# **Preparing, Exploring, and Cleaning up the Dataset**

We started by getting the names of over 4000 movie titles from the TMDB 5000 Movie Dataset from Kaggle. Next, we extracted information related to these movie titles using the API from the OMDB website (<url:https://www.omdbapi.com>). The Open Movie Database API allows you to search movies by title and returns JSON items for Title, Release Year, IMDB ID, Type (movie, series, or episode), and Movie Poster image. Within each of these JSON items, we extracted values from the following columns to arrive at the final data that we wanted to analyze and saved it into a csv file:

1. Title – Movie Title
2. Year – Movie released year
3. Rated – MPAA viewer rating
4. Runtime- Duration of Movie in minutes
5. Genre – movie genre
6. Director – director (s) of movie
7. Writer – movie writer (s)
8. Language – spoken languages in movie
9. Country – country/countries of production
10. Awards – awards received
11. IMDB Rating (score out of 10)
12. Revenue – movie revenue \*
13. Budget- movie budget \*

Note: Items 12 and 13 were extracted from our original TMDB Kaggle dataset which we merged with the data set we created with our API calls.

Some of the steps we took to clean the final CSV file included:

1. Dropping columns with a significant amount of missing data.
2. Reviewing columns for inconsistent data and either dropping or converting the data to the desired data type.
3. Filling missing data with either NA or 0 depending on the type.
4. Exporting a final cleaned .csv file.

For a more detailed explanation of the clean-up process, please see the project Jupyter Notebook.

~~There were also certain rows that contained lists of information, such as the “Genre”, “Language Columns,” “Country,” and “Awards.” The items in these lists needed to be parsed and stored in a larger list so that their values could be counted and otherwise manipulated.~~

Comment: Not sure if the above applies anymore since we didn’t use the Kaggle data set other than the Revenue and Budget information.