**Introduction**

Over the course of my lifetime, I have watched serious video game playing evolve from a “geeky” pursuit to a mainstream competitive event as culture and society has embraced the products of the information and data age. The variety of available themes, styles, and genres of games seems to grow exponentially every year as new ideas and concepts develop. Even those that are avidly not gamers – such as my wife – often find something to capture their interest, from puzzles to simulations to games that do all the work for them. With such an expansive variety of options, the opinions about such games – both positive and negative – are just as expansive

Is it possible to predict the general opinion of a game, however? Using data provided by the SteamSpy API, I plan to compare the actual numbers produced by Valve Corporation’s digital gaming service of Steam to learn what values all of the games currently available on the service to compare the results over time – in particular, since the data’s beginning starting in March 2009. I will be comparing such things as game playing time, purchases, cost of the game, number of owners, and the like to see what differences can be seen between different games and how this all compares to the games’ rankings of positive and negative scores.

**Preliminary Requirement**

As mentioned above, the primary plan for the data is to pull it from the SteamSpy API – this will be done through the use of Python with the urllib and requests modules. From there I plan to perform statistical tests such as logistical regression and multifactor analysis of variance to determine what factors are associated with particular scores. In order to help improve the understanding of the data, data visualizations based on time – such as line and bar charts – will also be implemented to help demonstrate the differences for all audiences, not only those that have an understanding of data analysis or statistics. Additionally, highlighting the data for the highest and lowest scores in these visualizations will also help to bring attention to the particular changes which have occurred.

While the initial data is fairly straightforward, metadata will need to be implemented to help expand the data’s information and clarity. For example, the differences between the totals off all the data from March 2009 to the present and the totals of the last couple of months will help to provide an overview of the gaming data for the past few years prior to the latest developments. Data provided by additional sources will help to clarify this information further.

Since the data provided by the SteamSpy API is a dynamic, live extraction, multiple pulls of that data and/or the acquiring of historical data pulls provided by other analysts publicly online can also help in the testing and review of the data, especially in the cases where there are major shifts in the rankings. Thus, this data will need to be prepared in a clear manner that differentiates the time factors of when the data was received in order to properly display the chronology of the information for comparisons.

Testing and evaluation will be performed on the basis of the individual months in 2020 to the same months in years past in order to minimize any additional factors that could account for increased variability in the data, such as holidays or seasonal sales. In addition, all such evaluations will be performed on a basis of the either same game or genre being compared against itself, or as a comparison of all games in the specified time periods against themselves. Additional research may be required to determine if any unique factors – such as the release of a highly anticipated game – has any effect on the results at specific periods of time as well as methods to counteract those issues.