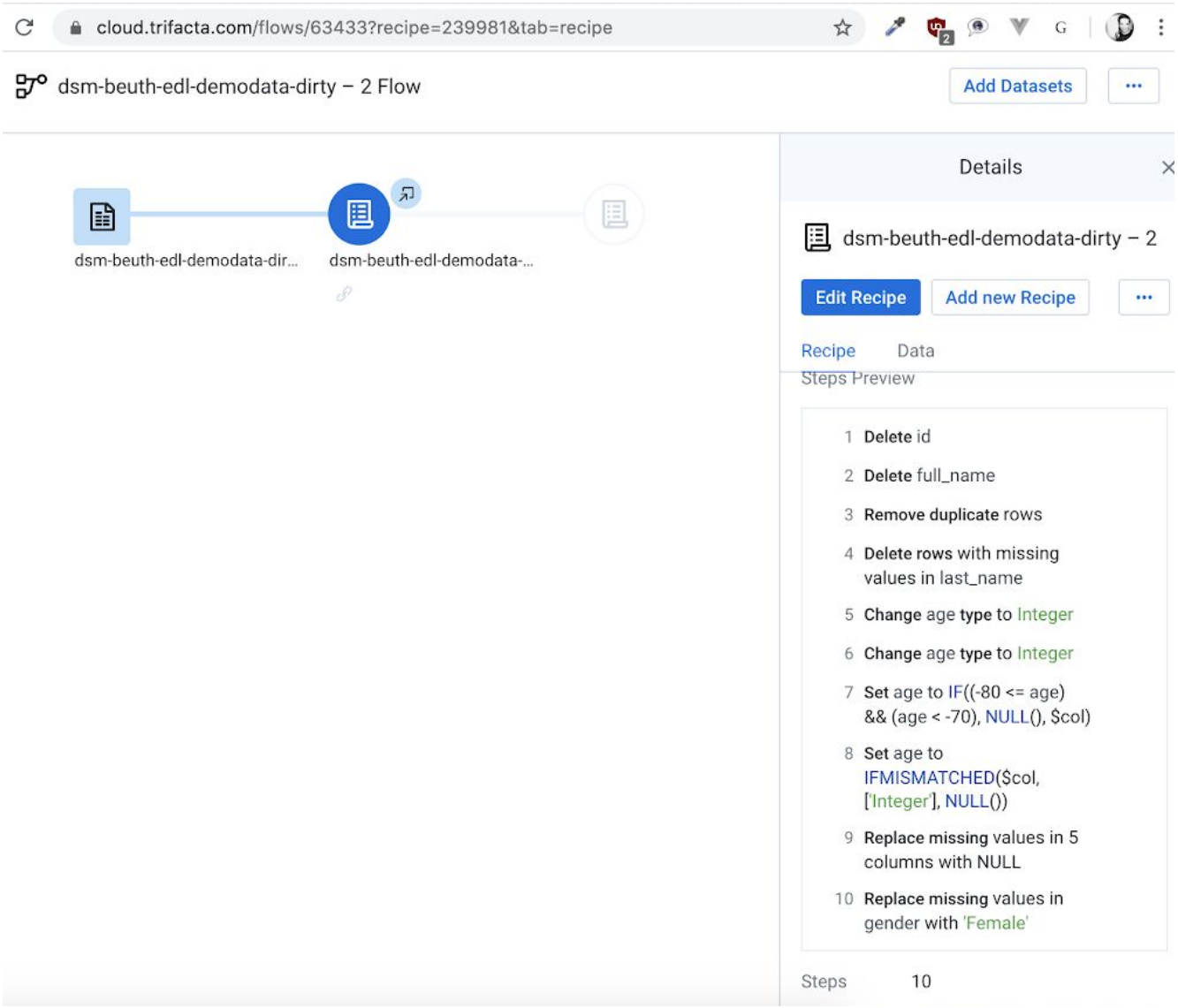
Data Science SS 2020

Tool Supported Data Cleaning

Sandra Tieben

A1: Clean the dsm-beuth-edl-demodata-dirty.csv mini csv from the first exercise with Trifacta Wrangler. Create a recipie to clean the data as good as you can (it must not be a general script). Try to upload only one file (e.g. with screenshots and the end result). (10 points)



Wie auf dem Screenshot zu erkennen wurden verschiedene Schritte durchgeführt, die ich im kurz erläutere:

Schritt	Erläuterung
Delete id	Entscheidung, dass die ID für uns nicht relevant ist
Delete full_name	Die Spalte “full_name” enthält redundate Informationen, die über “first_name” und “last_name” bereits vorliegen
Remove duplicate rows	Selbsterklärend.
Change age to integer	Der Typ der Spalte “age” wird zu einem Integer geändert. Alle fehlerhaften Daten werden zu NULL (“old” → NULL)
Changed age < 0	Die Spalte “age” darf m. M. n. nur Daten zwischen 0 und ca. 120 enthalten. Alle Werte, die nicht in diesem Wertebereich liegen werden mit NULL ersetzt. **Eine andere Alternative wäre es negative Werte (wie hier -78) durch den Wert ohne Vorzeichen zu ersetzen.**
Replace missing values in column “gender”	Da nur ein fehlender Wert in der Spalte “gender” vorlag, wurde dieser manuell durch den Wert “Female” ersetzt (https://www.names.org/n/hasty/about). **Das ist eher keine sinnvolle Strategie für sehr große Datenmengen!**

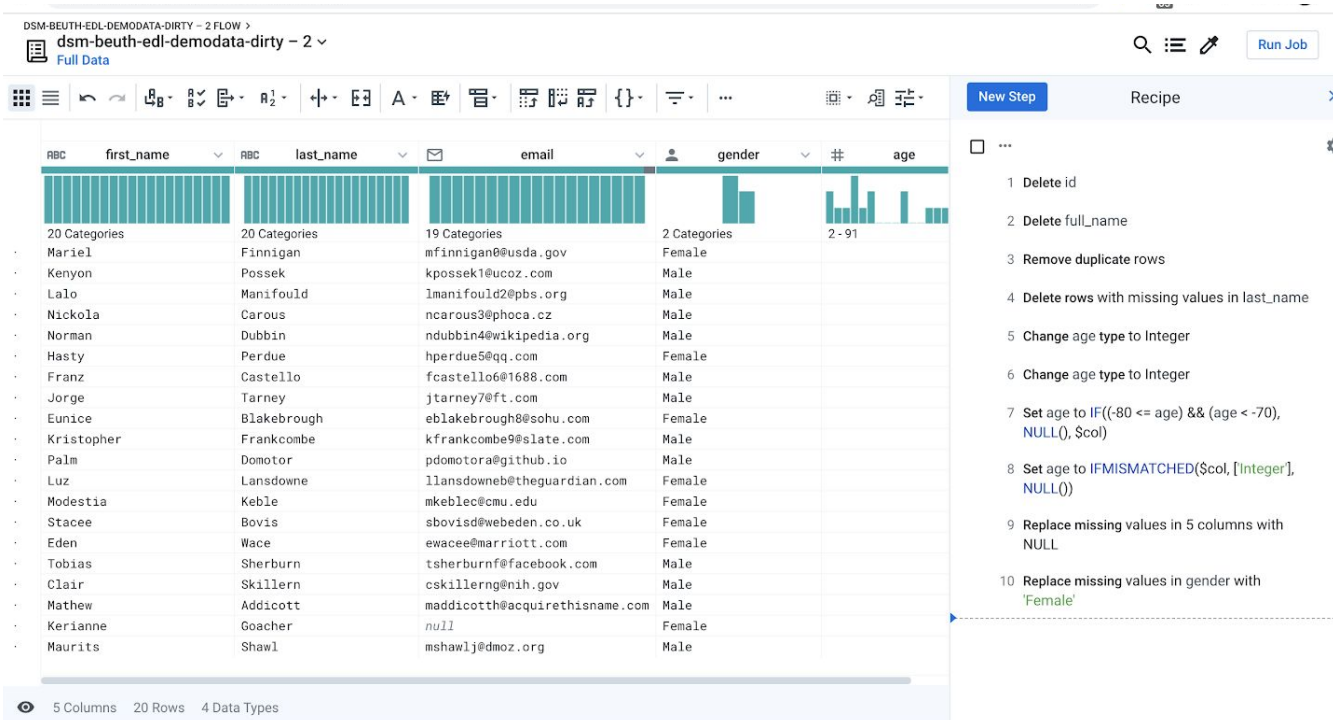
Weitere Möglichkeiten/Ideen:

Da wir aktuell kein besonderes Ziel haben, sind einige Werte als **NULL-Werte** in der Tabelle verblieben (ein fehlender Wert in “email”, zwei fehlende Werte in “age”). Je nach Ziel kann man **diese Werte ersetzen oder die Zeilen löschen**.

Als Ergebnis erhalten wir folgende, bereinigte Daten:

first_name	last_name	email	gender	age
Hasty	Perdue	hperdue5@qq.com	Female	77
Norman	Dubbin	ndubbin4@wikipedia.org	Male	17
Eden	Wace	ewacee@marriott.com	Female	16
Tobias	Sherburn	tsherburnf@facebook.com	Male	2
Franz	Castello	fcastello6@1688.com	Male	25
Eunice	Blakebrough	eblakebrough8@sohu.com	Female	45
Kristopher	Frankcombe	kfrankcombe9@slate.com	Male	
Stacee	Bovis	sbovisd@webeden.co.uk	Female	22
Luz	Lansdowne	llansdowneb@theguardian.com	Female	16
Kerianne	Goacher		Female	45
Lalo	Manifould	lmanifould2@pbs.org	Male	26
Modestia	Keble	mkeblec@cmu.edu	Female	91
Mathew	Addicott	maddicotth@acquirethisname.com	Male	65

Kenyon	Possek	kpossek1@ucoz.com	Male	12
Palm	Domotor	pdomotora@github.io	Male	6
Mariel	Finnigan	mfinnigan0@usda.gov	Female	60
Maurits	Shawl	mshawlj@dmoz.org	Male	72
Jorge	Tarney	jtarney7@ft.com	Male	77
Clair	Skillern	cskillerng@nih.gov	Male	
Nickola	Carous	ncarous3@phoca.cz	Male	4



A2: Load the Grid_Disruption_00_14_standardized - Grid_Disruption_00_14.csv Dataset from Kaggle: 15 YEARS OF POWER OUTAGES. Where are errors here? How would you clean this file? (5 Points)

Link zu Kaggle: https://www.kaggle.com/autunno/15-years-of-power-outages

Vorüberlegung: Welche Daten sind für mich relevant? Welche nicht? Davon abhängig muss man ggf. andere Schritte durchführen.

Erster Schritt: Überblick gewinnen (Aufbau, Inhalte, Größe)
→ Infos über Stromausfälle

Allgemeine Infos:
- 12 Spalten
- 1652 Reihen
- 4 Datentypen

Inhalt der Spalten:

Spalte	Datentyp	Beschreibung (aus Kaggle)
Event Description	String	Reason of the outage (e.g. Vandalistm, Severe Weather, etc)
Year	Date/Time (yyyy)	Year of outage
Date Event Began	Date/Time (mm*dd*yyyy)	Date of the outage
Time Event Began	Date/Time (hh:MMa)	Time the outage was registered
Date of Restauration	Date/Time (mm*dd*yyyy)	Date the outage was resolved
Time of Restauration	Date/Time (hh:MMa)	Time the outage was resolved
Respondent	String	The company that acted upon the outage
Geographic Areas	String	Region of the outage
NERC Region	String	NERC refers to the North American Electricity Reliability Corporation, formed to ensure the reliability of the grid
Demand Loss (MW)	String	How much energy was not transmited/consumed during the outage.
Number of Customers Affected	String	How many consumers (e.g. homes, offices, industry, etc) were left to their devices.

Tags	String	Summary event description

Nachdem wir einen Überblick über die Tabelle haben, suche ich nach **Duplicates**. Danach schaue ich die einzelnen Spalten genauer an, von denen einige Auffälligkeiten aufweisen:

Event Description:
Enthält ein oder mehrere Werte (z. B. “Severe Weather”, “Severe Weather - Thunderstorms”, “Vandalism”, “Physical Attack - Vandalism”…). Scheinbar werden die Ereignisse unterschiedlich genau beschrieben und es liegt keine einheitliche Beschreibung vor.Hier kann man sich mehrere Vorgehensweisen vorstellen:

1. Aufteilung in zwei Spalten mit Trennzeichen “-”
2. Wenn Trennzeichen “-” vorhanden ist, wird der vordere oder hintere Teil entfernt.

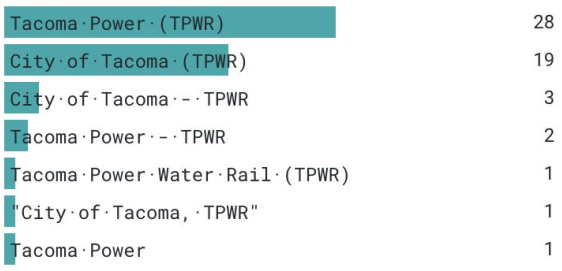
Hier ist viel manuelle Nacharbeit vermutlich nötig.

Time Event Began:
23 Zeilen enthalten ein fehlerhaftes Format (z.B. 12:00 a.m./midnight/Ongoing statt 12:00 AM), die korrigiert werden sollten.

Date of Restauration:
Hier liegen 35 nicht-valide Werte vor (z.B. NA, Ongoing, Unknown) → *Bereinigen*

Time of Restauration:
72 fehlerhafte Reihen. Teilweise falsches Format (z.B. 14:00 statt 14:00 PM) → Korrektur
Teilweise unbekannt (“Unknown”, “Ongoing”) → *Bereinigung*

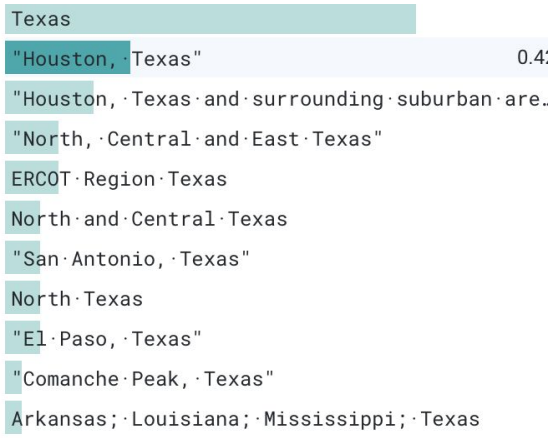
Responent:
Beim Betrachten der Zeilen fällt auf, dass einige Firmen in unterschiedlicher Schreibweise vorliegen, was bereinigt werden sollte. Beispiel (Suche nach “Tacoma”):



Hier kann eine Bereinigung z. B. auf Basis der Wort-Ähnlichkeit zwischen den Begriffen erfolgen.

Geographic Areas:
Dieses Feld enthält ein oder mehrere Orte. Teilweise wird nur der Bundesstaat aufgeführt, zum Teil auch eine Stadt und in manchen Fällen wird auch die NERC-Region angegeben.

Hier ist eine Bereinigung auch sinnvoll, z. B. nur der Staat oder nur die Stadt (siehe Abbildung). Die NERC-Region sollte entfernt werden.



NERC-Region:
Enthält ein oder mehrere Kürzel. Ist soweit okay (ggf. auf Fehler prüfen?)

Demand Loss:
Über 35% der Felder enthalten “Unknown” oder “N/A”. Auch relativ oft 0 (**Überlegung: Stimmt das wirklich oder wird hier 0 für fehlende Daten verwendet?**).

Number of Costumer Effected:

- Konvertieren zu Zahl
- Unknown, N/A zu Null konvertieren
- Auch hier wieder die Frage bei 0 bzw. 1: Stimmt das oder fehlen hier die Daten?
- Einige Werte scheinen sehr genau, andere gerundet (ggf. alle Werte runden?)

Tags

- enthält mehrere Elemente…
- ist oft ähnlich oder identisch zu “Event Description” → Spalte löschen

Weitere mögliche Schritte

- Inkonsistente Daten finden (z. B. Geographic Area und NERC passen nicht)
- Zusammengehörige Datum und Zeit zusammenfassen?
- Sind einige Ereignisse mehrfach vorhanden? Wie kann man das am besten herausfinden? (Start? Ende? Tags? Region? Tags?)

8/23/2013	7:30 AM	8/23/2013	7:31 AM	Delmarva Power & Light
8/22/2013	12:55 PM	8/22/2013	2:45 PM	Delmarva Power & Light

Ausschnitt: Zwei Spalten mit identischen Inhalten - nur Start-/Enddatum sind leicht unterschiedlich