**An Introduction**

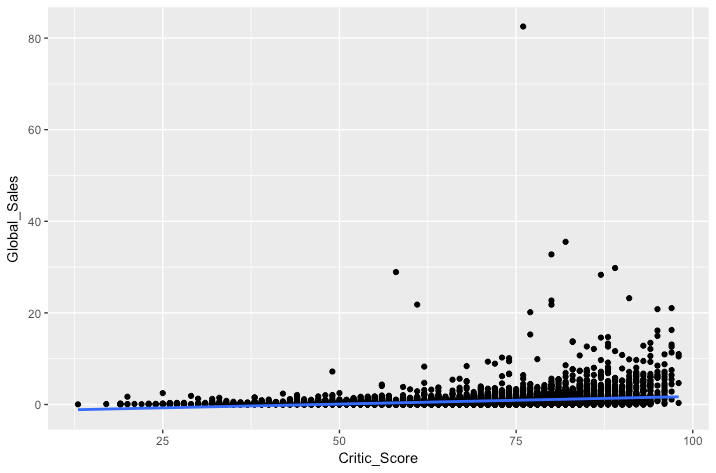
**The Dataset**

**Initial Dataset Cleaning**

**Data Visualization**

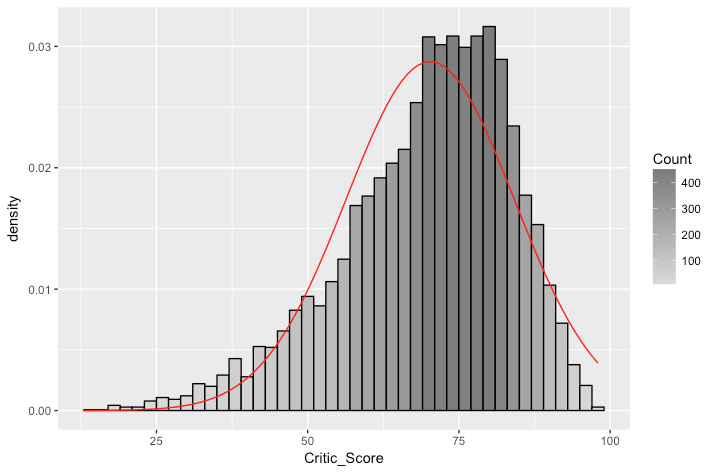
With the point of my Springboard capstone project being to simply see if review scores have an effect on game sales, I was intending to use this [Kaggle dataset](https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings), apply statistics, and then project 1-2 data visualizations (using ggplot2) to prove my overall point. If I was able to do that simply, I’d be done already. To some extent, and in order to learn more about data science and common methods, I’m expanding the goal of my capstone project to predict future outcomes with statistical models and machine learning. Further, I may expand or contract the goal of this project in the future as well to present a quality finding for this project, but thus far, I’ve definitely been having some trouble. In this report, I will document the output I’ve been capable of thus far with the goal of completing a specific assignment (first of all), but also with an invite for constructive criticism toward how I might better this project.

Regarding data visualization, the first graph I did was a scatterplot in ggplot2 with the Metacritic average review score on the X axis, and the total worldwide game sales on the Y axis. I also provided a line acquired from the linear regression model in order to show graphically (as opposed to numerically or mathematically) the increase of game sales as review scores go up:



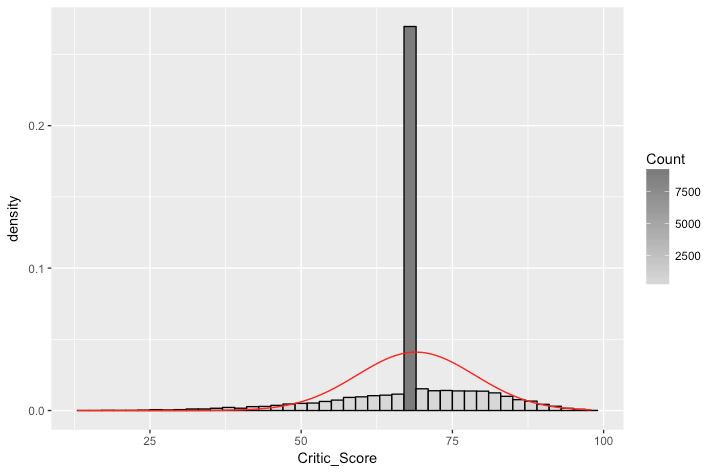
This graph is somewhat helpful, but it would definitely need to be constrained and/or have some of the outliers removed. Also, it would be helpful for a few more text objects or values to be added for the viewer, so they can see what’s going on here and hopefully so they could take some useful information away from this chart.

At first, I thought this chart clarified the main goal of my project. It comes close to showing generally how game sales relate to review scores. After showing this to my mentor for the course, I was told that you can’t use linear regressions directly unless a distribution is normal. The histogram of critic score averages per game is below with a normal curve, and a color change from light to dark depending on the amount of games reviewed with that score:



It must be noted as well, that I had to use a dataset with over half the values deleted to acquire this visualization. The values kept were those values where game sales and critic scores both existed. If games only had sales data and no review scores, I threw those rows out because I couldn’t compare sales against review scores. Likewise, if games only have review scores (whether critic or user) but no sales data, I threw those rows out as well. There were a few rows that had no publisher listed or that were missing published dates. This data is categorical and would have been unrelated to the goal of this project.

Further to the point of data visualization, when I use a dataset where average critic score values are replaced with the mean of all the other values (in order to preserve data), I get the following histogram:



The problem here is that there is a huge amount of scores right at the mean value of all the other review scores. There is likely some value to a dataset like this, but I need further clarification to determine what this dataset is useful for.

Going back to the 1st histogram though, I think this is a nice-looking chart except that density on the Y axis isn’t a great descriptor, though it is necessary to project the normal curve the way ggplot2 was done it. I’d need some help in getting the normal curve (if useful), but then changing the Y axis back to count in order for a viewer-friendly projection.

I also was able to provide plots for the user review score averages, but as these values are from 1-10, the histogram is much chunkier and doesn’t look nearly as nice.