

**TASK**

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

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

**Introduction**

The Forbes Richest Athletes (1990-2020) dataset is a multivariate dataset from Forbes. The dataset comprises of 7 features. The features are comprised of categorical and numerical types.

The dataset consists of 301 records of the top 10 earning athletes for each year. Each row has items for each column/feature, with unique datatypes of that specific feature.

**DATA CLEANING**

The data cleaning process comprised of identifying columns that are redundant or unnecessary namely: Previous Year Rank.

Furthermore, the data cleaning involved the identification and removal of the duplicated athlete entries (rows) in the dataset and assigning the correct datatype for each column/feature in the dataset.

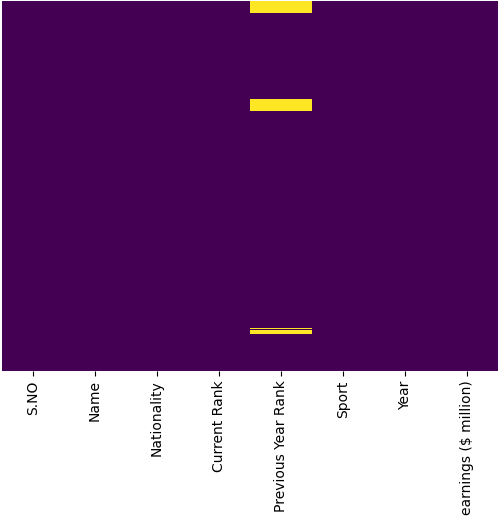
The column **Sport** had to be further formatted, to ensure that it consisted of unique sporting codes.

* **NBA** had to be replaced with **Basketball**.
* **NFL** had to be replaced with **American Football**.
* **Ice Hockey** replaced with **Hockey.**
* **F1 Motorsports**, **F1 Racing**, **Auto Racing** (**NASCAR**), **NASCAR** had to be replaced with **Auto Racing**.

**MISSING DATA**

The only missing values in the dataset, were contained in the **Previous Year Rank** feature, which was removed during data cleaning.

The heatmap visualisation below demonstrates and shows that all the missing data were in the **Previous Year Rank**.

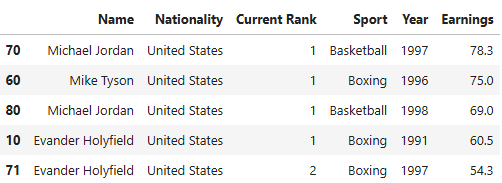


**DATA STORIES AND VISUALISATIONS**

The first analysis has been to establish the top earners during the periods, that are accompanied by the respective tables:

**Table: top 5 earners from 1990 – 2020**

  
  
**Table: top 5 earners from 1990 – 1999**

****

**Table: top 5 earners from 2000 – 2009**

A screenshot of a computer screen

Description automatically generated

**Table: top 5 earners from 2010 – 2020**

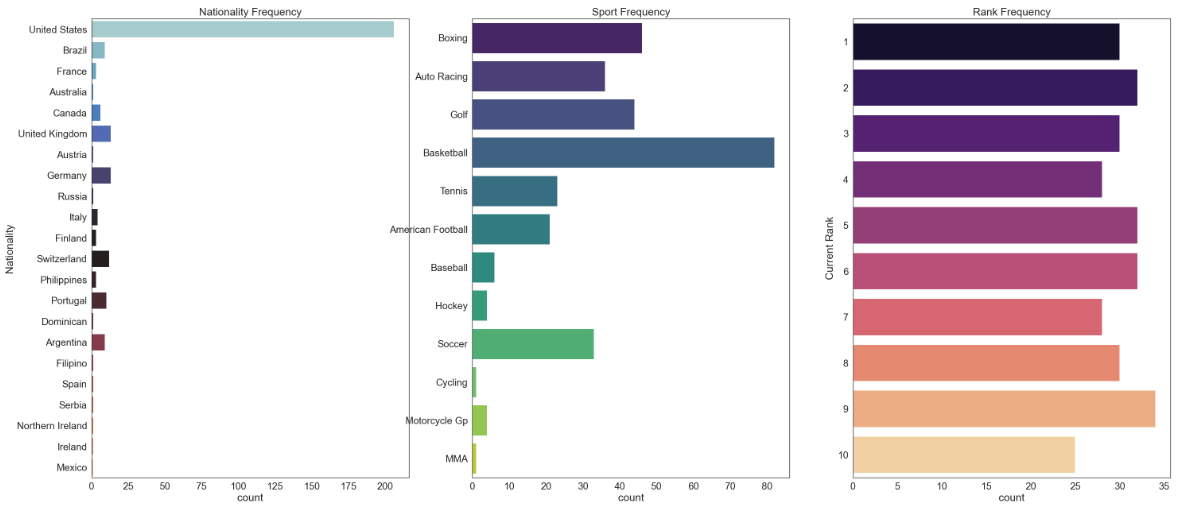
****

The most notable high earning athlete throughout the periods is American boxing legend Floyd Mayweather. Moreover, most of the periods are dominated by athletes from the United States.

Another interesting observation is that the sport of boxing features prominently in al the periods except during the early 2000’s, where American golfer Tyger woods single handedly dominated the early 2000’s period (2000 - 2009).

The highest top 5 earning athletes were either ranked first or second in the Forbes list for each period.

The horizontal bar chart below provides a visualisation of the frequency of Nationality, Rank and Sport code:

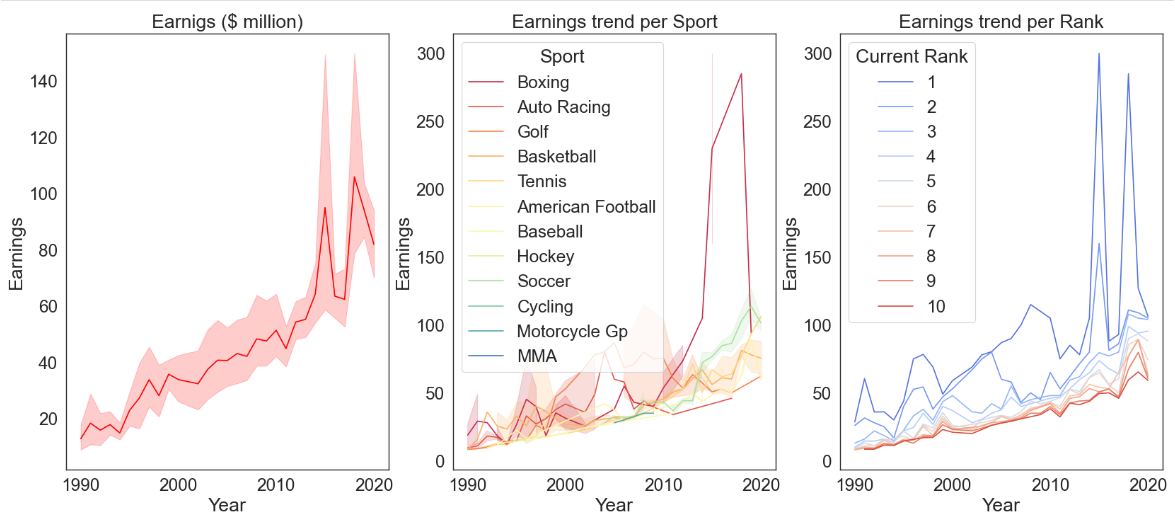


The United States dominates the list of countries in the list, meaning that an overwhelming majority of athletes are from the United States.

Basketball, Boxing, Golf, Auto Racing and Soccer lead the representation of sporting codes with the highest earners in the Forbes Richest Athletes list.

The Current Ranks are evenly presented in the list of rich athletes.

The below chart depicts the trend of the earnings over the period (1990 -2020), under various scenarios.



The earnings of the athletes have been on an upward trend over the period (1990 - 2020) in question. The earnings trend per Sport/Rank points to a possible outlier in the dataset during the last decade (2010 - 2020).

This can be attributed to boxer Floyd Mayweathers high earnings. However, it appears that the trend might be reverting to the mean post his retirement.

The following visualisations (box-plots) are meant to demonstrate the distribution of the earnings based on the three categorical features **Nationality**, **Current Rank** and **Sport**.

A graph showing different colored boxes

Description automatically generated with medium confidence

A diagram of different colored squares

Description automatically generated

A chart of different colored squares

Description automatically generated

The boxplots demonstrate the outliers in the earnings feature for Nationality/Current Rank/Sport.

The boxplots show there is enough variability for the Sport, Current rank and Nationality to play a part in a predictive model.

A method ought to be developed for the treatment of outliers, as demonstrated by the histogram (skewed bell shape due to outliers) below. This might be through some feature scaling be it either normalisation or standardisation.

A graph with a line going up

Description automatically generated

# ENSURE THIS DOCUMENT IS NEAT AND ADD IT TO YOUR PORTFOLIO

**THIS REPORT WAS WRITTEN BY: ZUKISA XOTYENI**

