Some Thoughts on the IMDB dataset

**Missing Values:**

**Type Vaiable** **NAs Missing Text Dealt With**

F color 0 19 Replaced “Color”

F director\_name 0 104

N num\_critic\_for\_reviews 50 NA

N duration 15 NA

N director\_facebook\_likes 104 NA

N actor\_3\_facebook\_likes 23 NA Removed Obs

F actor\_2\_name 0 13 Removed Obs

N actor\_1\_facebook\_likes 7 NA Removed Obs

N gross 884 NA

F genres 0 0

F actor\_1\_name 0 7 Removed Obs

~~F movie\_title 0 0~~

N num\_voted\_users 0 0

N cast\_total\_facebook\_likes 0 0

F actor\_3\_name 0 23 Removed Obs

N facenumber\_in\_poster 13 NA

~~F plot\_keywords 0 153~~

~~F movie\_imdb\_link 0 0~~

N num\_user\_for\_reviews 21 NA

F language 0 12

F country 0 5 “USA”

F content\_rating 0 303 “Not Rated” == “”

N budget 492 NA

F title\_year 108 NA

N actor\_2\_facebook\_likes 13 NA Removed Obs

N imdb\_score 0 0

~~F aspect\_ratio 329 NA~~

N movie\_facebook\_likes 0 0

**The Numerical Variables**

1. Concerned about the number of missing values in “Gross” and “Budget”

**The Factor Variables**

1. With regards to the factor variables, I think we can get rid of

a. Movie\_imdb\_link

b.Plot\_keywords and movie genre are very complicated.

2. The content\_rating variable has the following levels:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| "" | (303) | TV-PG | (13) | G | (112) |
| Approved | (55) | TV-14 | (30) | PG-13 | (1461) |
| Unrated | (62) | TV-MA | (20) | PG | (701) |
| Passed | (9) | TV-Y7 | (1) | GP | (6) |
| Not Rated | (116) | TV- G | (10) | R | (2118) |
|  |  | TV-Y | (1) | X | (13) |
|  |  | NC-17 | (7) | M | (5) |

From good old Wikipedia, I think we can

1. combine M, PG and GP – they are the same thing from different time periods
2. possibly combine “Not rated” and “Unrated” and “”
3. The “TV-“ ratings are designed specifically for US television. Comparing the descriptions to normal movie content ratings I think:
   1. G = TV-Y + TV-Y7 + TV-G
   2. PG = TV-PG
   3. R = TV-MA (possibly)
   4. I also feel like TV-14 and PG-13 are close enough to be the same thing.
   5. We can of course leave the movie and tv ratings separate if you think that is important.

<https://en.wikipedia.org/wiki/Motion_Picture_Association_of_America_film_rating_system>

<https://en.wikipedia.org/wiki/TV_parental_guidelines_(US)>

4. Also looked at Language and Country, but couldn’t find any obvious typos or double ups

3. Actor names