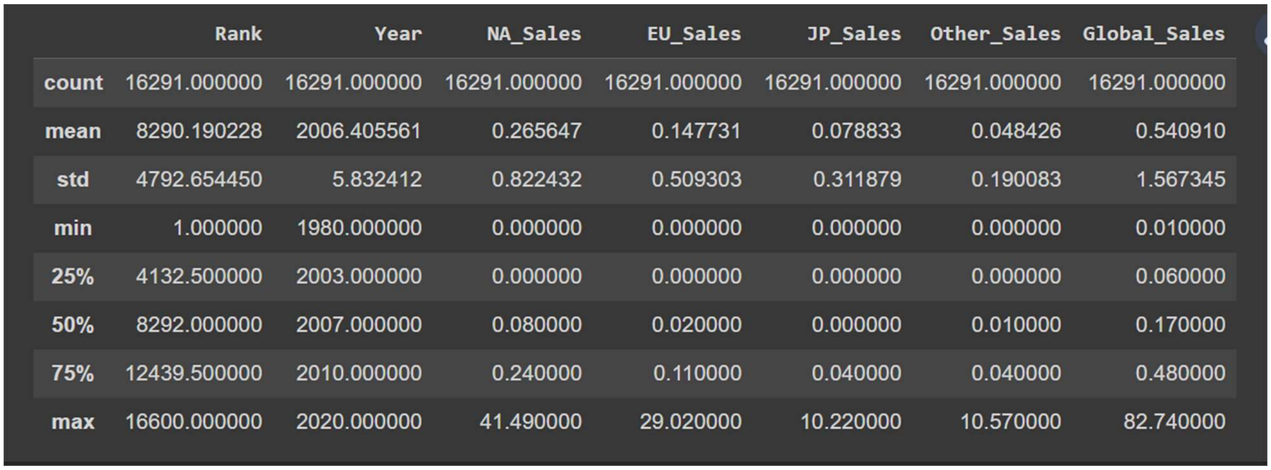
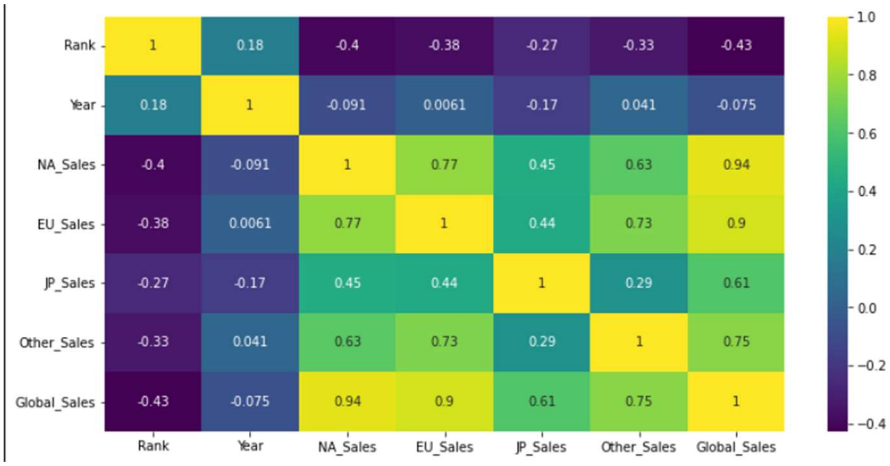
Upon starting the data analysis of our project, we had already had a big issue. Our original project intended to compare college majors and projected salaries to determine which college major had the best wage and unemployment rate and other analyses. When looking at the data file we found for this project, we discovered that the data did not have much of the information that we needed to continue with this idea, so we started back from square one and searched for a new data set. After taking some time to find a new data set, we found one that appeared to be both practical and interesting to analyze. Our new data set, global video game sales, allowed us to analyze a variety of variables with sales between different regions. With the new data set in hand, we decided that our new question would be: What are the best variables to predict high sales for a video game? We created a Google collab for a space to code due to familiarity and convenience to use.

Now back on track for the project, we had to start with basic data analysis. To start, we had to clean up the data from its original condition. Luckily, the only issue that we found was that a few entries were missing the year. While we could've manually entered the release date year for each of those entries, we decided to just cut that data out as there were too many entries to fill in, and our data set is large enough to save time and remove that data. After that, we decided to create a five-number summary for our integer data columns to get a better idea of how the data was spread out, which looks like this:



	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000	16291.000000
mean	8290.190228	2006.405561	0.265647	0.147731	0.078833	0.048426	0.540910
std	4792.654450	5.832412	0.822432	0.509303	0.311879	0.190083	1.567345
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4132.500000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8292.000000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12439.500000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.480000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Moving on, to keep things simple for this part of the project, we only did the correlational analysis on the data that is already integers, leaving the correlation between genre, publisher, and system unknown at the time. After writing the code that does the correlation calculations, we also made a heat map that helps visualize the correlation values, which looks like this:



From this heat map, already we noticed something interesting. Comparing the correlation between global sales and each four regional sales, we noticed that NA sales are the largest determining factor for global sales, suggesting that the American market is where a game needs to be sold to ensure success. Conversely, we noticed that the correlation between the year and NA and global sales is negative, suggesting that as the years go by, video game sales are slowly declining in each market. Going forward, we will find a way to include the other variables in our correlation analysis, whether using a counter method or transforming data types into integers to calculate the correlations. When all correlation calculations are done, we will be able to determine which variables would be the best predictors for high video game sales, as well as do a more in-depth analysis as to which variables lead to high sales in certain regions.

Looking forward to our timeline, we decided it to be best to start meeting weekly on Tuesdays in order to make sure our project moves along in a timely manner and to ensure we are always on the same page on decisions. By April 19th, we plan for our model to of been created, using the variable correlations we calculated and demonstrating what would be the best factors for high video game sales. We also plan to split the work on the research paper for part 3 of the project by allowing the person who codes an action to explain it in the paper (for example whoever codes a graph may be the one to explain it). And using the feedback from the class on our work so far, we will improve our model and paper in time for the final submission.