Assignment 1 - Exploratory Data Analysis

Instructions:

Welcome to Assignment 1! In this assignment, we will be diving into the exciting world of Exploratory Data Analysis (EDA) using Python, R, and Weka. Please read the following instructions carefully:

Part 1: Python (40 Points)

Task 1A (2 points): Display all the details about the dataset.

Task 1B (3 points): Check for null values within each column. Fill in values for more than 100 missing columns with a specific label such as "No Director" or "Country Unavailable." Drop the rows for the remaining missing columns.

Task 1C (5 points): Create a horizontal bar chart displaying the top 10 countries with the highest number of movies and TV shows.

Task 1D (4 points): Print the first row based on the longest duration time of a movie from each country. Include information such as the director, date added, release year, duration, and description of the movie.

Task 1E (4 points): Display the title, director, date added, and release date of movies where the official release date and the date added to the platform have the same year.

Task 1F (4 points): Display the director, release year, and the number of movies and TV shows directed by each director within a year. Sort the results in descending order based on the count.

Task 1G (3 points): Display the title, director, date added, and category of movies or TV shows from the Documentary/Docuseries category.

Task 1H (4 points): Display the title, date added, category, and description of Family Dramas. Use the description to identify the type of drama.

Task 1I (5 points): Plot the distribution of TV shows based on the number of seasons using a horizontal bar chart. Group the seasons into the following categories:

Less than 3 seasons
3 seasons
4 seasons
5 to less than 10 seasons
10 or more seasons

Task 1J (6 points): Display a side-by-side pie chart showing the distribution of movie and TV show ratings.

Movie Ratings:

Uncut/Not rated
Restricted
Parental guidance
General audience
Adults only

TV Show Ratings:

All Children
Older Children
Parental Presence
General audience
Mature

Part 2: R (40 Points) Perform all the tasks in R, using a different R notebook. The data and questions will be the same as those in Task 1.

Part 3: Weka (20 Points) For this task, use the "Employee_retention.csv" data file in Weka. Perform the following analyses:

Task 3A (3 points): Display a visualization for each column in the dataset.

Task 3B (3 points): Display the time spent by employees in the company vs. the occurrence of work accidents. Interpret the graph and provide your observations.

Task 3C (3 points): Display the job satisfaction level of employees vs. whether they received a promotion within the last 5 years. Interpret the graph and provide your observations.

Task 3D (3 points): Display the last evaluation score of employees vs. their job satisfaction level. Interpret the graph and provide your observations.

Task 3E (3 points): Display the average monthly working hours for employees vs. the number of projects they are working on. Interpret the graph and provide your observations.

Task 3F (5 points): Display the correlation between low salary levels and the occurrence of promotions within the last 5 years. Interpret the graph and provide your observations.

Programming Assignment Details:

Cite any resources used (books, internet) within the corresponding cell.
Do not rename the dataset files.
Include screenshots of the Weka analysis for each question in the submission folder. Name each image with the corresponding task (e.g., Task3B, Task3C).
Include the name and ID of each group member in the Jupyter notebook in the provided format.

Submission Details:

Name your submission files using the following format:
yourLastName_Last4digitsofyourID.ipynb (e.g.,
smith_1234_johnson_5678_assignment1.ipynb).
Only one team member should submit the file.

Good luck with your assignment! Explore the data, analyze, and enjoy the process of Exploratory Data Analysis using Python, R, and Weka.