TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The data set contains information for 4803 different movies. The information being stored is: Budget, genres, homepage, id, keywords, original_language, original_title, overview, popularity, production_companies, production_countries, release_date, revenue, runtime, spoken_languages, status, tagline, title, vote_average, vote_count

DATA CLEANING

To make it easier to analyse the data we have removed the following columns from the data set:
Keywords, Homepage, Status, Tagline, Original language, Overview, Production companies and Original title

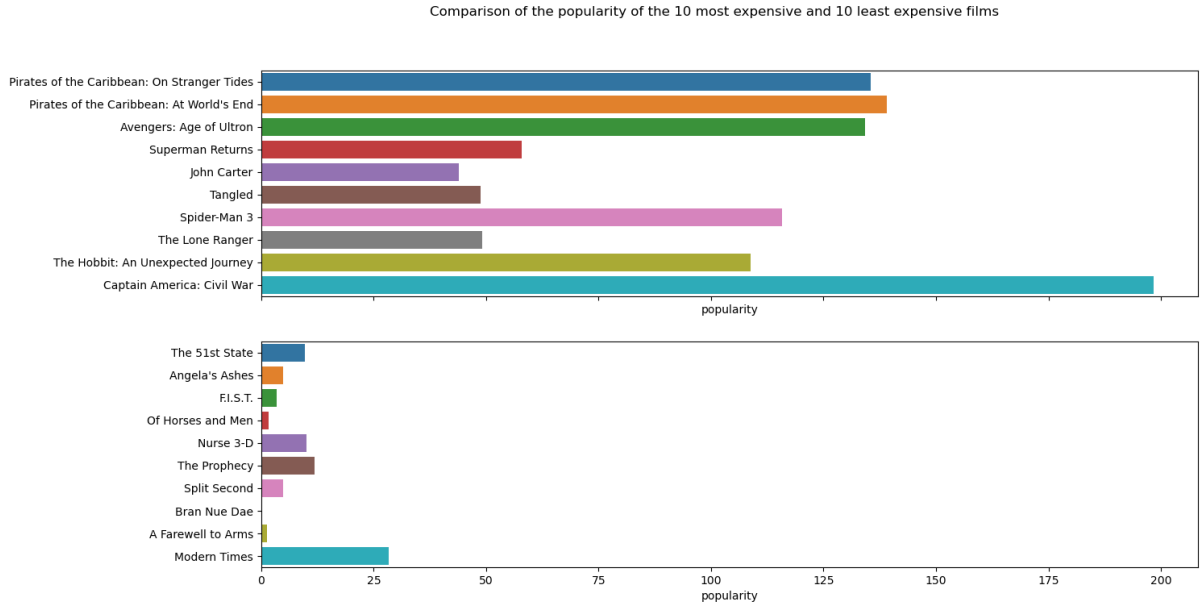
To manipulate the data easily we have made sure the release date is being read in as a DateTime and then we have extracted the release year into a new column.

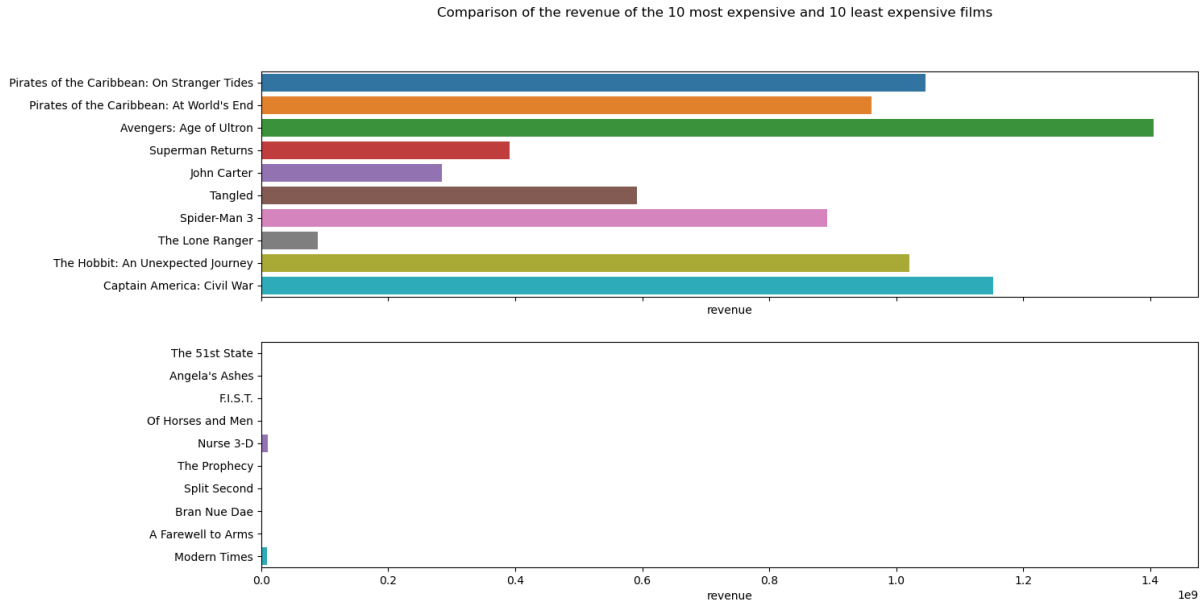
MISSING DATA

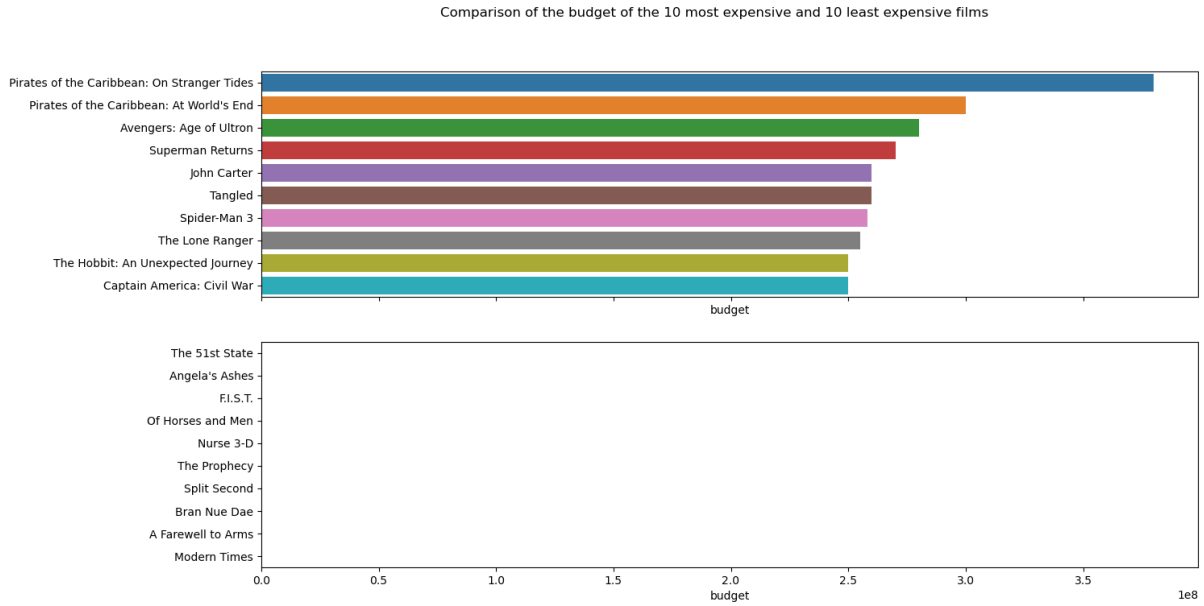
There were no duplicate rows that needed to be removed, but there are movies in the database that don't have a budget or revenue. For these films that are missing this information, we will remove them from the data set.

DATA STORIES AND VISUALISATIONS

The first explorative data analysis we did was to compare the 10 most expensive and least expensive films.

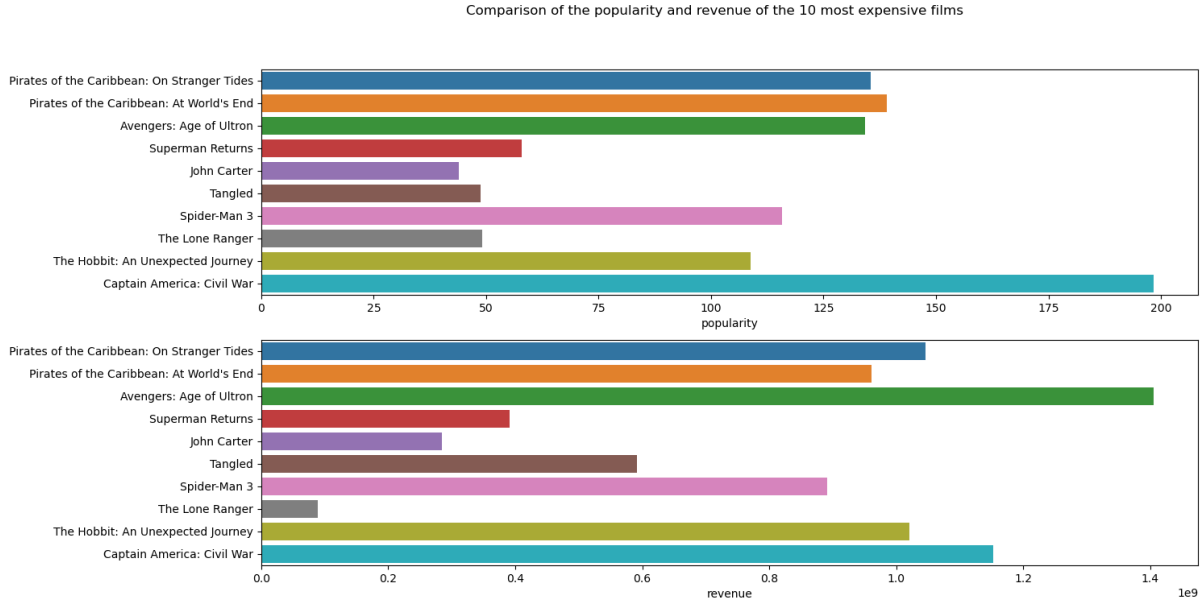






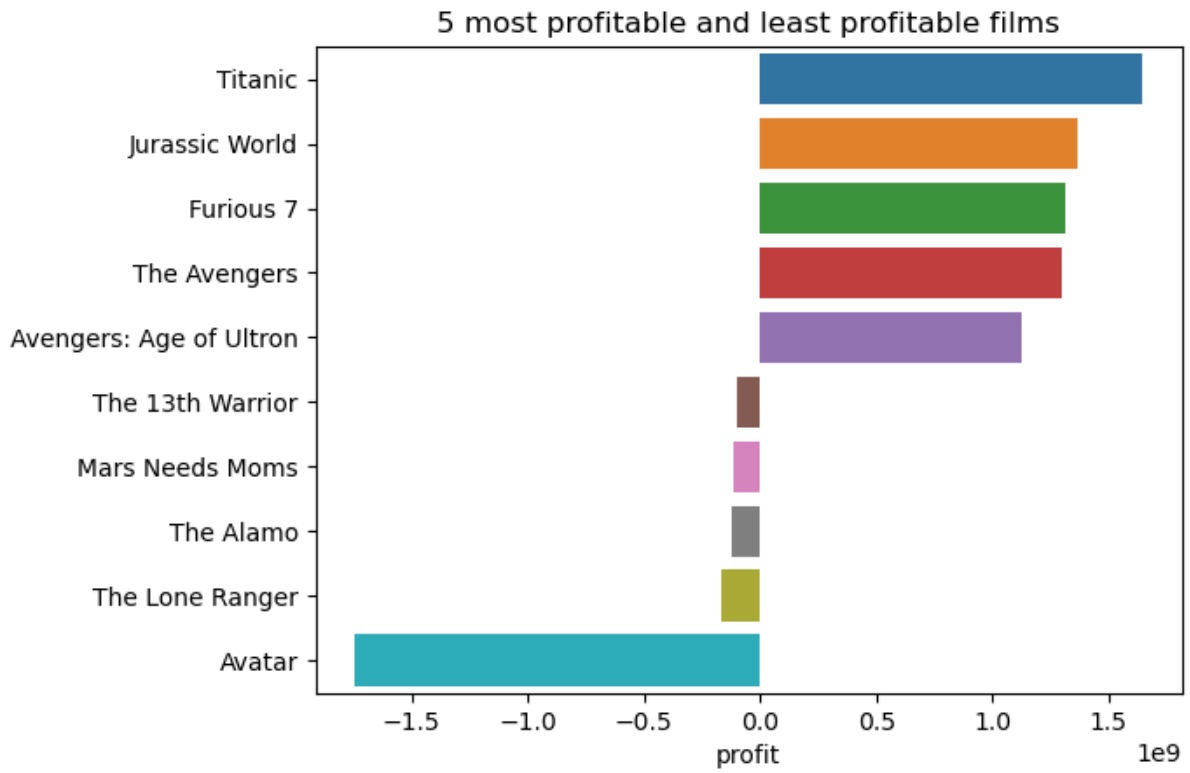
From these graphs, the main conclusion to draw from them is that the although there is a huge gap in the revenues and budgets there isn't the same gap in popularity. One particularly interesting thing is that the film Modern Times has the smallest budget of all the films in the data set but its popularity is only half of John Carter, this would imply that the money spent on John Carter wasn't well spent.

This leads to the next area we will explore which is whether the most expensive films were good value for money or not.

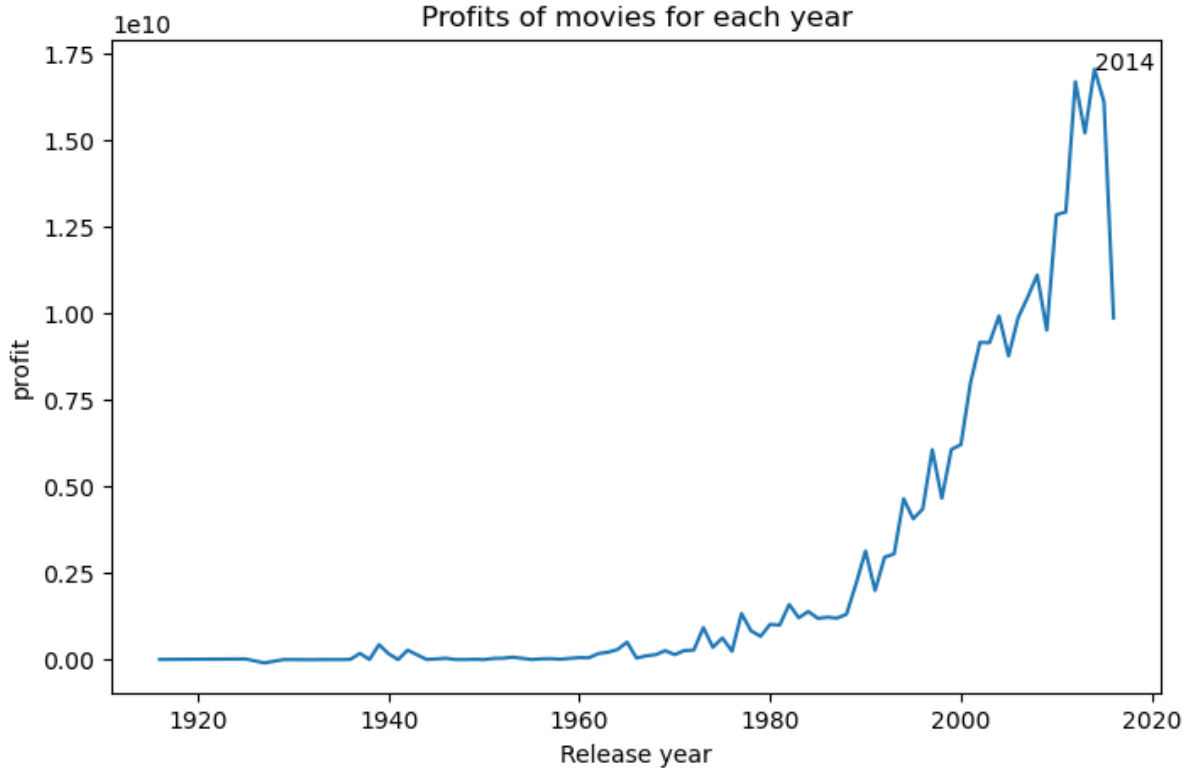


Here we can see that spending more on a movie does not translate into more revenue or a more popular movie. The lone ranger performs the worst for revenue with its revenue being considerably less than the other 9 films. John Carter is the least popular of the 10 movies even though it has the 5th largest budget. The film that seems to be the best value for money is Captain America: Civil War which has the highest popularity and 2nd highest revenue when it has the smallest budget of the 10 movies.

The next area we will explore is the profit of movies.

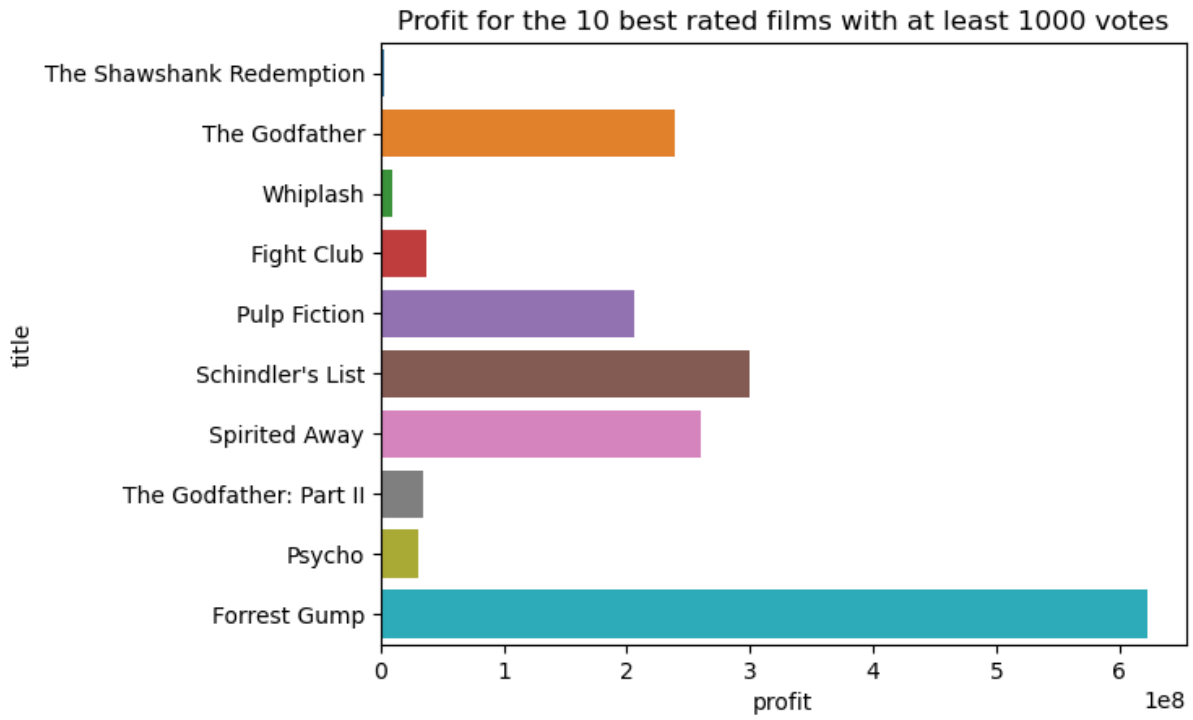


Here we can see that the money spent on producing The lone ranger was bad value as it has the 2nd largest loss of the entire data set. This again shows that spending more on a movie does not translate into a more successful movie.



Here we can see that profit has rapidly increased since 1980 and then it peaked in 2014.

Now we want to explore whether a film that is highly rated by viewers means that it was also a commercial success and generated a large profit.



Here we can see that a film that is highly rated by viewers does not translate into a larger profit generated for the film.

THIS REPORT WAS WRITTEN BY : Tom Anderson