# Project 2: ETL Challenge

*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 source was Kaggle.com with the data in CSV format.

## Database Preparation

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

Graphical user interface, application

Description automatically generated

The export function was used to create the file schema.sql and run in PostreSQL to create the tables in a new database called movies\_db.

## ETL Process

### 4.1 Extract

The data sources were as follows with the files downloaded from each in CSV format.

1. Kaggle.com – IMDb (Movies/Ratings)

<http://www.kaggle.com/stefanoleone992/imdb-extensive-dataset?select=IMDb+movies.csv>

**IMDb movies.csv**

**IMDb ratings.csv**

1. Kaggle.com – Golden Globe Awards

<https://www.kaggle.com/unanimad/golden-globe-awards>

**golden\_globe\_awards.csv**

The CSV files of the raw data can be found in the data folder.

### 4.2 Transform

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

**IMDb movies.csv**

1. Removed all unnecessary columns leaving the data with the three relevant columns: imdb\_ID, imdb\_rating and total\_votes.
2. Columns were renamed to be consistent with the ERD table.
3. All rows with any null values were removed from the dataset.

**IMDb ratings.csv**

**Golden\_globe\_awards.csv**

### Load

## Visualisations

## Conclusions