PART 1

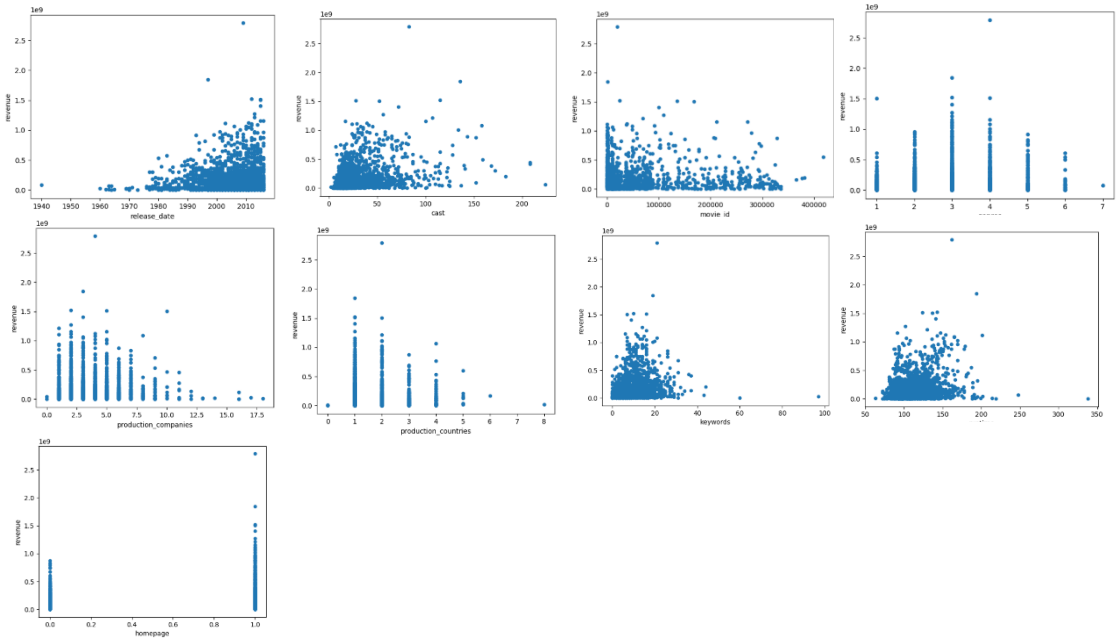
In the beginning, I plan to build a simple regression model, so I choose to develop a Linear Regression model. To create that model, I clean my data by

- converting all JSON formatted data to number (e.g. cast, crew)
- transferring release data to release year
- labelling no homepage to 0 and has a homepage to 1
- dropping text data (e.g. overview)

Then, I get an inaccurate model with lots of negative revenues.

After that, I try to improve my model. I compare four regression model (Linear Regression, Logistic Regression, Decision Tree and Random Forest) by putting data in them, and I find Random Forest has the best performance. So I choose Random Forest as my model for part one.

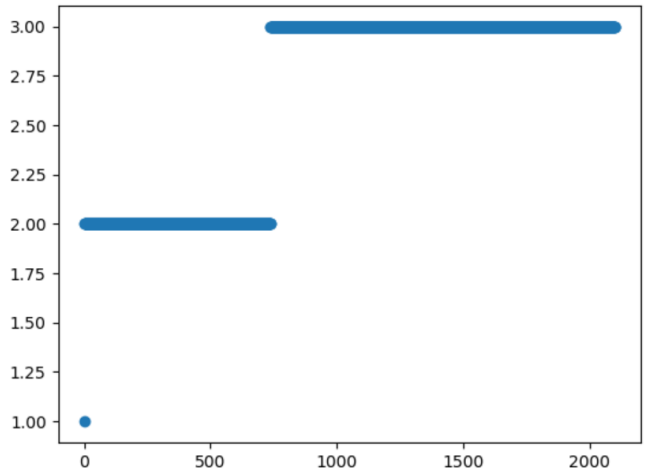
To do better data clean, I draw several scatter diagrams between revenue and other features.



From the graphs, we can find that some features have apparent relations with revenue, but some features like cast and crew do not have an evident relationship with revenue. I choose to drop those columns, who has no apparent relation with the revenue. There also has some features like genres, does not have an evident relationship with revenues on the graph, but they have connections with revenue in logic. However, there are so many genres for each movie, which is hard to clean them. Finally, I decide to define the first genre to be the only genre of the movie and give different genre different value to distinguish between them. Then, I finish my model.

PART 2

In the beginning, I draw a scatter graph to show the layout of rating in the training dataset. And it shows that there are only two values (2 and 3) for rating.



So I try KNN model at first. Then I begin to do data clean.

For Cast, I try to get all actor's and actress's names, because a cast sometimes can influence the rating. But there are so many different actor and actress, so I change to count the number of casts.

For Crew, because the number of crews sometimes can decide how well a movie be produced, so I count the number of crews.

For Genres, I give different numbers for different genres.

For Homepage, I give 0 for no homepage and 1 for has a homepage.

For Keywords, as keywords have so many types, which is hard and useless to clean, so I drop it.

For Original_language, Original_title, Spoken_languages, Tagline and Overview, I think they all useless and difficult to clean, so I drop them.

For Production companies and Production countries, I think the number of collaborators can reflect the quantity of a movie, so I count the number of them.

For Release date, I leave the year numbers.

For Runtime, I scale it to 1 and 0 to make the prediction more accurate.