CSIT 552 Final Project
Topic: Data Anatycis in Python

Dr. Boxiang Dong

1 Problem Description

Instructions. Please write code in a Python notebook to complete the following tasks.

- Task 0 (0 pts). Load the *netflix_titles* dataset. Use the magic command *%sh* and the Linux command *wget* to download the dataset from `https://msuweb.montclair.edu/~dongb/misc/netflix_titles.csv`.
- Task 1 (10 pts). Data Cleaning. The *duration* column describes the length of the movie/show. The rule is as follows: if it is a movie, the duration is described in the number of minutes; if it is a TV show, it is described in the number of seasons. Find the records that do not follow the rule and fix the error.
- Task 2 (20 pts). Data Transformation.
  - Task 2.1 (10 pts). The *country* column includes a string that lists the countries where the movie/show was produced. In case of multiple countries, they are concatenated with commas. Replace this column with single country.

    For example, show_id "s13" has *Germany, Czech Republic* in the country column. Replace that row with two rows, where one row stores *Germany* in country, and another row stores *Czech Republic*. All the other information in the row are simply duplicated into two rows.
  - Task 2.2 (10 pts). The *listed_in* column stores the movie/show categories. Similar to the *country* column, it may store multiple categories that are concatenated by comma. Similar to Task 1.1, create a new column named *genre* that stores a single category in each row.
- Task 3 (70 pts). Data Aggregation & Visualization.
  - Task 3.1 (10 pts). Count the total number of movies/shows by *release_year* and draw a lineplot to show the number of movies/shows since 2000.

- Task 3.2 (10 pts). Find the top-30 productive directors. The productivity of a director is measured by the number of movies/shows. Visualize the top-30 productive directors and their number of movies/shows with a barplot.
- Task 3.3 (10 pts). Make a lineplot that shows the average movie length and 95% confidence interval for every year since 2000. (x-axis is the year, y-axis is the length in minutes).
- Task 3.4 (20 pts). Make a lineplot that shows the number of movies/shows produced every year since 2000 in each of the following countries (United States, India, United Kingdom, Japan, South Korea) respectively. (x-axis is the year, y-axis is the number of products, and each country has a line). Make sure that you add the legend.
- Task 3.5 (20 pts). Find all the countries that produced more than 50 movies/shows in history. Make a mapplot where each country is plotted on the map as a circle and annotated. The circle size is based on the number of movies/shows it produced.

2 Submission Guideline

1. Work individually.
2. Please submit a .ipynb file.
3. Submit your solution on Canvas on time. A late penalty of 10 points for each late day applies. Any late for more than three days receives zero automatically.