
🧹 Data Assessment and Cleaning – Full Detailed Notes

📊 What is Data Assessment?

Data Assessment is the process of **examining a dataset** to:

- Understand its **structure** and **quality**
- Identify any **problems** (like missing, inconsistent, or duplicate data)
- Decide how much cleaning or preparation is required before analysis

It is the **first key step** in the data cleaning pipeline.

💡 Goals of Data Assessment

- Spot issues that may affect analysis or models
- Document data types, missing values, and unusual values
- Understand whether the dataset is ready for analysis

❌ Types of Unclean Data

Here are common types of “dirty” or problematic data:

Type	Description	Examples
Missing Data	Blank or null values	NaN, empty cells
Duplicate Rows	Exact copies of rows	Repeated entries
Inconsistent Values	Different formats for same thing	"male", "Male", "M"
Outliers	Very large or very small unexpected values	Age = 999
Wrong Data Types	Data stored in the wrong format	Date as string, price as text
Invalid Entries	Logically impossible values	Age = -5, Salary = "abc"
Misspelled Categories	Typos in labels	"Femle", "femlae" instead of "Female"
Mixed Units or Scales	Units not standardized	km vs miles

📥 Loading the Data

We usually use Python (e.g., pandas) to load the dataset:

```
import pandas as pd

df = pd.read_csv("your_data.csv")
```

Then start exploring:

```
df.head()        # Preview first few rows

df.shape         # Rows and columns

df.columns       # List of columns

df.dtypes        # Data types

df.info()        # Summary of nulls and types

df.describe()    # Stats for numerical columns
```

📝 Writing a Summary of the Dataset

Create a table to summarize key points:

Feature	Data Type	Missing Values	Unique Values	Min	Max	Mean
Age	Integer	4	55	0	90	36.4
Gender	Object	0	2			
Salary (USD)	Float	10	1000+	0	200k	55k

This helps to quickly see where problems lie.

🧾 Column Descriptions (Data Dictionary)

This is a **human-readable explanation** of what each column means. It's critical for future users and even for yourself later.

Column	Description
CustomerID	Unique ID for each customer
Age	Customer age in years
Gender	Male or Female

Column	Description
Salary	Estimated annual salary in USD
Purchase	1 if made a purchase, 0 otherwise

Include units, encoding, and any assumptions.

➕ Additional Data Information

Sometimes extra metadata is needed to fully understand the data:

- Units (e.g., income in USD, height in cm)
- Encoding (e.g., 1=Yes, 0=No)
- Transformations applied (e.g., log-transformed)
- Data source (survey, API, sensor, etc.)
- Data collection date (relevant for timeliness)

🧠 Types of Data Assessment

There are **two kinds** of assessment methods:

🔍 Manual Assessment (Visual / Google Sheets)

This is when **you inspect the data visually** — often using **Google Sheets or Excel**.

Examples of manual methods:

- Open CSV in Google Sheets
- Scroll through rows to spot missing values or formatting issues
- Use built-in sorting, filters, and charts to find problems
- Insert bar charts, histograms, or pivot tables manually

When to use:

- Small datasets
- Early exploration
- When working with non-programmers
- When visual understanding is more important

🤖 Automatic Assessment (Code-Based / Python)

This means **using Python or libraries like pandas** to programmatically inspect the data.

Common functions:

```
df.info()            # Types and null counts
df.describe()        # Summary stats for numeric
df.isnull().sum()     # Missing values per column
df.duplicated().sum()   # Total duplicate rows
df.nunique()          # Unique values in each column
df['Gender'].value_counts()  # Frequency of categories
```

When to use:

- Large datasets
- Reproducible workflows
- Automation/pipelines
- Part of EDA process

📐 Data Quality Dimensions

These are standard criteria for evaluating whether your data is “clean” or not:

Dimension	Meaning
Accuracy	Are values correct (true, verified)?
Completeness	Are values missing?
Consistency	Are values uniform across the dataset?
Validity	Do values follow the correct format or rules?
Uniqueness	Are duplicate entries avoided?
Timeliness	Is the data recent/up-to-date?

✅ Use these to **evaluate** your dataset and guide your cleaning steps.

🧼 What is Data Cleaning?

Once assessment is complete, data cleaning begins. It includes:

Task	**Example**
Fill or drop missing values	df.fillna(), df.dropna()
Remove duplicates	df.drop_duplicates()
Fix data types	Convert string to datetime
Standardize values	"Male", "male" → "Male"
Handle outliers	Remove or treat extreme values
Encode categories	Label Encoding, OneHot

⚠️ Always clean data based on **what you observed during assessment**.