1. Descriptive Statistics

Descriptive statistics involves summarizing and organizing the data so they can be easily understood.

Types of Descriptive statistics?

Descriptive statistics are broken down into two categories. **Measures of central tendency** and **Measures of variability (spread).**

Measure of Central Tendency

Central tendency refers to the idea that there is one number that best summarizes the entire set of measurements, a number that is in some way “central” to the set.

- **Mean / Average**
  Mean or Average is a central tendency of the data i.e. a number around which a whole data is spread out. In a way, it is a single number which can estimate the value of whole data set.
  Let’s calculate mean of the data set having 8 integers.

$$x = \frac{12+24+41+51+67+67+85+99}{8} = 55.75$$

- **Median**
  Median is the value which divides the data in 2 equal parts i.e. number of terms on right side of it is same as number of terms on left side of it when data is arranged in either ascending or descending order.
  Median will be a middle term, if number of terms is odd
  Median will be average of middle 2 terms, if number of terms is even.

$$12+24+41+51+67+67+85+99 = 59$$

  The median is 59 which will divide set of numbers into equal two parts. Since there are even numbers in the set, the answer is average of middle numbers 51 and 67.

- **Mode**
  Mode is the term appearing maximum time in data set i.e. term that has highest frequency.

$$12, 24, 41, 51, 67, 67, 85, 99$$

  In this data set, mode is 67 because it has more than rest of the values, i.e. twice.

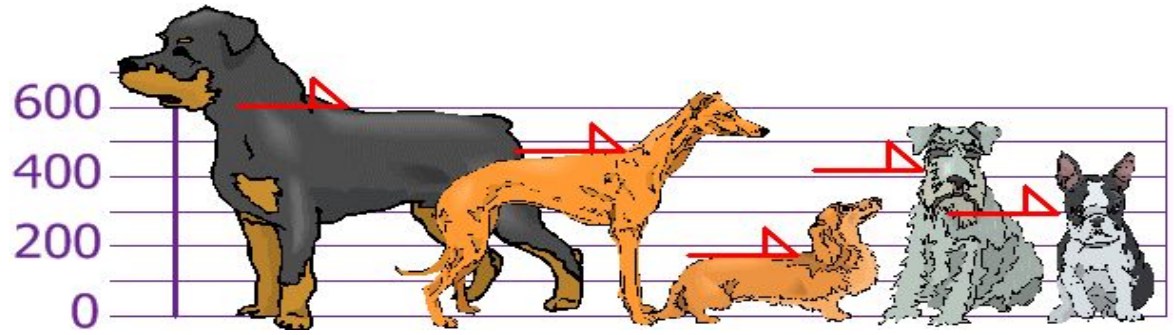
Measure of Spread

Measure of Spread refers to the idea of variability within your data.

- **Standard Deviation**
  The Standard Deviation is a measure of how spread out numbers are.
  Its symbol is σ (the greek letter sigma)
  The formula is easy: it is the square root of the Variance.
- **Variance**
  The Variance is defined as:
  The average of the squared differences from the Mean.

Example:

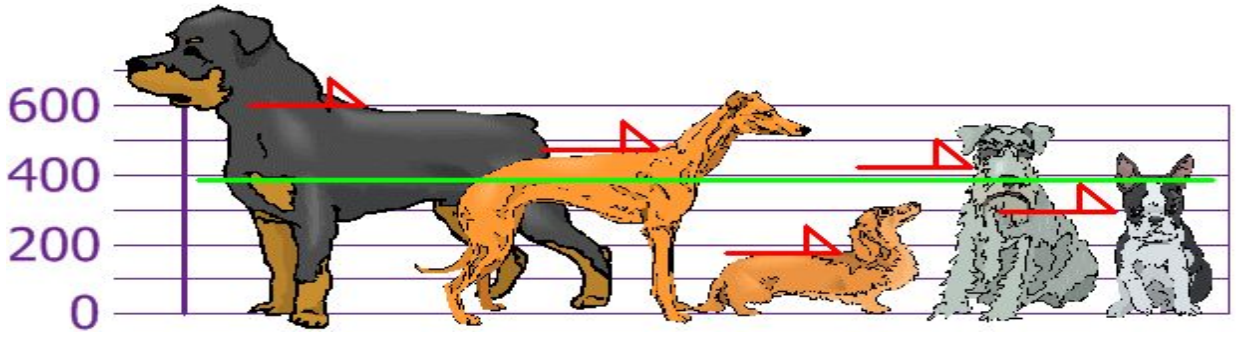
You and your friends have just measured the heights of your dogs (in millimeters):



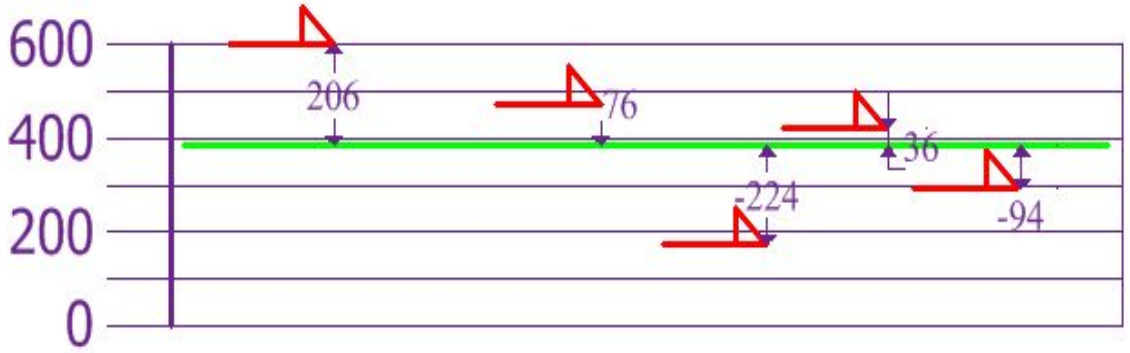
The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.
Find out the Mean, the Variance, and the Standard Deviation.
Your first step is to find the Mean:

Mean = (600 + 470 + 170 + 430 + 300)/5
= 394

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



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



To calculate the Variance, take each difference, square it, and then average the result:
Variance(σ^2) = (206^2+ 76^2+ $(-224)^2$+ 36^2+ $(-94)^2$)/5

= 108520/5
= 21704

So the Variance is **21,704.**

And the Standard Deviation is just the square root of Variance, so:

Standard Deviation

σ = √21704
= 147.32...
= 147

Standard Deviation is **147**.

- **Mean Deviation / Mean Absolute Deviation**
  Mean Deviation:
  Find the mean of all values ... use it to work out distances ... then find the mean of those distances!
  1. Find the mean of all values
  2. Find the distance of each value from that mean (subtract the mean from each value, ignore minus signs)
  3. Then find the mean of those distances

Example:

The Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

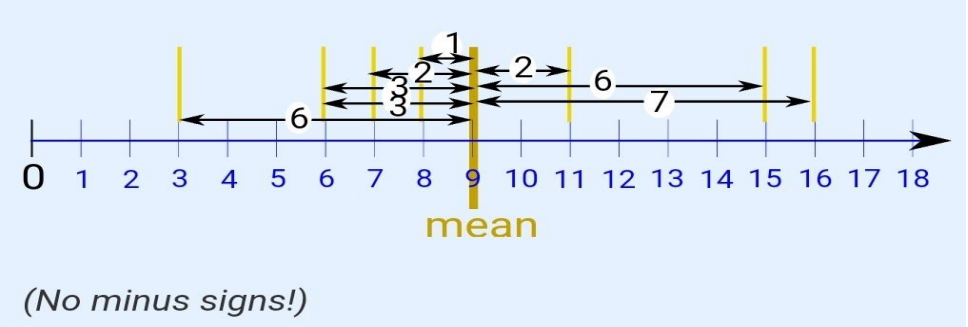
Step 1: Find the **mean**:

Mean = (3 + 6 + 6 + 7 + 8 + 11 + 15 + 16)/8
= 72/8 = 9

Step 2: Find the **distance** of each value from that mean:

Values	Distance from 9
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7

Which looks like this:



Step 3. Find the **mean of those distances**:
Mean Deviation = (6 + 3 + 3 + 2 + 1 + 2 + 6 + 7)/8 = 30/8 = 3.75
So, the **mean= 9**, and the **mean deviation = 3.75.**
For **deviation** just think **distance**

Formula:

$$\text{Mean Deviation} = \frac{\Sigma|x - \mu|}{N}$$

- **Σ** is Sigma, which means to sum up
- || (the vertical bars) mean Absolute Value, basically to ignore minus signs
- **x** is each value (such as 3 or 16)
- **μ** is the mean (in our example **μ = 9**)
- **N** is the number of values (in our example **N = 8**)

Mean Absolute Deviation:
Each distance we calculate is called an Absolute Deviation, because it is the Absolute Value of the deviation (how far from the mean).
To show "Absolute Value" we put "|" marks either side like this:
|-3| = 3
Absolute Deviation = |x - μ|
From our example, the value 16 has Absolute Deviation = |x - μ| = |16 - 9| = |7| = 7
And now let's add them all up …
Sigma:
The symbol for "Sum Up" is Σ (called Sigma Notation), so we have:

Example:
The Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

Step 1: Find the **mean**:

Mean = (3 + 6 + 6 + 7 + 8 + 11 + 15 + 16)/8

= 72/8 = 9

Step 2: Find the **distance** of each value from that mean:

Values	|x - μ|
3	6
6	3
6	3
7	2
8	1
11	2
15	6
16	7
	Σ|x - μ| = 30

Step 3. Find the **Mean Deviation:**
Mean Deviation = Σ|x - μ|N = 308 = 3.75

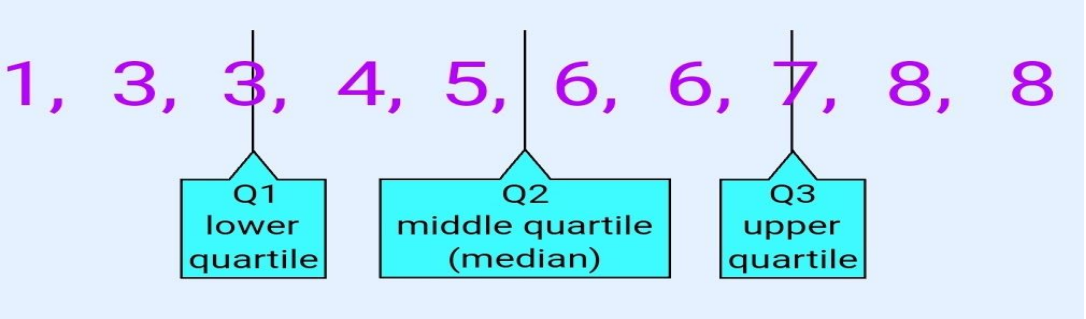
Note: the mean deviation is sometimes called the Mean Absolute Deviation (MAD) because it is the mean of the absolute deviations.

- **Range**
  The difference between the lowest and highest values.
  In {4, 6, 9, 3, 7} the lowest value is 3, and the highest is 9, so the range is 9 − 3 = 6.

- **Percentile**
  Percentile is a way to represent position of a values in data set. To calculate percentile, values in data set should always be in ascending order.
  12,24,41,51,67,67,85,99.
  The median 59 has 4 values less than itself out of 8. It can also be said as: In data set, 59 is 50th percentile because 50% of the total terms are less than 59.

- **Quartiles**
  Another related idea is Quartiles, which splits the data into quarters:
  Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are in order. Cut the list into quarters:



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