**EXPLORATORY DATA ANALYSIS**

**ON**

**Forbes Richest Athletes 1990-2020**

**[04/02/2023]**

**CAPSTONE PROJECT IV**

**INTRODUCTION**

This Exploratory Data Analysis (EDA) is based on a Forbes dataset entitled; ‘Richest Athletes (Forbes Richest Athletes 1990-2020)’.

It appears that Forbes is well regarded website/publisher as can be seen from this website: <https://en.wikipedia.org/wiki/Forbes>.

The subject matter of the dataset is the estimated wealth of athletes.

The dataset has the following column headings:-

* Name
* Nationality
* Current Rank
* Previous Year Rank
* Sport
* Year
* earnings ($ million)

The dataset consists of 301 entries.

The dataset consists of these 3 datatypes:-

* object
* int64
* float64

**DATA CLEANING**

To identify data values requiring cleaning these functions were applied to the dataset:-

* .unique()
* .isnull().sum()
* .head()
* .info()
* list()

**MISSING DATA**

An investigation uncovered the existing of blank entries in the dataset. Because of the missing values the csv file was reread into a dataframe and the blank values with NaN values, using the na\_values=’’ function.

The function .isnull().sum() was applied to the dataset to identify missing values. The result showed that the only column with NaN values was Previous Year Rank with 24 values missing.

Using the .unique() function on the dataset unearthed many values which were confusingly similar.

Using the .replace() function the similar values in the Sport column were streamlined into a single consistent value.

The remaining confusing values in the Sport column were further investigated to identify which column value should be applied to them. These confusing values were identified with the .loc() function to see which sportspeople were associated with these values.

The values which I found confusing were:-

* 'American Football / Baseball' 🡪 this was unusual, it was the only combination value in this Sport column.
* ‘Cycling’ 🡪 this looked similar to the ‘Motorcycling’ values.
* ‘Hockey’ 🡪 this looked similar to the ‘Ice Hockey’ values.

Once the sportsperson related to the confusing values were identified using .loc() then a web search clarified which value was correct. For example, ‘Cycling’ related to the sportsperson Lance Armstrong and the value was changed to ‘Pedal Cycling’ to make its meaning clearer.

The Name column was checked with the .unique() function and no similar/confusing values were identified.

**DATA STORIES AND VISUALISATIONS**

This graph shows that the USA dominates the wealth rankings of the wealthiest sportspeople as identified in this dataset.

The USA by far and away exceeds the closest competitor which is the UK. It would be interesting to investigate the reason behind the concentration of wealth in sports with the USA.

Chart, histogram

Description automatically generated

Below is a graph – a boxplot – which shows some extreme values when looking at the earnings of individual sportspeople.

There is an extreme point in the graph which relates to a sportsperson called Floyd Mayweather.

I have not heard of Floyd Mayweather and I was suspicious if this value was an error in the dataset because it seemed to be an extreme value.

Chart, histogram

Description automatically generated

When looking at the sports where wealth is most aggregated it can be seen that Basketball is by far and away the relatively richest sport in this dataset.

I have common knowledge that one of the richest sport-stars in the 30-years of this dataset was Michael Jordan, a famous basketball player with endorsements from Nike and others.

The degree of the distance between Basketball and other sports, such as Formula 1 Motorsports – an apparently wealthy sport – surprised me. I suspect that Basketball has more participants and contributors. It would be interesting to know how the wealth is accumulated in Basketball.

Chart, histogram

Description automatically generated

I was concerned that a previous graph identified an unknown sportsperson called Floyd Mayweather as the wealthiest individual in the dataset.

The scatterplot below shows Floyd Mayweather again as the wealthiest in the dataset. But, each dot shows an individual entry in the dataset. This graph shows that Floyd Mayweather has only a few entries in the dataset and the values of these dots is extreme.

Two of the Floyd Mayweather blue dots extend above the legend and are outlier values compared to all other sportspersons.

A picture containing text, receipt

Description automatically generated

Looking at the box and whisker plot below which shows how much money has flowed into the individual wealth of sportspeople over time shows an increasing trend of top sportspeople becoming richer.

In 2016 there is an outlier value because the diamond extends far above the box.

It would be interesting to see if overall wealth has increased in the Sports sector or if sportspersons have taken a larger proportion of Sports wealth in earnings.

I would be interested to know if the 2016 outlier value was related to Floyd Mayweather’s income which has fewer but larger payouts.

Chart

Description automatically generated

Looking at a pie chart of the proportion of top professional sports people’s wealth by Nationality it can be seen that the USA has the majority of the financial pie.

This is surprising because even though the USA has a very succesful, commercial sports market, USA sports might be considered most attractive to Americans. For example, Basketball, Amercian Football and Baseball are not considered popular international sports viewing like Soccer/Football.

Part of the sports wealth in the USA might be connected to its popularity in international fashion such as baseball caps. It would be insterested to know the source of Sports income.

Chart, pie chart

Description automatically generated

By comparing the wealth of individual sportspersons with Time it can be seen that there is another extreme outlier value for Floyd Mayweather which indicates that he has acquired his wealth in a shorter period of time.

The plotting for Floyd Mayweather far exceeds the nearest surrounding sportspersons.

Chart, histogram

Description automatically generated

Looking at a Word Cloud graph of the dataset shows that Tiger Woods and Michael Jordan have more frequent entries in the dataset which shows there staying power, consistency as well as wealth.

Text

Description automatically generated

This report was written by: Prabhjit Gill [04/02/2023]