Python for Data Analysis, 3rd edition

Data Wrangling with pandas, NumPy, and Jupyter

Stephen CUI[1]

January 4, 2022

[1]cuixuanStephen@gmail.com

Contents

Chapter 1

Getting Started with pandas

1.1 Essential Functionality

Arithmetic and Data Alignment

The internal data alignment introduces missing values in the label locations that don't overlap. Missing values will then propagate in further arithmetic computations.

Arithmetic methods with fill values

Using the add method on df1, I pass df2 and an argument to `fill_value`, which substitutes the passed value for any missing values in the operation.

Operations between DataFrame and Series

By default, arithmetic between DataFrame and Series matches the index of the Series on the columns of the DataFrame, broadcasting down the rows

If you want to instead broadcast over the columns, matching on the rows, you have to use one of the arithmetic methods and specify to match over the index.

Table 1.1: Flexible arithmetic methods

Method	Description
add, radd	Methods for addition (+)
sub, rsub	Methods for subtraction (-)
div, rdiv	Methods for division (/)
floordiv, rfloordiv	Methods for floor division (//)
mul, rmul	Methods for multiplication (*)
pow, rpow	Methods for exponentiation (**)

Table 1.2: Tie-breaking methods with rank

Method	Description
“average”	Default: assign the average rank to each entry in the equal group
“min”	Use the minimum rank for the whole group
“max”	Use the maximum rank for the whole group
“first”	Assign ranks in the order the values appear in the data
“dense”	Like method=“min”, but ranks always increase by 1 between groups rather than the number of equal elements in a group

Function Application and Mapping

frequent operation is applying a function on one-dimensional arrays to each column or row. DataFrame’s apply method does exactly this.

Element-wise Python functions can be used, too. You can do this with applymap. The reason for the name applymap is that Series has a map method for applying an element-wise function.

1.1.1 Sorting and Ranking

Ranking assigns ranks from one through the number of valid data points in an array, starting from the lowest value. By default, rank breaks ties by assigning each group the mean rank.

See Table 1.2 for a list of tie-breaking methods available.

1.1.2 Axis Indexes with Duplicate Labels

Chapter 2

Data Loading, Storage, and File Formats

Chapter 3

Data Cleaning and Preparation

Chapter 4

Appendix A