Exploratory Data Analysis using R

Eric Wu

2023-07-10

Background

The website "FlixPatrol" provides a variety of recommendations based on social media platforms and has a large movie data set that offers a regularly updated billboard for the most popular titles.

Figure 1: (Cited from FlixPatrol)

Skill Set Objectives

- Data scraping and wangling
- Data visualizations and chart design

Dataset

The primary data set is scraped from the November FlixPatrol billboard with the name of the movie (Movie_Name), where the movie is from (Country), date of insertion (Date), which platform is it from (Platform), votes (Vote), and other miscellaneous remarks (Misc.).

Research Question

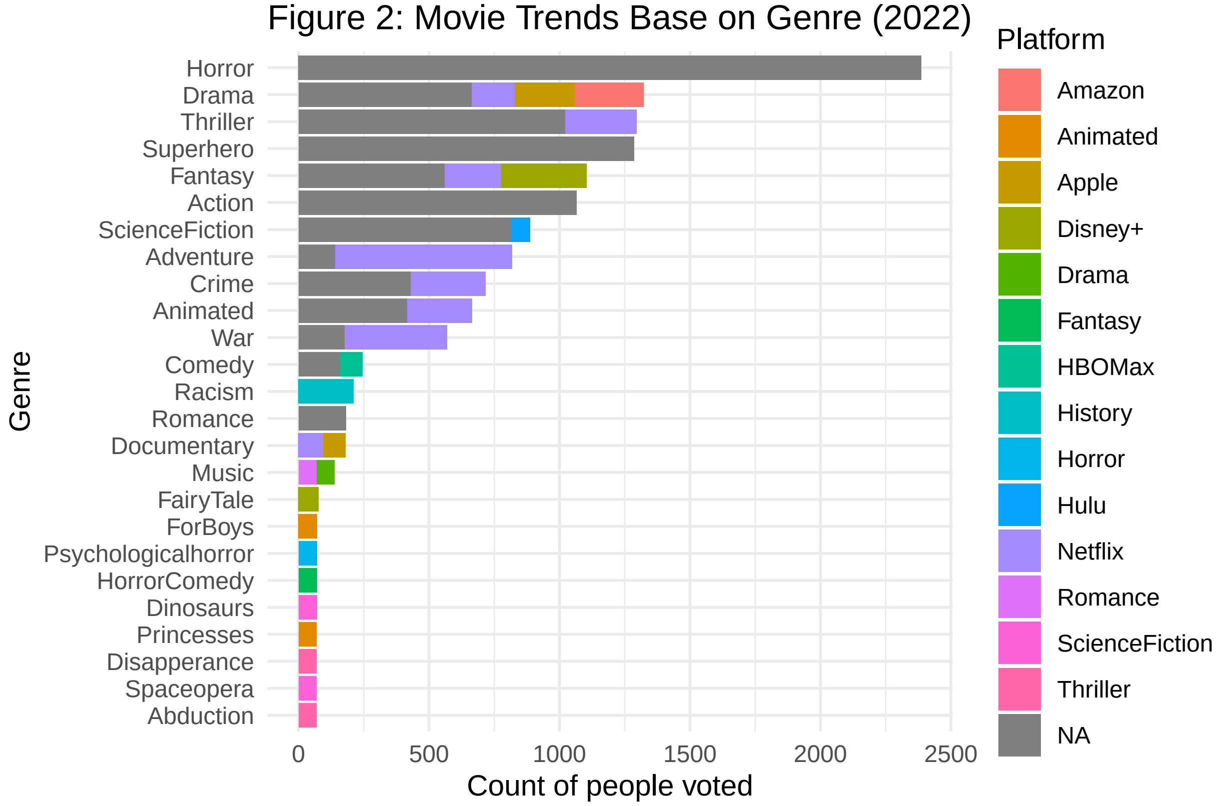
Here are some of the research questions to begin with:

1. What is the most popular genre of the month?
2. What are the three most popular movies within the genre?

Exploratory Data Analysis

Data Visualization

Figure 2: Movie Trends Base on Genre (2022)



Data from FlixPatrol

Figure 2 presents an organized bar graph with the rank of genres using the featured billboard from FlixPatrol. Based on the bar graph, “Horror” is the most favored genre of the month, followed by “Drama” and “Fantasy”. In addition, most movies are not released on only one platform, which is why most of the bar graphs are filled with gray color. Furthermore, it seems like other than N/A, the Netflix platform released the most number of movies. the actual number will be examined using a frequency table.

Table 1: Top Three Horror Movie

Movie_Name	Country	Platform	Genre	Misc.	Vote
Terrifier	United States	NA	Horror	NA	326
Halloween Ends	United States	NA	Horror	NA	284
X	United States	NA	Horror	NA	262

Table 1 presents the top three horror movies with the highest vote.

Table 2: Frequency between Platform and Genre

Platform/Genre	Action	Animated	Drama	Fantasy	Horror	ScienceFiction	Superhero	Thriller	Other	Total
Amazon	0 (0.00%)	0 (0.00%)	1 (1.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)
Animated	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (2.11%)	2 (2.11%)
Apple	0 (0.00%)	0 (0.00%)	1 (1.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	2 (2.11%)
Biopic	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
Disney+	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (2.11%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	3 (3.16%)
Drama	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
Fantasy	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
HBOMax	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
History	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
Horror	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
Hulu	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)
Netflix	0 (0.00%)	1 (1.05%)	2 (2.11%)	1 (1.05%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	5 (5.26%)	10 (10.53%)
Romance	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	1 (1.05%)	1 (1.05%)
ScienceFiction	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (2.11%)	2 (2.11%)
Thriller	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2 (2.11%)	2 (2.11%)
NA	7 (7.37%)	4 (4.21%)	6 (6.32%)	7 (7.37%)	14 (14.74%)	6 (6.32%)	6 (6.32%)	7 (7.37%)	8 (8.42%)	65 (68.42%)
Total	7 (7.37%)	5 (5.26%)	10 (10.53%)	10 (10.53%)	14 (14.74%)	7 (7.37%)	6 (6.32%)	8 (8.42%)	28 (29.47%)	95 (100.00%)

Table 2 provided a deeper understanding of the relationship between platform and genre. In addition, the table drops any genres that has less than 5.3% of data values and combine them into one single "other" column to prevent the final frequency table from being over-sized. From the table, it's proven to be true that Netflix released the most amount of movies that are on this billboard, aside from NA.

Conclusion

This rmd file discussed and analyzed a billboard scraped from a movie recommendation website, and presented several data visualization using the ggplot2, janitor, and kableExtra libraries. Some limitation of this report included a limited amount of attribute that does not support many other kinds of analysis and visualization, and the inability on using other analysis tools like clustering, etc. However, the cleaned data set will also be included as the output of this exercise.

Code Appendix

```r
knitr::opts_chunk$set(echo = FALSE)

rm(list = ls()) # Clean environment
# Load libraries
library(rvest)
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)
library(esquisse)
library(janitor)
library(kableExtra)
library(float)

# Scraping data from FlixPatrol
df <- read_html(
  x = "https://flixpatrol.com/popular/movies/movie-db/2022-11-10/") %>%
  html_elements(css = "table") %>%
  html_table()

flix_table <- bind_cols(df)

movies <- flix_table %>%
  separate( # Divide X2 into multiple individual columns
    col = X2,
    into = c("Movie_Name","Country","Date","Platform", "Genre", "Misc."),
    sep = "\\| "  # Separate the string using a regular expression (Regex) for "|"
  ) %>%
  dplyr::filter( # Filter movies without no genre
    !is.na(Platform)
  ) %>%
  select(-X1) %>% # Get rid of X1 column
  separate( # Separate the string using double spaces
    col = Movie_Name,
    into = c("Movie_Name", "Type"),
    sep = "  "
  )

# Removing extra spaces and characters
movies$Platform <- gsub(
    pattern = " ",
    replacement = '',
    x = movies$Platform)

movies$Genre <- gsub(
    pattern = " ",
    replacement = '',
    x = movies$Genre)

movies$X3 <- gsub(
    pattern = "[ p.]",
    replacement = '',
    x = movies$X3) %>%
  as.numeric() # Turn X3 into numerical data

# Create a vector of genre
genre <- movies$Genre
genre <- append(genre, c('Action', 'Horror', 'Romance','Superhero')) # append other genre to the list m

# Try to move "Platform" value to "Genre" column
for (i in 1:84)
```

```
  if (movies$Platform[i] %in% genre){ # Check if the "Platform" column contains "Genre" value
    value = movies$Platform[i]
    movies$Genre[i] <- value
    movies$Platform[i] <- NA
  }

# Change the name of x3 column
movies <- movies %>%
  rename(
    Vote = X3
  ) %>%
  select(-Type) # Remove "type" column

#esquisse::esquisser(movies)

movies %>%
 filter(!(Platform %in% "Action") | is.na(Platform)) %>% # Filter extra information
 filter(!is.na(Genre)) %>%
 ggplot() +
  aes(x = reorder(Genre, +Vote, sum), y = Vote, fill = Platform) + # Sort bar chart in ascending order
  geom_col() +
  scale_fill_hue(direction = 1) +
  labs(
    x = "Genre",
    y = "Count of people voted",
    title = "Figure 2: Movie Trends Base on Genre (2022)",
    caption = "Data from FlixPatrol"
  ) +
  coord_flip() +
  theme_minimal()

movies %>%
  select(-Date) %>%
  filter(Genre == "Horror") %>%
  arrange(desc(Vote)) %>%
  head(3) %>%
  kable(
    caption = "Top Three Horror Movie",
    align = c(rep('c', 7))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 14,
    latex_options = "HOLD_position"
  )

freq_table <- tabyl(movies, Platform, Genre) %>%
  mutate(
    Other = 0
  )

name_temp <- c() # Store column that needs to be drop
for (i in 2:27){
  if (sum(freq_table[[i]]) < 5){
    freq_table[28] <- paste(freq_table[[i]], freq_table$Other, sep = "-") # Store all values to the "Ot
    name_temp <- append(name_temp, i)
  }
}

freq_table <- freq_table[-name_temp]

for (i in 1:nrow(freq_table)){ # Convert "other" column to actual integer values
  temp <- c(freq_table$Other[i])
  temp <- strsplit(temp, "-")
  temp <- as.numeric(unlist(temp))
  freq_table$Other[i] = sum(temp)
}

freq_table$Other <- as.numeric(freq_table$Other)

movie_freq <- freq_table %>%
  untabyl() %>%
  adorn_totals(c('row', 'col')) %>%  # add the total
  adorn_percentages(denominator = 'all') %>%  # add the relative percentage
  adorn_pct_formatting(digits = 2) %>%  # adjust the decimal point
  adorn_title(
    placement = 'combined',
    row_name = 'Platform',
    col_name = 'Genre'
  )

formatAF <- attr(movie_freq, 'core') %>%
  adorn_totals(where = c('row', 'col')) %>%
  mutate(
    across(where(is.numeric), format, big.mark = ',')
  )

movie_freq_fin <- movie_freq %>%
  adorn_ns(position = 'front', ns = formatAF)

# Polish (Return a image)
movie_freq_fin %>%
  kable(
    caption = 'Frequency between Platform and Genre',
    booktabs = TRUE,
    align = c('l', rep('c', 17))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 4,
    latex_options = "HOLD_position"
  )

write.csv(movies,"C:\\Users\\wue77\\Documents\\R file\\Movies.csv", row.names = FALSE)
```