Data analysis and visualization in Python

Outline

- Array manipulation with Numpy
- Data manipulation with Pandas
- Data visualization with Matplotlib/Seaborn

Introduction to Numpy

Numpy is the core library for scientific computing.

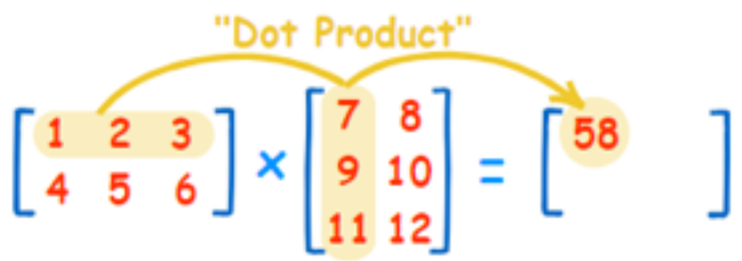
- Provide a high-performance multidimensional array object
- Create a array object: np.array()
- Modify a array
- Array indexing
- Create commonly used array
- Array slicing
- Boolean array indexing
- Tool for working with these arrays: datatype, statistical analysis
- Array math

Array:

- all of the same type
- have dimension: difference between a list and a bumpy array

Numpy: dot product

The column numbers of the first array is equal to the row numbers of second array:
for example below:
the first array: (2,3)
the second array: (3,2)

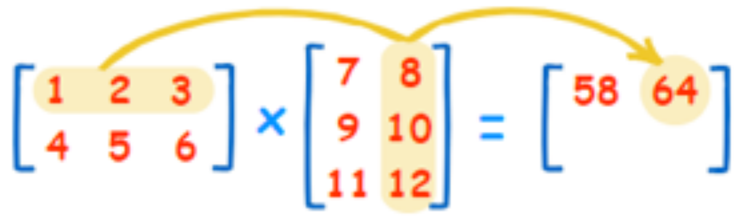


The "Dot Product" is where we **multiply matching members**, then sum up:

(1, 2, 3) • (7, 9, 11) = 1×7 + 2×9 + 3×11
= 58

We match the 1st members (1 and 7), multiply them, likewise for the 2nd members (2 and 9) and the 3rd members (3 and 11), and finally sum them up.

Want to see another example? Here it is for the 1st row and **2nd column**:



(1, 2, 3) • (8, 10, 12) = 1×8 + 2×10 + 3×12
= 64

Dot Product

$$\begin{bmatrix} a & b \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax+by \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} ax+by \\ cx+dy \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} w & x \\ y & z \end{bmatrix} = \begin{bmatrix} aw+by & ax+bz \\ cw+dy & cx+dz \end{bmatrix}$$

Introduction to Pandas

Pandas is a library that unifies the most common workflows that data analysts.

- Create data frame from dictionaries or arrays

- Data manipulation

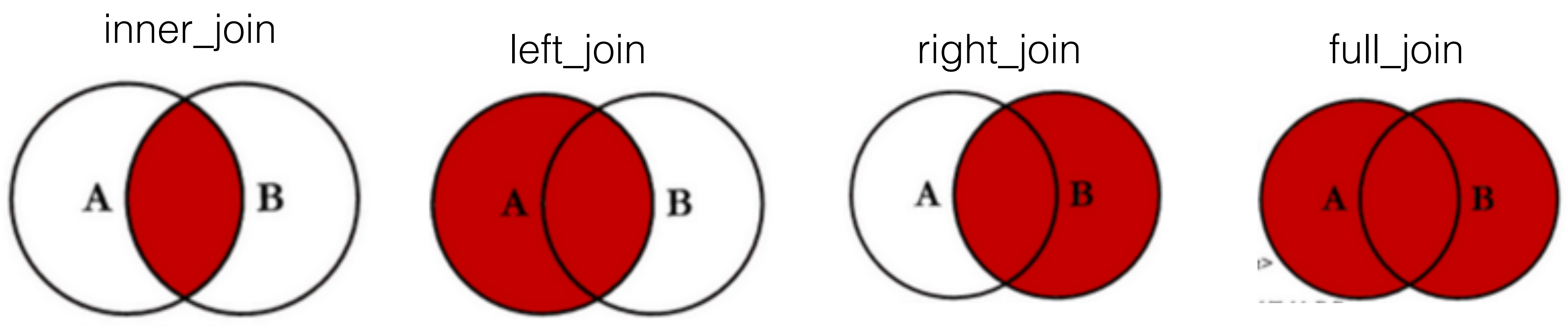
1. analysis, such as column names, dimension, data types, statistical summary
2. select columns/rows
3. subletting by conditional
4. data aggregation, such as sort, groupby and aggregation
5. change data type
6. add a new column
7. check missing values by column or row
8. combine data frames, such as concat and merge

Join data sets

Need information from 2 or more data frames

- Adding Columns: merge two data frames (datasets) horizontally
  1. inner
  2. left
  3. right
  4. outer

- Adding Rows: join two data frames (datasets) vertically



Introduction to Matplotib

Matplotib is a python 2D plotting library

- Line plot
- Scatterplot
- Histogram
- Barplot
- Piechart
- Boxplot

Matplotlib comes with a set of default settings that allow customizing all kinds of properties.
You can control the defaults of almost every property in matplotlib: figure size and dpi, line width, color and style, axes, axis and grid properties, text and font properties and so on.

Introduction to Seaborn

Seaborn provides an API on top of Matplotlib that offers same choices for plot style and color defaults, defines simple high-level functions for common statistical plot types, and integrates with the functionality provided by Pandas DataFrames.

- Histogram and distribution: plot histograms and joint distributions of variables
- Paris plot: generalize joint plots to datasets of larger dimensions and explore correlations between multidimensional data
- Faceted histograms: View data is via histograms of subsets
- Factor plots: view the distribution of a parameter within bins defined by any other parameter
- Bar plots
- Joint distributions: show the joint distribution between different datasets and do some automatic kernel density estimation and regression