DATA 8

Spring 2022

Lecture 11

Pivots and Joins

Announcements

Rows from Lists

(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
- collect=...:   (Optional) How to combine values

Group by **one** column (and minimize each other column):
`t.group("label", collect=min)` **or** `t.group("label", min)`

Group by **two** columns (and minimize each other):
`t.group(["label 1", "label 2"], collect=min)` **or**
`t.group(["label 1", "label 2"], min)`

PIVOT

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 the grid
  - Second: variable that forms row labels of the grid
- Two optional arguments (include **both** or **neither**)
  - `values`='column_label_to_aggregate'
  - `collect`=function_to_aggregate_with

(Demo)

Let's Practice

How to solve a table manipulation problem:

1. **Understand the result** — what information is needed to fill in a particular value of the goal.
2. **Describe the operations** — articulate (in English) the operations that will produce the result.
3. **Write Python expressions** — express the operations using table methods, functions, & array arithmetic.

Discussion Question

1. For each city, what's the height of the tallest building for each material?
2. For each city, what's the age difference between the oldest steel building and the oldest concrete building?

sky

(Demo)

name	material	city	height	age
Metropolitan Tower	concrete	New York City	218.24	35
Paul Hastings Tower	steel	Los Angeles	213.06	49
Barclay Tower	concrete	New York City	205.06	13
Westin Peachtree Plaza	concrete	Atlanta	220.37	44
Wells Fargo Plaza	steel	Houston	302.37	37

Challenge Question

Generate a table of the names of the oldest buildings for each material for each city:

city	concrete	mixed/composite	steel
San Francisco	Coit Tower	Transamerica Pyramid	Ferry Building
Baltimore	Charles Towers North Apartments		Emerson Tower
Detroit	Renaissance Center 400 Tower		Michigan Central Station
Minneapolis	River Towers A	IDS Tower	Soo Line Building
Columbus	Key Bank Building		Leveque Tower

Group or Pivot?

- Distribution of one categorical variable => .group()
- Cross-classification of two or more categorical variables:
  - One row per combination => .group()
  - One variable vertically, one horizontally => .pivot()

JOIN

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	Asha	5.5
Espresso	Strada	1.75
Latte	Strada	3.25
Espresso	FSM	2

discounts

Coupon	Location
10%	Asha
25%	Strada
5%	Asha

Cafe	Drink	Price	Coupon
Asha	Milk Tea	5.5	10%
Asha	Milk Tea	5.5	5%
Strada	Espresso	1.75	25%
Strada	Latte	3.25	25%

The joined column is sorted automatically

(Demo)

Table Review

Important Table Methods

`t.select(column, …)` *or* `t.drop(column, …)`

`t.take([row_num, …])` *or* `t.exclude([row_num, …])`

`t.sort(column, descending=False, distinct=False)`

`t.where(column, are.condition(...))`

`t.apply(function_name, column, …)`

`t.group(column)` *or* `t.group(column, function_name)`

`t.group([column, …])` *or* `t.group([column, …], function_name)`

`t.pivot(cols, rows)` *or* `t.pivot(cols, rows, vals, function_name)`

`t.join(column, other_table, other_table_column)`

http://data8.org/sp22/python-reference.html