

The Data Cleaning Process

Data Cleaning

- Also called data cleansing
- Steps to prepare data for analysis
- Often, spend more time on cleaning than analysis

Examples of Data Cleaning

- Missing data
 - Data entry mistake?
 - Data doesn't exist?
 - May need to remove or replace with 0 or an average
- Duplicate data
 - Data entry mistake?
 - Data updates?
 - Need to remove duplicate rows
- Inconsistent data
 - Multiple data types in the same column
 - Different formats for column labels
- Outliers
 - Extreme values
 - Determine if they need to be included or removed

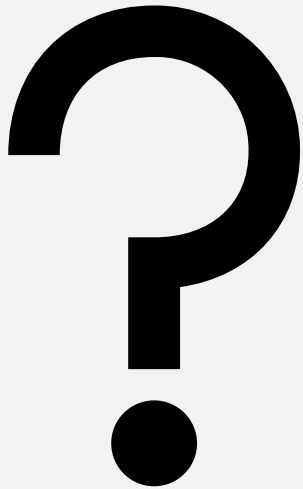
Examples of Data Cleaning

Missing Data

Duplicate Data

Inconsistent Data

Outliers



- Data entry mistake?
- Data doesn't exist?
- May need to remove or replace with 0 or an average

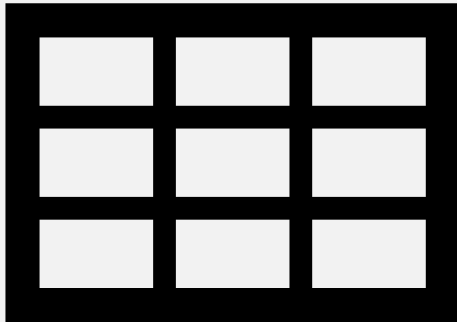
Examples of Data Cleaning

Missing Data

Duplicate Data

Inconsistent Data

Outliers



- Data entry mistake?
- Data updates?
- Need to remove duplicate rows

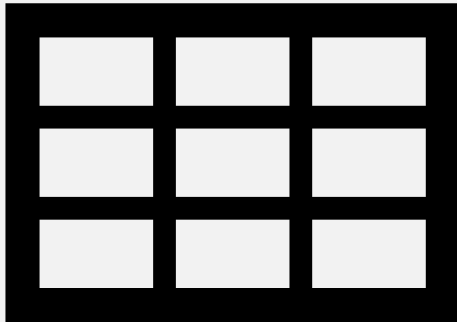
Examples of Data Cleaning

Missing Data

Duplicate Data

Inconsistent Data

Outliers



- Multiple data types in the same column
- Different formats for column labels

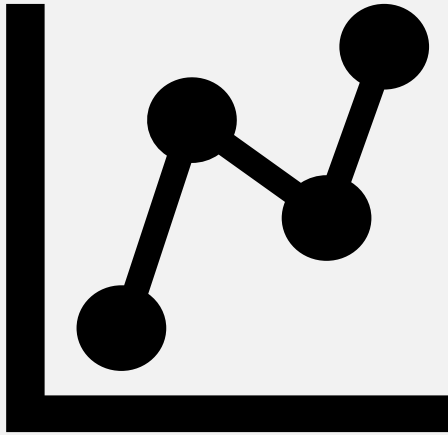
Examples of Data Cleaning

Missing Data

Duplicate Data

Inconsistent Data

Outliers



- Extreme values
- Determine if they need to be included or removed