A messy dataset, also known as dirty data, contains inaccuracies, inconsistencies, or errors that can hinder analysis or modeling. These issues can include incorrect data types, missing values, duplicates, misspellings, and logical inconsistencies. Cleaning messy data is crucial for ensuring data quality and reliability.

Here's a breakdown of what makes data messy and how to address it:

Common Issues in Messy Datasets:

Inconsistent Data Types: Data should be in the correct format (e.g., numbers as numbers, dates as dates).

Missing Values: Gaps in the data can be caused by various reasons and need to be handled appropriately.

Duplicates: Redundant entries can skew results and should be removed.

Outliers: Extreme values that deviate significantly from the norm can distort analysis and need to be addressed.

Inconsistent Formatting: Differences in capitalization, spacing, or abbreviations can create inconsistencies.

Incorrect Values: Data entry errors can lead to inaccurate information.

Logical Inconsistencies: Data that contradicts itself or other known information.

Cleaning Techniques:

Data Validation: Checking data against defined rules and constraints to identify errors.

Data Transformation: Converting data into a consistent format, such as standardizing date formats or recoding categorical variables.

Handling Missing Values: Removing rows or columns with missing data, imputing missing values with statistical measures (mean, median), or using more advanced techniques according to Stony Brook University.

Handling Outliers: Identifying and dealing with extreme values, potentially removing them or using robust statistical methods.

Removing Duplicates: Identifying and removing redundant entries according to a Reddit thread.

Standardizing Formats: Ensuring consistent capitalization, spacing, and abbreviations.

Data Profiling: Analyzing the data to understand its characteristics and identify potential issues.

Benefits of Cleaning Data:

Improved Data Quality: Ensures data is accurate and reliable.

More Accurate Analysis: Reduces errors and biases in analysis and modeling.

Better Decision Making: Enables more informed and reliable decisions based on clean data.

📓 Lecture: Data Cleaning Using Excel Midterm

🎯 Learning Objectives

By the end of this lecture, students will be able to:

1. Define data cleaning and explain its importance in analytics.
2. Identify common data quality issues such as missing values, duplicates, and outliers.
3. Apply Excel tools and functions for cleaning and preparing data.
4. Perform hands-on data cleaning using a sample dataset.

□ 1. Introduction to Data Cleaning

Definition:
Data cleaning is the process of detecting and correcting (or removing) errors and inconsistencies in data to improve its quality.

Why It's Important:

- Ensures **accuracy** in analysis results.
- Prevents **misleading insights**.

- Saves time during later stages of analysis.

Common Problems in Raw Data:

- Missing values
- Duplicate records
- Inconsistent formats
- Outliers
- Extra spaces or special characters
- Wrong data types

🔍 2. Common Data Cleaning Tasks in Excel

Below are common cleaning techniques and the Excel tools/functions to use:

Task	Excel Tool/Function	Example
Remove duplicates	**Data → Remove Duplicates**	Duplicate names in a list
Handle missing values	**Filter / IF / Fill Down**	Blank age column
Trim extra spaces	`TRIM()`	" Anna " → "Anna"
Standardize text case	`UPPER()`, `LOWER()`, `PROPER()`	"ROZAIDA" → "Rozaida"
Find & replace errors	**Find & Replace (Ctrl+H)**	"Filipinas" → "Philippines"
Correct data types	Format cells (Number, Date, Text)	Text dates → date format
Identify outliers	**Conditional Formatting**	Highlight ages > 100
Combine columns	`CONCATENATE()` or `&`	First Name + Last Name
Split columns	**Text to Columns**	Full name → First name, Last name

---

🛠 3. Step-by-Step: Data Cleaning in Excel

Example Dataset (Messy Data)

Name	Age	Salary	Join Date	Department
Anna	21	$50,000	1/3/2023	Sales
Ben		$55,000	Jan 5, 2023	Sales
Clara	22	N/A	2023-01-07	HR
Anna	21	$50,000	1/3/2023	Sales
dan	24	$58,000	1-9-2023	Finance
Ella	23	$61,000	1.12.2023	SALES

Cleaning Process:

1. **Remove duplicates**:
   - Go to **Data → Remove Duplicates** → Select columns.
2. **Handle missing values**:
   - Filter blanks and either fill manually, use averages, or mark as “Unknown.”
3. **Standardize department names**:
   - Use **PROPER()** or Find & Replace: "SALES" → "Sales".
4. **Fix inconsistent date formats**:
   - Select column → Format Cells → Date.
5. **Remove "N/A" and symbols from salary**:
   - Use **Find & Replace** or `SUBSTITUTE()` function.
6. **Trim spaces**:
   - Use `=TRIM(cell)`.
7. **Capitalize names**:
   - Use `=PROPER(cell)`.

💻 4. Hands-on Activity

Task:
Download a messy dataset (provided by instructor), clean it in Excel, and prepare it for analysis by:

- Removing duplicates
- Standardizing text formats
- Filling or marking missing data
- Correcting data types
- Removing unnecessary spaces or symbols

Expected Output:
A clean Excel sheet with consistent, accurate, and analysis-ready data.

📌 5. Best Practices in Data Cleaning

- Always keep a copy of the raw dataset before cleaning.
- Document every change made for reproducibility.
- Use Excel's **Filter** and **Sort** to inspect data before making changes.
- Validate data after cleaning to ensure accuracy.

Lecture: Data Cleaning Using Power BI

1. Learning Objectives

By the end of this lecture, students should be able to:

- Understand the role of data cleaning in the analytics process.
- Use Power BI's Power Query Editor to clean, transform, and prepare datasets.
- Handle missing values, duplicates, and outliers in Power BI.
- Apply common data transformation techniques such as merging, splitting, and formatting columns.

---

2. Introduction

- **Hook:** *"Data cleaning is like laundry for your dataset—before you can wear it in your analysis, you need to make sure it's fresh and neat."*
- Explain **why** data cleaning matters:
    - Dirty data leads to wrong insights.
    - In real-world datasets, issues like inconsistent formatting, missing values, and duplicates are common.

- Highlight that Power BI's **Power Query** is a visual, no-code/low-code tool for cleaning and shaping data.

---

3. Steps in Data Cleaning with Power BI

Step 1: Loading Data

- Open Power BI Desktop.
- Import data from:
  - Excel/CSV files
  - SQL database
  - Web or API
- Select **Transform Data** to open Power Query Editor.

Step 2: Removing Unnecessary Columns

- Identify irrelevant columns (e.g., unused IDs, metadata).
- Right-click column → **Remove** or use **Choose Columns**.

Step 3: Handling Missing Values

- Detect missing values:
  - Filter for null/blank values.
- Options for handling:
  - Replace nulls with a default value (**Replace Values**).
  - Remove rows with null values (**Remove Rows → Remove Blank Rows**).

Step 4: Removing Duplicates

- Highlight the relevant columns.
- **Remove Duplicates** option to keep unique records.

Step 5: Fixing Data Types

- Ensure numeric columns are set to **Whole Number/Decimal Number**.
- Dates set to **Date** type.
- Text columns set to **Text** type.

Step 6: Formatting Text

- Change text case (**Transform → Format → UPPERCASE/lowercase/Capitalize Each Word**).
- Trim unnecessary spaces (**Transform → Format → Trim**).

Step 7: Splitting and Merging Columns

- **Split Column** by delimiter (e.g., separating "Full Name" into First/Last name).
- **Merge Columns** to create combined fields (e.g., City + State).

Step 8: Filtering Outliers

- Sort column values and manually check for anomalies.
- Use **Filters** to remove extreme values if not relevant.

Step 9: Saving and Loading Clean Data

- Click **Close & Apply** to load cleaned data into Power BI.
- Ready for visualization.

Lecture: Data Cleaning Using Python (Pandas)

---

1. Learning Objectives

By the end of this lecture, students should be able to:

- Understand the importance of data cleaning in the data science workflow.
- Load and inspect datasets using **Pandas**.
- Handle missing values, duplicates, and inconsistent formats.

- Perform basic data transformations for analysis readiness.

---

2. Introduction

- **Hook:** *"80% of a data scientist's time is spent cleaning data, and the other 20% is complaining about it."* 😉
- Data cleaning ensures:
  - Accuracy of analysis
  - Consistency in reporting
  - Removal of errors and irrelevant data
- **Python's Pandas library** is one of the most popular tools for this purpose.

---

3. Basic Workflow in Data Cleaning (Python + Pandas)

```python
import pandas as pd

# Load dataset
df = pd.read_csv("data.csv")

# View first 5 rows
print(df.head())
```