Question 1

(a) We minimize the RSS function with respect to w_0 and w_1

$$\frac{\vartheta RSS(w_0, w_1)}{\vartheta w_0} = 0 \Rightarrow -2 \sum_{n=1}^{N} (y_n - w_0 - w_1 x_n) = 0$$

$$\Rightarrow N w_0 + \left(\sum_{n=1}^{N} x_n \right) w_1 = \sum_{n=1}^{N} y_n \quad (1)$$

$$\frac{\vartheta RSS(w_0, w_1)}{\vartheta w_1} = 0 \Rightarrow -2 \sum_{n=1}^{N} x_n (y_n - w_0 - w_1 x_n) = 0$$

$$\Rightarrow \left(\sum_{n=1}^{N} x_n \right) w_0 + \left(\sum_{n=1}^{N} x_{n1}^2 \right) w_1 = \sum_{n=1}^{N} x_n y_n \quad (2)$$

Combining (1) and (2) we get the 2×2 system of equations

$$\begin{bmatrix} N & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n^2 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} N & \sum_{n=1}^{N} y_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} x_n y_n \end{bmatrix}$$

We solve the above system using the determinants and we get:

$$w_1^* = \frac{\sum_{n=1}^{N} x_n y_n - N \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right) \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right)}{\sum_{n=1}^{N} x_n^2 - N \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right)^2}$$

$$w_0^* = \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right) - w_1 \left(\frac{1}{N} \sum_{n=1}^{N} x_n \right)$$

(b) Starting from the given expressions we get:

$$
\begin{aligned}
w_1^* &= \frac{\sum\limits_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum\limits_{n=1}^{N} (x_n - \bar{x})^2} \\
&= \frac{\sum\limits_{n=1}^{N} x_n y_n - \bar{y} \sum\limits_{n=1}^{N} x_n - \bar{x} \sum\limits_{n=1}^{N} y_n + N\bar{x}\bar{y}}{\sum\limits_{n=1}^{N} x_n^2 - 2\bar{x} \sum\limits_{n=1}^{N} x_n + N\bar{x}^2} \\
&= \frac{\sum\limits_{n=1}^{N} x_n y_n - \frac{1}{N} \sum\limits_{n=1}^{N} y_n \sum\limits_{n=1}^{N} x_n - \frac{1}{N} \sum\limits_{n=1}^{N} x_n \sum\limits_{n=1}^{N} y_n + N\frac{1}{N} \sum\limits_{n=1}^{N} x_n \frac{1}{N} \sum\limits_{n=1}^{N} y_n}{\sum\limits_{n=1}^{N} x_n^2 - 2\frac{1}{N} \sum\limits_{n=1}^{N} x_n \sum\limits_{n=1}^{N} x_n + N \left(\frac{1}{N} \sum\limits_{n=1}^{N} x_n \right)^2} \\
&= \frac{\sum\limits_{n=1}^{N} x_n y_n - N \left(\frac{1}{N} \sum\limits_{n=1}^{N} x_n \right) \left(\frac{1}{N} \sum\limits_{n=1}^{N} y_n \right)}{\sum\limits_{n=1}^{N} x_n^2 - N \left(\frac{1}{N} \sum\limits_{n=1}^{N} x_n \right)^2}
\end{aligned}
$$

It is straightforward to show for w_1^*.

(c) The weight w_1^*, which is the slope of the linear regression line, is proportional to the co-variance between the input feature and the outcome. The weight w_0^*, which is the bias term of the linear regression line, is equivalent to the expected value of the expression $Y - w_1 X$, if we assume that X and Y are probabilistic distributions, i.e. $w_0^* = \mathbb{E}(Y - w_0^* X) = \mathbb{E}(Y) - w_0^* \mathbb{E}(X)$

Question 2

(a) In the given Taylor series expansion, we substitute $f := J$, $\mathbf{x} := \mathbf{w}$ and $\mathbf{x_0} : \mathbf{w}(k)$ and we get the desired equation:

$$
J(\mathbf{w}) \approx J(\mathbf{w}(k)) + (\nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot (\mathbf{w} - \mathbf{w}(k)) + \frac{1}{2}(\mathbf{w} - \mathbf{w}(k))^T \cdot \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \cdot (\mathbf{w} - \mathbf{w}(k)) \tag{1}
$$

(b) Substituting $\mathbf{w} := \mathbf{w}(k+1)$ to equation (1), we get:

$$
\begin{aligned}
J(\mathbf{w}(k+1)) \approx & J(\mathbf{w}(k)) + (\nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot (\mathbf{w}(k+1) - \mathbf{w}(k)) + \\
& \frac{1}{2}(\mathbf{w}(k+1) - \mathbf{w}(k))^T \cdot \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \cdot (\mathbf{w}(k+1) - \mathbf{w}(k))
\end{aligned} \tag{2}
$$

Substituting $\mathbf{w}(k+1) - \mathbf{w}(k) = -\alpha(k) \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)}$ (from the gradient descent update) to equation (2), we get:

$$
\begin{aligned}
J(\mathbf{w}(k+1)) \approx & J(\mathbf{w}(k)) - (\nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot \alpha(k) \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)} + \\
& \frac{1}{2}(\alpha(k) \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \cdot (\alpha(k) \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)}) \\
= & J(\mathbf{w}(k)) - \alpha(k) \|\nabla J|_{\mathbf{w}=\mathbf{w}(k)}\|_2^2 + \frac{1}{2}(\alpha(k))^2 (\nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)}
\end{aligned} \tag{3}
$$

(c) We minimize (3) wrt $\alpha(k)$

$$\frac{\vartheta J(\mathbf{w}(k+1))}{\vartheta \alpha(k)} = -\|\nabla J|_{\mathbf{w}=\mathbf{w}(k)}\|_2^2 + \alpha(k)(\nabla J|_{\mathbf{w}=\mathbf{w}(k)})^T \cdot \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \cdot \nabla J|_{\mathbf{w}=\mathbf{w}(k)} = 0$$

$$\Rightarrow \alpha(k) = \frac{\left\|\nabla J|_{\mathbf{w}=\mathbf{w}(k)}\right\|_2^2}{\left(\nabla J|_{\mathbf{w}=\mathbf{w}(k)}\right)^T \mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \left(\nabla J|_{\mathbf{w}=\mathbf{w}(k)}\right)}$$

(d) The most expensive operations are the ones in the denominator involving multiplications between a D-dimensional vector and a $D \times D$ dimensional matrix, i.e. $\nabla J|_{\mathbf{w}=\mathbf{w}(k)} \in \mathbb{R}^D$ and $\mathbf{H}_J|_{\mathbf{w}=\mathbf{w}(k)} \in \mathbb{R}^{D\times D}$. The cost of a vector by matrix multiplication is $O(D^2)$, therefore the cost of computing $\alpha(k)$ in each iteration is $O(D^2)$.

HW1 Q3

a. Data Exploration

It's not obvious to see whether a variable is continuous or categorical. Go back to the definitions of these features:

- **X** and **Y** are continuous because they are coordinate values.
- **month** and **day** are categorical.
- The other features are all continuous variables.

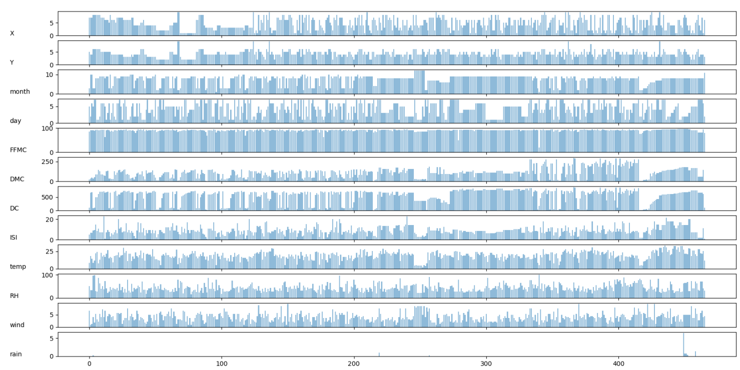

Figure 1: Plots of raw data.

b. KNN

```
# KNN implementation with Python
def KNN(n, x_train, y_train, testcase):
    dist = []
    for i in x_train:
        dist+=[np.linalg.norm(testcase-i)]
    neighbours = [y_train[x] for x in np.argpartition(dist, n)[:n]]
    return 1 if sum(neighbours)*2 >= n else 0
```

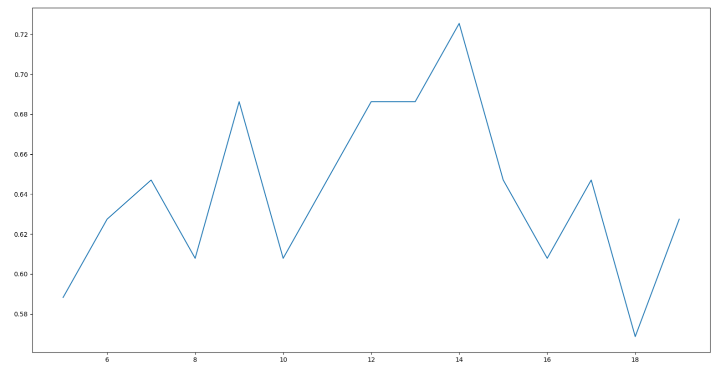

Figure 2: Cross-validation accuracy with different K.

c. Linear Regression

The log function can significantly reduce the impact of outliers as shown in Fig.3.

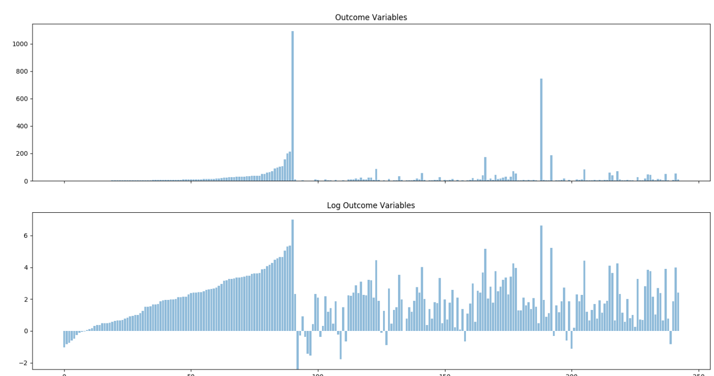

Figure 3: Outcome variables and log outcome variables.

The Pearson correlation of the linear regression and true values is 0.1652, which means there's weak correlation between these two sets of values.