Data analysis and visualization in Python

Outline

- Numpy
- Data Frame
- Data manipulation with Pandas
- Grouping and aggregation
- Matplotlib/Seaborn

Introduction to Numpy

Numpy is the core library for scientific computing.

- Provide a high-performance multidimensional array object
- Create a array object
- Array indexing and slicing
- Boolean array indexing
- Tool for working with these arrays
- Array math

Array:

- all of the same type
- have dimension

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 columns, shape, describe
2. subletting, such as filter
3. data aggregation, such as groupby and aggregation
4. missing values
5. combine data frames, such as concat and merge

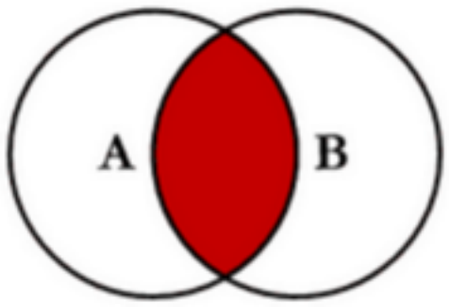
Join data sets

Need information from 2 or more data frames

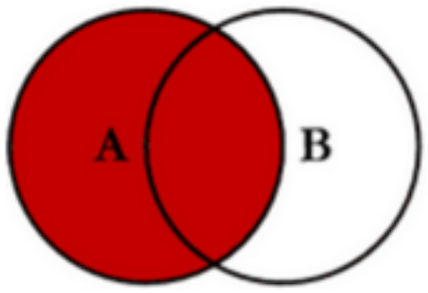
- Adding Columns: merge two data frames (datasets) horizontally
  1. inner_join
  2. left_join
  3. right_join
  4. full_join

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

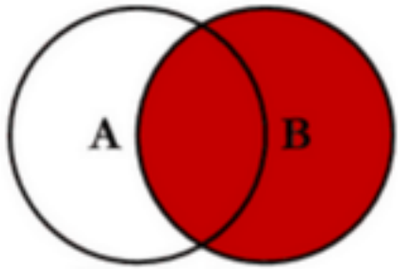
inner_join



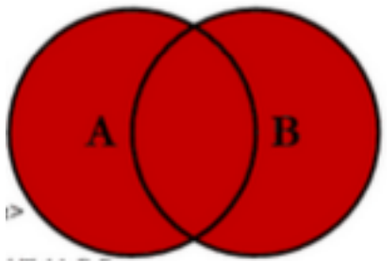
left_join



right_join



full_join



Introduction to Matplotib

Matplotib is a python 2D plotting library

- Basic 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 sane 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