Exploratory Data Analysis (Holiday Movies)

Md Nazmul Hossain (0843082)
A A Talha Talukder (XXXXXXX)
Nibedita Mallik (XXXXXXX)

Group number: 09
Project ID: 07

Introduction

Movie data introduction and the citation style we are going to use is IEEE [1]. The tidytuesday reference is here [2].

Data Collection

We have collected the data from tidytuesday [2].

The data we have collected is showing as below in Table 1.

Table 1: Top 10 renres by number of movies.

genres	Count
Comedy	1025
Drama	828
Romance	737
Family	707
Animation	268
Fantasy	185
Adventure	117
Documentary	101
Short	96
Music	91

Data Summaries

We are going to split this section into two sub-sections, as with numerical and graphical summaries.

Nemerical Summaries

The total number of observations from the survey received is: 2265. The table shows in Table 1.

Graphical Summaries

We will take help of several graphical plots here to describe the data. Listed as below.

Box Plot

Box plot details will go here ...

Bar Plot

Bar plot details will go here and sample with caption as below Figure 1.

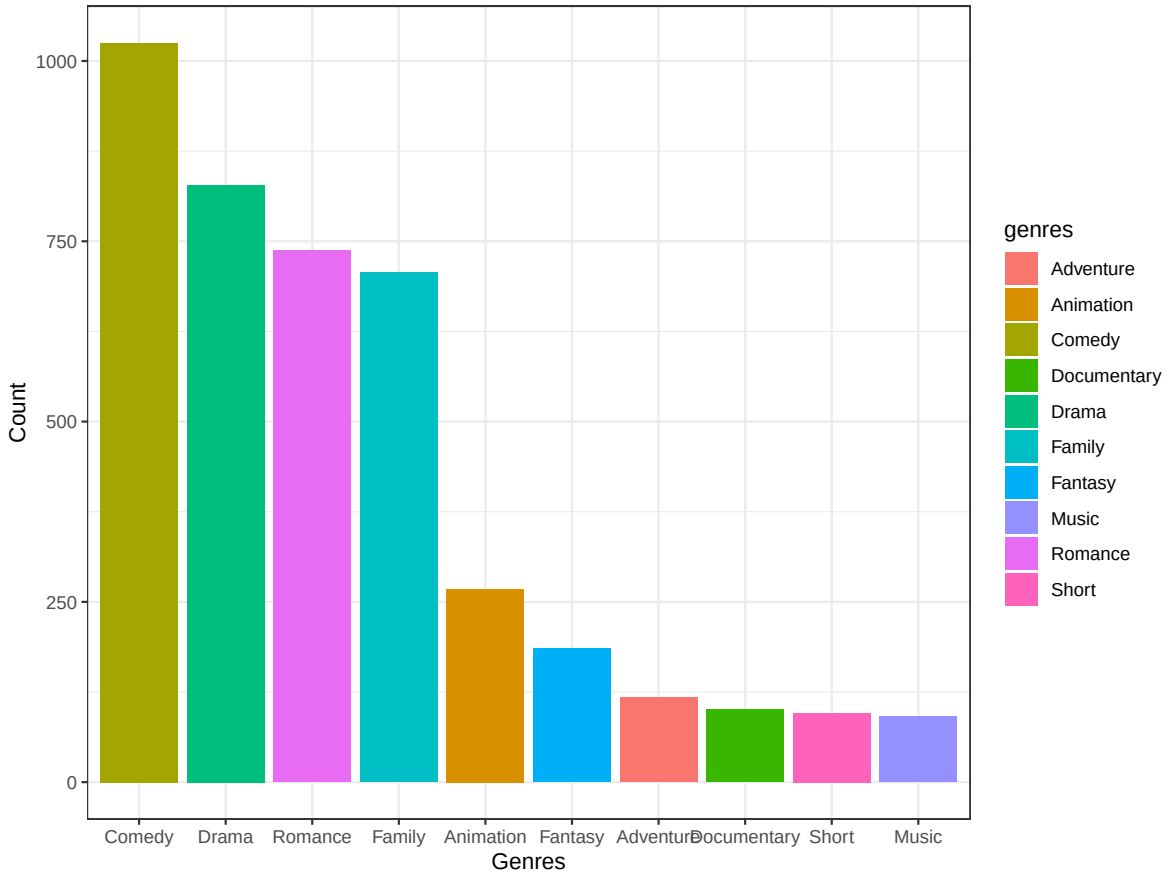

Figure 1: Top 10 genres by number of movies.

Linear Regression (e.g. with Scatter Plot)

Linear regression details.

Discussion

The discussion details.

Correlations (if applicable)

Correlation details.

Conclusion

Here goes the conclusion section!

References

[1] Q. He, Z. Feng, X. Wang, Y. Wu, and J. Yang, "A smart pen based on triboelectric effects for handwriting pattern tracking and biometric identification," *ACS applied materials & interfaces*, vol. 14, no. 43, pp. 49295–49302, 2022, doi: 10.1021/acsami.2c13714.

[2] D. S. L. Community, "Tidy tuesday: A weekly social data project." 2024. Accessed: Feb. 10, 2025. [Online]. Available: https://tidytues.day

Acknowledgements

- We would like to give thanks to the tidytuesday for the relevant data.
- Additionally, we would took help to debug tables, plot and data transformation codes in Stack overflow code snippets and Copilot.

Appendix

Code

```
# Some global settings, data preparation and variables to ease the usage of
 ↪  different variables
# without worrying to create or load in multiple places

# load tidyverse
library(tidyverse)
# load kableExtra
library(kableExtra)

# read data from github
holiday_movies <- readr::read_csv('holiday_movies.csv')
holiday_movie_genres <- readr::read_csv('holiday_movie_genres.csv')

# create a data frame where every genre name is a column
# and each will have value 0 or 1. Value 1 means it beolngs to that genre, 0
 ↪  means doesnt belongs to that
holiday_movies_ext <- holiday_movies %>%
  mutate(copy_genres = genres) %>%  # keep the original genres column intact
  separate_rows(copy_genres, sep = ",") %>%  # split genres into separate
   ↪  rows
  mutate(dummy = 1) %>%  # create a dummy variable for each genre
  pivot_wider(names_from = copy_genres, values_from = dummy, values_fill = 0)
   ↪  # pivot to wide format, fill with 0

# view the resulting data frame
# head(holiday_movies_ext)

# set theme
theme_set(theme_bw())
# get top 10 genres by movie count
top_10_gnr_mv_count <- holiday_movies %>%
  separate_rows(genres, sep = ",") %>%
  group_by(genres) %>%
  summarise(Count = n()) %>%
  arrange(desc(Count)) %>%
  head(10)
```

```r
# display the table
kable(top_10_gnr_mv_count, format = "latex", booktabs = TRUE)
# display the bar plot
ggplot(top_10_gnr_mv_count, aes(x = reorder(genres, -Count), y = Count, fill
 ↪  = genres)) +
  geom_bar(stat = "identity") +
  labs(x = "Genres", y = "Count")
```