Hyperiondev

TASK

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

Introduction

This is a dataset that contains the top 10 richest athletes of the year from 1990-2020, along with some details such as what sport they play, how much they earn and what year they made the list. Immediately I wanted to investigate a handful of things, like which sports provide the biggest payouts and how the distribution of wealth in these sports change over time.

DATA CLEANING

Now this dataset absolutely needed some elbow grease to get clean. Some aspects were fine, such as the data types and the lack of duplicate rows. The sheer amount of typos and variants however required a little extra time to put right.

First of all we had some typos in the 'Name' column, "Aaron Rodgers" and "Shaquille O'Neal". This got fixed in the exact same way as the rest of the typos and variants I am going to mention, simply just renamed all mistaken occurrences with the correct spelling.

Nationality also had a few discrepancies, for example both 'Filipino' and 'Philippines' were present, since the other entries were nouns and not an adjective like 'Filipino', I decided to change it to 'Philippines'. Then onto maybe the first controversial decision, both 'Northern Ireland' and the 'UK' were present. I consulted with an Irish friend of mine and they assured me that although 'Northern Ireland' isn't a fan of it, they are still a part of the UK. With this blessing, I replaced all instances of 'Northern Ireland' with 'UK'. I was however warned that if I were to do that with the purely 'Ireland' entries, there would be trouble. Which pleasantly ends our corrections in this column.

Now onto the chunky portion of this cleansing, the Sport column. Similar to the occurrences in the Nationality section, there were a plethora of variants that all meant the same thing in this section, namely with motorsports. Firstly however I noticed that there were a lot of entries that meant the same thing, but they had a different casing with some being upper, some lower and some title. To fix this, I just changed all entries to lowercase for the time being and this handled the majority. The following all existed in the sport column: auto racing (nascar), auto racing, f1 racing, f1 motorsports, nascar and motorcycle gp. This isn't including the ones that were case sensitive before, so I decided to merge them all under the heading of 'motorsports'. I initially tried to split it into nascar and f1, however I noticed that auto racing had a mix of f1 and nascar drivers, and took this as a sign to unite them

as one. There was also a strange occurrence on this list, there was an entry which was ‘American football / baseball’. Initially I thought this was a wind up, however after a bit of research, athlete Deion Sanders actually competed in both which is quite phenomenal, talk about an all-rounder (is that a baseball reference?). Since Deion participated in baseball when he was entered into the list however, I changed his sport to just baseball in the end.

To wrap this section up, I combined nfl and american football to be named the latter, then did the same to nba and basketball. I then changed MMA to Mixed Martial Arts and then changed the entire column to title case.

I also renamed the ‘earnings ($ million)’ to be title case too, since it wasn’t before and looked out of place next to the other dapper looking, title cased column names.

Last but not least, I also dropped the ‘Current Rank’ column, since I wasn’t interested in working with it.

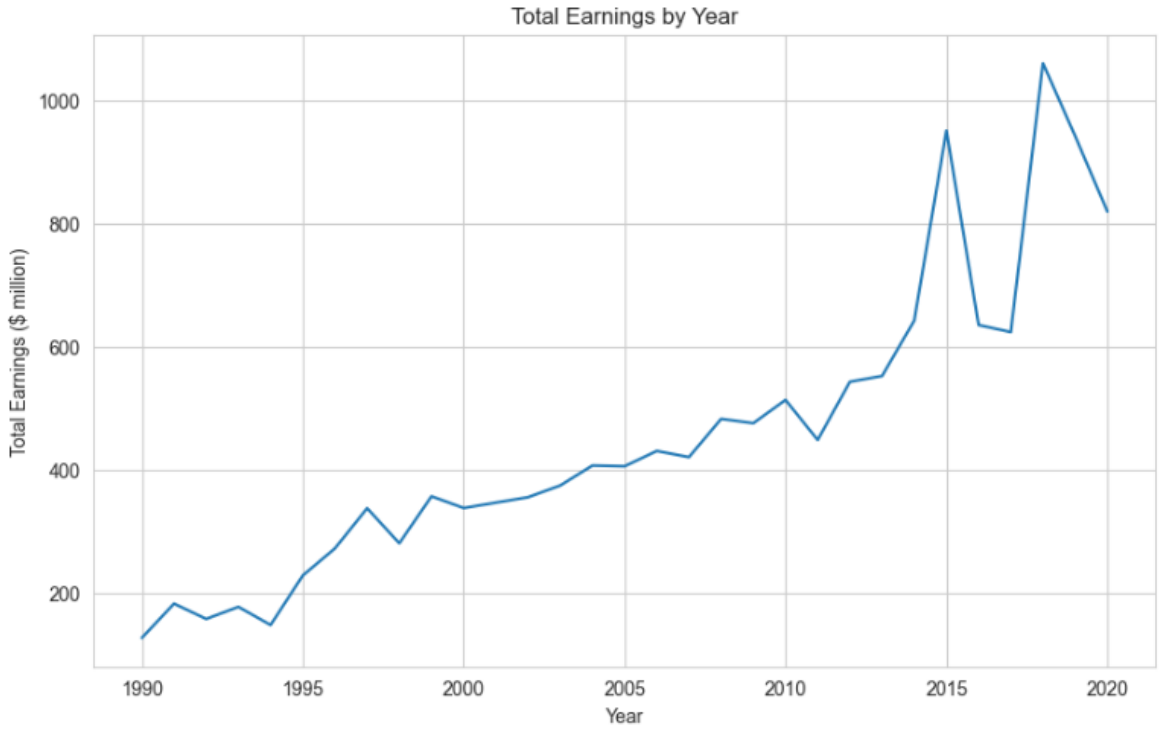
MISSING DATA

The only missing values present were all located in the ‘Previous Year Rank’ column, which was only the first sign of bad things to come with this column. There were Null values present, missing values, ‘??’ and approximations in the form of ‘>10’ or ‘>40’.

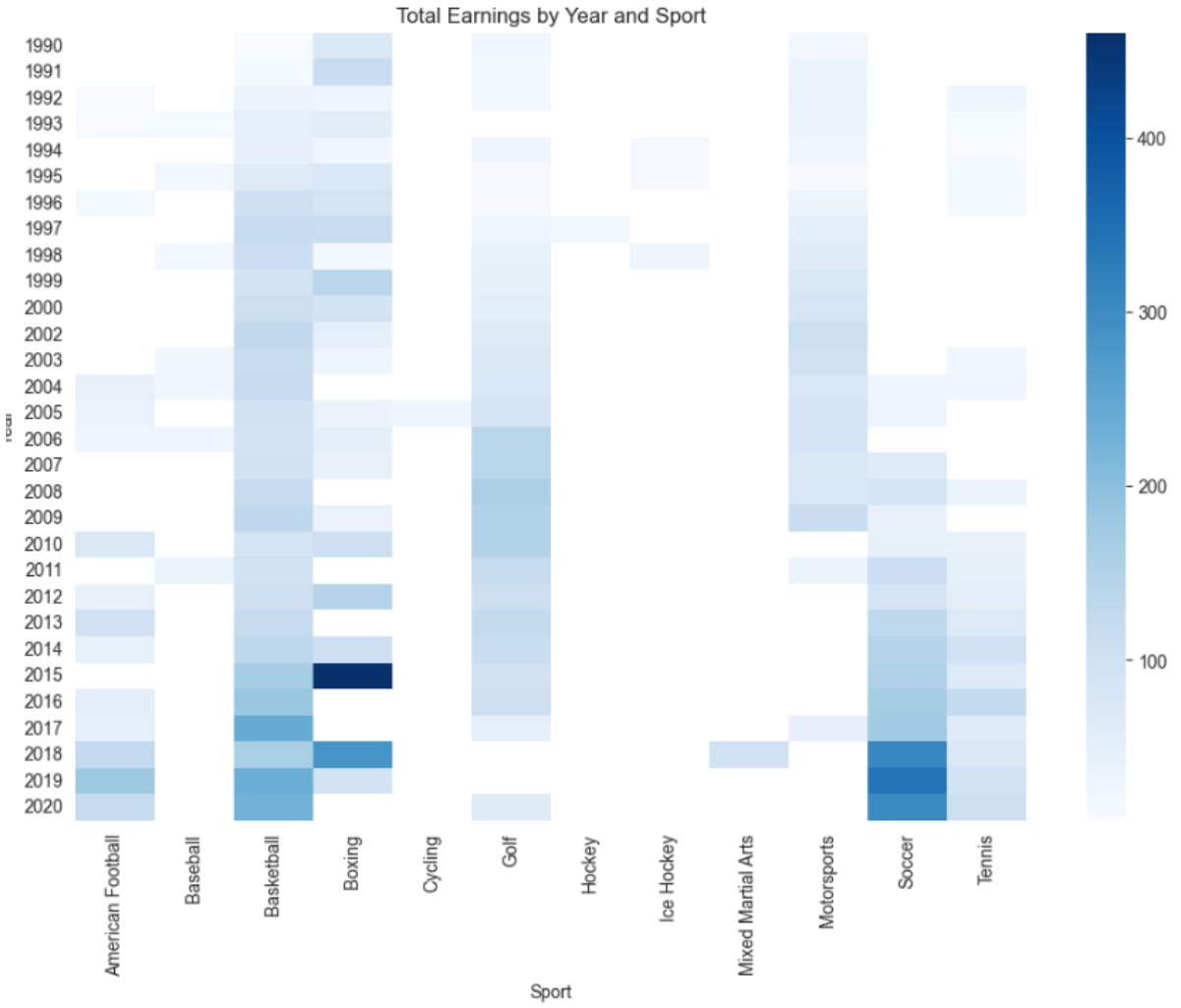
Since there was data already missing, the existing data being vague and the cherry on the cake being that I wasn’t even going to use the current rank column, I ended up also dropping this column too. Which very conveniently removed all the missing values for me.

DATA STORIES AND VISUALISATIONS

For this investigation, I was very much interested in how sports has changed over the years. Answering the questions like, is there a certain sport or person who gets consistently paid the most? Does sports interest change often? How about the diversification of the athletes? As scuba divers often say, let's dive in.



I thought I would keep it nice and plain to begin with, cue the humble line plot. This right here shows how the total earnings of the year has gradually increased throughout the years, at first it did so at quite a calm and steady rate, until the more recent years at 2015 when it all became a little hectic. Since we do not have enough data after 2015, I can't confidently say whether or not 2015 and 2018 established a new status quo when it comes to payments, or if they were just really good years for sports. The reason I say this is because for two years in between the earnings seem to nosedive before rising up again, having a look at where it lowers to it seems as if it goes back to where the steady increase would have also taken it in due time. My guess is that the rare and fabled blue moon rose in the clear night sky those years, that blue moon of course being boxing.



This is a heatmap that displays how much each sport got paid by the year, the darker the colour the more they get paid. I'll now explain my blue moon comment, as you can see the majority of the sports either have some reliability, or were one-offs on the Forbes list. Nothing has a reputation however for being sporadic, unless you're boxing that is. Over the years it seems very hit and miss, taking year long hiatus' almost consistently. When it does go ahead however, it does exceptionally well, just look at the last few times it has appeared, having some of the darkest coloured cells on the map. With this I conclude that the baseline increase in pay hasn't shot up out of nowhere, instead the boxing just happened to be on that year and either due to exemplary marketing or some other reasoning, the public just absolutely loved it.

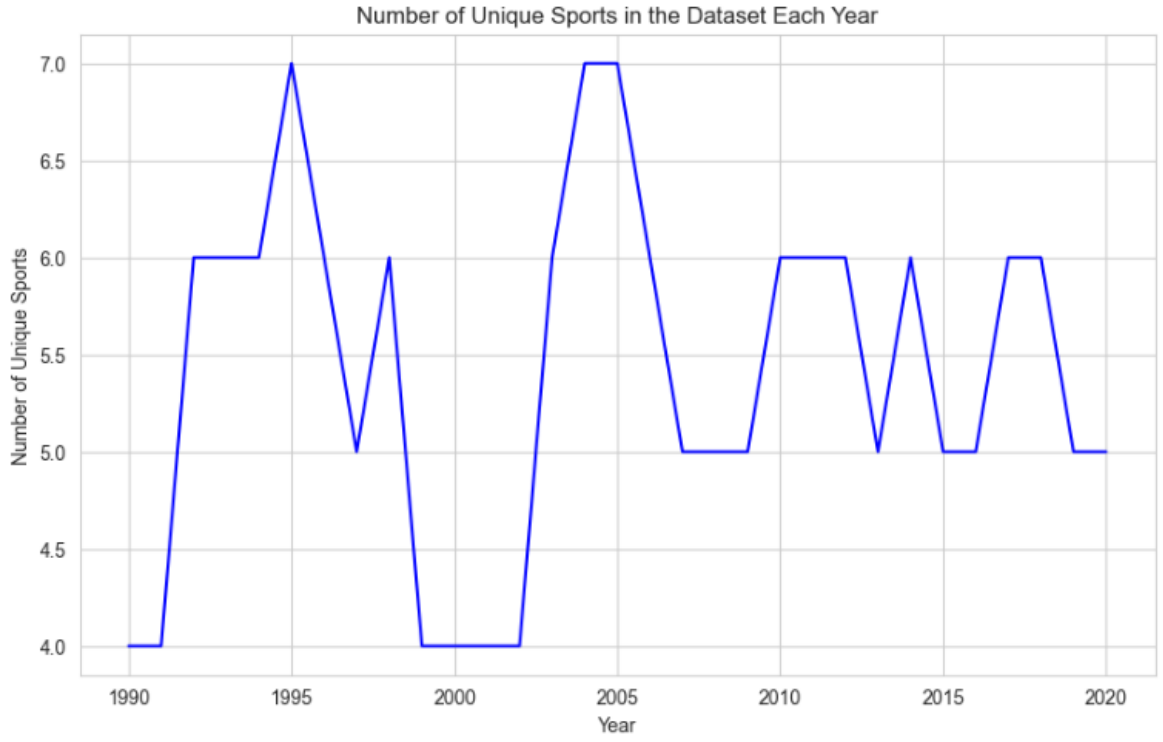
Some other observations we can make here could be that we definitely have some usual suspects, such as basketball and golf who are consistent members of this list, rarely missing a year. Sports such as soccer, tennis and American football however seem to be joining the fray too, by picking up a lot of steam at around the same

time (2011). This must be good news for the world of sports, who from the looks of it desperately needed some new content.

Soccer and basketball are in quite unique positions, since they have both received an extremely large pay rise in the last 3-5 years, possibly a hefty investment to capitalise on how much the public are eating up sports currently?

As for both motorsports and baseball however, it seems as if the best days are behind them now, with motorsports having a clear and consistent participation on the list and baseball struggling to even make it on there, but both fizzling out at around the late 2000s.

There are also a handful of other sports on this list, but they all seem to be one off additions during the late 90s and early 2000s, with the exception of MMA making an appearance in 2018.

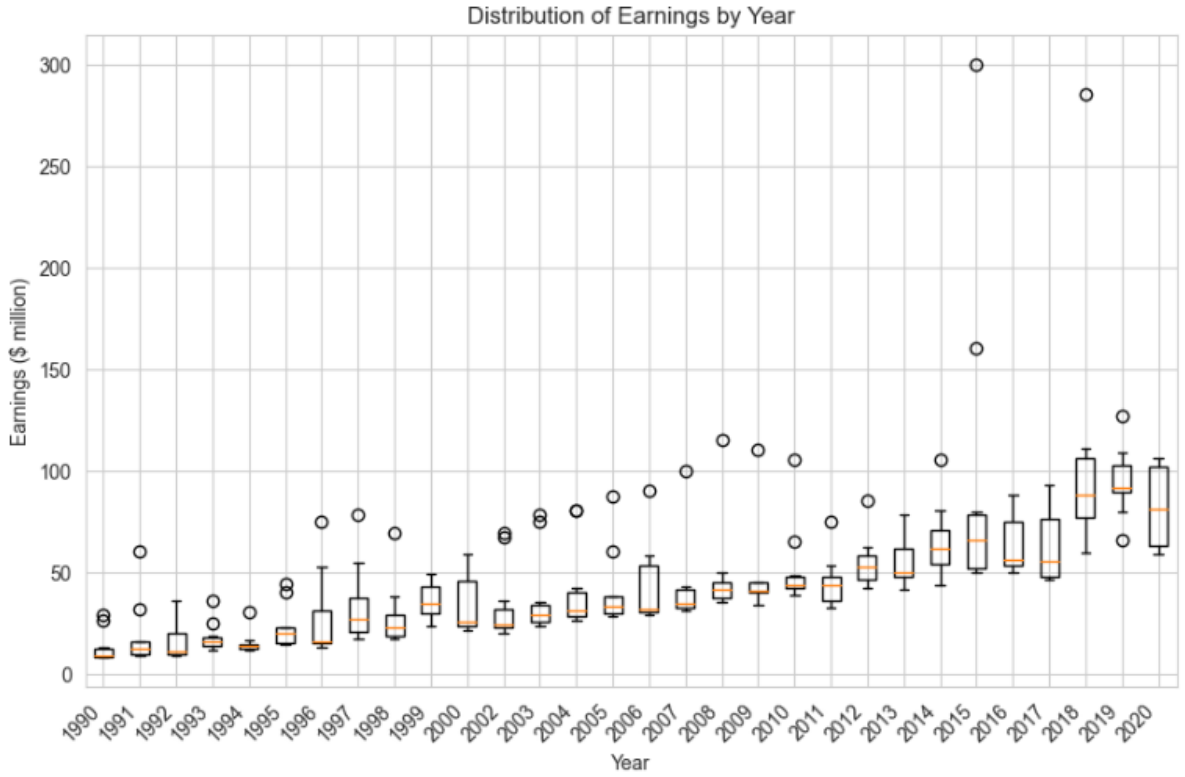


Next up we will be looking at the diversity of the sports on the list. It seems that the minimum seems to be that 4 different sports are active each year, this makes sense since you will usually have basketball, golf, boxing and some other sport; whether that be motorsports or soccer, which seemed to start picking up when motorsports died.

The average however lingers around the 5-6 range. After analysing the heatmap a little more I happened to notice that in the earlier years of this list, the sport that pushed it up to the 5-6 range tended to be baseball, american football and tennis

almost taking it in turns to be present, with the occasional appearance of some not so common sport like hockey. In more recent years however it doesn't seem to be as sporadic, with sports such as soccer, American football and tennis beginning to become a lot more secure in their positions.

There also seem to be much fewer outliers, with the only one since 2005 being MMA. Maybe this is a sign of sports entering an age that is a lot more stable, establishing its roots in cultures around the world to stay for the long run. This begs the question of whether or not sports are to become a lot less diversified now that a select few have outperformed so well.



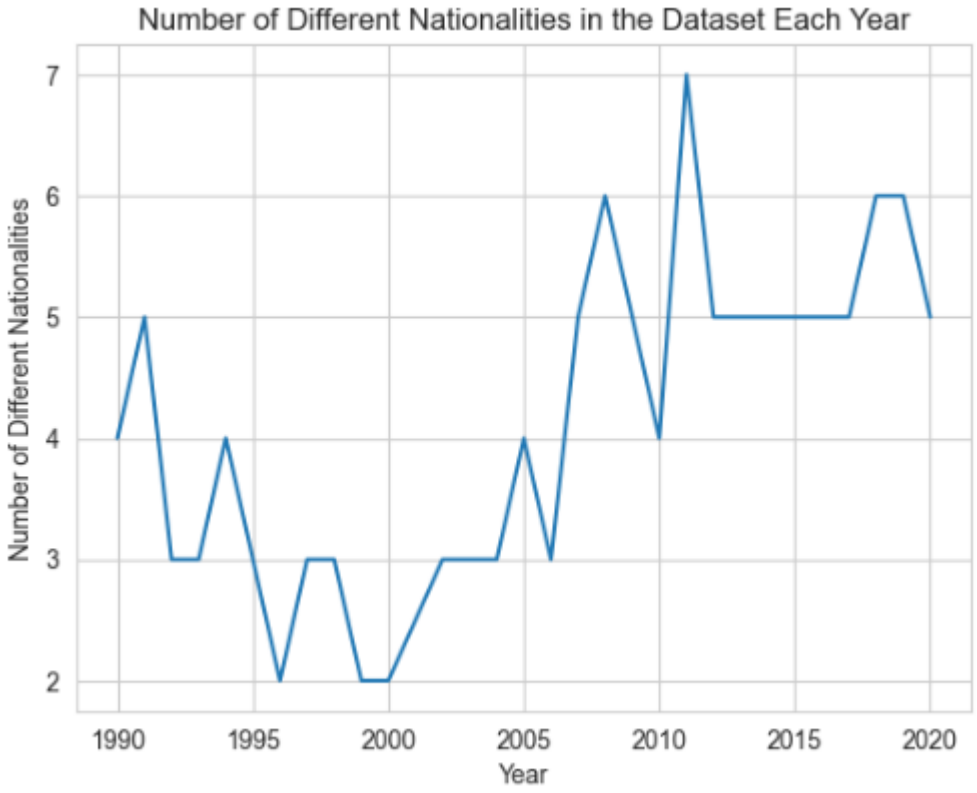
Here we have the distribution of earnings per year and interestingly enough it seems as if the wealth is mostly distributed towards the lower end of the y axis, with the big payouts almost exclusively being outliers in their times.

The minimum pay however is steadily increasing, there are only a handful of years towards the right hand side of the graph that seems to dip back down temporarily. This could be explained by either inflation or sports becoming much more popular since video media has become much more accessible.

There seems to be only one occasion where there is an outlier that appears underneath the minimum pay of that year, this occurs in 2019. From a glance at the heatmap it appears that this occurred in either boxing or tennis, which is very uncharacteristic for boxing since they're usually suckers for a big payday.

As for the outliers on the upper half of the graph, it seems as if they just emulate the line plot displayed earlier in this report.

#



Last but certainly not least however let's have a gander at the diversity of nationalities in the list over the years.

The minimum value we seem to have here seems to be 2 during the 1996-2000 period, which partially coincides with when the diversification of sports also had a temporary plummet. Basketball, Boxing, Golf and Motorsports reigned supreme during this time-frame, meaning that the USA possibly dominated the list during this time.

Very suddenly afterwards however the diversification began to shoot up, until the grande apex of 7 was reached in 2011, then plateaued out at around 5-6, establishing a nice level of variance in the list which seems to become the baseline from here on out.

CONCLUSION

I believe that the main take from this would be that in the earlier years, the sports industry went through some drastic change, with technological advances changing the way that sports was both displayed, advertised and discussed between fans and rivals. This came to the favour of some sports, and the downfall of others which couldn't quite keep the momentum going. Those who do establish their footing however will become the sports giants of tomorrow in the form of American Football, Soccer, Basketball and Tennis.

It is prevalent in some sports, like soccer and basketball, that they are aiming to keep the attention on themselves by signing on bigger names with an exceptionally large amount of money, and even branching out by hiring a more diverse set of athletes from around the world.

Some sports like boxing however who play more of a prestige approach will stick to the 'big fight' technique that comes so rarely, but knocks it out of the park consistently.

I will be interested to see however if the occasional outliers will continue to be snuffed out in the recently established shadow of the colossi, and what the next big thing to break the mould will be. Will sports such as soccer succumb to the same fate that both motorsports and Ozymandias did? And whilst we are here too, how did cycling get on this list?

THIS REPORT WAS WRITTEN BY : JOSHUA WATMOUGH