1. I determine which features to drop and it seems like some features are useless such as movie_id(useless), keywords(too specific), homepage(useless), original_title(useless), overview(useless), tagline(useless) and status(everything is released).
2. Create a list of features based on the cast, crew, production countries, production companies, genres, and spoken language since they are a list.
3. For the cast, crew, production countries, production, companies, and spoken language, I set a variable that determines which of them to include based on how many times they appear on the dataset. Different variables will result in different amount of features.
4. Repeat the steps for the validation dataset by using the same vectorizer and list of features from step 2.

I tried using several methods to preprocess the dataframe but I cannot seem to find the proper one since I kept on getting correlation = 1. I also tried tweaking the variable in step 3 to no avail.

After asking my friends and brother as well as researching and testing, I decided to use StandardSampler and RCA. For q1, I use Linear Regression and q2, gradient boosting.

Result:

Q1.

MSE = 7060649366606446

Correlation = 0.61

Q2.

Average precision = 0.59

Average Recall = 0.54

Accuracy = 0.69

I would say that I improved my result since at first I kept getting correlation = 1. I also keep on improving the correlation and MSE by tweaking the variables on step 3.

Problems:

There are several challenges that I face during the assignment. The first challenge is preparing the data and features. I was confused on how I should approach this but in the end after several tries (most of them results in either overfit or underfit), I finally decide to only take the data that appears several times. This results in a significant decrease in the amount of features.

The second challenge is finding the proper method to preprocess the dataframe and finding the proper model. I tried several feature selections such as SelectKBest and RFE but it doesn't seem to work. Then only after consulting with my friends and brother I decided to use StandardSampler and RCA. I also tried several models for Q2 such as gradient boosting and random classifier. In the end I decided to use gradient boosting since it provides higher precision, recall and accuracy.