Andrew Torrance

3 May 2017

DSEA-06

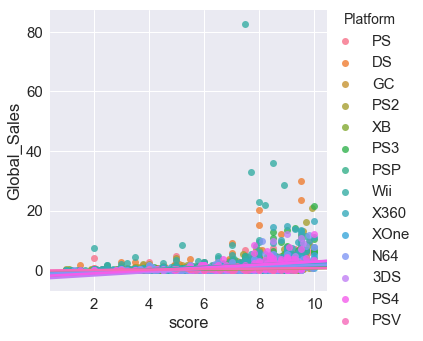
**Problem statement and hypothesis**

Video game publishers fight for consumer’s money. While a successful game can prove extremely profitable, video games with sizeable production and marketing budgets with even tepid reception can financially burden publishers. If IGN’s video game reviews are an indicator of the popularity of a video game, then the number of copies sold of a video game should be higher based on IGN’s rating.

**Data Sourcing**

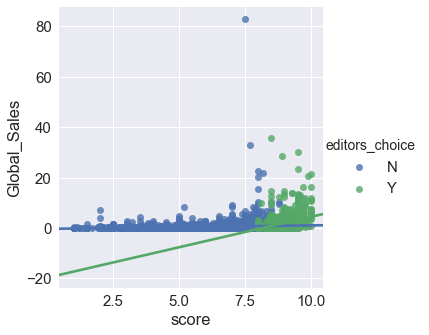
Two datasets were obtained on the website Kaggle and were processed through a multi-column merge based on the name of the video game and the platform on which they operated. The first dataset, Video Game Sales, contains columns of data conveying the title of the video game, its operating platform, release year, sales numbers broken down into the regions of North America, Japan, Europe, Other, and summed globally, genre, publisher and publishing year. The second dataset, titled “20 Year of Game Reviews”, Upon creating a scatterplot of the data using an x axis of a title’s IGN score and a Y axis of the title score’s scales in millions, it became apparent that there would be little linear correlation and any correlation between game sales and IGN score would likely be exponential.

**Data Cleaning and Exploration**

The IGN video game ratings and Video Game sales chart based on matching strings in their shared “Title” and “Platform” columns. The strings in the Sales chart had to be recoded to match those of the ratings chart. Once this was completed and the tables merged, np.isfinite was utilized on both the “Global\_Sales” and “score” features to remove any rows that contain null values. I used seaborn to produce a plot of video game title sales as a function of scores, taking note of different platforms. The fit lines produced by this Seaborn plot did not appear to have much positive slope which suggests a less powerful correlation between the two features than I expected. Using train/test/split on the whole of the data, I calculated a line of best fit with a slope of .275 indicating an average increase in sales of 276,000 units per 1-unit increase in score. 

Within the data, I did notice that after scores exceed 5.0, there is a greater range of sales figures for games. Additionally, there appears to be a higher degree of linear correlation between the categorical feature “editors\_choice”, which indicates whether or not a title was selected by IGN editors based on its innovations in gaming and other less quantifiable aspects of the title that would suggest it to be an enduring game.

As a means of narrowing the scope of the hypothesis, I decided to see if there was any difference between the success of Microsoft and Sony’s video game consoles and associated titles.



**Challenges and Successes**

The merged dataframes of IGN review scores and video game sales created an interesting set of features to explore, but unfortunately reduced the total number of rows of data from 16,598 to 6,480 complete rows of merged data. Through visualization I also noticed that most video games that received a score below 5.0 were very unlikely to have high volume sales. Due to this, I decided to remove them from the dataset and examine the correlation between sales and video games with a score of 5.0 or higher. Through the addition of the categorical feature of genre I was able to refine my regression a bit further, but the R squared value for the lines still fell very short of indicating any predictive correlation.

In further data exploration, it became apparent that additional forces affected video game sales beyond simply their review rating. Mean video game sales appeared to have a normal distribution over the lifespan of a console, with the games with the highest sales numbers releasing in the middle of the console’s lifespan.

**Sources**

Video Game Sales Data: <https://www.kaggle.com/gregorut/videogamesales>

20 Years of Games: <https://www.kaggle.com/egrinstein/20-years-of-games>