**Quality Data Report**

**1. Introduction**

This report analyzes the data quality of the following files:

* Equipments.xlsx
* Materials.txt
* Data Climate
* Notifications
* Suppliers.csv

The objective is to identify potential issues such as missing values, type errors, outliers, and duplicates in order to propose corrective actions that ensure data integrity.

**2. Data Analysis**

**2.1. Equipments.xlsx (Total: 867 equipment records)**

* Missing values detected in several columns:
  + *Equipment\_Name* (17 values)
  + *Category* (28 values)
  + *CO2\_Emissions\_kg* (22 values)
* Duplicate records:
  + *Equipment\_ID*: 174 duplicates
  + *Equipment\_Name*: 926 duplicates
  + *Category*: 935 duplicates
* Most frequent values:
  + *Equipment\_Name*: "Machine d'Inspection Visuelle"
  + *Category*: "Transformation"
  + *Maintenance\_Cycle*: "Mensuelle"

**2.2. Materials.txt (Total: 855 material records)**

* Missing values found in:
  + *Material\_ID* (20 values)
  + *Material\_Name* (9 values)
  + *Type* (8 values)
* 999 duplicate records identified.
* Most frequent values:
  + *Material\_Name*: "Eau déminéralisée"
  + *Type*: "Produit laitier"
  + *Supplier\_ID*: "SUPPLIER\_31"

**2.3. Data Climate**

* Missing values:
  + *Temperature*: 30 values
  + *Humidity*: 40 values
  + *CO2\_Concentration*: 35 values
* Outliers detected:
  + 257 anomalies in *Temperature*
  + 655 anomalies in *CO2\_Ambient\_Level*

**2.4. Notifications (Total: 780 notifications)**

* Missing values:
  + *Severity*: 15 values
  + *Category*: 10 values
* Most frequent values:
  + *Category*: "Stock"
  + *Description*: "Stock critique détecté" for material "Lait"

**2.5. Suppliers.csv (Total: 83 suppliers)**

* All columns are complete, but some recurring values raise concerns:
  + 22 suppliers are located in "Tunis"
  + 50 suppliers have *Transport\_Type* set to "Unknown"
* Outliers detected in *Renewable\_Energy\_Percentage*, with values below 2% or above 99%.

**3. Recommendations**

**3.1. Handling Missing Values**

* Replace numerical missing values with the mean or median.
* Fill categorical missing values with "Inconnu".

**3.2. Removing Duplicates**

* Eliminate redundant records to ensure data consistency.

**3.3. Managing Outliers**

* Replace extreme values with an adjusted mean.

**3.4. Data Validation**

* Convert dates into a standardized format and check their consistency.

**4. Implementation**

**4.1. Removing Duplicates (Python Example)**import pandas as pd

# Load files

df\_equipments = pd.read\_excel('Equipments.xlsx')

df\_materials = pd.read\_csv('Materials.txt', delimiter='\t')

# Remove duplicates

df\_equipments.drop\_duplicates(subset=['Equipment\_ID'], keep='first', inplace=True)

df\_materials.drop\_duplicates(subset=['Material\_ID'], keep='first', inplace=True)

# Save cleaned files

df\_equipments.to\_excel('Equipments\_cleaned.xlsx', index=False)

df\_materials.to\_csv('Materials\_cleaned.txt', index=False, sep='\t')

**4.2. Imputing Missing Values (Example with Mean)**

# Impute missing numerical values

df\_equipments['CO2\_Emissions\_kg'].fillna(df\_equipments['CO2\_Emissions\_kg'].mean(), inplace=True)

df\_materials['Carbon\_Footprint\_per\_Unit\_kgCO2e'].fillna(df\_materials['Carbon\_Footprint\_per\_Unit\_kgCO2e'].mean(), inplace=True)

# Impute categorical missing values

df\_equipments['Location'].fillna('Inconnu', inplace=True)

df\_materials['Supplier\_ID'].fillna('Inconnu', inplace=True)

# Save cleaned files

df\_equipments.to\_excel('Equipments\_cleaned.xlsx', index=False)

df\_materials.to\_csv('Materials\_cleaned.txt', index=False, sep='\t')  
**4.3. Handling Outliers**

# Load supplier data

df\_suppliers = pd.read\_csv('Suppliers.csv')

# Replace extreme values in Renewable\_Energy\_Percentage

df\_suppliers['Renewable\_Energy\_Percentage'] = df\_suppliers['Renewable\_Energy\_Percentage'].apply(

lambda x: df\_suppliers['Renewable\_Energy\_Percentage'].mean() if x < 2 or x > 99 else x

)

# Save cleaned file

df\_suppliers.to\_csv('Suppliers\_cleaned.csv', index=False)  
**4.4. Validating Data Types**

# Ensure numerical format for Estimated\_Lifetime\_Years

df\_equipments['Estimated\_Lifetime\_Years'] = pd.to\_numeric(df\_equipments['Estimated\_Lifetime\_Years'], errors='coerce')

# Convert dates to datetime format

df\_materials['Arrival Date'] = pd.to\_datetime(df\_materials['Arrival Date'], errors='coerce')

# Save cleaned files

df\_equipments.to\_excel('Equipments\_cleaned.xlsx', index=False)

df\_materials.to\_csv('Materials\_cleaned.txt', index=False, sep='\t')  
  
**5. Conclusion**

This analysis identified several data quality issues in the files *Equipments.xlsx*, *Materials.txt*, *Data Climate*, *Notifications*, and *Suppliers.csv*. Implementing the above recommendations, along with the provided practical solutions and code examples, will enhance the reliability of these datasets and facilitate future analyses.