Project Reflections

Nirav Bariya

Date: 08/02/2024

GamesCo, a video game company, wants to make data driven decisions for their future development of games. We are asked to do a descriptive analysis of the historical sales figures from North America, Europe, Japan, Others, and Global. The company's executive teams want the answer to the following questions:

- Are certain types of games more popular than others?
- What other publishers will likely be the main competitors in certain markets?
- How have their sales figures varied between geographic regions over time?

From my analysis, I believe, the Vice President of Marketing will get the answer about which genre is performing the best; Chief Financial Officer will get the answer about which publishers are our top competitors; Senior Vice President of the Sales will know how market share of different regions has been changed.

Before doing aggregation, summarization, or analysis, it was important to have a look at the data and understanding what kind of information do I have. This helped me set up my mind on what type information am I working with and be prepared. Further, doing this process made it clear what type of questions will I be able to answer and which questions will not be answered by this analysis. In other words, I discovered the limitations of the dataset. Moreover, to check for biases I explored the source of the dataset as it was important to look for any biases present in the dataset as biases often yields misleading conclusions.

Once, I have looked at the data, it was imperative to find out information that was missing, incomplete, misplaced, or not formatted correctly. Rectifying any errors related to the data is conducive to an accurate analysis.During the cleaning process these were my observations:

- there were some rows which have missing characteristics such as Name, Platform, Publisher, EU_sales, JP_sales, Other_sales, Global_sales
- Milestone S.r.l. has different ways of writing, multiple spelling, I figured the correct spelling and corrected it
- there was one redundant column called NA units with all entries being 0, I deleted this column as it was not contributing anything
- The Year, Publisher, Platform, name, and Genre column did have blank cells, which I filled with NA to indicate that the values were not available or unknown
- There was 1 duplicate entry found which I removed to avoid error in aggregation
- The columns of NA sales, JP sales, EU sales, Other sales, and Global sales with blank cells were imputed with mean values t
- there was 1 completely blank row which was deleted

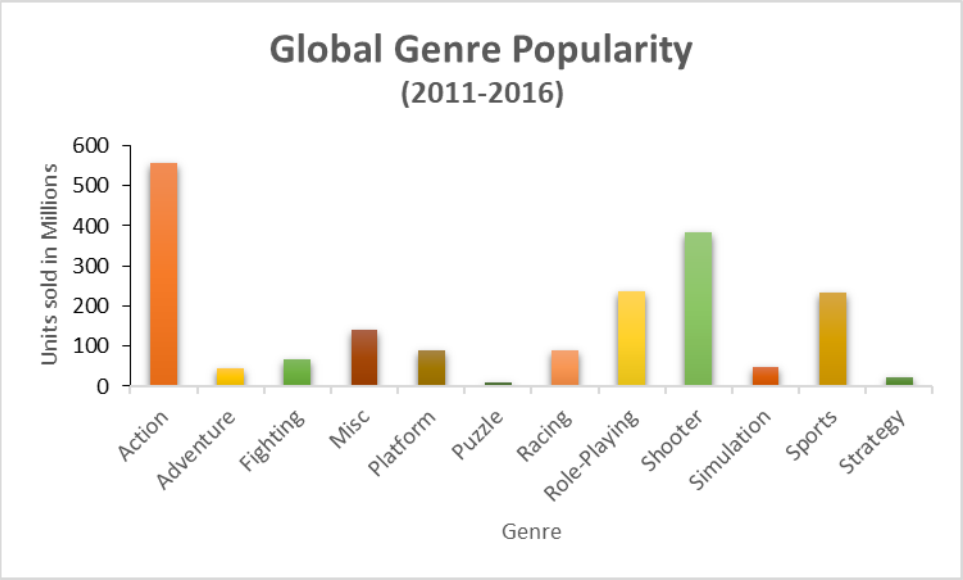
After the cleaning process, I tried to get more out of the dataset by summarizing and aggregating the values. This gave me a better understanding of the data at hand. When we look at the central tendency of sales in all the region, we get the following observation:

	North America	**Europe**	**Japan**	**Other**	**Global**
Mean	0.267	0.146705	0.0777977	0.049521907	0.53762054
Median	0.08	0.02	0	0.01	0.17
Observation	Mean > Median	Mean > Median	Mean > Median	Mean > Median	Mean > Median
Distribution	Right-Skewed	Right-Skewed	Right-Skewed	Right-Skewed	Right-Skewed
3rd Quartile	0.24	0.11	0.04	0.04	0.42

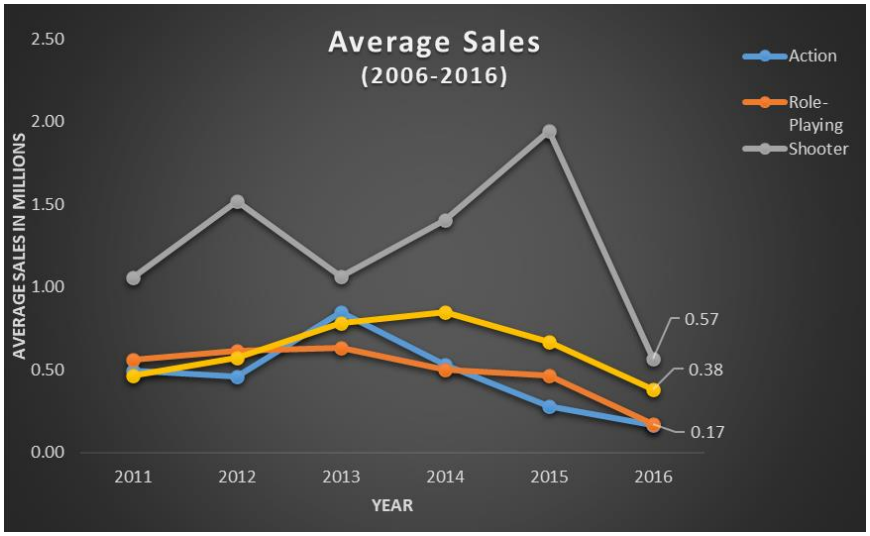
What this table suggest is that for all the regions, the outliers were pulling the mean up, making it larger than the median. And, hence the distribution is right-skewed. As we see from the table, median in all the cases were less than 0.1 million. Globally speaking, 75% of the titles seen a sale of less than 420,000 units. Moreover, any sale greater than 1.1 million was an outlier, globally.

To find out which genres are popular, I grouped the data according to genre by creating a pivot table and looked at the total global sales. I considered the sales data for between 2011-2016 to capture the latest trend.

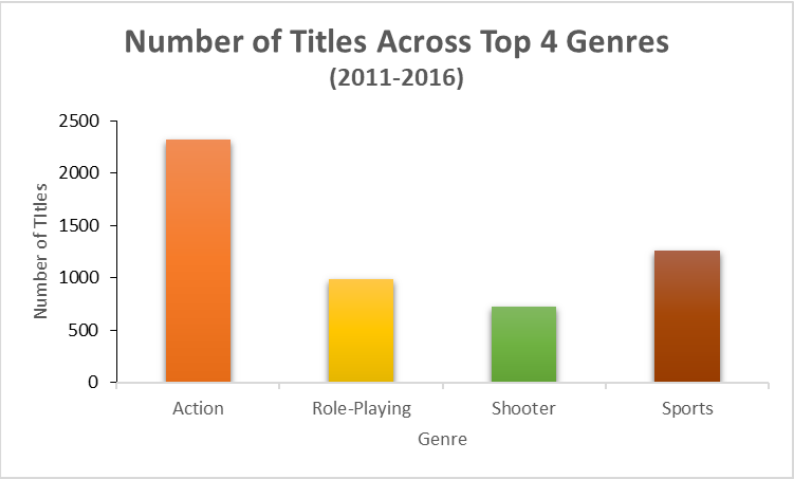
This led me to answer the first question the stakeholders wanted the answer to: **the most popular genre** in **Globally** was **Action** with a total sale of 555.85 million copies sold, the **second was Shooter** genre with total sale of 385.39 million copies sold, the **least favourite genre** was **Puzzle** with 10.06 million copies sold, and the **second last was Strategy** with a sale of 22.03 million copies sold. The following chart was used to gain this insight and show the comparison of genres:



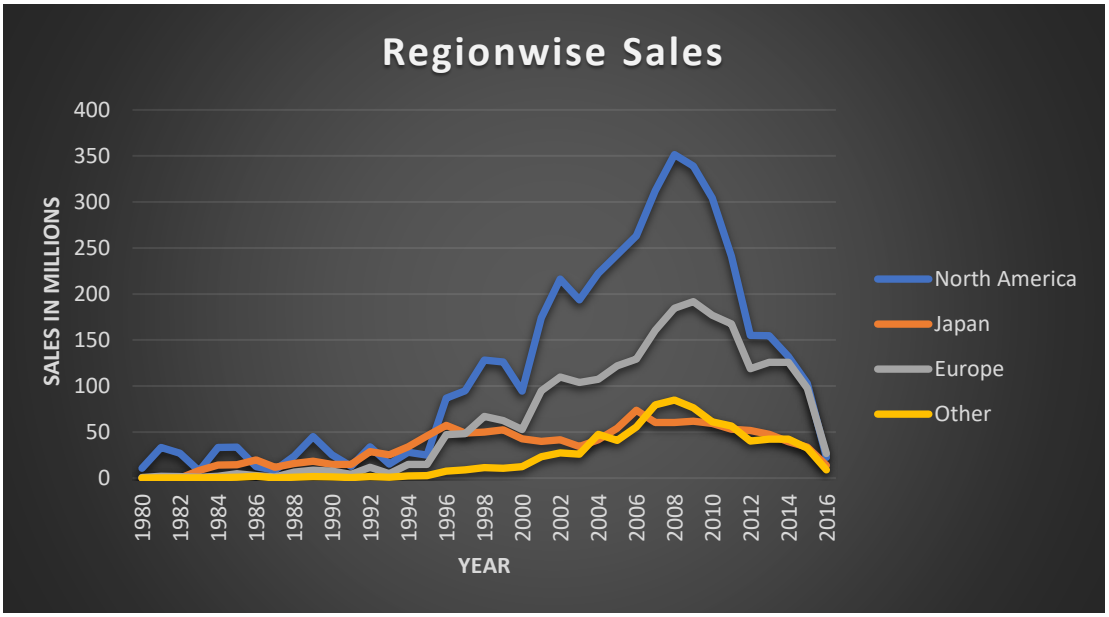
Further when we observe the average sales for the top performing genres like Action, Shooter, Role-playing, and Sports, we see that average sales of shooter were highest with **more than half a million** units and **average sales for Sports** were second highest with approximately 380 thousand units in 2016. Surprisingly, average sales for action genre were least among the top for genres with about 170 thousand units.



The average sales for action, sports, and role-playing genre decreased overtime. Whereas for Shooter genre it remained above half a million mark. This made me wonder why sales of Action genre was so much higher than other top 4 genres. I imagined that the sales were higher due to number of action games produced. This is confirmed by the following graph for number of titles released in last 5 years

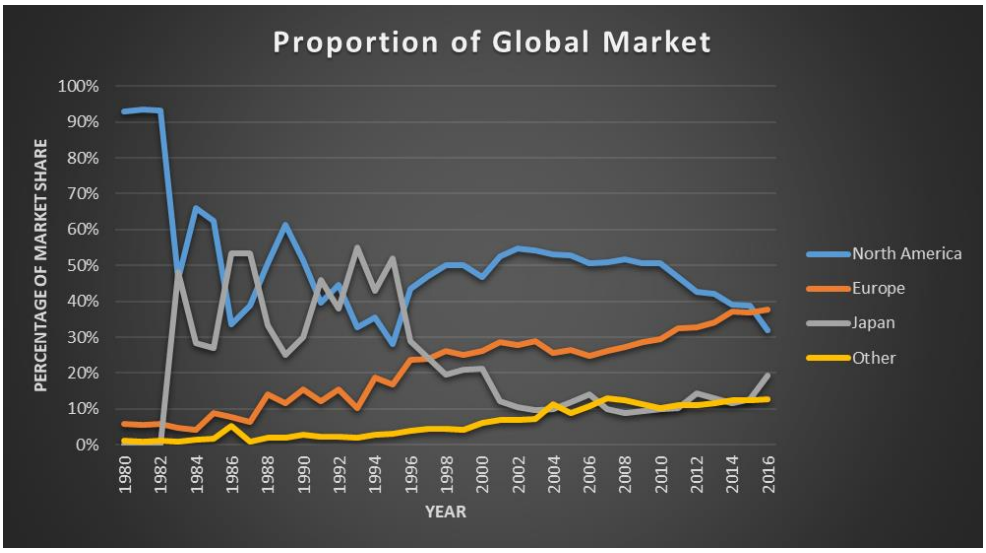


Next, I focused on market share of each geographical region over the time and noticed the changes. First, I looked at the sales figure for each of the regions and plotted a line graph to get a better understanding of the data.

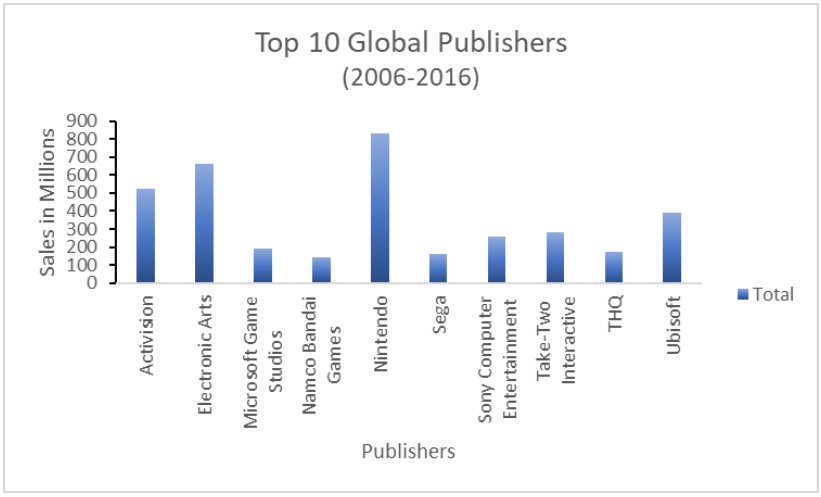


This graph shows and increase in sales for all regions till year 2008-2009. Then sales in all regions fell-dramatically in North America and Europe. There may be multitude of reasons why this could happen. One of them could be the increasing use of social media platforms for entertainment purposes, such as YouTube, Facebook, Instagram, and Tik-Tok. The other could be an implicit bias in the sales figures published by the VGchartz. May be the sales figures were not accurate enough because of increase in software sales across multiple platforms.

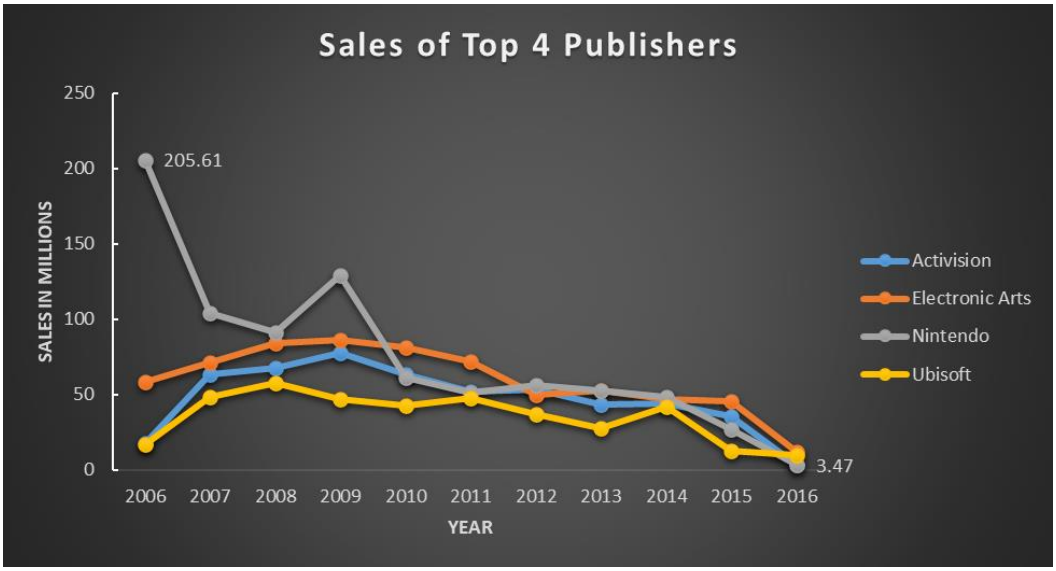
Then I looked at the market share of each of the regions under investigation. Upon plotting the line-graph, I found that sales in North American region had dropped from 93% to 32% whereas sales in Japan and Europe had increased. Moreover, present market share (2016) of Europe (38%) and North America (32%) is closer than ever in history.



It would also be helpful if we consider the competition that GameCo is facing. This will allow for potential opportunities for growth. For that, I had created a bar graph highlighting the top 10 competitors on a global level.



From the graph it becomes clear, in the past 10 years, Nintendo was the highest selling publisher, followed by Electronic Arts, followed by Activision, followed by Ubisoft. These 4 markets had the maximum sales. And, if we further track their sales record over the 10 years, we have the following line graph:



This line graph shows the sales for each of them fell considerably. The **surprising aspect of this graph is that the highest selling publisher became the least selling publisher** in 10 years span from 205.61 million sales to just 3.67 million sales. The total sales figure for **Nintendo** was **830.96** million between 2006-2016, out of which more than half the sales (474.39 million) came from three **genre namely, Sports, Platform, and Role-Playing**. As we have seen earlier, Sports and Role-Playing were among the top 4 genre by sales. This is an opportunity which GameCo can capitalize on.

Creating summaries, groupings and visualizations helped me draw a picture of the data that wasn't there in my mind before. It helped me develop a better understanding of what the data is trying to tell and how it went through changes over a period of time.

Answer 4. As shown earlier, I will use **line chart** to show trends over a period of time as line charts are **most suitable for trends and telling a story**. All the **charts are recreated**. It will help the stakeholders realise how market share for different regions has changed over the years.

The recreated bar charts will be different from the one that I created as it will contain data for specific time period and will be for global market.