

# **Data Visualisation Assessment/Coursework**

*CS5803- Data Visualisation  
Assessment/Coursework*

*Department of Computer Science, Brunel University*

*Student ID:*

*2022/2023*

# Introduction

This submission is for coursework on Data Visualization. A tableau dashboard is created as an infographic tool to tell two stories for two research questions asked by a marketing manager of a games-producing company. This dashboard view is redesigned in Power BI for comparison of how I achieved the dashboard to answer the research questions using both methods (Tableau and Power BI).

## Data

The Global Video Game Sales dataset was retrieved from the Kaggle website in 2022<sup>1</sup>. It was originally published by Gregory Smith on the Zenodo website<sup>2</sup> in 2022 who sourced the data from vgchartz.com. The dataset captures the top 100 video games worldwide along with their respective platforms, genres, publishers, and sales regionally and globally. Primarily, there are 11 columns and 16,600 rows in the dataset. However, additional columns are computed from the already existing variables to facilitate our storytelling. The data dictionary below shows the key original variables along with the new variables shown in bold:

Name	Description	Domain
Rank	Ranking of the game based on global sales	Numerical (Integer)
Name	Name of the game	Categorical (String)
Platform	Platform the game was released on	Categorical (String)
Year	Year the game was released (1980 – 2020)	Numerical (Integer)
Genre	Genre of the game	Categorical (String)
Publisher	Publisher of the game	Categorical (String)
NA_Sales	Sales of the game in North America	Numerical (Float)
EU_Sales	Sales of the game in Europe	Numerical (Float)
JP_Sales	Sales of the game in Japan	Numerical (Float)
Others_Sales	Sales of the game in other regions	Numerical (Float)
Global_Sales	Total sales of the game worldwide	Numerical (Float)
<b>Pct NA_Sales</b>	<b>Percentage of global sales sold in NA</b>	<b>Numerical (Float)</b>
<b>Pct EU_Sales</b>	<b>Percentage of global sales sold in EU</b>	<b>Numerical (Float)</b>
<b>Pct JP_Sales</b>	<b>Percentage of global sales sold in JP</b>	<b>Numerical (Float)</b>
<b>Pct Others_Sales</b>	<b>Percentage of global sales sold in others</b>	<b>Numerical (Float)</b>

## Persona and Questions

The user is a marketing manager for a sports game-producing company. He wants to create targeted advertising campaigns for upcoming games they are producing. He has asked the following questions:

1. In which year did the most popular sports games publisher record his highest growth in global sales and which platform influenced its growth in sales the most?
2. Between 2011-2016, was there a strong correlation in sales of sports games between any pair of regions and which was the top-selling sports game common to the regions?

## Requirements

Here we specify the main requirements in terms of the relationships that must be visualised to answer each question, followed by brief design ideas for representations and interactions.

1. To answer Q1, the user will need to see a Treemap showing the count of Genres per publisher.  
The following filter will be applied:
  - i. Genre (Sports)
2. To answer Q1, the user would need to see a line graph of global sales over the years. The year will be on the x-axis and global sales on the y-axis.  
The following filter will be applied:
  - i. Genre (Sports)
  - ii. Publisher (Most popular from R1)
3. To answer Q1, the user will need to visualize the relationship between global sales made per platform. This can be achieved with a bar plot of the platform (on the x-axis) against global sales (on the y-axis).  
The following filter will be applied:
  - i. Genre (Sports)

---

<sup>1</sup> <https://www.kaggle.com/datasets/thedevastator/global-video-game-sales>

<sup>2</sup> <https://zenodo.org/record/5898311#.ZASwVezP1QJ>

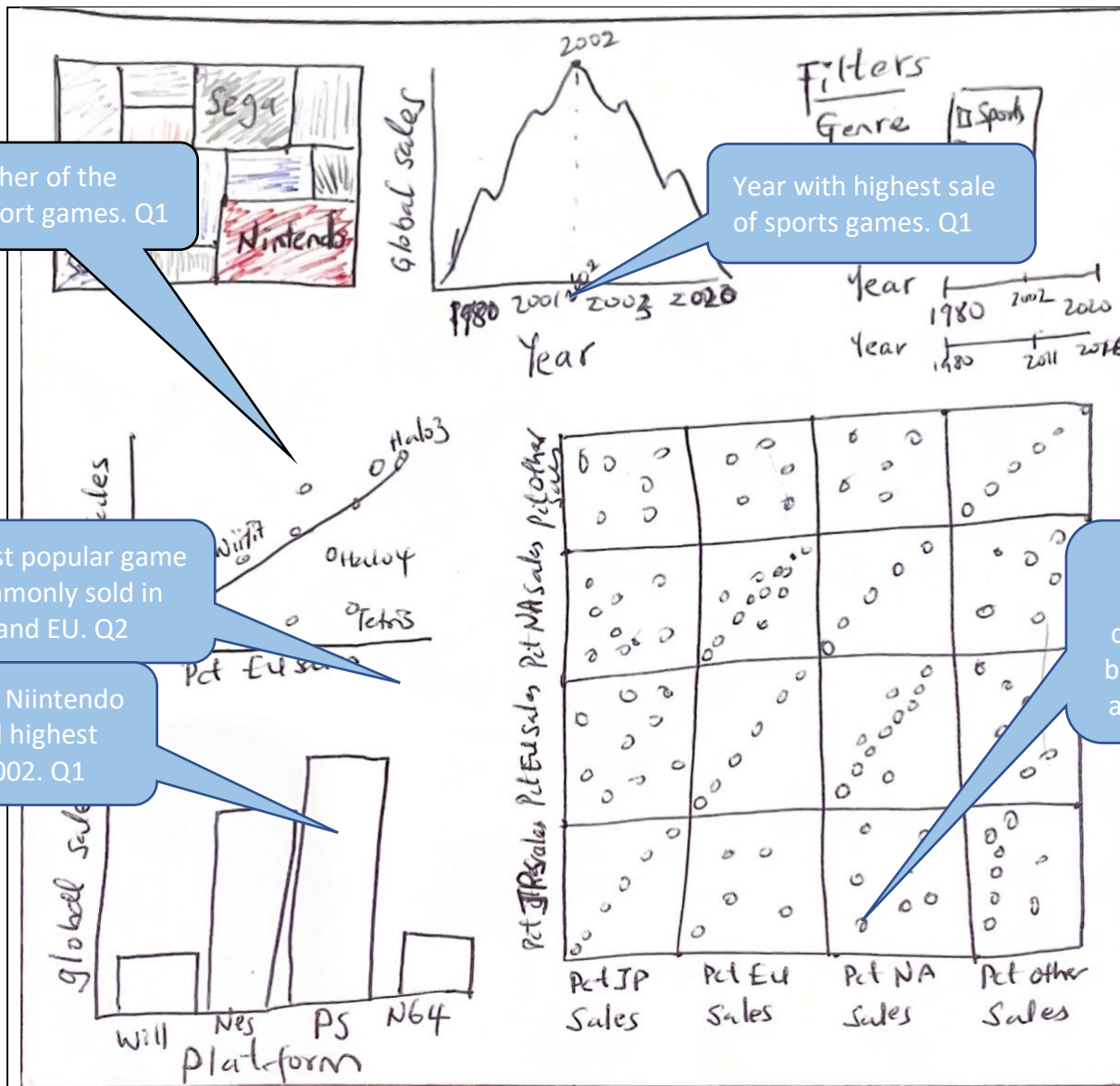
- ii. Publisher (Most popular from R1)
- 4. To answer Q2, the user will need to see a scatter plot matrix comparing the percentage of global sales of sports games between the regions.  
The following filter will be applied:
  - i. Year (2011 – 2016)
  - ii. Genre (Sports)
- 5. To answer Q2, the user will need to see a scatter plot comparing the sales of the correlating regions.  
The following filter will be applied:
  - i. Year (2011 – 2016)
  - ii. Genre (Sports)

## Design

This section includes, first a prototype design of how the solution to the research questions will look (following the requirements) and an illustration of how the questions are answered through the views. Second, screenshots of the final implementation on Tableau will be included with annotations.

### Paper Landscape

Global Video Game Sales
<b>Goal: To create targeted advertising campaigns for upcoming sports video games in production.</b> Q1. In which year did the most popular sports games publisher record his highest growth in global sales and which platform influenced its growth in sales the most? Q2. Between 2011-2016, was there a strong correlation in sales of sports games between any pair of regions and which was the top-selling sports game common to the regions?



### Result

1. The marketing manager can see Nintendo is the most experienced publishing company for sports video games.
2. He can also see that Nintendo recorded its highest sale of sports games in 2002.
3. He can also see that using PS platform produced the highest sales of sports games globally with Nintendo as the publisher in 2002.
4. The marketing manager can see that there is a positive correlation between sales of sports games in EU and NA in the 2011 - 2016.
5. He is able to identify Halo 3 as the most popular sports game sold commonly to EU and NA in 2011 - 2016.

# Implementation

## Tableau Implementation:

1. I started by loading my data into Tableau and confirmed the variables were correctly read in as measure values or measure names.
2. I created three calculated fields for per cent sales per global sales for each of the regions (JP, NA, EU, and others). The formula used was-  $(\text{Region sales}/\text{Global sales}) * 100$ .  
This field is required when sales are compared by regions to give more accurate results instead of comparing based on the prices of games. This also solves the problem of imbalance (discovered during the preliminary exploration of the dataset in R) in the sale prices as a few games are really expensive compared to the majority.

### Implementing Requirement 1 for Q1

- I placed the Genre on the Rows and changed the measure to “count” while the Publisher was placed on the columns tab.
- From the “Show Me” tab I chose the Treemap.
- I then dragged the Genre to the filter shelf and selected “Sports” I applied this filter to all worksheets using the data source.
- From the chart, we see Electronic Art is the most popular publisher of Sports games.
- I right-clicked on Electronic Arts to include point annotation and made the boundary bold and black to aid visualization.

### Implementing Requirement 2 for Q1

- I placed the year in the columns and the global sales in the rows (measure is sum)
- Genre (Sports) is automatically applied to the filter, so I then filtered for the Publisher (Electronic Arts).
- I added the year to the marks label.
- From the chart, the highest growth in sales occurred between 2009 to 2010. I annotated this again.

### Implementing Requirement 3 for Q1

- I placed the Global Sales on the Rows (measure is sum) and Platform on the column. Tableau immediately produces the bar plot.
- I filtered the Publisher to the top publisher from R1(Electronic Arts) and the year from R2 (2009 to 2010).
- Then I sorted the bars in descending order.
- From this view, we can see that Electronic Arts made most of its sales in 2009 -2010 through the PS3 platform.

### Implementing Requirement 4 for Q2

- I placed the 4 regions’ sales (Pct NA Sales, etc.) on both the row and columns per region to form a scatter plot matrix. Then from the analysis tab, I unchecked the aggregate measure.
- The region is filtered to the Year is filtered 2011- 2016. This filter is applied on selected worksheets (R4 & R5).
- Finally, from the Analytics tab, I added a trend line.
- From the plot, we can see that NA and EU seem to have a strong negative correlation.

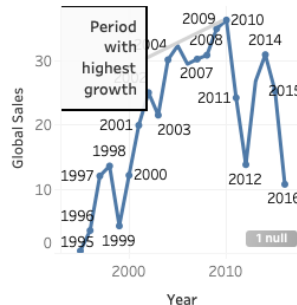
### Implementing Requirement 5 for Q2

- I dragged the NA sales and EU sales to the rows and columns respectively and from the Analysis tab, I unchecked the Aggregate measure. This produced a scatterplot as expected for comparing two numerical variables.
- From Analytics, I included a trend line. The filter was already applied automatically.
- I dragged the name of the game to the marks label.
- From the plot, I could get that, FIFA 15 was the top-selling sports game common to NA and EU from 2011 - 2016.

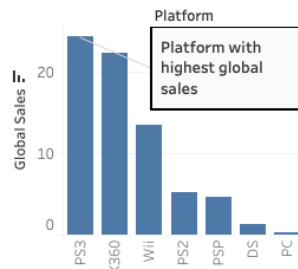
# Comparison of Tableau Final Dashboard to Proposed Design

## Video Games Sales Dashboard

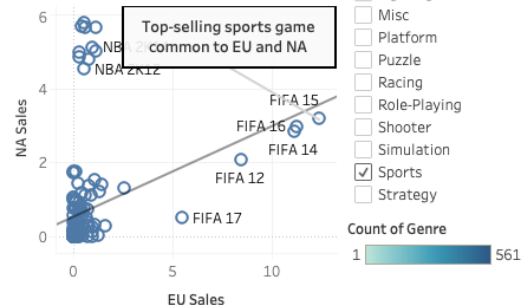
Which year recorded highest sales of Sports games?



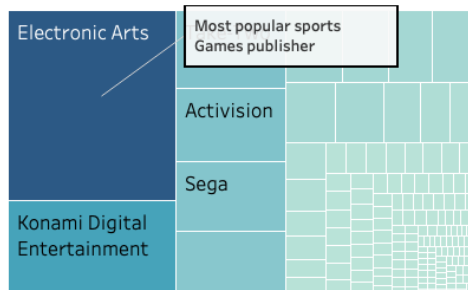
Through which platform did Electronic arts sell most sports games in 2009-2010?



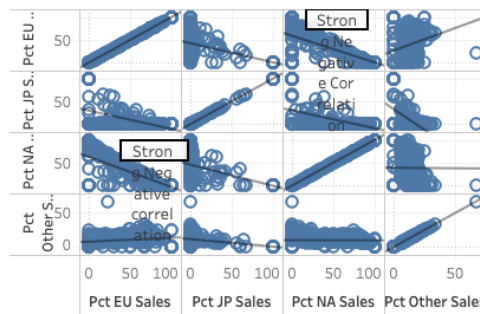
Which Games are the most commonly sold in EU and NA in the last 5 years?



Most Experienced Publisher of Sports games

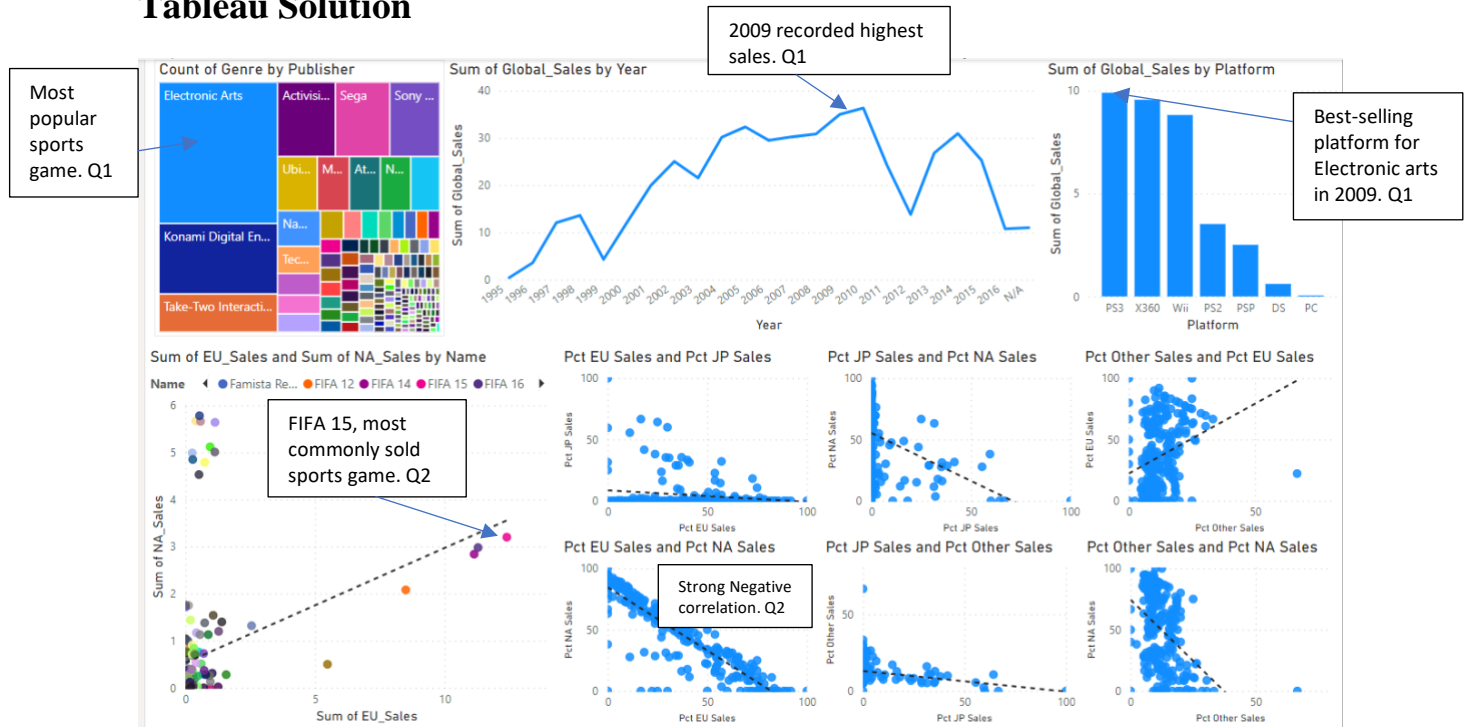


Is there a strong correlation in sales of Sports games between any regions in the last 5 years?



- All the views on the proposed design were implementable on Tableau.
- The colour code for the heat map was better to read using the colour density instead of the colour hue for the count of the genre.
- In the proposed design, I had compared the Global sales directly with the year and platform however, during implementation, I discovered I had to make the comparison with the aggregate “Sum” of global sales.
- I decided to sort the bar graph for global sales per platform in descending order. Sorting in Tableau aids correct visualization of the bar lengths, especially for cases when the lengths are close.
- There was no need to compare the percentage sales for NA and EU when looking for the common game. I decided to compare the sum of sales instead.

# Comparison of Implementation in Power BI to Proposed Design and/or Tableau Solution



Power BI was able to recreate 4 views quite easily except for the 4<sup>th</sup> requirement (Scatter plot matrix). In all views, the same results were arrived at with the Tableau dashboard. Some challenges/differences in implementation are outlined below.

- Power BI did not have the scatterplot matrix visual. I saw online that a correlation matrix was achievable in Power BI but after signing in for the pro subscription I was unable to find the visual. Eventually, I implemented the requirement, by manually plotting a scatter plot comparing each pair of regions instead of a matrix.
- In Power BI, the variables were pulled on the category pane and appropriately placed in the x-axis, and y-axis with the appropriate measures. Some of the visuals are displayed in the visualization pane and some extras can be imported if needed.
- In Power BI there was no need to create separate sheets for each chart as all views could be created directly on the dashboard and all filters are applied on the same sheet.
- I was unable to customize the chart title in Power BI as I did in Tableau as Power BI named the charts automatically.
- Power BI does not have a function to choose what measure should be represented on the labels. This affected the scatterplot of games common to NA and EU. The names would have been better displayed as labels. And for the sales per year. I would have preferred the year as labels, but this was not achievable.
- Power BI selected the colour code for the heat map. There was no feature to adjust the colour styles.
- While implementing Requirement 2 (line graph) and 3 (bar graph), Power BI specifically labelled the Global sales column as “Sum of Global Sales” whereas Tableau only called this “Global Sales” even though both cases represented the sum.
- In implementing requirement 5 (Scatter plot), Power BI required I encoded the names of the games with colour.

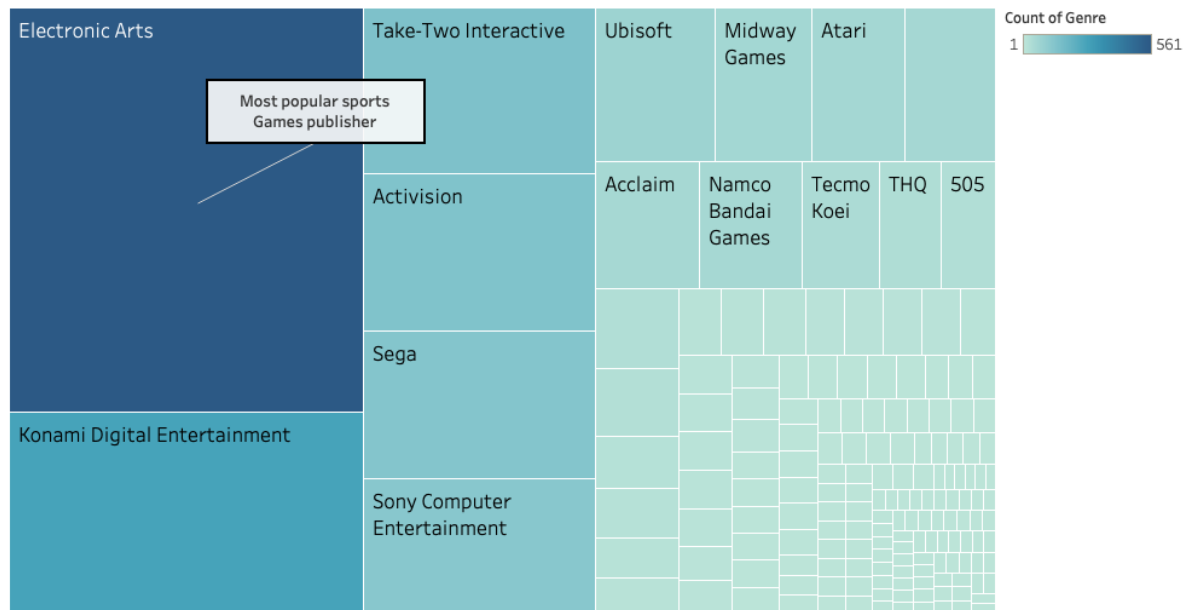
## WALKTHROUGH

From the implementation in both Tableau and Power BI, the research questions have been sufficiently answered. **Question 1** was answered by **Requirements 1,2 and 3** while **Question 2** was answered by **Requirements 4 and 5**. Referencing the Tableau Story, we will answer both research questions by walking through the annotated implementation of the requirements to arrive at the answers.

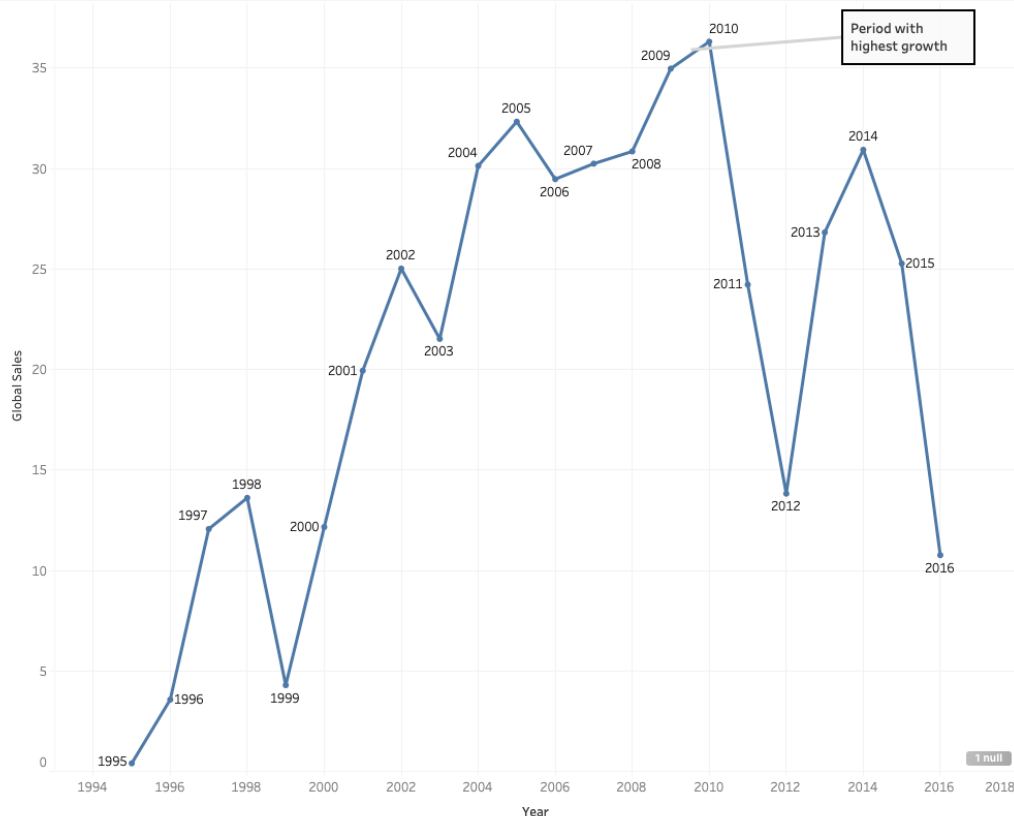
**ANSWERING QUESTION 1:** In which year did the most popular sports games publisher record his highest growth in global sales and which platform influenced its growth in sales the most?

**Answer:** From R1, R2, and R3, the year in which the most popular sports game publisher (Electronic Arts) had its highest sale was 2009 and the platform that influenced their popularity the most was PS3.

**For requirement 1(R1)**, the view shows the publisher that was most popular for publishing sports games. From the heatmap, The colour density shows the number of sports games published by different types of publishers. The darker blue represents the highest count of Genre as shown in the key. From our chart, **Electronic Arts** published the most sports games.

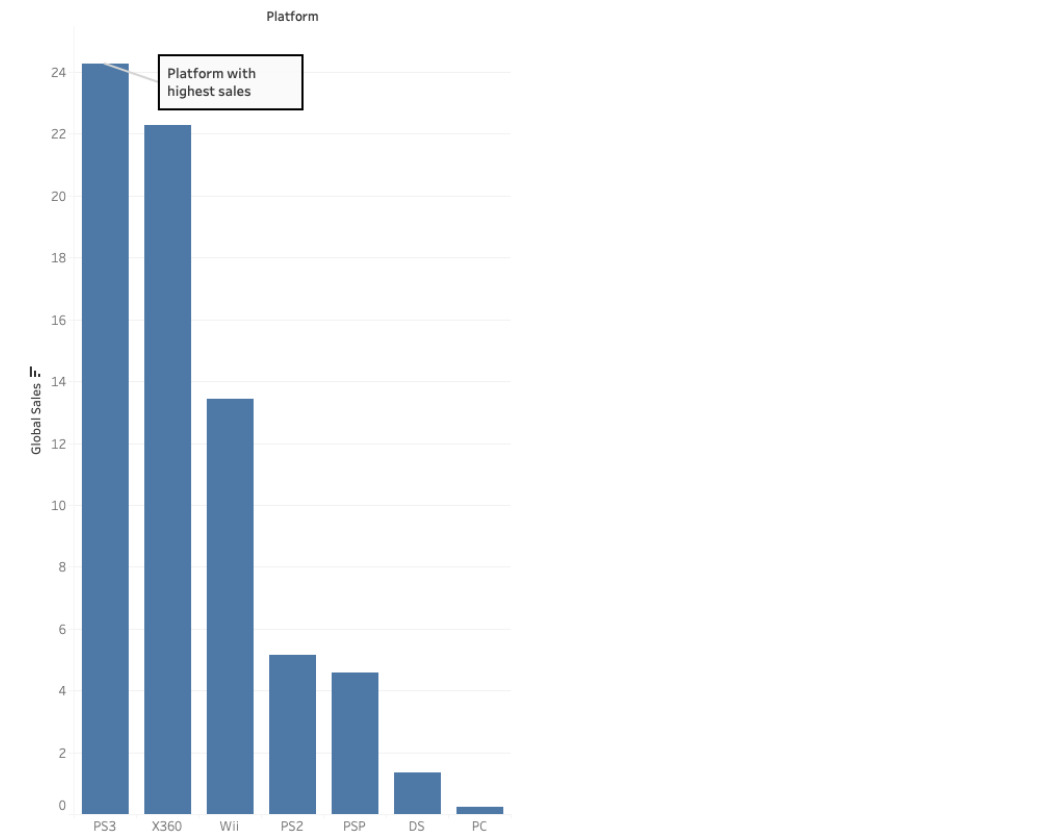


**For requirement 2 (R2)**, the view shows the period with the highest growth in sales of sports games. The highest peak of the line graph represents the year with the highest sales. From our chart, 2009 records the highest sales.



**For requirement 3 (R3)**, the view shows the global sales of games by the platform for sports games published by Electronic arts. The length of the bars shows the global sales of games for each platform and the longest bar is the best platform used for selling sports games. From our bar chart, PS3 was the best platform used to bring the highest sale.

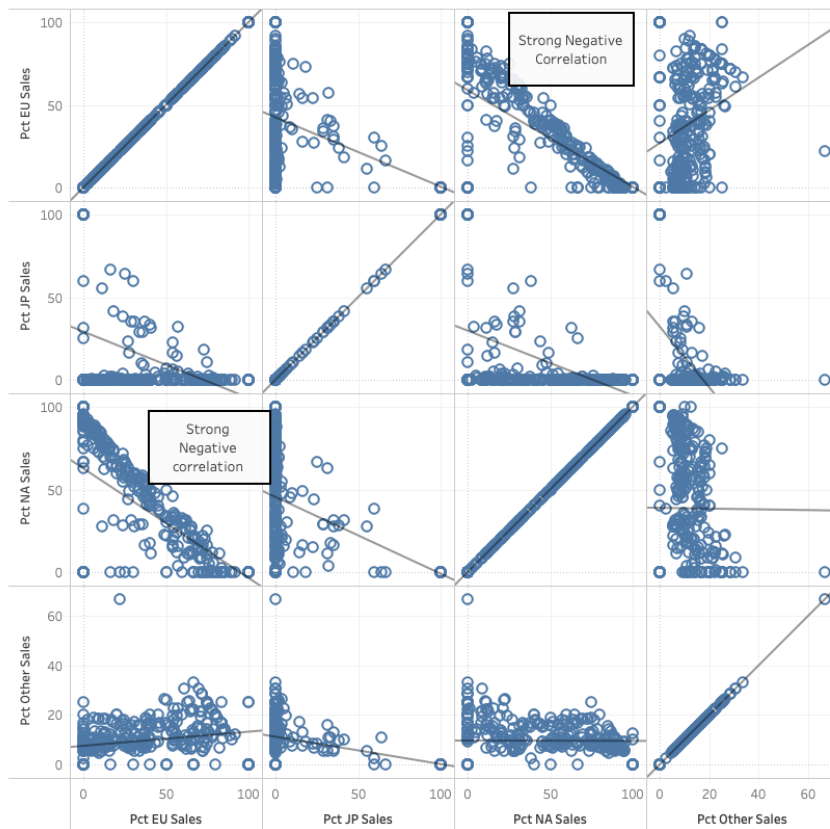




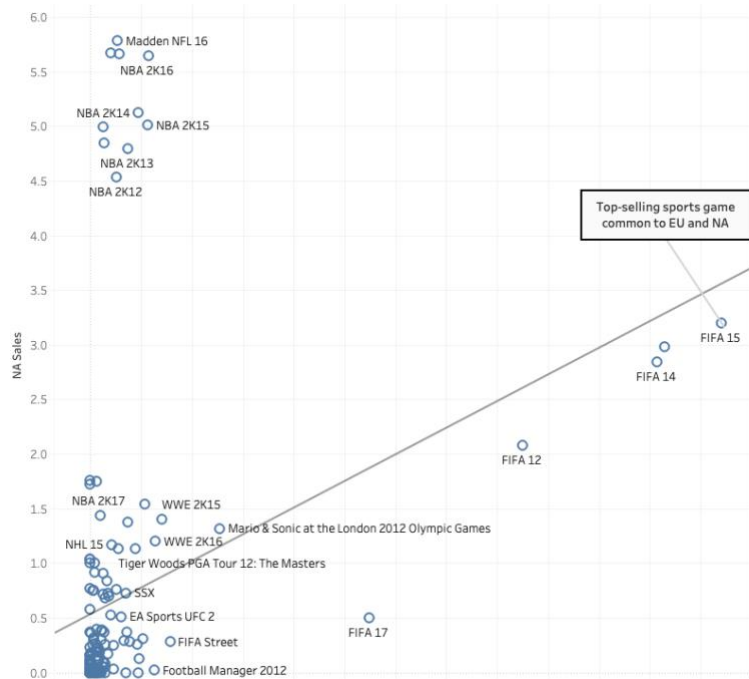
**ANSWERING QUESTION 2:** Between 2011-2016, was there a strong correlation in sales of sports games between any pair of regions and which was the top-selling sports game common to the regions?

**Answer:** From R4 and R5, we could see that there was a strong negative correlation between the EU(Europe) and NA (North America) and FIFA 15 was the top-selling game popular in the two regions.

**For requirement 4 (R4),** the view created shows the relationship between the sales of sports games across the regions in the year 2011-2016. A strong positive correlation will show in a diagonal cluster of points across the axes from the bottom left to the top right. A strong negative correlation shows a diagonal cluster of points falling from top left to bottom right across the axes. A random distribution of points or points forming a horizontal or vertical cluster indicates either weak or no correlation between the regions of the axes. From our scatter plot matrix, EU and NA are seen to have a strong negative correlation.



**For requirement 5 R5),** the view shows the most popular games common to EU and NA. The points represent the names of the games sold in the regions. The points closer to the trend line are more commonly sold in the two regions and the points farther away to the top right are the highest-selling common games. From the scatter plot, FIFA 15 was the top-selling sports game common to the EU and NA.



## Reflective Discussion

Implementing this project has brought to mind that data visualization is beyond the ability to create nice and colourful charts on software but being able to express the large information in data in a way that is concise, visual, appealing and user-controlled. Different persona would have different interests in the same data. During the stages of the project, I had to constantly consider the context of use to ensure information was correctly represented. For example, comparing sales by region as a percentage of global sales was to ensure comparison is based on the relative sales and not prices which may be greatly imbalanced for some regions.

Getting the right visual design starts by choosing the right software, choosing the right charts that are captivating and unambiguous, and then using features like filters, zoom, annotations, and brushing that enable users to interact with our data (Shneiderman, 1996). By implementing the solution of two different software (Tableau and Power BI), I noted the pros and cons of each and the context where each would be preferable to use. Tableau, for instance, availed a more user-friendly way to discuss data using readily available features like annotations on charts and a story sheet. A downside of Tableau, however, is that is expensive. I believe this is because Tableau is server-based. Power BI on the other hand is cloud-based (Khatuwal and Puri, 2022) and offers a free package which can be upgraded to access more sophisticated visuals. Power Bi also has capabilities that make it more suited for comparison and statistical analysis, especially with the embedded R and Python script features.

## Conclusion

The objective of the project to develop an interactive data visualization, using Tableau, to derive useful insight from a dataset was well achieved. The methodology of the project showed a practical application of User-Centered Design (UCD) guidelines to Human-Computer Interaction (HCI). Norman (1988) defined HCD (human-centered design) as “starting with a good understanding of people and the needs that the design intended to meet.” This project structure tactically summarized the four stages of UCD as Study/Analysis/Requirement, Design/Prototype, Build/Development, and Evaluation (Mithun and Yafooz, 2018). In the first stage, the human-centered process was planned by creating the persona, defining the context of use, and specifying the user requirements. In the second stage, a prototype of how the solution will be implemented was created. The third stage involved the development of the design and finally, the last stage of evaluation of results to ensure the user requirements were effectively and efficiently satisfied (González, et.al., 2010).

The implementation of this project has taught me many valuable lessons, but one challenge leads me to the quest for more studies. I expected to be able to produce a correlation plot using Power Bi to give more context to my visualization. Although there was information online that the view was achievable, I could not find the correlation visual on the software even after signing up for the pro version. I will take the Microsoft Power BI certification course to gain more insights on using its advanced features, especially statistical analysis.

## References

Gregory Smith. (2022). Video Games Sales [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5898311> (Accessed: 13<sup>th</sup> March 2023)

Shneiderman, B (1996) The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. <http://hcil2.cs.umd.edu/trs/96-13/96-13.html>

Norman, D. (1988) The Design (Psychology) of Everyday Things.

A. M. Mithun, A. M. Mithun and W. M. S. Yafooz, "Extended User Centered Design (UCD) Process in the Aspect of Human Computer Interaction," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Shah Alam, Malaysia, 2018, pp. 1-6, doi: 10.1109/ICSCEE.2018.8538388.

C. S. G. González, E. G. González, V. M. Cruz and J. S. Saavedra, "Integrating the Design Thinking into the UCD's methodology," IEEE EDUCON 2010 Conference, Madrid, Spain, 2010, pp. 1477-1480, doi: 10.1109/EDUCON.2010.5492352.

V. S. Khatuwal and D. Puri, "Business Intelligence Tools for Dashboard Development," 2022 3rd International Conference on Intelligent Engineering and Management (ICIEM), London, United Kingdom, 2022, pp. 128-131, doi: 10.1109/ICIEM54221.2022.9853086.