YData: Introduction to Data Science

Lecture 11: Joins

Overview

Grouping continued

Pivot Tables

Joining tables

Grouping

Grouping by one column

The tb.group() method aggregates all rows with the same value for a column into a single row in the resulting table.

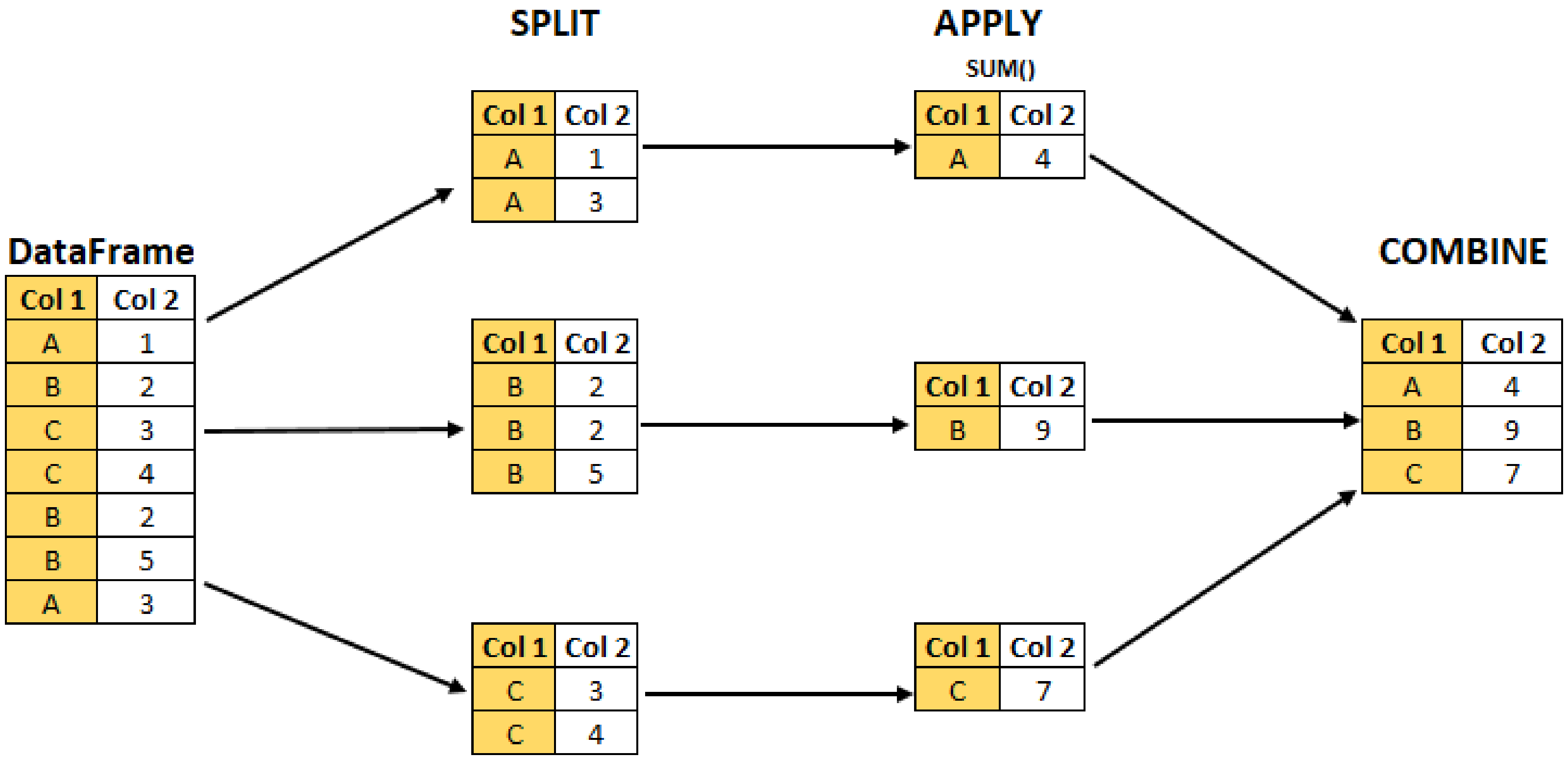
tb.group("grouping col", agg_function)

- "grouping col": column to group data by
- agg_function: function on how data in each group should be combined

Examples of aggregating functions:

- len: number of items in each group (default if no second argument is specified)
- list: list of all values in each group
- sum: total of all grouped values

Grouping: split-apply-combine



DataFrame

Col 1	Col 2
A	1
B	2
C	3
C	4
B	2
B	5
A	3

SPLIT

Col 1	Col 2
A	1
A	3

Col 1	Col 2
B	2
B	2
B	5

Col 1	Col 2
C	3
C	4

APPLY

SUM()

Col 1	Col 2
A	4

Col 1	Col 2
B	9

Col 1	Col 2
C	7

COMBINE

Col 1	Col 2
A	4
B	9
C	7

```
tb.group("Col 1", sum)
```

Grouping by multiple columns

The tb.group() method can also aggregate values in rows that share the combination of values in multiple columns

tb.group(["grouping col1", "grouping col2"], agg_function)

- ["grouping col1", "grouping col2"]: list of columns to group by
- agg_function: function on how data in each group should be combined

Let's explore this in Jupyter!

Pivot Tables

Pivot Tables

Pivot tables aggregate values according to two categorical variables but the results are in a table

- i.e., same as grouping by two categorical variables but puts one variable as the rows and the other as columns

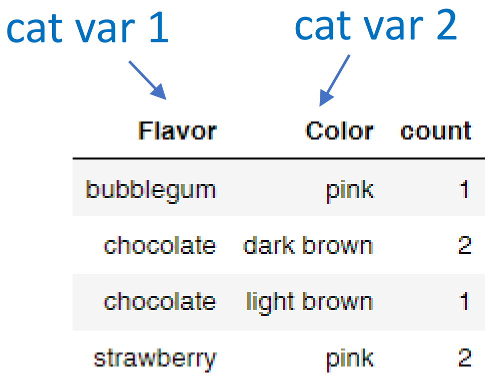
Produces a grid of counts or aggregated values two required arguments:

- First: variable that forms column labels of grid
- Second: variable that forms row labels of grid

Two optional arguments (include both or neither)

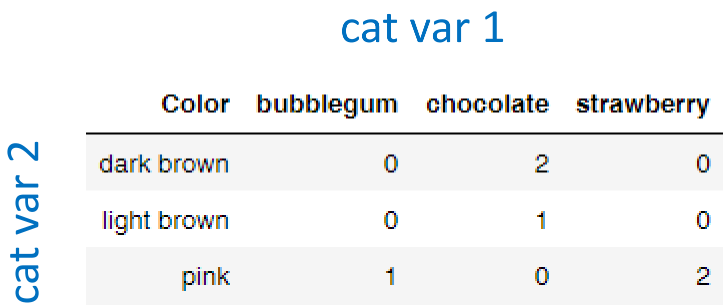
- values = `column label to aggregate'
- collect = function with which to aggregate

Grouping


cat var 1
cat var 2

Flavor	Color	count
bubblegum	pink	1
chocolate	dark brown	2
chocolate	light brown	1
strawberry	pink	2

Pivot Table


cat var 1
cat var 2

Color	bubblegum	chocolate	strawberry
dark brown	0	2	0
light brown	0	1	0
pink	1	0	2

Let's explore this in Jupyter!

Joins

Joining Two Tables

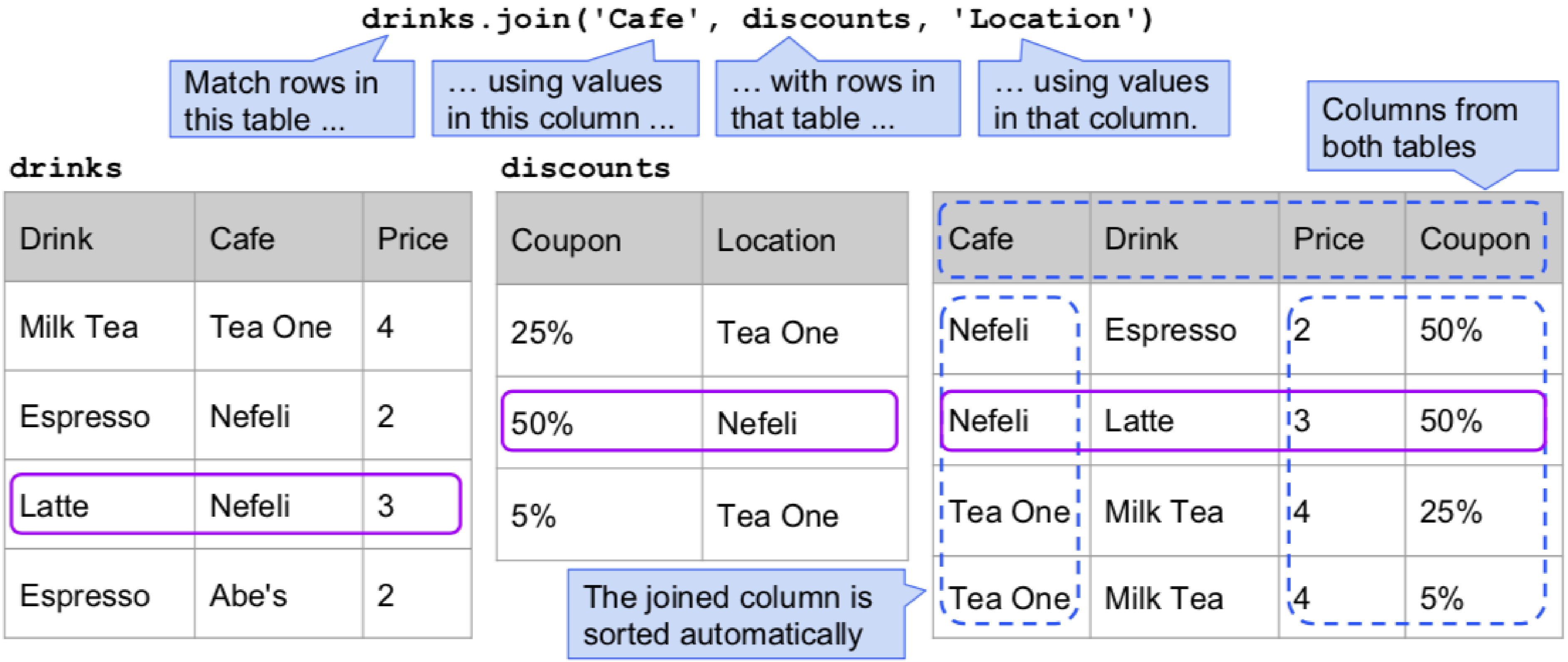
Joining involves combining the rows of two tables together into a new table

- A column in each table needs to be specified which indicates how the rows should be combined

tb1.join("col tb1", tb2, "col tb2")

- tb1: the first table
- "col tb1": a column in the first table
- tb2: the second table
- "col tb2": a column in the second table

Joining Two Tables



```
drinks.join('Cafe', discounts, 'Location')
```

Match rows in this table ...

... using values in this column ...

... with rows in that table ...

... using values in that column.

Columns from both tables

drinks

Drink	Cafe	Price
Milk Tea	Tea One	4
Espresso	Nefeli	2
Latte	Nefeli	3
Espresso	Abe's	2

discounts

Coupon	Location
25%	Tea One
50%	Nefeli
5%	Tea One

The joined column is sorted automatically

Cafe	Drink	Price	Coupon
Nefeli	Espresso	2	50%
Nefeli	Latte	3	50%
Tea One	Milk Tea	4	25%
Tea One	Milk Tea	4	5%

Let's explore this in Jupyter!