Exercise: 1

$\hat{y} = w^T x$
; where x is a feature vector of size (3x1) and w is a parameter vector of (3x1) (a parameter for each feature)

$MSE_{train} = \frac{1}{m} || \hat{y}_{train} - y_{train} ||_2^2$ the performance metric

And the design matrix X (the matrix representing the features (columns) and and data points (rows))

Here will say the design matrix X is (5, 3); 5 data points and 3 features

MSE where gradient is 0: $\nabla_w \frac{1}{m} || \hat{y}_{train} - y_{train} ||_2^2 = 0$; where $\nabla_w = \frac{\partial MSE_{train}}{\partial w}$

By substituting $\hat{y} = X_{train} W$ one gets $\frac{1}{m} \nabla_w || X_{train} w - y_{train} ||_2^2 = 0$

Here note that $\hat{y}$ is a (5x1) vector. Scalar prediction for each data point

Taking the inner product, $\nabla_w (X_{train} w - y_{train})^T (X_{train} w - y_{train}) = 0$

By doing the expansion of the equation one gets:

$\nabla_w ((X_{train} w)^T X_{train} w - (X_{train} w)^T y_{train} - y_{train}^T X_{train} w + y_{train}^T y_{train}) = 0$

Since it does not matter the order as long as the dimensions agree (its a vector product)
So we can simplify the equation as

$\nabla_w (w^T X_{train}^T X_{train} w - 2(X_{train} w)^T y_{train} + y_{train}^T y_{train}) = 0$

Taking the partial derivative of this equation one gets

$2 X_{train}^T X_{train} w - 2 X_{train}^T y_{train} = 0$

Give that $X_{train}^T X_{train}$ is not a singular matrix (Invertible) we can find the parameters w that minimises the MSE as

$w = (X_{train}^T X_{train})^{-1} X_{train}^T y_{train}$; This equation is called the normal equation