Pokemon ETL Project

Ty Pleasant, Eli Souvannarath, Sarah Cullen

9 December 2021

We chose to put together a database about Pokémon. The franchise has spanned so many years, different games, and shows that it is sometimes hard to keep up with all the data across all these platforms, when you are only interacting with a few of them. This database allows the user to have all the stats at their fingertips.

           Our database has two sources in the form of csv files that we found on Kaggle.com. The first dataset is called World of Pokémon by creator tensor. This dataset is the smaller of the two that we used, and its information is very concise and simple. It outlines the basic stats for each Pokémon as well as generation and legendary status. The second dataset is Complete Pokémon Dataset (Updated 16.04.21) by creator Mario Tormo Romero. This dataset goes much more in depth, and while it has good information, it is easy for a viewer to get lost. It contains much more in-depth stats related to game play and species description, as well as the names for the Pokémon in multiple languages.

           There was not much to clean in the World of Pokémon dataset. We kept as much of that set as we could, but we had to take out some of the secondary values for abilities since not all the Pokémon had values for that column, but every Pokémon had a primary ability listed. All 800 Pokémon listed in the original csv were kept for the database.

           The Complete Pokémon Dataset proved a bit more difficult to wrangle into a usable format. There were quite a few columns that did not have values for all the listed Pokémon (there were 1045 in this set). We made the choice to leave out any column with more than 5 missing values. The only one that was cut which was close was catch\_rate with 18 missing values; the rest were missing several hundred. We were trying to keep as many in as possible to be able to have as much cross over with the other dataset. In the end we ended up dropping only 4 rows, so this dataset contained 1041 Pokémon.

           We decided to use the relational database, SQL, for our project. We felt that it allowed us more ability to do certain analysis and observation. Since all Pokémon are fictional creatures created for a specific purpose, they tend to have very similar characteristics, because that is what is required to have them work in their games and shows. Because of these similarities, it is helpful to see them in relation to each other. Each dataset was made into a table in the database: World of Pokémon is pokemon\_new and Complete Pokémon Dataset is pokemon\_dex.

Our group did some sample joins that are included as possible ways that one could use this database (images are included in the repository so show these sample joins). Since there is some overlap in the simple statistics of the datasets, one must be careful to only call the column names that match up with the table they wish to pull from.

This database could be used for multiple reasons. As a basic fan, one could use it to look up statistics for the various specific Pokémon by name, but it could also be used for more in depth analysis. Two of the simplest routes to take would be to look at comparisons of Pokémon across generations and types. With so many numerical statistics it would be simple to run analyses across these groups.