Database TMDb movie data report

The used dataset is TMDB movies which have imported 4 libraries: pandas, numpy, matplotlib.pyplot, and seaboarn.

The reason why I choose this dataset?

I choose this data over the rest because of the variety of data itself. Which could let me investigate more widely than the others during the analysis phase.
So, the codes start with defining how many rows and columns we have in the data set.
And to statists analysis I describe it to know more about my dataset.

What is the suggestion question?

- does the budget affect the popularity of real movies?
- What are the top 5 movies with the highest popularity?
- which value has the highest/lowest overall value?

Questions investigation description

1- Does the budget affect the popularity of real movies?

The answer is YES.
As shown in the histogram graph, I can conclude that the more budget spent on the movies, then the more revenue would return. So, they are equal. Once the budget grows. The popularity of the movies would grow too.

2- What are the top 5 movies with the highest popularity?

Before I go to answer this question, I sort the dataset values to make the search easier.
The most popular movies we got is:
1- khosla ka ghosla!
2- Mon petit doigt m'a dit...

3- G.B.F.
4- The Hospital
5- North and south Book1

And I used here the barplot from the seaborn library to make the visualization clear.

3- which value has the highest/lowest overall value?

1- The budget and revenue column has a **strong positive correlation coefficient relationship**, meaning that more budget spending on the movie creation steps would also make the revenue high.
2- Now the popularity and release year **have uncorrelated positive relationships**. since the correlation coefficient of their relationship is near 0.
So, that means the release year doesn't affect the popularity of the films. And vice versa.
3- Between the revenue_adj and release year. is an **uncorrelated negative relationship** because it is minus nearest to 0.
And that means it is in an inverse relationship.

Data wrangling

First. The wrangling starts with dropping the column that is not needed or may it be not useful/unusable and full of missing values that cannot be modified.
So, I dropped into the first group of dropping 5 columns.
Get to check if the dataset is still cleared from null values. which again I dropped another 2 columns.
Also, I renamed a column to get the work easier which I take the idea from Kaggle when I start reading about the dataset.
Checking for duplication is necessary too, the TEKKEN row has a duplication and I dropped one of them.
To get check again about the dataset size it gets (10865, 14) since it was at first (10866, 21).

Conclusion

I will abbreviation conclusion in points:

1- The Average Movies Profit increased over time.
2- There is a Positive relationship between popularity and budget.
3- popularity and revenue have a strong correlation, which means the more the film becomes popular. More revenue will come.
4- release year and runtime have an Inverse relationship. So, we can say that over the years. People still want to watch a long movie runtime without being bored while watching.

Limitations

1- there is a lot of missing data. even after the cleaning phase, there are still a lot of NAN values, which means that the dataset was still insufficient.

2- since this dataset is rich in information. I had a lot of obstacles during the cleaning phase. because even if I filled in that missing data. the plotting information goes wrong. so I had to take different ways to make a correct investigation.