Data Science
Sommersemester 2022
Julia Friedrich, 7014416

Einsendeaufgabe 8 – Tool Supported Data Cleaning

1. Clean the dsm-beuth-edl-demodata-dirty.csv mini csv from the first exercise with Trifacta Wrangler (if you have no cloud access use the download version).

Not the red bars that give you instant feedback (on big datasets) where errors could be!

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)

Recipe

settype col: age lockDataType: true type: Integer

filter type: custom rowType: single row: 0 >= age action: Delete

filter type: custom rowType: single row: 'old' == age action: Delete

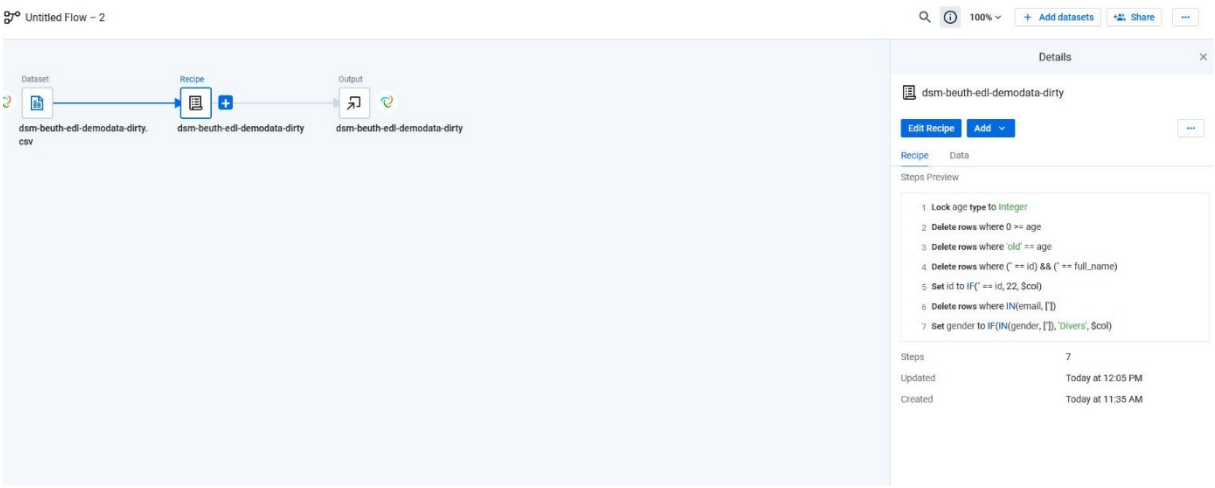
filter type: custom rowType: single row: ('' == id) && ('' == full_name) action: Delete

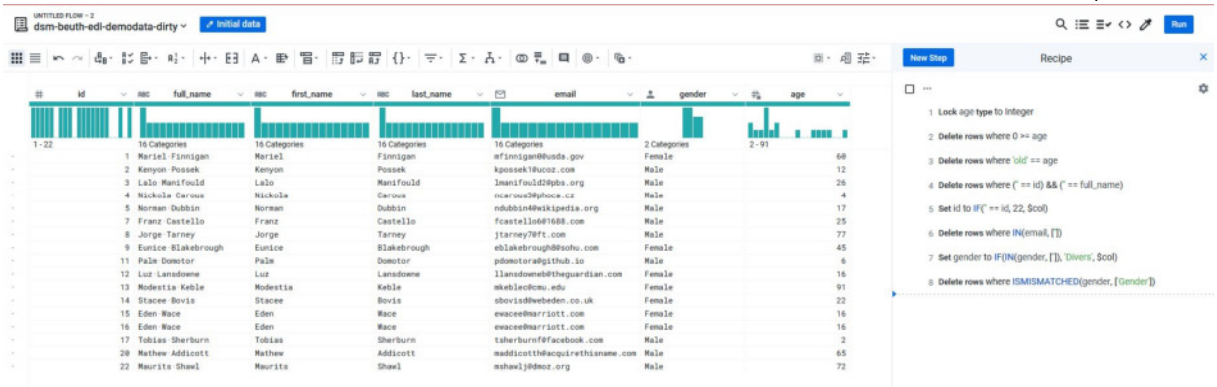
set col: id value: IF('' == id, 22, $col)

filter type: custom rowType: single row: IN(email, ['']) action: Delete

set col: gender value: IF(IN(gender, ['']), 'Divers', $col)

filter type: custom rowType: single row: ISMISMATCHED(gender, ['Gender']) action: Delete





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

- Die Spalte „Demand Loss (MW)“ hat sehr unterschiedliche Werte. Für nicht vorhandene Werte wird sowohl „Unknown“ als auch „N/A“ verwendet. Wenn der Wert bekannt ist wird dieser in unterschiedlicher Notation verwendet und ist daher nicht direkt vergleichbar. Die Daten müssten dazu zunächst skaliert werden
- Time of Restoration ich würde die Zeitangabe in ein 24 Stunden Format überführen um eine besser Vergleichbarkeit auf den ersten Blick zu haben
- Event Description, hier wird zum Teil der genauere Grund mit angegeben, zum Teil wird dieser nicht erwähnt. Ich würde für die Unterscheidung in Wetter bedingt sowie Menschen bedingt noch eine weitere Zeile einfügen
- Geografic Area, hier wird zum Teil keine genaue Angabe, zum Teil der Bundesstaat sowie zum Teil der Bundesstaat mit genauer Region angegeben. Ich würde die Daten dahingehend vereinheitlichen, dass in der Spalte nur keine Angabe oder Bundesstaat angegeben wird. Für die Angabe einer genauen Region würde ich eine weitere Spalte hinzufügen.