UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

STC 122 - Assignment 1 submission instructions

Instructions

- Answer the questions that follow and save all your code in a single `R` script. Name the script **Assignment1.R** where **Assignment1** is the filename and **.R** is the file extension.
- Two submissions are required for this Assignment, namely a code submission and an interpretation submission.

Submission 1: Code

- Submit your `R` script on Gradescope before **Tuesday, 15 August 2023, 23:59**.
- Multiple code submissions are allowed and your autograded results will be available shortly after each submission.
- **Ensure all variables are named correctly, as incorrectly named variables will not be awarded any marks.** (Remember variable names are case sensitive.)
- **Ensure your code does not contain any syntax errors.** If your code produces errors when run, the autograder will not be able to mark it.
- Any code commented out will be considered rough work and will not be marked.
- Once you have completed your submission, ensure the file is submitted on Gradescope with the correct file name, **Assignment1.R** where **Assignment1** is the filename and **.R** is the file extension. **The autograder will only be able to grade your submission if you use the correct filename.**
- Students must add **detailed** explanations of their working as comments in their code. This must be done for every question in the assignment. Students should also add citations for any resources they make use of in their assignment. Refer to the document `commenting_and_citing_in_your_code.pdf` on ClickUP.

Submission 2: Interpretation

- Submission 2 will be an invigilated assessment.
- Full details regarding the date, time and format of this assessment will be communicated as soon as possible on ClickUP.

Guidelines

- This assignment covers Exploratory data analysis.
- The assignment is based on Lab 1, Introduction to Data from the Openintro website. The lab is available on ClickUP.
- The data set for this assignment is available on ClickUP and is called `titles.csv`
- More details on the data set can be found here.
- Where applicable, answer the questions below by typing the appropriate **code** in the `R` script template provided on ClickUP. Some questions are theoretical and no coding is needed to answer those questions.

Questions

Question 1

Find the name of the title with the lowest IMDB score. Save your response into a variable called `Q1`

Question 2

Create a subset containing only the movies. Save your response into a variable called `Q2`

Question 3

Find the name of the **movie** with the highest TMDB popularity. Save your response into a variable called `Q3`

Question 4

Calculate the most common age certification of the movies. Save your response into a variable called `Q4`

Question 5

Create a histogram to investigate the distribution of the IMDB ratings. Use `breaks = 50`. Give your plot an appropriate main title and axis titles.

Investigate the impact of changing the argument `breaks`. Consider break values of 5, 10, 50 and 100.

Question 6

Calculate the mean runtime. Save your response into a variable called `Q6`

Question 7

Calculate the standard deviation for the runtime. Save your response into a variable called `Q7`

Question 8

Calculate the percentage of releases with a runtime within 1.4 standard deviations of the mean. Save your response into a variable called `Q8`

Question 9

Create a two-way contingency table for the age certification and the type of the release. Save your response into a variable called `Q9`

Question 10

Create a side-by-side boxplot to compare the IMDB ratings for the movies against the IMDB ratings for the TV shows. Make sure you add a legend to your plot. Give your plot an appropriate main title and axis titles.

Question 11

Create a bar chart of the age certifications. Give your plot an appropriate main title and axis titles.