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?

Framing the main question of my capstone project as a machine learning problem would simply be something like: If a game gets reviewed with an average metascore of X, what kinds of sales can we predict it to have in each of three worldwide territories and overall? There are more specific questions we could answer as well including the minimum metascore a game needs to reach a certain sales figure or which review score averages for a certain genre or from a certain publisher are needed in order to reach certain sales figures. All of these questions, though not without much more analysis, would be of use to a game publisher and developer.

The machine learning problem here would definitely be supervised, though some grouping or sorting of genre or publisher may be useful for further analysis down the road. That said, I'm looking most likely at a linear regression problem, though I might need a more complicated analysis type if I want to include genre and publisher among other variables in my machine learning algorithm.

2. What are the main features (also called independent variables or predictors) that you'll use?

The features I will make the most use of are the review score averages. I need to find some way to weight the averages based on the amount of reviews provided, but that will be the key feature for prediction of sales figures. If possible, I'll take into account user scores as well, but there is likely to be much more noise in the dataset there (possibly so much so that it wouldn't be worth using that data). Also, if possible, I'll make use of genre and publisher as features for predicting sales numbers.

3. Which machine learning technique will you use?

At least preliminarily, I'll use linear regression. The problem with lm is that it's only useful regarding numbers. I know more intricate machine learning techniques exist, but I might have to do some digging to find one that I can use with some categorical variables within my R dataset.

4. How will you evaluate the success of your machine learning technique? What metric will you use?

So far, I'm thinking I will use K-fold cross validation as an evaluation method to test my model's effectiveness. By splitting my dataset randomly and testing it repeatedly with randomly split datasets cut from my main dataset, I'll be able to tell if my model is predicting as accurately. If I get positive results repeatedly, it should mean the model is working properly. If I'm unsure after this model evaluation method, I can further select a different evaluation method to compare this one against, but currently, I believe my dataset to be large enough that I shouldn't need another method.