Lecture 5 – Visualizations

Tuesday 02/01/22
Announcements

- HW01 released tonight (due Monday 02/07)
- Lab02 (Data Types and Arrays) due (Monday 02/07)
- No class Tuesday 02/08
- t.sort(column) sorts rows in increasing order
- t.sort(column, descending=True) sorts 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, value) keeps all rows where a column’s value equals some particular value
  - Equivalent as t.where(column, are.equal_to(value))
Types of Attributes

All values in a column of a table 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 difference
Census Data
Every ten years, Census Bureau counts how many people there are in the U.S.

Census Bureau estimates how many people are in US during the other 9 years

U.S. Constitution Article 1, Section 2:
- “Representatives and direct Taxes shall be apportioned among the several States ... according to their respective Numbers ...”
- https://www2.census.gov/programs-surveys/popest/datasets/

- https://www2.census.gov/programs-surveys/popest/datasets/2010-2015/national/totals/

- demo
Plotting Numerical data

Line graph plot

Scatter plot scatter
Which is the x-axis?
- Year

Which is the y-axis?
- Number of Movies
Use **line plots** for sequential data if
- x-axis has an order
- sequential differences in y values are meaningful
- there’s only one y-value for each x-value
- usually: x-axis is time or distance

Use **scatter plots**
- when looking for associations
Bar Plots

- Display relationship between categorical variable and a numerical variable
- Display a categorical distribution
```python
import Table

top_movies = Table.read_table('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 Brothers</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>
```
: 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.