Final Project Proposal

Jie Zou

3/26/2022

Intro

I have navigated from different websites to find a specific data set that interests me. Instead of exploring covid data set again, I prefer something that relates to my spare time. Despite the time that I spend on watching movies, I found a data set from kaggle showing the top 1000 highest grossing movies. I observed the first few rows from the data set and found out seven out of ten movies that I've watched, in addition, some of the seven I even re-watched them. Therefore, I made the decision, this will be my dataset for the final project.

Data info

The link for dataset will be here

The dataset has 11 variables and 918 observations. Well, technically, it has 10 variables, the first variable is index which has been added into data by R, the structure is listed below

- Title(str): Movie name
- Movie info(str): Movie's brief introduction
- Distributor(str): Distributor name
- Release.Date(str): Original release date
- Domestic.Sales(in $)(int): Domestic sales of the movie
- International.Sales(in $)(int): International sales of the movie
- World.Sales(in $)(int): Total sales(domestic + international)
- Genre(list): Genres the movie belongs to
- Movie Runtime(str): Movie runtime
- License(str): License

Interests

As we all know, the selling point of a movie has a certain relationship with the combinations of theme, genre, producer, director, actor, etc. When these factors are accumulated together, it is possible to publish a movie with a good reputation and a high box office. Except for the actual human factors, in other words the actors, directors and so on, the base of the movie plays an important role. Therefore, I am going to visualize

1. The distribution of genres is and the gross each genre obtained

2. The relationship between domestic sales and international sales in different movies
3. The distribution of distributor in distinct genre

Technique

Instead of using dash/shiny app, I am going to try to use Streamlit since it becomes a new ‘star’ in my eyes to maximize user interactive graphs to ensure they can use selection box to explore and understand data better.