

Project 2

ETL Challenge



October 28, 2021

*Group 3: Ivana, Namita, Prabhdeep, Sean*

## Objectives

As a team of analysts for a new start-up streaming service for movies, we would like to explore the relationship between Golden Globe winners and IMDb ratings to help inform on which movie rights to chase. We aim to explore whether public perception (IMDb ratings) of movie quality and the officially given status of a movie (Golden Globe winners) are in line with one another.

## Data Sources

The data sources were obtained from Kaggle.com.

## Database Preparation

An ERD diagram of the tables to be developed was created on quickdatabasediagrams.com as per Figure 1 below.

Graphical user interface, application, Teams

Description automatically generated

Figure - ERD diagram for movies\_db

The export function was used to create the file table\_schema.txt and run in PostreSQL to create the tables in a new database called movies\_db. PostgreSQL was used due to the relational nature of our data.

## ETL Process

The code for the ETL process can be found in movies\_etl.ipynb.

The cleaned data in csv format is stored along with the raw data in the data folder.

### 4.1 Extract

The data sources are shown in Table 1 below with the files downloaded from each in CSV format.

Table - Data sources

|  |  |
| --- | --- |
| Data Source 1 | |
| Kaggle.com – IMDb (Movies/Ratings)  <http://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv> | |
| **File 1:** | IMDb movies.csv |
| **File 2:** | IMDb ratings.csv |
| Data Source 2 | |
| Kaggle.com – Golden Globe Awards  <https://www.kaggle.com/unanimad/golden-globe-awards> | |
| **File 3:** | golden\_globe\_awards.csv |

### 4.2 Transform

The three CSV files were imported into a Jupyter Notebook into Pandas data frame format and the following transformations conducted for each.

**IMDb movies.csv**

1. Created a filtered data frame with only the desired columns.
2. Re-named the column headers to match those in the ERD diagram and the newly created database movies\_db.
3. Cleaned the data by dropping the duplicates and null values.
4. The imdb\_id was set as the index.
5. The new transformed movies data was then saved in the data folder as IMDb\_movies\_transformed.csv.

**IMDb ratings.csv**

1. Created a filtered data frame with only the desired columns.
2. Columns were renamed to be consistent with the ERD table and the database movies\_db.
3. All rows with any null values were removed from the dataset.
4. The new transformed ratings data was then saved in the data folder as IMDb\_ratings\_transformed.csv.

**Golden\_globe\_awards.csv**

1. Created a filtered data frame with only the desired columns.
2. Columns were renamed to be consistent with the ERD table and the database movies\_db.
3. The data was cleaned by dropping rows with null values.
4. The new transformed golden globes data was then saved in the data folder as Golden\_globe\_transformed.csv.

### Load

A connection to the database was created and the three data frames were then successfully loaded into their respective tables in Postgresql. Namely, the data is stored in 3 tables in Postgresql:

* imdb\_movies,
* imdb\_ratings, and,
* golden\_globe.

## Visualisations

The source file for the following visualisations code is contained in movies\_etl.ipynb.

Chart, bar chart

Description automatically generated

Figure - Films that have won 5 or more golden globe awards

Chart, bar chart

Description automatically generated

Figure - Films that were nominated more than 15 times for a golden globe award

Chart, bar chart

Description automatically generated

Figure - Golden globe film nominations with highest total votes

Chart, bar chart

Description automatically generated

Figure - Golden globe nominations with highest critic reviews

Chart, bar chart

Description automatically generated

Figure - Films with highest IMDb ratings and >1 golden globes win

## Conclusions

Overall, our project was a great success returning some interesting findings. For instance, looking at the data, the top grossing films are not necessarily the top-rated movies on IMDb with only three of the top 20 highest grossing films in the top 20 rated films on IMDb. It was also discovered, that out of the top 20 rated films on IMDb, 11 of them also won golden globes. However, only one of those golden globe winners came in between 11-20th. This suggests that after the top handful of films, professional and public perception may start to diverge meaning we may want to chase rights for films that are highly rated by the public rather than the industry.