Midpoint Report - Video Game Analysis

My data explorations project is still an analysis of video game sales. More specifically how total game sales relate to other attributes such as critic or user ratings. My ultimate goal is to determine which factor is most correlated to total video game sales and then analyze why.

My overall goal is still the same as it was in the project plan. My approach is mostly the same, as I still want to create visualizations for the data that I've generated using the web scraping tool. However, I will need to spend significantly more time than I realized cleaning the data. My approach will likely consist of some trial and error where I create some test visuals to see if the data is sufficiently cleaned and then I will keep cleaning it until I'm satisfied with the results. This contrasts with my previous approach where I would have done all of the cleaning first and then created the visuals after.

I've made some progress on my milestones since I made the project plan. I've created the initial dataset from the website and started the cleaning process. I'm a bit behind where I thought I would be at this point mostly due to the fact that the dataset was harder to generate than I thought. The website scraping tool ended up being less effective than I thought it would be. I had multiple problems getting the code to even run at all. There were issues with the version of Python I was using and other errors regarding calls to external libraries. One of the biggest problems I had was how long the program took to run. It took several hours to complete and crashed on several occasions which afterward I had to change the code to output page by page instead of all at once so that if it crashed I would at least have everything that it fetched so far. In the end, I was able to get all of the data into one file and have begun cleaning it. The data is quite a bit more incomplete than I previously thought but I think I should be able to clean it up enough.

Since I created the project plan, I've thought more in-depth about exactly which visualizations I want to make. Since I want to focus on what affects total game sales I will put total game sales on the y-axis as the dependent variable while I put whatever I'm comparing it to on the x-axis as the independent variable. Right now I plan to make five graphs, which will consist of:

- Year of release vs total game sales
- Average critic score vs total game sales
- Average user score vs total game sales
- Publisher vs total game sales
- Platform vs total game sales

Measuring total game sales lends itself well to bar graphs, I'll be using bars graphs for all of these measurements. Using bar graphs I should be able to see which of these factors contribute most to total game sales. After I create these main five graphs, I can start to perform some deeper analysis of the trends of the graphs. It's one thing to figure out which factor influences game sales the most but it's another thing to figure out why.

The last part of my project is the analysis portion. After getting the results from the initial graphs I might create a second series of graphs showing trends in game sales. Maybe over the last few years total video game sales have been trending downward. Maybe one particular publisher is taking over and selling more and more games each year. In its simplest form, my overall goal with this project is to determine if video game quality as a whole is getting better or worse. Nowadays, I often hear people complain about every new game that comes out, so this project is a way to determine if games are actually getting worse and if so why?