Hyperiondev

TASK

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

Visit our website

Introduction

This dataset contains information about the top 10 highest-earning athletes for each year from 1990 to 2020 (inclusive). Their names, ranks, and earnings are provided in the dataset.

MISSING DATA

The only column containing missing data was the “Previous Rank” column. This is mainly because athletes appearing in the top 10 in one year may not have been within the Forbes ranking system at all in previous years. The data was quite “chaotic”, with missing values entered as any of the following: “nan”, “not ranked”, “none”, “?”, and “??”. Moreover, much of the data that was actually present was not an integer type, but rather included other symbols such as “>” in “>100”, which makes further numerical analysis difficult. In light of these reasons, this column was discarded.

DATA CLEANING

Aside from the “Previous Rank” column, the “S. NO” and “Name” columns were dropped as they were identified as not useful in this analysis.

The “Sport” column required considerable cleaning; mainly, because some entries pointing to the same sport were being recognised as different categories. Firstly, obvious duplicates were removed by making all entries lowercase in this column, and then those which were not rectified in this way were dealt with separately by calling the *.replace()* function as necessary. Finally, the entries were formatted appropriately using *.title()* and capitalising acronyms where needed.

By and large, the data in the “Nationality” column was input as a country name. However, there were two instances where the adjective was used in place of the country name (Filipino rather than Philippines, and Dominican rather than Dominica). These were corrected to the relevant country names.

Finally, the “earnings ($ million)” column was also reformatted (capitalised) in order to ensure consistency with the other column names.

DATA STORIES AND VISUALISATIONS

Nationalities & Sports

Athletes from 21 different countries were represented in the dataset (Figure 1). By far, most of the athletes in the dataset were from the USA, with 206 entries. This had a large lead from the next-highest countries, Germany, with 13 athletes, and Switzerland, with 12 athletes. It is unlikely that there is a data collection bias where the data collector favours the inclusion of American athletes in the dataset (i.e. given that Forbes is an American business magazine), but rather that the US has a greater provision for athletes' salaries.

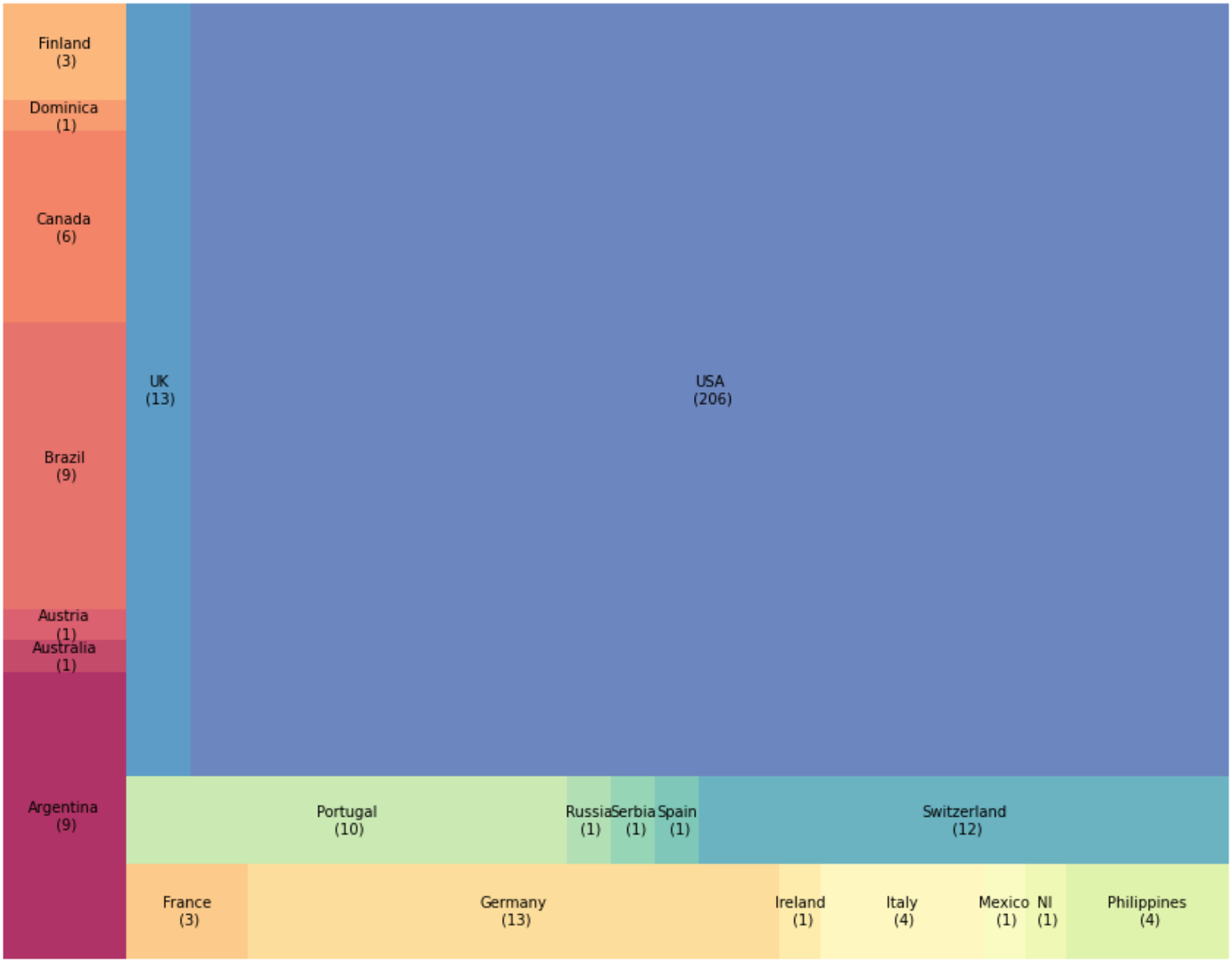

Figure 1: Treemap of athletes' nationalities in the dataset.

It is also worth examining the particular sports appearing in the dataset. By looking at Figure 2, we observe that not only is basketball the most common sport to appear in the dataset, but also that all basketball players are from the USA. From this we can conclude that basketball players are generally well-paid (enough to consistently make it into the top 10 of richest athletes). Since all basketball players in the data set are US citizens, it also follows that a large portion of the athletes would be from the USA also. It is also worth noting that the 2nd and 3rd most common sports (boxing and golf respectively) also comprise mainly US athletes, further bolstering the number of US athletes in the dataset. This supports the conclusion that athletes in the USA are generally well-paid.

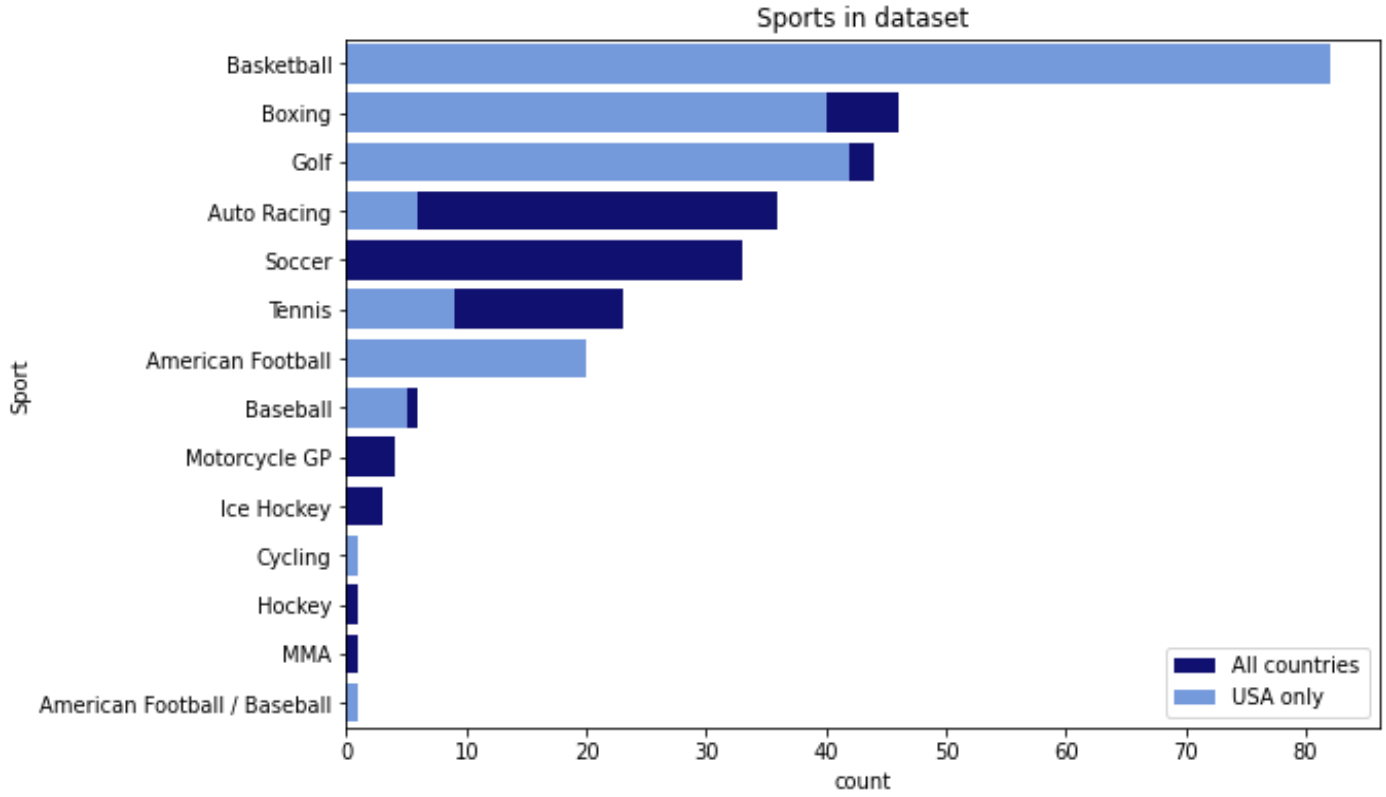

Figure 2: Incidence of athletes, grouped by sports, divided into USA (light blue) and non-USA athletes (navy).

It might be tempting to conclude that basketball, which appears frequently in the top 10, is perhaps the highest-paid sport. This is actually not the case; the best-paid sport on average is MMA (Figure 3). However, there is only one data point for this, so this result should be taken with a pinch of salt; MMA appears only once in the entire 20-year dataset, so this is likely to be an outlier in terms of general MMA earnings.

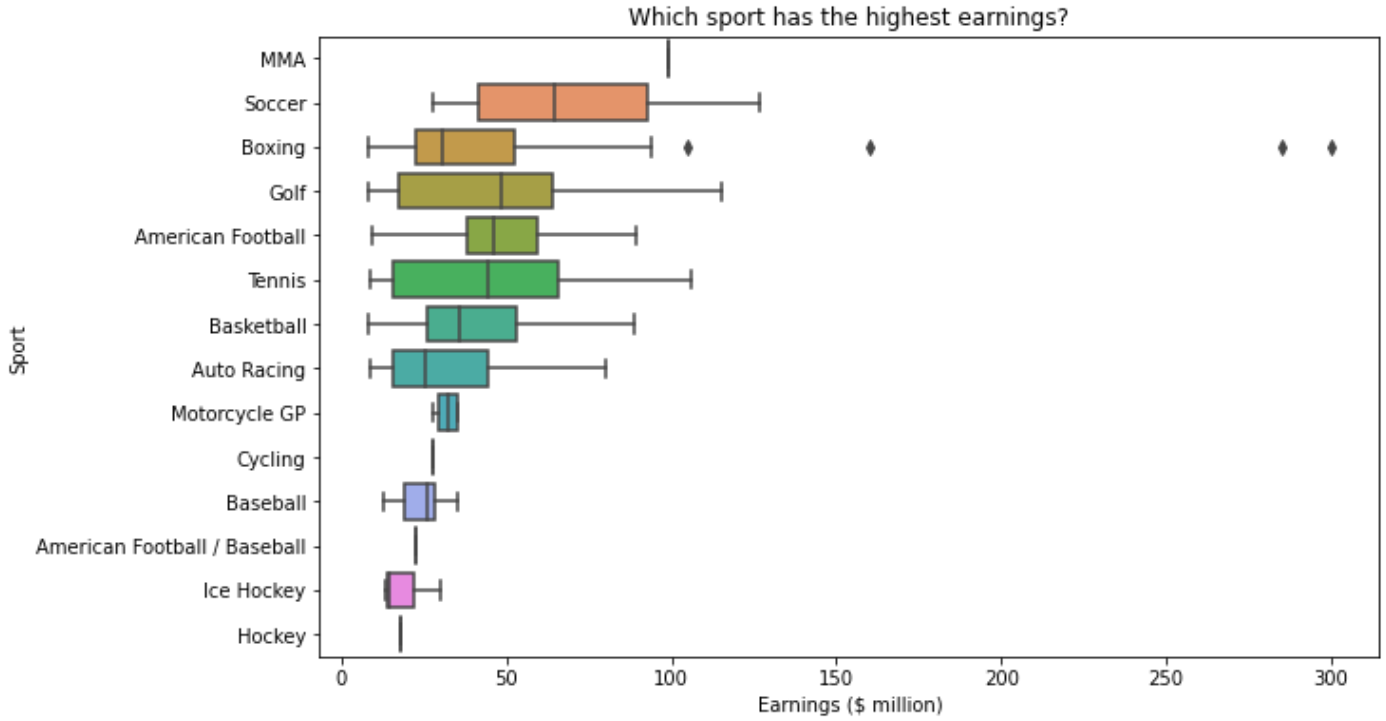

Figure 3: Boxplot of athletes' pay, grouped by sport type, and ordered by highest to lowest mean earnings.

It is actually boxing which has the top 3 highest-paid athletes in the entire dataset, much higher than the MMA mean. It also has the third-highest earnings by mean, after MMA and soccer, and precedes golf. We note from Figure 2 that the majority of boxers, golfers, and basketball players are US athletes, and these sports overall score quite highly in terms of mean earnings. Does this mean that US athletes are the best-paid athletes in the world?

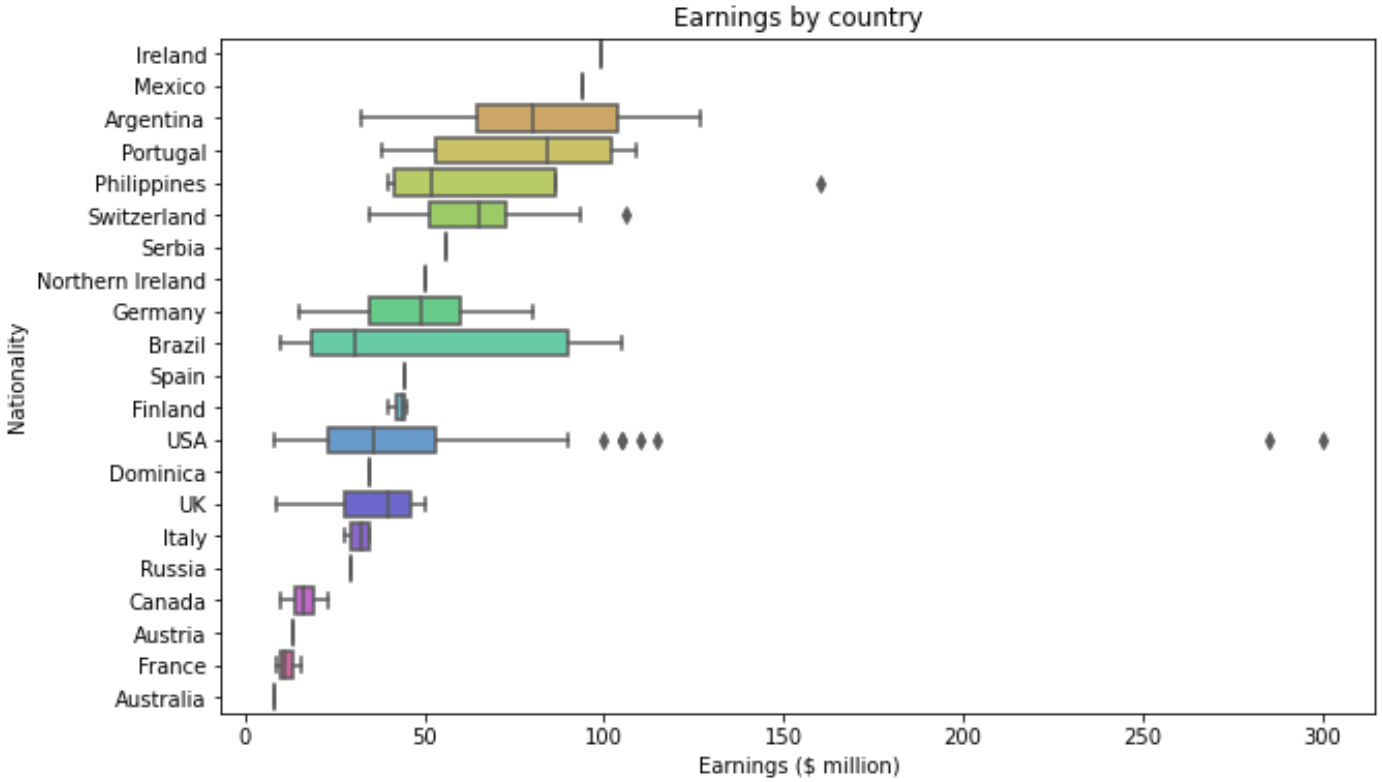

Figure 4: Boxplot of earnings, grouped by nationality, and ordered by highest to lowest mean earnings.

Figure 4 suggests otherwise. The mean earnings for the US is far from the top; however, the top two athletes in the dataset are indeed US citizens (we can infer from Figure 3 that they're both boxers). Ireland has the highest mean, but we can see that there is just one data point for Irish athletes. Fetching the relevant data from the dataset, we see that this is the MMA athlete with the exceptionally high earnings discussed in Figure 3.

We see that many countries only contribute one athlete each to the dataset (Ireland, Mexico, Serbia, Northern Ireland, Spain, Dominica, Russia, Austria, and Australia). Thus, perhaps these data should be analysed in a different way, given that they are not consistent contributors to the top 10 lists. A better comparison of earnings could be made using an e.g. Top 100 list, i.e. if these countries appeared more frequently.

Years & Ranks

It might be interesting to note how the earnings of the top 10 athletes fluctuate over time. There is a general increase in mean and total earnings with time, and the two lines follow the shape of one another quite closely (Figure 5). This is to expected with inflation and as the sports industries grow.

There are two years (2018 and 2015) in which both the mean and total earnings peak. Examination of these two years shows that in both years there was an athlete with substantially higher earnings than average (2018: Floyd Mayweather, boxing, $285 million; 2015: Floyd Mayweather, boxing, $300 million), which brought both the mean and total earnings up considerably. We further note that the top two earners in the dataset were not two separate athletes – but actually the *same* athlete, Floyd Mayweather.

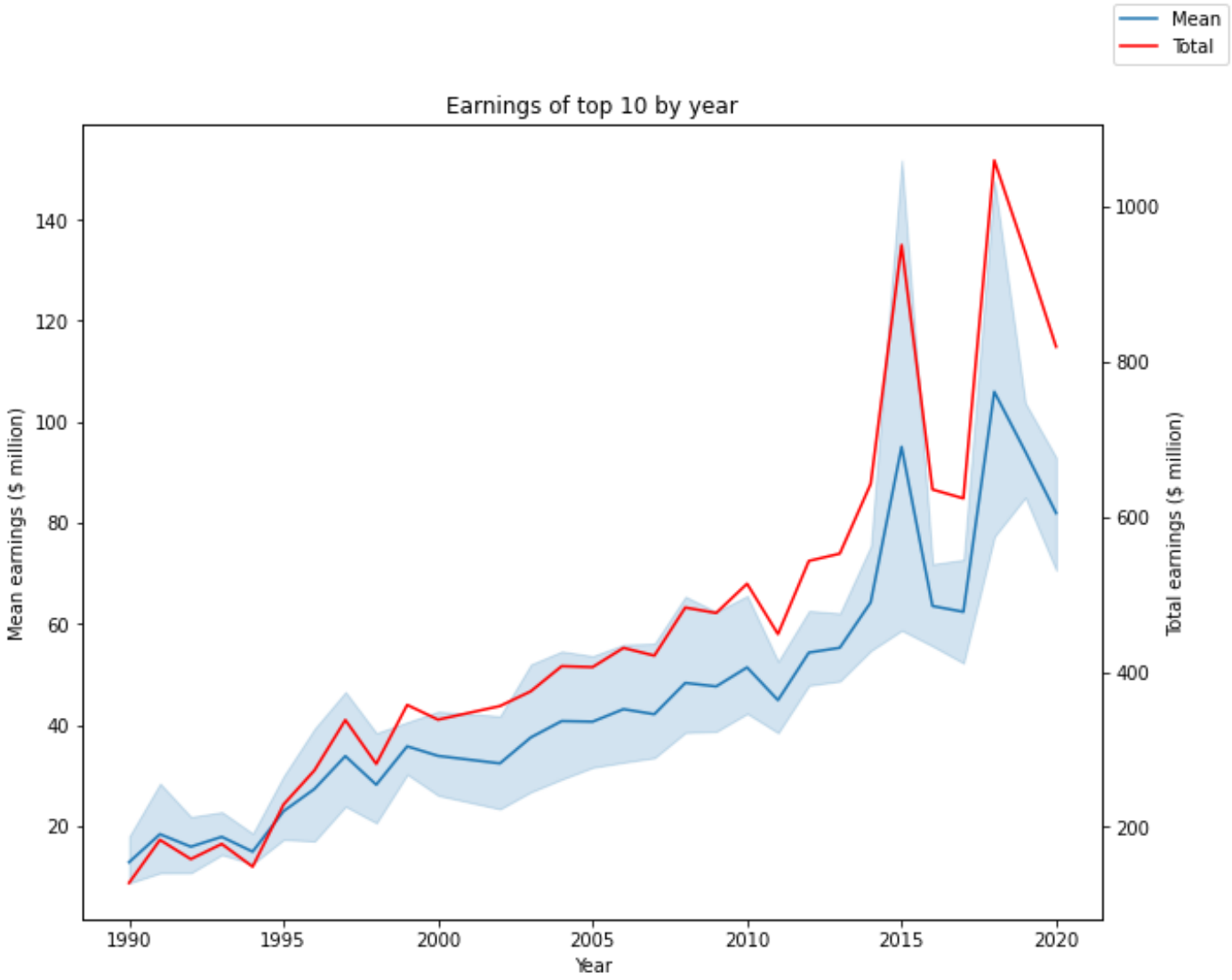

Figure 5: Mean and total earnings of top 10 athletes between 1990 and 2020 (inclusive).

Floyd Mayweather was the rank 1 athlete in both 2015 and 2018, outstripping the remaining athletes in his year substantially. Usually, however, there is not such a large difference in earnings between ranks (Figure 6).

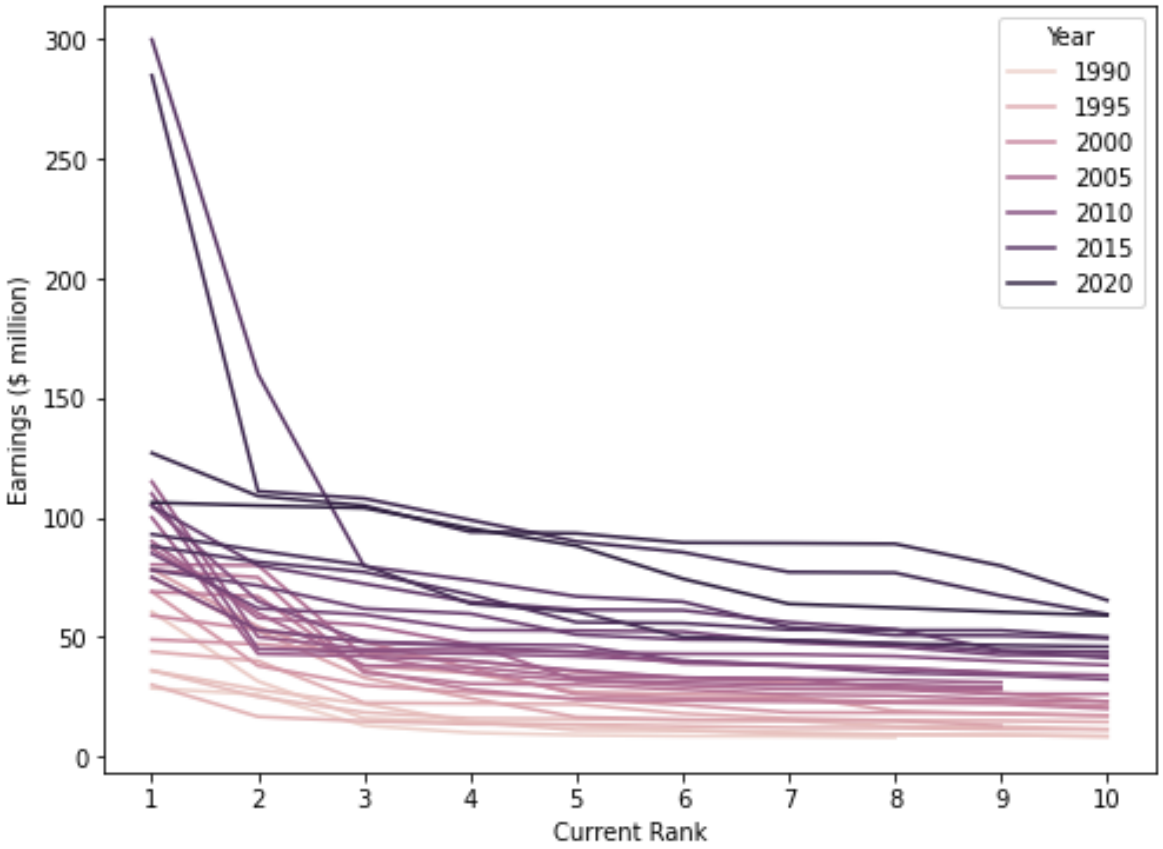

Figure 6: Earnings by rank for each year between 1990 and 2020 (inclusive). The darker lines indicate the more recent years, and vice versa for the light lines.

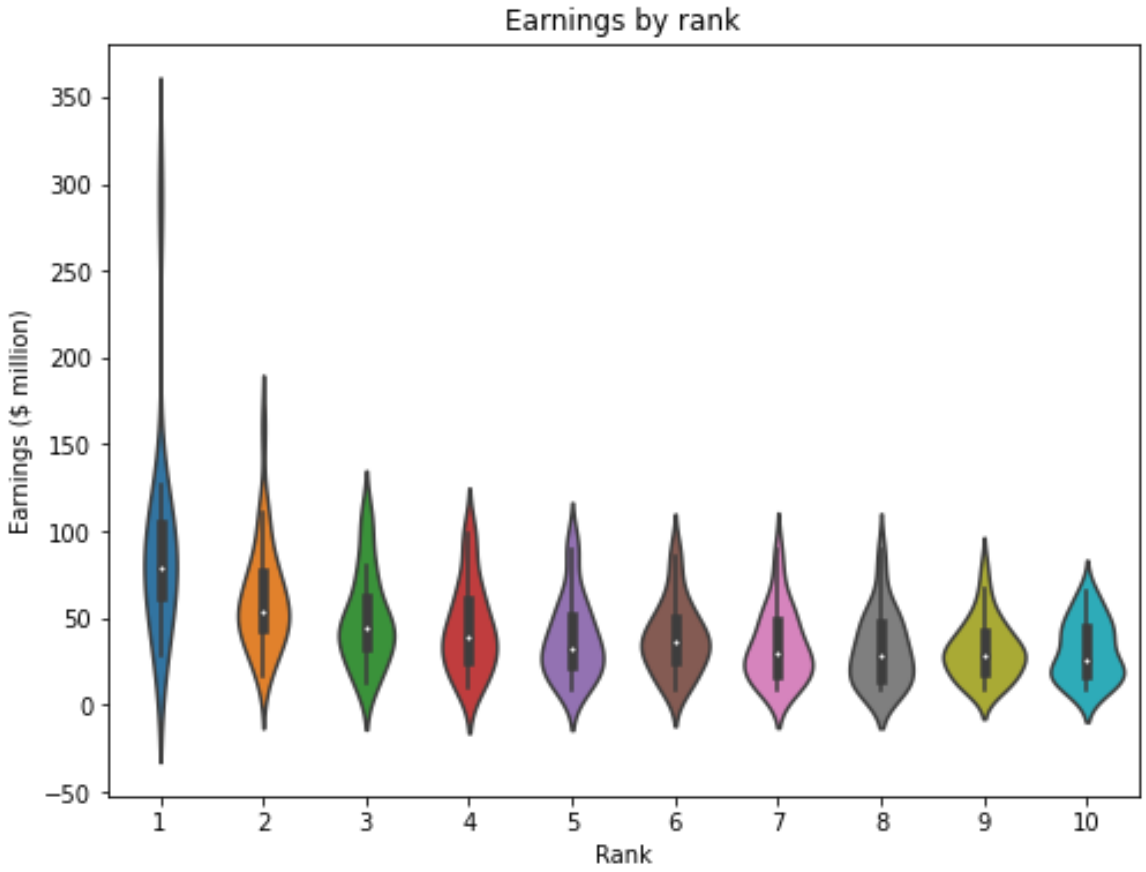

Figure 7: Violin plot of earnings according to rank.

The outliers in the data belong to the years 2015 and 2018 as discussed previously. However, it is true for many years that there is quite a bit of difference between rank 1 and rank 2 athletes (just not as pronounced as for the years 2015 and 2018). Aside from these differences, however, the earnings do not vary significantly between ranks in the top 10. This is further exemplified in Figure 7.

THIS REPORT WAS WRITTEN BY: VICTORIA DARAMY-WILLIAMS