Eine hohe Datenqualität ist in diesem Vorhaben sehr wichtig. Erstelle ein Konzept, wie der Datenbestand fortlaufend bearbeitet werden kann und die Datenqualität auf ein >97% Level kommt.

* Eingabeprüfung
* Sind die Daten konsistent

# Übung zu Datenqualität und Datenbereinigung 01

Q.1

Notiere Beispiele von Problemen der Datenqualität die eher schwer zu

entdecken sind.

A.1

* Inkonsistenzen in der Schreibweise von Daten (z. B. "Street" vs. "St.")
* Fehlerhafte Dateneingaben, die jedoch semantisch korrekt erscheinen (z. B. falsche Postleitzahlen, aber gültige Adressen)
* Versteckte Duplikate oder Inkonsistenzen in Beziehungen zwischen verschiedenen Datensätzen
* **Falsche Datentypen**: Daten, die in einem unerwarteten Datentyp gespeichert sind, was zu Inkonsistenzen und unerwartetem Verhalten führen kann.
* **Fehlende Werte**: Fehlende Daten, die nicht offensichtlich sind, da sie implizit durch NULL-Werte dargestellt werden.
* **Semantische Inkonsistenzen**: Unterschiedliche Repräsentationen desselben Konzepts in verschiedenen Tabellen oder Spalten, die schwer zu erkennen sind.
* **Historische Datenfehler**: Fehler, die sich im Laufe der Zeit ansammeln und schwer zu identifizieren sind, wenn keine umfassende Datenauditierung durchgeführt wird.
* **Veränderungen der Datenqualität im Laufe der Zeit**: Änderungen in den Datenqualitätsproblemen im Laufe der Zeit, die möglicherweise nicht erkannt werden, wenn keine kontinuierliche Überwachung erfolgt.
* Fehlende Daten

Q.2

Erkunde den Datenbestand der Datei customer.csv.

• Wie gehst Du vor?

• Was fällt Dir auf?

• Identifiziere alle vorkommenden Fehlerarten je Spalte und liste diese auf.

• Welche Struktur und Inhalt haben die Daten?

A.2

1. **Übersicht verschaffen:** Ich würde mir einen Überblick über die vorhandenen Spalten und Daten verschaffen, um die Struktur der Datei zu verstehen.
2. **Spaltenbeschreibung überprüfen:** Anhand der Spaltenüberschriften würde ich versuchen zu verstehen, welche Art von Informationen in jeder Spalte enthalten ist.
3. **Datenqualität überprüfen:** Ich würde die Daten in jeder Spalte auf Fehler oder Inkonsistenzen überprüfen, wie z.B. fehlende Werte, falsches Datenformat, Tippfehler usw.
4. **Häufigkeit von Werten überprüfen:** Ich würde mir ansehen, wie häufig bestimmte Werte in den Spalten vorkommen, um mögliche Muster oder Ungenauigkeiten zu erkennen.
5. **Beziehungen zwischen den Spalten analysieren:** Ich würde prüfen, ob es Beziehungen zwischen den Spalten gibt und ob diese konsistent sind.

Es können bestimmte Unregelmäßigkeiten oder Inkonsistenzen auftreten, die weitere Untersuchungen erfordern.

Alle vorkommenden Fehlerarten

* **Fehlende Werte**: Es gibt einige leere Zellen in den Spalten: fehlende Faxnummern, contact\_name, contact\_title, address, city, region, postal\_code
* **Inkonsistente Daten**: Letzte Spalte
* **inkonsistente Datenformatierung**: unterschiedliche Formate für city, postal\_code, country, phone, letzte Spalte
* **Ungültige Werte**: Es können ungültige Werte auftreten, z.B. falsch geschriebene Ländernamen: country.
* **fehlende Spaltennamen**

Die Struktur der Daten ist tabellarisch, wobei jede Zeile einen Datensatz darstellt und jede Spalte eine bestimmte Eigenschaft oder Information über den Kunden enthält, wie z.B. Kunden-ID, Firmenname, Kontaktname, Adresse, Stadt, Land usw.

Q.3

Welche Folgen können eine schlechte Datenqualität haben?  
  
A.3

* Falsche Geschäftsentscheidungen aufgrund ungenauer oder unvollständiger Daten.
* Verlust des Vertrauens von Kunden und Partnern aufgrund von Fehlern in den Daten.
* Zeit- und Ressourcenverschwendung durch manuelle Korrekturen und Nacharbeit.
* Rechtliche und regulatorische Probleme aufgrund von Datenschutzverletzungen oder Nichteinhaltung von Vorschriften.
* Fehlintepretation von Daten, die nicht klar benannt werden.

Q.4

Wo setzt Du an, um die Datenqualität langfristig zu verbessern? Bitte nenne

Beispiele.  
  
A.4

* Automatisierte Validierungsprozesse implementieren, um fehlerhafte Daten frühzeitig zu erkennen.
* Datenbereinigungs- und Standardisierungsregeln entwickeln und anwenden.
* Schulungen für Mitarbeiter zur korrekten Dateneingabe und -pflege durchführen.
* Regelmäßige Überprüfung und Aktualisierung der Daten durchführen.
* Die Einführung eines Data Governance-Frameworks zur Verantwortungsübernahme für Datenqualität und -integrität erwägen.

Q.5

Du möchtest einen Datensatz (siehe die Dateien customer.csv sowie orders.csv) hinsichtlich der Datenqualität begutachten.

• Macht der Einsatz beschreibender Verfahren der Statistik hierbei Sinn?

• Falls ja, welche Verfahren kannst Du bei customer.csv sowie orders.csv zum Einsatz bringen?

• Von was hängt deren Einsatz ab?

A.5

* Für customers.csv lohnt sich eine Nutzung von statistischen Auswertungen im Sinne von Minimalwerten vs. Maximalwerten nicht – eine Adresse wird sich nicht so häufig ändern.
* Für orders.csv stehen uns mehr quantitative Datensätze, ein Blick auf kleinste vs. größte Umsätze pro Bestellung könnte relevante Ergebnisse liefern. Ein Blick über die Geschäftsentwicklung könnte Bestellmuster erkennbar machen (z. B. „am Anfang war alles langsam, nun bekommen wir größere Orders.“)
* Erstens sollten quantitative Daten zur Verfügung stehen, die man auswerten kann. Einen Mittelwert kann man über Bestellungen und deren Wert vergleichen.

Q.6

Welche spezifischen Probleme können bei der Integration aus mehreren Datenquellen auftauchen? Auf was ist hierbei ggf. zu achten?

A.6

* Unterschiedliche Datenformate und -strukturen zwischen den Quellen.
* Fehlende oder inkonsistente Schlüssel zur Datenverknüpfung.
* Doppelte Datensätze oder fehlende Datenbereinigung vor der Integration.
* Unterschiedliche Datenqualität zwischen den Quellen.
* Fehlende oder unvollständige Daten.

Um diese Probleme zu minimieren, ist es wichtig:

* Einheitliche Datenstandards und -formate zu verwenden.
* Datenbereinigungs- und -standardisierungsprozesse einzurichten.
* Datenqualitätsprüfungen und -tests durchzuführen.
* Mechanismen für die Datenvalidierung und -überprüfung zu implementieren.
* Dokumentation der Integrationsprozesse und -entscheidungen zu führen.

Q.7

Weshalb ist es von Vorteil ein Data Dictionary für die betreffenden Daten zu

haben?

A.7

* Eine klare Definition der Datenattribute und ihrer Bedeutung liefert.
* Einheitliche Standards für die Dateneingabe und -pflege festlegt.
* Die Datenverständnis und -verwendung für alle Benutzer verbessert.
* Als Referenz für die Datenqualitätsprüfung und -verbesserung dient.
* Kann als „single source of truth“ wirken: klare Definition von Daten, inklusive was Daten nicht liefern können.

Q.8

Inwiefern ist die Entstehung bzw. die Erhebung der Daten aus Sicht der Datenqualität von Relevanz?

A.8

* Fehlerhafte Dateneingaben oder -sammlungen die Datenqualität von Anfang an beeinträchtigen können.
* Proaktive Maßnahmen während des Datenerfassungsprozesses helfen, die Qualität der Daten zu verbessern.
* Eine klare Datenstrategie und -richtlinien von Anfang an die Grundlage für hochwertige Daten legen.
* Die Entstehung und Erhebung von Daten ist entscheidend für deren Qualität, da sie den Grundstein für die Datenqualität

Q.9

Inwiefern können Data Understanding (Datenverständnis) und Datenqualität zusammenhängen?

A.9

Data Understanding und Datenqualität können zusammenhängen, da ein tieferes Verständnis der Daten dazu beiträgt, potenzielle Qualitätsprobleme zu erkennen und zu adressieren. Wenn Benutzer die Struktur, den Inhalt und die Einschränkungen der Daten verstehen, können sie besser beurteilen, wie vertrauenswürdig und nützlich die Daten für ihre Analyse oder Entscheidungsfindung sind.

**DATA Qualität des Zoo Pirmasens**

Data-Qualität ist entscheidend im Bereich Data Engineering. Hier sind die einigen wichtigsten Aspekte zu berücksichtigen:

1. **Übersicht verschaffen:** Es wäre wichtig, sich einen umfassenden Überblick über die vorhandenen Spalten und Daten zu verschaffen, um die Struktur der Datei besser zu verstehen
2. **Spaltenbeschreibung überprüfen:** Es ist wichtig, die Spaltenüberschriften genau zu prüfen, um zu verstehen, welche Art von Informationen in jeder Spalte enthalten ist.
3. **Häufigkeit von Werten überprüfen:** Es ist wichtig, die Häufigkeit bestimmter Werte in den Spalten zu analysieren, um mögliche Muster oder Ungenauigkeiten zu erkennen.
4. **Datenvalidierung**: Implementiere Datenvalidierungsregeln für jedes Attribut, um sicherzustellen, dass die eingegebenen Daten gültig sind.
5. **Referenzielle Integrität**: Stelle sicher, dass alle Beziehungen zwischen den Tabellen durch Fremdschlüsselbeziehungen festgelegt sind. Dadurch wird sichergestellt, dass nur gültige Werte in Beziehungstabellen gespeichert werden können.
6. **Duplikatserkennung**: Implementiere Mechanismen zur Erkennung von Duplikaten, insbesondere in Tabellen. Dies kann durch regelmäßige Überprüfungen und Bereinigungsprozesse erfolgen.
7. **Vollständigkeit der Daten**: Verwende NOT NULL-Constraints für alle erforderlichen Attribute, um sicherzustellen, dass keine wichtigen Informationen fehlen. Stelle sicher, dass alle erforderlichen Felder ausgefüllt sind, bevor Datensätze gespeichert werden können.
8. **Konsistenz der Daten**: Überprüfe regelmäßig die Konsistenz der Daten, insbesondere bei Beziehungen zwischen Entitäten wie Eltern-Kind-Beziehungen bei Tieren. Stelle sicher, dass die Zuordnungen korrekt sind und keine ungültigen Beziehungen vorhanden sind.
9. **Datenbereinigung**: Implementiere Prozesse zur regelmäßigen Bereinigung und Aktualisierung der Daten, um veraltete oder fehlerhafte Informationen zu entfernen. Dies kann durch automatisierte Skripte oder manuelle Überprüfungen erfolgen.
10. **Überwachung und Berichterstattung**: Implementiere Mechanismen zur Überwachung der Datenqualität und zur Erstellung von Berichten über potenzielle Probleme oder Unregelmäßigkeiten. Dies ermöglicht es, Probleme frühzeitig zu erkennen und zu beheben.

**Mitarbeiter:**

* Implementiere Validierungsregeln für Felder wie Telefonnummer, E-Mail-Adresse und Postleitzahlen, um sicherzustellen, dass sie korrekt formatiert sind.
* Stelle sicher, dass die Anstellungsdaten-sinnvoll sind und im akzeptablen Zeitrahmen liegen.
* Implementiere Dupliterkennung, um sicherzustellen, dass keine doppelten Mitarbeiterdatasäzte vorhanden sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.
* Implementiere Validierungsregeln für die Kontakinformationen der Mitarbeiter, um sicherzustellen, dass sie korrekt formatiert sind.

**Tier:**

* Überprüfen die Konsistenz der Daten zu den Eltern-Tierbeziehungen, um sicherzustellen, dass sie korrekt sind und keine fehlenden oder ungültigkeit Verweise enthalten.
* Implementiere Dupliterkennung, um sicherzustellen, dass keine doppelten Tierdatensäzte vorhanden sind.
* Überwache regelmäßig die Genauigkeit der Geburts- und Sterbedaten, um sicherzustellen, dass sie korrekt und aktuell sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.

**Tierpfleger:**

* Validiere die Zuordnung von Tierpflegern zu Tieren, um sicherzustellen, dass sie sinnvoll ist und keine ungültigen Verweise enthält.
* Stelle sicher, dass Tierpfleger nicht für zu viele Tiere gleichzeitig verantwortlich sind und die Arbeitsbelastung angemessen ist.
* Implementiere Dupliterkennung, um sicherzustellen, dass keine doppelten Tierpflegedatensäzte vorhanden sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.
* Implementiere Validierungsregeln für die Kontakinformationen der Tierpfleger, um sicherzustellen, dass sie korrekt formatiert sind.

**Gehege:**

* Validierung der Zuordnung von Tieren zu Gehegen, um sicherzustellen, dass sie den Bedürfnissen der Tiere entsprechen und keine Konflikte entstehen.
* Überprüfung der Gehegebelegung auf Überschneidungen oder Überkapazitäten, um eine effiziente Nutzung der Ressourcen sicherzustellen.
* Implementierung von Regeln zur Konsistenz der Gehegeattribute, um sicherzustellen, dass sie korrekt und vollständig sind.
* Überprüfen die Konsistenz der Gehegearten und Positionen, um sicherzustellen, dasss sie korrekt und kossistent sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.

**Tierarzt:**

* Überwache die Zuordnung von Tieräztrn zu Tieren, um sicherzustellen, dass sie sinnvoll ist und keine ungültigen Verweise enthält.
* Implementiere Validierungsregeln für die Kontakinformationen der Tierärzte, um sicherzustellen, dass sie korrekt formatiert sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.

**Lieferant:**

* Überprüfen regelmäßig die Konditionen und Rabatte der Lieferanten, um sicherzustellen, dass sie aktuell und konsistent sind.
* Implementiere Validierungsregeln für die Kontakinformationen der Tierärzte, um sicherzustellen, dass sie korrekt formatiert sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.

**Futterart / Futter:**

* Validierung der Lieferanteninformationen und Futterarten, um sicherzustellen, dass sie korrekt und aktuell sind.
* Überwachung der Futterlieferungen und -verbräuche, um sicherzustellen, dass genügend Futter vorhanden ist und keine Engpässe auftreten.
* Implementierung von Duplikaterkennung für Futterarten, um doppelte Datensätze zu vermeiden.
* Überprüfung auf fehlende oder ungültige Werte, um Datenintegrität sicherzustellen.
* Implementiere Dupliterkennung, um sicherzustellen, dass keine doppelten Liferantdatasäzte vorhanden sind.

**Futterverbrauch / Lager / Lagerbestand:**

* Überwachung des Futterverbrauchs und der Lagerbestände, um sicherzustellen, dass genügend Futter vorhanden ist und rechtzeitig nachbestellt wird.
* Implementierung von Validierungsregeln für Lagerdaten, um sicherzustellen, dass sie korrekt und aktuell sind.
* Überprüfung auf fehlende Werte oder Inkonsistenzen in den Lagerbeständen, um Datenintegrität sicherzustellen.

**Krankheit / Krankheitsverlauf:**

* Implementierung von Validierungsregeln für Diagnosen und medizinische Behandlungen, um sicherzustellen, dass sie korrekt formatiert und dokumentiert sind.
* Überprüfung der Konsistenz der Krankheitsverläufe, um sicherzustellen, dass sie chronologisch und vollständig sind.
* Überprüfung auf fehlende Werte oder ungültige Einträge, um die Datenintegrität zu gewährleisten.