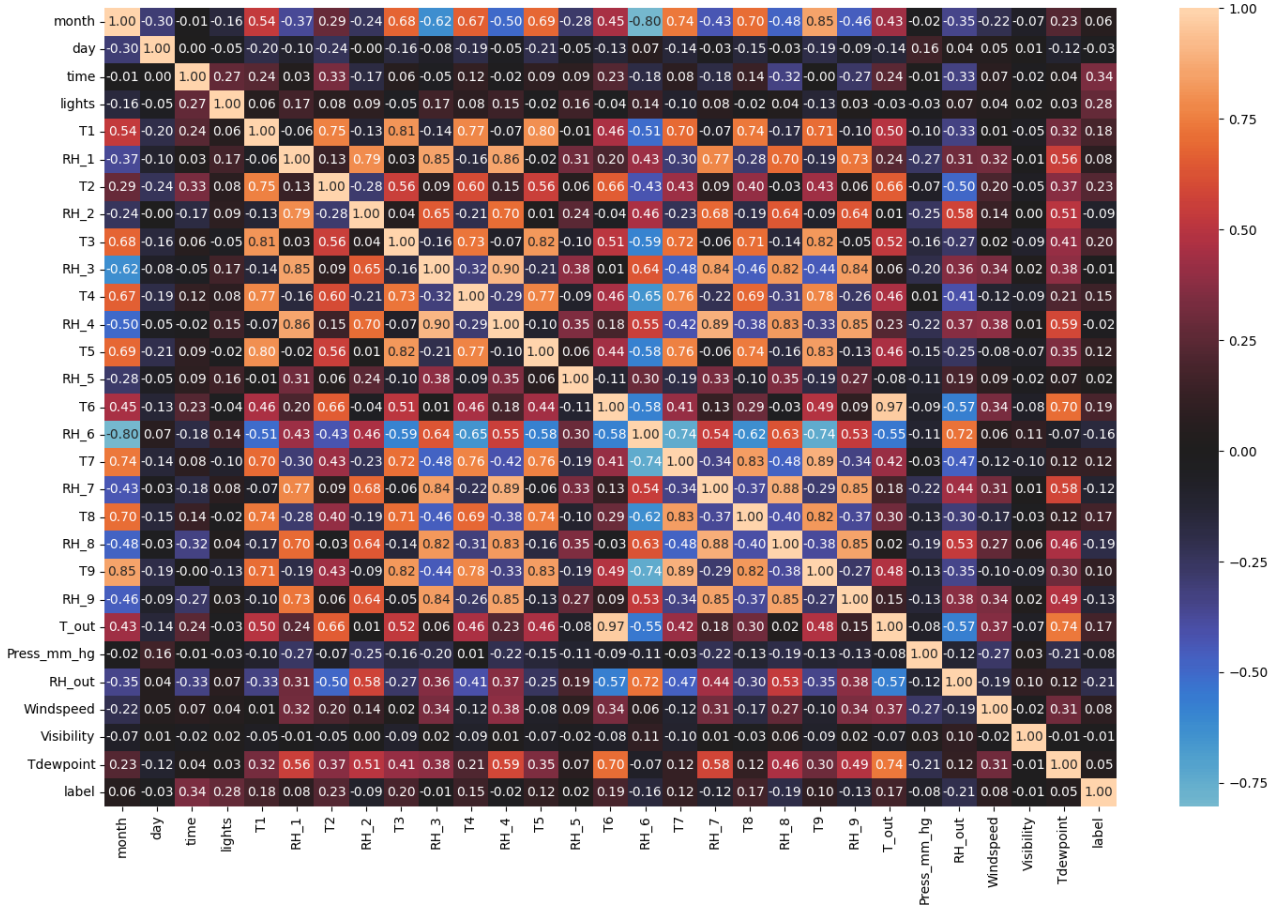
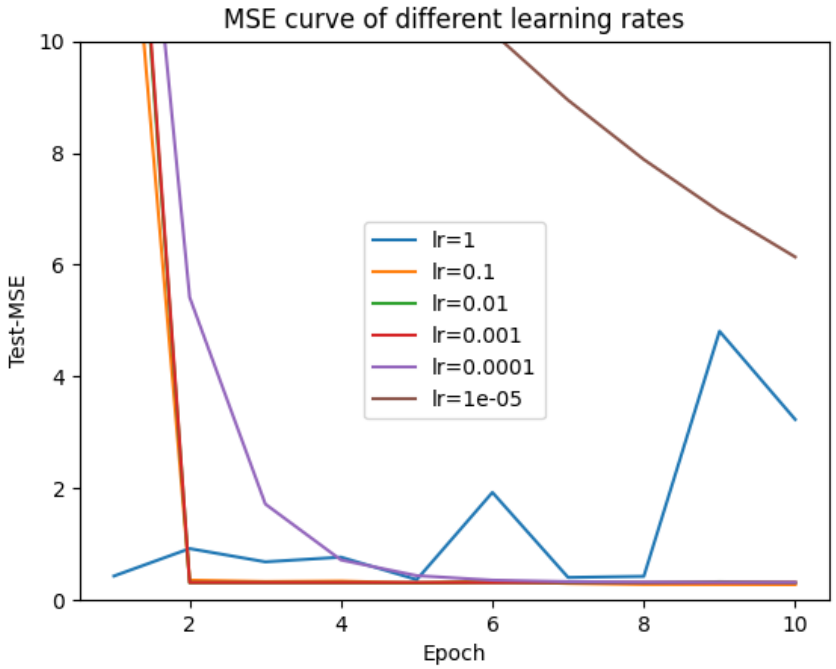
Q1
b)



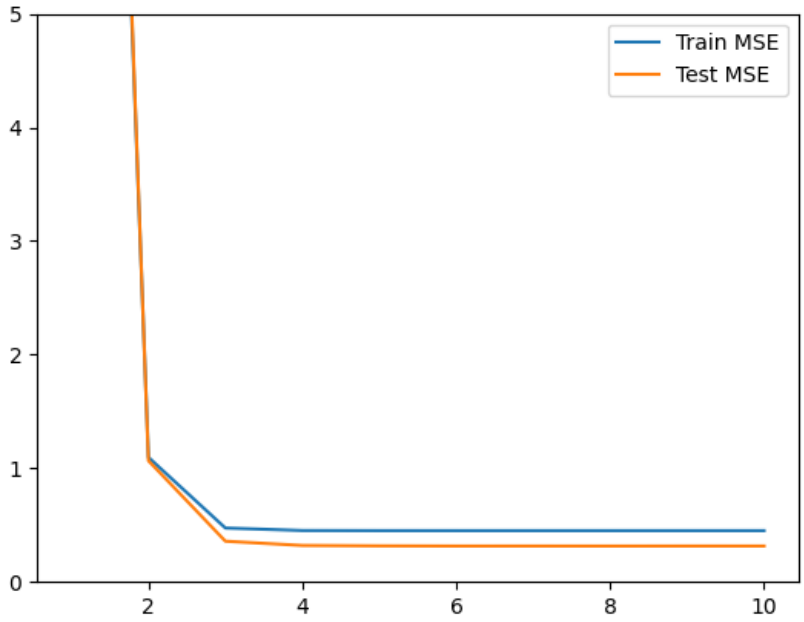
c) For each two features that have a high absolute Pearson correlation (>0.8), I dropped one of them. Since they are highly correlated, they provide similar information to the prediction process. Thus, training the model on both of them is redundant.

Q3
b)



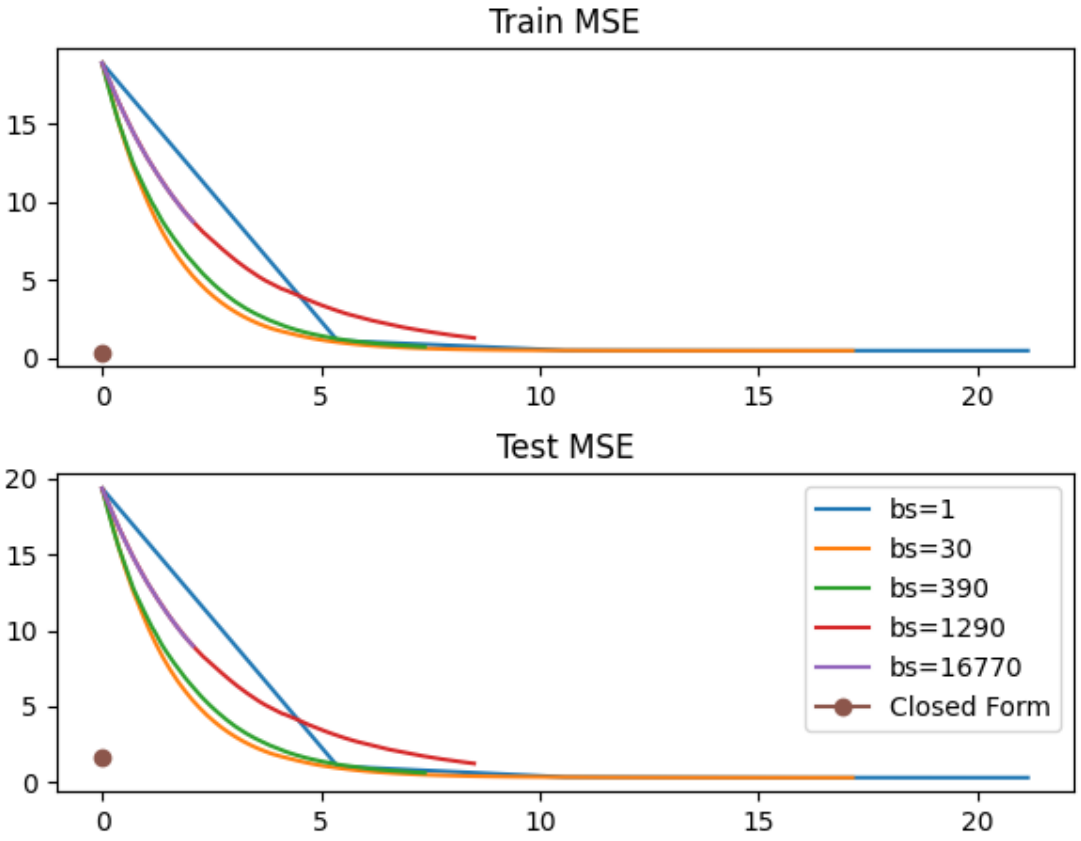
Based on the graph, lr=0.0001 seems to be the optimal learning rate, since the MSE curve does not have a sudden drop and then stop decreasing, and converges in an efficient rate.

Training the model on the whole training dataset with a learning rate of 0.0001:



Q4

a)



b) Based on the graph, increasing batch size would decrease training and testing time, but may converge on a higher MSE. As the batch size gets larger, the converged MSE is getting closer to closed form and even become lower.