Quiz: Data Displays in R

Ryan Cheng

10/19/2023

The following quiz is open book and open notes (anything I've posted on Canvas counts as notes!). You may use the internet and ChatGPT but you need to indicate when you are using it either as a comment in the code or right above or below the code chunk. For internet resources that you use, please include a link to the website. When you use ChatGPT to help you code, note specifically what you used it for. For example: "used ChatGPT to figure out why I was getting an error here", "asked ChatGPT how to rename the levels in a factor", "Asked ChatGPT to help me figure out how to subset here", "asked ChatGPT to help me figure out if this answer makes sense"). You may NOT give or receive help from any other students, teachers, or anyone else. You may NOT share this rmd and data with anyone else, even a blank version.

When you are finished please knit the file to pdf and upload it to Canvas. Every question should be answered in the R markdown file. **All answers should be in complete sentences unless the questions only asks for a value or graph to be printed by R.** All graphs should contain suitable titles, be fully labeled, and have colors. They should have appropriate axis lengths and be easy to read. Legends should be visible and not overlap any data. See the rubric on Canvas for more information about how this will be graded.

If you are stuck because of an error or other issue you may email me attaching the rmd. I will look at your code and see what the issue is and depending on the problem I may be able to give you some assistance but you may lose partial credit. I will let you know if it is an issue that would lose you partial credit before I provide advice so that you can decide if you would like help or not.

Read in the file `clean.movie.data.2022.csv`. To find out more about this dataset go to: https://www.kaggle.com/datacat0/all-us-movies-imdb-from-1972-to-2016 . Note that this dataset actually contains data through 2019, not 2016 as the website says.

```
clean.movie.data = read.csv("clean.movie.data.2022.csv")
head(clean.movie.data)
```

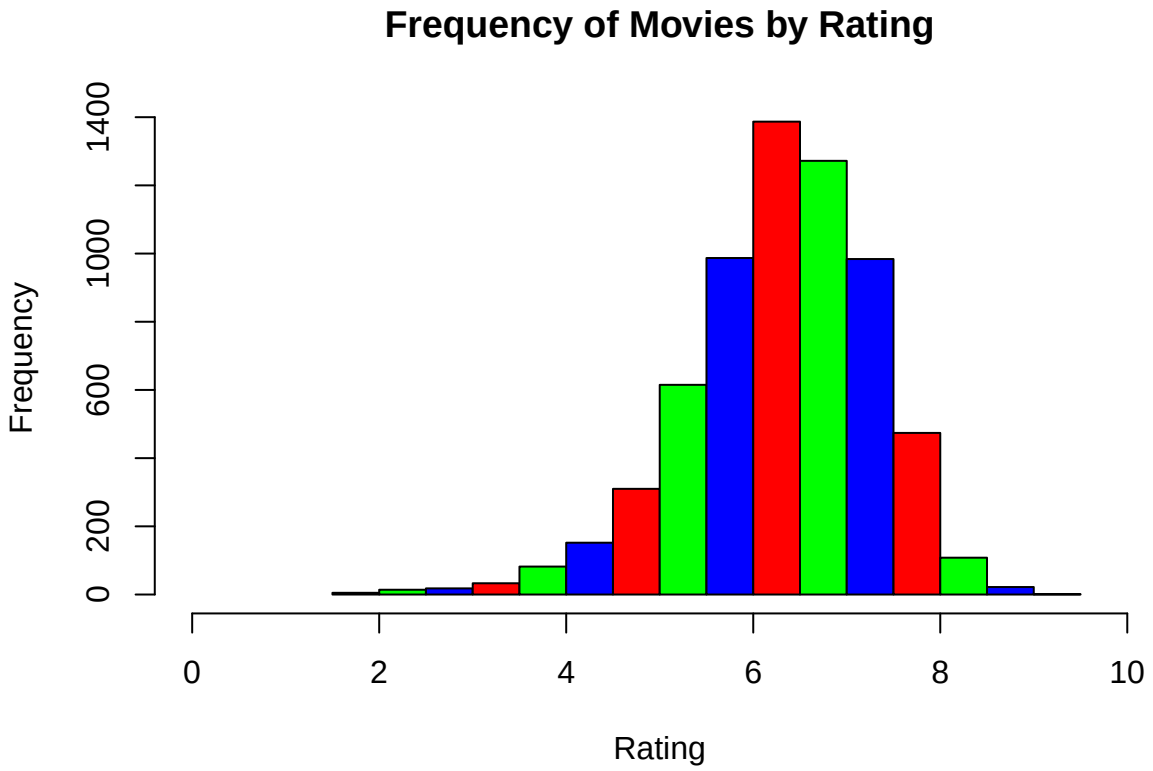
```
##   X.1 X                         title year runtime  genre1 certificate rating
## 1   2 2    Alrededor de la medianoche 1986     133   Drama          PG    7.4
## 2   3 3      Hasta que te encontr\xe9 1997     113 Romance       PG-13    4.8
## 3   4 4           Buenas noches madre 1986      96   Drama       PG-13    7.6
## 4   5 5 Nuestros maravillosos aliados 1987     106  Comedy       PG-13    6.7
## 5   8 8          10 la mujer perfecta 1979     122  Comedy           R    6.1
## 6   9 9          Calle Cloverfield 10 2016     103   Drama          PG    7.2
##             director
## 1 Bertrand Tavernier
## 2       Scott Winant
## 3          Tom Moore
## 4    Matthew Robbins
## 5      Blake Edwards
## 6   Dan Trachtenberg
##                                                                      cast
## 1    Dexter Gordon|Fran\xe7ois Cluzet|Gabrielle Haker|Sandra Reaves-Phillips
## 2              Jeanne Tripplehorn|Dylan McDermott|John Plumpis|Janel Moloney
## 3                          Sissy Spacek|Anne Bancroft|Ed Berke|Carol Robbins
```

```
## 4                    Hume Cronyn|Jessica Tandy|Frank McRae|Elizabeth Pe\xf1a
## 5                          Dudley Moore|Bo Derek|Julie Andrews|Robert Webber
## 6 John Goodman|Mary Elizabeth Winstead|John Gallagher Jr.|Douglas M. Griffin
##      gross  votes
## 1  3272600   4351
## 2  3478370   2558
## 3   441863   1991
## 4 32945797  29310
## 5 74865517  14280
## 6 72082998 268516
```

Please answer the following questions.

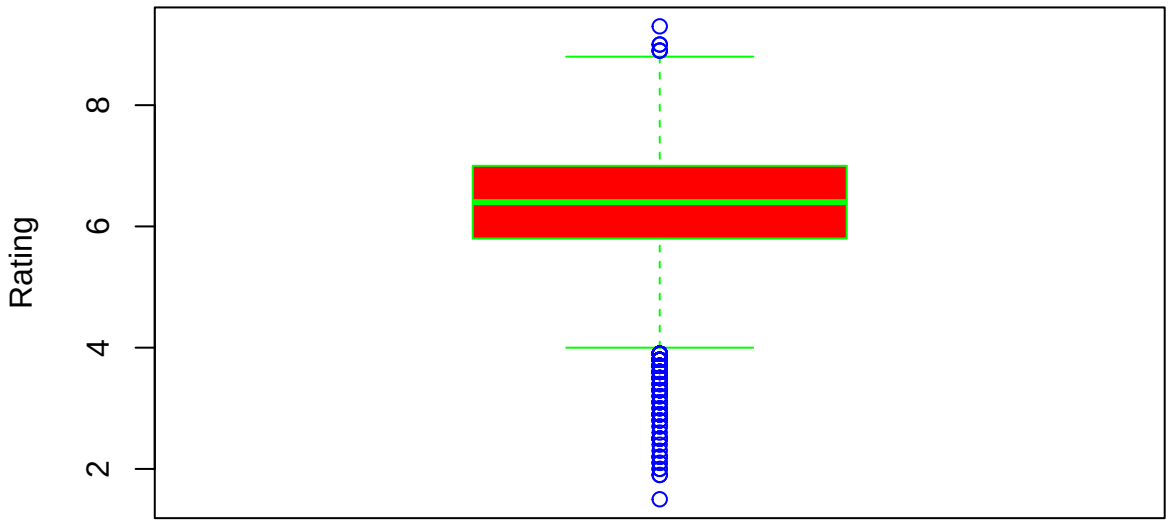
1. Create a histogram of the variable `rating`.

```
hist(clean.movie.data$rating,
     main = "Frequency of Movies by Rating",
     col = c("red", "green", "blue"),
     xlab = "Rating",
     xlim = c(0, 10))
```



2. Create a boxplot of the variable `rating`.

```
boxplot(clean.movie.data$rating,
        main = "Movies by Rating",
        ylab = "Rating",
        col = "red",
        border = "green",
        outcol = "blue")
```

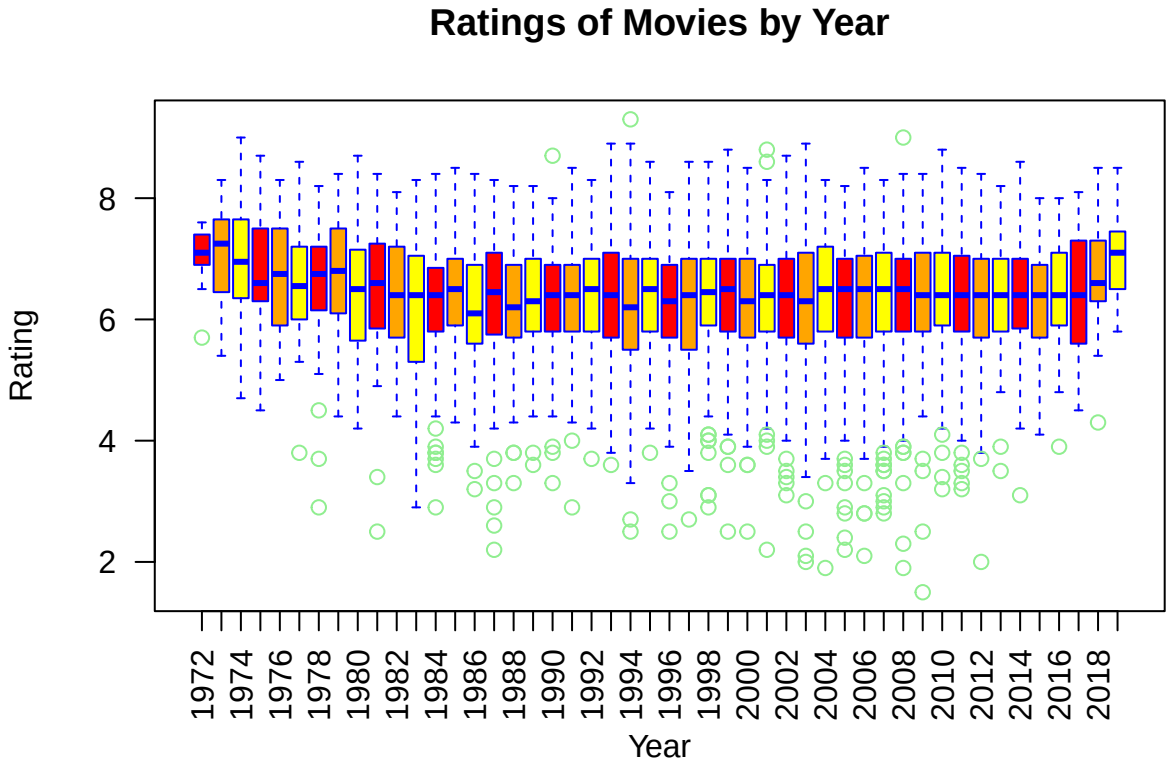


3. Describe the shape, center, and spread of the variable `rating`. Be thorough! Feel free to do any additional calculations in R but make sure you answer this question in full sentences.

The variable `rating` is skewed left. It is also unimodal. The median is approximately 6.5. The range is approximately 8, and the IQR appears to be roughly 1.25.

4. Make a comparative box plot of `rating` by `year`.

```
boxplot(clean.movie.data$rating~clean.movie.data$year,
        las = 2,
        main = "Ratings of Movies by Year",
        ylab = "Rating",
        xlab = "Year",
        col = c("red", "orange", "yellow"),
        border = "blue",
        outcol = "lightgreen")
```



5. What insight(s) do we get from the graph you made in problem 4? Give at least three specific insights, making sure to use proper vocabulary.

The median rating is relatively stable; it doesn't fluctuate massively. There are a lot more outliers on the lower end of the rating scale than on the higher end. The IQRs also seem to be relatively stable as in they don't fluctuate massively.

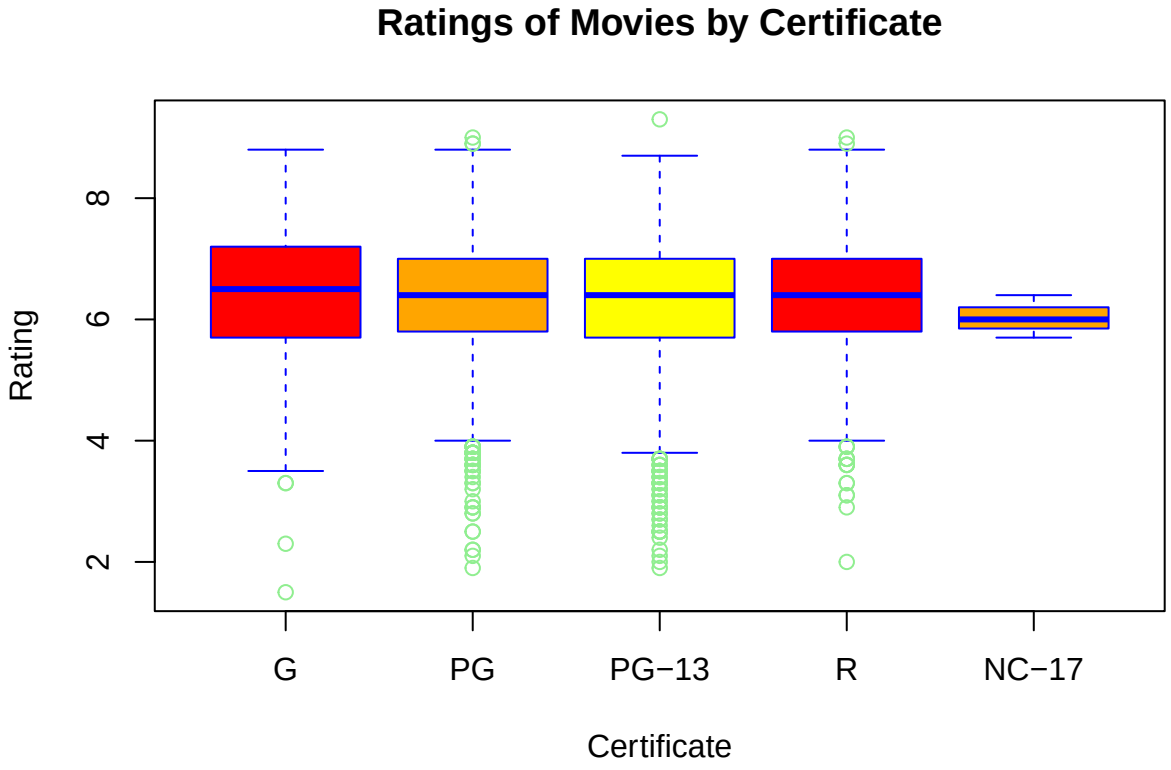
6. See parts a and b.

a. Make a comparative box plot of `rating` by `certificate`. Note that the rating scale G, PG, PG-13, R, NC-17 is not currently in the correct order. *For full credit the order must be changed.*

```
clean.movie.data$certificate = as.factor(clean.movie.data$certificate)

correct.order.certificates = c("G", "PG", "PG-13", "R", "NC-17")

clean.movie.data$certificate <- factor(clean.movie.data$certificate, levels = correct.order.certificates

levels(clean.movie.data$certificate)
```

```
## [1] "G"     "PG"    "PG-13" "R"     "NC-17"
```

```
boxplot(clean.movie.data$rating~clean.movie.data$certificate,
        main = "Ratings of Movies by Certificate",
        ylab = "Rating",
        xlab = "Certificate",
        col = c("red", "orange", "yellow"),
        border = "blue",
        outcol = "lightgreen")
```



b. What insight(s) do we get from this graph? Give at least three specific insights, making sure to use proper vocabulary.

There are no outliers in the NC-17 boxplot. The median rating among all certificate levels are very close to each other. The IQR for the NC-17 certificate is significantly smaller than those of the other certificates.

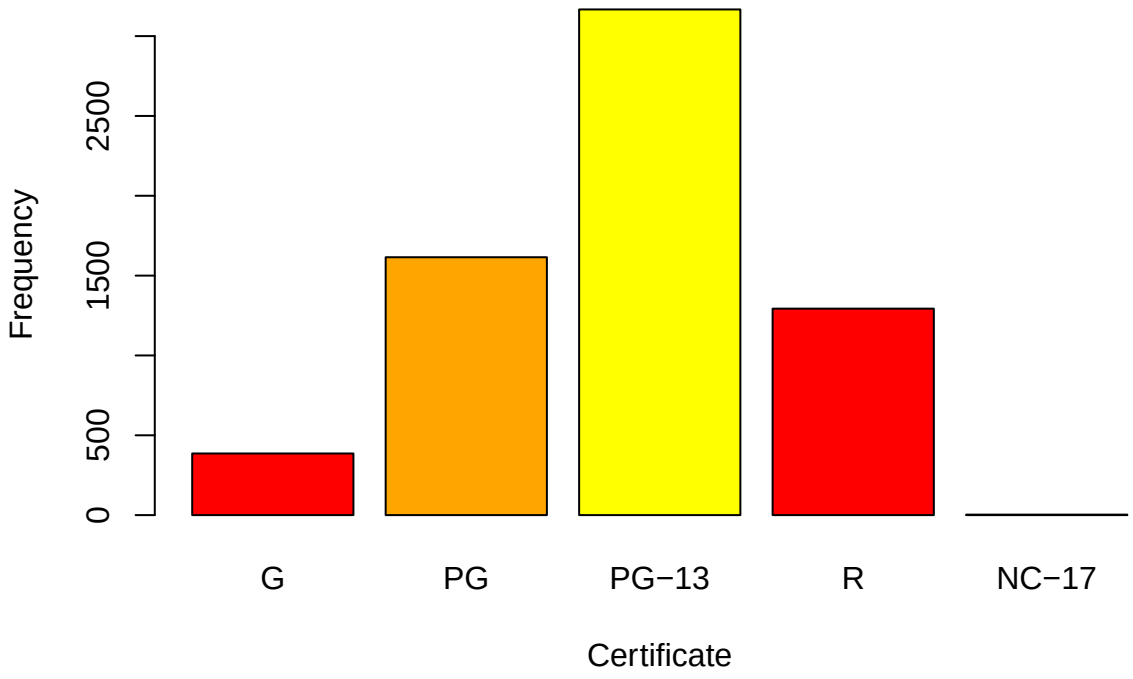
7. A move that grosses (earns) at least 100 million dollars is a *Blockbuster*. Less than 100 million but at least 10 million is a *Major Picture*, less than 10 million but more than 1 million is a *Disaster*, less than 1 million but at least 100 thousand is an *Indy Film*, and under 100 thousand is a *Passion Project*. Create a **new factor vector** in the data frame called gross.categories that classifies each movie as the appropriate type. Then print the first 10 rows of the updated data frame.

```
breaks.for.movie.gross = c(0, 100000, 1000000, 10000000, 100000000, 1000000000)

clean.movie.data$gross.categories = cut(clean.movie.data$gross, breaks = breaks.for.movie.gross, right =

levels(clean.movie.data$gross.categories) = c("Passion Project", "Indy Film", "Disaster", "Major Pictur

head(clean.movie.data$gross.categories, 10)
```

```
##  [1] Disaster       Disaster       Indy Film      Major Picture
##  [5] Major Picture  Major Picture  Passion Project Major Picture
##  [9] Indy Film      Disaster
## Levels: Passion Project Indy Film Disaster Major Picture Blockbuster
```

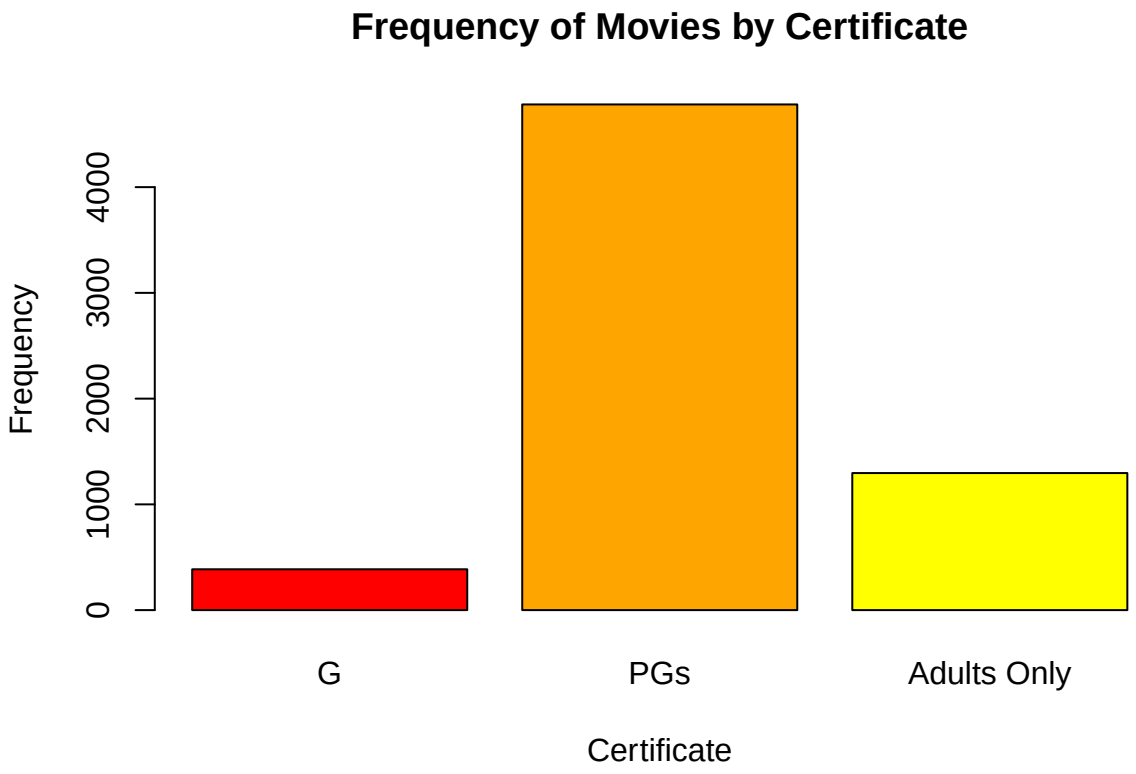
8. Make a bar graph of the variable `certificate`.

```
barplot(table(clean.movie.data$certificate),
        main = "Frequency of Movies by Certificate",
        xlab = "Certificate",
        ylab = "Frequency",
        col = c("red", "orange", "yellow"))
```



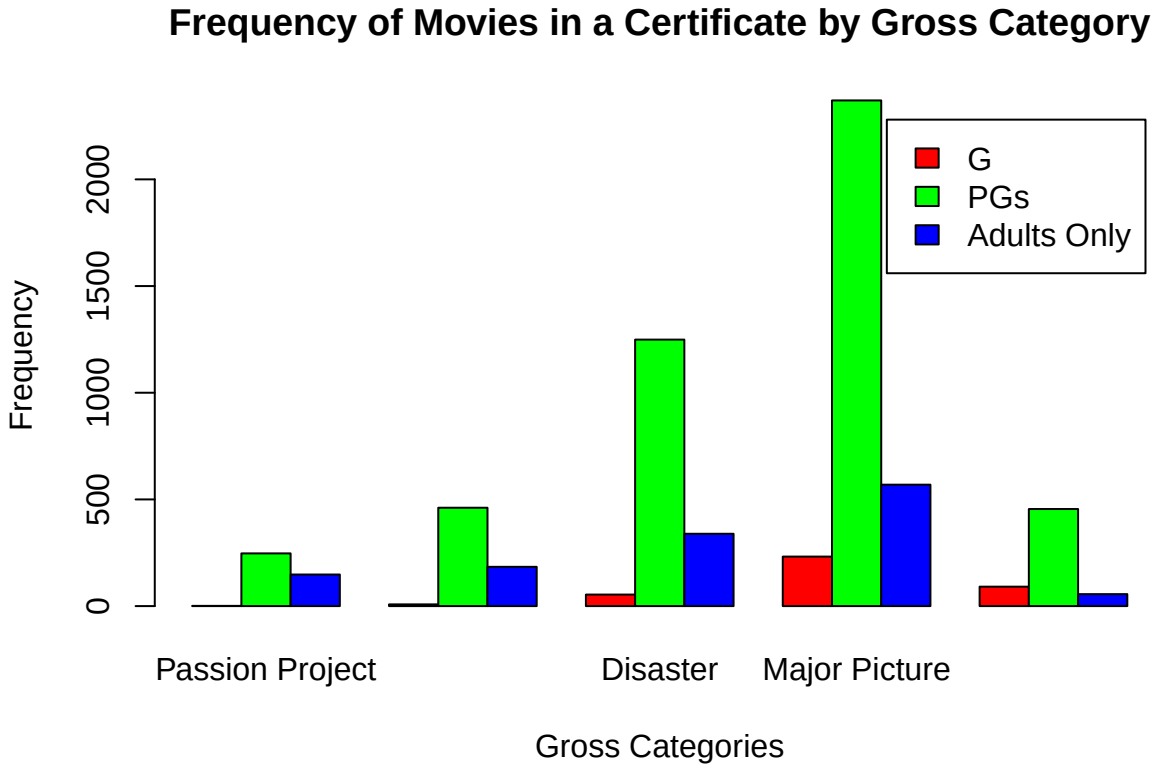
9. I don't understand the difference between PG and PG-13 or between R and NC-17. Combine PG and PG-13 and call the group *PGs*. Also combine R and NC-17 and call the group *Adults Only*. Then make a bar graph of the new data.

```
levels(clean.movie.data$certificate) = c("G", "PGs", "PGs", "Adults Only", "Adults Only")

barplot(table(clean.movie.data$certificate),
        main = "Frequency of Movies by Certificate",
        xlab = "Certificate",
        ylab = "Frequency",
        col = c("red", "orange", "yellow"))
```

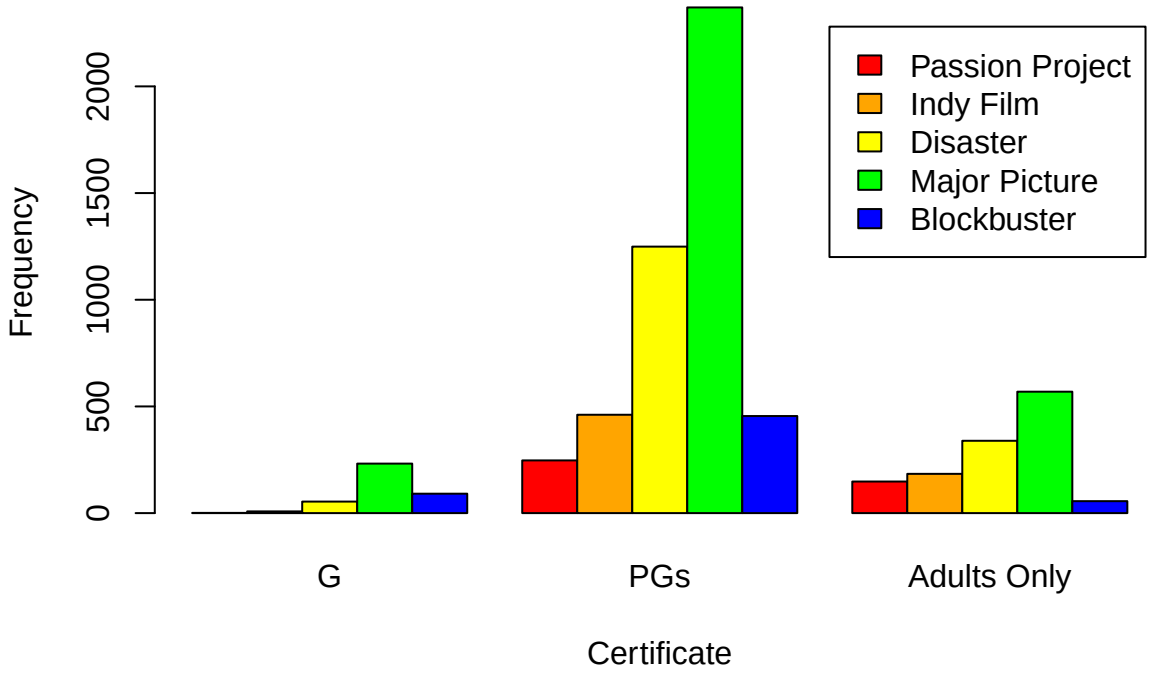


10. Make at least 3 and no more than 5 *different* bar charts that examine the variables `certificate` and `gross.factor`. At least one graph should be a proportion graph. If you are tempted to make more than 5, pick the ones that tell the best, clearest, most useful and/or interesting story.

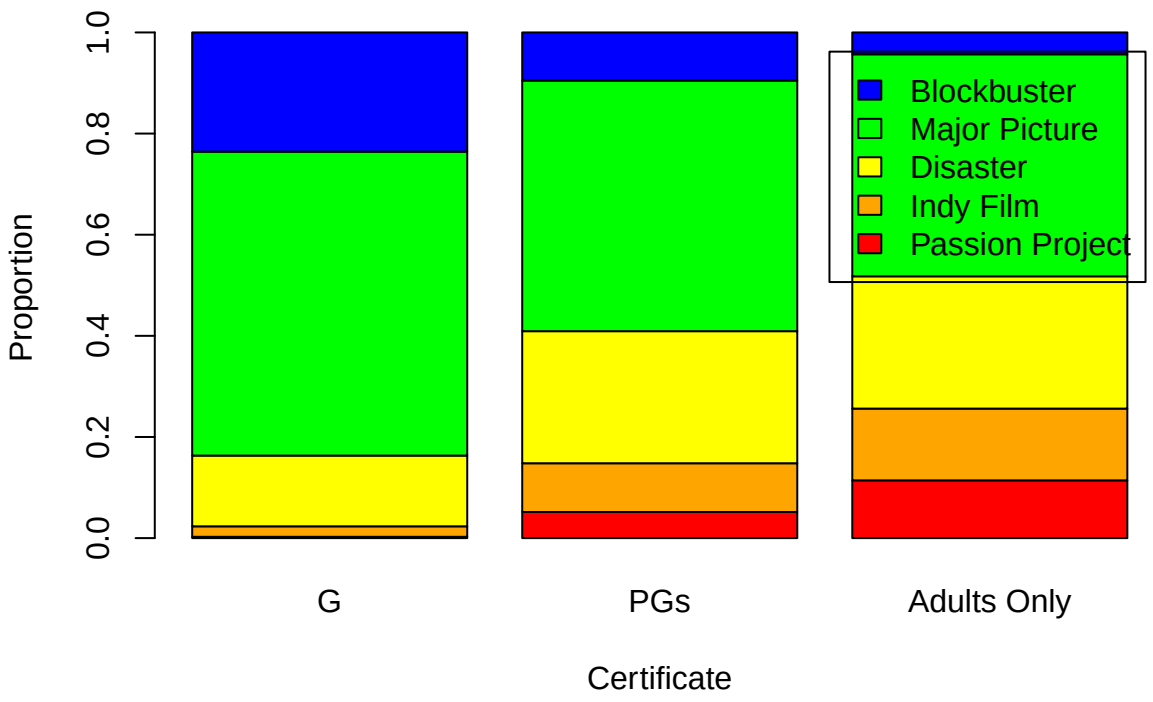
```
barplot(table(clean.movie.data$certificate, clean.movie.data$gross.categories),
        main = "Frequency of Movies in a Certificate by Gross Category",
        ylab = "Frequency",
        xlab = "Gross Categories",
        beside = TRUE,
        col = c("red", "green", "blue"),
        legend.text = TRUE)
```




```
barplot(table(clean.movie.data$gross.categories, clean.movie.data$certificate),
        main = "Frequency of Movies in a Gross Category by Certificate",
        ylab = "Frequency",
        xlab = "Certificate",
        beside = TRUE,
        col = c("red", "orange", "yellow", "green", "blue"),
        legend.text = TRUE)
```




```
barplot(prop.table(table(clean.movie.data$gross.categories, clean.movie.data$certificate), 2),
        main = "Proportion of Movies in a Gross Category by Certificate",
        ylab = "Proportion",
        xlab = "Certificate",
        col = c("red", "orange", "yellow", "green", "blue"),
        legend.text = TRUE)
```



11. Describing Graphs

a. For each bar graph that you made in problem 10, write a one sentence observation. You may write these sentences below each graph, or write them here in the same order as you made the graphs so it is clear which graph each observation corresponds to.

1. The overwhelming majority of major picture movies are PG-rated.
2. A majority of PG-rated movies are major pictures.
3. The smallest proportion of Adults Only movies are blockbusters.

b. Choose what you believe to be the most useful/helpful graph that you made in problem 10. Explain why it is useful (especially as compared to other graphs you made), and write a few sentences describing what the graph tells us - what "story" does it tell? If you think two graphs are equally useful/helpful, just choose one to describe.

I think the second graph is the most useful. It shows how movies in different gross categories compare to each other within a certificate as well as their frequencies, which is something the proportion barplot doesn't offer. This same ability to compare within a certificate is also what makes it better than the first graph. This second graph shows that a lot of movies are PG-rated, and within the PG rating, most are major pictures.

c. Choose the least useful/helpful graph that you made in problem 10. Explain why it is not useful.

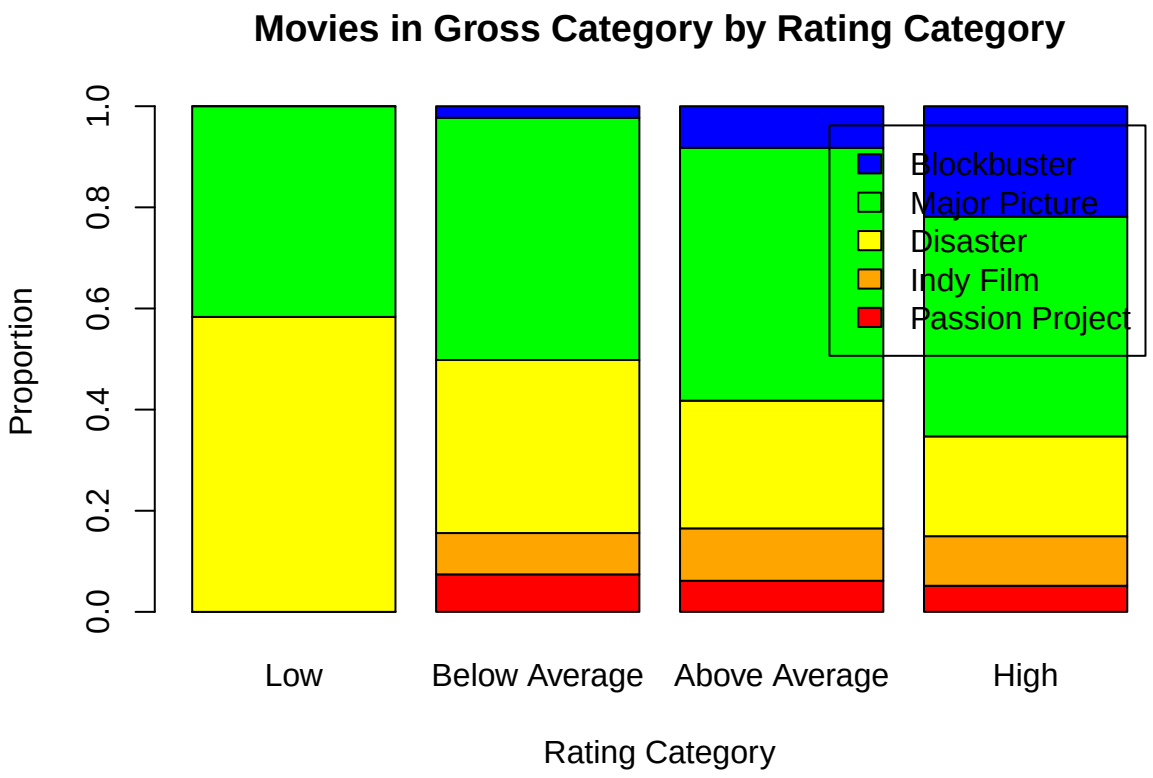
The least useful/helpful graph is the proportion graph because you can't determine frequency – only proportion – and it's easier for me to compare elements when they are side by side, not stacked.

12. Time to be creative... formulate a question about this data that involves a variable in the dataset that we did not work with on this quiz. Your question may also involve variables we have worked with if you'd like. Then answer your question with a suitable graph or graph(s) and a few sentences. If you also need to perform non-graph calculations in order to answer your question, go ahead, but your answer should include a graph. *Note: this does not need to be a complicated, in-depth exploration to earn full points. A fairly basic question that is answered succinctly and correctly with graphical evidence and proper vocabulary will earn full credit.*

Do higher ratings correlate with higher gross?

Yes, the higher the rating, the more likely it is to be a blockbuster, with the likelihood of it being a major picture remaining relatively the same. However, low ratings don't necessarily correlate with low gross. In fact, it appears ~40% of low rated films are actually major pictures. It's interesting to note that there are no indy films or passion projects in the low rating category, suggesting that these low rated films are produced by big companies who got it seriously wrong.

```
breaks.for.ratings = c(0, 2.5, 5, 7.5, 10)
clean.movie.data$rating.categories = cut(clean.movie.data$rating, breaks = breaks.for.ratings, right = F
levels(clean.movie.data$rating.categories) = c("Low", "Below Average", "Above Average", "High")
barplot(prop.table(table(clean.movie.data$gross.categories, clean.movie.data$rating.categories), 2),
        main = "Movies in Gross Category by Rating Category",
        col = c("red", "orange", "yellow", "green", "blue"),
        xlab = "Rating Category",
        ylab = "Proportion",
        legend.text = TRUE)
```



13. **Bonus:** What film is the highest-grossing film in this data set? What film is the highest-rated film in this data set? *In this data set*, which is a more notable/extraordinary achievement, the amount of money that the highest grossing film earned or the rating that the highest-rated film earned? Explain and show all of the work you did to answer these questions. Bonus Points will only be given if you correctly answer all three of these questions, and show work.

Highest-grossing film = Star Wars: El despertar de la fuerza (Star Wars: The Force Awakens) Highest-rated film = Cadena perpetua (The Shawshank Redemption)

The amount of money that the highest grossing film earned is more notable/extraordinary than the rating that the highes-rated film earned because it has a higher z-score, or in other words, it is more standard deviations away from the mean. (~14.9 vs ~2.9)

```r
sorted.movie.gross = sort(clean.movie.data$gross, decreasing = TRUE)
clean.movie.data$title[clean.movie.data$gross == sorted.movie.gross[1]]
```

```
## [1] "Star Wars: El despertar de la fuerza"
```

```r
sorted.movie.ratings = sort(clean.movie.data$rating, decreasing = TRUE)
clean.movie.data$title[clean.movie.data$rating == sorted.movie.ratings[1]]
```

```
## [1] "Cadena perpetua"
```

```r
(sorted.movie.gross[1] - mean(sorted.movie.gross)) / sd(sorted.movie.gross)
```

```
## [1] 14.97122
```

```r
(sorted.movie.ratings[1] - mean(sorted.movie.ratings)) / sd(sorted.movie.ratings)
```

```
## [1] 2.997389
```