Exploratory Data Analysis on the Forbes Richest Athletes (1990–2020) Dataset

Introduction

Forbes Richest Athletes 1990–2020: Key Insights The dataset captures the earnings of the world's highest-paid athletes over three decades. Through exploratory data analysis, several trends and patterns emerge:

Data cleaning

1. **Loaded the data safely**
    - Detected file encoding using chardet to ensure proper reading of CSV.
    - Handled any errors during file load to avoid crashes.
2. **Standardized Sport column**
    - Converted all entries to lowercase and removed extra spaces or non-alphabetic characters.
    - Consolidated similar sport names (e.g., "NBA" → "basketball", "Nascar" → "motorsports") for consistency.
    - This ensures all analyses group sports correctly.
3. **Standardized Nationality column**
    - Trimmed whitespace and unified different representations of the same nationality (e.g., USA → American, UK → British).
    - This avoids miscounts in nationality-based aggregations.
4. **Cleaned Previous Year Rank**
    - Replaced invalid entries like "nan", "not ranked", "??" with proper NaNs.
    - Flagged values with ">" and cleaned numeric values for analysis.
    - Created a column to flag ">" if it existed in the Previous Year Rank.
5. **Pre-processed earnings ($ million)**
    - Converted earnings to numeric, coercing invalid values to NaN.
    - This allowed proper aggregation, median, and sum calculations without errors.
6. **Handled missing data**
    - Visualized missing data to identify gaps in columns.
    - Ensured analyses (e.g., median earnings per year) account for missing values.

Missing data

The dataset contains some missing values, particularly in columns like Previous Year Rank and earnings ($ million). Missing data were handled as follows:

- For Previous Year Rank, invalid entries such as “nan”, “not ranked”, “none”, “?”, and “??” were replaced with NaN.
- For earnings ($ million), non-numeric or malformed values were coerced to NaN.
- Missing values were visualized using missingno bar plots to understand their distribution across the dataset.
- Analyses (e.g., median earnings by year, total earnings by sport) inherently ignore NaN values in aggregation, ensuring that summaries are based on valid data only.

Data stories and visualizations

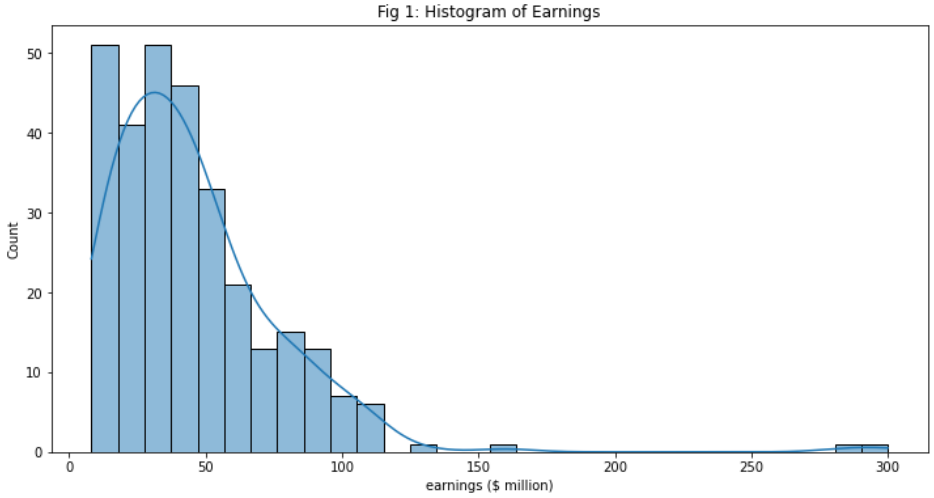
1. Earnings Distribution and Outliers

Most athletes earn below $100 million annually, but a small number of superstars, such as Floyd Mayweather, skew the distribution with extreme earnings in peak years like 2015 and 2018. These outliers highlight how a single athlete can dominate earnings each year.

Histogram of Earnings, Figure 3, displays the distribution of earnings among a group of athletes. The x-axis represents earnings in millions of dollars, and the y-axis shows the count of athletes within each earnings bracket.

Right-Skewed Distribution: The most significant feature of this histogram is its right-skewed distribution. The majority of athletes have earnings concentrated in the lower income brackets (under $100 million), forming a tall peak on the left side of the chart. The distribution then tails off to the right, with fewer athletes in the higher earnings brackets.

Fig 1: Histogram of Earnings

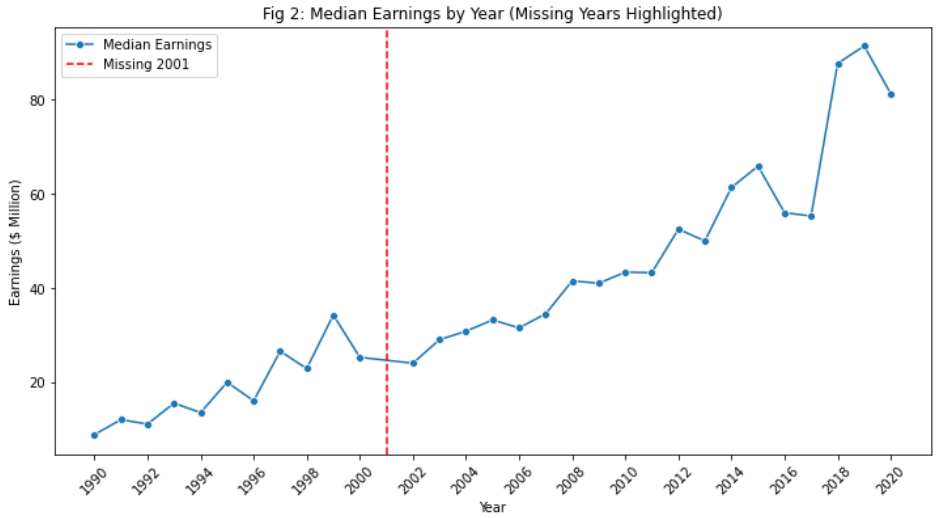


2. Trends Over Time

The median earnings have generally increased from 1990 to 2020, reflecting the growing commercial and sponsorship opportunities in sports. Spikes in specific years indicate extraordinary earning events, often tied to superstar performances or highly lucrative contracts.

- The spike in 2015 and 2018 were driven by boxing as can be seen in **figure 2** below, but this was a result of Floyd Mayweather who earned significantly more than all other sports people.
- There is data missing for 2001 and I believe this is MNAR (missing not at random), I believe this had something to do with September 11th when most major sport events were cancelled.

Fig 2: Median Earnings by Year (Missing Years Highlighted)



3. Earnings by Sport

Basketball, soccer, golf and boxing consistently appear among the highest-earning sports. Individual peak years in certain sports show how a single athlete or a group of athletes can dramatically influence yearly averages.

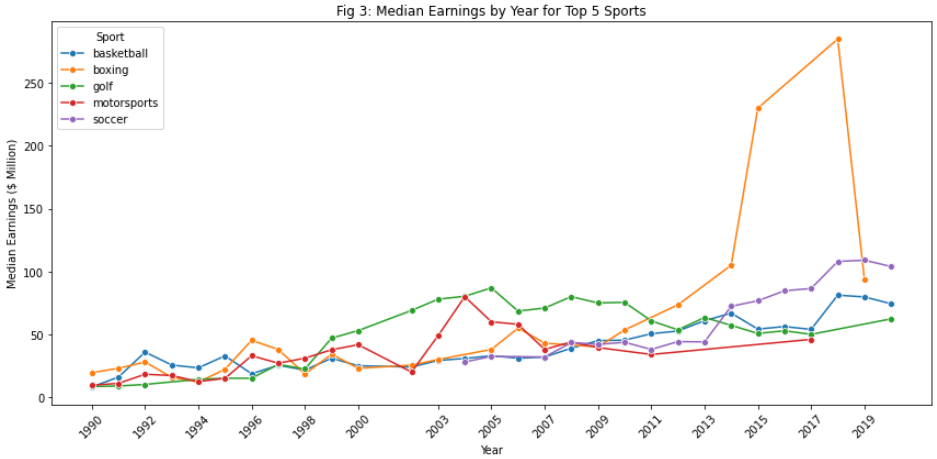
Figure 3 shows the median earnings for different sports tells a clear story: not all sports make money the same way.

Boxing: Earnings are usually low, but occasionally, a single big fight (like the one in 2015 and 2018) makes so much money that it sends the average earnings skyrocketing. This shows that in boxing, a few top athletes win almost everything, while most others earn very little.

Soccer and Golf: These sports show a steady, solid increase in earnings over the years. Tiger Woods in the mid-2000s shows how even in a consistently growing sport, one superstar can still dramatically boost the numbers.

Basketball: Earnings for basketball players have consistently gone up because the NBA brand is getting bigger and richer every year, both in the US and internationally.

Motorsports: Earnings in this sport have been much more stable and don't change much from year to year.



In short, team sports like soccer and basketball show a more reliable increase in earnings, while individual sports like boxing can have huge, but short-lived, spikes in income due to a single major event.

4. A Breakdown of Global Sports Earnings by Nationality and Sport

The bar chart provides a clear visual representation of the earnings distribution for top athletes across different nationalities, categorized by sport. The most striking insight is the distinct difference between the earning patterns of American athletes and those from the other nationalities shown (Argentine, German, Portuguese, and Swiss).

- **American Athletes:** Earnings are distributed across multiple major sports. Largest percentages: Basketball (39.1%), Boxing (21.5%), Golf (23.7%) and American Football (10.9%). Smaller contributions from Soccer, Cycling, Motorsports, and Tennis.
- **Top Athletes from other countries:** Earnings concentrated in a single sport (Soccer, Tennis or Motorsport). High degree of specialization.

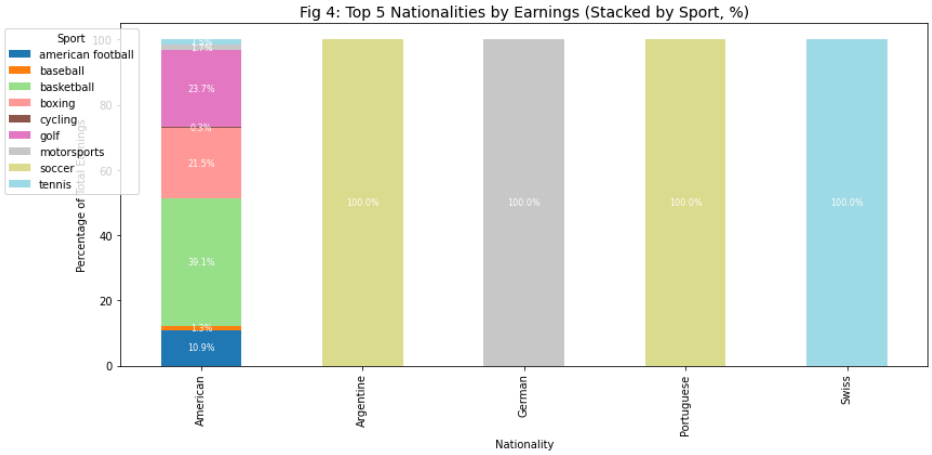


Figure 4 shows how a country's top athletes make their money. In the United States, top athletes earn money from many different sports, like basketball, boxing, and American football. In other countries like Argentina, Germany, Portugal, and Switzerland, the top athletes make almost all their money from just one sport.

5. Number 1 Ranked Athletes by Year

Tracking the highest-earning athlete each year reveals historical shifts in dominance across sports.

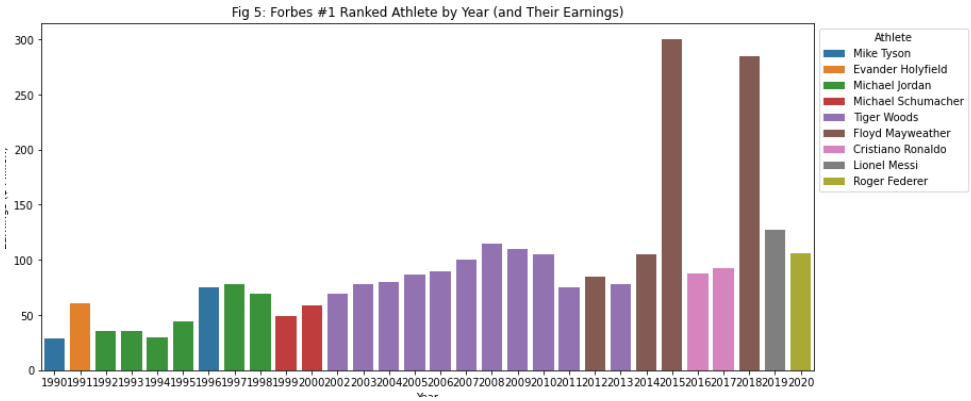
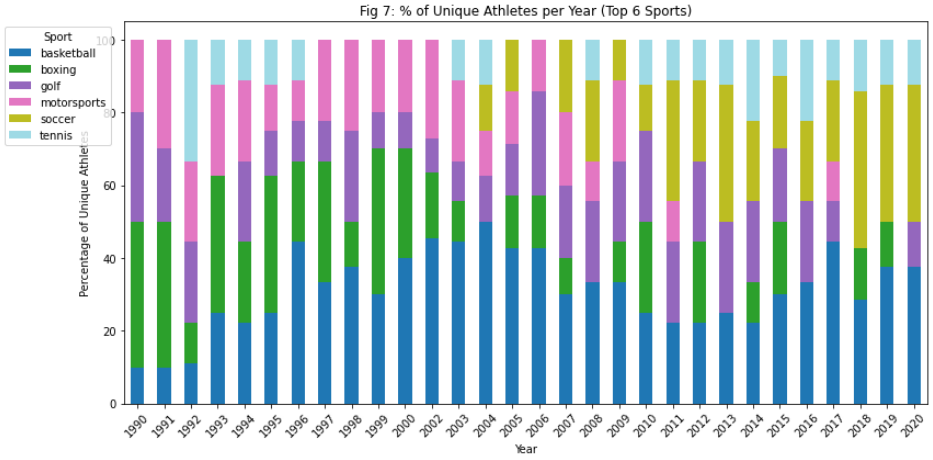


Figure 5 puts names to the outliers. It shows a dramatic increase in the earnings of the top-ranked athletes over time. Floyd Mayweather's earnings in 2015 and 2018 stand out, representing the "long tail" of the histogram. It also shows how a few individuals like Tiger Woods and Michael Jordan, consistently dominated the top spot, reinforcing the idea of a handful of superstars.

6. Globalization of Sports Earnings

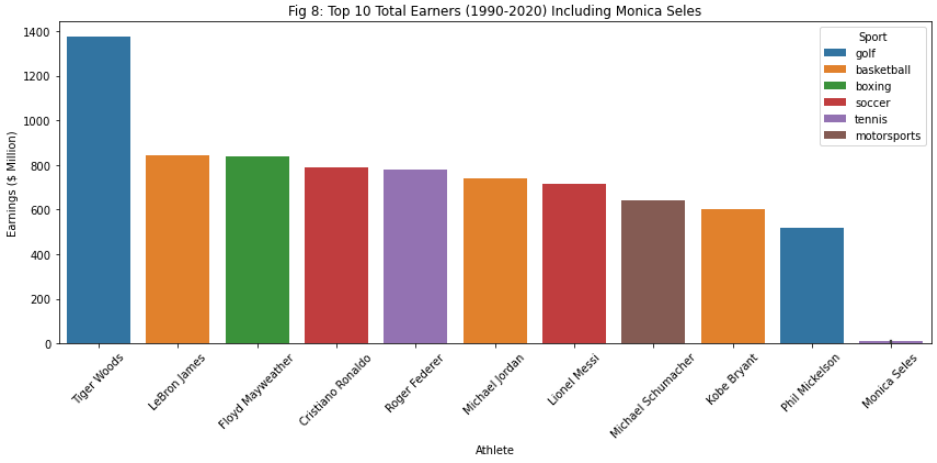
Figure 7 shows how the representation of athletes across different sports in Forbes' top rankings has shifted over time. In the 1990s, basketball and boxing dominated, largely driven by icons like Michael Jordan and Mike Tyson. Soccer, by contrast, was almost absent until the early 2000s, when the emergence of global superstars such as David Beckham, Cristiano Ronaldo, and Lionel Messi coincided with the sport's commercial globalization through television rights and sponsorship deals. This shift highlights how changing market dynamics and global fan bases have reshaped the composition of top-earning athletes.

In the most recent three years, basketball and soccer have consistently dominated the top rankings, while some sports such as motor racing have dropped out entirely for several years at a time, reflecting the declining share of athletes from those sports among the world's top earners.



7. Gender Pay Gap

Figure 8 shows a significant gender disparity in professional sports. No female athletes appear in the **number one ranking** across the 30-year period. Monica Seles is added for comparison (1992), her earnings were below the median, see figure 2 above. This highlights the persistent gender pay gap.



Conclusion

The analysis highlights the disproportionate impact of superstar athletes on earnings trends, the steady growth of athlete income over time, and the dominance of certain sports and nationalities in shaping the global earnings landscape.

One striking observation is the absence of female athletes from the number 1 ranking in the past 30 years. Despite increasing visibility and sponsorship opportunities for women's sports, a significant disparity remains at the highest levels of athlete earnings.

This report was written by: Juleiga Regal