Datenqualität ist nicht alles, aber alles ist nichts ohne Datenqualität

Georg Franzke

SAS Institute GmbH, Deutschland; Karin.Pees@sas.com

DQ als Voraussetzung in der BigData & Business Analytics Industrie:

Daten sind aus der heutigen Welt nicht mehr wegzudenken. Sie sind fester Bestandteil der gesellschaftlichen Prozesse geworden und Basis vieler Entscheidung.

ABER: Vieles in unserer Welt ist genormt, es gibt Angaben, wie groß Papier sein darf, wie Schrauben zu sein haben und wie krumm eine Banane zu sein hat. Nur Daten können sein wie sie wollen – und die Ergebnisse von Berechnungen, die auf Daten beruhen, sind dann ebenso beliebig.

Das darf nicht sein. Viele Initiativen beschäftigen sich mehr und mehr mit dem Thema Datenqualität. Und immer mehr Behörden verlangen den Nachweis, dass Firmen Ihre Datenqualität im Griff haben (z.B. Solvency 2 für Versicherungen).

Neben diesen Anforderungen ist aber auch vielfach festzustellen, dass Projekte scheitern, weil Projektverantwortliche Ihre Daten und deren Bedeutung nicht im Griff haben.

Die Datenqualität ist daher essenziell für alle Prozesse, die auf Daten beruhen.

Dieser Vortrag geht nun auf die Fassetten von Datenqualität ein und zeigt, wie die Datenqualität gemessen und teilweise auto. Verbessert werden kann.

Der Schwerpunkt liegt im Finden von Datenanomalien (Profiling), dem Messen von Datenanforderungen (Regeln), Ermittlung von doppelten Datensätzen (Clustern) und der Prüfung von Adressdaten. Anhand von live Demos werden dies Möglichkeiten und Verfahren dargestellt um zu verdeutlichen, wie man mit der richtigen Software schneller zum Ergebnis kommt.

Alle Punkte beziehen sich vor allem auf Daten, die manuell erfasst werden.

Die Datenqualität von automatisiert erfassten Daten wird mit Hilfe zusätzlicher statistischer Verfahren ermittelt, die nicht Bestandteil dieser Präsentation sind.