

**TASK**

**Exploratory Data Analysis on the \*\*\*\*\*\*\*\*\*\*\*\* Dataset**

[](http://www.hyperiondev.com/portal/)

***Forbes Richest Athletes 1990 - 2020***

This Csv data, contains such columns:

1. "S.NO" - a unique serial number which is assigned to each athlete.
2. "Name" - which represents athletes themselves.
3. "Nationality" - a nationality of each athlete.
4. "Current Rank" - a rank assigned to each athlete in the year specified by the "Year" column.
5. "Previous Year Rank" - a rank assigned to each athlete in the year previous to the "Year" column.
6. "Sport" - a type of sport each athlete practises.
7. "Year" - the definition of the current year.
8. "earnings ($ million)" - the amount of money quantified in millions earned by each athlete.

Each row in this data represents the column variables named above for each specific athlete name.

**DATA CLEANING**

* Make sure no information in the data is duplicant.

(function - *drop\_duplicates()*)

* Remove any "string" information about the missing data within the integer containing columns (e.g. "?"; "non"; "not stated") and replace it with an empty space.

(function - replace())

**MISSING DATA**

* Use a further function to replace any empty spaces with the data with 0 values. (function - *fillna()*)
* Check all the integers containing columns for values within them.

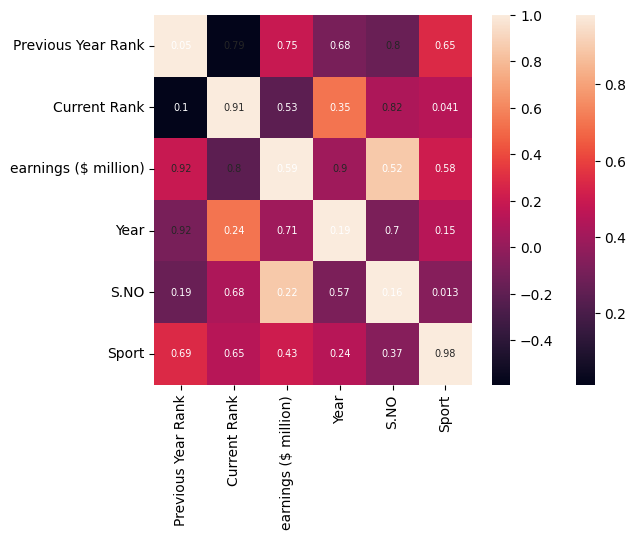
(function - *unique()*)

* Convert integer information into an integer format for further visualisation use (function - int64)

**DATA STORIES AND VISUALISATIONS**

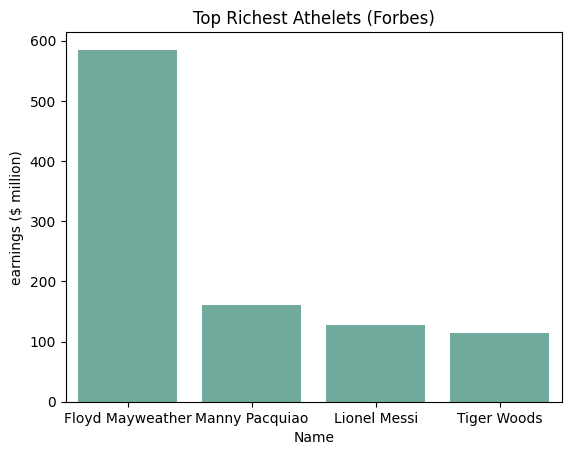
Represent data patterns within the Forbes provided data frame, for easy visual representation and interpretation:

*Visualisation 1*

****

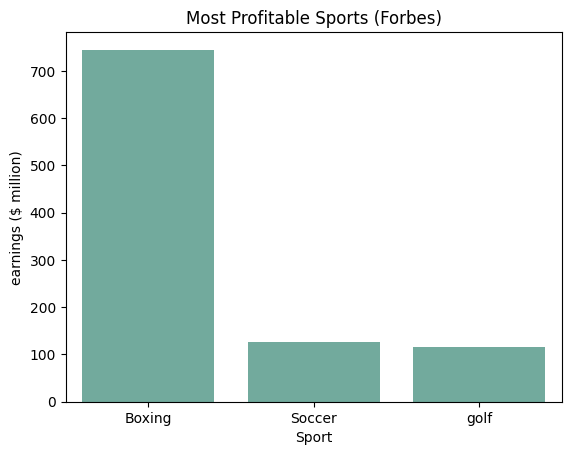
This Forbes data obtained heatmap, shows the relationship patterns of variables that could be of interest in terms of influence on the earnings column. Here the colours presented on the right handside of the map show the correlation degree, the closer the correlation colour is to 1 the stronger the relationship between the values and the closer the represented colour is to zero the weaker the relationship is. It can be noticed that in the middle across the diagonal of the map there is a line representing the strongest correlations, due to the same variable overlapping, such representation can be used to better understand what a strong data relationship should look like. Starting with such a visual data representation, helps understand the relationship patterns that have the potential to be further investigated (e.g. earning ($ million) and Sports).

*Visualisation 2:*

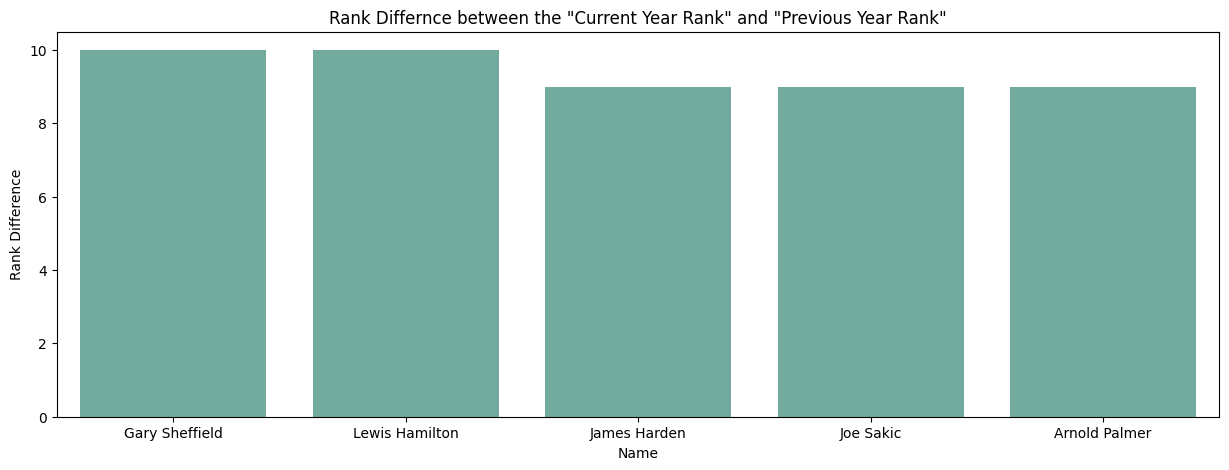


This barchart represents the top five of the richest athletes according to their earning amount in the dollar currency, which is expressed to the nearest million. It can be noticed that the order within this bar chart shows the richest athletes in the descending order, with the richest athlete being Floyd Mayweather.

*Visualisation 3:*

****

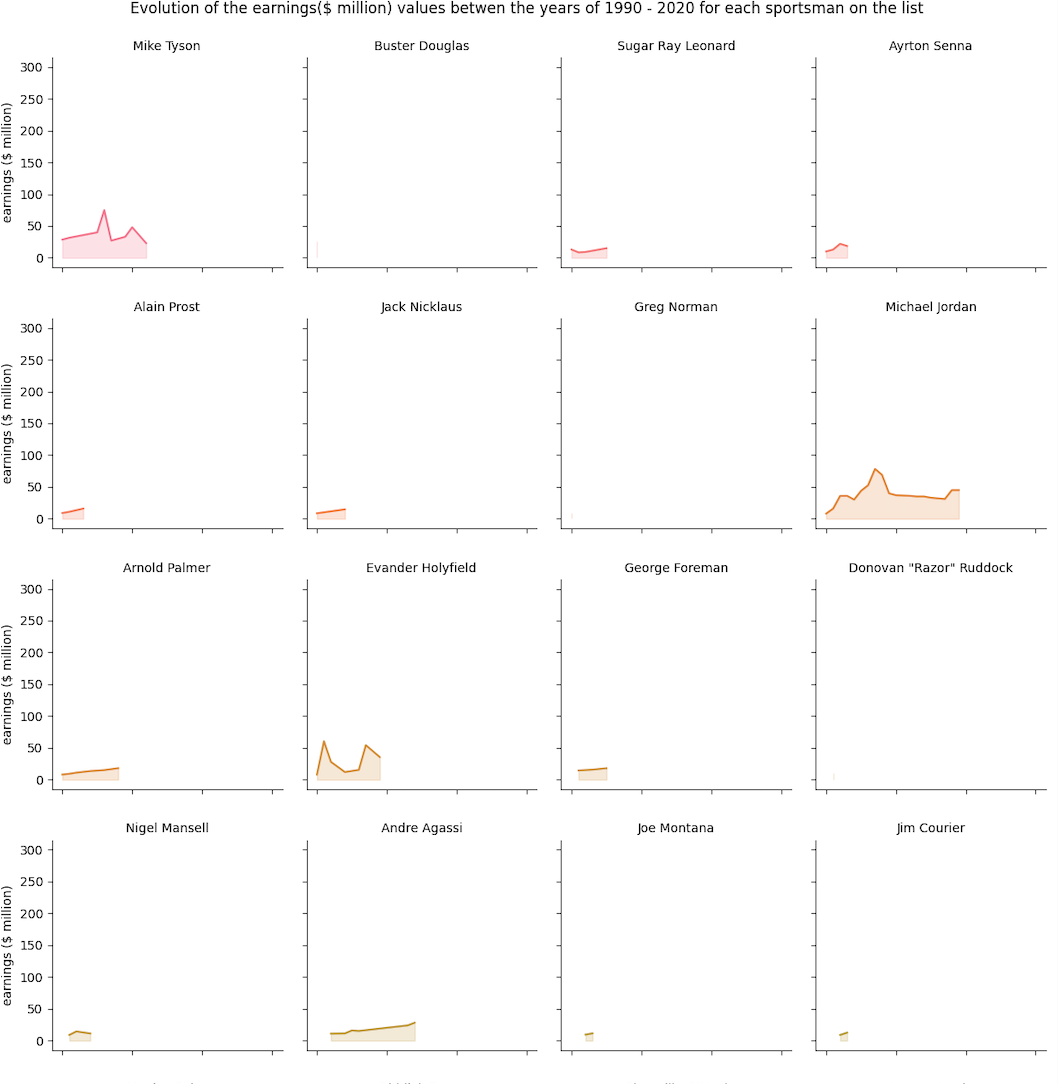
This barchart represents the top five of the most profitable sport types represented within the Forbes provided data. The order is sorted according to the earnings ($ million) of the individuals within the list, calculating the sport type which brings the highest earnings to the contributing sportsman. As it can be seen boxing is the highest paid sports type, with a large gap in the earnings with the next highest paid sports type - soccer. The difference between the second highest paid sport (soccer) and the third highest paid sport (golf) is way less significant, then the first highest paid sport (boxing) and the second (soccer).

*Visualisation 4:*

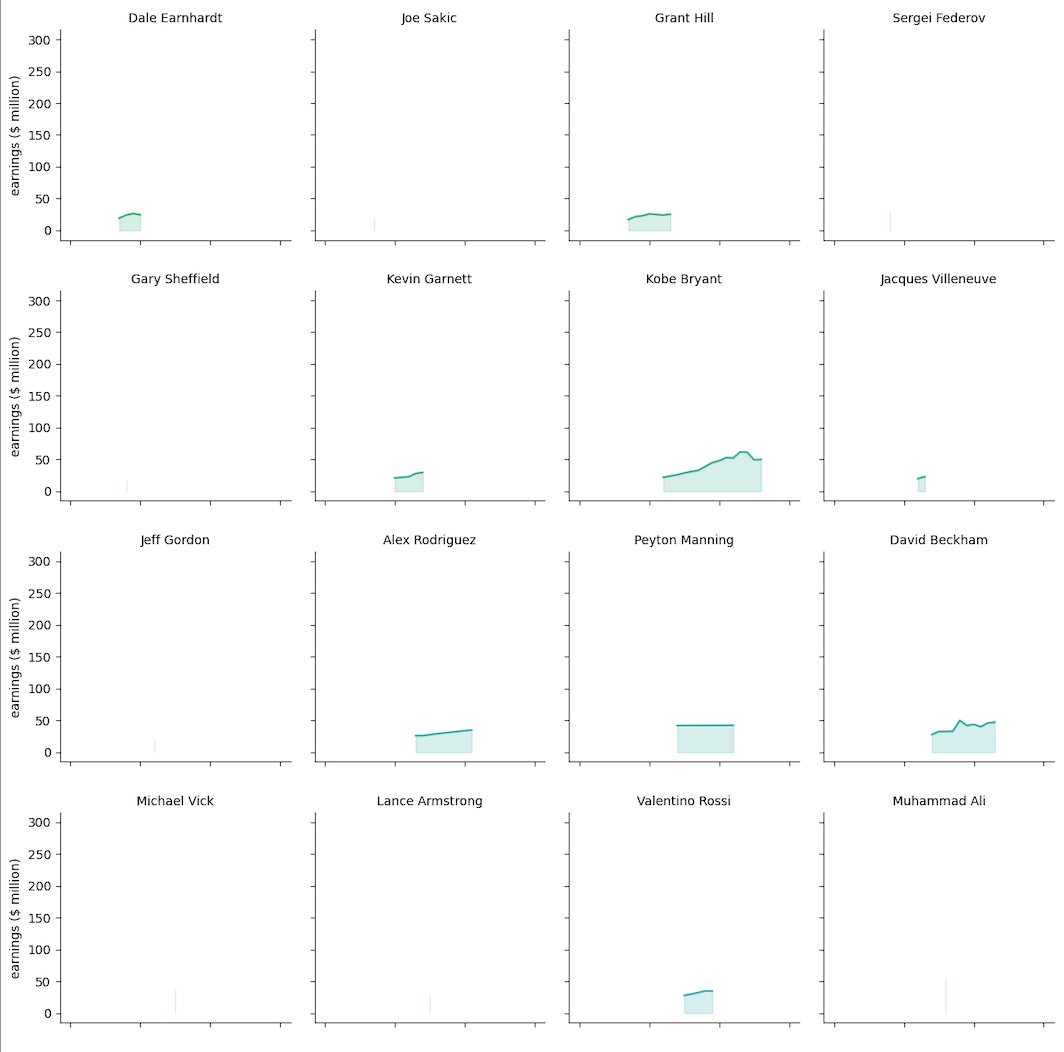
This barchart represents rank patterns within the data, as two rank columns are present within the Forbes data frame ( “Current Rank” and “Previous Year Rank”). This barchart presents the difference between the two ranks(“Current Rank” - “Previous Year Rank”), starting with the name of the sportsman who has the biggest difference between the rank columns, showing rank migration patterns between the years. The smaller the difference the closer the two ranks were to each other through those years, and vice versa. Such visualisation could be beneficial for the representation of the rank stability of a specific sportsman, highlighting the possible reliability of investment.

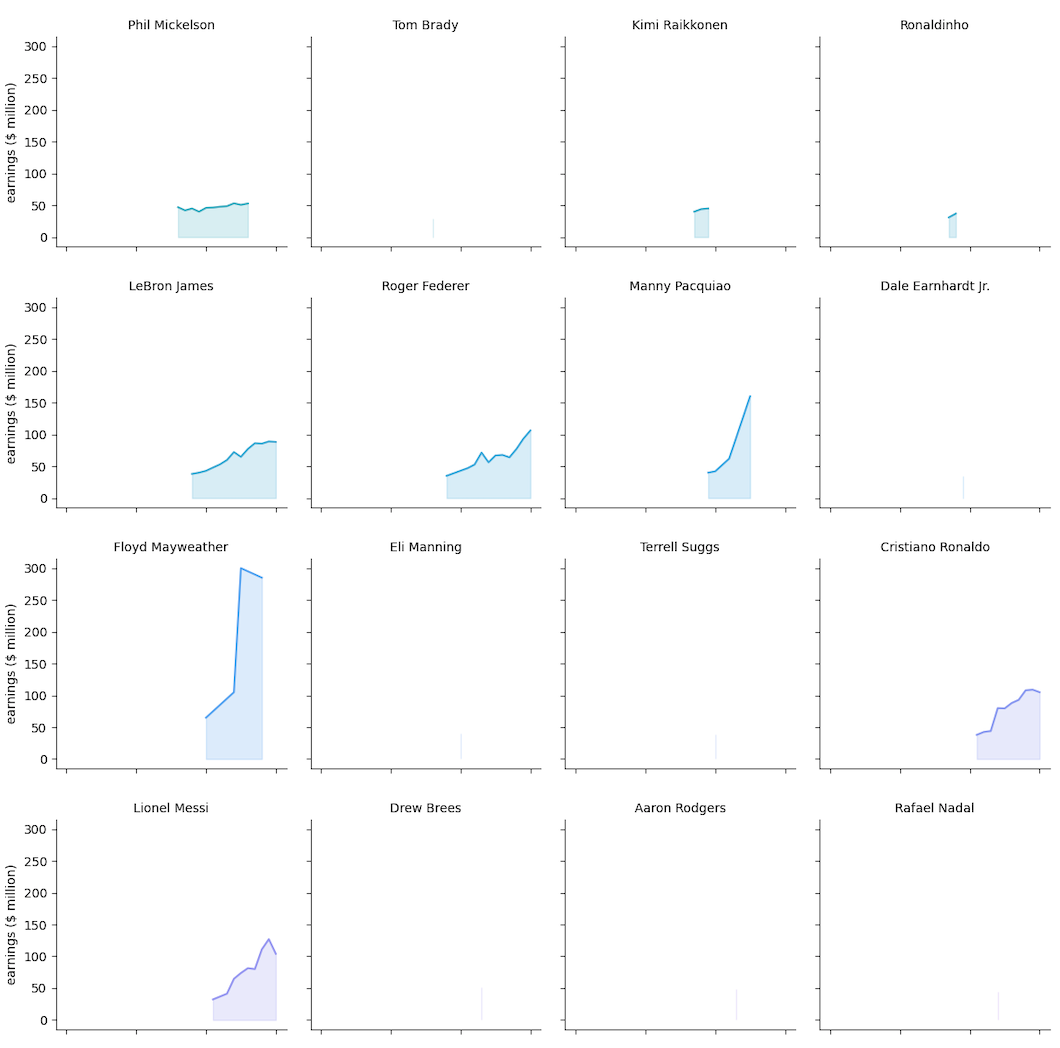
*Visualisation 5:*

The following set of time data, represents how the earnings ($ million) of each sportsman on the Forbes provided data fluctuated over the years. As it can be seen, due to the scale choice of the earnings ($ million) values, the fluctuations are seen as not that grand (e.g. Joe Montana; Jim Courier; Wayne Gretzky). As well as this, for the individuals with the earnings that have maintained stable over the years, a fade thin line represents such a pattern (e.g. Emmit Smith; Monica Seles; Sergei Federov). However, looking past those representation limitations an overall over-time trends can be clearly seen and individuals with the growth patterns in earning can be spotted ( e.g. Michael Schumacher; Tiger Woods; Kobe Bryant).



****



****

**THIS REPORT WAS WRITTEN BY : Vasilisa Usanova**

