Question 2. Chapter 2, Exercise 18

Given two vectors

x	0	1	0	1	0	1	0	0	0	1
y	0	1	0	0	0	1	1	0	0	0

Question a: Answer

Hamming distance is the number of bits that are different between two binary vectors (red color) as depicted in red color in Table 1 above. Thus, $d(x,y) = F10 + F01 = 3$

Jacaard similarity is a measure of the similarity between two binary vectors and is calculated by the following formulas: $J = F11 / (F01 + F10 + F11) = 2 / (1 + 2 + 2) = 0.4$

Question b: Answer

Recap:

SMC = number of matches / number of attributes = $(F11 + F00) / (F01 + F10 + F11 + F00)$

Jacaard = number of 11 matches / number of non-zero attributes = $(F11) / (F01 + F10 + F11)$

Hamming distance = $F01 + F10$

Cosine similarity: $\cos(d1, d2) = (d1 \bullet d2) / ||d1|| \, ||d2||$.

Jaccard is similar to cosine approach because they are both exclude the absence of an element in the formulas.

Hamming distance is similar to SMC approach but in reverse order. Hamming can be calculated as: Hamming dist = Number of attributes (1 - SMC).

Question c: Answer

Since we are only interested in the number of genes they share, similarity should be taken into account. Thus Jacaard is more appropriate (Hamming works on the difference.)

Question d: Answer

Because two human beings share > 99.9% of the same genes, dissimilarity information is more valuable. Thus, Hamming distance measure should be used.

Question 3. Chapter 3, Exercise 8

First, we need to understand how the boxplots are constructed, how to interpret them and finally answer the question based on the interpretation.

Median (red middle line -): divides the sorted data into two equal parts. Half of data values are greater than or equal to current value and half are less.

Interquartile range (IQR): the middle rectangle box represents half of the data values, it ranges from lower quartile to upper quartile.

Lower quartile (Q1): represents 25% of the data values that fall below lower quartile.

Upper quartile (Q3): represents 75% of the data values that fall below upper quartile
Whiskers: represents data values that fall outside the interquartile range.
Outliers: are values that much smaller or greater than the rest of data values and are identified by values > Q3 + 1.5 IQR or values < Q1-1.5IQR where Q1,Q3,IQR are the lower quartile, upper quartile and interquartile range respectively.

The purpose of boxplot is to measure the spread of data values and detect outliers. Short box indicates that data values are very close to each other, while long box (or whiskers) shows that data values are spread out. The values of data are considered to be symmetrically distributed when we see boxes are even in size.
Thus, Figure 3.11 shows that ***sepal length and sepal width are relatively symmetrically distributed.*** Long whiskers of sepal length and outliers of sepal width indicate some skewness of the tails. Petal length and width are relatively skewed since boxes are uneven.

Question 4: Answer
Team website: https://alex-nguyen.github.io/CS5331Team1/
Team repository: https://github.com/Alex-Nguyen/CS5331Team1
Source code:
https://github.com/Alex-Nguyen/CS5331Team1/tree/master/Projects/Source%20code
Data:
https://github.com/Alex-Nguyen/CS5331Team1/tree/master/Projects/Data