Streamlined Insights: A Descriptive Analysis of TV Shows and Movies on Amazon Prime Video





**Course title:** Introduction to Artificial Intelligence

**Course code:** CMP4294

**Submitted By:**

Diya Kharel

Id: 24152363

**Submitted to:**

Dr. Mariam Adedoyin-Olowe

Module Leader

**Submission date**: 4th October, 2024

Table of Contents

[I. Introduction 3](#_Toc10631)

[A. Purpose of Study 3](#_Toc10632)

[B. Domain Description 3](#_Toc10633)

[C. Scope and objective of data analysis in streaming services: 3](#_Toc10634)

[II. Literature Review 3](#_Toc10635)

[A. Background study 3](#_Toc10636)

[B. Current and Future Trends 3](#_Toc10637)

[C. Research Gap 4](#_Toc10638)

[D. Critical evaluation 4](#_Toc10639)

[III. Methodology 4](#_Toc10640)

[A. Dataset Description 4](#_Toc10641)

[B. Model Analysis 5](#_Toc10642)

[IV. Dataset Analysis and Result 7](#_Toc10643)

[A. Exploratory Data Analysis 7](#_Toc10644)

[B. Clustering 10](#_Toc10645)

[C. Association Rule Mining 11](#_Toc10646)

[D. Analysing Descriptive column with word cloud 11](#_Toc10647)

[V. Conclusion 12](#_Toc10648)

[VI. Reference 12](#_Toc10649)

[VII. Appendix 13](#_Toc10650)

***Abstract —*** ***The streaming industry is experiencing significant growth, with Amazon Prime Video becoming a major competitor to Netflix. This study examines the content offered on Amazon Prime Video to reveal viewer preferences, identify emerging trends, and explore potential opportunities for content creation. We utilize basic machine learning techniques and Exploratory Data Analysis (EDA) to extract insights from the dataset. By employing methods like Rule Mining, we create tailored rules, and k-means clustering is used to uncover relationships among different attributes. Our goal is to deepen the understanding of the platform's content landscape and guide strategic decisions related to viewer engagement and content acquisition.***

***Keywords-*** *Amazon Prime Video, Descriptive Machine Learning, Word Cloud, Clustering, Movies and TV show, Statistical Description*

# I. INTRODUCTION

## A. PURPOSE OF STUDY

## The streaming industry has experienced significant growth, with Amazon Prime Video becoming a major competitor to Netflix. One of the key advantages for Amazon is its strategic use of data analytics to improve content recommendations, boost user engagement, and enhance the overall viewing experience (Parrot Analytics, 2022). By examining the vast catalog on Amazon Prime Video, we can gain important insights into viewer preferences, emerging content trends, and potential opportunities for new content creation.

## This paper uses descriptive analysis and machine learning techniques to explore the dataset, uncover patterns, and derive insights through statistical methods and Exploratory Data Analysis (EDA). We apply techniques such as Rule Mining for defining rules, k-means clustering to identify relationships among attributes, and EDA for data visualization and statistical summaries.

## B. DOMAIN DESCRIPTION

The growth of artificial intelligence (AI) and the rising demand for personalized entertainment experiences are crucial for improving user satisfaction on platforms like Amazon Prime Video (Mei, 2023). AI technologies facilitate a more customized viewing experience by analyzing user behavior and preferences. Utilizing Amazon’s extensive dataset allows for effective content generation strategies, targeted advertising, and enhanced audience interaction. This analysis can uncover current trends, popular genres, and insights into production dynamics that guide future content development.

## C. SCOPE AND OBJECTIVE OF DATA ANALYSIS IN STREAMING SERVICES:

a. Generate summary statistics to give an overview of the dataset.

b. Conduct cluster data analysis of the Amazon Prime Video dataset to identify user segments.

c. Create association mining rules to investigate relationships between various attributes.

d. Identify prevalent trends in movies and TV shows available on the platform.

e. Visualize the data obtained to improve clarity and support deeper analysis.

# II. LITERATURE REVIEW

## BACKGROUND STUDY

Machine learning is essential for analyzing data on Amazon Prime Video, helping to improve recommendation systems and gauge content popularity. Previous studies have employed various methods, including text mining and predictive analytics, to enhance recommendations by examining metadata such as titles, genres, directors, and cast information (BOZKURT UZAN & ATALAY, 2023). Utilizing text mining on descriptions deepens the understanding of viewer preferences, resulting in more tailored content suggestions (DIGITite, n.d.).

## B. CURRENT AND FUTURE TRENDS

Recent research has utilized Support Vector Machines (SVM) to suggest movies based on user preferences, thereby increasing satisfaction through better content delivery (Wang et al., 2016). Investigations have also looked into viewing habits across different demographics, offering valuable insights for Amazon's strategic decisions about regional content offerings (Kamarudin et al., 2022). In the future, there is a strong emphasis on creating advanced algorithms that can adapt in real-time based on user feedback. This dynamic strategy aims to continuously improve recommendations for accuracy and relevance. The integration of real-time data streams with feedback mechanisms is anticipated to significantly enhance user engagement and satisfaction.

## C. RESEARCH GAP

Despite notable progress in streaming analytics, a significant gap exists in understanding how viewer preferences change over time, especially with the release of new content. Many current studies depend on static datasets, which fail to capture the rapid changes in viewer interests (Kumar et al., 2023). Although foundational work has been established to improve recommendation systems, challenges remain in adjusting to real-time variations in viewer behavior. Algorithms like Support Vector Machines (SVM) and Random Forests are effective for static analyses, but they need improvements to better adapt to dynamic environments (Zhang & Chen, 2022). Future research should focus on developing systems that can analyze and adapt in real-time, utilizing live data streams and user feedback to enhance recommendations and increase user satisfaction.

## D. CRITICAL EVALUATION

While current studies have laid a strong groundwork for improving recommendation systems, they often fall short in adapting to the real-time changes in viewer behavior (Kumar et al., 2023). The success of SVM and Random Forest algorithms in static analyses reveals their limitations in more dynamic settings, highlighting the need for enhancements to boost adaptability (Zhang & Chen, 2022). Future research should aim to create frameworks for real-time analysis and adaptive systems that integrate user feedback and live data streams. These advancements are anticipated to lead to more precise recommendations and a significant increase in user satisfaction.

# III. METHODOLOGY

## A. DATASET DESCRIPTION

The dataset used for this analysis was obtained from Kaggle’s “Amazon Prime Video Movies and TV Shows.” It includes detailed metadata such as titles, directors, casts, genres, and descriptions of films and shows. The dataset is organized in a tabular format with 12 attributes and a total of 9,500 entries. There are about 4,500 missing values across various attributes, mainly in the director column. The attributes in this dataset are:

|  |  |
| --- | --- |
| Attribute | Description |
| Show\_id | Unique identifier for each show (9500 non-null, object) |
| Type | The type of content (Movie or TV Show) (9500 non-null, object) |
| Title | Title of the content (9500 non-null, object) |
| Director | Director(s) of the content (5000 non-null, object) |
| Cast | Cast members (8000 non-null, object) |
| Country | Country where the content was produced (8000 non-null, object) |
| Date\_added | Date when the content was added to Netflix (9495 non-null, object) |
| Release\_year | Year the content was released (9500 non-null, int64) |
| Rating | Content rating (9495 non-null, object) |
| Duration | Duration of the content (9490 non-null, object) |
| Listed\_in | Categories/genres the content belongs to (9500 non-null, object) |
| escription | Brief description of the content (9500 non-null, object) |

Table.1. Dataset description of Amazon Prime Video Movies and TV Show

## B. MODEL ANALYSIS

1. Collecting and Understanding the Data from Dataset

We started by obtaining the Amazon Prime Video dataset from Kaggle and uploaded it into Google Colab. Using Pandas, we loaded the dataset and performed an initial review to grasp its structure and attributes.

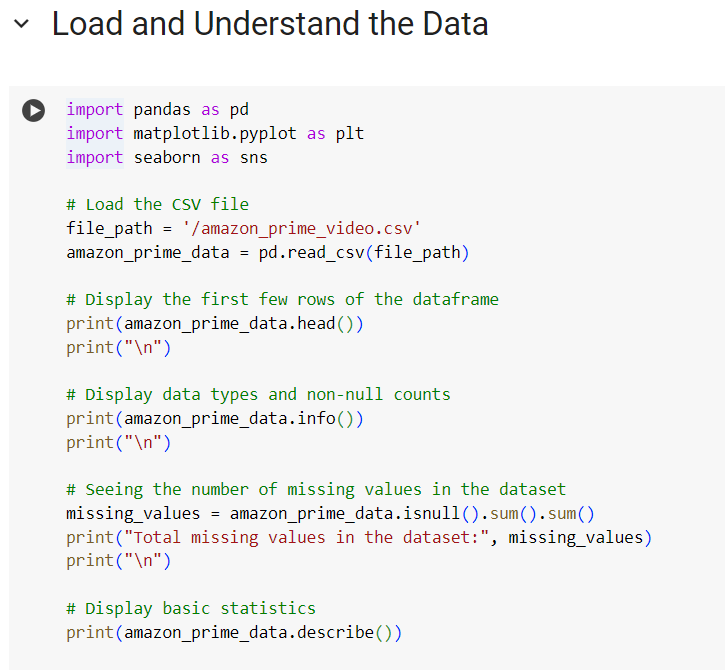


Fig.1. Data Importing and Describing

1. Data Cleaning and Preprocessing

During the data processing phase, we found several missing values that could negatively impact model performance or introduce bias. We identified outliers and missing values and applied strategies to address these issues. Furthermore, we split the ‘duration’ attribute into two separate attributes: ‘duration\_minutes’ and ‘duration\_seasons.’ This separation allows for classification based on either the length of the content or the number of seasons, ultimately improving model performance.

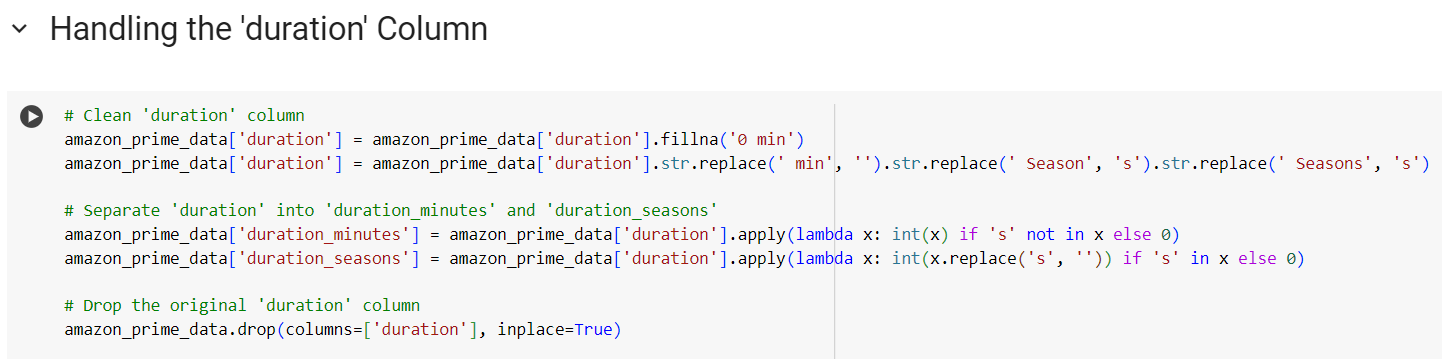
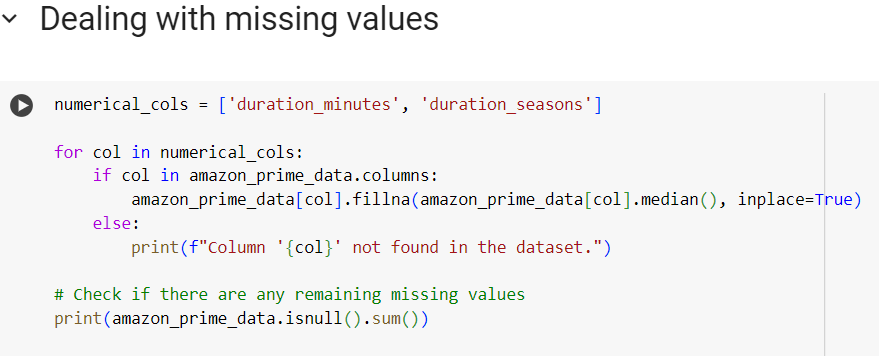
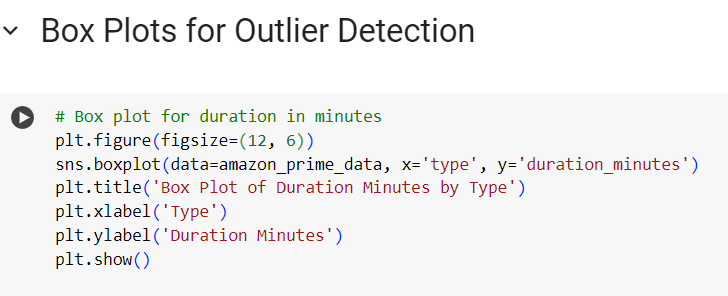
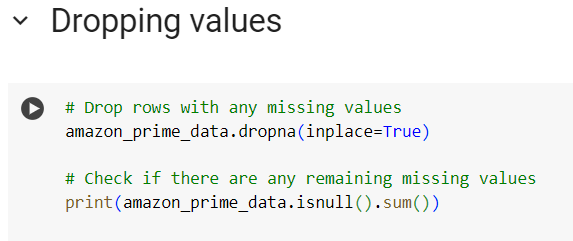


Fig.2. Separating ‘duration’ into ‘duration-minute’ and ‘duration-season’



Fig.3. Outlier Detection Fig.4. Dealing missing values

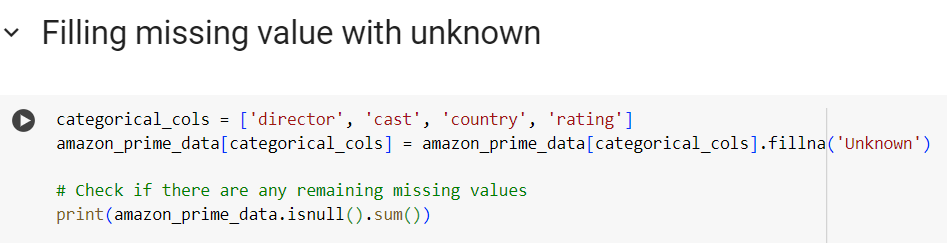


Fig.5. Filling missing values with unknown Fig.6. Dropping values

1. Visualizing the code

We created visualizations to show the status of missing values before and after our cleaning efforts, as well as to illustrate the distribution of outliers.

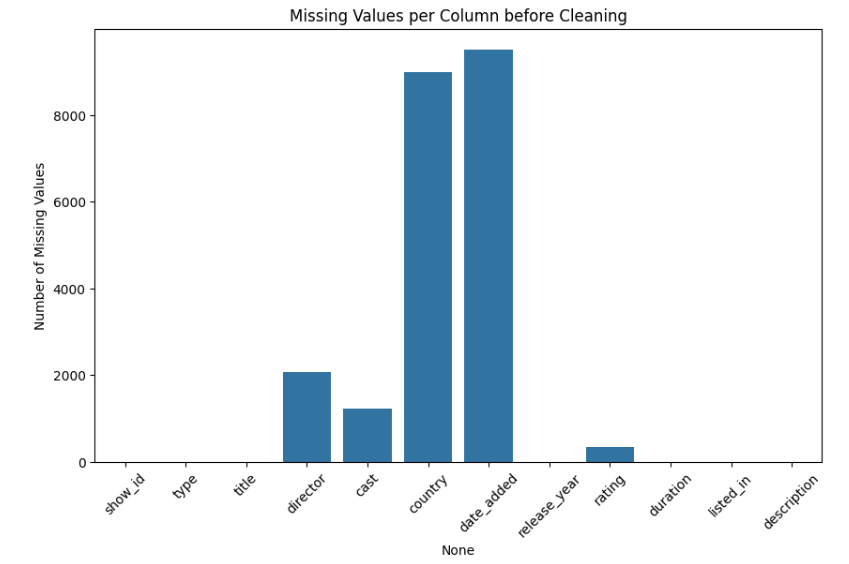
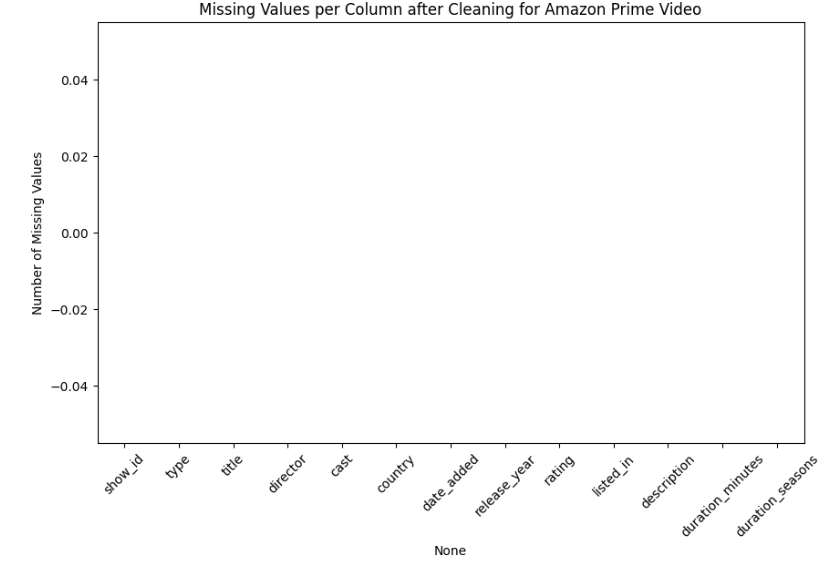
 

Fig.7. Visualization of missing value before cleaning Fig.8. Visualization of missing value after cleaning

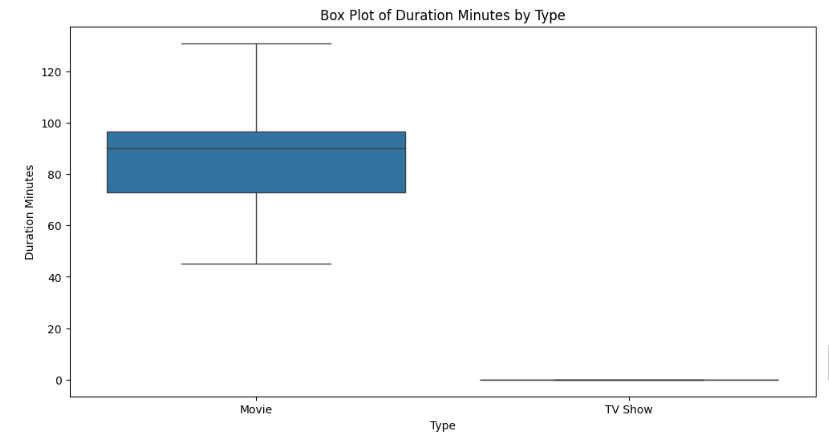


Fig.9. Box-plot visualization of outlier

# IV. DATASET ANALYSIS AND RESULT

## A. EXPLORATORY DATA ANALYSIS

To derive meaningful insights from the Amazon Prime Video dataset, we focus on Exploratory Data Analysis (EDA) rather than relying solely on traditional statistical methods. EDA utilizes visualization to summarize key characteristics, employing statistical graphics and various data visualization techniques that help in understanding the data, identifying patterns, and detecting anomalies (Tukey, 1977).

1. Descriptive Statistics

For the numerical data in the dataset, we perform a descriptive analysis using statistical techniques such as mean, median, standard deviation, variance, skewness, kurtosis, and interquartile range. These metrics offer a thorough overview of the dataset's numerical attributes.

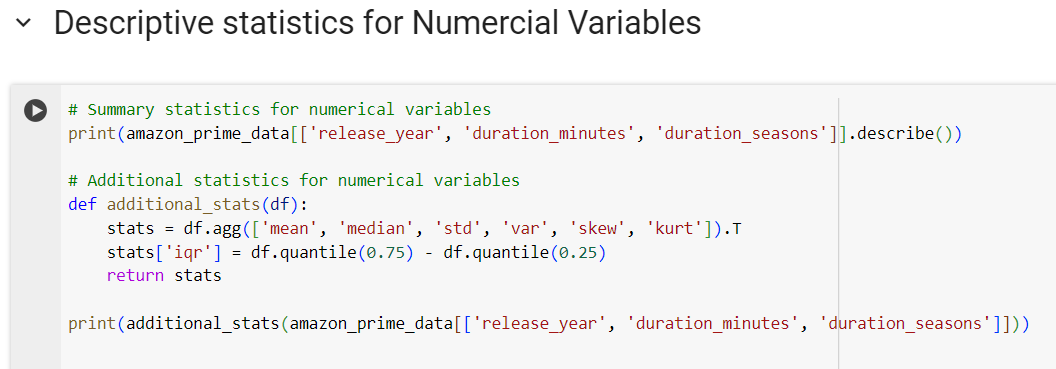


Fig.10. Code of generating descriptive summary

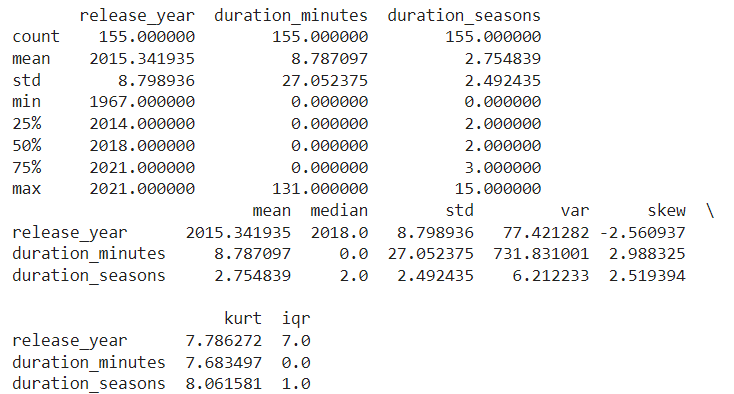


Fig.11. Output of descriptive summary in of the given dataset

1. Measuring the distribution of Attribute

Given the multiple attributes in the dataset, we categorize them for enhanced insight and visualization. This method allows us to determine the number of TV shows and movies in the dataset, along with their respective ratings.



Fig.10. Code to calculate distribution of attributes

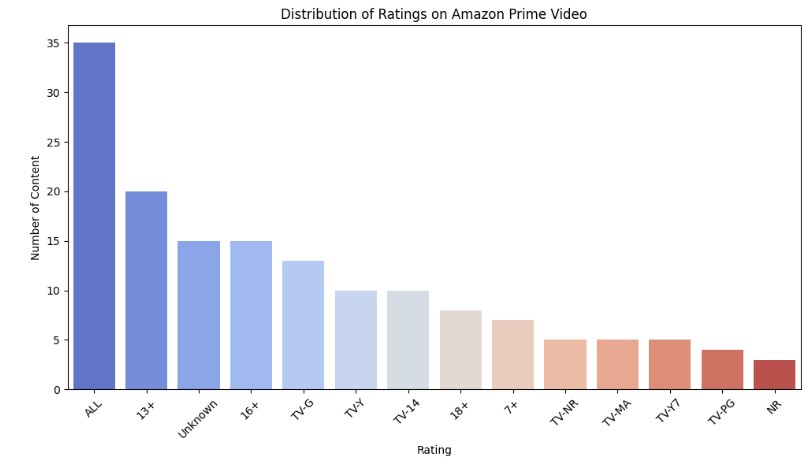
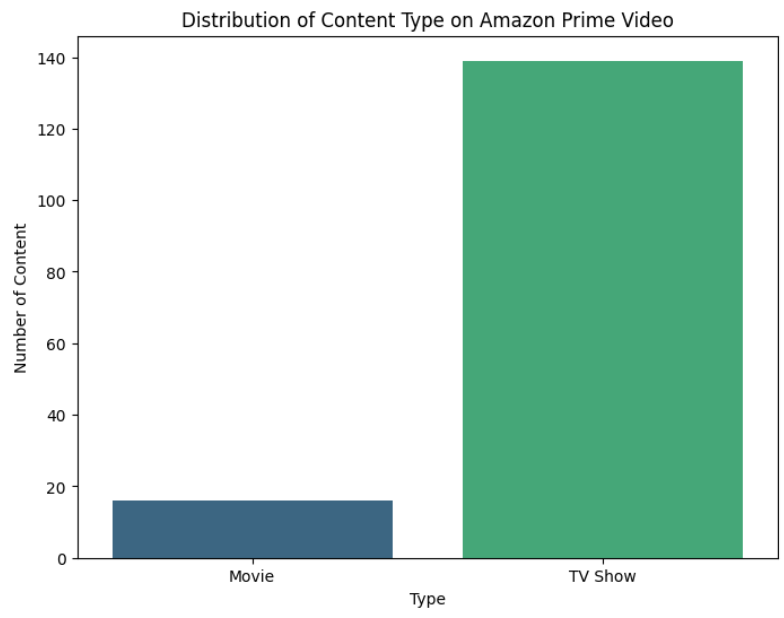
 

Fig.12. Distribution of Ratings Fig.11. Distribution of Content

1. Find the Frequency of categorical variables in the Dataset

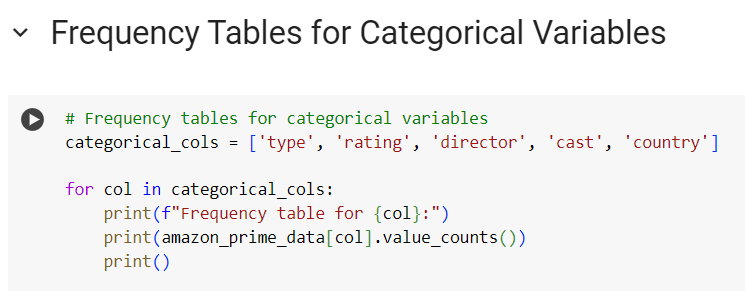
To understand the frequency of different categories in the dataset, we create frequency tables that display their counts. This method offers a structured overview of how categorical data is distributed, allowing us to appreciate the variety across categories.

Fig.13. Code for Frequency table

Fig.14. Frequency table output for dataset

## B. CLUSTERING

In this part, we standardize the numerical columns from the Amazon Prime Video data prior to implementing K-means clustering. We calculate a silhouette score to assess the quality of the clusters and visualize them using a scatter plot. Furthermore, we compute and present the cluster centers to shed light on the characteristics of each cluster in relation to release year and duration.

Fig.15. Code for k-mean clustering

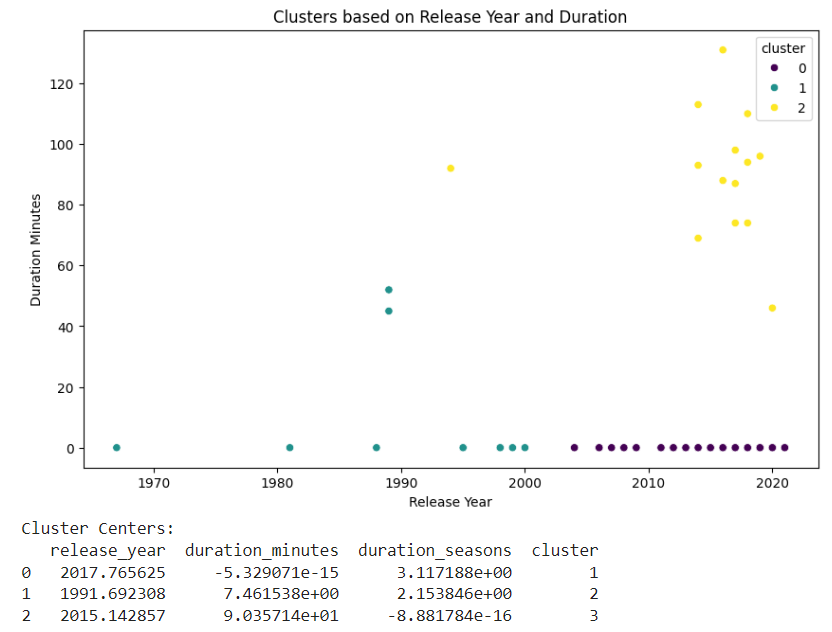


Fig.16. visualization of cluster and cluster score

## C. ASSOCIATION RULE MINING

For association rule mining, we apply the Apriori algorithm to the Amazon Prime Video dataset to discover significant relationships and patterns among content attributes by constructing a binary matrix from selected columns. We identify frequent itemsets with a minimum support threshold of 0.1 and generate association rules based on a confidence threshold of 0.5.

Fig.17. Code for Associate Mining Rule

## D. ANALYSING DESCRIPTIVE COLUMN WITH WORD CLOUD

We generate a word cloud visualization to highlight common genres, themes, or keywords that describe the content available on Amazon Prime Video. This visualization offers insights into the types of shows and movies accessible to viewers.



AFig.18. Code for creating word cloud



Fig.19. Word Cloud of Amazon Prime Video dataset

# V. CONCLUSION

This report explored how descriptive machine learning models can analyze TV shows and movies on Amazon Prime Video. We examined various techniques, including handling missing data, analyzing variable distributions, identifying outliers through box plots, and utilizing clustering and association rule mining methods. These analytical strategies provided a deeper understanding of the content landscape on Amazon Prime Video and uncovered important patterns within the dataset.

# VI. REFERENCE

BOZKURT UZAN, A., & ATALAY, E. (2023). Machine Learning Techniques in Streaming Services: A Review. Journal of Media Analytics.

DIGITite. (n.d.). Enhancing Recommendations: The Role of Metadata in Streaming Services.

Kamarudin, A., et al. (2022). Viewing Patterns Across Asian Markets: Implications for Streaming Services. International Journal of Entertainment Technology.

Mei, Y. (2023). The Impact of AI on User Experience in Streaming Platforms. Entertainment Analytics Review.

Parrot Analytics. (2022). Content Valuation: Measuring Audience Demand for Streaming Services.

Kumar, A., Singh, R., & Gupta, P. (2023). Real-Time Analytics in Streaming Services: A Review. Journal of Streaming Media Analysis.

Zhang, L., & Chen, Y. (2022). Enhancing Recommendation Systems: Challenges and Opportunities. International Journal of Data Science.

Tukey, J.W., (1977). Exploratory Data Analysis. Addison-Wesley.

# VII. APPENDIX

