Wrap-up of the EDA meeting

1. Delete the legend of question5 (no colors)
2. Recall that the dataset only contains 1000 movies with the highest domestic sales, it can’t represent anything of the Hollywood movie industry
3. For question8, each data point represents the average sales of movie with a certain run time, but the number of movies with that run time may differ greatly, so the graph may not be so indicative.
4. For final project: generate a predictive model for domestic sales and international sales with all the variable we have and observe the difference between the two models
5. For final project: calculate the proportion of each movie’s domestic sales over total sales and create a confidence interval with the data. Then observe if the confidence intervals overlap for different genres/ distributors