Some sketch answers and pointers for the sample exam. These are not the only ways to answer each question, but provide some idea of points to consider.

1a) XML allows one to define vocabulary of elements/attributes, whereas for HTML this is fixed and uses a vocabulary suitable for presenting in a browser. XML better for generic applications (e.g. ChemML) not dependent on presentation. HTML well suited to presentation. [Note that HTML is a subset of XML, so one could argue that when using HTML one is also using XML]

b) JSON relatively easier for data interchange by Web servers, more natural to represent hierarchical data than CSV.

c) No reliance on single point of control/authority, which is susceptible to failure/attack/scalability issues. Data is also more transparent on the blockchain (every node has a copy)-> increased trust.

d) Compute mean of column age, restricted to rows which are Male. Such a mean incorporates the background knowledge about the missing value's gender, providing a more specific estimate. However sample size is lower and estimation may be less reliable.

2a) From lecture notes on box plots

bi) Need to explain whether relationship between variables is linear or non linear and why. For a non linear relationship such as Age and Height, the (normalised) mutual information is more appropriate, since Pearson correlation won't be able to detect it. Could draw your estimate of the relationship (curve) between a person's Age and Height to support the reasoning

bii) Could argue either way here. The examples given are not very extreme, so throwing the information away could be viewed as harsh and might result in misleading analysis. It would of course depend on the population being analysed (the range of values for people in Japan would be very different from those in the USA)

biii) More reasonable computation of distance between objects (stop large scale features dominating) - useful in k-means clustering or k-nn. Feature scale may also be more interpretable.

3a) Allows immediate visualisation of the dataset. Helps show the cluster structure, helps show the overlap between classes, helps identify potential anomalies, extreme individuals from each class.

b) VAT might reveal more clearly how many clusters there are and their respective sizes.
More difficult to relate VAT info to class structure. VAT provides less idea about *why* an instance is different/similar to other instances.

c) The 90% estimate would be biased, since the testing data (class label info) was looked at when doing feature selection.
This provided information to the feature selection process that should not have been seen. (like seeing the final exam before it is held). Consequently the model that was trained using the results from the feature selection was developed on information that should not have been seen. The reported accuracy will thus likely be over optimisitic.

d) Could be domain knowledge, or evaluating accuracy using different choices of k and choosing the one that works best.

4) a) see lecture notes
b) Technique used to improve efficiency in record linkage of large datasets. Blocks are formed based on some property of each record (e.g. first letter of surname), then only blocks with matching properties are compared.
c)
A code which can be added to a document as a “signature”. This can be used to verify that a particular person signed/authorised the document (only the person who knew both the public key and the private key). Generation of the digital signature relies on public key cryptography, where the person signing has both a public key (known to all) and a private key (known only to them). Digital signatures facilitate trust and verifiability.

d) A model for data anonymisation, following on from k anonymity (an individual should be indistinguishable from at least (k-1) other individuals on the non sensitive attributes). Furthermore, there should be at least l different values for the sensitive attribute. This reduces the risk of privacy attacks on data which only satisfies k anonymity.
e) A table satisfies k- anonymity if every record in the table is indistinguishable from at least k − 1 other records with respect to every set of quasi-identifier attributes; such a table is called a k- anonymous table.

5a) Break each string into its two grams, e.g.

wrangling-> wr, ra, an, ng .....
wrapping-> wr, ra, ap ....
Use Dice coefficient for similarity

b) 2*6/(7+9)

c) Explain how string information is represented in bloom filter (generation of 2 grams, hashing of 2 grams with multiple hash functions). Explain how a single bloom filter might map to multiple possible input strings, can't easily reverse engineer.

d) Extra string that is appended to the information being encoded, so that hashed value is not susceptible to a dictionary attack (need to explain dictionary attack).   The two parties doing the linkage would agree on a salt, the 3rd party would not know it.