Summary

To start this project we initially important dependencies for the following three processes :

**Interact with data**

* PD
* Numpy

**Data Scrapping**

* BeautifulSoup
* Splinter
* Time

**Databasing**

* Python
* MongoDB
* Json

After establishing the proper connections and importing all the data from the proper API’s used:

* [www.omdbapi.com](http://www.omdbapi.com/)
* [www.boxofficemojo.com](http://www.boxofficemojo.com/)
* [www.the-numbers.com](http://www.the-numbers.com/)

We had to properly pull the data off the API’s using for loops to iterate through all the pages and pull all data off. We then renamed some titles for prepare a more readable file and ultimately JSONify the file to prepare for upload to mongoDB.

Running into a bit of a problem with [www.the-numbers.com](http://www.the-numbers.com/) being that it is a secured API. We ended up using splinter to mask our connection as a browser connection to continue working with this API and our data. Pulling the data from all the API into two ultimate selections we had to eventually use the merge function to merge all three data sets on the movie title.

**Cleaning up**

Now came the part to clean up all our data. To start we did a quick out put to see what the data looks like in a data frame and then selecting the specific columns that we needed for out objects. Selecting (Title, Production Budget, Worldwide, Domestic, Foreign, Year, Awards, Genre, Production, Rated, imdb rating.) and renaming the columns to make it look a bit neater.

Dropping the NA’s and making the data more presentable we did a count at the end of the data to see how many items are in each column to confirm each record has data across. To end it we stored the output in a CSV and saved it