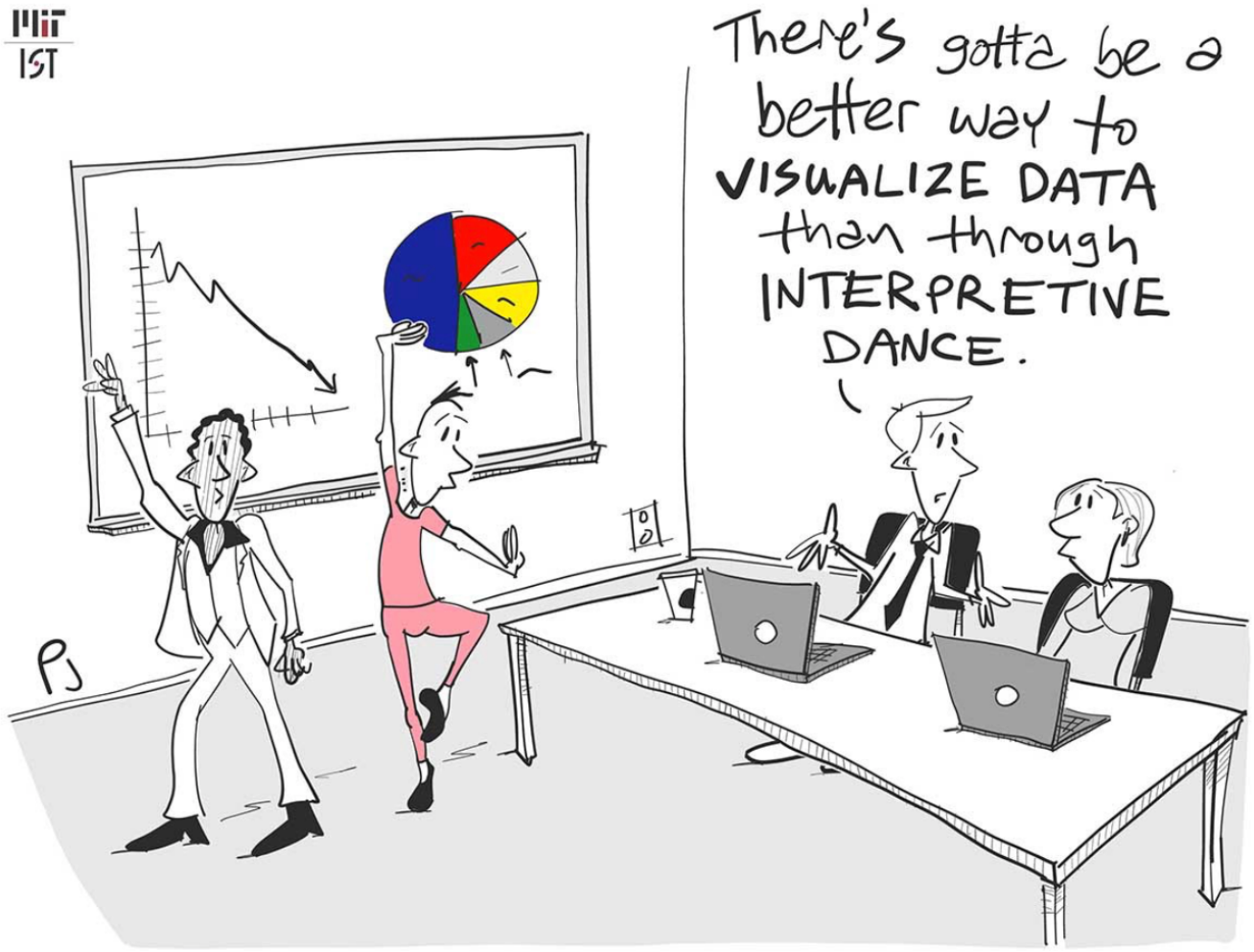
CSCI 3022: Intro to Data Science

Lecture 3: Exploratory Data Analysis, Data Visualization and Wrangling

Rachel Cox

Department of Computer Science



Alan Dabney & Grady Klein

Course Logistics – Before Next Class

- ❑ Read the Syllabus. You are responsible for knowing the information contained in this document.
- ❑ Make sure you can access the Canvas page.
- ❑ Check out the Piazza page.
- ❑ Install Anaconda (or other reliable Jupyter notebook method).
- ❑ Review and complete Numpy/Pandas tutorial
- ❑ Explore nb00 and nb01

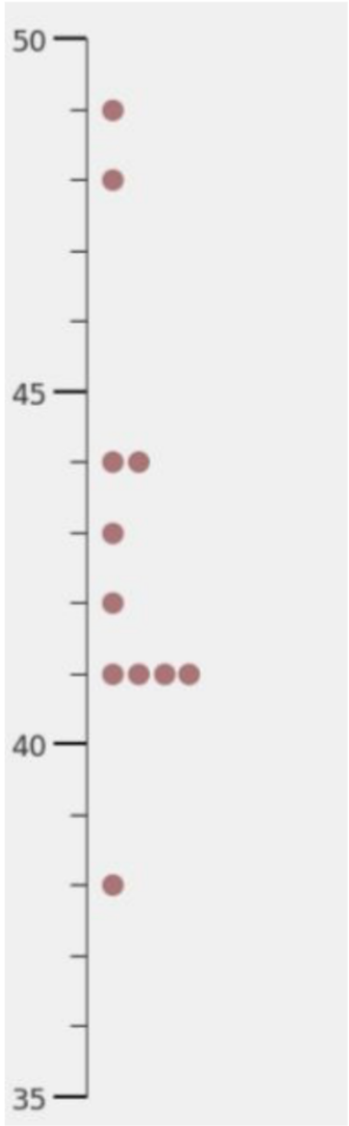
What will we learn today?

- ❑ Box-and-Whisker Plots
- ❑ Histograms

- ❑ *A Modern Introduction to Probability and Statistics, sections 15.1, 15.2, 16.4*

Quartiles

Example: Compute the quartiles and IQR of the data below.



$[38, 41, 41, 41, 41, \underline{42}, 43, 44, 44, 48, 49]$

$\tilde{x} = Q_2 = 42$

Q_1 = median of lower half $[38, 41, 41, 41, 41, 42]$

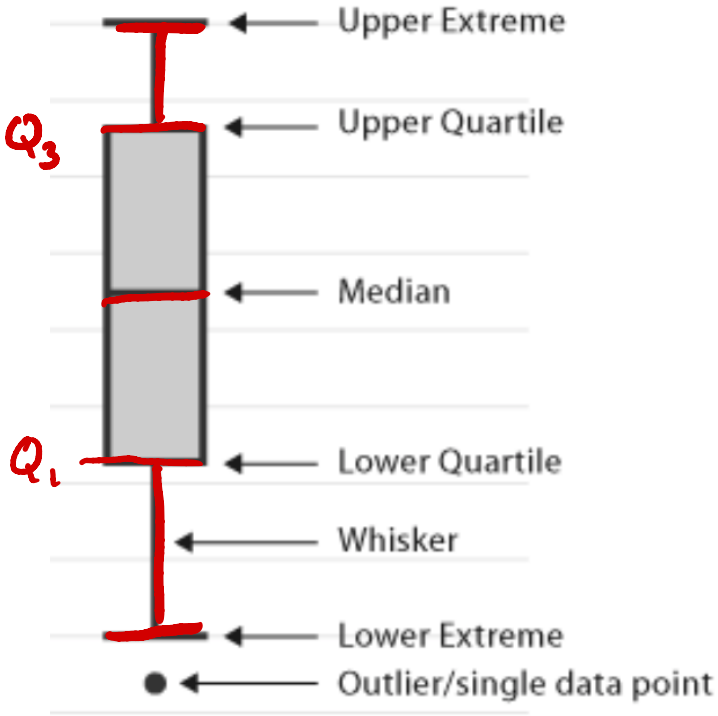
$Q_1 = 41$

Q_3 = median of upper half $[42, 43, 44, 44, 48, 49]$

$Q_3 = 44$

$IQR = Q_3 - Q_1 = 44 - 41 = 3$

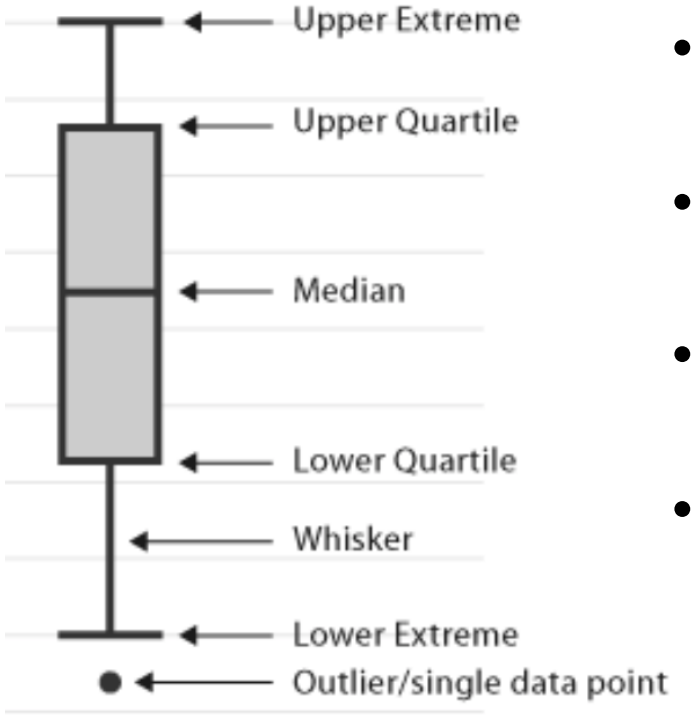
Box–and–Whisker Plots (aka Boxplots)



Box-and-Whisker plots are a convenient way to visualize data

- The **box** extends from Q_1 to Q_3
- The **median line** displays the median, $\tilde{x}$
- The **whiskers** extend to the farthest data point within $1.5 \times IQR$ of each quartile
- The fliers or outliers are any points outside of the whiskers.
- The width of the box is unimportant
- Boxplots can be horizontally or vertically oriented

Box–and–Whisker Plots (aka Boxplots)

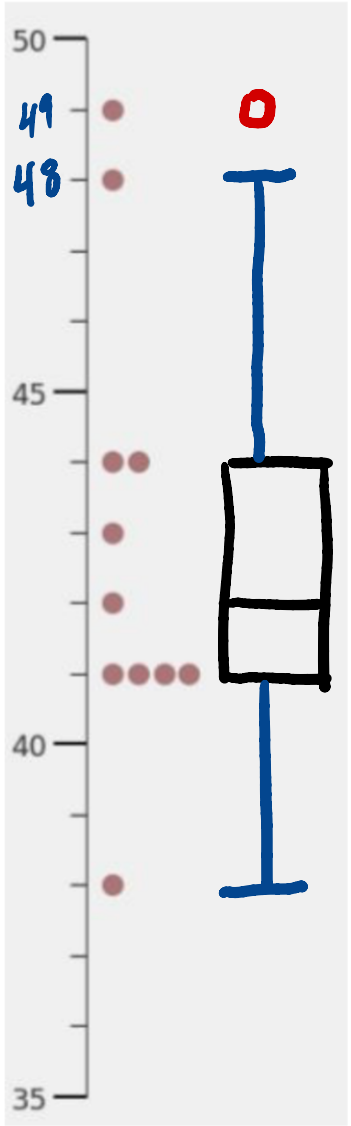


Box-and-Whisker plots are a convenient way to visualize data

- Depict the center of the data
- Depict the range and IQR
- Depict symmetry/skewness
- Show likely outlier
  - **When might a box-whisker plot be misleading?**
  - **When are box-whisker plots particularly useful?**

Box–and–Whisker Plots (aka Boxplots)

Example: Draw the box-and-whisker plot for the data on the left.



outlier

Q_3

median

Q_1

$Q_1 = 41$ ↑ lower side of box

$Q_3 = 44$ ↑ top of box

$\tilde{x} = 42$ ↑ inside box

$1.5 * IQR = 1.5 * 3 = 4.5$

longest upper whisker: $44 + 4.5 = 48.5$

longest lower whisker: $41 - 4.5 = 36.5$

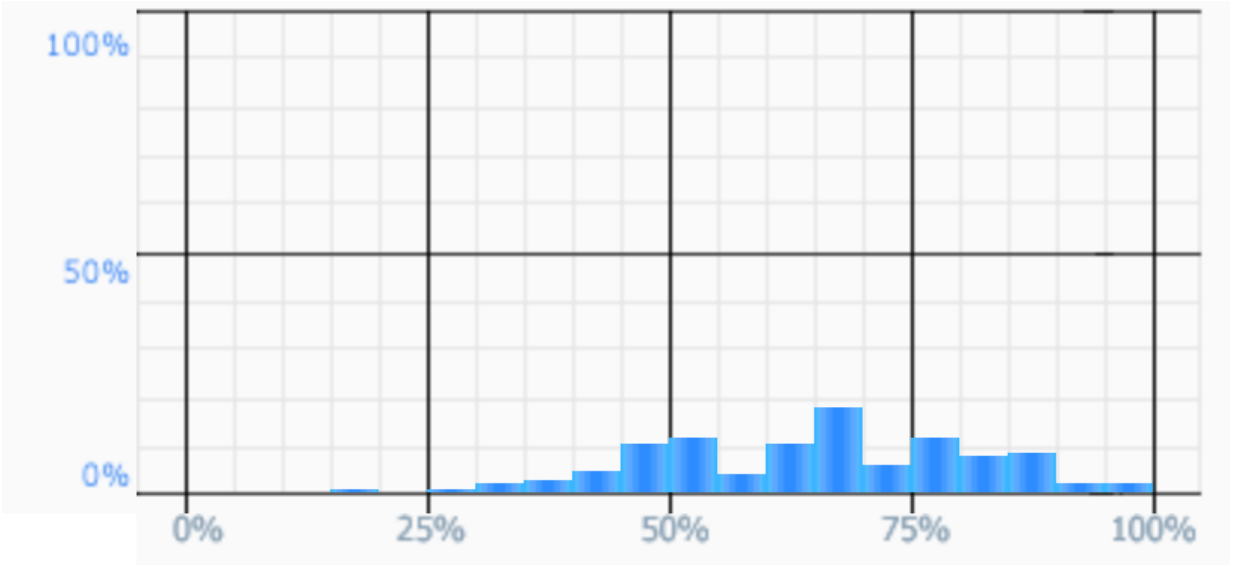
Histograms

The **histogram** is a graphical representation of the distribution of numerical data

Construction:

- Lump or “bin” the observed values of the Variable of Interest (VOI)
    - Bins typically are consecutive, non-overlapping, and equal in width

Example: Histogram of student grades on an exam.

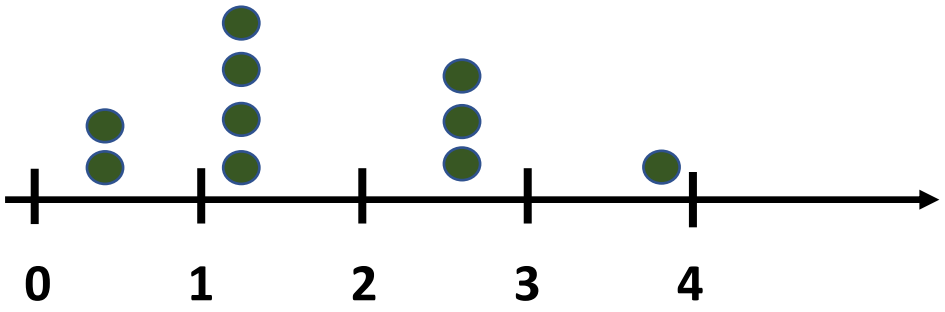


Histograms

For a **frequency histogram**: count the number of data values that fall into a bin and draw a rectangle over that bin with height equal to the count.

For a **density histogram**: count the number of data values that fall into a bin and adjust the height such that the sum of the area of all bins is equal to 1 (normalizing so that the sum of the heights = 1)

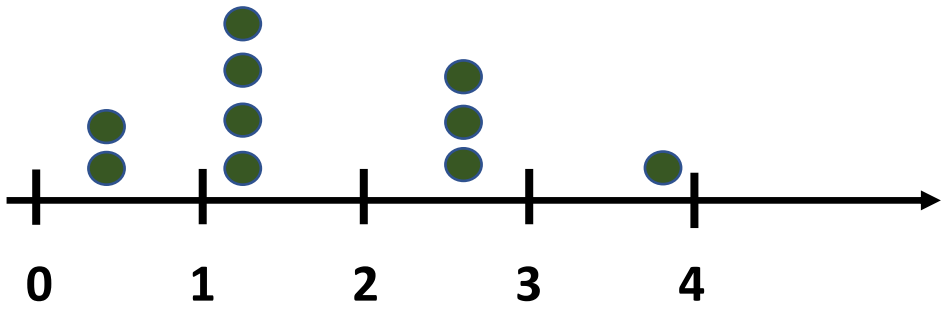
Example: Create a frequency histogram and a density histogram for the following data.



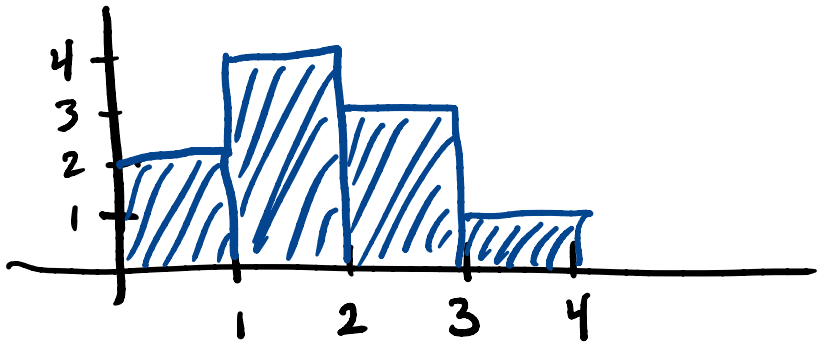
Histograms

Example: Create a frequency histogram and a density histogram for the following data.

0.3, 0.3, 1.2, ...



Frequency Histogram:

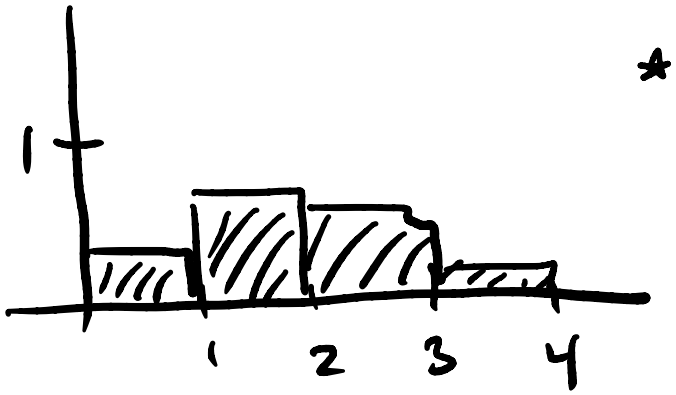


Density Histogram

Between		
(0, 1)	2/10	
(1, 2)	4/10	relative freq.
(2, 3)	3/10	
(3, 4)	1/10	

$\sum$ rectangle areas = 1

$\frac{2}{10} \cdot 1 + \frac{4}{10} \cdot 1 + \frac{3}{10} \cdot 1 + \frac{1}{10} \cdot 1 = \frac{10}{10} = 1$



* Density heights are not always = rel. freq.

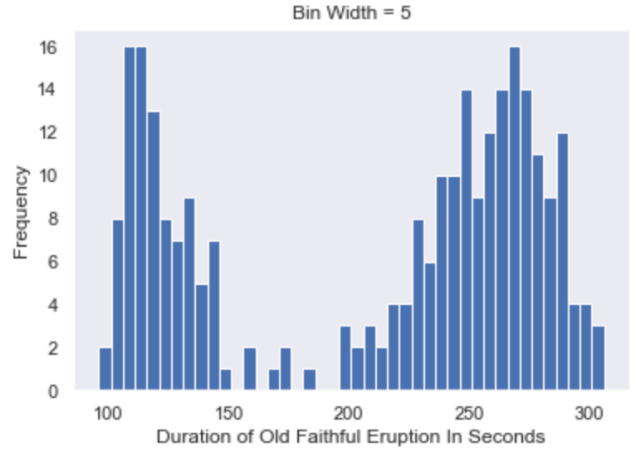
example Nose length (mm), $n = 60$

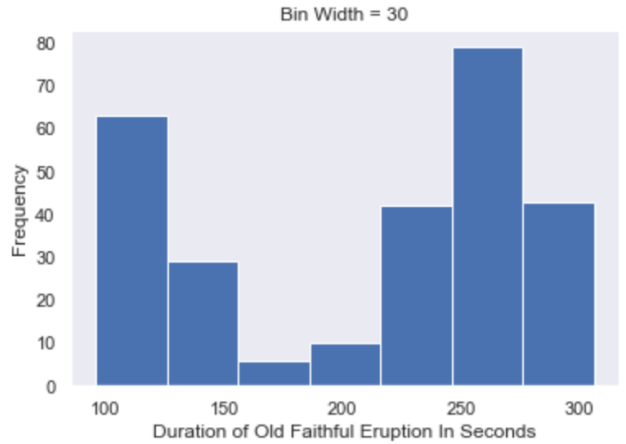
interval	Frequency	Relative Frequency	Density height
27.5 – 32.5	2	2/60	5(dens height) = 2/60
32.5 – 37.5	5	5/60	
37.5 – 42.5	17	17/60	
42.5 – 47.5	21	21/60	5(dens. height) = 21/60
47.5 – 52.5	11	⋮	
52.5 – 57.5	3		
57.5 – 62.5	0		
62.5 – 67.5	0		
67.5 – 72.5	1		
	f_i	f_i/n	density = $\frac{\text{Rel. freq.}}{\text{bin width}}$

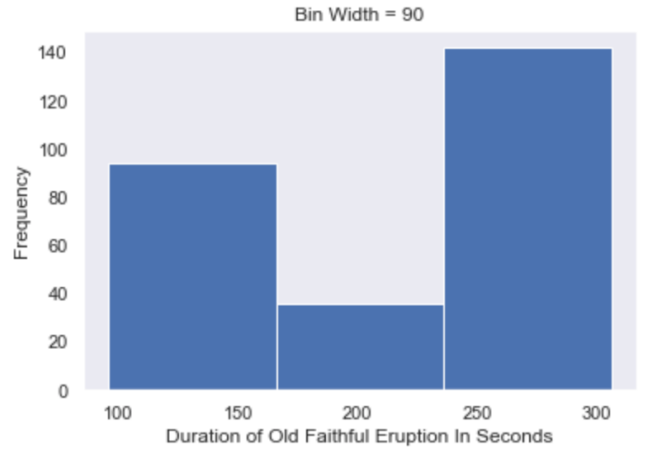
Histograms

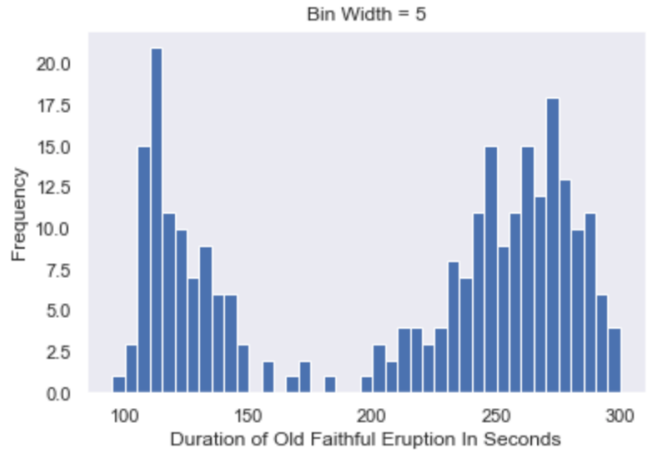
Note that choosing a different bin width can paint a very different picture of the data. Choosing the starting point for the bins also makes a difference.

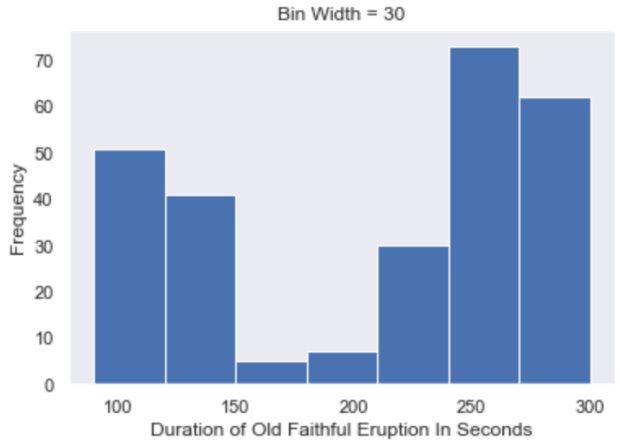
Example: Old Faithful eruption duration data (from MIPS section 15.1 p. 208)

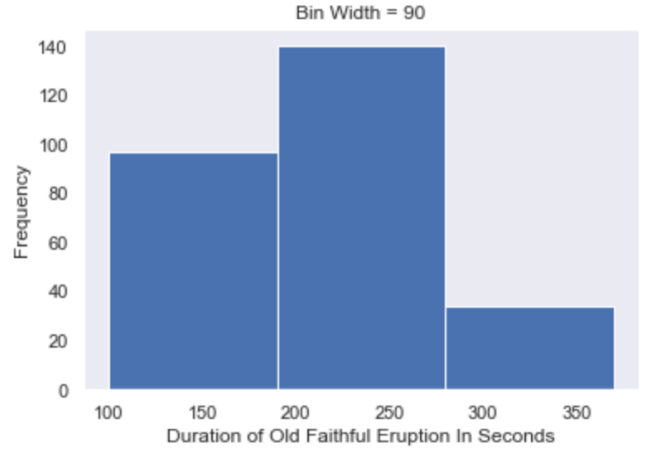






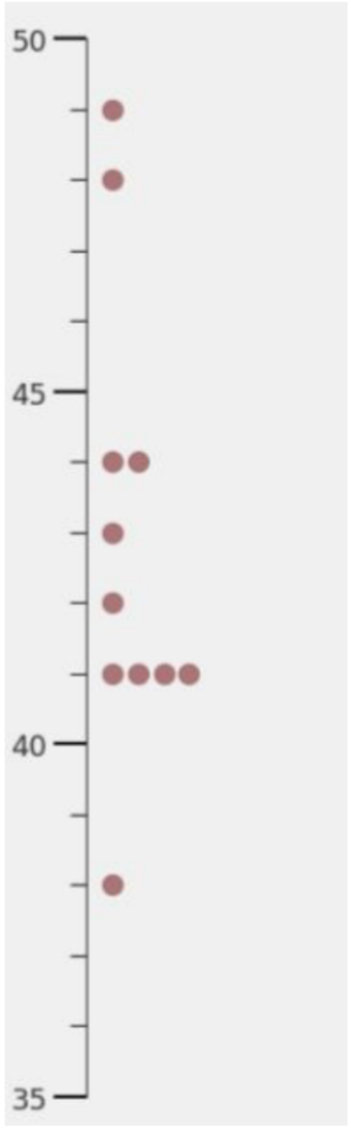


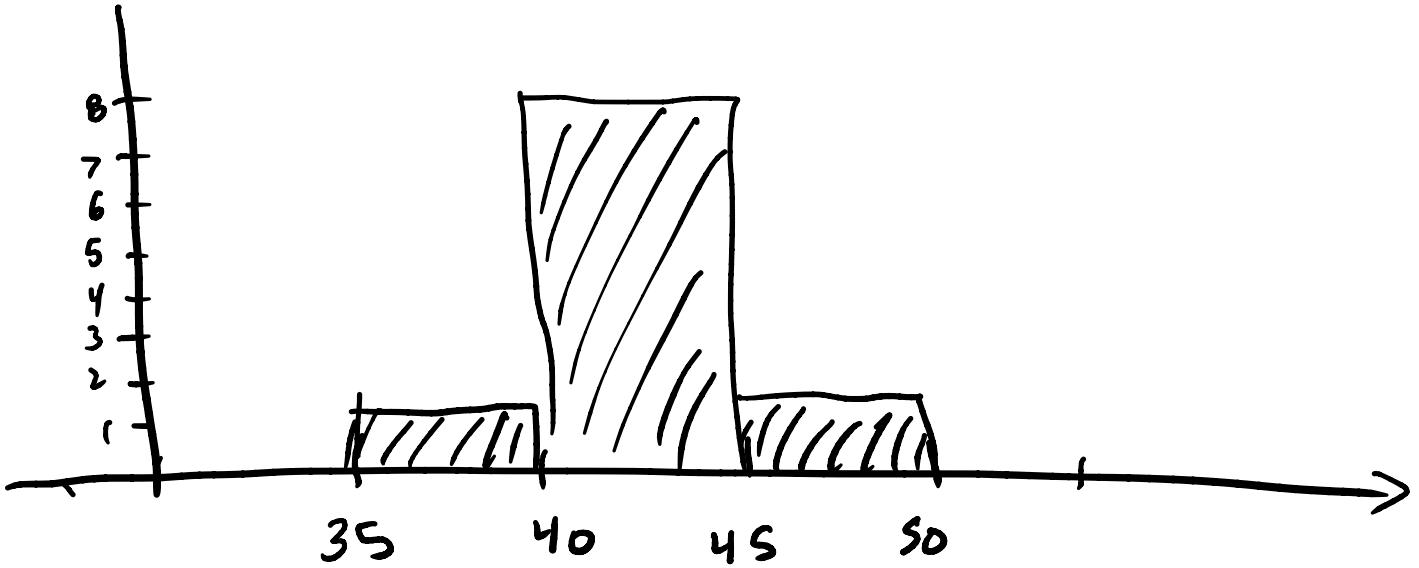




Histograms

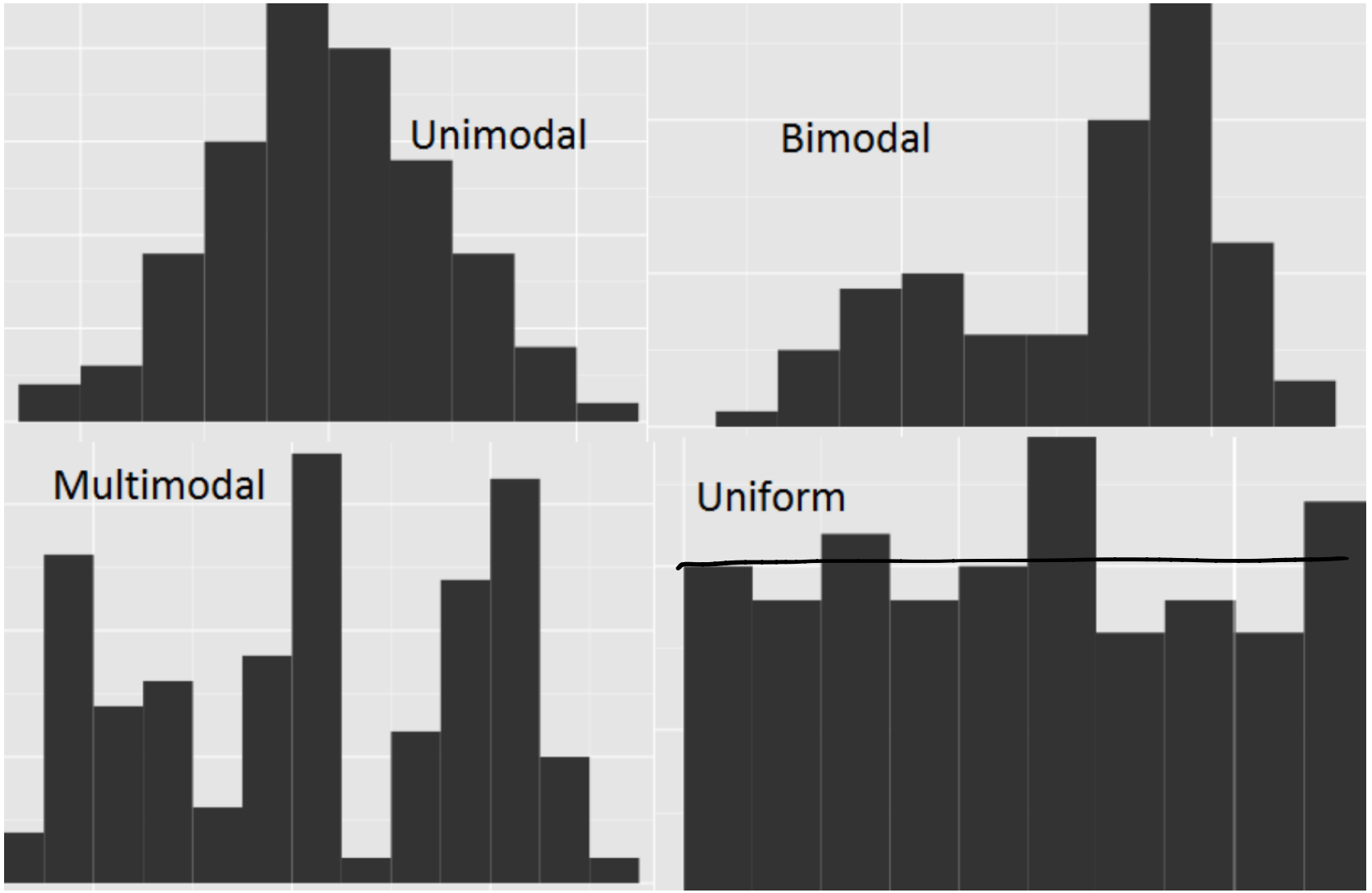
Example: Find the frequency histogram with bin width 5 of the data below, with left-most bin edge at 35.





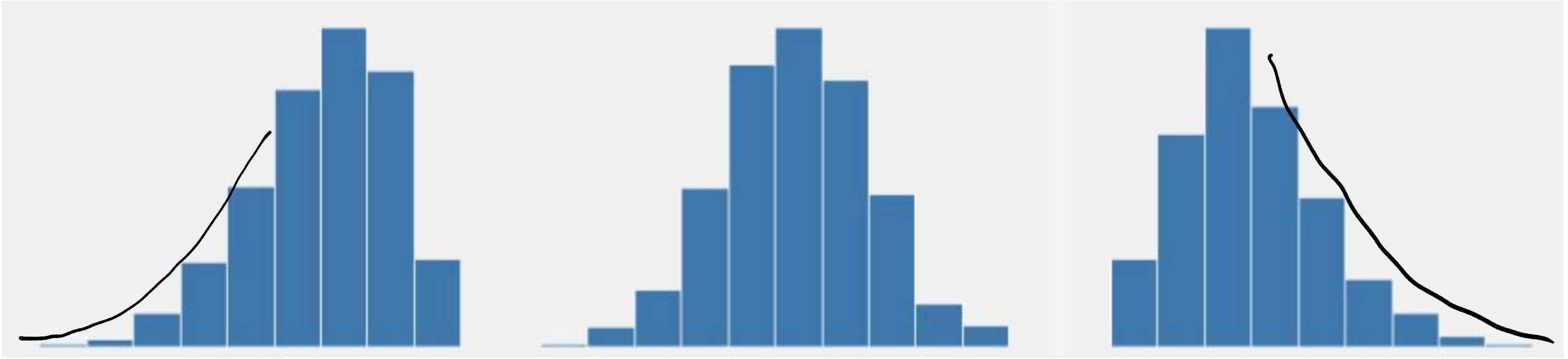
Histograms

Histograms come in a variety of shapes.



Histograms

Histograms come in a variety of shapes.



negative skew

symmetric

positive skew

left skew

Right skew

Next Time:

Probability