In the folder, **netflix1** is my dataset to work on

o The data and its source [1 point]
Dataset is obtained from https://www.kaggle.com/ and it contains information about
Various TV Shows and Movies that are released by Netflix these past years.
It also includes - director name, date_added, genres, release year, duration.

o A description of your data exploration and data cleaning steps [1 point]
After reading the file with pd.read_csv(), I tried getting info about data with
Df.info about number of rows, columns and datatypes of every column
For cleaning data, I found following things that needed to be taken care of:

1) date_added column is object type: changing it to datetime dtype
2) For a given duration, there are 2 kinds of values: Seasons and minutes. I need to put int values in these elements. Upon further inspection, I found that seasons are for the duration of tv_series and minutes for movies.
3) I need to first separate the df into 2: one with tv_series and other one with movies.then I can separate their duration that is mins and seasons
4)There are multiple genres per row in listed_in in some cases, these genres are repeating values in the rows. For ease of looking and analysis, I'll convert the rows having multiple genres into 3 separate columns, namely genre column 1, 2 and 3.

o Two clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed. [1 point]
1) Compute the total number of tv shows and movies that netflix has produced over the years and draw inferences on it.
I used matplotlib and seaborn to show this in form of bar chart
unit of analysis: Year
comparison values: Compute and compare number of tv shows and movies over the years
how they are computed: with matplotlib and seaborn, having x axis as years and
Projecting the numbers of tv shows and movies with a bar plot.
In order to differentiate tv shows and movies I selected hue as 'type'
Which is the column that contained these two values
2) Suppose Rajiv from Pakistan is looking for a TV series from pakistan. He is only looking for series to binge watch(series having 4 or more seasons)
unit of analysis: TV_show titles (from df_tv)
comparison values: country - pakistan, duration_seasons >= 4 (this is why I cleaned duration column)
how they are computed: with df indexing I put in the comparison values in df_tv

o A description of the program [1 point]
– Import Libraries
– read file
– analyze the data set in .info() , .head() and check for duplication
–4 things to clean

–after cleaning store the cleaned and separated dataframes
–Answer 2 questions, 1st one has output in form of graph, second in form of dframe.
I converted the second one to csv and just took screenshot of the graph

o A description of the output files [1 point]

1st one has output in form of graph so I just took screenshot of the graph
2nd answer is in form of dataframe, I saved this in csv type

o The source data file.

It contains information about various TV Shows and Movies that are released by Netflix these past years.
It also includes - director name, date_added, genres, release year, duration.