BIRMINGHAM CITY
University

CMP5352 Report - TITLE NEEDED

DRAFT VERSION

Lewis Higgins - Student ID 22133848

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.

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.
The aim of this report is to analyse a large dataset containing data about Netflix's content library, identifying and visualising factors that have considerable influence over content viewership.
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

Netflix is a massive service used by hundreds of millions of people worldwide. Therefore, it is important to identify what they have done correctly, and how they optimize their content to maximise viewership, revenue and profit.

1.1 Key questions conerning the data

- Which month of the year has the most successful releases?
- Which **content type** (movies / TV shows) is more popular?
- Which genres are the most popular?
-

Experimental results

```r
# Replace all instances of 'NA' as a string or a blank string "" to NA.
# The strings 'NA' or " " and NA are two different things to R, and only the
# non-string NA is detected by functions like is.na().
# Identify rows containing NA.
naRows <- dataDf[rowSums(is.na(dataDf)) > 0,]
nrow(naRows)

## [1] 0

dataDf[(dataDf == 'NA' | dataDf == "" | dataDf == "NaN")] <- NA

# Output a summary of the data.
summary(dataDf)

##    show_id              type               title              director
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      cast              country           date_added         release_year
##  Length:8807        Length:8807        Length:8807        Min.   :1925
##  Class :character   Class :character   Class :character   1st Qu.:2013
##  Mode  :character   Mode  :character   Mode  :character   Median :2017
##                                                           Mean   :2014
##                                                           3rd Qu.:2019
##                                                           Max.   :2021
##     rating             duration          listed_in          description
##  Length:8807        Length:8807        Length:8807        Length:8807
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##

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

## [1] 3475
```

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).