Netflix movie analysis

Project Information and data cleaning

- Data Source
- Data Profile
- Data cleaning steps
- Analysis Questions

Data Source

This data was collected by Shivam bansal, this person got the data from Netflix. The data seems really reliable to me, since a lot of other people already used it and the data origin is an official big company (Netflix). Additionally the data collector has a really professional profile on kaggle.

Data collection

The data was collected in the mid 2021. There is nothing said about how it was collected, but I reckon it either was scraped or netflix provided this data to him.

Data content

This data contains 8807 rows and has 12 columns. All of the Netflix TV shows and movies that were available in the middle of 2021 are listed in this tabular dataset.

limitations/ relevancy

This dataset is the only one I'm using in this analysis. This means it is crucial to this analysis. Limitation could be the use of only one streaming company in regards of geographic analysis.

Why this data?

To be very honest, the main reason I chose this dataset was that it met every requirement listed in the project brief. Although I had many ideas, none of the datasets met the requirements. Additionally, I enjoy Netflix a lot and am curious about some of its data.

Data Profile

Column name	Data Terminology			Description
show_id	Structured data	Quantitative data	Discrete data	Unique ID for every Movie / Tv Show
type	Structured data	Qualitative data	Nominal data	Identifier - A Movie or TV Show
title	Unstructured data	Qualitative data	Nominal data	Title of the Movie / Tv Show
director	Unstructured data	Qualitative data	Nominal data	Director of the Movie
cast	Unstructured data	Qualitative data	Nominal data	Actors involved in the movie / show
country	Structured data	Qualitative data	Nominal data	Country where the movie / show was
date_added	Structured data	Quantitative data	Discrete data	Date it was added on Netflix
release_year	Structured data	Quantitative data	Discrete data	Actual Release year of the move / show
rating	Structured data	Qualitative data	Ordinal data	TV Rating of the movie / show
duration	Structured data	Quantitative data	Discrete data	Total Duration - in minutes or number of seasons
listed_in	Structured data	Qualitative data	Nominal data	Genere
description	Unstructured data	Qualitative data	Nominal data	The summary description

Data cleaning steps

Mixed data types

I found 6 columns containing mixed data types: director, cast, country, date_added, rating, duration. The reason for them, are missing values.

null values

I found 6 columns containing null values: director, cast, country, date_added, rating, duration. I marked them and have to be aware of them.

duplicates

This dataframe does not contain any duplicates.

null value counts

director:	2634
cast:	825
country:	831
date_added:	10
rating:	4
duration:	3

Note: Kaggle already provided some basic statistics about the data which I additionally used to understand the data.

Analysis Questions

- Does Netflix has more focus on TV Shows than movies in recent years?
- Is there a trend in the country where the movies are produced?
- How long does it take for a movie to be published on netflix after it was published. Is there a correlation to another variable, like country or genre or director?
- Which genre has the best rating?
- Is there anything else interesting that might come up during the analysis

advanced Questions:

- Is there a pattern between the descriptions and a second variable, like rating, length or genre?