Lecture 6 – Histograms and Functions
Announcements

- **Lab02** *(Data Types and Arrays)*
  - Due tomorrow night (Friday 11/06)

- **HW02 - Table Manipulation & Visualization**:
  - Due Monday (11/09)

- **Checkpoint/Project 1**:  
  - Paired assignment that covers the previous section of the course material  
  - Released Wednesday (11/11) and due Wednesday (11/18)
Course Outline

- **Exploration**  
  Week 1 - 3  
  - Discover patterns in data  
  - Articulate insights (visualizations)

- **Inference**  
  Week 3 - 5  
  - Make reliable conclusions about the world  
  - Statistics is useful

- **Prediction**  
  Week 6-7  
  - Informed guesses about unseen data
3 checkpoints/projects:

- Flexible grading scheme – weight based on scores
  - Best performing one will count as “midterm”
  - Remaining two will count as 20% in “project” from first lecture

Exploration checkpoint/project

- Released 11/11 (maybe earlier)
- Due Wednesday 11/18
- HW3 & HW4 are on the shorter side
Visualization Review

- Line plots
  - Sequential data

- Scatter plots
  - Finding associations

- Bar plots
  - Categorical distributions
Bar Plots

- Display relationship between categorical variable and a numerical variable
- Display a categorical distribution
```python
import pandas as pd

top_movies = pd.read_csv('top_movies_2017.csv')
top_movies
```

<table>
<thead>
<tr>
<th>Title</th>
<th>Studio</th>
<th>Gross</th>
<th>Gross (Adjusted)</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gone with the Wind</td>
<td>MGM</td>
<td>198676459</td>
<td>1796176700</td>
<td>1939</td>
</tr>
<tr>
<td>Star Wars</td>
<td>Fox</td>
<td>460998007</td>
<td>1583483200</td>
<td>1977</td>
</tr>
<tr>
<td>The Sound of Music</td>
<td>Fox</td>
<td>158671368</td>
<td>1266072700</td>
<td>1965</td>
</tr>
<tr>
<td>E.T.: The Extra-Terrestrial</td>
<td>Universal</td>
<td>435110554</td>
<td>1261085000</td>
<td>1982</td>
</tr>
<tr>
<td>Titanic</td>
<td>Paramount</td>
<td>658672302</td>
<td>1204368000</td>
<td>1997</td>
</tr>
<tr>
<td>The Ten Commandments</td>
<td>Paramount</td>
<td>65500000</td>
<td>1164590000</td>
<td>1956</td>
</tr>
<tr>
<td>Jaws</td>
<td>Universal</td>
<td>260000000</td>
<td>1138620700</td>
<td>1975</td>
</tr>
<tr>
<td>Doctor Zhivago</td>
<td>MGM</td>
<td>111721910</td>
<td>1103564200</td>
<td>1965</td>
</tr>
<tr>
<td>The Exorcist</td>
<td>Warner Bros</td>
<td>232906145</td>
<td>983226600</td>
<td>1973</td>
</tr>
<tr>
<td>Snow White and the Seven Dwarves</td>
<td>Disney</td>
<td>184925486</td>
<td>969010000</td>
<td>1937</td>
</tr>
</tbody>
</table>
Bar plot

```python
: top10_adjusted.barh('Title', 'Millions')
```
Displaying a Categorical Distribution

- Distribution of a variable describes the frequencies of the values.

- The **group** method counts the number of values in the column.

- Bar chart displays the distribution of categorical variable:
  - One bar per category/value
  - Length of bar is the count of individuals in that category.
Areas should be proportional to values they represent

- If you represent 20% by
- 40% should be represented by
- and not by
Area Principle – 2016 Election Map
Area Principle – 2016 Election Map

https://www.wired.com/story/is-us--leaning-red-or-blue-election-maps/
Binning converts a numerical distribution to a categorical distribution.

Binning counts the number of numerical values that lie within a range, aka a bin.

Bins contain:
- A lower bound (inclusive)
- An upper bound (exclusive)
- Bins contain:
  - A lower bound (inclusive)
  - An upper bound (exclusive)

188, 170, 189, 163, 183, 171, 185, 168, 173, ...
Histogram

Chart that displays the distribution of a numerical variable

Uses bins; there is one bar corresponding to each bin

Uses the area principle:

• The area of each bar is the percent of individuals in the corresponding bin
Understanding Histograms

- Axes
- Height
- Area
By default, `hist` uses a scale (`normed=True`) that ensures the area of the chart sums to 100%.

The area of each bar is a percentage of the whole.

The horizontal (x-) axis is a number line (e.g., years), and the bins sizes don’t have to be equal to each other.

The vertical axis is a rate (e.g., percent per year).
Histogram Height (of a bin)

\[
\text{Height} = \frac{\% \text{ in bin}}{\text{width of bin}}
\]

- Height measures density
- the percent of data in the bin relative to the amount of space in the bin
- Units: percent per unit on the horizontal axis
Histogram Area (of a bar)

- Area tells us what percent of all data is in a bin

- Area of a bar = Height times width of a bin
  - “How many individuals in the bin?” Use area.
  - “How crowded is the bin?” Use height
Bar Chart or Histogram?

**Bar Chart**
- Distribution of categorical variable
- Bars have arbitrary (but equal) widths and spacings
- **height (or length)** and area of bars proportional to the percent of individuals

**Histogram**
- Distribution of numerical variable
- Horizontal axis is numerical: to scale, no gaps, bins can be unequal
- **Area** of bars proportional to the percent of individuals; **height** measures density
Anatomy of a Function

- Name
- Parameters / Argument Names
- Body
- Return Expression
```python
def sread(values):
    spread_val = max(values) - min(values)
    return spread_val
```
Example Function

```python
def sread(values):
    spread_val = max(values) - min(values)
    return spread_val
```
What does this function do?

```python
def f(s):
    return np.round(s / sum(s) * 100, 2)
```

- What kind of input does it take?
- What output will it give?
- What's a reasonable name?