

**Oscars, Ratings, & Films . . . Oh Py!**

**Project Team:**

**Zehra Tokatli, Chase Webb, Lauren Parrish**

**Table of Contents:**

[Project Objective: 3](#_Toc66968947)

[Extraction: 3](#_Toc66968948)

[Transformation: 4](#_Toc66968949)

[Loadings: 4](#_Toc66968950)

[Conclusions: 4](#_Toc66968951)

# **Project Objective:**

It seems like a bygone activity of yesteryear, but yes people did go to movie theaters. With the advent of streaming services and well Covid, most of us are enjoying film from the comfort of our couch. We have many movie options! How do we know if what we select is “quality” and if something is “quality” is it something the average streamer really wants to watch?

We extracted data from an Academy Awards Database and a Streaming Services database to try and determine:

* Do award winning films actually have high user ratings?
* Do awards impact film popularity on streaming services?
* Do award winning directors receive higher ratings?
* Do films with winning soundtracks have higher popularity?

# **Extraction:**

The first dataset was extracted from Kaggle.com in CSV. This data included details on movie titles, user ratings, streaming services supporting the film, director, and release year, genres, and runtime.

The team leveraged DataHub for the second dataset extraction in JSON format. This data set provided details year, award category, winner (T/F), and winner name.

Both datasets were downloaded and imported into Python for cleaning.

* https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney
* https://datahub.io/rufuspollock/oscars-nominees-and-winners

# **Transformation:**

The Streaming dataset contained 15 million rows of film data ranging across genres and four streaming service providers. The greatest challenge was that several cells contained multiple data points separated by commas. Some of these cells were not applicable for the team analysis. The decision was made to drop any additional data elements post the primary and secondary listing. The exception being the genres information which was split and added to eight new genre columns. The genre split alone took an hour to run.

**Summary of Streaming Data Cleanse:**

* Drop empty columns;
* Remove rows that did not contain rating information;
* Split First and Second Director, drop any additional director information;
* Convert streaming services cells to Booleans, then merge with main data frame. Drop redundant columns and remain headers as appropriate;
* Split Location cell into Primary and Secondary Location, drop other locations;
* Split Language into Primary and Secondary Language, drop other languages;
* Split Genres into eight separate columns:
* Validate column order for ease of use.

**Awards Data Cleanse:**

**Data Joins:**

**Data Analysis:**

# **Loading:**

# **Conclusions:**