Exploratory Data Analysis on the Movies Data Set

Introduction

This is a dataset made of 4803 rows and 20 columns and the topic is related to movies.

The columns in the dataset are 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

In the movies file, I decided to remove the duplicate rows, after that there were 436 rows and 12 columns.

I selected all the movies by the column “budget” and I removed the rows with a budget equal to zero, after that the rows were 368.

I replaced the format in column “release_date” with the correct datatype format using the datetime library.
I changed the “budget” and “revenue” columns format to an integer using NumPy’s int64 method.
I defined a function that parsed columns in JSON format to string format.

MISSING DATA

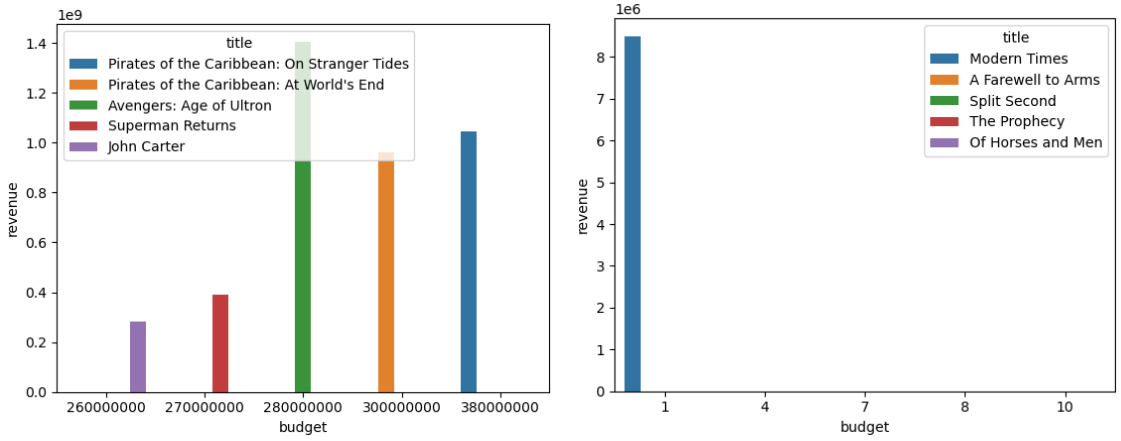
In the beginning, there were 3091 missing values in the column ”homepage” and 844 missing values in the column “tagline”.

After my data cleaning, there weren't missing values anymore

DATA STORIES AND VISUALISATIONS

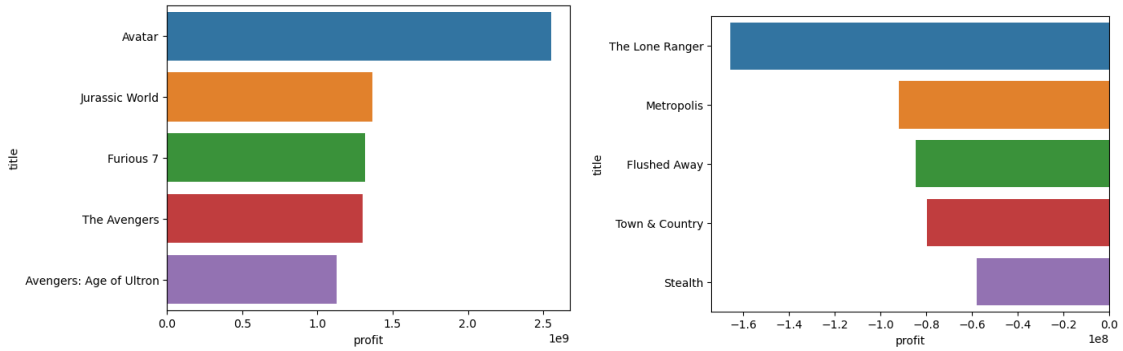
Expensive vs cheapest movies comparison:

As we can see in the graphs below, the 5 more costly movies are Pirates of the Caribbean: On Stranger Tides, Pirates of the Caribbean: At World's End,
Avengers: Age of Ultron, Superman Returns, and John Carter.
We notice that if we consider the cheapest movies, only Modern Times was able to generate revenue even without spending money.
Regarding the expensive movie, every movie was able to generate profit. So it is worth so invest money in the movie because it will generate more profit.



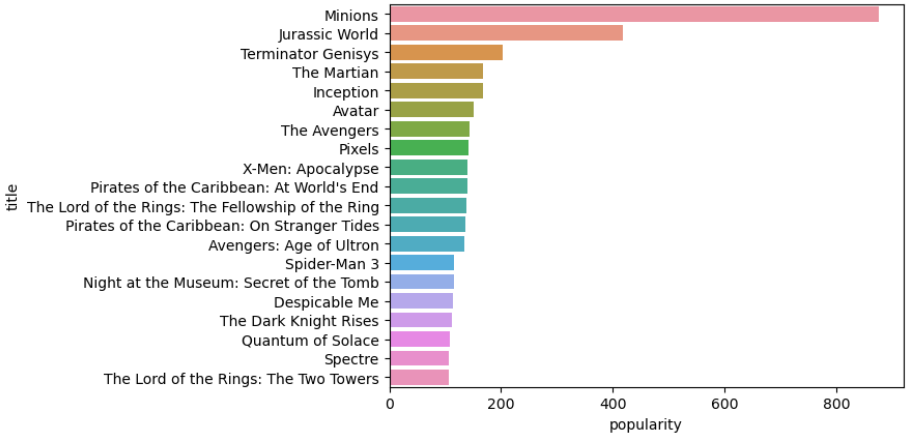
Most profitable vs less profitable movies:

The graphs below, show the most profitable vs the less profitable movies. We can see that the movie that generates more profit is Avatar and the one that generates less profit was The Lone Ranger which has lost money.



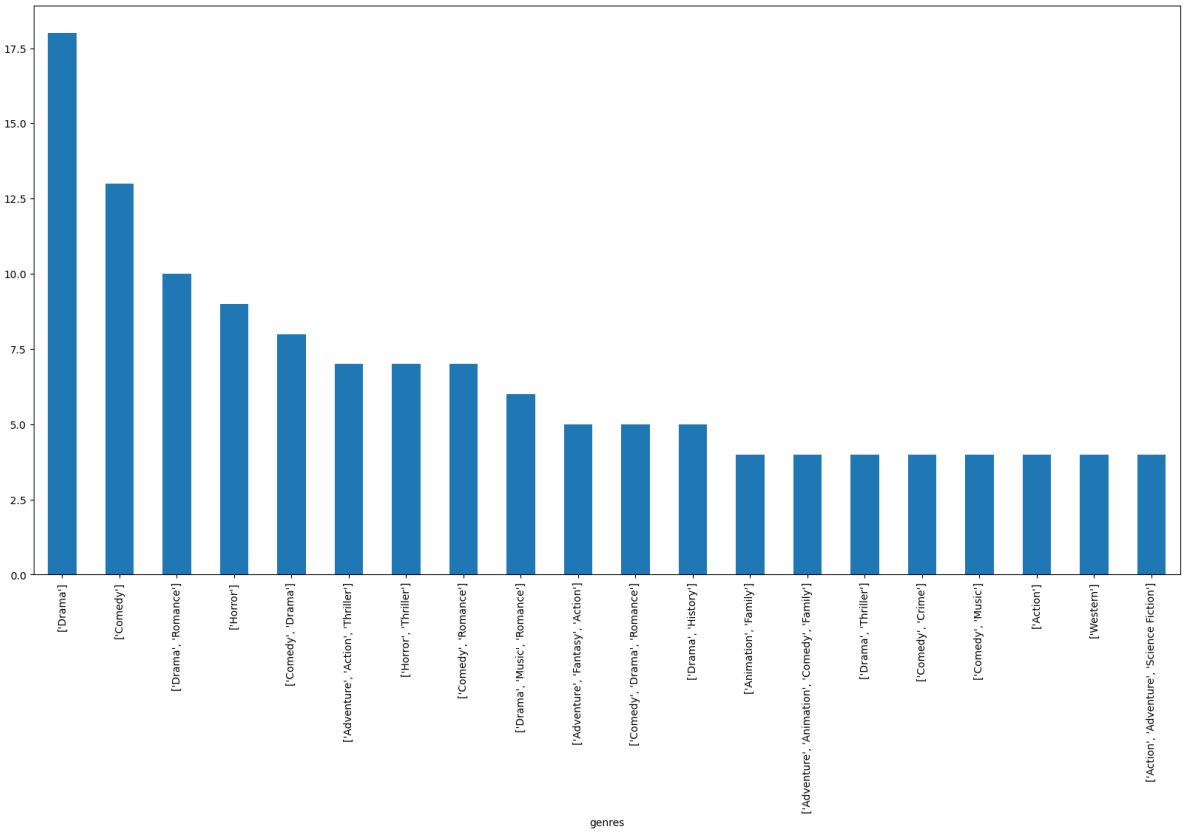
Most popular movie:

In the bar graph below, we can visualize the top 20 popular movies.
The most popular one is Minions



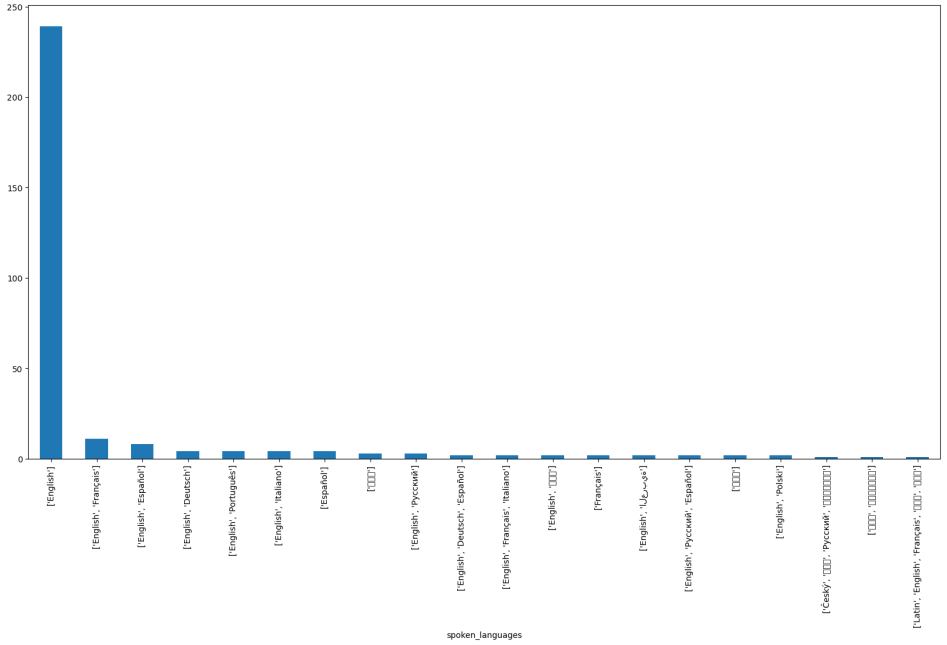
Most successful genre:

We visualize that the most successful genre was Drama because as we see in this bar graph, the majority of movies are related to this genre.



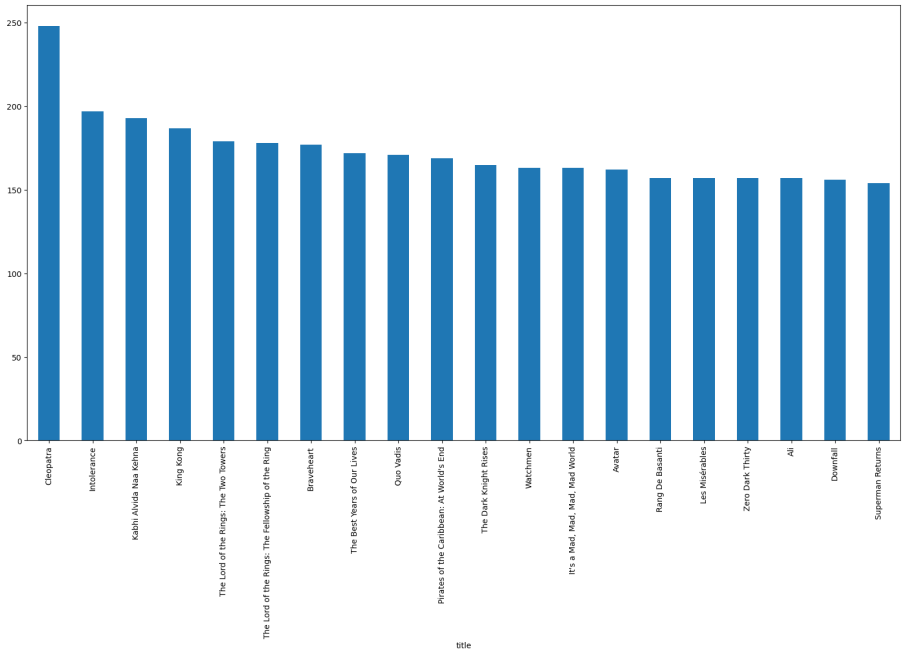
Most spoken languages:

In the bar graph below we can visualize that almost every movie is spoken in English.



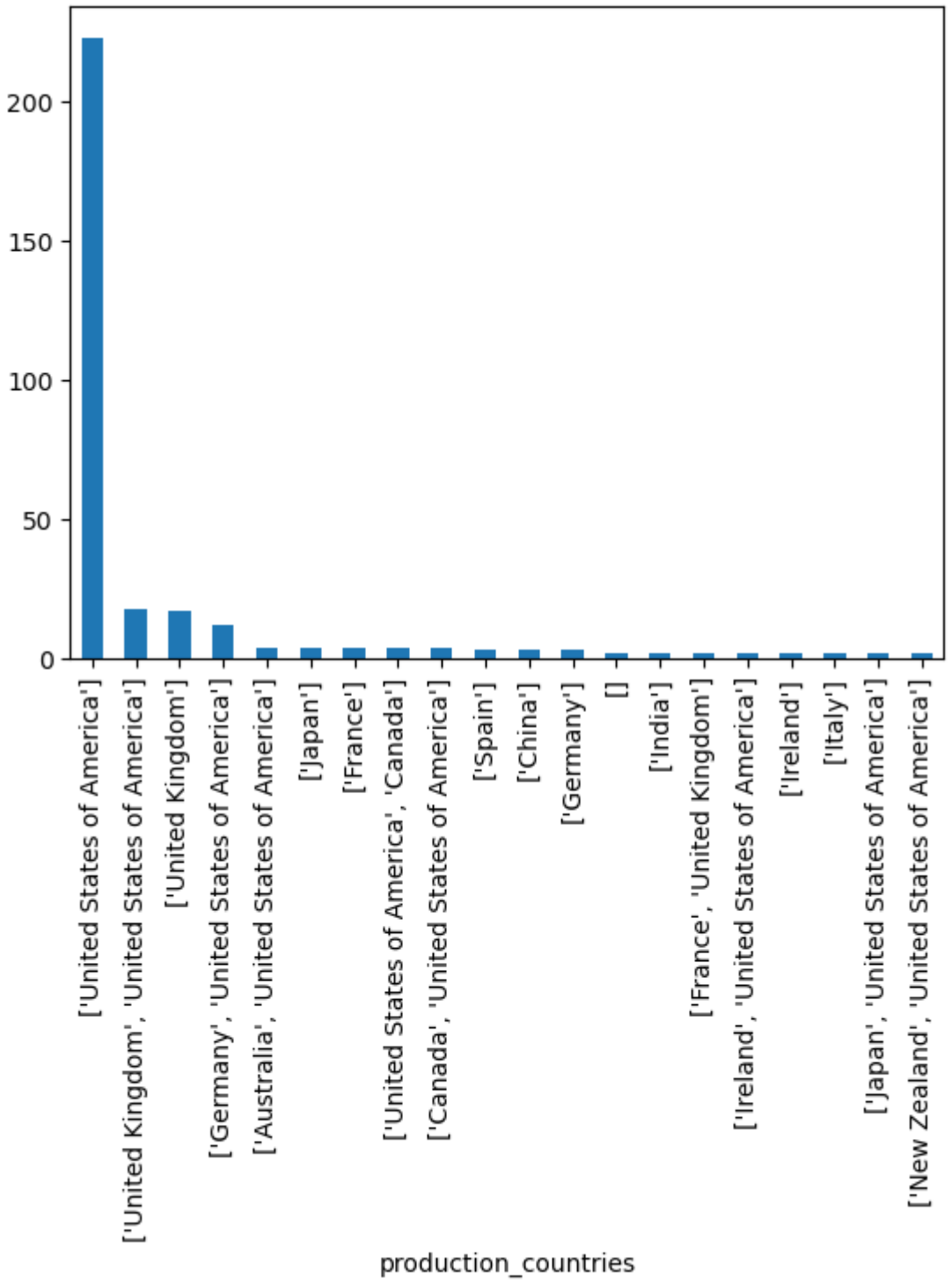
Longest runtime movies:

In the bar graph below, we visualize that the longest movie is Cleopatra with a runtime of 250 minutes.



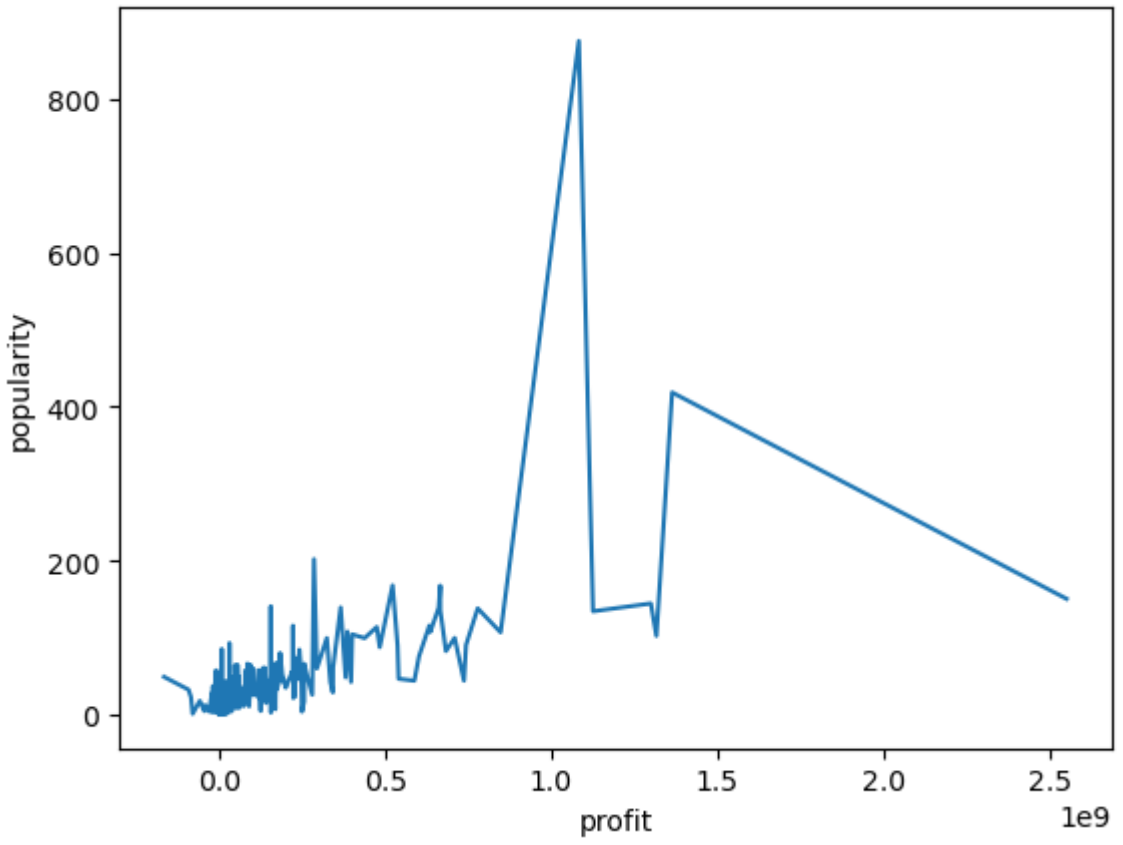
Production country movies:

In the bar graph below, we can see that the majority of movies were produced in the United States.



Relationship between profit and popularity:

In the graph below, can see the relationship between the popularity and the profit of each movie. Interesting to notice that the more popular movies didn`t make more profit.



THIS REPORT WAS WRITTEN BY: GAETANO LOPEZ
DATE: 23/12/2022