CPTR420 Test 1

October 5, 2023 12:30 – 1:45

For this exercise you may use a reference sheet, previous assignments and/or the R and RStudio cheat sheets available from the RStudio help menu.

You must not use other online help. You must not seek help from or give help to your colleagues.

You should download and extract the exam folder The data file you need for the exercise is in the folder.

After you have completed the assignment, please upload the r script learningHub.

The file Test1Movies.csv has data about several Walt Disney movies. The data was extracted from a more complete dataset found here. It is a list of films produced by and released under the Walt Disney Pictures banner, known as that since 1983. Films released before that were under the former name of the parent company, Walt Disney Productions

The data is stored in a comma-delimited (csv) text file. Missing data is indicated by empty cells.

Variable names are somewhat self-explanatory. The more complex ones are described below:

Title

Production_Company

Running_Time: Length of movie i.e. viewing time

Country: Principal country where the movie was released

Language_Abbr : International code for language used in the movie.

Budget: amount (USD) spent to produce the movie

Box_Office: earnings (USD) from the movie

Release_Date: Date the movie was released

Directed_By : Movie directors (may be more than one)

Produced_By: Movie Producers (may be more than one)

Music_By : Music writers (may be more than one)

Distributed_By: Distributors of the movie

Languages: Other languages of production for the movie

Countries : Additional countries where movie was released

The data will eventually be used to look at trends and perhaps make predictions about the movie industry. First it must be cleaned and prepared so that current trends could be observed.

In order to do that, you must complete the following exercises:

Write your code in an R script named **test2YourInitials**. The script should be well documented with your name, the date and the file name in an opening comment block. Each significant command or block of commands should be appropriately commented. You may use the numbered lines below to comment the code.

Your initial comment block must also include the following statement to indicate that you used only your notes, R cheat sheets, R help (available using ? or help) to complete the exercise

I ____________ declare that no unauthorized documents or person was used for help to complete this exercise.

Exercises

Write appropriate r statement(s) for each **5 points each**

SET UP THE ENVIRONMENT

# 1. Install, if necessary, and load the packages you will use to complete the exercise.

#2. Read the data from the comma-delimited file test1Movies.csv in your folder into an appropriately named dataframe. Missing data should be coded as NA. Note that missing data in the csv file is represented by blank cells.

Read the data from the Language_Codes file into an appropriately named dataframe

EXAMINE AND PREPARE THE DATA

# 3. What are the data types of the variables?

#4. Examine the data, show a sample, the first or last 5 observations.

#5. Convert the following to numbers if necessary:
# a. Budget
# b. Box_Office

# 6. Run basic stats to get an idea of the distribution of the data i.e. min, max, mean, median, quartiles, etc. and number of NAs for the variables

# 7. The variable Release_Date is shown as mm/dd/yy, split this into three separate columns: Month, Day, Year. Discard the original Release_date column.

#8: The Languages and Countries columns are not necessary for our analysis. Remove it.

#9 Remove all the rows where the language-codes.csv is not known.

#10. The Running_Time is currently type chr because it has the units attached e.g. 80 minutes. Write the code to create a column called Movie_Length that will have just the number of minutes and will be an integer. The new Movie_Length column should replace Running_Time and must be integer type.

#11. The Produced_by column has the names of the movie producers. Several movies have more than one producer. Split his into three new columns: Producer_1, Producer_2, Producer_3 that will have the name of the first three producers. Note that some movies may have fewer than 3. For those that have more than 3, just ignore the others. Remove all extra spaces from the data in these columns.

#12. Add a column called Language to the data. This column will have the name of the language represented by the Language_Abbr column. The codes csv file has corresponding language names for the codes.

ANALYSE:

#13. Which movie had the largest income (i.e. box_office)

#14. Which did Jordan Kerner produce? Show the Title, release date, and the producer.

#15 Which were the 5 most successful movies based on profit? Show the Title, Budget and Box Office values. Note that the budget information is missing for some movies.

#16 What is the number of movies, average budget and average box office by language. Note that the data might be missing for some movies.

#17. What is the most popular day for movie release? Show the day number, number of movies released, total budget, total box office.

Save your script and upload it to LearningHub