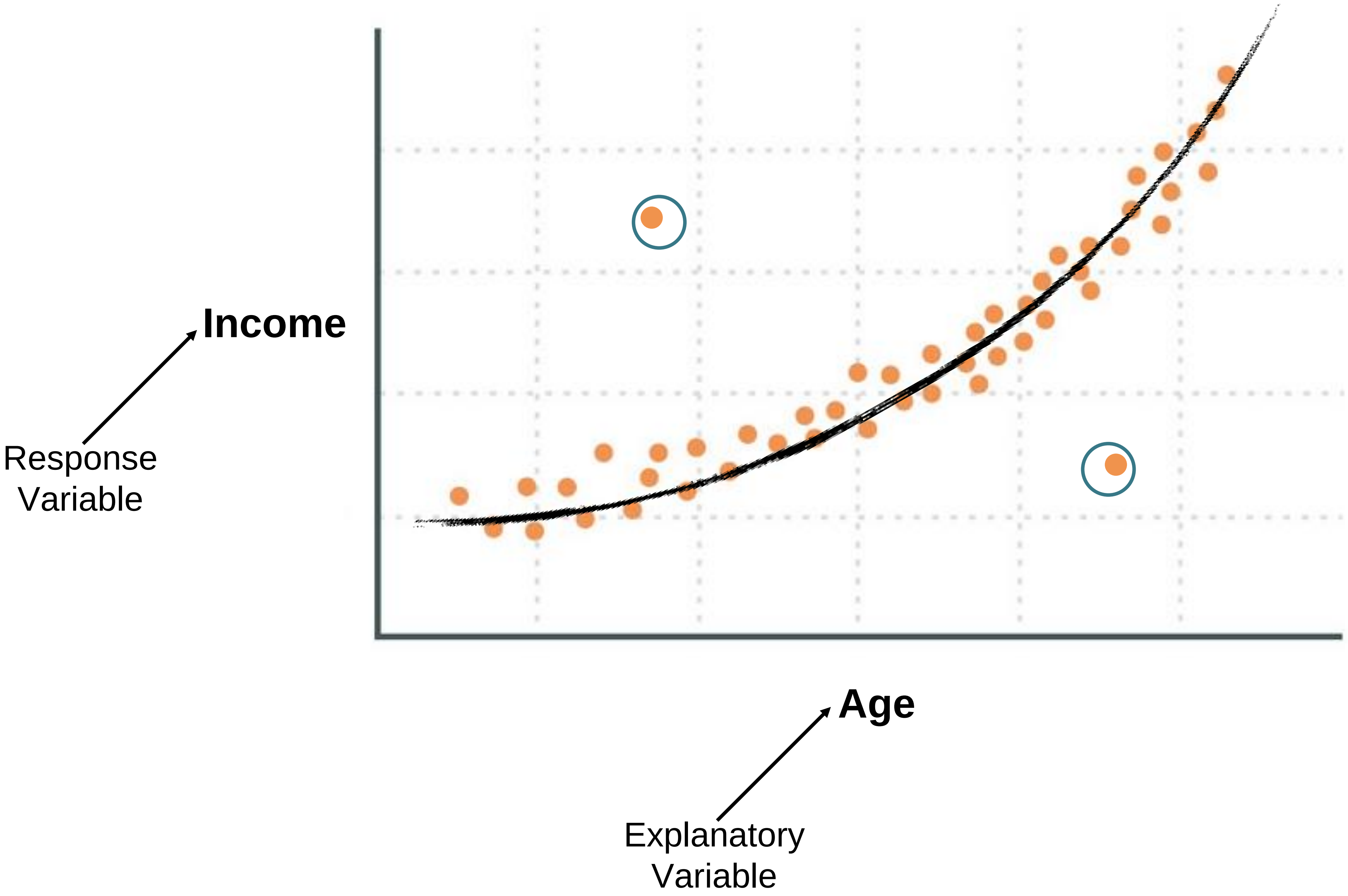
Data Visualisation

Visualising Numerical Data

Scatterplot



Income

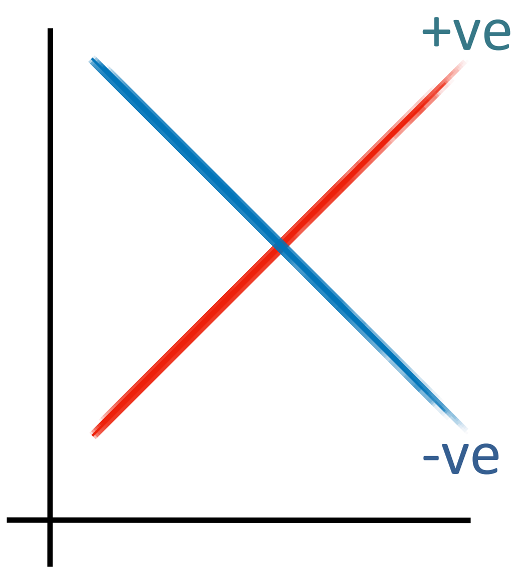
Response Variable

Age

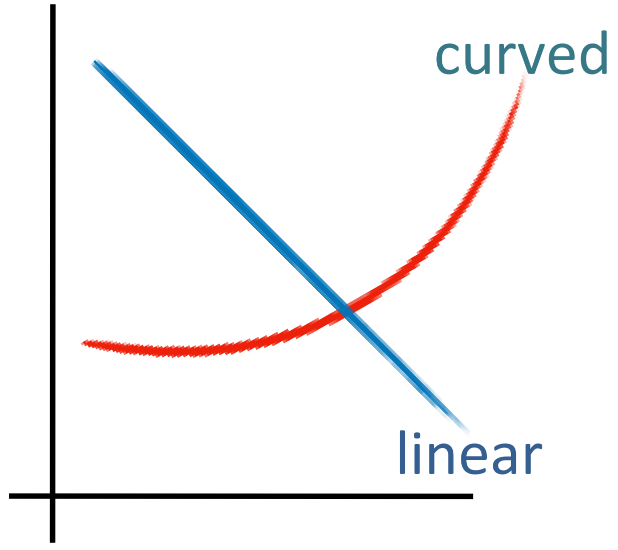
Explanatory Variable

Characteristics of Relationship

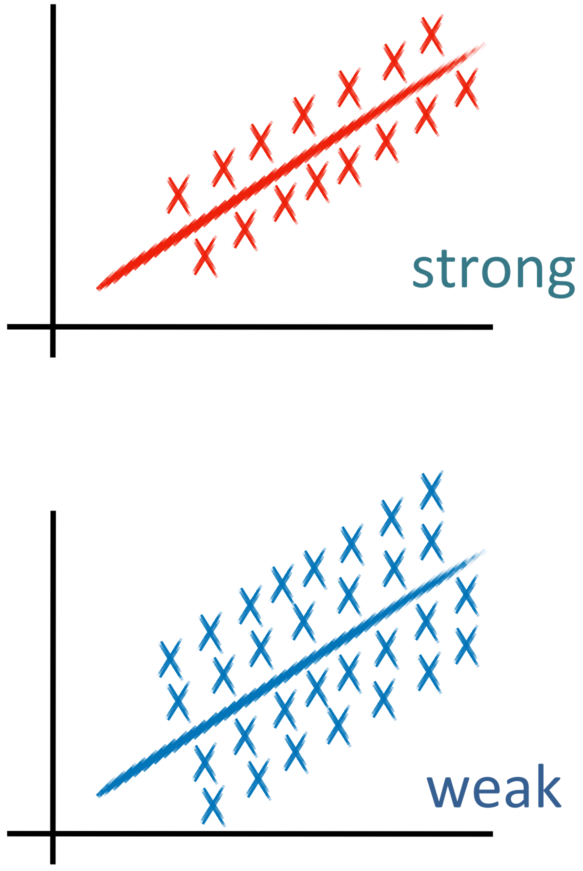
Direction



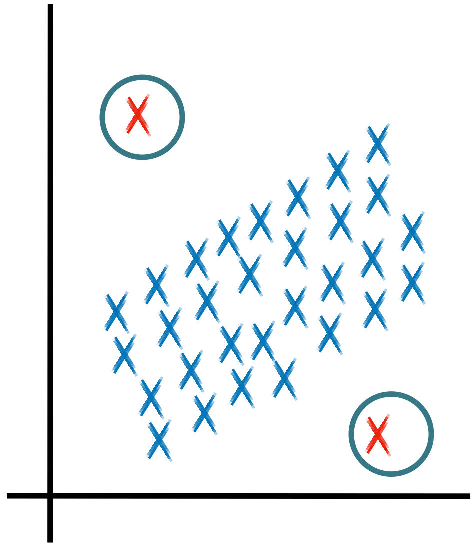
Shape



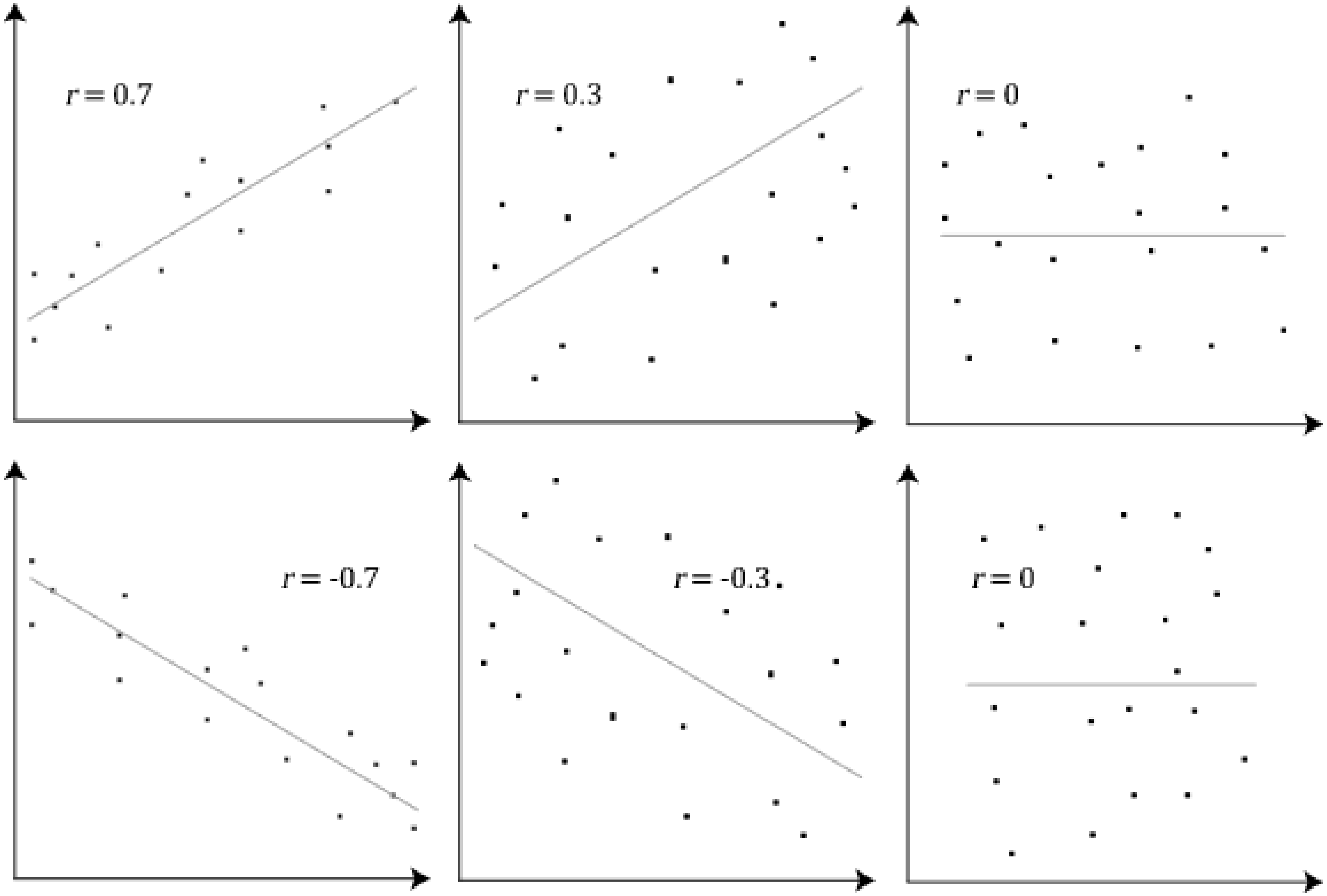
Strength



Outliers



Correlation (example)

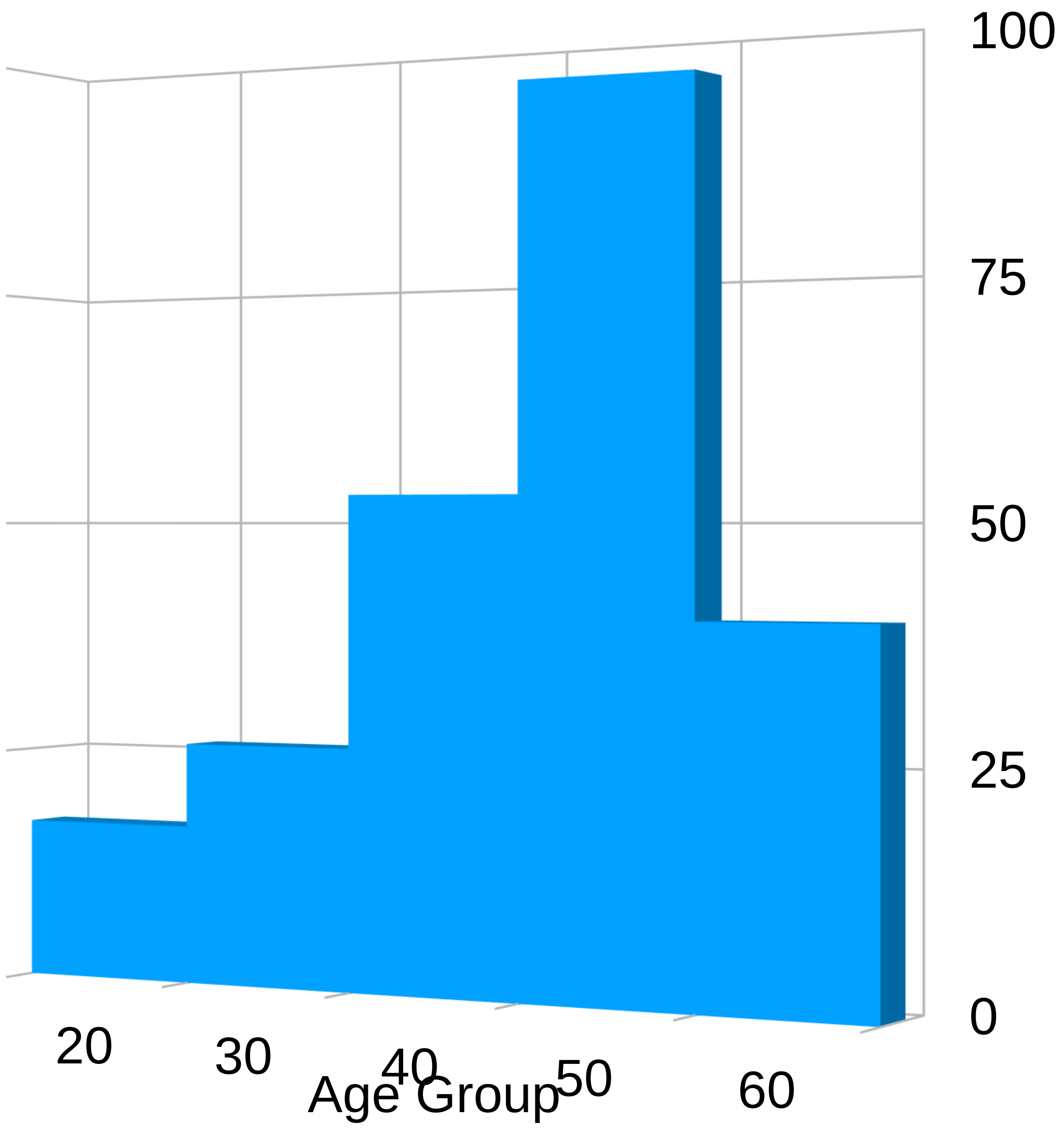


Histograms

- Help to view <u>*data density*</u>

- Help to see <u>*shape of distribution*</u>

1) Skewness
2) Modality

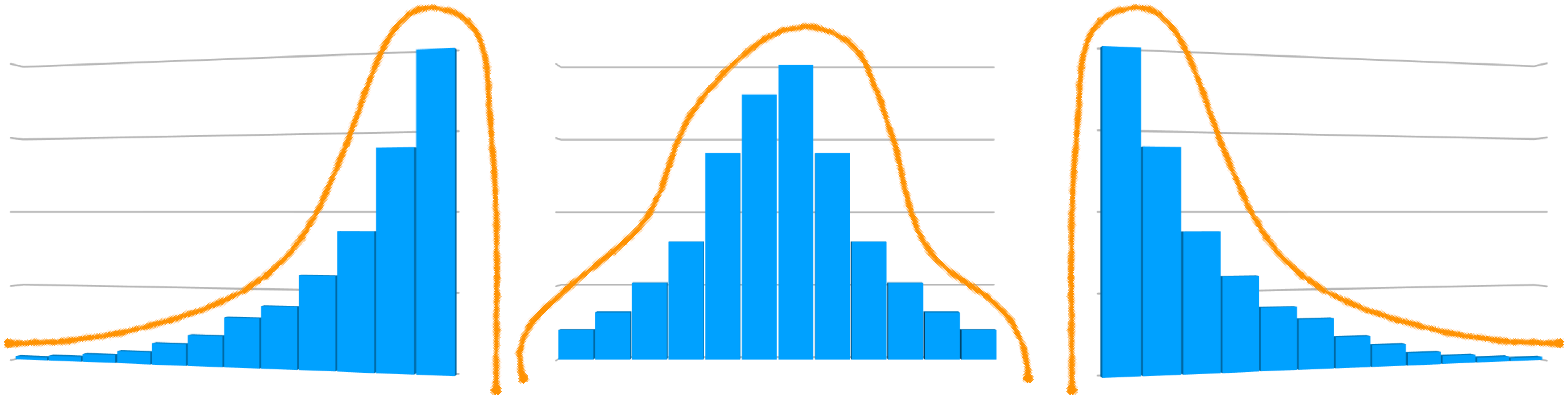


Skewness

Left Skewed

Symmetric

Right Skewed



-ve Skewness

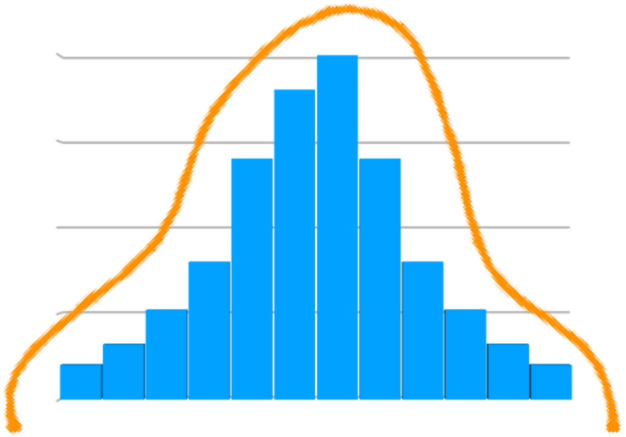
Zero Skewness

+ve Skewness

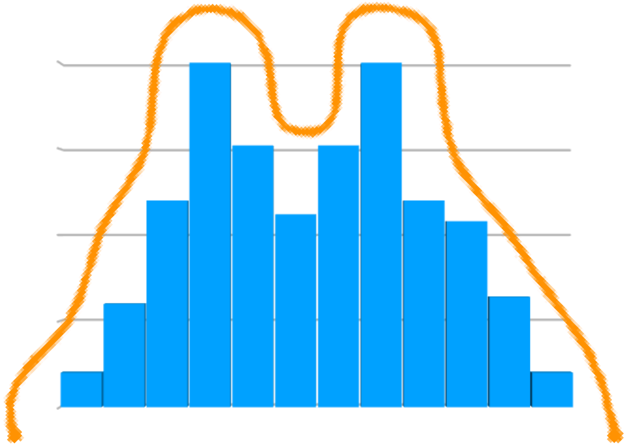
- Draw a smooth curve to see skewness
- Don’t rely on jagged edges

Modality

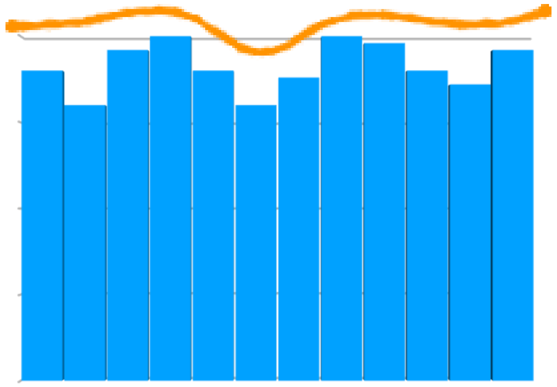
unimodal



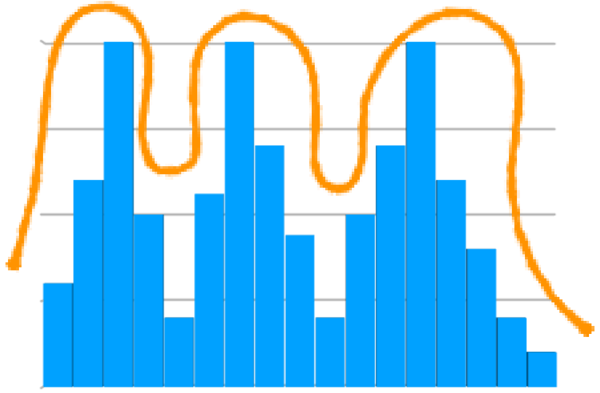
bimodal



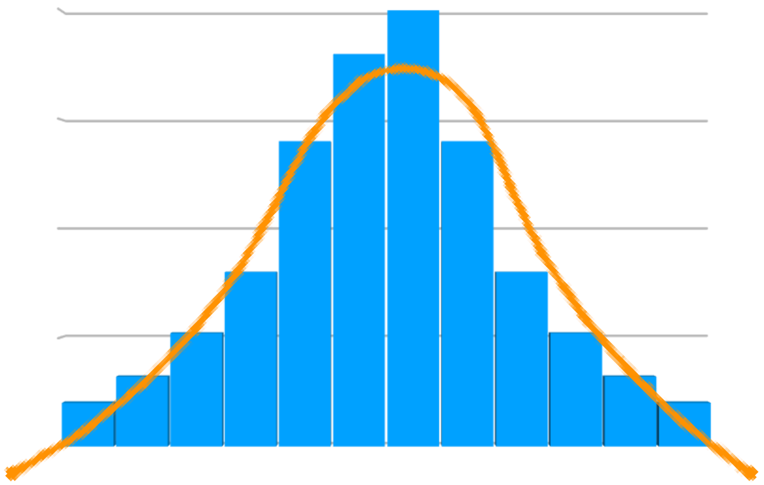
uniform



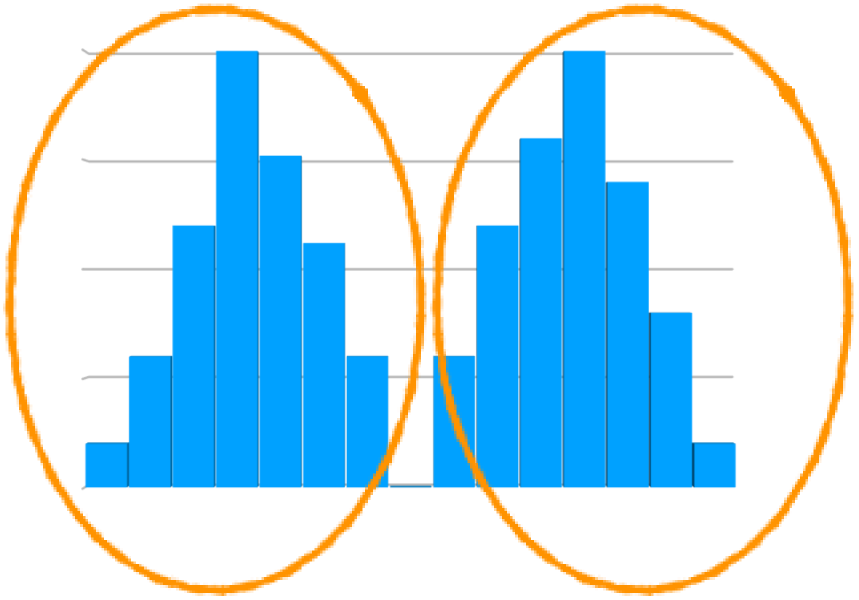
multimodal



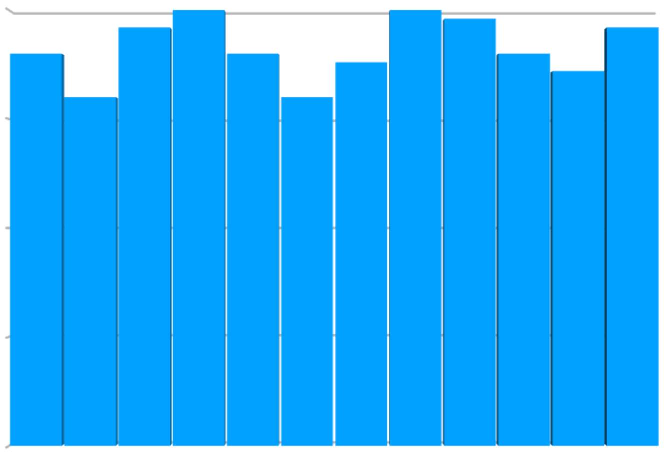
Modality (Example)



Normal Distribution

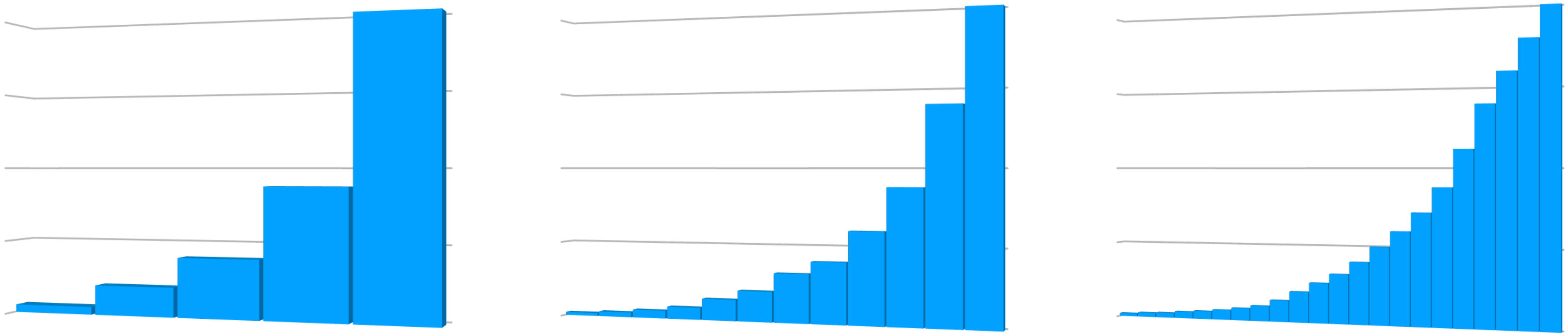


Two separate groups



No trend

Binwidth

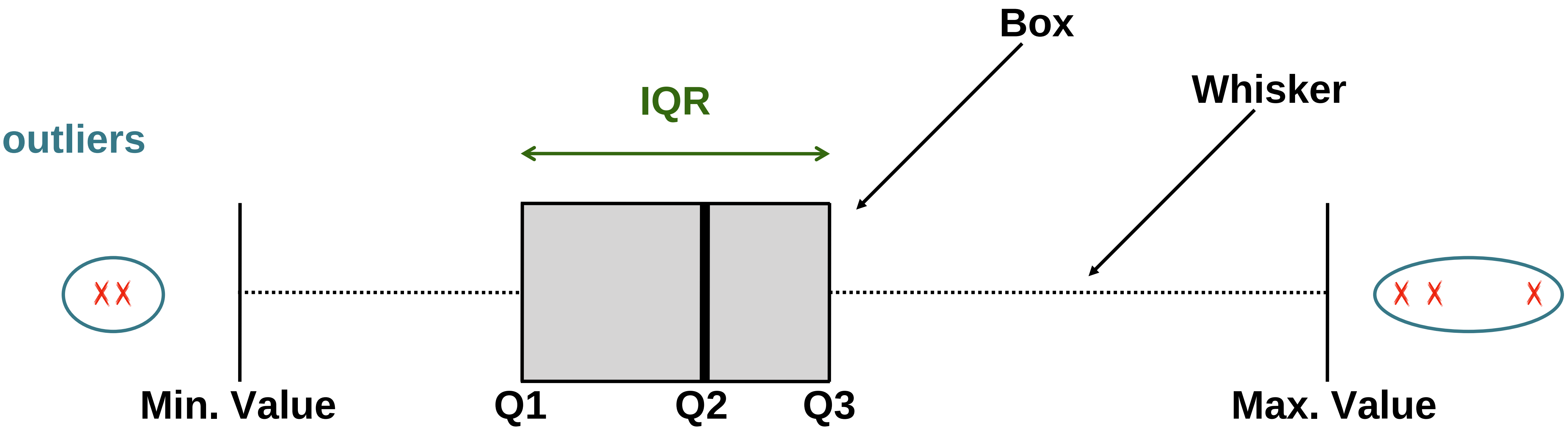


Measures of Center

Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79

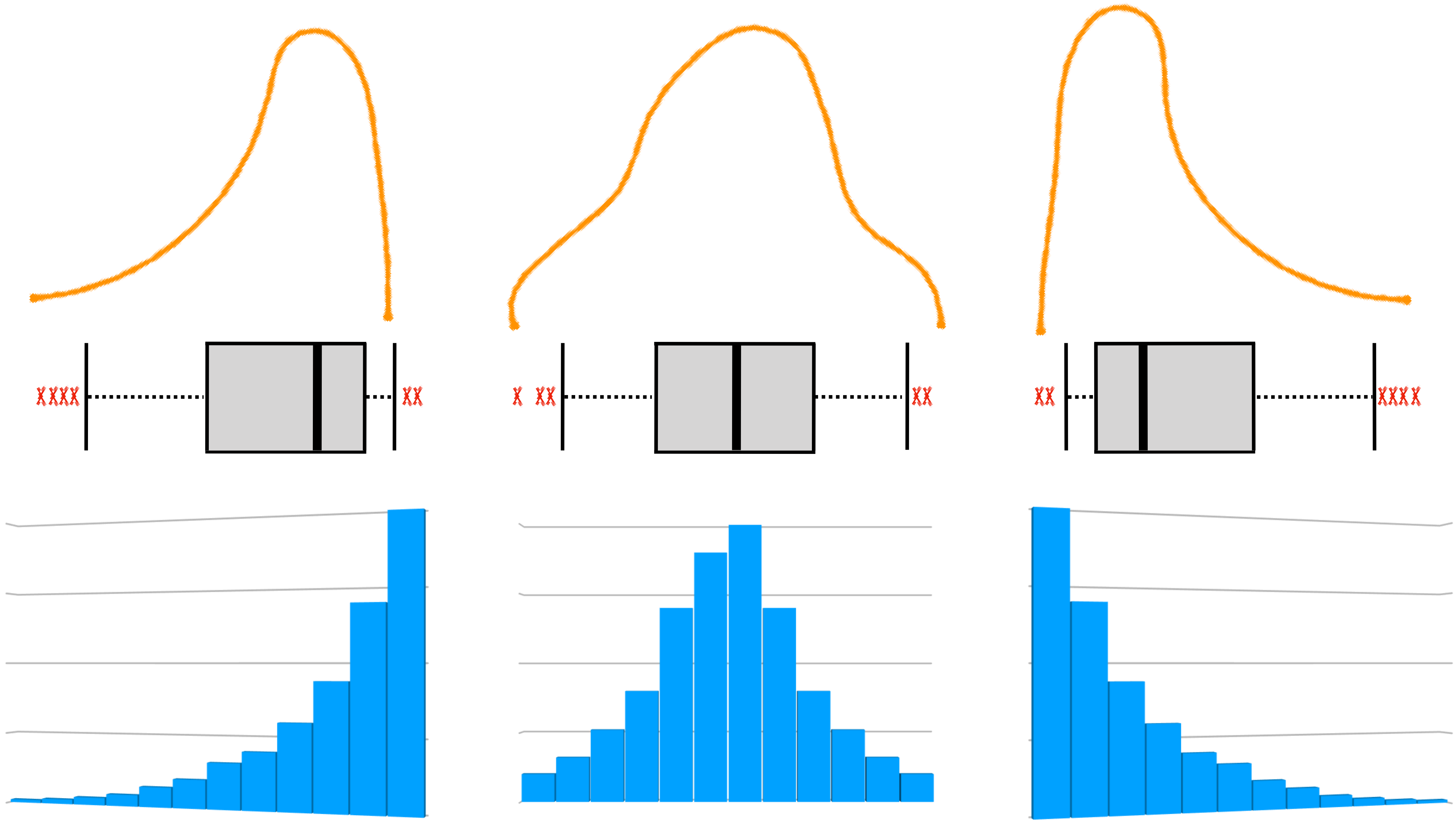
Mean	Arithmetic Average $\text{Mean} = \frac{56 + 87 + 34 + 65 + 77 + 62 + 90 + 45 + 77 + 79}{10}$ Mean = 67.2
Mode	Most frequent value/observation Mode = 77
Median	Midpoint of distribution (50th percentile) $\text{Median} = \frac{77 + 62}{2} = 69.5$

Box Plots



Min. Value :Lower Extreme (that's not an outlier)
Q1 :Lower Quartile (25% of observations)
Q2 :Median (50% of observations)
Q3 :Upper Quartile (75% of observations)
Max. Value :Upper Extreme (that's not an outlier)
IQR :Inter-Quartile Range = Q3 - Q1 (middle 50% of observations)

Box Plots & Skewness

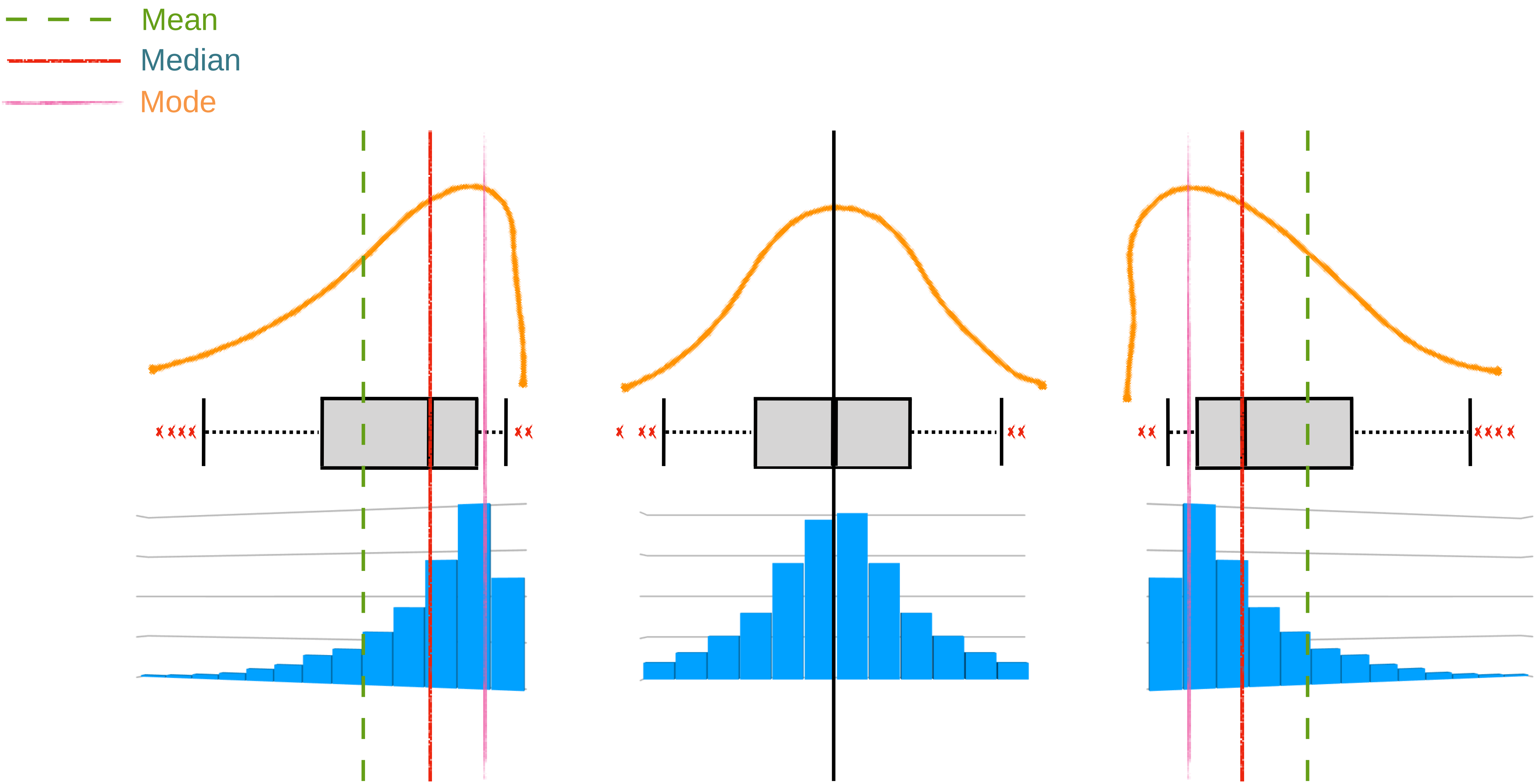


Left Skewed

Symmetric

Right Skewed

Skewness vs Measures of Center



Mean < Median < Mode

Mean = Median = Mode

Mean > Median > Mode

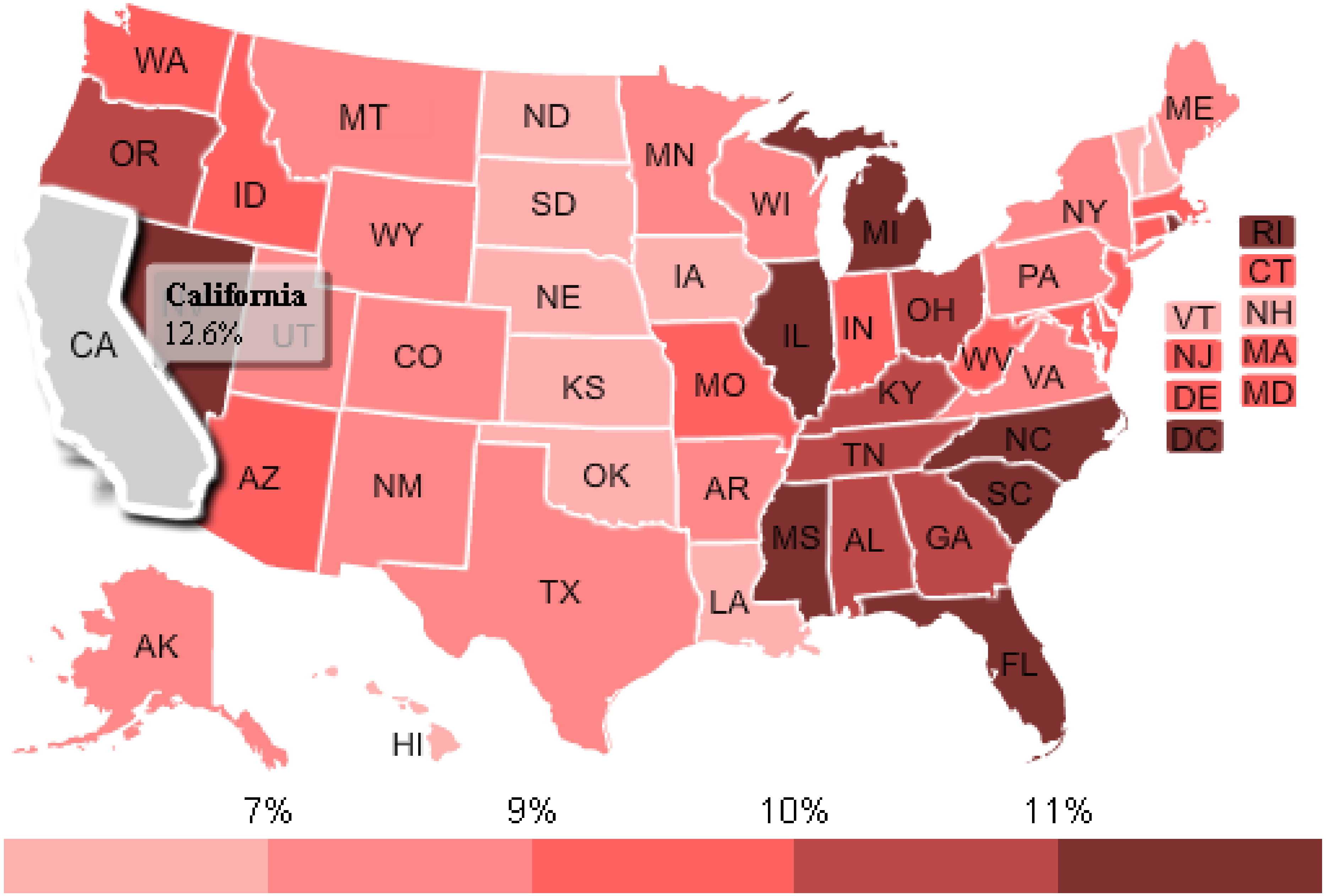
Left Skewed

Symmetric

Right Skewed

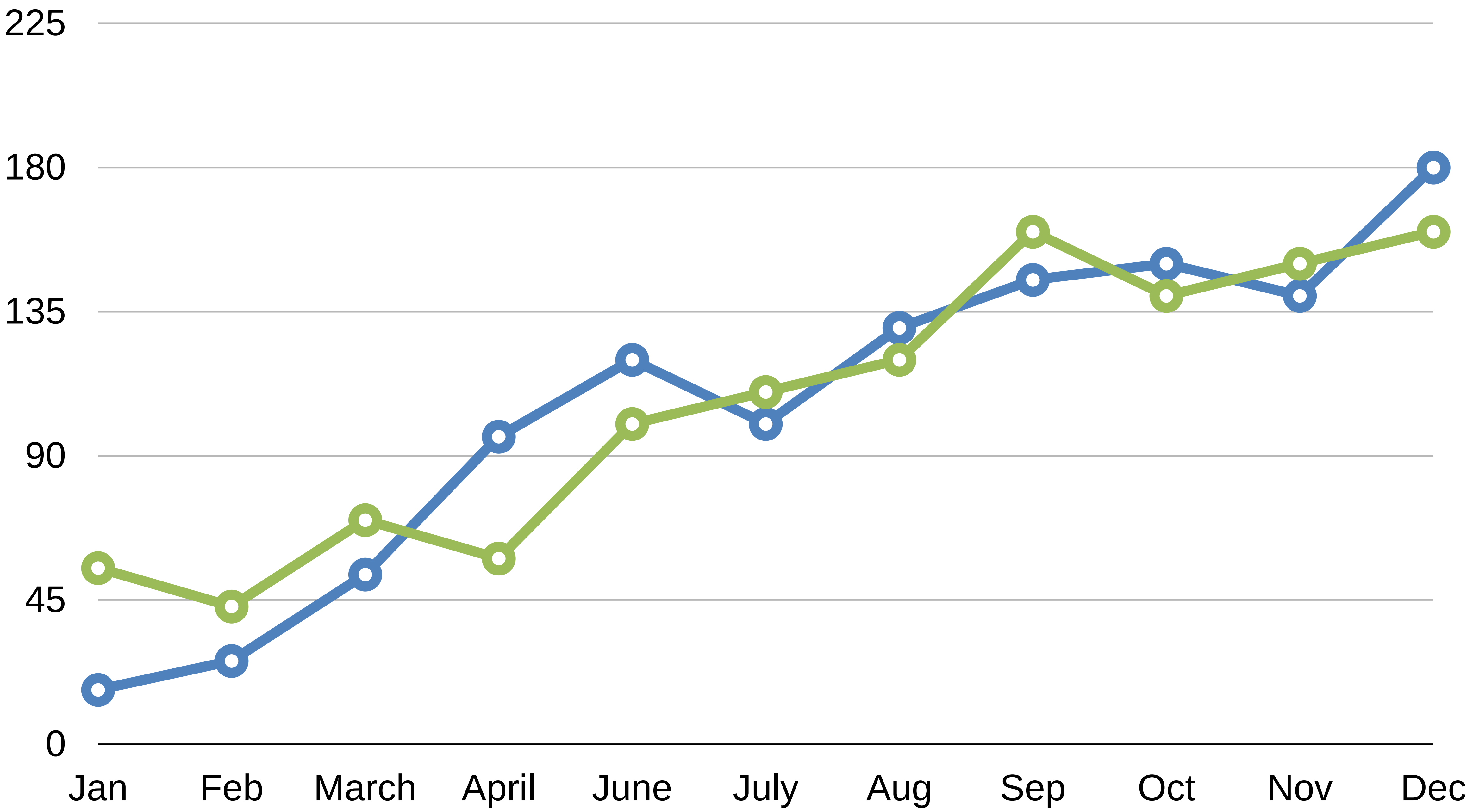
DICE ANALYTICS

Intensity/Heat Maps



DICE ANALYTICS

Time Plots



225
180
135
90
45
0
Jan
Feb
March
April
June
July
Aug
Sep
Oct
Nov
Dec

Measures of Spread

Range

Variance

Standard Deviation

Inter-quartile Range

Range

- Range = Max. Value - Min. Value
- **Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79**
- Range = 90 - 34 = 56

Variance

- A measure of how much data (a variable) varies; how spread out a data set is about the mean.
- Average squared deviation from mean; has squared units of the variable
- Sample Variance

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1}$$

- Population Variance

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Variance (Example)

- **Data : 56, 87, 34, 65, 77, 62, 90, 45, 77, 79**

$$s^2 = \frac{\boxed{\sum (X - \overline{X})^2}}{N-1} = \frac{\boxed{(56 - 67.2)^2 + (87 - 67.2)^2 + \ldots\ldots + (79 - 67.2)^2}}{10-1}$$

$$= \frac{2995.6}{9}$$

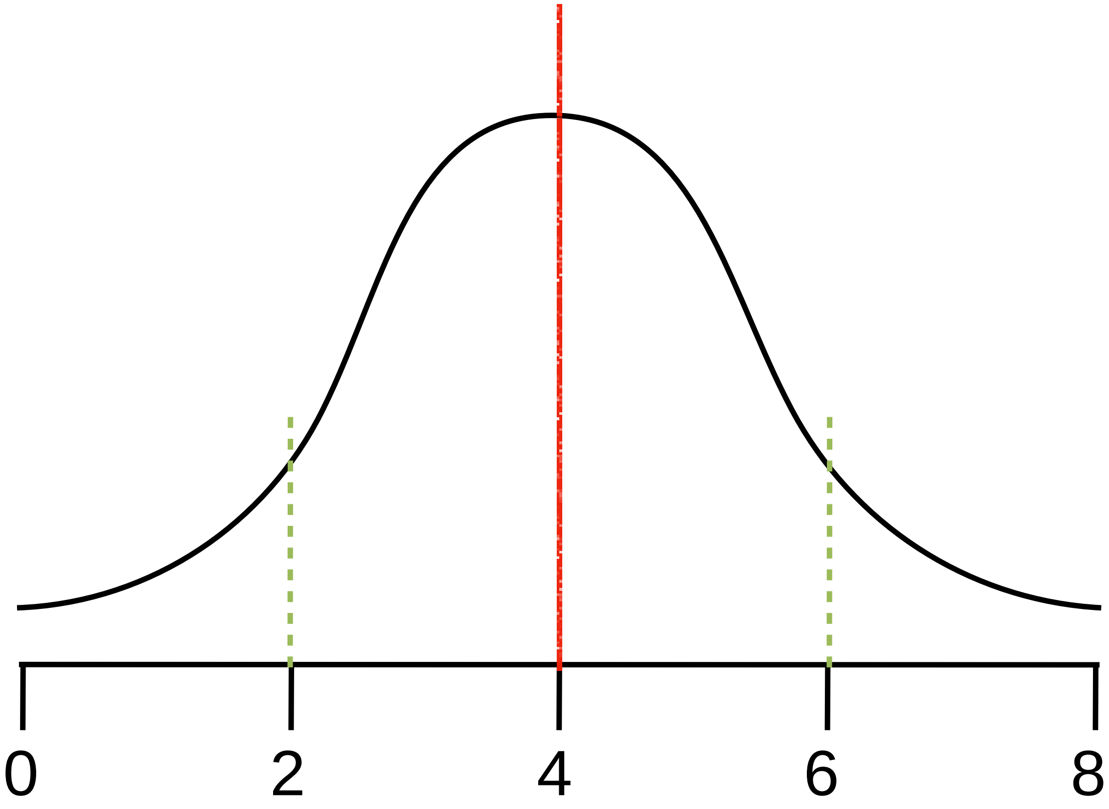
$$= 332.8$$

Sum of Squares

DICE ANALYTICS

Why Square The Differences?

- Get rid of negatives, so that the negatives and positives do not cancel each other during addition.
- Increase larger deviations more than smaller ones so that they are weighed more heavily.

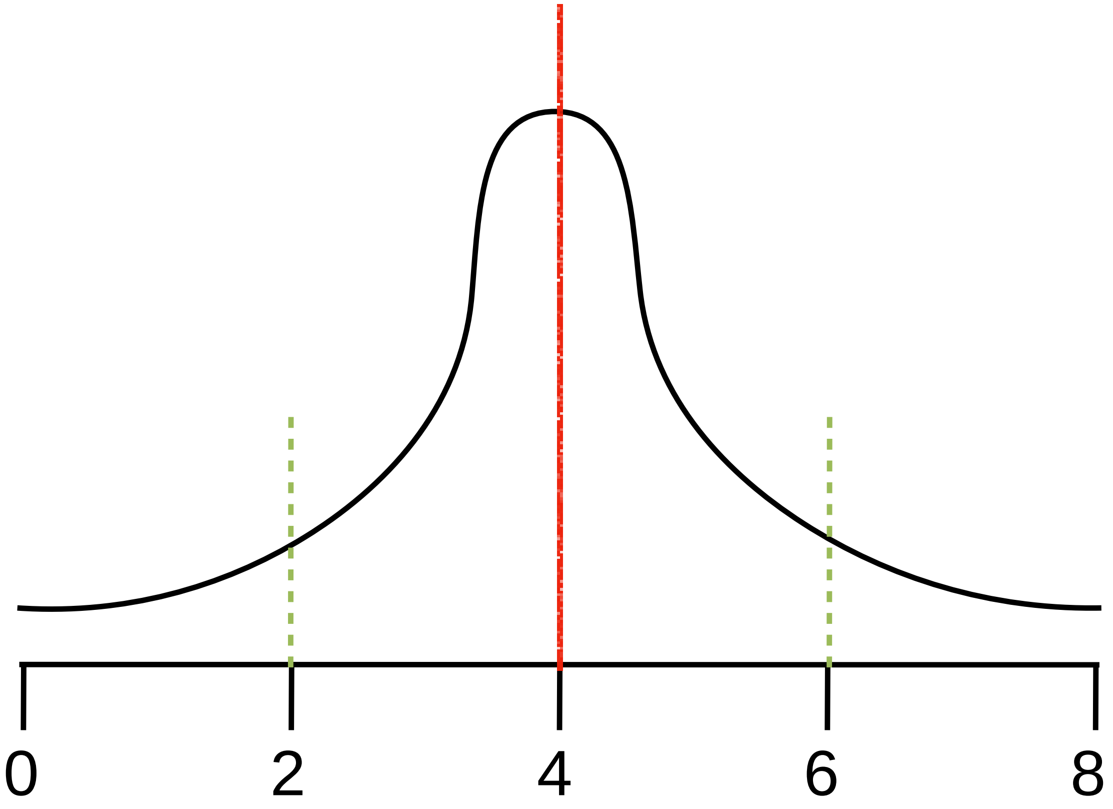


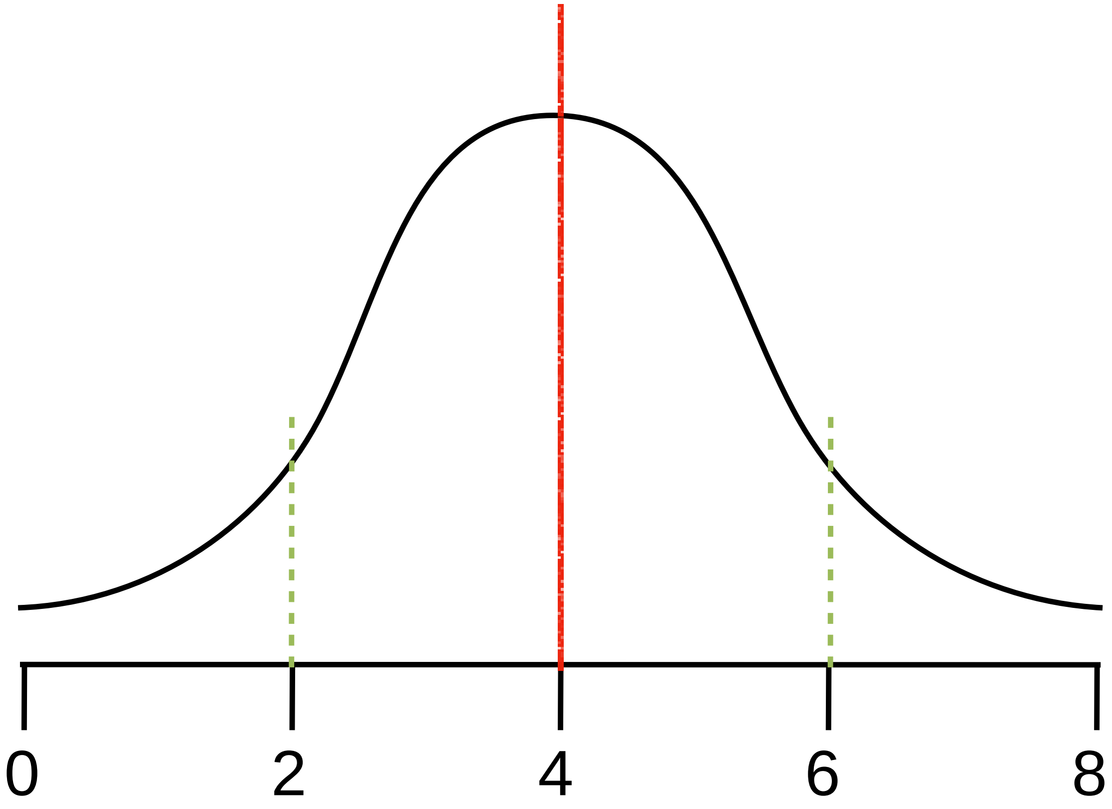
(2-4) + (6-4) = -2 + 2 = 0

Standard Deviation (SD)

- Square root of Variance
- It has the same units as the variable, which makes it useful in comparisons and calculations
- Sample SD $s = \sqrt{s^2} = \sqrt{\frac{\sum (X - \bar{X})^2}{N - 1}}$
- Population SD $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

Spread





Less Spread

Low Variance

Low Deviation

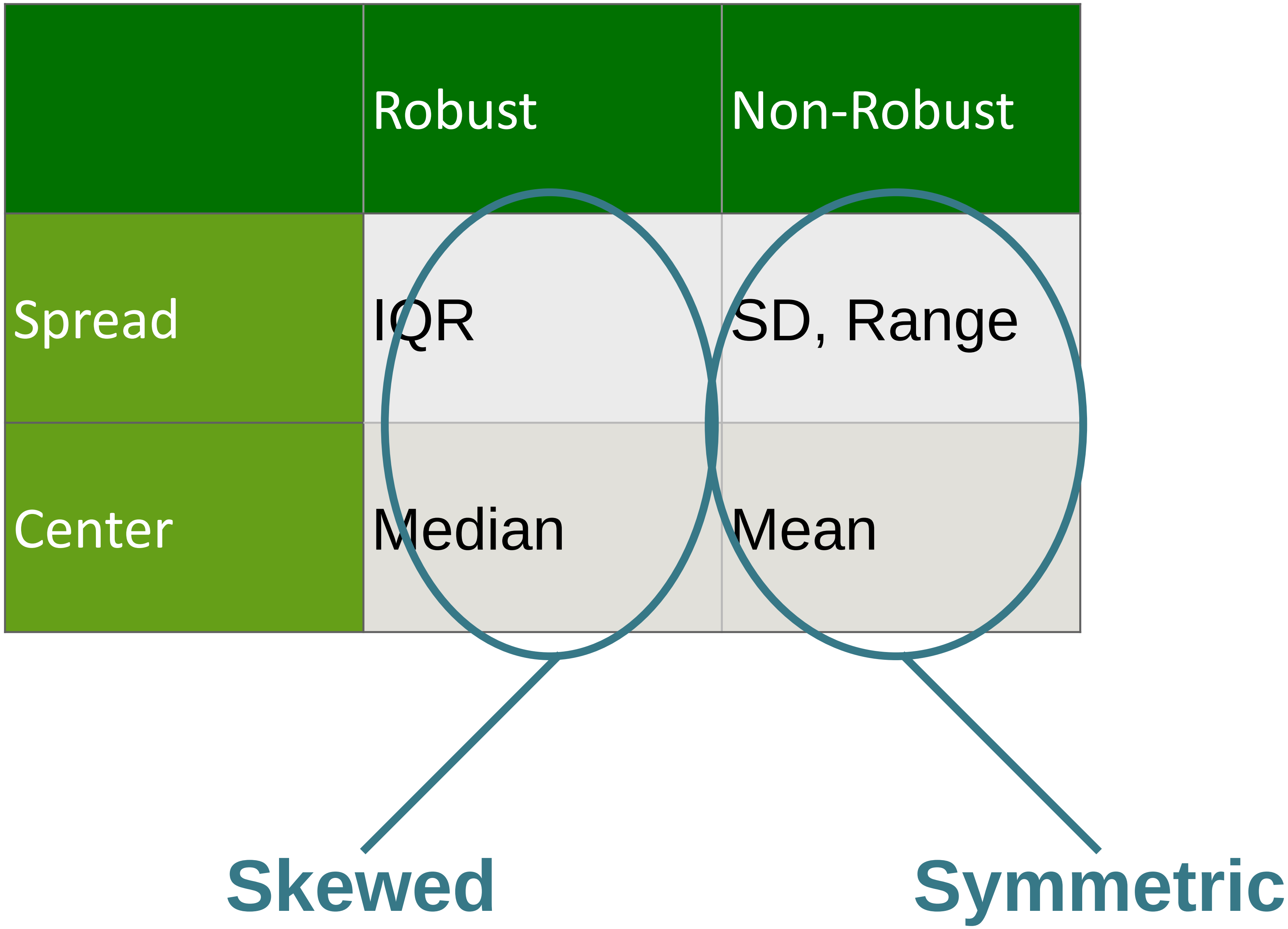
More Spread

High Variance

High Deviation

Robust Statistics

- Measures on which extreme observations or outliers have little effect



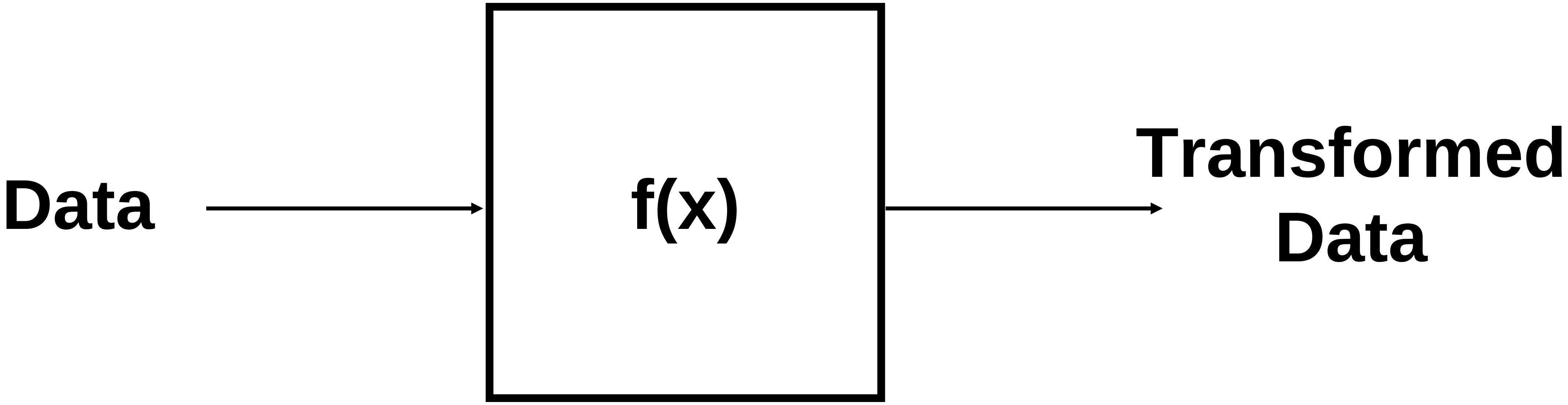
	Robust	Non-Robust
Spread	IQR	SD, Range
Center	Median	Mean

Skewed

Symmetric

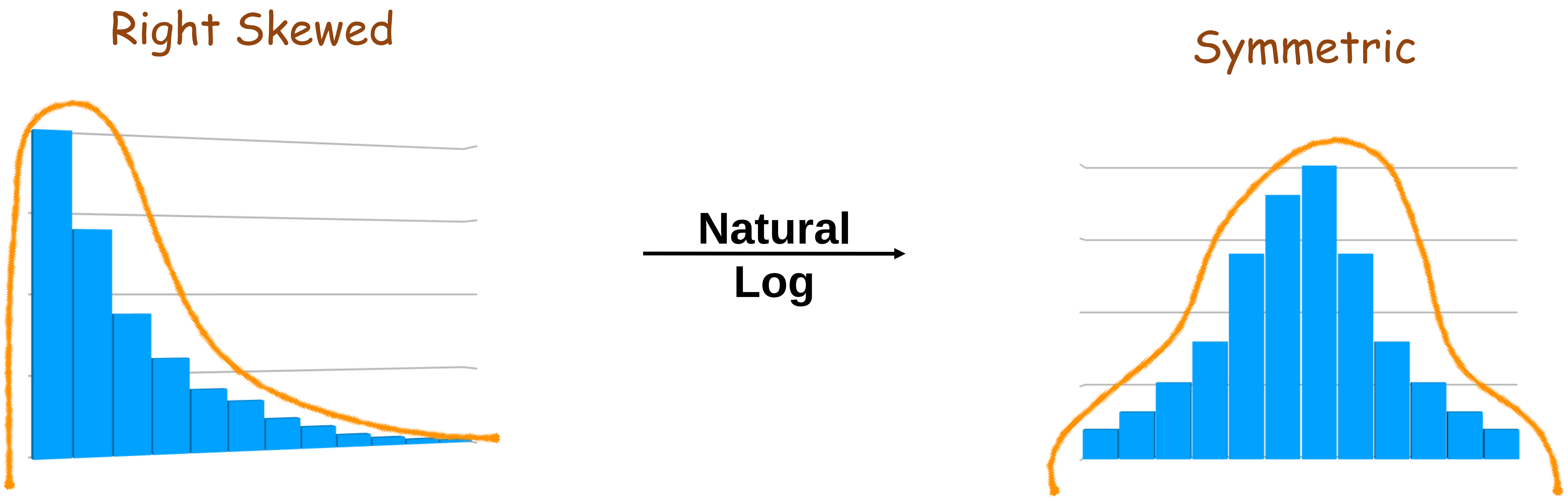
Data Transformations

- Applying a Function f(x) to adjust scales of data.
- Done usually when data is skewed, so that it becomes easier to perform *modelling.*
- Done to convert non-linear relationship into a linear relationship.



(Natural) Log Transformation

- To transform data that is positively skewed
- Usually done when data is concentrated near Zero (relative to the few large values in data)



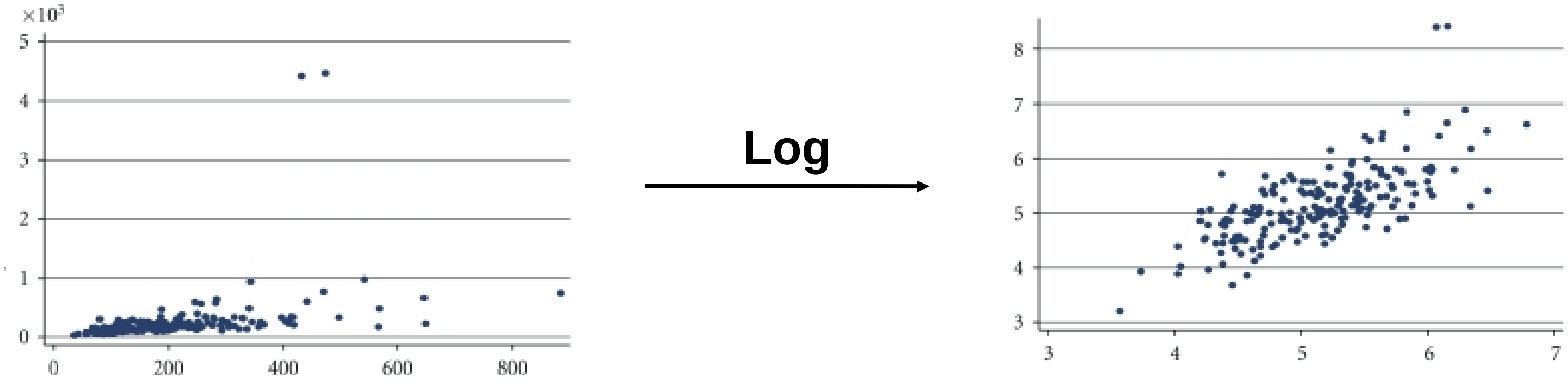
Right Skewed

Natural Log

Symmetric

Log Transformation

- To make the relationship between two variable more linear
- Most of the simple methods for modelling work only when relationship is linear

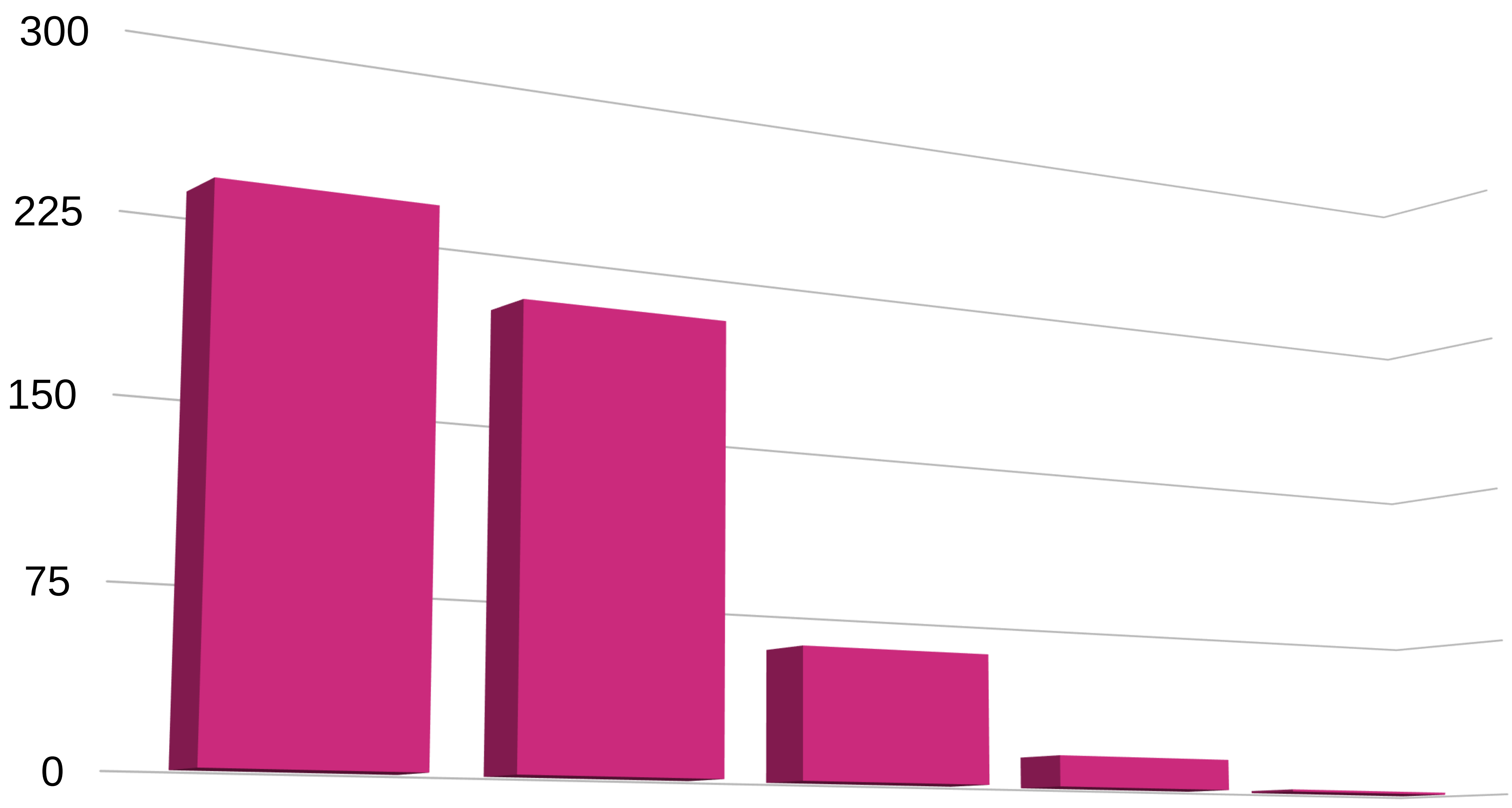


Other Transformation

- You may use other transformations or create of your own
- For instance: Square Root, Square, Inverse

Visualising Categorical Data

Bar Plot

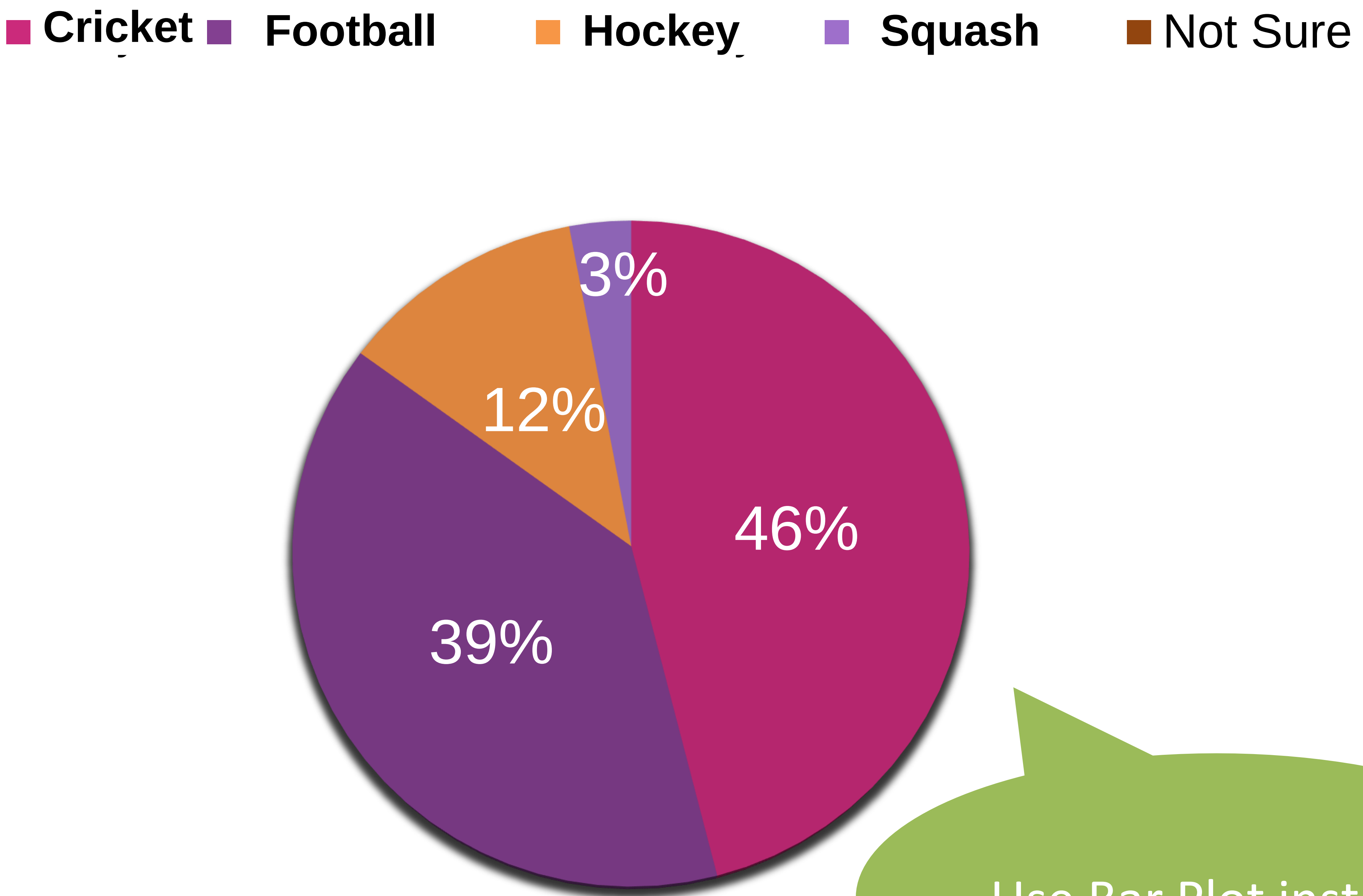


Frequency

Bar Plot vs Histogram

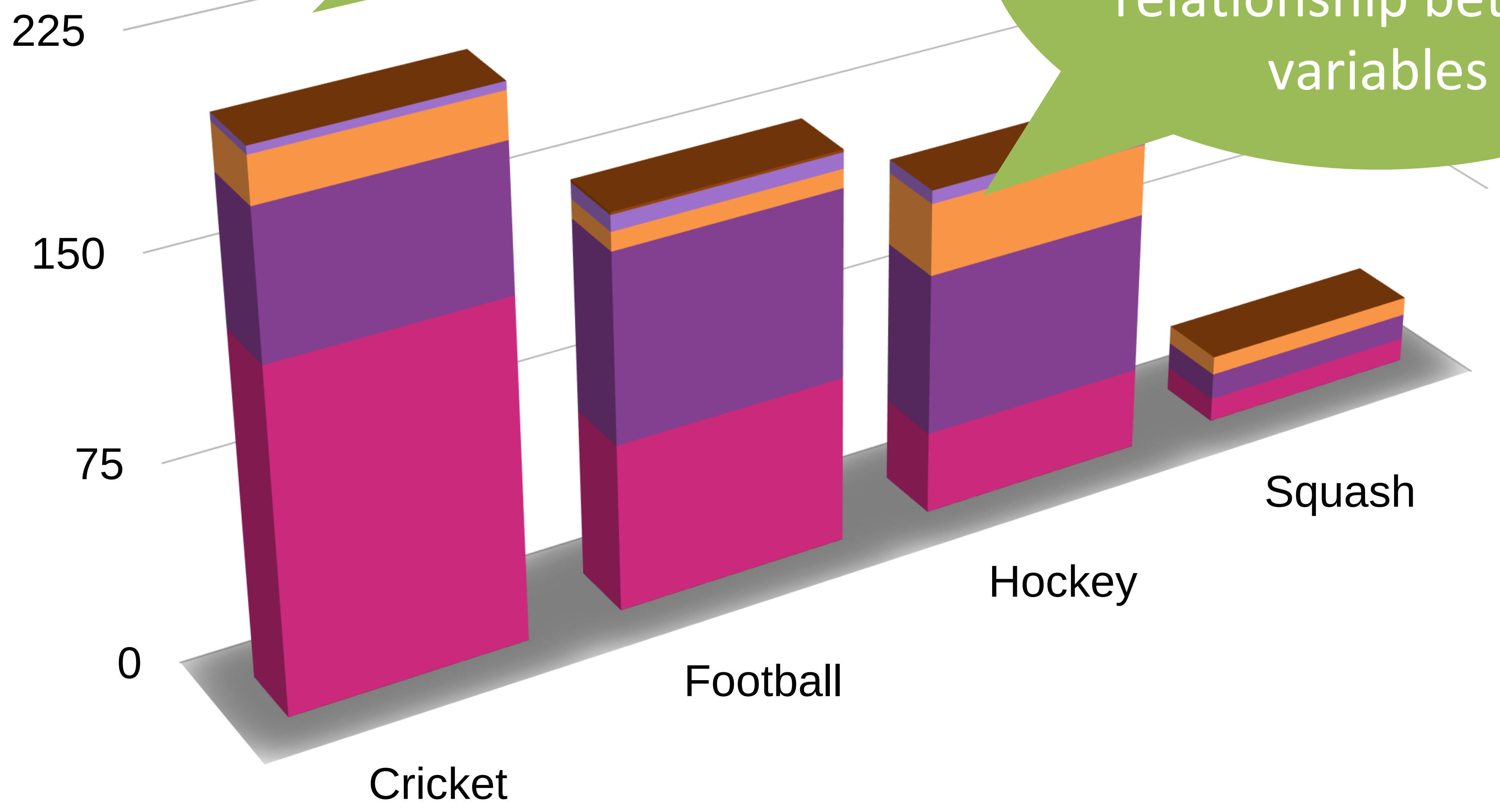
- Bar Plot for Categorical Variables, Histogram for Numerical Variables
- X-axis in Histogram must be a Number Line
- Ordering of bars is not interchangeable in Histogram as compared to Bar Plot

Pie Chart



Use Bar Plot instead

Segmented Bar Plot

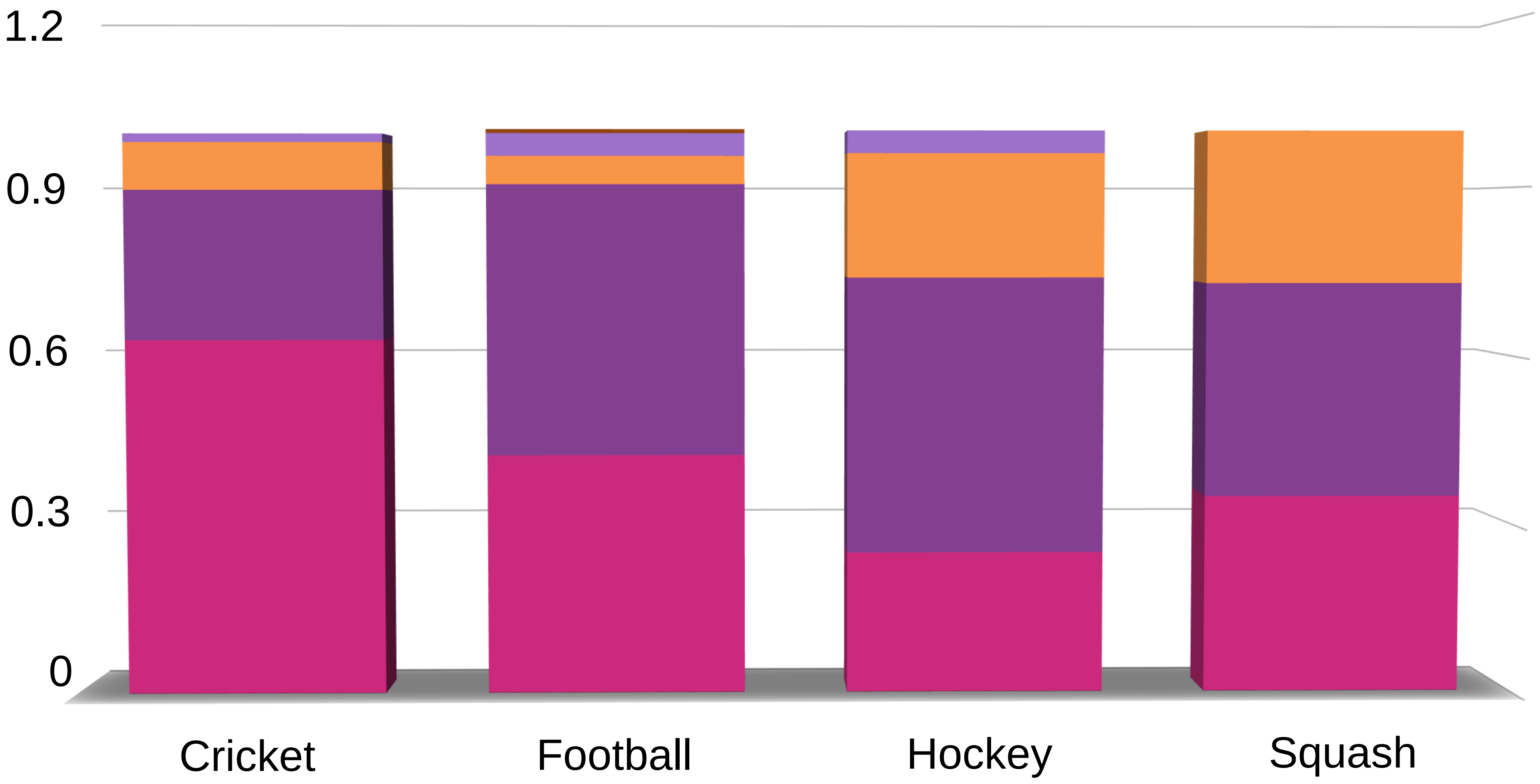


For visualising conditional frequency distributions

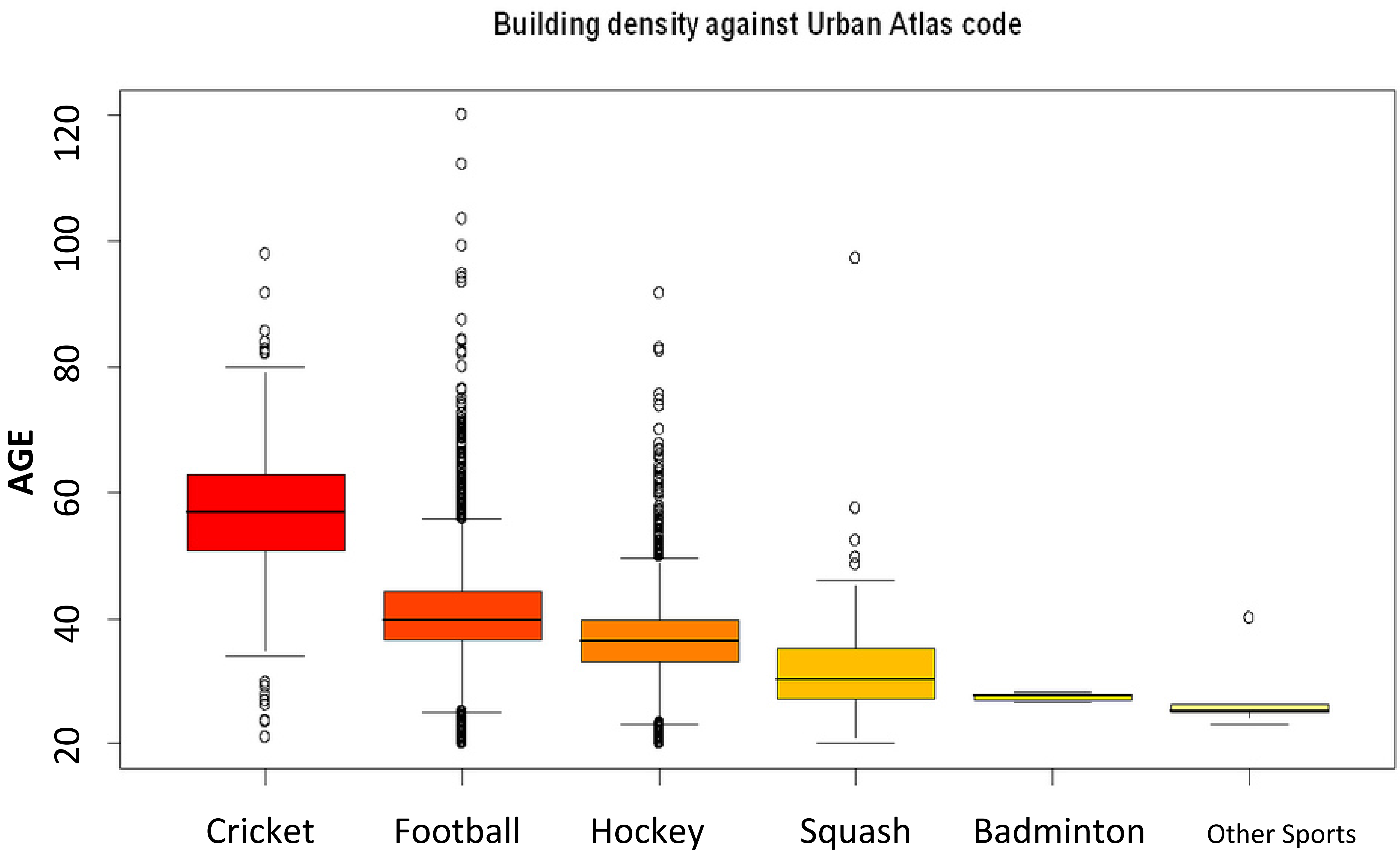
For comparing relative frequencies to explore relationship between variables

Sensitivity: Internal

Relative Frequency Segmented Bar Plot

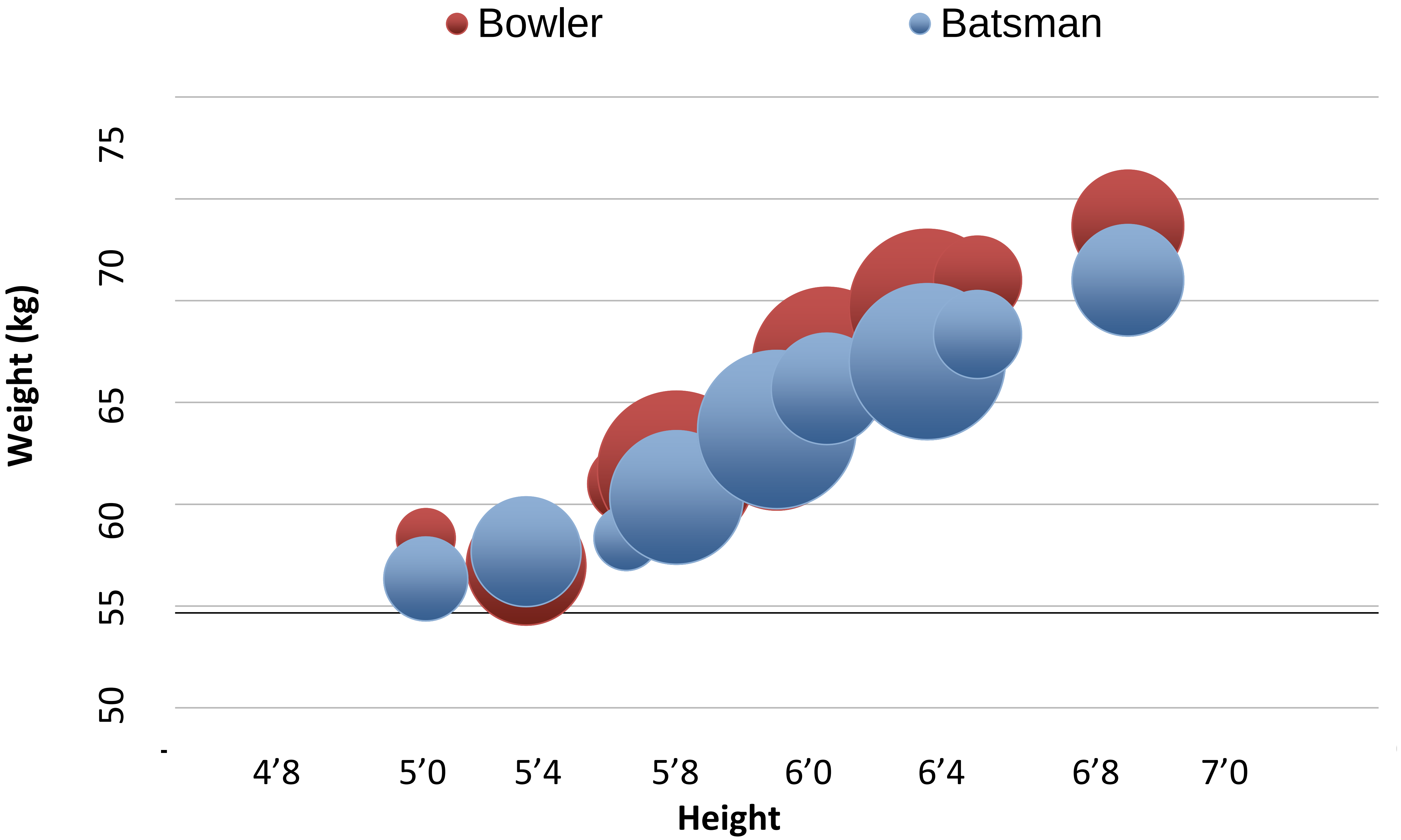


Side-by-Side Box Plots



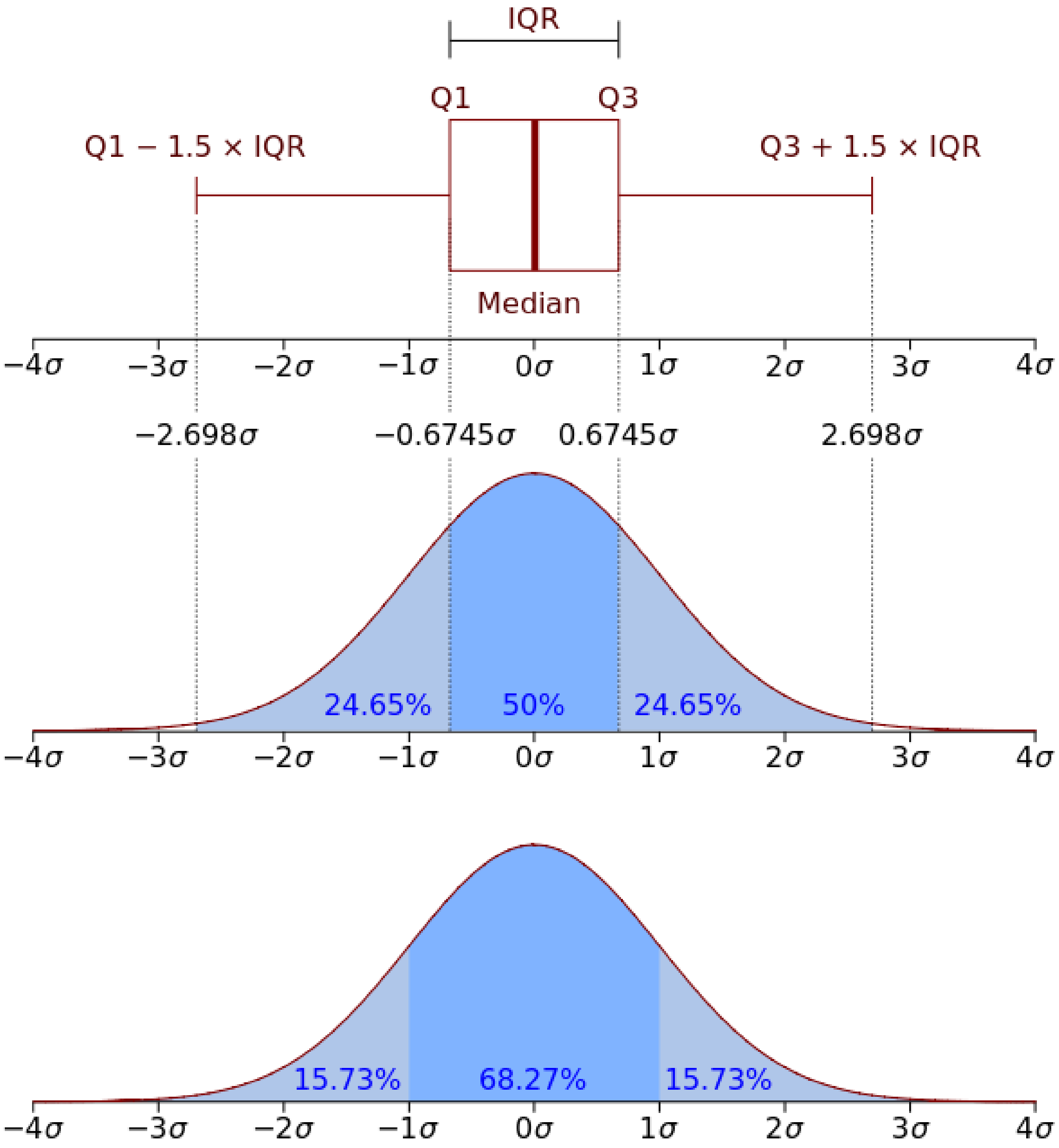
Sensitivity: Internal

Bubble Plot



Sensitivity: Internal

Outliers


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

Why do EDA

- To understand data properties
- To find patterns in data
- To suggest modelling strategies
- To "debug" analyses
- To communicate results

(From JHU)

Why do EDA

https://www.youtube.com/watch?v=jbkSRLYSojo