Homework 4

Data Cleaning

MATH 210-010 ◇ Fall 2024

November 12, 2024

Due: Thursday, November 21, 2024

Instructions: To complete a problem set, you must submit a zip file labeled `Yourlastname_HW#` to Dropbox no later than 11:59 PM on the due date above. For example, if I were to complete this assignment, my folder would be named `Emerick_HW4`. In this folder, a `py` file is to be submitted for each problem such that when the `py` file is executed, the output (as presented in Python) is the solution to the problem. Each `py` file must be saved as `Yourlastname_HW#_No#.py`. For example, if I were submitting the answer to Question Number 1 on Homework 4, the `py` file for that problem would be saved as `Emerick_HW4_No1.py`. Each `py` file should be well commented and be free of extraneous lines and commands. Also, each `py` file must output only what the problem asked to be outputted. Failure to abide by these simple homework submission guidelines may result in a deduction of points at my discretion.

Name:

Score:

For each problem below submit a separate `py` file with an initial comment that describes the objective of the `py` file. Always remember to begin your `py` file by importing appropriate libraries.

1.] Create a file that uploads the `RottenTomatoes.csv` data set. Perform the following changes in Python and then save a cleaned `csv` file.

   a.) In the column labeled Certificate, there is a rating for each movie (e.g. PG-13, R, NR, etc). For some of the entries, there is a justification for this rating in parentheses. For any entries with a justification, remove the justification but keep the rating.

   b.) In the column labeled Runtime, every entry has "minutes." Delete the string "minutes" and preserve the numerical value for the runtime. Then, convert all entries in this column to type `float`.

   c.) In the column labeled Director, there is a list of directors. Most movies only have one director. Preserve only the first director before the first comma. Delete any entry after the first comma.

   d.) In the Studio column, convert all entries to `string` type.

   e.) In the last four columns (RT Critics Score, RT Critic Reviews, RT Audience Score, and RT Audience Reviews), make sure all entries are numerical and converted to `float` type.

2.] Create file that uploads the `iMDb.csv` data set. Perform the following changes in Python and then save a cleaned `csv` file.

   a.) Write code that deletes the columns labeled Poster, Cast, Description, Review Title, and Review.

   b.) In the Genre column, only keep the first classification before the first comma. That is, any entries that have more than one genre listed, only keep the first genre in the list.

   c.) The rating system in this file is based on India's movie rating system. Some entries such as UA, A, or UA 16+ are not rated in the United States. Search for an equivalent rating for the unknown India ratings and convert any unknown strings to known U.S. ratings. If there is no equivalent, reduce the entry to a blank entry.

   d.) In the columns for iMDb Critics Score, iMDb Metascore, iMDb Audience Reviews, and iMDb Critic Reviews, be sure to convert them all to the same `float` type.

3.] Create a new dataset with the common movies of both cleaned data frames. In this merged data frame, the columns should include Title, Year, Runtime, Genre, Director, Studio, Certificate, and the eight numerical columns pertaining to the two websites critics scores, audience scores, number of critic reviews, and number of audience reviews.