The problem I'm aiming to solve in my capstone project is a fairly simple one and was something I'd always wondered about the specifics of: How do video game reviews relate to video game sales? As I was looking through possible datasets to analyze, I found a Kaggle dataset that interested me very much (this was prior to my having any concrete goal in mind for my project). The dataset was composed of scrapes from two websites, vgchartz.com for game sales information, and metacritic.com for weighted game review score averages. The dataset was acquired in December of 2016, so it's not the most up-to-date data that could possible be available, but the data is very extensive and it has proven to be rather useful in deepening my understanding of data cleaning and working with semi-large datasets.

Looking more closely at the dataset as it originally existed and how it's changed since I've begun manipulating it, answering a few questions is in order. Firstly, what important fields and information does the data set have? The most key column headers in answering my capstone project's question would be the following: Global_Sales, which lists the total sales of a game worldwide, Critic_Score, which is a weighted average of the review scores critics have given games from various websites across the web, Critic_Count, which provides the number of sites used in order to provide the Critic_Score value (if there are less than 4, Critic_Score was originally an NA value), User_Score, which is a mean of all the users of Metacritic as they have provided their own scores for reviews of a given game, and User_Count, which provides the number of users that have provided a user review of a game and follows the same rule for an NA value if there are less than 4). These five columns are the most important because they provide the information for the main comparison I want to make in this project: If Critic_Score is a higher number even as a weighted average, does that tend to correlate to higher game sales? Does that correlation hold true with User_Score? How much does the weight added by Critic_Count and User_Count (meaning that more critics or users were assigned to or cared to review a game) tend to affect sales? These main questions plus possibly others are hugely important in utilizing my dataset to effectively answer the question I'd posed with this project.

Aside from these main informational columns and possibly looking ahead to further cleaning or clarification of data, some other columns may be of importance: Year_of_Release, which lists the year of release of a particular version of a game, Platform, which names the console type in which a particular game appeared on (perhaps the Gamecube version of a game was better than the Xbox version or vice versa), NA_Sales, JP_Sales, EU_Sales, and Other_Sales each list the respective sales in the continent or "market" that video game companies tend to separate themselves by, Publisher, which lists the specific company that paid to put that game on a shelf or get it

pushed through a digital service and, finally, Genre, which lists the style of game presented be it RPG, Action, Adventure, or even Misc.

Another question worth answering about the dataset is relates to the specific limitations of the data provided. Questions that could not be answered by this specific dataset would be those related to price. Game prices change constantly after their initial release, and this dataset contains no information even related to a game's price at launch let alone the current cost of the game or how that value changed since its release. Asking whether game prices should be dropped or if a certain game from a certain published should launch at a certain price, again, would all be very difficult angles to gleam much info about at all.

Other questions that might be difficult to answer (if not impossible) would be if a particular game or console was profitable for the developer or publisher that created it and put it out. This dataset contains no information as to how much it cost to make a game or how long it took to make. Further, the dataset doesn't contain information related to developers (as opposed to publishers) at all. This is a key bit of information for answering questions within the games industry.

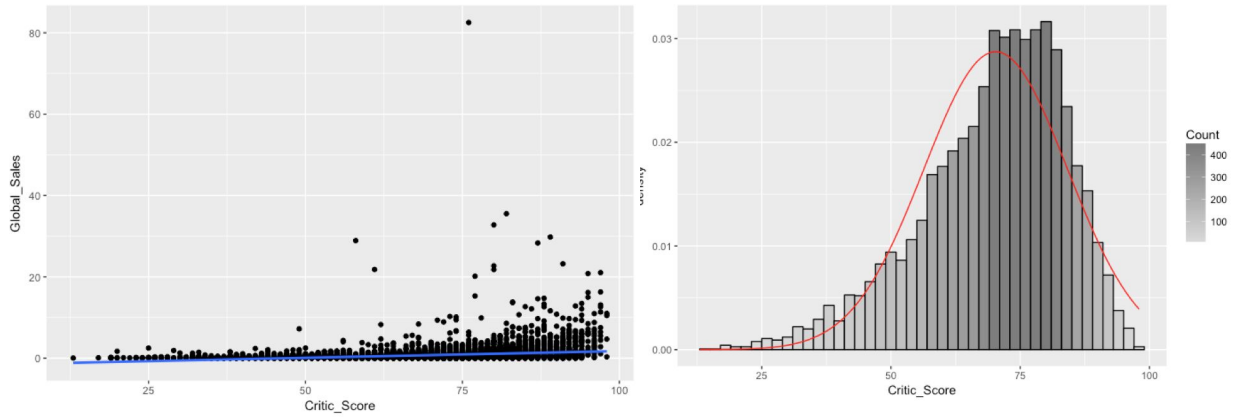
One last concern about the dataset relates to the way in which Metacritic provides review averages in the first place. Metacritic is a huge operator and of high importance in the games industry. There was a time (and that time might still be right now) when stories circulated of developers bonuses hinged largely on if a game broke the Metacritic average of 90 or not which is an incredibly difficult task. That said, the averages Metacritic provides are weighted meaning that some review outlets are given higher importance than others, but the respective weights given to an outlet are a closely guarded secret to Metacritic alone. The Metacritic FAQ on the site makes this clear: as an answer to the question "Can you tell me how each of the different critics are weighted in your formula?" the answer provided is simply: "Absolutely not." Also of concern is how Metacritic converts some sites review scores. For example, the now defunct site 1up.com switched to a grade scale for their reviews (A, B+, C-, etc.) and these grades would commonly be related to numbers in a certain fashion though I don't believe this was spelled out on their site (A is a 95, B+ is an 87, an F is any score below 60). Metacritic took these grades and divided them equally between 0-100: An A is 100, an F is 0, and all the other grades from D- to A- fell evenly divided in between those numbers. This goes against the spirit of 1up's reviews and eventually led to the removal of 1up's review scores from Metacritic but nevertheless points out a major issue with Metacritics review average system.

Regarding the cleaning of the original dataset that has been done so far, I published a somewhat R code-heavy data cleaning explainer on Github in July. In summary, most of what I did related to the NA values that existed in the original dataset. After writing a script that creates a dataframe which tells me in columns that match the dataset I'm using how many NA values exist in each column, I first set about removing two rows (from the more than 15,000 that exist in the dataset) because these two rows have NA values for Name, each game's title, they're from the early 90s, a time when Metacritic didn't exist and likely the site didn't go back and acquire every possible publication for whichever games these two were, and, most importantly, these rows don't have user or critic score values or counts, which makes them somewhat useless in solving the main problem I have.

Next, I set about the largest NA value issue in the dataset I'm using: almost every item in the dataset has sales data of some kind be it Global, NA, JP, or EU, but there are thousands of rows that have no critic or user scores. At first I was tempted to simply delete all rows that didn't have both sets of data that I wanted. This made the most sense to me because I wanted to compare review scores and sales data. If a game didn't have both, it was not of much value for my main comparison and the problem I was trying to solve. After speaking with my mentor, it was suggested that I replace NA review score values with an average of every other review score data point. It still seems problematic to me to think that creating data in this fashion is useful, but I'm definitely learning the ropes here. I've seen during the data visualization portion of this course that some histograms show many more than seems to make sense at the point where the average value replaced the NA value, but perhaps this specific graph could be used on the dataset where I deleted those rows for a cleaner image. Maybe not. That said, after having replaced the NA review average values with an average of the others, there were miniscule NA values left in the dataset. These values were replaced with either 0 values (for user or critic counts) or a string like "Unrated" instead of the NA value which can cause some R functions to not work properly. Finally, outside of some definite outliers in the dataset, the dataset appeared to be largely cleaned up. I'm sure I'm not finished messing with the dataset though regarding cleaning though. A lot of my visualizations don't seem to make sense, and I'm guessing that better or more thorough cleaning would fix some of these issues.

Moving on, I'd like to list some of the preliminary exploratory steps I've taken to attempt to visualize the dataset I'm working with and to make initial attempts at solving the problem I started this project with. I posted a data visualization explainer as part of this course on August 29th on Github which goes into detail a little bit further than I will here (it's a Word document), but mainly I've been attempting as many basic methods of

visualizing my dataset as I can think of. The most useful of the visualizations I've attempted seem to be the scatterplot with a linear model running through the data and a histogram of the Critic_Score with shading based on the review count for each time a game received a certain score average but also with a projected normal curve through the data (which my mentor has stated doesn't make sense as the distribution of data I'm using here is not normal. These two visualizations are shown below for reference. If they're a bit too small to see adequately, again, please check out the explainer above for a better idea of what I'm working with here.



Other than these two seemingly helpful visualizations, I've created many other histograms and scatterplots. I've used alpha transparencies for the data points so that the more that exist, the heavier the darkness of each spot would be. I've also attempted to use jittering to make the data better visible, but there seem to simply be too many data points on too small of a space, so this seems to do nothing helpful. Lastly, I've also tried a few basic attempts at making the data distribution normal (via taking the log of all values and creating a histogram of those values as well as attempting a box cox transformation), but I don't think my math skills are really adequate to know what I'm doing when attempting this, and, thus, my results haven't seemed very helpful or informative.

Finally, and based on everything else I've done up to this point and as summarized above, I'm going to keep pushing forward with attempting to answer my initial problem. How, at least basically, are review scores related to game sales? I'll likely need a bit more math and statistics training as well as R data visualization manipulation training, but I'm going to get this problem answered in some fashion. Past that, and almost entirely because I enjoy pushing myself but also to try and get as much as possible out of this course, I'm going to try to use machine learning algorithms to project what

theoretical games from a certain publisher or that sell in certain amounts, for example, might receive as review scores. This seems a cool extension of my initial project and one that would help me in my data science learning.

That said, my milestones are as follows:

1. P-value test - I'm going to conduct one or more P-value tests in order to assess whether some basic hypotheses I have regarding game sales and review scores make sense, and then, whether they're valuable indicators of other data or if they'd need to be thrown out. I'll try to use as little P-hacking as possible here to indicate a reasonable level of certainty.
2. Isolate other independent variables (correlation) - publisher? - I'm going to attempt, in conjunction with my P-value testing, to check for other independent variables. Publisher comes to mind as an early possibility. Genre may be another one.
3. Linear regression for prediction - Lastly thus far, I'll use linear regression as the basic machine learning algorithm to project what another hypothetical game released in some year, of some genre, and/or through some publisher might do sales or review score wise.

There may be future milestones set or changed slightly as I begin wrapping up my capstone project and this course, but I believe these three milestones are a good starting point if nothing else. Of course, any suggestions or input are welcome. I'm looking forward to making this project a solid first example of data science work from a broad set of industry knowledge and interest areas.

Thanks for reading.
-Darin