Approach and Intuition:

The purpose behind this task was to generate the probability distribution function for the set of variables (flow, speed and occupancy) pertaining to a particular zone. We assumed that if the data followed a gaussian distribution, we could calculate the probability distribution function for it. For computing the probability distribution of a single variable we take the mean and variance as parameters. In this case, there are multiple variables which need to be taken together (since there might be correlations between the different variables). To cope with multiple variables, we have assumed that our data follows a multivariate normal distribution (which is a generalization of the one-dimensional normal distribution to higher dimensions). The parameters needed to compute the probability distribution function were the mean vector and the covariances between the different variables.

Implementation Steps:

For computing the probability distribution function we used multivariate normal distribution. We computed the mean vector by taking the mean values of each of the variables (flow, speed and occupancy). We computed the covariances between the different fields based on the method cov() present for dataframes in pandas (a package in python). We fed the mean vector and covariances to the function multivariate_normal (in the scipy.stats package). This function computed the probability distributions for each of the vectors (of flow, speed and occupancy) present in the data given for each of the zones.

Challenges and Self Evaluation:

Some of the challenges we face if we follow this approach are:

- **Since we are taking all variables into consideration while computing the probability distribution function, there may be cases where a high probability may be assigned to a vector even if one (or more) of the variables are significantly different from the mean.**
- **If we follow this approach, there are some vectors which should get a value of 0 (for ex. If the flow is negative then it should get a value of 0 but it doesn't). To handle these cases, we need to have an idea of which vectors are definitely erroneous and we may need to manually assign a probability of 0 to those.**

- **We are assuming that the number of erroneous vectors are small. Hence we compute the mean of all the values. If the number of erroneous vectors are large, then the probability distribution function can go on the wrong track.**

Improvement of Final Results:

For improving the final results we manually assigned a probability of 0 to the vectors with negative flow, speed and occupancy values. We thought of assigning a lower value to the vectors with outliers (in one of its variables) but didn't do it since the outliers might be important.