Question 1

a)Continuous,quantitative,ratio

b)Discrete(Assuming people rank the brightness from 1-10),qualitative,ordinal

c)Continuous,quantitative,ratio

d)Continuous,quantitative,ratio

Question 2

(i)cosine = (x,y)/($||x||||y||$) = 0/($\sqrt{2} * \sqrt{3}$)=0

correlation = co-variance(x,y)/(standard deviation(x)* standard deviation(y)) = -1.2/$\sqrt{1.2 * 1.2} = -1$

Euclidean = $\sqrt{\sum(xi - yi)^2} = \sqrt{5}$

(ii)cosine = (x,y)/($||x||||y||$) = 14/$\sqrt{520}$ = 0.61

correlation = co-variance(x,y)/(standard deviation(x)* standard deviation(y)) = 0/0(undefined)

Euclidean = $\sqrt{\sum(xi - yi)^2} = 9$

b.

No, cosine similarity does not take the length of the two data objects into account when computing similarity.If the cosine similarity is 1, the angle between x and y is and x and y are the same except for length.

c.

When the mean of each of the two vectors is zero,for example,mean(x)=mean(y)=0,then corr(x,y)=cos(x,y).

Question 3

(a)

Jaccard.Because similarity depends on the number of characteristics they both share, rather than the number of characteristics they both lack. Jaccard is appropriate for such data.

(b)

(i)

Cosine. Because cosine is more suitable for sparse document data where only scaling is important, while correlation works better for time series, where both scaling and translation are important. Furthermore, similarity depends on the number of characteristics they both share and cosine is appropriate for such data.

(ii)

Strength: Cosine similarity does not take the length of the two data objects into account when computing similarity. But Euclidean distance might be a better choice when length(In this problem, the number of distinct books available) is important.Furthermore, Euclidean distance is sensitive to both the scaling and translation operations.

Weakness: Euclidean distance is most appropriate when two data vectors are to match as closely as possible across all components. But the vectors in this problem are sparse vectors, the performance of Euclidean distance might not be very well.

Question 4

(a)

If the months were in proper order, the temperature time series will be looked like a smooth mountain. High temperature in summer and low temperature in winter and spring. Since the months were not in proper order, there would be sudden rises or falls in temperature.

(b)

All the formulas for these three measures – correlation, L1 distance, and Euclidean distance – will not be changed by the order of the columns.

Question 5

The first scheme is stratified sampling and would be more stable. Same number of objects is guaranteed to get from each group. For the second scheme, which is the simple random sampling scheme, the number of objects from each group will be different. The variance of the first scheme is always smaller than the variance of the second scheme since the latter has to count the variance between the different groups.

Question 6

In order to be a metric, there are three properties need to be satisfied:

1. Positivity

2. Symmetry for all x and y.

3. Triangle Inequality for all points x , y , and z.

(a)

It is not a metric because d(x, y)=0 only if x=y is not always satisfy. When x = -1 and y = 1, $|x^2 - y^2| = |1 - 1| = 0$

(b)

It is not a metric because d(x, y)=0 only if x=y is not always satisfy. When x1 = x2 = 2 and y1 = y2 = 1, $(x1 * x2)^2$ + $(y1 - y2)^2$ = 16 + 0 = 16 $\neq$ 0.

(c)

It is a metric.1 and 2 are easy to prove. For the Triangle Inequality, d(x,z) $\leq$ d(x,y)+d(y,z).

If x = (x1,x2,$\cdots$,xn), y = (y1,y2,$\cdots$,yn), z = (z1,z2,$\cdots$,zn) , then d(x,z) counts at how many xi are different from zi.

Now we assume that x1 $\neq$ z1, it cannot happen that x1 = y1 and y1 = z1 at the same, so either x1 $\neq$ y1 or y1 $\neq$ z1. In either case, the RHS(d(x,y)+d(y,z)) also increases by one. So for each mismatch, LHS will increase by 1 and RHS will increase at least 1. We can conclude that this formula is a metric.

Question 7

(a)True. The mean and standard deviation are strongly affected by outliers.In the standardization, the mean is always replaced by the median and the standard deviation is always replaced by the absolute standard deviation

(b)False. The correlation between(1,1,1,1) and (2,2,2,2) is 0/0(undefined).

(c)False. It depends on the data type. Cosine is more suitable for sparse document data where only scaling is important, while correlation works better for time series, where both scaling and translation are important.

(d)True. Like correlation, mutual information is used as a measure of similarity between two sets of paired values that is sometimes used as an alternative to correlation, particularly when a nonlinear relationship is suspected between the pairs of values..