Lecture 5
Census & Charts
Announcements

- Homework 2 due Thursday 2/7
Tables Review
### Table Structure

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Code</th>
<th>Area (m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>CA</td>
<td>163696</td>
</tr>
<tr>
<td>Nevada</td>
<td>NV</td>
<td>110567</td>
</tr>
</tbody>
</table>
Table Methods

- Creating and extending tables:
  - `Table().with_columns` and `Table.read_table`
- Finding the size: `t.num_rows` and `t.num_columns`
- Referring to columns: labels, relabeling, and indices
  - `t.labels` and `t.relabeled`; column indices start at 0
- Accessing data in a column
  - `t.column` takes a label or index and returns an array
- Using array methods to work with data in columns
  - `a.item(row_index)` returns a value in an array
  - `a.sum()`, `a.min()`, `a.max()` or `sum(a)`, `min(a)`, `max(a)`
- 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 for which a column's value equals some particular value
- `t.with_row` makes a new table that has another row

(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 … according to their respective Numbers …”
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)

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?
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

(Demo)
How Do You Generate This Chart?

Top 10 highest grossing movies:

- Star Wars: The Force Awakens
- Avatar
- Titanic
- Jurassic World
- Marvel's The Avengers
- The Dark Knight
- Star Wars: Episode I - The Phantom Menace
- Star Wars
- Avengers: Age of Ultron
- The Dark Knight Rises

How long ago each one was released:
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 numerical scale
  - Numerical measurements are ordered
  - Differences are meaningful
- **Categorical** — Each value is from a fixed inventory
  - May or may not have an ordering
  - Categories are the same or different
“Numerical” Data

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

- Census example had numerical SEX code (0, 1, and 2)
- It doesn’t make sense to perform 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 for the categories
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
Categorical Data
Bar Charts of Counts

Distributions:

- The distribution of a variable (a column) describes the frequency of its different values.
- The **group** method counts the number of rows for each value in a column.

Bar charts can display the distribution of categorical values:

- Proportion of how many US residents are male or female.
- Count of how many top movies were released by each studio.

(Demo)
Categorical Distributions

Bar chart: `barh`

Displays a categorical distribution

(But when the values of the variable have a rank ordering, or fixed sizes relative to each other, more care might be needed.)