

**TASK**

**Exploratory Data Analysis on the Forbes’ Richest Athletes Dataset**

**Introduction**

This dataset consists of 301 entries and includes the top ten highest earning athletes from each year between 1990-2020. Each entry contains information regarding an athlete’s earnings for one particular year, as such, many celebrity athletes who had high value sponsorships or stayed at a high level for many years appear on the list many times, once for each year they appear in the yearly top 10 list from Forbes.

This list consists of 82 unique athletes. The dataset describes the Name, Nationality, Sport, Rank, Previous year’s rank and yearly earnings for each athlete on the dataset.

## Data Ranges:

**Name:** Athlete’s Given and Family name, in the case of athletes who go by two names (Mike Tyson also goes by the name Malik Abdul Aziz), the more familiar name is used.

**Sport:** Auto Racing, Boxing, Basketball, Golf, Tennis, American Football, Baseball, Ice Hockey, Formula One (referred to as F1 in this document), Nascar, Soccer, Cycling, Motorcycle GP, Mixed Martial Arts (referred to as MMA).

**Rank:** Continuous 1 to 10.

**Previous Year’s Rank:** Continuous 1 to 10.

**Yearly Earnings:** Continuous from 8.5 to 300.0 (in millions of USD).

**Year:** Continuous 1990 to 2020.

**DATA CLEANING**

There were some incidents of inconsistent data in the Sport field where some athletes would have the name of the sport they played, while others had the league in which they played. (e.g. basketball and NBA) there were also several different ways that f1 and Nascar were entered into the dataset. I decided to ensure that these were all set to consistent labels that clearly explain the sport the athlete participates in, as such all references to NBA in the Sport column were changed to basketball, NFL to American football, auto racing (Nascar) to “nascar”, f1 motorsports, f1, and f1 racing were changed to simply be “f1”. I have left “auto racing” without mention of f1 or Nascar to remain as other motorsports exist such as rally or touring car racing, and this entry can represent this. Another issue were case inconsistencies such as “Boxing” and “boxing.” I changed all values to be in lower case in the dataset to avoid python reading these as two separate sports.

In the column “Previous Year Rank” there were some athletes who were placed outside of the top ten who then rose into the top ten for that year. Some of these records were formatted as such “>40.”

This caused the column to have an object datatype and not be numeric which is a problem. To fix this, I approximated values that could be approximated. E.g., we only have data for ranks 1-10 of any year, so any ranks outside of the top ten are already vague. When listed as “rank: >40” this was changed to “40” so that we can keep the rank to the best of our knowledge but still keep it usable when we change this column into integer datatype.

**MISSING DATA**

There were 24 missing values recorded in this dataset. However, this initial count did not include mislabelled data. An investigation into the “Previous Year Rank” column found that as well as NAN values there were also values such as “None, No record, ??” that were in effect missing but in a format that python could not recognise as missing. To fix this, any string that was not a numeric rank was replaced with an empty entry and then all empty entries were replaced with Not A Number values so that the column could be changed into the integer datatype. After that, there was a consideration on how to address these null values. We had around 11% of “previous year’s rank” values missing which is more than can be dropped while preserving the dataset. Considering the context of this column, the important values are 1-10 anything else we do not have any information on anyway. While knowing the specific rank outside of the top 10 is still valuable, there is only so much value in an arbitrary value outside of the top ten.

The options were then, leaving the values as null, this would then cause the dataset to see these as essentially “no rank” which is what many of them were originally. Or to find athletes that have appeared before and after and use a regression line to predict this missing value. This seemed precarious as we are only keeping data for the years they have appeared in the top ten, and so if an athlete was, for example, injured and missed a year of competition, only to recover and rejoin the top ten list, our regression method would put them in and another the same rank when actually this would be very misleading.

In the end, using these null values as “not ranked” was the most reasonable action to take, doing this preserved all of the entries, while doing so in a way that did not decrease the overall amount of data in the set. Our knowledge remained the same by addressing the null values in this way.

**DATA STORIES AND VISUALISATIONS**

**Univariate Analysis:**

Firstly, it was valuable to show a distribution of sports among the highest earning athletes, this chart shows the total among all years covered in the dataset so consistently high earners will be contributing more to this pie chart than those who only briefly appear.

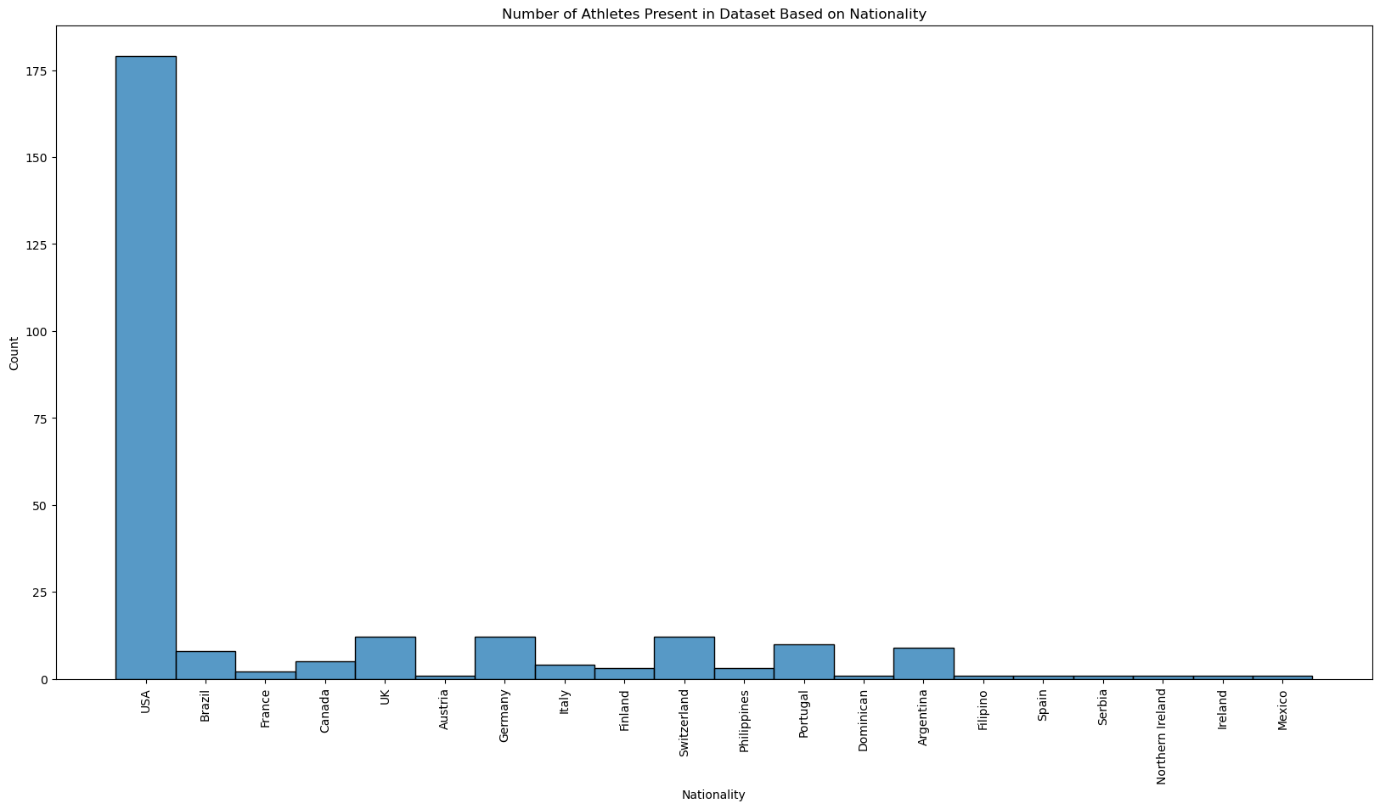
**图表, 饼图

描述已自动生成**

Figure

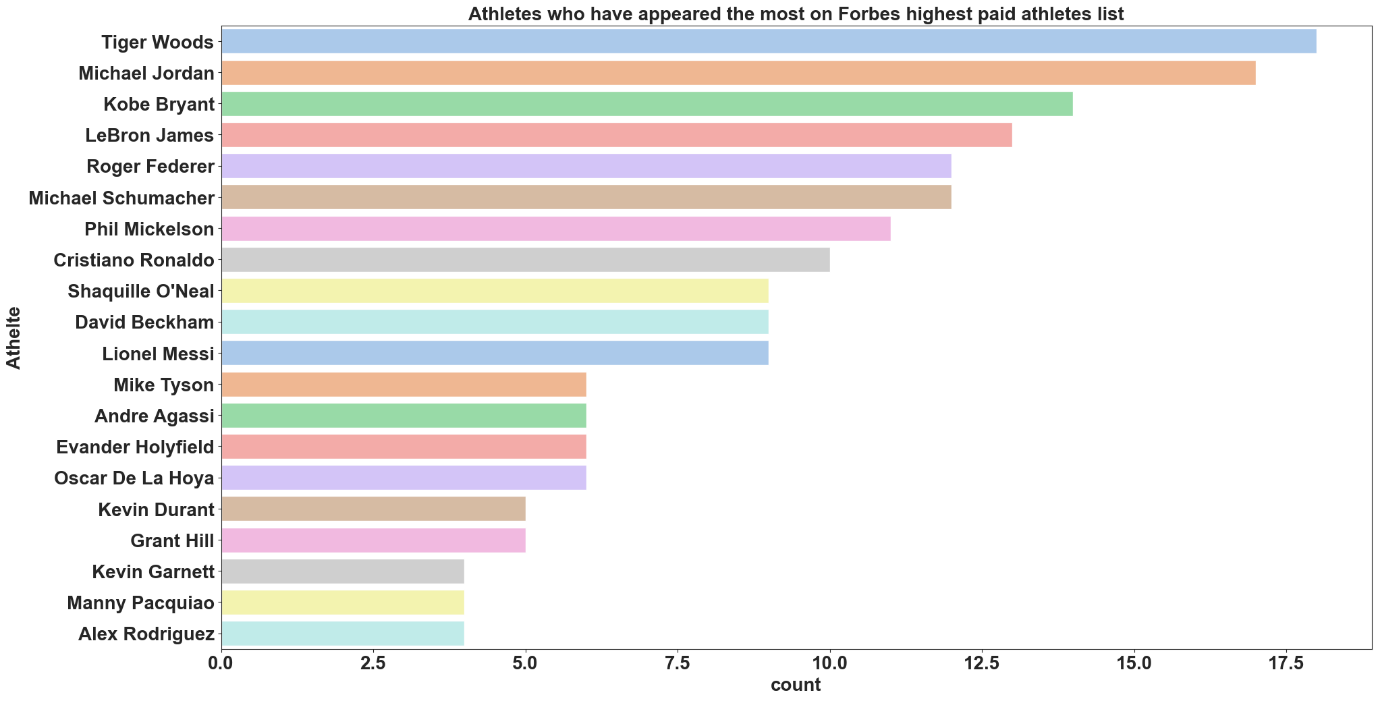
Next, a count of the total appearances of each country in the nationality column.

It is clear that the list is dominated by the USA, which could allude to the sports industry being considerable larger and more lucrative than that of other countries.

****

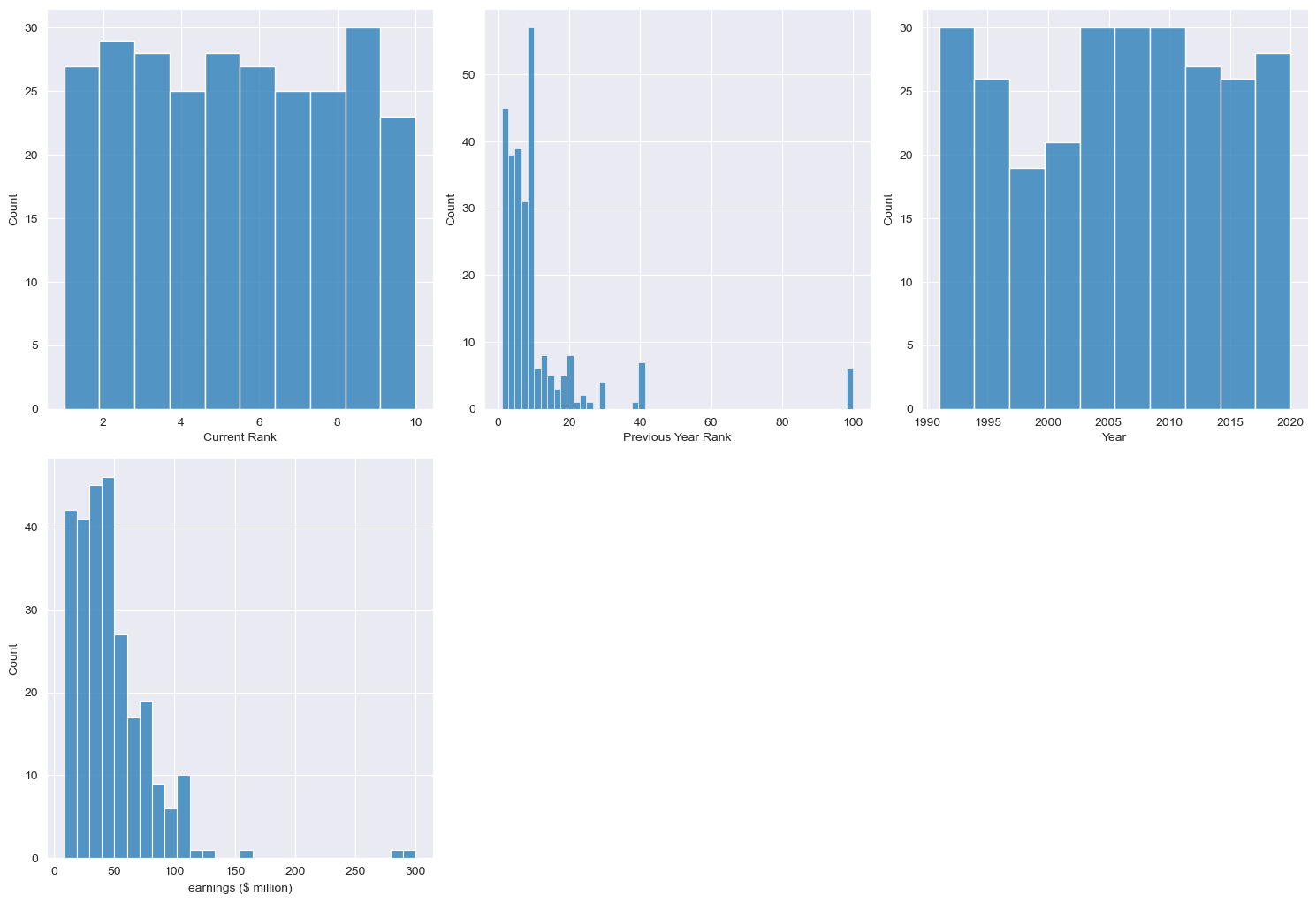
Figure

The following histogram visualises the number of appearances made by the top 20 athletes between 1990 and 2020 in the list. The three most frequent athletes to make it onto the list are: Tiger Woods, Michael Jordan, and Kobe Bryant. This reveals that not only are more Americans on the list, but also the highest earners even among the dataset of highest earners in sport are all American. They are also all celebrity figures who have very lucrative sponsorship deals. Although this is likely true for the majority of the athletes in this dataset.

****

Figure

The next is a multiplot of all the non-object columns. This was done mostly for the sake of completion, and on the chance that any new insights might be discovered.



Figure

Here, the first and third graphs do not show very much other than what was obvious, except that maybe the 9th rank is tied with two or more athletes more often than other ranks.

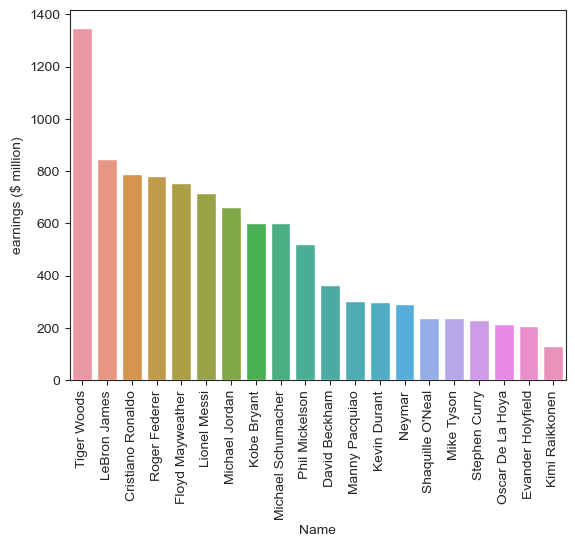
The second and fourth graphs show the previous rank of people on the list over all years and the distribution of the earnings of all athletes. For the second graph, it is interesting to see that overall people on the list stay on the list. I.e. if you rank within the top 10, you’re likely to maintain the top 10 rank in the following year. On occasion athletes can rise from as high as 40th into the top 10 and even on a few occasions from 100th into the top 10. The final graph shows that the majority of these athletes on this dataset earn between 10-100 million USD, with the average around 50million. One or more athletes have made as much as 290 million on rare occasions.

**Bivariate Analysis:**

Next for the bivariate analysis, visualisations of multiple variables at once.

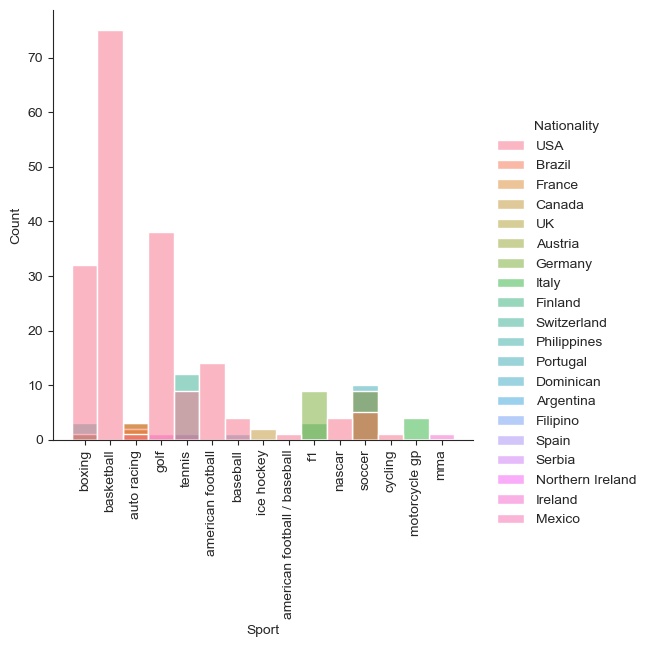
This following graph shows the gross earnings of the top 20 athletes over each year that they are present within the dataset. We can see that this has some similarities with the graph showing the most appearances on the Forbes list.

However, this isn’t a perfect correlation, with Michael Jordan appearing the 2nd most often, but earning only the 7th highest total over all of their appearances. Kimi Raikkonen the F1 driver also make it onto this graph, while not appearing among the 20 most common athletes.

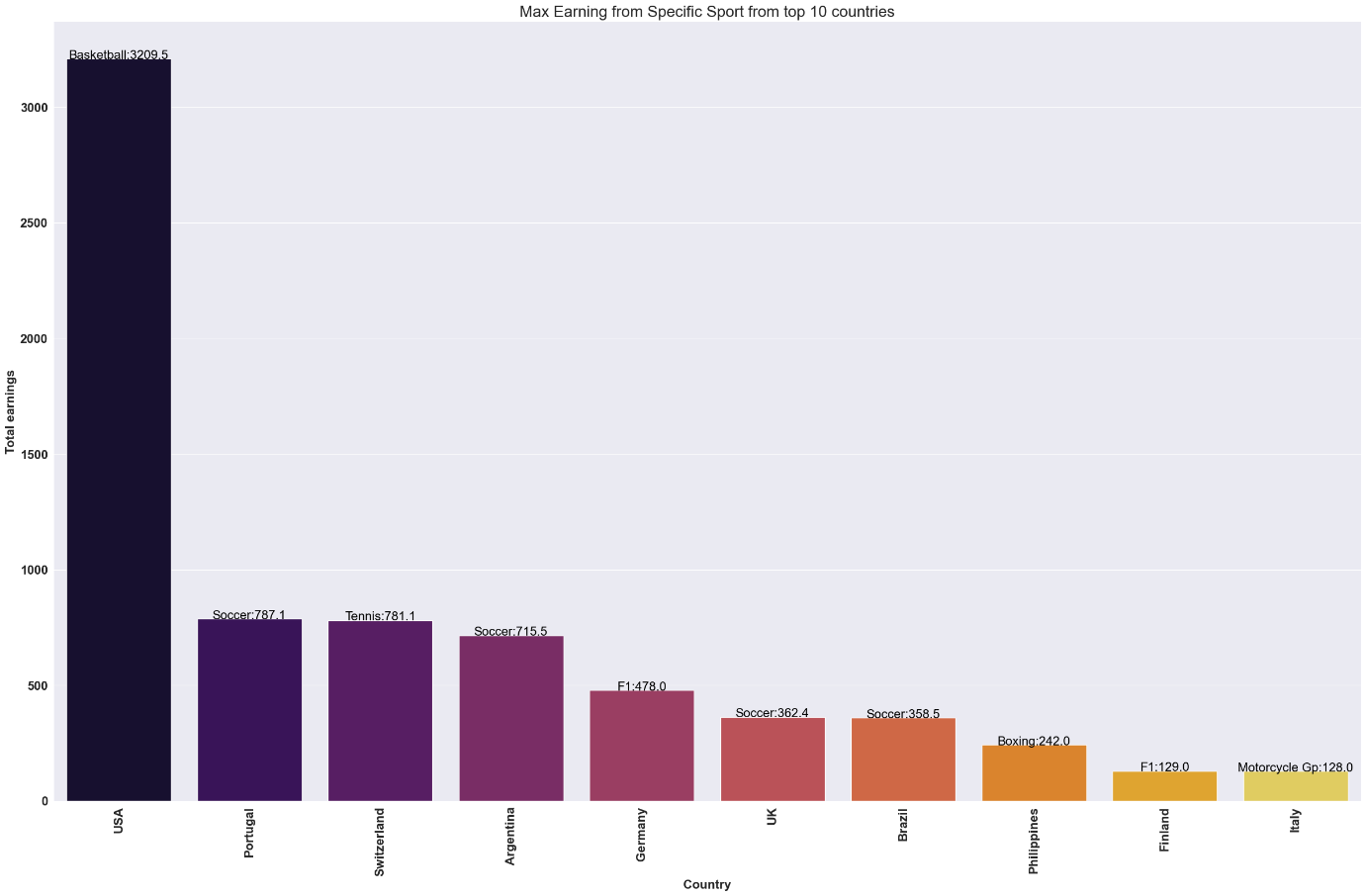


Figure

Next, a distribution of how common each sport was within the dataset divided into which nationalities make up these totals. It can be seen from this graph that the sports popular with Americans are dominated by Americans and do not have many athletes from other nationalities. This means that American sports have much more money into their most popular sports than other countries. Noticeably, sports with little or no Americans among the highest earners are ice hockey, auto racing, F1, soccer and motorcycle GP, Tennis, and MMA.

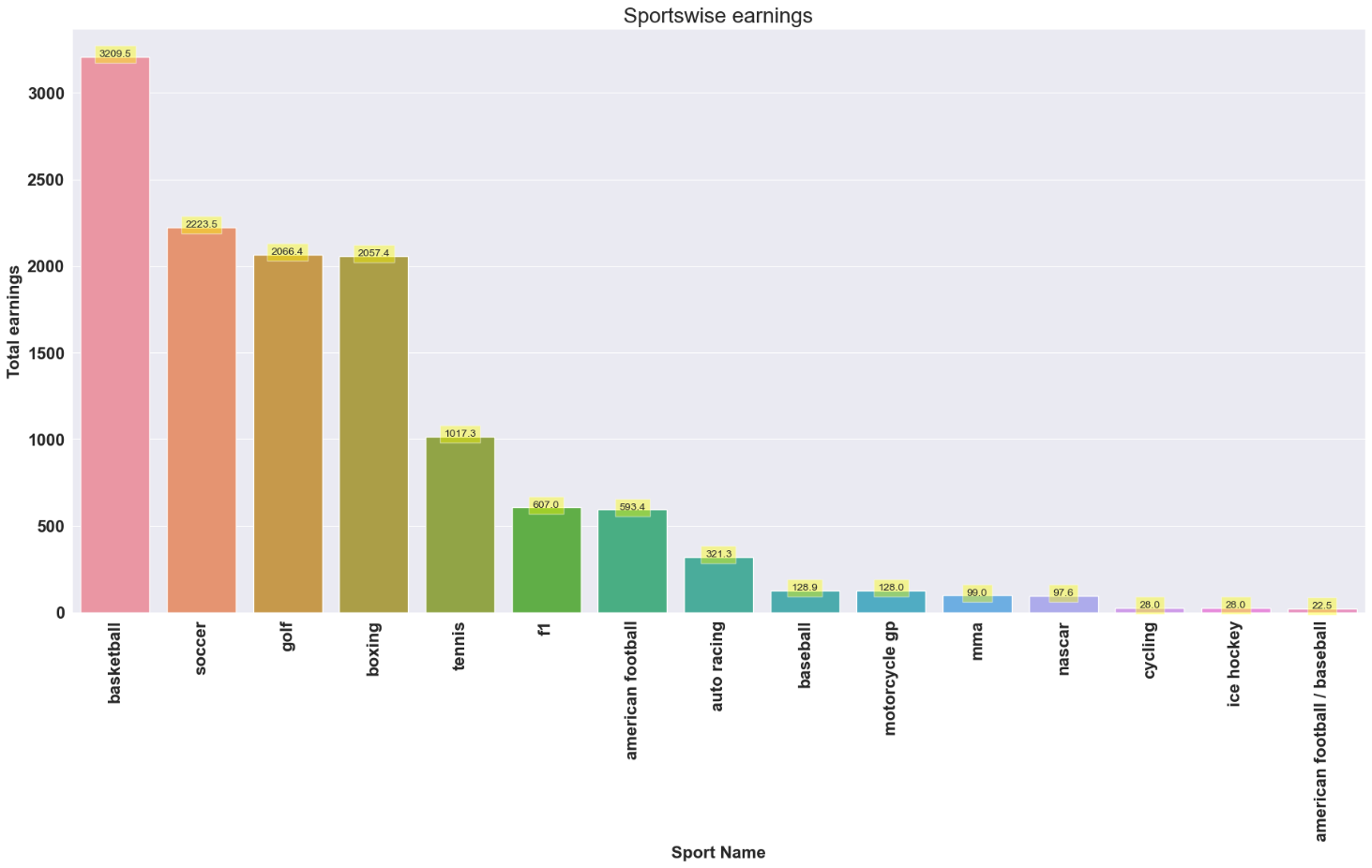


Figure

Another graph that shows similar, but more specific information as the above is below. For the top 10 highest earning countries, each country’s highest earning sport is shown. American basketball has by far the highest earning among each countries highest earning sport. Interestingly on this list, Portugal’s soccer has the next largest share of the Forbes list. This, as can probably be guessed is soley due to Cristiano Ronaldo’s 10 appearances on this list. As is the same situation with the third place Swiss Tennis which is made entirely by Roger Federer’s earnings. It is in fact the case that the USA’s total in basketball shown below is one of the few that consist of multiple athletes. 

Figure

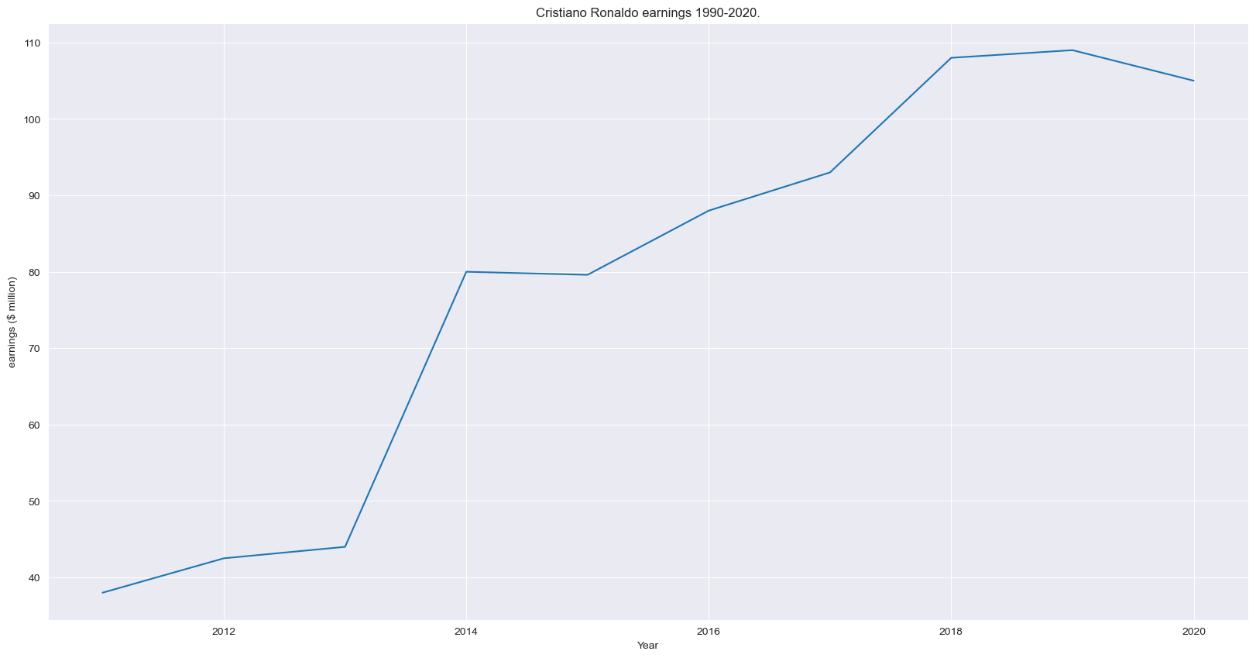
Following is the net earnings of all sports that appear on the list. This shares a lot of insights that the nationality breakdown of sports graph, but shows that many of the high value American sports dominate this list as well, the main exception is that soccer, the highest earning sport of many countries comes in at second but is significantly behind Basketball.



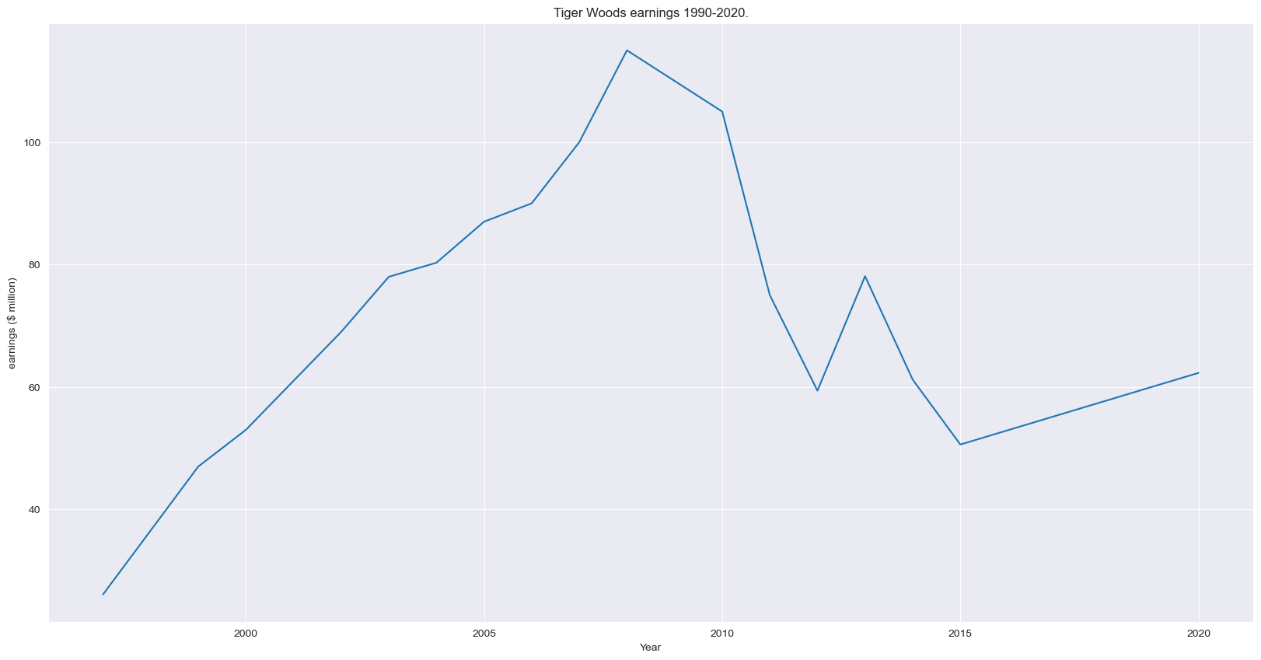
Figure

Finally, a function was built to show the earnings over time (as long as they made the top 10 ranking on the Forbes rich-list) for any athlete who appears at least once.

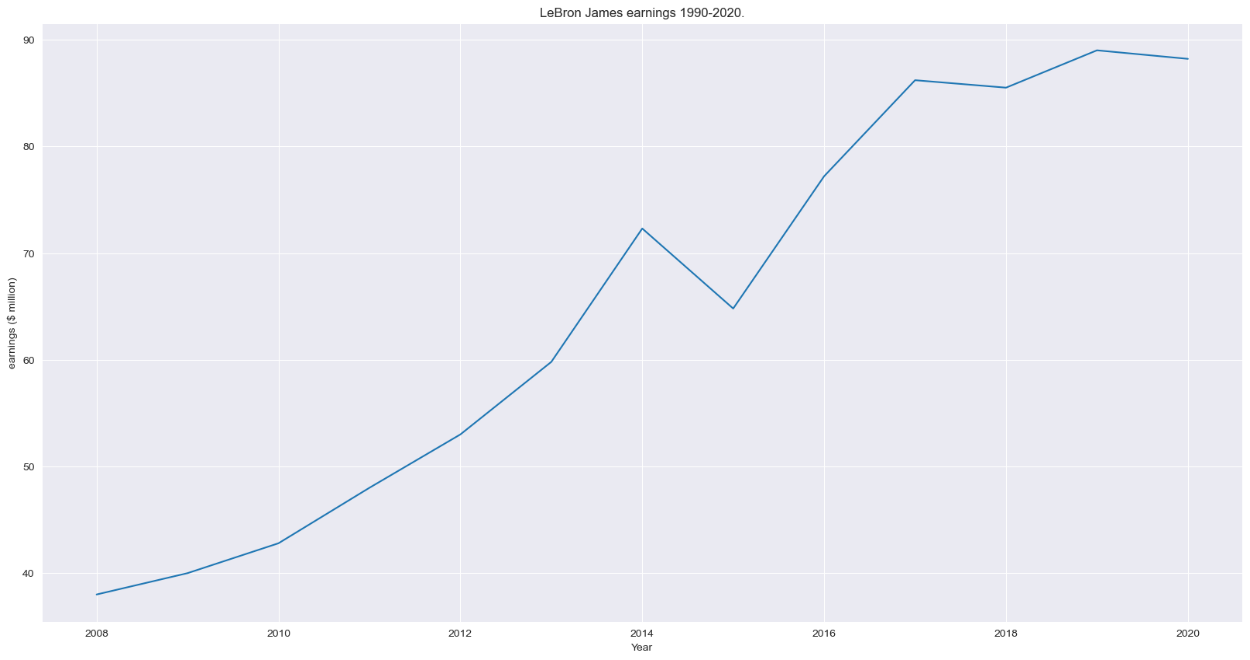
Some of the most prolific athletes’ earnings across their career are shown below.



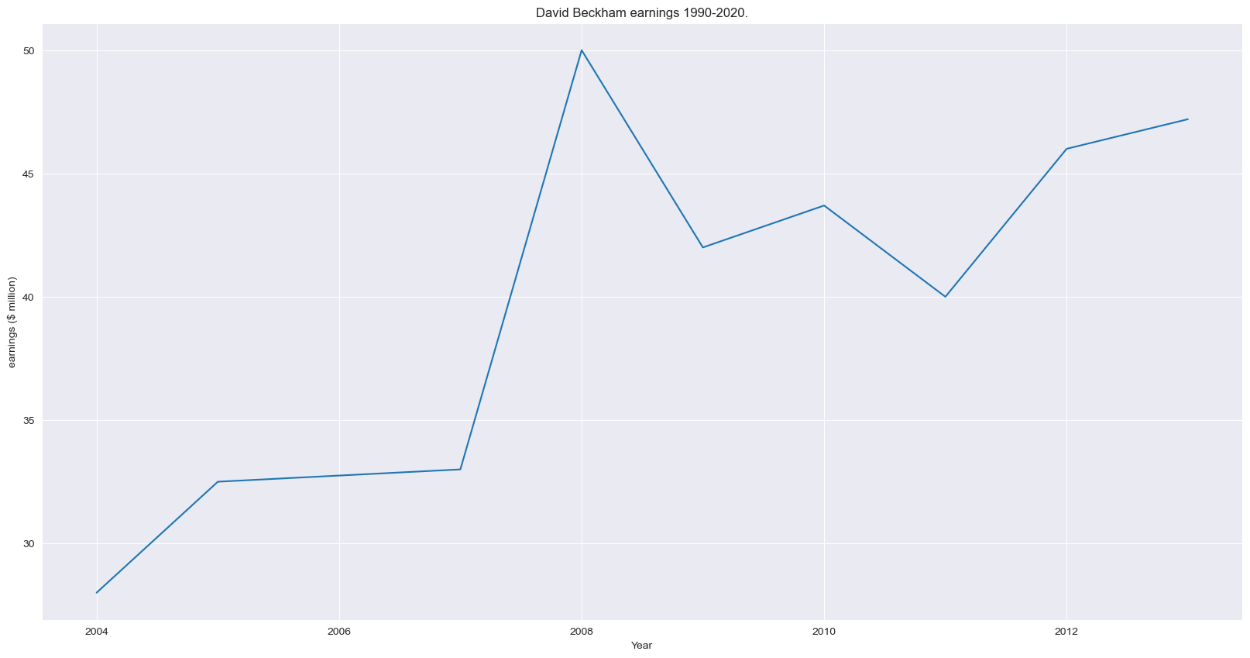
Figure



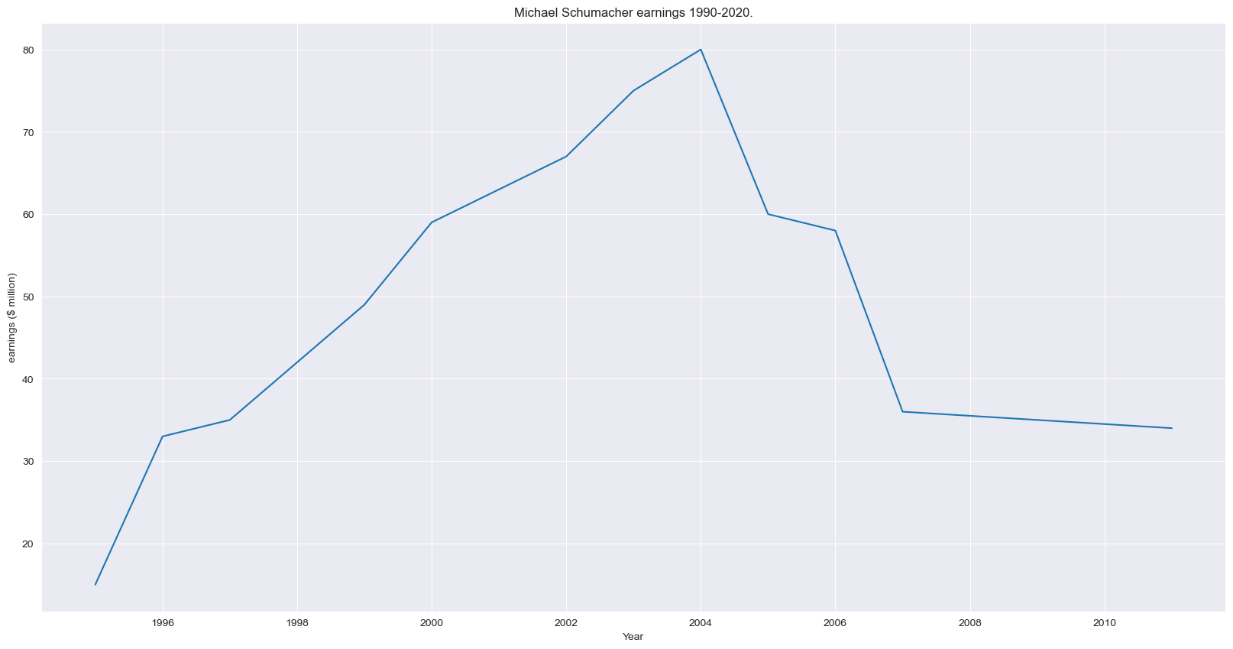
Figure



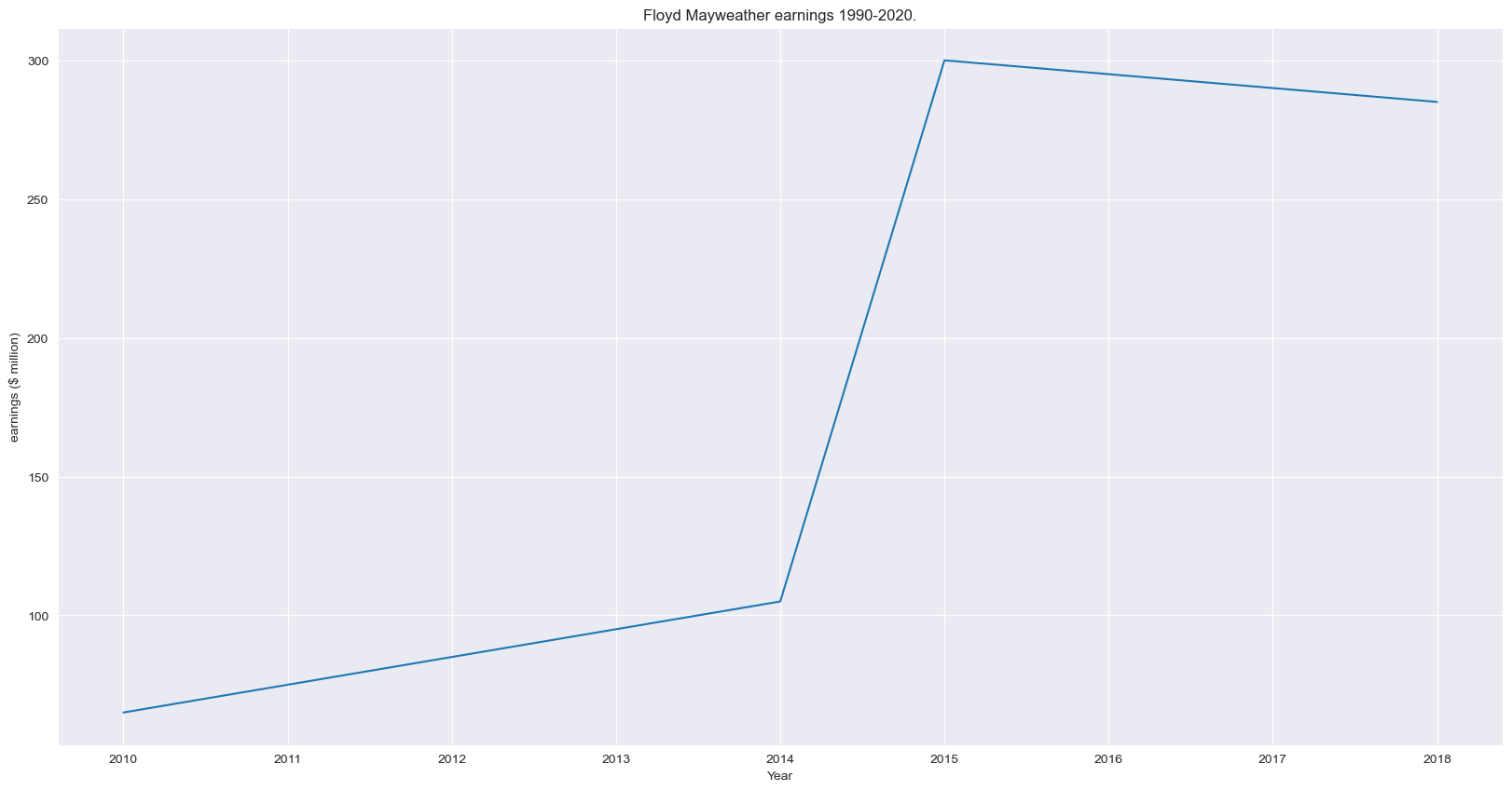
Figure



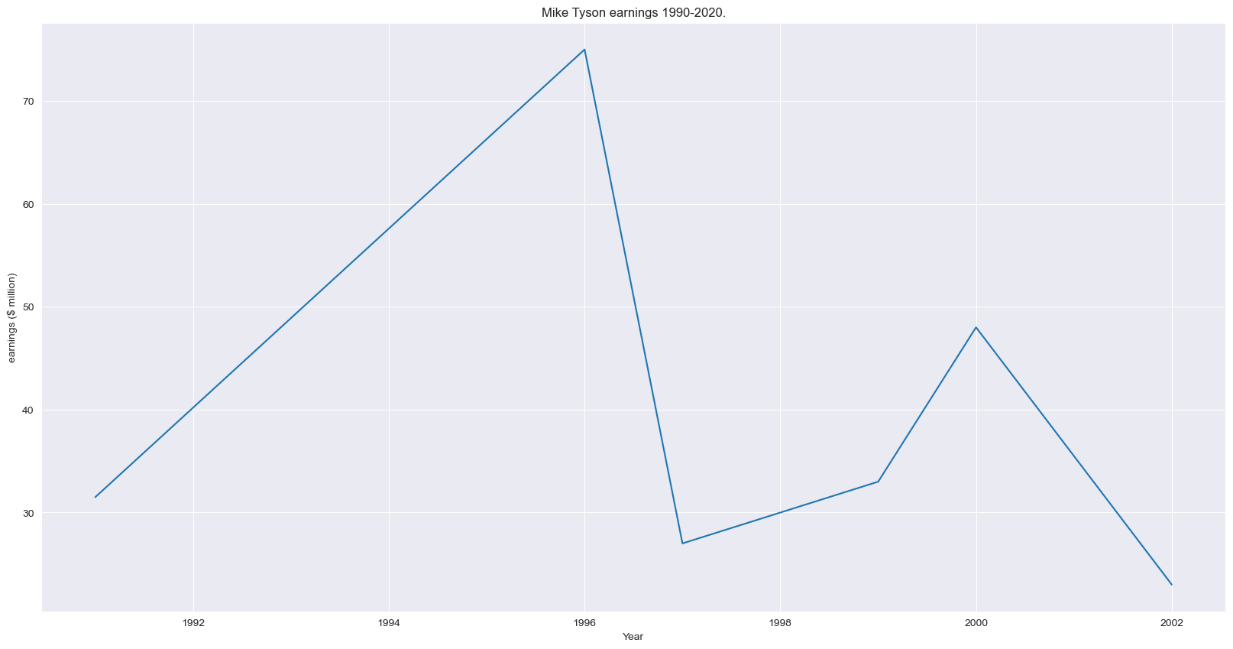
Figure



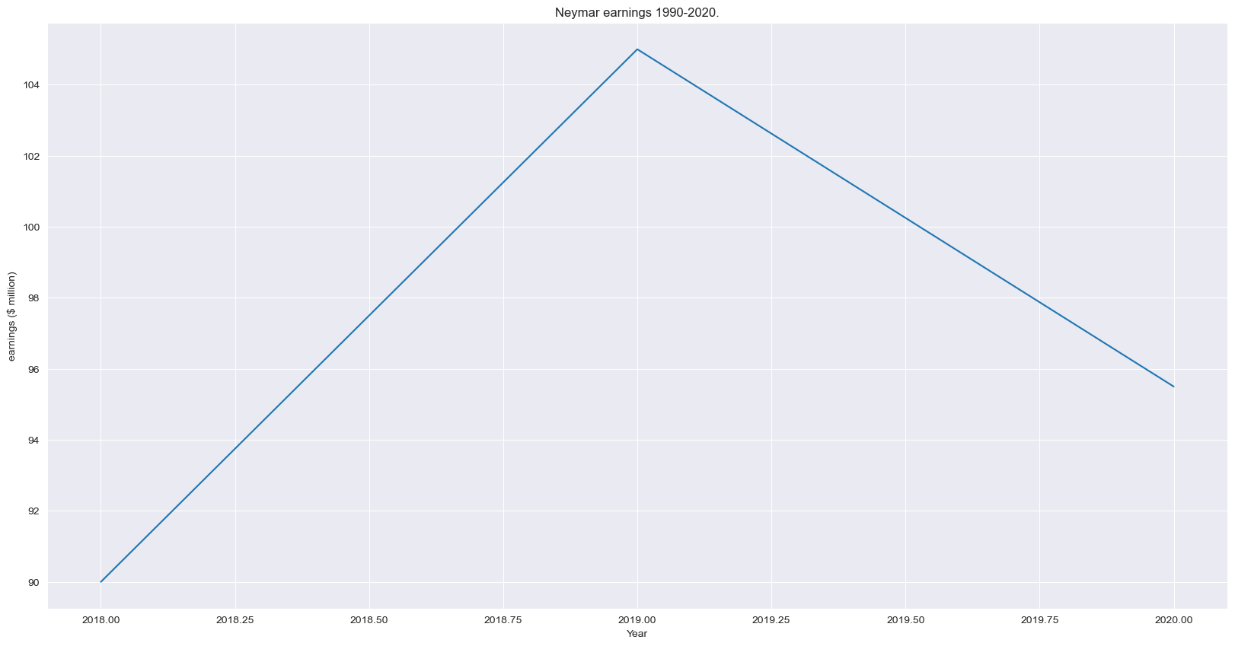
Figure



Figure



Figure



Figure

Here we can see some interesting stories. Firstly, it is rare for even the most famous / successful athletes to maintain these gigantic incomes throughout their career and only some are able to consistently increase their income, instead rather, there are often large swings that happen in one year and then decrease over time after that. A surprise was that for certain athletes, Tiger Woods and Mike Tyson in particular, that even well publicised scandals could not see them fall out of the top 10 ranks; their income significantly drops, but not enough to see them drop out of the top 10 rankings. This shows how dominant these athletes were as the face of their sports.

**CONCLUSION**

**Data Stories**

*Data Story 1: Does ranking top 10 once make it easier to continue ranking top 10?*

This is absolutely yes. The highest earners in sports are not receiving such a high income based off of their sporting performance alone, although that does contribute to them being able to maintain the other factors that can contribute to these high earnings. An athlete’s performance will dictate their playing salary and any prize money they receive, but being famous and the face of a brand also contributes to this. As shown in figure 4 and also through the individual earnings over time, figures 9-16, as a player becomes more prolific in their sport, their earnings increase. However, even if they stop playing or their ability in their sport declines as shown in figure 12, their earnings can continue to rise. In this case, David Beckham became the face of many brands during his time as a world-class player, and this image stayed even as he slowed down and eventually retired, he continued to reach the same heights as when he played. Because of factors like this, it is unlikely that an athlete will suddenly appear on this list and then suddenly disappear, although this does happen on occasion. What appears more common is that they will slowly drop off over many years. Even athletes who have had scandals see their income remain in the top 10. Mike Tyson after biting Evander Holyfield’s ear saw his earnings fall dramatically, but not enough to ever see him drop out of the top 10 ranks. Once an athlete has reached the level that they can appear on this list once, it seems rather straightforward to leverage that position into maintaining such high earnings.

*Data Story 2: Is Geography the most important factor in high earnings?*

Leading on from Data Story 1, does the location of a player significantly impact their odds of ranking in the top 10. Firstly, a major part of earning potential comes from how popular the sport is. This seems logical as more people interested in the sport means more money will be circulating in that sport’s bubble. However, according to google search trends[[1]](#footnote-1), Soccer received an overwhelming 39.7 relative search score over the period of 2016-2020 while basketball scored 7.8. In fact, cricket received 7.9 placing that as the second most popular sport based off of this metric. However, we see zero athletes reaching the top 10 highest earners from the sport of cricket (figure 1). Even among soccer and basketball, the gross earnings among those who play basketball is significantly higher than those who play soccer, despite soccer being overwhelmingly more popular globally. It continues to appear this way that sports popular in America see the highest earners compared to sports famous globally. This suggests that perhaps for world class athletes, geography could be the most influential factor in their earning potential.

Playing in the USA, in a sport that is popular in the USA can see your earnings soar in comparison to being an equal level athlete in more popular game globally, but with a smaller piece of American fans’ attention.

**THIS REPORT WAS WRITTEN BY: Ryan Pitt**



1. https://www.topendsports.com/world/lists/popular-sport/google-trend.htm [↑](#footnote-ref-1)