Also wir haben Spielchen gespielt, wir haben Muster gelernt, was haben wir noch gemacht? Ich zeige diesen Zyklus, vielleicht erinnert ihr euch an den. Wir haben schon den ersten Schritt, oder besser gesagt, die ersten zwei, die hier in einem Kasten zusammengefasst sind, schon gemacht. Erinnert ihr euch? Probleme formalisiert haben wir. Wir hatten ein ganz volles Whiteboard. Weiß noch jemand, was man machen muss, wenn man ein Problem macht?

was umgangssprachlich beschrieben ist, formalisieren will als Klassifikationsregelungsaufgabe. Und wir haben eine Instanz dazwischen. Genau, wir müssen fragen, was ist jetzt dran? Unsere Klasse 50, ich glaube. Ja. Klasse 50 und Zielwertung. Ist das ein Schifferpunkt? Nein, das war ein Bayern-Klub-Klub.

Genau, also Klassenergebung oder Zielvariable, das sind unterschiedliche Namen, die wir verwendet haben, je nachdem, ob es eben Klassifikation oder Regression wäre. Klassifikation, Klassifikation, Regression. Genau. Genau, also beim Schiedsgebiet hatten wir eine numerische Zielvariable. Wir wollten wissen, wie viele Leute kommen. Also Anzahl des U-Karates, das war...

Dann haben wir auch Beispiele zur Klassifikation. Weiß noch jemand, welche gegeben haben, die sich mit der Klassifikation problemlos beschäftigt haben? Das war dein Beispiel. Also komm, wir gucken sozusagen. Ja, genau. Also Kreditvergabe war es, ne?

Unser Klassener Tribut war entweder, ob wir den Kredit vergeben sollen, ja oder nein. Und dann hatten wir noch überlegt, vielleicht können wir sogar vorhersagen, ob der Kredit zurückgezahlt werden wird oder nicht. Und dann darauf unsere Entscheidungen basieren lassen. Und wir hatten noch zwei weitere Beispiele. Und eins davon werden wir heute nochmal ganz genau unter die Lupe nehmen. Also nicht, dass das besonders wichtig wäre. Es ist einfach ein Beispiel, anhand dessen wir...

die Data Preparation, Datenvorverarbeitung, uns nochmal ganz genau anschauen. Also das ist unser Thema für heute. Sagen wir mal, gehen wir in den anderen Folien, da hat die Folie ein Gefühl. Die Daten vorzubereiten und das hat zwei Teile. Also einerseits die Daten so zu transformieren, wie es unsere

Formalisierung vorgibt. Also wir definieren in der Formalisierung, was sind die Instanzen. Wir müssen also sicherstellen, dass in der Datei, die wir generieren und hinterher in Orange oder irgendein anderes Pool laden, wo Instanz genau eine Zeile existiert. Wir müssen sicherstellen, dass es eine Spalte gibt, wo unser Klassenattribut abgebildet ist, also mit der richtigen Überschrift und den richtigen Werten drin und wir müssen die anderen Attribute als Spalten in dieser Datei kreieren. Und dann werden wir noch ein bisschen diskutieren über weitere Dinge, wie

einem so entgegenkommen und es wird hart heute. Es ist auch in so einem Projekt, wenn man den Zyklus anschaut, der härteste und aufwendigste Schritt. Bevor wir das machen, gab es ja das Quiz, was du angesprochen hast, Jeremy. Also die Frage war ja, ob das verpflichtend ist. Nein, also wenn ich Hausaufgaben gebe, dann ist es auf jeden Fall sinnvoll, wenn ihr die macht. Ich werde es nicht irgendwie überprüfen.

Aber ich werde auch jedes Mal fragen, ob es dazu Fragen gibt, ob wir was anschauen sollen zusammen. Und wenn ihr es nicht gemacht habt, dann könnt ihr mich auch nicht fragen. Jetzt wäre sozusagen der Punkt, wir könnten mal gucken, bevor wir jetzt hier voll einsteigen. Hattet ihr irgendwo Zweifel oder war euch irgendwo was nicht klar beim Quiz? Alles sonnenklar? Alle Fragen easy? Ich kann auch noch... Ja, mach. Geben wir die Daumen.

Ich frage was? Also ich kann es einmal sozusagen durchgehen, wie ich es mir gedacht habe. Und dann kannst du sagen, wie es auch anders sein könnte deiner Ansicht nach. Dann können wir gucken. Vielleicht geht es ja auch anders. Also es ist manchmal so, dass es nicht die richtige Lösung gibt. Mal schauen. Also.

Es geht darum, Betrug in einer Unfallversicherung aufzudecken. Betrug in dem Sinne, dass der Leistungserbringer, also der Arzt, manchmal im Teamwork mit dem Patienten, Sachen abrechnet, die nicht im Zusammenhang mit dem Unfall stehen. Also die Unfallversicherung kriegt eine Schadensmeldung, dass jemand bei der Arbeit zum Beispiel einen Unfall gehabt hat und wartet dann auf Rechnungen von Ärzten, die diesen Betrug

dieses Problem behandeln. Und natürlich, also, dass es überhaupt Betrug gibt bei dieser Fragestellung, liegt daran, dass man in der Krankenversicherung einen Selbstbehalt hat und einen Teil der Kosten selber tragen muss im Allgemeinen, bei der Unfallversicherung nicht. Das heißt, da gibt es einen Anreiz, gewisse Probleme, die ich schon immer hatte, an meinem Körper auch noch behandelt zu lassen und das dann der Unfallversicherung unterzuschieben. Man muss nichts dafür bezahlen. So, das bisschen zum Hintergrund. Und

Jetzt ist die Frage, was sind hier die Instanten? Und wenn ihr euch jetzt überlegt, also es kann natürlich sein, dass man zum Arzt geht und dann werden verschiedene Sachen behandelt. Manche davon stehen im Zusammenhang mit dem Unfall und andere nicht. Und deswegen muss ich eigentlich hier auf die Rechnungsposition gehen und sagen, okay, da wurde jetzt, was weiß ich, der linke Arm in irgendeiner Weise geröntgt oder was, ist das diese einzelne Aktivität?

gerechtfertigt oder nicht, also gegenüber der Unfallversicherung abzuzeichnen. Gut, jetzt steht da, dass ein Klassenattribut festgelegt wird, also dass man heute schon darauf hin, dass es um eine Klassifikation geht. Jetzt kann man sich überlegen, was wird da vorhergesagt? Ja, eigentlich ist das wahrscheinlich offensichtlich, ob es eben Betrug ist oder nicht, die einzelne Rechnungsposition. Das heißt, wir haben eine Klassifikationsaufgabe, Betrug nicht Betrug ist unser Klassenattribut.

So, und dann gibt es noch Feed-Engineering-Fragen. Das ist natürlich schwierig, weil ich jetzt gesagt habe, wichtig, okay, da habe ich mich mit dem Satz hin, was ist wichtig als Attribut? Man weiß es manchmal nicht vorher und es können natürlich auch mehrere wichtige Attribute geben. Ich habe mir, glaube ich, Mühe gegeben, auch Sachen zu nehmen, die eher weniger relevant sind und zu gucken, dass ich vor allem ein Relevantes habe, der Code, genau, also

Man will sicher wissen, was wurde da behandelt. Und dann wird man das wahrscheinlich auf der anderen Seite irgendwie abgleichen mit, ich weiß nicht, ob das auch kodiert ist, wahrscheinlich die verschiedenen Arten von Schadensfällen oder Arten von Unfällen, die da als Schaden eingereicht werden. Also zu wissen, was da behandelt wird, ist sicher gut als Attribut. Ja, was ist nicht gut?

Also da ist man sich relativ sicher. Jeder Kunde hat seine eigene Versicherungsnummer. Okay. Wir können darüber diskutieren. Also natürlich könnte es sein, dass die Person schon mal auffällig geworden ist, dazu neigt, sowas zu machen. Aber so viele Unfälle hat man wahrscheinlich in seinem Leben nicht, dass man da wirklich Muster erkennen kann. Das wäre jetzt so meine Story, die ich da erzählen würde. Jetzt weiß ich nicht, Marc, wo genau du Zweifel hast. Oder es anders ist. Ja.

Lass mal gucken, ob ich meine eigene Aufgabe richtig gelöst habe. Das sieht gut aus. Genau, aber sozusagen...

Ist klar, das ist jetzt sozusagen granularer. Also ich kann berechtigte Kosten fordern und andererseits manchmal Betrug dabei haben, aber das ist dann eher auf der Ebene der einzelnen Behandlung. Das ist gerechtfertigt und das nicht. Macht Sinn? Okay. War noch irgendwo eine Frage vorhin? Ich hatte das Gefühl, ich scroll durch. War alles klar. Okay. Gut, dann hatten wir nochmal ein Beispiel. Ist auch gut als Wiederholung gewesen jetzt.

Jetzt geht's los. Also, ich würde gerne nochmal gucken auf dieses Beispiel. Wir können kurz die Folie hernehmen. Also unser Running Example wird die Firma Teleflow sein, also eine Telekom-Firma, die genau dieses Problem adressieren muss oder will. Thema Churn oder Kundenabwanderung. Und das war ja eines unserer Beispiele letzte Woche. Wir können jetzt einfach nochmal gucken, was wir an der Tafel hatten.

Wahrscheinlich in dem Fall, ich habe mal gecheckt, es ist auch noch gut zu gucken, was die Freitagsklasse hatte. Die hatte nämlich eins von den Attributen, was Teleflow oder was wir nachher noch konstruieren werden, auch gebrainstormt. Ihr hattet andere, aber auch eins der Attribute ist, glaube ich, auch dabei. Genau, Anzahl Tickets zum Beispiel. Also hier haben wir Verträge. Jeder Vertrag ist eine Instanz und das Klassenattribut wurde definiert als verlängert, ja oder nein.

Wir werden es umdrehen. Nachher werden wir sagen, okay, Vertrag wurde gecancelt. Ja oder nein? Aber das ist ja das Gleiche. Nicht alle von diesen Attributen werden wir haben, aber manche davon. Insbesondere zum Beispiel die Anzahl der Tickets, die jemand gehabt hat. Und wenn ich nochmal auf die Freitagsklasse schaue, da gibt es noch ein Attribut. Wie haben Sie es genannt? Manche Sachen sind

Ähnlich, Vertragsdauer bisher zum Beispiel wird auch noch eine Rolle spielen bei uns. Und die hatten hier noch den Leistungsumfang. Also welche Services nutze ich? Diese Daten haben wir auch bei Teleflow und vielleicht ist es ja eine gute Idee, wenn wir das auch nutzen. Vielleicht gibt es ja manche Services, die eher zu Problemen führen, wo die Leute dann unzufrieden sind. Hat es bei euch geklappt, Tableau Prep zu installieren? Sollen wir es mal hochstarten? Also ich habe es schon gestartet. Wo muss man denn den Schlüssel eingeben?

Wenn du es aufmachst, fragt er dich eigentlich. Wenn er dich nicht fragt, dann benutzt du Sonne-Schlüssel, aber normal, bevor du irgendwas machen kannst, fragt er dich eh danach. Dann probier es mal.

Produktflüssel verwalten. Hast du da eins? Ja, schon. Okay, super. Gut, also bei euch ist hier wahrscheinlich weniger. Das sind meine ganzen Workflows, die ich schon mal gemacht habe. Und ja, kurz vielleicht einführende Slide. Wir werden drei Dateien haben. Ach so, jetzt nochmal raus. Die Dateien findet ihr auf Moodle. Im Ordner

Teleflow-Daten, hier im Abschnitt 2, direkt unter den Infos zum Tableau Prep, findet ihr diesen Ordner, da sind die drei Dateien drin, wo ich einen Screenshot jeweils auf der Folie habe. Es geht los mit den Kontakten. Die werden als erstes laden, also das ist das, wie ihr auch vorgeschlagen hattet.

dass wir unsere Instanzen definieren. Jeder Vertrag eine Zeile. Wir können das mal laden hier, also ladet euch das runter. Ihr habt es schon irgendwo. Wenn ihr hier auf Verbindung oder ihr könnt auch oben den Knopf hier Verbindung zu Daten herstellen benutzen. Das ist eine CSV-Datei, für Tableau ist das Text. Also Textdatei müsst ihr hier auswählen. Und dann Contracts. So, ich brauche einen Spickzettel.

Damit alle soweit sind, seid ihr soweit schon? Noch ein paar Klicks entfernt. Also Tableau ist ein Tool zum Visualisieren. Mit dem kannst du Charts erstellen und Reports und Dashboards und sowas. Und das machen wir gar nicht. Also da gibt es vielleicht ein anderes Modul, wo ihr euch das anschaut.

Tablo Prep Builder ist wirklich zum Vorverarbeiten von Daten. Also das, was wir uns heute anbauen. Normalerweise ist der Ablauf so, dass du Daten irgendwo herziehst. Das bietet ja hier auch relativ viele Quellen an. Also du kannst auch aus einer Datenbank oder so das ziehen. Dann machst du hier dein Workflow, bearbeitest die irgendwie vor und werdet

Und wenn du dann fertig bist, dann lädst du sie in Tableau und visualisierst sie. Und wir werden sie nicht in Tableau laden, sondern in Orange, aber Datenvorverarbeitung ist sozusagen hier der Zweck von Tool. Okay, jetzt, was man immer machen kann, auch um sich die Daten ein bisschen anzuschauen, also ihr seht ja hier schon eine Vorschau, aber normalerweise braucht man erstmal so einen Aufbereitungsfluss. Da kann man auch schon anfangen, Spalten wie einen Nerven oder den man...

Kein Wert beim ist schon mal zu löschen. Also was heißt, du kriegst keine Verbindung, wenn du hier auf Textdatei klickst? Ja, genau. Okay. Ja. Kannst du nicht? Ja.

Die Anna hat ja auch schon gefragt, was ihr alles falsch habt.

Also es sieht wahrscheinlich auch fundamental aus, wenn die Tabellen installiert haben.

Also wir werden jetzt Feature Engineering machen.

Also Features löschen ist immer einfach. Also du kannst jetzt hier draufklicken und sagen, ob jemand Kinder hat zum Beispiel, interessiert mich nicht, entfernen. Also glaube ich nicht, dass das wichtig ist. Und was wir aber auch machen, was noch viel spannender ist, ist, dass wir noch Features dazu kreieren. Auch aus den anderen Dateien, die wir jetzt noch nicht geladen haben. Und aber auch aus den Daten, die wir da haben, noch per Berechnung noch neue Sachen kreieren. Okay.

Wie schaut es aus? Also, Diana hat es auch gerade schon gefragt. Ich sage es nochmal. Tableau ist ein Visualisierungstool. Das machen wir gar nicht hier im Machine Learning. Vielleicht macht ihr es in einem anderen Modul, vielleicht auch mit einem anderen Tool. Also ihr habt vielleicht auch ein Modul Business Intelligence oder sowas. Da braucht man solche Tools. Tableau Prep.

Also der Name Prep bezieht sich auf die Vorverarbeitung von Daten. Das ist das, was wir heute machen wollen. Wir werden die Daten vorverarbeiten und dann in Orange laden. Also nicht mehr heute, aber wir werden dann vorverarbeitete Daten immer wieder in Orange laden. Also zum Beispiel in eurem Semester-Assignment müsst ihr auch erst mit Prep. Also ich empfehle Prep. Man kann manche Sachen auch in Orange selbst machen. Aber Orange ist sehr, sehr limitiert, was das angeht.

Deswegen, Tableau Prep ist sehr viel besser geeignet. Es ist eins der möglichen Tools, die man für Datenvorbereitung nehmen kann. Und Tableau bietet es an, damit du deine Daten für Tableau vorbereiten kannst. Du kannst sie aber genauso gut vorbereiten und dann in irgendein anderes Tool laden. Und das ist das, was wir machen. Ist klar? Okay. Jetzt habt ihr alle Tableau Prep. Das ist schon mal gut. Und jetzt könnt ihr auch die Daten laden und diesen Aufbereitungsschritt erzeugen.

Gebt mir mal einen Daumen hoch oder sowas. Du brauchst noch. Okay, also solange du noch brauchst, möchte ich mal kurz eine Sache anschauen. Ich werde zwei Kunden hier immer wieder rauspicken, also Verträge rauspicken. Der eine ist der 7892 POO KP. Könnte es anders sein. Also zwei, deswegen einer von denen hat gekündigt, der andere nicht.

Und wir können die zwei Beispiele immer wieder benutzen, um so ein bisschen zu verifizieren, dass wir die richtigen Sachen machen. Einfach als Testbeispiele. Stimmt auch wieder. So, den hier meine ich. So, jetzt schauen wir uns den an. Der, nee, die, ist eine Frau, hat am 8.2.2020 den Vertrag bekommen.

Okay, dann sehen wir irgendwie, was das Zeug ist und Zahlungsmethode, wie viel zahlt sie pro Monat? Und hier, dieser Vertrag wurde gecancelt am 28.05.2022. So, jetzt müssen wir noch kurz auf die Folie zurückgehen und lesen. Die Firma Teleflow bietet ein Jahres mobile Punktverträge. Wenn wir jetzt anfangen wollen, Trainingsdaten zu kreieren aus Teleflow,

Sagen wir mal nur diesem einen Beispiel. Stellt euch irgendwas auf. Wir haben jetzt hier definiert als Instanzen letzte Woche Verträge. Und hier wird es seine Cancellation. Fangen wir schon wieder an in Englisch. Entschuldigung, das ist die Seite drunter. Und Attribute, ja, kommen ganz viele zusammen nachher noch.

Bei den Instanzen, da möchte ich gerne noch was nachschärfen. Habt ihr eine Idee, was ich meinen könnte? Wenn ihr jetzt mal auf diese Kundin schaut, die also am 2.8.2020 ihren Vertrag bekommen hat und dann etwas mehr als zwei Jahre später gekündigt hat und jetzt zusammen mit der Information, die hier steht, dass die Firma ein Jahresverträge anbietet. Länge raus sind die noch.

Genau, also die Kundin ist länger als ein Jahr mit ihrem Vertrag bei uns. Ist das jetzt immer noch eine Instanz oder ist es eigentlich so, dass wir hier aus dieser Zeile vielleicht mehrere generieren wollen, mehrere Instanzen? Also eigentlich ist es doch so, das erste Jahr hätte sie kündigen können. Immer noch ein Jahr kann man kündigen, weil es sind eigentlich zwei Jahre in den meisten Fällen der Praxis. Aber sagen wir mal, viele Fälle sind ein bisschen kontaktiert.

Hat sie aber nicht gemacht. Also am 08.02.2021 ist ihr Vertrag oder am 07.02. ausgelaufen und sie hat nicht gekündigt. Also das ist eigentlich ein Trainingsbeispiel, wo Sie sagen, okay, in dem Jahr war alles gut, nicht gekündigt. Und dann gab es noch ein Jahr bis zum 08.02.2022, da war auch alles gut. Und erst im dritten Vertragsjahr hat sie dann gekündigt. Also würdet ihr mitgehen? Ich würde gerne aus dieser einen Zeile drei Zeilen machen. Ich würde gerne drei Instanzen daraus machen.

Ich würde unsere Instanzen umbenennen in Vertragsjahre. Also jede Instanz beschreibt ein Jahr Vertrag sozusagen und wir gucken immer am Ende, wurde gekündigt oder nicht. Das heißt, wenn jemand mehrere Jahre da war, dann sind es immer die ersten paar Jahre haben immer dann ein No hier und nur das letzte Jahr hat ein Yes. Und Kunden, die gar nicht gekündigt haben, die haben nur No. Das heißt, wir werden

Ja, ist sowieso gut. Wir werden relativ viele Beispiele haben von Notion. Seid ihr dabei? Jetzt ist die Frage, wie machen wir das? Okay. Vielleicht erstmal uns nochmal bewusst werden, was jetzt passiert. Wir werden jetzt Instanzen konstruieren. Das heißt, wir haben jetzt hier ein Beispiel, wo die Daten, die wir bekommen haben,

nicht eine Instanz pro Zeile haben, zumindest nicht so, wie wir uns die Instanzen vorstellen. Also wir haben jetzt hier eine neue Vorstellung entwickelt, was die Instanzen sein sollen. Die Daten entsprechen unserer alten Vorstellung, aber mit unserer neuen Vorstellung müssen wir jetzt tätig werden. Wir müssen jetzt hier was machen. Und das ist eigentlich so, wenn man es versucht zu abstrahieren, manchmal hat man so Zustände von einem System,

oder man möchte so Art Snapshots machen. Also hier machen wir jedes Jahr einen Snapshot. Wir gucken jedes Jahr, ist der Vertrag noch, läuft der noch oder wurde der gecancelt? Und das

Ursprungssystem, aus dem wir die Daten haben, bildet das nicht ab. Das ist eigentlich implizit. Und dann muss ich, wie ich hier geschrieben habe, die Mittelzeitstempeln aufteilen. Also was bedeutet das jetzt? Habt ihr irgendeine Idee, was wir machen können? Also es gibt hier...

tatsächlich ein Schritt, der heißt Neue Zeilen. Der fügt irgendwie Zeilen ein. Wir können den mal kreieren. Und was wollen wir jetzt machen? Hier haben wir jetzt eine Wahlmöglichkeit. Wollen wir den Wertbereich aus einem oder aus zwei Feldern? Und tatsächlich muss man erst mal verstehen, was das bedeutet. Deswegen, ich helfe euch jetzt oder uns. Also was wir wollen, ist eigentlich...

Zwischen dem Start, also dem achten, zweiten, ich mache das mal auf die Achtung. Unsere Beispielkundin, 7892, die ist die, glaube ich. Genau, also, was wollen wir eigentlich am Ende haben? Wir wollen, habe ich gesagt, drei Zeilen haben. Und wir wollen wahrscheinlich irgendwie, also wir haben den Start,

von dem Vertrag. Der ist immer gleich, 8.2.2020. Also lasst uns mal definieren, was wir am Ende haben wollen. Dann haben wir ein Ende. Das haben wir noch nicht kreiert, aber das werden wir gleich brauchen. Was ist das Ende von dem Vertrag? Lass uns zurückgehen hier und wieder filtern. Da brauchen wir, glaube ich, so vielleicht. Also hat diese Geschäftsbeziehung ein Ende?

Sie waren noch seit 2020. Ja, also dieses Statement existiert eigentlich schon. Ja, aber jetzt...

Ich will die anders nennen. Also sozusagen diesen Stab, den werden wir später auch nochmal brauchen. Ich benenne das mal auch gleich um ins Stab, damit wir an der Tafel das gleiche haben wie hier. Einfach nur Stab.

Ich will das behalten. Ich will zum Beispiel nachher ausrechnen können, wie lange jemand schon bei uns ist insgesamt. Deswegen behalte ich das. Ich nenne das neue. Ich nenne es mal Period Start. Also Period meine ich Periode des Vertragsjahr. Am Ende habe ich noch einen besseren Quartal. Du wirst Ende 8, 2, 3 und schon. Also lass uns Period Start und Period End machen.

Mit Ende meine ich jetzt sozusagen wirklich das Ende der Geschäftsbeziehung. Wann ist das eigentlich? Das kann man hier sehen. 28.5.2022. Genau. Das kopiere ich jetzt auch dreimal. So möchte ich das am Ende haben. Okay, und jetzt hier, die Periode begann tatsächlich 2020, oder? Und wann hat das erste Vertragsjahr geendet? Ja, ich weiß nicht.

Achter, zweiter, also eigentlich achter, siebter, zweiter, aber wir machen es mal so, weil so kommt es raus, ist auch okay, oder? Also wollen wir jetzt nicht so... Und dann ging es hier achter, zweiter, 21 bis achter, zweiter, 22, richtig? Und dann nochmal achter, zweiter, 22, das war dann kein ganzes Jahr mehr, bis 28.05.22, ja? So waren die drei Instanzen oder so sollen die drei Instanzen werden, die wir jetzt trainieren, okay?

Jetzt brauchen wir erst mal das hier. Und dann können wir diese neue Zeilen, dann können wir sagen, es soll uns zwischen den zwei Werten hier, also zwischen dem 8.2.20 und dem 28.5.22, soll es uns noch Zeilen generieren. Ja, das ist der Trick. Versteht ihr? Ich kann es euch noch nicht zeigen. Wir müssen erst mal dieses Ende hier als Attribut generieren.

Also neue Werte generieren geht oft meistens über berechnetes Feld erstellen. Also berechnetes Feld erstellen, ihr kennt ja Excel, oder? Stellt es euch so vor, ihr habt eine Tabelle, die letzte Spalte ist noch leer, da macht ihr eine Überschrift und dann fügt ihr eine Formel ein und dann zieht ihr die einmal runter. Und das ist im Prinzip das, was hier passiert. Wir geben die Formel ein und der Blueprep erstellt eine neue Spalte und kopiert die.

die Formel darunter sozusagen. Lass uns das machen. Also wir werden es Endel nennen oder End, bleiben dabei Englisch. Gibt es eine Rolle, ob die Aufbereitung oder die Neuzeit? Nein, theoretisch gibt es wahrscheinlich in fast jedem dieser verschiedenen Schritte die Möglichkeit, ein Berechnungsfeld zu erstellen. Ich finde es aber sauberer, wenn man es in aufbereitenden Schritten macht. Überhaupt, also man kann natürlich

Also ich will nicht dagegen reden, dass man in Datenaufbereitung auch zum Beispiel mit Python macht. Was aber cool ist an so einem Tool ist, dass es wie eine direkte Dokumentation liefert. Ich kann auch diese Schritte noch benennen in irgendeiner sinnvollen Weise. Und, also wenn ich das jetzt, ihr seht jetzt hier schon, wird Änderungen 1 angezeigt. Das ist jetzt das, was ich gerade mache. Dass ich diese Formel erzeuge, kann ich hier ausplatten und kann ich sehen. Dann kann ich das auch...

Löchern, vielleicht kann ich mir anschauen, also sozusagen meine Dokumentation. Okay, also wie kriege ich das Ende her? Jetzt muss ich, in dem Fall ist das Ende gleich den Cancellation Received Wert da hinten. Also man sieht jetzt das nicht mehr, aber man sieht hier, dass bei manchen Verträgen nichts steht und bei anderen Verträgen steht das Datum, wann wir die Kündigung bekommen haben.

Das heißt, sollen wir uns erstmal um den Fall kümmern, dass da was steht in der letzten Spalte, dass da ein Datum steht. Also jetzt müssen wir so ein bisschen eine Formel konstruieren, wie programmieren schon fast. Und wir werden irgendwie mit wenn, dann operieren. Also, wie würdet ihr es formulieren? Wenn da hinten was steht, dann... 8.22 steht, dann kommt 8.22. Ja.

Also sozusagen, da müssen wir ja für jeden Vertrag extra eine Regel haben. Also was ich machen will, ich gucke hier hinten. Ich gucke, haben wir eine Kündigung bekommen? Wenn wir eine bekommen haben, dann war das Ende des Vertrages gleich diesem Datum, oder? Das heißt, ich kann irgendwie gucken, dass ich die Cancellation Received

dass ich dieses Feld benutze in meiner Formel. Ich würde sagen, if CastellationReceived ist nicht 0, also wenn es 0 ist, dann müssen wir noch überlegen, aber wenn es nicht 0 ist, dann nehme ich den Wert einfach, oder? Dann nehme ich diesen Wert, der da steht. Also in unserem Fall bei dieser Kundin wird es dieser Wert aus dem Mai 2022 sein. Was mache ich, wenn da 0 steht?

Ja, dann läuft sie der Betrag nicht. Genau. Das ist doof, weil wir wollen ja nachher diese Zeilen erzeugen zwischen dem Start- und dem Enddatum. Also lass uns irgendwas reinschreiben. Ich würde jetzt einfach mal vorschlagen, sagen wir mal, wir haben das Ende letztes Jahr diese Analyse gemacht oder wir machen sie bis dahin. Lass uns einfach mal den 31.12.24 einschreiben. So.

dann wissen wir innerlich läuft noch aber damit da was steht damit wir die zeilen kreieren können ok also lass uns mal so jetzt beim ersten mal mache ich das ok beim nächsten mal dürfte ja auch mal probieren zu überlegen und auch zu tippen natürlich also es gibt eine funktion die heißt ist nahe

Wir fangen jetzt also mit dem Fall an, dass hinten nichts steht. Und dann haben wir gesagt, also, wenn ihr jetzt hier zum Beispiel das D einfach nur eingibt, D-A, schlägt euch zum Fall Cancellation Received. Also, kann ich das jetzt mal übernehmen? Also, was jetzt da steht, ist, wenn in der letzten Spalte 0 steht, dann. Also, was machen wir dann? Was haben wir gesagt? Tragen wir...

Ja, also es ist nicht Return, sondern wir tragen jetzt direkt Cancellation Received. Nee, Stop. Wenn da 0 steht, wollten wir den 31. Ja, genau. Ich könnte jetzt 31.12. schreiben. Ich muss aber noch eine Date-Funktion voransetzen, damit er das auch als Datum interpretiert. Also Date und jetzt kann ich so

Hochkommas machen und das Datumsformat, was das am besten versteht, ist, dass man erst das Jahr und dann den Monat mit einem Minus 12, 31. So.

Okay, also 2024 minus 12 minus 31. Ihr könnt es auch mit anderen Datumsformaten versuchen, aber so ist es am sichersten. Also, wenn da was steht, also wenn es nicht 0 ist, dann, hatten wir gesagt, wollten wir einfach den Wert aus der Spalte Cancellation Receive. Also tippe ich einfach Cancellation Receive. So, jetzt ist er nicht zufrieden, ich muss am Ende noch End eingeben. Das ist so die Syntax.

Tableau Prep. Jetzt sagt der Berechnung, es ist gültig. Jetzt gucken wir mal, was passiert, oder? Insbesondere würde ich mir das jetzt gerne angucken für unsere Beispielkundin. Also, es hat geklappt, ja. Ich sehe jetzt hier Ende 28.05.2022. Jetzt ist vielleicht der Zeitpunkt, wo ich noch einen zweiten Vertrag dazunehme.

Das ist der Vertrag. Was sagt mein Spielzettel? 5575 GNVDE. Ist ja logisch. Lass uns den mal anschauen. Der wurde nicht gekündigt. Der hier. Also ihr seht, Start war 16. Mai 2021. Kleine Zeichen mit Lied. Hier hinten ist 0. Und hier steht jetzt der Wert, den wir uns überlegt haben. 31.12.2024. Also hat geklappt. Super. Okay. Und jetzt wollten wir, der Schritt ist schon angelegt. Ja.

Zeilen generieren für jedes Vertragsjahr. Also jetzt gehen wir vielleicht mal zurück zu unserem 7892, die OKP. Wir hatten gesagt, am Ende wollen wir diese drei Zeilen, die ich hier schon auf der Tafel angefangen habe zu drehen. Und jetzt gucken wir mal, ob das funktioniert hier. Folgendermaßen, wir können jetzt sagen, wir wollen das aus zwei Feldern. Und hier steht sozusagen zwischen welchen Werten. Also Start und Ende.

So, und dann, wie viele Zeilen er generieren soll, sozusagen um wie viel soll jeweils erhöht werden. Also ich fange beim Start an, um wie viel soll ich jeweils dazugeben für die neue Zeile. Und jetzt gehen wir hier auf zwölf Monate. Also Jahre gibt es nicht. Das Größte, was ich hier habe, ist Monat. Das heißt, wenn ich ein Jahr machen will, muss ich zwölf Monate eingeben. Und der Rest soll alles aus der anderen Zeile kopiert werden. Also alles, was da...

für diesen Vertrag, die im 892 linksbums steht, wird mit kopiert. Jetzt gucken wir uns an, ob es funktioniert. Wir werden gleich sehen, dass da ein paar Sachen passieren, die wir nicht wollten, aber die kriegen wir auch noch in den Griff. Also leider sind es nicht drei Zeiten, sondern vier. Gucken wir uns die mal an. Wir haben jetzt... Warte mal, ich wollte noch was. Wir nennen das hier mal Period Start. Wir gucken gleich nochmal.

Also sozusagen die Formel mit dem Ende, die habt ihr alle.

Übrigens, ich hatte das ja erwähnt, das wird jetzt hier angezeigt und wenn ihr jetzt zum Beispiel noch was an der Formel ändern wollt, könnt ihr die hier wieder logischerweise dann auch bearbeiten. Da findet ihr hier unter Änderungen links. Wenn man sich eine Formel bearbeitet, gibt es dann eine dritte Änderung? Also sieht man die Bearbeitung? Nein, man sieht nicht die Schicht. Ja, okay.

Also, jetzt nochmal zum Mitklicken. Was habe ich gemacht? Neue Zeilen habt ihr. Gut. Jetzt habe ich gesagt, ich möchte die Zeilen generieren mit Hilfe von zwei Feldern. Und dann habe ich Start und Ende genommen. Also ich möchte zwischen dem Startdatum, also zwischen diesem Beispiel hier jetzt, zwischen diesem Datum und diesem Datum, möchte ich Zeilen erzeugen. Jetzt kann ich noch angeben, sozusagen,

ein neues Feld, was erzeugt werden soll, um sozusagen dieses Vertragsjahr zu charakterisieren. Ich nenne es Period Start. Werden wir gleich sehen, ob das gut kommt. Also ich habe das ja hier schon

und hoffe, dass dann da diese Werte drinstehen. Mal gucken. Und dann sage ich, um wie viel jeweils von dem Startdatum aus hochgerechnet oder addiert werden soll, um zum nächsten Startdatum

zur nächsten zeile zu gehen und da stelle ich 12 und monat ein jahr und dann sage ich noch dass die werte aus der vorherigen zeile also aus der ursprünglichen zeile die wir hatten alle anderen werden sollen probiert werden ok also bis so weit wie es mit den anderen alle auf dem stand wir machen auch gleich mal eine pause dann wieder auf dem gleichen stand okay

Muss natürlich auch mein Berliner essen, aber... Jetzt lass mal schauen, was passiert ist. Also ich hatte ja schon gesagt, leider sind vier Zeilen generiert worden. Also man kann diese Spalten hier auch rumschieben. Ich will zum Beispiel Period Start mir nach vorne bringen. Ich kann das jetzt einfach per Drag & Drop, wenn mein Computer mich lässt. Ich habe hier dieses Gerät.

Das habe ich verloren, das habe ich jetzt gemacht. Also wenn ich das jetzt anklicke und dann gedrückt halte und schiebe, dann sollte es mitkommen.

Ich will, dass man diese ganzen Datumsangaben nebeneinander sieht. Also Contract ID. Dann möchte ich den Start haben und das Ende. Und dann so wie an der Tafel. Dahin und das will ich danach haben. Ihr müsst es nicht unbedingt so machen. Schaut jetzt lieber mal an die Wand. Und jetzt können wir wieder nochmal filtern.

Dann sehen wir Period Start. Also das müsste ja jetzt sozusagen, müssten diese Werte hier sein. Ja. 8.2.2020, 8.2.2021, 8.2.2022. Das ist tatsächlich auch so. Also sozusagen hat eine Zeile generiert für das erste Vertragsjahr. Period End haben wir noch nicht. Müssen wir gleich bauen. Gleich nach der Pause. Dann nächstes Vertragsjahr startet am 8.2.2021.

Dann gibt es ein Vertragsjahr, das startet am 8.2.2022. Und dann hat er mir noch eins generiert, wo Period Start gleich End ist. Das muss ich dann noch aufräumen gleich. Jetzt weiß ich auch nicht, warum das passiert nicht. Das wollen wir nicht. Wir haben ja gesagt, wir wollen nur diese 13. So, jetzt glaube ich, könnten wir mal eine Pause vertragen. Ich würde vielleicht...

Nein, wir machen erst zuvor. 20 nach machen wir weiter, okay? Wer braucht das dann vor? Ja.

Ja, ich glaube, das ist gut. Ich meine, es ist diese Formel, oder? Ja. Okay. Ja.

Ja.

...

Und jetzt? Vielleicht kannst du nur einen... Also du musst nicht alles tippen. Wenn du auf geht's, dann schieb du dir nach. Wenn du was vorhast, dann schiebst du dich nach. Dann ja auch schlussreich, dann geht man das nach draußen. Dann musst du nicht glauben. Dann musst du nicht glauben. Dann ist es nicht gelaufen.

Wenn die diesen Moment da sind, dann ist das ein Beispiel.

Jetzt schweifen.

... ...

Vielen Dank.

Ach so.

Jetzt bist du schon mal in Europa. Nein, ich habe nicht. Irgendwann habe ich das geklärt.

Bevor wir uns darum kümmern, dass da eine Zeile zu viel ist, haben wir ja noch die Theorie.

Was sind eure Ideen dazu? Ich glaube, wir machen einen neuen Schritt, Aufbereitungsschritt. Also jetzt müssen wir uns fragen, jetzt nehmen wir wieder das Beispiel, fehlt euch hier wieder, da fehlt das ja noch. Wie kommen wir dazu, dass hier genau diese Werte stehen? Also wir machen jetzt wieder gleich,

berechnet das Feld erstellen, gleich. Wir müssen den Filter wieder wegmachen gleich, aber dann können wir das machen. Und dann müssen wir wieder eine Formel eingeben. Lasst uns mal zusammen ein bisschen eine Idee entwickeln. Und dann lasse ich vielleicht mal selber ein bisschen probieren. Oder vielleicht machen wir das noch zusammen. Das ist beim nächsten Problem. Aber wir können also das Konzept. Wie könnte die Formel aussehen? Wo kommt dieser Wert her?

Also der ist jetzt noch nicht da, der soll jetzt gleich da erscheinen. Was ist die Logik? Ja, das stimmt in der letzten Zeile aber nicht. Also in den ersten zwei stimmt es. Solange dieser Wert plus ein Jahr kleiner gleich als das. So irgendwie. Lass uns mal gucken. Jetzt wird es irgendwie hinschreiben.

Also hier ist es einfach plus ein Jahr. Und hier ist es das Enddatum, weil wenn ich das hier plus ein Jahr nehme, dann bin ich beim 8.2.2023 und das schlägt über das Ende meines Vertrages hinaus. Das müssen wir in eine Formel gießen. Ist die Logik für alle ungefähr klar?

Ich mache das jetzt wieder weg, weil es ziemlich unübersichtlich aussieht. Also irgendwie brauchen wir sowas. Ist Period Start plus ein Jahr. Was hast du gesagt? Kleiner gleich. End. Dann Period Start plus ein Jahr, oder? Ja. Else. End, oder? Ja.

Also wenn das Ende früher eingetreten ist, als wenn wir hier ein Jahr drauf rechnen, dann nehmen wir das vorzeitige Ende. Also das flägt ja immer dann zu, wenn es ein vorzeitiges Ende gegeben hat. Okay, jetzt fangen wir das hier ein. Ich mache den Filter hier weg. Was ist jetzt los? Genau, also ihr müsst irgendwie den Filter wegkriegen und dann gehen wir wieder hier auf berechnetes Feld erstellen. Wir erstellen also wieder eine neue Spalte und wir nennen sie Period End.

Am Ende soll das alles so schön aussehen hier auf der Tafel. Mal gucken. Also, if. So, jetzt. Hier mit Start plus ein Jahr. Das können wir nicht so hinschreiben. Aber wir können eine Funktion benutzen. Die Funktion heißt DateAdd. Und da helfe ich euch jetzt mit der Syntax. Oder ihr schreibt einfach mit. Also, DateAdd seht ihr hier. Und jetzt kann man hier sehen, wie man die benutzt. Also, ich kann hier zum Beispiel angeben, in welchen Schritten ich...

hinzuaddieren will. Da könnte ich jetzt mal auf hier gehen und dann ist das hier sozusagen wie viele, in dem Fall Monate, bei mir dann Jahre. Da würde ich dann 1 sagen und hier kann ich zum Beispiel dann Period Start eintragen. Also Date Add hier, 1, also rechne ein Jahr drauf auf Period Start. So, was habe ich gesagt? Wenn das kleiner gleich

Kleiner gleich oder kleiner, ja, machen wir mal kleiner gleich. End ist, dann... Und dann nochmal das Gleiche. Date, Add. Es gibt noch ein kleines Problem. Also, wieder ein Ja. Bin ich zu schnell wieder? Else, End. End. Nee. So. Okay. Findet der nicht gut. Gott sei Dank weiß ich auch schon, warum. Also, das Problem ist...

Ich glaube, eigentlich ist es komisch. Er erwartet hier eigentlich, dass ein Datum rauskommt. Und wenn ich hier so ein Date-Add anwende, dann kommt kein Datum raus, sondern eine Kombination aus Datum und Uhrzeit. Und deswegen werde ich dem jetzt noch eine Funktion Date voranstellen, die diese Datums-Uhrzeit-Kombination einfach nur in Datum umwandelt. Und dann ist er zufrieden.

Okay, jetzt gebe ich euch kurz Zeit, nochmal drauf zu starren. Also ich sage nochmal, was da passiert. Wenn hier jetzt Datus ein Jahr kleiner ist als vor dem Enddatum liegt, dann nehme ich das. Also ein Jahr drauf. Wenn nicht, dann nehme ich das vorzeitige Enddatum. Also da, wo gekündigt wurde. Ich hätte mir gar nicht zuhören müssen. Wie ich das hinkriege, das ist die Formel. Also nochmals. Ich bin...

Also dass ihr in die Brut kommt. Genau, wir helfen euch gerne gegenseitig. Wie schaut euch das? So lange haben wir nicht selbst gekommen, wenn es andere so nicht war.

Gut, ich zeichne es auf. Ihr könnt es ja nachspielen. Hier bei euch dreien. Erste Reihe. Dennis kriegt Support. Wir machen es gleich nochmal. Nachvollziehbar.

Vielen Dank.

...

Ja.

Ja, das ist ein ganz interessantes Thema. Ja, das kann ich auch so sagen. Ja, das ist ein ganz interessantes Thema.

Ja, das ist ein guter Punkt.

...

Ja, das ist eine gute Frage.

Vielen Dank.

Ja, ich glaube, wir sind jetzt in der Folge, wir sind in der Folge, dann geht es um die Folge, dann geht es um die Folge, dann geht es um die Folge.

Ich weiß nicht, ob ich das so sagen kann. Ich weiß nicht, ob ich das so sagen kann.

Ja, wir wissen, dass wir hier nicht in der Lage sind.

Vielen Dank.

Vielen Dank.

Ja, das ist es.

Vielen Dank.

Genau.

Vielen Dank.

Vielen Dank.

Okay, jetzt. Das ist immer gut, wenn man weiß, was man da eigentlich tut. Also, was haben wir hier?

Wir wollen berechnen, wann das Vertragsjahr geendet hat. Und grundsätzlich endet das Vertragsjahr ein Jahr nach dem Beginn. Das heißt, grundsätzlich mal sagen wir, wir rechnen auf den Period Start ein Jahr drauf. Und das machen wir dann, wenn wir dabei nicht hinter dem Vertragsende landen. Also in den Fällen, wo vorher gekündigt wurde, landen wir, wenn wir ein Jahr drauf geben. Also wenn wir hier

ein Jahr drauf geben, dann sind wir bei 28,23. Und das liegt nach dem Ende des Vertrages. In dem Fall nehmen wir dann das Vertragsende, das vorzeitige. Das ist das, was da steht. Macht Sinn?

Anna? Ich bin einfach noch einmal gefragt, bei der zweiten Zeile hat man ja noch von Date-Ad noch ein Date hinzugeschrieben. Wie hat man das getan, damit die Datum-Datum ein Datum ist? Genau, also wenn du es nicht machst, wir können auch noch einmal kurz bei Date-Ad gucken, hier steht, was rauskommt, da siehst du Datum plus Uhrzeit, 12 Uhr mittags. Aha.

Also wenn ich auf ein Datum Date Add anwende, dann kommt eine Date-Time-Kombination raus und das muss ich dann wieder konvertieren zu Date, wenn ich Date haben will. Okay, gut. Jetzt schauen wir nochmal unsere Beispielkunden an, um zu gucken, was jetzt da passiert ist. Also wir sehen jetzt, holen wir uns mal das Period End an die richtige Stelle. Ich finde, die richtige Stelle ist hier. Also

Wieder filtern. Jetzt haben wir Period Start. Das ist eigentlich das, was an der Tafel steht, die ersten drei Zeilen. Und da haben wir noch diese letzte, da hatte ich gesagt, müssen wir uns noch drum kümmern. Also was wir machen können, die einfachste Art, das zu tun, ist jetzt wirklich, ich weiß auch nicht, wie ich es besser hinkriege, hier gibt es Werte filtern. Wir hatten das gerade schon. Das ist der Unterschied, wenn ich hier

bei Contract ID auf Filtern gehe, dann geht jedes Mal wieder weg. Habt ihr schon bemerkt. Jetzt kann ich aber auch wirklich filtern. Also ich kann wirklich Zeilen löschen. Und das ist das, was passiert, wenn ich hier drauf gehe. Und hier kann ich eine Bedingung eingeben. Also jetzt muss man es noch richtig rum haben im Kopf. Die Bedingung, die ich angebe, die sagt, welche Zeilen übrig bleiben sollen. Also ich möchte, dass Period Start ungleich Period End ist. Das ist meine Bedingung, weil

Bei dieser letzten Zeile, die wir gerade gesehen haben, da waren es gleich. Da war es beide Male 28.05.2022. So, jetzt habe ich was falsch gemacht. Ich habe beide Male Period End eingegeben. So. Gucken. Und jetzt müssen wir auch ein bisschen...

genauer noch gucken, jetzt ist mal alles gut, für das Beispiel, wo der Vertrag tatsächlich gepublikt wurde. Jetzt haben wir genau die drei Zeilen, die wir auch an der Tafel haben. Gibt es bei euch? Dann hast du es weggeschildert. Also, der Filter ist Period Start und dann kleiner, größer. Das heißt ungleich, Period End. Gut.

Also übrig lassen soll er alle Zeilen, bei denen Period Start nicht das gleiche wie bei Period Start. So, hier war er jetzt. Er sieht so ganz entspannt aus.

Ich habe noch eine Frage. Kann man irgendwo nachschreiben, was man als Wissen beinhaltet, wie die Wissen von den Menschen sind?

Sag es nochmal. Es gibt keine Übersicht, aber was du siehst, ist, dass über dem Schritt so ein Filtersymbol entstanden ist. Das bedeutet, dass da ein Filter ist und wenn du dann auf Änderungen gehst, dann kannst du die Filter sehen. Also ich glaube nicht, dass es eine Übersicht gibt. Ich wüsste nicht, wo, aber so findet man sie. Anna?

Das war beim Josef auch so.

Also ich glaube, in der nächsten Pause werde ich sowieso so machen, dass ich mal meinen Stand hochlade, obwohl, dass ich mich mal speichere. Und dann könnt ihr alle darauf aufsetzen. Dann müssen wir nicht alles wieder zurückverfolgen. Ich habe...

Ja. Ja, vielleicht. Wir gucken es gleich an. Einfach jetzt zu gucken. Ja. Okay, jetzt lasst uns noch einen Check machen für diese anderen Vertragsregeln und schauen, was da noch so steht. Mal schauen, ob ihr den findet, wenn ihr fehlt habt. 5575...

Da hat es begonnen am 16.05.2021. Das ist das erste Vertragsjahr, das zweite Vertragsjahr, das dritte Vertragsjahr. Und jetzt ist hier die Frage, wir haben jetzt hier noch eine vierte Zeile. Das haben wir ja so gewollt, dass wenn nicht gekündigt wurde, dass dann das Enddatum der 31.12.2024 ist.

Das ist immer dann der Fall, wenn dieser Vertrag eigentlich jetzt gerade noch läuft oder wenn wir annehmen, dass wir die Analyse Ende 2024 gemacht haben, dann waren die alle offen. Das heißt eigentlich, wenn wir jetzt gleich, das wird unser zweiter Schritt sein, wenn wir dann irgendwann endlich erfolgreich den Status kreiert haben, unser Klassenattribut zu kreieren, dass

Was ist dann das Kassenattribut für diese Periode hier? 16.05.2024 bis 31.12.2024? Okay, bis dahin nicht gekündigt. Aber eigentlich ist das Vertragsjahr erst am 16.05.2025 zu Ende. Also in der Zukunft. Das heißt, wir wissen es nicht. Also mein Vorschlag wäre, die Zeile zu löschen. Oder alle Zeilen, bei denen Period End 31.12.2024 ist.

Weil wir es da noch nicht wissen. Also, dann machen wir noch einen Filter hin, weil wir schon dabei sind. Werte filtern. Bedenkt wieder, wir müssen eine Bedingung angeben dafür, für das, was übrig bleiben soll. Also, das Period End soll nicht ungleich im, das kennen wir schon, 2024 minus 12 minus 31 und davor ein Date. Okay.

Mal schauen, was wir jetzt bewirkt haben. Ich suche wieder nach diesem Vertrag. Hat jetzt nur noch drei Zeilen. Also das letzte Vertragsjahr, von dem wir noch nicht wissen, wie es ausgehen wird. Und wir schauen jetzt aber auch noch sicherheitshalber nach der anderen Kundin. 7892. Wer es noch hat, da hat sich nichts verändert. Das ist gut. Also weiterhin drei. Das liegt ja alles in der Vergangenheit. Da wissen wir von allem, ob wir...

eine Cancellation bekommen haben oder nicht. Ich habe gesagt, es wird hart. Aber jetzt können wir einen Haken machen. Wir haben jetzt erfolgreich den Chancen. Wir haben jetzt für jedes Vertragsjahr eine Zeile. Ist das cool? Was wollen wir als nächstes machen?

Ich würde sagen, wir machen es so, ja, nee. Ich würde vorschlagen, wir machen es so, wie wir es letzte Woche gemacht haben. Klassener Punkt erst und dann die Abzüge. Okay, wir machen Klassener Punkt. Jetzt wollte ich eigentlich euch das machen lassen, aber vielleicht machen wir es nicht im Tool, weil jetzt nicht alle auf dem gleichen Stand sind. Einige sind ausgestiegen. Lasst uns einfach drüber nachdenken. Also,

Wie wollen wir das nennen, das Klassenattribut? Hier habe ich geschrieben "cancellation". So können wir es nennen, ja. Sollen wir es so nennen? Okay, was müssen wir machen? Also wir haben jetzt die ganzen Daten. Wir haben "period start", "period end" und hier hinten ist auch immer noch die Spalte, wann die Kündigung bei uns eingegangen ist.

Lass uns erstmal hier überlegen, was soll denn da rauskommen? Also unser Kassenattribut soll Cancellation heißen. Wir machen das jetzt für 7892.p.o.k.p. Was sollen hier für Werte hin?

Ja, oder yes or no. Aber was konkret jetzt? Diese erste Zeile zum Beispiel. Also das wurde nicht gekündigt. Zweite Zeile? No? Dritte Zeile? Okay. Das ist die Logik dahinter. Das ist doch die Frage. Eigentlich wollte ich jetzt

Kurz war und allen die Chance geben, sowas zu probieren, aber kein Publikum. Beim nächsten Mal. Sind alle dabei? Also müssen wir wieder... Okay, lass uns mal jetzt eine Formel hinschreiben, so grob. Oder irgendjemand anders vielleicht mal, so mit If und Else. Was hat der Dennis gesagt? Weiß noch jemand?

Also es muss jetzt rausgehen, falls end, period, end ist. Also ich fange mit if an, period, end, gleich end. Then yes, hast du gesagt? Ja. Und sonst no, oder? Ja, end. Das ist doch gar nicht so schwierig. Und dann noch end. Okay, lass uns mal probieren. Sollen wir einen neuen Aufbereitungsschritt machen, oder? Ja, komm.

Komm, wir machen einen Aufbereitungsschritt und dann mache ich hier umbenennen und schreibe Klassenattribut. Das ist Dokumentation, so wie wir es bisher auch schon gemacht haben. So, jetzt müssen wir auch unser Wort halten, berechnetes Feld erstellen. Also ich nenne es Cancellation und dann tippe ich das im Wesentlichen ab. If period end gleich end. Jetzt bin ich gerade selber unsicher, ob ich ein gleich gleich brauche. Ich glaube nicht. Dann ja, in

Anführungszeichen, sonst nein. Das ist schon mal gültig. Mal schauen, was passiert. Wir filtern wieder. Wer noch kann? Okay, wir haben drei Zeilen. No, no, yes. Perfekt. Hat funktioniert. Und schauen noch nach dem anderen Vertrag. Was muss da rauskommen? Genau, jetzt waren noch drei übrig. Und wir haben bisher immer noch keine Kündigung bekommen. Das heißt, dreimal no. Ja, so ist es.

Also es scheint zu funktionieren. Wir können noch einen Haken machen. Jetzt läuft es richtig schnell vorwärts. Also die schlechte Nachricht ist bei den Attributen, wenn wir uns noch ein bisschen mehr ausprobieren. Aber das Klassenergebnis haben wir jetzt. Also theoretisch könntet ihr das jetzt schon exportieren und in Orange laden und versuchen irgendwas zu lernen. Aber wir wollen noch bessere Attribute haben. Ja.

Ja, ja, ja, ich höre dich. Ich habe noch kurz eine Frage zu den Aufbereitungsstätten. So willst du feststellen, dass wir... Nein, es sind schon harte Modifikationen. Der Workflow wird damit enden, dass wir Resultate exportieren, also Ausgabe. Und da kommt dann...

CSV oder Excel-Datei raus, die wir in Orange laden können. Und die Modifikation, die wir machen, die sind jetzt da. Und diese DS-Controls aus Chrome, die ist quasi auf der Voraussetzung. Ja, also du lädst die Daten vorne rein und dann wird es sozusagen im Arbeitsspeicher alles modifiziert und am Ende hinten raus wieder in die Datei geschrieben. Genau, meine Frage ist dann, was auch...

Ich versuche es ein bisschen sauberer zu machen. Es geht wahrscheinlich viel besser. Aber ich versuche jetzt nicht berechnete Felder im Neue-Zeilen-Schritt zu erstellen. Was glaube ich auch geht. Du kannst filtern und alles geht auch, aber ich nehme dann immer einen Aufbereitungsschritt.

Und ich habe jetzt auch einen separaten fürs Klassenattribut angelegt, damit man so ein bisschen klar bleibt, was passiert da. Also ich könnte jetzt hier zum Beispiel noch sagen, bei Neue Zeilen Instanzen erstellen. Irgendwie so. Naja, passt schon wieder nicht hin. Ja, geben wir es auf. Okay, soweit sind wir. Das heißt, wir würden jetzt zu den Attributen kommen. Vielleicht nochmal kurz Blick auf die Folien. Die falsche. Also Instanzen konstruieren.

Haben wir gemacht. Manchmal, das ist jetzt bei uns nicht der Fall gewesen, kommt schon zum Erstellen der Instanzen, muss man schon aus mehreren Tabellensachen kombinieren. Das war jetzt bei uns nicht der Fall. Wir hatten eine Datei, wir haben Zeilen aufgesplittet. Also ich will nicht sagen, dass das, was hier auf den Folien ist, auch alle Fälle abdeckt, die passieren können. Okay, jetzt haben wir...

den Fall noch abgedeckt hier beim Klassenattribut, dass der Wert des Klassenattributs sich so ein bisschen implizit ergeben hat. Also es war implizit, halb implizit. Wir hatten das Cancellation Received, das heißt, ja, und das, da hatten wir vorher schon ein paar neue Attribute konstruiert, aber letztlich, letztlich zurückführen, war Cancellation Received irgendwo null oder nicht. Wenn es überall, also wenn es null ist, am Anfang gewesen ist, dann sind sowieso alle Vertragsjahre mit No

Und wenn da ein Datum steht, dann ist das letzte Vertrag ja ein Yes. Also das war jetzt ein bisschen komplex, habe ich hier so genannt. Ja, jetzt werden wir noch verschiedene Arten von Herausforderungen kennenlernen. Ich versuche euch vorzubereiten. Wir haben ja auch ein Semester-Assignment und da werde ich euch auch Daten geben. Lustigerweise sind es auch drei Dateien, glaube ich, wenn ich mich recht erinnere.

Und da werdet ihr dann vielleicht nicht alle von den Sachen machen müssen, die wir heute machen, aber einen Teil davon, sodass ihr es schon mal gesehen habt. Und wir zeichnen es ja auch auf, also nicht nervös werden jetzt, wenn ihr nicht alles gleich mitschneidet sozusagen. Dadurch, dass wir es aufzeichnen, könnt ihr auch nochmal, wenn ihr dann am Assignment arbeitet, zurückgehen und gucken, wie haben wir das gemacht und ist das hier vielleicht das Gleiche und kann ich das anwenden. Okay? Gut.

Jetzt habe ich mir hier aufgeschrieben, dass wir das nächste eventuell weglassen. Aber wir können vielleicht kurz darüber reden. Also ein Attribut, was ich weiß nicht mehr, ob es letzte Woche mit euch war oder mit der Freitagsklasse. Ein Attribut war zum Beispiel Dauer der Beziehung mit dem Kunden. Wie würdet ihr das ausrechnen? Wir machen das jetzt nicht, wir überspringen es. Nur, dass wir es kurz erwähnt haben. Also wie lange ist unsere Beziehung...

Wenn wir jetzt zum Beispiel diese Instanz hier nehmen, dieses Vertragsjahr mit dieser Kundin, wie lang war da unsere Beziehung zu Beginn des Vertragsjahres? Da hatten wir schon ein Jahr hinter uns. Also zwölf Monate hatten wir da schon. Und in diesem Jahr hatten wir dann schon 24 Monate Beziehung, also Kundenbeziehung mit dieser Kundin. Also das könnten wir auch aussetzen. Als Beispiel für Datumsangaben sind sehr häufig Beziehungen.

eine gute Quelle, um sich neue Attribute zu konstruieren. Wenn man darüber nachdenkt, was für Attribute könnten hilfreich sein. Aus Datumsangaben, ja, man kann einerseits gucken, wie lang ist etwas schon gewesen oder in welche, ja, ist es Sommer oder Winter? Also Monat lässt sich ableiten. Wochentag ist manchmal wichtig. Ist es Wochenende oder unter der Woche? Das hat oft einen Einfluss auf bestimmte Sachen. Oder war gerade Ferien. Genau. Was sehr häufig auftritt,

dass wir Informationen aus anderen Informationsquellen, aus anderen Tabellen zum Beispiel noch dazu laden wollen. Das erspare ich euch jetzt nicht. Kann auch gut sein, dass es euch begegnet im Semester Assignment. Genau, die Aggregation, die ist da gleich mit integriert. Manchmal müssen wir Sachen auch zusammenrechnen oder zusammenzählen. Worauf will ich hinaus? Also, habt ihr eine Idee? Jetzt bin ich im falschen Kurs.

Was könnten wir denn noch für Attribute konstruieren? Vielleicht auch, wenn wir uns nochmal anschauen, was eigentlich unsere Input-Daten sind. Da haben wir noch nicht so genau hingeguckt. Also, was wir da haben, außer der Datei, die wir jetzt geladen haben, sind noch zwei weitere Dateien. Services, da seht ihr immer eine Vertrags-ID und dann seht ihr hier zum Beispiel die Vertrags-ID 9305, tritt viermal auf.

mit verschiedenen Services. Das bedeutet also, dieser Vertrag nutzt oder hat diese Services dazu gebucht. Und hier sehen wir Tickets. Also hier zum Beispiel 0280 hat insgesamt vier Tickets gehabt. Alle im März 2023. Da war irgendwie der Grund drin. Okay. Inspiriert uns das? Wollen wir damit noch was machen? Dennis. Willkommen. Eingriff in den Bündnis. Börsenservice. Wie viele?

Also lass uns die zwei nicht unbedingt zusammen. Also welcher Service, ich glaube, das kriegen wir nicht hin. Aber welcher Vertrag gibt dir die Tickets? Also wir könnten das hier machen, oder? Wie viele Tickets? Genau.

Ja gut, wollen wir das machen? Habt ihr noch Kraft? Wollen wir erst Pause machen? Pause? Dann, ja genau, dann mache ich es so, wir sagen mal bis 15 Uhr 25. Ich speichere jetzt das hier und lade es auf Moodle und dann könnt ihr, können wir alle damit weitermachen. Dann sind wir wieder alle auf dem Stand, okay? Auch ihr könnt euch das noch schön runterladen und dann wieder mitmachen. Okay.

Gut, dann bis in einer guten Viertelstunde. Ja, was auch immer. Wir leben. Wir gehen. Wir gehen jeden Tag. Wir gehen heute. Am 20. ist es anders. So, kommt. Wir freuen uns doch darauf, weiterzumachen. Wieso gibt es kein Bild? Warte mal. Okay, jetzt wollten wir also die Tickets dazunehmen. Dann machen wir das direkt.

Ihr habt alle runtergeladen, sofern es nötig war. Okay, dann machen wir jetzt folgendes. Wir holen uns die Tickets. Also wenn ich hier klicke und wenn ihr alles an die gleiche Stelle runtergeladen habt, dann ist es bei euch vermutlich auch so. Dann geht es schnell. Wir holen uns die Datei 02-Tickets.csv. Wollen wir mal erst einen Aufbereitungsschritt machen? Ich weiß gar nicht, ob wir den brauchen. Ich glaube nicht. Ja.

Okay, jetzt können wir auch hier tatsächlich für diesen Vertrag, den wir immer als Beispiel nehmen, filtern. Okay. Was fällt euch auf? Zwei Einträge. Ich habe nicht mehr so wahnsinnig viel Platz auf meiner Tafel, aber das kriege ich noch hin. Also, was wollen wir? Wir haben drei Instanzen für 7892EUQP. Warte mal, Dennis. Ähm,

Wir haben drei Instanzen hier und wir wollen ein neues Attribut. Das heißt, dieses neue Attribut, Anzahl Tickets, das braucht für jede Instanz irgendeinen Wert. Jetzt seht ihr hier, dass es zwei Tickets gab. Eins am 17.02.2022 und eins am 28.05.2022. Am letzten Tag, da war jemand da.

Das ist ein Zunuss, jetzt war ich das jetzt gönnig. Was sind die Werte, die hier hingefahren sind? Ah, ich würde es dir einfach mal zahlen. Also was meinst du mit Ticketnummer? Du meinst diese Nummer? Ne, es gibt gar keine Nummer. Also wenn ich da so schreibe Anzahl Tickets, dann würde ich einfach nur zählen.

Also ich habe jetzt gedacht, wir können sozusagen diese Vertragsjahre dadurch beschreiben, dass wir gucken, wie viele Tickets gab es in dieser Verwendung. Willst du was fragen, Dennis? Nein. Komm, wir geben mal den anderen den Spaß. Also was würdet ihr, was für Werte solltet ihr einfahren? Eins, zwei, drei, vier, fünf, sechs.

Das sind alle Daten, die wir in dieser Tickets-Datei haben, zu diesem Vertrag. Also ich habe jetzt die Tickets-Datei gefiltert nach der Vertragsnummer. Das heißt, wir können jetzt konkret die richtigen Zahlen da eintragen.

Ah ja, genau. Also sozusagen alles ist 17.2. auch nach dem Period Start vom 3. Konntest du das gleiche sagen, Dennis? Ist klar für alle? Okay.

Okay, irgendeine Idee, wie wir das jetzt hinkriegen? Generell müssen wir da jetzt am Ende eine Datei produzieren. Da steht dann in diesen Zeilen die Zahl am Ende noch mit drin. Das heißt, wir müssen jetzt zwei Dateien miteinander verknüpfen.

Ihr habt doch schon so ein bisschen was über Datenbanken gelernt, oder? Ja, Dennis? Ja, genau. Wir machen jetzt ein Join. Und das, was du gerade gesagt hast, also hier heißt es verknüpfen, es sei denn, ihr habt die englische Version, also Left Join wollt ihr machen? Okay, gucken wir mal. Ne, was hat der jetzt gemacht? Sorry, da muss man immer ein bisschen aufpassen. Also

Ich mache so, dass ich auf das Plus klicke und das dann hier oben drauf ziehe. So, jetzt passt. Jetzt ist da ein Ausrufezeichen, weil wir noch diverse Sachen definieren müssen. Also erstens mal die Arthus-Joint. Hast du gesagt, Katharina, Left-Joint? Warum? Oder was ist das? Also, wenn... Ja, was mache ich? Also, ich habe...

Also ich klicke auf das Plus, halt gedrückt und dann ziehe ich und dann, ich darf es erst loslassen, wenn das Ding in der Mitte auch rot wird. Das ist ein bisschen tricky. Nicht zu früh loslassen. Ja, warum Left Join? Was ist denn Left Join? Also gut, denn es ist ja eine Umscheinung. Genau, also wir wollen alle Vertragsjahre behalten, auch wenn es keine Tickets gab. Gut. Also machen wir Left Join. Das könnt ihr hier einstellen, indem ihr auf das linke Ding hier klickt.

Jetzt ist es ein Left-Join. Was wir jetzt noch brauchen, ist die Bedingung für den Join. Also hier. Jetzt habe ich es wieder weggenommen. Jetzt ist es ein Inner-Join. Seht ihr diese Kreise hier? Die müsst ihr bei euch auch geben. Also ursprünglich ist es ein Inner-Join. Nur, dass es in der Mitte grau gefärbt wird. Und ich klicke dann auf den linken Kreis, dann wird es zum Left-Join. Das ist nicht intuitiv. So, also...

Genau. Jetzt steht hier auch Verknüpfungstyp links. Tja, hat es ja gegeben. Seid ihr soweit für die Join Condition hier? Also, wir können auch mehrere Verknüpfungsklauseln. Ich glaube, Dennis, du hattest auch schon mehrere genannt, oder? Also, ich denke mal, wir brauchen auf jeden Fall auf beiden Seiten die gleiche Contract ID, oder? Ja. Soweit ist schon mal klar. Und jetzt, also wenn wir nochmal zurückgehen und hier gucken, die Tickets haben ja ein Datum.

Und wir haben das Datum irgendwie benutzt. Also als wir jetzt gerade, oder Dalia, als du das gesagt hast, welche Zeile welchen Wert haben soll, hast du aufs Datum geschaut und hast gesagt, diese zwei Tickets sollen der dritten Zeile zu beobachten werden. Das können wir jetzt irgendwie in unseren Join einbauen, oder? Also wir können sagen, dieses Ticket soll zu einer Zeile dazu gejoined werden, wenn, und lass jetzt nehmen wir irgendwie das Datum, es muss dazwischen liegen, oder? Zwischen dem und dem.

Also hier ist es nicht der Fall, die zwei Daten, aber hier ist es der Fall. Jetzt wird es noch mal ein bisschen kompliziert, weil man das mit dem Kleiner und dem Größer hinkriegen muss. Also wenn, und ich fange mal hier rechts an, ja, lass es uns von rechts lesen. Also das Ticket Date soll größer als Period Start sein, richtig? Oder größer gleich? Und jetzt muss ich, wenn ich so rumlese, dann habe ich, also von rechts nach links habe ich jetzt ein Größer, ja. Also ich lese ja

Lest von hier, Ticket Date, größer gleich, Period Start. Und jetzt kommt noch eine dritte Bedingung. Wieder soll ich Ticket Date, soll aber kleiner als Period End sein, oder? Also jetzt muss ich mir wieder hier von rechts nach links lesen, kleiner, Period End. Jetzt habe ich diese drei Bedingungen hier. Ich gucke, dass die Tickets zum gleichen Vertrag gehören und dass sie innerhalb von dem Vertragsjahr liegen, was ich gerade anschaue.

Ja, das ist so ein kleiner Platz, denn wir haben ja durch den V-Aus und die Translation-Base auch ein Ticket dran gebrochen. Ja, das stimmt. Lass uns lieber so machen. Ja, jetzt ist es eigentlich auch nicht optimal. Wir ignorieren das Problem mal. Weil natürlich jetzt, wenn eins gerade auf der Grenze liegt und danach kommt noch ein Jahr, dann wird es zweimal gezählt, wenn es dumm läuft. Dann machen wir hier einen größeren. Also einen kleiner. Jetzt sollte jedes Ticket nur zu einem Jahr gehören, wenn es gut läuft.

Okay, sollen wir mal schauen, was rauskommt? Jetzt passiert natürlich was doofes. Ich weiß nicht, ob ihr es schon ahnt. Wir hatten doch vorher drei Zeilen. Wo bin ich? Also jetzt sind hier die Felder dazugekommen und jetzt, wenn ich mir das genauer anschaue, ist es gar nicht so einfach. Ich sollte die vielleicht wieder umsortieren, die Felder. Aber ich sehe jetzt hier diese zwei Tickets. Die haben jetzt, das passiert beim Join sozusagen, wenn ich...

auf der rechten Seite zwei Einträge habe. Die sind jetzt aus dem letzten Vertragsjahr zwei Zeilen geworden. Ja, Dennis? Genau. Okay, also was wir machen können, bevor wir was zählen, das ist vielleicht sinnvoll, wir können mal ein Attribut definieren.

wir nennen es vielleicht num tickets und wir könnten sagen wenn hier also wenn jetzt 0 sozusagen dann tragen wir die zahlen 0 ein damit es dann am ende auch eine 0 steht wenn das summiert wird und wenn dann datum steht dann 1 und dann können wir einen aggregierenden schritt nehmen ich zeige euch mal was ich meine also wollen wir einen aufbereitungsschritt machen ist doch sauberer dann machen wir berechnet das feld ich nenne es mal num tickets

Oder Tickets, am Ende wird es vielleicht Tickets heißen, das Aggregierte. So, und jetzt sage ich if, also wenn es 0 ist in der Zeile, was ist das? Ticket Date, oder? Dann mache ich 0 und sonst 1. Macht Sinn? Also wenn da ein Datum steht, ist ja ein Left Join sozusagen. Wenn ich zu einem Parktagsjahr kein Ticket finde, dann steht da 0. So, wir gucken nochmal, was jetzt passiert ist.

Mit der Formel, ich mache sie gleich nochmal auf, keine Sorge. Ich filtere, übrigens habe ich jetzt zwei Contract IDs, die zweite lösche ich mal. Ich filtere jetzt wieder für unsere Kundin hier und gehe wieder nach rechts. Also ich sehe jetzt sozusagen in den ersten zwei Vertragsjahren, wo es kein Ticket gab, habe ich jetzt für das neue Attribut eine Null und jetzt habe ich diese zwei Zeilen, die will ich jetzt gleich aggregieren und die zwei Einsen aufsummieren. Dann sollte da

die zwei auskommen. Genau, und jetzt können wir hier aggregieren machen. Seid ihr soweit? Überhaupt nicht. Okay. Dann mache ich nochmal zurück. Wo hängt, soll ich nochmal die Formel aufmachen? Also, ich sage nochmal, wenn das Ticket Date 0 ist, also if ist 0 Ticket Date, dann machen wir eine 0 rein und sonst eine 1. Läuft?

Okay, aggregieren, wie funktioniert das?

Also wenn ihr schon, ihr habt schon ein bisschen SQL gemacht, oder? Das ist so ein bisschen wie Sum und Group By. Also jetzt muss ich das Group By sozusagen definieren. Hattet ihr sowas? Group By und Sum? Okay. Also was will ich aggregieren? Was soll summiert werden? Das muss ich hier reinziehen. Was will ich addieren? Genau, und dann Neues.

Ding, num tickets, wo ist es? Hier. Nehmen wir jetzt, ziehen es hier rüber. Okay. Und jetzt ist es ein bisschen speziell, bei Tableau Prep muss ich tatsächlich jetzt alle Felder, die ich behalten will, bei gruppierte Felder reinziehen. Jetzt muss ich aufpassen. Ah, ist okay. Ich glaube, es ist okay. Nee, Ticket Date nicht. Ticket Date lasse ich jetzt hinter mir. Das will ich ja, das brauche ich nicht mehr. Ja, also ich zoomiere jetzt über die verschiedenen Tickets.

Dabei fällt das Ticket-Date raus. Ja, wenn ich danach auch gruppieren würde, dann würde es die nicht zusammen klatschen, die Zeilen mit den verschiedenen Tickets. Mal schauen, was passiert. Ich filter mal wieder. Ich glaube, es stimmt. Wenn ich es nach rechts ziehe, nein, kann ich nicht. Warum auch immer. Ich suche mal nach Period Start. Ist mir das abhanden gekommen.

Also das hier ist das erste Vertragsjahr, das ist das zweite, da haben wir jeweils eine Null stehen und im letzten Vertragsjahr haben wir unsere zwei. Also ich würde sagen, das ist ein relativ häufiger Fall, wenn man Attribute konstruiert, dass man, ja, so eine End-zu-End-Beziehung haben wir hier. Habt ihr über sowas gesprochen bei Datenbanken? Ja, es gibt ein Wiedererkennen. Also, ne, eigentlich 1 zu N, das End-zu-End-Beziehung.

Also jedenfalls kann ein Vertragsjahr mehrere Tickets haben und ich möchte das dann einfach aggregieren, diese verschiedenen Tickets. Und ja, also sowas passiert oft, dass man zu einer Instanz, die man kreieren will, zu einer Instanz mehrere Ereignisse oder irgendwas hat in einer anderen Tabelle und dann...

die einfach zählt oder wenn die noch irgendeinen Wert haben, zum Beispiel einen Preis oder so oder eben ein Verdienst, die ich dann auch summiere. Okay, jetzt habe ich noch ein Challenge für euch. Das letzte jetzt. Ja, unbedingt. Ja,

Das heißt, dass es neun verschiedene Werte gibt. Also das Maximum, was du siehst, sozusagen maximale Anzahl Tickets in einem Vertragsjahr waren acht Tickets. Und die meisten... Ah, das ist ja nur unser Beispiel. Also das ist sozusagen tatsächlich über alle... Also ich glaube, das passt sich nicht dem Filter an. Ja, das bleibt... Das bleibt...

Ja. Wie denn? Das stimmt schon. Was ich gemacht habe, ist tatsächlich, ich habe hier wirklich von links alles rübergezogen, außer dem Ticket-Date und der Anzahl an Zeilen. Alles, was hier noch ist, muss hier rein. Weil du das alles behalten willst. Das ist bei Tableau Prep so.

In SQL würde es ein bisschen anders aussehen. Ja, ich meine, letztlich ist das hier sozusagen dein Group-By-Clause. Du könntest einfach in SQL, das ist einfach ein Group-By-Contract-ID, and Period Start oder so. Und hier nimmst du einfach alle rein, die du behalten willst. Und hier das, was du aggregieren willst.

Und das Ticket-Date, hat sich erklärt, wollen wir nicht behalten, weil wir ja über Tickets aggregieren mit verschiedenen Daten.

Also solange die Werte da unten stimmen, bin ich besorgt. Also ich habe jetzt gefiltert, ne? Nach unserem Contract-ID. 7892. Wenn du nicht gefiltert hast, dann stehen da natürlich überall andere Werte.

Also wichtig ist jetzt sozusagen, dass du hier zweimal eine 0 und einmal eine 2 hast, wenn du einen Filter hast. Okay. Ich bin einfach aus Interesse. Weshalb hat es die Reihenfolge verändert, dass das letzte Jahr in der Mitte ist und nicht mehr am Plus? Keine Ahnung. Ich weiß wirklich nicht. Ja. Bei der Landkreiszeit, die haben sie auch 200. Wo haben die schon da? Jetzt hier? Ja.

Also hier? Also ich habe die zweite Contract-ID, die beim Join mitgenommen wurde, die habe ich hier rausgenommen. Gut.

Seid ihr bereit für den nächsten Challenge? Der letzte.

Die bleiben ja übrig bleiben. Die übrig bleiben. Das verstehst du nicht, warum die übrig bleiben? Oder weil so eine Summe, also Tickets. Du kannst es probieren. Wenn du jetzt Ticket Date auch noch hier

rüber nimmst, dann wirst du für 7892 PUU KP wieder vier Zeilen haben. Weil er ja jetzt auch Group by Ticket Date macht. Mhm.

Das heißt, die zwei Tickets, die an unterschiedlichen Daten generiert wurden, werden jetzt nicht zusammen aggregiert. Deswegen nehme ich es raus. Das will ich aber auch nicht behalten. Das interessiert mich ja nicht mehr. Ich will es ja aggregieren über die zwei Daten hinweg. Ja, also hier oben kann ich wählen sozusagen, wie. Ich kann auch einen Durchschnitt oder sowas machen.

Aber wenn ich ein Feld hier rüberziehe, wird automatisch aggregiert. Und hier oben steht sozusagen sum als Funktion, als Aggregationsfunktion. Und hier ist der groupByClause, wenn du es in SQL übersetzen willst. Ah, sum, wenn man so anzeigen würde, alles. Ah, okay, ja. So, jetzt müsste es wieder passen. Noch ein letzter Check, bevor wir jetzt wirklich versprochen den letzten Challenge uns vornehmen.

Genau. Seid ihr bereit? Okay, jetzt haben wir ja noch die andere Datei. Was machen wir denn mit der? Services. Sollen wir das auch einfach dazu joinen? Das wird es nicht sein am Ende, aber wir probieren es mal als Anfang, oder? Also, wir laden. Textdatei. Services. Jetzt kommt die da unten irgendwo hin. Ich ziehe die mal hier rüber. Machen wir mal Aufbereitungsschritt. Ist immer gut. Okay.

Mal schauen, welche Services unsere Kundin hat. Okay, die hat fünf verschiedene Services. Kann schon jemand vorhersehen, was passiert, wenn wir jetzt einen Left-Join mit unserem oberen Workflow machen? Sollen wir was machen? Gucken, was passiert? Ja. Okay, also Join, Left-Join wieder, oder? Dem gleichen Grund, wir wollen niemanden verlieren, der keine Zusatz-Services gebucht hat. Also mache ich hier gleich ein Left.

Und jetzt ziehe ich das mal hier hin. Was ist unsere neuen Bedingungen? Wir haben ja auf der rechten Seite nicht so viel Auswahl. Wir haben eigentlich nur die Contract-ID. Also nehmen wir die, oder? Die muss gleich sein. So, Achtung, ich suche nach unserer Kundin. Es explodiert. Ja, das ist schlecht. Wie viele Zeilen haben wir jetzt? Man muss es nicht zählen. Man kann es eigentlich wissen. Wir haben drei Vertragsjahre und fünf

Services. Ich behaupte, wir werden jetzt 15 Zeilen haben. Ihr könnt es zählen. Gucken wir, ob das stimmt. Aber ihr habt für jedes Vertragsjahr jetzt 5 Zeilen bekommen. Also Period Start, wo ist es? Hier. Ihr seht, das sind 5 Zeilen für das erste Vertragsjahr, was 2020 gestartet hat und alle 5 Services. Dann kommt das nächste Vertragsjahr. Okay, das ist jetzt das letzte in dem Fall.

fünfmal und mittlere vertragsjahr auch fünfmal was können wir jetzt mal wieder irgendwas aggregieren also der trick den man macht man möchte ja jedes vertragsjahr wirklich nur in einer zeile haben jetzt habe ich leider keinen platz mehr aber was man macht ist wie viele services gibt es insgesamt

7 man macht sieben neue spalten ja man macht eine für die weiß protection eine für multiple lines eine für online backup und so weiter also ich kann jetzt vielleicht einfach hier unten weiter schreiben und das mal als spalten anliegen also device protection muss bisschen kleiner scheint multiple lines online backup online security

Streaming Movies. Ich mache einfach nur Movies und TV und dann noch Text Support. Das sind die ganzen Zusatzpackages, die man buchen kann. Gefiltert nochmal für unsere Kunden. Also wie stellt man das jetzt dar? Hier sehen wir, welche Services sie gebucht hat. Und man macht das jetzt einfach, man nennt das auch One Hot Encoding. Nee, stimmt nicht.

Man nennt das pivotieren, so heißt es auch in Tableau Pressing. Also wir haben jetzt aus den Werten dieser Spalte Service, diesen Attributs, haben wir jetzt jeweils Spalten gemacht. Und jetzt gucken wir einfach, aha, die Kundin hat Tech Support, also machen wir hier nur einzeln. Sie hat Multiple Lines, sie hat Device Protection und Streaming Movies.

und TV. Was sie nicht hat, ist Online Security und Online Backup. Da macht Ugo das. Und jetzt seht ihr sozusagen, wir haben jetzt in einer Zeile die gleiche Information, wie wir sie hier in fünf Zeilen haben. Das ist die Tricks. Also auf der Folie hier gibt es das auch beschrieben. Wir sind jetzt schon weit gekommen hier. Ich nenne es das End-zu-End-Problem. Jetzt ist es wirklich End-zu-End. Also

Ein Vertrag kann mehrere Services haben und ein Service kann zu mehreren Verträgen dazu gebucht werden. Und die Lösung fangen wir uns jetzt an, die Portieren. Hier ist übrigens noch das Aggregieren, was wir gerade eben gemacht haben. Und natürlich habe ich jetzt, wir werden hinterher wieder drei Zeilen haben, für jedes Vertragsjahr und hier steht immer das Gleiche. Kann sich natürlich auch mal ändern. Es kann sein, dass jemand, die Information haben wir jetzt nicht, zum bestimmten Datum ein Service dazu bucht.

Dann kann hier auch mal vorkommen, dass nicht in allen drei Zeilen das Gleiche steht. Aber wir wissen jetzt nichts. Sorry, alles klar. Also, ich habe hier sozusagen eine Tabelle, wo ich die Verträge nachschlagen kann und gucken, welche Services die dazu gebucht haben. Und mein Ziel ist aber wirklich, pro Vertragsjahr, also pro Instanz,

die für dich, für das ich nachher Vorhersagen machen will, wirklich nur eine Zeile zu haben. Das muss so sein. Also Orange will das sonst nicht oder kommt damit nicht klar. Das heißt, du musst irgendwie in eine Zeile pressen. Und der Trick, den man da anwendet, ist dieses sogenannte Pivotieren. Das heißt, man nimmt also diese möglichen sieben Werte und macht für jedes eine Spalte und dann gucke ich einfach, aha, sie hat Tech Support, dann mache ich das bei Tech Support eine Eins. Ja.

Und die zwei, die ich nicht habe, diese zwei hier, Online Backup, Online Security, da mache ich eine Null hin. Okay? Alle dabei? Gut, dann machen wir es. Was ich jetzt machen sollte, ich mache nochmal einen Aufbereitungsschritt. Damit hier am Ende irgendwelche Zahlen stehen, muss ich jetzt noch einen Trick machen. Und zwar lege ich jetzt ein Feld an, also berechnet das Feld erstellen.

Ich nenne es einfach mal Temp. Oder nein, ich nenne es NumServices. Und das ist einfach immer ein. Ich hoffe, ich mache das gerade richtig. So, das steht einfach immer eins. Und jetzt, nee, Entschuldigung, ich wollte jetzt pivotieren. Also das gibt es hier auch. Auf Englisch heißt es Pivot, glaube ich, wenn ich es richtig weiß. So, und jetzt muss man ein paar Sachen wissen. Jetzt gut aufpassen. Das müssen wir jetzt behalten. Das brauchen wir jetzt. Genau.

Also, das Erste, was ich jetzt mache, ist, man kann anscheinend auch Spalten in Zeilen pivotieren, aber meistens will man Zeilen in Spalten pivotieren. Also wir haben jetzt zu viele Zeilen für unsere drei Vertragsjahre. Statt drei haben wir plötzlich 15 gehabt. Und wir wollen jetzt aus den Zeilen wieder Spalten machen. So ungefähr. Und jetzt steht hier, fällt das Zeilen in Spalten pivotiert. Okay. Eine Idee? Zeilen in Spalten.

Service, ja. Genau, das sind die sieben Werte, also die sieben verschiedenen Services, die wir, jetzt habe ich es falsch genommen. Service, komm jetzt, wieder nicht. Warum? Er nimmt immer Senior Citizen. So, jetzt sind da die sieben Werte, die ich brauche. Daraus werden jetzt sieben Spalten. Die sehen wir gleich jetzt hier auftauchen, wenn ich noch sage, was aggregiert werden soll. Er will auch immer irgendwas aggregieren. Und da nehmen wir jetzt nun Services.

Jetzt seht ihr, dass diese Spalten neu dazugekommen sind. Und da unten seht ihr, also nochmal, hier habe ich Service reingezogen, dann erscheinen da die sieben verschiedenen Services und hier unten Num-Services. Also ihr seht jetzt, alles von hier links, das heißt, es ist alphabetisch sortiert. Ja.

Ich habe es von hier nach da gezogen. Von hier links nach da unten rechts. Also vielleicht war ich hier zu schnell, oder?

Ich habe hier das Feld erstellt. Im Aufbereitungsschritt berechnen das Feld. Ich habe es so genannt und einfach eine 1 als Formel reingeschrieben. Wie meinst du, oben und unten? Hast du hier noch Spalten in Teilen vielleicht?

Du musst doch Zeilen in Spalten umstellen. Dann hast du beides. Okay, suchen wir wieder nach unserer Kundin und gucken, was passiert ist. Wo ist es denn jetzt? Es wird ein bisschen unübersichtlich langsam. Jetzt haben wir wieder drei Zeilen. Seht ihr es? Und bei Online Security steht immer Null. Bei Online Backup auch. Und sonst überall Einsen. Okay, es steht keine Null da, sondern Null.

Lassen wir jetzt mal so. Man könnte das noch. Dann müsste man noch mal sieben Formeln machen. Ist null, dann null. Okay. Das machen wir jetzt nicht. Das ist schon anstrengend genug gewesen.

Jetzt noch diese letzte Contract-ID hier weg. Okay, also wenn ihr dann irgendwann soweit seid, dann können wir eigentlich zur Ausgabe übergehen. Also wir sind jetzt fertig. Nicht mit dem Unterricht, aber mit den Weiterführungen.

... ...

Vielen Dank.

... ...

Ich muss jetzt einfach auf die Flache gehen.

Okay, wie gesagt, wir sind eigentlich fertig jetzt. Wir machen nur noch den Export, damit ihr den auch noch seht. Also jetzt kann ich hier Ausgabe wählen. Also es tut mir innerlich weh. Ich würde wahrscheinlich jetzt die Spalten noch umsortieren, weil das ist nicht schön. Also zum Beispiel würde ich das Klassenattribut nach ganz hinten sortieren. Können wir vielleicht noch machen, damit ich nicht ganz so viele Schmerzen leiden muss.

Also das nach hinten und die Contract-ID würde ich, glaube ich, ganz nach vorne machen, auch wenn die am Ende kein Attribut ist, was ich zum Lernen benutzen würde. Übrigens habe ich die Contract-ID zweimal im Join verwendet und ich habe sie jeweils rausgelöscht. Ich weiß nicht, ob ihr das mitgekriegt habt. Und wahrscheinlich, ja doch, jetzt räume ich hier ein bisschen auf, ich halte es nicht aus. Period Start und Period End würde ich auch nach vorne nehmen.

Wo die alle hingekommen sind. Also hinter die Contract-ID. Also natürlich kann hier jeder so aufräumen, wie er will, aber... So, jetzt gefällt es mir besser. Also hier gibt es jetzt sozusagen die Vorschau. Also ihr seht jetzt, wenn wir nochmal...

Wir können gleich, wir exportieren das jetzt mal, dann machen wir es in Excel auf und dann filtern wir mal nach unserem Beispiel. Ja, ja, natürlich. Das mache ich gleich. Wir machen doch mal gleich fünf Minuten Pause und dann, warte mal, halt. So, jetzt hier. Ich würde jetzt mal, sollen wir Excel machen? Machen wir Excel? Ja.

Bei Excel ist das Problem, jetzt muss ich hier noch ein Arbeitsblatt anlegen. Ich nenne das mal Churn. Genau, also hier habe ich Excel bei Ausgabetyp. Dann will er noch ein Arbeitsblatt von mir wissen. Da muss ich auf Erstellen klicken. Also ich mache es nochmal weg. Geht gar nicht. Also wenn ihr da noch nichts habt, könnt ihr hier was eingeben und dann fragt er euch, ob ihr es erstellen wollt. Und dann kann ich hier oben noch

Gucken, wo ich es hin speichern will. Mache ich vielleicht mal ein Output-Folder. Und dann nenne ich das mal Churn. Könnt ihr nennen, wie ihr wollt. So, jetzt ist noch nichts passiert. Ich muss erst auf den Pfeil dann drücken. Oder hier unten auf Schema ausführen. Dann wird es gemacht. Dauert kurz. Und dann können wir schauen. Ja.

Okay, jetzt filter ich nochmal nach unserer Kundin. Moment, wir machen es anders. So, drei Zeilen. Ich sehe jeweils hier jetzt Start und End, so wie am Whiteboard. Dann sehe ich, dass die Kundin alle Services außer Online-Services

Backup und Online Security gebucht hat. Ich sehe, wie viele Tickets sie hatte in jedem Vertragsjahr. 002, so wie am Whiteboard. Und die anderen Sachen, die vorher da waren, sind auch da. Und hier hinten ist mein Klassenattribut. Also jetzt könntet ihr das wirklich in Orange laden und gucken, ob ihr einen schönen Pattern findet. So wie wir es letzte Woche gemacht haben. Ihr seid noch nicht ganz da. Macht nichts. Wir können gleich noch gucken. Ich schlage vor, wir machen noch mal 10 Minuten Pause. Okay?

Und dann habe ich noch ein, zwei Folien, kleine Diskussionen. Diese Sachen hier, die drei, haben wir jetzt. Es gibt noch ein paar. Was machen wir, wenn Werte fehlen, zum Beispiel so ein paar Sonderfälle, die wir gleich kurz angucken. Holt euch noch einen Kaffee oder Energy Drink, damit ihr wieder stark seid.

Nein, das machen wir jetzt nicht auf.

Okay.

Dann schauen wir uns noch ein paar andere Sachen an, die einem begegnen können. Die Daten, die wir da jetzt verwendet haben, die waren nicht so realistisch. Tatsächlich sind auch die Daten, die ihr fürs Assignment bekommen werdet oder bekommen habt, eigentlich schon nicht so realistisch, weil diese Probleme, die hier beschrieben sind, einige davon...

Die meisten davon nicht auftreten. Zum Beispiel gibt es keine fehlenden Werte. Fangen wir mal damit an. Also ihr seht hier, es gibt Herausforderungen. Wir wollen die noch kurz uns anschauen. Was kann man machen, wenn Werte fehlen? Man kann sie einfach fehlen lassen und die meisten Algorithmen kommen damit klar, aber nicht alle. Also in Orange habe ich das Gefühl, die Implementierung da sind sehr robust, was fehlende Werte angeht. Die meisten Widgets, die hier fürs

trainieren von Modellen, euch wählt, werden damit klar. Wenn ihr später mal mit Python Scikit-Learn oder sowas arbeitet, dann sieht es manchmal anders aus. Dann braucht man Lösungen. Speziellen Wert sozusagen, es kommt darauf an, was es für ein Attribut ist. Wenn es kategorisch ist, dann ist 0 vielleicht kein nötiger Wert.

Wenn es ein numerisches Attribut ist, so was wie NumTickets, dann kann ich 0 reintun. Es ist natürlich, also wenn es null ist, dann ist es oft unbekannt. Also es fehlt. Und die Aussage ist nicht die gleiche wie 0. Also 0 heißt, es gab kein Ticket. Und wenn es leer ist, dann heißt das vielleicht, wir wissen nicht, ob es eins gab. Deswegen macht man oft nicht die 0,

Oft macht man sie auch. Man muss nur überlegen, was man da gerade tut. Aber ja, wenn es ein numerisches Attribut ist, kann man die fehlenden Werte durch Nullen ersetzen. In vielen Fällen ist das okay. Was viele Leute auch gerne machen, ist einfach Zeilen löschen, die fehlende Werte haben. Das ist eine ziemlich radikale Sache.

Oder man kann sogar, das ist fast noch radikaler, einfach die Spalten löschen. Insbesondere, wenn Sie viele fehlende Werte haben, sagt man, okay, die Information ist so unvollständig von dem Feature, von dem Attribut, lass uns einfach weglöschen. Das waren also sozusagen diese zwei Optionen hier. Und dann gibt es die Möglichkeit, dass man was berechnet. Das eine, was man machen kann, ist, dass man den, also irgendeine,

Summary einfügt, also zum Beispiel den Mittelwert oder den Median, wisst ihr, was ein Median ist? Oder bei einem kategorialen Attribut Modus, wisst ihr, was der Modus ist beim kategorialen Modus? Der häufigste Wert. Gerade bei numerischen Attributen kann das zu Problemen führen. Also nehmen wir mal an, der Mittelwert, also ihr habt ein Attribut Alter und da fehlen Werte.

Und sagen wir mal, der Mittelwert von Alter ist 40,79. Und wenn viele Werte fehlen, dann heißt es, dass ihr für viele Instanzen den Wert 40,79 einfügt. Und dann werden viele Machine Learning Modelle sich genau auf diesen Wert stürzen, weil sie dann irgendwie rausfinden, aha, das bedeutet, dass der Wert gefehlt hat. Und dann fangen die damit irgendwas an. Das ist nicht immer erwünscht, nicht immer gut.

Was oft besser ist, ist, dass man eine andere Instanz findet, die der aktuellen Instanz am ähnlichsten ist. Dafür muss man eine Definition davon haben, was ähnlich heißt. Also natürlich benutzt man dafür die anderen Attributwerte, wo die nicht fehlen und guckt, okay, gibt es eine Instanz, die vielleicht sogar die gleichen Attributwerte oder sehr ähnliche Attributwerte in den anderen Attributen hat und dann kann man von dieser Instanz den fehlenden Wert übernehmen.

Was ich jetzt machen will mit euch, ist eine kleine Diskussion anzetteln. Einfach nur, oh nein, mit den zwei löschen Optionen. Ich will einfach, dass ihr ein bisschen Bewusstsein dafür bekommt, dass es sehr unterschiedliche Auswirkungen haben kann, was man tut. Also, was seht ihr hier? Ihr seht da oben einen Beispieldatensatz. Den habt ihr, glaube ich, schon mal, oder? Die

unsere Wintercheck von Swissbikes. Habe ich letztes Mal auch als Beispiel schon genommen, oder? Was beobachten wir? Wir beobachten, da gibt es die Spalte Datum, letzte Reparatur und da fehlen drei Werte. Also zehn Kunden, das sind unsere Instanzen. Alle sind, das bleibt mal so gut, ist überall vorhanden. Also wir wissen, wer reagiert hat und wer nicht. Und bei vier Kunden, nee, drei Kunden, beim zweiten, siebten und zehnten Kunden, da fehlt

Das Datum der letzten Reparatur. Was man beobachtet, was auch irgendwie plausibel ist, ist, dass Leute, die in der Spalte von Anzahlreparaturen mehr, also größere Zahlen haben, da natürlich auch die letzte Reparatur eher weniger lange her ist. Also da gibt es eine gewisse Korrelation. Also wer mehr Reparaturen hatte, der hatte seine letzte auch vor weniger langer Zeit, tendenziell.

Und was man beobachtet ist, wenn man so ein bisschen die Mitbeute und das Klassenattribut anschaut, da gibt es auch eine Korrelation. Also wenn jemand viele Reparaturen hatte, kommt er auch eher zum Wintercheck. Ist auch plausibel. Wenn wir jetzt das alles mal in unserem Kopf behalten. Ich gebe euch jetzt nur zwei Optionen. Denkt darüber nach, was ihr lieber hättet, was ihr denkt, was sinnvoller ist. Entweder wir löschen diese drei Zeilen, wo die Werte fehlen, oder wir löschen die Spalte.

Datum, letzte Reparatur. Das sind beides relativ radikale und nicht unbedingt die besten Lösungen, aber ich gebe euch nur die zwei Optionen. Also wir nehmen uns jetzt zwei, drei Minuten. Sorry, Dennis. Und ihr diskutiert mal in kleinen Gruppen, einfach mit dem Nachbarn oder zu dritt, was ist besser und warum? Oder besser gesagt, was passiert in jedem Fall? Also was passiert, wenn ich die drei Zeilen lösche? Was passiert, wenn ich die Spalte lösche? Und was ist schlimmer?

Ist klar, du bist beginnend? Also.

Nehmen wir mal drei Minuten uns Zeit. Okay.

...

Ja, das ist ein guter Punkt.

Vielen Dank.

Vielen Dank.

Wenn man die Ziele durchlöscht, dann kommt man in die Bündel und die Fälle kommen. Wenn man die Fälle durchlöscht, dann ist das nicht einfach so, dass man sich das dann auch übernimmt, aber es ist auch eine ganz andere Sache. Ich denke, das ist ein sehr guter Autoreperatoren, der sich die Spuren ganz offen macht.

Vielen Dank.

Was wollen wir machen? Spalte oder Teilen löschen? Spalte. Warum? Weil das Daten von der letzten Reparatur nicht wichtig ist. Okay, also es ist nicht wichtig, weil...

Wir haben ja diese Spalte, die korreliert mit der und die ist auch schon gut, um vorherzusagen, ob jemand reagiert. Mir ist es ja egal, wenn die letzten Spalten gemacht wurden. Ich kann einfach dieses Video auch auf YouTube anstellen.

Also ich meine, die Idee ist, also das könnte schon ein nützliches Attribut sein. Je kürzer es her ist oder andersrum, je länger es her ist, desto unwahrscheinlicher wird es, dass die Person wiederkommt. Vielleicht ist sie ja schon weggezogen oder so. Das haben wir ja auch in der Anzahl. Ja, nicht ganz. Also es sind ja die letzten drei Jahre. Kann sein, dass alles vor zwei Jahren oder länger war. Ja, das ist auch ein Problem.

Also seht ihr das alle? Es gibt vier

Leute, die ihren Gutschein eingelöst haben. Und wenn wir die Zeilen löschen, dann verlieren wir drei von denen und haben nur noch einen übrig. Und dann, was wird dann passieren, wenn wir ein Modell lernen mit dem siebener Datensatz? Wie wird das sein im Vergleich zu dem anderen, wo wir die Spalte gelöscht haben? Das ist das Unterscheiden. Das heißt, die Kondens wird natürlich nicht wahr sein, dass niemand einen Gutschein hat.

Also das Modell wird pessimistisch sein. Es wird viel öfter Nein sagen als das andere Modell, was mehr Reaktionen von Kunden gesehen hat, mehr positive Antworten. Das ist wichtig. Das ist auch so sozusagen, selbst wenn wir jetzt, nehmen wir mal an, wir hätten 10.000 Kunden und nur 100 Zeilen hätten diese fehlenden Werte.

dann würde man sagen, naja, 100 von 10.000 können wir ruhig wegschmeißen. Aber ich würde vorher gucken. Gerade im Targeted Marketing gibt es oft nicht wie hier 40% Response Rate, sondern vielleicht 2%. Also bei 10.000

wären das dann 200. Wenn jetzt die 100, wo die Werte fehlen, ausgerechnet, und das ist dann wahrscheinlich kein Zufall, sondern da steckt ein bösartiges System dahinter, wenn das ausgerechnet diese 100, alles welche sind, die auch geantwortet haben und ich lösche die, dann habe ich zwar nur 1% der Daten gelöscht, was irgendwie irrelevant erscheint, aber ich habe 50% der Respondents gelöscht. Und dadurch wird mein Modell ein ganz anderes sein als das,

wenn ich was anderes tue, also zum Beispiel die Spaltung. Ja, also man sollte immer gucken, was man da löscht. Also es muss nicht immer falsch sein, Zeilen zu löschen. Also wenn man sehr viele Daten hat und man checkt, dass die Zeilen, wo fehlende Werte sind, quasi zufällig verteilt sind, dann okay, löscht. Also ich weiß nicht, wie es gemacht wird, aber kann man zwei Modelle machen, dass ein Modell ist mit den 100,

Und dann sieht man auch das Muster, okay, wie reagieren. Und ein Modell ohne diesen Faktor, dass man weiß, okay, was ist bei den anderen der ausgebende Faktor. Ja, ja, also du interessierst dich dafür, warum jetzt gerade bei den 100 die Werte fehlen. Also nein, bei den 100, wo die Werte fehlen,

Man sieht z.B. die Anzahl Reparaturen. Wenn die hoch ist, dann reagiert man eher. Dann gibt es noch andere Hunde, die haben reagiert, aber bei denen ist die Anzahl nicht hoch. Kann man dann die anderen ausschliessen, um sich nur einen anderen Faktor anzuschauen? Weisst du, was ich meine? Oder nicht so ganz?

Ich weiß nicht ganz genau, aber ich würde es, nein, nein, nein, würde ich nicht machen. Also ich würde die Entscheidung treffen, was ich mit den fehlenden Werten mache. Das hat einen Einfluss auf meinen Datensatz. Vielleicht werden es weniger Instanzen oder weniger Spalten. Und dann würde ich damit weiterarbeiten und nicht vergleichen. Also ich meine, du kannst es vergleichen, aber dann würdest du schauen, was besser funktioniert. Wir werden ja Metriken kennenlernen, um zu bewerten, wie ein Modell funktioniert.

Und spätestens dann, wenn du siehst, aha, das eine ist besser als das andere, dann die Entscheidung treffen, immer mit dem einen weiterzufahren. Okay, andere Fragen, Bemerkungen? Nein, jetzt haben wir es gleich geschafft. Genau, also manchmal hat man Ausreißer. Wie man Ausreißer definiert, ist natürlich auch mit mehreren Möglichkeiten. Also man kann zum Beispiel...

gucken, was mehrere Standardabweichungen vom Mittelwert entfernt ist oder so. Auch da muss man sich aber erst angucken, welche Verteilung die Daten aufweisen. Wenn sie nicht normal verteilt sind, aufpassen. Okay, wir wollen nicht zu sehr in die Mathematik einsteigen, beziehungsweise wir haben nicht die Zeit. Ja, das sind auch, ja, also ich meine, Ausreißer können zum Beispiel auftreten bei Preisen. Ist ja gern mal passiert, dass das da passiert.

das Dezimaltrendzeichen verrutscht. Also die Rappenbeträge, wenn man das nach hinten geschoben, das Komma sozusagen, dann ist es plötzlich hundertmal so hoch, der Preis. Oder man hat unterschiedliche Datumsformate oder oft hat man auch bei Strings Kodierungsprobleme mit Umlauten und so ein Zeug. Also kommt alles noch auf einen zu und meistens findet man pragmatische Lösungen.

Was manchmal passiert ist, dass insbesondere kategoriale Attribute so aussehen, als seien sie numerisch. Zum Beispiel könntet ihr es mal vorfinden, dass ein Attribut Monat heißt und statt dass da als Werte drinnen Januar, Februar bis Dezember sind, wird es vielleicht einfach mit Zahlen, also 1 bis 12, kodiert. Wenn das zum Beispiel ein CSV-File ist und ich lade das in Orange,

dann besteht die Gefahr, dass Orange denkt, dass es ein numerisches Attribut ist und fängt an, mit dieser Zahl zu rechnen. Aber es ist eigentlich keine Zahl, wenn ihr darüber nachdenkt. Und vielleicht ist sich ja der Dezember ähnlicher mit dem Januar als mit dem Juli. Aber wenn ich anfange, die numerisch zu vergleichen, ja, 1 und 12 sind maximal weit auseinander. Also das ist keine gute Idee. Dann sollte man das einfach ändern. Man kann das in Orange jederzeit, wenn man ein Pfeil... Mal schauen, habe ich gerade ein Beispiel.

Das ist jetzt alles irgendwelches Zeug, was euch nichts sagt, aber hier Age zum Beispiel. Age ist natürlich wirklich numerical, aber ich kann hier jeweils mit dem Dropdown das ändern auf categorical. Okay, das sollte ich nicht tun. Okay, wenn ihr feststellt, dass da was falsch angeordnet wird, dann könnt ihr es einfach ändern. Dann gibt es manchmal komplexere Typkonversionen. Also in dem Sinne, dass

eure Attribute entweder in wirklich, also die sind entweder, die sind wirklich kategorial oder wirklich numerisch, nur euer Algorithmus will sie nicht fressen. Also das ist eigentlich selten. Es gibt so gut wie keine Algorithmen mehr, die nur kategoriale Attribute haben wollen. Es gibt so ganz einfache Implementierungen von Entscheidungsbäumen, die das wollen, aber

Eigentlich können wir das überspringen. Was aber sehr häufig vorkommt, ist das Umgekehrte. Also ihr habt kategoriale Attribute, aber euer Algorithmus akzeptiert nur Zahlen, also nur numerische Attribute. Das ist zum Beispiel bei neuronalen Netzen so und die sind ja nun mal beliebt. Und dann ist der Trick, dass man die Werte einfach

so ähnlich wie beim Pivotieren, nimmt und für jeden dieser Werte eine neue Spalte erzeugt und dann mit 0 und 1 kodiert. Also wenn ihr jetzt hier, das sind jetzt irgendwelche Nahrungsmittel und dann gibt es vielleicht, ursprünglich hieß das Attribut Type und hatte die Werte Apple, Chicken und Brokkoli.

Und dann gibt es noch Calories als zweites Attribut. Und aus diesem ersten Type-Attribut haben wir jetzt anhand der vorhandenen Werte, Apple, Chicken, Brokkoli, haben wir drei neue Spalten generiert. Und ein Apple wird jetzt als 1 0 0 kodiert, ein Chicken als 0 1 0 und ein Brokkoli als 0 0 1. Das heißt One Hot, weil immer nur eine dieser Spalten jetzt eine Eins im Held. Das ist anders als bei der Pivotierung, wo wir mehrere Einsen haben können.

Hier, ja, wir gewinnen eigentlich überhaupt keine Information. Es wird eigentlich nur aufgeblasen, aber es sieht aus wie Zahlen. Und somit könnt ihr das auch einem neuronalen Netz vorwerfen. Orange ist da nicht so. Das macht es irgendwie für euch im Hintergrund. Da könnt ihr auch kategoriale Gebote reingeben, aber wenn ihr dann mit Python arbeitet, dann müsst ihr es selber machen. Und das ist aber auch mit eigentlich einer Zeile Code.

zu schaffen. Das war Not-Encoding. Ja, also ihr wendet das auf Spalten an und dann habt ihr plötzlich viel mehr Spalten, eben entsprechend den Werten eures kategorialen Attributs. Fragen? Letzte Folie. Wir haben es gleich geschafft. Dann gehe ich nochmal in Präsentationsmodus. Manchmal möchte man auch zum Beispiel, dass die Werte von numerischen Attributen alle zwischen 0 und 1 liegen.

Also es kann zum Beispiel bei Algorithmen gut sein, die irgendwie auf einer Ähnlichkeitsfunktion basieren, damit alle Attribute gleich wichtig sind, was sonst nicht der Fall ist. Also wenn ihr sie nicht normalisiert, dann werden die Attribute die Ähnlichkeit dominieren, die besonders große Werte haben. Und ja, da gibt es eine Formel, wie man das zum Beispiel macht. Min-Max-Normalisierung ist so ein Standard. Gibt es in Orange auch. Also mal gucken.

Ich nehme jetzt einfach hier irgendeinen Workflow. Ich muss euch nichts sagen. Hier gibt es also ein Pre-Process-Widget. Da habt ihr alle möglichen Möglichkeiten. Zum Beispiel Normalize-Features. Aber es gibt Normalize auch nochmal als eigenes Widget, soviel ich weiß. Lass mal schauen. Normalize. Nee. Nee, okay. Dann nicht. Dann machen wir es mit Pre-Process. Also das ist jetzt hier Impute Missing Value. Das machen wir nicht.

Darüber haben wir uns gerade unterhalten. Normalize Features. Also da gibt es verschiedene. Hier steht nicht, wie es passiert, aber normalerweise will man sie oder oft will man sie in das Intervall 0,1. Vermutlich wird da im Hintergrund auch Min-Max-Normalisierung passieren. Und ihr seht hier, ich habe das hier Skalierung genannt, sodass sie...

eine Normalverteilung mit einem Mittelwert von 0 und einer Standardweichung von 1 aufweisen. Das ist das hier. Standardize heißt es hier. Also gibt es auch. Ja, also das ist zum Beispiel insbesondere die Normalisierung, auch wenn wir zum Beispiel später bestimmte Algorithmen wie Logistic Regression, logistische Regression benutzen wollen, dann hat man Koeffizienten und die können, wenn man alles normalisiert hat, einem sagen,

wie wichtig ein Attribut ist. Wenn man es nicht normalisiert hat, dann darf man diesen Schluss nicht ziehen. Also es gibt manchmal einfach, werden wir noch lernen, Gründe, warum man das machen will. Jetzt sind wir am Ende. Also ich habe euch wieder ein Quiz gemacht. Bevor ihr euch anfangt zu langweilen, ich schicke euch entsprechend die E-Mail. Nächste Woche wird es wieder ein bisschen spaßiger, dann machen wir wieder was mit Orange und es ist nicht so anstrengend. Bis zum nächsten Video.

Dennis, please.