University of Essex

Pandas: DataFrame Descriptive Statistics and Grouping

Akitaka Matsuo
Essex IADS

DataFrame, descriptive statistics

University of Essex

- 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

University of Essex

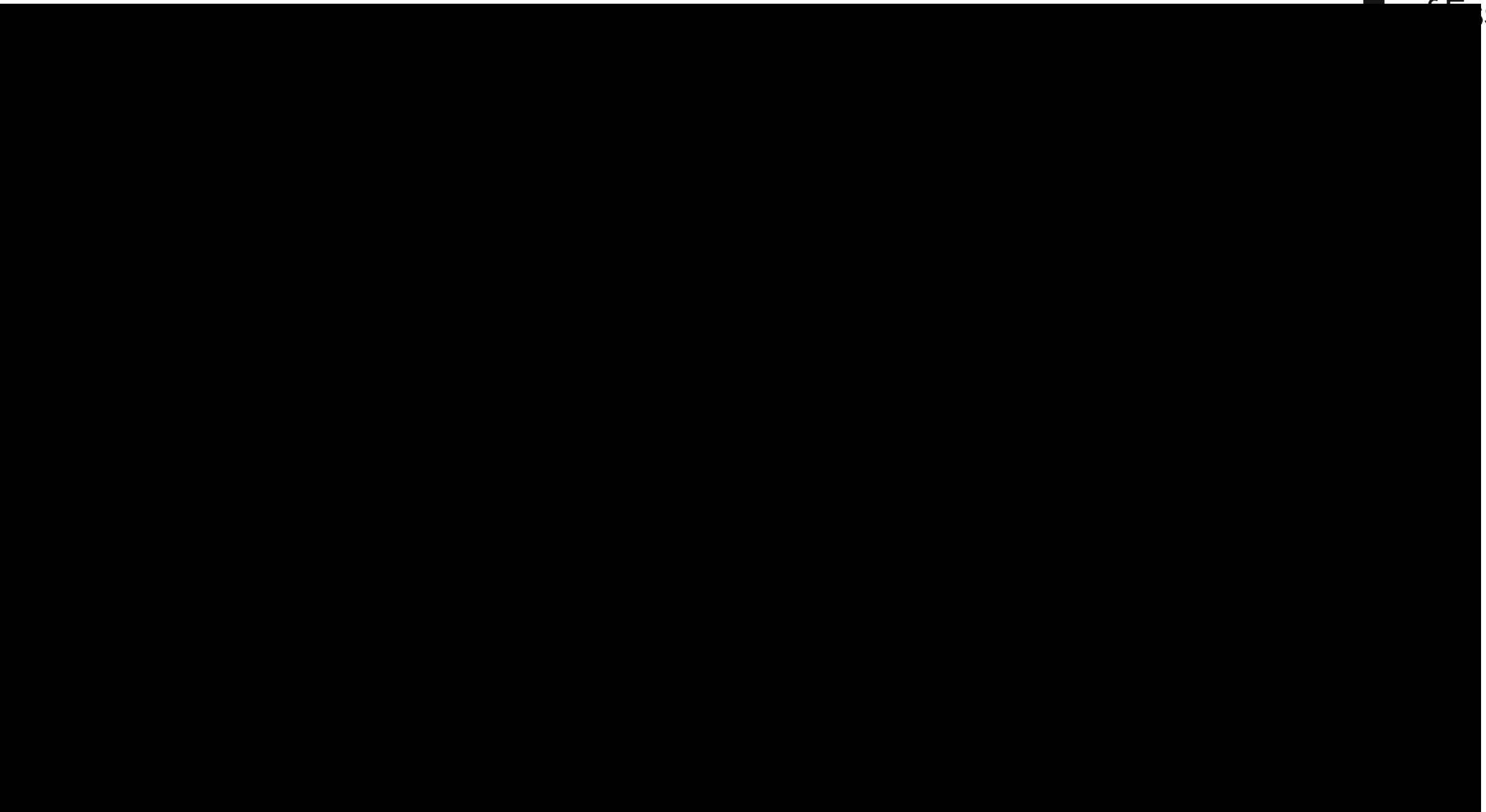
- Instead of get a summary statistics, you may want to get the frequency of values for a variable or two
- `value_counts()` provides the method for that

DataFrame, correlation

University of Essex

- 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)

Demo

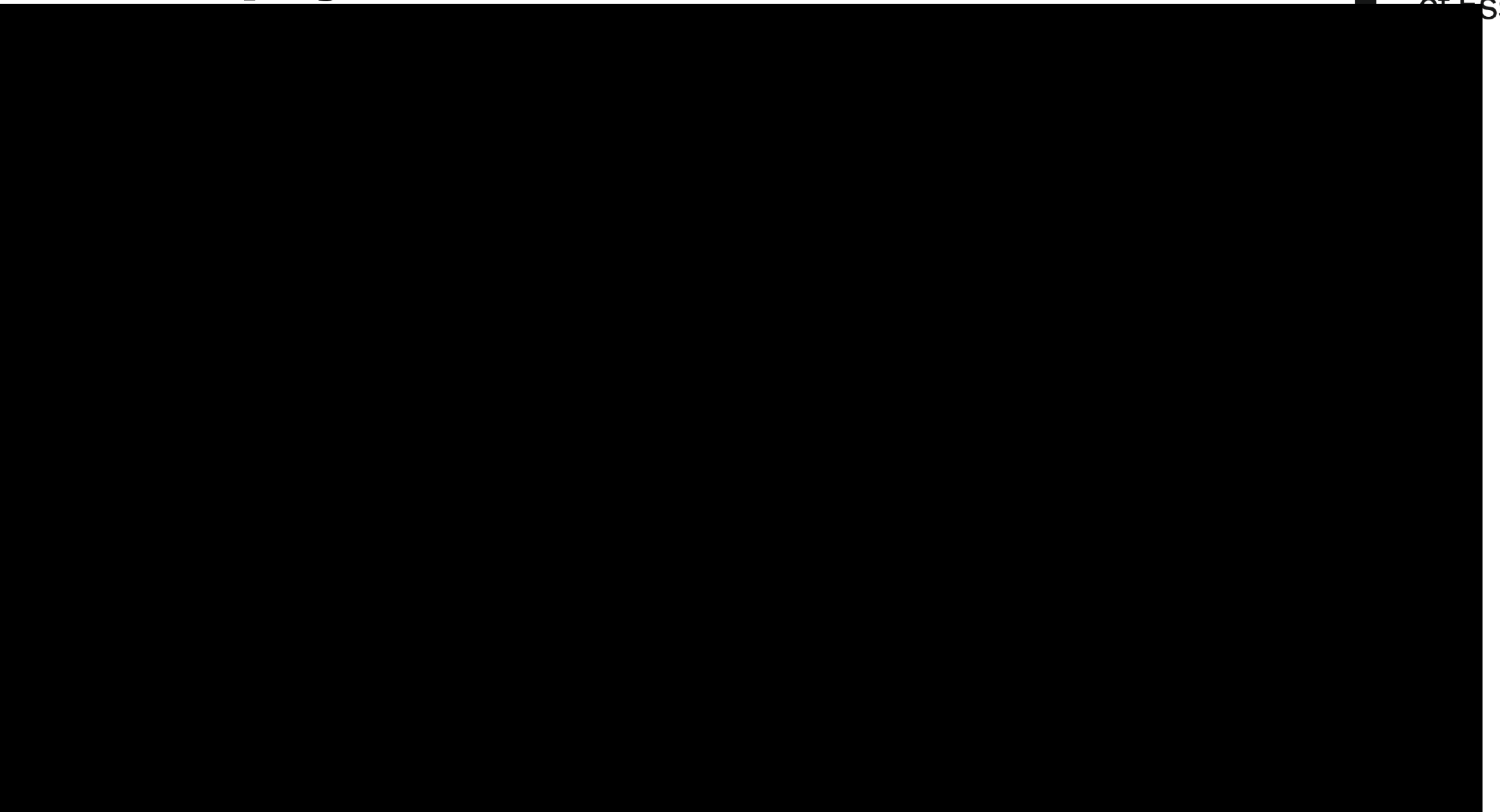


Group Summary

University of Essex

- 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

Groupby Demo



University of Essex