

**TASK 36**

**Exploratory Data Analysis on the Forbes Highest Paid Athletes 1990 - 2020 Data Set**

[](https://www.hyperiondev.com/)

**Introduction**

Forbes Highest paid Athletes 1990-2020 data set containing 8 columns and 300 rows of information. I imported varies libraries to assist my analysis and loaded the csv file into Jupyter notebook.

**DATA CLEANING**

In summary I carried out the following:

* I identified special characters were present, I replaced these with a blank and the figures before the special characters would now show up in the coming analysis
* I identified Previous Year Rank column was object data type so I changed this to integer in the data frame
* I removed the S.No column as this was not required
* I changed the Sport column which contained a mixture of lower/upper casing to string upper to assist with analysis.

**MISSING DATA**

In summary I carried out the following:

* I identified null values and replaced these with a 0 value

**DATA STORIES AND VISUALISATIONS**

In summary I carried out the following:

Nationality

* I looked at the number of athletes by country in this data frame

Chart, histogram

Description automatically generated

* I looked at the number of athletes country by earnings in this data frame

Chart, histogram, waterfall chart

Description automatically generated

* I looked at the overall spread of earnings across the whole data frame

Chart, histogram

Description automatically generated

Earnings

* I analysed the highest earnings by athlete and produced a matplotlib visualisation

Chart, bar chart

Description automatically generated

I analysed the lowest earnings by athlete and produced a matplotlib visualisation

Chart

Description automatically generated

* I analysed the highest earnings athlete and created a line plot to visualise.

Chart, line chart

Description automatically generated

Sport Type

* I analysed all the types of sports to find the highest earnings sport which was Basketball and visualised using a matplotlib visualisation

Chart, histogram

Description automatically generated

* I analysed all the types of sports to find the highest number of athlete in which sport and again this was Basketball. I produced a visualisation using a matplotlib plot.

Chart, histogram

Description automatically generated

Year

* I analysed the 3 decades and found the highest year per decade to visualise using matplotlib.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

* I also created a seaborn relplot to visualise this further and compare all 30 years and earnings over this timescale.

Chart, scatter chart

Description automatically generated

My findings were as follows:

1. The number of athletes by country was the highest from the USA.
2. The total earnings of athletes by country was highest from the USA.
3. The spread of earnings of athletes by count were clustered around the 10 – 50 million.
4. The top 10 highest earning athletes are also shown.
5. The lowest 10 earning athletes are also shown.
6. I analysed the top earning athlete, tiger woods in more detail.
7. For sport type basketball is the highest earning.
8. I visualised each decade with the highest athletes earnings and isolated the highest earning year.
9. All athlete earnings have increased over time.

Finally I will be presenting my findings on 2nd March 2023 at 7.30am to mentor Kenneth Mlimi

**THIS REPORT WAS WRITTEN BY: LOUISE RANDALL**

