ITEC 620

Fall 2022

**Assignment 1**

Due October 31, 5:15 PM

Please submit your assignment via Canvas, **as a single Word file**.

**For all questions that involve generating R output, you must include both the output itself and the command or sequence of commands used to generate it to receive credit. Include ONLY the commands used to generate that output; do not include entire scripts.**

For this assignment, you will need the “Damon” csv file. It contains data on movies from 2001-2017 featuring Matt Damon, an American actor. For each movie, the title, the average rating on IMDb (0-10, higher numbers are better), and the total amount of money the movie made at the box office worldwide ($M) are provided.

You will first need to import the dataset into R. Save the Damon.csv file in your working directory, and then use the following R command to store it in an object called Damon:

*Damon <- read.csv("Damon.csv")*

All results (except 2b) should be obtained using R. Several of them are easy to figure out manually, but the point is to practice with R!

**1. R and Descriptive Statistics**

**a**. Show the row containing the 12th movie in the dataset. (Note that headers do not count as a row in R; row 1 contains the 1st movie in the dataset.)

**b**. What is the average (mean) value of BoxOffice?

**c**. What is the lowest rating in the dataset?

**d**. Using only a single R command, display the rows of the dataset containing movies with a rating of 8 or higher.

**2. Clustering**

**a**. Normalize the dataset, and use *k*-means clustering to split the movies into three clusters. What are the resulting three centroids, in original units?

You will need to remove the title column before doing this. The command: “Damon <- Damon[,-1]” will remove the first column.

You will also need to run the code that defines the unscale function; it is not necessary to include that block of code with your answer.

**b**. What do the centroids from part a tell us about Matt Damon movies? (Please limit your answer to 100 words or less.)

By looking at the centroids for cluster 1, we can conclude that on average, all the observations have 5.9 ratings and made $104.4M at the box office worldwide. Similarly, for cluster 2, the average ratings for the observations within this cluster are 6.9 and they made $93.7M at the box office worldwide, and for cluster 3, there is an average of 7.7 ratings and $366.7M at the box office worldwide among the observations.

**c**. How many movies are in each of the three clusters from part a?

Cluster 1 has 7 movies

Cluster 2 has 15 movies

Cluster 3 has 9 movies