Springboard Machine Learning Exercise

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?
    a. My main question is one of classification; given the information available before the release of a movie (genre, movies in theaters at the same time, rating, planned release date, production budget, etc.), can a model accurately predict whether the film will be a commercial success or not? Which of the above factors contribute to the prediction? As a problem of classification, I will most likely use a logistic regression to understand the influence of the different factors on the probability that the film is a commercial success.
    b. After investigating how the different factors contribute to predictions on the entire dataset, it might be useful to cluster the movies and create separate models for each cluster.
2. What are the main features (also called independent variables or predictors) that you'll use?
    a. Is_comm_success - This is a binary categorical variable denoting that a given film exceeds a threshold of return and is classified as commercially successful.
    b. Rel_date_TH - The date of theatrical release for the film. This is mostly of interest as it pertains to seasonality (i.e. whether or not a film is slated for a “prime” window during the year). Is_rel_prime is a binary classification of whether or not a film was released in a “prime” window.
    c. Movies_TH - The average amount of time Americans report spending watching movies in theaters for the month of the release of the film (from the BLS American Time Use Survey).
    d. Infl_prod_budget - The production budget for the film, adjusted for inflation (using CPI data from the St. Louis Federal Reserve.
    e. MPAA_rating - The Motion Picture Association of America rating for the film
    f. Creative_type - A categorical variable denoting a classification about the film with respect to the origin of the story (e.g. Contemporary fiction, Science fiction, Historical fiction, etc.)
    g. Source - The original source material for the film (e.g. Original Screenplay, Based on TV, Spin-off, etc.)
3. Which machine learning technique will you use?
    a. Given the nature of the problem I will likely use logistic regression for creating the model for all the films in the dataset.
    b. Given how many films and features I would use for clustering, I will use k-means clustering.

4. How will you evaluate the success of your machine learning technique? What metric will you use?
    a. Since the case I will be trying to make is that certain features are not predictive of commercial success, my evaluation of the model will hinge on whether not there is a statistically significant relationship between the features and the outcome. Thus, the primary metric will be the p-value for the different features.