DSC 10, Spring 2018
Lecture 6

Census
Visualizations

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Announcements

- Please fill out survey about Tuesday’s guest lecture
  Only 9 responses □
- Another guest lecture after this, nextdoor
- HW2 due Sunday
- My office hours: tomorrow, 1:30-3:30pm, CSE 4204

Summary of Manipulating Rows

- `t.sort(column)`
  - sorts the rows in increasing order
- `t.take(row_numbers)`
  - keeps only specified rows (row numbers start 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

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that produces the same table without a row for `Cheeseburger.`

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that produces the same table without a row for `Cheeseburger.`

```
menu.where(‘Item’, are.not_equal_to(‘Cheeseburger’))
```

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Which line of code finds the number of items on the menu at this restaurant?

A. `menu.num_rows`
B. `menu.column(0).num_rows`
C. `menu.column(0).length`
D. `menu.column(1).size`
E. `More than one of the above`

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that evaluates to

a) the name of a menu item that has the lowest possible price.

b) **Challenge**: a table containing the name of **all** menu items that have the lowest possible price.

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that evaluates to

a) the name of a menu item that has the lowest possible price.

```
menu.sort(‘Price’).column(0).item(0)
```

b) **Challenge**: a table containing the names of **all** menu items that have the lowest possible price.

Practice

The table `menu` has a row for each item on a restaurant’s menu. The columns are `Item` and `Price`, in that order. One of the menu items is `Cheeseburger.`

Write one line of code that evaluates to

a) the name of a menu item that has the lowest possible price.

```
menu.sort(‘Price’).column(0).item(0)
```

b) **Challenge**: a table containing the names of **all** menu items that have the lowest possible price.

```
menu.sort(‘Price’).where(‘Price’,
menu.sort(‘Price’).column(‘Price).item(0)).select(‘Item’)
```

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.

Why estimate each year?

Why estimate 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 …”

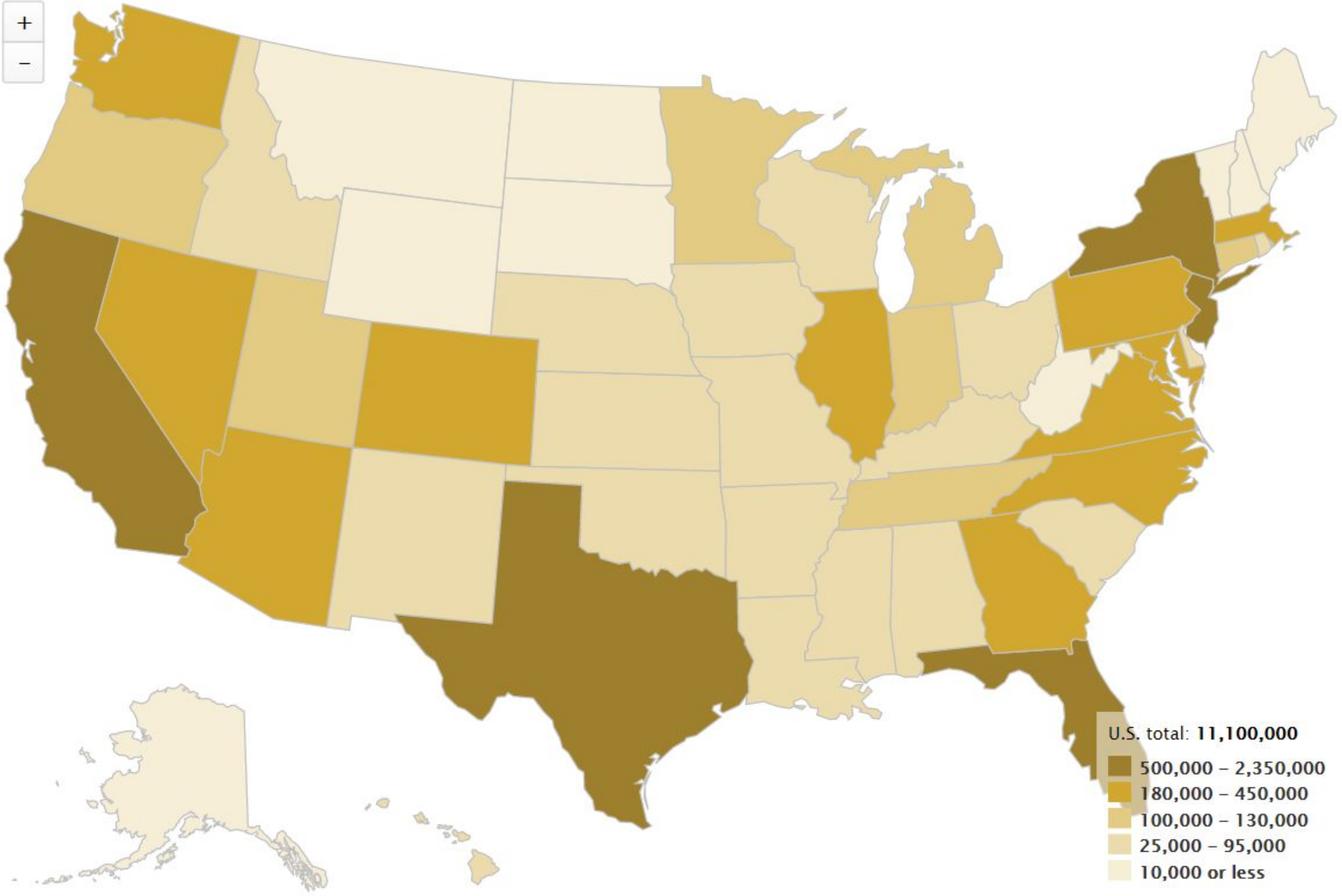
Why estimate each year?

- Article 1, Section 2 of the Constitution:
  - “Representatives and direct Taxes shall be apportioned among the several States … according to their respective Numbers …”

Which of these states would be most likely to want to adjust the census to correct undercount?

A. Hawaii
B. Wyoming
C. Texas
D. Vermont
E. New York

Estimated unauthorized immigrant population, by state, 2014



Pew Research Center

Census Table Description

- Interpretation of values in table depend on columns
  - The SEX column: 1 is *Male*, 2 is *Female*
  - The POPESTIMATE2010 column: *7/1/2010 estimate*

http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Census Table Description

- Interpretation of values in table depend on columns
  - 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

http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Census Table Description

- Interpretation of values in table depend on columns
  - 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

http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Census Table Description

- Interpretation of values in table depend on columns
  - 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)

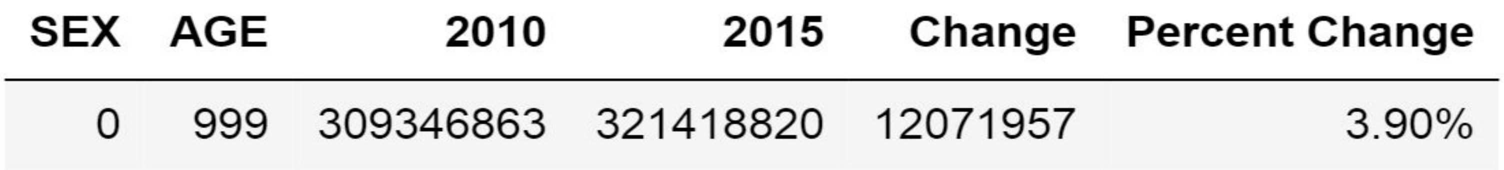
http://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/asrh/nc-est2015-agesex-res.pdf

Analyzing Census Data

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

(Demo)

Discussion Question



SEX	AGE	2010	2015	Change	Percent Change
0	999	309346863	321418820	12071957	3.90%

What does this code calculate?

```
(321418820/309346863) ** (1/5) - 1
```

A. The ratio of the population in 2015 to the population in 2010.
B. The percentage by which the population changed from 2010 to 2015.
C. The annual growth rate for the population from 2010 to 2015.
D. This code doesn’t compute a meaningful value.

(Demo)

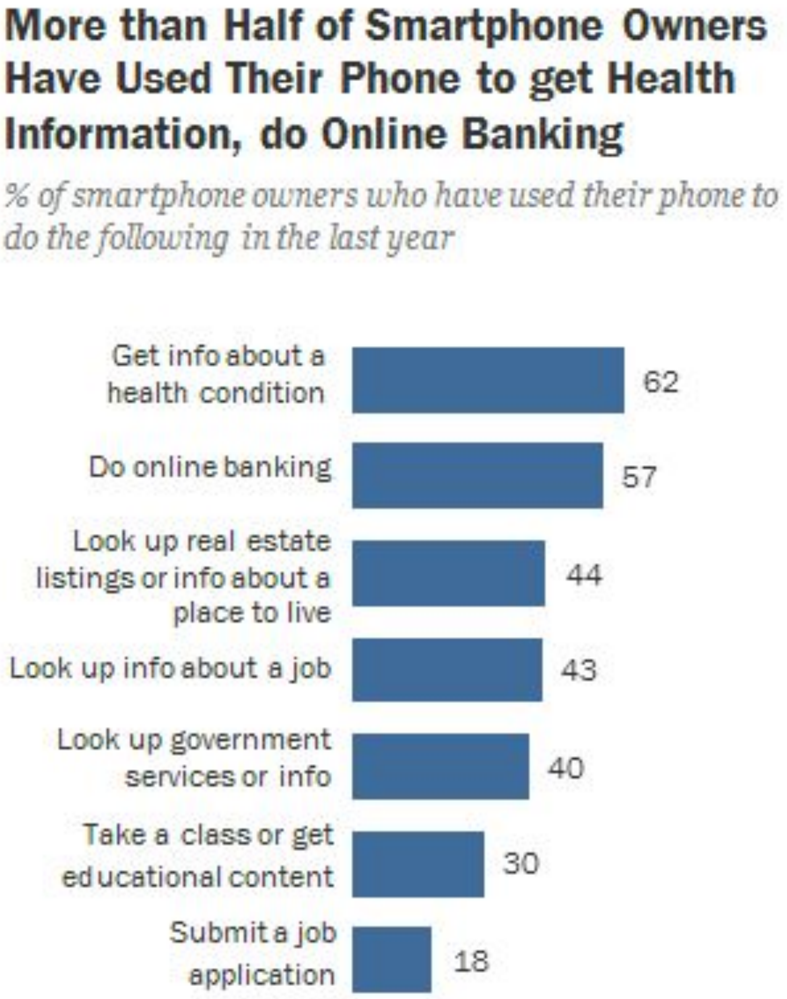
Data Visualization

Discussion Question

Which of the following questions can be answered by this chart?

Among survey responders...

- What proportion did **not** use their phone for online banking?
- What proportion either used their phone for online banking or to look up real estate listings?
- Did everyone use their phone for at least one of these activities?
- Did anyone use their phone for both online banking and real estate?



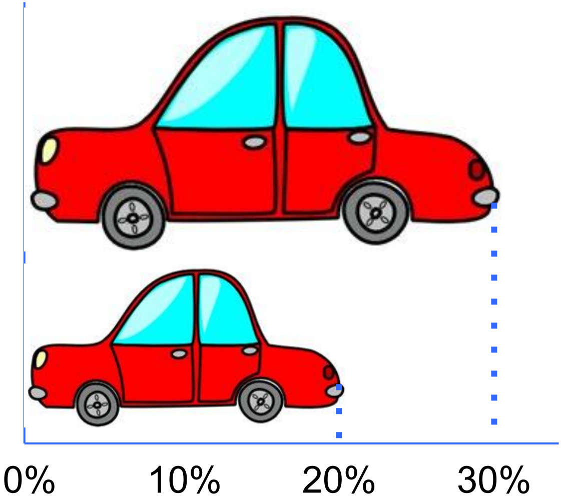
More than Half of Smartphone Owners Have Used Their Phone to get Health Information, do Online Banking

% of smartphone owners who have used their phone to do the following in the last year

Pew research center, 2014

Area Principle

Areas should be proportional to the values they represent



In 2013,

30% of accidental deaths of males were due to automobile accidents

20% of accidental deaths of females were due to automobile accidents

Example from Tian Zheng

Numerical Data

Types of Data

All values in a column should be both the same type **and** be comparable to each other in some way

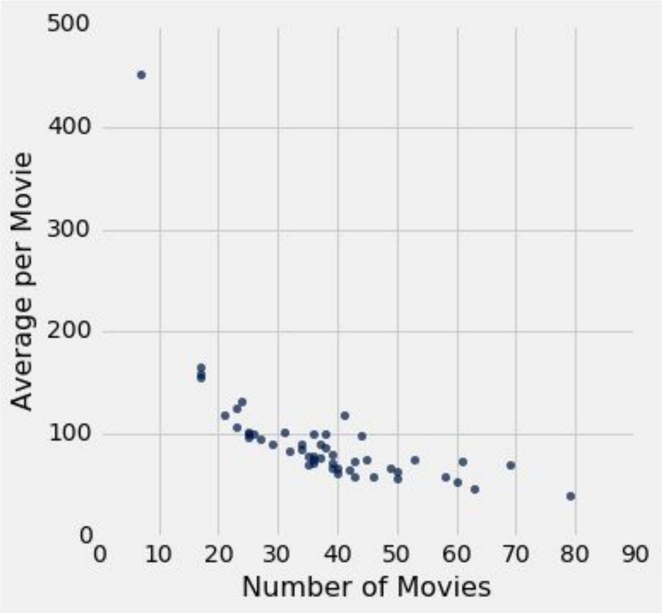
- **Numerical** — Each value is from a fixed scale
  - Numerical measurements are ordered
  - Differences are typically meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories are either the same or different

Terminology

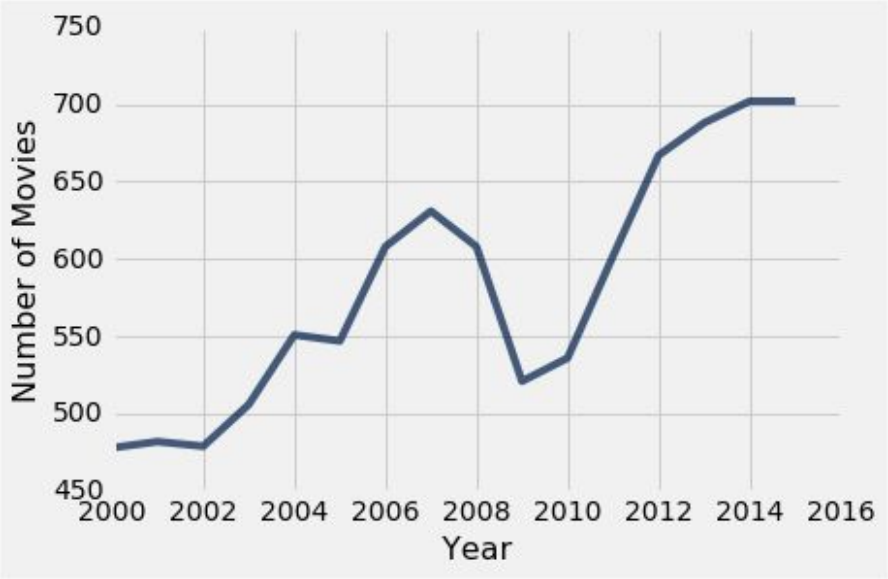
- **Individuals**: those whose features are recorded
- **Variables**: features; these vary across individuals
- Variables have different **values**
- Values can be **numerical**, or **categorical**, or of many other types
- **Distribution**: For each different value of the variable, the frequency of individuals that have that value
- Frequency is measured in counts. Later we will use proportions or percents.

Plotting Two Numerical Variables

Scatter plot: `scatter`



Line graph: `plot`



(Demo)

Categorical Data

Numerical or Categorical?

Just because the values are numbers, doesn’t mean the variable is numerical.

- Census example had numerical `SEX` code (0, 1, and 2).
- Doesn’t make sense to do arithmetic on these “numbers”, e.g. 1 - 0 or (0+1+2)/3 are nonsense here.
- The variable `SEX` is still categorical, even though numbers were used as codes.

Bar Charts

Compare some quantity across categories

- % of smartphone owners who have used their phone for the following in the last year: online banking, job search, etc.
- Gross ticket sales for individual movies

(Demo)