

**Oscar Winning Movies and Rotten Tomato Ratings**

**ETL Project**

**Technical Report**

Group Members: Melissa Agruda, Zachary Steele, Iretemi



Objective

Finding Data

3 original sources.

1. <https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv>. This file is a csv.
2. <https://data.world/crowdflower/academy-awards-demographics>. The file pulled from this link is a csv.
3. <https://www.rottentomatoes.com/browse/opening/>. The file pulled from here was obtained using splinter to web scrape the movie title and ratings which was subsequently transferred to a SQL table.

Data Cleanup and Analysis

1. We transformed the Oscar file by using and renaming 4 columns, “movie”, “award”, “birthplace”, and “date\_of\_birth”. We also dropped possible duplicates from “Movies” and set “Movies” as the index.
2. We transformed the Movie file by using and only keeping 4 columns, “homepage”, “release\_date”, “title” and “popularity”. We also dropped possible duplicates from “title” and set “title” as the index.
3. Obtaining Rotten Tomato movie title and rating
4. Joining 3 tables and loading into Pg-admin and postgresql database
5. Joined the 3 tables to analyze the Oscar winners, the movies that won and the rotten tomatoes ratings those movies received.

# Guidelines for ETL Project

## Finding Data

Your project must use 2 or more sources of data. We recommend the following sites to use as sources of data:

\* [data.world](https://data.world/)

\* [Kaggle](https://www.kaggle.com/)

You can also use APIs or data scraped from the web. However, get approval from your instructor first. Again, there is only a week to complete this!

## Data Cleanup & Analysis

Once you have identified your datasets, perform ETL on the data. Make sure to plan and document the following:

\* The sources of data that you will extract from.

\* The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

\* The type of final production database to load the data into (relational or non-relational).

\* The final tables or collections that will be used in the production database.

You will be required to submit a final technical report with the above information and steps required to reproduce your ETL process.

\* \*\*E\*\*xtract: your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc).

\* \*\*T\*\*ransform: what data cleaning or transformation was required.

\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.

Please upload the report to Github and submit a link to Bootcampspot.