Homework 5 Julia Fish

Task 1: Conceptual Questions

- Question 1: **What is the purpose of using cross-validation when fitting a random forest model?**
  Cross-validation helps estimate a model's performance with "new" data. When fitting a random forest model, cross-validation ensures that the model is not overfitting by evaluating its performance across multiple data splits (i.e. leaving one fold out at a time). Because of this, it gives a more reliable estimate of prediction error than a single train-test split.
- Question 2: **Describe the bagged tree algorithm.**
  The baged tree algorithm involves generating multiple training sets using bootstrapping techniques. A decision tree is trained on each of the bootstrap resamples, and predictions are averaged over. This reduces variance and increases model stability compared to fitting a single tree model.
- Question 3: **What is meant by a general linear model?**
  A general linear model is a linear model that can have different forms of response variable than just all real numbers. For example, binomial, poisson, gamma, etc. can be used in general linear models to help modify the response range as well as oher logistical factors in the model fit.
- Question 4: **When fitting a multiple linear regression model, what does adding an interaction term do?**
  Adding an interaction term to a MLR model allows the effect of one variable on the response to depend on the level of another variable. That is, the variables are allowed to communicate with one another to predict a response value instead of only having just additive effects.
- Question 5: **Why do we split our data into a training and test set?**
  Using a training and test set allows us to see model performance on "unseen" data with known response values. That way we can get a better idea of whether or not our model performs well or has just been overfit to the data it was given to train the model.