YData: Introduction to Data Science

Lecture 05: Arrays and Tables

Overview

Arrays continued

Ranges

Tables!

Additional table topics (if there is time)

- Extracting columns from Tables
- Creating a table from scratch

Announcements: Homework and additional practice exercises

Homework 1 is due on Sunday (2/6) at 11pm

- Submit the pdf to Gradescope and be sure to make the pages for each question!

For additional practice see:

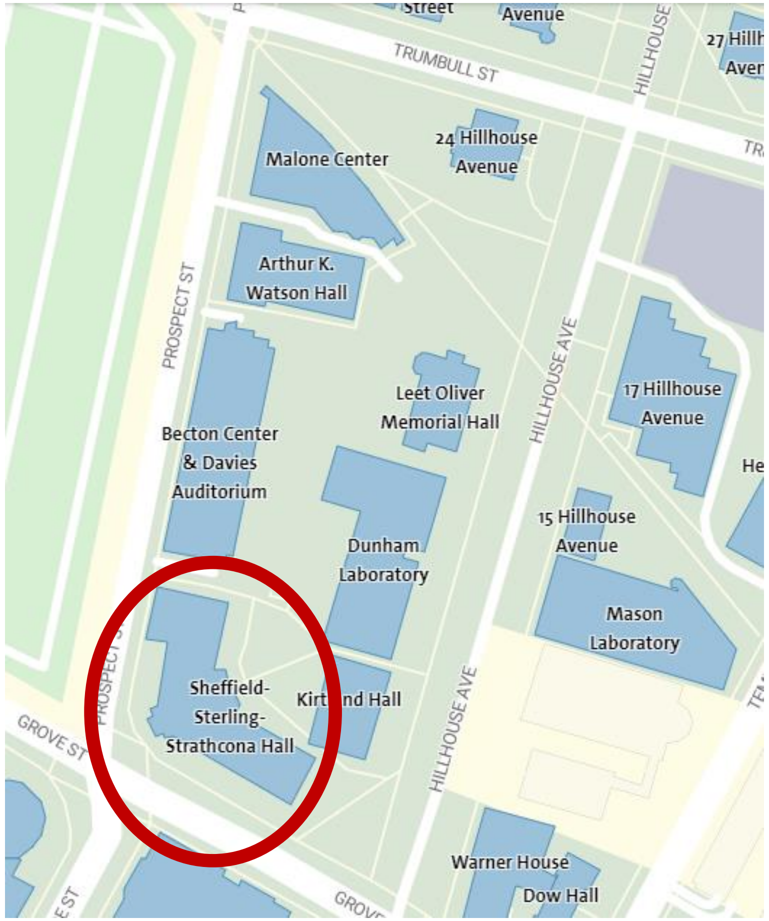
- Practice 01 and 02 Jupyter Notebooks on the class calendar site
- The class textbook has additional examples
- There are a few additional examples in today's lecture slide class folder
  - (we will cover similar techniques with different data next week)

Announcements: In person class

Next class (on Monday) is in person
in Sheffield-Sterling-Strathcona 114

Bring a laptop with Anaconda/Python and make sure your laptop is fully charged

- There are no power outlets in the classroom



Review and continuation of arrays

Array Review (i.e., NumPy ndarrays)

An array contains a sequence of values

- All elements of an array must have the same type
- We can apply fast operations to all elements of an array
  - E.g., we can add a number to all elements of a numeric array
- When two arrays are added corresponding elements are added in the result
  - Note, the two arrays must have the same size

Let's explore this in Jupyter!

Ranges

Ranges

A range is an array of consecutive numbers

An array of increasing integers from 0 up to end - 1

- np.arange(end)

An array of increasing integers from start up to end - 1

- np.arange(start, end)

A range with step between consecutive values

- np.arange(start, end, step)

The range always includes start but excludes end

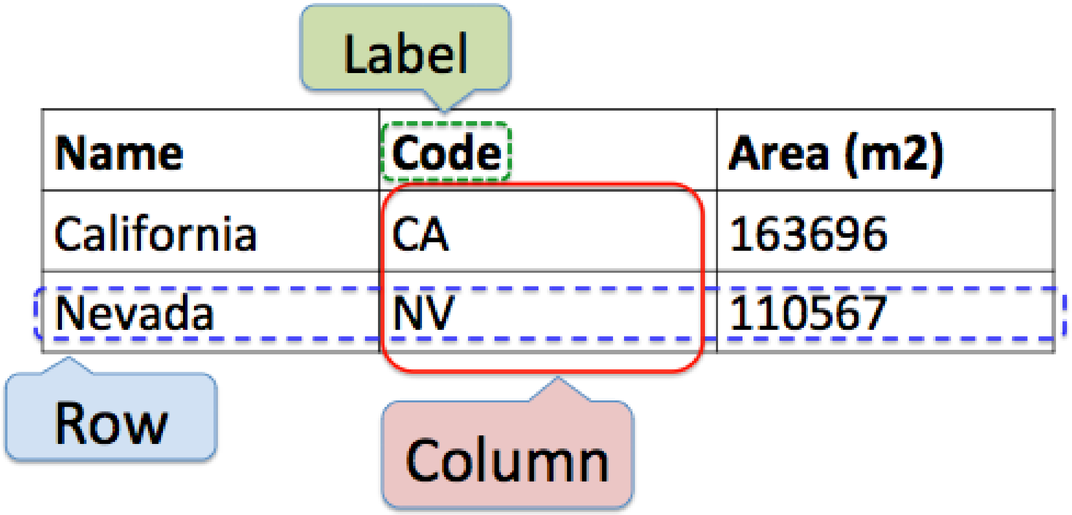
Let's explore this in Jupyter!

Tables

Table structure

A Table is a sequence of labeled columns

- Each row represents one individual case
- Data within a column represents one attribute



Name	Code	Area (m2)
California	CA	163696
Nevada	NV	110567

Some Table Operations

tb.select(label) - constructs a new table with just the specified columns

tb.drop(label) - constructs a new table in which the specified columns are omitted

tb.sort(label) - constructs a new table with rows sorted by the specified column

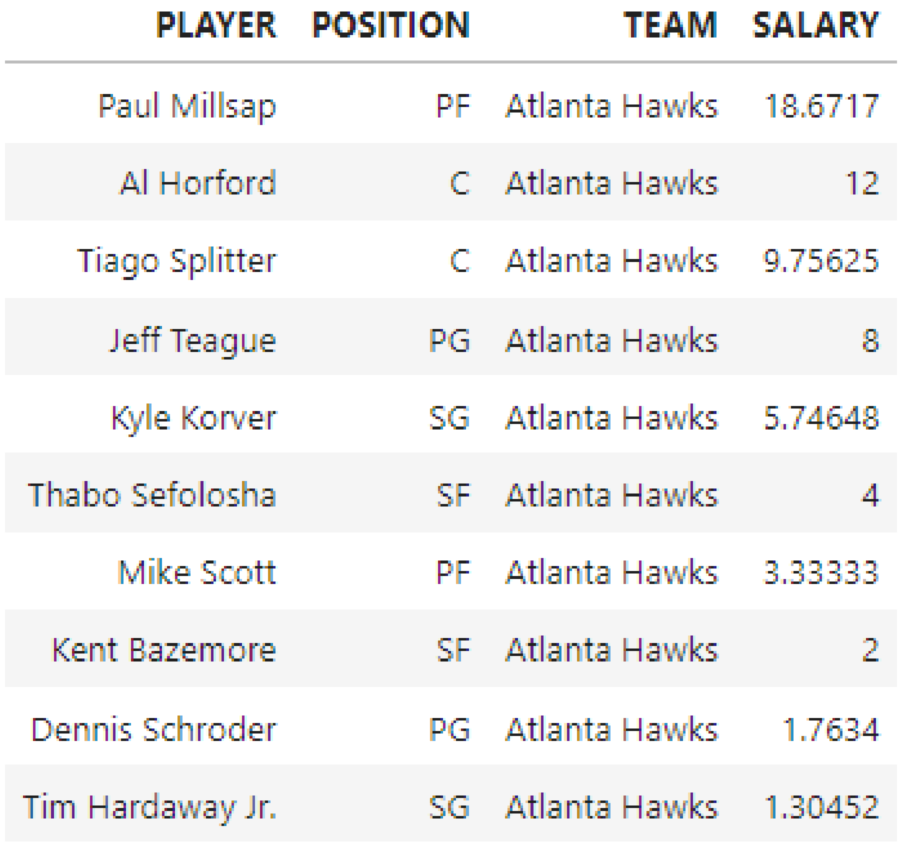
tb.where(label, condition) - constructs a new table with just the rows that match the condition

See Berkeley's documentation

Let's explore this in Jupyter!

Example: NBA salaries

Let's explore salaries of NBA players (from the 2015-2016 season)



PLAYER	POSITION	TEAM	SALARY
Paul Millsap	PF	Atlanta Hawks	18.6717
Al Horford	C	Atlanta Hawks	12
Tiago Splitter	C	Atlanta Hawks	9.75625
Jeff Teague	PG	Atlanta Hawks	8
Kyle Korver	SG	Atlanta Hawks	5.74648
Thabo Sefolosha	SF	Atlanta Hawks	4
Mike Scott	PF	Atlanta Hawks	3.33333
Kent Bazemore	SF	Atlanta Hawks	2
Dennis Schroder	PG	Atlanta Hawks	1.7634
Tim Hardaway Jr.	SG	Atlanta Hawks	1.30452

Pandas

FYI: The datascience package is a Berkeley product

It's a light wrapper on top of pandas

Hopefully at the end of the class we'll have time to discuss Pandas

Ways to create a Table

Table.read_table(filename) - reads a table from a spreadsheet

Table() - an empty table

We can build a Table ourselves by creating an empty Table and then adding columns

- Table().with_column("column_name", ndarray)

Let's explore this in Jupyter!