DATA 1202

Spring 2024

Lecture 12

Table Examples

Announcements

- **Lab 6** on Thursday due 2/16 @ 5pm
- **HW 6** due Wednesday 2/21 @ 11pm
- **Project 1** due Friday 3/8
  - Checkpoint due Friday 2/23
  - World population through data
  - Covers topics in lecture through today
- **Midterm** in-class, Wednesday 2/21

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)
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)
```

https://pages.mtu.edu/~lebrown/data1202-s24/reference/index.html

Table Practice

Join for Value Annotation

One common use of `t.join(_, u, _)`:

- A table `t` has a categorical variable x.
- A table `u` has one row per possible value of x that describes some properties of that value.
- The joined table has the same rows as `t`, but each row in `t` is now annotated with the properties of its x value.

(Demo)

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

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

Group or Pivot?

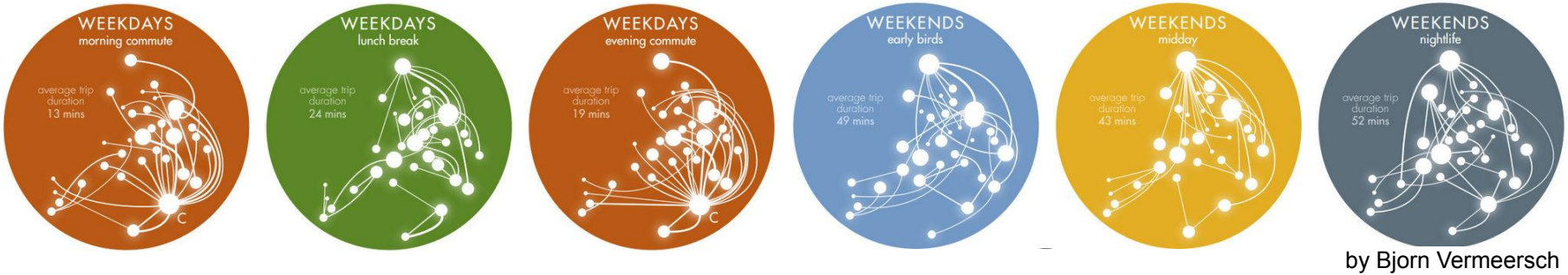
- When to **Group**:
  - aggregates of one categorical variable
  - aggregates of **many variables**
  - **Multiple** outputs (aggregate columns)

- When to **Pivot**:
  - Aggregates of **exactly two variables**
  - **Few unique values** for column variable
  - Interested in **every combination** of values

Bike Sharing in SF Bay Area

Hourly bike sharing in the Bay Area began with a pilot program in 2014-2015 that produced a public dataset.

- The SF Metropolitan Transportation Commission organized an Open Data Challenge in which participants visualized the dataset in interesting ways.



by Bjorn Vermeersch

(Demo)