Netflix Original Movies and IMDB scores

Ridhi Batra

School of Computer Science and Engineering,

Lovely Professional University, Punjab.

Abstract

Analyzing Netflix's datasets for upcoming TV series and movie releases on the service. In this project, we will examine the Kaggle dataset in order to determine the following: the average time it takes for a movie or TV show to be released on the Netflix platform; the number of releases within a given time period; the number of releases within the last ten years on the platform; and the top ten genres that the Netflix platform's audience enjoyed the most. From here, we want to use a machine learning strategy to completely comprehend the data and offer a fantastic answer for the platform's future. From our data analysis we conducted on R markdown, we have discovered that there were a wide variety of genres that movie directors produced worldwide and we have observed many cast members and genres they were in.

Keywords
Netflix TV shows, imdb popularity, release year, genres, seasons, description

Data Analysis on the Netflix Datasets

Motivation

The biggest movie and TV show streaming service on the internet is called Netflix. Numerous nations, including but not limited to the United States, India, South Korea, Japan, and many more, offer this service. The service was initially made available online as a DVD rental service, then afterwards, the creator and CEO of the business, Reed Hastings, introduced a novel method of streaming movies and TV series on its website, enabling a large number of users to instantly stream their preferred content on a variety of Internet-enabled devices, such as mobile phones, tablets, laptops, and desktop computers. Netflix's sales skyrocketed after it adopted a whole new strategy for distributing TV episodes and films. Since then, the website has developed a recommender system of its own to learn what kinds of TV series and movies viewers enjoy watching, what kind of cinematography they find most appealing, and how they watch their favorite series. Netflix saw an increase in users due to the power of data analysis, and many of these users now spend most of their time viewing movies and television series on the platform. Using this method, we want to investigate the dataset and comprehend the Netflix movie and TV show trends.

Introduction

Initially, we aimed to obtain a general overview of the dataset under consideration. Initially, we imported tidyverse for a basic analysis of data. We obtained the dataset from Kaggle and will make use of information available on the Kaggle website to comprehend the pattern of

films and television series that are released on the platform.
This collection of data includes We could see the column names in the CSV file by looking at the code. The following columns will be used to help us determine which films and TV series were released in a given year, as well as their genres, release dates, and audience ratings.

```
> summary(netflix)
   Title              Genre              Premiere            Runtime        IMDB.Score      Language
 Length:584         Length:584         Length:584         Min.   :  4.00   Min.   :2.500   Length:584
 Class :character   Class :character   Class :character   1st Qu.: 86.00   1st Qu.:5.700   Class :character
 Mode  :character   Mode  :character   Mode  :character   Median : 97.00   Median :6.350   Mode  :character
                                                          Mean   : 93.58   Mean   :6.272
                                                          3rd Qu.:108.00   3rd Qu.:7.000
                                                          Max.   :209.00   Max.   :9.000
>
```

Based on the dataset summary, the earliest year that a movie or TV show was filmed was 1925, and the most recent content available on the site is 2021. We can learn how long it takes Netflix to upload content to the platform, which genre is most popular, which actors and actresses are most popular, and how the popularity of a genre has affected the movies and TV shows that an actor or actress has appeared in over time. This information includes directors, actors and actresses, movies, TV shows, ratings, duration of each content, and more. We'll investigate the Netflix trend and use R to determine its cause.

Project Background

In order to create a new data exploration on the trend of movies on the platform, the paper offers a data analysis study and coordination activity one in the Netflix dataset found on Kaggle. When Netflix was debuted, it lacked data analysis to comprehend the previously described audience/user pattern on the platform. As time went on, the value of using data analysis became more apparent, and the most popular shows were launched on the platform only after careful examination of the user data the company had gathered.
The two-person team is working on the same process to comprehend the platform's trend and make an effort to determine how long it typically takes the platform to upload content, the top ten directors of films in the top ten nations where the platform is available for streaming, and the most popular genres that actors and actresses belong to. Next, using the dataset they obtained from Kaggle, the team will try to comprehend the general pattern of the Netflix platform.

Methodology

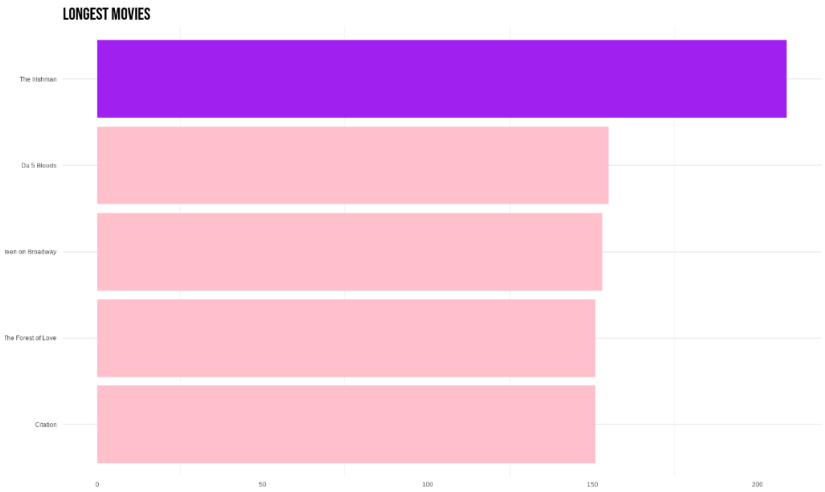
Techniques for Analyzing Data

We used R programming to combine descriptive statistics and data visualization approaches to examine the movie data and IMDb scores. The dataset, which contained details about movie names, genres, release years, IMDb ratings, and runtime lengths, was sourced from a reliable online movie database.

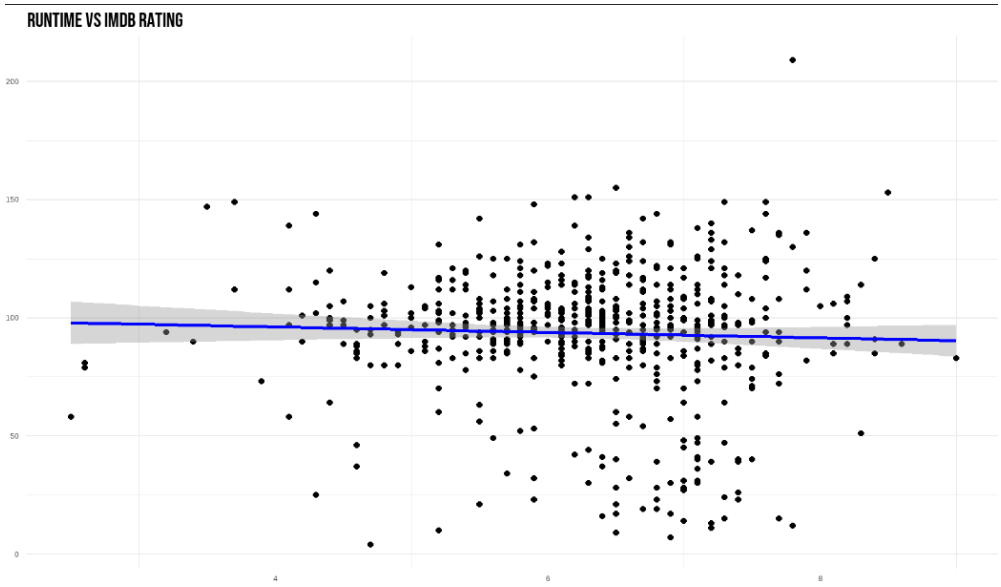
Descriptive Statistics: To learn more about the general distribution and features of the movie data and IMDb ratings, we started by performing descriptive statistics. This involved figuring out runtime durations and IMDb rating values' mean, median, standard deviation, minimum, and maximum values.

Statistical Analyses: To measure associations between variables like IMDb ratings and runtime lengths, we performed statistical analyses including correlation analysis in addition to descriptive statistics and data visualization. We were able to identify any noteworthy trends or correlations in the movie dataset thanks to our research.

```r
#Longest Movies
n10 <- netflix %>% arrange(desc(Runtime)) %>% head(5)

n10_graph <- ggplot(data=n10)+
  geom_col(mapping=aes(
    x=reorder(`Title`,`Runtime`),
    y=`Runtime`,
    fill=ifelse(Runtime==max(`Runtime`),"purple","pink")))+
  labs(title="Longest Movies")+
  theme_minimal()+
  scale_fill_manual(values=c("pink","purple"))+
  coord_flip()+
  theme(
    legend.position="none",
    plot.title = element_text(
      family="Bebas Neue",
      size=25,
      color="black"),
    axis.title.x=element_blank(),
    axis.title.y=element_blank(),
    panel.grid.major.x=element_blank()
  )

n10_graph
```




```r
n12_graph <- ggplot(data=netflix,aes(x = `IMDB.Score`, y = Runtime))+
  geom_point()+
  geom_smooth(method = "lm", color="blue")+
  labs(title="Runtime vs IMDB Rating")+
  theme_minimal()+
  scale_fill_manual(values=c("lightgreen","lightblue"))+
  theme(
    legend.position = "none",
    plot.title=element_text(
      family="Bebas Neue",
      size=25,
      color="black"),
    axis.title.x=element_blank(),
    axis.title.y=element_blank(),
    panel.grid.major.x=element_blank()
  )

n12_graph
```



These data analysis techniques allowed us to gain a comprehensive understanding of the movie data and IMDb scores, leading to informed insights that were subsequently incorporated into the R dashboard for effective visualization and exploration by end-users.

FutureWork

With more time, we would try to develop or enhance a number of ideas.Additional work may have been completed by applying deep learning to this issue.A few practical approaches could be utilized to learn unsupervisedly from numerical data and could also be applied to other neural networks to stop the formation of classification

Conclusion

In conclusion, we have created a somewhat curated examination to ascertain the genre of a particular television program and film.By analyzing and sifting through the elements of the Netflix dataset, we were able to determine which actors and actresses, directed by which, were popular in each of the top ten nations with the most viewership. We then fed this data to figure out the trend over time.
Using logistic regression and growing the dataset and feature set are two examples of steps that could potentially enhance future outcomes.

References

https://www.kaggle.com/datasets

Netflix challenge. http://www.netflixprize.com/.

Netflix leaderboard. http://www.netflixprize.com//leaderboard.

Netflix prize. http://en.wikipedia.org/wiki/Netflix_Prize.

https://datasetsearch.research.google.com/

https://www.researchgate.net/

https://scholar.google.com/