Analysis Plan

Cady Stringer and David Aarhus

1. Does rating (PG-13, R, G, etc.) have an impact on IMDB score?
    a. We can answer this with Lasso and Ridge **dimensionality reduction**
    b. We can see how much rating impacts the accuracy of our predictions of IMDB score first using Ridge, then check whether rating is one of the more important variables or instead if it's one of the variables Lasso would set to 0 and eliminate.
    c. Data visualizations
        i. We could start by making a side by side bar plot of the average IMDB score for each rating and compare them. This will help us identify which ratings are getting the higher scores from viewers.
        ii. Also, we can seperate the ratings up into "age appropriate" groups by labeling them a number (1-4 or 1-5) this would allow us to look at the average IMDB score across different age ratings by using a barplot as well.
2. Is Netflix increasingly focusing on TV rather than movies in recent years?
    a. We can use **data visualization and summary** to answer this question.
        i. We can create a bar plot colored by the variable "type" to see what percent of the total each year is movies versus TV shows, and see if there is an increase or decrease in either one over time.
        ii. We can create a line chart of the number of movies and number of TV shows over time to see if and when there are changes in which occurs the most frequently in the data set.
3. How well can we predict the century a film was made using both Netflix and IMDB data?
    a. We can use a **logistic regression model** to predict the century a film was made starting with all variables to give our model as much prediction power and possible, then use Lasso as a **dimensionality reduction** to select the most important variables and create a simpler model.
    b. We will use K Fold cross validation with 5 folds because we have a data set of 2600 data points, so LOOCV would be too computationally expensive and a test/training split wouldn't be thorough enough for the amount of data we have.
    c. We will standardize all continuous variables to make our model more interpretable and ensure that variables measured on different scales don't interfere with our model creation.
    d. Data visualizations
        i. First we'd create a confusion matrix to see how accurate our model was at predicting the century a film was made in.
        ii. Secondly, we could create a new column in our dataset of (1's and 0's) that identifies which century the movie was created in. After that we can create simple bar plot and fill it with colors that correspond with different variables (ratings, country, duration, etc)
        iii. We could make a scatterplot (duration vs IMDB score) with a "fill = factor(century)" which would help us see if there was a difference in

movies produced in the 20th or 21st century based on their duration, and IMDB score. Possibly no correlation at all.

4. How well can we predict the IMDB score of a film based on the genre, director, release year, rating, country, and duration?
    a. We can use a **linear regression** using all of these variables to predict IMDB score. Although this question specifies the variables we thought would be most valuable, we'll use as many variables as possible initially to give our model as much prediction power as possible, then use Lasso **dimensionality reduction** to eliminate the less important variables and create a simpler model.
    b. We will use K Fold cross validation with 5 folds because we have a data set of 2600 data points, so LOOCV would be too computationally expensive and a test/training split wouldn't be thorough enough for the amount of data we have.
    c. We will standardize all continuous variables to make our model and coefficients more interpretable.
    d. Data visualizations
        i. First we could create a new data column that labels each movie (0 or 1) identifying which movies were filmed in multiple countries or just one. Then create another column that labels each movie with a number that represents how high their IMDB score is (1 = lower, 2 = low, 3 = medium, 4 = high). This then would allow us to create a ggplot bar graph of the 4 rating totals and then color each portion of the bar that were filmed in different countries or just one.
        ii. Also would create a simple scatterplot with IMDB Score and duration to see if there is an obvious positive correlation between the two variables.
5. What content is similar?
    a. We can answer this question using an Agglomerative Hierarchical **clustering model**. We will use all continuous variables for this model because clustering works best with all numeric variables where you can measure the distance between them.
    b. We can test the silhouette scores of different numbers of clusters to pick the amount of clusters n that creates the most cohesion and separation.
    c. We will standardize all continuous variables because there are a variety of different metrics in the data and we don't want that to impact the way the model treats different values, or give any one category more weight.
    d. Data visualizations
        i. We can create a barplot of clusters and color the bars by the variable "type" to see how movies and TV shows are distributed among the clusters, and see whether that was a big factor the model used to separate.
        ii. We can create ggplots of different variables (like year, duration, IMDB score, etc.) and color the points by cluster to see how different clusters are separated.
6. What release years had the highest average IMDB rating?
    a. We can answer this using **data visualization and summary.**

i. We can group by the release year and create a ggplot of the average IMDB rating for each year, and this will show us both the year with the highest average and how IMDB ratings have changed over time.
ii. We can group by release year and create a bar plot of the average rating by year to see which year has the tallest bar and thus highest average rating.