DATA 8

Summer 2018

Lecture 09

Groups

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Contributions by Fahad Kamran (fhdkmrn@berkeley.edu) and Vinitra Swamy (vinitra@berkeley.edu)

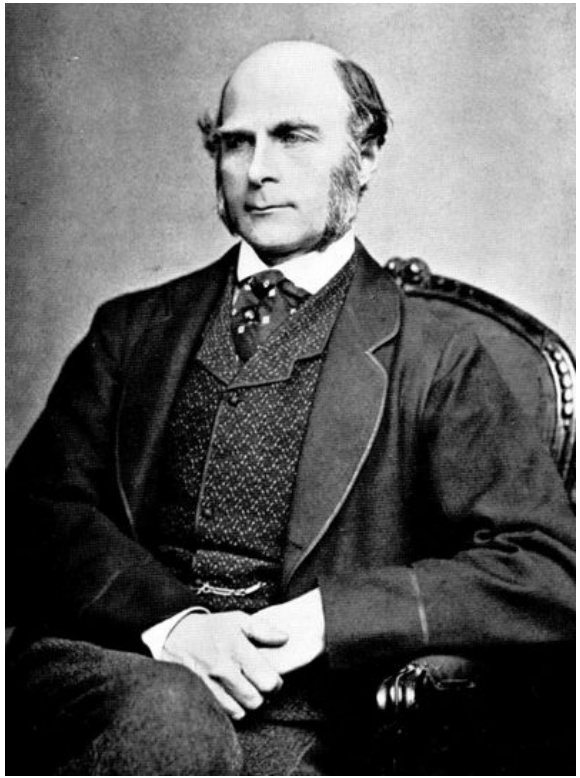
Announcements

Example: Prediction

Sir Francis Galton

- 1822 - 1911 (knighted in 1909)
- A pioneer in making predictions
- Particular (and troublesome) interest in heredity
- Charles Darwin's half-cousin

(Demo)



Apply with Multiple Columns

Apply

The `apply` method creates an array by calling a function on every element in one or more input columns

- First argument: Function to apply
- Other arguments: The input column(s)

```
table_name.apply(one_arg_function, 'column_label')
table_name.apply(two_arg_function,
                 'column_label_for_first_arg',
                 'column_label_for_second_arg')
```

`apply` called with only a function applies it to each row

(Demo)

Grouping by One Attribute

Grouping by One Column

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

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

(Demo)

Cross-Classification

Grouping By Multiple 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)

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 2017-2018?

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

Player	Position	Team	Salary
Stephen Curry	PG	Golden State Warriors	34.3826
LeBron James	SF	Cleveland Cavaliers	33.2857
Paul Millsap	PF	Denver Nuggets	31.2692

(Demo)

Take-Home Question

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

Team	C	PF	PG	SF	SG
Atlanta Hawks	Miles Plumlee	Mike Muscala	Dennis Schroder	Taurean Prince	Kent Bazemore
Boston Celtics	Aron Baynes	Al Horford	Kyrie Irving	Gordon Hayward	Jaylen Brown
Brooklyn Nets	Timofey Mozgov	Luis Scola	Jeremy Lin	DeMarre Carroll	Allen Crabbe
Charlotte Hornets	Dwight Howard	Marvin Williams	Kemba Walker	Michael Kidd-Gilchrist	Nicolas Batum
Chicago Bulls	Robin Lopez	Nikola Mirotic	Kris Dunn	Denzel Valentine	Justin Holiday
Cleveland Cavaliers	Kevin Love	Ante Zizic	George Hill	LeBron James	JR Smith
Dallas Mavericks	Dwight Powell	Josh McRoberts	J.J. Barea	Harrison Barnes	Wesley Matthews
Denver Nuggets	Mason Plumlee	Paul Millsap	Devin Harris	Wilson Chandler	Mike Miller
Detroit Pistons	Andre Drummond	Blake Griffin	Reggie Jackson	Stanley Johnson	Langston Galloway
Golden State Warriors	Zaza Pachulia	Draymond Green	Stephen Curry	Kevin Durant	Klay Thompson