Project 2 ETL – Zhuoli Ma, Rana Badri, Alysia Dugan

Extract:

<https://www.rottentomatoes.com/top/bestofrt/?year=2021>

<https://www.rottentomatoes.com/top/bestofrt/?year=2020>

<https://www.rottentomatoes.com/top/bestofrt/?year=2019>

<https://www.the-numbers.com/market/2021/top-grossing-movies>

<https://www.the-numbers.com/market/2020/top-grossing-movies>

<https://www.the-numbers.com/market/2019/top-grossing-movies>

Transform:

* On “The Numbers” website, we found the quickest way to pull and clean the data was to copy and paste the data into Excel and convert to CSV file. This also allowed us to quickly filter for films from prior years that were re-aired in our time frame and remove them.
* On the Rotten Tomatoes webpage, we used pandas\_read\_html to load the data into Jupyter Notebook.
  + Removed unnecessary string on the movie’s title using “replace” function.
  + Exported to CSV and subsequently appended the files together.

Load:

* Used Jupyter Notebook to connect with pgAdmin via sqlalchemy
* Created box\_office and rotten\_tomatoes tables in pgAdmin under a “movies” database
* Created two SQL queries to test that the tables were working correctly