WELCOME TO TRAINING

ADVANCE DATA SCIENCE (ADS)

Fireblaze AI School

STATISTICS

Fireblaze AI School

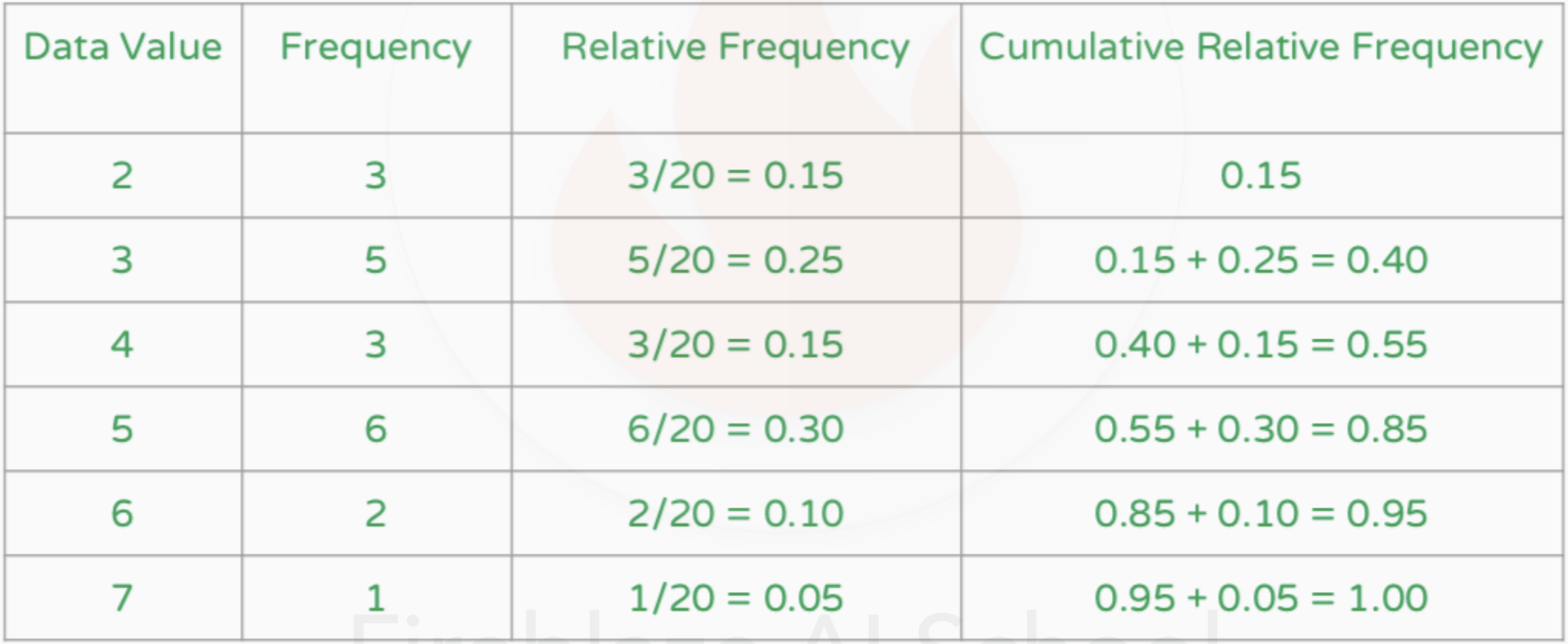
Important Terms

- **Frequency:** A frequency is the number of times a value of the data occurs.Example: 20 students were asked how many hours they worked per day. Their responses, in hours, are as follows: 5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3

- **Relative Frequency:** A relative frequency is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes.

- **Cumulative Relative Frequency:** Cumulative relative frequency is the accumulation of the previous relative frequencies.

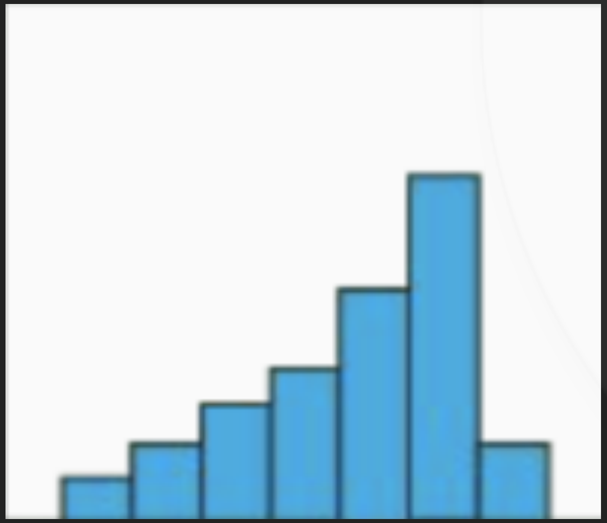
Frequency, Relative Frequency and CRF



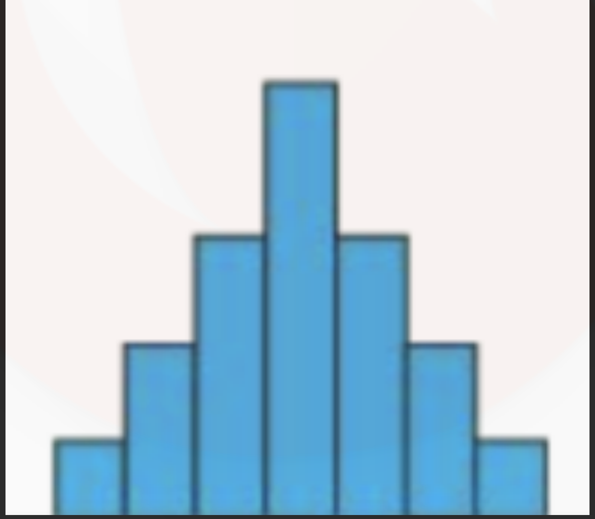
Data Value	Frequency	Relative Frequency	Cumulative Relative Frequency
2	3	3/20 = 0.15	0.15
3	5	5/20 = 0.25	0.15 + 0.25 = 0.40
4	3	3/20 = 0.15	0.40 + 0.15 = 0.55
5	6	6/20 = 0.30	0.55 + 0.30 = 0.85
6	2	2/20 = 0.10	0.85 + 0.10 = 0.95
7	1	1/20 = 0.05	0.95 + 0.05 = 1.00

Explore the Data

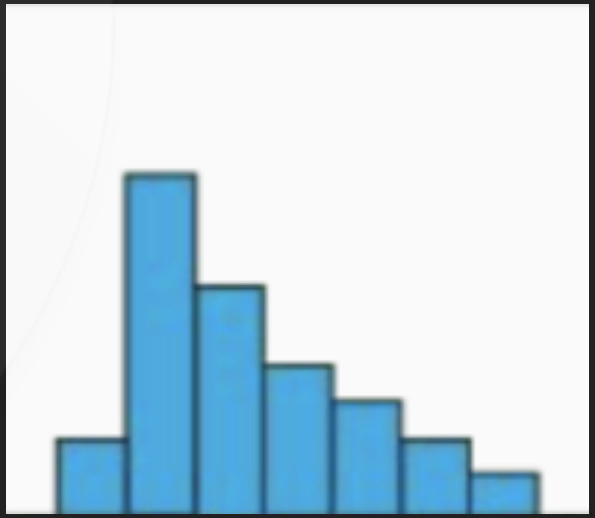
Skewed Left



Symmetric



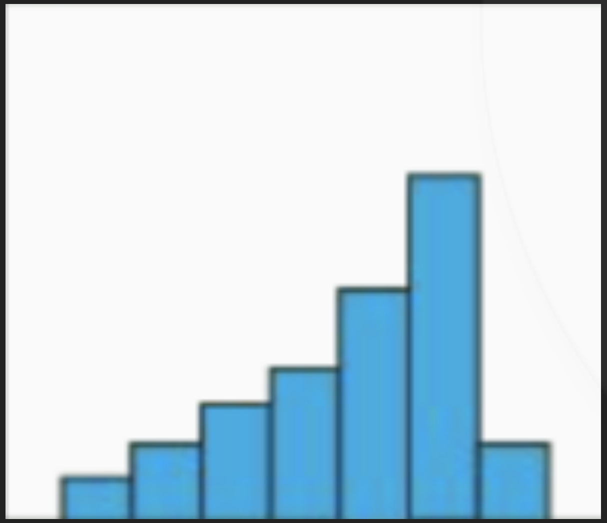
Skewed Right



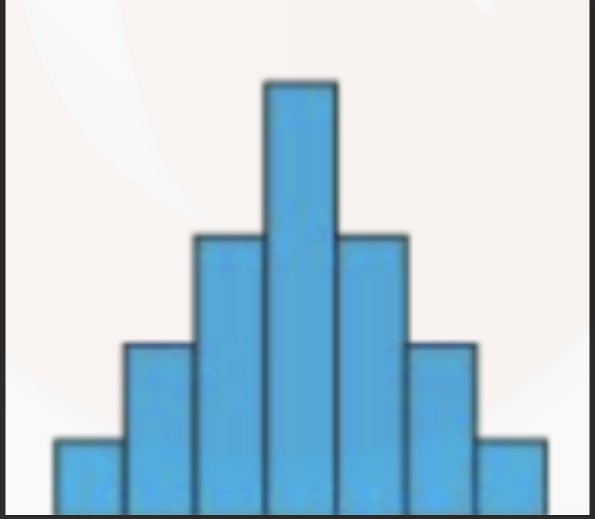
1. Middle one has the shape of a bell curve, has one peak, and is approximately symmetric.
2. Left one is left skewed and unimodal
3. Right one is right skewed and unimodal

Explore the Data

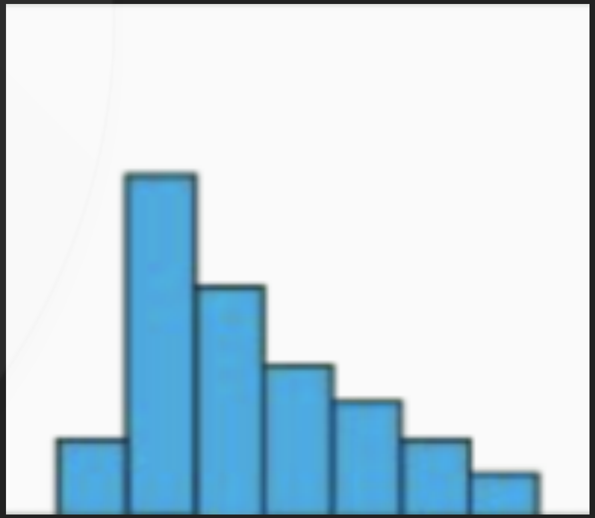
Skewed Left



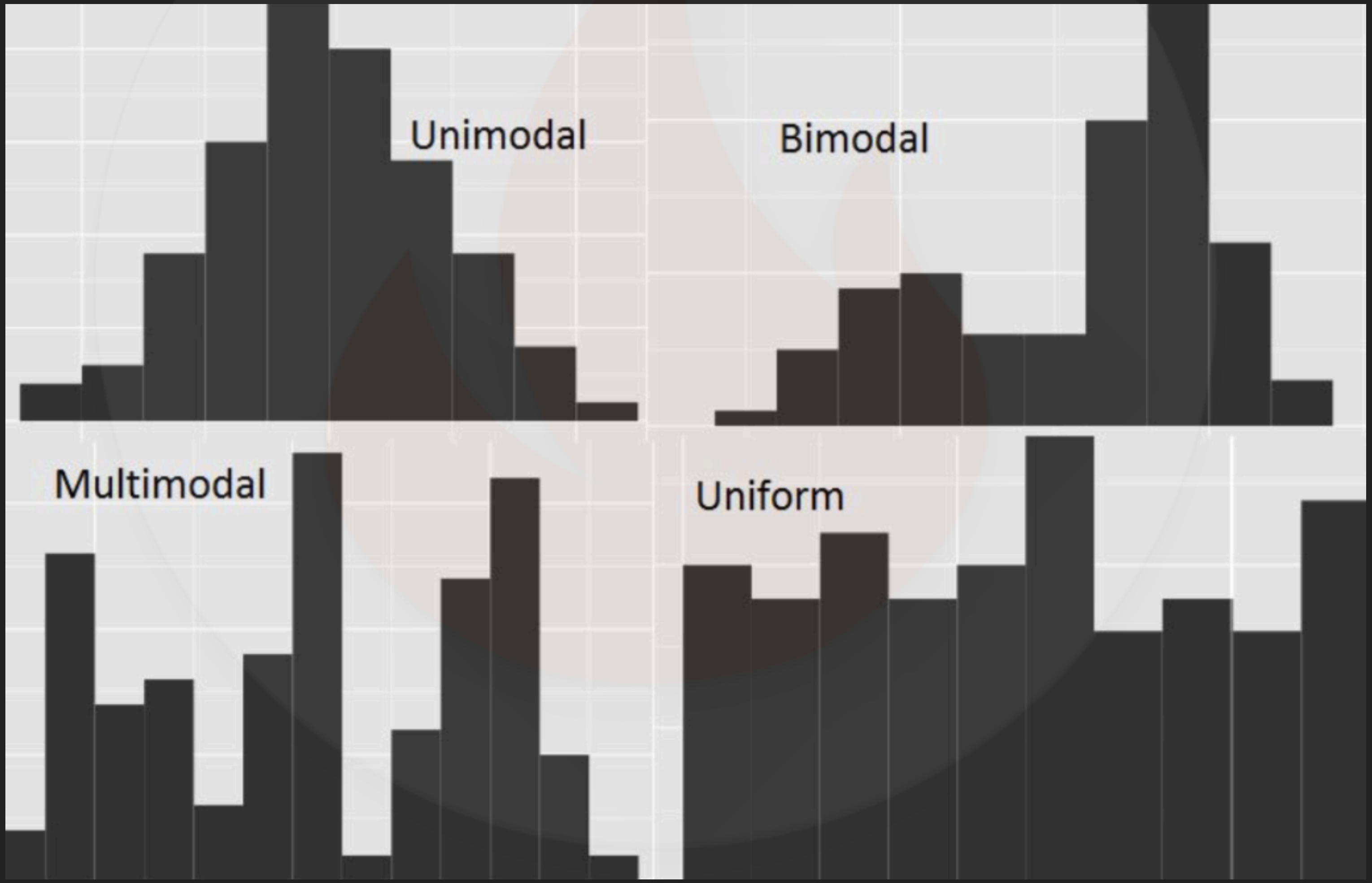
Symmetric



Skewed Right



Four kind of modalities



Unimodal: It has only one peak
Bimodal: It has two peak
Multimodal: It has many peak
Uniform: All are distributed uniformly

Measures of Central Tendency

It describes a whole set of data with a single value that represents the centre of its distribution.

There are three main measures of central tendency: the mode, the median and the mean.

Mean : Average Value
Median : Middle Value
Mode : Most Frequent Value

MEAN

The mean is the sum of all the values divided by the number of observations or sample size. It is nothing but the average value.

The mean of the values 5,6,6,8,9,9,9,9,10,10 is (5+6+6+8+9+9+9+9+10+10)/10 = 8.1

Limitation: It is affected by extreme values. Very large or very small numbers can distort the answer.

MEDIAN

The median is nothing more than the middle value of your observations when they are order from the smallest to the largest. It is the middle value. It splits the data in half. Half of the data are above the median; half of the data are below the median.

7, 8, 7, 6, 9, 8, 8 → 6, 7, 7, 8, 8, 8, 9 → 8 is the Median 7, 8, 7, 6, 9, 8, 8, 7 → 6, 7, 7, 7, 8, 8, 8, 9 → (7 + 8)/2 = 7.5 is the Median

Advantage : It is NOT affected by extreme values. Very large or very small numbers does not affect it.

MODE

Mode: It is the value that occurs most frequently. In other words, mode is the most common outcome. Mode is the name of the category that occurs more often. There is a chance of having more than one

5, 6, 5, 7, 5, 8, 9, 5 → 5 is the Mode
5, 6, 6, 5, 7, 6, 5, 6, 8, 9, 5, 6 → 5 and 6 are mode.

Mode Advantage : It can be used when the data is not numerical.

Disadvantage :
1. There may be no mode at all if none of the data is the same .
2. There may be more than one mode.

WHEN TO USE WHAT MEASUREMENT OF CENTRAL TENDENCY?

When to use what measurement of central tendency?

Mean – When your data is not skewed i.e Symmetric/Normally Distributed. In other words, there are no extreme values present in the data set (Outliers).

Median – When your data is skewed or you are dealing with ordinal (ordered categories) data.

Mode - When dealing with nominal (unordered categories) data.

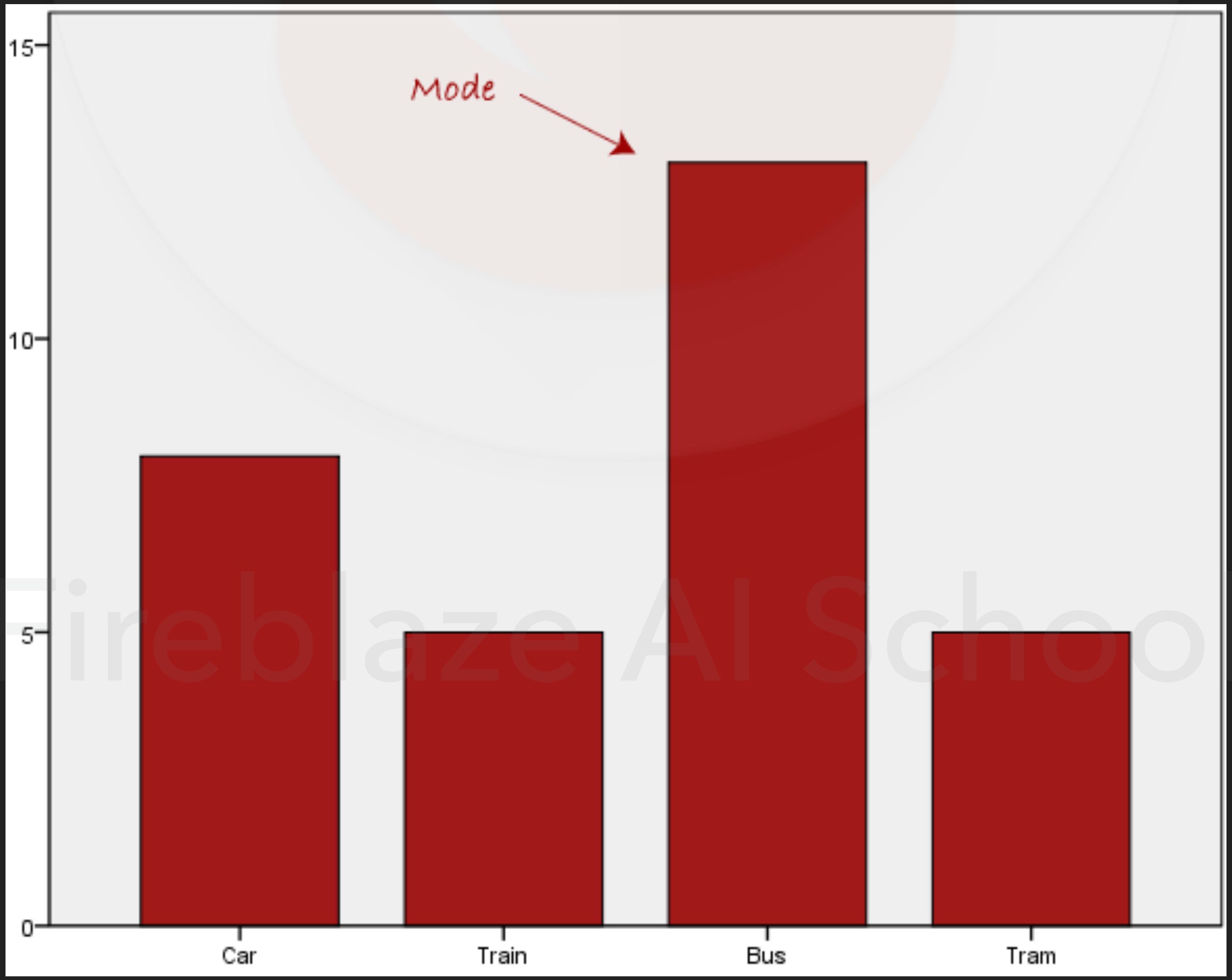
When to use Median instead of Mean

- If your data is quantitative then go for mean or median. Basically, if your data is having some influential outliers or data is highly skewed then median is the best measurement for finding central tendency. Otherwise go for Mean.

Eg : Salary 15k 18k 16k 14k 15k 15k 12k 17k 90k 95k
Mean is 30.1K where as most workers have salaries in the $12k to 18k range. Hence Median is to be preferred.

When to use Mode

- If data is Categorical (Nominal or Ordinal) it is impossible to calculate mean or median. So, go for mode.Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:



Measures of Variability

It is used to get good indication of how the values in a distribution are spread out i.e Measures of Variability.

Range: Difference between Maximum and Minimum value in a distribution.

Team1 → 2, 4, 10, 15, 24, 25, 40 → Range → 40 - 2 = 38

As ranges takes only the count of extreme values sometimes it may not give you a good impact on variability.

Disadvantage: It is very sensitive to outliers and does not use all the observations in a data set.

IQR (Interquartile Range)

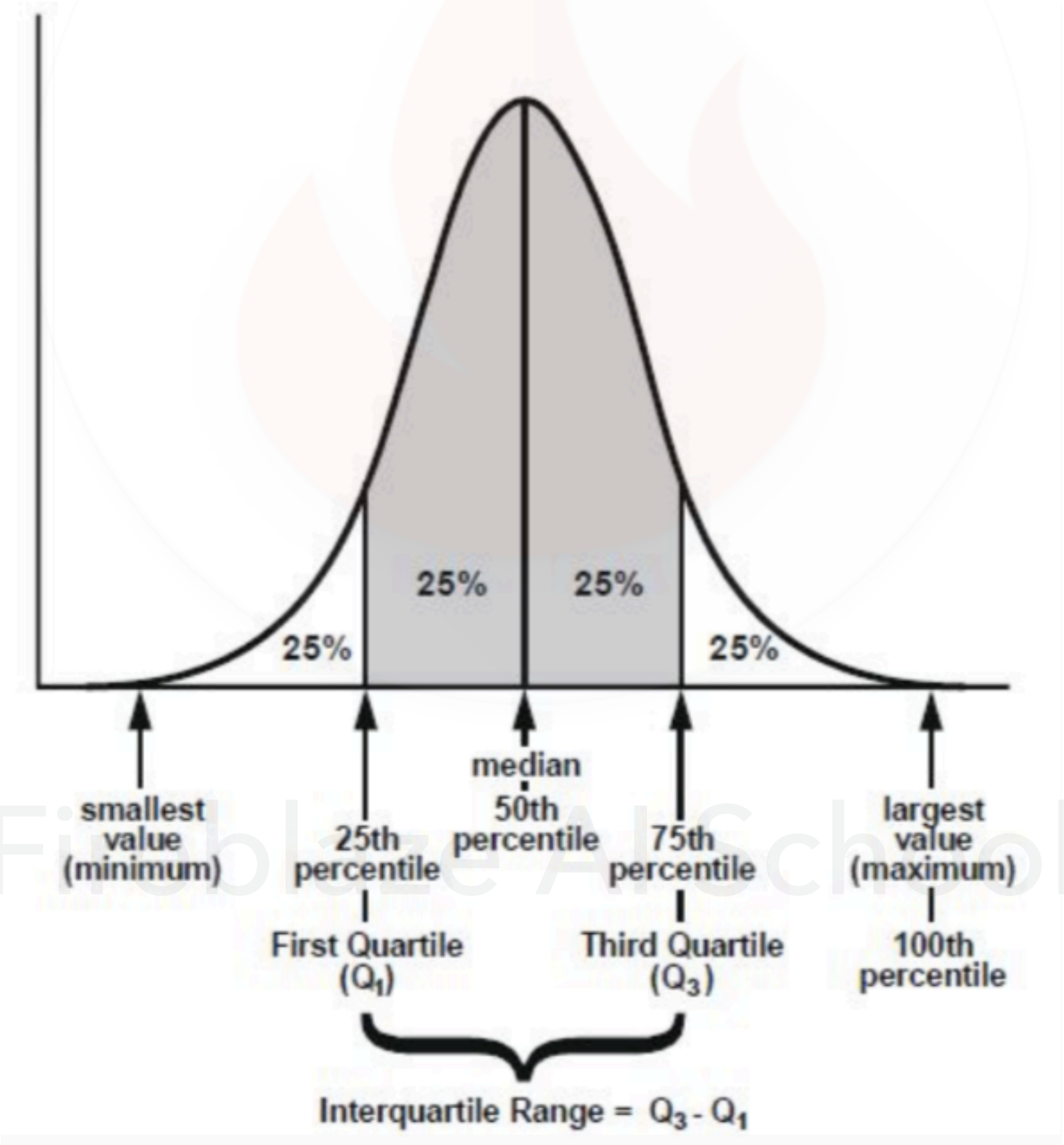
It equally divides the distribution into four equal parts called quartiles.

First 25% is 1st quartile (Q1), last one is 3rd quartile (Q3) and middle one is 2nd quartile (Q2) and it leaves out the extreme values.

2nd quartile (Q2) divides the distribution into two equal parts of 50%. So, basically it is same as Median.

The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1
IQR = Q3- Q1

IQR (Interquartile Range)



How to calculate IQR

Step 1: Arrange data in ascending order from low to high.
Step 2: Find the median or in other words Q2.
Step 3: Then find Q1 by looking the median of the left side of Q2.
Steps 4: Similarly find Q3 by looking the median of the right of Q2.
Steps 5: Now subtract Q1 from Q3 to get IQR.

How to calculate IQR

0, 12, 17, 4.5, 2.3, 17, 23, 14.6, 11, 10, 19.7, 20, 25 0, 2.3, 4.5, 10, 11, 12, 14.6, 17, 17, 19.7, 20, 23, 25

14.6 is the Middle Value or Median or Q2

Consider: 0, 2.3, 4.5, 10, 11, 12
→ 4.5 + 10 = 14.5/2 = **7.25 → Q1 Value**

Consider: 17, 17, 19.7, 20, 23, 25
→ 19.7 + 20 = 39.7/2 = **19.85 → Q3 Value**

→ Q3 Value IQR = Q3 - Q1 = 19.85 - 7.25 = **12.60 → IQR Value**

Advantage of IQR

- The main advantage of the IQR is that it is not affected by outliers because it doesn’t take into account observations below Q1 or above Q3.

- It might still be useful to look for possible outliers in your study.

- As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile. Outliers are values that "lie outside" the other values.

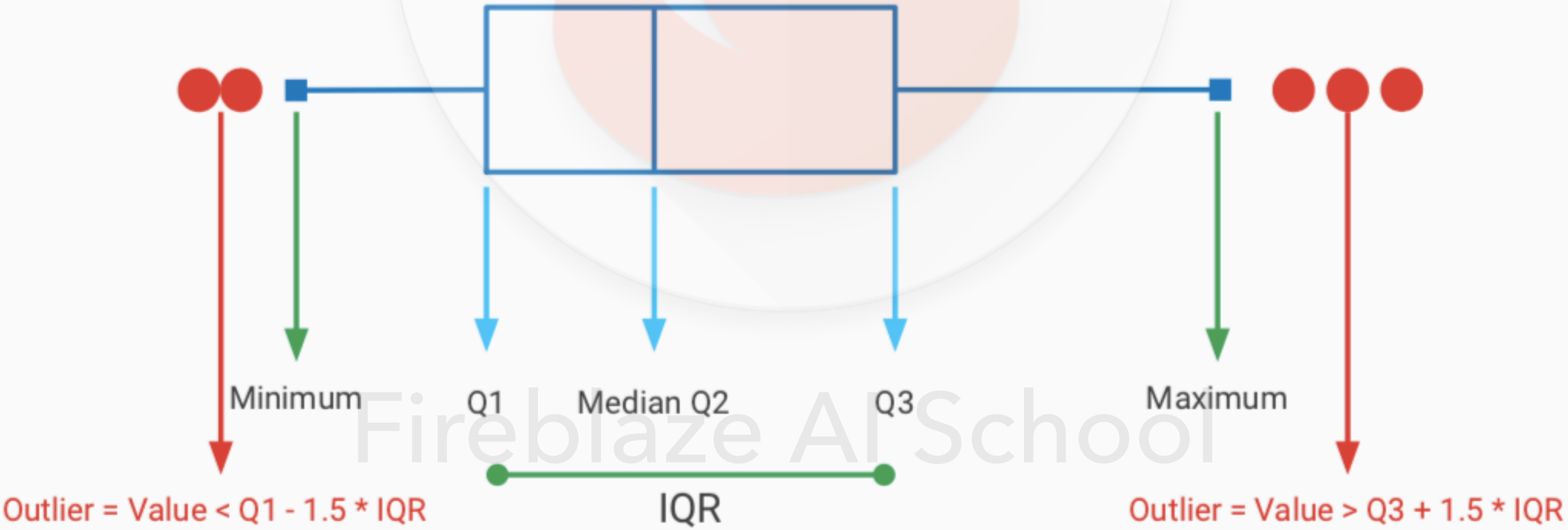
Outliers = Q1 - 1.5 * IQR

OR

Outliers = Q3 + 1.5 * IQR

Box Plot

There is one graph that is mainly used when you are describing centre and variability of your data. It is also useful for detecting outliers in the data.



Minimum

Q1

Median Q2

Q3

Maximum

Outlier = Value < Q1 - 1.5 * IQR

IQR

Outlier = Value > Q3 + 1.5 * IQR

Outliers

Outliers: "Outliers" are values that "lie outside" the other values.

Example: Long Jump A new coach has been working with the Long Jump team this month, and the athletes' performance has changed.

- Augustus: +0.15m
- Tom: +0.11m
- June: +0.06m
- Carol: +0.06m
- Bob: + 0.12m
- Sam: -0.56m

So here, Sam is outlier

Outliers

“Outliers" are values that "**lie outside**" the other values.

The mean is: Including “Sam” i.e. Outlier
(0.15+0.11+0.06+0.06+0.12-0.56) / 6 = -0.06 / 6 = -0.01m
So, on average the performance went **DOWN**.

The mean is: Excluding “Sam”
i.e. Outlier Mean = (0.15+0.11+0.06+0.06+0.12)/5 = 0.1 m
.So, on average the performance went **UP**.

Outliers

Outliers: "Outliers" are values that "lie outside" the other values.

The median ("middle" value):

- including Sam is: 0.085
- without Sam is: 0.11 (went up a little)

The mode (the most common value):

- including Sam is: 0.06
- without Sam is: 0.06 (stayed the same)

The mode and median didn't change very much.

Variance and Standard Deviation

Variance and Standard Deviation consider all the values of a variable to calculate the variability of the data.

These are called as Measures of Spread or Dispersion.

- **Standard Deviation:** The Standard Deviation is a measure of how spread out numbers are.
- **Variance:** The average of the squared differences from the Mean. i.e the Square of the Standard Deviation.

There are two types of variance and standard deviation in terms of Sample and Population.

Variance and Standard Deviation

For Samples:

variance = $s^2 = \Sigma(x - \mu)^2 / n - 1$

standard deviation $s = \sqrt{s^2}$

For Populations:

variance = $\sigma^2 = \Sigma(x - \mu)^2 / n$

standard deviation $= \sqrt{\sigma^2}$

- x is individual one value
- n is size of population
- μ is the mean of population or sample

Steps to calculate Standard Deviation

1. Work out the Mean (the simple average of the numbers)
2. Then for each number: subtract the Mean and square the result
3. Then work out the mean of those squared differences.
4. Take the square root of that and we are done!

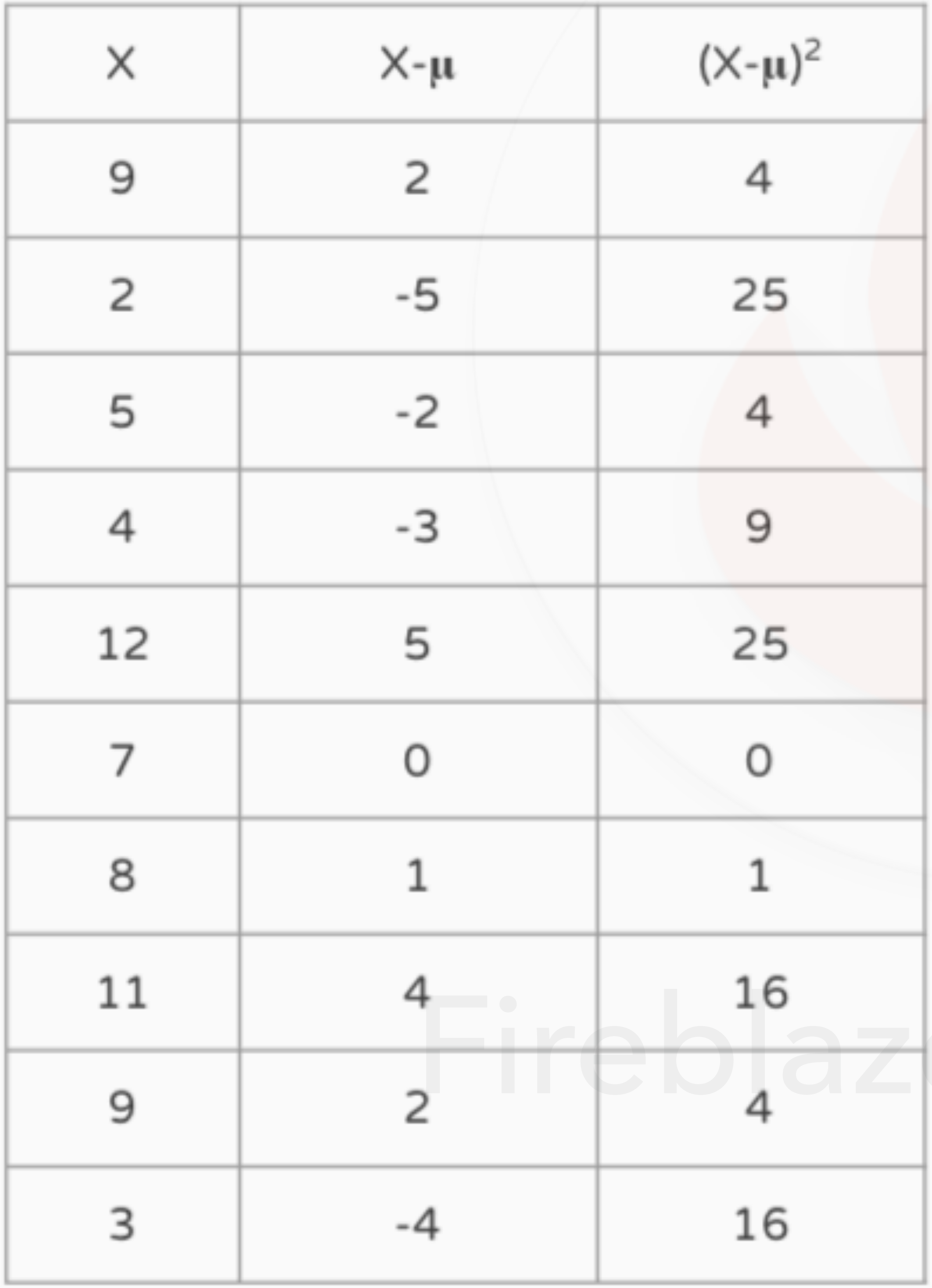
Steps to calculate Standard Deviation

Example: 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

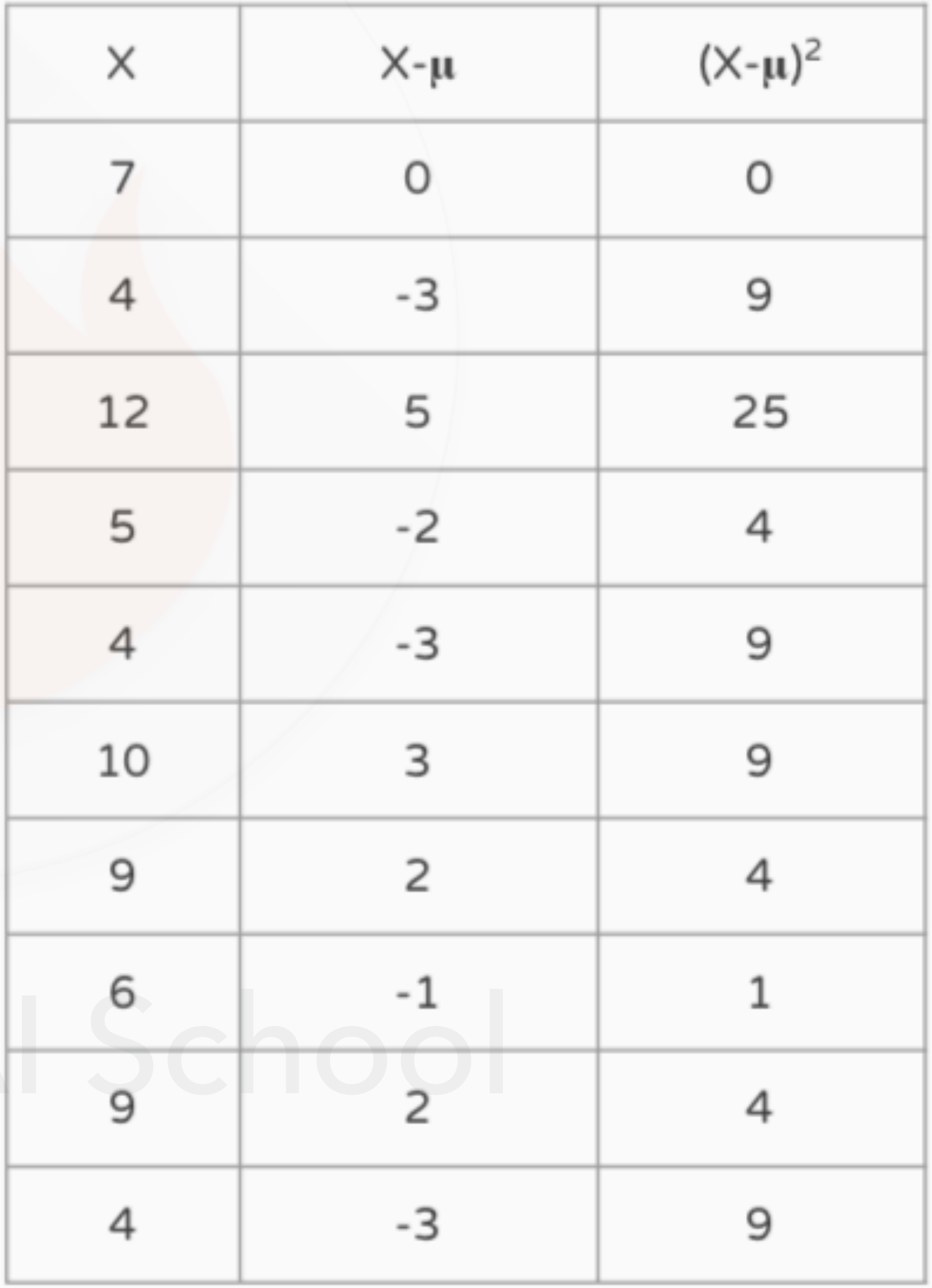
Work out the Standard Deviation.

Step 1: Work out the mean:
(9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4)/20 = 140/20 =7
So, $\mu = 7$

Step 2. Then for each number: subtract the Mean and square the result.



X	$X-\mu$	$(X-\mu)^2$
9	2	4
2	-5	25
5	-2	4
4	-3	9
12	5	25
7	0	0
8	1	1
11	4	16
9	2	4
3	-4	16



X	$X-\mu$	$(X-\mu)^2$
7	0	0
4	-3	9
12	5	25
5	-2	4
4	-3	9
10	3	9
9	2	4
6	-1	1
9	2	4
4	-3	9

Steps to calculate Standard Deviation

Step 3. Then work out the mean of those squared differences.

To work out the mean, add up all the values then divide by how many.

$=\Sigma 4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9$

$= 178$

Here, **$n = 20$**

Sample Variance (s^2) = 178/19 = 9.3

Population Variance (σ^2) = 178/20 = 8.9

Sample Standard Deviation (s) = $\sqrt{s^2} = \sqrt{9.3} = 3$

Population Standard Deviation (σ) = $\sqrt{\sigma^2} = \sqrt{8.9} = 2.9$

Intuition

1. If **variance is high**, that means you have **larger variability** in your dataset. In the other way, we can say more values are spread out around your mean value.

2. **Standard deviation** represents the average distance of an observation from the mean.

3. The **larger the standard deviation**, larger the variability of the data.

4. A low standard deviation indicates that the data points tend to be close to the **mean**

Z Score or Standard Score

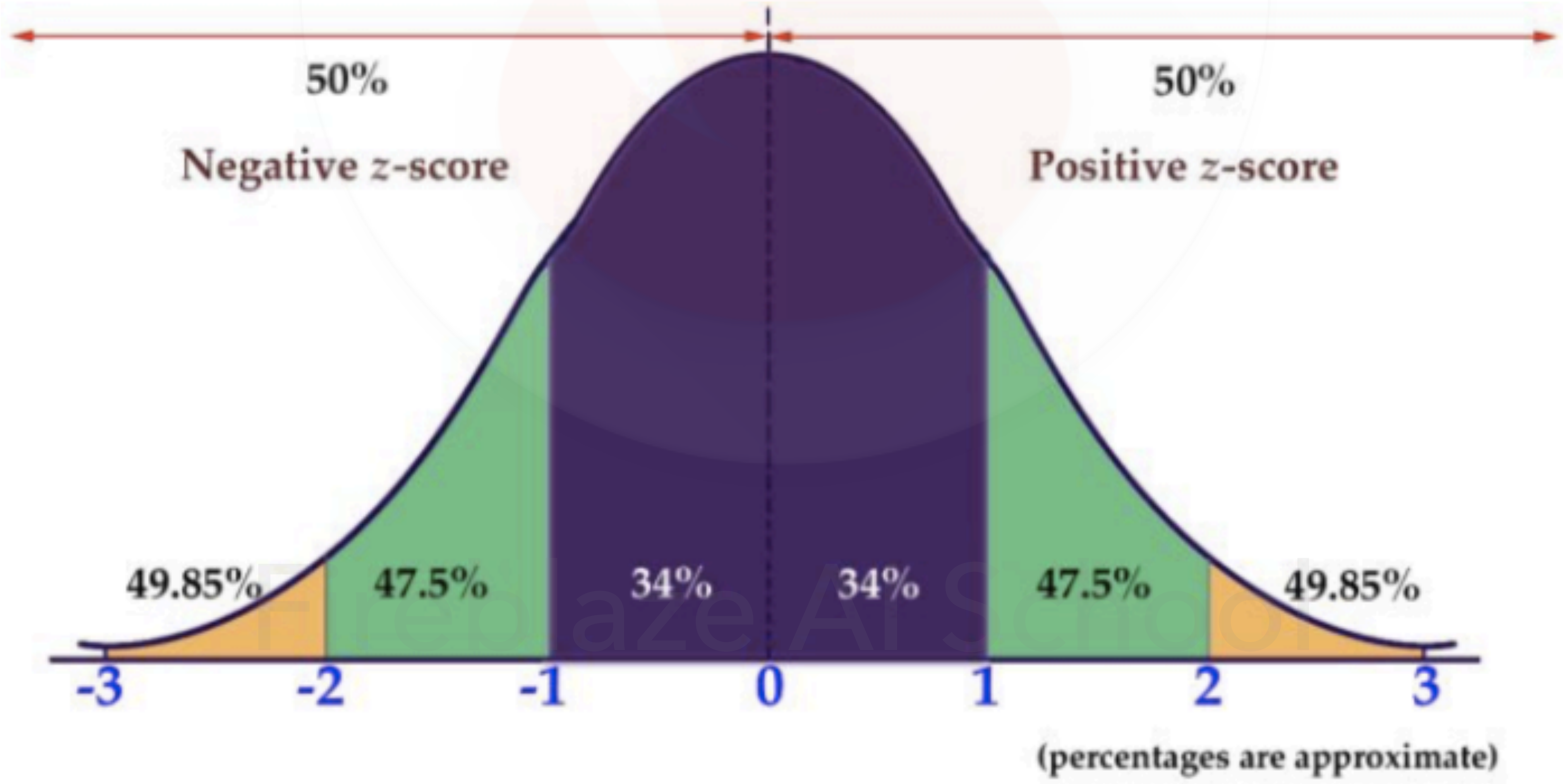
Z-score is the number of standard deviations from the mean a data point is.

Z Score = $(x - \mu) / \sigma$

x : Value of the element
μ : Population mean
σ : Standard Deviation

A z-score of zero tells you the values is exactly average while a score of +3 tells you that the value is much higher than average.

- **Bell Shape Distribution and Empirical Rule: If distribution is bell shape then it is assumed that about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; and about 99% have a z-score between -3 and 3.**

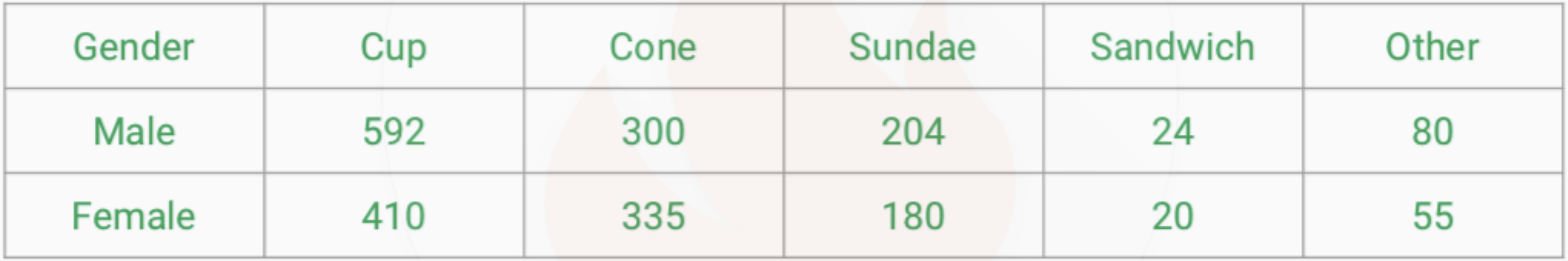


Contingency Table

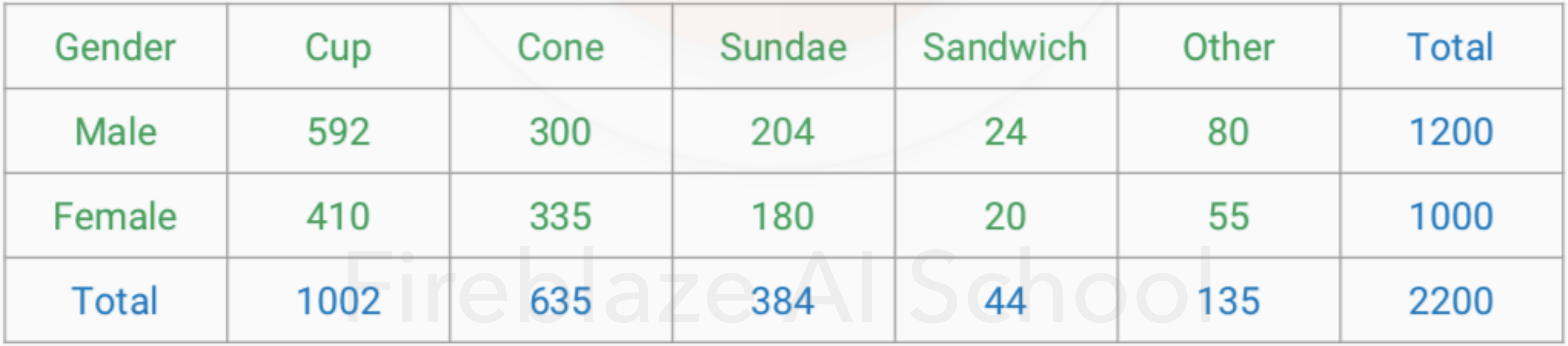
It is very similar to a frequency table. But the major difference is that a frequency table always concerns only one variable, whereas a contingency table concerns two variables.

To know the relationship between two ordinal or nominal variables then look for contingency table which displays this relationship.

Consider this sample, which shows gender and favourite way to eat ice cream.



Gender	Cup	Cone	Sundae	Sandwich	Other
Male	592	300	204	24	80
Female	410	335	180	20	55



Gender	Cup	Cone	Sundae	Sandwich	Other	Total
Male	592	300	204	24	80	1200
Female	410	335	180	20	55	1000
Total	1002	635	384	44	135	2200

Analysis

- There is 1002/2200 = 45.54% probability that the person prefers his ice cream in a cup.

- There is 24/1200 = 2% probability that a person prefers ice cream sandwich given that person is male.

These things are called Conditional proportion and Marginal proportion.

Univariate and Bivariate Data

➔ Univariate means “One Variable” (one type of data)

Example: Travel Time (in minutes): 15, 28, 9, 25, 36, 11, 45
The variable is Travel Time

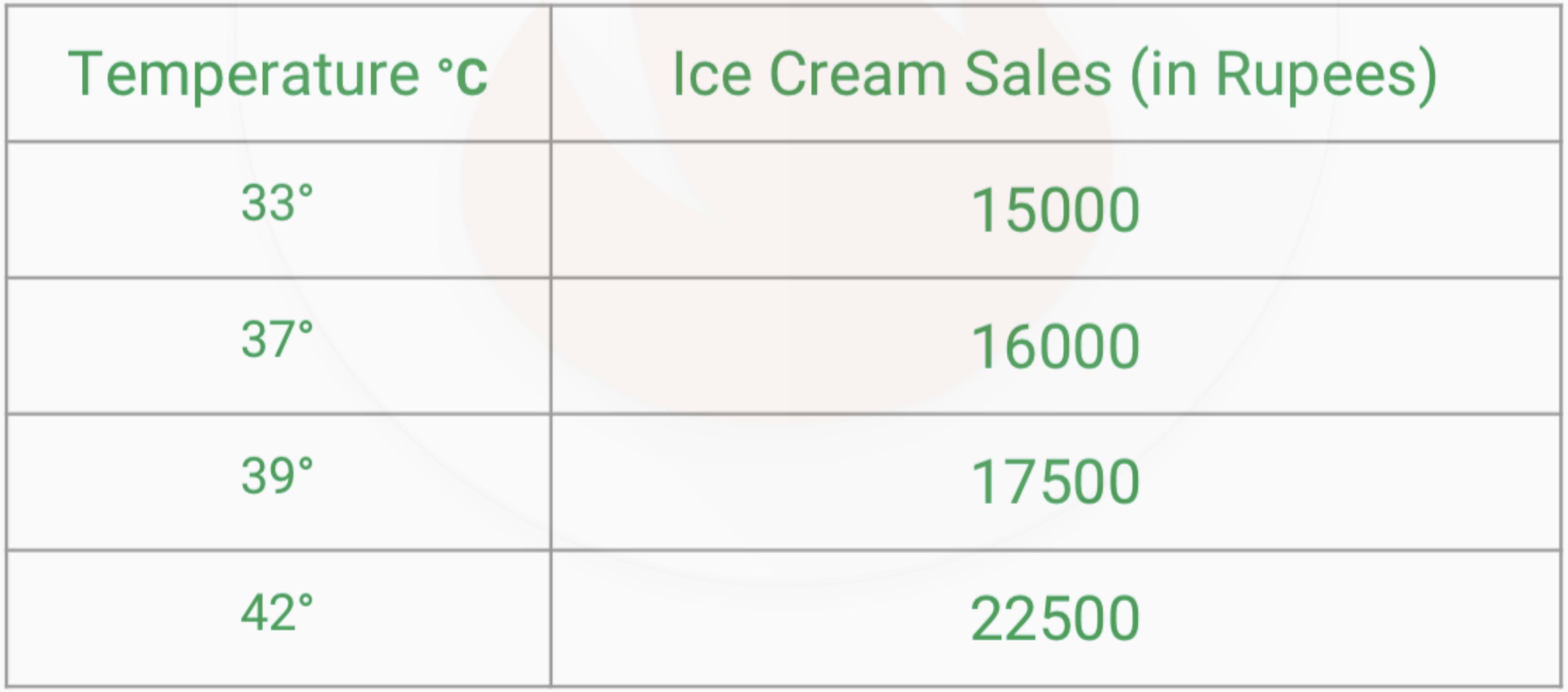
➔ Things to do with Univariate Data:

- Find a central value using mean, median and mode
- Find how spread out it is using range, quartiles and standard deviation
- Make plots like Bar Graphs, Pie Charts and Histograms

Univariate and Bivariate Data

➔ Bivariate means “Two Variables” (two types of data)

- Example: Here are two variables Ice Cream Sales and Temperature:

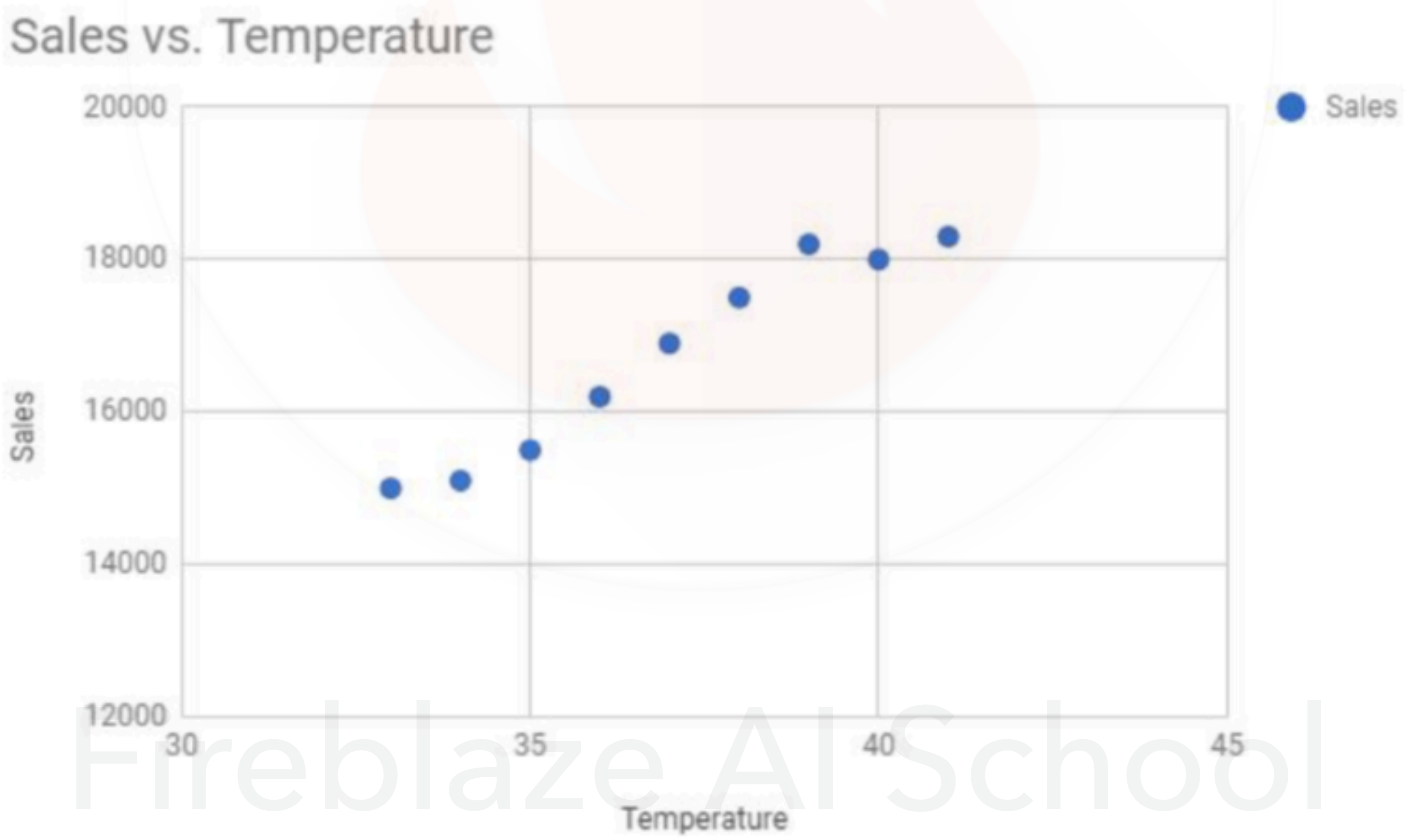


Temperature °c	Ice Cream Sales (in Rupees)
33°	15000
37°	16000
39°	17500
42°	22500

- With bivariate data: Comparing the two sets of data and finding any relationships. Use Tables, Scatter Plots, Correlation, Line of Best Fit to find out these relationships.

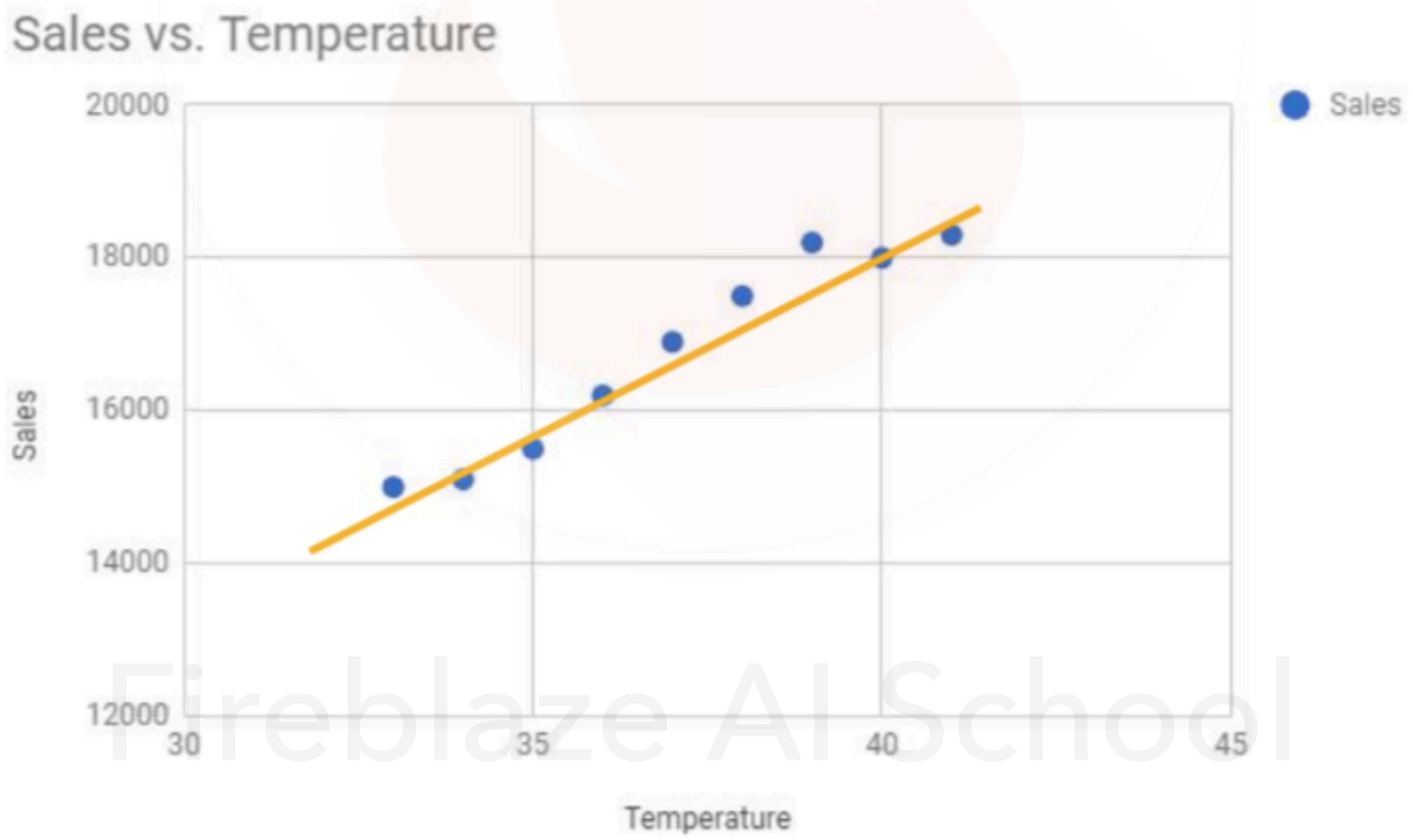
Scatter Plots

A Scatter(XY) Plot has points that show the relationship between two sets of data.



Line of Best Fit

The line as close as possible to all points and as many points above the line as below.



Correlation Coefficient

A contingency table is useful for nominal and ordinal variables, but not for quantitative variables. For quantitative variables, a scatterplot is more appropriate.

Scatter plot displays relation between two quantitative variables explanatory variable will be in X axis and Response variable will be in y axis.

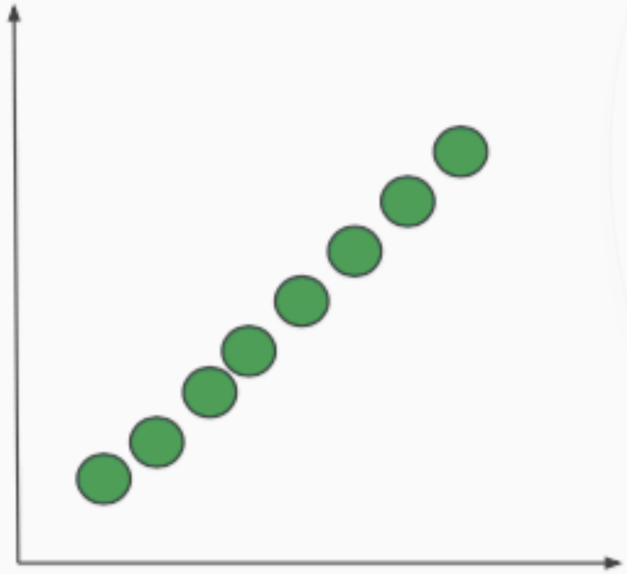
- Pearson’s r or Pearson Correlation: When two sets of data are strongly linked together, they have a High Correlation.

- The word Correlation is made of **Co**- (meaning "together"), and **Relation**
- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

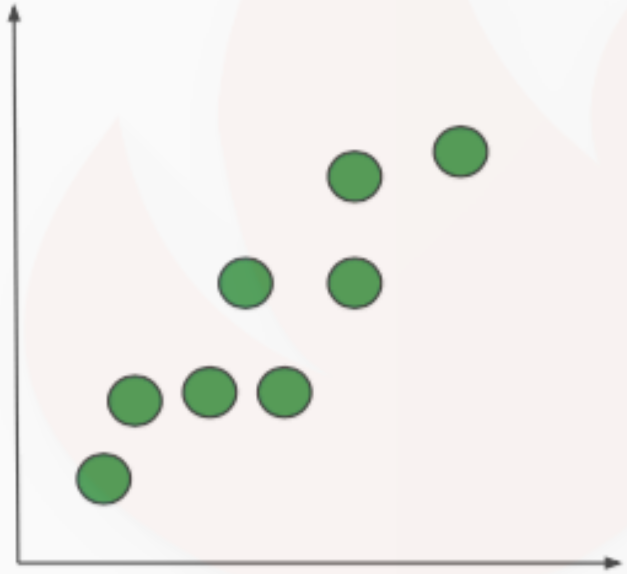
Formula: Pearson’s

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

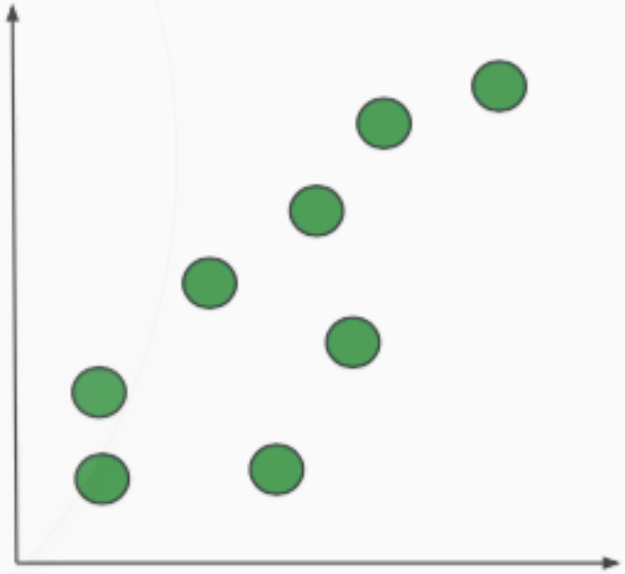
Scatter Plot



Perfect Positive Correlation
Value = 1

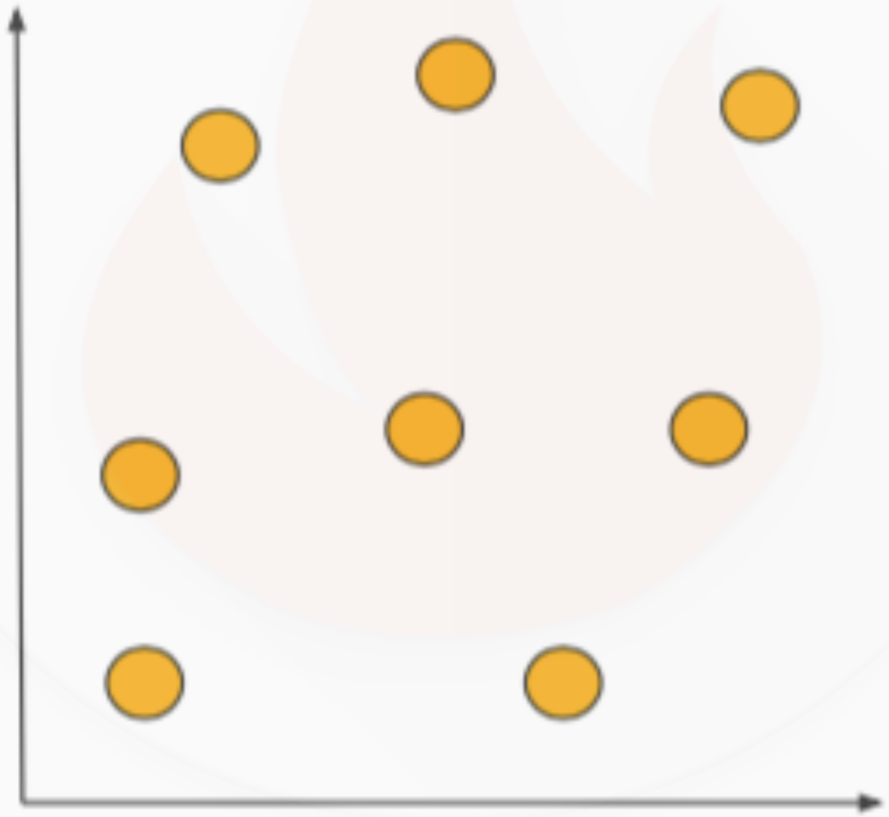


High Positive Correlation
Value = 0.9



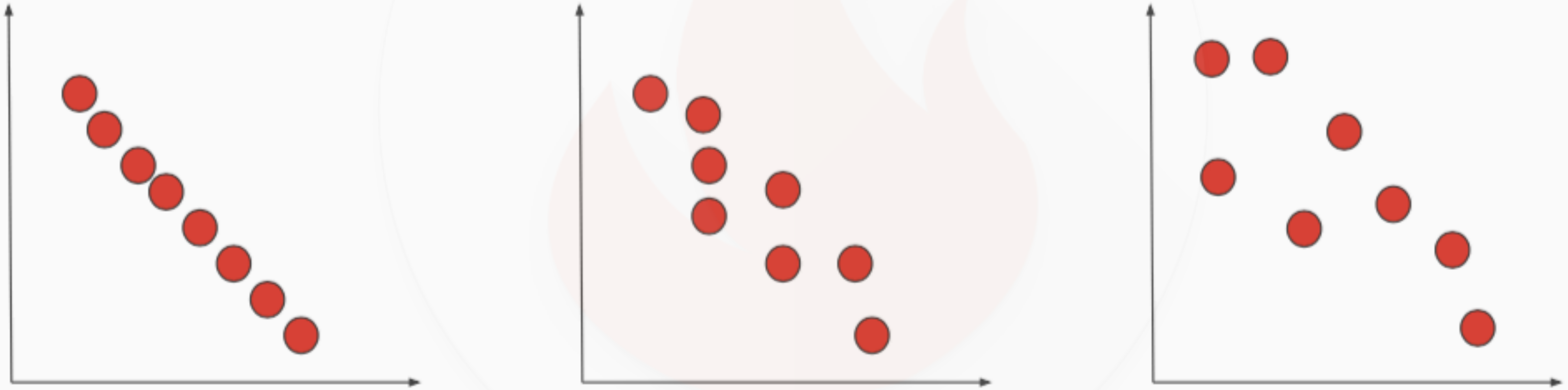
Low Positive Correlation
Value = 0.5

Scatter Plot



No Correlation
Value = 0

Scatter Plot



Perfect Negative Correlation
Value = -1

High Negative Correlation
Value = -0.9

Low Negative Correlation
Value = -0.5

Correlation Coefficient

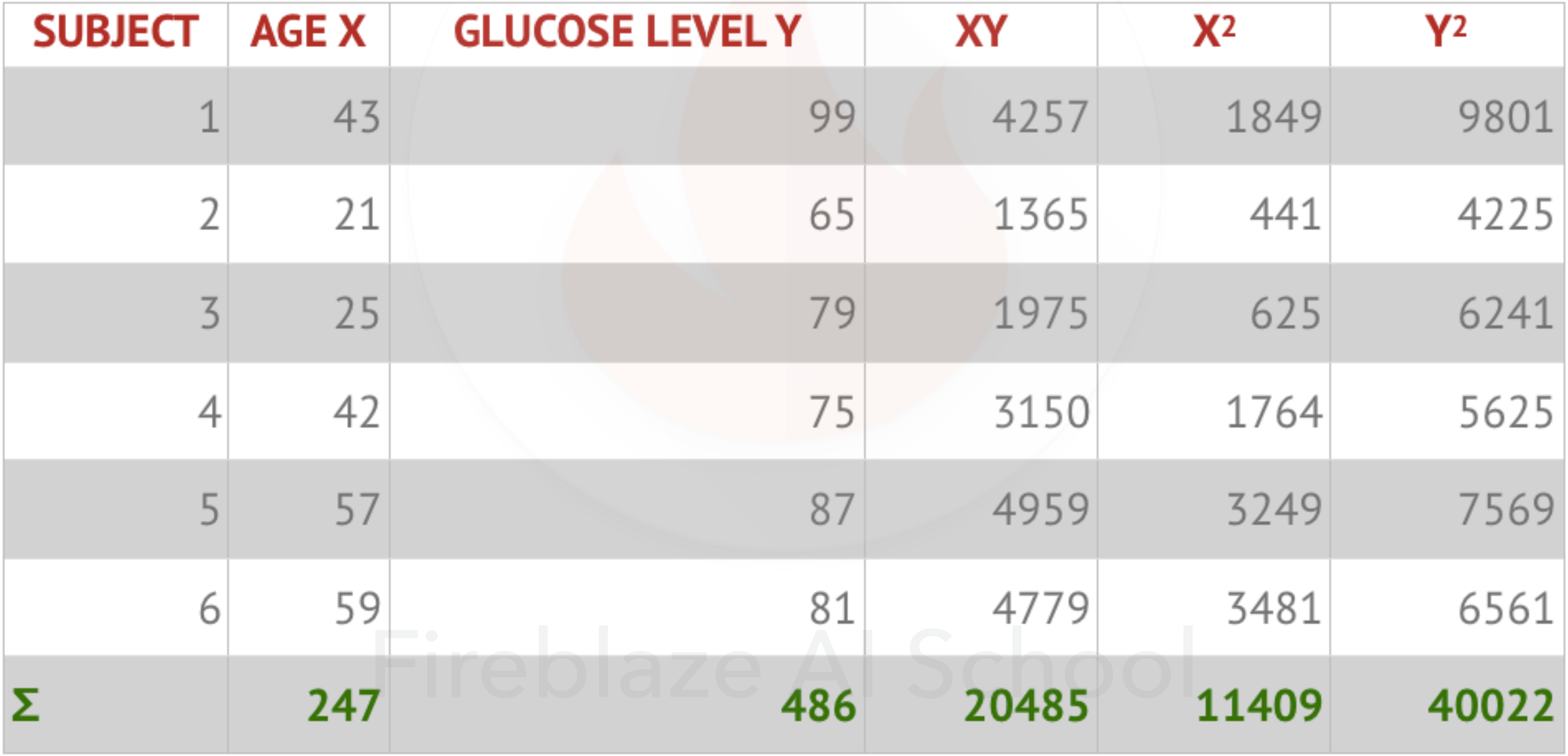
Correlation Coefficient can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation
- -1 is a perfect negative correlation
- The value shows how good the correlation is even if it is positive or negative.

Note: Correlation is not Causation

- "Correlation Is Not Causation" ... which says that a correlation does not mean that one thing causes the other (there could be other reasons the data has a good correlation).

Formula: Pearson's



SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	**247**	**486**	**20485**	**11409**	**40022**

Formula: Pearson’s

From our table:

$\Sigma x = 247$

$\Sigma y = 486$

$\Sigma xy = 20{,}485$

$\Sigma x^2 = 11{,}409$

$\Sigma y^2 = 40{,}022$

n is the sample size, in our case = 6

The correlation coefficient = 6(20,485) – (247 × 486) /

$[\sqrt{}[[6(11{,}409) - (247)^2] \times [6(40{,}022) - (486)^2]] = 0.5298$

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.

Fireblaze AI School

THANK YOU