ETL Project

Extract:

Data was pulled from the website <https://www.the-numbers.com/movie/budgets/all>, which provides a table of the most expensive films produced, along with their release date and gross income. This was accomplished by implementing splinter to move through the different pages and BeautifulSoup to scrape the table on the html page. This data was stored into a pandas DataFrame. To get the ratings movies received, the OMDB api was used. The names of the movies from the web scraping were passed to the api to get this data. The ratings were stored as a list and then added into the pandas DataFrame.

Transform:

Some of the movies during the api call did not return a value. These movies were dropped from the DataFrame. The production budgets, domestic and worldwide gross, and the ratings from IMDB, Metascore, and Rotten Tomatoes were converted from a string to a float type. The budget and gross income columns needed the ‘$’ symbol and ‘,’ removed from the columns first. For the ratings, the IMDB column needed to remove ’/10’ and Rotten Tomatoes column needed to remove ‘%’ and in some cases ‘/100’. After converting the IMDB ratings column, the column was multiplied by 10 to be more inline with the other rating columns (Metascore and Rotten Tomatoes are out of 100).

Load:

Once the cleaned DataFrame is created, the data was loaded onto MongoDB using Pymongo. After establishing a connection to the database, the DataFrame was looped through to create a dictionary for each movie using the .to\_dict() function from pandas and then inserting it into the database using Pymongo. The created database was then exported as a json file ready for analysis.