BIRMINGHAM CITY
University

Visualising Trends in Netflix's Content Library

DRAFT VERSION

Lewis Higgins - Student ID 22133848

CMP5352 - Data Visualisation

Word count: XXXX

Abstract

As of March 31, 2024, Netflix is the most popular television streaming service in the world (Nickinson, 2024), with over 269,000,000 active paid memberships. This report aims to analyse the library of content found on Netflix and identify key factors associated with the viewership of this content in the interests of furthering profit in the sector.

Contents

Introduction

Data visualisation is a field of data science wherein large datasets are parsed using code (most commonly written in Python or R) to produce clear visualisations interpretable to a wide audience, even if they do not have in-depth knowledge of the dataset.

This report specifically aims to produce visualisations based on an analysis of a large dataset based on the content library available on the Netflix online streaming service. The ever-evolving landscape of streaming services versus traditional television has positioned Netflix as a giant in the industry. Central to its success is a vast content library, encompassing a diverse range of movies, TV shows, documentaries, and more. This report delves into a comprehensive dataset of Netflix's content library, aiming to uncover valuable insights and trends. Through exploratory data analysis, the composition of the library will be examined, including factors such as content popularity by genre and release dates. This exploration will not only shed light on Netflix's content strategy but also provide potential indicators of current and future trends within the streaming service and the broader entertainment market.

This report is split across three sections:

- The **motivation and objectives** of this report.
- The **results from experiments** on the dataset.
- A **summary** of overall findings.

Motivation and objectives

The dominance of Netflix in the entertainment industry is undeniable; as the most popular online streaming service with over 269 million active paid memberships (Nickinson, 2024), it is essential to identify what they are doing correctly in the interests of furthering the industry and understanding the preferred content of their millions of subscribers.

The dataset in use is sourced from Kaggle, a public dataset-sharing website. It contains 8807 rows with twelve columns of data:

- show_id - A unique ID for each row of the dataset.
- type - "Movie" or "TV Show"
- title - The title of the content
- director - The director of the content
- cast - Actors featured in the content
- country - Country of production
- date_added - The date when Netflix added the content to the service
- release_year - The date when the content originally released
- rating - The age rating of the content
- duration - Duration in minutes (for movies) or seasons (for TV shows)
- listed_in - The genre of the content
- description - A large text description of the content.

1.1 Key questions conerning the data

- How has the content library grown over time?
- Which month of the year has the most releases?
- Which **content type** (movies / TV shows) is more frequent?
- Which genres are the most frequent?
- What is the age rating distribution of the content library?

Experimental results

2.1 Data wrangling

To be able to analyse the data, it is essential that the data is first in a state that can be analysed effectively. To do so, the data must be explored to identify any potential issues, and any that are found must be cleaned.

2.1.1 Data exploration

A good step to take in the initial exploration of the data is to identify if there are any missing values in certain columns. To do so, we can convert any instance of a blank string into R's recognised "NA" type and then use is.na() to identify how many there are.

```
# Convert blank strings to NA.
dataDf[dataDf == ""] <- NA

# Identify rows containing NA.
naRows <- dataDf[rowSums(is.na(dataDf) > 0),]
nrow(naRows)

[1] 3475

nrow(naRows)/nrow(dataDf)*100

[1] 39.46
```

We can identify that this dataset has 3475 rows where at least one column has missing data, equating to 39.46% of the dataset. To analyse this in further detail to see specifically which columns contain missing data, the naniar package can be used to generate a tibble of where the data is missing.

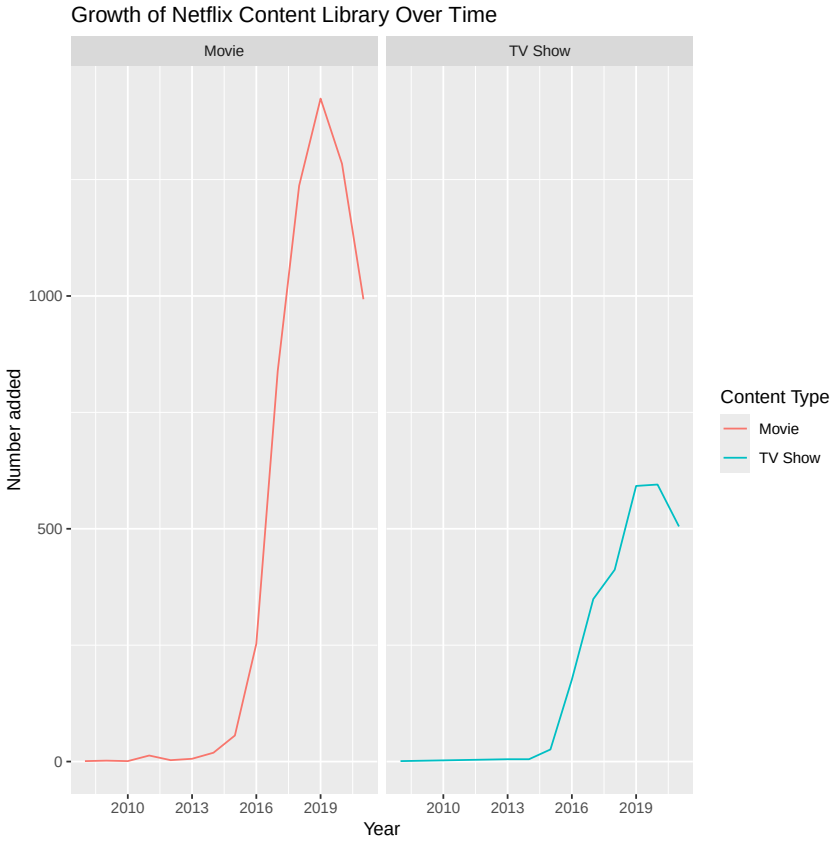
Table 2.1: Missing variable summary of the dataset

variable	n_miss	pct_miss
director	2634	29.9
country	831	9.44
cast	825	9.37
date_added	10	0.114
rating	4	0.0454
duration	3	0.0341
show_id	0	0
type	0	0
title	0	0
release_year	0	0
listed_in	0	0
description	0	0

```r
# Factorise categorical columns before the summary for cleaner output.
factorCols <- c("type", "country", "release_year", "rating", "duration")
dataDf <- dataDf %>%
    mutate(across(all_of(factorCols), as.factor))

# Convert date_added to a date format using Lubridate.
dataDf$date_added <- mdy(dataDf$date_added)


str(dataDf, vec.len = 1, width = 70, strict.width = "cut")

'data.frame': 8807 obs. of  12 variables:
 $ show_id     : chr  "s1" ...
 $ type        : Factor w/ 2 levels "Movie","TV Show": 1 2 ...
 $ title       : chr  "Dick Johnson Is Dead" ...
 $ director    : chr  "Kirsten Johnson" ...
 $ cast        : chr  NA ...
 $ country     : Factor w/ 748 levels ", France, Algeria",..: 604 42..
 $ date_added  : Date, format: "2021-09-25" ...
 $ release_year: Factor w/ 74 levels "1925","1942",..: 73 74 ...
 $ rating      : Factor w/ 17 levels "66 min","74 min",..: 8 12 ...
 $ duration    : Factor w/ 220 levels "1 Season","10 min",..: 211 11..
 $ listed_in   : chr  "Documentaries" ...
 $ description : chr  "As her father nears the end of his life, fil"..

# The 'cast' column contains Unicode characters that cause other miscellaneous
# errors later, so the Unicode characters in each row will be removed by
# running a gsub on the cast column to remove them.  dataDf$cast <-
# sapply(dataDf$cast, gsub, pattern = '<U\\+200B>', replacement= ' ')


# summary(dataDf)
```

2.1.2 Data cleaning

2.2 Visualisations


Growth of Netflix Content Library Over Time
Movie
TV Show
Number added
1000
500
0
2010
2013
2016
2019
Year
Content Type
Movie
TV Show

Summary and conclusion

aaaaa

Bibliography

Nickinson, Phil (Apr. 18, 2024). *The 10 most popular streaming services, ranked by subscriber count*. URL: https://www.digitaltrends.com/home-theater/most-popular-streaming-services-by-subscribers (visited on 04/23/2024).