**Members:**

* Alex Burch, David J. Brewer

**Data:** [Movies Dataset](https://www.kaggle.com/rounakbanik/the-movies-dataset#ratings.csv) (multiple csv files)

* [Ratings\_small.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#ratings_small.csv)
* [Links\_small.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#links_small.csv)
* [Movies\_metadata.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#movies_metadata.csv)

**Subject:** Combining movie data with ratings to analyze revenue

**EXTRACT:**

* Columns: Budget, id, title, revenue, vote\_average, rating from the following files
  + [Ratings\_small.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#ratings_small.csv)
  + [Links\_small.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#links_small.csv)
  + [Movies\_metadata.csv](https://www.kaggle.com/rounakbanik/the-movies-dataset/version/7#movies_metadata.csv)

**TRANSFORM:** (Data cleaning)

* Reduce indexes
* merged tables, dropped unwanted columns, dropped columns with missing info and duplicates
* Intended to use SQLite but ended up utilizing MongoDB
* Final table represents unique list of movie titles sorted by rating
  + Initial data had lots of duplicate movie titles, we used the drop function from pandas to remove the duplicates. This can be improved by merging the duplicate titles instead of dropping them as we would not lose the data from the other entries.
  + Then we drop useless columns, drop all zeros in the ‘revenue’ column, and sorted with a ‘nlargest’ command
  + Then, we created two html tables for top twenty revenue and top twenty rating, this way they can be uploaded onto a mongodb collection

**LOAD:** - Final table includes

* The final tables are two different tables, both their own collection on mongodb and both a top twenty list. The list includes columns such as ‘revenue’, ‘movie\_title’, ‘movieID’ (which was our joining column for information in our csv(s))
* Finally, mongodb database was created called “Top20Movies” with two collections; “Top20MoviesRevenue” & “Top20MoviesRating”. (Although their names on mongodb are just collection & collection2”.
* In conclusion, please see our jupyter notebook & mongodb screenshots for references