"This is not what I meant when I said 'we need better data cleansing!'"

Lecture 3: EDA and data visualization

Announcements and reminders

- **Canvas:** make sure you have looked over the syllabus and schedule

  https://canvas.colorado.edu/courses/24706

- **Piazza:** be on it, because no more emails, and I don't like Canvas very much!

  https://piazza.com/colorado/spring2019/csci3022/

- Get **Jupyter notebook / Anaconda Python** -- make sure you have a working install and check out the Numpy/Pandas tutorial (github/notebooks)

  https://www.anaconda.com/downloads

Quizlet 1

Due Wed @ 10 AM

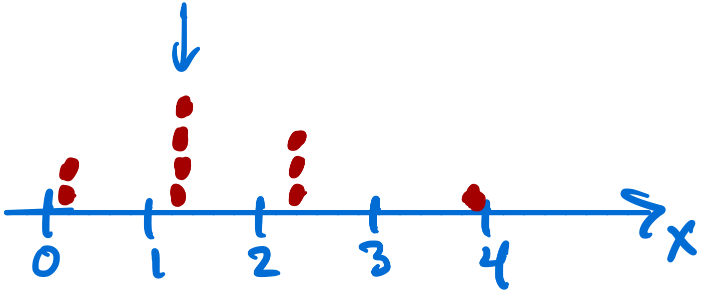
Histograms

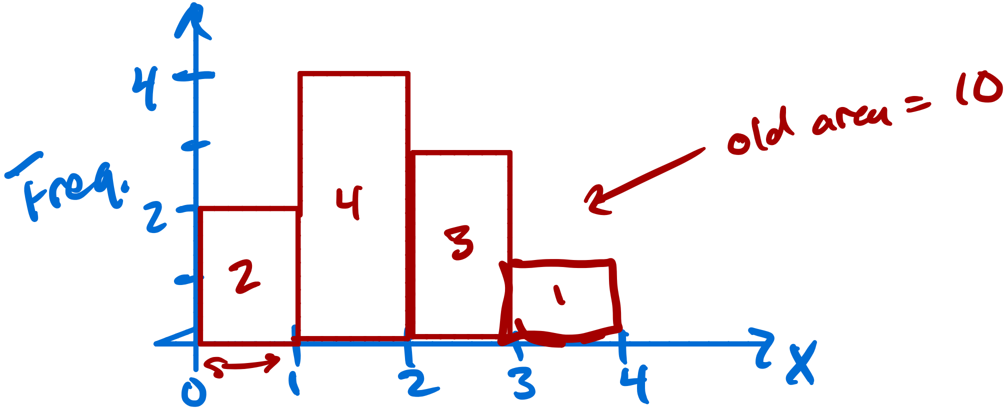
The **histogram** is a graphical representation of the **distribution** of numerical data

Construction:

- Lump or "bin" the observed values of the VOI
  - Bins typically consecutive, non-overlapping, and equal in length
- 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

Example:





Histograms

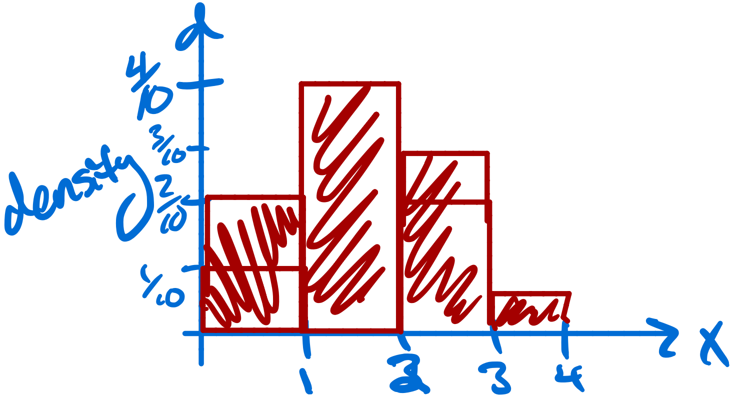
The **histogram** is a graphical representation of the **distribution** of numerical data

Construction:

- Lump or “bin” the observed values of the VOI
  - Bins typically consecutive, non-overlapping, and equal in length
- 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 sum = 1

Example:



$$\text{old Area} = \sum_{\text{boxes}} \overset{\text{old}}{\text{box}}\ \text{heights} \times \text{widths}$$

↓

$$\text{new area} = \sum_{\text{boxes}} \overset{\text{new}}{\text{box}}\ \text{heights} \times \text{widths} \overset{\heartsuit}{=} 1$$

↓

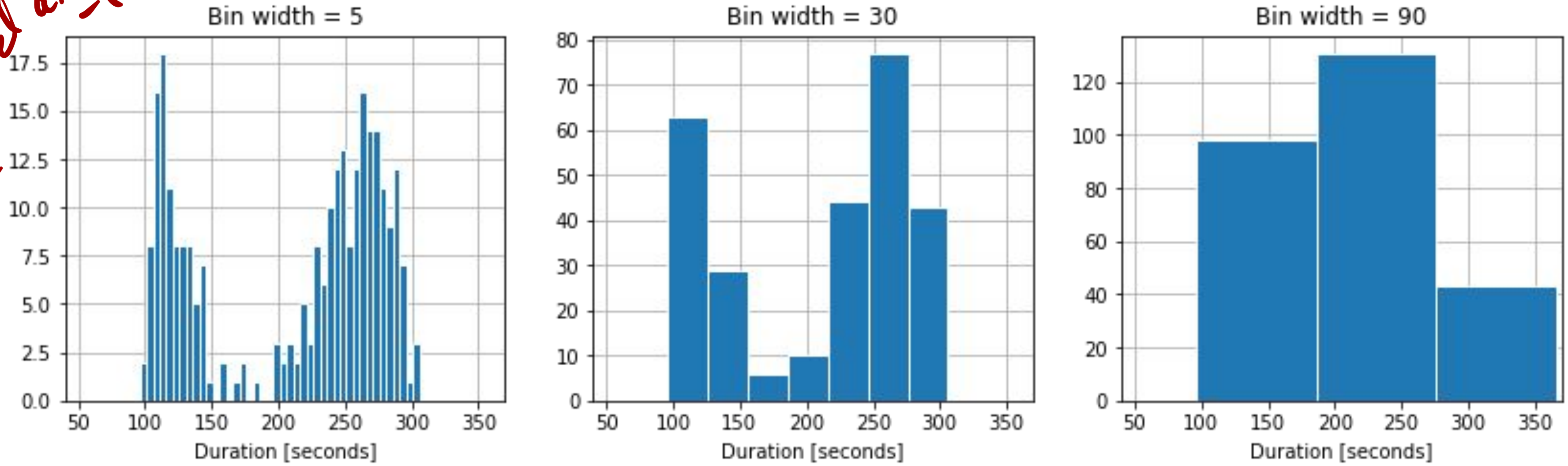
$$\text{pick new hgts} = \frac{\text{old hgts}}{\underbrace{\text{old area}}_{1}} = \boxed{\frac{\text{old hgts}}{10}}$$

$$= \sum_{\text{boxes}} \frac{\text{old hgts}}{\underset{\leftarrow \text{const}}{10}} \times \underline{\text{widths}} = \frac{1}{10} \sum \text{old hgts} = 1$$

Histograms

Note that choosing a different bin width can paint a very different picture of the data

Example: Old Faithful eruption duration data



label axis -1

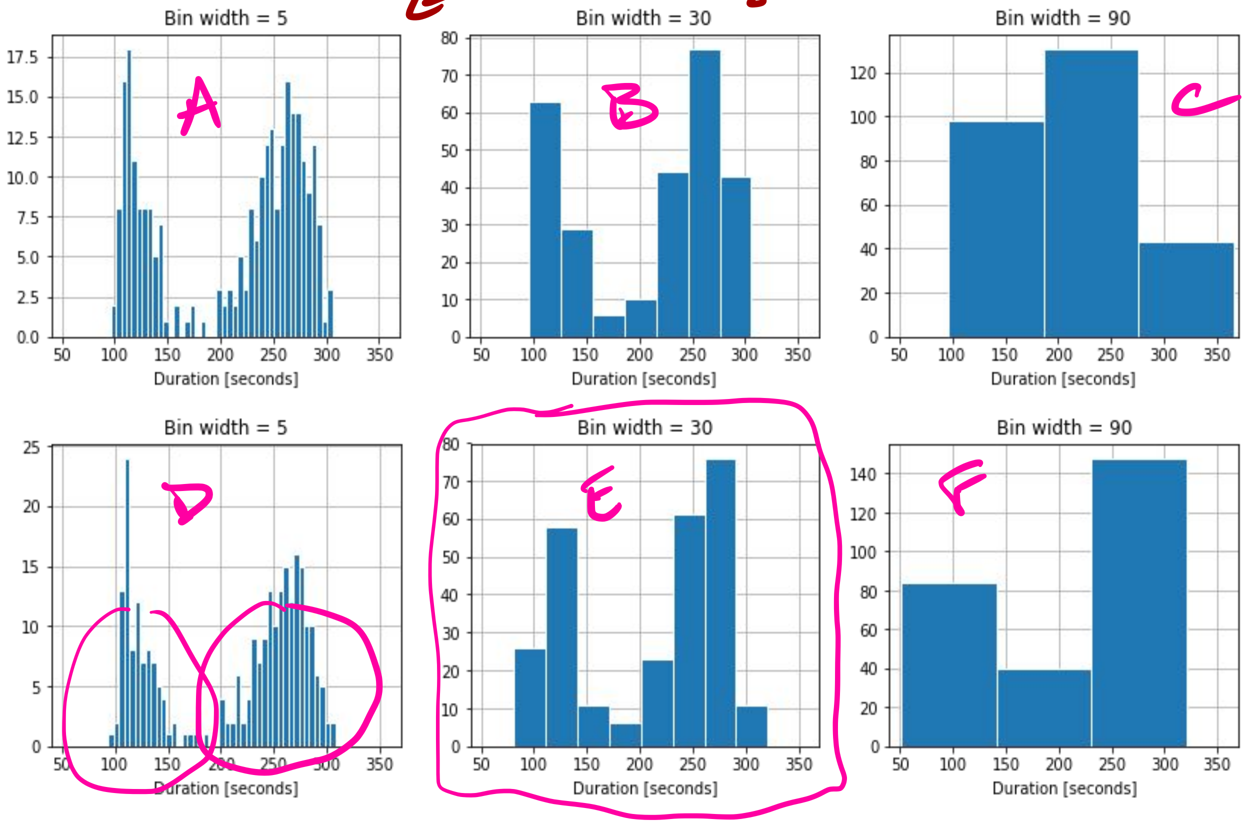
Common choice: Freedman-Diaconis Rule: $\text{bin size} = 2\,\frac{IQR}{n^{1/3}} = 2\,\frac{Q_3 - Q_1}{n^{1/3}}$

Histograms

In practice play around w/ the widths & starting/ending pts

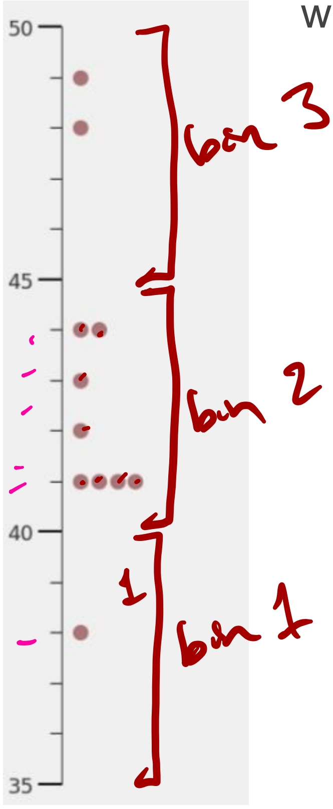
looking at smaller widths can give a sense of where data are

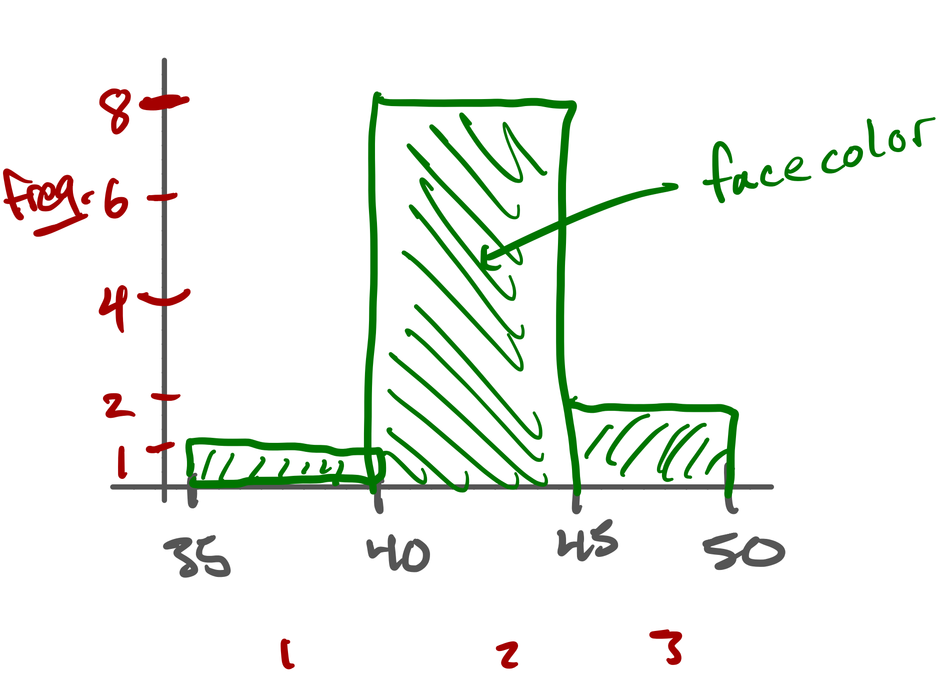
Also consider **where** the bins begin/end



Histograms

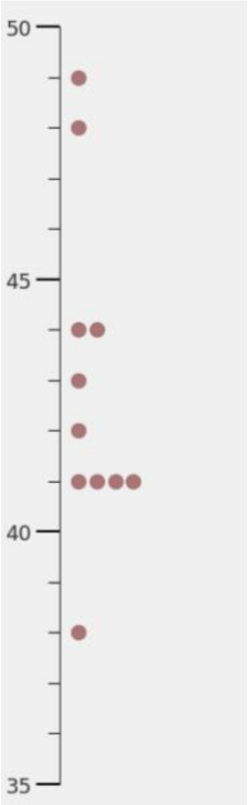
Example: Find the **frequency** histogram with bin width 5 of the data on left, with left-most bin edge at 35.



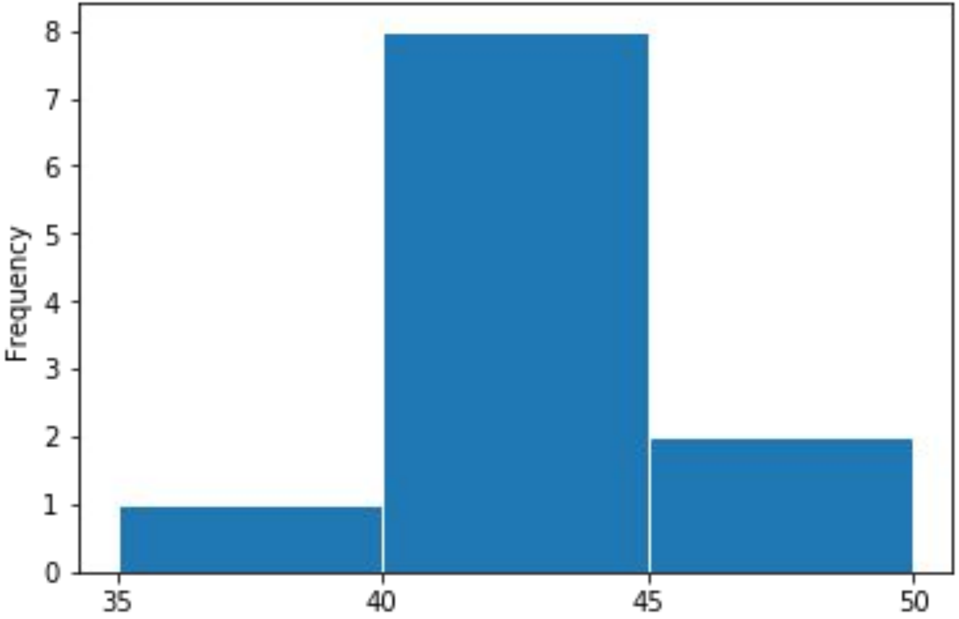


Histograms

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



Solution:



Histograms

mode ⟺ most (data)

Histograms come in a variety of shapes.

unimodal

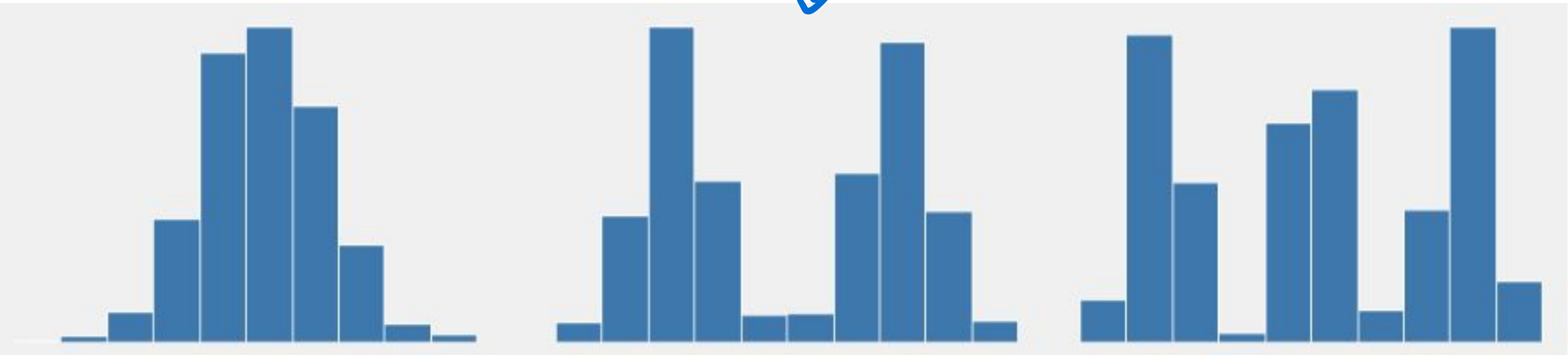
bimodal

multimodal

Histograms

Histograms come in a variety of shapes.

Ex: distribution of ages at a child's birthday party



unimodal bimodal multimodal

Question: what can you say about the data if the histogram is bimodal?

↳ may be sub-population

Histograms

Histograms come in a variety of shapes.

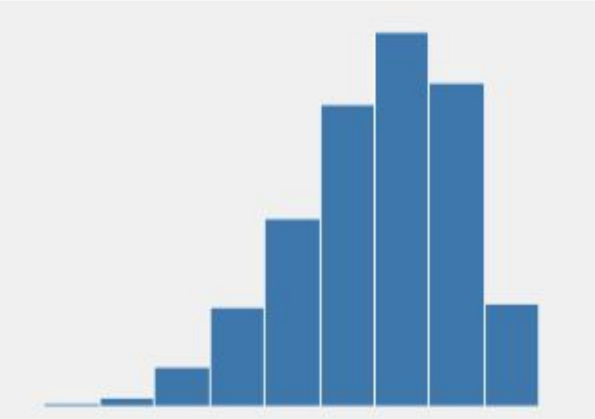
negative skew

symmetric

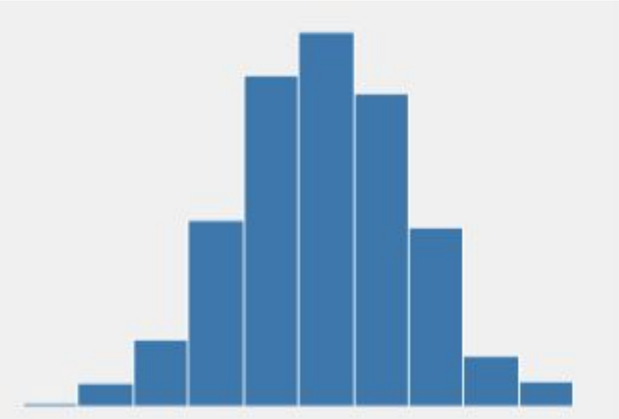
positive skew

Histograms

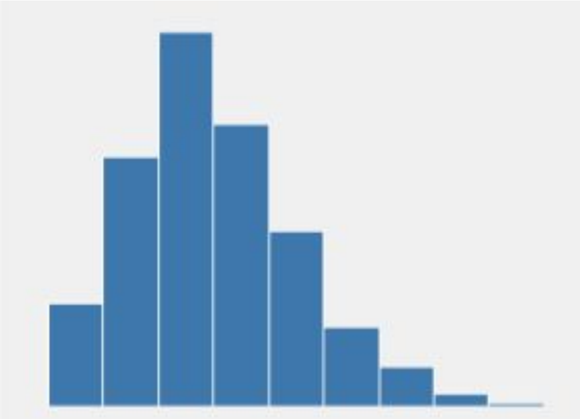
Histograms come in a variety of shapes.



negative skew



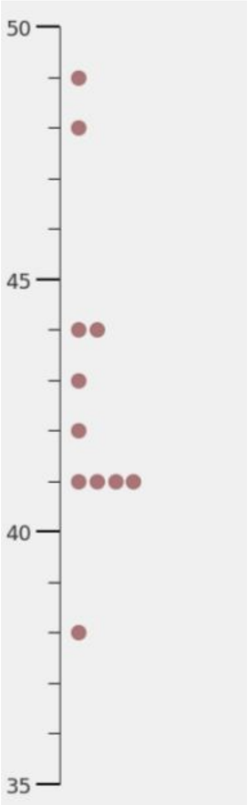
symmetric



positive skew

Quartile refresher

Example: Compute the quartiles and IQR of the data on the left:



1 2 3 4 5 6 7 10 11

8 9

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

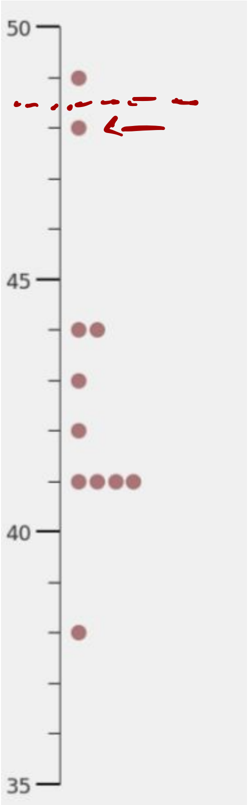
$\tilde{x} = 42$

$Q_1 = 41$

$Q_3 = 44$ ⟶ $\underline{IQR = 3}$

Quartile refresher

Example: Compute the quartiles and IQR of the data on the left:



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

n = 11, odd, so Q_2 is the middle value: Q_2 = 42

Compute Q_1 from first half: 38, 41, 41, 41, 41, 42

→ Q_1 = (41+41)/2 = 41

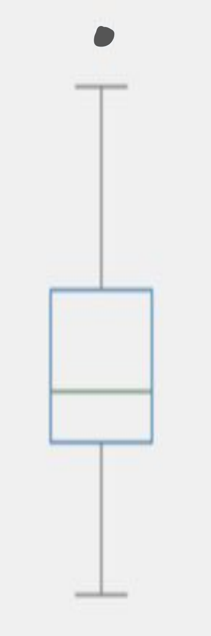
Compute Q_3 from second half: 42, 43, 44, 44, 48, 49

→ Q_3 = (44+44)/2 = 44

IQR = Q_3 - Q_1 = 44 - 41 = 3

Box-whisker plots (aka, boxplots)

Box-whisker plots are a convenient way to visualize data through quartiles

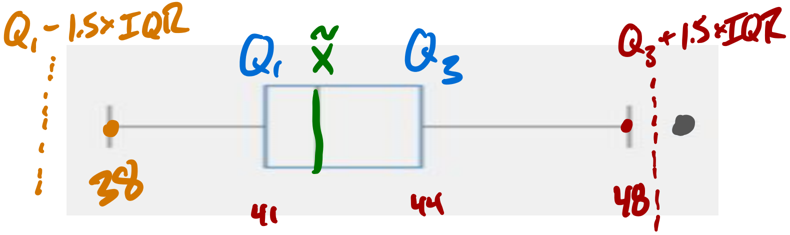


- The **box** extends from Q_1 to Q_3
- The **median line** displays the median $\tilde{x}$
- The **whiskers** extend to farthest data point within 1.5 × IQR of each **quartile**
- The **fliers** or outliers are any points outside of the whiskers
- The width of the box is unimportant
- Can be **horizontally**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

$1.5 \times IQR = 1.5 \times 3 = \underline{4.5}$

$44 + 4.5 = \underline{48.5}$ at most

$41 - 4.5 = \underline{36.5}$ at least

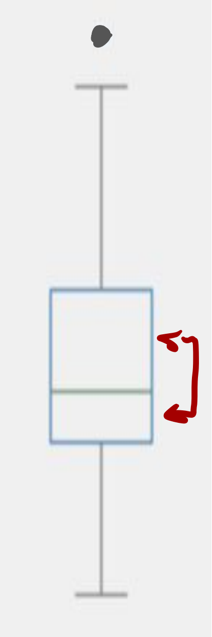


→ inc.

Box-whisker plots

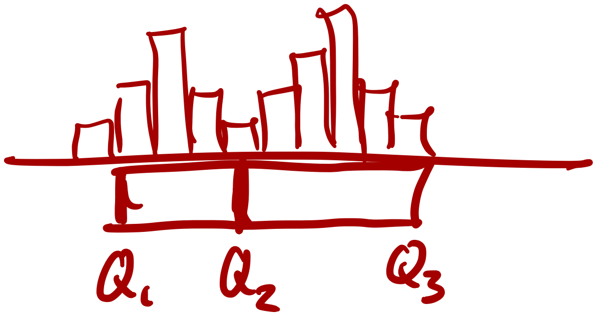
{ HW 1 posted → due next Friday (Feb 1) }

Box-whisker plots are a convenient way to visualize data through quartiles



Box-whisker plots are good because they

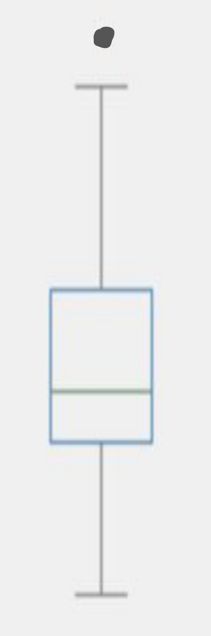
- Depict the center of the data
- Depict the range and IQR
- Depict symmetry / skewness
- Show likely outliers
- When might a box-whisker plot be **misleading**?



- When are box-whisker plots **particularly useful**?

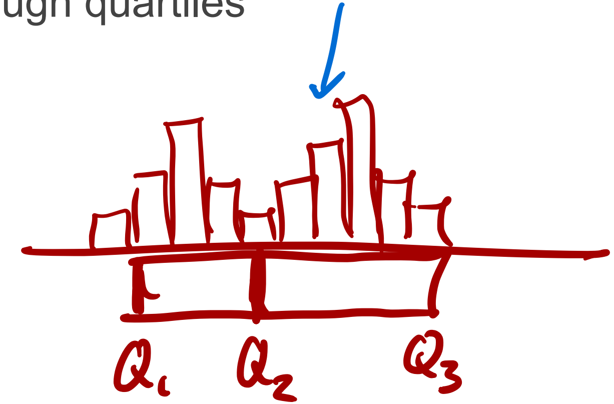
Box-whisker plots

Box-whisker plots are a convenient way to visualize data through quartiles



Box-whisker plots are good because they

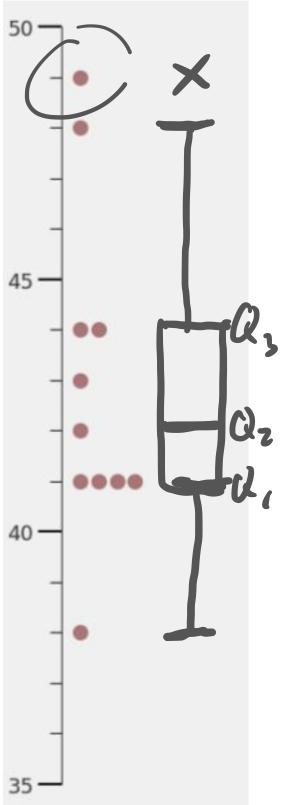
- Depict the center of the data
- Depict the range and IQR
- Depict symmetry / skewness
- Show likely outliers
- When might a box-whisker plot be **misleading**?
  - ✓ No indication of how data are **dispersed** (is there “no-man’s land”?)
  - ✓ # **modes**?
- When are box-whisker plots **particularly useful**? → { comparing data sets



Box-whisker plots

Do include $\tilde{x}$ in upper & lower values of data to get Q_1 & Q_3 if n is odd

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



$$\begin{array}{lccccccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ x = [& 38, & 41, & 41, & 41, & 41, & 42, & 43, & 44, & 44, & 48, & 49] \end{array}$$

$Q_1 = 41$

$Q_3 = 44$

$\tilde{x} = 42$

$Q_1 = 41$

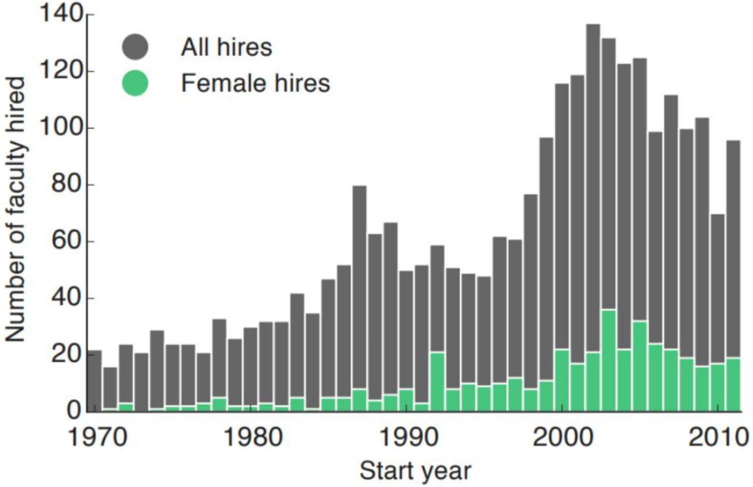
$Q_3 = 44$

$IQR = 3 \rightarrow 1.5 \times IQR = \pm 4.5 \rightarrow$ whiskers can extend out to

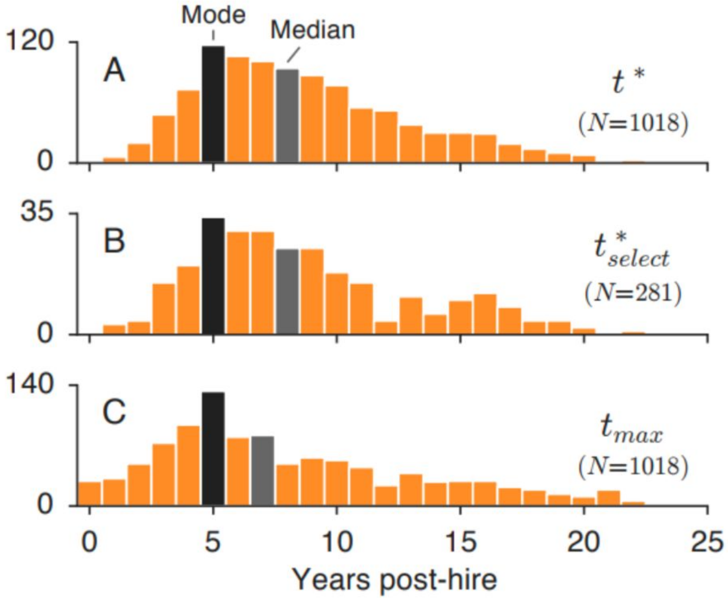
$44 + 4.5 = 48.5$

& $41 - 4.5 = 36.5$

Histograms and boxplots in the wild!

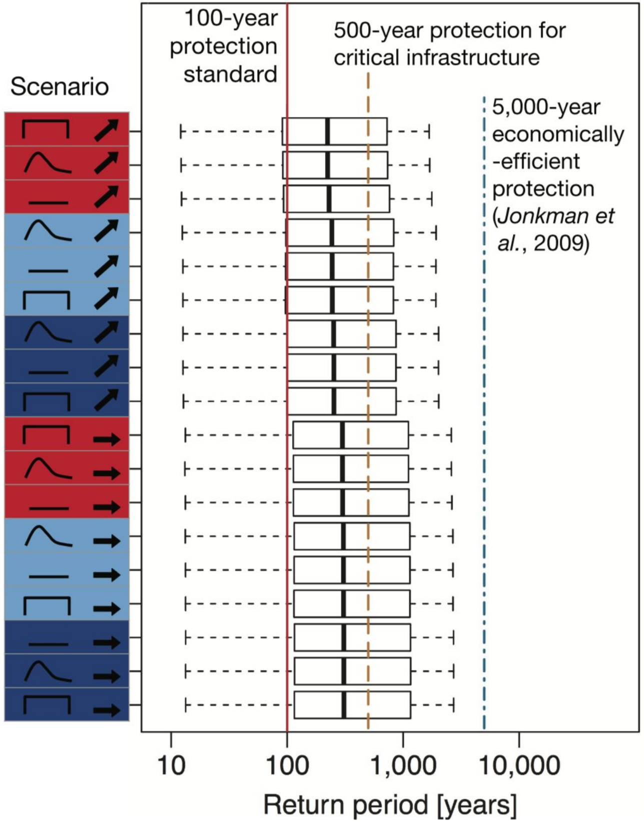


"Gender, Productivity, and Prestige in Computer Science Faculty Hiring Networks" Samuel F. Way, Daniel B. Larremore, and Aaron Clauset. Proc. 2016 World Wide Web Conference (WWW), 1169-1179 (2016).

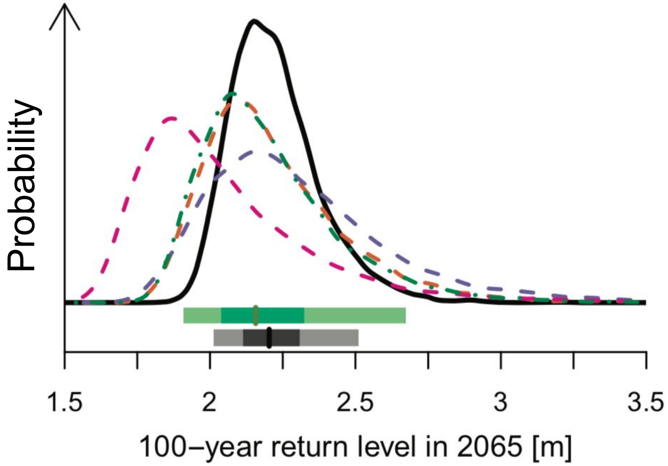


"The misleading narrative of the canonical faculty productivity trajectory" Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. (2016)

Histograms and boxplots in the wild!



Wong, T.E. and K. Keller (2017), Probabilistic Future Flood Risk Scenarios for New Orleans, *Earth's Future*, DOI: 10.1002/2017EF000607.



(flood height in Norfolk, VA)

Wong, T.E., A. Klufas, V. Srikrishnan, and K. Keller (2018), Neglecting Model Structural Uncertainty Underestimates Upper Tails of Flood Hazard, *Environmental Research Letters*, DOI: 10.1088/1748-9326/aacb3d.

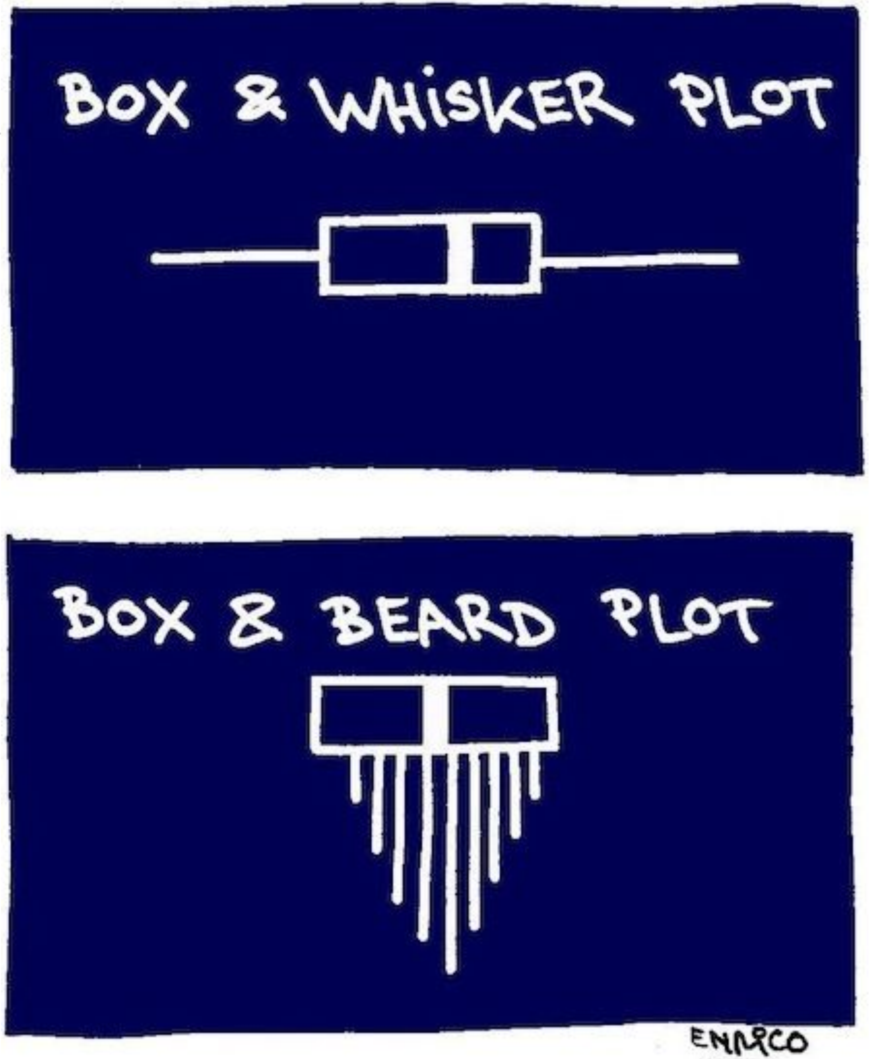
EDA and data visualization

Today we learned…

- How to represent data using a histogram and a box-whisker plot (boxplot)
- And some strengths/weaknesses of each

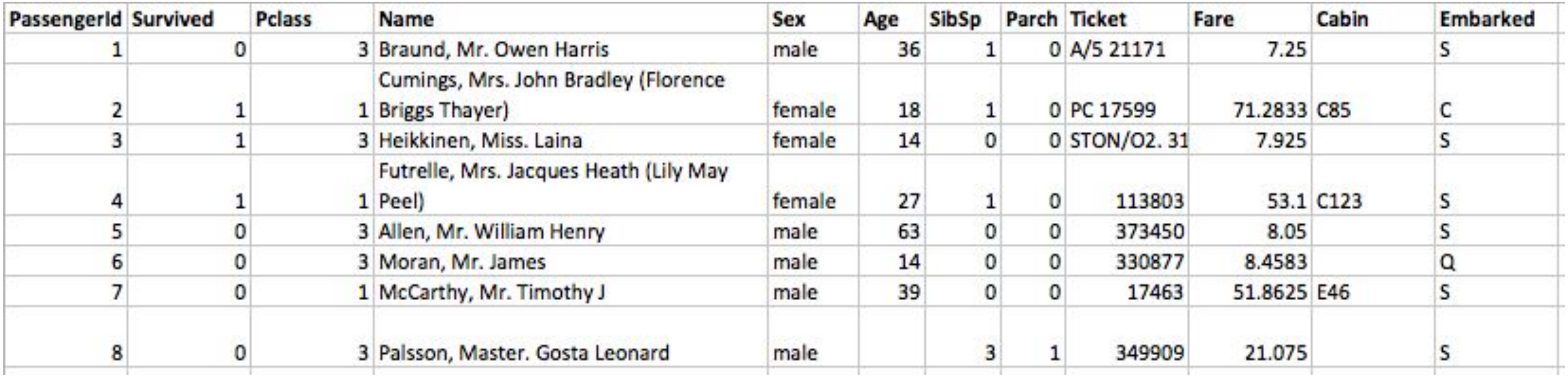
Next time…

- We talk box-beard plots! (Not really!)
- We talk probability! (Probably!)
- (no class/OH on Monday - ~~Labor Day~~)



Cleaning and wrangling data

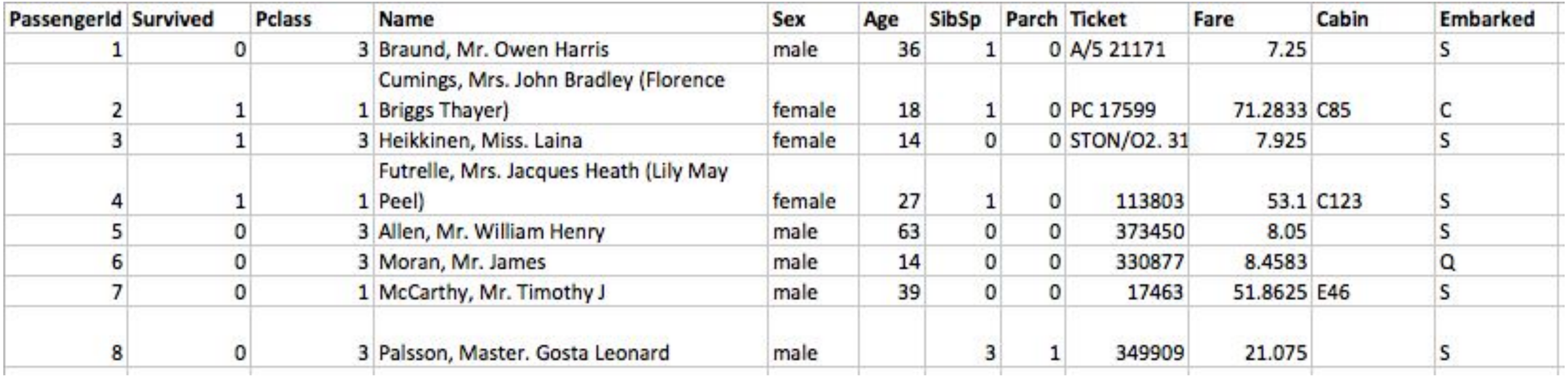
Example: Dirty Titanic data. What looks ***wrong*** to you?



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	36	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	18	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	14	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	27	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	63	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male	14	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	39	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male		3	1	349909	21.075		S

Cleaning and wrangling data

Example: Dirty Titanic data. What looks ***wrong*** to you?



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	36	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	18	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	14	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	27	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	63	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male	14	0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	39	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male		3	1	349909	21.075		S

Today's in-class notebook: **nb03**

1) Remove rows/columns with missing values
2) Creating new columns from old ones (using apply( ) and custom functions)
3) Replacing messy string values with numerical ones