Solution 1:

1) The conclusion is incorrect. Since the tree structure is built recursively, the algorithm does not necessarily identify the optimal tree with lowest empirical risk on the training data. This lies in the nature of greedy optimization procedures. Empirical risk minimization (ERM) is only performed to identify *the next* splitting rule, and not entire sets of subsequent splitting rules.

2) CART does automatically select features for splitting nodes if they lead to an expected reduction in empirical risk. Irrelevant features are therefore more likely to be picked less often for split rules in model construction. (Of course, the subject of assessing feature importance is left for the chapter on random forests.) However, one could gain a rough understanding of a feature's relevance by looking at how often it was picked for splitting a node. However, this kind of "split rule selection frequency" does not necessarily relate to a feature variable's contribution to ERM.

3) CART can perform automatic feature selection by remembering surrogate splits in an extra step in model construction. Per default, the `rpart` package retains up to 5 surrogate splits. For each split rule, a surrogate split rule that leads to sorting observations into child nodes in a similar way is retained. These surrogate splits can then be used to "guide" observations through the tree even if they have some missing feature values. Therefore, CART is generally-speaking well-suited to handle missing observations.

4) The number of possible split points evaluated per feature variable is equal to the number of different values the respective feature has in the training data minus 1, e.g., a numerical or categorical variable with 4 different values in the training data has 3 potential split points. (Actually, for a continuous feature there is an infinite amount of possible split points, but there are just 3 which lead to different results for the training data.) As each feature variable can be used for the split point, one needs to sum over the feature variables in the data set.

$$\text{Number of possible split points} = \sum_{j=1}^{p} (\text{number of different values in the training data}_j - 1) \leq 3 \cdot (n-1)$$