YData: An Introduction to Data Science

Lecture 11: Joins

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University

Spring 2021

Credit: data8.org

Announcements

- Homework assignment 04 is posted
- Project 1 is posted on our calendar. For the checkpoint you will have to submit your solution for the first 8 questions.

Project 1: World Progress

Video: www.gapminder.org/videos/dont-panic-the-facts-about-population/

Pivot Tables

Pivot

- Cross-classifies according to two categorical variables
- 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`

(DEMO)

Challenge Question

Which NBA teams spent the most on their “starters” in 2015-2016?

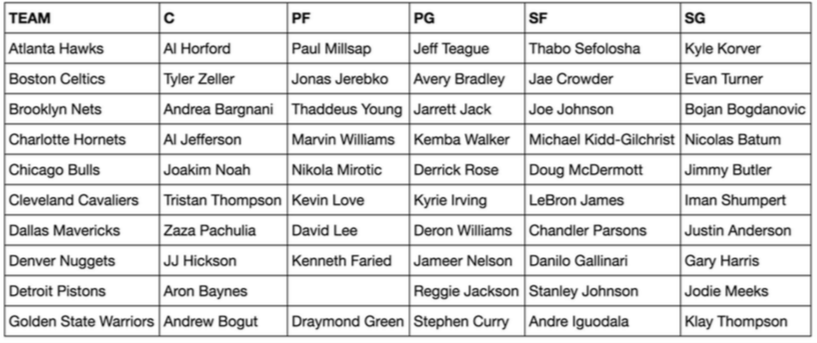
Assume the “starter” for a team & position is the player with the highest salary on that team in that position.

PLAYER	POSITION	TEAM	SALARY
Paul Millsap	PF	Atlanta Hawks	18.6717
Al Horford	C	Atlanta Hawks	12
Tiago Splitter	C	Atlanta Hawks	9.75625

(DEMO)

Take-Home Question

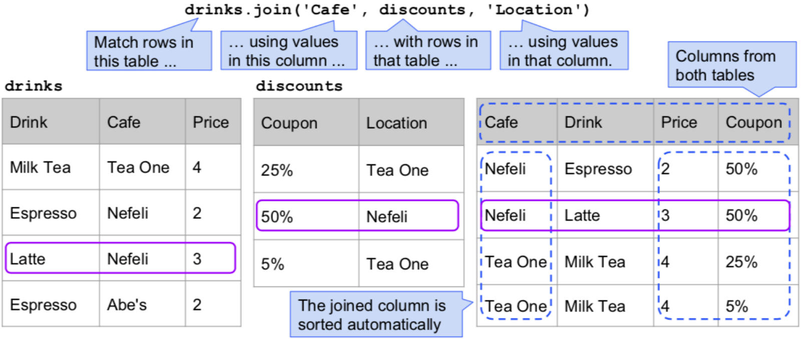
Generate a table of the names of the starters for each team



TEAM	C	PF	PG	SF	SG
Atlanta Hawks	Al Horford	Paul Millsap	Jeff Teague	Thabo Sefolosha	Kyle Korver
Boston Celtics	Tyler Zeller	Jonas Jerebko	Avery Bradley	Jae Crowder	Evan Turner
Brooklyn Nets	Andrea Bargnani	Thaddeus Young	Jarrett Jack	Joe Johnson	Bojan Bogdanovic
Charlotte Hornets	Al Jefferson	Marvin Williams	Kemba Walker	Michael Kidd-Gilchrist	Nicolas Batum
Chicago Bulls	Joakim Noah	Nikola Mirotic	Derrick Rose	Doug McDermott	Jimmy Butler
Cleveland Cavaliers	Tristan Thompson	Kevin Love	Kyrie Irving	LeBron James	Iman Shumpert
Dallas Mavericks	Zaza Pachulia	David Lee	Deron Williams	Chandler Parsons	Justin Anderson
Denver Nuggets	JJ Hickson	Kenneth Faried	Jameer Nelson	Danilo Gallinari	Gary Harris
Detroit Pistons	Aron Baynes		Reggie Jackson	Stanley Johnson	Jodie Meeks
Golden State Warriors	Andrew Bogut	Draymond Green	Stephen Curry	Andre Iguodala	Klay Thompson

Joins

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

(DEMO)

Bikes

(DEMO)

Shortest Trips

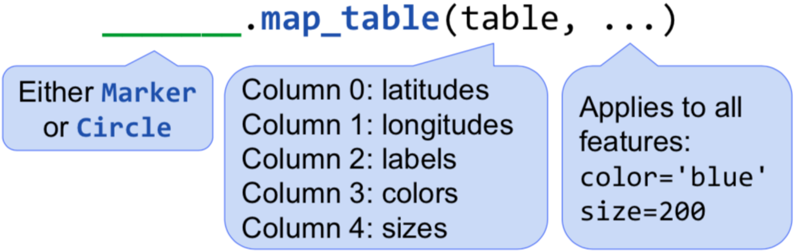
(DEMO)

Maps

(DEMO)

Maps

A table containing columns of latitude and longitude values can be used to generate a map of markers



```
______.map_table(table, ...)
```

Either `Marker` or `Circle`

Column 0: latitudes
Column 1: longitudes
Column 2: labels
Column 3: colors
Column 4: sizes

Applies to all features:
`color='blue'`
`size=200`