Hyperiondev

TASK

# Exploratory Data Analysis on the ************* Dataset

Visit our website

# Introduction

Summary of the data set

## DATA CLEANING

# SUMMARY OF THE METHODS AND VISUALISATIONS DONE DURING DATA CLEANING

In this section, I have prepared the dataset for analysis by standardising column formats, handling missing values, and converting data types. These steps were essential to ensure consistency and to make the data suitable for visualisation and statistical analysis

Summary of the cleaning:

- Renamed the folder Forbes Top Athletes
- Renamed columns to lowercase and replaced spaces with underscores for easier access in code.
- Converted 'Previous Year Rank' from string to numeric type, 'coericing' invalid values (e.g: 'NA') to NaN
- Converted 'Year" to integer for proper chronological analysis
- Handled missing values in 'Previous Year Rank' by filling with 0, assuming the athlete was new to the list that year.

Visualisation & Checks used during the cleaning:

- Df.info () - to inspect data types and missing values
- Df.isnull () - to count nulls per column
- Df.describe () - to view the summary statistics and check for anomalies

## MISSING DATA

# ANY MISSING DATA? HOW DID YOU HANDLE IT

A careful review of the dataset revealed that only one column contained missing values: 'Previous Year Rank'. This is expected since athletes who were new to the top 100 list in a given year would not have a rank from the previous year.

How was it handled?

- All missing values in 'Previous Year Rank' were replaced with 0.
- This approach treats 0 as an indicator that the athlete was newly ranked that year.
- No other columns contained missing values, so no rows were dropped during this process.

## DATA STORIES AND VISUALISATIONS

# THIS IS THE BULK OF THIS PROJECT. EXTRACT STORIES AND ASSUMPTIONS BASED ON VISUALISATIONS OF THE DATA

1. Most Frequently Featured Athletes
   - A bar chart highlighted athletes who appeared most often on the list.
   - Tiger Woods and Floyd Mayweather dominated in terms of consistent presence.
   - These Athletes not only performed at elite levels but also hand strong brand value and long-term sponsorship deals.

2. Total Athlete Earnings Over Time
   - A line plot of total earnings by year showed a general upward trend in pay, particularly from the early 2000s onward.
   - Peaks often aligned with global sports events (e.g: Olympics, World Cups, major Boxing matches).

3. Dominant Sports
   - Bar plots revealed the most represented sports over the years:
   - Boxing, Basketball, and Golf were among the most common.
   - This shows how certain sports consistently generate massive financial rewards for top athletes.

4. Average Earnings by Sport
   - A comparison of mean earnings showed that individual sports like Boxing and Golf often had higher average payouts, largely due to fewer competitors sharing revenue and larger personal endorsements.

5. Correlation Between Rank and Earnings
   - A heatmap indicated that moderate inverse relationships between ranks and earnings – lower numerical rank (closer to #1) generally aligned with higher income.
   - However, outliers existed: some athletes earned less but maintained high rankings, possibly due to non-financial criteria like popularity or media impact.

6. Key Takeaways:
   - Endorsements play a massive role in total earnings, especially in sports like Tennis and Golf.
   - Nationality clusters could be explored further – the USA dominates the dataset.
   - Gender data was not included, but it’s notable that female representation appears extremely limited or absent in top earnings

# ENSURE THIS DOCUMENT IS NEAT AND ADD IT TO YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY : Aashiq Ebrahim**