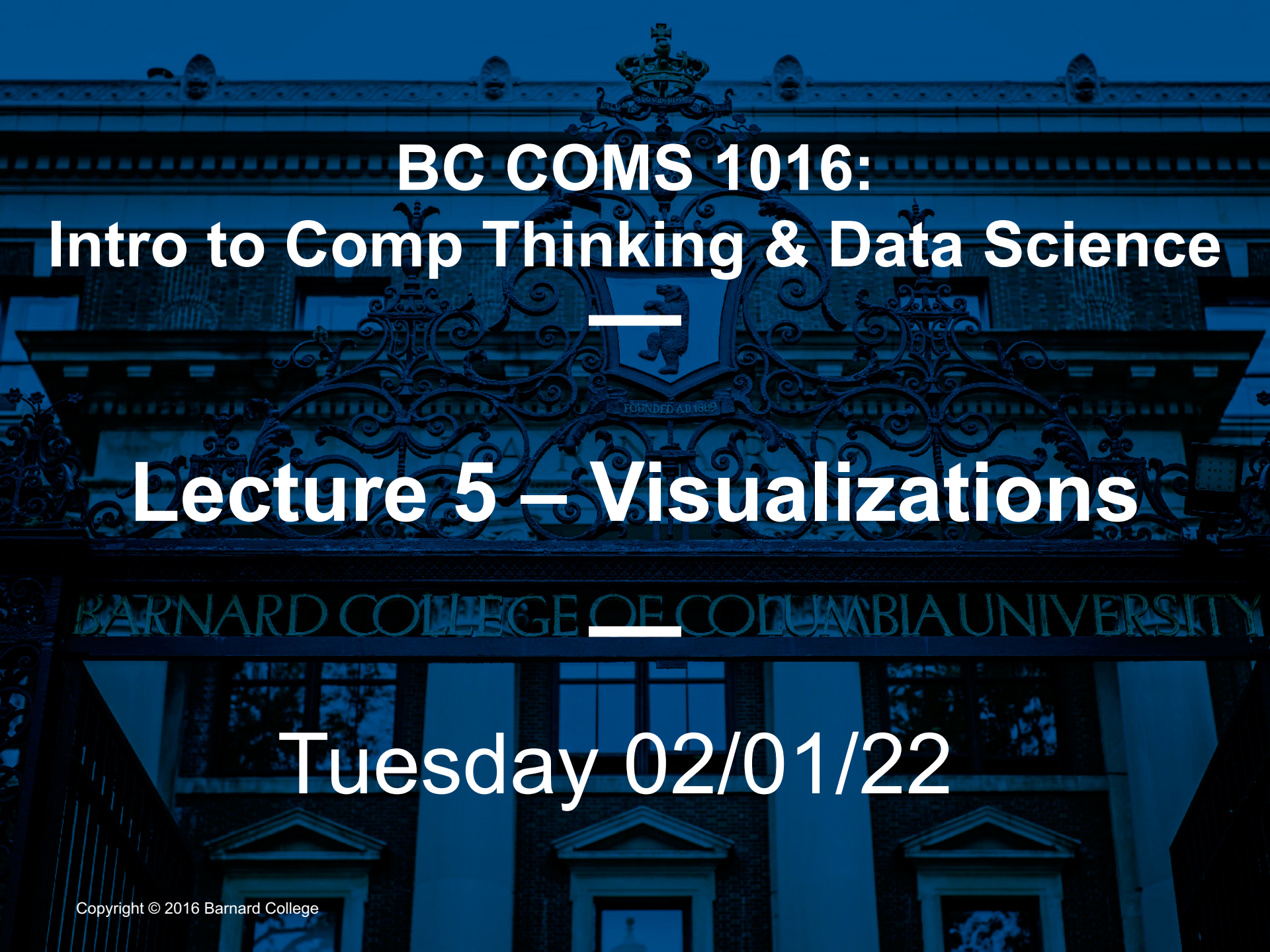


BC COMS 1016:
Intro to Comp Thinking & Data Science

—

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

Table Review

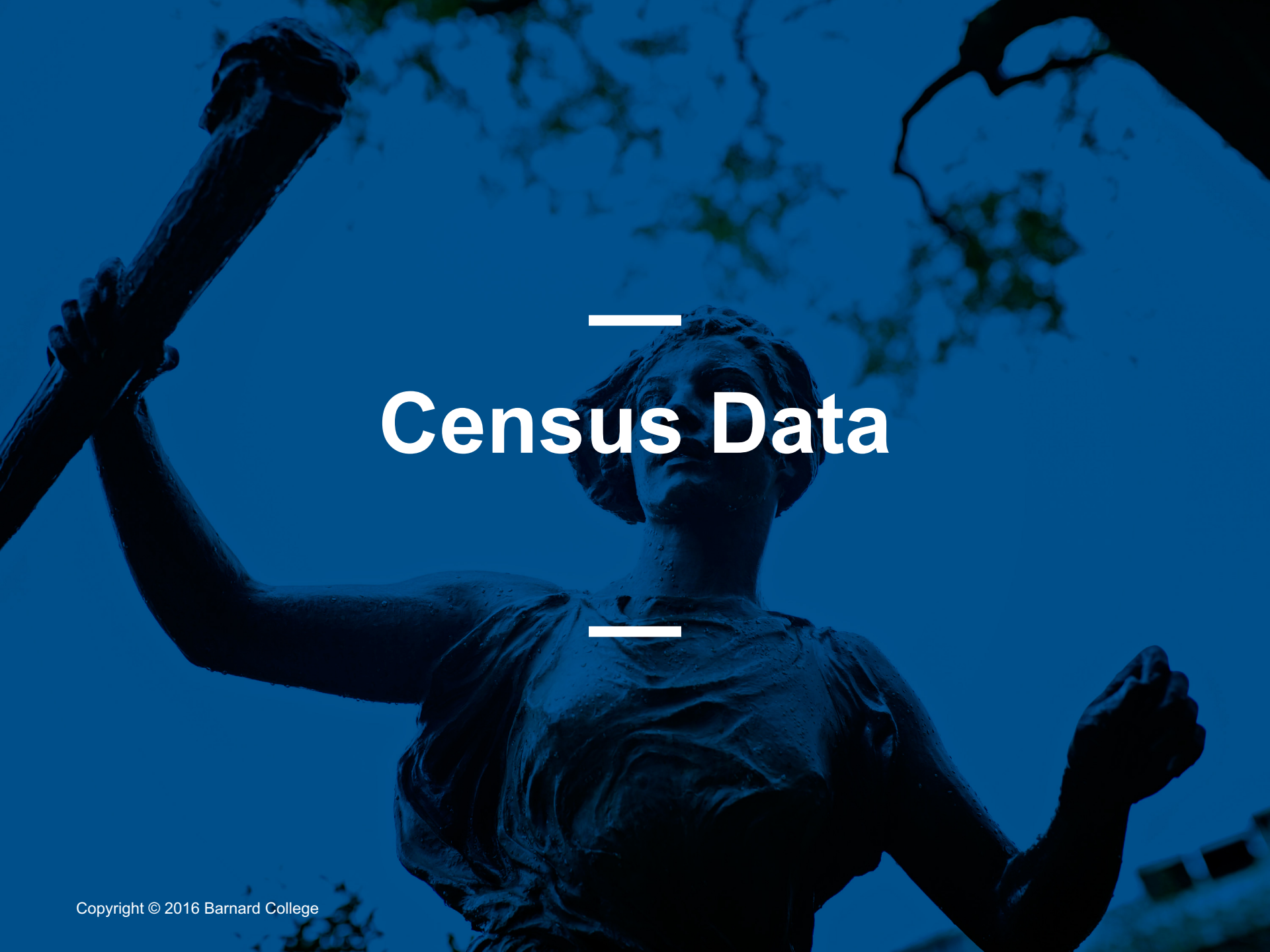
Table Review

- 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

The Decennial Census

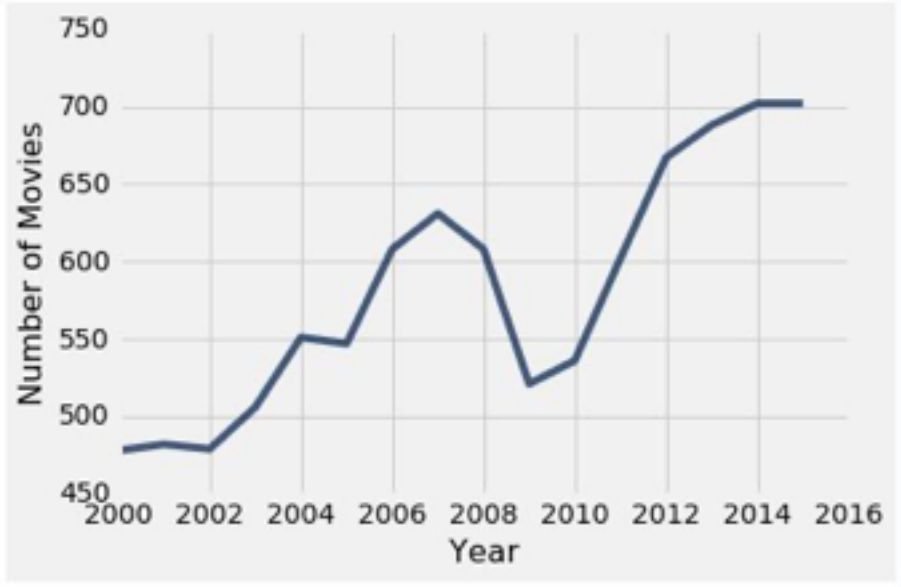
- Every ten years, Census Bureau counts how many people there are in the U.S.
- Census Bureau estimates how many people are in US during the other 9 years
- U.S. Constitution Article 1, Section 2:
  - “Representatives and direct Taxes shall be apportioned among the several States ... according to their respective Numbers ...”

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

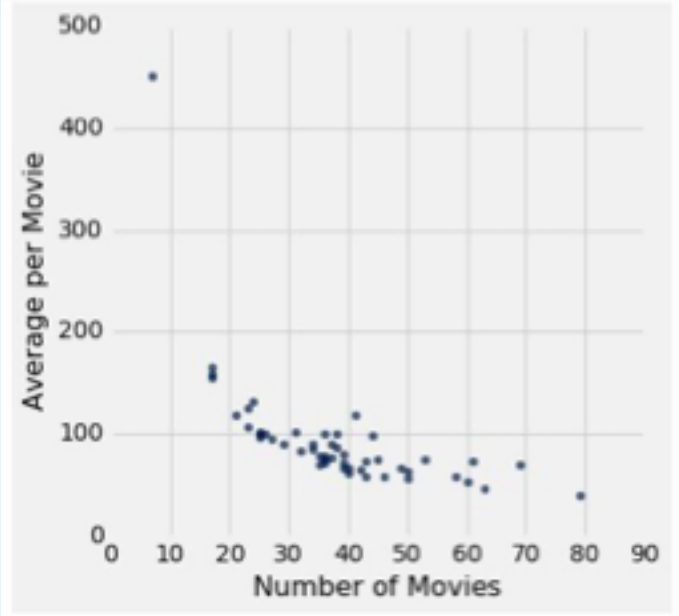
Graphs

Plotting Numerical data

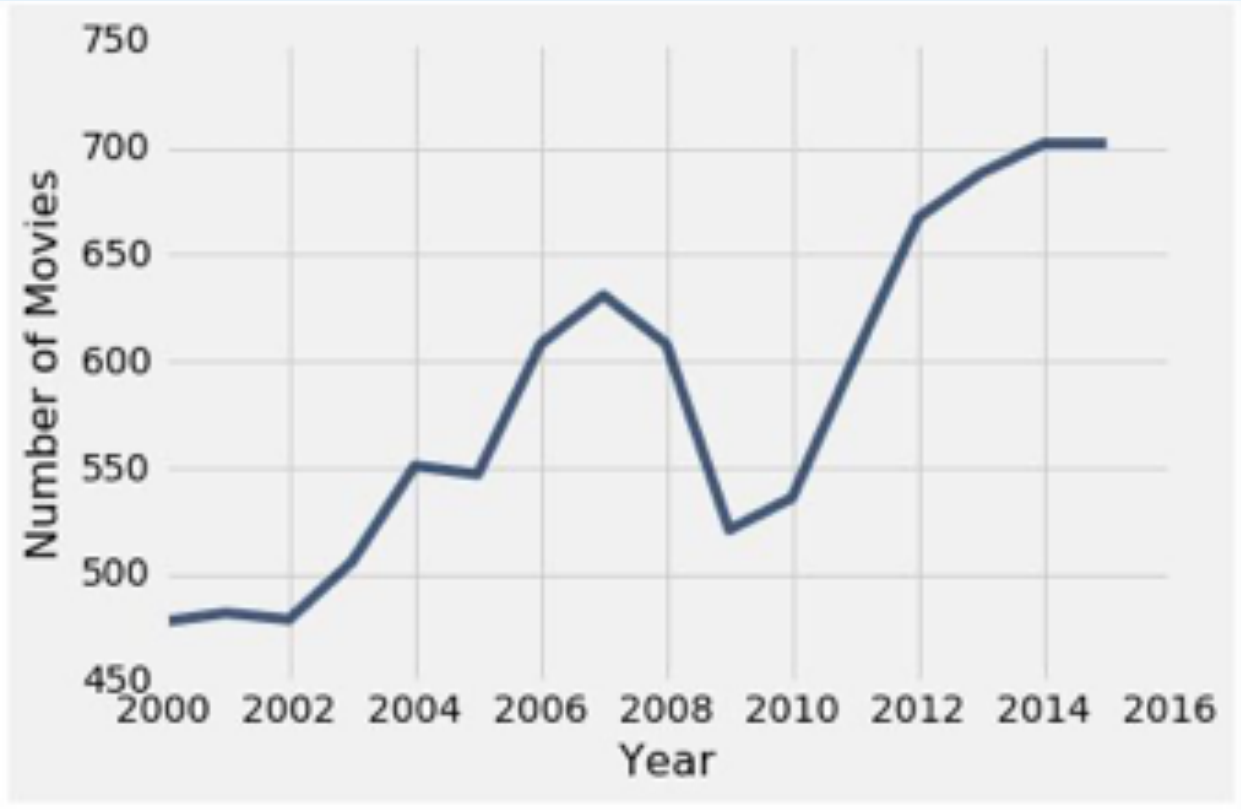
Line graph plot



Scatter plot scatter



x-axis and y-axis



Which is the x-axis?

- Year

Which is the y-axis?

- Number of Movies

Line vs Scatter plot: When to use which?

- 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

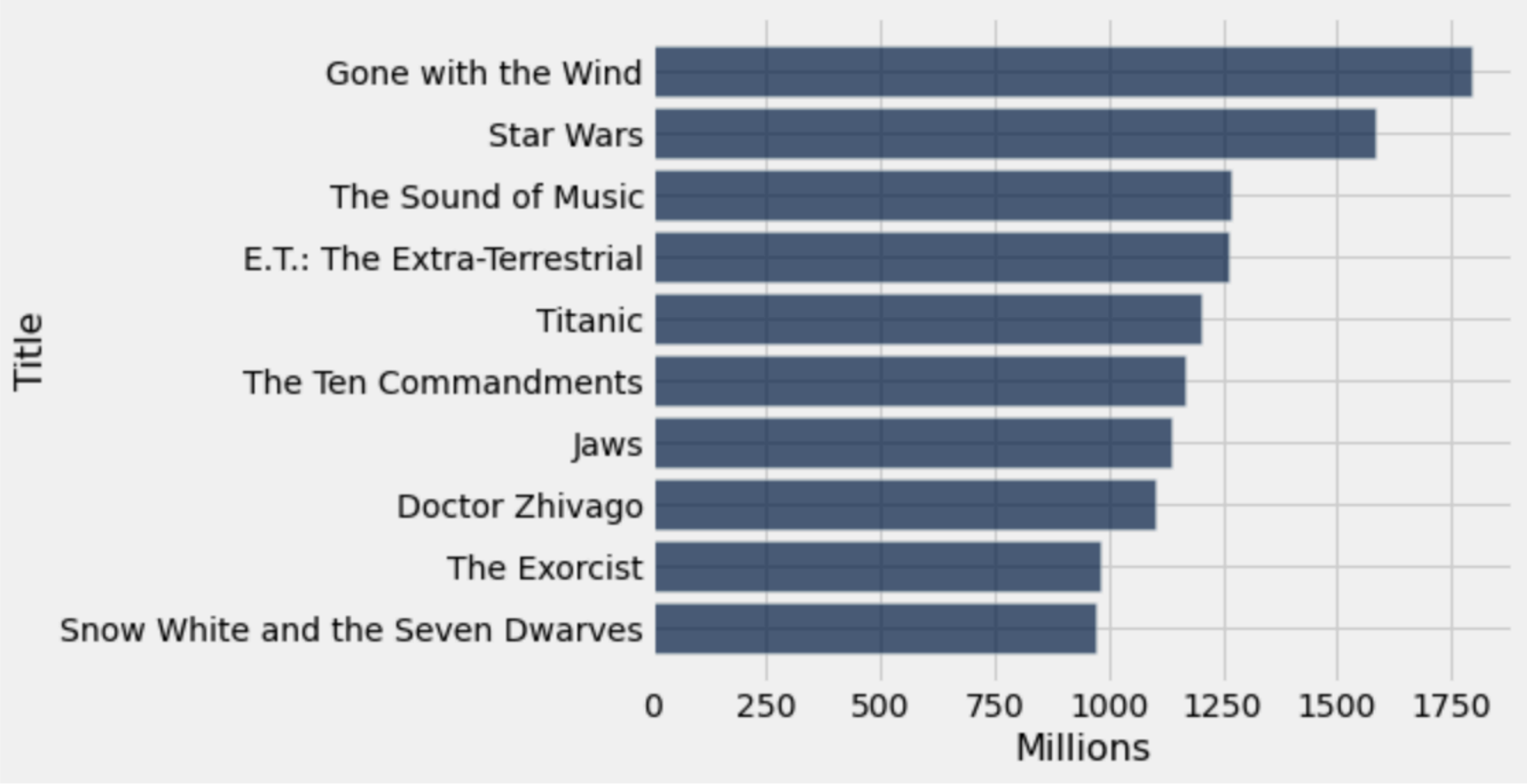
Bar Charts

```
top_movies = Table.read_table('top_movies_2017.csv')
top_movies
```

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

Bar plot

```
: 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