Christopher Pierson

Lab Prep 5

BAIS:3250

Data Cleaning "Cheat Sheet"

- Initial Inspection
  - Load the dataset and preview the first few rows using .head()
  - Check data types using .info()
  - Check dataset shape using .shape()
  - Generate descriptive statistics using .describe()
- Columns
  - Understand variables and their meanings
  - Rename unclear column names
- Standardization & Formatting
  - Ensure consistent date formats
  - Ensure consistent units of measurement
  - Convert columns to proper data types
  - Reduce irrelevant data
- Handling Inconsistent Data
  - Identify & correct misspellings
- Duplicates
  - Detect exact duplicates
  - Remove duplicates
  - Investigate near-duplicates
- Missing Data
  - Identify missing values
  - Investigate why data is missing
  - Decide how to handle (delete rows, fill with median or mean, leave alone)
- Outliers
  - Detect with statistical methods
  - Decide how to handle
- Categorical Data
  - Standardize categories and encode for analysis if needed
- Final Check
  - Check final time and save new dataframe