#### YData: An Introduction to Data Science

Lecture 06: Census

Elena Khusainova & John Lafferty Statistics & Data Science, Yale University Spring 2021

Credit: data8.org

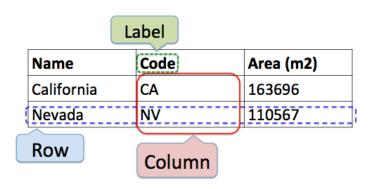


## Announcements

## 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().with\_column and Table.read\_table
- Finding the size: num\_rows and num\_columns
- Referring to columns: labels, relabeling, and indices
  - labels and relabeled; 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.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, value) keeps all rows containing a certain value in a column
- t.with\_row makes a new table that has another row

# Lists

### **Lists are Generic Sequences**

A list is a sequence of values (just like an array), but the values can all have different types

$$[2 + 3, \frac{\text{'four'}}{\text{, Table().with\_column('K', }} [3, 4])]$$

- 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

(DEMO)

### **Discussion Questions**

The table nba has columns NAME, POSITION, and SALARY.

a) Create an array containing the names of all point guards (PG) who make more than \$15M/year

```
nba.where(1, 'PG').where(2, are.above(15)).column(0)
```

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

```
nba.with_row(['Samosa', 'Mascot', 100])
nba.where('NAME', are.containing('Samo'))
```

## 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  $\dots$  according to their respective Numbers  $\dots$ "

## **Analyzing Census Data**

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

(DEMO)

### **Census Table Description**

- Values have column-dependent interpretations
  - The SEX column: 1 is Male, 2 is Female
  - 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 + Female)
  - 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/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.csv

### **Growth Rate**

- Growth rate = g (for example 3%, or 0.03)
- Initial value x, final value y after t periods of time

Value after 1 period 
$$= x + xg$$
  $= x * (1+g)$   
Value after 2 periods  $= x(1+g)(1+g)$   $= x * (1+g) ** 2$   
Value after t periods  $= y$   $= x * (1+g) ** t$ 

So 
$$(1+g)$$
 \*\*  $t = y/x$  and so  $1+g = (y/x)$  \*\*  $(1/t)$ 

So 
$$g = (y/x) ** (1/t) - 1$$