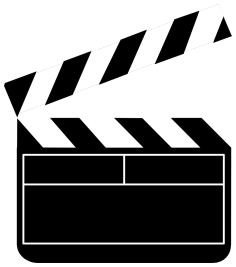

Exploring Movies

April 27, 2022

The dataset used for this challenge is titled 'movies.csv'.

What topic does the dataset cover?

This dataset contains information about the top 1000 highest grossing holywood films. It is up to date as of 10th January 2022.

Acknowledgements

This data has been scraped from multiple site and has been added together for performing various datat operations. The data has been taken from idmb, rotten tomatoes and many other sites. It can be found at https://www.kaggle.com/datasets/sanjeetsinghnaik/top-1000-highest-grossing-movies.

Variables

- title: Movie name
- movie_info: Information on the given movie
- distributor: distributor for movie
- release_date: When the movie was released
- domestic_sales: Domestic sales in dollars
- international_sales: International sales in dollars
- world_sales: World sales in dollars
- runtime: runtime of movie in minutes
- licence: licence of movie (G, PG, etc.)
- genre: Main genre of movie

Assignment

Complete each of the following questions using R Studio and submit your answers as a detailed report.

1. Visualize and describe the world sales of movies in millions of dollars.
2. Visualize and describe the distribution for movie runtimes in minutes.
3. Summarize how many movies each distributor has made.
   (a) Which 3 distributors have made the most movies.
   (b) Which 3 distributors have made the least movies?

4. Which three distributors have the highest average world sales?

5. Look at the world sales across the 5 distributors with the most amount of movies made. Does there appear to be a significant difference?

6. Count the number of movies that are in each genre. What is problematic about the way the data is coded?

7. Compare movie runtimes across the 4 genres that have made the most movies. Does there appear to be a significant difference? Use pairwise boxplots to support your answer.

8. Compare domestic sales across the 4 genres that produce the most movies. Does there appear to be a significant difference? Support your answer using pariwise boxplots.

9. Is there a linear relationship between the runtime of a movie and its world sales? Support your answer using a scatterplot.

10. Is there a linear relationship between the international sales of a movie and the domestic sales of a movie? Support your answer using a scatterplot.

11. If a movie made $750 million domestically, what would you predict it to make internationally?

12. Count the number of movies in each license category. Is there anything problematic with the data?

13. Summarize the runtime of movies across the various licence categories. Does there appear to be any significant difference?

14. What is your favourite movie and why?

Solution:

```
### Load packages and dataset
library(tidyverse)
library(ggplot2)
mov <- read.csv('/data/datasets/movies.csv')

### Take a look at the dataset
glimpse(mov)

### Question 1: Visualize the world sales of movies in millions of dollars
mov %>% mutate(mill = world_sales/1000000) %>%
  ggplot(aes(x=mill)) +
  geom_histogram(col='black', fill='green') +
  labs(x="Inrernational Sales (millions of dollars)")+
  theme_classic()

### Question 2: Visualize the distribution for movie runtime in hours
mov %>% ggplot(aes(x=runtime_minutes)) +
  geom_histogram(col='black', fill='red') +
  labs(x="Runtime in minutes") +
  theme_classic()

### Question 3A: Which 3 distributors make the most movies?
mov %>% group_by(distributor) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>% head(10)

### Question 3B: Which 3 distributors make the least movies?
mov %>% group_by(distributor) %>%
```

```
  summarise(count=n()) %>%
  arrange(count) %>% head(10)

### Question 4: Which three distributors have the highest average world sales?
mov %>% group_by(distributor) %>%
  summarise(mean = mean(world_sales)) %>%
  arrange(desc(mean))

### Question 5: International sales across 5 biggest distributors
top <- c("Warner Bros.", "Walt Disney Studios Motion Pictures",
         "Universal Pictures", "Twentieth Century Fox",
         "Sony Pictures Entertainment (SPE)")

mov %>% filter(distributor %in% top) %>%
  ggplot(aes(y=world_sales, fill=distributor)) +
  geom_boxplot() +
  theme_classic()

### Question 6: Number of movies in each genre
mov %>% group_by(genre) %>%
  summarise(count=n()) %>%
  arrange(desc(count))

### Question 7: Compare runtimes across top 4 genres
top <- c("Action", "Adventure", "Comedy", "Drama")
mov %>% filter(genre %in% top) %>%
  ggplot(aes(y=runtime_minutes, fill=genre)) +
  geom_boxplot() +
  theme_classic()

### Question 8: Compare domestric sales across top 4 genres
top <- c("Action", "Adventure", "Comedy", "Drama")
mov %>% filter(genre %in% top) %>%
  ggplot(aes(y=domestic_sales, fill=genre)) +
  geom_boxplot() +
  theme_classic()

### Question 9: Is there a relationship between runtime and world sales?
mov %>% ggplot(aes(x=runtime_minutes, y=world_sales)) +
  geom_point() +
  theme_classic()

### Question 10: Is there a relationship between international sales and domestic
    sales?
mov %>% ggplot(aes(x=domestic_sales, y=international_sales)) +
  geom_point() +
  geom_smooth() +
  theme_classic()

### Question 11 Eyeball prediction using smoothed line

### Question 12: Count number of movies in each license category
mov %>% group_by(license) %>%
  summarise(count = n())

### Question 13: compare the runtime for movies in each license category
mov %>% ggplot(aes(y=runtime_minutes, fill=license))+
  geom_boxplot() +
  theme_classic()


```