TASK

Exploratory Data Analysis on the Forbes Richest Athletes Dataset

Visit our website

Introduction

DATA CLEANING and SUMMARY of DATA

An initial view of the data shows we have 301 entries and 8 data items, one of which is a numeric ID. The data is split into 3 numeric and 4 object (text based) features. Of these, 'Sport' and 'Nationality' are categorical variables, as is 'Name', as there are separate entries for individual athletes (presumably one for each year they appeared in the list). 'Previous Year Rank' is storing a mix of numbers and characters and may contain nulls. Of the numeric features (ignoring 'S.NO', 'Current Rank' could also be regarded as categorical as could 'Year'. 'earning ($ million)' is float type and is the key data item of interest that will mostly be unique.)

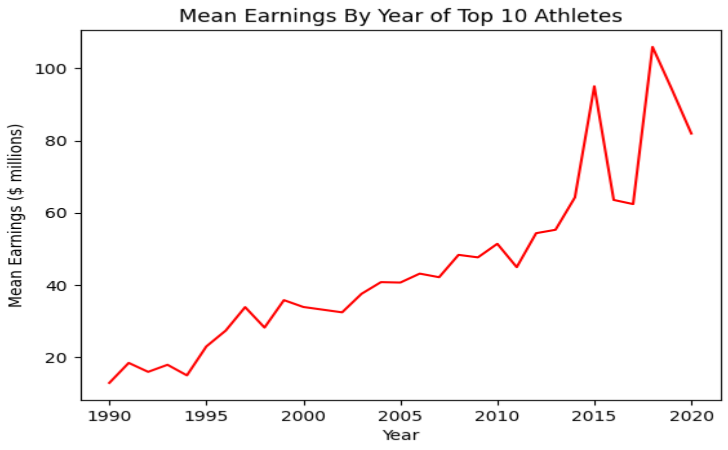
MISSING DATA

A quick look at nulls in the data shows there are considerable null values in the 'Previous Year Rank' feature. This is normal as this feature may well be blank if an athlete hasn't appeared before, and so therefore no imputation is needed here. For the remaining data, I am going to look for unique values in the categorical fields to see if they are consistent.

From looking at the full output in the text editor, I can see a few issues. In the 'Names' feature, it seems like 'Aaron Rodgers' and 'Aaron Rogers' may be the same person. In the 'Year' category, 2001 is missing. In the 'Sport' category there are a number of repeated categories because they haven't been capitalised. There are some entries listed with a nationality of 'Filipino' that should be 'Philippines'.

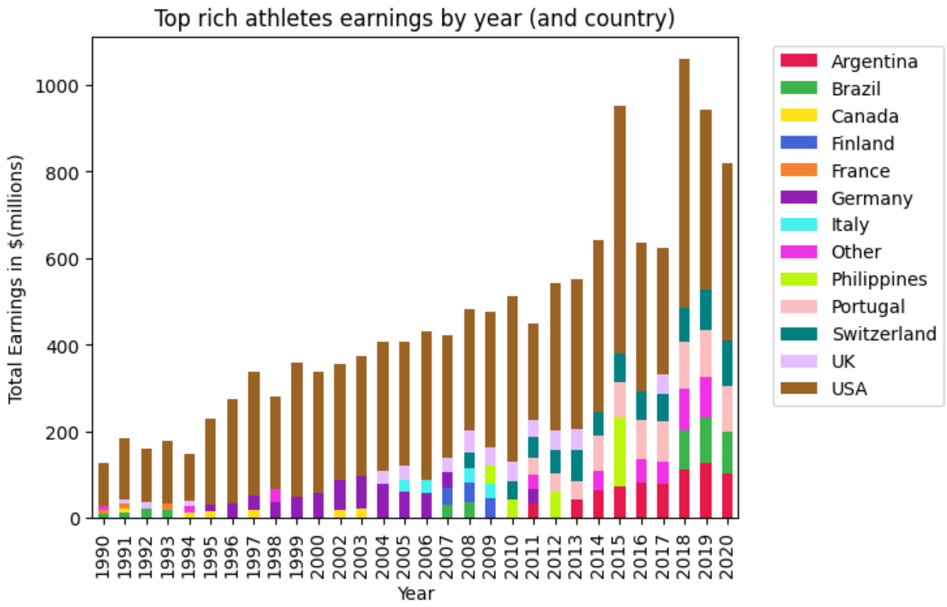
I am going to create a new dataframe and first capitalise the 'Sport' feature. I am also going to change 'Aaron Rogers' to 'Aaron Rodgers' having checked to see if this is an error, and I will change the ‘Nationality’ feature containing 'Filipino' to 'Philippines'. Having cleaned the data, but still with 2001 missing, I am going to proceed with the analysis.

DATA STORIES AND VISUALISATIONS

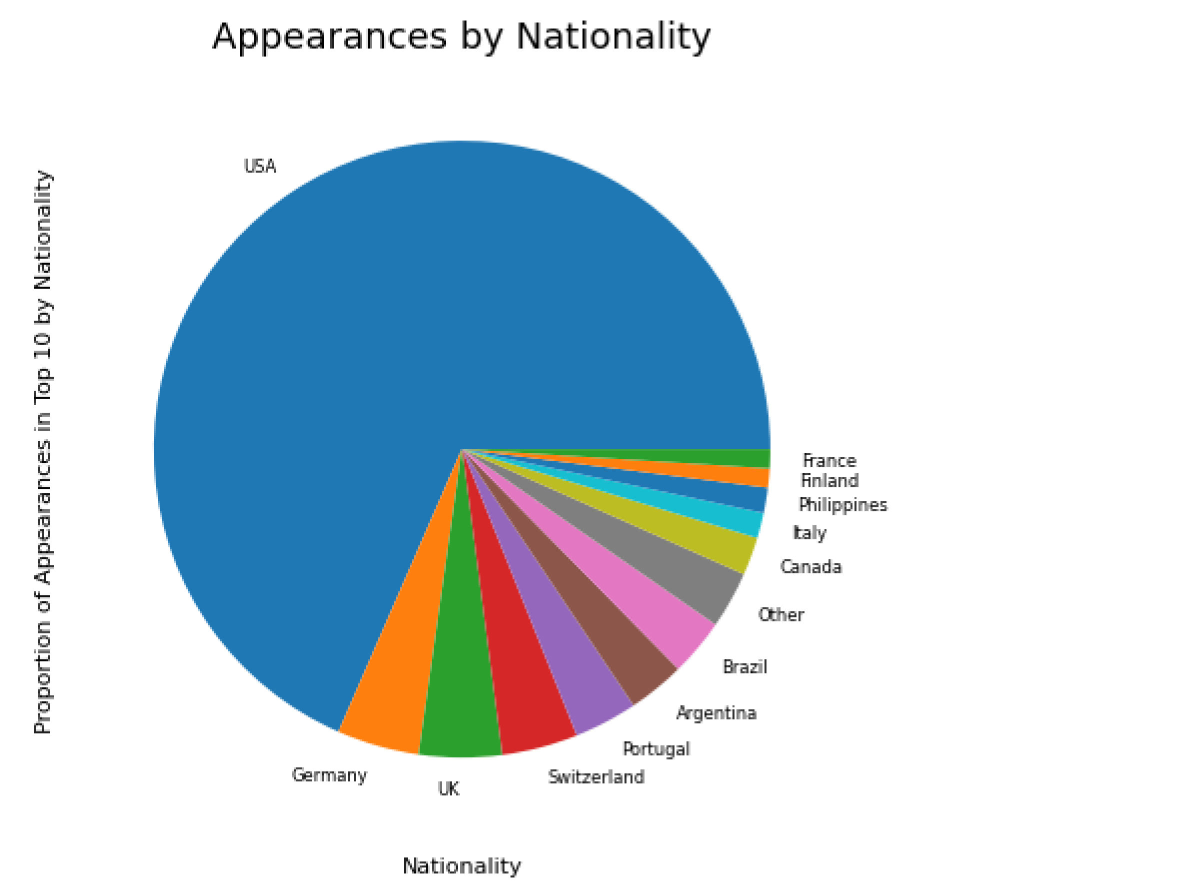


The graphic to the left shows a plot of the average earnings of the top 10 athletes over the 30-year period.

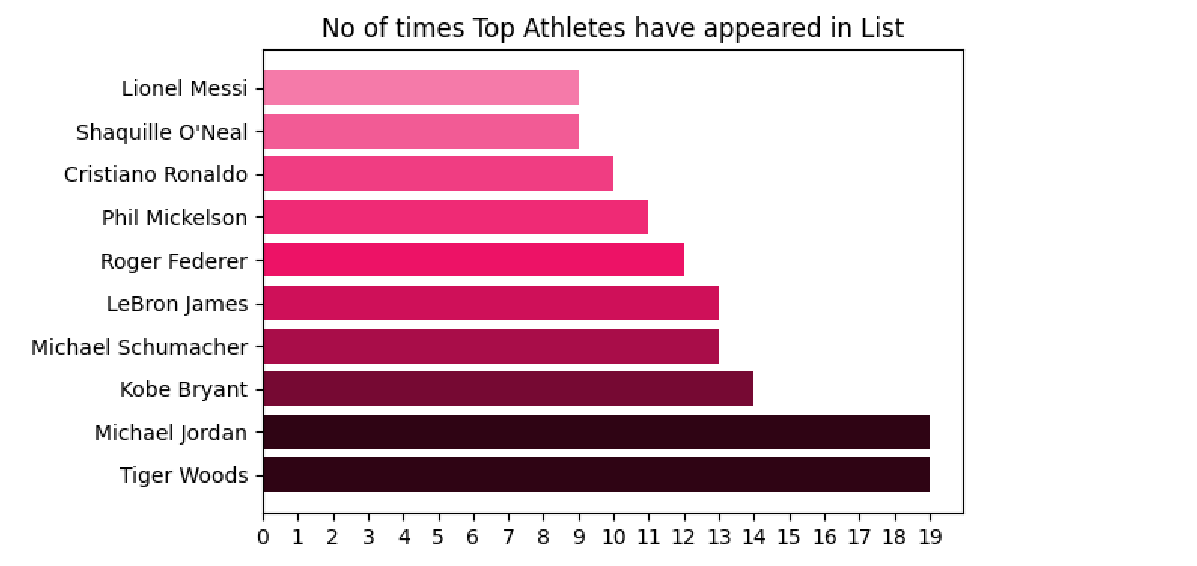
It's interesting to see the significant dips and rises in the periods spanning 2015-2020. The drop in 2020 may be explained by the COVID outbreak even though they may not directly coincide, but the other fluctuations may need further explanation but are probably similarly economically linked.



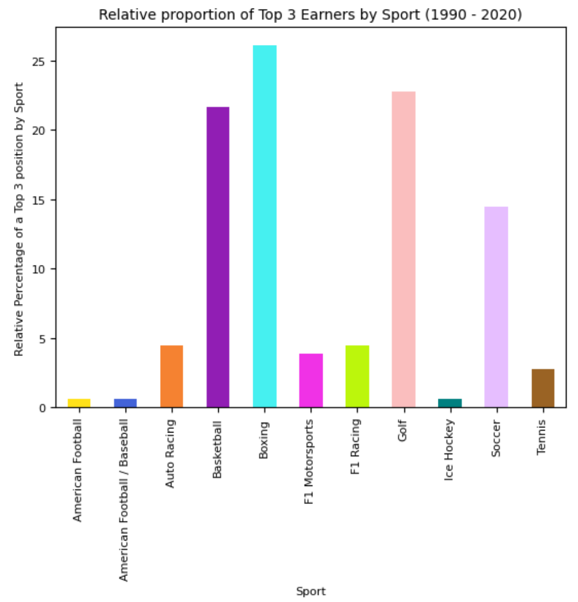
The above graphic shows the distributions of earnings between Nations over the 30-year period. It should follow the line graph above but with the added feature of seeing the nations with a higher proportion of the revenue. Out of the 20 nations listed, I have put the lowest 8 into the category of 'other'. Obviously, the USA is accounting for much of the revenue over the entire period but it is interesting to note how Brazil has appeared significantly in the last 3 years. Perhaps this has some political connections. Argentina has also fared well during the last 7 or eight years of the data span.



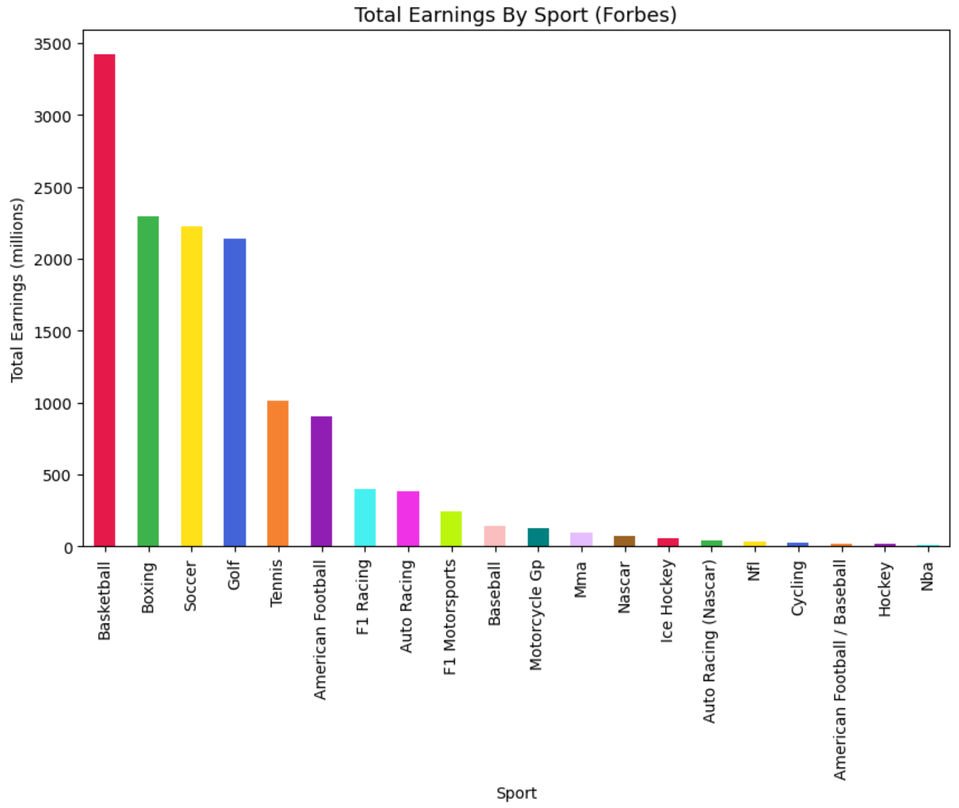
The pie chart above shows how heavily the data is weighted by USA nationality.



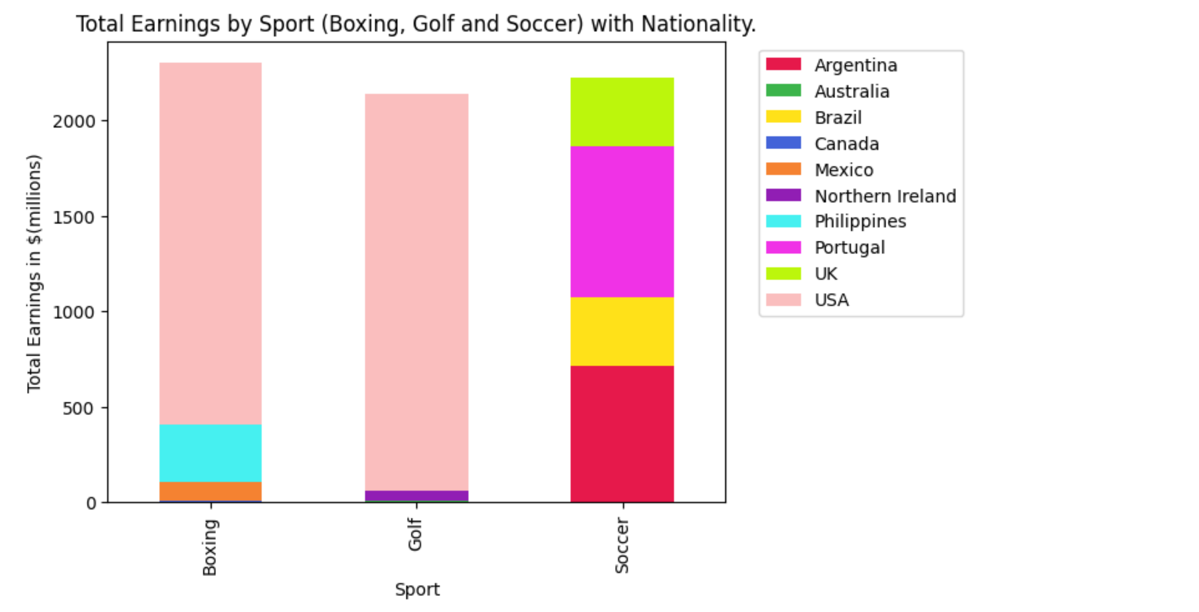
From the above Bar chart, we can see the athletes with the largest number of appearances in the list over the 30 years. Michael Jordan (basketball) and Tiger Woods (golf) have both appeared in the list a total of 19 different years (excluding 2001 of course which is missing from the data). One question raised from this is whether they were actually competing in all of these years or whether it just represents a massive income in that year by still being associated with the sport, as in Tiger Woods' case, he did have a while not playing.



This bar chart shows a measure of the top 3 athletes' (in wealth) nationalities over the 30-year period. If the athlete's ranking was 1 then they are given a score of 3, 2 if second and 1 if third. These are then totalled by sport and proportionality as a percentage. It doesn't necessarily give a proportion of total earnings but a proportion of how often an athlete representing that sport appears in the top 3 earners. So you will see some similarities between this graph and the graph below, but with differences. For example, here boxing appears much higher than basketball, implying that although there may be more money in basketball, boxing is still clearly well paid and there are likely to be more very wealthy ex-boxes than ex-basketball players.



Perhaps this is one of the most informative graphs in terms of the money in different sports as it is giving the total earnings over the 30-year period. The graph has a distinctly descending slope. Basketball is predominantly a U.S. sport and they are by far the most common nationality in the list so this is not a big surprise from the theme of the data. It is interesting to note that total earnings from the three sports of Boxing, Soccer and Golf are all considerably high and are more widely played internationally. The final graph on the next page shows a breakdown of these by nationality.



Contrary to what I said above, this graph shows the prominent earners in Boxing and Golf are of US nationality. However, what this graph does show is that soccer is the most rewarding sport that has minimal US participation. Portugal and Argentina feature strongly in this category probably due to Ronlado(Portugal) and Messi(Argentinian) both appearing in the list on many occasions. U.K and Brazil also feature significantly here and this can be explained by the fact many top Brazilian players play in the European leagues where wages are high, or in the UK which boasts one of the most entertaining and wealthiest leagues in the world.

THIS REPORT WAS WRITTEN BY: Paul Aughterson