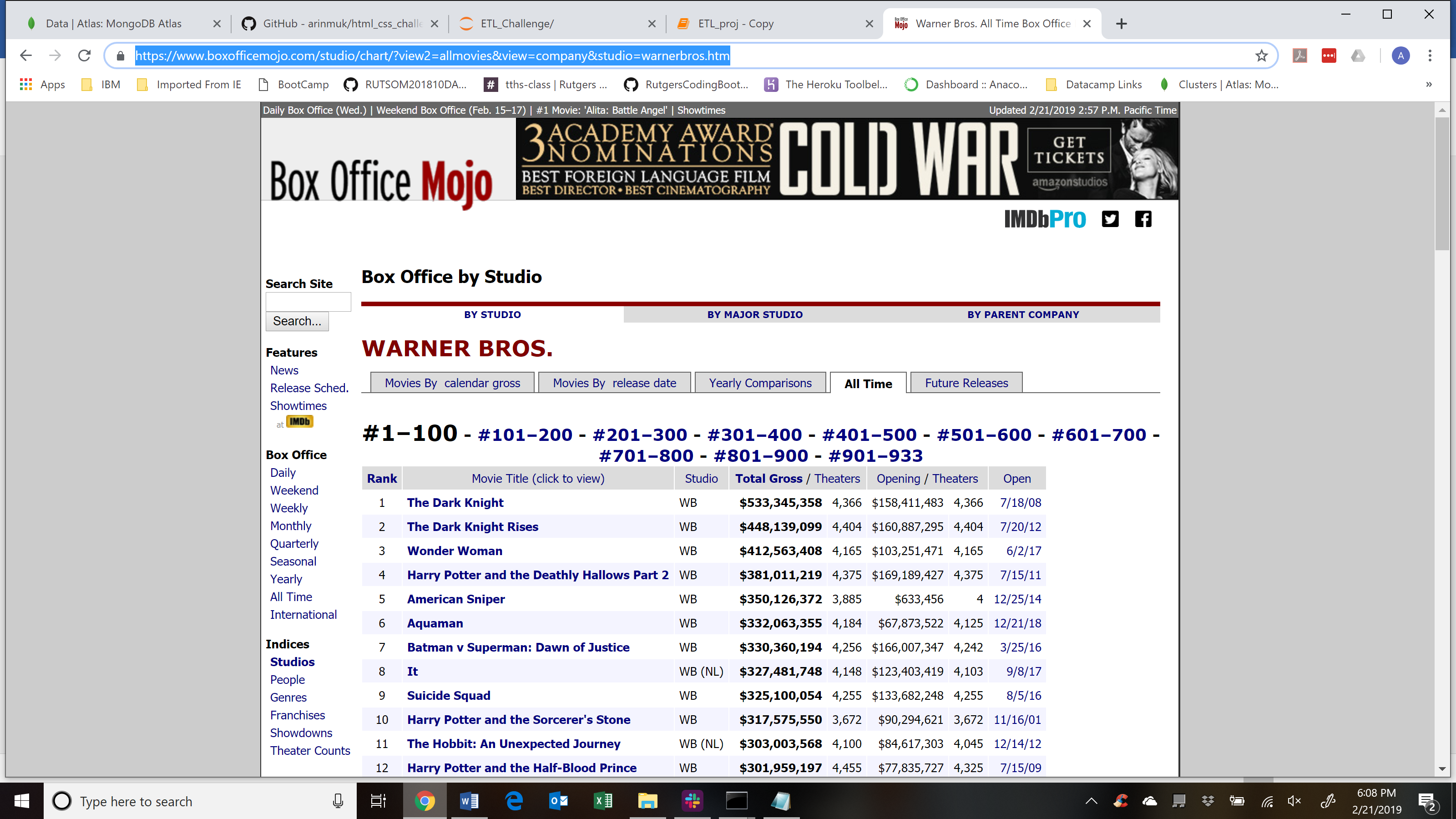
ETL Project

Data source 1: Panda scrapping a website for data.

The purpose is to collect the top grossing movies for each of the studios. The site provides information abut the highest grossing movies by studio as seen in the screenshot below



<https://www.boxofficemojo.com/studio/chart/?view2=allmovies&view=company&studio=warnerbros.htm>

The fields returned are Rank, Movie Title, Studio, Total Gross, Theater Count, Opening Day gross, Opening Day theater release count, and date it was released first. Use of splinter was not possible since each of the studios had 100s of releases over several years. Each page per studio has a hundred movie records and there were several pages of records for each studio as is also seen above in the screenshot.

The loop to scroll through the pages was set to 15 , expecting 14 pages per studio since there was no way to find out all the pages as I was not using splinter but a dynamic page number build.

The problems created extra data into the dataframe that needed cleaning. The data was also required to be cleaned in the mongo for nulls(in Movie Title value) being entered due to this. One the data came in the dataframe per page , I converted it( 100 rows for each page) into Jason and inserted them into Mongo. Once the full scrapping for all studios was completed, mongo data had to be cleaned up as noted above.

Data Source 2: Api Call to OMDB for each movie stored in Mongo to capture Rating, votes and runtimes

Using the OMDB API call , I added three more fields for each of the movies that can be used for aggregation purposes : Rating, Number of votes and Runtimes..

The final data was stored in Mongo and the output required for homework was stored in the excel format as required from the Mongo DB