**Lab1 - Initial Stages of Data**

**Data Collection:**  
The dataset used for this analysis was obtained from Kaggle and is titled "Sports Data Analysis." This dataset focuses on FIFA players and provides comprehensive information about their various skills and attributes, making it particularly valuable for machine learning applications. It offers a wealth of data that can be used for conducting in-depth analyses and creating machine learning models related to football player performance.

Data link: [FIFA Data for EDA and Stats](https://www.kaggle.com/datasets/mukeshmanral/fifa-data-for-eda-and-stats).

**Data Preprocessing:**

Initially, I examined the dataset and identified that it comprised more than 18,000 data points with approximately 57 columns. Within this dataset, I noticed that 49 columns contained a few missing values.

Subsequently, as I delved deeper into the data, I made an intriguing discovery. There were numerous rows that exhibited missing values, and notably, 42 specific columns had missing values in the same rows approximately 48 times. To address this issue effectively, I opted to remove these 48 rows from the dataset.

Furthermore, I encountered 7 additional columns with missing data. In an effort to handle this missing information sensibly, I proceeded to impute these missing values with appropriate data points that were most fitting in the context of the dataset.

For instance, one of the columns in the dataset was the "Release Clause" column, which contained null values. To address this, I decided to fill these missing values with the value '0'. This interpretation signifies that the players in question had no release clause specified for them.

Another column with missing data was the "Club" column. To handle these missing values, I chose to replace them with the string "Club not Mentioned." This approach helps to indicate that the club information for these players was not provided in the dataset.

**Data Understanding and Visualization:**

Following the data preprocessing steps, I employed various techniques to gain a deeper understanding of the dataset. One of the primary methods I used was the "describe" function, which provided statistical insights into each individual column.

Using the "describe" method, I was able to obtain statistical summaries for the dataset's columns. These summaries included key statistics such as the mean, standard deviation, minimum, maximum, and quartile values. This allowed me to get a sense of the central tendencies, variability, and distribution of the numerical attributes within the dataset.

Additionally, I likely examined data distributions through data visualization techniques, such as histograms, box plots, and scatter plots. These visualizations could reveal patterns, outliers, and relationships within the data that might not be immediately apparent from the summary statistics alone.

Furthermore, I may have conducted exploratory data analysis (EDA) to identify trends, correlations, or interesting patterns between different columns or groups of data points. EDA often involves creating plots, charts, and graphs to visually represent the data and uncover insights.

Overall, these techniques provided a comprehensive understanding of the dataset, enabling me to make informed decisions for subsequent analysis or machine learning tasks.

**Statistical Techniques:**

To gain further insights into the dataset, I applied statistical techniques like correlation analysis. This helped me assess the relationships between different columns. To visualize these correlations more effectively, I created a heatmap, which highlighted the strength and direction of correlations.

Through this analysis, I discovered that numerous columns exhibited high levels of correlation with each other. This information provided valuable insights into potential multicollinearity issues or dependencies among various attributes in the dataset, which can be crucial when building predictive models or conducting more in-depth analyses.