Data Story: Visualizations and EDA

The main goal of this capstone project was to look at whether or not there is a correlation between Metacritic scores and how well a game sells. By looking at a plot of the Metacritic scores and global units sold, it can be seen that higher scores tend to have higher amounts of video games sold. However, it looks like many data points are on the lower end of the global sales. When exploring deeper, taking the log of global sales may provide a better visualization. There may be a trend that was observed, but perhaps what contributes to video game sales is a bit more complex than just having higher scores. What else affects the global sales of a video game and/or the metacritic scores? Is it the difference between user and critic scores, the video game genres, and/or the publishers? Perhaps it is a combination of these factors or more.

In order to see how the genre affects global sales, the dataset had to be grouped by genre with an aggregation of the sum of global sales. A bar chart was plotted for these two variables and it is seen that the highest selling genre is action, by about a total of 300 million units sold more than the next two highest genre. The action genre is followed by the genre of shooters and sports.

Another big factor in how well video games sell is the publisher that releases the game. There were too many publishers to look at all of them for this initial analysis; the top 10 publishers that released the most games were looked at. In order to retrieve the top 10 publishers by game count, the value counts method was performed on the publisher column of the dataframe. Doing this returned a pandas data series with the index being publisher names and the column being the amount of times the publisher appeared in the dataset. The top 10 publishers were extracted with the head method as well as the index.values method. This returned an array which was used to filter the dataframe. This new, filtered, dataframe was then grouped by publisher with an aggregation on the sum of global sales. Both a pie plot and a bar chart were used to visualize the results. The main reason why both of these visualizations were performed was to show the difference between observing a pie plot and a bar chart. It was much more informative to look at the bar plot, rather than the pie plot. It can be clearly seen that both Nintendo and EA were significantly out performing the rest of the top 10 publishers by 200 million units over the next highest publisher. The relationship between count of games per publisher and global sales could be explored further.

So far, it was seen that both publishers and video game genre have an affect on global sales. Which video game genres perform the best for the top 5 publishers? The dataset was filtered by the top 5 publishers, grouped by publisher and genre, and then aggregated by the count of genre for each of those publishers. The most interesting observations that came from the bar chart visualization of genre counts per publisher were that EA released a very large amount of sports games and action games were the most produced genre. Also, Nintendo released more role playing games that any other of the genres they released. The results of the visualization shows that there may be some variance in the distribution of video game genres.

We now know that there are certain publishers that seem to be dominating the overall global video game sales and that they tend to release specific genre over others. How does the overall count of games per publisher relate to the global sales? The visualization of this shows the publishers who release more games tend to have more overall global sales. If this were to be looked into further, perhaps instead of looking at count of games per company, average metacritic scores per company could be investigated.

Data Story: Visualizations and EDA

By looking at the top 10 publishers and grouping by the average of both the critic scores and the user scores, it can be seen that Nintendo does slightly better in user scores overall and nearly ties with EA for critic scores. Overall, all the average scores for the top 10 publishers tend to be within the range of 65-80.

The last aspect of this dataset that was looked into was how many games with Metacritic scores were sold per year up to 2016. The dataset was grouped by year of release for the games and then aggregated by the count of video games released that year. A line chart of this time-series data shows that there was a skyrocket of games with Metacritic ratings sold in 2001 and a drop off in about 2009 and onwards. This could be investigated further to see what causes the rise and drop of this data.

In summary, the variables that seem to affect the global sales the most are the genres, the publishers, and the amount of games sold per publisher. Performing this initial exploratory data analysis allowed formulations of potential routes that could be explored further in the data. Now that there is an idea of what the dataset looks like, linear regressions, hypothesis testing, confidence interval calculations, and other inferential statistics techniques can be performed on this dataset of video games.