Dplyr Practice questions

2022-09-05

0.1 IMDB movies

1. How many movies are there in the dataset?
2. Each movie has a unique IMDB ID. For example, in the link http://www.imdb.com/title/tt0449088/?ref_=fn_tt_tt_1, the id is 0449088. Create a column that stores this ID.
3. ID should be unique for each movie. Check if that is true. Hint: use group_by, summarise.
4. Make a list of all keywords used in the dataset. Which keywords are repeated most often? (Note: this is not a dplyr based question.)
5. Create a column that contains the first genre in the list of genres used for a given movie.
6. Create a barplot for the genres, with the following conditions:

   a. Keep the panel background blank
   b. On x-axis, keep the text at an angle of 60, set the hjust and vjust arguments to 1
   c. Make axis titles bold
   d. Remove the grid.
   e. Fill the bars with red color

7. Find the average IMDB rating for each director. Which director has the highest average IMDB rating?
8. English movies are made in multiple countries. Is there a bias in terms of the IMDB score of English movies for a given country?

0.2 Quality of life dataset

1. How many cities are there in the dataset? How many quality indices?
2. Notice that there are spaces in the column names. Having spaces in varible names is not desirable. Replace spaces with "_".
3. Arrange the dataset based on the City.
4. A ranking has been assigned to each city. Which column was used to assign that? In other words, sorting by which column gives you the same dataframe?
5. Create column containing the country of each city.
6. Find the average Quality of life index for each country. Which country is better to live in?
7. Make scatter plot of the Quality of life index (y axis) against the other indices (x-axis, one plot for each index). Use the same conditions as above. Which index can explain the quality of life best?