Introduction to Importing Data in Python

Context manager (with statement): A Python construct (with ...) that opens a resource like a file or database connection for the duration of a code block and automatically closes or releases it afterwards

CSV (Comma-Separated Values): A common flat-file format where each field in a row is separated by a comma and rows are separated by newlines

DataFrame: A two-dimensional, labeled data structure in pandas with rows (observations) and columns (variables) that can hold different data types per column

Delimiter: The character or sequence (for example a comma or tab) used to separate fields within each row of a flat file

Flat file: A basic plain-text file that stores tabular records (rows) with fields (columns) but without relationships between multiple tables or files

h5py: A Python library that interfaces with HDF5 files, exposing their hierarchical groups and datasets with a dictionary-like API for reading and writing arrays

HDF5: A binary file format and data model designed for storing and organizing large, complex, hierarchical numerical datasets efficiently

Function body: The indented block of code beneath the function header that performs the function's work and may include a return statement

FacetGrid: The Seaborn object returned by functions that support subplots (like relplot and catplot), which contains one or more AxesSubplot objects and manages row/column faceting and the overall figure

MATLAB .mat (scipy.io.loadmat): The .mat file is MATLAB's workspace file format containing variables

Missing value: A placeholder indicating absent or undefined data in a dataset (commonly represented as NaN in numeric arrays or NaT/None in pandas), which libraries provide tools to detect and handle

NumPy: A core Python library for numerical computing that provides the ndarray data structure and efficient operations on numeric arrays

numpy.genfromtxt: A NumPy function similar to loadtxt but more tolerant of missing data and mixed types, providing flexible parsing of text files into arrays

numpy.loadtxt: A NumPy function to load numerical data from a text file into a NumPy array, with options for delimiter, skipping rows, and selecting columns

pandas: A Python library for data manipulation and analysis that provides labeled, table-like data structures and many import/export utilities

pandas.read_csv: A pandas convenience function that reads a CSV (and other delimited) file directly into a DataFrame, handling headers, missing values, and parsing options

pandas.read_sql_query: A pandas function that executes an SQL query against a database engine or connection and directly returns the query result as a pandas DataFrame

Pickle / serialization: The process of converting a Python object into a binary byte stream (pickling) so it can be saved to disk and later restored (unpickled), typically opened with mode 'rb' or 'wb'

Plain text file: A file that contains human-readable characters and line breaks (for example .txt) and can be opened and read as plain text without special parsing

Primary key: A column or set of columns in a database table whose values uniquely identify each row, used to index and relate records across tables

Record: A single row in a table or flat file that represents one instance of data and contains a fixed set of fields or attributes

Relational database: A database organized into multiple tables of rows and columns where relationships between tables are expressed via keys, and data is queried using SQL

scipy.io.loadmat loads a .mat file into Python as a dictionary mapping variable names to arrays or objects

SQL (Structured Query Language): A standardized language for querying, filtering, aggregating, and updating data stored in relational databases using statements like SELECT and WHERE

SQLAlchemy (create_engine): SQLAlchemy is a Python SQL toolkit and ORM, and create_engine is its function to create a database engine object that manages connections to a specific database

Statistical software files (SAS .sas7bdat and Stata .dta): Native file formats from SAS and Stata that store datasets and can be imported into Python (often via specialized packages) as pandas DataFrames