Data Engineering

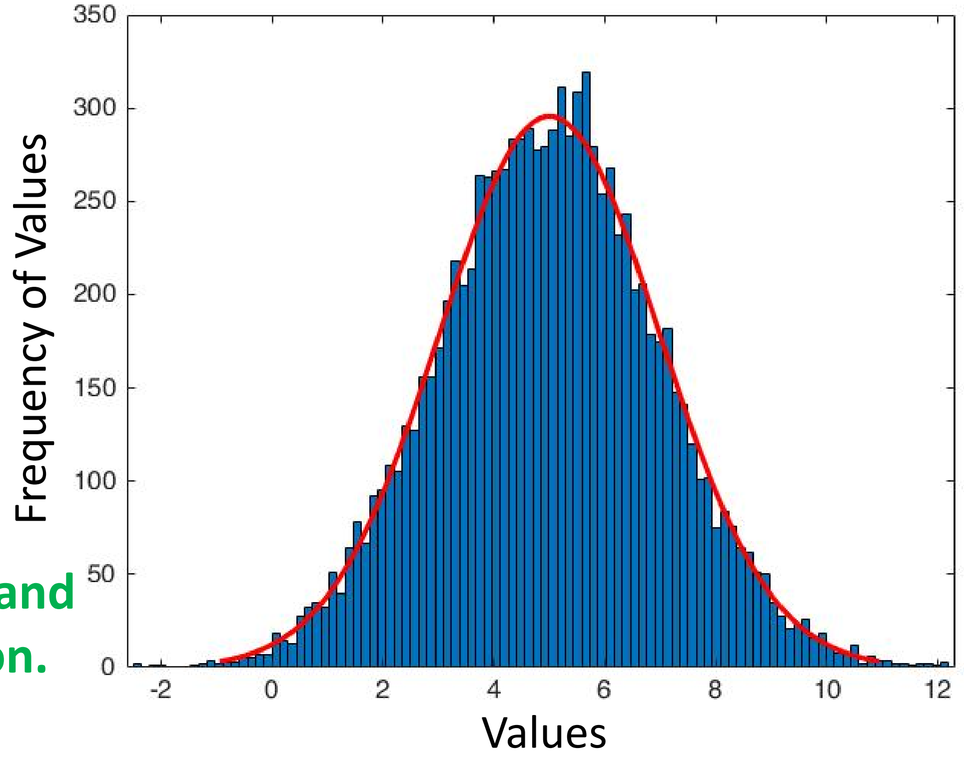
204426

Data Exploration &
Data Visualization

Distribution Shape

Mean, Median and Mode

	X_1	X_2	…	X_{10}
$\mathbf{x}_1$				
…				
$\mathbf{x}_n$				



A distribution in statistics is a function that shows:

- the possible values for a variable (x-axis)
- how often they occur (y-axis).

We can slice a feature/variable and describe it as a data distribution.

Distribution Shape

Mean

- A measure of a central or typical value for a probability distribution.
- The sum of all measurements divided by the number of observations in the data set.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Mean of job performance:

$$\bar{x} = \frac{7+10+11+15+10+10+12+14+16+12}{10} = \frac{117}{10} = 11.7$$

Distribution Shape

Median

- Reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliners.
- The middle value that separates the higher half from the lower half of the data set.
- To compute the middle value, we need to arrange all the numbers from smallest to greatest.
- Then,

$$\tilde{x} = \begin{cases} x_{\frac{(n+1)}{2}}, & if\ n\ is\ odd, \\ \dfrac{\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)}{2}, & if\ n\ is\ even, \end{cases}$$

Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Median of job performance:

n = 10. So, n is even

$$\tilde{x} = \frac{x_5 + x_6}{2} = \frac{11+12}{2} = 11.5$$

11.5

7	**10**	**10**	**10**	**11**	**12**	**12**	**14**	**15**	**16**
				x_5	x_6				

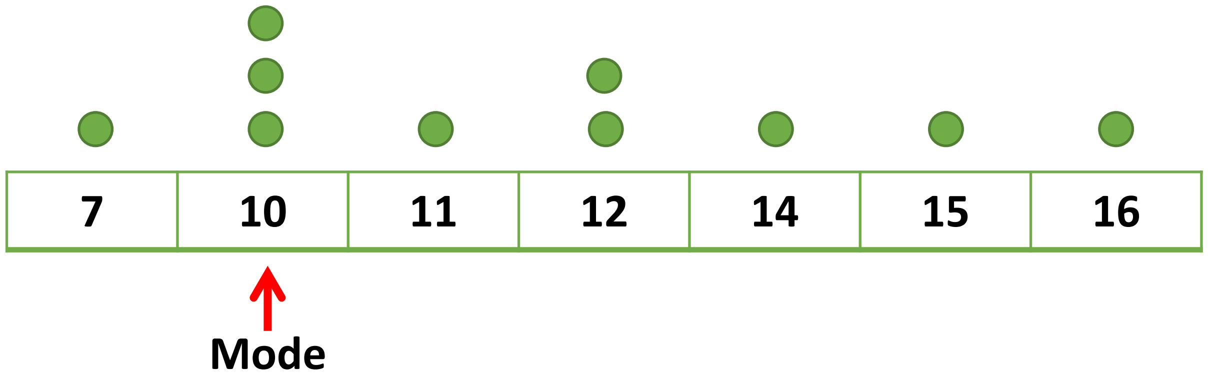
Distribution Shape

Mode

- The most frequent value in the data set.

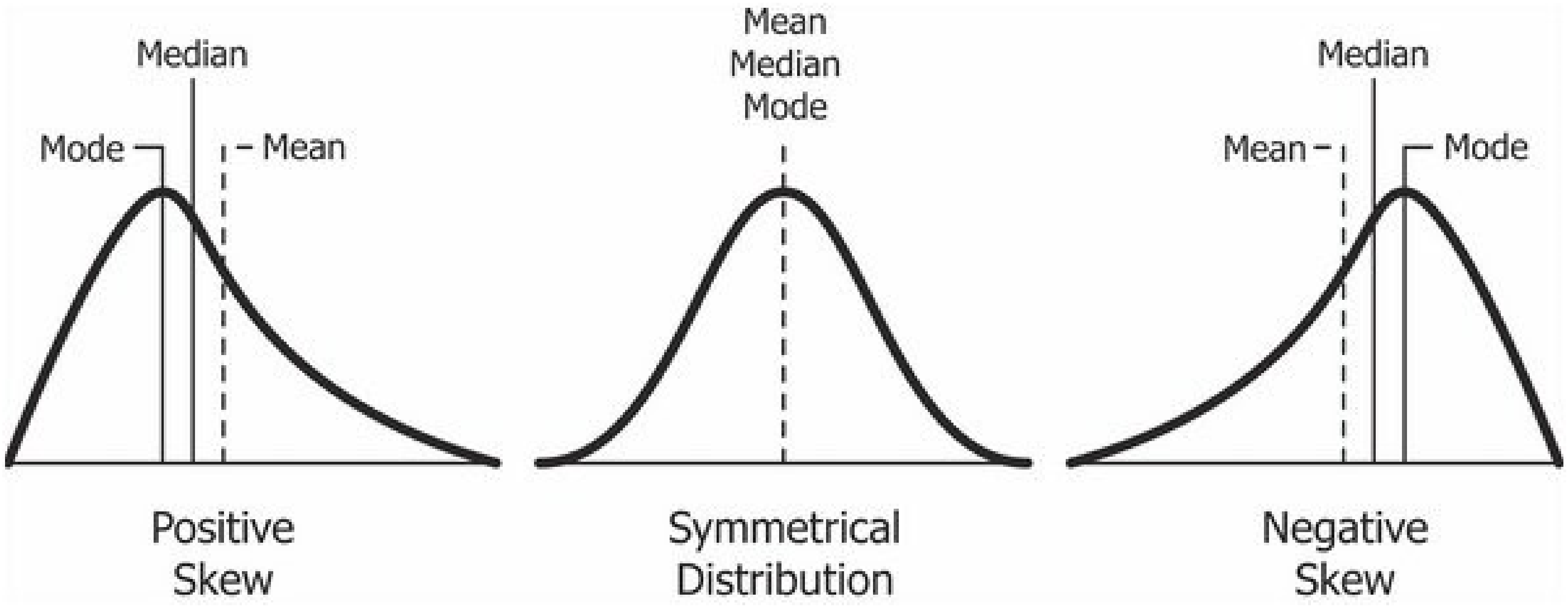
Example:

- Job performance: 7, 10, 11, 15, 10, 10, 12, 14, 16, 12
- Mode of job performance:



Distribution Shape

Geometric visualization of the mode, median and mean of an arbitrary probability density function

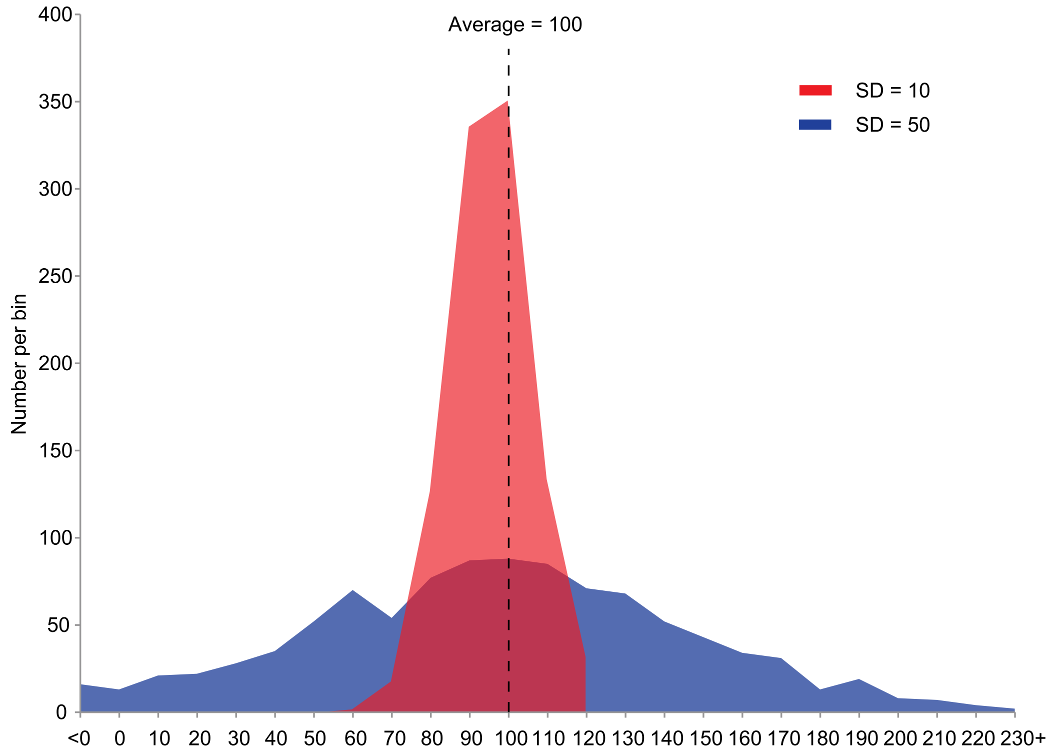


Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Distribution Shape

Standard Deviation (SD, s)

- A measure that is used to quantify the amount of variation or dispersion of a set of data values.
- A low standard deviation indicates that the data points tend to be close to the mean.
- A high standard deviation indicates that the data points are spread out over a wider range of values.



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Comparison_standard_deviations.svg

Distribution Shape

Standard Deviation (SD, s)

The formula for the sample standard deviation is

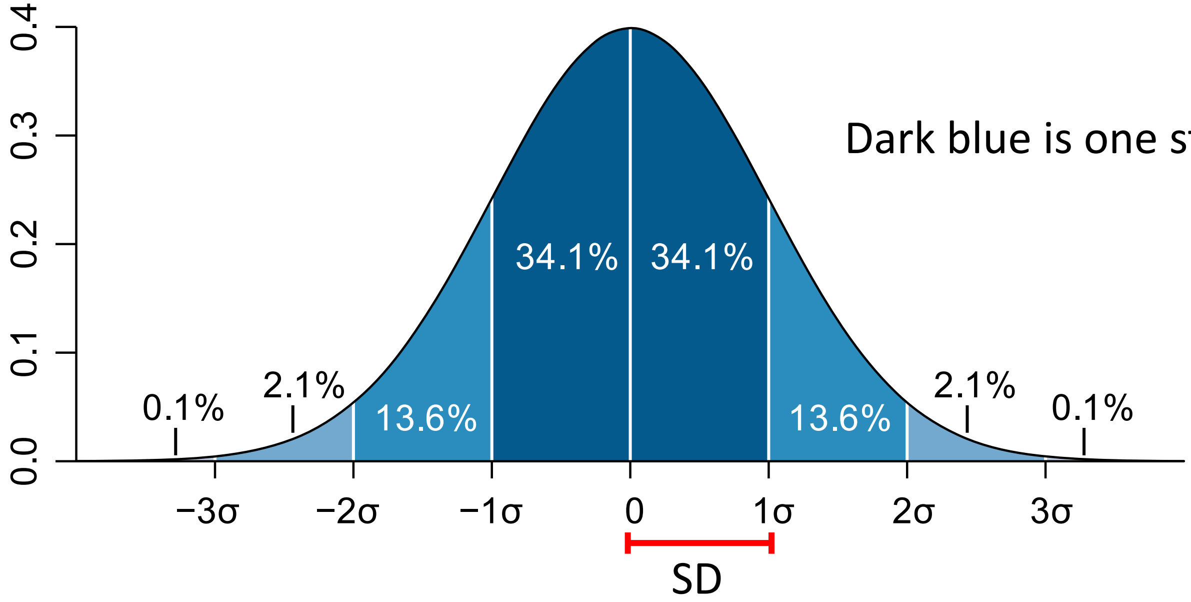
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The formula for the population standard deviation is

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Distribution Shape

Standard Deviation (SD, s)



Dark blue is one standard deviation on either side of the mean.

Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Distribution Shape

Variance

- How far a set of numbers are spread out from their average value.
- It is the square of the standard deviation
- The formula for the sample variance is

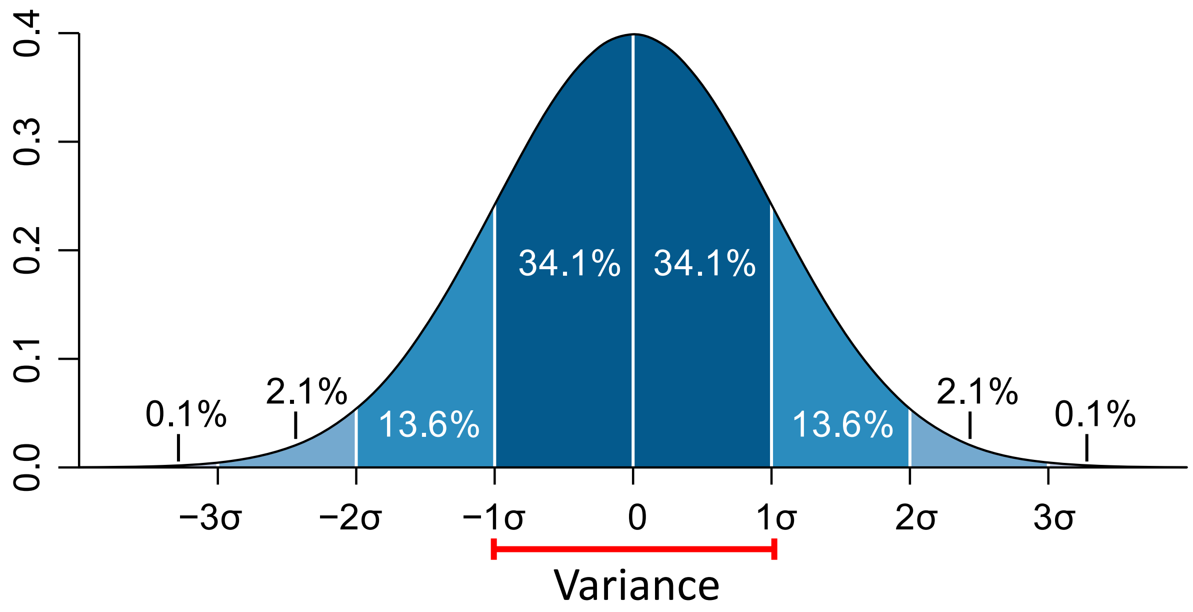
$$s^2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}^{\,2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- The formula for the population variance is

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Distribution Shape

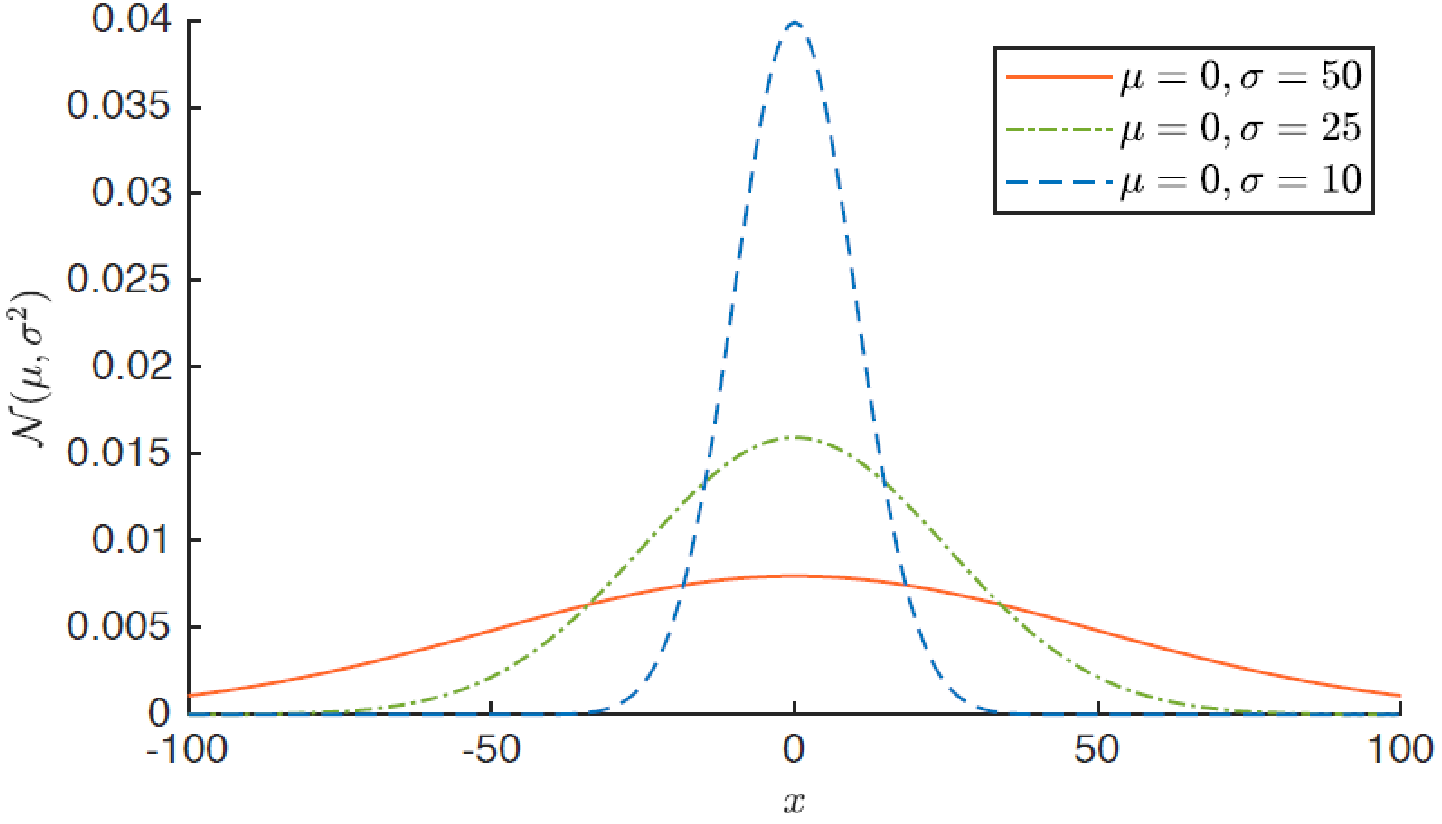
Variance



Source:
https://en.wikipedia.org/wiki/Standard_deviation#/media/File:Standard_deviation_diagram.svg

Distribution Shape

Standard Deviation and Variance

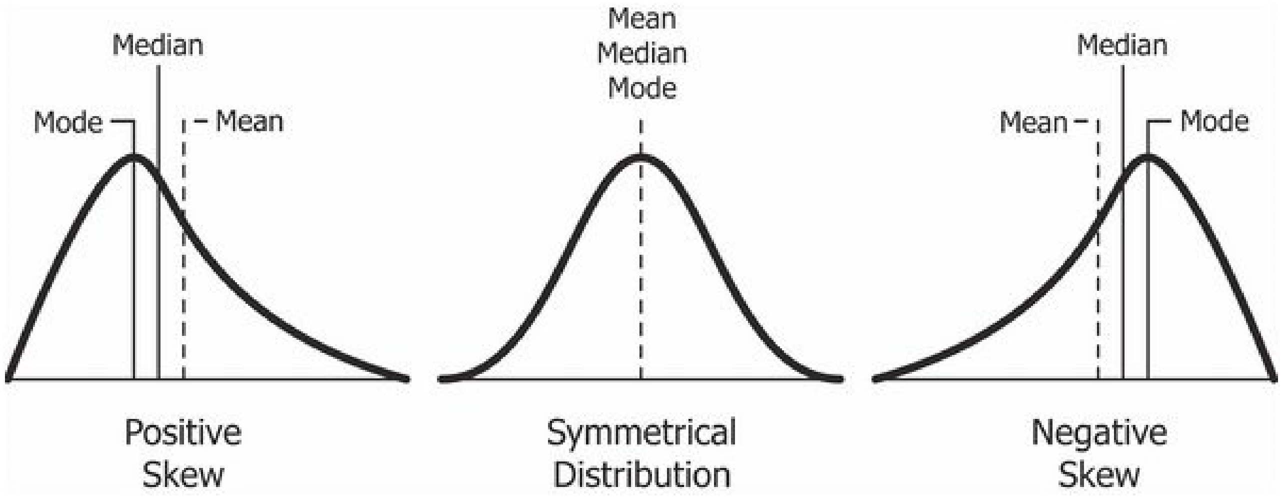


Distribution Shape

Skewness

- Skewness is usually described as a measure of a **dataset's symmetry** – or lack of symmetry.
- The normal distribution has a skewness of 0.
- The skewness can be calculated by

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{2}{3}}}$$



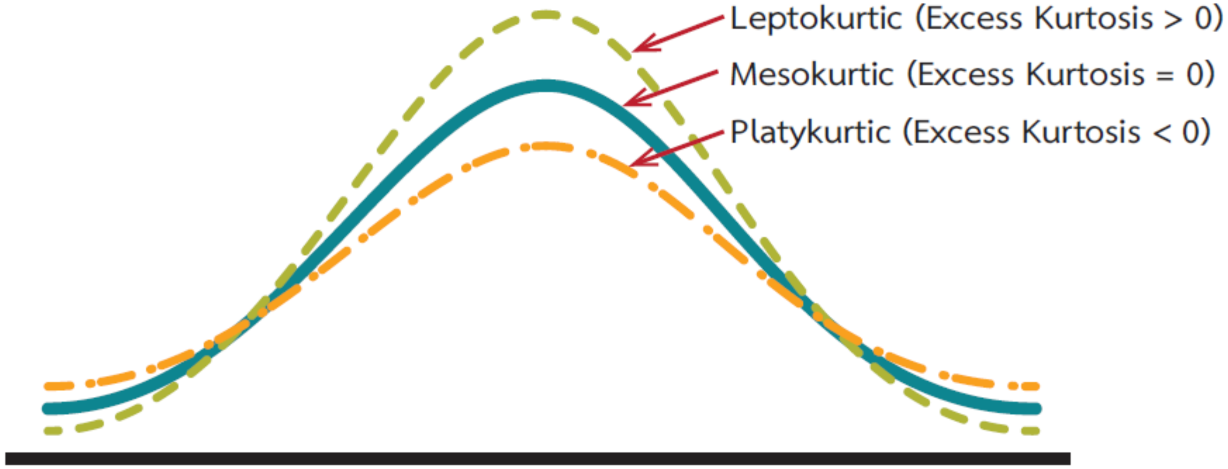
Source: https://codeburst.io/2-important-statistics-terms-you-need-to-know-in-data-science-skewness-and-kurtosis-388fef94eeaa

Distribution Shape

Kurtosis

- Measures the **tail-heaviness of the distribution**.
- The excess kurtosis for a standard normal distribution is 0.
- The excess kurtosis can be calculated by

$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} - 3$$



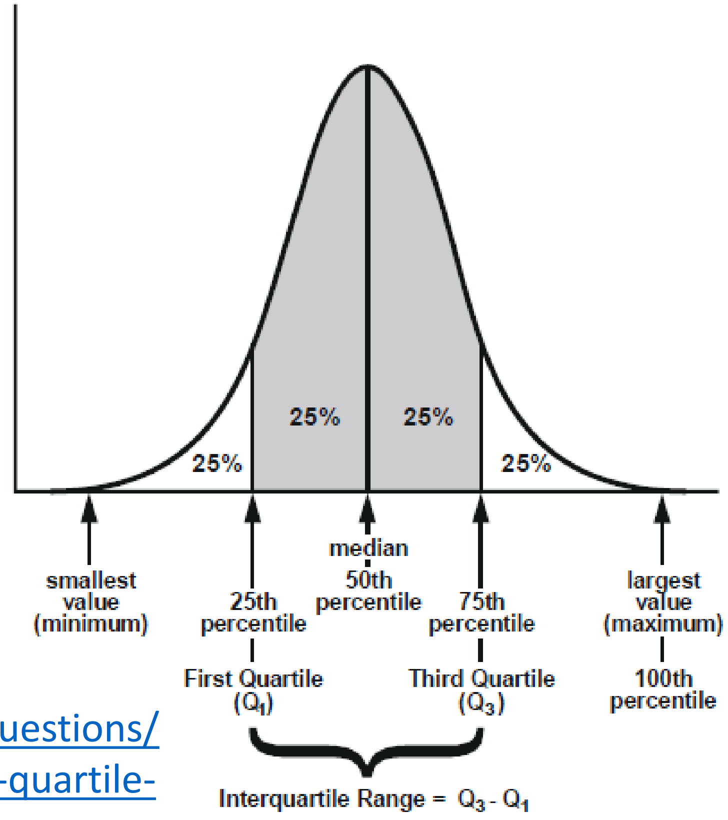
Distribution Shape

Interquartile Range

- The difference between the first and the third sample quartiles.
- This give the range of middle 50% of the data
- It is found from the following

$$IQR = q(0.75) - q(0.25)$$

- **Lower limit**: $LL = q(0.25) - 1.5 \times IQR$
- **Upper limit**: $\mathrm{U}L = q(0.75) + 1.5 \times IQR$
- Observations outside these limits are **potential outliers**.



Source: https://stats.stackexchange.com/questions/470193/representing-quantile-like-quartile-in-form-of-normal-distribution-curve

Distribution Shape

Histogram

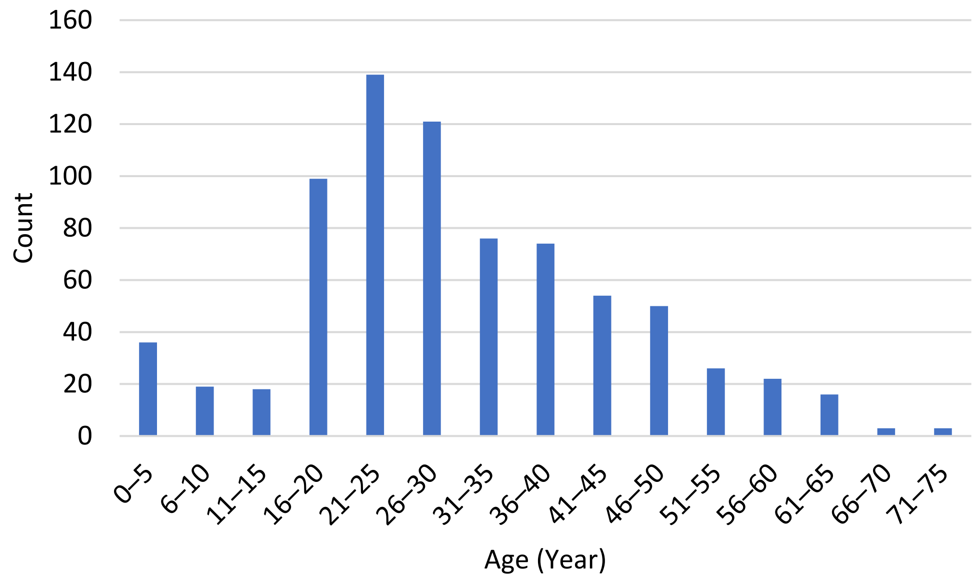
- A histogram takes as input a **numeric variable** only.
- The variable is cut into several bins
- The number of observation per bin is represented by the height of the bar.

	Age	…	…	…
$\mathbf{x}_1$	35			
$\mathbf{x}_2$	25			
…	…			
$\mathbf{x}_n$	62			

$Age \in \{0,75\}$

Data

Bin of Age	Count
0-5	36
6-10	19
11-15	18
…	…
61-65	16
66-70	3
71-75	3



Histogram

Distribution Shape

Histogram bin widths

- **Sturges' rule**

$$k = 1 + \log_2 n$$

  - k is the number of bins.
  - The bin width *h* is obtained by taking the range of the sample data and dividing it into the requisite number of bins.
- **Normal reference rule**

$$h \approx 3.5 \times \sigma \times n^{-\frac{1}{3}}$$

- **Scott's rule**

$$h = 3.5 \times s \times n^{-\frac{1}{3}}$$

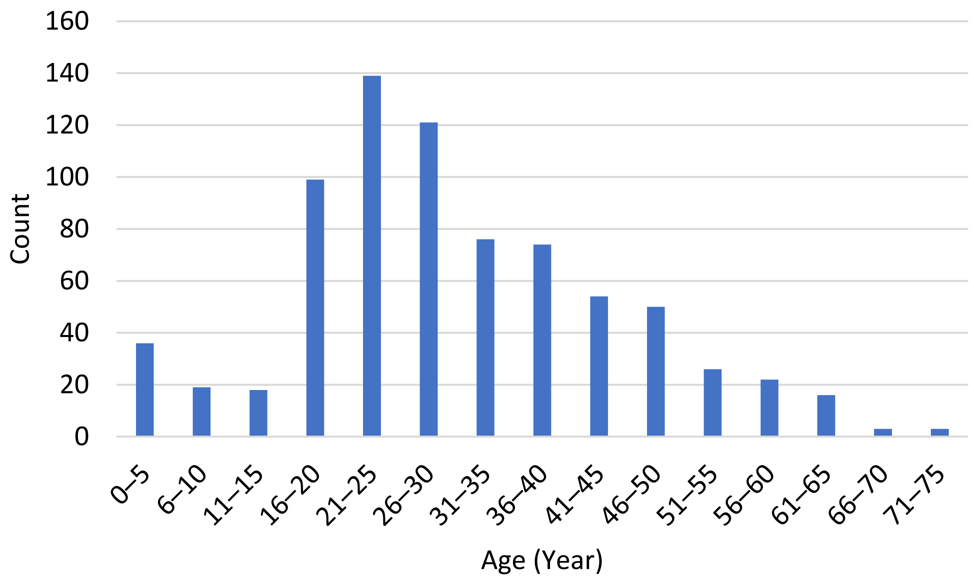
- **Freedman-Diaconis rule**

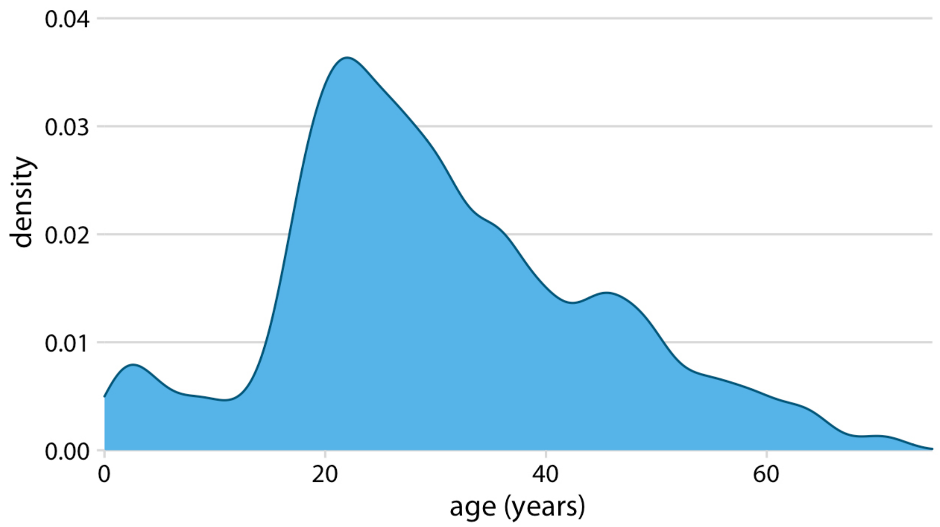
$$h = 2 \times IQR \times n^{-\frac{1}{3}}$$

Distribution Shape

Density

- Visualize the underlying probability distribution of the data by drawing an appropriate **continuous curve**.
- This curve needs to be estimated from the data using **kernel density estimation**.





Density plot

Distribution Shape

Kernel density estimation

- The univariate kernel estimator is given by

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- $K(t)$ is a kernel
- h > 0 is a smoothing parameter called the bandwidth, and can be estimated by

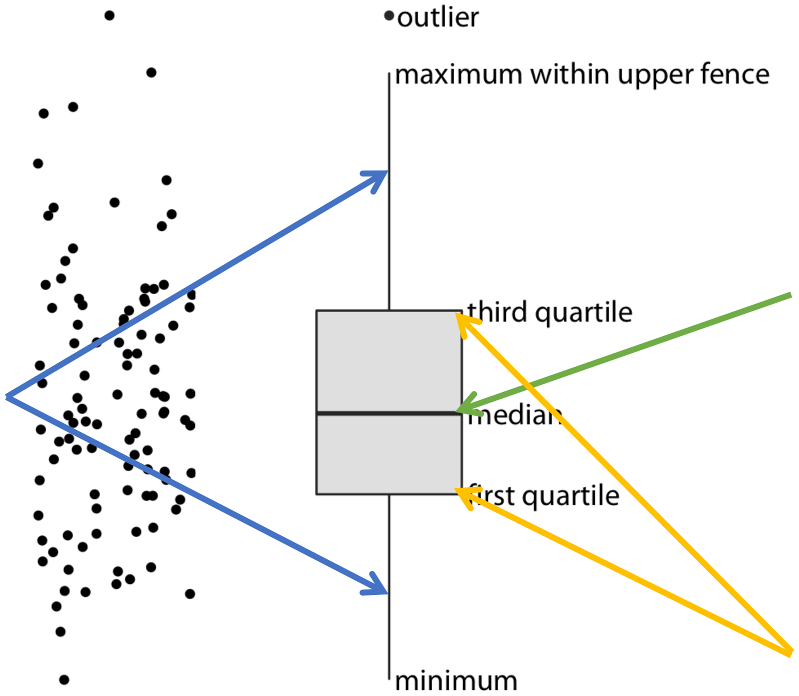
$$h = 0.786 \times IQR \times n^{-\frac{1}{5}}$$

Distribution Shape

Boxplot

Boxplot gives a nice summary of <u>one or several distributions</u>.

Extreme lines show the highest and lowest value excluding outliers



Anatomy of a boxplot

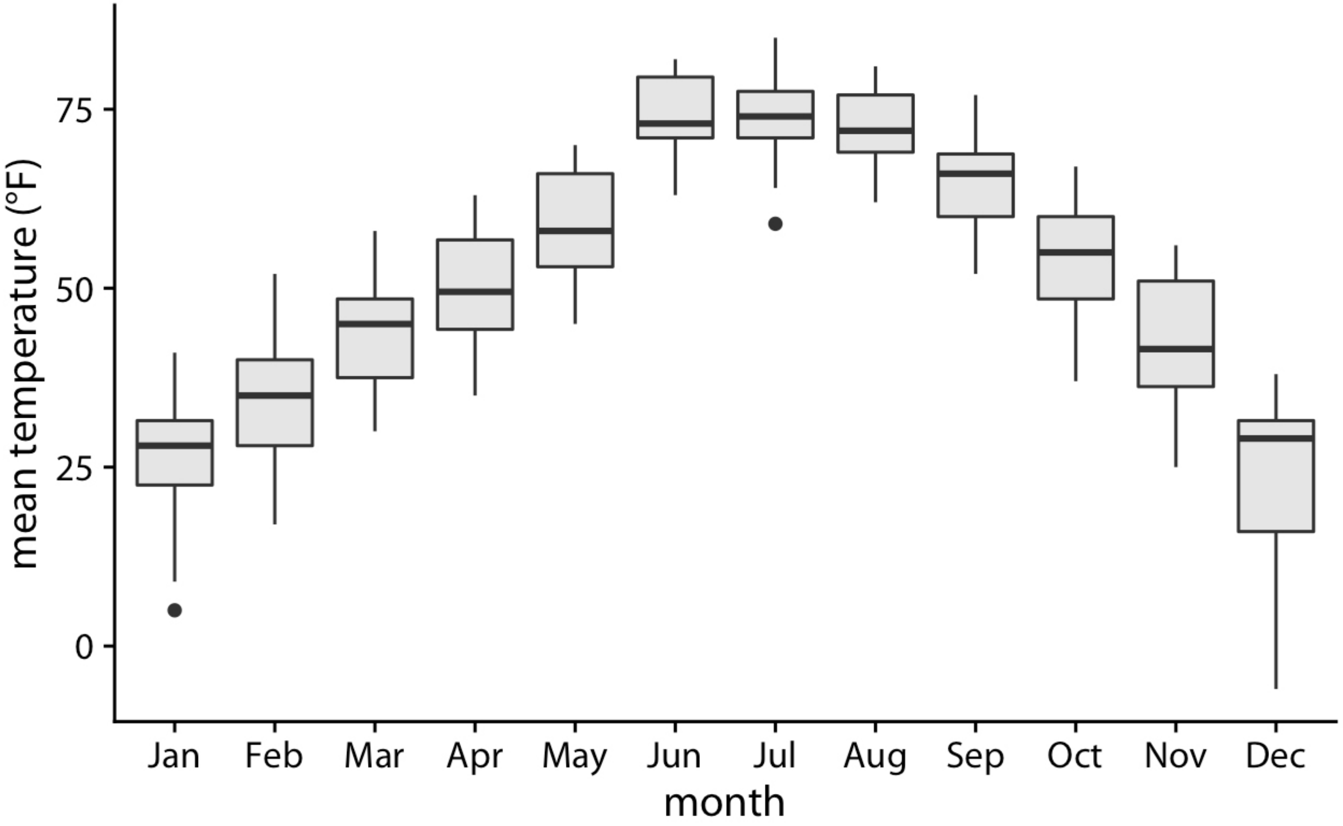
Line that divides the box into 2 parts represents the median of the data

The end of the box shows the third and first quartiles.

Image Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Distribution Shape

Boxplot

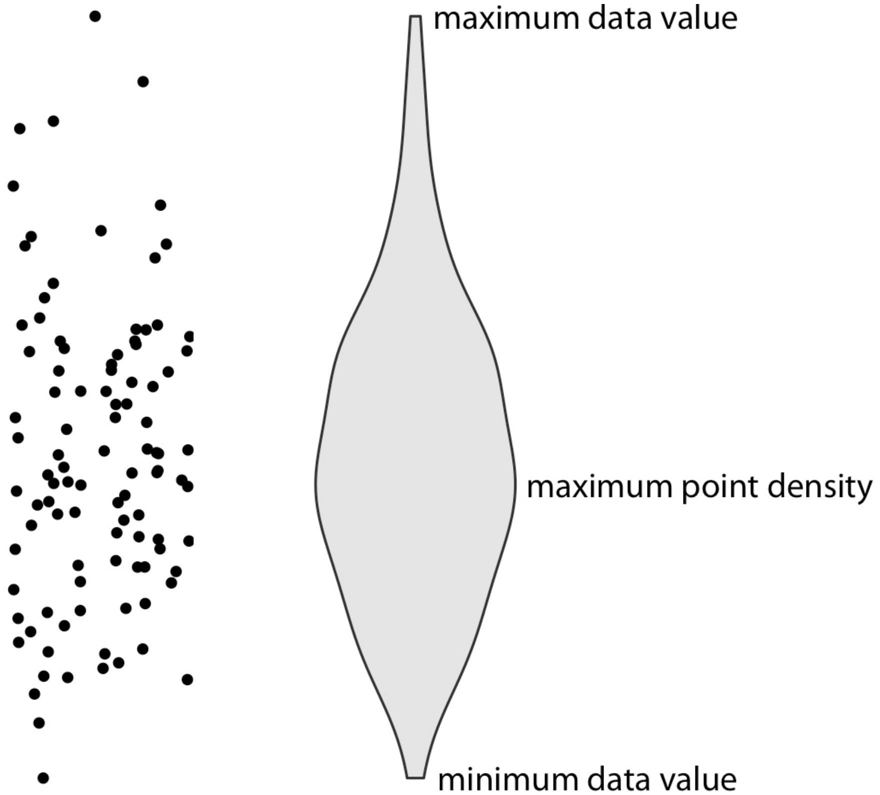


Mean daily temperatures in Lincoln, NE, visualized as boxplots.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

Distribution Shape

Violin Plot



Anatomy of a violin plot

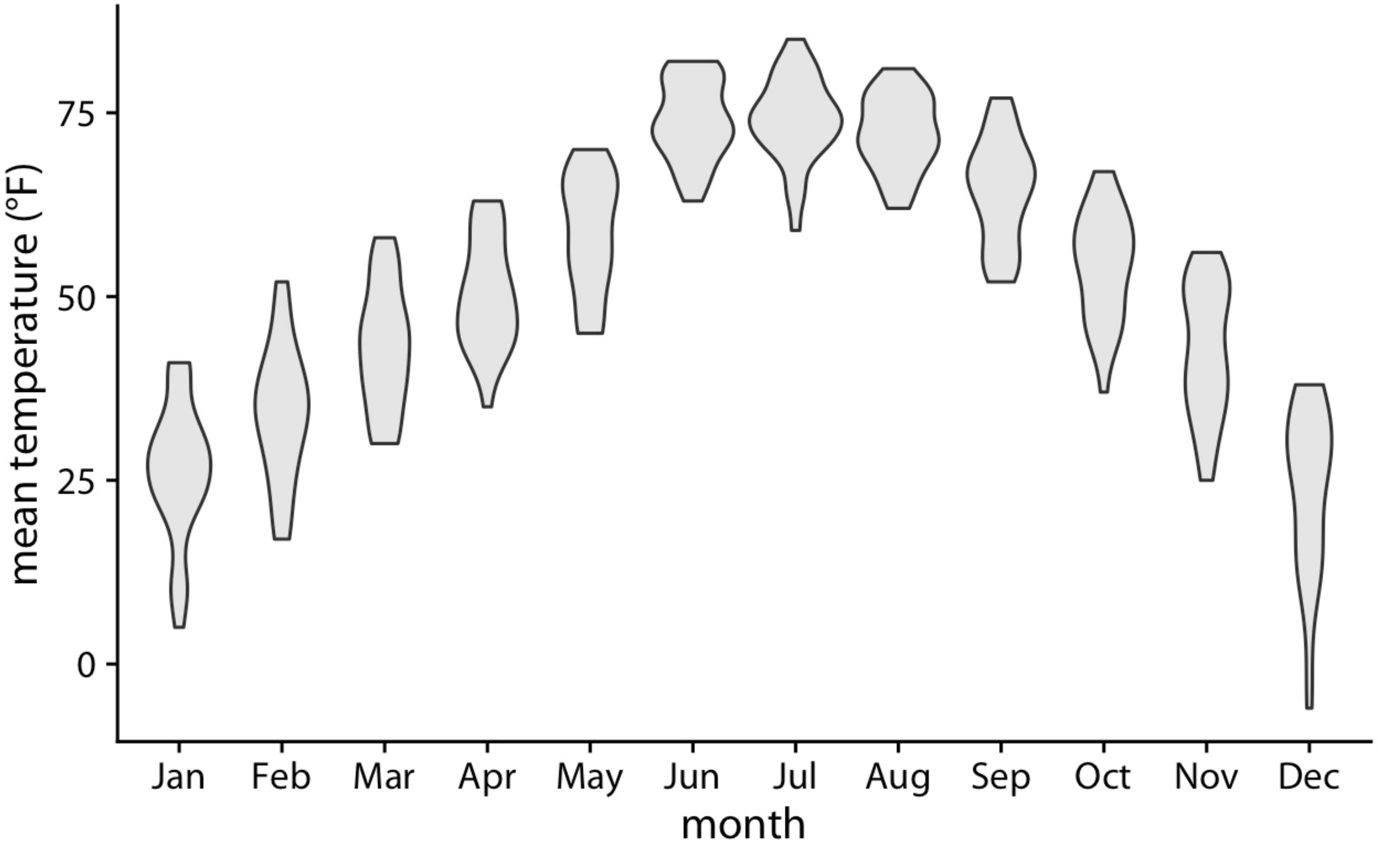
Image Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O’Reilly Media, Inc.

- Visualize the distribution of a numeric variable for one or several groups.
- It is really close from a **boxplot** <u>but</u> allows a deeper understanding of the distribution.

- Violins are particularly adapted when
  - the amount of data is huge
  - showing individual observations gets impossible.

Distribution Shape

Violin Plot



Mean daily temperatures in Lincoln, NE, visualized as violin plot.

Source: Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O'Reilly Media, Inc.

References

- Joanes, D. N. and C. A. Gill (1998). “Comparing measures of sample skewness and kurtosis.” In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.1, pp. 183–189.
- Raymond H. Meyers, Ronald E. Walpole and, Sharon L. Meyers, and Keying E. Ye (2012). *Probability & Statistics for Engineers & Scientists*. 9th. USA: Prentice Hall.
- Westfall, Peter H. (2014). “Kurtosis as Peakedness” In: *The American Statistician* 68.3, pp. 191–195.
- Claus O. Wilke (2019), Fundamentals of Data Visualization. USA: O’Reilly Media, Inc.
- Martinez, Mendy L., Martinez, Angel R. and Solka, Jeffrey L. (2017). Exploratory Data Analysis with MATLAB. USA: CRC Press.