

Fakultät Informatik

# Analyse und Visualisierung von Datenqualität innerhalb eines Data Warehouse.

Bachelorarbeit im Studiengang Informatik

vorgelegt von

Lukas Welsch

Matrikelnummer 3201095

Erstgutachter:	Prof. Dr. Korbinian Riedhammer
Zweitgutachter:	Prof. Dr. Jens Albrecht
Betreuer:	Dipl. Ing. Andreas Sachs
Unternehmen:	Sopra Financial Technology GmbH

© 2020

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.



## Prüfungsrechtliche Erklärung der/des Studierenden

Angaben des bzw. der Studierenden:

Name: Welsch

Vorname: Lukas


Matrikel-Nr.: 3201095

Fakultät: Informatik



Studiengang: Informatik



Semester: Wintersemester  2020/21

### Titel der Abschlussarbeit:

Analyse und Visualisierung der Datenqualität innerhalb eines Data Warehouse

Ich versichere, dass ich die Arbeit selbständig verfasst, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Muss noch ausgefüllt

Ort, Datum, Unterschrift Studierende/Studierender

## Erklärung zur Veröffentlichung der vorstehend bezeichneten Abschlussarbeit

Die Entscheidung über die vollständige oder auszugsweise Veröffentlichung der Abschlussarbeit liegt grundsätzlich erst einmal allein in der Zuständigkeit der/des studentischen Verfasserin/Verfassers. Nach dem Urheberrechtsgesetz (UrhG) erwirbt die Verfasserin/der Verfasser einer Abschlussarbeit mit Anfertigung ihrer/seiner Arbeit das alleinige Urheberrecht und grundsätzlich auch die hieraus resultierenden Nutzungsrechte wie z.B. Erstveröffentlichung (§ 12 UrhG), Verbreitung (§ 17 UrhG), Vervielfältigung (§ 16 UrhG), Online-Nutzung usw., also alle Rechte, die die nicht-kommerzielle oder kommerzielle Verwertung betreffen.

Die Hochschule und deren Beschäftigte werden Abschlussarbeiten oder Teile davon nicht ohne Zustimmung der/des studentischen Verfasserin/Verfassers veröffentlichen, insbesondere nicht öffentlich zugänglich in die Bibliothek der Hochschule einstellen.

Hiermit ☒ genehmige ich, wenn und soweit keine entgegenstehenden Vereinbarungen mit Dritten getroffen worden sind,  
☐ genehmige ich nicht,

dass die oben genannte Abschlussarbeit durch die Technische Hochschule Nürnberg Georg Simon Ohm, ggf. nach Ablauf einer mittels eines auf der Abschlussarbeit aufgebrachten Sperrvermerks kenntlich gemachten Sperrfrist

von Jahren (0 - 5 Jahren ab Datum der Abgabe der Arbeit),

der Öffentlichkeit zugänglich gemacht wird. Im Falle der Genehmigung erfolgt diese unwiderruflich; hierzu wird der Abschlussarbeit ein Exemplar im digitalisierten PDF-Format auf einem Datenträger beigelegt. Bestimmungen der jeweils geltenden Studien- und Prüfungsordnung über Art und Umfang der im Rahmen der Arbeit abzugebenden Exemplare und Materialien werden hierdurch nicht berührt.

Ort, Datum, Unterschrift Studierende/Studierender

Formular drucken

**Datenschutz:** Die Antragstellung ist regelmäßig mit der Speicherung und Verarbeitung der von Ihnen mitgeteilten Daten durch die Technische Hochschule Nürnberg Georg Simon Ohm verbunden. Weitere Informationen zum Umgang der Technischen Hochschule Nürnberg mit Ihren personenbezogenen Daten sind unter nachfolgendem Link abrufbar: <https://www.th-nuernberg.de/datenschutz/>



## Kurzdarstellung

Kurze Zusammenfassung der Arbeit, höchstens halbe Seite. Deutsche Fassung auch nötig, wenn die Arbeit auf Englisch angefertigt wird.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

## Abstract

*Only if thesis is written in English.*

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.



# Inhaltsverzeichnis

<b>1. Einleitung</b>	<b>1</b>
<b>2. Grundlagen</b>	<b>3</b>
2.1. Definition von Datenqualität	3
2.1.1. Datenqualitätsdimensionen	3
2.1.2. Definition / Metriken der Dimensionen	4
2.1.3. Visualisierungen / Charts zur Darstellung der Metriken	4
<b>3. Daten</b>	<b>5</b>
<b>4. Methoden</b>	<b>7</b>
4.1. Stakeholderanalyse	7
4.2. Metriken	9
4.3. Auf Datenebene	10
4.3.1. Machine Learning	10
4.3.2. Visualisierung	10
4.4. Sicherstellung der korrekten Prozesse	11
<b>5. Experimente</b>	<b>13</b>
5.1. ML-Experimente (evtl in genaues Verfahren umbenennen)	13
5.2. Fragebogen	14
5.3. Visualisierung	14
<b>6. Ausblick</b>	<b>15</b>
<b>7. Fazit</b>	<b>17</b>
<b>A. Supplemental Information</b>	<b>19</b>
<b>Abbildungsverzeichnis</b>	<b>21</b>
<b>Tabellenverzeichnis</b>	<b>23</b>
<b>Auflistung</b>	<b>25</b>
<b>Literaturverzeichnis</b>	<b>27</b>

**Glossar . . . . . 29**



# Kapitel 1.

## Einleitung

You can also write footnotes.<sup>1</sup>

Wie kann die Daten Qualität aufbereitet und visualisiert werden, sodass Entscheider schnell und gezielt Maßnahmen einleiten können? -> Wie visualisiert man am besten? -> Wie kann man die DQ messen? (Am besten ohne großen Aufwand durch die Experten, also möglichst viel automatisch machen) ->

versch. Arten visualisieren und dann selbst entscheiden bzw Entscheidungen von Experten mit einfließen lassen Anpassung der Adressen aus den letzten Jahren (Durchschnitt wie oft wird Kundenstammdaten aktualisiert)

It is possible to reference glossary entries as library as an example.

---

<sup>1</sup>Footnotes will be positioned automatically.



# Kapitel 2.

## Grundlagen

### 2.1. Definition von Datenqualität

Datenqualität wird in der Literatur auf verschiedene Arten definiert. Viele Datenqualitätsstudien verwenden Korrektheit als einziges (oder als Haupt-) Datenqualitätsmerkmal. [Wang 96] Datenqualität umfasst jedoch nicht nur die Korrektheit von Daten, sondern auch anderen Dimensionen. Ein Paper von 2011 zeigt den Einfluss von anderen Dimensionen, wie zb. Vollständigkeit auf die Erkennungsrate von Klassifikatoren [Espi 11] Die Autoren schlagen vor den Aufbereitungsprozess und vor allem Datenqualität nicht nur auf die Korrektheit von Daten zu beziehen sondern auch auf die anderen Datenqualitätsdimensionen zu achten.

#### 2.1.1. Datenqualitätsdimensionen

Die Dimensionen Richtigkeit, Vollständigkeit, Konsistenz und Aktualität werden in den meisten Veröffentlichungen genannt, allerdings gibt es keinen Standard, weder im Bezug auf die verwendeten Dimensionen, noch die Definition der Dimensionen. [Scan 02] **Begründung finden! In dieser Arbeit werden diese Definition von Datenqualität verwendet, da es sich am Besten anbietet?** Jede dieser Datenqualitätsdimensionen umfasst eine Facette der Datenqualität.

**Richtigkeit** Richtigkeit wird als Gleichheit zwischen zwei Werten definiert, sodass die Daten die Wirklichkeit korrekt repräsentieren. Beispiel: Eine Person mit dem Namen Max Mustermann ist als Mx Mustermann abgespeichert. Die gespeicherten Daten repräsentieren nicht die Wirklichkeit, sie sind somit nicht richtig.

**Aktualität / temporäre Richtigkeit / Zeitnähe** Als Spezialisierung der Richtigkeit kann die temporäre Richtigkeit gesehen werden, die richtige Repräsentation zu einem bestimmten Zeitpunkt oder Zeitspanne beschreibt. Die temporäre Richtigkeit zum aktuellen Zeitpunkt wird auch als Aktualität bezeichnet.

**Vollständigkeit** Die Daten sind umfangreich genug für die jeweilige Aufgabe. [Wang 96] Da ein Data Warehouse aus relationalen Datenbanken besteht, wird die Vollständigkeit im Folgenden auf relationale Datenbanken bezogen. Die Vollständigkeit wird von [Pipi 02] in drei Klassen kategorisiert.

- Schema: Grad der Daten, die nicht im Schema fehlen
- Spalte: Anzahl der fehlenden Werte innerhalb einer Spalte
- Population: Wenn eine Spalte alle Institute (36,70..) beinhalten sollte, aber welche fehlen, dann herrscht Populationsunvollständigkeit

**Konsistenz** Die gleichen (redundanten) Datensätze haben die gleichen Werte in verschiedenen Tabellen. Ein Beispiel für Konsistenz ist die referentielle Integrität, die sicherstellt, dass Datensätze nur auf existierende Datensätze verweisen.

**Appropriate Amount of Data, Believability,**

### 2.1.2. Definition / Metriken der Dimensionen

Um diese eher abstrakten Definitionen messbar zu machen haben sich einige Verfahren etabliert. Diese lassen sich grob in die Kategorien subjektiv und objektiv aufteilen. [Pipi 02] Bei den subjektiven Verfahren wird die vorherrschende Datenqualität durch Experten geschätzt. Dieses Verfahren ist allerdings nicht nur zeitaufwendig für die Experten, sondern ist auch fehleranfällig. **HIERZITATANGEBEN** []

Eine weitere Möglichkeit besteht darin objektive Verfahren zu verwenden, bei denen mit Hilfe von mathematischen Funktionen die Datenqualität berechnet oder geschätzt wird.

Des weiteren gibt es ein kombiniertes Verfahren, bei denen Grenzwerte von Experten geschätzt und anschließend durch mathematische Funktionen abgeglichen werden. Das kombinierte Verfahren kann durch ein Quadrat visualisiert werden, das an der y-Achse die subjektive Bewertung und auf der x-Achse die objektive Bewertung zeigt. Bei einer guten Datenqualität liegt das Ergebnis im IV. Quadrant. [Pipi 02]

### 2.1.3. Visualisierungen / Charts zur Darstellung der Metriken

Sollte zum Beispiel sehr gut interpretierbar sein. Genaue Vorgaben zu den Visualisierungen müssen allerdings von den Stakeholdern erfragt werden

## Kapitel 3.

### Daten

Die in der Arbeit verwendeten Daten stammen aus dem Data Warehouse einer deutschen Bank. Aufgrund des Datenschutzes wurden personenbezogene Daten pseudonymisiert und anonymisiert.

Ein- und Ausschlusskriterien für die Daten:

grobe Anzahl des Datenbestandes (evtl. auf dem produktiv System) berechnen

Kontakt Daten Stammdaten Kunden

Im vorliegenden Projekt wird zunächst mit den Entwicklungsdaten gearbeitet, da diese im wesentlichen den produktiven Daten entsprechen. Der Vorteil an Entwicklungsdaten liegt darin, dass mit diesen Daten mehr experimentiert werden kann, da diese schon anonymisiert worden sind.

Zunächst sollte sich eine Übersicht über die Daten beschafft werden:

```
1 SELECT VALID_DT, COUNT(VALID_DT)
2 FROM SCHEMA.STAMMDATEN
3 GROUP BY VALID_DT
4 ORDER BY VALID_DT
```

Listing 3.1: Überblick über die Daten verschaffen

Beispiel: (Auszug)

bsp: 1234, 1, 1, -1234, 1000, -2345, 2430 2000-01-01, 2020-10-10, 2020-10-10, 34

Des weiteren befindet sich in der Tabelle der Zinssatz, wenn es sich um ein Darlehen handelt und verschiedene IDs.

Visualisierung auf zwei Arten: Major critical (Sonar) 1. Visualisierung der Metriken (Ergebnisse) 2. Visualisierung der Daten selbst, zb Null Values gelb markiert



## Kapitel 4.

### Methoden

Um die nachfolgenden Methoden zur Analyse und Visualisierung von Datenqualität, zu vergleichen werden sowohl qualitative als auch quantitative Untersuchungen durchgeführt. Für eine Datenqualität, die sich in einem Unternehmen etablieren kann, ist es nötig diese von den Gesichtspunkten aller Stakeholdern zu betrachten. Anhand einer Stakeholderanalyse wird es möglich sein die Probleme der Datenqualität auf zwei Ebenen zu betrachten. Zum einen die der Business-User, die gute Datenqualität benötigen, um Kampagnen (Werbung) gezielt an die richtigen Kunden zu senden. Zum anderen benötigen die Entwickler eine Validierung (Überprüfungsmöglichkeit) für die ETL-Prozesse, die sie entwickelt haben, um sicherzustellen, dass diese keine (neuen) Datenqualitätsprobleme erzeugen.

#### 4.1. Stakeholderanalyse

Mit Hilfe einer Stakeholderanalyse ist es möglich, die negativen Einflüsse zu erkennen und die positiven Einflüsse zu nutzen. Zum Anderen können die Erwartungen der einzelnen Stakeholder erkannt werden und die Projektziele richtig gewichtet werden.

Es besteht die Möglichkeit durch einen Fragebogen die vorherrschende Datenqualität mit Hilfe der Stakeholder zu berechnen. [Pipi02] Dies fällt unter den subjektive Messbarkeit der Datenqualität und *wird im Kapitel 5.Experimente durchgeführt und dargestellt.*

1. Identifikation 2. Information und Analyse 3. Aktionsplanung 4. Monitoring

- Der Benutzer von Daten - Der Entwickler - Der, der zur Einhaltung der korrekten Datenqualität da ist

Nach [Pipi02] gibt es drei Stakeholder: the collector, custodians and consumers of data products.

Aufgrund der Analyse des mittelständischen Unternehmens in der Bankenbranche können folgende Stakeholder identifiziert werden:

Als Stakeholder lassen sich zunächst die Entwickler identifizieren. Diese sind für die Einhaltung einer guten Datenqualität unmittelbar wichtig, da diese die Extraktionsprozesse entwickeln, die zu Fehlern in der Datenqualität führen können. Auf der anderen Seite gibt es die Business User, die Kampagnen für die Banken entwickeln, um beispielsweise neue Kunden zu gewinnen. Auch für diese Gruppe von Personen ist es wichtig, dass die Daten fehlerfrei sind, da sonst (evtl.) Kunden angesprochen werden, die gar nicht relevant für die Kampagne sind. Eine weitere Rolle ist die des Datenmanagers, der überprüft, ob die Datenqualität gut genug ist und diese im Überblick behält und gegebenenfalls Maßnahmen zur Besserung einleitet. Des weiteren ist der Kunde ein Bestandteil einer datenverarbeitenden Abteilung. Dieser erwartet nicht nur, dass die Daten fehlerfrei sind, sondern auch, dass diese immer verfügbar sind. Der Aspekt der Verfügbarkeit der Daten hat nur bedingt etwas mit der Datenqualität zu tun und wird deshalb nicht weiter in dieser Arbeit betrachtet. Für ein umfassendes Projekt, bei dem alle Stakeholder zufriedengestellt werden, müsste dieser Aspekt in die Konzeption einfließen.

Die beschriebenen Stakeholder werden mit ihren Erwartungen an das Projekt in der folgenden Grafik dargestellt.

ID	Wer	Betroffenheit	Erwartung	Macht	Einstellung	Maßnahmen
S1	Entwickler	m	Fehler in der Entwicklung werden angezeigt; Wenig Programmieraufwand	g	Neutral	Erklärung der Notwendigkeit, Zeitvorteil aufzeigen
S2	Business User	h	Daten sollten fehlerfrei sein	g	Positiv	-
S3	Datenmanager	h	Überprüfung der Qualität muss jederzeit und einfach möglich sein	g	Positiv	-
S4	(Bank)-Kunde	h	Daten müssen immer verfügbar und richtig sein	h	Positiv	-

Welchen Stakeholder interessiert welche Dimension?



Die Entwickler benötigen einen Mechanismus oder ein Programm, dass ihnen zeigt, ob durch eine Neuentwicklung Fehler in den Daten entstehen, hier ist besonders die Datenqualitätsdimension Vollständigkeit wichtig.

Stakeholder	Dimension	Begründung
Entwickler	Vollständigkeit	(Wurden alle Daten vom Quellsystem abgeholt)
Business User	Alle Dimensionen	
Datenmanager		
Kunde	Keine Dimension	Geht davon aus, dass die Datenqualität schon geprüft wurde

## 4.2. Metriken

Nach [Pipi02] besteht die Schwierigkeit nicht darin die Metriken zu formulieren, sondern die Datenqualitätsdimension zu definieren, die auf den spezifischen Anwendungsbereich des Unternehmens passt. Aus BI-Sicht ist es enorm wichtig, dass die Daten korrekt sind. Beispielsweise würde es unnötige Kosten verursachen einem Kunden eine Werbung zu zusenden, wenn dieser schon umgezogen ist. Für diese Fehler kann als Indiz die Aktualität verwendet werden und Kunden können nicht angeschrieben werden, wenn die Wahrscheinlichkeit groß ist, dass diese schon umgezogen sind. Nicht nur die Daten selbst müssen den Datenqualitätsanforderungen stimmen, sondern es muss auch sichergestellt werden, dass die ETL-Prozesse fehlerfrei ablaufen. Sonst würden durch die Extraktionen und Anreicherungen neue Datenqualitätsfehler in den Daten entstehen.

**Simple Ratio:** Die Simple Ratio kann für die Dimensionen Richtigkeit, Vollständigkeit und Konsistenz verwendet werden und ist wie folgt aufgebaut: [Pipi02]

$$1 - \frac{\text{erwarteteAnzahl}}{\text{Gesamtzahl}}$$

Das Problem an der Simple Ratio liegt darin, dass nicht unbedingt bekannt ist, wie viele Daten zu erwarten sind. Würden nur die null-Werte gezählt werden ist nicht klar, ob die Daten tatsächlich fehlen, unbekannt sind oder nicht existent. Bei einer Telefonnummer könnte null bedeuten, dass diese nicht bekannt ist, dass derjenige kein Telefon und somit keine Telefonnummer besitzt oder, dass nicht bekannt ist, ob es derjenige eine Telefonnummer besitzt. Diese Metrik kann gut für die Überprüfung der Prozessqualität verwendet werden, da in diesem Fall die erwartete Anzahl und die Gesamtzahl bekannt sind.

Desired outcomes to total outcomes -> geeignet für Richtigkeit, Vollständigkeit, Konsistenz

Maximum operation -> geeignet für Aktualität

Weighted Average.

**Wahrscheinlichkeitsverteilung zur Schätzung der Aktualität** Die Dimension Aktualität ist besonders Interessant für die Datenqualität, da diese als Wahrscheinlichkeit angegeben werden kann. Hierfür gibt es verschiedene Ansätze, der aktuellste liegt darin eine Wahrscheinlichkeits- bzw. Dichtefunktion zu schätzen und anhand dieser konkrete Wahrscheinlichkeiten zu berechnen, ob die Daten schon veraltet sind. Um diese Dichte zu schätzen können externe Daten verwendet werden (wie oft ziehen Menschen um, wie viele Eheschließung gibt, wie oft lassen sich Paare scheiden). Da in diesem Fall auf einen vollständig historisierten Datensatz zurückgegriffen werden kann, kann die Dichtefunktion anhand der durchschnittlichen Lebensdauer der Attribute berechnet werden. Dabei ist sehr kritisch zu betrachten, dass die Daten aus denen die Funktion geschätzt wird, selbst Fehler enthalten können.

Aktualität: - Externe Daten verwenden zb. Anzahl der Eheschließungen / Scheidungen) - historische Daten aus dem Data Warehouse, wie lange ist im Schnitt eine Adresse gültig  
 - Wie lange ist im Schnitt ein Attribut gültig -> Achtung! Daten, die zur Berechnung genommen werden können selbst von schlechter DQ sein -> Das ist kritisch zu betrachten

Stichproben für Schätzung

Vollständigkeit: Metrik entwickeln!

### 4.3. Auf Datenebene

Richtigkeit Um die Richtigkeit zu berechnen gibt es zwei verschiedene Möglichkeiten.

#### 4.3.1. Machine Learning

- Unüberwachtes Lernen findet mögliche Fehlerquellen - Fehlerquellen werden identifiziert - Mit neuen Daten (Labels) können Experimente durchgeführt werden

#### 4.3.2. Visualisierung

- Kibana

## 4.4. Sicherstellung der korrekten Prozesse

Ein weiterer wichtiger Aspekt guter Datenqualität besteht darin die Prozesse der Extraktionen so zu gestalten, dass diese fehlerfrei sind.

Vollständigkeit auf Datensatzebene Vollständigkeit auf Attributwerte (es kommen keine null-values hinzu)

Aktualität die Daten werden schnell genug abgeholt Richtigkeit die Daten werden so abgeholt, dass sie fehlerfrei sind

Ideen: - Source und Target vergleichen - historisch vergleichen, wie viel zu erwarten ist -

Die Daten müssen innerhalb einer vorgelegten Range liegen, damit sichergestellt wird, dass die Daten in dem Zielsystem richtig ankommen.

Zuverlässigkeit, Protokollierung, Dokumentation, Audit der Prozesse Verfügbarkeit, Wartbarkeit, Nachvollziehbarkeit

You can also include listings from a file directly:

```
1 x = 1
2 if x == 1:
3     # indented four spaces
4     print("x is 1.")
```

Listing 4.1: This is an example of included listing



## Kapitel 5.

### Experimente

- Aufbau des ETL-Prozesses -

ML ML -> BASE -> SRC -> BIZ

#### 5.1. ML-Experimente (evtl in genaues Verfahren umbenennen)

- In den Datensatz werden die typischen Fehler eingebaut:

Vollständigkeit - Felder sind null - Zeilen fehlen

Richtigkeit - Daten haben Fehler zb. Umsatz ist zu groß - Prozentzahl des Zinssatzes sind falsch - Datum ist invalide

Aktualität - Daten sind zu Alt

Konsistenz -> Hinzufügen von einigen fiktiven Daten - Daten werden summiert und anhand der Summen wird erkannt, ob diese sehr abweichen

-

Aufbau Experimente: Ziele\* Aufbau\* Ergebnisse\* Interpretation\* Threats\*to\*Validity (Seite 93 <https://userpages.uni-koblenz.de/laemmel/esecourse/slides/perf.pdf>)

Ideen: - Komplexe Funktionen mit Stakeholdern basteln, zb wenn verheiratet dann Alter > 18 - Daten für zb Aktualität müssen definiert werden, ob sie beispielsweise überhaupt verfallen können. Zb Geburtsdatum ändert sich nie; Alter schon - Ist es möglich solche Regeln mit Hilfe von ML abzuleiten oder funktioniert das gar nicht? - Daten vor einem Monat berechnen, wie viele sich ändern müssten (aufgrund von zb Timeliness, correctness) und dann nachschauen wie viele sich tatsächlich geändert haben - Big Data Quality A Quality Dimension evaluation hat zwei konkrete Experimente, dort kann man sich gute Ideen holen. Es wird auch ein Experte zu Rate gezogen, der beispielsweise angibt, welche Daten gelöscht werden können (zb wenn 80% der Attribute fehlen). Es hat auch einige Visualisierungen - Mit SQL: <https://dataform.co/blog/advanced-data-quality-testing>

## 5.2. Fragebogen

- Fragebogen an die Stakeholder (Es besteht die Möglichkeit durch einen Fragebogen die vorherrschende Datenqualität mit Hilfe der Stakeholder zu berechnen. [[Pipi 02](#)])

Auf die verschiedenen Ebenen Aktualität, Richtigkeit, Vollständigkeit und Konsistenz eingehen! Oft ist es besser die Daten nachzufordern, anhand eines möglichen Fehlers kann nicht der Originalzustand wiederhergestellt werden

## 5.3. Visualisierung

Kibana, Graphana

Reduktion der Datenmenge

- Welche Visualisierungen bieten sich an? - Gibt es evtl Visualisierungen, die DQ-Probleme aufzeigen?

## **Kapitel 6.**

### **Ausblick**





## **Kapitel 7.**

### **Fazit**

Datenqualität ist unsichtbar, wenn alles richtig gemacht wird.



## Anhang A.

### Supplemental Information

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Das hier ist der zweite Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Und nun folgt – ob man es glaubt oder nicht – der dritte Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Nach diesem vierten Absatz beginnen wir eine neue Zählung. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

## Abbildungsverzeichnis



## Tabellenverzeichnis





## Auflistung

3.1. Überblick über die Daten verschaffen . . . . .	5
4.1. This is an example of included listing . . . . .	11



## Literaturverzeichnis

- [Espi 11] R. Espinosa Oliva, J. Zubcoff, and J.-N. Mazón. *A Set of Experiments to Consider Data Quality Criteria in Classification Techniques for Data Mining*. June 2011.
- [Pipi 02] L. L. Pipino, Y. W. Lee, and R. Y. Wang. “Data Quality Assessment”. *Communications of the ACM*, Vol. 45, No. 4, pp. 211–218, Apr. 2002.
- [Scan 02] M. Scannapieco and T. Catarci. “Data Quality under a Computer Science Perspective”. *Journal of The ACM - JACM*, Vol. 2, Jan. 2002.
- [Wang 96] R. Y. Wang and D. M. Strong. “Beyond Accuracy: What Data Quality Means to Data Consumers”. *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5–33, March 1996.



# Glossar

**library** A suite of reusable code inside of a programming language for software development. i, 1

**shell** Terminal of a Linux/Unix system for entering commands. i