Data Visualization Project: Blockbuster Films

Charlotte Kerr (217456559), Samantha Gylys (210177764), and

Oluwanifemi Osundina (216752099)

Department of Psychology, York University

PSYC 3031 A: Intermediate Statistics Laboratory

Dr. Monique Herbert

November 4, 2020

Data Visualization Project: Blockbuster Films

The original dataset contains information about the top ten highest grossing films each year, from 1975 to 2018. Various variables were included in the data, but for the purpose of this project, we selected certain parts of the data to be used. The variables of interest to us include the main genre of the film, its IMDb rating, and its Motion Picture Association rating.

The genres of the original data include the following: action, adventure, animation, comedy, crime, drama, family, fantasy, history, horror, music, romance, sci-fi, sport, thriller, and war. However, we chose three popular genres for our visualization: Action, comedy, and drama. The IMDb rating is measured on a 10-point scale. A film's rating is determined by the reviews of audience members and film critics, and the computed rating is just an average of all the ratings left for the movie. A rating of 1 indicates that the film was poorly received by critics and the audience, while a rating of 10 indicates that critics and the audience left positive reviews and enjoyed the film.

The Motion Picture Association rating includes four types: G, PG, PG-13 & R (in ascending order, from suitable for all viewers to restricted).

G: General Audiences

PG: Parental Guidance Suggested

PG-13: Parents Strongly Cautioned

R: Restricted

Typically, more violent or suggestive movies, such as dramas and action movies have a stronger rating (PG-13 and R). Depending on the specific movie, comical movies tend to require little to no viewers discretion, making it suitable for younger audiences (PG, G), still retaining appeal for older audiences.

Movies with a G rating are accessible to all ages of viewers, and there is a possibility that it appeals to everyone, leading to a less critical analysis by the viewers and a higher IMDb rating.

Movies with an R rating are more restricted, and the possibility exists that depending on the audience, more explicit movies might be rated lower. Taking all this into consideration, there might be a relationship between these variables, and this is the rationale behind our visualization question.

Visualization Question

Does a film's genre and its Motion Picture Association rating have an effect on the film's IMDb rating? Other factors to consider include the following:

a) What genre has the highest IMDb rating in relation to its rating?

b) For the different genres, does the IMDb rating depend on the movie rating?

c) Does the movie rating (G, PG, PG-13, R) depend on the genre of the movie? For example, is an action movie less likely to be rated G?

d) Is there a consistent trend in the ratings of movies in relation to the IMDb rating across all genres? For instance, does a certain rating (G, PG, PG-13, R) consistently receive a higher IMDb rating across all genres?

Goals of the Visualization

By visualizing this data, the research group aims to view the relationship between a film's genre, IMDb rating, and its Motion Picture Association rating. Do movies with a rating of

R have a higher IMDb rating? Do movies with a rating of G have a lower IMDb rating? Is there a noticeable trend in IMDb ratings when genre and rating are considered?

The visualization will make it easier to detect patterns, trends, and outliers in the dataset. This way, the research group can determine whether the genre and Motion Picture Association rating have any noticeable effect on the IMDb rating and whether this potential relationship has a direction. This will ultimately be a useful feedback for filmmakers when considering the content of their work.

R Analysis

Final_data_visualization.R

morol

2020-11-04

```
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------------
- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(tibble)
library(dplyr)
```

```r
library(readr)
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.0.3
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

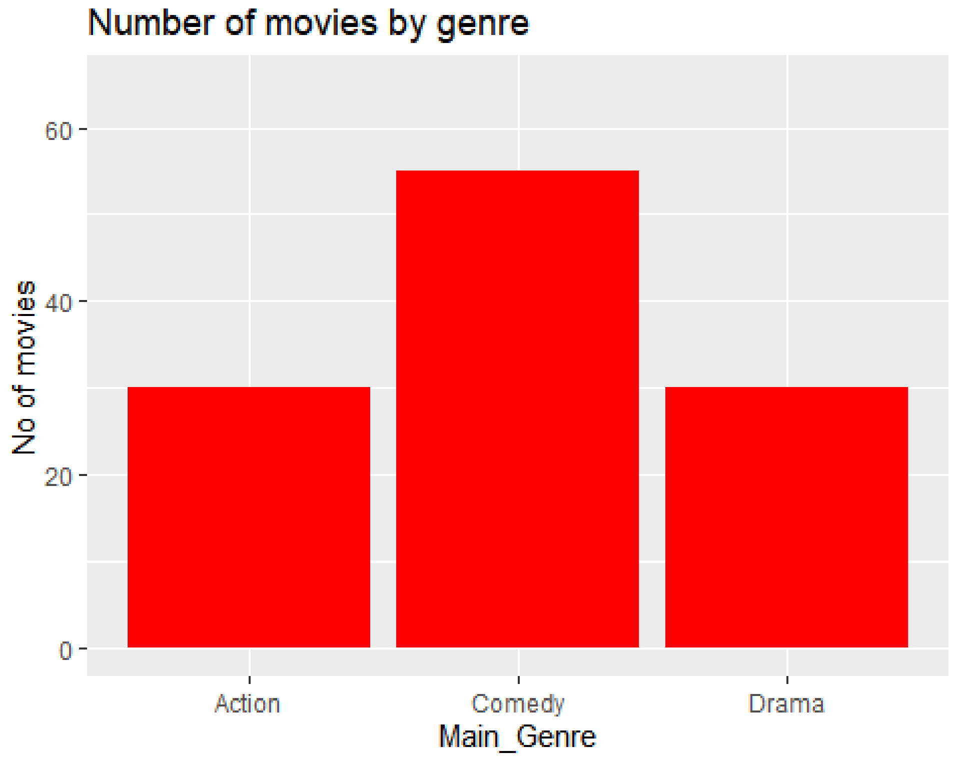
```
## The following object is masked from 'package:purrr':
##
##     transpose
```

```r
blockbusters_ <- read_csv("data/blockbusters .csv")
```

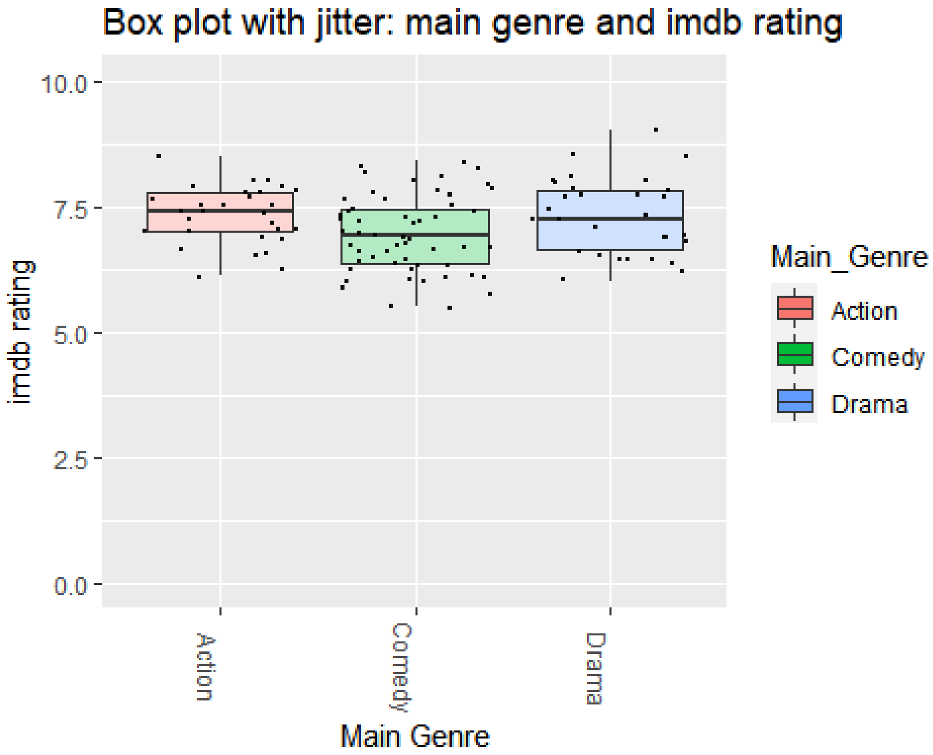
```
## Parsed with column specification:
## cols(
##   Main_Genre = col_character(),
##   Genre_2 = col_character(),
##   Genre_3 = col_character(),
##   imdb_rating = col_double(),
##   length = col_double(),
##   rank_in_year = col_double(),
##   rating = col_character(),
##   studio = col_character(),
##   title = col_character(),
##   worldwide_gross = col_character(),
##   year = col_double()
## )
```

```r
View(blockbusters_)

#Filter out desired variables of interest: Main genre, rating, imdb rating
blockbusters_ <- select(blockbusters_, Main_Genre,rating, imdb_rating, year)
view(blockbusters_)

#For some reason, cannot filter together as it says data needs to be logical
#we tried mutating and converting data unsuccessfully
#we chose to independently filter to bypass this, then combine filtered resul
ts

#Filter out: Action
Action <- filter(blockbusters_, Main_Genre == "Action")
#Filter out: Comedy
Comedy <- filter(blockbusters_, Main_Genre == "Comedy")
#filter out: Drama
Drama <- filter(blockbusters_, Main_Genre == "Drama")
```

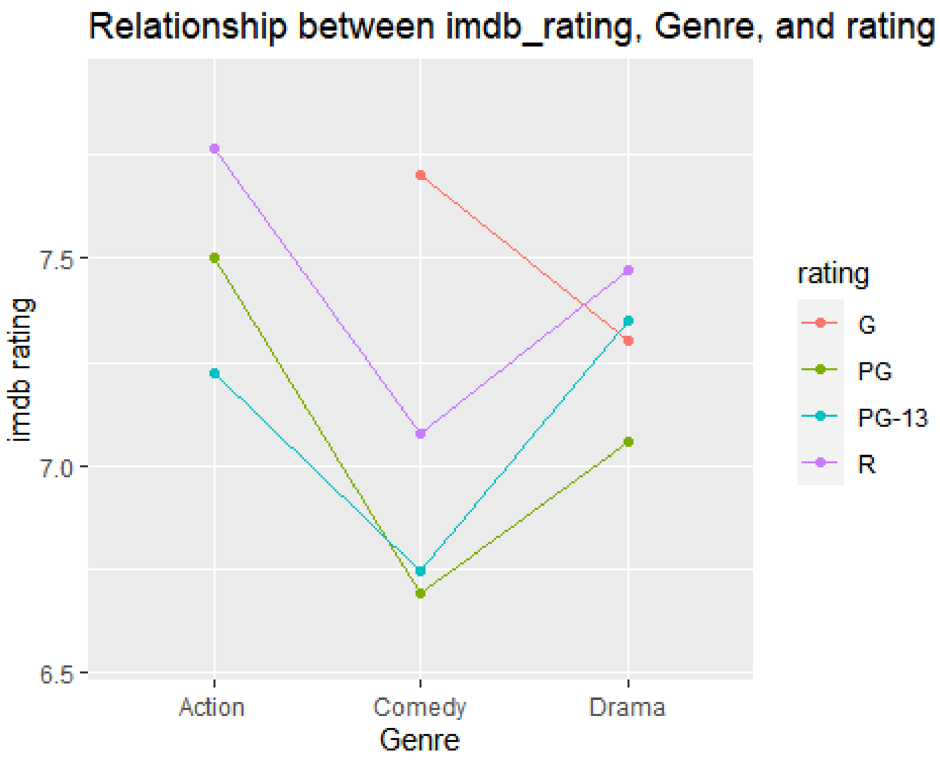
```
#Bind rows and assign to object "Comedy Action Data"
Comedy_action_data <- bind_rows(Action, Comedy, Drama)
View(Comedy_action_data)

#duplicating the data so that when we plot our last graph by grouping then fi
nind the average, it would not
#group our main data and prevent us from using .functions
Comedy_Action_data2<- Comedy_action_data

#interested in the number of movies per genre.
Genre_Count <- count(Comedy_action_data, Main_Genre)
View(Genre_Count)

#add ID column to keep track when viewing data
Comedy_action_data <- tibble::rowid_to_column(Comedy_action_data, "ID")
view(Comedy_action_data)

#Plot a bar graph showing number of movies for each genre (Graph 1)
ggplot(data = Comedy_action_data,
       mapping = aes(x = Main_Genre)) +
  geom_bar(fill = "Red") +ylim(c(0,65)) + labs(title = "Number of movies by g
enre") + labs(y ="No of movies")
```



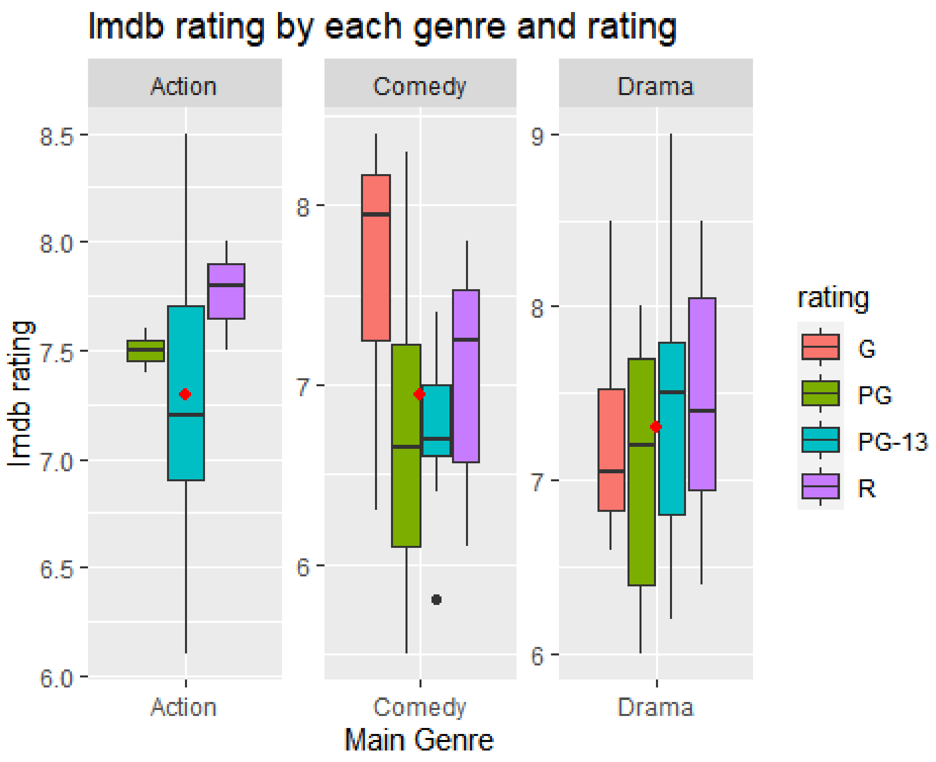
```
#Boxplot comparing genre and rating (Graph 2)
#Add Jitter so we can view distribution of data across these categories
ggplot(data = Comedy_action_data,
       mapping = aes(x = Main_Genre, y = imdb_rating)) +
  geom_boxplot(aes(fill = Main_Genre)) + theme(axis.text.x =element_text(angl
e = 270)) + labs(title =

"Box plot with jitter: main genre and imdb rating") +
  labs(x = "Main Genre", y = "imdb rating") + ylim(c(0,10)) + geom_boxplot(al
pha=0.7) + geom_jitter(color="black", size=0.3, alpha=0.9)
```



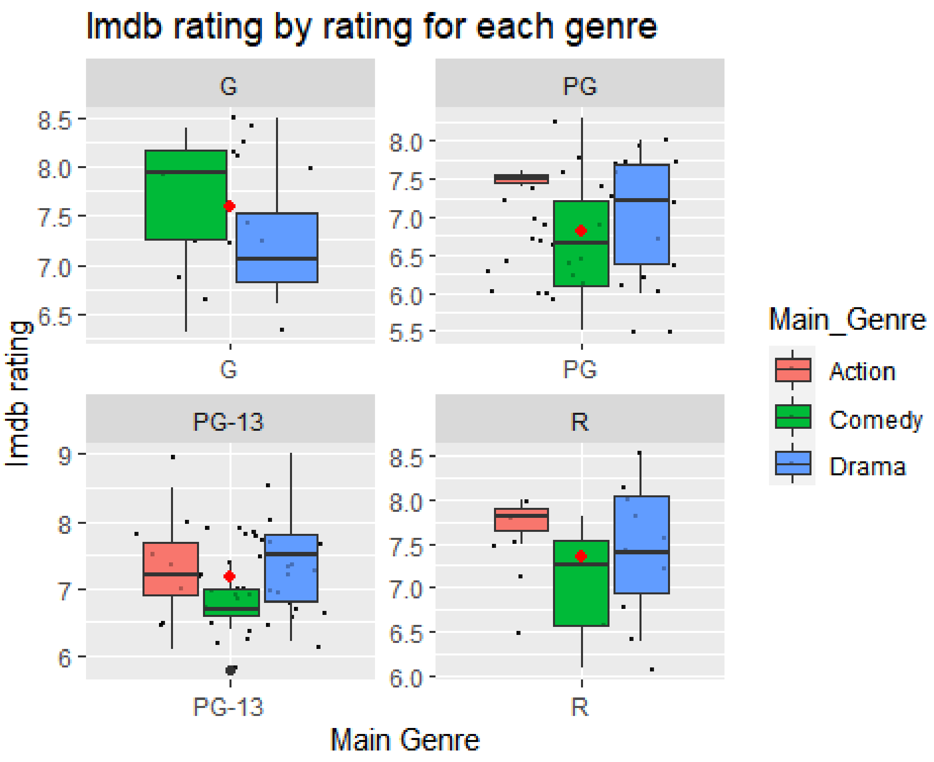
```
#line graph showing the relationship between imdb_rating, Genre, and rating (
graph 3)
ggplot(data=Comedy_action_data, aes(x=Main_Genre, y=imdb_rating, group=rating
, color=rating))+
  geom_line(stat='summary') +
  geom_point(stat='summary') + labs(title = "Relationship between imdb_rating
, Genre, and rating ") + labs(x="Genre", y="imdb rating")

## No summary function supplied, defaulting to `mean_se()`

## No summary function supplied, defaulting to `mean_se()`
```



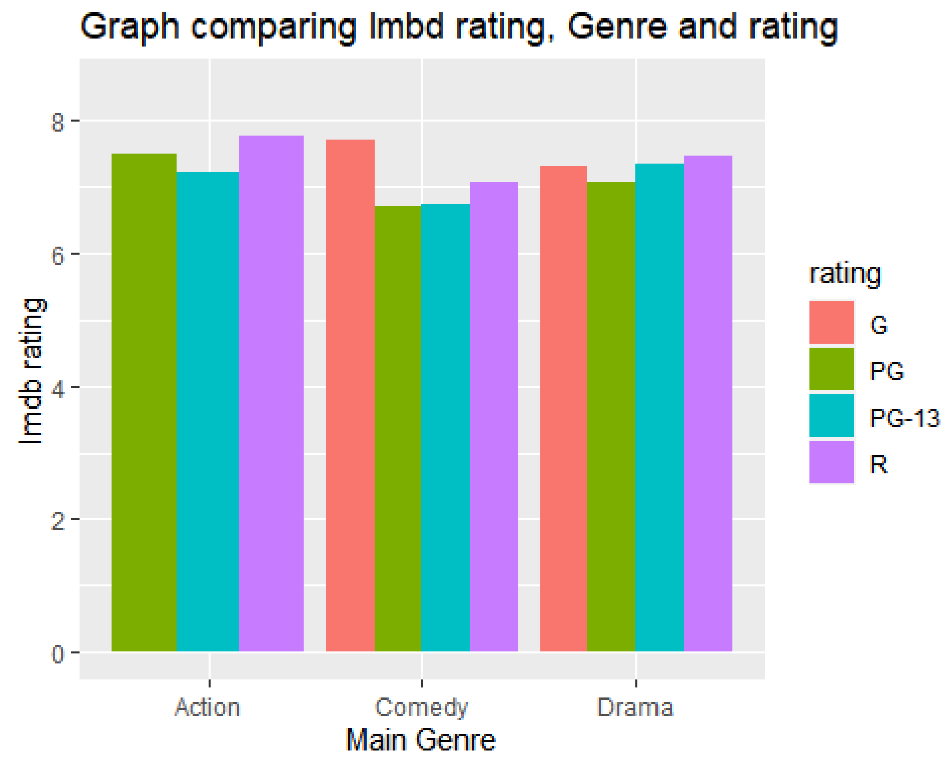
```
#Create a box plot for each Genre type comparing rating and imdb rating (Graph 4)
ggplot(data=Comedy_action_data, aes(x=Main_Genre, y=imdb_rating, fill=rating)) +
  geom_boxplot() +
  facet_wrap(~Main_Genre, scale="free") + labs(title = "Imdb rating by each genre and rating", y = "Imdb rating", x = "Main Genre") + geom_boxplot(alpha=0.7) + stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red")
```



```
#Unfortunately we could not figure out how to locate the means for each ratin
g, so we just have means for each genre

#we can also view this data the other way around, therefore
#Box plot for each rating, comparing Main Genre and imdb rating (Graph 5)
ggplot(data=Comedy_action_data, aes(x=rating, y=imdb_rating, fill=Main_Genre)
) +
  geom_boxplot() +
  facet_wrap(~rating, scale="free") + labs(title = "Imdb rating by rating for
each genre", y = "Imdb rating", x = "Main Genre") +  geom_jitter(color="black
", size=0.3, alpha=0.9)+geom_boxplot(alpha=0.7) + stat_summary(fun=mean, geom
="point", shape=20, size=3, color="red", fill="red")
```



```
#Multiple barchart visualizing the relationship between the three variables (
Graph 6)
#first, find the average IMDB ratings of the ratings (P, PG, PG-13 and R) by
Genre
#loaded library data table earlier on we will use data.table to do the groupi
ng. We got stuck using "Group_by"

#Find the averages
Average_ratings <- as.data.table(Comedy_Action_data2)[, mean(imdb_rating), by
= .(Main_Genre, rating)]
view(Average_ratings)

#Plot graph
ggplot(data = Average_ratings, aes(x = Main_Genre, y = V1, fill = rating)) +
  geom_col(position = position_dodge()) + labs(title = "Graph comparing Imbd
rating, Genre and rating", x = "Main Genre", y = "Imdb rating") + ylim(0,8.5)
```



Limitations

One limitation of the analysis is that only three genres were analyzed. Some of the omitted genres could potentially have shown a stronger or weaker interaction. Furthermore, some genres had a wider range of Motion Picture Association ratings than others. It is possible that genres with stricter ratings might have a smaller audience and therefore, have fewer IMDb reviews from audience members and critics. Additionally, films that are oriented towards families may have a wider viewership and consequently, have more IMDb reviews, giving a clearer picture of how the film was received.

Another limitation is that films can have multiple genres. There is no official system for determining whether a film, for example, is more of a comedy and less of a romance. Thus, it is possible that some films should have been included in the analysis but were labelled as another genre or that some films were included but should have been omitted.

The IMDb rating is an average and is not always an accurate measure of the overall sentiment of viewers. Sometimes, a film that is loved by audience members is poorly received by critics, skewing the average rating. There is also the possibility that some of the reviews are paid ones and not a true reflection of the viewer's feelings about the film, or that the film was given poor reviews in retaliation for a controversial story or topic. In addition, the rating is determined by a small subset of the population who register on the IMDb website. Thus, the rating is not necessarily an accurate picture of how the film was received by wider audiences.

Regarding the data, there is an unequal number of movies in each genre, this could slightly offset the reported average IMDb ratings.

The use of box plots is also a limitation, as boxplots emphasize the tails of distributions. We were able to see the highest and lowest values for this dataset. We were unable to successfully code the means of each level into the box plots, giving us less certain information to work with as means were harder to locate. However, we were able to code the means by the groups as indicated by the red dot.

Graphs

Graph 1: Bar chart showing the number of movies for each genre. It is seen that there is an even number of Action and drama movies, with 30 movies each, while Comedy has the highest number of movies, 55.

Graph 2: Box plot comparing both genre and rating. This reflects the minimum and maximum ratings for each of the genres, as well as the medians. We see from the graph that drama has both the minimum rating, as well as the maximum rating compared to Action and comedy genres. The medians of all three genres do not appear to be extremely skewed.

Graph 3: Line graph showing the relationship between IMDb rating, Genre, and rating. With the line graph, it is indicated that

- G rated movies had the highest IMDb rating for comedy than for drama
- R rated movies have a highest IMDb rating for action than for drama and comedy
- PG-13 movies had the highest IMDb rating when the genre was Drama rather than Action
- PG movies had the highest IMDb rating when they were action than when they were Drama or comedy. Although, they were rated higher for drama than for comedy

Graph 4: Create a box plot for each Genre type comparing rating and IMDb rating: This graph reflects the IMDb ratings for the different viewer ratings by genre. We have included the means by group, as well as jitter to reflect distribution of the data

Graph 5: Box plot for each rating, comparing Main Genre and IMDb rating: The data was looked over in another way, this time we looked at the IMDb ratings for the different viewer ratings by genre. We have included the means by group, as well as jitter to reflect distribution of the data

Graph 6: Multiple bar chart visualizing the relationship between the three variables. This graph shows the average IMDb rating of the different movie genres (drama, action, and comedy) categorized by their Motion Picture Association rating.

References

Chanda, B. (2018). *Top 10 Highest Grossing Films (1975-2018)* (Version 1) [dataset and table chart]. Retrieved from https://www.kaggle.com/bidyutchanda