Mandatory Assignment #2

Overview

For this mandatory assignment, you will work with data on Netflix tv shows and movies. You will work with a total of 3 datasets. The Netflix dataset (Netflix.csv) contains data on tv shows and movies available on Netflix. The IMDB datasets (IMDB_movies) and (IMDB_votes) contain data on movies with IMDB votes. Your goal in this assignment will be to build a clean and well-documented workflow that finds 10 top voted comedy movies on Netflix.

Submission details:

- You can use Alteryx or Python to complete this assignment.
- Submit a PDF document with the answers along with one of the following:
    - Alteryx: yxmd file
    - Python: Jupyter notebook

Tasks

1. Pull in Netflix.csv file. Perform EDA- as a minimum, provide summary statistics and 3 visualizations to show distribution of numerical and categorical variables of your choice.
2. Filter for and find the total number of movies in the Netflix dataset.
3. Add a continent column to the dataset in order to not only see the country of the movie but also the corresponding continent.

Note that to answer question 3, you will need to find a dataset online with country and continent mapping.

4. Pull in IMDB_movies and IMDB_votes files and join them. Check for and handle duplicate values in IMDB_movies dataset. Describe why you chose to handle the duplicates in such a way. What would be the alternative(s)?
5. Join IMDB datasets with the Netflix.csv file on title and director columns. Before joining the datasets, clean the title and director columns. What cleaning operations did you perform and why?
6. Find all movies that appear both in the Netflix and IMDB movies datasets.
7. Filter on movies that were categorized as comedies. Find the top 10 comedies on Netflix.
    a. What year(s) were they released in?
    b. What continent do these movies come from?
8. Assuming that you were to put this workflow in production, what kind of quality checks and tests would you implement?
9. Describe documentation and design principles you followed when building this workflow. You are welcome to use bullet points.

Data dictionary

Netflix dataset

Column name	Data type	Description
Show_id	V_String	Unique ID of the tv show or movie
Type	V_String	
Title	V_String	
Director	V_String	
Cast	V_String	
Country	V_String	
Date_added	V_String	date it was added on Netflex
Release_year	V_String	Original release year of the movie
Rating	V_String	
Duration	V_String	Total duration of the show/ movie
Listed_in	V_String	Categories/ genres
Description	V_String	

IMDB_movies dataset

Column name	Data type	Description
IMDB_title_id	V_WString	Unique ID of the movie
Title	V_WString	
Director	V_WString	

IMDB_votes dataset

Column name	Data type	Description
IMDB_title_id	V_WString	Unique ID of the movie
Avg_vote	V_WString	