Elizabeth Earl

DSC540-T301

Week3&4: Milestone 1

* 3 data sources:
  + Text

    Description automatically generatedThe CSV file I will be using contains top 1000 films made in the 2000’s:
  + API being used can be found: [IMDb API Documentation - IMDb API (imdb-api.com)](https://imdb-api.com/api#Top250Movies-header) which includes the top 250 movies of all time
  + The HTML style website I will use: [The Numbers - Top-Grossing Movies of 2021 (the-numbers.com)](https://www.the-numbers.com/market/2021/top-grossing-movies#:~:text=Top%20Grossing%20Movies%20of%202021%20%20%20,%20%20Action%20%2021%20more%20rows%20) contains the top ~300 films of 2021 so far
* Finding a relationship between datasets will rely on whether any of the films appear across different platforms. For example, a top film in 2021 may/may not indicate it is considered a top film in general or of all time. Movie titles will be used to find rank relationships. Because there is not many relationships between the datasets currently I may have to do some extra research to help add columns to some datasets to create new relationships.
* To accomplish the entire project, I will have to rely heavily on cleaning up the data. Because different datasets abbreviate certain movie titles differently, I will have to consider all possibilities. This is to say titles like: The Lion King can be written as Lion King or Lion King, The across different platforms and I have to ensure that I am correctly matching movie titles to one another to have correct relationship mapping. Across the different datasets I also noticed they have different measure of what makes a movie a top movie. While The Numbers dataset is using revenue the other datasets are using an arbitrary rating which by the looks of it seems to be a out of a perfect 10 in the CSV file as well in the IMDb API. Once datasets are properly merged based on title, I will use their rank number to create visuals of each film’s popularity. Because every dataset has a rank which cannot be used as a relationship between datasets this rank will work to help determine the film’s success. The rank/success will be portrayed in visuals which will make it clearer to the audience which film was #1 (if any) across the board. In terms of what type of visuals, I am currently planning on having a line chart for a handful of “top” films to see their growth by dataset type. I also plan on using a histogram chart using again the “top” films to compare the revenue each film made.