

# 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 recipe 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)

The screenshot displays the Trifacta Wrangler web interface. The browser address bar shows the URL: `cloud.trifacta.com/flows/63433?recipe=239981&tab=recipe`. The main header indicates the current flow is "dsm-beuth-edl-demodata-dirty - 2 Flow". Below the header, a visual workflow diagram shows a sequence of steps: a document icon, a blue circular node with a document icon and a small blue circle with a plus sign, and another document icon. The right-hand sidebar is titled "Details" and contains the following information:

- Flow name: dsm-beuth-edl-demodata-dirty - 2
- Buttons: Edit Recipe, Add new Recipe, and a three-dot menu.
- Tabs: Recipe (selected) and Data.
- Section: Steps Preview
- List of 10 steps:
  - 1 Delete id
  - 2 Delete full\_name
  - 3 Remove duplicate rows
  - 4 Delete rows with missing values in last\_name
  - 5 Change age type to Integer
  - 6 Change age type to Integer
  - 7 Set age to `IF((-80 <= age) && (age < -70), NULL(), $col)`
  - 8 Set age to `IFMISMATCHED($col, ['Integer'], NULL())`
  - 9 Replace missing values in 5 columns with NULL
  - 10 Replace missing values in gender with 'Female'
- Footer: Steps 10

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 redundante 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. <b>Eine andere Alternative wäre es negative Werte (wie hier -78) durch den Wert ohne Vorzeichen zu ersetzen.</b>
Replace missing values in column "gender"	Da nur ein fehlender Wert in der Spalte "gender" vorlag, wurde dieser manuell durch den Wert "Female" ersetzt ( <a href="https://www.names.org/n/hasty/about">https://www.names.org/n/hasty/about</a> ). <b>Das ist eher keine sinnvolle Strategie für sehr große Datenmengen!</b>

#### 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	<a href="mailto:hperdue5@qq.com">hperdue5@qq.com</a>	Female	77
Norman	Dubbin	<a href="mailto:ndubbin4@wikipedia.org">ndubbin4@wikipedia.org</a>	Male	17
Eden	Wace	<a href="mailto:ewacee@marriott.com">ewacee@marriott.com</a>	Female	16
Tobias	Sherburn	<a href="mailto:tsherburnf@facebook.com">tsherburnf@facebook.com</a>	Male	2
Franz	Castello	<a href="mailto:fcastello6@1688.com">fcastello6@1688.com</a>	Male	25
Eunice	Blakebrough	<a href="mailto:eblakebrough8@sohu.com">eblakebrough8@sohu.com</a>	Female	45
Kristopher	Frankcombe	<a href="mailto:kfrankcombe9@slate.com">kfrankcombe9@slate.com</a>	Male	
Stacee	Bovis	<a href="mailto:sbovisd@webeden.co.uk">sbovisd@webeden.co.uk</a>	Female	22
Luz	Lansdowne	<a href="mailto:llansdowneh@theguardian.com">llansdowneh@theguardian.com</a>	Female	16
Kerianne	Goacher		Female	45
Lalo	Manifould	<a href="mailto:lmanifould2@pbs.org">lmanifould2@pbs.org</a>	Male	26
Modestia	Keble	<a href="mailto:mkeblec@cmu.edu">mkeblec@cmu.edu</a>	Female	91
Mathew	Addicott	<a href="mailto:maddicott@acquirethisname.com">maddicott@acquirethisname.com</a>	Male	65

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

DSM-BEUTH-EDL-DEMOMDATA-DIRTY - 2 FLOW >

dsm-beuth-edl-demodata-dirty - 2 Full Data Run Job

New Step Recipe

first_name	last_name	email	gender	age
Mariel	Finnigan	mfinnigan0@usda.gov	Female	
Kenyon	Possek	kpossek1@ucoz.com	Male	
Lalo	Manifould	lmanifould2@pbs.org	Male	
Nickola	Carous	ncarous3@phoca.cz	Male	
Norman	Dubbin	ndubbin4@wikipedia.org	Male	
Hasty	Perdue	hperdue5@qq.com	Female	
Franz	Castello	fcastello6@1688.com	Male	
Jorge	Tarney	jtarney7@ft.com	Male	
Eunice	Blakebrough	eblakebrough8@sohu.com	Female	
Kristopher	Frankcombe	kfrankcombe9@slate.com	Male	
Palm	Domotor	pdomotora@github.io	Male	
Luz	Lansdowne	llansdowne@theguardian.com	Female	
Modestia	Keble	mkeblec@cmu.edu	Female	
Stacee	Bovis	sbovisd@webbeden.co.uk	Female	
Eden	Wace	ewacee@marriott.com	Female	
Tobias	Sherburn	tsherburnf@facebook.com	Male	
Clair	Skillern	cskillerng@nih.gov	Male	
Mathew	Addicott	maddicott@acquirethisname.com	Male	
Kerianne	Goacher	null	Female	
Maurits	Shawl	mshawlj@dmoz.org	Male	

- Delete id
- Delete full\_name
- Remove duplicate rows
- Delete rows with missing values in last\_name
- Change age type to Integer
- Change age type to Integer
- Set age to IF((-80 <= age) && (age < -70), NULL(), \$col)
- Set age to IFMISMATCHED(\$col, [Integer], NULL())
- Replace missing values in 5 columns with NULL
- Replace missing values in gender with 'Female'

5 Columns 20 Rows 4 Data Types

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. Vandalism, 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 <a href="#">North American Electricity Reliability Corporation</a> , 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 Restoration:

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

#### Time of Restoration:

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"):

Tacoma · Power · (TPWR)	28
City · of · Tacoma · (TPWR)	19
City · of · Tacoma · - · TPWR	3
Tacoma · Power · - · TPWR	2
Tacoma · Power · Water · Rail · (TPWR)	1
"City · of · Tacoma, · TPWR"	1
Tacoma · Power	1

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.

```
Texas
"Houston, Texas" 0.4:
"Houston, Texas and surrounding suburban are..
"North, Central and East Texas"
ERCOT Region Texas
North and Central Texas
"San Antonio, Texas"
North Texas
"El Paso, Texas"
"Comanche Peak, Texas"
Arkansas; Louisiana; Mississippi; Texas
```

### 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