Lab 3

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong, Khushi Shah

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore the dataset `netflix` that you will upload from my `GitHub`. Let's first load the packages we will need to complete this lab:

```r
# Load the package
library(tidyverse)
```

Then let's import our dataset using `read_csv()`:

```r
# Upload data from GitHub
netflix <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//netflix.csv")
```

This dataset contains information about 532 Netflix movies: title, genre, year of release, running time (in minutes), IMDB score, and language.

```r
# Take a quick look with tail() which shows the last 6 observations
tail(netflix)
```

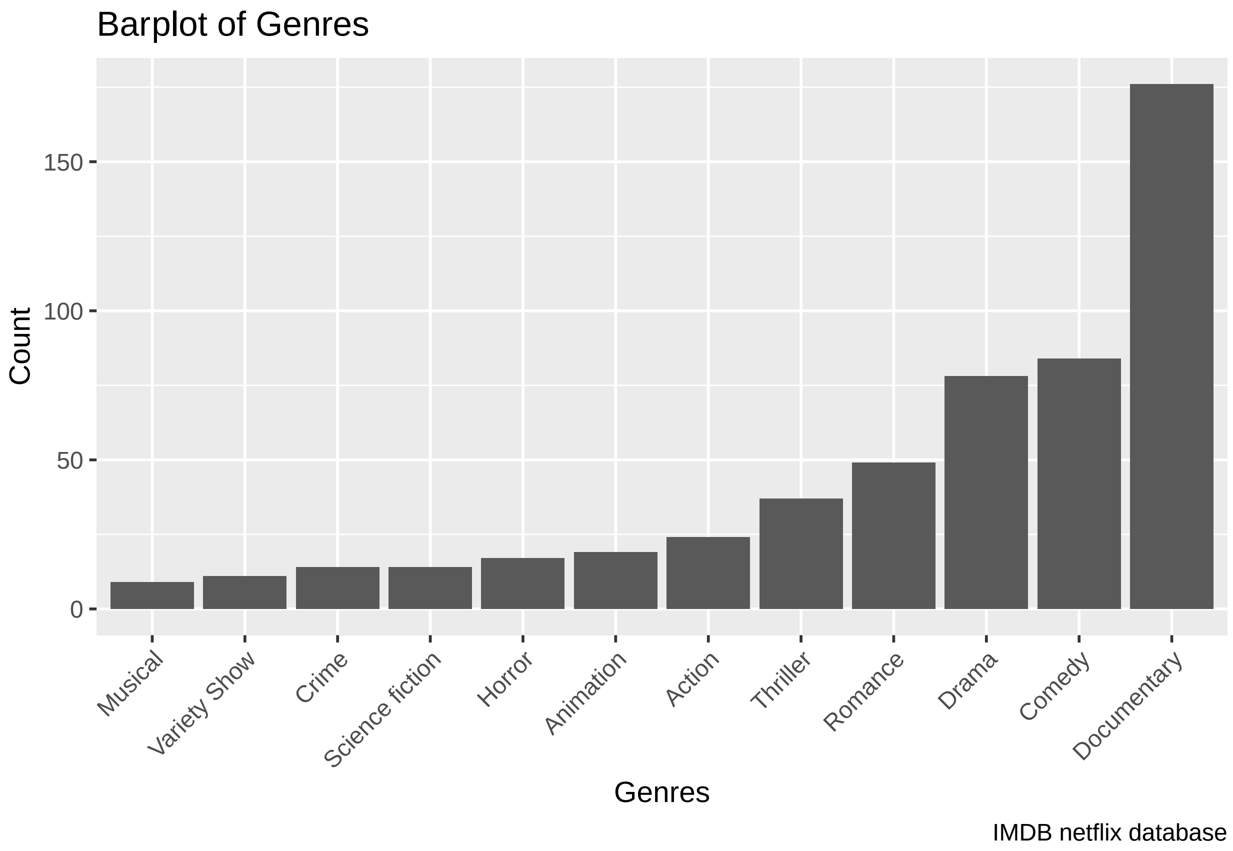
```
## # A tibble: 6 x 6
##   Title                                 Genre        Year Runtime  IMDB Language
##   <chr>                                 <chr>       <dbl>   <dbl> <dbl> <chr>
## 1 "To All the Boys: Always and Forever" Romance      2021     109   6.3 English
## 2 "Tony Parker: The Final Shot"         Documentary  2021      98   6.8 French
## 3 "Tribhanga \x96 Tedhi Medhi Crazy"    Drama        2021      95   6.1 Hindi
## 4 "What Would Sophia Loren Do?"         Documentary  2021      32   6.6 English
## 5 "Why Did You Kill Me?"                Documentary  2021      83   5.6 English
## 6 "Yes Day"                             Comedy       2021      86   5.7 English
```

The goal of the lab is to investigate how some characteristics of a movie affects its IMDB score, focusing on some genres.

Question 1: (3 pts)

Explore the `Genre` variable with a `ggplot` visualization. *Make sure to add labels and that it is readable.* What is the most common genre?

```r
# This code creates the bar chart for Genre and reorders it by greatest to least
ggplot(data = netflix) +
    geom_bar(aes(x = reorder(Genre, table(Genre)[Genre]))) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(x = 'Genres',
         y = 'Count',
         title = 'Barplot of Genres',
         caption = 'IMDB netflix database')
```


Barplot of Genres

IMDB netflix database

The most common movie is the Documentary Genre.

Question 2: (4 pts)

You will focus on a subset of genres: each group member selects a different genre. Index/filter the `netflix` dataset to only keep the genres you all selected and save it as a new dataset using the initials of all team members. You will work with this new dataset for the next two questions.

```r
# This code subsets the netflix data frame to our conditions
czgbad <- netflix[netflix$Genre == 'Drama' | netflix$Genre == 'Documentary' | netflix$Genre == 'Comedy'
```

How many movies are in your new dataset?

```r
# structure of the subsetted data frame
str(czgbad)
```

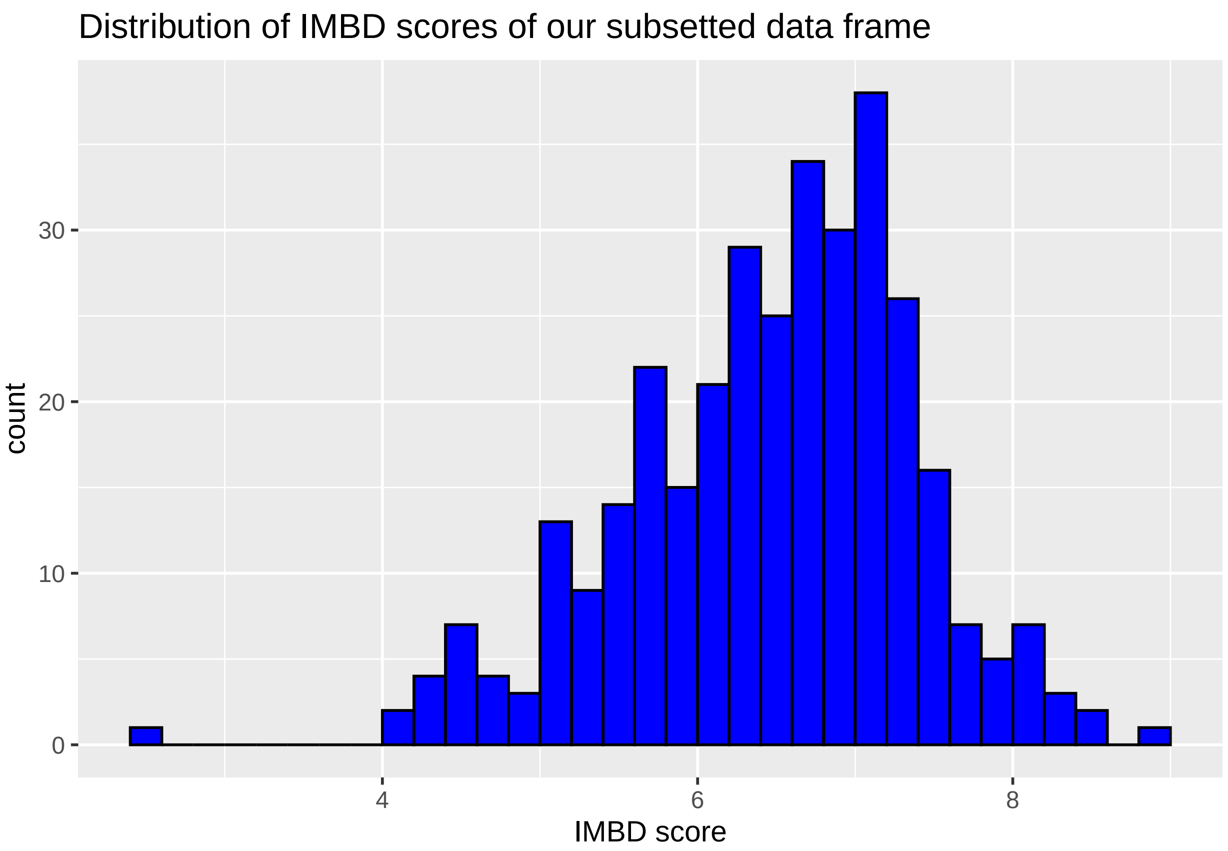
```
## tibble [338 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Title   : chr [1:338] "My Own Man" "Beasts of No Nation" "Hot Girls Wanted" "Keith Richards: Unde
##  $ Genre   : chr [1:338] "Documentary" "Drama" "Documentary" "Documentary" ...
##  $ Year    : num [1:338] 2014 2015 2015 2015 2015 ...
##  $ Runtime : num [1:338] 81 136 84 81 83 119 80 84 91 100 ...
##  $ IMDB    : num [1:338] 6.4 7.7 6.1 7.1 7.3 4.8 7.4 7.6 8.4 8.2 ...
##  $ Language: chr [1:338] "English" "English" "English" "English" ...
```

There are 338 movies in our new dataset.

Question 3: (5 pts)

Using your dataset, explore the distribution of the `IMDB` variable with a `ggplot` histogram and calculate some descriptive statistics. *Make sure to add basic labels and adjust the bins so that we can identify the intervals. You can also add a color to make it prettier!* Describe the shape, center, and spread of the distribution, including appropriate statistics.

```r
# This code creates the histogram of the IMDB of our subsetted netflix data frame
ggplot(data = czgbad) +
    geom_histogram(aes(x = IMDB), color = 'black', fill = 'blue', binwidth = 0.2, center = 0.1) +
    labs(title = 'Distribution of IMBD scores of our subsetted data frame',
         x = 'IMBD score',
         y = 'count')
```



```
summary(czgbad$IMDB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.500   5.900   6.650   6.519   7.100   9.000
```

```
sd(czgbad$IMDB)
```

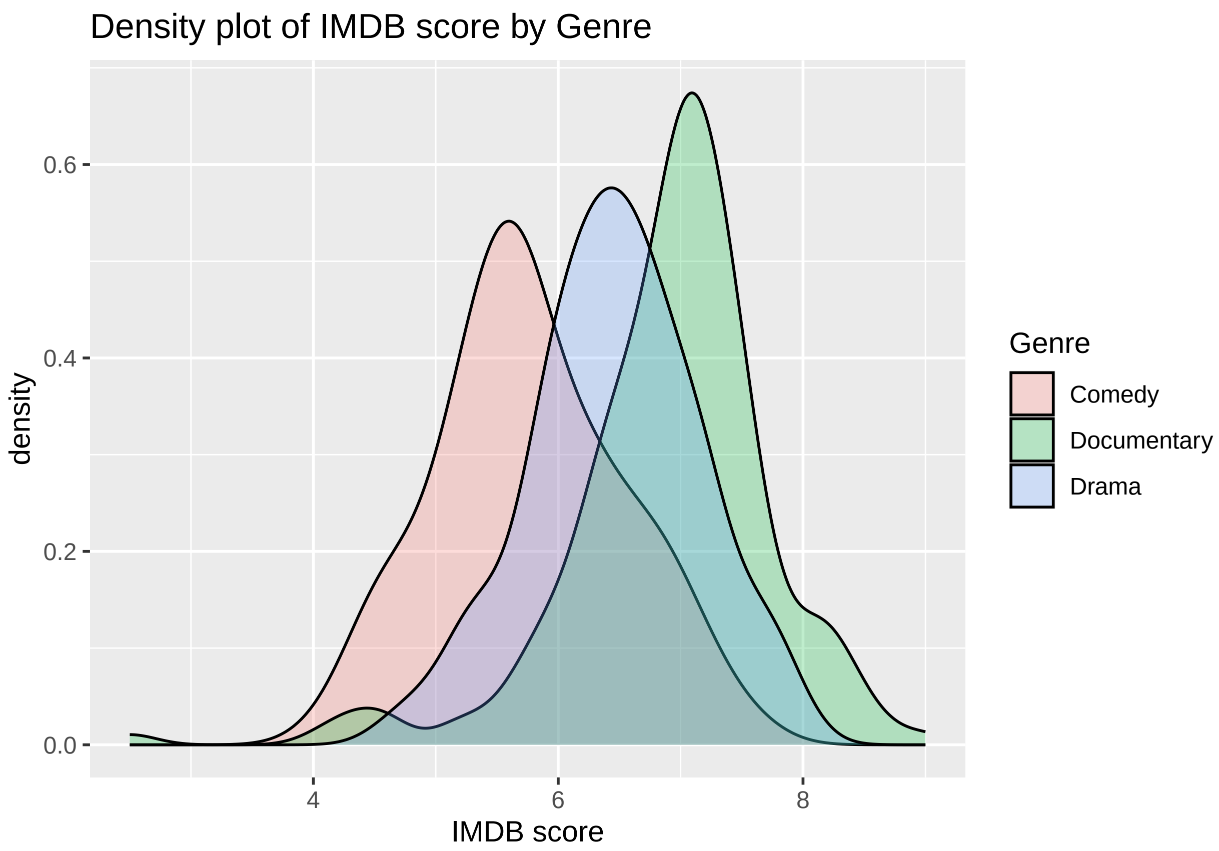
```
## [1] 0.9171929
```

The shape of the distribution is normal with the exception of one outlier. The distribution is centered around 6.519 and the standard deviation is 0.9171929.

Question 4: (5 pts)

Using your dataset, investigate the relationship between `IMDB` and `Genre` with a `ggplot` visualization. *Make sure to add labels. You can also add a color to make it prettier!* Does there seem to be a difference in IMDB score across these genres? Justify your answer by referring to your visualization.

```
# This code creates the density plot of the IMDB score by genre
ggplot(data = czgbad) +
    geom_density(aes(x = IMDB, fill = Genre), alpha = 0.25) +
    labs(title = 'Density plot of IMDB score by Genre',
         x = 'IMDB score')
```



The IMDB score for Documentaries is higher on average than Comedies and Dramas. Comedies have the lowest IMDB score on average.

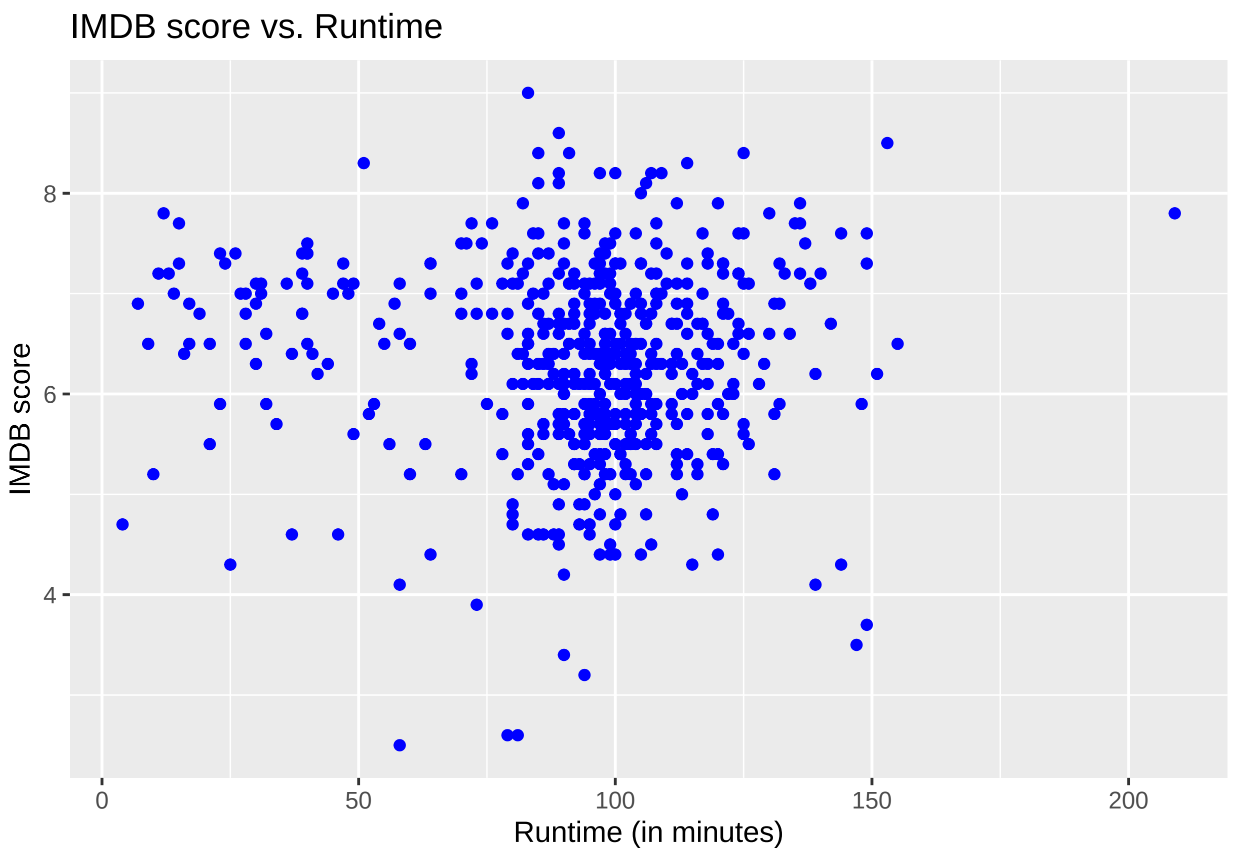
Question 5: (6 pts)

Using the entire `netflix` dataset, choose another variable that you think might affect the IMDB score (could be a variable you selected in the intro lab). Write a question you would like to investigate using these variables:

Does a shoter runime in minutes of a movie lead to a lower IMDB score?

Create an appropriate visualization to answer your research question using a `ggplot` (include a title and labels).

```r
# This code creates a scatter plot of runtime and IMDB score
ggplot(data = netflix) +
    geom_point(aes(x = Runtime, y = IMDB), color = 'blue') +
    labs(title = 'IMDB score vs. Runtime',
         x = 'Runtime (in minutes)',
         y = 'IMDB score')
```



How would you interpret what you see in this visualization?

There is no distinct relationship between run time and IMDB score so there is a very low correlation between the 2 numeric variables so runtime is not a good indicator of IMDB score.

Question 6: (1 pt)

After investigating how some variables relate to IMDB score, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

When comparing the relationship between genres and IMDB score, we found Comedy to have the lowest average IMDB score. This was expected since Comedies are not as insightful as other genres. In contrast, when we analyzed run time and IMDB score, we found no relationship between the two variables. This was surprising to us as we expected a shorter runtime to correlate with a lower IMDB score

---

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.