Data Cleaning in Tableau Prep

Einleitung

Zur Verfügung stehen nach den 3 Crawlings drei .csv Dateien. Welche in einem nächsten Schritt mit Hilfe von Tableau Prep bereinigt und zur Auswertung zusammengefügt werden. Der gesamte Prozess soll zu einem späteren Zeitpunkt automatisiert in laufen. Manuelle Cleanings werden daher nur bei vereinzelten Datenbausteinen angewendet. Und zwar bei jenen, wo offensichtlich ein Lesefehler vorhanden ist. Ein gutes Beispiel dafür ist: «24 years» statt «24 months». Es ist in diesem Fall schlichtweg unrealistisch, dass ein Studium 24 Jahre dauern soll.

Vorgehen

Die gecrawlten Informationen konnten bereits mit Pandas vollständig automatisiert in eine sinnvolle .csv Struktur gebracht werden. So kommen die Daten bereits in Spalten daher und vereinfachen eine erste Übersicht in Tableau Prep deutlich. Im bevorstehenden Cleaning Prozess geht es also vornehmlich darum, die einzelnen Spalten so zu bereinigen, dass die drei Datenquellen zusammengeführt werden können. Dazu mussten diverse Inkonsistenzen bereinigt werden, welche nachfolgend erklärt sind.

Schritt 1: Vereinigen

Die beiden Datenquellen «test\_master.csv» und «test\_bachelor.csv» werden zusammengeführt. Da diese beide von der gleichen Quelle stammen und daher ihrer Logik nach den gleichen Aufbau haben, macht dies ohne vorhergehende Bereinigungsschritte Sinn.

Schritt 2: Bereinigung von inkonsistenten Spalten

Schrittweise werden jetzt mittels kurzer Befehle (Calculated Field) die einzelnen Spalten bereinigt. Dabei fällt der Bereinigungsprozess mehr oder weniger Komplex aus. So muss zur Bereinigung der Spalte «location» nur an der Kommastelle getrennt werden. Für andere wie «Duration» muss ein mehrstufiger Prozess geschrieben werden. Insgesamt müssen 4 von 10 Spalten bereinigt werden. Die detaillierten Schritte dazu können über Tableau Prep nachverfolgt werden.

Schritt 3: Vereinigung mit der Datenquelle «Ratings.csv»