DATA 8

Spring 2020

Lecture 11

Pivots and Joins

Weekly Goals

- **Monday**
  - Review of graphs
  - Writing our own functions
- **Wednesday**
  - Making predictions
  - Aggregating data using `group`
- **Today**
  - `group` and `pivot`
  - Combining tables using `join`

Lists

(Demo)

Lists are Generic Sequences

A list is a sequence of values (just like an array), but the values can all have different types

```
[2+3, 'four', Table().with_column('K', [3, 4])]
```

- Lists can be used to create table rows.
- If you create a table column from a list, it will be converted to an array automatically.
- Lists can even contain other lists.

Grouping

(Demo)

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
  - etc.

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_to_aggregate_with

(Demo)

Challenge Question

1. For each city, what's 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`

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

(Demo)

Take-Home 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?

For cross-classification:

Pivot Table

- One combo of grouping variables **per entry**
- **Two** grouping variables: columns and rows
- Aggregate values of **values column**
- Missing combos **= 0 (or empty string)**

Grouped Table

- One combo of grouping variables **per row**
- **Any number** of grouping variables
- Aggregate values of **all other columns** in table
- Missing combos **absent**

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)