The purpose of the ETL pipeline we built was to compare box office data of movies and their respective IMDb ratings to find if there was a correlation between how a movie was critically acclaimed and its success in generating revenue.

The biggest challenge was gathering reliable IMDb ratings that were accurate. We first tried to use OMDb API, but it did not allow for a large amount of requests. This made it difficult to reach an accurate conclusion. We also tried web scraping, but IMDb's changing HTML structure caused trouble in terms of consistency. We then found IMDbPY, which utilized objects for the movies. However, there were issues with aligning the titles from the CSV file to the IMDb search results, which we solved by matching the title of the movie and its release year to guarantee that we retrieved the correct rating.

Once the data was retrieved, it was then cleaned and merged. The pandas and seaborn libraries made data analysis and visualization quite straightforward, with the exception of learning the functions. Correlation and summary statistics did not require many lines of code like we thought they might, and converting data between output formats using pandas was also easier than expected.

Retrieving accurate ratings proved to be the most complex. There were movies with remakes, duplicates, or even television shows with the same names, which made it difficult to match movie titles to their accurate IMDb entries. IMDbPY looks for movie objects by searching by the name of the movie, but sometimes the first entry that would show in the search would not be the desired movie, resulting in either an inaccurate rating or no result returned. Accurate data was then ensured by not just requiring a match in titles through the search but also a verification of matching years.

This project could be useful in that we learned to merge two different types of datasets into one, with the output varying in file type depending on the user's desires. The use of libraries and functions to visualize and manipulate the data is certainly a skill that will prove to be useful in future data projects. Looking for APIs that fulfilled the desired functionalities for the program was an interesting task that required trial and error. In conclusion, the project helped develop reusable skills such as data cleaning, transformation of databases, and merging datasets.