Question 1

In this histogram, we can see that there are instances that are out of 3 STDs without the need to calculate. These instances are the ones that have a value under 0 and over 100. Because these values are impossible these instances should be deleted. However, there are a few values that are outside of 2 STDs but, in this case, they are realistic values so they should be kept as part of the dataset.

Question 2

To start with, we can see that there are students that submitted their questionnaire in 0 seconds which means that they either didn't have a time and those instances were filled with zeros or that they closed the questionnaire as soon as they opened it. Either way, I think these instances should be deleted if the rest of their attributes don't provide any relevant information. In the case that they do, I would impute them. We can also see that many instances are out of 2 STDs but I believe that they should be kept because they could represent students with a learning difficulty. If you deleted these instances, this minority wouldn't be represented in the analysis of the data.

Question 3

In this histogram, we can see that there are 6 instances with an impossible value (at this moment in time). The rest of the values are within 2 STDs and are representative. However, these 6 days that had a temperature of 100ºC, should be deleted because they aren't representative.

Question 4

In the history of the NBA, there is only one player under 5ft 5in, so it is certain that those instances under 5ft 5in should be marked if they aren't Muggsy Bogues (his height is 5ft 3in). Given that these instances represent less than 5% percent of the population, it would be okay to delete them from the dataset.

Question 6

Despite the majority of movies on Netflix having a very similar length of around 95 minutes, there are some values outside of 2 and 3 STDs. However, having investigated a little, it looks like there is a possibility of there being at least 100 short movies under 50 minutes across the whole of Netflix's library. So, in this case, I think that all the values are representative and should be kept in the dataset.