THIS IS THE PDF FILE THAT DOCUMENTS MY DATA ANALYSIS PROCESS

HERE IS THE DATASET I USED:

(https://video.udacity-data.com/topher/2022/November/6375a7af_tmdb-movies.csv/tmdbmovies.csv.zip)
(THE MOVIES ONE)

FIRST

I explored the data and noticed that there is an important column missing so I added it, it was the profit column

SECOND

I put the questions that I wanted to answer

1- I wanted to calculate and visualize the correlation between the profit and the popularity (which I think may help the business to see how much the marketing will help their next movie)
2- I wanted to see how well and trending a specific genre was in the past few years Then I visualized the outcome
3- I wanted to have the ability within my code to visualize the relation between any column and the profit column

NOTE:during all this process I have cleaned the data by replacing some nans with empty string which I thought it wouldn't matter in the analysis

REFRENCES: there is about 40% percent of my code from stack overflow this was my only recourse to find a solution for the proplems I met (mostly the matplot and the seaborn)

CONCLUSION:

I answered all the questions that are listed above and I made some good insights from it ,for example I noticed that the drama genre has not been succeeding in the in the current time contrasting the comedy who has been in a good shape

Even tho I learned a lot from this data but it appears not to be perfect from my perspective here is why:

First, The column [run time] was hard for me to understand and till now I still don't understand the point of it

Second, The column home page had too much nan values so I choose to ignore it and not use it in my analysis