Tutorial 2

1. a. It can be observed that, for data points that have an L_2 length of 1, as the cosine similarity increases, the Euclidean distance will decrease correspondingly. In particular, when the cosine similarity is 1, the corresponding Euclidean distance is 0. On the other hand, when the cosine similarity is 0, the corresponding Euclidean distance is about 1.4.

   b. Let **u** and **v** be two vectors where each vector has an L_2 length of 1. The relationship between the Euclidean distance and cosine similarity is derived as follows:

$$\begin{aligned} d(\mathbf{u},\mathbf{v}) &= \sqrt{\sum_{k=1}^{n}(u_k - v_k)^2} \\ &= \sqrt{\sum_{k=1}^{n}(u_k^2 - 2u_k v_k + v_k^2)} \\ &= \sqrt{1 - 2\cos(\mathbf{u},\mathbf{v}) + 1} \\ &= \sqrt{2(1-\cos(\mathbf{u},\mathbf{v}))} \end{aligned}$$

2. a. Suppose we denote the first attribute as x, and the second attribute as y.

$$\bar{x} = \frac{2+9+7+5}{4} = 5.75, \bar{y} = \frac{19+6+15+12}{4} = 13$$

$$\sigma_x^2 = \frac{1}{3}[(2-5.75)^2 + (9-5.75)^2 + (7-5.75)^2 + (5-5.75)^2] = 0.897$$

$$\sigma_y^2 = \frac{1}{3}[(19-13)^2 + (6-13)^2 + (15-13)^2 + (12-13)^2] = 30$$

$$\begin{aligned} covariance(x,y) = \frac{1}{3}&[(2-5.75)(19-13) + (9-5.75)(6-13) \\ &+(7-5.75)(15-13) + (5-5.75)(12-13)] \\ =& -14 \end{aligned}$$

The covariance matrix is given by $\begin{bmatrix} 8.917 & -14 \\ -14 & 30 \end{bmatrix}$