Project : Video Game Data ETL Process

**Team Members:**

* Moore, Ashley
* Johnson, Joseph
* Patel, Vinika
* Bedwell, Jenny
* Benedict, Robert
* Bilal, Mohamed
* Zhu, Valentina

**Project Description/Outline:**

The team project focuses on creating a dataset following the extract, transform, and load process (ETL). The team plans to upload the data into the jupyter notebook via pandas, merging the two csv files, cleaning and converting to JSON. Once cleaned and converted, the JSON file is then loaded into MongoDB.

**Datasets/Sources:**

* Metacritic: Top Video Games 1995-2021

https://www.kaggle.com/datasets/deepcontractor/top-video-games-19952021-metacritic

* VGChartz: Video Games Sales Dataset

<https://www.kaggle.com/datasets/gregorut/videogamesales>

**Analysis:**

The purpose of this group project is to extract, transform, and load data into a SQL/NoSQL database. The sources (refer to links above) of this data is based on video games of all platforms from 1980 to present. First, the two CSVs were imported into a Jupyter notebook utilizing Pandas. The two CSVs (video game sales and metacritic reviews) were merged into one Pandas dataframe. Next began the cleaning processes such as renaming columns, converting columns to the respective data types, and checking for null values. However, there was one issue that the group had to overcome. The issue was User\_Review column returned as a string type instead of a float. To resolve this issue the team checked for duplications which returned false. The team then searched for a string value by sorting the rows using the .head() and .tail() functions to check the first and last rows for any odd values. It was discovered that a few of the rows under User\_Review column had a string value “tbd” instead of a float. To finalize the cleaning process, rows of “tbd” were queried and dropped then converted the column into the appropriate data type. As a bonus, the team decided to split up the release date column (yyyy/mm/dd) splitting into three separate columns individually by year, month, and date. Lastly the columns in the dataframe were regrouped respectively. The final Pandas dataframe then was converted into JSON format as a dictionary. As the final step in the ETL process, a client connection was created using pymongo and loaded Pandas dataframe into MongoDB.