Origin and General Info:

The dataset was obtained from kaggle, https://www.kaggle.com/datasets/rohitgrewal/netflix-data, and it contains around 8k rows and 11 columns. These columns include Category (Movie or Tv show), title, director, cast, country, release date, rating duration, and type. Some of these columns depend on the category:

- Duration shows runtime for movies and amount of season for tv shows
- Type (genre) has specific types for tv vs movies (The genres TV Dramas and Dramas)

According to the data card on kaggle: "This dataset is collected from Flixable which is a third-party Netflix search engine"; along with this, the dataset is rated as having a 100% for credibility.

Cleaning/preprocessing:

Plenty of cleaning and preprocessing went into altering the data to fit the needs of my visualizations. To list them:

- Immediately drop the Title, Director, Cast, and Description columns as they will go unused for what I'm visualizing
- Update the release date to just be a column of the year the show/movie was put on Netflix, rather than an exact date
- Then convert the duration to be ints. For movies this int will represent the length in minutes, for shows it'll represent the number of seasons

At this point, we can render most of the charts with matplotlib. However, the last chart showing the genre required quite a few more steps:

- Create a new pandas series from the Type (genre) column of the dataset
- Remove missing values and split the string that lists the genres

- Using the new series, which is just a list of every genre that appears in the dataset we get the count of each unique genre that appears in the dataset.
- Then drop a few genres that aren't really what we are looking for (like British TV Shows or Movies as this info was already previously covered in past charts)
- Lastly we need to merge some of the genres, for example Horror Movies and TV Horror can just go under the genre Horror

To make sure some genres weren't double counted, I had some code look through the dataset to see if any row had types like "Horror Movies" and "TV Horror" as that row would've then been double counted. Though I was unable to find any.

Now all the data has been cleaned and formatted.

Visualization choices: why you used a bar chart vs. a line chart, why certain colors, etc.

- The first graph that appears within the portfolio is a line chart. Simply chosen because I needed to show the change in the amount of titles Netflix added to their streaming service over time.
- Then I chose a pie chart to show the ratio of Movies to TV shows. I thought that given the small number of categories (just TV shows and Movies) a pie chart would be most apt to display it.
- Then comes two pretty basic bar charts. I wanted to display categorical data in descending order (greatest to smallest) and a bar chart seemed to be the best choice.
- I then chose to use a histogram to show the duration of the movies. I liked how it turned out.
- And finally, the genre graph. I was initially stuck on how to exactly parse and display this so with a help from chatGPT, then google because chatGPT lied, I was able to get it

to work (mostly related to accurately cleaning the data). It just so happens that I liked the sky blue color chatGPT chose and it kinda stuck.

Reflection:

For the most part, I think all the charts were well made to display the information I was trying to convey. That is, for the charts I included in the report. One chart that turned out to be a disappointment was the histogram for the number of seasons TV shows have. Looking back it should've just been super obvious it was going to start high at 1 season and decrease from there. Along with this was a chart showing how the duration of movies added to Netflix changed over time. There was an increase but not by a real noticeable amount, and I wasn't really sure what I could tell with the chart and its wider narrative point.

The main thing I would've liked to change is making the graphs look nicer. When creating them I was mostly concerned with the information they were telling rather than the stylist way they looked. Looking back at the last graph, I like the change in color and think I should've added more variety.

Code:

The code is in the data_visualization.py file within the project_1 directory in the github repository. It can be found here:
https://github.com/nOchsner/no/blob/main/project_1/data_visualization.py