Info 573: Data Science I: Theoretical Foundations
Lab 2: EDA and exploring probability distributions
Instructor: Ben Althouse; Contact: bma85@uw.edu

In this lab you will practice basic EDA on movie ratings and users and examine some common probability distributions and look at the sums of random variables. Either RStudio or the console may be used for this lab.

You may work with a partner on this lab, however, you will be asked to submit a copy of your analysis code to Canvas at the end of class - each individual must submit their own version of their code – *please be sure to put your name in the code!* Keep track of all the commands you run using a text editor or R script.

You should comment your code as you run through this exercise. You can do this in R using the # character. Please answer the questions posed in the exercise/lab by adding comments to your R script.

Dataset descriptions: Table below gives the variables in the two datasets. Data source: https://grouplens.org/datasets/movielens/

"ratings" dataset:

userId	ID of the user doing the rating
movieId	ID of the movie, links to the other dataset
rating	Numerical rating of the movie
year	Year the movie was made
genre	Genre of the movie

"movies" dataset:

movieId	ID of the movie, links to the other dataset
title	Title of the movie
genres	Genre of the movie
year	Year the movie was made

PART 1

1. Load the data.

a) what are the dimensions of the data? Ie, how many ratings? How many raters?

b) what are the mean, median, and standard deviation of the ratings?

c) plot a histogram of the ratings. What patterns do you see? Do people seem to prefer round numbers?

2. Link the two datasets using movieId. The function `match()` is useful:

```
ix <- match(ratings$movieId, movies$movieId)
```

How can you verify that you have the correct link? Hint read the help of what `match()` is doing and then use `head()` to compare the datasets.

3. Exploring Relationships I: Visually explore relationships between your variables.

a) Plot the relationship between ratings and year. Try a scatter plot and a box plot. What do you see?

b) It is messy, try coarse-graining the data using `cut()` on the years.

4. Exploring Relationships II:

a) Do the ratings vary by genre? Create a 'comedy' column and draw a box plot of ratings for comedy versus others:

```
ratings$comedy <- rep(F, nrow=ratings)

ratings$comedy[grep("comedy",ratings$genre, ignore.case = T)] <- T
```

b) Run a t-test to see if the differences in ratings for comedy versus non-comedy (note, we haven't learned yet what this is, but we will soon!).

```
t.test(ratings$rating, ratings$comedy)
```

5. Extra credit: what's the "best" movie in the dataset? Come up with a metric to assess quality and find the top 10 movies.

PART 2

1) Examine several distributions. Plot a histogram of 1000 samples from an exponential and uniform distributions with default mean and variance, and a Poisson distribution with lambda = 1/2 (look up `rexp()`, `runif()`, and `rpois()`).

2) Look at the distributions of sums of these samples. Here's some code to look at sums of the exponential variables:

```
N <- 1000  # number of exponential draws
n.samp <- 30 # number of sums to take

M <- matrix(NA, nrow=N, ncol=n.samp) # create an empty
matrix to fill with samples
for(j in 1:n.samp) M[,j] <- rexp(N) #generate the samples

hist(rowSums(M), freq = F) # plot a histogram of the sums
across rows of our matrix M
```

What distribution does the histogram of sums look like? Add this curve for a helpful comparison:

curve(dnorm(x, mean(rowSums(M)), sd(rowSums(M))), min(rowSums(M)), max(rowSums(M)), add=T, col="red", lwd=2)

3) Repeat this for the uniform and Poisson distributions. (Hint: the code can be copied and one single function can be replaces – which is it?) What pattern has emerged?

4) Further explorations. What happens when you change the number of samples from the distribution? What happens when you change the number of sums taken?