

# CSCI 3022-002 Intro to Data Science

## Visual Exploratory Data Analysis

**Opening Zoom Example:** *Calculate the Mean and Standard Deviation of the data set:*  
Data (units in dollars): 2,4,3,5,6,4.

## Opening Sol'n

**Example:** Calculation of the SD

Data (units in dollars): 2, 4, 3, 5, 6, 4.

std. distance from mean

$$\bar{X} =$$

$$\frac{2 + 4 + 3 + 5 + 6 + 4}{6} =$$

$$\frac{\sum_{i=1}^n X_i}{n}$$

$$= 24/6 = 4$$

$$X = 4$$

$$X = [4, 4]$$

$$S^2 = \frac{\sum_{i=1}^n (\text{squared dist. from mean})}{n-1 \quad (\text{avg})} =$$

## Opening Sol'n

**Example:** Calculation of the SD

Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2 + 4 + 3 + 5 + 6 + 4}{6} = \frac{24}{6} = 4$$

$X$	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
2	-2	4
4	0	0
3	-1	1
5	1	1
6	2	4
4	0	0

sum:  $s^2$

$$\frac{4 + 1 + 1 + 4}{6 - 1} = \frac{10}{5} = 2$$

$$\begin{aligned} \text{var}(X, \text{dof}=1) \\ &= \frac{\sum (X_i - \bar{X})^2}{n - \text{dof}} \end{aligned}$$

## Opening Sol'n

**Example:** *Calculation of the SD*

Data (units in dollars): 2,4,3,5,6,4.

Since we mean business, we need the average first.

$$\bar{X} = \frac{2 + 4 + 3 + 5 + 6 + 4}{6} = \frac{24}{6} = 4$$

Now let's compute the deviations...

vectorized deviations

$$\overbrace{[(X - \bar{X})^2]}^{\text{vectorized deviations}} = [(2 - 4)^2, (4 - 4)^2, (3 - 4)^2, (5 - 4)^2, (6 - 4)^2, (4 - 4)^2]$$

# Opening Sol'n

**Example:** Calculation of the SD

Data (units in dollars): 2, 4, 3, 5, 6, 4.

(4)

sd: \$1.41

$$x_1 = 2 \quad x_2 = 4 \quad \dots \quad x_6 = 4$$

Since we mean business, we need the average first.

$$\bar{X} = \frac{2 + 4 + 3 + 5 + 6 + 4}{6} = \frac{24}{6} = 4$$

Now let's compute the deviations...

vectorized deviations

$$[(X - \bar{X})^2] = [(2 - 4)^2, (4 - 4)^2, (3 - 4)^2, (5 - 4)^2, (6 - 4)^2, (4 - 4)^2]$$

$s^2$ : "variance"

and sum and "average" those!

$s$ : std. dev:  $\sqrt{2}$

$$s^2 = \frac{4 + 0 + 1 + 1 + 4 + 0}{6} = 2$$

# Announcements and To-Dos

## Announcements:

1. HW 1 Posted, due Friday!
2. Another nb day this Friday!

## Last time we learned:

1. Loading, beginning to manipulate data in Python.

## To do:

1. Start that HW! Ensure you can load the data and work with it. Practice your TeX/markdowns!

M.H. Form

1) Predictions

c) Jupyter setup.

3) (1)

## Last Time Recap:

We talked about two big types of measures for a data set  $X_1 \dots X_n$ : centrality and dispersion.

Measures	Stat	Calculation	Advantages
Centrality	Mean	$\frac{\sum_{i=1}^n X_i}{n}$	Uses all data, behaves 'nicely' not pulled by single outliers 'indicative' of true data value
	Median	middle value	
	Mode	most common value	
Dispersion	Variance	$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$	Squared distance can be nice, <u>no radical</u> <u>units</u> same as data shows extremes like median: avoids outliers
	SD	$\sqrt{s^2}$	
	Range	Max minus min	
	IQR	$Q_3 - Q_1$	

## Means and Medians:

One way we conceptualize the mean and the median as data scientists is we ask the question: “what single number is **closest** to our data.” This requires us choose a definition of distance: squared or absolute?

**We prove:** Show that the *sample mean* of data  $X_1, X_2, \dots, X_n$  is the unique **minimizer**  $c$  of the function

$$f(c) = \sum_{i=1}^n (X_i - c)^2$$

**NB:** The *median* of data  $X_1, X_2, \dots, X_n$  is the possibly non-unique minimizer  $c$  of the function

$$f(c) = \sum_{i=1}^n |X_i - c|$$

$x = [1, 2]$   
 1.3 dist: .3, .7  
 1.5 dist: .5, .5



**Proof:** Differentiating yields

$$f(c) = \sum_{i=1}^n (x_i - c)^2 \quad \frac{\partial f}{\partial c} = \sum_{i=1}^n \frac{\partial f}{\partial c} (x_i - c)^2$$

↪

$$\frac{\partial f}{\partial c} = \sum_{i=1}^n 2 \underbrace{(x_i - c)}_{\frac{dx}{dc}(x_i - c)} (-1)$$

Set = 0  
= 0

$$0 = \sum_{i=1}^n [(-2x_i) + (2c)] = \sum_{i=1}^n (x_i - c) = \sum_{i=1}^n x_i - \sum_{i=1}^n c$$

**Proof:** Differentiating yields

$$f'(c) = \frac{df}{dc} \sum_{i=1}^n (X_i - c)^2 = \sum_{i=1}^n -2(X_i - c).$$

Setting  $f'(c) = 0$  gives

**Proof:** Differentiating yields

$$f'(c) = \frac{df}{dc} \sum_{i=1}^n (X_i - c)^2 = \sum_{i=1}^n -2(X_i - c).$$

Setting  $f'(c) = 0$  gives

$$\begin{aligned} 0 &= \sum_{i=1}^n -2(X_i - c) \\ &= 2nc - 2 \sum_{i=1}^n X_i \\ \Rightarrow c &= \frac{\sum_{i=1}^n X_i}{n} = \bar{X} \end{aligned}$$

Handwritten notes:

$$0 = \sum_{i=1}^n X_i - n \cdot c$$

$$nc = \sum_{i=1}^n X_i$$

$$c = \sum X_i / n$$

# Histograms

**Definition:** A *histogram* is a graphical representation of the distribution of numerical data.

To construct a histogram:

1)

"Bin" the measured values of the Vol. (The bins are typically consecutive, non-overlapping, and are usually equal size.)

height = "H" of times  
width = data inside interval

intervals in the range.

*Frequency histogram:* count how many values fall into each bin/interval and draw accordingly.

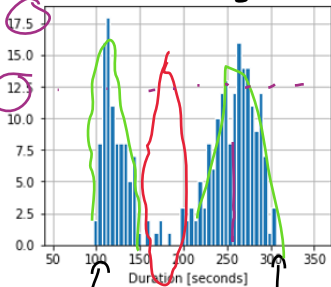
*Density histogram:* count how many values fall into each bin, and adjust the height such that the sum of the area of all bins equals 1. Equivalently: construct a Frequency histogram and divide the  $y$  axis by the total data count.

# Old Faithful Histogram

The number of bins chosen may lead to very different pictures of the data!

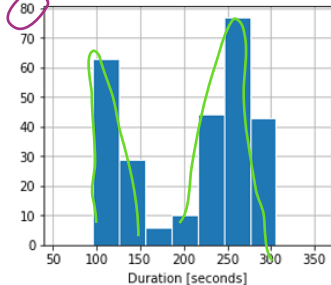
*Smaller*

Bin width = 5

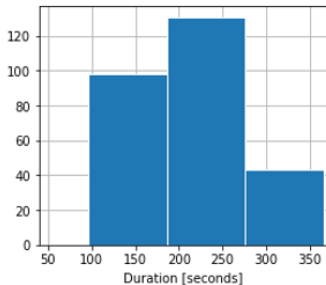


*larger*

Bin width = 30



Bin width = 90



One such choice:

Friedman-Diaconis: bin width =  $2 \frac{IQR}{\sqrt[3]{n}} = 2 \frac{Q_3 - Q_1}{n^{1/3}}$

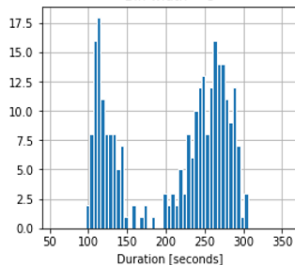
*range of data*

*more data  $\rightarrow$  smaller bins*

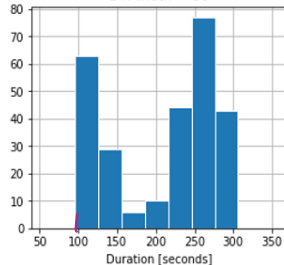
*2 important things:*  
 1) gap at 150-200  
 2) more often 200-300 than 100-150

Where bins begin and end may also matter!

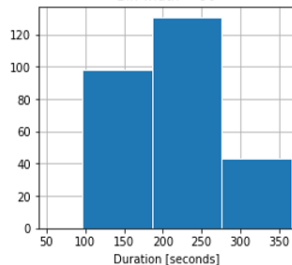
Bin width = 5



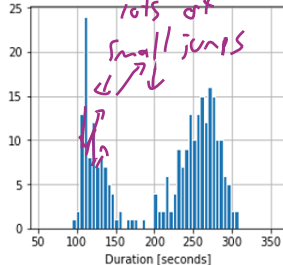
Bin width = 30



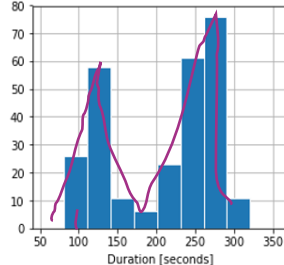
Bin width = 90



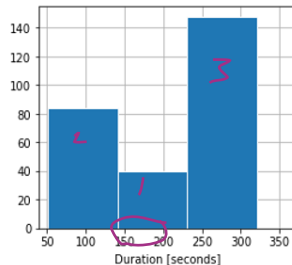
Bin width = 5



Bin width = 30



Bin width = 90



## How many bins?

A lot of statisticians advise different rules or sanity checks for histogram bins.

Textbook:

$$n_{bins} = 1 + 3.3 \log_{10}(n)$$

$$w_{bins} = \frac{3.49s}{n^{1/3}}$$

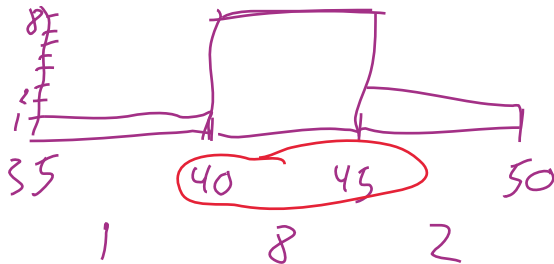
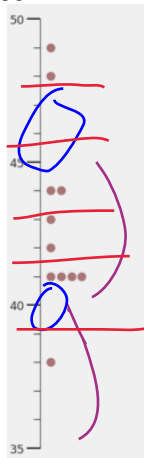
*5! std. dev.*

Don't memorize these. My heuristic for binning: start with "too many" bins at first if you have to, and slowly expand the bin size to ensure:

1. The data starts to "smooth" out a little... but
2. We don't smooth over what appear to be distinct multiple modes

## Histogram Example

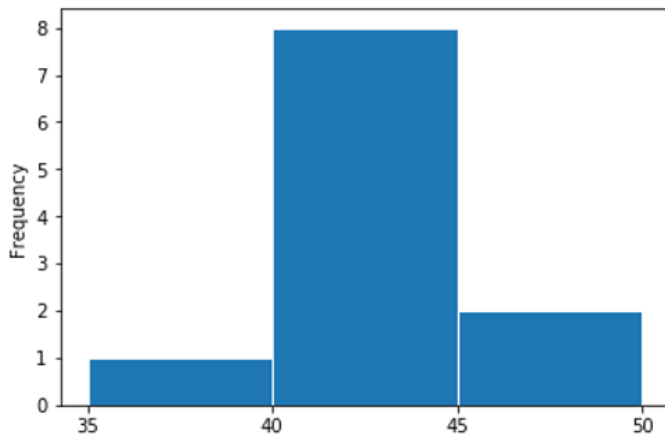
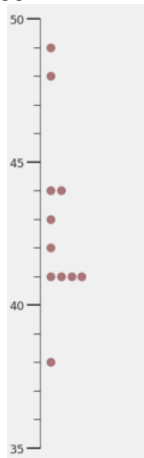
Find the frequency histogram with bin width 5 of the data on left, with left-most bin edge at 35.





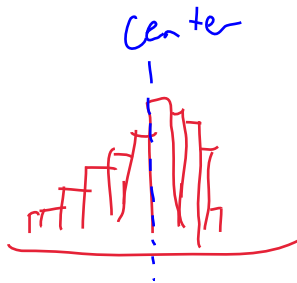
## Histogram Example

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



# Histogram Descriptives

Histograms come in a variety of shapes.



Negative Skew

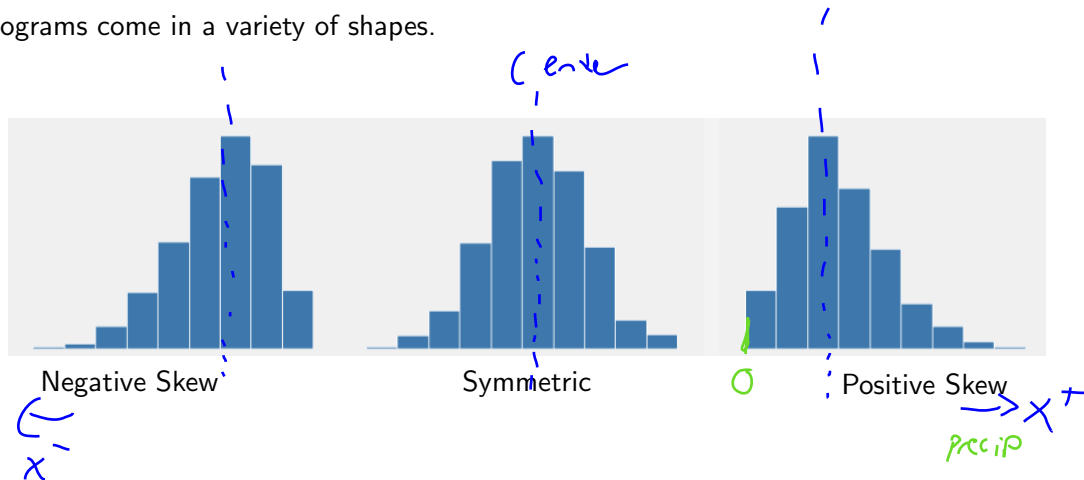
outlier in  
the "negative x"

Symmetric

Positive Skew

# Histogram Descriptives

Histograms come in a variety of shapes.



# Histogram Descriptives

Histograms come in a variety of shapes.

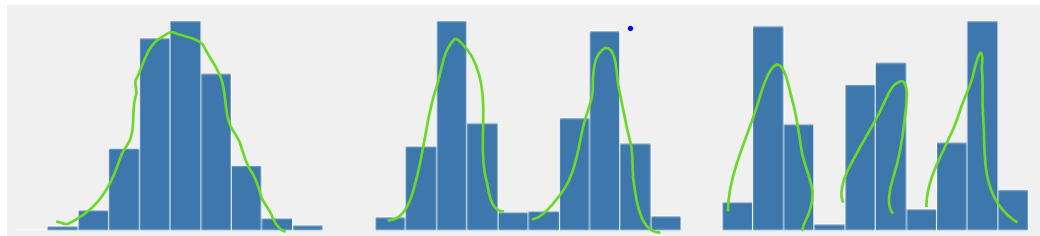
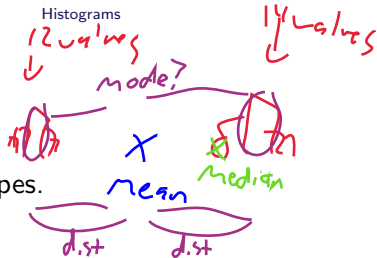
Unimodal

Bimodal

Multimodal

# Histogram Descriptives

Histograms come in a variety of shapes.



Unimodal

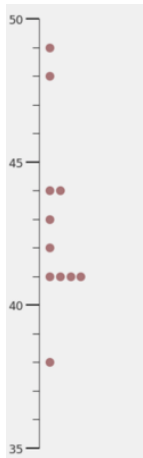
Bimodal

Multimodal

## Quartiles, Day 2

Compute the Quartiles and the IQR of the data to the left, with

$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$



$Q_1$

$q_1$

$Q_2$

$q_2$

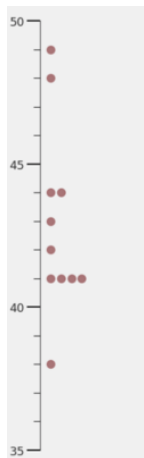
$Q_3$

$q_3$

## Quartiles, Day 2

Compute the Quartiles and the IQR of the data to the left, with

$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$



$n = 11$  is odd, so  $Q_2$  or the median is the 6th sorted value of 42. Then 41 and 44 divide the the halves in half, and are the 3rd and 9th sorted data points.

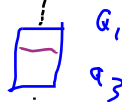
$$x = [38, 41, 41, 41, 41, 42, 43, 44, 44, 48, 49]$$

This makes the  $IQR = 44 - 41 = 3$

# Boxplots

A boxplot is a convenient way of graphically depicting groups of numerical data through the five number summary: minimum, first quartile, median, third quartile, and maximum.

1. The box extends from Q1 to Q3
2. The median line displays the median
3. The whiskers extend to farthest data point within  $1.5 \times IQR$  of each quartile
4. The fliers or outliers are any points outside of the whiskers
5. The width of the box is unimportant
6. Can be horizontally or vertically oriented





## Boxplots

Why do we use box plots?

1. They depict centrality via the median.
2. They depict dispersion through both the range and the IQR
3. Major outliers are shown
4. The median's location within the IQR suggests skewness; so too may lopsided whisker lengths or outliers

When might a box-whisker plot be misleading?

- 

When might a box-whisker plot be particularly useful?

-

## Boxplots

Why do we use box plots?

1. They depict centrality via the median.
2. They depict dispersion through both the range and the IQR
3. Major outliers are shown
4. The median's location within the IQR suggests skewness; so too may lopsided whisker lengths or outliers

When might a box-whisker plot be misleading?

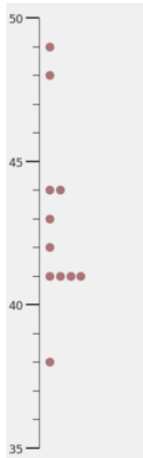
- No indication of how data are dispersed (is there “no-man’s land”?)

When might a box-whisker plot be particularly useful?

- Comparing medium numbers of variables or columns quickly (say, 3-10); and much easier than histograms

## Boxplot Example

Draw the box-whisker plot for the data to the left.



# Title

Today we learned

1. How to represent data with histograms and box-whisker plots (boxplots)
2. Some strengths and weaknesses of each

Moving forward:

- No class Monday for Labor Day.
- Notebook day: making some histograms, boxplots, and playing around with data frames.
- 3 of the next 5 course meetings are notebook days.

Next time in lecture (Friday):

- We probably talk about probability!

