**Case Study - Probability Distribution for Data Science**

**Introduction**

Akshay is a researcher working on the health data obtained from different Australian athletes. He wants to know how various factors interact with one another in different athletes. His firm wants to market sports beverages to these athletes, and therefore, he wants to understand what are the most common and uncommon health factors (like haemoglobin, RBC count, BMI, etc.) in athletes.

Akshay also has another dataset from which he wants to analyze the income of athletes. Akshay also wants to find both the most probable combinations of factors and outliers in the data to help his firm decide which features to target while creating the sports beverage.

**Problem Statement**

Akshay decided to use probability distributions to find the outliers and other useful information about the athletes while also understanding the spread of the dataset. He wanted to use basic distributions such as normal distributions, binomial distributions, uniform distributions, etc. Akshay has another dataset from which he hopes to examine the earnings of different athletes. Akshay also wants to uncover the outliers in the data as well as the most likely combinations of elements.

**Example Dataset**

Below is the example of the given dataset, which contains various characteristics of the athlete's blood varied with sport, body size, and sex.

**rcc** - Red Blood Cell Count, in 1012-1

**Wcc** - White Blood Cell Count, in 1012 per liter

**Hc** - Hematocrit, percent

**Hg** - Haemoglobin Concentration, in g per decaliter

**Ferr** - Plasma Ferritins, ng dl-1

**Bmi** - Body Mass Index, kg cm-2102

**ssf** - sum of skin folds

**pcBfat** - Percent Body Fat

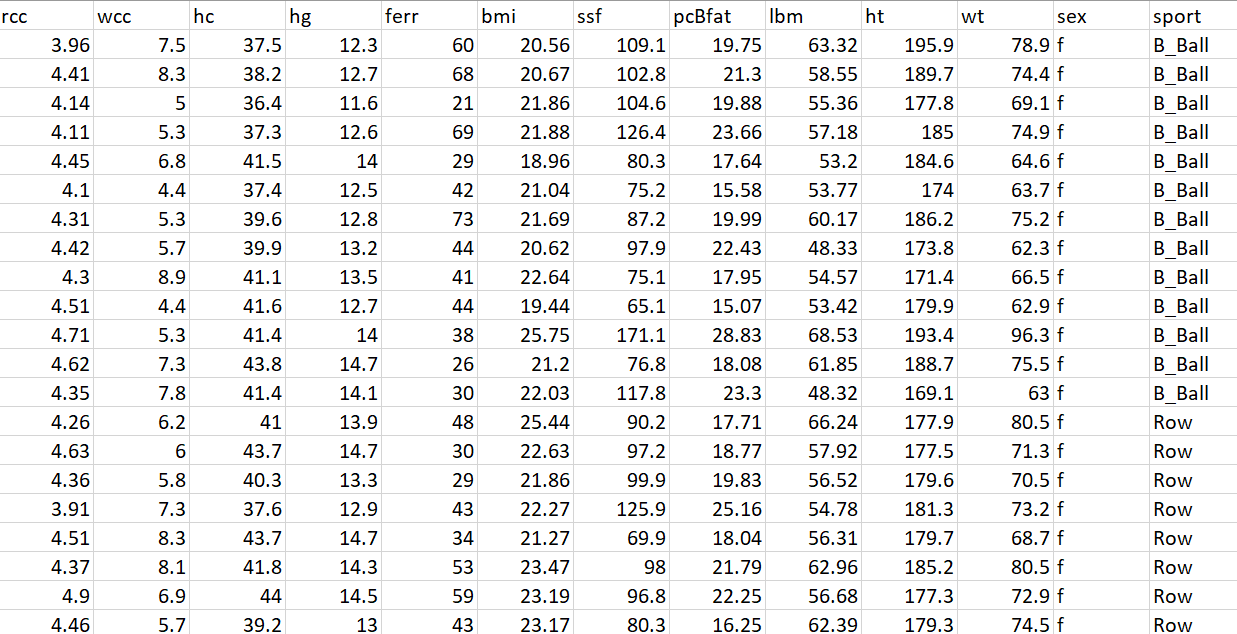
**Lbm** - Lean Body Mass, kg

**Ht** - Height, cm

**Wt** - Weight, kg

**Sex** - a factor with levels f m

**Sport** - This includes a variety of sports fields (B\_Ball, Row, Netball, Swim, Field, T\_400m, T\_Sprnt, Tennis, Gym, and W\_Polo).

****

**Case Study Questions**

Analyze the dataset using the following distributions:

1. Normal distribution
2. Uniform distribution
3. Binomial distribution
4. Bernoulli distribution
5. Poisson distribution
6. Exponential distribution

**Next steps**

* Identifying the outliers and resolving them.
* You can observe how the random variables in a specific dataset are behaving.
* You can fit a random variable to generalize its behavior.
* Selection of the Machine Learning models.