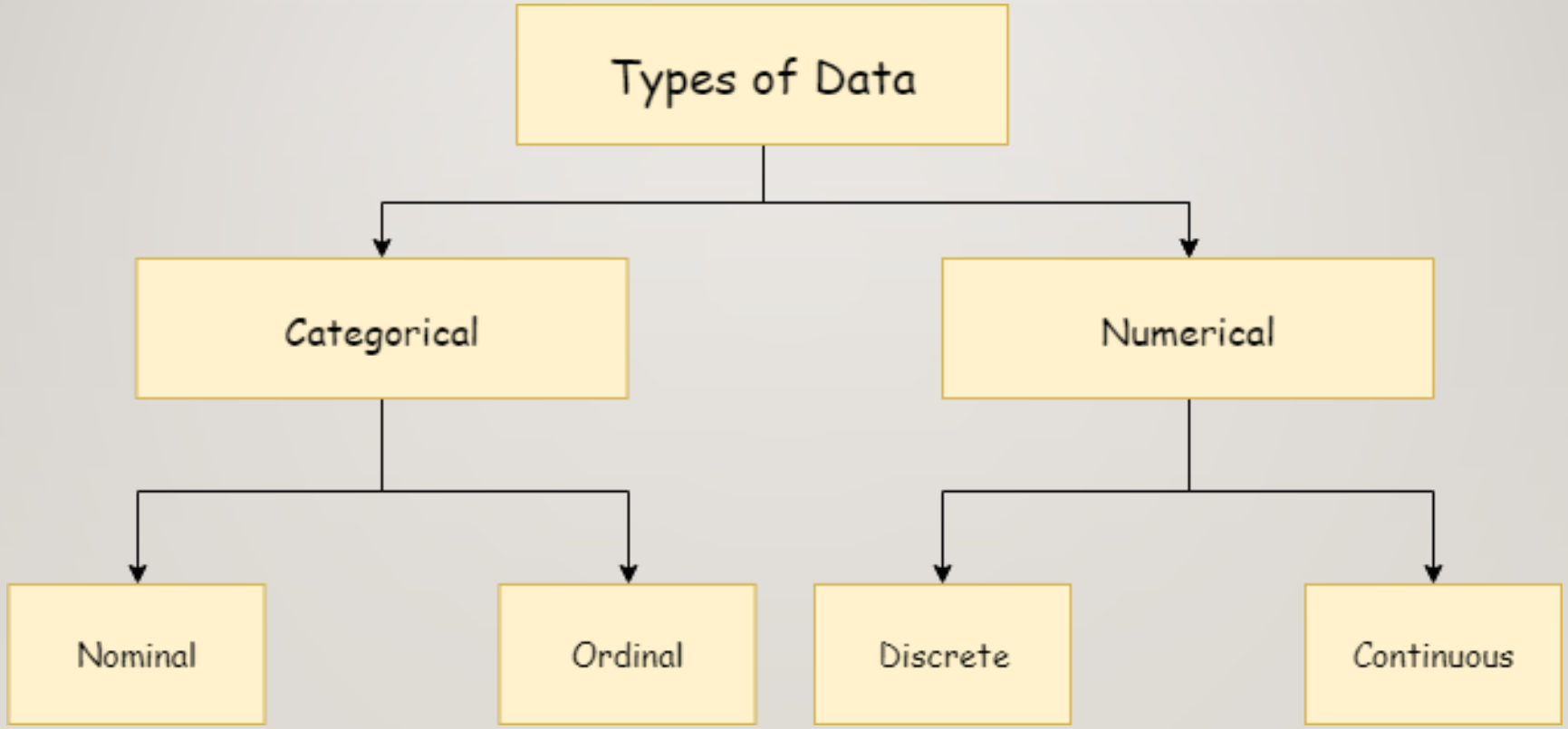
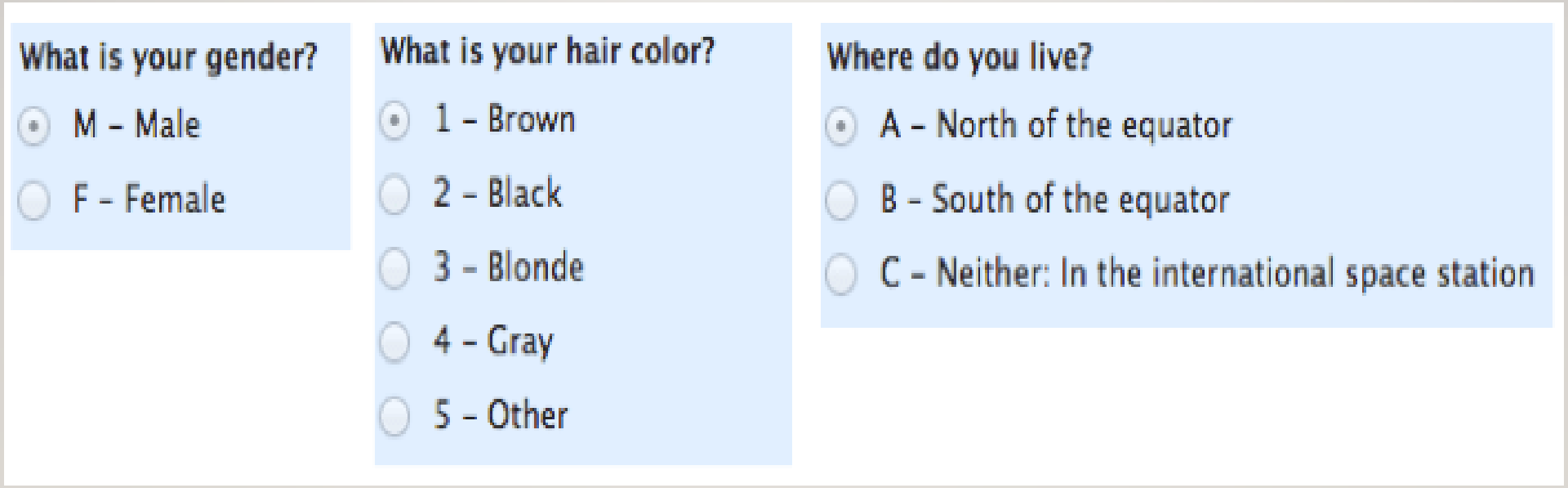
RECAP ON DESCRIPTIVE STATISTICS

LAXMINARAYEN

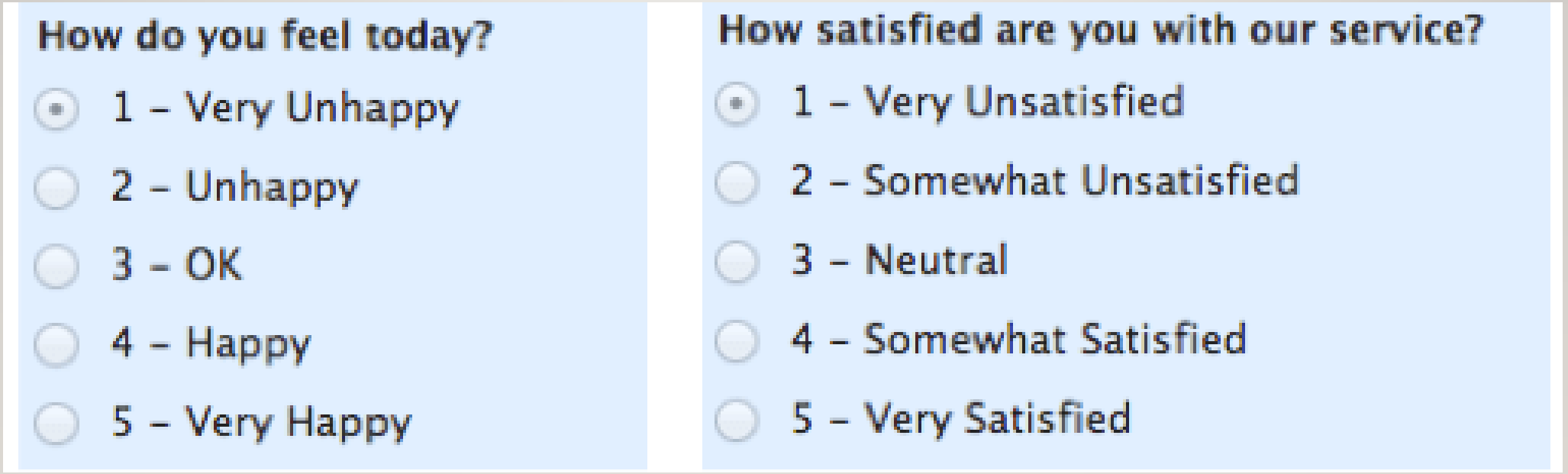
TYPES OF DATA



NOMINAL DATA

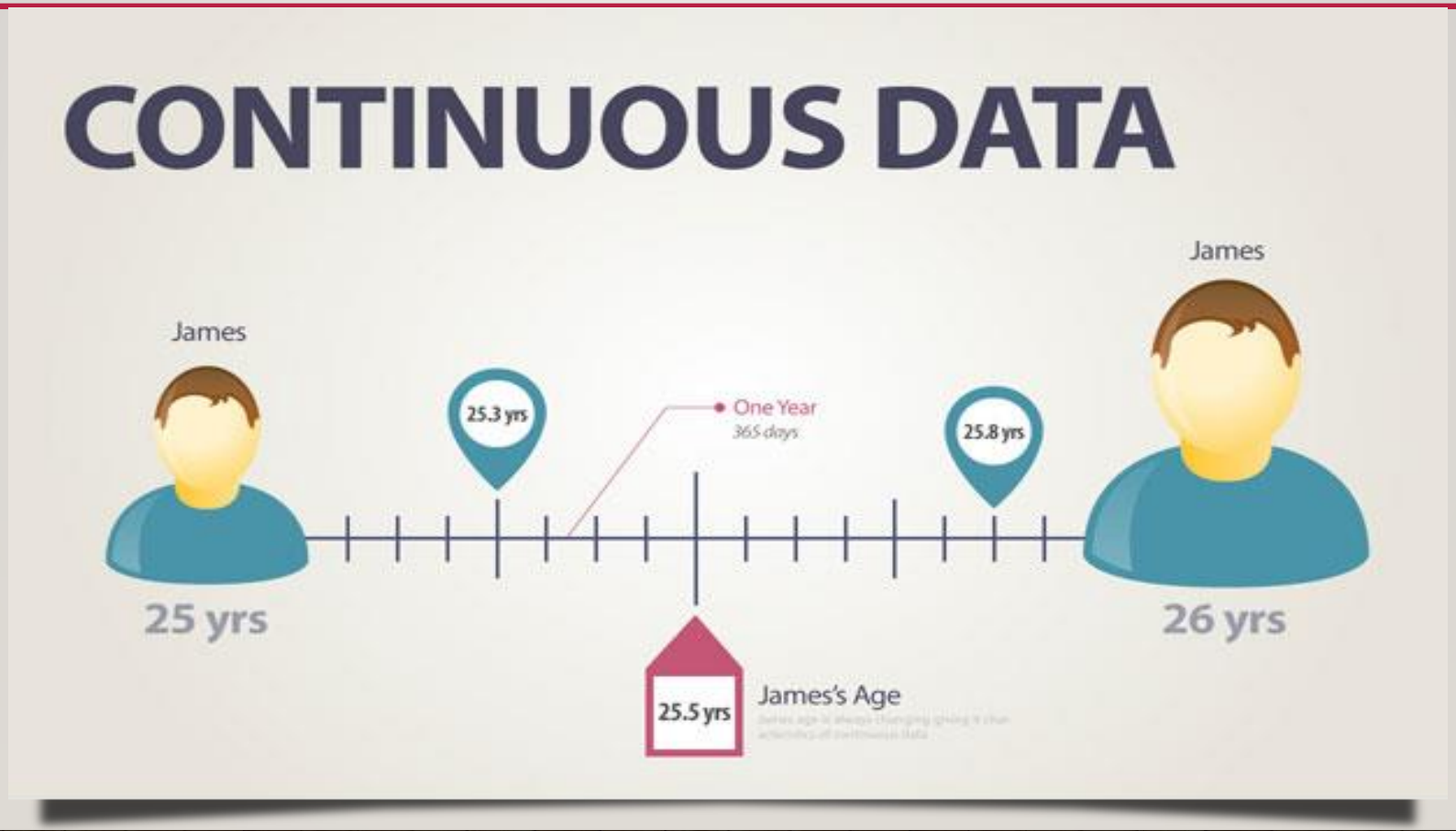


ORDINAL DATA



DISCRETE DATA

CONTINUOUS DATA



FREQUENCY

- **Frequency** is how often something occurs.

FREQUENCY EXAMPLE

- Example: Sam played football on:
- Saturday Morning,
- Saturday Afternoon
- Thursday Afternoon

FREQUENCY EXAMPLE

- Example: Sam played football on:
- Saturday Morning,
- Saturday Afternoon
- Thursday Afternoon
- The frequency was 2 on Saturday, 1 on Thursday and 3 for the whole week.

FREQUENCY DISTRIBUTION

- By counting frequencies we can make a **Frequency Distribution** table.

FREQUENCY DISTRIBUTION EXAMPLE

- Sam's team has scored the following numbers of goals in recent games
- 2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

FREQUENCY DISTRIBUTION EXAMPLE

- Sam's team has scored the following numbers of goals in recent games
- 2, 3, 1, 2, 1, 3, 2, 3, 4, 5, 4, 2, 2, 3

Scores:
1,1,2,2,2,2,2,3,3,3,3,4,4,5

Score	Frequency
1	2
2	5
3	4
4	2
5	1

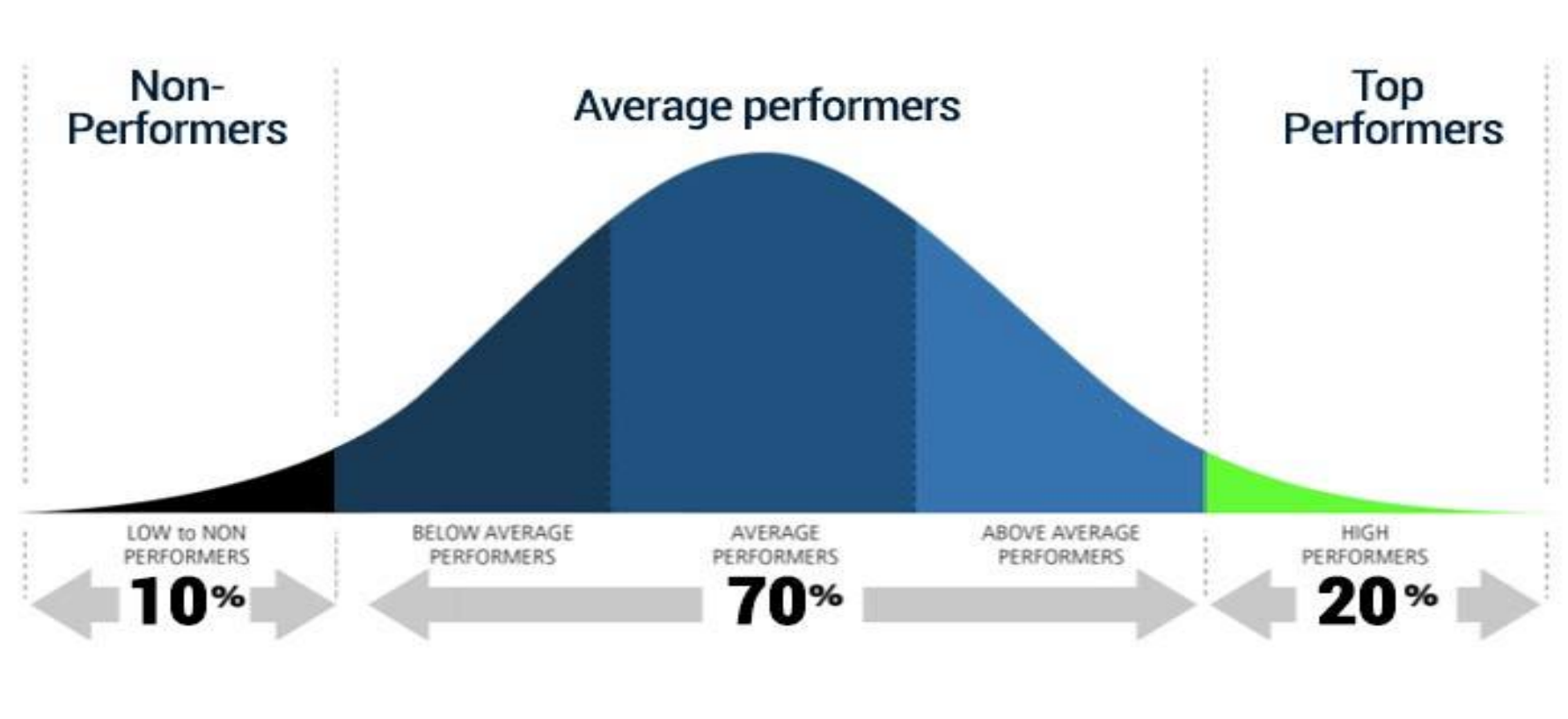
NORMAL DISTRIBUTION

- Many characteristics in this world are distributed throught in a '**normal**' manner

- They have **well defined statistical properties**

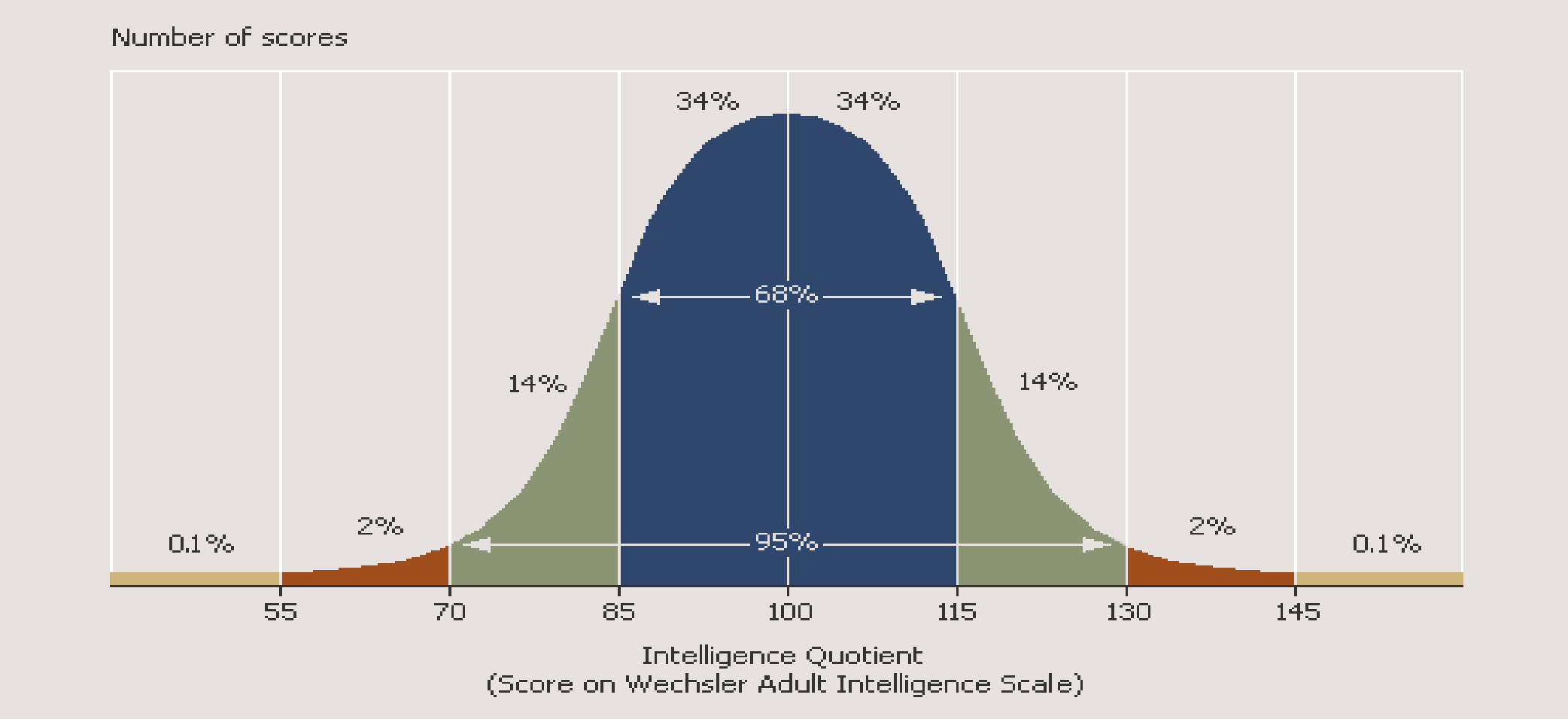
EXAMPLE FOR NORMAL DISTRIBUTION

- **Use Of Bell Curve In Performance Appraisals – Good Or Bad?**



EXAMPLE OF NORMAL DISTRIBUTION

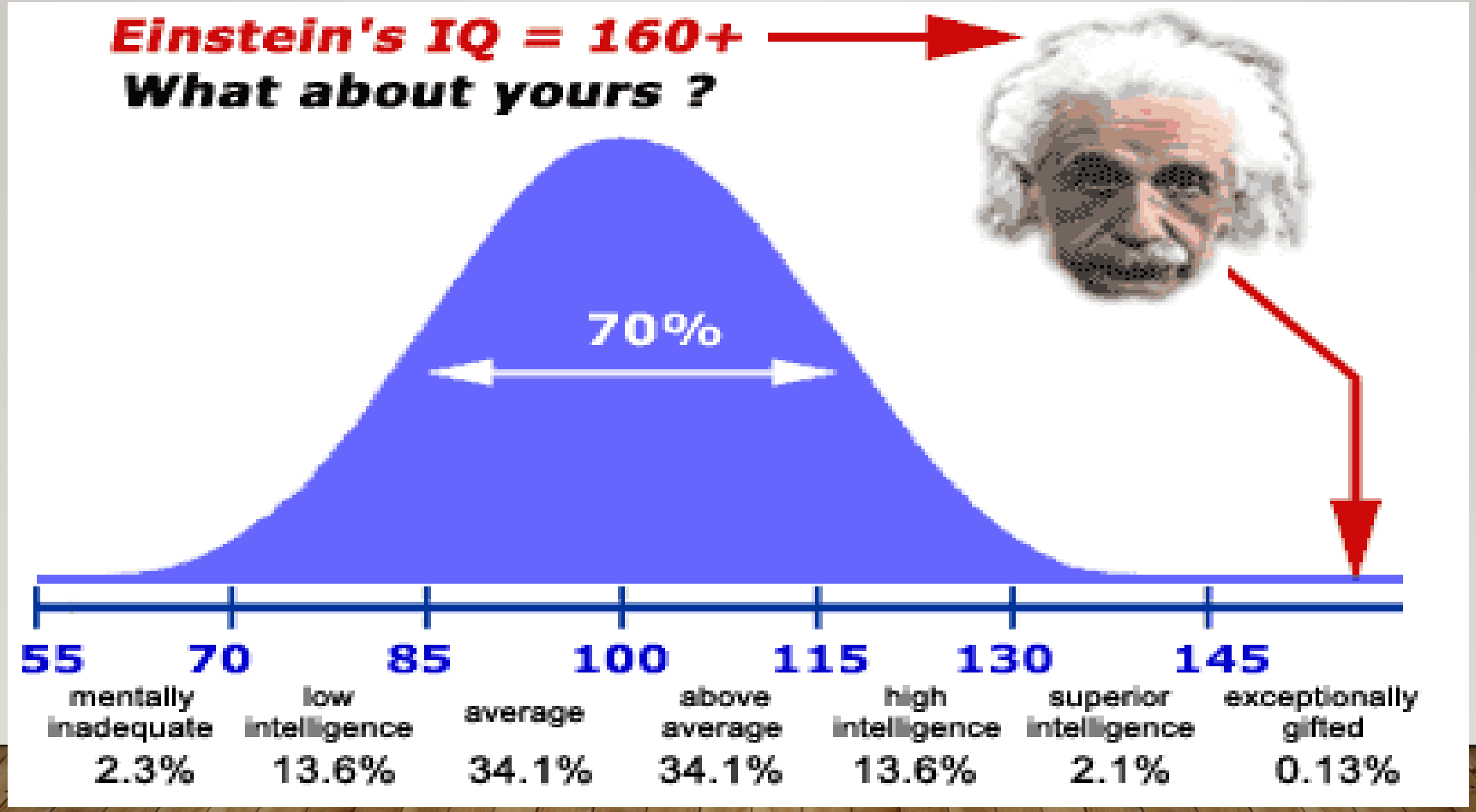
- **I.Q. distribution**



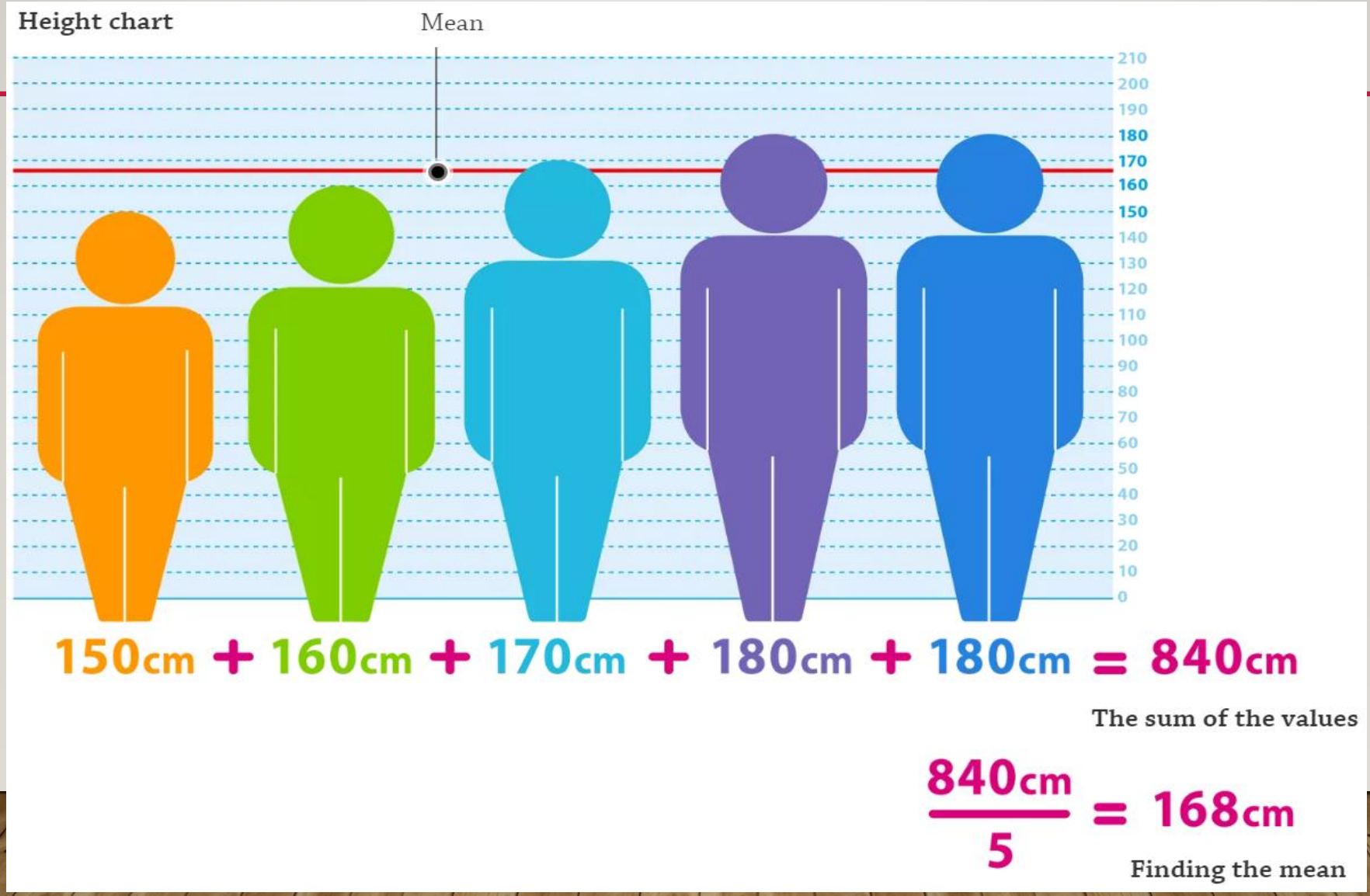
MEASURES OF CENTRAL TENDENCY

- **MEAN**
- **MEDIAN**
- **MODE**

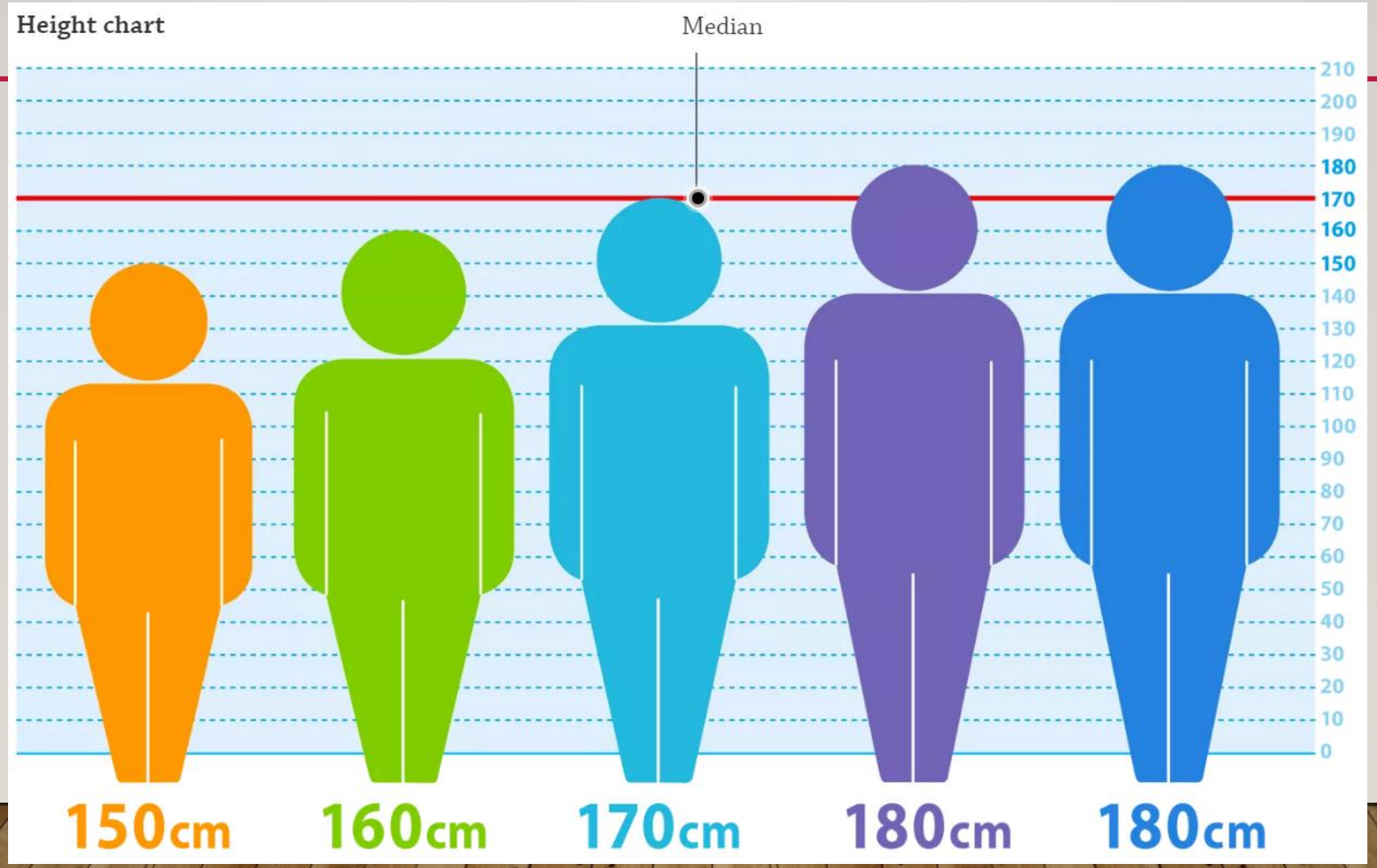
EXAMPLE FOR MEAN



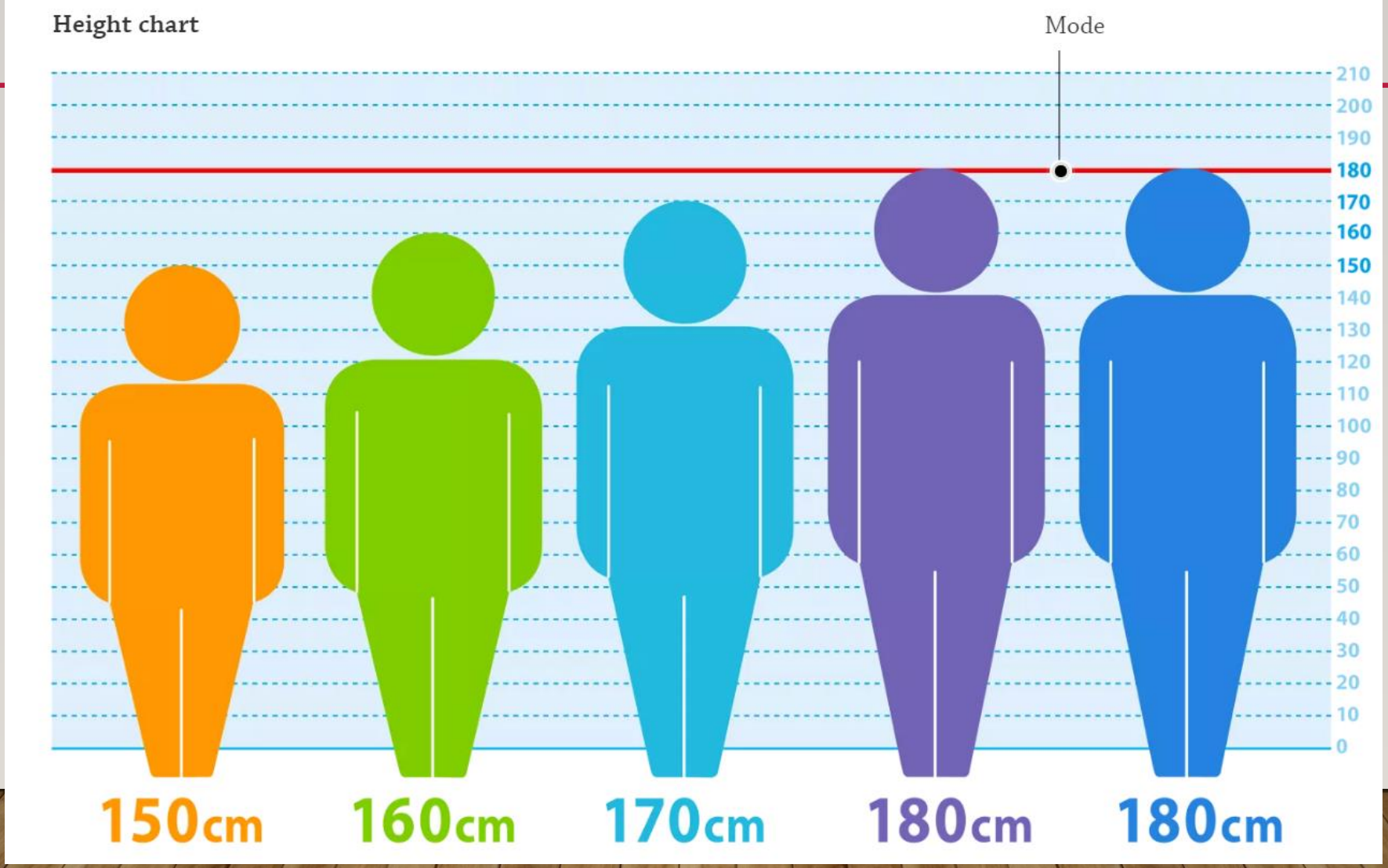
EXAMPLE FOR MEAN



EXAMPLE FOR MEDIAN



EXAMPLE FOR MODE



WHY DID WE LOOSE??

Players: CSK	Out/Not	Runs	Balls	Avg. runs
FaF du plessis	Out	6	11	32.4
Shane Watson	Out	10	13	39.64
Suresh Raina	Out	5	7	37.08
Murali Vijay	Out	26	26	12
Ambati Rayudu	Not Out	42	37	43
MS Dhoni	Not Out	37	29	75.83
TOTAL		131/4		

WHY DID WE LOOSE??

Players: CSK	Out/Not	Runs	Balls	Avg. runs
FaF du plessis	Out	6	11	32.4
Shane Watson	Out	10	13	39.64
Suresh Raina	Out	5	7	37.08
Murali Vijay	Out	26	26	12
Ambati Rayudu	Not Out	42	37	43
MS Dhoni	Not Out	37	29	75.83
TOTAL		131/4		

1. **Low Score**
2. **Why Low score? Because most of the players gave a below average performance**

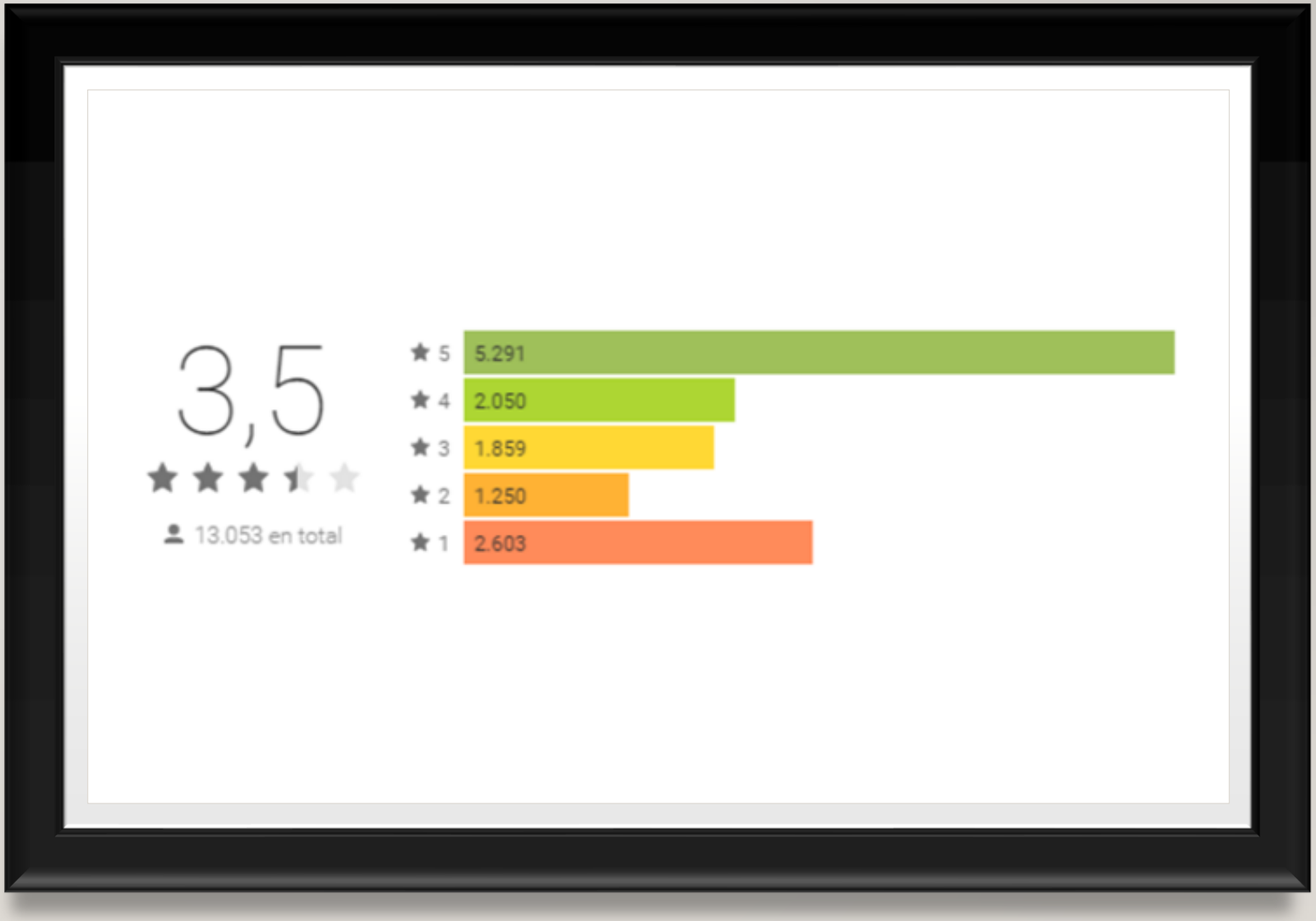
WHICH ONE IS COOL?

- Let us take another example to explain Mean.

Which one would you download??

WHICH ONE IS COOL?

IN THE BELOW EXAMPLE IF YOU ADD HERE ALL THE RATING GIVEN BY USERS IT WILL COME TO 45685 YOU DIVIDE IT BY 13053 YOU WILL GET 3.5.



AVERAGE INDIVIDUAL SALARY

- We take a set of Employees from a Company
- Set = { Junior employee, Assistant Manager, Director}
- Salary = { 23,000, 50,000, 3,00,000}

AVERAGE INDIVIDUAL SALARY

- We take a set of Employees from a Company
- Set = { Junior employee, Assistant Manager, Director}
- Salary = { 23,000, 50,000, 3,00,000}
- Mean = 124,333.333

No WAY!!!!!!!!

AVERAGE INDIVIDUAL SALARY

- We take a set of Employees from a Company
- Set = { Junior employee, Assistant Manager, Director}
- Salary = { 23,000, 50,000, 3,00,000}

MEDIAN = 50,000 YES!!!!!!!

MEDIAN EXAMPLE

- Let's say you are reading in a class and one of them is the son of Bill Gates, the students of the class get the following pocket money

Students	Monthly Pocket Money in $
1	50
2	60
3	70
4	80
5	90
6	100
7	110
8	120
9	130
10	140
11	2350

MEDIAN EXAMPLE

- Here total Students = 11, and the total amount of pocket money = 3300$

- The mean of the Pocket money = 3300/11 = 300$
- You see nobody other than the son of bill gates get that much
- So we go with median = 100$

EXAMPLE FOR MODE

- Deciding the salary for the new role opening.
- Salary for the same role given by 4 of your competitors

Company A: Rs. 15,000

Company B: Rs. 23,000

Company C: Rs. 23,000

Company D: Rs. 40,000

EXAMPLE FOR MODE

- Deciding the salary for the new role opening.
- Salary for the same role given by 4 of your competitors

Company A: Rs. 15,000

Company B: Rs. 23,000

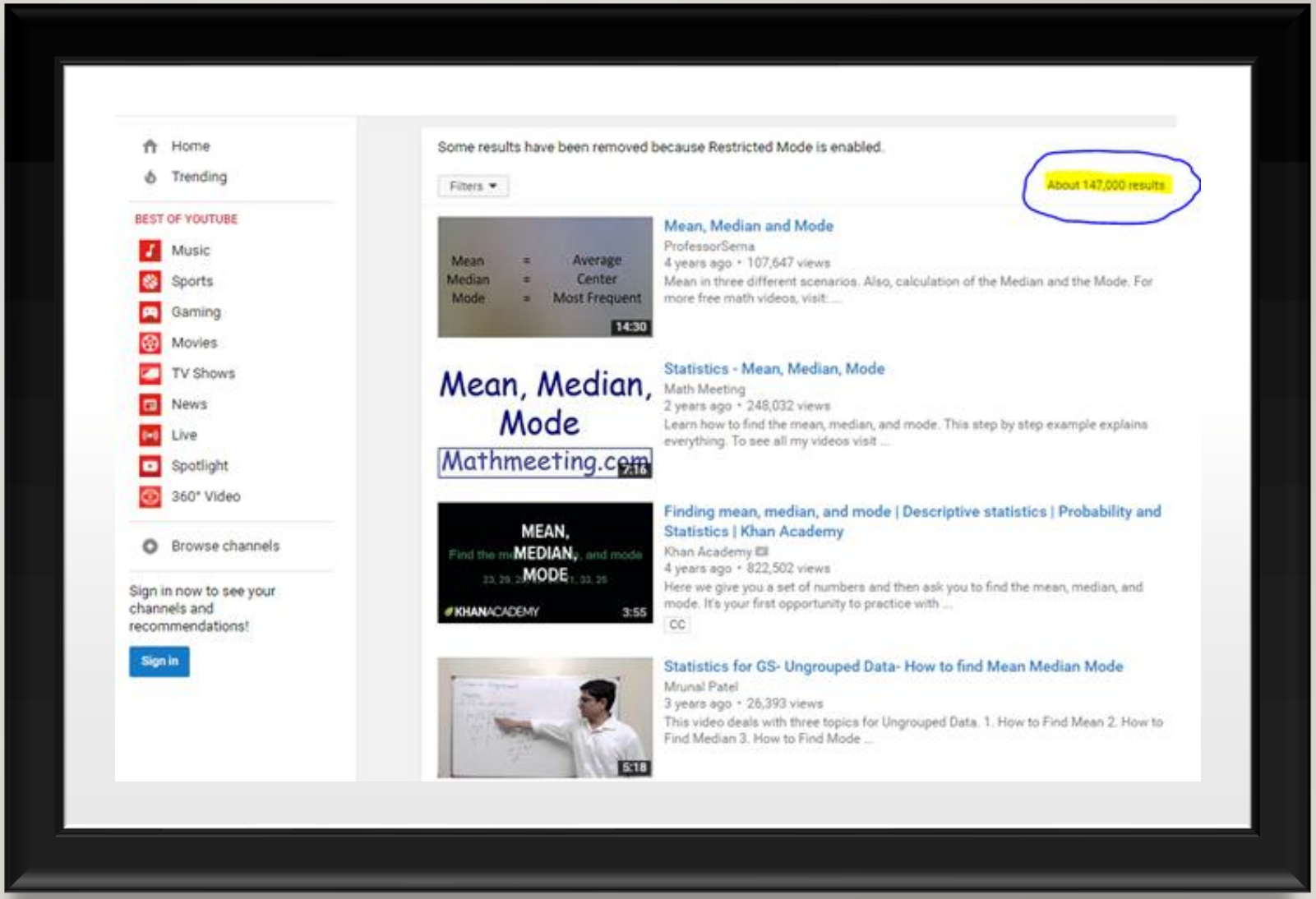
Company C: Rs. 23,000

SO YOU DON'T NEED TO OVERPAY OR UNDERPAY

Company D: Rs. 40,000

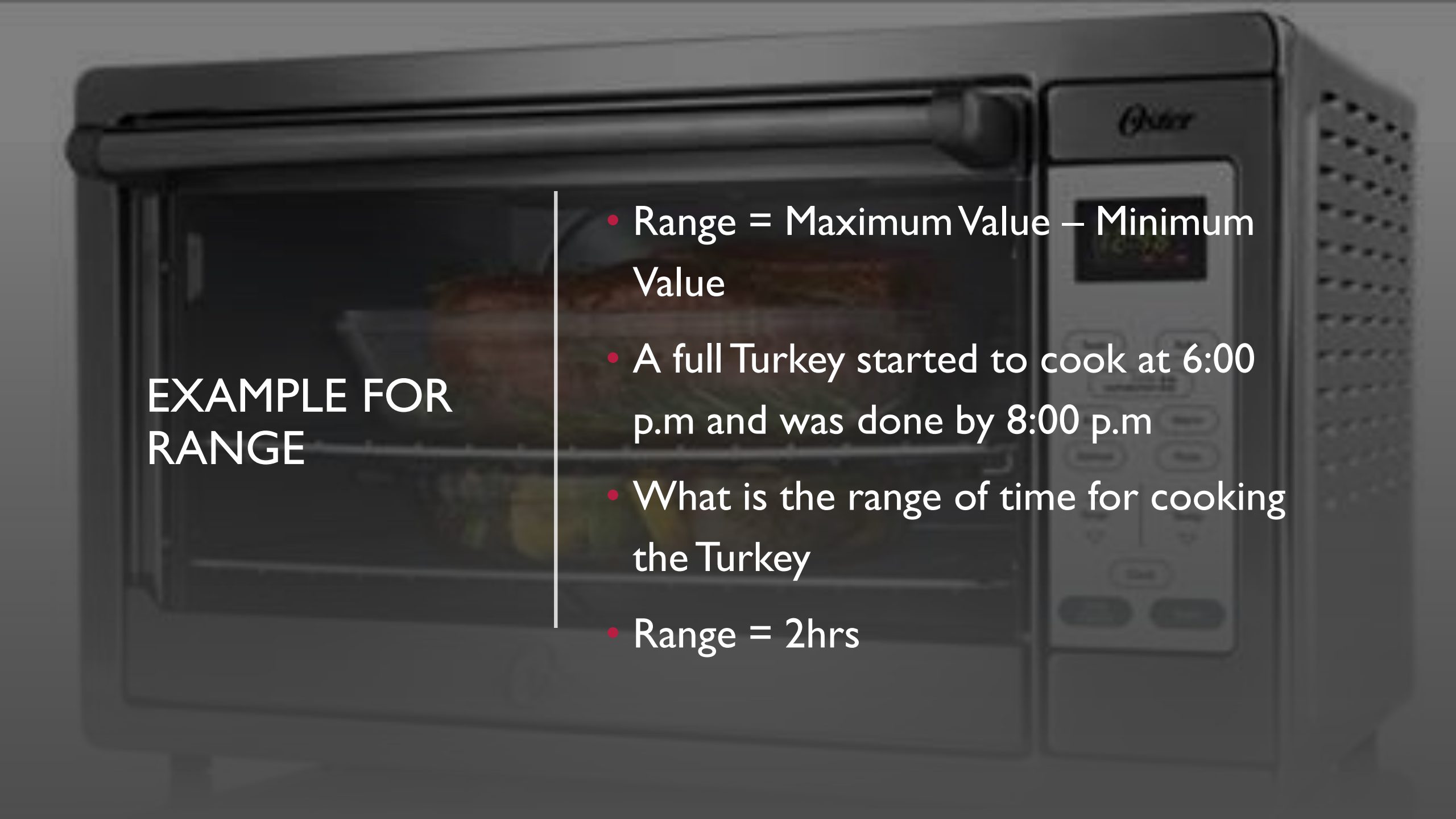
Company E: Rs. 30,000

EXAMPLE FOR MODE



MEASURES OF DISPERSION

- Range
- Standard Deviation
- Variance



EXAMPLE FOR RANGE

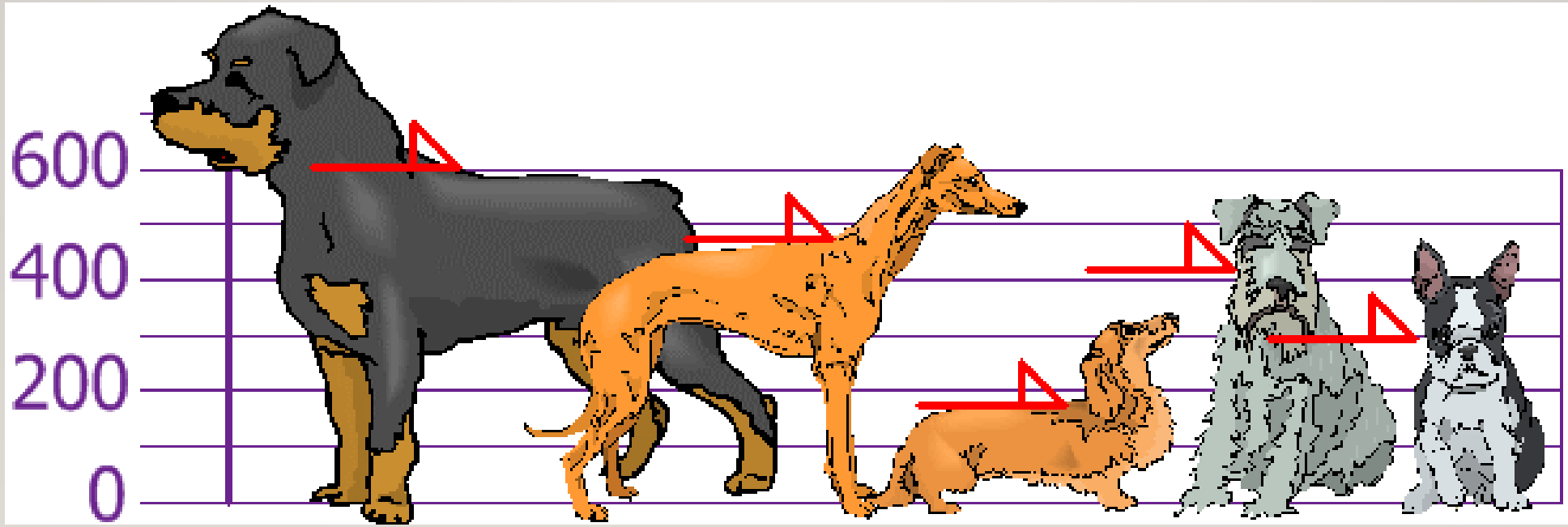
- Range = Maximum Value – Minimum Value
- A full Turkey started to cook at 6:00 p.m and was done by 8:00 p.m
- What is the range of time for cooking the Turkey
- Range = 2hrs

WHAT DO I MEAN RANGE?

- When I have numbers {1,2,3,4}
- The range is 4-1 = 3, i.e.,
- 1-2 range 1
- 2-3 range 2
- 3-4 range 3

EXAMPLES FOR VARIANCE AND SD

- Research on the heights of dogs:
- Say you calculate the height of dogs from their shoulders



EXAMPLES FOR VARIANCE AND SD

- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

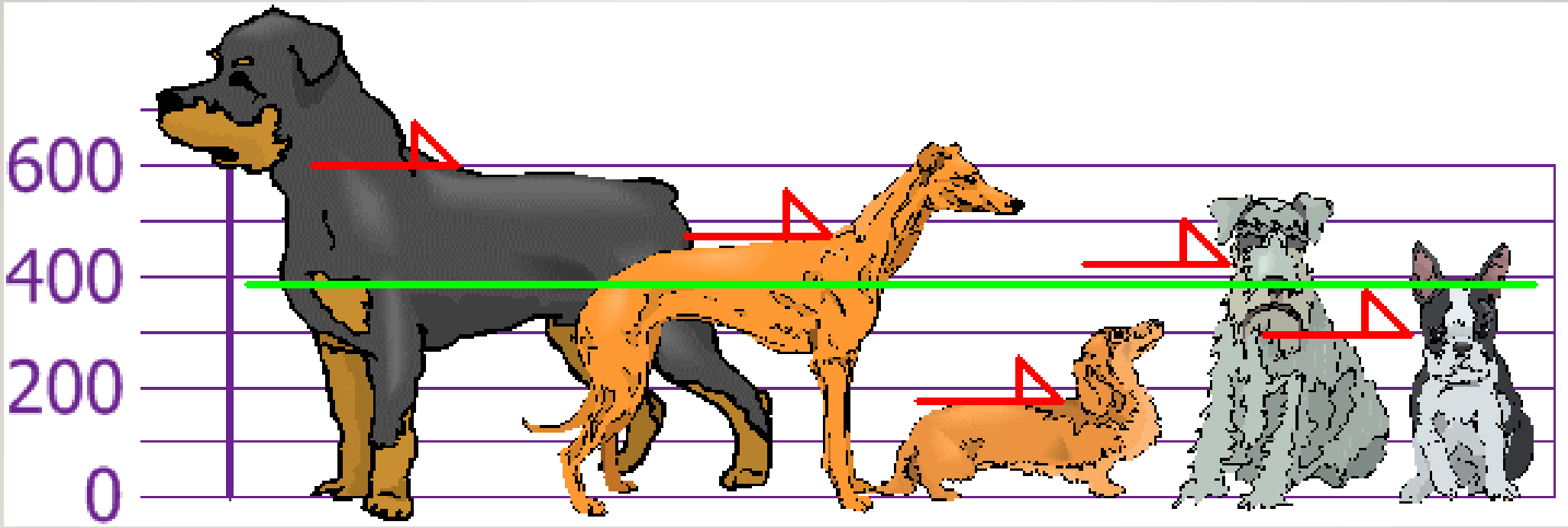
$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= \frac{1970}{5}$$

$$= 394$$

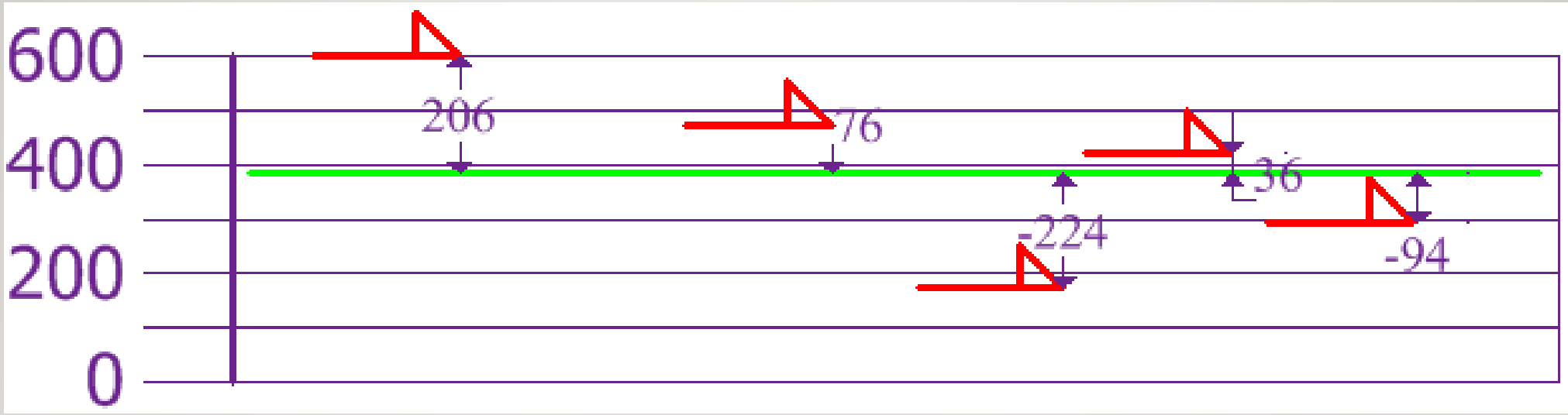
EXAMPLES FOR VARIANCE AND SD

- so the mean (average) height is 394 mm. Let's plot this on the chart:



EXAMPLES FOR VARIANCE AND SD

- Now we calculate each dog's difference from the Mean:



EXAMPLES FOR VARIANCE AND SD

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

EXAMPLES FOR VARIANCE AND SD

Standard Deviation

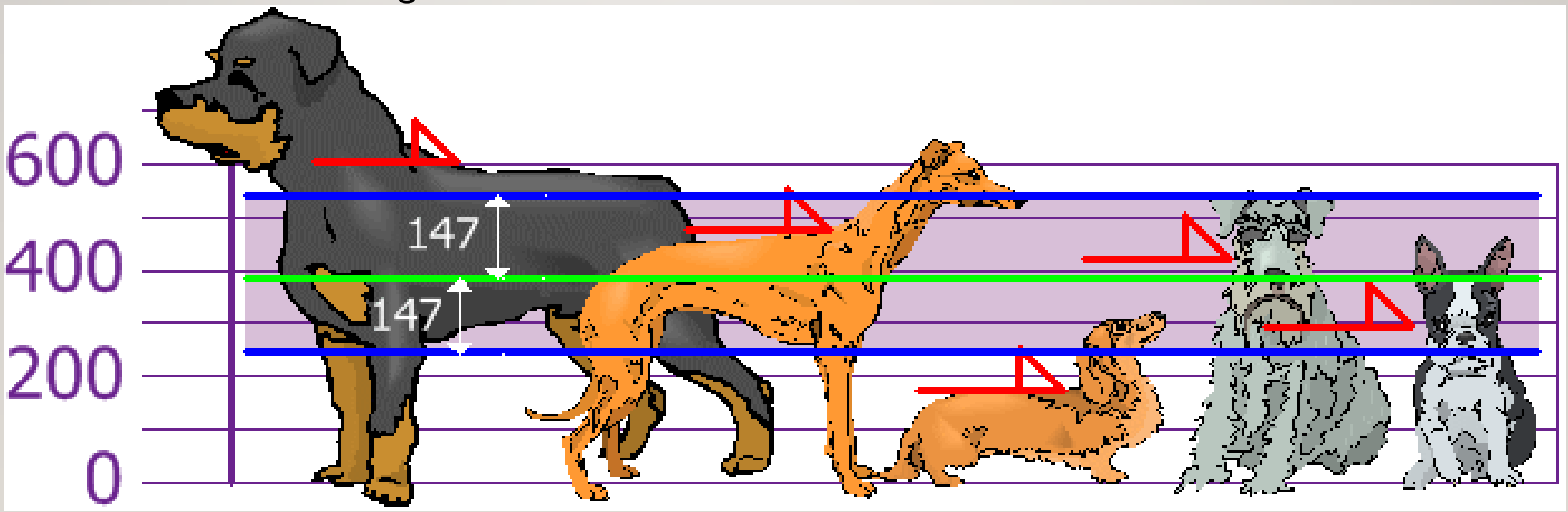
$\sigma = \sqrt{21704}$

$= 147.32...$

$=$ **147** *(to the nearest mm)*

EXAMPLES FOR VARIANCE AND SD

- So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.



EXAMPLE FOR VARIANCE AND SD

- Market research result of recent customer survey on product X

Customer	Rating for product X
A	1
B	5
C	1
D	4
E	2
F	4
G	5

EXAMPLE FOR VARIANCE AND SD

- Mean = 3.14
- Standard Deviation = 1.77

The deviation is high. Which means the Survey is not reliable.

EXAMPLE OF COEFFICIENT OF VARIATION

- Fred wants to find safe investment that provides stable returns. He considers the following options for investment:
- **Stocks:** Fred was offered stocks of ABC Corp. It is a mature company with the strong operational and financial performance. The Volatility (SD) of the stock is 10% and the expected return (mean) is 14%.
- **Mutual Funds:** which offers an expected return (mean) of 13% with a Volatility (SD) of 7%.
- **Bonds:** Bonds with return (mean) of 3% with a 2% Volatility (SD).

EXAMPLE OF COEFFICIENT OF VARIATION

$$\text{Coefficient of Variation (Stock)} = \frac{10\%}{14\%} \times 100\% = 71.4\%$$

$$\text{Coefficient of Variation (ETF)} = \frac{7\%}{13\%} \times 100\% = 53.8\%$$

$$\text{Coefficient of Variation (Bond)} = \frac{2\%}{3\%} \times 100\% = 66.7\%$$

EXAMPLE OF COEFFICIENT OF VARIATION

There are two Experiments being carried out on two products and we need to find out which is precise

Experiment	Product	Result Deviation	Mean
X	A	4.0	100
Y	B	4.8	120

EXAMPLE OF COEFFICIENT OF VARIATION

Calculate the CV to see:

$$CV\ (\%) = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100$$

$$\frac{4.8\ (SD)}{120\ (Mean)} \times 100 = 0.04\%\ (CV) \qquad \frac{4.0\ (SD)}{100\ (Mean)} \times 100 = 0.04\%\ (CV)$$

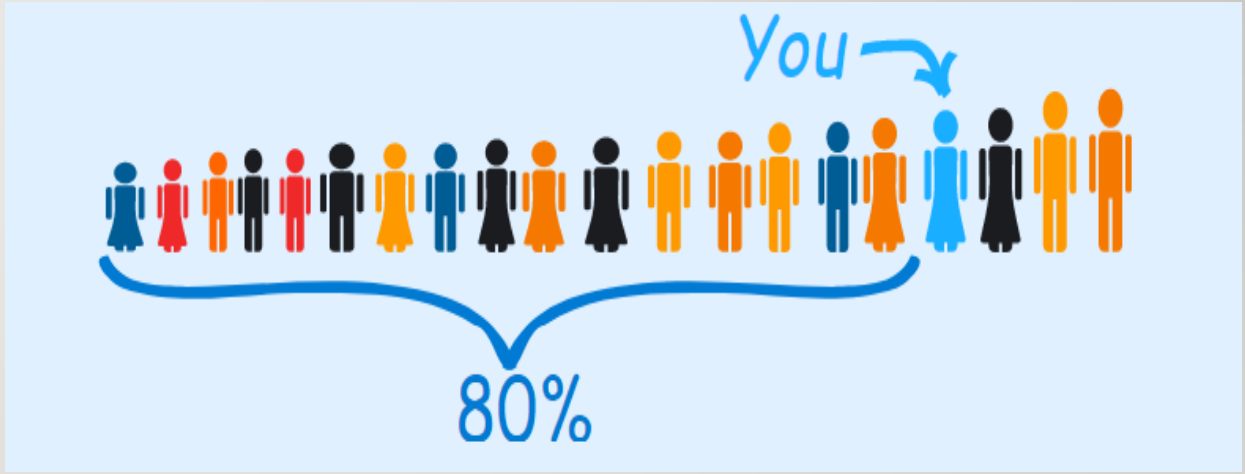
SO BOTH ARE EQUALLY PRECISE OR RELIABLE

PERCENTILE

- Percentile means "Percentage below"
- If there are 2,00,000 people appearing for examination, you are in top 1% i.e. Top 2000.

PERCENTILE

- You are the fourth tallest person in a group of 20
- 80% of people are shorter than you:



PERCENTILE

- The percentile rank is calculated using the formula
- R= P/100(N)

WHERE P IS THE DESIRED PERCENTILE AND N IS THE NUMBER OF DATA POINTS

EXAMPLE FOR PERCENTILE

- If the scores of a set of students in a math test are 20 , 30 , 15 and 75 what is the percentile rank of the score 30?

EXAMPLE FOR PERCENTILE

- If the scores of a set of students in a math test are 20 , 30 , 15 and 75 what is the percentile rank of the score 30?

Arrange the numbers in ascending order and give the rank ranging from 1 to the lowest to 4 to the highest.

NUMBER	15	20	30	75
RANK	1	2	3	4

EXAMPLE FOR PERCENTILE

- Use the formula now,
- 3 = (P/100)4
- 75 = P
- Therefore, the score 30 has 75^{th} percentile

EXAMPLE FOR PERCENTILE

- Determine the percentile of the sales of a new product across countries is given. Find the Rank of India in the sales chart. (What percentile India is at?)

Country	USA	China	India	Australia	Japan	Germany	Russia
Sales	3	4	10	12	14	15	20
Rank	1	2	3	4	5	6	7

EXAMPLE FOR PERCENTILE

3 = (P/100)*7 = 42.85 percentile

There are 43.8% countries below India

POPULATION CHART OF COUNTRIES (TOP 10 LISTED OUT OF 195 COUNTRIES)

1. China	1,389,618,778	6. Brazil	210,301,591
2. India	1,311,559,204	7. Nigeria	208,679,114
3. United States	331,883,986	8. Bangladesh	161,062,905
4. Indonesia	264,935,824	9. Russia	141,944,641
5. Pakistan	210,797,836	10. Mexico	127,318,112

POPULATION CHART OF COUNTRIES (TOP 10 LISTED OUT OF 195 COUNTRIES)

- If we put the population in **ascending order**
- Rank of India = 194
- 194 = (P/100)195
- = 99.487 Percentile
- While means 99% of the countries have low population than India

BOX-AND-WHISKER PLOTS

- **A Five Number Summary includes:**

Minimum

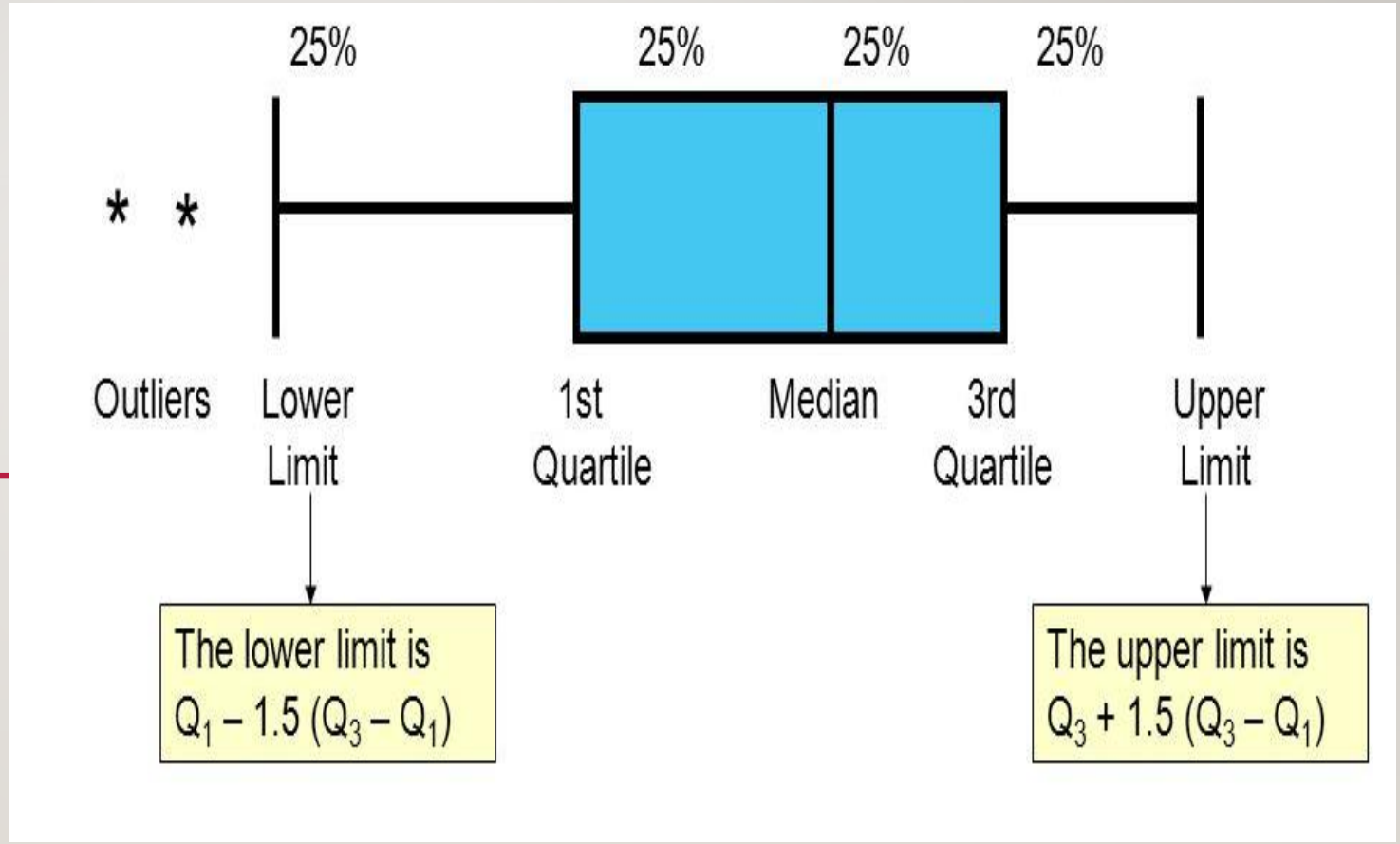
First Quartile

Median (Second Quartile)

Third Quartile

Maximum

BOX-AND-WHISKER PLOTS

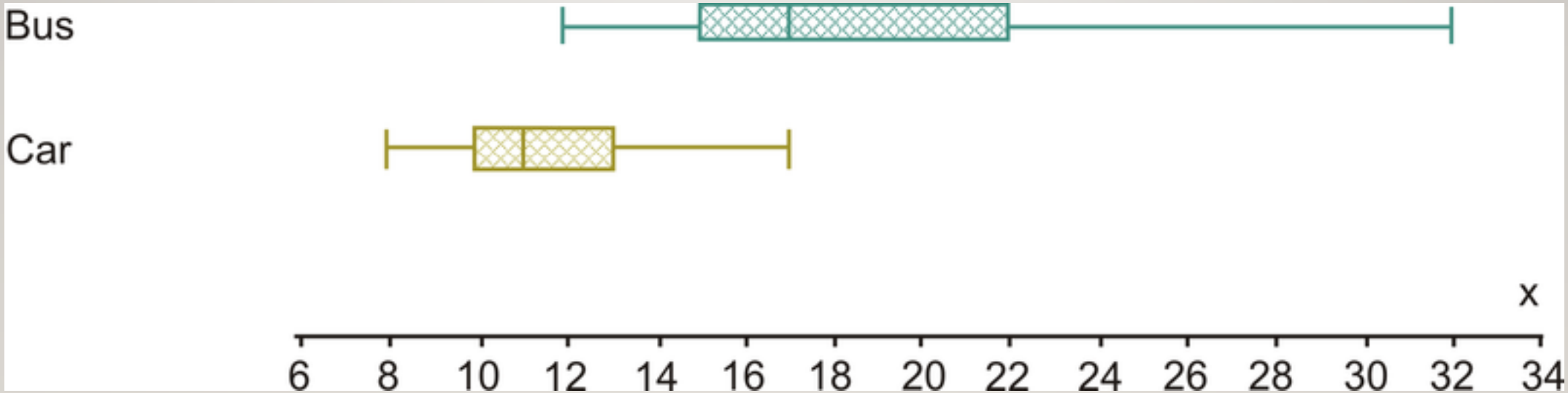


BOX AND WHISKER PLOT

- A data scientist conducted a survey of times it takes for him to reach to the office from his home. He drove through Car and recoded the times and went through bus and recorded the time

BUS (min)	**12**	**14**	**16**	**16**	**17**	**18**	**22**	**25**	**32**
CAR (min)	8	9	10	10	11	11	12	14	17

BOX AND WHISKER PLOT



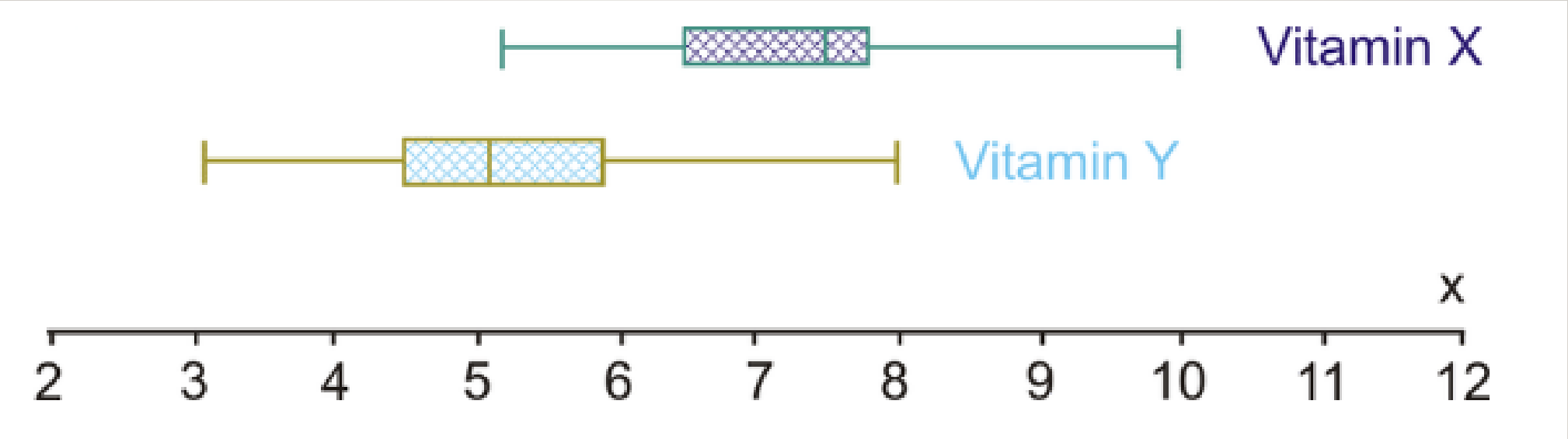
INFER FROM THE FOLLOWING

- The drug company wanted to see which of the 2 vitamins had the greatest impact on lowering people's cholesterol.

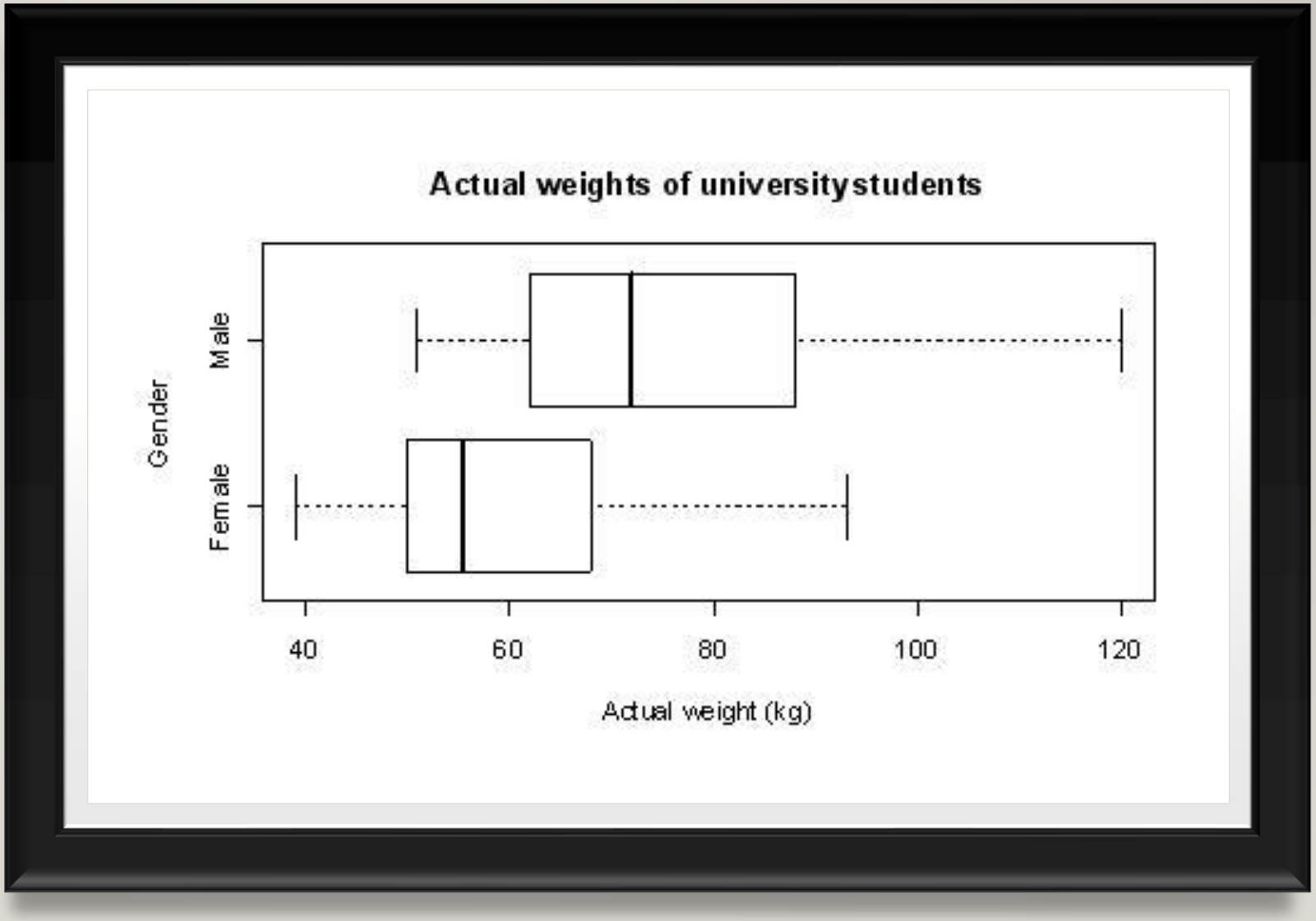
min X	7.2	7.5	5.2	6.5	7.7	10	6.4	7.6	7.7	7.8	8.1	8.3	7.2	7.1	6.5
min Y	4.8	4.4	4.5	5.1	6.5	8	3.1	4.6	5.2	6.1	5.5	4.2	4.5	5.9	5.2

15 people chosen at random to take Vitamin X for 2 months and then have their cholesterol levels checked.
15 different people were randomly chosen to take Vitamin Y for 2 months and then have their cholesterol levels checked.
All 30 people had cholesterol levels between 8 and 10 before taking one of the vitamins.

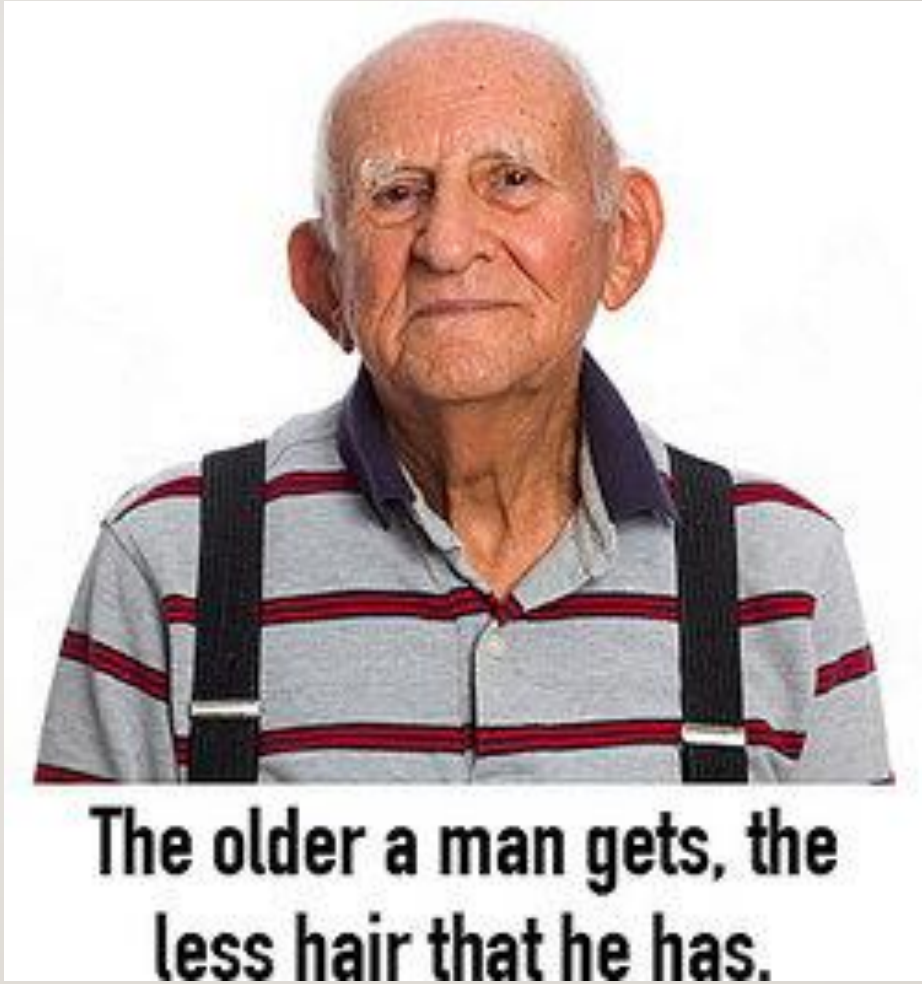
INFER FROM THE FOLLOWING



MAKE AN INFERENCE-EXERCISE



CORRELATION



CORRELATION

- A student who has many absences has a decrease in grades.

CORRELATION

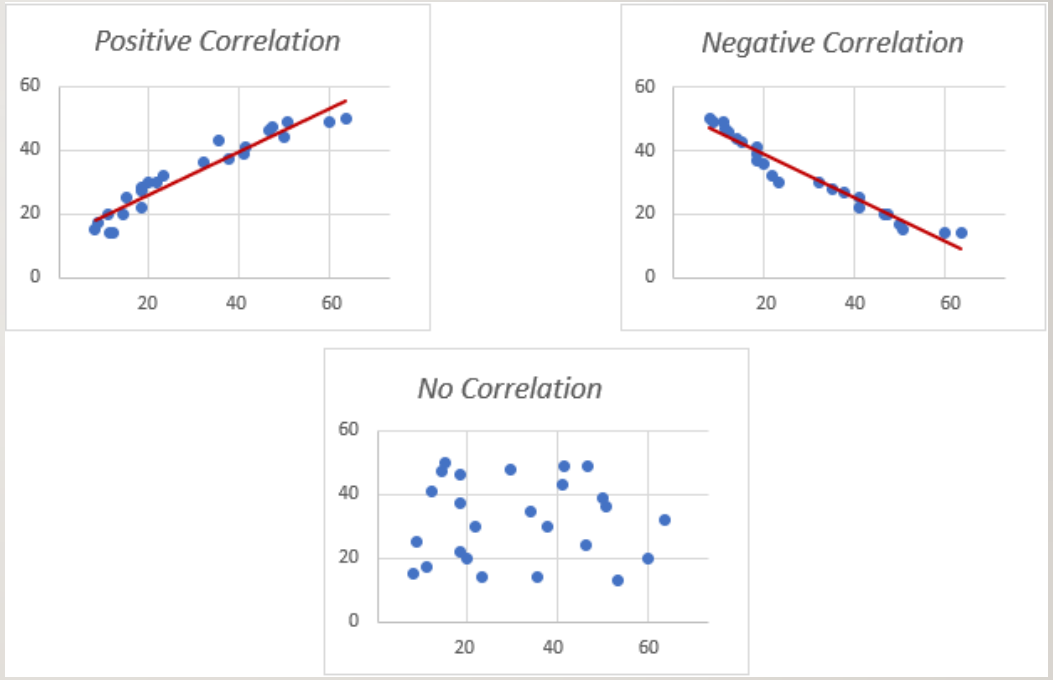
- As age increases your salary also increases

CORRELATION

- While travelling as time increases the more you go towards your destination also increases

SO WHAT IS CORRELATION

A MUTUAL RELATIONSHIP OR CONNECTION BETWEEN TWO OR MORE THINGS.



POSITIVE CORRELATION

NEGATIVE CORRELATION

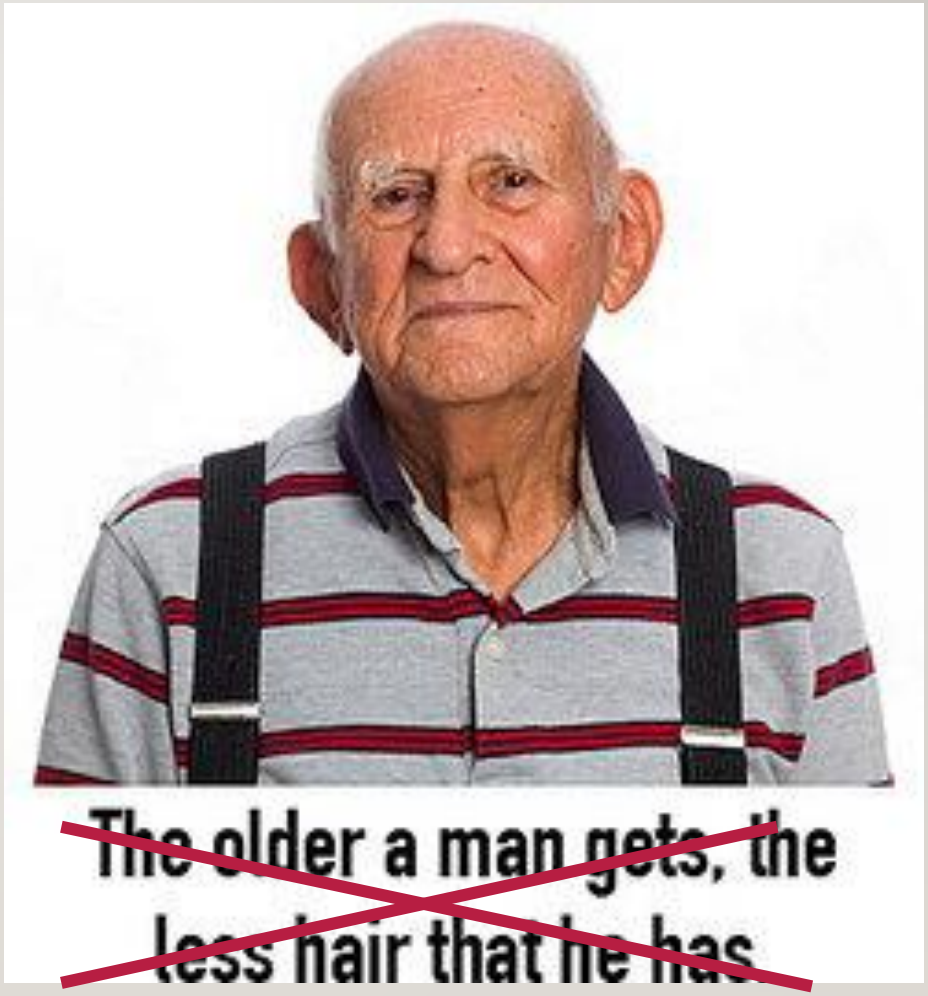
- If a car tire has more air, the car may use less petrol per km.

CAUSATION

- I accept correlation is mutual relationship or connection between two variables.
- But does it explain causation?

BUT WAIT THERE IS SOMETHING WRONG!!

Age is not the reason for hair loss. It might be heredity, some disorder, stressed life, No activity., etc.,



BUT WAIT THERE IS SOMETHING WRONG!!

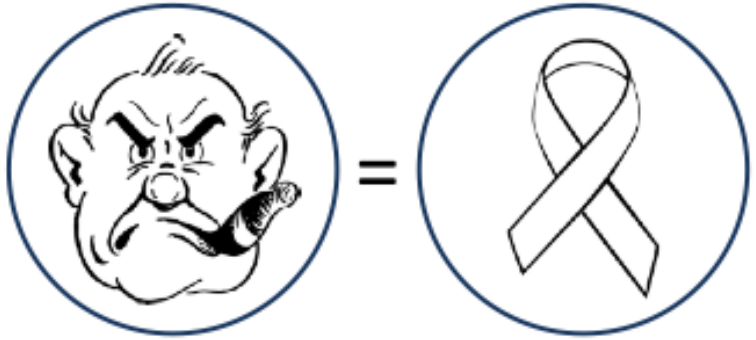
- A student who has many absences has a decrease in grades.
- NO!!!!!!!
- A student who does not prepare well will have low score not who is absent a lot!!

BUT WAIT THERE IS SOMETHING WRONG!!

- As age increases your salary also increases
- No!!!!!!
- Not the age implies your salary your Experience does, the work you do does, promotions, etc.,

BUT WAIT THERE IS SOMETHING WRONG!!

- While travelling as time increases the more you go towards your destination also increases
- NO!!!!!!!!!!
- The speed with which you drive does! If u travel faster you will reach the destination sooner

	Relationship	Definition
	Many people who smoke also drink.	Correlation
	Smoking has been proven to cause lung cancer	Causation

CORRELATION ≠ CAUSATION

COVARIANCE

COVARIANCE IS A MEASURE OF THE JOINT VARIABILITY OF TWO RANDOM VARIABLES.

FORMULA FOR COVARIACE

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

x = the independent variable
y = the dependent variable
n = number of data points in the sample
$\bar{x}$ = the mean of the independent variable x
$\bar{y}$ = the mean of the dependent variable y

FORMULA FOR CORRELATION

$$r_{(x,y)} = \frac{COV(x,y)}{s_x s_y}$$

$r_{(x,y)}$ = correlation of the variables x and y

$COV(x, y)$ = covariance of the variables x and y

s_x = sample standard deviation of the random variable x

s_y = sample standard deviation of the random variable y

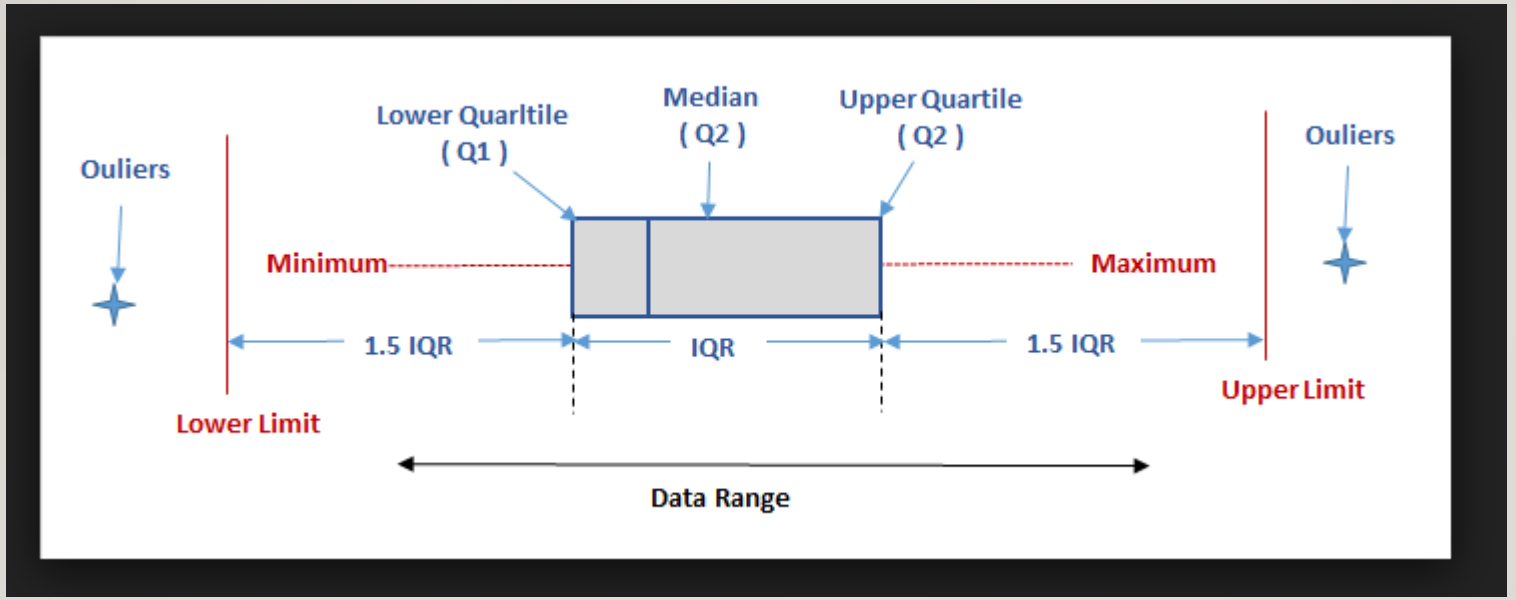
COVARIANCE AND CORREALTION

BASIS FOR COMPARISON	COVARIANCE	CORRELATION
Meaning	Covariance is a measure indicating the extent to which two random variables change in tandem.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	Lie between $-\infty$ and $+\infty$	Lie between -1 and +1

DATA VISUALIZATION - PLOTS

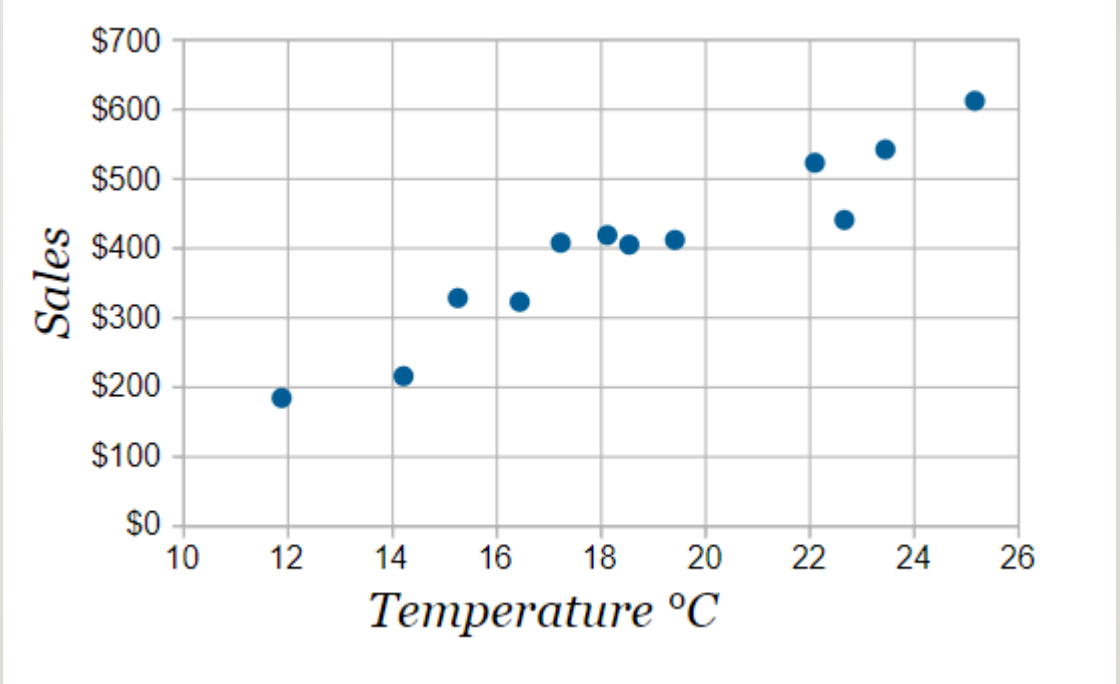
1. *Box Plot*
2. *Scatter plot*
3. *Histogram*
4. *Density Plot*

BOX PLOT - SHOWS THE DATA SPREAD FOR INDIVIDUAL COLUMNS



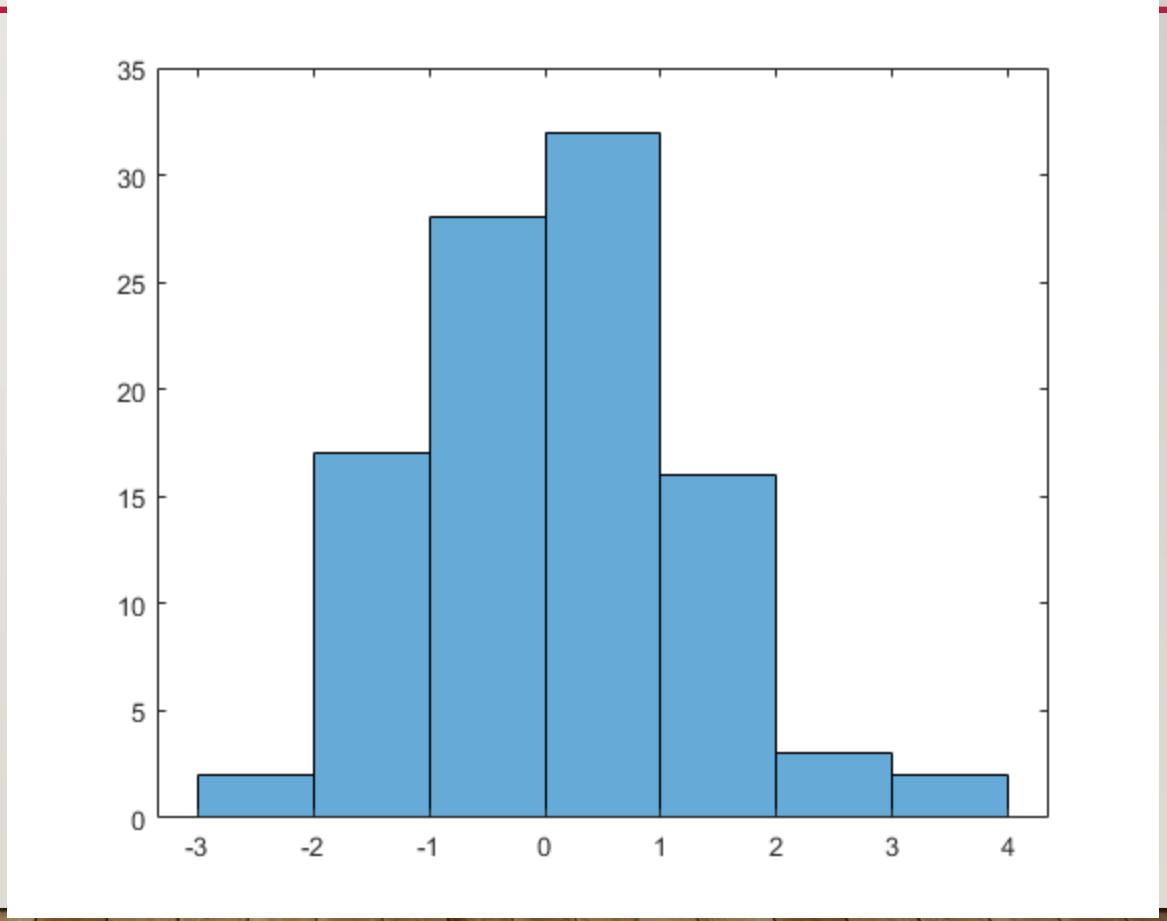
SCATTER PLOT - SHOWS RELATIONSHIP BETWEEN 2 COLUMNS

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

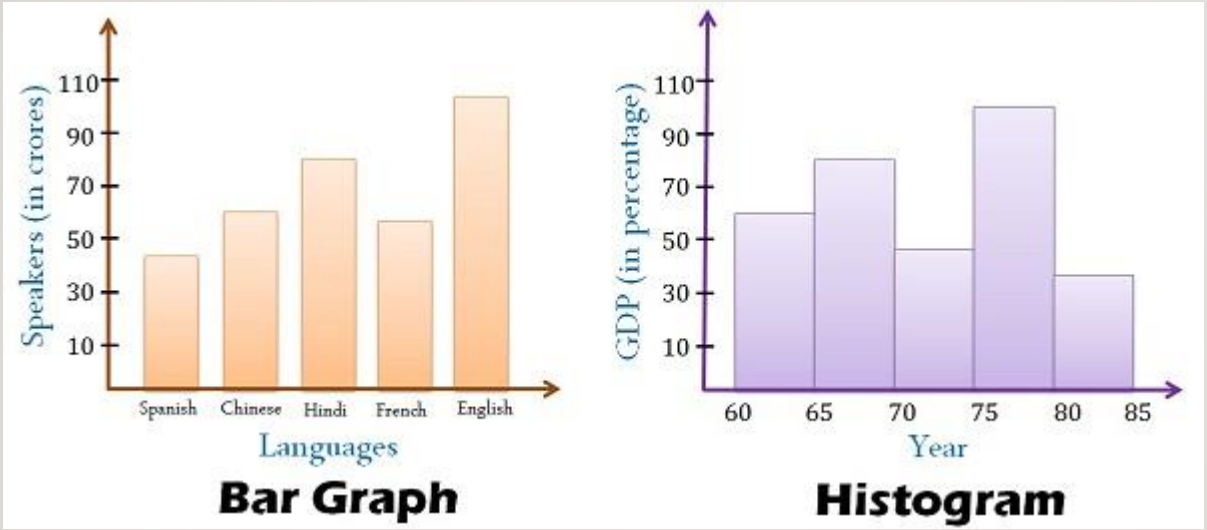


HISTOGRAMS

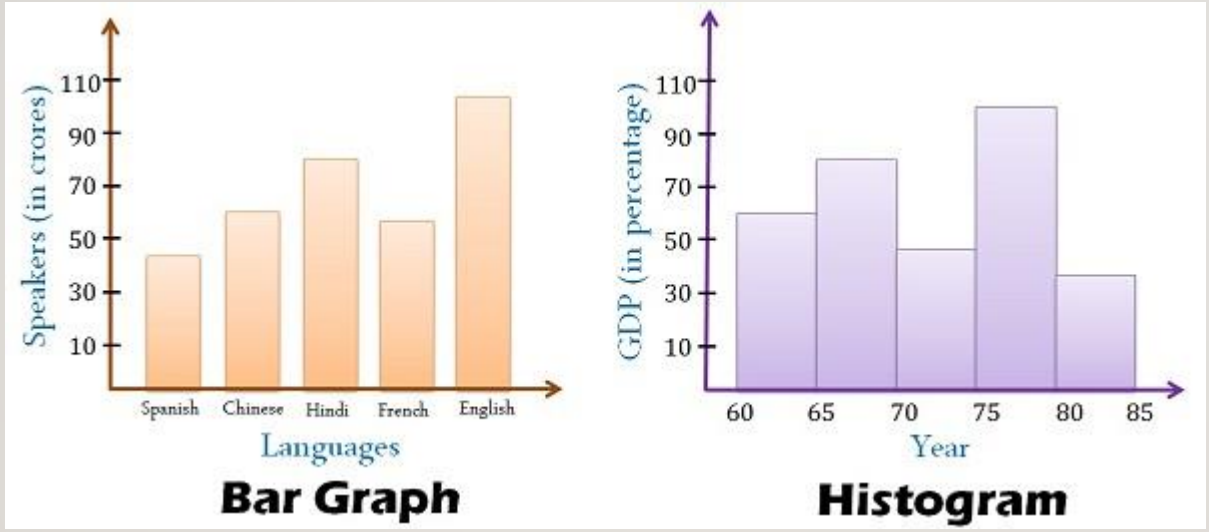
A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable



DIFFERENCE BETWEEN HISTOGRAM AND BAR GRAPH?



DIFFERENCE BETWEEN HISTOGRAM AND BAR GRAPH?



A **histogram** represents the frequency distribution of continuous variables. Conversely, a **bar graph** is a diagrammatic comparison of discrete variables. Histogram presents numerical data whereas **bar graph** shows categorical data.

DENSITY PLOT - SHOWS THE DISTRIBUTION OF DATA