Nathan Reed

DSC 530

Prof. Jim

06/03/2023

Final Project Response

I feel like the EDA that I completed did pretty well in most aspects. I was able to get results on every question. I was able to apply the code I learned throughout the semester and apply it to my own data. I was able to find out that critic score is more effective than user score on global sales. I definitely thought that while critic score is useful, that user score would make more sense as I tend to trust my peers more for what games I should get than a video game critic.

While I was able to make some conclusions on the data, I did feel like that I probably could have cleaned the data a bit better than I ended up doing. While I was able to get rid of a lot of the data that was not complete, I was unable to make data like global sales fit with the data. There was very little spread of the data due to the data being in turns of millions. I tried to make attempts at changing it to being in terms of hundreds of thousands, but that made the data not function as well. I also could have tried to use a prediction to fill in the missing values in order to keep all of the data given.

There was not an exact variable that I felt was missing, but I do again feel like some of variables could have been changed to better fit the data. In the previous section I mentioned that the sales data did not fit the other variables as well as they could had I been able to successfully find a way to better fit the data. There is also the option that a transformation could have been better served on the sales for each region.

One assumption that I made that could be incorrect is that I could use the Japan sales data and compare that to the North American Sales data for the PMF as well as for the hypothesis testing. I was unable to get a p-value that I thought made sense for the hypothesis test. I also feel like the graph for the PMF did not work very well. The hypothesis test might be working, because of the population in North America being around 4 times larger than in Japan.

I feel like the biggest challenge I faced was being able to make the data work without having a great spread of data with global sales, Japanese sales, and North American sales. I think that a transformation of these variables may have been able to help spread the data and get better interpretations from the data. I am unsure of what is the best way of knowing which transformation to use and when. Learning when to use which transformation to use at a specific time would be the one thing that I would like to learn better.