Video Game Sales

Rahul Bharadwaj Mysore Venkatesh

23/08/2020

Introduction -

Having played tons of games on PC and Playstation2 as a kid, it was a moment of nostalgia as I found a Kaggle dataset on Video Game Sales. I couldn't wait to get a hands on overview of this data which has over 15000 observations and sales for most of the games on all platforms. Note that this data is all about the number of copies sold and not about the revenue generated through sales. After all, its not always about money and this is an effort to analyze the popularity of games through number of copies sold! All values are in millions of copies sold. Let's dive in and check out which game, publisher, and platforms were most popular among gaming fans!

- We first load the libraries required for our analysis.

```r
#loading libraries
library(tidyverse)
library(visdat)
library(kableExtra)
library(ggpubr)
```

- We now read our data into R environment from a source file
- The raw data in the form of csv looks as follows.

```r
#reading video games sales data from csv file
vgsales <- read.csv("vgsales.csv")
#displaying dimensions of data and sample observations
glimpse(vgsales)
```

```
## Rows: 16,598
## Columns: 11
## $ Rank         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,...
## $ Name         <chr> "Wii Sports", "Super Mario Bros.", "Mario Kart Wii", "...
## $ Platform     <chr> "Wii", "NES", "Wii", "Wii", "GB", "GB", "DS", "Wii", "...
## $ Year         <chr> "2006", "1985", "2008", "2009", "1996", "1989", "2006"...
## $ Genre        <chr> "Sports", "Platform", "Racing", "Sports", "Role-Playin...
## $ Publisher    <chr> "Nintendo", "Nintendo", "Nintendo", "Nintendo", "Ninte...
## $ NA_Sales     <dbl> 41.49, 29.08, 15.85, 15.75, 11.27, 23.20, 11.38, 14.03...
## $ EU_Sales     <dbl> 29.02, 3.58, 12.88, 11.01, 8.89, 2.26, 9.23, 9.20, 7.0...
## $ JP_Sales     <dbl> 3.77, 6.81, 3.79, 3.28, 10.22, 4.22, 6.50, 2.93, 4.70,...
## $ Other_Sales  <dbl> 8.46, 0.77, 3.31, 2.96, 1.00, 0.58, 2.90, 2.85, 2.26, ...
## $ Global_Sales <dbl> 82.74, 40.24, 35.82, 33.00, 31.37, 30.26, 30.01, 29.02...
```

Initial Data Analysis -

- Initial Data Analysis is a process which helps one get a feel of the data in question. This helps us have an overview of the data and gives insights about potential Exlporatory Data Analyis (EDA).
- Initial data analysis is the process of data inspection steps to be carried out after the research plan and data collection have been finished but before formal statistical analyses. The purpose is to minimize the risk of incorrect or misleading results. Link for more info
- IDA can be divided into 3 main steps:
  - Data cleaning is the identification of inconsistencies in the data and the resolution of any such issues.
  - Data screening is the description of the data properties.
  - Documentation and reporting preserve the information for the later statistical analysis and models.

visdat

1. The visdat package in R helps us get a visual overview of the data in the form of plots. The vis_dat() function helps us get a glimpse of the data types for all variables in our dataset.

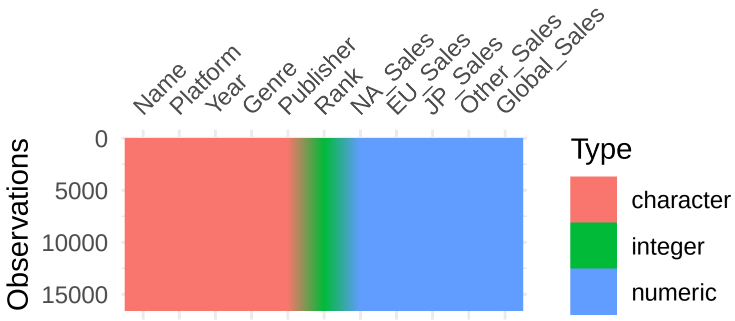


Figure 1: Visulaization of Data Types of the data

- We can observe that there are only three Types of data in our dataset viz, character, integer, and numeric. This makes it pretty straightforward and simple to conduct analysis.

visguess

2. The vis_guess() function tries to predict the kind of data in each cell of our dataset.

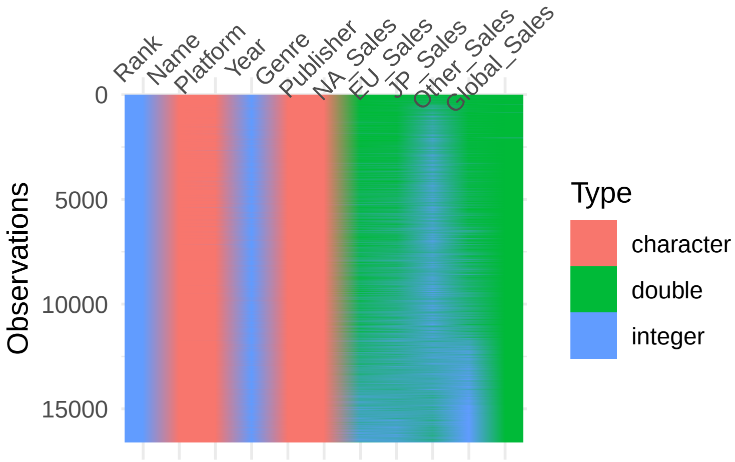


Figure 2: Data Type for each cell in dataset

- We can thus oobserve the following from the dataset.
  - Rank is integer Type.
  - Name is character Type.
  - Platform is character Type with an exception of one cell value which might have an integer.
  - Year is integer Type with some excpetion that might look like character.
  - Genre and Publisher are character Type.
  - The rest of the sales variables are either integer or double.

- Note that this is a cell-wise interpretation and the actual data will have only one type for one column.

vismiss

3. The vis_miss() function give an overall visual of the missing data in our data set.

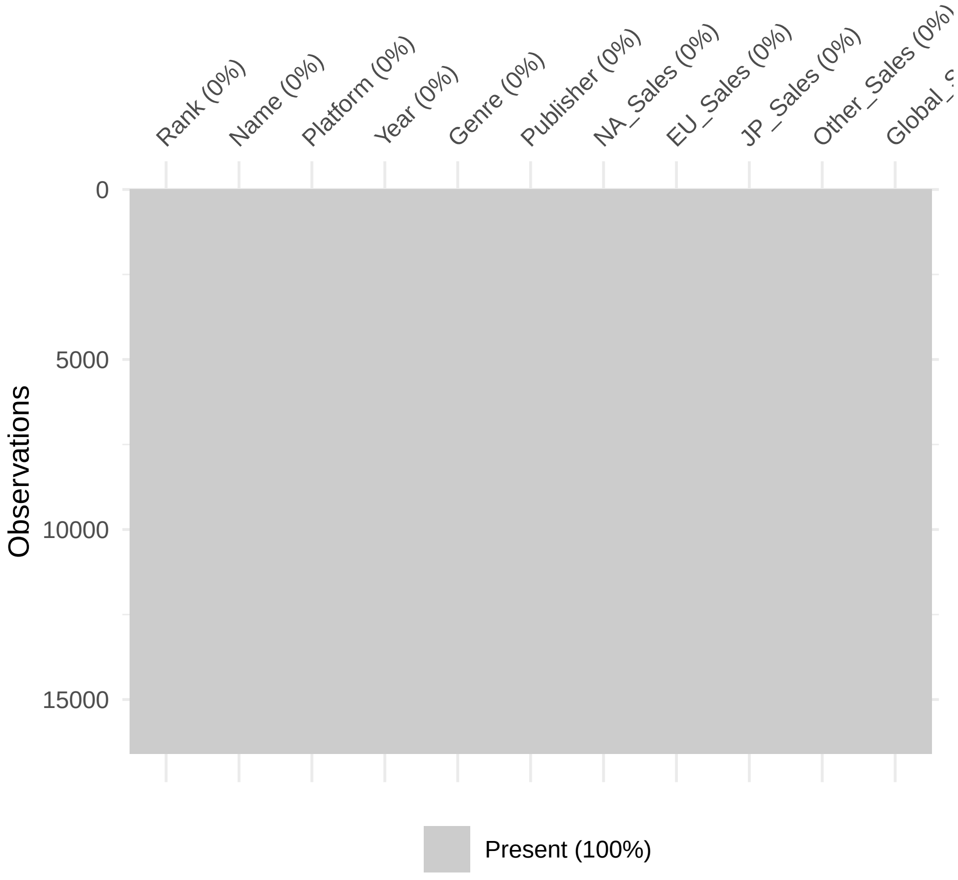


Figure 3: Missing Values Plot

- We can observe that there is no missing values in our data. Present 100% indicates the same. This means we do not have to deal with missing values.

viscor

4. The vis_cor() function gives us a visual plot of the correlation between variables in our dataset. An important thing to note here is that it takes only numeric variables. We have already established this, thanks to vis_dat(). So we select only the numeric columns for this function.

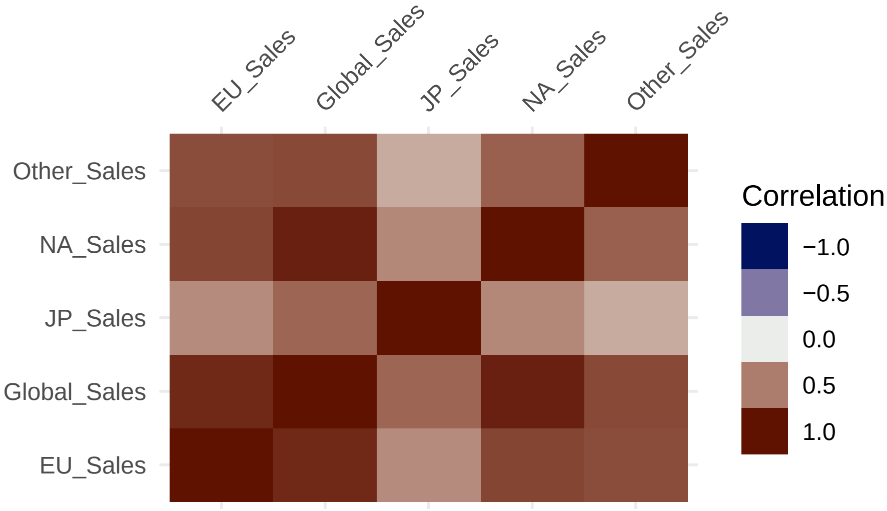


Figure 4: Correlation Plot for variables

- The above figure shows correlation as a range between +1.0 and -1.0 for the different variables.

gathercor

5. The gather_cor() function gives us the exact values for the same instead of a range.

row_1	row_2	value
Global_Sales	NA_Sales	0.9410474
Global_Sales	EU_Sales	0.9028358
Global_Sales	JP_Sales	0.6118155
Global_Sales	Other_Sales	0.7483308
Global_Sales	Global_Sales	1.0000000

- The above table shows the exact correlations between the variables in our dataset. This sample table is filtered to show the correlation between Global Sales and Sales in each major locations of the world.

Exploratory Data Analysis -

- In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task (Wikipedia)

Regional Sales -

- As the data has regional sales, this was the first question I could think of. What regions drove most of the global sales?
- Let us consider the Correlation between sales in different regions and global sales as mentioned in the previous table. We plot with Regional Sales on x-axis and Global Sales on y-axis and check how each regional sales drive global sales.

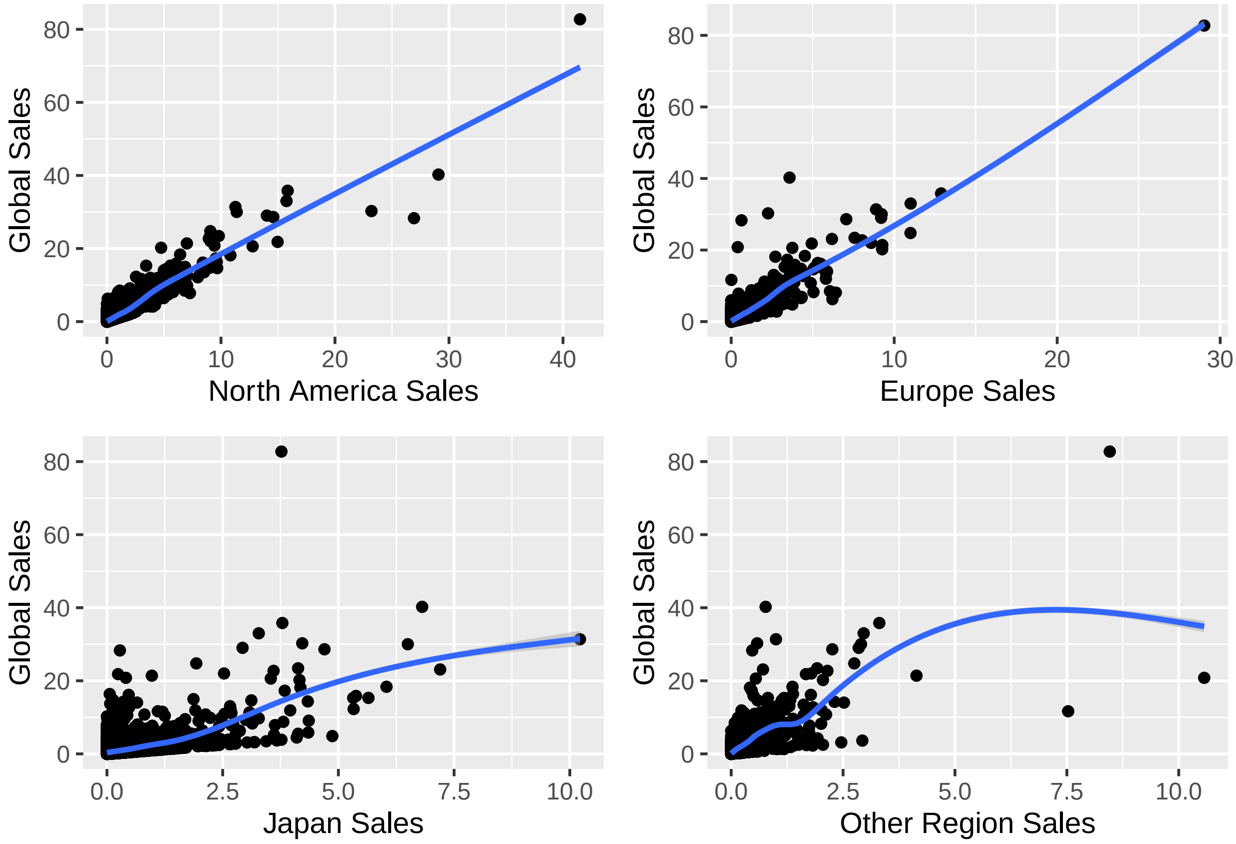


Figure 5: Smoothened curves for Regional Sales and Global Sales

- The above plot is a smoothened curve of all the observations and how the tend to change.

- The following correlation displays the above table as a plot with all the exact values.

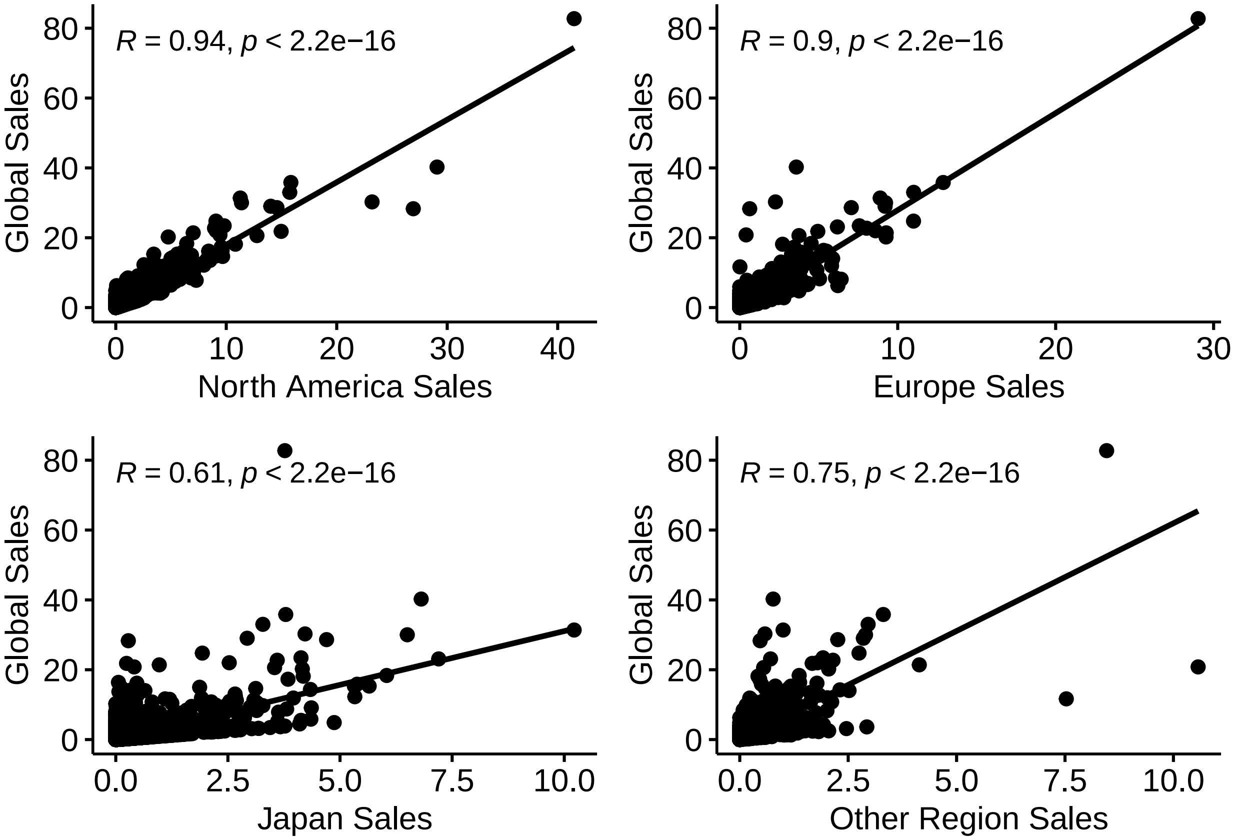


Figure 6: A plot of Correlation between Regional Sales and Global Sales

- We can observe that most of the global sales share is from North America and Europe. Other regions of the world has a lesser share and Japan has the least share of global sales.

Famous Gaming Franchises -

- It's time to compare some of my favorite franchises from my childhood! Which franchise was more popular and on which platform?
- Let us consider some of the most famous game franchises and check what franchise sold most on what gaming platform. **The considered franchises are Grand Theft Auto, FIFA, Pokemon, Mario, Need for Speed, and Call of Duty.** These are the ones I could think of instantaneously.

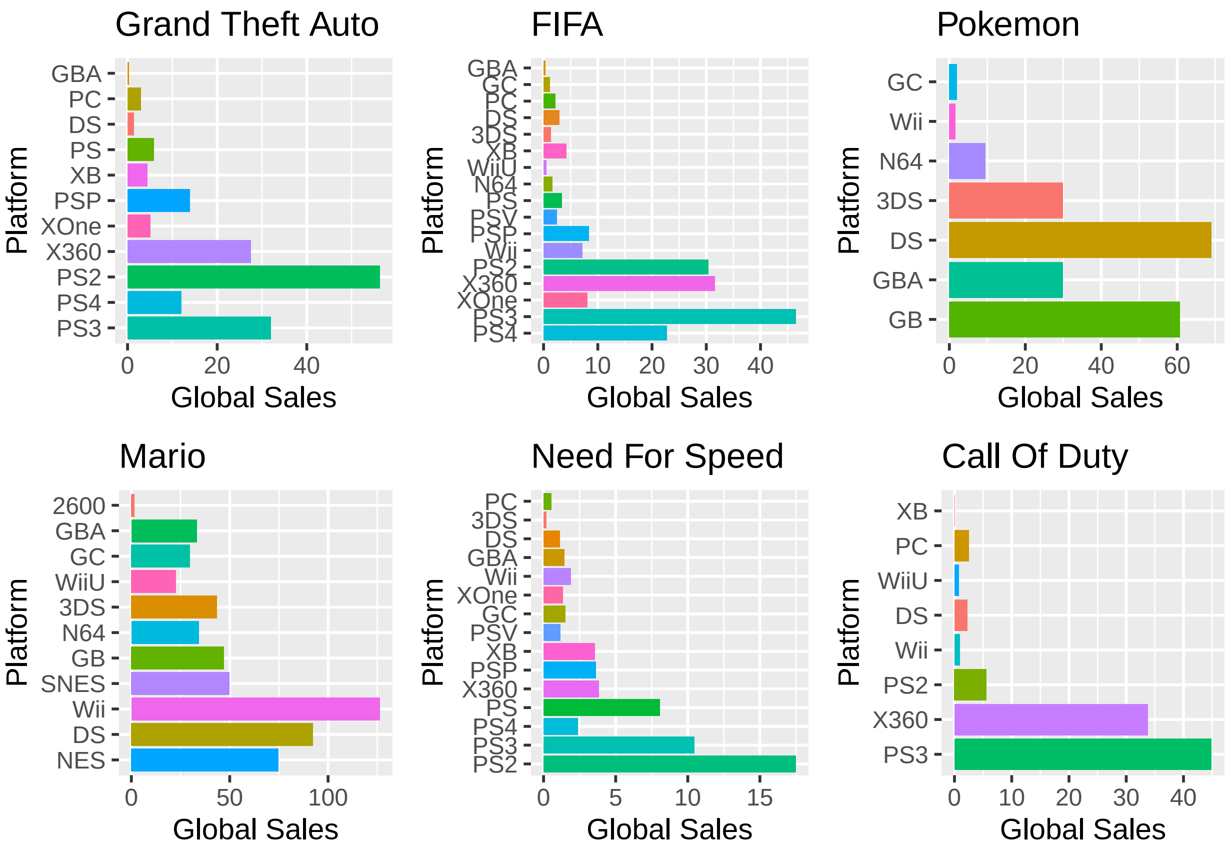


Figure 7: All Time Global Sales (millions)

- The above barplot displays global sales on x-axis and platform of game on y-axis. We can get an overview of which franchise of games was sold on what platforms and also about the number of copies sold in millions. This can be compared side-by-side and it clearly shows which game was most popular on what platform.
- Notice that the scales are different for each franchise and **Mario has sold the most with more than 100 million copies** while **Need For Speed has sold a little over 15 million only.**
- In terms of revenue and profits generated, this might not hold good. A cheaper videogame sold in more numbers might still generate less revenue than an expensive videogame sold in lesser numbers. Something to keep in mind! Also the cost of the games definitely affect the number of copies sold.

UbiSoft Franchises -

- The following graph shows a dot plot of two popular game franchises of UbiSoft namely, **Assassin's Creed and Prince of Persia**. It displays global sales on x-axis and game title on the y-axis. The different colors of dots account to the type of platform the game belonged to.
- Prince of Persia (PoP) was one of my most favorite gaming franchise as a kid and I remember having played it for days together, endlessly, during my vacations. There was a time when the PoP games saw a decline over Assassin's Creed. I wanted to know why! And thus, this comparison.

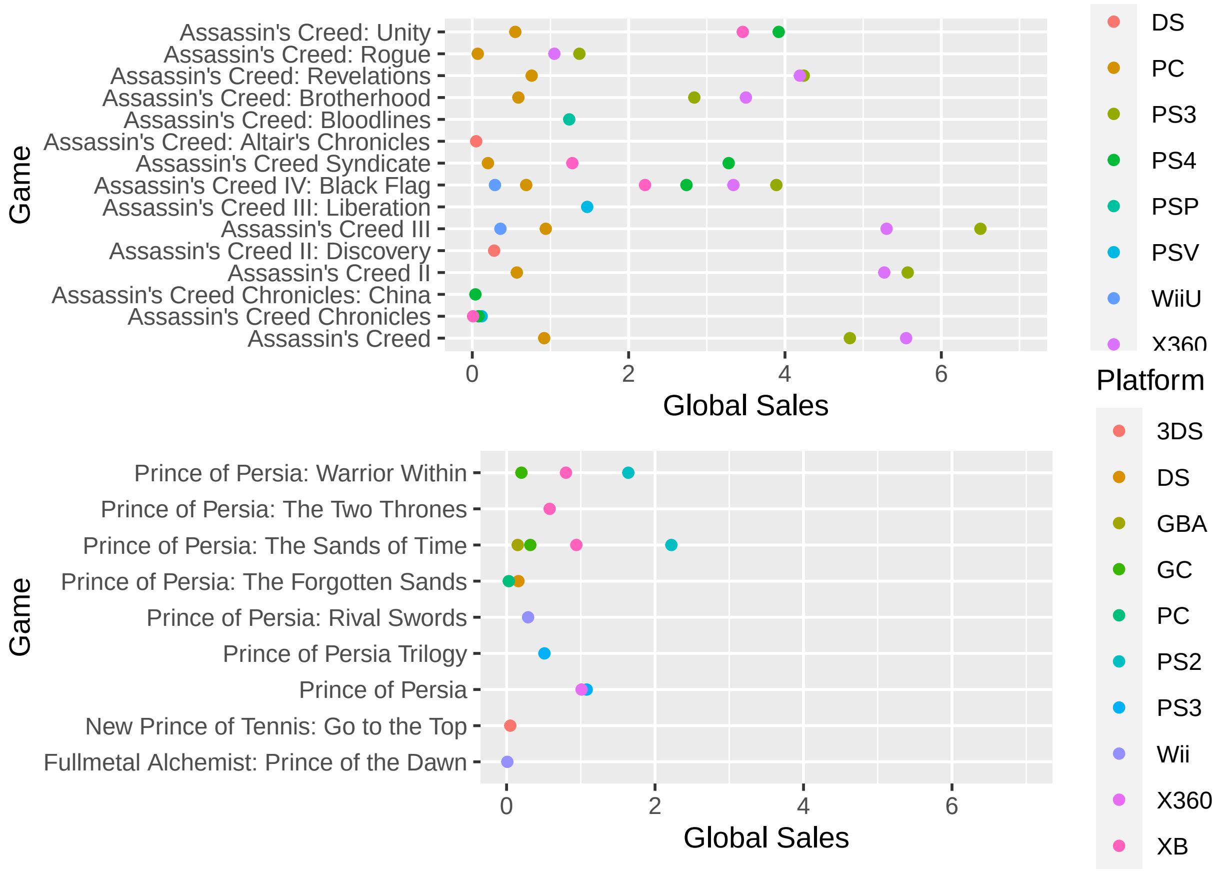


Figure 8: Assassin's Creed and Prince of Persia Sales (millions)

- We can observe that there were **lesser PoP titles released compared to AC titles**. Also, only two games from the PoP franchise crossed sales of 1.5 million. On the other hand, AC has sold more than 2 million copies of most of their titles.
- This might explain why there are lesser games in PoP franchise compared to AC. UbiSoft slowly shifted their focus on the franchise that was more popular among gaming fans.
- Turns out after all, it's all about the money for the game producers… Fair enough from a business point of view. But I would've loved to have more PoP releases!

Publisher Performances -

- Next, I wanted to know how each Publisher performed in terms of number of copies sold. Which publisher sold most copies? This analysis can be extended to any Publisher on the dataset.
- The following table shows a summary of the **EA games Global Sales.**

TotalTitles	TotalCopiesSold	MeanGS	MedianGS	MaxSold	MinSold
1353	1110.74	0.820946	0.48	8.49	0.01

- The following table shows a summary of the **Activision games Global Sales.**

TotalTitles	TotalCopiesSold	MeanGS	MedianGS	MaxSold	MinSold
1005	734.9	0.7312438	0.28	14.76	0.01

- The following table shows a summary of the **UbiSoft games Global Sales.**

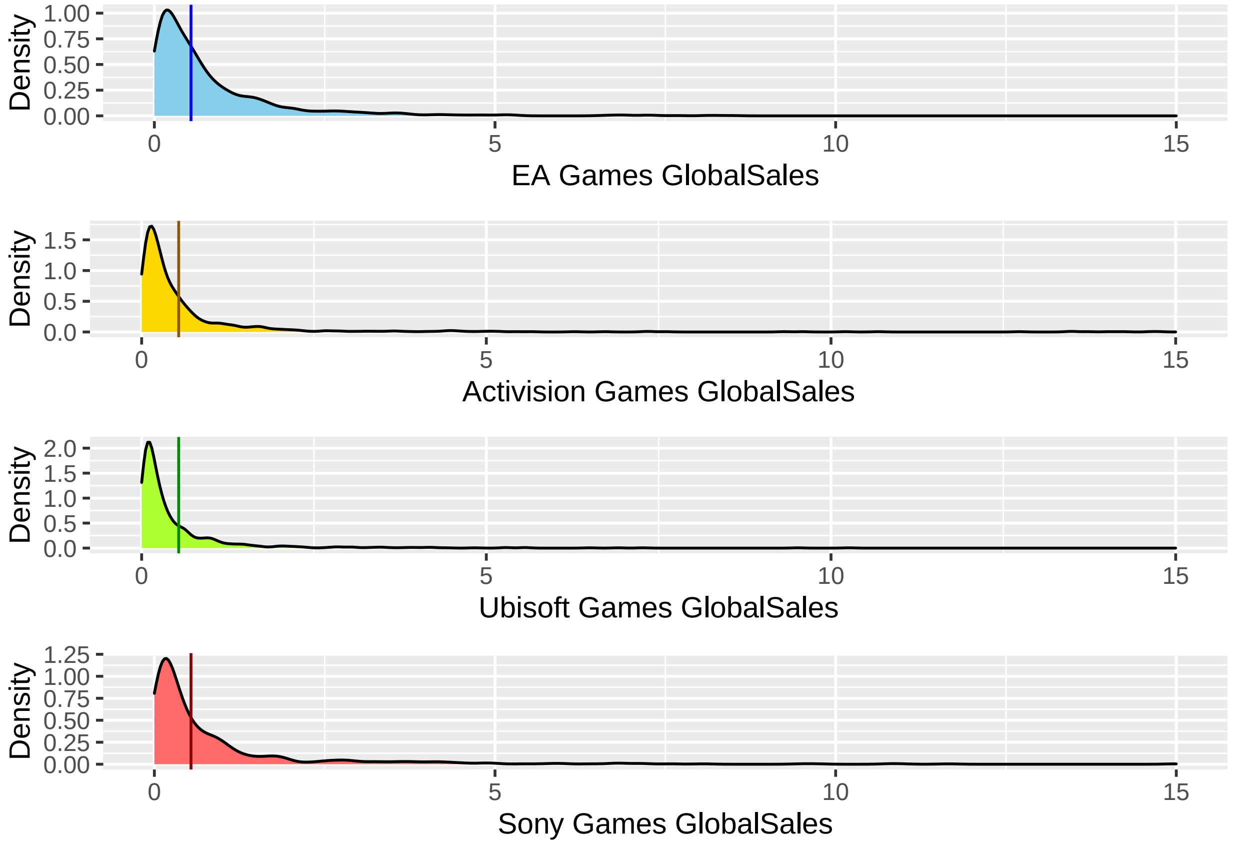
TotalTitles	TotalCopiesSold	MeanGS	MedianGS	MaxSold	MinSold
935	479.18	0.512492	0.21	10.26	0.01

- The following table shows a summary of the **Sony games Global Sales.**

TotalTitles	TotalCopiesSold	MeanGS	MedianGS	MaxSold	MinSold
701	632.57	0.9023823	0.34	14.98	0.01

Publisher Global Sales Comparison -

- This density plot helps us get an understanding as to which publisher sold how many copies.



- Sony seems to have the best mean sales followed by EA. Activision and UbiSoft have almost the same mean sales. Sony having sold the least number of titles still have an upper hand in the number of copies sold.
- With different questions asked, the answer is different. And it is debatable as to who was best.

Conclusion -

- This is just the start of answering questions about the data. There can be numerous questions asked and appropriate analysis conducted with suitable visual representations that can effectively answer the questions.
- Dig deep and find out answers to the most burning questions you have in mind, if you're a gaming enthusiast.

References -

- IDA reference
- Correlation Plots Reference
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Tierney N (2017). "visdat: Visualising Whole Data Frames." *JOSS*, *2*(16), 355. doi: 10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL: http://dx.doi.org/10.21105/joss.00355>.
- Hao Zhu (2019). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra
- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpub