Project Reflections

Tara Perrige

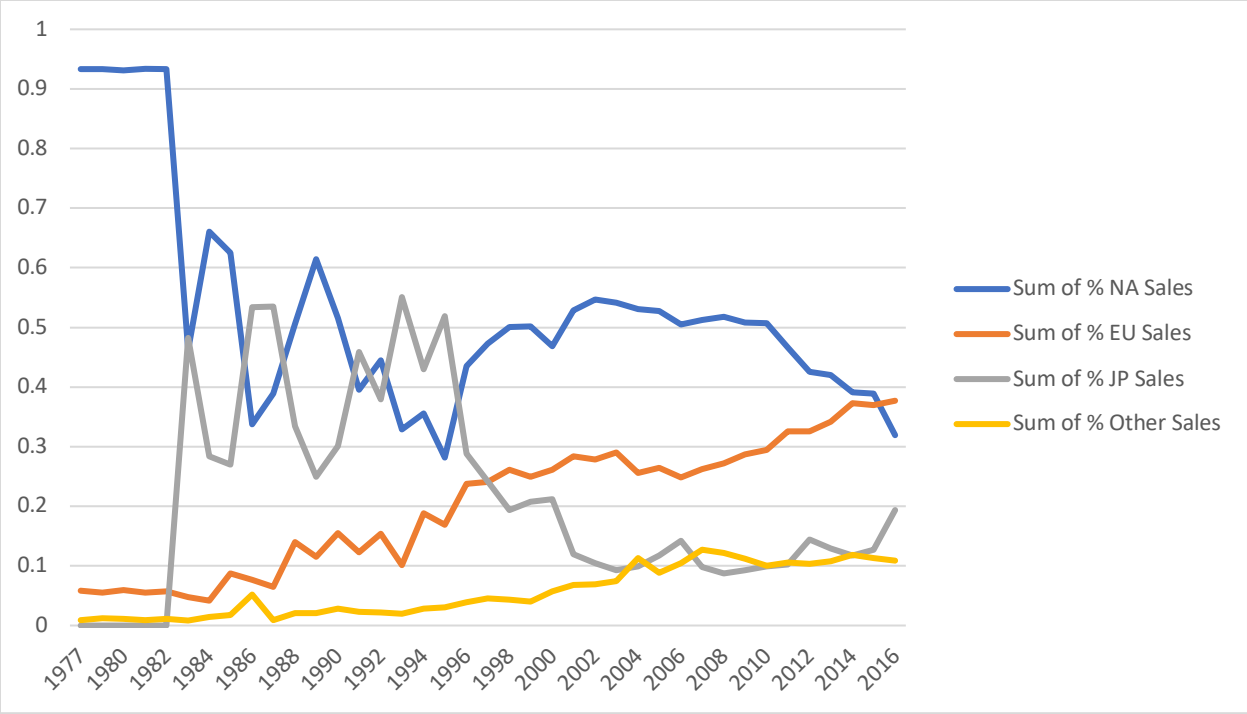
Step 1.

First, I needed to clean the data. The “Rank” column is unnecessary, so I removed it. Then I found that 271 rows under “Year” had N/A instead of a year. After sorting global sales in descending order, I decided to research and input the correct years for the games that had 0.5 or more million units sold worldwide (53 rows) and discard the rest. This was an arbitrary decision, but I didn’t want to spend too much time getting data on my own and 500,000 units sold globally seemed like a good cut-off point. For the games that released in different years in different regions, I used the year in which the game was released in the most regions – these games tended to be released during the fall in every region except Japan, which received their release early the following year (7 rows). There were 3 rows for 2017 and 1 row for 2020. I researched these four and updated them to show their correct release years. There were a number of N/A and Unknown written under “Publisher”, but it was not important to know the publisher for what I was looking for, so I left them in.

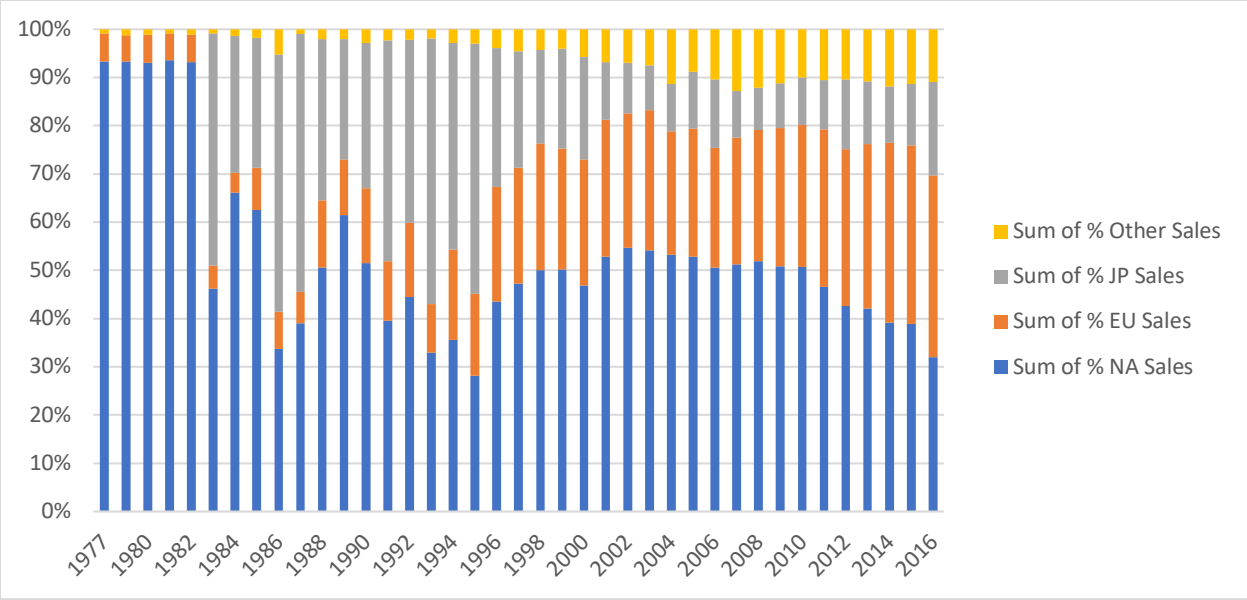
At the end of cleaning the data, there were 16,379 rows of data, and the summary statistics for the data set consisted of the following (note that sales are millions of units sold):

	NA Sales	EU Sales	JP Sales	Other Sales	Global Sales
Mean	0.27	0.15	0.08	0.05	0.54
Median	0.08	0.02	0	0.01	0.17
Mode	0	0	0	0	0.02
Min	0	0	0	0	0.01
Max	41.49	29.02	10.22	10.57	82.74
Range	41.49	29.02	10.22	10.57	82.73
IQR	0.24	0.11	0.04	0.04	0.42

I needed to check the data to see if the executives’ assumption that the proportion of sales per region stayed the same over time was correct or not. In order to find that out, I wanted to create a line chart that would show the proportion of sales per region over time. So, I made a pivot table with the years as the rows, and I created new columns that would show the proportion of sales per region. Then I created a line chart to show how they compare to each other.

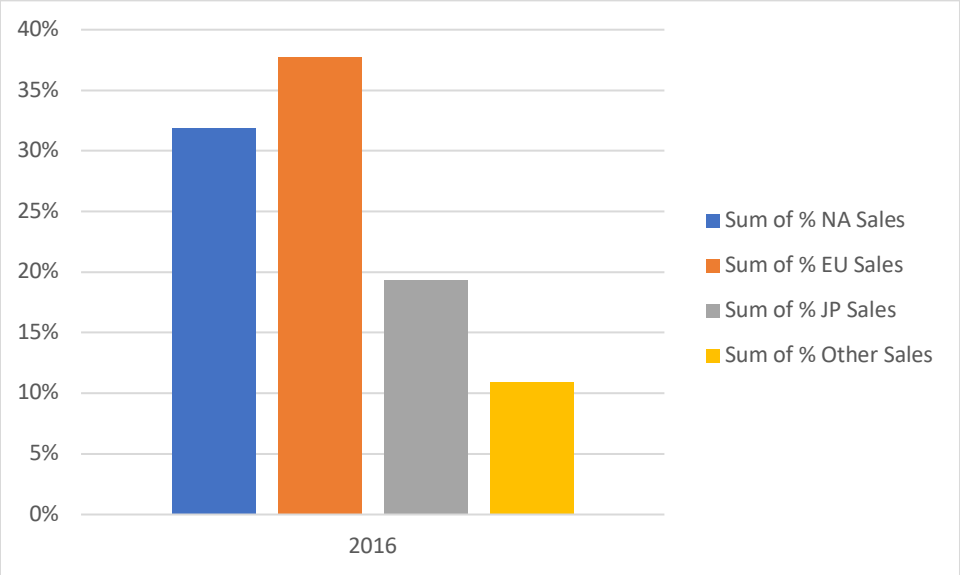


This chart clearly shows that the proportion of sales per region has not stayed the same over time, and does in fact change, which challenges the executives' assumption. I also created a 100% stacked bar chart from that same Pivot Table to see how that showed the data.

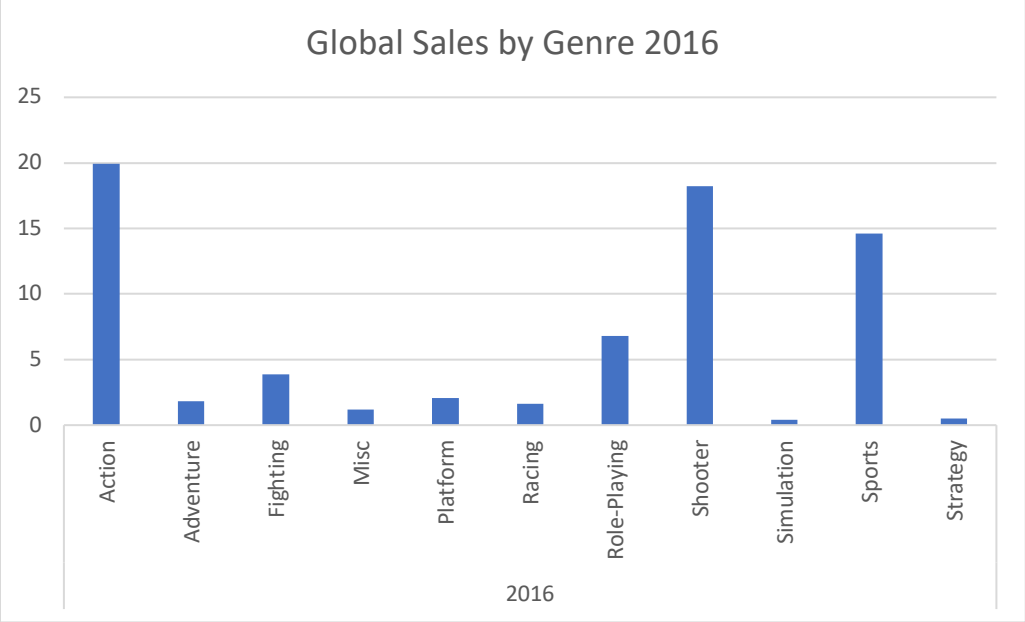


I looked at how much data I had for each year by creating a Pivot Table with the count for each year. I decided that I would only use data from 1994-2016 for the presentation to the executives, because the previous years had much fewer records.

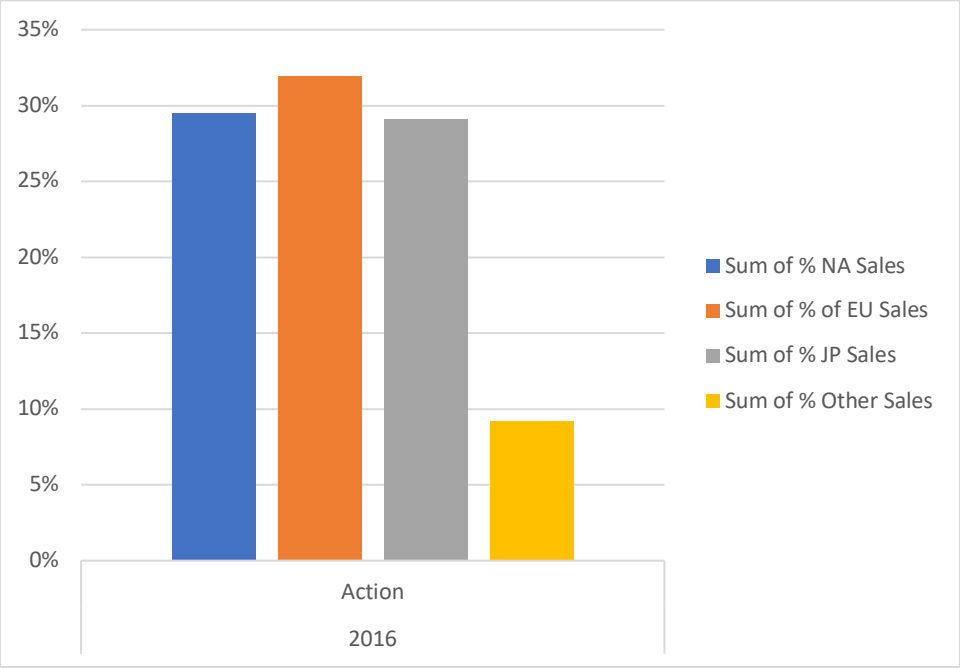
After I showed the executives the polished data visualizations that challenge their expectations, I had to show them how to revise their understanding and how they can use the data from 2016 to inform their decision for 2017. So, I created the same Pivot Table as earlier, but added a slicer to only show 2016. I reformatted the columns to percentages. Then I created a histogram from this data.

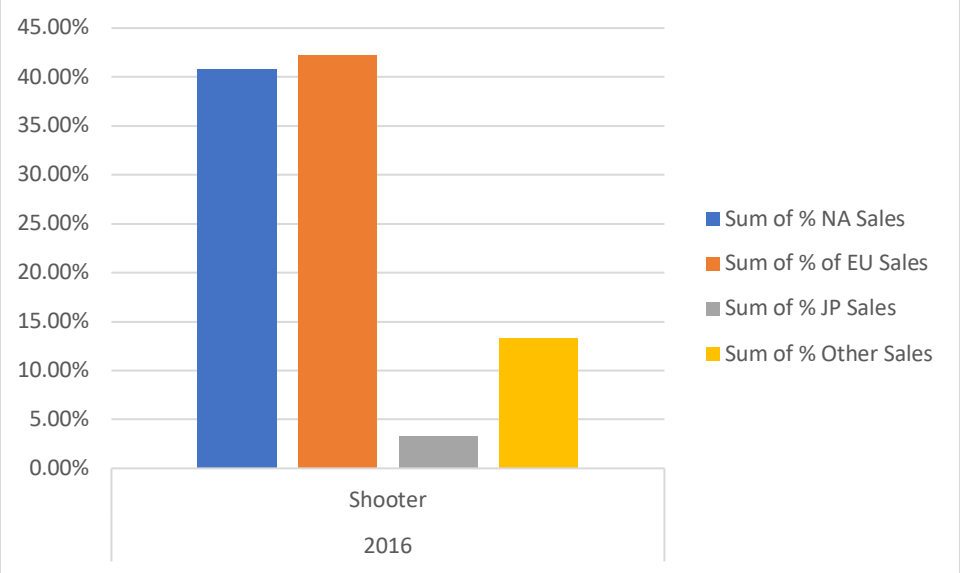


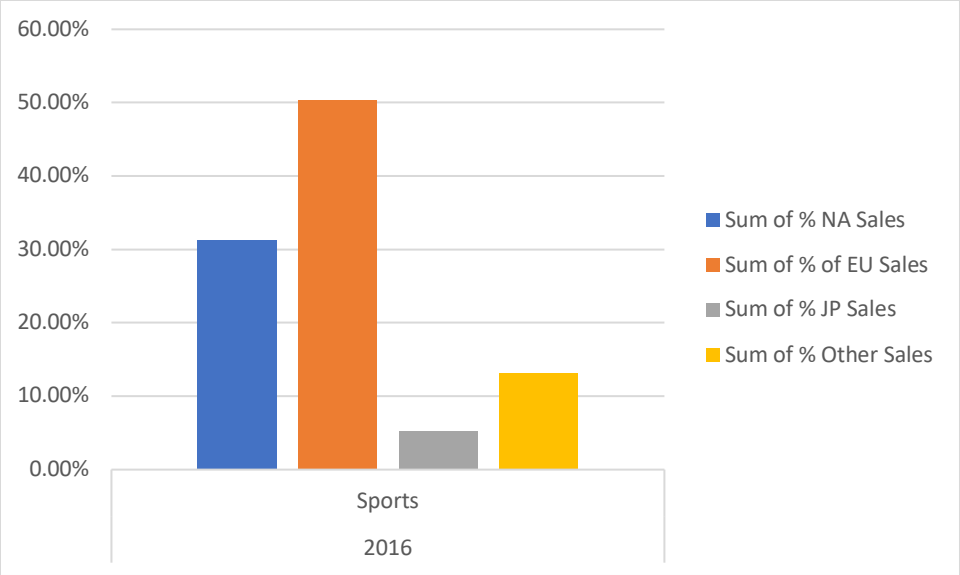
Next I wanted to show the executives what the most popular game genres were in 2016. I made a Pivot Table of sum of global sales by genre by year and used a slicer to show only 2016. I then made a histogram from this data. The three most popular genres were action, shooter, and sports.



I decided to go a little further and look at the percentage of sales for each region for the three most popular game genres in 2016. In order to do this, I created three new Pivot Tables with the same rows (year and genre) but instead used the columns for percentage of global sales by North America, Europe, Japan, and other. For each Pivot Table, I used a slicer to only include data for action, shooter, and sports games in 2016. Then I made a histogram for each.







Step 4.

I chose a line chart that shows the percentage of global sales from each region. I think this was the best way to show the executives how each region compares over time. I looked at both a line chart and a stacked bar chart in step 1, but I decided to use a line chart. I thought the line chart was easier to read and understand. The line chart in the presentation is very similar to the line chart in step 1, except I only used the years 1994-2016. I thought the years before 1994 had too few data points. I also reformatted the columns to be percentages. I also cleaned up the chart by adding a title, axis titles, and renamed the titles in the legend. This made the line chart easier to understand.

I chose a histogram in order to show the executives the percentage of global sales from each region from 2016. I cleaned it up from the one above by adding a title, axis titles, getting rid of the legend to label each column, and adding a data label for each column with the exact percentage. This made the histogram easier to understand.

I chose to use a histogram to show the executives the amount of global sales by genre in 2016. I wanted to show them what game genres were most popular worldwide. I cleaned it up from the one above by adding axis titles and a data label for each column with the exact percentage. This made the histogram easier to understand.

I chose to use histograms to show the percentage of global sales by geographic region for the three genres with the most units sold worldwide (action, shooter, and sports). I thought it was important to show the executives how preferences can vary widely per region. I cleaned them up from the ones above by adding titles, axis titles, getting rid of the legends to label each

column, and adding data labels for each column with the exact percentages. This made the histograms easier to understand.