

Data Analysis Process: Asking Questions → Data Wrangling → EDA → Drawing Conclusions

Source of data

- CSV
- Excel
- Text
- JSON
- Database
- API
- Web Scraping

- Gathering Data
- Assessing Data
- Cleaning Data

↓
present Findings
or
Results

Types of Unclean Data

- Quality Issues
- Structural Issues

- Completeness
- Validity
- Accuracy
- Consistency

↑
Increasing
severity

* Write a summary of the dataset

* Write column descriptions

* Any additional information

Types of Assessment

- Manual
- Programmatic

(Discover & Document)

↳ Iterative Process

Data Cleaning Order



Define → Code → Test

* Always create a copy b4
cleaning.

- Quality (Completeness)
- Structural
- Quality (Validity)
- Quality (Accuracy)
- Quality (Consistency)

Numerical Categorical

* The type of data will dictate what type of analysis can be performed upon it.

* Univariate Analysis - Focuses on analyzing each feature independently.

* Bi-variate Analysis

* Multi-variate Analysis

* There is no set method; highly iterative & intertwined process.