1. **Shape, Center, Spread**: These are the basic characteristics of a data set. The shape refers to the overall pattern of the data distribution, the center refers to the middle value of the data set, and the spread refers to the variability in the data.
2. **Histogram**: A graphical representation of data using bars of different heights. It groups numbers into ranges and the height of each bar depicts the frequency of each range.
3. **Stem Plot**: Also known as a stem-and-leaf plot, it provides a visual distribution of a dataset.
    1. **Dot Plot**: A statistical chart consisting of data points plotted on a fairly simple scale. Dot plots are used for continuous, quantitative, univariate data.
4. **Boxplot**: A standardized way of displaying the dataset based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.
5. **Skewed Distribution**: A distribution that is not symmetric. This could be either positively skewed (long tail on the right) or negatively skewed (long tail on the left).
6. **Symmetric Distribution**: A distribution where the left-hand side is a mirror image of the right-hand side.
7. **Interquartile Range (IQR)**: A measure of statistical dispersion, or in simple terms, where most of your data lies in your dataset.
8. **Outliers**: An observation that lies an abnormal distance from other values in a random sample from a population.
9. **Scatter Plot**: A graphical display of statistical relationships with two variables plotted along two axes.
10. **Correlation**: A statistical measure that describes the size and direction of a relationship between two or more variables.
11. **Correlation Coefficient**: A numerical measure that quantifies the degree and direction of correlation.
12. **Regression Line**: A straight line that describes how a response variable y changes as an explanatory variable x changes.
13. **Standard Deviation**: A measure of the amount of variation or dispersion in a set of values.

SHAPE, CENTER, SPREAD

They are fundamental concepts in statistics that help us understand and describe data

CENTER

it represents the typical or representive value in a dataset, we can measure it in several ways, it's like having a bunch of apples and choosing the one that represents the average apple (mean), or the middle apple (median) or the apple that shows up a lot (mode)

Mean ($\bar{x}$) is the average $\frac{sum}{count}$ or $\bar{x} = \frac{\sum x}{n}$

Median (M / $\tilde{x}$) is when you arrange the values and take the middle

2, 4, 6, 7, 8 -> 6 is the median

1,2,3,4,5,6 -> 3.5 is the median

Its denoted as $\tilde{x}$

Mode () is the number that repeated the most

1, 4, 6, 1, 6, 1, 1, 5 -> 1 is the mod

It is referred to as a sample mode

SHAPE

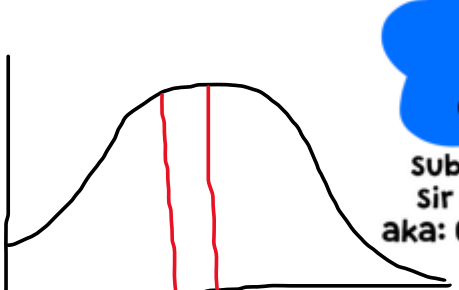
it tells you the appearance of the data point distribution, gives you trends and patterns

The relation between shape and center

Mean is $\frac{\Sigma(each\ value \times its\ frequency)}{total\ frequency}$

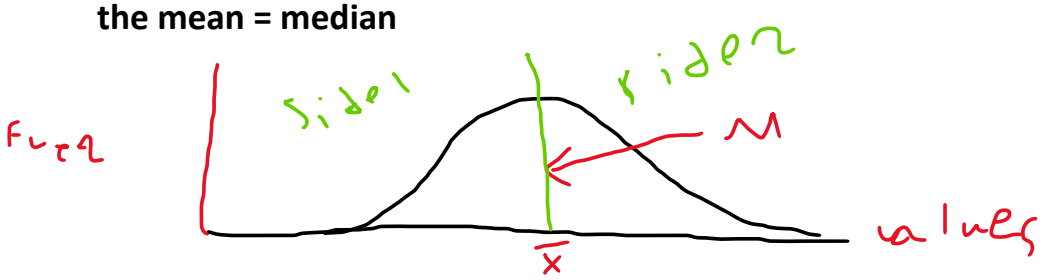
Median is the middle value of a dataset when it's arranged in ascending or descending order

Mode is the value with the highest frequency



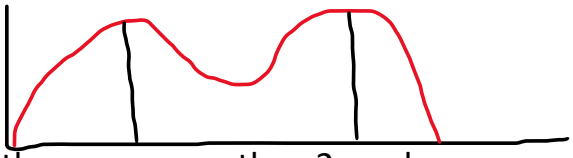
like

- **Symmetric** when both sides are equal or approximability similar, here **the mean = median**

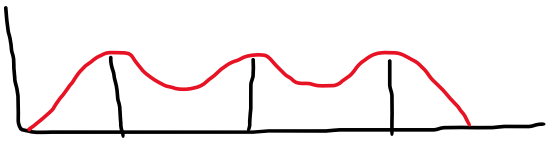


- **Unimodal** when there is only one mode
  the mean = median = mode

- **Bimodal** when ther are 2 modes



- **Multimodal** when there are more than 2 modes

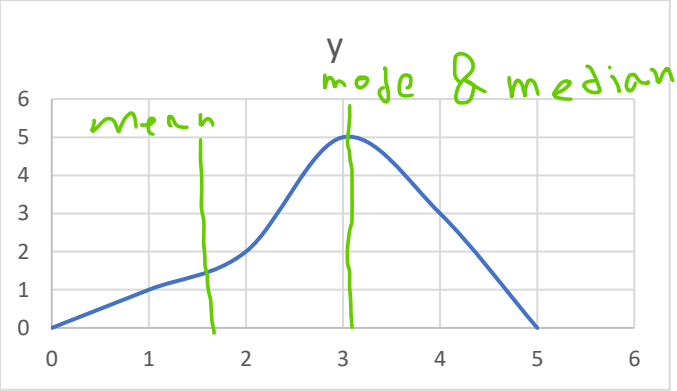


- **Uniform** when all values have very similar frequencies

- **Right tailed is probably Right Skewed** **Left tailed is probably Left Skewed**

for example, being skewed to the left means that your **mean** is to the **left** of your **median** & **mode**

x	0	1	2	3	4	5
y	0	1	2	5	3	0



Your mean is 1.7, ur mode and median are 3

When you're right tailed or right skewed **the mean > the median**

And if you're left tailed or left skewed **the mean < the median**

Spread

it tells us the variability/versitality in the dataset and the difference between the max and min values, it's like seeing the range of colors an apple can get, it has 3 key elements

Range is the range from the highest to lowest value so

range = max – min

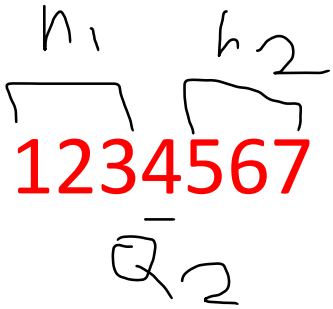
Interquartile Range (IQR) is the thing that measures the range within the middle 50% of the values, it calculated as the difference between the first quartile or quarter (25%) and third quartile or quarter (75%)

IQR = $Q_3 - Q_1$

Q_3 the median of the second half
Q_1 the median of the first half
so if the values were 123456

Q1 = median first half = 2

Q3 = median second half = 5
IQR = Q3 – Q1 = 3



so if the values were 1234567

Q1 = median first half = 2

Q2 = the whole median = 4

Q3 = median second half = 5
IQR = Q3 – Q1 = 3

Standard Deviation (SD / σ) measures the average difference between the values and the mean (average value)

$$SD = \sqrt{\frac{\sum|x - \mu|^2}{N}}$$

You calculate the SD using this formula, let's explain it

$\sum$ means the sum of
x is a value in the dataset

μ is the mean of the dataset

N is the number of values

so what this thing means is that you get the sum of all the differences between each value and the mean in the dataset, then you divide

them by the number of values and then you take the square root

how to solve step by step:

- Get the mean
- Go across each value In the data set and get the difference between it and the mean
- Combine all the values
- Divide them by the number of values
- Take the square root

Variance is the average of the squared differences from the mean

It is like σ but remove the root

$$\boldsymbol{Varience = \frac{\sum|x-\mu|^2}{N}}$$

Coefficient of variation is the ratio of the Standard deviation to th mean

CV = SD / μ

Coefficient of Skewness:

- **Equation:** $\frac{3\times(\text{Mean}-\text{Median})}{\text{Standard Deviation}}$
- **Interpretation**: Skewness measures the asymmetry of the probability distribution of a dataset around its mean. Positive skewness indicates a right-leaning tail (right-skewed), while negative skewness indicates a left-leaning tail (left-skewed).

Measures of Dispersion:

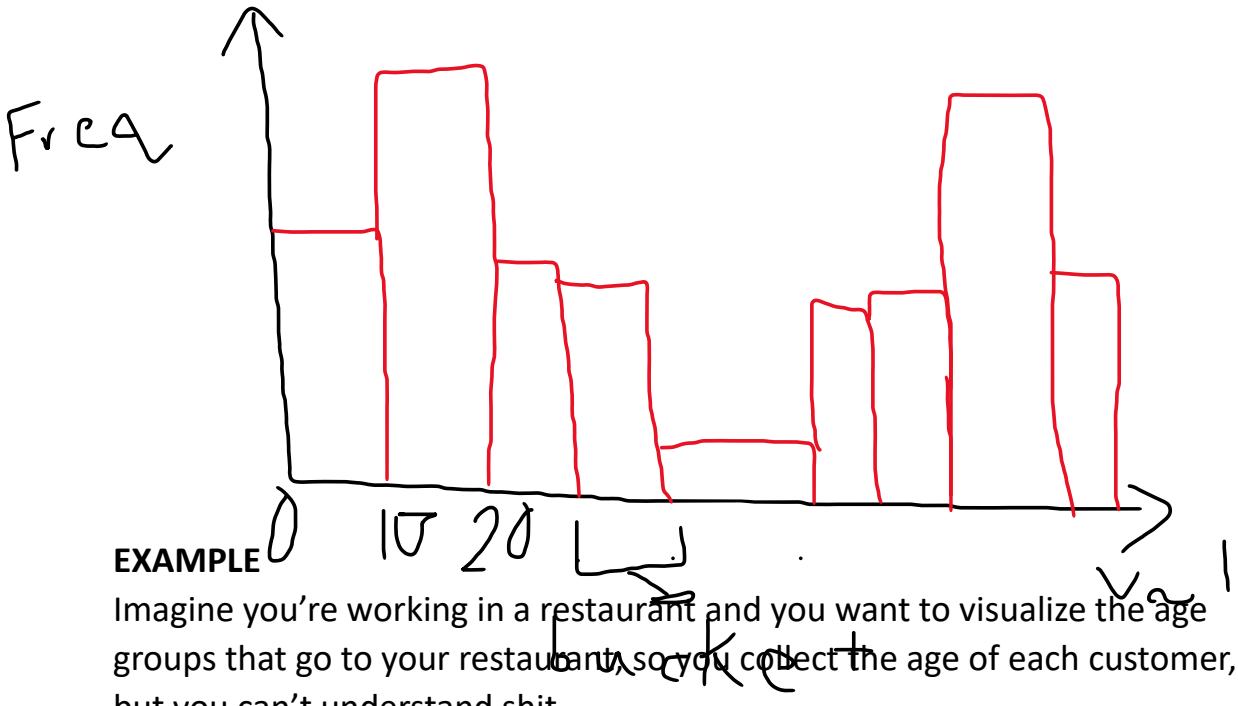
- **Interpretation:** These measures describe the spread or dispersion of data points around the mean. Variance and standard deviation indicate the average distance of data points from the mean. Range and IQR give a sense of the spread of values from minimum to maximum and within the middle 50% of the data, respectively.

Algebraic Sum of Deviations:

- **Equation:** $$\sum(X_i - \bar{X}) = 0$$
- **Interpretation:** The sum of deviations of individual data points from the mean is always zero. This property is fundamental in statistical calculations and demonstrates the balancing effect of deviations above and below the mean in a dataset.

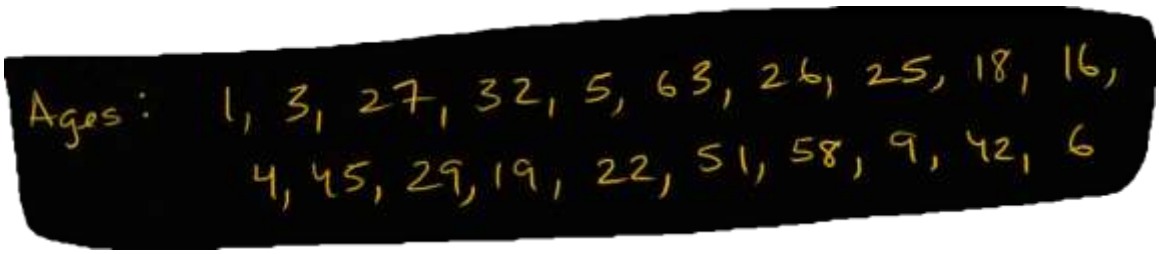
HISTOGRAMS

It's a graphical display of data using bars of different height

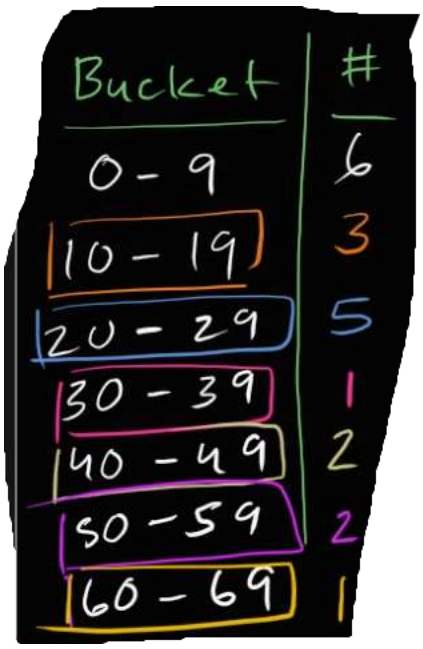


EXAMPLE

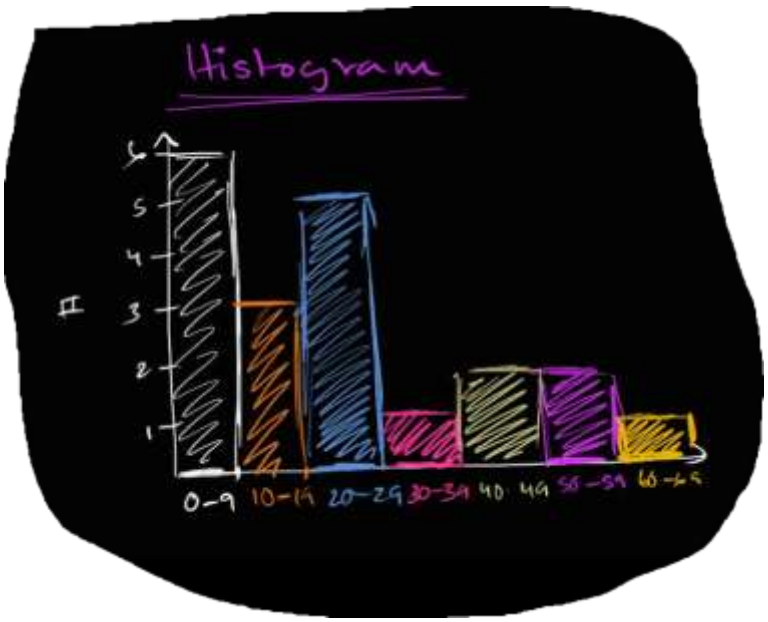
Imagine you're working in a restaurant and you want to visualize the age groups that go to your restaurant so you collect the age of each customer, but you can't understand shit



So now we turn those ages into categories (**buckets**) and numbers



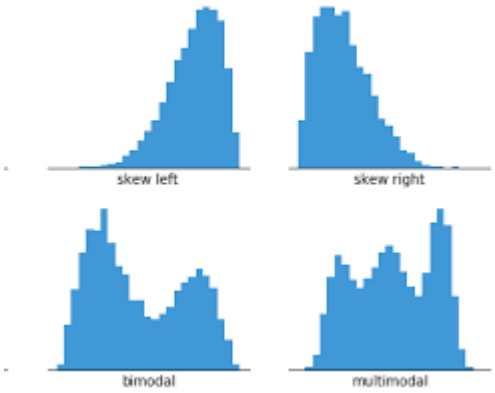
And then we take the buckets as the values and numbers as the frequencies



Histogram is used when visualizing multiple groups of data so instead of doing how many 1 and 2 and 3 and so on so on year olds like dot plots, we put them into buckets/groups and do it that way

Sure, here are the four types of histograms with photos under each:

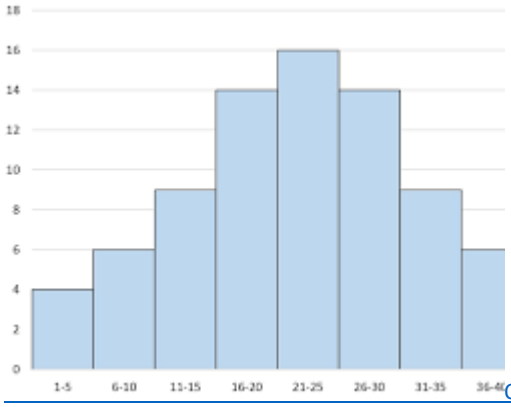
1. Uniform histogram



Opens in a new window

A uniform histogram is a histogram that has all of its bars the same height. This means that the data is evenly distributed across all of the bins. Uniform histograms are often used to represent data that is randomly distributed.

2. Symmetric histogram

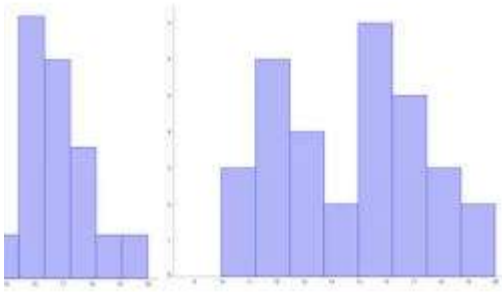


Opens in a new window

A symmetric histogram is a histogram that is the same on both sides of the center. This means that the data is evenly distributed around the mean. Symmetric histograms are often used to represent data that is normally distributed.

3. Bimodal histogram

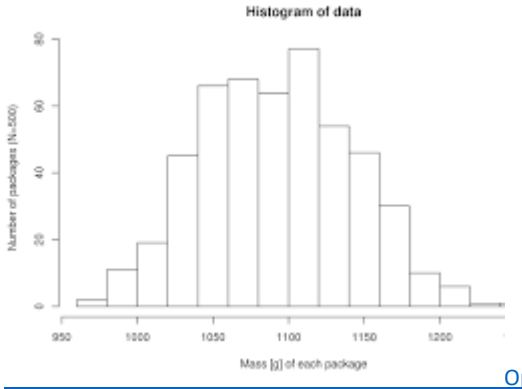
imodal Histogram



Opens in a new window

A bimodal histogram is a histogram that has two peaks. This means that the data is not evenly distributed, but instead has two clusters of values. Bimodal histograms are often used to represent data that has two distinct groups.

4. Probability histogram



Opens in a new window

A probability histogram is a histogram that shows the probability of each value in the data set. This means that the height of each bar is equal to the probability of that value occurring. Probability histograms are often used to represent the distribution of a random variable.

STEM PLOT

It’s a special table where each data value is split into a "stem"

(the first digit or digits) and a "leaf" (usually the last digit). Like in this **example:**

Median order = (N+1)/2

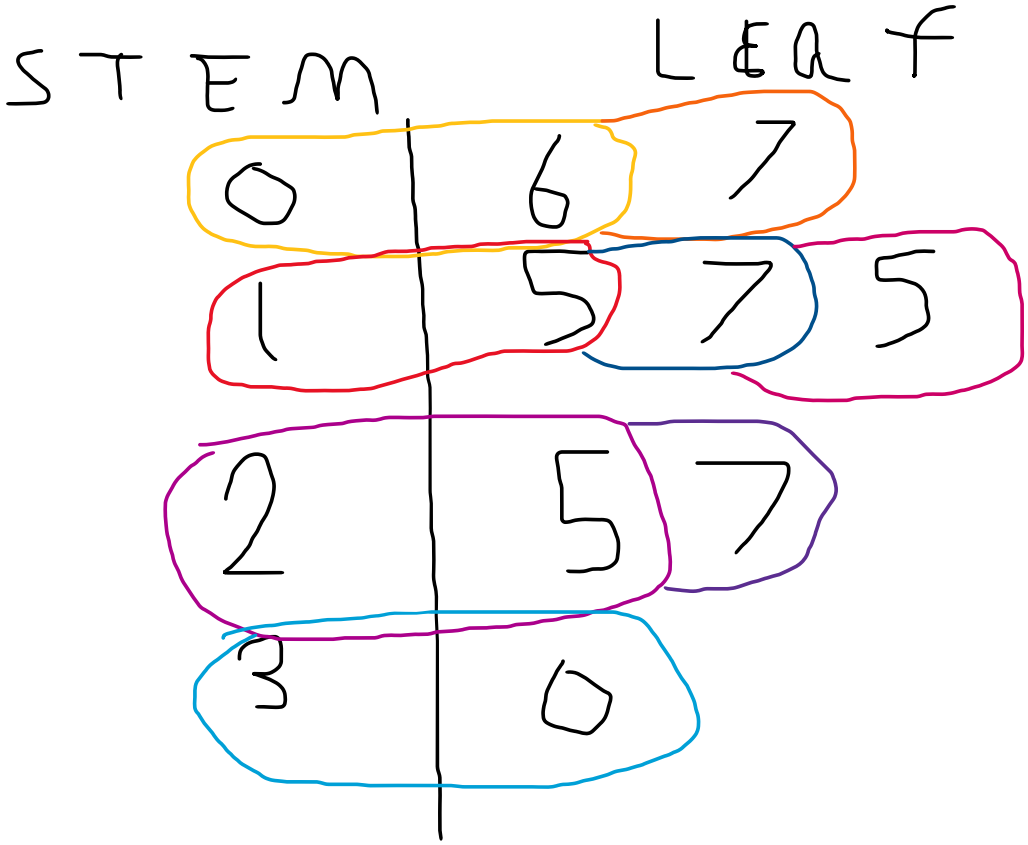
32 is split into "3" (stem) and "2" (leaf).

132 is split into "13" (stem) and "2" (leaf).

8 is split into "0" (stem) and "8" (leaf).

24.85 -> 2485 is split into "248" (stem) and "5" (leaf).

So let's say we have 6, 15, 25, 7, 17, 15, 27, 30

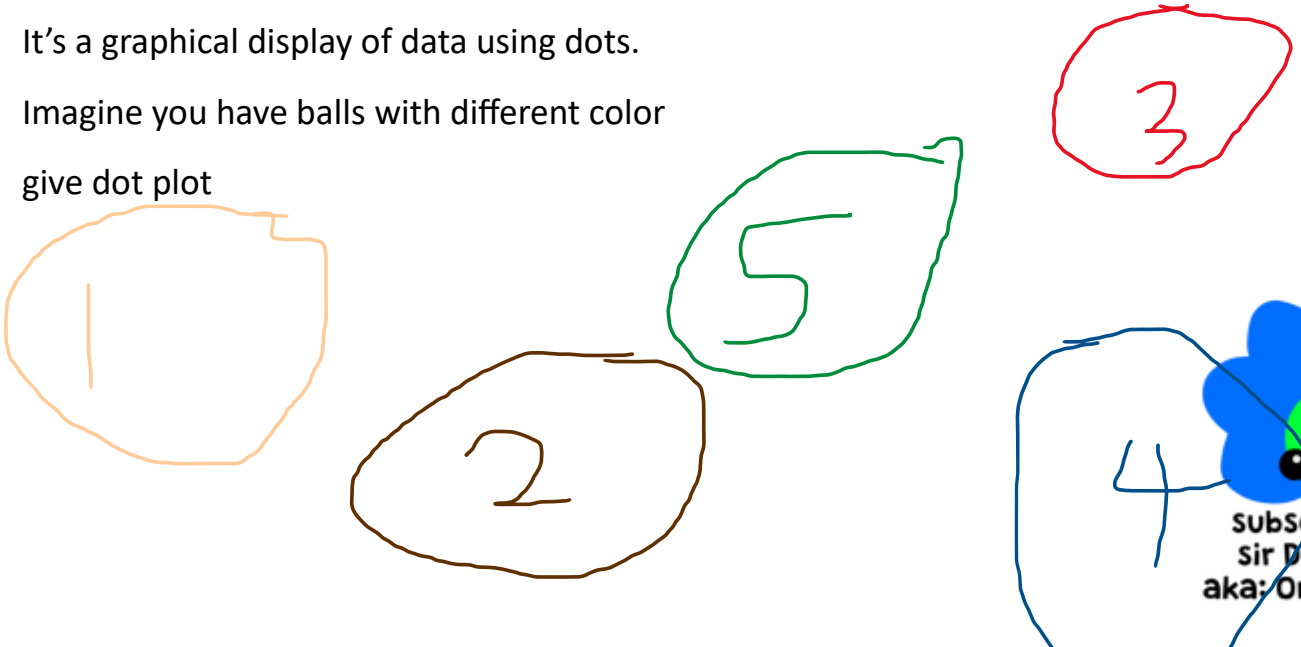


DOT PLOT

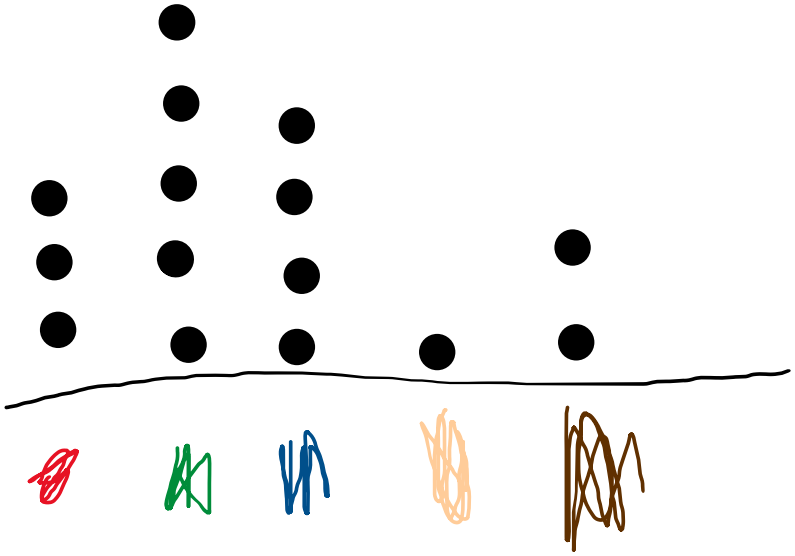
It's a graphical display of data using dots.

Imagine you have balls with different color

give dot plot



Firstly, you arrange them into a value frequency table



Here, the mode is green, the mean is between green & blue, and the median is green

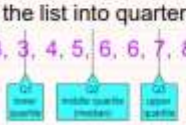
Quartiles are the values that divide a list of numbers into quarters:
-Put the list of numbers in order
-Then cut the list into four equal parts
-The Quartiles are at the "cuts"

Example: 5, 7, 4, 4, 6, 2, 8
Put them in order: 2, 4, 4, 5, 6, 7, 8
Cut the list into quarters:

Quartiles of 2, 4, 4, 5, 6, 7, 8
And the result is:
Quartile 1 (Q1) = 4
Quartile 2 (Q2), which is also the Median, = 5
Quartile 3 (Q3) = 7

Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8
The numbers are already in order

Cut the list into quarters:



Quartiles

In this case Quartile 2 is half way between 5 and 6:
Q2 = (5+6)/2 = 5.5
And the result is:
Quartile 1 (Q1) = 3
Quartile 2 (Q2) = 5.5
Quartile 3 (Q3) = 7

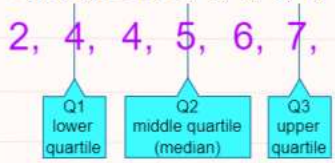
Q1 is the median of the first half

Q2 is the median of everything

Q3 is the median of the last half

Interquartile Range (IQR)
The "Interquartile Range" is from Q1 to Q3:
Interquartile Range

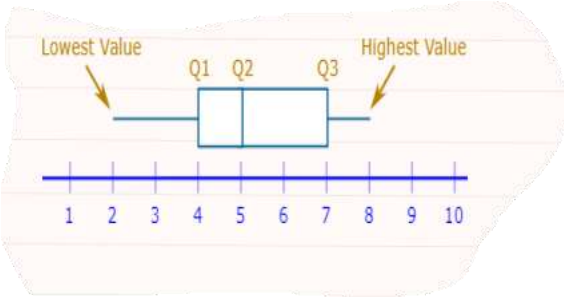
To calculate it just subtract Quartile 1 from Quartile 3
Example:
Quartiles of 2, 4, 4, 5, 6, 7, 8



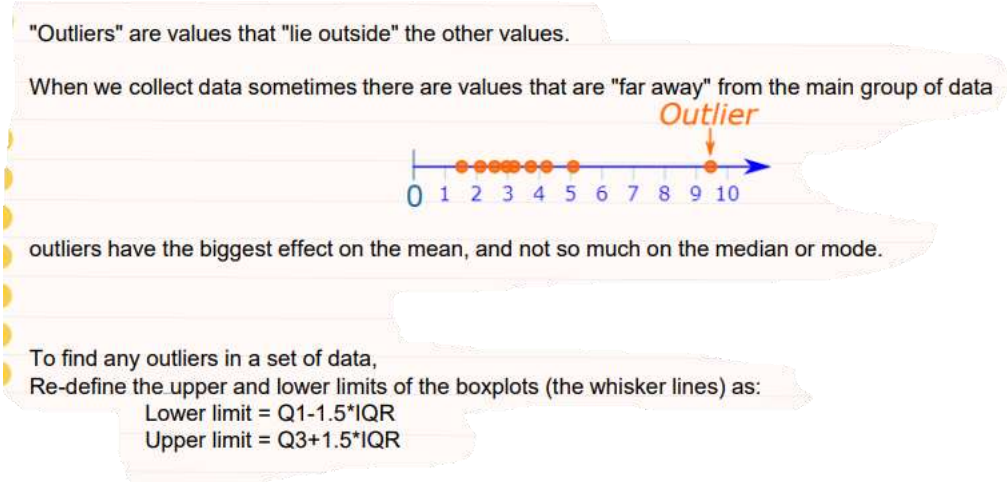
The Interquartile Range is:
Q3 − Q1 = 7 − 4 = 3

BOX PLOT

It shows important values



You need these 5 values to do a box plot



Outliers are areas that are far than the main cluster of data, they affect the mean, not the median nor mode

To define if a value is an outlier in a box plot

You have to get the Limits

$$\textbf{Down limit} = Q_1 - 1.5 * IQR$$
$$\textbf{Up Limit} = Q_2 + 1.5 * IQR$$

[Down Limit , Up Limit]

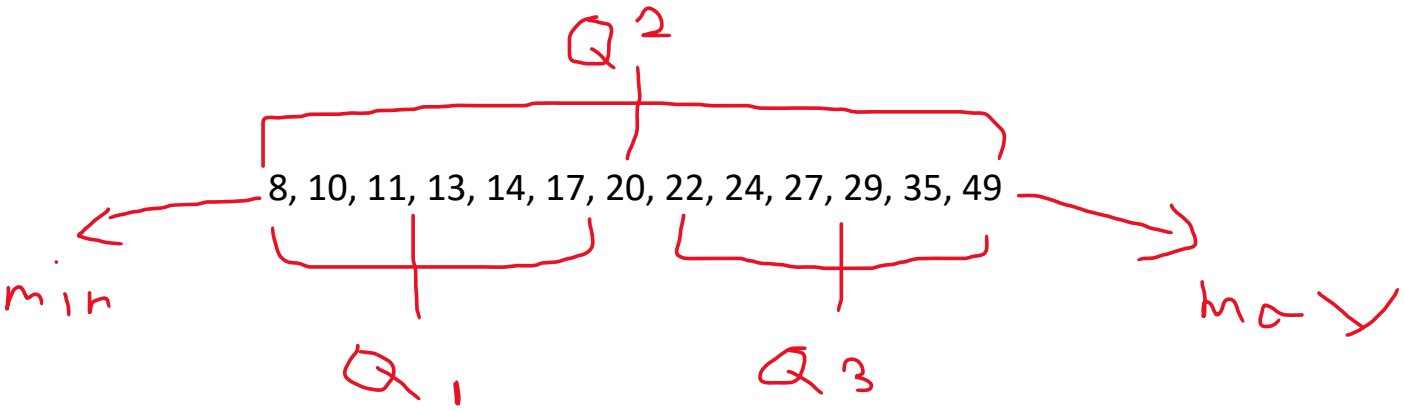
If a value is outside that range, then it is an **outlier**

EXAMPLE 1 -> no outlier

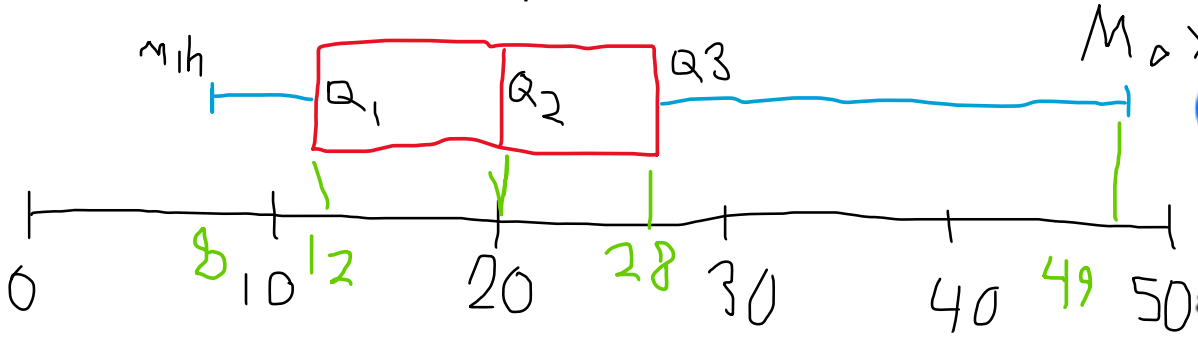
Let's say you have these

11, 22, 20, 14, 29, 8, 35, 27, 13, 49, 10, 24, 17

1. Arrange them



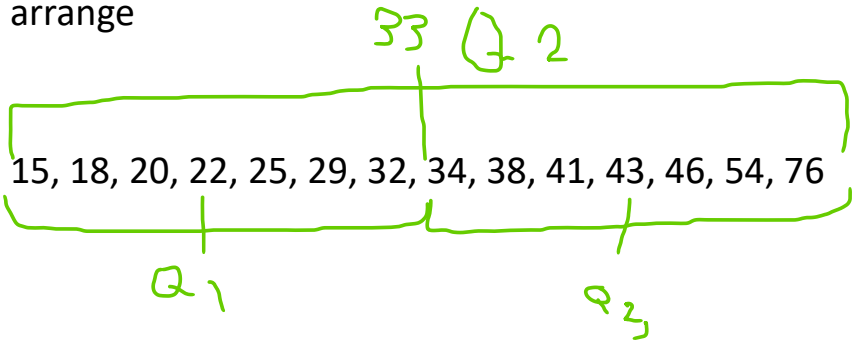
2. Q_1 = median of half 1 = 11+13/2 = 12
   Q_2 = median of all = 20
   Q_3 = median of half 3 = 27+29/2 = 28
   IQR = $Q_3 - Q_1$ = 16
   min = 8
   max = 49
3. Check if the min and max are not outliers, because if they are, they will exist out of the box & whisker plot
4. Determine the limits
   down limit = Q1 – 1.5 * IQR = -12
   up limit = Q3 + 1.5 * IQR = 52
5. The min and max fall into the limits so they are not outliers
6. Draw the box & whiskers plot



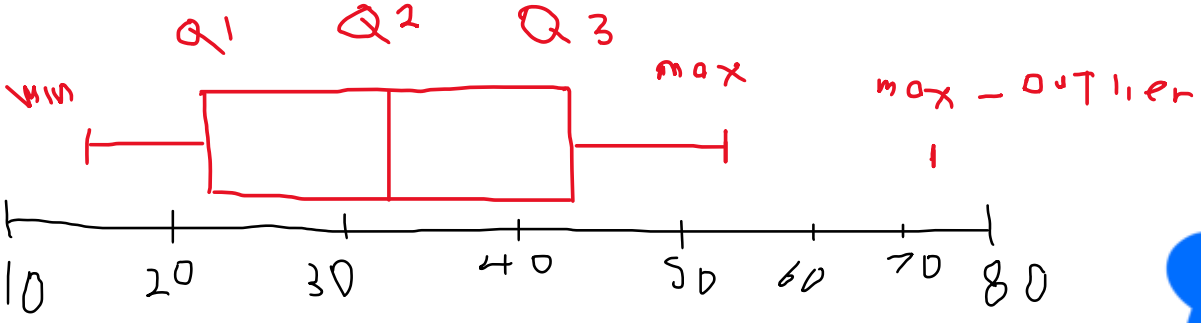
EXAMPLE 2 -> outlier

18, 34, 76, 29, 15, 41, 46, 25, 54, 38, 20, 32, 43, 22

1. arrange

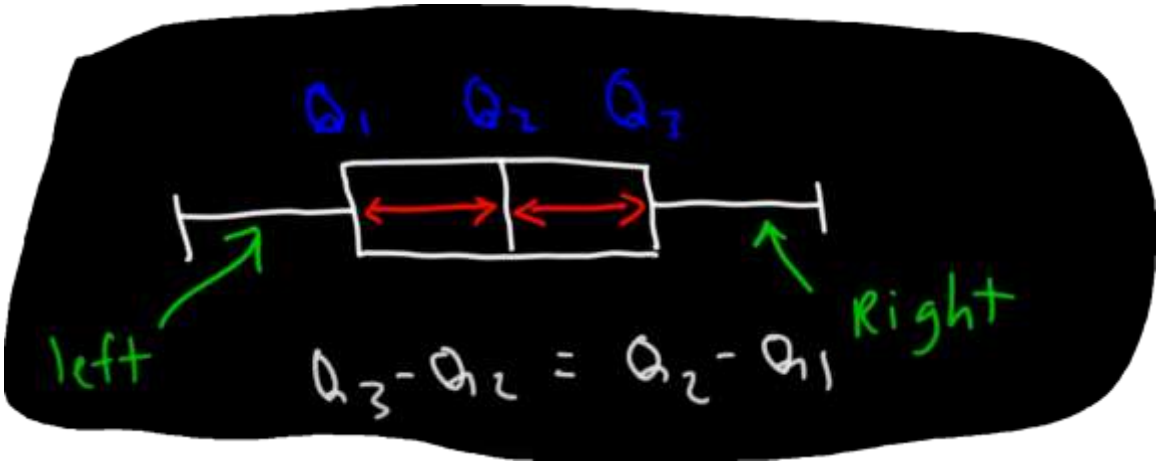


2. get values
   Q2 = median of all = 33
   Q1 = median of half 1 = 22
   Q3 = median of half 2 = 43
   min = 15
   max = 76
   IQR = Q3 – Q1 = 21
   Up limit = 43 + 1.5 x 21 = 74.5
   Down limit = 22 – 1.5 x 21 = -9.5
3. as you can see, the limits are [-9.5, 74.5] and the max is out of bounds so it is an outlier, so the maximum not outlier value is 54
4. draw

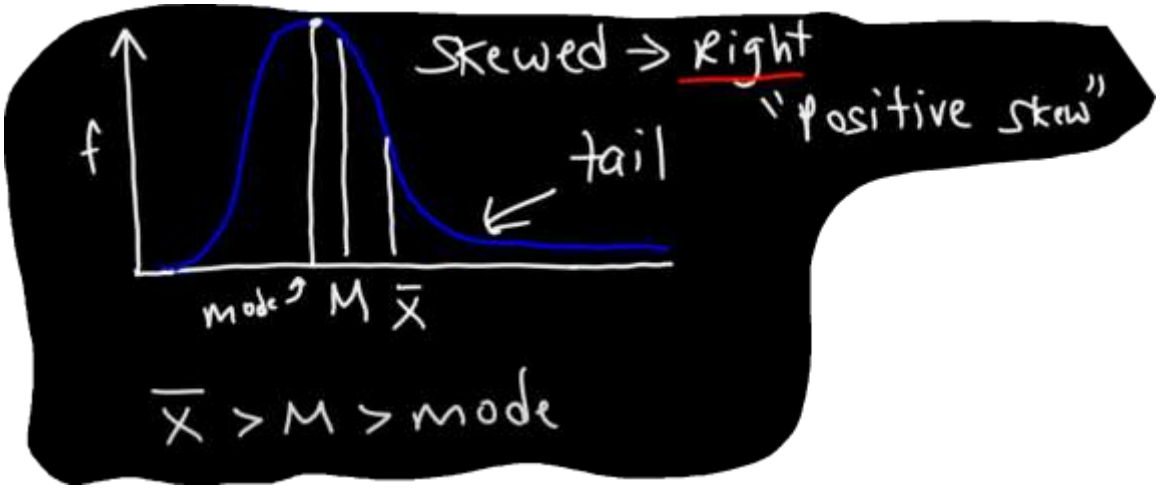


$$\text{Coefficient of Skewness} = \frac{Q3+Q1-2\times\text{Median}}{Q3-Q1}$$

DISTRIBUTIONS

- 

symmetric

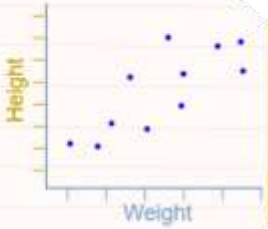
- 

skewed to the right

SCATTER PLOT

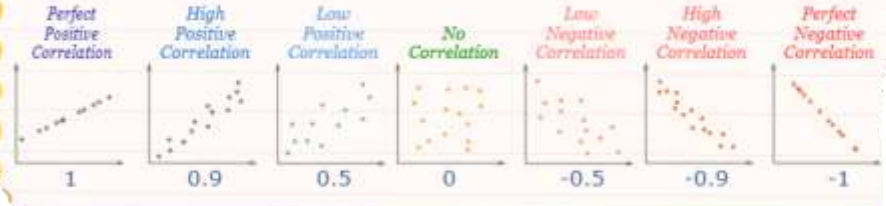
A Scatter (XY) Plot has points that show the relationship between two sets of data.

Correlation
When the two sets of data are strongly linked together we say they have a High Correlation.



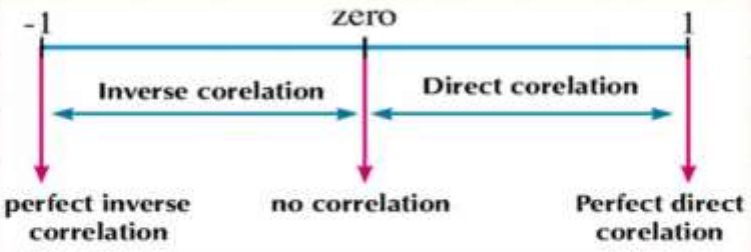
The word Correlation is made of Co- (meaning "together"), and Relation

Correlation is Positive when the values increase together, and
Correlation is Negative when one value decreases as the other increases
Like this:



CORRELATION COEFFICIENT

A correlation coefficient (r) is a number between -1 and 1 that tells you the strength and direction of a relationship between variables.



If r = Zero then its no correlation between two variables.
If $0 < r < 0.4$ then its weak correlation.
If $0.4 \leq r < 0.6$ then its intermediate correlation.
If $0.6 \leq r < 1$ then its strong correlation.
If r = I then its perfect correlation.

How to calculate the correlation coefficient

1 – Pearson's correlation

Step 1	Find the mean of x, and the mean of y
Step 2	Subtract every x from the value mean of x (call them "a"), and subtract every y value from the mean of y (call them "b")
Step 3	Calculate ab, a^2, and b^2 for every value
Step 4	Sum up ab, sum up a^2, and sum up b^2
Step 5	Divide the sum of ab by the square root of [(sum of a^2) × (sum of b^2)]

$$r = \frac{n \times \sum(xy) - \sum(x) \times \sum(y)}{\sqrt{n \times \sum(x^2) - \sum(x)^2} \times \sqrt{n \times \sum(y^2) - \sum(y)^2}}$$

Where

- $\sum(x)$ is the sum of x
- $\sum(y)$ is the sum of y
- $\sum(xy)$ is the sum of (x*y)
- $\sum(x^2)$ is the sum of x^2
- $\sum(y^2)$ is the sum of y^2
- N is the number of elements

2 – Spearman correlation

You turn the x and y to their ranks “from highest to lowest” so if x is 5.2, 9, 3, 4, the R(x) is going to be 2, 1, 4, 3
do the same for y and then get the difference for each x amd its respective y (d_i)

Then square each difference and combine them

Then do this equation

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

Here observation mean each x and its respective y

EXAMPLE 1

1. Calculate the correlation coefficient between the two variables x and y shown below:

X:	1	2	3	4	5	6
Y:	2	4	7	9	12	14

Step 1 : make a table with x, y, xy, x^2, y^2

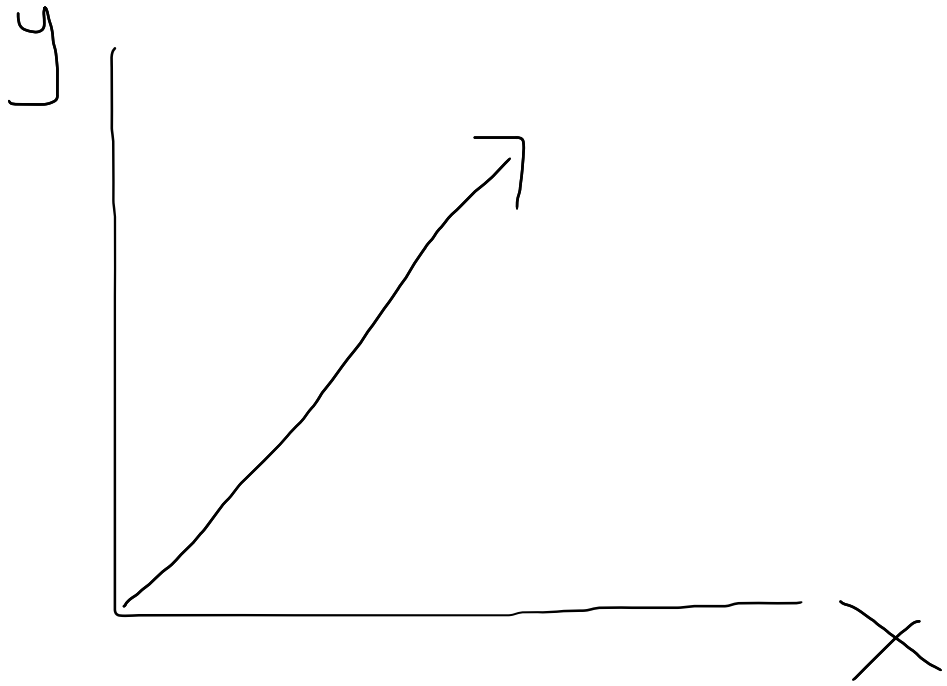
x	y	xy	x^2	y^2
1	2	2	1	4
2	4	8	4	16
3	7	21	9	49
4	9	36	16	81
5	12	60	25	144
6	14	84	36	196
21	**48**	**211**	**91**	**490**

Use this

$$r = \frac{n \times \sum(xy) - \sum(x) \times \sum(y)}{\sqrt{n \times \sum(x^2) - \sum(x)^2} \times \sqrt{n \times \sum(y^2) - \sum(y)^2}}$$

$$R = \frac{6*211-21*48}{\sqrt{6*91-(21)^2}*\sqrt{6*490-(48)^2}} = .998$$

This indicates a very strong direct/linear relationship between x and y



EXAMPLE 2

1. Calculate the correlation coefficient between the two variables x and y shown below:

X:	1	2	3	4	5	6
Y:	2	4	7	9	12	14

Step 1 : make a table with x, y, r(x), r(y), d_i, d_i^2

x	y	R(x)	R(y)	d	d^2
1	2	6	6	0	0
2	4	5	5	0	0
3	7	4	4	0	0
4	9	3	3	0	0
5	12	2	2	0	0
6	14	1	1	0	0
21	**48**			**0**	**0**

Usethis

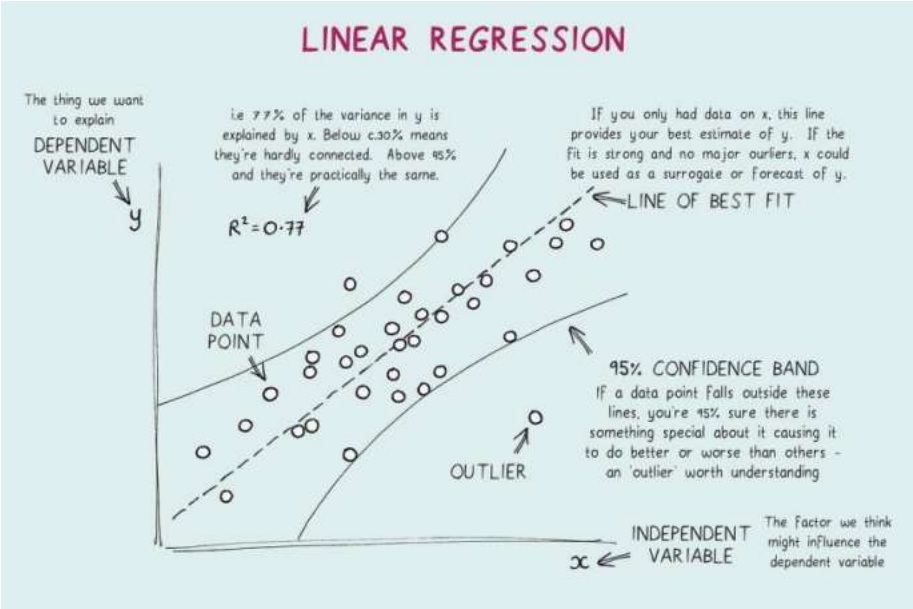
$$p = 1 - \frac{6 \times \Sigma(d^2)}{n(n^2 - 1)}$$

$$P = 1 - \frac{6*0}{6(6^2-1)} = 1$$

Regression line

displays the connection between

scattered data points in any set



Regression line equation:

$$y = a + bx$$

$$b = \frac{n \times \sum(xy) - \sum(x) \times \Sigma(y)}{n \times \sum(x^2) - (\sum x)^2}$$

$$a = \frac{\Sigma y - b \Sigma x}{n}$$

The regression line equation of y on x is used for:

1- predicting the value of Y if the value of X is known.

2- identifying the error which can be identified by the relation:

Error = | Table value - the value satisfying the regression equation |