Investigate a dataset – IMDB movies

Dataset

I analyzed the IMDB movies dataset

Questions asked:

Which year had the highest number of movies released?
Which year had the highest budget?
Which genres occur the most over the years?
Which actor was most casted?
Is there any relationship between budget and revenue
Is there any relationship between budget and runtime?
Is there any correlation between popularity and runtime?

Steps taken to solve them

I split the genres and the actors into a list and counted each one of them using a dictionary to know the frequently occurred ones.
I used a bar plot to show adequately the most occurring genres and the most casted actors.
I plotted the relationships between some attributes using the matplotlib scatterplot.

Data Wrangling and cleaning

I had to remove duplicates rows.
I removed rows with null values including rows with a budget value of zero or a revenue value of zero.
I removed columns that were not needed for the analysis.
I created two new data frames for the actors and genres to properly visualize them