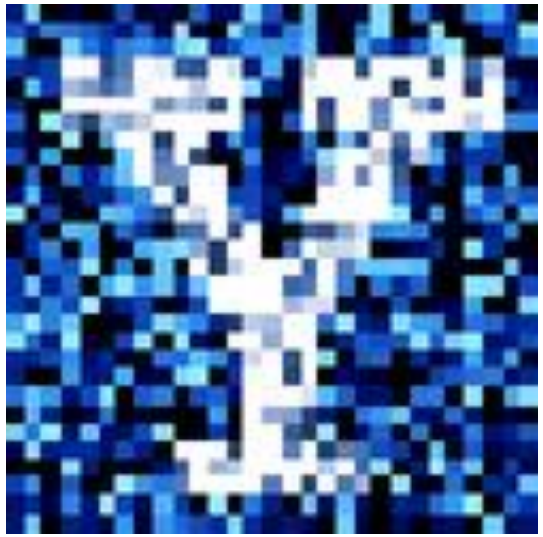
YData: Introduction to Data Science



Class 05: Descriptive statistics and plots continued

Overview

Very quick review of categorical data

Statistics and data visualizations for quantitative data continued:

- Measures of central tendency: mean median
- Measures of spread (standard deviation)
- z-scores
- Scatter plots and correlation

Announcement: Homework 2

Homework 2 has been posted!

```
import YData
YData.download_homework(2)
```

It is due on Gradescope on Sunday September 14th at 11pm

- **Be sure to mark each question on Gradescope!**

Notes:

- There is an ~18 page reading from the book "Data and the American Dream" that you need to do, so I recommend you get started on this soon.

Very quick review

Last class we spoke about comparisons and Booleans, and more string methods

- 3 < 5 # TRUE
- "123".isnumeric()
- f"The number {my_num} in a string"

We discussed structure data

Categorical Variable

Quantitative Variable

title	clean_test	binary	budget	domgross	budget_2013	domgross_2013
21 & Over	notalk	FAIL	13000000	25682380.0	13000000	25682380.0
Dredd 3D	ok	PASS	45000000	13414714.0	45658735	13611086.0
12 Years a Slave	notalk	FAIL	20000000	53107035.0	20000000	53107035.0
2 Guns	notalk	FAIL	61000000	75612460.0	61000000	75612460.0
42	men	FAIL	40000000	95020213.0	40000000	95020213.0

Categorical data main statistic:

- Proportion = $\frac{\text{number in category}}{\text{total number}}$

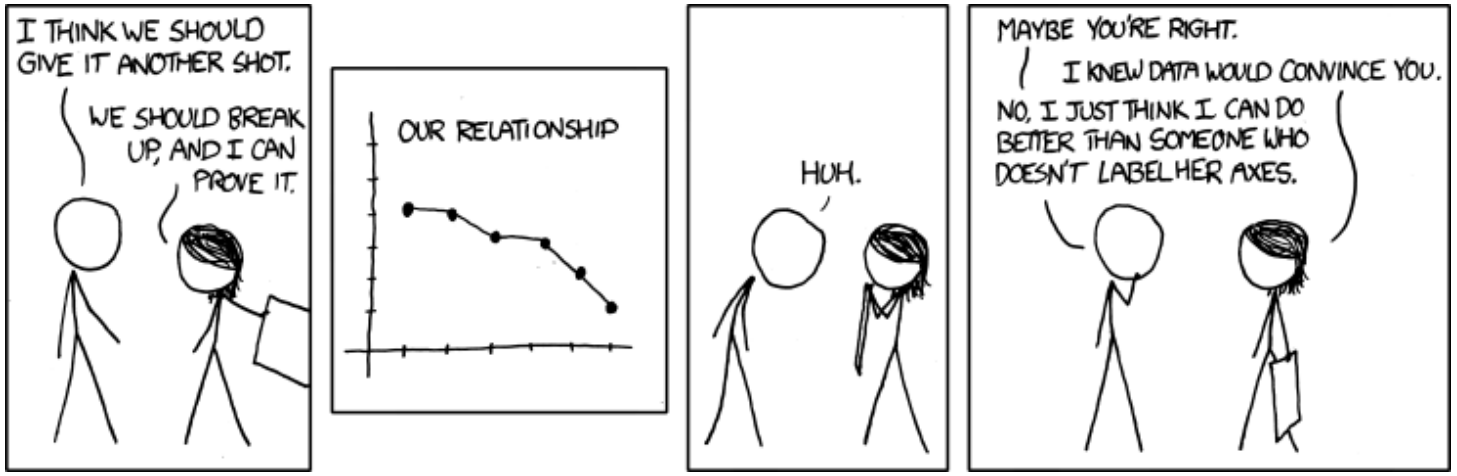
bechdel.count("PASS")/len(bechdel)

Very quick review

Visualize categorical data

```
import matplotlib.pyplot as plt
plt.bar(labels, data)
plt.pie(data)
```

If you don't want exes, label you axes!



Quantitative data

To explore quantitative data, let's look at how much revenue each movie made in the United States in (2013) inflation adjusted dollars

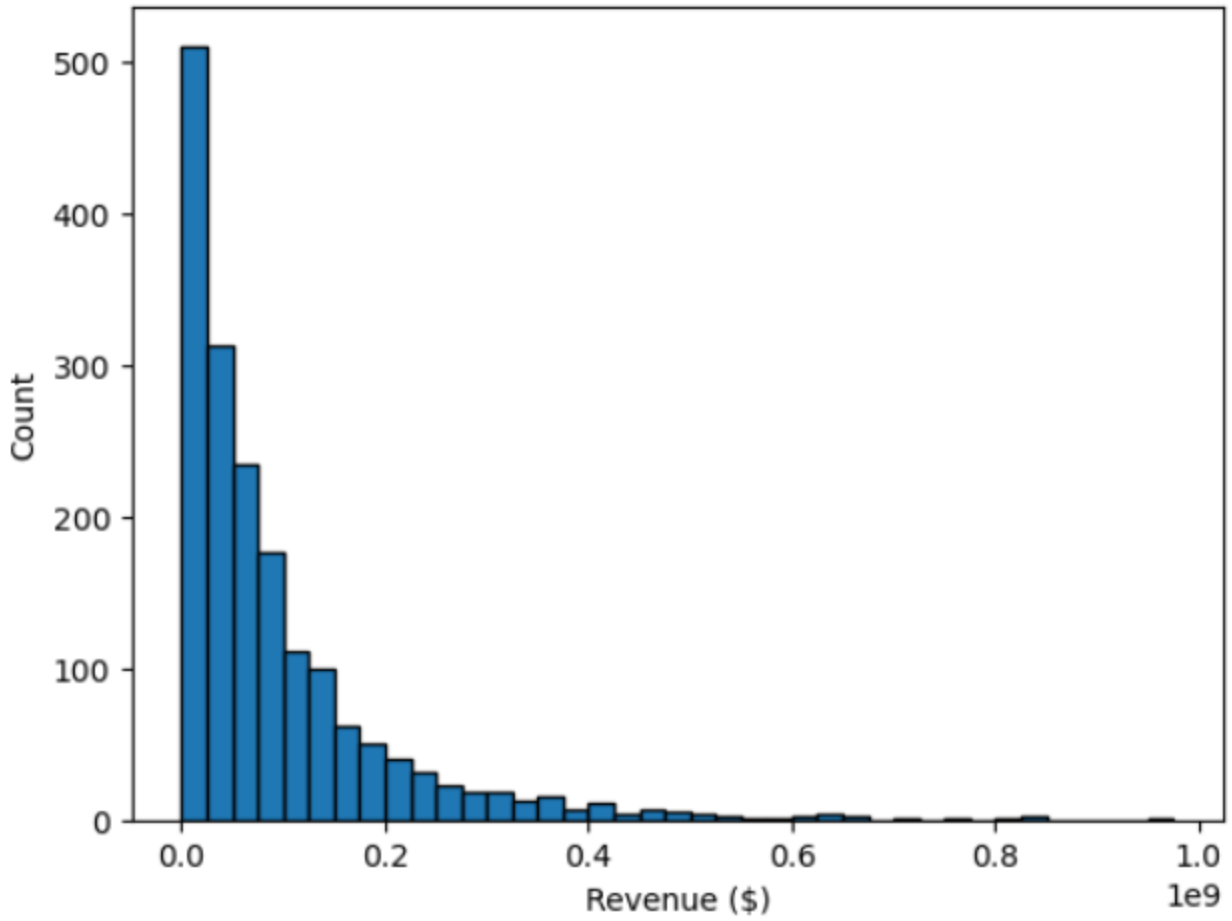
- domgross_2013

Quantitative Variable

title	clean_test	binary	budget	domgross	budget_2013	domgross_2013
21 & Over	notalk	FAIL	13000000	25682380.0	13000000	25682380.0
Dredd 3D	ok	PASS	45000000	13414714.0	45658735	13611086.0
12 Years a Slave	notalk	FAIL	20000000	53107035.0	20000000	53107035.0
2 Guns	notalk	FAIL	61000000	75612460.0	61000000	75612460.0
42	men	FAIL	40000000	95020213.0	40000000	95020213.0

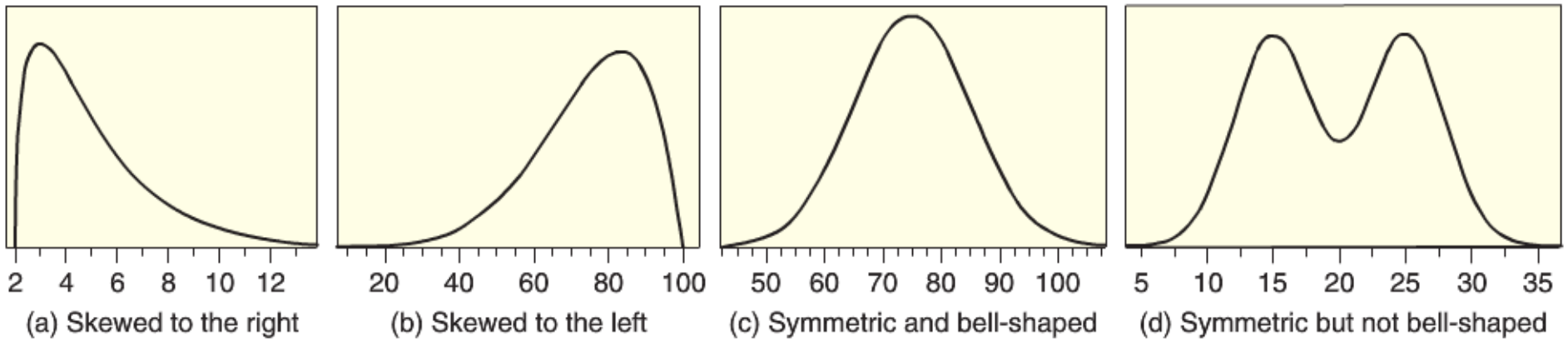
Review Histograms – movie US revenue

Domgross range	Frequency Count
(0 – 25]	510
(25 – 50]	312
(50 – 75]	234
(75 – 100]	176
(100 – 125]	111
(125 – 150]	99
(150 – 175]	62
(175 – 200]	51
(200 – 225]	40
(225 – 250]	32

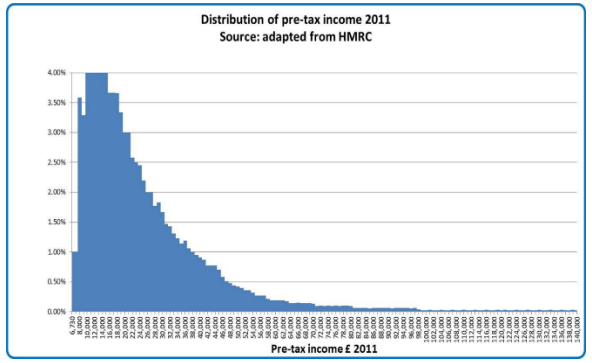


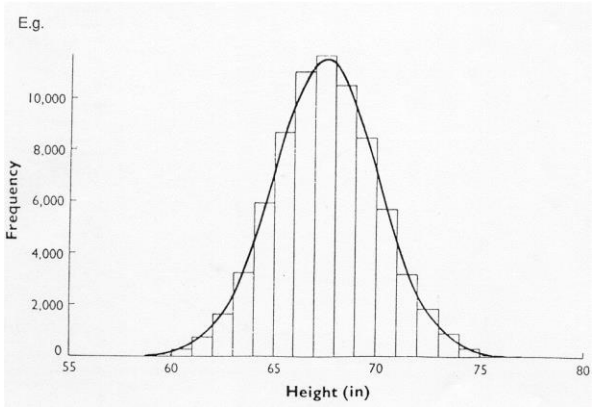
Matplotlib: `plt.hist(data)`

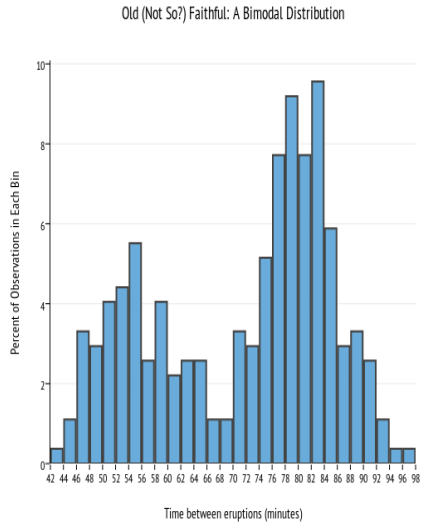
Review common shapes of data distributions

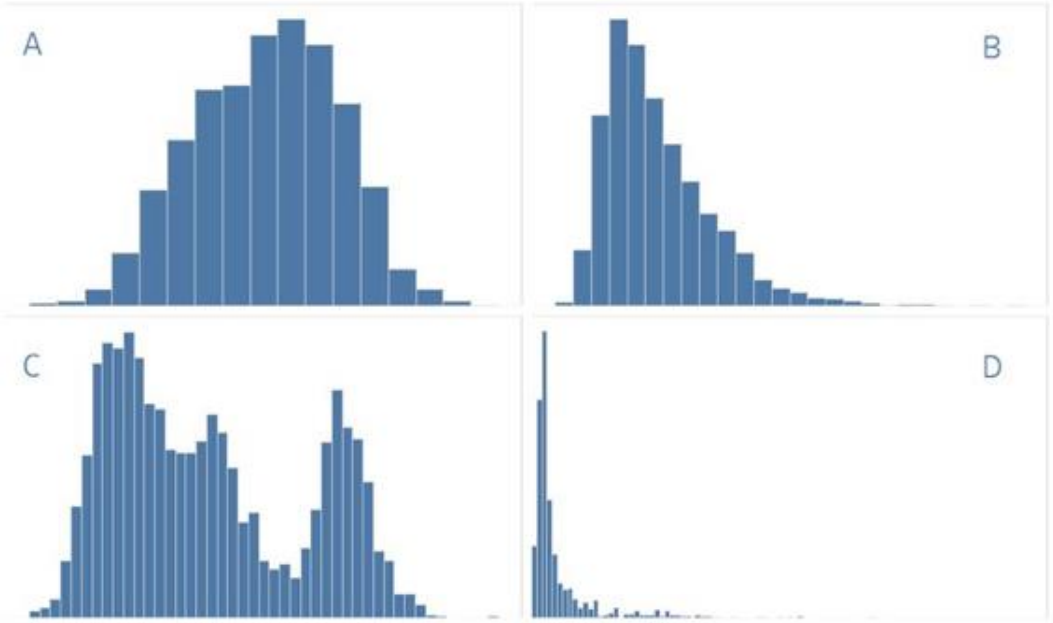


Income distribution









???

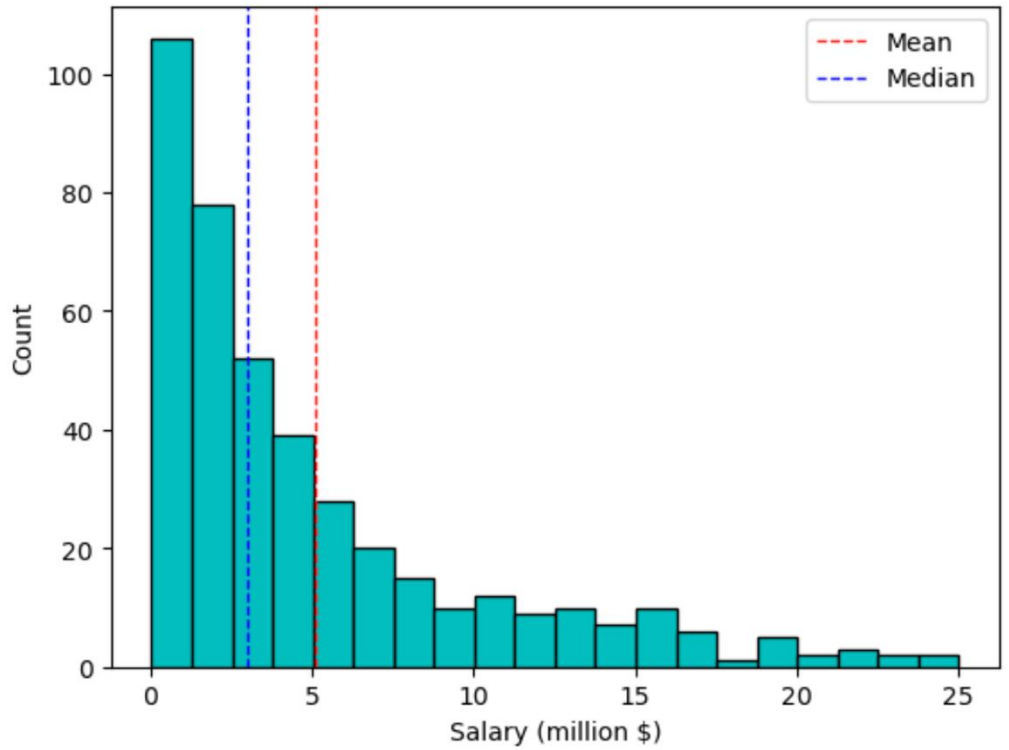

```
plt.ylabel("y label")

plt.xlabel("x label")

plt.title("my title")
```

Quantitative data: statistics for central tendency

Two statistics for measuring the “central value” of a sample of quantitative data are the ***mean*** and the ***median***



The mean

$$\text{Mean} = \frac{\text{Sum of all data values}}{\text{Number of data values}}$$

$$\text{Mean} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

```
import statistics
statistics.mean(data_list)
```

The median

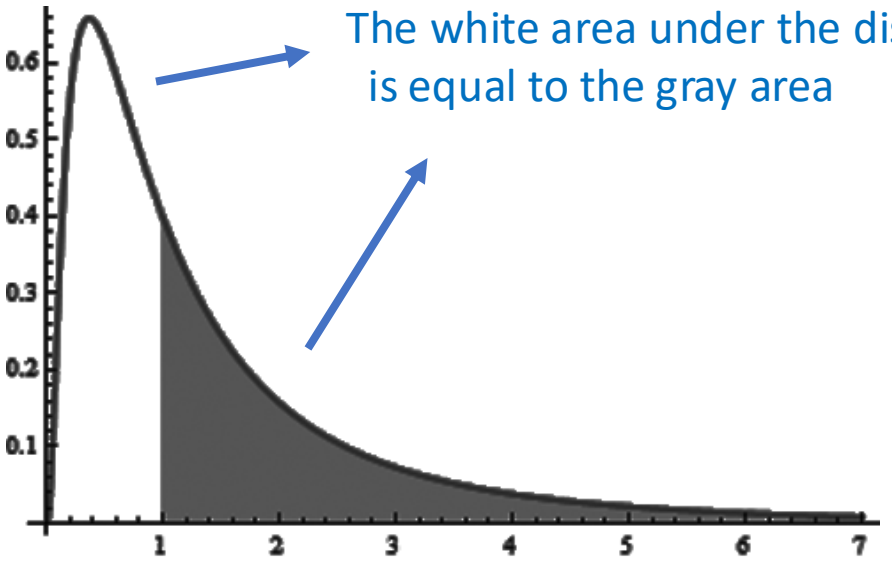
The **median** is a value that splits the data in half

- i.e., half the values in the data are smaller than the median and half are larger

To calculate the median for a data sample of size *n*, sort the data and then:

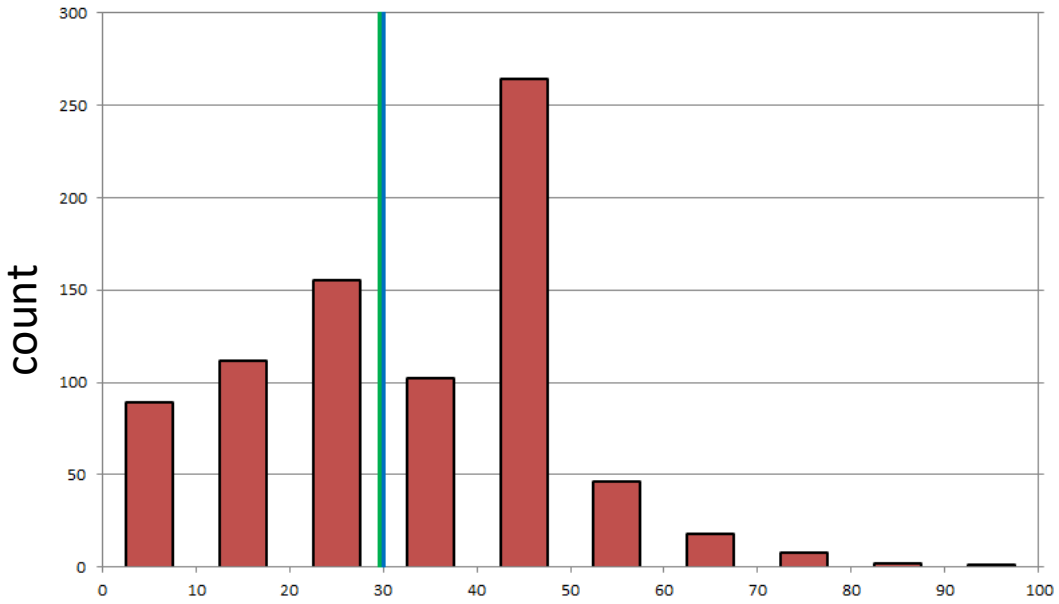
- If n is **odd**: The middle value of the sorted data
- If n is **even**: The average of the middle two values of the sorted data

The median



The white area under the distribution is equal to the gray area

The sum of the heights of the bars on the left is equal to the sum of the heights of the bars on the right

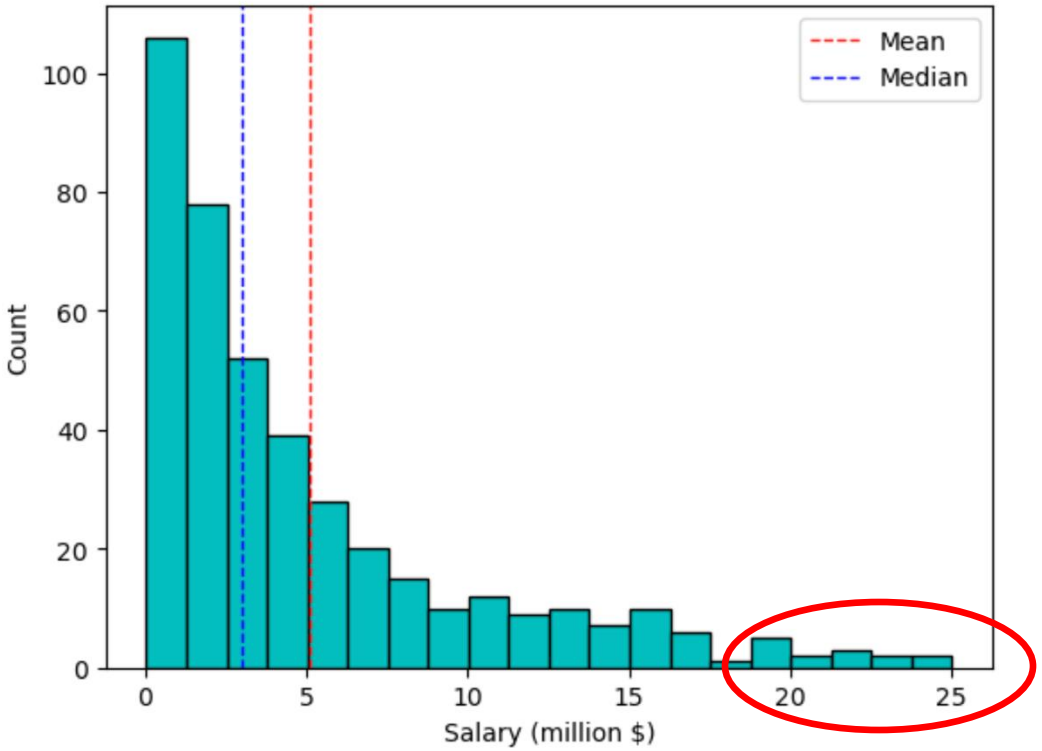



```
import statistics
statistics.median(data_list)
```

Let's explore this in Jupyter!

Outliers

An **outlier** is an observed value that is notably distinct from the other values in a dataset by being much smaller or larger than the rest of the data.

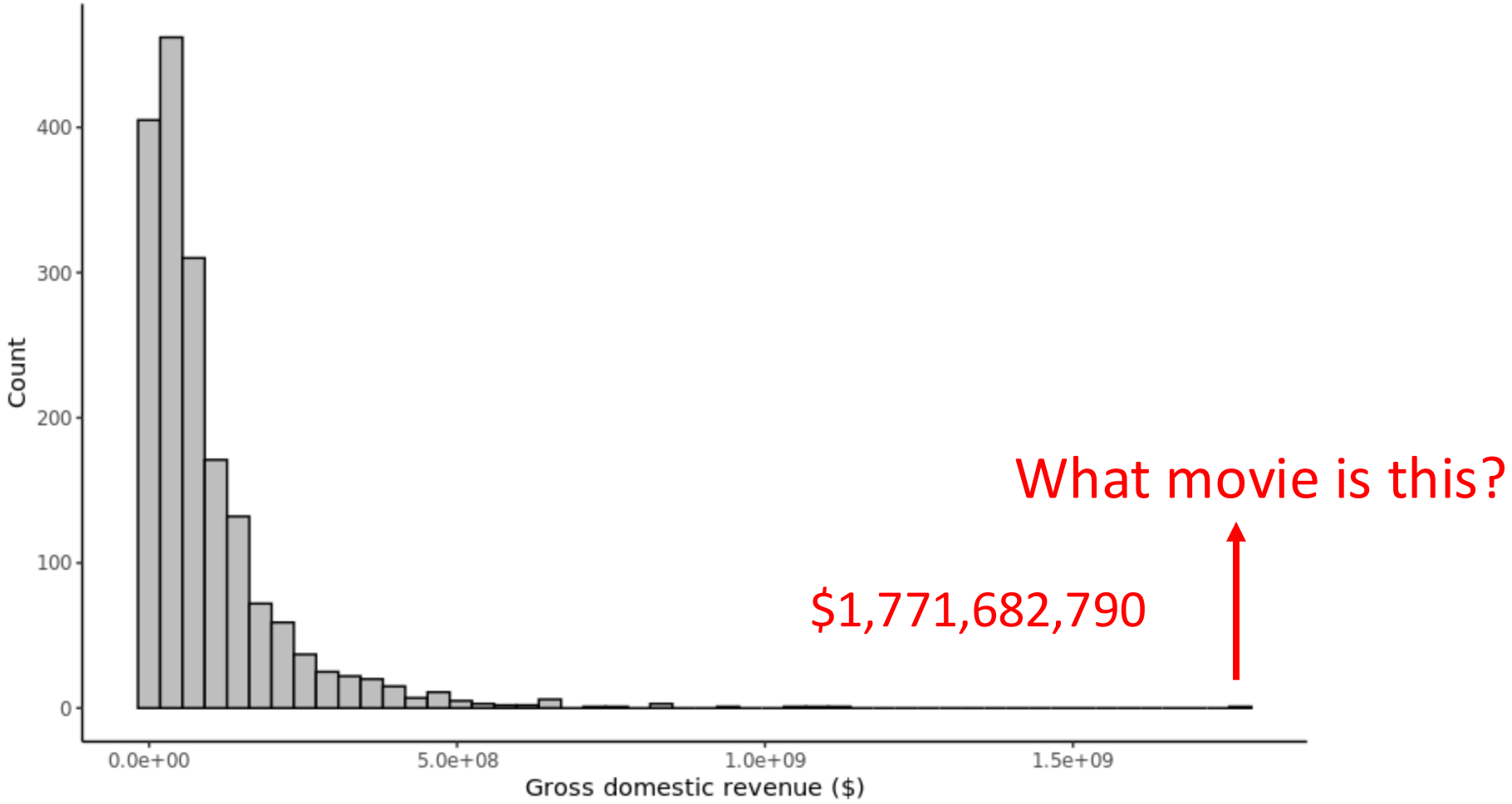


Outliers can potentially have a large influence on the statistics you calculate

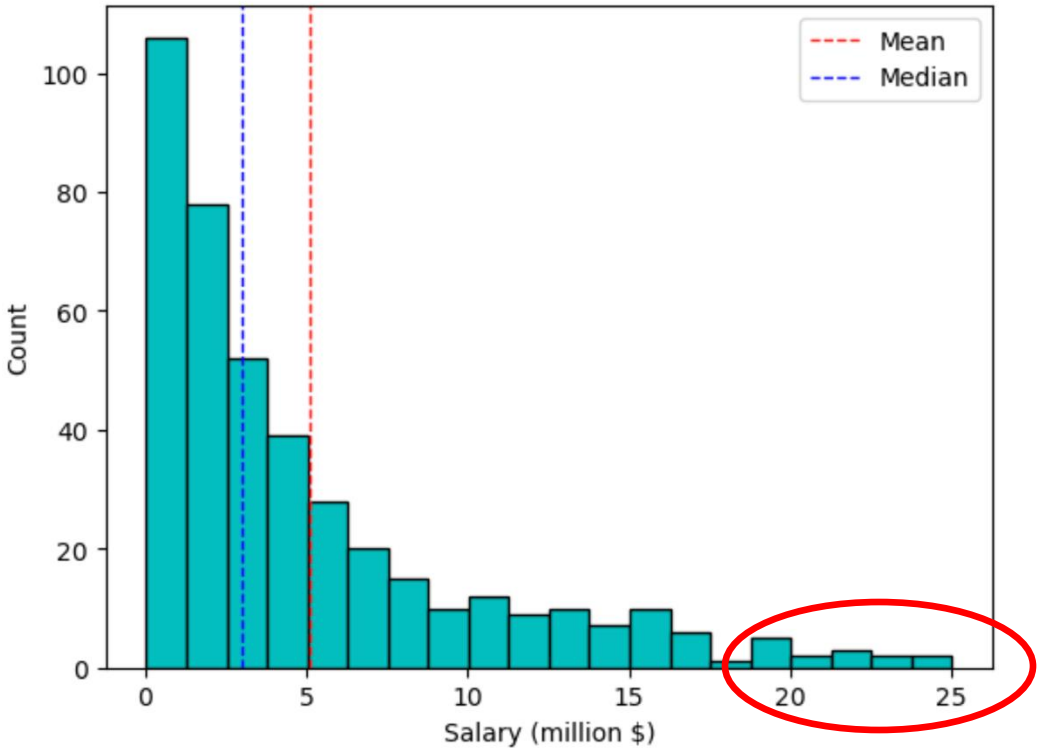
One should examine outliers to understand what is causing them

- If there are due to an error, remove them
- Otherwise, need to think about how to treat them
  - Could be interesting phenomenon
  - Could restrict data to a particular range of values
  - Etc.

Bechdel outliers



Outliers' impact on mean and median



The median is *resistant* to outliers

- i.e., not affected much by outliers

The mean is not resistant to outliers

What is the mean and median of the data: 1, 2, 3, 4, 990?

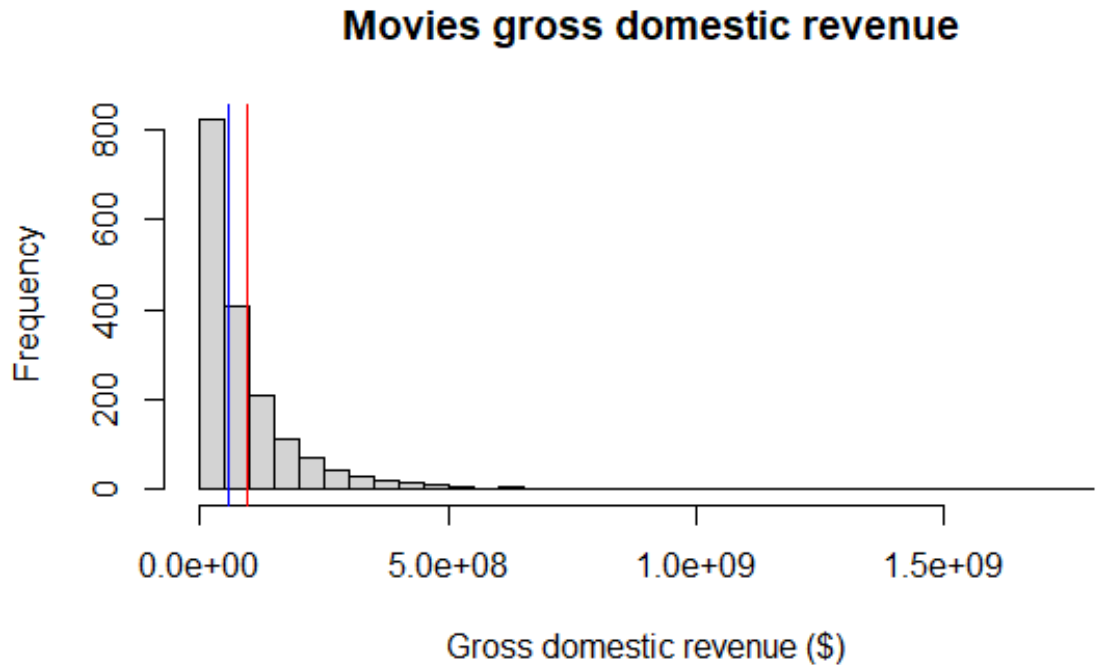
- Mean = 200
- Median = 3


ANY
QUESTIONS?

Measures of spread

Characterizing the spread

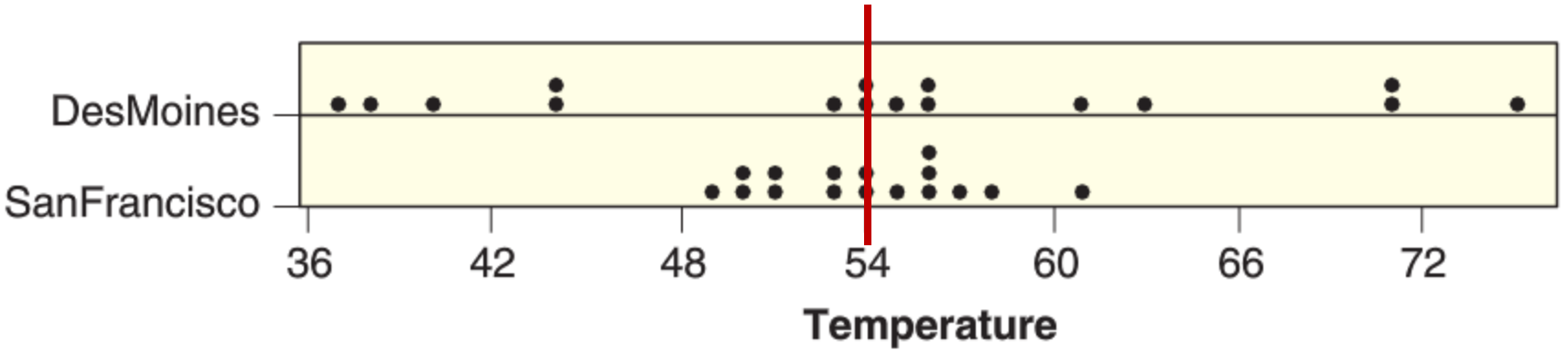
The mean and median are numbers that tell us about the center of a distribution



We can also use numbers to characterize how data is spread

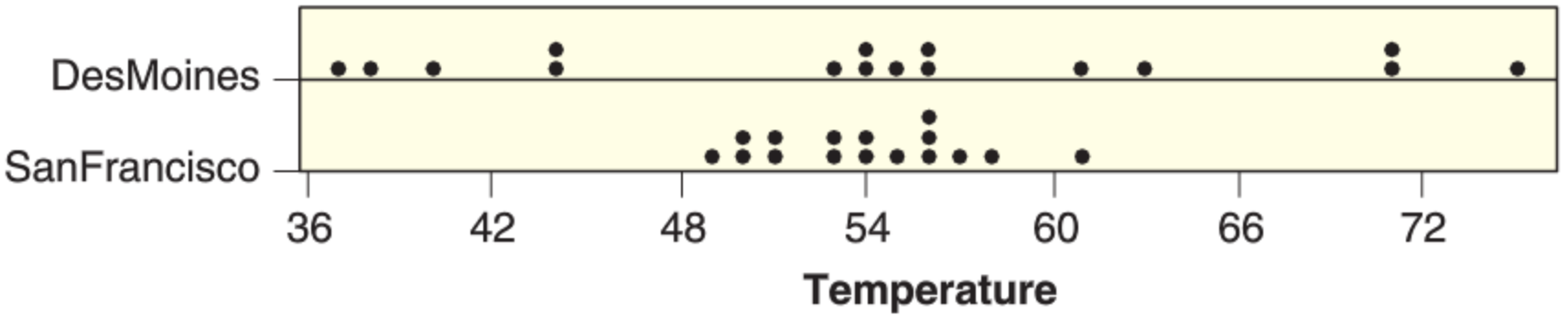
Average monthly temperature: Des Moines vs. San Francisco

Data measured on April 14th from 1997 to 2010:



Mean temperature (°F): Des Moines = 54.49 San Fran = 54.01

Which has the larger standard deviation?



s_{DM} = 11.73 °F s_{SF} = 3.38 °F

The standard deviation

The standard deviation can be computed using the following formula:

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Standard deviation measures roughly how far the data are from their average

Example: computing the standard deviation

Suppose we had a sample with n = 4 points:

$x_1 = 8, \quad x_2 = 2, \quad x_3 = 6, \quad x_4 = 4,$

We can compute the mean using the formula:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad = \tfrac{1}{4} \cdot (x_1 + x_2 + x_3 + x_4) \quad = \tfrac{1}{4} \cdot (8 + 2 + 6 + 4)$$

The standard deviation can be computed using the formula:

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(remember order of operations!)

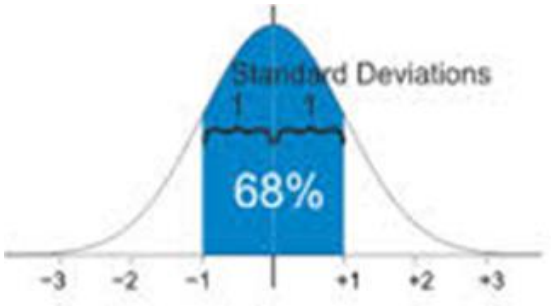
```
statistics.stdev(data_list)
```

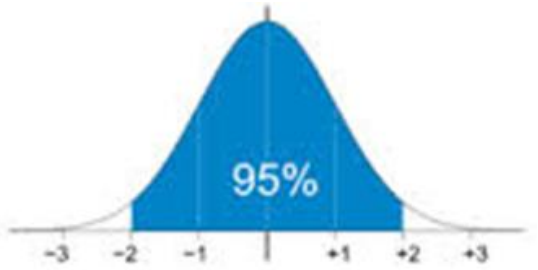
Normally distributed data

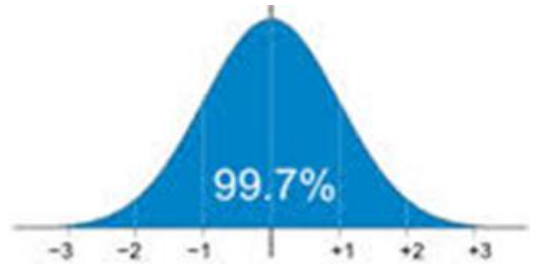
The bulk of the data are in the range "average ± a few SDs"

If the data is “normally distributed” (bell shaped distribution) than the following holds:

Range	Proportion
Average ± 1 SDs	68% of the data
Average ± 2 SDs	95% of the data
Average ± 3 SDs	99.7% of the data







Chebyshev's Inequality

The bulk of the data are in the range "average ± a few SDs"

Chebyshev's Inequality: No matter what the shape of the distribution, the proportion of values in the range "average ± $z \cdot$ SDs" is at least $1 - 1/z^2$

Range	Proportion
Average ± 2 SDs	at least 1 - 1/4 (75%)
Average ± 3 SDs	at least 1 - 1/9 (88.88…%)
Average ± 4 SDs	at least 1 - 1/16 (93.75%)
Average ± 5 SDs	at least 1 - 1/25 (96%)

Let’s briefly explore standard deviations in Jupyter!

Z-scores

Standardized units

Item in the world are often measured on very different scales

How can we create a standard scale to quantify unusual/large/impressive values?

Z-scores measure how many SDs a value is from average:

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

- Negative z: value below average
- Positive z: value above average
- z = 0: value equal to average

Which Accomplishment is most impressive?

LeBron James is a basketball player who had the following statistics in 2011:

- Field goal percentage (FGPct) = 0.510
- Points scored = 2111
- Assists = 554
- Steals = 124

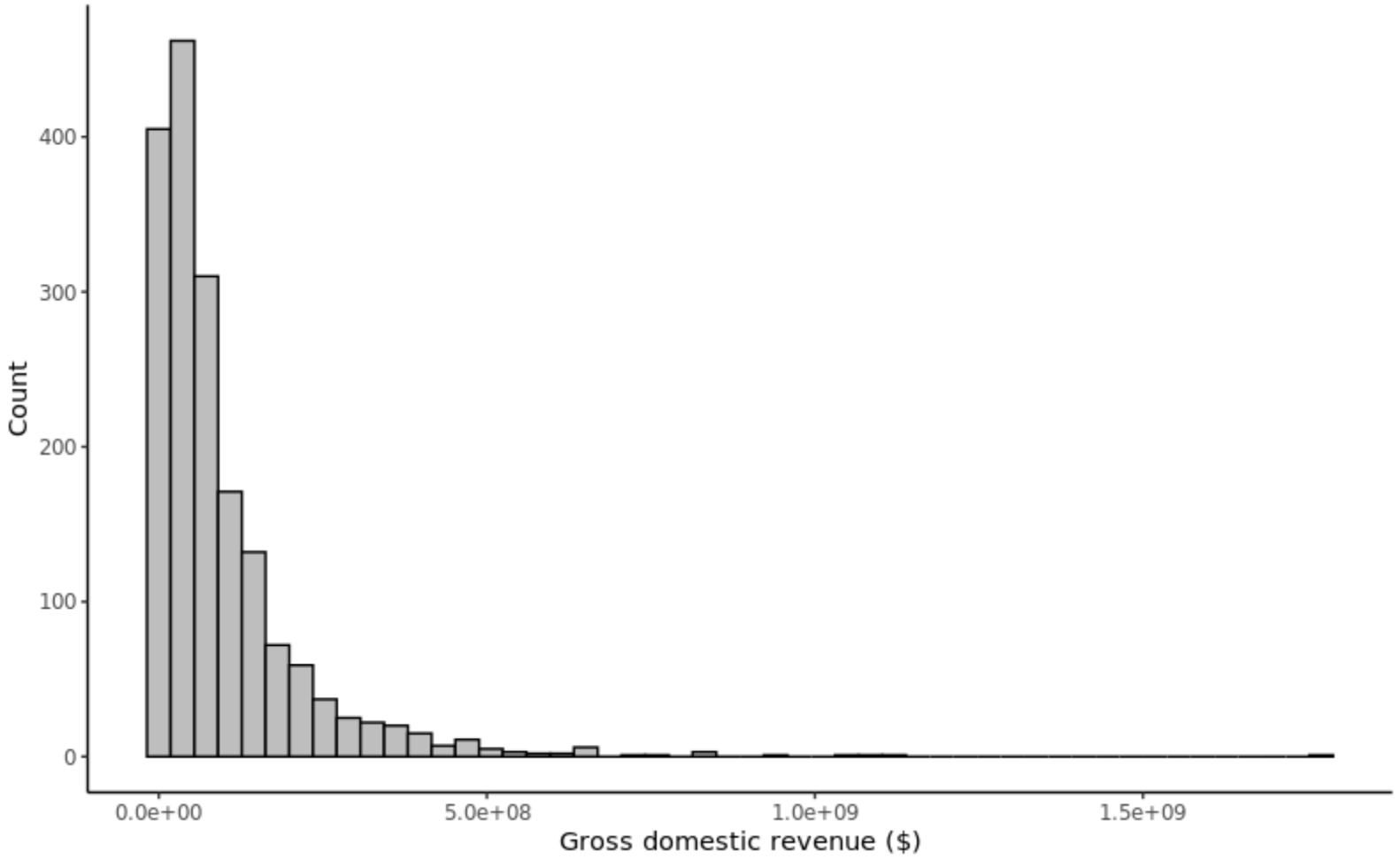
The summary statistics of the NBA in 2011 are given below

$$\text{z-score}(x_i) = \frac{x_i - \bar{x}}{s}$$

	Mean	Standard Deviation
FGPct	0.464	0.053
Points	994	414
Assists	220	170
Steals	68.2	31.5

Question: Relative to his peers, which statistic is most and least impressive?

What is star wars' z-score?



$1,771,682,790

Let’s try it in Python!

Relationships between two quantitative variables

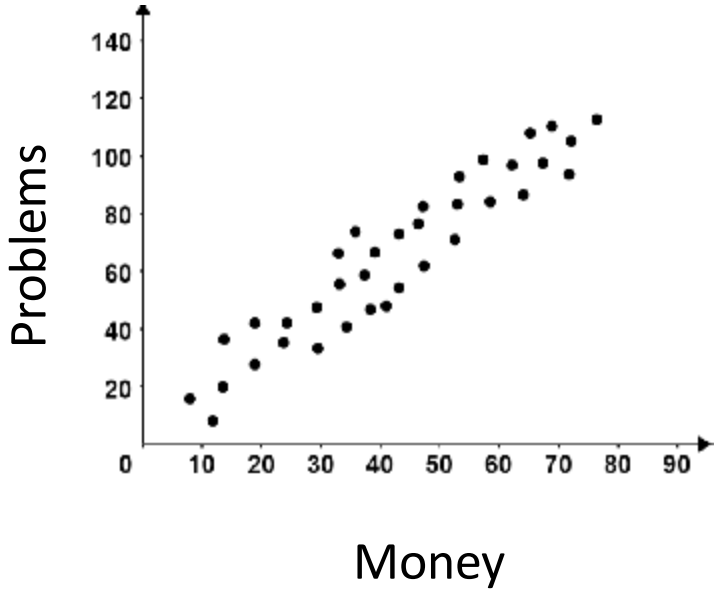
Do movies with larger budgets make more money?

Q: How could we visualize the data to see if this is true?

Visualizing two quantitative variables: scatterplots

A **scatterplot** graphs the relationship between two variables

- Each axis represents the value of one variables
- Each point the plot shows the value for the two variables for a single data case



If there is an explanatory and response variable, then the explanatory variable is put on the x-axis and the response variable is put on the y-axis

```
plt.plot(x, y, '.')
```

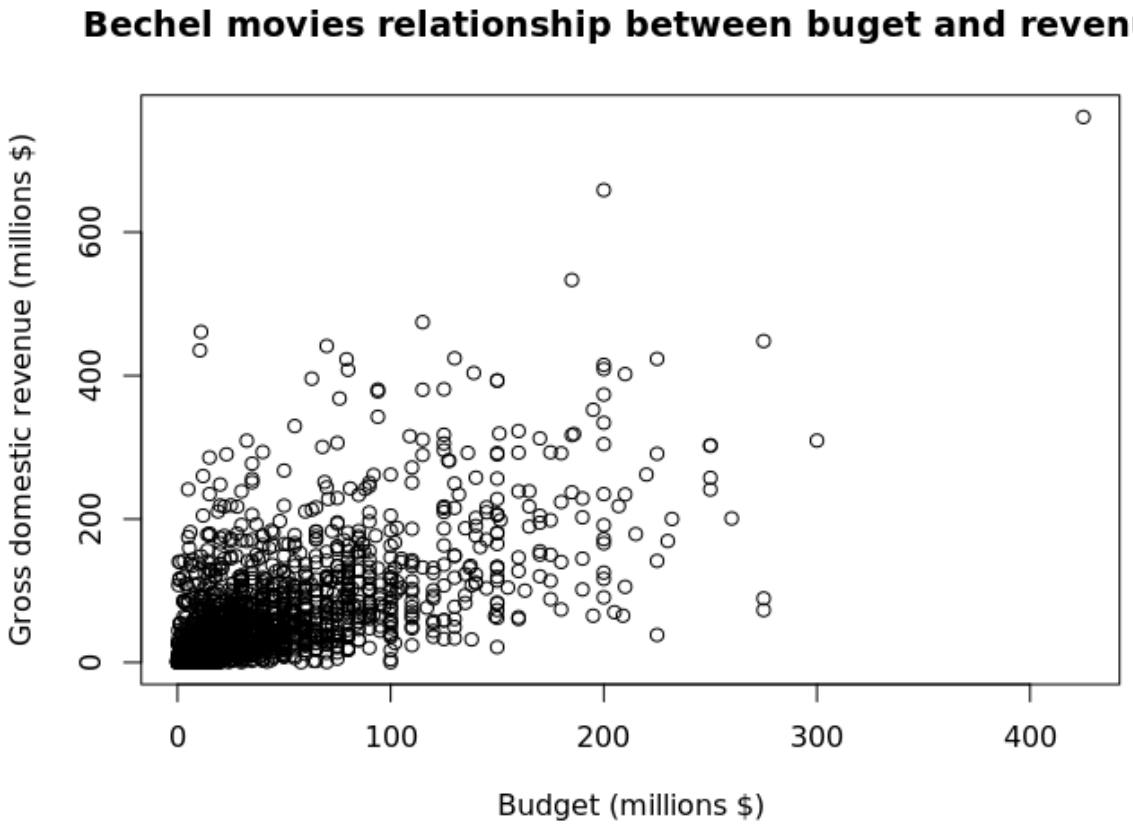
Do movies with larger budgets make more money?

Q: If we want to create a scatter plot to address whether movies with larger budgets make more money, what variables should we use in our plot?

A:

- budget_2013
- domgross_2013

Relationship movie money spent and made



Do movies with larger budgets make more money?

Do most movies make money?

- How could we create a more informative scatter plot of this data?

Matplotlib: `plt.plot(x, y)`

Questions when looking at scatterplots

Do the points show a clear trend?

- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Questions when looking at scatterplots

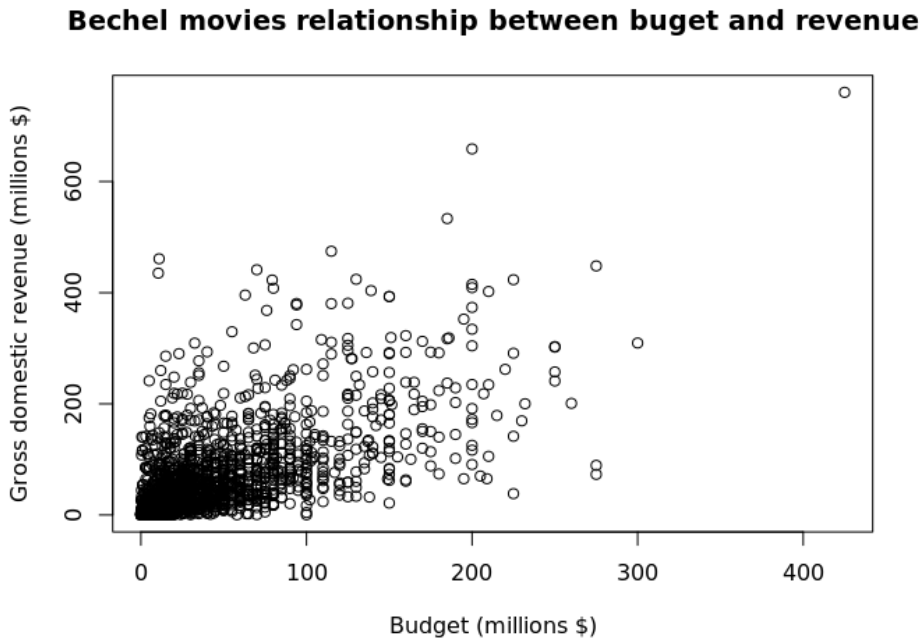
Do the points show a clear trend?

- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?

Budget and revenue



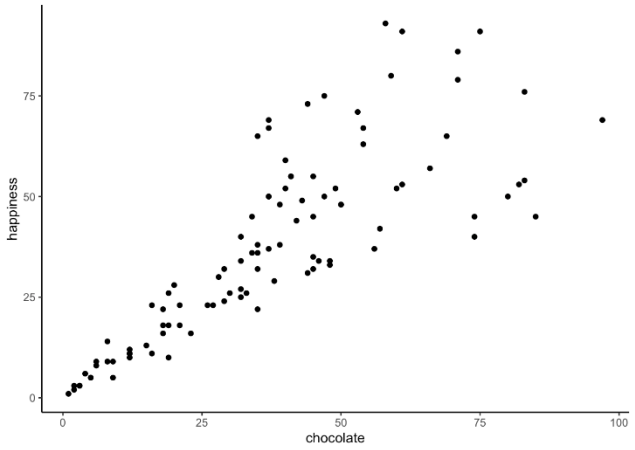
Positive, negative, no correlation

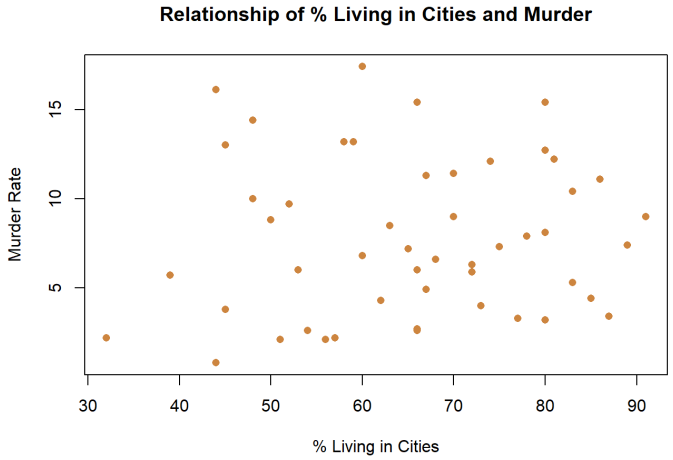
Do the points show a clear trend?

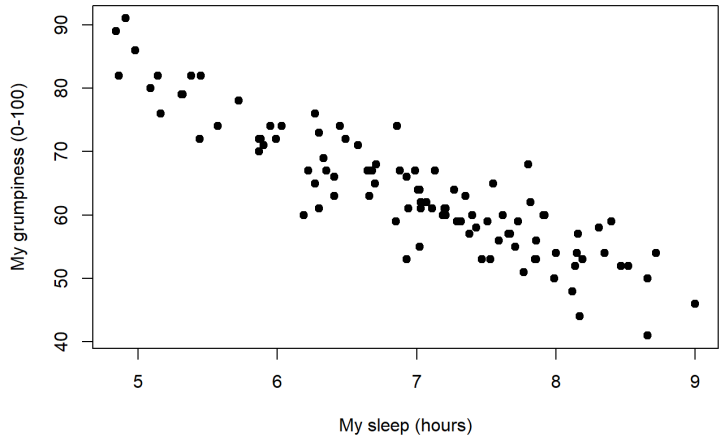
- Does it go upward or downward?
- How much scatter around the trend?

Does the trend seem be linear (follow a line) or is it curved?

Are there any outlier points?







The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

```
statistics.correlation(x, y)
```

Properties of the correlation

Correlation as always between -1 and 1: $-1 \leq r \leq 1$

The sign of r indicates the direction of the association

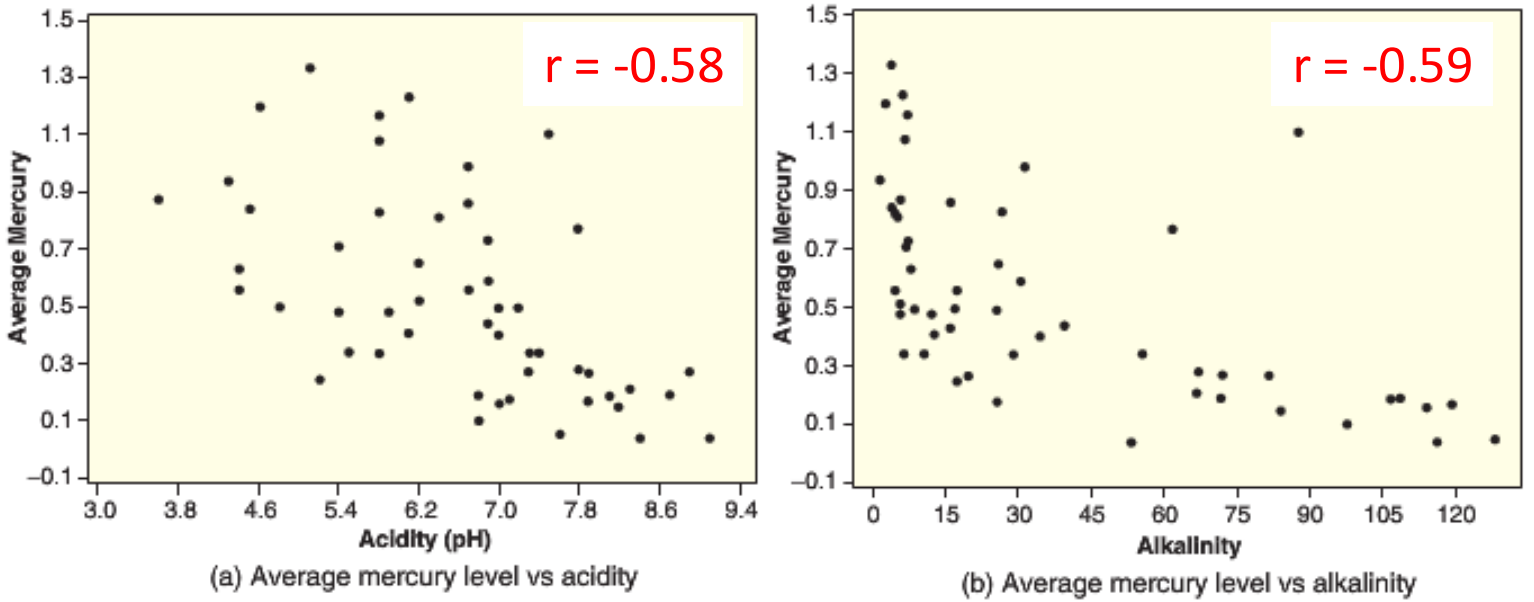
Values close to ± 1 show strong linear relationships, values close to 0 show no linear relationship

Correlation is symmetric: r = cor(x, y) = cor(y, x)

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

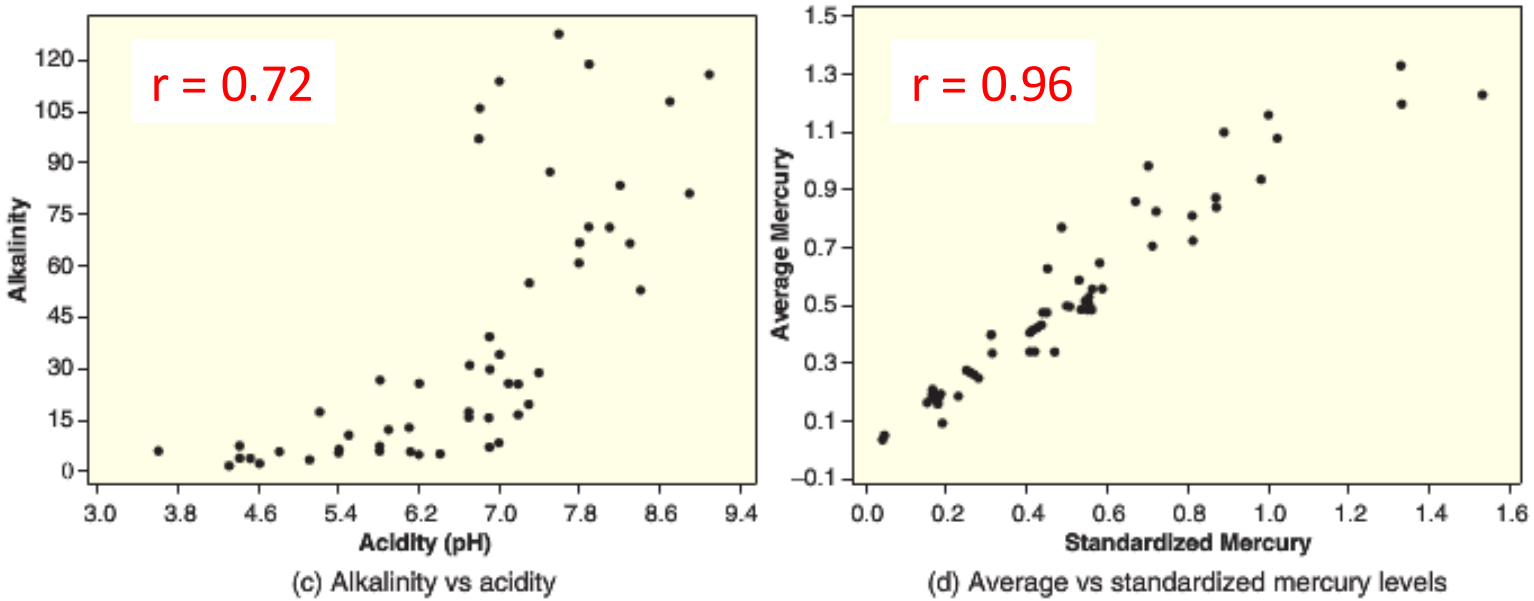
Florida lakes

[Correlation game]



(a) Average mercury level vs acidity

(b) Average mercury level vs alkalinity



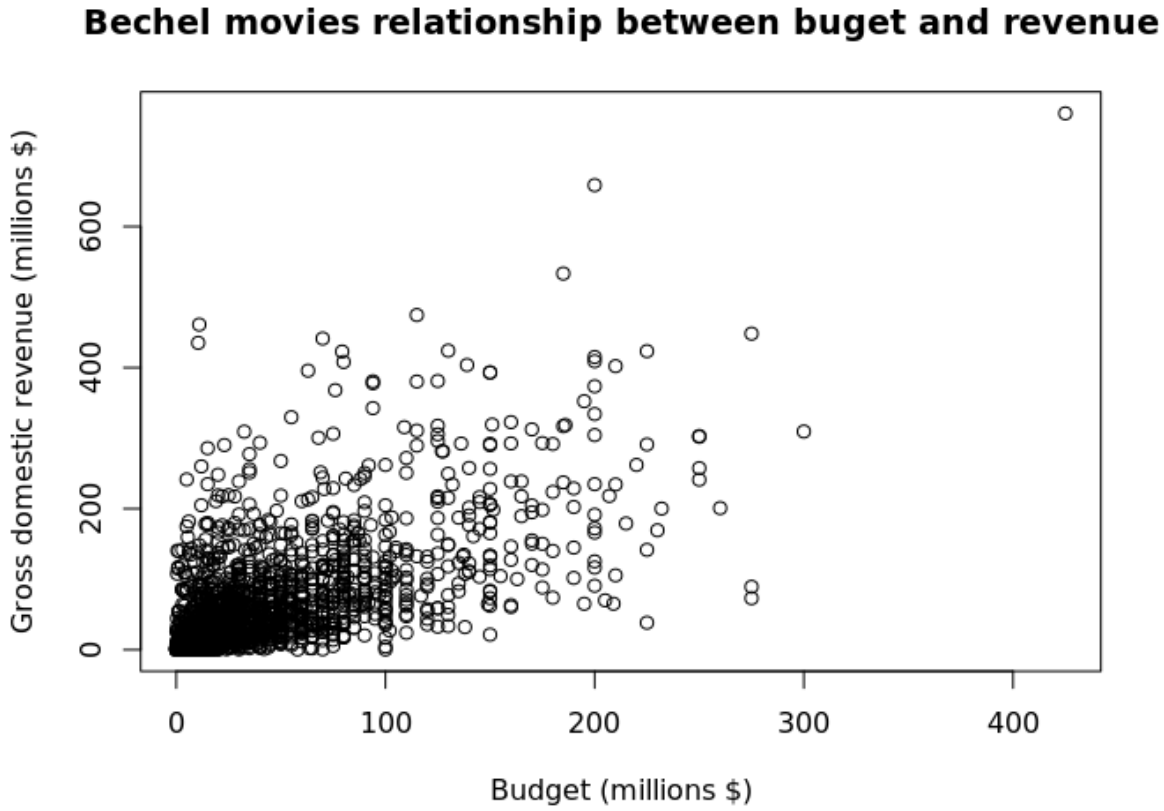
(c) Alkalinity vs acidity

(d) Average vs standardized mercury levels

Movie budget and revenue correlation?

The **correlation** is measure of the strength and direction of a linear association between two variables

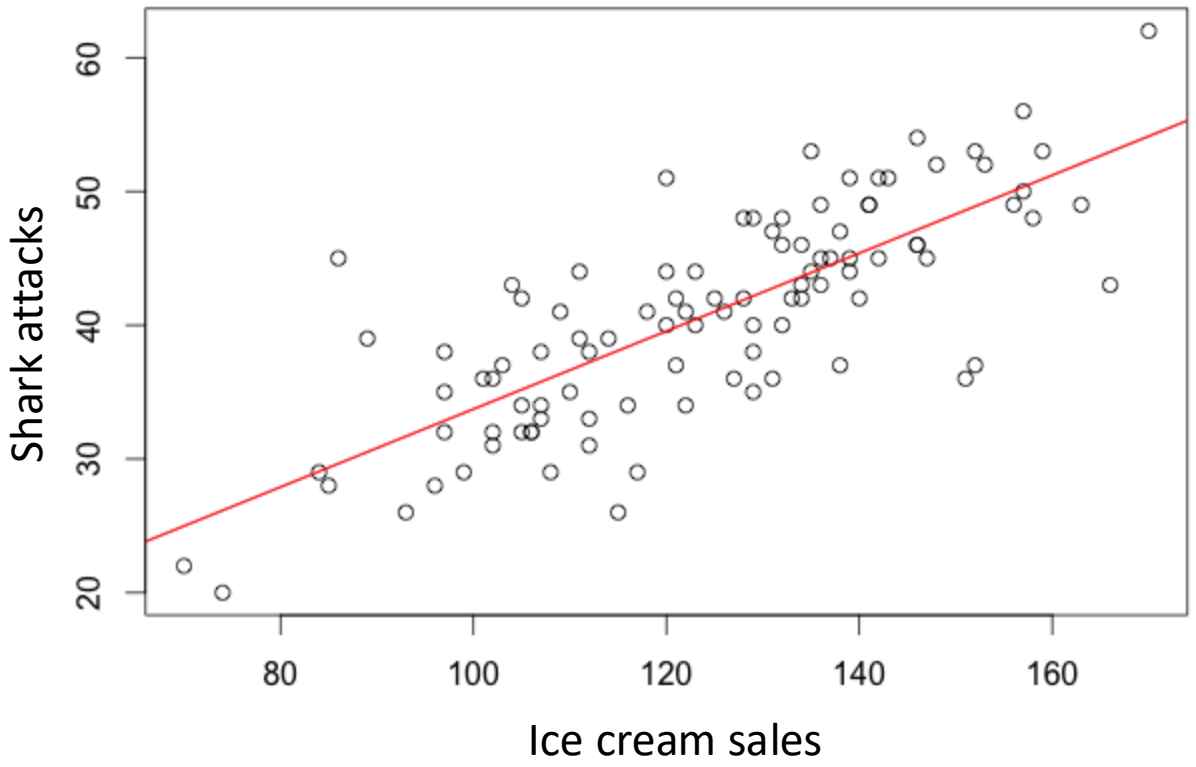
r = ?



Let's calculate the correlation in Python!

Correlation caution #1

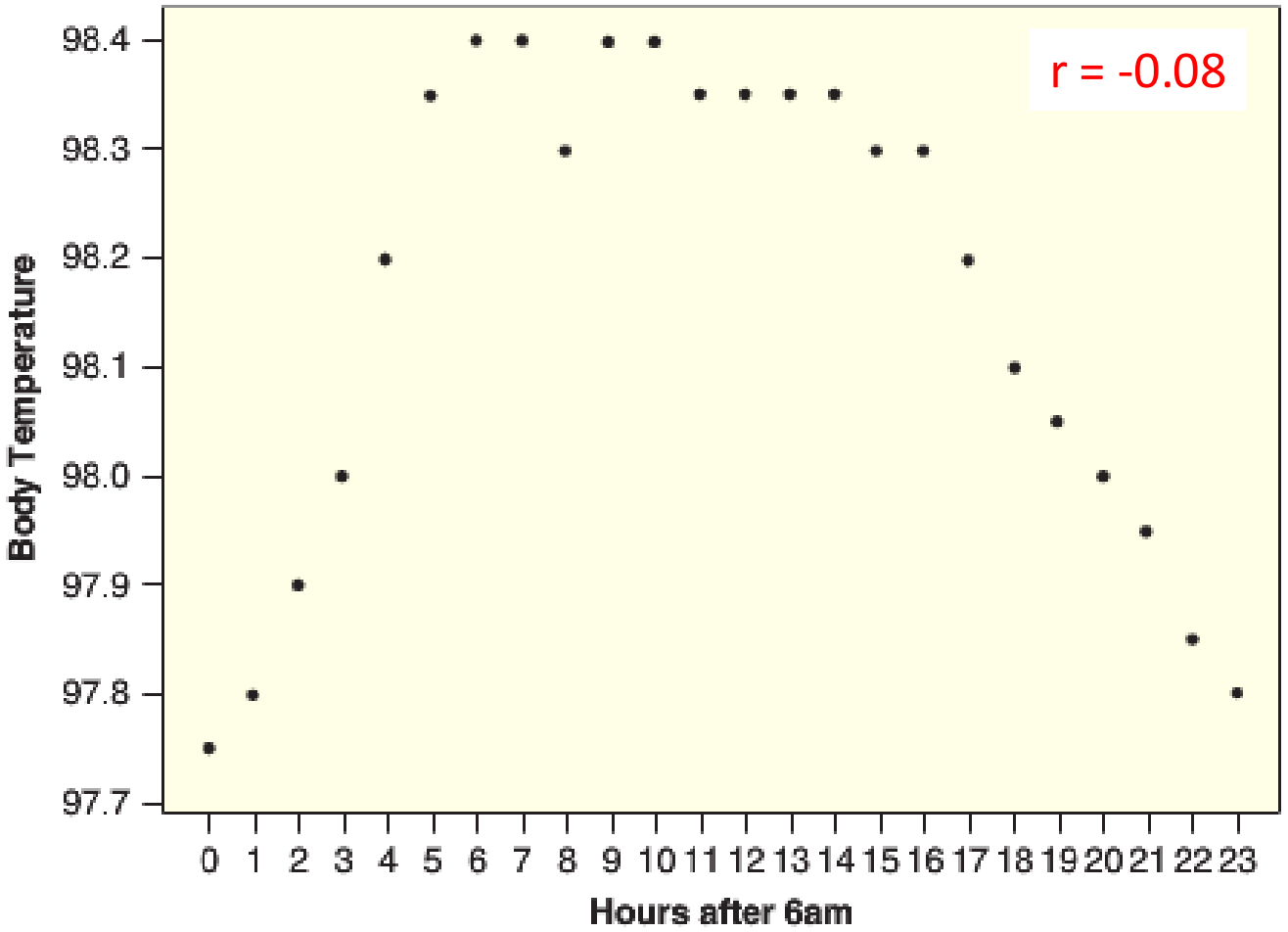
A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



Correlation caution #2

A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a linear relationship.

Body temperature as a function of time of the day

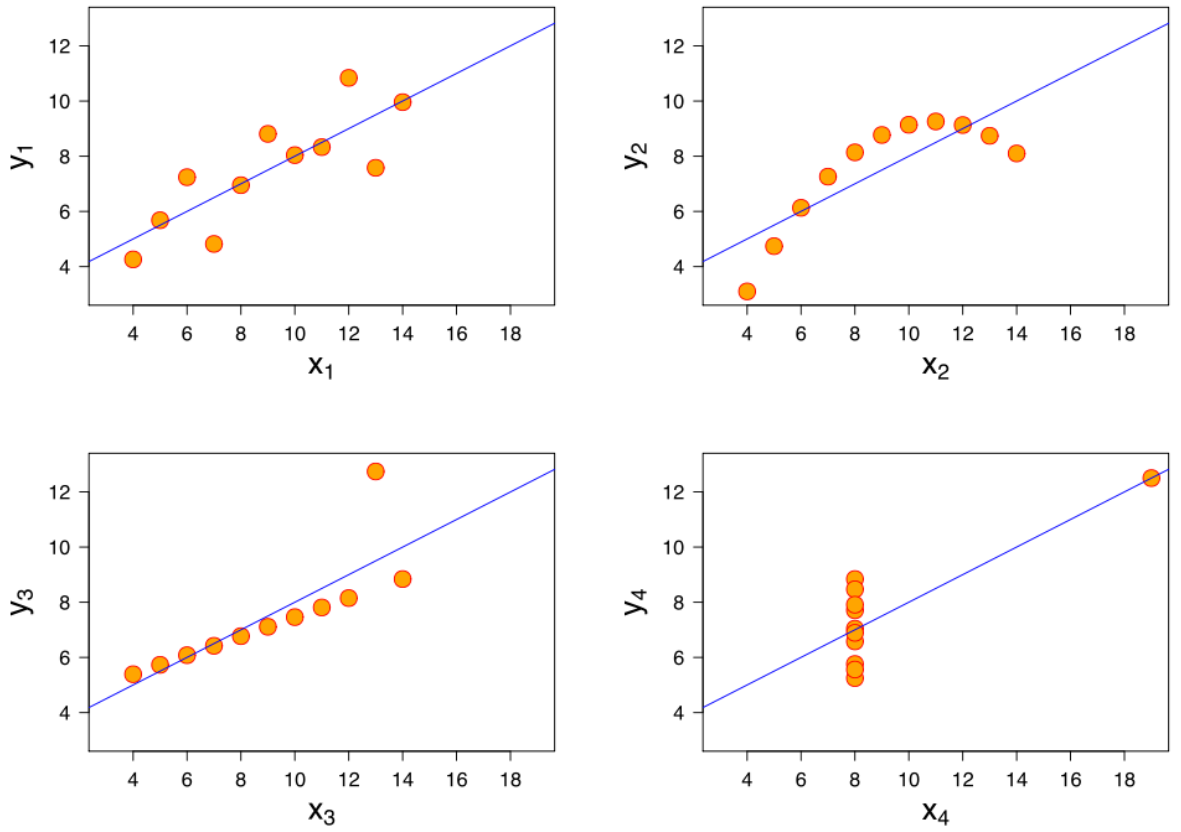


Correlation caution #3

Correlation can be heavily influenced by outliers. Always plot your data!

Example:
"Anscombe's quartet"

r = 0.81 for all plots



Next week: array computations…

Homework 2 has been posted!

```
import YData
YData.download_homework(2)
```

It is due on Gradescope on Sunday September 14th at 11pm

- **Be sure to mark each question on Gradescope!**

Notes:

- There is an ~18 page reading from the book "Data and the American Dream" that you need to do, so I recommend you get started on this soon.