Programming with Data
Summer 2025

Day 2

Cleaning Data for Data Analytics

Applying Tidy Data Principles

Why Data Cleaning Matters

- 80% of data analysis time is spent cleaning and preparing data
- Garbage in → Garbage out: insights depend on data quality
- Foundation for valid, actionable analytics

Common Problems with Raw Data

- Missing values
- Duplicates
- Inconsistent formats
- Mixed data types in one column
- Wide vs. long formats

Understanding Data Types in Spreadsheets

- Text (String): e.g., "Product Name"
- Number (Integer/Float): e.g., 42 or 3.14
- Date/Time: e.g., 6/3/2025 or 12:45 PM
- Boolean (TRUE/FALSE): e.g., “Is Active?”
- Categorical: e.g., ["Small", "Medium", "Large"]


Tidy Data
Based on the Paper by Hadley Wickham

What is Tidy Data?

- Tidy data sets are easy to manipulate, model, and visualize.
- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Why Tidy Data Matters

- Simplifies data analysis workflows.
- Integrates well with statistical software.
- Promotes consistency across datasets.
- Reduces cognitive load on analysts.

Messy Data Problems

- Column headers are values, not variable names.
- Multiple variables stored in one column.
- Variables stored in both rows and columns.
- Multiple types of observational units in the same table.

Tidy Data Principles

1. Each variable forms one column.
2. Each observation forms one row.
3. Each type of observational unit forms one table.

The image below shows the contrast between messy and tidy data formats.

Visualizing Tidy Data Structure

messy

	id	city	hwy
1	car1	19	24
2	car2	20	30
3	car3	29	35

tidy

	id	roadtype	mpg
1	car1	city	19
2	car2	city	20
3	car3	city	29
4	car1	hwy	24
5	car2	hwy	30
6	car3	hwy	35

Example: Untidy Data

ID | Income2023 | Income2024

A1 | 45000 | 47000

A2 | 52000 | 54000

Tidying the Example

ID | Year | Income
A1 | 2023 | 45000
A1 | 2024 | 47000
A2 | 2023 | 52000
A2 | 2024 | 54000

Steps to Clean Data

1. Standardize Column Names
2. Fix Data Types
3. Handle Missing Values
4. Remove Duplicates
5. Split or Combine Columns
6. Reshape Data (wide ↔ long)

Handling Missing Data

Methods:

- Delete rows/columns (only if sparse)
- Impute (mean, median, forward fill)
- Always document your approach

Data Consistency

- Normalize values:
  - “NY” vs “New York”
  - “Yes” vs “Y” vs “TRUE”
- Use controlled vocabularies when possible

Tools for Cleaning Data

- Excel/Google Sheets
- Python (Pandas): dropna(), fillna(), melt()
- R (tidyr/dplyr)
- OpenRefine
- Power BI Power Query

Cleaning Example in Google Sheets

Demo using:

- SPLIT(), FILTER(), QUERY()
- Before/after dataset cleanup using built-in tools

Google Sheets: SPLIT() Function

=SPLIT(text, delimiter)

- Used to break a text string into separate values based on a delimiter.
- Example:

  =SPLIT("John,Doe,Marketing", ",")

  → Returns: John | Doe | Marketing

Google Sheets: FILTER() Function

=FILTER(range, condition1, [condition2], ...)

- Returns rows in a range that meet specified conditions.
- Example:

  =FILTER(A2:B10, B2:B10 > 5000)

  → Returns rows where column B values are greater than 5000

Google Sheets: QUERY() Function

=QUERY(data, query, [headers])

- Uses SQL-like syntax to analyze and manipulate datasets.
- Example:

  =QUERY(A1:C10, "SELECT A, B WHERE C > 1000", 1)

  → Selects columns A and B where column C is greater than 1000

Summary

- Clean data enables better analysis
- Know your data types before cleaning
- Apply tidy data principles:
    - Variables = columns
    - Observations = rows
- Use tools effectively and document changes

Exploratory Data Analysis (EDA)

An Introduction for Beginners

What is Exploratory Data Analysis?

- EDA is the process of examining data sets to summarize their main characteristics.
- Often involves visual methods.
- Helps uncover patterns, spot anomalies, and test assumptions.

Why EDA is Important

- Understand the structure of your data.
- Identify missing or incorrect data.
- Choose the right tools and models.
- Develop intuition about data behavior.

Steps in EDA

1. Understand the context of the data.
2. Load and inspect the dataset.
3. Clean the data (remove nulls, fix formats).
4. Summarize statistics (mean, median, etc.).
5. Visualize distributions and relationships.

Common Techniques

- Descriptive statistics
- Data visualization
- Correlation analysis
- Outlier detection
- Data transformations (Tidy Data)

Descriptive Statistics

- Mean, Median, Mode
- Min and Max
- Standard Deviation

Data Visualization Tools

- Histograms: distribution of a variable
- Box plots: detect outliers
- Scatter plots: relationships between variables
- Bar charts: categorical comparisons

Example: Sales Dataset

- Count missing values in revenue column.
- Use histogram to view revenue distribution.
- Use scatter plot to see relationship between price and quantity sold.

Tools for EDA

- Excel / Google Sheets
- Python (Pandas, Matplotlib, Seaborn)
- R (ggplot2, dplyr)
- Power BI / Tableau

Tips for Beginners

- Ask questions about the data.
- Visualize before modeling.
- Clean data early.
- Document your observations.
- Keep it simple at first.

Conclusion

- EDA is the foundation of all data analysis.
- Helps make informed decisions.
- Start small and build your skills with practice.