horizontal line

**Emir Dogan**

Finding and Cleaning the Data

**10th October 2024**

# Finding the Data

The dataset for this analysis was sourced from Kaggle, specifically from a project by [Omar Sobhy](https://www.kaggle.com/datasets/omarsobhy14/redbull-and-heart-rates/data). The dataset includes valuable insights related to the consumption of Red Bull and its effects on heart rate, collected through a structured survey methodology. Key details about the dataset are as follows:

* **Provenance**: The data was collected via face-to-face interviews.
* **Sample Size**: The original study had a sample size of 281 volunteers; however, our analysis is based on a subset of 120 participants.
* **Demographics**: The survey was conducted on Egyptian volunteers, providing a specific cultural context.
* **Survey Questions**: The survey included questions regarding participants' sex, age group, and before-and-after blood pressure readings.
* **Data Collection Period**: The data was collected over a period of approximately two months.
* **Collection Methodology**: The survey was conducted under open database licensing, ensuring accessibility and transparency.

# Cleaning the Data

* We confirmed that the dataset comprises 120 participants, with no missing values present. The absence of missing data is a crucial factor for maintaining the integrity of our analysis.
* We will check for the presence of outliers in the dataset using Python code. This will help us ensure that extreme values do not unduly influence our results. The method for outlier detection will involve statistical techniques such as the IQR method or Z-score analysis, which will be executed in the data analysis phase.

# 