NETFLIX MOVIES AND TV SHOWS

DATA SOURCING

The data for this project is open-source and can be downloaded from the following link: https://www.kaggle.com/datasets/dgoenrique/netflix-movies-and-tv-shows

DATA COLLECTING

This dataset was intended to list all Netflix shows and movies. It was gathered from JustWatch in March 2023, and it contains data available in the United States.

DATA CONTENT

This dataset contains two files, one for the titles and the other for the cast of each movie and show on the platform.

DATA PROFILE

The "titles" file has 6137 rows and 15 columns, while the "credits" file has 81355 rows and 5 columns. In order to avoid creating several duplicates of the same movie or TV show for each actor and the director, I chose not to include the "credits" dataset in my analysis.

After quality checks, my dataset includes 6137 rows and 14 columns:

Index	Columns	Description	Time Variant/ Invariant	Data Type
1.	id	The title ID on JustWatch	Invariant	Qualitative
2.	title	The title of the movie/show	Invariant	Qualitative
3.	type	TV show or movie	Invariant	Qualitative
4.	release_year	The release year of the show/movie	Variant	Quantitative
5.	length	The length of the episode (if it's a show) or movie.	Variant	Quantitative
6.	genres	The genre (or genres) of the show/movie	Invariant	Qualitative
7.	production_countries	The country (or countries) where the movie/show was produced	Invariant	Qualitative
8.	seasons	Number of seasons (if it's a show)	Variant	Quantitative
9.	imdb_score	The score of the	Invariant	Quantitative

		series/movie on IMDB		
10.	imdb_votes	The votes of the series/movie on IMDB	Invariant	Quantitative
11.	tmdb_popularity	The popularity of the show/movie on TMDB	Invariant	Quantitative
12.	tmdb_score	The score of the show/movie on TMDB	Invariant	Quantitative
13	genre	The genre of the show/movie	Invariant	Qualitative
14	production_country	The country (or countries) where the movie/show was produced	Invariant	Qualitative

LIMITATIONS AND ETHICS

Limitations: The data set had a lot of missing information, therefore handling the incomplete data was fairly challenging. I decided to drop a potentially valuable column (age certification) since there were too many missing values (2743)

Ethics: in terms of ethics, the data set does not contain any kind of personal information.

QUESTIONS TO EXPLORE

1) Which titles have the highest ratings on IMDB?
2) Which titles garnered the most votes on IMDB?
3) What are top highest rated titles on TMBD?
4) What are the most well-liked titles on TMBD?