Retrieving Data

You can retrieve data from multiple sources:

- SQL databases
- NoSQL databases
- APIs
- Cloud data sources

The two most common formats for delimited data flat files are comma separated (csv) and tab separated (tsv). It is also possible to use special characters as separators.

SQL represents a set of relational databases with fixed schemas.

Reading in Database Files

The steps to read in a database file using the sqlite library are:

- create a path variable that references the path to your database
- create a connection variable that references the connection to your database
- create a query variable that contains the SQL query that reads in the data table from your database
- create an observations variable to assign the read_sql functions from pandas package
- create a tables variable to read in the data from the table sqlite_master

JSON files are a standard way to store data across platforms. Their structure is similar to Python dictionaries.

NoSQL databases are not relational and vary more in structure. Most NoSQL databases store data in JSON format.

Data Cleaning

Data Cleaning is important because messy data will lead to unreliable outcomes.Some common issues that make data messy are: duplicate or unnecessary data, inconsistent data and typos, missing data, outliers, and data source issues.

You can identify duplicate or unnecessary dataCommon policies to deal with missing data are:remove a row with missing columns, impute the missing data, and mask the data by creating a category for missing values.

Common methods to find outliers are: through plots, statistics, or residuals.

Common policies to deal with outliers are: remove outliers, impute them, use a variable transformation, or use a model that is resistant to outliers.