1885

DATA 1202

Spring 2024

Lecture 6

Charts

Announcements

- **HW 3** due Wednesday (1/31) at 11pm
- **Lab 4** is due Friday at 5pm

Weekly Goals

- Monday (**Today)**
    - Attribute Types
    - Visualizing data: Relationships
    - Distributions
- Wednesday
    - Visualizing Data: Histograms
    - Height as Density

Table Review

Table Methods

- Creating and extending tables:
  - `Table.read_table` and `Table().with_columns`
- Finding the size: `num_rows` and `num_columns`
- Referring to columns: by labels or indices
  - column indices start at 0
- Accessing data in a column
  - `column` takes a label or index and **returns an array**
- Using array methods to work with data in columns
  - `item`, `sum`, `min`, `max`, and so on
- Creating new tables containing some of the original columns:
  - `select`, `drop`

Manipulating Rows

- `t.sort(column, descending=True)` sorts the 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, are.equal_to(value))` keeps all rows for which a column's value equals some particular value
  - Shorter form: `t.where(column, value)`

Discussion Questions

The table `nba` has columns `PLAYER`, `POSITION`, and `SALARY`.

a) Create an array containing the names of all point guards (`PG`) who made more than $15M

```
guards = nba.where('POSITION', 'PG')
guards.where('SALARY',
             are.above(15000000)).column('PLAYER')
```

b) After evaluating these two expressions in order, what's the result of the second one?

```
nba.drop('POSITION')
nba.num_columns
```

(Demo)

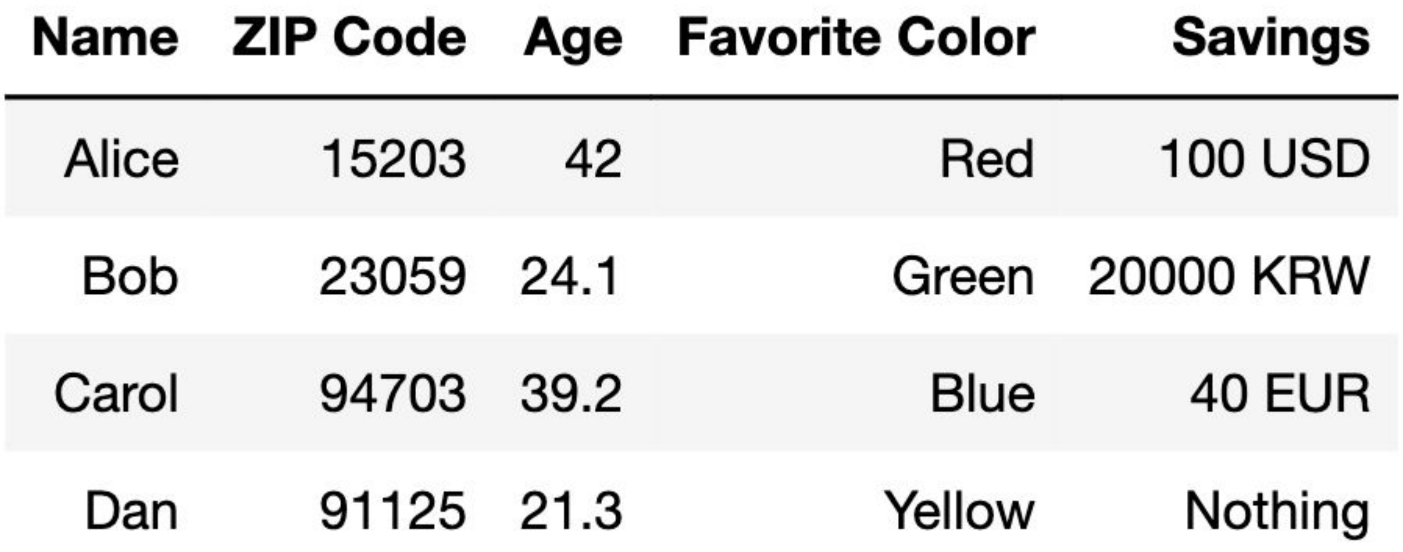
Attribute Types

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
    - Examples of ordered categorical values?
  - Categories are the same or different

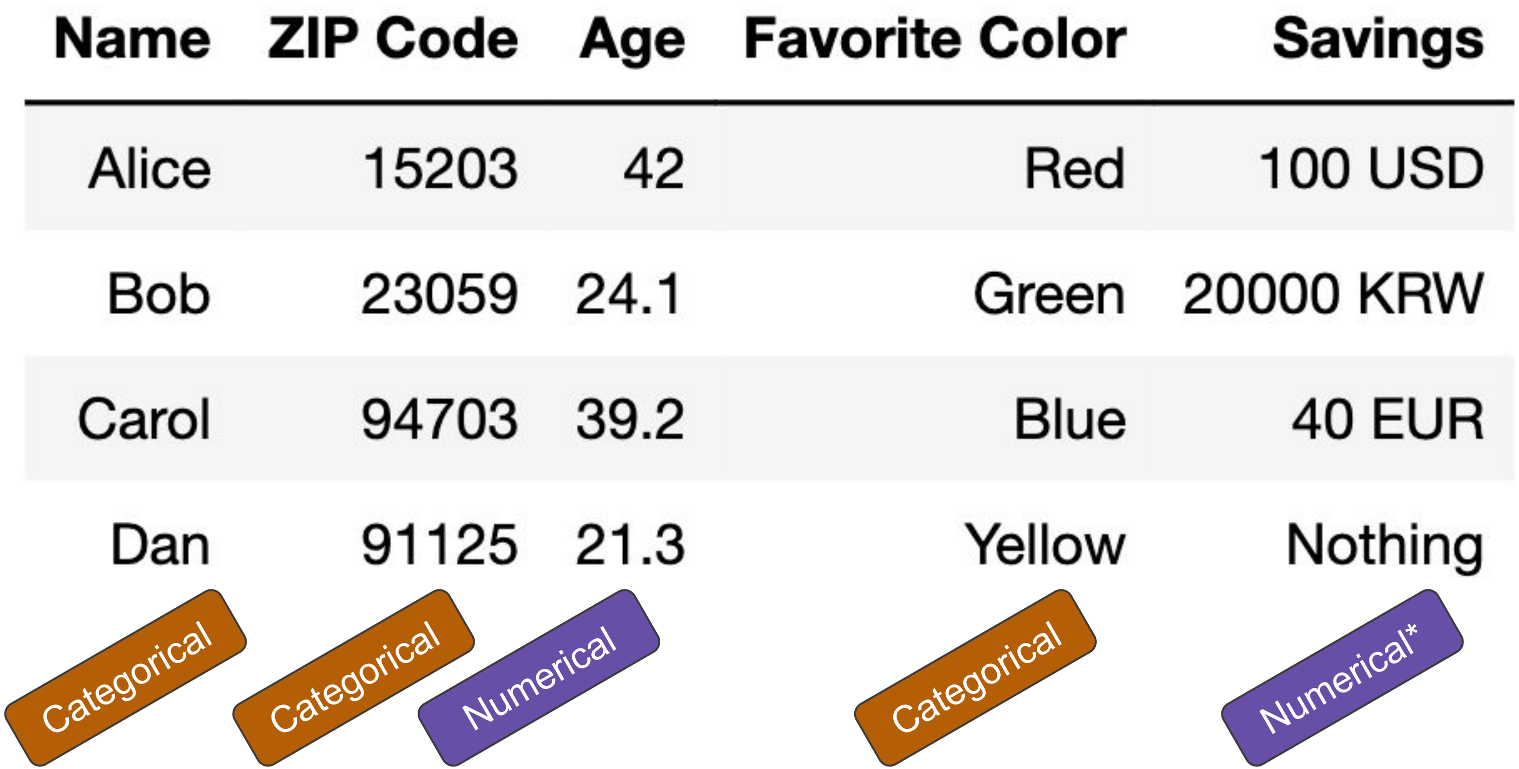
Attribute Types ≠ Python Types



Name	ZIP Code	Age	Favorite Color	Savings
Alice	15203	42	Red	100 USD
Bob	23059	24.1	Green	20000 KRW
Carol	94703	39.2	Blue	40 EUR
Dan	91125	21.3	Yellow	Nothing
str	int	float	str	str

The Python type doesn’t fully convey the meaning of the data.
In this class, we will talk about Attribute Types to describe the “kind of data”.

Attribute Types ≠ Python Types



Name	ZIP Code	Age	Favorite Color	Savings
Alice	15203	42	Red	100 USD
Bob	23059	24.1	Green	20000 KRW
Carol	94703	39.2	Blue	40 EUR
Dan	91125	21.3	Yellow	Nothing
Categorical	Categorical	Numerical	Categorical	Numerical*

“Numerical” Attributes

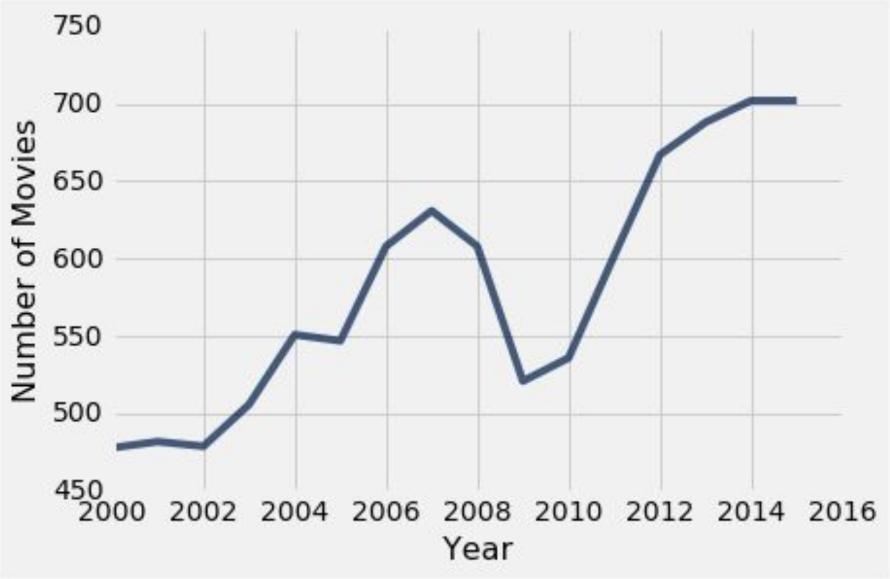
Just because the values are numbers, doesn’t mean the attribute is numerical

- Census example has numerical `SEX` code (0, 1, and 2)
- It doesn’t make sense to perform arithmetic on these “numbers”, e.g. (0+1+2)/3 is meaningless
- The attribute `SEX` is still categorical, even though numbers were used for the categories

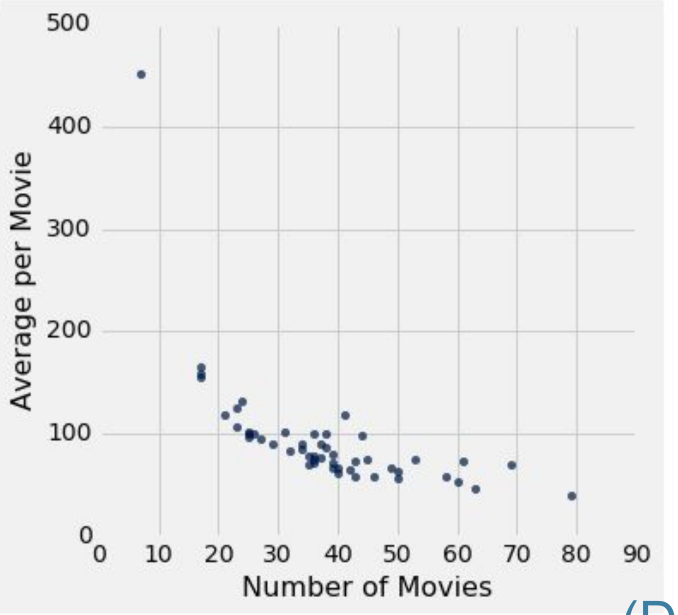
Numerical Data

Plotting Two Numerical Variables

Line plot: `plot`

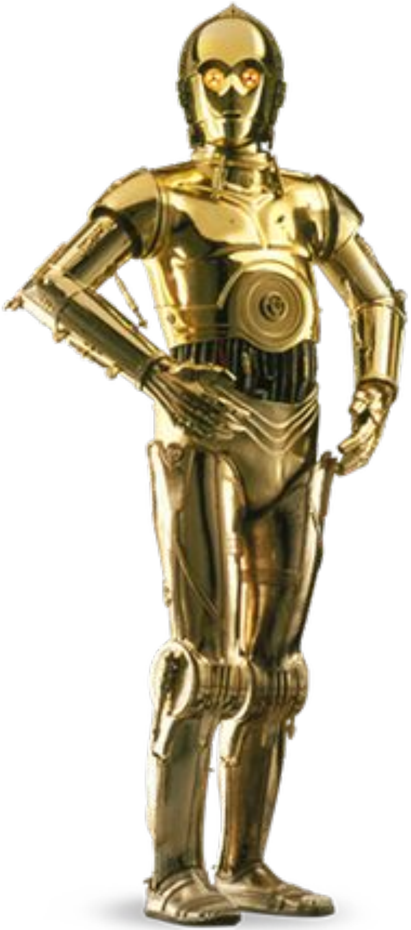


Scatter plot : `scatter`



(Demo)

Anthony Daniels, actor



https://en.wikipedia.org/wiki/C-3PO

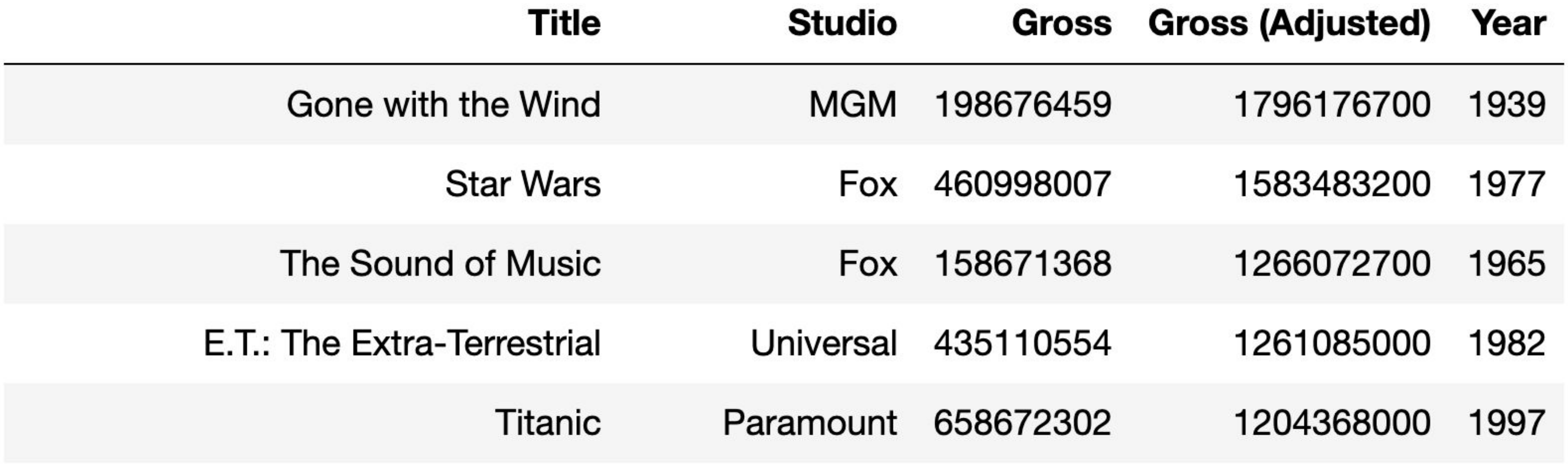
Line vs Scatter Plot

- `t.plot(x_label, y_label)`
- `t.scatter(x_label, y_label)`

- Use line plots for sequential quantitative data: if...
  - ...your x-axis has an order
  - ...sequential differences in y values are meaningful
  - ...there’s only one y-value for each x-value
  - Often: x-axis is **time** or **distance**
- Use scatter plots for non-sequential quantitative data
  - If you are looking for associations

Categorical and Numerical Variables

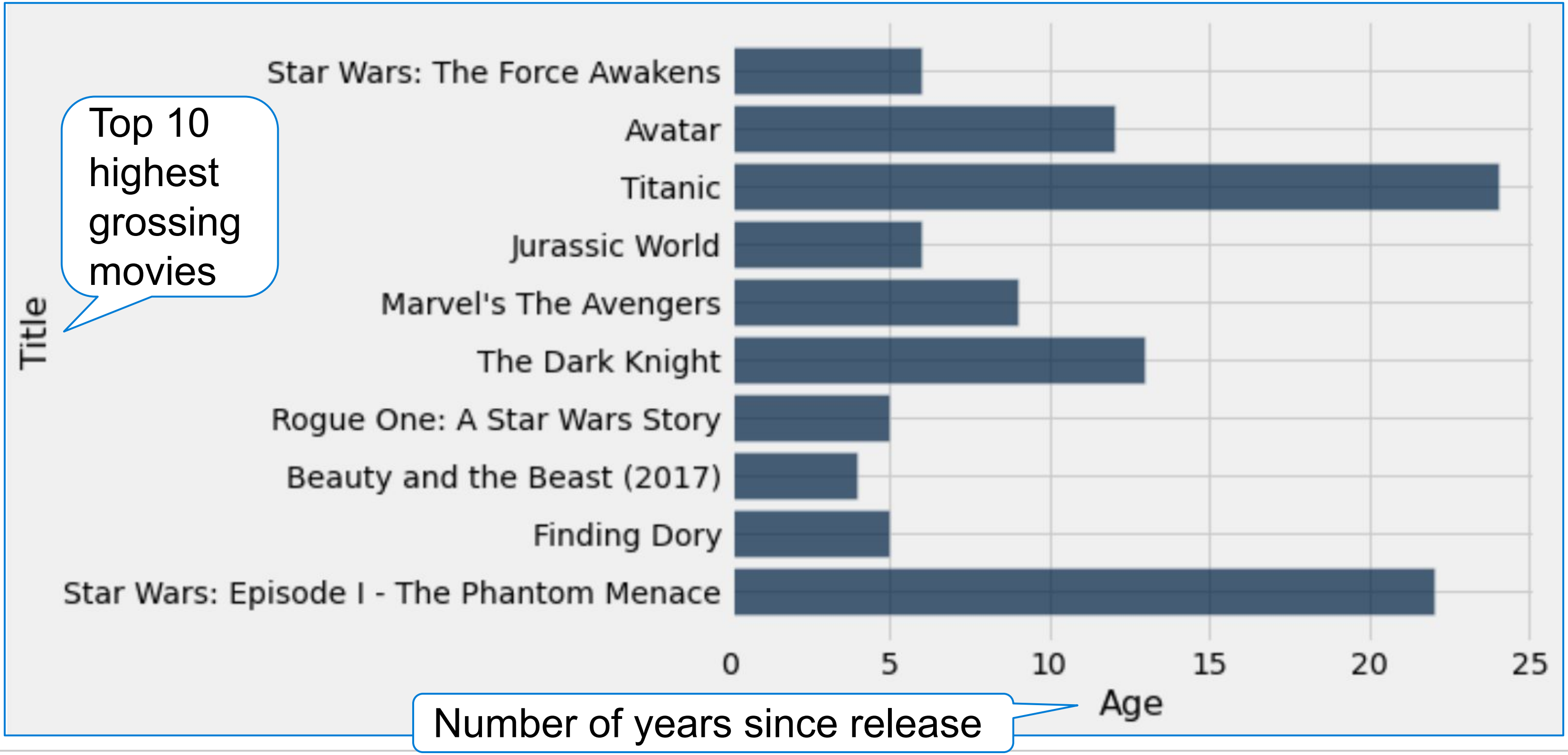
Highest Grossing Movies as of 2017



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

(Demo)

How Do You Generate This Chart?

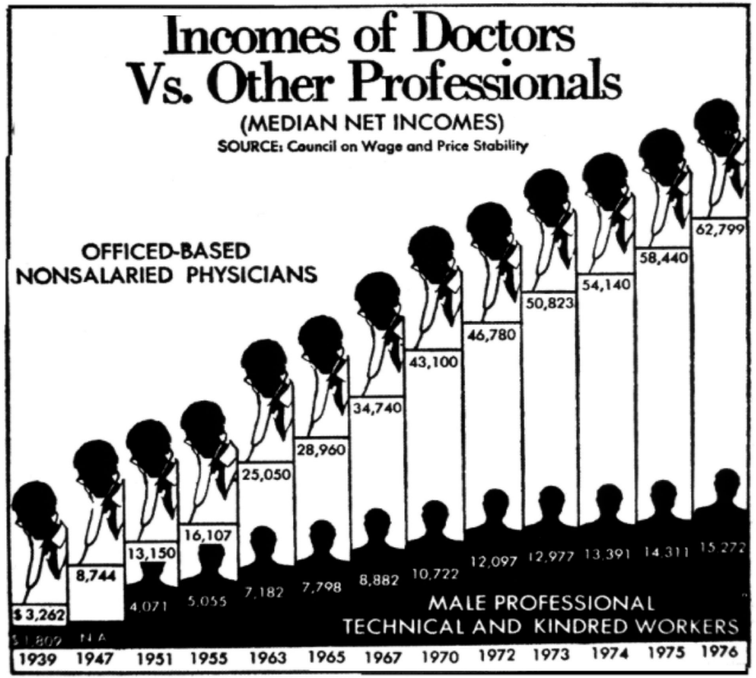


Top 10 highest grossing movies

Number of years since release

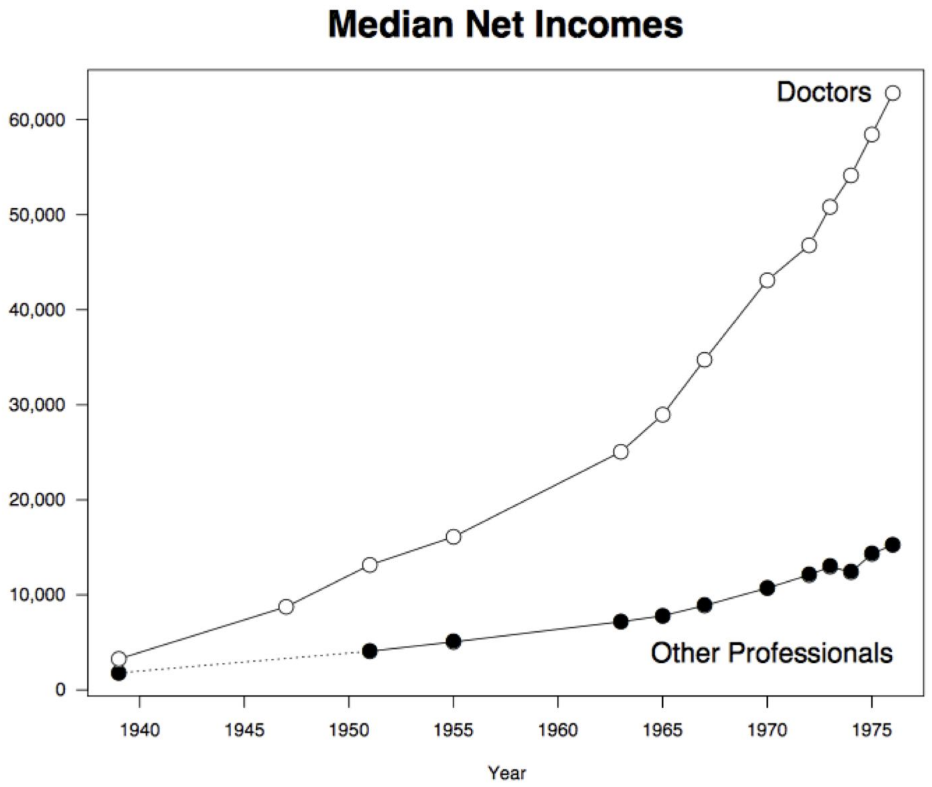
Visualization Fundamentals

Don’t Do This



Source: Washington Post

Do This Instead



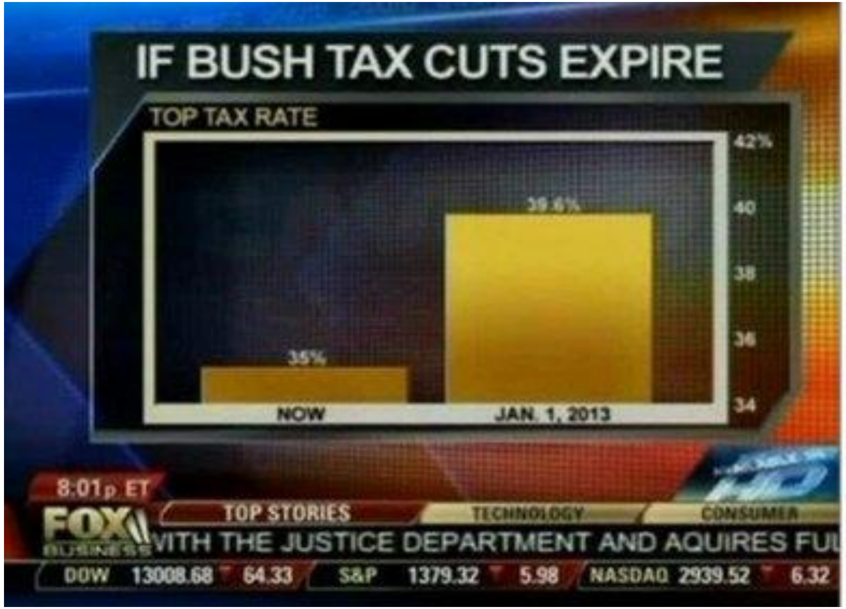
Source: Ross Ihaka

Good Practices

- Less can be more
  - Minimize decoration
  - Choose colors carefully
    - Minimize the number of different colors
- If data are numerical, preserve their relative values and distances between them

See Edward Tufte’s “The Visual Display of Quantitative Information”

Importance of the Y-Axis


IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
WITH THE JUSTICE DEPARTMENT AND AQUIRES FUL
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32