**Fundamentals of Data Science**

**Student Name: Muhammad Ahmed**

**Student id: 22031216**

## **Dataset Description:**

The dataset consisted of newborn babies' weights born in some regions of Europe. There are 400 values provided; however, the calculation metric (i.e., kg or pound) is unknown. The goal is to perform different statistical operations to understand the story of data.

## **Data Distribution:**

According to distribution, we have the following statistics:

* The minimum weight of a baby is 2.35, while the maximum value recorded is 4.47
* The average weight of babies is 3.40.
* According to the plotted distribution, many babies (high frequency) are born with weights ranging from 3.25 to 3.75. Thus, in that European region, there is a concentration of weights with a specific range, which can help understand the qualities of these newborns

## **Calculating W:**

According to the information, W is the average weight of babies. The average represents a dataset's expected or typical value, a measure of central tendency. The standard is calculated as follows:

From the given distribution, the value obtained is: **3.406**

## **Calculating X:**

In the given dataset, "X" represents a value such that 10% of the newborns have weights above X. It is like a threshold weight that separates the top 10% of newborns in terms of their weights. We use the cumulative distribution function (CDF), which tells us the proportion of data points that are less than or equal to a given value. X is the value where the CDF reaches 90% (or 0.90), indicating that 10% of the newborn weights in the dataset are above X. In simpler terms, X is a weight value that marks the cutoff for the dataset's heaviest 10% of newborns. The value obtained is **2.88**.

## **Plotting X and W**

A histogram is a graphical representation of the distribution of data. It is an estimate of the probability distribution of a continuous variable. Histograms represent the data by dividing it into intervals and showing how many values fall into each interval. Histograms can be useful in determining the shape of the distribution, identifying outliers, and examining the spread of the data.

The provided code creates a histogram of a given array of newborn weights. It sets the histogram to be normalized, so that it represents the density of the data. It also adds a label to the histogram, sets its color and opacity, and defines the number of bins for the data. The code also adds two vertical lines to the plot, one indicating the mean of the data and another indicating the 10th percentile.

The code uses the Matplotlib library, which is a data visualization library for Python. It is a popular tool for creating various types of graphs, charts, and plots, including histograms.

