4.

a. 1) Hamming Distance: It can be used as it is applied in the situation having to find similarity between 2 vectors of the same length. Hence in the case of two binary vectors, it can be implemented to find how similar or dissimilar they are, therefore concluding how far they are from each other.

   2) Covariance: The property, displaying the change in an attribute in comparison with another attribute is termed as covariance. In covariance, the sign is considered more important than the magnitude of the value. In case of variables A, B and C, in:

   Case 1: Covariance of A and B is -1: This depicts that on increasing one variable, the other variable is affected inversely, i.e. on increasing A, B will decrease and vice versa.

   Case 2: Covariance of B and C is 20: This depicts that on increasing one variable, the other variable will increase as well, i.e on increasing B, C will also increase and vice versa.

   3) Case Study: Data duplication is repetition of data points in a sample. Let us consider a scenario of a fintech company, where the HR department is conducting a survey for medical insurance policy. The data is collected in various manners, interview, paper forms, online document form, etc. The data collected was then integrated into a database for analytical purposes and health insurance selection; at the time of integration the company found many points of intersection in the data points.

   Cause of duplication: This could be because of employees taking the survey form, interviews with an HR representative or filling out the same form multiple times with different ids (work and personal).

   Detection: We can detect the duplicate data by observation of the data points, identifying primary keys (identification numbers).

   Solution: We can reduce or eliminate the problem of data duplication by applying constraints on attributes that are specific to a person (Employee id), so while entering data into the data store, we must check the value in the database, if the entry doesn't exist in the database, we add it else, ignore it.

   Pros: Backup of data, better knowledge retention (data collected from multiple sources can help in deducing and better understanding)

   Cons: More storage utilization, Hindrance in analysis, repetition could bias the output of an analysis, It costs lot of utility to handle data duplication

b. 1) Noise: it is described as data that has been modified or corrupted and deviates from the usual data points expected in the data pool due to tampering of the data by outside interference.

Noise in theory is not a desirable outcome in any data, while in the real world scenario can become a necessary factor.
Justification:
If we are training a model to function on real world data, as in the real world computation, there are scenarios in which we can't avoid noise coming into the data. Thus, in these cases, training the data with noise, helps in the model to understand and adjust accordingly for real world scenarios.

2) Outlier: an outlier is a data point that is considerably at a larger distance from the other data points in a sample of data points.These data points are natural and have not been modified by external tampering.

Outliers can be detected by multiple methods:
I. Sorting: In this we can arrange the data points and then flag the ones which show extreme values.
II. Calculation of Z Scores: By calculating this statistical metric, a person can calculate how far the given data point is from the others and be able to deduce the ones which are outliers.
III. Clustering algorithms: We can use unsupervised clustering algorithms in order to group data points and then understand the groupings and detect the points lying far from the centers of the clusters, and can also work with visualization for better understanding.
Example: K-Means
IV. Interquartile Range: We can calculate the interquartile range and declare outliers to the points that fall out of it

Noise Vs Outliers:
Noise is addition of meaningless data in the actual data. Noisy data is data that is corrupted, or distorted, or has a low signal-to-noise ratio.
Noise is undesirable or unwanted distortion in the data, it may or may not be random.
Noise can be used for analysis of data as it is present in the data collection and may be useful for real time analysis eg: Hindrance in call recordings

An outlier is an actual data point that is observed while data collection that lies at an abnormal distance from the distribution of the points in the dataset.
An outlier data point is a value that differs significantly from other data points in the dataset
Outliers can be disregarded as it differs significantly from other data points.

3) In this specific case, the data generated would be noise as for some of the measurements, the result was hindered by an external source. A noise is generated when the data is externally modified, in this case, the ground reducing the electric charge is the factor causing the creation of noisy data.