Data Wrangling

Data Wrangling

Data Wrangling Operations Can Be Grouped In To 3 Important Parts

- Combine
- Join
- Rearrange

Data Wrangling

Hierarchical Indexing

- It allows To have multiple (two or more) index levels on an axis
- It allows us to work with higher dimensional data in a dimensional form
- It is also referred as MultiIndex level
- Each MultiIndexed Object can be partially indexed with its label name
- It plays an important role in reshaping and group-based operations on data like pivote table
- We can use ‘stack’ and ‘unstack’ to group & ungroup the indexes respectively

Data Wrangling

Hierarchical Indexing

Reordering and Sorting Levels

swaplevel

- It returns the new object with the index levels Interchanged.
- Data values are unaltered.

sort_index

- It sorts the data using only the values in a single level.
- The object is lexicographically sorted by the indicated level.

Index		**Columns**	
Index_1	Index_2	R	Y
a	1		
	2		
b	1		
	2		
c	1		

Some values

Data Wrangling

Hierarchical Indexing

Summary Statistics by Level

‘sum’ method

With the help of ‘sum’ method we can get the summary and descriptive statistics by index levels on either rows or columns.

Index		Columns	
Index_1	Index_2	R	Y
a	1		
	2		
b	1		
	2		
c	1		

Some values

Data Wrangling

Hierarchical Indexing

Indexing with a DataFrame’s columns

‘set_index’ and ‘reset_index’ methods

- ‘set_index’ method will create a new DataFrame object with the existing values by using its columns as the new index objects

- ‘reset_index’ method will do the opposite work of set_index

Index		Columns	
Index_1	Index_2	R	Y
a	1		
	2		
b	1		
	2		
c	1		

Some values

Data Wrangling

Hierarchical Indexing

Indexing with a DataFrame's columns

`‘set_index’  and ‘reset_index’ methods`

- ‘set_index’ method will create a new DataFrame object with the existing values by using its columns as the new index objects

- ‘reset_index’ method will do the opposite work of set_index

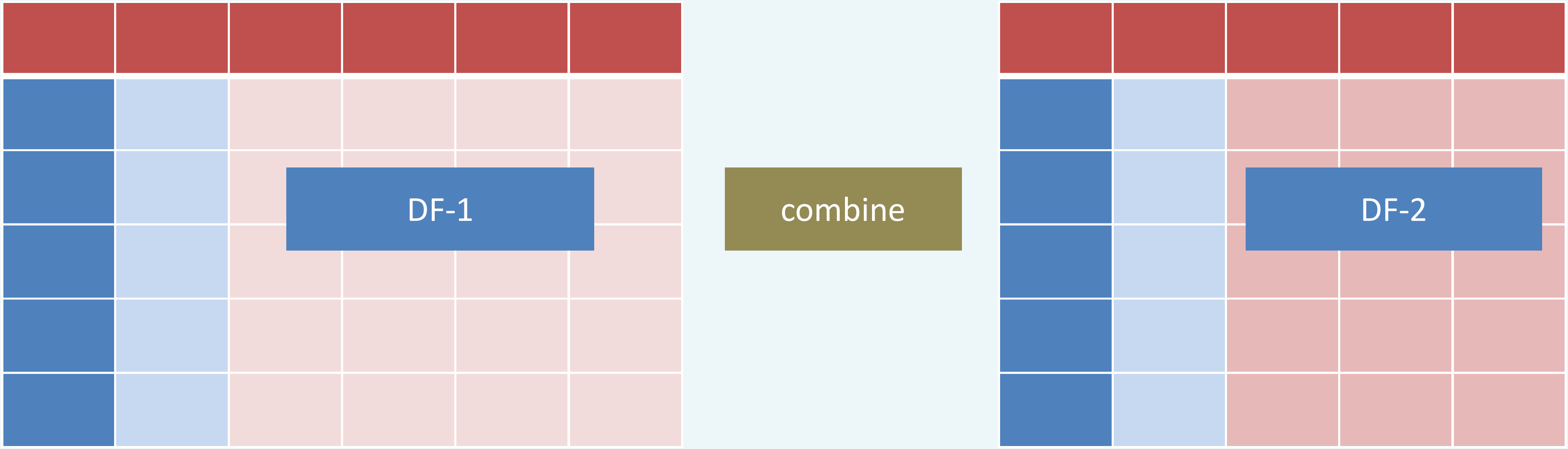
Index		Columns	
Index_1	Index_2	R	Y
a	1		
	2		
b	1		
	2		
c	1		

Some values

Data Wrangling

Combining and Merging Datasets

Pandas Provides Many Ways To Combine Data



DF-1

combine

DF-2

Data Wrangling

Combining and Merging Datasets

Pandas Provides Many Ways To Combine Data

‘merge’ method:

- ‘pandas.merge’ connects rows in DataFrames based on one or more index keys

‘concat’ method:

- ‘pandas.concat’ concatenates or “stacks” together objects along an axis.

‘combine_first’ method:

- This method enables splicing together overlapping data to fill in missing values in one object with values from another

Data Wrangling

Combining and Merging Datasets

Database-Style DataFrame Joins

‘merge’ method

- Merge or join operations combine datasets by linking rows using one or more keys.
- Merge uses the overlapping column names as the keys.
- By default merge operation do an 'inner' join; to do outer join we can specify `how=‘outer’.`
- The outer join takes the union of the keys, combining the effect of applying both left and right joins.
- To merge with multiple keys, we can pass a list of column names to the merge option. These columns act as index keys for merge operation.

Data Wrangling

Combining and Merging Datasets

Database-Style DataFrame Joins

`‘merge’ method`

- Merge has a suffixes option for specifying strings to append to overlapping names in the left and right DataFrame objects.

- Merge method has many other arguments to combine and return new data object.

All these merge operations are broadly grouped in to two main categories:

- Many – to – one merge method
- Many – to – many merge method

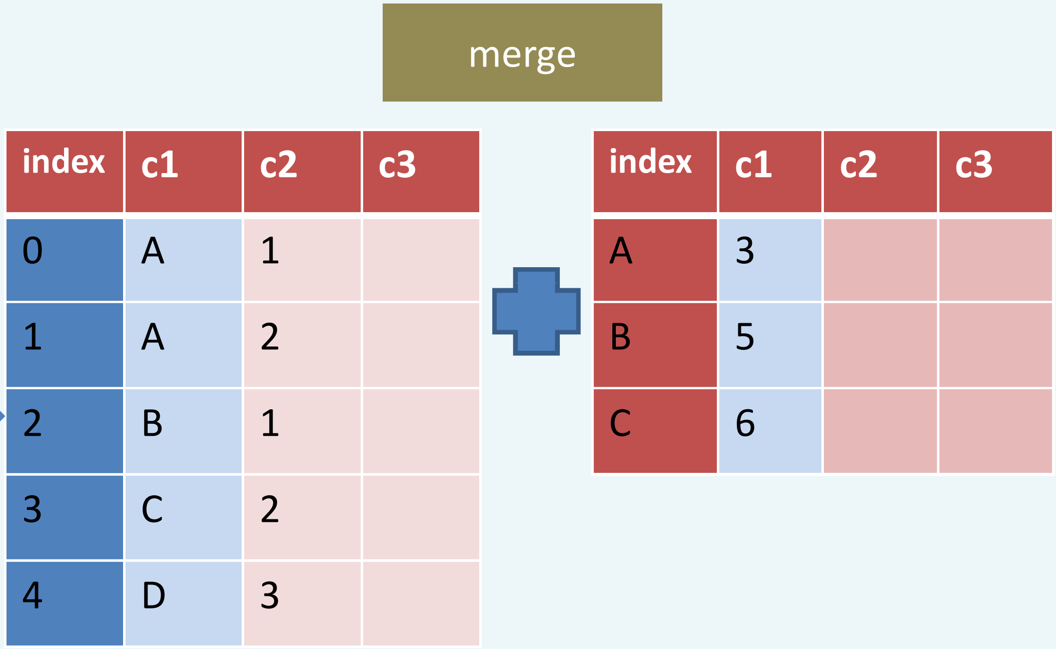
Data Wrangling

Combining and Merging Datasets

Merging on Index

When the merge index of DataFrame found in its index still we can merge.

We can pass left_index=True or right_index=True (or both) to merge method.



merge

index	c1	c2	c3
0	A	1	
1	A	2	
2	B	1	
3	C	2	
4	D	3	

index	c1	c2	c3
A	3		
B	5		
C	6		

Data Wrangling

Combining and Merging Datasets

When the merge index of DataFrame found in its index still we can merge.

We can also use ‘join’ method instead of merge or merging by index

It can also be used to combine together many DataFrame objects having the same or similar indexes but non-overlapping columns

Merging on Index

join

index	c1	c2	c3
0	A	1	
1	A	2	
2	B	1	
3	C	2	
4	D	3	

+

index	c1	c2	c3
A	3		
B	5		
C	6		

Data Wrangling

Combining and Merging Datasets

Concatenating Along an Axis

pandas.concat

Things to know about pandas.concat method

- If the objects are indexed differently on the other axes, should we combine the distinct elements in these axes or use only the shared values (the intersection)?
- Do the concatenated chunks of data need to be identifiable in the resulting object?
- Does the “concatenation axis” contain data that needs to be preserved?

Data Wrangling

Combining and Merging Datasets

Concatenating Along an Axis

`pandas.concat`

Things to know about pandas.concat method

- By default concat works along axis=0, producing another Series.
- If you pass axis=1, the result will instead be a DataFrame (axis=1 is the columns).

Data Wrangling

Combining and Merging Datasets

Combining Data with Overlap

`pandas.combine_first`

- DataFrame’s, ‘combine_first’ do the same thing column by column as the Numpy’s ‘where’ function does.
- It also allows panda's usual data alignment logic to align the data based on index alignment.

Data Wrangling

Combining and Merging Datasets

Reshaping and Pivoting

Rearranging the tabular data is referred as reshaping or pivot operation.

Data Wrangling

Combining and Merging Datasets

Reshaping and Pivoting

Reshaping with Hierarchical Indexing

`stack`

- This “rotates” or pivots from the columns in the data to the rows

`unstack`

- This pivots from the rows into the columns

Data Wrangling

Combining and Merging Datasets

Reshaping and Pivoting

Reshaping and Pivoting Methods

Long To wide; Long or stacked format: it uses ‘`pandas.pivot`’ method

- A general way to store multiple time series in databases like MySQL and CSV.

Wide To long; An inverse operation to pivot for DataFrames. It uses ’`pandas.melt`’ method

- It merges multiple columns into one, producing a DataFrame that is longer than the input.