01 CORE DATA CONCEPTS

- **Data** = Collected observations and information on something.
- Data can be continuous vs. discrete (infinite values between data points?), nominal vs. ordinal (can we order them), structured vs. unstructured (is it tabular?), population vs. sample (does it cover all individuals?).
- **Central tendency** : "Where is the center of this data set?"
  - **Arithmetic Mean** : The balancing point.

$$\mu = \frac{\sum_{i=1}^{N} (x_i)}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} (x_i)}{n}$$

  - **Weighted Mean**: Used when some data points contribute more to the final average:

$$W = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - **Median**: When we order the data set, median is the value in the middle (helpful when data is skewed).

$$Med(X) = \begin{cases} X\left[\frac{n+1}{2}\right] & \text{if } n \text{ is odd} \\ \frac{X\left[\frac{n}{2}\right] + X\left[\frac{n}{2}+1\right]}{2} & \text{if } n \text{ is even} \end{cases}$$

  - **Mode** : The value that occurs the most often in the data set.
- **Dispersion** : "How much is the data spread out around the mean?"
  - Standard Deviation $= \sqrt{\text{Variance}}$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

  - Why choose standard deviation over variance?
    - Because the units of σ and s match the units of the data.

- * Why divide by N for population but by $(n-1)$ for sample?
  - * To try to unbalance the bias we introduced when we minimized the deviation in our sample by calculating $\bar{x}$. ($\bar{x}$ is biased because it depends on the sample we use.)

- * Five-Number Summary : Gives us an idea of the center and the spread of our data set at one glance.

Min.	Q_1	Median (Q_2)	Q_3	Max.
Minimum value	25% percentile	50th percentile	75th percentile	Maximum value

- * IQR : Interquartile Range $= Q_3 - Q_1$

- * Outlier : A data point that differs significantly from others. One common formula is:

$$\text{Outlier}(x) = \begin{cases} \text{True} & \text{if } x < Q_1 - 1.5 \times \text{IQR} \text{ or } x > Q_3 + 1.5 \times \text{IQR} \\ \text{False} & \text{otherwise} \end{cases}$$