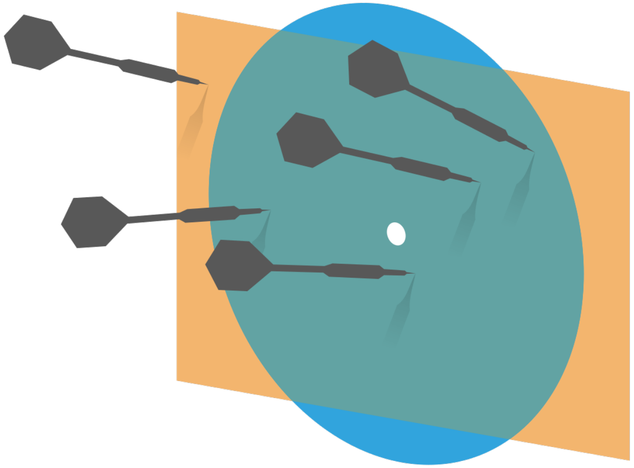
Probability and Statistics for Computer Science



Credit: wikipedia

"The statement that "The average US family has 2.6 children" invites mockery" – Prof. Forsyth reminds us about critical thinking

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 8.27.2021

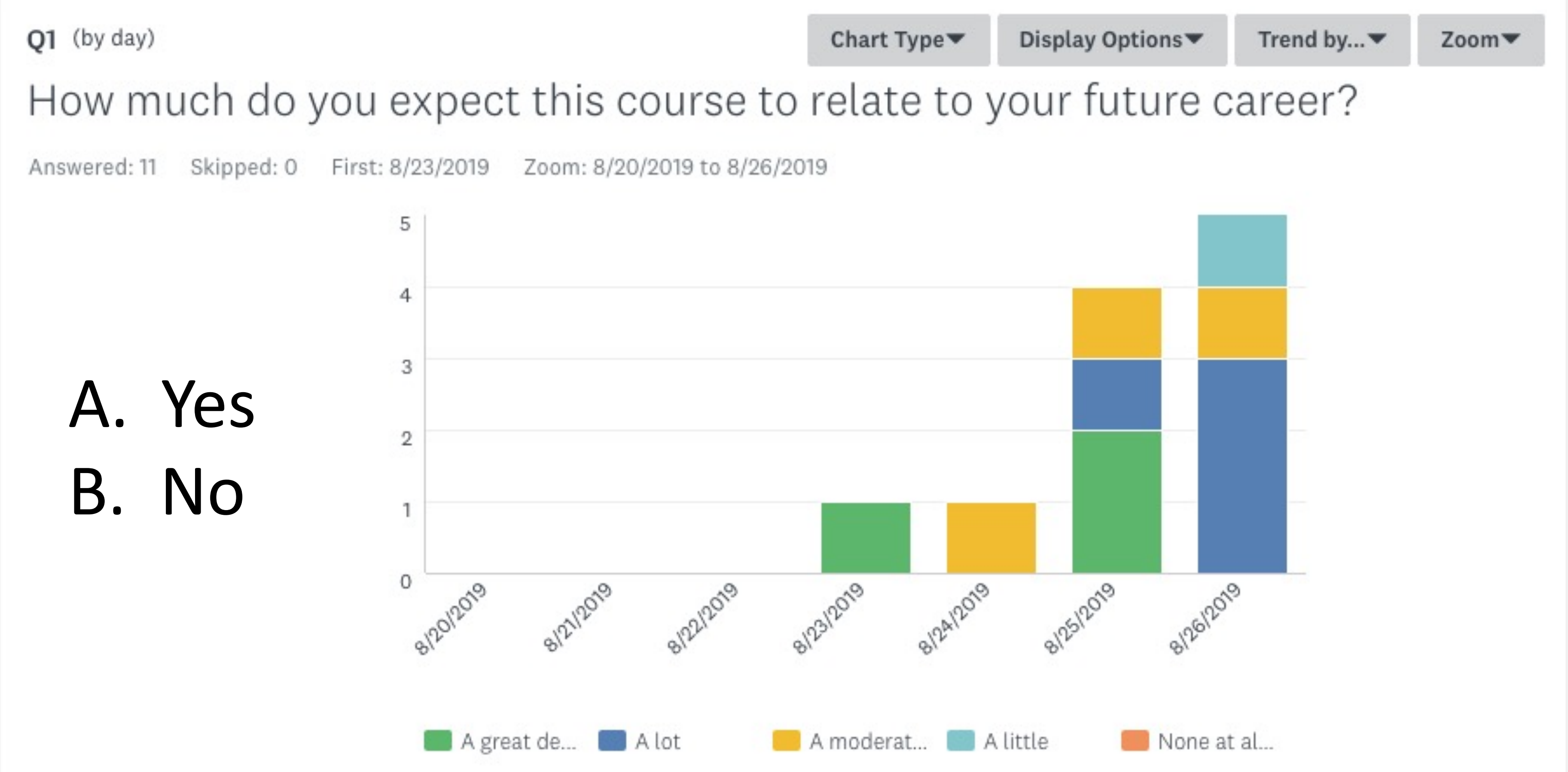
Last lecture

- Welcome/Orientation
- Big picture of the contents
- Lecture 1 - Data Visualization & Summary (I)
- **Orientation quiz due today**

Warm up question:

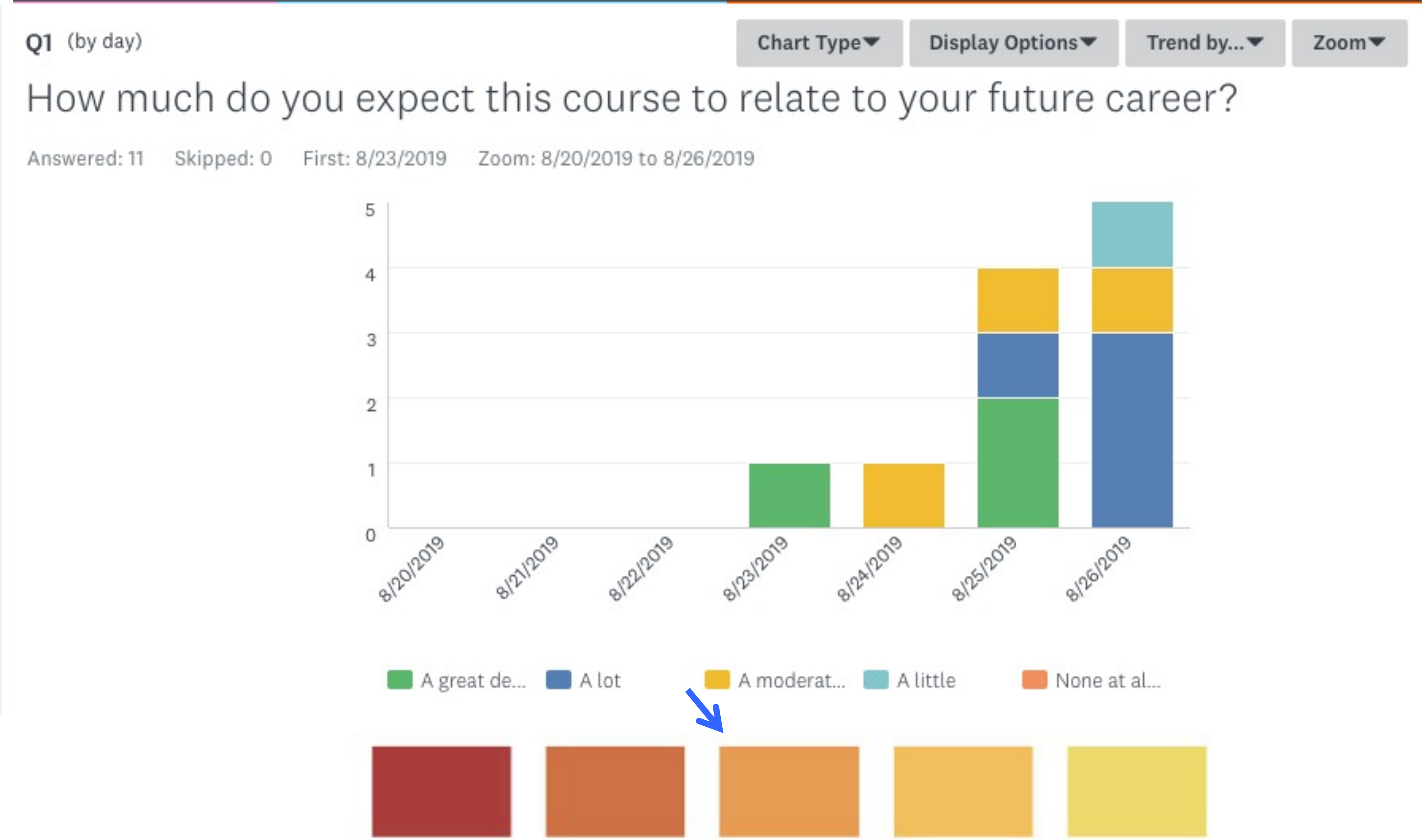
- What kind of data is a letter grade?
- What do you ask for usually about the stats of an exam with numerical scores?

Q: Is this a good bar chart?



A. Yes
B. No

How about using a color scale



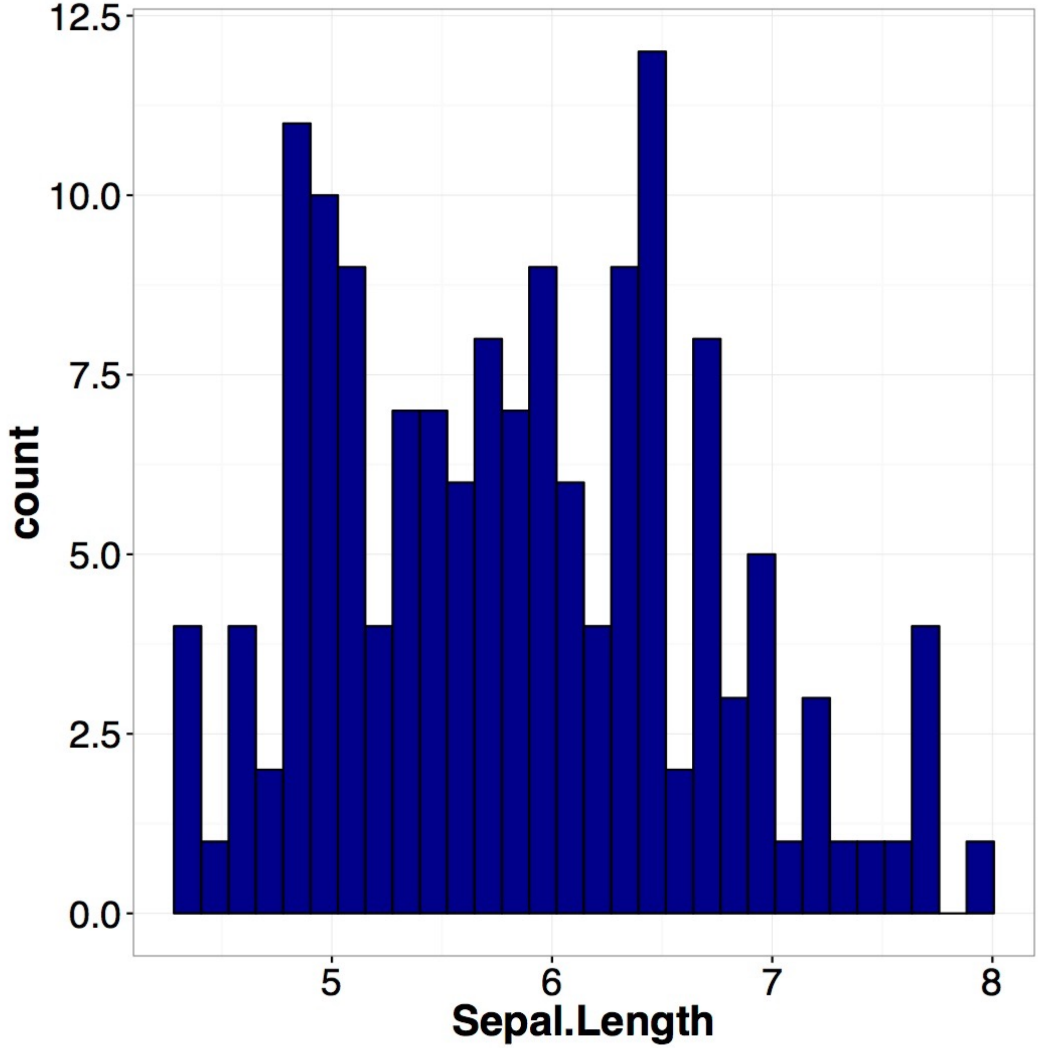
Objectives

- **Histograms**
- **Grasp** Summary Statistics
- Learn more Data Visualization for **Relationships**

Visualizing Data with Histogram

- Histogram

A set of bars that are organized by bins that contains numerical data (discrete or continuous)



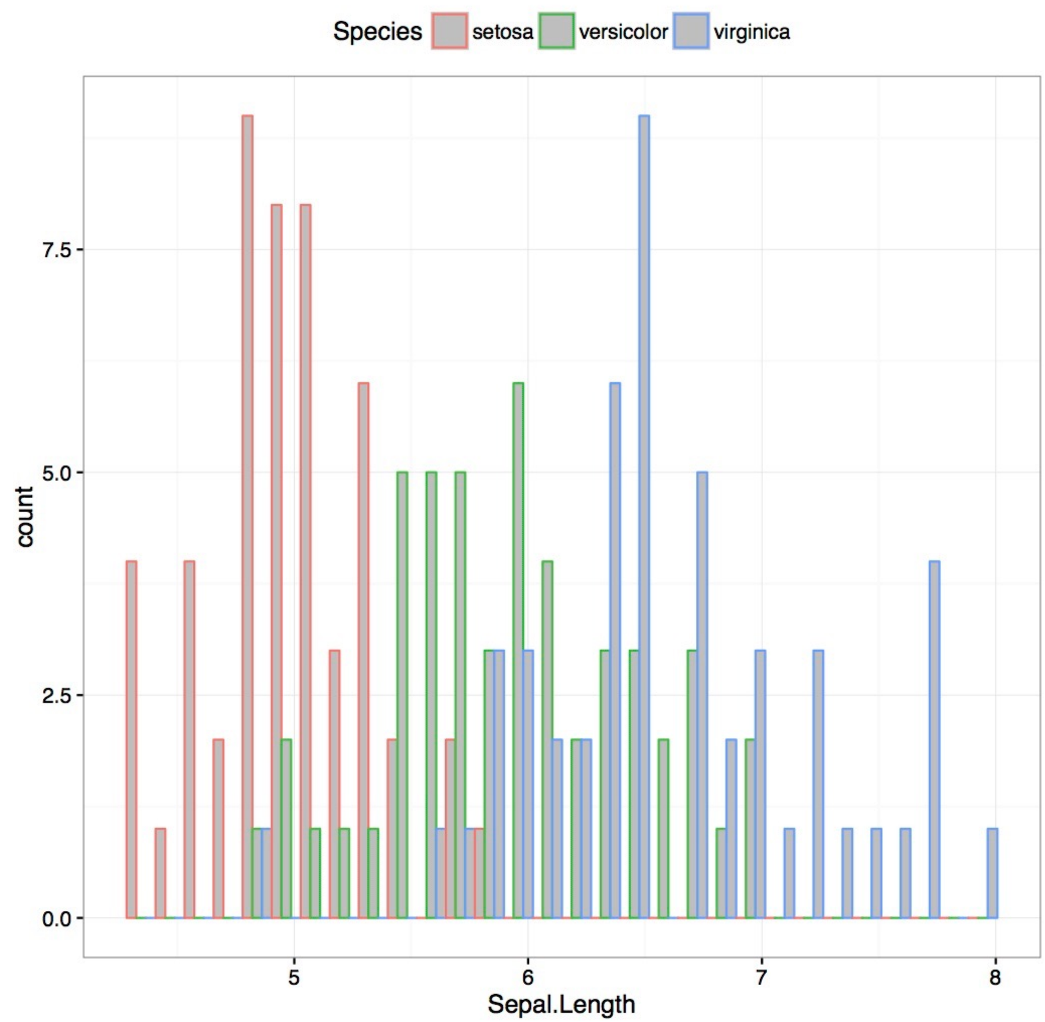
Data: “iris”

Visualizing Data with Histogram (II)

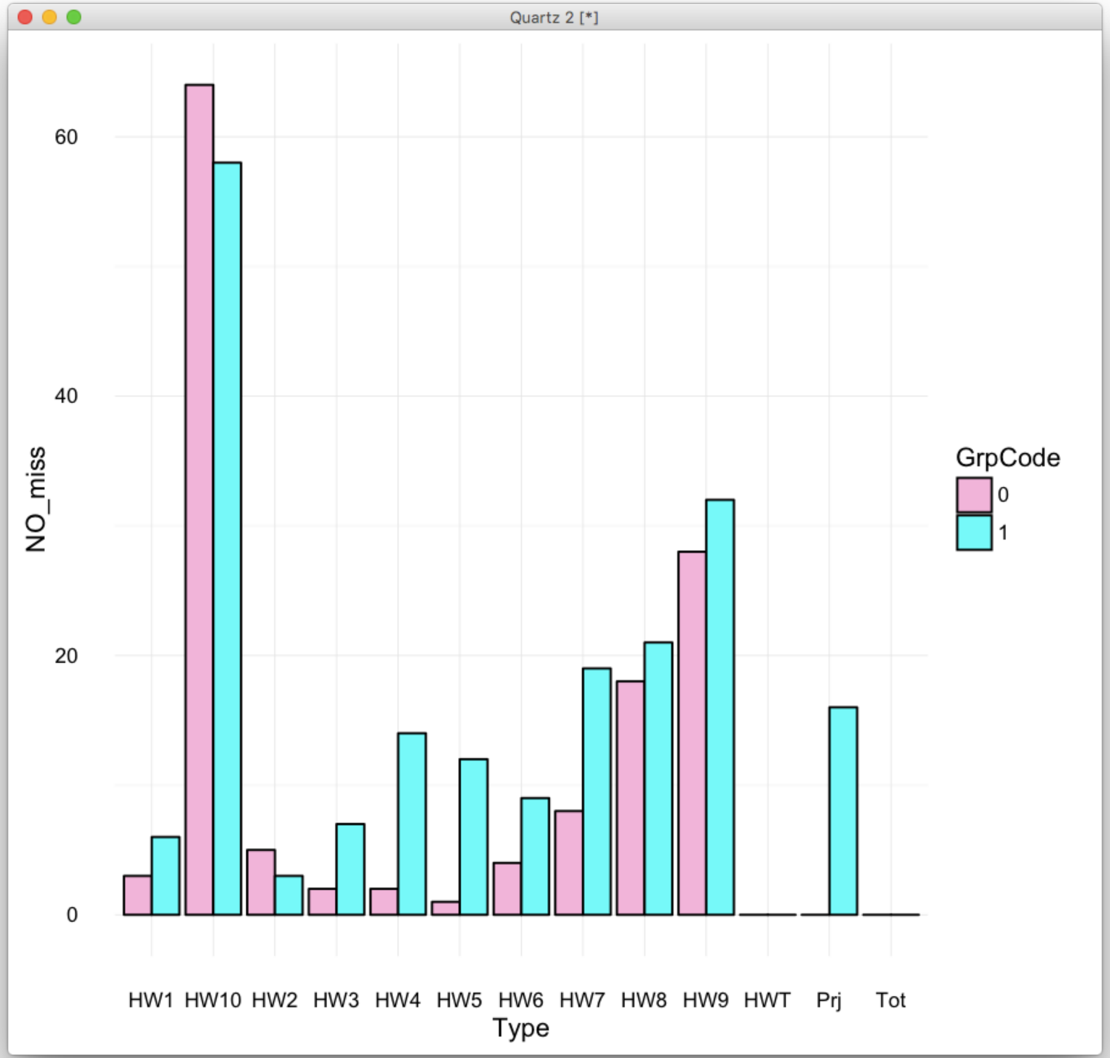
- Conditional histogram

Histogram generated by subsets of the data

Data: “iris”



Which group has the higher total scores?



Summarizing 1D continuous data

For a data set {x} or annotated as $\{x_i\}$, we summarize with:

- Location Parameters

- Scale parameters

Summarizing 1D continuous data

- Mean

$$mean(\{x_i\}) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

It’s the centroid of the data geometrically,
by identifying the data set at that point, you find
the center of balance.

Properties of the mean

- Scaling data scales the mean

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

- Translating the data translates the mean

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

Less obvious properties of the mean

- The signed distances from the mean

  sum to 0
  $$\sum_{i=1}^{N}(x_i - mean(\{x_i\})) = 0$$

- The mean minimizes the sum of the squared distance from any real value

$$\underset{\mu}{argmin} \sum_{i=1}^{N}(x_i - \mu)^2 = mean(\{x_i\})$$

Q1:

- What is the answer for

  $mean(\{mean(\{x_i\})\})$?

  A. $mean(\{x_i\})$ B. unsure C. 0

Standard Deviation (σ)

- The standard deviation

$$std(\{x_i\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - mean(\{x_i\}))^2}$$

$$= \sqrt{mean(\{(x_i - mean(\{x_i\}))^2\})}$$

Q2. Can a standard deviation of a dataset be -1?

A. YES
B. NO

Properties of the standard deviation

- Scaling data scales the standard deviation

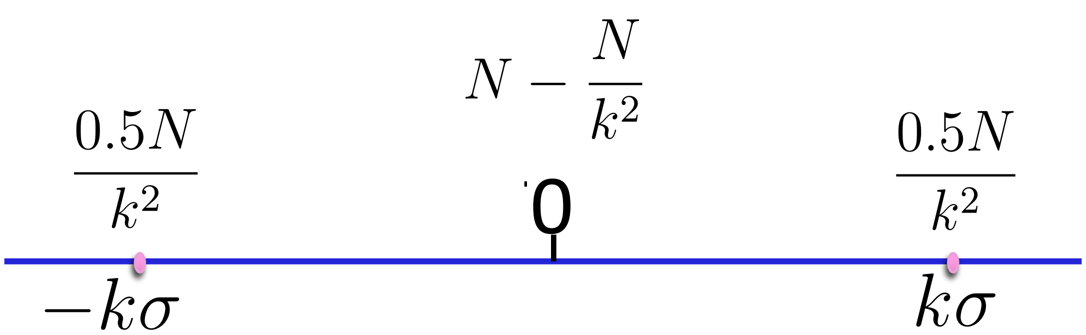
$$std(\{k \cdot x_i\}) = |k| \cdot std(\{x_i\})$$

- Translating the data does **NOT** change the standard deviation

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

Standard deviation: Chebyshev's inequality (1st look)

- At most $\frac{N}{k^2}$ items are k standard deviations (σ) away from the mean
- Rough justification: Assume mean =0



$$std = \sqrt{\frac{1}{N}[(N - \frac{N}{k})0^2 + \frac{N}{k^2}(k\sigma)^2]} = \sigma$$

Variance (σ^2)

- Variance = (standard deviation)^2

$$var(\{x_i\}) = \frac{1}{N}\sum_{i=1}^{N}(x_i - mean(\{x_i\}))^2$$

- Scaling and translating similar to standard deviation

$$var(\{k \cdot x_i\}) = k^2 \cdot var(\{x_i\})$$

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

Q3: Standard deviation

- What is the value of

  $std(\{mean(\{x_i\})\})$?

  A. 0 B. 1 C. unsure

Standard Coordinates/normalized data

- The *mean* tells where the data set is and the *standard deviation* tells how spread out it is. If we are interested only in comparing the shape, we could

  define:

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

- We say $\{\widehat{x_i}\}$ is in standard coordinates

Q4: Mean of standard coordinates

- mean($\{\widehat{x_i}\}$) is:

  A. 1 B. 0 C. unsure

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

Q5: Standard deviation (σ) of standard coordinates

- Std($\{\widehat{x_i}\}$) is:

  A. 1 B. 0 C. unsure

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

Q6: Variance of standard coordinates

- Variance of $\{\widehat{x_i}\}$ is:

  A. 1 B. 0 C. unsure

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

Q7: Estimate the range of data in standard coordinates

- Estimate as close as possible, 90% data is within:

  A. [-10, 10]

  B. [-100, 100]

  C. [-1, 1]

  D. [-4, 4]

  E. others

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

Standard Coordinates/normalized data to $\mu=0$, $\sigma=1$, $\sigma^2=1$

- Data in standard coordinates always has

  mean = 0; standard deviation =1;

  variance = 1.

- Such data is unit-less, plots based on this sometimes are more comparable

- We see such normalization very often in statistics

Median

- We first sort the data set $\{x_i\}$
- 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

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

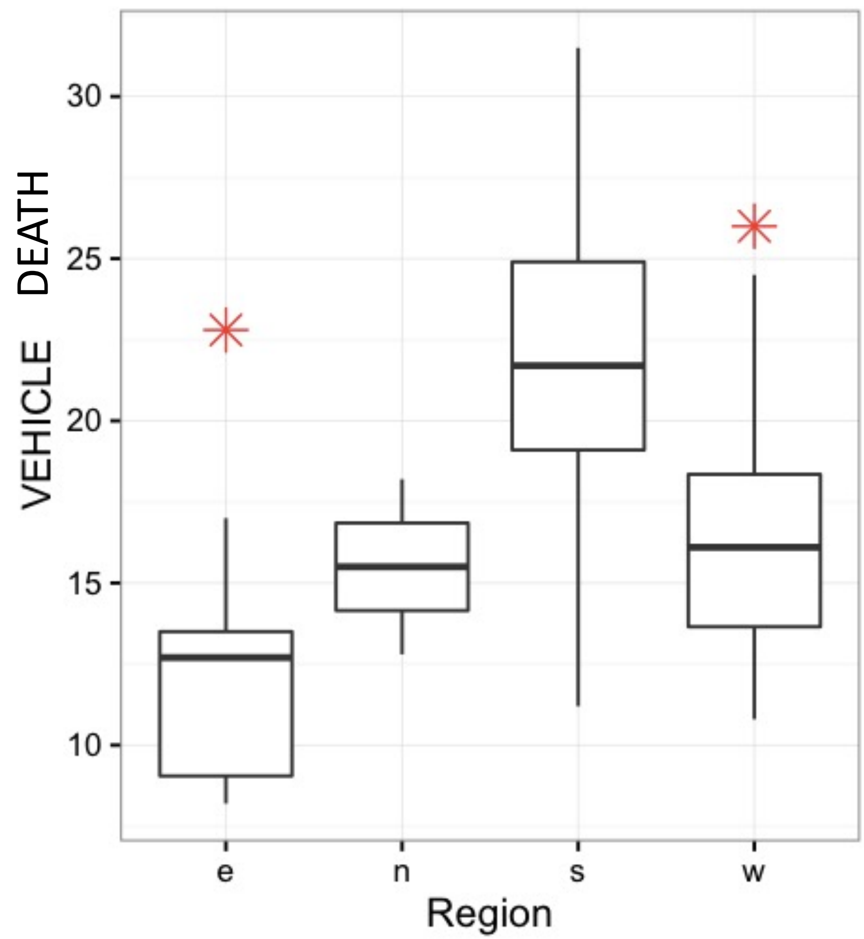
- Translating data does **NOT** change the interquartile range

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

Box plots

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

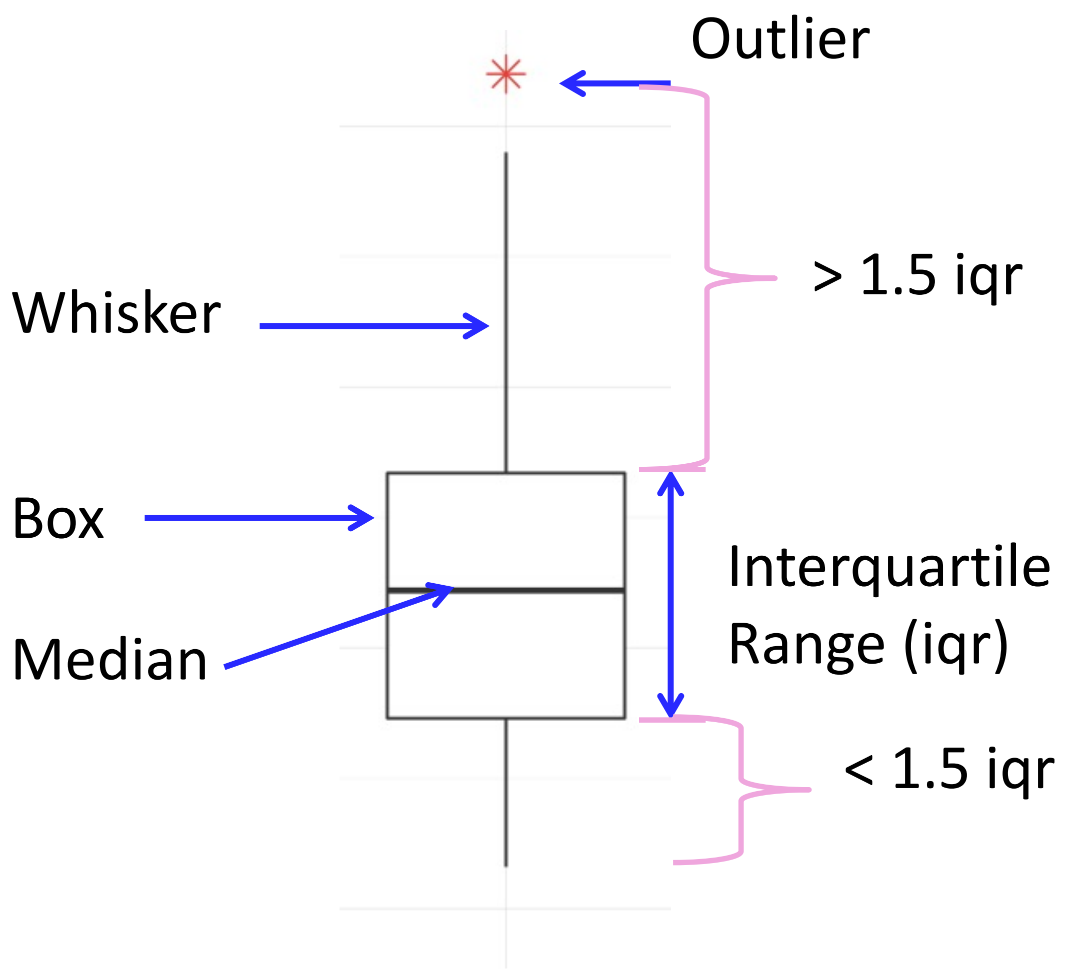
Vehicle death by region



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

Boxplots details, outliers

- How to define outliers? (the default)



Sensitivity of summary statistics to outliers

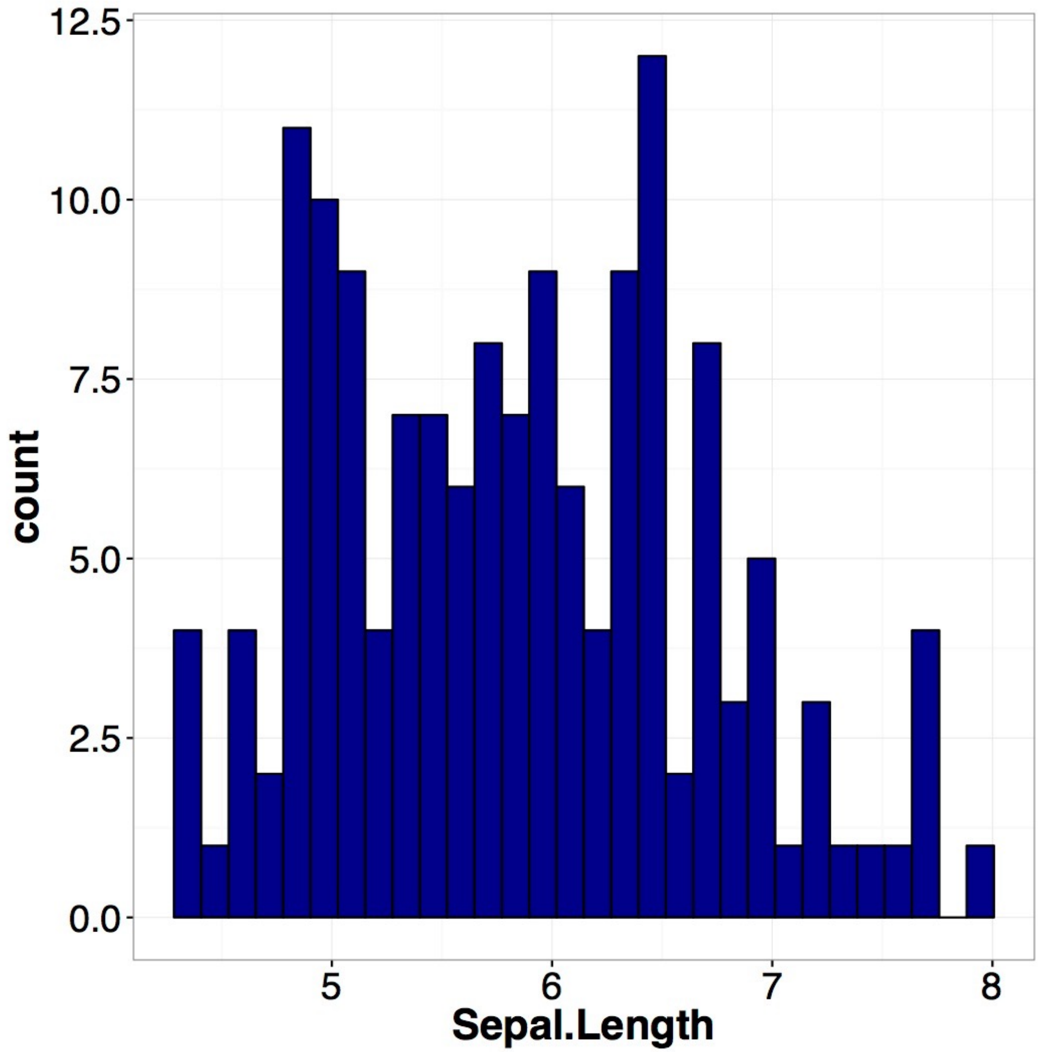
- mean and standard deviation are very sensitive to outliers
- median and interquartile range are not sensitive to outliers

Modes

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

Multiple modes

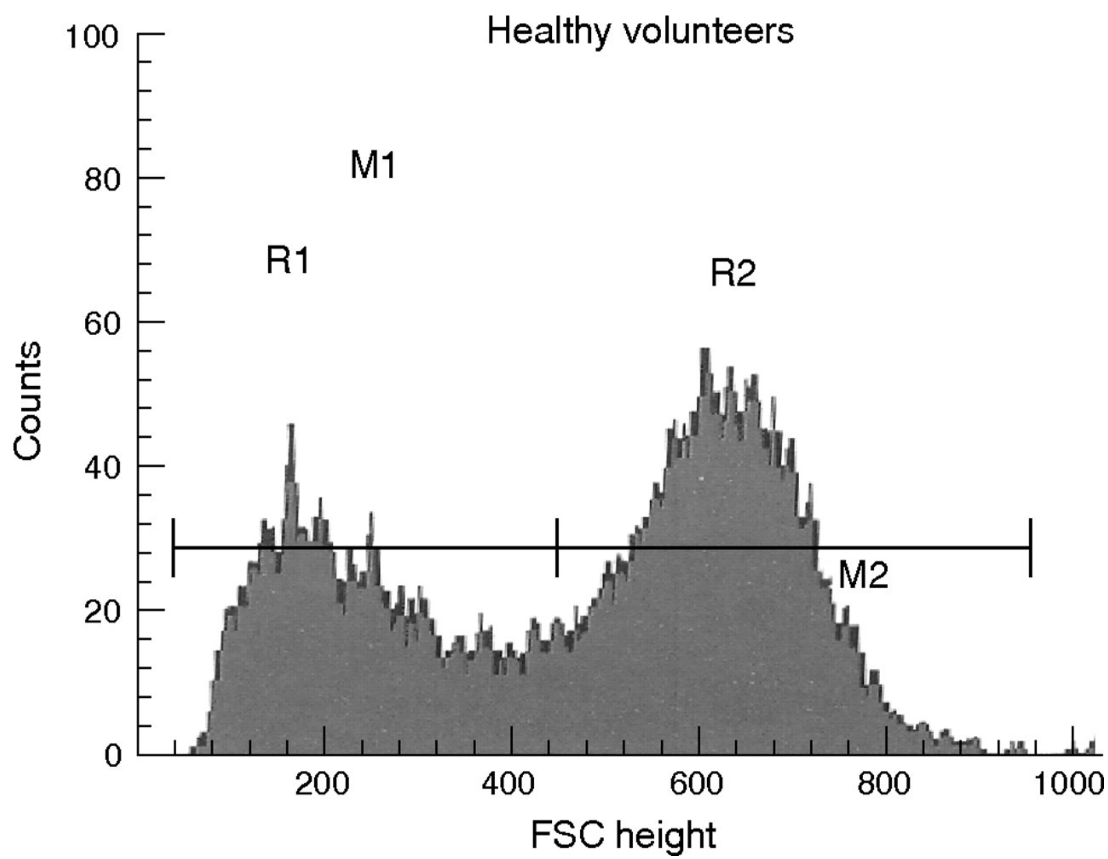
- We have seen the “iris” data which looks to have several peaks



Data: “iris” in R

Example Bi-modes distribution

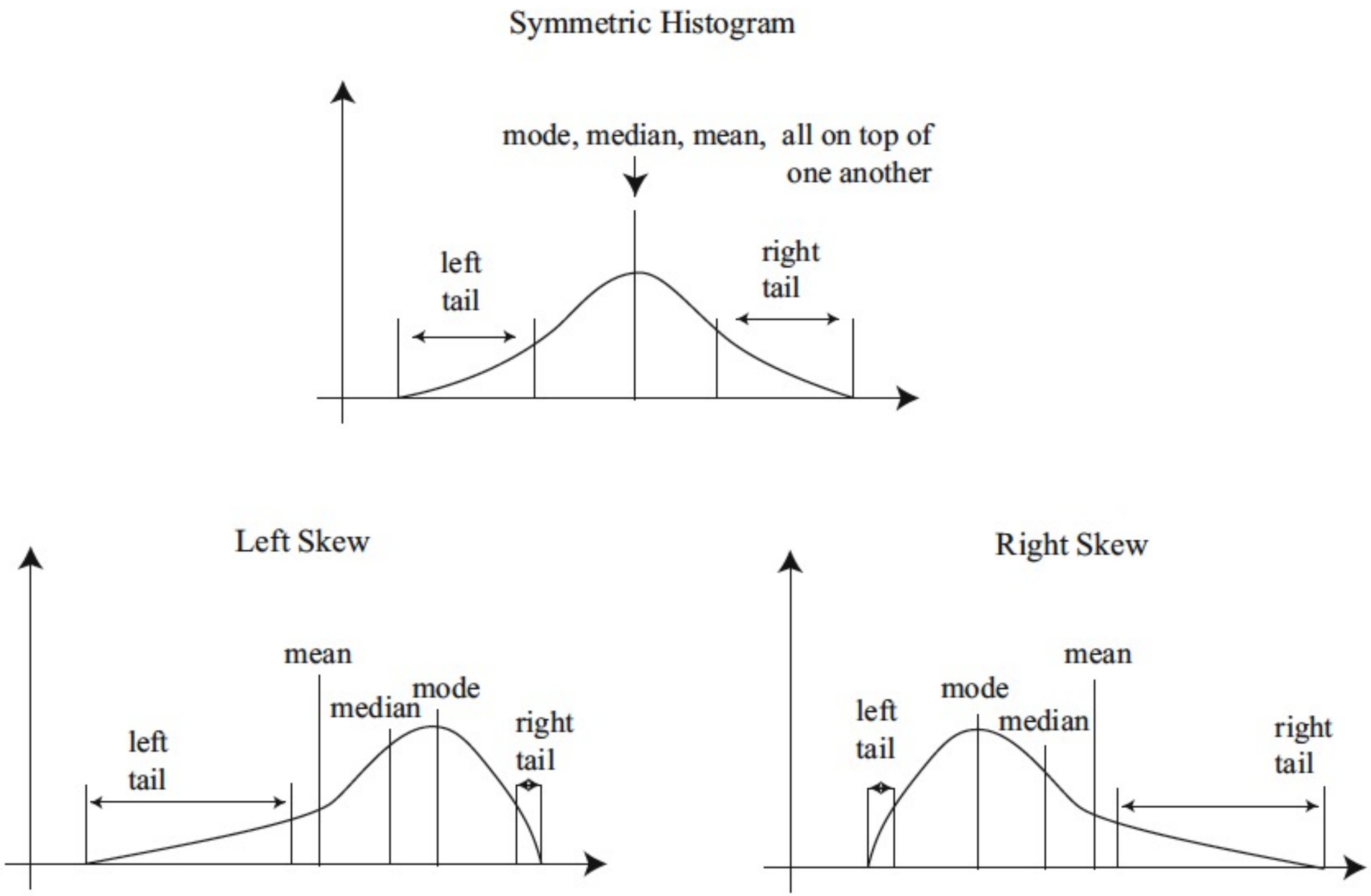
- Modes may indicate multiple populations



Data: Erythrocyte cells in healthy humans

Piagnerelli, JCP 2007

Tails and Skews



Credit: Prof.Forsyth

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 science

Heatmap

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

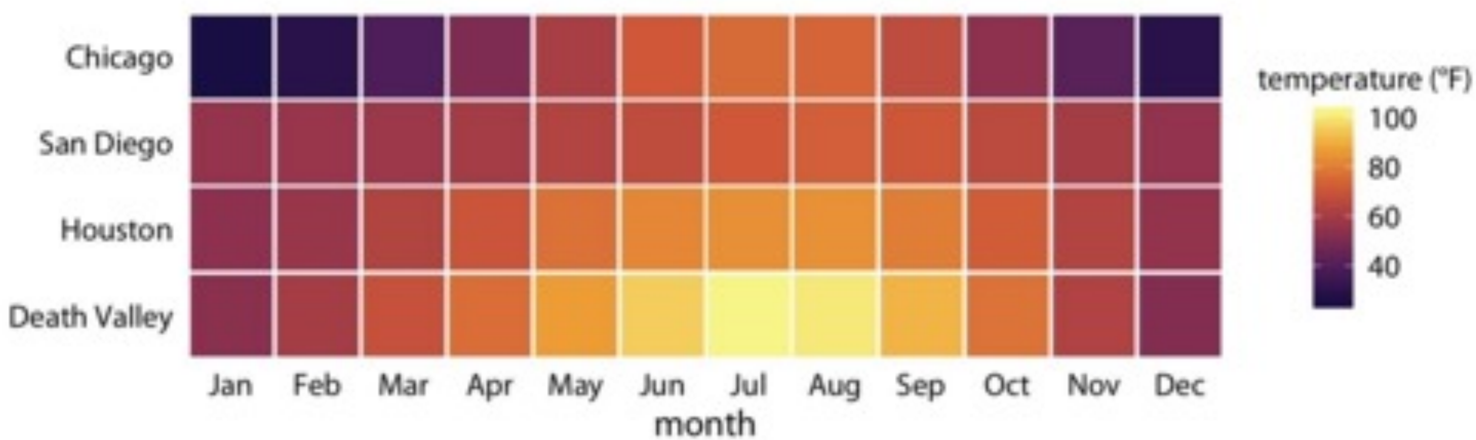
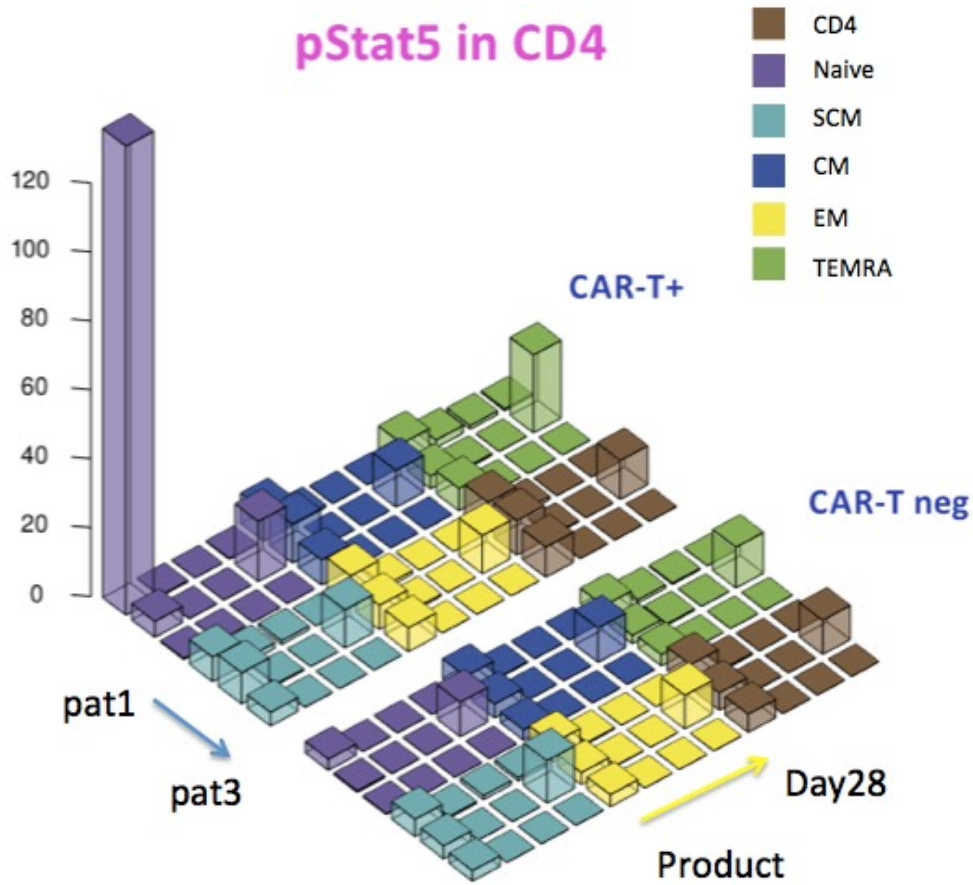


Figure 2-4. Monthly normal mean temperatures for four locations in the US. Data source: NOAA.

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

Relationship between data features

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

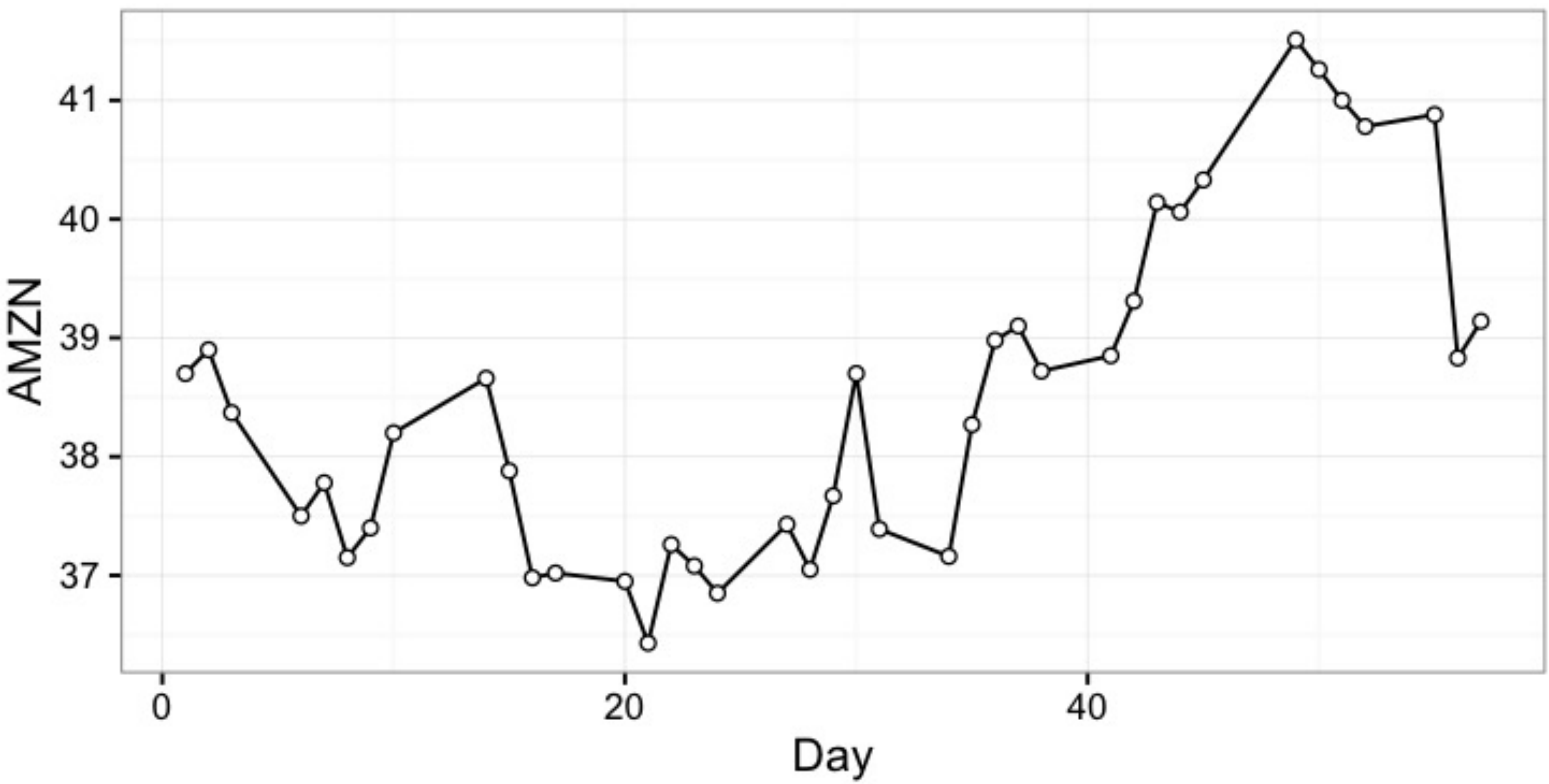
IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT
1	12.6	1.0708	23	154.25	67.75
2	6.9	1.0853	22	173.25	72.25
3	24.6	1.0414	22	154.00	66.25
4	10.9	1.0751	26	184.75	72.25
5	27.8	1.0340	24	184.25	71.25
6	20.6	1.0502	24	210.25	74.75
7	19.0	1.0549	26	181.00	69.75
8	12.8	1.0704	25	176.00	72.50
9	5.1	1.0900	25	191.00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

- x : HIGHT, y: WEIGHT

The visual way for continuous features

- Time series plot
- Scatter plot

Time Series Plot: Stock of Amazon

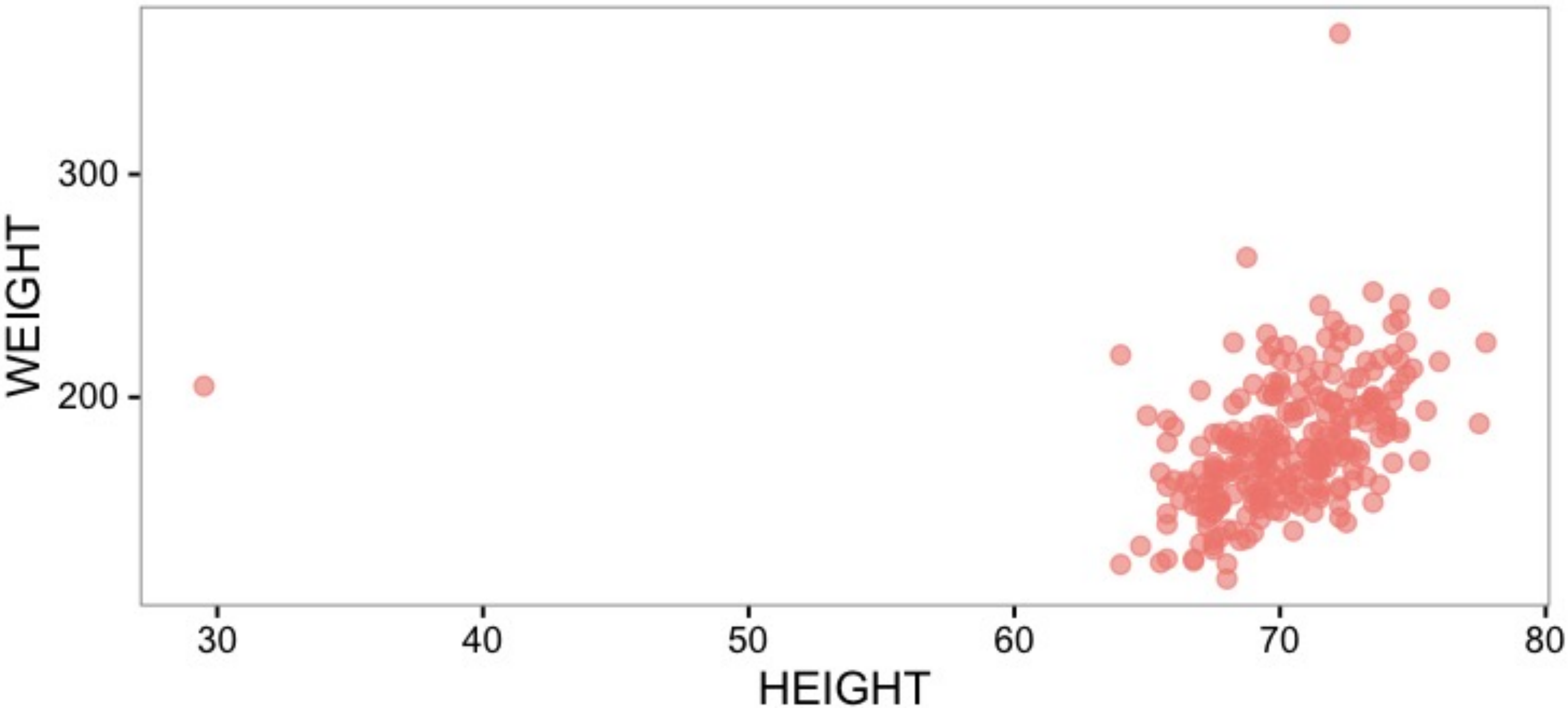


Scatter plot

- A most effective tool for geographic data and 2D data in general. It should be your first step with a new 2D dataset.

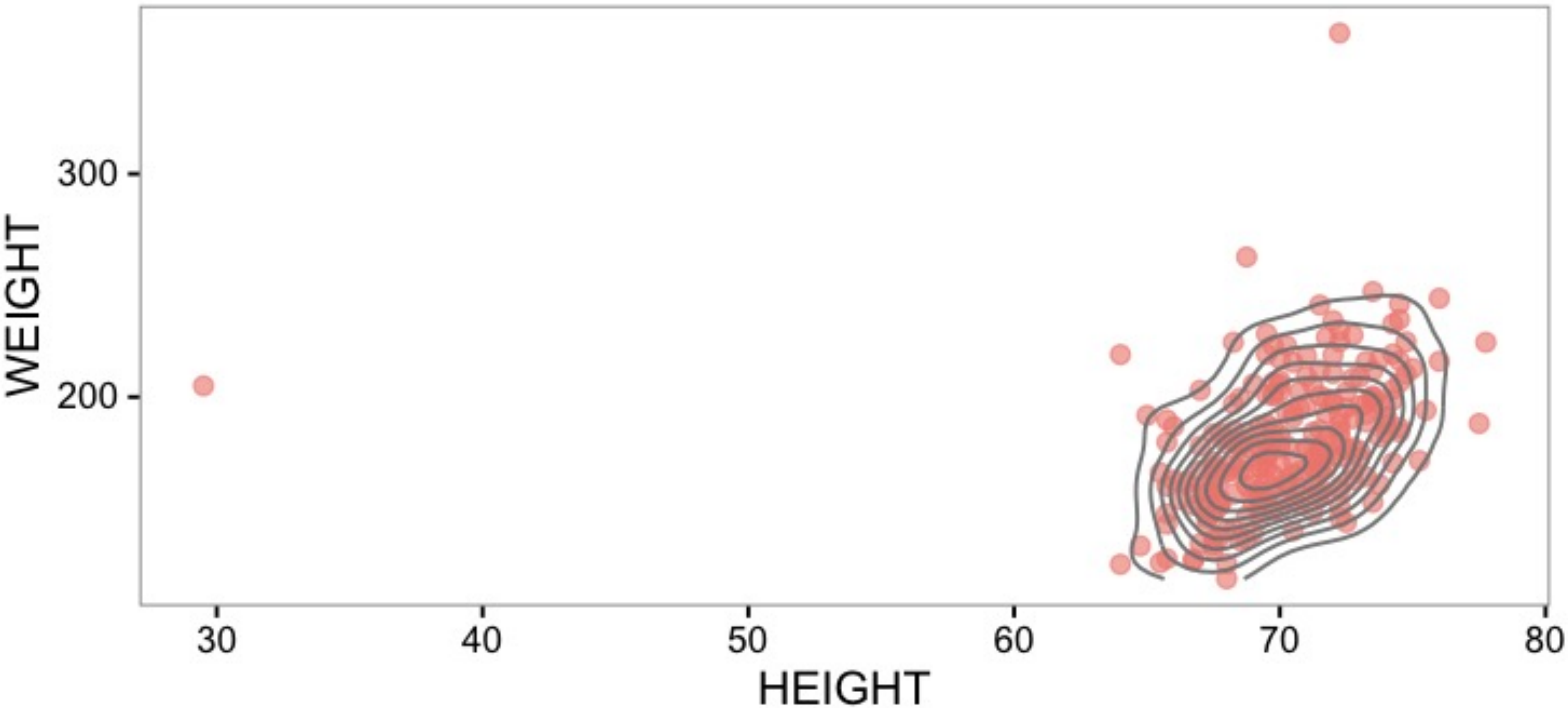
Scatter plot

- Body Fat data set



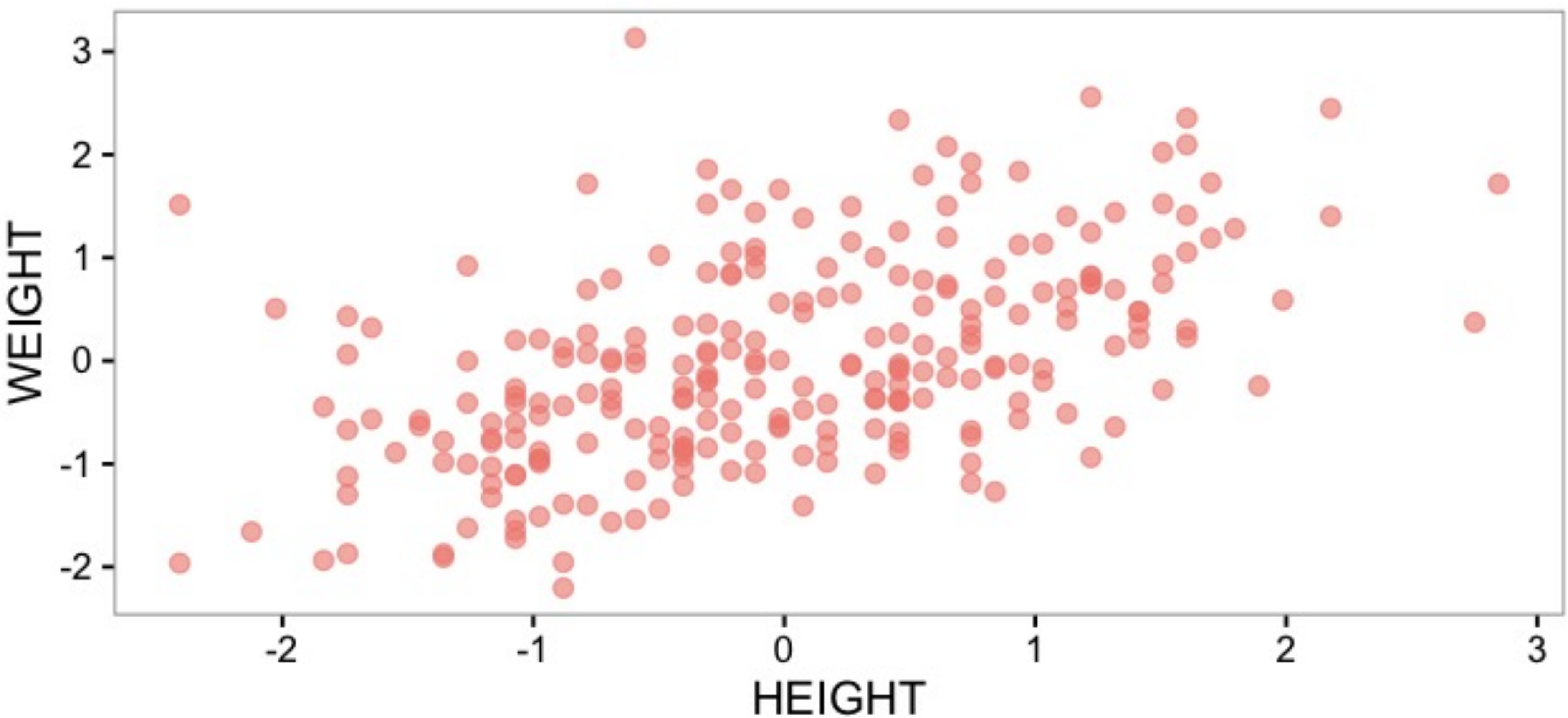
Scatter plot

- Scatter plot with density



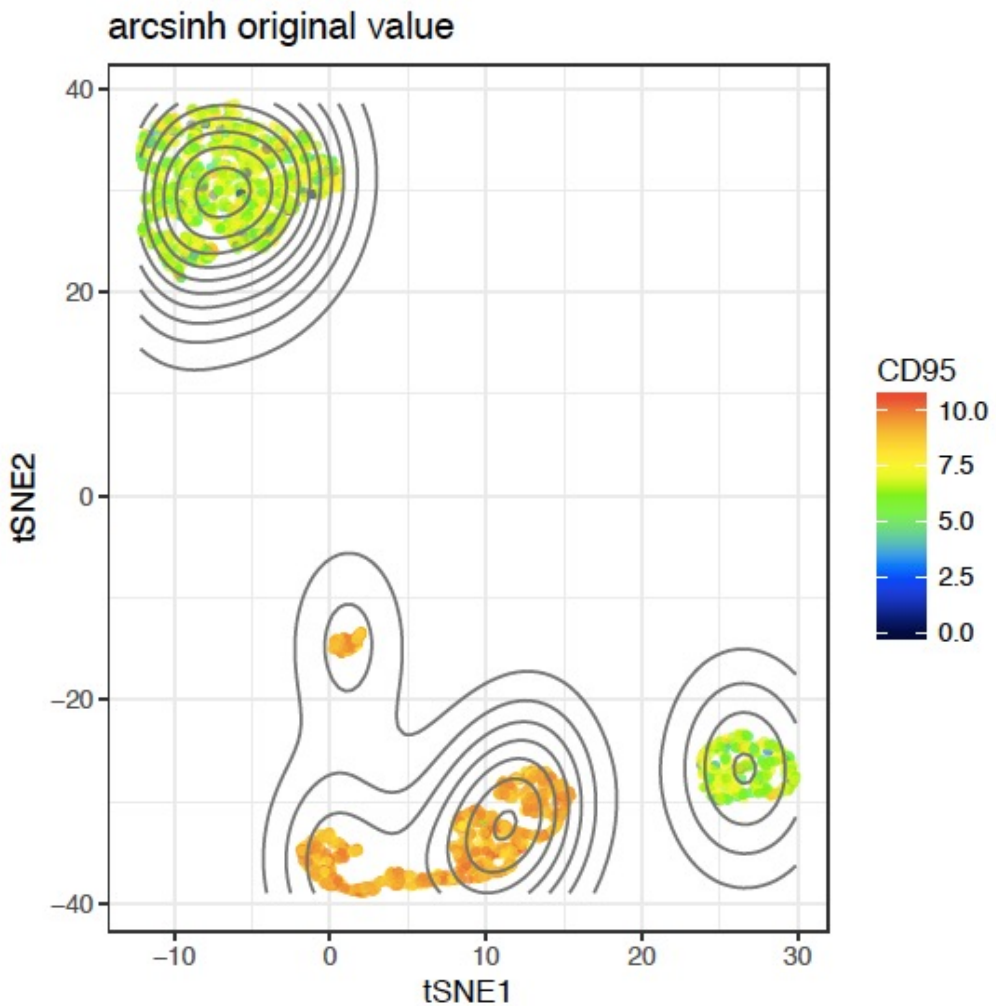
Scatter plot

- Removed of outliers & standardized



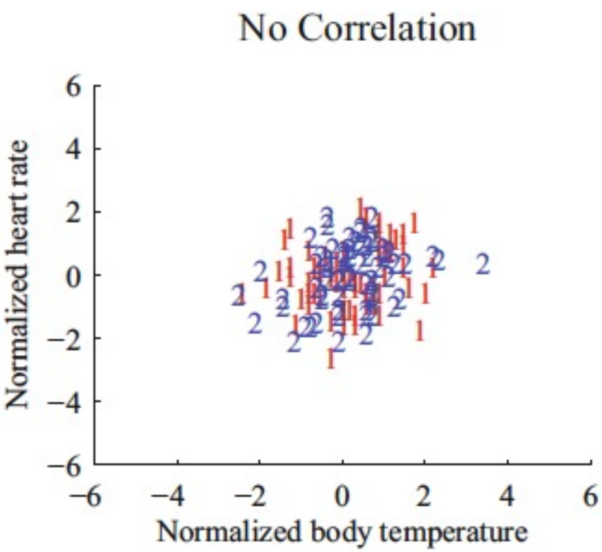
Scatter plot

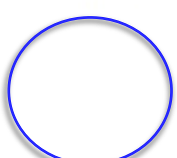
- Coupled with heatmap to show a 3rd feature



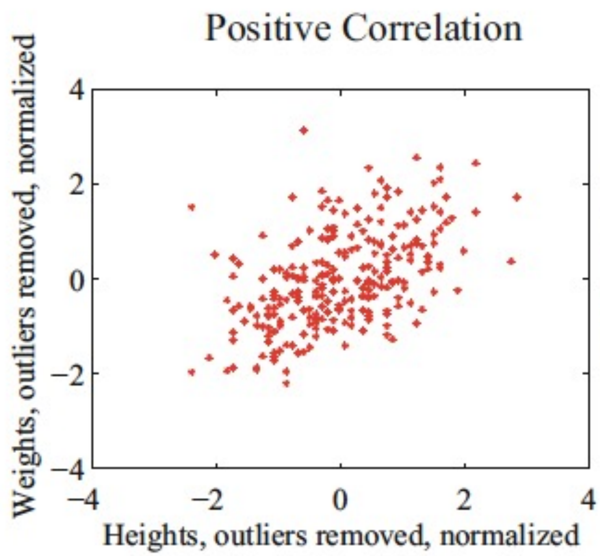
Correlation seen from scatter plots

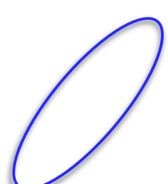
Zero Correlation



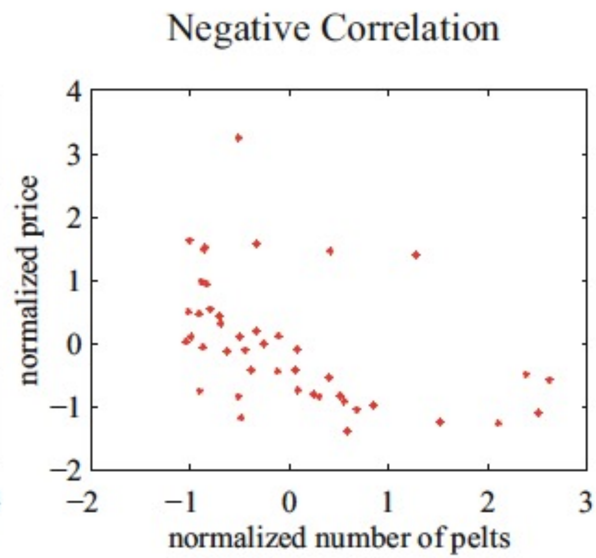


Positive correlation





Negative correlation



Credit: Prof.Forsyth

What kind of Correlation?

- line of code in a database and number of bugs
- 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.

Assignments

- **HW1** due Mon. Sept. 6.
- **Quiz 1 (open 4:30pm today until Mon. next week)**
- Reading upto Chapter 2.1
- Next time: the quantitative part of correlation coefficient

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!