Grundlagen der KI

Apps-Analyse mit KNIME

Moritz Dick,

Thaddäus Friedel

Gliederung

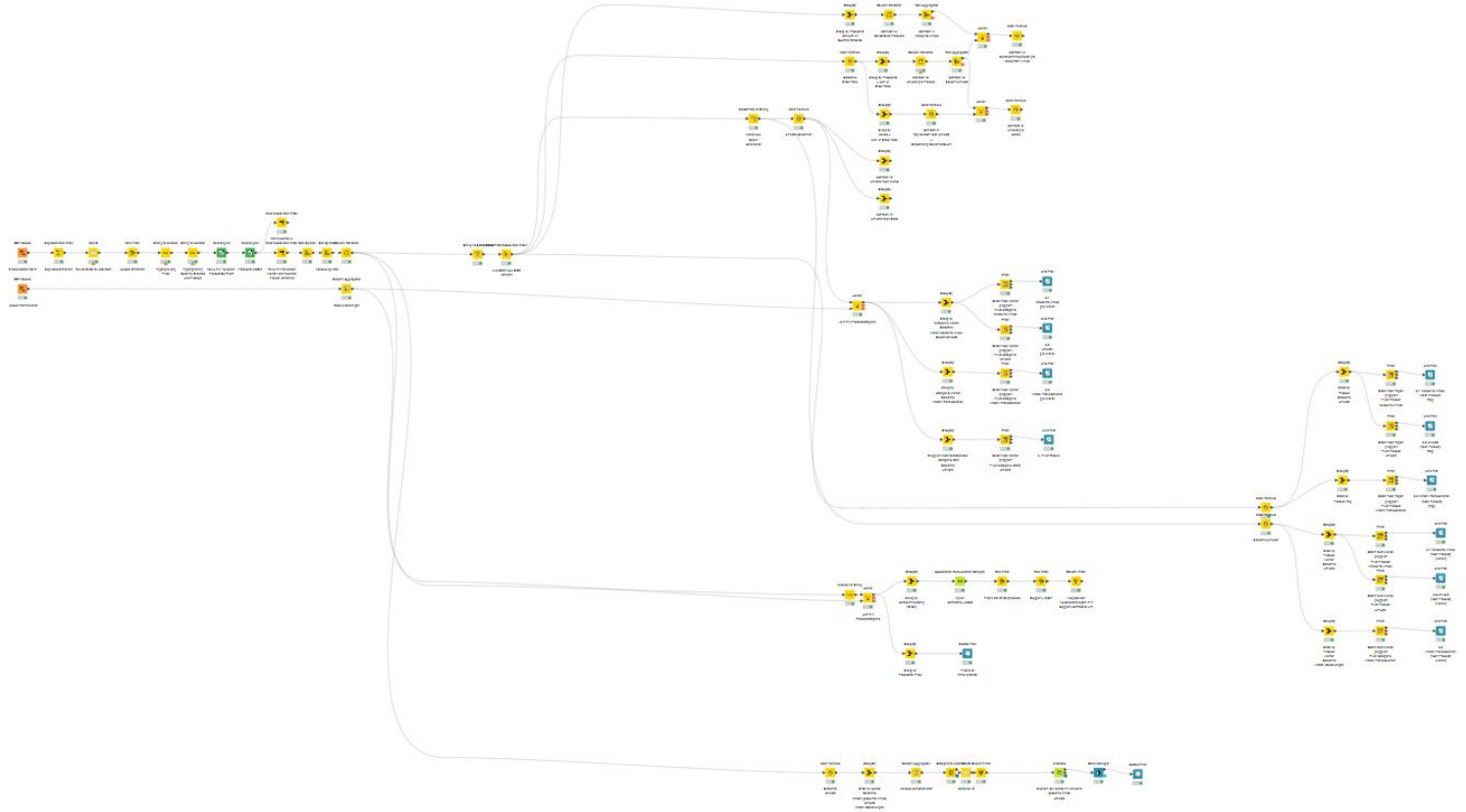
- Projektüberblick
- Datensatz
- Kennzahlen
- Visualisierung & Pivot Tabellen
- Apriori Algorithmus
- Clustern der Kunden

Projektüberblick

- Analyse zu Geschäftsdaten eines Handels für Elektroartikel
  - Gewinn von Informationen zur Optimierung der Verkaufsstrategien

- Verwendung von KNIME zur Datenverarbeitung
  - Nodes können in Workflows verbunden werden
    - Inputs und Outputs der Nodes werden verknüpft
  - Ermöglicht schrittweise Verarbeitung von Daten
  - Visuelle Darstellung des Datenflusses
    - Übersichtliche Benutzeroberfläche

KNIME Workflow



Datensatz

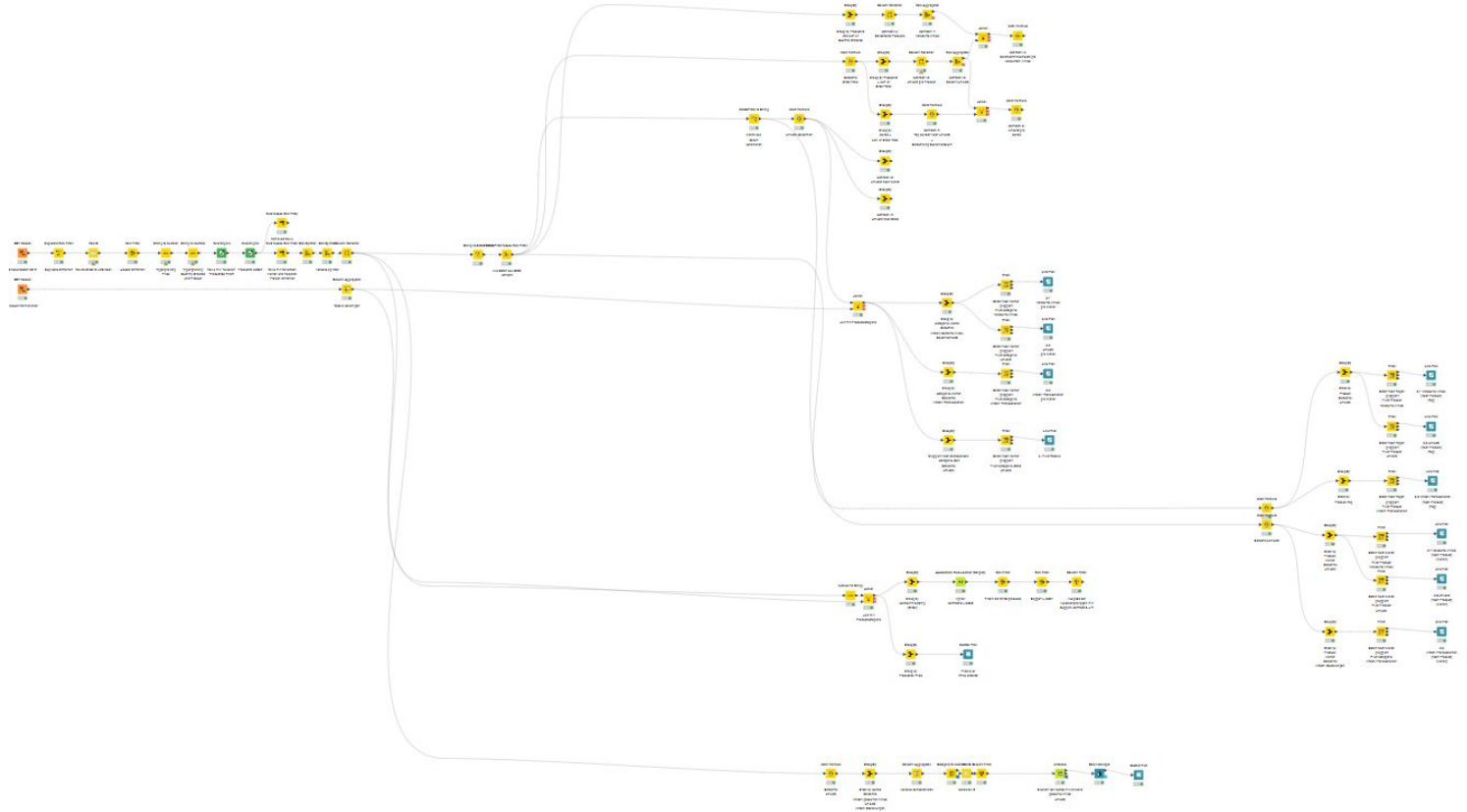
- Gegebene Datensätze
  - 12 sales.csv-Dateien mit Verkaufsdaten (Insgesamt ca 186.000 Datensätze)

Row ID	Order ID	Product	Quantit...	Price Each	Order Date	Purchase Address
Row514	177052	19	2	11.95	04/02/19 09:30	532 Walnut St, San Franci...
Row515	177053	21	1	11.99	04/24/19 20:45	5 Adams St, Boston, MA 0...
Row516	177054	8	1	150	04/09/19 19:18	800 Jackson St, Atlanta, G...
Row517	Order ID	Product	Quantity Or...	Price Each	Order Date	Purchase Address
Row518	177055	15	1	14.95	04/09/19 12:37	59 Forest St, Atlanta, GA ...

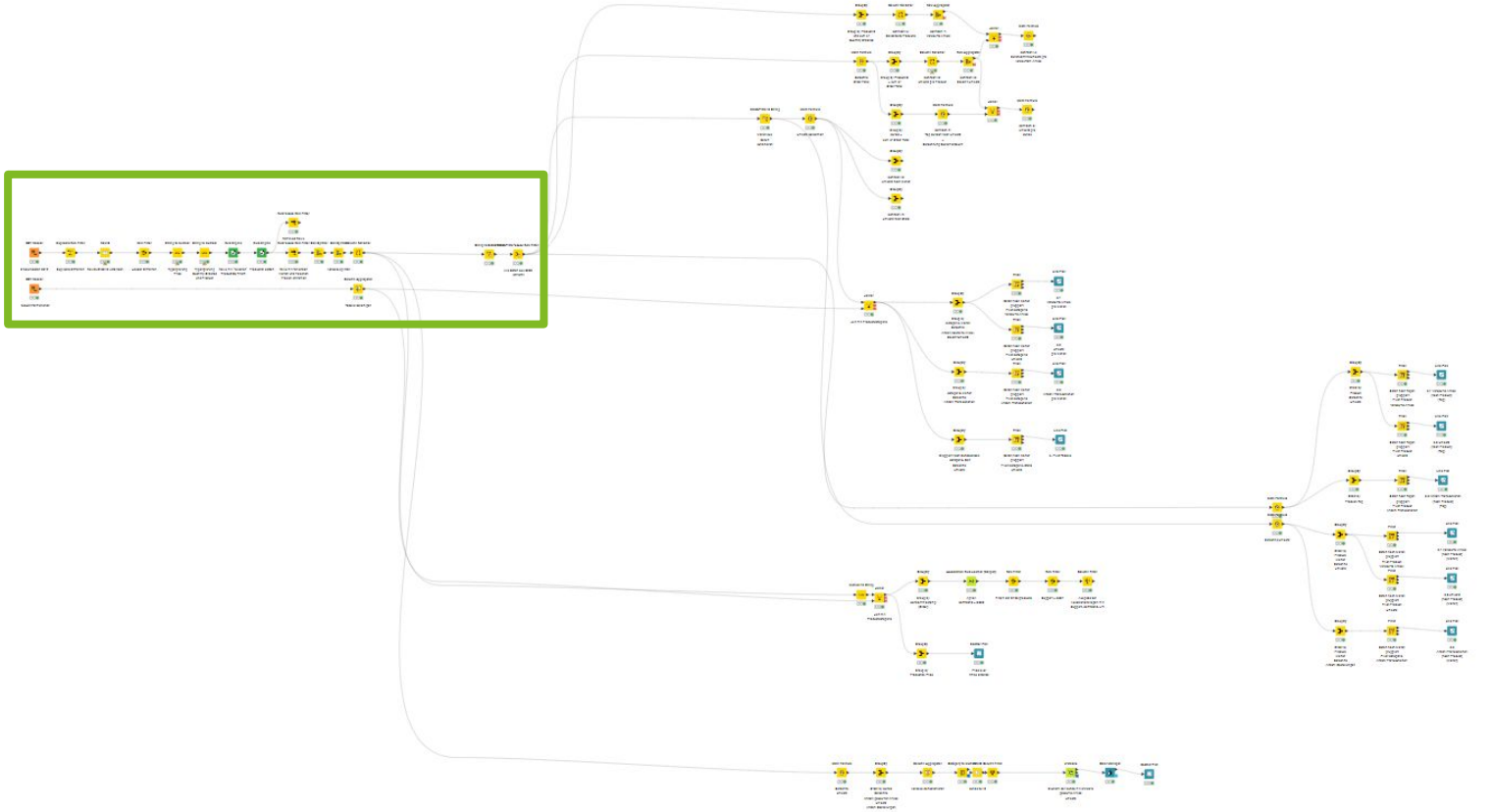
  - 1 products.csv-Datei mit Produktinformationen (21 unterschiedliche Produkte)

Row ID	produkt	Column1	Column2	Column3	produkt...
1	17in_Monitor	?	?	monitor	?
2	20in_Monitor	?	?	monitor	?
3	27in_4K_Ga...	?	monitor	?	?

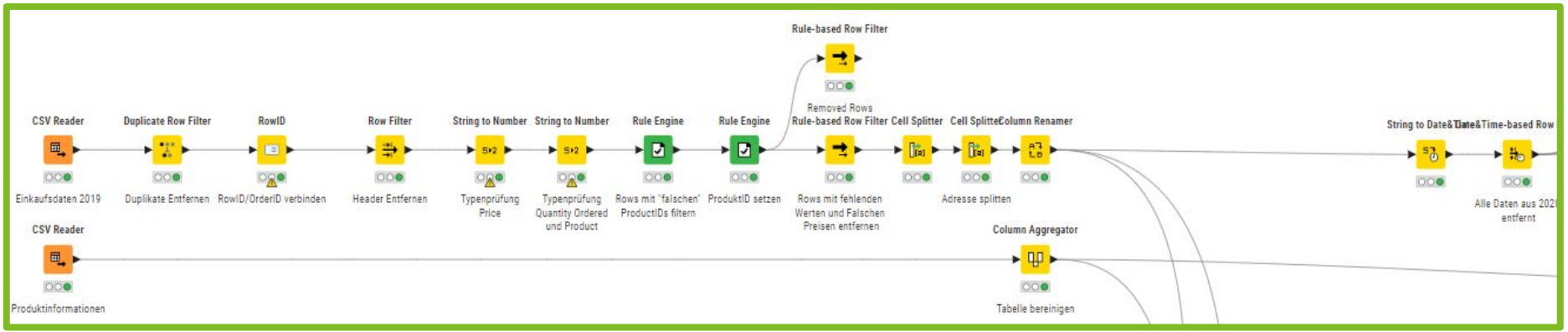
KNIME Workflow



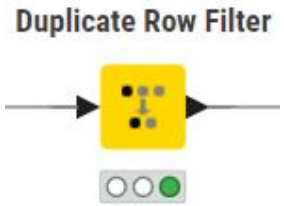
KNIME Workflow - Datensanierung



KNIME Workflow - Datensanierung

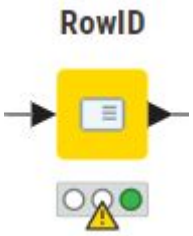


Datensanierung



- CSV Reader
  - Liest Datensätze aus einer CSV-Datei ein
  - Gibt Datensätze in Tabellen als Output weiter
- Duplicate Row Filter
  - Entfernt doppelte Datensätze
  - ca. 600

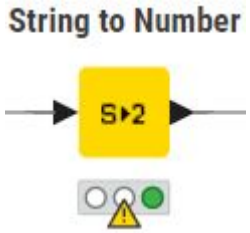
Datensanierung



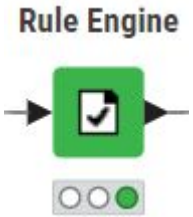
- RowID
  - Ersetzt RowID mit OrderID

- Row Filter
  - Entfernen der Header Zeile

Datensanierung

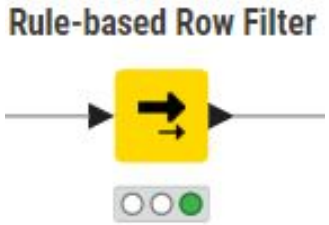


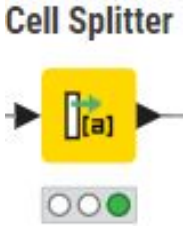
- String to Number
  - Umwandeln der Eingabedaten von Strings in Int bzw. Double
  - Typenprüfung



- Rule Engine
  - Definieren von Bedingungen um Datensätze mit falscher, bzw. fehlender ProduktID und/oder Preis zu berichtigen

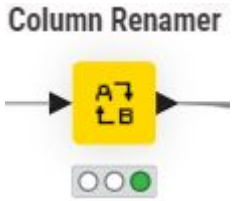
Datensanierung



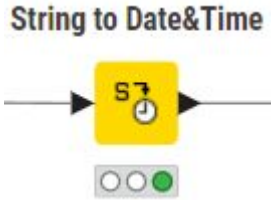


- Rule-based Row Filter
  - Definieren von Regeln, um übrig gebliebene, nicht verwertbare Datensätze zu filtern
  - 7 entfernte Datensätze
- Cell Splitter
  - Aufspalten der Adresse in ihre Bestandteile

Datensanierung

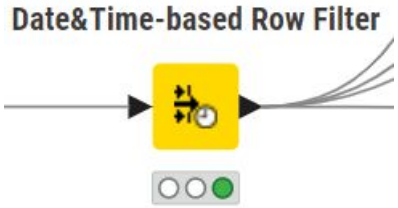


- Column Renamer
  - Umbenennen von Spalten

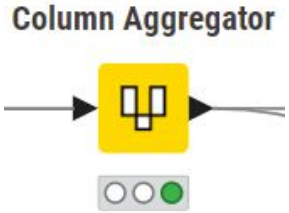


- String to Date&Time
  - Zeitstempel von String in Date Format umwandeln

Datensanierung



- Date&Time-based Row Filter
  - ca. 30 Datensätze aus 2020 entfernen, um Kennzahlen, die sich auf 2019 beziehen nicht zu verfälschen



- Column Aggregator
  - Spalten aus products.csv bereinigen

Datensanierung

RowID	Order ID *String*	Product ↓ *String*	Quantity Ordered *String*	Price Each *String*	Order Date *String*	Purchase Address *String*
Row...	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
Row...	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
Row...	Order ID	Product	Quantity Ordered	Price Each	Order Date	Purchase Address
Row...	236744	24	1	2.99	08/09/19 20:21	11 Wilson St, Atlanta, GA 30301
Row...	278836	22	1	11.99	11/26/19 20:48	989 12th St, New York City, NY 100...
Row3	176560	21	1	11.99	04/12/19 14:38	669 Spruce St, Los Angeles, CA 90...
Row4	176561	21	1	11.99	04/30/19 09:27	333 8th St, Los Angeles, CA 90001

RowID	Product *Number (integer)*	Quantity Ordered *Number (integer)*	Price Each *Number (double)*	Order Date *Local Date*	Street *String*	City *String*	State *String*	ZIP *Number (integer)*
1765...	19	2	11.95	2019-04-19	917 1st St	Dallas	TX	75001
1765...	9	1	99.99	2019-04-07	682 Chestnut St	Boston	MA	2215
1765...	11	1	600	2019-04-12	669 Spruce St	Los Angeles	CA	90001
1765...	21	1	11.99	2019-04-12	669 Spruce St	Los Angeles	CA	90001

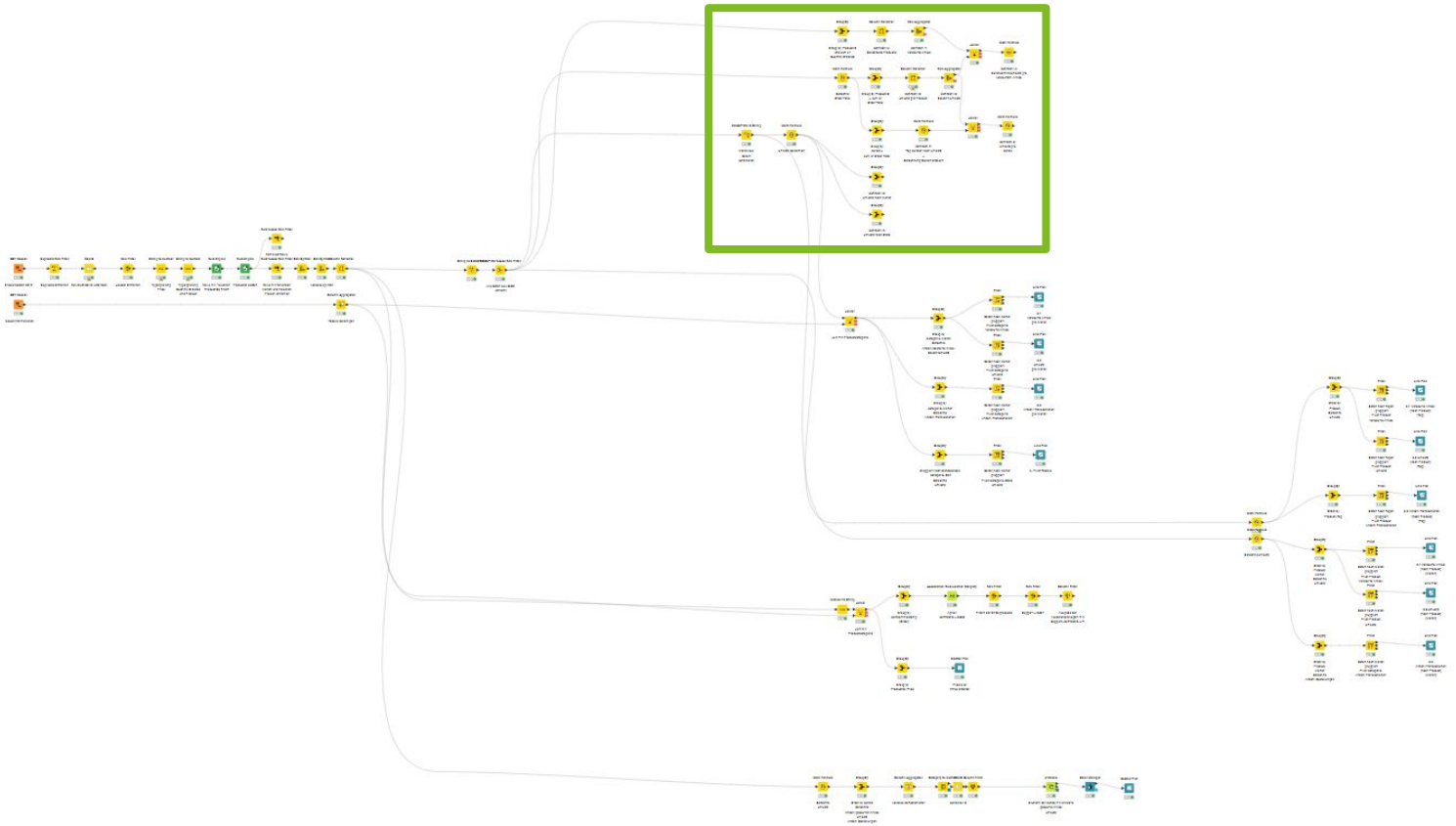
➢ Ca. 700 Fehlerhafte Datensätze entfernt bzw. angepasst

Datensanierung

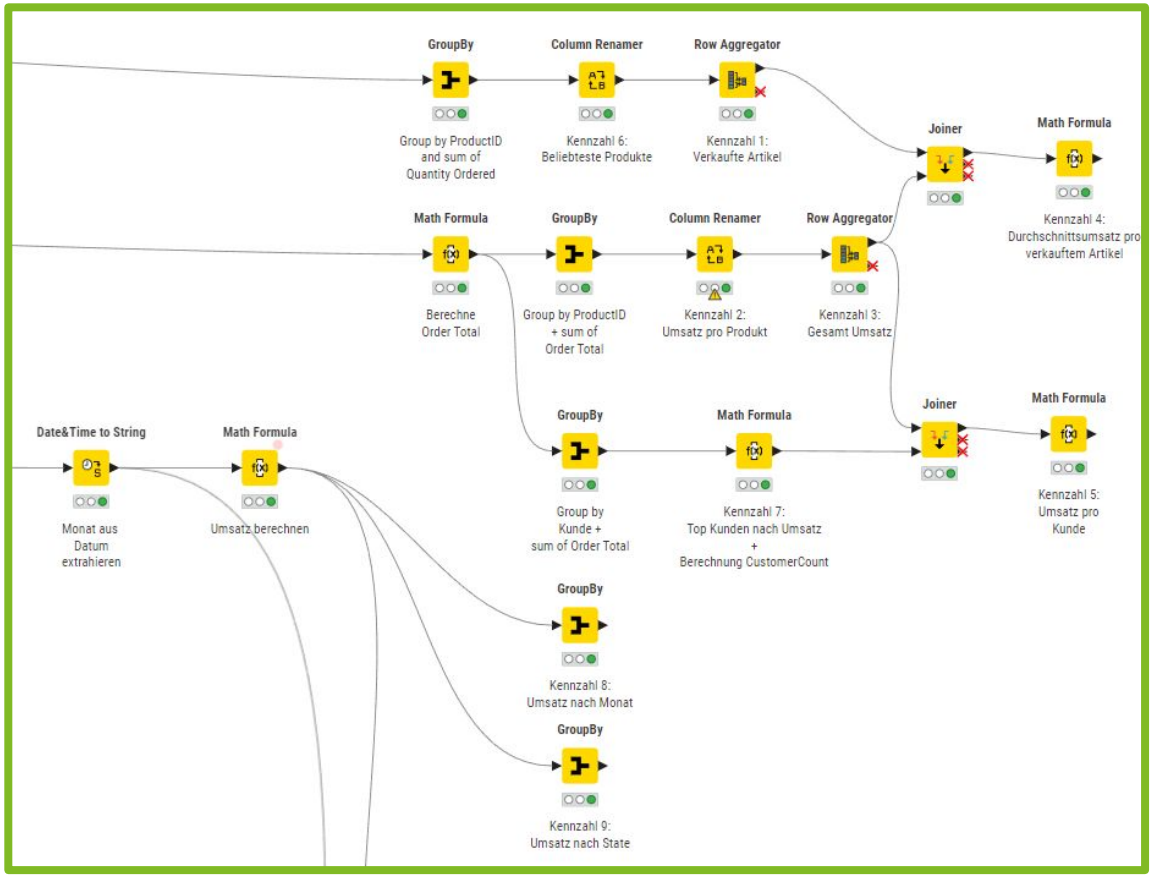
RowID	produkt String	Column1 String	Column2 String	Column3 String	produktkategory String
1	17in_Monitor	?	?	monitor	?
2	20in_Monitor	?	?	monitor	?
3	27in_4K_Gaming_Monitor	?	monitor	?	?
4	27in_FHD_Monitor	?	monitor	?	?
5	34in_Ultrawide_Monitor	?	monitor	?	?
6	AAA_Batteries_(4-pack)	?	battery	?	?

RowID	produkt String	produktkategory String
1	17in_Monitor	monitor
2	20in_Monitor	monitor
3	27in_4K_Gaming_Monitor	monitor
4	27in_FHD_Monitor	monitor
5	34in_Ultrawide_Monitor	monitor
6	AAA_Batteries_(4-pack)	battery

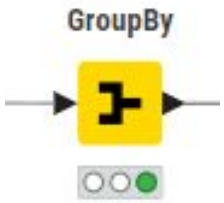
KNIME Workflow - Kennzahlen

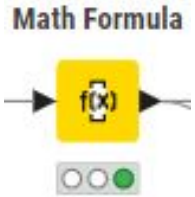


KNIME Workflow - Kennzahlen



Kennzahlen

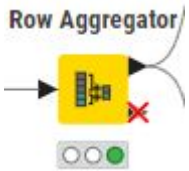




- GroupBy
  - Gruppieren der Datensätze nach Attribut/en
  - Aggregationsfunktionen wie Sum() oder Count() möglich
- Math Formula
  - Definition von Formeln zur Berechnung von Werten
  - z.B. Einzelpreis*Anzahl = Gesamtpreis

Kennzahlen

- Joiner
  - Ermöglicht verbinden zweier Datensätze



- Row Aggregator
  - Aggregationsfunktion durchführen

Kennzahlen

Verkaufte Artikel
- 208.672

Durchschnittlicher Umsatz pro Artikel
- 165,05

Gesamtumsatz
- 34.456.563,85

Durchschnittlicher Umsatz pro Kunde
- 244,78

Kennzahlen

Beliebteste Produkte

- AAA Batterien 4-Pack
  - 30977
- AA Batterien 4-Pack
  - 27615
- USB-C Ladekabel
  - 23926

Umsatzstärkste Produkte

- Macbook Pro
  - 8.030.800
- iPhone
  - 4.791.500
- ThinkPad Laptop
  - 4.125.958

Kennzahlen

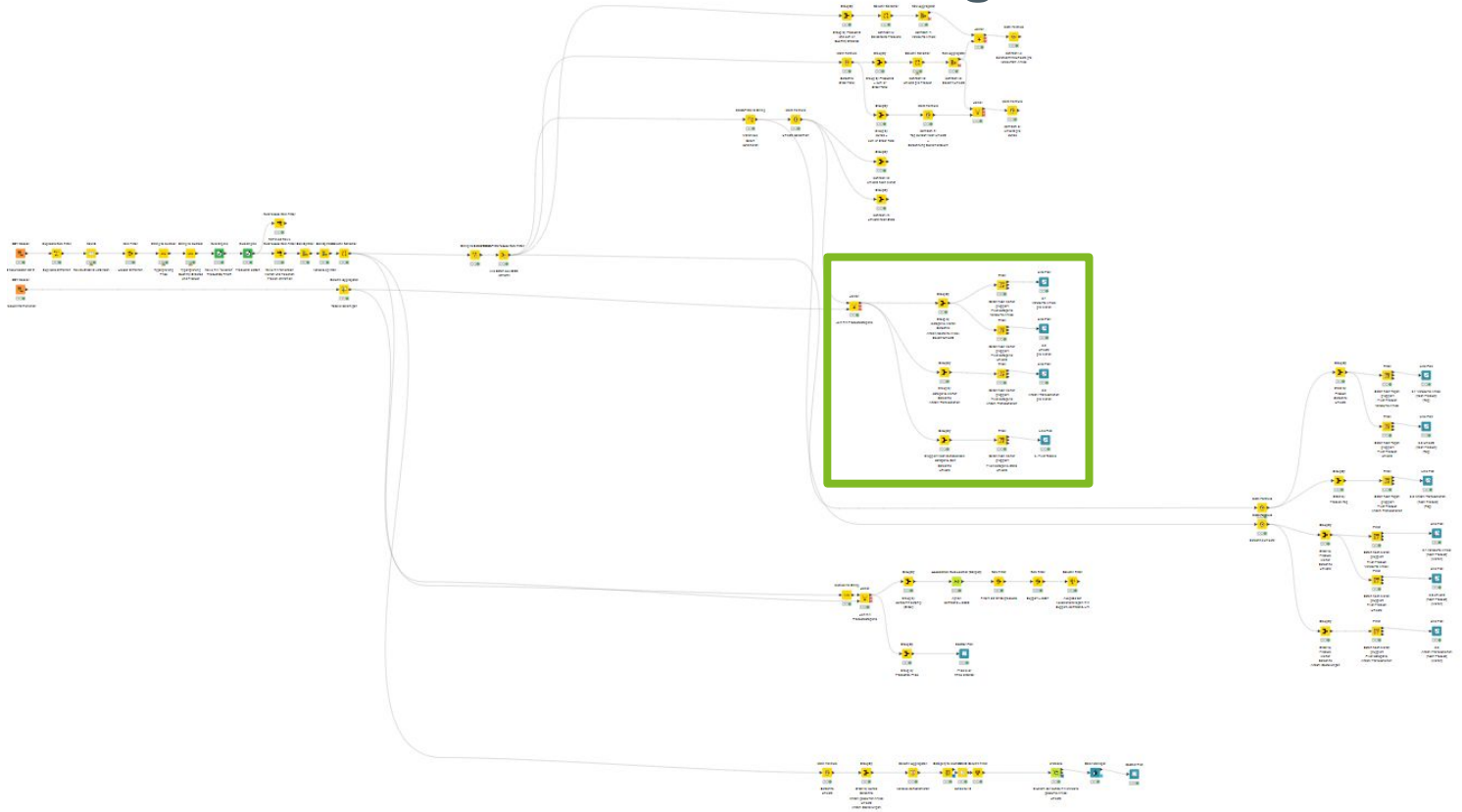
Umsatz nach US-Bundesstaat

State *String*	Sum(Price Order) *Number (double)*
CA	13,699,301.02
NY	4,660,502.6
TX	4,581,203.36
MA	3,657,300.76
GA	2,794,199.07
WA	2,744,878.09
OR	1,869,857.57
ME	449,321.38

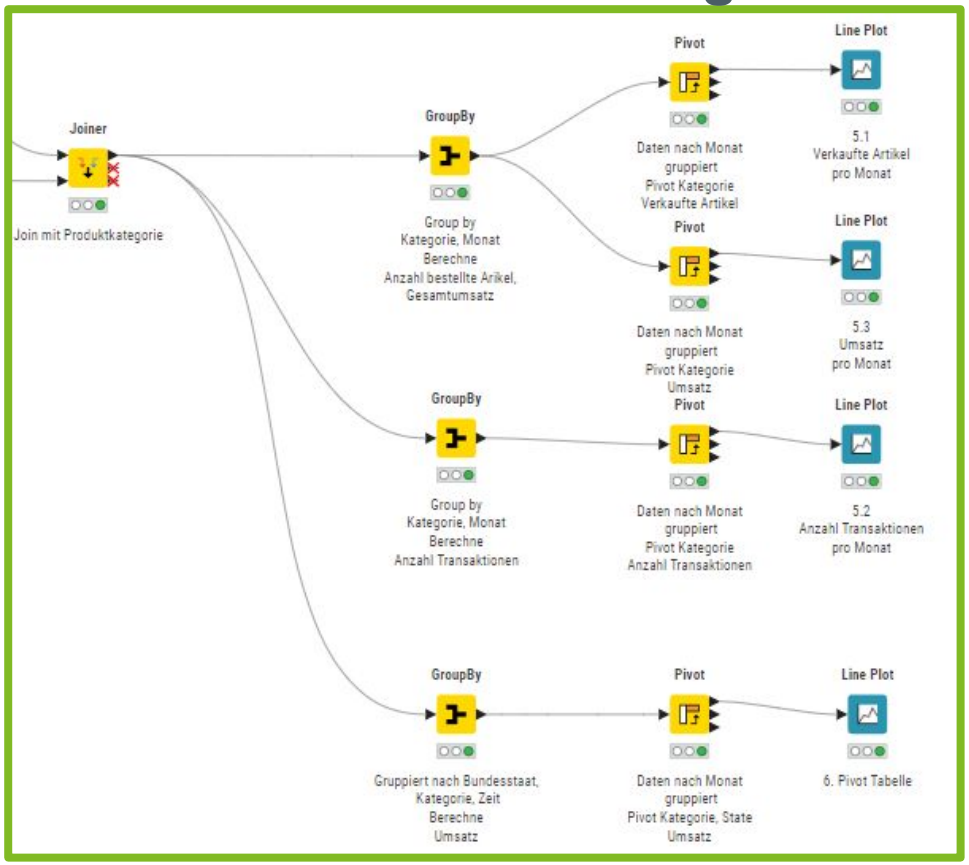
Umsatz nach Monat

Order D... *String*	Sum(Price Order) *Number (double)*
01	1,812,727.92
02	2,200,078.08
03	2,804,964.38
04	3,389,117.99
05	3,150,616.23
06	2,576,280.15
07	2,646,311.32
08	2,241,083.37
09	2,094,465.69
10	3,734,747.97
11	3,197,875.05
12	4,608,295.7

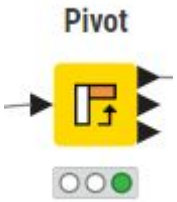
KNIME Workflow - Visualisierung & Pivot Tabellen



KNIME Workflow - Visualisierung & Pivot Tabellen



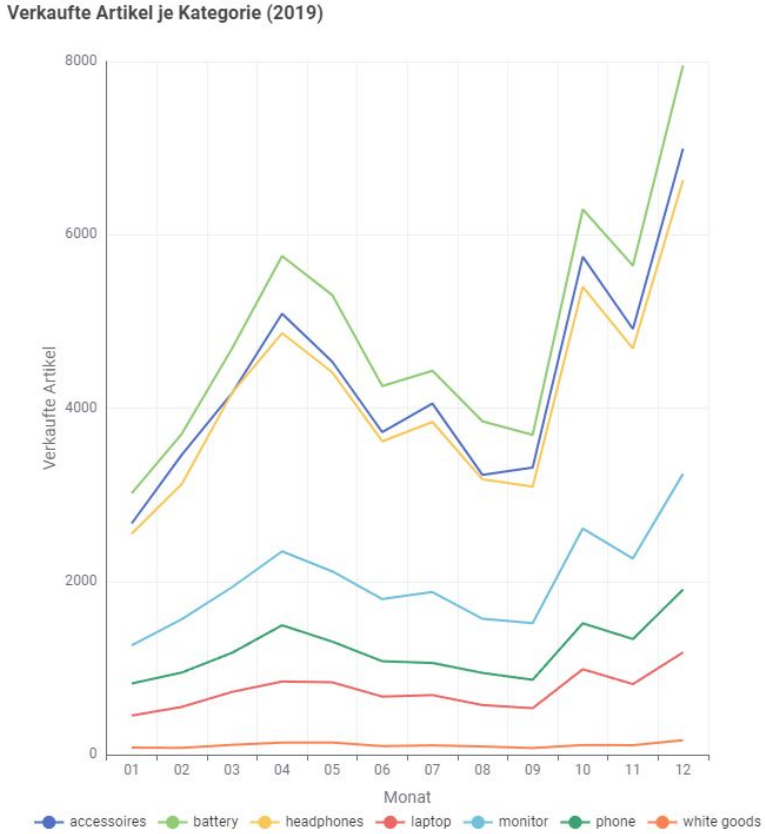
Visualisierung & Pivot Tabellen



- Pivot
  - Erlaubt andere Darstellung von Tabellen
  - Möglichkeit große Datenmenge in überschaubare Tabellen zu formen

- Line Plot
  - Ermöglicht Darstellung von Daten als Liniendiagramm

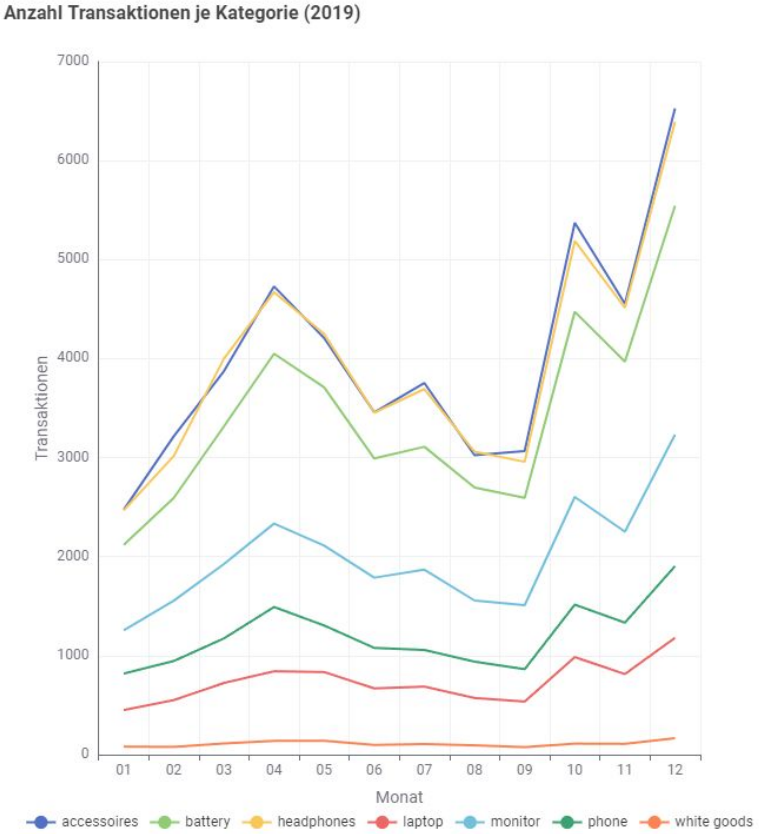
Visualisierung & Pivot Tabellen



- Produktkategorien aus denen mit Abstand am meisten Artikel verkauft wurden
  - Batterien
  - Accessoires
  - Kopfhörer

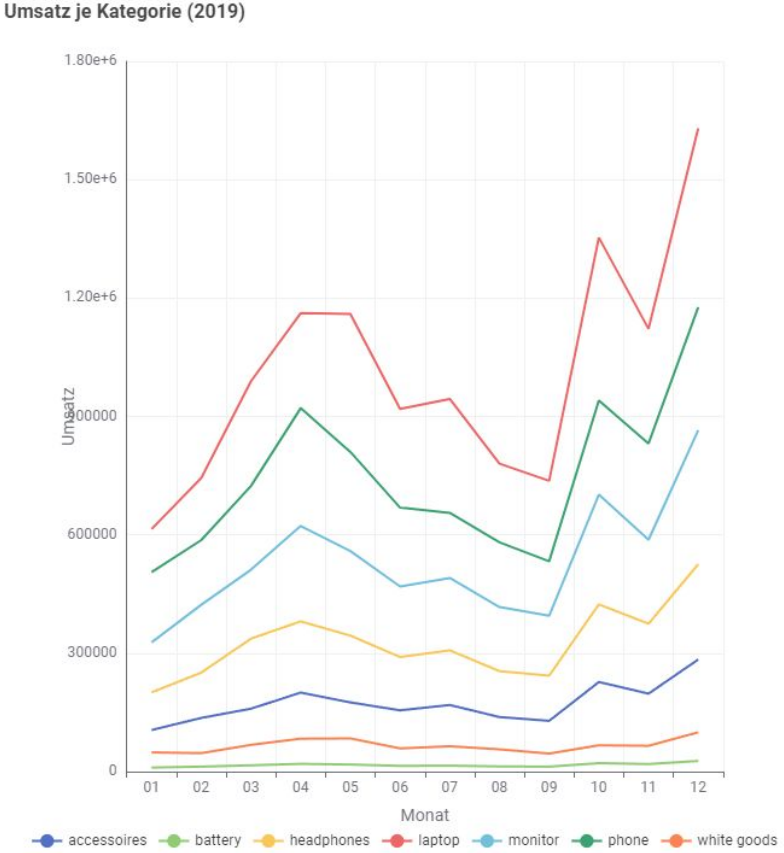
- Mit Abstand am wenigsten verkauft wurden ‘white goods’
  - Waschmaschinen & Trockner

Visualisierung & Pivot Tabellen



- Produktkategorien mit den meisten Transaktionen
  - Accessoires
  - Kopfhörer
  - Batterien
- Auch mit Abstand am wenigsten Transaktionen gab es bei ‘white goods’

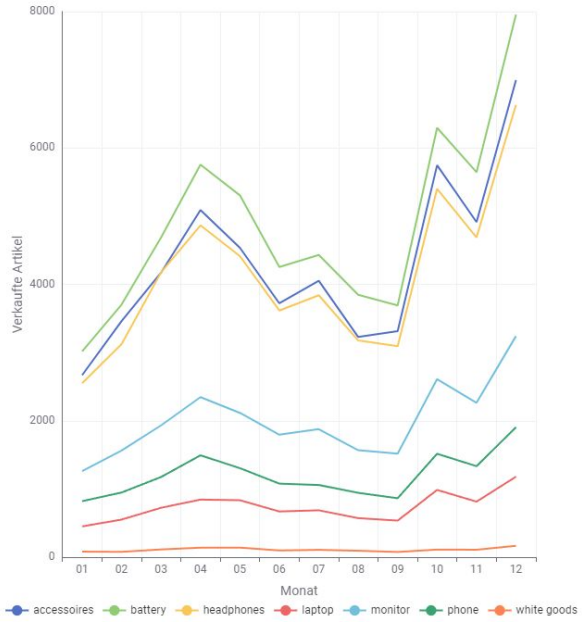
Visualisierung & Pivot Tabellen



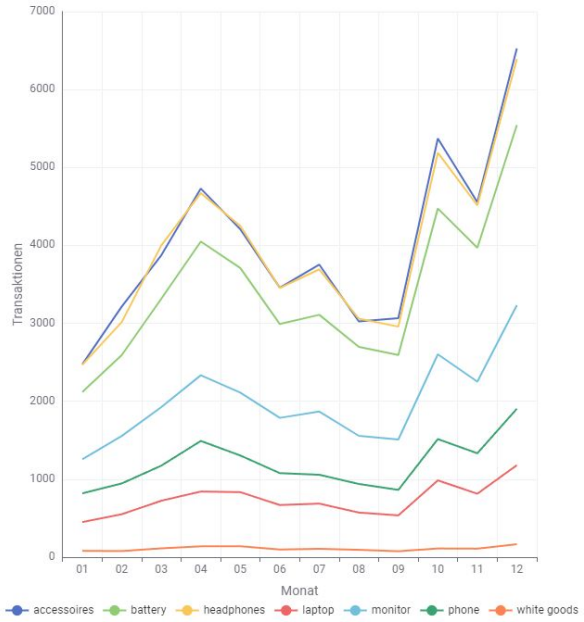
- Produktkategorien mit dem meisten Umsatz
  - Laptops
  - Smartphones
  - Monitore
- Am wenigsten Umsatz entstand durch Batterien und ‘white goods’

Visualisierung & Pivot Tabellen

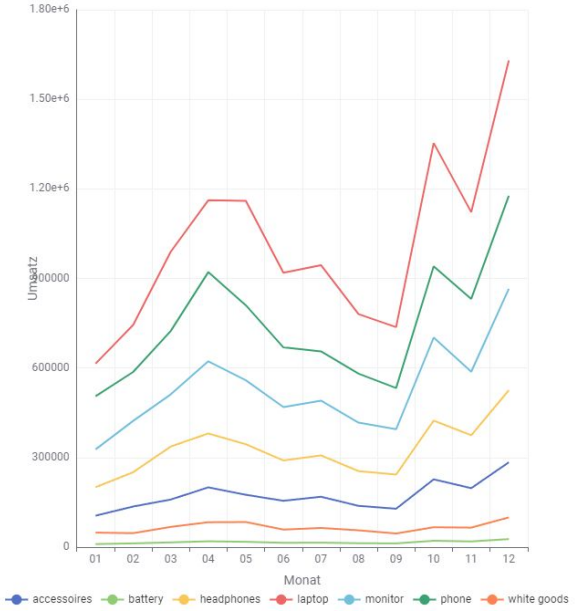
Verkaufte Artikel je Kategorie (2019)



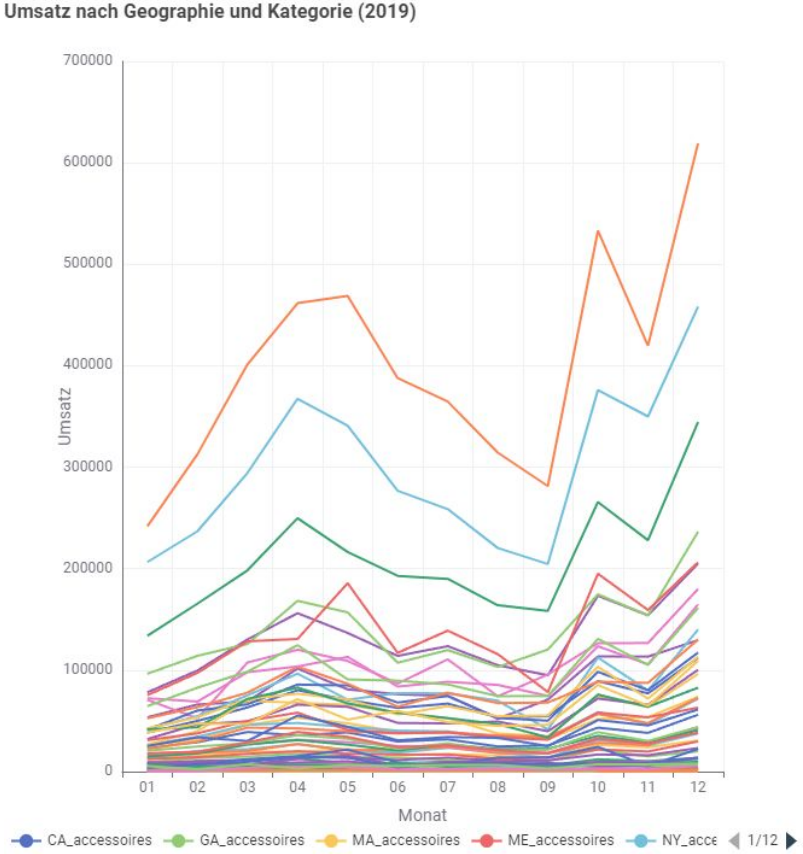
Anzahl Transaktionen je Kategorie (2019)



Umsatz je Kategorie (2019)

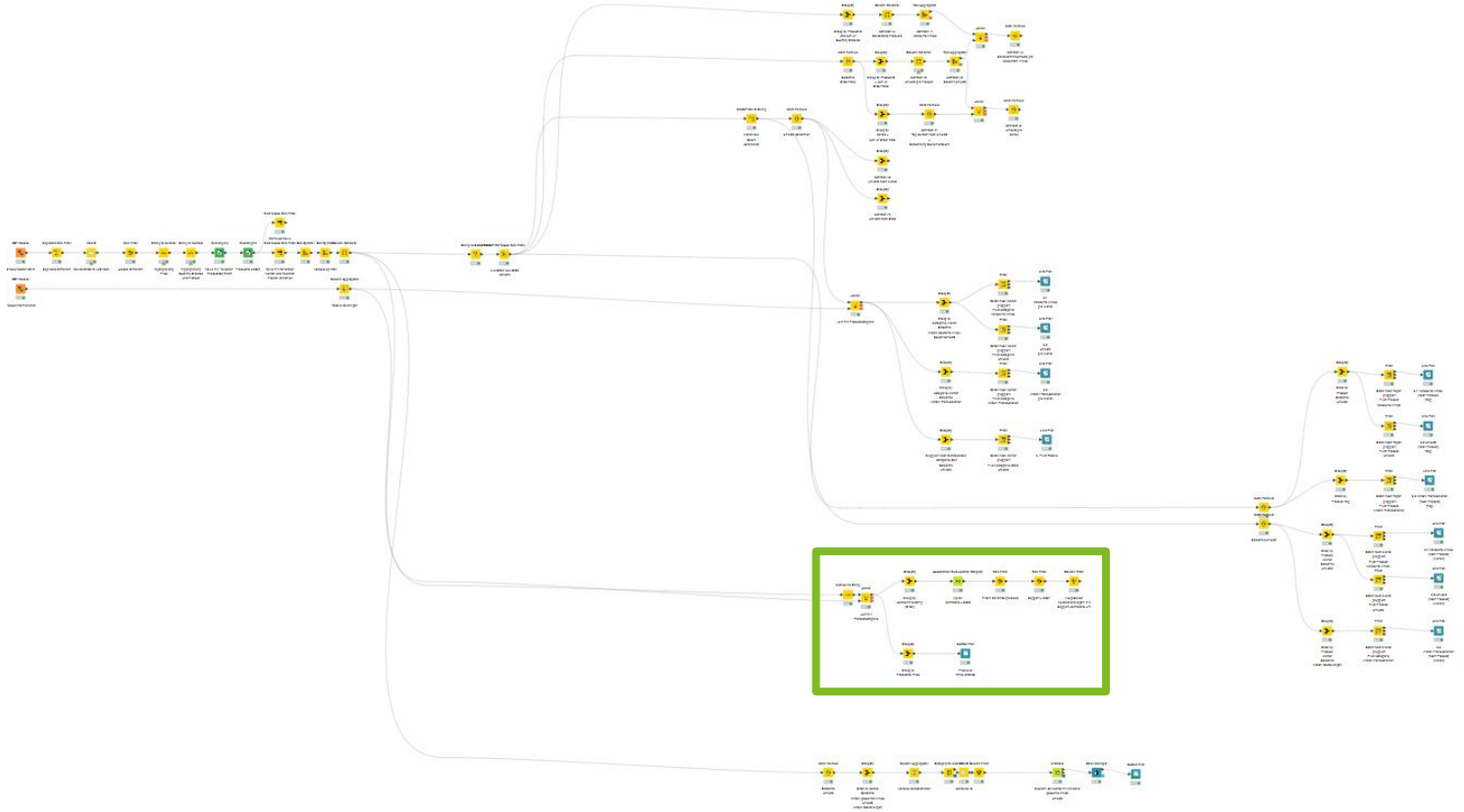


Visualisierung & Pivot Tabellen

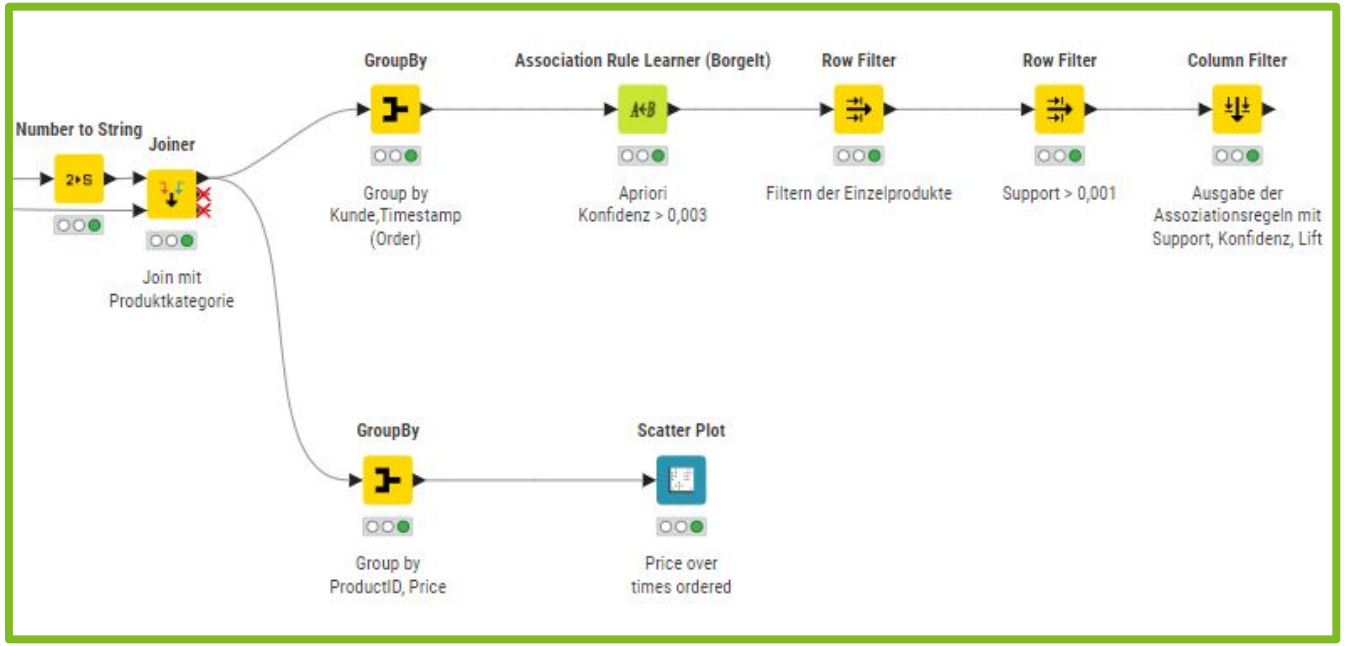


- Am meisten Umsatz
  - CA Laptop
  - CA Smartphone
  - CA Monitor

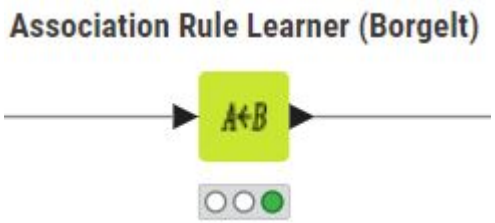
KNIME Workflow - Apriori



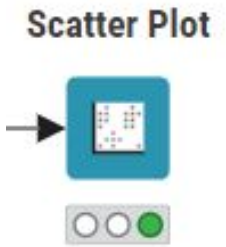
KNIME Workflow - Apriori Algorithmus



Apriori Algorithmus



- Association Rule Learner
  - Wendet Apriori Algorithmus auf Input Datensatz an
  - Liefert Assoziationsregeln



- Scatter Plot
  - Ermöglicht Darstellung von Daten als Streudiagramm

Apriori Algorithmus

- Input Daten für ‘Association Rule Learner’ sind alle Bestellungen
  - Datensätze mit gleichem Timestamp und Adresse in eine Bestellung zusammengefasst, mit Liste der bestellten Artikel

- Die Output Assoziationsregeln gefiltert nach
  - Konfidenz, Support, Lift

Consequent *String*	Antecedent *List*	RelativeItemSetSupport% *Number (double)*	RuleConfidenc... ↓ *Number (double)*	RuleLift *Number (double)*
iPhone	[Lightning_Charging_Cable]	0.566	4.68	1.22
Google_Phone	[USB-C_Charging_Cable]	0.559	4.56	1.474
iPhone	[Wired_Headphones]	0.259	2.45	0.639
iPhone	[Apple_Airpods_Headphones]	0.209	2.4	0.627
Google_Phone	[Wired_Headphones]	0.237	2.24	0.724
Google_Phone	[Bose_SoundSport_Headphones]	0.128	1.71	0.554
Vareebadd_Phone	[USB-C_Charging_Cable]	0.206	1.68	1.455
USB-C_Charging_Cable	[Wired_Headphones]	0.114	1.08	0.088
Wired_Headphones	[USB-C_Charging_Cable]	0.114	0.929	0.088

Apriori Algorithmus - Erkenntnisse

Google_Phone	[USB-C_Charging_Cable]	0.559	4.56	1.474
Vareebadd_Phone	[USB-C_Charging_Cable]	0.206	1.68	1.455
iPhone	[Lightning_Charging_Cable]	0.566	4.68	1.22

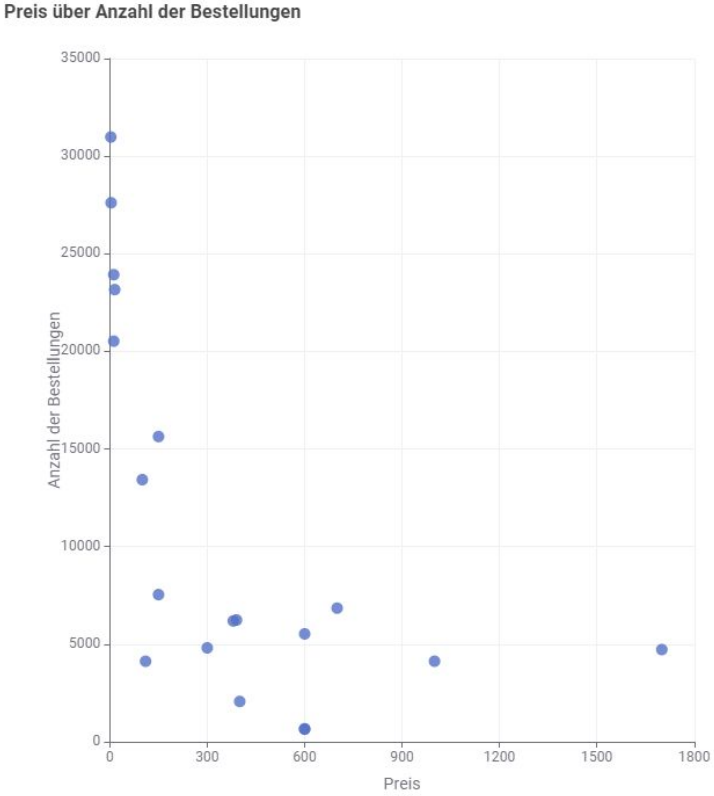
- Smartphones werden häufig zusammen mit Ladekabeln gekauft
  - Hat ein Kunde ein USB-C Ladekabel “Im Warenkorb”, kauft der Kunde ca. 1,5x häufiger auch ein Google Phone
  - Hat ein Kunde ein Lightning Ladekabel “Im Warenkorb”, kauft er mit fast 5% Wahrscheinlichkeit auch ein iPhone

Apriori Algorithmus - Erkenntnisse

Google_Phone	[USB-C_Charging_Cable,Wired_Headphones]	0.049	42.9	13.848
iPhone	[Lightning_Charging_Cable,Wired_Headphones]	0.035	48.8	12.74
iPhone	[Apple_Airpods_Headphones,Lightning_Charging_Cable]	0.026	40.5	10.569
Google_Phone	[Bose_SoundSport_Headphones,USB-C_Charging_Cable]	0.02	34.3	11.088
Vareebadd_Phone	[USB-C_Charging_Cable,Wired_Headphones]	0.018	16.3	14.047
Google_Phone	[Bose_SoundSport_Headphones,Wired_Headphones]	0.013	32.4	10.48
Vareebadd_Phone	[Bose_SoundSport_Headphones,USB-C_Charging_Cable]	0.009	15.7	13.554
iPhone	[Apple_Airpods_Headphones,Lightning_Charging_Cable,Wir...	0.002	100	26.086

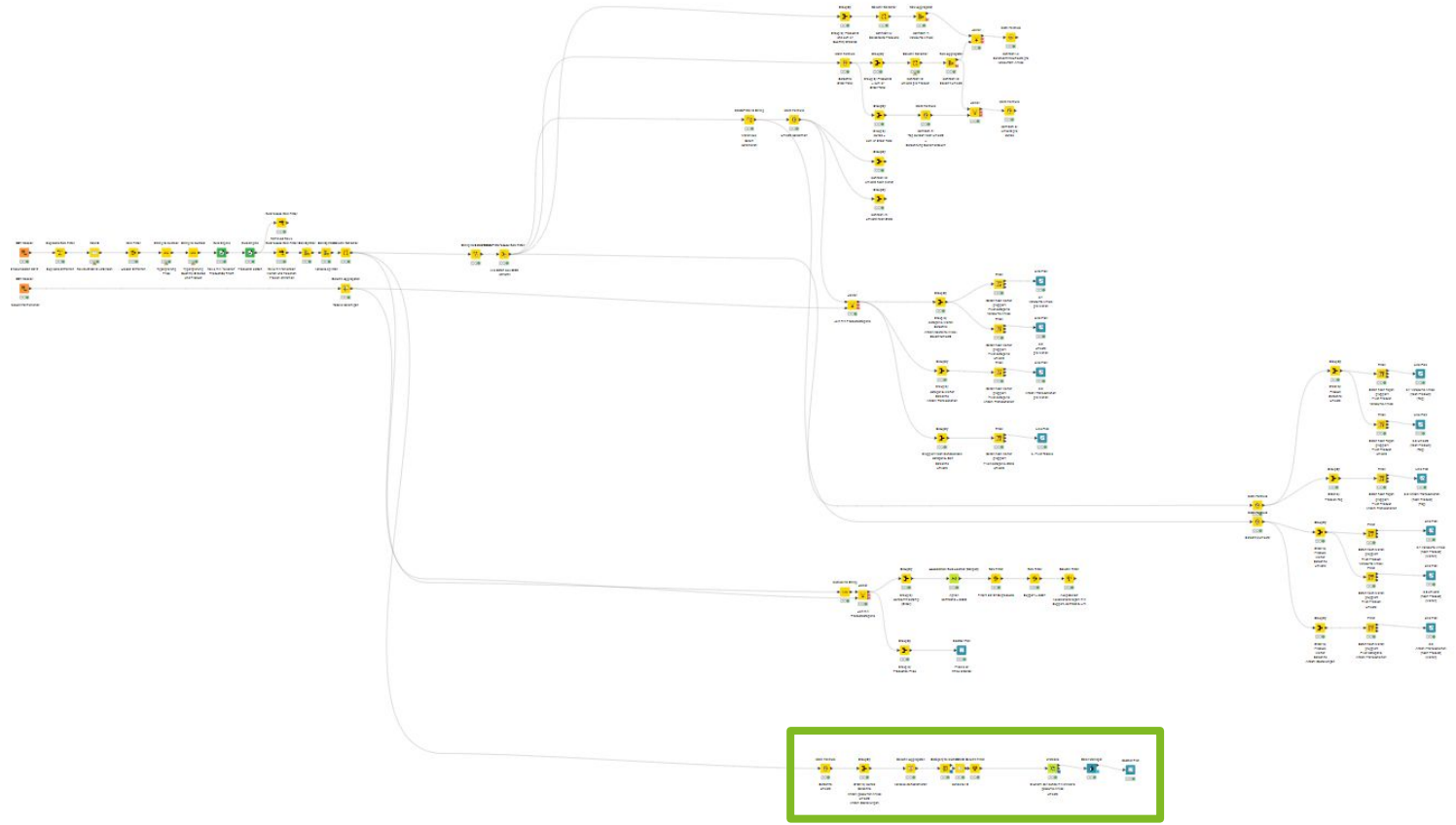
- Wenn Kunde bereits Ladekabel **und** Kopfhörer gekauft hat, so kauft er sehr wahrscheinlich auch ein Smartphone
  - Lift > 10
    - Kunde kauft dann >10x häufiger ein Smartphone

Apriori Algorithmus

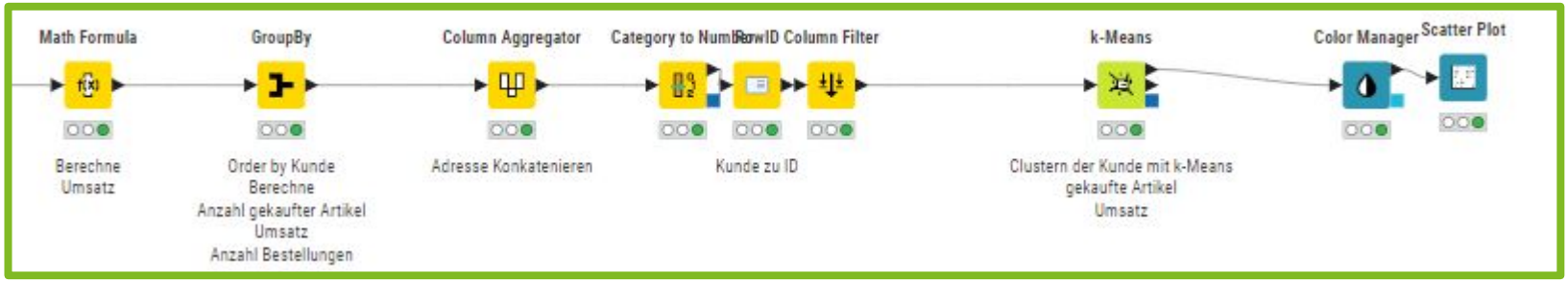


- Kostengünstige Produkte werden deutlich häufiger bestellt als teure Produkte

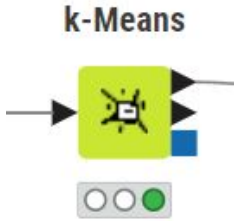
KNIME Workflow - Clustern der Kunden



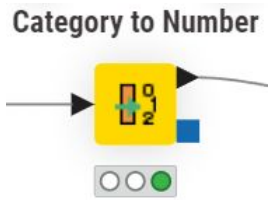
KNIME Workflow - Clustern der Kunden



Clustern der Kunden

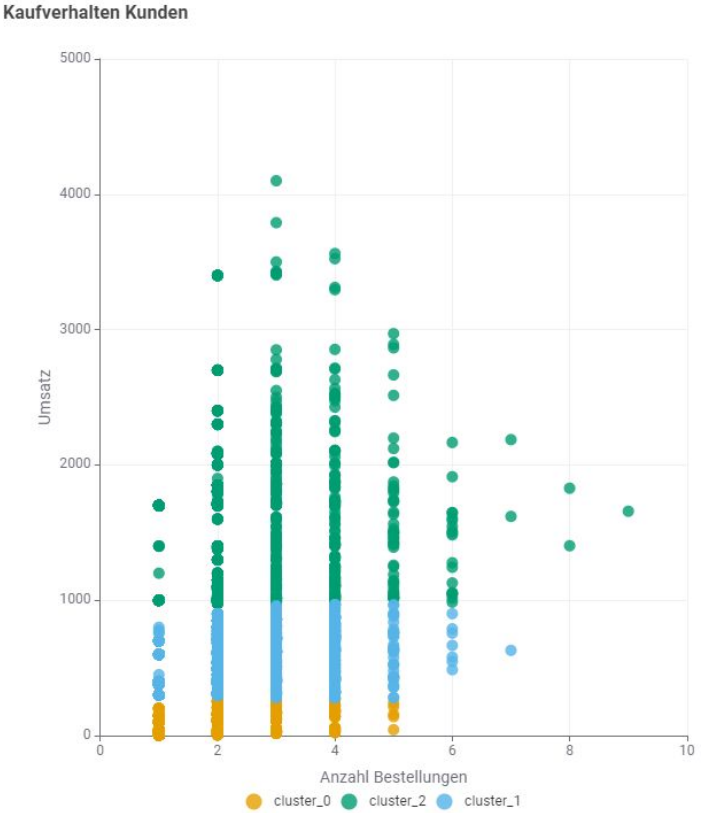


- k-Means
  - Wendet k-Means Algorithmus auf Datensatz an
  - Input Daten gruppiert nach Kunden



- Category to Number
  - Verwendet, um Kunden bzw. Adressen in KundenIDs umzuwandeln, um k-Means Algorithmus anzuwenden

Clustern der Kunden



- Wir haben uns für 3 Cluster entschieden
- Orange
  - Kunden mit wenigen (meist 1-3) Bestellungen und wenig Ausgaben (<300)
- Blau
  - Kunden mit mittleren Bestellungen (meist 2-4) und mittleren Ausgaben (300-1000)
- Grün
  - Kunden mit meist mehreren und teuren Bestellungen (<1000)

Vielen Dank für ihre Aufmerksamkeit!