Problem 1

An example where this might post an issue would be a dataset where the data portrayed by the scatterplot appears to be curved. As mentioned in class, a dataset where this would be an issue would be one mapping human ages, to their heights. Given that people stop growing after a certain period in time, the graph would likely resemble a concave down curve. This is because children rapidly grow up until their 'teens', but then stop growing as much. This poses an issue for any linear regression analysis, given that the relationship between the height and age variables are not entirely straightforward, and cannot be interpreted as linear. For this reason, in that case, it would be best to use polynomial regression analysis.

Another example would be a dataset comprised of mostly outliers (i.e., unexpected values skewing the results). A dataset consisting of days of each month and their average temperatures for that day would be a good example. Because of global warming, there have been instances of unexpected, record highs and lows in temperature. This would skew the data and affect the overall analysis.

Lastly, one final example would be that of the Trovan drug, tested by Pfizer in Nigeria back in 1995. In February 1996, the leading physician of Pfizer's Trovan project, Scott Hopkins, came to realize that Nigeria was currently experiencing the largest meningitis outbreak ever recorded in their region. Seeing this as an opportunity to test their Trovan pills on Nigerian children sick with the disease, as well treat those who were ill, Hopkins nonetheless was presented with an issue. Although testing their drug in Nigeria could yield significant results, it was still a risky opportunity to explore– especially because, as Trovan belonged to a class of antibiotics known as quinolones. Although effective, quinolones were also found to have many extreme side effects– such as the eroding joint cartilage, impairing bone growth, rupturing tendons, as well as liver damage. Moreover, some research also suggested that quinolones could impair skeletal development in young children, which could potentially cause the many children sick with meningitis to be faced with this impaired growth. As Trovan could therefore cause all of these side effects, and potentially kill some of the patients (particularly children) being treated with it, it certainly would not have been morally correct to inflict those individuals to a haphazard, rushed, likely unsafe drug trial. Ultimately, however, Pfizer went through with testing the drug in Nigeria, resulting in many deaths due to the medication's side effects. For this reason, this is another example of how outliers can pose an issue to linera regression analysis. Although the outliers in this dataset were the severe side effects and reactions to the drug, since the majority of users did not not experience these, analysis would have shown that the drug would be safe enough to use– a conclusion which ultimately cost many lives.

Problem 2

The basis for this argument is that there is no way to ethically use statistical methods, specifically, linear regression, given that some outliers in data will be given more weights than others. The respondent to the author then goes on to state that the question of using weights is of "moral significance" especially if each outlier represents a person – thus implying that different people would therefore be weighted differently. However, the original author goes on to close the discussion post by saying that typically, these statistical methods are used for "mathematical considerations rather than applied considerations", and that therefore, this issue isn't of the most pressing significance.

Personally, I agree with the respondent's argument. Although it is true that statistical methods are typically used more for mathematical considerations, it is still important to always take into account the ethics of any computation. The results of a decision problem can have negative effects on any group of people, no matter how small, and it is thus important to always take this into consideration– especially when weights are being used on outliers. Because, it is true, an outlier could come to represent a specific individual, and there are always moral repercussions to take into account when weighting groups of people differently. I feel that being this ethically aware is the mark of a great statistician – a sense of social responsibility, aside from mathematical.

Problem 3

a. I think that it's honestly very difficult to pinpoint a sole individual or team at fault for this. Although the machine learning researcher (and their team) would certainly be at fault for testing their model appropriately to ensure this wouldn't happen, or at the very least have taken this possibility into account, I also do believe that the company is, even in the slighter bit, at fault. Although they did hire the machine researcher to do their job, as the brand backing the software based on the model, it is ultimately their responsibility to fully assess any possible repercussions it could have, given that it would only reflect poorly on the company. A team of statisticians should have examined the model in depth, as it is difficult for all of this to only fall upon one person. It is difficult to take every edge case into account, especially without any help or supervision.

b. I believe that the responsibility falls on both the corporation and individual(s) creating the automated systems. Given that the corporation oversees the creation of that system, and is ultimately responsible for negative impacts (if any) it might have on their users, they most definitely are responsible for ensuring that the systems are consistently checked for being morally up to standards. Additionally, I believe that the machine learning researchers or statisticians working on the system are also responsible, given that they

themselves are the reason for its creation. They also have the responsibility to consider any ethical implications their system may have.