1. B
2. B
3. B, Gradient Descent
4. C
5. D
6. A
7. A
8. C
9. A, B, & C
10. B
11. C
12. We could use batch gradient descent, stochastic gradient descent, or mini-batch gradient descent. SGD and MBGD would work the best because neither of them need to load the entire dataset into memory in order to take 1 step of gradient descent. Batch would be ok with the caveat that we have enough memory to load all the data.

    The normal equations method would not be a good choice because it is computationally inefficient. The main cause of the computational complexity comes from inverse operation on an (n x n) matrix.

13. Feature scaling is required for the various gradient descent algorithms. Feature scaling will help gradient descent converge quicker.

    The normal equations method does not require normalizing the features, so it remains unaffected by features in the training set having very different scales.