lässt sich die Datenqualität in einem Unternehmen verbessern?

Datenqualität und Datenpflege

Die Attribute der Datenqualität lassen sich in vier Kategorien einteilen:

- Glaubwürdigkeit
- Zeitlicher Bezug
- Nützlichkeit
- Verfügbarkeit

Glaubwürdigkeit

Merkmal	Beschreibung
Korrektheit	Die Daten stimmen inhaltlich mit der Datendefinition überein und sind empirisch korrekt.
Datenherkunft	Die Datenherkunft und die vorgenommenen Datentransformationen sind bekannt.
Vollständigkeit	Alle Daten sind gemäß Datenmodell erfasst.
Widerspruchsfreiheit	Die Daten weisen keine Widersprüche zu Integritätsbedingungen (Geschäftsregeln, Erfahrungswerte) und Wertebereichsdefinition auf (innerhalb des Datenbestands, zu anderen Datenbeständen, im Zeitverlauf).
Syntaktische Korrektheit	Die Daten stimmen mit der spezifischen Syntax (Format) überein.
Zuverlässigkeit	Die Glaubwürdigkeit der Daten ist konstant.

Zeitlicher Bezug

Merkmal	Beschreibung
Aktualität	Datenwerte sind in Bezug auf den gegenwertigen Zeitpunkt erfasst.
Zeitliche Konsistenz	Alle Datenwerte bezüglich eines Zeitpunktes sein gleichermaßen aktuell
Nicht-Volatilität	Die Daten sind permanent und können zu einem späteren Zeitpunkt wieder abgerufen werden.

Nützlichkeit

Merkmal	Beschreibung
Relevanz	Die Datenwerte können auf einen relevanten Datenausschnitt beschränkt werden.
Zeitlicher Bezug	Die Datenwerte beziehen sich auf den benötigten Zeitraum
Verständlichkeit	Die Datensätze müssen in ihrer Begrifflichkeit und Struktur mit den Vorstellungen der Informationsempfänger (z.B. Fachbereiche) übereinstimmen

Verfügbarkeit

Merkmal	Beschreibung
Zeitliche Verfügbarkeit	Die Daten stehen rechtzeitig zur Verfügung.
System-Verfügbarkeit	Das Gesamtsystem ist verfügbar.
Transaktionsverfügbarkeit	Einzelne benötigte Transaktionen sind ausführbar, die Zugriffszeit ist akzeptabel und gleichbleibend
Zugriffsrechte	Die benötigten Zugriffsrechte sind ausreichend.

Wie wird eine gute Datenqualität erreicht?

Um eine gute Datenqualität in Unternehmen sicherstellen zu können, bedarf es eines **aktiven Qualitätsmanagements**.

Der Aufbau und die Pflege eines Qualitätsmanagements ist ein kontinuierlicher Prozess, der verschiedene Schritte umfasst.

1. Analyse der Datenbestände: Identifizieren von Fehlern und Widersprüchen aufgrund von z.B. Dubletten oder fehlerhaften Daten.
2. Bereinigung von Mängeln: In diesem Schritt werden die zuvor identifizierten Mängel durch automatisierte Prozesse oder manuelle Korrekturen behoben.
3. Überwachung der Datenprozesse: Nur eine kontinuierliche Überwachung der Datenprozesse hilft die Qualität der Datenbestände über größere Zeiträume zu bewahren. Regelmäßige Berichte schaffen zudem Transparenz und Vertrauen in die Daten.

Checkliste für eine gute Datenqualität

1. Werden alle Mitarbeitererebenen für das Thema Datenqualität sensibilisiert?
2. Erfolgt die Beurteilung der Datenqualität durch eine Analyse der Datenbestände?
3. Gibt es Regeln für Beschaffenheit und Relevanz von Datenbeständen?
4. Erfolgt eine eindeutige Kompetenzzuweisung? Wer ist verantwortlich für die Datenpflege?
5. Gibt es Standards und Strukturen für eine korrekte Datenerfassung?
6. Gibt es automatisierte Workflows oder erfolgt die Datenerfassung manuell?
7. Werden Mitarbeiter regelmäßig für das Thema Datenqualität sensibilisiert und geschult?

1. Beantworten Sie die Punkte auf der Checkliste zur Datenqualität in Hinblick auf Ihr Unternehmen. Wie steht es um Ihre Datenqualität?

2. Welche Maßnahmen wurden in Ihrem Unternehmen durchgeführt, um die Datenqualität zu verbessern?

Ergänzungen

Datenqualität prüfen

Wie wichtig die Datenqualität für die Datenprozesse ist, wurde bereits erläutert. Es sollte beim Entwurf der Workflows darauf geachtet werden, wie die **Datenqualität systematisch sichergestellt** werden kann.

Dafür eignen sich 3 Vorgehensweisen:

1. Daten und Workflow beim Aufbau überprüfen

- Nach Ausführung eines Knotens immer die **Daten-Outputs anzeigen** und sicherstellen, dass sie den geforderten Standards entsprechen.
- Die Konfigurationen der Knoten sorgfältig den Anforderungen anpassen. Was am Anfang zunächst wie eine belanglose Abweichung aussieht, kann später zum massiven Showstopper werden.
- **Fehler erkennen und beheben** bevor man weiterarbeitet. Auch hier können kleine Problem später sehr groß werden.

Daten prüfen

Nach dem Datenimport:

Habe ich die richtigen Daten eingelesen?

Was für Daten habe ich?

Welche Spalten benötige ich?

Was für Daten brauche ich?

Habe ich weitere Datenquellen desselben Typs?

→ Überprüfen und Experten /
Kollegen fragen

Wie groß ist meine Datenmenge?

→ Zu groß, dann reduzieren

Sind die Datentypen korrekt?

→ Datentypen korrigieren

Sind meine Daten korrekt eingelesen?

→ Überfliegen der Daten

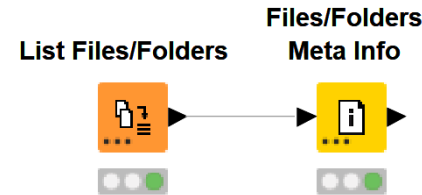
2. Einbau von Kontroll- und Monitoring-Strukturen im Workflow

Automatische Prüfmechanismen helfen die Integrität der Workflows zu überwachen!

Beispiel:

a) Überwachen der Metadaten (Pfad und Zeitstempel).

- Wurden die richtigen Daten verwendet?
- Entsprechen sie der beabsichtigten Auswertungsperiode?



Knoten „Files/Folder Meta Info“ in Verbindung mit „List Files/Folders“.

Es werden zu jeder Datei in dem untersuchten Ordner die Meta-Daten ausgelesen und in eine Tabelle geschrieben. Diese kann dann mit den Sollwerten verglichen werden.

▲ Output Table - 0:15:12 - Files/Folders Meta Info

Table "default" - Rows: 4 Spec - Columns: 7 Properties Flow Variables								
Row ID	P Path	B Directory	L Size	S Size (h...	Last modified date	Creation date	B Exists	
Row0	../_01_Daten/Input/titanic_idclass_data.x...	false	36203	35 KB	2020-03-09T14:33:28...	2021-02-22T11:40:01.385...	true	
Row1	../_01_Daten/Input/titanic_idclass_surviv...	false	4440	4 KB	2020-03-09T14:33:28...	2021-02-22T11:40:01.389...	true	
Row2	../_01_Daten/Input/titanic_rest_data.xlsx	false	72767	71 KB	2020-03-09T14:33:28...	2021-02-22T11:40:01.391...	true	
Row3	../_01_Daten/Input/titanic_rest_survival...	false	7102	6 KB	2020-03-09T14:33:28...	2021-02-22T11:40:01.393...	true	

Einbau von Kontroll- und Monitoring-Strukturen im Workflow

b) Arbeiten mit Grenzwerten und Abgleichen:

Häufig bewegen sich Daten in bestimmten Bereichen. Überwacht man diese Bereiche und kommt es zu Abweichungen, so kann das ein Hinweis für eine Veränderung der Datenqualität sein.

Zur Überwachung eignen sich:

- Minimum, Maximum und Durchschnitt
- Aber auch Anzahl und Häufigkeit von Fehlenden Werten oder Fehleinträgen
- „Checksums“ – Abgleichungen von Rechenwerten

Fast man diese Überprüfungen zu kurzen Berichten zusammen, ist ein durchgängiger Nachweis der Datenintegrität sehr einfach.

3. Regelmäßig Stichproben durchführen und den Workflow kontrollieren

Auch mit den zuvor getroffenen Maßnahmen kann es zu unvorhergesehenen Abweichungen kommen. Daher ist es ratsam, in regelmäßigen Abständen, den Workflow auf Richtigkeit zu überprüfen.

Ist der Workflow sehr komplex und eine vollständige Prüfung aufwendig, können auch gezielte Kontrollpunkte in den Workflow integriert werden, die einen repräsentativen Überblick über den Gesamtworkflow erlauben.

A dark blue rounded rectangle with a white border, containing the title "Explorative Datenanalyse". The background of the slide features a wireframe globe with glowing nodes and connecting lines, set against a blurred cityscape at night. A hand is visible at the bottom left, reaching towards the globe.

Explorative Datenanalyse

Was ist eine explorative Datenanalyse?

Das Ziel der explorativen Datenanalyse

Für Data Analysten ist die explorative Datenanalyse eine Möglichkeit, **verborgene Strukturen oder Auffälligkeiten aufzudecken** und die in den Daten enthaltenen Informationen zu verdichten.

Auf diese Weise können die **wesentlichen Inhalte verdeutlicht** und dargestellt werden und davon ausgehend **Hypothesen abgeleitet** werden.

Durch die explorative Datenanalyse wird also eine große Menge an unbekannten Daten anschaulich und verständlich.

Beschreiben Sie die Vorgehensweise in den einzelnen Teilschritten:

1. Variablen und Datentypen identifizieren
2. Statistische Zusammenfassung
3. Grafische Analyse
4. Ausreißer entdecken
5. „Missing Values“ identifizieren
6. Korrelationen berechnen
7. Hypothesen aufstellen

1. Variablen und Datentypen identifizieren

Variablen			
numerisch [#]		kategorisch [abc]	
diskret [Anzahl]	kontinuierlich [4,68 €]	ordinal [Januar]	nominal [männlich]
Ganze Zahlen (Integer - i) Bsp.: Anzahl Geschwister	Gleitkommazahlen (Double - d) Bsp.: Ticketpreis	Geordneter Text (String - s / Integer - i) Bsp.: Ticketklasse	Ungeordneter Text (String - s) Bsp.: Herkunftsland

Um Daten analysieren und miteinander vergleichbar machen zu können, müssen die Datentypen zunächst identifiziert und klassifiziert werden. Der Datentyp entscheidet darüber, wie mit den Daten weiter verfahren wird.

2. Zusammenfassung in Kennzahlen

Wurden alle Daten klassifiziert, kann mittels der folgenden Werte ein erster statistischer Überblick gewonnen werden:

Allgemeiner Überblick:

- Größe des Datensatzes

Statistik für die einzelnen Variablen:

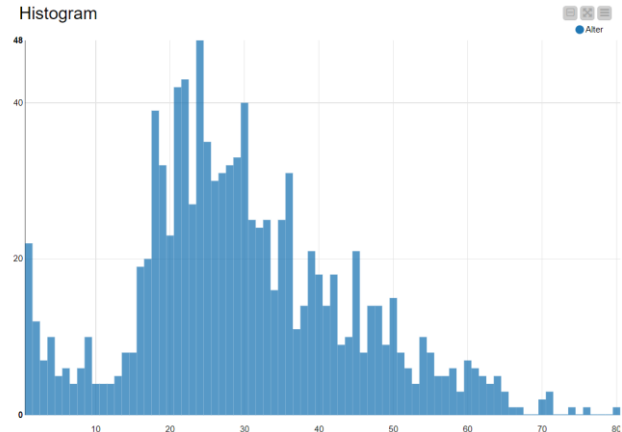
- Anzahl unterschiedlicher Werte
- Minimum, Maximum
- Mittelwert
- Quantile
- Varianz
- Anzahl fehlender Werte (Missing Values)

3. Graphische Darstellung erleichtert die Aufnahme von Informationen

Muster lassen sich leichter erkennen als Zahlenreihen.

Können unbekannte Strukturen anhand von Tabellen und Fließtext nur schwer herausgearbeitet werden, so zeigt die grafische Darstellung auf einen Blick, wo sich Ansammlungen, Zusammenhänge oder Ausreißer befinden.

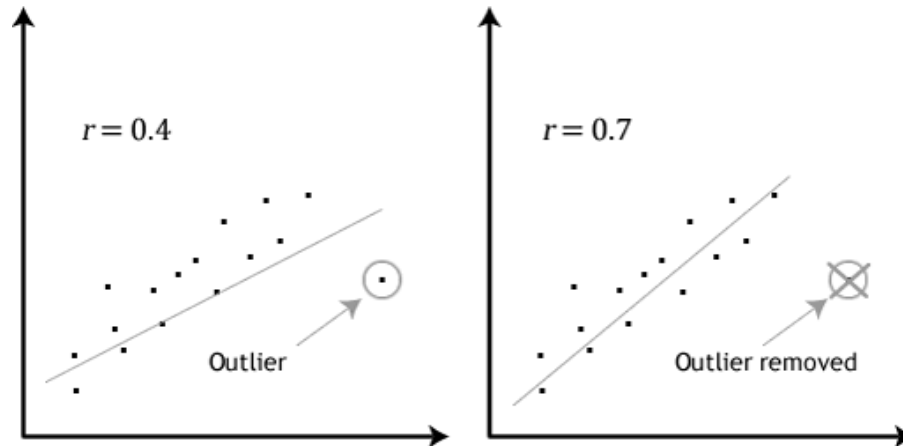
S_Name	S_Geschle...	I_Passag...	I_Klasse	D_Alter	I_Geschle...	S_Ticketnr	D_Preis	S_Kabine	S_Ausgan...	S_Versandt...	I_Them_...	I_Uberleit
Budnell, Mrs. William Robert (Emma Eliza Ward)	female	1	0	60	0	12813	36.292	D15	C	Philadelphia, Pa	0	1
Beckwith, Mr. Richard Leonard	male	4	1	37	1	11751	52.554	D35	S	New York, NY	0	1
Bent, Mr. William Edward	male	7	1	36	1	12258	57.75	C78	C	New York, NY	0	1
Atkins, Mr. William	male	9	1	71	0	PC 17029	46.504	?	C	Montreal, ...	0	0
Hager, Mr. Henry Stepan	male	10	1	48	1	PC 17572	26.729	D33	C	New York, NY	0	1
Carr, Mr. Howard Brown	male	14	0	58	0	10924	24	?	S	Amst., Berkh...	0	0
Burns, Miss. Elizabeth Margaret	female	23	1	41	0	10986	134.5	D40	C	?	0	1
Astor, Col. John Jacob	male	27	1	47	1	PC 17557	227.525	C82 C84	C	New York, NY	0	0
Astor, Mr. Frederick	male	29	0	61	0	20653	32.251	D50	S	Hudson River, NY	0	0
Graham, Mrs. William Thompson (Edith Jenkins)	female	44	1	58	0	PC 17582	155.463	C125	S	Greenwich, CT	1	1
Lindberg, Lind, Mr. Erik Gustaf (The Edward L.)	male	47	1	42	0	17475	26.55	?	S	Stockholm, S...	0	0
Freeman, Mr. Arthur Leonard	male	49	1	61	1	PC 17608	242.175	D17 D19 D61	C	New York, S...	0	0
Crosby, Miss. Margaret R.	female	53	1	34	0	WEP 5735	71	D22	S	Missoula, MT	0	1
Beattie, Mr. Thomas	male	55	1	36	0	12050	75.242	C6	C	Winnipeg, MB	0	0
Beckwith, Mrs. Richard Leonard (Sally Margaret)	female	60	1	47	1	11751	52.554	D35	S	New York, NY	1	1
Charles, Mrs. Norman Campbell (Bertha G.)	female	65	1	33	1	12306	53.1	D9	S	New York, NY	0	1
McGough, Mr. James Robert	male	69	1	36	0	PC 17475	26.288	D25	S	Philadelphia, Pa	0	1
Carras, Mr. Jose Pedro	male	70	1	127	0	12029	47.1	?	S	Montreal, ...	0	0
Frauenthal, Mr. Isaac Gerald	male	72	1	43	1	17765	27.721	D40	C	New York, NY	0	1
Colley, Mr. Edward Pomeroy	male	73	1	47	0	5727	25.587	D58	S	Victoria, BC	0	0
Frauenthal, Mrs. Henry William (Carla Hendrich)	female	75	1	7	1	PC 17611	133.65	?	S	New York, NY	0	1
Hopash, Mrs. Louis Albert (Sally Sophia Fitch)	female	76	1	44	0	11541	33.679	D18	C	Chicago, Ill.	1	1
Crosby, Capt. Edward Gifford	male	82	1	70	1	WEP 5735	71	D22	S	Missoula, MT	0	1
Lundin, Miss. Elie	female	83	1	58	0	PC 17669	146.521	D60	C	?	0	1
Barnes, Miss. Elizabeth	female	87	0	58	0	11783	26.55	C83	S	Brickley, Eng...	0	1
Lindholm, Mrs. Carl Johan (Sigrid Persen)	female	89	1	55	0	112377	27.721	?	C	Stockholm, S...	0	1
Fortune, Mr. Charles Alexander	male	97	1	19	0	10950	263	C33 C25 C27	S	Winnipeg, MB	0	0
Duff Gordon, Dr. Cyrus Edmund (The Morgan)	male	98	1	49	1	PC 17465	96.629	D40	C	London, Paris	0	0
Weyer, Mrs. Edgar Joseph (Leila Sel)	female	103	1	7	1	PC 17604	82.171	?	C	New York, NY	0	0
Carmichael, Mr. Tyrrel William	male	104	1	36	1	10877	78.85	C46	S	Little One Hall...	0	0
Hedden, Mr. John	male	107	0	?	?	PC 17557	227.525	?	C	?	0	0
Leader, Dr. Alice (Franklin)	female	119	1	49	0	17465	25.629	D17	S	New York, NY	0	1
Volmer, Mr. George Durbin	male	127	1	50	1	11203	211.5	C80	C	Elmira Park, Pa	1	0
Ward, Mrs. George Andrew (Dorothy Annan)	female	130	1	29	1	12763	35.442	D10	C	Boston, NY	0	1
Cheney, Mr. Paul Rosamine	male	136	1	45	0	PC 17594	26.7	D6	C	Paris, France	0	1
Sempick, Miss. Augusta	female	137	1	30	0	11270	21	?	C	?	0	1
Shaperson, Mrs. Walter Bertram (Martha G.)	female	141	1	52	1	10147	78.267	D20	C	New York, Pa	0	1



4. Ausreißer erkennen und bearbeiten

Ausreißer können nachfolgende Verarbeitungsschritte behindern/verfälschen:

- Sie verändern Reports, Diagramme, Algorithmen, Trends
- Sie können durch fehlende, falsche oder korrupte Daten entstehen



Umgang mit Ausreißern

- a. Zunächst muss geprüft werden, ob es sich um echte Mess- bzw. Datenwerte oder um Fehler handelt. Ist es ein Fehler sollte zur Verbesserung der Datenqualität für weitere Messungen bzw. Datenerhebungen die Fehlerquelle geprüft und der Fehler behoben werden.

Beispiel:

Hat der Offizier auf der Titanic beim Notieren des Alters eines Passagiers 224 statt 24 geschrieben.

- b. Handelt es sich um echte Werte, wird in einem nächsten Schritt geprüft, ob es sich um zufällige oder systematische Ausreißer handelt.

Beispiel:

Hat der Offizier auf der Titanic im Bordbuch eine falsche Spalte befüllt, ohne es zu bemerken, ist das ein systematischer Fehler.

Wenige zufällige Ausreißer können bereinigt werden. Systematische Ausreißer sollten genauer überprüft und gegebenenfalls durch eine zusätzliche Analyse erklärt werden.

5. Fehlende Werte identifizieren

Oft werden Systemeingaben, Fragebögen etc. unvollständig oder fehlerhaft befüllt – das führt zu fehlenden Werten im Datensatz.

Es finden sich dann folgende Tabelleneigenschaften:

→ Wert vs. 0 / Leere Zelle / NA (auch: NULL, NaN, ?)

Fehlende Werte können dennoch Informationen vermitteln!

- Bei Einkommensangaben werden Geringverdiener die Spalte ggf. nicht befüllen.
- Testteilnehmer erscheinen nur, wenn sie bestehen können.
- Teilnehmer medizinischer Studien verlassen diese aufgrund negativer Ereignisse.

Das bedeutet: Wir müssen berücksichtigen, ob fehlende Werte rein zufällig oder in Abhängigkeit von Drittfaktoren entstehen und entsprechend damit umgehen.

In welche Kategorien lassen sich fehlende Werte einteilen und wie arbeitet man mit fehlenden Werten?

Fehlende Werte lassen sich in drei Kategorien einteilen:

1. Völlig zufällig (Missing completely at random, MCAR):

Das Auftreten von fehlenden Werten ist rein zufällig und ist vollkommen unabhängig von den Eigenschaften der Werte, der Subjekte und der Quellen der Daten.

2. Bedingt zufällig (Missing at random, MAR):

Das Auftreten von fehlenden Werten ist zufällig in Bezug der Eigenschaften der Werte jedoch nicht zufällig in Bezug auf Subjekte oder Quellen von Daten.

Beispiel: Unterschiede in der Häufigkeit von fehlenden Daten bei Befragungen von Männern und Frauen.

3. Nicht zufällig (Missing not at random, MNAR):

Das Auftreten von fehlenden Werten hängt von den Eigenschaften der Werte sowie der Subjekte und Quellen der Daten ab.

Beispiel: Männer beantworten weniger persönliche Fragen zu Depressionen als Frauen.

Arbeiten mit fehlenden Werten

Bei der Bewertung fehlender Werte muss also berücksichtigt werden, wie der gesamte Datensatz zustande gekommen ist und welche Einflüsse auf die Datenerhebung eingewirkt haben. Statt einzelne Werte leichtfertig zu löschen, sollten diese gründlich analysiert und interpretiert werden.

Möglichkeiten zum Umgang mit fehlenden Werten:

1. Löschen von Daten:

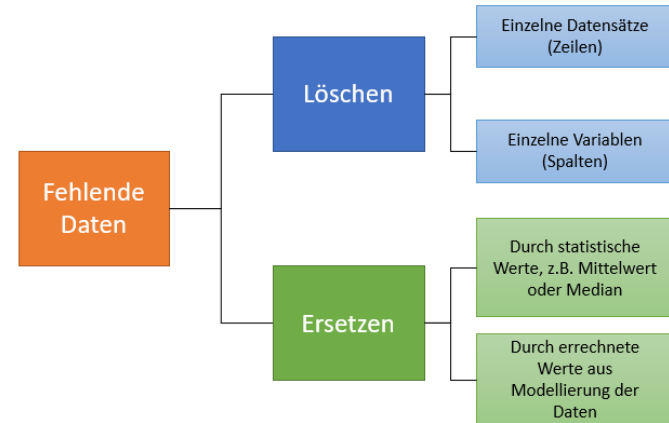
Löschen einzelner Datensätze (Zeilen) oder Variablen (Spalten).

2. Ersetzen durch statistische Werte:

Ergänzen der fehlenden Werte durch Mittelwerte, Mediane, oder Modus bzw. lineare Interpolationen.

3. Ersetzen durch Analyse- oder modellbasierte Daten

Anwendung von Analysemethoden und Modellen zur Ergänzung der fehlenden Werte. Wichtig ist hier, geeignete Modelle zu wählen, d.h. bei MCAR und MAR Modelle für unabhängige Variablen und bei MNAR Modelle für abhängige Variablen.



6. Korrelation

Was ist Korrelation?

Die Korrelation ist ein Maß aus der Statistik, das ausdrückt, inwieweit zwei Merkmale in einer linearen Beziehung zueinanderstehen. "Linear" heißt, sie verändern sich in einem festen Verhältnis zueinander: Wenn sich die Anzahl der Passagiere verdoppelt, verdoppelt sich auch die Anzahl der benötigten Essensportionen. Korrelationen beschreiben also Daten, die sich zusammen verändern.

Wie wird Korrelation gemessen?

Die Stärke der linearen Beziehung zwischen den Variablen wird durch den Korrelationskoeffizient „ r “ angegeben. Korrelationen werden auch auf statistische Signifikanz überprüft.

Der **Korrelationskoeffizient** ist ein einheitsloses Maß und reicht von -1 bis +1

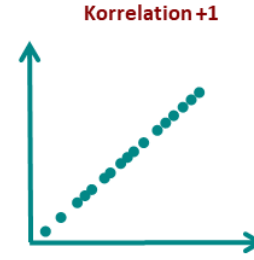
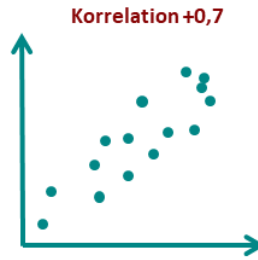
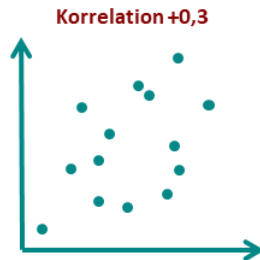
Je näher „ r “ bei Null liegt, desto schwächer ist der lineare Zusammenhang.

Positive r -Werte zeigen eine positive Korrelation an, d.h. die Werte beider Variable steigen gemeinsam an. **Negative r -Werte** zeigen eine negative Korrelation an, d.h. die Werte einer Variable steigen an, wenn gleichzeitig die Werte der anderen Variablen fallen.

Aussage von Korrelationsdaten

- Anhand des **Korrelationseffizient r** kann man sehen, wie stark zwei Variablen voneinander abhängig sind.
- Besonders deutlich wird dies, wenn man beide Variablen gegeneinander in einem Streudiagramm aufträgt.
- Je größer der Zusammenhang zwischen den beiden Variablen wird, desto kleiner die Streuung der Datenpunkte. Bei $r=1$ wird die mathematische Funktion exakt beschrieben, d.h. bei einer linearen Korrelation sind alle Datenpunkte auf einer Geraden.

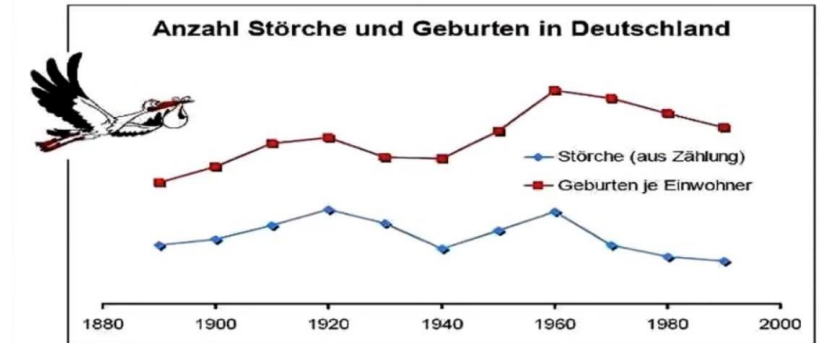
r	Zusammenhang
0,0 - 0,1	Kein Zusammenhang, willkürliche Streuung
0,1 - 0,3	Geringer Zusammenhang
0,3 - 0,5	Mittlerer Zusammenhang
0,5 - 0,7	Hoher Zusammenhang
0,7 - 1,0	Sehr hoher Zusammenhang



Korrelation und Kausalität

- Die Korrelation betrachte nur die statistische Beziehung der untersuchten Variablen und berücksichtigt weder das Vorhandensein noch den Effekt anderer Faktoren.
- Von einer Kausalität spricht man, wenn zwischen zwei Merkmalen ein Zusammenhang aus Ursache und Wirkung besteht.
- Korrelationen sind **ein Hinweis aber kein Beweis für Kausalitäten**, also bewiesene Ursachen- und Wirkungszusammenhänge.

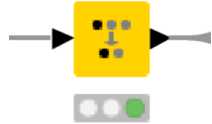
Korrelationen, die keine direkte Aussage zu Ursache und Wirkung haben, nennt man **Scheinkausalitäten**.
Ein gutes Beispiel hierfür ist die Korrelation zwischen Anzahl der Störche und der Geburtenrate in Deutschland.



Quelle: Statistisches Bundesamt

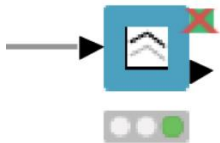
Weitere Knoten zur EDA

**Duplicate
Row Filter**



- Erkennt Dubletten innerhalb von Spalten
- Dubletten können gekennzeichnet oder gefiltert werden
- Kennzeichnung der „Originalzeile“ möglich

**Parallel
Coordinates Plot**



- Stellt die Verteilung alle Werte der ausgewählten Spalten nacheinander dar
- Durch Verbindungslinien können zusammenhänge zwischen Spalten gezeigt werden

Übung Explorative Datenanalyse

Daten bearbeiten und transformieren

1. Was sind die Merkmale von relationalen Daten?

2. Wie werden Daten in Tabellen miteinander verknüpft?

Wiederholung: Relationale Daten

Die Struktur von relationalen Datenbanken und Tabellen ist so angelegt, dass jede Zeile einer Tabelle einen zusammenhängenden Datensatz darstellt, der über eine ID miteinander verknüpft ist.

Diese ID kann dabei in verschiedenen Formen gewählt sein. Häufig wird eine Zeilen-ID gesetzt, um die Zeilen zu identifizieren.

Es können aber auch inhaltliche IDs wie z.B. Kundennummern, Verzeichnis IDs, Personalnummern, etc. sein. Eine weitere Möglichkeit besteht darin, sie beispielsweise über einen zeitlichen Verlauf zu identifizieren. Zeitstempel oder Datum sind hier gebräuchlich Formen der Identifizierung.

I PassagierID	S Name	S Geschle...	S Ausgangshafen
1	Bucknell, Mrs. William Robert (Emma Eliza Ward)	female	C
4	Beckwith, Mr. Richard Leonard	male	S
2	Karaic, Mr. Milan	male	S
3	Funk, Miss. Annie Clemmer	female	S
5	Reynaldo, Ms. Encarnacion	female	S

Vertikales aneinanderhängen von Tabellen

Sind die Tabellen in ihrer Attribute- Struktur (Spalten) nahezu übereinstimmend und die IDs der Tabelle untereinander nicht mehrfach vergeben, können Tabellen zusammengeführt werden, indem eine Tabelle mit weiteren vertikal erweitert wird. Deren Zeilen werden dabei einfach an die der ersten Tabelle angehängt.

Dabei kann man wählen, ob alle Attribute in die neue Tabelle übernommen werden soll (Union) oder nur die Attribute zusammengeführt werden sollen, die bei allen Tabellen gemeinsam enthalten sind (Intersection):

Verketteten – vertikales Aneinanderhängen von Tabellen

Tabelle 1

I	PassagierID	S	Name	S	Geschle...	S	Ausgangshafen
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)		female		C
4			Beckwith, Mr. Richard Leonard		male		S
2			Karaic, Mr. Milan		male		S
3			Funk, Miss. Annie Clemmer		female		S
5			Reynaldo, Ms. Encarnacion		female		S

Tabelle 2

I	PassagierID	S	Name	S	Geschlecht
7			Mock, Mr. Philipp Edmund		male
9			Artagaveytia, Mr. Ramon		male
10			Harper, Mr. Henry Sleeper		male
6			Culumovic, Mr. Jeso		male
8			Trout, Mrs. William H (Jessie L)		female



Tabelle 1+2, Intersection

I	PassagierID	S	Name	S	Geschlecht
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)		female
4			Beckwith, Mr. Richard Leonard		male
2			Karaic, Mr. Milan		male
3			Funk, Miss. Annie Clemmer		female
5			Reynaldo, Ms. Encarnacion		female
7			Mock, Mr. Philipp Edmund		male
9			Artagaveytia, Mr. Ramon		male
10			Harper, Mr. Henry Sleeper		male
6			Culumovic, Mr. Jeso		male
8			Trout, Mrs. William H (Jessie L)		female

Tabelle 1+2, Union

I	PassagierID	S	Name	S	Geschlecht	S	Ausgangshafen
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)		female		C
4			Beckwith, Mr. Richard Leonard		male		S
2			Karaic, Mr. Milan		male		S
3			Funk, Miss. Annie Clemmer		female		S
5			Reynaldo, Ms. Encarnacion		female		S
7			Mock, Mr. Philipp Edmund		male	?	
9			Artagaveytia, Mr. Ramon		male	?	
10			Harper, Mr. Henry Sleeper		male	?	
6			Culumovic, Mr. Jeso		male	?	
8			Trout, Mrs. William H (Jessie L)		female	?	

Tabellen horizontal erweitern - Join

Beim Zusammenführen von Tabellen werden die IDs genutzt, um Attribute aus den ursprünglichen Tabellen im neuen Kontext zusammenzuführen und damit eine neue vereinte Tabelle zu schaffen.

Auch hier ist eine Tabelle führend, die Haupttabelle und die verknüpften Tabellen werden in diese integriert. Die Haupttabelle wird laut Konvention als „Linke Tabelle“ bezeichnet.

Es gibt 4 Typen von „Join“, die je nach ihrer Anwendung unterschiedliche Ergebnisse erbringen:

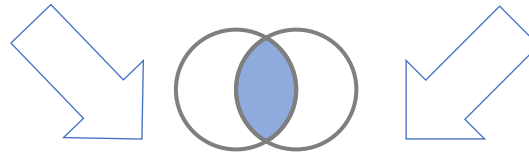
Wie funktioniert ein „JOIN“ von Tabellen?

Inner Join

Der „Inner Join“ führt nur die Daten zusammen, für die bei beiden Tabellen eine gemeinsame IDs vorhanden ist. Sie bilden eine gemeinsame Schnittmenge:

I	PassagierID	S	Name	S	Geschlecht
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)		female
4			Beckwith, Mr. Richard Leonard		male
2			Karaic, Mr. Milan		male
3			Funk, Miss. Annie Clemmer		female
5			Reynaldo, Ms. Encarnacion		female

I	PassagierID	D	Alter	D	Preis
4			37		52.554
7			30		57.75
3			38		13
5			28		13
6			17		8.662



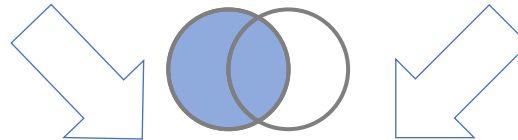
I	Passagi...	S	Name	S	Geschle...	D	Alter	D	Preis
4			Beckwith, Mr. Richard Leonard		male		37		52.554
3			Funk, Miss. Annie Clemmer		female		38		13
5			Reynaldo, Ms. Encarnacion		female		28		13

Left Outer Join

Beim „Left Outer Join“ werden alle Daten von der Haupttabelle übernommen und nur die Daten aus der 2. Tabelle übernommen, die eine übereinstimmende ID besitzt:

I	PassagierID	S	Name	S	Geschlecht
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)	female	
4			Beckwith, Mr. Richard Leonard	male	
2			Karaic, Mr. Milan	male	
3			Funk, Miss. Annie Clemmer	female	
5			Reynaldo, Ms. Encarnacion	female	

I	PassagierID	D	Alter	D	Preis
4			37	52.554	
7			30	57.75	
3			38	13	
5			28	13	
6			17	8.662	



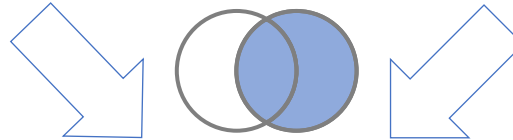
I	PassagierID	S	Name	S	Geschlecht	D	Alter	D	Preis
4			Beckwith, Mr. Richard Leonard	male		37		52.554	
3			Funk, Miss. Annie Clemmer	female		38		13	
5			Reynaldo, Ms. Encarnacion	female		28		13	
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)	female		?		?	
2			Karaic, Mr. Milan	male		?		?	

Right Outer Join

Der “Right Outer Join“ funktioniert analog zum „Left Outer Join“. Hier werden die Daten der 2. Tabelle vollständig übernommen und die der Haupttabelle nur bei übereinstimmenden IDs:

I	PassagierID	S	Name	S	Geschlecht
1			Bucknell, Mrs. William Robert (Emma Eliza Ward)		female
4			Beckwith, Mr. Richard Leonard		male
2			Karaic, Mr. Milan		male
3			Funk, Miss. Annie Clemmer		female
5			Reynaldo, Ms. Encarnacion		female

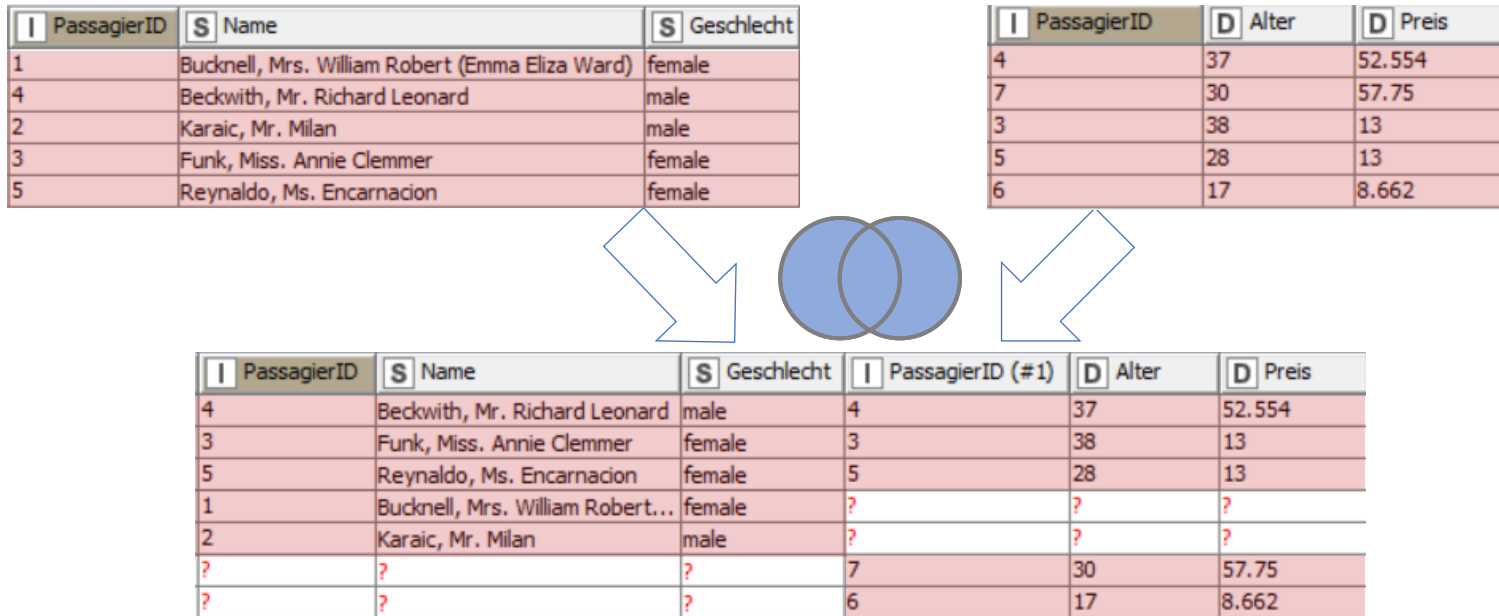
I	PassagierID	D	Alter	D	Preis
4			37		52.554
7			30		57.75
3			38		13
5			28		13
6			17		8.662



I	PassagierID	S	Name	S	Geschle...	I	PassagierID (#1)	D	Alter	D	Preis
4			Beckwith, Mr. Richard Leonard		male	4			37		52.554
3			Funk, Miss. Annie Clemmer		female	3			38		13
5			Reynaldo, Ms. Encarnacion		female	5			28		13
?			?		?	7			30		57.75
?			?		?	6			17		8.662

Full Outer Join

Beim „Full Outer Join“ werden alle Daten der Haupt- und Nebentabelle übernommen und über die ID verknüpft. Auch hier können die IDs unterschiedlich übernommen werden.



1. In welchen Datentypen liegen digitale Daten vor und was sind ihre Merkmale?

2. Was ist bei ihrer Konvertierung und Wertetransformation zu beachten?

Datentypen

Der Datentyp hat einen wichtigen Einfluss auf die Verarbeitung der in ihm enthaltenen Daten.

Dabei gibt der Typ an, welcher Art die Daten sind, in welcher Struktur sie vorliegen und wie groß die Daten maximal sein können.

Datentypen sind Repräsentationen von Binärstrukturen. Dabei gibt die Anzahl an Binärstellen (Bits) an, wie groß eine Variable eines Datentyps maximal gewählt werden kann:

Beispiel:

Integer: Ganze Zahlen von -128 bis 127 **bei 8 Bit**, -32768 bis 32767, **bei 16 Bit**, usw.

Die Zahl „5“ würde bei 8 Bit im Binärcode „0 0000 0101“ lauten, wobei die erste Stelle für das Vorzeichen genutzt wird.

Weitere Beispiele für Datentypen sind:

- **Gleitkommazahlen** (z.B. 3,434) wie Float (32 Bit) oder Double (64 Bit).
- **Zeichen** (z.B. Text) wie Byte (8 Bit) Short (16 Bit) oder Long (64 Bit)
- **Boolesche (Boolean) Variable** (1 Bit): 1 = Wahr, 0 = Falsch

Typkonvertierung

Bei der Bearbeitung von Daten kommt es häufiger vor, dass ein Datentyp in einen anderen umgewandelt werden muss.

Beispiel:

Daten, die aus Textdateien wie CSV-Dateien importiert werden, enthalten keine Kennzeichnung, um welchen Datentyp es sich handelt. Einige Anwendungen verfügen über eine Typerkennung beim Datenimport und schlagen dem Anwender den wahrscheinlichsten Datentyp vor. Diese Automatisierung sollte jedoch kritisch genutzt und genau geprüft werden.

Die Umwandlung des Datentyps geschieht in der Regel, indem die Variable bzw. das Attribut (Spalte) dem neuen Datentyp zugewiesen wird.

Zu beachten ist hier, dass dieser Vorgang von der Anwendung konsequent umgesetzt wird, egal ob die Daten mit dem neuen Datentyp kompatibel sind oder nicht. Zahlen lassen sich in der Regel ohne Probleme in Zeichen umwandeln, umgekehrt geht dies jedoch nur, wenn ausschließlich Zahlenzeichen verwendet wurden, Buchstaben führen zu Fehlern.

Sondertyp: Datum und Zeit

In vielen Anwendungen stellt Datum und Zeit den Nutzer vor große Herausforderungen:

- Sie sind weder dezimal noch durchgängig gleichförmig:
 - Datum: Jahr – Quartal – Monat – Kalenderwoche – Tag
(Wochentag, Julianischer Tag, Kalendertag...)
 - Zeit: Stunde – Minute – Sekunde – dezimale Untereinheiten der Sekunde
- Die Einheiten haben teilweise unterschiedliche Werte (z.B. Monatsdauer, Schaltjahre)
- Sie werden in sehr vielen verschiedenen Formate dargestellt: offizielle, abgewandelte, nationale und regionale, softwaresystembasierte

Datum und Zeit

Beispiele für unterschiedliche Lösungen:

Excel:

- Zählt Tage in ganzen Zahlen ab Ursprung (1.1.1900, aber nur in Windows Excel, Ursprung im MacOS ist 1.1.1904)
- Nachkommastellen geben Fraktionen des Tags → Umrechnung von dezimal in 60iger System
- Das angezeigte Datum wird mit einer Formel im Hintergrund berechnet

Knime:

- Datum und Zeit als Zeichenfolge: yyyy-MM-dd, oder dd.MM.yyyy hh:mm:ss
- Zeitdauer z.B. in ISO-8601-Form P1Y2M3D4H (1 year 2 months 3 days 4 hours)

Berechnungen mit Datum und Zeit sind häufig in der Anzeigeform nicht möglich und es empfiehlt sich, Elemente des Datums in Zeichen oder Zahlen umzuwandeln, um mit diesen gesondert Berechnungen durchzuführen:

Wertetransformation

1. Zeichen

Hier werden Zeichensätze gelöscht, überschrieben, geteilt, ergänzt, usw. Dabei beschreibt die Formel, welcher Teil des Zeichensatzes verändert werden soll und in welcher Weise.


Beispiel:

Ein Index besteht aus 3 Buchstaben und einem Datum:

ABC01042021

Durch die Formel wird festgelegt, dass die ersten 3 Zeichen entfernt werden, um das Datum zu erhalten:

→ **01042021**



S Index	S Datum
abc01042021	01042021
cba02042021	02042021
def03042021	03042021
ghi04042021	04042021
jkl05042021	05042021

Das extrahierte Text-Datum kann anschließend in ein Datumsformat umgewandelt und für weitere Berechnungen genutzt werden.

Wertetransformation

2. Zahlen

In der Zahlentransformation werden die gebräuchlichen mathematischen Formeln verwendet, um den Zahlenwert zu verändern:

Addition, Subtraktion, Multiplikation,


Dies kann mit Fixwerten oder aber auch mit Variablen anderer Attribute durchgeführt werden

Beispiel Provisionsberechnung:

Verkaufte Ware: 20.000€

Provisionssatz 10%

Provision: 1.000 €



L Erlös	D Provisionssatz	D Provision
10000	0.1	1,000
20000	0.15	3,000
30000	0.2	6,000

Wertetransformation

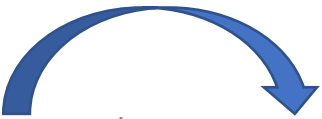
3. Regelbasierte Wertetransformation:

Diese Transformation bearbeitet analog zu den 2 vorherigen Transformationen Zeichen und Zahlen. Der Unterschied besteht jedoch darin, dass die Transformation nur ausgeführt wird, **wenn eine vorher festgelegte Bedingung erfüllt wird**.

Beispiel aus den Titanic-Daten:

Für die Berechnung der Korrelation wurde das Attribut „Geschlecht“ in die Zahl 1 umgewandelt, für die Bedingung: **Attributwert (Geschlecht) = „weiblich“**

Analog die Umwandlung für das Geschlecht „männlich“ in die Zahl 0



S Geschlecht	I Geschlecht, numerisch
female	1
male	0
male	0
male	0
male	0
male	0
female	1
male	0

Wie lassen sich Tabellen
transformieren?

Tabellentransformation

1. Sortieren

Neben der Transformation von Datenwerten nach Attribut können auch ganze Tabellen verändert werden. Dies wirkt sich auf die Gestaltung der Reihen wie auch auf die der Spalten aus.

Eine der einfachsten Formen der Tabellentransformation ist das **Sortieren von Reihen und Spalten**. Dabei werden Reihen in der Regel nach den Werten bestimmter Attribute sortiert (z.B. nach aufsteigender ID, absteigenden Kosten, usw.) während Spalten nach einer definierten Reihenfolge angeordnet werden, um die Daten übersichtlich zu gestalten.

Eine besondere Form der Neusortierung ist das **Transponieren**, wobei Zeilen in Spalten und Spalten und Zeilen umgewandelt werden. Diese Funktion wird vor allem dann notwendig, wenn Daten aus frei gestaltbaren Tabellenkalkulationsprogrammen importiert werden, die nicht der Logik der reihenweise aufgebauten Datensätze folgen

Tabellentransformation

2. Filtern

Eine wichtige Funktion beim Bearbeiten von Daten ist die Auswahl der relevanten Daten und die Begrenzung des Datenumfangs. Dies wird durch Filter erreicht. Auch hier können sowohl die Spalten als auch die Reihen bearbeitet werden. Bei Reihenfiter wird in der Regel entsprechend eines Attributwertes gefiltert.

Beispiel Reihenfiter:

Attributeigenschaft: Alle Personen die älter als 18 sind.

Die Reihen der Tabelle, die diese Werte nicht enthalten, werden entfernt oder gekennzeichnet.

Bei Spaltenfiltern werden die relevanten Spalten vom Anwender ausgewählt und die übrigen entfernt.

Tabellentransformation

3. Aufteilen

So wie man Tabellen zusammenführen kann, kann man sie auch wieder trennen. Diese Funktion führen sogenannte „Splitter“ durch.

Ein Spalten-Splitter verteilt die Spalten auf zwei Tabellen gemäß der Auswahl des Anwenders, beim Reihensplitter werden sie anhand eines Attributwertes verteilt.

Tabellentransformation

4. Aggregationen

Bei der Datenaggregation werden Informationen gesammelt und in einer zusammenfassenden Form ausgedrückt, um beispielsweise Kennzahlen zu generieren.

Bei der Aggregation geht es darum, mehr Informationen über eine spezielle Zielgruppe zu erhalten, beispielsweise über das Alter, den Beruf oder das Einkommen.

Wie werden Werte in Tabellen
aggregiert?

Aggregationen


1. Gruppieren

Zusammenfassen (Aggregation) von Daten zu Meta- bzw. Kenndaten.

- **Gruppen (auch als Dimensionen zu betrachten):** Spalte, aus der für jeden einzigartigen Wert eine Zeile entsteht
- **Werte:** über Aggregationsformeln berechnete Daten (Summe, Mittelwert, ...)

Beispiel:

Berechnen des mittleren Alters der Titanic-Passagiere nach Kabinenklasse



S	Name	S	Geschle...	I	Klasse	D	Alter
	Abbing, Mr. Anthony		male	3			42
	Abbott, Master. Eugene Joseph		male	3			13
	Abbott, Mr. Rosmore Edward		male	3			16
	Abbott, Mrs. Stanton (Rosa Hunt)		female	3			35
	Abelseth, Miss. Karen Marie		female	3			16
	Abelseth, Mr. Olaus Jorgensen		male	3			25
	Abelson, Mr. Samuel		male	2			30
	Abelson, Mrs. Samuel (Hannah Wilzoky)		female	2			28
	Abrahamsson, Mr. Abraham August Johannes		male	3			20
	Abraham, Mrs. Joseph (Sophie Halaut Easu)		female	3			18
	Adahl, Mr. Mauritz Nils Martin		male	3			30
	Adams, Mr. John		male	3			26
	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)		female	3			40
	Aks, Master. Philip Frank		male	3			0.833
	Aks, Mrs. Sam (Leah Rosen)		female	3			18
	Albimona, Mr. Nassef Cassem		male	3			26
	Aldworth, Mr. Charles Augustus		male	2			30
...	Alexander, Mr. William		male	3			26



I	Klasse	D	Mean(Alter)
1			39.16
2			29.507
3			24.816

Gruppe/Detail:
Aggregation:

Kabinenklasse
Durchschnitt

Aggregationen

2. Pivot

Die nächste Stufe der Aggregation ist das Erstellen einer Kreuztabelle oder auch Pivot-Tabelle. Der Einsatz einer Pivot – Tabelle ist sinnvoll, um Beziehungen in einer große Datenmenge mit vielen Attributen aufzudecken.

Dabei werden die Daten nach zwei Attribute aggregiert (Gruppe und Pivot), wobei das eine Attribut die Spalten und das andere Attribut die Zeilen bildet. Der Wert, über den die Aggregation gebildet wird, wird zu Kennwerten wie Mittelwert, Summe oder Anzahl berechnet.

Es können auch mehrere Attribute in Gruppe und Pivot verwendet werden, jedoch sollte dabei darauf geachtet werden, dass diese in einem sinnvollen Verhältnis zueinander stehen, beispielsweise in einer Hierarchie.

Pivoting

S ▲ Name	S Geschle.	I Klasse	D Alter
Abbing, Mr. Anthony	male	3	42
Abbott, Master. Eugene Joseph	male	3	13
Abbott, Mr. Rossmore Edward	male	3	16
Abbott, Mrs. Stanton (Rosa Hunt)	female	3	35
Abelseth, Miss. Karen Marie	female	3	16
Abelseth, Mr. Olaus Jorgensen	male	3	25
Abelson, Mr. Samuel	male	2	30
Abelson, Mrs. Samuel (Hannah Witosky)	female	2	28
Abrahamsson, Mr. Abraham August Johannes	male	3	20
Abraham, Mrs. Joseph (Sophie Halaut Easu)	female	3	18
Adahl, Mr. Mauritz Nils Martin	male	3	30
Adams, Mr. John	male	3	26
Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	female	3	40
Aks, Master. Philip Frank	male	3	0.833
Aks, Mrs. Sam (Leah Rosen)	female	3	18
Albimona, Mr. Nassef Cassem	male	3	26
Aldworth, Mr. Charles Augustus	male	2	30
Alexander, Mr. William	male	3	26

Erstellen einer Kreuztabelle:

- **Gruppen** werden **Zeilen**
- **Pivots** werden **Spalten**
- **Aggregationen** werden **Zellen**

I Klasse	D female +Mean(Alter)	D male +Mean(Alter)
1	37.038	41.029
2	27.499	30.815
3	22.185	25.962

Gruppe :

Kabinenklasse

Pivot:

Geschlecht

Aggregation:

Durchschnitt (Alter)

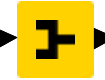
Pivoting vs. GroupBy

- Kreuztabelle vs. Liste

Pivoting



GroupBy



Pivot table - 12:1 - Pivoting

File Hilite Navigation View

Table "default" - Rows: 28 Spec - Columns: 15 Properties Flow Variables

Row ID	S BR	S Iracem...	S Hamba...	S Uuskau...	S Ho Chi ...	S Fa. SFT...	S Bremen...	S Valmet...	S Sindelf...	S OKD-Fe...	S Tuscalo...	S Novo M...	S East Lo...	S Rastatt...	S OKD Fe...
Row0	220	G	?	?	?	?	?	?	?	?	?	?	?	?	?
Row1	227	?	B, G	?	?	?	?	?	?	?	?	?	?	?	?
Row2	248	?	B, A	?	?	?	?	?	?	?	?	?	?	?	?
Row3	284	?	B, A	?	?	?	?	?	?	?	?	?	?	?	?
Row4	294	?	E, F, G	E	?	?	?	?	?	?	?	?	?	?	?
Row5	314	?	?	?	D, B	?	?	?	?	?	?	?	?	?	?
Row6	328	?	?	?	?	G, B	?	?	?	?	?	?	?	?	?
Row7	464	?	F	G	?	?	F	G, D, B, F	?	?	?	?	?	?	?
Row8	490	?	?	?	?	?	?	?	G, F	?	?	?	?	?	?
Row9	504	?	?	?	B	?	?	?	?	?	?	?	?	?	?
Row10	524	?	?	?	?	?	?	B	?	?	?	?	?	?	?
Row11	528	?	?	?	?	?	?	?	?	B, A	F	?	?	?	?
Row12	607	?	?	?	?	D	?	?	?	?	?	?	?	?	?
Row13	608	?	F, D	F	?	?	?	?	?	?	?	?	?	?	?
Row14	642	?	?	?	?	?	?	?	B	?	?	?	?	?	?
Row15	644	?	?	?	?	?	?	C	?	?	?	?	?	?	?
Row16	660	F	?	?	?	?	?	?	?	?	?	?	?	?	?
Row17	661	G	?	?	?	?	?	?	?	?	?	?	?	?	?
Row18	674	?	F, D	?	?	?	?	?	?	?	?	?	?	?	?
Row19	710	?	?	?	?	?	?	C	?	?	?	?	?	?	?
Row20	719	?	?	G	?	?	?	G, B	?	?	?	?	G	?	?
Row21	809	?	?	?	?	?	?	?	E	?	?	?	?	?	?
Row22	842	?	?	F	?	?	F	D, F, B, A	F	?	?	?	?	E	F
Row23	873	?	E, F	?	?	?	?	?	?	?	?	?	?	?	?
Row24	875	?	?	?	?	?	?	B, A	?	?	?	?	?	?	?
Row25	912	?	?	?	?	?	?	?	A, B	?	?	?	?	?	?
Row26	945	?	B	?	?	?	?	?	?	?	?	?	?	?	?
Row27	950	?	B	?	?	?	?	?	?	?	?	?	?	?	?

Group table - 12:2 - GroupBy

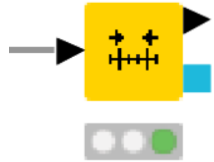
File Hilite Navigation View

Table "default" - Rows: 43 Spec - Columns: 3 Properties Flow Variables

Row ID	S BR	S WERK rand	S Unique conca...
Row0	220	Iracemapolis (PKW)	G
Row1	227	Hambach	B, G
Row2	248	Hambach	B, A
Row3	284	Hambach	B, A
Row4	294	Hambach	E, F, G
Row5	294	Uusikaupunki Finnland	E
Row6	314	Ho Chi Minh City	D, B
Row7	328	Fa. SFT (Graz)	G, B
Row8	464	Bremen	F
Row9	464	Hambach	F
Row10	464	Uusikaupunki Finnland	G
Row11	464	Valmet	G, D, B, F
Row12	490	Sindelfingen	G, F
Row13	504	Ho Chi Minh City	B
Row14	524	Valmet	B
Row15	528	OKD-Fertigung USA	B, A
Row16	528	Tuscaloosa	F
Row17	607	Fa. SFT (Graz)	D
Row18	607	Iracemapolis (PKW)	D
Row19	608	Hambach	F, D
Row20	608	Uusikaupunki Finnland	F
Row21	642	Sindelfingen	B
Row22	644	Valmet	C
Row23	660	Iracemapolis (PKW)	F
Row24	660	Novo Mesto	F
Row25	661	Iracemapolis (PKW)	G
Row26	674	Hambach	F, D
Row27	710	Valmet	C
Row28	719	East London	G
Row29	719	Uusikaupunki Finnland	G
Row30	719	Valmet	G, B
Row31	809	Rastatt	F
Row32	809	Sindelfingen	E
Row33	842	Bremen	F
Row34	842	OKD Fertigung Deutschland	F
Row35	842	Sindelfingen	F
Row36	842	Uusikaupunki Finnland	F
Row37	842	Valmet	D, F, B, A
Row38	873	Hambach	E, F
Row39	875	Valmet	B, A
Row40	912	OKD-Fertigung USA	A, B
Row41	945	Hambach	B
Row42	950	Hambach	B

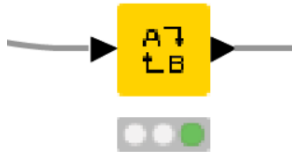
Weitere Knoten zur Bearbeitung von Daten

Normalizer



- Normalisiert Zahlenwerte
- Zielwertebereich ist wählbar: min/max, Gauß, exponentiell

Column Rename



- Ersetzt Spaltenbezeichnungen

Übung: Daten bearbeiten und transformieren

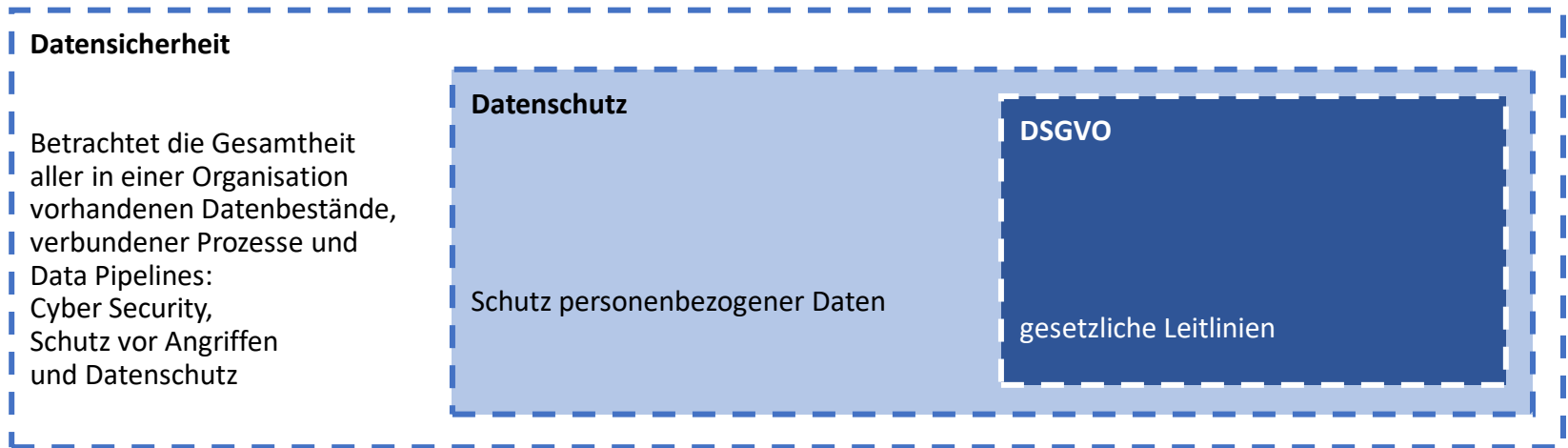
The background of the slide features a blue-toned image of a hand reaching out to touch a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and digital technology. The hand is positioned in the lower-left foreground, with fingers extended towards the globe.

Datensicherheit und Datenschutz

Was ist Datensicherheit und
Datenschutz? In welcher Beziehung
stehen sie zueinander?

Datensicherheit und Datenschutz

Der Datenschutz befasst sich mit personenbezogenen Daten (besonders schützenswerte Daten) und ist nur ein Teilaspekt der Datensicherheit.



Datensicherheit

Die Datensicherheit hat das primäre, technische Ziel, Daten jeglicher Art gegen Manipulation, Verlust, unberechtigte Kenntnisnahme und andere Bedrohungen zu sichern.

Hierunter fallen damit auch reine Unternehmensdaten, also Daten von juristischen Personen.
Das oberste Ziel der Datensicherheit besteht in der Gewährleistung der:

- **Vertraulichkeit:** Nur autorisierte Personen können auf die Daten zugreifen.
- **Integrität:** Daten können nicht unbemerkt verändert werden.
- **Verfügbarkeit:** Der technische Zugriff auf Daten wird gewährleistet.

Vereinfacht könnte man sagen, dass es sich hier um die praktischen Sicherheitsmaßnahmen oder Ansätze zum Schutz von Daten handelt (z.B. Maßnahmen zur Datensicherung, technischer Schutz vor Datenverlust usw.).

Maßnahmen zur Gewährleistung der Datensicherheit



Technische Maßnahmen:

- Verschlüsselung, Firewalls, Zugriffsrechte, etc.
→ verhindern nicht-authentifizierte Zugriffe auf Systeme beziehungsweise auf die Systemarchitektur.
- Backupsystem gegen Datenverlust



Organisatorische Maßnahmen:

- Standards und Richtlinien
- Schulung der Anwender
→ Verhaltensweisen etablieren, die innerhalb einer Organisation Datensicherheit fördern.



Physische Maßnahmen:

- Einschränkung des Zugriffs auf physische Datenträger einrichten.

Was ist Datenschutz?

- Unter Datenschutz versteht man den Schutz von personenbezogenen Daten. Hierunter fallen alle Daten, die sich auf eine natürliche Person beziehen.
- Ziel des Datenschutzes ist der Schutz des allgemeinen Persönlichkeitsrechts der betroffenen natürlichen Personen.
- Normen hierzu finden sich in der Datenschutzgrundverordnung (DSGVO) und dem Bundesdatenschutzgesetz (BDSG).
- Der Datenschutz dient somit dem Zweck, natürliche Personen und ihre Grundrechte und Grundfreiheiten zu schützen.

Wie ist der Datenschutz geregelt?

Das Recht auf den Schutz persönlicher Daten ist gesetzlich verankert. Jeder kann selbst darüber entscheiden, was mit seinen personenbezogenen Daten geschieht und ob diese verarbeitet werden dürfen oder nicht. Dieses Recht auf informationelle Selbstbestimmung ist eine besondere Ausprägung des allgemeinen Persönlichkeitsrechts.

Die für den Datenschutz relevanten Bestimmungen sind:

- DSGVO (Datenschutzgrundverordnung, EU-weit)
- BDSG (Bundesdatenschutzgesetz)
- ePrivacy-Verordnung



Der Datenschutz ist meist auf nationaler Ebene geregelt und in seiner Ausprägung sehr unterschiedlich. So wird auch die DSGVO länderspezifisch umgesetzt.

Der Datenschutzbeauftragte

Die Datenschutzbeauftragten sind dafür zuständig, die Verantwortlichen dabei zu unterstützen, die Vorgaben der Datenschutzgesetze zu erfüllen. Es wird dabei zwischen zwei Arten von Datenschutzbeauftragten unterschieden:

- **behördliche Datenschutzbeauftragte** (zum Beispiel der Landesdatenschutzbeauftragte, der für die öffentlichen Stellen des jeweiligen Bundeslandes zuständig ist, oder der Bundesdatenschutzbeauftragte)
- **betriebliche Datenschutzbeauftragte** (ein Angestellter eines Unternehmens oder ein externer Datenschutzbeauftragter, der für den betrieblichen Datenschutz zuständig ist).

Wann ein Unternehmen einen Datenschutzbeauftragten (DSB) bestellen muss, hängt von verschiedenen Kriterien ab und lässt sich nicht pauschal beantworten. Eine Rolle spielen sowohl die **Anzahl der Personen** in einem Unternehmen, welche **regelmäßig mit persönlichen Daten arbeiten**, als auch die **Kategorien der persönlichen Daten** sowie bestimmte **Kerntätigkeiten des Unternehmens**. Genaue Bedingungen, die erfüllt sein müssen, finden sich in der DSGVO.

Welches sind die Kernpunkte der DSGVO?

Verarbeitung personenbezogener Daten

Die DSGVO befasst sich mit Grundsätzen für die Verarbeitung personenbezogener Daten:

- **Datenminimierung:** Es sollen nur so viele Daten verarbeitet werden, wie notwendig.
- **Speicherbegrenzung:** Personenbezogene Daten dürfen nur so lange gespeichert werden, wie es für die Zwecke, für die sie verarbeitet werden, erforderlich ist.
- **Zweckbindung:** personenbezogene Daten dürfen nur für festgelegte, eindeutige und legitime Zwecke erhoben werden.
- **Richtigkeit:** Die verarbeiteten Daten müssen korrekt sein.
- **Rechenschaftspflicht:** Der Verantwortliche muss die Einhaltung der Grundsätze nachweisen können.
- **Rechtmäßigkeit:** Die Daten wurden auf rechtmäßige und für die betroffene Person nachvollziehbare Weise verarbeitet.

Entwicklungsvorgaben für Datenverarbeitung

Privacy by Design und Privacy by Default

Datenschutz muss integraler Bestandteil der Entwicklung von Produkten, Diensten oder Anwendungen sein.

Maximaler Datenschutz muss „serienmäßig“ sein und nicht mehr die Option, die der Betroffene aktiv anwählen muss. Wichtig ist, dass insbesondere die „Privacy by Design“-Anforderungen die verantwortlichen Stellen treffen und nicht die Hersteller von Produkten, Diensten und Anwendungen.

**Datenschutz durch
Technikgestaltung**
(Privacy by Design)

Die Technikgestaltung orientiert sich in allen Bereichen an den Datenschutzanforderungen und berücksichtigt diese von Anfang an.

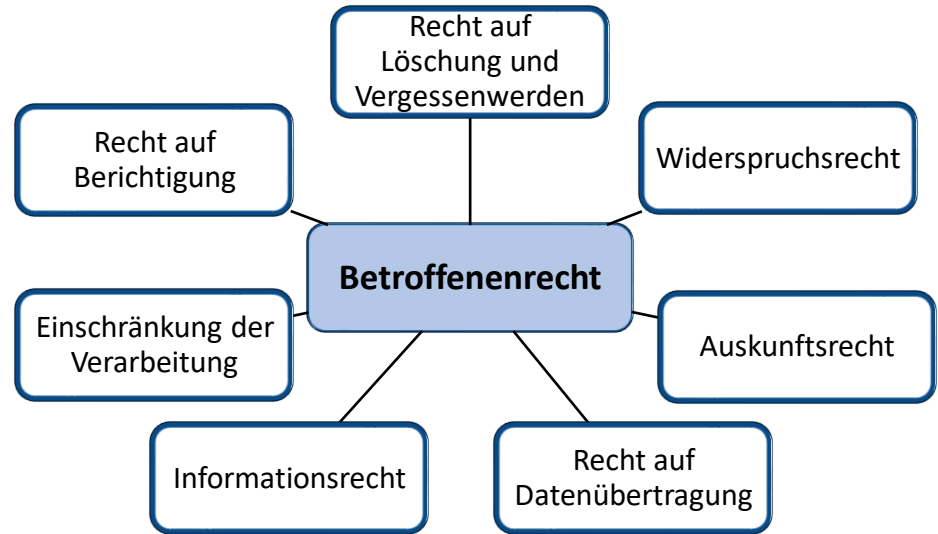
**Datenschutz durch
datenschutzfreundliche
Voreinstellungen**
(Privacy by Default)

Software, Hardware und Services müssen bei Auslieferung datenschutzfreundlich voreingestellt sein. Datenschutz ist keine Option, die der Betroffene aktiv anwählen muss.

Betroffenenrechte

Die Verarbeitung personenbezogener Daten ist grundsätzlich verboten, außer die Betroffenen stimmen der Nutzung ihrer Daten ausdrücklich zu.

Damit jeder den Umgang mit seinen personenbezogenen Daten kontrollieren und steuern kann, gibt es die Betroffenenrechte der DSGVO.



Die „Rechte der betroffenen Person“ werden in Kapitel III der DSGVO beschrieben:

(<https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679&from=DE> ab S. 39).



Diskussion:

- Welchen Einfluss hat Datensicherheit und -schutz auf Ihre Tätigkeit?
- Welche Mittel und Maßnahmen setzen Sie aktiv ein, um Datensicherheit und -schutz zu gewährleisten bzw. zu verbessern?

The background of the slide features a blue-toned image of a hand reaching out towards a wireframe globe. The globe is composed of a network of white lines and dots, symbolizing global connectivity and data. The hand is positioned in the lower-left foreground, with fingers slightly curled as if about to touch the globe.

Datenexport

Worauf sollte beim Datenexport
geachtet werden?

Datenexport vorbereiten

Beim Datenexport stellen sich ähnlich (aber umgekehrt) wie beim Datenimport folgende Fragen:

- In welcher Form liegen die Daten am Ende des Workflows vor?
- Wofür dienen die Daten im Zielsystem?
- Welches Format bzw. welchen Datentyp bearbeitet das Zielsystem?
- Welche Schnittstellen hat das Zielsystem?

Die Antwort auf die erste Frage ist insofern relevant, dass sie den Ausgangspunkt definiert, von dem aus eventuell weitere Bearbeitungsschritte notwendig sind.

Die übrigen drei Fragen helfen dabei, das richtige Datenformat und ein geeignetes Übertragungsformat zu finden.

Reporting und Berichte

Sollen die Daten als abgeschlossener Bericht (z.B. Journale, Ergebnisberichte, Performanceübersichten, etc.) verteilt werden, wird das Format in der Regel von der Art des Berichtes und der benötigten Informationen bestimmt:

- Tabellarische Auflistung von Kennwerten
- Charts und Grafiken
- Bilder, Karten, Infoboxen, etc.

Bei periodisch erstellen Berichten empfiehlt es sich, das Format so zu wählen, dass Informationen themenbezogen an den gleichen Stellen zu finden sind bzw. eine feste Struktur für den Bericht zu wählen.

Reporting und Berichte

Sollen die Daten nur fix dargestellt werden oder sollen sie noch bearbeitbar sein?

Für den ersten Fall eignet sich ein unveränderliches Format wie pdf oder html. Die Berichte werden als nur-lesbare Dokumente erstellt und können vom Nutzer nicht mehr verändert werden.

Wenn der Bericht für den Nutzer noch bearbeitbar sein soll, eignet sich ein Format für Tabellenkalkulationsprogramms wie etwa Excel als Datentyp. Diese Formate sind sehr variabel gestaltbar und erfordern neben dem Export der Daten noch Informationen zur Tabellengestaltung (z.B. Zeilen- und Spaltengröße, Farben, Textformatierungen, etc.)

Datenschnittstellen und Datenbanken

Sind die zu exportierenden Daten Quell- oder Rohdaten für das Zielsystem, z.B. für weiterführende Analysen und Visualisierungen in BI-Anwendungen, sollte auch ein leicht übertragbares und leicht verarbeitbares Format gewählt werden.

Arbeitet das Zielsystem mit einer Datenbank, so sollten nach Möglichkeit die Daten direkt in diese geschrieben werden.

Dies ist allerdings nicht immer möglich, sei es aus technischen oder aber auch sicherheitsrelevanten Gründen.

Häufig werden daher Daten auch als formatierte Text-Dateien übertragen, für die es in den meisten Systemen eine Schnittstelle gibt.

Datenvolumen – Größe von Tabellen und Dateien

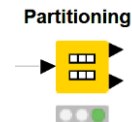
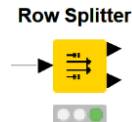
Sollen die Daten in einer einzigen Tabelle oder Datei übertragen bzw. exportiert werden?

Bei der Bearbeitung der Daten wurde mitunter ein erheblicher Aufwand betrieben, die Daten in ein einheitliches Format zu übertragen und in einer Tabelle zu vereinen. So erscheint es auch logisch, dass man dieses Format beibehält und an das Zielsystem weitergibt.

- Auch hier sollte man aber prüfen, wie die exportierten Daten genutzt werden sollen und welche Kapazitäten dem Zielsystem bzw. bei der Übertragung zur Verfügung stehen.

Eine Tabelle mit beispielsweise vielen Bild-Dateien kann hier zu einer großen Herausforderung werden und mitunter das Zielsystem stark beeinflussen oder sogar blockieren. Bei großen Datensätzen besteht zudem die Gefahr, dass sie bei der Übertragung beschädigt werden.

Es sollten also Maßnahmen ergriffen werden, um effizienten und sicheren Datentransfer zu gewährleisten. Eine Methode wäre z.B. die Daten zu partitionieren und paketweise zu übertragen.



Übung Datenexport

The background of the slide features a stylized, wireframe globe of Europe and Africa, composed of white lines and dots. A hand is visible at the bottom left, reaching towards the globe. The overall color scheme is blue and white.

Dokumentation und Workflow- organisation

Warum ist Dokumentation so wichtig
und wie sollte sie gestaltet werden?

Jeder Schritt muss verständlich beschrieben sein!

- Was wird gemacht?
- Wie soll es umgesetzt werden?
- Welchem Zweck soll es dienen?

Neben der Dokumentation im Skript und einer Übersicht von Programmen mit Funktion und Pfaden, ist es grundsätzlich empfehlenswert, ein Arbeitsbuch zu führen (wenn möglich innerhalb einer Teamwork-Umgebung).

Dort sollten Aufgaben, Notizen, Gedanken und Ideen festgehalten werden.

Dokumentation in visuellen Analytics


Anwendungen

Workflow Metainformation

Die grundlegendste Form der Dokumentation sind die begleitenden Informationstexte zu einem Workflow. Diese sind in der Regel als „Beschreibung“, „Inhalt“ oder „Workflow-Information“ im Menü des Workflows abrufbar.

Je nach Gestaltung können hier alle wichtigen Informationen beschrieben und dokumentiert werden.

- Klick auf Workflow im Explorer
- In Description Fenster editieren

02_EDA 

Title 03_EDA

Description
Workflow zur Übung der explorativen Datenanalyse für den Kurs Data Analyst

Tags
EDA Data Analyst

Links
No links have been added yet.

Creation Date 2020-12-21

Author Jörg Endter

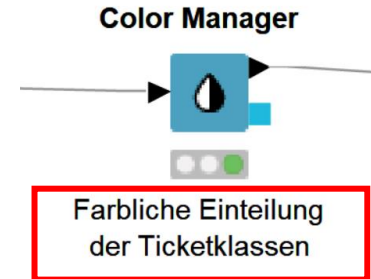
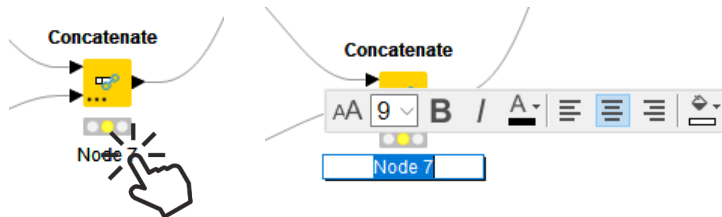
Beschreibung des Workflows in einer Infobox

Dokumentation in visuellen Analytics Anwendungen

Beschreibung der Knoten

Jeder Bearbeitungsschritt sollte eine kurze Information enthalten, was in ihm durchgeführt wird und welches Ziel erreicht werden soll.

In KNIME kann man dafür die Textboxen unterhalb der Knoten verwenden.

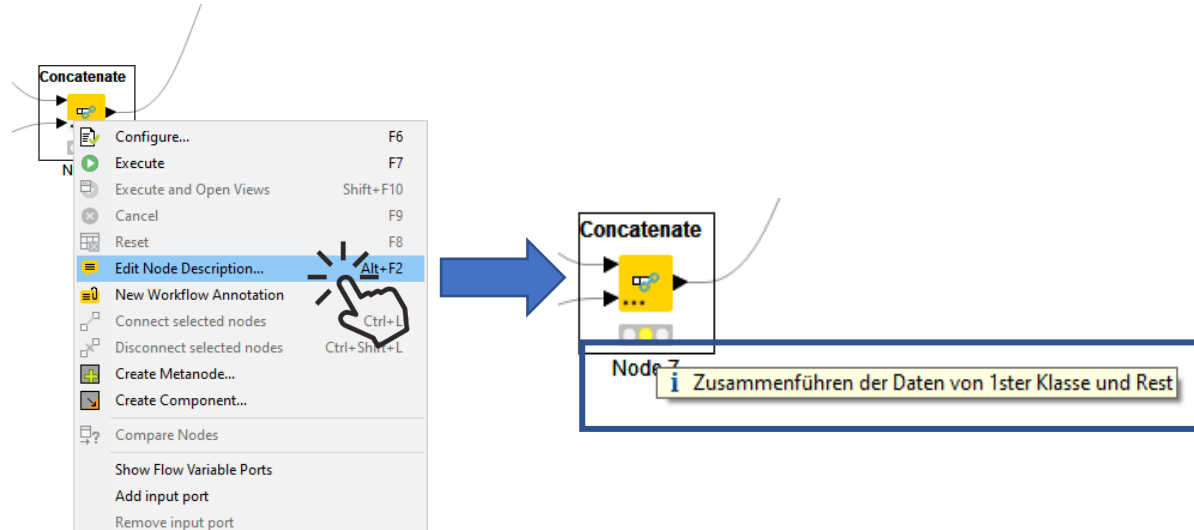


Funktionsbeschreibungen von Knoten

Dokumentation in visuellen Analytics Anwendungen

Zusätzliche Knoteneschreibungen mit MouseOver Funktion

- Rechtsklick auf Knoten
- Edit Node Description



Dokumentation in visuellen Analytics Anwendungen

Notizen und Beschreibungen zu Knotengruppen

Knoten lassen sich meistens in einem Vorgang zusammenfassen. Beispielsweise das Laden aus Datenquellen, die Bereinigung der Rohdaten, das Erlernen eines Modells, etc. Dies geschieht durch eine Gruppe von abgrenzbaren Knoten.

Die Oberfläche des Workflows ermöglicht es, dies auch durch grafische Mittel darzustellen. Die jeweiligen Nodes sollten sichtbar zusammenstehen und von weiteren Prozessen abgesetzt sein.

Beschreibungen und Markierungen heben die Funktion hervor und unterstützen die Dokumentation.



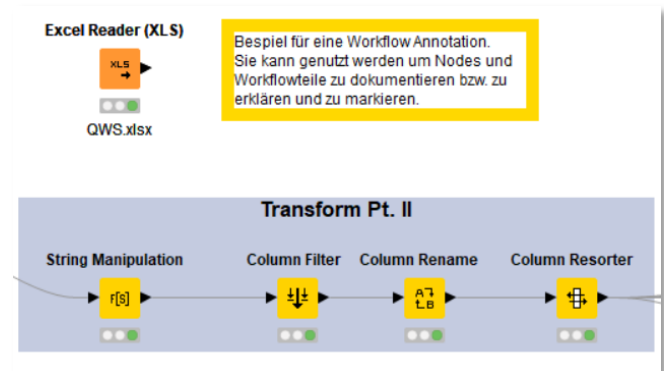
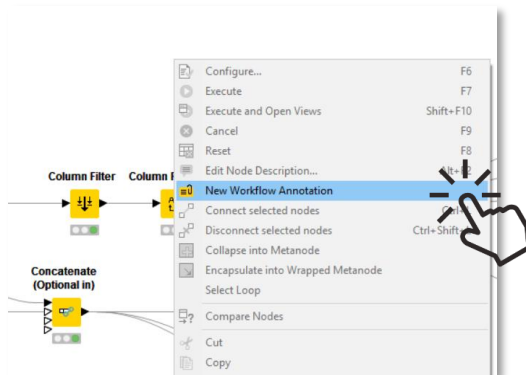
Dokumentation im Workflow:
Farbliche Markierung der Abschnitte und
Beschreibung der Aktionen

Dokumentation in visuellen Analytics Anwendungen

Notizen und Beschreibungen zu Knotengruppen

„Workflow Annotation“

- Rechtsklick im Editor
- „New Workflow Annotation“



Workflow-Organisation

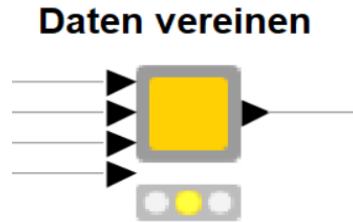
Die Gestaltung der Arbeitsfläche in einem Workflow hat einen wesentlichen Einfluss auf die Verständlichkeit des Datenprozesses sowie die Effizienz mit ihm zu arbeiten und ihn zu verstehen.

Ein paar wichtige Regeln und Maßnahmen helfen einen Workflow übersichtlich und effizient zu gestalten:

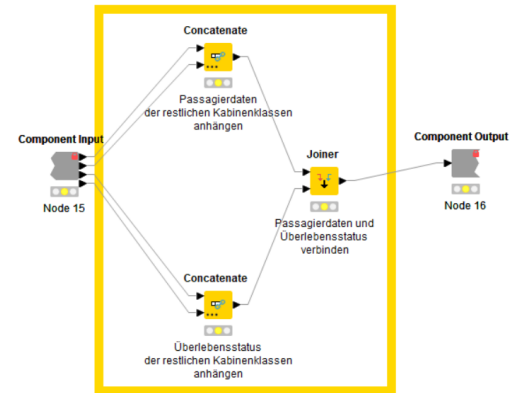
1. Strukturieren Sie ihren Workflow in klar abgrenzbare Gruppen nach Funktion
2. Halten Sie die Flussrichtung bei, um deutlich Anfang und Ende darzustellen
3. Lassen sie die Datenverbindungen parallel laufen und vermeiden Sie Überschneidungen
4. Wenn die Anwendung es erlaubt, nutzen Sie die Möglichkeit, funktionelle Gruppen in Meta-Nodes und Components zusammenzufassen und somit die Zahl der Nodes zu reduzieren

Meta-Nodes und Komponenten

Die Möglichkeit, Meta-Nodes und Komponenten zu erstellen, hat eine wichtige Aufgabe in Workflows. Hiermit können funktionelle Gruppen modular eingesetzt und verwaltet werden. Einige Anwendungen geben diesen Modulen weitere Funktionen, die sie wie kleine Programme im Workflow handhaben lassen.

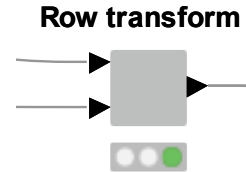


Die Komponente „Daten vereinen“ besteht aus 3 Knoten, die die verschiedenen Eingangsdaten zu einer Tabelle zusammenfassen.



Workflow Organisation

Aufräumen des Workflows und verpacken von Knoten

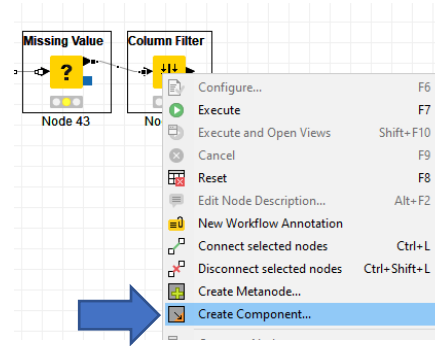


Metanode	Komponente (Component)
Einen Teil des Workflows zusammenfassen	Einen Teil des Workflows zusammenfassen
	Hat alle Eigenschaften eines regulären Knotens: Konfigurationsmenü, Beschreibung, Views, interaktive Elemente
	Kann als Vorlage gespeichert und geteilt werden
Anzeige des Inhalts per Doppelklick	Anzeige des Inhalts mit Strg+Doppelklick

Workflow Organisation

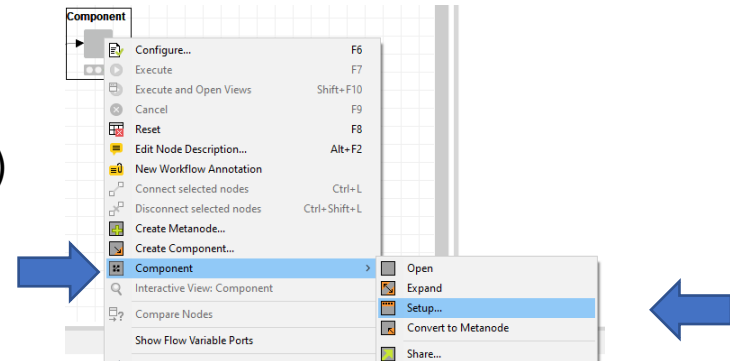
Erstellen von Komponenten:

- Knoten markieren,
- Rechtsklick auf markierten Knoten
→ Create Component



Ändern von Komponenten:

- Rechtsklick auf Komponente → Component
 - Setup (Ändern der Ports und des Namens)
 - Open (Knoten in der Komponente bearbeiten)
 - Expand (Komponente wieder entfalten)
 - Share (Speichern und Teilen)



Workflow Organisation

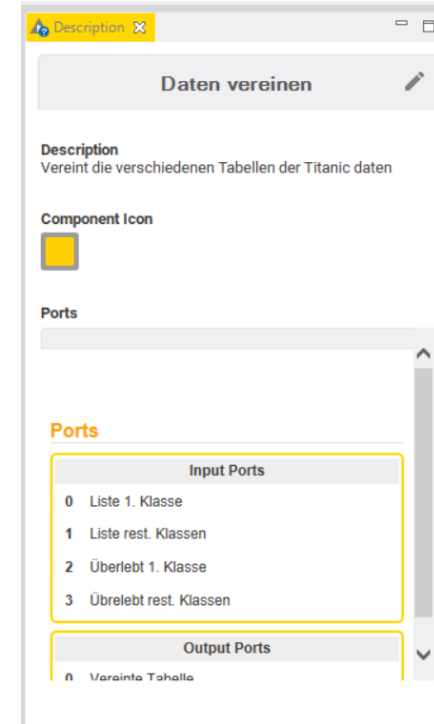
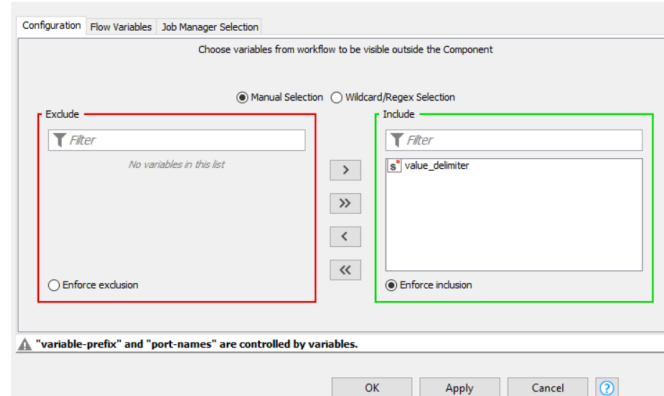
Beschreibung in „Component Description“:

Öffnen der Komponente

- Beschreibung der Komponente und Ports ändern

Doppelklick auf Anfang/Endknoten im Inneren

- Konfigurieren der Übergabe“ von Flow Variables an den Workflow



Praxis-Tipp: Workflow Organisation

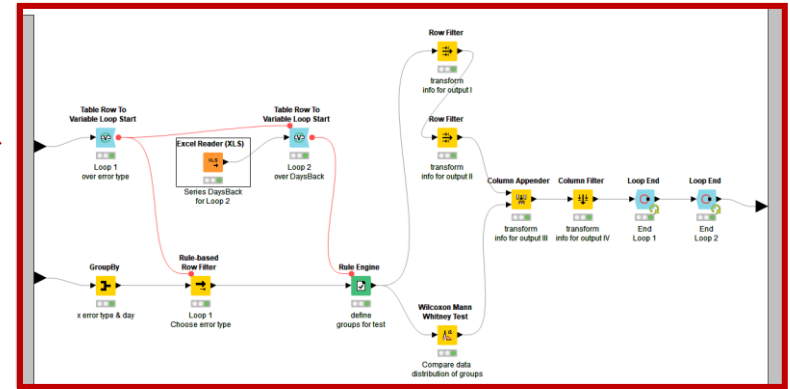
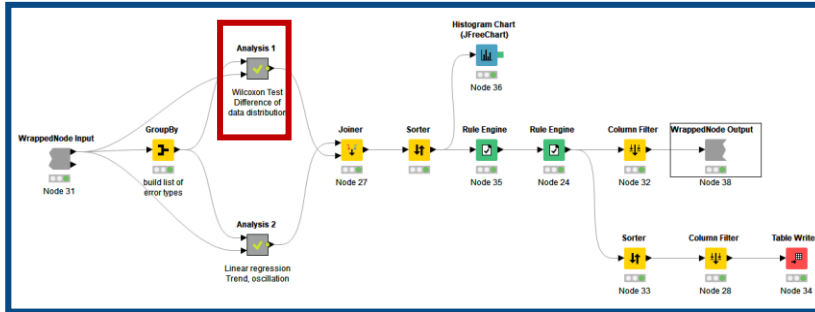
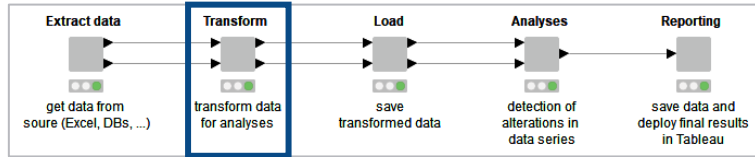
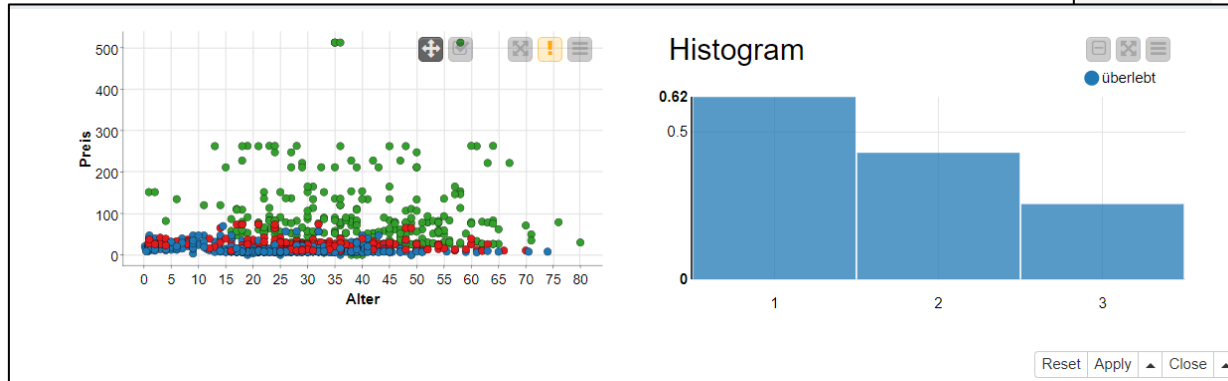
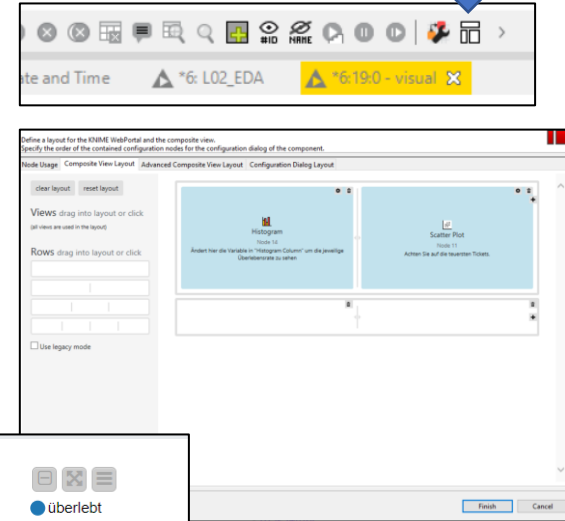


Diagramme verbinden

- Diagramme, die in einer Komponente verbunden werden, können gemeinsam angezeigt werden.
- Filter und Markierungen wirken sich auf alle Plots aus.



Konfiguration



Übung Dokumentation und Workflow Kontrolle

The background of the slide is a futuristic, blue-toned image. It features the silhouettes of four people standing in a row, facing right. Overlaid on these silhouettes and the background are various digital elements: glowing lines, circular patterns, and semi-transparent rectangular boxes containing text and icons. One box in the center lists business functions: Administration, Human Resources, Legal, Accounting, Finance, Marketing, Publicity, Promotion, Research, Business, Hospitality, Engineering, and Management. Another box on the left lists: Administration, Human Resources, Legal, Accounting, Finance, Marketing, Publicity, Promotion, Research, Business, Hospitality, Engineering, and Management. The overall aesthetic is high-tech and digital.

Vielen Dank!