Feedback on Background/Methods and Data Analysis:

Background

- You have great info on metacritic score; this is a nice level of detail, and there are important implications for your research findings. You will likely want to expand/reflect on this info in your conclusion, as you discuss how your results tie in to the background information if you find an association or not

- The bullet point format ok for a draft, but in the final paper, you need to have paragraph format with in-line source citation

- Please add more info, as discussed in the feedback on the project proposal. For example, I know you have looked up info on Steam and Steam Spy, but it isn’t listed here.

- Also, you mention reading some articles that look at “associations” but don’t discuss what those associations are or what the implications of their findings are, in general or for your research questions. Did these papers look at metacritic scores? Or at genres?

- You may also want to discuss the work of the original authors in more detail, to describe how you are building on what they have already done (or more specifically, how you are doing something new and useful, not just repeating their work with the same data)

- The background section should provide articles and literature to help motivate your own research questions. For example, what articles have led you to posit that there may be an association between number of screen shots and sales? This may not be specific to steam games or even to video games— it could be about marketing with images in general, if you can’t find something specific to games. However, the idea with this project is that you aren’t simply making up questions. You are generating questions of interest that build on existing literature or fill in gaps in the literature. I think your questions satisfy these requirements, but you haven’t demonstrated that in your writing. This is feedback I already gave you in the project proposal, but I’m still not seeing it in the current background.

Methods

- It’s definitely important to be clear about the data collection process. You may want to comment on this with vocabulary from class. Is this data population or sample? Can it be used to generalize to a broader population?

- You discuss this a bit when you claim that because you have all games from Steam, you can generalize to all games for PC (prior to 2016). This is an interesting claim and requires support. Not all PC games are on Steam. Just because Steam is the “number one game distribution service” doesn’t mean that this data can be used to generalize about non-Steam games. That is generally an indication that your sample is representative of the entire population. Is that the case here? It may be, but if so, you need to argue it with better evidence.

- This ties in to your proposed methods of data analysis. Explain why you chose to focus on descriptive and visual measures.

- You may run into issues using discrete variables for regression and scatterplots. It is still the best choice from the options we have learned, but you will only have limited output, which could make regression not helpful. Count data can also end up being heavily skewed or have serious outliers, which can also affect the usability of scatterplots and regression.

- This is where it is also helpful to draw on your background reading results. Based on your background reading (not just on your own knowledge!), do you expect to encounter any confounding or lurking variables? If so, how will you deal with that? How will you visualize the impact of lurking variables, if you find any?

Data analysis

- What happened to question 1? Also, I’m not sure “qualitative” is the word you want here.

- for Question 2:

+ You should consider transforming one or both of your variables to deal with the outliers. This can be as simple as taking the log or square root of the variable. Transformations can help linearize relationships, though they do affect the interpretability of any output from regression.

+ Speaking of regression, you need to assess whether the conditions for regression are met before trying to interpret the p-values. Is there a linear association? Are the residuals normally distributed? Is there constant variance around the regression line at each x-location? Are the data independent (random)?