Movies

Juliana Rosa

March 2, 2019

Source: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

Cleaning the Data

```
#Reading the csv file that contains the raw data and storing it into
"movies_data"
setwd("~/DATA110/Exam_1")
movies_data <- read.csv("movies_rawdata.csv")

#Loading the dplyr package
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

#Taking a look at the data
View(movies_data)

#Removing unwanted columns
movies_data <- select(movies_data, -(ends_with("likes")), -
(starts_with("num")), -(facenumber_in_poster), -(movie_imdb_link), -
(aspect_ratio), -(plot_keywords))

#Reordering columns
movies_data <- select(movies_data, movie_title, title_year, duration,
genres, content_rating, director_name, country, language, imdb_score,
color, budget, gross, actor_1_name, actor_2_name, actor_3_name)

#Renaming some columns
movies_data <- rename(movies_data, movie=movie_title,
duration_in_min=duration, year=title_year)
```

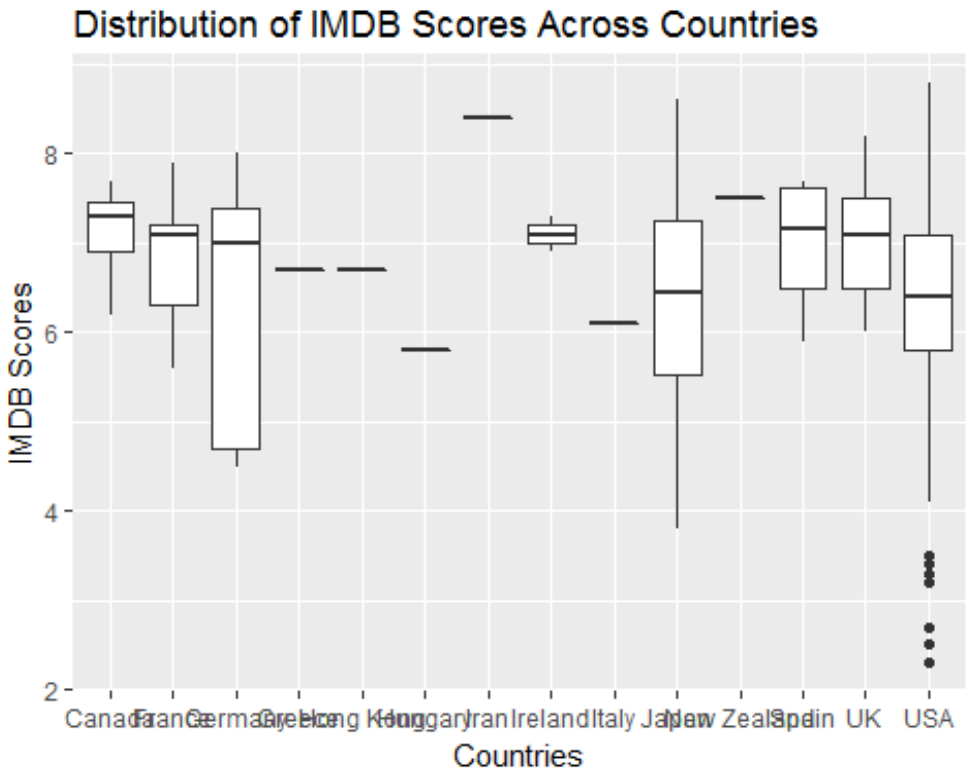
```r
#Removing rows with missing values
movies_data <- filter(movies_data, complete.cases(movies_data))
movies_data <- filter(movies_data, movie!="", year!="",
duration_in_min!="", genres!="", content_rating!="", director_name!="",
country!="", language!="", imdb_score!="", color!="", budget!="",
gross!="", actor_1_name!="", actor_2_name!="", actor_3_name!="")

#Correcting typos in the movie column
movies_vector <- as.character(movies_data$movie)
movies_list <- list()
for (i in 1:3830){
  movies_list <- append(movies_list, movies_vector[i])
}
library(stringr)
for (i in 1:3830){
  movies_list[i] <- str_sub(movies_list[i], start=1,
end=(str_length(movies_list[i])-2))
}
movies_vector <- as.character(movies_list)
movies_factor <- as.factor(movies_vector)
movies_data$movie <- movies_factor

#Simplifying the values of color
color_vector <- as.character(movies_data$color)
color_list <- list()
for (i in 1:3830){
  color_list <- append(color_list, color_vector[i])
}
for (i in 1:3830){
  if (color_list[[i]]=="Color"){
    color_list[i] <- 1
  }
  else{
    color_list[i] <- 0
  }
}
color_vector <- as.character(color_list)
color_factor <- as.factor(color_vector)
movies_data$color <- color_factor

#Creating a profit column
movies_data <- movies_data%>%
  mutate(profit=gross-budget)%>%
  select(movie, year, duration_in_min, genres, content_rating,
director_name, country, language, imdb_score, color, gross, budget,
profit, actor_1_name, actor_2_name, actor_3_name)

#Organizing in cronological order
movies_data <- arrange(movies_data, year)
```

```r
#Taking another look at the data
View(movies_data)
```

Creating Visualizations

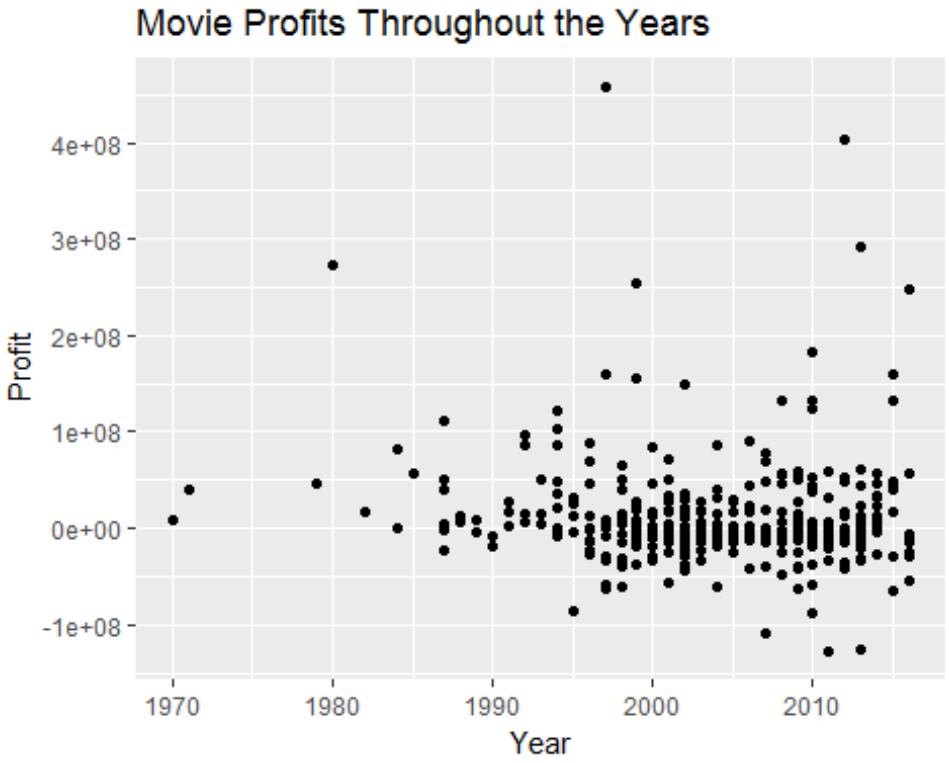
I'm going to use samples because the original datset is too big.

```r
#Taking a sample
set.seed(1)
my_sample <- movies_data[sample(1:nrow(movies_data), 383), ]

#Loading the ggplot2 package
library(ggplot2)

#Creating a boxplot
ggplot(my_sample, aes(country, imdb_score))+
  geom_boxplot()+
  ggtitle("Distribution of IMDB Scores Across Countries")+
  xlab("Countries")+
  ylab("IMDB Scores")
```

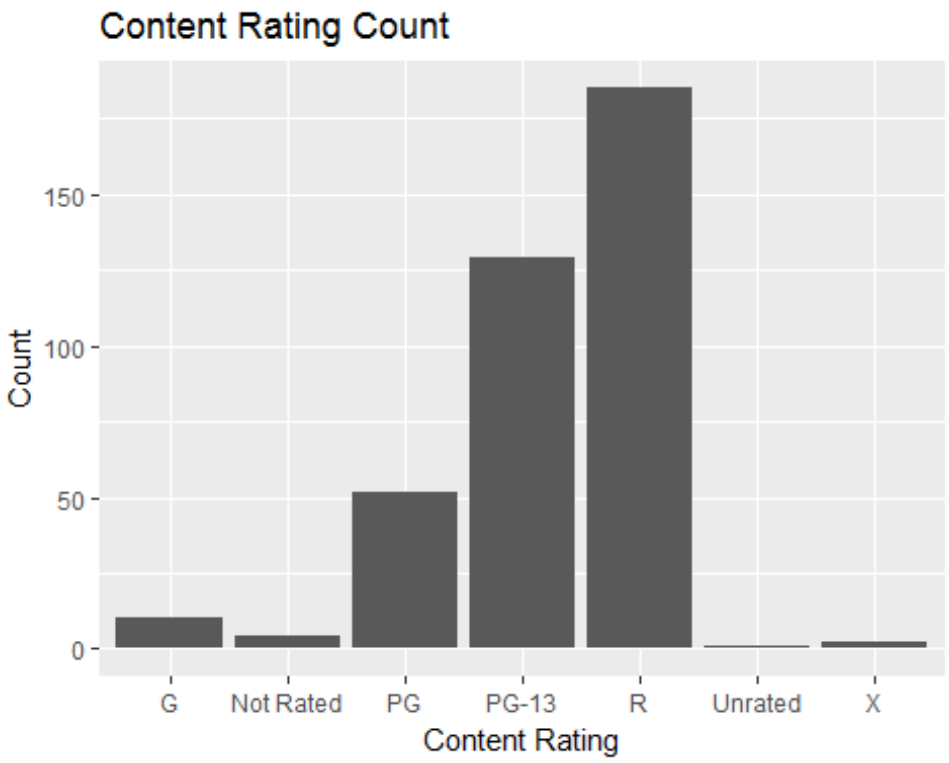



```r
#Creating a scatterplot
ggplot(my_sample, aes(year, profit))+
  geom_point()+
  ggtitle("Movie Profits Throughout the Years")+
```

```
  xlab("Year")+
  ylab("Profit")
```




```
#Creating a bar chart
ggplot(my_sample, aes(content_rating))+
  geom_bar()+
  ggtitle("Content Rating Count")+
  xlab("Content Rating")+
  ylab("Count")
```



Essay

I wanted to use a dataset that was about something that interests me, so I picked the topic "movies". I had to look at a lot of data before choosing the IMDB 5000 Movie Dataset at Kaggle. The raw data contained twenty eight variables, but after I cleaned it there were only sixteen left, ten of them being categorical (movie, genres, content_rating, country, language, color, actor_name_1, actor_name_2, and actor_name_3) and the other six quantitative (year, duration_in_min, imdb_score, gross, budget, and profit).

Since the dataset was so large (5044 in its raw form), there was a lot of cleaning to do. For that I used the dplyr package in R: removed unwanted columns and reordered the remaining ones with the "select" function, renamed columns using the "rename" function, removed rows with missing values (either in NA form, or as blank spaces) using the "filter" function, corrected typos and changed values using for loops and if/else conditions, created a new column with "mutate", and at last reordered the rows of the data using the "arrange" function.

When the data was ready, I took some samples and used them to create visualizations with ggplot2 in R. The country x imdb_score boxplot shows the distribution of movie scores for each country. We can see that the US has the wider range of scores and, curiously, is the only country that has outliers. Both of these things are probably due to the United States being one of the biggest film production countries in the world (more movies mean more scores variety). The boxplots that are

being represented by a single line are movies from countries like Iran or Hungary that are not that big on film production, but happened to be selected in the random sampling process. It is also interesting to note that the data is centered somewhere around score 7, with not that much spread.

The year x profit scatterplot displays the movies profit along the years. It is very clear from this plot how much the cinematographic industry has grown, since there weren't many movies until the 90's. After 1990, a pattern can be seen: most movies are actually centered on 0. That doesn't mean that there are a lot of films that have zero profit, it just means that most of them have low profit, or actually loses a bit of money (negative numbers for the profit). What really caught my attention were the outliers, so I took the liberty to check three of them: the two highest profits were from Titanic in 1997, followed by The Avengers in 2012, and the other one is the only movie we have from 1980, which has a very high profit in comparison to the other movies from close years or even to the most recent ones. It is, of course, a Star Wars movie, more specifically Episode V- The Empire Strikes Back.

The content_rating bar chart shows the movie count for each rating category. We may say that the biggest movie audience is composed by adults or young adults, since the count for R rated movies is way higher than any other category. Generally speaking, the more restricted the rating is, the higher its count is, and the unrated or not rated movies are not common at all. Besides that, the X rated category only contains two movies and, as I guessed, they are relatively old, since this rating is no longer being used by MPAA. I checked and the two movies are A Nightmare on Elm Street 3 (1987) and Beyond the Valley of the Dolls (1970).

In short, I decided to create more than one visualization because this way I would be able to work with many different variables without adding too much information to one plot and making it confusing. Although it would have been nice to include colors or shapes in one of the graphs, my qualitative variables all had too many categories, so these aesthetic wouldn't have worked very well. I also wanted to make a contingency table, but ended up not doing it because it's harder to interpret than graphs. Besides that, I pretty much added everything I wanted and was happy with the interesting and informative results.