Pandas (2)

Akitaka Matsuo
Essex IADS

MERGING

Merging Motivation

- There are countless situation when you want to combine two or more datasets
- Example:
  - – Main dataset: Election results at constituency level
  - – We have other datasets at the same level
    - Demography
    - Previous election results
    - Economic conditions
  - – If we have a unique identifier, we can combine two datasets
    - Name? (not ideal, e.g. Newcastle-upon-Tyne)
    - ID code? (ideal)

Merging

- In Pandas, there are two methods for
  1. `df.merge()`: Combining two datasets based on some ID variables
     - Simpler (as we don't have to set index)
  2. `df.join()`: Combining two datasets based on index
     - Tedious but much faster once we set the index
- We will use mainly 1. in the demo

DESCRIPTIVE STATISTICS AND GROUPING

DataFrame, descriptive statistics

- Pandas offers a number of methods for descriptive statistics (mostly) for columns
- Example:
  - `min(), max(), sum(), mean(), std()`
  - `describe()`
- You can get the group summary by the method described below

DataFrame, tabulation

- Instead of get a summary statistics, you may want to get the frequency of values for a variable or two
- `value_counts()` provides the method for that

DataFrame, correlation

- Correlation (or Pearson correlation coefficient):
  - "a measure of the strength of the association between the two variables."
  - It is a simple way to check the relations between two variables
  - Domain: [-1, 1]
    - 1: perfectly positive linear relationship
    - -1: perfectly negative linear relationship
  - c.f.: http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442
- Pandas method: `df.corr()` (after selecting variables to use)

Group Summary

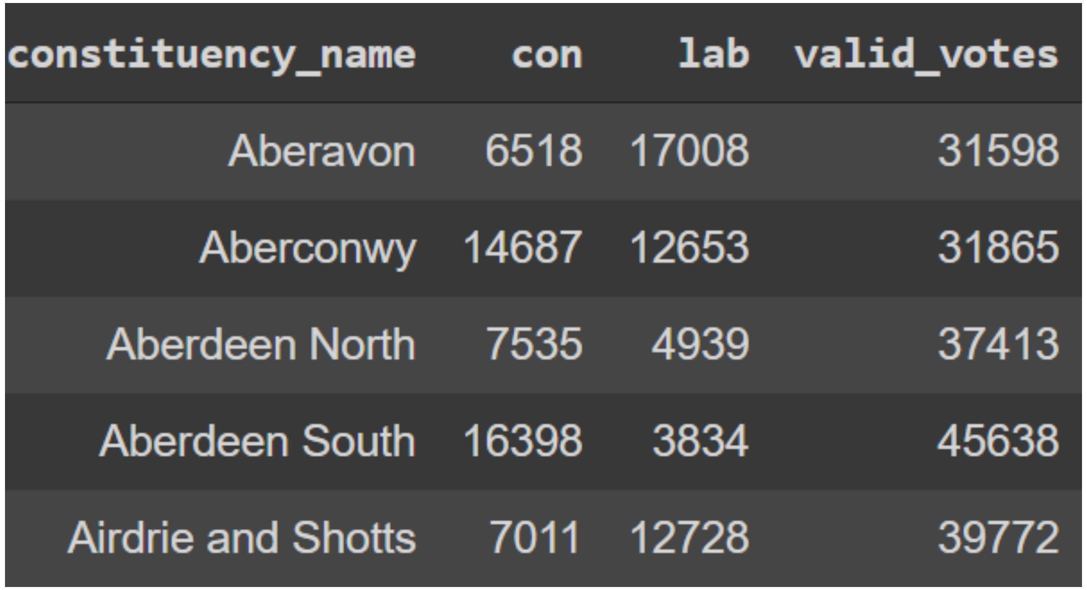
- You may want to get the summary statistics by group (e.g. group mean)
- There are several ways to do that in Pandas
  1. Use index (set index and use index option in applying a method)
  2. Group the data with `groupby()`, then apply methods
- Results are the same

PIVOTING

“Wide” and “Long” DataFrame

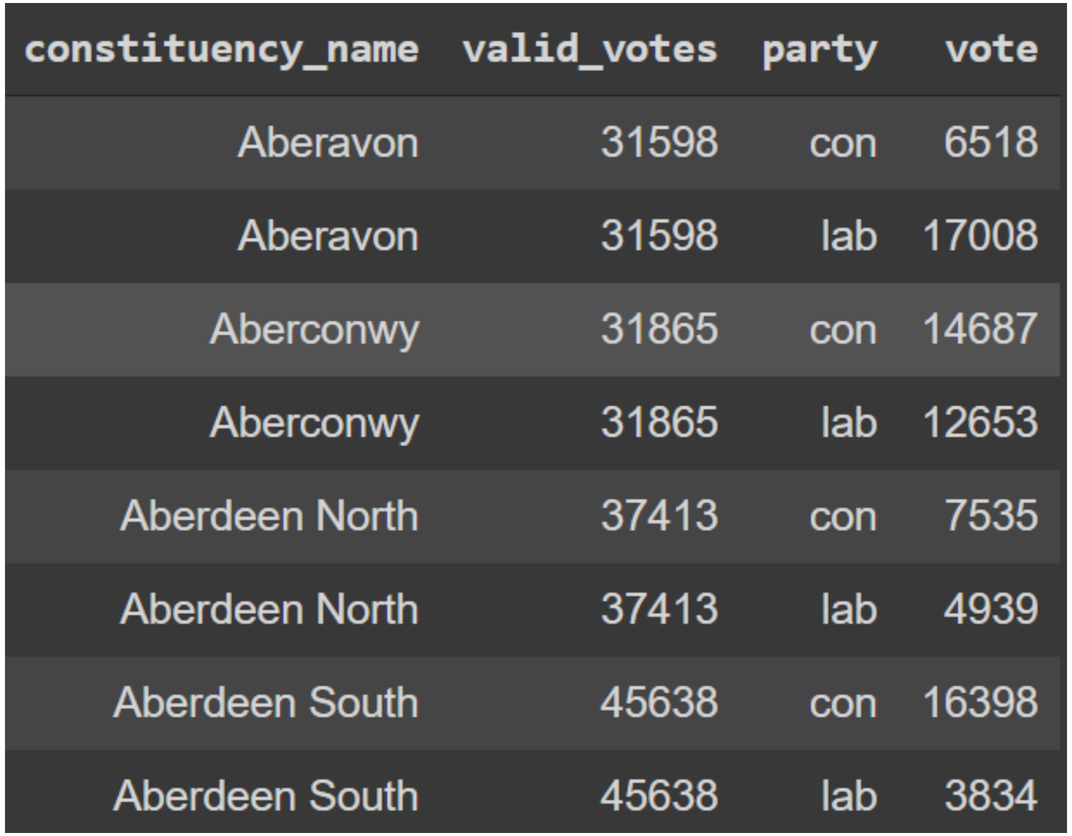
University of Essex

Wide Format



constituency_name	con	lab	valid_votes
Aberavon	6518	17008	31598
Aberconwy	14687	12653	31865
Aberdeen North	7535	4939	37413
Aberdeen South	16398	3834	45638
Airdrie and Shotts	7011	12728	39772

Long Format



constituency_name	valid_votes	party	vote
Aberavon	31598	con	6518
Aberavon	31598	lab	17008
Aberconwy	31865	con	14687
Aberconwy	31865	lab	12653
Aberdeen North	37413	con	7535
Aberdeen North	37413	lab	4939
Aberdeen South	45638	con	16398
Aberdeen South	45638	lab	3834

“Wide” and “Long” DataFrame

- “Wide”: There are several variables of the same masure for different units
  - e.g. Unemployment data. Multiple columns for unemployment rates in different months
- You many want to convert between these formats
  - In pandas, you can achieve that with
    - `melt()`: “Wide” to “Long”
    - `pivot()`: “Long” to “Wide”