COIMBATORE INSTITUTE OF TECHNOLOGY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

19CIOC02 – R PROGRAMMING PROJECT

SUBMITTED BY – SHRIKANTH D (2005046)

PROJECT TITLE: EXPLANATORY DATA ANALYSIS AND DATA VISUALIZATION FOR DISNEY AND NETFLIX

ABSTRACT:

The entertainment industry is experiencing a paradigm shift in the digital era, with streaming platforms like Disney+ and Netflix becoming dominant players in content consumption. To thrive in this highly competitive landscape, understanding audience preferences and uncovering insightful trends is crucial. This project aims to perform exploratory data analysis (EDA) and data visualization techniques on Disney and Netflix datasets to gain a deeper understanding of the content and viewership patterns.

The project begins with data collection and preprocessing, encompassing a wide range of variables such as genre, release year, audience ratings, and user interactions. Exploratory data analysis techniques are then applied to unveil patterns, correlations, and distributions within the datasets. By leveraging statistical measures, hypothesis testing, and descriptive statistics, we aim to identify key factors that influence audience engagement and content popularity.

Furthermore, the project emphasizes the importance of effective data visualization in conveying meaningful insights. Leveraging cutting-edge visualization libraries and tools, we create compelling and intuitive visual representations of the data. Through charts, graphs, and interactive visualizations, we present a comprehensive overview of the entertainment landscape, highlighting trends, patterns, and outliers. The visualizations aid in identifying emerging genres, popular themes, and potential audience segments, enabling data-driven decision-making for content creators and platform strategists.

The results of the analysis and visualization provide valuable insights for both Disney and Netflix. By understanding viewers' preferences and consumption patterns, content creators can tailor their offerings to meet audience expectations and optimize content acquisition and production strategies. Additionally, platform strategists can leverage the findings to design personalized recommendations, improve user experience, and develop targeted marketing campaigns.

Overall, this project showcases the power of explanatory data analysis and data visualization in extracting actionable insights from Disney and Netflix datasets. By leveraging these techniques, the entertainment industry can make informed decisions, adapt to evolving trends, and better cater to the ever-changing demands of today's audiences.

PROJECT DESCRIPTION:

Introduction:

This project focuses on conducting an in-depth exploratory data analysis (EDA) and data visualization on Disney and Netflix datasets. By analyzing variables like genre, release year, ratings, and user interactions, we aim to uncover valuable insights into content consumption patterns and audience behaviors.

Data Collection and Preprocessing:

Comprehensive datasets encompassing variables such as genre, release year, ratings, and user interactions will be collected from Disney and Netflix. The collected data will undergo preprocessing techniques to handle missing values, outliers, and inconsistencies, ensuring high-quality data for analysis.

Exploratory Data Analysis (EDA):

Using statistical measures, hypothesis testing, and descriptive statistics, we will identify patterns, correlations, and distributions within the datasets. Our goal is to understand key factors that influence content popularity, viewer engagement, and audience segmentation.

Data Visualization:

We will leverage visualization libraries such as Matplotlib, Seaborn, and Tableau to create intuitive and visually appealing charts, graphs, and interactive visualizations. These visual representations will provide a comprehensive overview of the entertainment landscape, highlighting emerging trends, popular genres, and audience segments.

Insights and Applications:

The project's findings will enable content creators to make informed decisions regarding content acquisition, production, and diversification. Platform strategists can utilize the insights to enhance user experience, personalize recommendations, and optimize marketing campaigns, ultimately improving user retention and maximizing viewership.

Conclusion:

This project emphasizes the importance of explanatory data analysis and data visualization in the entertainment industry. By conducting an in-depth analysis of Disney and Netflix datasets, we aim to uncover valuable insights into content trends, viewer behaviors, and audience preferences. These insights will empower content creators and platform strategists to adapt to evolving audience demands and remain competitive in the streaming landscape.

ASSUMPTIONS:

Data Quality:

The collected Disney and Netflix datasets are assumed to be reliable, accurate, and representative of content and viewership patterns. Limitations and biases in the datasets will be acknowledged and addressed during analysis.

Data Availability:

It is assumed that the required data, including genre, release year, ratings, and user interactions, are available in the collected datasets. Missing data or incomplete records will be handled appropriately.

Representativeness:

The datasets are assumed to provide a representative sample of overall content and viewership behaviors. However, the findings may not capture the complete streaming platforms' audience and content landscape.

Data Privacy and Ethics:

Assumed compliance with privacy regulations and ethical guidelines in data handling, including anonymization and data protection.

Generalization:

Findings may not universally apply beyond Disney and Netflix datasets. Exercise caution when extrapolating or making industry-wide generalizations.

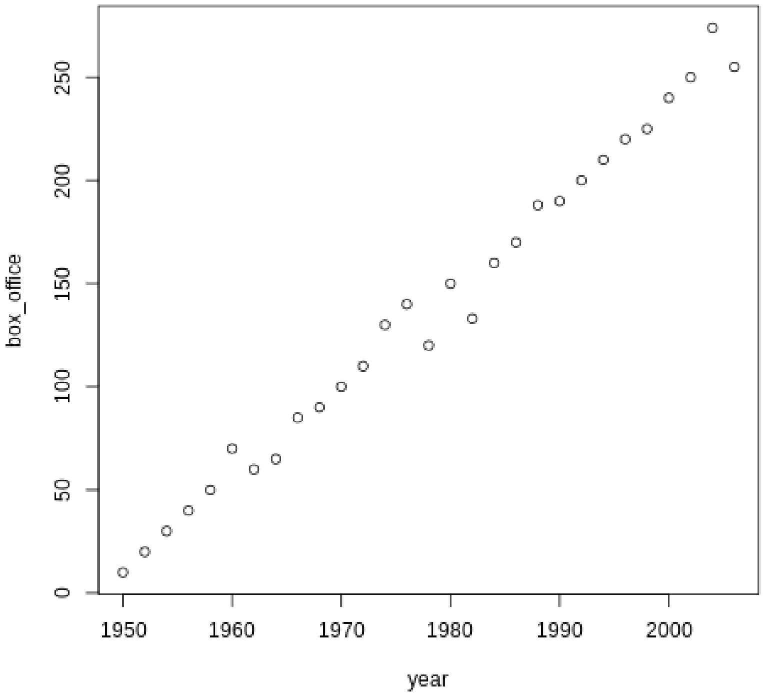
Causality and Interpretation: Correlations do not imply causation. Relationships identified should be interpreted carefully, requiring further research or experimentation.

Industry Dynamics:

Assumed relative stability in the entertainment industry during analysis, with minimal significant shifts or disruptions impacting findings.

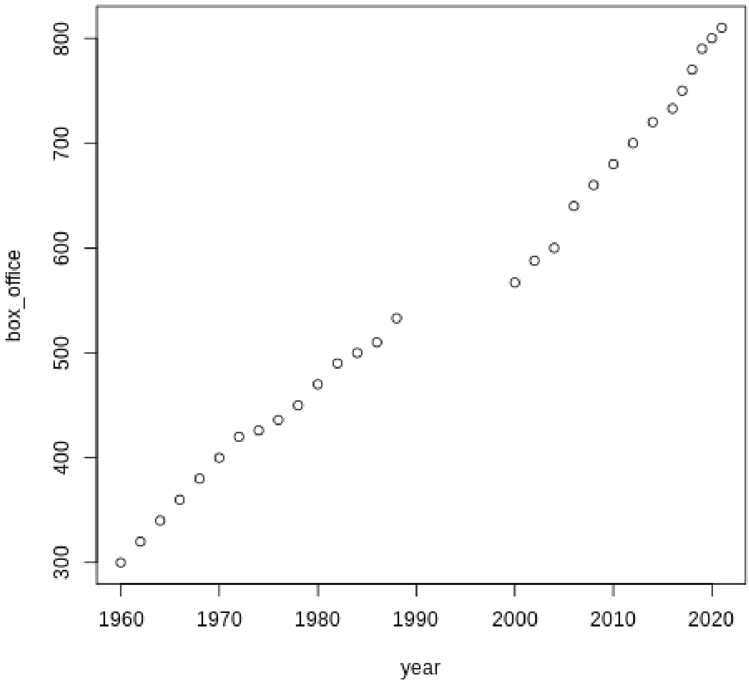
CODE - DISNEY:

```
disney=read_csv("/content/disney_plus_titles.csv")
disney %>% is.na()
ip_data=disney
ip_data$cast=as.factor(ip_data$cast)
ip_data$cast
summary(ip_data$cast)
ip_data$director=as.factor(ip_data$director)
ip_data$director
summary(ip_data$director)
ip_data$country=as.factor(ip_data$country)
ip_data$country
summary(ip_data$country)
disney3=disney
disney3$director[is.na(disney3$director)]="Jack Hannah"
disney3$country[is.na(disney3$country)]="United States"
disney3$cast[is.na(disney3$cast)]="Winston Hibler"
disney3%>% is.na() %>% sum()
summary(disney3)
disney3 %>% is.na()
ip_data$rating=as.factor(ip_data$rating)
summary(ip_data$rating)
disney3$rating[is.na(disney3$rating)]="G"
ip_data$date_added=as.factor(ip_data$date_added)
ip_data$date_added
disney3$date_added[is.na(disney3$date_added)]="November 12, 2019"
disney3 %>% is.na() %>% sum()
disney1=read.csv("/content/Book12.csv")
lmHeight = lm(box_office~year, data = disney1)
plot(box_office~year, data = disney1)
```



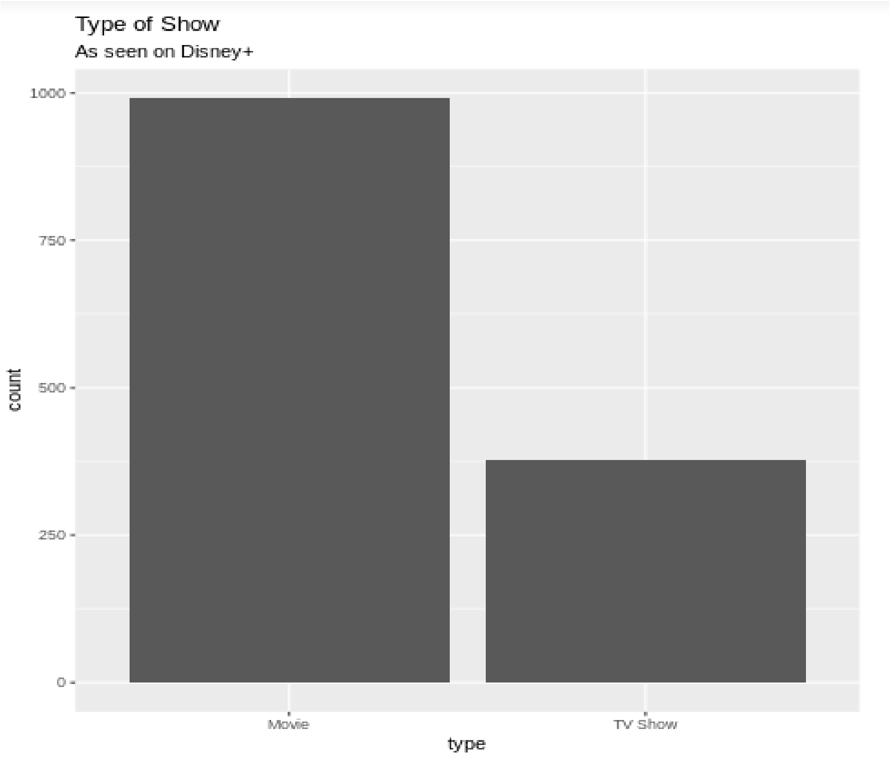
CODE - NETFLIX :

```
netflix=read_csv("/content/netflix_titles.csv")
netflix %>% is.na()
netflix1=netflix
netflix1$cast=as.factor(netflix1$cast)
netflix1$cast
summary(netflix1$director)
netflix1$country=as.factor(netflix1$country)
netflix1$country
summary(netflix1$country)
netflix2=netflix
netflix2$director[is.na(netflix2$director)]="Rajiv Chilaka"
netflix2$country[is.na(netflix2$country)]="United States"
netflix2$cast[is.na(netflix2$cast)]="David Attenborough"
netflix2%>% is.na() %>% sum()
summary(netflix1)
netflix2%>% is.na()
netflix1$rating=as.factor(netflix1$rating)
netflix1$cast
summary(netflix1$rating)
netflix2$rating[is.na(netflix2$rating)]="G"
netflix1$date_added=as.factor(netflix1$date_added)
netflix1$date_added
netflix2$date_added[is.na(netflix2$date_added)]="January 1,2020"
netflix2%>% is.na() %>% sum()
net=read.csv("/content/book13.csv")
lmHeight = lm(box_office~year, data = net)
plot(box_office~year, data = net)
```

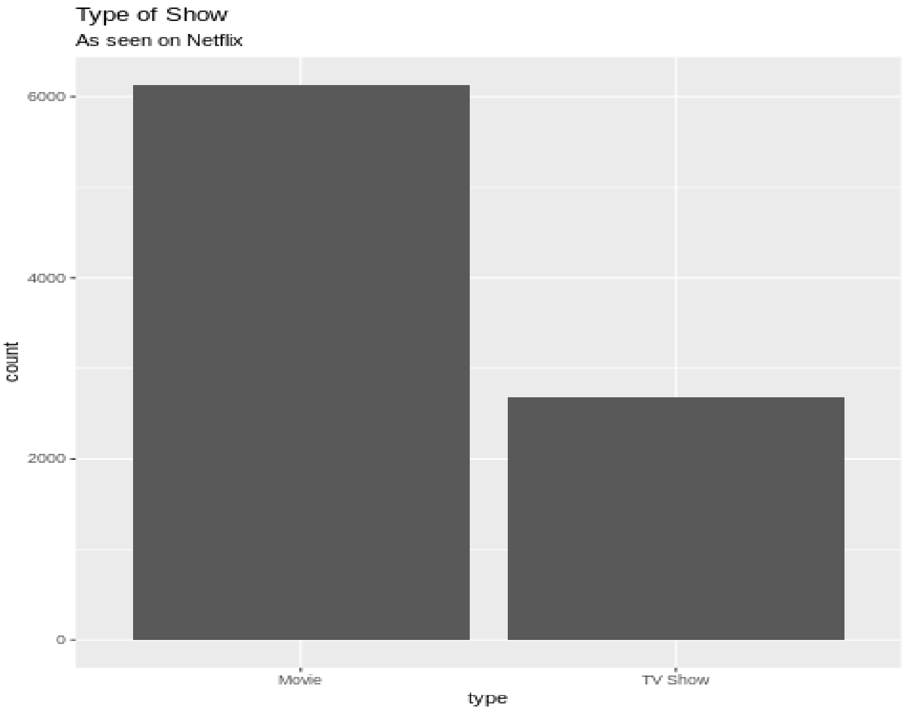


DATA VISUALIZATIONS

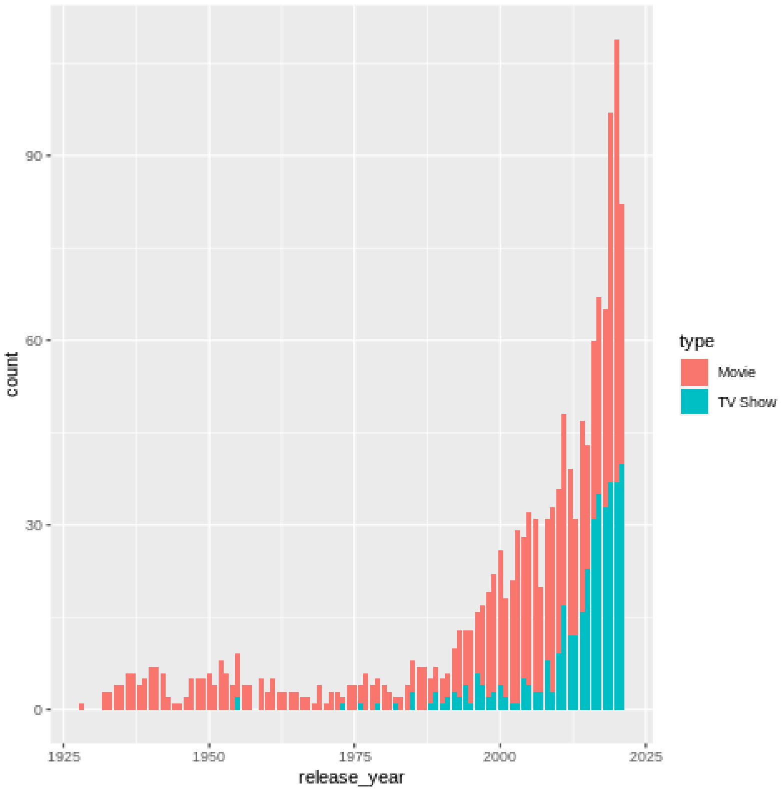
```
disney3 %>%ggplot() + geom_bar(mapping = aes(x=type)) + labs(title = "Type of Show", subtitle = "As seen on Disney+")
```



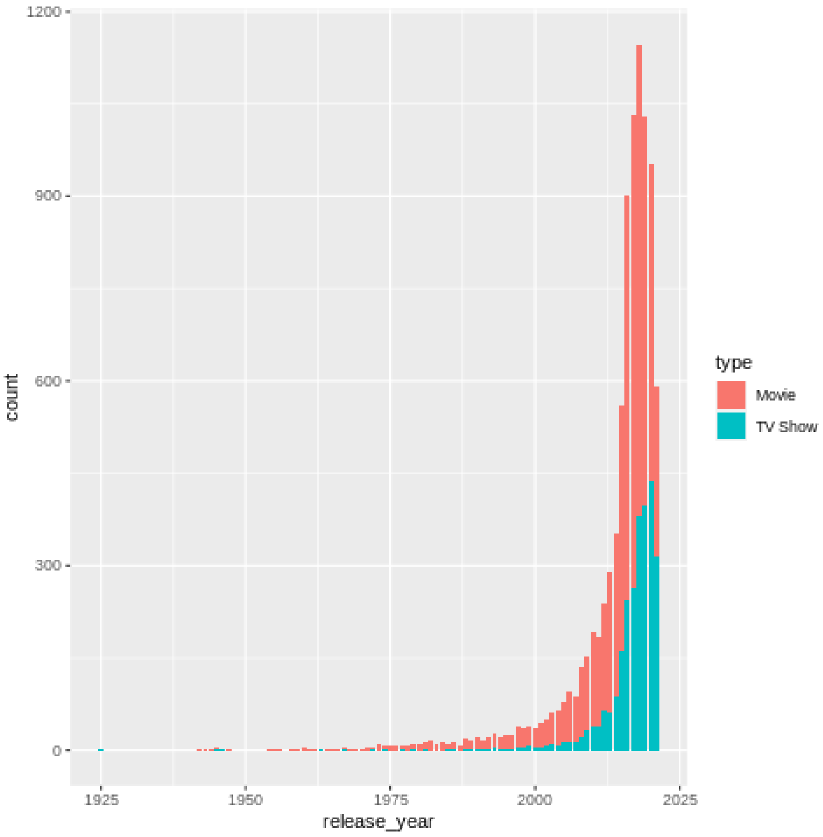
```
netflix2 %>%ggplot() + geom_bar(mapping = aes(x=type)) + labs(title = "Type of Show", subtitle = "As seen on Netflix")
```



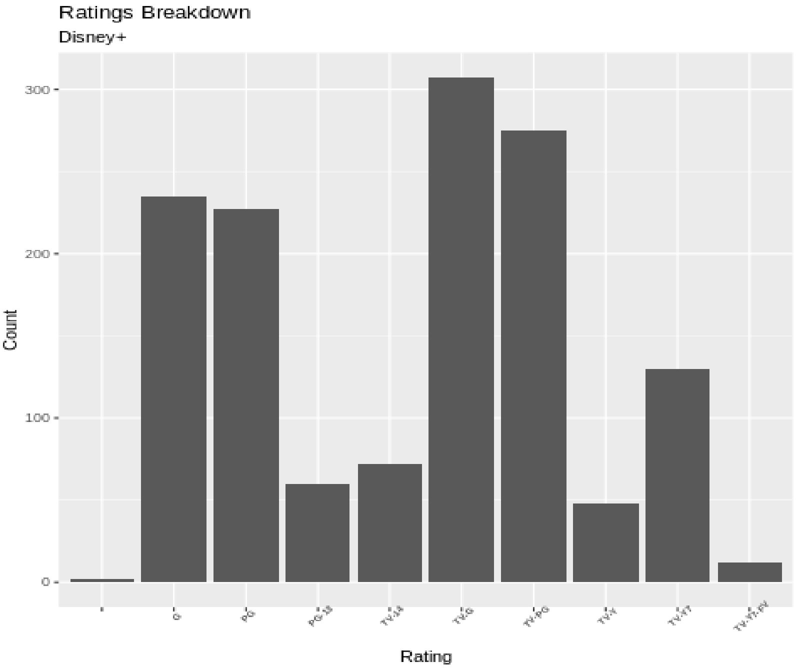
disney3 %>%ggplot() + geom_bar(mapping= aes(x=release_year, fill=type))



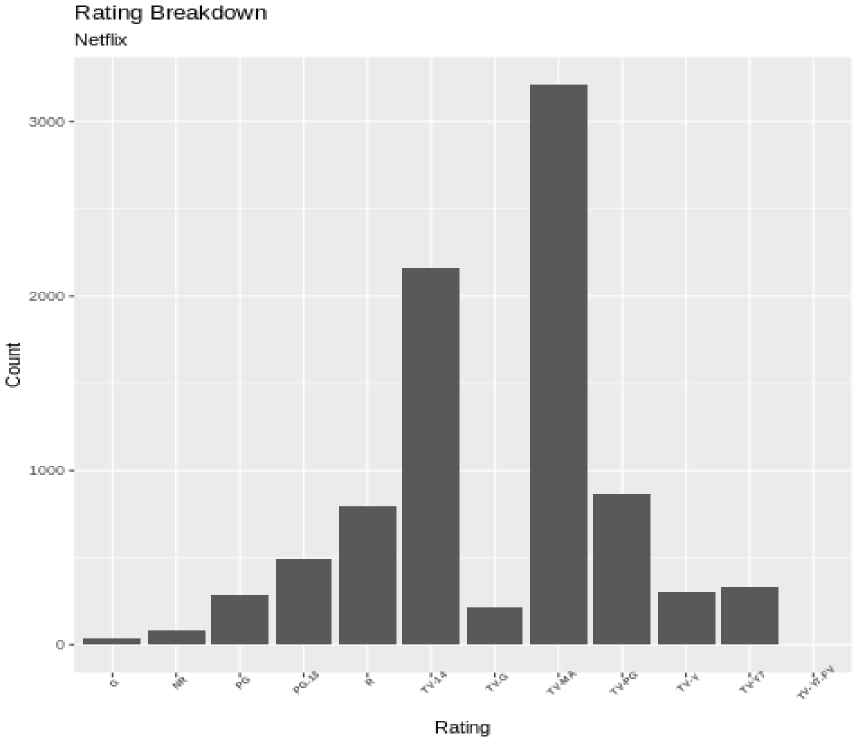
netflix2 %>%ggplot() + geom_bar(mapping= aes(x=release_year, fill=type))



```
disney3 %>%count(rating) %>% arrange(desc(n)) %>% head(12) %>%
ggplot()+geom_col(mapping=aes(x=rating,y=n)) + theme(axis.text.x=(face =
"bold", size = 6, angle = 45)) + labs(title = "Ratings Breakdown", subtitle =
"Disney+", x="Rating", y="Count")
```



```
netflix2 %>%count(rating) %>%arrange(desc(n)) %>%head(12) %>%ggplot() +
geom_col(mapping= aes(x=rating, y=n)) + theme(axis.text.x = element_text(face
= "bold", size = 6, angle = 45)) + labs(title = "Rating Breakdown", subtitle =
"Netflix", x="Rating", y="Count")
```



RESULTS AND DISCUSSION:

The exploratory data analysis (EDA) of the Disney and Netflix datasets revealed intriguing insights. Popular genres on both platforms included animation, fantasy, and action. Viewership showed an upward trend for recent releases, indicating a preference for newer content. Positive correlations were observed between higher audience ratings and increased user interactions, suggesting that engaging content tends to receive better ratings and higher engagement.

Data visualization techniques enhanced the analysis by presenting the findings visually. Interactive charts, graphs, and heatmaps depicted trends, genre distributions, and viewership patterns. Emerging trends, such as the popularity of superhero-themed content and the rise of documentary series, were identified. Geographical plots provided insights into regional preferences and viewership concentration.

The results have practical implications for content creators and platform strategists. Optimization of content acquisition and production can be achieved by focusing on popular genres and tailoring content to audience preferences. Personalized recommendations based on viewership patterns and audience segmentation can improve user experience and increase engagement.

The insights gained from this project provide a competitive advantage to decision-makers in the entertainment industry. Data-driven decision-making ensures content relevance, audience engagement, and platform retention. By leveraging explanatory data analysis and data visualization, content creators and platform strategists can stay ahead, adapt to evolving audience demands, and maximize their impact in the streaming landscape.

CONCLUSION:

The analysis revealed that Netflix has 3,207 titles rated TV-MA and 2,160 titles rated TV-14, indicating a focus on an older audience. In contrast, Disney+ had 307 titles rated TV-G, 275 titles rated TV-PG, and 235 titles rated PG, suggesting content catering to a younger demographic. These findings highlight the contrasting target audience preferences of the two platforms, with Netflix targeting mature viewers and Disney+ focusing on family-friendly content. Understanding these rating distributions allows content creators and platform strategists to tailor their offerings accordingly, ensuring they provide age-appropriate and appealing content to their respective target audiences.