Approach and Intuition:

The purpose behind this task is to try to predict the correct flow value for flow values which we think are erroneous. For this task, there were certain intuitions which helped us make the decision. We thought that the flow values for an erroneous detector might be close to flow values of nearby detectors. This is based on the assumption that on an average around the same amount of flow would be along different lanes. Thereby if a detector saved erroneous values, a nearby detector might have saved a value which is close to what the actual flow value of the erroneous detector might be. Another consideration we took into account was that the closeby flow values (based on time) might help in detecting the correct value of an erroneous flow detected by a detector. If the detector has not been working for a long time, this might not help though. We also considered that the value might actually be correct. So there isn't much of a necessity to change it at all. In fact, it has been assumed that most of the data is correct. There are only a few erroneous values which need to be corrected.

Implementation Steps:

We implemented three methods and assigned different confidence levels to each of the methods we implemented. Based on the confidence level of a method for a value, a higher priority was assigned to the flow value of that method while we tried to predict the correct flow value.

- We implemented a linear regression on the nearby lane value to predict the flow value. The confidence [C1] was assigned based on the nearby flow value's probability of being correct.
- We assigned weights w1 [= c1/(c1+c2)] and w2 [= c2/(c1+c2)] to the preceding row's flow value (whose probability is c1) and the following row's flow value (whose probability is c2) [since we assumed that the timestamps of nearby rows are usually in increasing order]. Based on the weights we calculated the new flow value. We assigned the confidence [C2] to be the minimum of the probabilities of the preceding row and following row.
- We also considered the fact that the actual flow value might be correct. So we assigned the confidence [C3] to be the original probability of that row.

- We merged the multiple methods by assigning weights to each of the methods. [W1 = C1/(C1+C2+C3) etc.]. Based on the weights, we gave a higher priority to those methods which have a higher probability value.
- Since most of the flow values are correct, we considered to correct only extreme cases (and those flow values which were above a certain threshold - based on the histogram we had plotted for flow values).

Challenges and Self Evaluation:

- For the nearby timestamp method, if a detector wasn’t working for a long period of time, very low confidence results for this method was obtained.
- For the nearby lane method, erroneous values often contained very high values, so the predicted value also tended to be high (since it was a linear regression).
- We assumed that we didn’t need to correct all the values since most of the flow values are correct and that we could correct values for some extreme cases. Because of this, some erroneous flow values might be missed.

Improvement of Final Results:

- For certain cases, the values predicted by all the methods were very low. In such cases, we thought we could replace it with the median flow value.
- We made a manual correction for some cases:
    - If occupancy was 100 and speed was 0, we made the flow as 0 ○
      If occupancy was 0 and speed was 0, we made the flow as 0