In the folder, My_dataset_netflix_json 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, description.

o A description of your data exploration and data cleaning steps [1 point]

After reading the file with pd.read_json(), I tried getting info about data with functions: my_df.info, my_df.shape etc 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) use .fillna() to fill up any null values in dataframe
3) sweeping out null values from new director and actor df that I create in program

o Two clearly stated comparison questions with the unit of analysis, the comparison values and how they are computed. [1 point]

1) Calculate top 5 Directors on netflix based on the number of content produced over the years. I also did a bar plot of top 5 directors' output data
unit of analysis: director
comparison values: Compute and compare number of content (rows having same director name) over the years
how they are computed: grouping the df by director column which indicates their total number of content produced. Then finding top 5 director names and depicting in form of bar plot

2) Suppose someone wants to watch content of most famous actor, analyze the df and find top actors to recommend based on the number of content they produce.
unit of analysis: actor
comparison values: number of tv show or movies in which each actor has starred in
how they are computed: group by function based on actor and finding number of rows containing their name.

3) Perform Sentiment Analysis of all content produced since 2010 based on three sentiments: Negative, Neutral and Positive.
unit of analysis: description of content, and release year (after 2010)
comparison values: words used in description and inferring the kind of sentiment it depicts
how they are computed: using library TextBlob, I used .sentiment.polarity function to find sentiment of each description by iterating it row wise. Then used a bar plot to show number of content produced and their sentiment over the years after 2010 (also save the output to csv).

o A description of the program [1 point]

– Import Libraries
– read json file
– analyze the data set in .info() , .shape
– clean data
–Answer 3 questions

o A description of the output files [1 point]

Each question’s answer is saved in form of csv file and also depicted as graph in .ipynb file