Data Wrangling Report

This project entails three different datasets that need to be inspected and cleaned before going into further analysis. The first set is the all-time video game sales list from kaggle that includes not only the names of the games but also the sales for different areas, the consoles they were on, critic as well as user scores and even more. Following that dataset is the video game Metacritic user reviews, also from kaggle, that contains hundreds of thousands of reviews to cypher through with natural language processing. Finally, the last dataset is from the University of Portsmouth, which has the U.S. sales info of video games from 2004 to 2010. Most of this will be used in the machine learning aspect as a comparison to the models made with the other two datasets. Although none of these sets of data were overly messy, there were several necessary steps for cleaning before heading into exploratory data analysis.

When it came to the all-time sales list, most of the columns and rows were clean. The dataset was pretty tidy with each row being an observation (game) and each column was a feature. However, there was one issue that needed to be resolved. For many of the older games, there was missing data especially nonexistent review info (critic or user). This is mainly because they were around before the Internet was created as well as before reviews became common. And unfortunately, with reviews especially, there is no simple statistical measurement that would be helpful for the data because the scores can range wildly in either direction. Because of this, the only solution was to drop any row that contained an "NaN." While many classics were lost, this was necessary in order to have clean data.

Also, to help make this dataset better, the features needed to be further enhanced. First off, the release years were changed from floats to integers so they would look cleaner and hopefully avoid any upcoming confusion. Secondly, each specific genre was made into its own binary column, which will be helpful when it comes to plotting. Lastly, the different systems that video games came out for were also made into individual binary columns for the same reason as the genre category. With these new columns, the dataset was now ready to be saved into a newer and cleaner CSV that would be used for exploratory data analysis.

For the second dataset, the few "NaN"s were dropped. Then, for later merging, many of the platform names were changed to match with those of the first dataset. Because games could come out on multiple consoles, these needed to be matched correctly. The main cleaning took place with the natural language processing. First, all of the reviews that contained symbols or just numbers needed to be removed because they would cause errors with tokenization. To do that, a function was created to remove any non-letters and replace them with spaces followed by minimizing any extra white space. Then, only the reviews that had words were kept. Finally, the reviews were whittled down more so that only those in English remained.

After the cleaning, this DataFrame was merged with the first one on the name of the game as well as the platform and together there was no missing data. After renaming a couple columns and dropping unnecessary ones, it was time for the actual natural language processing. Because of how many reviews and words were available, the matrix would become too large if everything was run together and my computer could not handle it. To deal with that, a sample of 10,000 reviews were taken and put through a function to help find the most common words. The function made sure all of the words were lowercase, tokenized each word, removed the stop words, broke them down to the stems and lastly counted each word in each review. This made a dictionary that contained dictionaries for each review and the count of each word.

Now this data had to be transformed into a DataFrame with each row being a review and each column being an individual word that could be in the review. The values would then be how many times that word showed up in the review. In order to do that, the dictionary of dictionaries was turned into a list before then being transformed into a DataFrame. Because each column is an individual word, there were many “NaN”s. These were changed into zeroes, and then the DataFrame was put through a TF-IDF Matrix, which is better at weighting words by importance. With this new DataFrame, the most important words were chosen as the ones that had a mean value greater than 0.001. This created series of a little over 1,000 words that were considered the most important and would help with tokenizing the entire dataset.

The same process of tokenization was used on all of the reviews except a change in the stemming portion where only the important words from the earlier sample were stemmed and kept. From there, all of the DataFrame and TF-IDF transformations were the same as they created a new DataFrame that contained all of the reviews and how often the most common words occurred in each review. This was merged with the earlier DataFrame and then modified to only have the mean for each word for all reviews, grouped by each game and platform along with all of the other important features from the first dataset. Lastly, the DataFrame was saved into a CSV that would be used for the modeling section in the machine learning portion of the project.

Finally, the last dataset, which contained U.S. sales info from 2004-2010, was very clean and did not need any work. Each row was a different game and there were many different features as columns including the different genres for a game similar to what was done with the first set of data. Also, there was no missing data so no cleaning was needed. Now there were three clean DataFrames that would be used throughout the rest of the data process to find the deeper trends inside the sale of video games.