**Project 2 Proposal**

Project team 1

(Chisimnulia, Sofia, Shannon)

**Netflix film and TV show analysis**

Our project focuses on two datasets sourced from Kaggle. We chose these datasets as we love to watch Netflix ... but we had run out of tv shows and films to watch. We wanted to see how many of IMDBs top 1000 films and tv series were on Netflix. Surely the most rated and top grossing films are the most interesting? So we also wanted to know which films/tv shows and how many were not on Netflix.

**Datasets**

We chose these datasets because we thought it would best illustrate what we wanted to find.

Netflix Movies and TV Shows

<https://www.kaggle.com/datasets/shivamb/netflix-shows>

IMDB Movies Dataset

<https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows>

**Research Questions:**

How many of IMDB’s top 1000 films are currently on Netflix?

What is the corresponding IMDB score for these films, has Netflix missed any major top rated films?

What release year are most common in IMDB’s top 1000? Possible suggestions for films to be added next month?\*

**Extract**

2 datasets csv files

**Transform & Load**

Left joins: flag the 1000 and name the nulls? Or join on film/show name?

* Would involve a lot of cleaning - may need to eliminate dates from titles.

Load to SQLpostgres

**Report**

* Loaded onto a website?

**Project task allocation / timeframes**

Examining Data - Tues-Thurs

Create Github repo - Wed

ETL - Thurs

Report - Sat