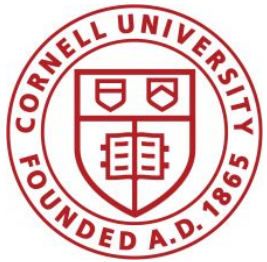

DSFA

Spring 2020

Lecture 8

Groups, Joins, and Maps

Announcements

- Project 1 out this Saturday AM.
  - Can work on it with a partner from your same lab section (or by yourself if you prefer).
  - Note: Only work on one copy of the notebook at a time!
- Prelim 1 is Thursday, Feb. 27. More info early next week.

What we'll do: Citibike visualization

Learn enough computing to do our own visualizations and observations to identify patterns in big data sets.

Grouping Rows

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

(Demo)

Grouping By Two Columns

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

- First argument: A list of which columns to group by
- Second argument: (Optional) How to combine values

(Demo)

Challenge Question

Which NBA teams spent the most on their starters in 2016?

- Each team has one *starter* per position
- 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

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)

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')
```

Keep all rows in the table that have a match ...

… for the value in this column ...

… somewhere in this other table's ...

… column that contains matching values.

drinks

Drink	Cafe	Price
Milk tea	Panda Tea	4
Espresso	Gimme	2
Latte	Gimme	3
Espresso	Cafe Gola	2

discounts

Coupon	Location
25%	Panda Tea
50%	Gimme
5%	Gimme

The joined column is sorted automatically

Cafe	Drink	Price	Coupon
Gimme	Espresso	2	50%
Gimme	Espresso	2	5%
Gimme	Latte	3	50%
Gimme	Latte	3	5%
Panda Tea	Milk Tea	4	25%

(Demo)

Bikes

Maps

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`