SOC - Security Analyst Interview Exercise

The following is a short exercise to allow us to understand how you problem solve. This exercise should take no more than one (1) hour. Documenting your thinking is more important than the answer. Please include any scripts or visualisations that you may have used in determining your answers.

Included with this exercise are two (2) files which contain movie related data pulled from IMDB:

**title\_basics\_2018.csv** - A file containing all movies released in 2018, including the following columns:

|  |  |  |
| --- | --- | --- |
| Field | datatype | description |
| tconst | string | alphanumeric unique identifier of the title |
| primaryTitle | string | the more popular title / the title used by the filmmakers on promotional materials at the point of release |
| originalTitle | string | original title, in the original language |
| year | YYYY | represents the release year of a title |
| runtimeMinutes | int | primary runtime of the title, in minutes |
| genres | string array | includes up to three genres associated with the title |

**title\_ratings.csv** - the ratings for ALL movies, made up of the following columns:

|  |  |  |
| --- | --- | --- |
| Field | datatype | description |
| tconst | string | alphanumeric unique identifier of the title |
| averageRating | float | weighted average of all the individual user ratings |
| numVotes | int | number of votes the title has received |

Please answer the following questions, explaining your reasoning:

1. According to the provided dataset, how many 2018 films were categorized as a Comedy?
2. According to the provided dataset, how many 2018 films got a score of 8.0 or higher? (Note that this will require joining the two datasets together)
3. What was the best film of 2018?
4. Do audiences prefer longer films, or shorter films? You may choose to simply outline your methodology to approach this problem.

