# ETL Project Report

## By Chris and Pam Zhao

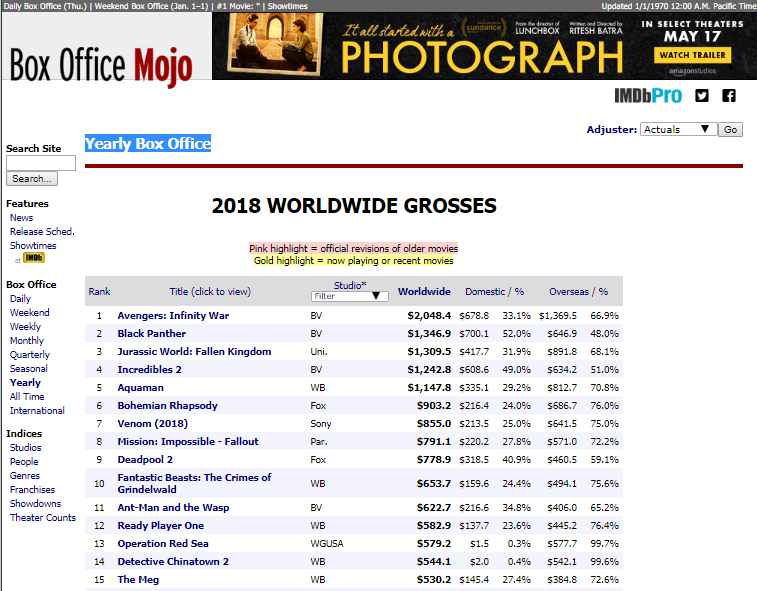
We decided to look at all 2018 movie data. We found two data sources, an API has movies basic information and a box office website has the movies grosses.

Pam got a list of movies titles by scraping the box office webpage, and Chris used API to pull the movies information based on Pam’s list.

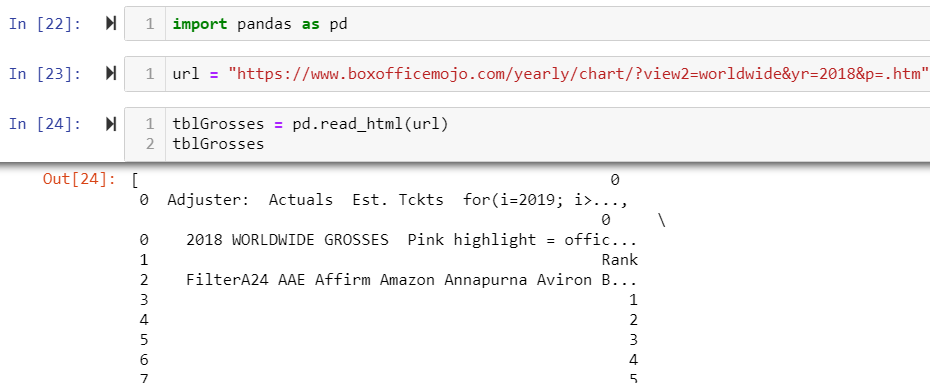
**Extract Data**

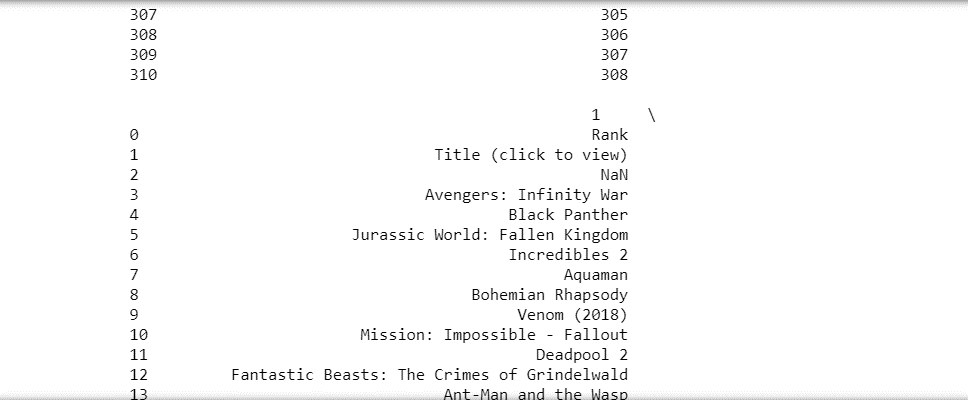
Scrape Box Office Mojo webpage – Pam Zhao

<https://www.boxofficemojo.com/yearly/chart/?view2=worldwide&yr=2018&p=.htm>

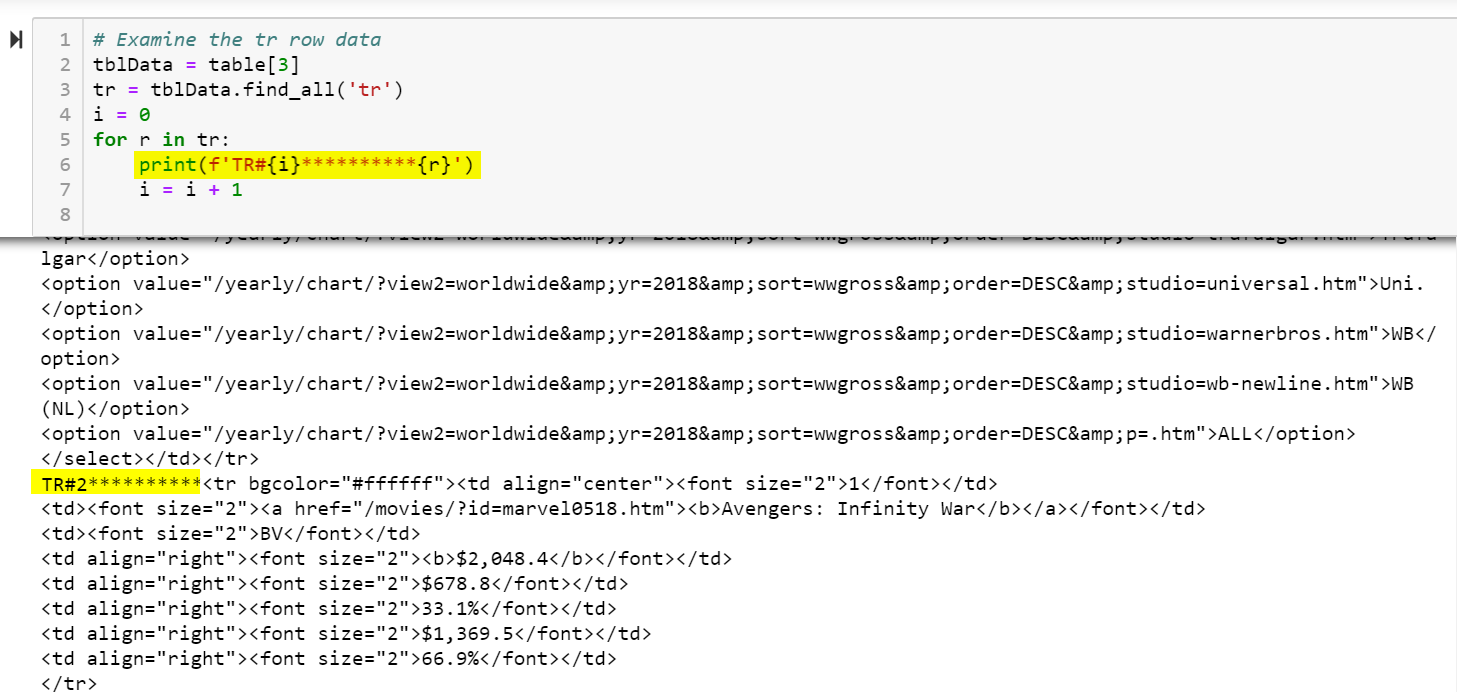


When I saw the table on the website (see above screenshot), I thought it would be very easy to scrape the table by using Panda scraping because there is only one big table in the webpage. However, the format was off a lot when I used Pandas scraping.





Then I inspected the HTML for the webpage and decided to use BeautifulSoup to scrape the table. There are more than one table on the webpage, and I was able to locate the table that has the movies grosses data by going through each table data. I examined the table by printing out the table row number in order to determine where to start to get the data.



I iterated through the html table and append the data into a Pandas DataFrmae. Then I saved the DataFrame to a CSV file. My teammate, Chris, used my list of movies title to pull the data from OMBD API.

**Transform Data**

**Load Data**