Authors:

- Name: Mhd Jawad Al Rahwanji - MatrikelNummer: 7038980 - Email: mhal00002@stud.uni-saarland.de
- Name: Christian Singer - MatrikelNummer: 7039059 - Email: chsi00002@stud.uni-saarland.de

Exercise 3.4 - Validation set and Cross Validation

a) The advantage cross validation (CV) has over a holdout set is allowing the model to train on multiple train-test splits. That gives us a better idea about the performance by utilizing all of the available data. That is particularly useful when the available dataset is small. Also, with CV we eliminate split dependency resulting in a more realistic performance score for the model. But bear in mind that CV is computationally more expensive.

b) The process would go as follows:

- For each hyperparameter (order of the polynomial = 1, 5, 9), we train linear regression using the augmented features with a *k*-fold cross validation (CV) (*k* = 5).
- Each time (for example, order = 1) we train 5 linear regressions each trained on different train-test splits. The 5 models are scored (using MSE for example) and then their scores are averaged to one final score for the current parameter.
- As for the means of comparison, we sort the final scores achieved by the different hyperparameters (the averaged scores of the 5 linear regressions we trained).
- We pick the hyperparameter with the best overall final score.

c) Both runs will score 35. In LOOCV, we train *n* models (*n* = # of observations). Each iteration we train on *n-1* observations and test on the remaining observation. As a result, no matter how we shuffle the dataset we will end up with the same final score.