MATH1401

Fall 2021

Lecture 11

Groups

Class Checklist

- **Lab 4 - Due Date** : Thursday: 9/30 – 9 PM
  - Graded Questions:1.1-1.5, 2.1-2.4, 3.1-3.6, 4.2
- **Quiz 8** – Tuesday: 9/28 – Covers Chapter 8

Review of Lecture 10

Apply

The `apply` method creates an array by calling a function on every element in input column(s)

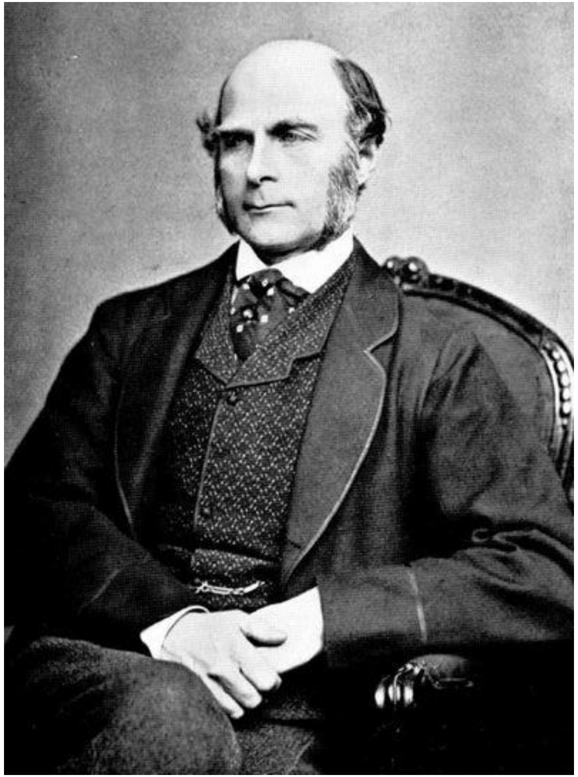
- First argument: Function to apply
- Other arguments: The input column(s)

```
table_name.apply(function_name, 'column_label')
```

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)



Lecture 11 Checklist – Chapter 8.2, 8.3

- **Apply** – apply function to column
- **Group** – find categorical distribution

Apply functions and make table

- **Group** – find groups based on categorical variables
- **Pivot** – Display groups of categorical variables

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

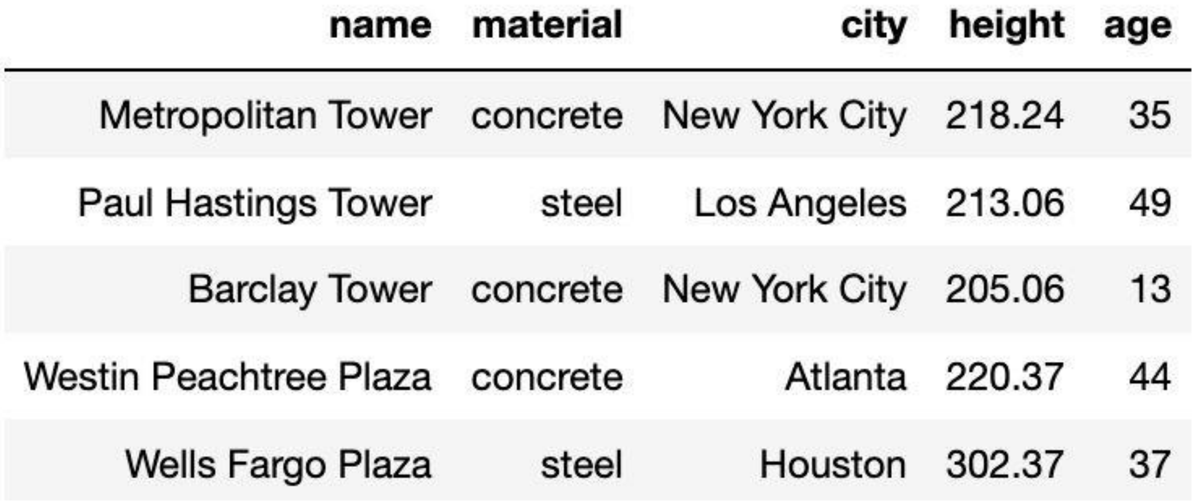
- Cross-classifies according to two categorical variables
- Produces a grid of counts or aggregated values
- Two required arguments:
  - First: variable that forms column labels of grid
  - Second: variable that forms row labels of grid
- Two optional arguments (include **both** or **neither**)
  - `values`=’column_label_to_aggregate’
  - `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)

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**