**Capstone Report**

Jelly P. Edwards

**Intro:**

This will be a report discussing topics that on the data science process that I will take for the project. For this project, I wanted to examine and focus on a subject matter that I was interested in. I decided that I wanted to analyze data that focuses on the live-streaming and video game industries. I have always been passionate about video games, and I am also a contributor and consumer in the live-streaming industry. In live-streaming, users record and broadcast live video of their gameplay for others to watch and commentate on. Live-streaming became very entwined with the video game industry in the past decade or so and has since become integral in gaming culture with barely any signs of leaving. The most popular platform for live-streaming is Twitch, which focuses mostly on video games but includes other genres such as music, art, and much more.

I wanted to analyze Twitch data and Metacritic data to answer various questions that I have about live-streaming. Are there games that viewers are more likely to watch that contributors are not streaming as much as others? As a user who has contributed to Twitch myself, I know that it is difficult to receive and keep viewership while broadcasting my gameplay. It can be assumed that games that fit the scenario mentioned above would have high probability to increase viewership. Are there games that have increased or decreased viewership over the years? Does a video game’s score from a review site influence how popular that game is to watch or stream on Twitch? Based on that review, can we predict if that game will be popular to watch or stream? These are questions I hope to answer while I examine the data.

**Github and Python Information**

My project will have its files and Jupyter-Notebook in a version control repository in GitHub. This will help manage any changes needed to the project, and will provide a platform for the project to live in. I will be using Jupyter-Notebook to interact and analyze the data. In this notebook, I will various python libraries to help manipulate and examine the data. I will use Pandas, Pandas\_profiling, Matplotlib, Numpy, and Seaborn libraries to solve the problem stated. Other libraries may be added later in the project, but these are the main 5 libraries that I know I will need for this analysis.

**Data**

I have chosen two data sets that I found on Kaggle for my analysis. The first data set is Ran Kirsh’s dataset called, “Top Games on Twitch 2016-2021”. I am using the CSV on Twitch Game data. This data provides viewership and streaming observations for the many games and categories that are in Twitch per month in 2016-2021. The second data set is Deep Contractor’s dataset called, “Top Video Games 1995-2021 Metacritic”. This dataset lists the ratings for numerous video games from the Metacritic website, which is a popular site for reviews on media. It provides both the Meta Score – a weighted average of professional critic reviews - and User Score – the average rating from users who rated the game on the Metacritic site. Links to both Kaggle datasets are provided in the Jupyter-Notebook. The data needs to go through the pre-processing and cleaning process before analysis. The two datasets will also need to be merged based on Video Game name to perform a full analysis later.

**Data Science Processes**

The data science process needs to be defined early on to help guide the project in the best direction possible. Once the data science process has been defined, I will begin the exploratory data analysis. This will bring insight on the data and hopefully answer a few questions that were originally proposed. It will also help identify any features that can be used during the machine learning process. Once identified, I will use those features to begin the predictive modeling. I will use cross validation to identify the best algorithm for the model, and then create the model. The model will be evaluated and tested to see if it truly answers the questions stated.