# Video Game Sales Analysis

<https://datascience.fm/video-game-sales-analysis/>

Analysing the 'Video Game Sales Analysis' dataset with the help of pandas library and learn the fundamentals of plotly by visualizing the result using simple yet interactive graphs.

Welcome back to another pandas posts with us. We will be analysing the video games sales analysis using pandas and visualizing the results using plotly.

Who doesn't like video games? Video games have become a favourite pastime for most of us. Video games are a billion-dollar business and have been for many years. Now imagine being able to analyse the sales trends for the popular Nintendo games and many more such platforms. How cool would it be!

In this module, we will be laying the foundation for our analysis by processing and exploring a large amount of data on video game sales. The dataset contains information regarding the sales of video games across various regions like North America, Europe, Japan and also globally, while also giving information regarding the Names, Publishers and Platforms. This dataset has been made available thanks to Kaggle which is the home for many such datasets and competitions.

You are a data scientist in the gaming industry and want to analyse the sales pattern for different features. You want to investigate different features and approach the game producer with a thorough report. You want to find the answers to the following questions:

1. Which region has performed the best in terms of sales?
2. The top gaming consoles are Microsoft (Xbox), Sony (PlayStation) and Nintendo, with Google acting as a new competitor. Does the dataset also back this information? Analyse with respect to different regions.
3. What are the top 10 games currently making the most sales globally?
4. What are the top games for different regions?
5. Are there any games with release year older than 2000 that are still making high sales? What are they?
6. What are the top gaming genres that are making high sales?
7. Does the publisher have any impact on the regional sales?
8. Is there any region that has out-performed global average sales?

We will be exploring the questions above using data, analytics and visualization techniques one at a time.

So let's get started!

## Learning Objectives

After this module, we should be able to do the following:

1. Use pandas library to analyse different features of the dataset, which includes,

* Read the dataset.
* Use functions like df.query and df.groupby.
* Analyse the sales for different regions with respect to various features.

1. Use plotly library to visualize the given results.
   * Plot graphs like bar graphs and pie charts.

## Understanding the dataset

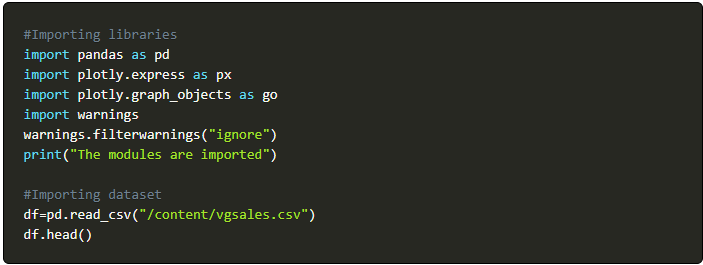
This dataset contains a list of video games with sales greater than 100,000 copies. Each row in the dataset represents a game, whereas, each column contains different attributes.

The data set includes information about:

* Rank - The overall rank of the game.
* Name - The name of the game.
* Platform- The platform on which the game was released.
* Year- The year in which the game was released.
* Genre-The genre of the game.
* Publisher-The publisher of the game.
* NA\_Sales, EU\_Sales, JP\_Sales , Other\_Sales, Global\_Sales - Sales made by a particular game in North America, Europe, Japan, Other regions and Globally (in millions).

## Pre-processing the dataset

To get started, we need to import some useful libraries that will help us import the dataset into our python environment, manipulate and analyse the same and later help us to visualize it.





Using df.shape function, we see that our dataset has 16598 rows and 11 columns.

Now, we will look for missing values using the df.isnull().sum() function. We observe that the column Year has 271 missing values and the column Publisher has 58 missing values. We will treat the missing values as follows:

* For Year, we use the code

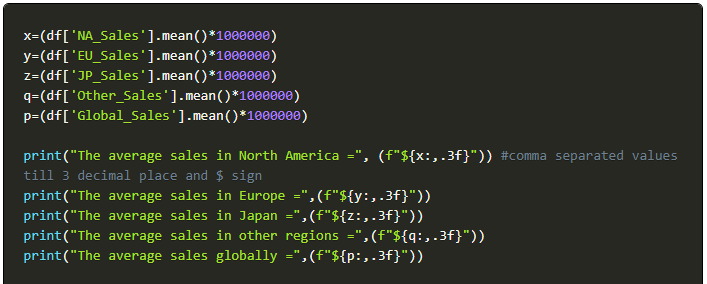
df["Year"].fillna(df["Year"].quantile(0.50),inplace = True)

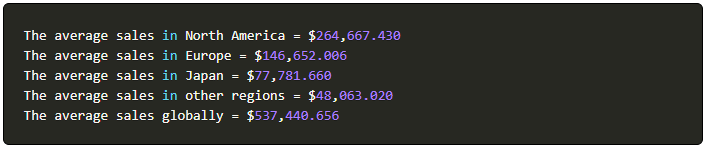
* ForPublisher, we use the codedf["Publisher"].dropna(inplace = True)

Now that we are done, it's time to jump to the questions.

## 1. Which region has performed the best in terms of sales?

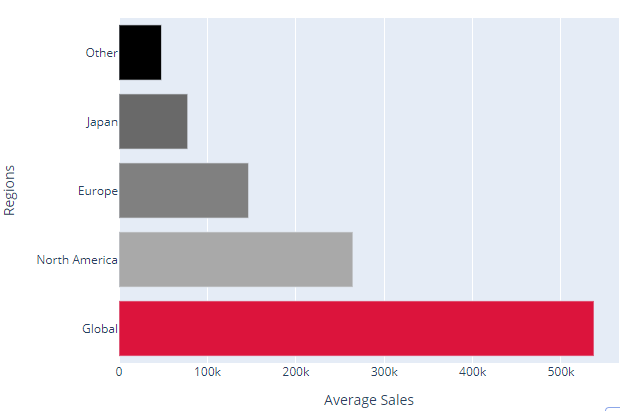
We will utilize the average sales made per region and compare the results. Before we do that, let's make sure we know how to calculate the average. We will be using a very simple line of code for this, i.e., df['Region'].mean(), where 'region' depicts the different regions in the dataset and mean() function is used to calculate the mean. We observe that our output is coming in decimals, to convert the values in millions, let's multiple the result with 10,00,000. The final code should look like this.





Let us now plot our findings on a bar chart.





We have successfully used a bar graph to demonstrate the results. It is clear now which region has made the most sales (expect global sales) in terms of video game. We observe that **North America** has the highest average sales of **$264,667.430**. It's clear now that we can use the bar graph to answer questions like, which country sells the most video games in the world? This will be help us make some strategy changes.

The above result also answers our 8th questions, i.e., **Is there any region that has out-performed global average sales?**

The answer is '**No'**. There is no region that has out-performed the global average sales. The average global sales is **$537,440.656 million.**

## 2. The top gaming consoles are Microsoft (Xbox), Sony (Playstation) and Nintendo, with Google acting as a new competitor.

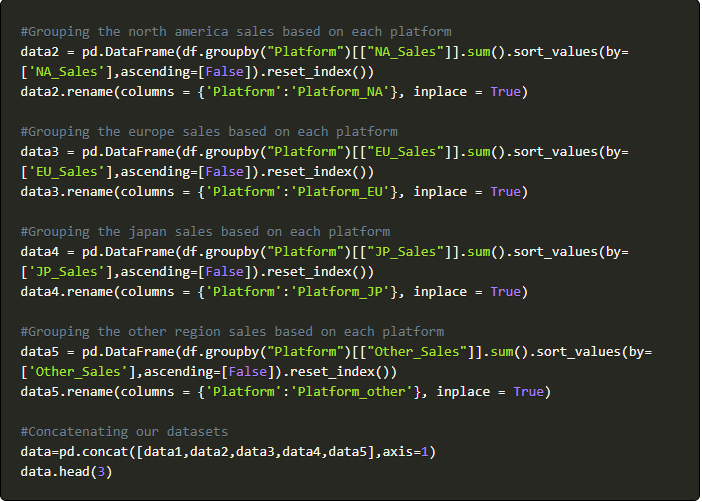
### Does the dataset also back this information? Analyze with respect to different regions and also, globally.

The world is getting more connected every day, and more gaming trends and preferences are emerging around the globe. We want to know which console is preferred the most based on different regions. We will compile results from all over the world and determine if any regional trends might shape the gaming industry as a whole.

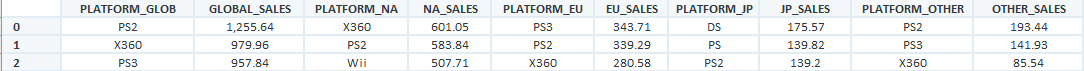
We will use the df.groupby function of pandas for our analysis.

* ***groupby(): This function is used to group or combine large amounts of data and compute operations on these groups.***

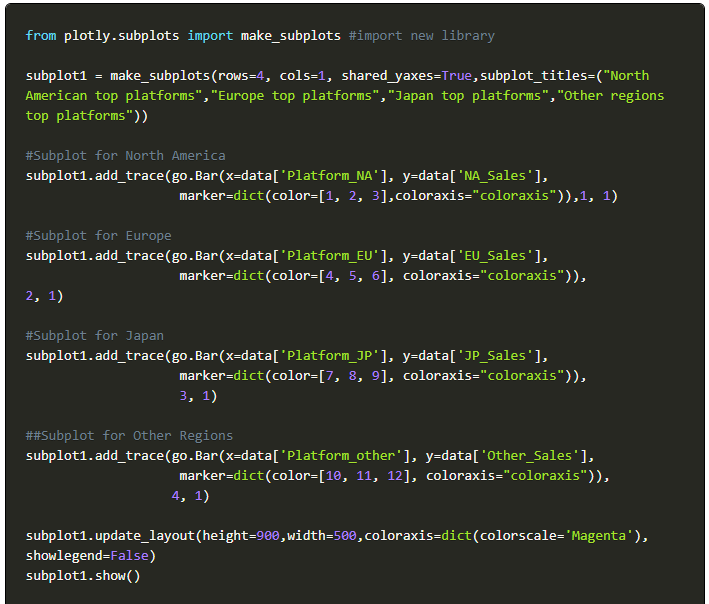
Have a look at the code given below.

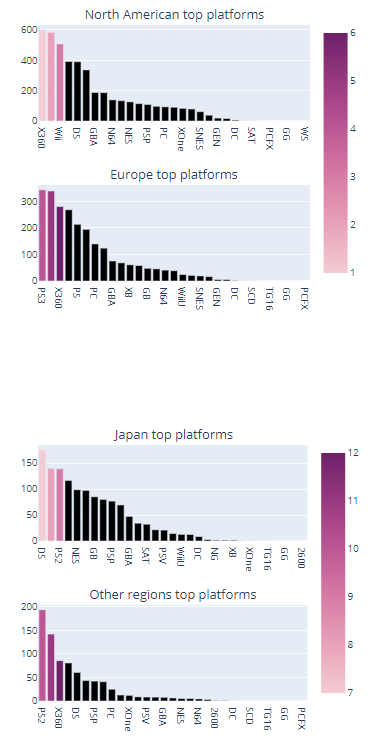


The dataframe data we created will provide us with the total sales (global, NA, EU, JP, Others) based on different platforms. The dataframe will look like this:



We realize that the dataframe alone will make it very hard for us to analyze the result. Therefore, we plot the result using a line chart.





The graph shows us the top platforms preferred by users in different regions.

We observe the following:

* **X360** (Microsoft) is the top preferred console by users in North America making a total of **$601.05 million**.
* **PS3** (Sony) is the top preferred console by users in Europe making a total of **$343.71 million**.
* **DS** (Nintendo) is the top preferred console by users in Japan making a total of **$175.57 million**.
* **PS2** (Sony) is the top preferred console by users in other regions making a total of **$193.44 million**.

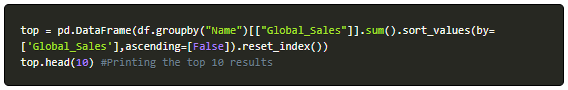
This implies that our assumption about Sony, Nintendo and Microsoft among the top consoles was correct.

We can use the same method to compare the top games and publishers for different regions.

Moving on to the next question.

## 3. What are the top 10 games currently making the most sales globally?

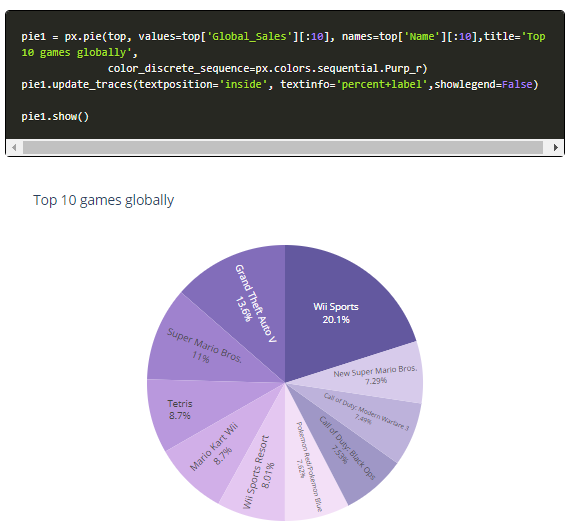
We will use a similar approach by grouping the games with respect to the global sales and observe the top 10 games.





We see the most played game is **Wii Sports** making a total of **$82.74 million** globally.

We will plot the above using a pie chart.

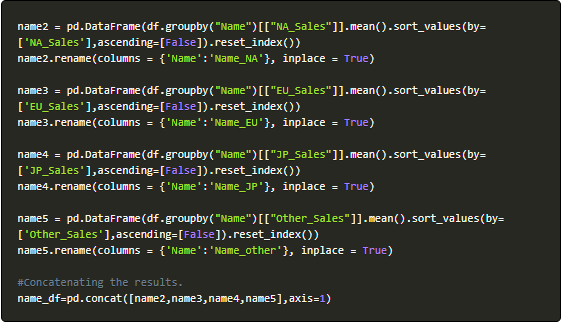


The pie chart also shows the proportion of sales each game holds, while also depicting the results.

## 4. What are the top games for different regions?

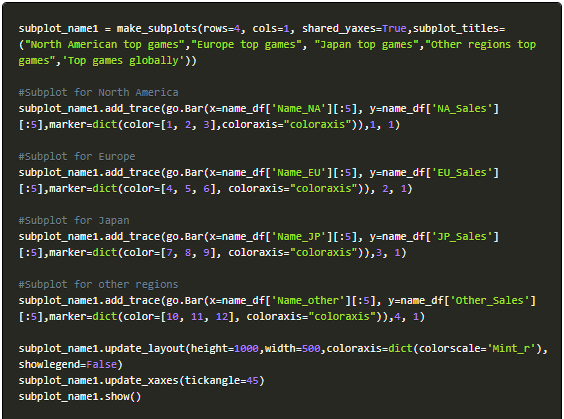
We will have to compare the sales made by different games regionally. We will use the same approach we did while analysing the platform.

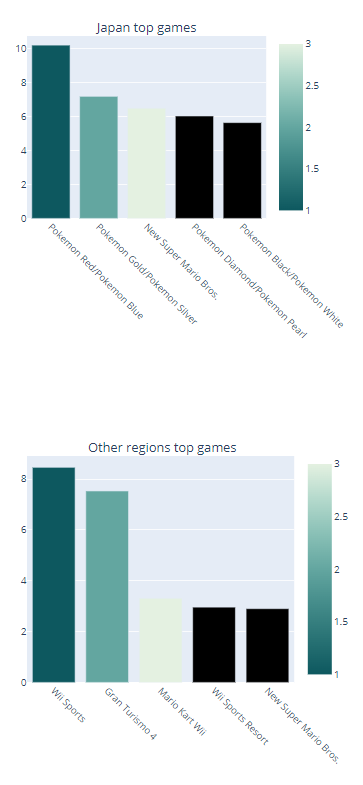
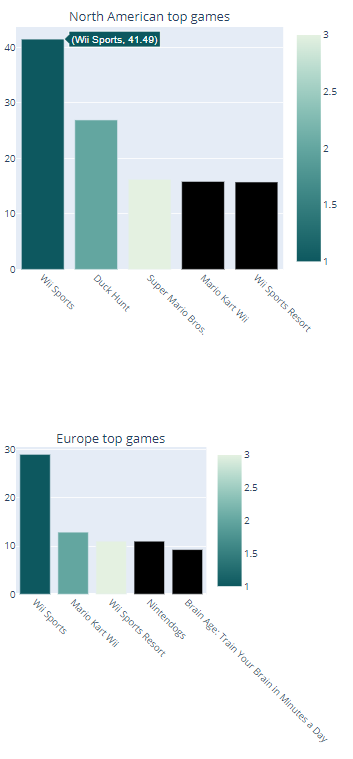
Try answering this on your own, if you get stuck you can always refer to the lines of code below.



The dataframe name\_df will give us the required result.

Plotting it in a similar fashion to understand the results obtained.





The graph shows us the top games preferred by users in different regions and also globally. We observe the following:

* **Wii Sports** has been the top game in North America, Europe, other regions.
* **Pokemon Red/Pokemon Blue** is the top game in Japan.

## 5. Are there any games with release year older than 2000 that are still making high sales? What are they?

To achieve this we will first segregate the Year less than 2000 from the rest of the dataframe. Doing this using the df.query() function.

* ***query()- This is a filtering function that enables to select and filter the columns of a dataFrame with a boolean expression.***



The above code returns a dataframe old\_games that has all the games with release date older than 2000. We can also see the first 5 rows using the function df.head(). Moreover, using the function df.shape we observe that the total number of such games are 1974. However, to answer this particular question, we will only focus on the old games making the most sales.

We are considering Global Sales to analyze the overall sales. Remember that you can use the same approach to find the sales for any other region as well.

We will target the data exceeding the 99th percentile to get only the games making the highest sales.



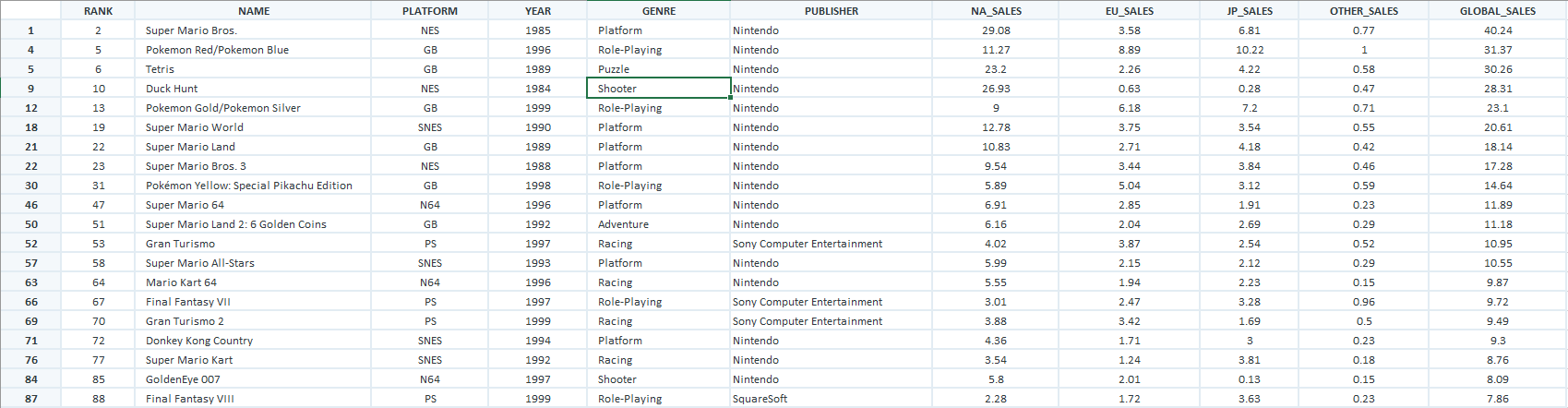
Let's consider global sales greater than $7.82 million as high and proceed with the analysis.



Here, we have only targeted those games which makes global sales more than $7.82 million.

Note that we have used this value to help us simplify our analysis. This can change according to your definition of high sales.

The final output should look like this:



**Super Mario Bros.** released in **1985** is the top old game still making global sales of **$40.24 million.** This is followed by Pokemon Red/Pokemon Blue (1996), Tetris (1989) and more.

Let's quickly hop on to the next question.

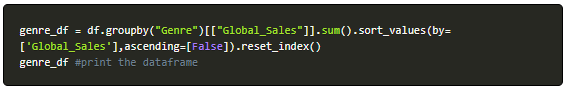
## 6. What are the top gaming genres that are making high sales?

The result can help game developers find out which genre is preferred in the market. It can also be used to compare gameplay across different games as one tool. These can be analysed and used to develop strategies for future game releases.

To analyze the sales with respect to genre, we will consider global sales.

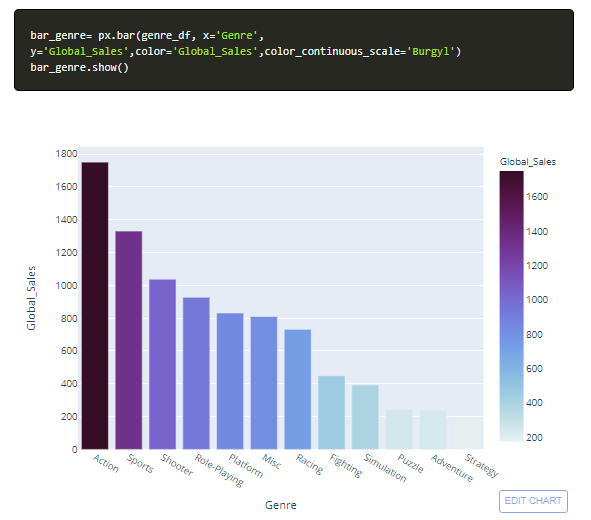
Give this question a try before looking at the code below. We will simply use the groupby function on global sales and later visualize our result using a bar graph.

This is how the code should look like:



The dataframe genre\_df will return the most preferred genres.

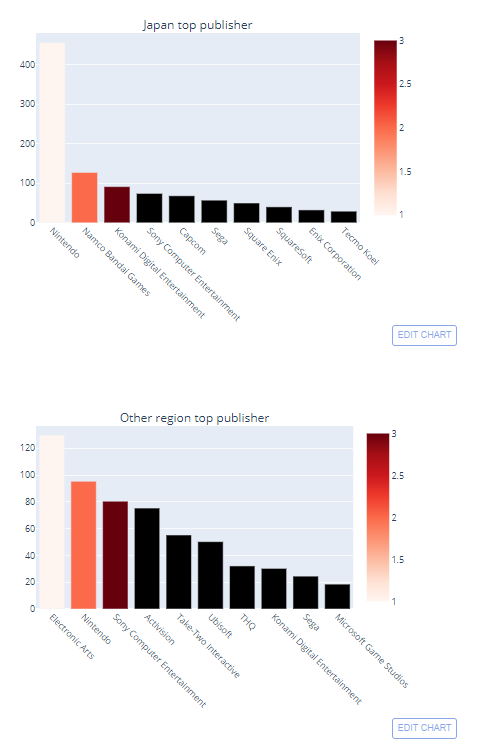
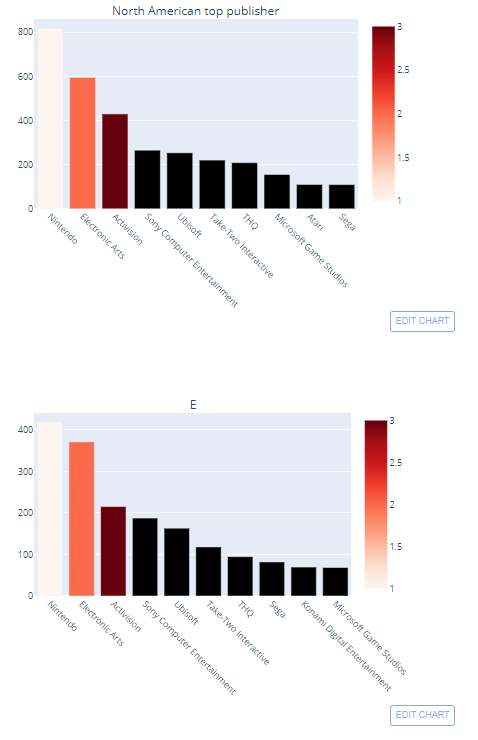




The bar graph clearly shows the preference of users for different genres. We observe that **Action**games are the most preferred across the globe, making a total of **$1751.18 million.**

## 7. Does the publisher have any impact on the regional sales?

We can quickly extract the top publishers for different regions using the same method we used in games and platforms. We can take the top 10 publishers for our analysis. Try comparing the publishers for different regions yourself and compare!



We observe the following from the bar graphs:

* **Nintendo** is the top publisher in North America, Europe and Japan making a total of **$816.87 million, $418.74 million and $455.42 million** respectively.
* For other regions, **Electronic Arts** is the top publisher making **$129.77 million.**

## 8. Is there any region that was out-performed global average sales?

We have already answered the above while answering the 1st question. We found that no region has out-performed global average sales.

You can look at the detailed explanation in the 1st question above.

And with that, we come to an end of our analysis.

## Conclusions

We have managed to find the answer to each question, and we can now share our report with the game producer. This report will contain the following information:

1. **North America** is the top region has the highest average sales of **$264,667.430**.
2. The most preferred platforms are as follows:
   * ***X360*** (Microsoft) is the top preferred console by users in North America making a total of ***$601.05 million***.
   * ***PS3*** (Sony) is the top preferred console by users in Europe making a total of ***$343.71 million***.
   * ***DS*** (Nintendo) is the top preferred console by users in Japan making a total of ***$175.57 million***.
   * ***PS2*** (Sony) is the top preferred console by users in other regions making a total of ***$193.44 million***.
3. The top most played game is **Wii Sports** making a total of **$82.74 million**       globally.
4. The top games are as follows:
   * ***Wii Sports*** has been the top game in North America, Europe, other regions.
   * ***Pokemon Red/Pokemon Blue*** is the top game in Japan.
5. **Super Mario Bros.**released in **1985** is the top old game still making global sales of **$40.24 million.** This is followed by Pokemon Red/Pokemon Blue (1996), Tetris (1989) and more.
6. **Action**games are the most preferred across the globe, making a total of **$1751.18 million.**
7. The top publishers are as follows:
   * ***Nintendo*** is the top publisher in North America, Europe and Japan making a total of ***$816.87 million, $418.74 million and $455.42 million*** respectively.
   * For other regions, ***Electronic Arts*** is the top publisher making ***$129.77 million.***
8. There is no region that has out-performed the global average sales. The average global sales are **$537,440.656 million.**

This report summarizes the top games, platforms, publishers, genres are more attributes for different regions. The results will be beneficial in improving both sales and customer satisfaction.

We hope you find this analysis useful in addressing your plans for improving the next version of your game and have a clearer vision of how to work on datasets in similar sectors while also handling graphs using plotly.