**Forbes Highest Paying Athletes Analysis**

**Introduction**

The sports industry is valued in US$ 500BN worldwide and is grows at a 5.9% annual rate. Just it sports sponsorships, it is estimated that brands will spend US$ 50BN this year and the single most valuable sporting event (the NFL Superbowl) is valued at $ 379 M. With all of these facts, we are curious to understand what the top figures in sport make, if their earnings depend on what sport they play and how their earnings have evolved over the years.

Research Questions:

1. Are American athletes paid more than their foreign counterparts? Does this differ by sport?
2. Have top athlete earnings increased in real value over the past 30 years?
3. What sport pays its top athletes the most?

To analyze the data for these research question, we used a different Kaggle database with monthly CPI indices to calculate inflation with a 1990 as the base year and joined it to our main database. We used the inflation to adjust nominal earnings to 1990 US$ and performed the analysis using the adjusted earnings to eliminate inflation bias from the analysis (for example, in research question 1, more recent earnings would have skewed the data vs past earnings).

**Datasets Used**

For our project we utilized two datasets, the first “Forbes Highest Paid Athletes 1990-2020” was a dataset that included the ranking, pay, nationality, and sport of the top 10 athletes for each year from 1990 to 2020. It can be found at: <https://www.kaggle.com/datasets/parulpandey/forbes-highest-paid-athletes-19902019>. The compiled list pulled from reported earnings for each athlete for each of the years included which may have been a result of massive events, signings, and more. The dataset was helpful but it could have included more information as it was very limited at only 301 rows overall, the 301 results from a tie in pay one year between two individuals on the list. The second dataset used was, “U.S. Inflation Data” which included U.S. inflation indexes and cpi for over a hundred years starting in 1913 and going all the way to 2021. It can be found at: <https://www.kaggle.com/datasets/varpit94/us-inflation-data-updated-till-may-2021/data>. The data itself is sourced from the U.S. Bureau of Labor Statistics. We utilized this dataset by comparing the reported earnings in the form dataset and using 1990 as a base year then calculating the inflation of each year there after in order to represent athlete earnings based on what a dollar was worth in 1990.

**Data Cleaning**

At the beginning phases of our project upon setting up our initial Jupyter Notebook and loading in the first csv containing the dataset, “Forbes Highest Paid Athletes 1990-2020” we noticed that there was an issue as far as the veracity of our dataset goes. There were several different versions of the same sport that were represented in different ways. For example, NFL and American football were represented as two different variations of the same sport. NHL and Ice Hockey were another example. Nascar and F1 Motorsports were also not distinguished. Our first step in data cleaning was to identify these issues and group them appropriately so that they could be properly represented in our further analysis. A second issue that we noticed is that there were a few outliers that performed multiple sports, such as Deion Sanders and Conor McGregor. We had to make a decision on how we would assign them to a particular sport. As a result, we decided to assign them to their most publicly identified sport that they were involved with, Deion Sanders to American Football and Conor McGregor to MMA, as a bulk of their careers were in the two sports respectively. The changes that we made here were able to be evident in the form of a better representation in our visualizations.

**Are American athletes paid more than their foreign counterparts? Does this differ by sport?**

Our hypothesis for this research question was that American athletes would earn more than their international counterparts in all sports based on the size of the sporting industry in the USA vs the rest of the world. However, we found that this is not necessarily true.

A graph of multiple colored bars

Description automatically generated

As the graph above shows, there is a clear difference in top earnings between sports. Sports like basketball, boxing, Golf and American football have earnings going to predominantly American athletes while sports like Soccer, F1 Motorsports and Tennis have top earners being predominantly international. This matches the sport’s geographical popularity as F1, Soccer and Tennis are sports that have historically had an international audience, while the NBA, NFL, Boxing and PGA have a very large American audience and most of the tournaments/matches happen in the USA.

The American sporting sport industry is the largest in the world. This appears to correlate with athlete salaries as they are the stars and the events, TV rights, merchandise, etc. depends on them performing well, however, we need more information to carry out a deeper analysis into salary drivers in the industry. It would be interesting to understand what drives salaries in the USA vs the rest of the world and how this varies by sport. It is interesting to note that sports with higher international earners are actively trying to break into the American market with F1 launching two new circuits in the past two years (Miami and Las Vegas) and the Inter Miami buying Messi. Both these events broke the news and elevated ticket process for the events considerably so it would be interesting to understand how this correlates with salaries in the future.

**Have top athlete earnings increased in real value over the past 30 years?**

Because the sporting industry has grown above inflation, we expect athlete earnings to have increase over the past 30 years, even when adjusted for inflation. For this analysis, we removed some anomalies that we found in 2015 and 2018 by using the 99th quantile.

A graph showing the difference between an average and an average

Description automatically generated

As we can see, there is a positive correlation between years and adjusted earnings. R2 for this model however is 0.31, which means that only 31% of the variability in the data is explained by the model. When looking more closely at the data, we found that in the majority of years, the top earner had a considerably higher reported earning than the second ranked athlete. In this sense, we further cleaned the data considering the top earners for every year as anomalies, and re-plotted the linear regression model.

A graph showing the difference between an average and an average

Description automatically generated

As we can see from the second model, the positive correlation remains and it even became slightly more positive (0.88x) and in this case, R2 for the model is 0.52, which means that more than half of the variability in the data is explained by the linear regression model.

In conclusion, top earners are earning more in real terms over the years.

**What sport pays its top athletes the most?**

Based on the first research question and our individual experiences, we believe that NBA and NFL athletes will be the top earners. To analyze this question, we decided to look at per capita earnings instead of aggregated ones to distinguish individual earnings.

A graph of a chart

Description automatically generated with medium confidence

If we look at the scatterplot of adjusted earnings vs time with all of the data (anomalies included), we can see that there are 2-3 points that catch our attention. When deep diving into the data, we found that these points corresponded to boxing legend Floyd Mayweather. Looking into pop culture, 2015’s fight between Floyd Mayweather and Manny Pacquiao was deemed the fight of the century, and Mayweather made a record US$ 275M on a single night according to Forbes.com. Interestingly enough, this fight also made Manny Pacquiao an incredible US$ 100M, making boxing the top sport in earnings that year. A similar situation occurred in 2018 when Mayweather fought Connor McGregor. That year, Mayweather toped the earnings rank with US$ 285M, nearly triple of what the second ranked athlete made that year.

A graph of a bar graph

Description automatically generatedA graph with green bars

Description automatically generated

If we look at the aggregated data, Basketball has the total aggregated earnings, however, they also have a very high number of entries and therefore the per capita earnings are actually much lower and basketball ranks 5th on the per average adjusted earnings list by sport. The top earning sport in this case is MMA but this is because there is a single athlete in the list from that sport and he is Connor McGregor who fought Floyd Mayweather in his sport in 2018.

In general terms, individual sports appear to pay more than group sports. This makes sense as the events have to pay more athletes for a similar event time in a group sport.

**Limitations and Next Steps**

Because our database only included top earners for every year 10, we constantly ran into the issue of athletes falling off the top 10 and therefore their earnings not being recorded. This alters the data slightly, especially when looking at aggregate earnings. This is why we decided to look at yearly earnings instead of lifetime.

It is also important to recognize that analyzing only top earners means that we are by default analyzing skewed data. In order to get a better idea of the reality of the sporting industry, it would be interesting to obtain a larger sample of salaries from athletes in various earning levels.

Furthermore, we lacked industry data (ex. TV rights, number of spectators, team valuations) to carry out a correlation analysis to understand key drivers for earnings. It would be interesting to understand what the key drivers for yearly earnings are and how these vary by sport and nationality.

Finally, we found a limitation when looking at earnings by nationality. When looking for correlations and drivers for earnings, we looked into nationality, however this is not a variable for where the athlete is playing the sport and therefore some data might be altered, especially in the NBA, NFL and MLB, where the league happens entirely in the USA but some athletes are international.

**Summary**

This dataset was very interesting in the sense that it provided a clear indication that athletes are indeed being paid at a higher rate than their 1990 equivalents. It also paints a clear picture that athletes that belong to certain sports such as Boxing and Basketball earn more worldwide compared to individuals of less popular sports. It is also evident that athletes that are from the United States tend to earn more than the other athletes who hail from different places around the world. That being said, we can tell that is the case for the highest earners. There is a clear limitation in our observation of such in that we cannot account for the total earnings of every member of a given sport with such a limited dataset. We also wouldn’t be able to confidently predict what the highest earners would be earning down the road. In the end though we can say that this was an interesting thought experiment and that it provided a nice way to compare the athletes who are on top.

**Team Members involved with this project:**

* Fernanda Valdez
* Jason Cisneros
* Uzor Francis

**Faculty**

* Alexander Booth
* Sherhone Grant