DSC 10, Spring 2018
Lecture 10

Group and Join

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Grouping Rows

(Demo)

Group

The `group` method aggregates all rows with the same value for a column into a single row in the result

- First argument: Which column to group by
- Second argument: (Optional) How to combine values
  - `len` — number of grouped values (default)
  - `sum` — total of all grouped values
  - `list` — list of all grouped values

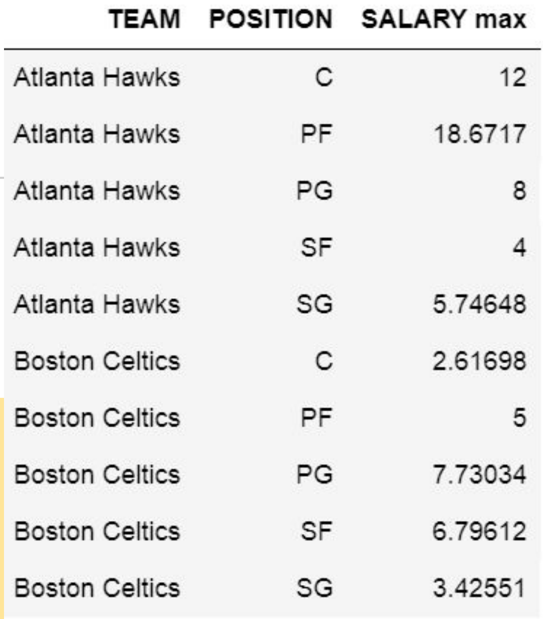
Group with Multiple Columns

The `group` method can also aggregate all rows that share the combination of values in multiple columns

- First argument: List or array of which columns to group by
- Second argument: (Optional) How to combine values

Discussion Question

- A *starter* for a team is the player with the highest salary on that team in that position.
- The name of the table shown is *starters*.



TEAM	POSITION	SALARY max
Atlanta Hawks	C	12
Atlanta Hawks	PF	18.6717
Atlanta Hawks	PG	8
Atlanta Hawks	SF	4
Atlanta Hawks	SG	5.74648
Boston Celtics	C	2.61698
Boston Celtics	PF	5
Boston Celtics	PG	7.73034
Boston Celtics	SF	6.79612
Boston Celtics	SG	3.42551

Which will rank the teams in order of their highest-paid starter?

A. `starters.group(‘TEAM’, max).sort(1, descending = True)`

B. `starters.drop('POSITION').group('TEAM', max).sort(1, descending = True)`

C. `starters.select('TEAM', 'SALARY').group('TEAM', max).sort(1, descending=True)`

D. `starters.select('TEAM', 'SALARY max').group('TEAM', max).sort(1, descending = True)`

E. More than one of the above

Joining Tables

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

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

The joined column is sorted automatically

Random Selection

Random Selection

`np.random.choice`

- Selects at random
- with replacement
- from an array
- a specified number of times

```
np.random.choice(some_array, sample_size)
```

Discussion Question

```
d = np.arange(6) + 1
```

What happens when we evaluate the following 2 expressions?

- `np.random.choice(d, 1000) + np.random.choice(d, 1000)`
- `2 * np.random.choice(d, 1000)`

A. Gives the same result; Describing the same process
B. Gives the same result; Describing different processes
C. Gives different results; Describing the same process
D. Gives different results; Describing different processes
E. None of the above