DATA 8

Spring 2022

Lecture 6

Census

Announcements

- HW 2 is due Thursday, 2/3
  - Submit Wednesday for a bit of extra credit
- HW 1 and Lab 2 solutions released tonight
  - Lab 1 regrades due this Wednesday
- **Reminder:** Tutoring sections start this week!
  - There are still spots available
- Look out for Ed post on midterm/final exam conflicts!

Weekly Goals

- **Today**
  - Table review
  - Working with Census data
- Wednesday
  - Visualizing data
  - Distributions
- Friday
  - Visualizing two kinds of distributions
  - Proportions as areas

Table Review

Table Structure

- A Table is a sequence of labeled columns
- Labels are strings
- Columns are arrays, all with the same length

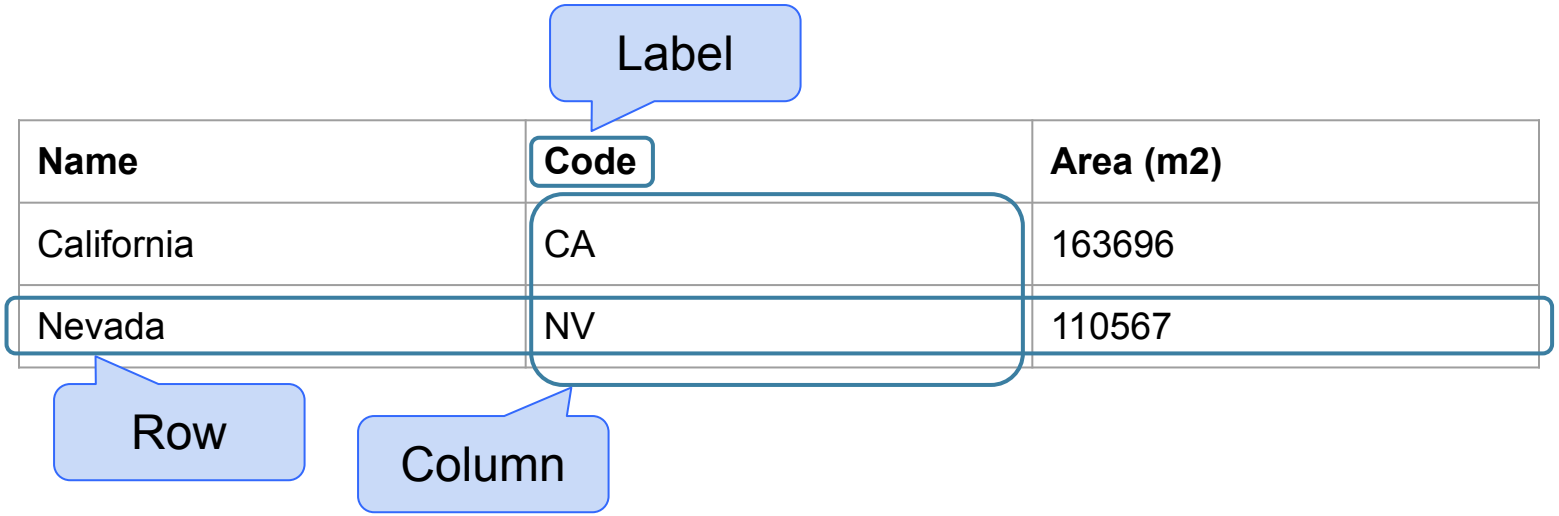

Table Methods

- Creating and extending tables:
  - `Table.read_table` and `Table().with_columns`
- Finding the size: `num_rows` and `num_columns`
- Referring to columns: by labels or indices
  - column indices start at 0
- Accessing data in a column
  - `column` takes a label or index and **returns an array**
- Using array methods to work with data in columns
  - `item`, `sum`, `min`, `max`, and so on
- Creating new tables containing some of the original columns:
  - `select`, `drop`

Manipulating Rows

- `t.sort(column)` sorts the rows in increasing order
- `t.sort(column, descending=True)` sorts the rows in decreasing order
- `t.take(row_numbers)` keeps the numbered rows
  - Each row has an index, starting at 0
- `t.where(column, are.condition)` keeps all rows for which a column's value satisfies a condition
- `t.where(column, are.equal_to(value))` keeps all rows for which a column's value equals some particular value
  - Shorter form: `t.where(column, value)`

(Demo)

Discussion Questions

The table **`nba`** has columns **`PLAYER`**, **`POSITION`**, and **`SALARY`**.

a) Create an array containing the names of all point guards (**`PG`**) who made more than $15M

```
guards = nba.where('POSITION', 'PG')
guards.where('SALARY', are.above(15)).column('PLAYER')
```

b) After evaluating these two expressions in order, what's the result of the second one?

```
nba.drop('POSITION')
nba.num_columns
```

(Demo)

Census Data

The Decennial Census

- Every ten years, the Census Bureau counts how many people there are in the U.S.
- In between censuses, the Bureau estimates how many people there are each year.
- Article 1, Section 2 of the Constitution:
  - “Representatives and direct Taxes shall be apportioned among the several States … according to their respective Numbers …”

Census Table Description

- Values have column-dependent interpretations
  - The SEX column: 1 is *Male*, 2 is *Female* (only two categories)
  - The POPESTIMATE2010 column: *7/1/2010 estimate*
- In this table, some rows are sums of other rows
  - The SEX column: 0 is *Total* (of *Male* and *Female* categories)
  - The AGE column: 999 is *Total* of all ages
- Numeric codes are often used for storage efficiency
- Values in a column have the same type, but are not necessarily comparable (AGE 12 vs AGE 999)

https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2019/nc-est2019-agesex-res.pdf

Analyzing Census Data

Leads to the discovery of interesting features and trends in the population

(Demo)

Binary Code in SEX Column

- Pretty much the same question since 1790
- “Male” (code 1), “Female” (code 2)
  - Have to select one of these
  - No other options
- Sex categorization is more complex.
- Historical form of the question can lead to
  - Non-response
  - Inaccurate response (Demo)
- This can make Census data hard to interpret and use.
- We will use the Census data keeping these caveats in mind.