SOC - Security Analyst Interview Exercise

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.

Luis Imlauer

THOUGHT PROCESS AND SOLUTION

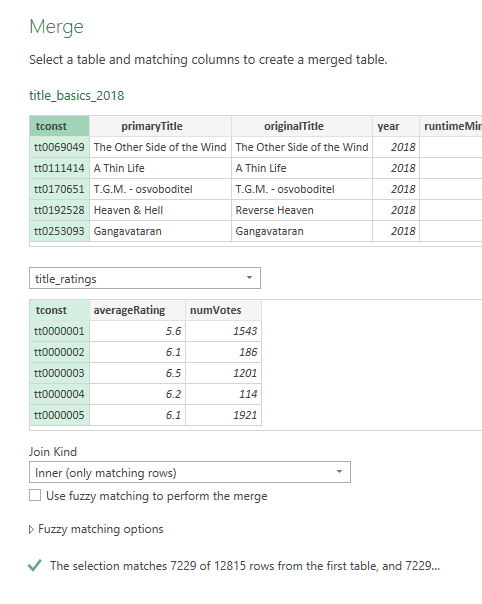
Thanks for the opportunity! Let's start

I open every file, and see that there are 12816 movies in total, and a file that contains almost a million rows but only rating and number of votes.

13k is not a lot of rows so we can easily do this with excel and don't need to use python or R.

I read every data type and format, and looking at title\_ratings.csv I am also thinking that we should measure a 'weight' for the number of votes and average rating, could be interesting.

First we go to "title\_basics\_2018" (from now on: Title basics) and change \N to Uncategorized in the genres column.

Primary Title and Original Title look correct, and every movie is from 2018.

Duration has some strange values, like 1 or 2. I checked some of the films and they are shorts or mini documentaries (10 minutes), so I won't be changing any data as this is supposed to take only an hour; but know that this may change the conclusions in a "real world case" and I would modify the information or at least check with the person in charge of the data sources.

Now we need to merge the two files, this would be an inner join (using tconst) as we need to know the movie and its rating. If we don't have the rating then we don't need to keep the movie, hence the inner join and not left join.

ANSWERS

**1\_** According to the provided dataset, how many 2018 films were categorized as a Comedy?

To see the films categorized as Comedy I simply filter them in excel where Genre contains Comedy. In total, there are 1665 Comedy films (of which 566 ONLY contain Comedy as the genre).

**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)

As I already joined the files, I can just filter 8+ score films. In total, 780 films from 2018 got a score of 8.0 or higher.

**3\_** What was the best film of 2018?

By Rating, there is only one film with a 10 score, with only 5 votes: “Exteriores: Mulheres Brasileiras na Diplomacia”. It’s a documentary.

As we already mentioned, we need to compensate the low amount of votes, so for this answer we will be doing a weighted average.

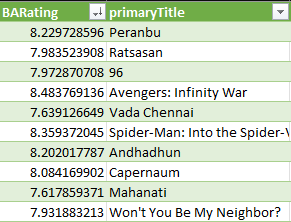
Let’s calculate the 95 percentile (result: 5725.2) and use that as our minimum amount of votes for a movie to be considered.

We filter by movies with only +5725.2 upvotes and calculate the mean average rating (C).

We will be using Bayesian Average.

**BARating = ( (v / (v + m)) \* R ) + ( (m / (v + m)) \* C )**

Where:

v: is the number of votes

m: minimum amount of votes

R: average rate

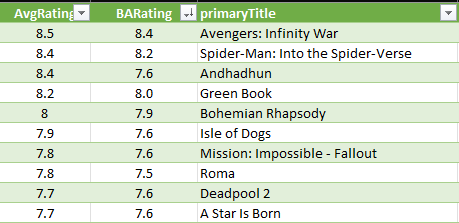
C: overall average rating of every movie

C = 6.445

m = 5725.2

With those values, the top 10 would be:

With **Peranbu** in first place.

If we wanted to see only very popular movies (measured by the amount of votes) then we could change m to be the 99 percentile (result: 50352.6); which would give us an overall average of 6.867

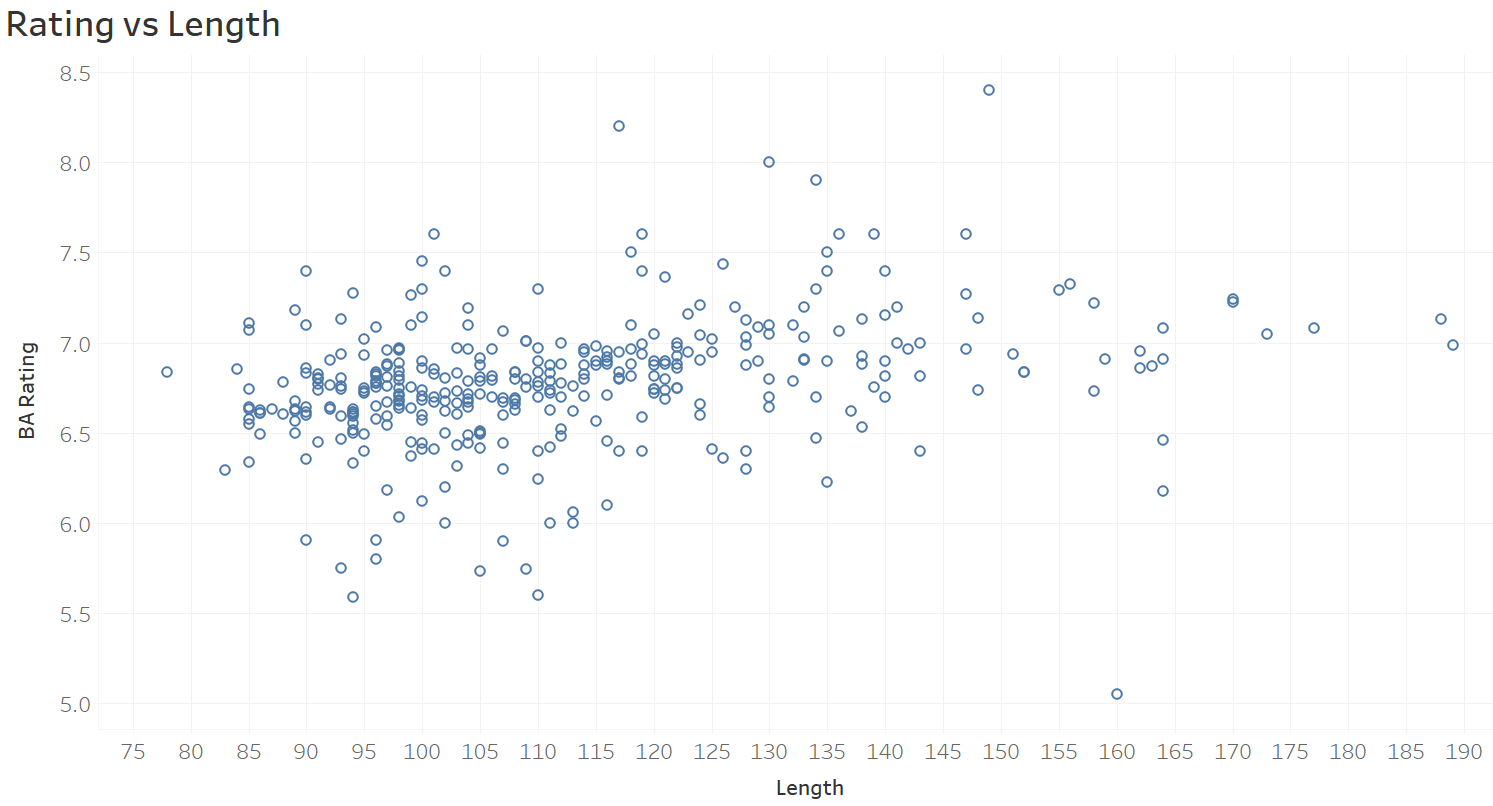
So, with C = 6.867 and m = 50352.6 we would get this top 10:

Where **Avengers: Infinity War** is the top rated movie.

**4\_** Do audiences prefer longer films, or shorter films? You may choose to simply outline your methodology to approach this problem.

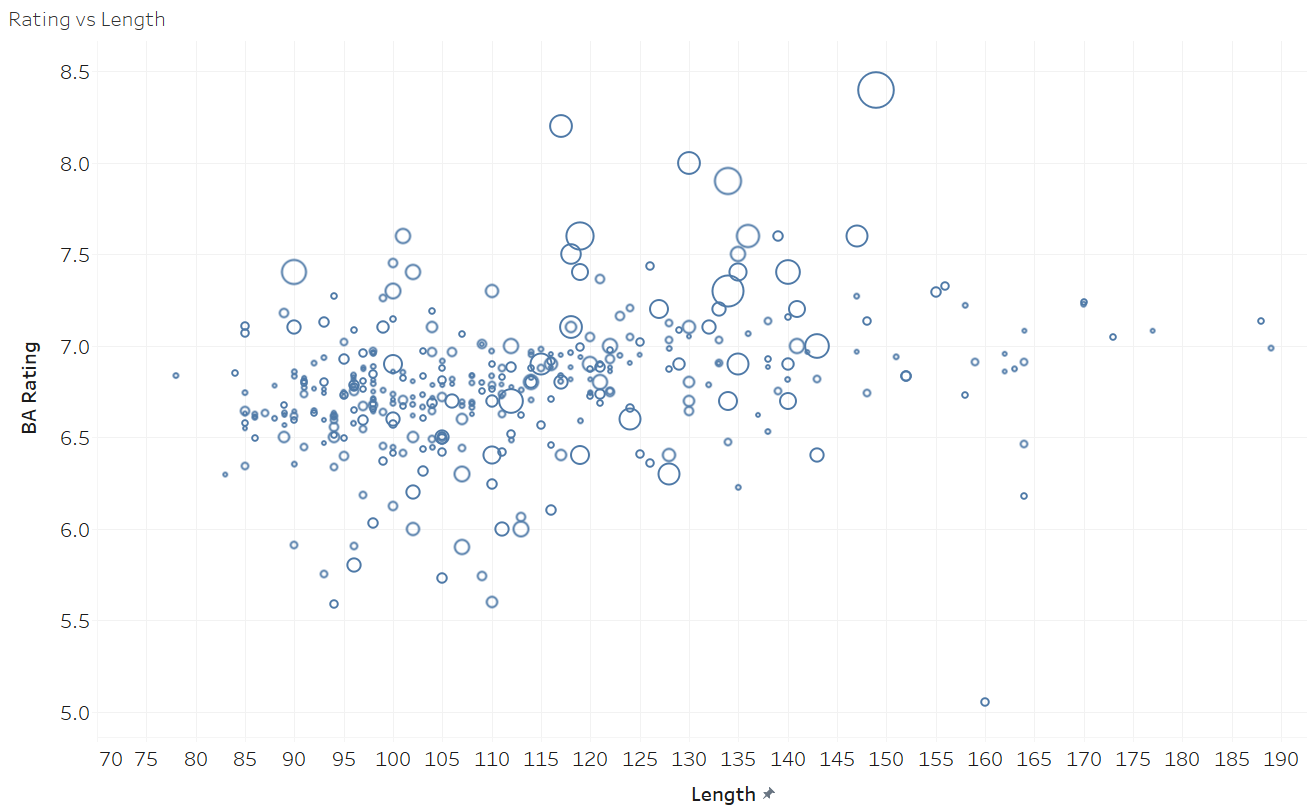
For this question we should use a correlation measure. Let’s consider our new “BA Rating” and “Length. We will be using Tableau, to draw a scatter plot to show the relationships between the variables, and Excel to calculate the correlation between BA Rating and Length.

*I’m only going to export movies with +5725.2 votes.*



In this plot we can clearly see there is no correlation between the two variables, as the dots are located around the center of BA Rating for all values of Length.

In the next plot, I filtered by amount of votes; which gives us a bigger circle depending on the amount of votes the movie got. This does not provide any new information.



In conclusion, there is **no correlation** between BA Rating (our new variable measuring movie rating) and Length of the film. Which means that audiences don’t tend to rate movies higher or lower based on their duration.