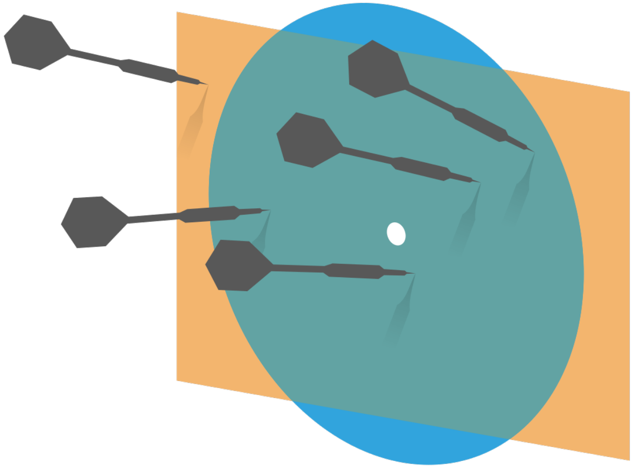
Probability and Statistics for Computer Science



Credit: wikipedia

"Correlation is not Causation"
but Correlation is so beautiful!

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 9.1.2020

Last time

- Mean
- Standard deviation
- Variance
- Standardizing data

Objectives

- Median, Interquartile range, box plot and outlier
- Scatter plots, Correlation Coefficient
- Visualizing & Summarizing *relationships*

  Heatmap, 3D bar, Time series plots,

Median

- To organize the data we first sort it
- Then *if* the number of items N is odd

  median = middle item's value

  if the number of items N is even

  median = mean of middle 2 items' values

Properties of Median

- Scaling data scales the median

$$median(\{k \cdot x_i\}) = k \cdot median(\{x_i\})$$

- Translating data translates the median

$$median(\{x_i + c\}) = median(\{x_i\}) + c$$

Percentile

- k^{th} percentile is the value relative to which k% of the data items have smaller or equal numbers
- Median is roughly the 50^{th} percentile

Interquartile range

- iqr = (75th percentile) - (25th percentile)
- Scaling data scales the interquartile range

$$iqr(\{k \cdot x_i\}) = |k| \cdot iqr(\{x_i\})$$

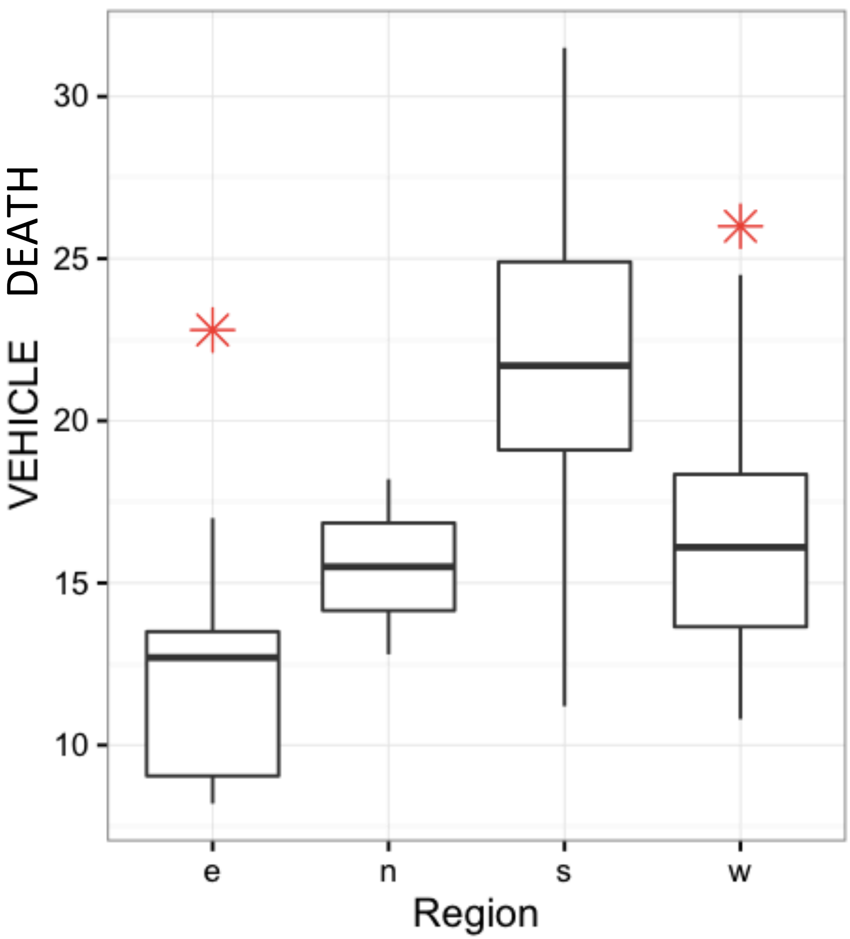
- Translating data does **NOT** change the interquartile range

$$iqr(\{x_i + c\}) = iqr(\{x_i\})$$

Box plots

- Boxplots
  - Simpler than histogram
  - Good for outliers
  - Easier to use for comparison

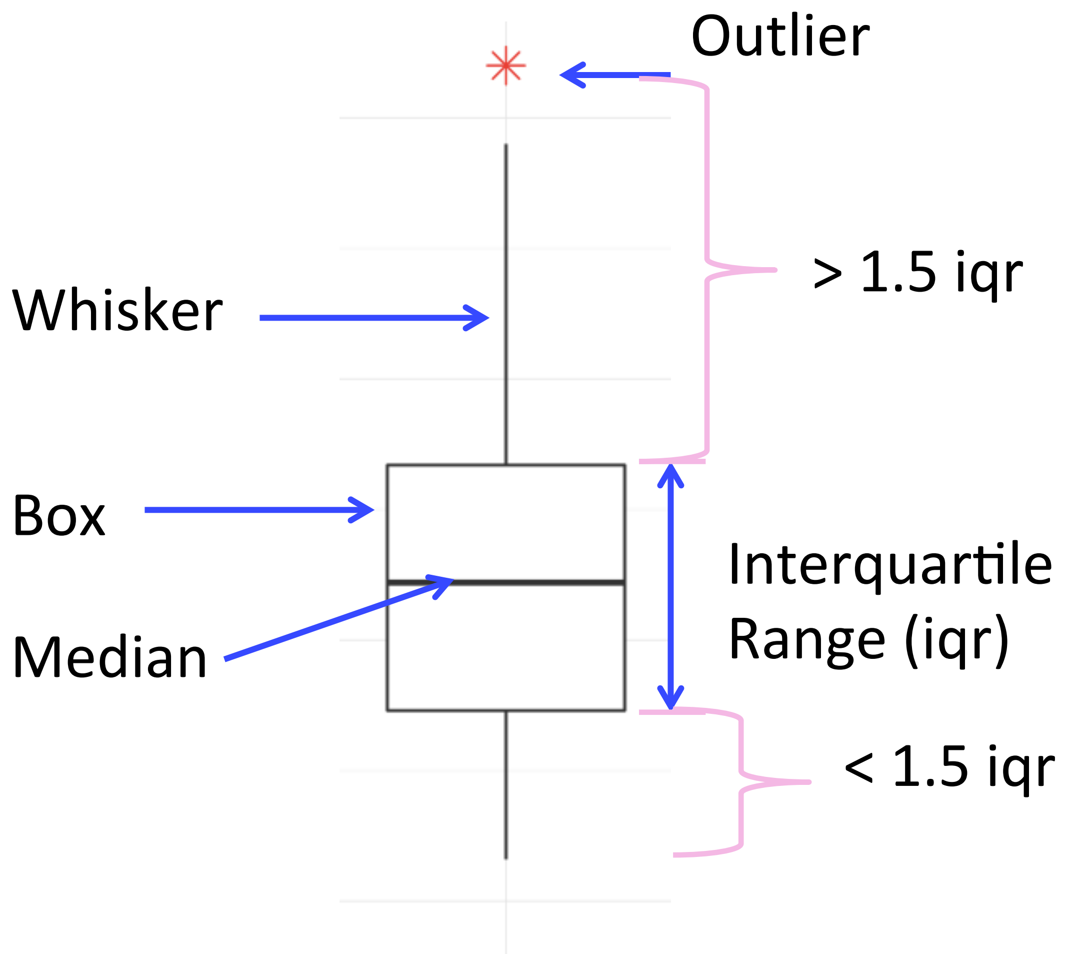
Vehicle death by region



Data from https://www2.stetson.edu/~jrasp/data.htm

Boxplots details, outliers

- How to define outliers? (the default)



Sensitivity of summary statistics to outliers

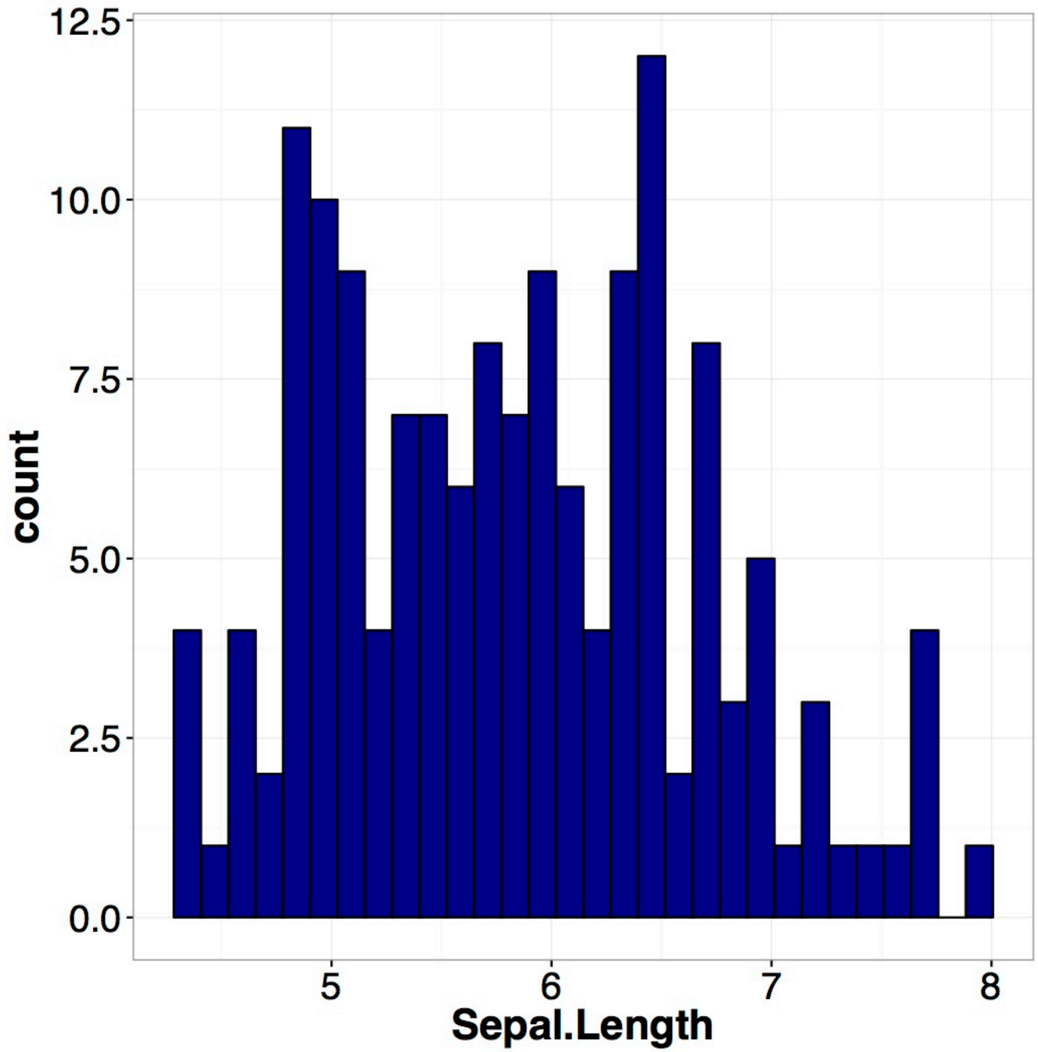
- mean and standard deviation are very sensitive to outliers
- median and interquartile range are not sensitive to outliers

Modes

- Modes are peaks in a histogram
- If there are more than 1 mode, we should be curious as to why

Multiple modes

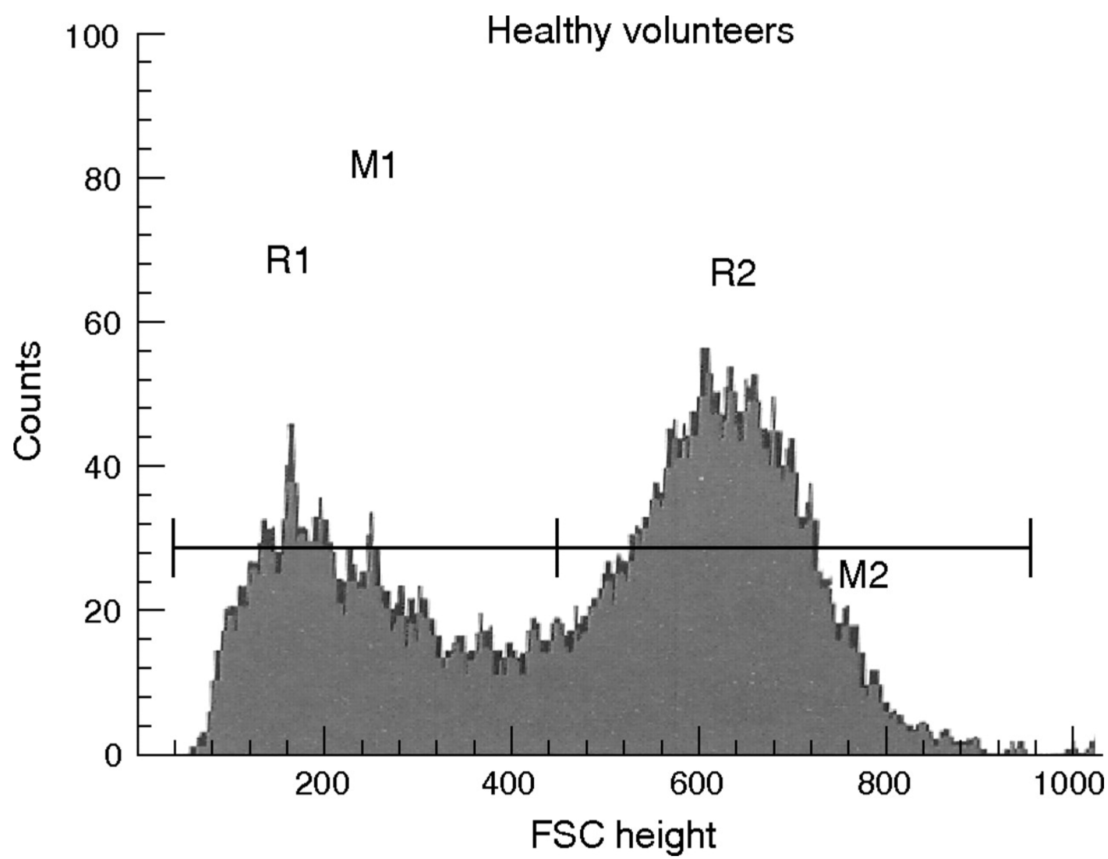
- We have seen the "iris" data which looks to have several peaks



Data: "iris" in R

Example Bi-modes distribution

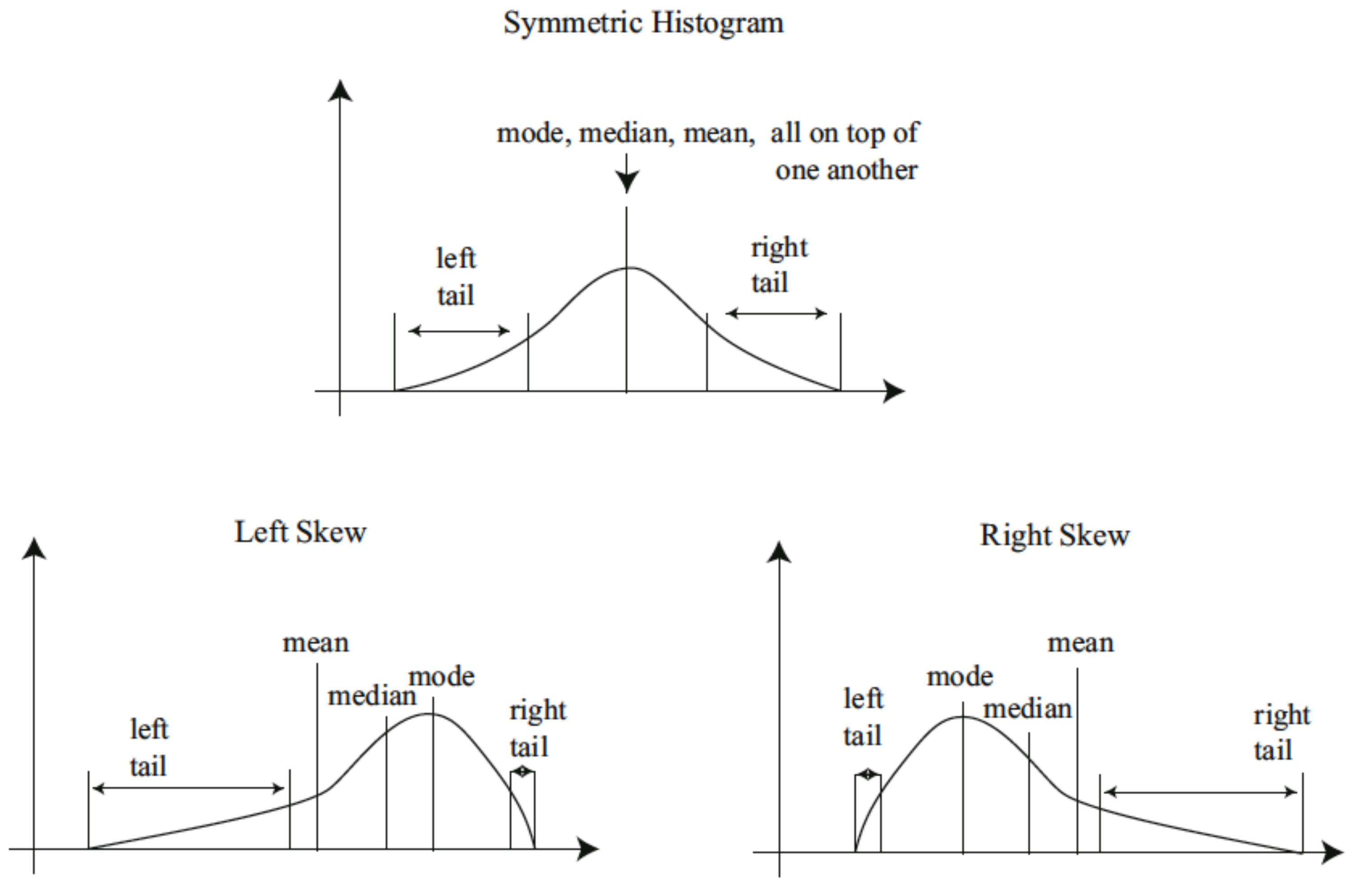
- Modes may indicate multiple populations



Data: Erythrocyte cells in healthy humans

Piagnerelli, JCP 2007

Tails and Skews

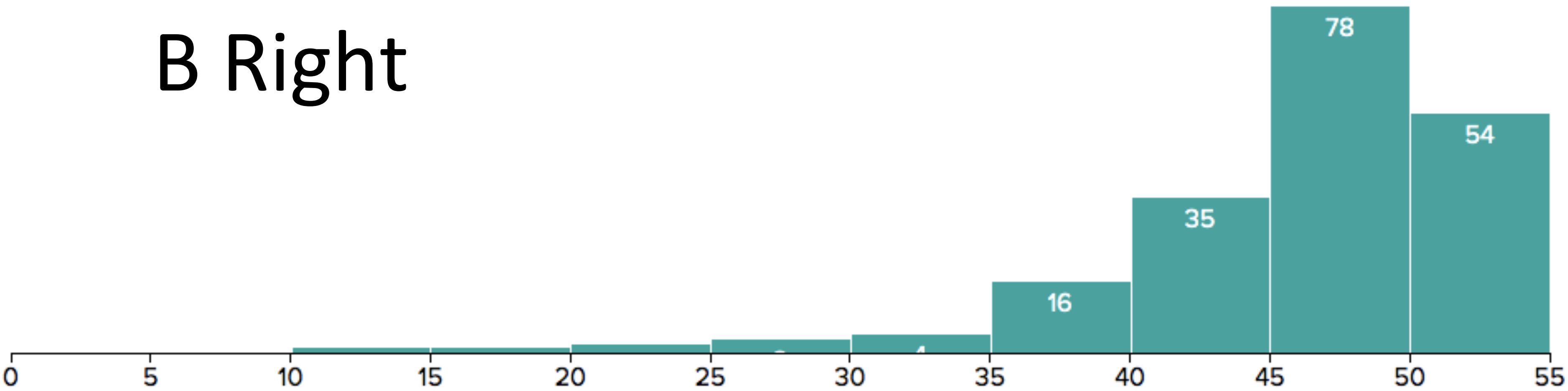


Credit: Prof.Forsyth

Q. How is this skewed?

A Left

B Right



Median = 47

Looking at relationships in data

- Finding relationships between features in a data set or many data sets is one of the most important tasks in data analysis

Relationship between data features

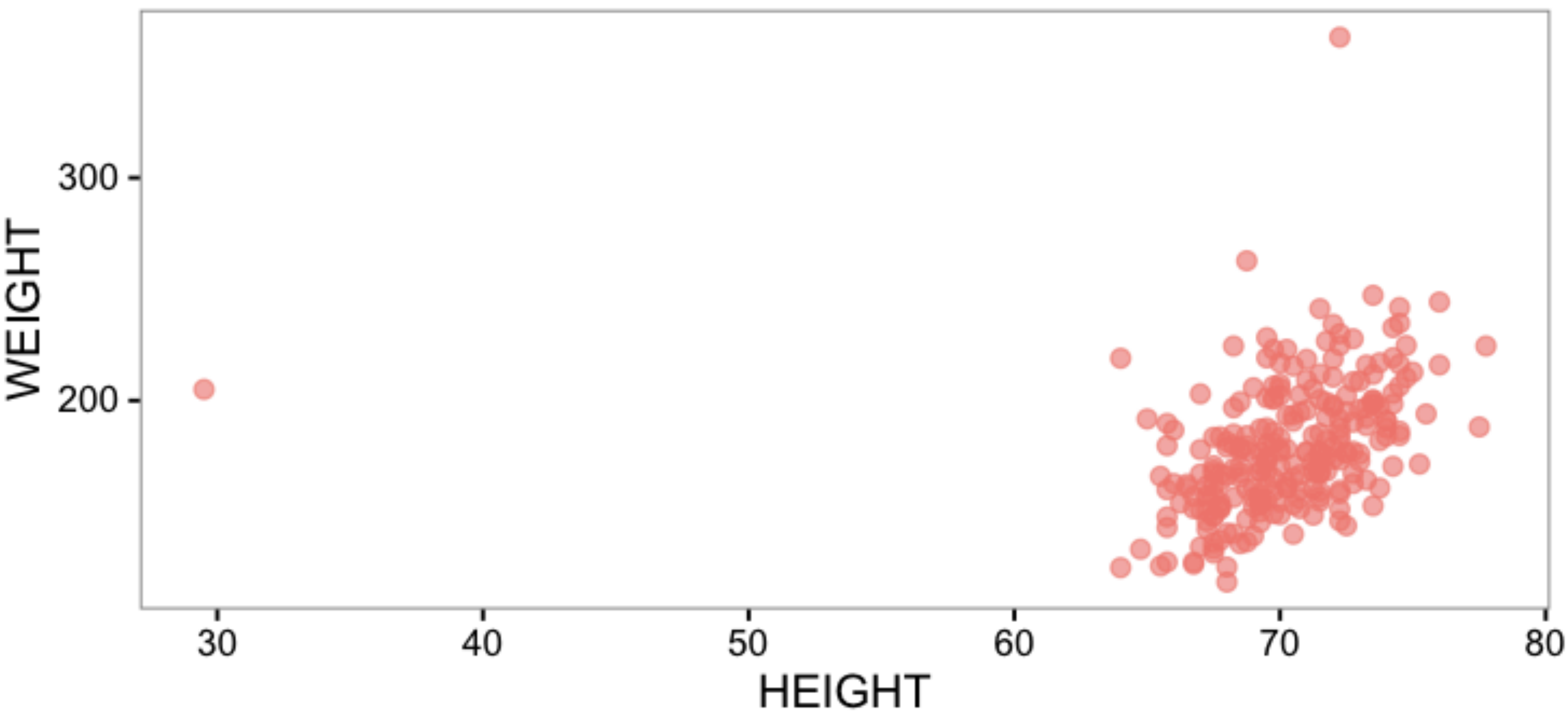
- Example: does the weight of people relate to their height?

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT
1	12.6	1.0708	23	154.25	67.75
2	6.9	1.0853	22	173.25	72.25
3	24.6	1.0414	22	154.00	66.25
4	10.9	1.0751	26	184.75	72.25
5	27.8	1.0340	24	184.25	71.25
6	20.6	1.0502	24	210.25	74.75
7	19.0	1.0549	26	181.00	69.75
8	12.8	1.0704	25	176.00	72.50
9	5.1	1.0900	25	191.00	74.00
10	12.0	1.0722	23	198.25	73.50

- x : HIGHT, y: WEIGHT

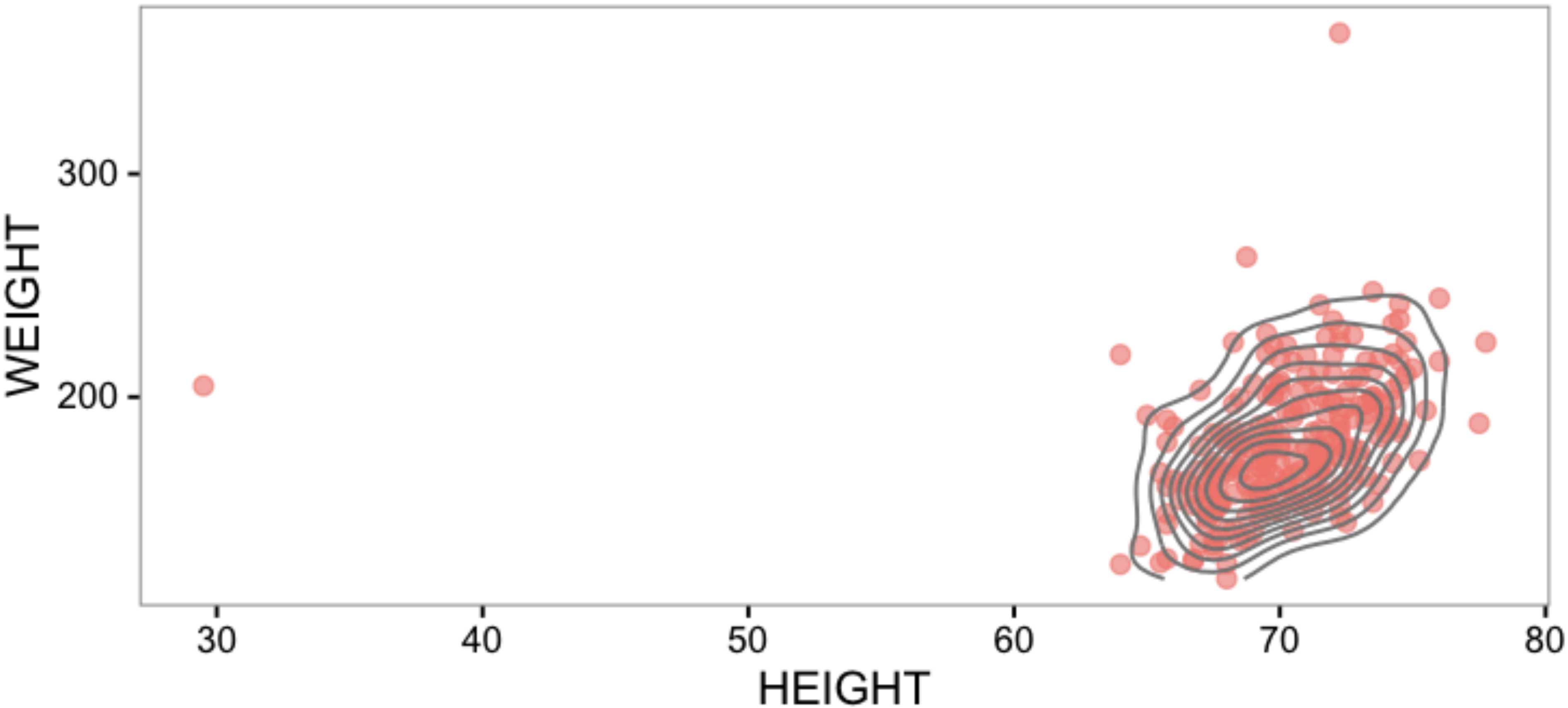
Scatter plot

- Body Fat data set



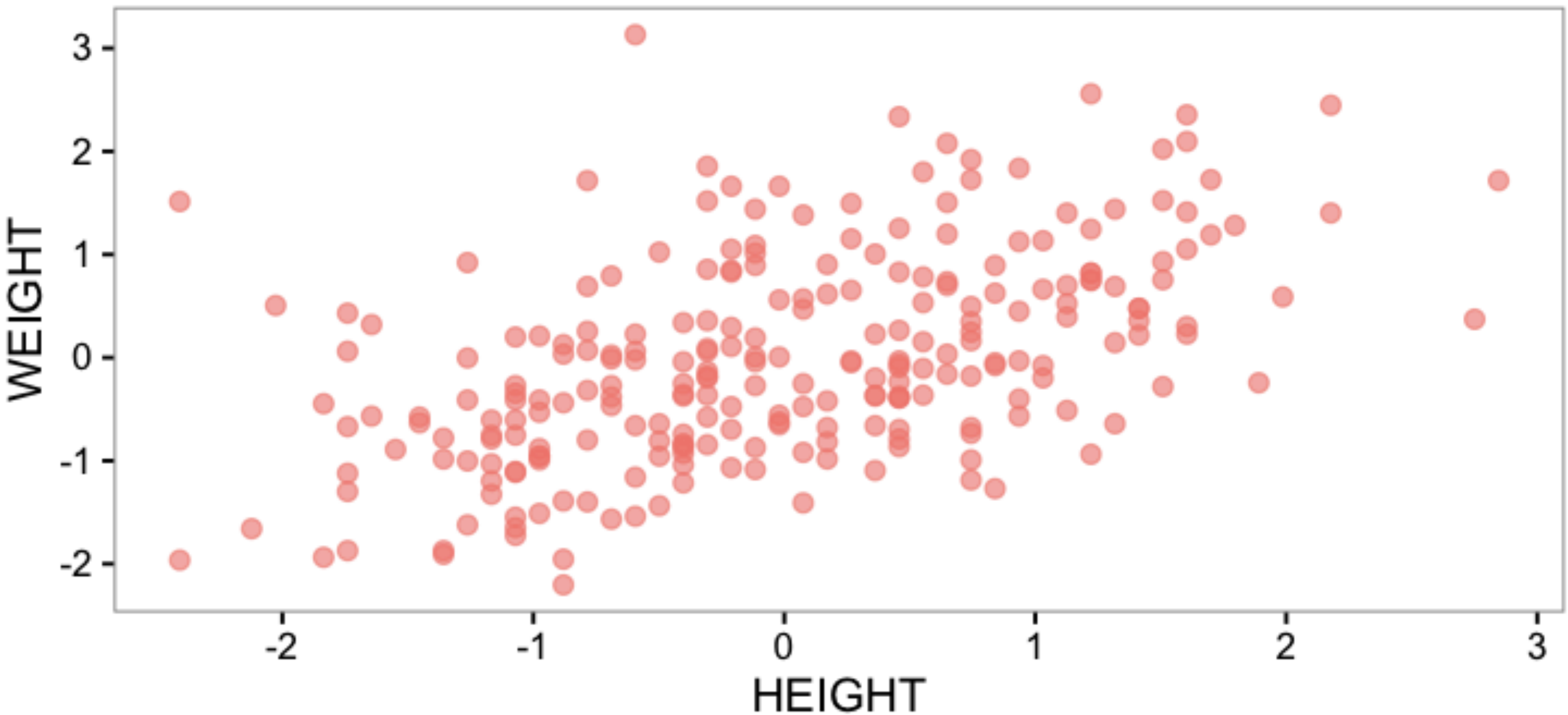
Scatter plot

- Scatter plot with density



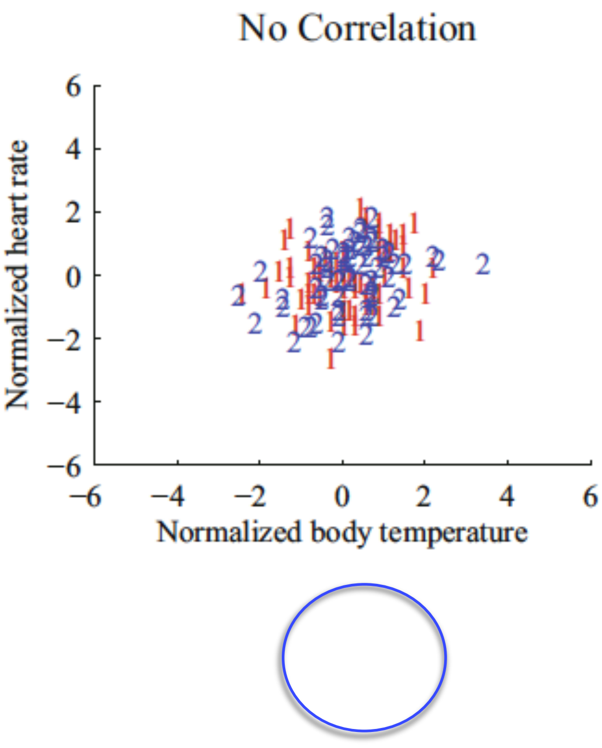
Scatter plot

- Removed of outliers & standardized

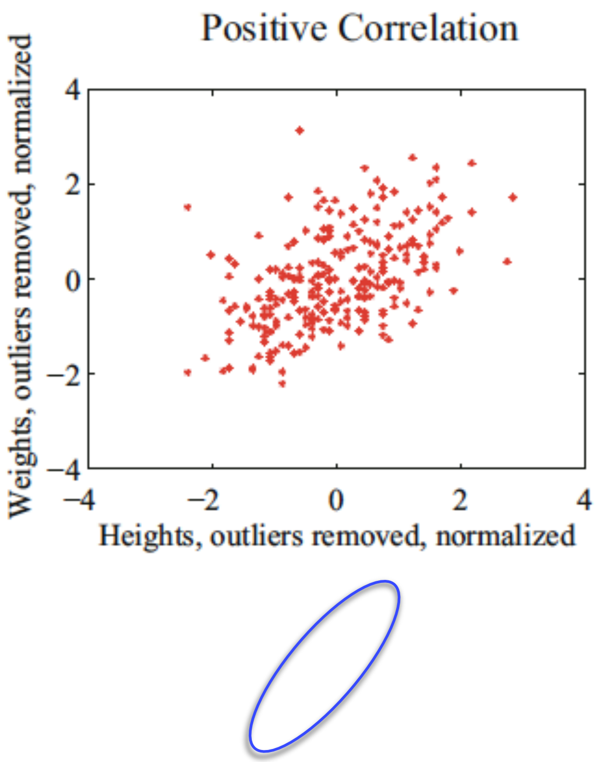


Correlation seen from scatter plots

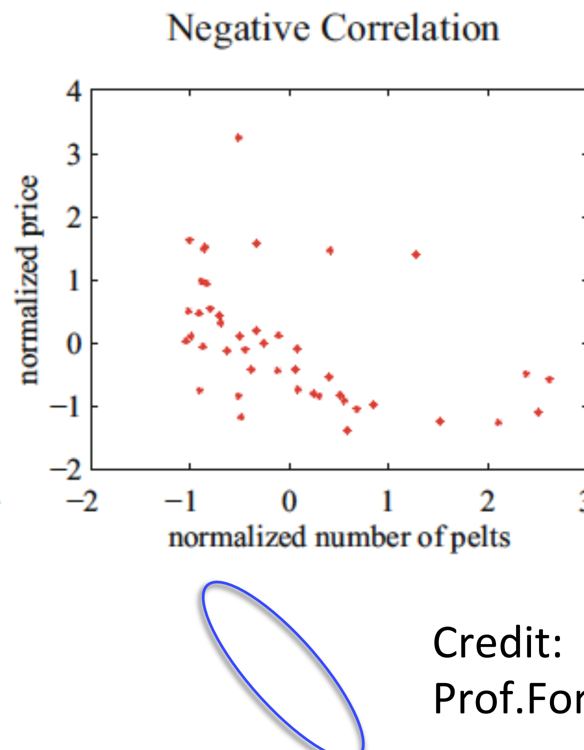
Zero Correlation



Positive correlation



Negative correlation



Credit: Prof.Forsyth

What kind of Correlation?

- Line of code in a database and number of bugs
- Frequency of hand washing and number of germs on your hands
- GPA and hours spent playing video games
- earnings and happiness

Credit: Prof. David Varodayan

Correlation doesn’t mean causation

- Shoe size is correlated to reading skills, but it doesn’t mean making feet grow will make one person read faster.

Correlation Coefficient

- Given a data set $\{(x_i, y_i)\}$ consisting of items (x_1, y_1) ... (x_N, y_N),
  - Standardize the coordinates of each feature:

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})} \qquad \widehat{y_i} = \frac{y_i - mean(\{y_i\})}{std(\{y_i\})}$$

  - Define the correlation coefficient as:

$$corr(\{(x_i, y_i)\}) = \frac{1}{N}\sum_{i=1}^{N} \widehat{x_i}\widehat{y_i}$$

Correlation Coefficient

$$\widehat{x_i} = \frac{x_i - mean(\{x_i\})}{std(\{x_i\})} \qquad \widehat{y_i} = \frac{y_i - mean(\{y_i\})}{std(\{y_i\})}$$

$$corr(\{(x_i, y_i)\}) = \frac{1}{N} \sum_{i=1}^{N} \widehat{x_i}\widehat{y_i}$$

$$= mean(\{\widehat{x_i}\widehat{y_i}\})$$

Q: Correlation Coefficient

- Which of the following describe(s) correlation coefficient correctly?

  A. It's unitless

  B. It's defined in standard coordinates

  C. Both A & B

$$corr(\{(x_i, y_i)\}) = \frac{1}{N}\sum_{i=1}^{N} \widehat{x}_i \widehat{y}_i$$

A visualization of correlation coefficient

https://rpsychologist.com/d3/correlation/

In a data set $\{(x_i, y_i)\}$ consisting of items (x_1, y_1) ... (x_N, y_N),

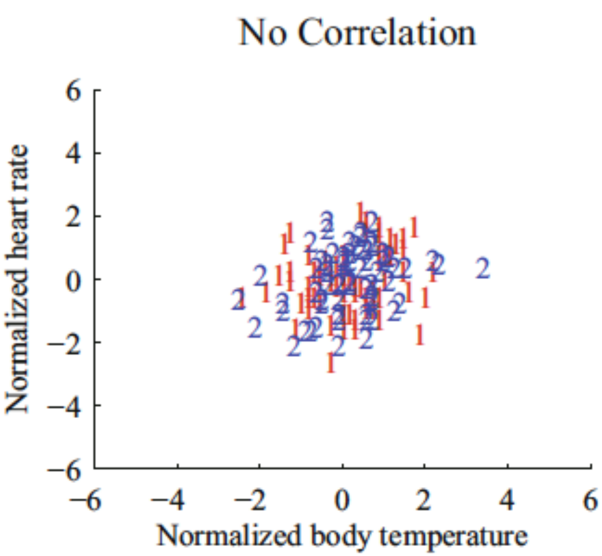
$corr(\{(x_i, y_i)\}) > 0$ shows positive correlation

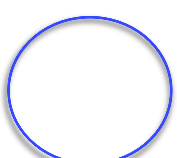
$corr(\{(x_i, y_i)\}) < 0$ shows negative correlation

$corr(\{(x_i, y_i)\}) = 0$ shows no correlation

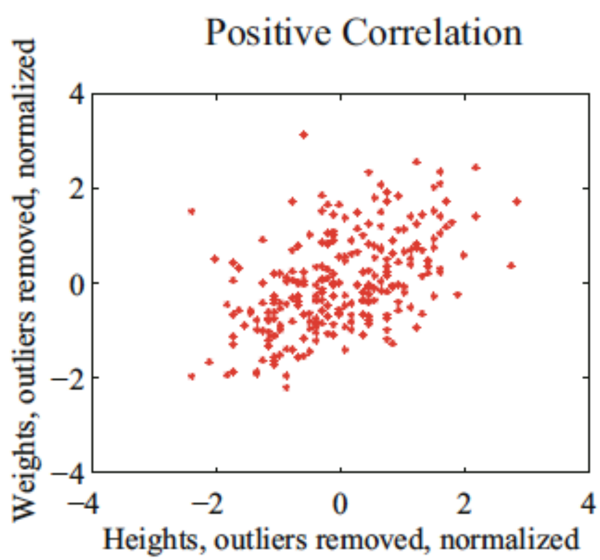
Correlation seen from scatter plots

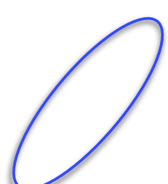
Zero Correlation



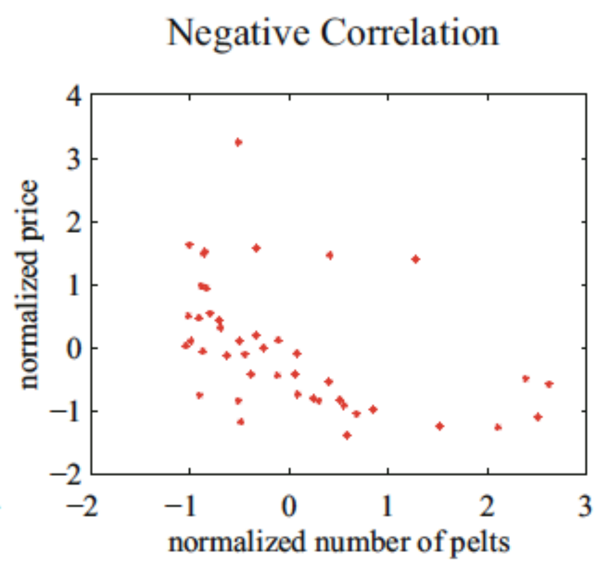


Positive correlation





Negative correlation



Credit: Prof.Forsyth

The Properties of Correlation Coefficient

- The correlation coefficient is symmetric

$$corr(\{(x_i, y_i)\}) = corr(\{(y_i, x_i)\})$$

- Translating the data does **NOT** change the correlation coefficient

The Properties of Correlation Coefficient

- Scaling the data may change the sign of the correlation coefficient

$$corr(\{(a\ x_i + b,\ c\ y_i + d)\}) = sign(a\ c)corr(\{(x_i, y_i)\})$$

The Properties of Correlation Coefficient

- The correlation coefficient is bounded within [-1, 1]

$corr(\{(x_i, y_i)\}) = 1$ if and only if $\widehat{x_i} = \widehat{y_i}$

$corr(\{(x_i, y_i)\}) = -1$ if and only if $\widehat{x_i} = -\widehat{y_i}$

Concept of Correlation Coefficient's bound

- The correlation coefficient can be written as $$corr(\{(x_i, y_i)\}) = \frac{1}{N}\sum_{i=1}^{N}\widehat{x_i}\widehat{y_i}$$

$$corr(\{(x_i, y_i)\}) = \sum_{i=1}^{N}\frac{\widehat{x_i}}{\sqrt{N}}\frac{\widehat{y_i}}{\sqrt{N}}$$

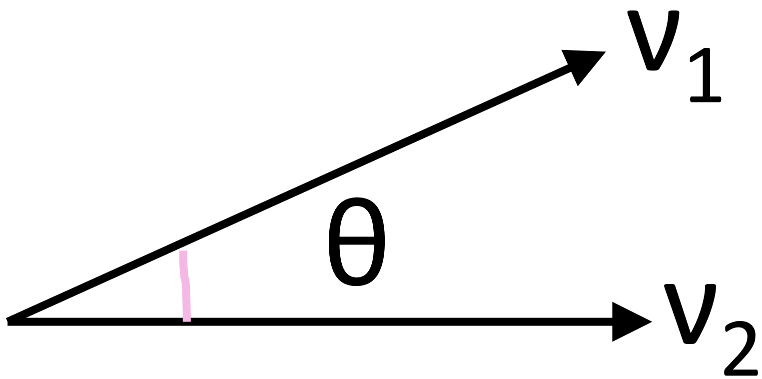
- It's the inner product of two vectors

$$\left\langle \frac{\widehat{x_1}}{\sqrt{N}}, \quad \ldots \quad \frac{\widehat{x_N}}{\sqrt{N}} \right\rangle \text{ and } \left\langle \frac{\widehat{y_1}}{\sqrt{N}}, \quad \ldots \quad \frac{\widehat{y_N}}{\sqrt{N}} \right\rangle$$

Inner product

- Inner product's geometric meaning:

$$|\nu_1|\,|\nu_2|\,cos(\theta)$$



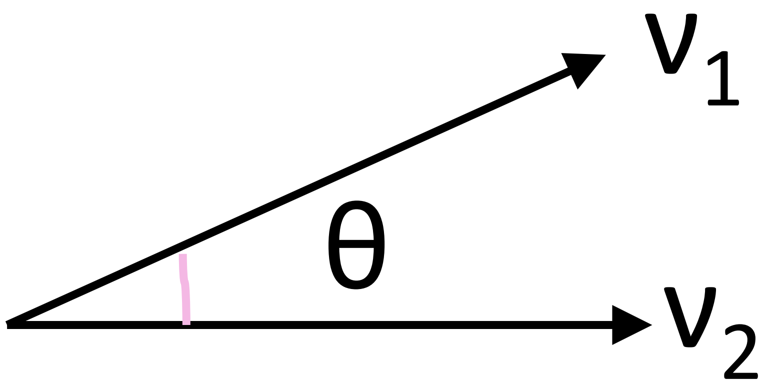
- Lengths of both vectors

$$v_1 = \left\langle \frac{\widehat{x_1}}{\sqrt{N}}, \quad \cdots \quad \frac{\widehat{x_N}}{\sqrt{N}} \right\rangle \qquad v_2 = \left\langle \frac{\widehat{y_1}}{\sqrt{N}}, \quad \cdots \quad \frac{\widehat{y_N}}{\sqrt{N}} \right\rangle$$

are 1

Bound of correlation coefficient

$$|corr(\{(x_i, y_i)\})| = |cos(\theta)| \leq 1$$



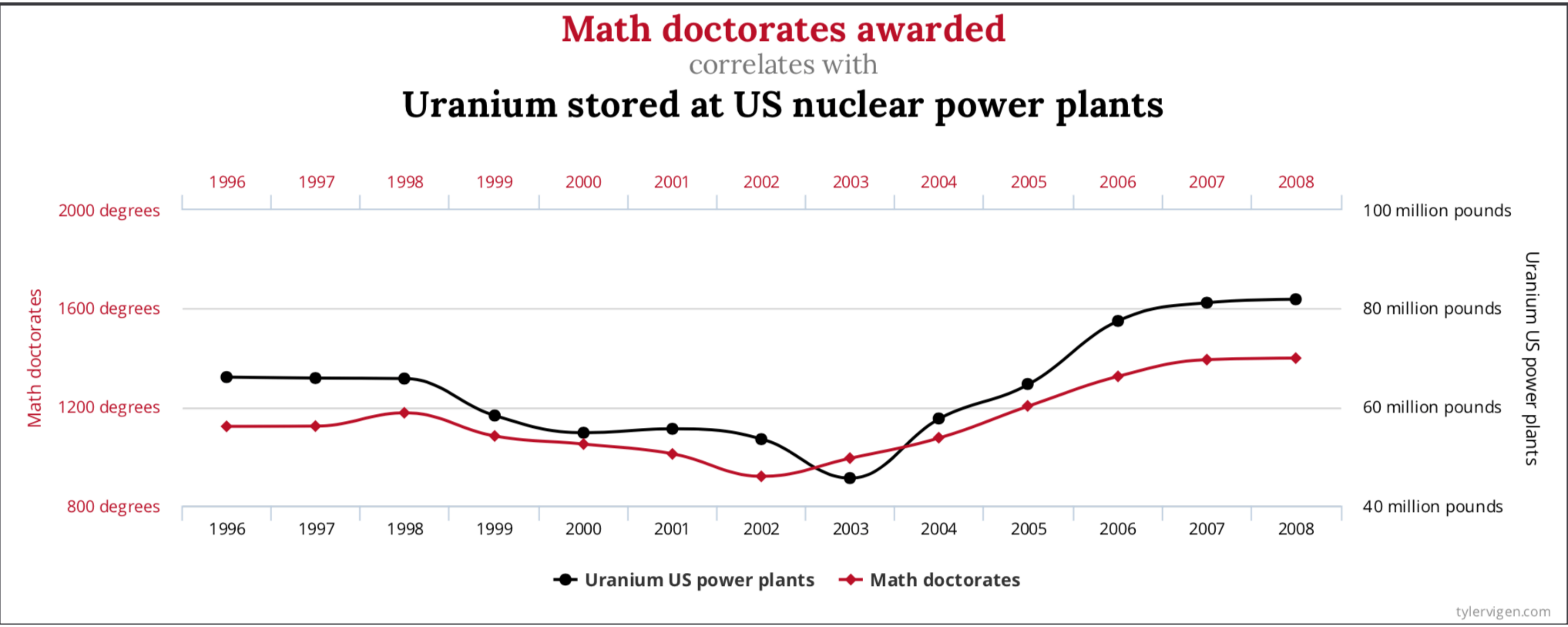
$$v_1 = \left\langle \frac{\widehat{x_1}}{\sqrt{N}}, \quad \cdots \quad \frac{\widehat{x_N}}{\sqrt{N}} \right\rangle \quad v_2 = \left\langle \frac{\widehat{y_1}}{\sqrt{N}}, \quad \cdots \quad \frac{\widehat{y_N}}{\sqrt{N}} \right\rangle$$

The Properties of Correlation Coefficient

- Symmetric
- Translating invariant
- Scaling only may change sign
- bounded within [-1, 1]

Using correlation to predict

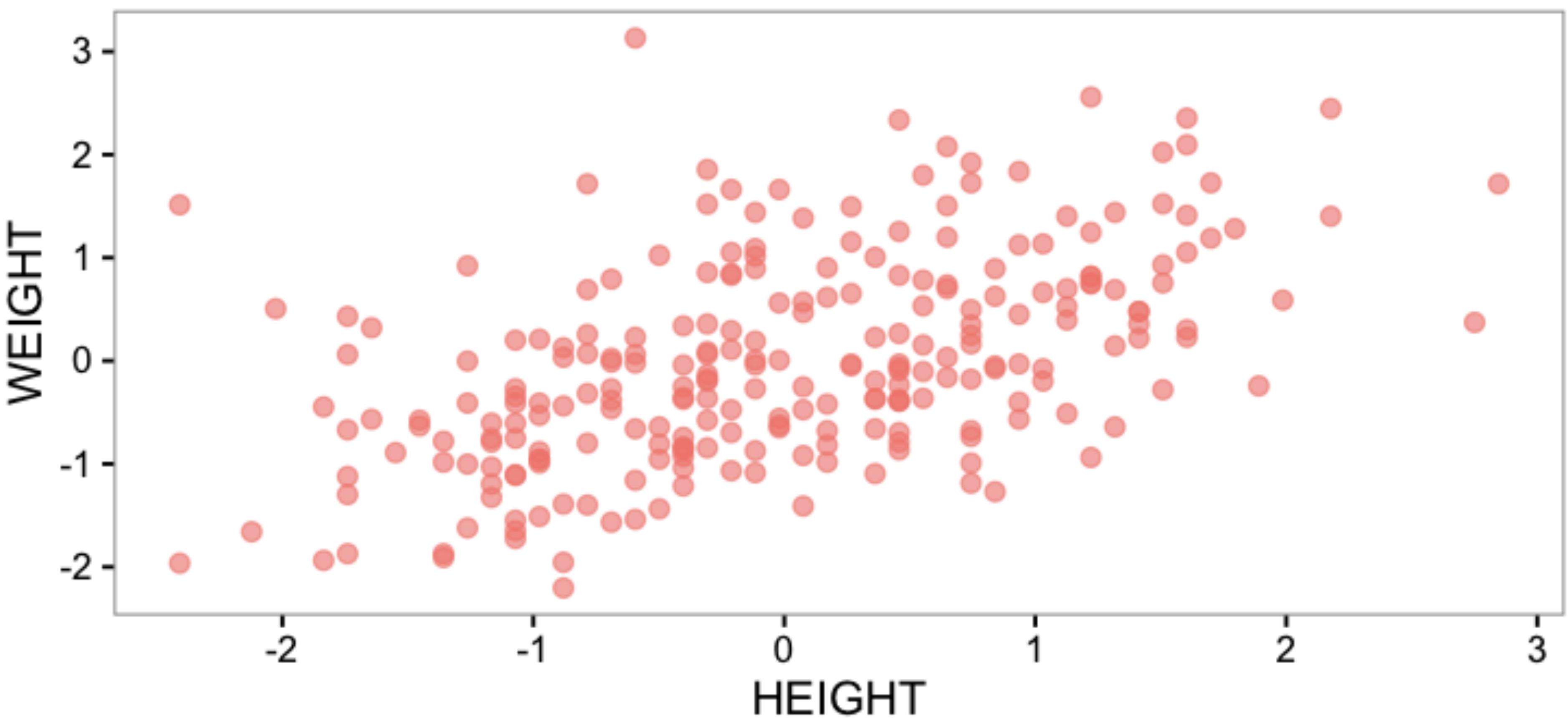
- **Caution**! Correlation is **NOT** Causation



Credit: Tyler Vigen

How do we go about the prediction?

- Removed of outliers & standardized



Using correlation to predict

- Given a correlated data set $\{(x_i, y_i)\}$

  we can predict a value ${y_0}^p$ that goes with a value x_0

- In standard coordinates $\{(\widehat{x}_i, \widehat{y}_i)\}$

  we can predict a value ${\widehat{y_0}}^p$ that goes with a value $\widehat{x_0}$

Q:

- Which coordinates will you use for the predictor using correlation?

  A. Standard coordinates

  B. Original coordinates

  C. Either

Linear predictor and its error

- We will assume that our predictor is linear

$$\widehat{y}^{\,p} = a\ \widehat{x} + b$$

- We denote the prediction at each $\widehat{x_i}$ in the data set as $\widehat{y_i}^p$

$$\widehat{y_i}^{\,p} = a\ \widehat{x_i} + b$$

- The error in the prediction is denoted u_i

$$u_i = \widehat{y_i} - \widehat{y_i}^{\,p} = \widehat{y_i} - a\ \widehat{x_i} - b$$

Require the mean of error to be zero

We would try to make the mean of error equal to zero so that it is also centered around 0 as the standardized data:

Require the variance of error is minimal

Require the variance of error is minimal

Here is the linear predictor!

$$\widehat{y}^{\,p} = r\ \widehat{x}$$

↓

Correlation coefficient

Prediction Formula

- In standard coordinates

$$\widehat{y_0}^{\,p} = r\ \widehat{x_0} \quad \text{where} \quad r = corr(\{(x_i, y_i)\})$$

- In original coordinates

$$\frac{y_0^p - mean(\{y_i\})}{std(\{y_i\})} = r\frac{x_0 - mean(\{x_i\})}{std(\{x_i\})}$$

Root-mean-square (RMS) prediction error

- $$Given \qquad var(\{u_i\}) = 1 - 2ar + a^2$$
  $$\& \qquad a = r$$

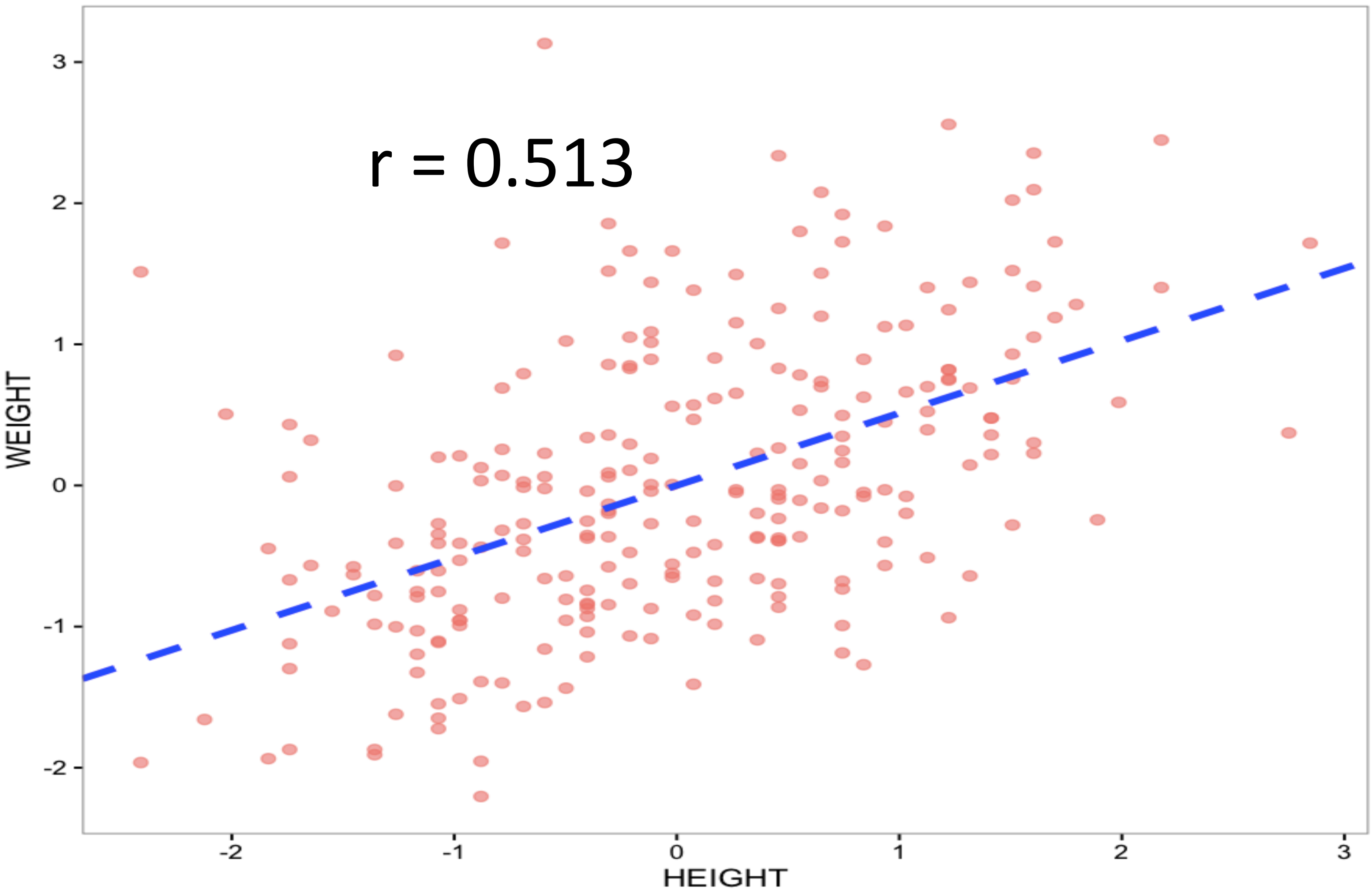
  $$var(\{u_i\}) = 1 - r^2$$

- $$RMS \;\; error = \sqrt{mean(\{u_i^2\})} = \sqrt{var(\{u_i\})} = \sqrt{1 - r^2}$$

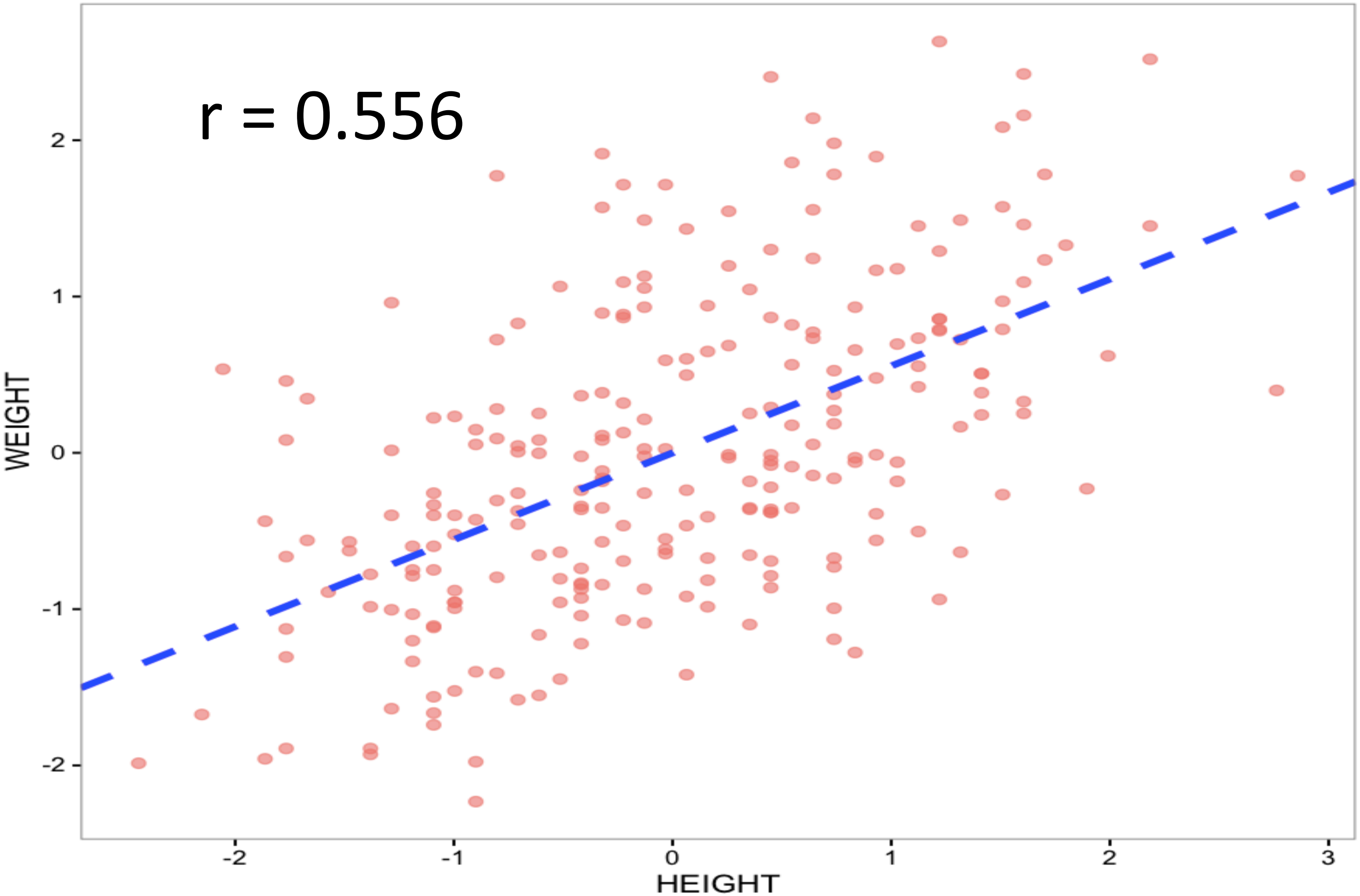
See the error through simulation

https://rpsychologist.com/d3/correlation/

Example: Body Fat data



Example: remove 2 more outliers



Heatmap

- Display matrix of data via gradient of color(s)

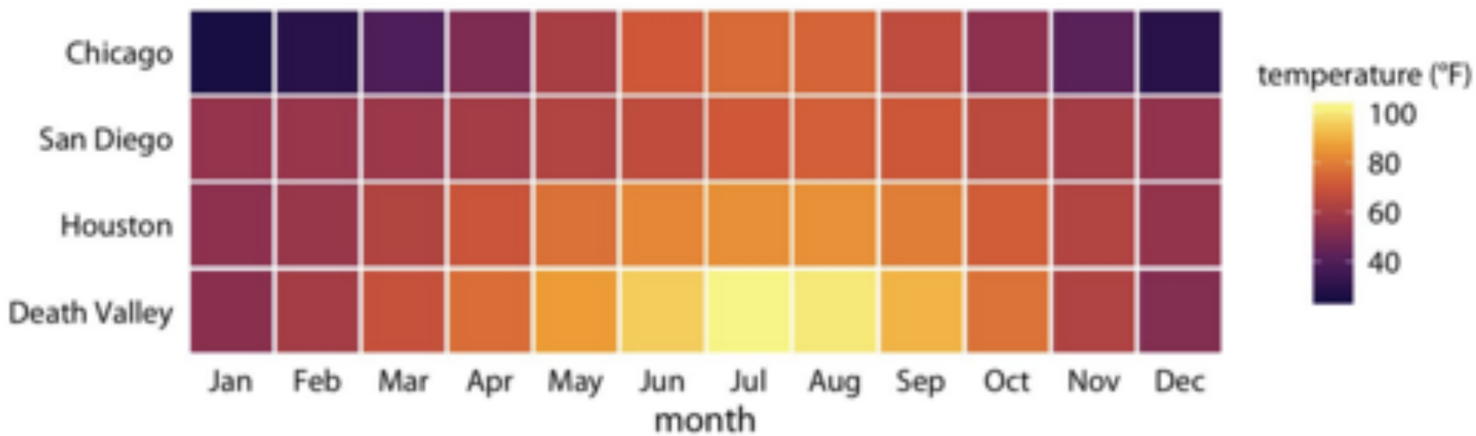
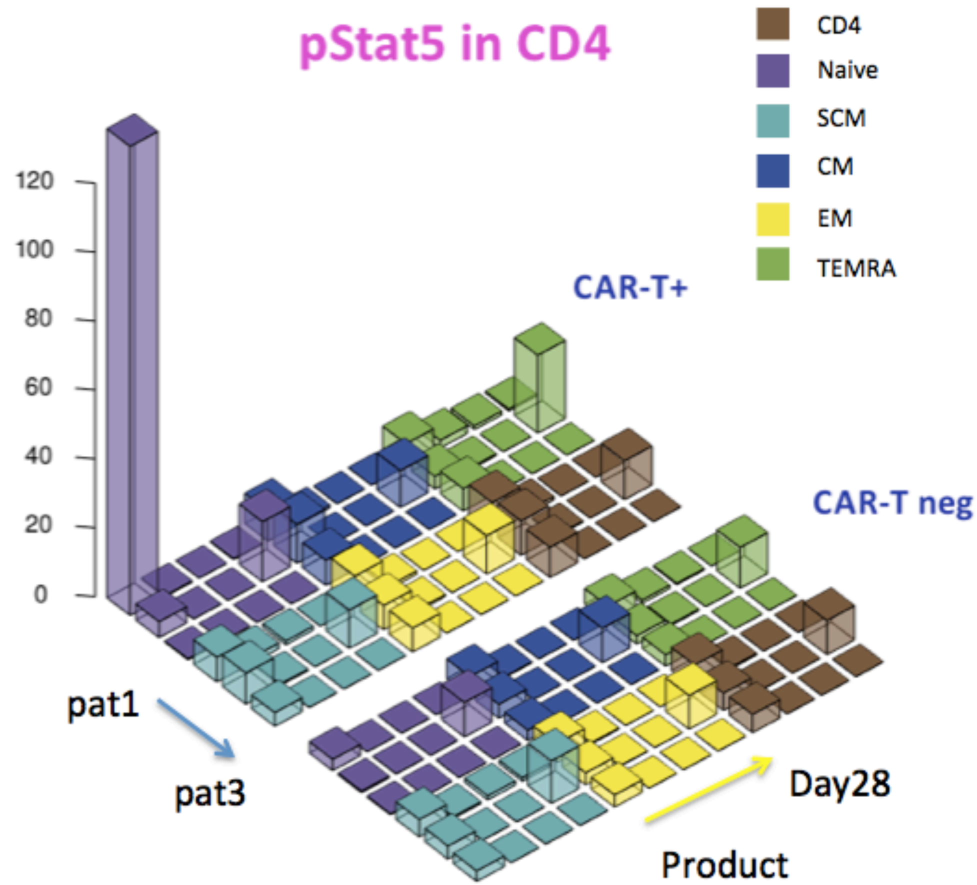


Figure 2-4. Monthly normal mean temperatures for four locations in the US. Data source: NOAA.

Summarization of 4 locations' annual mean temperature by month

3D bar chart

- Transparent 3D bar chart is good for small # of samples across categories



Relationship between data feature and time

- Example: How does Amazon's stock change over 1 years?

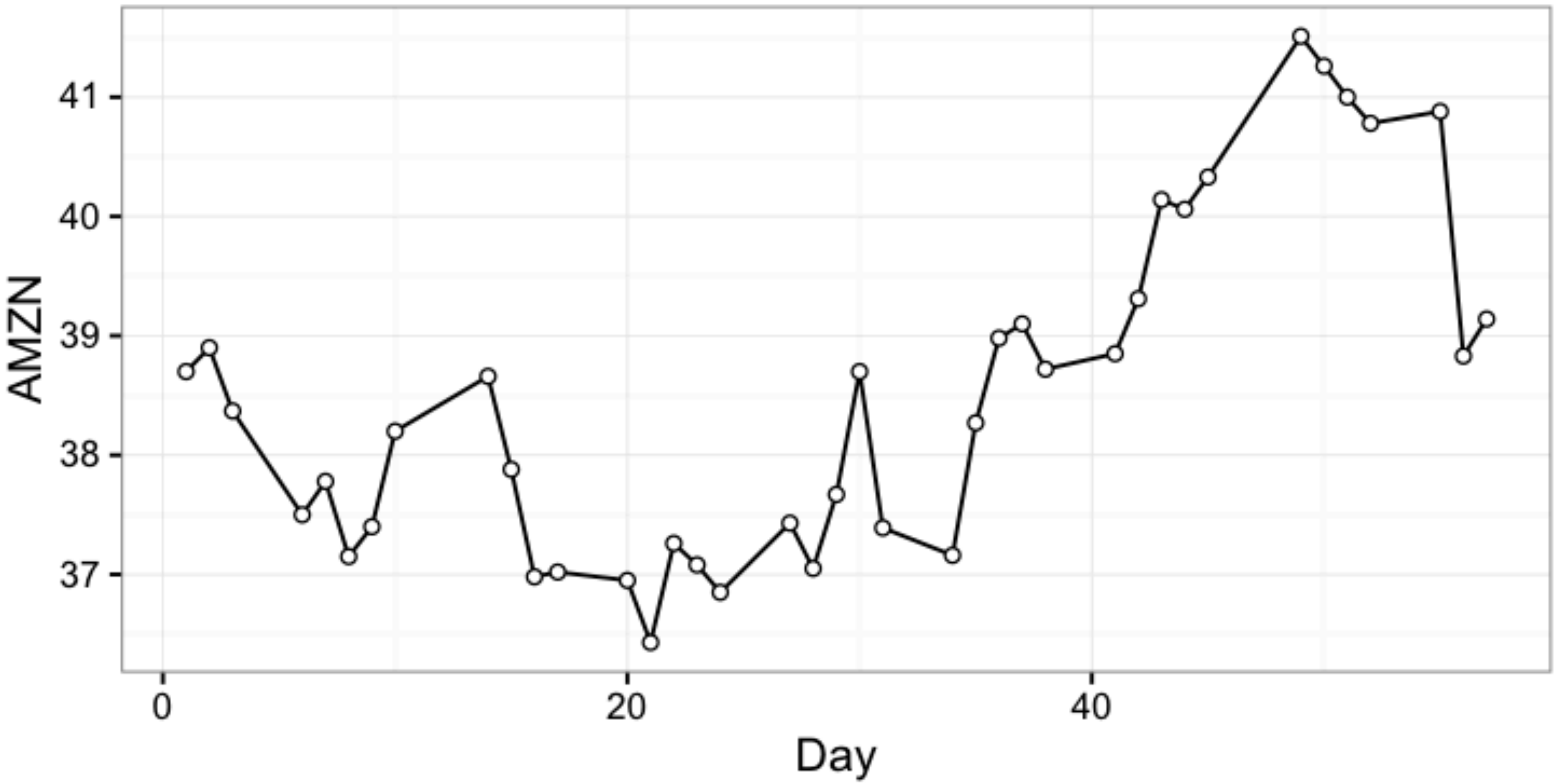
  take out the pair of features

  x: Day

  y: AMZN

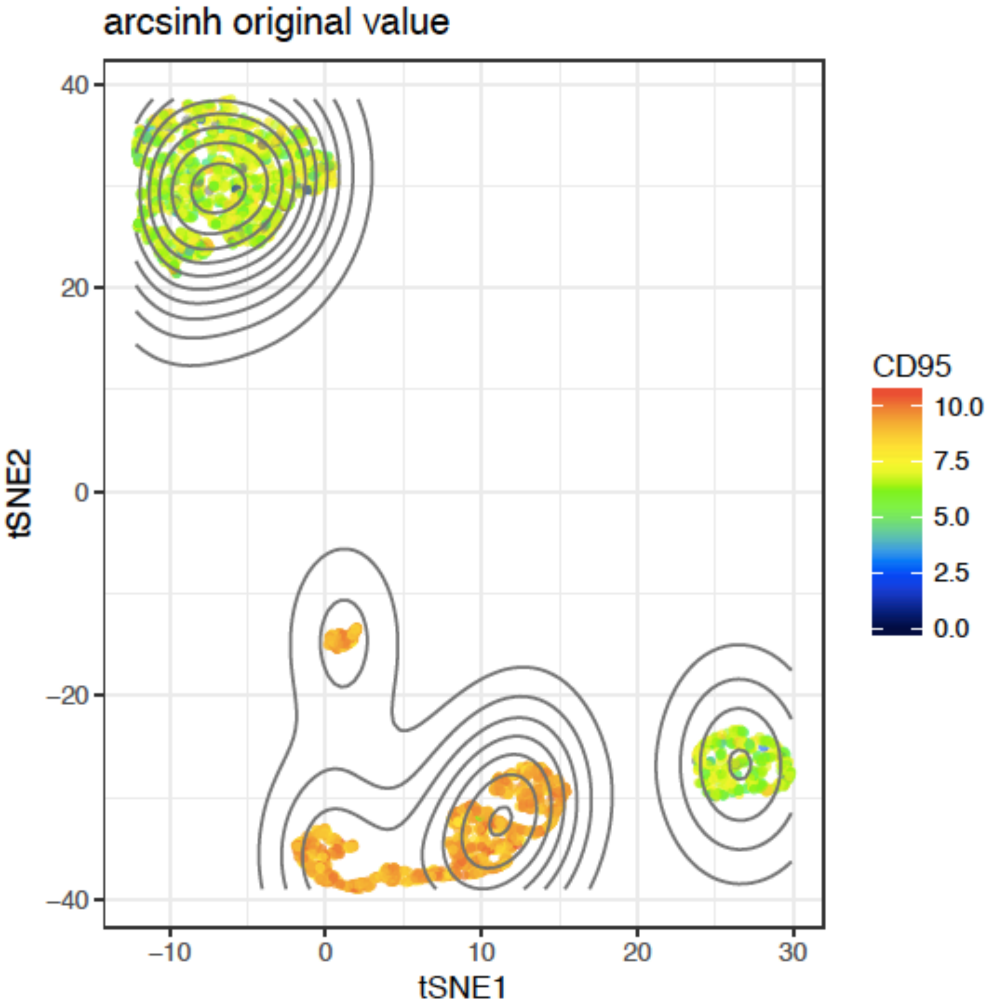
Day	AMZN	DUK	KO
1	38.700001	34.971017	17.874906
2	38.900002	35.044103	17.882263
3	38.369999	34.240172	17.757161
6	37.5	34.294985	17.871225
7	37.779999	34.130544	17.885944
8	37.150002	33.984374	17.9117
9	37.400002	34.075731	17.933777
10	38.200001	33.91129	17.863866
14	38.66	34.020917	17.845469
15	37.880001	33.966104	17.882263
16	36.98	34.130544	17.790276
17	37.02	34.240172	17.757161
20	36.950001	34.057458	17.672533
21	36.43	34.112272	17.705649
22	37.259998	34.258442	17.709329
23	37.080002	34.569051	17.639418
24	36.849998	34.861392	17.598945

Time Series Plot: Stock of Amazon



Scatter plot

- Coupled with heatmap to show a 3rd feature



Assignments

- Finish reading Chapter 2 of the textbook
- Next time: Probability a first look

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability"
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics"

See you next time

See You!