W3.1

2020年3月23日 星期一 下午7:56

week3

Pandas Data Structures

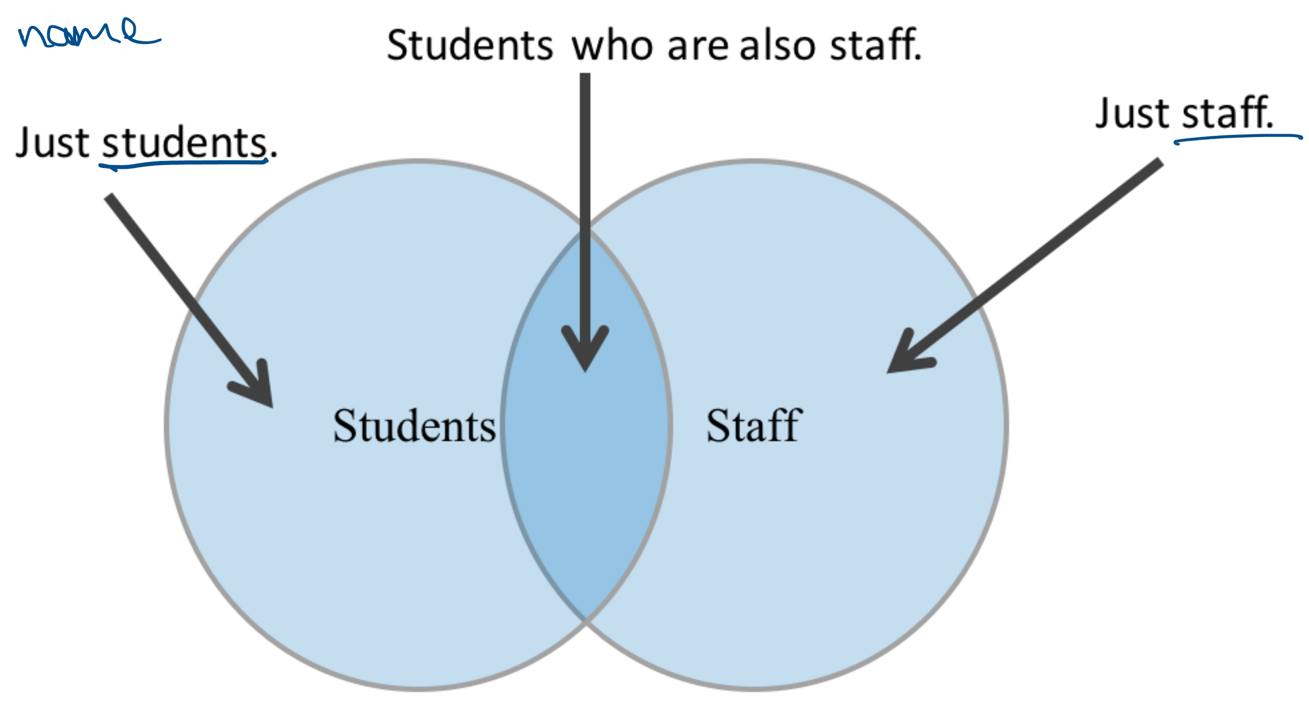
- Series Object (1 dimensional, a row)
- DataFrame Object (2 dimensional, a table)
- Querying
  - *iloc[], for querying based on position*
  - *loc[], for querying rows based on label*
  - *Querying the DataFrame directly*
    - *Projecting a subset of columns*
    - *Using a boolean mask to filter data*

Setting Data in Pandas

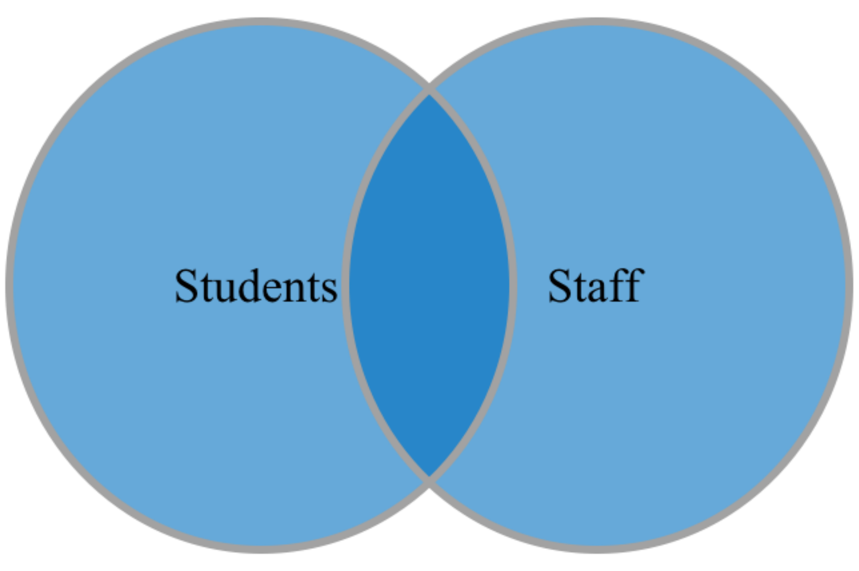
- To add new data
  - *df[column]=[a,b,c]*
- To set default data (or overwrite all data):
  - *df[column]=2*

> population – index
in separate DF
label: person name

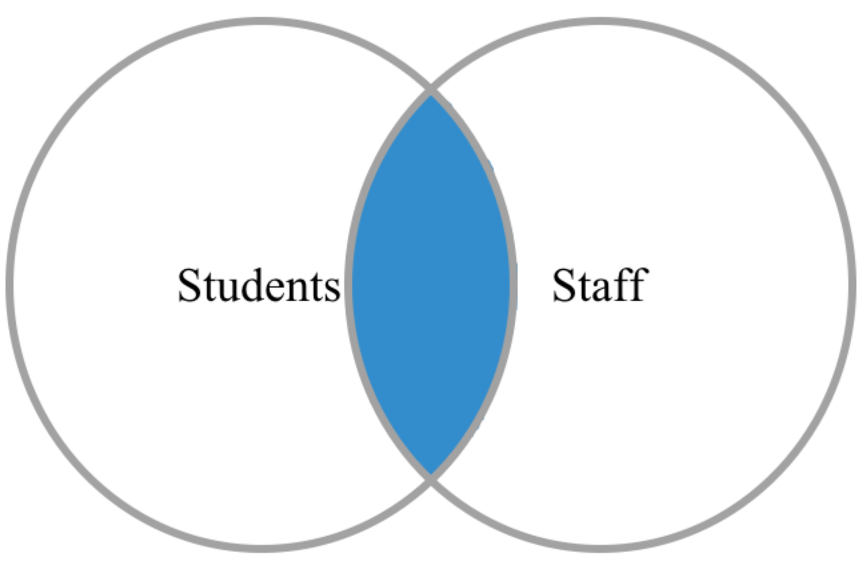
Venn Diagram



Full outer join (union)



Inner join (intersection)



- Chain Indexing:
  - *df.loc["Washtenaw"]["Total Population"]*
  - *Generally bad, pandas could return a copy of a view depending upon numpy*
- Code smell
  - *If you see a][you should think carefully about what you are doing (Tom Augspurger)*

④ (a,b) (c,d): Scales

DF: students ~~academic level~~ grade 1 2 3 difference

- **Ratio scale:** – difference × apparent – clear in mind
  - measurement *units are equally spaced.*
  - *mathematical operations of +-/* are all valid*
  - *E.g. height and weight*
- **Interval scale:**
  - *units are equally spaced, but there is no true zero* no clear absence of value, 0 degree meaningful (× /) × Ex: temperature, Fahrenheit direction on a campus – 0 degree ≠ lack
- **Ordinal scale:**
  - *the order of the units is important, but not evenly spaced.*
  - *Letter grades such as A+, A are a good example*
- **Nominal scale:** (categorical data)
  - *categories of data, but the categories have no order with respect to one another.*
  - *E.g. Teams of a sport.* × math function
  - Team names
  - only 2 possible value → binary

importance: statistics / ML.

pd. functions – convert between measurement scales