The EDA Playbook: Part 1 Cheatsheet

Your essential guide to loading and inspecting your first datasets with Pandas.

Reading Data Files

This is your first step in any project—getting the data into a DataFrame.

- **Read a CSV File:**
  - Loads data from a comma-separated values file.
  - `df = pd.read_csv('your_file.csv')`
- **Read a Messy CSV File:**
  - Use parameters to handle custom separators, headers, and decimals.
  - `df = pd.read_csv('messy_file.csv', sep='|', header=None, decimal=',')`
- **Read an Excel File:**
  - Loads data from an `.xlsx` file. By default, it reads the first sheet.
  - `df = pd.read_excel('your_file.xlsx')`
- **Read a Specific Excel Sheet:**
  - Use the `sheet_name` parameter to target a specific tab in the workbook.
  - `df = pd.read_excel('your_file.xlsx', sheet_name='Sheet2')`

Inspecting Your DataFrame

Once the data is loaded, get a high-level overview.

- **Technical Summary (`.info()`):**
  - Provides crucial info: total rows, total columns, column data types, and the count of non-missing values.
  - `df.info()`
- **View First Rows (`.head()`):**
  - Shows the first 5 rows of your DataFrame. Good for a quick look.
  - `print(df.head())`
- **View Random Rows (`.sample()`):**
  - Shows a random sample of rows. Better for understanding data diversity than `.head()`.
  - `print(df.sample(5))`

Summarizing Your Data

Get a quick statistical and categorical summary.

- **Statistical Summary (`.describe()`):**
  - Provides a summary for **numeric columns**: count, mean, standard deviation, min, max, and quartiles.
  - `print(df.describe())`
- **Categorical Summary (`.describe(include='object')`):**
  - Provides a summary for **non-numeric (text) columns**: count, number of unique values, the top (most frequent) value, and its frequency.
  - `print(df.describe(include='object'))`
- **Full Summary (`.describe(include='all')`):**
  - Shows a combined summary for all columns, filling in non-applicable fields with `NaN`.
  - `print(df.describe(include='all'))`