DICE
ANALYTICS

Data Science and Machine Learning

https://www.facebook.com/diceanalytics/

https://pk.linkedin.com/company/diceanalytics

DATA ORGANIZATION

Data is stored in the form of a ***Data Matrix***

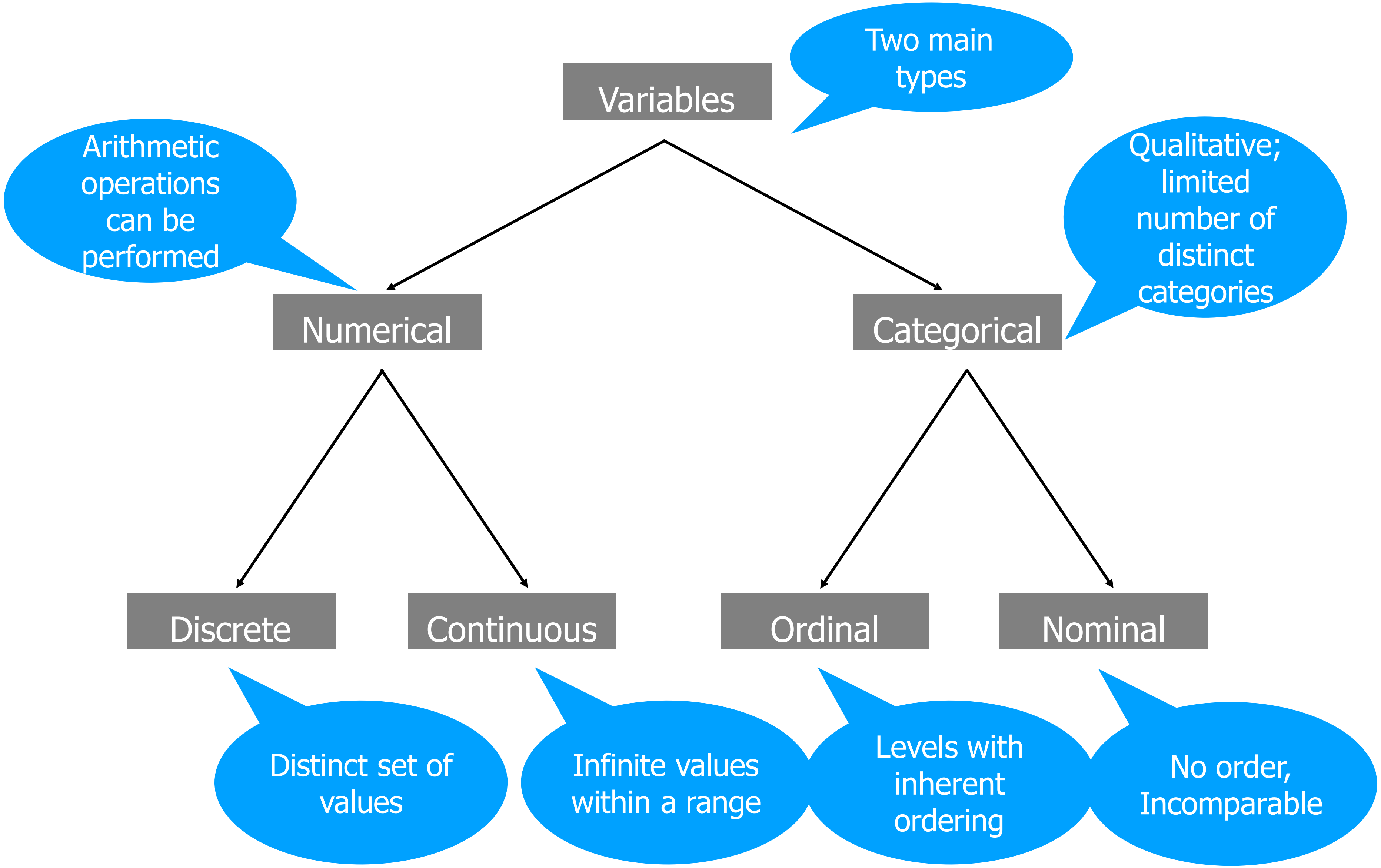
OrderDate	Region	Rep	Item	Units	Cost	Total
1/6/10	East	Jones	Pencil	95	1.99	189.05
1/23/10	Central	Kivell	Binder	50	19.99	999.50
2/9/10	Central	Jardine	Pencil	36	4.99	179.64
2/26/10	Central	Gill	Pen	27	19.99	539.73
3/15/10	West	Sorvino	Pencil	56	2.99	167.44
4/1/10	East	Jones	Binder	60	4.99	299.40
4/18/10	Central	Andrews	Pencil	75	1.99	149.25
5/5/10	Central	Jardine	Pencil	90	4.99	449.10
5/22/10	West	Thompson	Pencil	32	1.99	63.68

Variable Names

Observation (Row)

Variable (Column)

TYPES OF VARIABLES



TYPES OF VARIABLES

https://www.statisticshowto.com/probability-and-statistics/types-of-variables/

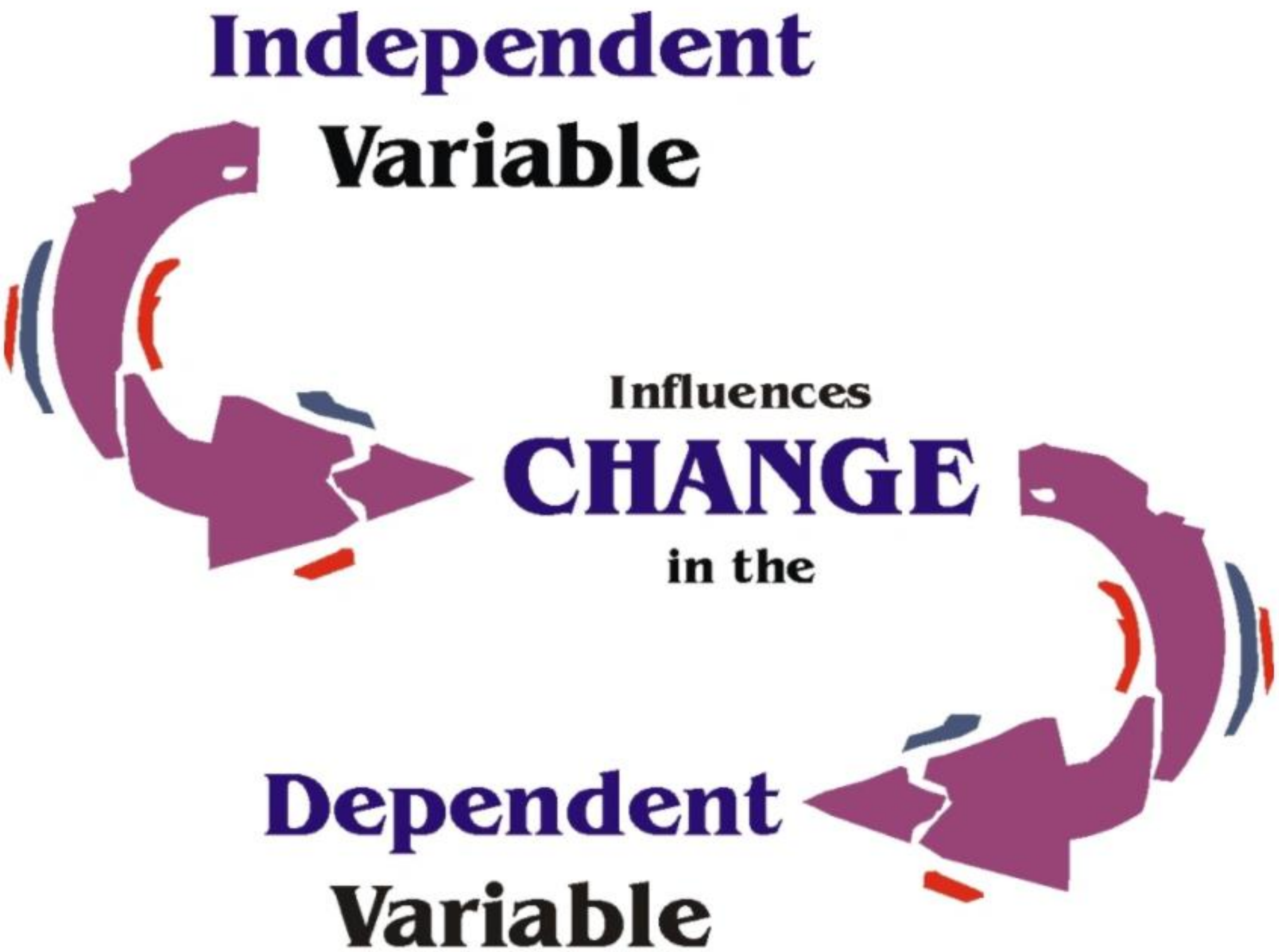
Types of Variables in Statistics and Research

TYPES OF VARIABLES

- *Response Variable*: It is the focus of a question in a study or experiment. It is the variable we want to predict or observe. It is the dependent variable.

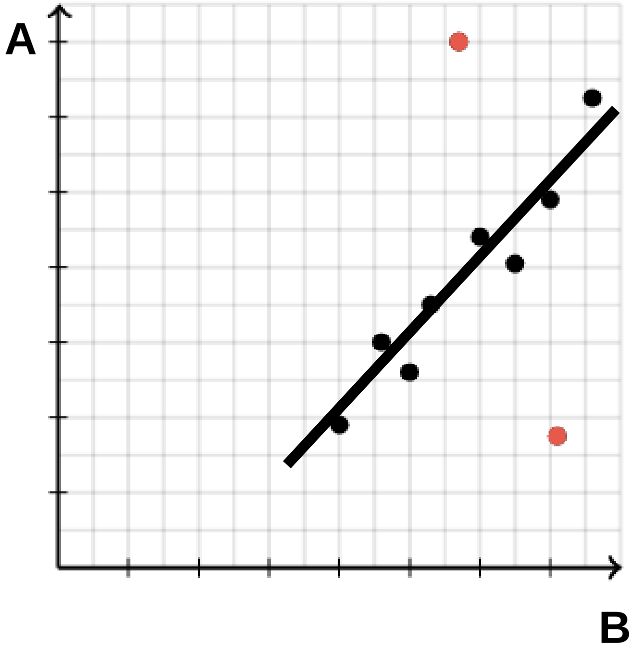
- *Explanatory Variable*: It is the variable on whom the response variable depends, or the variable which ‘explains’ the response variable. It is assumed to be independent variable.

http://www.statisticshowto.com/types-variables/

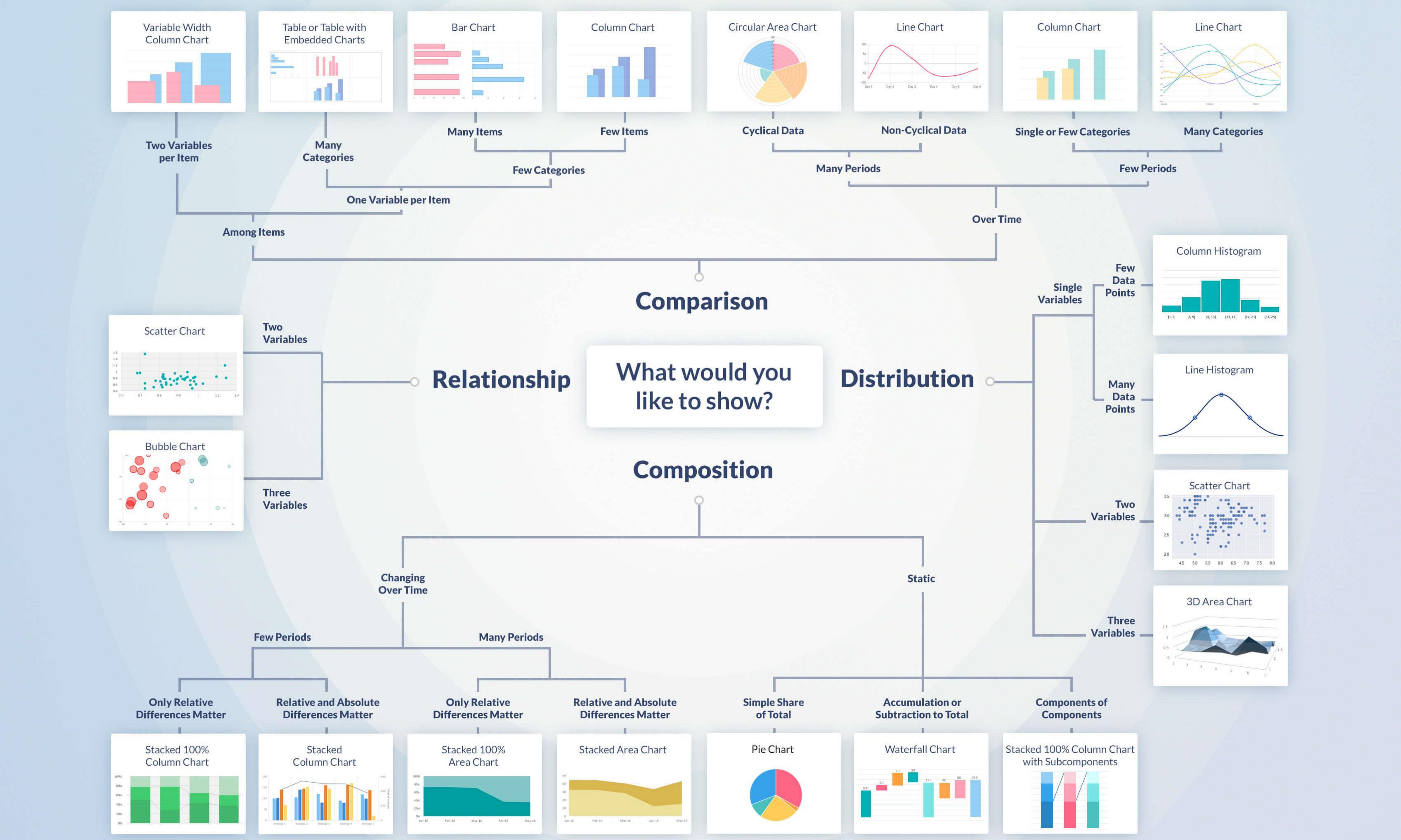

Independent
Variable
Influences
CHANGE
in the
Dependent
Variable

RELATIONSHIP B/W VARIABLES

- Two variables that show connection with each other are called ***Associated/Correlated (Dependent)***
- Two variables that do not show connection with each other are called *Independent*
- An observation that is away that is not close to majority of data is called *Outlier*



DATA VISUALIZATION


What would you like to show?
Comparison
Among Items
Two Variables per Item
Variable Width Column Chart
One Variable per Item
Many Categories
Table or Table with Embedded Charts
Few Categories
Many Items
Bar Chart
Few Items
Column Chart
Over Time
Many Periods
Cyclical Data
Circular Area Chart
Non-Cyclical Data
Line Chart
Few Periods
Single or Few Categories
Column Chart
Many Categories
Line Chart
Relationship
Two Variables
Scatter Chart
Three Variables
Bubble Chart
Distribution
Single Variables
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Two Variables
Scatter Chart
Three Variables
3D Area Chart
Composition
Changing Over Time
Few Periods
Only Relative Differences Matter
Stacked 100% Column Chart
Relative and Absolute Differences Matter
Stacked Column Chart
Many Periods
Only Relative Differences Matter
Stacked 100% Area Chart
Relative and Absolute Differences Matter
Stacked Area Chart
Static
Simple Share of Total
Pie Chart
Accumulation or Subtraction to Total
Waterfall Chart
Components of Components
Stacked 100% Column Chart with Subcomponents

And
Human Brain
Works
Different

DICE ANALYTICS

HOW MANY 9 ?

3	3	0	3	0	1	8	7	6	8	2	1	4	0	3	8	3	7	7	2	0	5	2	3	2	7	0	2	0
7	1	4	6	0	2	1	3	2	7	6	0	2	5	6	3	2	5	7	6	3	3	0	2	0	3	0	7	2
8	7	5	7	2	8	3	8	7	7	8	2	0	7	7	5	2	3	1	1	5	6	3	8	4	7	8	2	0
0	5	0	5	1	6	1	7	5	6	8	0	4	4	6	7	4	7	1	4	0	0	8	4	4	3	0	3	2
2	4	3	1	3	5	4	9	5	0	7	6	0	7	4	3	1	8	2	7	3	4	6	0	2	4	8	2	3
8	6	2	2	6	5	4	6	7	0	7	6	0	0	3	9	0	2	4	7	1	7	2	3	3	5	8	7	0
0	8	4	5	1	3	1	7	6	4	5	4	1	2	4	5	3	3	5	4	9	6	7	7	6	3	4	2	5
4	7	7	0	2	2	0	1	1	7	7	7	0	2	6	6	4	7	5	8	6	1	4	3	7	8	5	4	6
4	3	6	6	4	6	6	2	8	4	8	5	3	7	8	8	1	3	8	5	4	5	7	4	0	3	2	8	4
5	5	0	3	5	3	5	3	8	3	2	3	8	2	3	1	6	2	7	2	4	6	3	6	4	4	3	2	5
4	4	0	2	1	7	2	4	4	7	4	1	9	2	4	5	2	5	0	4	0	0	5	3	6	3	3	6	7
7	4	6	6	8	7	5	7	9	2	0	2	8	8	8	8	3	2	4	2	6	4	0	4	6	3	7	2	1
0	1	7	1	5	9	1	4	2	8	7	3	7	1	4	5	1	8	7	8	0	5	1	7	0	5	8	8	1
2	8	5	2	1	2	8	7	7	6	2	5	6	2	6	4	1	5	1	6	1	2	1	1	0	5	6	4	0
2	1	1	7	7	2	0	0	1	8	7	0	2	9	0	2	8	5	7	8	4	6	0	6	5	0	7	1	2
0	5	2	4	1	5	3	3	1	5	5	1	4	0	1	6	4	3	3	9	8	8	3	4	6	8	4	8	6
7	3	7	5	2	4	0	2	7	6	3	8	5	5	4	5	8	8	7	5	5	6	5	6	7	9	7	7	4
0	3	2	8	1	4	4	6	0	8	2	3	0	1	3	4	6	2	0	5	7	7	3	6	1	8	7	3	5
4	4	8	3	3	3	5	0	1	0	3	8	6	3	2	0	5	0	6	1	3	3	4	3	6	1	5	8	6
1	0	2	2	7	6	3	3	0	8	8	0	3	1	8	8	1	2	1	7	5	2	9	3	5	8	3	2	5

HOW MANY 9 ?

3	3	0	3	0	1	8	7	6	8	2	1	4	0	3	8	3	7	7	2	0	5	2	3	2	7	0	2	0
7	1	4	6	0	2	1	3	2	7	6	0	2	5	6	3	2	5	7	6	3	3	0	2	0	3	0	7	2
8	7	5	7	2	8	3	8	7	7	8	2	0	7	7	5	2	3	1	1	5	6	3	8	4	7	8	2	0
0	5	0	5	1	6	1	7	5	6	8	0	4	4	6	7	4	7	1	4	0	0	8	4	4	3	0	3	2
2	4	3	1	3	5	4	9	5	0	7	6	0	7	4	3	1	8	2	7	3	4	6	0	2	4	8	2	3
8	6	2	2	6	5	4	6	7	0	7	6	0	0	3	9	0	2	4	7	1	7	2	3	3	5	8	7	0
0	8	4	5	1	3	1	7	6	4	5	4	1	2	4	5	3	3	5	4	9	6	7	7	6	3	4	2	5
4	7	7	0	2	2	0	1	1	7	7	7	0	2	6	6	4	7	5	8	6	1	4	3	7	8	5	4	6
4	3	6	6	4	6	6	2	8	4	8	5	3	7	8	8	1	3	8	5	4	5	7	4	0	3	2	8	4
5	5	0	3	5	3	5	3	8	3	2	3	8	2	3	1	6	2	7	2	4	6	3	6	4	4	3	2	5
4	4	0	2	1	7	2	4	4	7	4	1	9	2	4	5	2	5	0	4	0	0	5	3	6	3	3	6	7
7	4	6	6	8	7	5	7	9	2	0	2	8	8	8	8	3	2	4	2	6	4	0	4	6	3	7	2	1
0	1	7	1	5	9	1	4	2	8	7	3	7	1	4	5	1	8	7	8	0	5	1	7	0	5	8	8	1
2	8	5	2	1	2	8	7	7	6	2	5	6	2	6	4	1	5	1	6	1	2	1	1	0	5	6	4	0
2	1	1	7	7	2	0	0	1	8	7	0	2	9	0	2	8	5	7	8	4	6	0	6	5	0	7	1	2
0	5	2	4	1	5	3	3	1	5	5	1	4	0	1	6	4	3	3	9	8	8	3	4	6	8	4	8	6
7	3	7	5	2	4	0	2	7	6	3	8	5	5	4	5	8	8	7	5	5	6	5	6	7	9	7	7	4
0	3	2	8	1	4	4	6	0	8	2	3	0	1	3	4	6	2	0	5	7	7	3	6	1	8	7	3	5
4	4	8	3	3	3	5	0	1	0	3	8	6	3	2	0	5	0	6	1	3	3	4	3	6	1	5	8	6
1	0	2	2	7	6	3	3	0	8	8	0	3	1	8	8	1	2	1	7	5	2	9	3	5	8	3	2	5

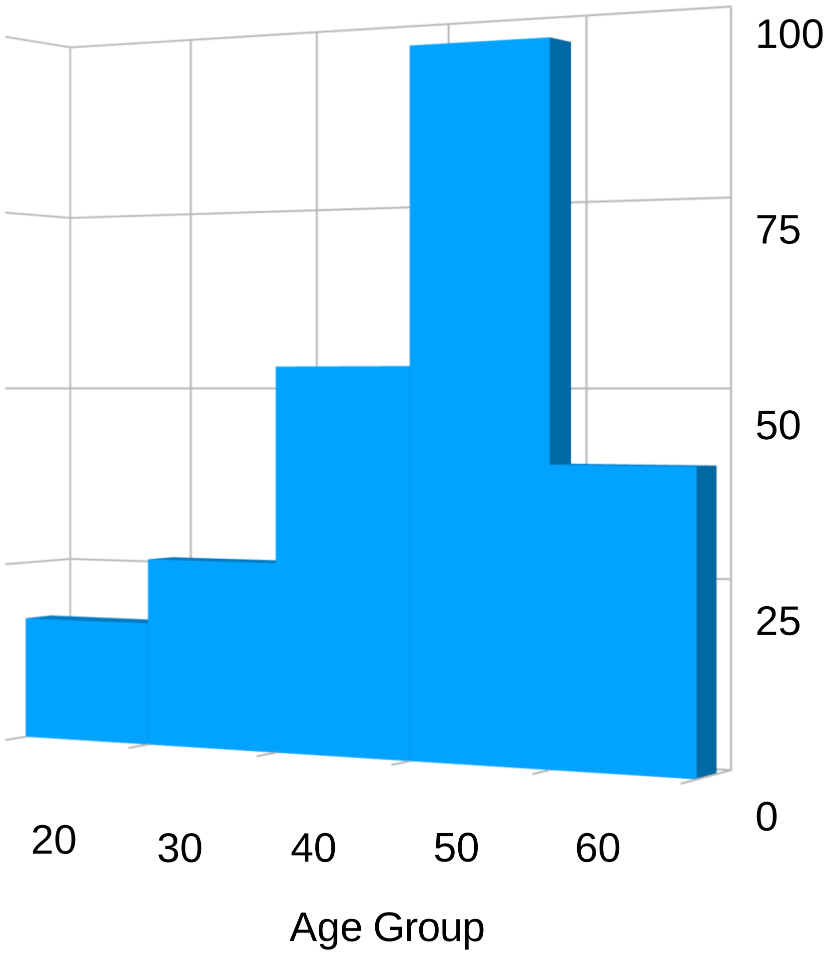
The Human Visual System is Powerful

Visualizing Numerical Data

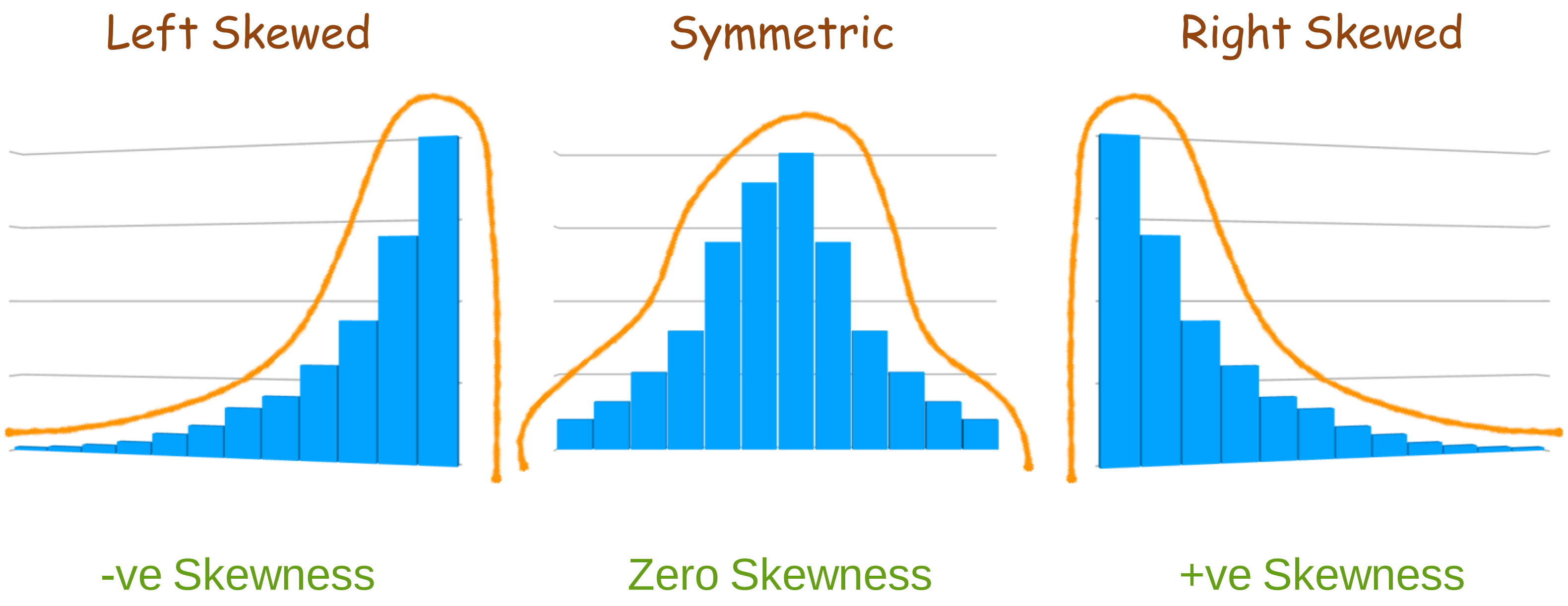
HISTOGRAMS

- Help to view ***data density***
- Help to see ***shape of distribution***

  1) Skewness
  2) Modality

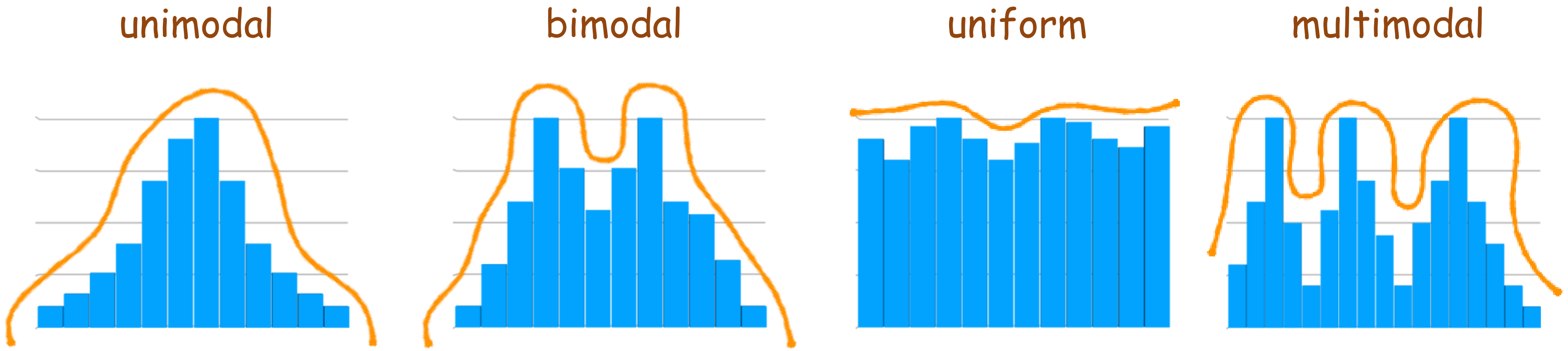


SKEWNESS

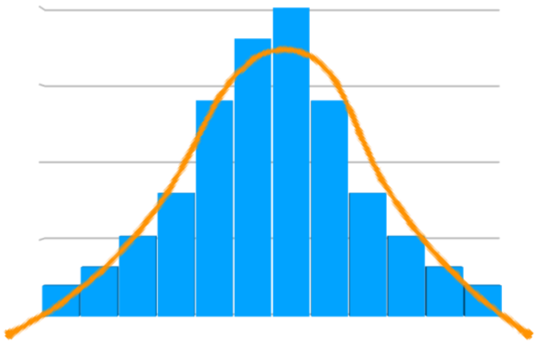


- Draw a smooth curve to see skewness
- Don’t rely on jagged edges

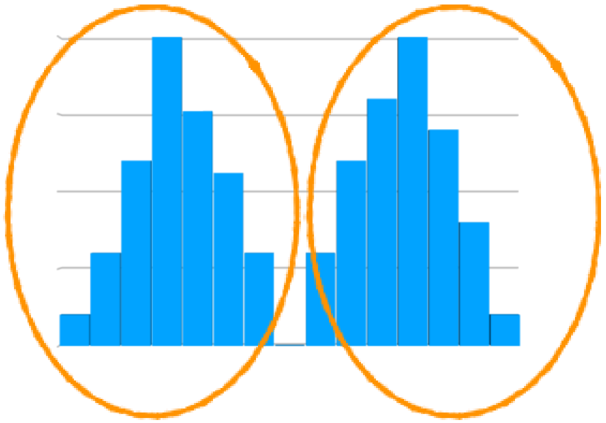
MODALITY



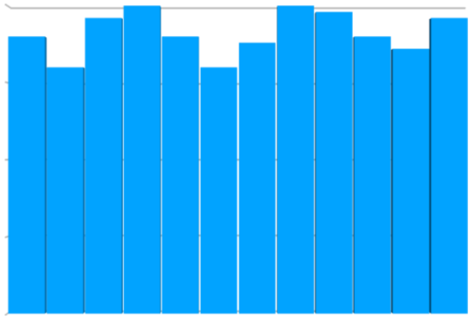
MODALITY (EXAMPLE)



Normal Distribution

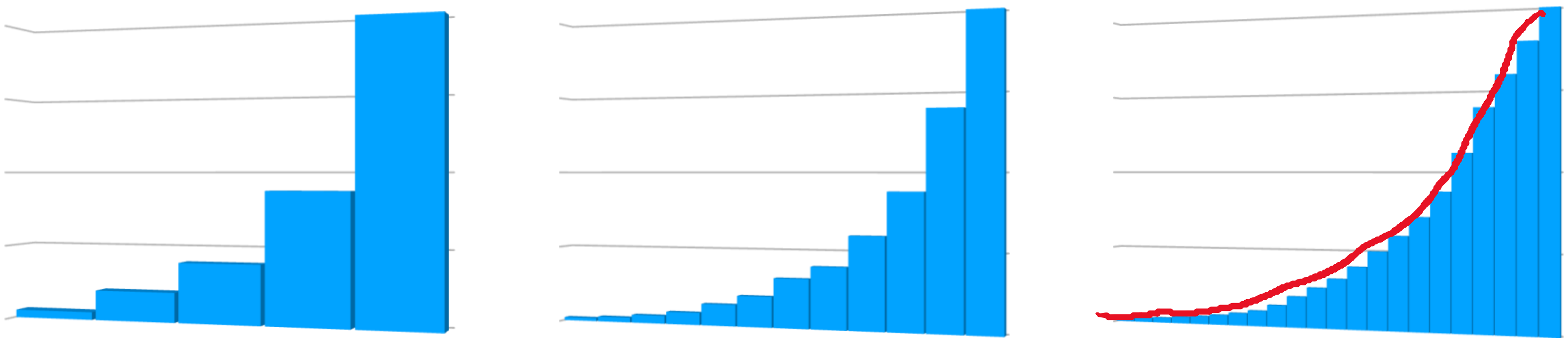


Two separate groups

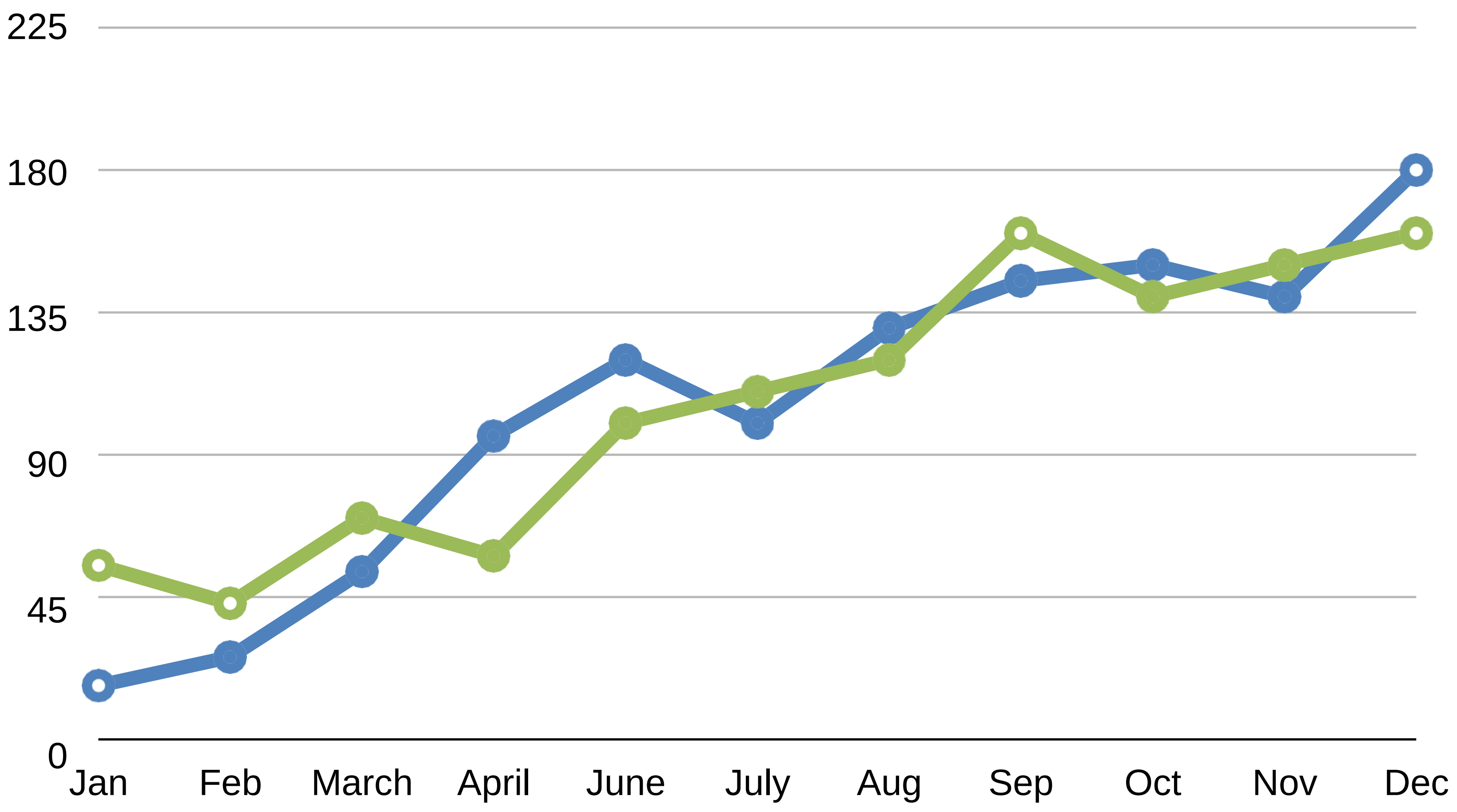


No trend

BINWIDTH

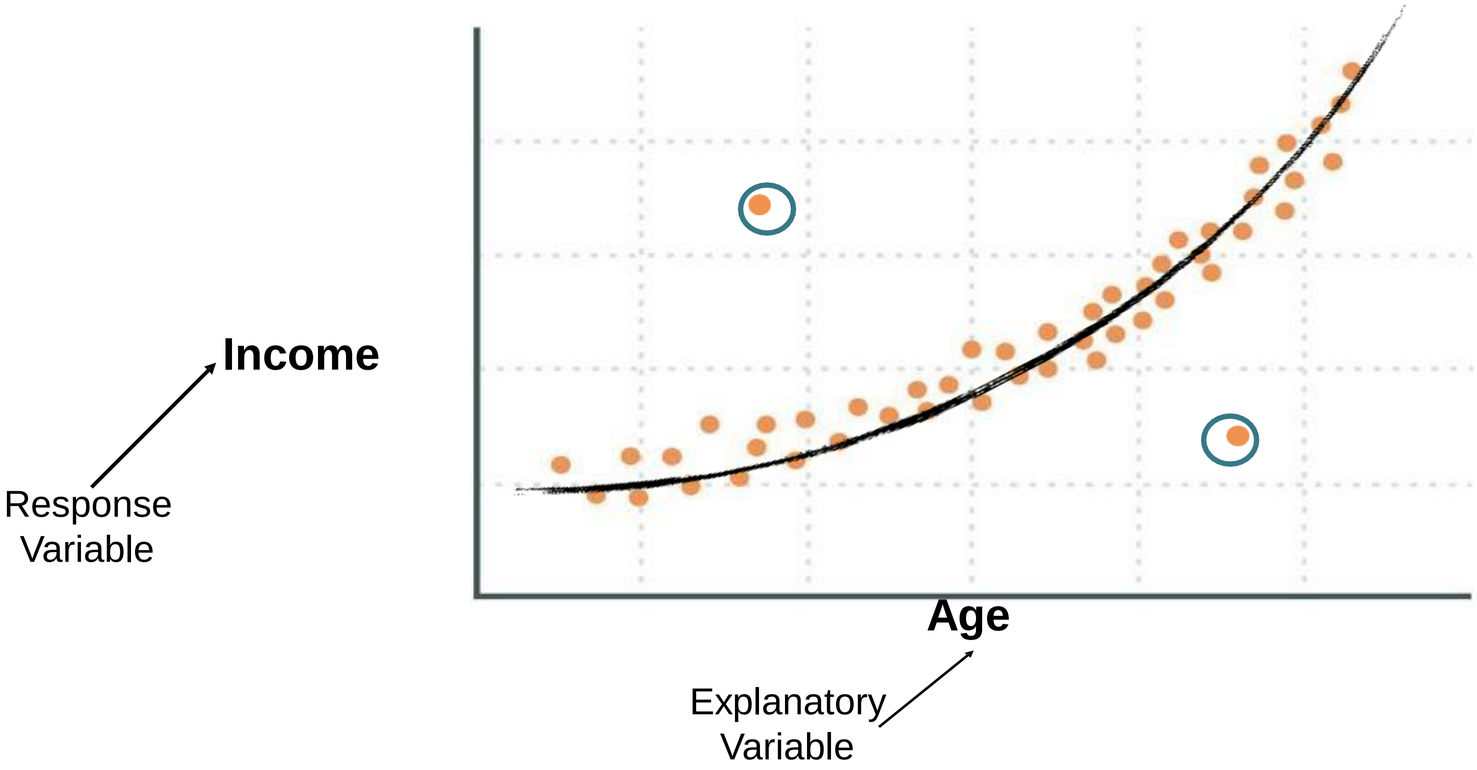


TIME PLOTS



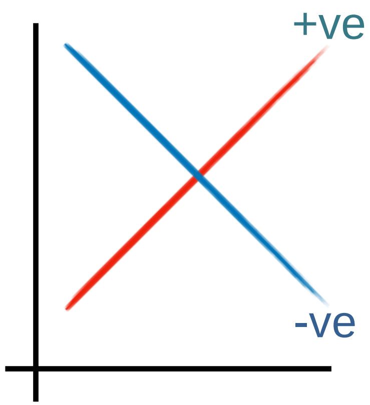
225
180
135
90
45
0
Jan
Feb
March
April
June
July
Aug
Sep
Oct
Nov
Dec

SCATTERPLOT

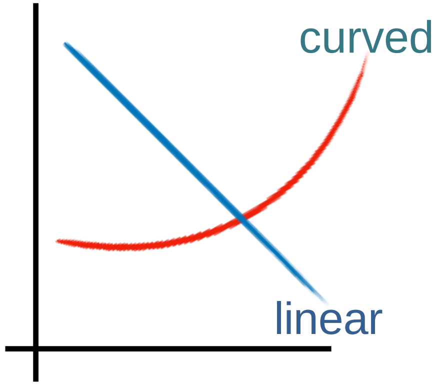


CHARACTERISTICS OF RELATIONSHIP

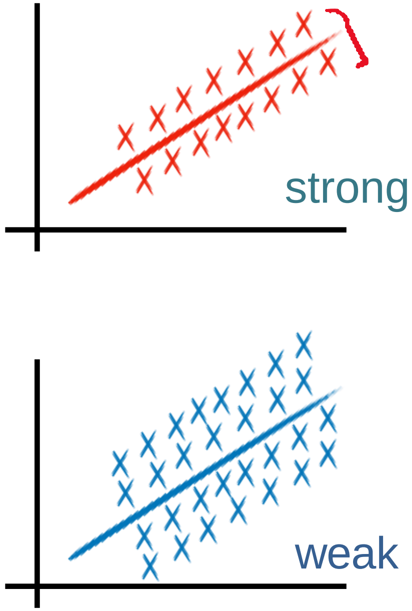
Direction



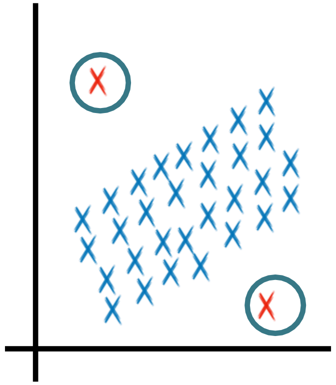
Shape



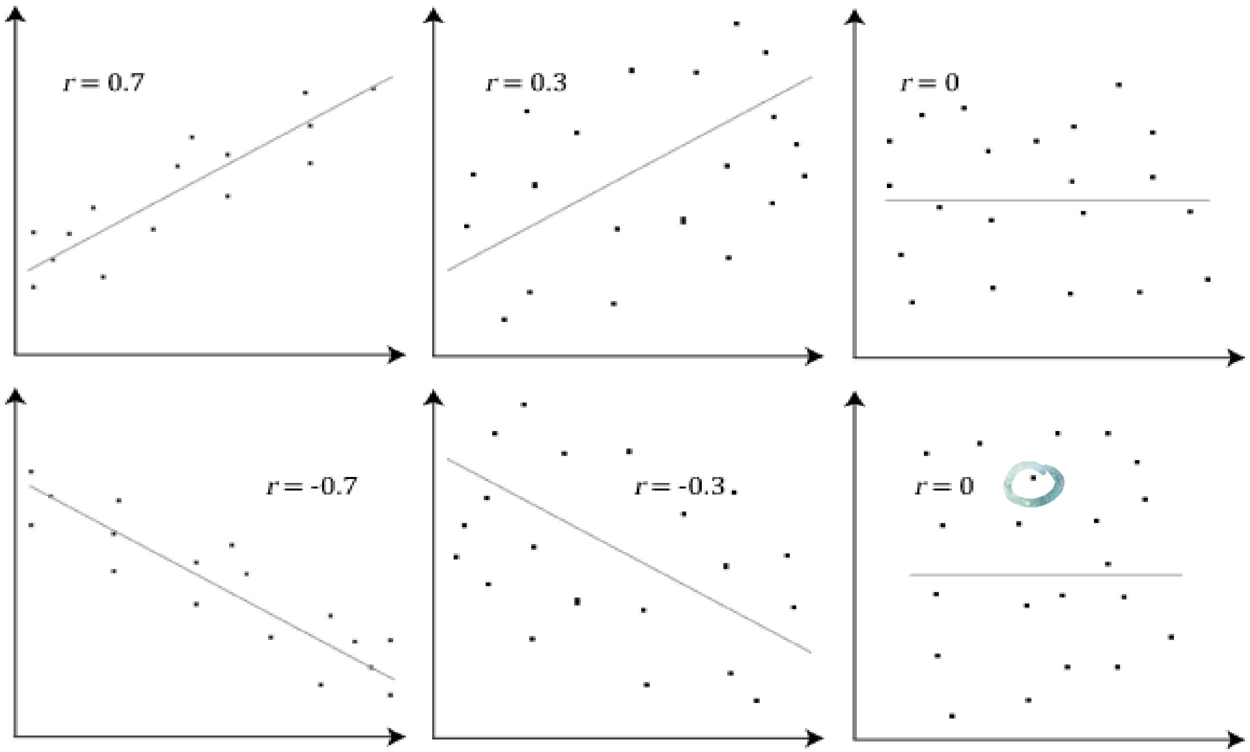
Strength



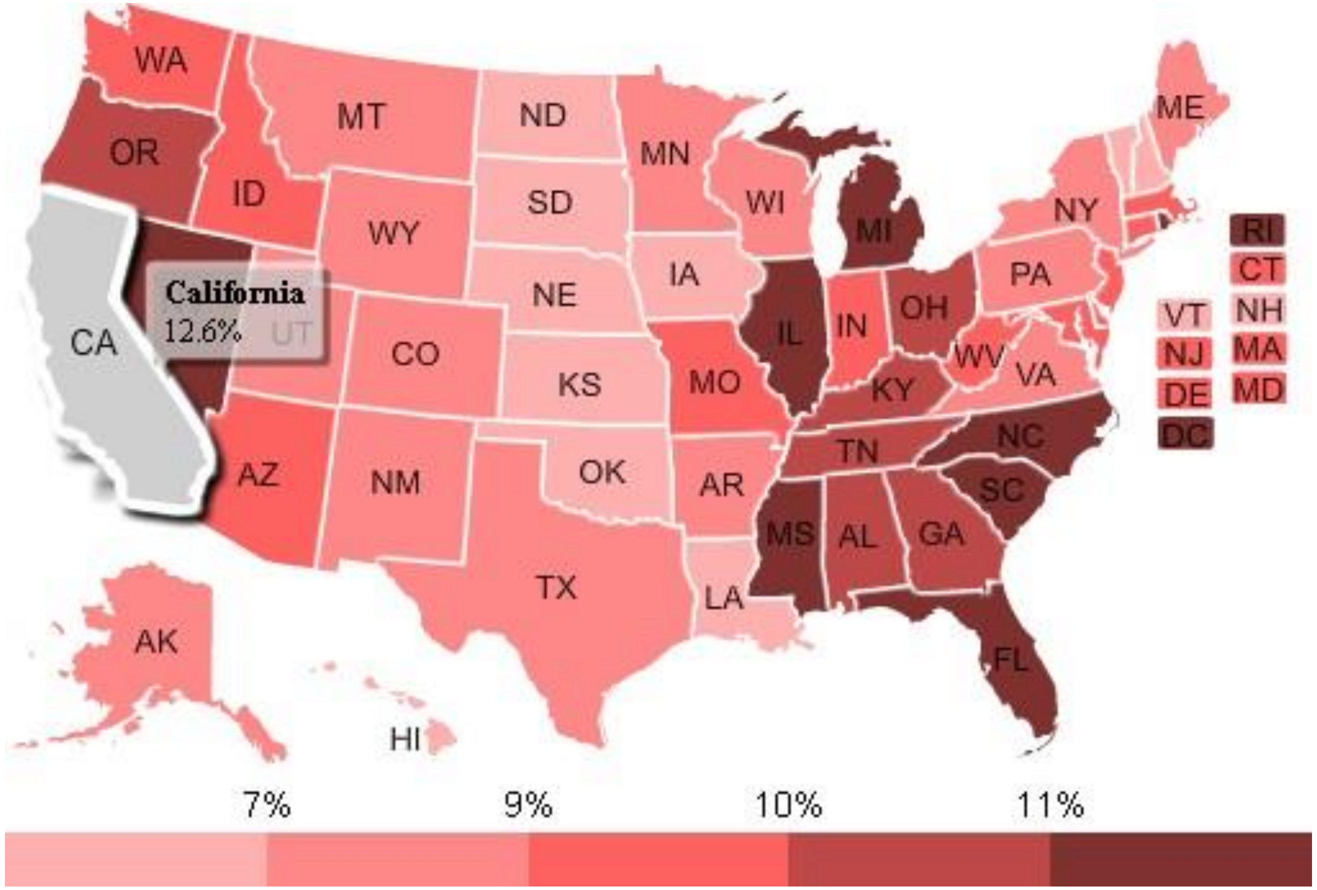
Outliers



CORRELATION (EXAMPLE)

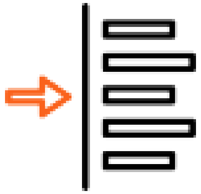


INTENSITY/HEAT MAPS



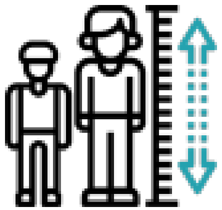
DESCRIPTIVE STATISTICS

DATA EXPLORATION ASPECTS – DESCRIPTIVE STATISTICS



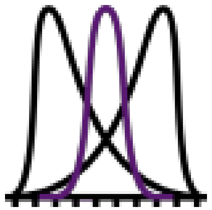
Measures of Central Tendency

- Mean
- Median
- Mode



Measures of Dispersion

- Range
- Variance
- Standard deviation



Shapes of Distribution

- Skewness
- Kurtosis

MEASURES OF CENTER

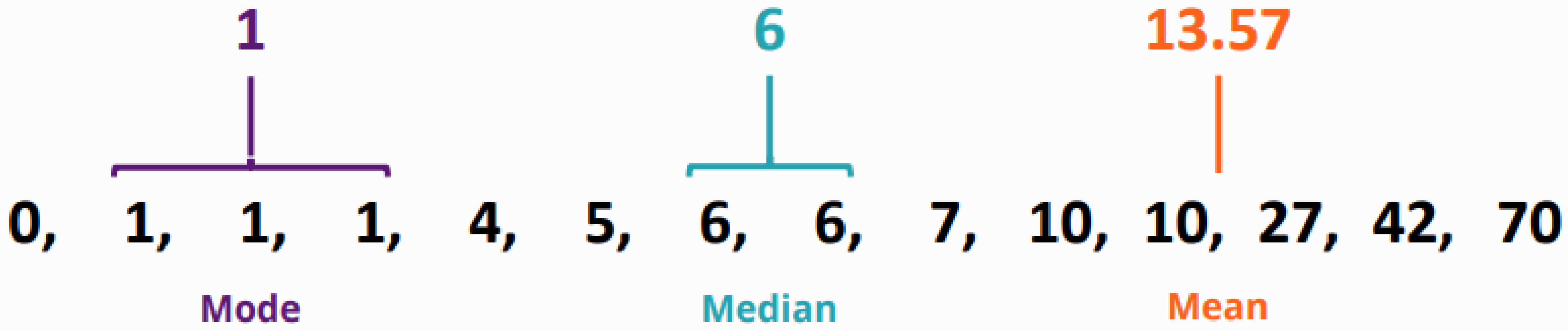
Mean	• The average value for the data. • The sum of all of the items divided by the number of items in the set.
Median	• The middle value when the data are in numerical order, or the mean of the two middle values if there are an even number of items.
Mode	• The value or values that occur most often in a data set. There can be more than one mode, or no mode.

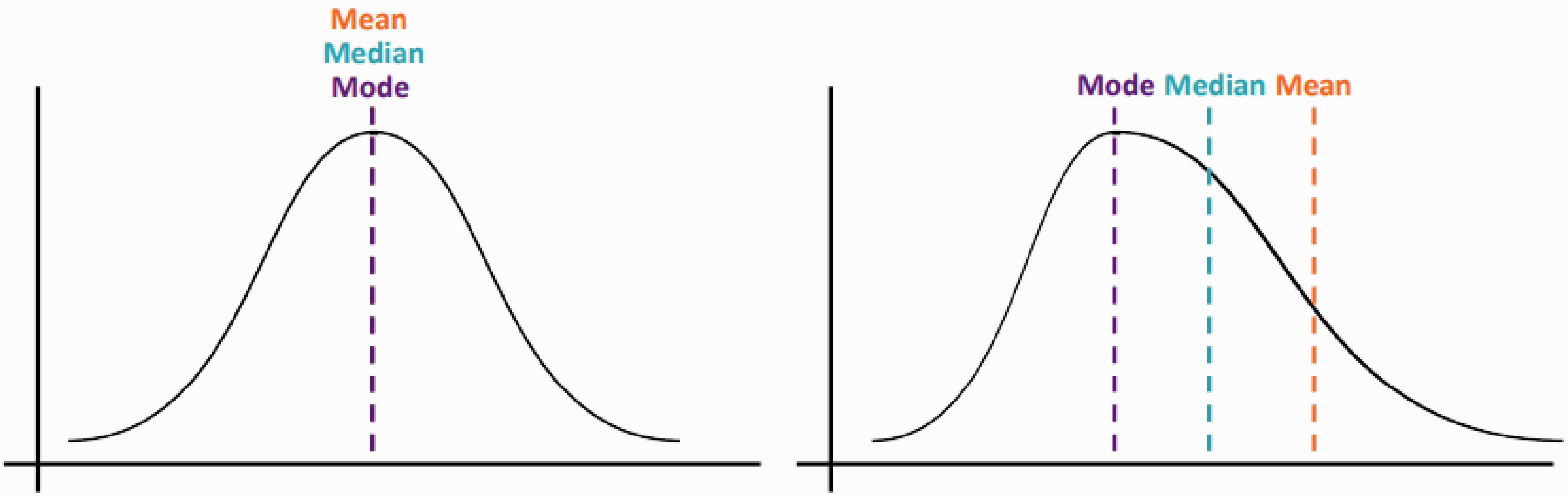
MEASURES OF CENTER

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 (50^{th} percentile) $$\text{Median} = \frac{77 + 62}{2} = 69.5$$

Measures of central tendency are single values that attempt to describe the central position of a set of data.





Mean, Median, Mode under normal distribution

Mean, Median, Mode under skewed distribution

MEASURES OF SPREAD

Range

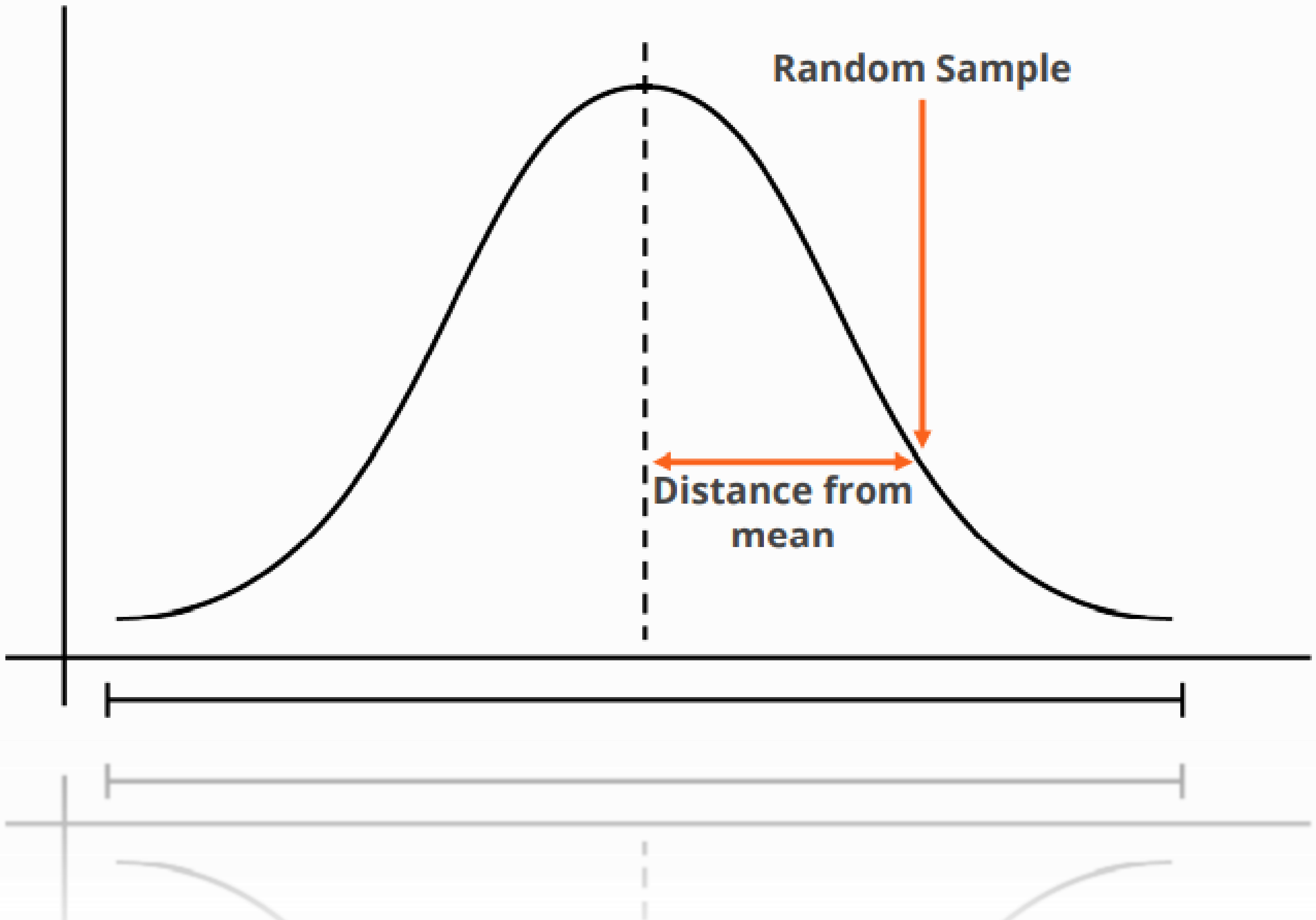
Variance

Standard Deviation

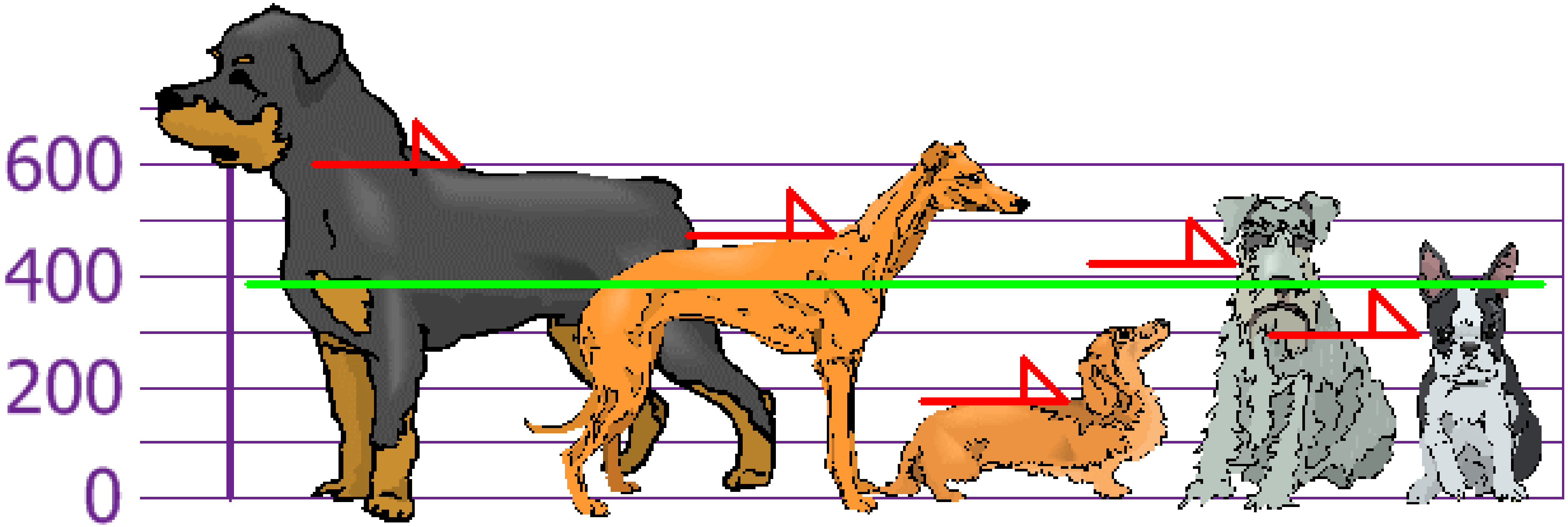
Interquartile Range

Mode, Median, Mean, Range, and Standard Deviation (1.3) - YouTube https://www.mathsisfun.com/data/standard-deviation.html

Measures of dispersion describe how much our data is either spread out or squeezed. The two most important dispersions we will cover are variance and standard deviation.

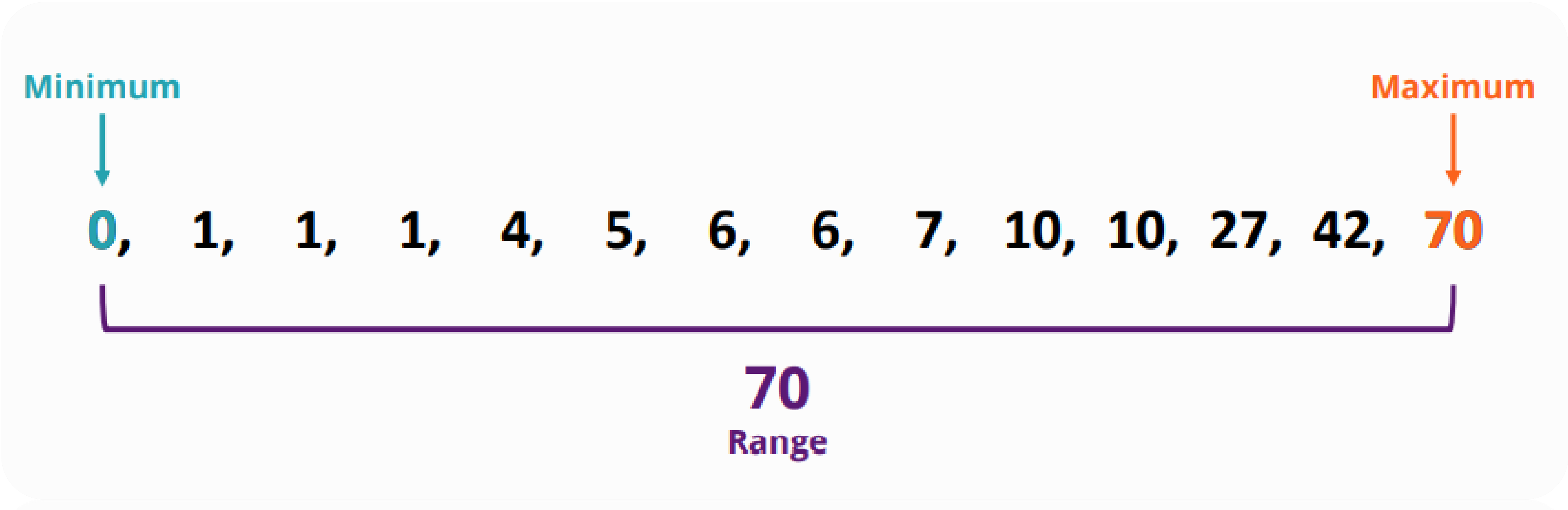


www.mathisfun.com



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**

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

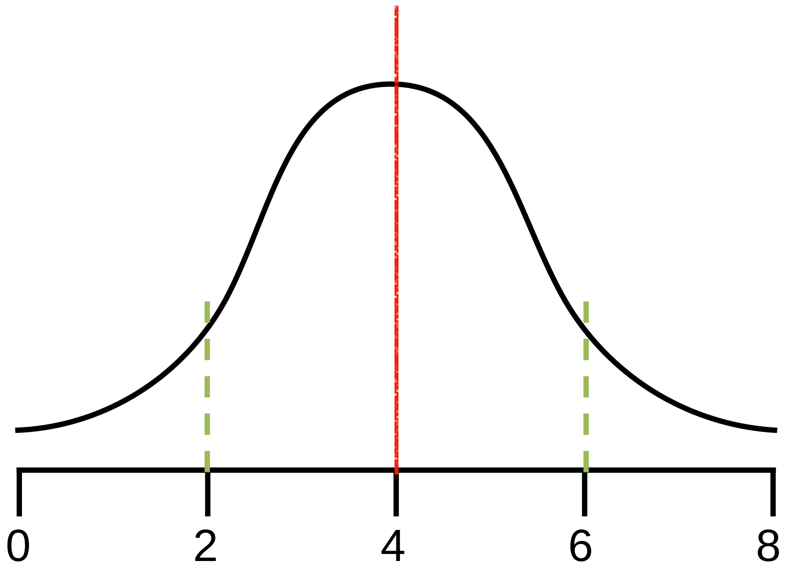
$$= \frac{2995.6}{9}$$

$$= 332.8$$

Sum of Squares

WHY SQUARE THE DIFFERENCES? IN STANDARD DEVIATION

- Get rid of negatives, so that the negatives and positives do not cancel each other during addition.



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

- Increase larger deviations more than smaller ones so that they are weighed more heavily.

STANDARD DEVIATION (SD)

- Square root of Variance
- It has the same units as the variable, which makes it useful in comparisons and calculations

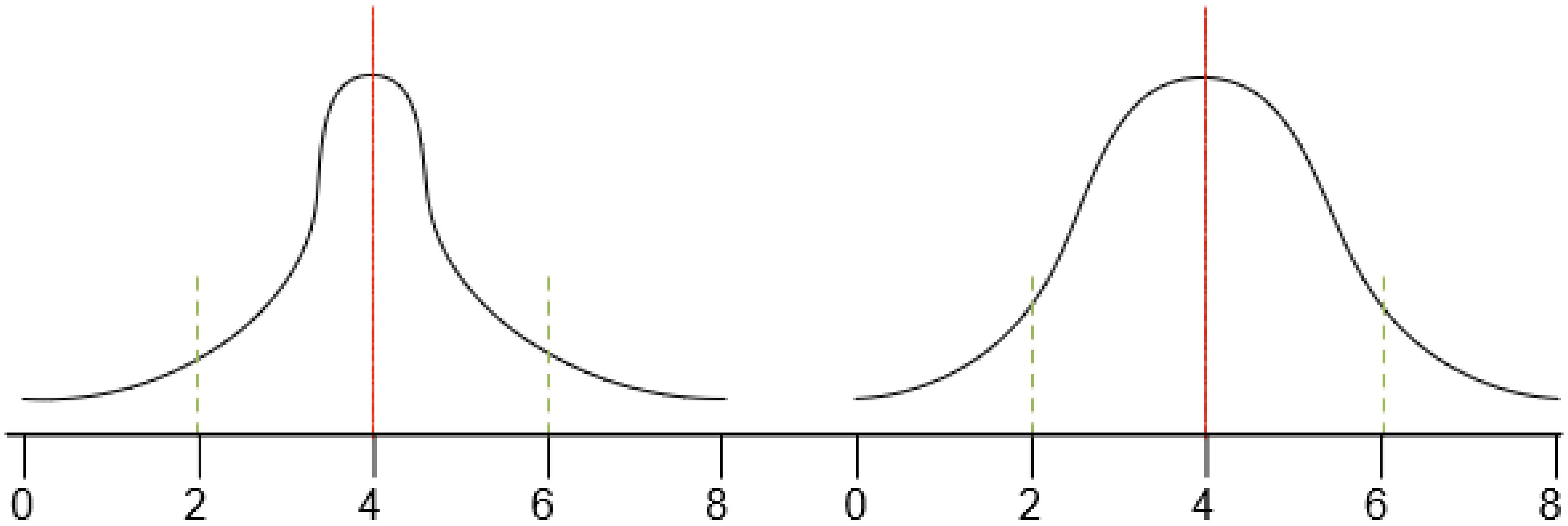
- Sample SD

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

- 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**

ROBUST STATISTICS – IQR IS ROBUST

Scneario #1			Scneario #2		
Quartiles	**Data**		**Quartiles**		**Data**
Q1	1		1		***Q1***
	2		2		
	3		3		
Q2	4		4		***Q2***
	5		5		
	6		6		
Q3	7		7		***Q3***
	8		8		
	9		9		
Q4	10		10		***Q4***
	11		11		
	12		12		
			1000		
	Average	***6.5***	***Average***	***82.9231***	
	IQR	6	***IQR***	6	

An Outlier of 1,000 had a limited impacted on IQR but extreme Impact on Average..

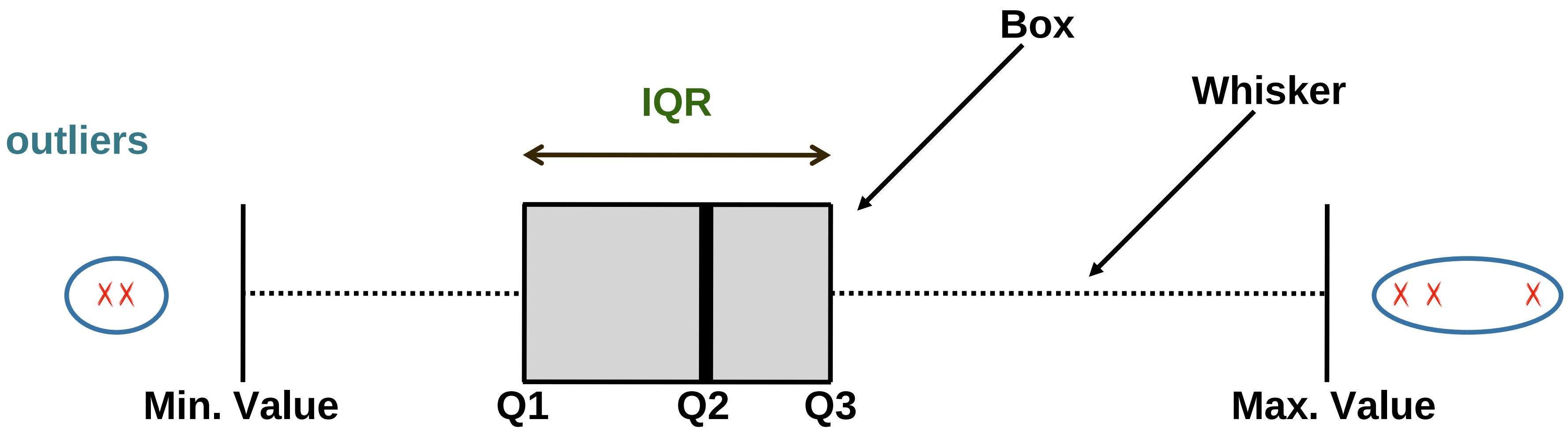
So Average is not a Robust Statistics Parameter
While IQR is a Robust Statistics Parameter

https://www.youtube.com/watch?v=tpToLyZibKM&t=1s

https://www.coursera.org/lecture/probability-intro/robust-statistics-ssktR

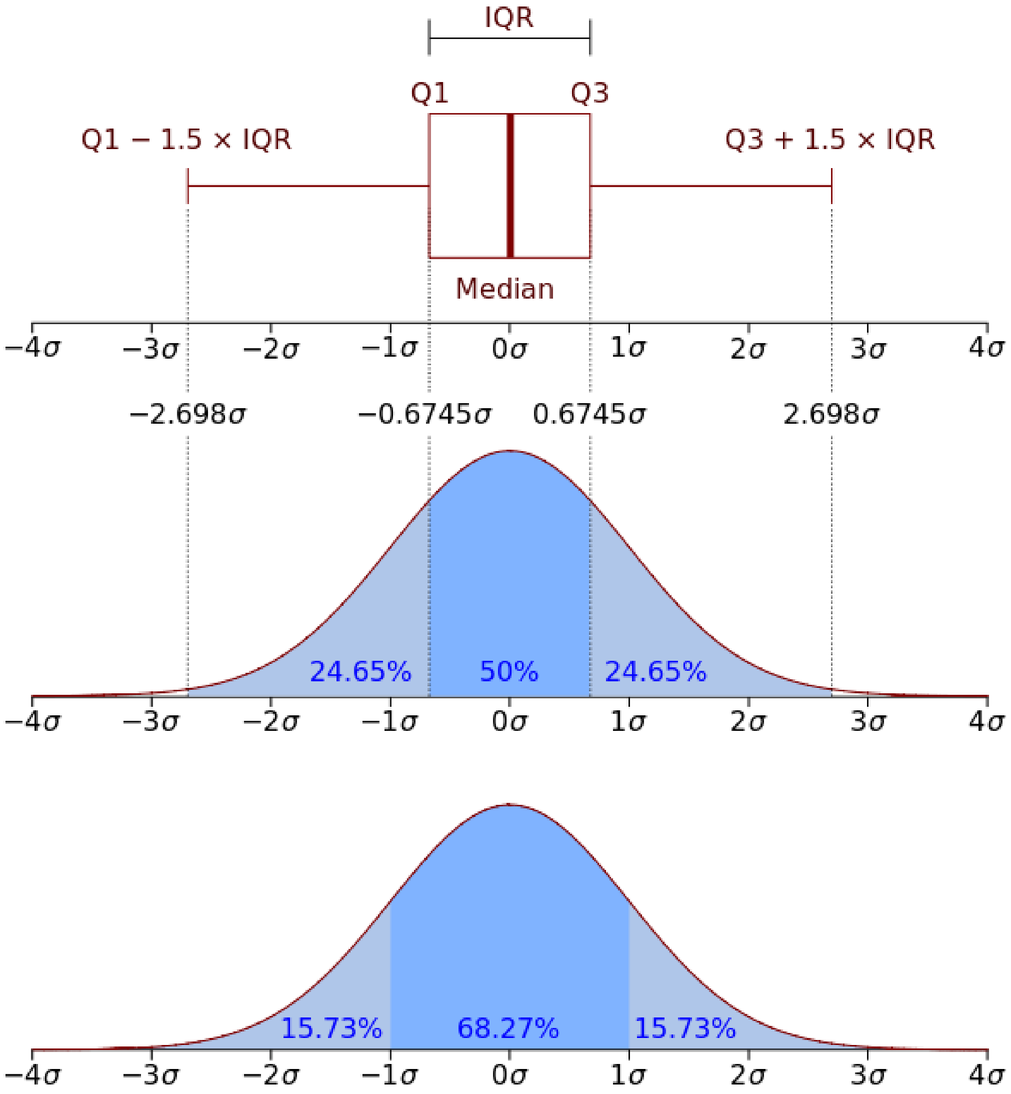
https://www.youtube.com/watch?v=PaRZge3njm4

BOX PLOTS

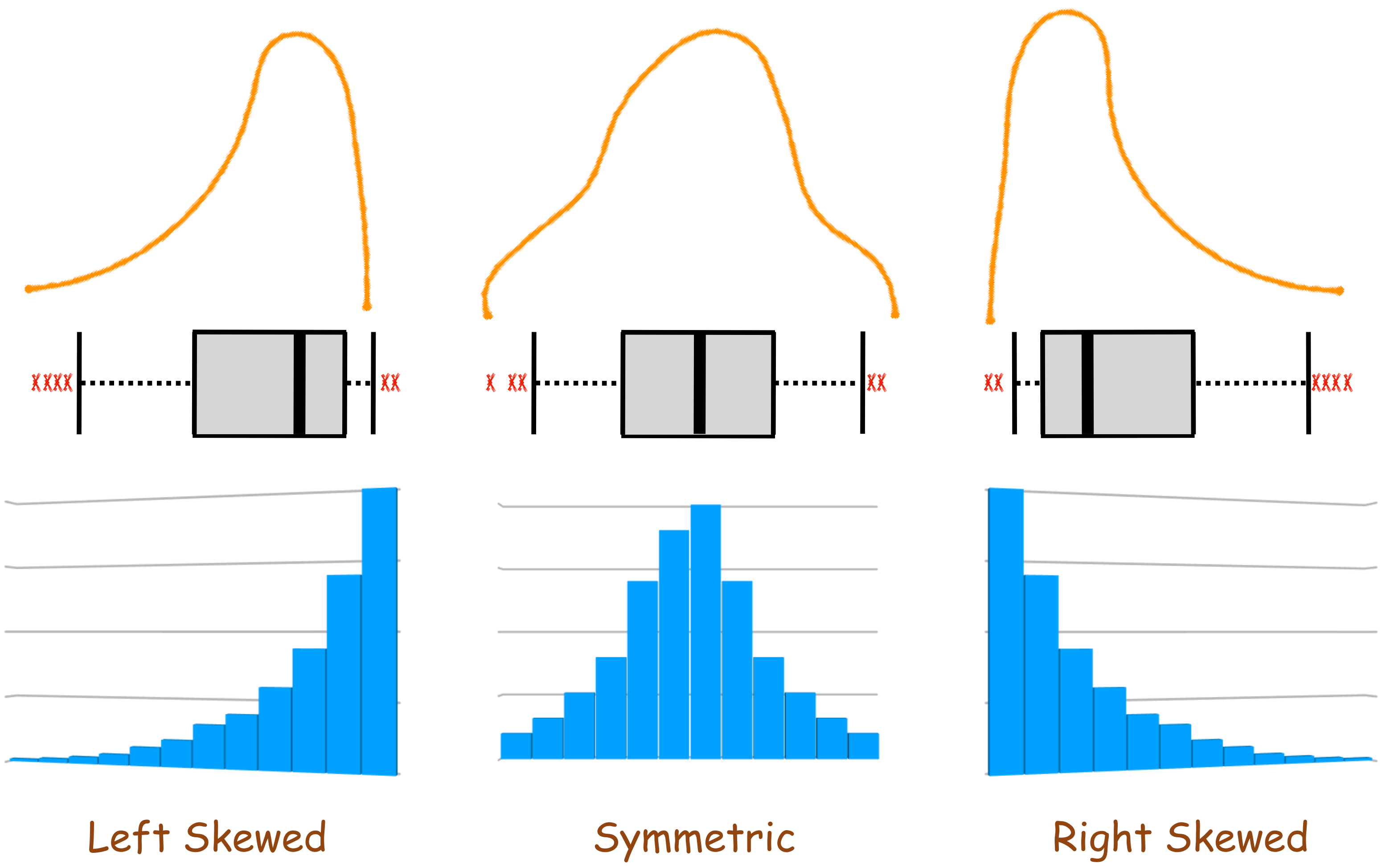


Min. Value :Lower Extreme (that’s not an outlier)
Q1 :Lower Quartile (25% of observations)
Q2 :Median (50% of observations)
Q3 :Upper Quartile (75% of observations)
Max. Value :Upper Extreme (that’s not an outlier)
IQR :Inter-Quartile Range = Q3 - Q1 (middle 50% of observations)

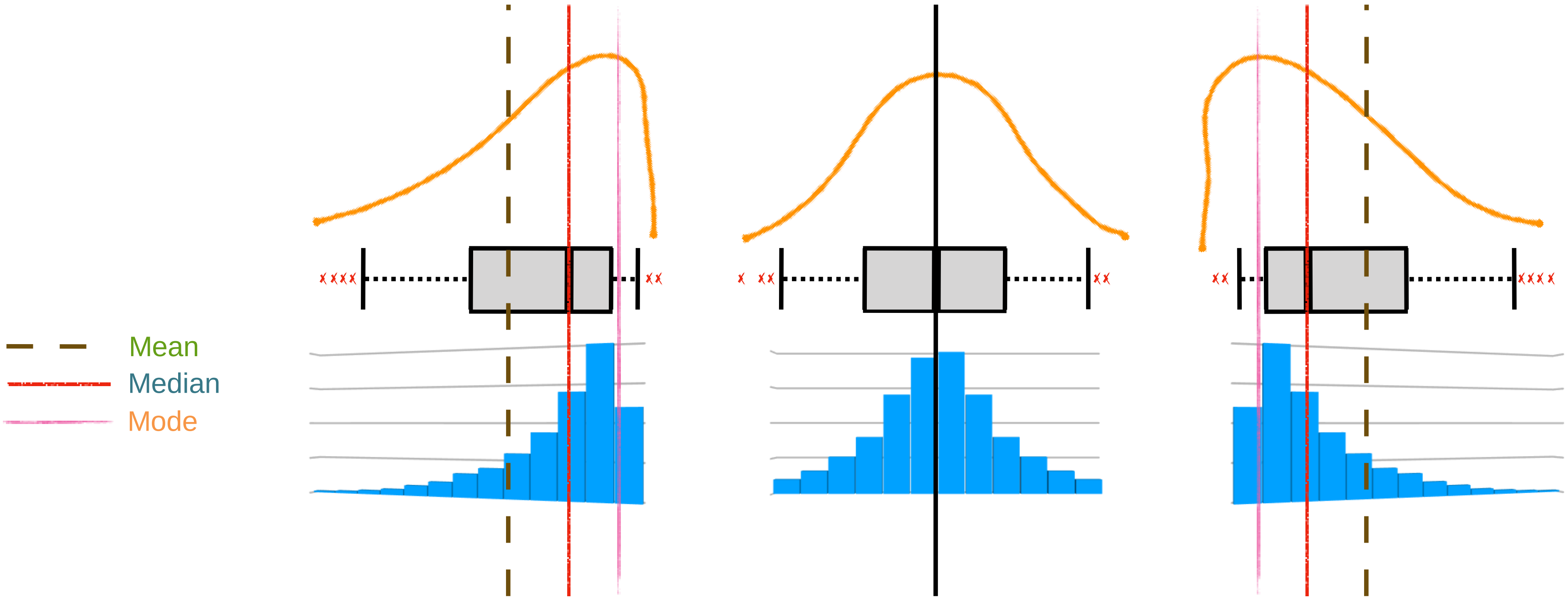
OUTLIERS


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%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73% 68.27% 15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

BOX PLOTS & SKEWNESS



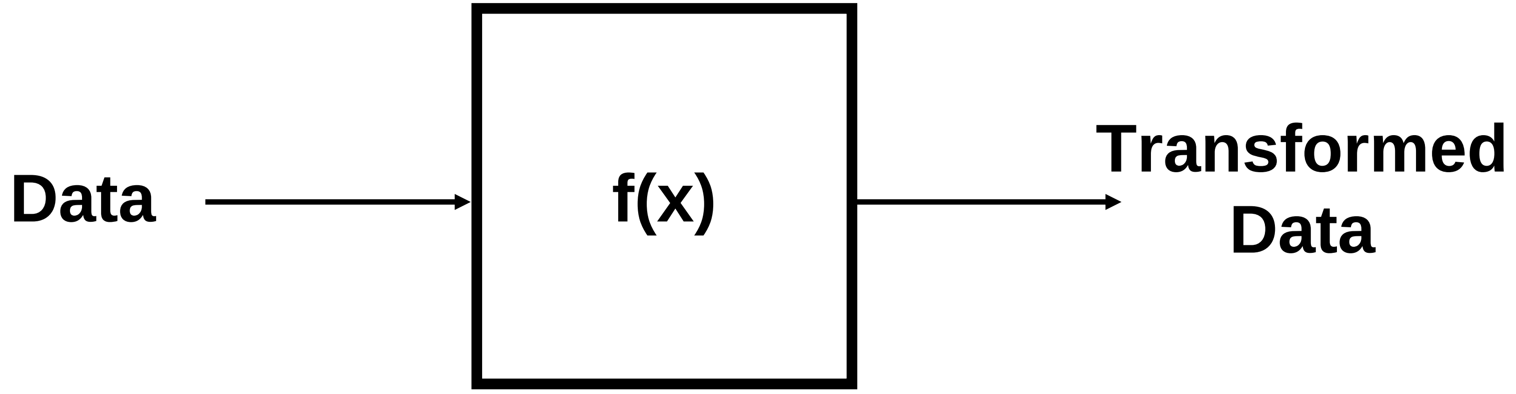
SKEWNESS VS MEASURES OF CENTERS



Mean < Median < Mode	Mean = Median = Mode	Mean > Median > Mode
Left Skewed	Symmetric	Right Skewed

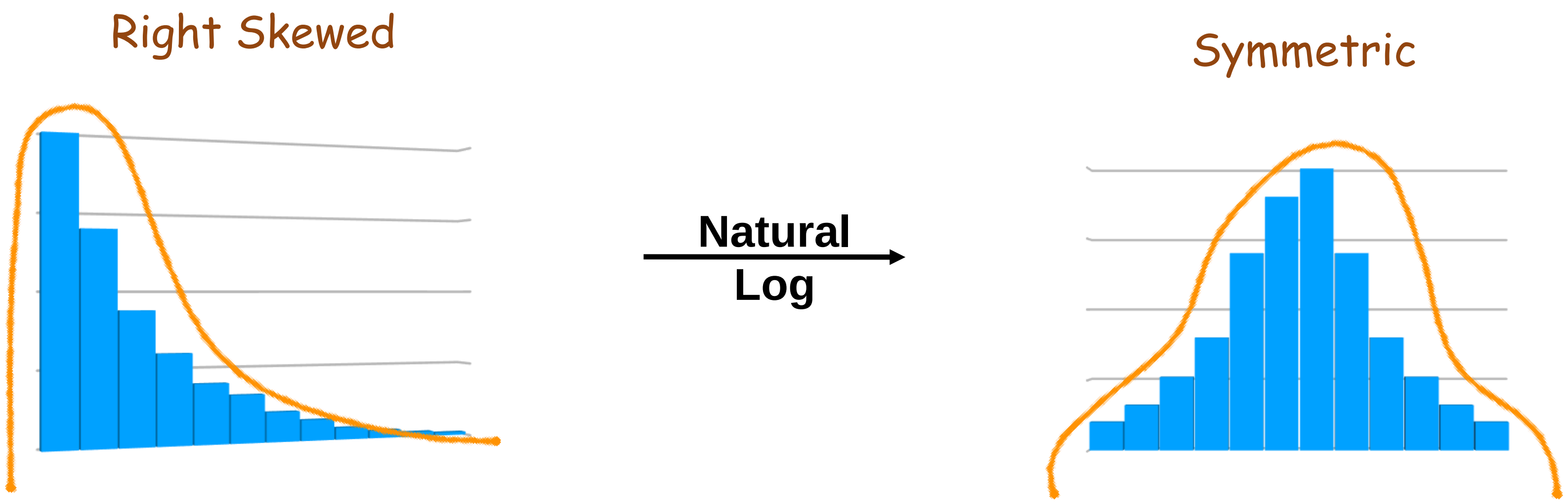
DATA TRANSFORMATIONS

- 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.



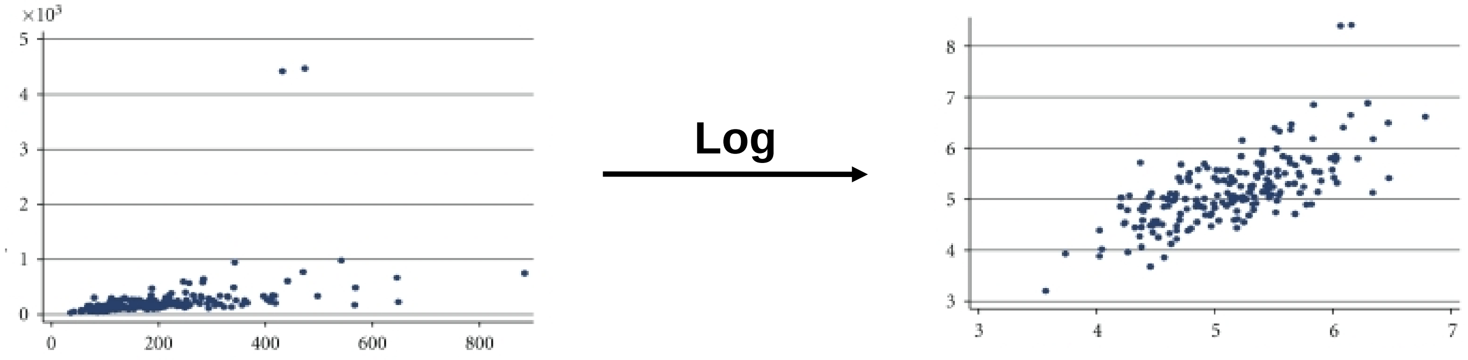
(NATURAL) LOG TRANSFORMATION

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



LOG TRANSFORMATION

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

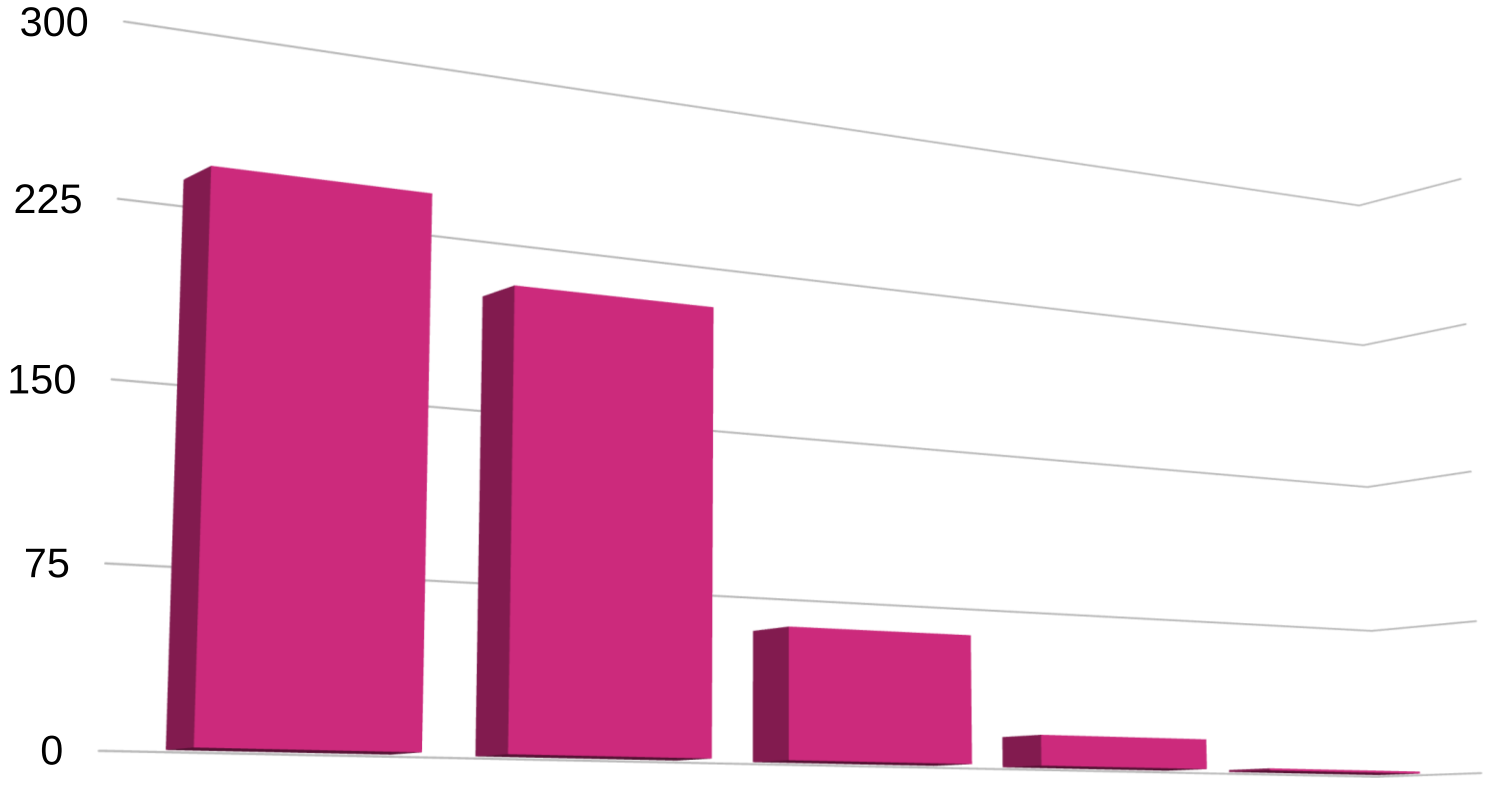


OTHER TRANSFORMATIONS

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

VISUALIZING CATEGORICAL DATA

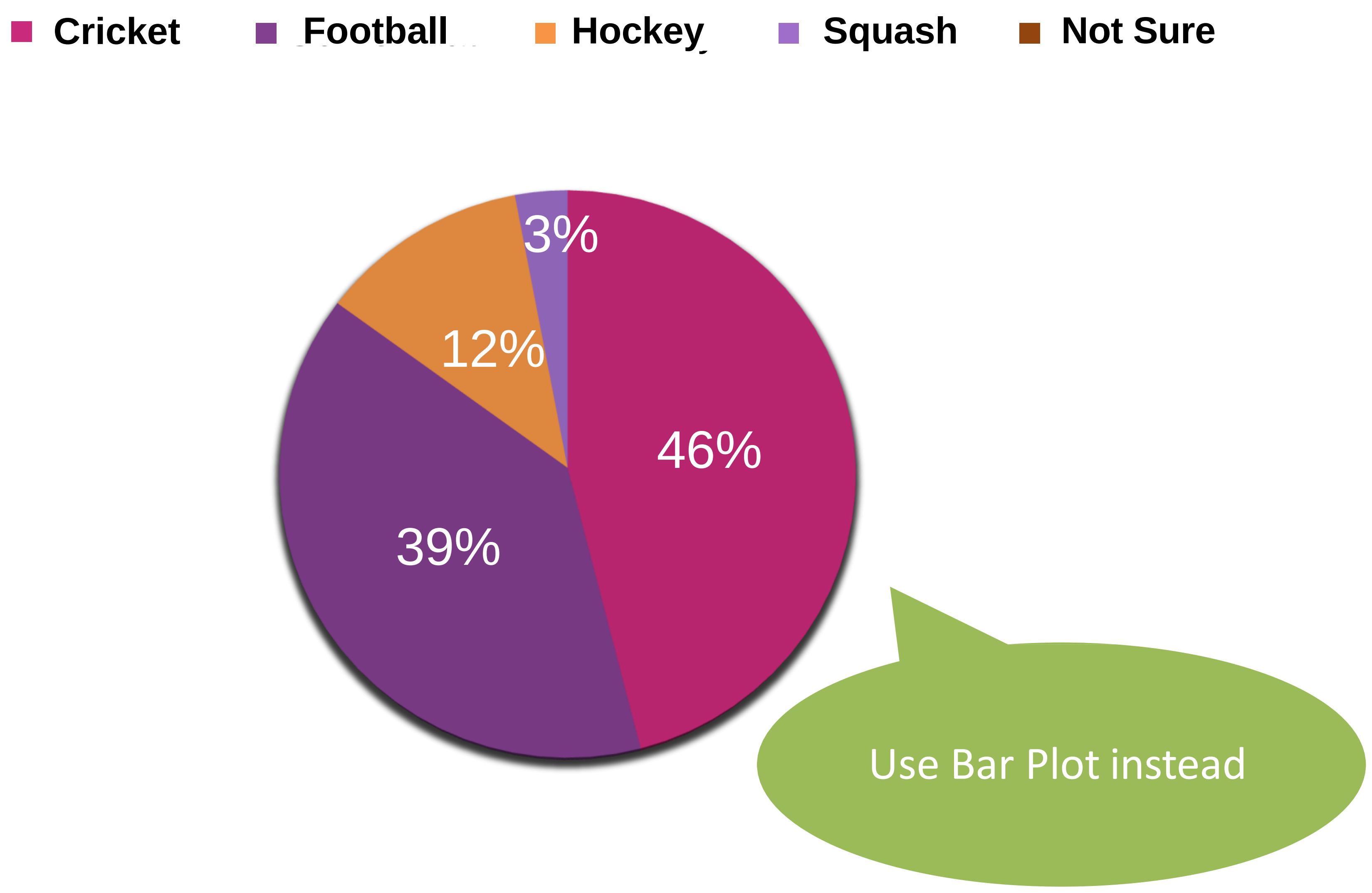
BAR PLOT



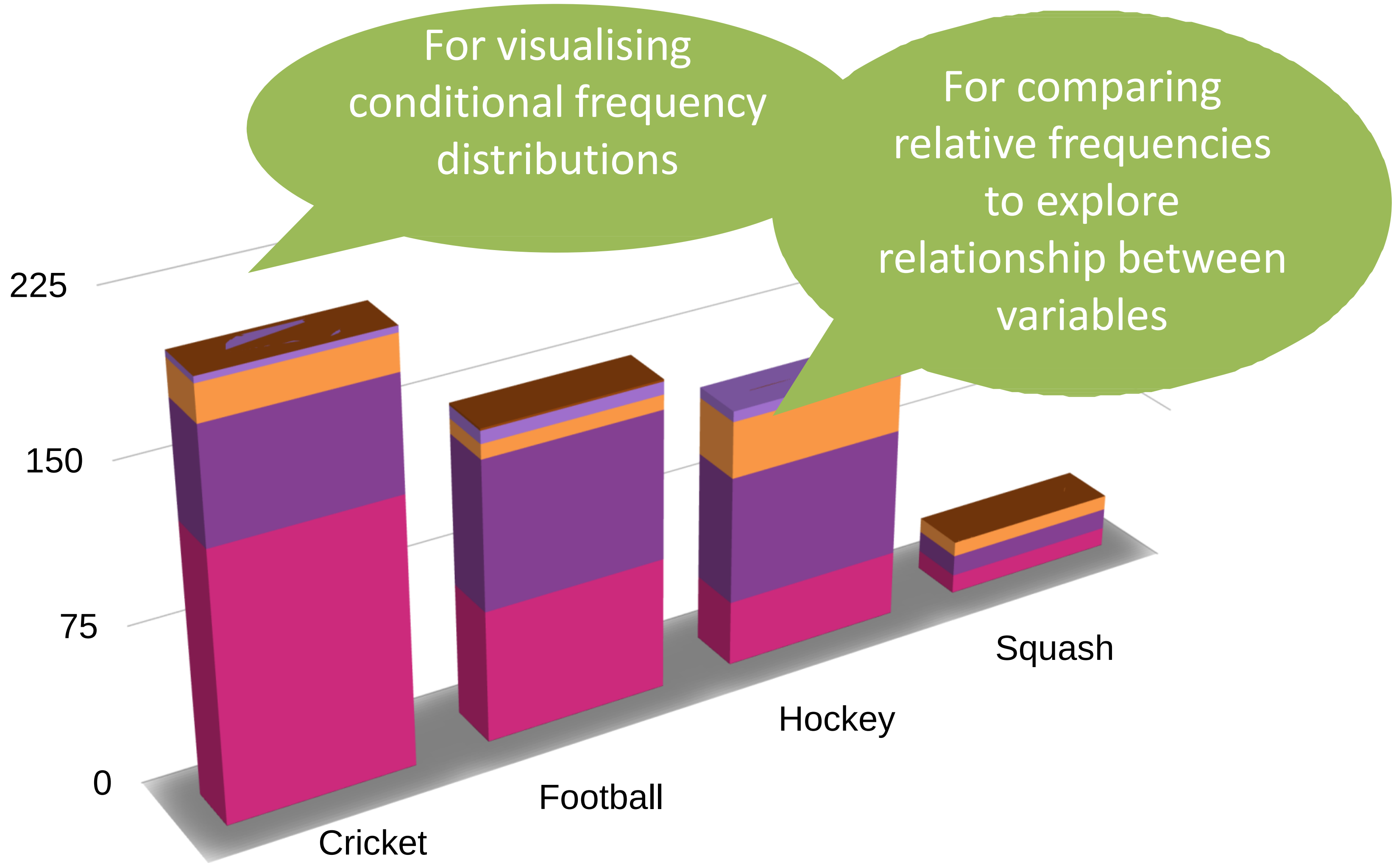
BAR PLOT VS HISTOGRAM

- 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

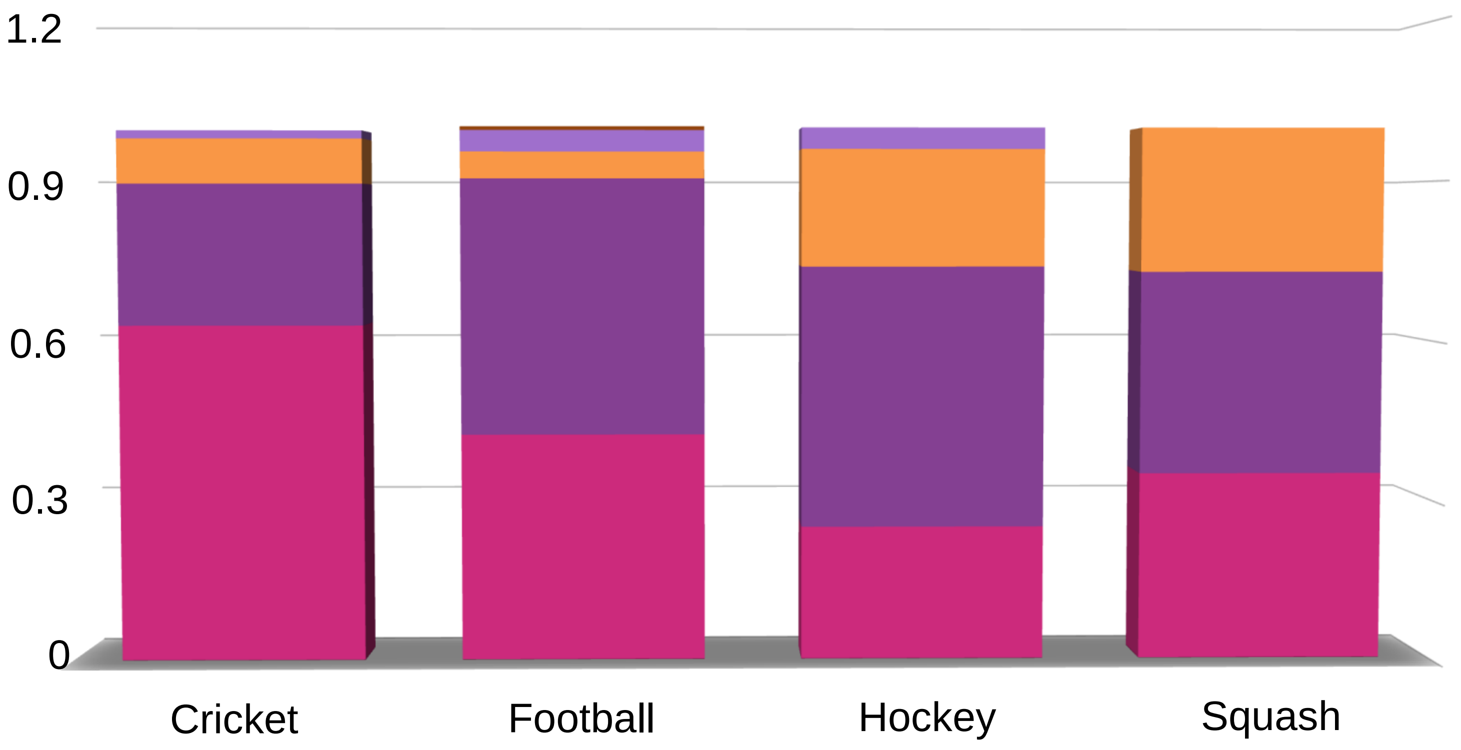
PIE CHART


Cricket
Football
Hockey
Squash
Not Sure
3%
12%
46%
39%
Use Bar Plot instead

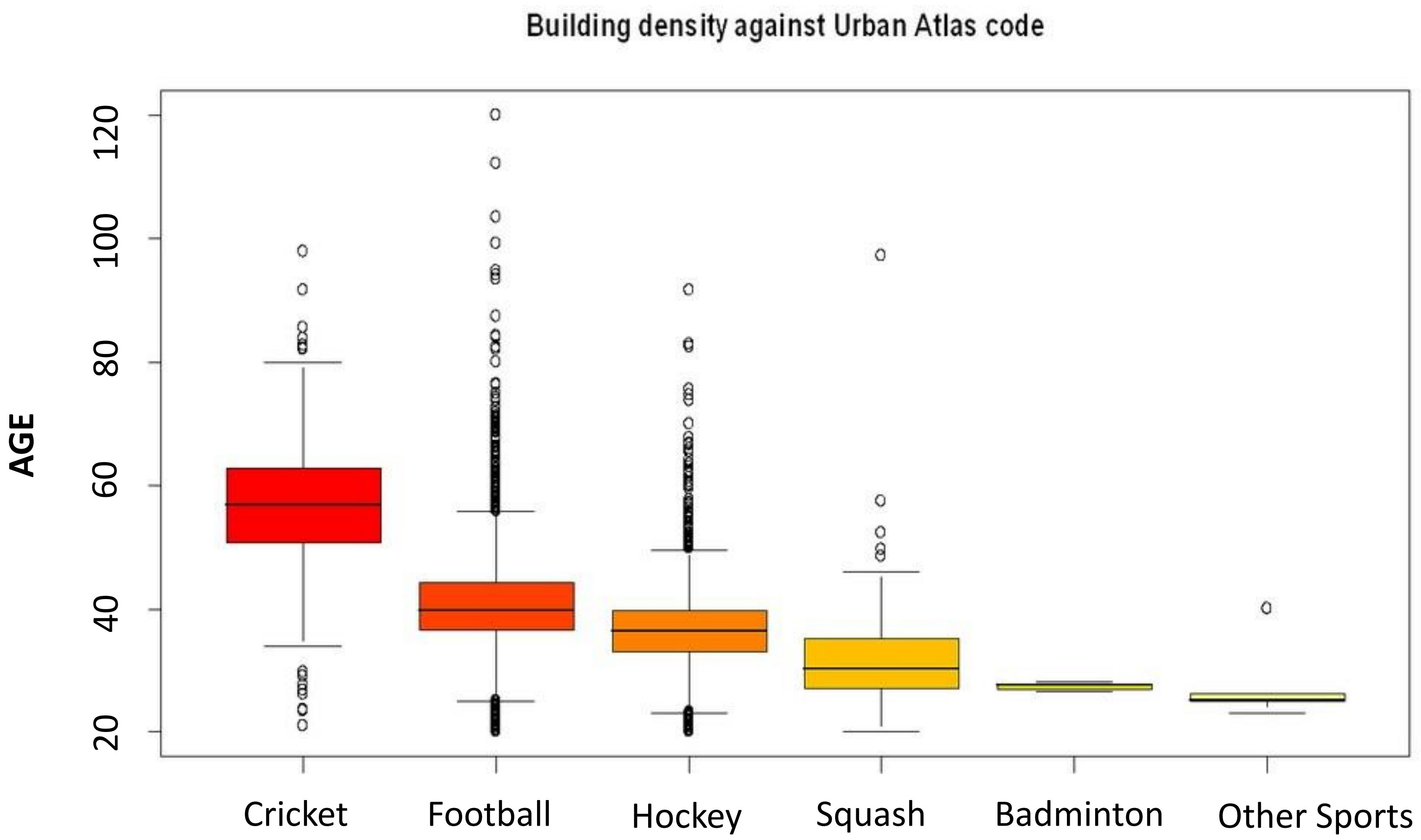
SEGMENTED BAR PLOT



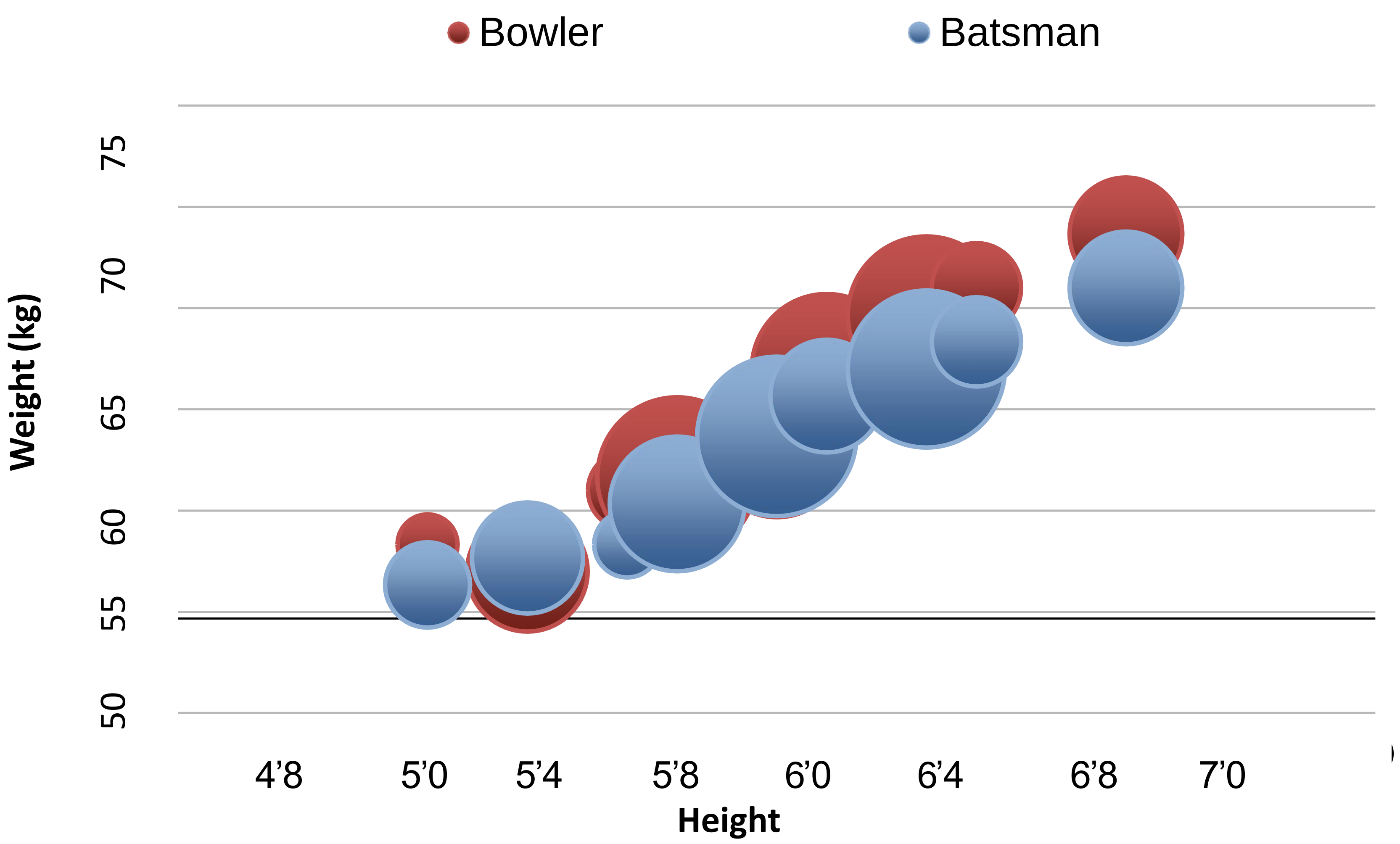
RELATIVE FREQUENCY SEGMENTED BAR PLOT



SIDE-BY-SIDE BOX PLOTS



BUBBLE PLOT



Bowler
Batsman
Weight (kg)
75
70
65
60
55
50
4’8
5’0
5’4
5’8
6’0
6’4
6’8
7’0
Height

PRINCIPLES OF VISUAL DESIGN

https://kevinlanning.github.io/DataSciLibArts/principles-of-data-visualization.html

https://www.inzonedesign.com/blog/6-principles-of-design/

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