**ETL Project Report**

*Group 5: Christine Mitchell, Byron Allen, Kyle Peterson, Nick Allen*

As a group we were tasked with finding data of multiple sources which we could perform an ETL analysis on. The first file we agreed upon using was a SQLite database containing level data from over ##### Super Mario Maker Levels. After looking into our first dataset we decided to find country information to assist in analysis on the popularity of Super Mario Maker in various populous nations. For that country information we were able to find a CSV on Kaggle containing the population data we were after along with various rates that, while interesting, would not really factor into our envisioned analysis. Lastly, we realized that our two datasets had no direct way to be joined, so some sort of “bridge” would be needed to connect the two. We were luckily able to find a JSON file containing the two letter country codes present in our SQLite database and the full country name contained in our CSV information.

The first task to tackle was “E” - Extract. Utilizing the pandas library we were able to directly bring the different data sets into jupyter notebooks for further use. Extraction presented no real issues.

Once our files were read in it was time for the “T” step: Transform. Both of our main data sets (SQLite and CSV) contained more data than was entirely necessary for our purposes. For the Super Mario Maker SQLite database we first needed to combine the two tables the file contained. Once that was accomplished the cleaning began. The combined data frame was trimmed down to hold just the country codes along with the associated medal totals, total number of plays, and total level count for each