Lecture 10

Joins and Table Examples
Announcements
Project 1: World Progress

Video: https://www.gapminder.org/videos/dont-panic-the-facts-about-population/
Pivot
**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)
Joins
## Joining Two Tables

**drinks.join('Cafe', discounts, 'Location')**

- **drinks** table:
  - **Drink** | **Cafe** | **Price**
  - Milk Tea | Tea One | 4
  - Espresso | Nefeli | 2
  - Latte | Nefeli | 3
  - Espresso | Abe's | 2

- **discounts** table:
  - **Coupon** | **Location**
  - 25% | Tea One
  - 50% | Nefeli
  - 5% | Tea One

**Columns from both tables**

- **Cafe** | **Drink** | **Price** | **Coupon**
  - Nefeli | Espresso | 2 | 50%
  - Nefeli | Latte | 3 | 50%
  - Tea One | Milk Tea | 4 | 25%
  - Tea One | Milk Tea | 4 | 5%

The joined column is sorted automatically.

(Demo)
Bikes

(Demo)
Shortest Trips

(Demo)
Maps

(Demo)
A table containing columns of latitude and longitude values can be used to generate a map of markers.

 Either **Marker** or **Circle**

 Column 0: latitudes  
 Column 1: longitudes  
 Column 2: labels  
 Column 3: colors  
 Column 4: sizes

 Applies to all features:  
 color='blue'  
 size=200
Break
Combining Table Methods
Important Table Methods

t.select(column, ...) or t.drop(column, ...)
t.take([row, ...]) or t.exclude([row, ...])
t.sort(column, descending=False, distinct=False)
t.where(column, are.condition(...))
t.apply(function, column, ...)
t.group(column) or t.group(column, function)
t.group([column, ...]) or t.group([column, ...], function)
t.pivot(cols, rows) or t.pivot(cols, rows, vals, function)
t.join(column, other_table, other_table_column)

http://data8.org/datascience/tables.html
Discussion Question

Generate a table with one row per cafe that has the name and discounted price of its cheapest discounted drink

<table>
<thead>
<tr>
<th>drinks</th>
<th>discounts</th>
<th>cheapest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drink</td>
<td>Cafe</td>
<td>Price</td>
</tr>
<tr>
<td>Milk Tea</td>
<td>Tea One</td>
<td>4</td>
</tr>
<tr>
<td>Espresso</td>
<td>Nefeli</td>
<td>2</td>
</tr>
<tr>
<td>Coffee</td>
<td>Nefeli</td>
<td>3</td>
</tr>
<tr>
<td>Espresso</td>
<td>Abe's</td>
<td>2</td>
</tr>
</tbody>
</table>

(Demo)
(b) (8 pt) Each row of the trip table from lecture describes a single bicycle rental in the San Francisco area. Durations are integers representing times in seconds. The first three rows out of 338,343 appear below.

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ferry Building</td>
<td>SF Caltrain</td>
<td>765</td>
</tr>
<tr>
<td>San Antonio Shopping Center</td>
<td>Mountain View City Hall</td>
<td>1036</td>
</tr>
<tr>
<td>Post at Kearny</td>
<td>2nd at South Park</td>
<td>307</td>
</tr>
</tbody>
</table>

Write a Python expression below each of the following descriptions that computes its value. The first one is provided for you. You may use up to two lines and introduce variables.

- The average duration of a rental.
  
  ```python
  total_duration = sum(trip.column(2))
  total_duration / trip.num_rows
  ```

- The name of the station where the most rentals ended (assume no ties).

- The number of stations for which the average duration ending at that station was more than 10 minutes.
<table>
<thead>
<tr>
<th>Team</th>
<th>C</th>
<th>PF</th>
<th>PG</th>
<th>SF</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta Hawks</td>
<td>Miles Plumlee</td>
<td>Mike Muscala</td>
<td>Dennis Schroder</td>
<td>Taurean Prince</td>
<td>Kent Bazemore</td>
</tr>
<tr>
<td>Boston Celtics</td>
<td>Aron Baynes</td>
<td>Al Horford</td>
<td>Kyrie Irving</td>
<td>Gordon Hayward</td>
<td>Jaylen Brown</td>
</tr>
<tr>
<td>Brooklyn Nets</td>
<td>Timofey Mozgov</td>
<td>Luis Scola</td>
<td>Jeremy Lin</td>
<td>DeMarre Carroll</td>
<td>Allen Crabbe</td>
</tr>
<tr>
<td>Charlotte Hornets</td>
<td>Dwight Howard</td>
<td>Marvin Williams</td>
<td>Kemba Walker</td>
<td>Michael Kidd-Gilchrist</td>
<td>Nicolas Batum</td>
</tr>
<tr>
<td>Chicago Bulls</td>
<td>Robin Lopez</td>
<td>Nikola Mirotic</td>
<td>Kris Dunn</td>
<td>Denzel Valentine</td>
<td>Justin Holiday</td>
</tr>
<tr>
<td>Cleveland Cavaliers</td>
<td>Kevin Love</td>
<td>Ante Zizic</td>
<td>George Hill</td>
<td>LeBron James</td>
<td>JR Smith</td>
</tr>
<tr>
<td>Dallas Mavericks</td>
<td>Dwight Powell</td>
<td>Josh McRoberts</td>
<td>J.J. Barea</td>
<td>Harrison Barnes</td>
<td>Wesley Matthews</td>
</tr>
<tr>
<td>Denver Nuggets</td>
<td>Mason Plumlee</td>
<td>Paul Millsap</td>
<td>Devin Harris</td>
<td>Wilson Chandler</td>
<td>Mike Miller</td>
</tr>
<tr>
<td>Detroit Pistons</td>
<td>Andre Drummond</td>
<td>Blake Griffin</td>
<td>Reggie Jackson</td>
<td>Stanley Johnson</td>
<td>Langston Galloway</td>
</tr>
<tr>
<td>Golden State Warriors</td>
<td>Zaza Pachulia</td>
<td>Draymond Green</td>
<td>Stephen Curry</td>
<td>Kevin Durant</td>
<td>Klay Thompson</td>
</tr>
</tbody>
</table>