YData: An Introduction to Data Science

Lecture 10: Groups

Jessi Cisewski-Kehe
Statistics & Data Science, Yale University
Spring 2020

Credit: data8.org

Announcements

- HW03 due Thursday by 11:59 PM
- Friday, 2/7, lecture rescheduled for TODAY (Wed.), 4 - 4:50 PM in OML 202

  It will be recorded and posted on Canvas
- Project 1 is posted on our calendar

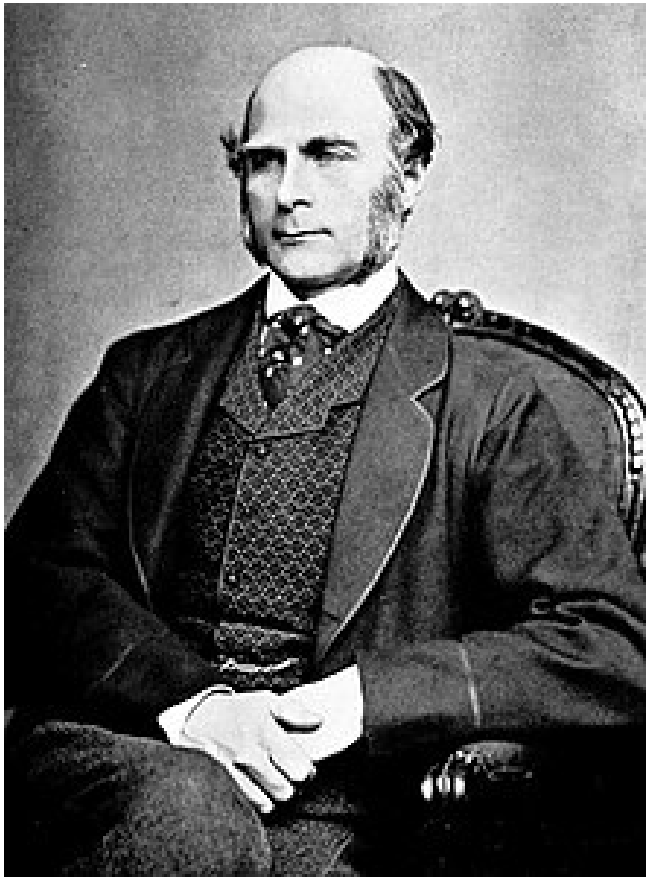
  Checkpoint Fri 2/14; Due Fri 2/21

Example: Predictions

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 input column(s)

- 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
- Other arguments: (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 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