Marshal Ruzvidzo mruzvidz@andrew.cmu.edu

Cleaning the dataset

- The director's column was separated into multiple columns using the piping character as the delimiter. In the end, each movie had multiple columns of directions
- The above splitting was done for the cast and producing companies' columns
- The data set was sorted in descending order of popularity

The questions for this data set were

1. The revenue and budget for the top ten popular movies
2. The cast for top ten popular movies
3. The directors of top ten popular movies
4. The producing companies for top ten popular movies

Question1 : The revenue and budget for top ten popular movies

Firstly, the top ten popular movies across the dataset was found and then then the dataset was sorted in ascending order of popularity.

A combined bar graph of revenue and budget for each movie in the top ten popularity ranking was plotted as shown below.

t

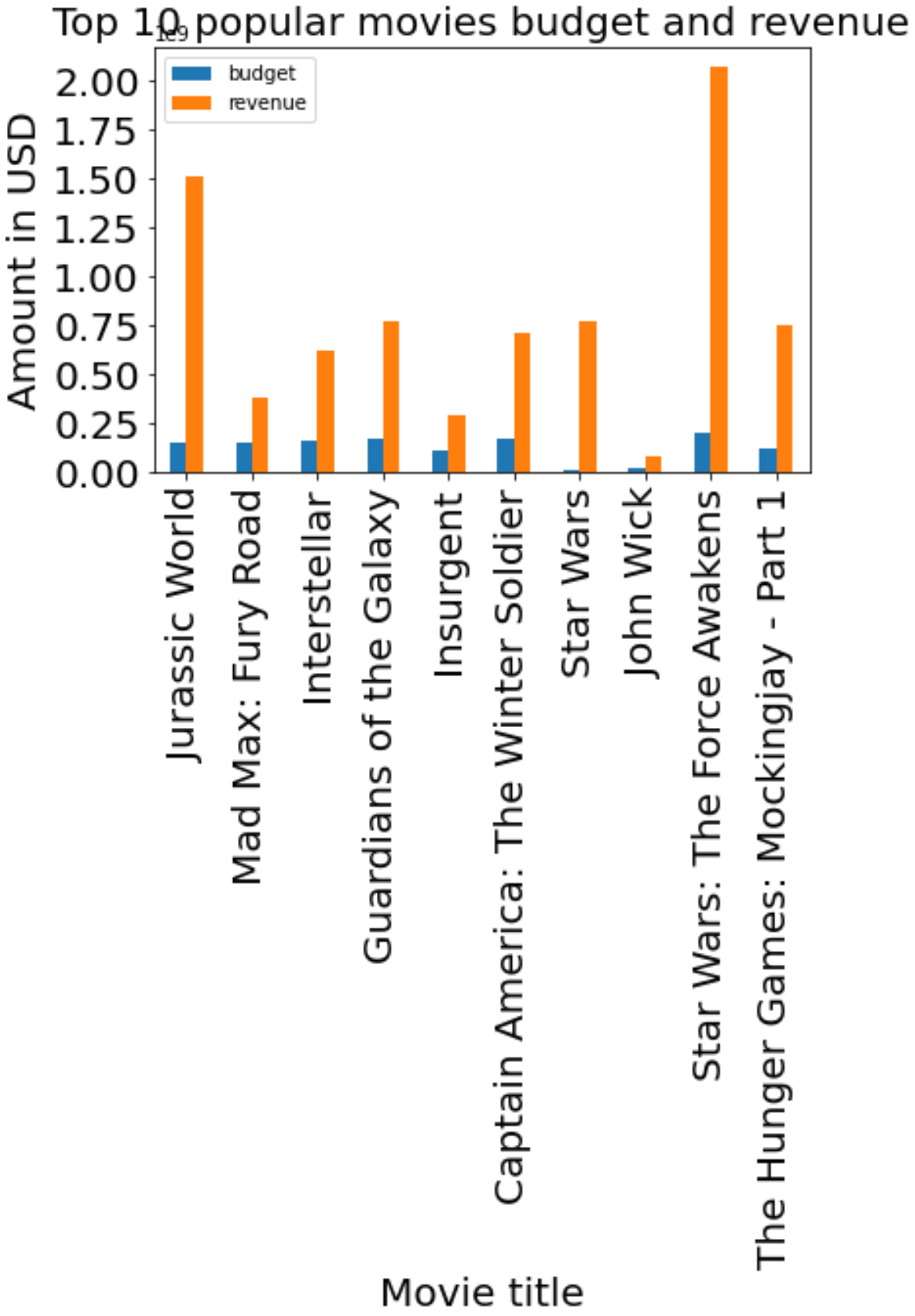

Figure 1 Top ten popular movies budget and revenue

From figure 1 it can be concluded that Jurassic World is the most popular movie in the dataset. Also it can be noted that all top ten popular movies had almost similar revenues. John Wick is one of the popular movies with lower budget and lower revenue. Its popularity could be attributed to the good story line of the movie. Hence it can be concluded that popularity of a movie cannot be attributed to the budget or revenue it generates

Question 2: Who were the cast for top ten popular movies?

As mentioned in question 1 the first ten popular movies in the data set were selected and sorted in ascending order. These movies had many actors in one movie and the cast column was split into multiple columns. Then the number of appearances of each actor in top ten popular movies was counted and plotted on a bar graph below.

p

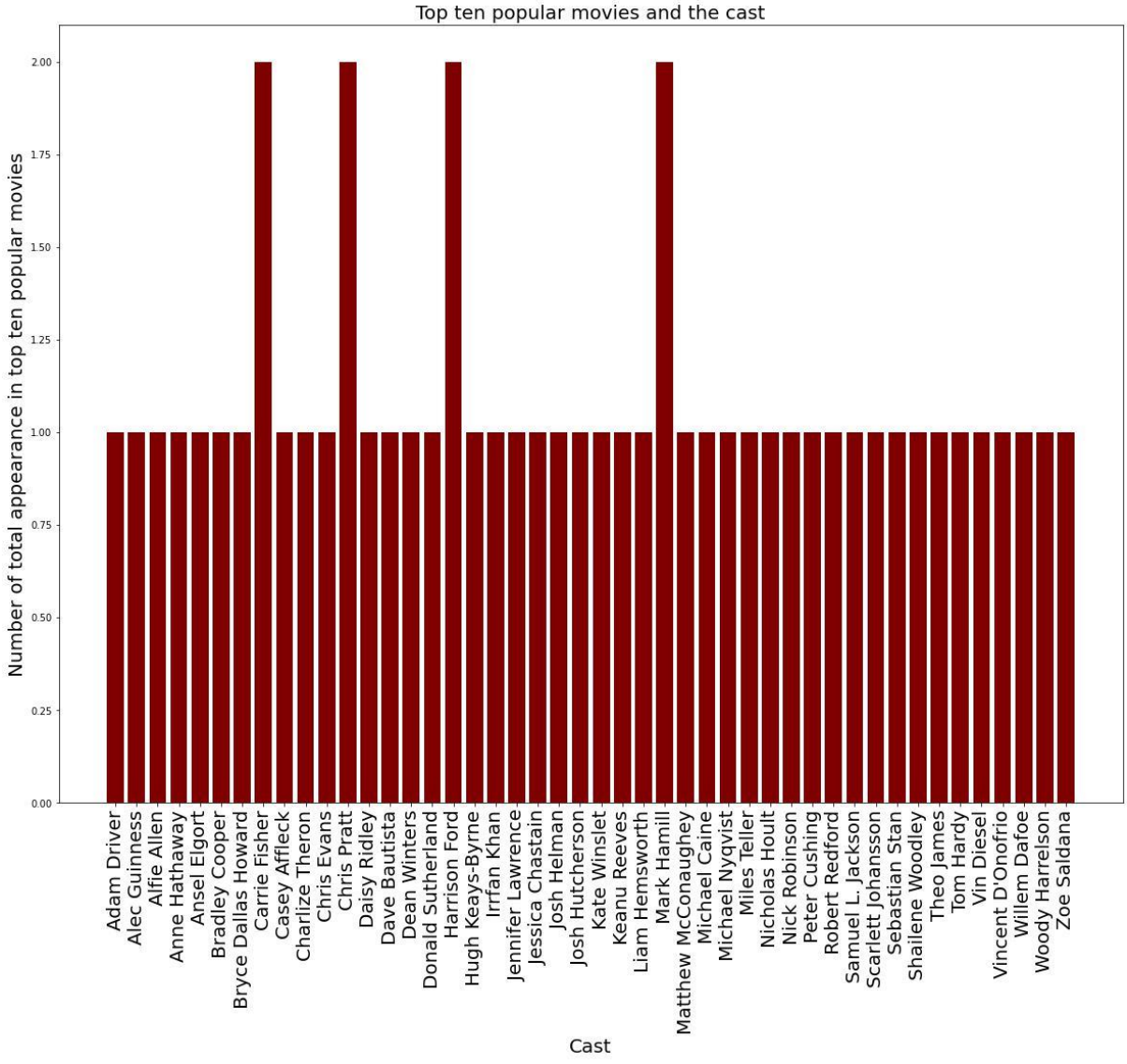

Figure 2 cast for top ten popular movies

From figure 2 it can be concluded that 3 actors (Carrier Fisher, Harrison Ford and Mathew McConaughey) appeared more than once in top ten popular movies. Hence it can be concluded that for a movie to be popular cast does not play a bigger role.

Question 3: Who directed the directors of top ten popular movies?

Firstly, the top ten popular movies were sorted in ascending order. Some movies had more than one director and the director column was split into multiple columns so that each director has his/her column. Then the number of appearances of each director was counted in the top ten popular movies and the results were plotted on the bar graph below.

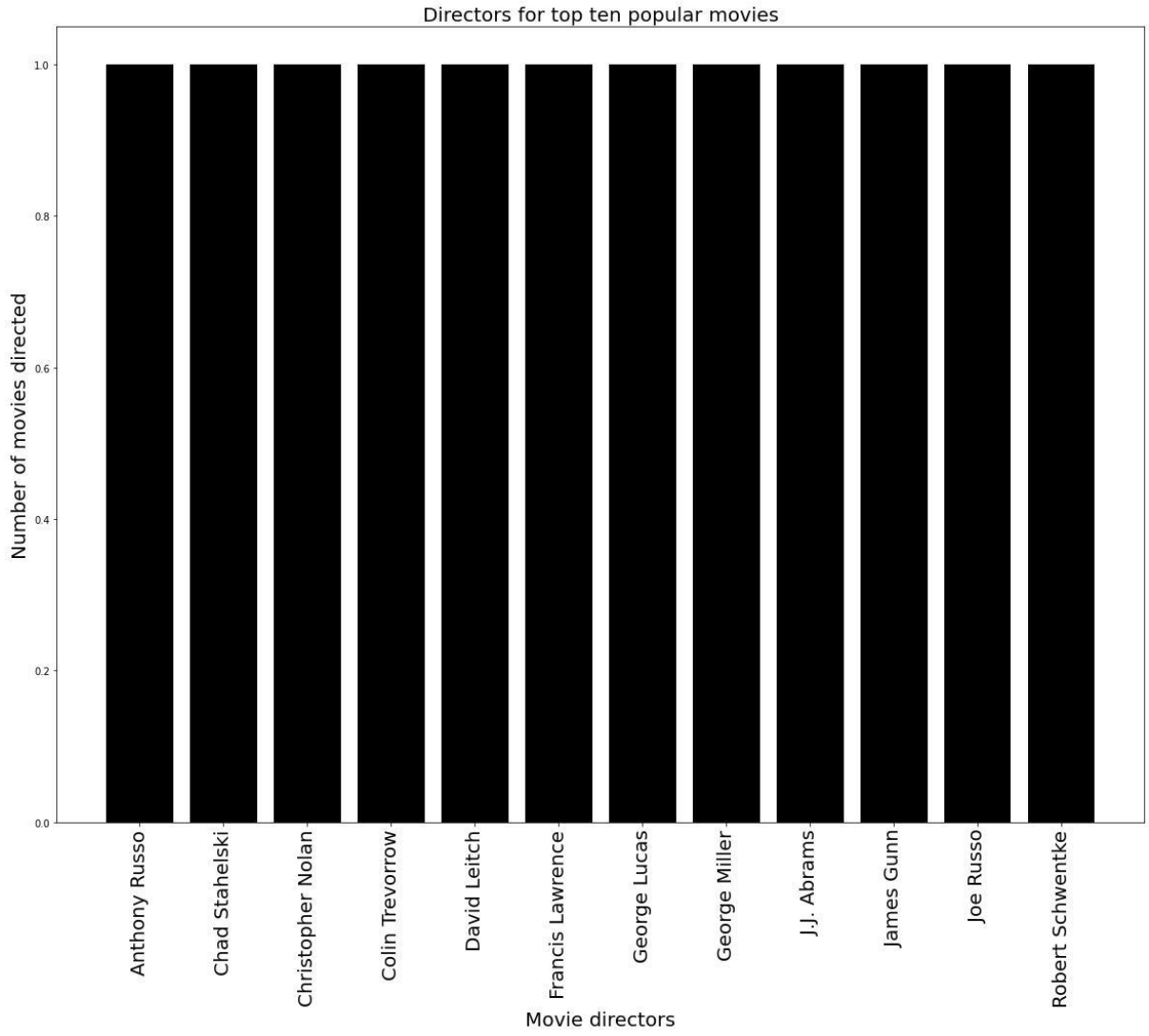

Figure 3 directors for top ten movies

From fig 3 it can be concluded that all top ten popular movies had different directors

Question 4: Who were the producing companies for top ten popular movies?

Firstly, the top ten popular movies were sorted in ascending order. Some movies had more than one producing companies and the producing_companies column was split into multiple columns. Then the number of appearances of each producing company was counted in the top ten popular movies and the results were plotted on the bar graph below.

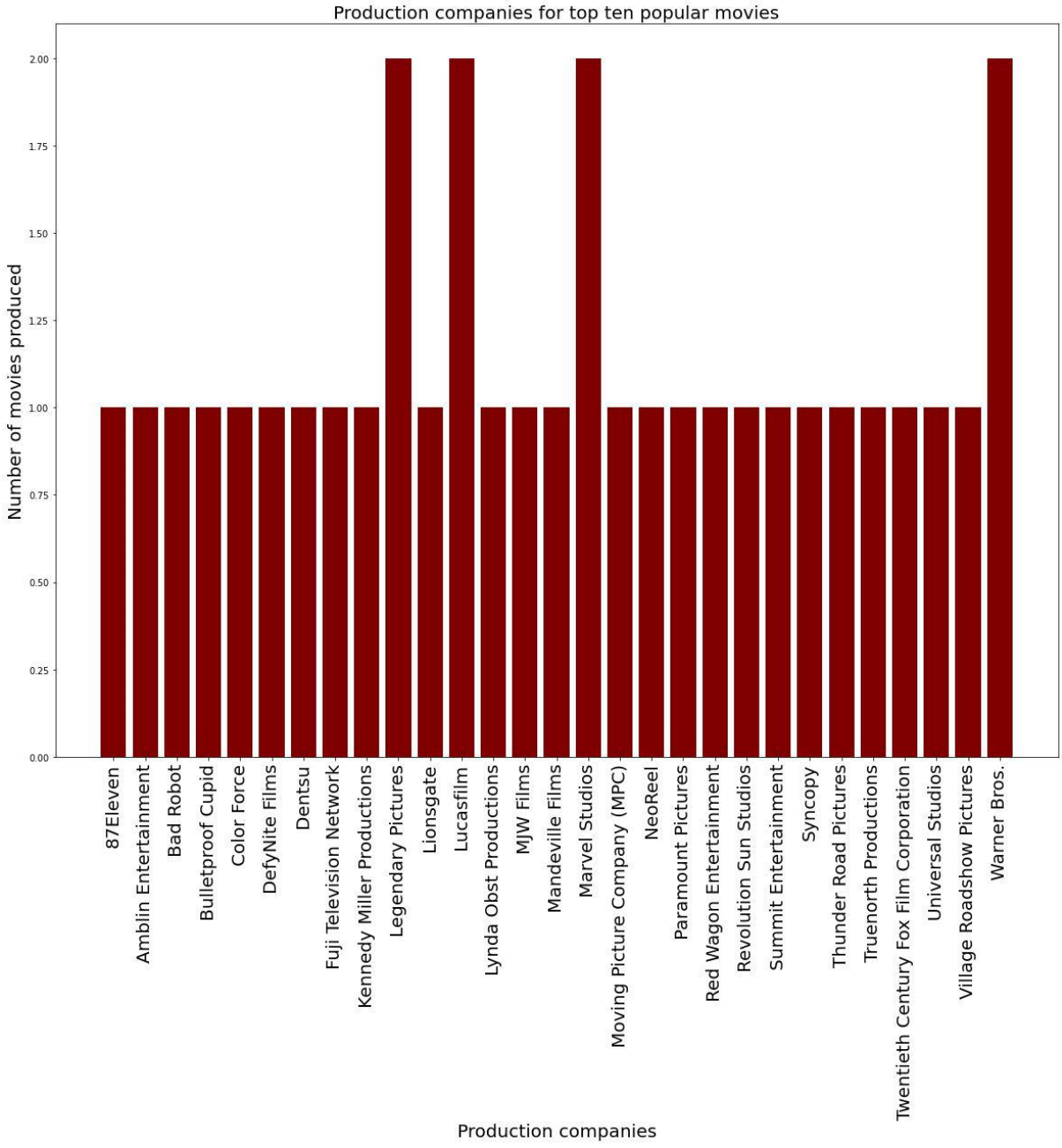

Figure 4 producing companies for top ten popular movies

From fig 4 it can be concluded that 4 companies (Legendary pictures, Lucasfilm, Marvel studio and Warner Bros) produced more than one movie that were in the top ten popular movies

Limitations of the conclusions derived

The limitations of the conclusions include but are not limited to the fact that this dataset assumes the same weight of popularity as the early years of the dataset. For example, a movie can have a popularity score of 70% in the 1960s, but this does not mean it will have the same popularity score in the 2000s. This might be attributed to the change of taste of the audience. Hence a weighing or smoothing function would have been implemented to scale the popularity as a function of time.