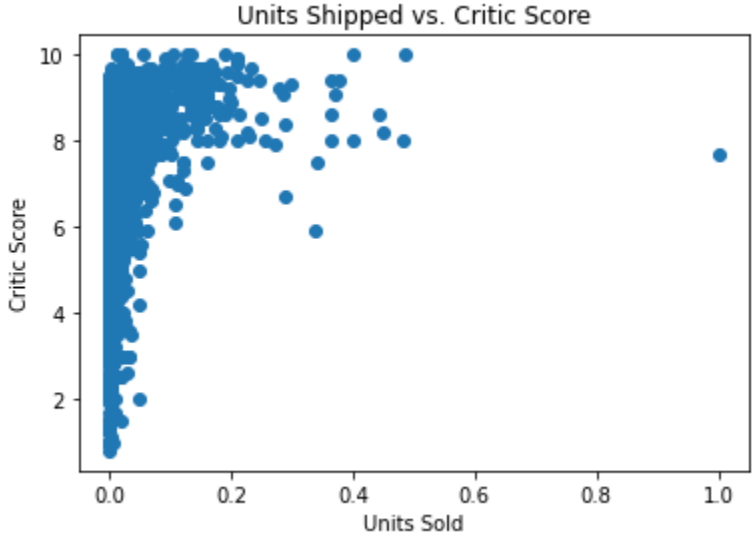
Data Mining Final Project

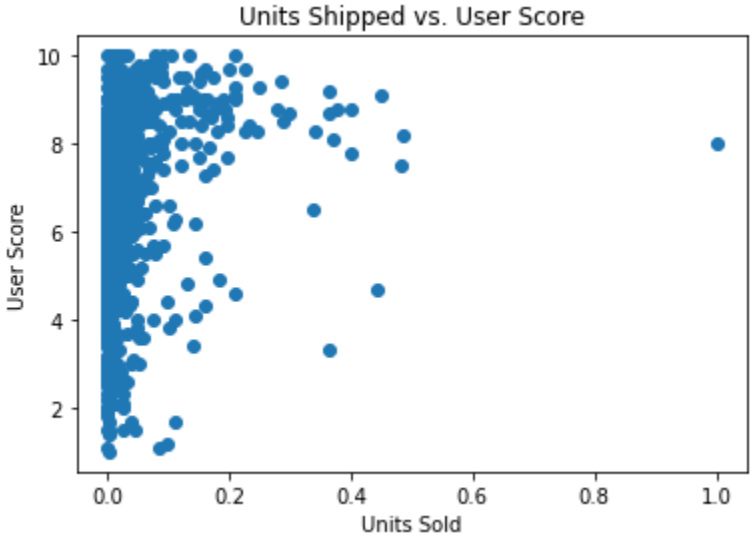
Jay Pfister

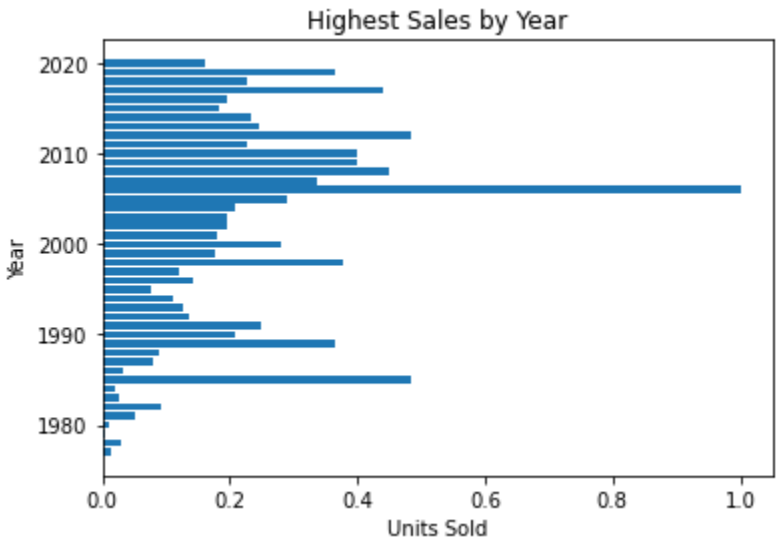
The heart of this project is the comparison of video game sales data vs. their critical reviews from both end users and industry critics found on Metacritic.com. The initial hypothesis is that that critical acclaim correlates directly to sales figures on the back end. In order to test this, I compared the closest metric to sales figures I could reliably find (Total number of shipped copies that I then 0 mean normalized) against those critical reviews.

A potential flaw in my data set is that I don't have actual sales numbers for any of these games. Getting them for 19000 titles would be nearly impossible, so I used total shipped as the best proxy available. With the actual sales data, a more fine-toothed examination could be done but I would predict the results would be similar, accounting for outliers like blockbuster perennial titles that always sell well.

The first question I encountered when looking at this data was what parts of it were really necessary to achieve the goals I had. In evaluating these features, I used python to take a walk through them examine them and their potential relationships to one another.


Units Shipped vs. Critic Score
Critic Score
Units Sold
10
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0


Units Shipped vs. User Score
User Score
Units Sold
10
8
6
4
2
0.0
0.2
0.4
0.6
0.8
1.0



Viewing the data in a Pandas Data frame helped make that clearer. The data frame contained a number of fields that I just didn't need for my analysis. Examples of those fields are things like rank (which is really just the title's rank by overall shipped numbers among the list), and the year the game was published.

An interesting further analysis I would do given the time is observing trends across time periods using the year data. Ultimately, the features I really needed were the review scores and the sales data. I didn't even necessarily need the game titles as I was just looking at the big picture.

At that point, I had my features that I would need to explore this hypothesis, so I went about shaping the data in a usable way by dropping features I didn't need and grouping the data itself by developer before purging any NaN values.

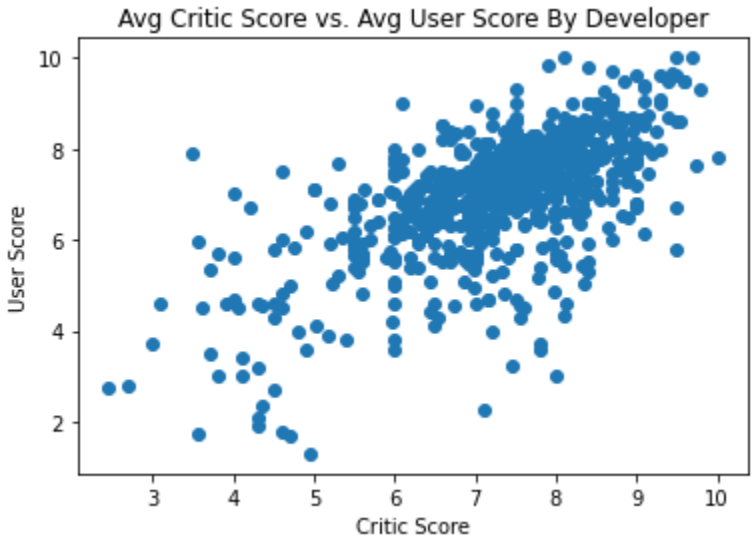
The next issue I encountered was the shipped data was in a format that wasn't particularly neat or scaled in any way. The solution I found was to use a scaling library to 0 means scale the values between 0 and 1. Once I had accomplished that, it was simply a matter of taking the data and performing a linear regression to determine if there were in fact meaningful relationships between the features I had chosen.

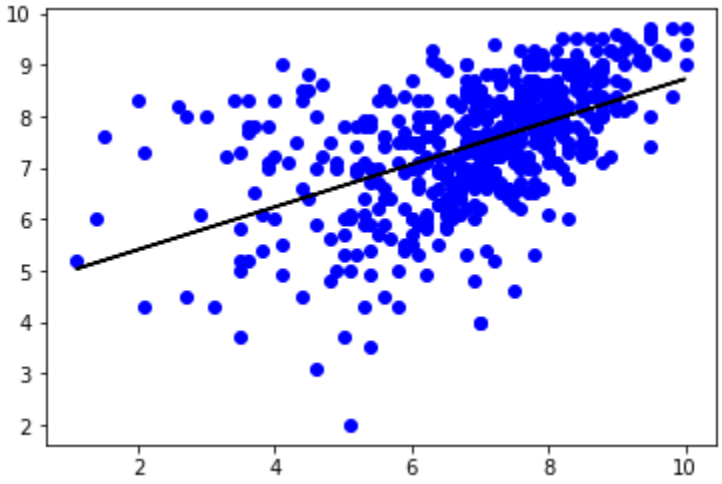
There was a fair amount of trial and error in this process even though it sounded simple at the outset. Things just sort of kept popping up that I hadn't considered. I'll provide a few examples of "those things". To start with, I didn't have the actual sales revenue data, so I had to use a proxy for it which is a good stand in but isn't the real picture. In a perfect world, I'd have had the sales data for all the games on the list, but that simply wasn't realistic. Ideally, if I could gather the sales data for each title from each developer, I could make determinations regarding the way they report their sales and if necessary, normalize the data to be able to fit cleanly into a regression model.

I also noticed that there were almost no scores below 5 for the industry critics and chose build the experiment in such a way as to omit scores lower than 5 for both columns to avoid warping the data with odd user scores and help to normalize the scoring.

At this point, the units shipped data wasn't really normalized in any way, the numbers were in millions and the accounting varied in its level of precision. The solution I came to was to scale the data using a Zero Means Scalar library, which as the name implies makes the mean value zero and uses a standard deviation of one. Doing this made it much easier to work with the total units shipped data without distorting its proportionality when comparing it to itself.

Ultimately, I found that in general terms, critical review has almost no impact on the sales of a given game. This applies to both end user review and industry critic review, which themselves have a fairly significant overlap. The correlation between those numbers and total shipped units is negligible. What this tells me is that people just don't care about reviews for titles they were already interested in. There are so many other factors that account for this that I didn't have time to examine all of them, but one that I would consider is pre-order purchasing incentives which lead consumers to purchase titles in advance of their release before any review can be produced.





Truth be told, this was sort of the position I had at the outset, but I chose to approach this assignment from the perspective that I was wrong and see if the data bore it out. As someone fairly involved in the “playing games” space, I know who my peers are and whether they read reviews or not. They don’t, ever.