BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session 1:
Overview of the course
& Descriptive Statistics
(Session 1: 7th /8th May 2022)

Overview of the course

TEXT BOOK

Probability and Statistics for Engineering and Sciences,

8th Edition, Jay L Devore, Cengage Learning

Overview of the course

- Descriptive Statistics
- Probability
- Conditional Probability
- Random Variables
- Probability Distributions – Univariate & Joint
- Sampling & Estimation
- Testing of Hypothesis – mean , proportions
- Regression
- Time Series Analysis

Contact Session	List of Topic Title	Reference
CS - 1	Descriptive Statistics: Data Visualisation, Measures of Central Tendency, Measures of Variability	T1:Chapter 1

Evaluation Components

- Assignment 1– 7%
- Assignment 2 – 8%
- Mid – 30%
- Compre – 45%

- Assignment submission is individual

"Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write"

H G Wells

!!!!

A famous statistician would never travel by airplane, because she had studied air travel and estimated the probability of there being a bomb on any given flight was 1 in a million, and she was not prepared to accept these odds.

One day a colleague met her at a conference far from home.

"How did you get here, by train?"

"No, I flew"

"What about the possibility of a bomb?"

"Well, I began thinking that if the odds of one bomb are 1:million, then the odds of TWO bombs are (1/1,000,000) x (1/1,000,000) = 10^{-12}. This is a very, very small probability, which I can accept. So, now I bring my own bomb along!"

Statistics may be defined as science that is employed to

- Collect the data
- Present and organize the data in a systematic manner
- Analyse the data
- Infer about the data
- Take decision from the data.

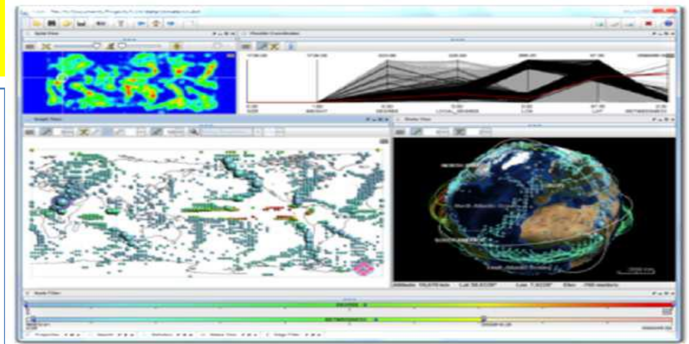
Statistics may be defined as numerical data with a view to analyse it.

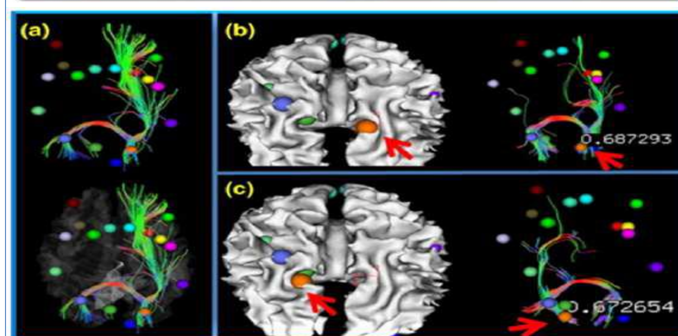
Need for Data Visualization

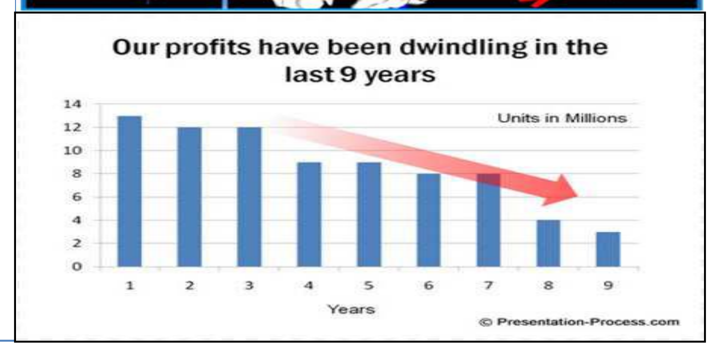
Tool to enable a user get insight into data

Broadly three types of goals:

- To **explore**:
  - *Nothing is known*
  - *Required to get an insight*
- To **analyze** :
  - *There are hypotheses*
  - *Used for verification or falsification*
- To **present**:
  - *We have the required information*
  - *Used for communication of result*







Source: Google images

Hits/Game	Number of Games	Relative Frequency	Hits/Game	Number of Games	Relative Frequency
0	20	.0010	14	569	.0294
1	72	.0037	15	393	.0203
2	209	.0108	16	253	.0131
3	527	.0272	17	171	.0088
4	1048	.0541	18	97	.0050
5	1457	.0752	19	53	.0027
6	1988	.1026	20	31	.0016
7	2256	.1164	21	19	.0010
8	2403	.1240	22	13	.0007
9	2256	.1164	23	5	.0003
10	1967	.1015	24	1	.0001
11	1509	.0779	25	0	.0000
12	1230	.0635	26	1	.0001
13	834	.0430	27	1	.0001
				19,383	1.0005

Statistical Visualization

A picture is worth a thousand words!

- Bar chart / graph
- Histogram
- Box plot
- Pie chart
- Density plot
- Line chart
- Frequency polygons
- Scatter plots

Chart Types

Line charts are great when it comes to displaying patterns of change across a continuum.

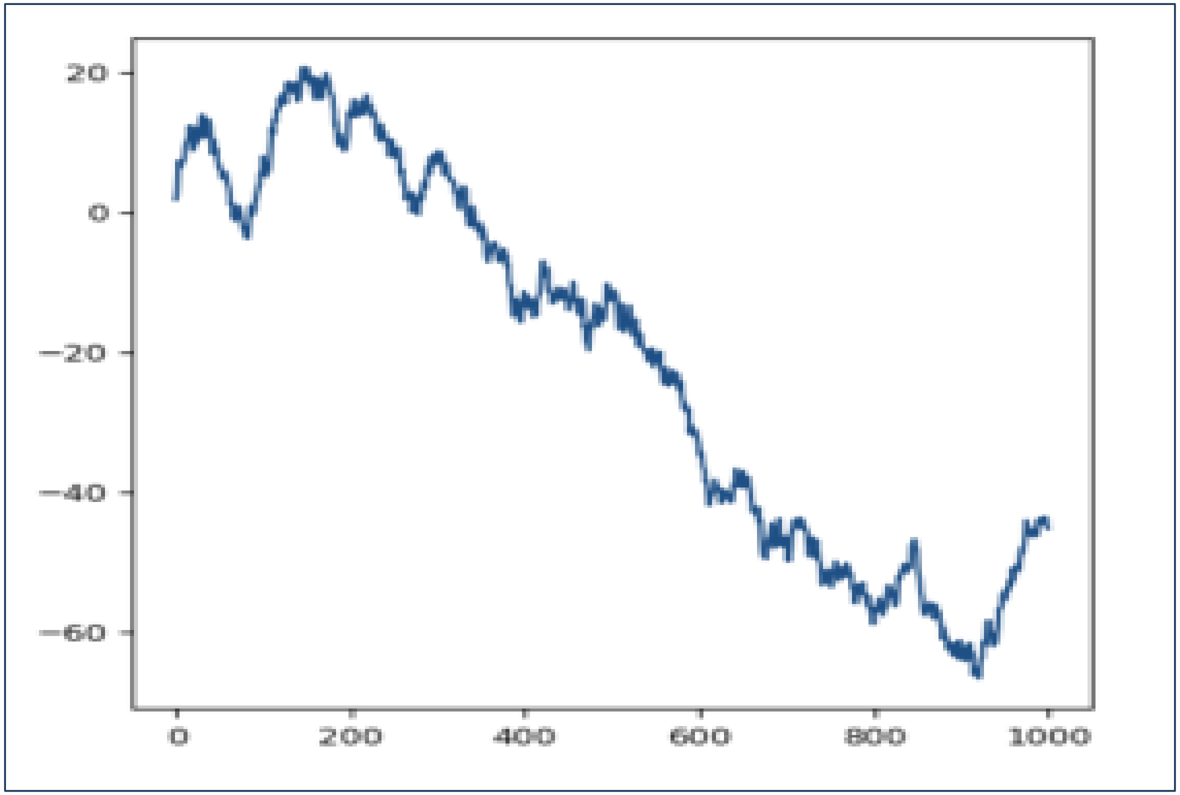
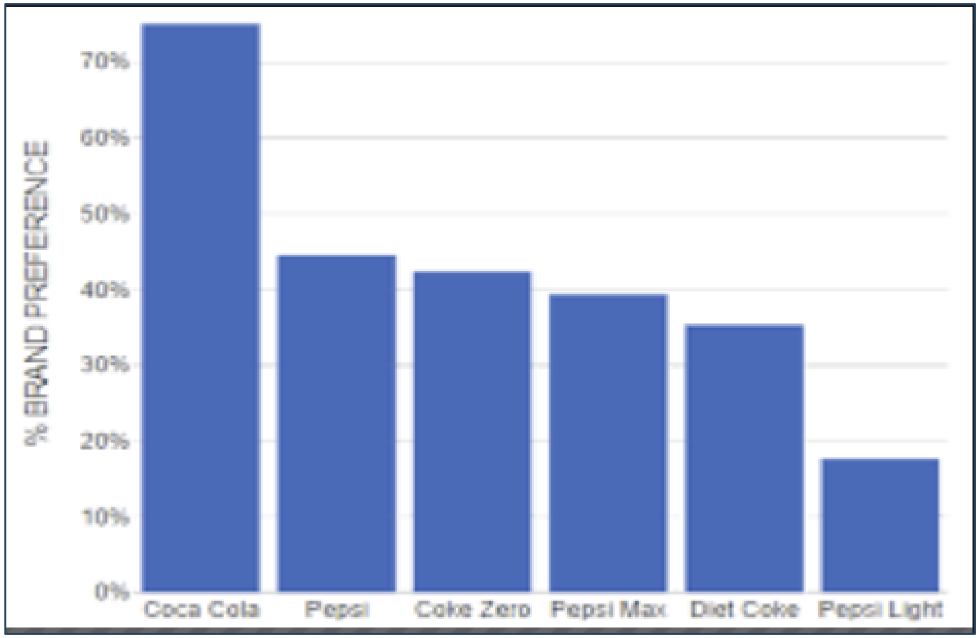
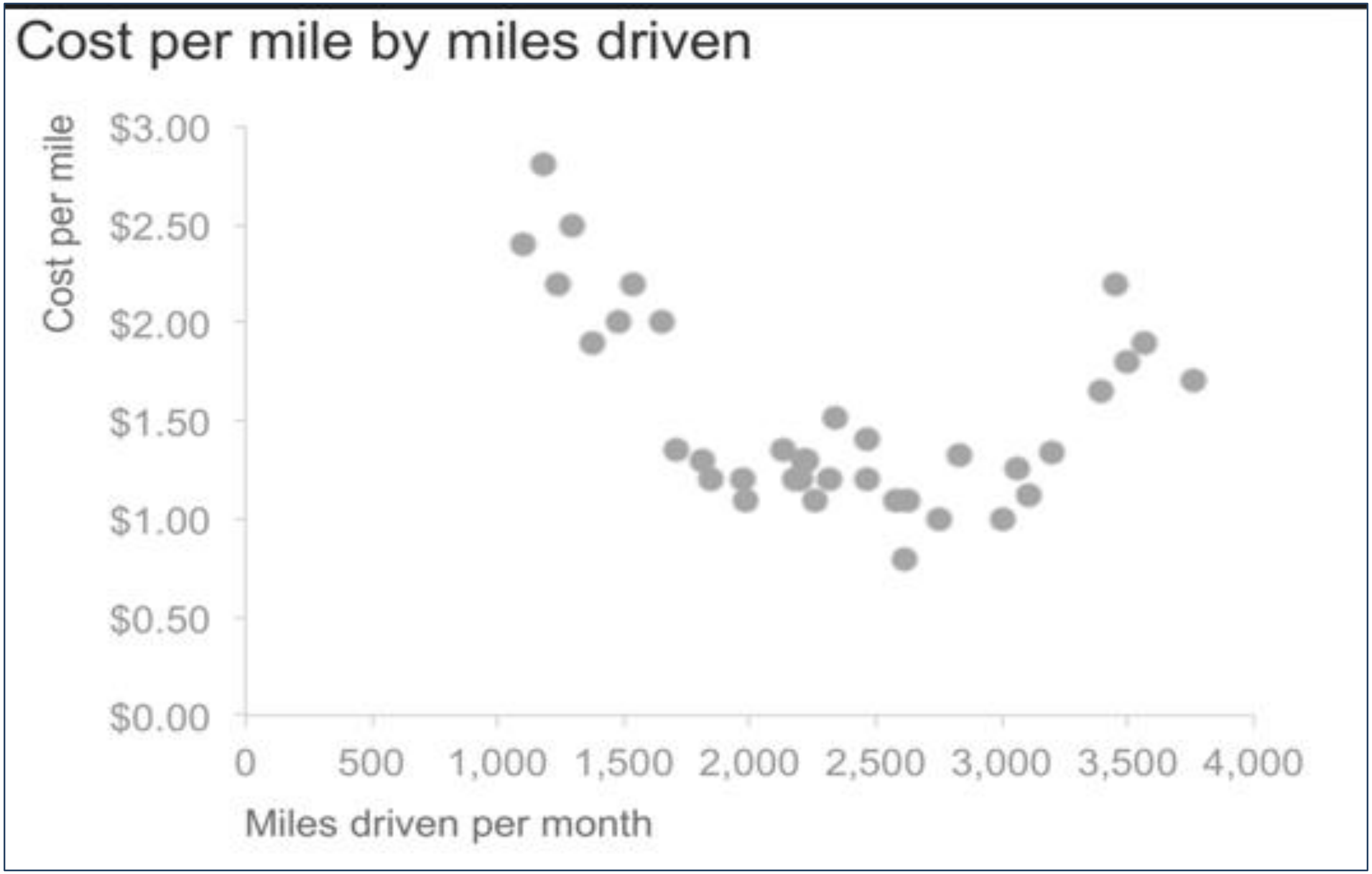


Chart Types

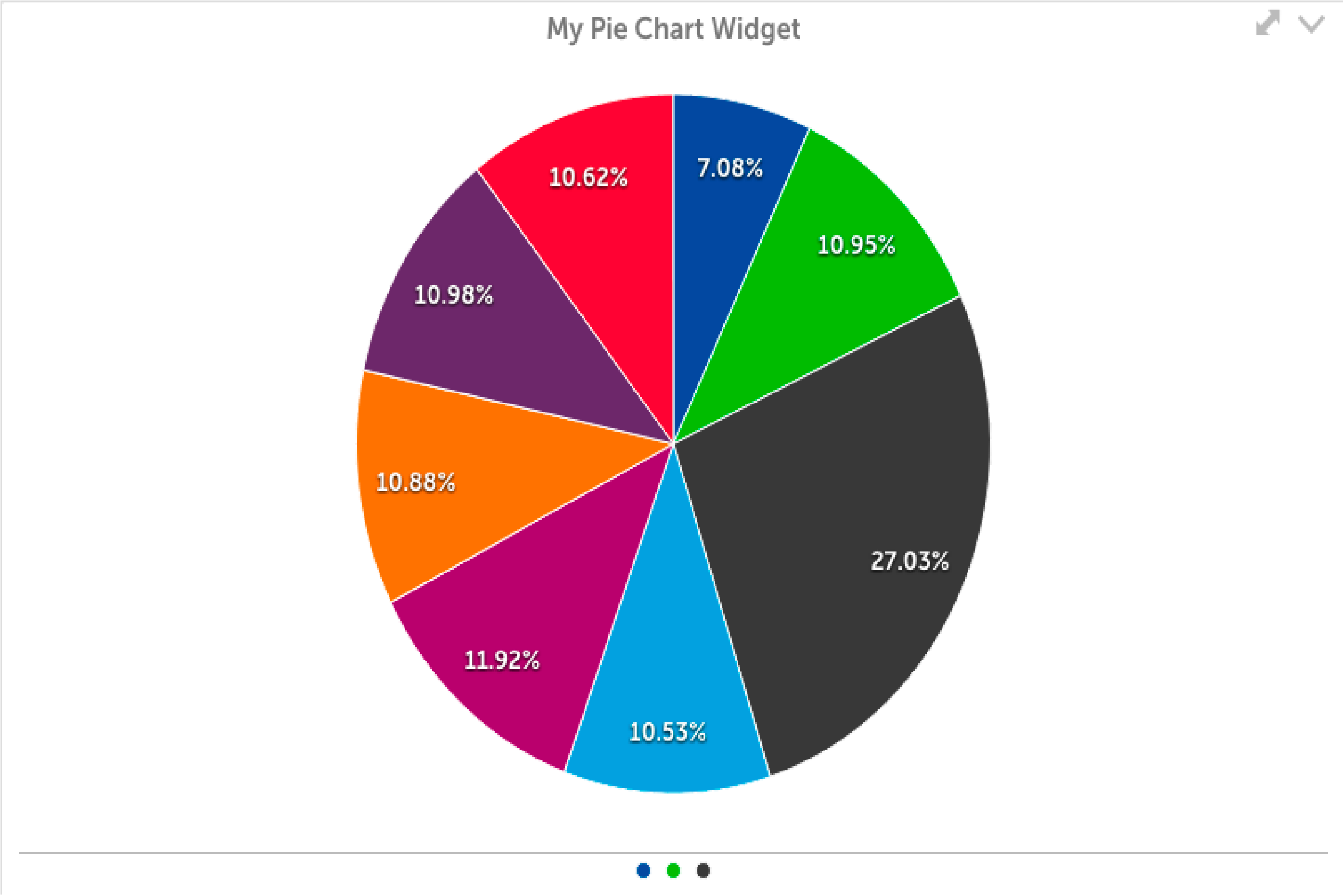
- Choose **bar charts** if you want to compare items in the same category.
- The objective is not just to compare but also show how much one is better or worse than the rest.

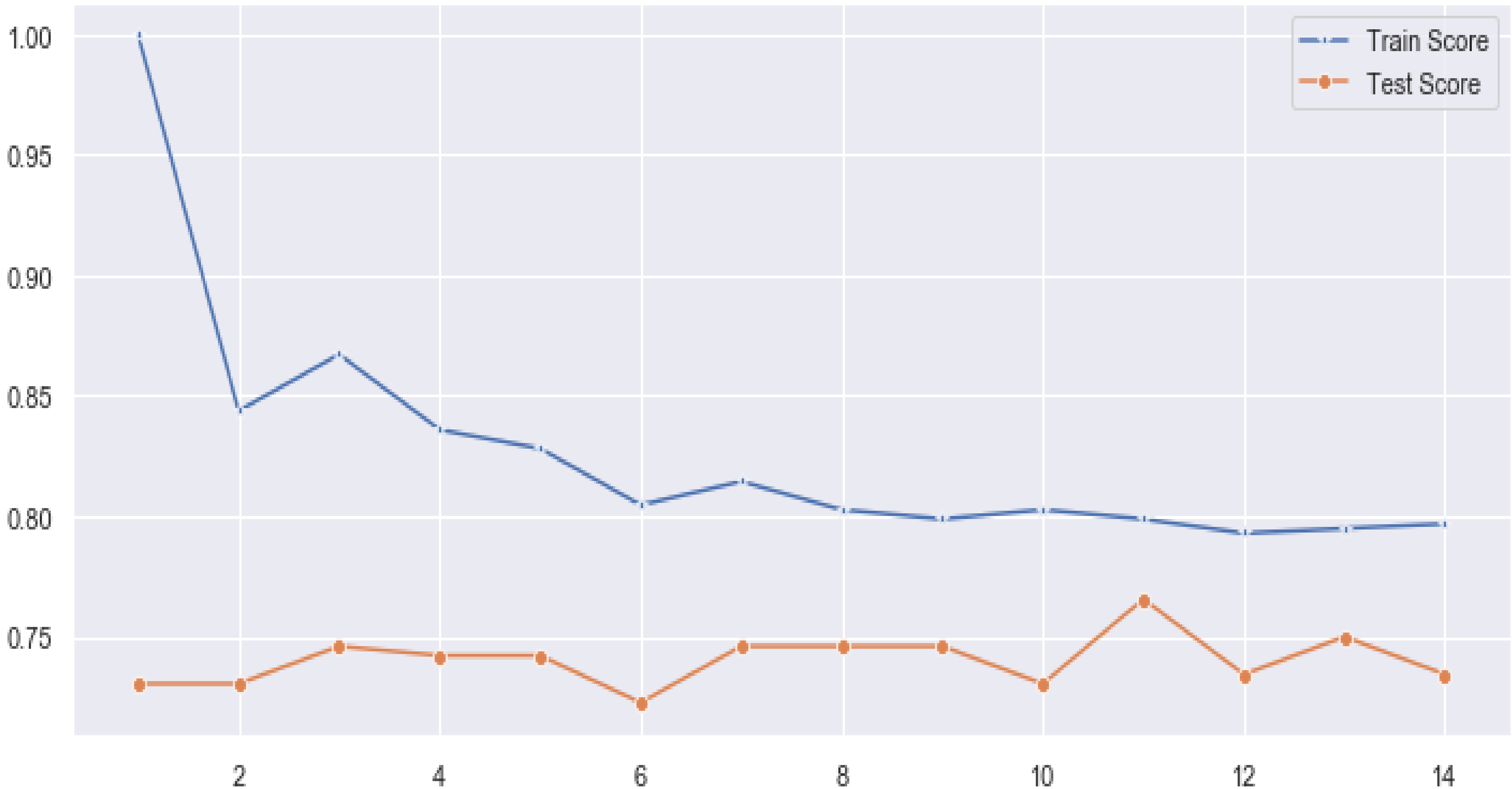


scatterplots.

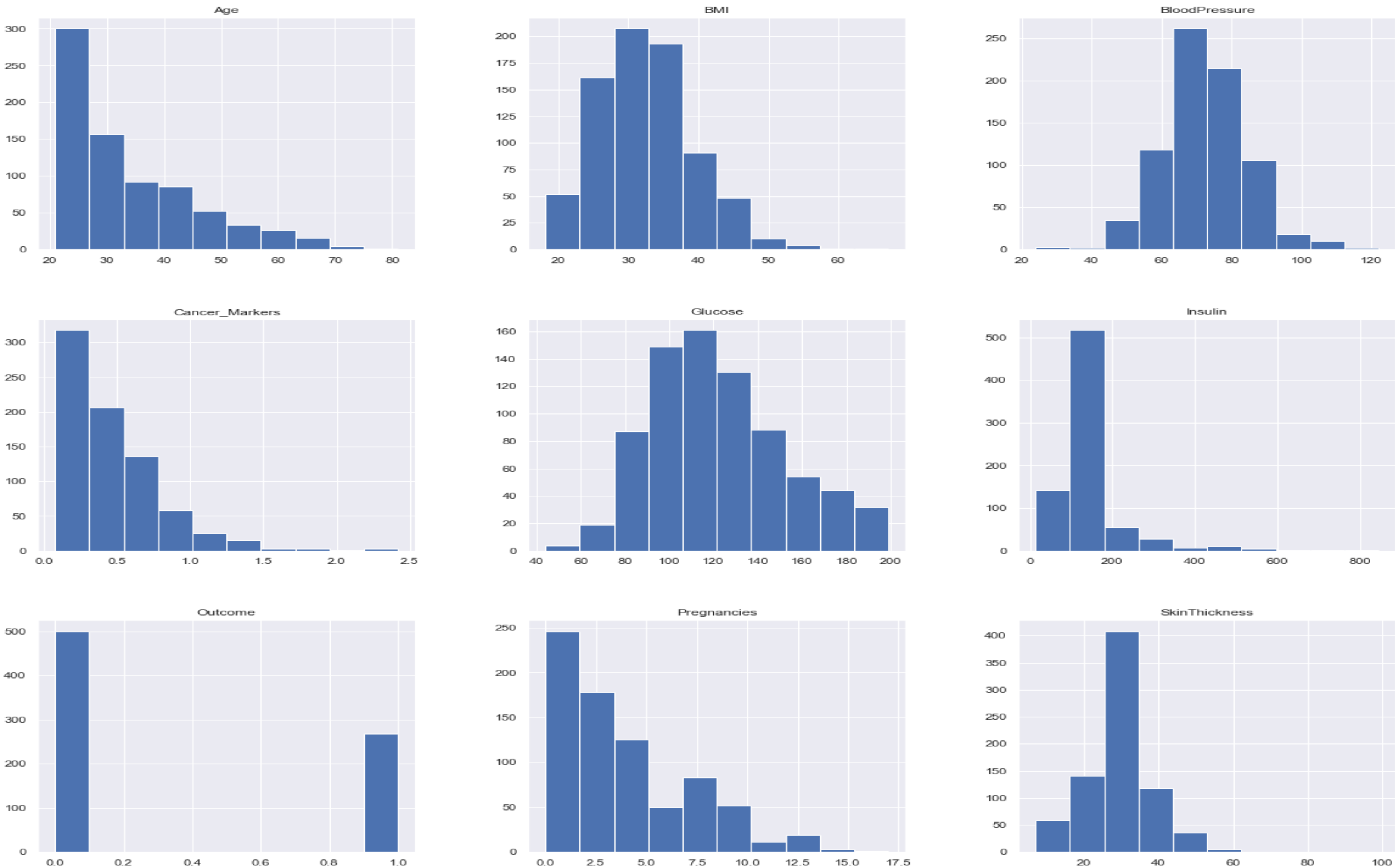


Pie Chart

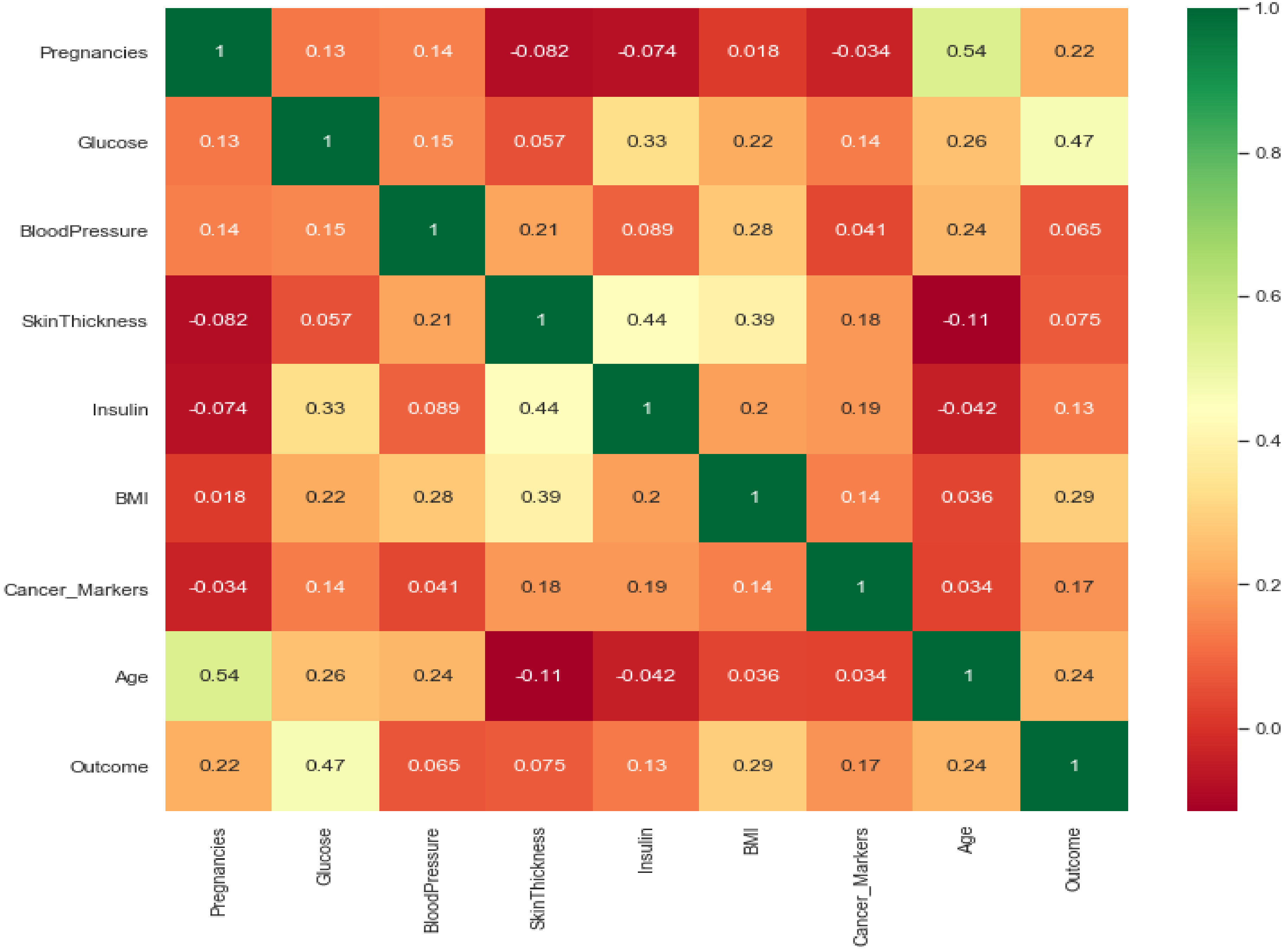




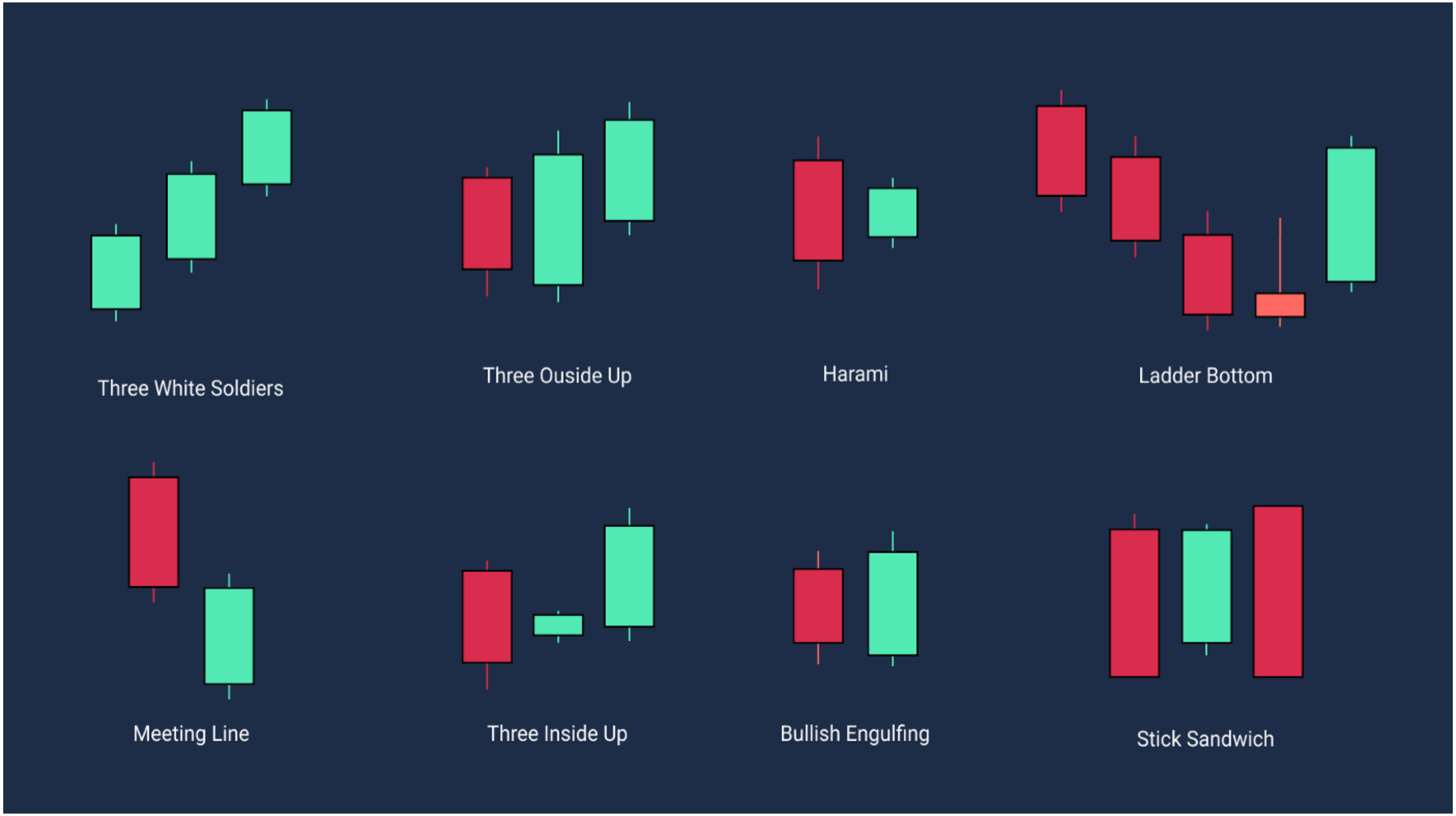
Train Score
Test Score
1.00
0.95
0.90
0.85
0.80
0.75
2
4
6
8
10
12
14

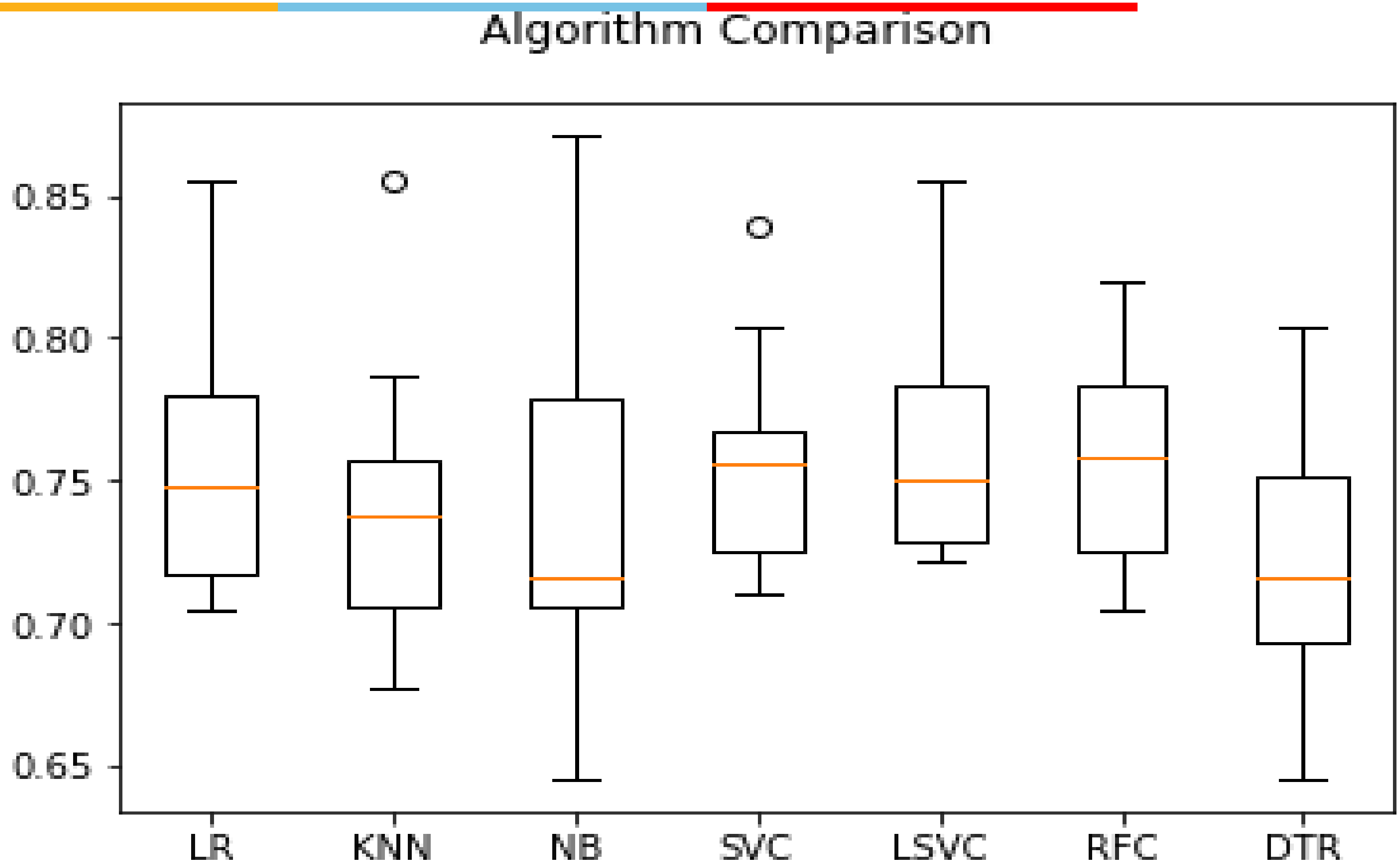

Age
BMI
BloodPressure
Cancer_Markers
Glucose
Insulin
Outcome
Pregnancies
SkinThickness


Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
Cancer_Markers
Age
Outcome
0
1


Pregnancies 1 0.13 0.14 -0.082 -0.074 0.018 -0.034 0.54 0.22
Glucose 0.13 1 0.15 0.057 0.33 0.22 0.14 0.26 0.47
BloodPressure 0.14 0.15 1 0.21 0.089 0.28 0.041 0.24 0.065
SkinThickness -0.082 0.057 0.21 1 0.44 0.39 0.18 -0.11 0.075
Insulin -0.074 0.33 0.089 0.44 1 0.2 0.19 -0.042 0.13
BMI 0.018 0.22 0.28 0.39 0.2 1 0.14 0.036 0.29
Cancer_Markers -0.034 0.14 0.041 0.18 0.19 0.14 1 0.034 0.17
Age 0.54 0.26 0.24 -0.11 -0.042 0.036 0.034 1 0.24
Outcome 0.22 0.47 0.065 0.075 0.13 0.29 0.17 0.24 1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
Cancer_Markers
Age
Outcome
1.0
0.8
0.6
0.4
0.2
0.0

u


Three White Soldiers
Three Ouside Up
Harami
Ladder Bottom
Meeting Line
Three Inside Up
Bullish Engulfing
Stick Sandwich


Algorithm Comparison
0.85
0.80
0.75
0.70
0.65
LR
KNN
NB
SVC
LSVC
RFC
DTR

Statistical Summary

Cost	Weight	Weight1	Length	Height	Width	
count	159.000000	159.000000	159.000000	159.000000	159.000000	159.000000
mean	398.326415	26.247170	28.415723	31.227044	8.970994	4.417486
std	357.978317	9.996441	10.716328	11.610246	4.286208	1.685804
min	0.000000	7.500000	8.400000	8.800000	1.728400	1.047600
25%	120.000000	19.050000	21.000000	23.150000	5.944800	3.385650
50%	273.000000	25.200000	27.300000	29.400000	7.786000	4.248500
75%	650.000000	32.700000	35.500000	39.650000	12.365900	5.584500
max	1650.000000	59.000000	63.400000	68.000000	18.957000	8.142000

Measures of Central Tendency
Measures of Variability

Measures of Central Tendency

- Measure of central tendency provides a very convenient way of describing a set of scores with a single number that describes the **PERFORMANCE** of the group.

- Also defined as a single value that is used to describe the "**center**" of the data.

- Three commonly used measures of central tendency:
  1. Mean
  2. Median
  3. Mode

Mean

- Also referred as the "**arithmetic average**"
- The most commonly used measure of the center of data
- Numbers that describe what is average or typical of the distribution
- Computation of Sample Mean:

$$\bar{Y} = \frac{\sum Y}{N}$$ = Σ Y/ N = (Y1 + Y2 + Y3 + … Yn) / N where

"Y bar" equals the sum of all the scores, Y, divided by the number of scores, N.

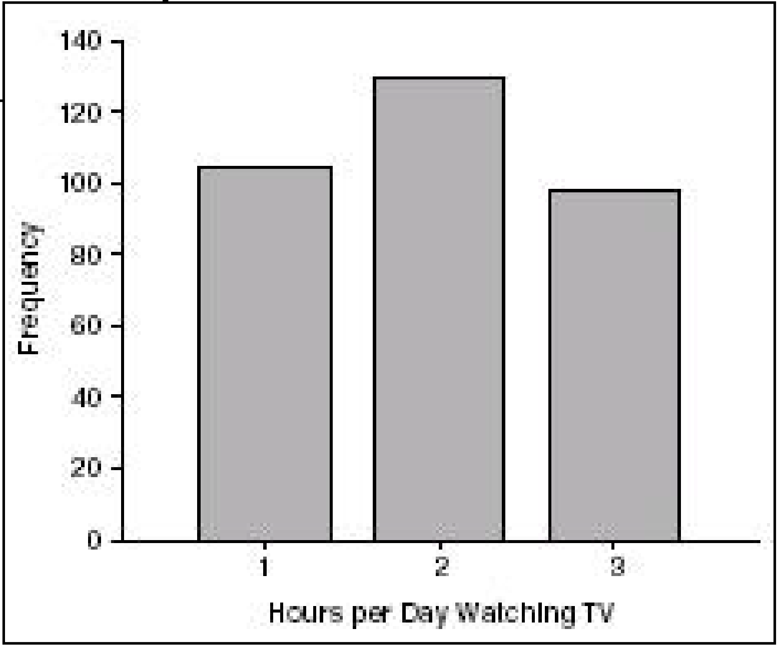
- Computation of the Mean for grouped Data

$$\bar{Y} = \frac{\sum f Y}{N}$$ Where f Y = a score multiplied by its frequency

Mean: Grouped Scores

Hours Spent Watching TV	Frequency (f)	fY	Percentage	C%
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$



Data of Children watching TV in Bengaluru

Mean

Properties

- It measures stability. Mean is the most stable among other measures of central tendency because every score contributes to the value of the mean.
- It may easily affected by the extreme scores.
- The sum of each score's distance from the mean is zero.
- It can be applied to interval level of measurement
- It may not be an actual score in the distribution
- It is very easy to compute.

Mean

When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness

The Mode

- The category or score with the largest frequency (or percentage) in the distribution.
- The mode can be calculated for variables with levels of measurement that are: nominal, ordinal, or interval-ratio.

Example:

- Number of Votes for Candidates for Lok Sabha MP. The mode, in this case, gives you the "central" response of the voters: the most popular candidate.

  - Candidate A – 11,769 votes
  - Candidate B – 39,443 votes
  - Candidate C – 78,331 votes

The Mode: "Candidate C"

Mode

Properties

- It can be used when the data are qualitative as well as quantitative.
- It may not be unique.
- It is affected by extreme values.
- It may not exist.

When to Use the Median

  - When the “typical” value is desired.
  - When the data set is measured on a nominal scale

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.

- The median is the middle score, or average of middle scores in a distribution.
  - Fifty percent (50%) lies below the median value and 50% lies above the median value.
  - It is also known as the middle score or the 50th percentile.

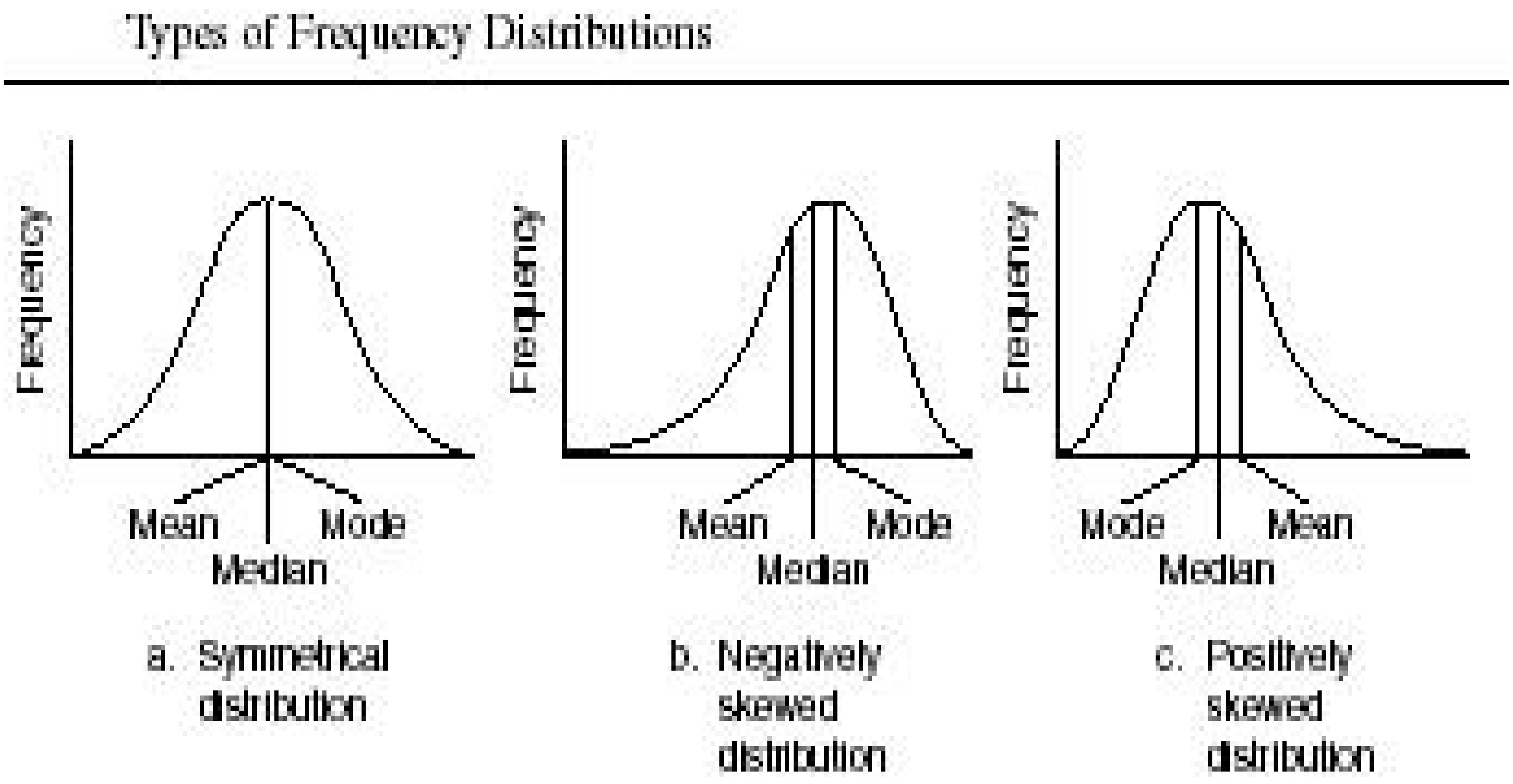
Measures of central tendency

- The mean
- the median
- the mode

Shape of the Distribution

- Symmetrical : mean is about equal to median
- Skewed
  - Negatively :   mean < median
  - Positively   : mean > median
- Bimodal : has two distinct modes
- Multi-modal : has more than 2 distinct modes)

Distribution Shape



Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

Statistical measures	Group 2
Mean	5
Median	5
Mode	5

Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Statistical measures	Group 1 & 2
Mean	5
Median	5
Mode	5

Do we need any other measure?

Answer: Yes

Measures of variability

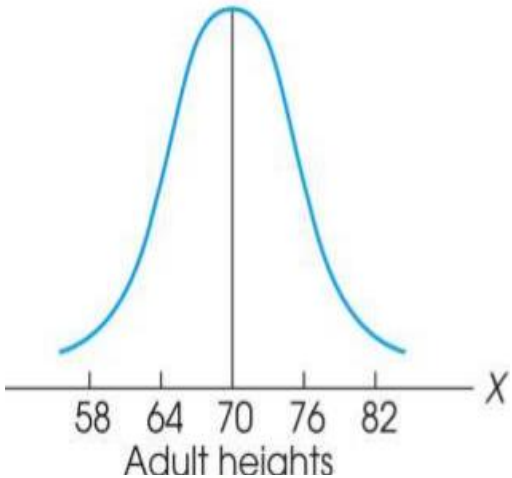
Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

Measure of Variability

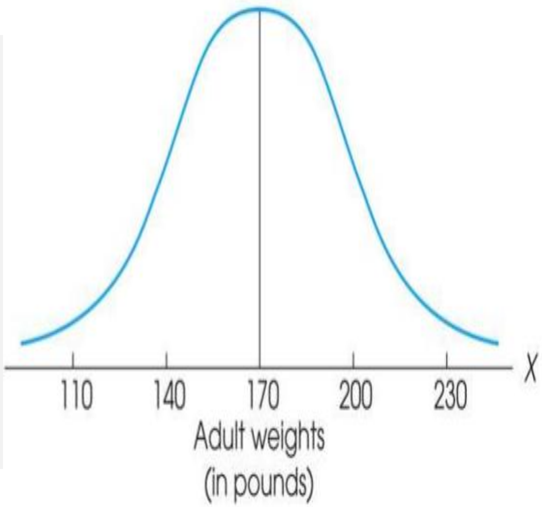
Variability can be defined several ways:

- A quantitative distance measure based on the differences between scores
- Describes distance of the spread of scores or distance of a score from the mean



Purposes of Measure of Variability:

- Describe the distribution
- Measure how well an individual score represents the distribution



The Three Measures

Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

The Ranges

- The distance covered by the scores in a distribution – From smallest value to highest value
- For continuous data, real limits are used

$$\text{Range} = \text{URL for } X_{max} - \text{LRL for } X_{min}$$

- Based on two scores, not all the data – An imprecise, unreliable measure of variability

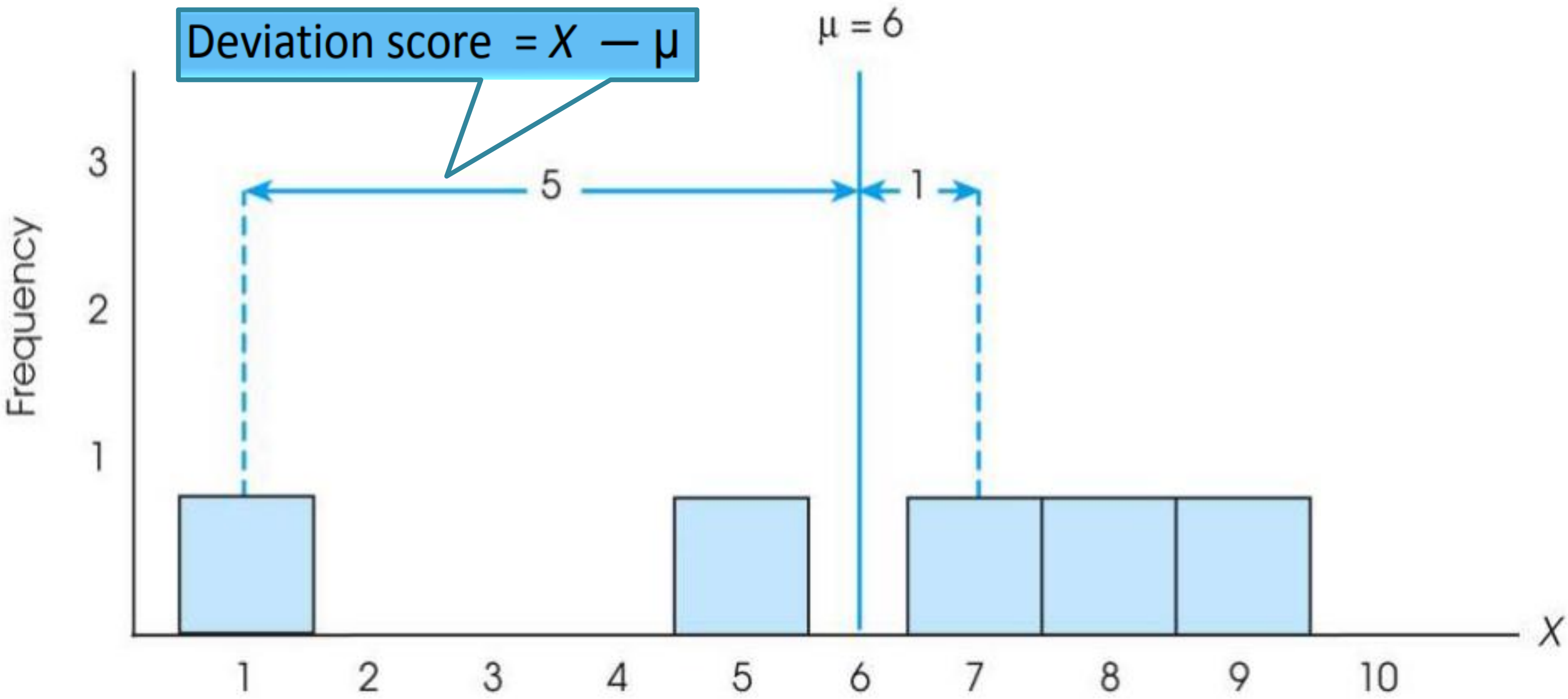
Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

Range = Highest Score minus Lowest score = 7 - 2 = 5

The Standard Deviation

- Most common and most important measure of variability is the standard deviation
  - A measure of the standard, or average, distance from the mean
  - Describes whether the scores are clustered closely around the mean or are widely scattered
- Calculation differs for population and samples
- Variance is a necessary *companion concept* to standard deviation but *not the same* concept

The Standard Deviation



Exercise : Find out the deviations of all the data points with the mean....and then find the 'mean deviation'.

The Standard Deviation

- Mean deviations will always be ‘zero’ !
  (because Mean is a balance point)

Then, how do you find ‘Standard Deviation’ ?

Need a new strategy

The Standard Deviation

New Strategy :

a) First square each deviation score
b) Then sum the Squared Deviations (SS)
c) Average the squared deviations

- Mean Squared Deviation is known as **"Variance"**
- Variability is now measured in squared units

$$Standard\ Deviation = \sqrt{Variance}$$

The Variance

Variance equals mean (average) squared deviation (distance) of the scores from the mean

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

where $$SS = \sum (X - \mu)^2$$

The Population Variance

- Population variance equals mean (average) squared deviation (distance) of the scores from the population mean

- Variance is the average of squared deviations, so we identify population variance with a lowercase Greek letter sigma squared: σ 2

- Standard deviation is the square root of the variance, so we identify it with a lowercase Greek letter sigma: σ

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum\left(x - \overline{x}\right)^2}{n-1}}$$

$$S = \sqrt{\frac{44}{8}} = 2.345$$

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum (x - \overline{x})^2}{n-1}}$$

$$S = \sqrt{\frac{134}{8}} = 4.093$$

Learning Check

a) If all the scores in a data set are the same, the Standard Deviation is equal to 1.00

True / False ?

Select the correct option

b) The standard deviation measures …
   (1) Sum of squared deviation scores
   (2) Standard distance of a score from the mean
   (3) Average deviation of a score from the mean
   (4) Average squared distance of a score from the mean

Solution

a) If all the scores in a data set are the same, they are equal to the mean and hence the deviation from mean = 0 therefore, Standard Deviation is equal to **zero**

False

b) The standard deviation measures …

(1) Sum of squared deviation scores

(2) Standard distance of a score from the mean

(3) Average deviation of a score from the mean

(4) Average squared distance of a score from the mean

Standard Deviation and Variance for a Sample

- Goal of inferential statistics:
  - Draw general conclusions about population
  -
  - Based on limited information from a sample

- Samples differ from the population
  - Samples have less variability

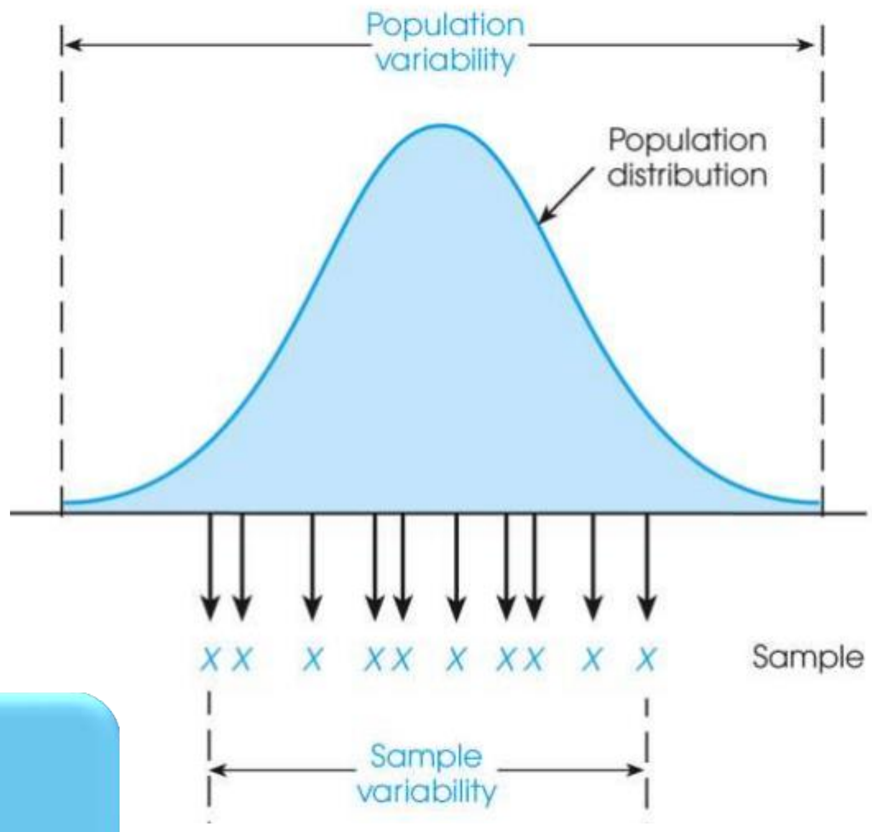
  - Computing the Variance and Standard Deviation in the same way as for a population would give a biased estimate of the population values

Sample Standard Deviation and Variance

- Sum of Squares (SS) is computed as before
- Formula for Variance has n-1 rather than N in the denominator
- Notation uses s instead of σ

$$variance\ of\ sample = s^2 = \frac{SS}{n-1}$$

$$standard\ deviation\ of\ sample = s = \sqrt{\frac{SS}{n-1}}$$



Population of Adult Heights

Degrees of Freedom

- Population variance
  - Mean is known
  - Deviations are computed from a known mean

- Sample variance as estimate of population
  - Population mean is unknown
  - Using sample mean restricts variability

- Degrees of freedom
  - Number of scores in sample that are independent and free to vary
  - Degrees of freedom (df) = n – 1

Learning Check

Select the correct option

a) A sample of four scores has SS = 24. What is the variance?
   (1) The variance is 6
   (2) The variance is 7
   (3) The variance is 8
   (4) The variance is 12

b) A sample systematically has less variability than a population

True / False ?

c) The standard deviation is the distance from the Mean to the farthest point on the distribution curve

True / False ?

Solution

Select the correct option

a) A sample of four scores has SS = 24. What is the variance?
   (1) The variance is 6
   (2) The variance is 7
   (3) The variance is 8
   (4) The variance is 12

b) Extreme scores affect variability, but are less likely to be included in a sample

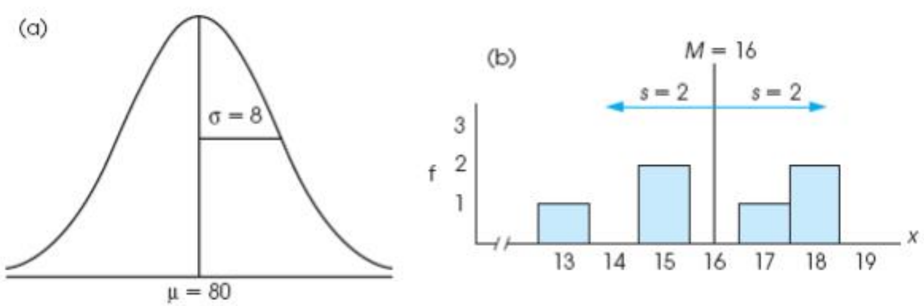
c) The standard deviation extends from the mean approximately halfway to the most extreme score

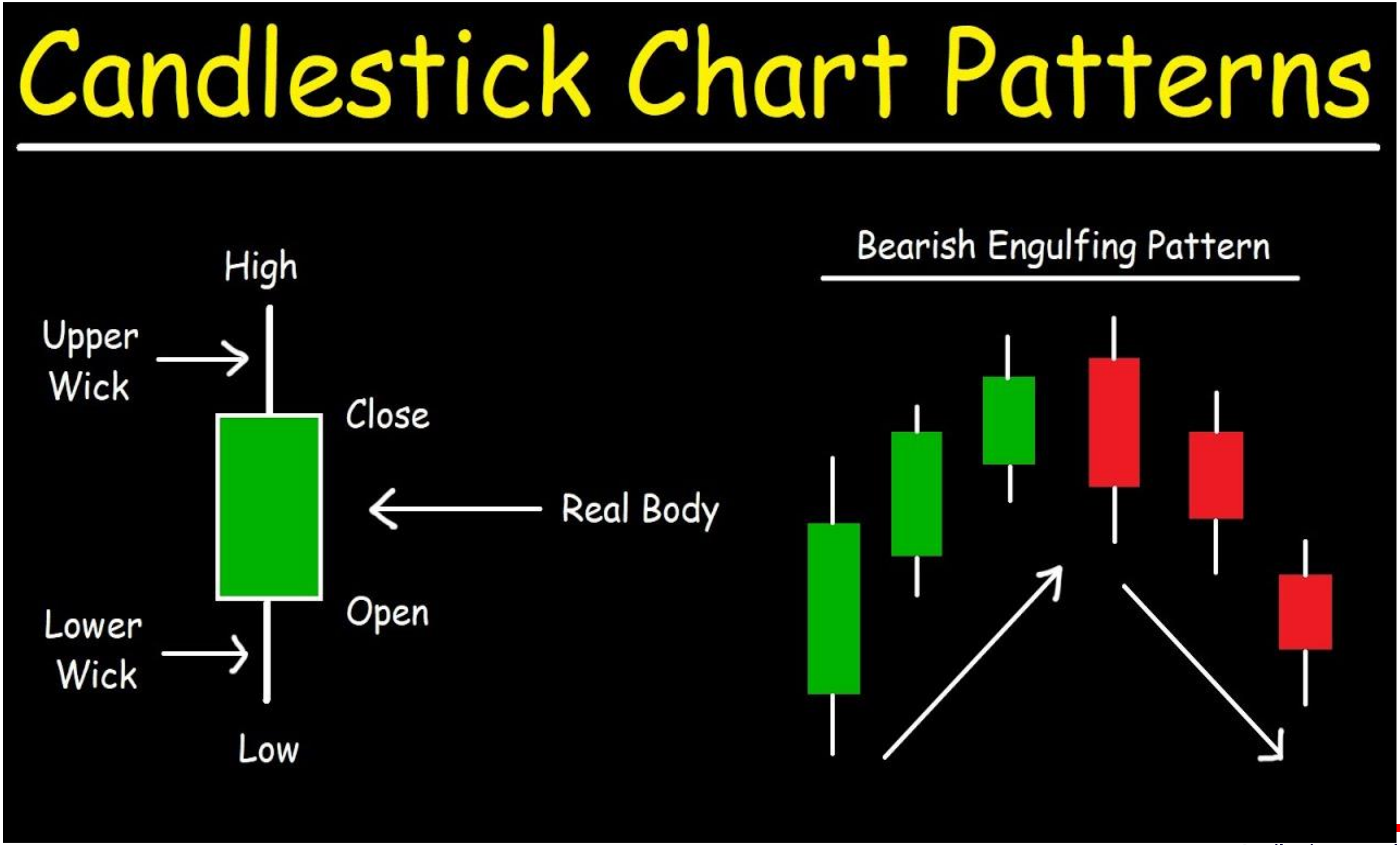
True

False

Descriptive Statistics

- A standard deviation describes scores in terms of distance from the mean
- Describe an entire distribution with just two numbers (M and s)
- Reference to both allows reconstruction of the measurement scale from just these two numbers
- Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions




Candlestick Chart Patterns
High
Upper Wick
Close
Real Body
Open
Lower Wick
Low
Bearish Engulfing Pattern

Interquartile range (IQR)

- Measure of Variation
- Also Known as Midspread: Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values

$$\text{Interquartile Range} = Q_3 - Q_1$$

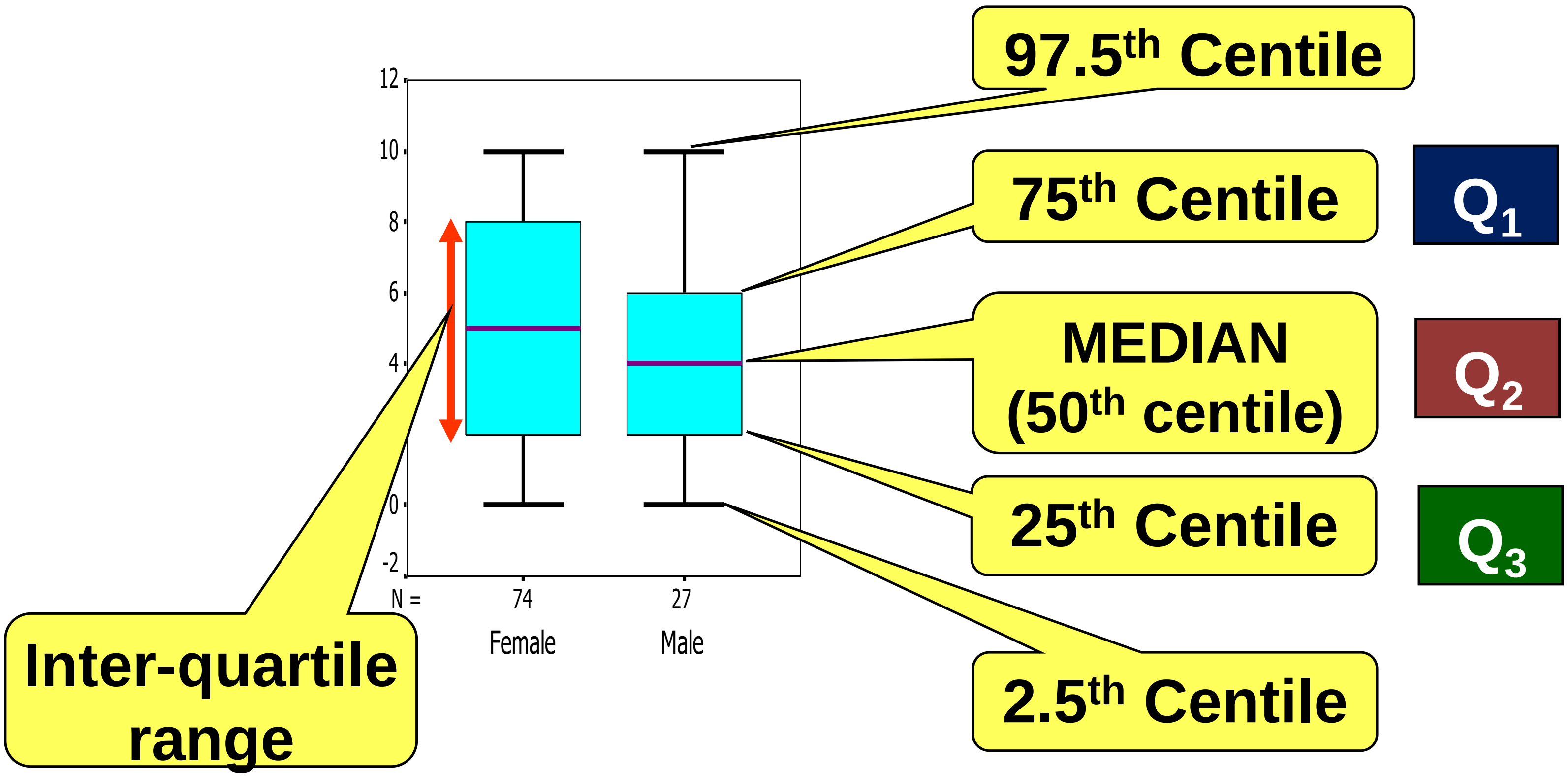
Data in Ordered Array: 11 12 13 16 16 17 17 18 21

$$\text{Position of } Q_1 = \frac{1 \cdot (9+1)}{4} = 2.50, \qquad Q_1 = 12.5$$

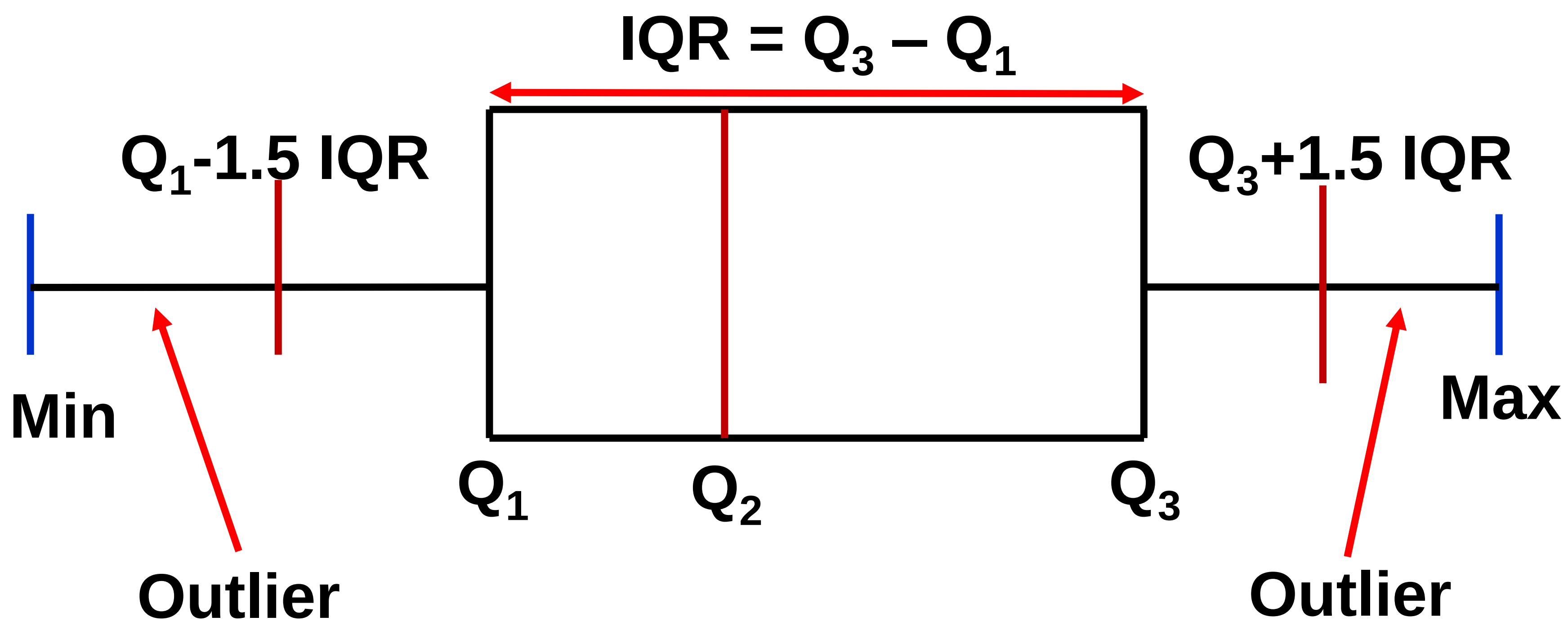
$$\text{Position of } Q_3 = \frac{3 \cdot (9+1)}{4} = 7.50, \qquad Q_3 = 17.5$$

$$\text{Interquartile Range} = Q_3 - Q_1 = 17.5 - 12.5 = 5$$

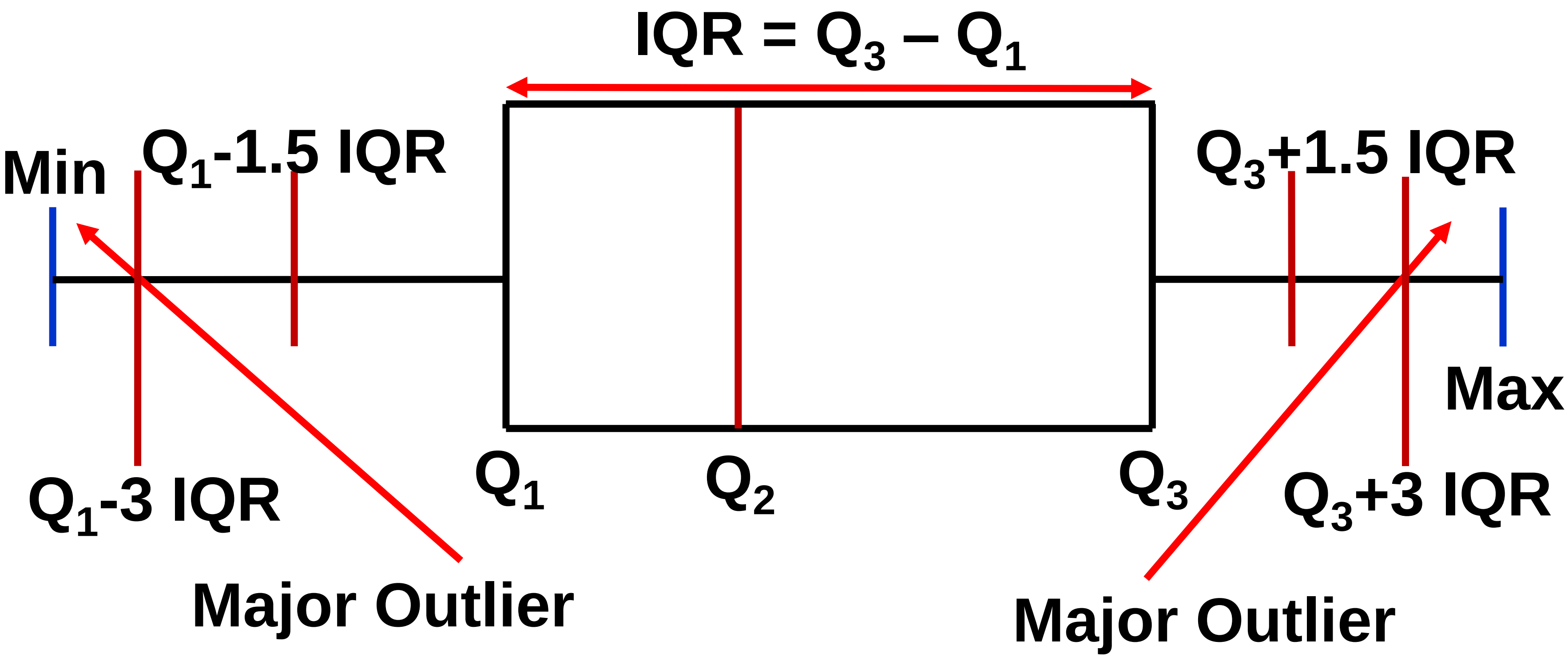
Box and Whisker plot



Box and Whisker plot



Box-and-Whisker plot



Thanks