To make it easier in grading the assignment, please delete the English/Japanese part that is not being used. 課題の採点を容易にするため、使用していない英語／日本語の部分を削除してください。

Assignment Lesson 3

- Data Exploration -

Data Exploration is an important step to know your data better before performing any machine learning model.

In this assignment, you will try to do Data Exploration steps and give some comments/ analysis. You are **free to use any dataset you are interested in** as long as it satisfies the following conditions:

- Contains at least **3 regular numerical** attributes
- Contains at least **1 regular categorical** attribute (different than the label)
- Contains at least **40 examples**
- Contains **a categorical label attribute**
- File size less than 5 MB

Feel free to priorly adjust the data (discretization or just select a few attributes) so it will fit those requirements. If you are struggling to find one, you can look at this and this site for inspiration. To complete this assignment feel free to choose any statistical tool you preferred.

[!] *Three students who submit their assignments by the deadline with the most interesting data and data exploration results will receive a* **Google Play / Apple Gift / Amazon Gift Card worth 3000 yen**. *Winners will be announced approximately two weeks after the deadline. By clicking the boxes,* ***cross two*** *of the following items that you wish* ***not to get****:*

- ~~Google Play Gift Card~~
- ~~Apple Gift Card~~
- Amazon Gift Card

Prior Knowledge

● Data Science Objective 5/5

Using the data you have, write down the data science objective (such as what kind of prediction to do).

Possible questions:

- Is the video game consumption behavior linked to our geographical zone?
- Does the most popular genre depend on the used platform?
- Does the most popular genre depend on the geographical region?
- Has a publisher the monopole of the most popular genre?
- Over the last two decades, did the most popular genre change?

We decided to focus on

- **Is the video game consumption behavior linked to our geographical zone?**

Upload the data to your Google Drive and paste the link here. It can be in a **csv or xls** format (not compressed like zip). Please make sure the sharing setting allows Tohoku University members to see it.

Link: https://drive.google.com/file/d/1xtYVJoYKxe1rmPWoe7QJP5QdQudojncO/view?usp=sharing

Original data came from here: https://www.kaggle.com/datasets/gregorut/videogamesales

Data adjustments:

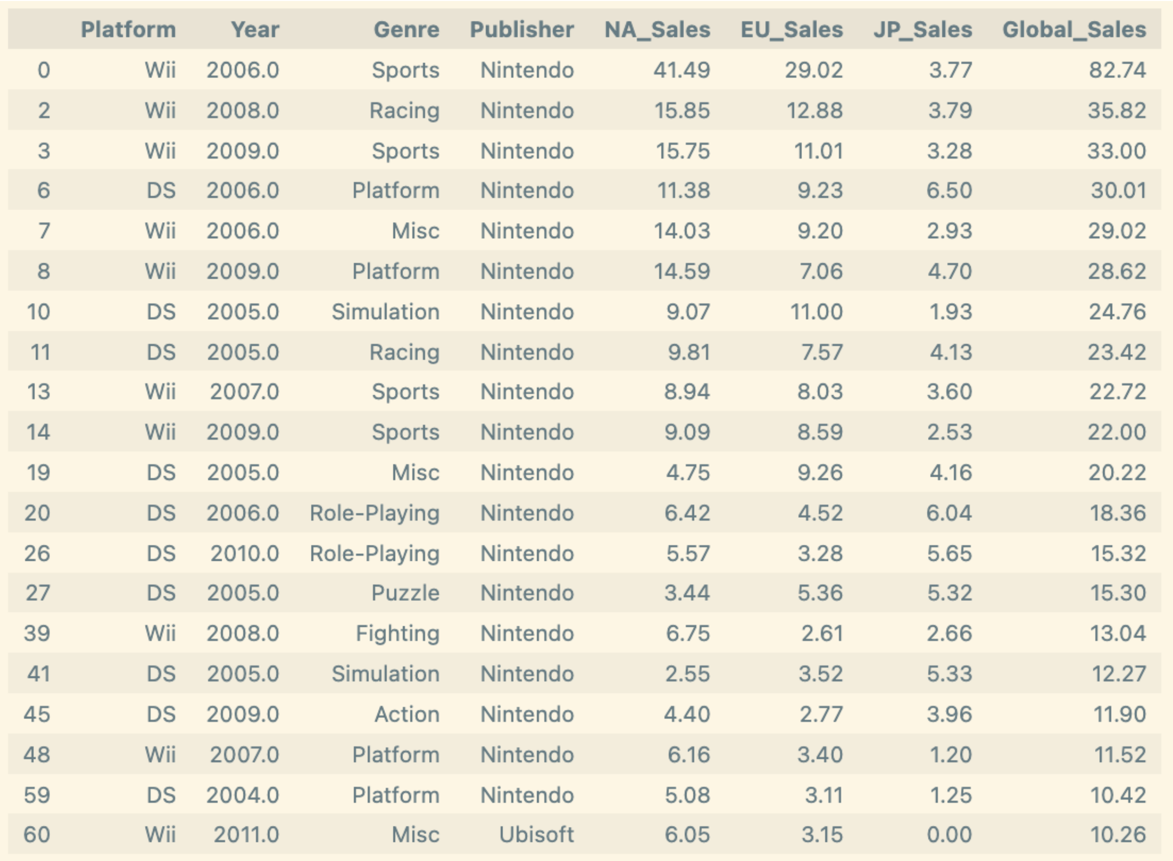
- Year >= 2000
- Publishers with less than 500 occurrences were discarded. We only kept nine classes out of 20 (`"Electronic Arts"`, `"Activision"`, `"Ubisoft"`, `"Namco Bandai Games"`, `"Konami Digital Entertainment"`, `"THQ"`, `"Sony Computer Entertainment"`, `"Nintendo"`, `"Sega"`)
- Platforms were grouped with this pattern:
    - `Playstation`: Home consoles made by Sony
    - `PSP`: Portative consoles made by Sony
    - `Wii`: Home consoles made by Nintendo
    - `DS`: Portative consoles made by Nintendo
    - `XBox`: Home consoles made by Microsoft
    - `PC`: Computer games
    - Other consoles were discarded

At the end the file contains 2456 rows.

Data Exploration

Organize the Dataset 5/5

Give a screenshot of your dataset. The top 10-20 rows of the data will suffice.



	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Global_Sales
0	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	82.74
2	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	35.82
3	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	33.00
6	DS	2006.0	Platform	Nintendo	11.38	9.23	6.50	30.01
7	Wii	2006.0	Misc	Nintendo	14.03	9.20	2.93	29.02
8	Wii	2009.0	Platform	Nintendo	14.59	7.06	4.70	28.62
10	DS	2005.0	Simulation	Nintendo	9.07	11.00	1.93	24.76
11	DS	2005.0	Racing	Nintendo	9.81	7.57	4.13	23.42
13	Wii	2007.0	Sports	Nintendo	8.94	8.03	3.60	22.72
14	Wii	2009.0	Sports	Nintendo	9.09	8.59	2.53	22.00
19	DS	2005.0	Misc	Nintendo	4.75	9.26	4.16	20.22
20	DS	2006.0	Role-Playing	Nintendo	6.42	4.52	6.04	18.36
26	DS	2010.0	Role-Playing	Nintendo	5.57	3.28	5.65	15.32
27	DS	2005.0	Puzzle	Nintendo	3.44	5.36	5.32	15.30
39	Wii	2008.0	Fighting	Nintendo	6.75	2.61	2.66	13.04
41	DS	2005.0	Simulation	Nintendo	2.55	3.52	5.33	12.27
45	DS	2009.0	Action	Nintendo	4.40	2.77	3.96	11.90
48	Wii	2007.0	Platform	Nintendo	6.16	3.40	1.20	11.52
59	DS	2004.0	Platform	Nintendo	5.08	3.11	1.25	10.42
60	Wii	2011.0	Misc	Ubisoft	6.05	3.15	0.00	10.26

Analysis (in **20-50 words**)

(Hint: Explains the types of attributes in the data and which one will be used to help in answering the data science objective?)

- Genre - Genre of the game: categorical label attribute
- Platform - Platform of the release of the game (i.e., PC, PS4, etc.): regular categorical
- Publisher - Publisher of the game: regular categorical
- Year - Year of the game's release: regular numerical
- NA_Sales - Sales in North America (in millions): regular numerical
- EU_Sales - Sales in Europe (in millions): regular numerical
- JP_Sales - Sales in Japan (in millions): regular numerical
- Global_Sales - Total worldwide sales (in millions): regular numerical Global_Sales is the sum of the EU_Sales, NA_Sales, JP_Sales, and the other sales.

The columns Genre, Platform, NA_Sales, EU_Sales, JP_Sales, and Global_Sales are going to be helpful to answer the data science objective.

● Find central points and spread 10/10

Give a screenshot of the **mean**, **median, standard deviation, min, max** for numerical data, **mode**/most occurring element for categorical attributes.

- Genre

  Mode = Action with 432 out of 2456 occurrences (18%)

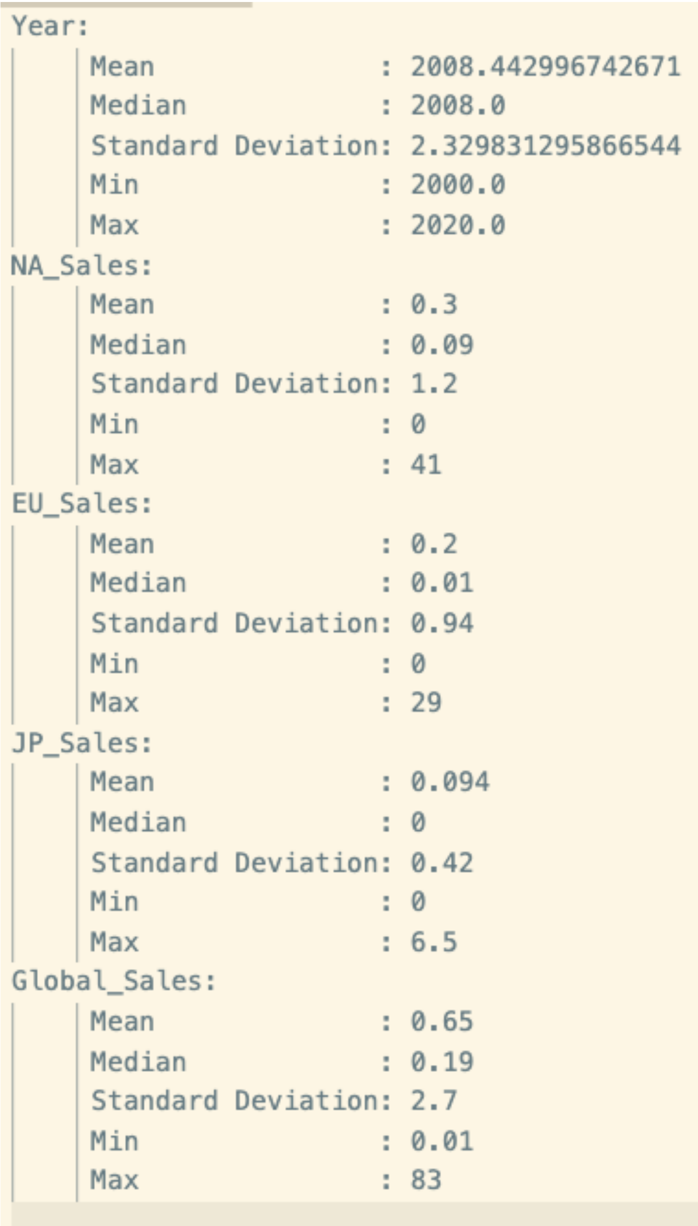
  Number of classes = 12
- Platform

  Mode = DS with 942 out of 2456 occurrences (38%)

  Number of classes = 4
- Publisher

  Mode = Ubisoft with 410 out of 246 occurrences (17%)

  Number of classes = 9



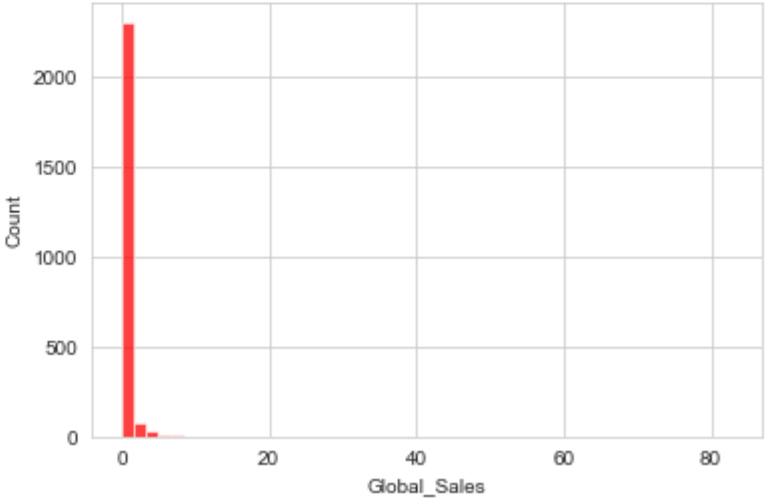
```
Year:
|    | Mean              : 2008.442996742671
|    | Median            : 2008.0
|    | Standard Deviation: 2.329831295866544
|    | Min               : 2000.0
|    | Max               : 2020.0
NA_Sales:
|    | Mean              : 0.3
|    | Median            : 0.09
|    | Standard Deviation: 1.2
|    | Min               : 0
|    | Max               : 41
EU_Sales:
|    | Mean              : 0.2
|    | Median            : 0.01
|    | Standard Deviation: 0.94
|    | Min               : 0
|    | Max               : 29
JP_Sales:
|    | Mean              : 0.094
|    | Median            : 0
|    | Standard Deviation: 0.42
|    | Min               : 0
|    | Max               : 6.5
Global_Sales:
|    | Mean              : 0.65
|    | Median            : 0.19
|    | Standard Deviation: 2.7
|    | Min               : 0.01
|    | Max               : 83
```

Analysis (in **30-100 words**)
(Hint: Explains the difference in mean and median of the numerical data? Say something about how the data are spread. Also, do the classes in the categorical attributes equally distributed or not?)

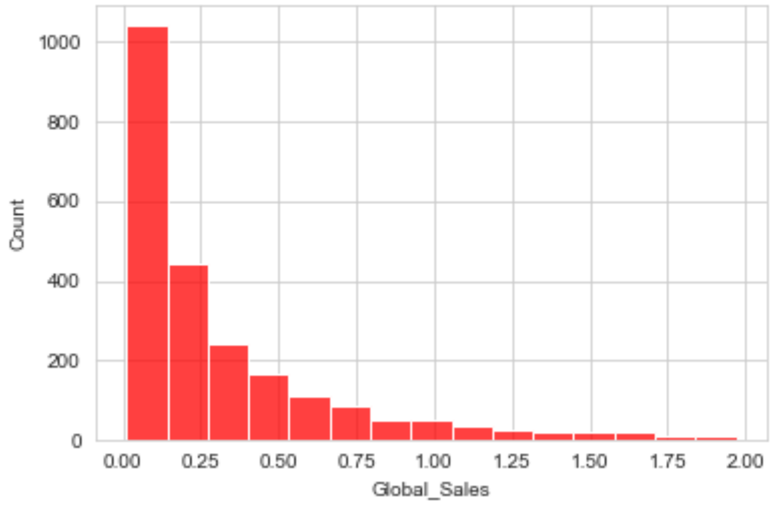
- The mean and the median of the attribute **Year** are very similar, which means that the values are equally distributed around the mean value. Apparently, there are no outliers for the Year column. However, the values are not equally distributed between the oldest year (2000) and the most recent year (2020) because if it were the case, the mean would be around 2010.
- For the **sales columns**:
  - we can see that the median is lower than the mean. It's because few games have a high number of sales, whereas many games have a number of sales between 0 and 1. *The sales are not equally spread, and there are outliers.*
  - We can see that for each geographical zone, the minimal sale is 0. This is not the case for global sales, *which means that a game is at least sold somewhere around the world.*
  - The maximal value of global sales is more significant than the maximal value of each geographical, meaning that apparently, *when a game is popular, it is popular in multiple geographical zones.*
- **Genre**: the frequency of the mode is way two times higher than if the data were equally distributed (100/12 = 8%). The data are not equally distributed
- **Platform**: the frequency of the mode is higher than 100/4 = 25%, which means the data are not equally distributed.
- **Publisher**: the frequency of the mode is a little bit higher than 100/9 = 11%, which means the data are almost equally distributed.

● Visualize the distribution of each attribute 14/15

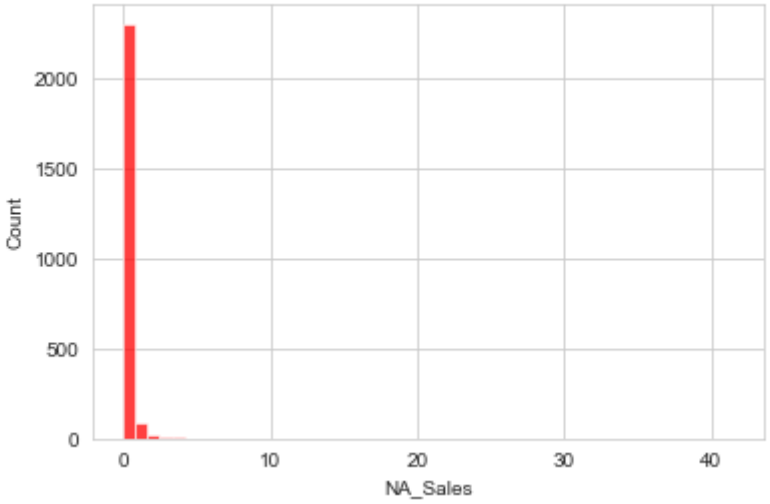
Give a screenshot of the **histogram** and/or **distribution plots** for the numerical attributes.



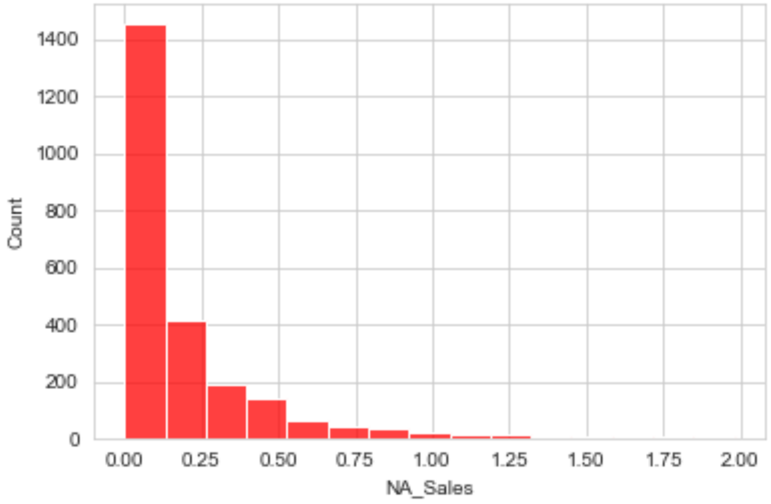
Global Sales:



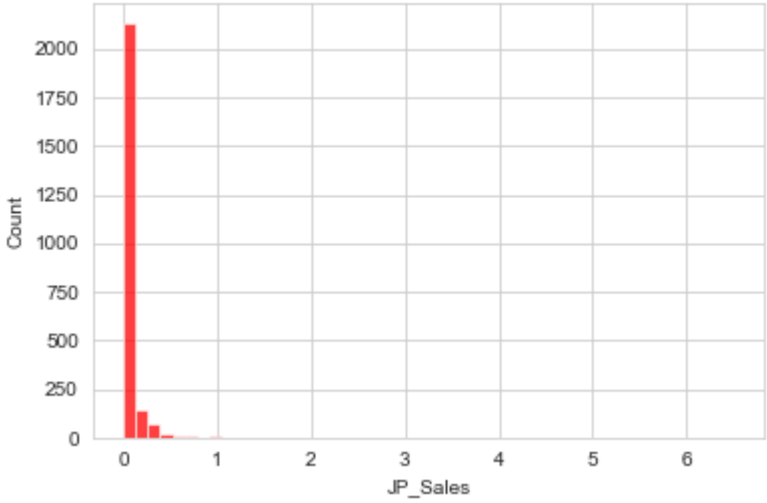
Zoom of Global Sales:



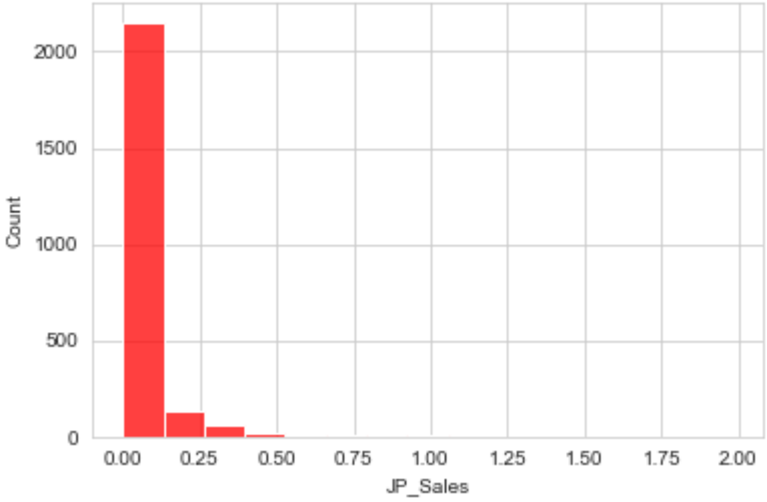
NA Sales:



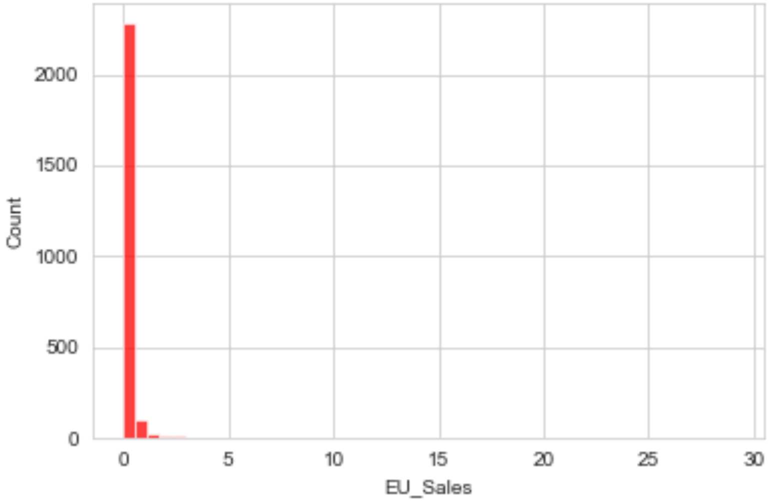
Zoom of NA Sales:



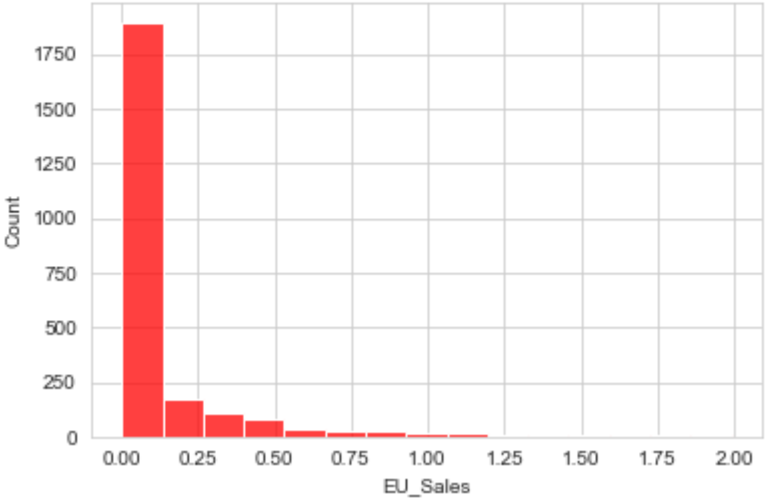
JP Sales:



Zoom of JP Sales:



EU Sales:



Zoom of EU Sales:

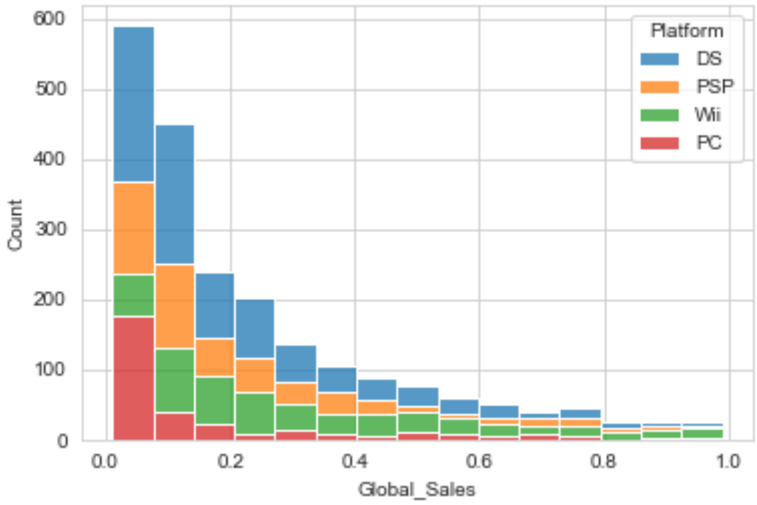
Analysis (in **30-100 words**)
(Hint: What can you say about the overall data distribution? Does it follow a particular distribution or not?)

The histograms validate our first impression of the data distribution of numerical values:

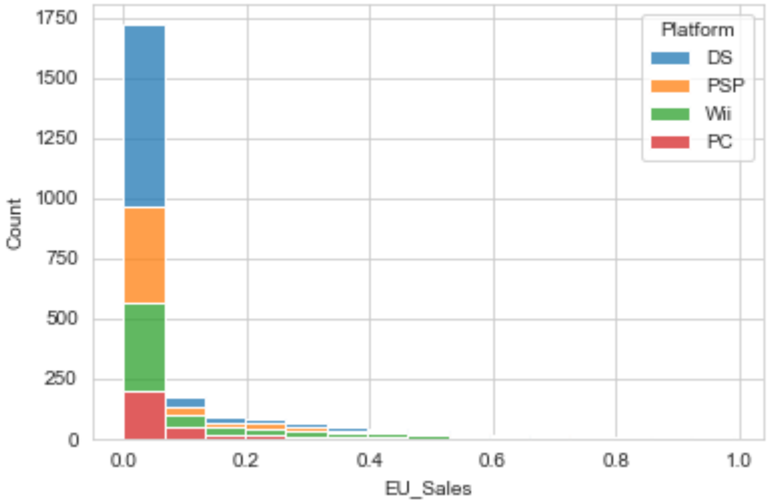
- The columns NA_Sales, EU_Sales, JP_Sales, and Global_Sales contain **outliers,** and the data is **not equally distributed**. Indeed we can see that in NA_Sales, EU_Sales, JP_Sales, and Global_Sales, most of the values are between 0 and 1, but they have high singular values..
- The **Year values are equally distributed** with a center around 2008. The plot shape looks like a Gaussian. We can see that there are very few values after 2016
- Each distribution seems to be decreasing and exponential

Give a screenshot of the **class-stratified histogram** and/or **distribution plots** for the numerical attributes. You can use any categorical column to perform the separate class.

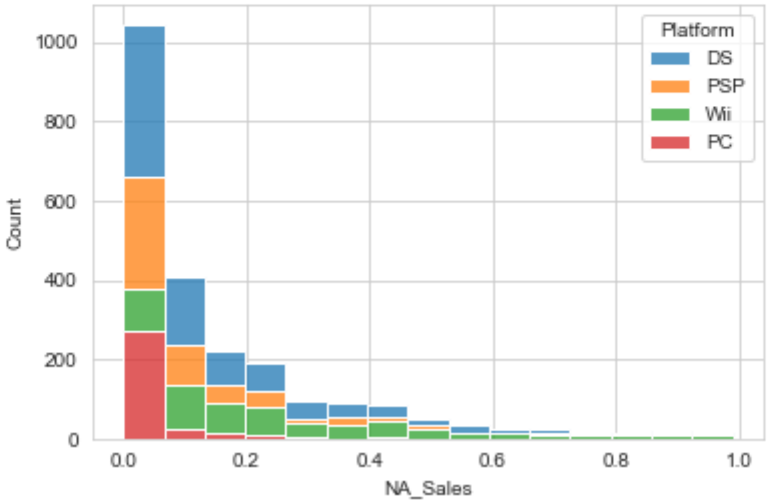
We only kept video games sold for less than 1 Million dollars.



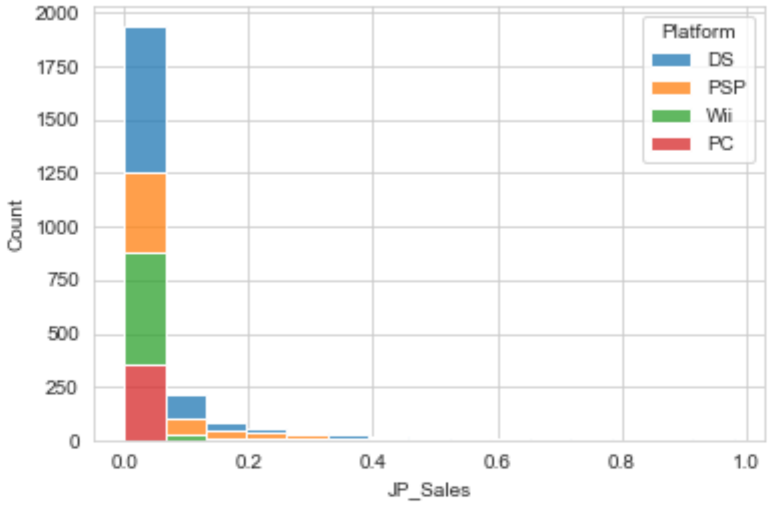
Global Sales:



EU Sales:



NA Sales:



JP Sales:

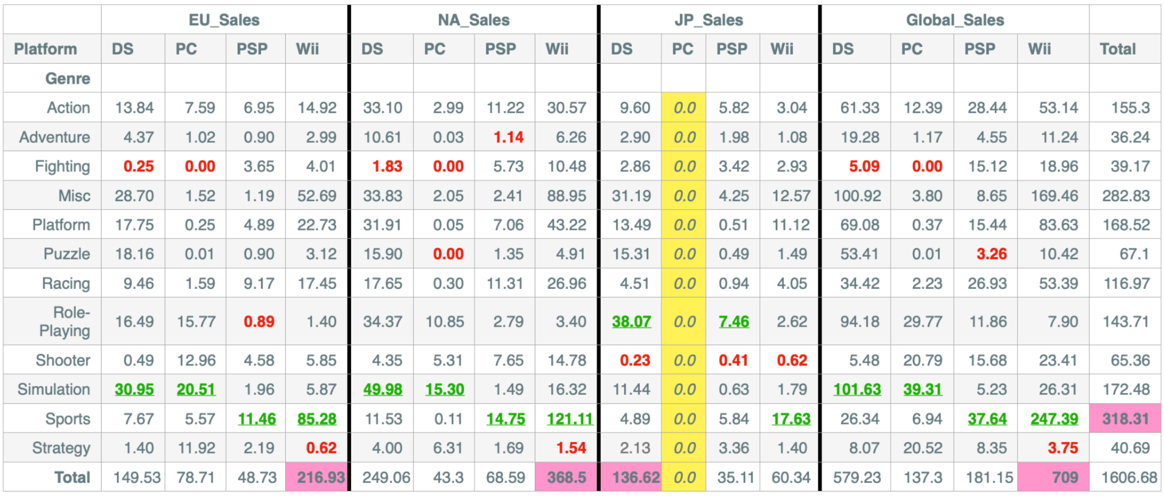
Analysis (in **30-100 words**)

(Hint: What hidden/interesting information you can see that is not easily captured by using the normal histogram/distribution plot?)

- The PSP was more prevalent in 2005-2007. It seems to disappear after 2013.
- The DS was the most popular between 2007 and 2011. After that period, the DS seems to disappear.
- The PC is never trendy, but its popularity has been stable over time.
- For each geographical zone, the DS is the most used platform.

● Pivot the data 15/15

Give a screenshot of a **pivot table** from the data. You are free to choose the attributes involved as long as it helps in answering the data science objective.



	EU_Sales				NA_Sales				JP_Sales				Global_Sales				
Platform	**DS**	**PC**	**PSP**	**Wii**	**DS**	**PC**	**PSP**	**Wii**	**DS**	**PC**	**PSP**	**Wii**	**DS**	**PC**	**PSP**	**Wii**	**Total**
Genre																	
Action	13.84	7.59	6.95	14.92	33.10	2.99	11.22	30.57	9.60	*0.0*	5.82	3.04	61.33	12.39	28.44	53.14	155.3
Adventure	4.37	1.02	0.90	2.99	10.61	0.03	**1.14**	6.26	2.90	*0.0*	1.98	1.08	19.28	1.17	4.55	11.24	36.24
Fighting	**0.25**	**0.00**	3.65	4.01	**1.83**	**0.00**	5.73	10.48	2.86	*0.0*	3.42	2.93	**5.09**	**0.00**	15.12	18.96	39.17
Misc	28.70	1.52	1.19	52.69	33.83	2.05	2.41	88.95	31.19	*0.0*	4.25	12.57	100.92	3.80	8.65	169.46	282.83
Platform	17.75	0.25	4.89	22.73	31.91	0.05	7.06	43.22	13.49	*0.0*	0.51	11.12	69.08	0.37	15.44	83.63	168.52
Puzzle	18.16	0.01	0.90	3.12	15.90	**0.00**	1.35	4.91	15.31	*0.0*	0.49	1.49	53.41	0.01	**3.26**	10.42	67.1
Racing	9.46	1.59	9.17	17.45	17.65	0.30	11.31	26.96	4.51	*0.0*	0.94	4.05	34.42	2.23	26.93	53.39	116.97
Role-Playing	16.49	15.77	**0.89**	1.40	34.37	10.85	2.79	3.40	**38.07**	*0.0*	**7.46**	2.62	94.18	29.77	11.86	7.90	143.71
Shooter	0.49	12.96	4.58	5.85	4.35	5.31	7.65	14.78	**0.23**	*0.0*	**0.41**	**0.62**	5.48	20.79	15.68	23.41	65.36
Simulation	**30.95**	**20.51**	1.96	5.87	**49.98**	**15.30**	1.49	16.32	11.44	*0.0*	0.63	1.79	**101.63**	**39.31**	5.23	26.31	172.48
Sports	7.67	5.57	**11.46**	**85.28**	11.53	0.11	**14.75**	**121.11**	4.89	*0.0*	5.84	**17.63**	26.34	6.94	**37.64**	**247.39**	318.31
Strategy	1.40	11.92	2.19	**0.62**	4.00	6.31	1.69	**1.54**	2.13	*0.0*	3.36	1.40	8.07	20.52	8.35	**3.75**	40.69
Total	149.53	78.71	48.73	216.93	249.06	43.3	68.59	368.5	136.62	*0.0*	35.11	60.34	579.23	137.3	181.15	709	1606.68

The numerical values were aggregated by sum.
The biggest value is in green, underlined, and bold.
The smallest value is in red and bold.

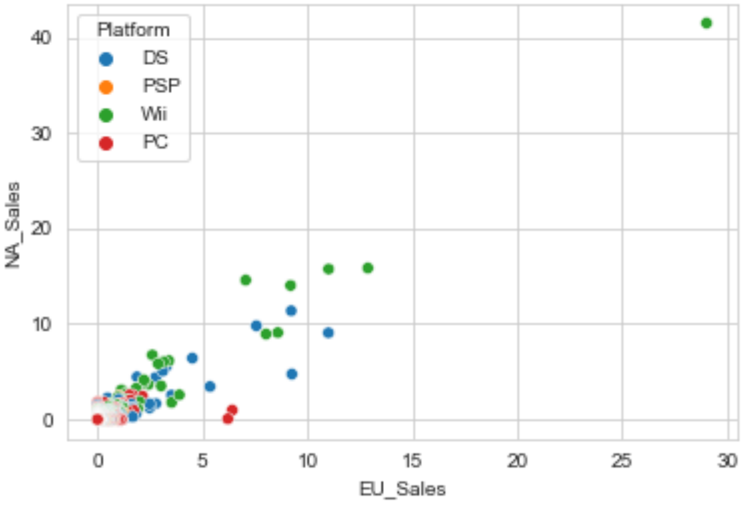
Analysis (in **30-100 words**)

(Hint: What can you say about the information in the aggregated columns that is difficult to obtain by just using visualization?)

- PC games were not sold in Japan in our dataset.
- Europe and North America have similar tastes regarding video games (Genre and Platform)
- Europe and North America represent most of the sales, so the global sales also have the same trends as those two geographical zones.
- Japan prefers Role-Playing games, whereas Europe and North America prefer Simulation and Sport.
- The least popular game genre in Japan is Shooter, whereas, in Europe and North America, it is Fighting. However, Fighting is not very popular in Japan as well.
- The Wii seems to be a platform known for Sports Games.
- Apart from Japan, the Wii is the preferred platform. In Japan, the DS is where people spend more money buying games. Japanese people tend to prefer portative consoles instead of home consoles.
- Overall, Sports game is the favorite genre all over the world.

● Watch out for outliers 10/10

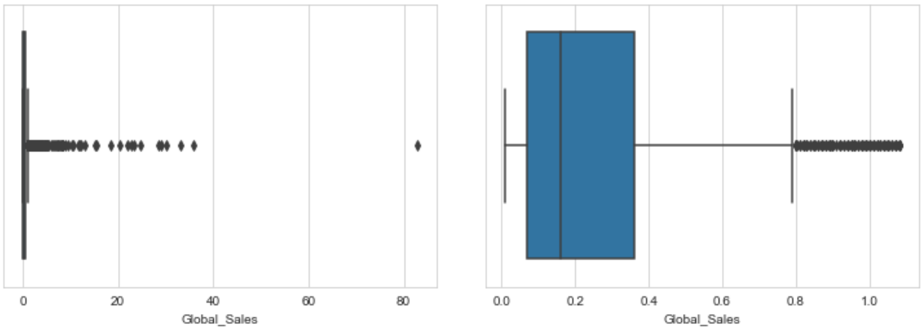
Give a screenshot of a **scatterplot** between any of the numerical attributes or **box plots** of some of the attributes. You can color differently the data points that you suspected to be the outlier.



Scatterplot of North American sales and European sales by different platforms

Box Plot with all data

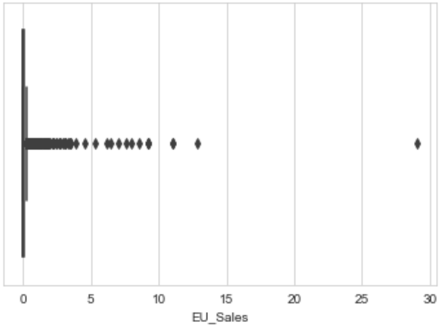
Box Plot with only data between Q1 and Q3

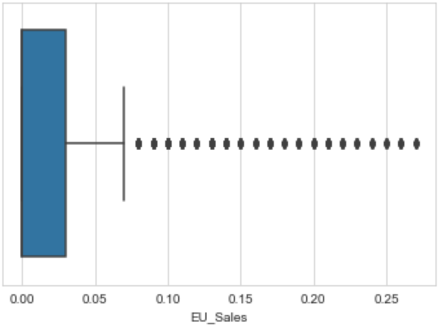


Global Sales

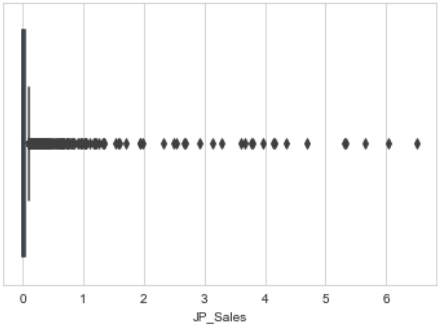
Box Plot with all data

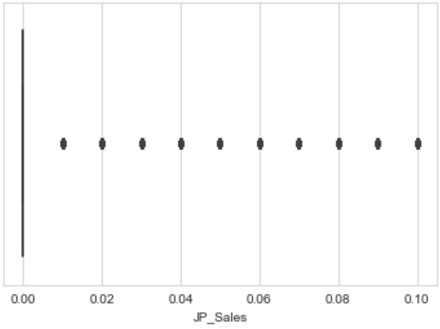
Box Plot with only data between Q1 and Q3



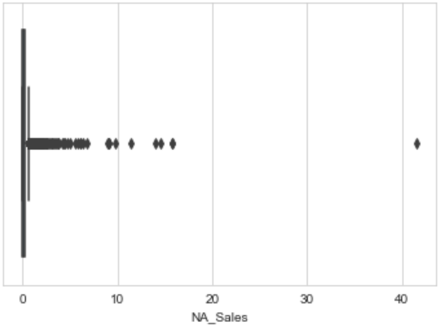


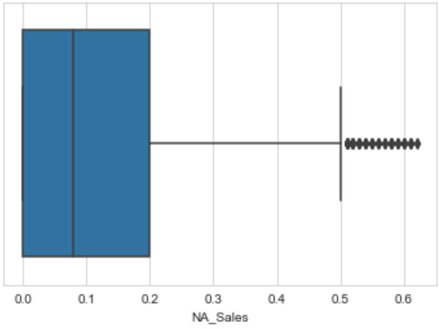
EU Sales



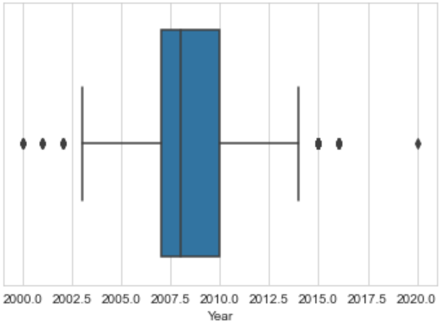


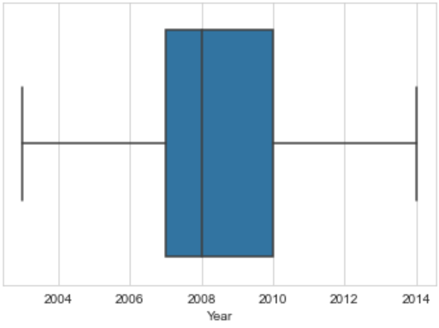
JP Sales





NA Sales





Year

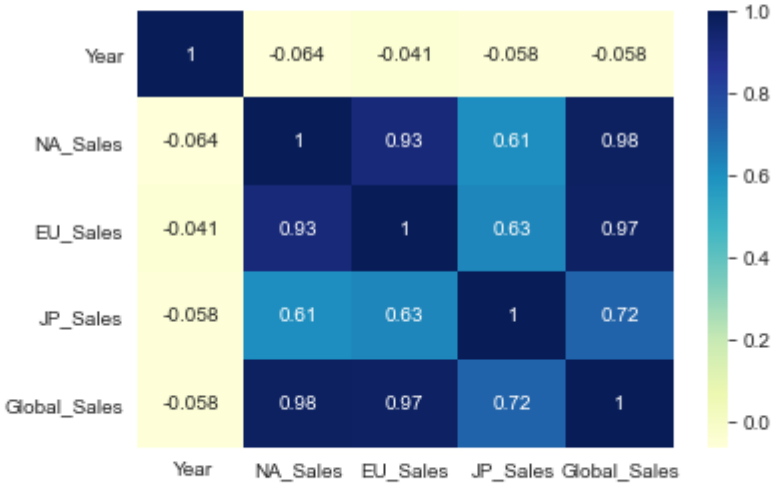
Analysis (in **30-100 words**)

(Hint: What can you say about the existence of outliers by just looking at the scatterplot or box plots?)

We can see some outliers in the sales columns again. We can also observe that video games were bought between 2003 and 2014. The rest is considered to be outliers. With the first scatterplot, we can see a there seems to be a linear relation between sales in Europe and North America. We can also see that one Wii video game has especially been sold a lot of times.

● Understand the relationship between attributes

Give a screenshot of a **correlation matrix** of the numerical attributes.



Analysis (in **30-100 words**)

(Hint: What can you say about attributes that are correlated/dependent on each other? If can, investigate why they are dependent.)

This correlation matrix validates our previous hypotheses (cf parts “Watch out for outliers” and “Pivot the data”). We can see a correlation between

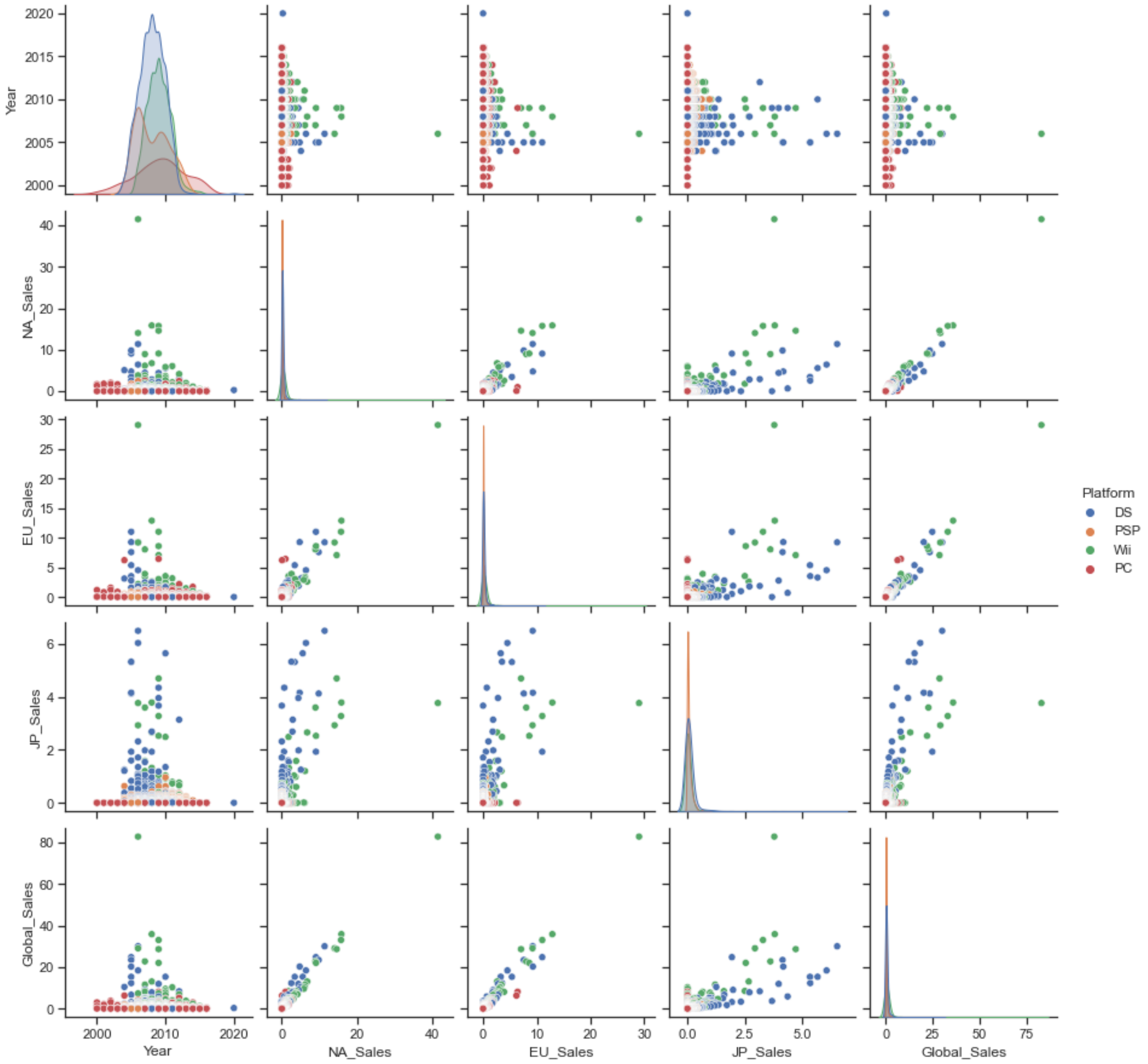
- NA_Sales and EU_Sales
- NA_Sales and Global_Sales
- EU_Sales and Global_Sales

NA_Sales and EU_Sales are probably correlated due to cultural reasons (Western culture). Global_Sales is the sum of the EU_Sales, NA_Sales, JP_Sales, and the other sales. As the European Union and North America represent an absolute majority of sales, it is not surprising to observe the same trends between the European Union, North America, and Global Sales.

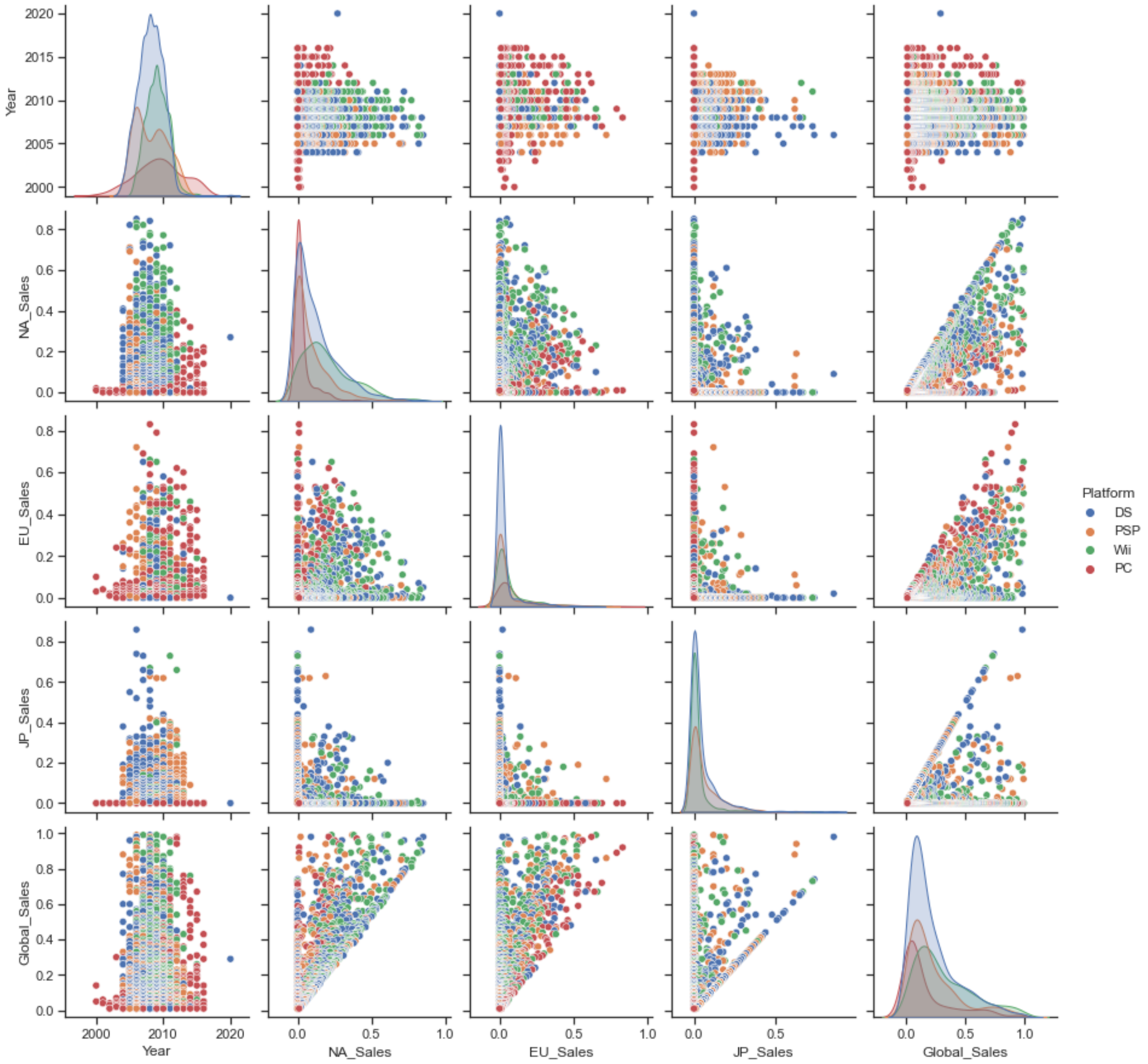
● Visualize the relationship between attributes 15/15

Give a screenshot of a **scatter matrix** (color-coded by any categorical attribute)

Visualisation of all the data (2456 elements)



Visualisation of sales less than 1 Million (2171 elements)



Analysis (in **30-100 words**)

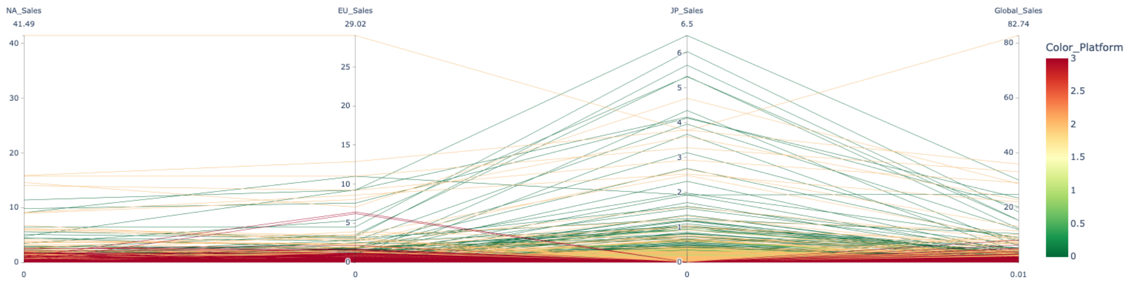
(Hint: What other information can you get from the scatter matrix that you haven't figured out before? You can also conclude most of the interesting findings so far that can be seen in the given scatter matrix.)

- Once again, we can see a linear correlation between NA_Sales, EU_Sales, and Global_Sales.
- As the Global_Sales is the sum of NA_Sales, EU_Sales, JP_Sales, and Other_Sales, the values can only be bigger. That is why all the values are between the lines x = y and x = 0 (or symmetrical).
- When JP_Sales is compared with EU_Sales, we can see two distinct lines: x= 0 and y=0. This implies that video games sold a lot in Japan are not that sold in Europe. Same observations between JP_Sales and NA_Sales, but it is a bit less flagrant (probably due to the amount of data)
- Based on the scatter matrix, we cannot identify Platform clusters.

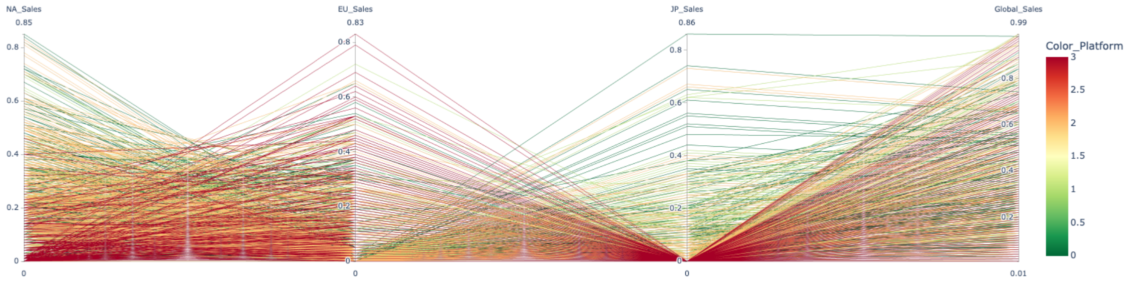
● Visualize high-dimensional datasets 10/10

Give a screenshot of a **deviation chart**, **parallel charts,** or **parallel coordinates** for the numerical attributes. You can color-code it using any categorical column.

Visualisation of all the data (2456 elements)



Visualisation of sales less than 1 Million (2171 elements)



Color corresponds to the following genre:

"PC": 3 → Red

"Wii": 2 → Orange

"PSP": 1 → Light Green

"DS": 0 → Dark Green

Analysis (in **30-100 words**)

(Hint: What can you say about the similarity of the data points? Do some attributes separate the classes in the label nicely?)

Data with outliers:

- There is no clear separation between the different Platforms.
- DS and PSP are more sold in Japan. They are indeed Japanese brands.
- Western people tend to buy more home console games (Wii, PC) whereas Japanese people tend to buy more portative console games (PSP, DS).
- We can clearly see one outlier in NA_Sales and in EU_Sales which belongs to the Wii category.

Data without outliers:

- Europeans are the ones most attracted to PC games.
- North Americans are the ones most attracted to Wii games. But they also used a lot of DS and PSP.